UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

TEMA:

Análisis de la demanda de clientes bajo factores significativos – Cluster

AUTOR (ES):

Criollo Montiel, Sonia Elena

Galvez Guillen, Kristel Elizabeth

Trabajo de titulación previo a la obtención del título de

LICENCIADA EN NEGOCIOS INTERNACIONALES

TUTOR:

Lcd. Lucin Castillo, Virginia Carolina

Guayaquil. Ecuador

22 de agosto del 2026

UNIVERSIDAD CATÓLICA

DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de integración curricular fue realizado en su totalidad por **Criollo Montiel, Sonia Elena** y **Galvez Guillen, Kristel Elizabeth,** como requerimiento para la obtención del título de **Licenciados en Negocios Internacionales**

TUTOR (A):

f. 

Eco. Virginia Carolina Lucin Castillo, Mgs.

DIRECTORA DE LA CARRERA

f. _____

Ing. Hurtado Cevallos, Gabriela Elizabeth, Mgs.

Guayaquil, a los 22 del mes de agosto del año 2026

UNIVERSIDAD CATÓLICA

DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Nosotro/as, **Criollo Montiel, Sonia Elena** y **Galvez Guillen, Kristel Elizabeth**

DECLARAMOS QUE:

El trabajo de Integración Curricular, **Análisis de la demanda de clientes bajo factores significativos – Cluster,** previo a la obtención del título de **Licenciadas en Negocios Internacionales**, ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias bibliográficas. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Titulación referido.

Guayaquil, a los 22 del mes de agosto del año 2026

LOS AUTORES:

f. _____

Criollo Montiel, Sonia Elena

f. _____

Galvez Guillen, Kristel Elizabeth

UNIVERSIDAD CATÓLICA

DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

AUTORIZACIÓN

Nosotro/as, **Criollo Montiel, Sonia Elena** y **Galvez Guillen, Kristel Elizabeth**

Autorizamos a la Universidad Católica de Santiago de Guayaquil a la **Publicación** en la biblioteca de la institución del Trabajo de Integración Curricular, **Análisis de la demanda de clientes bajo factores significativos – Cluster,** cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 22 del mes de agosto del año 2026

LOS AUTORES:

f. _____

Criollo Montiel, Sonia Elena

f. _____

Galvez Guillen, Kristel Elizabeth

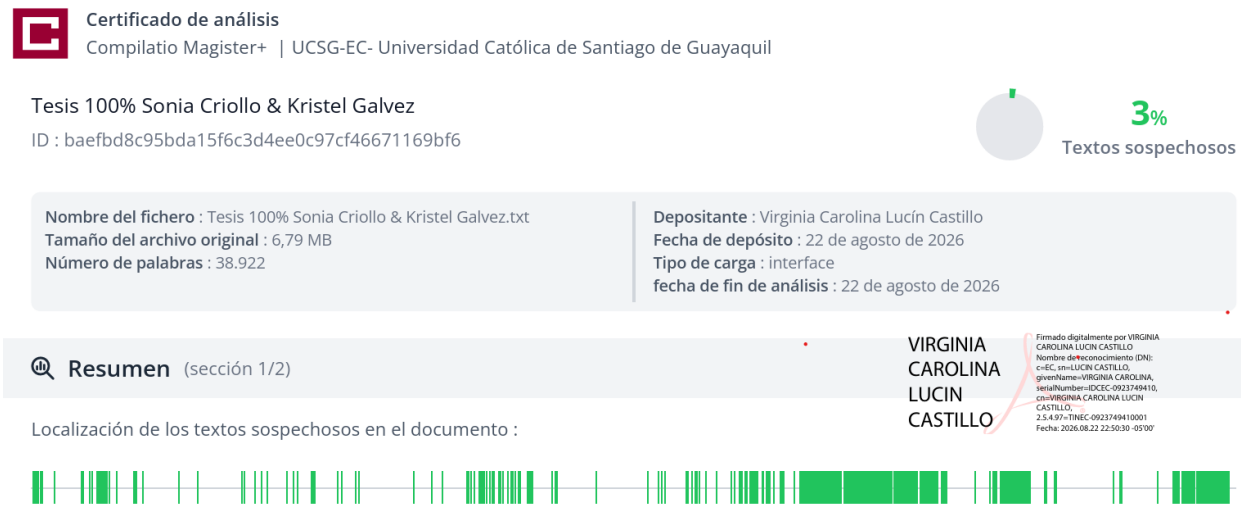
UNIVERSIDAD CATÓLICA

DE SANTIAGO DE GUAYAQUIL

FACULTAD DE ECONOMÍA Y EMPRESA

CARRERA DE NEGOCIOS INTERNACIONALES

REPORTE COMPILATIO



Eco. Virginia Carolina Lucin Castillo, Mgs

AGRADECIMIENTO DE KRISTEL GALVEZ GUILLEN

Agradezco a Dios por permitirme seguir este camino, por darme fuerza en los momentos difíciles

A mis padres Jorge Gálvez y Fátima Guillén, a quienes debo mi más profundo amor y gratitud, por confiar siempre en mí, por enseñarme el valor de la responsabilidad, la perseverancia y el esfuerzo, y por acompañarme incondicionalmente en cada etapa de mi vida y de mi formación profesional. Este logro también es de ustedes.

A mi abuela Mercedes Guillén y a mis hermanos Jorge Gálvez y Santiago Gálvez también por todo su apoyo y por estar en cada etapa de mi camino.

También agradezco a mi tutora Carolina Lucín y al Ing. Félix Carrera, por su orientación, paciencia, conocimiento y apoyo en el desarrollo. Sus observaciones y sugerencias fueron cruciales para completar este proceso académico.

Gracias a las personas que estuvieron en este viaje conmigo durante este tiempo, especialmente a mis amigas de toda la vida y amigos de la universidad en diferentes momentos de este viaje en particular. Gracias por las charlas, risas, apoyo, y por hacer que incluso los días más difíciles fueran más fáciles de vivir.

A esa persona que ha sido parte de mi vida desde el año 2024, y ha sido muy importante para mí,

Y finalmente, a mi compañera de tesis, Sonia Criollo, por haber compartido todo esto conmigo hasta ahora. Lo hicimos juntas y hoy tenemos el mejor resultado de todo ese esfuerzo y meses de trabajo.

A cada una de estas personas, gracias por ser parte de esto y por un logro que marca el final de una etapa de aprendizaje, desafíos, crecimiento y recuerdos que siempre llevaré conmigo.

DEDICATORIA DE KRISTEL GALVEZ GUILLEN

Dedico este trabajo con todo mi amor a mis padres, Jorge Gálvez y Fatima Guillen quienes han sido mi ejemplo, mi fortaleza y mi apoyo incondicional en cada etapa de mi vida. Gracias por creer en mí, por animarme a seguir adelante y por enseñarme que, con esfuerzo y perseverancia, los sueños pueden hacerse realidad.

A mi abuela, Mercedes Guillén, por su amor, sus consejos y por ser una persona tan importante en mi vida.

Este logro lleva una parte de cada uno de ustedes, y se los dedico con todo mi corazón.

AGRADECIMIENTOS DE SONIA CRIOLLO MONTIEL

En primer lugar, agradezco a Dios por permitirme pasar por esta gran experiencia universitaria.

A mi padre Juan Criollo que a pesar de que al principio no creía que esta carrera y esta universidad sea para mí siempre me apoyó, a mi madre Glennis Montiel por ser mi inspiración de nunca rendirme.

A mis hermanos Freddy, Jeison y Frank por aconsejarme y a pesar de que no vivamos juntos han sido parte esencial para seguir adelante.

Quiero agradecer a mi profesor Félix Carrera por enseñarme las lecciones que trae la vida y demostrar que uno puede con más y a mi tutora Carolina por estar pendiente en este proceso.

Agradezco a mis amigos que conocí en la universidad, especialmente a mis chicos de independientes economía por enseñarme a trabajar en equipo, amar la política universitaria y formar una identidad.

A mi amiga Pole, por permitirme conocerla, escuchar mis problemas, conocer a su familia y darme la mano cuando más lo necesitaba.

Agradezco a mis amigas Ámbar y Ana por estar presente en los momentos tristes, felices y apoyarme en mis locuras.

A mi amigo Sebastian, por ser como un hermano para mí y aconsejarme de la manera más sincera posible.

A mi compañera de tesis Kristel porque esto lo hicimos juntas y sin ella no hubiera sido lo mismo, ella es mi inspiración y este logro lo tendré siempre en mi corazón.

DEDICATORIA DE SONIA CRIOLLO MONTIEL

Dedico este trabajo a mis abuelos, especialmente a mi abuelo Manuel Santos sé que él hubiera estado orgulloso de este logro, le dedico mi esfuerzo y dedicación a ese hombre que me cantaba canciones de pequeña y me defendía cada vez que podía, un beso al cielo mi querido abuelito.

ÍNDICE

Tabla de contenido

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Esta investigación analiza la demanda de los clientes mediante el agrupamiento de factores significativos en dos sucursales de un supermercado ubicado en la provincia de El Oro. Esta investigación se motiva por el hecho de que los consumidores no tienen el mismo comportamiento de compra independientemente del volumen comprado, la frecuencia de compra, el uso de descuentos y créditos, el tipo de cliente y otras características comerciales. Solo las ventas totales son suficientes para entender el rendimiento de las ventas, pero no los perfiles que realmente sostienen y explican la demanda. Por esta razón, la idea principal es estudiar la demanda a través de factores agrupados importantes para comprender las características del consumidor e identificar el segmento con mayor impacto en las ventas. La investigación se lleva a cabo en un marco cuantitativo y analítico basado en la información numérica de las transacciones comerciales de las dos sucursales. La base de datos está compuesta por ventas, clientes, productos, descuentos, créditos y características de las transacciones. Antes de aplicar los modelos, se realizó una preparación y limpieza de los datos que incluyó la identificación de registros duplicados, operaciones canceladas, errores de digitalización, valores atípicos, organización y estandarización de las variables. Esto nos permitió tener información consistente para realizar análisis estadísticos y aplicar técnicas de aprendizaje automático. Finalmente, realizamos un análisis exploratorio de datos para entender la distribución y el comportamiento de las variables comerciales. Para la segmentación, utilizamos K-Means, un método de aprendizaje no supervisado que ayuda a la clasificación de observaciones de calidad similar. Las variables fueron estandarizadas para evitar que sus diferentes escalas afectaran la distancia desde la cual se utilizó el algoritmo. Como resultado, establecimos tres grupos de clientes en los algoritmos. Los grupos 0 y 1 mostraron un comportamiento relativamente similar con clientes siendo principalmente individuales y con bajo

volumen de ventas y cantidad de productos. Sin embargo, el grupo 2 tenía un mayor volumen de ventas y más productos, alto uso de crédito y clientes corporativos. En otras palabras, este es el segmento que es más intensivo y comercial por naturaleza. También se utilizó el Análisis de Componentes Principales (PCA) para mostrar los segmentos gráficamente y visualmente para ver cómo difieren los diferentes segmentos. Los resultados revelaron que hay tres grupos distintos, aunque la dispersión del grupo 2 es más dispersa de los otros grupos y se superpone con algunos de ellos. Luego, utilizamos modelos de regresión lineal múltiple del grupo con el volumen de ventas como variable dependiente y variables como precio de venta, cantidad de producto, descuento, crédito, factor tiempo, tipo de cliente y ubicación de ventas como variables explicativas para indicar que los factores asociados con la demanda son más importantes para el segmento del estudio. En conclusión, los resultados muestran que los clientes no deben ser analizados como un solo grupo ya que hay varios patrones de comportamiento en la demanda. Al combinar K-Means y PCA, podemos identificar perfiles comerciales y entender los factores que influyen en las ventas. Esta información se utiliza para crear estrategias diferenciadas para la lealtad, promociones, crédito, inventario y planificación comercial; apuntando a clientes de alto valor y creando incentivos específicos para esos segmentos de clientes que tendrán potencial de crecimiento. A través de técnicas analíticas y conocimientos, los datos transaccionales se convierten en la base para una toma de decisiones más precisa basada en el comportamiento real de los consumidores.

Palabras clave: segmentación de clientes; análisis de la demanda; K-Means; análisis de componentes principales; regresión lineal múltiple; supermercados; comportamiento del consumidor.

ABSTRACT

This research analyzes customer demand by grouping significant factors across two branches of a supermarket located in El Oro Province, Ecuador. The study is motivated by the fact that consumers do not exhibit the same purchasing behavior regardless of purchase volume, purchase frequency, use of discounts and credit, customer type, and other commercial characteristics. Although total sales provide information about sales performance, they are not sufficient to understand the customer profiles that actually sustain and explain demand. Therefore, the main objective is to analyze demand through the grouping of relevant factors in order to understand consumer characteristics and identify the segment with the greatest impact on sales. The research follows a quantitative and analytical approach based on numerical information from commercial transactions at the two branches. The database includes sales, customers, products, discounts, credit, and transaction-related characteristics. Prior to applying the models, data preparation and cleaning were performed, including the identification of duplicate records, canceled transactions, data entry errors, outliers, and the organization and standardization of variables. This process ensured consistent information for statistical analysis and the application of machine learning techniques. An exploratory data analysis was also conducted to examine the distribution and behavior of the commercial variables. For customer segmentation, K-Means was applied as an unsupervised learning method for grouping observations with similar characteristics. The variables were standardized to prevent differences in scale from affecting the distances used by the algorithm. As a result, three customer groups were identified. Groups 0 and 1 showed relatively similar behavior, consisting mainly of individual customers with low sales volumes and product quantities. In contrast, Group 2 presented higher sales volumes and larger product quantities, greater use of credit, and a predominance of corporate customers, making it the most commercially

intensive segment. Principal Component Analysis (PCA) was also used to graphically represent the segments and examine the differences among them. The results revealed three distinct groups, although Group 2 showed greater dispersion and some overlap with the other groups. Subsequently, multiple linear regression models were applied to the selected group, using sales volume as the dependent variable and variables such as selling price, product quantity, discount, credit, time factor, customer type, and sales location as explanatory variables to determine the factors most strongly associated with demand within the study segment. In conclusion, the results show that customers should not be analyzed as a single homogeneous group, since different demand behavior patterns can be identified. By combining K-Means and PCA, it is possible to identify commercial profiles and understand the factors associated with sales. This information can support the development of differentiated strategies for customer loyalty, promotions, credit management, inventory, and commercial planning, focusing on high-value customers and establishing specific incentives for customer segments with growth potential. Thus, analytical techniques and data-driven insights transform transactional data into a foundation for more accurate decision-making based on actual consumer behavior.

Keywords: customer segmentation, demand analysis, K-Means, principal component analysis, multiple linear regression, supermarket, consumer behavior.

INTRODUCCIÓN

Actualmente, en el contexto comercial, las empresas se enfrentan a desafíos como expandirse de manera eficiente en un sector competitivo al momento de adquisición de posibles clientes, en consecuencia, las empresas enfrentan una necesidad constante de adaptación e innovación estratégica buscando herramientas que posibiliten la obtención de información detallada y análisis de comportamiento del mercado (Kotler & Keller, 2016).

Las empresas tienden a experimentar una creciente competencia donde la oferta del mercado y la retención de clientes es cada vez más complicada. Esto ha desencadenado en exigencia por parte de los consumidores los cuales cada vez encuentran otras opciones que puede satisfacer sus necesidades, al mismo tiempo la importancia no solo recae en el diseño de producto y el precio sino en cómo estos factores influyen en la atracción de nuevos consumidores. En Ecuador se presenta una alta competitividad en la industria en donde las compañías buscan fortalecer vínculo con los clientes y búsqueda de nuevo grupo de consumidores. Esto ha causado la exigencia de comprender las preferencias de consumo, con el fin de desarrollar decisiones comerciales más sostenibles (Kotler & Keller, 2016; Kumar & Reinartz, 2018).

En un informe de estudio de tendencias de los consumidores de Ecuador 2025 explica las principales intenciones de compra de los consumidores en donde lidera: Viajar fuera del país con 66% y comprar electrodomésticos para su hogar con 65% (Advance Consultora & MarketWatch Ecuador, 2025).

En adición, este estudio menciona las nuevas formas de compra mediante tiendas virtuales las cuales resultan en ser una forma más viable para los consumidores de adquirir productos. Esta nueva forma de compra ha comprometido a varias empresas a entrar en proceso de adaptación.

Ejemplificando lo anteriormente mencionado, el supermercado Mi comisariato debido a esta actualización realizó una alianza estratégica conjunto a la plataforma Pedidosya, organización líder dedicada en la venta digital de productos y entrega personalizada a domicilio, resultando en una estrategia en donde el supermercado y la plataforma amplíen su presencia comercial (Zambrano, 2022).

Por ende, desde una perspectiva analítica del consumidor, esta plataforma genera grandes volúmenes de datos clave, tales como las preferencias de productos, horarios de consumo, ubicación geográfica, etc. Este mismo método puede ser implementadas por las empresas para identificar patrones de comportamiento, segmentar consumidores y desarrollar estrategias de fidelización más efectivas (Kumar & Reinartz, 2018; Ngai et al., 2009).

Estos avances han permitido a empresas a evolucionar y adaptarse y por otro lado una nueva idea de negocio permitió la creación de varias plataformas que facilita la compra a los consumidores. Es importante que las empresas puedan analizar y segmentar mejor a sus consumidores para tomar mejores decisiones estratégicas que ayuden a ampliar datos de esos clientes sabiendo exactamente los deseos y necesidades actuales.

Por consiguiente, desde un punto de vista comercial, la demanda significa el interés de los consumidores por obtener ciertos productos o servicios, siempre y cuando dicha inclinación esté respaldada por la capacidad de compra. En esta perspectiva las organizaciones comerciales deben evaluar más de una variable las cuales son las preferencias de los consumidores y la intención de compra real (Kotler & Keller, 2016).

Por consiguiente, entender la demanda constituye en un eje principal para el desarrollo de planificación comercial dado que permite identificar las preferencias y comportamientos dentro de

un mercado seleccionado. Este conocimiento de la demanda facilita la evaluación de oportunidades comerciales y mejora el tipo de método orientado a cubrir las expectativas de los clientes de manera más perspicaz (Ludeña, 2022).

Aunque esta información parece simple de comprender lleva meses de preparación de análisis de mercado para entender las respuestas de las personas a través de sus emociones, mentalidad y comportamiento que influyen directamente en las decisiones de compra y en los patrones de consumo observados dentro de un mercado (Anderson, 2024).

De esta manera se llega al siguiente punto que es entender lo que es el análisis de comportamiento del consumidor utilizada como un método relevante para interpretar cómo las personas escogen, utilizan y evalúan los productos disponibles. Este proceso conlleva diferentes elecciones de compra. Este proceso requiere diversos factores económicos, sociales y conductuales.

El entendimiento del análisis adquiere una importancia estratégica dentro de las empresas debido a que tienen la información necesaria sobre las variables esenciales para realizar una compra, así como la compresión de las dinámicas del mercado. Asimismo, el primer análisis disminuye la incertidumbre empresarial y mejora los procesos de planificación y gestión comercial (Kumar & Reinartz, 2018; Ngai et al., 2009).

La planificación empresarial está vinculada con la capacidad de las organizaciones para comprender el entorno del mercado y anticiparse ante el cambio de comportamiento de las personas. En consecuencia, el conocimiento del concepto de demanda, utilizar un análisis de mercado e identificar los patrones de consumo contribuyen al desarrollo del crecimiento y fortalecimiento empresarial (Ludeña, 2022).

Teniendo en cuenta estos conceptos, el orden es esencial para llegar al éxito en una empresa debido a que el conocer la demanda, los clientes, análisis de mercado y segmentación de mercado que se explicará a continuación.

La segmentación de clientes se ha convertido en un instrumento común para agrupar a los compradores según sus características, conductas y actitudes. Este procedimiento permite a las compañías entender y percibir con mayor precisión a los diversos grupos que integran su mercado y fijar estrategias comerciales más adecuadas para cada uno (Kotler & Keller, 2016; Mailchimp, 2023).

En un principio, los métodos de segmentación utilizan criterios demográficos, geográficos, socioeconómicos y actitudinales (Fischer & Espejo, 2020).

Las metodologías de segmentación tradicionales se han basado en variables demográficas, geográficas, socioeconómicas y conductuales. Sin embargo, el crecimiento de las tecnologías informáticas y el incremento de la disponibilidad de datos ha favorecido una evolución importante en las técnicas y procedimientos de segmentación. En la actualidad, las compañías pueden estudiar información vinculada con hábitos de compra, frecuencia de consumo y patrones de comportamiento, obteniendo una interpretación más detallada de sus clientes(admin, 2022; Ed Lorenzini, 2023).

Para concluir este apartado, esta evolución ha permitido que las empresas entiendan mejor el comportamiento de los consumidores y el porqué de sus elecciones al momento de elegir la compra de un producto o servicio, por lo tanto, es de suma importancia realizar un uso debido de esta herramienta conociendo como la era digital y otro tipo de herramientas ayudaría de una forma más rápida a identificar los tipos de consumidores por grupos con el propósito de crear estrategias

prácticas que enganchen hasta un grupo de consumidores prospectos a ser fieles a una marca o producto.

El análisis de datos se ha transformado en una herramienta imprescindible para las organizaciones debido a que permite entender las variables que influyen en las decisiones de compra de los consumidores. La información extraída de transacciones comerciales es evidencia sobre preferencias y necesidades de los clientes que a largo tiempo contribuyen a una mejor comprensión del mercado.

A partir del procesamiento de la información de los clientes, es factible identificar tendencias y comportamientos de consumo, este conocimiento facilita el diseño de estrategias de comportamiento de compra, así como mejora la competitividad aumentando la tasa de éxito en ventas, fortaleciendo el vínculo entre la organización y los consumidores.

Los usuarios prefieren marcas que ofrecen experiencias únicas, esto aumenta las opciones de compra son esenciales para comprender el comportamiento del consumidor los grandes volúmenes de datos ayudan a analizar estos comportamientos y encontrar patrones (Dois Z, 2025).

El análisis de datos genera beneficios competitivos y permite que las marcas establezcan oportunidades de mejora y de esta forma encaminarse hacia el éxito.

Los supermercados están empleando el uso de BIG DATA para entender mejor a los consumidores y brindar un valor segmentado al de la competencia, estos llegan a diferenciarse por el mercado retail Latinoamericano debido a su continuo desarrollo y evolución (PREDIK, 2025).

Las tendencias actuales del consumo demuestran una elección por realizar las compras en un único lugar, dada la importancia que los usuarios consideran a la comodidad y optimización del tiempo, lo que reduce la preferencia de desplazarse hacia otros centros de compra, esto ha generado

un reto importante para las empresas de sector retail, los cuales se enfrentan a una gran competencia de formatos de compra integral, definidos por ofrecer una amplia variedad de productos en el mismo punto de venta, con el objetivo de abastecer múltiples necesidades de los clientes en una sola visita (Wanatop, 2022).

En la actualidad, la optimización de la gestión del inventario constituye un elemento determinante para el desempeño de las compañías comerciales, ya que garantiza la capacidad de respuesta de los productos precisados por los consumidores en el momento adecuado en este contexto, la analítica de datos cumple un papel importante al facilitar el equilibrio entre las variaciones de la demanda, los patrones estacionales y los tiempos de reposición de los productos (Pacifico, 2026; Tardivon, 2022).

Mediante la integración del sistema de punto de venta. (PSO). Los supermercados tienen la capacidad de recopilar y analizar información proveniente de las transacciones realizadas por los clientes, lo que les permite obtener indicadores relevantes para la gestión de inventarios entre éstos se encuentran la identificación de niveles excesivos de inventario, la velocidad de rotación de los distintos artículos el tiempo promedio de caducidad de los productos el comportamiento de la demanda durante eventos especiales y la influencia de la estacionalidad sobre los patrones de consumo (PREDIK, 2025).

El aprovechamiento de herramientas de analítica de datos contribuye a una administración más eficiente de los recursos, permitiendo optimizar la disponibilidad de productos y mejorar la capacidad de respuesta de las empresas ante necesidades cambiantes de los consumidores (Pacifico, 2026).

La industria de productos de consumo masivo en América Latina ha experimentado importantes desafíos debido al incremento sostenido de los niveles de inflación registrados en los últimos años ha generado cambios significativos en el comportamiento de los consumidores y en las decisiones económicas adaptadas a los diferentes agentes del mercado (Dekimpe & van Heerde, 2023).

En un escenario caracterizado por la inflación constante, los consumidores evidencian una mayor participación en sus decisiones de compra, priorizando aspectos como el precio, las promociones y la relación entre calidad y costo simultáneamente. Las compañías pertenecientes al sector de consumo masivo se han visto en la necesidad de adaptar sus estrategias y reevaluar las ofertas que ofrecen para adecuarse a las circunstancias económicas actualizadas (Enrique Manzur et al., 2013)

La identificación de oportunidades y el conocimiento de las tendencias del mercado adquieren una mayor relevancia para las organizaciones ya que les permiten fortalecer su capacidad de adaptación y desarrollar estrategias que contribuyan a enfrentar de manera más eficiente los desafíos presentes en la industria del consumo masivo.

El sector de los supermercados ha atravesado importantes procesos de transformación y reorganización frente a la creciente participación de grandes cadenas y la comercialización de alimentos, las empresas del sector respondieron mediante estrategias, integración y concentración empresarial se forman redes de abastecimiento y de mayor tamaño fortaleciendo los procesos de fusión y adquisición los cambios en los sistemas de transporte y distribución impulsan el desarrollo de la industria logística más eficiente, permitiendo que los productos alimenticios lleguen al consumidor a través de cadenas de suministro cada vez más amplias y sofisticadas (Thomas Reardon et al., 2003).

Por ello, el estudio tiene como finalidad analizar y segmentar los clientes de cada sucursal de la misma cadena de supermercados que están ubicados en diferentes sectores de la ciudad de Pasaje, en la provincia del Oro con esta segmentación buscamos las características y comportamientos de compra de los clientes o consumidores con la finalidad de comprender los diferentes comportamientos de ambas sucursales y generar información que ayude a la implementación de estrategias comerciales más efectivas.

OBJETIVOS

Objetivo general

Analizar la demanda de clientes mediante factores significativos clusterizados en las dos sucursales de un supermercado, para reconocer características de consumidores y establecer los segmentos con mayor influencia en las ventas.

Objetivos específicos

- Determinar las variables comerciales relevantes que influyen en la demanda de los clientes de las dos sucursales del supermercado.
- Segmentar a los clientes por técnicas de *clustering*, teniendo en cuenta sus características de compra, consumo, descuentos, crédito y participación en ventas.
- Interpretar los segmentos obtenidos para establecer su relación con el comportamiento de la demanda y orientar estrategias comerciales de fidelización y estímulos comerciales.

Pregunta de investigación

¿Cómo influye la identificación de perfiles de clientes mediante técnicas de clusterización en el análisis de la demanda y el pronóstico de ventas de las dos sucursales del supermercado?

PROBLEMÁTICA

La industria de los supermercados se encuentra actualmente en una industria de intensa competencia, cambios constantes en las preferencias de los consumidores y comportamientos de compra más diversos. Ahora hay múltiples opciones para productos de consumo masivo, por lo que factores como el precio, la variedad, las promociones, la disponibilidad y la facilidad de compra impactan en el comportamiento de compra de los consumidores. Por lo tanto, entender el comportamiento de la demanda es crucial para la planificación y la toma de decisiones en este sector empresarial.

La actividad de los supermercados tiene la particularidad de que sus clientes no tienen el mismo comportamiento de compra. Algunos a menudo compran menos, pero compran más y más de ello, y así tienen un mayor volumen de ventas. También hay compradores que solo compran cosas en descuento y promociones y algunos que son más o menos constantes en sus hábitos de consumo. Estas diferencias significan que la demanda está compuesta por diferentes perfiles de clientes que pueden ser muy diferentes en su comportamiento.

En este sentido, analizar solo las ventas totales de un cierto período nos permite conocer el rendimiento general del mercado, pero no necesariamente explica las características de los consumidores responsables de generar estas ventas. El mismo nivel de ingresos puede estar compuesto por clientes con comportamientos completamente diferentes en términos de frecuencia de compra, cantidad adquirida, tipos de productos, volumen de consumo o respuesta a promociones.

La demanda, en el sector de los supermercados, está influenciada por factores económicos y comerciales como precios, descuentos, promociones y el poder adquisitivo de los consumidores. Las personas pueden disminuir la cantidad que compran, cambiar de marca, optar por productos

de menor precio o interesarse más en las promociones ante las fluctuaciones económicas. Por lo tanto, el análisis de la demanda debe tener en cuenta diferentes factores que afectan las decisiones de compra.

El comportamiento del consumidor también es muy diferente con respecto a la frecuencia y el gasto. Hay clientes que aumentan su frecuencia de compra pero realizan compras menos valiosas, mientras que otros compran con menos frecuencia pero gastan más. Así que identificar diferentes perfiles de consumidores nos permite ver qué grupos están más involucrados en las ventas y cuáles pueden necesitar apoyo comercial para mantener su lealtad alta.

El problema es cuando estos consumidores se ven de manera general y no se diferencian los diferentes patrones que componen la demanda. La falta de perfiles puede dificultar saber qué grupos realizan compras frecuentes, cuáles adquieren mayores volúmenes, qué consumidores son sensibles a los descuentos y cuáles mantienen más continuidad en sus compras.

En la actualidad, los supermercados generan mucha información sobre sus operaciones comerciales. Los registros de ventas pueden incluir cosas como la fecha de compra, productos comprados, cantidades, precios, tasas de descuento y características del cliente. Esta es una buena fuente de información para examinar el comportamiento de la demanda en los supermercados, ya que permite el análisis de características reales y observables de las decisiones de compra.

Sin embargo, los datos no necesariamente significan que se utilizarán para entender al consumidor en detalle. Cuando la información transaccional se analiza solo de forma agregada, es posible identificar cuánto se vende o qué productos tienen mayor demanda, pero es más difícil determinar qué tipos de clientes generan estos resultados y qué características tienen en común.

Esto deja claro que debemos pasar de un análisis general de la demanda a una segmentación basada en el comportamiento real de los consumidores. La segmentación tradicional consideraría factores geográficos, demográficos o socioeconómicos, pero los datos transaccionales pueden usarse para incluir variables directamente relacionadas con el comportamiento de compra como frecuencia, cantidad adquirida, volumen de ventas, categorías de productos, descuentos y condiciones de pago.

De esta manera, puede haber consumidores frecuentes con alto valor comercial, compradores de alto volumen que son menos frecuentes, consumidores sensibles a las promociones y consumidores ocasionales que pueden aumentar su consumo. Cada uno de estos grupos puede tener diferentes características y puede ejercer una influencia diferente en la demanda.

El problema, por lo tanto, no es solo cuánto compran los consumidores, sino los factores que permiten diferenciarlos y cómo estos factores se relacionan con el comportamiento de la demanda. Variables como la frecuencia de compra, cantidad de productos, tipo de productos, volumen de ventas, precio, descuentos y crédito pueden proporcionar información importante para identificar patrones y establecer grupos de consumidores con atributos similares.

Basado en esto, las técnicas de agrupamiento son una forma alternativa de analizar la heterogeneidad entre los consumidores. Las técnicas de agrupamiento permiten formar grupos según las similitudes encontradas en los datos sin la necesidad de categorías. De esta manera, se pueden crear perfiles basados en patrones de comportamiento reales y no en clasificaciones generales.

Las técnicas de aprendizaje automático que se pueden usar en la segmentación tratarán con variables como la recurrencia de compra, la cantidad gastada y el volumen comprado para identificar grupos de consumidores sin definirlos de antemano. Esto nos permite identificar

patrones que no son visibles desde el análisis de ventas y luego pueden ayudarnos a entender la demanda con mayor precisión.

Finalmente, encontrar estos segmentos también puede ayudar a entender el comportamiento de las ventas. Algunos grupos pueden tener una mayor participación en el volumen vendido, mientras que otros pueden presentar potencial para aumentar su consumo. Entender cómo se crea la demanda y qué aspectos están más relacionados con las variaciones comerciales puede ayudar a identificar los tipos de demanda y qué características tienen mejores vínculos con la variación comercial de esos segmentos.

Por lo tanto, el sector de los supermercados se enfrenta a la dificultad de entender la diversidad de comportamientos de compra cuando los consumidores se analizan de manera general. La ausencia de segmentación basada en factores comerciales clave impide la identificación de perfiles, patrones de consumo y grupos con diferente participación en las ventas.

Con esto en mente, es necesario analizar la información transaccional utilizando técnicas de agrupamiento, donde los consumidores se agrupan según características similares y luego se determina cómo estos perfiles se relacionan con la demanda. De esta manera, el análisis puede pasar de una visión general de los consumidores a una interpretación más detallada de quiénes componen la demanda, cómo se comportan, qué factores los diferencian y qué segmentos tienen un mayor impacto en las ventas.

JUSTIFICACIÓN

La importancia de la presente investigación radica en el entorno caracterizado por la globalización y competitividad, los constantes cambios en las preferencias de los consumidores y las organizaciones requieren conocer de manera precisa las características y necesidades de sus clientes para diseñar estrategias comerciales más efectivas. Desde un enfoque de marketing, las organizaciones requieren conocer con mayor precisión a sus clientes para diseñar estrategias comerciales que respondan a sus necesidades, hábitos de compra y nivel de contribución a la demanda. El análisis de la demanda de clientes bajo factores significativos clusterizados contribuye a una herramienta muy relevante, ya que ésta permite identificar qué grupos de consumidores y qué características tienen comportamientos similares facilitando una comprensión del mercado y una toma de decisiones adecuada (Kotler & Keller, 2016; Kumar & Reinartz, 2018).

Aplicar técnicas de agrupamiento y clusterización adquieren gran importancia debido a que posibilita tener una demanda segmentada a partir de las variables significativas, permitiendo que las organizaciones optimicen la gestión de inventarios. Mejorar la disponibilidad de sus productos, fortalecer las relaciones con los clientes y desarrollar estrategias de marketing más específicas y eficientes. Asimismo, el avance de las tecnologías con información y una disponibilidad creciente de los datos que favorecen la evolución de métodos tradicionales y segmentación brindando nuevas oportunidades para tener un mejor análisis y una mejor interpretación de esta información (Ngai et al., 2009; Pacifico, 2026).

El adaptar estas metodologías analíticas y avances tecnológicos determinan la competitividad empresarial, puesto que todas las organizaciones deben responder de una manera eficiente. Las exigencias del mercado en este sentido el estudio busca tener conocimientos relevantes sobre este comportamiento de la demanda de los usuarios, teniendo información que

permita a las empresas o supermercados identificar ciertos patrones de consumo y tener ventajas competitivas en el sector en el que se desarrolla. Este enfoque permite que las estrategias de marketing respondan de mejor manera a las características reales de los clientes, debido a que se fundamentan en información observable sobre sus hábitos de compra, preferencias y niveles de consumo. El ámbito del marketing analítico utiliza los datos comerciales para que las organizaciones puedan aumentar su capacidad de responder, consolidar su posición en la competencia y guiar la toma de decisiones mediante evidencia concreta (Anderson, 2024; Kumar & Reinartz, 2018).

Las modificaciones de factores significativos y la información de clústeres aceptan una alternativa que combine el rigor estadístico con los análisis de datos que permiten obtener resultados más precisos, estas agrupaciones de los consumidores con características semejantes tienen la facilidad y el reconocimiento de tener patrones de compra y de que los usuarios decidan por preferencia, lo que también contribuye al diseño de ciertas estrategias de mercado que se basan en satisfacer las necesidades de cada usuario. Desde la perspectiva del marketing relacional, conocer los segmentos de clientes también permite fortalecer los procesos de fidelización (Durán Barros & Nieto, 2025; Jain, 2010).

Por consiguiente, esto posee una relevancia muy práctica, ya que ciertos resultados pueden servir como referencia de nuevas investigaciones para organizaciones interesadas en mejorar su rendimiento, sus procesos de planificación y gestión comercial, de esta manera el análisis de mercado y de la demanda de clientes bajo factores significativos clusterizados permitan tener una fortaleza en la toma de decisiones. En entorno de marketing estratégico, la investigación contribuye al diseño de estrategias más coherentes con el comportamiento real de los clientes, al

fortalecimiento de la planificación comercial y al desarrollo de ventajas competitivas sostenibles (Ludeña, 2022; Pérez, 2006).

MARCO TEÓRICO

Análisis de la demanda en empresas comerciales

La demanda de clientes constituye un concepto fundamental que conlleva el análisis de varios factores puesto que se estudia el nivel de consumo existente del mercado, comprender el comportamiento del consumidor en el área de los supermercados. Por ende, la demanda se puede analizar mediante las preferencias de compra, la frecuencia de consumo, el volumen adquirido y la respuesta frente a precios y descuentos (FasterCapital, 2014).

Por consiguiente, el conocimiento del tipo de mercado conlleva entender el tipo de consumidores al que se enfrenta una empresa. En el caso de los supermercados, los mercados de consumo son aquellos conformados por individuos y familias que realizan transacciones de bienes y servicios para satisfacer necesidades personales o familiares (Pérez, 2006).

La compresión del mercado implica reconocer lo que compone al consumidor ya que su participación en las actividades comerciales es primordial. En los supermercados, el mercado de consumo implica individuos e integrantes de familias que realizan compras de bienes destinados a satisfacer sus propias necesidades o las de su residencia. Según Pérez (2006) redacta que el mercado de consumo indica que los compradores finales son los que llevan a cabo las compras dirigidas a la utilización o consumo directo de los productos. Este punto explica que los supermercados van más allá de las transacciones individuales y cumplen con las necesidades recurrentes del consumo de comida, limpieza y cuidado personal al igual que alimentos, medicina y comodidad.

En el análisis de la demanda, se debe diferenciar quienes intervienen en la compra y el consumo, dado que no siempre la misma persona que compra el producto es la que finalmente lo consume, Según Hoyer et al. (2016), durante el proceso pueden elegir diferentes funciones y desempeñarse en diferentes roles, tales como quién hace la elección, quien hace la compra y quien la consume después, Dentro del contexto de las tiendas de supermercado, esa diferencia es relevante porque una persona puede comprar los artículos para ocultar las necesidades de su hogar , lo que significa que otros miembros de la familia consumen el producto adquirido , Como resultado, la demanda tiene que ser evaluada desde el punto de vista del acto de compra y también con respecto al modo en que esta demanda se satisface y la probabilidad de un eventual reclamo de devolución o venta

Teniendo en cuenta los principales conceptos, la demanda en los supermercados se refleja a través de las decisiones de compra por lo cual, se explicará el comportamiento del consumidor el cual según Solomon (2019), se enfoca en el estudio de acciones y decisiones de compra, opciones de compra, el uso del producto considerando sus necesidades, preferencias y hábitos, estas variables a su vez influyen en el área de los supermercados.

El comportamiento de los consumidores sucede desde un niño de ocho años que exige un juguete a un adulto que decide comprar su caja de cereal favorita. Las necesidades de las personas van más allá de las básicas que son comer y vestirse. Además, los deseos de los consumidores son amplios teniendo varias opciones de compra de distintas categorías.

El consumidor tiene fases antes, durante y después de la compra. La primera fase es porque el consumidor decide comprar cierto producto o servicio, como el usuario percibe cierto producto, luego cuál es la experiencia que obtiene luego de adquirirlo, si cumple con las expectativas, por último, como se desecha finalmente el producto y si cumplió su función que pretende.

Las empresas comerciales, en particular en el sector de supermercados, adoptan el comportamiento de sus clientes basándose en la demanda observada a partir de los registros de compra generados por los propios consumidores. La demanda se refleja en la frecuencia de compra, el monto gastado, el número de productos comprados, el nivel de repetición en las visitas y el patrón de consumo desarrollado a lo largo del tiempo, y es una manifestación del comportamiento del mercado (Ngai et al., 2009; Tardivon, 2022).

El análisis de la demanda ofrece información útil para la toma de decisiones empresariales, debido a que permite anticipar tendencias de consumo y conocer las demandas del cliente, de tal forma que las organizaciones pueden diseñar mejor sus estrategias comerciales y también, claro está, adaptarse a las transformaciones del mercado. Para Arellano et al. (2010), el análisis del consumidor demanda considerar las diferencias existentes entre personas y grupos, puesto que estas diferencias condicionan la forma de compra y valoración de los productos.

Factores que afectan la demanda del cliente.

La demanda del consumidor tiende a influenciarse por factores económicos externos. Estos cambios se reflejan en las empresas de consumo masivo cuando los consumidores cambian su volumen de compra, de productos, buscan otras marcas o responden de manera diferente a un descuento o promoción. En consecuencia, el análisis de la demanda debe considerar los factores que afectan la toma de decisiones de compra y se hallan a través de transacciones comerciales de la compañía. Una de las variables más relevantes es el precio, puesto que es uno de los factores que los consumidores consideran al momento de la compra de un producto (Ailawadi et al., 2001; Bell et al., 1998).

En el contexto de los supermercados, los compradores muestran interés al costo, presentación, lugar cuando se trata de una misma marca o tienda. Además, cuando se implementan descuentos o promociones en un producto, la demanda aumenta puesto que es una estrategia atractiva que atrae a los clientes. El comportamiento de compra de los consumidores va más allá de la adquisición de un bien o servicio, la disponibilidad de fondos, promociones, variedad y facilidad de obtener un producto en un punto de venta influyen en la decisión de compra. Según Rolando Arellano Cuevas et al., (2010) explican que es esencial realizar un estudio del consumidor abarcando como el marketing afecta en el comportamiento de compra. Por ejemplo, en los supermercados existe una variedad de marcas, localización conveniente esto ayuda al crecimiento de la demanda de un solo producto y mejora la satisfacción de compra.

Comportamiento del consumidor en los supermercados.

El comportamiento del consumidor es la unión de acciones planificadas, rutinarias e impulsivas en los supermercados. Así como en otros sectores, los supermercados establecen una relación constante con los consumidores por la razón de la venta de productos básicos y de uso diario (Solomon, 2019).

En los supermercados, las fases de compra que son el antes, durante y después de la adquisición de un bien y/o servicios, se ejemplifican mejor cuando las personas se abastecen, comparan marcas y aprovechan los descuentos con la finalidad de acción de compra, sin embargo, está la otra alternativa de la decisión de compra en una tienda donde se muestra el mismo producto por un precio diferente o una locación más cercana de sus viviendas. Existen tipos de consumidores y comportamientos de compra en los supermercados, se conforman por individuos que compran productos básicos frecuentemente, suelen ser consumidores como foráneos, estudiantes universitarios, etc. Por otro lado, están las familias que realizan compras en mayor cantidad

frecuentemente para cubrir las necesidades del hogar. Por último, los consumidores que compran ciertos productos que se pueden encontrar solo en ciertos supermercados (Arellano et al., 2010; Solomon, 2019).

Por ende, la demanda suele ser cambiante y saber qué tipo de cliente puede desempeñar un papel diferente en el rendimiento comercial. Estos comportamientos son visibles y facilitan la identificación de tendencias. Un grupo de clientes puede presentar un incremento de la tasa de compra, pero una reducción en la tasa de compra media del pedido, por otro lado, existen consumidores que adquieren productos con menos frecuencia, pero sus gastos son mayores. Cabe recalcar que el análisis de tipos de clientes ayuda a identificar qué perfiles mejoran el negocio y quién necesita ser incentivado para seguir siendo leal (Durán Barros y Nieto, 2025; Kumar & Reinartz, 2018).

Segmentación de clientes basada en el comportamiento y valor del consumidor

La segmentación de clientes es una estrategia utilizada para el manejo de la demanda comercial en la que se identifican y agrupan consumidores que son similares o semejantes entre sí. A través de esta herramienta, se puede reconocer la heterogeneidad del mercado y desarrollar diferentes estrategias según las necesidades y hábitos de consumo de cada grupo. En el sector del marketing, la segmentación es una herramienta para la toma de decisiones comerciales al permitir que las empresas adapten sus acciones a las características de los consumidores (Kotler & Keller, 2016).

En el supermercado, esta herramienta es particularmente relevante porque la demanda tiene diferentes patrones de compra. La caracterización de los consumidores puede establecerse a partir de variables geográficas, demográficas, socioeconómicas y conductuales. Sin embargo, si una

empresa tiene información de sus actividades comerciales, entonces el análisis puede extenderse mediante variables directamente relacionadas con el comportamiento de compra, como la frecuencia de compra, la cantidad comprada, el volumen de ventas, las categorías de productos, los descuentos, el método de pago y la sucursal donde se realiza la compra (Bell et al., 1998; Thomas Reardon et al., 2003).

Tabla 1.

Tipos de segmentación de clientes

Tipo de segmentación	**Características consideradas**	**Aplicación en el estudio**
Geográfica	Ubicación o zona	Sucursal
Demográfica	Características del consumidor	Tipo de cliente
Socioeconómica	Capacidad y condiciones económicas	Crédito, gasto
Conductual	Frecuencia, compras, descuentos	Comportamiento transaccional
Por valor	Contribución económica	Volumen y monto de ventas

Nota. Elaboración propia a partir de Kotler & Keller (2016); Kumar & Reinartz (2018) y las fuentes citadas en el estudio.

De esta manera, la segmentación de datos puede ayudar a superar una clasificación basada en características generales del consumidor. Dos clientes en la misma área geográfica o grupo socioeconómico pueden comprar cosas diferentes, y es importante tener en cuenta la información de sus transacciones comerciales. La información se utiliza para identificar patrones de consumo y crear grupos de clientes con características demográficas similares, para comprender mejor la composición de la demanda (Ed Lorenzini, 2023; Mailchimp, 2023).

En este enfoque, la segmentación conductual permite clasificar a los consumidores según su comportamiento real hacia la empresa. A diferencia de los criterios tradicionales, esta perspectiva considera la frecuencia de compra, los productos seleccionados, el gasto, la respuesta a promociones y el tipo de relación que el consumidor tiene con la empresa. Como tal, es de particular interés para las empresas con bases de datos comerciales si es capaz de identificar patrones relacionados con las operaciones comerciales (Anderson, 2024).

En un supermercado, estos patrones pueden reflejarse en diferentes perfiles de consumidores. Algunos clientes pueden realizar compras frecuentes de bajo valor, mientras que otros pueden realizar compras menos frecuentes pero de mayor volumen. Al mismo tiempo, existen consumidores que estén en ciertas categorías de productos y compren en promociones y descuentos. Esta comprensión de estas diferencias ayuda a entender la distribución de la demanda entre diferentes grupos y qué grupo tiene más ventas (Bell et al., 1998; Thomas Reardon et al., 2003).

Esta perspectiva conductual se basa en la segmentación del valor del cliente, es decir, el análisis de la segmentación del valor del cliente, que tiene como objetivo encontrar el valor económico de cada grupo para la empresa, y de esta manera, no solo la frecuencia de compra sino también el volumen de compras, el valor monetario del proceso de transacción, la cantidad de

ofertas en términos de descuento del precio recibido, el acceso al crédito y la continuidad del consumo. Esto se puede utilizar para diferenciar una variedad de clientes en función de la contribución económica actual y potencial y las estrategias de gestión de la lealtad y relación con los consumidores (Kumar & Reinartz, 2018; Ngai et al., 2009).

La combinación de la dimensión conductual y el valor del cliente permite una segmentación más completa de la demanda. Por ejemplo, se pueden identificar consumidores frecuentes con alto valor comercial, clientes que realizan compras de gran volumen pero con menos frecuencia, consumidores sensibles a los descuentos o clientes ocasionales con potencial para aumentar su nivel de consumo. Esta clasificación permite estrategias comerciales más específicas y evita las mismas acciones para toda la base de clientes (Kotler & Keller, 2016)

Además, la segmentación permite traducir la información de ventas en decisiones comerciales. Los clientes frecuentes pueden ser considerados en programas de lealtad; aquellos con alto volumen de compra pueden obtener diferentes beneficios, y los consumidores sensibles al precio pueden responder mejor a promociones específicas. Los clientes ocasionales también pueden ser un grupo con potencial de crecimiento a través de estrategias para aumentar su frecuencia de compra. Así, la segmentación ayuda a optimizar promociones, suministros, precios y gestión de relaciones con los clientes (Fischer & Espejo, 2020; Kumar & Reinartz, 2018).

Es relevante para este estudio porque permite analizar la demanda basada en información observable en la operación comercial del supermercado. Variables como la fecha de venta, la cantidad y tipo de producto, el volumen de compra, el precio, los descuentos, el crédito, el tipo de cliente y la sucursal pueden utilizarse para identificar grupos con comportamientos similares.

Se pueden utilizar técnicas de agrupamiento para identificar estos segmentos sin tener que establecer categorías rígidas, y así los patrones de consumo y las diferencias en la contribución económica de los clientes (Jain, 2010).

Por lo tanto, la segmentación de clientes basada en el comportamiento y valor del consumidor es una herramienta útil para el análisis de la demanda porque ayuda a pasar de una visión general de los consumidores a una identificación más detallada de sus patrones de compra y contribución comercial. Su aplicación ayuda a comprender la composición de la demanda y guiar la toma de decisiones en promoción, lealtad, suministro y gestión de clientes para crear una estrategia comercial que esté más cerca de las dinámicas reales del mercado (Arellano et al., 2010; Kumar & Reinartz, 2018)

ANÁLISIS DE DATOS APLICADO AL MARKETING

Actualmente, las decisiones de marketing y gestión comercial no se sustentan exclusivamente en la experiencia acumulada, la percepción o la visualización general del mercado. Los supermercados diariamente generan grandes cantidades de información a través de sus propias actividades, como riesgo de facturación, bases de clientes, ventas por producto, ventas por sucursal, montos de compra, transacciones, aplicación de descuentos y consumo frecuente. Cuando los datos están organizados e interpretados de manera adecuada, se crea un conocimiento estratégico para entender de mejor forma la conducta de los consumidores y alinear la toma de decisiones comerciales (Ngai et al., 2009; Pacifico, 2026; PREDIK, 2025).

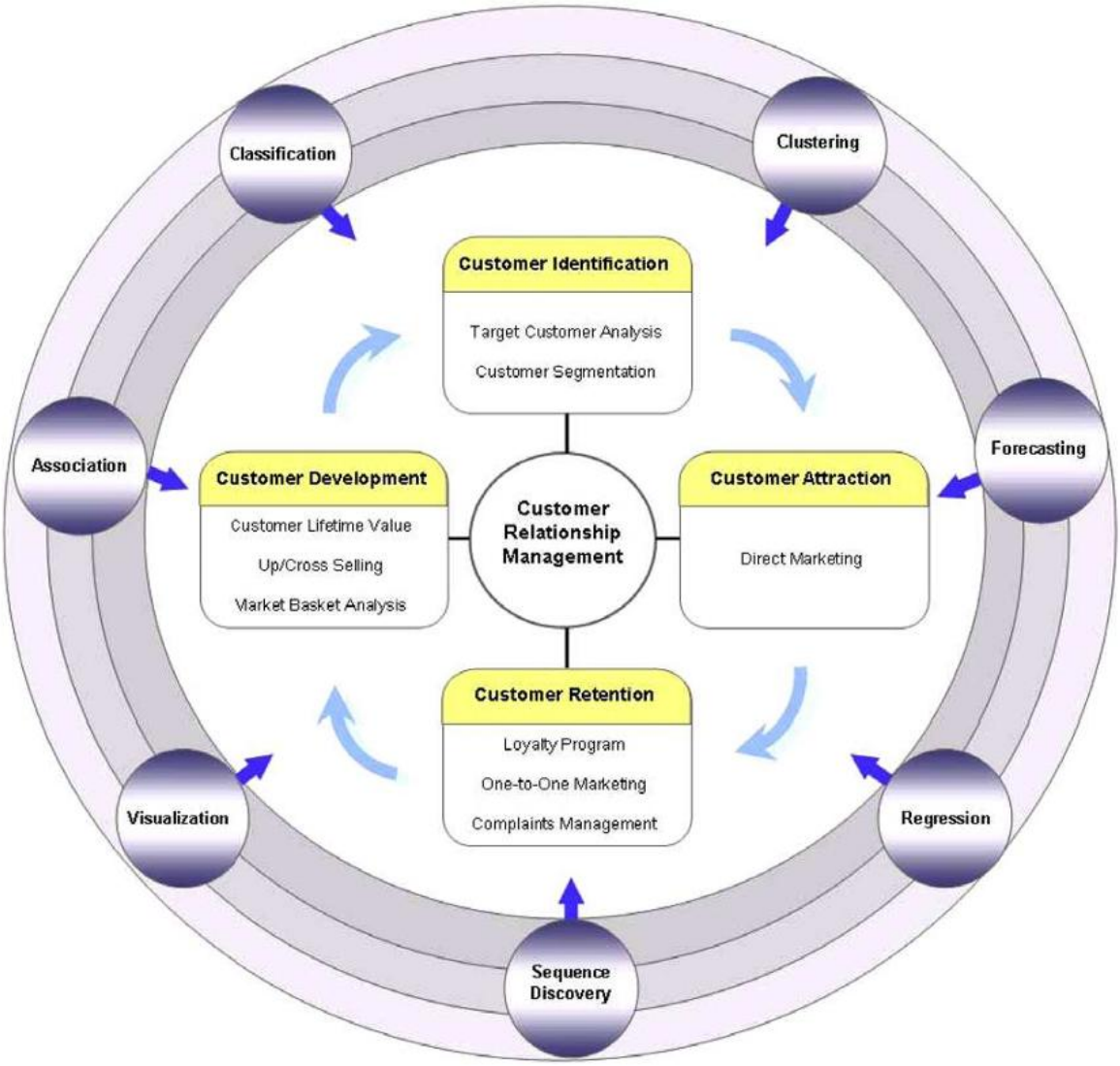
Bajo este enfoque, el análisis de datos aplicado al marketing permite sustituir acciones amplias por estrategias más ajustadas a los supermercados. Por medio de procedimientos de análisis descriptivo, estrategias de segmentación y modelos estadísticos, los supermercados o

empresas pueden detectar comportamientos recurrentes, saber qué productos tienen más demanda, identificar qué usuarios representan mayor valor comercial y conocer cómo pueden mejorar sus promociones, suministros, fidelización o precios. Puesto que el procesamiento de datos no solo permite identificar lo que paso en periodos pasados, sino también emplear tendencias próximas con base en evidencia (Ngai et al., 2009; Pacifico, 2026)

El marketing analítico y la inteligencia comercial desempeñan un rol importante en este proceso. Puesto que permite traducir los datos operativos en información estratégica para la administración de la empresa. Como alternativa a utilizar campañas o acciones estratégicas comerciales de forma generalizada, las organizaciones utilizan estrategias que se ajustan a los patrones de compra de los consumidores. La gestión de clientes mejora drásticamente, ya que permite identificar los perfiles de compradores, entender cuáles son sus necesidades y dirigir los recursos hacia iniciativas con un mayor resultado positivo. De este modo, el análisis de datos se convierte en una estrategia vital para incrementar la competitividad de las empresas. A través del estudio de la información recogida durante las ventas, las operaciones comerciales y la comunicación entre clientes, las instituciones tienen la capacidad de formular decisiones comerciales más objetivas, mitigar el desafío y potenciar su estrategia comercial (Ngai et al., 2009; Pacifico, 2026).

Figura 2.

Clasificación de las técnicas de minería de datos en la gestión de relaciones con clientes



Nota. *Tomado de Ngai et al. (2009)*

De esta manera, los datos dejan de ser simples registros operativos y se convierten en una base útil para detectar oportunidades, ajustar la oferta de productos y diseñar estrategias de marketing alineadas con la demanda y el comportamiento real de los consumidores (Pacifico, 2026).

Datos transaccionales y variables comerciales para el análisis de clientes

El análisis comercial, los datos transaccionales son información obtenida por ventas u operaciones diarias de alguna empresa, esta información permite evaluar la demanda de los clientes a partir de datos comerciales objetivos y observables reales, los artículos de mayor demanda, la frecuencia de compra, cantidad de productos y la sucursal de preferencia del consumidor (Tardivon, 2022).

Una base transaccional puede registrar variables ya sea con la fecha de compra, número de factura, código del cliente, tipo de producto, cantidades, monto de venta, sucursales, categoría de productos y tipo de pago, esta información permite clasificar las ventas y vinculadas con el comportamiento comercial de cada cliente (PREDIK, 2025).

Tabla 2.

Variables contenidas en la información transaccional

Variable	Tipo	Utilidad para el análisis
Fecha de venta	Temporal	Analizar comportamiento termporal
Cliente	Identificación	Agrupar transacciones
Producto	Categórica	Identificar preferencias
Cantidad	Numérica	Medir volumen vendido
Precio	Numérica	Analizar relación precio-

		demanda
Descuento	Binaria	Evaluar efecto promocional
Crédito	Binaria	Caracterizar modalidad de compra
Sucursal	Categórica	Comparar establecimientos
Tipo de cliente	Categórica	Segmentar consumidores

Nota. Elaboración propia a partir de PREDIK, (2025); Tardivon, (2022).

En funcionamiento de estos registros se pueden desarrollar indicadores importantes, como la recurrencia de compra, monto total comprado, ticket promedio, recencia, frecuencia, número de categorías adquiridas y local principal de compra. Estos parámetros ayudan a encontrar patrones de consumo y diferentes perfiles de clientes (Kumar & Reinartz, 2018; Ngai et al., 2009).

Para que los resultados sean precisos y confiables, la base debe estar filtrada, organizada y anonimizada, esto requiere eliminar registros repetidos, tratar datos faltantes, revisar valores atípicos, estandarizar variables, así como asegurar la confidencialidad de la información sensible antes de desarrollar procesos de segmentación. Según Anderson (2024), el análisis del comportamiento del cliente permite comprender los patrones de compra, las interacciones con la marca y las motivaciones que influyen en sus decisiones, lo que ayuda a las empresas a diseñar estrategias comerciales más efectivas

El análisis del comportamiento del cliente permite transformar los datos generados por las interacciones y operaciones comerciales en información útil para anticipar necesidades, mejorar la experiencia del cliente y apoyar la toma de decisiones estratégicas (Anderson, 2024).

Técnicas de *machine learning* aplicadas a la segmentación de clientes

La segmentación de clientes basada en *machine learning* permite categorizar consumidores según patrones de comportamiento reales, como recurrencia de compra, monto gastado y volumen adquirido. Una de las técnicas más utilizadas es el *clustering* o agrupamiento no supervisado ya que facilita la clasificación de clientes sin necesidad de tener grupos establecidos con anterioridad entre los algoritmos más utilizados están *K-means,clustering* jerárquico, DBSCAN, *gaussian mixture* model, BIRCH. Estos métodos facilitan clasificar clientes en grupos de consumidores recurrentes, alto valor, compradores ocasionales, compradores atraídos por descuento o clientes con baja recurrencia de compra, en lo que es aplicación de *retail* se comparan *K-Means*, GMM, DBSCAN, *clusterig* aglomerativo y BIRCH utilizando variables RFM, y GMM que obtienen buen desempeño (Bell et al., 1998; Durán Barros & Nieto, 2025; Jain, 2010).

Otro método significativo que destaca es RFM, que considera tres componentes importantes de la relación comercial con el cliente, frecuencia, cuántas veces compra, valor monetario y cuánto dinero aporta al negocio, este modelo se complementa con algoritmos de *clustering* para obtener la segmentación de clientes y que sean más precisos, algunos ejemplos son en estudios de segmentación bancaria y *retail*, el RFM se han implementado junto con aprendizaje no supervisado para distinguir clientes con mayor contribución al negocio, clientes en riesgo de abandono y clientes que pueden desarrollar una relación estable con la empresa (Kumar & Reinartz, 2018; Ngai et al., 2009).

Además, puede emplearse *K-Means*, que agrupa clientes con el objetivo de reducir la distancia interna entre los datos de cada bloque. Es un método de fácil aplicación y frecuente cuando se busca separa una base de consumidores en un número establecido de segmentos, la

documentación de *Scikit-learn* explica que *K-Means* separa los datos en grupos buscando mejorar la variación interna de cada clúster (Bell et al., 1998).

Clustering como técnica de segmentación

El *clustering* es un método de análisis no supervisado que permite diferenciar grupos de observaciones que comparten características similares dentro de un conjunto de datos. A diferencia de otras técnicas que se encargan de categorías previamente definidas, este método busca detectar estructuras internas utilizando los datos disponibles (Jain, 2010; Tan et al., 2021).

Desde un punto de vida metodológico, el *clustering* se aplica cuando el propósito no solo consiste en predecir una categoría ya establecida sino en encontrar grupos de elementos que tengan características similares. Según Jain (2010) explica que el análisis de agrupar se ha empleado para agrupar datos en conjuntos más accesibles de interpretar sobre todo cuando existen una gran cantidad de registros. Este uso es válido para la clasificación de clientes debido a que hay compañías que disponen de bastante información de datos de ventas con la desventaja de tener una clara clasificación de los perfiles que satisfacen su demanda.

Las ventajas del *clustering* en la estrategia comercial se da debido a la capacidad de emplearlo para crear perfiles demográficos, específicamente para la planificación de la segmentación mercadológica y la formulación de estrategias orientadas a fidelizar a los clientes. Los clusters se forman de manera aleatoria o siguiendo ciertos parámetros que se definen por variables que influyen en la decisión de compra de los consumidores, por ejemplo, frecuencias de compra, el valor adicional de los productos, etc. Las técnicas de *clustering* utilizadas son variadas, siendo principalmente los métodos clúster k-means los más populares (Durán Barros & Nieto, 2025; Tan et al., 2021).

Así mismo, seleccionar las variables, también es muy importante preparar la información antes de ejecutar el algoritmo. En técnicas de *clustering*, las variables podrían enseñarse en diferentes escalas; tales como el dinero que se ha gastado podrían ser elevado, mientras que una variable binaria relacionada con un descuento sólo tomará valores como si o no.

Si estos elementos no se normalizan, se da prioridad a las variables con valores más elevados durante el procedimiento de agrupamiento, lo que nos da el resultado de segmentos poco fiables. Por lo tanto, subrayan la importancia del método de los datos, la identificación y eliminación de *Outliers* y la modificación de las variables para ejecutar las técnicas de minería de datos (Tan et al., 2021).

La estandarización de los datos permite que las variables sean comprables y los grupos se forman basándose en similitudes verdaderas entre los grupos, En este caso particular ,es especialmente útil en estudios de clientes , ya que se agrupan variables monetarias , numéricas y categóricas , Se debe revisar la validez y una calidad de los datos para minimizar errores y garantizar la confiabilidad de los resultados dados así, el método del *clustering* comienza después de organizar y depurar la base de datos *comercial*. Finalmente, una vez obtenidos los grupos, no basta con hacer un recuerdo del número de clusters (Hair et al., 2022; Tan et al., 2021).

Es indispensable entender el contenido de cada grupo y exhibir con una comprensión clara. Las técnicas como la segmentación se valoran cuando los grupos obtenidos se facilitan medidas comerciales específicas dirigidas a diferentes segmentos de mercado, En el caso específico de la empresa de supermercados, la explicación puede señalar a identificar cuáles son los compradores más valiosos para el negocio, aquellos que ofrecen potencial para un crecimiento y, finalmente, quienes requieren incentivos adicionales para reforzar su lealtad (Durán Barros & Nieto, 2025; Kumar & Reinartz, 2018)

A través de la admiración comercial, los clusters pueden emplearse como un medio para progresar estrategias más precisas, Un bloque integrado por clientes recurrentes puede beneficiarse de medidas destinadas a incrementar su recurrencia; un conjunto de compradores ocasionales con altos montos puede obtener ayuda económica para sus futuras compras; y un cliente sensible a promociones se gestionaría mejor con descuentos especializados. Así, el *clustering* permite que las tomas de decisión comerciales se centren en observaciones detalladas y no simplemente en generalizaciones y así se basan en información derivada del comportamiento real de los compradores.

Además, el *clustering* ayuda a entender la participación de cada grupo dentro del mercado total. En compañías comerciales, no todos los clientes contribuyen al mismo nivel a las ventas ni reaccionan igual ante precios, ofertas o financiamiento. Por eso, agrupan clientes permite destacar las diferencias significativas entre ellos y estudiar cuál es su relación con la eficiencia de ventas del supermercado, Este tipo de datos pueden usarse para adaptar estrategias de fidelidad, publicidad, ofertas y atención personalizada (Kumar & Reinartz, 2018).

En el campo del *customer relationship management* (CRM), la minería de datos ha contribuido para ampliar el entendimiento de los clientes y reforzar las estrategias relativas a la relación entre el negocio y el mercado. Ngai et al. (2009) segmentación, la retención, la personalización y el análisis, Por lo tanto, el *clustering* resulta apropiado para investigación que tiene por objetivo transformar un soporte de datos comerciales en información estratégica para la dirección de la sociedad.

En consecuencia, el *clustering*, no debe limitarse a un recurso estadísticamente para crear categorías, sino debe emplearse como una metodología capaz de desvelar la diversidad de prácticas consumistas presentes en una población de clientes.

En esta indagación, la aplicación del *clustering* nos permitirá agrupar clientes con propiedades comerciales similares y, después de dichas agrupaciones, analizarlas como perfiles de necesidad o demanda, Por ende, estos datos ofrecerán información valiosa para comprender que el mercado está activo y que compradores merecen una atención y tratamiento basado en políticas y clientes adecuados para manejarse utilizando tácticas de fidelidad y beneficios comerciales

En la validación de clusters además de adecuar la base de datos para el análisis, es indispensable evaluar si los grupos de la base son coherentes y útiles para el análisis, el *clustering* no basta únicamente con procesar los datos mediante el algoritmo por lo que también se requiere reconocer los grupos de clientes que representan variaciones diferenciadas entre sí y rasgos compartidos entre sus integrantes, por ello se emplean variables de referencia como el método del codo, ayuda a determinar el número óptimo de *clústeres* y el coeficiente de silueta, que analiza en que clúster se encuentra cada cliente, esta comprobación de fiabilidad contribuye a evitar agrupaciones forzadas o con baja capacidad de representación dando mayor solidez a los resultados obtenidos (Rousseeuw, P.J, 1987; Thorndike, 1953).

Sin embargo, en las limitaciones del *clustering* se deben considerar dentro del análisis, los hallazgos varían según los criterios de análisis seleccionados, la calidad de los datos, la existencia de los valores atípicos y el número de segmentos definidos por el investigador. De igual manera algoritmos, como *K-Means*, pueden verse afectados por la escala de las variables y la manera en que contribuyen los datos, por consecuencia los hallazgos no deben entenderse de manera automática, sino que necesitan una revisión de análisis y revisión comercial que posibilita comprobar si los segmentos extraídos tienen sentido dentro de la realidad del negocio (Hair et al., 2022; Tan et al., 2021).

En el contexto de esta investigación, el *clustering* significa una herramienta adecuada para evaluar la demanda de consumidores del supermercado, ya que permite categorizar a los consumidores según los comportamientos reales de consumo. Al tomar en cuenta factores con frecuencia de compra, volumen de venta, tipo de producto, descuentos, crédito, precio y tipo de cliente, es fácil reconocer grupos con varios niveles de valor comercial. Bajo esta perspectiva los supermercados pueden saber que consumidores son habituales.

La importancia central del *clustering* no se reduce a la creación de grupos estadísticos, no únicamente a su capacidad para reformular datos comerciales en información útil para su administración. Desde los segmentos identificados, los supermercados pueden diseñar estrategias específicas para cada tipo de cliente, esto también permitirá que las estrategias de marketing se encaminan de manera más eficiente minimizando medidas generales que rara vez responden al comportamiento real del mercado.

Métodos de *clustering* aplicables al estudio

El *clustering* es una parte del aprendizaje no supervisado y se utiliza para agrupar observaciones que tienen características similares en una base de datos. A diferencia de los métodos supervisados, el *clustering* no parte de una categoría predefinida, sino que intenta descubrir patrones internos basados en la información disponible. Por lo tanto, este método es adecuado en este estudio ya que permite segmentar a los clientes del supermercado según variables comerciales observables como frecuencia, volumen adquirido, monto de ventas, descuentos, crédito, tipo de cliente y tipo de producto. De esta manera, los grupos no se establecen de manera subjetiva, sino basados en el comportamiento real registrado en la base de datos (Jain, 2010).

La técnica principal que se utilizará en este estudio es *K-means*, que es bien conocida por su uso en segmentación y variables numéricas. El algoritmo utiliza un conjunto de grupos o clusters y se asegura de que los clientes en el mismo segmento sean similares entre sí pero diferentes de los clientes en otros grupos (Tan et al., 2021). Para el ejemplo del supermercado, K-means podría ayudar a identificar grupos de consumidores como compradores frecuentes, altos montos de compra, consumidores sensibles a las promociones, clientes que son menos activos o que usan más crédito.

Una de las características clave de *K-means* es que se debe determinar de antemano el número de clusters a obtener. La decisión del número de grupos a seleccionar no debe ser arbitraria, ya que un número muy bajo de grupos ocultaría las diferencias entre los clientes, y un número excesivo de grupos haría que los segmentos fueran difíciles de entender. Para evitar este problema, se puede utilizar el método del codo y observar si agregar más clusters no está mejorando significativamente la agrupación. El coeficiente de silueta también se utiliza para verificar si cada cliente está en el lugar correcto en su grupo y si el segmento está suficientemente separado de otras partes (Rousseeuw, P.J, 1987; Thorndike, 1953).

Además de los factores estadísticos, el número de clusters debe considerarse para tener en cuenta el valor comercial de los segmentos. No solo podemos calcular matemáticamente grupos válidos de clientes, sino también permitir que se interpreten en el contexto del supermercado. Por ejemplo, una población podría consistir solo en clientes regulares y bajos montos de compra, mientras que otra podría ser compradores ocasionales y de mayor valor. O, otra población serían los clientes que hacen más descuentos o promociones. Se puede interpretar de esta manera y los resultados del *clustering* pueden utilizarse en decisiones comerciales.

Otro método alternativo es el *clustering* jerárquico pues permite explorar la estructura de agrupamiento de los datos sin un número fijo de clúster desde el comienzo. Este método organiza las observaciones progresivamente y puede representarse mediante un dendrograma que expone cómo se combinan los clientes según su nivel de similitud. En este estudio, el *clustering* jerárquico puede ser utilizado como una herramienta de apoyo para evaluar los resultados obtenidos por el K-means y comprobar si el modelo de la segmentación está bien formulado (Jain, 2010).

La ventaja principal de este método es la exploración inicial, ya que nos permite observar las relaciones existentes entre clientes o grupos antes de elaborar un segmento final. Puede ayudarnos a determinar si existen agrupaciones naturales en la base de datos y si estos segmentos destacan en particular. Sin embargo, al igual que el método *K-means*, el *clustering* jerárquico requiere elecciones metodológicas como medidas de distancias, tipo de enlace y punto de corte del diagrama de caída. Por lo tanto, sus resultados deben analizarse con criterios técnicos y comerciales, evitando interpretaciones automáticas (Tan et al., 2021)

Antes de aplicar cualquiera de estos métodos, es necesario procesar la base de datos. En esta etapa, los registros incompletos, datos duplicados, errores de tipeo, valores atípicos y variables pueden distorsionar el análisis. También es importante estandarizar las variables cuando están en diferentes escalas. Por ejemplo, el monto de ventas puede expresarse en valores monetarios altos mientras que el descuento puede ser valores binarios. Si estas diferencias no se corrigen, algunas variables pueden tener más peso en el algoritmo y afectar la formación de los clusters (Tan et al., 2019).

En la gestión de relaciones con clientes o CRM, el *clustering* también es crucial ya que nos ayuda a comprender mejor el comportamiento de los clientes y a adaptar nuestras estrategias de segmentación, retención y personalización para satisfacer sus necesidades. Los grupos que son

similares en aspectos comerciales pueden ayudar a la empresa a tomar decisiones más adecuadas para cada segmento. Los datos de ventas no solo reflejan compras pasadas; además, proporcionan información para fortalecer lealtad, implementar campañas promocionales, gestionar suministros y proveer atención personalizada (Ngai et al., 2009).

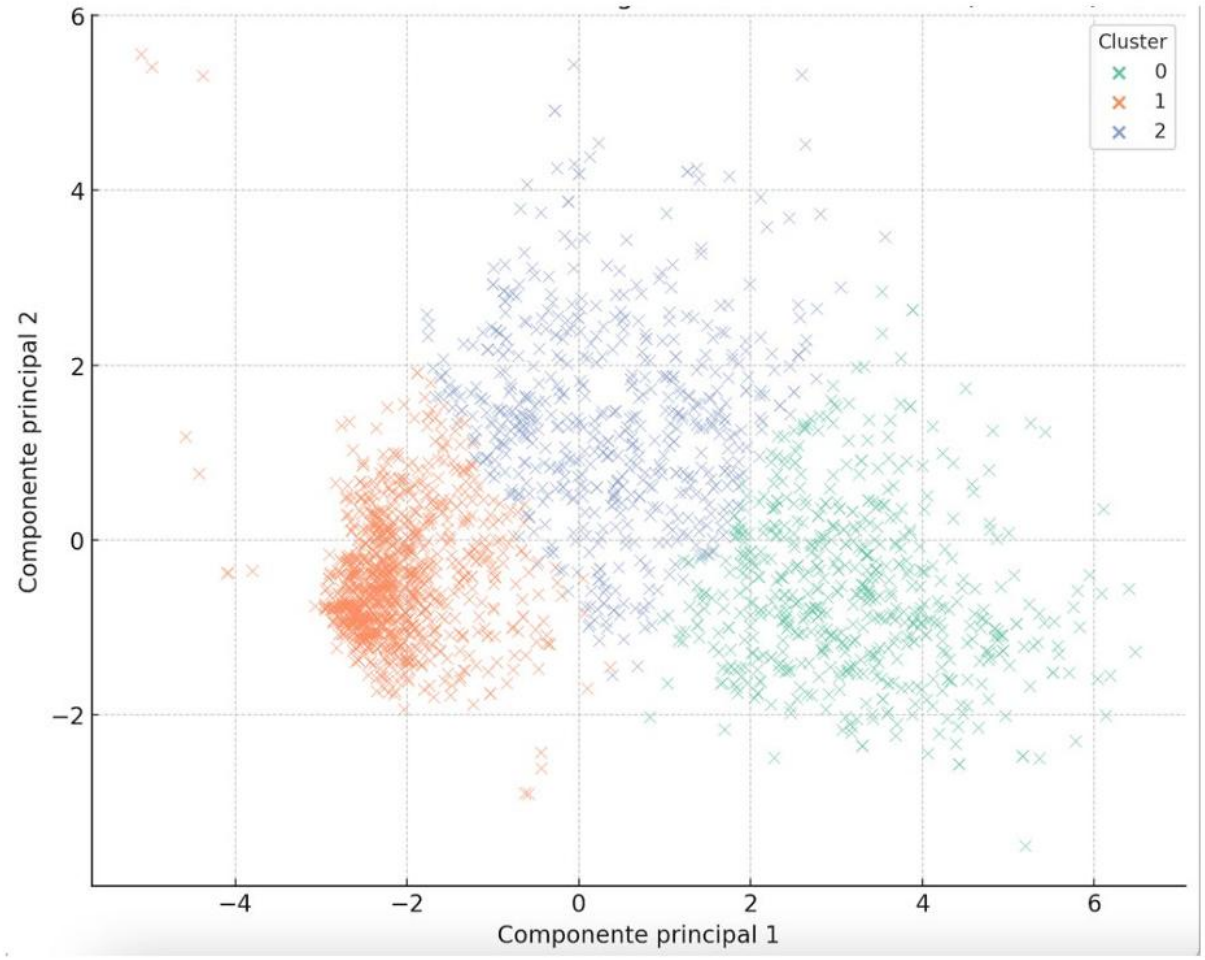
En conclusión, este estudio propone usar el método K-means como la técnica principal debido a su aplicación en bases de datos con variables numéricas y su facilidad de interpretación en estudios de segmentación de clientes. El *clustering* jerárquico podría emplearse como técnica adicional para explorar la estructura de los datos y comparar la consistencia de los clusters encontrados. El número de clusters debe estar respaldado por el método del codo, el coeficiente de silueta, el tamaño de cada grupo y la interpretación comercial de los segmentos. Por lo tanto, el *clustering* no solo se trata de clasificar clientes, sino también de convertir datos comerciales en información estratégica para la gestión del supermercado.

Aplicaciones del *clustering* en *retail* y supermercados

El *clustering* se ha utilizado en contextos comerciales como una técnica de segmentación para una mejor comprensión del comportamiento del cliente basado en datos reales. En la industria minorista, esta herramienta es valiosa ya que generalmente hay información demográfica, transaccional y promocional, pero no siempre se utiliza para desarrollar estrategias comerciales en el contexto analítico y empresarial. En este sentido, Durán Barros, Ernesto José & Cecilia (2025) señalan que la segmentación mediante *clustering* no supervisado permite la identificación de grupos homogéneos de clientes y personalizar las campañas de marketing en entornos minoristas multicanal.

Figura 3.

Visualización PCA de clientes K-means.



Nota. *Tomado de Durán Barros & Nieto (2025).*

Un importante uso de los clústeres en la venta minorista es encontrar compradores de gran valor. Al entender los ingresos, el gasto en una categoría específica de productos, la frecuencia de compra y las preferencias de categoría de producto, podemos detectar sectores que podrían producir mayores beneficios para la organización. En este estudio, se descubrió un grupo denominado Premium Multicanal, que tiene altos ingresos, un alto gasto en un producto específico y una tendencia de compra diversificada por canales. Este tipo de hallazgo sugiere que el *clustering*

es efectivo para diferenciar a los clientes más valiosos y se deben utilizar estrategias de fidelización más específicas para satisfacer sus necesidades (Durán Barros & Nieto, 2025).

Otra aplicación importante es la distinción entre clientes frecuentes, ocasionales y de bajo consumo. En las empresas de supermercados, no todos los consumidores están dispuestos a hacer lo mismo cuando se trata de ventas; algunos realizan compras repetidas, otros compran esporádicamente y algunos tienen baja conversión o baja rentabilidad. El *clustering* ayuda a distinguir estos perfiles para garantizar que la compañía no aplique el mismo movimiento comercial a cada cliente. El estudio de Durán Barros, Ernesto José & Cecilia (2025), detectó un segmento con bajos niveles de consumo y gasto y bajo nivel de conversión, lo que indica la utilidad del *clustering* para identificar clientes que necesitan diferentes estrategias o un seguimiento más exhaustivo.

Además, el *clustering* se aplica al desarrollo de promociones individuales. Si una empresa sabe qué grupos responden mejor a descuentos, cupones, beneficios digitales o programas de fidelización, entonces puede asignar los recursos de marketing de manera más eficiente y evitar campañas generales que tienen costos sustanciales y menor efectividad. En el estudio revisado, se menciona que los segmentos obtenidos pueden usarse para definir promociones específicas, asignar mejor el presupuesto de marketing y anticipar la respuesta promocional de cada grupo de clientes (Durán Barros & Nieto, 2025).

De la misma manera, esta perspectiva es beneficiosa para las estrategias de fidelización. Un cliente con más valor comercial puede recibir diferentes beneficios, como programas de fidelización, planes de descuento especiales, promociones por volumen o trato preferencial. Los clientes de menor consumo también pueden ser motivados mediante incentivos de bajo costo siempre que no alteren el margen de beneficio. Durán Barros y Nieto Quiñonez (2025)

recomiendan que los sectores más valiosos deben ser tratados como clientes estratégicos mientras que las fracciones de menor valor deberían ser manejadas con medidas de control limitado y revisarlas periódicamente.

Por otra parte, el *clustering* también puede proporcionar beneficios en la investigación de las ventas si la base de datos incluye variables geográficas como cantón, ciudad, sector u oficina; aunque el documento revisado se enfoca en un entorno multicanal, la metodología descrita puede utilizarse para analizar territorialmente los supermercados. Cuando combinamos las variables de ubicación con datos sobre compras, volumen de ventas, tipos de productos y frecuencia, la empresa puede identificar áreas donde los clientes son especialmente rentables, sectores donde la actividad es baja o lugares donde es recomendable aumentar la promoción o la distribución. Esta práctica sería adecuada para supermercados con múltiples sucursales o que sirven a clientes del mismo cantón.

El *clustering* hace que las decisiones de marketing sean más estratégicas ya que convierte los datos comerciales en información estratégica. En lugar de lanzar una campaña generalizada, una empresa se enfoca en un segmento específico y actúa específicamente con base en el perfil del segmento. Los clientes frecuentes pueden ser dirigidos con recompensas de fidelización, los clientes de alto valor pueden ser dirigidos con campañas personalizadas, a los compradores sensibles a las ofertas se les pueden ofrecer descuentos especiales y al usuario de bajo gasto se pueden examinar las posibilidades de reactivación. Este enfoque se alinea con la investigación revisada, que indica que los clusters pueden ayudar en el avance del marketing personalizado y basado en evidencia (Durán Barros & Nieto, 2025).

En consecuencia, las aplicaciones del *clustering* en el comercio minorista muestran que esta técnica tiene un trasfondo académico y una utilidad real en los negocios. Tiene más que ver

con clasificar a los clientes, con perfiles de valor, promociones, fidelización, selección de productos y orientación de la estrategia empresarial según el comportamiento del cliente. Por lo tanto, la industria de supermercados es relevante porque se puede hacer desde un punto de vista de datos y se pueden realizar acciones comerciales más precisas para cada grupo de clientes.

Análisis de la demanda en empresas comerciales

La demanda de clientes desempeña un papel crucial para las empresas comerciales dado que le brinda la oportunidad de conocer tanto el volumen del consumo presente en un mercado como la reacción de los consumidores ante la disponibilidad de mercancías y servicios. Desde el punto de vista de la publicidad, la demanda está vinculada a las necesidades y aspiraciones de los compradores cuando éstas se sostienen con la capacidad financiera, lo que facilita entender oportunidades comerciales y tomar mejores decisiones empresariales Kotler, P & Keller, K. L., (2016). El análisis de la demanda en los supermercados adquiere especial relevancia puesto que las compras son recurrentes e influenciadas por artículos de consumo diario.

Las empresas pueden evaluar la demanda utilizando variables medibles como las preferencias de compra, la recurrencia del consumo, el volumen adquirido, la cantidad vendida y la reacción de los clientes ante la oferta de precios, descuentos u otros incentivos comerciales. Así, comprendemos que no solo miden las ventas totales de un período específico, sino que también capturan tendencias en los segmentos de clientes (FasterCapital, 2014).

El análisis de la demanda dentro de una firma comercial requiere contemplar cuáles son las prácticas de compra, cuáles productos escogen los usuarios, con qué regularidad retornan a la compañía y cuáles criterios comerciales influyen en sus opciones de compra. (Solomon, 2019).

El conocimiento sobre el mercado comprende la identificación de las personas y familias que participan en la actividad empresarial. Los supermercados experimentan principalmente un mercado de consumo, en el cual individuos y familias compran productos para satisfacer sus propias necesidades o de hogar. Según Pérez (2006), el mercado de consumo involucra compradores finales que efectúan transacciones para consumo o uso directo de bienes y servicios. La definición implica que los supermercados más allá de proporcionar intercambios individuales incluyen necesidades recurrentes relacionadas con alimentación, limpieza, higiene personal, suministro familiar y autonomía.

El análisis de la demanda debe distinguir entre las diversas funciones implicadas en la compra y la utilidad del producto, ya que la persona que realiza la compra puede no ser la misma que luego consume el artículo.

Hoyer et al. (2016) destacan que los procesos de decisión a veces implican diversos perfiles, incluyendo el individuo que forma la opinión, el comprador y el usuario del producto. En las tiendas de conveniencia, este aspecto adquiere relevancia por la posibilidad de que una persona pueda comprobar las necesidades familiares mientras que otros miembros del hogar utilicen los artículos comprados. Por tanto, la demanda no se debe estudiar exclusivamente con referencia al acto de compra sino también teniendo en cuenta cómo los bienes cubren necesidades y provocan una renovación de compra.

El consumo minorista de alimentos se expresa mediante compras reales. Entonces, el estudio de este comportamiento está muy ligado al comportamiento del consumidor, tal como se define por Solomon (2019). Según él, el comportamiento del consumidor se refiere a todos los procesos involucrados cuando las personas escogen, compran, usan o abandonan productos y servicios con el propósito de satisfacer necesidades y deseos. Para un punto de vista

latinoamericano, Rolando Arellano Cuevas et al. (2010) destaca que el comportamiento del cliente debe considerarse en relación con su entorno social, cultural y económico, porque estos factores modifican cómo ve los productos y la elección entre uno y otro. Así, es necesario observar el comportamiento de los clientes en tiendas de supermercados con referencia a las acciones reales realizadas en la operación comercial.

En el contexto de los supermercados, los diferentes momentos del proceso de compra pueden manifestarse de varias formas. Antes de realizar una compra, el consumidor identifica una necesidad, evalúa las opciones disponibles, se fija en aspectos como el precio, la marca y la disponibilidad. Durante el proceso de compra, deciden qué productos quieren comprar e incluso cuántos, mientras que, tras la compra, evalúan si el producto ha cubierto sus expectativas y si estarían dispuestos a realizar otra compra en el futuro. Así, la demanda no surge solamente cuando hay mercancía disponible; también surgió gracias a la interacción entre las necesidades del consumidor, las percepciones y las experiencias y los aspectos comerciales que influyeron en su decisión final (Solomon, 2019).

En el día a día de los supermercados, este comportamiento se manifiesta de forma clara. Un comprador podría adquirir productos básicos debido a su urgente necesidad, comprarse artículos suficientes para durar varios días o aprovechar promociones y, por ejemplo, comprar una marca habitual simplemente porque le gusta. Por tanto, la demanda puede detectarse por la frecuencia con la que el consumidor entra a comprar, el número total de productos que compra, el importe en dólares y euros que cada compra representa y la repetición de esos patrones a lo largo de todo el tiempo. Estas cifras ofrecen al supermercado pistas sobre si el comprador tiene una compra rutinaria, cuáles son sus preferencias en productos o si está más influenciado por los descuentos ofrecidos y los beneficios que ofrecen los proveedores.

Las empresas comerciales, y en particular las organizaciones enfocadas en la venta de productos de alta rotación y consumo masivo pueden emplear la base de datos de ventas para analizar la demanda existente. Cada operación comercial proporciona datos acerca de los productos comercializados, los niveles de compra, la fecha del evento de compra, los precios que se aplicaron, así como atributos del comprador, Esta información permite pasar de una visión general de ventas a un análisis más exhaustivo del mercado. Según Kotler, P & Keller, K. L. (2016). Es fundamental conocer al cliente para crear ofertas de valor, dado que las firmas deben entender que quieren los compradores y cómo reaccionar al producto o servicio comercializado

El análisis de la petición también revela que los clientes actúan de manera desigual. En tiendas de gran superficie, existen compradores que realizan frecuentes compras de suministro, compradores que compren de forma esporádica con mayor volumen y otros que compran principalmente para asegurarse de tener los recursos necesarios o compradores que se centran en productos determinados. Esto demuestra que una organización puede tratar diversos comportamientos de demanda, razón por la cual debe examinar los patrones que definen cada grupo. Según Rolando Arellano Cuevas et al. (2010), el análisis de los consumidores exige analizar diferencias entre individuos y grupos, porque dichas diferencias afectan la manera en que seleccionan y estiman los bienes y servicios.

Las compras de baja importancia muy frecuentes suelen estar asociadas a productos de uso cotidiano o necesidades inmediatas. Así, el cliente recurre a la tienda supermercadista con mayor habitualidad, pero los importes de sus operaciones pueden ser bajos. La mayoría de los casos ocurre cuando las personas compran pan, leche, frutas, productos de limpieza o bienes de primeras necesidades frecuentemente. En términos de demanda, estos clientes ofrecen confiabilidad y estabilidad para el negocio. Su aporte monetario por venta generalmente es pequeño. En cuanto al

valor de comercialización de un cliente frente a una pequeña empresa, la repetición es una variable relevante.

De otra parte, las compras de mayor monto ocasional pueden estar ligadas a abastecimientos familiares, compras semanales mensuales, festividades, eventos o reposiciones múltiples de productos, estos clientes no acuden regularmente a la tienda, pero, al comprar generas tickets más elevados, este tipo de comportamiento es importante porque nos permite considerar la demanda según

Otro motivo de compra es la comodidad, donde el cliente prioriza aspectos tales como proximidad, velocidad, facilidad de acceso al producto o disponibilidad inmediata. Estas tomas de resoluciones suelen estar menos asociadas a un análisis detallado de las opciones y más conectadas con la necesidad de afrontar una circunstancia especifica. Según Solomon (2019), muchas elecciones de consumo están determinadas tanto por el entorno como por la experiencia adquirida durante el proceso de compra. Así, atributos como la ubicación de la tienda, el tiempo disponible y la facilidad con que funciona el proceso podrían condicionar la conducta del comprador. En los supermercados, este tipo de comportamientos podría explicar compras rápidas o incluso visitas urgentes

Otro comportamiento observable es la compra agrupada de ciertas categorías de productos. Algunos consumidores pueden mostrar preferencia por alimentos ,otros por artículos domésticos, bebidas, frutas frescas u objetos como marca .Estas concentraciones permiten identificar los intereses y los patrones de consumo que nos ayudan a entender la demanda de los bienes del punto de vista comercial , saber lo que mueve mayor tráfico indica cuales se deben incluir en el surtido, para cual hay que ofrecer promociones, cuánto debe abastecerse y cómo debe comunicarse , El

análisis de las categorías ayuda también a destacar que producto tienen un papel esencial en la conexión que establece el consumidor con el supermercado

Los precios y los descuentos incluyen en la forma en que se declara la demanda .Las empresas de consumo masivo utilizan los precios para compararlos, responder a ofertas y adaptar sus compras según un beneficio económico .En otras palabras, el precio representa una variable fundamental en la propuesta de valor , pues modifica la percepción del consumidor y su disposición para comprar ,Es común ver este fenómeno en supermercados, cuando ciertos productos gana en rotación durante periodos de promoción o cuando algunos durante periodos de promoción o cuando algunos clientes responden de forma diferente a descuentos específicos (Ailawadi et al., 2001).

Las reservas de mercadería es otro aspecto que puede influir en la demanda percibida. Si un almacén no dispone de los artículos que la persona quiere, la compra puede ser menor o pasar por otros locales. Además, una variedad amplia y disponible de productos puede incrementar el compromiso del consumidor y elevar la posibilidad de hacer negocios. Desde la publicidad, la satisfacción del comprador está ligada al grado mediante el cual la organización responde a las expectativas del cliente y entrega valor de modo consistente Kotler, P & Keller, K. L. (2016). Por tanto, la demanda no puede evaluarse únicamente en función del deseo del consumidor, sino también en relación con la habilidad de la compañía para proveer artículos adecuados en la situación indicada.

El argumento de la demanda proporciona información relevante para las decisiones empresariales, ya que ayuda a reconocer tendencias de consumo e incluir predicciones sobre posibles cambios en el comportamiento de los clientes. Al saber cuáles son los clientes que más compran ,los productos más populares , los montos más consumidos y los elementos que afectan

la compra de forma particular , una empresa puede desarrollar estrategias comerciales más acordes al entorno económico ,Este tipo de análisis ayuda a evitar las decisiones comerciales hechas desde la intuición y fundamentadas en evidencia, lo cual permite tener una dirección empresarial más sólida, En consecuencia, la información sobre la demanda se transforma en un valor estratégico para incrementar la capacidad competitiva de la organización (Kotler & Keller, 2016).

En negocios comerciales que manejan bases de datos de ventas, la demanda puede observar a través de indicadores tales como la frecuencia, la recencia, el total comprado, el ticket medio, el volumen adquirido y la recurrencia. Dichos parámetros permiten rastrear la relación del cliente con la organización a lo largo del tiempo y facilitan la localización de tendencias de compra. A pesar de que los factores implicados son cuantitativos, su significación depende de las interpretaciones de un punto de vista empresarial: los números sólo adquieren valor en cuanto permiten explicar perfiles (Kumar & Reinartz, 2018).

MARCO LEGAL

La Constitución de la República del Ecuador específica el marco jurídico que regula derechos, deberes y obligaciones del país y orienta la protección de la actividad productiva y el ejercicio empresarial. El artículo 321 reconoce y protege el derecho de propiedad tanto pública, como privada, comunitaria, estatal, asociativa, cooperativa, mixta. La propiedad también debe cumplir funciones sociales y ambientales según lo indica: "El Estado reconoce y garantiza el derecho a la propiedad en sus formas pública, privada, comunitaria, estatal, asociativa, cooperativa, mixta, y la debe cumplir su función social y ambiental." (Constitución de la República del Ecuador, 2008, art. 321).

Dicha disposición determina la creación de activos para una nueva marca por el hecho de proteger la propiedad de los locales, equipos, envases y derechos de la asociación, además exige que la responsabilidad social y la sostenibilidad se integren en la estrategia comercial.

Otro aspecto relevante son los límites y responsabilidades del Estado en cuanto a la delegación de funciones entre la administración pública y privada, así como en la gestión de sectores estratégicos. Así, el artículo 316 de la Constitución define la atribución del Estado sobre este tema, además de su compromiso con áreas estratégicas. Esta disposición indica que el Estado tiene la capacidad de delegar a empresas mixtas con participación mayoritaria de acciones estatales; en circunstancias extraordinarias, a la iniciativa privada y a la economía popular y solidaria en la medida que la legislación así lo dispusiera; y a la vez puede delegar la participación en sectores estratégicos y servicios públicos en las empresas mixtas, si el Estado es el dueño de una mayoría de sus acciones. Además, la atribución se debe hacer considerando el interés nacional y dentro del plazo y límites que le fije la ley por cada sector estratégico. El Estado podrá, de forma extraordinaria, delegar a la iniciativa privada y a la economía popular y solidaria, el ejercicio de estas actividades, en los casos que establezca la ley. Nota: Por Resolución de la Corte Constitucional No. 1, publicada en el Registro Oficial Suplemento 629 de 30 de enero del 2012, se interpreta estos artículos distinguiendo la gestión de la administración, regulación y control por el Estado y determina el rol de las empresas públicas delegatarias de servicios públicos. (Constitución de la República del Ecuador, 2008, art. 316).

Esa norma es relevante para el proyecto ya que establece restricciones para el acceso a infraestructuras críticas de telecomunicaciones y logística; estos componentes son esenciales para implementar una estrategia omnicanal que implique redes, autorizaciones y servicios que están sujetos a regulación estatal. La Constitución garantiza el derecho al trabajo y establece principios

de protección laboral que inciden en la organización empresarial y en la relación con la fuerza laboral (Constitución de la República del Ecuador, 2008).

El artículo 326 consagra el derecho al trabajo y enumera principios que sustentan su ejercicio, entre ellos la promoción del empleo pleno, la irrenunciabilidad de los derechos laborales y la igualdad de remuneración por trabajo de igual valor:

El derecho al trabajo se sustenta en los siguientes principios:

1. El Estado impulsará el pleno empleo y la eliminación del subempleo y del desempleo.
2. Los derechos laborales son irrenunciables e intangibles. Será nula toda estipulación en contrario.
3. Los derechos laborales son irrenunciables e intangibles. Será nula toda estipulación en contrario.
4. En caso de duda sobre el alcance de las disposiciones legales, reglamentarias o contractuales en materia laboral, estas se aplicarán en el sentido más favorable a las personas trabajadoras.
5. A trabajo de igual valor corresponderá igual remuneración.
6. Toda persona tendrá derecho a desarrollar sus labores en un ambiente adecuado y propicio, que garantice su salud, integridad, seguridad, higiene y bienestar.
7. Toda persona rehabilitada después de un accidente de trabajo o enfermedad, tendrá derecho a ser reintegrada al trabajo y a mantener la relación laboral, de acuerdo con la ley.
8. Se garantizará el derecho y la libertad de organización de las personas trabajadoras, sin autorización previa. Este derecho comprende el de formar sindicatos, gremios, asociaciones y otras formas de organización, afiliarse a las de su elección y desafiliarse

libremente. De igual forma, se garantizará la organización de los empleadores. (Constitución de la República del Ecuador, 2008, art. 326)

Tales postulados obligan a incorporar prácticas laborales formales en la microempresa que operará la marca, a dimensionar costes de personal dentro del modelo de negocio y a diseñar políticas de cumplimiento que reduzcan riesgos legales y reputacionales.

Ley de Comercio Electrónico, Firmas Electrónicas y Mensajes de Datos

Con el propósito de otorgar seguridad jurídica a la actividad comercial electrónica, regula los mensajes de datos válidos, las transacciones electrónicas, la firma electrónica y la salvaguarda de los usuarios en espacios digitales. Según la Ley de Comercio Electrónico, (2002 , art. 1) define el objetivo de la norma y precisa su ámbito de aplicación, lo cual posibilita que 42 una empresa omnicanal sustente contratos, ofertas y comunicaciones en mensajes de datos conforme a la legislación:

Esta Ley regula los mensajes de datos, la firma electrónica, los servicios de certificación, la contratación electrónica y telemática, la prestación de servicios electrónicos, a través de redes de información, incluido el comercio electrónico y la protección a los usuarios de estos sistemas. Ley de Comercio Electrónico, (2002, art. 1).Esa seguridad jurídica facilita la implementación de procesos de venta electrónica y la formalización contractual con proveedores y clientes. La ley reconoce la equivalencia jurídica entre mensajes de datos y documentos escritos, criterio central para la implementación de procesos comerciales digitales.

El artículo 2 otorga igual valor jurídico a los mensajes de datos que a los documentos impresos siempre que se cumplan los requisitos legales: “Los mensajes de datos tendrán igual valor jurídico que los documentos escritos. Su eficacia, valoración y efectos se someterá al

cumplimiento de lo establecido en esta Ley y su reglamento." (Ley de Comercio Electrónico, 2002, art. 2).

Y el artículo 6 dispone que cuando la ley exija información escrita dicho requisito quedará satisfecho mediante un mensaje de datos accesible para consulta posterior: "Cuando la Ley requiera u obligue que la información conste por escrito, este requisito quedará cumplido con un mensaje de datos, siempre que la información que éste contenga sea accesible para su posterior consulta." Ley de Comercio Electrónico, (2002, art. 6). Ese tipo de pronósticos ayuda con la facturación electrónica, las políticas de privacidad y el almacenamiento electrónico de comprobantes en la operación de la empresa; también determinan los requisitos técnicos para la custodia y el acceso.

La normativa establece reglas respecto a la firma electrónica y a la validez de los contratos firmados por medios telemáticos, aspectos cruciales para garantizar la seguridad jurídica en transacciones omnicanal. En ese sentido, la Ley de Comercio Electrónico, (2002 , art. 15), establece requisitos para la firma electrónica, como la verificación de autoría, el reconocimiento del titular y métodos seguros para su creación y comprobación: Requisitos de la firma electrónica. - Para su Validez, la firma electrónica reunirá los siguientes requisitos, sin perjuicio de los que puedan establecerse por acuerdo entre las partes:

a) Ser individual y estar vinculada exclusivamente a su titular;
b) Que permita verificar inequívocamente la autoría e identidad del signatario, mediante dispositivos técnicos de comprobación establecidos por esta Ley y sus reglamentos;
c) Que su método de creación y verificación sea confiable, seguro e inalterable para el propósito para el cual el mensaje fue generado o comunicado.

d) Que, al momento de creación de la firma electrónica, los datos con los que se creare se hallen bajo control exclusivo del signatario; y,
e) Que la firma sea controlada por la persona a quien pertenece (Ley de Comercio Electrónico, 2002, art. 1)

Y el artículo 45 afirma que los contratos instrumentados mediante mensajes de datos tendrán validez y fuerza obligatoria: "Los contratos podrán ser instrumentados mediante mensajes de datos. No se negará validez o fuerza obligatoria a un contrato por la sola razón de haberse utilizado en su formación uno o más mensajes de datos." (Ley de Comercio Electrónico, 2002, art. 45).

Tales mandatos obligan a seleccionar proveedores de certificación acreditados, implementar controles de autenticidad y diseñar procesos contractuales que protejan intereses comerciales y del consumidor en todos los canales.

Ley Orgánica de Defensa del Consumidor

Tiene por objeto normar las relaciones entre proveedores y consumidores y asegurar equidad y seguridad jurídica en el mercado, criterio que orienta la oferta comercial de productos alimentarios. El Artículo 1 establece que la norma es de orden público y se aplicará en el sentido más favorable al consumidor, lo que impone un marco protector para prácticas comerciales y marca obligaciones de transparencia: Las disposiciones de la presente Ley son de orden público y de interés social, sus normas por tratarse de una Ley de carácter orgánico prevalecerán sobre las disposiciones contenidas en leyes ordinarias. En caso de duda en la interpretación de esta Ley, se la aplicará en el sentido más favorable al consumidor. El objeto de esta Ley es normar las relaciones entre proveedores y consumidores promoviendo el conocimiento y protegiendo los derechos de los consumidores y procurando la equidad y la seguridad jurídica en las relaciones

entre las partes. (Ley de Comercio Electrónico, 2002, art. 1). Por tanto, las empresas deben estructurar políticas de información, reclamos y cumplimiento que garanticen derechos de los clientes urbanos.

El Artículo 17 determina obligaciones específicas del proveedor relativas a la entrega de información veraz, suficiente, clara, completa y oportuna sobre bienes y servicios: "Es obligación de todo proveedor, entregar al consumidor información veraz, suficiente, clara, completa y oportuna de los bienes o servicios ofrecidos, de tal modo que éste pueda realizar una elección adecuada y razonable." (Ley de Comercio Electrónico, 2002, art. 17). Cuando se trata de alimentos, esta exigencia exige etiquetas que incluyan la lista de ingredientes, la información sobre origen, y advertencias sanitarias, además de publicidad verificable. La normativa requerida implica implementar sistemas internos de control de calidad, registros de lotes y seguimiento de la trazabilidad, así como documentación probatoria necesaria para respaldar afirmaciones de salud o funciones en jugos naturales (Reglamento de Etiquetado de Alimentos Procesados para Consumo Humano, 2014).

El régimen de publicidad prohíbe expresamente toda modalidad de publicidad engañosa o abusiva y tipifica conductas que inducen al error, entre ellas la atribución falsa de origen y la exageración de beneficios del producto: "Quedan prohibidas todas las formas de publicidad engañosa o abusiva, que induzcan a error en la elección del bien o servicio que puedan afectar los intereses y derechos del consumidor." (Ley Orgánica de Defensa del Consumidor, 2000, art. 6). Esa regulación condiciona la formulación de mensajes en campañas omnicanal y establece riesgo administrativo y civil frente a afirmaciones no demostradas. Además, el Artículo 31 fija plazos de prescripción de acciones civiles en doce meses, dato que obliga a mantener registros de transacciones y comunicaciones durante ese periodo para responder eventuales reclamaciones.

Ley Orgánica de Protección de Datos Personales

Busca garantizar el ejercicio del derecho a la protección de datos personales y dotar de seguridad jurídica al tratamiento en entornos electrónicos y presenciales: 45 El objeto y finalidad de la presente ley es garantizar el ejercicio del derecho a la protección de datos personales, que incluye el acceso y decisión sobre información y datos de este carácter, así como su correspondiente protección, Para dicho efecto regula, prevé y desarrolla principios, derechos, obligaciones y mecanismos de tutela. El Artículo 1 delimita el objeto y ámbito de aplicación, lo que incluye tratamientos relacionados con la oferta de bienes y servicios al territorio nacional, criterio crucial para actividades omnicanal que recaban datos por ventas, programas de fidelidad y análisis de comportamiento urbano (Ley Orgánica de Protección de Datos Personales, 2021, art. 1).

El Artículo 9 reconoce los requisitos que deben cumplirse cuando el tratamiento de datos personales se sustenta en el interés legítimo, precisando que solo pueden emplearse aquellos datos indispensables para cumplir con lo requerido, asegurando al mismo tiempo que el titular cuente con información clara sobre dicho tratamiento y que sea posible evaluar los riesgos existentes para preservar los derechos fundamentales:

Cuando el tratamiento de datos personales tiene como fundamento el interés legítimo:

a) Únicamente podrán ser tratados los datos que sean estrictamente necesarios para la realización de la finalidad.
b) El responsable debe garantizar que el tratamiento sea transparente para el titular.
c) La Autoridad de Protección de Datos puede requerir al responsable un informe con riesgo para la protección de datos en el cual se verificará si no hay amenazas concretas a las

expectativas legítimas de los titulares y a sus derechos fundamentales. (Ley Orgánica de Protección de Datos Personales, 2021, art. 9)

La operación omnicanal requiere procedimientos que permitan atender reclamaciones en plazos razonables, documentar respuestas y ejecutar rectificaciones en todos los puntos de contacto, tanto en tiendas físicas como en plataformas digitales. Por lo tanto, la arquitectura de datos de las marcas debe incorporar políticas para proteger los datos y las interfaces administrativas que garantizan el respeto a los derechos (Ley Orgánica de Protección de Datos Personales, 2021).

El cuerpo normativo establece una serie de principios de tratamiento que requieren finalidad específica, proporcionalidad, calidad, transparencia, seguridad y legitimidad. En particular, la normativa asegura principios de tratamiento que guían toda actividad relacionada con la gestión de datos personales, lo que implica asumir responsabilidades tanto organizativas como técnicas vinculadas a la minimización de datos, la seguridad de la información y el control de accesos.

El titular tiene el derecho a recibir del responsable del tratamiento, sus datos personales en un formato compatible, actualizado, estructurado, común, interoperable y de lectura mecánica, preservando sus características; o a transmitirlos a otros responsables. La Autoridad de Protección de Datos Personales deberá dictar la normativa para el ejercicio del derecho a la portabilidad (Ley Orgánica de Protección de Datos Personales, 2021, art. 17).

MARCO CONCEPTUAL

En el presente estudio toma como fundamento los siguientes conceptos asociados con el análisis de la demanda comercial, los hábitos de compra de los clientes, con el uso de *clustering* y segmentación de consumidores para determinar características de consumidores entre las dos

sucursales de los supermercados, estos términos facilitan la comprensión de datos comerciales que permiten interpretar la información indispensable para la toma de decisiones, la lealtad del consumidor y diseño de estrategias comerciales más específicas.

Demanda del cliente

Nos referimos a la demanda del cliente como el interés real de los consumidores en comprar productos o servicios siempre que tengan poder adquisitivo. Para los supermercados, la demanda se refleja en el número de clientes que compran productos en los supermercados, la frecuencia con las que los compran, la cantidad de ventas que realizan y la respuesta a los precios, descuentos o promociones. Así que analizar la demanda no solo es revisar las ventas totales, sino también lo que impulsa el comportamiento de compra del consumidor y cómo varía entre diferentes tipos de clientes (Greenlaw et al., 2022)

Comportamiento del Consumidor

El comportamiento del consumidor se refiere a las acciones, decisiones y hábitos que toman las personas antes, durante y después de comprar un producto. En los supermercados, esto se manifiesta en forma de compras frecuentes, compras por necesidad, compras por conveniencia, compras familiares, compras impulsivas o compras motivadas por descuentos. Solomon (2019), afirma que este es el proceso de toma de decisiones del consumidor en el que el consumidor tiene que comprar, usar y evaluar productos según sus necesidades, preferencias y experiencias. Por lo tanto, el comportamiento del consumidor es muy importante cuando se trata de entender la demanda real en un supermercado.

Cliente y Consumidor

En el análisis comercial, un cliente no es lo mismo que un consumidor. El cliente es quien realiza la compra o transacción en el establecimiento, y el consumidor es quien usa o consume el producto. Para los supermercados (no estamos hablando de comprar productos para vender en tu tienda, sino para los miembros de la familia), esta diferencia nos muestra que la demanda (no solo para el comprador directo, sino también para el miembro de la familia y el grupo familiar o el entorno al que pertenecen) no es tan simple (Hoyer et al., 2016).

Segmentación de Clientes

La segmentación de clientes es la segmentación del mercado en diferentes tipos de consumidores (similitud en necesidades, características o comportamientos) y cómo se dividen. La industria de los supermercados utiliza esta herramienta para identificar diferentes tipos de clientes basándose en su frecuencia de compra, cantidad gastada, productos adquiridos, uso de descuentos, acceso a crédito o participación en ventas. A diferencia de tratar a los consumidores de la misma manera, la segmentación permite diseñar estrategias de marketing individualizadas para cada grupo con promociones, programas de lealtad, incentivos comerciales o atención personalizada (Kotler & Keller, 2016)

Segmentación Conductual

La segmentación conductual clasifica a los clientes según sus patrones de compra reales. Este tipo de segmentación observa con qué frecuencia compran los consumidores, cuánto gastan, qué productos les gustan y cómo reaccionan a las promociones o descuentos. En este estudio, se puede utilizar para construir perfiles basados en datos reales del supermercado como fecha de

venta, cantidad de producto, tipo de producto, volumen de ventas, precio, descuento, crédito y tipo de cliente (Kotler & Keller, 2016).

Segmentación por Valor del Cliente

La segmentación por valor del cliente es útil para identificar qué grupos de consumidores contribuyen más a la empresa. No todos los clientes son iguales en ventas; algunos compran con frecuencia, otros compran en grandes cantidades, algunos responden mejor a los descuentos y otros tienen potencial para la lealtad Kumar & Reinartz (2018) indican que el valor del cliente es un componente clave que ayuda a las marcas a diferenciar a los consumidores según su contribución actual y potencial a la empresa. Así, el supermercado puede identificar clientes de alto valor, clientes frecuentes, clientes ocasionales y clientes que necesitan incentivos para aumentar su lealtad.

Datos Transaccionales

Los datos transaccionales son los registros de las actividades de ventas de una empresa. En un supermercado, estos datos pueden incluir fecha de compra, nombre o código del cliente, tipo de producto, cantidad vendida, monto de ventas, sucursal, descuento, crédito y tipo de cliente. Estos datos son esenciales para el análisis de la demanda desde una perspectiva objetiva y basada en evidencia. La empresa puede ver patrones de consumo, tendencias de compra y diferencias entre clientes en cada sucursal organizando, limpiando y procesando esta información (C. B. Chen et al., 2015).

Marketing Analítico

El marketing analítico es la aplicación de datos comerciales para mejorar la toma de decisiones de marketing y la gestión empresarial. Como resultado del análisis de datos, es posible

identificar para las empresas qué productos tienen más demanda, qué clientes generan más ingresos, qué promociones son efectivas y qué segmentos serían más impactados por estrategias específicas. En este estudio, el marketing analítico ha hecho que el supermercado pase de tomar decisiones generales a tomar acciones basadas en datos reales sobre sus consumidores (France & Ghose, 2019).

Clustering

El Clustering es un método de aprendizaje no supervisado que clasifica datos en grupos o clústeres basados en su similitud. A diferencia de otros métodos, el clustering no requiere categorías específicas, sino la identificación de patrones en la base de datos. Como señala Jain, (2010), el clustering se puede usar para organizar grandes cantidades de información en grupos más fácilmente interpretables. En el supermercado, el clustering podría usarse para categorizar a los clientes con comportamientos de compra, consumo, descuento, crédito y participación en ventas similares.

K-means

K-means es uno de los métodos de clustering más populares para la segmentación de clientes. Este algoritmo agrupa observaciones de manera que los elementos en el mismo clúster sean similares y diferentes de otros grupos. K-means puede ayudar a identificar perfiles de clientes específicos, por ejemplo, clientes frecuentes, compradores de alto valor, clientes sensibles a descuentos, clientes de baja recurrencia o clientes que usan crédito. Para aplicar esta técnica correctamente, se necesita definir el número de clústeres utilizando el método del codo y el coeficiente de silueta para crear el perfil (Han et al., 2012; Jain, 2010).

Técnicas de winsorización

Winsorización es una técnica estadística que reemplaza los valores extremos o atípicos de un conjunto de datos por otros valores menos extremos fijados en un percentil específico, en lugar de eliminar dichos datos por completo. Su nombre proviene del bioestadístico Charles P. Winsor (Reifman & Garrett, 2022).

Lealtad del Cliente

La lealtad del cliente es la noción de mantener una buena relación con los clientes que son la base de la lealtad del cliente. La lealtad se puede lograr en los supermercados en forma de ofertas de descuento personalizadas, promociones de frecuencia, beneficios para clientes de alto valor, especializaciones de clientes y/o incentivos comerciales. Al segmentar los segmentos de clientes de manera más específica, el supermercado podría tener un enfoque más personalizado para aumentar la lealtad del cliente y mejorar la relación con los clientes (Kumar & Reinartz, 2018).

Pronóstico de Ventas

El pronóstico de ventas es una estimación futura del comportamiento comercial de una empresa basada en patrones de consumo pasados y futuros. En este trabajo, el pronóstico está relacionado con la identificación de perfiles de clientes; qué segmentos están comprando más, con qué frecuencia y cuánto, y cuándo es mejor para predecir la demanda. Esto ayuda al supermercado a reorganizar su estrategia comercial, ajustar el inventario, diseñar promociones y tratar con sus clientes de manera más eficiente (Chen et al., 2020).

Relación Conceptual del Estudio

Juntos, estos conceptos nos ayudan a entender que la demanda de los clientes en el supermercado no debe analizarse sólo como una función de las ventas totales, sino también

basándose en variables comerciales significativas. Los datos transaccionales nos dan una idea del comportamiento real de los consumidores; la segmentación nos permite agruparlos en grupos con características comunes; y el clustering nos permite identificar perfiles de clientes de una manera más objetiva. Por lo tanto, es importante analizar la demanda bajo factores significativos agrupados para ver qué partes del mercado tienen el mayor impacto en las ventas y qué estrategias se pueden tomar para mejorar la lealtad, los incentivos comerciales y la gestión de cada sucursal.

METODOLOGÍA

Enfoque de la investigación

El presente estudió se centra en datos cuantitativos puesto que trabaja con datos numéricos en este caso: las ventas, clientes, montos, créditos y tipo de productos. Además, contiene un alcance analítico debido a que busca patrones de compra de los consumidores y poder identificar relaciones entre variables comerciales como frecuencia de compra, descuentos y créditos. Este enfoque permite aplicar métodos estadísticos y técnicas de aprendizaje automático no supervisado que en este caso es el *clustering* con la finalidad de segmentar clientes según comportamientos similares con el propósito de generar información relevante para la toma de decisiones relevantes (Hernández-Sampieri, 2018).

Fuente de la información y base de datos

La información y base de datos fueron recopiladas de un supermercado de la provincia de El Oro la cual tiene dos sucursales, la base de datos contiene las ventas desde el mes de noviembre del 2025, así como el tipo de productos, sucursal, etc.

Preparación y depuración de datos

Ahora bien, antes de utilizar la técnica de *clustering* se preparan los datos, se tiene en cuenta la cantidad y las variables que se utilizarán. Además de la preparación y depuración de los datos para garantizar la calidad y la consistencia de los datos para el uso en el análisis (James, G., 2021; Tan et al., 2021)

Este procedimiento es necesario debido a que usualmente los archivos de datos se realizan para fines operativos y no para análisis estadístico, por ende, es crucial un buena organización y manejo.

El proceso de depuración se llevará a cabo en eliminar los datos duplicados, revisar los registros anulados, corregir errores en la digitalización e identificar los valores atípicos. La preparación de datos es una fase esencial en la investigación de mercados puesto que permite transformar registros administrativos en información analizable (Kahle & Malhotra, 1994).

Tabla 3.

Proceso de preparación y depuración de datos

Etapa	Procedimiento realizado	Finalidad
1	Identificación de registros duplicados	Evitar que una misma transacción sea contabilizada más de una vez.
2	Revisión de registros anulados	Excluir operaciones que no representen ventas efectivas.

3	Revisión de errores de digitalización	Corregir inconsistencias en los registros.
4	Identificación de valores atípicos	Detectar observaciones que puedan distorsionar el análisis.
5	Organización de variables	Estructurar la información para su posterior análisis
6	Estandarización de variables	Llevar las variables a escalas comparables antes de aplicar K-means.

Nota. Elaboración propia a partir del procedimiento metodológico planteado en la investigación.

Adicionalmente, el enfoque cuantitativo necesita la definición clara de las variables y asegurar la calidad de los datos para una mejor interpretación (Hernández-Sampieri, 2018).

Posteriormente, las transacciones por parte del cliente se organizan con el fin de construir indicadores de la demanda, ya que el objetivo es obtener patrones de compra.

Análisis exploratorio de datos (EDA): uso y desarrollo matemático

Una de las herramientas esenciales para el manejo de los datos es el EDA, se usa para analizar e investigar conjuntos de datos y resumir sus características principales, a menudo empleando métodos de visualización de datos (IBM, 2025).

En el presente estudio, el análisis exploratorio de datos consiste en reconocer la estructura de la base de datos. En este caso, cada fila representa un registro comercial vinculado a una venta

o movimiento de producto. Para llegar a esto, se verifica que las variables se encuentren correctamente clasificadas: la fecha de venta como variable temporal, el precio de venta, volumen de venta y cantidad de producto como variables cuantitativas y el tipo de cliente y tipo de producto como variables categóricas (Salazar & Castillo, 2018; Tukey, 1977).

Seguidamente, se evalúa la presencia de datos faltantes debido a que la ausencia de valores puede afectar los resultados descriptivos. En cada variable se calculará el porcentaje de valores faltantes. Si una variable presenta un porcentaje elevado de registros vacíos, se deberá ser corregida, excluida o analizada con precaución (Hair et al., 2022).

Donde:

$PF_j =$ porcentaje de datos faltantes en variable j

$n_{faltantes,j} =$ número de registros vacíos en esa variable j.

$n =$ número total de observaciones de la base de datos.

En la presente investigación, esta operación matemática permitirá revisar la calidad de variables cuantitativas y categóricas ya mencionadas.

Seguidamente, se realiza la calidad de los datos, se describen las variables cuantitativas de la base. En este estudio, las principales variables

numéricas son el precio de venta, la cantidad promedio vendida por registro o el volumen promedio generado por cada operación comercial. Para esto, se calcula la media aritmética, utilizada para obtener el valor promedio de todas las variables cuantitativas (Salazar & Castillo, 2018).

La expresión matemática de la media aritmética

$$\underline{x} = \frac{\sum_{i=1}^{n} \quad x_i}{n}$$

En la ecuación, x_i representa cada valor observado en la base de datos, mientras que "n" corresponde al número total de registros analizados.

Por consiguiente, se calcula la desviación estándar para analizar la dispersión de los datos.

La desviación estándar muestral es la definición más popular es la raíz cuadrada de la media aritmética de las desviaciones cuadráticas con relación a la media (Salazar & Castillo, 2018)

Este parámetro permite conocer qué tan lejos se encuentran los valores individuales respecto a la media aritmética o conocida también como promedio. En esta investigación se podría utilizar en el volumen de venta debido que una desviación estándar elevada indicaría diferencias importantes entre registros de ventas.

$$s = \sqrt{\frac{\sum_{i=1}^{n} \quad \left(X_i - \underline{x}\right)^2}{n-1}}$$

El siguiente paso es el coeficiente de variación que permite evaluar la dispersión relativa entre variables registradas en diferentes rangos de medida. En el caso del supermercado, se utiliza para comparar la variabilidad de precio de venta frente a la variabilidad de la cantidad de producto o de volumen de venta. Esta comparación facilita identificar cuál es la variable que tiene mayor incertidumbre relativa con respecto al comportamiento comercial (Anderson, 2024).

$$CV = \frac{s}{\underline{x}} \times 100$$

Además, se utiliza el rango intercuartílico para poder identificar los posibles valores atípicos. Este procedimiento ayuda a detectar registros que están muy por encima o por debajo del

comportamiento normal de la variable que se analiza. El método del rango intercuartílico (IQR) sirve para identificar valores atípicos y fue desarrollado por John Tukey en los inicios del análisis exploratorio de datos, en donde el análisis se realizaba manualmente sobre el conjunto de datos pequeños (Oracle, 2026).

A continuación, la expresión matemática es la siguiente:

$$IQR = Q_3 - Q_1$$

$$LI = Q_1 - 1.5(IQR) \;\; ; \;\; LS = Q_3 + 1.5(IQR)$$

Donde:

$Q_1 =$ primer cuartil

$Q_3 =$ tercer cuartil

$LI =$ límite inferior

$LS =$ límite superior

En este estudio, los registros ubicados debajo del límite inferior o por encima del límite superior serán revisados para determinar si corresponden a errores de digitación, operaciones extraordinarias o comportamientos comerciales reales que deben conservarse en el análisis.

Seguidamente, se analizan las variables categóricas que son: tipo de cliente y tipo de producto, se calcula las proporciones o participaciones porcentuales. Este procedimiento permite conocer qué categoría representa mayor presencia dentro de la base de datos y mejora la comparación entre grupos.

$$P_c = \frac{n_c}{n} \times 100$$

Donde:

$P_c =$ Proporción porcentual por categoría

$n_c =$ Número de observaciones que pertenecen a una categoría analizada

$n =$ Número total de observaciones del conjunto de datos

$100 =$ Factor utilizado para expresar el resultado en porcentaje

En esta investigación, se utiliza la expresión matemática para explicar la distribución de las variables cualitativas existentes en la base de datos de supermercados, como el tipo de clientes, el cantón de compra, la existencia de descuento y el uso de crédito. El cálculo de estas proporciones permite identificar cuáles categorías tienen mayor representación y comprender la composición de la demanda antes de realizar los análisis de segmentación mediante técnicas de agrupamiento (*clustering*). Esta aproximación nos ofrece una primera comprensión de cómo interactúan los clientes con los productos disponibles en el negocio y nos permite observar las características principales en la muestra de la base de datos.

A continuación, luego de haber calculado las proporciones por categorías, el EDA avanza a la agregación temporal de la demanda. Es de suma importancia seguir este paso puesto que la base de datos contiene fecha de venta, lo que permite observar el comportamiento de las ventas por día, semana o mes. Según Gujarati y Porter (2010), redactan que el análisis de datos ordenados en el tiempo permite identificar variaciones y tendencias antes de aplicar los modelos explicativos. En el caso del supermercado, la demanda se medirá mediante la cantidad de producto vendida en cada período.

$$D_t = \sum_{i=1}^{n_t} Q_{it}$$

Donde:

$D_t =$ Representa la demanda total del periodo t

$Q_{it} =$ Representa a la cantidad de producto vendida en el registro "i" durante ese periodo

$n_t =$ Representa el número de registros asociados al periodo analizado.

En este caso, la expresión matemática permitirá sumar la cantidad de productos vendidos por día, semana o mes, con la finalidad de observar si la demanda aumenta, disminuye o se mantiene estable con el tiempo.

El siguiente paso es calcular el volumen total de venta por periodo, con la siguiente expresión matemática.

$$VT_t = \sum_{i=1}^{n_t} V_{it}$$

Donde:

$VT_t =$ Volumen total de venta generado en el periodo t

$V_{it} =$ Representa el volumen de venta registrado en cada observación

$n_t =$ Representa el total de registros de período.

El último parámetro permite analizar el valor comercial generado por las ventas. Además, se necesita para observar no solo cuántos productos se vendieron, sino también cuánto valor económico representaron esas ventas en cada periodo.

El siguiente paso es calcular el precio promedio ponderado que se utiliza para analizar el comportamiento del precio considerando la cantidad vendida de cada producto. En este caso el promedio ponderado es mejor que el promedio simple puesto que los registros tienen cantidades diferentes con referente a otorgar mayor peso a los productos con mayor volumen comercializado.

$$PP_t = \frac{\sum_{i=1}^{n_t} P_{it}Q_{it}}{\sum_{i=1}^{n_t} Q_{it}}$$

Donde:

$PP_t =$ Representa el precio promedio ponderado del periodo t

$P_{it} =$ Corresponde al precio de venta registrado

$Q_{it} =$ Representa la cantidad de producto vendida

En este estudio, este cálculo permite analizar el precio promedio real de venta considerando el peso de cada cantidad comercializada.

Posteriormente, se realizan tablas cruzadas para relacionar variables categóricas con medidas de demanda, se realiza este paso puesto que permite observar cómo se distribuye la cantidad vendida según el tipo de cliente y el tipo de producto. En las tabulaciones cruzadas, (Kahle & Malhotra, 1994) señala que son útiles en investigación de mercado debido que permite comparar el comportamiento entre grupos, en este caso ayuda a identificar que productos tienen mayor peso dentro de cada tipo de cliente.

$$Part_{jk} = \frac{Q_{jk}}{\sum_{j=1}^{m} Q_{jk}} \times 100\backslash$$

Donde:

$Part_{jk} =$ Representa la participación porcentual del tipo de producto "j" dentro del grupo de clientes k.

$Q_{jk} =$ Cantidad vendida del producto j al tipo de cliente k

$\sum_{j=1}^{m} \quad Q_{jk} =$ Sumatoria del denominador representa el total de cantidad vendida dentro del grupo k.

Consecutivamente, se calcula la variación porcentual para identificar cambios relativos entre un periodo u otro. Esta medida resulta útil para observar aumentos o disminuciones en la cantidad vendida. En este presente estudio, se aplica a la demanda total, al volumen de venta o a la cantidad vendida por tipo de producto, siempre que existan registros suficientes por fecha.

$$\Delta\%_t = \frac{D_t - D_{t-1}}{D_{t-1}} \times 100$$

Donde:

$\Delta\%_t =$ Representa la variación porcentual de la demanda en el periodo actual.

$D_t =$ Es la demanda del periodo actual.

$D_{t-1} =$ Es la demanda del periodo anterior.

$D_t - D_{t-1} =$ Representa la diferencia entre la demanda actual y la demanda anterior.

$\frac{D_t - D_{t-1}}{D_{t-1}} =$ Permite comparar esa diferencia con respecto al periodo anterior

Este indicador permitirá el comportamiento temporal de las ventas sin afirmar todavía relaciones casuales.

Como último paso del EDA permite relacionar preliminares entre variables cuantitativas mediante el coeficiente de correlación de Pearson. Este proceso facilita la identificación de si dos variables cambian en la misma dirección, en dirección opuesta, en sentido contrario o sin relación lineal evidente. En la base de datos del supermercado se puede utilizar para analizar la relación

entre precio de venta y cantidad de producto y volumen de venta o precio de venta y volumen de venta (Gujarati & Porter, 2010).

$$r = \frac{\sum_{i=1}^{n} (x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \underline{x})^2 \sum_{i=1}^{n} (y_i - \underline{y})^2}}$$

Donde:

$r =$ Representa lo que es el coeficiente de correlación. Su valor siempre estará entre -1 y 1

$x_i =$ Es cada valor individual de la primera variable ya sea precio de venta.

$y_i =$ Es cada valor individual de la segunda variable ya sea cantidad o producto vendida.

$\underline{x} =$ Es el promedio de la primera variable

$\underline{y} =$ Es el promedio de la segunda variable

$x_i - \underline{x} =$ Muestra cuánto se aleja cada valor de x respecto a su promedio

$y_i - \underline{y} =$ Muestra cuánto se aleja cada valor de y respecto a su promedio

En el presente estudio, este paso permitirá identificar relaciones preliminares que posteriormente pueden orientar el uso de regresión lineal o regresión múltiple.

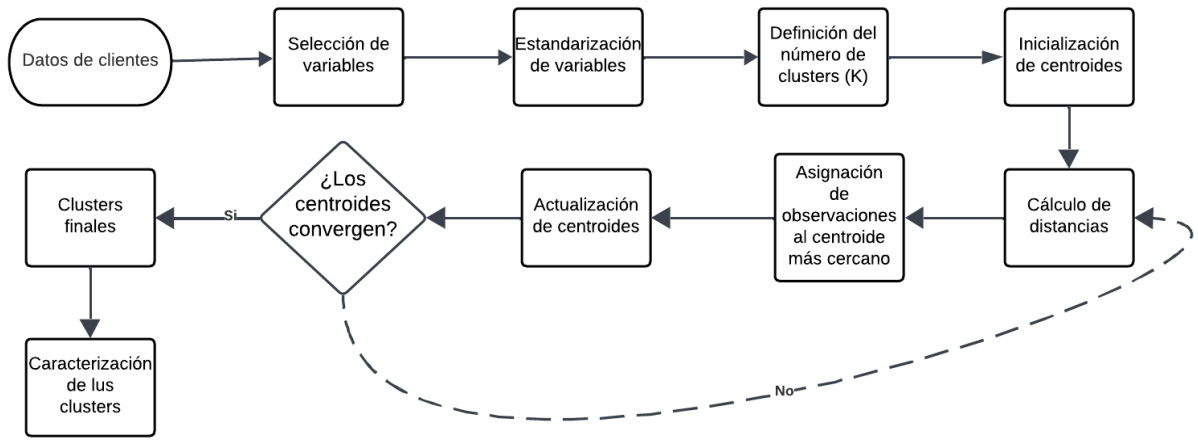
En conclusión, este es uno de los primeros pasos de EDA que permite preparar la base de datos para poder realizar un mejor análisis, estos pasos son de suma importancia porque permiten conocer el comportamiento de la información antes de aplicar un modelo más avanzado. Además, al estar vinculados con las variables reales de la base, los resultados del EDA contribuirán a que la metodología mantenga coherencia con el objetivo de analizar la demanda del supermercado a partir de información observable (Tukey, 1977).

Aprendizaje no supervisado mediante *K-means*

El aprendizaje no supervisado es útil cuando no existe una variable objetivo previamente definida. En este caso el algoritmo *k-means*, el procedimiento tiene como objetivo formar grupos de observaciones similares a partir de variables numéricas o variables categóricas transformadas en datos binarios (James, G., 2021; Tan et al., 2021).

Figura 1.

Proceso de agrupamiento mediante K-means



Nota. Elaboración propia a partir del procedimiento de K-means descrito en James, G., (2021); Tan et al. (2021)

1. Definición de la unidad de análisis para *K-means*

En el ámbito empresarial, *K-means* se puede utilizar en segmentación de mercado para identificar grupos de consumidores con necesidades similares, considerando cada consumidor con necesidades similares, considerando a cada consumidor con necesidades similares, considerando a cada consumidor como una observación y a cada variable como característica relevante para el análisis (Jain, 2010; Tan et al., 2021).

El primer paso consiste en saber qué tipo de variable utilizar en la agrupación del algoritmo. Esta decisión es importante porque *K-means* no interpreta nombres o etiquetas por sí mismo, sino valores comparables entre observación. Por ende, cada unidad de análisis deberá resumirse mediante variables cuantitativas que representen la demanda en este caso son cantidad total vendida, volumen total de venta y precio promedio ponderado (Tan et al., 2021).

Por ende, la expresión matemática utilizada es la siguiente:

$$Q_g = \sum_{i \in g} \quad q_i$$

Donde:

$Q_g =$ Representa la cantidad total vendida dentro del grupo de análisis "g" tales como "clientes jurídicos que compraron productos de arroz o sal o clientes naturales que compraron alimentos básicos durante una semana.

$q_i =$ Representa la cantidad registradas en cada fila de la base y la sumatoria permite obtener el total demandado por grupo

Seguidamente, el siguiente parámetro ayuda a conocer el valor comercial generado por cada combinación de periodo, tipo de cliente y tipo de producto. En la tesis, este indicador sirve para diferenciar grupos que pueden vender muchas unidades con respecto al precio.

$$VT_g = \sum_{i \in g} \quad V_i$$

Donde:

$VT_g =$ Representa el volumen total de venta del grupo g.

Como tercer paso, se calcula el precio promedio ponderado, este indicador es más conveniente que un promedio simple cuando los productos tienen diferentes cantidades vendidas, porque otorga mayor peso a los precios de los productos que tuvieron mayo movimiento (Anderson, 2024). Usándolo en el caso del supermercado, esta medida permite observar si determinados perfiles de demanda se relacionan con precios promedio más altos o bajos.

$$PP_g = \frac{\sum_{i \in g} \quad p_i q_i}{\sum_{i \in g} \quad q_i}$$

2. Construcción de la matriz de características

Luego de haber definido las unidades de análisis, se construye una matriz de características. En K-means, cada observación se representa como un vector de variables, este vector resume los atributos que el algoritmo utiliza para medir similitudes y diferencias entre grupos (Jain, 2010; Tan et al., 2021).

Para la base de esta investigación, las variables numéricas principales pueden ser cantidad total vendida, volumen total de venta y precio promedio ponderado. Por otro lado, las variables del tipo de cliente o tipo de producto también se pueden añadir aplicando codificación binaria a condición de que se requiera el algoritmo, se tiene en cuenta las diferencias que existan entre categorías (Hair et al., 2022).

La fecha de venta se utilizará en este caso para agrupar por periodos (día, mes, año) antes de la agrupación

$$X_g = (Q_g\ , VT_g\ , PP_g\ , TC_{g1}, \dots, TC_{gc}, TP_{g1}, \dots, TP_{gr})$$

Donde:

$X_g =$ representa el vector de características del grupo g

Q_g =Representa la cantidad total de productos vendidos en el grupo g

VT_g =Volumen total de venta

PP_g = Promedio ponderado del grupo g

Los términos TC representan categorías del tipo de cliente transformadas en variables binarias, mientras que TP representa categorías de tipo de producto, también codificadas. De este modo, el algoritmo puede trabajar con información numérica y categórica convertida en formato analizable (Hair et al., 2022).

El siguiente paso, se define la variable tipo de cliente con la siguiente expresión:

$$D_{ij} = 1 \; si \; i \in j \; ; \; D_{ij} = 0 \; si \; i \notin j$$

La explicación se basa en la codificación binaria, en el tipo de clientes tiene la categoría de natural y jurídico, se codifica la variable que toma el valor 1 cuando la observación pertenece a una categoría y 0 cuando no pertenece a ella. Este procedimiento permite incorporar variables cualitativas y de aprendizaje automático, puesto que haciendo esto el algoritmo interpreta las categorías como números ordinales sin significado real.

3. Estandarización de variables

Por consiguiente, se realiza lo que es la estandarización de variables, este paso se aplica antes del k-means, las variables deben estar en escalas comparables. En el caso del supermercado, el volumen de venta puede estar expresado en valores monetarios, la cantidad de productos en unidades y el precio de venta en valores unitarios. Es esencial realizar esta estandarización puesto que si una variable con números grandes puede dominar el cálculo de distancia y afectar deliberadamente la formación del clúster. Por ende, en análisis multivariante se recomienda

estandarizar las variables cuantitativas cuando presentan diferentes unidades de medida (Hair et al., 2022; Jain, 2010).

$$Z_{gi} = \frac{X_{gj} - \underline{x}_j}{s_j}$$

Donde:

$Z_{gi} =$ Representa el valor estandarizado de la observación o grupo g de la variable j. Por ejemplo, el valor estandarizado del volumen de venta para un grupo específico.

$X_{gj} =$ Representa el valor original de la variable j en el grupo g. Por ejemplo, volumen de venta.

$\underline{x}_j =$ Representa el promedio de la variable j en toda la base. Por ejemplo, promedio general del volumen de venta.

$s_j =$ Representa la desviación estándar de la variable j. Por ejemplo, el volumen de venta.

La estandarización es importante puesto que esta transformación convierte cada variable en escala común, donde los valores indican cuántas desviaciones estándar se encuentran por encima o por debajo del promedio. Este procedimiento resulta necesario debido a que en la base de datos se encuentran diferentes unidades de medida tales como precio de venta, volumen de venta y cantidad de producto. La transformación permite en cada variable aporte de manera equilibrada el cálculo de distancias utilizado por *K-means*.

4. Procedimiento de similitud mediante distancia euclidiana

En el cuarto paso, *k-means* agrupa observaciones mediante medidas de distancia. Una de las más utilizadas es la distancia eucaliana, puesto que calcula que tan disperso se encuentra una observación respecto al centroide de un *cluster*. Mientras menor sea la distancia, mayor será la

similitud entre la observación y el grupo. En minería de datos, esta medida se utiliza con frecuencia para algoritmos basados en proximidad, como *k-means* (Han et al., 2012; Tan et al., 2021).

$$d(X_g, \mu_k) = \sqrt{\sum_{j=1}^{p} (X_{gj} - \mu_{kj})^2}$$

Donde:

$d(X_g, \mu_k) =$ Representa la distancia entre el grupo de análisis y el centroide del cluster k.

$X_g =$ Representa una observación o grupo de análisis en tu base

$\mu_k =$ Representa el centroide del cluster k, es otras palabras, el punto promedio que representa a ese grupo.

$(X_{gj} - \mu_{kj})^2 =$ Representa el cálculo de la diferencia entre la observación y el centroide para cada variable elevada al cuadrado para evitar valores negativos.

$\sum_{j=1}^{\rho} \quad =$ Representa la sumatoria de la primera variable j =1 y el número total de variables usadas en el *k-means* representado como p y la variable del modelo como cantidad de producto representado por j.

El algoritmo *k-means* calcula la distancia entre cada registro y los centroides disponibles, asignando cada observación al grupo más cercano. De esta manera, los registros con comportamientos comerciales similares quedan agrupados dentro del mismo clúster.

5. **Asignación de observaciones y actualización de centroides.**

Primero se define un número inicial de clusters K. Después, cada observación se asigna al cluster cuyo centroide este más cercano. Luego, el centroide de cada grupo se calcula como el promedio de las observaciones asignadas a ese cluster. Este ciclo se repite hasta que las

asignaciones se estabilizan o hasta que el cambio en los centroides es mínimo (James, G., 2021; Tan et al., 2021).

$$C_k = \{X_g : d(X_g, \mu_k) \leq d(X_g, \mu_k), \forall h = 1, \dots, K\}$$

Donde:

$C_k =$ Es el cluster o grupo número k.

$\{X_g : d \dots\} =$ Significa el conjunto de observaciones X_g que cumplen con cierta condición

$\forall h = 1, \dots, K$ = Representa para todo cluster h, desde 1 hasta K, es decir, el algoritmo compara la observación con todos los clústeres disponibles.

Esta expresión matemática se interpreta como una observación X_g se asigna al cluster C_k si la distancia entre esa observación y el centroide μ_k es menor o igual que la distancia hacia cualquier otro centroide. En este estudio, da paso a que cada grupo de registros comerciales se ubicará dentro del segmento que tenga mayor similitud en término de cantidad vendida, volumen total, precio promedio y categorías codificadas.

Por consiguiente, la siguiente expresión matemática calcula el centroide del cluster k, es decir, el punto promedio que representa a todos los registros que pertenecen a ese grupo

$$\mu_k = \frac{1}{n_k} \sum_{x_g \in C_k} x_g$$

Donde:

n_k= Es el número total de observaciones que pertenecen al cluster k.

$\Sigma_{x_g \in C_k} =$ Representa la sumatoria de todas las observaciones x_g que pertenecen al cluster C_k.

En este caso, el centroide resume las características comerciales promedio de cada grupo de registros, considerando variables como cantidad de producto, volumen de ventas, etc. A partir de este cálculo, k-means actualiza a posición de cada cluster y permite interpretar los grupos obtenidos como perfiles de demanda dentro del supermercado (Jain, 2010; Tan et al., 2021).

6. Función objetivo y selección de número de clusters

Ahora bien, el objetivo matemático del k-means es minimizar la variabilidad interna de los clusters. En otras palabras, el algoritmo busca que las observaciones dentro de un mismo grupo sean lo más parecidas posibles entre sí. La expresión matemática se basa en la suma de las distancias cuadradas entre cada observación y centroide de su cluster (Tan et al., 2021).

$$min \sum_{k=1X}^{k} \sum_{\in C_k} \| X_g - \mu_k \|^2$$

Donde:

Min = Representa minimizar el resultado sea lo más pequeño posible.

$\Sigma_{k=1}^{k}$**=** Representa que suman las distancias de todos los *clusters*, desde el *cluster* 1 hasta el *cluster* K.

$\boldsymbol{\Sigma_{x9 \in c_k}}$ = Representa la sumatoria de todas las observaciones Xg que pertenecen al *cluster* Ck.

$\|X_g - \mu_k\|^2$ **=** Representa la distancia al cuadrado entre la observación y el centroide de su *cluster*. Se eleva al cuadrado para evitar valores negativos y dar más peso a las observaciones que están más alejadas al centroide.

En esta investigación, cada observación está compuesta por variables comerciales. Por tanto, el algoritmo agrupa registros que presentan comportamientos similares de demanda, procurando que las observaciones dentro de un mismo cluster sean lo más homogéneas posibles.

Por consiguiente, para seleccionar el número adecuado de clusters, se puede utilizar el método del codo. Este procedimiento compara la suma de cuadrados dentro de los grupos para diferentes valores de K Tan et al. (2021). Explica que el método del codo permite evaluar distintos valores de K mediante la variabilidad interna de los grupos, seleccionado aquel punto donde la reducción adicional comienza a ser poco significativa.

$$WCCS(K) = \sum_{k=1}^{k} \sum_{\in C_k} \| X_g - \mu_k \|^2$$

WCSS(K) = Representa la suma de cuadrados dentro de los *clusters* para un número determinado de grupos K.

En esta investigación, el WCSS para diferentes valores de K, se identifica el punto en el que la reducción de la variabilidad interna deja de ser significativa, con la finalidad de elegir una cantidad de grupos que mantenga el equilibrio entre ajuste estadístico y para su interpretación comercial.

7. Regresión lineal y regresión lineal múltiple con enfoque en *machine learning*.

La regresión lineal múltiple se incorpora como una técnica de aprendizaje supervisado para analizar y estimar el volumen de venta a partir de las variables comerciales disponibles en la base de datos. Mientras k-means permite formar grupos de registros con características similares, la regresión permite cuantificar la asociación entre una variable dependiente y varias variables explicativas de manera simultánea. En esta investigación, el volumen de venta constituye la

variable dependiente, debido a que representa el comportamiento de la demanda registrado en las transacciones comerciales. El modelo se estima de forma diferenciada para cada clúster previamente obtenido mediante k-means, de modo que la relación entre las variables comerciales y el volumen de venta pueda analizarse dentro de cada segmento (Gujarati & Porter, 2010; Wooldridge, 2020).

7.1. Definición de variables y estructura del modelo

La regresión se especifica considerando las variables incluidas en el script diseñado para el análisis. Como variable dependiente (y) se utiliza el volumen de venta, mientras que la matriz de variables explicativas (x) reúne el precio de venta, la cantidad de productos, el descuento, el crédito, el año, el mes, el día de la semana, el día del mes, la condición de fin de semana, el tipo de cliente y el local de venta. La variable clúster queda fuera de los predictores, dado que su función se limita a identificar el segmento al que pertenece cada observación, sin aportar a la explicación de la variable dependiente. Por su parte, el volumen de venta no se incluye en X debido a que corresponde a la variable objetivo del modelo.

Tabla 4.

Variables y representación matemática del modelo de regresión

Tipo de variable	**Variable en la investigación**	**Representación en el modelo**	**Función**
Dependiente	Volumen de venta	Y_{ik}	Variable que se busca explicar o predecir.

Explicativa numérica	Precio de la venta	X_{1ik}	Permite medir su asociación con el volumen de venta.
Explicativa numérica	Cantidad de productos	X_{2ik}	Representa la cantidad vendida en el registro comercial.
Explicativa binaria	Descuento	X_{3ik}	Indica si la venta tuvo descuento: 1= sí, 0= no.
Explicativa binaria	Crédito	X_{4ik}	Indica si la venta estuvo relacionada con crédito: 1= sí, 0= no.
Explicativa temporal	Año	X_{5ik}	Controla diferencias entre periodos anuales.
Explicativa temporal	Mes	X_{6ik}	Permite observar variaciones mensuales de la demanda.

Explicativa temporal	Día de la semana	X_{7ik}	Captura diferencias de comportamiento según el día.
Explicativa temporal	Día del mes	X_{8ik}	Controla posibles variaciones dentro del mes.
Explicativa binaria temporal	Fin de semana	X_{9ik}	Diferencia ventas de fin de semana frente a días laborales
Explicativa categórica transformada	Tipo de cliente	X_{10ik}	Variable dummy creada mediante codificación binaria.
Explicativa categórica transformada	Local de venta	X_{11ik}	Variable dummy creada mediante codificación binaria.

Nota. Elaboración propia a partir de las variables definidas en la investigación.

Para incorporar las variables categóricas al modelo, estas se codifican de forma binaria. La variable Tipo de cliente_Natural adopta el valor 1 cuando el registro pertenece a un cliente natural, y 0 cuando corresponde a la categoría de referencia, entonces Local de venta_Principal se asume el valor 1 si la venta se realizó en el local principal, y 0 si pertenece a la categoría de referencia.

Debido a esta transformación es posible integrar información de naturaleza cualitativa dentro de una estructura numérica compatible con la estimación del modelo (James, G., 2021).

En cada clúster k se define “y” como el vector que representa el volumen de venta, y X como la matriz de variable explicativas conformada por los predictores disponibles para ese segmento, excluidas las columnas las correspondencia con el procedimiento implementado en el script y se obtiene un modelo particular para cada grupo determinado mediante K-means.

7.2. Especificación matemática de la regresión

Si bien la regresión lineal simple funciona como referente conceptual para explorar la relación entre el volumen de ventas y un único factor explicativo, el eje del estudio radica en la regresión lineal múltiple. Esta última técnica permite examinar el comportamiento de la demanda evaluando el impacto simultáneo de diversos factores observables (Gujarati & Porter, 2010; Wooldridge, 2020).

$$Volumen_i = \beta_0 + \beta_1 Precio_i + \varepsilon_i$$

En esta expresión, $Volumen_i$ representa la cantidad vendida en el registro “i”, $Precio_i$ representa el precio de la transacción, β_0 representa el intercepto, β_1 es el coeficiente relativo al precio, y ε_i constituye el término de error. Con esta fórmula, β_1 mide la variación esperada en el volumen de ventas por cada cambio unitario en el precio. Por su parte, la especificación principal del estudio evalúa en conjuntos las variables comerciales, temporales y categóricas transformadas empleadas en el script

$$Volumen_{\text{i}} = \beta_0 + \beta_1 Precio_{\text{i}} + \beta_2 Cantidad_{\text{i}} + \beta_3 Descuento_{\text{i}} + \beta_4 Crédito_{\text{i}} + \beta_5 Año_{\text{i}}$$

$$+ \beta_6 Mes_{\text{i}} + \beta_7 DíaSemana_{\text{i}} + \beta_8 DíaMes_{\text{i}} + \beta_9 FinSemana_{\text{i}} + \beta_{10} ClienteNatural_{\text{i}}$$

$$+ \beta_{11} LocalPrincipal_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

Dado que la estimación se realiza al interior de cada segmento definido por K-means, la especificación integra el subíndice "k" para identificar el clúster respectivo.

$$Volumen_{\mathrm{ik}} = \beta_{0\mathrm{k}} + \beta_{1\mathrm{k}} Precio_{\mathrm{ik}} + \beta_{2\mathrm{k}} Cantidad_{\mathrm{ik}} + \beta_{3\mathrm{k}} Descuento_{\mathrm{ik}} + \beta_{4\mathrm{k}} Crédito_{\mathrm{ik}}$$

$$+ \beta_{5\mathrm{k}} Año_{\mathrm{ik}} + \beta_{6\mathrm{k}} Mes_{\mathrm{ik}} + \beta_{7\mathrm{k}} DíaSemana_{\mathrm{ik}} + \beta_{8\mathrm{k}} DíaMes_{\mathrm{ik}} + \beta_{9\mathrm{k}} FinSemana_{\mathrm{ik}}$$

$$+ \beta_{10\mathrm{k}} ClienteNatural_{\mathrm{ik}} + \beta_{11\mathrm{k}} LocalPrincipal_{\mathrm{ik}} + \varepsilon_{\mathrm{ik}}$$

En la ecuación, el subíndice "i" denota cada registro comercial y "k" indica el clúster al que este pertenece. Así, $\beta_{0\mathrm{k}}$ representa el intercepto correspondiente al clúster "k", cada βj_{k} refleja el coeficiente de una variable explicativa dentro de dicho segmento, y $\varepsilon_{\mathrm{ik}}$ Constituye el término de error, que capta la porción de volumen de ventas no explicada por los factores considerados. Esta estructura permite medir la variación esperada en las ventas ante el cambio en un predictor específico, asumiendo que el resto de las variables del modelo se mantiene constante.

7.3. Parámetros del modelo y significado de los coeficientes

Los coeficientes calculados identifican el sentido y la magnitud de la variación prevista en el volumen de ventas para cada variable explicativa, asumiendo constantes los demás predictores. Tratándose de las variables binarias transformadas, el parámetro expresa la diferencia esperada con respecto a la categoría de referencia. La interpretación conceptual de estos valores se sintetiza en el cuadro correspondiente.

Tabla 5.

Parámetros y variables del modelo de regresión.

Parámetro	**Variable**	**Significado**
β_{0k}	Intercepto	Valor esperado del volumen de venta cuando las variables explicativas toman valor cero o se encuentran en sus categorías de referencia.
β_{1k}	Precio de la venta	Cambio esperado en el volumen de venta ante una variación de una unidad en el precio, manteniendo constantes las demás variables.
β_{2k}	Cantidad de productos	Cambio esperado en el volumen de venta asociado con una variación de una unidad en la cantidad de productos, manteniendo constantes los demás factores.

β_{3k}	Descuento	Diferencia esperada del volumen de venta entre registros con descuento y registros sin descuento, manteniendo constantes las demás variables.
β_{4k}	Crédito	Diferencia esperada del volumen de venta entre registros asociados a crédito y registros sin crédito, manteniendo constantes las demás variables.
β_{5k} - β_{9k}	Variables temporales	Cambios asociados con año, mes, día de la semana, día del mes, condición de fin de semana, controlando los demás predictores.
β_{10k}	Tipo de cliente_Natural	Diferencia esperada respecto de la categoría de referencia del tipo de cliente.
β_{11k}	Local de venta_PRINCIPAL	Diferencia esperada respecto

		de la categoría de referencia del local de venta.
ε_{ik}	Error	Componente del volumen de venta que no es explicado por las variables incluidas en el modelo.

Nota. Elaboración propia a partir de la especificación del modelo de regresión planteado en la investigación.

8. Criterio de estimación y métricas de evaluación

8.1. Criterio de estimación mediante mínimos cuadrados ordinarios

La estimación de los coeficientes en cada modelo se realiza mediante mínimos cuadrados ordinarios, método que determina los parámetros encargados de minimizar la suma de los errores cuadrados entre las ventas reales y las estimadas. Dentro del *script*, este cálculo se ejecuta empleando el módulo LinearRegression de scikit-learn.

$$min\ \Sigma_i\ (Volumen_{ik}\ -\ VolumenEstimado_{ik})^2$$

Este criterio determina los coeficientes que optimizan el ajuste del modelo, basándose en la minimización del error cuadrático observado dentro del conjunto de datos empleado para el entrenamiento.

8.2. Predicción del volumen de venta

Tras la estimación de los coeficientes, el modelo proyecta el volumen de ventas para cada registro a partir de los valores observados en las variables explicativas. Dichas predicciones sirven

como base para contrastar las observaciones reales con los datos estimados en el conjunto de pruebas.

$$VolumenEstimado_{ik} = \beta_{0k} + \beta_{1k}Precio_{ik} + \beta_{2k}Cantidad_{ik} + \beta_{3k}Descuento_{ik} + \beta_{4k}Crédito_{ik} + \beta_{5k}Año_{ik} + \beta_{6k}Mes_{ik} + \beta_{7k}DíaSemana_{ik} + \beta_{8k}DíaMes_{ik} + {}_{9k}FinSemana_{ik} + \beta_{10k}ClienteNatural_{ik} + \beta_{11k}LocalPrincipal_{ik}$$

8.3. División de datos para evaluación predictiva

Con el fin de evaluar la capacidad predictiva, los datos de cada clúster se dividen en subconjuntos de entrenamiento y prueba. Un 80% de las observaciones se destina al ajuste del modelo, mientras que el 20% restante se reserva para medir la precisión sobre registros que no intervinieron en el cálculo de los coeficientes. Esta partición permite valorar el desempeño del algoritmo frente a datos no vistos previamente, en línea con el enfoque de aprendizaje automático supervisado

8.4. Coeficiente de determinación R^2

El coeficiente de determinación R^2 indica qué proporción de la variabilidad en el volumen de ventas resulta explicada por las variables integradas en el modelo:

$$R^2 = 1 - [\Sigma_i(Volumen_{ik} - VolumenEstimado_{ik})^2 / \Sigma_i(Volumen_{ik} - VolumenPromedio_k)^2]$$

El coeficiente de determinación cercano a 1 indica una mayor capacidad explicativa, si este es negativo (como ocurre en uno de los clusters del script), indica que la capacidad predictiva del modelo resulta inferior a la estimada, basada únicamente en el promedio del volumen de ventas.

8.5. Error cuadrático medio (MSE)

El MSE calcula el promedio de las desviaciones de predicción elevadas al cuadrado. Al elevar los residuos a esta potencia, las imprecisiones de mayor magnitud adquieren un peso proporcionalmente superior en el indicador resultante.

$$MSE = (1/n)\,\Sigma_{\mathrm{i}}(Volumen_{\mathrm{ik}} - VolumenEstimado_{\mathrm{ik}})^2$$

8.6. Raíz del error cuadrático medio (RMSE)

La raíz del error cuadrático medio se obtiene calculando la raíz cuadrada del MSE, lo que permite expresar el margen de error en la misma unidad de medida que el volumen de ventas.

$$RMSE = \sqrt{MSE}$$

Al mantener la misma escala de la variable dependiente, el RMSE simplifica la interpretación sobre la magnitud del error predictivo.

8.7. Residuos del modelo

El residuo representa la diferencia entre el volumen de ventas observado y el valor estimado por la regresión. Esta medida permite medir el error individual correspondiente a cada observación, lo que complementa la evaluación sobre el desempeño del modelo.

$$e_{\mathrm{ik}} = Volumen_{\mathrm{ik}} - VolumenEstimado_{\mathrm{ik}}$$

Los residuos se utilizarán como una medida complementaria para contrastar los valores observados con los estimados dentro de cada clúster, sin incluir en esta etapa una interpretación causal de las diferencias obtenidas.

8.8. Significancia estadística mediante p-values

El script utiliza statsmodels para calcular los p-values de los coeficientes estimados. Dichos valores se unen como un criterio complementario de evaluación estadística, pues permiten comprobar si la relación identificada para cada variable posee evidencia estadística en el modelo correspondiente. Su análisis se realizará por clúster y de forma conjunta con los coeficientes calculados y las métricas predictivas, sin incluir cifras concretas dentro del apartado metodológico.

ANÁLISIS DE RESULTADOS

Primeramente, para realizar el proceso de análisis de la base de datos de los supermercados se utilizó la herramienta Google collab el cual es un servicio gratis que se ubica en la nube de google el cual ayuda a cualquier usuario a ejecutar mediante el código python desde cualquier rede de internet (Galicia, 2024).

En esta investigación se utilizó esta herramienta para el conjunto de datos del supermercado de consumo masivo.

Paso 1. Creación del título de trabajo e importación de las librerías.

```
Análisis de la demanda de clientes bajo factores significativos clusterizados

#Cargo las librerías

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

import plotly.express as px

import plotly.graph_objects as go

#Kmeans para la clasificación

from sklearn.linear_model import LinearRegression
```

```
from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import silhouette_score, mean_squared_error

from sklearn.model_selection import train_test_split

from scipy.spatial.distance import cdist

import statsmodels.api as sm # Importar statsmodels
```

Interpretación:

Al momento de escribir el código de importar librerías, estas son cajas de herramientas que se usan para hacer graficas, leer archivos, etc. Ahora bien, cada línea importa un librería distinta, por ende se explicara cada una:

Pandas (pd) : Librería que maneja base de datos en forma de tablas llamadas DataFrame. Esta librería es utilizada para revisar columnas, filtrar datos, en unir documentos de excel, etc.

Numpy (np) : Librería encargada de operaciones numéricas y arreglos matemáticas. Además permite usar la winsorización para reemplazar valores extremos mediantes estados numéricos.

seaborn (sn) : Librería para visualizar estadística basada en matplotlib, se usa para gráficos exploratorios y otras funciones.

matplotlip.pyplot (plt) : Librería utilizada para crear gráficos en python, ayuda a graficar estadísticamente los datos, además de visualizaciones de clusters y predicciones versus valores reales.

plotly.express y graph_objects : Libreria para gráficos interactivos, permite visualizar patrones de datos de forma más dinámica. aunque los principales para el analisis de graficos son matplotlib/ seaborn

LinearRegression : Modelo de regresión lineal de scikit-learn, entrena regresiones por cluster para predecir o explicar el volumen de venta.

Scikit-learn (sklearn) : Es parte de la librería de python para el aprendizaje automatizado, permite aplicar los algoritmos de aprendizaje supervisado y no supervisado.

KMeans : Es un algoritmo de aprendizaje no supervisado, agrupa registros comerciales en clusters según similitud de variables.

StandardScaler : Método de estandarización, escala las variables para que K-means compare distancias sin que una variable define por tener mayor magnitud.

Silhoutte_score : Métrica para evaluar separación de clusters, aunque no se utilizó en todo el proceso se añade como validación complementaria

mean_squared_error : Métrica de error promedio cuadrático, calcula el MSE de cada regresión por cluster.

train_test_split : Función para dividir datos en entrenamiento y prueba, separa el 80% de los datos para entrenar y 20% para evaluar la regresión.

cdist : Función para calcular distancias entre puntos, se usa para cálculos de distancia.

statsmodels.api (sm) : Libreria estadística para modelos econometricos

Paso 2. Integración y consolidación de la información comercial

El procesamiento de la información comercial se inició mediante Google Colab, utilizando el archivo Excel que contiene los registros de las ventas del supermercado. Para acceder al documento desde el entorno de trabajo, primero se estableció la conexión con Google Drive mediante el siguiente código:

```
#Montamos los datos en drive

from google.colab import drive

drive.mount('/content/drive')
```

Interpretación:

El código permite conectar Google Colab con Google Drive, donde se encuentra almacenado el archivo utilizado para el análisis. La función `drive.mount()` habilita el acceso al contenido de la unidad de almacenamiento desde el entorno de Python. Este procedimiento permite trabajar con la información comercial directamente desde el archivo utilizado en la investigación.

Una vez establecida la conexión, se indicó la ubicación del archivo Excel y se procedió a leer las dos hojas que contienen los registros de la base de datos:

```
# Ruta del archivo Excel en Colab

excel_path = '/content/drive/MyDrive/BASE_UNIFICADA_SUPERMERCADO
SCRIPT.xlsx'

# Cargar la primera hoja en un DataFrame

df_sheet1 = pd.read_excel(excel_path, sheet_name=0)
```

```
# Cargar la segunda hoja en un DataFrame

df_sheet2 = pd.read_excel(excel_path, sheet_name=1)
```

Interpretación:

Dentro de este segmento del código, la variable `excel_path` define la ruta donde se ubica el archivo Excel que será utilizado. A partir de ahí, la función `pd.read_excel()` se encarga de importar el contenido de cada hoja para que pueda procesarse dentro del entorno de Python. El parámetro `sheet_name=0` hace referencia a la primera hoja del archivo, en tanto que `sheet_name=1` corresponde a la segunda.

El contenido de cada hoja queda almacenado temporalmente en estructuras independientes, nombradas `df_sheet1` y `df_sheet2`, lo que permite manipular los registros en formato de tabla y facilita su posterior integración y tratamiento.

Una vez leídas ambas hojas, se procedió a unificar sus registros en una sola estructura de datos por medio de la función `concat()`:

```
# Unificar los DataFrames (asumiendo que tienen las mismas columnas o
se desea apilar)

df_unificado = pd.concat([df_sheet1, df_sheet2], ignore_index=True)

# Mostrar las primeras 5 filas del DataFrame unificado

display(df_unificado.head())
```

Interpretación:

Con la función `pd.concat()` se logró unir los registros de `df_sheet1` y `df_sheet2` en una única base única, denominada `df_unificado`. El parámetro `ignore_index=True` se encargó de reorganizar el índice de los registros, de modo que no se conservará la numeración independiente que traía cada hoja.

A continuación, mediante `display(df_unificado.head())` se visualizaron las primeras cinco filas, lo que permitió verificar que la integración se hubiera realizado correctamente y revisar la estructura inicial de las variables. La base resultante quedó conformada por 1.452.837 registros y 15 columnas, entre las que figuran nombre del cliente, local de venta, cantón, cantidad y tipo de productos, fecha de venta, tipo de cliente, volumen de venta, precio, descuento, crédito y tipo de crédito.

Gracias a esta integración fue posible consolidar toda la información en una única base de trabajo, sentando así las bases para continuar con las etapas de preparación y análisis necesarias para la segmentación de clientes.

	Nombre del cliente	cedula	no.factura	Local de venta	Cantón	Cantidad de productos	Tipo de productos	Fecha de venta	Tipo de cliente	Volumen de venta	Precio de la venta	Descuento (variable binaria)	Tipo de descuento	Crédito (variable binaria	Tipo de crédito
0	PEREZ YUNGA ANA PRISCILLA	NaN	NaN	PRINCIPAL	Pasaje	1.0	CAFE	22/09/2025	Jurídico	11.99	10.4342	0	NaN	1.0	160.0
1	PRADO PAUCAY JOSE ARTURO	NaN	NaN	PRINCIPAL	Buena vista	1.0	CAFE	16/09/2025	Jurídico	599.98	521.7200	0	NaN	1.0	6000.0
2	PRADO PAUCAY JOSE ARTURO	NaN	NaN	PRINCIPAL	Buena vista	4.0	CAFE	16/09/2025	Jurídico	126.01	27.3920	0	NaN	1.0	6000.0
3	ORDOÑEZ JARA ROSA ANGELICA	NaN	NaN	PRINCIPAL	Pasaje	1-0	CAFE	18/09/2025	Jurídico	11.99	10.4342	0	NaN	1.0	200.0
4	ARMIJOS ZHIGUE ANGEL ARMANDO	NaN	NaN	PRINCIPAL	Pasaje	20.0	CAFE	18/09/2025	Natural	8.00	0.3478	0	NaN	1.0	250.0

Con referente a la tabla, el resultado es una pequeña muestra del excel sobre las variables que conforman el estudio del supermercado con sus respectivos datos menos en el número de cédula y número de factura puesto a que esa información no fue proporcionada.

Paso 3. Limpieza de Datos (Análisis Exploratorio de Datos)

```
# Mostrar información general del DataFrame

df_unificado.info()

# Mostrar estadísticas descriptivas de las columnas numéricas

display(df_unificado.describe())
```

Interpretación:

Desde este paso del EDA se empieza a analizar la base de datos así como conocer la estructura antes de explicar algún modelo. En el script se utiliza los siguientes códigos:

`df_unificado.info() :` Se utiliza para revisar número de registros, columnas, valores no nulos y tipo de datos de la base de datos unificada.

`display() :` Se utiliza para mostrar el resultado en forma de tabla en google colab.

`display(df_unificado.describe()) :` Se interpreta como la descripción numérica de la base de datos unificada, puesto que `describe` es una función que calcula estadísticas descriptivas de las columnas numéricas.

Resultados :

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1452837 entries, 0 to 1452836
```

```
Data columns (total 15 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   Nombre del cliente              1452837 non-null  object
 1   cedula                          0 non-null        float64
 2   no. factura                     0 non-null        float64
 3   Local de venta                  1452837 non-null  object
 4   Cantón                          1019839 non-null  object
 5   Cantidad de productos           1452402 non-null  float64
 6   Tipo de productos               1452834 non-null  object
 7   Fecha de venta                  1452837 non-null  object
 8   Tipo de cliente                 1444677 non-null  object
 9   Volumen de venta                1452837 non-null  float64
 10  Precio de la venta              1452837 non-null  float64
 11  Descuento (variable binaria)    1452837 non-null  int64
 12  Tipo de descuento               0 non-null        float64
 13  Crédito (variable binaria)      1444677 non-null  float64
 14  Tipo de crédito                 1444677 non-null  float64
```

```
dtypes: float64(8), int64(1), object(6)
```

	cedula	no. factura	Cantidad de productos	Volumen de venta	Precio de la venta	Descuento (variable binaria)	Tipo de descuento	Crédito (variable binaria)	Tipo de crédito
count	**0.0**	**0.0**	**1.452402e+06**	**1.452837e+06**	**1.452837e+06**	**1.452837e+06**	**0.0**	**1.444677e+06**	**1.444677e+06**
mean	**NaN**	**NaN**	**2.163232e+00**	**3.550676e+00**	**2.064710e+00**	**1.931669e-02**	**NaN**	**1.778155e-01**	**5.819433e+02**
std	**NaN**	**NaN**	**9.471300e+00**	**2.582808e+01**	**5.100378e+00**	**1.376356e-01**	**NaN**	**3.823575e-01**	**5.188587e+03**
min	**NaN**	**NaN**	**4.700000e-01**	**0.000000e+00**	**7.350000e-02**	**0.000000e+00**	**NaN**	**0.000000e+00**	**0.000000e+00**

25%	**NaN**	**NaN**	**1.000000e+00**	**9.000000e-01**	**6.500000e-01**	**0.000000e+00**	**NaN**	**0.000000e+00**	**0.000000e+00**
50%	**NaN**	**NaN**	**1.000000e+00**	**1.500100e+00**	**9.000000e-01**	**0.000000e+00**	**NaN**	**0.000000e+00**	**0.000000e+00**
75%	**NaN**	**NaN**	**2.000000e+00**	**2.700000e+00**	**1.652200e+00**	**0.000000e+00**	**NaN**	**0.000000e+00**	**0.000000e+00**
max	**NaN**	**NaN**	**1.000000e+04**	**2.481394e+04**	**1.464000e+03**	**1.000000e+00**	**NaN**	**1.000000e+00**	**8.225000e+04**

Interpretación:

En primer lugar se analiza los tipos de variables que encontró el script, las variables tipo Object son textuales o categoricas, en el resultado son:

- Nombre de cliente
- Local de venta
- Cantón
- Tipo de productos
- Fecha de venta
- Tipo de cliente

Los datos de esta variables se utilizan para identificar características comerciales, geográficas, temporales y de cliente.

Las variables de tipo float 64 son numéricas con decimales, las cuales son:

- Cantidad de productos
- Volumen de venta
- Precio de venta
- Crédito
- Tipo de crédito

La variable de tipo int64 la cual es numérica entera son:

- Descuento (variable binaria)

Los valores nulos identificados del resultado de la descripción de la base de datos son:

- Cedula tiene 0 datos válidos puesto no hay valores en la base de datos
- No. Factura tiene 0 datos válidos por la misma razón
- El tipo de descuento tiene 0 datos válidos.

Se interpretan como columnas vacías y no aportan información útil para el análisis.

A continuación, se explican los valores válidos de las otras variables.

- Cantón : tiene 1.019.839 datos válidos, pero la base total tiene 1.452.837 registros.
- Cantidad de productos: tiene 1.452.402 datos válidos
- Tipo de cliente, crédito y tipo de crédito: tiene 1.444.677 datos válidos.

Interpretación de la tabla:

A continuación, se explicara el análisis de las variables en las tabla mostrada.

Cantidad de productos

Esta variable contiene un promedio aproximado de 2 productos por registro. Sin embargo existen valores atípicos como 10.000.

Volumen de venta

Esta variable tiene un promedio aproximado de 3.55 con la mediana de 1.50 y el valor máximo llega a 24.813,94 los cuales son valores atípicos.

Precio de la venta

Los datos de esta variable son el promedio aproximado es 2.06, una mediana de 0.90 y un máximo de 1.464,00 demostrando que la mayoría de los precios son bajos pero existen registros de precios altos.

Descuento (variable binaria)

Esta variable es binaria, donde 0 indica ausencia de descuento y 1 indica la presencia de descuento. Su promedio de 0.0193 indica que el uso de descuento es bajo ya que solo una proporción pequeña de registro presenta descuento.

Crédito (variable binaria)

La variable presenta mayor frecuencia que el descuento, con un promedio de 0.1778. Esto indica que una parte relevante de la ventas se relaciona con el crédito del consumidor.

Tipo de crédito

La variable tiene un promedio de 581.94 pero un máximo de 82.250 que indica valores muy dispersos. Esto refleja que existen distintos montos de crédito asignados a clientes.

Este paso sirve para justificar la necesidad de realizar una limpieza de datos y tratamiento de valores externos antes de aplicar modelos de clusterización y regresión.

Paso 4. Limpieza de datos numéricos con el método del rango intercuartílico (IQR) y Winsorización

En este paso se manejan los valores atípicos en las variables numéricas utilizando el método de rango intercuartílico que usa la diferencia entre el tercer cuartil y el primer cuartil. Luego, se aplica la técnica de Winsorización que reemplaza los valores de datos en los extremos de una distribución con valores menores extremos, generalmente en un percentil específico.

```
# Función para aplicar Winsorización basada en el IQR

def winsorize_iqr(df, column):
```

```
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    # Calcular límites para Winsorización
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Aplicar Winsorización: los valores por debajo del límite inferior se reemplazan por el límite inferior,
    # y los valores por encima del límite superior se reemplazan por el límite superior.
    df[column] = np.where(df[column] < lower_bound, lower_bound, df[column])
    df[column] = np.where(df[column] > upper_bound, upper_bound, df[column])
    return df

# Identificar columnas numéricas para aplicar la Winsorización
```

```
# Excluir 'Descuento (variable binaria)' y 'Crédito (variable
binaria)' si son de tipo entero y ya están limpias/categorizadas,
# ya que la Winsorización podría no ser apropiada para variables
binarias.
# También excluimos las columnas de fecha que hemos creado, ya que son
categóricas por naturaleza o rangos definidos (año, mes, día).

numeric_cols =
df_unificado.select_dtypes(include=np.number).columns.tolist()

# Excluir columnas binarias o de fecha/tiempo ya tratadas
exclude_cols = [
    'Descuento (variable binaria)',
    'Crédito (variable binaria)',
    'Ano', 'Mes', 'Dia_Semana', 'Dia_Mes', 'Es_Fin_Semana', 'Cluster'
]

# Filtrar las columnas numéricas para aplicar Winsorización
```

```
columns_to_winsorize = [col for col in numeric_cols if col not in
exclude_cols]


print(f"Columnas numéricas a las que se aplicará Winsorización:
{columns_to_winsorize}")


# Aplicar Winsorización a las columnas seleccionadas
for col in columns_to_winsorize:
    df_unificado = winsorize_iqr(df_unificado, col)
    print(f"Winsorización aplicada a la columna: {col}")


print("\nValores estadísticos del DataFrame después de la
Winsorización:")
display(df_unificado[columns_to_winsorize].describe())
```

Resultados:

Columnas numéricas a las que se aplicará Winsorización: ['cedula', 'no. factura', 'Cantidad de productos', 'Volumen de venta', 'Precio de la venta', 'Tipo de descuento', 'Tipo de crédito']

Winsorización aplicada a la columna: cedula

Winsorización aplicada a la columna: no. factura

Winsorización aplicada a la columna: Cantidad de productos

Winsorización aplicada a la columna: Volumen de venta

Winsorización aplicada a la columna: Precio de la venta

Winsorización aplicada a la columna: Tipo de descuento

Winsorización aplicada a la columna: Tipo de crédito

Valores estadísticos del DataFrame después de la Winsorización:

	cedula	no. factura	Cantidad de productos	Volumen de venta	Precio de la venta	Tipo de descuento	Tipo de crédito
count	0.0	0.0	1.452402e+06	1.452837e+06	1.452837e+06	0.0	1444677.0
mean	NaN	NaN	1.540028e+00	2.059832e+00	1.255752e+00	NaN	0.0
std	NaN	NaN	9.052119e-01	1.574956e+00	8.825328e-01	NaN	0.0
min	NaN	NaN	4.700000e-01	0.000000e+00	7.350000e-02	NaN	0.0
25%	NaN	NaN	1.000000e+00	9.000000e-01	6.500000e-01	NaN	0.0
50%	NaN	NaN	1.000000e+00	1.500100e+00	9.000000e-01	NaN	0.0
75%	NaN	NaN	2.000000e+00	2.700000e+00	1.652200e+00	NaN	0.0

max	NaN	NaN	3.500000e+00	5.400000e+00	3.155500e+00	NaN	0.0

Interpretación

Primero, la función `def winsorize_iqr(df, column):` significa que se aplicará la winsorización a una columna específica del DataFrame. En este caso, df representa la base de datos y column representa la variable numérica que se va a limpiar.

Segundo, el script utiliza el primer cuartil Q1, el tercer cuartil Q3 y el rango intercuartílico IQR. Esta función sirve para conocer donde se concentra la parte central de los datos. En otras palabras, se identifica el rango donde se encuentra la mayoría de valores normales de la variable.

Tercero, se calcula los límites para la winsorización mediante la función `lower_bound = Q1 - 1.5 * IQR ; upper_bound = Q3 + 1.5 * IQR` se interpreta como límites inferior y superior. Si un dato está por debajo del límite inferior o por encima del limite superior, se considera un posible valor atípico.

Cuarto, se aplica la winsorización que no elimina los valores extremos, sino que los reemplaza por el límite permitido. En este caso, supongamos que un valor de volumen de venta es demasiado alto, no se borra, sino que se reduce hasta el límite superior. Esto permite conservar los registros sin permitir que valores extremos distorsionen el modelo.

Quinto, se empieza con la selección de columnas numéricas en el cual se identifica las columnas numéricas el código `numeric_cols df_unificado.select_dtypes(include=np.number).columns.tolist()` esto busca todas las columnas que son numéricas dentro de la base, como cantidad de productos, volumen de venta, precio de venta, crédito, descuento, etc.

Sexto, se define columnas que se deben excluir las cuales son las variables descuento (variable binaria), crédito (variable binaria), Año, mes, etc. Estas se excluyen porque no conviene aplicar la winsorización a variables binarias o temporales.

Séptimo, se filtra las columnas por limpiar con el siguiente código `columns_to_winsorize = [col for col in numeric_cols if col not in exclude_cols]` , esto crea una lista con las columnas numéricas que si recibiran winsorización.

Octavo paso, se aplica la winsorización que se aplica el proceso a cada columna seleccionada, con el siguiente código `for col in columns_to_winsorize:`

```
    df_unificado = winsorize_iqr(df_unificado, col)

    print(f"Winsorización aplicada a la columna: {col}")
```

Esto significa que el algoritmo revisa una por una las variables numéricas seleccionadas y corrige sus valores extremos.

En tabla mostrada de los resultados se aplicó la winsorización a las siguientes variables: Cedula, no. factura, cantidad de producto, volumen de venta, precio de la venta, tipo de descuento y tipo de crédito.

Sin embargo, cédula y no.factura y tipo de descuento aparecen con 0 datos válidos o Nan, por ende, no aportan información útil para el análisis, las variables esenciales en la tesis son; cantidad de productos, volumen de venta, precio de la venta y tipo de crédito.

Por último, en este paso se explica el resultado después de la winsorización con el código `display(df_unificado[columns_to_winsorize].describe())` , esto permite comparar cómo quedaron las variables después de controlar valores extremos. Por ejemplo, antes

el valor máximo de cantidad de productos era 10.000, pero después de la winsorización al máximo bajó a 3.5. Eso significa que el script redujo los valores extremos para que no afecten el análisis. En volumen de venta, el valor máximo quedó en 5.4 y en precio de venta, el máximo quedó en 3.1555. Esto indica que los valores demasiado altos fueron ajustados a límites más razonables según el método IQR.

Paso 5. Revisión de valores nulos

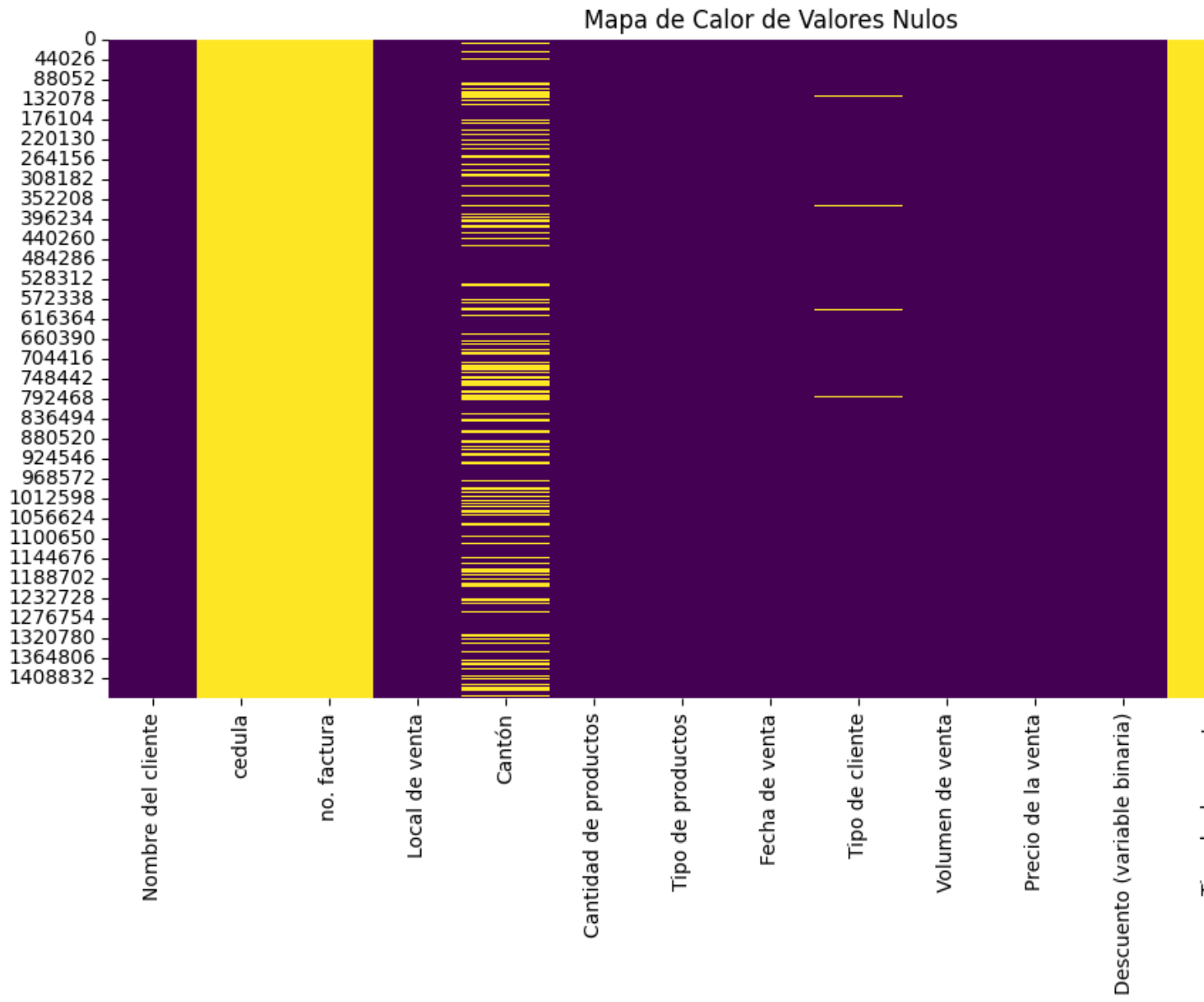
```
missing_values = df_unificado.isnull().sum()

missing_values = missing_values[missing_values >
0].sort_values(ascending=False)


print("Valores nulos por columna:")

display(missing_values)


# Visualizar valores nulos usando un mapa de calor

plt.figure(figsize=(12, 6))

sns.heatmap(df_unificado.isnull(), cbar=False, cmap='viridis')

plt.title('Mapa de Calor de Valores Nulos')

plt.show()
```

Resultado:

Valores nulos por columna:

	0
cedula	**1452837**
no. factura	**1452837**
Tipo de descuento	**1452837**
Cantón	**432998**
Tipo de cliente	**8160**
Crédito (variable binaria)	**8160**
Tipo de crédito	**8160**
Cantidad de productos	**435**
Tipo de productos	**3**

dtype: int64



Interpretación del código:

La primera línea : `missing_values = df_unificado.isnull().sum()`

Esta línea revisa toda la base de datos unificadas y cuenta cuántos valores nulos tiene cada columna, por ende .`isnull()` identifica las celdas vacías o nulas y `.sum()` que significa la suma de cuantos nulos hay por cada columna.

La segunda línea: `missing_values = missing_values[missing_values > 0].sort_values(ascending=False)`

Esta línea filtra solo las columnas que sí tienen valores nulos. En este caso, `missing_values > 0` muestra solo columnas con datos faltantes, `sort_values(ascending=False)` ordena de mayor a menor cantidad de nulos.

Esto permite identificar rapidamente cuales variables tienen mas problemas de información completa.

En la tercera linea, este bloque genera un mapa de calor para visualizar los valores nulos, `plt.figure(figsize=(12, 6))` define el tamaño del grafico, `sns.heatmap(...)` crea el mapa de calor; `df_unificado.isnull()` marca visualmente donde hay valores nulos; `cbar=False` elimina la barra lateral de colores; `cmap='viridis` se usa una escala de color; `plt.title` colocal el titulo del grafico; `plt.show()` muestra la imagen final.

Interpretación de los resultados del grafico de valores nulos

En la tabla y en el gráfico se observa que la variable cedula, numero de factura y tipo de descuento esta completamente vacia de datos, por otro lado el cantón tiene bastante datos faltantes pero no esta completamente vacia, tipo de cliente tiene pocos datos faltantes frente al total, credito tambien tiene poco datos faltantes, cantidad de producto tiene muy pocos datos faltantes a comparación de las otras variables y por ultimo, tipo de producto tiene solo tres datos faltantes siendo la mas completa.

Las columnas totalmente vacias de color amarillo deben eliminarse porque no aportan información, las columnas con poco nulos pueden limpiarse, las columnas con muchos nulos, como cantón, deben revisarse con cuidado para decidir si se usan o no en el modelo.

Las variables usadas para el K-means y regresión no deben tener datos vacíos, porque pueden generar errores o afectar los resultados.

Paso 6. Examinación de cantidad de valores unicos en cada columna.

En este paso se examina la cantidad de valores en cada variable para entender la cardinalidad de las variables, esto sirve para identificar que columnas podrian no ser utiles para el modelado directamente.

```
# Mostrar el número de valores únicos por columna

print("Número de valores únicos por columna:")

display(df_unificado.nunique().sort_values(ascending=False))
```

Resultado:

```
Número de valores únicos por columna:
```

	0
Nombre del cliente	11469
Volumen de venta	2597
Precio de la venta	930

Tipo de productos	746
Fecha de venta	271
Cantón	127
Cantidad de productos	98
Descuento (variable binaria)	2
Crédito (variable binaria)	2
Local de venta	2
Tipo de cliente	2
Tipo de crédito	1
no. factura	0
cedula	0
Tipo de descuento	0

```
dtype: int64
```

Interpretación

En la primera linea `df_unificado.nunique()`cuenta cuantos valores unicos tiene cada columna del DataFrame, en otras palabras revisa cuantas categorias o datos diferentes existen en cada variable.

Luego, `sort_values(ascending=False)` ordena los resultados de mayor a menor, para ver primero las columnas con más variedad de datos.

Finalmente, `display(...)` muestra el resultado como tabla.

Interpretacion de los resultados

En el caso del estudio, el resultado muestra que la con mas valores unicos es nombre de cliente, con 11.469 valores diferentes. Esto indica que la base tiene muchos clientes registrados.

Despues la variable Volumen de venta con 2.597 son valores unicos y Precio de venta, con 930 valores unicos. Esto se interpreta que las variables presentan alta variabilidad, lo cual es util para el analisis puesto que permiten diferenciar comportamiento de compra.

La variable Tipo de producto tiene 746 valores unicos, indicando una gran variedad de productos vendidos en el supermercado. Es de suma importancia este dato debido que permite analizar la demanda según categorías o tipos de producto.

La variable Fecha de venta tiene 271 valores unicos, lo cual muestra que la base contiene ventas registradas en 271 fechas diferentes, Esto permite analizar el comportamiento de la demanda en el tiempo.

La variable cantón tiene 127 valores unicos, lo cual muestra diversidad geografica en los registros, aunque tiene muchos valores nulos por lo que debe limpiarse o revisarse antes de usarla.

La variable Cantidad de producto tiene 98 valores únicos, indicando que existen diferentes cantidades vendidas por registro.

La variable Descuento, Credito, Local de venta y Tipo de cliente tien 2 valores unicos, lo cual confirma que son variables categoricas o binarias.

Finalmente, tipo de credito tiene solo 1 valor unico, y las columnas numero de factura, cedula y tipo de descuento tiene cero valores unicos, puesto que estan vacias o no contienes datos valido, por ende, deben eliminarse puesto que no aportan información.

Paso 7. Ingeniera de caracteristicas temporales

Se utiliza la columna Fecha de venta para crear nueva caracteristicas que podrian influir en la demanda, como el año, el mes, el dia de la semana y si es fin de semana.

```
df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de venta'])

# Extraer año, mes, día de la semana y día del mes

df_unificado['Ano'] = df_unificado['Fecha de venta'].dt.year

df_unificado['Mes'] = df_unificado['Fecha de venta'].dt.month

df_unificado['Dia_Semana'] = df_unificado['Fecha de venta'].dt.dayofweek # Lunes=0, Domingo=6
```

```
df_unificado['Dia_Mes'] = df_unificado['Fecha de venta'].dt.day


# Crear una variable binaria para indicar si es fin de semana
df_unificado['Es_Fin_Semana'] =
df_unificado['Dia_Semana'].apply(lambda x: 1 if x >= 5 else 0)


# Eliminar la columna 'Fecha de venta' original si ya no es necesaria,
o mantenerla según se considere.
# Por ahora, la mantendremos para referencia, pero la eliminaremos
antes del modelado si no es una característica directa.


print("Primeras filas del DataFrame con las nuevas características
temporales:")
display(df_unificado[['Fecha de venta', 'Ano', 'Mes', 'Dia_Semana',
'Dia_Mes', 'Es_Fin_Semana']].head())
```

/tmp/ipykernel_984/843679481.py:1: UserWarning: Parsing dates in %d/%m/%Y format when dayfirst=False (the default) was specified. Pass `dayfirst=True` or specify a format to silence this warning.

```
  df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de venta'])
```

Primeras filas del DataFrame con las nuevas características temporales:

	Fecha de venta	Año	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana
0	2025-09-22	2025	9	0	22	0
1	2025-09-16	2025	9	1	16	0
2	2025-09-16	2025	9	1	16	0
3	2025-09-18	2025	9	3	18	0
4	2025-09-18	2025	9	3	18	0

Interpretación

El codigo en la primera linea es `df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de venta'])` convierte la columna Fecha de venta a formato de fecha reconocido por python. Es decir, pandas deja de leerla como texto y la trasnforma en una variable tipo fecha.

Luego el script crea nuevas columnas en donde separa año, mes, dia de la semana y dia del mes. En especifico Dia_Semana se codifica asi:

lunes : 0

Martes: 1

Miercoles: 2

jueves: 3

Viernes: 4

Sabado: 5

Domingo: 6

Luego, la siguiente linea es `df_unificado['Es_Fin_Semana'] = df_unificado['Dia_Semana'].apply(lambda x: 1 if x >= 5 else 0)` esto significa la creación de una variable binaria llamada Es_Fin_Semana, se interpreta si la venta fue sabado o domingo, coloca 1. Si fue de lunes a viernes, coloca 0.

Esto sirve para analizar si el comportamiento de compra cambia entre dias normales y fines de semana.

Interpretación de la tabla

El resultado muestra las primeras filas del Dataframe con las nuevas variables temporales. Por ejemplo, la fecha 2025-09-22 en el script identifica que:

- El año es 2025
- el mes es 9, es decir, septiembre
- el dia de la semana 0, es decir, lunes
- el dia del mes es 22
- en la nueva variables Es_Fin_Semana marca 0 , por lo cual no es fin de semana

Paso 8. Manejo de valors nulos restantes

Se procede a imputar los valores nulos que aun persisten en el DataFrame. Se utiliza la mediana para columnas numéricas y la moda para columnas categóricas o binarias.

```
# Imputar valores nulos en 'Cantidad de productos' con la mediana

df_unificado['Cantidad de productos'] = df_unificado['Cantidad de
productos'].fillna(df_unificado['Cantidad de productos'].median())



# Imputar valores nulos en 'Tipo de crédito' con la mediana

df_unificado['Tipo de crédito'] = df_unificado['Tipo de
crédito'].fillna(df_unificado['Tipo de crédito'].median())



# Imputar valores nulos en 'Cantón' con la moda

df_unificado['Cantón'] =
df_unificado['Cantón'].fillna(df_unificado['Cantón'].mode()[0])



# Imputar valores nulos en 'Tipo de productos' con la moda

df_unificado['Tipo de productos'] = df_unificado['Tipo de
productos'].fillna(df_unificado['Tipo de productos'].mode()[0])
```

```
# Imputar valores nulos en 'Tipo de cliente' con la moda

df_unificado['Tipo de cliente'] = df_unificado['Tipo de
cliente'].fillna(df_unificado['Tipo de cliente'].mode()[0])


# Imputar valores nulos en 'Crédito (variable binaria)' con la moda

df_unificado['Crédito (variable binaria)'] = df_unificado['Crédito
(variable binaria)'].fillna(df_unificado['Crédito (variable
binaria)'].mode()[0])


print("Valores nulos después de la imputación:")

display(df_unificado.isnull().sum()[df_unificado.isnull().sum() > 0])
```

Resultado

Valores nulos después de la imputación:

	0
cedula	**1452837**

no. factura	**1452837**
Tipo de descuento	**1452837**

dtype: int64

Interpretación

En la primera linea de este paso se usa el siguiente codigo `df_unificado['Cantidad de productos'] = df_unificado['Cantidad de productos'].fillna(df_unificado['Cantidad de productos'].median())` en este caso se interpreta como los valores nulos de la columna cantidad de productos se reemplazan por la mediana de esa misma variable.

La mediana representa el valor central de los datos y se utiliza debido que es menos sensible a valores extrems que el promedio.

Por consiguiente, se utiliza el mismo procedimiento pero con la variable tipo de credito con el siguiente codigo `df_unificado['Tipo de crédito'] = df_unificado['Tipo de crédito'].fillna(df_unificado['Tipo de crédito'].median())` esta completa los valores faltantes de esa columna con su valor central.

En esta primera parte, se manejan las variables numericas en las cuales se utiliza la mediana por la razón que esta medida permite reemplazar valores faltantes sin verse tan afectada por los

valores extremos. Para las variables cantidad de productos y tipo de credito es esencial porque pueden presentar alta dispersión.

La segunda parte es la imputación de variables categoricas con la moda con las variables de cantón, tipo de producto, tipo de cliente y credito. Por ejemplo, el codigo para la variable cantón es `df_unificado['Cantón'] = df_unificado['Cantón'].fillna(df_unificado['Cantón'].mode()[0])` esto significa que los valores vacios de canton se reemplazan por el cantón que mas se repite en la base.

Asimismo, se utilizo con las otras variables categoricas y se utiliza la moda debido que estas variables no son continuas.

Tercero, la revisión despues de la imputación, la linea final es `display(df_unificado.isnull().sum()[df_unificado.isnull().sum() > 0])` esto significa que el algoritmo cuenta los valores faltantes y solo muestra las columnas que todavia tienen nulos.

Interpretación del resultado

El resultado muestra que despues de la imputación todavia quedan nulos en la columna de cedula, numero de factura y tipo de descuento.

Sin embargo, el resultado indica que la imputación funcionó para las variables que si tenian información parcial que son las variables númericas y categoricas.

La imputación se realiza para evitar que los modelos posteriores, como K-means y regresión lineal multiple, fallen o pierdan registros por tener datos vacios.

Paso 9. Limpieza y Preprocesamiento de Datos

Basandose en el analisis exploratorio de datos, el primero paso es eliminar las columnas que tiene 100% de valores nulos debido que no aportan información.

```
# Columnas a eliminar por tener todos sus valores nulos

columns_to_drop = ['cedula', 'no. factura', 'Tipo de descuento']



# Eliminar las columnas del DataFrame

df_unificado = df_unificado.drop(columns=columns_to_drop)



print(f"Columnas eliminadas: {columns_to_drop}")

print("Información del DataFrame después de eliminar columnas:")

df_unificado.info()
```

Resultado

```
Columnas eliminadas: ['cedula', 'no. factura', 'Tipo de descuento']

Información del DataFrame después de eliminar columnas:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1452837 entries, 0 to 1452836

Data columns (total 17 columns):
```

```
 #   Column                        Non-Null Count    Dtype
---  ------                        --------------    -----
 0   Nombre del cliente            1452837 non-null  object
 1   Local de venta                1452837 non-null  object
 2   Cantón                        1452837 non-null  object
 3   Cantidad de productos         1452837 non-null  float64
 4   Tipo de productos             1452837 non-null  object
 5   Fecha de venta                1452837 non-null  datetime64[ns]
 6   Tipo de cliente               1452837 non-null  object
 7   Volumen de venta              1452837 non-null  float64
 8   Precio de la venta            1452837 non-null  float64
 9   Descuento (variable binaria)  1452837 non-null  int64
 10  Crédito (variable binaria)    1452837 non-null  float64
 11  Tipo de crédito               1452837 non-null  float64
 12  Ano                           1452837 non-null  int32
 13  Mes                           1452837 non-null  int32
 14  Dia_Semana                    1452837 non-null  int32
 15  Dia_Mes                       1452837 non-null  int32
```

```
 16  Es_Fin_Semana                1452837 non-null  int64

dtypes: datetime64[ns](1), float64(5), int32(4), int64(2), object(5)

memory usage: 166.3+ MB
```

Interpretación

En la primera linea, `columns_to_drop = ['cedula', 'no. factura', 'Tipo de descuento']` significa donde se colocan las columnas que se van a eliminar. En este caso, se eliminan cédula, nuemero de factura y tipo de descuento porque en los pasoa anteriores se observo que no contenian datos validos en otras palabras vacios.

Segundo, la siguiente linea `df_unificado = df_unificado.drop(columns=columns_to_drop)` se elimina esas columnas en la base de datos unificado. Es decir, el dataframe queda solo con las variables que si tiene información útil para el análisis.

Tercero, el script utiliza la ultimas lineas en las cuales se interpretan en las siguientes lineas

```
print(f"Columnas eliminadas: {columns_to_drop}")

print("Información del DataFrame después de eliminar columnas:")

df_unificado.info()
```

Esto sirve para confirmar que las columnas fueron eliminadas correctamente y para revisar nuevamente la estructura de la base después de la limpieza.

Intepretacion del resultado

Despues de la eliminación, la base queda 1.452.837 registros y 17 columnas. Además, todas las columnas restantes aparecen con la misma cantidad de datos no nulos, lo que significa que ya no hay valores faltantes en las variables que se van a utilizar para el analisis. Las variables actualizadas son: Nombre de cliente, local de venta, cantón, cantidad de productos, tipo de productos, fecha de venta, tipo de cliente, volumen de venta, precio de venta, descuento, credito, tipo de credito, año, mes, dia de semana, dia del mes, etc.

Paso 10. Convertir la columna 'Fecha de venta' al tipo de dato 'datetime'

En paso se realiza para poder extraer información temporal importante de la 'Fecha de venta'.

```
# Convertir 'Fecha de venta' a datetime

df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de venta'])


print("Tipo de dato de 'Fecha de venta' después de la conversión:")

print(df_unificado['Fecha de venta'].dtype)


# Mostrar las primeras filas con la columna de fecha convertida

display(df_unificado.head())
```

Resultados

```
Tipo de dato de 'Fecha de venta' después de la conversión:
```

```
datetime64[ns]
```

	Nombre del cliente	Local de venta	Cantón	Cantidad de productos	Tipo de productos	Fecha de venta	Tipo de cliente	Volumen de venta	Precio de la venta	Descuento (variable binaria)	Crédito (variable binaria)	Tipo de crédito	Año	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana
0	PEREZ YUNGA ANA PRISCILLA	PRINCIPAL	Pasaje	1.0	CAFE	2025-09-22	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	0	22	0

1	PRADO PAUCA Y JOSE ARTURO	PRINCIPAL	Buenavista	1.0	CAFE	2025-09-16	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	1	16	0
2	PRADO PAUCA Y JOSE ARTURO	PRINCIPAL	Buenavista	3.5	CAFE	2025-09-16	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	1	16	0

3	ORDOÐEZ JARA ROSA ANGELICA	PRINCIPAL	Pasaje	1.0	CAFE	2025-09-18	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	3	18	0
4	ARMIJOS ZHIGUE ANGEL ARMANDO	PRINCIPAL	Pasaje	3.5	CAFE	2025-09-18	Natural	5.4	0.3478	0	1.0	0.0	2025	9	3	18	0

Interpretación

En la primera linea el codigo empieza con `df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de venta'])` significa que convierte la variable Fecha de venta al formato datetime. Antes, la fecha podia estar guardada como texto u objeto; al convertirla a datetime, el lenguaje de python la puede reconocer como una fecha real y trabajar correctamente.

La segunda linea es `print("Tipo de dato de 'Fecha de venta' después de la conversión:")`

`print(df_unificado['Fecha de venta'].dtype)` y `display(df_unificado.head())` estas tres lineas muestras el tipo de dato de la columna despues de la conversión y display muestra las primeras filas del Dataframe para verificar visualmente que la base quedo procesada correctamente despues de la limpieza y transformación de la variables.

Interpretación de resultados

El resultado en la tabla demuestra que la fecha de venta ahora tiene tipo de dato `datetime64[ns]`, lo cual permite trabajar con fechas dentro del analisis. Además, al mostrar las primeras cinco filas, se observa que la base ya conserva las variables necesarias para los modelos.

Paso 11. Preparación de Datos para K-Means Clustering.

Seleccionaremos las características relevantes, codificaremos las variables categóricas y escalaremos las características numéricas para la clusterización.

```
# Create the datetime features before selecting them

# Extraer año, mes, día de la semana y día del mes

df_unificado['Ano'] = df_unificado['Fecha de venta'].dt.year

df_unificado['Mes'] = df_unificado['Fecha de venta'].dt.month

df_unificado['Dia_Semana'] = df_unificado['Fecha de
venta'].dt.dayofweek # Lunes=0, Domingo=6

df_unificado['Dia_Mes'] = df_unificado['Fecha de venta'].dt.day

# Crear una variable binaria para indicar si es fin de semana

df_unificado['Es_Fin_Semana'] =
df_unificado['Dia_Semana'].apply(lambda x: 1 if x >= 5 else 0)

# Seleccionar características para la clusterización

# Incluimos las variables numéricas y las categóricas que consideramos
relevantes para la demanda de clientes

features = [

    'Volumen de venta',

    'Precio de la venta',

    'Cantidad de productos',

    'Descuento (variable binaria)',

    'Crédito (variable binaria)',

    'Tipo de cliente',

    'Local de venta',

    'Año',

    'Mes',
```

```
    'Dia_Semana',
    'Dia_Mes',
    'Es_Fin_Semana'
]
df_cluster_features = df_unificado[features].copy()
# Identificar columnas categóricas para One-Hot Encoding
categorical_features = ['Tipo de cliente', 'Local de venta']
# Aplicar One-Hot Encoding a las variables categóricas seleccionadas
df_cluster_features = pd.get_dummies(df_cluster_features,
columns=categorical_features, drop_first=True)


print("Primeras filas del DataFrame con características para
clusterización (después de One-Hot Encoding):")
display(df_cluster_features.head())
```

Primeras filas del DataFrame con características para clusterización (después de One-Hot Encoding):

	Volumen de venta	Precio de la venta	Cantidad de productos	Descuento (variable binaria)	Crédito (variable binaria)	Año	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana	Tipo de cliente_Natural	Local de venta_PRINCIPAL
0	5.4	3.1555	1.0	0	1.0	2025	9	0	22	0	False	True
1	5.4	3.1555	1.0	0	1.0	2025	9	1	16	0	False	True
2	5.4	3.1555	3.5	0	1.0	2025	9	1	16	0	False	True
3	5.4	3.1555	1.0	0	1.0	2025	9	3	18	0	False	True
4	5.4	0.3478	3.5	0	1.0	2025	9	3	18	0	True	True

Interpretación

Las características de agrupación de clientes para este paso debían estar preparadas en esta etapa. A partir de la variable Fecha de Venta, encontramos nuevas variables temporales: Año, Mes, Día de la semana, Día del mes y Es fin de semana. Estas variables nos permiten interpretar el momento y si existen patrones de consumo en ciertos días o períodos.

La variable **`Día_Semana`** fue codificada en el rango de 0 a 6 (0 es lunes y 6 es domingo). También se creó la variable binaria **`Es_Fin_de_Semana`**, que toma el valor 1 cuando corresponde a sábado o domingo y 0 cuando corresponde a un día laborable. Esto también nos permite comparar cómo gastamos dinero durante la semana y los fines de semana.Luego seleccionamos las variables que son relevantes para analizar la demanda: volumen de ventas, precio de venta, número de productos, descuento, crédito, tipo de cliente, ubicación de ventas y variables temporales. También consideramos los aspectos económicos, comerciales y temporales de la compra.

Las variables categóricas Tipo de Cliente y Ubicación de Ventas fueron transformadas utilizando la técnica de Codificación **`One-Hot`**, porque el algoritmo K-means necesita trabajar con información numérica. Por lo tanto, se generaron variables como Tipo de **`Cliente_Natural`** y Ubicación de **`Ventas_PRINCIPAL`** en forma de Verdadero y Falso, respectivamente 1 y 0.

Interpretación de resultados

Los resultados demuestran que la base de datos fue preparada para el algoritmo K-means. Los datos categóricos que incluían tipo de cliente y ubicación de ventas fueron convertidos en datos

numéricos y la fecha permitió generar año, mes, día de la semana y fin de semana. Estas variables también difieren en características como volumen de ventas, precio y número de productos y, por lo tanto, muestran diferentes comportamientos de compra. Esta información ayudará al modelo a identificar grupos de clientes que son similares.

Ahora, escalaremos las características numéricas para asegurar que todas contribuyen equitativamente al algoritmo de clustering.

```
# Escalar las características numéricas

scaler = StandardScaler()

df_scaled = scaler.fit_transform(df_cluster_features)

# Convertir el array escalado de nuevo a un DataFrame para
mantener los nombres de las columnas

df_scaled = pd.DataFrame(df_scaled,
columns=df_cluster_features.columns)

print("Primeras filas del DataFrame con características
escaladas:")

display(df_scaled.head())
```

```
Primeras filas del DataFrame con características escaladas:
```

	Volumen de venta	Precio de la venta	Cantidad de productos	Descuento (variable binaria)	Crédito (variable binaria)	Ano	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana	Tipo de cliente_Natural	Local de venta_PRINCIPAL
0	2.120788	2.152647	-0.596691	-0.140325	2.157954	-1.126314	0.665422	-1.459906	0.737241	-0.826204	-2.807594	0.508274
1	2.120788	2.152647	-0.596691	-0.140325	2.157954	-1.126314	0.665422	-1.013044	0.060899	-0.826204	-2.807594	0.508274

2	2.120788	2.152647	2.164904	-0.140325	2.157954	-1.126314	0.665422	-1.013049	0.060899	-0.826204	-2.807594	0.508274
3	2.120788	2.152647	-0.596691	-0.140325	2.157954	-1.126314	0.665422	-0.119321	0.286347	-0.826204	-2.807594	0.508274
4	2.120788	-1.028774	2.164904	-0.140325	2.157954	-1.126314	0.665422	-0.119321	0.286347	-0.826204	0.356177	0.508274

Interpretación

Se aplicó StandardScaler para estandarizar las variables utilizadas en la agrupación. Esto transforma los datos para que tengan una escala similar, de modo que los valores más altos de las variables no influyan en K-means. Los valores positivos reflejan observaciones por encima del promedio, y los valores negativos representan aquellas por debajo del promedio.

Interpretación de resultados

Los resultados muestran que las variables fueron correctamente estandarizadas. Por ejemplo, valores como 2.12 en volumen de ventas y 2.15 en precio de ventas son registros por encima del promedio, mientras que -0.59 en el número de productos está por debajo del promedio. Con esta transformación, todas las variables están en una escala uniforme, y estamos listos para aplicar el algoritmo K-means.

Paso 12. Determinación del número óptimo de clusters (K).

Para K-Means, es crucial determinar el número de clusters (K) más adecuado. Utilizaremos dos métodos populares:

1. **Método del Codo (Elbow Method)**: Evalúa la suma de los cuadrados de las distancias (WCSS) dentro de los clusters para diferentes valores de K. El punto "codo" en el gráfico sugiere un K óptimo.
2. **Coeficiente de Silueta**: Mide qué tan similar es un objeto a su propio cluster en comparación con otros clusters. Un valor más alto indica una mejor separación de los clusters.

```
# Método del Codo para encontrar el número óptimo de clusters

wcss = []

# Se prueban con un rango de K, por ejemplo, de 1 a 10. Ajusta según sea necesario.

for i in range(1, 11):

    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=42)

    kmeans.fit(df_scaled)

    wcss.append(kmeans.inertia_)

# Graficar el Método del Codo

plt.figure(figsize=(10, 6))

plt.plot(range(1, 11), wcss, marker='o', linestyle='--')

plt.title('Método del Codo para K Óptimo')

plt.xlabel('Número de Clusters (K)')

plt.ylabel('WCSS')

plt.grid(True)

plt.show()
```

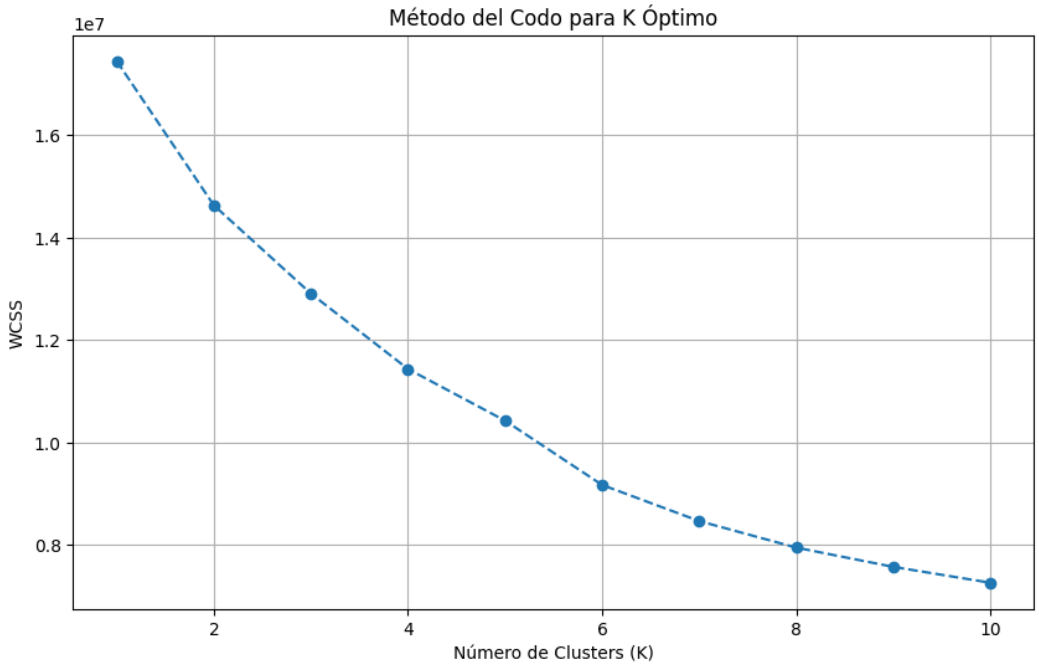
Interpretación

En esta etapa, nuestro objetivo es determinar el número de clústeres a utilizar en el modelo K-means. Para este propósito, el código calcula el WCSS para valores de K entre 1 y 10. El WCSS

indica cuán agrupadas están las observaciones en cada clúster; cuanto menor es el valor, más cohesivas son en su conjunto.

Interpretación de resultados

El código genera el gráfico del Método del Codo, para que podamos elegir el valor de K donde la disminución del WCSS comienza a ser menos pronunciada. En esta imagen, el gráfico de resultados aún no aparece, por lo que todavía no es posible establecer el número óptimo de clústeres solo con esta captura.



Interpretación.

El gráfico del Método del Codo muestra que el WCSS disminuye a medida que aumenta el número de clusters. La reducción es más fuerte en los valores iniciales de K y comienza a desacelerarse aproximadamente desde K = 6. El valor de K = 6 puede considerarse una buena opción porque a partir de ahí, más clusters llevarán a una mejora menor en el WCSS. Sin

embargo, dado que el "codo" no está completamente pronunciado, es útil confirmar este resultado con el coeficiente de Silhouette.

Paso 13. Aplicación de K-Means Clustering

Una vez que hemos analizado el Método del Codo y el Coeficiente de Silueta, elegiremos un número óptimo de clusters (K). Para este ejemplo, asumiremos K=3 basándonos en la tendencia de ambos gráficos (el "codo" en el WCSS y un valor razonable en la silueta antes de que decrezca significativamente).

Nota: Esta elección de K puede ajustarse si un análisis más detallado o específico del negocio sugiere otro valor.

```
# Aplicar K-Means con el número óptimo de clusters (K=3 como ejemplo)
# Ajusta n_clusters según el análisis del método del codo y silueta.
optimal_k = 3
kmeans_model = KMeans(n_clusters=optimal_k, init='k-means++', max_iter=300, n_init=10, random_state=42)
kmeans_model.fit(df_scaled)
# Asignar los clusters al DataFrame original
df_unificado['Cluster'] = kmeans_model.labels_
print(f"Se han creado {optimal_k} clusters.")
print("Conteo de clientes por cluster:")
display(df_unificado['Cluster'].value_counts())
```

```
print("Primeras filas del DataFrame con la columna de Cluster asignada:")
display(df_unificado.head())
```

Interpretación

En este punto, se utiliza el algoritmo K-Means con K = 3, por lo que los registros se agruparán en tres clústeres según la similitud de sus características estandarizadas. Luego, el número de clúster asignado a cada observación se incluye en la base de datos a través de la variable Clúster.

Interpretación de resultados

El modelo nos permitirá identificar tres grupos con comportamientos de compra similares y saber cuántos registros pertenecen a cada uno. Sin embargo, en esta imagen, aún no aparecen los resultados finales y el número de observaciones por clúster.

Importante: K = 6 fue una mejor elección en el gráfico del Método del Codo en el pasado. Como resultado, K = 3 solo debe mantenerse si el coeficiente de Silueta muestra que 3 clústeres proporcionan una mejor separación.

Se han creado 3 clusters.

Conteo de clientes por cluster:

	count
Cluster	
1	688694

0	557708
2	206878

dtype: int64

Primeras filas del DataFrame con la columna de Cluster asignada:

	Nombre del cliente	Local de venta	Cantón	Cantidad de productos	Tipo de productos	Fecha de venta	Tipo de cliente	Volumen de venta	Precio de la venta	Descuento (variable binaria)	Crédito (variable binaria)	Tipo de crédito	Año	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana	Cluster
0	PEREZ YUNGA ANA PRISCILLA	PRINCIPAL	Pasaje	1.0	CAFE	2025-09-22	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	0	22	0	2
1	PRADO PAUCA	PRINCIPAL	Buenavista	1.0	CAFE	2025-	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	1	16	0	2

	Y JOSE ARTURO					09-16												
2	PRADO PAUCA Y JOSE ARTURO	PRINCIPAL	Buenavista	3.5	CAFE	2025-09-16	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	1	16	0	2
3	ORDOÐEZ JARA ROSA	PRINCIPAL	Pasaje	1.0	CAFE	2025-09-18	Jurídico	5.4	3.1555	0	1.0	0.0	2025	9	3	18	0	2

	ANGELICA																			
4	ARMIJOS ZHIGUE ANGEL ARMANDO	PRINCIPAL	Pasaje	3.5	CAFE	2025-09-18	Natural	5.4	0.3478	0	1.0	0.0	2025	9	3	18	0	2		

Interpretación:

El algoritmo K-Means clasificó los registros en 3 clústeres según las similitudes de las variables analizadas. Cada observación recibió una etiqueta de clúster (0, 1 o 2), permitiendo separar distintos patrones de comportamiento de compra.

Interpretación de resultados:

De un total de 1.453.280 registros, el Clúster 1 concentra 688.694 (47,4%), el Clúster 0 557.708 (38,4%) y el Clúster 2 206.878 (14,2%). Esto indica que el Clúster 1 representa el grupo más numeroso y el Clúster 2 el menos frecuente.

Sin embargo, todavía no se puede definir qué caracteriza a cada clúster por ejemplo, alto valor, frecuentes o sensibles a descuentos hasta analizar los promedios o centroides de cada grupo.

Paso 14. Análisis de los clusters

Para entender mejor los segmentos de clientes creados por K-Means, analizaremos las características promedio de cada cluster. Esto nos permitirá perfilar cada grupo y observar qué define a cada uno.

```
# Analizar las características de cada cluster
# Agrupar el DataFrame unificado por la columna 'Cluster' y calcular la media de las características numéricas relevantes.
# Seleccionamos las mismas características que se usaron para el clustering, pero ahora del df_unificado original
# (o df_cluster_features antes de escalar, si se quiere ver valores originales)
```

```
cluster_analysis_features = [col for col in
df_cluster_features.columns if 'Tipo de cliente' not in col and
'Local de venta' not in col] # Excluimos las one-hot encoded para
simplificar la vista

# Incluir las columnas numéricas originales y la columna
'Cluster'

# Es importante asegurarse de que las columnas que se analizan
tengan sentido para la interpretación de los clusters.

# Usaremos el df_unificado para ver los valores originales, no
los escalados.

cluster_summary = df_unificado.groupby('Cluster')[['Volumen de
venta', 'Precio de la venta', 'Cantidad de productos', 'Descuento
(variable binaria)', 'Crédito (variable binaria)', 'Ano', 'Mes',
'Dia_Semana', 'Dia_Mes', 'Es_Fin_Semana']].mean()


print("Características promedio de cada cluster:")

display(cluster_summary)

# Para las características categóricas, podemos ver las modas o
distribuciones

print("Distribución de 'Tipo de cliente' por cluster:")

display(pd.crosstab(df_unificado['Cluster'], df_unificado['Tipo
de cliente'], normalize='index'))

print("Distribución de 'Local de venta' por cluster:")

display(pd.crosstab(df_unificado['Cluster'], df_unificado['Local
de venta'], normalize='index'))
```

Interpretación:

En esta etapa, se analizan las características de cada clúster aplicando K-Means. Los registros se agrupan según la variable Clúster y se calcula el promedio de variables como volumen de ventas, precio, número de productos, descuento, crédito y características temporales.También se examina la distribución de variables categóricas como Tipo de Cliente y Ubicación de Ventas para identificar qué categoría tiene mayor presencia dentro de cada grupo.

Interpretación de resultados:

Este proceso nos permitirá determinar qué características diferencian a los clústeres y asignarles una interpretación comercial. Por ejemplo, será posible identificar si un grupo tiene un mayor volumen de ventas, un mayor número de productos, un mayor uso de crédito o una mayor participación de ciertos tipos de clientes. En esta imagen, solo queda el código, por lo que aún no podemos describir las características de cada clúster hasta que tengamos la tabla de resultados.

Características promedio de cada cluster:

Cluster	Volumen de venta	Precio de la venta	Cantidad de productos	Descuento (variable binaria)	Crédito (variable binaria)	Ano	Mes	Dia_Semana	Dia_Mes	Es_Fin_Semana
0	1.773010	1.165879	1.453558	0.002500	0.058287	2025.000000	10.576852	3.285784	15.547320	0.425535
1	1.759618	1.174034	1.431994	0.000996	0.049073	2026.000000	3.064027	3.446536	15.505309	0.459769
2	3.833039	1.769876	2.133771	0.125601	0.921350	2025.599266	6.089347	2.618819	15.071989	0.172140

Distribución de 'Tipo de cliente' por cluster:

Tipo de cliente	Jurídico	Natural
Cluster		
0	0.042657	0.957343
1	0.036842	0.963158
2	0.553210	0.446790

Distribución de 'Local de venta' por cluster:

Local de venta	MATRIZ	PRINCIPAL
Cluster		
0	0.122229	0.877771
1	0.223495	0.776505
2	0.368700	0.631300

Interpretación de resultados

Los clústeres tienen perfiles claramente diferentes. Los clústeres 0 y 1 se comportan de manera similar (tanto en volumen de ventas como en cantidad de productos menos) pero principalmente basados en clientes individuales y compras en la ubicación principal.

Por otro lado, el clúster 2 tiene un mayor volumen de ventas (3.83), una gama más amplia de productos (2.13), un mayor uso de crédito (92.1%) y un mayor número de clientes corporativos (55.3%). Por lo tanto, este grupo podría considerarse como el de mayor intensidad y valor de compra.

Paso 15. Visualización de los Clusters Agrupados (usando PCA)

```
# Aplicar PCA para reducir a 2 componentes para visualización
pca = PCA(n_components=2)
df_pca = pca.fit_transform(df_scaled)
# Crear un DataFrame con los componentes PCA y las etiquetas de cluster
df_pca = pd.DataFrame(df_pca, columns=['PC1', 'PC2'])
df_pca['Cluster'] = df_unificado['Cluster']
# Visualizar los clusters
plt.figure(figsize=(10, 8))
sns.scatterplot(x='PC1', y='PC2', hue='Cluster', data=df_pca, palette='viridis', s=50, alpha=0.7)
plt.title('Clusters de Clientes (PCA 2D)')
plt.xlabel('Componente Principal 1')
plt.ylabel('Componente Principal 2')
plt.legend(title='Cluster')
plt.grid(True)
```

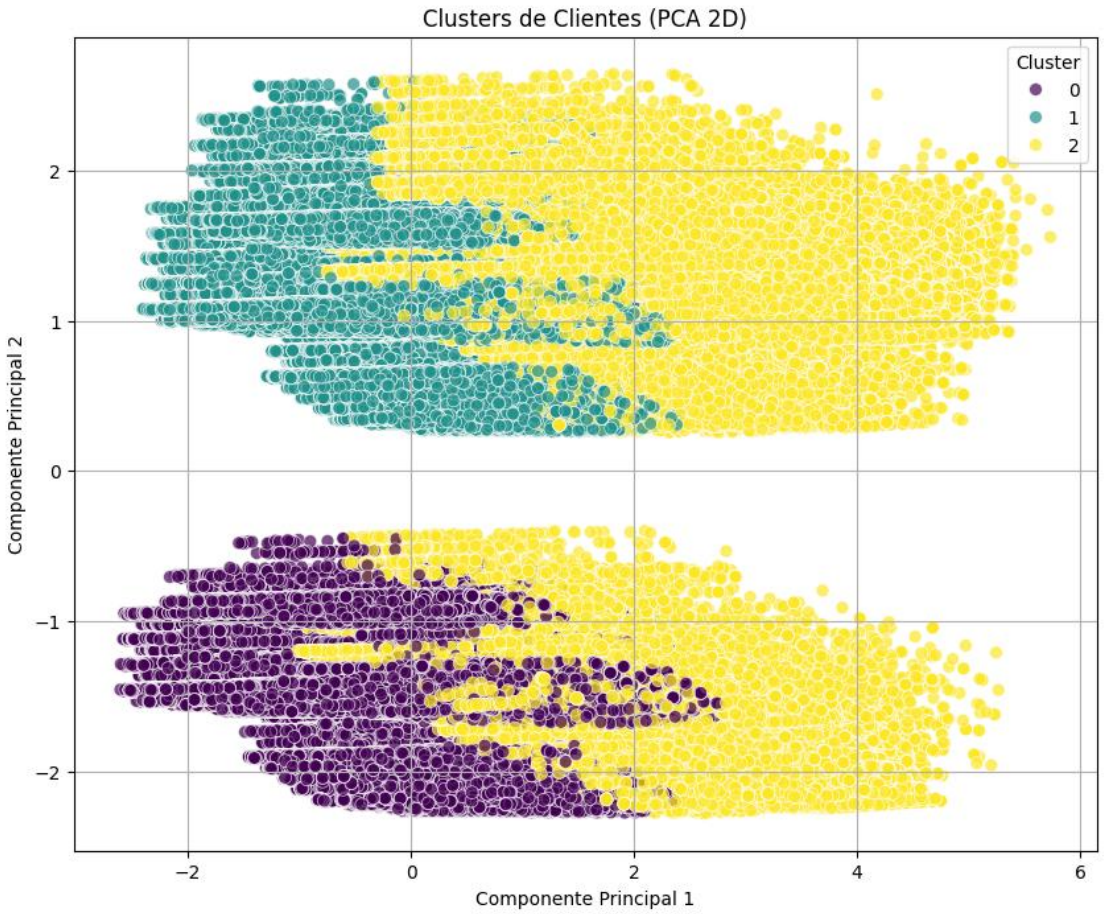
```
plt.show()
```

Interpretación

En esta etapa, se realiza un ACP (Análisis de Componentes Principales) para reducir las variables del modelo a dos componentes principales (CP1 y CP2) y visualizar los grupos utilizando K-Means.

Interpretación de resultados

El gráfico nos permitirá observar visualmente qué tan cerca o lejos están los tres grupos. En esta imagen, solo es visible el código, por lo que podemos interpretar la separación entre los grupos cuando se visualiza el gráfico de ACP.



Interpretación

El gráfico de PCA nos permite visualizar la distribución de los tres grupos en 2D. Podemos ver que los grupos 0 y 1 tienen una separación bastante clara, principalmente debido al Componente Principal 2.

Interpretación de resultados

El grupo 2 está más disperso y se superpone con los grupos 0 y 1. Esto indica que este grupo comparte algunas características con los otros, pero es de su propio tipo. En general, la visualización muestra que hay tres grupos distintos, pero que hay cierto grado de similitud entre algunas observaciones.

Paso 16. Regresión Lineal Múltiple por Cluster para Predecir el Volumen de Venta.

Ahora, aplicaremos un modelo de regresión lineal múltiple a cada cluster para predecir el 'Volumen de venta'. Esto nos permitirá identificar los factores más influyentes en la demanda para cada segmento de clientes. Primero, necesitamos preparar los datos para la regresión, utilizando las características que ya hemos codificado y limpiado.

Es importante recalcar que entrenaremos un modelo de regresión distinto para cada cluster, ya que los factores que influyen en la demanda podrían variar significativamente entre los diferentes segmentos de clientes.

```
# Añadir la columna de 'Cluster' al DataFrame de características para poder filtrar por cluster
df_reg_features = df_cluster_features.copy()
df_reg_features['Cluster'] = df_unificado['Cluster']
# Diccionario para almacenar los modelos y sus resultados por cluster (sklearn)
```

```
regression_models_sklearn = {}

regression_results_sklearn = {}

# Diccionario para almacenar los modelos de statsmodels y sus P-values

regression_models_sm = {}

regression_results_sm = {}

# Iterar sobre cada cluster para entrenar un modelo de regresión lineal

for cluster_id in sorted(df_reg_features['Cluster'].unique()):

    print(f"\n--- Entrenando modelo para el Cluster {cluster_id} ---")

    # Filtrar los datos para el cluster actual

    cluster_data = df_reg_features[df_reg_features['Cluster'] == cluster_id].copy()

    # Definir las características (X) y la variable objetivo (y)

    # 'Volumen de venta' es la variable objetivo.

    # Excluimos 'Cluster' y 'Volumen de venta' de las características.

    X = cluster_data.drop(columns=['Volumen de venta', 'Cluster'])

    y = cluster_data['Volumen de venta']

    # Convertir columnas booleanas a enteros para statsmodels, ya que puede tener problemas con True/False

    for col in X.select_dtypes(include=['bool']).columns:

        X[col] = X[col].astype(int)

    # Dividir los datos en conjuntos de entrenamiento y prueba
```

```
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

    # --- Modelo con scikit-learn (para R-squared y MSE) ---

    model_sklearn = LinearRegression()

    model_sklearn.fit(X_train, y_train)

    regression_models_sklearn[cluster_id] = model_sklearn

    y_pred_sklearn = model_sklearn.predict(X_test)

    r_squared_sklearn = model_sklearn.score(X_test, y_test)

    mse_sklearn = mean_squared_error(y_test, y_pred_sklearn)

    regression_results_sklearn[cluster_id] = {'R-squared':
r_squared_sklearn, 'MSE': mse_sklearn, 'y_test': y_test,
'y_pred_sklearn': y_pred_sklearn}

    print(f"R-squared (sklearn) para el Cluster {cluster_id}:
{r_squared_sklearn:.4f}")

    print(f"MSE (sklearn) para el Cluster {cluster_id}:
{mse_sklearn:.4f}")

    # Mostrar los coeficientes del modelo de sklearn

    print("Coeficientes del Modelo (sklearn):")

    coefficients_df_sklearn = pd.DataFrame({'Feature': X.columns,
'Coefficient': model_sklearn.coef_})

    display(coefficients_df_sklearn.sort_values(by='Coefficient',
ascending=False))

    # --- Modelo con statsmodels (para P-values) ---

    # Se añade una constante para el término de intercepción
(intercepto)

    X_train_sm = sm.add_constant(X_train)
```

```
    model_sm = sm.OLS(y_train, X_train_sm)
    results_sm = model_sm.fit()
    regression_models_sm[cluster_id] = results_sm
    print("\nP-values del Modelo (statsmodels):")
    p_values_df = pd.DataFrame(results_sm.pvalues, columns=['P-
value'])
    display(p_values_df)
print("\n--- Resumen de los resultados de regresión por Cluster
(sklearn) ---")
for cluster_id, metrics in regression_results_sklearn.items():
    print(f"Cluster {cluster_id}: R-squared={metrics['R-
squared']:.4f}, MSE={metrics['MSE']:.4f}")
```

Interpretación

En esta etapa aplicamos una regresión lineal de cada clúster con el volumen de ventas como la variable dependiente, y el resto de las variables como factores explicativos.

Interpretación de resultados

El modelo calcula para cada clúster el R^2, el MSE, los coeficientes y los valores p para determinar qué variables tienen una mayor relación e influencia en el volumen de ventas. Los resultados pueden interpretarse cuando se visualizan los valores obtenidos para cada clúster.

Entrenando modelo para el Cluster 0

R-squared (sklearn) para el Cluster 0: 0.8323

MSE (sklearn) para el Cluster 0: 0.2964

Coeficientes del Modelo (sklearn):

	Feature	Coefficient
0	Precio de la venta	1.469797e+00
1	Cantidad de productos	1.006372e+00
9	Tipo de cliente_Natural	1.098972e-01
5	Mes	2.302797e-03
6	Dia_Semana	1.720605e-03
7	Dia_Mes	2.907764e-04
4	Ano	8.326673e-17
8	Es_Fin_Semana	-9.103581e-04
10	Local de venta_PRINCIPAL	-1.490682e-02
2	Descuento (variable binaria)	-5.559541e-02
3	Crédito (variable binaria)	-2.043862e-01

P-values del Modelo (statsmodels):

	P-value

Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00
Descuento (variable binaria)	6.327822e-04
Crédito (variable binaria)	0.000000e+00
Ano	0.000000e+00
Mes	2.549430e-03
Dia_Semana	1.522756e-02
Dia_Mes	1.456000e-03
Es_Fin_Semana	7.811402e-01
Tipo de cliente_Natural	9.405110e-164
Local de venta_PRINCIPAL	7.195934e-09

Entrenando modelo para el Cluster 1

R-squared (sklearn) para el Cluster 1: 0.8385

MSE (sklearn) para el Cluster 1: 0.2863

Coeficientes del Modelo (sklearn):

	Feature	Coefficient
0	Precio de la venta	1.469992e+00
1	Cantidad de productos	9.854791e-01
9	Tipo de cliente_Natural	1.217967e-01
2	Descuento (variable binaria)	1.147856e-02
6	Dia_Semana	4.995399e-03
5	Mes	3.096130e-04
7	Dia_Mes	7.340043e-05
4	Ano	5.551115e-17
10	Local de venta_PRINCIPAL	-1.884389e-02
8	Es_Fin_Semana	-2.714590e-02
3	Crédito (variable binaria)	-1.453875e-01

P-values del Modelo (statsmodels):

	P-value
Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00
Descuento (variable binaria)	6.175868e-01
Crédito (variable binaria)	0.000000e+00
Ano	0.000000e+00
Mes	5.465144e-01
Dia_Semana	5.517227e-15
Dia_Mes	3.722848e-01
Es_Fin_Semana	1.025814e-20
Tipo de cliente_Natural	1.326444e-217
Local de venta_PRINCIPAL	8.694959e-27

Entrenando modelo para el Cluster 2

R-squared (sklearn) para el Cluster 2: 0.6803

MSE (sklearn) para el Cluster 2: 0.9948

Coeficientes del Modelo (sklearn):

	Feature	Coefficient
0	Precio de la venta	1.369662
1	Cantidad de productos	1.020171
9	Tipo de cliente_Natural	0.203754
6	Dia_Semana	0.031979
7	Dia_Mes	-0.000534
5	Mes	-0.016812
10	Local de venta_PRINCIPAL	-0.075871
8	Es_Fin_Semana	-0.105869
3	Crédito (variable binaria)	-0.156539
4	Ano	-0.305661
2	Descuento (variable binaria)	-0.693491

P-values del Modelo (statsmodels):

	P-value
const	1.123850e-87
Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00
Descuento (variable binaria)	0.000000e+00
Crédito (variable binaria)	9.629616e-63
Ano	6.839879e-88
Mes	3.219403e-19
Dia_Semana	7.479198e-67
Dia_Mes	5.480620e-02
Es_Fin_Semana	5.281757e-29
Tipo de cliente_Natural	6.962623e-291
Local de venta_PRINCIPAL	1.134499e-34

Resumen de los resultados de regresión por Cluster (sklearn)

Cluster 0: R-squared=0.8323, MSE=0.2964

Cluster 1: R-squared=0.8385, MSE=0.2863

Cluster 2: R-squared=0.6803, MSE=0.9948

Paso 17. P-values del Modelo (statsmodels) para Cluster 1 y Cluster 2

```
print("\nP-values del Modelo (statsmodels) para el Cluster 0:")
display(pd.DataFrame(regression_models_sm[0].pvalues,
columns=['P-value']))
print("\nP-values del Modelo (statsmodels) para el Cluster 1:")
display(pd.DataFrame(regression_models_sm[1].pvalues,
columns=['P-value']))
print("\nP-values del Modelo (statsmodels) para el Cluster 2:")
display(pd.DataFrame(regression_models_sm[2].pvalues,
columns=['P-value']))
```

Interpretación:

En este punto, se muestran los valores p de las variables para cada clúster utilizando statsmodels para determinar qué variables son estadísticamente significativas en la explicación del volumen de ventas.

Interpretación de resultados

Cuando el valor p es menor que 0.05, la variable se considera significativa en el modelo y si es mayor que 0.05, el efecto no es estadísticamente significativo. En esta imagen, aún no tenemos los valores, por lo que todavía no es posible encontrar qué variables son significativas en cada clúster.

P-values del Modelo (statsmodels) para el Cluster 0:

	P-value
Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00
Descuento (variable binaria)	6.327822e-04
Crédito (variable binaria)	0.000000e+00
Ano	0.000000e+00
Mes	2.549430e-03
Dia_Semana	1.522756e-02
Dia_Mes	1.456000e-03
Es_Fin_Semana	7.811402e-01
Tipo de cliente_Natural	9.405110e-164
Local de venta_PRINCIPAL	7.195934e-09

P-values del Modelo (statsmodels) para el Cluster 1:

	P-value
Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00

Descuento (variable binaria)	6.175868e-01
Crédito (variable binaria)	0.000000e+00
Ano	0.000000e+00
Mes	5.465144e-01
Dia_Semana	5.517227e-15
Dia_Mes	3.722848e-01
Es_Fin_Semana	1.025814e-20
Tipo de cliente_Natural	1.326444e-217
Local de venta_PRINCIPAL	8.694959e-27

P-values del Modelo (statsmodels) para el Cluster 2:

	P-value
const	1.123850e-87
Precio de la venta	0.000000e+00
Cantidad de productos	0.000000e+00
Descuento (variable binaria)	0.000000e+00
Crédito (variable binaria)	9.629616e-63
Ano	6.839879e-88

Mes	3.219403e-19
Dia_Semana	7.479198e-67
Dia_Mes	5.480620e-02
Es_Fin_Semana	5.281757e-29
Tipo de cliente_Natural	6.962623e-291
Local de venta_PRINCIPAL	1.134499e-34

Interpretación de resultados

Los valores p menores de 0.05 indican que la variable tiene una relación estadísticamente significativa con el volumen de ventas.

Cluster 0: el precio, la cantidad de productos, el descuento, el crédito, el año, el mes, el día de la semana, el día del mes, el tipo de cliente y la ubicación de ventas son significativos. El fin de semana no es significativo.

Cluster 1: el precio, la cantidad de productos, el crédito, el año, el día de la semana, el fin de semana, el tipo de cliente y la ubicación de ventas son significativos. El descuento, el mes y el día del mes no son significativos.

Cluster 2: en la parte visible, el precio de venta y la cantidad de productos son significativos, con valores p prácticamente iguales a 0.

Esto demuestra que los factores que explican el volumen de ventas varían según el perfil de cada cluster.

Paso 18. Visualización: Predicciones vs. Valores Reales por Cluster

```
import matplotlib.pyplot as plt
plt.figure(figsize=(18, 5))
for cluster_id in sorted(regression_results_sklearn.keys()):
    results = regression_results_sklearn[cluster_id]
    y_test = results['y_test']
    y_pred = results['y_pred_sklearn']
    r_squared = results['R-squared']
    mse = results['MSE']
    plt.subplot(1, len(regression_results_sklearn), cluster_id + 1)
    plt.scatter(y_test, y_pred, alpha=0.3, s=10)
    plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
    plt.xlabel('Volumen de Venta Real')
    plt.ylabel('Volumen de Venta Predicho')
    plt.title(f'Cluster {cluster_id}\n(R-squared: {r_squared:.2f}, MSE: {mse:.2f})')
    plt.grid(True)
plt.tight_layout()
plt.show()
```

Interpretación

Esta etapa compara el volumen de ventas real con el volumen de ventas predicho por el modelo de regresión para cada grupo.

Interpretación de resultados

Cuanto más cerca estén los puntos de la línea diagonal, mejor será la capacidad predictiva del modelo. Además, R^2 indica qué tan bien el modelo explica las variaciones en el volumen de ventas, y el MSE muestra el nivel de error en las predicciones. Los gráficos finales aún no aparecen en esta imagen.

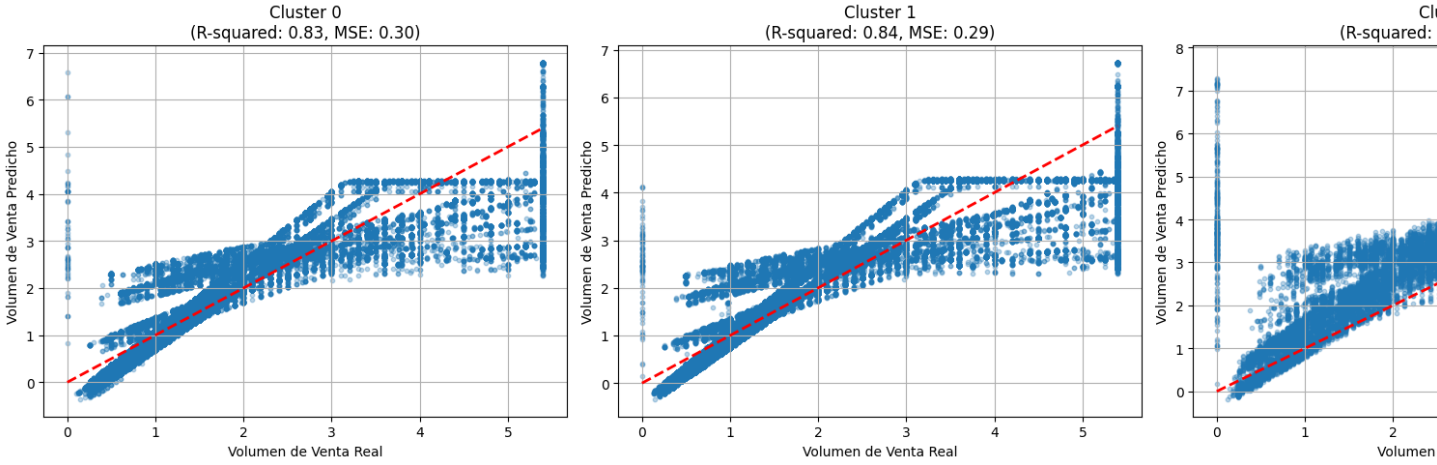


Imagen 1 Clúster 0: El modelo tiene un $R^2 = 0.83$, lo que representa aproximadamente el 83% de la variación en el volumen de ventas. El MSE de 0.30 significa que el error es relativamente bajo y muestra una buena capacidad de predicción.

Imagen 2 Clúster 1: Muestra el mejor rendimiento de los tres grupos con un $R^2 = 0.84$ y un MSE = 0.29. Esto significa que el modelo explica aproximadamente el 84% de la variación en el volumen de ventas con una buena predicción y, por lo tanto, realiza predicciones que están bastante cerca de los valores reales.

Imagen 3 Clúster 2: El modelo tiene un R^2 de 0.68, por lo que explica aproximadamente el 68% de la variación en el volumen de ventas. Su MSE de 0.99 es más alto que el de los otros clústeres, lo que indica que es menos preciso y hay una mayor dispersión entre

los valores reales y los predichos.

El clúster 2 presenta un ajuste menor, con $R^2 = 0,68$ y MSE = 0,99, por lo que sus predicciones son menos precisas. En general, el clúster 1 es el que muestra el mejor desempeño del modelo.

CONCLUSIÓN DEL ANÁLISIS DE RESULTADOS

Los resultados que obtuvimos nos permitieron identificar tres segmentos de clientes con diferentes comportamientos de compra. Los clústeres 0 y 1 son algo similares, principalmente en términos de volumen de ventas y número de productos comprados, pero el clúster 2 tiene un mayor volumen de ventas, un mayor número de ventas de productos y uso de crédito. La visualización de PCA confirmó que los grupos son diferentes en su distribución, pero hay un cierto grado de similitud entre algunas observaciones. Esto muestra que los clientes no pueden ser considerados como un solo grupo porque hay diferentes patrones en la demanda.Por otro lado, los modelos de regresión mostraron que los factores asociados con el volumen de ventas varían según cada clúster. El precio, el número de productos, el crédito, el tipo de cliente, el punto de venta y los factores temporales tienen diferente importancia dependiendo del segmento estudiado. Para concluir, el uso de K-Means, PCA y regresión permitió obtener una mejor comprensión de la demanda e identificar diversos perfiles de comportamiento comercial. Estos datos pueden ser utilizados para influir en las decisiones de fidelización de clientes en cuanto a lealtad, promociones, crédito, inventario y planificación de ventas basadas en las características específicas de cada segmento.

DISCUSIÓN

Los resultados metodológicos de la presente investigación pueden discutirse a partir de estudios previos que aplican tecnicas analiticas de datos para comprender el comportamiento de los clientes en contextos comerciales. En el primer articulo de Cam Gensollen (2022) tiene similitud puesto que analiza un proyecto de big data aplicado a una cadena de supermercados de Europa, en el cual se propone la segmentación de clientes medianste el algoritmo k-means y un sistema de recomendación. Este enfoque guarda relación directa con esta investigación, ya que ambos estudios parten de la necesidad de transformar los registros comerciales en información util para diferenciar perfiles de consumo. En esta situación referente a la tesis, la base de datos del supermercado se compone de variables tales como volumen de venta, precio de venta, cantidad de productos, descuento, crédito, tipo de cliente, local de venta y variables temporales que permiten segmentar a los clientes en relación con su verdadero comportamiento en cuanto a la demanda. De esta manera el análisis no se limita solamente a las ventas sino que esta orientado hacia el segmentar grupos de similitud comerciales.

Una conexión importante entre Cam Gensollen (2022) y este estudio es el enfoque en el analisis exploratorio de datos y el preprocesamiento de datos. El autor enfatiza la necesidad de una correcta definición del problema y una correcta organización de la infraestructura de datos antes de aplicar modelos de segmentación. Esta suposición corresponde al proceso de escribir la tesis, el cual incluye una revision preliminar de la estructura de la base de datos, la comprobación de valores nulos, la búsqueda de valores únicos, la conversión de la fecha de venta, la suplantación de valores faltante, y el manejo de los valores atípicos usando la Winsorización con el rango intercuartílico. Por lo tanto, todas las acciones no son triviales sino que deben realizarse para evitar la influencia

en el agrupamiento y la regresion de columnas vacías, fechas mal entendidas y valores extremos. Por lo tanto se puede concluir que la validez de los resultados no solo dependerá de los algoritmo K-means sino tambien de la calidad del preprocesamiento de los datos.

En cuanto a la segmentación Cam Gensollen (2022) aplica el algoritmo K-means para segmentar a los clientes en el contexto de un supermercado. Esto justifica la elección del k-means en la tesis ya que el objetivo aqui no es clasificar a los clientes en categorías preestablecidas sino más bien derivar patrones de comportamiento a partir de las variables. En cuento a la investigación de la tesis el k-means permite agrupar las transacciones comerciales que comparten características comunes en cuanto al volumen, los precios, la cantidad, los descuentos, el credito, el tipo de cliente, las sucursales y las variables relacionadas con el tiempo. La razón detras de esta metodologia es que la demanda en los supermercados no es uniforme; hay registros de transacciones con un volumen promedio bajo, otras con volúmenes altos y dependiendo de si involucran el uso de crédito, el tipo de cliente o la sucursal de venta. Por consiguiente el resultado de esta segmentación es una mejor compresión de la demanda que la de agrupar las ventas.

El segundo artículo titulado inteligencia de negocios aplicada a la segmentación de clientes: Modelo RFM y Análisis de Clúster, también fortalece la discusión porque plantea que los datos de transacciones de facturación contienen información básica para describir el comportamiento de consumidores, especialmente mediante recencia, frecuencia y magnitud monetaria. Aunque la presente tesis no proporciona un modelo RFM completo debido a la imposibilidad de identificar a cada individuo dentro de la base de datos, si refleja la idea principal del uso de los datos transaccionales para construir nuestro conocimiento sobre la demanda. En este caso los factores como la cantidad de producto , el volumen de ventas, el precio, el crédito, el descuento, el tipo de cliente y el lugar de venta cumplen la misma función que los indicadores comerciales,

ayudandonos a diferenciar entre varios grupos y a tomar decisiones con respecto a los programas de marketing y de lealtad.

La diferencia clave entre el estudio RFM y esta investigación de tesis radica en la unidad de analisis. El modelo de RFM supone la identificación del cliente en particular y la medición de cuanto tiempo ha pasado desde que realizo una compra, con que frecuencia compra y cuanto dinero ha gastado. Por el contrario la investigaciónde tesis funciona en el marco donde el analisi se basa en los datos proporcionados por los registros de negocios y las variables de venta. Por lo tanto la diferencia no significa que la metodologia sea inválida sino que el alcance de la metodologia es limitado. Mientras que el modelo RFM proporciona una forma de caracterizar a cada cliente individualmente la metodolofia propuesta en esta tesis ayuda a encontrar patrones de la demanda, el tipo de cliente, productos, precios, cantidades, credito y descuentos.

Otro aspecto que se debe mencionar es la utilidad de los grupos desde la perspectiva empresarial. En los estudios analizados el objetivo de la segmentación es potenciar el proceso de toma de decisiones, la personalización de la estrategia y la mejora de la relación con los clientes. Este objetivo coincide con el de la tesis ya que el objetivo es interpretar los segmentos identificados y revelar la conexión entre ellos y la demanda y orientar las estrategias comerciales de acuerdo a ello. En el contexto del supermercado analizado el uso de los grupos puede ayudar a segmentar a los clientes con un volumen promedio menor, los clientes con un alto volumen de ventas y los clientes con un alto volumen de uso de credito. Esta información se puede utilizar para tomar decisiones acerca de las promociones, los incentivos, el suministro, la lealtad y la diferenciación de acuerdo a la sucursal.

La inclusión del análisis de regresión lineal múltiple en la tesis aumenta el rango de metodologías empleadas en el estudio porque además de agrupar tambien busca identificar que variables afectan

el volumen de las ventas en cada grupo. En otras palabras la tesis intenta complementar la metodología de agrupamiento con la de predicción y explicación. Usando el agrupamiento de K-means primero se realizar un análisis de regresión en cada grupo tomando el volumen de ventas como la variable dependiente mientras que las variables independientes incluyen el precio de venta, la cantidad de productos, el descuento, el crédito, el mes, el dia de la semana, el día del mes, el fin de semana, el tipo de cliente y el lugar de venta. Es a través de esta metodología que podemos notar que las variables que influyen la demanda pueden comportarse de manera diferente en diferentes segmentos de clientes. Por ejemplo, el precio puede relacionarse con el volumen en un segmento con bajo consumo de manera diferente a otro segmento con alto volumen o con alto uso de crédito.

Tambien es importante hablar de las limitaciones. Otros estudios similares muestran la importancia de la necesidad de bases de datos integrales, bien estructuradas y confiables. En la tesis durante el Analisis Exploratorio de Datos se tomaron nota columnas como ID, número de factura y el tipo de descuento así como la falta de datos en cantón, tipo de cliente, crédito y la cantidad de productos. Aunque se superaron las limitaciones a través de la eliminación de columnas vacías e imputación de datos estos siguen siendo factores a tener en cuenta al interpretar los resultados. Además, la falta de alguna factura o identificación de cliente totalmente utilizable impide la posibilidad de calcular métricas tradicionales como la frecuencia individual, la recencia y la acumulación monetaria por cliente. Por lo tanto es importante interpretar los resultados como perfiles de la demanda basados en datos de negocios y no una segmentación definitiva de cada consumidor individualmente.

Finalmente, de puede afirmar que los estudios revisados respaldan la pertinencia metodológica de la tesis. Según Cam Gensollen (2022) la segmentación K-means puede ser aplicada en el caso de

un supermercado y se requiere el preprocesamiento de los datos para obtener resultados creíbles. En cuanto al estudio de la inteligencia de Negocios el RFM y el analisis de agrupamiento demuestran que los datos transaccionales pueden ser analizados para el perfilado de los clientes y la toma de decisiones. Es posible comparar estos trabajos con la tesis y concluir que el análisis de la demanda a través del agrupamiento ayuda a comprender la diversidad de los registros comerciales. Por lo tanto, los resultados obtenidos tienen tanto un valor estadístico como práctico ya que ayudan a comprender que segmentos tienen diferentes niveles de contribución, comportamiento y lealtad.

RECOMENDACIONES.

1. **Mantener la segmentación de clientes con clústeres.**

El supermercado puede mantener y actualizar la segmentación de clientes mediante K-Means y actualizarla periódicamente a medida que se añaden nuevas transacciones a la base de datos. La segmentación no debe tratarse como un resultado estático, ya que varía entre diferentes compradores en términos de la frecuencia con la que compran, el consumo y cuánto gastan, el uso de crédito, el descuento o la tasa de descuento y la frecuencia de compra. La empresa puede utilizar el mismo procedimiento que en el guion de investigación para reprocesar la información y ver si la segmentación de clientes sigue siendo la misma y si los clientes aún permanecen dentro del mismo clúster o si hay características que los hacen más o menos similares a otros clientes de otro grupo. De esta manera, la clasificación de clientes se capturará para que coincida con el comportamiento real de los clientes en los datos.

Porque los consumidores tienen diferentes niveles de variables para el volumen de ventas, número de productos, crédito, descuentos, tipo de cliente y punto de venta. Mantenemos los clústeres actualizados y tomamos decisiones a la luz de la información más reciente y no tratamos a todos los clientes por igual.

2. **Usar los clústeres para la mejora del inventario y la planificación de locales.**

Se debe utilizar los resultados de la segmentación de clientes como una herramienta estratégica para mejorar la gestión del inventario e incluso la planificación operativa para cada ubicación de supermercado. Los clústeres han llevado a una mejor comprensión de los clientes en cada ubicación; qué productos están en demanda y cómo cambian los hábitos de compra de los clientes en relación con los perfiles de los consumidores. A partir de esta información, cada ubicación

puede adaptar su suministro de productos a lo que es la mayoría de sus clientes. Si un clúster realiza más compras al por mayor, por ejemplo, la ubicación donde este grupo tiene muchos clientes siempre tendrá más productos de alta rotación. Si un clúster realiza menos compras al por mayor pero compra artículos con frecuencia, la estrategia de inventario será mantener un stock constante de productos básicos. Los clústeres también pueden usarse para informar el plan comercial de cada establecimiento, como promociones y la distribución de productos en el estante y lo que se está haciendo, así como los recursos operativos. También ayuda a identificar las diferencias basadas en la ubicación, lo que ayudará a entender por qué algunos lugares de venta funcionan mejor que otros.

Dado que variables como el número de productos comprados, el volumen de ventas, el tipo de cliente y la ubicación de ventas nos permiten identificar los grupos de individuos con comportamientos claramente diferenciados. Y esta segmentación no solo describe a los clientes, sino también los patrones de consumo de cada establecimiento. Y por lo tanto, al utilizar esta información podemos gestionar el inventario (escasez o exceso de inventario) y optimizar el aspecto comercial de cada ubicación.

3. **Utilizar los modelos de regresión como apoyo para el pronóstico de ventas.**

Sugerimos continuar utilizando los modelos de regresión desarrollados en el trabajo como herramienta de apoyo para el análisis y la previsión del volumen de ventas y mantener un modelo independiente para cada clúster identificado. Además, estos modelos deben actualizarse regularmente a medida que se añadan nuevos datos para mantenerlos precisos y relevantes.

Utilizar modelos separados por clúster permite analizar el comportamiento de la demanda más profundamente, ya que cada grupo de clientes es diferente. Por ejemplo, algunos clústeres se ven

más afectados por el precio y otros por la cantidad de productos o el crédito. Esta diferencia permite crear modelos más precisamente adaptados a la realidad de cada segmento.

Es beneficioso utilizar los resultados de los modelos como herramienta complementaria para la toma de decisiones en la planificación de ventas, la proyección de ingresos y la gestión de inventarios. Las predicciones obtenidas pueden compararse con datos reales para evaluar la precisión del modelo y ajustarlo cuando sea necesario.

Ya que muestran los resultados de la investigación, los factores que afectan el volumen de ventas no son los mismos para todos los clientes, sino que dependen del grupo al que pertenecen. Por lo tanto, dado que el modelo de regresión es diferente para cada segmento, los scripts desarrollados para cada segmento pueden capturar mejor estas diferencias. El mismo método permite un mejor proceso de pronóstico que permite realizar mejores previsiones, una mejor comprensión de la dinámica de la demanda y la toma de decisiones basada en datos.

CONCLUSIÓN

La demanda de los clientes de las dos sucursales ahora puede explicarse como una función de los diversos factores comerciales y de comportamiento. Los más relevantes son el precio de venta, la cantidad de productos comprados, el uso de crédito, el tipo de cliente y la ubicación donde se realiza la compra. Además, encontramos que los descuentos y ciertos factores temporales no afectan a los mismos clientes, sino que tienen un valor distinto dependiendo del segmento al que pertenece cada uno de ellos. Por lo tanto, está claro que la demanda del supermercado no puede evaluarse sólo a partir de la cantidad de ventas, sino que también es necesario considerar las características de compra de los clientes y cómo se comportan. Al utilizar la técnica de agrupamiento en K-Means, podemos separar a los clientes en tres grupos con diferentes

características comerciales y así observar los diferentes patrones de consumo entre los clientes. Los clústeres 0 y 1 tienen algunos comportamientos similares, consistentes principalmente en clientes naturales con menor volumen de ventas y cantidad de productos comprados. Por otro lado, el clúster 2 tiene un comportamiento comercial más intenso con un mayor volumen de ventas, más productos comprados, mayor uso de crédito y un alto número de clientes legales. Por lo tanto, este último grupo es el que tiene el mayor valor comercial en el análisis. Además, el PCA nos permitió visualizar las diferencias entre los grupos y mostró que, aunque hay algunas similitudes en algunas observaciones, los clientes son lo suficientemente diferentes como para que cada grupo pueda ser segmentado. Finalmente, los segmentos obtenidos nos permitieron conectar los perfiles de los clientes y el comportamiento de la demanda y mostrar que cada grupo requiere diferentes estrategias comerciales. Los clientes que pertenecen al segmento de mayor valor pueden ser gestionados a través de la lealtad, beneficios de volumen de compra, facilidades de crédito y la relación comercial. Por otro lado, los segmentos con menor intensidad de compra pueden ser estimulados mediante promociones específicas, incentivos, estrategias para aumentar la frecuencia de compra y acciones para incrementar la cantidad de productos comprados. Como resultado, el agrupamiento nos proporciona información precisa para adaptar la promoción, el crédito, el inventario, la lealtad y la planificación de ventas a las características reales de cada segmento con el fin de obtener una mejor comprensión de la demanda de las dos sucursales.

BIBLIOGRAFÍA

admin. (2022, febrero 14). Segmentación tradicional vs segmentación Predictiva. *Marketing Predictivo*. https://marketingpredictivo.io/segmentacion-tradicional-vs-segmentacion-predictiva/

Anderson, R. (2024). *Customer Behavior Analysis: A Complete Guide - Qualtrics*. https://www.qualtrics.com/articles/customer-experience/customer-behavior-analysis/

Arellano, Romeo Vite López, & Pedro Cantú Elizondo. (2010). *Marketing. Enfoque America Latina*.

Bell, D. R., Ho, T.-H., & Tang, C. S. (1998). Determining Where to Shop: Fixed and Variable Costs of Shopping. *Journal of Marketing Research*, *35*(3), 352–369. https://doi.org/10.1177/002224379803500306

Chen, C. B., Agrawal, D., & Kumara, S. (2015). Retail analytics: IIE Annual Conference and Expo 2015. *IIE Annual Conference and Expo 2015, IIE Annual Conference and Expo 2015*, 1034–1042.

Chen, C., Liu, Z., Zhou, J., Li, X., Qi, Y., Jiao, Y., & Zhong, X. (2020). *How Much Can A Retailer Sell? Sales Forecasting on Tmall* (arXiv:2002.11940). arXiv. https://doi.org/10.48550/arXiv.2002.11940

Dois Z. (2025, enero 24). El impacto de la sostenibilidad en las preferencias de los consumidores modernos. *Dois Z Publicidade Agência de Publicidade e Propaganda Marketing digital Goiânia*. https://doisz.com/es/blog/preferencias-dos-consumidores-modernos/

Durán Barros & Nieto. (2025). *SEGMENTACIÓN DE CLIENTES EN EL SECTOR RETAIL MEDIANTE CLUSTERING NO SUPERVISADO: UN ANÁLISIS DE COMPORTAMIENTO, DEMOGRAFÍA Y RESPUESTA PROMOCIONAL*.

Ed Lorenzini. (2023). *From basic demographics to advanced AI/ML, consumer segmentation has evolved, aided by AR, VR, & IoT. Discover what these techniques can offer the industry*. https://greenbook.org/insights/researc-technology-restech/the-evolution-of-b2c-consumer-segmentation-in-the-digital-age

Fischer & Espejo. (2020). *MERCADOTECNIA*.

France, S. L., & Ghose, S. (2019). Marketing Analytics: Methods, Practice, Implementation, and Links to Other Fields. *Expert Systems with Applications*, *119*, 456–475. https://doi.org/10.1016/j.eswa.2018.11.002

Greenlaw, S. A., Shapiro, D., MacDonald, D., Greenlaw, S. A., Shapiro, D., & MacDonald, D. (2022, diciembre 14). *3.1 Demand, Supply, and Equilibrium in Markets for Goods and Services—Principles of Economics 3e | OpenStax*. OpenStax. https://openstax.org/books/principles-economics-3e/pages/3-1-demand-supply-and-equilibrium-in-markets-for-goods-and-services

Gujarati & Porter. (2010). *Econometría*. https://gc.scalahed.com/recursos/files/r161r/w25589w/(2010)Econometria.pdf

Hair, J. F., Hult, G. T. M., Sarstedt, M., & Ringle, C. M. (2022). *Book Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R*.

https://www.researchgate.net/publication/362062761_Book_Partial_Least_Squares_Structural_Equation_Modeling_PLS-SEM_Using_R

Han, J., Kamber, M., & Pei, J. (2012, enero 1). *Mining Pattern—An overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/computer-science/mining-pattern?utm_source=

Hernández-Sampieri, D. R. (2018). *METODOLOGÍA DE LA INVESTIGACIÓN: LAS RUTAS CUANTITATIVA, CUALITATIVA Y MIXTA*.

Hoyer, W., Pieters, R., & MacInnis, D. (2016). *Consumer Behavior*.

IBM. (2025). *¿Qué es el Análisis Exploratorio de Datos? | IBM*. https://www.ibm.com/think/topics/exploratory-data-analysis

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

James, G. (2021). *An Introduction to Statistical Learning: With Applications in R | Springer Nature Link*. https://link.springer.com/book/10.1007/978-1-0716-1418-1

Kahle, L. R., & Malhotra, N. K. (1994). Marketing Research: An Applied Orientation. *Journal of Marketing Research*, *31*(1), 137. https://doi.org/10.2307/3151953

Kotler & Keller. (2016). *Dirección en Marketing—Kotler y Keller 15va edición*. https://www.academia.edu/37145555/Direcci%C3%B3n_en_Marketing_Kotler_y_Keller_15va_edici%C3%B3n

Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management*. Springer. https://doi.org/10.1007/978-3-662-55381-7

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2, Part 2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Oracle. (2026). *Oracle® Fusion Cloud EPM Trabajo con Planning*. https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_IQR.htm#PFUSU-GUID-CF37CAEA-730B-4346-801E-64612719FF6B

PREDIK. (2025). *Big Data aplicado a Supermercados – Caso de estudio—PREDIK Data-Driven ES*. https://predikdata.com/es/big-data-aplicado-a-supermercados-caso-de-estudio/

Reifman, A., & Garrett, K. (2022). Winsorize. En *The SAGE Encyclopedia of Research Design* (2nd ed. (2ª edición), Vol. 4, pp. 1821–1822). SAGE Publications, Inc. https://doi.org/10.4135/9781071812082.n683

Salazar & Castillo. (2018). *Principios básicos de estadística*. https://gc.scalahed.com/recursos/files/r161r/w24899w/Trabajo%20Final/Fundamentos_Basicos_de_Estadistica.pdf

Tan, A. L., Jamaludin, A., & Hung, D. (2021). In pursuit of learning in an informal space: A case study in the Singapore context. *International Journal of Technology and Design Education*, *31*. https://doi.org/10.1007/s10798-019-09553-1

Tardivon, A. (2022, mayo 19). ¿Cómo los supermercados pueden utilizar la data analytics para mejorar sus operaciones? *Liora*. https://liora.io/es/data-analytics-supermercados

Thomas Reardon, C. Peter Timmer, Christopher B. Barret, & Julio Berdegué. (2003). *The Rise of Supermarkets in Africa, Asia, and Latin America—Reardon—2003—American Journal of Agricultural Economics—Wiley Online Library*. https://onlinelibrary.wiley.com/doi/10.1111/j.0092-5853.2003.00520.x

Tukey, J. (1977). *Exploratory data analysis* (Addison-Wesley).

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (Seventh edition). Cengage.

ANEXOS

Anexo 1. Artículo científico sobre Big Data, segmentación de clientes y sistemas de recomendación en el sector retail

Fuente: Revista Ingeniería Industrial, Universidad de Lima.

Enlace:

Vista de Big data en el mundo del retail: segmentación de clientes y sistema de recomendación en una cadena de supermercados de Europa

Anexo 2. Behaviorally Interpretable Transactional Features for Customer Segmentation Using K-Means in Grocery Retail

Fuente: Edumatic Journal.

Enlace:

View of Behaviorally Interpretable Transactional Features for Customer Segmentation Using K-Means in Grocery Retail

Anexo 3. Script utilizado para el procesamiento y Análisis de datos

Código fuente:

```
#Cargo las librerías
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
import plotly.graph_objects as go
#Kmeans para la clasificación
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score, mean_squared_error
from sklearn.model_selection import train_test_split
```

```
from scipy.spatial.distance import cdist

import statsmodels.api as sm # Importar statsmodels

#Montamos los datos en drive

from google.colab import drive

drive.mount('/content/drive')

# Ruta del archivo Excel en Colab

excel_path = '/content/drive/MyDrive/BASE_UNIFICADA_SUPERMERCADO
SCRIPT.xlsx'


# Cargar la primera hoja en un DataFrame

df_sheet1 = pd.read_excel(excel_path, sheet_name=0)


# Cargar la segunda hoja en un DataFrame

df_sheet2 = pd.read_excel(excel_path, sheet_name=1)


# Unificar los DataFrames (asumiendo que tienen las mismas columnas o
se desea apilar)

df_unificado = pd.concat([df_sheet1, df_sheet2], ignore_index=True)


# Mostrar las primeras 5 filas del DataFrame unificado

display(df_unificado.head())

# Mostrar información general del DataFrame

df_unificado.info()


# Mostrar estadísticas descriptivas de las columnas numéricas

display(df_unificado.describe())

# Función para aplicar Winsorización basada en el IQR
```

```
def winsorize_iqr(df, column):
    Q1 = df[column].quantile(0.25)
    Q3 = df[column].quantile(0.75)
    IQR = Q3 - Q1

    # Calcular límites para Winsorización
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Aplicar Winsorización: los valores por debajo del límite inferior se reemplazan por el límite inferior,
    # y los valores por encima del límite superior se reemplazan por el límite superior.
    df[column] = np.where(df[column] < lower_bound, lower_bound, df[column])
    df[column] = np.where(df[column] > upper_bound, upper_bound, df[column])
    return df

# Identificar columnas numéricas para aplicar la Winsorización
# Excluir 'Descuento (variable binaria)' y 'Crédito (variable binaria)' si son de tipo entero y ya están limpias/categorizadas,
# ya que la Winsorización podría no ser apropiada para variables binarias.
# También excluimos las columnas de fecha que hemos creado, ya que son categóricas por naturaleza o rangos definidos (año, mes, día).

numeric_cols = df_unificado.select_dtypes(include=np.number).columns.tolist()
```

```
# Excluir columnas binarias o de fecha/tiempo ya tratadas
exclude_cols = [
    'Descuento (variable binaria)',
    'Crédito (variable binaria)',
    'Ano', 'Mes', 'Dia_Semana', 'Dia_Mes', 'Es_Fin_Semana', 'Cluster'
]

# Filtrar las columnas numéricas para aplicar Winsorización
columns_to_winsorize = [col for col in numeric_cols if col not in
exclude_cols]

print(f"Columnas numéricas a las que se aplicará Winsorización:
{columns_to_winsorize}")

# Aplicar Winsorización a las columnas seleccionadas
for col in columns_to_winsorize:
    df_unificado = winsorize_iqr(df_unificado, col)
    print(f"Winsorización aplicada a la columna: {col}")

print("\nValores estadísticos del DataFrame después de la
Winsorización:")
display(df_unificado[columns_to_winsorize].describe())
# Contar valores nulos por columna
missing_values = df_unificado.isnull().sum()
missing_values = missing_values[missing_values >
0].sort_values(ascending=False)
```

```
print("Valores nulos por columna:")
display(missing_values)

# Visualizar valores nulos usando un mapa de calor
plt.figure(figsize=(12, 6))
sns.heatmap(df_unificado.isnull(), cbar=False, cmap='viridis')
plt.title('Mapa de Calor de Valores Nulos')
plt.show()
# Mostrar el número de valores únicos por columna
print("Número de valores únicos por columna:")
display(df_unificado.nunique().sort_values(ascending=False))
# Extraer año, mes, día de la semana y día del mes
df_unificado['Ano'] = df_unificado['Fecha de venta'].dt.year
df_unificado['Mes'] = df_unificado['Fecha de venta'].dt.month
df_unificado['Dia_Semana'] = df_unificado['Fecha de
venta'].dt.dayofweek # Lunes=0, Domingo=6
df_unificado['Dia_Mes'] = df_unificado['Fecha de venta'].dt.day

# Crear una variable binaria para indicar si es fin de semana
df_unificado['Es_Fin_Semana'] =
df_unificado['Dia_Semana'].apply(lambda x: 1 if x >= 5 else 0)

# Eliminar la columna 'Fecha de venta' original si ya no es necesaria,
o mantenerla según se considere.
# Por ahora, la mantendremos para referencia, pero la eliminaremos
antes del modelado si no es una característica directa.
```

```
print("Primeras filas del DataFrame con las nuevas características temporales:")

display(df_unificado[['Fecha de venta', 'Ano', 'Mes', 'Dia_Semana', 'Dia_Mes', 'Es_Fin_Semana']].head())

# Imputar valores nulos en 'Cantidad de productos' con la mediana

df_unificado['Cantidad de productos'] = df_unificado['Cantidad de productos'].fillna(df_unificado['Cantidad de productos'].median())


# Imputar valores nulos en 'Tipo de crédito' con la mediana

df_unificado['Tipo de crédito'] = df_unificado['Tipo de crédito'].fillna(df_unificado['Tipo de crédito'].median())


# Imputar valores nulos en 'Cantón' con la moda

df_unificado['Cantón'] = df_unificado['Cantón'].fillna(df_unificado['Cantón'].mode()[0])


# Imputar valores nulos en 'Tipo de productos' con la moda

df_unificado['Tipo de productos'] = df_unificado['Tipo de productos'].fillna(df_unificado['Tipo de productos'].mode()[0])


# Imputar valores nulos en 'Tipo de cliente' con la moda

df_unificado['Tipo de cliente'] = df_unificado['Tipo de cliente'].fillna(df_unificado['Tipo de cliente'].mode()[0])


# Imputar valores nulos en 'Crédito (variable binaria)' con la moda

df_unificado['Crédito (variable binaria)'] = df_unificado['Crédito (variable binaria)'].fillna(df_unificado['Crédito (variable binaria)'].mode()[0])


print("Valores nulos después de la imputación:")
```

```
display(df_unificado.isnull().sum()[df_unificado.isnull().sum() > 0])

# Columnas a eliminar por tener todos sus valores nulos

columns_to_drop = ['cedula', 'no. factura', 'Tipo de descuento']


# Eliminar las columnas del DataFrame

df_unificado = df_unificado.drop(columns=columns_to_drop)


print(f"Columnas eliminadas: {columns_to_drop}")

print("Información del DataFrame después de eliminar columnas:")

df_unificado.info()

# Convertir 'Fecha de venta' a datetime

df_unificado['Fecha de venta'] = pd.to_datetime(df_unificado['Fecha de
venta'])


print("Tipo de dato de 'Fecha de venta' después de la conversión:")

print(df_unificado['Fecha de venta'].dtype)


# Mostrar las primeras filas con la columna de fecha convertida

display(df_unificado.head())

# Create the datetime features before selecting them

# Extraer año, mes, día de la semana y día del mes

df_unificado['Ano'] = df_unificado['Fecha de venta'].dt.year

df_unificado['Mes'] = df_unificado['Fecha de venta'].dt.month

df_unificado['Dia_Semana'] = df_unificado['Fecha de
venta'].dt.dayofweek # Lunes=0, Domingo=6

df_unificado['Dia_Mes'] = df_unificado['Fecha de venta'].dt.day
```

```
# Crear una variable binaria para indicar si es fin de semana
df_unificado['Es_Fin_Semana'] =
df_unificado['Dia_Semana'].apply(lambda x: 1 if x >= 5 else 0)



# Seleccionar características para la clusterización
# Incluimos las variables numéricas y las categóricas que consideramos
relevantes para la demanda de clientes
features = [
    'Volumen de venta',
    'Precio de la venta',
    'Cantidad de productos',
    'Descuento (variable binaria)',
    'Crédito (variable binaria)',
    'Tipo de cliente',
    'Local de venta',
    'Ano',
    'Mes',
    'Dia_Semana',
    'Dia_Mes',
    'Es_Fin_Semana'
]

df_cluster_features = df_unificado[features].copy()

# Identificar columnas categóricas para One-Hot Encoding
categorical_features = ['Tipo de cliente', 'Local de venta']
```

```
# Aplicar One-Hot Encoding a las variables categóricas seleccionadas

df_cluster_features = pd.get_dummies(df_cluster_features,
columns=categorical_features, drop_first=True)


print("Primeras filas del DataFrame con características para
clusterización (después de One-Hot Encoding):")

display(df_cluster_features.head())

# Escalar las características numéricas

scaler = StandardScaler()

df_scaled = scaler.fit_transform(df_cluster_features)


# Convertir el array escalado de nuevo a un DataFrame para mantener
los nombres de las columnas

df_scaled = pd.DataFrame(df_scaled,
columns=df_cluster_features.columns)


print("Primeras filas del DataFrame con características escaladas:")

display(df_scaled.head())

 Método del Codo para encontrar el número óptimo de clusters

wcss = []

# Se prueban con un rango de K, por ejemplo, de 1 a 10. Ajusta según
sea necesario.

for i in range(1, 11):

    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300,
n_init=10, random_state=42)

    kmeans.fit(df_scaled)

    wcss.append(kmeans.inertia_)
```

```
# Graficar el Método del Codo

plt.figure(figsize=(10, 6))

plt.plot(range(1, 11), wcss, marker='o', linestyle='--')

plt.title('Método del Codo para K Óptimo')

plt.xlabel('Número de Clusters (K)')

plt.ylabel('WCSS')

plt.grid(True)

plt.show()

# Aplicar K-Means con el número óptimo de clusters (K=3 como ejemplo)

# Ajusta n_clusters según el análisis del método del codo y silueta.

optimal_k = 3

kmeans_model = KMeans(n_clusters=optimal_k, init='k-means++',
max_iter=300, n_init=10, random_state=42)

kmeans_model.fit(df_scaled)


# Asignar los clusters al DataFrame original

df_unificado['Cluster'] = kmeans_model.labels_


print(f"Se han creado {optimal_k} clusters.")

print("Conteo de clientes por cluster:")

display(df_unificado['Cluster'].value_counts())


print("Primeras filas del DataFrame con la columna de Cluster
asignada:")

display(df_unificado.head())

# Analizar las características de cada cluster

# Agrupar el DataFrame unificado por la columna 'Cluster' y calcular
la media de las características numéricas relevantes.
```

```
# Seleccionamos las mismas características que se usaron para el clustering, pero ahora del df_unificado original

# (o df_cluster_features antes de escalar, si se quiere ver valores originales)

cluster_analysis_features = [col for col in df_cluster_features.columns if 'Tipo de cliente' not in col and 'Local de venta' not in col] # Excluimos las one-hot encoded para simplificar la vista


# Incluir las columnas numéricas originales y la columna 'Cluster'

# Es importante asegurarse de que las columnas que se analizan tengan sentido para la interpretación de los clusters.

# Usaremos el df_unificado para ver los valores originales, no los escalados.


cluster_summary = df_unificado.groupby('Cluster')[['Volumen de venta', 'Precio de la venta', 'Cantidad de productos', 'Descuento (variable binaria)', 'Crédito (variable binaria)', 'Ano', 'Mes', 'Dia_Semana', 'Dia_Mes', 'Es_Fin_Semana']].mean()


print("Características promedio de cada cluster:")

display(cluster_summary)


# Para las características categóricas, podemos ver las modas o distribuciones

print("Distribución de 'Tipo de cliente' por cluster:")

display(pd.crosstab(df_unificado['Cluster'], df_unificado['Tipo de cliente'], normalize='index'))


print("Distribución de 'Local de venta' por cluster:")
```

```
display(pd.crosstab(df_unificado['Cluster'], df_unificado['Local de
venta'], normalize='index'))

from sklearn.decomposition import PCA


# Aplicar PCA para reducir a 2 componentes para visualización

pca = PCA(n_components=2)

df_pca = pca.fit_transform(df_scaled)


# Crear un DataFrame con los componentes PCA y las etiquetas de
cluster

df_pca = pd.DataFrame(df_pca, columns=['PC1', 'PC2'])

df_pca['Cluster'] = df_unificado['Cluster']


# Visualizar los clusters

plt.figure(figsize=(10, 8))

sns.scatterplot(x='PC1', y='PC2', hue='Cluster', data=df_pca,
palette='viridis', s=50, alpha=0.7)

plt.title('Clusters de Clientes (PCA 2D)')

plt.xlabel('Componente Principal 1')

plt.ylabel('Componente Principal 2')

plt.legend(title='Cluster')

plt.grid(True)

plt.show()

# Añadir la columna de 'Cluster' al DataFrame de características para
poder filtrar por cluster

df_reg_features = df_cluster_features.copy()

df_reg_features['Cluster'] = df_unificado['Cluster']
```

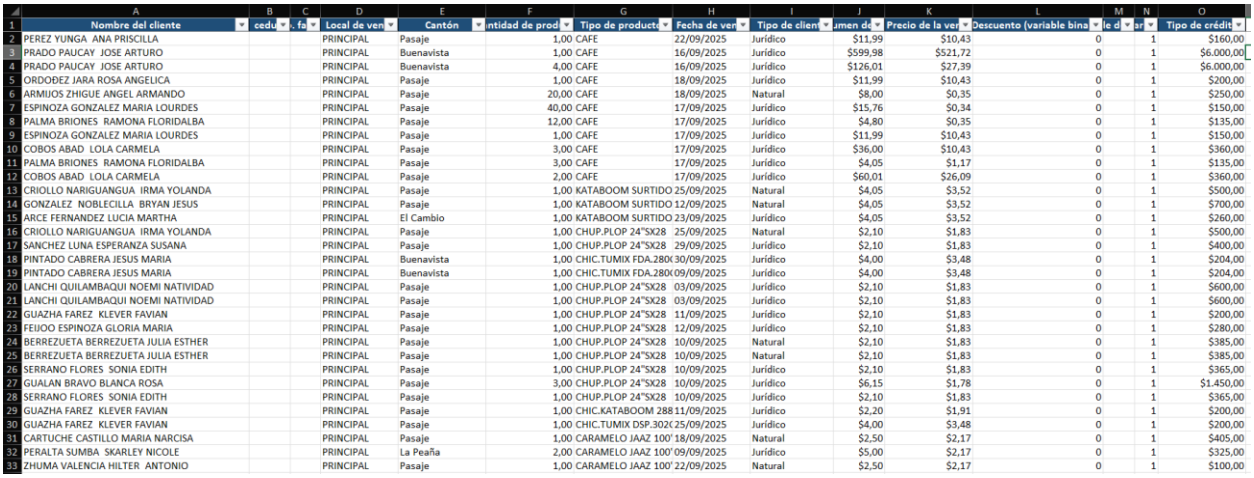
```
# Diccionario para almacenar los modelos y sus resultados por cluster (sklearn)
regression_models_sklearn = {}
regression_results_sklearn = {}

# Diccionario para almacenar los modelos de statsmodels y sus P-values
regression_models_sm = {}
regression_results_sm = {}

# Iterar sobre cada cluster para entrenar un modelo de regresión lineal
for cluster_id in sorted(df_reg_features['Cluster'].unique()):
    print(f"\n--- Entrenando modelo para el Cluster {cluster_id} ---")

    # Filtrar los datos para el cluster actual
    cluster_data = df_reg_features[df_reg_features['Cluster'] == cluster_id].copy()

    # Definir las características (X) y la variable objetivo (y)
    # 'Volumen de venta' es la variable objetivo.
    # Excluimos 'Cluster' y 'Volumen de venta' de las características.
    X = cluster_data.drop(columns=['Volumen de venta', 'Cluster'])
    y = cluster_data['Volumen de venta']

    # Convertir columnas booleanas a enteros para statsmodels, ya que puede tener problemas con True/False
    for col in X.select_dtypes(include=['bool']).columns:
        X[col] = X[col].astype(int)
```

```
    # Dividir los datos en conjuntos de entrenamiento y prueba
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)


    # --- Modelo con scikit-learn (para R-squared y MSE) ---
    model_sklearn = LinearRegression()
    model_sklearn.fit(X_train, y_train)
    regression_models_sklearn[cluster_id] = model_sklearn


    y_pred_sklearn = model_sklearn.predict(X_test)
    r_squared_sklearn = model_sklearn.score(X_test, y_test)
    mse_sklearn = mean_squared_error(y_test, y_pred_sklearn)
    regression_results_sklearn[cluster_id] = {'R-squared':
r_squared_sklearn, 'MSE': mse_sklearn, 'y_test': y_test,
'y_pred_sklearn': y_pred_sklearn}


    print(f"R-squared (sklearn) para el Cluster {cluster_id}:
{r_squared_sklearn:.4f}")
    print(f"MSE (sklearn) para el Cluster {cluster_id}:
{mse_sklearn:.4f}")


    # Mostrar los coeficientes del modelo de sklearn
    print("Coeficientes del Modelo (sklearn):")
    coefficients_df_sklearn = pd.DataFrame({'Feature': X.columns,
'Coefficient': model_sklearn.coef_})
    display(coefficients_df_sklearn.sort_values(by='Coefficient',
ascending=False))
```

```
    # --- Modelo con statsmodels (para P-values) ---
    # Se añade una constante para el término de intercepción (intercepto)
    X_train_sm = sm.add_constant(X_train)
    model_sm = sm.OLS(y_train, X_train_sm)
    results_sm = model_sm.fit()
    regression_models_sm[cluster_id] = results_sm

    print("\nP-values del Modelo (statsmodels):")
    p_values_df = pd.DataFrame(results_sm.pvalues, columns=['P-value'])
    display(p_values_df)

print("\n--- Resumen de los resultados de regresión por Cluster (sklearn) ---")
for cluster_id, metrics in regression_results_sklearn.items():
    print(f"Cluster {cluster_id}: R-squared={metrics['R-squared']:.4f}, MSE={metrics['MSE']:.4f}")
print("\nP-values del Modelo (statsmodels) para el Cluster 0:")
display(pd.DataFrame(regression_models_sm[0].pvalues, columns=['P-value']))

print("\nP-values del Modelo (statsmodels) para el Cluster 1:")
display(pd.DataFrame(regression_models_sm[1].pvalues, columns=['P-value']))

print("\nP-values del Modelo (statsmodels) para el Cluster 2:")
display(pd.DataFrame(regression_models_sm[2].pvalues, columns=['P-value']))
```

```
import matplotlib.pyplot as plt


plt.figure(figsize=(18, 5))


for cluster_id in sorted(regression_results_sklearn.keys()):
    results = regression_results_sklearn[cluster_id]
    y_test = results['y_test']
    y_pred = results['y_pred_sklearn']
    r_squared = results['R-squared']
    mse = results['MSE']


    plt.subplot(1, len(regression_results_sklearn), cluster_id + 1)
    plt.scatter(y_test, y_pred, alpha=0.3, s=10)
    plt.plot([y_test.min(), y_test.max()], [y_test.min(),
y_test.max()], 'r--', lw=2)
    plt.xlabel('Volumen de Venta Real')
    plt.ylabel('Volumen de Venta Predicho')
    plt.title(f'Cluster {cluster_id}\n(R-squared: {r_squared:.2f},
MSE: {mse:.2f})')
    plt.grid(True)


plt.tight_layout()
plt.show()
```

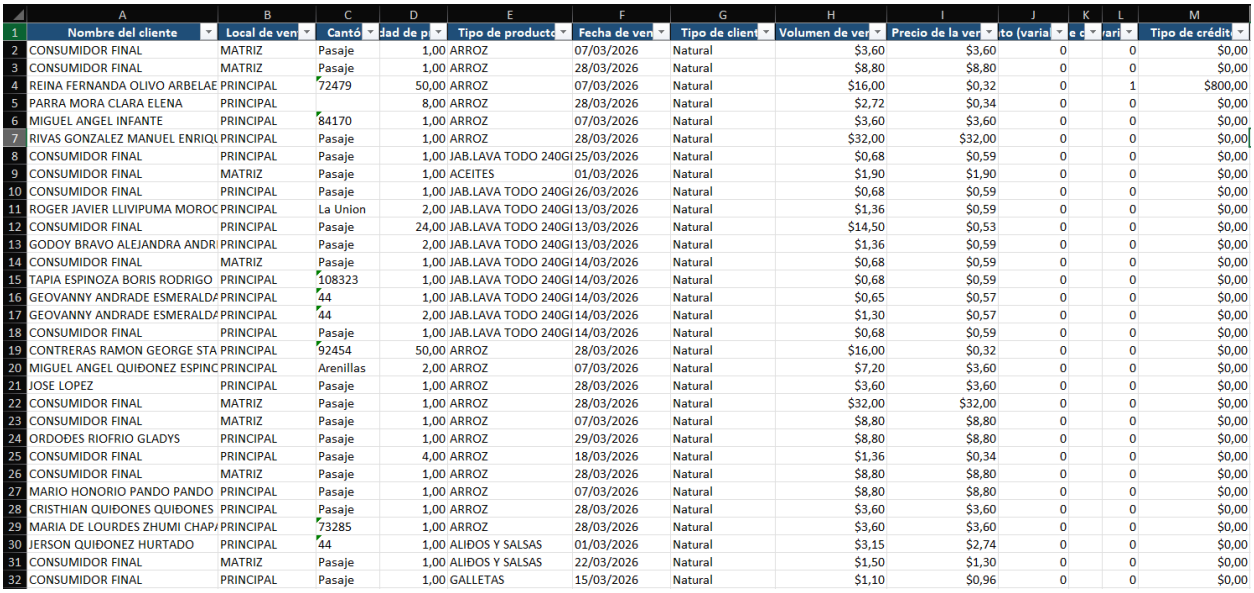
Anexo 4. Base de datos de clientes y registros de ventas



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Nombre del cliente	cedu	fa	Local de ven	Cantón	ntidad de prod	Tipo de producto	Fecha de ven	Tipo de client	umen de	Precio de la ver	Descuento (variable bina	le d	ar	Tipo de crédit
2	PEREZ YUNGA ANA PRISCILLA			PRINCIPAL	Pasaje	1,00	CAFE	22/09/2025	Jurídico	$11,99	$10,43	0		1	$160,00
3	PRADO PAUCAY JOSE ARTURO			PRINCIPAL	Buenavista	1,00	CAFE	16/09/2025	Jurídico	$599,98	$521,72	0		1	$6.000,00
4	PRADO PAUCAY JOSE ARTURO			PRINCIPAL	Buenavista	4,00	CAFE	16/09/2025	Jurídico	$126,01	$27,39	0		1	$6.000,00
5	ORDOÐEZ JARA ROSA ANGELICA			PRINCIPAL	Pasaje	1,00	CAFE	18/09/2025	Jurídico	$11,99	$10,43	0		1	$200,00
6	ARMIJOS ZHIGUE ANGEL ARMANDO			PRINCIPAL	Pasaje	20,00	CAFE	18/09/2025	Natural	$8,00	$0,35	0		1	$250,00
7	ESPINOZA GONZALEZ MARIA LOURDES			PRINCIPAL	Pasaje	40,00	CAFE	17/09/2025	Jurídico	$15,76	$0,34	0		1	$150,00
8	PALMA BRIONES RAMONA FLORIDALBA			PRINCIPAL	Pasaje	12,00	CAFE	17/09/2025	Jurídico	$4,80	$0,35	0		1	$135,00
9	ESPINOZA GONZALEZ MARIA LOURDES			PRINCIPAL	Pasaje	1,00	CAFE	17/09/2025	Jurídico	$11,99	$10,43	0		1	$150,00
10	COBOS ABAD LOLA CARMELA			PRINCIPAL	Pasaje	3,00	CAFE	17/09/2025	Jurídico	$36,00	$10,43	0		1	$360,00
11	PALMA BRIONES RAMONA FLORIDALBA			PRINCIPAL	Pasaje	3,00	CAFE	17/09/2025	Jurídico	$4,05	$1,17	0		1	$135,00
12	COBOS ABAD LOLA CARMELA			PRINCIPAL	Pasaje	2,00	CAFE	17/09/2025	Jurídico	$60,01	$26,09	0		1	$360,00
13	CRIOLLO NARIGUANGUA IRMA YOLANDA			PRINCIPAL	Pasaje	1,00	KATABOOM SURTIDO	25/09/2025	Natural	$4,05	$3,52	0		1	$500,00
14	GONZALEZ NOBLECILLA BRYAN JESUS			PRINCIPAL	Pasaje	1,00	KATABOOM SURTIDO	12/09/2025	Natural	$4,05	$3,52	0		1	$700,00
15	ARCE FERNANDEZ LUCIA MARTHA			PRINCIPAL	El Cambio	1,00	KATABOOM SURTIDO	23/09/2025	Jurídico	$4,05	$3,52	0		1	$260,00
16	CRIOLLO NARIGUANGUA IRMA YOLANDA			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	25/09/2025	Natural	$2,10	$1,83	0		1	$500,00
17	SANCHEZ LUNA ESPERANZA SUSANA			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	29/09/2025	Jurídico	$2,10	$1,83	0		1	$400,00
18	PINTADO CABRERA JESUS MARIA			PRINCIPAL	Buenavista	1,00	CHIC.TUMIX FDA.280(	30/09/2025	Jurídico	$4,00	$3,48	0		1	$204,00
19	PINTADO CABRERA JESUS MARIA			PRINCIPAL	Buenavista	1,00	CHIC.TUMIX FDA.280(	09/09/2025	Jurídico	$4,00	$3,48	0		1	$204,00
20	LANCHI QUILAMBAQUI NOEMI NATIVIDAD			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	03/09/2025	Jurídico	$2,10	$1,83	0		1	$600,00
21	LANCHI QUILAMBAQUI NOEMI NATIVIDAD			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	03/09/2025	Jurídico	$2,10	$1,83	0		1	$600,00
22	GUAZHA FAREZ KLEVER FAVIAN			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	11/09/2025	Jurídico	$2,10	$1,83	0		1	$200,00
23	FEIJOO ESPINOZA GLORIA MARIA			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	12/09/2025	Jurídico	$2,10	$1,83	0		1	$280,00
24	BERREZUETA BERREZUETA JULIA ESTHER			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	10/09/2025	Natural	$2,10	$1,83	0		1	$385,00
25	BERREZUETA BERREZUETA JULIA ESTHER			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	10/09/2025	Natural	$2,10	$1,83	0		1	$385,00
26	SERRANO FLORES SONIA EDITH			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	10/09/2025	Jurídico	$2,10	$1,83	0		1	$365,00
27	GUALAN BRAVO BLANCA ROSA			PRINCIPAL	Pasaje	3,00	CHUP.PLOP 24"SX28	10/09/2025	Jurídico	$6,15	$1,78	0		1	$1.450,00
28	SERRANO FLORES SONIA EDITH			PRINCIPAL	Pasaje	1,00	CHUP.PLOP 24"SX28	10/09/2025	Jurídico	$2,10	$1,83	0		1	$365,00
29	GUAZHA FAREZ KLEVER FAVIAN			PRINCIPAL	Pasaje	1,00	CHIC.KATABOOM 288	11/09/2025	Jurídico	$2,20	$1,91	0		1	$200,00
30	GUAZHA FAREZ KLEVER FAVIAN			PRINCIPAL	Pasaje	1,00	CHIC.TUMIX DSP.302(	25/09/2025	Jurídico	$4,00	$3,48	0		1	$200,00
31	CARTUCHE CASTILLO MARIA NARCISA			PRINCIPAL	Pasaje	1,00	CARAMELO JAAZ 100'	18/09/2025	Natural	$2,50	$2,17	0		1	$405,00
32	PERALTA SUMBA SKARLEY NICOLE			PRINCIPAL	La Peaña	2,00	CARAMELO JAAZ 100'	09/09/2025	Jurídico	$5,00	$2,17	0		1	$325,00
33	ZHUMA VALENCIA HILTER ANTONIO			PRINCIPAL	Pasaje	1,00	CARAMELO JAAZ 100'	22/09/2025	Natural	$2,50	$2,17	0		1	$100,00

Nota. La imagen muestra una sección de la base de datos utilizada para el análisis, incluyendo información de clientes, productos, fechas de venta, cantidades, precios, descuentos y formas de crédito. Debido a su amplitud (aproximadamente 1.000.000 de registros), se presenta únicamente una sección representativa utilizada para el análisis.

Anexo 5. Base de datos de ventas y transacciones



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Nombre del cliente	Local de ven	Cantó	dad de p	Tipo de producto	Fecha de ven	Tipo de client	Volumen de ver	Precio de la ver	to (varia	e c	ari	Tipo de crédit
2	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ARROZ	07/03/2026	Natural	$3,60	$3,60	0		0	$0,00
3	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ARROZ	28/03/2026	Natural	$8,80	$8,80	0		0	$0,00
4	REINA FERNANDA OLIVO ARBELAE	PRINCIPAL	72479	50,00	ARROZ	07/03/2026	Natural	$16,00	$0,32	0		1	$800,00
5	PARRA MORA CLARA ELENA	PRINCIPAL		8,00	ARROZ	28/03/2026	Natural	$2,72	$0,34	0		0	$0,00
6	MIGUEL ANGEL INFANTE	PRINCIPAL	84170	1,00	ARROZ	07/03/2026	Natural	$3,60	$3,60	0		0	$0,00
7	RIVAS GONZALEZ MANUEL ENRIQ	PRINCIPAL	Pasaje	1,00	ARROZ	28/03/2026	Natural	$32,00	$32,00	0		0	$0,00
8	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	1,00	JAB.LAVA TODO 240GI	25/03/2026	Natural	$0,68	$0,59	0		0	$0,00
9	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ACEITES	01/03/2026	Natural	$1,90	$1,90	0		0	$0,00
10	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	1,00	JAB.LAVA TODO 240GI	26/03/2026	Natural	$0,68	$0,59	0		0	$0,00
11	ROGER JAVIER LLIVIPUMA MOROC	PRINCIPAL	La Union	2,00	JAB.LAVA TODO 240GI	13/03/2026	Natural	$1,36	$0,59	0		0	$0,00
12	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	24,00	JAB.LAVA TODO 240GI	13/03/2026	Natural	$14,50	$0,53	0		0	$0,00
13	GODOY BRAVO ALEJANDRA ANDR	PRINCIPAL	Pasaje	2,00	JAB.LAVA TODO 240GI	13/03/2026	Natural	$1,36	$0,59	0		0	$0,00
14	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	JAB.LAVA TODO 240GI	14/03/2026	Natural	$0,68	$0,59	0		0	$0,00
15	TAPIA ESPINOZA BORIS RODRIGO	PRINCIPAL	108323	1,00	JAB.LAVA TODO 240GI	14/03/2026	Natural	$0,68	$0,59	0		0	$0,00
16	GEOVANNY ANDRADE ESMERALDA	PRINCIPAL	44	1,00	JAB.LAVA TODO 240GI	14/03/2026	Natural	$0,65	$0,57	0		0	$0,00
17	GEOVANNY ANDRADE ESMERALDA	PRINCIPAL	44	2,00	JAB.LAVA TODO 240GI	14/03/2026	Natural	$1,30	$0,57	0		0	$0,00
18	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	1,00	JAB.LAVA TODO 240GI	14/03/2026	Natural	$0,68	$0,59	0		0	$0,00
19	CONTRERAS RAMON GEORGE STA	PRINCIPAL	92454	50,00	ARROZ	28/03/2026	Natural	$16,00	$0,32	0		0	$0,00
20	MIGUEL ANGEL QUIÐONEZ ESPINC	PRINCIPAL	Arenillas	2,00	ARROZ	07/03/2026	Natural	$7,20	$3,60	0		0	$0,00
21	JOSE LOPEZ	PRINCIPAL	Pasaje	1,00	ARROZ	28/03/2026	Natural	$3,60	$3,60	0		0	$0,00
22	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ARROZ	28/03/2026	Natural	$32,00	$32,00	0		0	$0,00
23	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ARROZ	07/03/2026	Natural	$8,80	$8,80	0		0	$0,00
24	ORDOÐES RIOFRIO GLADYS	PRINCIPAL	Pasaje	1,00	ARROZ	29/03/2026	Natural	$8,80	$8,80	0		0	$0,00
25	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	4,00	ARROZ	18/03/2026	Natural	$1,36	$0,34	0		0	$0,00
26	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ARROZ	28/03/2026	Natural	$8,80	$8,80	0		0	$0,00
27	MARIO HONORIO PANDO PANDO	PRINCIPAL	Pasaje	1,00	ARROZ	07/03/2026	Natural	$8,80	$8,80	0		0	$0,00
28	CRISTHIAN QUIÐONES QUIÐONES	PRINCIPAL	Pasaje	1,00	ARROZ	28/03/2026	Natural	$3,60	$3,60	0		0	$0,00
29	MARIA DE LOURDES ZHUMI CHAP	PRINCIPAL	73285	1,00	ARROZ	28/03/2026	Natural	$3,60	$3,60	0		0	$0,00
30	JERSON QUIÐONEZ HURTADO	PRINCIPAL	44	1,00	ALIÐOS Y SALSAS	01/03/2026	Natural	$3,15	$2,74	0		0	$0,00
31	CONSUMIDOR FINAL	MATRIZ	Pasaje	1,00	ALIÐOS Y SALSAS	22/03/2026	Natural	$1,50	$1,30	0		0	$0,00
32	CONSUMIDOR FINAL	PRINCIPAL	Pasaje	1,00	GALLETAS	15/03/2026	Natural	$1,10	$0,96	0		0	$0,00

Nota. La base de datos contiene los registros de las transacciones realizadas, incluyendo información del cliente, local de venta, cantón, cantidad de productos, tipo de producto, fecha de venta, tipo de cliente, volumen de venta, precio de venta, descuentos y tipo de crédito. Debido a su amplitud (aproximadamente 400.000 registros), se presenta únicamente una sección representativa utilizada para el análisis.

DECLARACIÓN Y AUTORIZACIÓN

Nosotros, Galvez Guillen, Kristel Elizabeth con C.C: #0750411480 y Criollo Montiel, Sonia Elena, con C.C: #0943938225 autores del trabajo de integración curricular**: Análisis de la demanda de clientes bajo factores significativos – Cluster**, previo a la obtención del título de **Licenciado de Negocios Internacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

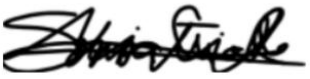
2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 22 de agosto de 2026

f. _____

Galvez Guillen, Kristel Elizabeth

C.C: 0750411480



f. _____

Criollo Montiel, Sonia Elena

C.C: 0943938225

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN			
TEMA Y SUBTEMA:	Análisis de la demanda de clientes bajo factores significativos – Cluster		
AUTOR:	Criollo Montiel, Sonia Elena Galvez Guillen, Kristel Elizabeth		
REVISOR/TUTOR:	Lucin Castillo, Virginia Carolina		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Facultad de Economía y Empresa		
CARRERA:	Negocios Internacionales		
TITULO OBTENIDO:	Licenciada en Negocios Internacionales		
FECHA DE PUBLICACIÓN:	22 de agosto de 2026	No. DE PÁGINAS:	208
ÁREAS TEMÁTICAS:	Comportamiento de compra, estrategias comerciales, mercadeo		
PALABRAS CLAVES/ KEYWORDS:	Segmentación de clientes; análisis de la demanda; K-Means; análisis de componentes principales; regresión lineal múltiple; supermercados; comportamiento del consumidor.		

RESUMEN/ABSTRACT (150-250 palabras): Esta investigación analiza la demanda de los clientes mediante el agrupamiento de factores significativos en dos sucursales de un supermercado ubicado en la provincia de El Oro. Esta investigación se motiva por el hecho de que los consumidores no tienen el mismo comportamiento de compra independientemente del volumen comprado, la frecuencia de compra, el uso de descuentos y créditos, el tipo de cliente y otras características comerciales. Solo las ventas totales son suficientes para entender el rendimiento de las ventas, pero no los perfiles que realmente sostienen y explican la demanda. Por esta razón, la idea principal es estudiar la demanda a través de factores agrupados importantes para comprender las características del consumidor e identificar el segmento con mayor impacto en las ventas. La investigación se lleva a cabo en un marco cuantitativo y analítico basado en la información numérica de las transacciones comerciales de las dos sucursales. La base de datos está compuesta por ventas, clientes, productos, descuentos, créditos y características de las transacciones. Antes de aplicar los modelos, se realizó una preparación y limpieza de los datos que incluyó la identificación de registros duplicados, operaciones canceladas, errores de digitalización, valores atípicos, organización y estandarización de las variables. Esto nos permitió tener información consistente para realizar análisis estadísticos y aplicar técnicas de aprendizaje automático. Finalmente, realizamos un análisis exploratorio de datos para entender la distribución y el comportamiento de las variables comerciales. Para la segmentación, utilizamos K-Means, un método de aprendizaje no supervisado que ayuda a la clasificación de observaciones de calidad similar. Las variables fueron estandarizadas para evitar que sus diferentes escalas afectaran la distancia desde la cual se utilizó el algoritmo. Como resultado, establecimos tres grupos de clientes en los algoritmos. Los grupos 0 y 1 mostraron un comportamiento relativamente similar con clientes siendo principalmente individuales y con bajo volumen de ventas y cantidad de productos. Sin embargo, el grupo 2 tenía un mayor volumen de ventas y más productos, alto uso de crédito y clientes corporativos. En otras palabras, este es el segmento que es más intensivo y comercial por naturaleza. También se utilizó el Análisis de Componentes Principales (PCA) para mostrar los segmentos gráficamente y visualmente para ver cómo difieren los diferentes segmentos. Los resultados revelaron que hay tres grupos distintos, aunque la dispersión del grupo 2 es más dispersa de los otros grupos y se superpone con algunos de ellos. Luego, utilizamos modelos de regresión lineal múltiple del grupo con el volumen de ventas como variable dependiente y variables como precio de

venta, cantidad de producto, descuento, crédito, factor tiempo, tipo de cliente y ubicación de ventas como variables explicativas para indicar que los factores asociados con la demanda son más importantes para el segmento del estudio. En conclusión, los resultados muestran que los clientes no deben ser analizados como un solo grupo ya que hay varios patrones de comportamiento en la demanda. Al combinar K-Means y PCA, podemos identificar perfiles comerciales y entender los factores que influyen en las ventas. Esta información se utiliza para crear estrategias diferenciadas para la lealtad, promociones, crédito, inventario y planificación comercial; apuntando a clientes de alto valor y creando incentivos específicos para esos segmentos de clientes que tendrán potencial de crecimiento. A través de técnicas analíticas y conocimientos, los datos transaccionales se convierten en la base para una toma de decisiones más precisa basada en el comportamiento real de los consumidores.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:** +593	E-mail:
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Nombre:** Freire Quintero, Cesar Enrique	
	Teléfono: +593 990090702	
	E-mail: cesar.freire@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	