Trabajo previo a la obtención de título de
Magister en

Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

Cristhian Javier Castro Gaibor

Deysi Estefania Jaque Intriago

Fernando Mauricio Siguenza Sarmiento

Ricardo Andres Cartagena Cueva

Verónica Nathaly Chicaiza Moreira

TUTOR/ES:

Karla Estefanía Mora

Fernanda Paulina Vizcaíno

Segmentación inteligente de clientes de un e-commerce florícola mediante análisis de comportamiento de compra en EE.UU. y Canadá

2025

REINVENTEMOS EL FUTURO

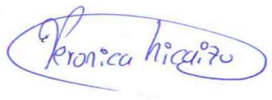
Certificación de autoría

Nosotros, **Cristhian Javier Castro Gaibor, Deysi Estefania Jaque Intriago, Fernando Mauricio Siguenza Sarmiento, Ricardo Andres Cartagena Cueva, Verónica Nathaly Chicaiza Moreira** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma

Deysi Estefania Jaque Intriago



Firma

Verónica Nataly Chicaiza Moreria

Firma

Fernando Mauricio Sigûenza Sarmiento

Firma

Cristhian Javier Castro Gaibor

Firma

Ricardo Andres Cartagena Cueva

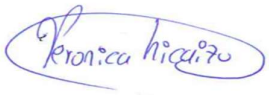
Autorización de Derechos de Propiedad Intelectual

Nosotros, **Cristhian Javier Castro Gaibor, Deysi Estefania Jaque Intriago, Fernando Mauricio Siguenza Sarmiento, Ricardo Andres Cartagena Cueva, Verónica Nathaly Chicaiza Moreira**, en calidad de autores del trabajo de investigación titulado ***Segmentación inteligente de clientes de un e-commerce florícola mediante análisis de comportamiento de compra en EE.UU. y Canadá***, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, diciembre 2025

Firma

Deysi Estefania Jaque Intriago



Firma

Verónica Nataly Chicaiza Moreira

Firma

Fernando Mauricio Sigûenza Sarmiento

Firma

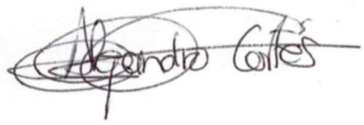
Cristhian Javier Castro Gaibor

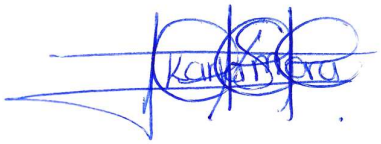
Firma

Ricardo Andres Cartagena Cueva

Aprobación de dirección y coordinación del programa

Nosotros, **Alejando Cortés Director EIG y Karla Mora Coordinadora UIDE,** declaramos que: **Cristhian Javier Castro Gaibor, Deysi Estefania Jaque Intriago, Fernando Mauricio Siguenza Sarmiento, Ricardo Andres Cartagena Cueva, Verónica Nathaly Chicaiza Moreira,** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.





Alejandro Cortés López
Director de la
Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

---

Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de Datos y Maquinas de Aprendizaje con Mención en Inteligencia Artificial

DEDICATORIA

El presente trabajo es resultado del esfuerzo en conjunto, la responsabilidad y el compromiso académico con quienes compartimos un proceso de aprendizaje marcado por el esfuerzo constante, la responsabilidad y el compromiso académico asumido por cada uno de los integrantes del equipo. La colaboración, el respeto mutuo y la disposición para afrontar los desafíos que surgieron a lo largo del desarrollo de esta investigación fueron fundamentales para alcanzar los objetivos planteados. Este trabajo es reflejo del trabajo en equipo, la perseverancia y el intercambio de conocimientos que permitieron consolidar este proyecto académico.

AGRADECIMIENTOS

Queremos expresar nuestro más sincero agradecimiento a nuestros docentes, quienes con su guía, conocimiento y paciencia orientaron cada etapa de esta investigación, proporcionando valiosos consejos que fortalecieron la calidad académica de este trabajo. Asimismo, agradecemos al personal de la institución por el apoyo brindado en recursos, acceso a información y facilidades logísticas que fueron fundamentales para el desarrollo del proyecto. De igual manera, reconocemos y valoramos la colaboración de nuestros compañeros de grupo, cuyo esfuerzo, responsabilidad y trabajo conjunto hicieron posible la culminación exitosa. Este proyecto no solo refleja la aplicación de conocimientos adquiridos, sino también el compromiso y la dedicación colectiva que caracterizan nuestro trabajo en equipo.

Finalmente, agradecemos a todas las personas e instituciones que, de manera directa o indirecta, contribuyeron al logro de este objetivo académico, reafirmando la importancia del apoyo, la cooperación y la orientación en el proceso de formación profesional.

RESUMEN

Para la empresa Ecuador Direct Roses, se establece como objetivo estratégico la segmentación del mercado en grupos de consumidores que presenten características, necesidades o patrones de comportamiento homogéneos. Esta clasificación permitirá identificar clientes potenciales con el objetivo de orientar las acciones comerciales hacia aquellos segmentos con mayor potencial, optimizando así la toma de decisiones y el impacto de las iniciativas corporativas

Los resultados de la facturación permitieron que se registre la información básica y necesaria para identificar el comportamiento de los clientes recurrentes, estacionales, mayoristas y esporádicos, revelando el consumo de sus productos.

Tras un proceso de limpieza, imputación, codificación y normalización de variables categóricas y numéricas con patrones de compra estacionales nos conducen a nuevas perspectivas de negocio basada en los datos.

El proyecto tiene como objetivo principal generar una segmentación de los clientes de la empresa Ecuador Direct Roses a través de un análisis descriptivo de datos que se construyen en base a los indicadores derivados de las variables RFM las mismas que se ejecutan en un modelo de clústering K-Means, donde se evalúa métricas como silhouette score, Davies-Bouldin y técnicas de codo, seleccionando la configuración que mejor representa la estructura de los datos.

Palabras Clave: segmentación de los clientes, modelo RFM, modelo K-means, análisis de clúster.

ABSTRACT

The project presents the development of a customer segmentation system based on a K-Means model and unsupervised learning techniques, using transactional data from the e-commerce platform of the Ecuadorian company Direct Roses, which exports flowers to the United States and Canada. In response to the challenges of the floriculture sector, such as demand seasonality, export logistics, and increasing international competition, the goal is to transform historical sales information into strategic knowledge through an analytical model capable of identifying purchasing behavior patterns and grouping customers optimally for strategic use. A three-year transactional dataset is used, containing categorical and numerical variables related to the customer, product characteristics, and operational order attributes. Derived variables such as RFM (Recency, Frequency, Monetary) and seasonal purchasing patterns are constructed. After a cleaning, imputation, encoding, and normalization process using Python (Pandas, Scikit-learn), derived variables like RFM are generated and analyzed using clústering techniques to classify customers into homogeneous groups according to their behavior. The performance of each model is evaluated using metrics such as silhouette score and Davies-Bouldin index, selecting the optimal configuration to represent the structure of the data. The results make it possible to identify key segments such as recurring, seasonal, wholesale, and sporadic customers, providing valuable information about their purchasing frequency and loyalty. Finally, the findings are integrated into an interactive Power BI dashboard that facilitates the interpretation of patterns and supports data-driven decision-making to optimize commercial strategies, campaign personalization, and loyalty management.

Keywords: customer segmentation, RFM model, cluster analysis

TABLA DE CONTENIDOS (Índice)

LISTA DE TABLAS (Índice de tablas)

LISTA DE FIGURAS

1. Introducción

1.1. Definición del proyecto

El presente proyecto consiste en la construcción de un modelo de segmentación de clientes en función de su comportamiento de compra a través del e-commerce de la empresa Ecuador Direct Roses, cuya especialidad es la venta de rosas en línea al mercado de Estados Unidos, y Canadá.

El objetivo central del proyecto de titulación es aplicar, algoritmos de aprendizaje no supervisado para analizar, comprender y clasificar la base de clientes en distintos grupos según su comportamiento de compra.

Este estudio transformará datos transaccionales en una visión estructurada y profunda de sus clientes, con la definición de características, que permitan conocer acerca de la estacionalidad, frecuencia, lealtad a ciertos productos. El conocimiento generado se materializa en un análisis detallado de cada segmento, y apoyado en plataformas de análisis de datos como Python (Pandas y Scikit-learn) para el procesamiento y segmentación, y Power BI para la visualización interactiva de los resultados.

El propósito es que Ecuador Direct Roses adopte un enfoque basado en la comprensión profunda y segmentada de sus clientes, optimizando las estrategias comerciales y orientando las promociones hacia los segmentos de mayor potencial. Esta nueva visión permitirá tomar decisiones fundamentadas en datos, reflejándose en una comunicación más efectiva, mayor lealtad del cliente y una mejor eficiencia en las estrategias de marketing.

1.2. Justificación e importancia del trabajo de investigación

La floricultura es uno de los sectores de mayor impacto económico en el Ecuador, siendo un rubro clave dentro de las exportaciones no tradicionales. Donde las condiciones climáticas y agrícolas del país se convierten en factores clave para el cultivo de flores con características sobresalientes, respecto al tamaño, color y durabilidad.

El crecimiento del e-commerce en los últimos años ha abierto nuevas oportunidades para conectar directamente a productores florícolas y compradores internacionales, siendo Estados Unidos uno de los principales países de destino de las exportaciones florícolas ecuatorianas. Además, el entorno digital brinda una gran cantidad de datos sobre el comportamiento de los clientes de este mercado que puede ser aprovechado.

Sin embargo, el sector enfrenta desafíos que exigen una transformación en la manera de entender al consumidor. Estos están relacionados con: la estacionalidad de la demanda, evidenciada en los registros históricos de ventas de la empresa, que muestran picos durante fechas con alta carga emocional y comercial como San Valentín, y Día de la Madre; la logística compleja de exportación, determinada por los tiempos de transporte y conservación del producto; y una competencia internacional creciente, junto con la expansión del mercado nacional.

En este contexto, la segmentación de clientes basada en datos digitales permitirá obtener una herramienta para anticipar patrones de compra según la temporada y destino, diseñar estrategias de marketing personalizadas para clientes de alto valor y programas de fidelización.

La importancia de este trabajo radica en la integración de la ciencia de datos con el conocimiento del sector florícola para transformar la manera en que este mercado comprende y atiende las necesidades de sus clientes, abriendo paso al camino de la personalización de campañas. Además, el contar con datos de dos mercados internacionales como Estados Unidos y Canadá, permite tener una base de referencia para la adopción de estrategias en otros países con dinámicas de consumo similares.

1.3. Alcance

En el presente trabajo de titulación se explorará el uso de algoritmos de agrupamiento aplicados a los datos históricos de ventas de un e-commerce dedicado a la venta de flores

(rosas y flores de verano) en Estados Unidos y Canadá. En una primera etapa se realizará la segmentación de clientes mediante técnicas de clústering.

Para lograrlo, se utilizarán los datos de los últimos tres años, los cuales contienen información de las órdenes, características de los clientes (país, estado, ciudad, origen del cliente), detalles de los productos adquiridos (producto, color, temporada, largo de tallo, cantidad de tallos, precio unitario) y variables transaccionales (total, fecha_vuelo, estado_orden).

El análisis se enfocará en la construcción de variables derivadas que permitan una comprensión multidimensional del comportamiento del consumidor. En primera instancia, se implementará el modelo RFM (Recency, Frequency, Monetary), el cual facilita la segmentación de los clientes a partir de la recencia de su última compra, la frecuencia de sus órdenes y el nivel de gasto promedio efectuado. Complementariamente, el estudio integra la identificación de preferencias de producto, analizando características como el tipo de flor (rosas y flores de verano), color, temporada, y largo de tallo. Finalmente, se profundiza enfocándose en los patrones de compra, diferenciando entre el consumo recurrente, la estacionalidad relacionada a fechas clave como San Valentín o el Día de la Madre, y las compras mayoristas definidas por el movimiento de grandes volúmenes de tallos.

Además de la segmentación, el análisis incluirá la clasificación de los clientes según su tipo de comportamiento, empleando técnicas de ciencia de datos como algoritmos de agrupamiento (K-Means, DBSCAN); esta combinación permitirá identificar y anticipar comportamientos clave.

En primer lugar, se buscará distinguir a los clientes Wholesaler, quienes realizan compras menos frecuentes, pero en grandes volúmenes, lo que los convierte en actores estratégicos para la empresa, son conocidos como clientes mayoristas adquiriendo cantidades considerables de producto con fines de reventa como arreglos florales, o al por mayor hacia

sus propios clientes. Para este grupo se deberían diseñar acuerdos comerciales especiales con políticas diferenciadas y estrategias para largo plazo.

Por otro lado, se identifican los clientes estacionales, los cuales concentran la mayor parte de sus compras en periodos específicos del año como San Valentín y el Día de las Madres; este segmento, compuesto principalmente por emprendedores o floristerías pequeñas, requiere la implementación de estrategias enfocadas en la anticipación de la demanda y ofertas temporales.

Asimismo, el análisis agrupa a los clientes nuevos, definidos como aquellos que realizan su compra por primera vez en el portal y cuya identificación requiere verificar la inexistencia de cuentas previas relacionadas a la misma persona o empresa.

De manera distinta, se clasifican los clientes esporádicos, caracterizados por una alta recencia, baja frecuencia y bajo gasto. En la mayoría de los casos corresponden a clientes que realizaron una única compra y no volvieron a interactuar con el portal. Este comportamiento puede ser por diversos factores como calidad del producto, percepciones del precio, tiempo de entrega u otros aspectos del servicio. Este grupo representa el segmento de menor valor por lo que se sugiere realizar alguna estrategia de remarketing y analizar la causa de su abandono.

Finalmente, se encuentran los clientes en riesgo, quienes mantienen un gasto y frecuencia medios, pero presentan una recencia en aumento lo que indica una disminución progresiva de la actividad de compra. Estos clientes son considerados en riesgo de abandono y requieren una intervención oportuna mediante alguna acción comercial con el objetivo de reactivar su comportamiento y evitar su pérdida definitiva.

En conclusión, el alcance de este proyecto comprende la aplicación de métodos de clústering sobre los datos históricos del e-commerce para obtener una segmentación robusta que permita tanto la identificación de preferencias de producto como la clasificación de tipos

de clientes. Los resultados serán útiles para la definición de campañas de marketing personalizadas y el diseño de estrategias de fidelización adaptadas a los distintos perfiles de clientes.

1.4. Objetivos

1.4.1. Objetivo general

Segmentar a los clientes de la florícola según su comportamiento de compra, con el fin de identificar patrones de consumo que orienten recomendaciones estratégicas para el área de marketing y el equipo comercial, presentadas mediante un dashboard interactivo y un informe ejecutivo.

1.4.2. Objetivos específicos

- Recopilar información histórica de las transacciones realizadas de los clientes de la florícola.
- Depurar y preprocesar los datos mediante técnicas de limpieza, normalización y transformación, garantizando su calidad y consistencia para el análisis posterior.
- Analizar datos históricos de los clientes para identificar las variables relevantes del comportamiento de compra y del mercado de destino. Donde se obtenga la frecuencia, precio y tipos de productos adquiridos.
- Aplicar técnicas de segmentación mediante modelos estadísticos o de machine learning (RFM y técnicas de clústering) para clasificar a los grupos de consumidores en grupos homogéneos según su comportamiento de compra.
- Evaluar y validar los modelos generados mediante métricas cuantitativas de desempeño (por ejemplo, silhouette score, Davies-Bouldin index, o precisión en clasificación), con el fin de garantizar la solidez de los resultados.
- Identificar los patrones de consumo dentro de cada segmento generados con el fin de entender sus preferencias, necesidades y hábitos de compra.

2. Revisión de Literatura

El presente capítulo aborda el marco teórico y la revisión de literatura que sustenta el análisis propuesto. Se examinan los antecedentes científicos y técnicos sobre la segmentación de clientes, las metodologías de ciencia de datos y los algoritmos de agrupamiento aplicados al sector florícola. Asimismo, se incorporan estudios nacionales e internacionales que evidencian el uso de técnicas como K-Means o DBSCAN en la segmentación de mercados y la predicción del comportamiento de compra.

2.1. Estado del arte

Diversas investigaciones han demostrado la efectividad de las técnicas de aprendizaje no supervisado en la segmentación de clientes. Villacrés (2023) aplicó los algoritmos K-Means y DBSCAN en una cadena de farmacias ecuatoriana, logrando identificar patrones de compra y validar la calidad del agrupamiento mediante el índice de Silhouette. Por su parte, Cáceres (2019) evaluó K-Means y K-Medoids con métricas de Davies–Bouldin, concluyendo que el modelo K-Means presentó una mejor consistencia interna para segmentar consumidores. A nivel internacional, investigaciones como las de Catota et al. (2023) y Torroba (2020) muestran el uso de modelos de clústering en agricultura de precisión y floricultura, donde la estacionalidad y la logística influyen significativamente en la demanda.

Las empresas que han usado RFM exitosamente es Amazon utiliza RFM y análisis de comportamiento para recomendar productos y promociones personalizadas, aumentando la frecuencia de compra y el ticket promedio.

2.2. Marco teórico

2.2.1. Industria de la floricultura

La floricultura ecuatoriana es uno de los principales sectores de exportación no petroleros, generando divisas, empleo y encadenamientos productivos. De acuerdo con

PROECUADOR (2022) indicó que Ecuador se posiciona como el tercer exportador mundial de flores, principalmente rosas. Los mercados de destino más importantes son Estados Unidos y Canadá, caracterizados por una demanda estacional en fechas como San Valentín y Día de la Madre. La integración de analítica de datos en este sector permite optimizar estrategias comerciales, prever picos de demanda y diseñar campañas promocionales alineadas con la estacionalidad (Catota et al., 2023).

En este contexto, la analítica de datos ha emergido como una herramienta fundamental para optimizar la competitividad del sector. Su aplicación permite predecir la demanda, diseñar campañas de marketing basadas en el comportamiento del consumidor, mejorar la eficiencia logística y reducir pérdidas económicas asociadas a sobreproducción o problemas de conservación, la floricultura ecuatoriana enfrenta desafíos como la necesidad de diversificar mercados internacionales, implementar prácticas sostenibles que cumplan estándares globales y aprovechar al máximo las oportunidades que ofrecen la innovación tecnológica y la inteligencia de datos. Estas acciones son esenciales para mantener la competitividad del país en un mercado global altamente demandante y estacional.

2.2.2. E-commerce

El comercio electrónico ha transformado la dinámica de los mercados tradicionales, permitiendo a los productores florícolas reducir intermediarios y vender directamente al consumidor final. Kotler, P., y Keller, K. L. (2016) destacan que la digitalización de los procesos comerciales exige sistemas de gestión de clientes basados en datos permitiendo llegar a clientes de diferentes países, rompiendo barreras geográficas donde su automatización de procesos lleva la trazabilidad de los pedidos.

La integración del E-commerce en la floricultura ecuatoriana no solo fortalece la competitividad del sector, sino que también ofrece un canal innovador para diversificar mercados y aumentar ingresos. Combinado con la analítica de datos, puede transformar la

forma en que los productores planifican cultivos, gestionar inventarios y diseñan campañas comerciales, potenciando la eficiencia y sostenibilidad del sector.

2.2.3. Segmentación de mercado

La segmentación de clientes en e-commerce permite identificar grupos con comportamientos homogéneos y diseñar estrategias personalizadas de marketing relacional y fidelización (Berry, 2002).

2.2.4. Método RFM

El modelo RFM (Recencia, Frecuencia y Valor Monetario) constituye una técnica ampliamente utilizada para evaluar el valor de los clientes. Según Fader et al. (2005), estas tres variables permiten clasificar a los consumidores según su frecuencia de compra y contribución económica, así como la toma de decisiones basadas en datos concretos sobre quién es más valioso para la empresa. En el contexto de la floricultura, esta metodología posibilita identificar clientes estratégicos con alta recurrencia durante fechas estacionales a través de tres dimensiones fundamentales. En primer lugar, la recencia para identificar clientes que han comprado, por ejemplo, aquellos que adquirieron flores en la última temporada de San Valentín o Día de la Madre. Complementariamente, la frecuencia para detectar clientes que compran con regularidad, como quienes envían arreglos florales cada mes o en fechas especiales recurrentes. Finalmente, el valor monetario se orienta a reconocer a los clientes que generan mayores ingresos, como empresas o clientes que compran grandes cantidades para eventos o regalos corporativos.

Se pueden mencionar tres beneficios clave en la industria de la floricultura: enviar promociones personalizadas antes de fechas clave a los clientes más frecuentes y valiosos, recompensar a clientes de alto valor con descuentos, programas de lealtad o servicios exclusivos, y prever la demanda basada en patrones de compra de los clientes más valiosos.

2.2.5. Aprendizaje no supervisado y análisis de clústeres (clústering)

El aprendizaje no supervisado busca descubrir patrones ocultos en grandes volúmenes de datos sin etiquetas predefinidas. Han et al. (2012) señalan que el análisis de clústeres es una de las técnicas más empleadas para la segmentación, permitiendo agrupar objetos con características similares. Entre los algoritmos más relevantes se encuentran K-Means y DBSCAN, ambos aplicados en proyectos ecuatorianos recientes (Villacrés, 2023; Cáceres, 2019).

De igual manera es importante mencionar la diferencia entre los algoritmos de clústering, basado en densidad, o distancia, estos enfoques son claves al momento de aplicarlos.

El clústering basado en densidad, como explica Sanchez (2024), apoyado en la terminología empleada por Sander (2011), se fundamenta en que, dado un conjunto de puntos de datos, se define una estructura que refleja con precisión la densidad subyacente, entendiendo por densidad como el número de puntos dentro de un radio específico, el algoritmo que más destaca en este concepto es DBSCAN.

Por otro lado, el clústering basado en distancia, tiene como objetivo encontrar grupos de objetos o puntos de datos, en función de la distancia y un punto central, en este aspecto tenemos los algoritmos: K-Means, K-Medoids.

2.2.5.1. Distancia euclidiana.

La distancia euclidiana dentro de un clúster es una medida que indica que tan lejos se encuentran dos puntos entre sí en un espacio multidimensional de esta forma nos permite medir similitudes. Este enfoque facilita la transformación de cada cliente en un vector numérico tridimensional, representado como:

$$Cliente_i = (R_i, F_i, M_i) \quad (1)$$

La distancia euclidiana es la distancia recta entre dos puntos, convierte la base de clientes en puntos dentro de un espacio tridimensional, es posible aplicar técnicas de minería de datos como K-means, cuyo propósito es agrupar clientes que exhiben comportamientos similares. Para determinar qué tan parecidos o disímiles son los clientes entre sí, el algoritmo utiliza la distancia euclidiana, una métrica que calcula la separación geométrica entre dos puntos en el espacio:

$$d(Cliente_{\mathrm{i}}, Cliente_{\mathrm{i}}) = \sqrt{(Ri - Rj)^2 + (Fi - Fj)^2 + (Mi - Mj)^2} \qquad (2)$$

Dentro del algoritmo de **K-means** se usa para asignar un punto al clúster más cercano dentro del centroide cuya distancia sea menor, además calcula los centroides (punto promedio) y recalcula interactivamente con clústeres con menor distancia que los hace más compactos y homogéneos. K-means trata de organizar los puntos de manera que cada grupo sea compacto y los puntos estén lo más cerca posible del centro.

La aplicación conjunta del análisis RFM y K-means resulta especialmente útil en industrias con fuerte componente estacional, como la floricultura. En este sector, permite identificar: clientes con alta recurrencia en fechas especiales (p. ej., San Valentín, Día de la Madre), clientes ocasionales de alto valor utilizados para eventos o compras corporativas, y finalmente clientes inactivos o esporádicos que requieren estrategias de reactivación.

De esta manera, el uso de distancia euclidiana dentro del modelo RFM permite construir grupos homogéneos basados en evidencia cuantitativa, optimizando el diseño de estrategias comerciales, de fidelización y gestión del inventario. La técnica proporciona una segmentación robusta, reproducible y alineada con principios estadísticos aplicados al comportamiento del consumidor.

2.2.5.2. Algoritmo K-Means.

K-Means divide el conjunto de datos en grupos. El funcionamiento en el que se basa es el cálculo de centroides, en un inicio punto aleatorios dentro del conjunto de datos, marcados por el valor de k, posteriormente emplea la medida euclidiana para determinar la distancia de cada punto a estos centroides, y con esto asignar al que pertenecen, el proceso se repite dentro de cada clúster para recalcular la posición del centroide utilizando promedio de los elementos dentro del clúster, esto es un proceso recurrente hasta lograr el máximo de iteraciones o los centroides converjan (Caicedo 2022). Su eficiencia lo convierte en un método preferido para la segmentación de clientes en e-commerce, aunque requiere definir previamente el número de clústeres. Villacrés (2023) reportó una precisión del 95,2% en la identificación de segmentos utilizando este algoritmo en el sector farmacéutico.

2.2.5.3. Algoritmo K-Medoids.

K-Medoids, como explica Caicedo (2022), nace como una generalización de K-means, su principal diferencia es el uso de medoide como alternativa al centroide, la introducción de este término hace uso de un dato real dentro del conjunto, deja de un lado los promedios empleados en K-means, esto hace que sea más resistente a datos atípicos, porque el valor del medoide no se verá afectado por promedio. Además, el hecho de que use un punto real como medioide hace que su interpretación sea más clara.

2.2.5.4. Algoritmos DBSCAN.

DBSCAN (Density-Based Spatial Clústering of Applications with Noise) agrupa puntos basándose en densidad, permitiendo identificar patrones de comportamiento incluso en presencia de ruido o outliers. Cáceres (2019) y Villacrés (2023) resaltan su utilidad para reconocer clientes atípicos con alto valor de compra.

2.2.5.5. Métricas de validación.

Las métricas de validación interna son fundamentales para evaluar la calidad de los agrupamientos. El coeficiente de Silhouette (Rousseeuw, 1987) mide la cohesión y separación entre clústeres, mientras que el índice de Davies–Bouldin (Davies and Bouldin, 1979) evalúa la compactación y distancia relativa entre ellos. Ambas métricas han sido aplicadas exitosamente en estudios ecuatorianos de segmentación de clientes (Cáceres, 2019; Villacrés, 2023).

2.2.5.6. Método del codo.

El método del codo es una técnica que se utiliza en el análisis de datos con el fin de determinar el número óptimo de clústeres **(K)**, este método es usado para encontrar el grupo que mejor se ajuste a los datos sin sobre ajustar. Su objetivo es identificar el punto a partir del cual se pueda añadir más clústeres en la calidad de la segmentación.

El método del codo usa la métrica SSE (Sum of Squared Errors) o inertia que se define en rendimientos decrecientes a medida que añade más clústeres, cada nuevo clúster explica cada vez menos variación adicional funciona minimizando la suma total de cuadrados dentro del clúster (Within-Clúster Sum of Squares, WCSS), definida como:

$$SSE = \sum_{i=1}^{k} \quad \sum_{x \in Ci} \quad \left|\left|x - \mu_i\right|\right|^2 \tag{3}$$

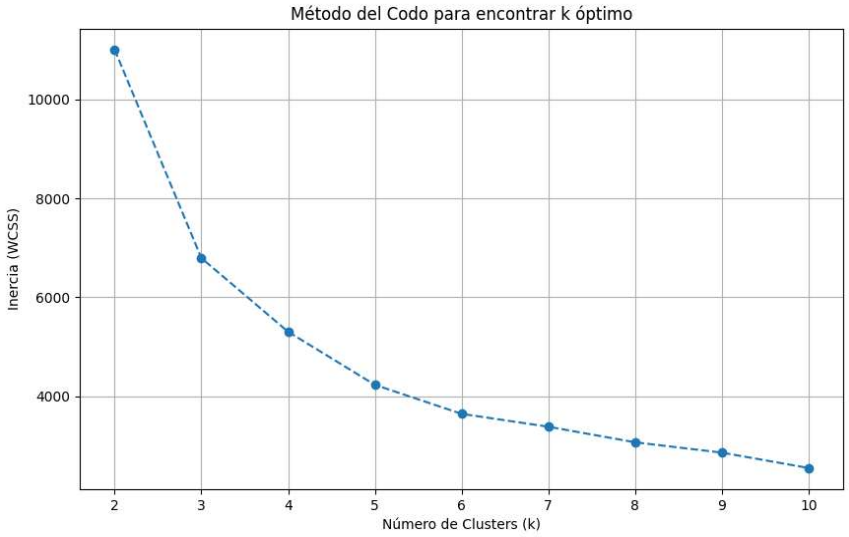
Variables:

- Ci= clúster i
- μ_i= centroide del clúster
- $|| x - \mu_i ||^2$= distancia cuadrática de cada punto a su centroide

Cuando más grande es el número de clúster, la varianza tiende a disminuir. Cuanto menor es la distancia entre clústeres, los clústeres son más compactos y por ende se maximiza la varianza, donde el método busca el valor de K que sea óptimo con un incremento.

Figura 1

Método del codo para encontrar k óptimo



Nota. El gráfico muestra la relación entre el número de clústeres (k) y la inercia (WSS). El punto de inflexión alrededor de k = 3 sugiere el número óptimo de clústeres para el análisis dado que incrementos posteriores de K proporcionan mejoras marginales en la reducción del WCSS.

2.2.6. *Metodología CRISP-DM*

El modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) propuesto por Chapman et al. (2000) estructura los proyectos de ciencia de datos en seis fases: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y despliegue. Este marco metodológico ha sido adoptado ampliamente en proyectos de segmentación en Ecuador (Villacrés, 2023; Cáceres, 2019), garantizando replicabilidad y trazabilidad en el análisis.

2.2.7. *Python*

Python es un lenguaje de programación interpretado, de alto nivel y orientado a objetos, ampliamente utilizado en ciencia de datos por su sintaxis sencilla, su gran

comunidad, y porque ofrece un conjunto de herramientas unificadas que facilitan el flujo de trabajo en el análisis científico (VanderPlas, 2017, p. 5), posee librerías especializadas (NumPy, Pandas, Scikit-learn, entre otras), esto lo convierten en una herramienta esencial en el desarrollo de algoritmos de aprendizaje automático. En este proyecto, Python constituye el entorno principal para el procesamiento, transformación y modelado analítico de los datos.

2.2.8. *Sklearn*

Scikit-learn es una biblioteca de Python diseñada para implementar algoritmos de aprendizaje automático, tanto supervisado como no supervisado. Esta librería ofrece algoritmos avanzados para clasificación, regresión, clústering, preprocesamiento, métricas de evaluación y selección de modelos dentro de una API estandarizada. En este proyecto, Sklearn se emplea para ejecutar los algoritmos de clústering K-Means y DBSCAN, así como para la normalización y evaluación de modelos.

2.2.9. *Pandas*

Según McKinney (2018, p. 4), la librería Pandas se ha convertido en el estándar de facto para la manipulación de datos estructurados en Python, proporciona estructuras como DataFrame y Series, que permiten trabajar de manera eficiente con datos de forma tabulares, facilitando tareas de limpieza, depuración, transformación, agrupación y combinación de datos. En este proyecto, Pandas se emplea para la carga de datos, limpieza inicial y construcción de las variables del modelo RFM.

2.2.10. *Power BI*

Power BI es una herramienta de inteligencia de negocios desarrollada por Microsoft que permite transformar datos en visualizaciones interactivas, dashboards y reportes dinámicos orientados a la toma de decisiones. En el presente proyecto, Power BI se utiliza para visualizar los resultados del clústering e interpretar perfiles de clientes.

2.2.11. *Google Collaboratory*

Google Collaboratory es un entorno de notebooks en la nube que permite ejecutar código Python sin necesidad de instalación local. Ofrece integración con Google Drive, acceso a CPU, GPU y TPU, además de un entorno colaborativo ideal para equipos de análisis de datos y machine learning. En este proyecto, Google Colab se utilizó como entorno principal para ejecutar el pipeline analítico.

3. Desarrollo

3.1. Descripción de la base de datos

La base de datos utilizada en este estudio proviene del sistema transaccional de ventas de la empresa Ecuador Direct Roses (EDR) e incluye información histórica relacionada con órdenes de compra, clientes y productos. La selección del dataset responde a dos criterios fundamentales: relevancia para el objetivo de análisis.

Torroba (2023) enfatiza que un dataset es adecuado para segmentación siempre que integre información transaccional detallada, identidades de clientes y atributos temporales que permitan construir indicadores derivados como frecuencia, inversión o recencia. De forma similar, Villacrés (2023) señala que la elección del dataset debe reflejar el comportamiento real del cliente y contener la granularidad suficiente para aplicar algoritmos de clústering y análisis descriptivo robusto.

El dataset seleccionado posee estas características, pues incluye 18 variables:

- Relacionadas con clientes
  - cliente_id: Código del cliente en la base de datos.
  - cliente: Nombre del Cliente que realizó la compra.
  - ciudad: Ciudad de residencia del cliente que realiza la compra.
  - estado: Estado de residencia del cliente que realiza la compra.
  - país: País de residencia del cliente que realiza la compra.
  - origen_cliente: Indica como se le contactó al cliente por primera vez.
- Operaciones
  - numero_orden: Número de orden que realizó un cliente.
  - fecha_vuelo: Fecha para la que se solicitaron la entrega del producto.
  - agencia: Indica la agencia de carga se enviaron los productos.
- Productos

  - producto_id: Código del producto que se vendió.
  - producto: Nombre de la variedad que compró el cliente.
  - tallos: Número de tallos que compró un cliente.
  - largo: El largo del tallo del producto que se vendió.
- Economía
  - precio_unitario: El precio por tallos que se vendio
  - total: La multiplicación del precio unitario por el número de tallos vendidos
- Por gestión comercial
  - estado_orden: Indica el estado de las órdenes.

    1 = Pending Payment

    2 = Processing

    3 = Shipped

    4 = Cancelled

    5 = On Hold

    6 = Refunded
  - vendedor: nombre corto del vendedor que realizó la venta.
  - usuario_id; código del usuario que creó la orden de venta.

Además, Cáceres (2019) destaca que datasets provenientes de sistemas institucionales o empresariales garantizan integridad estructural y consistencia, elementos esenciales para la fase de modelado. Anexo 1. Limpieza, preprocesamiento y transformación de datos.

Siguiendo la metodología CRISP-DM, esta etapa corresponde a la fase de Preparación de los Datos, considerada por Villacrés (2023) como el punto crítico que determina la calidad final del modelo, incluso por encima de la selección del algoritmo. Torroba (2023) coincide en que el preprocesamiento debe integrar procedimientos de depuración, imputación,

estandarización y derivación de variables para asegurar que el dataset final represente fielmente el fenómeno a modelar.

3.1.1. Creación de API de conexión y extracción de datos desde la base de datos

Se crea una API de conexión a la base de datos que realiza una solicitud de autenticación al servidor mediante protocolo HTTP POST, enviando las credenciales de acceso en el cuerpo de la petición.

Con la conexión realizada correctamente se extrae la respuesta en formato JSON mediante un token de acceso, este token funciona como un mecanismo de autenticación basado en sesiones.

La extracción de los datos se genera mediante un servicio web de tipo API REST, esta API proporciona información relacionada con registros de ventas, la cual se encuentra protegida mediante un mecanismo de autenticación basado en tokens (Bearer Token) el cual realiza consultas de datos de ventas por año y con los mismos se construye un dataset consolidado.

3.1.2. Limpieza de datos

3.1.2.1. Eliminación de registros.

Durante la revisión inicial del dataset se identificaron transacciones que no representan comportamiento real de compra, sino movimientos administrativos propios del sistema comercial. Entre estas se encuentran las categorías “DESCUENTOS VENTAS”, “CRÉDITOS EN VENTAS”, “MUESTRA”, así como líneas asociadas a productos reales, pero con tallos negativos, precio unitario negativo o montos totales negativos. Estas observaciones corresponden a devoluciones, notas de crédito, reversos de inventario, correcciones internas o entrega de muestras, ninguna de las cuales se interpreta como una venta efectiva.

Se tomaron solo las órdenes que se encuentran en estado igual a 3, que son las órdenes procesadas.

Mantener estas transacciones dentro del dataset analítico distorsiona métricas esenciales (como recencia, frecuencia y valor monetario) y puede generar clientes con consumo negativo o artificialmente reducido. Siguiendo el criterio metodológico señalado por Villacrés (2023), este tipo de operaciones debe ser excluido del dataset preparado para clústering, ya que introduce ruido y sesgos en los algoritmos de agrupamiento.

Por ello se eliminaron todas las filas asociadas a estas categorías, así como cualquier registro con valores negativos en tallos, precio_unitario o total. Estas observaciones se conservaron únicamente en un dataset auxiliar destinado a análisis contable, mientras que el dataset final utilizado en el modelado contiene exclusivamente ventas efectivas, garantizando consistencia estadística y validez en la segmentación.

3.1.2.2. Estandarización de formatos.

Con el fin de asegurar uniformidad y evitar discrepancias en el procesamiento posterior, se convirtió la variable fecha_vuelo al tipo datetime, se normalizaron las variables categóricas cliente, ciudad, estado, país, vendedor, producto, agencia y origen_cliente, transformándose a texto limpio con formato homogéneo (mayúsculas y sin espacios adicionales), y finalmente se comprobaron los tipos de datos del resto de variables para asegurar consistencia en el tratamiento numérico.

3.1.2.3. Tratamiento de valores faltantes.

Como parte de la depuración las variables críticas (cliente_id, producto_id, tallos, total, fecha_vuelo) se depuraron para garantizar 0% de nulos, eliminando cualquier registro incompleto que comprometa el cálculo de RFM o el modelado. Además, las variables secundarias (usuario_id, vendedor, agencia, origen_cliente) se imputaron con la categoría

"DESCONOCIDO" en los casos donde presentaban valores faltantes, evitando la pérdida innecesaria de registros válidos.

3.2. Descripción arquitectura

La arquitectura propuesta para el desarrollo del presente proyecto se fundamenta en un pipeline modular de ciencia de datos, diseñado para asegurar la trazabilidad, reproducibilidad y escalabilidad del proceso analítico. Este enfoque se alinea con los modelos contemporáneos de arquitectura de datos descritos en la literatura especializada (Biswas et al. 2021; Raja, 2025; Krantz, T., y Jonker, A., 2024), donde la analítica se estructura en capas funcionales que permiten separar responsabilidades y optimizar el flujo de información desde la adquisición hasta el consumo de resultados.

Siguiendo estas recomendaciones, la arquitectura empleada se compone de siete capas principales: adquisición de datos, almacenamiento, preprocesamiento, ingeniería de características, modelado analítico, validación y visualización. Esta estructura refleja buenas prácticas adoptadas en pipelines modernos y garantiza la consistencia metodológica para proyectos de segmentación mediante aprendizaje automático no supervisado.

Además, está alineada con la propuesta de Yen, L. (2025), que identifica estas mismas etapas como componentes esenciales de cualquier arquitectura de datos orientada a machine learning y analítica avanzada.

Capa de adquisición y fuentes de datos, comprende la recolección de los datos transaccionales provenientes del sistema de e-commerce de la empresa la cual está en un base de datos Sql Server. De acuerdo con Krantz, T., y Jonker, A. (2024), la arquitectura de datos inicia con una capa de ingesta que debe garantizar integridad y disponibilidad. En este proyecto, los datos se obtuvieron mediante exportaciones estructuradas en formato CSV, lo que permite un fácil manejo y su incorporación en etapas posteriores del pipeline.

Capa de almacenamiento y entorno de ejecución, siguiendo las recomendaciones de Raja (2025) sobre arquitecturas resilientes, el almacenamiento inicial se realizó una base de datos Postgresql, a la cual se accede mediante APIS consumiendolos con el lenguaje Python integrándose con el entorno de ejecución en Google Colab. Este diseño permite un procesamiento distribuido, reproducible y basado en herramientas abiertas, acorde a los principios de escalabilidad mencionados por Biswas et al. (2021).

Capa de preprocesamiento y limpieza de datos, esta capa se centra en la depuración y preparación inicial del dataset. Las tareas asociadas a esta etapa se detallan en el apartado 3.1.1 de este documento.

Capa de ingeniería de características, en esta fase se construyó el modelo RFM (Recency, Frequency, Monetary) el desarrollo completo del modelo RFM se presenta en el apartado 3.3.1.

Capa de modelado analítico, se aplicaron algoritmos como K-Means y DBSCAN, seleccionados por su capacidad para revelar patrones de agrupamiento no supervisado. Los métodos utilizados se detallan en el apartado 3.4.1.

Capa de validación y evaluación, cuyo fin es el de evaluar la calidad de los clústeres se emplearon Silhouette Score (apartado 3.4.2.2) y el método del codo (apartado 3.4.2.1). Estas métricas permiten incorporar validación cuantitativa en la arquitectura y garantizan la confiabilidad del pipeline analítico.

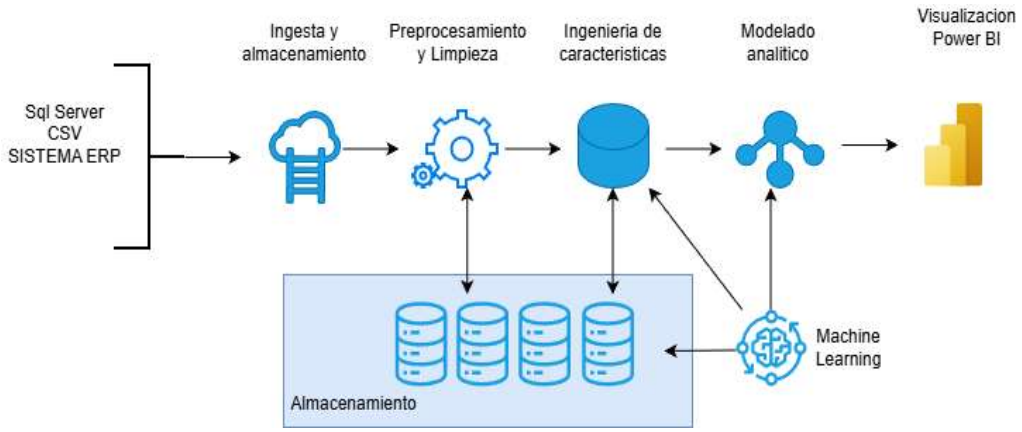
Capa de visualización y consumo de resultados, fueron almacenados en la misma base de datos de Postgres mediante una importación desde un archivo csv con esto logramos estar integrados en Power BI, cumpliendo la función de capa de presentación o Business Intelligence. Según Krantz, T., y Jonker, A. (2024) la arquitectura de datos debe finalizar con mecanismos de acceso e interpretación que permitan a los responsables de negocio tomar decisiones basadas en evidencia. El dashboard desarrollado permite: Visualizar la

composición de los segmentos, analizar variables clave de comportamiento, facilitar la toma de decisiones comerciales orientadas a clientes, y la visión integral de la arquitectura reflejada.

La arquitectura completa puede resumirse conceptualmente como un pipeline lineal y modular: Fuente de datos → Ingesta → Limpieza → Ingeniería de características → Clústering → Validación → Visualización.

Figura 2

Arquitectura del Flujo del proceso de ingesta y almacenamiento del modelado analítico



Nota. La figura muestra el proceso completo desde la obtención de datos (SQL Server, CSV y ERP), seguido por las etapas de ingesta, limpieza, ingeniería de características y modelado analítico, hasta su visualización final en Power BI. El diagrama ilustra la interacción entre los módulos y la fase de almacenamiento central que apoya el trabajo de machine learning.

3.3. Modelado de datos

3.3.1. Análisis RFM

Con el dataset limpio se generaron nuevas variables derivadas, destinadas a capturar patrones de comportamiento del cliente, siguiendo las recomendaciones de Torroba (2023) método RFM y segmentación por valor (Villacrés 2023):

- Recencia: número de días desde la última compra registrada.
- Frecuencia: número de órdenes emitidas por cada cliente durante el periodo de análisis.
- Monetary: suma total gastada por cliente.
- Estacionalidad: compras realizadas en fechas especiales (San Valentín, Día de la madre)

Con el fin de aplicar estas características, se definió las siguientes reglas a nivel de cada uno de los clientes:

La recencia, toma como referencia la fecha de la última transacción del dataset, más un día y menos la fecha (fecha vuelo) más reciente entre las transacciones del cliente.

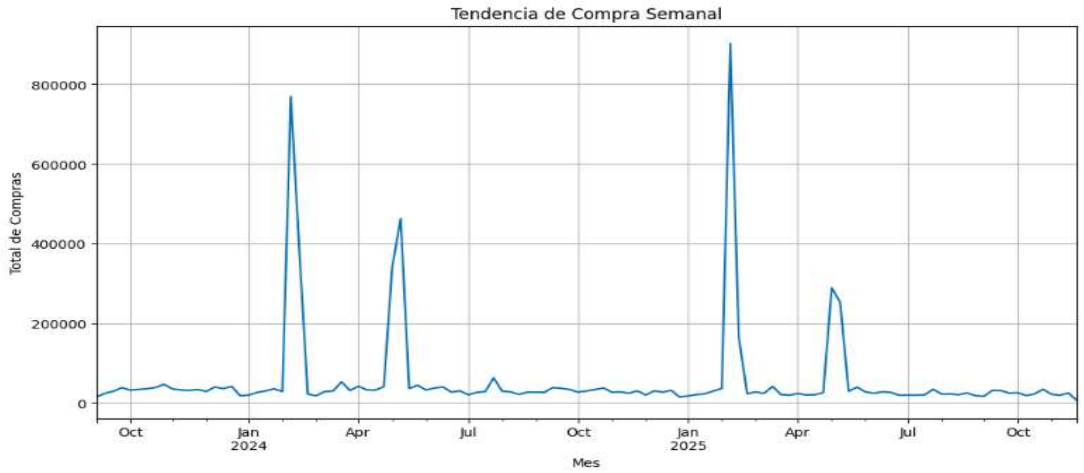
Para la frecuencia se sumó el número de pedidos por cliente, para ello se empleó la columna (numero_orden) y una sumatoria de sus valores únicos en cada cliente.

Respecto al gasto monetario, está dado por la suma del total de cada registro de ventas presente en el dataset, que se encuentra en la variable (total).

Finalmente, la variable estacionalidad, emplea una característica adicional (semana), que nace a partir de la fecha del pedido (fecha_vuelo), hace referencia al número de semana en el año en la que se realizó la compra. Con este dato y conociendo que las semanas 5, 6, 7 (San Valentín) y 18, 19 (Día de la madre), son semanas pico dentro del mercado florícola, consideradas como fechas especiales donde la concurrencia y cantidad de compras aumenta, se realizó un análisis de tendencia de compra semanal ilustrado en la Figura 4, el mismo que evidenció los picos de ventas durante estas semanas, lo cual sustenta el componente estacional en el mercado.

Figura 3

Tendencia de compra semanal a lo largo de los años



Nota. Gráfico muestra la variación semanal del total de compras entre los años 2024 y 2025.

El cálculo y análisis de la variable (matriz_semanas) se orientó a evaluar la concentración de compras en las denominadas semanas_pico. Estas semanas corresponden a pedidos del año en las cuales se observa un aumento significativo de compra, y crea una matriz de clientes por semanas donde cada fila es un cliente y cada columna contiene el número de órdenes que el cliente hizo en esa semana, se calcula las compras en semana pico con la siguiente fórmula donde *i* es cada cliente:

$$compras\ pico_i = \sum_{\beta \in semanas\ pico} \quad ordenes_{u,8} \tag{4}$$

Calcula compras totales:

$$compras\ totales_i = \sum_{s=1}^{52} \quad ordenes_{i,s} \tag{5}$$

Calcula el peso de semanas pico:

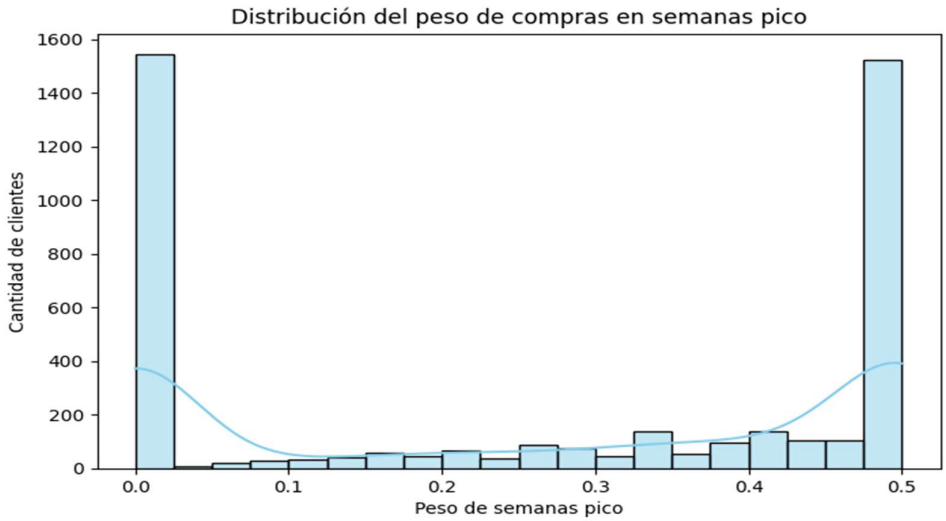
$$peso\ picos_i = \begin{cases} \frac{compras\ pico}{compras\ totales_i} & si\ compras\ totales_i \gg 0 \\ 0, & si\ compras\ totales_i = 0 \end{cases} \tag{6}$$

- **1.0** → cliente compra *solo* en semanas pico

- **0.5** → la mitad de sus compras son en semanas pico
- **0.0** → compra *nunca* en semanas pico

Figura 4

Tendencia de compra semanas pico



Nota. El histograma representa la distribución del peso de compras durante las semanas pico, mostrando concentraciones elevadas en los valores extremos y una dispersión menor en la zona central.

En la Figura 4 se observa la distribución del peso de compras correspondiente a las semanas pico. El gráfico revela que la mayor parte de los clientes se concentra en los valores extremos del indicador, específicamente cerca de 0 y 0.5, lo que indica patrones de compra muy bajos o muy elevados durante dichos periodos. En contraste, los valores intermedios presentan una menor frecuencia, lo que sugiere comportamientos menos comunes en torno a niveles de compra moderados. El cálculo permitió identificar clientes que compran principalmente en semanas pico, medir el comportamiento estacional y esto facilitó la segmentación según su dependencia a eventos temporales.

La combinación del modelo RFM con el análisis de estacionalidad proporciona una visión más completa del comportamiento del cliente. Mientras que RFM permite segmentar

clientes según su valor y actividad, el componente temporal de estacionalidad permite identificar clientes que concentran sus compras en periodos específicos, clientes cuya actividad es constante a lo largo del año, comportamiento cíclico útil para predicciones y campañas.

Una vez calculadas las variables RFM y estacionalidad, se procede a realizar el análisis estadístico, enfocado en los resultados obtenidos y que se visualizan en la Tabla 1.

Tabla 1

Análisis estadístico de RFM estacional

Estadístico descriptivo	Recencia	Frecuencia	Monetario	Estacional
count	4250.0000	4250.0000	4250.0000	4250.0000
mean	377.1845	4.2826	1565.6073	0.3513
std	223.1778	12.3720	5691.8175	0.4774
min	1.0000	1.0000	52.0000	0.0000
25%	199.0000	1.0000	175.0000	0.0000
50%	305.0000	1.0000	432.0000	0.0000
75%	564.0000	3.0000	1154.7500	1.0000
max	815.0000	244.0000	184497.0000	1.0000

La recencia, revela un dato crudo, donde la media de 377 días sugiere que la empresa tiene una mayor cantidad de clientes que han comprado hace más de un año, incluso el 25% o superior de clientes del cuartil uno, dos, y tres (Q1, Q2, Q3) no han comprado en más de 6 meses, y la mediana que corresponde al cuartil dos (Q2) se encuentra cercano a 1 año, podría tener una explicación en la estacionalidad del mercado, pero es un dato llamativo para el análisis.

La frecuencia, si bien existe una media de 4, en la mediana representada en el cuartil dos (Q2) se aprecia que es igual a 1 lo que nos indica que al menos el 50% de los clientes han comprado una única vez, pero un dato importante aquí es que existe un cliente con frecuencia 244 (max) el mismo que marca una diferencia enorme con la mediana, de esta medida se puede deducir la presencia de dos tipos de clientes: consumidor final, y mayorista.

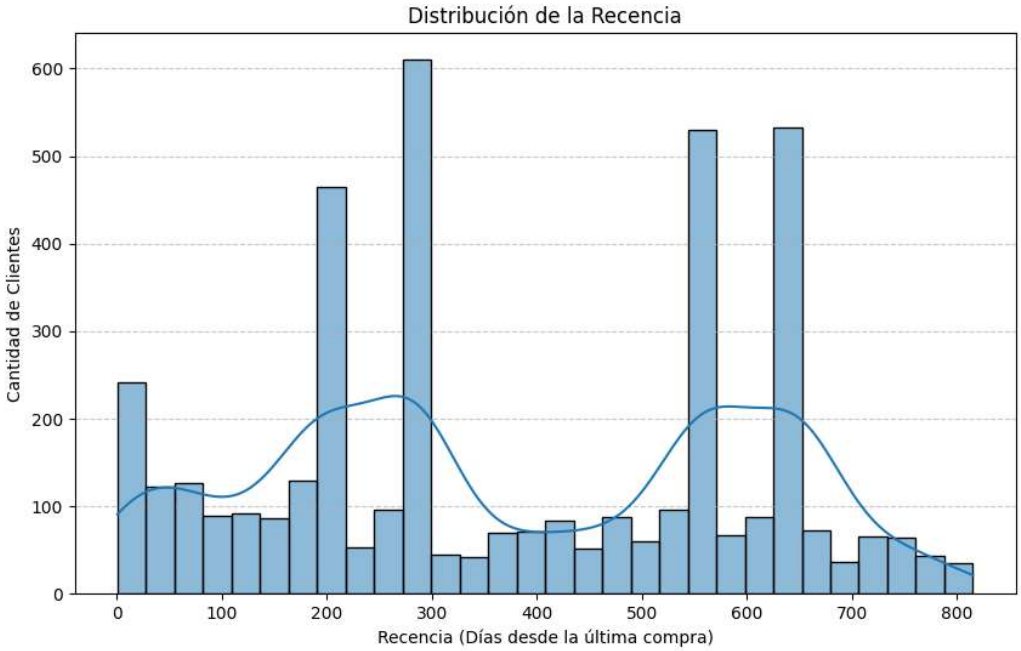
El valor monetario sigue siendo igual de revelador que las características anteriores, tenemos que nuestra mediana ubicada en el cuartil dos (Q2) el 50% de los clientes han gastado 432 dólares. Pero de igual forma apreciamos que nuestro valor máximo es 184497 dólares, lo cual confirma un segmento de clientes “de alto valor” con gastos enormes respecto a la mediana.

La estacionalidad nos marca una tendencia en nuestro cuartil 3 (Q3) hacia ser marcadas como compras en fechas especiales, nuestra cantidad de registros ubicados en el 75% son compras en fechas aleatorias.

Las variables RFM no tienden a seguir una distribución normal lo cual en este tipo de variables es común según explican Schmittlein et al. (1987) y Fader et al. (2005), esto debido a que las transacciones de comportamiento humano en este caso compras suelen seguir otro tipo de distribuciones: Binomiales Negativas (Frecuencia) y Gamma/Lognormales (Monto Monetario) caracterizadas por colas pesadas, es decir donde los eventos extremos o valores atípicos son más probables que en una distribución normal.

Figura 5

Distribución de la recencia



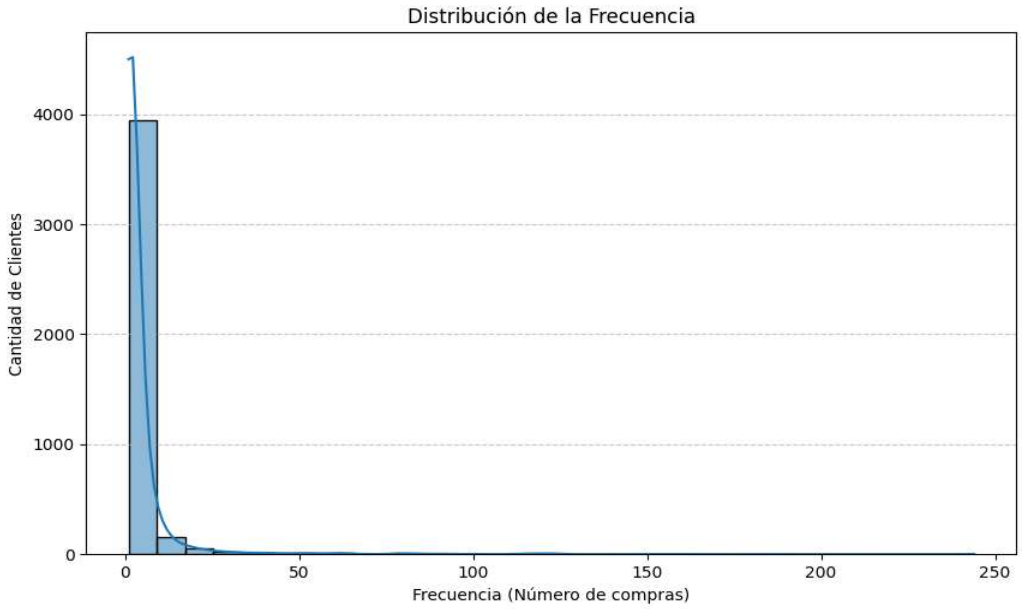
Nota. La figura muestra la distribución de la recencia, expresada en días desde la última compra. El histograma permite identificar patrones de inactividad entre los clientes y picos de frecuencia asociados a distintos periodos de tiempo sin compra.

En la Figura 5 se presenta la distribución de la recencia, medida como el número de días transcurridos desde la última compra de cada cliente. El histograma evidencia que existen agrupaciones en diferentes rangos de inactividad, lo cual sugiere que los clientes no siguen un patrón uniforme de compra. Se observan picos marcados alrededor de los 100, 300 y 600 días, lo que indica que ciertos segmentos tienden a presentar periodos prolongados sin realizar transacciones.

Asimismo, la dispersión en los valores refleja la heterogeneidad del comportamiento de compra dentro de la base de clientes, un aspecto relevante para posteriores análisis de segmentación, pronóstico y estrategias de retención.

Figura 6

Distribución de la frecuencia



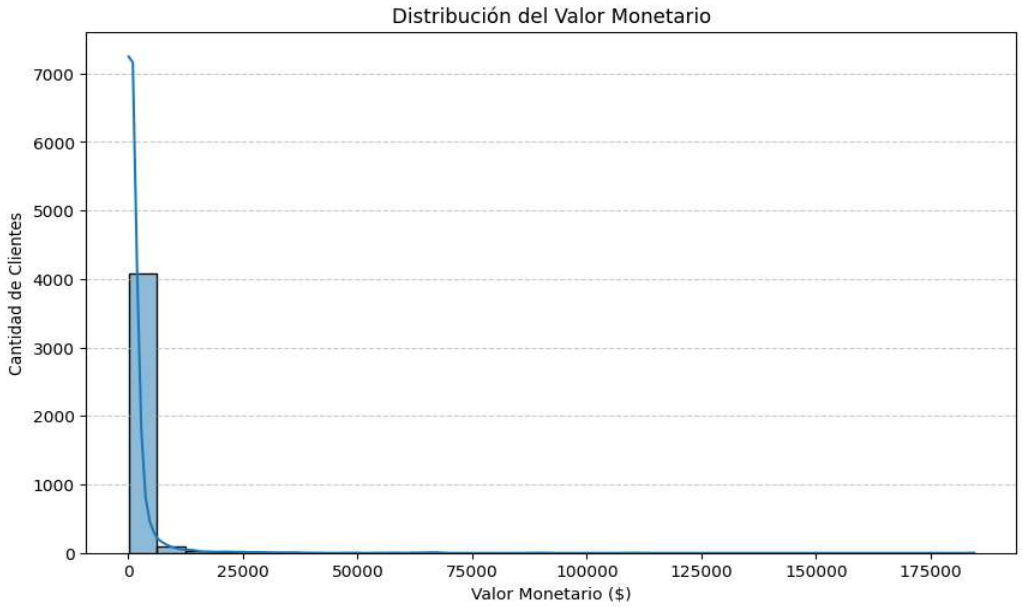
Nota. La figura muestra la distribución del número de compras realizadas por los clientes. La escala evidencia una alta concentración en valores bajos de frecuencia, seguida de una larga cola con pocos clientes de alta actividad.

En la Figura 6 se presenta la distribución de la frecuencia de compra, medida como el número total de transacciones realizadas por cada cliente. El gráfico revela una marcada concentración de clientes con niveles muy bajos de actividad, evidenciada por el gran pico en el rango de 0 a 10 compras. Esto indica que la mayor parte de los clientes compra de manera ocasional o muy esporádica.

A medida que la frecuencia aumenta, la cantidad de clientes disminuye abruptamente, configurando una distribución altamente sesgada a la derecha. La presencia de una cola larga sugiere que existe un grupo muy reducido de clientes que realiza compras frecuentes, lo cual tiene implicaciones directas en estrategias de segmentación

Figura 7

Distribución del valor monetario



Nota. La figura muestra la distribución del valor monetario asociado a los clientes. Se observa una alta concentración en valores bajos y una cola larga hacia la derecha, característica típica de variables financieras con alta dispersión.

En la Figura 7 se presenta la distribución del valor monetario, medido como el monto total gastado por cada cliente. El histograma evidencia que la mayoría de los clientes registra valores relativamente bajos de compra agregada, con una gran acumulación en el tramo inicial de la escala. Esto indica que el comportamiento dominante corresponde a clientes con niveles de gasto moderados o reducidos.

A medida que los valores monetarios aumentan, la frecuencia disminuye drásticamente, configurando una distribución fuertemente sesgada a la derecha. La presencia de una cola larga refleja la existencia de un grupo pequeño pero significativo de clientes de alto valor, cuyos montos de compra son considerablemente superiores al promedio.

La visualización de la distribución para las variables RFM, nos permiten detectar un problema que puede afectar nuestro clústering más adelante, para el caso de frecuencia en la masa más grande de clientes están agrupados en valores de 1 a 10, lo cual dista de los clientes de alto valor que en frecuencia tienen números más altos, pero en cantidad son menor. Similar problema ocurre en la distribución del valor monetario donde tenemos gran cantidad de clientes que gastan poco, y menos clientes que gastan mucho que de igual forma son opacados por el primer segmento.

Con el fin de prevenir un problema al momento de modelar el algoritmo de clústering y evitar que trate a los clientes como una masa enorme de iguales características sesgado por los valores extremos que tienen los clientes de alto valor se realiza una transformación logarítmica. John Tukey (1977), el padre del análisis de datos moderno establece que cuando los datos sufren de asimetría positiva fuerte la transformación logarítmica es la herramienta primaria para expresar los datos y conseguir simetría, esto se detalla a continuación en la fórmula:

Para una variable X:

$$Y = log(X) \text{ ó } Y = log(1 + X) \quad (7)$$

- Log1p(x)

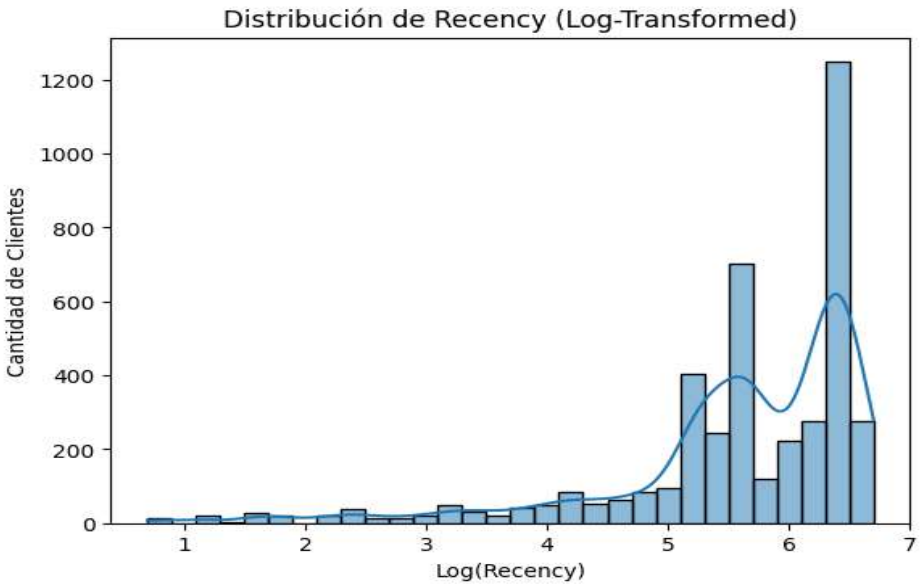
La misma que permite reducir simetría en las métricas económicas o de compras que tienen distribuciones sesgadas con muchos valores pequeño o grandes, el log aplana la distribución haciendo los datos más simétricos además de reducir el impacto en valores externos, facilita algoritmos que asumen normalidad o distancias euclidianas para mejorar las variables de transformación.

Después de haber realizado este paso importante en pro de la consistencia y ejecución del modelo, se puede apreciar en la Figura 8, Figura 9, y Figura 10. que los datos se han

expandido y tomaron una distribución normal, donde ya se puede distinguir entre valores: bajos, medios, altos.

Figura 8

Distribución de recencia, con transformación logarítmica



Nota. La figura muestra la distribución de la recencia luego de aplicar una transformación logarítmica. Esta transformación reduce la asimetría de la variable original, permitiendo observar con mayor claridad los patrones subyacentes de frecuencia en períodos de inactividad.

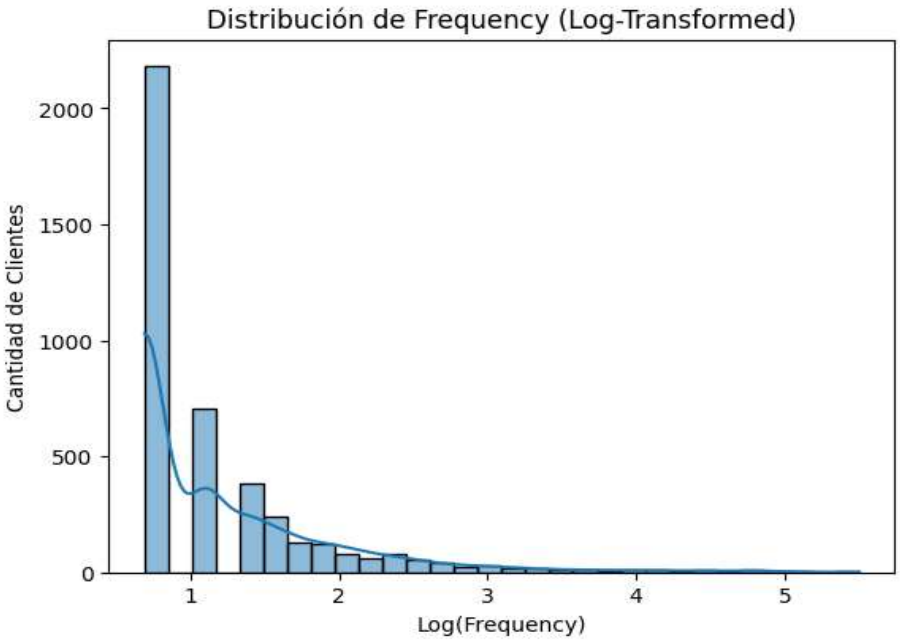
En la Figura 8 se presenta la distribución de la recencia después de aplicar una transformación logarítmica a la variable original. Esta transformación se utiliza para reducir la fuerte asimetría observada previamente y mejorar la interpretabilidad de los datos. Como consecuencia, los valores quedan distribuidos de forma más compacta, permitiendo identificar de manera más clara los rangos donde se concentran los clientes según el tiempo transcurrido desde su última compra.

El histograma resultante muestra una mayor densidad de clientes en los valores comprendidos aproximadamente entre 5 y 6 en la escala logarítmica, indicando que gran

parte de la base presenta períodos de inactividad relativamente homogéneos cuando se analizan en términos log-transformados.

Figura 9

Distribución de frecuencia con transformación logarítmica



Nota. La figura muestra la distribución de la frecuencia luego de aplicar una transformación logarítmica. Esta transformación comprime la escala y reduce la asimetría extremadamente pronunciada de la variable original, facilitando su análisis visual y estadístico.

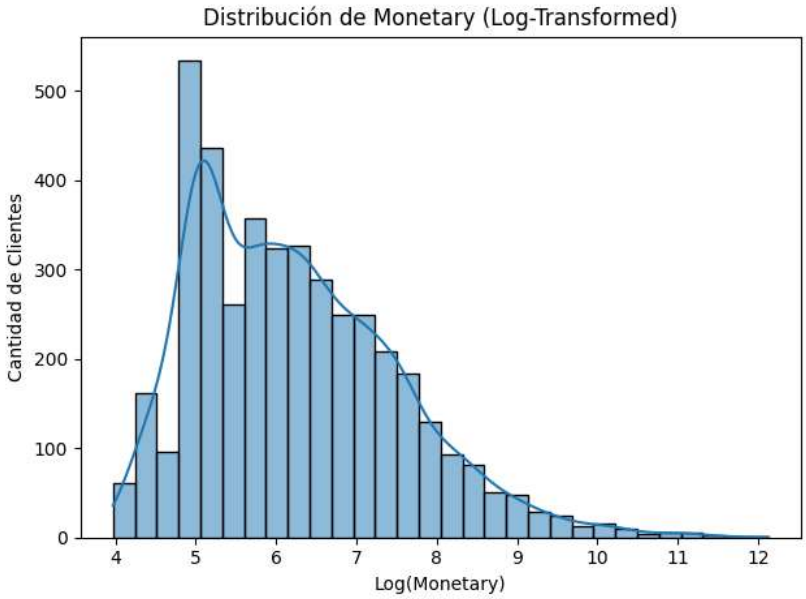
En la Figura 9 se presenta la distribución de la frecuencia de compra tras aplicar una transformación logarítmica. Este procedimiento es habitual en variables fuertemente sesgadas, como la frecuencia, ya que permite reducir la asimetría y visualizar de manera más clara la estructura de los datos.

El histograma resultante evidencia que, incluso después de la transformación, la mayor parte de los clientes continúa concentrada en los valores más bajos, ubicados cerca del

log(1). A medida que aumenta la frecuencia, la densidad decrece gradualmente, lo cual confirma la existencia de un número muy reducido de clientes con altos niveles de actividad.

Figura 10

Distribución de monetario, con transformación logarítmica



Nota. La figura presenta la distribución del valor monetario después de aplicar una transformación logarítmica, lo que permite corregir la asimetría extrema de la variable original y visualizar con mayor claridad la concentración de clientes según sus niveles de gasto.

En la Figura 10 se muestra la distribución del valor monetario a partir de la transformación logarítmica de dicha variable. Este procedimiento se utiliza con el fin de reducir la marcada asimetría observada en la distribución original, dado que el valor monetario suele presentar una concentración muy alta en montos bajos y una cola larga hacia valores elevados.

Tras la transformación, se obtiene una distribución más compacta y manejable, en la que es posible identificar patrones más claros. El histograma evidencia que gran parte de los clientes se ubica entre valores logarítmicos cercanos a 5 y 7, esto representa rangos

moderados de gasto. A partir de estos valores, la frecuencia disminuye gradualmente, marcando la presencia de un grupo reducido de clientes con niveles de gasto considerablemente más altos.

Las variables numéricas (recencia, frecuencia, monetario, estacionalidad) presentan escalas y magnitudes distintas. Para evitar que dicha disparidad sesgue el cálculo de distancias en el clústering se aplicó Standard Scaler, que transforma cada variable a una distribución con media cero y desviación estándar unitaria.

Torroba (2023) destaca que el escalado es fundamental para mejorar la estabilidad de los centroides y prevenir que las variables de mayor magnitud dominen los resultados del algoritmo.

3.4. Construcción e implementación del modelo

Una vez preparado los datos aplicando el análisis RFM + Estacional, procedemos con la construcción del modelo de clústering en función de cumplir con el objetivo del proyecto. Para esto se decidió utilizar dos algoritmos, uno basado en densidad (DBSCAN) y otro en distancia (KMEANS).

3.4.1. Parametrización y ejecución DBSCAN

El algoritmo DBSCAN, trabaja con dos parámetros fundamentales, épsilon (*eps*) o radio de vecindad y la cantidad de puntos mínimos para conformar una vecindad (*min_samples*) los mismos que se explican a continuación.

3.4.1.1. Selección de valor min_samples óptimo.

Para el valor óptimo de *min_samples*, se recomienda aplicar la regla basada en las columnas/dimensiones (D) del dataset que estamos analizando, está dado por la fórmula:

$$min\_samples = 2 \cdot D \quad (8)$$

Para el caso de nuestro dataset, teóricamente el valor óptimo sería 8, sin embargo por la magnitud de los datos, y la presencia de outliers que marcan grandes diferencias entre los

valores de clientes de alto valor, con clientes regulares, hace que nuestro min_samples óptimo sea ajustado con un valor superior, para que el modelo sea más exigente al formar los clústeres, además usar un valor bajo como 8 en un conjunto de datos extenso haría que se formen micro clústeres, y para el objetivo de negocio tener demasiados clústeres con pocos integrantes hacen inviable aplicar campañas de marketing. Basado en esta explicación se decide fijar el parámetro min_samples en 100.

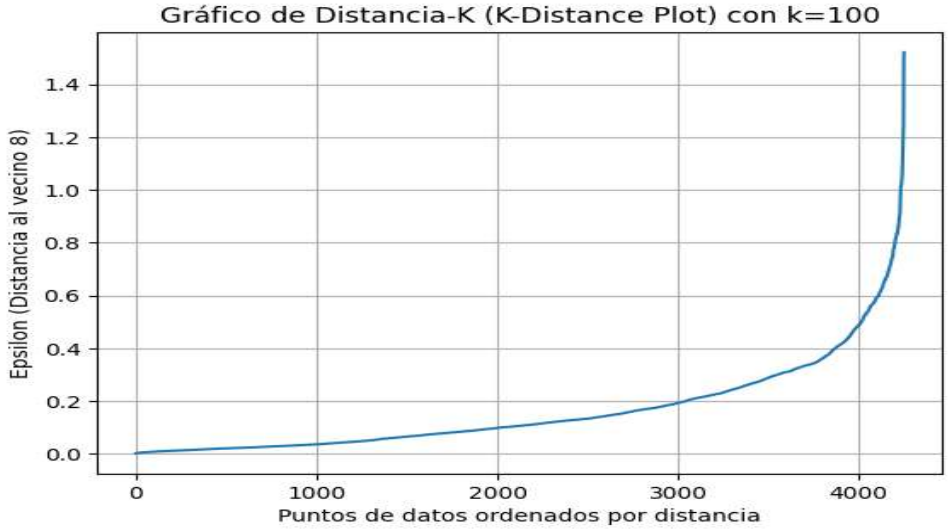
3.4.1.2. Selección de valor épsilon óptimo.

Para este motivo se empleó el método de *k-distancia*, en función de la distancia de cada punto a sus "k" vecinos más cercanos. La Figura 11 es el resultado de este método, en el análisis se quiere encontrar el punto donde existe el quiebre o la máxima curvatura (codo).

Aplicando la teoría, se puede observar que un valor óptimo para el modelo se encuentra en el rango de (0.4 a 0.5).

Figura 11

Gráfico de distancia k



Nota. La figura muestra el gráfico de distancia-K para k = 100, utilizado para identificar el valor óptimo de epsilon en el algoritmo DBSCAN. El punto de inflexión indica el umbral adecuado para separar ruido de clústeres.

3.4.1.3. Ejecución modelo DBSCAN.

Para la ejecución del modelo DBSCAN se emplearon los parámetros obtenidos de *eps*=0.5, y *min_samples*=100, se utilizó la librería sklearn de Python y su adaptación del algoritmo, otro de los parámetros usados en esta implementación es la distancia euclidiana configurada por defecto por DBSCAN, y con lo cual se asegura la coherencia geométrica en el espacio vectorial.

El universo de clientes de la empresa lo conforman 4250 elementos, los mismos que fueron agrupados por el modelo en los siguientes clústeres detallados en la Tabla 2.

Tabla 2

Clústeres resultado DBSCAN

Clúster	Cantidad de clientes
-1	1151
0	1003
1	629
2	946
3	399
4	122

Nota. clúster = identificador numérico del clúster.

Es importante recalcar que una de las funcionalidades principales de DBSCAN es la de presentar un clúster denominado -1, en el cual agrupa los elementos que no pertenecen a

ninguno de los otros clústeres, por lo general son valores catalogados como ruido. En este sentido el modelo DBSCAN proporcionó 5 clústeres etiquetados del 0 al 4.

3.4.1.4. Parametrización KMEANS.

El algoritmo KMEANS, emplea el parámetro k, el mismo que define la cantidad de clústeres que forma el modelo, para este fin se emplean principalmente dos técnicas, el método del codo, y el método Silhouette Score.

3.4.1.5. Método del codo.

Como dicta la teoría, establecimos el rango (2 a 10) de clústeres con el que se ejecuta el algoritmo y obtuvimos los valores de inercia en cada iteración, detallados en la Tabla 3.

Tabla 3

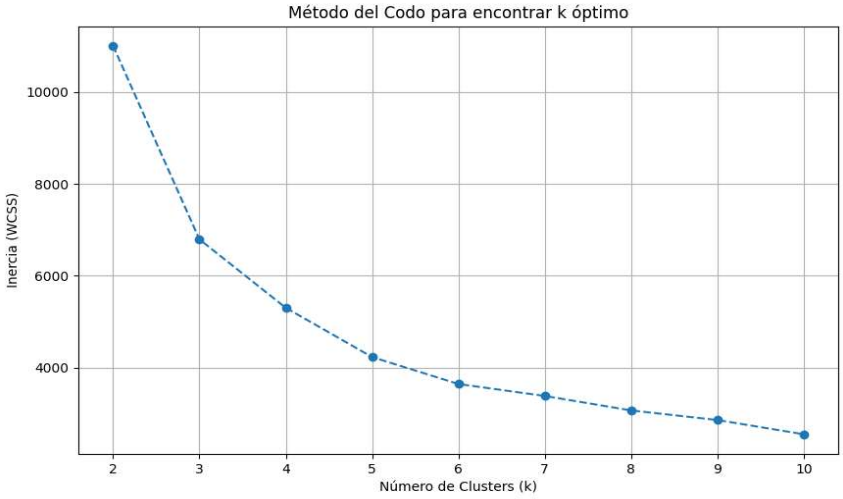
Valores de inercia para k en el rango 2 a 10

k	Inercia
2	11005.3176
3	6793.0648
4	5304.5487
5	4232.5362
6	3646.2492
7	3386.5382
8	3071.6769
9	2862.9130
10	2552.6387

Nota. k = número de clústeres

Figura 12

Gráfico del resultado obtenido con el método del codo



Nota. El gráfico muestra la variación de la inercia (WCSS) en función del número de clústeres para identificar el valor óptimo de k mediante el método del codo se observa que disminuye progresivamente a medida que se incrementa el número de clústeres en la gráfica determinamos que el valor óptimo de k estaría en el rango de 3 a 5, para establecer el valor final que empleamos en el modelo hacemos uso de otro método con el mismo fin.

3.4.1.6. Método de Silhouette Score.

De igual forma que en el método anterior, se estableció el rango de iteración (2 a 10) para el valor de k, después en cada ejecución del algoritmo dentro del rango se obtuvo el valor Silhouette Score, listados en la Tabla 4.

Tabla 4

Valores de silhoutte score para k en el rango 2 a 10

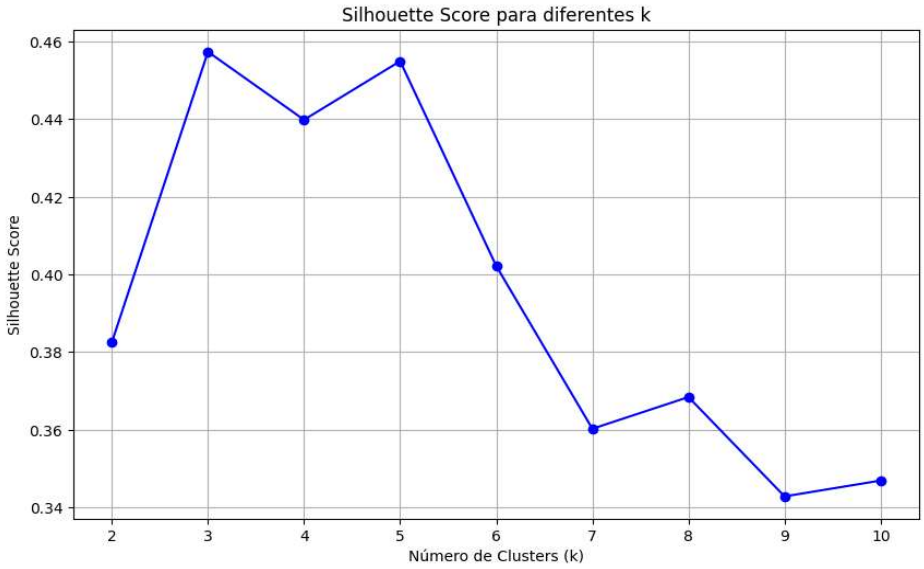
k	Silhouette Score
2	0.3825
3	0.4573
4	0.4398
5	0.4549
6	0.4021
7	0.3602
8	0.3684
9	0.3428
10	0.3469

Nota. k = número de clústeres

Como se apreció en la aplicación del método del codo, el k óptimo se encuentra entre 3 a 5 para nuestro fin elegimos los valores 3 y 5, por ser los más altos y prácticamente similares. En la Figura 13, se pueden ver los puntos en el plano para apreciar las diferencias existentes.

Figura 13

Gráfico comparativo de silhouette score para k en rango (2-10)



Nota. La figura presenta los valores del Silhouette Score para diferentes cantidades de clústeres (k), utilizados para evaluar la calidad de la segmentación generada por el algoritmo k-means.

Si bien el análisis arroja una ligera ventaja matemática del punto k=3 sobre k=5, al ser poco significativa se prioriza la utilidad en el negocio, tener 5 clústeres ofrece mayor granularidad al permitir tener clústeres con flexibilidad, pero sin ser inmanejables, y tampoco clústeres que generalicen a nuestros clientes manteniéndolos ocultos en una gran masa. En este sentido la decisión final fue usar k=5.

3.4.1.7. Ejecución modelo KMEANS.

Para la ejecución del modelo KMEANS, se empleó la librería sklearn de Python y su adaptación del algoritmo. De acuerdo con el análisis previo se obtuvo un valor k óptimo igual a 5, a este se unen los parámetros que el modelo usa por defecto como son: primero el método de inicialización *init=kmeans++*, para garantizar una selección inicial inteligente de

los centroides, y segundo el parámetro de número de iteraciones *n_init=10* con el fin de que el modelo internamente se ejecute diez veces variando la distribución randómica inicial de los centroides, y con esto evalúe el mejor resultado obtenido.

Al aplicar el modelo sobre el universo de clientes, y tomando en cuenta que KMEANS tiene como característica que el 100% de casos serán clasificados en un clúster, se obtuvo la distribución de estos detallados en la Tabla 5.

Tabla 5

Clústeres resultado KMEANS

Clúster	Cantidad de clientes
0	1645
1	301
2	906
3	1224
4	174

Nota. clúster = identificador numérico del clúster.

3.5. Generación de dashboard en Power BI

A partir de la información del modelo generado en los apartados anteriores, se desarrolló un dashboard en Power BI como herramienta de Business Intelligence. La elección de este software se genera por la facilidad de integrar diversos conjuntos de datos y ofrecer objetos visuales dinámicos y modernos, útiles en el análisis y toma de decisiones.

3.5.1. Preparación de datos

La construcción del dashboard se estructuró en torno a tres datasets:

- Dataset limpio: contiene información de tres años del e-commerce.

- Dataset derivado del modelo RFM + Kmeans: incluye la segmentación de clientes y datos de RFM.
- Dataset crudo

La preparación incluyó procesos de limpieza y generación de medidas personalizadas en DAX, teniendo un control de la trazabilidad de cada transformación dentro del software de Power BI. Entre las medidas implementadas se encuentran cálculos específicos, como: Promedio Monetario por Número de Ventas, Promedio de Frecuencia, Promedio de tallos por Número de Ventas, entre otros.

3.5.2. Publicación

Luego de la elaboración del dashboard en Power BI Desktop, se procedió a publicarlo en una cuenta con acceso a Power BI Pro en la versión online de este software. Después, se trabajó generando una página web que soporte la visualización de este tablero y a la cual se puede acceder con credenciales del Anexo 4.

4. Análisis de Resultados

4.1 Pruebas de concepto

Una vez se han construido los modelos KMEANS y DBSCAN, es importante revisar la calidad del clústering a través de métricas.

4.1.1. Silhouette Score

Se decidió obtener el Silhouette Score en ambos modelos, los resultados se detallan en la Tabla 6.

Tabla 6

Resultado Silhouette Score de modelo KMEANS y DBSCAN

Modelo	Resultado
DBSCAN	0.4100
KMEANS	0.4549

Dada la naturaleza de DBSCAN de generar un grupo catalogado como ruido que se identifica con código -1, es importante mencionar que dicho grupo fue excluido del cálculo de Silhouette Score, con el fin de no distorsionar su resultado.

En base a la teoría el Silhouette Score mide qué tan similar es un elemento al resto de elementos dentro del clúster y que tan diferente a los elementos del clúster vecino, su puntaje se encuentra en un rango de -1 a 1, donde valores más cercanos a 1 serán ideales, bajo esta premisa KMEANS tiene un mejor resultado para la métrica.

Ambos métodos muestran clústeres moderadamente definidos, mejor cohesión y mayor separación en cada clúster, sin embargo, aunque es útil detectar ruido y outliers DBScan tiene un resultado más bajo.

El índice Silhouette más alto es de 0.4549 que representa al modelo de KMEANS donde se identifica los puntos mejor agrupados dentro de los clústeres sin embargo se detectó que el modelo de DBSCAN detecta ruido y outliers que bajaron Silhouette.

Según el índice Silhouette k-Means representa una mejor calidad de clústering indicando una mayor separación entre los clústeres y una mejor cohesión interna.

4.1.2. Davies Boulding

Esta métrica busca conocer qué tan compactos están los clústeres, y que tan separados están de sus centros. La dificultad que se enfrenta es que la métrica asume una forma esférica en los clústeres, y es ahí donde DBSCAN por su origen de hallar formas irregulares basándose en la densidad tiene una desventaja frente a KMEANS para el puntaje Davies Boulding. En la Tabla 7 se listan los resultados de cada modelo.

Tabla 7

Resultado Davies Boulding

Modelo	Resultado
DBSCAN	0.9670
KMEANS	0.8540

Al igual que en la métrica anterior, para el caso de DBSCAN se excluyó el grupo del ruido, basados en la teoría para este puntaje un valor cercano a 0 resulta ser óptimo, lo cual hace que KMEANS vuelva a tener ventaja sobre DBSCAN.

4.1.3. Análisis de clúster*es*

En esta sección se presenta el análisis de los clústeres generados por cada modelo, no se busca comparar uno a uno los clústeres ya que no tienen equivalencia, sino más bien explicar la utilidad e importancia para nuestro objetivo.

En la Tabla 8 se puede apreciar como el modelo DBSCAN presenta un gran problema y es que 1151 clientes fueron catalogados como ruido, lo que representa en nuestro total de clientes el 27%, este grupo de ruido no puede ser considerado un clúster real, en él se encuentran elementos con valores atípicos que no pudieron ser asignados a un clúster válido, en términos de cumplir con el objetivo del negocio el no incluir esta gran cantidad de clientes en futuras campañas de marketing hará que nuestro análisis pierda objetividad y no sea útil.

Tabla 8

Listado de clúster y cantidad de elementos

Clúster	Cantidad DBSCAN	Cantidad KMEANS
-1	1151	NA
0	1003	1645
1	629	301
2	946	906
3	399	1224
4	122	174

Nota. clúster = identificador numérico del clúster.

Es por este motivo que se decide descartar el uso de DBSCAN para nuestro análisis, el modelo no es el adecuado para adaptarse a la estructura u origen de los datos, y el cumplimiento de los objetivos del proyecto. En contraste KMEANS con su forma rigurosa de asignar el cien por ciento de los elementos a un clúster específico, es de mayor utilidad en pro de conseguir un análisis total de nuestra base de clientes.

4.1.4. Análisis de Resultados

Este apartado, se centra en el análisis de los resultados y perfilamiento de los clústeres generados por el modelo KMEANS, para ello se presenta una serie de gráficos e ilustraciones que permitan comprender la situación actual de los clientes de la empresa.

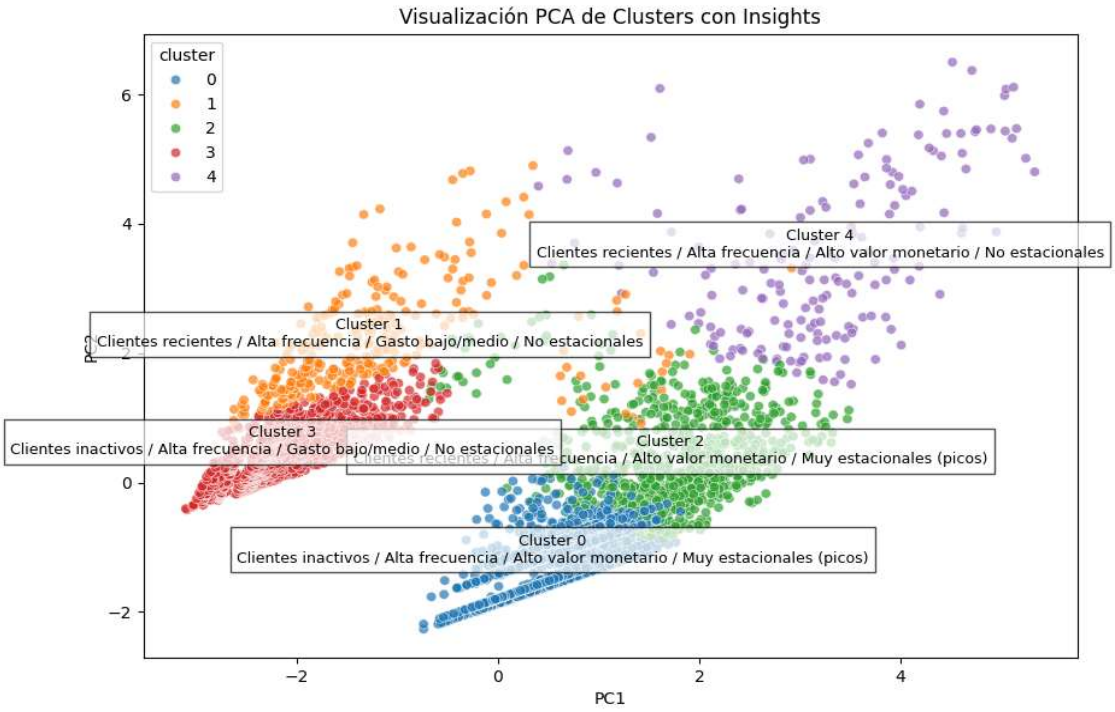
4.1.5. Resultados Técnicos

Para la representación gráfica de los segmentos de clientes obtenidos, se aplicó la técnica de Análisis de Componentes Principales (PCA) con el objetivo de reducir la dimensionalidad del conjunto de datos original (RFM y estacionalidad). Dado que la visualización directa en un espacio multidimensional es inviable para la interpretación humana, el PCA nos permite proyectar la información en un plano bidimensional conservando la mayor cantidad posible de la varianza original. Como explican Hair et al. (2007) y Pérez López (2004) en sus estudios acerca de análisis de multivariables, este procedimiento simplifica la complejidad de los datos originales, al transformar las variables en un conjunto menor de componentes, lo que resulta en una visualización clara al ojo humano. En la Figura 14, se muestra claramente la separación entre los clústeres, lo que evidencia que la segmentación obtuvo grupos distinguibles y con comportamientos diferenciados. Además, se han incorporado etiquetas descriptivas que sintetizan los rasgos más característicos de cada clúster, permitiendo contextualizar su comportamiento, los ejes representan los dos componentes principales, lo que facilita visualizar los clústeres generados y la cohesión de los elementos en cada uno de ellos.

Por ejemplo, ciertos grupos se asocian con clientes recientes y de alta frecuencia, mientras que otros representan clientes inactivos o altamente estacionales. Esta visualización resulta fundamental para comprender cómo se distribuyen los segmentos y para validar la coherencia entre los patrones identificados y los objetivos del análisis, particularmente en términos de valor del cliente, estacionalidad y frecuencia de compra.

Figura 14

Visualización PCA de clústeres con insights generados



Nota. La figura muestra la proyección bidimensional generada mediante Análisis de Componentes Principales (PCA), donde los puntos representan clientes y los colores indican los clústeres identificados. Se incluyen etiquetas con los principales insights descriptivos de cada grupo.

4.1.6. Perfilamiento

Por medio de la Tabla 9 se presenta un análisis individual por clúster, para el modelo KMEANS, en ella se realizó el promedio de las variables del RFM Estacional aplicado para cada clúster generado.

Tabla 9

Análisis de Clústeres KMEANS (Valores Promedio por Clúster)

Clúster	Cantidad	R (días)	F (cantidad)	M (dólar)	E (índice)
0	1645	456.89	1.48	517.8	0.47
1	301	48.91	2.03	370.77	0
2	906	272.97	6.46	2679.48	0.35
3	1224	472.9	1.3	259.74	0
4	174	36.34	44.33	16702.1	0.18

Nota. Clúster = identificador numérico del clúster; R = recencia; F = frecuencia, M = monetario; E = estacionalidad.

Respecto al clúster 0, su alto valor en recencia e índice de estacionalidad nos permite identificar un comportamiento estacional, es decir en ellos se encuentran los clientes que compran mayoritariamente en fechas especiales, 1 o 2 veces por año esto explica la frecuencia y la recencia desde su última compra, por este motivo a este grupo lo denominamos “Clientes Estacionales”.

Con el clúster 1, destaca la recencia en promedio han transcurrido 48 días a partir de su última compra y tienen una frecuencia de 2 pedidos, la frecuencia relativamente alta en función de la recencia deduce que se trata de clientes ingresando a probar la marca, los denominaremos “Clientes Nuevos”.

Al clúster 2, lo llamamos “Clientes en Riesgo”, lo curioso es que tienen un valor promedio de gasto alto $2679.48, frecuencia alta de 6 pedidos, pero su recencia de 272 días desde que compraron por última vez esto expone una problemática, ha transcurrido mucho tiempo desde su última compra, hay un riesgo de fuga, tiene cierto grado de estacionalidad que podría explicar porque su recencia es alta pero no es la principal causa.

El clúster 3, su recencia de más de 1 año sin comprar, su baja frecuencia 1 pedido con gasto promedio bajo, se puede concluir que fueron clientes que compraron una única ocasión, en resumen, son “Clientes Esporádicos”.

Finalmente, el clúster 4, baja recencia 1 mes en promedio, alta frecuencia alrededor de 44 pedidos, y un valor altísimo de compra en comparativa a los anteriores clústeres, es decir aquí se encuentran los mejores clientes de la empresa sus “Clientes Wholesaler”.

Resumiendo, la exploración de resultados en cada clúster que se puede apreciar en la Tabla 10, donde vemos el nombre con el que se cataloga a cada grupo el mismo que será de utilidad de aquí en adelante para su identificación.

Tabla 10

Denominación de clústeres

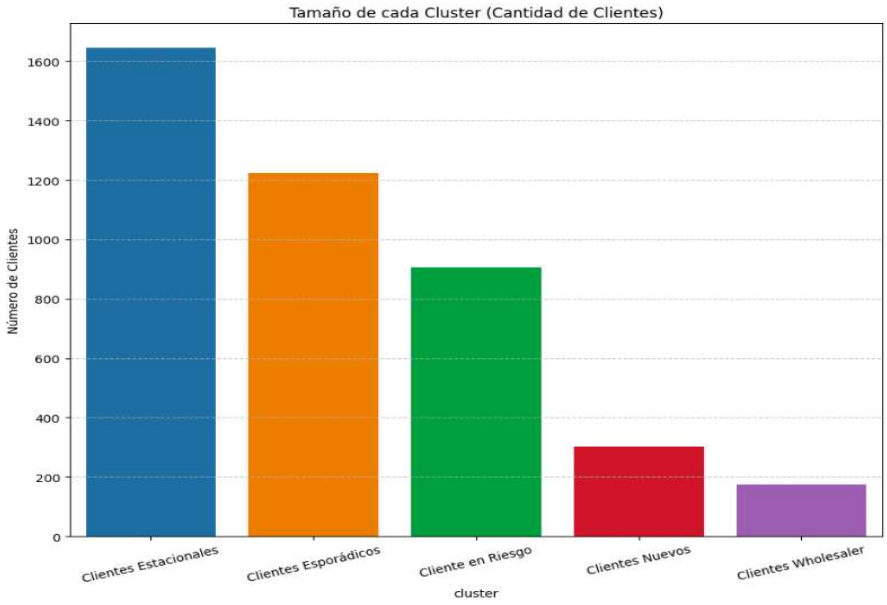
Número del clúster	Nombre del clúster
4	Clientes Wholesaler
3	Clientes Esporádicos
2	Clientes en Riesgo
1	Clientes Nuevos
0	Clientes Estacionales

Nota. Número = identificador numérico del clúster; Nombre = etiqueta asignada para su uso en el análisis.

4.1.7. Análisis de variables

Figura 15

Gráfico de barras, cantidad de clientes



Nota. La figura muestra la cantidad de clientes pertenecientes a cada clúster identificado en el análisis de segmentación, representados mediante un gráfico de barras.

La Figura 15 se presenta el tamaño de cada clúster generado en el proceso de segmentación. El gráfico permite visualizar claramente la distribución de los clientes entre los distintos grupos formados. Se observa que el clúster más numeroso corresponde a los Clientes Estacionales, seguido por los Clientes Esporádicos, lo cual indica que una parte significativa de la base presenta comportamientos de compra no constantes pero recurrentes en ciertos periodos. la distribución de la cartera de clientes, como se puede apreciar existe disparidad en el volumen de cada clúster. Mientras la mayoría de población se encuentra aglutinada en segmentos masivos, el clúster de “Clientes Wholesaler” apenas representan al 4% de la misma, esto confirma su carácter selectivo y exclusivo.

Si bien la sección de “Clientes Nuevos” es baja representan ya al 7% de la población, pero son clientes con potencial crecimiento a futuro, en contraste los “Clientes Estacionales”

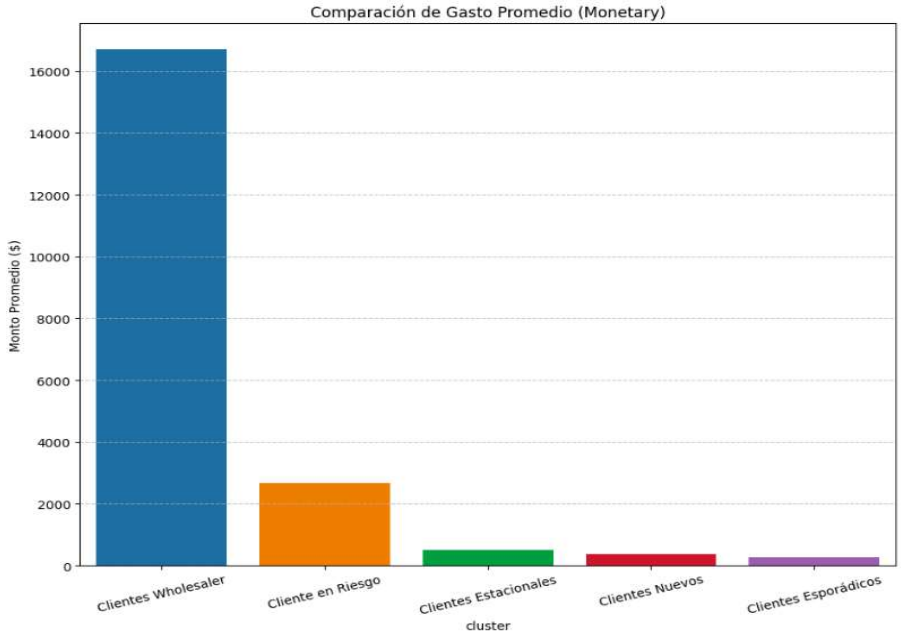
es el segmento de mayor cantidad 39%, su comportamiento de compra aporta al negocio y es fundamental su identificación en función de los intereses de la empresa.

Finalmente, los focos de atención se centran en la cantidad de “Clientes en Riesgo” el 21% es un nicho en el que se deben concentrar esfuerzos de retención, y los “Clientes Esporádicos” con el 29%, su actuar en el historial de compras ha sido revelador para la empresa, un segmento que no aporta mucho valor a la empresa.

Esta distribución aporta una visión global de la composición del mercado y permite priorizar acciones de marketing, servicio y fidelización según el tamaño y características de cada clúster.

Figura 16

Gráfico de barras, comparación de gasto promedio



Nota. La figura muestra la comparación del gasto promedio entre los distintos clústeres identificados, evidenciando diferencias significativas en el valor monetario asociado a cada segmento.

En contraste con su tamaño mínimo, la Figura 16 revela la influencia del valor monetario en el análisis. Visualmente el segmento de “Clientes Wholesaler” sobresale drásticamente sobre el resto, lo que evidencia la dependencia de este grupo para la sostenibilidad de la empresa.

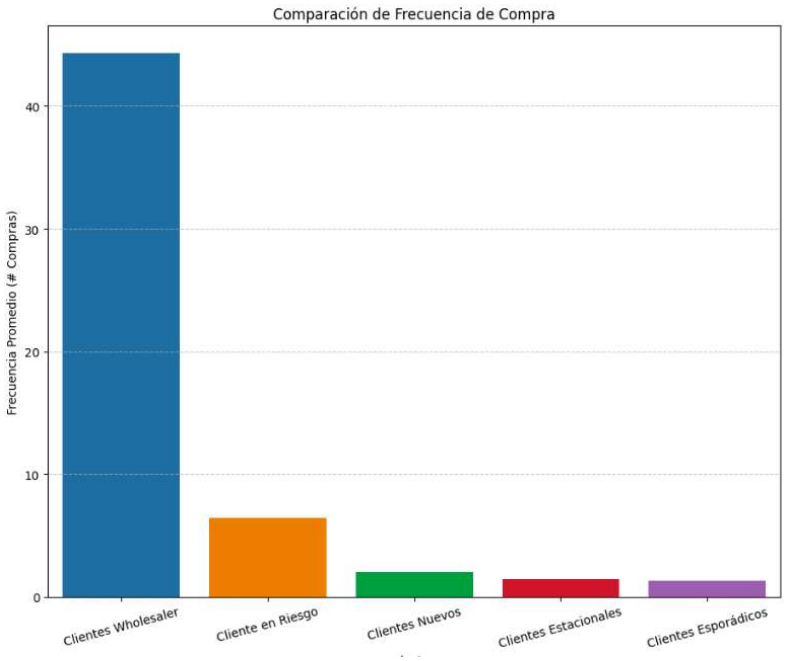
Esta variable ha puesto en evidencia una vez más la atención prioritaria que se debe dar al segmento de “Clientes en Riesgo”, históricamente tienen un gasto promedio medio, pero debe existir un motivo por el que se han alejado de la marca, recuperar clientes de este segmento es el objetivo.

Los “Clientes Nuevos”, y “Clientes Estacionales”, cuyo promedio en gasto es bajo con relación al de “Clientes Wholesaler”, representan un desarrollo progresivo a futuro.

Como punto final se encuentran los “Clientes Esporádicos”, el análisis monetario termina de sustentar la decisión de no agotar esfuerzos en este segmento, su gasto promedio histórico es el más bajo de todos.

Figura 17

Gráfico de barras, comparación de frecuencia de compra



Finalmente, en la Figura 17 se confirma la correlación directa que existe entre el gasto monetario y la frecuencia de compra, los “Clientes Wholesaler” lideran esta gráfica lo que valida que su alto gasto promedio es el resultado de un comportamiento de compra leal y consolidado.

El análisis en el resto de grupos, revela ciertos matices en su conducta: los “Clientes en Riesgo” muestran una frecuencia histórica media, que confirma que existió una rutina de compra, y algo hizo que la misma se vea afectada, por el contrario la intermitencia de los “Clientes Estacionales” avala su consumo por la necesidad puntual mas no por fidelidad, mientras que la baja frecuencia de los “Clientes Nuevos” se ve explicada en su reciente ingreso al radar de la empresa, pero perfectamente diferenciados de los “Clientes Esporádicos” cuya nula recurrencia respalda un poco interés en la marca.

En contraste, los clústeres de Clientes Estacionales, Clientes Nuevos y Clientes Esporádicos exhiben valores de gasto promedio considerablemente menores, lo cual es coherente con comportamientos de compra ocasionales o de menor intensidad. Por su parte, el clúster de Clientes en Riesgo presenta un gasto intermedio, lo que podría indicar que estos clientes anteriormente tenían niveles de gasto más elevados, pero han reducido su actividad recientemente.

Esta comparación resulta clave para priorizar estrategias comerciales y de marketing, ya que permite identificar los segmentos con mayor valor económico y aquellos que requieren acciones de reactivación o fidelización.

4.2 Herramienta BI

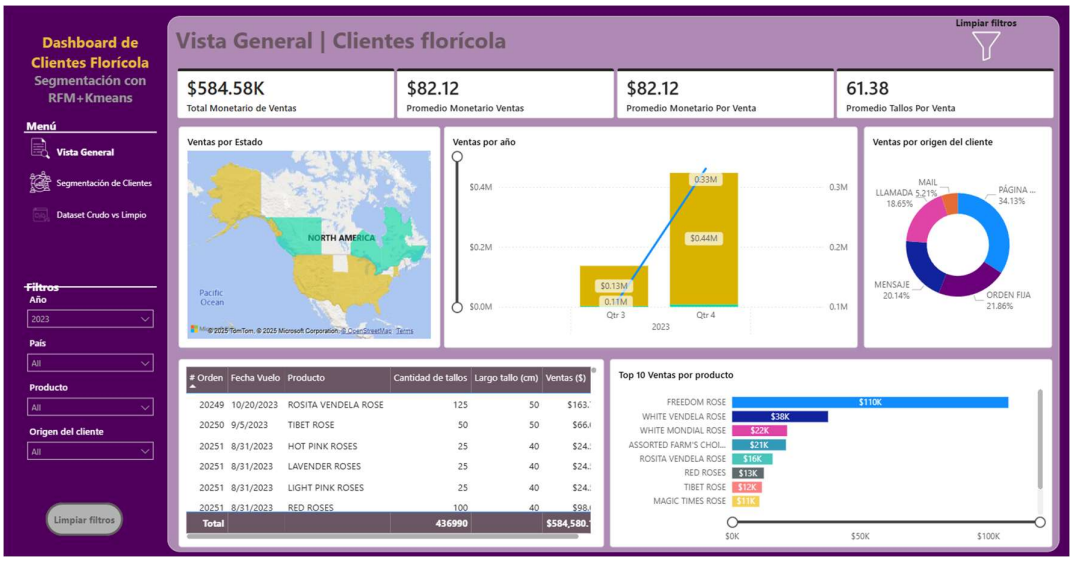
El dashboard se planteó para contener indicadores de ventas dentro de tarjetas, paneles de filtros, gráficos de barras por productos y tendencia. Este fue publicado en una página web desde donde se puede acceder a todas las funciones de visualización, el proceso

de acceso es detallado en el Anexo 4. En general, se trabajó con 3 pestañas debido al contenido de información:

- Visión general: contiene indicadores sobre los datos de los años 2023 a 2025 de ventas del e-commerce obtenidas del dataset limpio o procesado.

Figura 18

Ventana de Visión general



En la Figura 18, se pueden observar en filtros de año, país, producto y origen del cliente (correspondiente al medio del que vino a realizar la compra). De la misma manera, se presentan tarjetas con métricas de valor monetario, promedio de frecuencia, valor monetario y recencia. También se observan que los cuartiles Q2 y Q4 del año contienen los meses pico de venta monetario y de número de tallos, el origen de venta mayor corresponde a la página web y el top 1 de productos es Freedom Rose.

- Segmentación de clientes: incluye los datos del modelado utilizando RFM + KMEANS, donde se identifican los clústeres diferenciados por recencia, frecuencia y valor monetario.

Figura 19

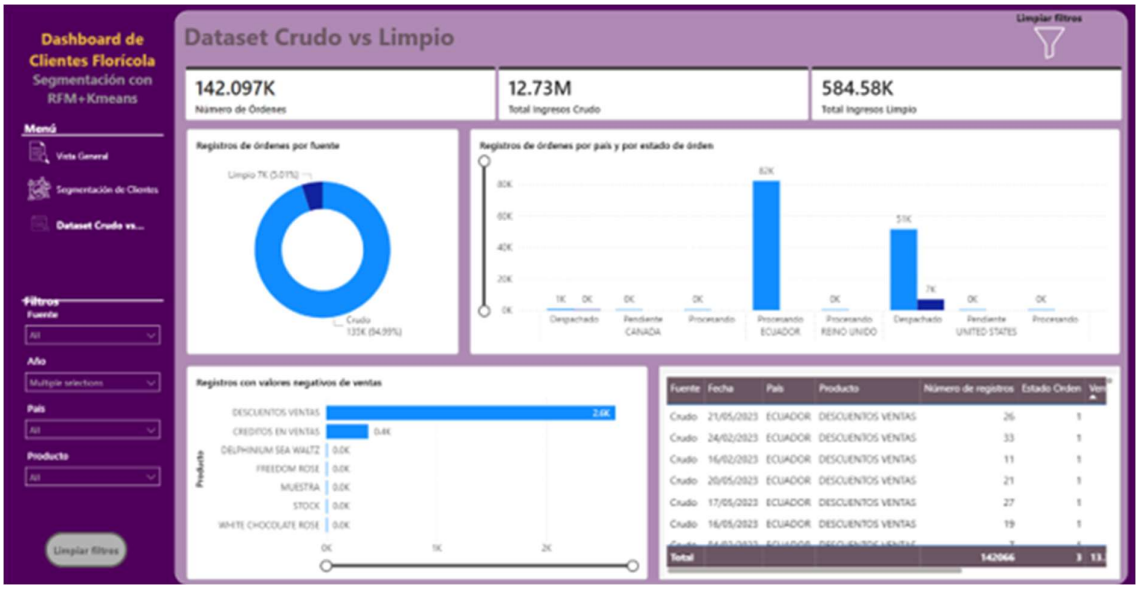
Ventana de Segmentación de clientes

En la Figura 19, se presentan filtros de nombre de clúster, ID del cliente y rango de recencia, junto con tarjetas de métricas como total de clientes, total valor monetario, promedio de frecuencia, valor monetario y recencia. Asimismo, se visualiza la distribución de clientes por clúster, una tabla resumen, tres gráficas de análisis de influencia según valor RFM. Esta información permite verificar que los "Clientes Wholesaler" y los otros clústeres están bien asignados. Este análisis se complementa con las gráficas de influencia en cada variable RFM, que permiten identificar dinámicamente que afecta en el aumento o disminución de los indicadores.

- Dataset crudo vs limpio: permite realizar comparaciones para evidenciar diferencias con el dataset procesado, especialmente valores faltantes, órdenes no procesadas, valores negativos, entre otros explicados en los apartados de desarrollo.

Figura 20

Ventana de Dataset Crudo vs Limpio



En la Figura 20, se observan tarjetas de métricas relacionadas a la comparación de los datasets pre y post limpieza, centrándose en la distribución de los registros según fuente, país, estado de orden. De manera complementaria, en esta ventana se visualizan gráficas de los puntos que se eliminaron por tener valores negativos y una tabla resumen consolidando información relevante de que registros fueron descartados o mantenidos en el dataset procesado.

5. Conclusiones y Recomendaciones

5.1. Conclusiones

Se realizó la compilación de todas las transacciones del año 2023 a 2025 del e-commerce utilizando lecturas a la base de datos de manera directa, que permitió contar con un dataset con registros representativos útiles para el análisis y la segmentación de clientes.

El procesamiento de los datos del e-commerce permitió contar con registros limpios, transformados y de calidad mediante un pipeline de limpieza generado en python para el posterior análisis de la información.

El análisis realizado demuestra que la combinación del modelo RFM con la estacionalidad permitió caracterizar de manera integral el comportamiento de compra de los clientes, revelando una alta heterogeneidad en términos de actividad, valor económico y dependencia a fechas especiales. Los resultados estadísticos evidencian una base de clientes mayoritariamente ocasional, con baja frecuencia y bajo gasto, coexistiendo con un grupo reducido, pero altamente rentable, caracterizado por compras recurrentes y montos elevados.

La fuerte asimetría y presencia de valores extremos justificó la aplicación de transformaciones logarítmicas y estandarización, pasos fundamentales para garantizar la correcta ejecución del algoritmo de K-means, evitando sesgos en el cálculo de distancias y permitiendo una segmentación más estable, interpretable y representativa del comportamiento real de los clientes.

Se concluye que el algoritmo K-Means es el más idóneo para nuestro conjunto de datos. K-Means logró el mejor equilibrio entre la cohesión interna y en la separación de grupos, esto respaldado por sus métricas en el Índice Davies-Bouldin y un Silhouette Score sólido, su fundamento en la distancia euclidiana a un centroide es indiferente a la densidad, y pudo capturar mejor la estructura global del negocio.

DBSCAN resultó ser inadecuado para la segmentación de nuestros datos, al ser un algoritmo basado en densidad, la cual en nuestro conjunto de datos es heterogénea: muchos clientes con gasto bajo (alta densidad) y pocos clientes con alto gasto (baja densidad) clasificó como ruido a los clientes de alto valor por su dispersión en el espacio, en nuestro objetivo los outliers positivos no podían ser tratados como ruido a fin de mantener un análisis total y de utilidad para el negocio.

La integración del modelo de segmentación con una herramienta de Análisis de Datos como Power BI permitió traducir los resultados analíticos en información comprensible y accesible para los equipos comerciales y de marketing. Esta capa de visualización consolidó el valor práctico del proyecto, evidenciando que los modelos de ciencia de datos alcanzan su máximo impacto cuando se integran en gráficos que ayudan a la decisión empresarial y no se limitan únicamente al análisis técnico.

5.2. Recomendaciones

Se recomienda a la empresa Florícola implementar estrategias comerciales diferenciadas por segmento, priorizando acciones de fidelización para los clientes, como por ejemplo alguna campaña que permitan a los clientes ganar puntos por cada compra. Esta segmentación permitirá optimizar el presupuesto de marketing y mejorar la efectividad de las campañas.

Se recomienda diseñar un cronograma de reentrenamiento del modelo de segmentación utilizando la información de las nuevas transacciones que se registran en el e-commerce con el objetivo de mejorar las métricas de desempeño y su precisión.

Finalmente, se sugiere que los resultados de la segmentación RFM–estacional sean integrados en los procesos de gestión comercial de la empresa, utilizando los clústeres como base para el diseño de estrategias diferenciadas, campañas personalizadas y optimización de

recursos, de modo que la analítica de datos no solo tenga un aporte académico, sino también un impacto práctico y medible en el desempeño organizacional.

Referencias

Berry, L. L. (2002). Relationship Marketing of Services: Perspectives from 1983 and 2000. *Journal of Relationship Marketing*, *1*(1), 59–77. https://doi.org/10.1300/J396v01n01_04

Biswas, S., Wardat, M., & Rajan, H. (2021). The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. arXiv. https://doi.org/10.48550/arXiv.2112.01590

Cáceres Lobato, D. D. (2019). Análisis de exactitud de los algoritmos de clústering aplicados en la base de datos del sistema académico de la UNACH [Tesis de grado, Universidad Nacional de Chimborazo]. Repositorio Digital UNACH. http://dspace.unach.edu.ec/handle/51000/6114

Caicedo Dorado, A. (2022). Agrupamiento básico k-means y k-medoids. https://doi.org/10.48713/10336_46969

Campbell, J. H., & Campbell, B. L. (2025). Cut flower purchasing and market segments within the US flower industry. HortTechnology, 35(3), 286–296. https://doi.org/10.21273/HORTTECH05584-24

Catota-Mesías, V. D., Ramírez-Proaño, M. S., Toscano-Ramos, E. S., & Licto-Vergara, A. E. (2024). Growth marketing y comunicación en el sector florícola: Desarrollo de estrategias. Revista Sigma, 11(2), 144–152. https://doi.org/10.24133/na76yf17

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research, 42(4), 415-430.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (2007). Análisis multivariante (5.ª ed.). Pearson Prentice Hall.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Helu, M. (2024). *Scalable data pipeline architecture to support the integration and analytics of heterogeneous data sources*. National Institute of Standards and Technology..

Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2023). Customer profiling, segmentation, and sales prediction using AI in direct marketing. arXiv. https://doi.org/10.48550/arXiv.2302.01786

Kotler, P., & Keller, K. L. (2016). Marketing Management (15a edición). Pearson.

Krantz, T., & Jonker, A. (2024)**.** *¿Qué es la arquitectura de datos?* IBM . Nombre del sitio web. https://www.ibm.com/es-es/think/topics/data-architecture

McKinney, W. (2018). Python para análisis de datos: Manipulación de datos con pandas, NumPy y IPython (2.ª ed.). Anaya Multimedia.

Naphade, D. (2025). *The evolution and modernization of data pipeline architectures*. *European Journal of Computer Science and Information Technology*, 13(6), 42–53. https://doi.org/10.37745/ejcsit.2013/vol13n64253

Pérez López, C. (2004). Técnicas de análisis multivariante de datos: Aplicaciones con SPSS. Pearson Educación.

Raja, M. S. (2025). Architecting data pipelines for scalable and resilient data processing workflows. International Journal of Emerging Research in Engineering and Technology, 6(1), 1–9.

Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). Minería de datos: Conceptos y tendencias. Revista Iberoamericana de Inteligencia Artificial.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65.

Sánchez Gutiérrez, I. (2024). *Uso del clústering espacial basado en densidad de aplicaciones con ruido en la minería de datos: Creación de una aplicación web con R para clasificar o predecir datos reales* [Trabajo de fin de grado, Universidad de Salamanca]. Repositorio Gredos. https://gredos.usal.es/bitstream/handle/10366/164345/TFG_Isabel_S%C3%A1nchez_Guti%C3%A9rrez.pdf?sequence=1&isAllowed=y

Sander, J. (2011). *Density-based clustering*. En C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning*. Springer. https://doi.org/10.1007/978-0-387-30164-8_488

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? Management Science, 33(1), 1-24.

Torroba Moreno, J. (2023). Análisis de segmentación de clientes en ventas minoristas mediante aprendizaje automático [Tesis de maestría, Universidad Internacional de Andalucía]. Repositorio de la Universidad Internacional de Andalucía.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media.

Villacrés Venegas, E. J. (2023). *Aplicación de las ciencias de datos para identificar segmentos de clientes en una cadena de farmacias* [Tesis de maestría, Pontificia Universidad Católica del Ecuador]. https://repositorio.puce.edu.ec/handle/123456789/41131

Yen, L. (2025)**.** *Data pipeline architecture: A comprehensive guide.* Datamation. https://www.datamation.com/big-data/data-pipeline-architecture

Anexo 1. Repositorio código fuente y recursos del proyecto

El desarrollo práctico del proyecto, incluyendo los scripts de extracción, transformación y carga de datos, los notebooks (Jupyter Notebooks) y los resultados de los modelos, se encuentra alojados en un repositorio de control de versiones en GitHub. A continuación, se detalla la guía de acceso y la estructura del proyecto para garantizar la reproducibilidad de los resultados presentados en el Capítulo 4.

- Enlaces de acceso

El repositorio es de acceso público y se encuentra disponible en la siguiente dirección URL: https://github.com/ccastro1992/clústering_rfm

Además se creó un enlace directo en cada notebook del repositorio para ser desplegado en Google Colab:

(KMEANS)

https://colab.research.google.com/github/ccastro1992/clústering_rfm/blob/main/notebooks/ModeloFinal.ipynb

(DBSCAN)

https://colab.research.google.com/github/ccastro1992/clústering_rfm/blob/main/notebooks/ModelamientoDBSCAN.ipynb

- Estructura del repositorio

Los archivos están organizados de la siguiente manera:

  - /notebooks: Contiene los análisis principales.
  - /src: Contiene los archivos de funciones extras en Python.
  - README.md: Información del proyecto.
  - .gitignore: Archivos excluidos del control de versiones.

- Instrucciones de ejecución

Este proyecto utiliza Google Colab Secrets para manejar las credenciales de la API de forma segura.

- Abrir el Notebook: Haz clic en el botón "Open in Colab" de arriba o abre el archivo .ipynb directamente en Google Colab.
- Configurar Credenciales: Para que el código funcione, es necesario el API_KEY. En Google Colab, en el menú lateral izquierdo, haz clic en el ícono de la llave (Secrets).
- Agrega una nueva variable con el nombre: API_KEY, coloca la clave en el campo "VALUE" y activa el interruptor de permisos.
- Ejecutar: En el menú "Entorno de ejecución" > "Ejecutar todas".

- API_KEY - Empleada en el desarrollo del proyecto

  E53&\K21If4h

Anexo 2. Descripción de los datos

Descripción de las variables

Tabla 11

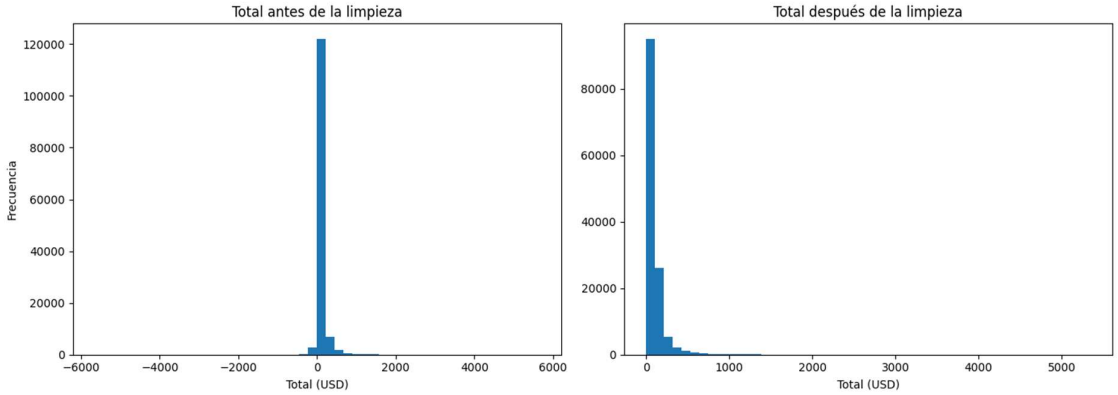
Atributos de la tabla de Ventas

Atributos	Tipo	Descripción
numero_orden	Entero	Código único de cada pedido
cliente_id	Categórica	Identificador del cliente
cliente	Texto	Nombre del cliente
fecha_vuelo	Fecha	Fecha de embarque/logística
pais / estado / ciudad	Categóricas	Ubicación del cliente
estado_orden	Categórica	Completado, cancelado, pendiente
vendedor	Categórica	Nombre del ejecutivo
producto_id	Categórica	Identificador del producto
producto	Texto	Descripción del producto
tallos	Numérica	Cantidad de tallos por orden
precio_unitario	Numérica	Precio por tallo
total	Numérica	Precio total (tallos × precio)
largo	Numérica	Largo del tallo (40–90 cm)
agencia	Categórica	Courier o transporte
origen_cliente	Categórica	Canal de entrada
usuario_id	Categórica	Usuario que procesó la orden

Anexo 3. Figuras del procesamiento y análisis exploratorio

Figura 21

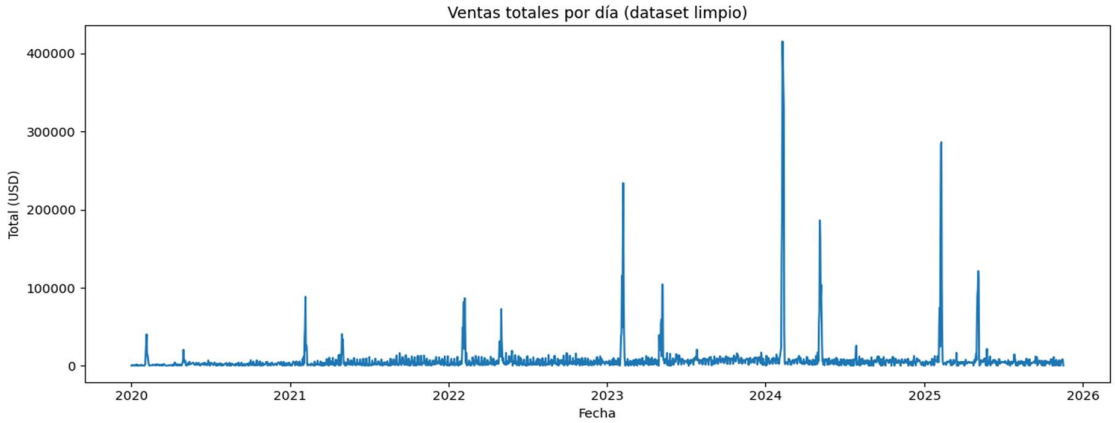
Comparación de histogramas del valor total antes y después de la limpieza.



Nota. La figura muestra la distribución del total monetario en dos etapas del procesamiento: antes de la limpieza y después de la limpieza, evidenciando el impacto del tratamiento de datos en la reducción de valores atípicos y errores.

Figura 22

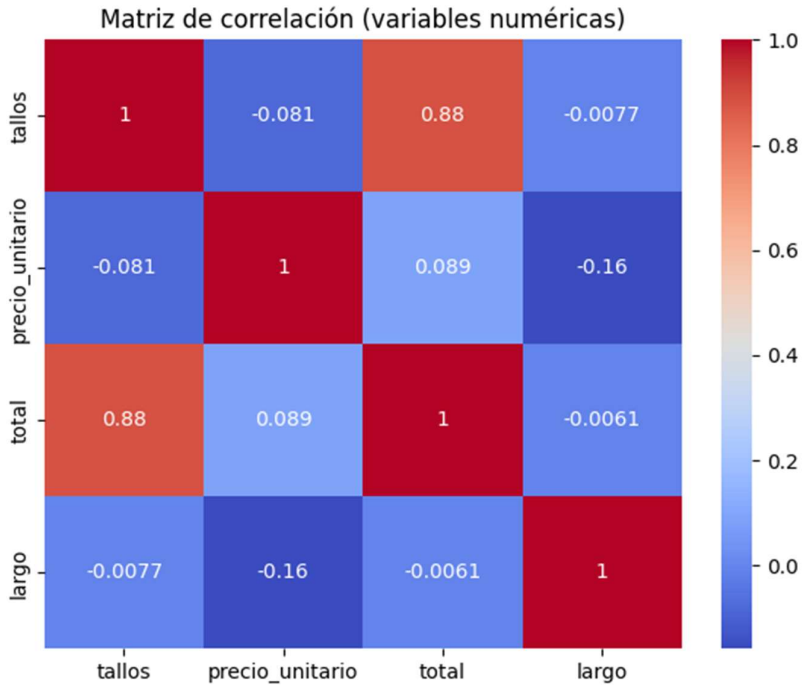
Serie temporal de ventas diarias en el periodo de estudio (dataset limpio).



Nota. La figura muestra la evolución de las ventas diarias tras el proceso de limpieza de datos, permitiendo observar picos de venta y patrones temporales relevantes.

Figura 23

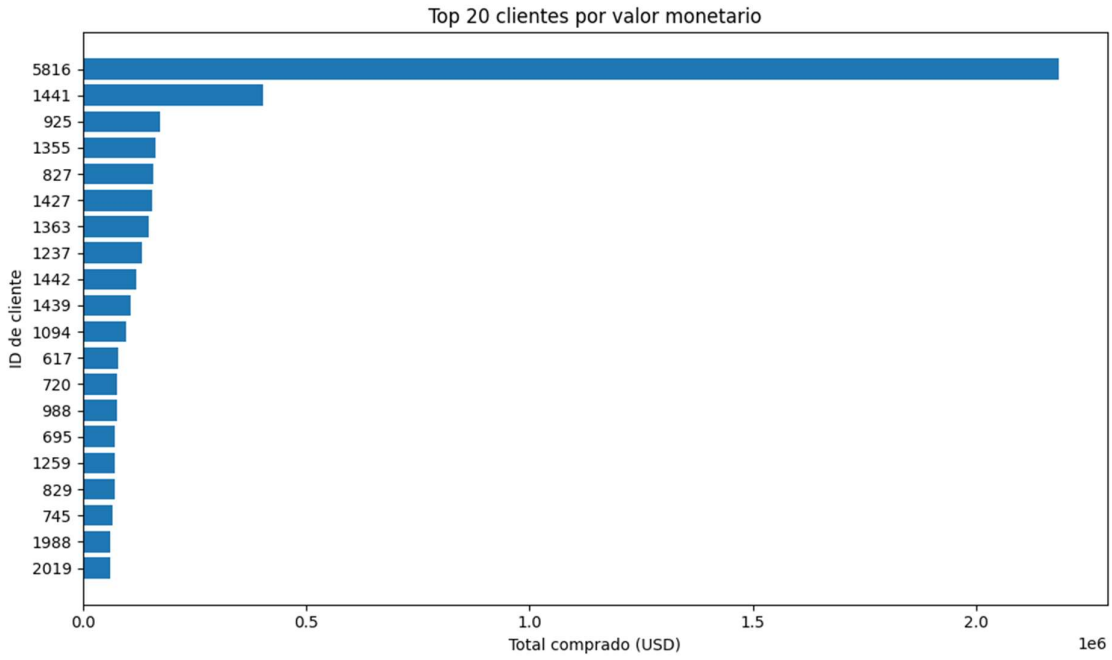
Matriz de correlación entre variables numéricas (tallos, precio unitario, total, largo).



Nota. La figura presenta la matriz de correlación entre las variables numéricas del dataset, utilizando una escala de color que permite identificar rápidamente la intensidad y dirección de las asociaciones.

Figura 24

Top 20 clientes por valor monetario en el periodo de estudio.



Nota. La figura muestra a los 20 clientes con mayor valor monetario total durante el período analizado, ordenados de mayor a menor según el monto acumulado de compra.

Anexo 4. Herramienta BI - dashboard del proyecto

El dashboard desarrollado en Power BI Desktop fue publicado en la versión online de la plataforma y habilitado para su visualización desde una página web, que a su vez fue generada para este proyecto. El acceso se realiza mediante un enlace web, ingresando el usuario y la contraseña indicados a continuación.

- Enlace de acceso

La página web es de acceso público y se encuentra disponible en la siguiente dirección URL: http://maestria.optimusec.com

- Acceso

El usuario para ingresar es: maestria y la clave de acceso: E53&\K21If4h