Maestría en

CIENCIA DE DATOS Y MÁQUINAS DE APRENDIZAJE CON MENCIÓN EN INTELIGENCIA ARTIFICIAL

Trabajo previo a la obtención de título de Magister en Ciencia de Datos y Máquinas de Aprendizaje con mención en Inteligencia Artificial

AUTORES:

ALVAREZ MENA PABLO ANDRES

MANTILLA BOLAÑOS ELVIS DAVID

MARIN NICOLALDE JERSON ESTEBAN

TRUJILLO SARZOSA IVAN ENRIQUE

TUTORES:

Alejandro Cortés, M.Sc.

Iván Reyes Chacón, Mgtr.

Segmentación de clientes de una empresa ecuatoriana de pulpa de fruta usando K-Means

Quito – Ecuador

julio-2025

REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, **Pablo Andrés Alvarez Mena, Elvis David Mantilla Bolaños, Iván Enrique Trujillo Sarzosa, Jerson Esteban Marín Nicolalde** 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.

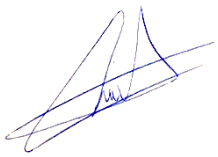
Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma del graduando

Pablo Andrés Alvarez Mena

Firma del graduando

Elvis David Mantilla Bolaños



Firma del graduando

Ivan Enrique Trujillo Sarzosa

Firma del graduando

Jerson Esteban Marin Nicolalde

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Pablo Andrés Alvarez Mena, Elvis David Mantilla Bolaños, Ivan Enrique Trujillo Sarzosa y Jerson Esteban Marín Nicolalde** en calidad de autores del trabajo de investigación titulado **Segmentación de clientes de una empresa ecuatoriana de pulpa de fruta usando K-Means**, 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.

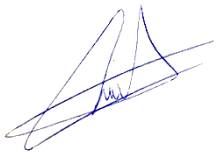
D. M. Quito, 24 de julio 2025

Firma del graduando

Pablo Andrés Alvarez Mena

Firma del graduando

Elvis David Mantilla Bolaños



Firma del graduando

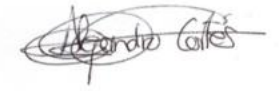
Ivan Enrique Trujillo Sarzosa

Firma del graduando

Jerson Esteban Marin Nicolalde

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés Msc.** e **Iván Reyes Ch. Mgtr.**, declaramos que: Pablo Andrés Alvarez Mena, Elvis David Mantilla Bolaños, Iván Enrique Trujillo Sarzosa y Jerson Esteban Marín Nicolalde son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés Msc.

Director de la Maestría en Ciencia de datos y Maquinas de Aprendizaje mención Inteligencia Artificial EIG

Iván Reyes Ch. Mgtr.

Coordinador de Posgrados Escuela de Ciencia de la Computación UIDE

DEDICATORIA

A mis padres por su ejemplo de vida, esfuerzo y apoyo

A mi esposa por su fuerza y dedicación

A mis hijas que son mi motivo de avanzar

Pablo Andrés Alvarez Mena

Dedico el presente trabajo de titulación, en primer lugar, a mi familia, quienes han sido un pilar fundamental en cada etapa de mi vida. En especial, a mi madre Silvana Bolaños, quien, con esfuerzo, sacrificio, amor y una inmensa tolerancia, me ha guiado y apoyado incondicionalmente para alcanzar cada uno de mis objetivos. De igual manera, extiendo esta dedicatoria a mi querida hermana Odalis Mantilla, por ser siempre una fuente de apoyo constante, motivación y confianza. A ambas les atribuyo no solo el profesional en el que me he convertido, sino, sobre todo, la persona que soy hoy.

Elvis David Mantilla Bolaños

Dedico este trabajo a mi mamá y a mi papá por su motivación para completar mis metas.

A mis hermanos por su compañía, por animarme en los momentos difíciles y celebrar cada logro.

A mi tía por su apoyo en cada etapa de mi vida, Gracias por estar ahí.

Ivan Enrique Trujillo Sarzosa

Dedico este trabajo a mi familia, especialmente a mi hermano mayor Eduardo por haberme guiado dentro del mundo de la ciencia de datos. A mi amada esposa Kate y mi hija Romy por todo su amor, a mis padres Eduardo y Flor y mi hermano Erick por su inquebrantable apoyo en todas las instancias de mi vida.

Jerson Esteban Marín Nicolalde

AGRADECIMIENTOS

A mi familia, por aguantar los sacrificios del trabajo y estudios, a mis padres por su constante apoyo en todos los momentos de mi vida. A los tutores de la maestría, quienes han promovido un aprendizaje aplicado en ciencia de datos.

Pablo Andrés Alvarez Mena

A mi familia, por ser parte fundamental de mi crecimiento profesional y personal. En especial, a mi madre Silvana Bolaños y a mi hermana, quienes con su acompañamiento y amor incondicional han sido una fuente constante de motivación para mi superación.

A mis compañeros de proyecto, con quienes he compartido esta valiosa etapa de aprendizaje y desarrollo. Agradezco su compromiso, guía y la disposición para compartir conocimientos, lo cual enriqueció profundamente esta experiencia académica.

Elvis David Mantilla Bolaños

A mi familia que siempre me ha acompañado y apoyado en mi desarrollo personal y profesional, a mi Mamá por siempre motivarme a salir adelante.

A mis compañeros por compartir su conocimiento y acompañarme durante el desarrollo de la maestría.

Ivan Enrique Trujillo Sarzosa

Agradezco a mi querida esposa Kate por estar presente incondicionalmente durante todo este proceso apoyándome, a mi hija Romy por inspirarme a seguir adelante cada día y a mi familia, a mis padres y hermanos Flor, Eduardo, Edu y Erick por ser el motor de mi vida.

Jerson Esteban Marín Nicolalde

RESUMEN

El propósito de este proyecto de titulación fue aplicar métodos de segmentación de clientes mediante el algoritmo K-means, y compararlo con un método clásico de segmentación, denominado RFM, que se basa en tres métricas del comportamiento de compra de un cliente: recencia (R), frecuencia (F) y valor monetario (M).

Estos métodos fueron aplicados sobre información histórica de ventas de la empresa Proalva, que produce y comercializa pulpa de fruta congelada en Quito, Ecuador.

Se realizó una evaluación de las variables disponibles, análisis exploratorio de datos y un proceso de limpieza y depuración. Posteriormente se generó un subconjunto de datos, en los que se agregó la información del comportamiento de compras a nivel de cliente, destacándose un total de 340 clientes que fueron clasificados mediante los dos métodos.

El estudio se enfocó en un periodo de 24 meses. Se determinaron métricas de comportamiento para cada cliente, tales como R, F, M, volumen de producto, diversidad de productos y cantidad de presentaciones adquiridas.

A partir del modelo RFM se generó una categorización de clientes en seis grupos, utilizando rangos ordinales establecidos por quintiles de las tres variables (R, F y M). Este método presentó limitaciones importantes vinculadas a su carácter discrecional y a la omisión de otras dimensiones que caracterizan el comportamiento de compra.

La segmentación mediante K-Means utilizó seis variables. La elección del número ideal de grupos se llevó a cabo a través del método del codo y la evaluación de silueta, concluyendo que cinco grupos proporcionaban un adecuado balance entre la cohesión interna y la separación entre grupos. El estudio demostró que K-Means pudo detectar patrones más sofisticados e identificar grupos de clientes que el modelo RFM agrupaba junto a perfiles de distinta importancia.

Palabras Claves: Segmentación de clientes, Modelo RFM, K-means

ABSTRACT

The purpose of this thesis project was to apply customer segmentation methods using the K-Means algorithm and compare it with a classical segmentation approach known as RFM. This method is based on three metrics of customer purchasing behavior: Recency (R), Frequency (F), and Monetary value (M).

These methods were applied to historical sales data from Proalva, a company that produces and markets frozen fruit pulp in Quito, Ecuador.

An evaluation of the available variables was conducted, along with exploratory data analysis and a data cleaning process. Subsequently, a subset of data was generated, incorporating customer-level purchase behavior. A total of 340 customers were classified using both methods.

The study focused on a 24-month period. Behavioral metrics were determined for each customer, including R, F, M, product volume, product diversity, and the number of packaging formats purchased.

Based on the RFM model, customers were categorized into six groups using ordinal ranges derived from quintiles of the three variables (R, F, and M). This method presented significant limitations related to its discretionary nature and the omission of other dimensions that characterize purchasing behavior.

Segmentation using K-Means incorporated six variables. The optimal number of groups was determined through the elbow method and silhouette analysis, concluding that five groups provided an appropriate balance between internal cohesion and group separation.

The study showed that K-Means was able to detect more sophisticated patterns and identify customer segments that the RFM model grouped together despite having different levels of importance.

Keywords: Customer segmentation, RFM model, K-Means

Tabla De Contenidos

Lista De Tablas

Lista De Figuras

Capítulo 1: Introducción

Antecedentes

En Ecuador, las micro, pequeñas y medianas empresas (MIPYME) representan un papel importante en la creación de empleo y las actividades productivas en general. De acuerdo con el Registro Estadístico de Empresas, en el país se registran 863.681 empresas activas, de las cuales el 94 % son microempresas, el 4 % pequeñas y cerca de 0.5 % grandes empresas (INEC, 2023). El sector manufacturero, que genera valor agregado a la producción primaria, representó el 12.6 % del empleo formal y generó el 24 % del total de ventas registradas ante el Servicio de Rentas Internas (SRI) año (INEC, 2023).

Pese a su importancia, la mayor parte de las MIPYMEs se ven limitadas por varios obstáculos en el desarrollo de estrategias comerciales basadas en evidencia. Los principales incluyen escasos conocimientos o herramientas para realizar análisis de datos, limitada cultura de uso de datos y falta de conocimientos sobre metodologías que permitan la clasificación y comprensión de las relaciones comerciales con sus clientes (Uquillas Granizo, 2024; Valencia, 2024). Esta realidad restringe la habilidad para diseñar estrategias de mercadeo fundamentadas en patrones observables del comportamiento de compra, repercutiendo negativamente en la competitividad y eficiencia en el uso de los recursos (Uquillas Granizo, 2024).

La falta de procedimientos predefinidos para segmentar a los clientes basándose en su comportamiento de compra dificulta que las MIPYMEs puedan distinguir entre consumidores de alto valor y los de menor influencia comercial.

El uso de modelos de segmentación, que usan variables provenientes de los sistemas regulares de contabilidad o de monitoreo de transacciones de ventas, permite reconocer perfiles específicos en la cartera de clientes de una empresa (Clarke & Freytag, 2023). Uno de los métodos más usados es el modelo RFM (Recency, Frequency, Monetary), el cual posibilita valorar de manera cuantitativa el comportamiento de los clientes

y categorizarlos en función a su comportamiento de recencia, frecuencia de transacción y valor monetario acumulado (Ramkumar et al., 2025a).

Otros métodos incluyen el uso de técnicas de aprendizaje no supervisado, como el clustering con K-means, pues permiten identificar agrupaciones desde una óptica no lineal, a diferencia de RFM, y permite también la incorporación de otras métricas que caracterizan al cliente o su comportamiento comercial. Existen varios estudios de segmentación de clientes que comparan o complementan los ejercicios de agrupación con ambos modelos (Christy et al., 2021a; Ramkumar et al., 2025a; Shirole et al., 2021).

Investigaciones previas en el sector alimentario muestran que aplicar técnicas de segmentación de clientes, que pueden ser basadas en la sensibilidad al precio o las preferencias de consumo, contribuye a la preparación de estrategias de marketing efectivas (Kotler & Keller, 2016a).

Información general de la empresa

Proalva es una empresa ecuatoriana familiar, con sede en Quito, categorizada como pequeña empresa por el SRI. La empresa fue fundada en el año 2002, se especializa en la producción y comercialización de pulpa de fruta congelada y otros productos derivados del sector agroindustrial, provenientes de frutas y vegetales. A lo largo de los años, ha logrado consolidarse en el mercado, principalmente en Quito, y sus principales clientes se encuentran en el segmento de hoteles, restaurantes y servicios de catering (horeca).

La empresa se encuentra en proceso de elaboración de un plan de marketing y ventas para el mercado local, lo que abre la posibilidad de incorporar técnicas de segmentación de clientes. Esta práctica facilitará la identificación de grupos con características similares, permitiendo así adaptar las campañas publicitarias a los perfiles específicos de cada grupo de clientes. Además, brindará una mayor precisión en la estimación de los presupuestos requeridos, al definir con claridad el público objetivo.

Definición del proyecto

Este proyecto de titulación tiene como objetivo aplicar técnicas de aprendizaje no supervisado para segmentar la cartera de clientes de una empresa dedicada a la producción y comercialización de pulpa de fruta congelada, utilizando datos históricos de ventas generados por sus sistemas administrativos y contables. La segmentación obtenida permitió establecer perfiles diferenciados de clientes, sobre los cuales se podrán diseñar estrategias comerciales y de marketing focalizadas.

La investigación se centra en la implementación del algoritmo K-means para la clasificación de clientes, a partir de variables derivadas del comportamiento de compra, además, se complementa el análisis con una segmentación usando métodos más clásicos como el modelo RFM (Recency, Frequency, Monetary).

El desarrollo metodológico del proyecto incorpora herramientas de aprendizaje automático y minería de datos, cuya aplicación es posible por el acceso al conjunto de datos de ventas, y que, por fines normativos, podría estar disponible en la mayoría de MIPYMEs.

Justificación e importancia del trabajo de investigación

Muchas PYMES en Ecuador se constituyeron y crecieron en distintos entornos, caracterizados por una limitada estructura en el manejo y madurez de sus datos, lo que dificultó el desarrollo de estrategias comerciales basadas en evidencia y derivó en una falta de segmentación de clientes o a su vez en una segmentación ineficaz (Armijos De La Cruz & Palacios Meléndez, 2024; Gabriela Jara & Sayonara Solorzano, 2024). Esta carencia limita su capacidad para mejorar la competitividad y el crecimiento sostenido en el mercado (Fredyna et al., 2019).

La aplicación de un método combinado de segmentación RFM y agrupación no supervisada no solo ofrece una perspectiva más integral del comportamiento de los clientes, sino que también ofrece una estructura de decisión que puede ser replicada en otras MIPYMEs con circunstancias parecidas. La presencia de datos estructurados en sistemas de gestión y el acceso a herramientas de análisis de libre uso son factores propicios para la

expansión de estas prácticas en empresas de pequeño y mediano tamaño en Ecuador (Tavakoli et al., 2018).

Alcance

El proyecto de titulación utiliza técnicas de minería de datos para el análisis y preparación de datos, aplica un modelo de segmentación clásico y utiliza técnicas de aprendizaje no supervisado para la segmentación de clientes desde la ciencia de datos, partiendo de un dataset de ventas no preparado, proveniente de los sistemas administrativos de una MIPYME ecuatoriana.

Objetivos

Objetivo General

Desarrollar un modelo de segmentación de clientes utilizando el algoritmo de aprendizaje no supervisado K-Means y el modelo RFM, a partir de los datos de facturación de una empresa ecuatoriana de pulpa de fruta congelada, con el fin de identificar grupos específicos de clientes que permitan orientar estrategias de marketing diferenciadas.

Objetivos específicos

- Realizar un análisis exploratorio y proceso de limpieza de los datos de la empresa.
- Calcular las métricas RFM (recencia, frecuencia y valor monetario) para cada cliente y clasificar segmentos.
- Aplicar el algoritmo K-means sobre variables de comportamiento de compra del cliente y evaluar la calidad de los grupos obtenidos.
- Comparar los segmentos obtenidos con ambos enfoques para identificar similitudes y complementariedades que puedan informar el diseño de acciones comerciales diferenciadas.
- Perfilar los clientes de la empresa en función a los clusters de clientes obtenidos por k-means y diseñar posibles estrategias de marketing.

Capítulo 2: Revisión De Literatura

Estado Del Arte

Segmentación de clientes en el contexto empresarial actual

La segmentación de clientes ha experimentado una importante evolución, desde modelos centrados en variables sociodemográficas hasta otros basados en el comportamiento de las transacciones que se producen en el marco de las relaciones comerciales. En este sentido, el estudio de las transacciones comerciales ha cobrado importancia por ser el mecanismo que permite capturar de forma directa el valor que el cliente aporta a una empresa, particularmente en los mercados en los que existe una alta concentración de ventas de manera recurrente (V. Kumar & Reinartz, 2018). Este enfoque ha sido aplicado de forma cada vez más creciente en aquellas industrias en las que hay ciclos de compra claros, como el comercio, el sector de la alimentación o aquéllos que se centran en la venta personalizada.

En el caso de las micro, pequeñas y medianas empresas (MIPYMEs) de los países en desarrollo, existen múltiples barreras de acceso a técnicas de segmentación de clientes basadas en datos, como son el bajo nivel de madurez en el manejo de los sistemas de información, la limitada capacitación técnica de las personas en contacto con los clientes o la ausencia de estrategias comerciales que estén, de forma efectiva, fundamentadas en la evidencia empírica (Laudon & Laudon, 2020). Así, muchas MIPYMEs desarrollan acciones de marketing basadas en estrategias generales o intuitivas y son incapaces de poder identificar de manera adecuada, previo a la implementación de estas campañas, las distintas tipologías de clientes y sus necesidades específicas (Dini et al., 2021; Martínez et al., 2024).

En este sentido, el análisis del comportamiento del cliente ha podido ser una opción válida para poder llegar a superar las limitaciones que se han presentado. Las técnicas como la analítica del modelo RFM o la aplicación de algoritmos de clustering pueden llegar a generar conocimiento aplicable a partir de registros administrativos que son accesibles

desde sistemas de facturación o ERP. De tal manera que su aplicación se ha mostrado como una práctica habitual en los entornos de menos sofisticación tecnológica, probablemente gracias al uso de software de código abierto y la existencia de metodologías replicables con bajo coste computacional (Hiziroglu, 2013; Tavakoli et al., 2018).

Aplicaciones recientes de RFM y K-means en diferentes sectores

La combinación del modelo RFM y técnicas de clustering como el algoritmo K-means permite la segmentación de clientes a partir de información de transacciones, tal y como han demostrado algunos estudios. Este enfoque ha sido aplicado en sectores como el comercio electrónico (Alves Gomes & Meisen, 2023; Wulansari & Heikal, 2024), el turismo, o el sector de alimentos y bebidas (Sarkar et al., 2024a), gracias a su habilidad para determinar grupos homogéneos de clientes con diferente comportamiento de consumo.

Sarkar et al. (2024), estudió una base de datos de un supermercado de Bangladesh para mostrar cómo RFM puede codificar a los clientes y cómo el algoritmo K-means los puede identificar en clusters. La investigación concluyó tras identificar cinco segmentos claramente diferenciados con respuestas diferentes a las promociones. También, Christy et al. (2021) llevaron a cabo un estudio en el que K-means se asoció a variables RFM y de comportamiento de navegación en una tienda de e-commerce, teniendo como resultado mejoras significativas en cuanto a la tasa de conversión de las campañas de marketing de la tienda, gracias a la adaptación llevada a cabo entre clientes segmentados (Christy et al., 2021b).

En América Latina, las aplicaciones empíricas de segmentación derivadas de esa combinación han empezado a documentarse, aunque siguen siendo escasas. Un ejemplo es el trabajo de Cuadrados López et al. (2017), que llevó a cabo la segmentación de una empresa dedicada a los productos naturales levantando RFM e incluyendo clustering jerárquico. La segmentación concluyó al identificar una serie de patrones de compra estacionales, patrones de recompra en función de promociones mensuales, que fueron determinantes para rediseñar el plan de fidelización de la empresa.

En la tesis “Segmentación de Clientes automatiza a partir de técnicas de minería de datos (K-means clustering)” el autor plantea la segmentación de clientes de la empresa Tiendacol S.A (empresa de moda colombiana), por el método K-means asociando los datos poblacionales y sus hábitos de adquisición de productos, identificando tres clusters: mejor cluster, cluster medio y peor cluster. La diferencia entre la población de cada cluster radica en la frecuencia de adquisición de productos, la cantidad de producto que adquieren y los créditos que poseen (Cálad Noreña, 2015a).

En la tesis “Aplicación de Minería de datos para la segmentación de clientes que compran materias primas derivadas del Maíz para la generación de estrategias de comunicación” los autores proponen agrupar clientes por tendencias de compra usando el método no supervisado K-means. Dando como resultado cinco clusters, segmento minorista, margarinero, Industria de alimentos, laboratorios y belleza. La diferencia entre los clusters radica en los productos que adquieren y su finalidad (Plazas Cárdenas & Plazas Cárdenas, 2013).

Asimismo, han sido exploradas algunas variantes del enfoque clásico, como puede ser el uso de RFM fuzzy o adaptar pesos dinámicos en función del sector en el que se usan, así como los canales de comercialización, para conseguir captar la heterogeneidad de los clientes. Estos modelos aportan a reducir las limitaciones del modelo tradicional RFM, siendo los más relevantes la sensibilidad a valores extremos en su análisis y la incapacidad para incorporar otras variables de comportamiento (Rungruang et al., 2024).

Gap en la literatura aplicado al caso Proalva

A pesar del avance que se ha llegado a observar en la literatura, se hacen evidentes vacíos que justifican la realización de estudios aplicados al contexto ecuatoriano. Hasta la fecha, no se han encontrado trabajos de investigación publicados que recojan la aplicación de técnicas de segmentación combinadas RFM y K-means en el ámbito del sector agroindustrial del país, sobre todo en compañías con un enfoque B2B como la de Proalva.

Este aspecto impide replicar buenas prácticas de sectores donde los patrones de consumo de los clientes son distintos a los de las prácticas del retail tradicional.

Por otro lado, un número bajo de trabajos han utilizado variables adicionales a las que capta el marco RFM, que hagan referencia a la variedad del portafolio de productos comprados por un cliente o la dispersión temporal de las transacciones. Disciplinas que son especialmente relevantes en sectores como la agroindustria, donde la variabilidad del producto, la presentación o la cadencia de comercialización pueden tener implicaciones en la gestión de la demanda o en la planificación de la producción.

El presente proyecto se plantea ayudar a cerrar parcialmente esta brecha, aportando evidencia empírica de un enfoque metodológico replicable y adecuado al contexto de las MIPYMEs ecuatorianas y sustentado en herramientas de análisis de datos accesibles para su implementación.

Marco Teórico

Fundamentos de la segmentación de clientes

El proceso de segmentación de clientes es un proceso no sólo analítico sino también de carácter estratégico, que consiste en dividir clientes heterogéneos en grupos homogéneos, para atender de modo diferente a sus características, comportamientos o necesidades específicas (Wedel & Kamakura, 2000), esta es una práctica fundamental en la gestión comercial moderna, para utilizar de forma eficiente los recursos, intensificar las acciones de marketing o maximizar el valor de la transacción (Kotler & Keller, 2016b).

Históricamente, son cuatro las aproximaciones que se han seguido para la segmentación de clientes: de carácter demográfico, geográfico, psicográfico y conductual (Dolnicar et al., 2018; Mora Cortez et al., 2021; Wedel & Kamakura, 2000). Las de carácter demográfico y geográfico son las más clásicas constituidas a partir de variables como edad, sexo, ingresos, ubicación geográfica, etc., aunque han mostrado limitaciones en un contexto en el que las decisiones para la compra las condicionan más el comportamiento que hacer referencia a características observables de la individualidad de cada cliente (Wedel &

Kamakura, 2000). La segmentación conductual o basada en el comportamiento se ha convertido en la alternativa de mayor aceptación en los últimos años, sobre todo, por la creciente disponibilidad de datos tanto transaccionales como digitales. La caracterización de este tipo de segmentación está basada en el comportamiento de los clientes en la interacción con la empresa: cómo compran, con qué frecuencia, cuánto gastan, en qué canal o a qué hora. Los análisis propuestos dan cuenta de una representación más apropiada del valor real de cada cliente cuando se relaciona al negocio (V. Kumar & Reinartz, 2018).

No se puede escoger el tipo de segmentación de forma arbitraria, sino que debe corresponder a la naturaleza del mercado, la disponibilidad de datos, la complejidad de la oferta y la organización operativa de la empresa. Aquellos contextos en los que existen ventas recurrentes y relaciones con los clientes en el mediano plazo, como las organizaciones B2B del ámbito alimentario, el análisis de comportamiento transaccional construye perfiles más estables y orientados hacia acciones empíricas de marketing (Tsiptsis & Chorianopoulos, 2010).

La segmentación no debe ser considerada como un esquema estático, sino que debe ser revisada o reformulada periódicamente, para poder captar los cambios en la dinámica comercial de la empresa. Numerosas organizaciones utilizan modelos que son dinámicos o consideran variables contextuales y temporales en sus prácticas (Wang et al., 2024).

Respecto a las MIPYMEs, si bien la segmentación ha sido considerada históricamente una práctica ausente, múltiples trabajos evidencian que su presencia activa además de mejorar la rentabilidad, potencian la fidelización y permiten generar propuestas de valor más específicas a los diversos tipos de cliente, incluso con unos recursos analíticos limitados (Pinho & Prange, 2016).

Modelos clásicos como RFM, y algoritmos de clustering tienen una ventaja en términos de simplicidad y accesibilidad, pues existe la posibilidad de convertir datos

administrativos en información utilizable para diseñar estrategias comerciales (Christy et al., 2021b; Rahim et al., 2021; Safari et al., 2016).

Modelo RFM: fundamentos y aplicaciones

El modelo RFM (Recency, Frequency, Monetary) constituye una técnica muy reconocida de segmentación de clientes con arreglo al análisis de las transacciones comerciales realizadas por los mismos. Este modelo fue propuesto por primera vez para el marketing directo y ha sido adaptado para diversos mercados por su sencillez, efectividad y fácil generación de perfiles comportamentales de los consumidores (Anitha & Patil, 2022; Christy et al., 2021b; Ramkumar et al., 2025b).

Este modelo parte de la hipótesis de que son mucho más propensos a responder positivamente a futuras acciones de marketing los clientes que han comprado hace poco, que lo hacen con más frecuencia y que han gastado más dinero (Rahim et al., 2021). En este sentido, el modelo describe el valor de cada cliente a partir de las tres dimensiones observables en los datos históricos de compra:

- Recency (R): tiempo desde la última compra del cliente. Se comprende que cuanto menor sea el valor, más probable será la compra.
- Frequency (F): número total de transacciones que realiza un cliente en un periodo determinado. La frecuencia es directamente proporcional a la fidelidad o hábito de compra de un determinado cliente.
- Monetary (M): valor total de los gastos que realiza un cliente sobre las transacciones realizadas en el mismo periodo. Refleja el impacto o la importancia económica que ejerce cada cliente sobre la empresa.

La combinación de estos tres indicadores permite construir un perfil numérico o score compuesto que sintetiza la importancia de cada cliente para el negocio. En su aplicación clásica, cada una de las variables se agrupa en quintiles, deciles u otros intervalos definidos, resultando en una escala ordinal que puede ser posteriormente

utilizada para poder agrupar a los clientes en segmentos con determinadas características (Rajagukguk, 2025).

El atractivo de la metodología RFM radica en su escasa exigencia técnica, así como en la rapidez de su implementación a partir de técnicas estadísticas básicas o lenguajes de programación como Python, R o incluso Excel. No necesita variables de tipo demográfico ni contextual, hecho que lo hace especialmente adecuado a aquellas empresas que únicamente poseen histórico de transacciones (Sarkar et al., 2024b; Tavakoli et al., 2018).

En diversos estudios se han puesto en evidencia limitaciones importantes de este modelo. En primer lugar, la selección de umbrales o rangos suele ser libre y arbitraria, lo que lleva a sesgos subjetivos que se trasladan a la segmentación. En segundo lugar, el modelo no contempla relaciones no lineales ni interacciones entre las variables, lo cual puede limitar su capacidad para describir patrones complejos de comportamiento (Wang et al., 2024).

Para subsanar estas limitaciones, se han propuesto variantes metodológicas. Una de ellas es la extensión de modelo RFM fuzzy, que presenta lógica difusa para describir patrones característicos de cada cliente en relación a criterios R, F y a lo largo de la historia, lo cual permite una descripción más flexible a la asignación de grupos (Wang et al., 2024). Otras alternativas incluyen la utilización de pesos adaptables en función del tipo de mercado o canal de comercialización (Rungruang et al., 2024; Yahya et al., 2024).

El modelo RFM permite considerar una primera segmentación de clientes a partir de datos históricos del comportamiento básico de compra, sin embargo, la efectividad del modelo puede ir más allá de este uso cuando se combina con técnicas de clustering, ya que, a partir de datos del perfil del cliente, se pueden explorar otras variables que pueden aportar a una mejor segmentación, o a su vez considere también la interacción entre las variables, sin la necesidad de establecer manualmente umbrales.

Aprendizaje no supervisado y clustering

El aprendizaje no supervisado constituye una subcategoría del aprendizaje automático (machine learning) cuyo objetivo principal es encontrar estructuras ocultas o patrones en conjuntos de datos que no poseen etiquetas o clasificaciones predefinidas (Everitt et al., 2011a; Jain, 2010). El aprendizaje no supervisado usa un conjunto de observaciones en las que no hay un dataset de entrenamiento, que contenga valores de salida previamente conocidos, por lo que se genera exclusivamente en el entorno de las características observadas, buscando relaciones de forma implícita en el conjunto de variables. (Everitt et al., 2011a)

Dentro del conjunto de técnicas más utilizadas en este modelo se encuentran los algoritmos de agrupamiento o clustering, que constituyen una técnica para dividir el conjunto de observaciones en subconjuntos (clusters) de manera que los elementos que aparecen en un mismo grupo son miembros similares y, a la vez, son diferentes de los que forman parte de otros grupos (Jain, 2010). La noción de similitud está dada muchas veces por una medida de distancia, en la cual la distancia euclídea es una de las más empleadas.

Los algoritmos de clustering han alcanzado mucha popularidad, son utilizados en áreas como el marketing, biología, tratamiento de imágenes, análisis de textos, y cada vez más en la ciencia de los datos para el sector empresarial. En concreto, en ámbitos comerciales, las técnicas de clustering permiten identificar perfiles homogéneos a partir de grandes cantidades de datos, y sin la necesidad de categorizaciones previas (Jain, 2010).

Existen múltiples enfoques de clustering, que pueden agruparse en cuatro tipos principales:

- Clustering de partición: divide los datos en un número de grupos fijo. El ejemplo clásico es el método K-means.
- Clustering jerárquico: a partir de un conjunto de agrupaciones iniciales, va añadiendo sus fusiones o divisiones de forma sucesiva hasta alcanzar el número de agrupaciones deseadas.

- Modelos basados en densidad: este tipo, como el DBSCAN, son aquellos que determinan regiones donde la densidad se incrementa formando clusters de forma arbitraria.
- Modelos probabilísticos: como los modelos de mezcla gaussiana GMM, que buscan la estimación de la distribución estadística que subyace a cada cluster.

Cada enfoque tiene sus ventajas y desventajas. Por ejemplo K-means es eficiente en grandes volúmenes de datos y fácil de interpretar, pero requiere definir a priori el número de clusters y es sensible a la escala de las variables (López, 2024). El DBSCAN, por el contrario, no requiere un número de grupos predefinidos, pero su desempeño se deteriora para datasets con densidades muy diferentes. Esto implica que la elección del algoritmo a utilizar debe considerar los datos, la escala, la dispersión y la finalidad del análisis (Everitt et al., 2011a).

En contextos empresariales como la segmentación de clientes, las técnicas de clustering tiene un valor estratégico, al permitir una clasificación de los consumidores de manera empírica, a partir de su propio comportamiento, sin imponer restricciones de segmentación basadas en criterios previos o arbitrarios. Esta capacidad de encontrar estructuras latentes ha sido destacada en diferentes trabajos como una forma de identificar las oportunidades comerciales, optimizar las campañas de marketing y definir mejores ofertas de valor (Cálad Noreña, 2015b).

Un punto importante con respecto al uso de estas técnicas es la preparación de datos, incluyendo la estandarización de las variables, el tratamiento de los valores atípicos y la elección de las variables de interés, ya que eso puede tener una repercusión directa en el resultado de la agrupación (Tsiptsis Antonios Chorianopoulos WILEY, n.d.; Tsiptsis & Chorianopoulos, 2010). El escalamiento de las variables es un paso importante en algoritmos como el K-means, ya que la distancia euclidiana se ve afectada por las unidades y los rangos de cada una de las variables.

El aprendizaje no supervisado y las técnicas de clustering son una herramienta muy potente para descubrir de una manera empírica segmentos de clientes, y permitir captar patrones complejos y en múltiples dimensiones de los datos, lo que viene a complementar a la perfección, modelos más rígidos como el RFM.

Algoritmo K-Means

El algoritmo K-means es probablemente la técnica de agrupamiento que más se usa hoy en día en el campo del aprendizaje no supervisado por su rapidez de computación, su sencillez de implementación en la práctica y su solidez al dividir conjuntos de datos numéricos en clústeres homogéneos (Jain, 2010). El objetivo que persigue el K-means consiste en dividir un conjunto de observaciones en k clústeres de forma que se minimice la distancia total en los grupos entre cada observación y cada centroide correspondiente.

La base del algoritmo es el uso iterativo de dos pasos importantes correspondientes a (i) la asignación de los puntos al centroide más cercano en función de la distancia euclidiana y (ii) la actualización del centroide de cada clúster como el punto medio de todos los puntos que han sido asignados a él (Lloyd, 1982). Este proceso se repite hasta la convergencia, ejecutándose hasta que las asignaciones de los puntos no cambian de forma perceptible entre cada iteración. Este criterio de optimización indica que los clusters resultantes serán esféricos, y de dimensiones parecidas, lo que supone una limitante en los contextos donde la estructura es más compleja (López, 2024; Zhao, 2024).

En síntesis, dado un conjunto de datos $X = \{x_1, x_2, \ldots x_n\}$, donde cada x_i es un punto en un espacio de dimensión *d*, el algoritmo sigue los siguientes pasos:

- Selección aleatoria de *k* centroides iniciales.
- Asignación de cada punto de datos al centroide más cercano, según la distancia euclidiana.
- Recalcular los centroides como el promedio de los puntos asignados a cada grupo.
- Repetición de los pasos anteriores hasta que las asignaciones no cambien o el cambio sea mínimo.

El objetivo del algoritmo es minimizar la siguiente función:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - \mu_j \right\|^2$$

Donde:

C_j representa el conjunto de puntos asignados al clúster *j*,

μ_j es el centroide del clúster *j*,

$\left\| x_i - \mu_j \right\|^2$ es la distancia euclidiana al cuadrado entre el punto x_i y su centroide correspondiente.

Una ventaja relevante del algoritmo K-means es su escalabilidad. En comparación de otros métodos de agrupamiento, como los jerárquicos, es un algoritmo que puede ser aplicado de forma eficiente en datasets con gran tamaño. Adicionalmente, su simplicidad hace que pueda ser utilizado fácilmente en las herramientas de análisis empresarial, a la hora de utilizar herramientas como Python, R o software especializado en CRM (Andy Hermawan et al., 2024; Tsiptsis Antonios Chorianopoulos WILEY, n.d.).

Sin embargo, K-means también tiene algunas limitaciones, y de manera importante destaca la necesidad de definir el número de clusters k, lo cual puede llegar a ser bastante arbitrario si uno no utiliza una métrica objetiva para determinarlo. El método del codo, el puntaje de silueta y el índice de Calinski-Harabasz son algunas de las métricas más comunes en la selección de k (Kodinariya & Makwana, 2013a). Todas estas técnicas se ocupan de hallar un equilibrio entre la cohesión dentro del mismo cluster y la separación entre clusters diferentes.

Otra limitación significativa que puede surgir del uso de K-means es su dependencia a la inicialización de los centroides. Si estos son elegidos de forma aleatoria, el algoritmo podría converger a valores mínimos locales de la función objetivo. Para evitarlo se han propuesto algunas variantes como K-means++, que escoge probabilísticamente las inicializaciones más dispersas obteniendo así agrupamientos más resistentes y de mejor calidad (Arthur & Vassilvitskii, 2007). También, el algoritmo es sensible a la escala de las

variables utilizadas, por lo que, antes de ponerse en marcha, resulta necesario estandarizar o normalizar previamente los datos que se van a utilizar, más aún si las variables son de distintas unidades o de diferentes órdenes de magnitud (López, 2024). De no hacerlo, el cálculo de distancias podría verse distorsionado por las variables que toman mayores valores.

El K-means es una buena técnica de agrupamiento en aquellos escenarios donde se busca: que sea explícitamente sencillo de interpretar; su bajo coste computacional; su capacidad de trabajar con variables múltiples. A pesar de sus limitaciones, si se realiza una buena parametrización y validación se pueden acabar alcanzando resultados útiles en cualquier entorno empresarial.

Uso de Python, librerías para manipulación de datos y scikit-learn para k-means.

Python se ha convertido en una de las herramientas más utilizadas en ciencia de datos y el machine learning, gracias a su sintaxis relativamente sencilla frente a otros lenguajes, así como por la facilidad de obtener librerías especializadas para el procesamiento de datos y análisis de datos estructurados (Oliphant, 2007). En particular, el trabajo entre bibliotecas de Python como pandas, NumPy y scikit-learn permite implementar flujos de trabajo reproducibles en todas las actividades que van desde la preparación de datos, la selección de características, hasta la aplicación de modelos de machine learning supervisados o no.

La interoperabilidad de estas librerías producidas con Python hace que sean fácilmente adoptadas en procesos de segmentación de clientes que requieren la combinación de diferentes fuentes de datos, la transformación de datos a estructuras tabulares y la aplicación de algoritmos de agrupamiento (Oliphant, 2007).

En las fases de preprocesamiento, la librería pandas, ofrece la posibilidad de cargar, filtrar, transformar y agregar datos fácilmente, utilizando estructuras de datos de tipo DataFrame y funciones de manipulación vectorizada. Esta librería resulta muy útil en investigaciones empíricas debido a su eficiencia a la hora de llevar a cabo tareas de

limpieza y organización, antes de aplicar algoritmos de minería de datos (McKinney, 2010). Por su parte, NumPy brinda soporte para operaciones de álgebra lineal, esto resulta fundamental cuando las variables tienen que ser escaladas antes de utilizar métodos que están basados en la distancia, como el del K-means (Harris et al., 2020).

Para la codificación de algoritmos de clustering, scikit-learn es una de las librerías más estandarizadas y robustas que existen en el ecosistema Python. Esta biblioteca ofrece una interfaz sencilla para distintos algoritmos, entre los que se cuentan técnicas de agrupamiento como K-means, que requieren que los parámetros sean ajustados, como el número de grupos (k), el tipo de inicialización y el número de reinicios de un algoritmo (Pedregosa et al., 2012). La función KMeans de scikit-learn permite fijar parámetros, que pueden estar predefinidos por defecto o pueden ser cambiados. También permite calcular métricas que permiten evaluar la calidad del agrupamiento, como su inercia o los índices de validación interna como el coeficiente de silueta.

A su vez, la biblioteca scikit-learn incluye funciones de escalado como StandardScaler(), las cuales son necesarias si las variables tienen distintas unidades o escalas, pues los algoritmos como K-means son sensibles a las magnitudes relativas de los atributos (Andy Hermawan et al., 2024; McKinney, 2010). El escalado previo asegura que las distancias euclidianas mantengan relaciones equitativas entre las variables. Este enfoque ha sido empleado en investigaciones recientes para segmentar clientes a partir de las variables tipo RFM y otras características del comportamiento del proceso de compra, demostrando su aplicabilidad en contextos de negocio que intentan identificar patrones no lineales y heterogéneos en las bases de datos transaccionales (Andy Hermawan et al., 2024).

Capítulo 3: Desarrollo

Descripción del dataset

Se utilizó información sin procesar de transacciones de ventas de Proalva, que registró datos desde el 2 de mayo del 2020 hasta el 7 de marzo del 2025. Este dataset tiene un formato estructurado, xlsx, consta de 31 columnas y 141,035 observaciones. En la Tabla 1 se presenta una muestra del contenido de las variables.

La variable correspondiente al identificador del cliente (RUC o cédula), fue pseudonimizada por la empresa previo a la entrega al grupo de trabajo. La variable Cliente_codigo presentó un identificador numérico único para cada cliente, esto con el fin de mantener la reserva empresarial y cumplir con la Ley de Protección de Datos Personales.

La empresa indicó que desde el año 2020 cuenta con un Sistema de planificación de recursos empresariales (ERP por sus siglas en inglés), desde donde provino la información. Este tipo de software les permite tener de forma centralizada la información de ventas a distintos niveles y con un número importante de variables.

Tabla 1

Listado de variables representativas del dataset y ejemplo de un registro

Variable	Ejemplo
Cliente_codigo	105
CodProducto	K15011
DescripcionProducto	PULPA MANZANA 1KG
CantidadItem	6.0
Unidad	UNIDAD
PrecioItem	1.59
TotalItem	9.54
IdentificadorVenta	00201000000389920230228
Fecha	2023-02-28 00:00:00
Tipo	CRED
Estado	COBRADO
Iva	0
Total	332.95
NomSabor	MANZANA
GRUPO O CLASIFICACION	Pulpa de fruta
MARCA	Propia
SUB CLASIFICACION	1kg

Variable	Ejemplo
CantKgPulpaItem	6.0
mesanio	2023-02

En cuanto a las 31 variables disponibles, se destacan identificadores de cliente “Cliente_codigo” y de producto “CodProducto”, así como variables que clasifican o caracterizan al ítem vendido. Entre estas se encuentran “DescripcionProducto” que hace referencia al nombre del ítem vendido, “NomSabor” indica la fruta con la que fue elaborada el producto, “MARCA” identifica si es marca propia de la empresa o maquila (empaque proporcionado y gestionado por el cliente), “GRUPO O CLASIFICACION” clasifica al item según la línea de producción a la que pertenece, y “SUB CLASIFICACION” detalla la presentación del producto.

La información cuantitativa de cada registro está representada en variables como la cantidad de producto vendido (CantidadItem), el precio unitario (PrecioItem), el total por ítem de la transacción (TotalItem), y el volumen de pulpa en kilogramos (CantKgPulpaItem). Esta última variable transforma las unidades vendidas según el volumen que contenga la presentación, y las estandariza a kilogramos, solo aplica para las líneas de producción de pulpa de fruta.

La variable “Total” representa el valor monetario final por cada factura, representada a su vez por la variable “IdentificadorVenta”. Previo al total por factura, se incluyen también columnas con información de “Descuento”, “Iva”, y otros campos como “Subt0”, “SubtIva” y “Retencion”. Según informó la empresa, la mayoría de los productos, al ser semielaborados, no gravan IVA, esto se corroboró en el dataset.

Todas las transacciones tienen información sobre la fecha de la facturación, descompuestas en variables temporales como “Fecha”, “FechaDet” que incluye fecha y hora, “anio”, “mes”, “día” y “mesanio”.

Adicionalmente, el dataset incluye información sobre el estado y tipo de las transacciones, “Estado” y “Tipo”, lo cual permite distinguir entre ventas efectivas, anuladas u

otros eventos contables, o si la venta fue con crédito. En el dataset recibido no se encontraron transacciones anuladas, sino dos categorías emitido y cobrado. Las variables Asiento y DescItem corresponden a registros contables y descuentos específicos por ítem vendido.

Análisis exploratorio de datos (EDA)

El EDA inició con una evaluación del formato de las variables disponibles en el dataset, se observaron tres formatos, numérico (int64 y float64), texto (object) que principalmente representa variables categóricas y fecha (datetime64) que representan fecha y hora. La variable correspondiente al código del cliente fue importada como numérica, sin embargo, se consideró que debería ser de tipo carácter o texto. Los detalles del formato de cada variable se encuentran en el Apéndice A.

Como segundo aspecto se revisó la presencia de valores faltantes, en la Tabla 2 se presenta las variables que presentan valores faltantes en sus registros. Se considera que las cinco primeras no son de gran relevancia para la segmentación de clientes, pues incorporan información cuantitativa de tipo contable para facturación. En cuanto a la variable “CantKgPulpaItem” se identifica que presenta valores perdidos en los productos que no pertenecen a la línea de producción de pulpa de fruta, al igual que las variables que caracterizan al ítem vendido.

Tabla 2

Frecuencia de valores perdidos por nombre de variable

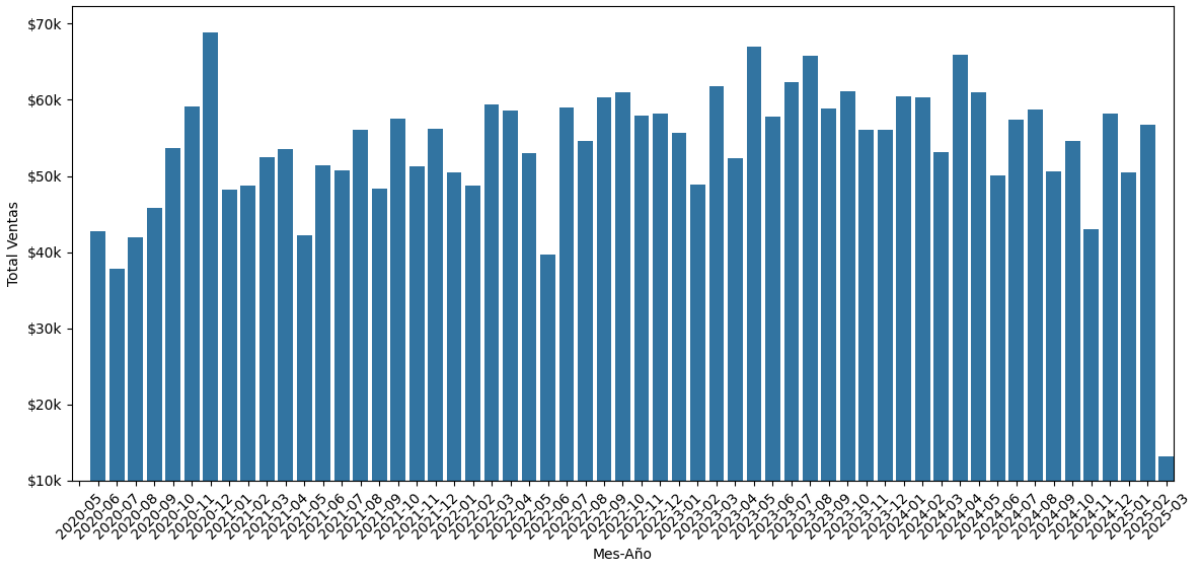
Variable	Observaciones con valores faltantes
Subt0	59,872
Descuento	59,872
Retencion	42,998
SubtIva	42,998
DescItem	42,998
CantKgPulpaItem	1,076
MARCA	117
SUB CLASIFICACION	65
NomSabor	44

Variable	Observaciones con valores faltantes
GRUPO O CLASIFICACION	43
CodSabor	43
NOMBRE2	43

Se analiza la evolución de las ventas mensuales de la empresa, acumulando el valor monetario de las ventas de todos los ítems en cada mes. En la Figura 1 se observa que el mes de marzo 2025 no está completo, por lo que este mes no se utilizará en análisis posteriores. En general las ventas promedio de la empresa se encuentran entre los 40 y 65 mil dólares al mes.

Figura 1

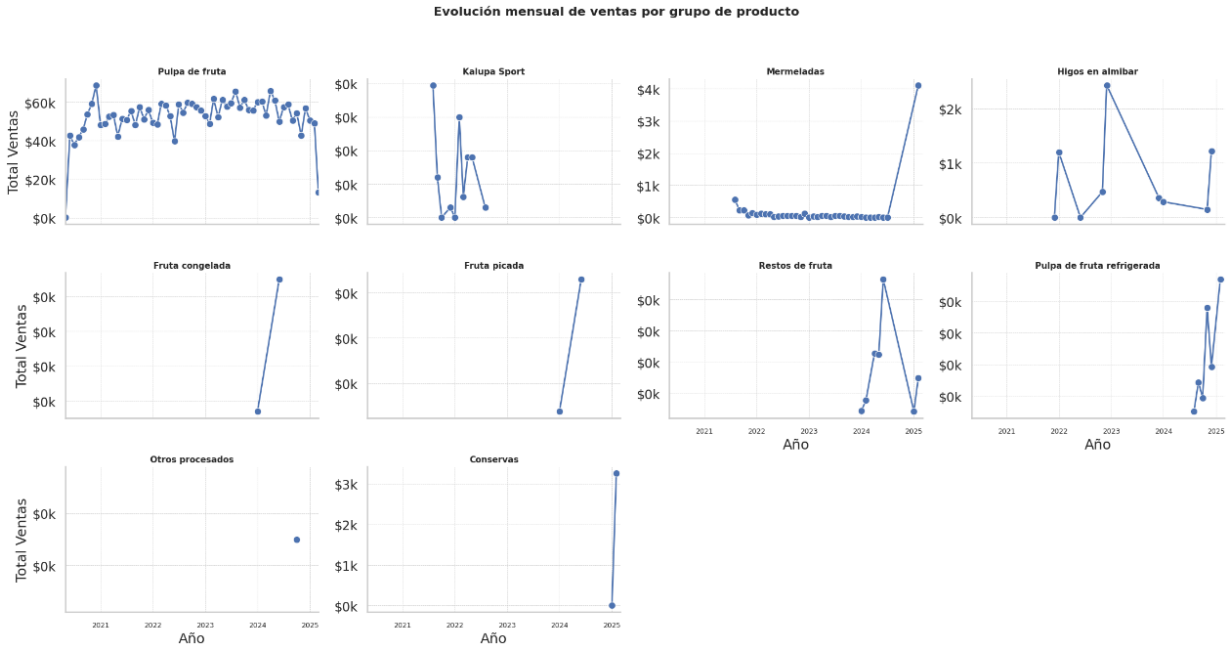
Ventas mensuales de la empresa, en dólares



Al revisar las ventas según línea de producción, Figura 2, se observa que la única línea de producción que presenta datos constantes a lo largo del periodo del dataset es la de pulpa de fruta. La empresa ha manifestado que el resto de los productos se venden bajo demanda, como por ejemplo para exportaciones esporádicas, o a su vez son productos descontinuados.

Figura 2

Ventas mensuales según línea de producción, en dólares



En cuanto a otras variables, el dataset muestra que existen dos categorías para la variable correspondiente a marca, propia y maquila, siendo la principal propia, en el Apéndice C se muestra la distribución de esta columna. La variable sabor registra 36 categorías, y describe la fruta con la que se elabora el producto, la que registra mayor frecuencia son mora, guanábana y maracuyá. En cuanto al número de productos, en total la empresa registra 147 productos, la Tabla 3 muestra el número de productos según la línea de producción. La mayoría se concentran en Pulpa de fruta, seguido de mermeladas y pulpa de fruta refrigerada.

Tabla 3

Número de productos por la línea de producción

Línea de producción	Número de productos
Pulpa de fruta	113
Mermeladas	8
Pulpa de fruta refrigerada	4
Kalupa Sport	4
Conservas	3
Fruta picada	2
Otros procesados	2
Higos en almíbar	2
Fruta congelada	1
Restos de fruta	1

Preparación y limpieza de datos

Antes de realizar la segmentación de clientes bajo los métodos propuestos, se debe realizar la limpieza de datos, este proceso ayuda a eliminar errores comunes en las bases de datos como: valores sin formato, faltantes o duplicados, datos irrelevantes o redundantes y detección de outliers (Chu et al., 2016; Zhang et al., 2022).

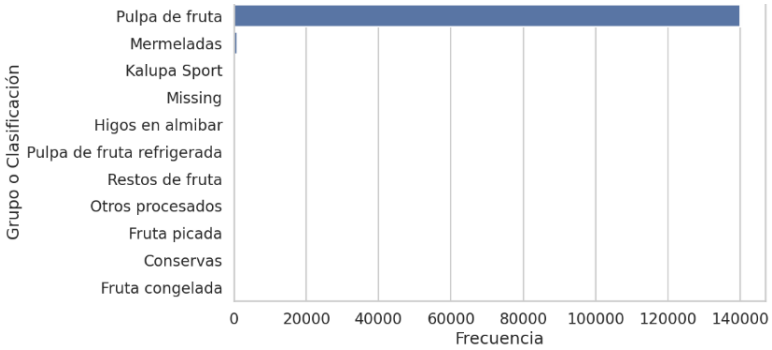
En la limpieza y preparación del dataset de transacciones de ventas de Proalva, se determinó que el código de cliente es un valor único, no se empleara en operaciones matemáticas, además es un dato categórico, se optó por asignarle el tipo objeto.

Se eliminó las columnas: “Asiento”, ”NOMBRE2”, ”CodSabor”, ”FechaDet”, ”anio”, ”mes”, ”dia”, ”Subt0”, ”Descuento”, ”SubtIva”, ”Retencion” y ”DescItem” debido a que contenían información contable o duplicada sobre las transacciones, mantenerlas representaría una pérdida de recursos ya que el procesamiento de los datos tomaría más tiempo, además que no aportarían a los objetivos del modelado para segmentación de clientes.

Se eliminaron los registros donde el valor de “TotalItem”, correspondiente a ventas por producto, era igual a cero, esto puede suceder en casos de anulaciones en las ventas, también se eliminaron las observaciones correspondientes a marzo de 2025 ya que no se cuenta con suficientes datos de ese mes, con el fin de tener disponibles datos de meses completos.

Figura 3

Frecuencia de registros por la línea de producción



Proalva ofrece múltiples líneas de productos, sin embargo, en el dataset se figura como principal la pulpa de fruta, no se encuentra suficientes datos de los otros productos, como se muestra en Figura 3, por lo que se decidió filtrar el dataset para trabajar solo con la línea de pulpa de fruta.

Se generó un nuevo conjunto de datos con el propósito de aplicar tanto el modelo RFM como el algoritmo de clustering K-means. El conjunto de datos considera un periodo de observación de 24 meses, seleccionado por su proximidad temporal y por ofrecer una representación actualizada del comportamiento de compra de los clientes (Shirole et al., 2021). Esta decisión responde a la naturaleza estática de los modelos de segmentación utilizados, los cuales requieren una representación consolidada del comportamiento transaccional del cliente durante un intervalo temporal definido (Konstantinos & Antonios, 2009).

El conjunto de datos agregado incluye información resumida por cliente para el periodo señalado, tras el proceso de limpieza, que mantuvo un total de 340 clientes. Para cada cliente se calcularon las siguientes variables: identificador del cliente, Recency (número de días desde la última compra), Frequency (número de transacciones realizadas), Monetary (valor total en dólares de las ventas en el periodo), fecha de la última compra (utilizada con fines de validación de la recencia), número de sabores distintos adquiridos, cantidad total en kilogramos de pulpa de fruta comprada, número de presentaciones diferentes adquiridas, número de días distintos de la semana en que se realizaron compras (por ejemplo, de lunes a viernes equivaldría a cinco), y variables binarias que indican la presencia de compras por cada sabor (dummy por tipo de fruta) y por marca. En el Apéndice B se presenta el detalle de las variables de este dataset generado.

Tabla 4

Estadísticos descriptivos de las variables en el dataset agregado por cliente

Variable	registros	media	desvEst	min	P5	P15	P50	P85	P95	max
Recency	340	200.58	223.37	0	0.95	4	98	539.6	625.4	724
Frequency	340	30.19	117.22	1	1	1	8	33.3	85.3	1724

Variable	registros	media	desvEst	min	P5	P15	P50	P85	P95	max
Monetary	340	3998.85	26359.06	0.01	12.88	29.82	204.70	1578.62	5901.10	351623.81
NumSabores	340	8.47	4.81	1	2	4	8	14	18	25
TotalKg	340	1877.62	13831.80	1	4	10	68.5	511.05	2354.68	188145.4
NumPresentaciones	340	1.43	0.69	1	1	1	1	2	3	4

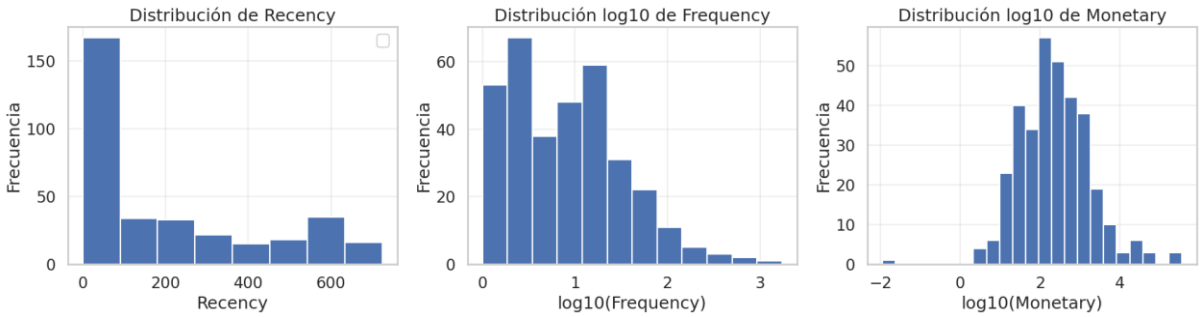
La Tabla 4 presenta los estadísticos descriptivos del conjunto de datos agrupados por cliente. La variable Recency, que indica el número de días desde la última compra, muestra una media de 200.58 días y una alta desviación estándar (223.37), lo que indicaría una alta heterogeneidad en los días transcurridos desde la compra más reciente. Su distribución, sesgada positivamente, tiene una mediana equivalente a 98 días y un percentil 95 de 625,4 días, lo que significaría que existe un porcentaje de clientes inactivos durante mucho tiempo. En cuanto a la variable Frequency, que contabiliza el número de transacciones por cliente, tiene una media de 30.19 y una alta desviación estándar (117.22), lo que da cuenta de que se da una gran acumulación de clientes que compran poco (mediana de 8) y algunos valores extremos (máximo de 1.724 transacciones), de clientes con alta frecuencia de compra.

La variable Monetary, que indica el valor total de compra por cliente, tiene una media de $3998.85 y una desviación estándar muy elevada ($26359.06), por lo que sería probable encontrar clientes con niveles de gasto excepcionalmente altos (máximo de $351,623.81). Estos valores fueron validados con la empresa, quienes confirmaron que efectivamente tiene un grupo muy pequeño de clientes grandes. La mediana ($204.70) y los percentiles 5 y 95 ($12.88 y $5901.10, respectivamente) dan constancia sobre la asimetría elevada de la distribución. Variables complementarias como NumSabores y TotalKg, que indican la diversidad de sabores comprados y el volumen de compra total en kilogramos, también tienen una alta dispersión, hasta máximos de 25 sabores y más de 188 mil kilogramos, con una mediana de valores medianos (8 sabores y 68,5 kg). Finalmente, la variable NumPresentaciones tiene una distribución más contenida con una media de 1,43 formatos diferentes y un máximo de 4, de tal forma que la mayoría de los clientes compran unos

formatos distintos de 1 o 2. Todo ello da cuenta de la existencia de subgrupos de clientes con comportamientos de compra diferenciados, lo que justifica la aplicación de métodos de segmentación de tipo clustering.

Figura 4

Distribución de las variables R, F y M en el dataset agregado por cliente



La Figura 4 muestra los perfiles de las distribuciones de las variables R, F y M que se agregó en el dataset por cliente, en el caso de Frequency y Monetary, por su alta dispersión (ver Tabla 4), fueron transformadas con logartimo base 10, con el fin de poder interpretar mejor su distribución. La variable Recency, días desde la última compra, presenta una distribución asimétrica hacia la derecha, donde la mayoría de los clientes tienden a realizar compras recientes (valores cercanos al cero), al contrario, se observa un pequeño número de clientes que no han comprado en periodos largos de tiempo. Este comportamiento hace suponer la existencia de una proporción de clientes inactivos o perdidos, y que podría estar relacionado con limitada fidelización.

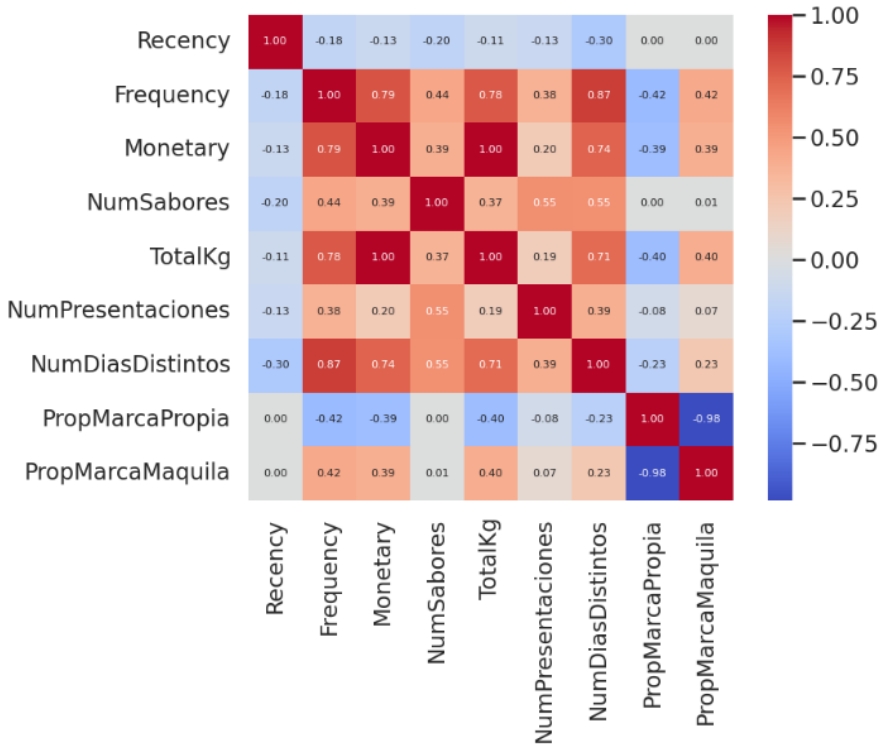
En el caso de la variable Frecuency, se observa que un grupo de clientes realizó pocas transacciones en el periodo analizado, mientras que otro grupo menor de clientes realiza compras con alta frecuencia, pudiéndose observar, así, la existencia de compradores ocasionales y de clientes muy activos. Esto genera también una idea de posibles estrategias de marketing o fidelización según el tipo de comportamiento.

La variable Monetary muestra una distribución con un comportamiento parecido a la de Frecuency, se observa una concentración mayoritaria en niveles bajos o medios de ventas (entre 0 a 4 en el gráfico), y unos pocos clientes que representan ventas superiores.

Este hecho también se debe al conocido long-tail, en donde una pequeña minoría de los clientes genera unos altos niveles de ingresos, favoreciendo, así, estrategias de retención y personalización específicas para este grupo.

Figura 5

Matriz de correlaciones entre variables numéricas del dataset para segmentación



La Figura 5 presenta un análisis de correlación entre las variables numéricas agregadas por cliente, con el fin de identificar relaciones en los comportamientos de compra. La variable Frequency muestra una correlación positiva con Monetary (r = 0.79), TotalKg (r = 0.78), y NumDiasDistintos (r = 0.87), lo que sugiere que los clientes que compran con mayor frecuencia también tienden a realizar compras de mayor valor, adquirir mayores volúmenes y comprar en más días distintos. Recency se correlaciona negativamente con Frequency (r = -0.18) y NumDiasDistintos (r = -0.30), lo cual sugiere que los clientes que han comprado recientemente también lo hacen con mayor frecuencia y regularidad.

Segmentación de clientes mediante el marco RFM

Se creó un conjunto de datos, denominado df_rfm, con el propósito de transformar y codificar las métricas RFM previamente calculadas, facilitando así su interpretación y posterior uso en este modelo de segmentación.

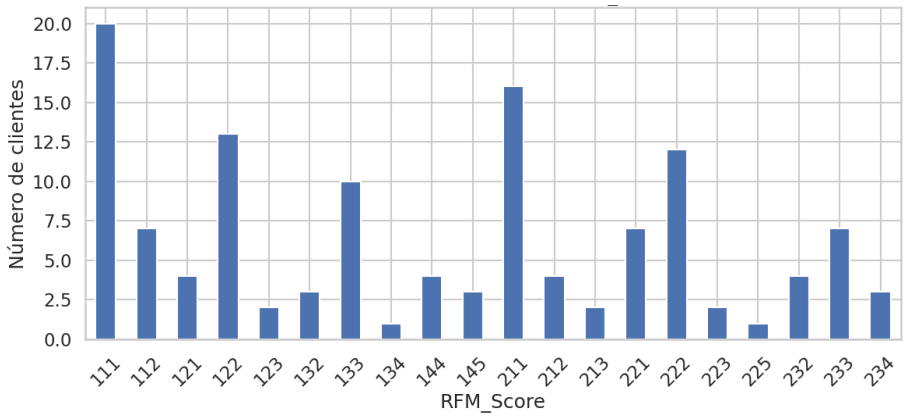
Se aplicaron transformaciones sobre las variables Recency (R), Frequency (F) y Monetary (M), utilizando la función qcut para asignar valores ordinales del 1 al 5 según el comportamiento de cada cliente.

En el caso de la recencia, se asignaron puntuaciones más altas a valores más bajos, dado que una compra reciente es más valiosa, mientras que para la frecuencia y el valor monetario se otorgaron mayores puntuaciones a valores más altos (Khan et al., 2022).

Finalmente, se construyó una nueva variable denominada RFM_Score, mediante la concatenación de los tres rangos asignados, creando así una variable única que sintetiza el valor RFM de cada cliente, y que permite perfilar mediante una métrica única (Cheng et al., 2023).

Figura 6

Número de clientes en los scores RFM más bajos



Se aprecia en la Figura 6 las veinte combinaciones más bajas de RFM Score, estos clientes que se encuentran en el primer quintil de recencia, de frecuencia y de valor monetario, es decir realizan compras que representan un bajo valor monetario, pocas transacciones, en periodos largos de tiempo.

Partiendo del modelo RFM (Recencia, Frecuencia y Valor Monetario), a partir de los valores ordinales para R, F y M, se asignaron categorías según el perfil del cliente al que podrían representar esas combinaciones.

La primera categoría se denominó “Estrella” (R ≥ 4, F ≥ 4, M ≥ 4), que muestran compras recientes, frecuencia elevada y gasto significativo. El grupo “Prometedor” incluye clientes con alta recencia (R ≥ 4) y al menos una puntuación moderada en frecuencia o valor monetario (F ≥ 3 o M ≥ 3), lo que sugiere una relación comercial reciente con potencial de consolidación (Wulansari & Heikal, 2024). El segmento “Leal pero inactivo” se define por alta frecuencia y gasto (F ≥ 4, M ≥ 4), pero baja recencia (R ≤ 2), lo que indica una trayectoria de compras intensiva en el pasado, pero con signos de descontinuación reciente (Andy Hermawan et al., 2024). Por su parte, los clientes clasificados como “Perdido” presentan valores bajos en las tres dimensiones (R ≤ 2, F ≤ 2, M ≤ 2), lo que sugiere una desvinculación comercial, posiblemente definitiva (Rajagukguk, 2025). Se identificó también un segmento “Buen cliente”, caracterizado por frecuencia y gasto moderadamente altos (F ≥ 3, M ≥ 3), aunque sin criterios sobre recencia, lo cual representa a clientes con contribución económica relevante. Finalmente, el grupo “Regular” incluye a los clientes que no cumplen con ninguna de las condiciones anteriores, y por tanto presentan patrones intermedios en sus comportamientos de compra.

Esta clasificación permitó construir una estructura de segmentos jerárquica, para un perfilamiento y segmentación inicial que permita comparar con los resultados de la aplicación del modelo de clustering no supervisado con K-means (Rajagukguk, 2025; Wulansari & Heikal, 2024).

Segmentación de clientes mediante clustering K-means

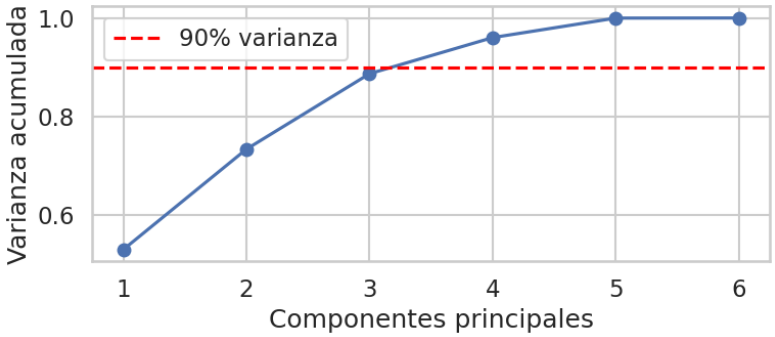
Para la implementación del modelo de clustering K-means se seleccionaron seis variables numéricas que describen diferentes dimensiones del comportamiento de compra de los clientes: recencia (Recency), frecuencia (Frequency), valor monetario (Monetary), total de kilogramos de pulpa de fruta comprados (TotalKg), número de sabores comprados

(NumSabores) y número de presentaciones adquiridas (NumPresentaciones). Estas variables fueron elegidas con base en criterios teóricos y prácticos que reflejan tanto la intensidad como la diversidad de las compras realizadas, y se ajustan a prácticas establecidas en la literatura para segmentación de clientes en contextos de consumo de productos (Christy et al., 2021a; Ramkumar et al., 2025a; Shirole et al., 2021).

Antes de aplicar el modelo K-means, se procedió al escalamiento estándar de las variables, dado que estas se encuentran en distintas unidades y rangos. El uso de la transformación StandardScaler() permite centrar las variables en cero y escalar su varianza a uno, lo que evita que aquellas con mayor magnitud dominen la distancia euclidiana utilizada por K-means para asignar puntos a los centroides (Durugkar et al., 2022). Este paso es importante para asegurar que todas las variables contribuyan de forma equitativa en la formación de los grupos de clientes.

Figura 7

PCA- Varianza acumulada explicada por los componentes principales



Como parte de la validación de las variables seleccionadas, se realizó un Análisis de Componentes Principales (PCA). La Figura 7 muestra que las primeras cuatro componentes principales explican más del 95% de la varianza total, superando así el umbral del 90% comúnmente recomendado para asegurar una buena representación de la información original en espacios de menor dimensión (Jolliffe & Cadima, 2016a).

Tabla 5

Cargas de los componentes principales

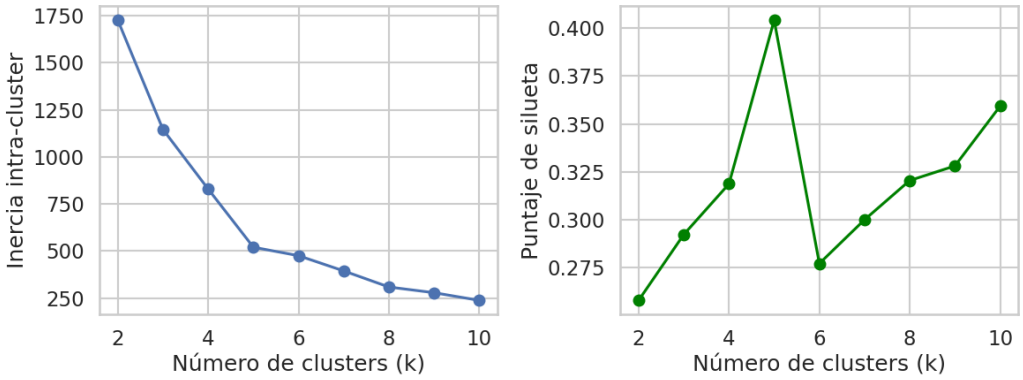
Variable	PC1	PC2	PC3	PC4	PC5	PC6
Recency	-0.15	-0.35	0.92	0.06	0.06	0
Frequency	0.5	-0.11	0	-0.28	0.81	-0.02
Monetary	0.51	-0.33	-0.03	0.03	-0.33	0.72
TotalKg	0.51	-0.35	-0.03	0.03	-0.36	-0.7
NumSabores	0.36	0.49	0.19	0.77	0.1	-0.01
NumPresentaciones	0.28	0.63	0.34	-0.58	-0.29	0

Respecto a las cargas del PCA, la Tabla 5 muestra que las variables Frequency, Monetary y TotalKg tienen cargas similares y elevadas sobre el primer componente principal (PC1), con valores cercanos a 0.5, lo que sugiere que comparten una dimensión común, que podría estar relacionada con la intensidad de compra. Por otro lado, NumSabores y NumPresentaciones presentan las cargas más altas en PC2 (0.49 y 0.63, respectivamente), indicando que capturan otra dimensión del comportamiento de compra, posiblemente asociada a la diversidad de productos adquiridos. La variable Recency, presenta su mayor carga en PC3 (0.92), por lo que abarca una dimensión distinta respecto a las anteriores. Este patrón de cargas sugiere una baja colinealidad entre las variables principales y resalta su capacidad para representar distintos aspectos del comportamiento del cliente (Abdi & Williams, 2010; Jolliffe & Cadima, 2016b).

A pesar de que se identificaron correlaciones relativamente altas entre algunas variables en el análisis del dataset agregado por cliente (Figura 5), se decidió conservar todas en la segmentación por K-means, con el fin de preservar las distintas dimensiones de comportamiento que fueron evidenciadas mediante el análisis de componentes principales. Esta decisión está respaldada por estudios que recomiendan mantener variables con aportes complementarios al espacio multidimensional en procesos de agrupamiento no supervisado (Everitt et al., 2011b; Wold et al., 1987).

Figura 8

Método del codo y puntaje de silueta con variables seleccionadas para el modelo de clustering



Para la identificación del número óptimo de clusters se utilizaron dos métricas complementarias: la inercia intra-cluster (método del codo) y el puntaje de silueta (silhouette score). La inercia mide la suma de las distancias cuadradas entre los puntos de cada cluster y su centroide, y su disminución progresiva refleja mejoras en la cohesión interna a medida que se incrementa el número de clusters (Kodinariya & Makwana, 2013b). En la Figura 8 se observa una disminución pronunciada de la inercia entre k = 2 y k = 5, seguida de una desaceleración que indica un punto de inflexión o “codo” en k = 5. Este patrón sugiere que cinco clusters logran un balance razonable entre complejidad del modelo y homogeneidad interna de los grupos.

De manera complementaria, el análisis del índice de silueta permite evaluar simultáneamente la cohesión interna y la separación entre clusters. El puntaje más alto de silueta se obtiene en k = 5 (aproximadamente 0.41), lo cual refuerza la elección sugerida por el método del codo. Este valor indica una separación razonable entre los grupos y una asignación consistente de los puntos dentro de sus respectivos clusters (Rousseeuw, 1987). Aunque para valores mayores de k también se observan puntajes aceptables, estos no superan el máximo alcanzado en k = 5, por lo que se selecciona este valor como óptimo para la segmentación por k-means.

Parámetros de configuración del algoritmo K-means

Figura 9

Código para la implementación del modelo k-means

```
kmeans = KMeans(n_clusters=5, random_state=92, n_init='auto')
df_segmentacion['cluster'] = kmeans.fit_predict(X_scaled)
```

Para llevar a cabo la implementación del modelo de segmentación de clientes se utilizó el algoritmo K-means, que se encuentra en la biblioteca scikit-learn en Python, el cual permite clasificar el conjunto de datos agregado a nivel de cliente, en un número predefinido de grupos k, minimizando la suma de las distancias cuadradas entre cada punto y el centroide de su grupo (Jain, 2010). En este proyecto, tal como muestra la Figura 9, el modelo fue configurado con cinco clusters (n_clusters=5), procurando generar perfiles de clientes que presenten patrones de comportamiento de compra similares entre sí. La semilla aleatoria fue fijada mediante el parámetro randomstate=92, de esta manera se obtuvo la reproducibilidad de los resultados frente a los procesos estocásticos que se dan en la inicialización de los centroides y en la selección de las observaciones (Pedregosa et al., 2012).

Otro de los parámetros definidos fue ninit='auto', este parámetro fue introducido a partir de scikit-learn 1.4., y ajusta automáticamente la cantidad de inicializaciones del algoritmo dependiendo del método de inicialización utilizado(Scikit-learn developers, 2024). Para el caso del presente proyecto, se mantuvo el método por defecto, que es el de init='k-means++', por esta razón no se lo define el código.

La técnica k-means++, es una propuesta que se plantea como adaptación del método K-means tradicional y que tiene como objetivo aumentar la probabilidad de obtener una solución próxima al óptimo global. En este caso la primera de la elección de los centroides se hace de forma aleatoria, mientras que las siguientes se construyen con una probabilidad que es proporcional al cuadrado de la distancia más corta de cada punto al centroide más cercano ya elegido. El procedimiento que se sigue pretende maximizar la dispersión inicial de los centroides de forma que se logra una mejor cobertura del espacio los datos y se acelera la convergencia alcanzada por el método (Arthur & Vassilvitskii, 2007).

Bajo este parámetro, ninit se establece internamente en 10, lo cual implica que el algoritmo se ejecuta diez veces con diferentes configuraciones iniciales y se retiene la mejor solución según la inercia intra-cluster (Scikit-learn developers, 2024). Esta configuración es preferible a valores de ninit muy bajos, pues incrementan la probabilidad de converger a mínimos locales no óptimos (Arthur & Vassilvitskii, 2007).

El resultado de esta segmentación se incorporó al dataset bajo una nueva variable categórica denominada cluster, la cual permite distinguir a qué grupo pertenece cada cliente dentro del espacio multivariado de características consideradas. Esta segmentación proporciona una estructura que será analizada en el siguiente capítulo para caracterizar los perfiles de clientes resultantes.

Capítulo 4: Análisis De Resultados

En el presente capítulo se exponen los resultados obtenidos a partir de la aplicación del modelo RFM (Recencia, Frecuencia y Valor Monetario) y del algoritmo de segmentación K-Means sobre los datos de clientes de la empresa Proalva. Se analiza el comportamiento de cada modelo de forma individual, destacando los patrones y perfiles de clientes identificados, y se realiza una comparación entre ambos enfoques para evaluar su efectividad en la generación de grupos. Esta comparación permite determinar cuál de los modelos ofrece mayor utilidad para la toma de decisiones comerciales y la formulación de campañas de marketing focalizadas.

Pruebas De Concepto Para Segmentación Por K-means

La agregación del dataset a nivel de cliente, posibilitó disponer de un conjunto significativo de variabes para describir el comportamiento de compra de los clientes, lo que permitió, simultáneamente, trabajar con el modelo clásico de RFM y con K-means.

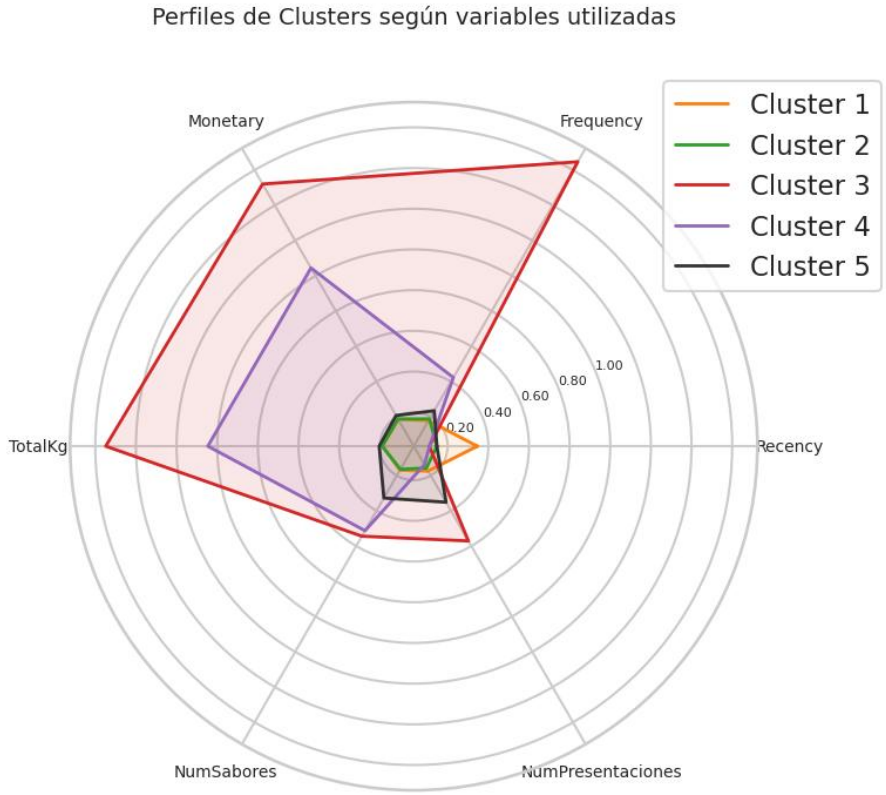
La calidad obtenida en la segmentación de clientes por K-means fue evaluada mediante dos métricas ampliamente utilizadas: el puntaje de silueta y el índice de Calinski-Harabasz. El puntaje de silueta alcanzó un valor de 0.404, lo que indica una estructura bien definida, con una separación razonable entre clusters y buena cohesión interna (Rousseeuw, 1987). Por otro lado, el índice de Calinski-Harabasz fue de 244.93, lo que sugiere que los clusters presentan una buena proporción de varianza entre grupos respecto a la varianza interna, siendo este índice particularmente sensible a la dispersión entre los centroides y al número de observaciones por grupo (Arbelaitz et al., 2013). Estos resultados, en conjunto, indican que la partición con cinco clusters genera una estructura interpretable y adecuada para los objetivos de este proyecto.

Con el fin de generar mejor interpretabilidad sobre la segmentación obtenida por k-means, le Figura 10 muestra un gráfico de radar, en la que se revisa el comportamiento de los clientes en cada cluster y para cada variable utilizada. Previamente se estandarizaron las variables usando StandardScaler de sklearn, con el fin de facilitar la comparación, pues

representan valores con rangos distintos. El Cluster 2 muestra los valores más elevados en Frequency, Monetary y TotalKg, el Cluster 0 presenta, casi de forma exclusiva, relación con la variable Recency, que representa los días desde la última compra. El Cluster 3, en menor medida que el 2, considera las variables Totalkg y Monetary. Por otro lado, en la mayoría de las dimensiones, los Clusters 1 y 4 muestran valores bajos, con algunas variaciones puntuales en NumSabores y NumPresentaciones.

Figura 10

Perfiles de clientes por cluster según variables de segmentación



Para la segmentación de clientes de la empresa Proalva se implementaron dos metodologías que permiten clasificar de forma precisa y estratégica a los consumidores: el modelo RFM y el algoritmo de agrupamiento K-Means. A continuación, se describen los resultados obtenidos con cada enfoque.

Análisis de Resultados

Modelo y clasificación RFM

Como resultado de la propuesta de clasificación de los rankings R, F y M, se generaron seis segmentos que representan diferentes niveles de valor para la empresa, clasificados desde los clientes más valiosos hasta los de menor relevancia comercial, desde la óptica de las tres variables.

Tabla 6

Distribución de clientes según clasificación RFM

Categoría RFM	Descripción	No. Clientes
Estrella	R ≥ 4, F ≥ 4, M ≥ 4	84
Perdido	R ≤ 2, F ≤ 2, M ≤ 2	83
Regular	Cualquier otra combinación que no cumpla las condiciones anteriores	64
Buen Cliente	F ≥ 3, M ≥ 3, sin cumplir condiciones previas	56
Prometedor	R ≥ 4 y (F ≥ 3 o M ≥ 3), sin cumplir la condición de Estrella	38
Leal pero inactivo	F ≥ 4, M ≥ 4, R ≤ 2	15
Total		**340**

Como se muestra en la Tabla 6, entre los 340 clientes analizados, la categoría Estrella concentra el mayor número de casos, con un total de 84 clientes. Le sigue de cerca el grupo Perdido, compuesto por 83 clientes que se caracterizan por valores bajos en recencia, frecuencia y volumen de compra. En tercer lugar se ubica el grupo Regular, con 64 clientes cuyo comportamiento no permite una clasificación precisa dentro de las categorías propuestas. Las categorías Buen cliente y Prometedor se posicionan en cuarto y quinto lugar, con 56 y 28 clientes respectivamente. Finalmente, el grupo Leal pero inactivo, que guarda cierta similitud con la categoría Perdido, reúne 25 clientes y ocupa la última posición en cuanto a tamaño.

Tabla 7

Promedios de cada variable según clasificación RFM

Categoría RFM	Recency	Frecuency	Monetary	TotalKg	NumSabores	NumPresentaciones
Estrella	9.15	95.23	15,165.47	7,212.28	10.93	1.74
Perdido	454.01	1.94	46.01	15.06	6.14	1.24
Regular	150.97	3.2	136.92	49.55	6.75	1.31
Buen Cliente	231.21	20.39	763.74	306.77	9.86	1.5
Prometedor	14.87	10.95	348.36	117.17	8.37	1.29
Leal pero inactivo	438.07	22.8	1,141.43	433.68	9.87	1.33

A partir de los valores promedios presentados en la Tabla 7, se observa que el grupo Estrella reúne a los clientes con mayor frecuencia de compra (95.23), mayor gasto (15,165.47) y volumen adquirido (7,212.28 kg), además de registrar tiempos recientes de compra (9.15 días en promedio) y una diversidad considerable tanto en sabores como en presentaciones. En contraste, el grupo Perdido refleja lo opuesto: una recencia muy alta (454.01 días), baja frecuencia (1.94) y montos de compra casi nulos, lo cual sugiere una relación prácticamente inactiva con la empresa. El grupo Regular presenta valores intermedios, con niveles bajos de frecuencia y volumen, sin un patrón claro de comportamiento. Por su parte, los clientes clasificados como Buen Cliente y Leal pero inactivo muestran compras más frecuentes y de mayor volumen que el promedio, aunque en el segundo caso, la alta recencia (438.07) indica una desconexión reciente. Finalmente, el grupo Prometedor destaca por su cercanía temporal (14.87 días), frecuencia moderada (10.95) y niveles de compra aún en desarrollo, lo que lo convierte en un grupo con potencial de crecimiento.

Si bien el modelo RFM facilita una primera aproximación al comportamiento de los clientes a partir de tres dimensiones clave, presenta ciertas limitaciones metodológicas. La clasificación depende de umbrales definidos de forma discrecional, lo que puede introducir sesgos al momento de asignar categorías. Además, asume homogeneidad dentro de cada grupo, aunque al observar la tabla de promedios es evidente que existen diferencias

importantes en variables adicionales como el volumen total comprado o la diversidad de productos adquiridos. Por ejemplo, aunque los clientes Leales pero inactivos muestran un gasto acumulado y volumen elevado, su alta recencia los ubica junto a perfiles con bajo nivel de actividad reciente, como el grupo Perdido. Estas inconsistencias revelan que el modelo RFM, al centrarse únicamente en tres variables, podría omitir patrones relevantes presentes en el resto de los datos, o a su vez no considera patrones no identificables a simple vista. Esto refuerza la necesidad de aplicar un método como K-means, que permite integrar múltiples dimensiones de forma simultánea, y agrupar clientes en función de su similitud general sin requerir reglas predefinidas.

Segmentación por K-Means

La segmentación realizada mediante el algoritmo K-Means, detallada en el Capítulo 3, dio como resultado cinco grupos, el número de clientes por cada grupo se muestra en la Tabla 8. Por motivos visuales se renombró a los clusters del 1 al 5. El clúster 2 concentra el mayor número de clientes, con un total de 181, le sigue el clúster 1, con 91 clientes, mientras que el clúster 5 agrupa a 65. En contraste, los clústeres 3 y 4 presentan 1 y 2 clientes respectivamente, estos podrían estar asociados a comportamientos atípicos o extremos que el modelo no logró agrupar con otros perfiles.

Tabla 8

Segmentación de clientes por método K-Means

Clúster	No. Clientes
1	91
2	181
3	1
4	2
5	65
Total	**340**

La Tabla 9 presenta los valores promedio de las variables utilizadas en la segmentación por este método, para cada cluster. Los clústeres 3 y 4 presentan valores que podrían considerarse extremos en todas las variables, con frecuencias superiores a las mil

transacciones y volúmenes de compra que superan los 100 mil kilogramos de pulpa de fruta, lo que sugiere que son grandes clientes. Por el contrario, el clúster más representativo en términos de tamaño fue el clúster 2, con 181 clientes, caracterizado por compras relativamente recientes, frecuencia moderada y un volumen de adquisición considerable. El clúster 1, con 91 clientes, presenta compras esporádicas y de bajo volumen, y el clúster 5, que reúne a 65 clientes, presenta un comportamiento intensivo tanto en frecuencia como en diversidad de productos.

Tabla 9

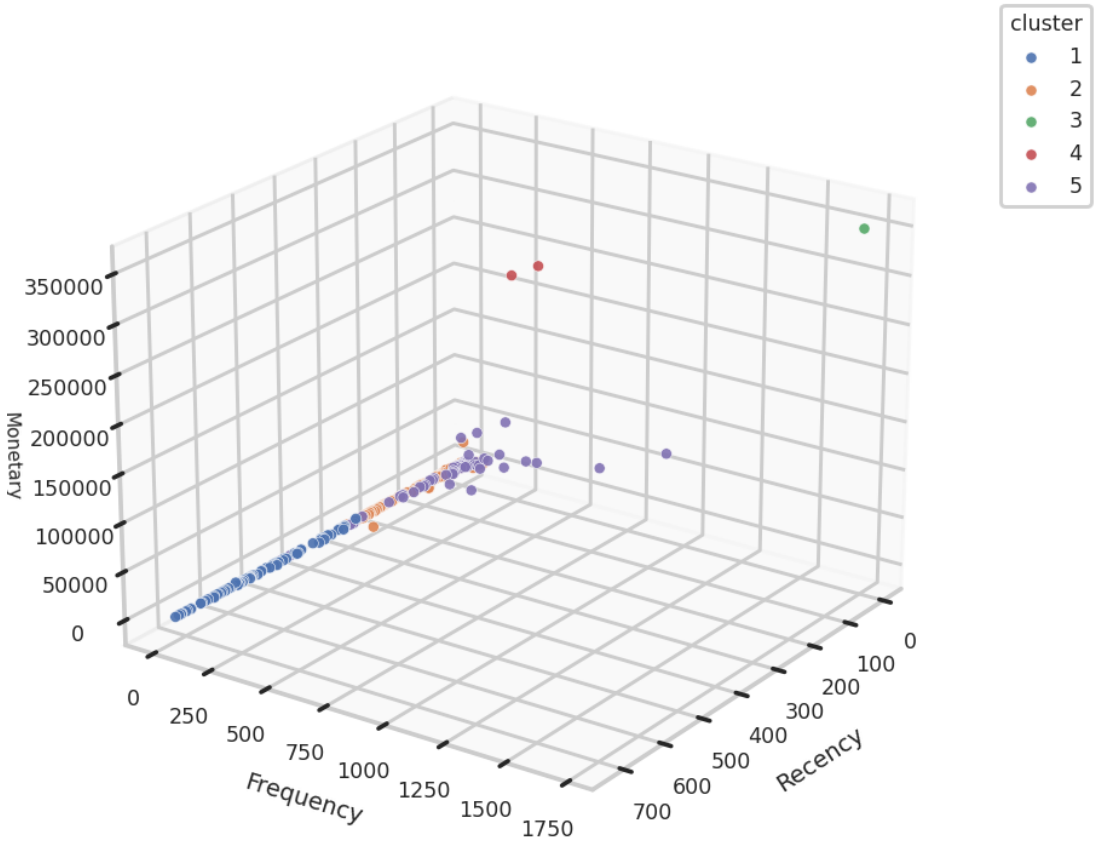
Promedios de cada variable según clúster de K-means

Clúster	Recency	Frecuency	Monetary	TotalKg	NumSabores	NumPresentaciones
1	525.91	5.18	240.35	92.24	6.99	1.25
2	82.73	15.98	811.42	297.04	6.75	1.13
3	0	1724	351623.81	188145.4	25	4
4	2	291	226290.66	118757	23.5	1
5	82.49	70.69	5,948.71	2316.49	14.58	2.48

A diferencia de la categorización mediante RFM, el modelo K-means no depende de umbrales definidos manualmente, lo que reduce el riesgo de sesgos en la segmentación. Además, al considerar simultáneamente múltiples variables, permite identificar patrones complejos y no evidentes a simple vista, capturando estructuras internas en los datos que reflejan el comportamiento de compras del cliente. La segmentación no supervisada permitió diferenciar a los granes clientes de la empresa en clusters específicos, generando espacio para agrupar al resto de clientes de forma más detallada e interpretable.

Figura 11

Distribución de clientes según cluster y variables R, F y M



La Figura 11 presenta las tres dimensiones R, F y M y la distribución de los 340 clientes, representados en un color diferente según el cluster al que pertenezcan. En el eje X se encuentra la variable Recency (días respecto a la última compra), en el eje Y, Frequency (número de compras a lo largo de un periodo de tiempo) y en el eje Z, Monetary (valor monetario total de las compras). El patrón de dispersión de los puntos nos indica una heterogeneidad significativa respecto al comportamiento de los clientes con un gran número de ellos situados en frecuencias y montos bajos o medios y unos pocos casos extremos que se corresponden con montos monetarios altos.

Existe un número de clústeres que están alejados del resto; por ejemplo, el clúster número 3 (color verde) y 4 (rojo) quedan en áreas alejadas del resto, lo que indica que existen clientes con comportamientos muy distintos como alta frecuencia, monto muy alto y recencia baja (compras recientes).

Por su parte, el clúster 1 (en azul) concentra un importante número de clientes, los mismos que son definidos por una baja frecuencia de compra, un alto grado de recencia y un monto monetario bajo, lo que explicaría la representación de clientes inactivos o de escaso valor para la empresa.

El clúster 5 (marcado en color púrpura) y en menor medida el cluster 2 (naranja) reúnen a clientes con valores intermedios-altos tanto en la dimensión Frequency como en Monetary, aunque con una dispersión mayor en la dimensión Recency. Estos clusters dan lugar al descubrimiento de la existencia de clientes que, si bien han realizado compras de forma relativamente frecuente y con montos económicos importantes, no han realizado compras muy recientemente. Es posible que estos dos clusters se diferencien por otras de las variables usadas, y no estrictamente por las del gráfico.

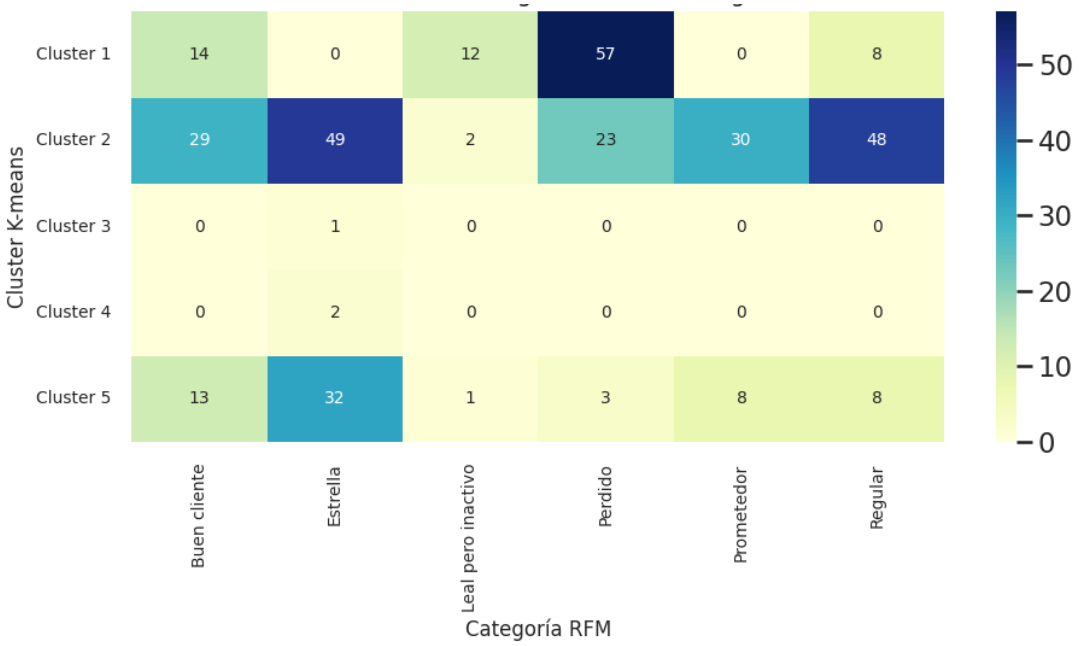
Este hallazgo en la combinación de estas dimensiones da paso a caracterizar este segmento como clústers de alto potencial pues identifica clientes con un historial de compras muy valioso, que podrían ser reactivadas mediante estrategias de fidelización o reengagement, se podría promover estrategias que, por un número de compras al mes, entreguen descuentos o productos gratis. Este grupo podría ser uno de los que se enfoque el presupuesto de marketing y ventas. (N. Kumar, 2025; V. Kumar & Reinartz, 2018)

Comparación de los dos modelos

En la Figura 12 se presenta una matriz comparativa de la distribución de clientes en función a los dos métodos de segmentación. El clúster 1, capta principalmente al grupo de clientes clasificados en el modelo RFM como perdido o leal pero inactivo. El clúster 2 presenta mayor dispersión, pues abarca a clientes de todos los grupos clasificados en RFM. Los clústeres 3 y 4 agrupan únicamente a tres clientes, categorizados como Estrella, en línea con su peso comercial identificado anteriormente, estos son los clientes más grandes de la empresa. El clúster 5 incluye clientes Estrella y Buen Cliente, lo que indicaría que esta agrupación retoma características de clientes con buenas métricas, pero que no llegan a los niveles de los 3 grandes clientes.

Figura 12

Distribución de clientes según método de segmentación



Aunque existe solapamiento entre los dos métodos de segmentación, K-means logra reorganizar los perfiles de los clientes con base en patrones más complejos, integrando dimensiones que el modelo RFM trata de forma independiente o que a su vez no considera.

Perfilamiento de los clientes y estrategias de marketing

Clúster 1. Clientes inactivos con un escaso valor comercial

Este grupo presenta una recencia más elevada, lo que hace suponer que han dejado de comprar desde hace bastante tiempo. Por otro lado, la frecuencia de compra de este grupo clientes es baja y presenta un escaso valor de compras, lo que refleja bajas perspectivas económicas para la compañía. Asimismo, el volumen total de productos adquiridos y la variedad de productos son bajos, lo que incide en su escasa relevancia comercialmente.

Se podría evaluar una campaña de reactivación dirigida, que mezcle comunicaciones personalizadas con incentivos específicos. Lo que se sugiere es llevar a cabo una campaña por email o chat para los clientes con una ventana de inactividad superior a seis meses. La campaña debería ser de bajo costo y sin mayor involucramiento

de recursos humanos, pues el retorno para la empresa podría ser limitado. La promoción podría ser por ejemplo, un porcentaje de descuento o un producto gratis en la siguiente compra. Para poder aumentar la eficiencia del gasto en esta campaña, se puede filtrar el grupo de clientes en función del valor histórico de compra, priorizando a los que en el pasado han dado más señales de fidelización.

Clúster 2. Clientes poco rentables con un comportamiento esporádico

En estos clientes se refleja una recencia media-baja, frecuencia baja-moderada y gasto monetario relativamente bajo. Si bien no se encuentran inactivos, sus patrones muestran como resultado compras en momentos muy concretos y de bajo nivel asociadas a una muy reducida diversificación de los sabores y presentaciones. Su valor es bajo por lo que es fácil que, por el tamaño del grupo (181 clientes), sean considerados para campañas de mantenimiento o para desarrollar estrategias de conversión a clientes recurrentes mediante promociones o paquetes personalizados.

Se puede plantear un programa de acumulación de puntos de acuerdo con la frecuencia mensual de compra, llegando a establecer un sistema que reconozca la superación de puntos por umbrales y que, de ese modo, permita acceder a descuentos, productos exclusivos o entrega sin coste adicional. Además, se puede crear un programa de empaquetado de productos, donde se agrupe los productos de referencia dentro del cluster (sabores más vendidos, presentaciones estándar), para disminuir la resistencia en el proceso de compra y favorecer el hábito de recompra. También se recomienda la comunicación automatizada con recomendaciones personalizadas en función de la trayectoria de compra, lo cual ha demostrado incrementar la tasa de conversión en otros contextos (Gorgoglione et al., 2019).

Clúster 3. Cliente institucional con alto valor (B2B)

Este cliente tiene la recencia más baja, la frecuencia más alta, el valor de compra más alto, ha comprado más de 188 toneladas de producto y también tiene una buena diversidad de sabores y presentaciones. La empresa ha identificado a este cliente como un

distribuidor mayorista, que explicaría esta serie de patrones. Se recomienda mantener una relación personalizada con este cliente, probablemente mediante un acuerdo de colaboración o condiciones ventajosas preferenciales.

Se propone una estrategia de retención proactiva centrada en la gestión de cuentas claves (Key Account Management) a la que se acompañan algunas características: uno de los elementos sería el contacto con una persona comercial responsable de la cuenta, el desarrollo de condiciones contractuales preferenciales, el acceso prioritario a nuevos productos y un soporte logístico personalizado.

En lo que respecta al ámbito operativo, puede llegarse a establecer un canal de atención exclusivo, planificación de la demanda en conjunto con entregas programadas y también incentivos por volumen de compra. También se podría analizar cómo se apalanca la integración del cliente en el ciclo de innovación de productos, por ejemplo en eventos de feedback de nuevos sabores o de nuevas presentaciones, por cuanto su volumen de compra lo convierte en un cliente importante para evaluar el portafolio de productos (Woodburn & McDonald, 2012).

Clúster 4. Grandes clientes industriales

También presentan recencia baja, frecuencia alta y un gasto elevado, este grupo de clientes constituiría una segunda categoría de clientes de alto valor. Disponen de baja diversidad de presentaciones y alta variedad de los sabores; estas últimas son indicativas de una posible necesidad de transformación del producto en el sector industrial. También son importantes en términos de volumen. Para desarrollar estrategias orientadas a este grupo de clientes se podría pensar en acuerdos de abastecimiento continuo, o un soporte personalizado en términos logísticos.

Se podría plantear una estrategia de consolidación y abastecimiento programado mediante la firma de contratos. Estos contratos podrían considerar precios escalonados en función del volumen comprado, con el agregado de servicios logísticos, tales como entrega asegurada de productos y almacenamiento temporal para despachos en función a la

demanda, así como asistencia posventa especializada. También se puede considerar esquemas de descuentos por pronto pago o descuentos por compromisos de fidelización anual.

Dada la naturaleza de estos clientes, se sugiere la exclusión de este grupo de campañas tradicionales, y poder gestionar estas estrategias a partir de una comunicación directa con los representantes de las áreas encargadas.

Clúster 5. Clientes de alto potencial

Este grupo presenta recencia en un nivel medio, pero destaca por una frecuencia alta y un alto gasto. También muestra una buena diversificación en los sabores y las presentaciones. A diferencia del clúster 3 y del clúster 4, su comportamiento parece corresponder con clientes minoristas o mayoristas en fase de crecimiento. El comportamiento de este grupo les convierte en clientes de alto potencial, casos que justifican una estrategia orientada a la fidelización y a la mejora del vínculo comercial del cliente mediante incentivos, atención preferencial o programas de fidelización.

Se propone una estrategia de fidelización y engagement relacional. Consistiría en diseñar un programa de beneficios, exclusivo para clientes de este segmento, que premie la recurrencia y la compra de diferentes tipos de productos. La inclusión de los clientes en campanas de comunicación personalizadas, ofertas para diferentes épocas del año, anticipo en los nuevos lanzamientos y participación de pruebas o focus group de nuevos sabores o productos. Estas pueden ser acciones que se enmarcan en incrementar el vínculo relacional de los clientes con la marca.

Además, se podrían elaborar encuestas de satisfacción periódicas y un sistema de descuentos sucesivos por cada referencia comercial exitosa. También podría considerarse la opción de crear una comunidad cerrada de los clientes preferenciales con quienes la empresa compartiría contenido educativo y recetas, o les permita interactuar con el producto, con la consigna de mejorar el sentido de pertenencia y de valor agregado percibido (Ali & Shabn, 2024).

Integración del modelo de segmentación con sistemas empresariales

El presente apartado describe una propuesta de flujo operativo para la implementación del modelo de segmentación K-means dentro de los sistemas de gestión de Proalva. La secuencia de actividades comienza con la recolección semestral de datos desde el sistema ERP, y finaliza con la actualización de tableros de visualización y la puesta en marcha de campañas comerciales en base a los segmentos obtenidos.

La actividad parte de la extracción de transacciones de ventas registradas de los propios sistemas durante los últimos 24 meses, en formato estructurado. Dicha extracción incluirá variables como el código del cliente, la fecha de compra, la cantidad comprada, el valor monetario en euros, el tipo de presentación, el sabor del producto o bien algún otro tipo de atributo relacionado con la operación comercial. Realizada la extracción de datos se procederá a la limpieza de los datos obtenidos eliminando registros duplicados, anulados o irrelevantes para el propio análisis. También se realizarán estandarizaciones de las variables cuantitativas.

Posteriormente, se obtiene un conjunto de datos agregado a nivel de cliente, que incluye las métricas usadas en este proyecto. Este conjunto de datos de entrada será usado para aplicar el algoritmo K-Means, que ejecutamos a partir de los parámetros fijados en el entrenamiento inicial; número de clusters, método de inicialización, semilla aleatoria, etc.

El mismo K-Means asigna a cada cliente una etiqueta de grupo (variable cluster) y permite calcular estadísticas descriptivas por segmento.

Estas salidas las pasamos a procesar mediante (i) un informe con los valores medios de las variables por grupo, y (ii) un dashboard de ventas que introduce la variable cluster al registro de los clientes.

Se puede acoplar esta opción a los sistemas de visualización operativa existentes en la actualidad, como los tableros de ventas, la empresa comentó que utiliza TableaU para hacer seguimiento de las ventas a lo largo del tiempo.

Una vez se ha realizado la segmentación, y han sido actualizados los sistemas, se pueden elaborar las estrategias de marketing en función del modelo de perfiles que haya sido definido. La variable cluster puede ser incorporada al sistema CRM como un campo más que estará disponible en las campañas de comunicación, en las promociones, o en las ofertas. De este modo, cada una de las acciones orientadas a un determinado segmento podrá quedar registrada y asociada a los clientes que hayan sido impactados, con el fin de permitir un posterior análisis. Para dicho análisis es conveniente registrar ciertos indicadores, como la frecuencia de recompra después de la campaña, el valor monetario de compras medio, o la modificación en el número de productos comprados.

Desde la perspectiva tecnológica de la infraestructura, esta secuencia de pasos puede ejecutarse en un servidor virtual configurado a tal efecto, programado para prenderse y ejecutar el código cada seis meses De este modo se consigue evitar la ejecución permanente del sistema, y se establece una rutina programada de procesamiento.

La secuencia de pasos descritos está en consonancia con prácticas de análisis recurrente propias de entornos empresariales que utilizan modelos de segmentación por comportamiento, integrados con plataformas de visualización y ejecución de la actividad comercial.

Limitaciones y potenciales mejoras metodológicas

El procedimiento de segmentar clientes mediante el modelo RFM y el algoritmo K-Means, utilizado junto con la base de datos ofrecida por la empresa Proalva, ha permitido encontrar patrones relevantes sobre el comportamiento de compra de sus clientes, ahora bien, también es cierto que existen unas limitaciones relacionadas con los procedimientos empleados y sería necesario indicar algunas mejoras que favoreciesen la capacidad de los modelos para captar la complejidad de los datos comerciales.

De las limitaciones inherentes al modelo RFM vale la pena mencionar que es de tipo univariado para cada dimensión (recencia, frecuencia y valor monetario), es decir que los vínculos entre las dimensiones no son tenidos en cuenta en el proceso de segmentación. A

su vez, a partir de particiones en percentiles que son elegidas de manera arbitraria por el equipo, puede existir sesgo en la asignación a segmentos. Dicha dependencia en reglas preestablecidas reduce la capacidad que tiene el modelo de responder a la variabilidad de los datos naturales o a cambios en el comportamiento de los clientes a lo largo del tiempo.

Por su parte, el algoritmo K-Means, si bien permite trabajar de forma simultánea con varias variables numéricas y es muy utilizado gracias a su eficiencia computacional, presenta retos relacionados con la forma de las agrupaciones. Este modelo supone que los clusters son esféricos y de iguales tamaños, algo que no se adecúa a la naturaleza real de los datos observados, destacando en este caso los extremos o distribuciones asimétricas. En este estudio, se identificaron dos grupos de tamaño reducido (n=1 y n=2), lo que sugiere que el algoritmo podría ser sensible a datos atípicos o a observaciones con comportamientos extremos. Aunque se decidieron mantener estas observaciones por su relevancia analítica, es importante considerar su influencia en la estabilidad general del modelo.

Otro aspecto clave es la sensibilidad del algoritmo K-Means en relación a la elección inicial de los centroides. A pesar de utilizar la técnica de inicialización k-means++, el proceso de segmentación sigue siendo influenciado por la aleatoriedad inherente. Esto puede dar lugar a particiones que varían ligeramente en diferentes ejecuciones en caso de que se use una semilla aleatoria diferente.

Además, utilizar la distancia euclidiana como principal métrica para asignar observaciones a los grupos puede no ser la mejor opción en todas las circunstancias. Esta medida es especialmente sensible a las escalas de las variables, por lo cual se realizó un escalamiento estándar previo. Sin embargo, si hay variables con distribuciones no gaussianas o con outliers extremos, podría ser beneficioso considerar otras métricas de distancia, como la distancia de Mahalanobis o Manhattan, que pueden ofrecer resultados más adecuados.

En lo que respecta a mejoras de orden metodológico se plantea la posibilidad de incluir nuevos algoritmos de agrupamiento alternativos menos dependientes de supuestos tan estrictos sobre la forma de los clusters. El algoritmo DBSCAN (Density-Based Spatial Clustering of Applications with Noise) puede identificar clusters con forma arbitraria y manejar de una forma más robusta la presencia de ruido y outliers, sin necesidad de tener que definir antes un número de grupos. Modelos de mezcla gaussiana (Gaussian Mixture Models, GMM) podría ser otra opción a explotar para capturar la heterogeneidad del comportamiento de los clientes bajo supuestos probabilistas, generando estimaciones de pertenencia a cada uno de los clusters antes que asignaciones de tipo determinístico.

Otra mejora y propuesta a tener en cuenta sería la de implementar modelos de segmentación dinámica o temporal, que permitan captar los cambios en el comportamiento de los clientes a lo largo del tiempo. En este sentido es posible que métodos como el clustering evolutivo o técnicas con ventanas deslizantes pudiesen adaptarse más a contextos donde las relaciones comerciales varían estacionalmente, en campañas o bien en función de ciclos de producción. Esto podría soportar la posibilidad de construir modelos más acordes con los contextos operativos de la empresa y con un mayor poder predictivo sobre cambios a introducir en la composición de la base de clientes.

Por último, se recomienda valorar un análisis de validación cruzada del modelo de clustering, aplicando técnicas de resampling o bootstrap con el fin de estimar las propiedades de estabilidad y consistencia de las particiones obtenidas. Hacer esto permitiría evaluar la robustez de los segmentos generados frente a variaciones en el dataset de clientes y fortalecer la fiabilidad de los resultados antes de su uso en decisiones estratégicas.

Capítulo 5: Conclusiones y Recomendaciones

Conclusiones

La metodología de segmentación de clientes propuesta para la empresa Proalva fue ejecutada de forma eficiente tomando en cuenta sus datos históricos de ventas, y que podrían estar disponibles para la mayoría de MIPYMES formales en el país.

La aplicación de modelo RFM y el algoritmo K-Means sobre los datos agregados por cliente, caracterizados a partir de las transacciones de ventas y comprendidos en un periodo de 24 meses, pudo generar la posibilidad de identificar patrones relevantes en el comportamiento de compra de los clientes de la empresa, lo cual es de mucha utilidad al momento de generar y optimizar las estrategias comerciales.

Mediante el modelo clásico RFM, fue posible segmentar los clientes en 6 grupos (Estrella, Prometedor, Buen Cliente, Regular, Leal pero inactivo y Perdido) partiendo de las tres variables clásicas: recencia, frecuencia y valor monetario.

Si bien la facilidad de uso de este modelo representa una ventaja en términos de capacidades técnicas de una empresa, también implica limitaciones. El modelo RFM asume una estructura uniforme dentro de cada grupo, sin considerar otras dimensiones relevantes del comportamiento de compra como la diversificación de la cartera de productos que consumen los clientes. Los umbrales utilizados para definir las categorías parte de decisiones discrecionales de parte del equipo técnico que las define.

El uso del algoritmo K-Means permitió tener una segmentación más granular ya que el modelo pudo identificar clusteres diferentes, incluyendo grupos de alto volumen de compra que no se pudieron identificar mediante el uso del RFM. Esta capacidad es un factor destacable del modelo K-Means y la importancia de su uso para crear estrategias personalizadas de marketing.

El análisis comparativo entre ambos métodos mostró ciertas coincidencias, pero también reveló discrepancias importantes, que evidencian el valor añadido de incorporar técnicas de aprendizaje no supervisado.

Es importante señalar que ambos enfoques se complementan en su uso para el propósito de crear segmentos de clientes. Este uso complementario permite hacer una exploración a detalle de los patrones de compra, revelando información útil para guiar una toma de decisiones basada en datos.

El modelo no supervisado logró detectar a aquellos clientes con comportamientos extremos en cuanto a volumen o frecuencia de compra, cosa que no era posible siguiendo el esquema discrecional del modelo tradicional RFM. Esta diferencia invita a reforzar la necesidad de recurrir a modelos complementarios a la hora de estudiar el comportamiento de los clientes.

La caracterización de perfil de los clientes también permitió adoptar líneas estratégicas de marketing para cada grupo de clientes, dando como resultado un enfoque aplicable a un contexto real de tipo empresarial, con información limitada pero estructurada, la cual puede ser la que proviene de sistemas tipo ERP, haciendo ver que el enfoque puede ser útil también para otras MIPYMES que quieran reforzar sus capacidades analíticas y comerciales.

Recomendaciones.

Aplicar una estrategia comercial por clúster, dado que la segmentación que proporciona el modelo K-Means nos ofrece una base empírica para diseñar acciones comerciales específicas a cada grupo. Por ejemplo, para los clientes del clúster 5 (alto potencial) se puede sugerir la aplicación de programas de fidelización o incentivos por volumen, mientras que para los clientes del clúster 1 (bajo valor y alta recencia) podrían plantearse campañas reactivadoras específicas únicamente cuando existan señales de compromiso previas.

Medir periódicamente los segmentos definidos. Dado que los hábitos de compra pueden cambiar en el tiempo se recomienda refrescar la segmentación, al menos, una vez cada 6-12 meses, dependiendo de la periodicidad con la que se produzcan las ventas: se

podría evaluar, en este proceso, el paso de clientes de un clúster a otro, el nivel de aparición de nuevos clientes o gestionar variables adicionales para la base de clientes.

Ampliar el análisis con variables contextuales externas. Al incorporar información adicional de los clientes, como pueden ser su localización geográfica, su clasificación tributaria, su sector económico o su comportamiento en el contexto de pago, es posible obtener una segmentación más precisa y, por tanto, también una mejor capacidad del modelo; este proceso podrá llevarse a cabo utilizando información pública (la correspondiente al SRI, por ejemplo) o mediante encuestas complementarias.

Evaluar la efectividad de las campañas ejecutadas. Tras haber puesto en práctica las estrategias diferenciadas, resulta aconsejable establecer indicadores de desempeño por clúster (como puede ser la tasa de recompra, la variación en la facturación promedio, la frecuencia de recompra, la conversión de clientes inactivos, etc.). Esta retroalimentación permitirá ajustar las estrategias o criterios del modelo.

Explorar otras maneras de segmentar o complementarias. En futuras aplicaciones, se pueden considerar modelos de clustering basados en densidad (como DBSCAN) o probabilísticos (como Gaussian Mixture Models), con el fin de contar con un mayor grado de flexibilidad frente a distribuciones que no son esféricas o presentan ruido. Junto con esto, se podría incorporar una herramienta de reducción dimensional con la intención de ayudar a validar visualmente los agrupamientos generados.

Integrar la lógica de la segmentación en los sistemas de la compañía. Se aconseja el desarrollo de herramientas internas (por ejemplo, dashboard interactivos o servidores pequeños en nube) que permitan automatizar el cálculo de los segmentos y que analicen el comportamiento de los clientes en periodos de tiempo definido. Un flujo de trabajo de este tipo permitiría una toma de decisiones más rápida basada en la información disponible.

Entrenar al personal comercial en cómo utilizar las segmentaciones obtenidas de este proyecto. Para poder sacar todo el potencial de este tipo de modelos, es importante que los equipos de ventas, atención al cliente o marketing tengan conocimiento de cómo

funciona la lógica de la segmentación y sean capaces de interpretar los resultados para tomar decisiones óptimas.

Referencias Bibliográficas

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Ali, N., & Shabn, O. S. (2024). Customer lifetime value (CLV) insights for strategic marketing success and its impact on organizational financial performance. *Cogent Business & Management*, *11*(1). https://doi.org/10.1080/23311975.2024.2361321

Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and E-Business Management*, *21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

Andy Hermawan, Nila Rusiardi Jayanti, Aji Saputra, Army Putera Parta, Muhammad Abizar Algiffary Thahir, & Taufiqurrahman Taufiqurrahman. (2024). Leveraging the RFM Model for Customer Segmentation in a Software-as-a-Service (SaaS) Business Using Python. *Maeswara : Jurnal Riset Ilmu Manajemen Dan Kewirausahaan*, *2*, 77–89. https://doi.org/10.61132/maeswara.v2i5.1283

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, *46*, 243–256. https://doi.org/10.1016/j.patcog.2012.07.021

Armijos De La Cruz, B. A., & Palacios Meléndez, J. G. (2024). Madurez Digital del E-Commerce y Desarrollo Empresarial MIPYMES, Provincia de Santa Elena. *Ciencia Latina Revista Científica Multidisciplinar*, *8*, 11275–11291. https://doi.org/10.37811/cl_rcm.v8i5.14509

Arthur, D., & Vassilvitskii, Sergei. (2007). K-Means++: The Advantages of Careful Seeding. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms.* , 1027–1035.

Cálad Noreña, F. (2015). *Segmentación de clientes automatizada a partir de técnicas de minería de datos (K-means clustering)* [Tesis de pregrado]. Escuela de Ingeniería de Antioquia.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, *26-June-2016*, 2201–2206. https://doi.org/10.1145/2882903.2912574

Clarke, A. H., & Freytag, P. V. (2023). Implementation of new segments in small- and medium-sized enterprises (SMEs). *Journal of Business and Industrial Marketing*, *38*, 930–942. https://doi.org/10.1108/JBIM-01-2021-0053

Dini, M., Patiño, A., & Gligo S., N. (2021). *Transformación digital de las mipymes: elementos para el diseño de políticas*.

Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis* (pp. 11–22). https://doi.org/10.1007/978-981-10-8818-6_2

Durugkar, S. R., Raja, R., Nagwanshi, K. K., & Kumar, S. (2022). Introduction to Data Mining. In *Data Mining and Machine Learning Applications* (pp. 1–19). Wiley. https://doi.org/10.1002/9781119792529.ch1

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley. https://doi.org/10.1002/9780470977811

Fredyna, T., Ruíz-Palomo, D., & Dieguez, J. (2019). Entrepreneurial orientation and product innovation. The moderating role of family involvement in management. *European Journal of Family Business*, *9*, 128–145. https://doi.org/10.24310/ejfbejfb.v9i2.5392

Gabriela Jara, A., & Sayonara Solorzano, S. (2024). Estrategias Comerciales en las Pymes Ecuatorianas del Sector Agroindustrial, para pertenecer a los Clústeres de Franquicias Internacionales en el Ecuador en el Año 2024. *Ciencia Latina Revista Científica Multidisciplinar*, *8*, 13423–13452. https://doi.org/10.37811/cl_rcm.v8i4.13663

Gorgoglione, M., Panniello, U., & Tuzhilin, A. (2019). Recommendation strategies in personalization applications. *Information & Management*, *56*(6), 103143. https://doi.org/10.1016/j.im.2019.01.005

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, *40*(16), 6491–6507. https://doi.org/10.1016/j.eswa.2013.05.052

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

INEC. (2023). *Registro Estadístico de Empresas 2022 Principales Resultados*. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Economicas/Registro_Empresas_Establecimientos/2022/Semestre_II/Principales_Resultados_REEM_2022.pdf

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kodinariya, T. M., & Makwana, P. R. (2013). Review on Determining of Cluster in K-means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(6), 90–95.

Konstantinos, T., & Antonios, C. (2009). *Data Mining Techniques in CRM: Inside Customer Segmentation*. John Wiley & Sons, Ltd.

Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th edition). Pearson.

Kumar, N. (2025). Intelligent customer segmentation: unveiling consumer patterns with machine learning. *Journal of Umm Al-Qura University for Engineering and Architecture*. https://doi.org/10.1007/s43995-025-00180-7

Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management Concept, Strategy, and Tools* (3rd ed.). Springer .

Laudon, K., & Laudon, J. (2020). *Management Information Systems* (16th ed.). Pearson Education.

Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, *28*, 129–137. https://doi.org/10.1109/TIT.1982.1056489

López, O. G. (2024, February). Análisis de Segmentación de Clientes Usando K-Means: Una Guía Práctica. *Https://Medium.Com/@oriolgilabertlopez/An%C3%A1lisis-de-Segmentaci%C3%B3n-de-Clientes-Usando-k-Means-Una-Gu%C3%ADa-Pr%C3%A1ctica-72ff735ba9e7*.

Mahdee, N. (2022). *Customer Segmentation Using K-means* [B.Sc. in Computer Science and Engineering]. BRAC University.

Martínez, J., Romo, L., & Riascos, S. (2024). Avances en la transformación digital de las MiPymes impulsadas por la pandemia COVID-19. *Journal of Technology Management & Innovation*, *19*(1), 52–65. https://doi.org/10.4067/S0718-27242024000100052

Mora Cortez, R., Højbjerg Clarke, A., & Freytag, P. V. (2021). B2B market segmentation: A systematic review and research agenda. *Journal of Business Research*, *126*, 415–428. https://doi.org/10.1016/j.jbusres.2020.12.070

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2012). Scikit-learn: Machine Learning in Python. . *Journal of Machine Learning Research.* , *12*.

Pinho, J. C., & Prange, C. (2016). The effect of social networks and dynamic internationalization capabilities on international performance. *Journal of World Business*, *51*(3), 391–403. https://doi.org/10.1016/j.jwb.2015.08.001

Plazas Cárdenas, L. P., & Plazas Cárdenas, J. E. (2013). *Aplicación de minería de datos para la segmentación de clientes que compran materias primas derivadas del maíz para la generación de estrategias de comunicación* [Tesis de pregrado]. Universidad Piloto de Colombia.

Rahim, M. A., Mushafiq, M., Khan, S., & Arain, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation. *Journal of Retailing and Consumer Services*, *61*, 102566. https://doi.org/10.1016/j.jretconser.2021.102566

Rajagukguk, R. (2025). Implementation of Loyalty Program Theory Based on Recency Frequency Monetary Score in Information Systems to Increase Customer Loyalty. *Journal of Information System Exploration and Research*, *3*, 45–52. https://doi.org/10.52465/joiser.v3i1.538

Ramkumar, G., Bhuvaneswari, J., Venugopal, S., Kumar, S., Ramasamy, C. K., & Karthick, R. (2025). Enhancing customer segmentation: RFM analysis and K-Means clustering implementation. In *Hybrid and Advanced Technologies* (pp. 70–76). CRC Press. https://doi.org/10.1201/9781003559139-9

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. https://doi.org/10.1016/0377-0427(87)90125-7

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K., & Muangprathub, J. (2024). RFM model customer segmentation based on hierarchical approach using FCA. *Expert Systems with Applications*, *237*, 121449. https://doi.org/10.1016/j.eswa.2023.121449

Safari, F., Safari, N., & Montazer, G. A. (2016). Customer lifetime value determination based on RFM model. *Marketing Intelligence & Planning*, *34*(4), 446–461. https://doi.org/10.1108/MIP-03-2015-0060

Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024a). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies*, *6*, 54–60. https://doi.org/10.32996/jbms.2024.6.2.5

Scikit-learn developers. (2024). *Scikit-learn Documentation (Version 1.4).* Https://Scikit-Learn.Org/Stable/Whats_new/v1.4.Html.

Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer Segmentation using RFM Model and K-Means Clustering. *International Journal of Scientific Research in Science and Technology*, 591–597. https://doi.org/10.32628/ijsrst2183118

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. *Proceedings - 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018*, 119–126. https://doi.org/10.1109/ICEBE.2018.00027

Terra, J. (2024, November). Why Use Python for Data Science? *Https://Pg-p.Ctme.Caltech.Edu/Blog/Data-Science/Why-Use-Python-for-Data-Science*, *Caltech Bootcamp*.

Tsiptsis Antonios Chorianopoulos WILEY, K. (n.d.). *Data Mining Techniques in CRM: Inside Customer Segmentation*.

Tsiptsis, K., & Chorianopoulos, A. (2010). *Data Mining Techniques in CRM*. Wiley. https://doi.org/10.1002/9780470685815

Uquillas Granizo, G. G. (2024). Dificultades para la internacionalización de negocios y empresas en el Ecuador. *Perspectivas Sociales y Administrativas*, *2*(1), 16–26. https://doi.org/10.61347/psa.v2i1.62

Valencia, J. A. (2024, October). Marketing y Ventas: ¿un problema o una oportunidad? *Revista Perspectiva*. https://perspectiva.ide.edu.ec/investiga/wp-content/uploads/2024/10/Perspectiva-2024-10-1.pdf

Wang, S., Sun, L., & Yu, Y. (2024). A dynamic customer segmentation approach by combining LRFMS and multivariate time series clustering. *Scientific Reports*, *14*(1), 17491. https://doi.org/10.1038/s41598-024-68621-2

Wedel, M., & Kamakura, W. A. (2000). *Market Segmentation* (Vol. 8). Springer US. https://doi.org/10.1007/978-1-4615-4651-1

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Woodburn, D., & McDonald, M. (Eds.). (2012). *Key Account Management*. Wiley. https://doi.org/10.1002/9781119207252

Wulansari, S., & Heikal, J. (2024). Analysis Of Customer Segmentation In The Top Three Most Visited E-Commerce Platforms In Indonesia In 2023 Using RFM Model And Clustering Techniques. *Jurnal Scientia*, *13*(03).

Yahya, M., Parenreng, J. M., Fathahillah, F., Wahid, A., Wahid, M. S. N., & Fajar B, M. (2024). A Machine Learning Model for Local Market Prediction Using RFM Model. *Elinvo (Electronics, Informatics, and Vocational Education)*, *9*(1), 24–37. https://doi.org/10.21831/elinvo.v9i1.58671

Zhang, C., Lin, S., & Zhang, D. (2022). A Data Cleaning Method for Industrial Data Flow Based on Multistage Combinational Optimization of Rule Set. *ACM International Conference Proceeding Series*, 77–81. https://doi.org/10.1145/3561801.3561814

Zhao, J. (2024). Customer segmentation application based on K-Means. *Applied and Computational Engineering*, *47*, 242–247. https://doi.org/10.54254/2755-2721/47/20241400

Apéndice

Apéndice A. Nombre de las variables del dataset y descripción

Tabla A1

Columnas, número de observaciones nulas y formato de la columna del dataset original

Index	Column	Non-Null Count	Dtype
1	Cliente_codigo	141035 non-null	object
2	CodProducto	141035 non-null	object
3	DescripcionProducto	141035 non-null	object
4	CantidadItem	141035 non-null	float64
5	Unidad	141035 non-null	object
6	PrecioItem	141035 non-null	float64
7	TotalItem	141035 non-null	float64
8	IdentificadorVenta	141035 non-null	object
9	Fecha	141035 non-null	datetime64[ns]
10	Tipo	141035 non-null	object
11	Estado	141035 non-null	object
12	Retencion	98037 non-null	object
13	Asiento	141035 non-null	object
14	DescItem	98037 non-null	object
15	Subt0	81163 non-null	float64
16	SubtIva	98037 non-null	object
17	Descuento	81163 non-null	float64
18	Iva	141035 non-null	object
19	Total	141035 non-null	float64
20	FechaDet	141035 non-null	datetime64[ns]
21	anio	141035 non-null	int64
22	mes	141035 non-null	int64
23	dia	141035 non-null	int64
24	NOMBRE2	140992 non-null	object
25	CodSabor	140992 non-null	float64
26	NomSabor	140991 non-null	object
27	GRUPO O CLASIFICACION	140992 non-null	object
28	MARCA	140918 non-null	object
29	SUB CLASIFICACION	140970 non-null	object
30	CantKgPulpaItem	139959 non-null	float64
31	mesanio	141035 non-null	object

Apéndice B. Nombre de las variables del dataset para la aplicación de los modelos de segmentación

Tabla A2

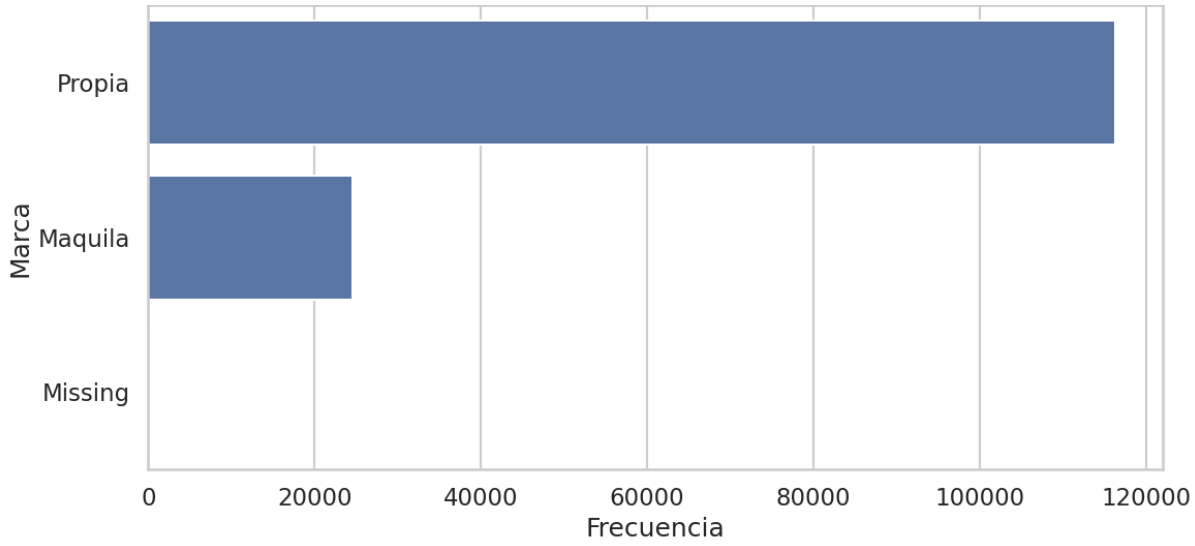
Columnas, número de observaciones nulas y formato de la columna del dataset agregado por cliente

Index	Column	Non-Null Count	Dtype
1	Cliente_codigo	340 non-null	int64
2	Recency	340 non-null	int64
3	Frequency	340 non-null	int64
4	Monetary	340 non-null	float64
5	FechaUltimaCompra	340 non-null	datetime64[ns]
6	NumSabores	340 non-null	int64
7	TotalKg	340 non-null	float64
8	NumPresentaciones	340 non-null	int64
9	NumDiasDistintos	340 non-null	int64
10	PropMarcaPropia	340 non-null	float64
11	PropMarcaMaquila	340 non-null	float64
12	ProdFav_ARAZA	340 non-null	bool
13	ProdFav_BABACO	340 non-null	bool
14	ProdFav_COCO	340 non-null	bool
15	ProdFav_DURAZNO	340 non-null	bool
16	ProdFav_FRUTILLA	340 non-null	bool
17	ProdFav_GUANABANA	340 non-null	bool
18	ProdFav_GUAYABA	340 non-null	bool
19	ProdFav_LIMON	340 non-null	bool
20	ProdFav_MANGO	340 non-null	bool
21	ProdFav_MANZANA	340 non-null	bool
22	ProdFav_MARACUYA	340 non-null	bool
23	ProdFav_MIX FRUTAL	340 non-null	bool
24	ProdFav_MORA	340 non-null	bool
25	ProdFav_NARANJILLA	340 non-null	bool
26	ProdFav_PIÑA	340 non-null	bool
27	ProdFav_TAMARINDO	340 non-null	bool
28	ProdFav_TOMATE	340 non-null	bool

Apéndice C. Frecuencia de registros según variable “Marca”

Figura 13

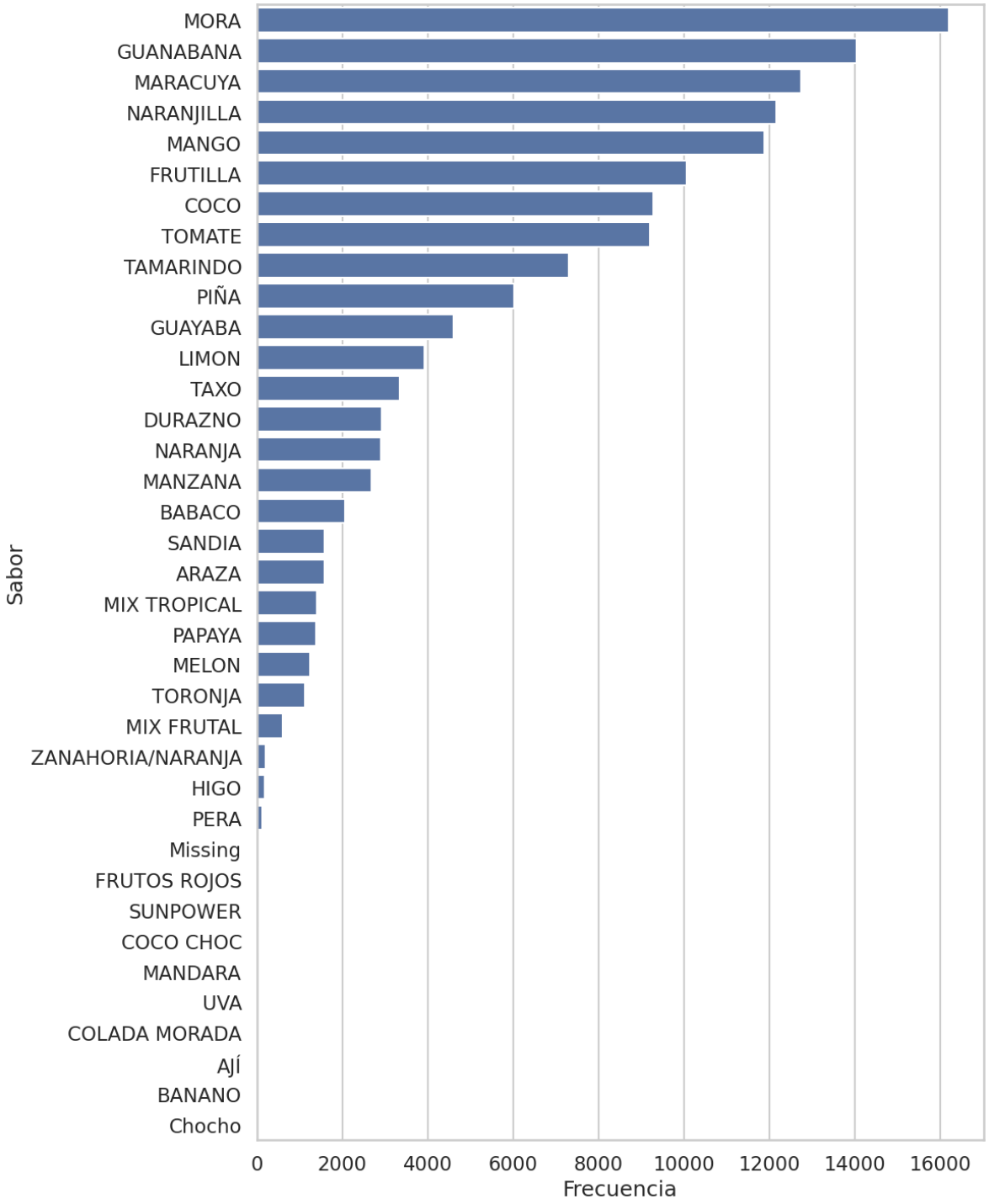
Frecuencia de registros de ventas según la marca del producto



Apéndice D. Frecuencia de registros según variable “NomSabor”

Figura 14

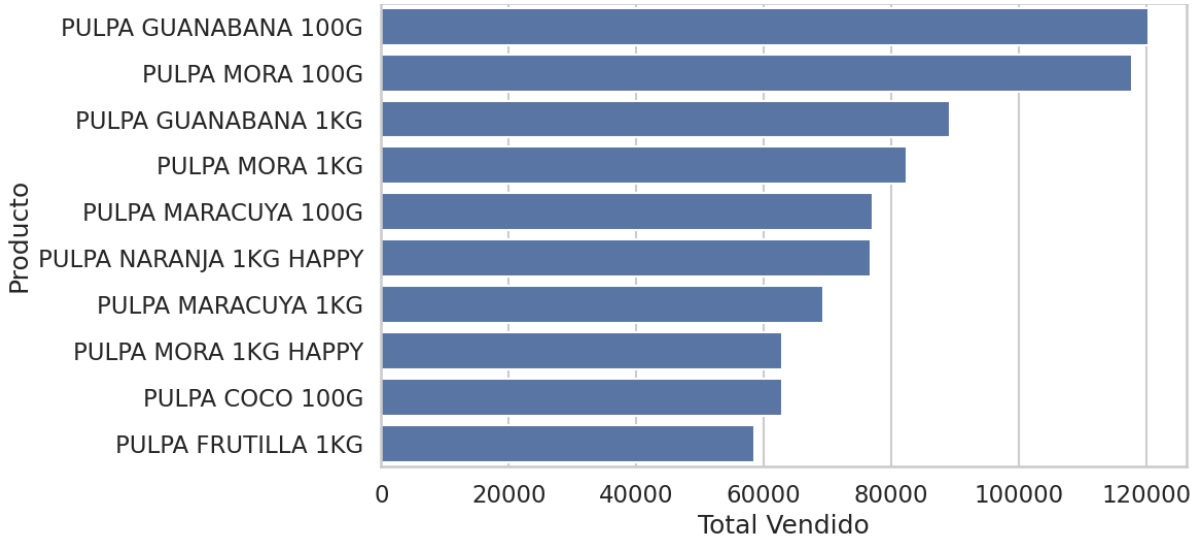
Frecuencia de registros de ventas según el sabor (fruta) del producto



Apéndice E. Top 10 productos por valor vendido

Figura 15

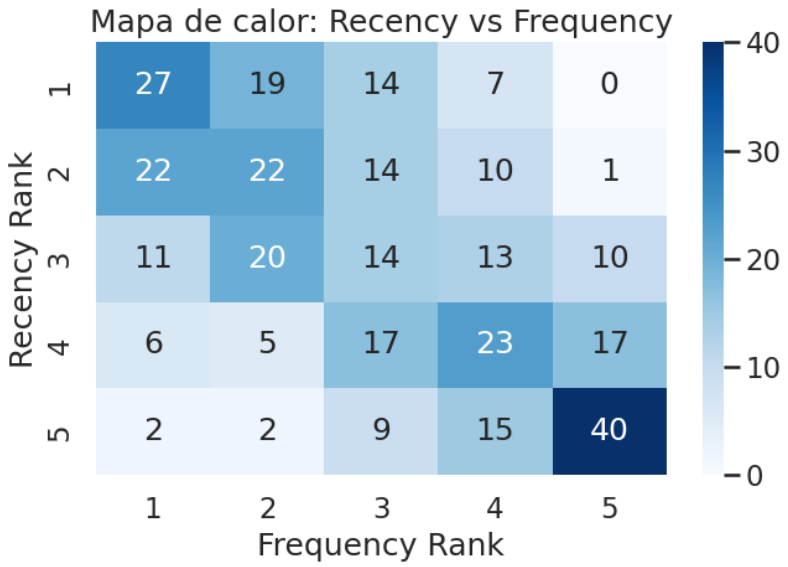
10 principales productos vendidos en función al valor monetario



Apéndice F. Clasificación según quintiles de R y F

Figura 16

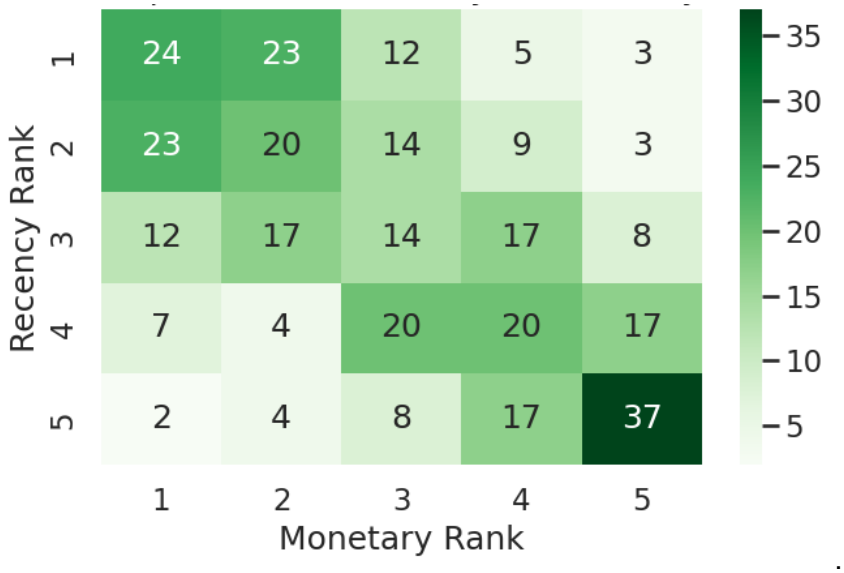
Mapa de calor de clientes según quintil R y F



Apéndice G. Clasificación según quintiles de R y M

Figura 17

Mapa de calor de clientes según quintil R y M



Apéndice H. Clasificación según quintiles de F y M

Figura 18

Mapa de calor de clientes según quintil F y M