## TÍTULO

# ANÁLISIS DE SEGMENTACIÓN DE CLIENTES EN VENTAS MINORISTAS MEDIANTE APRENDIZAJE AUTOMÁTICO

## AUTOR

José Torroba Moreno

**Esta edición electrónica ha sido realizada en 2024**

| | |
|---|---|
| Tutores | Dr. D. Diego Marín Santos ; Dr. D. Manuel Emilio Gegúndez Arias |
| Institución | Universidad Internacional de Andalucía |
| Curso | *Máster de Formación Permanente en Big Data (2022/23)* |
| © | José Torroba Moreno |
| © | De esta edición: Universidad Internacional de Andalucía |
| Fecha documento | 2023 |

Universidad Internacional de Andalucía


Para más información:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en

# Análisis de Segmentación de Clientes en Ventas Minoristas mediante Aprendizaje Automático

José Torroba Moreno

A thesis submitted in conformity with the requirements
for the MSc in Big Data

International University of Andalusia

October 2023

# Análisis de Segmentación de Clientes en Ventas Minoristas mediante Aprendizaje Automático

José Torroba Moreno

Máster en Big Data

Supervisores: Dr. Diego Marín Santos, Dr. Manuel Emilio Gegúndez Arias
Universidad Internacional de Andalucía

2023

## Abstract

In the retail sector, customer segmentation plays a pivotal role in understanding buying behavior and designing effective marketing strategies. This study addresses an analysis of consumer goods sales in a retail establishment using unsupervised machine learning techniques. With a dataset containing over 64.000 transactions from 22.000 customers over the course of a year, algorithms like K-Means, Hierarchical Clustering, and DBSCAN are applied to segment the store's audience based on recency, frequency, and investment. A valuable insight into purchase patterns and customer loyalty is provided, offering potential for enhancing business strategies in the retail sector.

**JEL classification**: C4, I25, J28, L81, M31, O52.

**Key words**: RFM Analysis, Customer Segmentation, Retail Sales, Machine Learning, K-Means Clustering, Hierarchical Clustering, DBSCAN Clustering.

# Resumen

En el ámbito minorista, la segmentación de clientes desempeña un papel crucial para comprender el comportamiento de compra y diseñar estrategias de marketing efectivas. Se aborda un análisis de ventas de bienes de consumo en un establecimiento minorista, utilizando técnicas de aprendizaje automático no supervisado. Con un conjunto de datos que abarca más de 64.000 transacciones de 22.000 clientes durante un año, se aplican algoritmos como K-Means, Hierarchical Clustering y DBSCAN, para segmentar el público del comercio según su recencia, frecuencia e inversión. Se proporciona una visión valiosa de los patrones de compra y la lealtad del cliente, aportando potencial para mejorar las estrategias comerciales en el sector minorista.

**Clasificación JEL**: C4, I25, J28, L81, M31, O52.

**Palabras clave**: Análisis RFM, Segmentación de Clientes, Ventas Minoristas, Aprendizaje Automático, Agrupamiento K-Means, Agrupamiento Jerárquico, Agrupamiento DBSCAN.

# Tabla de Contenidos

Tabla de Contenidos ....4

Lista de Tablas ....5

Lista de Figuras ....6

1 Propuesta TFM ....7

1.1 Motivación TFM ....7

1.2 Objetivos Generales y Específicos ....7

1.3 Estructura del documento ....8

2 Introducción ....9

3 Marco teórico ....11

3.1 Aprendizaje automático: supervisado vs no supervisado ....11

3.2 Técnicas de agrupamiento seleccionadas ....13

3.2.1 K-Means ....13

3.2.2 Hierarchical Clustering ....14

3.2.3 DBSCAN ....15

4 Materiales ....16

4.1 Descripción de la base de datos ....16

4.2 Análisis de variables ....17

5 Experimentación ....19

5.1 Preprocesamiento de datos ....19

5.2 Resultados ....23

5.3 Análisis de resultados ....36

6 Conclusiones ....37

7 Referencias ....38

# Lista de Tablas

Tabla 1. Estructura de los datos. ..... 18

Tabla 2. Estadística descriptiva del conjunto de datos. ..... 18

Tabla 3. Conjunto de datos RFM ..... 20

Tabla 4. Grupos obtenidos aplicando K-Means y cantidad de clientes por grupo. ..... 24

Tabla 5. Grupos obtenidos aplicando K-Means y cantidad de clientes por grupo. ..... 29

Tabla 6. Grupos obtenidos aplicando DBSCAN y cantidad de clientes por grupo ..... 35

# Lista de Figuras

Figura 1. Distribución de las variables Inversión, Frecuencia, Recencia obtenidas. ..... 21

Figura 2. Resultados del cálculo de la inercia y el Silhouette Score para valores de k comprendidos entre 2 y 8. ..... 23

Figura 3. Distribución de la variable Frecuencia en los grupos de clientes obtenidos (K-Means). ..... 24

Figura 4. Distribución de la variable Inversión en los grupos de clientes obtenidos (K-Means). 25

Figura 5. Distribución de la variable Recencia en los grupos de clientes obtenidos (K-Means). 26

Figura 6. Representación tridimensional de los grupos de clientes obtenidos mediante K-Means. ..... 27

Figura 7. Dendograma obtenido al aplicar Hierarchical Clustering. ..... 28

Figura 8. Distribución de la variable Recencia en los grupos de clientes obtenidos (Hierarchical Clustering)..... 30

Figura 9. Distribución de la variable Frecuencia en los grupos de clientes obtenidos (Hierarchical Clustering)..... 31

Figura 10. Distribución de la variable Inversión en los grupos de clientes obtenidos (Hierarchical Clustering)..... 32

Figura 11. Representación tridimensional de los grupos de clientes obtenidos mediante Hierarchical Clustering. ..... 33

Figura 12, Representación tridimensional de los grupos de clientes obtenidos mediante DBSCAN. ..... 34

# 1 Propuesta TFM

## 1.1 Motivación TFM

Dada la constante evolución y la creciente competitividad en el sector comercial, impulsada en gran medida por la digitalización, en la actualidad los competidores comerciales se enfrentan a un entorno empresarial altamente dinámico, donde la capacidad de recoger, almacenar y procesar datos se ha convertido en un recurso de alto valor.

La aplicación de técnicas de aprendizaje automático se ha vuelto factible gracias a esta capacidad de manejar grandes volúmenes de datos. Esto permite a las empresas comprender y segmentar a sus clientes de manera más precisa, lo que a su vez tiene un impacto positivo en la aplicación de técnicas de marketing y estrategias comerciales. Al conocer mejor a sus clientes, las empresas pueden personalizar sus ofertas, optimizar sus estrategias de ventas y, en última instancia, maximizar sus beneficios.

En este Trabajo de Fin de Máster se lleva a cabo un análisis de comportamiento de clientes, en el que se emplean técnicas de aprendizaje automático para ilustrar el valor potencial que existe en los datos cotidianos en un establecimiento minorista común.

## 1.2 Objetivos Generales y Específicos

El objetivo general de este trabajo consiste en encontrar los grupos similares de clientes que pudieran existir en datos analizados de un establecimiento concreto, por medio de la aplicación de algoritmos de aprendizaje automático no supervisado.

Se plantean los siguientes objetivos específicos:

- Análisis previo de los datos según RFM (Recencia, Frecuencia, Inversión) de los clientes.
- Aplicación de algoritmos de aprendizaje automático no supervisado
- Obtención de clientes segmentados e interpretación de hallazgos en los resultados obtenidos.

En primer lugar, se realiza un análisis RFM (Recencia, Frecuencia, Inversión) al conjunto bruto de datos, lo que permite una comprensión de la relación de los clientes con el establecimiento

minorista. Este enfoque permite determinar la frecuencia y la actualidad de las visitas de los clientes y evaluar el valor monetario de sus compras. Al aplicar el análisis RFM, se realiza un tratamiento de las fechas de compra de los clientes y su cantidad invertida en compras, resultando en un conjunto de datos que se utiliza en la aplicación de los algoritmos.

Luego, se aplican técnicas de aprendizaje automático no supervisado, incluyendo algoritmos como K-Means, Hierarchical Clustering y DBSCAN. Se utiliza Python para la aplicación a nivel práctico de estos algoritmos y visualizar los resultados que separan a los clientes en grupos homogéneos.

La interpretación de estos resultados proporciona información valiosa para la toma de decisiones estratégicas más eficaces, de aplicación en campañas publicitarias y de marketing.

## 1.3 Estructura del documento

Además de este primer capítulo de propuesta de Trabajo de Fin de Máster, este documento se ha organizado en los siguientes capítulos:

- Capítulo 2. Introducción. En este capítulo se hace una introducción contexto, explorando literatura relevante sobre la importancia del análisis de ventas minoristas y las técnicas de aprendizaje automático utilizadas en este estudio y las aplicaciones que otros autores han realizado en la segmentación de clientes.

- Capítulo 3. Marco Teórico. En este capítulo se describen el Aprendizaje Supervisado y el No Supervisado. Se explican con detalle los algoritmos contenidos en este último tipo y que se aplican en este Trabajo de Fin de Máster.

- Capítulo 4. Materiales. En este capítulo se describe la base de datos, se explica el análisis previo y el preprocesamiento que se aborda. Se analizan las variables para entender el conjunto de datos que se trata y su origen.

- Capítulo 5. Experimentación. En este capítulo se analizan e interpretan los resultados obtenidos, dándole significado a la segmentación de clientes obtenida en cada caso.

- Capítulo 6. Conclusiones. En este capítulo se plasman las conclusiones, denotando la importancia de la segmentación de clientes y cómo el uso de los algoritmos de Clustering aporta valor en el sector.

El estudio está organizado abordando primeramente una introducción de contexto, explorando literatura relevante sobre la importancia del análisis de ventas minoristas y las técnicas de aprendizaje automático utilizadas en este estudio.

Más adelante se profundiza en los fundamentos del aprendizaje automático, destacando las diferencias entre el aprendizaje supervisado y no supervisado, y se presenta una descripción detallada de los algoritmos de agrupamiento seleccionados.

Seguidamente se profundiza en los fundamentos del aprendizaje automático, destacando las diferencias entre el aprendizaje supervisado y no supervisado, y se presenta una descripción detallada de los algoritmos de agrupamiento seleccionados

Posteriormente se revisan los materiales, donde se aborda la descripción de la base de datos utilizada y se analizan sus características También se lleva a cabo un análisis detallado de las variables involucradas.

Luego se presenta la metodología utilizada en el análisis, incluyendo la preparación de datos y la aplicación de los algoritmos de aprendizaje automático. Se discuten los resultados obtenidos, seguidos de un análisis de los mismos.

Por último, se resumen los hallazgos conseguidos en el estudio y se extraen conclusiones relevantes.

# 2 Introducción

En un mundo cada vez más interconectado y digitalizado, la información se ha convertido en un activo alto valor. Los datos, su análisis y la interpretación de la información recogida, se han convertido en un recurso esencial para la toma de decisiones en una amplia gama de campos [1].

La Ciencia de Datos en general y el Aprendizaje Automático en particular, han experimentado un gran auge en los últimos años. Este incremento de uso y aplicabilidad, ha sido impulsado por una combinación de factores, que incluyen la disponibilidad masiva de datos, el aumento en el poder de cómputo en casi cualquier dispositivo, los avances en algoritmos, y la creciente conciencia de su potencial [2]. El Aprendizaje Automático en concreto, destaca como herramienta que habilita

el poder extraer conocimiento y tomar decisiones informadas a partir de datos masivos y complejos.

Una de las tendencias más notables es el surgimiento del Aprendizaje Profundo (Deep Learning), que han demostrado un rendimiento excepcional en tareas de visión por computadora, procesamiento de lenguaje natural y más. Redes neuronales profundas, como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), han superado con creces las expectativas en aplicaciones que van desde la detección de fraudes [3] hasta la traducción automática.

Una de las aplicaciones más relevantes del Aprendizaje Automático no supervisado, consiste en la segmentación de clientes a partir de datos sin clasificar para detectar patrones similares entre los clientes y lograr separarlos según sus hábitos y comportamiento [4]. Esto resulta en generación de conocimiento de alto valor para impulsar estrategias de marketing y retención más efectivas [5], [6].

La segmentación utilizando esta metodología analiza y agrupa los comportamientos de los clientes. Esto se logra por medio de algoritmos populares y extendidos en el ámbito del Aprendizaje Automático; K-Means resulta ser generalmente una buena aproximación [7] ampliamente conocida. Sin embargo, es una buena práctica intentar múltiples algoritmos diferentes, ya que diferentes propiedades de los datos pueden quedar más claras al aplicar más de una técnica de Clustering [8]. También hay que notar que, cada algoritmo puede mostrar limitaciones o ventajas: K-Means es eficiente en conjuntos de datos grandes y multidimensionales, pero no trabaja demasiado bien si existe ruido en los datos, cuestión con la que no tiene dificultades DBSCAN [9]. El Hierarchical Clustering por otro lado, es capaz de manejar datos con formas más irregulares, pero resulta más costoso computacionalmente y la jerarquía de grupos puede ser difícil de visualizar en conjuntos de datos complejos [10].

# 3 Marco teórico

## 3.1 Aprendizaje automático: supervisado vs no supervisado

En Aprendizaje Automático podemos encontrar dos metodologías diferentes, basadas en las diferencias fundamentales en la forma en que los algoritmos de aprendizaje interactúan con los datos y aprenden de ellos. Entre ellas, se distinguen el aprendizaje supervisado y no supervisado.

En el Aprendizaje Supervisado (Supervised Learning) los algoritmos se entrenan utilizando un conjunto de datos etiquetado. En este proceso, el conjunto de datos de entrenamiento contiene ejemplos que consisten en pares de entrada y salida esperada. Estas salidas esperadas pueden ser etiquetas o valores conocidos relacionados con las entradas correspondientes. El objetivo principal del Aprendizaje Supervisado es que el algoritmo aprenda una función o patrón que pueda mapear de manera efectiva las entradas a las salidas esperadas.

Una vez que el modelo se ha entrenado utilizando este conjunto de datos etiquetado, se vuelve capaz de hacer predicciones precisas sobre nuevos datos no etiquetados. En otras palabras, el algoritmo busca identificar patrones y relaciones entre las entradas y las salidas conocidas durante el entrenamiento y luego aplicar ese conocimiento para realizar predicciones en situaciones en las que las salidas no están etiquetadas previamente. Esta aproximación es ampliamente utilizada en una variedad de aplicaciones, como clasificación de correos electrónicos como spam o no spam, detección de fraudes en transacciones financieras y diagnóstico médico, entre otros. Ejemplos de técnicas en aprendizaje supervisado incluyen Regresión Lineal, Regresión Logística, Máquinas de Soporte Vectorial (SVM), y Redes Neuronales Artificiales (ANN).

Sus principales ventajas son:

- Precisión en las Predicciones: Debido a que se dispone de etiquetas de referencia, los modelos supervisados tienden a ofrecer predicciones altamente precisas.
- Amplia Aplicación: Este enfoque se aplica ampliamente en tareas de clasificación y regresión en una variedad de campos, como la medicina, la industria financiera y reconocimiento de imágenes.

- Interpretación Intuitiva: Los resultados del aprendizaje supervisado son fáciles de interpretar, ya que el modelo se entrena para predecir etiquetas específicas.

Desventajas:

- Dependencia de Datos Etiquetados: El principal inconveniente es la necesidad de disponer de un conjunto de datos etiquetados de alta calidad, que a menudo pueden ser costosos y difíciles de obtener.
- Sobreajuste (Overfitting): Los modelos pueden ser propensos al sobreajuste, lo que significa que pueden adaptarse demasiado a los datos de entrenamiento y no generalizar bien a nuevos datos. Esto puede ocurrir si el modelo es muy complejo o si el conjunto de datos de entrenamiento es pequeño.

Por otro lado, el Aprendizaje No Supervisado se centra en la identificación de patrones y estructuras ocultas en un conjunto de datos sin la guía de etiquetas o valores de salida predefinidos. En este tipo de aprendizaje automático, el algoritmo analiza los datos y busca relaciones naturales entre ellos sin necesidad de información externa que indique cómo deben agruparse o categorizarse. El objetivo principal del Aprendizaje No Supervisado es descubrir la estructura subyacente en los datos sin etiquetas previas. Ejemplos de técnicas de aprendizaje no supervisado son Clustering (Agrupamiento), Reducción de Dimensionalidad o Análisis de Componentes Principales (PCA).

Ventajas del Aprendizaje No Supervisado:

- Descubrimiento de Patrones Ocultos: El aprendizaje no supervisado es excepcional para descubrir patrones y estructuras ocultas en datos no etiquetados, lo que puede revelar información valiosa.
- Exploración de Datos Complejos: Es útil cuando se trabaja con datos complejos y no estructurados, como imágenes, texto no etiquetado o datos de sensores.

Desventajas:

- Dificultad para Evaluar la Calidad de Resultados: Dado que no se dispone de etiquetas de referencia, la evaluación de la calidad de los resultados puede ser subjetiva y complicada.

- Interpretación Menos Directa: En comparación con el aprendizaje supervisado, los resultados del aprendizaje no supervisado pueden carecer de una interpretación directa, lo que requiere un análisis más profundo.

Este Trabajo de Fin de Máster se centra en la aplicación de técnicas de aprendizaje no supervisado. A continuación, se describen a nivel teórico los distintos algoritmos de agrupamiento que se aplican.

## 3.2 Técnicas de agrupamiento seleccionadas

Se han seleccionado tres técnicas de agrupamiento ampliamente reconocidas: K-Means, Hierarchical Clustering y DBSCAN, debido a sus enfoques diferentes y posibilidad de variabilidad en los resultados que pueden ser obtenidos.

### 3.2.1 K-Means

K-Means [11], conocido por su simplicidad y eficiencia, este algoritmo agrupa datos intentando separar las muestras en grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. El algoritmo requiere que se especifique el número de grupos a separar. Escala bien a un gran número de muestras y se ha utilizado en una amplia gama de áreas de aplicación en diferentes campos.

El algoritmo K-Means divide un conjunto de N muestras X en K clústeres disjuntos C, cada uno descrito por la media $\mu_j$ de las muestras en el clúster. Comúnmente, a estas medias se les llama "centroides" de los clústeres; ten en cuenta que, en general, no son puntos de X, aunque se encuentran en el mismo espacio.

$$\sum_{i=0}^{n} \min_{\mu_j \subseteq C} \left( \left| x_i - \mu_j \right|^2 \right)$$

La inercia puede considerarse como una medida de cuán coherentes son internamente los clústeres, pero trae algunas desventajas:

- La inercia asume que los clústeres son convexos e isotrópicos, lo que no siempre es el caso. Responde mal a clústeres alargados o a variedades con formas irregulares.

- La inercia no es una métrica normalizada: solo sabemos que valores más bajos son mejores y cero es óptimo. Pero en espacios de muy alta dimensión, las distancias euclidianas tienden a inflarse (esto es un ejemplo del llamado "problema de la maldición de la dimensionalidad").

En términos básicos, el algoritmo consta de tres pasos. El primer paso elige los centroides iniciales aleatoriamente. Después de la inicialización, K-Means consiste en repetir los dos pasos restantes en un bucle. El primer paso asigna cada muestra al centroide más cercano. El segundo paso crea nuevos centroides tomando el valor promedio de todas las muestras asignadas a cada centroide anterior. Se calcula la diferencia entre los centroides antiguos y los nuevos, y el algoritmo repite estos dos últimos pasos hasta que este valor sea inferior a un umbral. En otras palabras, se repite hasta que los centroides no se muevan significativamente.

Transcurrido tiempo suficiente, K-Means siempre convergerá, aunque esto puede llevarlo a un mínimo local. Esto depende en gran medida de la inicialización de los centroides. Para sofocar esto, a menudo se realiza el cálculo varias veces con diferentes inicializaciones de los centroides. Un método para abordar este problema es el esquema de inicialización K-Means++, utilizado en este Trabajo. Esto inicializa los centroides de manera que estén (en general) distantes entre sí, lo que probablemente conduzca a mejores resultados que la inicialización aleatoria [11].

### 3.2.2 Hierarchical Clustering

La elección de Hierarchical Clustering [12] complementa el abordaje de K-Means, ya que no requiere la especificación previa del número de clusters. En su lugar, construye una jerarquía de clusters que puede ser útil para comprender las relaciones entre grupos de clientes. Esta técnica proporciona flexibilidad al explorar la jerarquía en diferentes niveles de detalle.

El algoritmo construye grupos anidados en el conjunto de datos, obteniendo subconjuntos más pequeños. La división es en función de sus similitudes, hasta que las diferencias son cada vez más pequeñas. En última instancia, el último grupo dividido resulta ser un grupo individual.

A nivel técnico, el algoritmo comienza considerando cada punto de datos como un grupo individual y luego procede a combinar gradualmente los grupos más cercanos en grupos más grandes. Esto se hace mediante la aplicación de una medida de similitud, como la distancia euclidiana, entre los puntos de datos y la fusión de los dos grupos más cercanos en uno solo. Este

proceso se repite hasta que todos los puntos de datos estén en un solo grupo o hasta que se cumpla algún otro criterio predefinido, como un número máximo de grupos o una distancia máxima entre grupos.

Una de las características distintivas de Hierarchical Clustering es que genera una estructura de árbol llamada dendrograma, que muestra cómo se agrupan los datos a diferentes niveles de similitud. Esto permite una visualización intuitiva de la jerarquía de clusters y facilita la identificación de patrones en los datos.

Entre las ventajas de este algoritmo se incluye la capacidad de manejar datos con formas irregulares y la capacidad de proporcionar una visión detallada de la estructura de clusters. Sin embargo, puede ser computacionalmente costoso en conjuntos de datos grandes y puede ser difícil determinar el número óptimo de clusters en la jerarquía.

### 3.2.3 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [13] es una selección incorporada por su capacidad para detectar clusters (grupos) de formas irregulares y su robustez ante ruido y valores atípicos. Dado que el conjunto de datos podría contener comportamientos de compra inusuales o no homogéneos, DBSCAN es valioso para identificar grupos de clientes que no se ajustan a patrones convencionales. Además, no requiere definir previamente el número de clusters y puede identificar automáticamente la densidad de datos.

DBSCAN es un algoritmo de agrupación de datos que se basa en la densidad de los puntos en el espacio de variables/características del conjunto de datos. En lugar de requerir una predefinición del número de clústeres, DBSCAN identifica automáticamente los clústeres en función de cómo se agrupan los puntos en áreas densas y cómo se separan de áreas menos densas. Utiliza dos parámetros clave: la distancia máxima que un punto puede estar de otro para considerarse parte del mismo clúster (épsilon) y el número mínimo de puntos requeridos para formar un clúster (minPoints).

Para cada punto del conjunto de datos, se aplica un escaneo a su alrededor. El escaneo se hace según un radio de acción marcado por la variable establecida épsilon. Si en la zona escaneada, resulta haber el número establecido de minPoints, se genera un clúster. Mientras que el parámetro minPoints controla principalmente la tolerancia del algoritmo hacia el ruido (en conjuntos de datos

ruidosos y grandes, puede ser deseable aumentar este parámetro), el parámetro épsilon es crucial ya que controla cuántos puntos vecinos forman un clúster.

# 4 Materiales

## 4.1 Descripción de la base de datos

El conjunto de datos elegido proporciona una visión detallada de las transacciones de ventas en un establecimiento minorista de bienes de consumo. Estos datos se originan en el escaneo de los códigos de barras de productos individuales en el momento de la compra por parte de los clientes en el establecimiento.

La fuente de donde se han extraído el conjunto de datos proviene de kaggle.com [14]

Este conjunto de datos abarca un total de 64.682 transacciones de venta, lo que proporciona un volumen significativo de información para su análisis. En estas transacciones, se encuentran involucradas 5.242 SKU's (Stock Keeping Units), que representan una variedad diversa de productos en el inventario del minorista.

Un aspecto importante a destacar es que los datos han sido previamente sometidos a un proceso de anonimización, en el que los identificadores clave, como el SKU Category ID y el SKU ID, están codificados alfanuméricamente, lo que garantiza la confidencialidad y la privacidad de la información del producto.

Este conjunto de datos abarca un período de tiempo de un año, específicamente el año 2016. Además, se ha registrado un total de 22.625 clientes que realizaron compras en el establecimiento minorista.

A continuación, se ofrece una descripción de las variables:

Date (Fecha de la Transacción de Venta): Registra la fecha en que se realizó una transacción.

Customer_ID (ID del Cliente) identifica de manera única a cada cliente.

Transaction_ID (ID de la Transacción) proporciona un identificador único que representa una venta/transacción, por cliente por día, y puede contener uno o más productos diferentes, con sus respectivas cantidades.

SKU_Category (ID de la Categoría de SKU) codifica la categoría a la que pertenece cada producto (anonimizado).

(SKU) ID de SKU codifica de manera única cada producto (anonimizado).

(Quantity) Cantidad Vendida indica la cantidad del mismo producto vendido en cada transacción.

(Sales_Amount) Importe de Ventas este valor es el resultado del precio unitario multiplicado por la cantidad vendida. Los precios unitarios pueden estar registrados como precios netos sin impuestos, ya que no siguen el redondeo típico de precios de venta al público.

En el siguiente apartado se realiza un análisis de los valores de estas variables.

## 4.2 Análisis de variables

El conjunto de datos contiene 131.706 instancias con 8 variables. Por medio de código de Python utilizando la librería Pandas, se detecta que en el conjunto de datos no hay datos nulos ni faltantes. Se detecta que la primera variable, 'Unnamed: 0', resulta ser tan sólo la numeración de las instancias, por lo que se elimina.

Lo siguiente que se plantea es la eliminación de valores atípicos. Sin embargo, como se cuenta a continuación, no se considera que estos existan.

Se añade la variable 'Price_Unit', para obtener el precio unitario de cada producto, que resulta de dividir 'Sales_Amount' entre ' Quantity'. La motivación de añadir esta variable es estudiar valores atípicos en el precio unitario de un producto que pudiera quedar enmascarado en las variables 'Sales_Amount' y ' Quantity'. De esta forma, podría detectar algún valor atípico que pasara desapercibido de otra forma.

**Tabla 1.** Estructura de los datos.

| Date | Customer_ID | Transaction_ID | SKU_Category | SKU | Quantity | Sales_Amount |
|---|---|---|---|---|---|---|
| 02/01/2016 | 2547 | 1 | X52 | 0EM7L | 1.0 | 3.13 |
| 02/01/2016 | 822 | 2 | 2ML | 68BRQ | 1.0 | 5.46 |
| 02/01/2016 | 3686 | 3 | 0H2 | CZUZX | 1.0 | 6.35 |
| 02/01/2016 | 3719 | 4 | 0H2 | 549KK | 1.0 | 5.59 |
| 02/01/2016 | 9200 | 5 | 0H2 | K8EHH | 1.0 | 6.88 |

En la tabla 1 se refleja la estructura de los datos. En la tabla 2 se realiza un estudio de la estadística descriptiva del conjunto después de añadir la variable 'Price_Unit:

**Tabla 2.** Estadística descriptiva del conjunto de datos.

| | Customer_ID | Transaction_ID | Quantity | Sales_Amount | Price_Unit |
|---|---|---|---|---|---|
| count | 131706.000000 | 131706.000000 | 131706.000000 | 131706.000000 | 131706.000000 |
| mean | 12386.450367 | 32389.604187 | 1.485311 | 11.981524 | 9.692429 |
| std | 6086.447552 | 18709.901238 | 3.872667 | 19.359699 | 14.944890 |
| min | 1.000000 | 1.000000 | 0.010000 | 0.020000 | 0.015000 |
| 25% | 7349.000000 | 16134.000000 | 1.000000 | 4.230000 | 3.680000 |
| 50% | 13496.000000 | 32620.000000 | 1.000000 | 6.920000 | 6.070000 |
| 75% | 17306.000000 | 48548.000000 | 1.000000 | 12.330000 | 10.100000 |
| max | 22625.000000 | 64682.000000 | 400.000000 | 707.730000 | 693.800000 |

Se realiza una revisión en busca de valores atípicos en el conjunto de datos, pero, aunque se han encontrado valores que podrían considerarse atípicos, estos no deben ser necesariamente eliminados. Los valores atípicos en un conjunto de datos pueden ser puntos de datos que se desvían significativamente del comportamiento general de los demás, pero eso no implica que sean datos erróneos o inadecuados para el análisis. En este caso, los valores atípicos pueden representar

comportamientos de compra únicos o transacciones excepcionales que son relevantes para comprender la diversidad de la base de clientes, pudiendo aportar información valiosa sobre segmentos de clientes únicos, por ello se decide contar con ellos.

# 5 Experimentación

## 5.1 Preprocesamiento de datos

Una vez se realiza el entendimiento del conjunto de datos, se analiza su estadística descriptiva, se verifica que el conjunto no tiene valores faltantes ni nulos, y se determina que la eliminación de valores atípicos aplica en este caso, se procede a trabajar con el conjunto de datos hacia la aplicación de los algoritmos mencionados.

A continuación, se prepara un nuevo conjunto de datos a partir del original, donde se calculan las tres métricas fundamentales en un análisis de Recencia, Frecuencia, Inversión (RFM), con el objetivo de proporcionar una representación más precisa y detallada del comportamiento de cada cliente en términos de sus hábitos de compra. Estas métricas son:

- Recency ("Recencia"): Tiempo transcurrido desde la última compra. Tiene como resultado el número de días transcurridos desde la última compra del cliente en el establecimiento, es decir, la actualidad de su última compra. Se calcula haciendo la diferencia de días entre 01 enero de 2017 con la fecha que marca la variable 'Date".
- Frequency (Frecuencia): Cantidad de veces que el cliente ha comprado, en este caso, durante el año de estudio. Se obtiene haciendo un conteo del total de las transacciones de cada cliente durante el año de estudio.
- Monetary (Inversión): Cantidad total de dinero invertido en compras por el cliente durante el año de estudio. Calculada a partir de la suma del valor monetario de todas las transacciones que ha realizado cada cliente.

El tratamiento destaca por su simplicidad, pero es sumamente importante las conclusiones a las que se puede llegar segmentando estos datos.

El conjunto de datos resultante del análisis de RFM es el que se muestra en la tabla 3. Como se puede observar, cada cliente tiene ahora tan solo las tres variables descritas. Estas instancias de

puntos de tres variables se representan posteriormente en visualizaciones tridimensionales, con los grupos de clientes segmentados obtenidos.

**Tabla 3.** Conjunto de datos RFM

| Customer_ID | Inversion | Frecuencia | Recencia |
|---|---|---|---|
| 0 | 1 | 16.29 | 2 |
| 1 | 2 | 22.77 | 2 |
| 2 | 3 | 10.92 | 3 |
| 3 | 4 | 33.29 | 5 |
| 4 | 5 | 78.82 | 5 |
| ... | ... | ... | ... |
| 22620 | 22621 | 9.69 | 2 |
| 22621 | 22622 | 6.07 | 1 |
| 22622 | 22623 | 128.01 | 2 |
| 22623 | 22624 | 19.60 | 2 |
| 22624 | 22625 | 83.62 | 9 |

En la siguiente figura, Figura 1, se muestra la distribución las variables obtenidas:

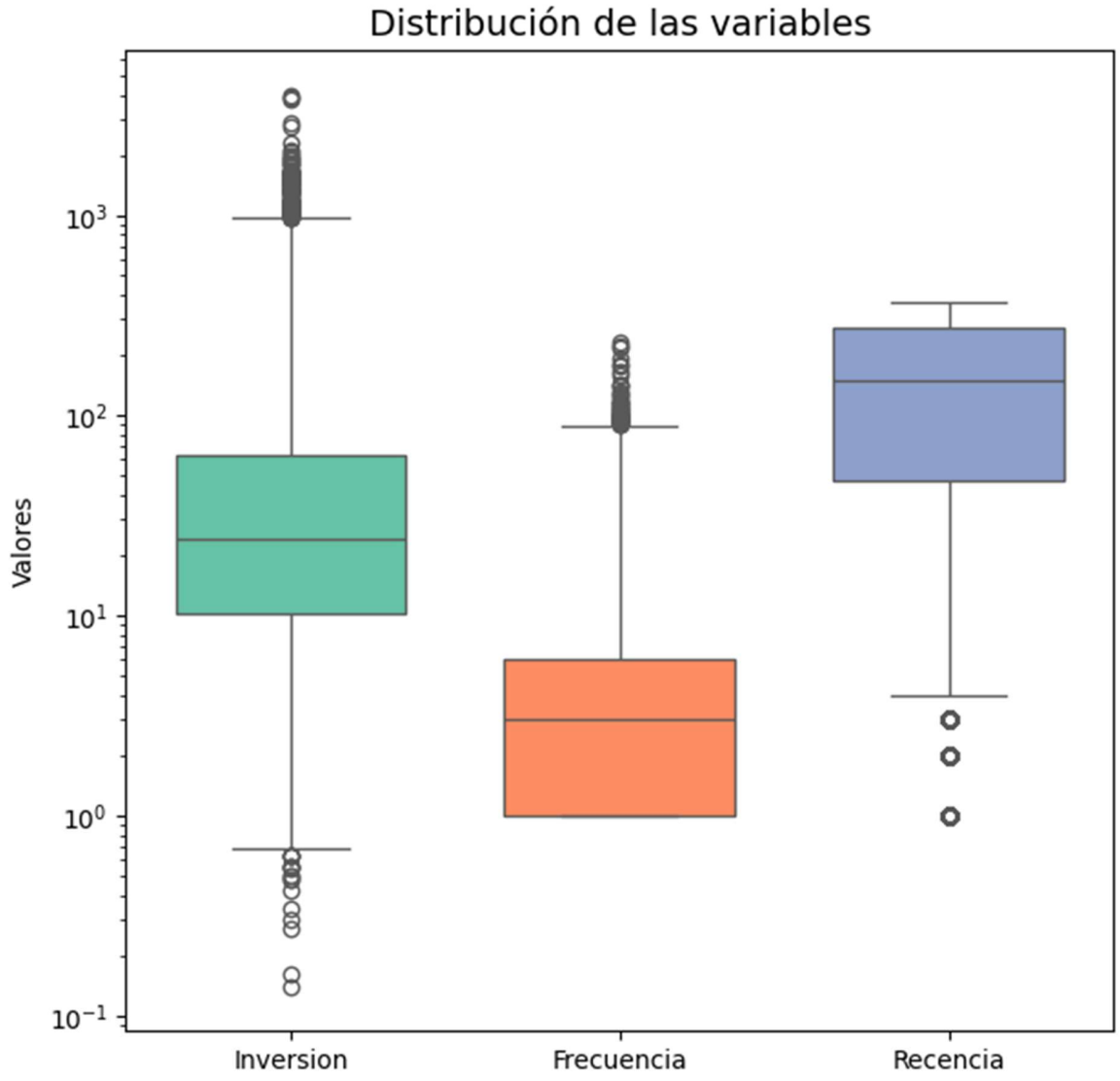


**Figura 1.** Distribución de las variables Inversión, Frecuencia, Recencia obtenidas.

Después del análisis RFM y se procede a la normalización/estandarización de variables, para asegurar que no hay impacto en el desempeño de los algoritmos debido a diferencias de escalas o magnitudes. La estandarización de datos que se aplica consiste en hacer la media cero y la desviación estándar uno. Para cada dato de cada variable, se recalcula según:

$$z = (x - \mu) / s$$

En la ecuación anterior, μ es la media aritmética y *s* corresponde con la desviación estándar del conjunto de datos de la variable calculada.

El siguiente paso consiste en aplicar cada algoritmo escogido, pues el conjunto de datos resultante es apto para ser tratado.

En K-Means primero se considera cuántos grupos deben ser fijados para que el algoritmo haga converger los centroides. Como se ha mencionado previamente, se escoge la inicialización 'k-means++', que inicializa los centroides de manera se distancien entre sí, y que probablemente asegure obtener mejores resultados que con inicialización aleatoria. Con ayuda de las medidas resultantes de aplicar la inercia y el Silhouette Score, evaluadas para hasta 8 grupos, se concluye que la mejor segmentación es la de 4 grupos.

Cuando se aplica Hierarchical Clustering, se aplica el dendrograma de la librería sklearn [14], en el que se aplica una visualización personalizada para un mejor entendimiento para el lector. Se utiliza el método 'Ward', que fusiona clústeres de manera que se minimice la varianza dentro de los clústeres resultantes, siendo estos lo mas homogéneos posible entre sí en términos de similitud de los datos del clúster. Por último, se establece un umbral de distancia euclidea en la que la segmentación queda hecha para 3 clústeres lo suficientemente diferentes entre sí.

Al aplicar DBSCAN, se prueban numerosas combinaciones de épsilon y minPoints, pero los puntos están poblando de forma muy densa la frecuencia y la inversión en sus valores más bajos y, lamentablemente es la única zona densa del conjunto de datos. Por lo que este algoritmo da los resultados esperados, pero no tienen demasiada aplicación ni desvelan nada novedoso en los datos que no pudiese verse antes de aplicar el algoritmo.

La fase final consiste en la conclusión los resultados obtenidos y se destacan las implicaciones prácticas de los hallazgos en el contexto de ventas minoristas.

## 5.2 Resultados

Dados todos estos pasos previos, se procede a aplicar los algoritmos de Clustering seleccionados:

- K-Means

Al aplicar K-Means, se obtienen 4 grupos de clientes diferentes. Para determinar el número de grupos (clusters), se emplean dos métricas clave: la inercia (distorsión) y el Silhouette Score.

La inercia se utiliza para evaluar cuán compactos son los clusters, y el Silhouette Score mide la cohesión y la separación entre clusters. Al ejecutar K-Means con una variedad de valores de 'k', se calcula la inercia y el Silhouette Score para cada caso.

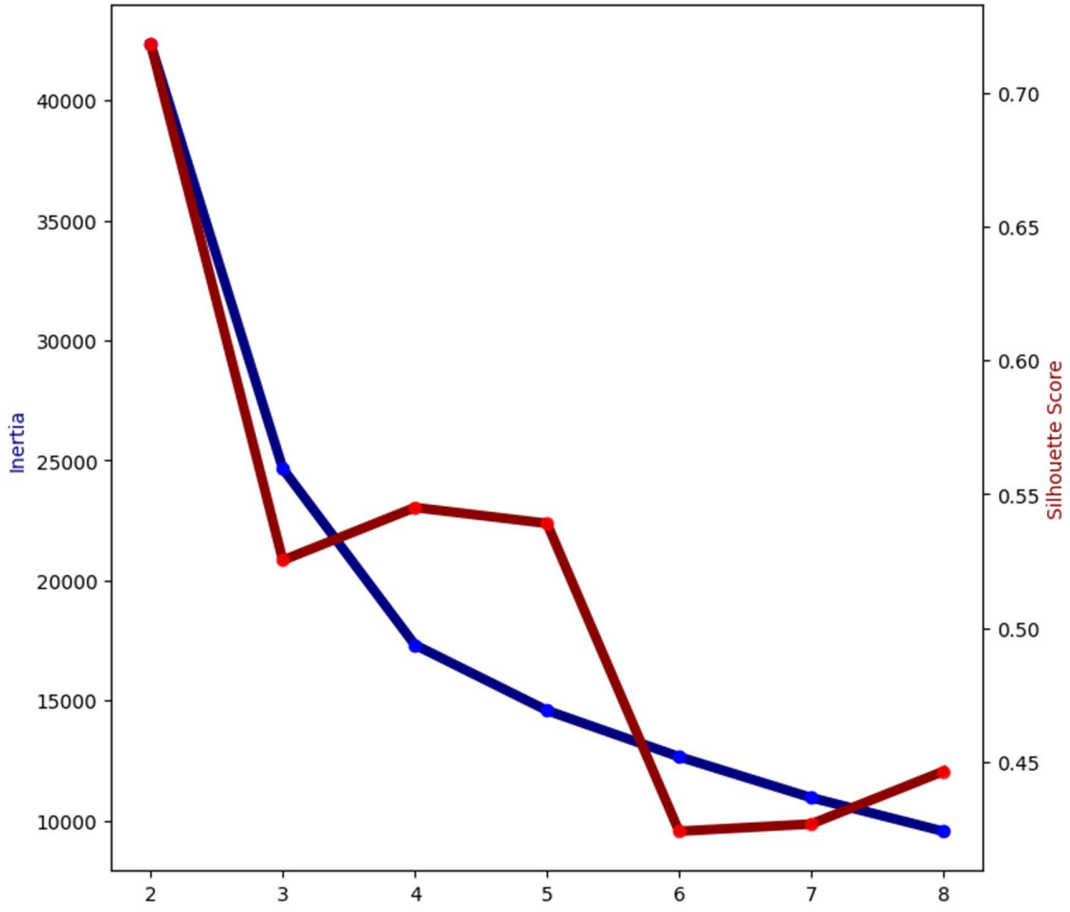


**Figura 2.** Resultados del cálculo de la inercia y el Silhouette Score para valores de k comprendidos entre 2 y 8.

Después de un análisis detallado de ambas métricas, se observa que el valor de 'k' que maximiza el Silhouette Score y, simultáneamente, muestra un punto de inflexión más pronunciado en la curva de la inercia es igual a 4. Establecidos los grupos, en la tabla 4 se muestra la cantidad de clientes comprendidos en cada uno.

**Tabla 4.** Grupos obtenidos aplicando K-Means y cantidad de clientes por grupo.

| Cluster | Count |
|---|---|
| 0 | 9903 |
| 1 | 10205 |
| 2 | 255 |
| 3 | 2262 |

En siguientes figuras (Figura 2-4) se representa la distribución de las variables por grupos:

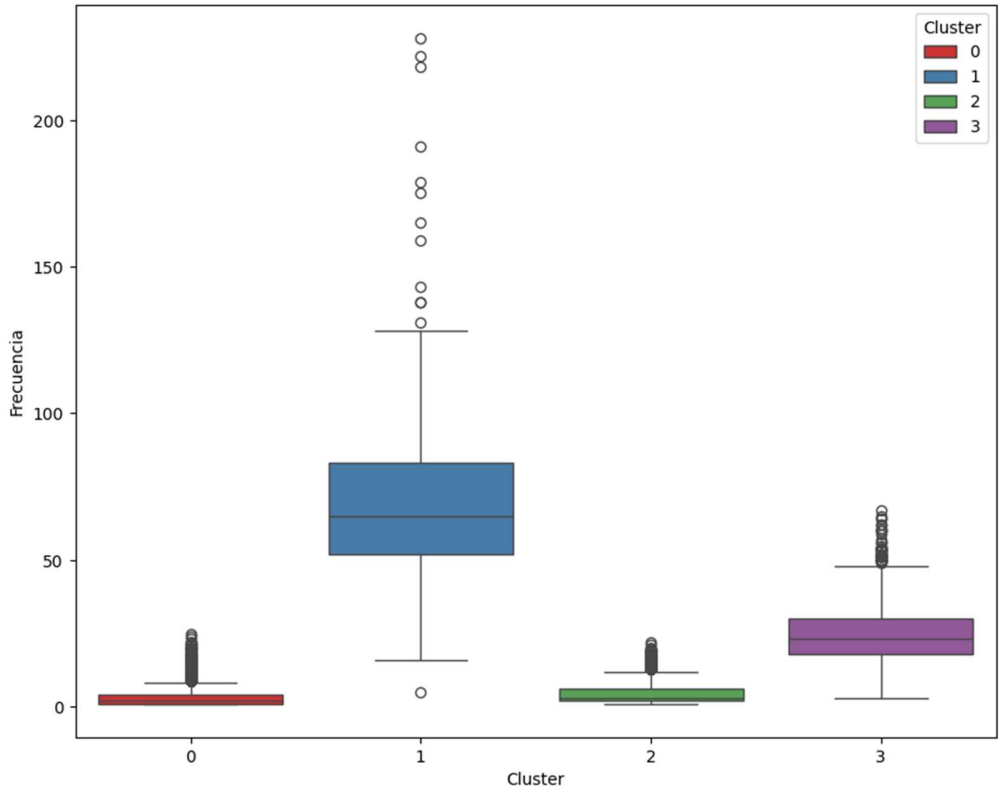


**Figura 3.** Distribución de la variable Frecuencia en los grupos de clientes obtenidos (K-Means).

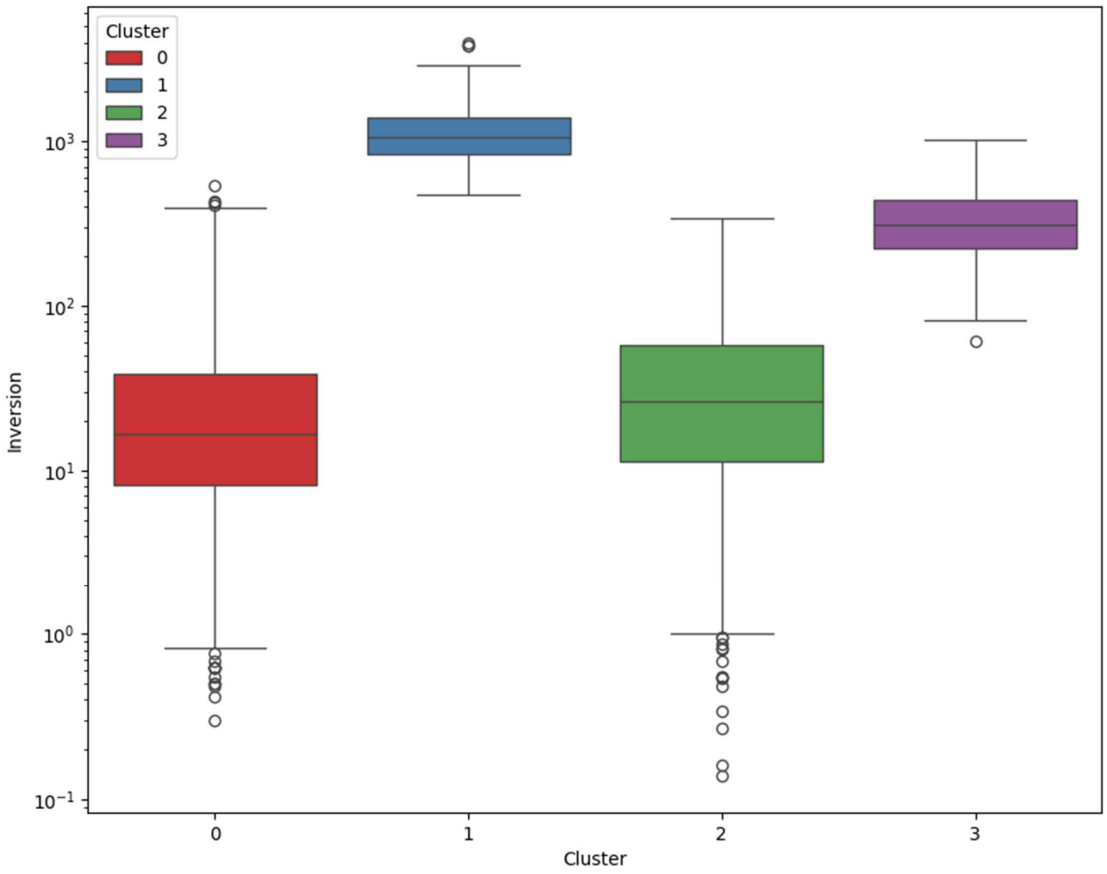


**Figura 4.** Distribución de la variable Inversión en los grupos de clientes obtenidos (K-Means).

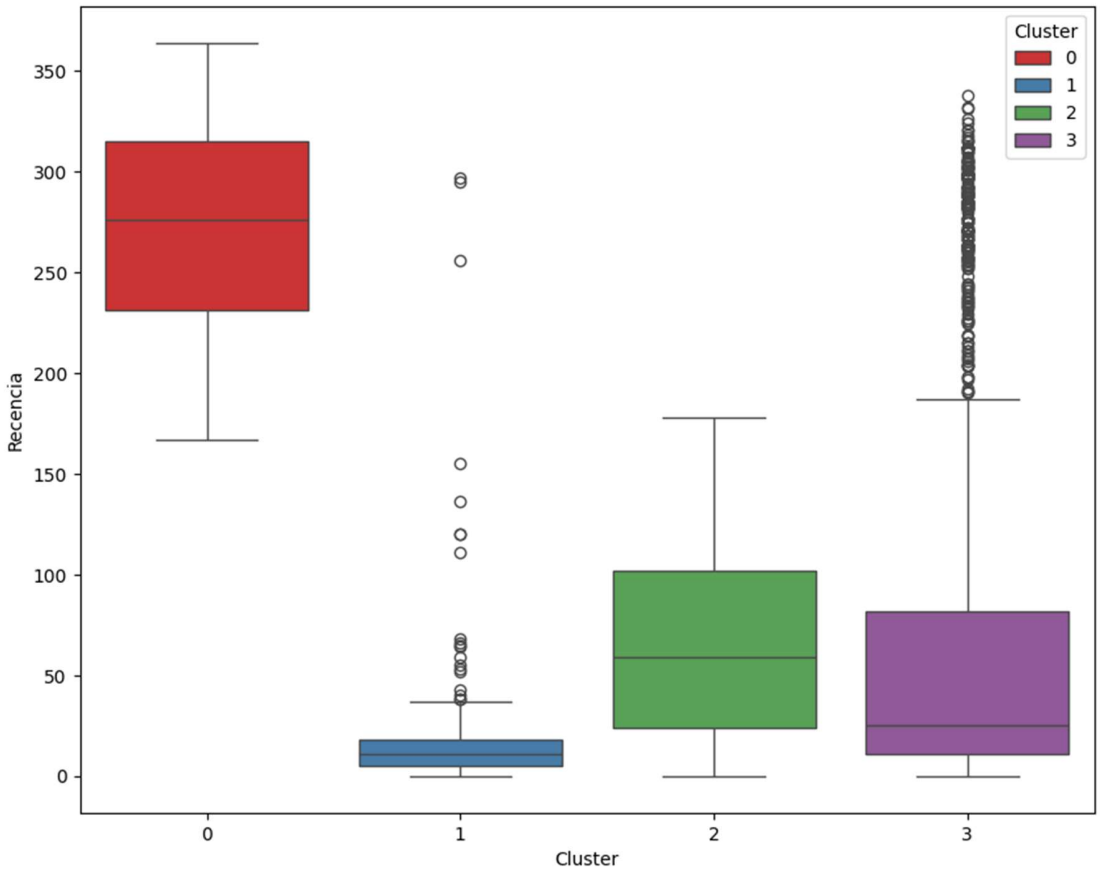


**Figura 5.** Distribución de la variable Recencia en los grupos de clientes obtenidos (K-Means).

Por último se muestran los clusters en una representación tridimensional para dar mejor idea de cómo han quedado segmentados los clientes en este caso:

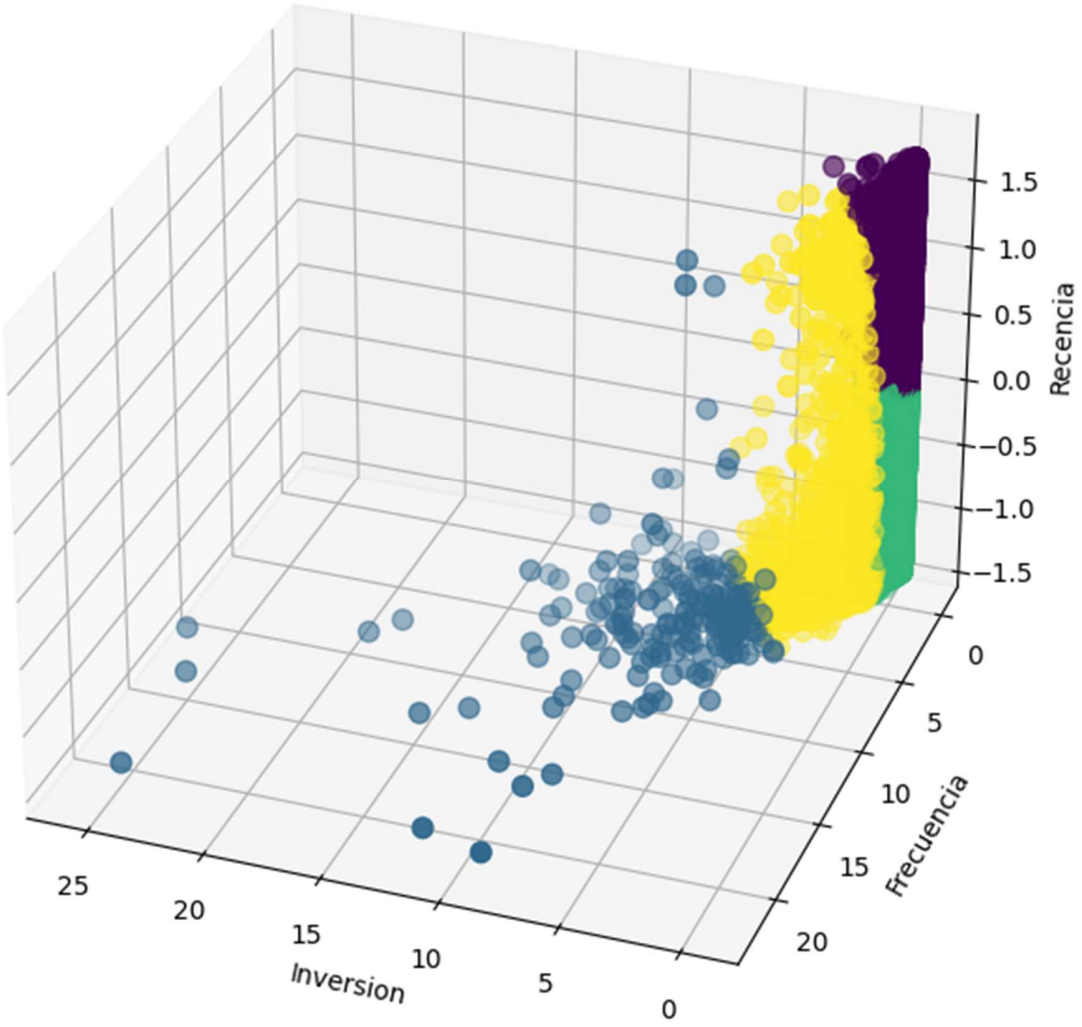


**Figura 6.** Representación tridimensional de los grupos de clientes obtenidos mediante K-Means.

- Hierarchical Clustering

Al aplicar Hierarchical Clustering en el conjunto de datos, se observa que este algoritmo segmenta en 3 grupos en lugar de 4, como se muestra en el dendograma de la figura 7.

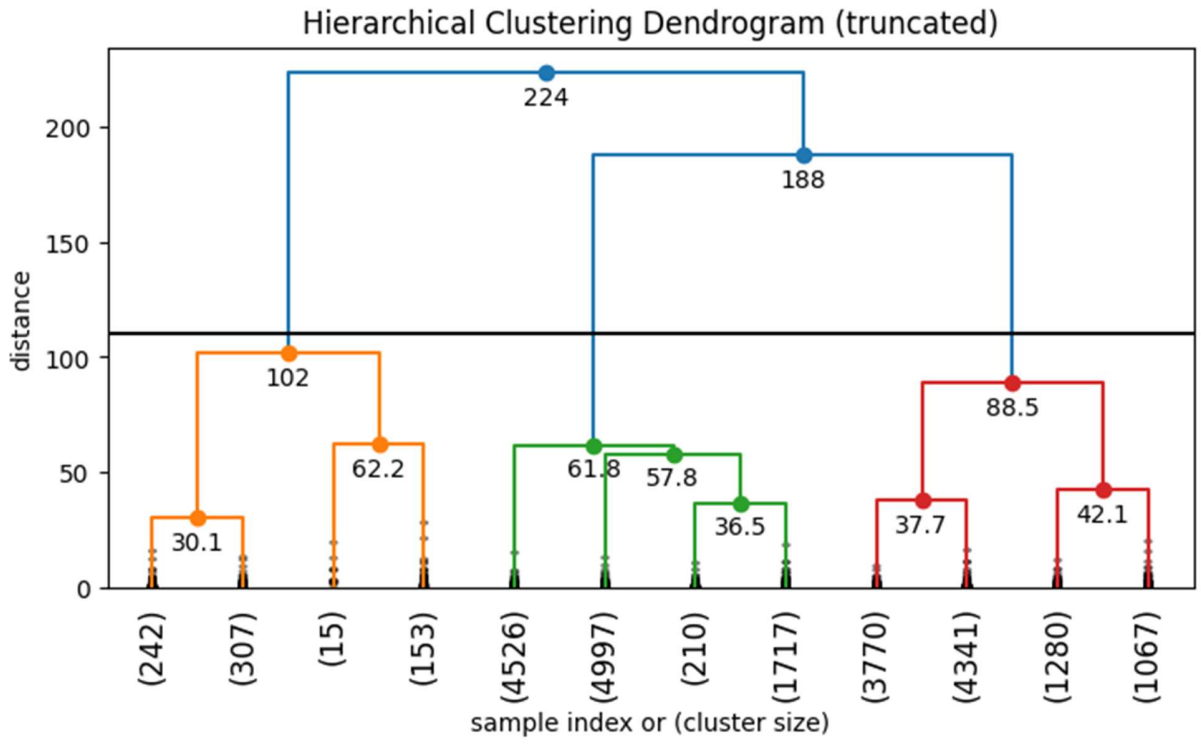


**Figura 7.** Dendograma obtenido al aplicar Hierarchical Clustering.

El dendrograma se presenta como una representación gráfica de la jerarquía de clusters y las relaciones de similitud entre los puntos de datos.

Para seleccionar el número adecuado de clusters, se realiza un análisis detenido del dendrograma. La clave radica en la observación de las divisiones verticales en el dendrograma, que indican la formación de nuevos clusters a medida que se desciende en la jerarquía. La altura en la que se corta el dendrograma determina el número de clusters.

Al examinar el dendrograma, se busca el punto donde un corte horizontal revele una estructura de clusters clara y distintiva. Normalmente, esto implica identificar una altura específica en el dendrograma donde los clusters se vuelven significativamente diferentes entre sí. Este punto se convierte en el número de clusters seleccionado.

En este caso, se determina que un corte en una altura específica genera 3 clusters claramente definidos y coherentes en términos de similitud entre los clientes.

En la tabla 5 se muestran la cantidad de clientes contenidos en cada grupo.

**Tabla 5.** Grupos obtenidos aplicando K-Means y cantidad de clientes por grupo.

| Cluster | Count |
|---|---|
| 1 | 717 |
| 2 | 11450 |
| 3 | 10458 |

En siguientes figuras (Figura 8-10) se representa la distribución de las variables por grupos:

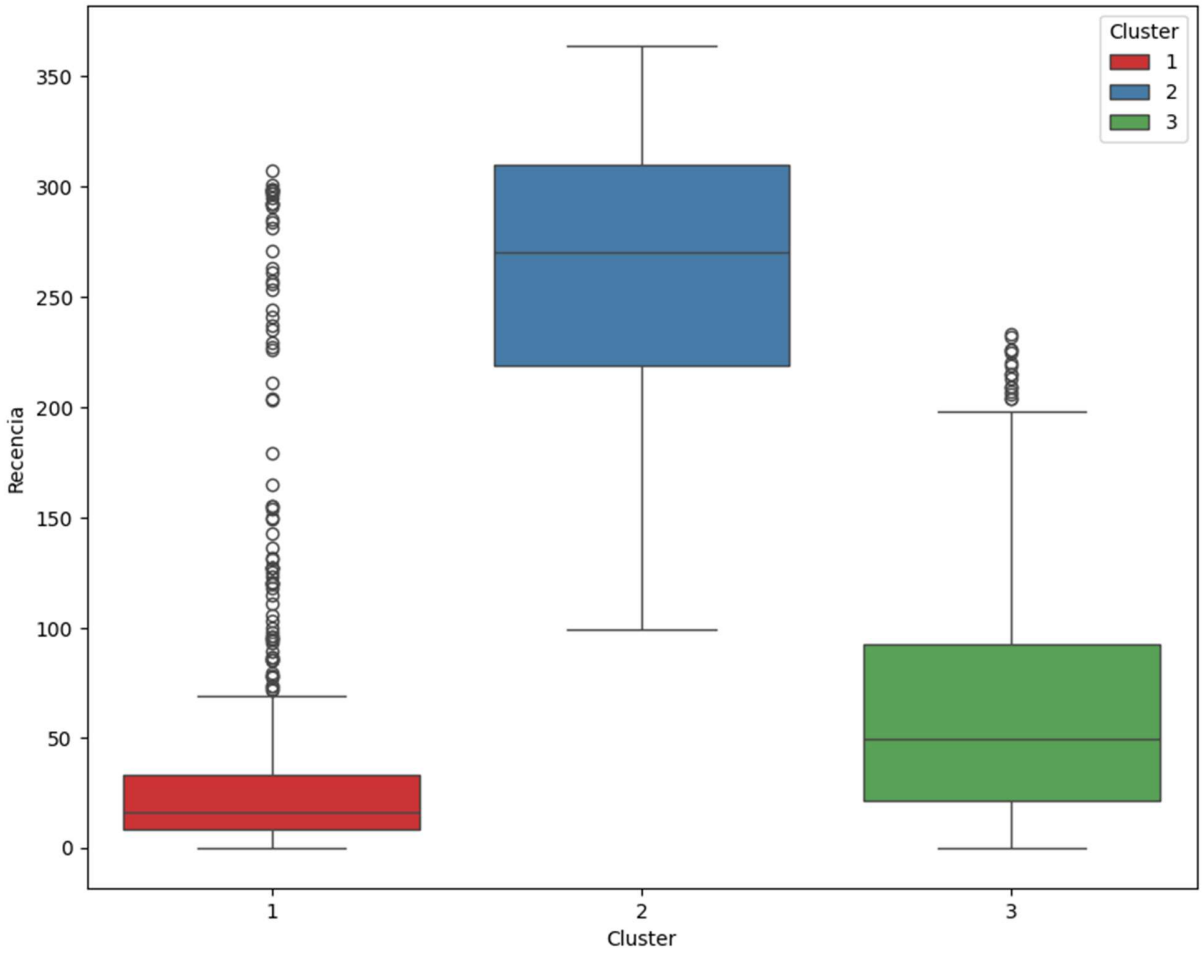


**Figura 8.** Distribución de la variable Recencia en los grupos de clientes obtenidos (Hierarchical Clustering).

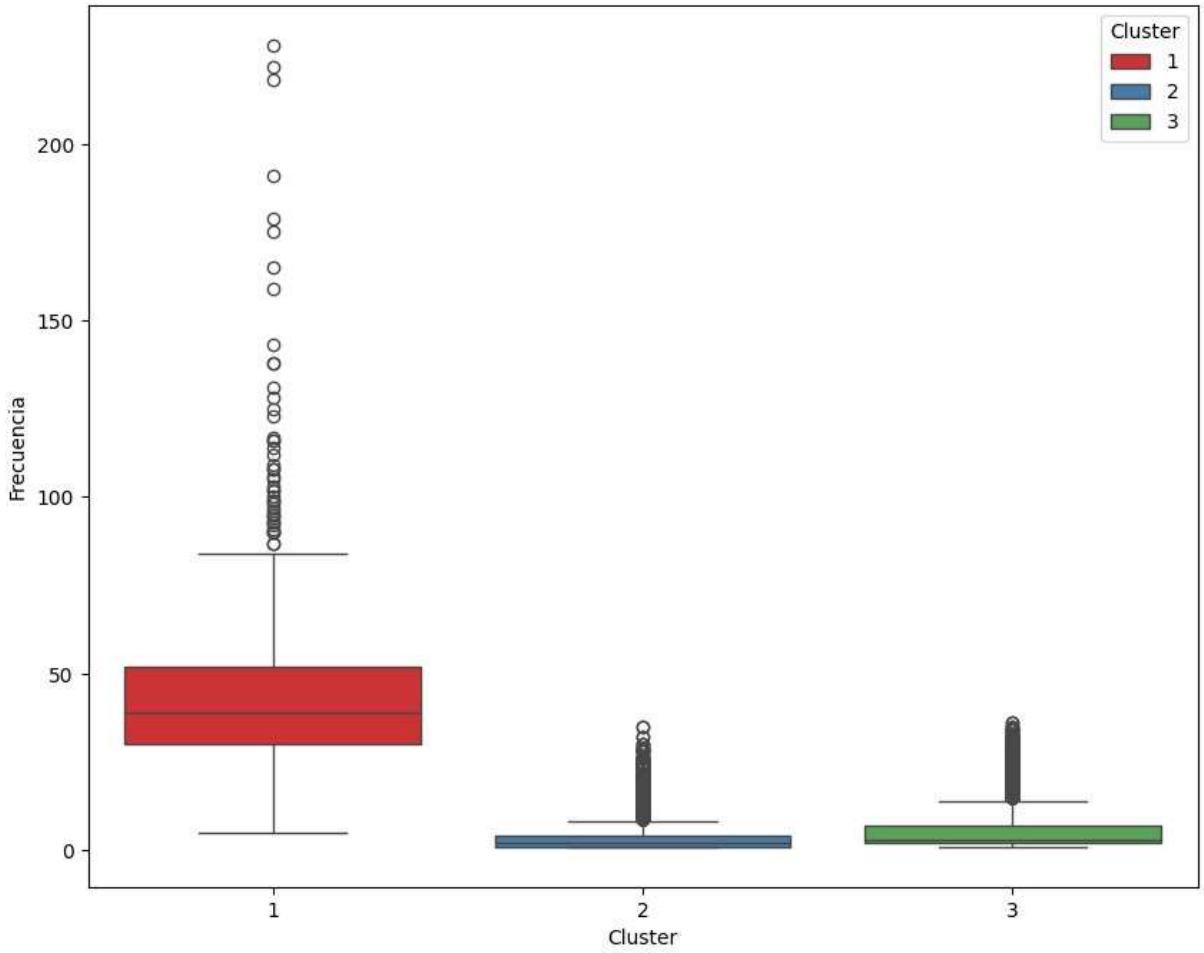


**Figura 9.** Distribución de la variable Frecuencia en los grupos de clientes obtenidos (Hierarchical Clustering).

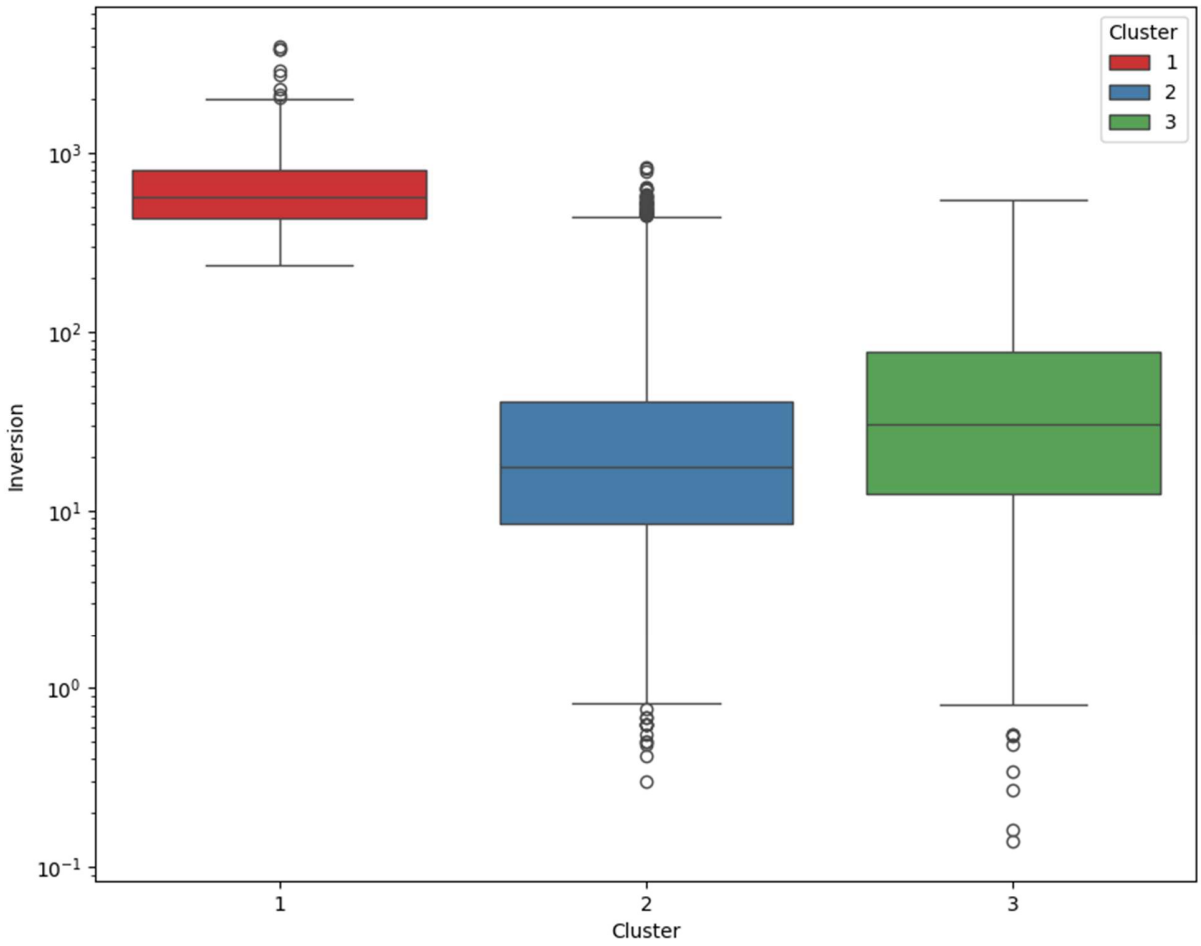


**Figura 10.** Distribución de la variable Inversión en los grupos de clientes obtenidos (Hierarchical Clustering).

Finalmente, se presentan los clusters en una representación tridimensional para proporcionar una mejor comprensión de cómo se ha realizado la segmentación de los clientes en este caso:

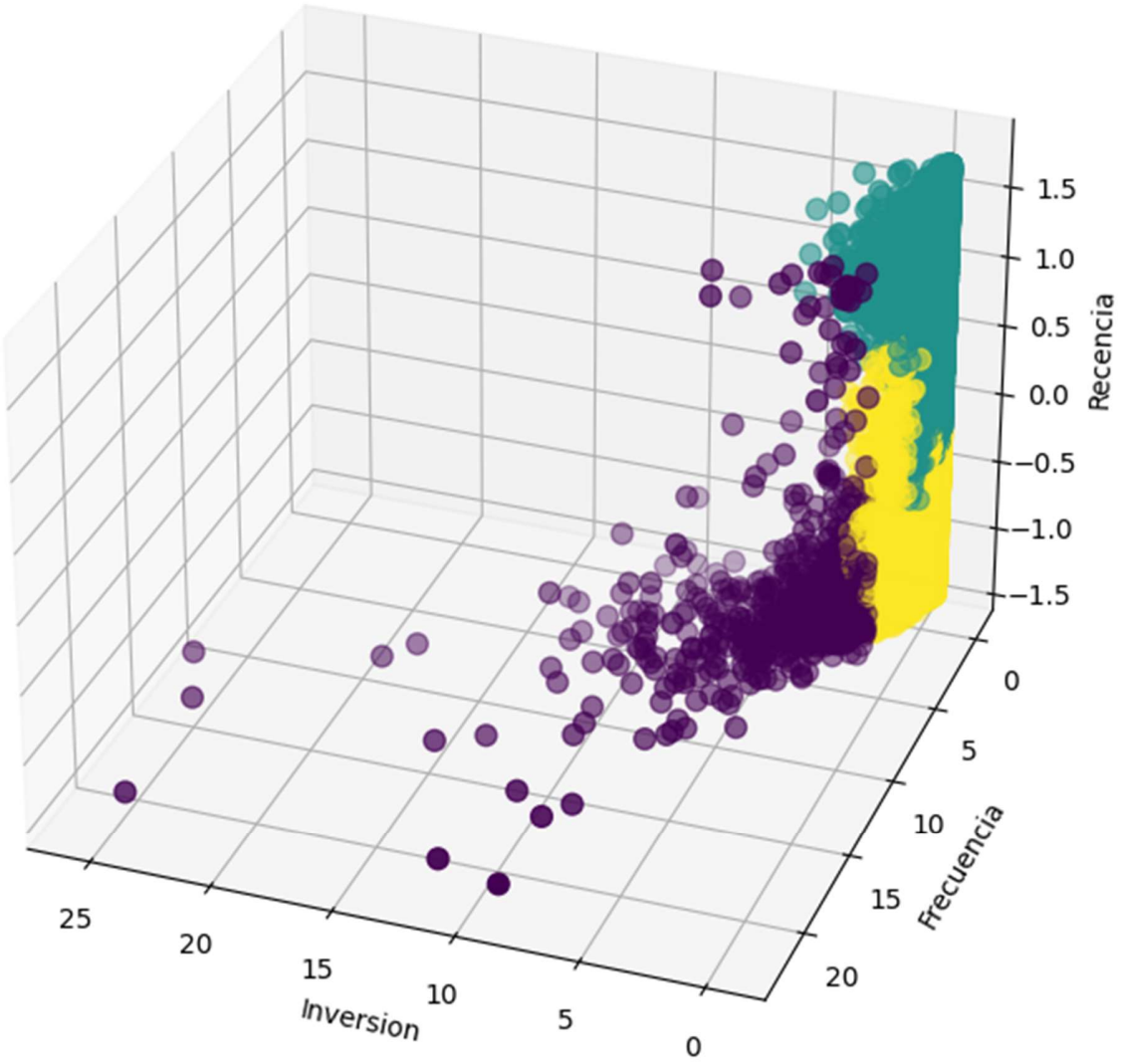


**Figura 11.** Representación tridimensional de los grupos de clientes obtenidos mediante Hierarchical Clustering.

- DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) es un algoritmo que se basa en la densidad de puntos en el espacio de características. Su enfoque consiste en identificar regiones densas de puntos que están separadas por regiones más dispersas.

Se han explorado diversas combinaciones de parámetros. No obstante, debido a la alta densidad de las variables de recencia y frecuencia en el conjunto de datos, se observa que DBSCAN no logra separar a los clientes de manera efectiva en grupos significativos, como se aprecia en la Figura 12.

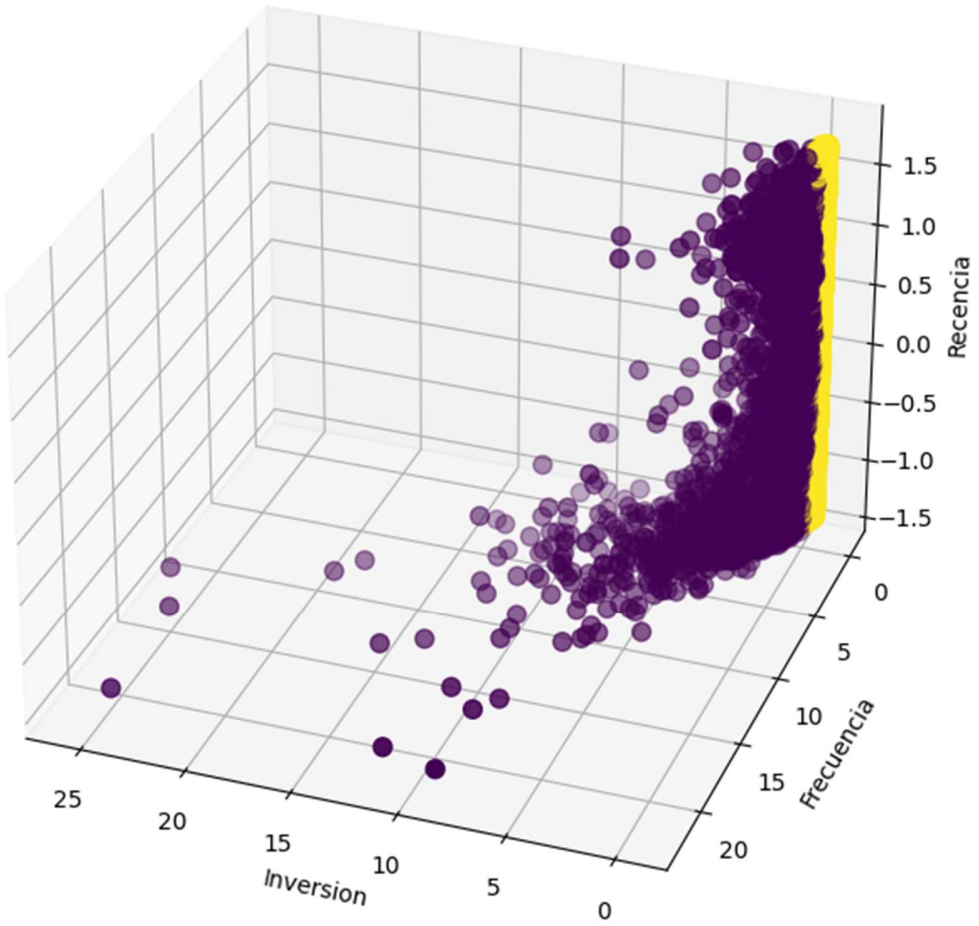


**Figura 12,** Representación tridimensional de los grupos de clientes obtenidos mediante DBSCAN.

**Tabla 6.** Grupos obtenidos aplicando DBSCAN y cantidad de clientes por grupo

| Cluster | Count |
|---|---|
| 0 | 18392 |
| -1 | 4233 |

Estos resultados son coherentes y previsibles, ya que DBSCAN suele enfrentar dificultades en conjuntos de datos con densidades variables o cuando las densidades son uniformes. En este caso particular, la distribución específica de los datos, con un alto grado de solapamiento en las variables de recencia y frecuencia, contribuye a la limitación en la capacidad de DBSCAN para realizar una segmentación precisa de los clientes.

## 5.3 Análisis de resultados

En el proceso de segmentación de clientes utilizando K-Means, se obtienen 4 clusters distintos.

Se observa un conjunto de clientes muy denso que compran muy frecuentemente y con pequeñas inversiones, este conjunto K-Means lo divide en 2 grupos atendiendo a su recencia, que es lo que los diferencia. Esta diferenciación tiene valor, pues a estos clientes que compran con pequeños desembolsos pero frecuentemente y lo han hecho hace poco (clúster 1) son los más importantes para el establecimiento, son clientes leales. Los clientes del cluster 3 están en la misma situación pero hace más tiempo que visitaron por última vez, por lo que captarlos de nuevo puede ser muy interesante si aumentamos esa 'lealtad' al establecimiento.

Los clientes del clúster 0, son parecidos a los que se acaban de mencionar, pero están algo más dispersos en cuanto a frecuencia e inversión, y no es un grupo tan denso, por lo que realizar una campaña única a todo el conjunto puede ser complicado e ineficiente, aunque su mayor fuerte es que compran frecuentemente. Por ello, no hay que desestimarlo, puesto que se tratan de clientes potenciales.

El último clúster, clúster 3, se identifica un grupo de clientes inactivo. Hace mucho desde su última visita y sus frecuencias e inversiones son muy dispares. Por tanto, si se lanza alguna estrategia para captarlos, tiene que ser adaptada a ellos y diferente de la de los otros grupos de clientes.

En el caso de Hierarchical Clustering, se obtienen 3 clusters de clientes segmentados:

Se observa también dos grupos aglomerados en baja frecuencia y baja inversión a lo largo de la recencia, clústeres 2 y 3, pero realmente están muy poblados, y más que los grupos obtenidos en K-Means. Y el Hierarchical Clustering va un poco más allá y logra agrupar a clientes con una recencia y frecuencia más variada en estos dos clústeres. Esta aproximación puede ser interesante ya que añade algo de variabilidad a los clientes segmentados, haciendo que la estrategia de marketing hacia estos clientes, incluya levemente a clientes que en K-Means fueron clasificados de forma diferente. Estas decisiones tienen que ser tomadas por expertos en la materia de marketing y publicidad, pero es interesante ver como la segmentación ha sido ligeramente diferente. El último clúster, clúster 1, si es más similar al clúster 3 de K-Means. Hierarchical Clustering también separa

a un grupo de clientes menos denso y con inversiones más dispares. Sobre todo estos clientes hace mucho que no visitan el establecimiento.

En cuanto a DBSCAN, como se mencionó anteriormente, no se obtienen resultados significativos debido a la alta densidad de las variables de recencia y frecuencia en el conjunto de datos. El algoritmo es incapaz de arrojar luz en los datos, ya que se obtiene un gran grupo muy denso o un grandísimo número de grupos de pocos clientes en cada uno, lo que hace inviable lanzar campañas de captación tan situacionales.

# 6 Conclusiones

La habilidad para segmentar a los clientes en grupos que comparten patrones de comportamiento similares tiene un alto valor para empresas para adaptar sus estrategias de marketing y retención con mayor precisión y eficacia. Al tener una visión sobre las necesidades, preferencias y tendencias de cada grupo, en datos que antes no mostraban tanta información, las organizaciones pueden personalizar sus ofertas, promociones y servicios de manera más eficaz para satisfacer las expectativas de sus clientes, y ser más competitivos.

Es crucial el trabajo en conjunto del analista de datos con los expertos en el sector del comercio y de captación de clientes, que son los que pueden finalmente decidir qué línea de resultados seguir. Habrá dependencias de la capacidad de realizar ofertas específicas y pertinentes a cada grupo, pero otro punto a favor de las técnicas vistas, es que hay cabida para el afinamiento y desde luego la decisión final de los clientes segmentados la tienen los expertos. Pero la aplicación de estas técnicas es un muy buen punto de partida si no bien la base fundamental en la que afinar la estrategia que se decida.

# 7 Referencias

[1] University of Cambridge, «What is the value of data? A review of empirical methods,» [En línea]. Available: https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/07/policy-brief_what-is-the-value-of-data.pdf.

[2] D. M. a. A. Heinzl, «Towards Machine Learning as an Enabler of Computational Creativity,» *EEE Transactions on Artificial Intelligence,* vol. 2, nº 6, pp. 460-475, 2021.

[3] H. L. H. S. W. Yaya Heryadi, «Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, Stacked LSTM, and CNN-LSTM,» *IEEE,* pp. 84-89, 2017.

[4] Ş. Ozan, «A Case Study on Customer Segmentation by using Machine Learning Methods,» de *International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018.

[5] N. Sharma, «analytics vidhya,» 14 Septiembre 2023. [En línea]. Available: https://www.analyticsvidhya.com/blog/2023/03/machine-learning-for-marketing/.

[6] V. A. Brei, «Machine Learning in Marketing: Overview, Learning Strategies, Applications, and Future Developments,» *Foundations and Trends® in Marketing,* vol. 14, nº 3, pp. 173-236, 2020.

[7] N. P. e. al, «Customer Segmentation Using Machine Learning,» *Recent Trends in Intensive Computing,* nº DOI:10.3233/APC210200, pp. 239-244, 2021.

[8] A. Abdulhafedh, «Incorporating K-means, Hierarchical Clustering,» *Journal of City and Development,* vol. 3, nº 1, pp. 12-30, 2021.

[9] D. &. M. J. Zakrzewska, «Clustering Algorithms for Bank Customer Segmentation,» de *Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 2005.

[10] N. T. T. L. a. N. D. N. Phan Duy Hung, «Customer Segmentation Using,» de *In Proceedings of the 2nd International Conference on Information Science and Systems (ICISS '19). Association for Computing Machinery*, New York, 2019.

[11] scikit-learn.org, «scikit-learn Machine Learning in Python - KMeans,» [En línea]. Available: https://scikit-learn.org/stable/modules/clustering.html#k-means.

[12] scikit-learn.org, «scikit-learn Machine Learning in Python - HC,» [En línea]. Available: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering.

[13] scikit-learn.org, «scikit-learn Machine Learning in Python - DBSCAN,» [En línea]. Available: https://scikit-learn.org/stable/modules/clustering.html#dbscan.

[14] kaggle.com, «Kaggle,» [En línea]. Available: https://www.kaggle.com/datasets/marian447/retail-store-sales-transactions/data.

[15] scikit-learn.org, «scikit-learn.org - AD,» [En línea]. Available: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html.