

**UNIVERSIDAD NACIONAL AGRARIA  
LA MOLINA  
FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**



**"SEGMENTACIÓN RFM DE CLIENTES DE UN PROGRAMA DE FIDELIZACIÓN EN UNA EMPRESA RETAIL CON EL ALGORITMO DE ANÁLISIS CLÚSTER PAM"**

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO  
DE  
INGENIERO ESTADÍSTICO INFORMÁTICO**

**ALONSO ARMANDO DIEZ RIOS**

**LIMA - PERÚ  
2022**

---

**La UNALM es titular de los derechos patrimoniales de la presente investigación  
(Art. 24 - Reglamento de Propiedad Intelectual)**

## Document Information

---

Analyzed document	TSP_ALONSO_DIEZ RIOS_turnititng.pdf (D143841328)
Submitted	2022-09-09 19:08:00
Submitted by	ivan soto rodriguez
Submitter email	ivans@lamolina.edu.pe
Similarity	0%
Analysis address	ivans.unalm@analysis.arkund.com

## Sources included in the report

---

<b>SA</b>	<b>Clustering.pdf</b> Document Clustering.pdf (D90069832)	 1
-----------	--	---

---

## Entire Document

---

RESUMEN Hoy en día las empresas se enfrentan a un gran desafío y es que cada día, estas generan mayor cantidad de datos y el reto está en aprovecharlos, transformándolos en información valiosa que generen insights relevantes para la toma de decisiones y el alineamiento de la estrategia. Una de las técnicas más utilizadas han sido los análisis de agrupamiento, teniendo injerencia en diversos campos como la biología, ingeniería, psicología, entre otros, pero donde habría resultado de mayor utilidad es en el campo de los negocios, específicamente en la investigación de mercados, abarcando temas de segmentación de clientes, identificación de hábitos de compra entre otros. En este marco, la empresa de retail de la que se habla en la presente monografía, se encuentra bajo un enfoque centrado en el cliente, por lo cual busca un mayor entendimiento del mismo, a través del aprovechamiento de los datos que dispone. Es así como el área de Inteligencia de Negocios interviene como encargado del tratamiento y análisis de los datos, proponiendo una segmentación RFM a través de un análisis clúster, que permitió agrupar a un conjunto de clientes objetivo para la empresa, con el fin de generar un entendimiento sobre estos y que finalmente sirvió como insumo para el despliegue del programa de fidelización que la empresa planeó lanzar. El algoritmo aplicado para el agrupamiento fue el PAM, el cual consiste en un proceso de partición iterativo similar al algoritmo k-means, con la diferencia de que el PAM utiliza como centroide la mediana y no la media, haciendo a este método más robusto ante valores atípicos. El algoritmo se basa en minimizar la suma de las diferencias entre una observación y la mediana. Todo el análisis y procesamiento de la data se realizó mediante el software R-Studio, donde finalmente se obtuvieron 3 grupos que describen como se comportan transaccionalmente los clientes objetivos de la empresa. Estos segmentos fueron denominados Vip, Plus y Regular, donde estarían ordenados de manera jerárquica, siendo los Vip los clientes más Top, los Plus los clientes intermedios y los Regulares los de menor rendimiento. Finalmente, estos segmentos le sirvieron a la empresa para entender cómo era el comportamiento de estos clientes, permitiendo sentar las bases para el programa de fidelización. Palabras claves: análisis clúster, algoritmo PAM, segmentación RFM, programa de fidelización, algoritmo k-means.

**UNIVERSIDAD NACIONAL AGRARIA LA MOLINA**

**FACULTAD DE ECONOMÍA Y PLANIFICACIÓN**

**"SEGMENTACIÓN RFM DE CLIENTES DE UN PROGRAMA DE FIDELIZACIÓN EN UNA EMPRESA RETAIL CON EL ALGORITMO DE ANÁLISIS CLÚSTER PAM"**

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERO ESTADÍSTICO INFORMÁTICO**

**Presentado por:**

**ALONSO ARMANDO DIEZ RIOS**

---

Dr. Cesar Higinio Menacho Chiok  
**PRESIDENTE**

---

Mg. Sc. Ivan Dennys Soto Rodríguez  
**ASESOR**

---

Mg. Grimaldo José Febres Huamán  
**MIEMBRO**

---

Mg. Sc. Rolando Jesús Salazar Vega  
**MIEMBRO**

## **Dedicatoria**

A mi madre Wilma y mi padre Dante, por el apoyo y motivación constante.

## **Agradecimiento**

Al profesor Mg. Sc. Ivan Dennys Soto Rodríguez, por su asesoría en la elaboración de la tesis, además de mi novia y tíos Ramón y Raquel por su motivación constante.

## ÍNDICE GENERAL

I.	TITULO .....	1
II.	INTRODUCCIÓN .....	2
	2.1 Problemática.....	2
	2.1.1. El área de acción .....	3
	2.1.2 Problema identificado .....	4
	2.2 Objetivo general .....	4
	2.3 Objetivos específicos.....	4
III.	REVISIÓN DE LITERATURA.....	5
	3.1 Marco Teórico.....	5
	3.1.1 Análisis Clúster .....	5
	3.1.2 Algoritmo K-means.....	8
	3.1.3 Algoritmo PAM.....	9
	3.1.4 Segmentación RFM.....	11
	3.1.5 Programas de Fidelización .....	13
IV.	METODOLOGÍA DEL TRABAJO.....	15
	4.1 Situación inicial.....	15
	4.2 Plan de trabajo y aplicación de competencias profesionales.....	16
	4.3 Tipo de Investigación .....	17
	4.4 Diseño de Investigación .....	18
	4.5 Población y muestra .....	18
	4.6 Instrumentos de colecta de datos.....	18
	4.6.1 Variables a Analizar .....	19
	4.7 Procedimiento del análisis de datos .....	20
	4.7.1 Identificación y tratamiento de datos atípicos .....	20
	4.7.2 Estandarización de los datos.....	20
	4.7.3 Análisis Clúster mediante el algoritmo PAM.....	20
V.	RESULTADOS Y DISCUSIÓN.....	22
	5.1 Identificación y tratamiento de datos atípicos .....	22
	5.2 Análisis descriptivo de las variables .....	23
	5.3 Estandarización de los datos .....	24
	5.4 Análisis Clúster mediante el algoritmo PAM.....	24

V. CONCLUSIONES .....	32
VI. RECOMENDACIONES .....	34
VII. REFERENCIAS BIBLIOGRÁFICAS .....	36

## ÍNDICE DE FIGURAS

<i>Figura 1.</i> Proceso iterativo en el análisis clúster mediante k-means.....	9
<i>Figura 2.</i> Variables RFM con escalas en quintiles .....	11
<i>Figura 3.</i> Relación Cliente - Vendedor.....	16
<i>Figura 4.</i> Relación 360 con el cliente .....	17
<i>Figura 5.</i> Proceso de recolección de datos para el análisis.....	19
<i>Figura 6.</i> Diagrama de cajas de la variable Recencia antes y después del procesamiento. ....	22
<i>Figura 7.</i> Diagrama de cajas de la variable Frecuencia antes y después del procesamiento. ....	22
<i>Figura 8.</i> Diagrama de cajas de la variable Monto antes y después del procesamiento. ....	23
<i>Figura 9.</i> Gráfico de Silhouette .....	25
<i>Figura 10.</i> Gráfico de Elbow .....	25
<i>Figura 11.</i> Visualización de la distribución de clientes en los clústers.....	26
<i>Figura 12.</i> Gráfica de silueta media de los clústers formados .....	27
<i>Figura 15.</i> Diagramas de cajas de la recencia por clúster. ....	30

## ÍNDICE DE TABLAS

<i>Tabla 1.</i> Matriz de actitud relativa y coeficiente de repetición .....	14
<i>Tabla 2.</i> Estadísticos descriptivos de las variables .....	24
<i>Tabla 3.</i> Distribución de clientes en los clústers PAM.....	26
<i>Tabla 4.</i> Análisis de variancia para las variables agrupadas con el algoritmo PAM.....	28
<i>Tabla 5.</i> Promedios de cada variable de acuerdo a los clústers .....	28
<i>Tabla 6.</i> Top 3 Categoría de productos más compradas por clúster .....	31

## RESUMEN

Las empresas enfrentan el desafío de manejar gran cantidad de datos, aprovecharlos y transformarlos en información valiosa que generen insights relevantes para toma de decisiones y alinear la estrategia. Una de las técnicas más utilizadas ha sido el análisis de agrupamiento en biología, ingeniería, psicología, y otras disciplinas, pero donde habría resultado de mayor utilidad es en los negocios, específicamente en investigación de mercados, desde la segmentación de clientes, identificación de hábitos de compra y otros. Así, la empresa de retail de la que se trata en este trabajo, opera bajo un enfoque centrado en el cliente, por lo cual busca mayor entendimiento del mismo, aprovechando los datos que dispone. El área de Inteligencia de Negocios es responsable del tratamiento y análisis de datos, proponiendo una segmentación RFM con un análisis clúster, que permitió agrupar a un conjunto de clientes objetivo para la empresa, para entenderlos mejor y sirvió como insumo para el programa de fidelización que la empresa lanzó. El algoritmo aplicado para el agrupamiento fue el PAM, que consiste en un proceso de partición iterativo similar al algoritmo k-means, pero que utiliza como centroide la mediana y no la media, haciendo al PAM más robusto ante valores atípicos. El algoritmo se basa en minimizar la suma de las diferencias entre una observación y la mediana. El análisis y procesamiento de datos se realizó con el software R-Studio, y se obtuvieron 3 grupos que describen como se comportan transaccionalmente los clientes objetivo de la empresa. Los segmentos fueron denominados Vip (clientes Top), Plus (clientes intermedio) y Regular (de menor rendimiento), ordenados jerárquicamente. Estos segmentos le sirvieron a la empresa para entender el comportamiento de estos clientes, sentando las bases para el programa de fidelización.

**Palabras clave:** algoritmo k-means, análisis clúster, algoritmo PAM, programa de fidelización, segmentación RFM,

## ABSTRACT

Companies face the challenge of managing a large amount of data, taking advantage of it and transforming it into valuable information that generates relevant insights for decision making and aligning the strategy. One of the most used techniques has been cluster analysis in biology, engineering, psychology, and other disciplines, but where it would have been most useful is in business, specifically in market research, from customer segmentation, identification of habits shopping and others. Thus, the retail company discussed in this work operates under a customer-centric approach, which is why it seeks a better understanding of it, taking advantage of the data it has. The Business Intelligence area is responsible for the treatment and analysis of data, proposing an RFM segmentation with a cluster analysis, which allowed grouping a set of target customers for the company, to understand them better and served as input for the loyalty program that the company launched. The algorithm applied for the grouping was the PAM, which consists of an iterative partitioning process similar to the k-means algorithm, but which uses the median and not the mean as the centroid, making the PAM more robust against outliers. The algorithm is based on minimizing the sum of the differences between an observation and the median. The analysis and processing of data was carried out with the R-Studio software, and 3 groups were obtained that describe how the target customers of the company behave transactionally. The segments were called Vip (Top customers), Plus (intermediate customers) and Regular (lowest performance), arranged hierarchically. These segments helped the company to understand the behavior of these customers, laying the foundations for the loyalty program.

**Keywords:** k-means algorithm, cluster analysis, PAM algorithm, loyalty program, RFM segmentation.

## **I. TITULO**

Segmentación RFM de clientes de un programa de fidelización en una empresa retail con el Algoritmo de Análisis Clúster PAM

## II. INTRODUCCIÓN

### 2.1 Problemática

La coyuntura originada por la pandemia del COVID-19 ha golpeado al mundo en muchos aspectos. Donde la economía se vio especialmente afectada, sobre todo en Latinoamérica, que en el 2020 se contrajo un -7.7% (Jones, Palumbo & Brown, 2021). La situación en Perú no fue diferente a esta realidad, ya que al cierre del 2020 sufrió una fuerte contracción de -11.1%. (Informe del BCRP).

En este marco, el sector comercio cerró el segundo trimestre del 2020 con una caída de 47.5% respecto al mismo periodo del 2019, resultado que se vio influenciado por el impacto negativo que tuvieron los retails, quienes cayeron en 23% en el mismo periodo, esto debido al cierre de tiendas por la inmovilización social que se dio en el país. (Perú Retail, 2020).

Sin embargo; el sector retail comenzó a mostrar signos de recuperación. De acuerdo con lo informado por PRODUCE, en el año 2021 las ventas de este sector en Perú ascendieron a S/47 mil 398 millones, representando un crecimiento de 18.4% versus el 2020 y de 17.2% respecto al 2019, año pre-pandemia (Plataforma digital única del Estado Peruano, 2022).

En línea con esta recuperación, un factor importante ha venido acompañando toda esta coyuntura, y es el comportamiento del cliente, el cual ha sufrido cambios importantes entre los que destacan 4 tendencias a resaltar.

- La primera hace referencia a la consolidación del hábito del ahorro y reducción del gasto no esencial, comportamiento que se reforzó sobre todo en los meses de confinamiento.
- La segunda es la nueva tendencia de relación de confianza con las marcas y la sostenibilidad que tienen estas en la vida de los consumidores.
- La tercera hace referencia a la aceleración vertiginosa que ha tenido la tecnología y digitalización.

- Mientras que la última es la relevancia que las personas le están dando a sus hogares (Gomez & Marin, 2021).

Por este motivo, las empresas de retail, conscientes de estos cambios, le han dado una mayor relevancia a un aspecto ya conocido desde hace algún tiempo y que debido a esta situación toma un mayor impulso; la estrategia de cliente al centro. En ese sentido como parte de mantenerse a flote y conseguir mayor éxito, las empresas de retail adoptan un enfoque centrado en el cliente, en el que las necesidades, preferencias y valores de este se convierten en un aspecto fundamental.

El objetivo de este enfoque está en brindar una gran experiencia para los clientes en cada etapa de su viaje, centrándose en lo que necesitan y en sus medios preferidos de comunicación e interacción. En ese sentido lograr la satisfacción y la lealtad de los clientes se convierte en la mejor manera de garantizar el crecimiento a largo plazo del negocio (Affde, 2021).

### **2.1.1. El área de acción**

Bajo este marco, la empresa de retail de la cual se habla en el trabajo de suficiencia profesional, no es ajena a esta realidad, por lo que, en búsqueda del aumento de su participación en el mercado, el área de Marketing y Comercial decide crear un programa de fidelización. Cabe mencionar que el área de Marketing y Comercial ya tenía un programa de atención personalizada que contaba con una cartera de clientes, los cuales eran frecuentes en la tienda.

Este programa no tenía una gestión directa y muy elaborada, sin embargo; esta cartera de clientes obtenía buenos resultados a nivel de venta y denotaba un potencial. Fue así que se determina lanzar el nuevo programa de fidelización. Con esta iniciativa, la empresa se enfocará en obtener un mayor conocimiento del cliente para poder medirlo y gestionarlo de una manera óptima, mediante la puesta en marcha de estrategias más personalizadas que vayan en línea con sus necesidades y obtener un mayor beneficio.

El mencionado programa de fidelización se basará en niveles o jerarquías, a los cuales el cliente puede alcanzar en función al comportamiento transaccional que posea, y de acuerdo con el nivel al que pertenezca, la empresa le otorga una serie de beneficios. De esta manera, brindando recompensas alcanzables en el corto plazo, así como recompensas aspiracionales en el mediano

y largo plazo, es como la empresa enfocará esfuerzos para generar una conexión bastante cercana con el cliente.

### **2.1.2 Problema identificado**

En este escenario es donde la empresa requiere entender más a profundidad al cliente identificando posibles tendencias en su comportamiento transaccional, que le permita obtener un insumo para categorizarlo en los niveles o jerarquías que fueran necesarios para el programa de fidelización.

Ante esta necesidad, se planteó realizar una segmentación de clientes a través de un análisis clúster utilizando el algoritmo PAM, en base a tres de las principales variables transaccionales (RFM) que se manejan dentro del área: El monto de compra, la frecuencia y la recencia.

Para dicha tarea, se utilizaron los datos de los clientes del antiguo programa de atención personalizada, que posteriormente pasarán a formar parte del programa de fidelización, en el periodo de enero a diciembre del 2021, los cuales se extrajeron del datalake de la empresa que se encuentra alojada en Google Cloud Platform, para después analizarlos mediante el software R.

### **2.2 Objetivo general**

Segmentar a los clientes pertenecientes a un programa de fidelización basados en variables RFM, utilizando un análisis clúster mediante el algoritmo PAM

### **2.3 Objetivos específicos**

- Plantear un número de segmentos en el que se dividan a los clientes del programa de fidelización que le sirva como insumo a la empresa para determinar los niveles o jerarquías necesarias.
- Obtener los diferentes segmentos del programa de fidelización en base a su comportamiento transaccional.
- Analizar y describir los patrones de comportamiento de compra en función a las variables en estudio, que tiene cada segmento encontrado que le permita a la empresa generar acciones direccionadas.

### **III. REVISIÓN DE LITERATURA**

#### **3.1 Marco Teórico**

##### **3.1.1 Análisis Clúster**

Según Thrun (2018) muchos métodos de minería de datos se basan en algún concepto de similitud entre los datos de interés que se piensan estudiar. Se han aplicado varios nombres a estos métodos de agrupación, dependiendo en gran medida del campo de aplicación, por ejemplo, en biología, psicología, investigación de mercados, sociología, economía, ingeniería, marketing, entre otros. Sin embargo; en el fondo su nombre se utiliza para definir una serie de técnicas cuyo objetivo es la búsqueda de grupos similares de individuos que se van agrupando en clústers o conglomerados de acuerdo a la relación natural que existe entre los individuos, es decir si se tiene una muestra de observaciones de ellos, el análisis clúster sirve para agruparlos en conglomerados lo más homogéneos posible en base a las variables que se definen.

En esa línea, de acuerdo a Gómez (2002) el origen de esta técnica se encuentra en la biología y la botánica, ya que investigadores de estos campos necesitaban agrupar las diversas especies de animales y vegetales, en categorías o familias que resultaran lo más homogéneas.

El análisis clúster se conoce desde hace varias décadas, donde en los años 30 se desarrollaron algunas pequeñas aplicaciones, aunque cuando comenzó a adquirir verdadero auge fue en los años 50. Teniendo mayores niveles de uso y proporcionando mayores beneficios en el campo de la segmentación de mercados y que en combinación con otras técnicas multivariadas permite desarrollar distintos tipos de estrategias.

Por su parte Singh & Rai (2010) menciona que el análisis clúster es una técnica que tiene como principal objetivo identificar grupos dentro de un conjunto de observaciones, en donde se busca que las observaciones dentro de los grupos sean lo más homogéneos posible, mientras que las observaciones entre diferentes grupos sean lo más heterogéneos posible. Para esto, se necesita utilizar alguna técnica de agrupamiento y dentro de este proceso, cuantificar el grado de similitud o disimilitud entre los datos, las cuales pueden ser obtenidas a través de diversas métricas de similitud, en donde los resultados obtenidos principalmente dependerán de la métrica utilizada. Las medidas de similitud se usan con frecuencia para datos categóricos y para datos que contienen variables continuas y categóricas, mientras que las medidas de disimilitud (medidas de distancia) se usan comúnmente para datos continuos.

Asimismo, según Csizér & Jamieson (2012) el análisis de clúster en comparación de otras técnicas estadísticas que se enfocan en reducción de datos, como por ejemplo un análisis factorial o un análisis discriminante, se distingue por el hecho de que no asume un modelo estadístico subyacente, por tanto, el investigador necesita tomar varias decisiones relacionadas al método de cálculo.

En ese sentido de acuerdo a Milligan & Cooper (1997) el análisis clúster se aplica siguiendo siete pasos

1. Selección los elementos u observaciones que serán agrupados.
2. Selección de las variables del conjunto de datos sobre las que agrupar las observaciones. En donde cada variable de contener suficiente información que permita el agrupamiento de los elementos.
3. Estandarización de las variables, con el objetivo de tener distancias unificadas, de lo contrario, las variables con valores más grandes contribuirán más a la medida de distancia que las variables con valores más pequeños.
4. Selección de una medida de similitud o disimilitud.
5. Selección del método de agrupamiento.

6. Determinación del número de clústers.

7. Interpretación de los clústers, la cual requiere del conocimiento y experiencia de la disciplina en el contexto aplicado por parte del investigador. Posteriormente está la prueba de los resultados, que involucra determinar si hay un agrupamiento significativo o una partición arbitraria de datos provocado por un ruido aleatorio. Finalmente, la replicación, en el sentido de determinar si la estructura de conglomerados resultante se puede replicar en otras muestras.

### **3.1.1.1 Medidas de distancia**

Para agrupar adecuadamente las observaciones, previamente se deberá determinar el tipo de distancia que se utilizará, estas pueden ser medidas de similitud o disimilitud.

Las medidas de similitud son una forma de determinar cuan relacionadas o cercanas se encuentran las observaciones, mientras que las medidas de disimilitud determinan cuan diferentes o alejados lo están. En ese sentido las medidas de distancia son el principio fundamental para el agrupamiento y la técnica utilizada dependerá en gran medida de la situación que se esté trabajando. Entre ellas se pueden encontrar la distancia Euclidiana, Manhattan, Canberra, Chebyshev, Mahalanobis, entre otras (Mahmud Harmouch, 2021).

### **3.1.1.2 Clasificación de los métodos de agrupamiento**

El análisis clúster es aplicado a diferentes ámbitos y realidades, por esta razón existen diversos métodos de agrupamiento, cuya elección dependerá del objetivo del análisis que se tenga. De acuerdo con Kaufman & Rousseeuw (2005) los métodos de agrupamiento pueden dividirse en dos grandes tipos:

- **Métodos de partición**

Clasifica las observaciones de un conjunto de datos en un determinado número de grupos ( $k$ ), los cuáles deben cumplir ciertos requisitos.

- Cada grupo debe contener al menos una observación.
- Cada observación debe pertenecer sólo a uno de los grupos, es decir, grupos diferentes no pueden tener observaciones en común.

Donde  $k$  debe definirse a priori y las observaciones se van intercambiando de manera iterativa entre cada grupo hasta encontrar la partición más adecuada, en el sentido de obtener observaciones más similares dentro de cada grupo y observaciones más diferentes entre grupos. Entre estos métodos se pueden encontrar a los algoritmos K-means, K-prototypes, K-modes, K-medoids, entre otros.

- **Métodos Jerárquicos**

Los algoritmos jerárquicos son un método alternativo el cual no requiere que se defina a priori el número determinado de grupos ( $k$ ). En ese sentido según Everitt, Landau, Leese & Stahl (2011) consiste en generar una serie de particiones, que pueden ir desde un solo grupo que contiene a todos los individuos, hasta  $n$  grupos, cada uno de los cuales contiene un solo individuo, con el objetivo de ir agrupando clúster para crear nuevos (métodos aglomerativos) o separar algún clúster ya formados para dar origen a otro dos (métodos divisivos) de modo que se logre una alta medida de similitud.

### **3.1.2 Algoritmo K-means**

El análisis clúster a través del algoritmo k-means es de bastante utilidad y popular en análisis exploratorios y minería de datos dentro de cualquier campo de investigación, dado a su facilidad de implementación, eficiencia computacional y bajo consumo de memoria, en un marco donde el poder computacional ha crecido a un ritmo a la par de la aparición de grandes volúmenes de datos (Morissette & Chartier, 2013).

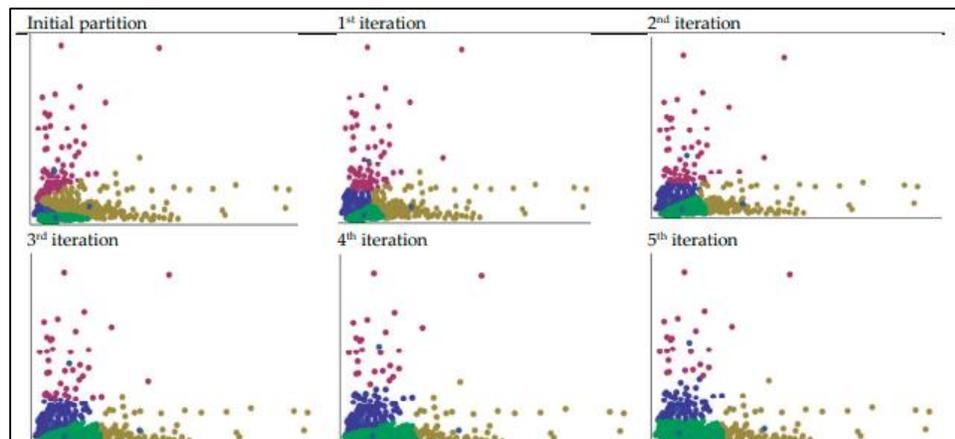
Este algoritmo es un método de partición iterativo el cual consiste en dividir en  $k$  grupos o conglomerados en donde  $k \geq 2$ , el valor de  $k$  deberá ser definido por el investigador.

El método primero selecciona de manera aleatoria  $k$  observaciones de los datos, los cuales serán los centros del conglomerado o lo que comúnmente se denomina centroide. Luego, se calcula la distancia de cada una de las demás observaciones versus cada centroide se calcula utilizando alguna métrica de distancia. De manera que se va asignando cada caso al centro más próximo. Este proceso se hace de manera iterativa, tomando como centroide en las próximas iteraciones al promedio de los grupos, hasta que no se observen más cambios en los grupos (Imron, Hasanah & Umaid, 2020)

En la figura 1 se observa un ejemplo de cómo cambian las observaciones en cada asignación a los centroides dentro de cada iteración.

**Figura 1.**

*Proceso iterativo en el análisis clúster mediante k-means*



FUENTE: Morissette & Chartier, 2013

### 3.1.3 Algoritmo PAM

Según Bhat (2014), Partition around medoids (PAM) es uno de los algoritmos para aplicar el método de clustering K-medoids, cuyo funcionamiento es similar al algoritmo de K-means, con la principal diferencia de la elección del centroide. Mientras en el K-means se selecciona la media, en el K-medoids se selecciona la mediana. En ese sentido el proceso comienza con la selección aleatoria de los centroides (medianas), que serán los representantes de los k grupos a formar y posteriormente las demás observaciones del conjunto de datos se van asignando al grupo cuya mediana es más cercana a ellos.

A partir de entonces, se determina un nuevo centroide que puede representar mejor el grupo e igualmente todas las observaciones restantes se asignan una vez más a los grupos que tienen la mediana más cercana. Este proceso se va realizando de manera iterativa, donde en cada iteración las medianas alteran su ubicación hasta que se logre minimizar la suma de las diferencias entre cada observación y su mediana correspondiente (Costo total del cambio). Esto marca el final del proceso y tenemos los grupos finales resultantes con sus medianas definidas, formándose los k grupos que se centran alrededor de las medianas y todas las observaciones se colocan en el clúster apropiado según la mediana más cercana.

Los pasos para el algoritmo son los siguientes:

- a. Seleccionar de manera aleatoria k observaciones que serán los centroides (medianas) de cada grupo.
- b. Para cada observación  $x_i$  no seleccionada como centroide se deberá calcular su similaridad o disimilaridad con el centroide  $m_k$ , de esta manera se busca asociar  $x_i$  al centroide más cercano.
- c. Calcular el costo total de intercambio  $S(x_i, m_k)$  de la observación  $m_k$  con cada una de las observaciones  $x_i$ .
- d. Si el costo total aumenta se deshace el intercambio y se regresa al paso 2.
- e. De lo contrario, se selecciona un nuevo centroide  $m_k$ , volver al paso 2, iterando así hasta que no haya más observaciones para agrupar.

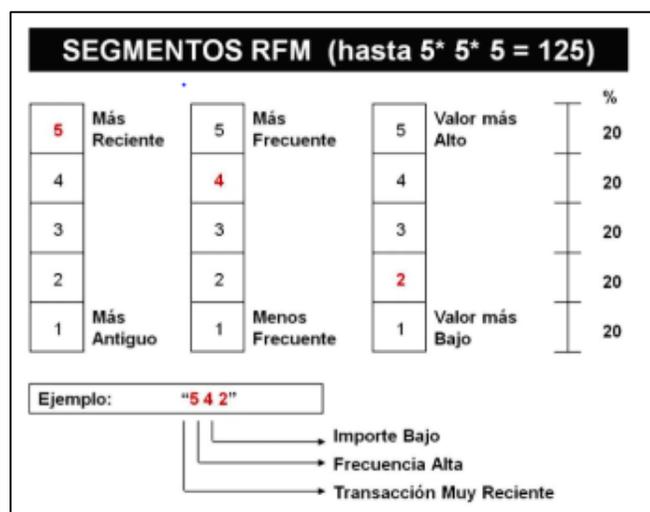
### 3.1.4 Segmentación RFM

La segmentación RFM es un método simple pero efectivo que se aplica a diversos campos de acción del mercado. Este análisis utiliza la información del comportamiento de compra del consumidor en un determinado periodo pasado, el cual se basa en tres variables que dan origen a su nombre. (R) Recencia, variable que hace referencia el periodo de tiempo transcurrido desde la última transacción realizada por el consumidor hasta la fecha actual. (F) Frecuencia, variable que representa la cantidad de transacciones realizadas por el cliente en un periodo de tiempo determinado. (M) Monto de compra, que es el dinero total gastado por el cliente en un periodo de tiempo determinado (Gustriansyah, Suhandi & Antonio, 2020).

De acuerdo con Rojas (2020), la segmentación RFM consiste en la construcción de escalas, basadas en las tres variables mencionadas al comienzo, dándole a cada cliente un valor según el percentil donde se encuentra. Donde lo más habitual suele ser realizar las escalas por quintiles, es decir a todas las variables se le asigna un valor entre el 1 al 5, donde 1 es la peor puntuación y 5 la mejor.

**Figura 2.**

*Variables RFM con escalas en quintiles*



FUENTE: Rojas, 2020

En la figura 2 se puede apreciar las escalas en quintiles para la recencia, frecuencia y monto de compra, donde los clientes de mayor valor serían los etiquetados como “5-5-5”, ya que serían los compradores más frecuentes y con un total de compras importante.

En línea con lo anterior, Srivastava (2016) menciona que las segmentaciones RFM parten del principio de la ley de Pareto donde el 20% de los principales clientes representan el 80% de los ingresos, en donde los clientes que muestran una puntuación alta de RFM normalmente debería realizar más transacciones y generar mayores ingresos. En ese sentido esta segmentación es de utilidad para diversos negocios ya que permite aumentar la comprensión del comportamiento de los clientes con el fin de generar una mejor gestión de estos en pro de su fidelización.

Adicionalmente, según Rahma, Fahd & Kankamol (2020) se utilizan bastante a menudo tres métodos de agrupamiento aplicados en los modelos de segmentación RFM, por su sencillez de entendimiento, estos son el método de K-means, K-medoids y DBSCAN.

El algoritmo K-means es sensible a los valores atípicos, por lo que puede distorsionar sustancialmente la distribución de datos, debido a que toman la media como centroide. En el algoritmo K-medoids se utiliza en cambio la mediana como centroide, lo cual no lo hace sensible a valores atípicos. El método DBSCAN utiliza los parámetros de punto de entrada mínimo (minpts) y epsilon (eps), el proceso de determinar los valores de los parámetros se basa en la prueba y error, lo que significa que la determinación de los valores de los parámetros debe probarse varias veces para obtener varios grupos.

Con la aplicación de estos métodos de agrupamiento se busca mejorar la segmentación RFM para que la determinación de los clústers sea más precisa, óptima y de mayor calidad (no necesariamente dividida en 5 grupos que representan el 20% de clientes cada uno, como la segmentación predeterminada en el análisis RFM), con lo que se conseguiría mayor homogeneidad dentro de los clúster y mayor heterogeneidad entre los clústers.

### **3.1.5 Programas de Fidelización**

Los programas de fidelización son claves en la estrategia de gestión de las relaciones con los clientes y tienen ventajas múltiples (Dominguez, 2017). Entre ellas, se tiene una disminución de la competencia en precios, la construcción de una base de datos de consumidores, que permite conocerles mejor y ofrecerles productos y servicios adaptados a sus necesidades, y la mejora de las tasas de retención de clientes y la intensificación de las relaciones con ellos (Lewis, 2004). Asimismo, existe la posibilidad de que los miembros de los programas de fidelización de una empresa olviden experiencias negativas, no la comparen con la competencia y compren exclusivamente en dicha empresa (Bolton, Kannan & Bramlett, 2000).

Los programas de fidelización empresariales se diferencian por la combinación de recompensas que ofrecen. Pueden ser tangibles, económicas, blandas o financieras; que se entregan a través de incentivos en forma de descuentos o regalos que se obtienen mediante la acumulación de compras. Por otro lado, las recompensas intangibles, duras, no económicas o no financieras no tienen un valor económico sino una orientación emocional, y tratan de brindar a los clientes una sensación de reconocimiento o que se sientan especiales en comparación con otros clientes (Dominguez, 2017).

De las distintas recompensas que ofrecen, ya sean tangibles como intangibles, se pueden derivar una serie de beneficios percibidos por parte de los miembros. Estos beneficios se definen como “el valor percibido que los consumidores obtienen por las recompensas del programa, es decir, lo que el programa puede proporcionar a sus miembros” (Mimouni-Chaabane & Volle, 2010, p. 32).

**Tabla 1.**

**Matriz de actitud relativa y coeficiente de repetición**

		ALTA	BAJA
ACTITUD RELATIVA	ALTA	Fidelidad o lealtad verdadera	Fidelidad latente
	BAJA	Fidelidad espuria o falsa lealtad	No fidelidad

FUENTE: Dick & Basu, 1994

Dick & Basu (1994) construyen una matriz en la que combinan la actitud relativa y el coeficiente de repetición para identificar diferentes niveles de fidelidad.

La verdadera fidelidad exige la existencia de las dos dimensiones de referencia: los comportamientos de consumo y las actitudes favorables del consumidor (Trinquecoste, 1996). La fidelidad latente y la fidelidad espuria vienen determinadas por la existencia de factores situacionales y normas sociales que restringen respectivamente, a una fidelidad comportamental (Castañeda, 2005).

## **IV. METODOLOGÍA DEL TRABAJO**

El presente trabajo tiene como finalidad mostrar el aporte otorgado por el área de Business Intelligence perteneciente a la gerencia de Marketing y Comercial, área de la cual el profesional que suscribe la monografía forma parte. Todo esto en un marco donde los datos están tomando mayor relevancia día a día y la gran cantidad de estos pone en evidencia la necesidad de explotarlos y obtener el máximo beneficio posible.

En ese sentido, la empresa enfoca una estrategia centrada en el cliente, en el que busca generar el mayor entendimiento del mismo y es en esta instancia donde el área de Business Intelligence entra a brindar soporte.

### **4.1 Situación inicial**

La empresa de retail, tenía un programa de atención personalizada para un grupo de clientes, que empezó a inicios del 2021; los cuales eran captados y atendidos directamente por el personal de las tiendas, dicho programa estaba implementado solo en determinadas tiendas, específicamente en tres. Era justamente el personal de la tienda los que basados en su criterio, experiencia y conocimiento de los clientes más asiduos; los invitaban a participar en el programa, en el que se les ofrecía una atención más personalizada, que consistía en una estrecha relación cliente-vendedor. Es así como, cuando los clientes llegaban a las tiendas, buscaban a estos vendedores para que los atiendan y asista en sus compras, donde adicionalmente existía un contacto remoto para separación de productos y posterior recojo en tienda.

Todo este contacto y atención hacia los clientes, comenzó a ganar relevancia y aceptación, hasta que se consiguió una cartera de ellos que se fue volviendo bastante frecuente. Sin embargo; si bien el personal de la tienda se encargaba de velar por la captación, satisfacción y la atención personalizada, la escala que estaba tomando el programa no hacia sostenible el proyecto a lo largo del tiempo. Además, no existía un conocimiento profundo del comportamiento de estos clientes que permitiera llevar su gestión a otro nivel y posterior abordaje bajo una estrategia 360.

**Figura 3.**

*Relación Cliente - Vendedor*



FUENTE: Elaboración propia

Es así como a finales del 2021, el personal de las tiendas ya no se daba abasto para poder asumir una gestión de calidad sobre una cartera de clientes en crecimiento, sumado a los buenos resultados a nivel de ventas y contribución que se venían obteniendo es que se decide abordar la situación desde una óptica de mayor escala.

#### **4.2 Plan de trabajo y aplicación de competencias profesionales**

Ante el contexto anteriormente mencionado, es que el equipo de Marketing y Comercial decide transformar este programa de atención personalizada en un programa de fidelización. Para esto primero necesitaban obtener un conocimiento e información relevante que les permita sentar las

bases y reglas del programa. En esta instancia y en base a un trabajo en conjunto con el área de Business Intelligence, este último propone trabajar un criterio de segmentación de clientes basado netamente en su comportamiento transaccional, sustentado en lineamientos corporativos y de estrategias de comunicación de cara al cliente.

En línea con esto se realizó una segmentación RFM, mediante el algoritmo de análisis clúster PAM, logrando así obtener una visibilidad de en cuantos segmentos de valor se dividían naturalmente los clientes y cómo se distribuían dentro de cada segmento. Aportando de esta manera un conocimiento sobre los clientes que le sirvió al área de Marketing para que pudiera sentar las bases y estrategias de comunicación 360, con el que se buscaba ampliar la cartera de clientes e incrementar el consumo y frecuencia de estos.

**Figura 4.**  
*Relación 360 con el cliente*



FUENTE: Elaboración propia

**4.3 Tipo de Investigación**

El tipo de investigación fue descriptiva ya que se especificaron en todos sus componentes una realidad, en este caso se buscó identificar distintos segmentos de clientes, describiéndolos en base a patrones de comportamiento transaccional.

#### 4.4 Diseño de Investigación

Mientras que el diseño de investigación fue de tipo no experimental y transversal, debido a que los datos para la realización de la investigación se tomaron en un solo momento. En este caso se obtuvieron y analizaron los datos de ventas y transacciones de los clientes en un momento determinado.

#### 4.5 Población y muestra

- **Población:** Todos los clientes que son parte del anterior programa de atención personalizada, de los cuales se llevaba un registro.
- **Muestra:** 10,783 clientes que realizaron compras en el periodo de enero a diciembre del 2021.

#### 4.6 Instrumentos de colecta de datos

Los datos de los clientes del antiguo programa de fidelización eran administrados por el personal de la tienda, de modo que primero se debía proceder con la migración de estas identificaciones al datalake de la empresa que se encuentra en Google Cloud Platform (GCP), el cual es un conjunto de servicios de computación, almacenamiento, redes, macrodatos, aprendizaje, entre otros, dentro de la nube ofrecidos por Google (Bigelow, 2022). Para esto se hizo uso de una carpeta FTP, mediante la cual se almacenaron estos datos y posteriormente se generó un flujo para trasladar los datos hacia el GCP.

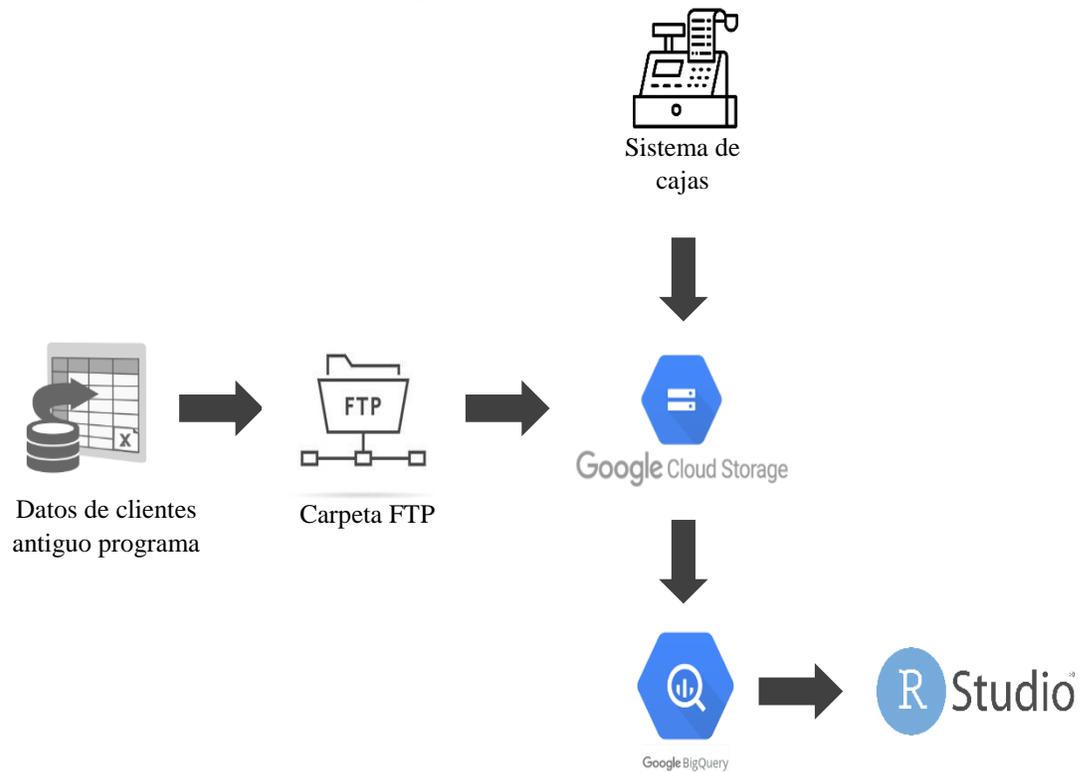
Asimismo, también se hace uso de las transacciones realizadas por estos clientes, que pasan por el sistema de cajas, viajan y se almacenan en el datalake.

En ese sentido, la recolección de los datos se realizó consultando el comportamiento transaccional de los clientes del antiguo programa de fidelización a través de BigQuery, que según Sato (2012) es un servicio de consulta interactivo basado en la nube y completamente

administrado para conjuntos de datos masivos. A través de este servicio se obtuvieron los datos para los 10,783 clientes con compras en el periodo de enero a diciembre del 2021. Finalmente, estos datos serán llevados al software R para su posterior análisis.

**Figura 5.**

*Proceso de recolección de datos para el análisis*



FUENTE: Elaboración propia

#### 4.6.1 Variables a Analizar

- **Recencia:** Variable cuantitativa que representa el tiempo transcurrido desde la última compra hasta la actualidad.
- **Frecuencia:** Variable cuantitativa que representa el número de días que los clientes fueron a comprar en el periodo de análisis.
- **Monto de compra:** Valor monetario de las compras realizadas por los clientes en el periodo de análisis.

## **4.7 Procedimiento del análisis de datos**

Para realizar la segmentación de clientes en base a las variables RFM, se ejecutó un análisis clustering mediante el algoritmo PAM, para lo cual se procedió de la siguiente manera:

### **4.7.1 Identificación y tratamiento de datos atípicos**

Ya que todas las variables en análisis son cuantitativas, se procedió al tratamiento de valores atípicos para cada una de ellas. Primero se realizaron diagramas de cajas para identificar posibles valores atípicos superiores o inferiores. Posteriormente los valores atípicos superiores se reemplazaron por el percentil 97, mientras que los valores atípicos inferiores se reemplazaron por el percentil 3.

### **4.7.2 Estandarización de los datos**

Debido a que los datos para cada variable estaban en escalas diferentes, para que alguna, debido a su escala, no interfiera o distorsione el algoritmo de segmentación, se procedió a estandarizarlas, el cual es un proceso de transformación de escala en la distribución de cada variable que permite hacer comparaciones entre ellas eliminando efectos de influencia de los datos (Rodó, 2019)

### **4.7.3 Análisis Clúster mediante el algoritmo PAM**

#### a) Selección de la medida de distancia

Se utilizó la distancia euclidiana, función de distancia más común utilizada para variables numéricas.

#### b) Elección del número óptimo de clúster

Se utilizó el coeficiente de Silueta para estimar el número óptimo de clúster.

#### c) Realización del algoritmo PAM

Con la cantidad óptima de clúster obtenida en el punto anterior, se procedió a ejecutar el algoritmo PAM.

#### d) Análisis de los resultados

Luego de obtener los segmentos y la distribución de los clientes dentro de cada uno de ellos, se procedió a analizar y describir de qué manera se comporta cada segmento en base a las variables en estudio.

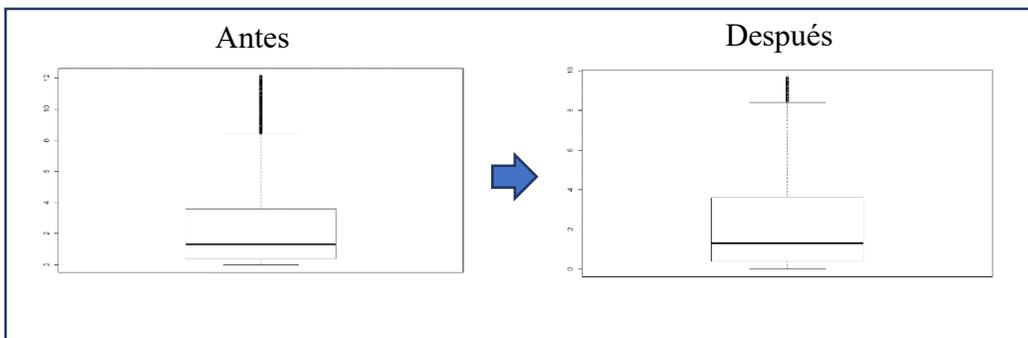
## V. RESULTADOS Y DISCUSIÓN

### 5.1 Identificación y tratamiento de datos atípicos

Primero para identificar los datos atípicos se realizaron diagramas de cajas para analizar cada variable, encontrando que las tres variables contienen valores atípicos superiores por lo que estos se reemplazaron por los respectivos percentiles 97.

**Figura 6.**

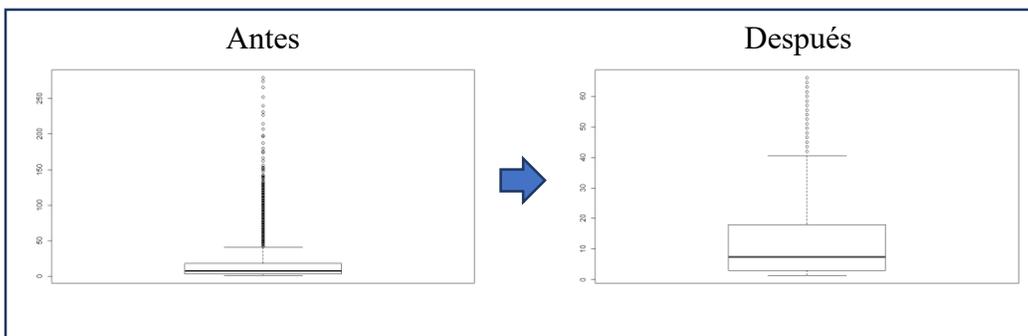
*Diagrama de cajas de la variable Recencia antes y después del procesamiento.*



FUENTE: Elaboración propia

**Figura 7.**

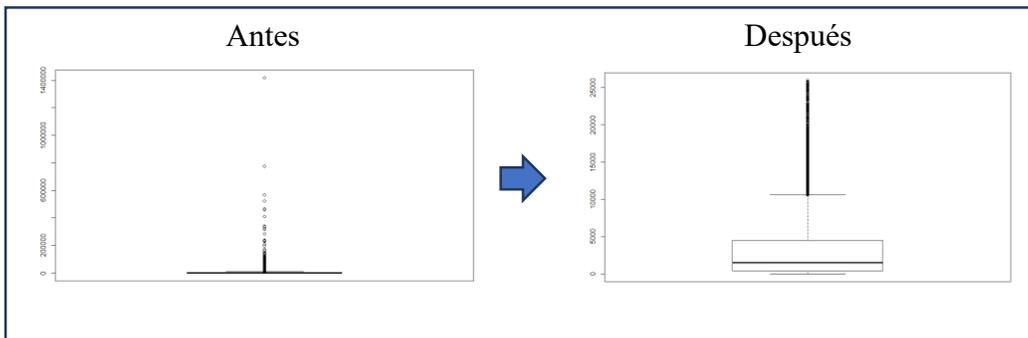
*Diagrama de cajas de la variable Frecuencia antes y después del procesamiento.*



FUENTE: Elaboración propia

**Figura 8.**

*Diagrama de cajas de la variable Monto antes y después del procesamiento.*



FUENTE: Elaboración propia

## 5.2 Análisis descriptivo de las variables

En esta sección se presenta un análisis descriptivo de 3 variables cuantitativas en estudio.

Tal y como se observa en la tabla 2, la recencia tiene un mínimo de 0 y un valor máximo 12, siendo el valor más frecuente de 1.3, lo que equivale a decir que estos clientes visitaron la tienda con una anterioridad de poco más de un mes. Adicionalmente se puede mencionar que hay un 25% de los clientes que tienen una recencia de al menos 3.6.

Respecto a la frecuencia, esta tiene un mínimo de 1.5 y un máximo de 66 días, siendo el valor más frecuente el de 7.5 días, con un percentil 25 de 3, lo que quiere decir que existen al menos 25% de clientes con una frecuencia de hasta 3. Y un percentil 75 de 18, que nos dice que hay un 25% de clientes que va a la tienda como mínimo 18 veces al año.

Finalmente, en cuanto al monto, este tiene un mínimo de 2.2 y un máximo de 19,489.2 soles, con una mediana de 1,100.2 soles. Además, un 25% de clientes gastan como mínimo 3,361.3 soles al año.

**Tabla 2.**

*Estadísticos descriptivos de las variables*

<b>Estadísticos</b>	<b>Recencia</b>	<b>Frecuencia</b>	<b>Monto</b>
Mínimo	0.0	1.5	2.2
Percentil 25	0.4	3.0	301.7
Mediana	1.3	7.5	1100.2
Media	2.4	13.9	2900.4
Percentil 75	3.6	18.0	3361.3
Máximo	9.6	66.0	19489.2

FUENTE: Elaboración propia

### **5.3 Estandarización de los datos**

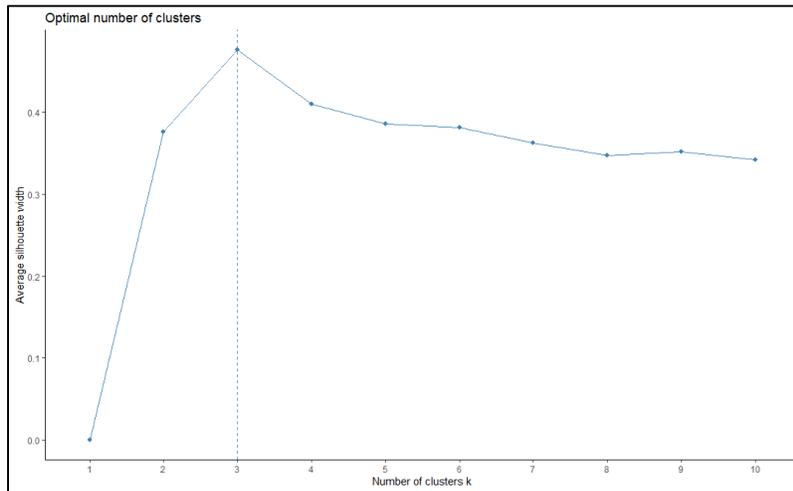
Con el objetivo de eliminar cualquier ruido en los datos que podrían afectar el agrupamiento, debido a que se encuentran en diferentes escalas, se buscó igualar de alguna forma las 3 variables mediante la estandarización con la función scale del software R.

### **5.4 Análisis Clúster mediante el algoritmo PAM**

Para realizar el agrupamiento, previamente se definió la cantidad óptima de clústers, para esto se utilizó la librería factoextra, la cual brinda la función fviz\_nbclust que permite realizar el gráfico de Silhouette y también el gráfico de Elbow. En ese sentido luego de aplicar dichas funciones se obtuvieron los gráficos representados en la figura 9 y 10, donde se observa que el número óptimo de clústers es de 3, es decir que este agrupamiento deberá trabajarse con 3 segmentos de clientes.

**Figura 9.**

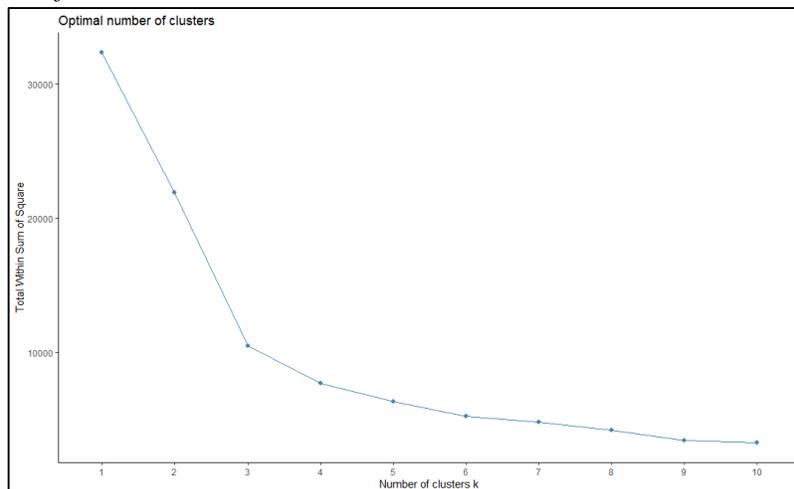
*Gráfico de Silhouette*



FUENTE: Elaboración propia

**Figura 10.**

*Gráfico de Elbow*



FUENTE: Elaboración propia

Luego de haber determinado el número de clústers óptimo, se ejecutó el análisis de agrupamiento mediante el algoritmo PAM, obteniéndose los 3 segmentos con distinta cantidad de clientes en cada uno. En el siguiente cuadro se puede observar la distribución de clientes en cada segmento.

**Tabla 3.**

*Distribución de clientes en los clústers PAM*

Clúster	Clientes	%
1	6,152	57%
2	2,058	19%
3	2,573	24%

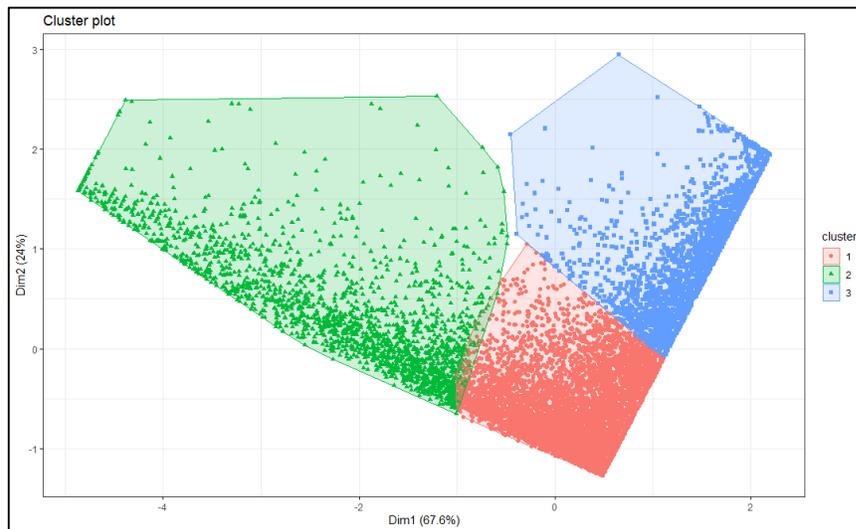
FUENTE: Elaboración propia

Como se puede observar la mayor cantidad de clientes cae en el clúster 1 que agrupa a 6,152, representado a más de la mitad de los clientes (57%), mientras que la menor cantidad de clientes se encuentra en el clúster 2 con 2,058 (19%).

Adicionalmente usando la función `fviz_cluster`, se ilustró la distribución de los clientes en cada clúster.

**Figura 11.**

*Visualización de la distribución de clientes en los clústers*



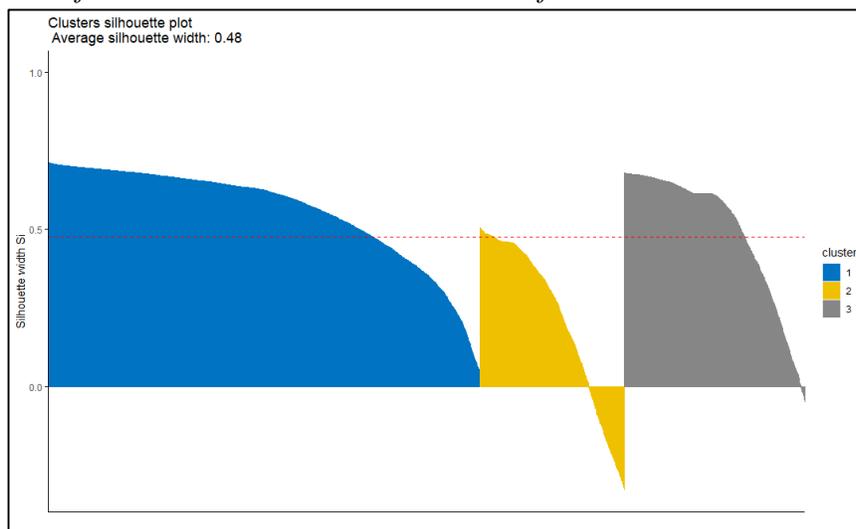
FUENTE: Elaboración propia

Asimismo, si se observa el gráfico de silueta de cada clúster en la figura 12, donde la silueta promedio es 0.48, siendo una media aceptable, en un marco donde este valor puede estar entre -1 a 1 y mientras más cerca de 1 se encuentre representa que los grupos son bastante densos y bien separados, que es lo que se busca idealmente, sin embargo; no es usual encontrarse con valores muy altos de una silueta media, por lo cual se menciona que un valor de 0.48 sería

aceptable, adicional a esto se tiene que no hay ningún clúster que tenga una silueta por debajo del promedio, en ese sentido los clústers formados serían relativamente aceptables.

**Figura 12.**

*Gráfica de silueta media de los clústers formados*



FUENTE: Elaboración propia

Además, con el objetivo de validar la segmentación se realizó un análisis de variancia para las 3 variables cuantitativas en estudio considerando como niveles del factor, los 3 clústers creados. Con esto se quiere probar la hipótesis que afirma que las medias de los grupos son iguales. De acuerdo a la tabla 4, las pruebas resultaron altamente significativas para las 3 variables, rechazándose de esta manera la hipótesis de igualdad de medias, con lo cual es posible afirmar que la variancia entre grupos es mayor que la variancia dentro de los grupos. Demostrando de esta manera que el agrupamiento realizado es válido y cumple con el principio que se busca de heterogeneidad entre los clústers y homogeneidad dentro de ellos

**Tabla 4.***Análisis de variancia para las variables agrupadas con el algoritmo PAM*

Variable	Varianza	gl	Suma de cuadrados	Media de cuadrados	F-value	P-value	Sig
Recencia	Entre grupos	2	56880	28440.1	17079	< 2.2e-16	***
	Dentro de los grupos	10780	17951	1.7			
	Total	10782	74831				
Frecuencia	Entre grupos	2	1753552	876776	10573	< 2.2e-16	***
	Dentro de los grupos	10780	893910	83			
	Total	10782	2647462				
Monto	Entre grupos	2	1.2482E+11	6.24E+10	8298.9	< 2.2e-16	***
	Dentro de los grupos	10780	8.11E+10	7.52E+06			
	Total	10782	2.05887E+11				

FUENTE: Elaboración propia

Posteriormente se analizó el comportamiento de cada clúster respecto a las variables en estudio, donde al observar tanto la tabla 5 como la figura 9,10 y 11 se puede afirmar que los clientes agrupados en el clúster 2 son los que poseen mayor monto de compra, frecuencia y la menor recencia, seguido del clúster 1, mientras que el clúster 3 es el de menor rendimiento.

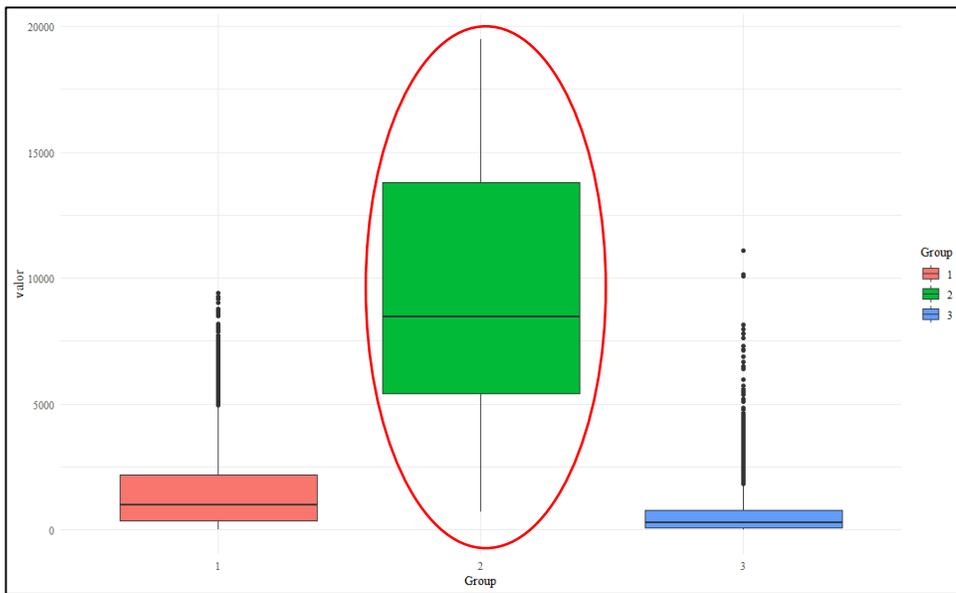
**Tabla 5.***Promedios de cada variable de acuerdo a los clústers*

Clúster	Recencia	Frecuencia	Monto
1	1.3	9.5	1498
2	0.6	39.8	9871
3	6.5	4.1	678

FUENTE: Elaboración propia

**Figura 13.**

*Diagramas de cajas del monto por clúster*

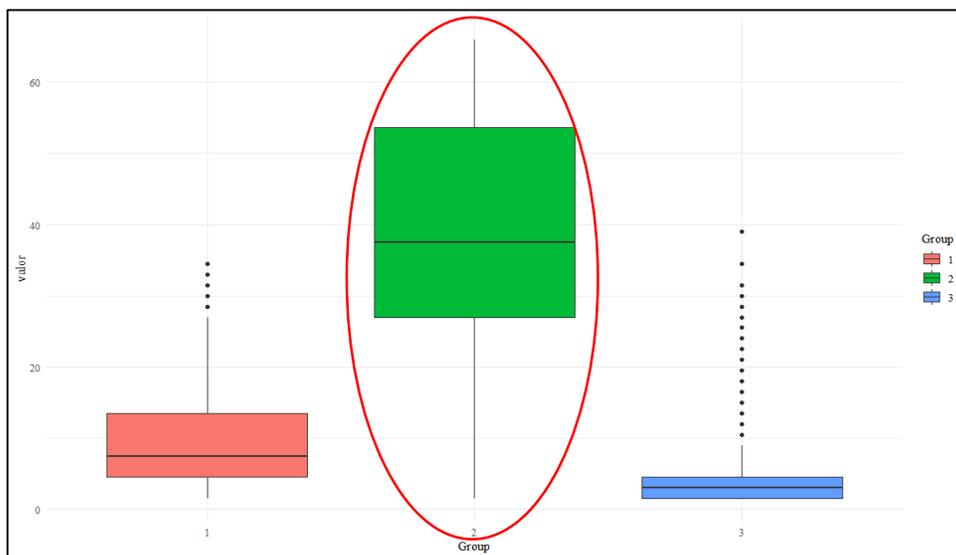


FUENTE: Elaboración propia

En la figura 13 se observa el diagrama de cajas para la variable venta por cada clúster, donde el clúster número 2 es el que tiene una mayor mediana, seguido por el número 1.

**Figura 14.**

*Diagramas de cajas de la frecuencia por clúster*

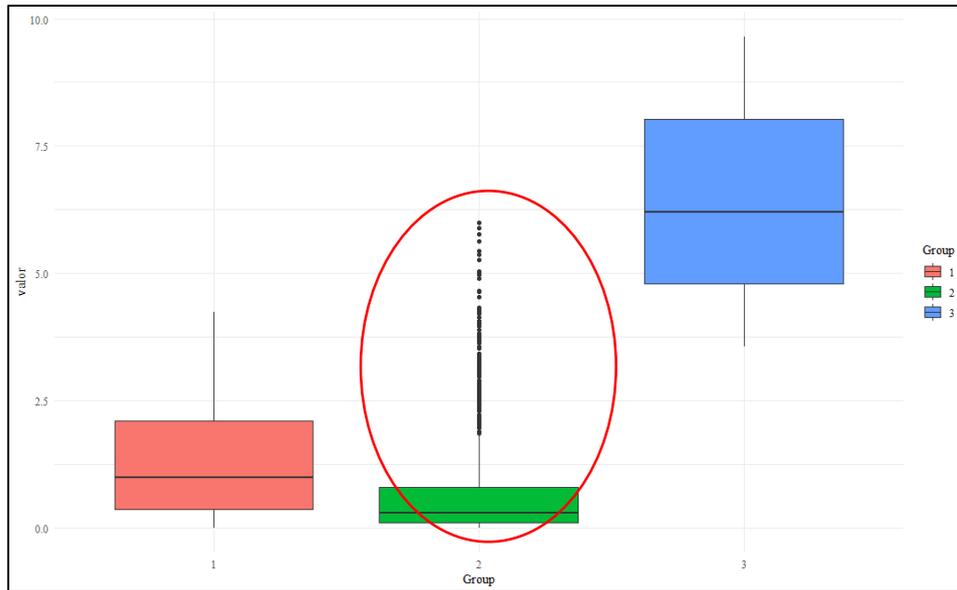


FUENTE: Elaboración propia

En la figura 14 se observa el diagrama de cajas para la variable frecuencia por cada clúster, donde el clúster número 2 es el que tiene una mayor mediana, seguido por el número 1.

**Figura 15.**

**Diagramas de cajas de la recencia por clúster.**



FUENTE: Elaboración propia

En la figura 15 se observa el diagrama de cajas para la variable recencia por cada clúster, donde el clúster número 2 es el que tiene una menor mediana, seguido por el número 1.

Finalmente, con los clústers obtenidos se realizó un cruce para identificar cuál es el top 3 de categorías más compradas por cada clúster. De manera que, si se observa la tabla 6, se tiene que la categoría de productos más comprada por el clúster 1 es la categoría A, para el clúster 2 es la categoría B y para el clúster 3 es la categoría F.

En base a esto, se obtiene un insumo adicional, ya que se adquiere el conocimiento de los productos más comprados por cada clúster con lo que se puede por un lado dirigir mejor la comunicación de cara a los clientes, ofreciéndoles productos a sus necesidades, y por otro tomar esta información como palanca para poder el planteamiento de los beneficios del programa de fidelización.

**Tabla 6.***Top 3 Categoría de productos más compradas por clúster*

Clúster	Categoría productos	Venta
1	Categoría A	S/ 1,074,256
1	Categoría B	S/ 1,054,253
1	Categoría C	S/ 724,623
2	Categoría B	S/ 5,143,118
2	Categoría D	S/ 4,326,575
2	Categoría E	S/ 3,099,750
3	Categoría F	S/ 245,699
3	Categoría G	S/ 193,982
3	Categoría D	S/ 153,753

FUENTE: Elaboración propia

## V. CONCLUSIONES

1. Luego de realizar el análisis para la segmentación basado en las 3 variables de comportamiento transaccional: Recencia, Frecuencia y Monto. Primero se logró determinar la cantidad óptima de clústers para el conjunto de datos en estudio, en este caso se identificaron 3.
2. Mediante el algoritmo PAM se logró el agrupamiento de los 3 segmentos, dando con una distribución de 57% para el clúster número 1, 19% para el 2 y 24% para el 3.
3. A través del indicador de silueta media se obtuvo un valor de 0.48, lo que nos indica que los clústers formados son aceptables. Además, mediante un análisis de variancia para cada una de las 3 variables en estudio, donde todas las pruebas resultaron altamente significativas, se validó la heterogeneidad entre clústers y homogeneidad dentro de ellos, demostrando la consistencia del agrupamiento.
4. En base a los 3 clústers encontrados podemos describir el comportamiento que caracteriza a cada uno, estos fueron etiquetados como clientes Vip, Plus y Regular.
  - Clúster 1 (Plus): Estos clientes son aquellos cuyo rendimiento de compras es intermedio, los cuáles en un año gastan en promedio S/ 1,498, van a la tienda 10 veces y su recencia promedio es de poco más de un mes.
  - Clúster 2 (Vip): Estos son los clientes de mayor valor, los cuáles en el periodo de un año gastan en promedio S/ 9,871, además de ser clientes que van a la tienda con bastante frecuencia, en promedio 40 veces al año, aproximadamente entre 3 o 4 veces al mes y tienen una recencia de compra promedio de menos de un mes. Este segmento de clientes representa el 19% del total, siendo el más pequeño, pero con gran rendimiento.

- Clúster 3 (Regular): Estos son los clientes con más bajo rendimiento, quienes gastan en promedio en un año S/ 678, con una frecuencia de visita a la tienda de 4 y una recencia promedio de 7 meses.
5. Con esto se logra tener un insumo que le sirve a la empresa para obtener un conocimiento de cómo se distribuye naturalmente su cliente objetivo y a la vez obtener un conocimiento del comportamiento de cada uno que le permitirá sentar las bases para el programa el despliegue del programa de fidelización
  6. Finalmente, el análisis realizado permite concluir que los clientes objetivo del programa de fidelización tiene una madurez intermedia, ya que se ha identificado que la masa de clientes actualmente se encuentra en los clientes Plus, que son los de jerarquía intermedia, los cuales serían más cercanos y fáciles de migrar al segmento Vip que los clientes Regulares. En ese sentido se tendrán que aplicar acciones dirigidas a los clientes de nivel bajo e intermedio para que puedan aumentar de valor y también a los de mayor nivel para que puedan mantenerse.

## VI. RECOMENDACIONES

1. Para los clientes Vip, que es el segmento más top, los cuales son clientes que van regularmente a la tienda y tienen un consumo alto. Los beneficios que el equipo de marketing podría entregar son los siguientes:

- Descuentos superiores en productos, principalmente en los de la categoría A.
- Acumulación superior de puntos para canjear cualquier producto.
- Envío de muestras de productos nuevos con el objetivo de demostrarles exclusividad y a la vez promover su consumo.
- Vales gratuitos para actividades de recreación (Cines, restaurantes, etc)
- Flete de envío de productos gratuito.
- Reposición o renovación gratuita de productos seleccionados por mantenerse en el nivel por un determinado tiempo.

El enfoque con estos clientes debería estar en acciones que generen sensación de exclusividad para promover el mantenimiento del nivel. Donde habría que tener un contacto constante con ellos y una atención personalizada en tiendas.

2. Respecto a los clientes Plus, estos son los que tienen un comportamiento de compra intermedio. En ese sentido los beneficios que el equipo de marketing podría entregar son los siguientes:

- Descuentos intermedios en productos, principalmente en los de la categoría B.
- Acumulación intermedia de puntos para canjear productos seleccionados.
- Descuento en el flete de envío de productos.
- Sorteo de muestras de productos nuevos.

Estos representan la mayor proporción de clientes objetivo, estos clientes si bien no tienen un comportamiento de consumo top, ya existiría un relacionamiento establecido con la empresa y que no sería de mucha dificultad que pasen al siguiente nivel.

3. Los clientes regulares que son lo de menor rendimiento y el equipo de marketing podría entregarles los siguientes beneficios:
  - Descuentos regulares en productos, principalmente en productos de la categoría F.
  - Acumulación intermedia de puntos para canjear productos seleccionados.

Estos clientes probablemente compran de manera esporádica ya que tienen un consumo compartido con tiendas de la competencia. Aquí la tarea está en resaltarles los beneficios que obtendrían por comprar en tiendas de la marca, a través de acciones de marketing directo. Además de realizar encuestas de percepción y satisfacción que permitan identificar posibles dolores.

4. Teniendo ya la información de los segmentos de valor de los clientes en base a su comportamiento transaccional. Posteriormente se podrían cruzar estos resultados con otras variables como datos sociodemográficos (Edad, Sexo, Lugar de Residencia, etc), con forme se vayan captando, para obtener una mirada adicional de cómo se distribuyen los clústers.
5. Ejecutar un modelo de propensión de compra para las distintas categorías de productos que tendrían los clientes de cada clúster, con el fin de generar comunicación aún mejor, direccionada al cliente.

## VII. REFERENCIAS BIBLIOGRÁFICAS

Affde. (10 de agosto de 2021). Estrategia centrada en el cliente: cómo transformar su organización en 2021. Recuperado de <https://bit.ly/3xzVobu>

Jones et al (2021). Coronavirus: 8 gráficos para entender cómo la pandemia ha afectado a las mayores economías del mundo. BBC News. Recuperado de <https://www.bbc.com/mundo/noticias-55802814>

Perú Retail (29 de septiembre de 2020). Sector retail muestra signos de recuperación. Recuperado de <https://bit.ly/3M7F1qz>

Plataforma digital única del Estado Peruano (1 de marzo de 2022). Ventas del sector retail alcanzan los S/ 47 mil millones y superan nivel prepandemia. Recuperado de <https://bit.ly/381PISk>

Gomez et al (2021) Incidencia del COVID-19 en el comportamiento del Consumidor y la adaptación del Sector Retail. Trabajo presentado por fin de Grado, Universidad Pontificia Comillas. Recuperado de <https://bit.ly/3JQiTIX>

Singh et al (2010). A Survey of Clustering Techniques. International Journal of Computer Applications 7(12): 1-5. Recuperado de <https://bit.ly/3MdLW1d>

Harmouch, M. (13 de Marzo de 2021). 17 tipos de medidas de similitud y disimilitud utilizadas en la ciencia de datos [Mensaje en un blog]. Recuperado de <https://bit.ly/3JNIUR6>

Kaufman et al (2005). Finding Groups in Data An Introduction to Cluster Analysis [versión digital]. Recuperado de <https://bit.ly/3xu55rQ>

Everitt et al (2011). Cluster Analysis [versión digital]. Recuperado de <https://bit.ly/37jEWRX>

Bhat, A. (2014). K-Medoids Clustering using partitioning around medoids for performing face recognition. Department of Electrical Engineering , IIT Delhi, Hauz Khas, New Delhi 3 (3). Recuperado de <https://bit.ly/3xF5nwd>

Gustriansyah et al (2020). Optimización de agrupamiento en análisis RFM Basado en k-Means. Revista Indonesia de Ingeniería eléctrica y ciencias de la computación 18(1). doi: <http://doi.org/10.11591/ijeecs.v18.i1.pp470-477>

Rojas, A. (2020) Análisis de datos para identificar perfiles de clientes por características similares de hábitos de compras. Trabajo presentado por fin de Máster, Universidad Politécnica de Madrid. Recuperado de <https://bit.ly/3rz3CfV>

Srivastava, R. (Marzo de 2016). Identificación de clústeres de clientes usando RFM Modelo: un caso de clasificación diversa de compradores. 18th International Conference on Business Intelligence, Analytics, and Knowledge Management. Conferencia llevada a cabo en Singapur. Recuperado de <https://bit.ly/37Yq5vT>

Rahma et al (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. Lontar Komputer 11(1). doi: 10.24843/LKJITI.2020.v11.i01.p04

Bigelow, S. (Marzo de 2022). Nube de Google [Mensaje en un blog]. Recuperado de <https://bit.ly/3xA3u3J>

Sato, K. (2012). An Inside Look at Google BigQuery. Recuperado de <https://bit.ly/3EmPZ8N>

- Rodó, P. (Julio de 2019). Normalización estadística. Recuperado de: <https://bit.ly/3zzhzOe>
- Csizér, K.; Jamieson J. (Noviembre de 2012). Cluster Analysis. Recuperado de <https://bit.ly/3RjxoAa>
- Milligan et al (1987). Methodology Review: Clustering Method 11(4): 329-354. Recuperado de <https://bit.ly/3wWeW8I>
- Thrun, C. (Enero de 2018). Approaches to Cluster Analysis. Recuperado de <https://bit.ly/3QiKJYn>
- Gómez, M. (Enero de 2002). El análisis clúster en investigación de marketing: Metodología y crítica. Recuperado de <https://bit.ly/3TFLI7L>
- Imron et al (2020). Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping 3(1). Recuperado de <https://bit.ly/3eczHzY>
- Dominguez, A. (2017). Los programas de fidelización online: un estudio desde la perspectiva del Engagement Marketing y las comunidades de marca. (Tesis doctoral). Universidad de Valladolid, España.
- Lewis, M. (2004). The influence of loyalty programs and short-term promotions on customer retention. *Journal of Marketing Research*, 41(3), 281-292.
- Bolton et al (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, 28(1), 95-108.
- Mimouni-Chaabane et al (2010). Perceived benefits of loyalty programs: Scale development and implications for relational strategies. *Journal of Business Research*, 63(1), 32-37

Dick et al (1994): “Customer loyalty: Toward an integrated conceptual framework”. Journal of the Academy of Marketing Science; vol. 22; nº2; págs. 99-113.

Trinquecoste, J. (1996): “Fidéliser le consommateur: Un objectif marketing prioritarie”. Decisions Marketing; nº 7; enero-abril.

Castañeda, J. (2005): La fidelidad en Internet, en Gutiérrez, A.M. y Sánchez, M.J. (Eds.), Marketing en Internet. Estrategia y Empresa. Pirámide Madrid.

Morissette et al (2013). The k-means clustering technique: General considerations and implementation in Mathematica 9(1): 15-24. Recuperado de <https://bit.ly/3x0GvO5>