



# Aplicación de técnicas de análisis multivariado para la segmentación de clientes en una entidad financiera colombiana

Alejandro Abril Acosta

Universidad del Valle  
Facultad de Ingeniería, Escuela de Estadística  
Santiago de Cali, Colombia  
2023



# Aplicación de técnicas de análisis multivariado para la segmentación de clientes en una entidad financiera colombiana

Alejandro Abril Acosta

Trabajo de grado presentado como requisito parcial para optar al título de:  
**Estadístico**

Director:  
Jaime Mosquera Restrepo

Codirectora:  
Claudia Johanna Polo Yepes

Universidad del Valle  
Facultad de Ingeniería, Escuela de Estadística  
Santiago de Cali, Colombia  
2023



## Dedicatoria

A mis queridos padres

Su apoyo y ejemplo son fundamentales para este nuevo paso. Agradezco a la vida por tenerlos en este momento y poder compartir con ustedes este logro. Sé que les debo mucho más y espero que se sientan orgullosos cada día más de su hijo.

A ti, Ana.



# Agradecimientos

Agradezco a la Universidad del Valle, a la Facultad de Ingeniería y a la Escuela de Estadística, así como a todos los profesores que forman parte de ella, por haber dejado una huella invaluable en mi formación. También, quiero expresar un especial reconocimiento a mi director, el profesor Jaime Mosquera, y a mi co-directora, Claudia Polo, su orientación, paciencia y motivación fueron esenciales en el desarrollo de este trabajo de grado.

Le agradezco a mis padres y hermanos, su amor ha sido mi mayor motivación, y esta etapa de mi vida no estaría completa sin su presencia. A ustedes les dedico este logro, con la certeza de que sin su amor y aliento, este camino habría sido mucho más difícil. Gracias por ser mi mayor fuente de inspiración y por acompañarme con tanto amor en cada paso que doy.



## Resumen

Desde una perspectiva financiera, social y transaccional, los titulares de tarjetas asociados a una entidad financiera colombiana son sorprendentemente diversos. Al analizar las transacciones desde diciembre de 2021 hasta diciembre de 2022, se encuentran diferencias notables en las tendencias y decisiones de consumo de estos consumidores. En este trabajo, se lleva a cabo una detallada segmentación de los tarjetahabientes utilizando información transaccional suministrada por la entidad financiera, identificando las diferencias de sus patrones de consumo y comportamiento financiero. Para realizar esta segmentación, se han seleccionado 2 grupos de variables: Sociales y Transaccionales, conformado cada grupo por variables numéricas y categóricas, las cuales representan las principales características de los 26.307 tarjetahabientes vinculados a la entidad financiera. La información de estos grupos de variables es resumida mediante un Análisis Factorial Múltiple (AFM). Con los resultados del AFM, se realizó un Análisis de Conglomerados (AC), con el propósito de clasificar a los tarjetahabientes en categorías basadas en semejanzas y particularidades, resultando en seis clases bien diferenciadas. Los grupos conformados tienen como objetivo fidelizar a los clientes actuales e incrementar el uso de las tarjetas entre ellos.

**Palabras Clave:** Análisis de Componentes Principales, Análisis Factorial Múltiple, Análisis Clúster. Segmentación Tarjetahabientes.

## Abstract

From a financial, social, and transactional perspective, the cardholders associated with a Colombian financial entity are surprisingly diverse. When analyzing transactions from December 2021 to December 2022, notable differences are found in the consumption trends and decisions of these consumers. In this study, a detailed segmentation of cardholders is carried out using transactional information provided by the financial entity, identifying the differences in their consumption patterns and financial behavior. To carry out this segmentation, 2 groups of variables have been selected: Social and Transactional, each group being comprised of numerical and categorical variables, which represent the main characteristics of the 26,307 cardholders linked to the financial entity. The information from these groups of variables is summarized through a Multiple Factorial Analysis (MFA). With the results of the MFA, a Cluster Analysis (CA) was performed, with the purpose of classifying the cardholders into categories based on similarities and particularities, resulting in six well-differentiated classes. The groups formed aim to retain current customers and increase the use of cards among them.

**Keywords:** Principal Component Analysis, Multiple Factor Analysis, Cluster Analysis. Cardholder segmentation



# Contenido

|                                                                  |             |
|------------------------------------------------------------------|-------------|
| <b>Agradecimientos</b>                                           | <b>VII</b>  |
| <b>Resumen</b>                                                   | <b>IX</b>   |
| <b>Lista de figuras</b>                                          | <b>XIII</b> |
| <b>Lista de tablas</b>                                           | <b>XIV</b>  |
| <b>1. Introducción y Planteamiento</b>                           | <b>2</b>    |
| 1.1. Introducción . . . . .                                      | 2           |
| 1.2. Planteamiento del Problema . . . . .                        | 4           |
| 1.3. Justificación . . . . .                                     | 6           |
| 1.4. Objetivos . . . . .                                         | 7           |
| 1.4.1. Objetivo General . . . . .                                | 7           |
| 1.4.2. Objetivos Específicos . . . . .                           | 7           |
| <b>2. Antecedentes</b>                                           | <b>8</b>    |
| 2.1. Antecedentes Nacionales . . . . .                           | 8           |
| 2.2. Antecedentes Internacionales . . . . .                      | 10          |
| <b>3. Marco Teórico</b>                                          | <b>14</b>   |
| 3.1. Marco Contextual . . . . .                                  | 14          |
| 3.2. Marco Teórico Estadístico . . . . .                         | 16          |
| 3.2.1. Análisis de Componentes Principales (ACP) . . . . .       | 16          |
| 3.2.2. Análisis Factorial Múltiple (AFM) . . . . .               | 22          |
| 3.2.3. Análisis de Conglomerados . . . . .                       | 28          |
| <b>4. Metodología</b>                                            | <b>36</b>   |
| 4.1. Población Objetivo . . . . .                                | 37          |
| 4.2. Extracción, Transformación y Carga (ETL) de Datos . . . . . | 38          |
| 4.2.1. Extracción . . . . .                                      | 39          |
| 4.2.2. Transformación . . . . .                                  | 39          |
| 4.2.3. Carga . . . . .                                           | 40          |
| 4.3. Análisis Descriptivo y Exploratorio . . . . .               | 41          |
| 4.4. Análisis Factorial Múltiple (AFM) . . . . .                 | 41          |

|                                                                                      |           |
|--------------------------------------------------------------------------------------|-----------|
| 4.5. Clusters . . . . .                                                              | 42        |
| 4.6. Validación e Interpretación de los Clúster . . . . .                            | 42        |
| <b>5. Resultado</b>                                                                  | <b>43</b> |
| 5.1. Análisis Exploratorio . . . . .                                                 | 43        |
| 5.2. Análisis Multivariado . . . . .                                                 | 48        |
| 5.2.1. Análisis Factorial Múltiple (AFM) . . . . .                                   | 48        |
| 5.2.2. Análisis de Conglomerados. . . . .                                            | 53        |
| 5.2.3. Validación de la Estabilidad de los Grupos . . . . .                          | 55        |
| 5.3. Caracterización de los Conglomerados y Descripción de Características . . . . . | 56        |
| <b>6. Conclusiones</b>                                                               | <b>64</b> |
| <b>7. Recomendaciones</b>                                                            | <b>66</b> |
| <b>A. Anexo: Tablas Caracterizacion Grupos</b>                                       | <b>67</b> |
| A.1. Variables transaccionales Formula . . . . .                                     | 67        |
| A.1.1. KPI'S año de cierre . . . . .                                                 | 68        |
| A.1.2. KPI'S mes de cierre . . . . .                                                 | 70        |
| <b>B. Anexo: Tablas Caracterizacion Grupos</b>                                       | <b>73</b> |
| B.1. Grupo 1 . . . . .                                                               | 73        |
| B.2. Grupo 2 . . . . .                                                               | 74        |
| B.3. Grupo 3 . . . . .                                                               | 75        |
| B.4. Grupo 4 . . . . .                                                               | 76        |
| B.5. Grupo 5 . . . . .                                                               | 77        |
| B.6. Grupo 6 . . . . .                                                               | 78        |
| <b>Bibliografía</b>                                                                  | <b>79</b> |

# Lista de Figuras

|                                                                                            |    |
|--------------------------------------------------------------------------------------------|----|
| 3-1. Ejemplo de Screeplot . . . . .                                                        | 18 |
| 3-2. Ejemplo la nube de individuos. . . . .                                                | 20 |
| 3-3. Ejemplo la nube de variables. . . . .                                                 | 21 |
| 3-4. Ejemplo de Biplot. . . . .                                                            | 22 |
| 3-5. Matriz X conformada por la yuxtaposición de las tablas . . . . .                      | 23 |
| 3-6. Ejemplo de representación de los grupos variables. . . . .                            | 25 |
| 3-7. Ejemplo de representación de los individuos. . . . .                                  | 26 |
| 3-8. Ejemplo de representación de los individuos parciales. . . . .                        | 27 |
| 3-9. Ejemplo de representación método del codo. . . . .                                    | 32 |
| 3-10.Ejemplo de representación coeficiente de silueta. . . . .                             | 34 |
| 4-1. Secuencia Metodológica . . . . .                                                      | 36 |
| 4-2. Diagrama de proceso para la obtención de datos. . . . .                               | 40 |
| 5-1. Gráfico de la matriz de correlación de las variables consolidadas . . . . .           | 47 |
| 5-2. Histograma de Valores Propios . . . . .                                               | 48 |
| 5-3. Representación de los Grupos Dimensión 1-2 . . . . .                                  | 49 |
| 5-4. Representación de los Grupos Dimensión 1-3 . . . . .                                  | 49 |
| 5-5. Círculo de correlaciones componentes 1 - 2 . . . . .                                  | 50 |
| 5-6. Círculo de correlaciones componentes 1 - 3 . . . . .                                  | 50 |
| 5-7. Representación contribución variables numéricas Dim1 . . . . .                        | 51 |
| 5-8. Representación contribución variables numéricas Dim2 . . . . .                        | 51 |
| 5-9. Representación contribución variables categóricas Dim1 . . . . .                      | 51 |
| 5-10.Representación contribución variables categóricas Dim2 . . . . .                      | 51 |
| 5-11.Relaciones entre Categorías Originales y Componentes 1 - 2 . . . . .                  | 52 |
| 5-12.Relaciones entre Categorías Originales y Componentes 1 - 3 . . . . .                  | 52 |
| 5-13.Grafico Elbow. . . . .                                                                | 53 |
| 5-14.Coficiente Silueta y Dendograma Via Ward . . . . .                                    | 54 |
| 5-15.Coficiente Silueta para los 6 grupos. . . . .                                         | 54 |
| 5-16.Proyección de la nube de puntos para los 6 grupos en los 3 primeros factores. . . . . | 55 |
| 5-17.Tablero Estadísticas Grupo 1. . . . .                                                 | 56 |
| 5-18.Tablero Estadísticas Grupo 2. . . . .                                                 | 57 |
| 5-19.Tablero Estadísticas Grupo 3. . . . .                                                 | 58 |
| 5-20.Tablero Estadísticas Grupo 4. . . . .                                                 | 59 |

|                                                    |    |
|----------------------------------------------------|----|
| <b>5-21.</b> Tablero Estadísticas Grupo 5. . . . . | 60 |
| <b>5-22.</b> Tablero Estadísticas Grupo 6. . . . . | 61 |

# Lista de Tablas

|                                                                               |    |
|-------------------------------------------------------------------------------|----|
| 2-1. Variables por antecedentes . . . . .                                     | 13 |
| 4-1. Recategorización de Variables . . . . .                                  | 40 |
| 5-1. Variables Sociodemográficas Numéricas . . . . .                          | 43 |
| 5-2. Variables Sociodemográficas Categóricas . . . . .                        | 44 |
| 5-3. Variables Transaccionales Tarjeta Numéricas . . . . .                    | 45 |
| 5-4. Variable Transaccional Tarjeta Categóricas . . . . .                     | 46 |
| 5-5. Valores Propios del AFM . . . . .                                        | 49 |
| 5-6. Indicador de estabilidad de los grupos. . . . .                          | 55 |
| A-1. Información de Individuos . . . . .                                      | 67 |
| A-2. Registro detallado de Transacciones . . . . .                            | 67 |
| A-3. Variables creadas con un año de Transacciones (Individuo -Año) . . . . . | 72 |
| B-1. Caracterización Categorías Grupo 1 . . . . .                             | 73 |
| B-2. Caracterización Numérica Grupo 1 . . . . .                               | 73 |
| B-3. Caracterización Categorías Cluster 2 . . . . .                           | 74 |
| B-4. Caracterización Numérica Grupo 2 . . . . .                               | 74 |
| B-5. Caracterización Categorías Grupo 3 . . . . .                             | 75 |
| B-6. Caracterización Numérica Grupo 3 . . . . .                               | 75 |
| B-7. Caracterización Categorías Grupo 4 . . . . .                             | 76 |
| B-8. Caracterización Numérica Grupo 4 . . . . .                               | 76 |
| B-9. Caracterización Categorías Grupo 5 . . . . .                             | 77 |
| B-10. Caracterización Numérica Grupo 5 . . . . .                              | 77 |
| B-11. Caracterización Categorías Grupo 6 . . . . .                            | 78 |
| B-12. Caracterización Numérica Grupo 6 . . . . .                              | 78 |

# 1. Introducción y Planteamiento

## 1.1. Introducción

Actualmente, la personalización y adaptación de productos y servicios se ha vuelto esencial debido a la creciente demanda de experiencias a medida por parte de los consumidores y/o clientes. Las empresas, reconociendo esta tendencia, están invirtiendo en comprender profundamente las necesidades y deseos de sus clientes. Esta estrategia no solo se orienta hacia una mayor satisfacción del cliente, sino también hacia su fidelización y la maximización del valor que aportan a la organización.

Las tarjetas de crédito se han consolidado como una herramienta fundamental en la dinámica económica de un país. Al ofrecer a los consumidores un acceso más ágil a la compra de bienes y servicios, no solo impulsan la demanda en diversos sectores del mercado, sino que también han llegado a integrarse en la planificación financiera de muchos hogares, convirtiéndose en un componente esencial de su presupuesto doméstico (Rodríguez, 2014).

Actualmente, las tarjetas de crédito se han consolidado como el principal medio de pago a crédito, lo que ha llevado a una reducción en el uso del dinero en efectivo. Además, en muchos casos, su uso viene acompañado de beneficios y privilegios para el consumidor. Por otro lado, las entidades bancarias, principales emisoras de estos instrumentos, proporcionan a los usuarios una herramienta que simplifica y agiliza sus procesos de compra.

Las entidades financieras enfrentan una mayor competencia para satisfacer y retener a sus usuarios. En un mercado con múltiples ofertas de tarjetas de crédito, el hecho de proporcionar acceso a una tarjeta de crédito no resulta suficiente. En este caso, factores diferenciadores como tarifas competitivas, programas de fidelización y descuentos especiales se vuelven cruciales. Estos elementos no solo determinan la elección del cliente al seleccionar una entidad financiera, sino que también influyen en la continuidad y la profundidad de su relación con el banco a largo plazo (Castro Puche and Díz Garcés, 2014).

En este contexto, la entidad financiera colombiana objeto de estudio, entre sus múltiples productos, ofrece el servicio de tarjetas de crédito, en el cual tiene 26.307 usuarios con una tarjeta de crédito de la franquicia MasterCard. Así, para esta entidad financiera, la gestión óptima de las relaciones con sus tarjetahabientes se convierte en una prioridad. El

objetivo de este trabajo radica en comprender las características y comportamientos de los clientes con base en su historial de transacciones y atributos sociodemográficos. Al hacerlo, no solo aseguran la fidelidad del cliente, sino que también potencian el uso y la valoración del producto ofrecido.

Para llevar a cabo la clasificación de tarjetahabientes según sus características, es fundamental emplear técnicas de clasificación multivariante. Una de las técnicas más adecuadas para este propósito es el Análisis Factorial Múltiple (AFM). Según Escofier and Pagès (1992), el AFM es una técnica estadística multivariante diseñada para analizar datos en los que se presentan múltiples grupos de variables. Es especialmente útil cuando se busca analizar diversos conjuntos de variables de manera simultánea. Para implementar el AFM, se recurre a la función MFA del paquete *Ade4* en la librería FactoMineR, siguiendo las orientaciones de Francois et al. (2013), utilizando el software estadístico **R**. Una vez obtenidos los resultados del AFM, se prosigue con el Análisis Clúster mediante la librería *FactomineR*. Este análisis permite determinar agrupaciones de tarjetahabientes con base en sus similitudes conforme a las variables estudiadas.

En conclusión, el profundo análisis de las características y comportamientos de los tarjetahabientes permite a la entidad crear campañas de marketing focalizadas. Estas campañas están diseñadas para resonar con las necesidades e intereses específicos de diferentes grupos de clientes. Este enfoque personalizado no solo alienta un uso más frecuente de la tarjeta de crédito, sino que también fortalece la relación a largo plazo con el cliente, contribuyendo así a su fidelización. De esta manera, la organización no solo maximiza el valor que cada cliente aporta, sino que también se posiciona como un referente en la entrega de soluciones financieras que son tanto personalizadas como competitivas.

## 1.2. Planteamiento del Problema

La entidad financiera cuenta con una segmentación de sus tarjetahabientes realizada partir de modelos simples RFM (Recency, Frequency, Monetary Value o en Recencia, Frecuencia y Valor monetario), esta técnica se basa en tres variables que evalúan el comportamiento de los clientes: Recencia (la última vez que un cliente realizó una compra), Frecuencia (cuán a menudo realiza compras) y Valor Monetario (el monto total gastado por el cliente) Du et al. (2010). Usando estos tres factores, los clientes se pueden segmentar de manera en diversos grupos, esta segmentación permitirá a la entidad financiera adaptar sus estrategias de manera específica para abordar las necesidades y preferencias de cada grupo.

La segmentación RFM brinda a las empresas herramientas valiosas para potenciar sus relaciones con los clientes. Permite un marketing personalizado al ajustar estrategias de comunicación según las necesidades específicas de cada segmento, garantizando interacciones más pertinentes y efectivas. Esta segmentación también refuerza la fidelización al identificar aquellos tarjetahabientes con compras esporádicas y diseñar campañas que fortalezcan su lealtad. Además, al enfocarse en tarjetahabientes de alto valor mediante ofertas y comunicados exclusivos, la entidad no solo incentiva un mayor gasto de estos usuarios, sino que también optimiza las ventas y eleva sus ingresos (Garitano, 2023)

Pese a sus beneficios, esta segmentación presenta ciertas limitaciones. Ofrece una perspectiva corta del comportamiento y las preferencias del cliente al no incorporar elementos, tales como sus características sociodemográficas y perfil transaccional, lo cual dificulta el diseño y la gestión de estrategias comerciales precisas y eficientes para la fidelización de estos clientes.

Ante las limitaciones identificadas por la entidad en la segmentación RFM, surge la necesidad de evaluar técnicas estadísticas avanzadas que prometen una segmentación más precisa de los clientes. Estas técnicas multivariantes son capaces de analizar múltiples variables de diferente naturaleza simultáneamente, estableciendo las interrelaciones entre ellas. Lo cual facilita la identificación de patrones de comportamiento distintivos en los clientes. Como resultado, se pueden formular estrategias de marketing mucho más afinadas, direccionadas a satisfacer las particularidades y necesidades de cada segmento identificado.

Para la entidad financiera en cuestión, el objetivo es realizar una segmentación a sus tarjetahabientes de la franquicia Mastercard. Considerando una amplia gama de variables sociodemográficas, incluyendo: la edad, ingresos, nivel educativo, estado civil, estrato, puntaje crediticio y la antigüedad en la entidad, además se incorporan también variables transaccionales tales como: plazos, número de transacciones, tiempo entre transacciones y categoría de la misma. Esta amplia variedad de variables contribuirá a una segmentación más refinada, permitiendo la creación de grupos que reflejen una complejidad de características

---

mucho más detallada. Bajo lo cual asegura la creación de estrategias de marketing y atención más específicas y adaptadas a las diferencias de cada grupo.

Para efectuar esta clasificación, debido a la magnitud del conjunto de datos (26,307 clientes y un total de 786,642 registros de transacciones en un año), se requiere un método de análisis robusto y eficiente que permita extraer patrones y segmentaciones representativas. Representa un desafío analítico y computacional para el proceso de segmentación. Se emplean métodos de clasificación multivariante. Se utiliza el Análisis Factorial Múltiple (AFM), es una técnica estadística multivariante utilizada para analizar conjuntos de datos en los que existen múltiples grupos de variables. Se utiliza cuando se quiere analizar simultáneamente varios conjuntos de variables.

### 1.3. Justificación

En el caso de la entidad financiera, la metodología actual de segmentación RFM, aunque ampliamente reconocida, ha mostrado limitaciones para establecer efectivas estrategias comerciales en el contexto de la entidad financiera. Cuatro grupos se generan bajo esta técnica de segmentación, pero su utilidad práctica se ve cuestionada por la falta de consistencia en las características de estos grupos. Un problema es la clasificación errónea de individuos con perfiles similares en diferentes conglomerados, lo que disminuye el propósito central de la segmentación: diseñar estrategias comerciales basadas en características comunes. Desde una perspectiva más técnica, los índices de consistencia de esta segmentación son insatisfactorios, con un coeficiente de Silhouette promedio inferior a 0.05. Esto indica una baja cohesión intra-grupo y amplias diferencias inter-grupos. Esta inconsistencia, combinada con la riqueza de datos disponibles sobre los clientes, en términos de variables sociodemográficas, resalta la necesidad de reevaluar y adaptar la metodología de segmentación en uso.

El bajo rendimiento de las campañas de marketing actuales, reflejado en bajos promedios de facturación y fidelización, es un claro indicativo de la inadecuación de la segmentación actual. Se requiere un enfoque más sofisticado y detallado, que pueda incorporar tanto las variables sociodemográficas como las transaccionales. Esta adaptación no solo permitirá un diseño más eficaz de las campañas de marketing, sino que también ofrecerá una comprensión más profunda de las preferencias y necesidades de los clientes, aumentando la calidad en la experiencia del cliente y mejorando los resultados comerciales.

Desde el punto de vista estadístico, es importante resaltar que la metodología RFM, al centrarse únicamente en la facturación, ignora otras dimensiones importantes en la comprensión del comportamiento del cliente, como sus características sociodemográficas. Esto resulta en una perspectiva limitada que desaprovecha la riqueza de datos disponibles. A diferencia del Análisis Factorial Múltiple (AFM), que permite incorporar una gama más amplia de variables tanto transaccionales como sociodemográficas. Esto no solo optimizará el diseño de campañas de marketing más focalizadas y eficaces, sino que también favorecerá una mayor comprensión de las complejidades inherentes al cliente, mejorando así la calidad de la experiencia del cliente y, a largo plazo, sus niveles de fidelización.

## 1.4. Objetivos

### 1.4.1. Objetivo General

Implementar un modelo para la segmentación de los tarjetahabientes de una entidad financiera en Colombia, a través de la identificación de patrones en dimensiones sociodemográficas y transaccionales, considerando variables cuantitativas y cualitativas para cada dimensión.

### 1.4.2. Objetivos Específicos

1. Implementar un proceso ETL (*Extract, Transform, Load*) para la extracción, consolidación y preprocesamiento de los datos disponibles sobre los clientes de una entidad financiera colombiana, considerando las dimensiones sociodemográficas y transaccionales, considerando variables tanto cuantitativas como cualitativas para cada dimensión.
2. Realizar una caracterización multivariante de los clientes a través de la implementación de la metodología AFM con el fin de identificar los principales factores que explican la variabilidad en los clientes.
3. Proponer una nueva segmentación de clientes basada en los resultados obtenidos en el AFM, utilizando técnicas complementarias de clustering.
4. Evaluar estadística y operativamente la pertinencia de los segmentos de clientes generados.

## 2. Antecedentes

La literatura sobre la segmentación de tarjetahabientes es amplia, pero este estudio ha identificado una notable carencia en investigaciones que integren técnicas multivariantes para analizar simultáneamente variables categóricas y numéricas en el ámbito de los perfiles de los tarjetahabientes, aunque existen investigaciones que se centran en el análisis descriptivo para estudiar estas variables. Este capítulo, por lo tanto, se centrará en esas investigaciones específicas, destacando las variables claves y las técnicas estadísticas utilizadas.

Para un análisis más enriquecedor y contextual, los antecedentes se desglosarán en dos grupos: investigaciones nacionales e internacionales. Este enfoque permitirá comparar el nivel de avance y las prácticas en segmentación de tarjetahabientes dentro del país con las prácticas globales. La importancia de esta comparación se acentúa por la notable carencia de información en el ámbito nacional, atribuida en parte a la limitada publicación o ejecución de estudios de este tipo en el país.

### 2.1. Antecedentes Nacionales

A continuación se presentará una serie de estudios sobre la segmentación de clientes haciendo uso de datos de entidades colombianas. Se enfatiza en la técnica empleada y las variables empleadas en el análisis.

**Castellanos (2004):** El estudio presentado se basó en una profunda segmentación de mercado utilizando una base de datos de 35,000 tarjetahabientes de un banco colombiano. Esta base de datos cuenta con una amplia variedad de variables, que comprenden desde datos demográficos (sexo, antigüedad, edad, ciudad) hasta detalladas métricas de comportamiento y transaccionales: estado de la tarjeta, rentabilidad, número de transacciones al año, volumen de gasto al año, lugar de transacción, saldo de cartera, número de avances, valor de los mismos, cuota de manejo, compras totales, valor de las mismas, intereses totales y porcentaje de utilización. El objetivo principal era trazar un mapa de rentabilidad para identificar la eficacia de cada segmento de cliente.

La técnica adoptada para segmentar los datos fue el Análisis de Conglomerados utilizando la herramienta **SPSS**. La segmentación se decidió finalmente en cinco conglomerados distintos,

basándose en la dispersión y facilidad de interpretación.

A continuación se resumen las conclusiones de cada conglomerado:

- Los Pasivos: Estos clientes son menos activos en comparación con el portafolio general. No muestran una fuerte preferencia por ninguna categoría de comercio y realizan menos transacciones anuales que el promedio general.
- Las Tradicionales: Este grupo se caracteriza por tarjetahabientes que son responsables en sus pagos y tienen un comportamiento de compra centrado en Almacenes con y sin supermercado.
- Los Valiosos: A pesar de ser el grupo más pequeño, son de gran importancia para el banco debido a su alta actividad y margen de utilidad.
- En Aprietos: Este grupo es interesante, ya que muestra un comportamiento que indica posibles problemas financieros. El hecho de que todos los miembros del segmento realicen avances sugiere una posible necesidad urgente de efectivo, lo que puede interpretarse como una señal de que están atravesando dificultades económicas.
- Los Comelones: Se distinguen por sus gastos en restaurantes. Este grupo, en su mayoría hombres, también muestra una diversidad en sus gastos, indicando una variedad de intereses y comportamientos.

Entre las conclusiones, el banco pudo adaptar y personalizar sus ofertas, campañas y comunicaciones para atender de manera eficiente a cada uno de estos segmentos. Para el segmento denominado “En Aprietos”, el banco realizó programas educativos sobre gestión financiera y proporcionó opciones de refinanciamiento. Por otro lado, para el segmento “Los Comelones”, se le dieron al banco recomendaciones de asociarse con restaurantes para ofrecer descuentos u ofertas especiales.

**Barbosa Ariza and González Jiménez (2018):** En este estudio, se buscan identificar los perfiles predominantes de tarjetahabientes de una entidad financiera en Bogotá, con el propósito de orientar estrategias de marketing que potencien la adquisición y el adecuado uso de sus productos financieros. Se llevó a cabo una investigación de carácter descriptivo-analítico con un enfoque mixto, utilizando una base de datos que incluye variables como edad, ingresos promedio, nivel educativo, estado civil, tipo de vivienda, límite de la tarjeta de crédito y monto adeudado.

A partir de esta base, que contenía 500 registros, se seleccionó una muestra representativa de 219 clientes de tarjetas de crédito. Para el tratamiento de las variables numéricas, se aplicó el método de componentes principales, lo que facilita la reducción de la dimensionalidad y

permite representar la mayor variabilidad en la información con menos variables, destacando patrones y correlaciones significativas entre las mismas. Por otro lado, para las variables categóricas se llevó a cabo un análisis descriptivo, determinando frecuencias y porcentajes, para entender la distribución y tendencias en estas variables.

Esta información permitió encontrar las características socioeconómicas de los tarjetahabientes, identificar los productos o servicios de mayor demanda y observar tendencias en relación con sus hábitos de pago. La combinación de componentes principales y análisis descriptivo brindó una visión completa y detallada sobre los perfiles asociados con un manejo óptimo de la tarjeta.

En conclusión, el artículo presentó resultados que señalan que los hábitos de consumo de los tarjetahabientes están fuertemente influenciados por sus ingresos. Se identificó un perfil predominante de individuos entre 30 y 59 años con educación universitaria. Además, las variables relacionadas con ingresos, límite de tarjetas de crédito y valor de la deuda muestran una fuerte asociación, lo que facilita la segmentación precisa de clientes para las actuales estrategias de marketing.

## 2.2. Antecedentes Internacionales

A continuación se presentará una serie de estudios sobre la segmentación de clientes haciendo uso de datos de entidades internacionales. Se enfatiza en la técnica empleada y las variables empleadas en el análisis.

**Li et al. (2010):** Los autores usan datos reales de un banco comercial chino, utilizaron el algoritmo de clustering  $K$ -means para clasificar a los titulares de tarjetas de crédito en cuatro tipos de clientes: clientes de alta calidad, clientes de alta calidad potenciales, clientes comunes y clientes desfavorables. Las variables usadas para la segmentación de los clientes fueron: Ingreso personal, Ingreso familiar mensual, Cantidad promedio mensual de uso de la tarjeta de crédito y calificación crediticia personal. Los autores indican que fue posible identificar características específicas de los clientes que estaban asociadas con un mayor potencial de generación de ganancias.

El estudio también utilizó diferentes métodos de minería de datos, como C5.0, redes neuronales, detector automático de interacciones chi-cuadrado, árboles de clasificación y regresión, para construir modelos de pronóstico basados en la información de los titulares de tarjetas de crédito. Estos modelos proporcionaron información útil para comprender las características de los diferentes clientes y para implementar estrategias de marketing dirigidas.

**Tripathi (2018):** En el artículo se destaca la importancia de la segmentación de clientes como una herramienta importante para las empresas, especialmente en la industria bancaria, señalan que dicha segmentación dependerá de la adecuada elección de las variables y técnicas usadas. En el desarrollo del artículo se emplean diversos modelos para segmentar los clientes utilizando técnicas de agrupación como son  $K$ -means y clustering Jerárquico. Los autores destacan que cada técnica de agrupamiento tiene sus propios méritos y limitaciones, siendo  $K$ -Means es más rápido computacionalmente y más adecuado para conjuntos de datos grandes, mientras que el Clustering Jerárquico ayuda en la visualización de datos y es más adecuado para conjuntos de datos pequeños. Adicional, mencionan la posibilidad de desarrollar una solución híbrida mediante la combinación de las dos técnicas, que tenga la capacidad de superar a los modelos individuales.

En el caso aplicado se usó 200 observaciones de clientes de un Centro Comercial a los cuales se les tiene 5 atributos denominados: Cliente, Género, Edad, Ingreso Anual y Puntaje de Gasto (las decide y calcula la empresa, en función de sus hábitos de consumo). En el caso del método de  $K$ -Means se necesita conocer previamente el número de clústeres, para lo cual, usaron el método del codo, mientras que para el método de clustering Jerárquico se empleó el dendograma y determinaron que eran necesarios 5 grupos. Los autores no presentan conclusiones propias del caso aplicado, sino que es usado para ejemplificar el uso de diferentes técnicas de agrupación.

**Umuhoza et al. (2020):** Se centran en la segmentación de clientes con tarjeta de crédito en África, utilizando técnicas de aprendizaje automático no supervisado. Los autores utilizaron datos de transacciones de tarjetas de crédito de una institución financiera en Ruanda, que comprenden un total de 2355 transacciones realizadas por 256 titulares de tarjetas de crédito. En cuanto a las variables utilizadas para la segmentación de clientes, los autores analizaron *la cantidad total de transacciones, la cantidad promedio gastada por transacción, el número de días entre transacciones y la cantidad total gastada*.

El estudio aplicó técnicas de aprendizaje no supervisado, específicamente la técnica de agrupamiento, utilizando el algoritmo  $K$ -means para identificar patrones de comportamiento de los titulares de tarjetas de crédito. Los resultados del estudio revelaron la existencia de 4 segmentos de clientes con diferentes patrones de comportamiento, se presentan sus perfiles y una estrategia de marketing para cada segmento.

En cuanto a las implicaciones prácticas del estudio, los autores destacaron la importancia de comprender los patrones de comportamiento de los titulares de tarjetas de crédito para desarrollar estrategias de marketing efectivas y mejorar la satisfacción del cliente. En general, este estudio demuestra la eficacia de las técnicas de aprendizaje automático no

supervisado para la segmentación de clientes con tarjeta de crédito en un contexto africano y cómo las variables utilizadas pueden ayudar a comprender mejor el comportamiento de uso de los titulares de tarjetas de crédito.

**Raj et al. (2023):** Los autores proponen el algoritmo *K*-means para realizar la segmentación de clientes basada en el análisis de datos de tarjetas de crédito. Lograron identificar cuatro tipos de clientes, permitiendo ofrecer diferentes experiencias a cada grupo de clientes y desarrollar estrategias más efectivas para aumentar los beneficios. En la segmentación de clientes se utilizaron diferentes variables. Algunas de las variables incluidas en la segmentación fueron: Frecuencia de compras, Frecuencia de pagos de cuotas, Límite de crédito y Frecuencia de avances en efectivo.

Los autores mencionan que esta investigación ilustra cómo se puede confiar en los algoritmos de agrupamiento empleados por las técnicas de minería de datos para proporcionar información valiosa sobre el comportamiento y las características del consumidor. Se puede decir que los ingresos, la educación y la disciplina fiscal indican el estatus socioeconómico de los consumidores, que está estrechamente relacionado con su solvencia crediticia.

La existencia de múltiples estudios que utilizan variables sociales y transaccionales para crear segmentos de tarjetahabientes respalda la idea de que este enfoque es ideal para entender y atender de manera efectiva las necesidades y comportamientos de diferentes grupos de clientes. Estas investigaciones demuestran que la segmentación basada en datos tanto sociales como transaccionales permite a las instituciones financieras ofrecer productos y servicios más personalizados, mejorando así la satisfacción del cliente y la rentabilidad del negocio.

Finalmente, se presenta una tabla de resumen que evidencia las distintas variables usadas en la segmentación de clientes:

| <i><b>Autores</b></i>       | <i><b>Variables incluidas</b></i>                                                                                                                                                                                                                                                                                 |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Castellanos (2004)          | Sexo, Antigüedad, Edad, Ciudad, Estado de la tarjeta, Rentabilidad, Número de transacciones al año, Volumen de gasto al año, Lugar de transacción, Saldo de cartera, Número de avances, Valor de los mismos, Cuota de manejo, Compras totales, Valor de las mismas, Intereses totales y Porcentaje de utilización |
| Li et al. (2010)            | Ingreso personal, Ingreso familiar mensual, Cantidad promedio mensual de uso de la tarjeta de crédito y calificación crediticia personal                                                                                                                                                                          |
| Barbosa and Gonzalez (2018) | Edad, Ingresos promedio, Nivel educativo, Estado civil, Tipo de vivienda, Límite de la tarjeta de crédito y Monto adeudado                                                                                                                                                                                        |
| Tripathi (2018)             | Género, Edad, Ingreso Anual y Puntaje de Gasto                                                                                                                                                                                                                                                                    |
| Umuhoza et al. (2020)       | La cantidad total de transacciones, número de días entre transacciones, cantidad total gastada y promedio gastado por transacción                                                                                                                                                                                 |
| Raj et al. (2023)           | Frecuencia de compras, Frecuencia de pagos de cuotas, Límite de crédito y Frecuencia de avances en efectivo                                                                                                                                                                                                       |

**Tabla 2-1.:** Variables por antecedentes

## 3. Marco Teórico

En este Capítulo. Se presenta el Marco Teórico sobre el cual se soporta el desarrollo del proyecto. Este Marco Teórico se divide en dos secciones. En la primera sección, se exploran los conceptos esenciales que permiten una comprensión adecuada de los términos relacionados con las entidades financieras. La segunda sección se concentra en la explicación de las técnicas estadísticas multivariadas empleadas en el desarrollo del proyecto.

### 3.1. Marco Contextual

#### **Tarjeta de Crédito**

Es un documento de material plástico o metal emitido por un banco o institución especializada a nombre de una persona, que podrá utilizarla para efectuar compras sin tener que pagar en efectivo y pudiendo, además, llevar el pago de los productos a períodos futuros. Esta tarjeta está respaldada y regulada por una franquicia, que es una marca o red global (como Visa, MasterCard, American Express, etc.) que establece y regula las normas de transacción para las tarjetas emitidas bajo su marca. Estas franquicias proporcionan la infraestructura y tecnología necesaria para procesar y autorizar transacciones con tarjetas de crédito de manera segura y eficiente. (Gil, 2016).

#### **Tarjetahabiente**

Es el nombre que recibe la persona que posee una tarjeta de crédito o débito. Con dicha tarjeta, la persona tiene acceso a realizar compras tanto en tiendas electrónicas en internet como en comercios presenciales que posean una máquina POS que captura el pago (Continuum, 2018).

#### **Calificación crediticia (Puntaje Acierta)**

De acuerdo con Definición del Diccionario de Cambridge, la calificación o puntajes crediticia es un número que se basa en la información de un informe de crédito (un registro de la deuda de alguien y el pago de la deuda), que las organizaciones financieras consideran al tomar decisiones sobre préstamos de dinero. El puntaje se calcula en función del historial de

crédito de una persona, que incluye la cantidad de cuentas abiertas que tiene, su historial de pago y la cantidad de deuda que tiene.

La calificación crediticia es un método estadístico utilizado para predecir la probabilidad de que un solicitante de préstamo, un prestatario existente o una contraparte incumplan o se vuelvan morosos. Proporciona una estimación de la probabilidad de incumplimiento o morosidad, que se utiliza ampliamente para préstamos de consumo, tarjetas de crédito y préstamos hipotecarios (Investopedia, 2023).

Datacrédito es un operador de información que es la entidad encargada de recopilar, administrar y transmitir datos sobre el comportamiento crediticio de las personas. El puntaje en Datacrédito va entre los 150 y los 950 puntos, eso quiere decir que entre más alto el puntaje, menos riesgo presenta para las entidades financieras. Si tienes más de 600 o estás cerca de los 950 tu puntaje es bueno y no representas un riesgo para las entidades, por lo que les será más fácil concederte los créditos que solicites (R5, 2023).

## **Merchant Category Code**

El Merchant Category Code o Código de Categoría Mercante (MCC por sus siglas en inglés) es un código numérico de cuatro dígitos utilizado en la industria de pagos electrónicos para clasificar la actividad principal de un comerciante. El MCC se asigna a cada establecimiento comercial y proporciona información sobre el tipo de productos o servicios que ofrece el comerciante. Estos códigos se utilizan principalmente en transacciones con tarjetas de crédito y débito para facilitar el procesamiento y el análisis de datos relacionados con las transacciones comerciales (Carter, 2023).

## **Avance**

Es el tipo de anticipo de efectivo más popular, es pedir prestado en una línea de crédito a través de una tarjeta de crédito. El dinero se puede retirar en un cajero automático o, dependiendo de la compañía de la tarjeta de crédito, de un cheque que se deposita o cobra en un banco. Los adelantos en efectivo con tarjeta de crédito suelen tener una tasa de interés más alta que la tasa de las compras regulares (Kagan, 2023).

## 3.2. Marco Teórico Estadístico

El análisis multivariado, subcampo del análisis estadístico, se ocupa de la exploración conjunta de dos o más atributos (variables) registrados en un grupo de elementos (personas, objetos o entidades) (Kachigan, 1982). Aunque se apoyan en diversas técnicas, en este trabajo se centrarán en dos de ellas: el Análisis de Componentes Principales (ACP), el Análisis Factorial Múltiple (AFM) y el Análisis de Clustering.

### 3.2.1. Análisis de Componentes Principales (ACP)

El Análisis de Componentes Principales (ACP) es una herramienta de la rama de la estadística multivariada diseñada para reducir la dimensión de un conjunto de datos con múltiples variables,  $X_{n \times p}$ . Donde  $n$  es la cantidad de individuos, representando las filas como  $x_1, \dots, x_n$  y  $p$  es el número de variables, las cuales representan las columnas de la matriz de datos, denotadas como  $x^1, \dots, x^p$ . Su objetivo es sintetizar la información manteniendo la mayor cantidad de la variabilidad original y, al mismo tiempo, reduciendo el número de dimensiones  $q$  ( $q < p$ ). Para ello, el ACP utiliza una combinación lineal de las variables originales para crear nuevos componentes principales o factores, los cuales son independientes entre sí. De esta manera, se puede obtener una visión más clara y simplificada de los datos, lo que facilita su interpretación y análisis (Gurrea, 2000).

Existen dos opciones de matriz de insumo para el cálculo del ACP: la matriz de correlación, que se utiliza cuando todas las variables tienen diferentes escalas de medida y se desean tratar con la misma importancia, y la matriz de covarianzas, que se utiliza cuando todas las variables tienen la misma escala de medida y se busca analizar la variabilidad de las variables. Los requerimientos previos para la aplicación del ACP son que las variables sean continuas, lineales, que el número de individuos sea mayor que el número de variables y que la base de datos sea completa, sin presencia de datos faltantes (Muñoz, 2018).

La matriz de datos  $X$ , luego de ser estandarizada, se conoce como  $Z$ . La matriz de datos a diagonalizar es  $Z'NZ$ , la cual es la matriz de correlación que tiene como características que es simétrica y positiva. Por lo tanto, se busca maximizar la correlación de los nuevos ejes con las variables originales.  $N_{n \times n}$  es la matriz de pesos de los individuos. La maximización de la forma cuadrática  $u'Z'NZu$  se realiza mediante el cálculo del lagrangiano  $L(u)$ , que garantiza un valor máximo para  $\lambda$ .

$$L(u) = u'Z'NZu - \lambda(u'u - 1)$$

$$L(u) = u'Z'NZu - \lambda u'u + \lambda$$

$$\frac{\partial L}{\partial u} = 2Z'NZu - 2\lambda u$$

$$\frac{\partial L}{\partial u} = 2(Z'NZu - \lambda u)$$

Al igualar a cero, se obtienen los valores y vectores propios, donde  $Z'NZ = \text{cor}(x_j, x_{j'})$

$$Z'NZ = \lambda u$$

La anterior es un sistema de valores y vectores propios, el cual garantiza que las componentes principales  $\psi = Zu$  sean ortogonales y sus valores propios  $\lambda$  sean decrecientes. Es importante mencionar que la matriz  $Z'NZ = \text{Cov}(X_i, X_j)$  y es una matriz simétrica.

### Indicadores para las interpretación del ACP

Las siguientes definiciones fueron tomadas de Muñoz (2018):

- **Contribuciones de los individuos:** Se define la contribución de un individuo en un eje de inercia como la parte de inercia total a lo largo del eje que se debe al individuo.

$$CTR(i, \alpha) = \frac{\psi_{i,\alpha}^2}{n\lambda_\alpha}$$

Donde  $\alpha = 1, 2, \dots, p$  representa el número de la componente principal. La contribución de un individuo a la inercia de un eje mide la importancia de dicho individuo en el eje.

- **Contribución de las variables:** La contribución de una variable en un determinado eje de inercia es una medida de la importancia de la variable en dicho eje y, en concreto, determina que parte de la inercia total que existe a lo largo del eje se debe a la variable.

$$CTR(j, \alpha) = \frac{\varphi_{j\alpha}^2}{\lambda_\alpha} = u_{j\alpha}^2$$

- La calidad de la representación presenta una idea acerca de qué tan cercano se encuentra un punto a un componente principal. Su cálculo se realiza mediante los cosenos cuadrados entre cada punto y el componente principal de la siguiente forma:

#### Individuos

$$\cos^2(i, \alpha) = \frac{\psi_{i\alpha}^2}{d^2(i, G)}$$

#### Variables

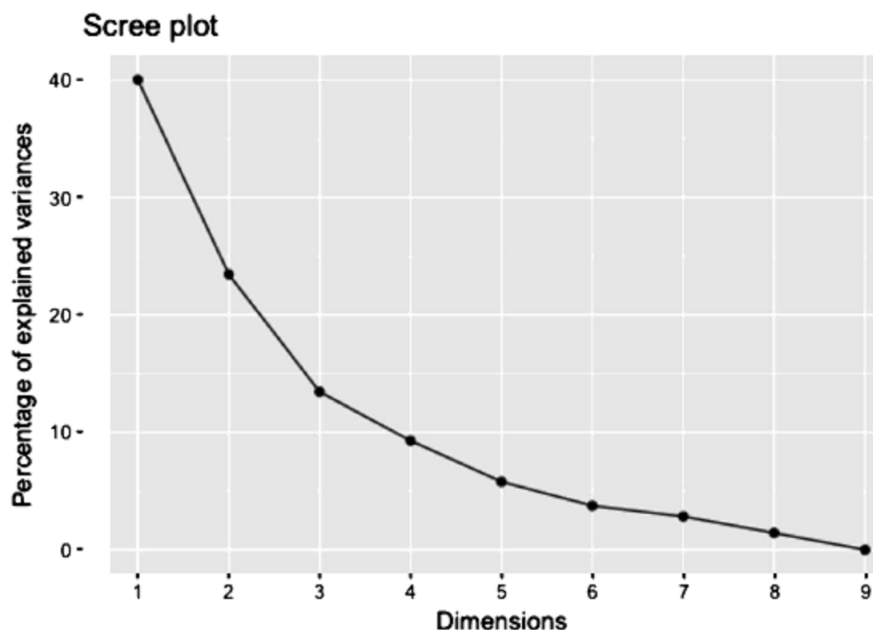
$$\cos^2(j, \alpha) = \varphi_{j\alpha}^2 = \text{cor}^2(j, \alpha)$$

### Número de componentes principales a extraer

Las siguientes definiciones fueron tomadas de Aldás Manzano and Uriel Jiménez (2017).

En este punto es importante mencionar que la decisión a tomar es cuanta información (varianza) estamos dispuestos a sacrificar para ganar la facilidad de interpretación. Algunas reglas para orientar esta decisión pueden ser:

- **Criterio de valores propios superior a la unidad:** Cuando las variables están estandarizadas  $Z$  el criterio es retener aquellas componentes cuyo valor propio sea superior a la unidad. Recordemos que el valor propio es la varianza extraída por cada componente. Parece lógico que esta varianza sea, como mínimo, la correspondiente a la de una variable. Si una componente no es capaz de explicar más información que una variable, no va a facilitar la reducción de dimensión, es decir, facilitar la interpretabilidad de la información.
- **Criterio del gráfico de sedimentación (Screeplot):** Este es un gráfico que representa los valores propios correspondientes a las distintas componentes. El criterio propone en retener aquellas componentes anteriores a donde la línea comienza a nivelarse.



**Figura 3-1.:** Ejemplo de Screeplot  
fuente *Aldás Manzano and Uriel Jiménez (2017)*

### Análisis de la nube de los individuos

En  $R^p$  con datos originales, la distancia euclidiana entre parejas de individuos está dada por  $d_2(i, l) = \sum_j (x_{ij} - x_{lj})^2$ . Dos puntos individuos están cerca si son caracterizados o toman valores casi iguales en cada una de las variables. La idea del ACP es representar la nube de puntos en un espacio de dimensión reducida de un modo óptimo. Es decir, deformando lo menos posible las distancias entre individuos.

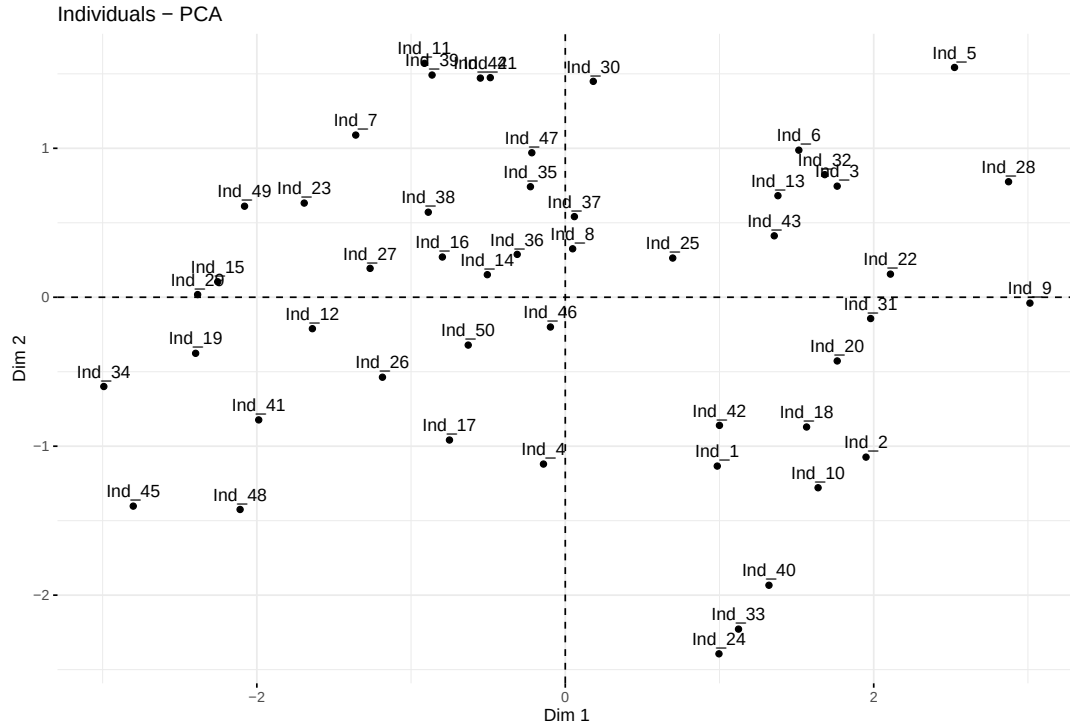
Para lograr la mejor representación, se busca inicialmente el subespacio en  $R^p$  de dimensión uno, un vector  $u$  en el eje  $\alpha$ , sobre el que se maximice la suma de los cuadrados de las distancias entre todas las parejas de puntos. De acuerdo con Lebart et al. (1994), encontrar el máximo de la suma de cuadrados de las distancias entre todas las parejas de individuos es lo mismo que maximizar la suma de cuadrados de las distancias entre cada punto y el centro de gravedad de la nube.

### Coordenadas de los individuos

Las coordenadas de los  $n$  individuos proyectados sobre el eje “canónico”  $u_\alpha$  son los componentes del vector (factor)  $\psi_\alpha = Zu_\alpha$ , el cual es una combinación lineal de las variables iniciales.

$$\begin{aligned}\psi_1 &= u_{11}Z_1 + u_{12}Z_2 + \dots + u_{1p}Z_p && \text{(Primera Componente)} \\ \psi_2 &= u_{21}Z_1 + u_{22}Z_2 + \dots + u_{2p}Z_p && \text{(Segunda Componente)} \\ &\vdots \\ \psi_p &= u_{p1}Z_1 + u_{p2}Z_2 + \dots + u_{pp}Z_p && \text{(P-ésima Componente)}\end{aligned}$$

Donde cada elemento  $u_{ij}$  es la ponderación (peso) que tiene la variable  $Z_j$  sobre la componente  $i$ . Debido a que la nube de individuos está centrada sobre el centro de gravedad, la media del factor es nula, esto es  $E(\psi_\alpha) = 0$  y la varianza es  $V(\psi_\alpha) = \psi'_\alpha N \psi_\alpha = \lambda_\alpha$ , lo cual implica que cada componente explica una proporción de la variabilidad total.



**Figura 3-2.:** Ejemplo la nube de individuos.  
Elaboración propia

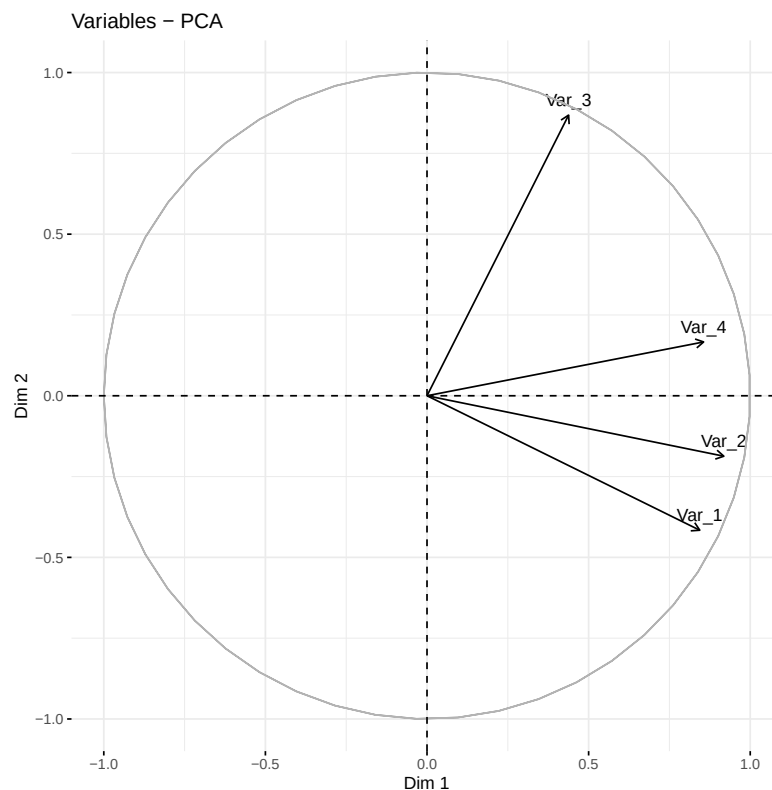
### Análisis de la nube de variables (Estandarizadas)

Consideremos ahora la tabla de datos a través de variables. Una variable es una columna de la tabla, un conjunto de  $n$  valores numéricos, asimilable a un vector de dimensión  $n$  evolucionando en un espacio vectorial de  $n$  dimensiones denotado  $R^n$  (llamado espacio de variables). El conjunto de estos vectores constituye la nube de variables. El primer plano de ajuste cortará la hipersfera, siguiendo círculos al interior de los cuales se encuentran proyectados los puntos variables y su estructura de correlaciones; los puntos mejor representados son los más próximos al círculo de correlaciones.

### Coordenadas de las variables

Las relaciones de transición permiten relacionar el espacio de los individuos con el de las variables. A través de ellas se obtiene que las coordenadas de las  $p$  variables están dadas por

$$\varphi_{\alpha} = \sqrt{\lambda_{\alpha}} u_{\alpha}.$$



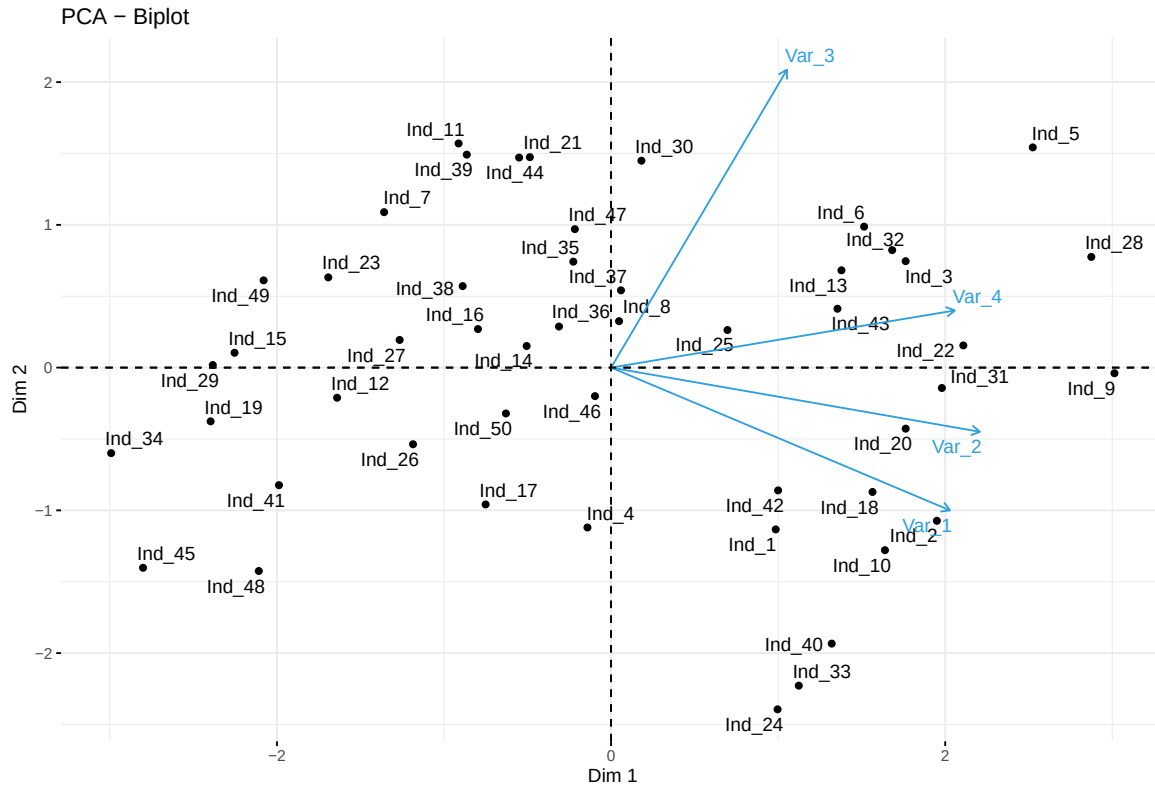
**Figura 3-3.:** Ejemplo la nube de variables.  
Elaboración propia

### Representación simultánea

En el ACP las dos representaciones (Individuos y Variables) están definidas en espacio distinto, con orígenes distintos y bases distintas. Para la nube de individuos, el origen es el centro de gravedad de los individuos, que es de dimensión  $p$ . Para la nube de variables, el origen de coordenadas es de dimensión  $n$ . Los puntos fila y los puntos columna están, pues, en espacios distintos, por lo que es imposible poderlos visualizar en un solo espacio que respete las proximidades internas de las dos nubes. Sin embargo, si se considera la dirección de las variables en  $R_p$  es posible realizar la representación simultanea en éste espacio. Las relaciones entre los objetos (individuos) y las variables pueden apreciarse en gráficos **biplots** (Muñoz, 2018).

- Los antiguos ejes unitarios  $(e_1, e_2, \dots, e_j, \dots, e_p)$  correspondientes a las  $p$  variables al comienzo del análisis son  $e_j = (0, 0, \dots, 1, 0, \dots, 0)$ ,  $j = 1, \dots, p$ , y constituyen el sistema de ejes de referencia para las coordenadas iniciales de los individuos.
- Los nuevos ejes unitarios  $u_\alpha$  ( $\alpha = 1, \dots, p$ ) constituyen los ejes factoriales.

Es posible entonces comparar las posiciones relativas de dos individuos respecto al conjunto de variables, o dos variables con respecto al conjunto de los individuos.



**Figura 3-4.:** Ejemplo de Biplot.  
Elaboración propia

Un biplot se usa para proveer una representación bidimensional de una matriz de datos  $X$ . Se emplean únicamente dos dimensiones para hacer más fácil el gráfico (Díaz-Monroy, 2002).

### 3.2.2. Análisis Factorial Múltiple (AFM)

Con frecuencia se tienen datos tabulares para los cuales se requiere que un conjunto de individuos sea caracterizado a través de varios grupos de variables. En el marco del análisis factorial, tener en cuenta diferentes grupos de variables lleva al problema de balancear los diferentes grupos. Para que este problema sea solucionado, más allá de los resultados del análisis factorial, es necesario contar con herramientas para comparar la estructura inducida por los diferentes grupos de variables sobre los individuos. Este es el objetivo del Análisis Factorial Múltiple (AFM) para dichos datos tabulares (Pagès, 2004).

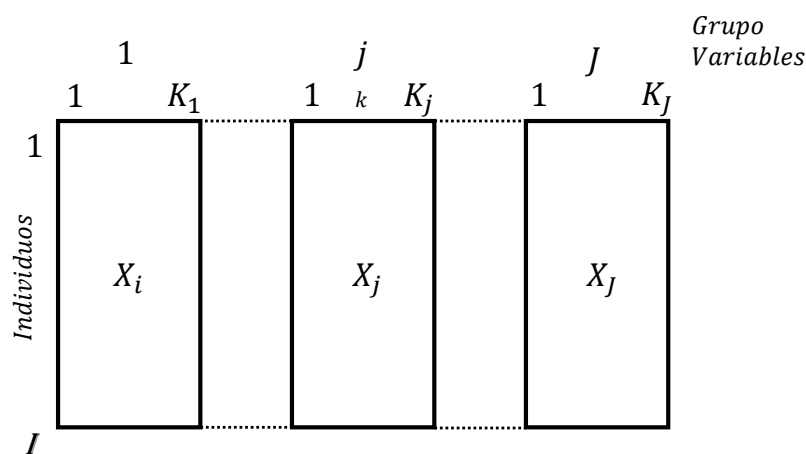
El AFM, definido por Escofier and Pagès (1992), es una técnica factorial empleada a partir de tablas de datos que describen un conjunto de individuos a través de grupos de variables,

las cuales pueden ser de naturaleza cuantitativa o cualitativo, con la restricción de ser de la misma naturaleza (Becerra, 2010).

El AFM tiene en cuenta:

- $R^K$ : Nube de individuos
- $R^I$ : Nube de variables
- $R^{I^2}$ : Nube de grupos de variables

Un ejemplo de un arreglo de datos en múltiples subgrupos se presenta en la siguiente imagen:



**Figura 3-5.:** Matriz X conformada por la yuxtaposición de las tablas  
fuente Escofier and Pagès (1992)

En la Figura 3-5 las matrices  $X_1, \dots, X_j, \dots, X_J$  conforman subgrupos de variables, cada uno conformado por distinto número y tipo de variables. A continuación, se expresa la notación de Escofier and Pagès (1992):

- $X$  : Tabla completa de datos.
- $X_j$  : Tabla del grupo  $j$ .
- $I$  : Conjunto de individuos.
- $K$  : Conjunto de variables de todos los grupos.
- $J$  : Conjunto de subtablas.
- $K_j$  : Conjunto de variables del  $j$  – esimo grupo.

El AFM puede considerarse como la generalización del ACP en el que la influencia de los grupos de variables está equilibrada. El objetivo general del análisis es explorar la existencia de estructuras comunes a todas, o aparte, de las nubes parciales. El AFM usa la metodología del ACP y lo hace en dos pasos:

1. **Análisis Separado:** Realizar un ACP para cada uno de los  $X_j$  tablas de variables cuantitativas o un ACM a las variables cualitativas. En el método AFM, el tratamiento para los conjuntos de variables nominales y ordinales es el mismo. En esta etapa, se retiene en cada uno de los análisis el mayor valor propio (llamado primer valor propio) denotado por  $\lambda_{j1}$  (Escofier and Pagès, 1992).
2. **Análisis Global:** Se realiza un ACP de la tabla global, en la cual se asigna una ponderación a cada grupo de tablas  $X_j$ . Lo que busca esto es encontrar el equilibrio entre las tablas. Esta ponderación equilibrada se obtiene calculando el inverso de cada valor propio de cada una de las tablas. Esta ponderación mantiene la estructura de cada tabla, ya que la inercia máxima en cada una de las nubes de individuos es igual a uno en cualquier dirección (Abascal Fernández and Landaluce Calvo, 2002).

### Ponderación de los grupos de las variables

Inicialmente, se debe obtener el peso de cada variable del grupo  $j$ , por lo cual se divide por  $\lambda_{j1}$  (Denotando por  $\lambda_{j1}$  primer valor propio del ACP o ACM separado del grupo  $j$ ). Esta ponderación se realiza con el objetivo de equilibrar el papel de los grupos en todos los aspectos del análisis, ya que ninguno determina por sí solo la primera componente principal del análisis (Escofier and Pagès, 1992).

Posteriormente, se realiza un ACP de la tabla general (estandarizada y ponderada), se obtienen los valores y vectores propios que dan lugar al análisis general en el que la inercia de cada factor está conformada proporcionalmente por las variables activas de cada grupo. Estas inercias se utilizan como coordenadas para graficar una “interestructura” que pone de manifiesto el tipo de los grupos (Gonzalez et al., 2009).

### Representación de los grupos de variables

Uno de los objetivos principales al estudiar varios grupos de variables es realizar una comparación global entre estos grupos. En este contexto, se introduce un espacio  $R^I$ , que representa una transformación del plano original. En este espacio transformado, los puntos correspondientes se alinean con un punto fijo y mantienen una razón constante en sus distancias a dicho punto. Este espacio  $R^I$  no solo sirve como una base para la representación que puede facilitar la interpretación del ACP de la tabla completa  $X$ , sino que también tiene su propia optimización.

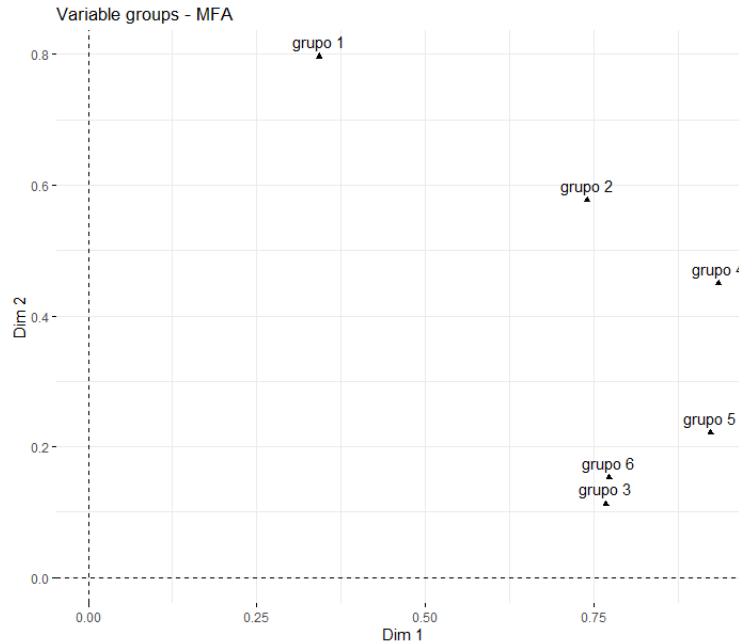
La nube  $NJ$  se conforma por las distancias entre los grupos de variables, las cuales se derivan del producto escalar  $L_g = \langle W_j D, W_l D \rangle = \text{traza}(W_j D W_l D)$ , donde  $W_j = X_j M_j X_j'$  y  $D$  es la matriz diagonal de los pesos de los individuos.

El producto escalar se interpreta como una medida de relación entre los grupos. A partir de este producto escalar, se construye el coeficiente de correlación vectorial  $RV$ .

$$RV = \frac{\langle W_j D, W_l D \rangle}{\|W_j D\| \|W_l D\|} = \frac{\text{traza}(W_j D W_l D)}{\|W_j D\| \|W_l D\|}$$

La norma de  $W_j D$  se expresa como  $\|W_j D\| = \sqrt{\sum_s \lambda_s^j}$ , donde  $\lambda_s^j$  es el  $j$ -ésimo valor propio del análisis de la tabla  $j$ .

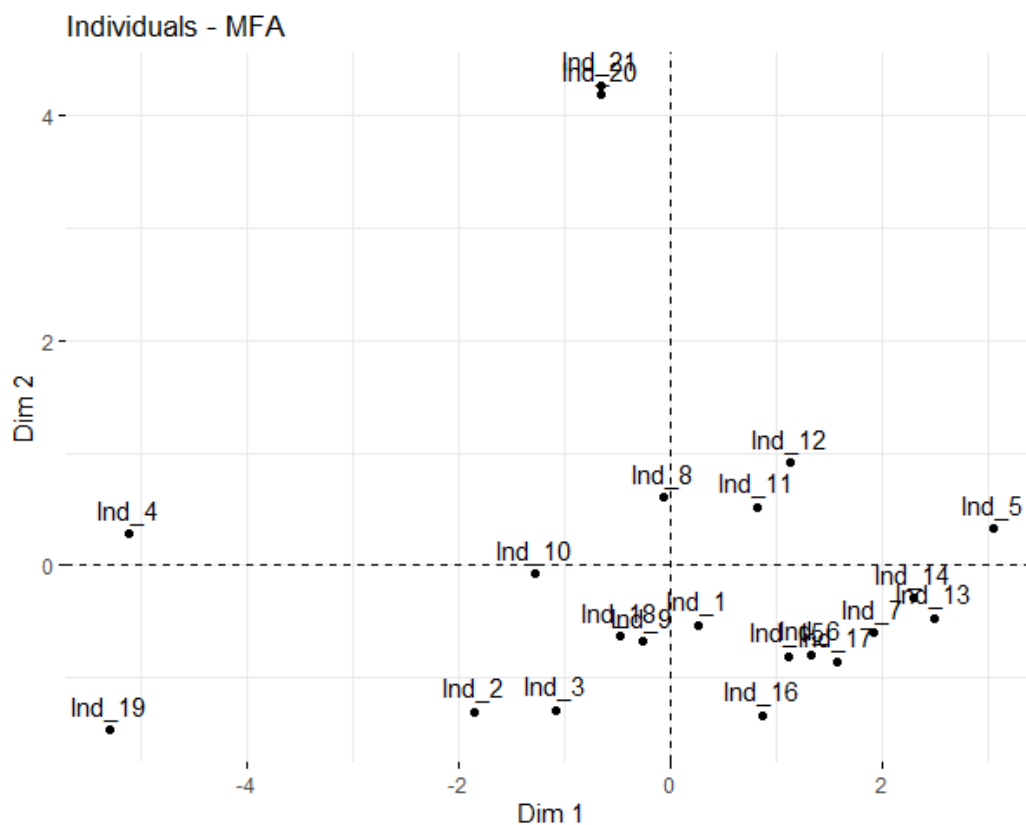
La Figura 3-6 representa un ejemplo con datos de *wine* sobre la representación gráfica de los grupos de variables. Es importante mencionar que los valores altos de las coordenadas en los grupos indican que el factor está asociado a las variables de los grupos. Por el contrario, valores bajos de las coordenadas indican poca relación del factor con las variables de los grupos.



**Figura 3-6.:** Ejemplo de representación de los grupos variables.  
Elaboración propia

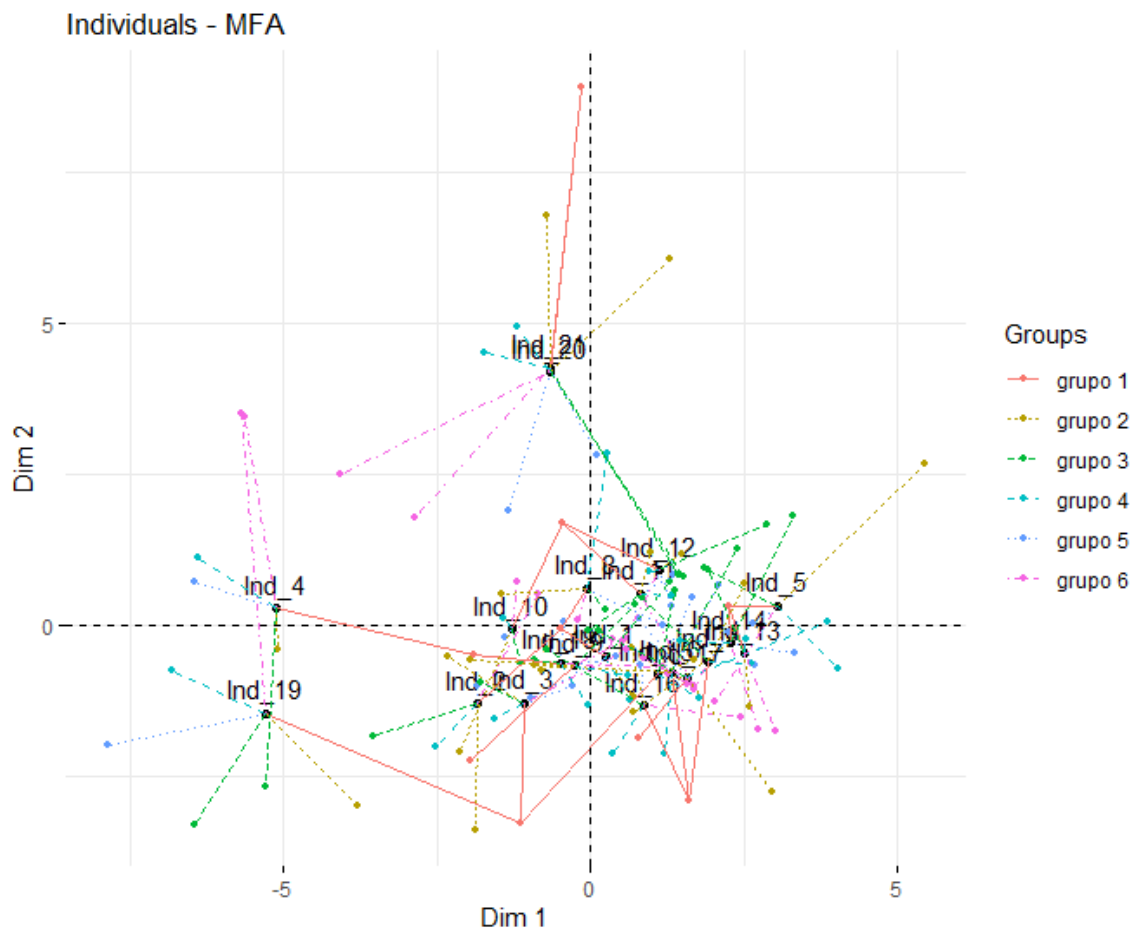
### Representación de los individuos

Las  $J$  nubes de individuos, llamadas “nubes parciales”, se proyectan como ilustrativas sobre los ejes. La Figura 3-7 es una representación gráfica ficticia que ilustra como se verían los individuos en el primer plano factorial.



**Figura 3-7.:** Ejemplo de representación de los individuos.  
Elaboración propia

Un individuo considerado desde el punto de vista de un solo grupo se llama individuo parcial. En la Figura 3-8, para un individuo dado, el punto corresponde al individuo medio o al centro de gravedad de los puntos parciales del individuo. Es decir, el individuo visto por todos los grupos de variables. Para un individuo dado, hay tantos puntos parciales como grupos de variables. La gráfica de individuos parciales representa cada vino visto por cada grupo y su baricentro.



**Figura 3-8.:** Ejemplo de representación de los individuos parciales.  
Elaboración propia

### Métricas para interpretación en el AFM

Al igual que en el ACP, para identificar el concepto subyacente que representan los ejes factoriales del AFM y comprender las relaciones entre individuos, variables y grupos, se calcula la contribución de cada uno de estos objetivos de análisis a los ejes factoriales.

#### ■ Contribución de los individuos a una dimensión

$$ctr_{i,s} = \frac{m_i (F_s^J)^2}{\lambda_s}$$

Donde  $m_i$  es el peso del  $i$ -ésimo individuo y  $F_s^J$  la coordenada del  $i$ -ésimo individuo en el  $s$ -ésimo eje factorial.

- **Contribución de las variables a una dimensión** Al igual que con los individuos, se puede encontrar la importancia de las variables para una dimensión dada.

Formalmente, la contribución de la  $k$ -ésima variable a la  $s$ -ésima dimensión está dada por:

$$ctr_{k,s} = \frac{\varphi_s^2}{\lambda_s^1}$$

- **Contribución de un grupo de variables a una dimensión** Está definida como la sumatoria de las contribuciones del  $j$ -ésimo grupo de variables en el  $s$ -ésimo eje factorial. Formalmente:

$$ctr_{j,s} = \sum_{k=1}^{K_j} ctr_{k,s}$$

Cabe resaltar que la contribución toma valores entre 0 y 1. Para un eje factorial dado, la suma de las contribuciones de todos los individuos (o variables) es igual a 1. Individuos (o variables) con altas contribuciones y con coordenadas factoriales con diferentes signos pueden entonces contrastarse para ayudar a interpretar el componente (Abdi et al., 2013).

### Selección del número de ejes factoriales

Como se mencionó anteriormente, el AFM se puede considerar una generalización del ACP. Por lo tanto, la selección del número de ejes factoriales a retener puede seguir los mismos criterios que se utilizan en el ACP, tales como la varianza explicada o la elección de valores propios superiores a 1.

### 3.2.3. Análisis de Conglomerados

El análisis de conglomerados, también conocido en algunos contextos como 'clustering', tiene como objetivo común la formación de clases o grupos de individuos o de variables que presenten características similares entre sí. Esta técnica se basa en la premisa de que es posible segmentar un conjunto de datos de manera que, dentro de cada grupo o clase, los elementos sean lo más semejantes posible entre sí (variabilidad intraclass baja), mientras que entre diferentes grupos, la disparidad sea máxima (alta variabilidad entre clases) (Jiménez, 2004).

### Medidas de Similitud

Las medidas de similitud son métricas utilizadas para evaluar el grado de similaridad entre dos conjuntos de datos. Estas medidas permiten comparar características, propiedades o atributos. Entre las más frecuentes se encuentra la distancia *Euclidiana*, que está definida por la siguiente ecuación:

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2}$$

- $d_{ij}$  representa la distancia euclidiana entre los puntos  $i$  y  $j$ .
- $X_{ik}$  y  $X_{jk}$  son las coordenadas del punto  $i$  y  $j$  respectivamente en la variable  $k$ .

La distancia euclidiana es una medida de la distancia lineal directa entre los puntos en un espacio  $p$ -dimensional. Cuanto menor sea la distancia euclidiana, más cercanos estarán los puntos y mayor será su similitud en términos de su posición en el espacio.

Otra medida es la distancia  $D^2$  de Mahalanobis, también conocida como distancia generalizada, la cual está definida por:

$$D^2 = d_{ij} = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)^T \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_j)}$$

Donde  $\Sigma$  es la matriz de varianzas y covarianzas de los datos, y  $\mathbf{X}_i$  y  $\mathbf{X}_j$  son los vectores de las mediciones que identifican los dos objetos  $i$  y  $j$ .

### Métodos de Agrupamiento

Luego de obtener la matriz de distancias, se emplean métodos de agrupación con el objetivo de identificar individuos que se encuentren cercanos o lejanos entre sí, con el fin de generar grupos homogéneos de individuos. Estas técnicas se pueden clasificar en agrupaciones jerárquicas y no jerárquicas (Hernández Meneses, 2015).

### Métodos Jerárquicos

Los métodos jerárquicos son técnicas diseñadas para construir una jerarquía de grupos. Estos métodos inician el proceso considerando cada individuo como un grupo individual. A medida que avanzan, se combinan los dos grupos más cercanos en uno más grande o se dividen grupos ya existentes en otros más pequeños. La elección entre estas dos estrategias depende de si el método utilizado es aglomerativo o divisivo.

En el enfoque aglomerativo, parte de grupos individuales que contienen un solo individuo y se procede a fusionar iterativamente los pares de grupos más cercanos hasta que todas los individuos pertenecen a un único grupo o se alcanza algún criterio de parada. Por otro lado, el enfoque divisivo comienza con un solo grupo que abarca todas las observaciones y se enfoca en dividirlo en grupos más pequeños, también de manera iterativa, hasta que cada observación constituye su propio grupo o hasta cumplir un criterio de parada específico.

Los criterios de parada en métodos jerárquicos pueden variar y suelen incluir el alcanzar un número predefinido de grupos, llegar a un umbral de distancia en el dendrograma, o satisfacer métricas de calidad como la cohesión y la separación de los grupos.

### Método de Ward

El método de Ward, originalmente propuesto por Ward y Wishart, se emplea para agrupar grupos en cada etapa de manera que se minimice la suma total de los cuadrados de las diferencias entre ellos. Esta estrategia busca encontrar la combinación de clusters que deben fusionarse para resultar en la menor pérdida de homogeneidad dentro de los grupos conformados.

En cada paso del método de Ward, se evalúa la distancia euclidiana al cuadrado entre cada individuo y los  $K$  centroides. El centroe de un grupo representa el punto promedio de todos los individuos que lo componen Gallardo (2011).

Al unir los dos grupos que generan el menor incremento en la suma de los cuadrados de las diferencias, se obtiene una nueva agrupación que reduce la dispersión total de los datos. Este proceso de unión se repite iterativamente hasta que todos los individuos estén agrupados en un único grupo o se cumpla un criterio de parada predefinido.

Denotado por:

- $X_{ki,j}$ : el valor de la  $j$ -ésima variable sobre el  $i$ -ésimo individuo del  $k$ -ésimo cluster, suponiendo que dicho cluster posee  $n_k$  individuos.
- $\bar{X}_{kj}$ : la media de la  $j$ -ésima variable dentro del  $k$ -ésimo clúster.
- $E_k$ : la suma de cuadrados de los errores del clúster  $k$ , es decir, la distancia euclidiana al cuadrado entre cada individuo del clúster  $k$  y su centroe.

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^p (x_{ki,j} - \bar{x}_{kj})^2 = \sum_{i=1}^{n_k} \sum_{j=1}^p (x_{ki,j})^2 - n_k \sum_{j=1}^p (\bar{x}_{kj})^2$$

- $E$ : la suma de cuadrados de los errores para todos los clústeres, es decir, si suponemos que hay  $h$  clústeres.

$$E = \sum_{k=1}^h E_k$$

El proceso comienza con  $m$  grupos, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del grupo y por lo tanto en este

primer paso se tendrá  $E_k = 0$  para cada grupo y con ello,  $E = 0$ . El objetivo del método de Ward es encontrar en cada etapa aquellos dos grupos cuya unión proporcione el menor incremento en la suma total de errores,  $E$ .

### Métodos No Jerárquicos

La característica principal de estos métodos es que se tiene conocimiento de un número determinado de grupos inicial  $K$ , por lo cual se toman los individuos y se intercambia entre los grupos para obtener la mejor ubicación para el mismo, de tal forma que se maximiza la homogeneidad de las características frente a individuos del mismo grupo asignado y se obtiene una diferencia significativa frente a los rasgos de los otros grupos. Estos métodos no involucran relaciones de orden como lo hacen los métodos jerárquicos (Peña, 2002).

### K-means

De acuerdo con Peña (2002), el método  $k$ -means dispone de  $n$  individuos con  $p$  variables, los cuales se agrupan en  $k$  grupos y se ejecuta un algoritmo de cuatro etapas.

1. Seleccionar  $k$  puntos como centros de los grupos iniciales. Esto puede hacerse:
  - Asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos así formados.
  - Tomando como centros los  $k$  puntos más alejados entre sí.
  - Construyendo unos grupos iniciales con información a priori y calculando sus centros, o bien seleccionando los centros a priori.
2. Calcular las distancias de cada individuo a los  $k$  centroides y asignarles a cada uno de los individuos el grupo que presente el centroide más cercano. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas del nuevo centro del grupo.
3. Definir un criterio de optimalidad, que puede ser minimizar la suma de cuadrados dentro de los grupos y comprobar si reasignando alguno de los elementos mejora el criterio.

$$\text{suma de cuadrados} = SCDG = \sum_{k=1}^K \sum_{j=1}^p \sum_{i=1}^{n_k} (x_{ijk} - x_{jk})^2$$

Donde  $X_{ijg}$  es el valor de la variable  $j$  en el elemento  $i$  del grupo  $g$  y  $X_{jg}$  es la media de esta variable en el grupo

4. Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

### Determinación del Número de Grupos

La elección del número final de grupos es la última pregunta a responder para determinar la correcta y óptima selección de los grupos o clústeres. De la correcta selección de la cantidad de grupos, estos tendrán una óptima caracterización entre las características de los individuos. Existen diversas técnicas y criterios para tomar esta decisión.

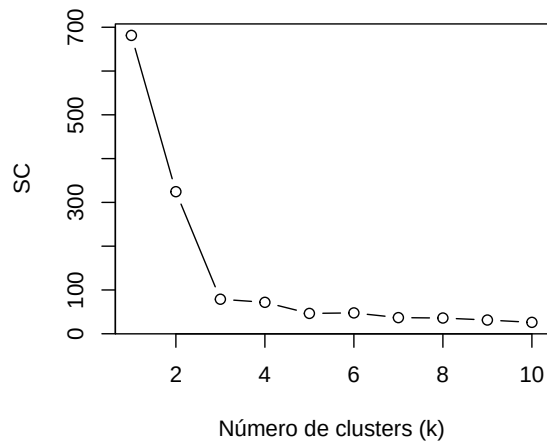
### Elbow method (Método del codo)

Este método consiste en trazar la suma de las distancias al cuadrado entre los puntos de datos y sus centroides en función del número de clústeres. Se selecciona el número de clústeres en el punto donde se observa un cambio significativo en la pendiente de la curva o “codo”.

$$SC = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

Donde:

- $SC$  representa la suma de cuadrados intra-cluster.
- $k$  es el número de clústeres,  $C_i$  es el  $i$ -ésimo clúster y  $x$  es un punto de datos en el clúster  $C_i$ .
- $\mu_i$  es el centroide del clúster  $C_i$ .
- $\|x - \mu_i\|^2$  es la distancia euclidiana al cuadrado entre el punto de datos  $x$  y el centroide  $\mu_i$ .



**Figura 3-9.:** Ejemplo de representación método del codo.  
Elaboración propia

### Coeficiente de silueta (Silhouette coefficient)

Este coeficiente se calcula para cada punto de datos y mide la calidad de su asignación al grupo correspondiente. Se puede calcular para diferentes números de grupos y se selecciona aquel número que maximice el coeficiente de silueta promedio. Para cada objeto  $i$  se denota por  $A$  la agrupación a la que pertenece (Struyf et al., 1997).

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j)$$

Distancia promedio de  $i$  para todos los demás objetos de  $A$ .

Consideremos ahora cualquier grupo  $C$  diferente de  $A$  y calcular:

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j)$$

Distancia promedio de  $i$  para todos los objetos de  $C$ .

Después de calcular  $d(i, C)$  para todos los grupos  $C \neq A$ , tomamos el más pequeño de ellos:

$$b(i) = \min_{C \neq A} d(i, C)$$

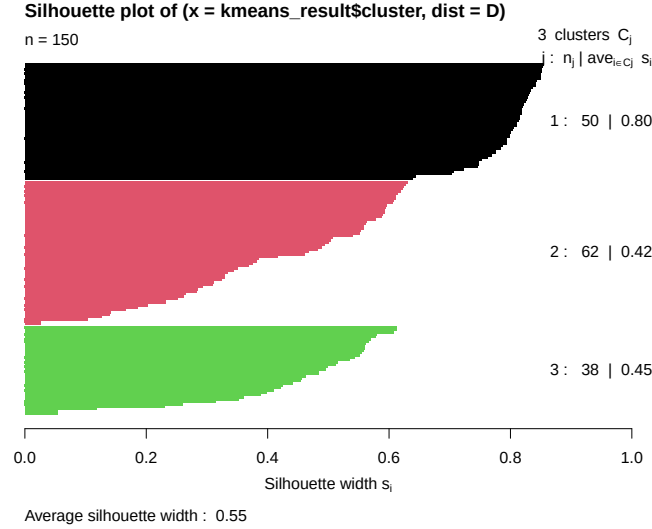
El grupo  $B$  que alcanza este mínimo se llama el vecino del objeto  $i$ . Este es el segundo mejor grupo para el objeto  $i$ .

El valor silueta  $s(i)$  del objeto  $i$  se define como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Claramente,  $s(i)$  siempre se encuentra entre -1 y 1. El valor de  $s(i)$  puede interpretarse de la siguiente manera:

- $s(i) \approx 1$ : el objeto  $i$  está bien clasificado (en  $A$ ).
- $s(i) \approx 0$ : el objeto  $i$  se encuentra intermedio entre dos grupos ( $A$  y  $B$ ).
- $s(i) \approx -1$ : el objeto  $i$  está mal clasificado (más cercano a  $B$  que a  $A$ ).



**Figura 3-10.:** Ejemplo de representación coeficiente de silueta.  
Elaboración propia

### Caracterización de los grupos (Valores test):

Para finalizar con la creación de los grupos, es necesario llevar a cabo la caracterización de las variables continuas. En este contexto, se consideran los parámetros:  $X$  representa una variable continua;  $\bar{X}$  se refiere a la media general de esta variable;  $\bar{X}_k$  denota la media de las observaciones de  $X$  que pertenecen a la clase  $k$ ; y finalmente,  $s_k^2(X)$  simboliza la varianza general de la variable  $X$ , definida como  $s_k^2(X) = (n - n_k)s^2(X)$ . A partir de estos parámetros, se obtiene la siguiente expresión para caracterizar la relación de la variable dentro de cada clase:

$$t_k = \frac{\bar{X}_k - \bar{X}}{s_k(X)}$$

donde  $s^2(X)$  es la varianza de la variable  $X$  en todo el conjunto de datos,  $n$  es el número de individuos clasificados y  $n_k$  es el número de individuos dentro de la clase  $k$ . Por lo tanto,  $|t(X_1)_k| > 2$  indica que la variable  $X_1$  es relevante en la clase  $k$ .

### Population Stability Index (PSI)

Es una métrica estadística que se utiliza comúnmente para evaluar el rendimiento de los modelos de puntuación de riesgo, como los modelos de crédito o fraude. El objetivo del PSI es medir cuánto ha cambiado una distribución de puntuación cuando se compara un conjunto de datos de muestra con un conjunto de datos de referencia.

Matemáticamente, el PSI se calcula generalmente de la siguiente manera:

$$\text{PSI} = \sum_{i=1}^n (p_i - e_i) \ln \left( \frac{p_i}{e_i} \right)$$

Donde:

- $n$  es el número de grupos o “bins” en los que se dividen los datos.
- $p_i$  es la proporción de observaciones en el  $i$ -ésimo grupo en el conjunto de datos de muestra.
- $e_i$  es la proporción de observaciones en el  $i$ -ésimo grupo en el conjunto de datos de referencia.

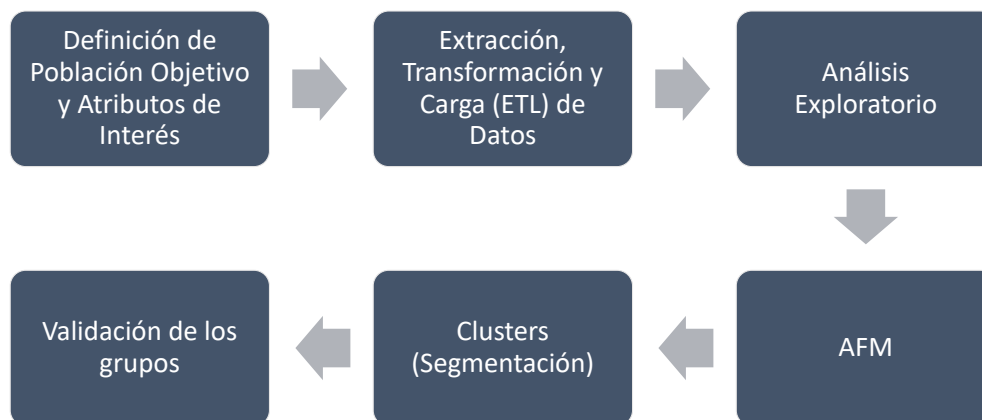
Un valor de PSI cercano a cero indica que los dos conjuntos de datos (muestra y referencia) son muy similares en términos de la distribución de puntuación, lo que generalmente se considera una buena señal de que el modelo está funcionando como se esperaba a lo largo del tiempo o en diferentes poblaciones. Valores de PSI más altos sugieren que la distribución de puntuación ha cambiado, lo que podría indicar problemas con la estabilidad del modelo. Aquí hay algunas pautas generales para interpretar el valor del PSI:

- $\text{PSI} < 0,1$ : Muy poco cambio, excelente.
- $0,1 \leq \text{PSI} < 0,25$ : Algunos cambios menores, todavía aceptable.
- $\text{PSI} \geq 0,25$ : Cambios significativos en la distribución, se requiere revisión del modelo.

El PSI es ampliamente utilizado en industrias como la banca y las finanzas para monitorear la eficacia de los modelos de puntuación de riesgo y para cumplir con las regulaciones que requieren una validación y revisión regular de estos modelos.

## 4. Metodología

Este capítulo especifica la metodología empleada para el cumplimiento con los objetivos establecidos en el trabajo de grado. El problema se analizó desde el desarrollo de las etapas expuestas en la Figura 4-1



**Figura 4-1.:** Secuencia Metodológica

## 4.1. Población Objetivo

Con la formación de un equipo de trabajo establecido por la entidad, compuesto por un estadístico máster, un ingeniero de datos y un estadístico sénior, se llevaron a cabo reuniones con el área de gestión comercial. Estos encuentros resultaron fundamentales tanto para la elección de la población objeto de estudio como para la extracción y selección de los datos. Como resultado de este proceso colaborativo, se decidió enfocar el estudio en todos los tarjetahabientes pertenecientes a la **franquicia Mastercard**, debido a que se mantiene una relación estrecha con ellos y se dispone de una vasta cantidad de información. Por ello, se establecieron los siguientes criterios de inclusión:

- Tarjetahabientes bajo la franquicia Mastercard.
- Que tengan al menos una transacción en el período evaluado.
- El período en que se llevará a cabo el estudio abarca desde Dic 2021 hasta Dic 2022.

A partir de estos tres criterios de selección para la población objetivo, se determinó que un total de **26.307 clientes** cuentan con una tarjeta de la franquicia Mastercard y han realizado al menos una transacción en el período de evaluación.

### Definición de variables

#### *Grupo 1: Variables Sociodemograficas Categóricas*

**Estrato:** Nivel socioeconómico.

**Estado Civil:** El estado civil del cliente.

**Nivel Académico:** Grado de formación y conocimientos que una persona ha adquirido.

**Puntaje Acierta:** Calificación crediticia.

#### *Grupo 2: Variables Sociodemograficas Numéricas*

**Total Ingresos SMMLV:** El total de los ingresos en salarios mínimos legales vigentes.

**Edad:** La edad en años.

**Antigüedad en el Banco:** El tiempo que ha estado asociado al banco.

#### *Grupo 3: Variables Transaccionales Anuales Numéricas*

**Porcentaje plazo 1-3:** Porcentaje de transacciones asignados entre 1 y 3 cuotas.

**Porcentaje plazo 3-6:** Porcentaje de transacciones asignados entre 3 y 6 cuotas.

**Porcentaje plazo 36:** Porcentaje de transacciones asignados hasta 36.

**Porcentaje almacenes:** Porcentaje de transacciones en Almacenes.  
**Porcentaje avance:** Porcentaje de transacciones en avances.  
**Porcentaje combustibles:** Porcentaje de transacciones en combustibles.  
**Porcentaje comidas:** Porcentaje de transacciones en comidas rápidas.  
**Porcentaje correo:** Porcentaje de transacciones en correos y encomiendas.  
**Porcentaje droguerías:** Porcentaje de transacciones en droguerías.  
**Porcentaje plataformas :** Porcentaje de transacciones en plataformas.  
**Porcentaje restaurantes:** Porcentaje de transacciones en restaurantes.  
**Porcentaje servicios:** Porcentaje de transacciones en servicios públicos.  
**Porcentaje tiendas:** Porcentaje de transacciones en tiendas.

**Días máximo usado:** Días desde el último uso de la TC.  
**Días última compra:** Tiempo en días desde la última compra.  
**Días último avance:** Tiempo en días desde el último avance.  
**Indicador uso TC:** Uso la tarjeta en el último mes.  
**Monto avance último mes:** Monto de avances en el mes de ejecución.  
**Monto compra último mes:** Monto de compras en el mes de ejecución.  
**Monto transacciones último mes:** Monto de transacciones en el mes de ejecución.  
**Número transacciones último mes:** Cantidad de transacciones en el mes.  
**Número avances último mes:** Cantidad de avances en el mes de ejecución.  
**Número compras último mes:** Cantidad de compras en el mes de ejecución.  
**Número compras:** Cantidad de compras.  
**Número avances:** Cantidad de avances.  
**Valor avances:** Valor total de avances.  
**Valor compras:** Valor total de compras.  
**Plazo promedio compras:** Plazo promedio de las compras.  
**Porcentaje uso mes:** Porcentaje de Uso mensual.

#### ***Grupo 4: Variables Transaccionales Anuales Categóricas***

**Realiza compras:** Realizo al menos una compra en el último año.  
**Realiza avances:** Realizo al menos un avance en el último año.  
**MCC más usado:** MCC más usado junto con su categoría.  
**MCC mayor monto:** MCC de mayor monto junto con su categoría.

## **4.2. Extracción, Transformación y Carga (ETL) de Datos**

Como paso importante en la recopilación y obtención de la base de datos final, se hace uso del proceso de ETL (Extract, Transform, Load). En esta sección, se definirá cada paso

realizado de ETL en el trabajo. Este proceso es esencial para enfrentar el desafío de trabajar con datos provenientes de diversas fuentes y formatos.

Al aplicar la extracción, se obtienen los datos necesarios para abordar los objetivos de investigación planteados. Posteriormente, la transformación de los datos garantiza su coherencia y calidad. Finalmente, la carga y depuración de los datos en un repositorio estructurado facilita su análisis. Detallar cada una de estas etapas en el contexto de este trabajo permitirá comprender cómo se han tratado los datos desde su origen hasta su preparación para el análisis.

### 4.2.1. Extracción

La recolección de datos se realizó mediante un proceso de extracción de la información contenida en las bases de datos de la entidad. Este proceso estuvo a cargo de un equipo especializado, con conocimiento sobre la ubicación de las variables empleadas en el estudio. Dichas variables se encuentran en dos repositorios distintos: uno que contiene información de los usuarios de la entidad y que incluye las variables contenidos en el Anexo **A-1**.

La segunda base de datos conserva registros de transacciones a lo largo de un año, actualizándose cada mes para siempre reflejar un historial de doce meses consecutivos. Las variables presentes en esta base pueden observarse en la Tabla **A-2**. A partir de esta información, generaremos un nuevo conjunto de datos a nivel de cliente, que se unirá con la Tabla **A-1** usando una llave de identificación única por cliente. El resultado final será una base de datos en la que cada fila representará información individual de un cliente, acompañada de indicadores de transacción calculados a partir de registros anuales.

### 4.2.2. Transformación

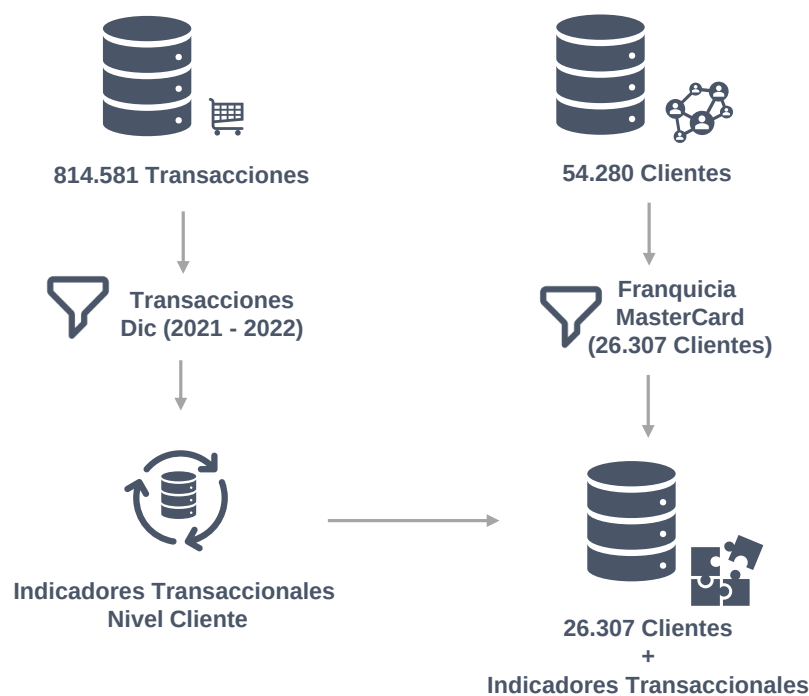
#### Recategorización y creación de variables

Durante el proceso de análisis se identificó que las variables 'Estado Civil', 'Nivel Académico', 'Estrato' y 'Puntaje Acierta' presentaban un alto número de categorías que agregar complejidad al análisis dentro del **AFM**. En una reunión preliminar, el equipo evaluó estas variables y consideró oportuno reformular las categorías. Esta decisión no solo estuvo basada en criterios técnicos, sino también en el conocimiento profundo del negocio. Con ello, se buscó simplificar y agilizar el proceso de análisis, permitiendo una mejor interpretación y operatividad de los datos dando como resultado las nuevas categorías que se observan en la Tabla **4-1**. Adicionalmente, se utilizó la tabla de registros anuales de transacciones para calcular nuevas variables, las cuales están detalladas en el Anexo **A-3**.

| Variable        | Categorías Originales                           | Nuevas Categorías |
|-----------------|-------------------------------------------------|-------------------|
| Estado Civil    | Unión Libre, Casado                             | Con Vínculo       |
|                 | Otros                                           | Sin Vínculo       |
|                 | Sin Información                                 | Sin Información   |
| Nivel Académico | Primaria, Bachiller                             | Bajo              |
|                 | Técnico, Tecnólogo                              | Medio             |
|                 | Profesional, Especialización, Máster, Doctorado | Alto              |
|                 | Sin Información                                 | Sin Información   |
| Estrato         | Bajo-Bajo, Bajo                                 | Bajo              |
|                 | Medio-Bajo, Medio                               | Medio             |
|                 | Medio-Alto, Alto                                | Alto              |
| Puntaje Acierta | Numérica (0 - 1000)                             | Hasta 700         |
|                 |                                                 | 700 - 800         |
|                 |                                                 | Superior a 800    |
|                 | Sin Información                                 | Sin Información   |

**Tabla 4-1.:** Recategorización de Variables

### 4.2.3. Carga



**Figura 4-2.:** Diagrama de proceso para la obtención de datos.

Finalmente como se aprecia en la Figura 4-2, la base de datos que se cargará está originalmente conformada por **814.581 registros** de transacciones de **54.280 tarjetahabientes**. Sin embargo, debido a que no todos los tarjetahabientes realizaron al menos una transacción durante el periodo evaluado, que abarca desde diciembre de 2021 hasta diciembre de 2022, no fue posible calcular las variables de transacciones para todos ellos. Como resultado, un total de **27.973 tarjetahabientes fueron excluidos del análisis**. Después de llevar a cabo este proceso de depuración, se logró obtener un conjunto final de **26.307 tarjetahabientes** con los cuales se llevará a cabo el trabajo de análisis y modelado.

### 4.3. Análisis Descriptivo y Exploratorio

Se realiza un análisis descriptivo y exploratorio de los datos, aplicando técnicas estadísticas para examinar las relaciones entre las variables e identificar características relevantes entre los individuos. Para las variables cuantitativas, se emplean medidas de tendencia central y dispersión. En el caso de las variables categóricas, se utilizan tablas de frecuencia y análisis de asociación. Los resultados obtenidos a través de procedimientos computacionales se detallan en la sección de Resultados.

Como criterio de elección de variables, se llevará a cabo un **AFM** preliminar con todas las variables por grupo, evaluando la relación que presenta con las primeras 10 componentes principales. A manera de filtro se seleccionarán aquellas variables que presenten una mayor grado de relación con estas componentes para su inclusión en el análisis subsiguiente.

### 4.4. Análisis Factorial Múltiple (AFM)

El **AFM** se lleva a cabo bajo el software estadístico **R**, utilizando las librerías *FactoMineR* (Lê et al., 2008) y *Factoextra* (Kassambara and Mundt, 2020), las cuales permiten el cálculo y la visualización de los resultados. Utilizando cuatro tablas de datos: *Variables Sociodemográficas Numéricas*, *Variables Transaccionales Anuales Numéricas*, *Variables Transaccionales Anuales Categóricas* y *Variables Sociodemográficas Categóricas*. Cada tabla contiene 3, 20, 3 y 4 variables, respectivamente. Tras depurar la base de datos, se obtuvo un total de 30 variables y 26,307 individuos. Para determinar el número de ejes factoriales, se empleó el porcentaje de varianza acumulada, revelando que las 10 primeras componentes las cuales explican el 58.17 % de la varianza. Las representaciones de los grupos y los individuos se realizaron gráficamente, donde los grupos se presentaron mediante el plano factorial y los individuos por la proyección de individuos, las variables se representaron en el círculo de correlaciones.

Se estimaron las correlaciones de los grupos en los dos primeros ejes, los coeficientes RV y las contribuciones de las variables, para valorar las relaciones entre los grupos y los aportes que las variables realizan a cada eje factorial o dimensión. En etapas posteriores, se toman las coordenadas de los individuos derivadas del análisis factorial para crear una nueva hoja de datos. Este nuevo conjunto de datos sirve como la base para la formación de clusters.

## 4.5. Clusters

En el siguiente paso del análisis, se emplea el algoritmo de clustering *K*-means para crear los clusters de individuos basados en los scores generados por las primeras 10 componentes principales. La elección de *K*-means se justifica por su simplicidad y eficacia, lo que es crucial para la entidad que requiere que estos grupos se generen y evalúen de manera periódica. Además de los métodos tradicionales como el método del codo y el índice de la silueta para determinar la calidad del clustering, se introduce un criterio adicional: el Population Stability Index (PSI), o Índice de Estabilidad de la Población. Este índice se evalúa mensualmente para asegurar que los grupos sean estables a lo largo del tiempo, lo cual es vital para que los resultados del análisis sean útiles y confiables en el mediano y largo plazo.

## 4.6. Validación e Interpretación de los Clúster

La validación e interpretación de los clústeres se llevan a cabo en dos etapas esenciales. En primer lugar, se realiza un análisis estadístico empleando pruebas de valores (t-tests) para caracterizar los grupos formados. Este enfoque estadístico aporta una primera capa de rigor al análisis. Sin embargo, la etapa más crítica del proceso de validación es la revisión final llevada a cabo en colaboración con el área comercial de la entidad. En estas reuniones, se presentan y discuten los resultados del análisis de clustering. La aprobación del área comercial es fundamental para dar por concluido el proyecto. Este enfoque combinado asegura que los clústers no solo sean estadísticamente válidos, sino que también sean prácticos y aplicables a las necesidades comerciales de la entidad.

## 5. Resultado

En este capítulo se presentan los resultados obtenidos a través de un análisis exploratorio por grupos de variables. El objetivo de este análisis es conocer las características de cada grupo.

### 5.1. Análisis Exploratorio

La antigüedad de los individuos se encuentra entre los 0 y 44 años, con una media de 10.25 años. La dispersión es amplia (desviación estándar de 8.79), junto al CV de 0.86, indicando una amplia variabilidad en la antigüedad. La edad de los clientes se encuentra en el rango de 18 a 104 años. El mínimo de edad tiene sentido, pero el máximo sugiere la presencia de valores atípicos. Al analizar la variable del total de ingresos en SMMLV, el máximo de 5700 da una fuerte evidencia de un valor atípico. Sin embargo, en reuniones con la entidad financiera, descartaron la idea de que estos valores lo sean. Por el contrario, estos individuos son un porcentaje minoritario que no supera el 1 % del total. Esta variable, debido a estos datos atípicos, toma un CV de 2.72 que se considera elevado.

| Variables                       | Min | Max | Mean  | Std.Dev | CV   |
|---------------------------------|-----|-----|-------|---------|------|
| Antigüedad en la Entidad (Años) | 0   | 44  | 10.25 | 8.79    | 0.86 |
| Edad (Años)                     | 18  | 104 | 47.18 | 14.45   | 0.31 |
| Total Ingresos (SMMLV)          | 0   | 912 | 6.61  | 17.95   | 2.72 |

**Tabla 5-1.:** Variables Sociodemográficas Numéricas

De acuerdo a la Tabla 5-1, La variable de puntaje Acierta, más de la mitad de la población tiene un puntaje crediticio superior a los 800 puntos. Estos individuos cuentan en su mayoría con un nivel académico profesional (62 %). Además, mas de un 71 % pertenece a estratos medio.

| Variable               | Categorías   | Frecuencia (%) |
|------------------------|--------------|----------------|
| <b>Puntaje Acierta</b> | Superior 800 | 15650 (59.4 %) |
|                        | 700 - 800    | 6635 (25.2 %)  |
|                        | Hasta 700    | 2856 (10.8 %)  |
|                        | Sin Info     | 1166 ( 4.6 %)  |
| <b>Estado Civil</b>    | Con Vinculo  | 10675 (40.6 %) |
|                        | Sin Vinculo  | 15632 (59.4 %) |
| <b>Nivel Academico</b> | E.M.D.O      | 5880 (22.4 %)  |
|                        | Profesional  | 16393 (62.3 %) |
|                        | B.T.T        | 4034 (15.3 %)  |
| <b>Estrato</b>         | Alto         | 5689 (21.6 %)  |
|                        | Medio        | 18701 (71.1 %) |
|                        | Bajo         | 1917 ( 7.2 %)  |

**Tabla 5-2.:** Variables Sociodemográficas Categóricas

En la Tabla **5-2** la antigüedad de la tarjeta medida en meses, se tienen datos de tarjetas nuevas, hasta un máximo de 86 meses, con una media de 46 meses. Para los días desde que se ha realizado una última compra o avances, los mínimos y máximos toman valores iguales. Esto nos dice que se tienen datos de tarjetas que han realizado transacciones tanto recientes como tarjetas inactivas en alguna de las dos categorías, recordando que los datos están condicionados a que las tarjetas en la base de datos deben tener al menos una transacción en el periodo de un año. Las métricas de dispersión resultan de alta variabilidad para ambas variables. La media de estos días revela que las compras son más realizadas que los avances.

Las variables de conteo, como lo son número de avances o compras, número de avances y compras en el último mes, confirman que en la base de datos hay presencia de tarjetas que pueden usarse en uno de los dos tipos de transacciones. Esto se aclara por el hecho de que las tasas de interés son diferentes en estos dos tipos de transacción. Por otra parte, la variable de plazos promedios va de un mínimo de 0 a 36 plazos, ya que estos están establecidos como límites de plazos.

Los porcentajes de plazos manejados están establecidos por la entidad como los de mayor uso. De estos porcentajes, los de mayor uso, de acuerdo con la media 48 %, son los plazos de 0 a 1 mes, lo que sugiere que una proporción significativa de las compras se pagan dentro del primer mes.

|                                         | Min  | Max              | Mean         | Std.Dev       | CV   |
|-----------------------------------------|------|------------------|--------------|---------------|------|
| <b>Días desde Última Compra</b>         | 2.00 | 730.00           | 136.33       | 221.15        | 1.62 |
| <b>Días desde Último Avance</b>         | 1.00 | 730.00           | 448.01       | 317.17        | 0.71 |
| <b>Número Avances</b>                   | 0.00 | 127.00           | 2.49         | 5.44          | 2.19 |
| <b>Número Compras</b>                   | 0.00 | 905.00           | 27.37        | 46.27         | 1.69 |
| <b>Número Compras Último Mes</b>        | 0.00 | 100.00           | 2.32         | 4.67          | 2.01 |
| <b>Número Avances Último Mes</b>        | 0.00 | 24.00            | 0.21         | 0.77          | 3.65 |
| <b>Porcentaje Uso Mes</b>               | 0.00 | 3.04             | 0.10         | 0.20          | 1.91 |
| <b>Porcentaje Plazo 0 - 1</b>           | 0.00 | 1.00             | 0.47         | 0.37          | 0.79 |
| <b>Porcentaje Plazo 1 - 3</b>           | 0.00 | 1.00             | 0.12         | 0.19          | 1.58 |
| <b>Porcentaje Plazo 3 - 6</b>           | 0.00 | 1.00             | 0.07         | 0.16          | 2.21 |
| <b>Porcentaje Plazo 36</b>              | 0.00 | 1.00             | 0.24         | 0.32          | 1.37 |
| <b>Porcentaje de Avances</b>            | 0.00 | 1.00             | 0.21         | 0.34          | 1.62 |
| <b>Porcentaje Almacenes</b>             | 0.00 | 1.00             | 0.15         | 0.20          | 1.35 |
| <b>Porcentaje Combustibles</b>          | 0.00 | 1.00             | 0.04         | 0.11          | 2.75 |
| <b>Porcentaje Comida Rápidas</b>        | 0.00 | 1.00             | 0.03         | 0.06          | 2.44 |
| <b>Porcentaje Plataformas Digitales</b> | 0.00 | 1.00             | 0.11         | 0.22          | 2.10 |
| <b>Porcentaje Restaurantes</b>          | 0.00 | 1.00             | 0.02         | 0.07          | 2.64 |
| <b>Porcentaje Tiendas</b>               | 0.00 | 1.00             | 0.07         | 0.14          | 1.93 |
| <b>Valor Avances</b>                    | 0.00 | 321,800,000.00   | 1,822,392.64 | 5,188,189.26  | 2.85 |
| <b>Valor Compras</b>                    | 0.00 | 1,447,036,263.00 | 4,900,104.31 | 12,422,897.60 | 2.54 |

**Tabla 5-3.:** Variables Transaccionales Tarjeta Númericas

Además, el porcentaje de plazos de 36 meses es también notable. En promedio, el 24 % de las compras se realizan hasta un plazo de 36 meses. Aunque esta proporción es menor que la del primer mes, es significativamente mayor que para los otros plazos. La desviación estándar de 0.32 y el coeficiente de variación de 1.37 indican una variabilidad moderada en esta variable.

La mayoría de las tarjetas se utilizan más frecuentemente para compras que para avances en efectivo, como lo demuestra el hecho de que el Porcentaje de Avances es solo del 21 %. Las tarjetas se utilizan con mayor frecuencia para compras en plataformas digitales, almacenes y tiendas, como lo indican sus porcentajes de uso más altos en comparación con otras categorías. Por otro lado, las categorías de gasto que presentan un uso más bajo de tarjetas de crédito incluyen comida rápida, restaurantes, servicios públicos y transacciones por correo. Esto podría sugerir que los titulares de tarjetas prefieren utilizar otros métodos de pago para este tipo de transacciones.

En general, estos datos sugieren que, si bien las tarjetas se utilizan con mayor frecuencia para compras que para avances de efectivo, cuando se realizan avances de efectivo, suelen ser por

montos considerables. Además, aunque el monto medio de las compras y las transacciones es relativamente alto, también existe una gran variabilidad en estos montos.

| Variable               | Categorías                       | Frecuencia (%) |
|------------------------|----------------------------------|----------------|
| <b>MCC Más Usado</b>   | Avance                           | 7272 (27.6 %)  |
|                        | Almacenes Con Y Sin Supermercado | 6270 (23.8 %)  |
|                        | Otras                            | 4279 (29.3 %)  |
|                        | Plataformas Virtuales            | 2107 ( 8.0 %)  |
|                        | Servicios Públicos               | 1599 ( 6.1 %)  |
|                        | Tiendas De Ropa                  | 1378 ( 5.2 %)  |
| <b>Realiza Compras</b> | No realiza compras               | 2775 (10.5 %)  |
|                        | Si realiza compras               | 23566 (89.5 %) |
| <b>Realiza Avances</b> | No realiza avances               | 14393 (54.6 %) |
|                        | Si realiza avances               | 11948 (45.4 %) |

**Tabla 5-4.:** Variable Transaccional Tarjeta Categóricas

La Tabla 8-4 indica las variables que son del uso de la tarjeta, como lugares y tipo de transacciones. Una categoría la cual segmenta los clientes en nuevos a partir de 3 meses de antigüedad de la tarjeta. Esta indica que el 6 % pertenece a esta categoría y la idea es encontrar un grupo que contenga esta categoría para la generación de campañas diferentes al resto. Esto también para conocer el uso y sus características. Los MCC de más uso son en su mayoría con un 27 % para la categoría Avance, seguido de Almacenes con y sin supermercado con un 23.8 %, que sumados dan más de un 50 % de clientes. Los avances son una categoría del MCC que maneja una tasa diferente a las demás y a esta se genera una nueva variable que es si realiza al menos 1 avance con la tarjeta. Por lo que se observa que tiene una división casi homogénea entre los que realizan avances, que son el 54 %, y los que no lo hacen, que son el 45.4 %. Al igual que si realiza compras, donde se observa un 89 % que realiza compras con la tarjeta. Por último, la variable Avanceros clasifica los clientes que usaron su tarjeta en más de un 70 % de las veces en el año para realizar Avances.

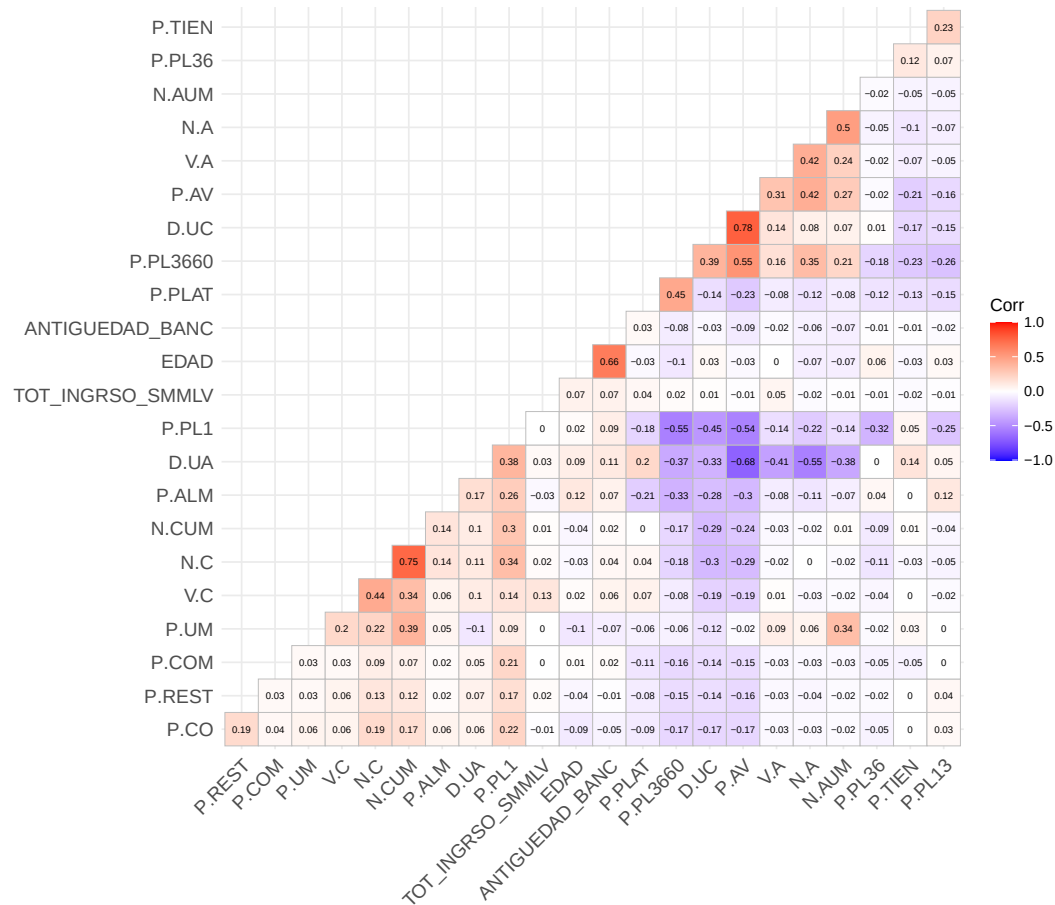
De acuerdo con la Figura 5-1 la fuerte correlación negativa entre 'D.UA' (días desde el último avance) y 'N.A' (número de avances) de -0.55 sugiere que cuantos más días han pasado desde el último avance, es probable que haya habido menos avances. Esto tiene sentido ya que, si ha pasado mucho tiempo desde el último avance, es probable que la frecuencia de estos avances sea menor.

También notamos una fuerte correlación negativa entre 'D.UA' y 'P.AV' (Porcentaje de Avances) de -0.68. Esto podría indicar que si han pasado más días desde el último avance, es probable que el porcentaje de avances en relación con otras transacciones sea menor.

Por otro lado, 'P.AV' (Porcentaje de Avances) y 'D.UC' (días desde la última compra) tienen una fuerte correlación positiva de 0.78. Esto es particularmente interesante porque sugiere que cuanto más tiempo ha pasado desde la última compra, más probable es que una mayor proporción de transacciones sean avances en lugar de compras. Este patrón podría ser motivo de una investigación más profunda, ya que podría estar relacionado con el comportamiento financiero de los clientes.

'N.C' (Número de Compras) y 'N.CUM' (Número de Compras el Último Mes) tienen una correlación fuerte de 0.75, lo cual es lógico. Si una persona ha realizado muchas compras en el último mes, es probable que también haya realizado muchas compras en general.

Curiosamente, 'TOT\_INGRSO\_SMMLV' no muestra correlaciones significativas con otras variables. Esto podría implicar que el nivel de ingreso no tiene un impacto notable en el comportamiento de compra o en la toma de avances de un individuo en este conjunto de datos.

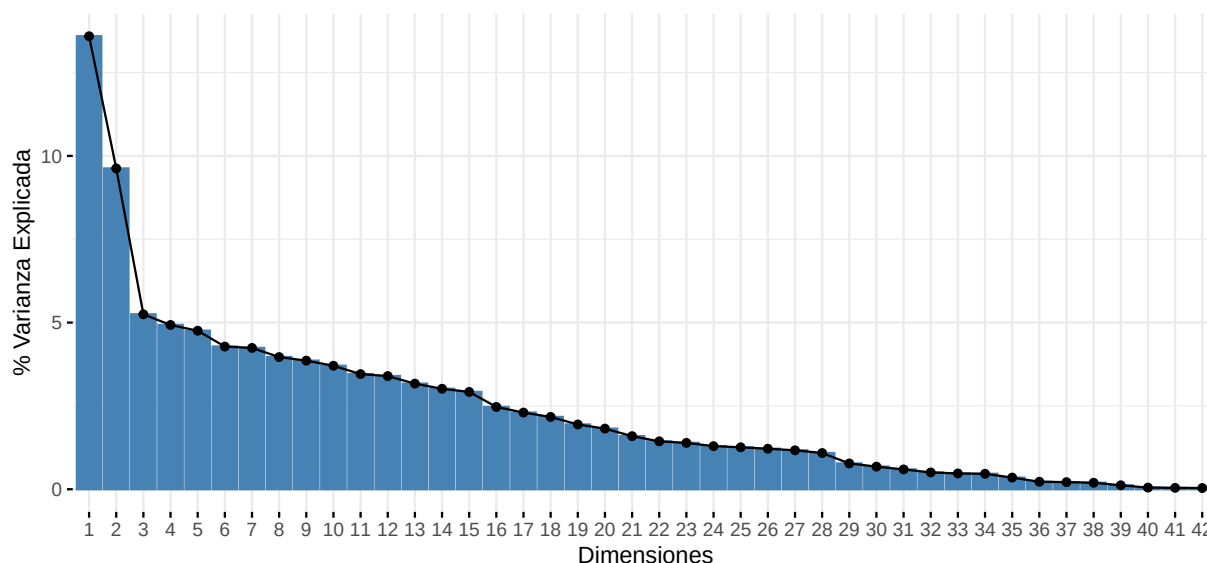


**Figura 5-1.:** Gráfico de la matriz de correlación de las variables consolidadas

## 5.2. Análisis Multivariado

### 5.2.1. Análisis Factorial Múltiple (AFM)

El AFM es una técnica que permite la reducción de la dimensionalidad inicial de los datos y se basa en el ACP. Su objetivo es explicar la naturaleza de los individuos en función de agrupaciones o tablas de variables (SocioDemo Cat, SocioDemo Num, Transac Num, Transac Cat). Cada grupo se analiza individualmente a través de un ACP, que se pondera en cada tabla por el inverso del primer valor propio. Este enfoque busca encontrar un equilibrio en el peso de los grupos, evitando depender del número y calidad de las variables usadas. Posteriormente, se calculan las coordenadas de las variables en cada uno de los grupos con respecto a los factores para obtener la representación global. A continuación, se describe la relación entre los grupos de variables y los factores del AFM.



**Figura 5-2.:** Histograma de Valores Propios

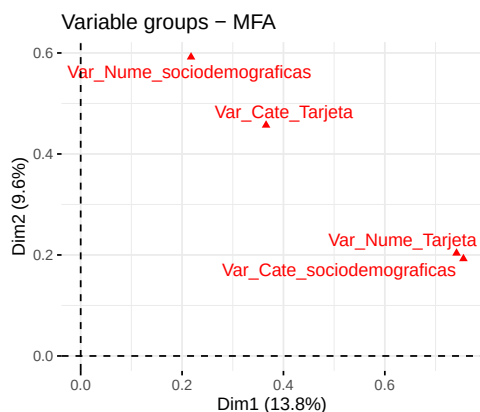
la Figura 5-2, representa los valores propios del AFM. Se toma la decisión de utilizar los primeros 10, que retienen el 58.2% de la variabilidad de los datos, para lograr una representación adecuada de los individuos, como se observa en la Tabla 8-5.

En la representación de los cuatro grupos de variables, Figura 5-3 y 5-4, se puede observar que el plano de los componentes (1-2) recoge el 23.42% de la inercia y su dimensión 1 está determinada por la información del grupo de variables Var.Cat.sociodemograficas y Var.Num.Tarjeta; mientras que la dimensión 2 prácticamente está determinada por la información del grupo de las variables del Var.Num.sociodemograficas y Var.Cat.Tarjeta, pues estos dos grupos están bastante alejados del grupo Social. El plano de los

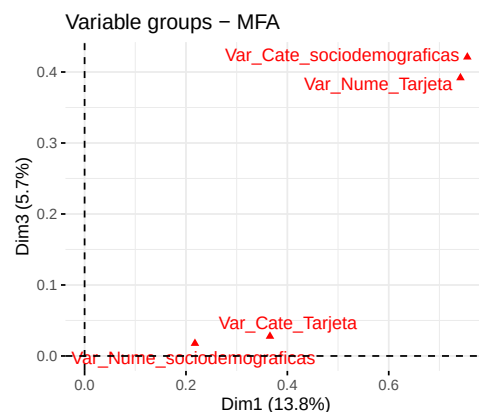
| Componentes | Valores Propios | % Varianza | % Varianza Acumulada |
|-------------|-----------------|------------|----------------------|
| Comp 1      | 2.08            | 13.82      | 13.82                |
| Comp 2      | 1.45            | 9.60       | 23.42                |
| Comp 3      | 0.86            | 5.70       | 29.12                |
| Comp 4      | 0.73            | 4.83       | 33.95                |
| Comp 5      | 0.69            | 4.55       | 38.50                |
| Comp 6      | 0.65            | 4.32       | 42.83                |
| Comp 7      | 0.64            | 4.28       | 47.11                |
| Comp 8      | 0.61            | 4.05       | 51.15                |
| Comp 9      | 0.60            | 3.96       | 55.11                |
| Comp 10     | 0.58            | 3.86       | 58.98                |

**Tabla 5-5.:** Valores Propios del AFM

componentes (1-3) recoge el 19.52% de la inercia y se observa que la dimensión 1 está determinada por la información del grupo de variables de Var\_Cat\_sociodemograficas y Var\_Num\_sociodemograficas, mientras que la dimensión 3 está determinada por los grupos de variables Var\_Cat\_Tarjeta y Var\_Num\_Tarjeta.



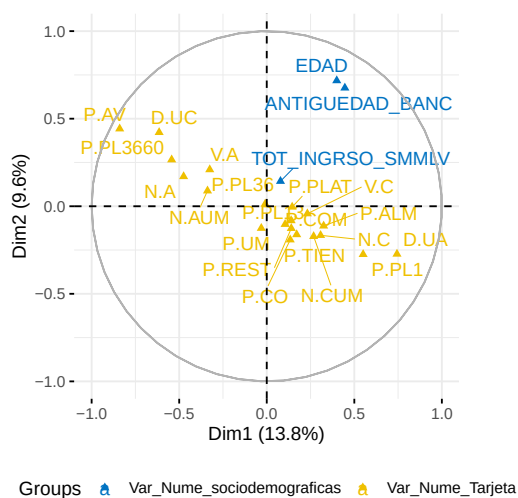
**Figura 5-3.:** Representación de los Grupos Dimensión 1-2



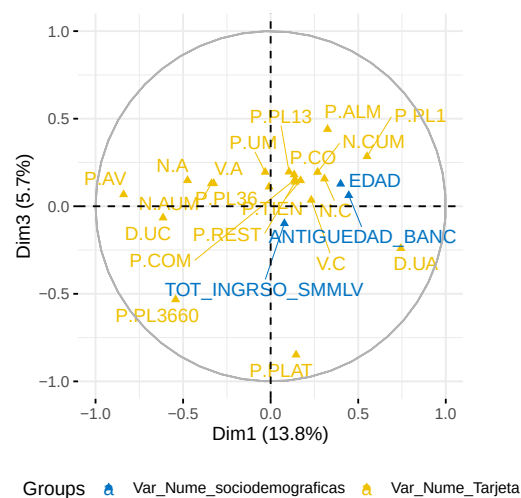
**Figura 5-4.:** Representación de los Grupos Dimensión 1-3

Para comprender la relación entre los primeros factores y las variables, se recurre al círculo de correlaciones 5-5. En este, se muestra cómo quedan representadas las variables en las dos primeras dimensiones del Análisis de Componentes Principales (ACP) global. Se puede apreciar que todas las variables de Tarjetas quedan representadas en el primer eje factorial, mientras que la mayoría de las variables Sociodemograficas están representadas en el segundo eje. En el círculo de correlaciones 5-6, correspondiente a los ejes factoriales 1 y 3, se observa que variables de Tarjetas contribuyen significativamente a la formación del tercer componente.

Por lo tanto, se puede concluir que los componentes 1 y 3 son indicadores de las variables asociadas al grupo de Var\_Num\_Tarjeta, y el componente 2 representa las variables del grupo de Var\_Num\_Sociodemográficas.

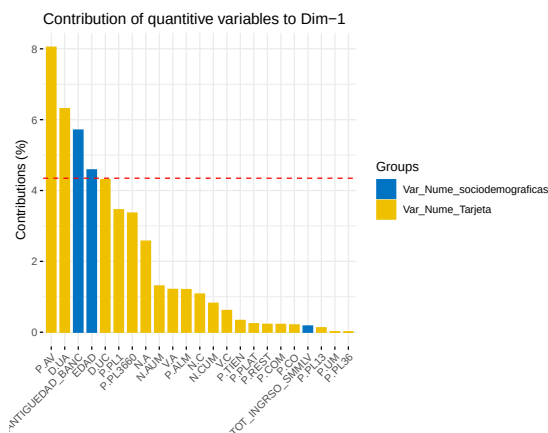


**Figura 5-5.:** Círculo de correlaciones componentes 1 - 2

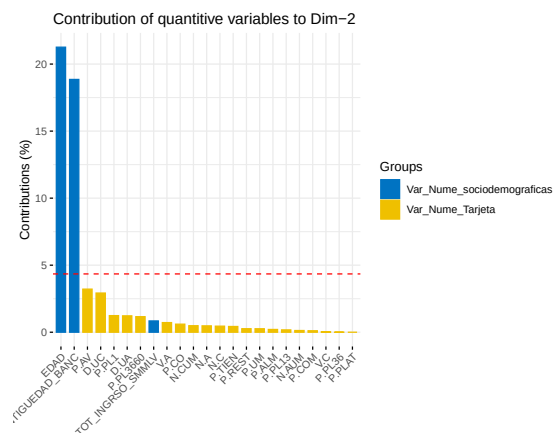


**Figura 5-6.:** Círculo de correlaciones componentes 1 - 3

En las Figuras 5-7 y 5-8, se presenta cómo las variables contribuyen a la formación de las primeras 2 dimensiones. En la primera dimensión, las variables numéricas que más influencia tienen son el porcentaje de avances, los días desde la última compra, la antigüedad y la edad. Por otro lado, para la segunda dimensión, las variables que más contribuyen son la edad y la antigüedad.

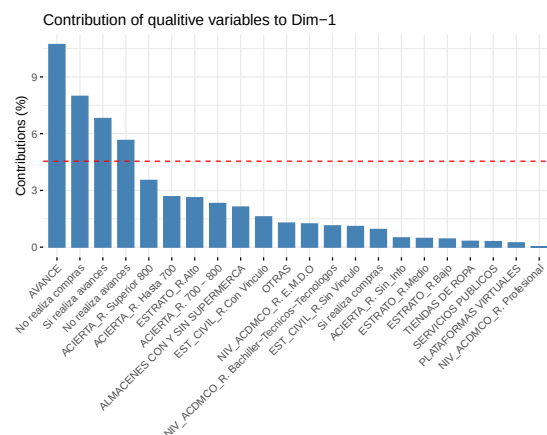


**Figura 5-7.:** Representación contribución variables numéricas Dim1

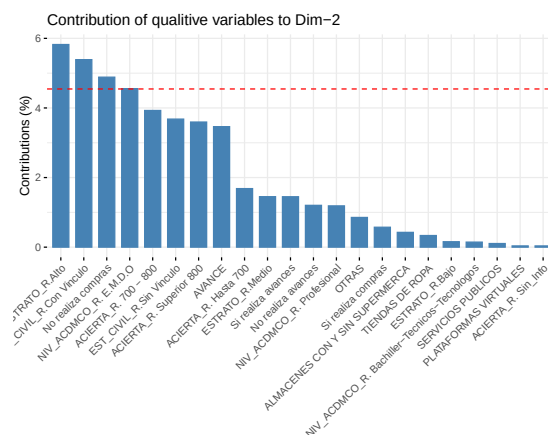


**Figura 5-8.:** Representación contribución variables numéricas Dim2

En las Figuras 5-9 y 5-10, se muestra la manera en que las variables categóricas contribuyen a la construcción de las primeras dos dimensiones. Para la primera dimensión, las variables categóricas de mayor relevancia son 'Avance', 'no realiza compras', 'sí realiza avances' y 'no realiza avances'. En cuanto a la segunda dimensión, las variables que más contribuyen son 'estrato alto', 'estado civil con vínculo' y 'no realiza compras'. Este análisis proporciona un entendimiento más claro sobre qué variables categóricas son más significativas en cada una de las dimensiones estudiadas.



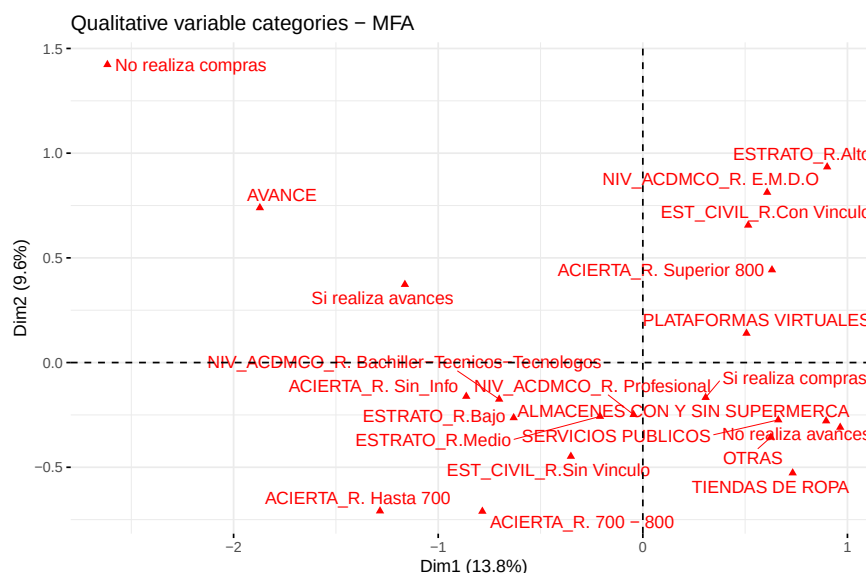
**Figura 5-9.:** Representación contribución variables categóricas Dim1



**Figura 5-10.:** Representación contribución variables categóricas Dim2

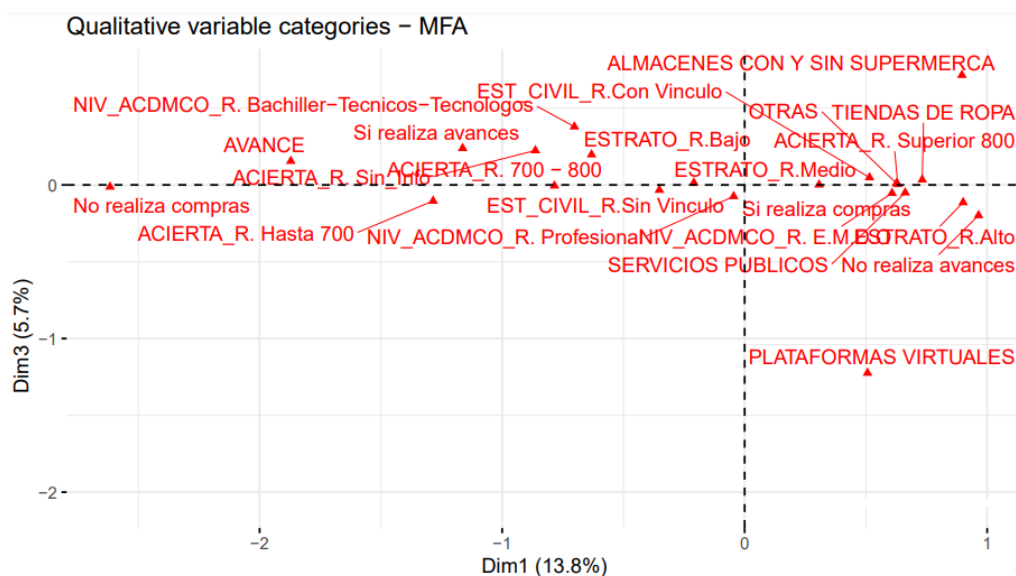
Las Figuras 5-11 y 5-12 revelan varias interconexiones interesantes entre diferentes categorías de variables. La primera figura muestra una asociación clara entre los estratos altos, niveles de estudios E.M.D.O, estado civil con vínculo y puntajes Acierta superiores a

800. Además, se nota que la categoría del MCC más usado para avances está vinculada con los individuos que no realizan compras.



**Figura 5-11.:** Relaciones entre Categorías Originales y Componentes 1 - 2

La Figura 5-12, por su parte, destaca la relación entre el nivel académico de bachiller técnico y tecnólogo, los puntajes Acierta de 700 y las personas que no realizan compras pero que optan por hacer avances. Estas conexiones nos proporcionan información valiosa sobre cómo se correlacionan distintas variables, lo cual es esencial para entender más profundamente el comportamiento de los clientes.



**Figura 5-12.:** Relaciones entre Categorías Originales y Componentes 1 - 3

### 5.2.2. Análisis de Conglomerados.

Para la clasificación de los tarjetahabientes, ante las variables sociodemográficas y transaccionales, se tomaron como entrada el score de los primeros diez factores que acumulan el 58.98 % de la variabilidad de los datos originales.

Luego se procede a realizar el análisis de clasificación, que acuerdo a la Figura 5-13, la selección del número de grupos se encuentra entre los valores del 4 al 6, ya que dentro de estos se encuentra el mayor grado de inclinación que se busca visualmente al detallar el gráfico.

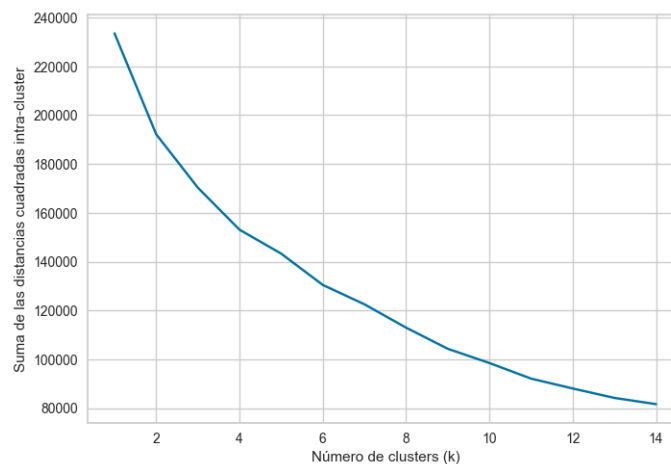
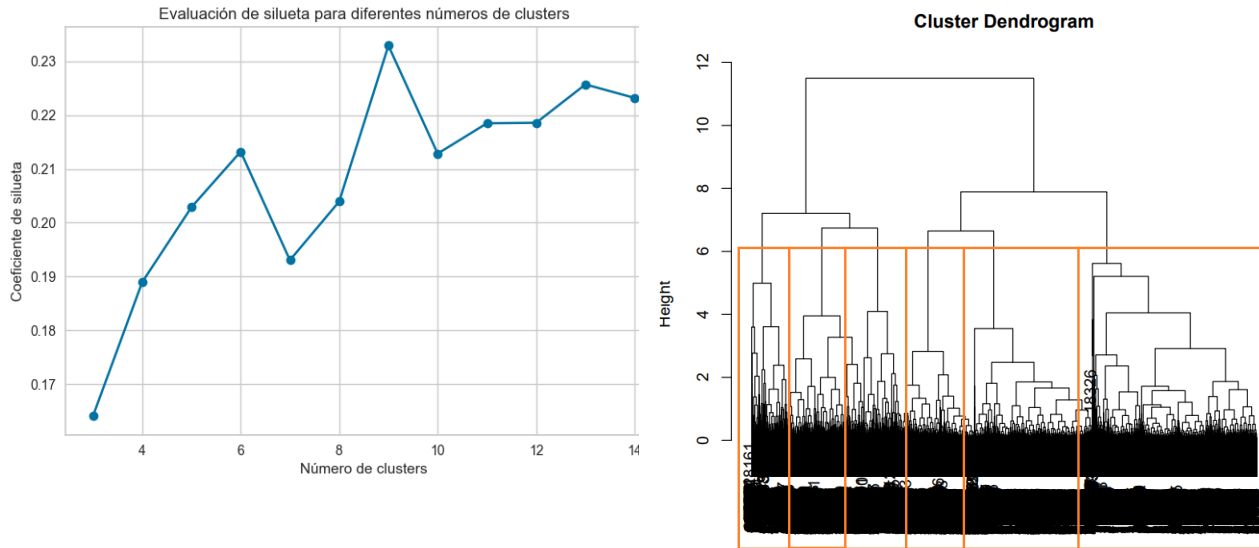


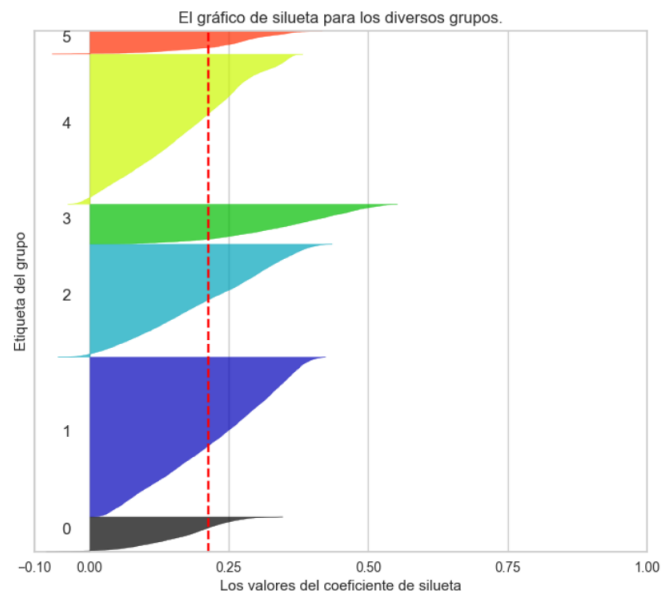
Figura 5-13.: Grafico Elbow.

En la Figura 5-14 se busca que el coeficiente de los grupos sea lo más cercano a 1, siendo el número de grupos con mayor coeficiente alcanzado el seis y el nueve. Como el aumento de este coeficiente es mínimo entre uno y otro, se toma el seis como valor de referencia para la selección del número de grupos. Para entrar en detalle se realizó el gráfico de la silueta para los 6 grupos e identificar la consistencia de los mismo.



**Figura 5-14.:** Coeficiente Silueta y Dendograma Via Ward

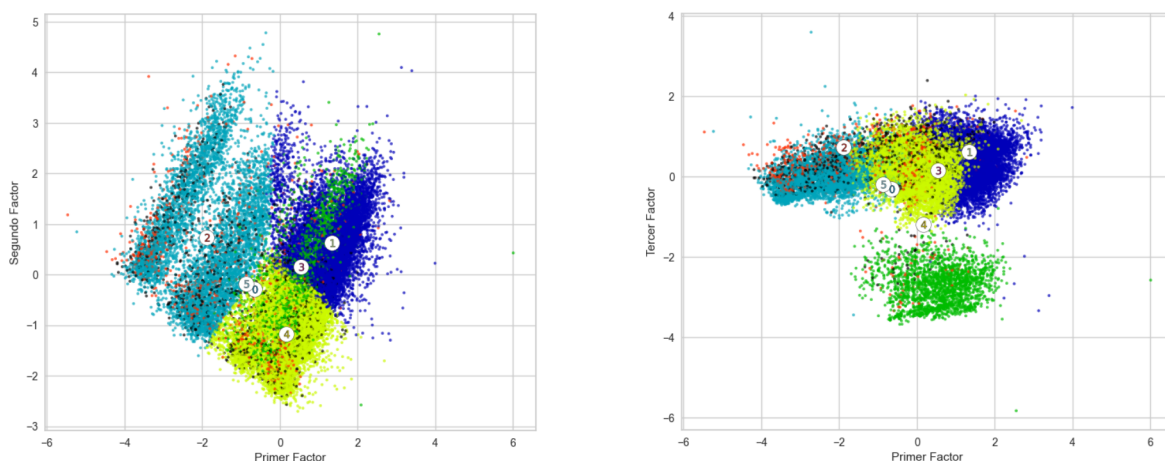
Basados en lo anterior, se tomarán 6 clústeres. Con esto, se pasa al análisis de la Figura 5-15, la cual nos detalla que el grupo que alcanzó el mayor coeficiente es el grupo 3, siendo este mayor a 0.5. Por el contrario, el grupo cero obtuvo un valor superior a 0.25. Esto explica la homogeneidad en la similitud de las características de los grupos.



**Figura 5-15.:** Coeficiente Silueta para los 6 grupos.

Se finaliza la ejecución con la clasificación del algoritmo *K*-means, el cual requiere el número de grupos que, en este caso, son 6 para realizar la segmentación. Esta consiste en ubicar

$K$  centroides en el conjunto de datos hasta obtener la partición que minimice la inercia intraclases.



**Figura 5-16.:** Proyección de la nube de puntos para los 6 grupos en los 3 primeros factores.

Los gráficos de la Figura 5-16 permiten identificar la conformación de 4 de los 6 grupos con mayor robustez, basados en los tres primeros factores del AFM. Los grupos 1, 2 y 4 se visualizan mejor con los primeros factores, y el grupo 3 se aprecia mejor cuando se gráfica reemplazando el segundo factor por el tercero.

### 5.2.3. Validación de la Estabilidad de los Grupos

Antes de proceder con la caracterización de los grupos, se lleva a cabo una comprobación de su estabilidad utilizando el indicador recomendado por la entidad en cuestión (PSI). Se espera que los valores de este indicador sean inferiores a 0,1 para asegurar una estabilidad adecuada. Tras realizar este análisis, se ha confirmado que los grupos efectivamente cumplen con este criterio y se encuentran dentro del margen de estabilidad deseado.

| Segmento | Enero   | Febrero | Marzo   | Abril   |
|----------|---------|---------|---------|---------|
| 1        | 0,00000 | 0,00000 | 0,00000 | 0,00001 |
| 2        | 0,00001 | 0,00011 | 0,00001 | 0,00001 |
| 3        | 0,00001 | 0,00002 | 0,00001 | 0,00003 |
| 4        | 0,00107 | 0,00155 | 0,00092 | 0,00185 |
| 5        | 0,00036 | 0,00071 | 0,00007 | 0,00000 |
| 6        | 0,00056 | 0,00065 | 0,00063 | 0,00028 |
| Total    | 0,00202 | 0,00305 | 0,00164 | 0,00218 |

**Tabla 5-6.:** Indicador de estabilidad de los grupos.

### 5.3. Caracterización de los Conglomerados y Descripción de Características

Para la caracterización de los grupos, que implica determinar qué variables tienen una mayor relación con uno de los grupos, se utilizan los valores de prueba, para las variables numéricas se realiza el cálculo de los cuantiles de la distribución normal estándar. Aquí, se calcula la media de la variable dentro del grupo y, si resulta ser superior a la media global, entonces la variable representa una característica del grupo en cuestión, en caso de las variables con categorías se realiza con la frecuencia de la categoría, se calcula la frecuencia de la categoría dentro del grupo y si su resultado es mayor que la frecuencia total, la categoría es una característica del grupo. Además, se lleva a cabo un análisis exploratorio que busca respaldar la prueba. Las tablas se pueden observar en el Anexo B.

#### Grupo 1

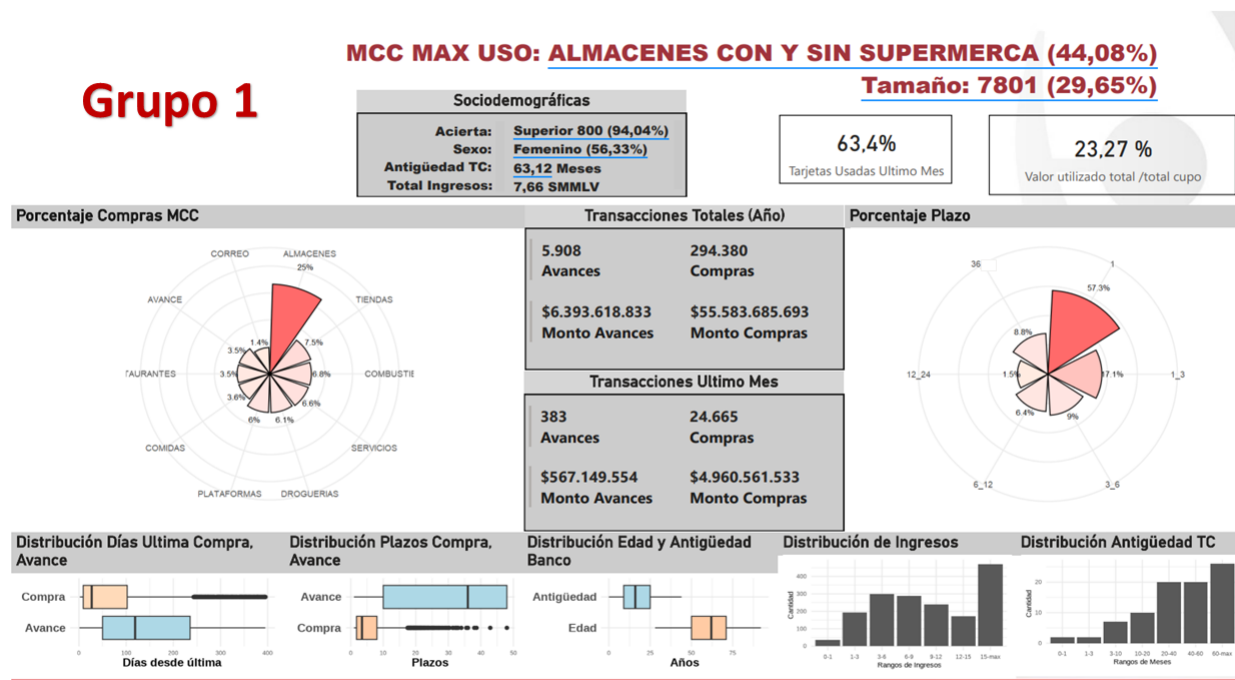


Figura 5-17.: Tablero Estadísticas Grupo 1.

El comportamiento transaccional, todos los miembros de este grupo se dedican principalmente a realizar compras con sus tarjetas, en lugar de avances de efectivo. De hecho, el 100 % de los integrantes ha efectuado al menos una compra con su tarjeta durante

el año. En orden de importancia, los lugares en los que realizan compras con mayor frecuencia son Almacenes (44.08 %), Otros (43.19 %) y Servicios Públicos (9.28 %). Por mes, realizan en promedio 3.16 compras, lo que asciende a un total aproximado de 37.74 compras en el transcurso del año. Suelen elegir opciones de pago a corto plazo, dejando el pago de las transacciones a 1 plazo, y de 1 a 3 plazos en el 65 % y 14 % de los casos respectivamente. En cuanto a los aspectos sociodemográficos, los miembros de este grupo pertenecen principalmente al estrato Alto, que constituye el 42.01 % del grupo. La mayoría están casados o en una relación, representando el 65.26 % del grupo. En términos de educación, un porcentaje considerable, el 38,87 %, tiene un nivel académico superior al profesional. Adicionalmente, se puede destacar que la edad promedio de los miembros del grupo es de 58 años. También muestran una antigüedad promedio con la entidad de 10 años. En términos de ingresos, cuentan con un total promedio de 7.6 veces el salario mínimo mensual legal vigente (smmlv) y un puntaje acierta superior a 800 puntos para el 94 %.

## Grupo 2

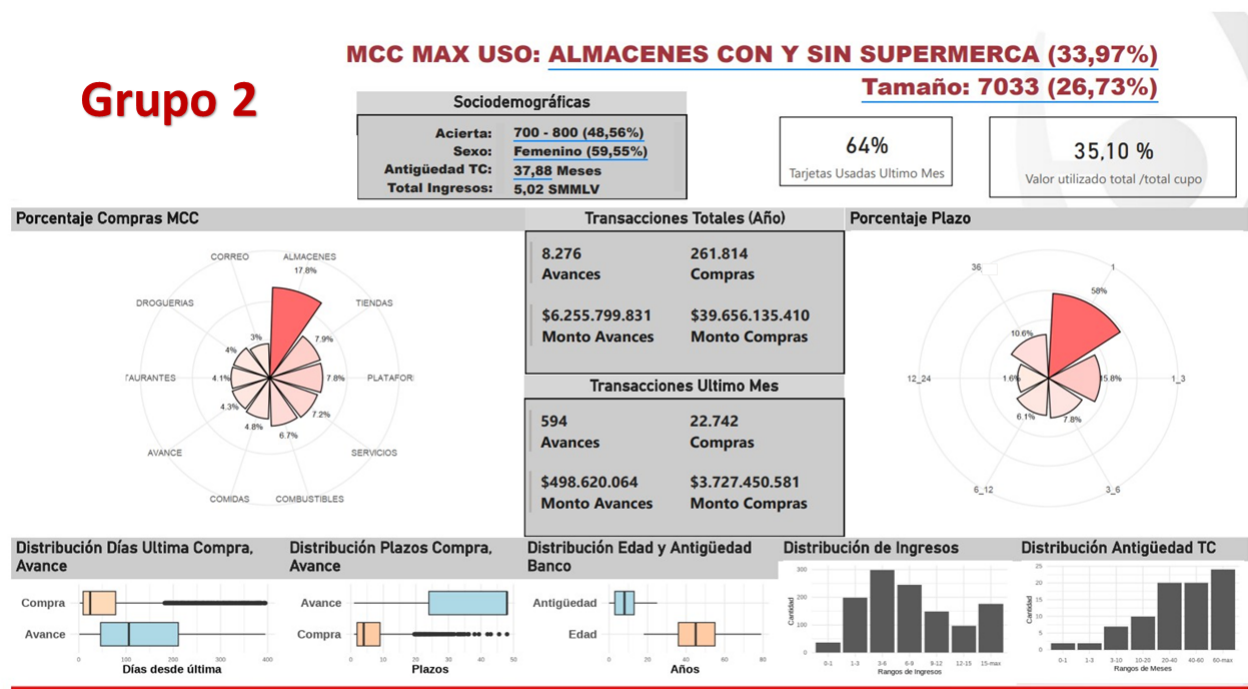


Figura 5-18.: Tablero Estadísticas Grupo 2.

De acuerdo con la actividad transaccional, todos los tarjetahabientes de este grupo se dedican principalmente a realizar compras con sus tarjetas. Siendo que el 100 % de los mismos ha

efectuado al menos una compra. Los lugares donde realiza con mayor frecuencia las compras, en orden de importancia, son Otros (52.37%), Almacenes (33.97%) y Servicios Públicos (10.89%). En promedio, realiza 3.23 compras al mes, lo que asciende a un total aproximado de 37.22 compras anuales. El valor promedio anual de estas compras es de 5.6 millones. Donde toma los plazos de pago a 1 y de 1 a 3 plazos, con un promedio del 61% y 14% respectivamente. Es de resaltar que el promedio de días desde su último avance es de 533 días, lo que indica un uso se restringe a solo realizar compras.

Desde sus variables sociodemográficas, los miembros de este grupo pertenecen a los estratos Medio y Bajo, representando el 82.94% y 10.25% respectivamente. Además, la mayoría están casados o tienen algún tipo de vínculo, siendo el 79,17% del grupo. En términos de educación, una proporción significativa, el 75.44%, tiene un nivel académico profesional. Además, la edad promedio de los miembros de este grupo es de 38 años y muestran una antigüedad promedio con la entidad de 5 años y un puntaje acierta entre 700 - 800 puntos para el 48.56%.

### Grupo 3

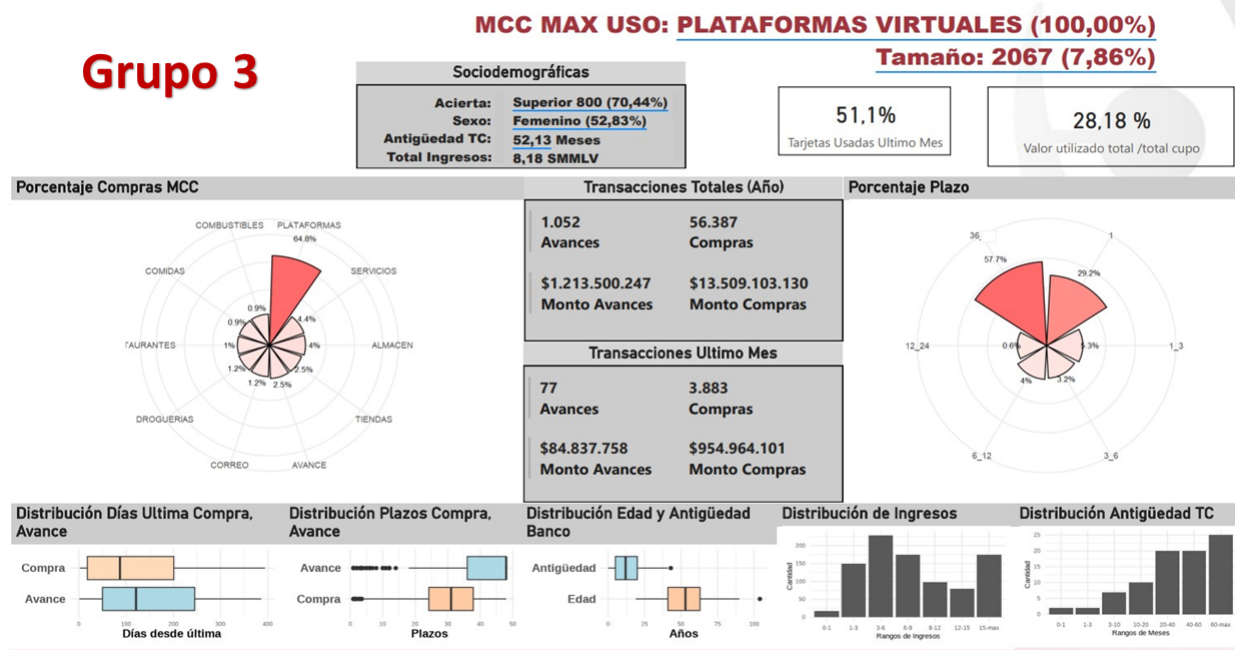


Figura 5-19.: Tablero Estadísticas Grupo 3.

El comportamiento transaccional de este grupo presenta que, el 100 %, realizan compras con sus tarjetas, mientras que el 84.47 % de ellos no realizan avances de efectivo. Los lugares donde realiza las compras con mayor frecuencia son las Plataformas Virtuales, siendo el 100 % del grupo. El valor promedio de estas compras es de 6.5 millones al año. Además, los plazos de pago elegidos con frecuencia son de 36 plazos, donde el 67 % de sus transacciones anuales se realizan en estos plazos. Resalta que el promedio de días desde su último avance es de 639 días. Es importante resaltar que el 74 % de sus transacciones anuales se realizan en plataformas virtuales, esto da a conocer una preferencia por el comercio electrónico.

Desde la perspectiva sociodemográfica, los miembros de este grupo pertenecen mayormente al estrato Alto, representando el 29.66 %. Además, una proporción significativa, el 45.48 %, están casados o tienen algún tipo de vínculo. El 27.38 % ha alcanzado el nivel de Educación superior a la profesional (E.M.D.O), mientras que el 62.75 % posee un nivel académico profesional. Además, la antigüedad promedio de los miembros de este grupo con la entidad es de 11 años. En términos de ingresos, cuentan con un total promedio de 8.17 veces el salario mínimo mensual legal vigente (smmlv) y un puntaje acierta superior a 800 puntos para el 70.44 %.

## Grupo 4

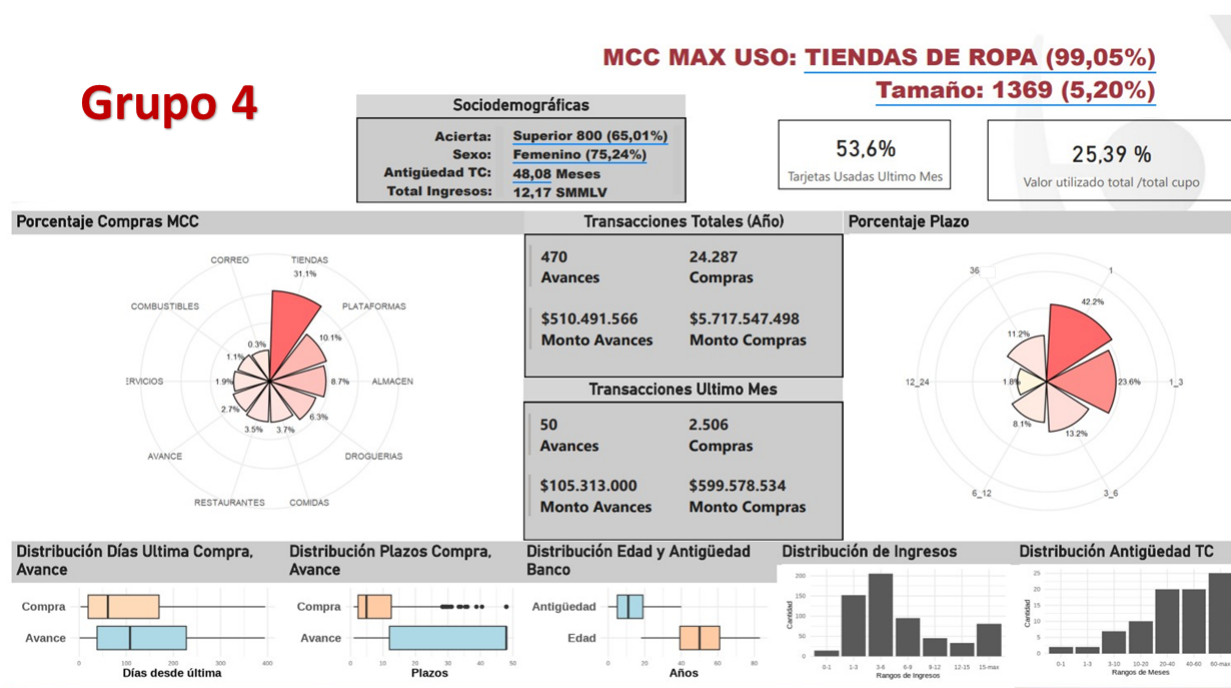


Figura 5-20.: Tablero Estadísticas Grupo 4.

En términos de comportamiento transaccional, el 99.93 % de tarjetahabientes, realizan compras con sus tarjetas, mientras que un 85.03 % de ellos no realizan avances de efectivo. Los lugares donde realiza compras con mayor frecuencia son las Tiendas de Ropa, que es el lugar de elección para el 99.05 % del grupo. Además, el 44 % de sus transacciones anuales se realizan en estas tiendas de ropa. Los plazos de pago, que eligen con frecuencia son de 1 a 3 plazos (22 %), 3 a 6 plazos (12 %) y 1 plazo (49 %). Resalta que el promedio de días desde su último avance es de 641 días, lo que indica un poco uso de esta actividad.

Una proporción significativa del grupo, el 43.90 %, está casada o tiene algún tipo de vínculo. En términos de educación, el 66.25 %, posee un nivel académico profesional, los ingresos tienen un total promedio de 12 millones y un puntaje acierta superior a 800 puntos para el 65.01 %. Esto sugiere un poder adquisitivo notable.

## Grupo 5

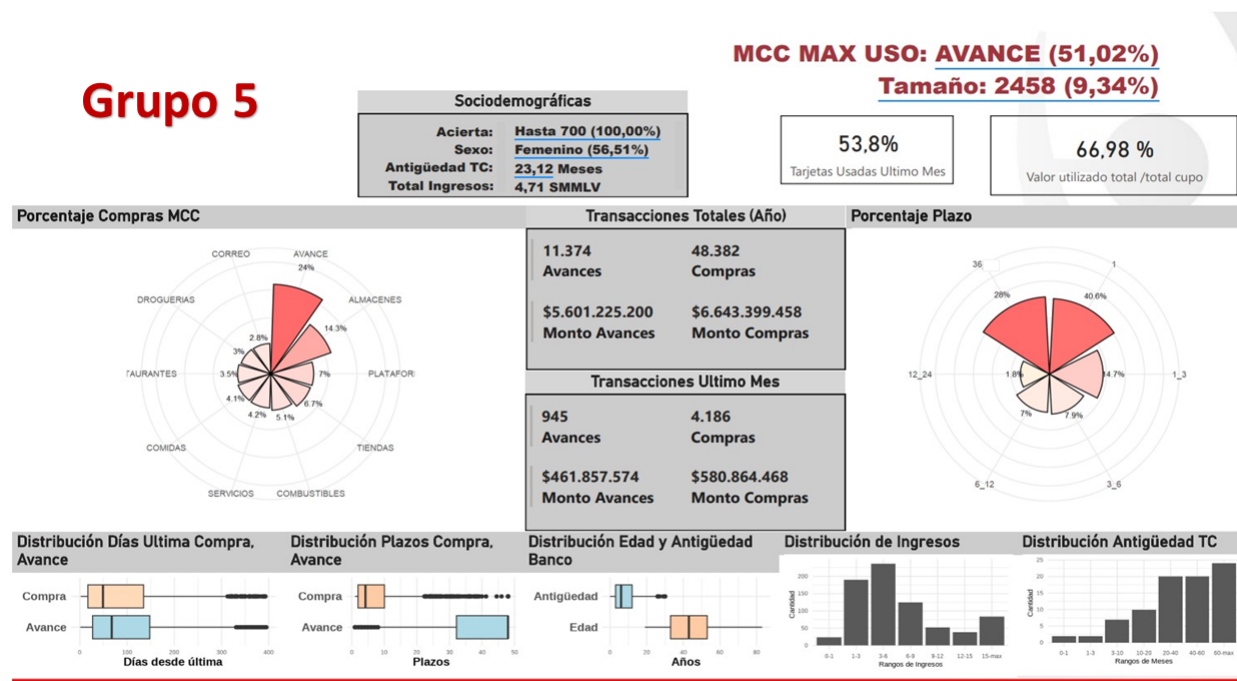


Figura 5-21.: Tablero Estadísticas Grupo 5.

En términos de comportamiento transaccional, un 70.91 % de los tarjetahabientes de este grupo realiza avances de efectivo, mientras que un 16.80 % no realiza compras. El total de avances realizados en el año equivale al 37 % de todas sus transacciones, con un valor promedio total de 2.2 millones. El número de avances en el año es de 4.62, y en el último

mes es de 0.3. En cuanto a los plazos de pago, la mayor frecuencia es de 36 plazos, siendo el 33 % de las transacciones las que se dejan a estos plazos. El promedio de días desde su última compra es de 182 días, esto indica un bajo uso de la tarjeta para compras.

Este grupo se caracteriza sociodemográficamente por no tener ningún tipo de vínculo 83.32 %. Los miembros de este grupo pertenecen mayormente a los estratos medio y bajo, representando el 83.65 % y 9.32 % respectivamente. El nivel educativo alcanzado por el grupo está dividido entre aquellos con formación de Bachiller-Técnicos-Tecnólogos (22.95 %) y aquellos con nivel académico profesional (66.48 %) y un puntaje acierta inferior a 700 puntos para el 100 %.

## Grupo 6

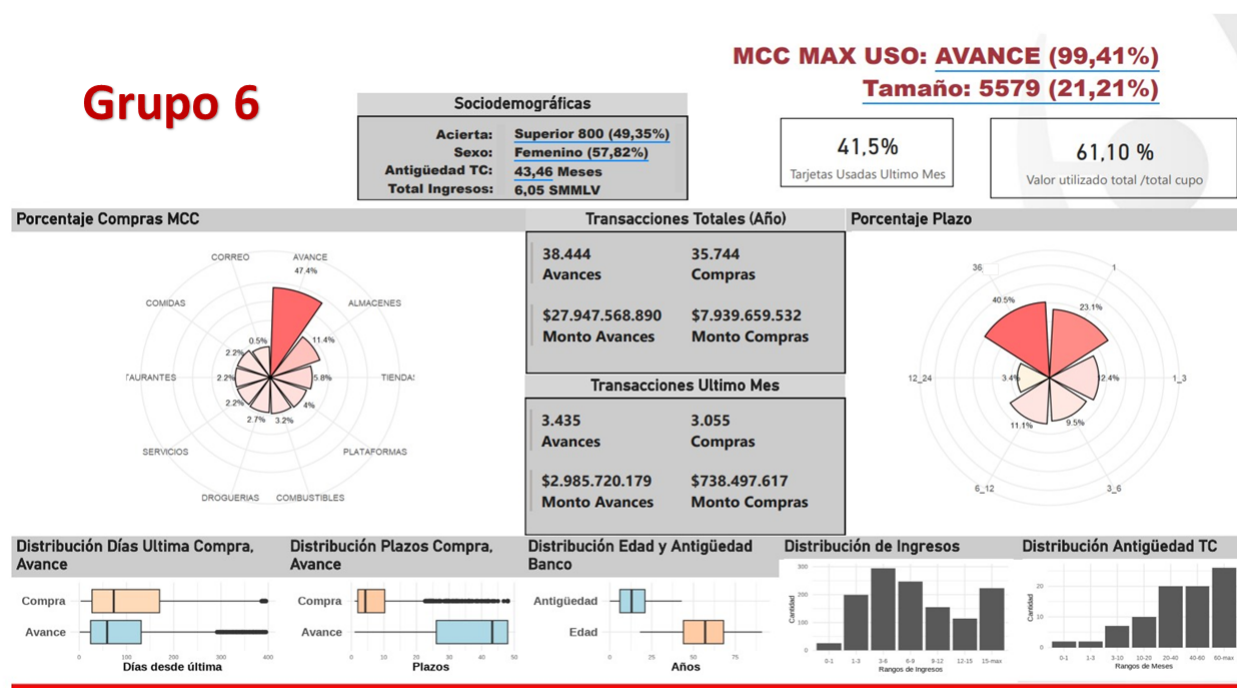


Figura 5-22.: Tablero Estadísticas Grupo 6.

El 100 % los tarjetahabientes de este grupo realizan avances de efectivo con sus tarjetas, mientras que un 42.09 % no realiza compras. El total de avances realizados en el año equivale al 71 % del total de sus transacciones, con un valor promedio de 5 millones. El número de avances en el año es de 6.8, y en el último mes es de 0.6. En cuanto a los plazos de pago, tiende a plazos considerablemente largos de 36 plazos (49 %) y de 3 a 6 plazos (7 %). El promedio de días desde su última compra es notablemente alto, con 365 días.

Desde la perspectiva sociodemográfica, el grupo se compone en su mayoría de individuos sin vínculo, constituyendo un 61.14 %. Este grupo pertenecen principalmente a los estratos medio y bajo, representando el 20.67 % y 9.52 % respectivamente. En términos de educación, un 20.67 % posee un nivel académico de Bachiller-Técnicos-Tecnólogos. Además, el grupo presenta una edad promedio de 47 años y un puntaje acierta entre a 700 - 800 puntos para el 40.65 %.

## Validación Operativa de los Grupos

Tras realizar la validación operativa de los grupos, se han obtenido importantes conclusiones que pueden orientar las futuras estrategias de marketing. A continuación, se presentan los que hasta el momento se contemplan como posibles campañas de marketing basadas en la información recopilada sobre los grupos identificados:

### Segmento Premium

Dado que este grupo se compone principalmente de individuos de estrato alto, con altos niveles de ingresos y educación, se puede desarrollar una campaña de marketing dirigida a productos de lujo y exclusividad.

**Estrategia:** Se pueden lanzar campañas publicitarias enfocadas en productos de alta gama, resaltando su calidad y exclusividad, y ofreciendo opciones de pago a corto plazo para adaptarse a las preferencias de pago del grupo.

### Segmento Digital

Dado el alto porcentaje de transacciones realizadas en plataformas virtuales y la preferencia por plazos de pago a largo plazo, se puede desarrollar una campaña de marketing enfocada en promover servicios y productos digitales, como suscripciones a servicios de streaming o compras en línea.

**Estrategia:** Se pueden crear campañas publicitarias que destaquen la conveniencia y seguridad de las compras en línea, ofreciendo descuentos especiales y opciones de financiamiento a largo plazo para incentivar las compras digitales.

### **Segmento de Moda**

Dado que este grupo muestra una preferencia por realizar compras en tiendas de ropa y utiliza principalmente plazos cortos de pago, se puede diseñar una campaña de marketing centrada en promover las últimas tendencias de moda y ofrecer opciones de financiamiento flexibles.

**Estrategia:** Se pueden lanzar campañas publicitarias que resalten la diversidad de productos de moda disponibles en la tienda, junto con promociones especiales y facilidades de pago atractivas para impulsar las ventas en este segmento.

### **Segmento Avanzado**

Dado que este grupo muestra un alto uso de avances de efectivo y plazos de pago largos, se puede desarrollar una campaña de marketing enfocada en promover servicios financieros avanzados, como inversiones y productos de ahorro a largo plazo.

**Estrategia:** Se pueden diseñar campañas publicitarias que destaquen los beneficios de planificar financieramente para el futuro, ofreciendo asesoramiento personalizado y soluciones adaptadas a las necesidades de inversión y ahorro del grupo.

## 6. Conclusiones

Para lograr una comprensión completa del comportamiento de los tarjetahabientes, es fundamental analizar diversas dimensiones. Estas van desde las preferencias generales en compras y avances de efectivo hasta la diversidad en factores financieros como puntajes de crédito y antigüedad con la entidad financiera. Cada dimensión aporta una perspectiva única y enriquecedora que, cuando se considera junto con las demás, ofrece un perfil más completo y detallado de los distintos grupos de usuarios. El objetivo es entender no solo los patrones de gasto, sino también cómo estos se relacionan con variables sociodemográficas y otros indicadores financieros. Al examinar estos comportamientos desde diferentes ángulos, podemos desarrollar estrategias de marketing más enfocadas y efectivas, que se alineen mejor con las necesidades y deseos específicos de cada grupo.

### **Preferencias Generales en Compras y Avances de Efectivo**

En general, la mayoría de los tarjetahabientes en todos los grupos muestran una preferencia por realizar compras en lugar de avances de efectivo. Sin embargo, es fundamental destacar que cada grupo tiene sus propias particularidades en términos de los lugares donde más frecuentemente realizan sus compras, la frecuencia de estas transacciones y los plazos de pago que eligen.

### **Vinculación con Características Sociodemográficas**

El comportamiento de compra está fuertemente relacionado con las características sociodemográficas de los diferentes grupos. Como ejemplo ilustrativo, el Grupo 1, predominantemente de estrato alto y con una edad promedio de 58 años, suele enfocar sus gastos en Almacenes y Servicios Públicos. En contraste, el Grupo 3, también de estrato alto pero con una clara inclinación hacia el comercio electrónico, efectúa la mayoría de sus transacciones en plataformas virtuales.

### **Diversidad en Factores Financieros**

Los grupos presentan una variabilidad considerable en factores como la edad promedio, la antigüedad con la entidad financiera y el puntaje de crédito. Estos factores son cruciales ya que podrían estar correlacionados con la cantidad de ingresos y, por ende, con la capacidad

de gasto. Por ejemplo, el Grupo 5, cuyo puntaje Acierta es inferior a 700 puntos para todos sus miembros, tiende a hacer un uso más moderado de sus tarjetas de crédito.

## **Variabilidad en Plazos de Pago**

Los plazos de pago elegidos también muestran una variabilidad significativa entre los grupos. Esta diversidad puede interpretarse como un reflejo de diferentes estrategias o enfoques de gestión financiera. Algunos grupos, por ejemplo, optan por plazos de pago más cortos, lo que podría indicar una mayor estabilidad financiera o una mayor aversión al riesgo.

## **Relación entre Nivel Educativo y Comportamiento Financiero**

Es evidente que existe una relación entre el nivel educativo y el comportamiento financiero de los grupos. Los que tienen un nivel más alto de educación tienden a mostrar un comportamiento más activo y diversificado en el uso de sus tarjetas. En este sentido, tanto el Grupo 2 como el Grupo 3, que cuentan con un alto porcentaje de miembros con educación superior, muestran una mayor actividad transaccional.

En resumen, la segmentación de los tarjetahabientes en estos grupos nos proporciona un mapa detallado de distintas conductas financieras y preferencias de consumo. Esta información es invaluable para entender mejor a los clientes y diseñar estrategias de marketing más efectivas y personalizadas.

## 7. Recomendaciones

- Asegurar que la información sociodemográfica de los clientes esté actualizada es crucial para la validez de los resultados. Factores como puntajes crediticios, estratos o ingresos pueden cambiar con el tiempo y, si están desactualizados, pueden llevar a fallas en el análisis que son difíciles de corregir posteriormente.
- Revisar la ponderación aplicada en el AFM es otro aspecto importante. El modelo actual pondera los grupos de variables para equilibrar su impacto en el análisis, pero no ajusta de la misma manera para los individuos. Esto podría resultar en que ciertos individuos ejerzan una influencia desproporcionada sobre los resultados.
- Abordar la presencia de valores atípicos que no se trataron en el análisis inicial es vital desde un punto de vista estadístico. Estos valores pueden influir significativamente en indicadores de calidad del grupo, como el valor de la silueta, y su omisión podría comprometer la precisión de las conclusiones del estudio.

## A. Anexo: Tablas Caracterizacion Grupos

| Variable          | Definición                                                                        |
|-------------------|-----------------------------------------------------------------------------------|
| Actividad Laboral | Independiente o Asalariado                                                        |
| Antigüedad        | Numero de años con vinculo en la entidad                                          |
| Edad              | Numero años cumplidos                                                             |
| Estrato           | Bajo-Bajo, Bajo, Medio-Bajo, Medio, Medio-Alto, Alto                              |
| Ingresos          | Ingresos en SMMVL                                                                 |
| Menores a cargo   | Numero de menores bajo obligación                                                 |
| Nivel educativo   | Especialización, maestría, doctorado<br>Pregrado Tecnico, Tecnologo, bachillerato |
| Regional          | Bogota, Cali, Caribe, Eje Cafetero, Medellin, Palmira                             |
| Sexo              | Femenino o Masculino                                                              |
| Estado Civil      | Casado, Union libre, Viudo, Soltero, Divorciado, Separado                         |

**Tabla A-1.:** Información de Individuos

### A.1. Variables transaccionales Formula

| Variable | Descripción                                                                |
|----------|----------------------------------------------------------------------------|
| CODMCC   | Código MCC del establecimiento donde se realizó la transacción.            |
| VRDESEMB | Valor desembolsado para llevar a cabo la transacción.                      |
| CATMCC   | Recategorización establecida por la entidad financiera para el código MCC. |
| PLAZO    | Número de cuotas asignadas para el pago de la transacción.                 |
| TASA     | Tasa de interés aplicada a la transacción.                                 |
| PERIODO  | Mes en el que se llevó a cabo la transacción.                              |
| VLRCUPO  | Valor del cupo disponible en la tarjeta de crédito.                        |
| VLRUTIL  | Valor total utilizado en el mes en transacciones con la tarjeta.           |
| PERCORTE | Contempla la información de 1 año de transacciones.                        |

**Tabla A-2.:** Registro detallado de Transacciones

### A.1.1. KPI'S año de cierre

Cada PERCORTE tiene la información de 1 año, todos los cálculos deben hacerse por PERCORTE.

- MCC más usado junto con su categoría: De la tabla FUENTE DE DATOS, identificar el CODMCC más utilizado por el cliente en un PERCORTE y extraer el campo CATBANCO.
- MCC de mayor monto junto con su categoría: De la tabla FUENTE DE DATOS, identificar el CODMCC con el mayor VRDESEMB por el cliente en un PERCORTE y extraer el campo CATBANCO.
- Cantidad de transacciones del MCC y categoría más usado: Sumar la cantidad de registros del CODMCC más utilizado por el cliente en un PERCORTE.
- Días desde el último uso del MCC y categoría más usado: En la tabla FUENTE DE DATOS, calcular la diferencia en días de la fecha de cierre versus FECDESEMB del último uso del CODMCC más usado.
- Suma del monto del MCC y categoría de mayor monto: Sumar el campo VRDESEMB del CODMCC más usado por el cliente en un PERCORTE.
- Participación del MCC y categorías más usados dentro del total de transacciones realizadas:

$$\frac{\sum \text{registros CODMCC más usado del cliente}}{\sum \text{registros totales del cliente}} \times 100$$

- Participación de los montos del MCC y categorías dentro del monto total de transacciones:

$$\frac{\sum \text{VRDESEMB CODMCC de mayor monto del cliente}}{\sum \text{VRDESEMB total del cliente}} \times 100$$

- Plazo promedio de las compras: Promedio del campo PLAZO en la tabla FUENTE DE DATOS, solo para transacciones con CODMCC diferente de 1.
- Cantidad de compras: En la tabla FUENTE DE DATOS, contar la cantidad de transacciones con CODMCC diferente de 1.
- Valor de compras: En la tabla FUENTE DE DATOS, sumar el campo VRDESEMB con CODMCC diferente de 1.
- Cantidad de avances: En la tabla FUENTE DE DATOS, contar la cantidad de transacciones con CODMCC = 1.

- Valor de avances: En la tabla FUENTE DE DATOS, sumar el campo VRDESEMB con CODMCC = 1.
- Tiempo en días desde la última compra: En la tabla FUENTE DE DATOS, calcular la diferencia en días de la fecha de cierre versus el FECDESEMB más reciente del CODMCC diferente de 1.
- Tiempo en días desde el último avance: En la tabla FUENTE DE DATOS, calcular la diferencia en días de la fecha de cierre versus el FECDESEMB más reciente del CODMCC = 1.
- Porcentaje de Uso promedio mensual de la TC en el último año: Para que sea del último año, se debe promediar los 12 PERIODOS del PERCORTE.

$$\frac{\sum \text{VRDESEMB del periodo}}{\text{VLRCUPO}} \times 100$$

- Tasa promedio de los usos en compras último año: Para que sea del último año, se debe calcular por CODMCC diferente de 1.

$$\frac{\text{TASA}}{12}$$

- Valor promedio por transacción: En la tabla FUENTE DE DATOS, calcular el promedio del campo VRDESEMB por PERCORTE.

- Proporción de transacciones con un plazo exacto de 1 mes.

$$\frac{\sum (PLAZO = 1)}{\text{Total de Transacciones}}$$

- Proporción de transacciones con un plazo mayor a 1 mes y hasta 3 meses.

$$\frac{\sum (1 > PLAZO \leq 3)}{\text{Total de Transacciones}}$$

- Proporción de transacciones con un plazo mayor a 3 meses y hasta 6 meses.

$$\frac{\sum (3 > PLAZO \leq 6)}{\text{Total de Transacciones}}$$

- Proporción de transacciones con un plazo mayor a 12 meses y hasta 24 meses.

$$\frac{\sum (12 > PLAZO \leq 24)}{\text{Total de Transacciones}}$$

- Proporción de transacciones con un plazo mayor a 36 meses.

$$\frac{\sum(36 > PLAZO \leq 60)}{\text{Total de Transacciones}}$$

- Proporción de transacciones categorizadas como 'ALMACENES CON Y SIN SUPERMERCA'.

$$\frac{\sum(CATBANCO = 'ALMACENES CON Y SIN SUPERMERCA')}{\text{Total de Transacciones}}$$

- Proporción de transacciones en 'PLATAFORMAS VIRTUALES'.

$$\frac{\sum(CATBANCO = 'PLATAFORMAS VIRTUALES')}{\text{Total de Transacciones}}$$

- Proporción de transacciones categorizadas como 'AVANCE'.

$$\frac{\sum(CATBANCO = 'AVANCE')}{\text{Total de Transacciones}}$$

- Proporción de transacciones en 'COMBUSTIBLES Y GASOLINA'.

$$\frac{\sum(CATBANCO = 'COMBUSTIBLES Y GASOLINA')}{\text{Total de Transacciones}}$$

- Proporción de transacciones en 'COMIDAS RÁPIDAS'.

$$\frac{\sum(CATBANCO = 'COMIDAS RÁPIDAS')}{\text{Total de Transacciones}}$$

- Proporción de transacciones en 'CORREO Y ENCOMIENDAS'.

$$\frac{\sum(CATBANCO = 'CORREO Y ENCOMIENDAS')}{\text{Total de Transacciones}}$$

- Proporción de transacciones en 'DROGUERÍAS'.

$$\frac{\sum(CATBANCO = 'DROGUERÍAS')}{\text{Total de Transacciones}}$$

### A.1.2. KPI'S mes de cierre

Campo periodo = mes cierre

- Indicador de uso de la tarjeta: En la tabla FUENTE DE DATOS, si en el periodo de cierre el VRDESEMB > 0 entonces 1, caso contrario 0.

- Cantidad de transacciones en el mes de ejecución: En la tabla FUENTE DE DATOS, contar la cantidad de transacciones realizadas en un mismo PERIODO.
- Monto de transacciones en el mes de ejecución: En la tabla FUENTE DE DATOS, sumar el campo VRDESEMB en un mismo PERIODO.
- Cantidad de compras en el mes de ejecución: En la tabla FUENTE DE DATOS, contar la cantidad de transacciones con CODMCC diferente de 1 en el mismo PERIODO.
- Monto de compras en el mes de ejecución: En la tabla FUENTE DE DATOS, sumar el campo VRDESEMB con CODMCC diferente de 1 en el mismo PERIODO.
- Cantidad de avances en el mes de ejecución: En la tabla FUENTE DE DATOS, contar la cantidad de transacciones con CODMCC = 1 en el mismo PERIODO.
- Monto de avances en el mes de ejecución: En la tabla FUENTE DE DATOS, sumar el campo VRDESEMB con CODMCC = 1 en el mismo PERIODO.
- Valor usado en el mes de ejecución: En la tabla FUENTE DE DATOS, sumar el VRDESEMB del mismo PERIODO.
- Porcentaje de Uso mensual:

$$\frac{\sum \text{VRDESEMB del periodo}}{\text{VLRCUPO}} \times 100$$

| <b>Variables Transacciones Anuales</b> | <b>Etiqueta</b> | <b>Tipo</b> |
|----------------------------------------|-----------------|-------------|
| Días máximo usado                      | D.MU            | Numérica    |
| Días última compra                     | D.UC            | Numérica    |
| Días último avance                     | D.UA            | Numérica    |
| Monto avance último mes                | M.AUM           | Numérica    |
| Monto compra último mes                | M.CUM           | Numérica    |
| Monto transacciones último mes         | M.TUM           | Numérica    |
| Número avances                         | N.A             | Numérica    |
| Número avances último mes              | N.AUM           | Numérica    |
| Número compras último mes              | N.CUM           | Numérica    |
| Número compras                         | N.C             | Numérica    |
| Número transacciones último mes        | N.TUM           | Numérica    |
| Plazo promedio compras                 | P.PC            | Numérica    |
| Porcentaje MCC máximo usado            | P.MMU           | Numérica    |
| Porcentaje MCC mayor monto             | P.MMM           | Numérica    |
| Porcentaje uso mes                     | P.UM            | Numérica    |
| Porcentaje uso promedio mes            | P.UPM           | Numérica    |
| Porcentaje almacenes                   | P.ALM           | Numérica    |
| Porcentaje avance                      | P.AV            | Numérica    |
| Porcentaje combustibles                | P.COM           | Numérica    |
| Porcentaje comidas                     | P.CO            | Numérica    |
| Porcentaje correo y encomiendas        | P.COR           | Numérica    |
| Porcentaje droguerías                  | P.DRO           | Numérica    |
| Porcentaje plataformas                 | P.PLAT          | Numérica    |
| Porcentaje plazo 1                     | P.PL1           | Numérica    |
| Porcentaje plazo 1-3                   | P.PL13          | Numérica    |
| Porcentaje plazo 12-24                 | P.PL1224        | Numérica    |
| Porcentaje plazo 3-6                   | P.PL36          | Numérica    |
| Porcentaje plazo 36-60                 | P.PL3660        | Numérica    |
| Porcentaje restaurantes                | P.REST          | Numérica    |
| Porcentaje servicios públicos          | P.SERV          | Numérica    |
| Porcentaje tiendas                     | P.TIEN          | Numérica    |
| Cantidad MCC máximo usado              | C.MU            | Numérica    |
| Valor avances                          | V.A             | Numérica    |
| Valor compras                          | V.C             | Numérica    |
| Valor uso último mes                   | V.UM            | Numérica    |
| Indicador uso TC                       | I.UTC           | Categorica  |
| MCC máximo usado                       | MCC.MU          | Categorica  |
| MCC mayor monto                        | MCC.MM          | Categorica  |

**Tabla A-3.:** Variables creadas con un año de Transacciones (Individuo -Año)

# B. Anexo: Tablas Caracterizacion Grupos

## B.1. Grupo 1

| Variable        | Categoria                     | Cla/Mod | Mod/Cla | Global | v,test |
|-----------------|-------------------------------|---------|---------|--------|--------|
| Puntaje Acierta | Superior 800                  | 46,87   | 94,03   | 59,48  | Inf    |
| Realiza Avances | No                            | 41,74   | 76,94   | 54,66  | Inf    |
| Realiza Compras | Si                            | 33,13   | 100,00  | 89,50  | Inf    |
| MCC Max Usado   | Almacenes y SuperM            | 54,87   | 44,08   | 23,82  | Inf    |
| Estrato         | Alto                          | 57,60   | 42,01   | 21,63  | Inf    |
| Nivel Academico | E,M,D,O                       | 51,56   | 38,87   | 22,35  | Inf    |
| Estado Civil    | Con Vinculo                   | 47,69   | 65,26   | 40,58  | Inf    |
| MCC Max Usado   | Otros                         | 43,72   | 43,19   | 29,29  | 31,57  |
| MCC Max Usado   | Servicios Publicos            | 45,34   | 9,28    | 6,07   | 13,66  |
| Nivel Academico | Bachiller-Tecnicos-Tecnologos | 20,80   | 10,76   | 15,33  | -13,77 |
| Estrato         | Bajo                          | 12,00   | 2,95    | 7,29   | -19,00 |
| Nivel Academico | Profesional                   | 23,97   | 50,38   | 62,31  | -25,74 |
| MCC Max Usado   | Tiendas de Ropa               | 0,65    | 0,12    | 5,24   | -30,06 |
| Estrato         | Medio                         | 22,96   | 55,04   | 71,09  | -36,52 |
| Puntaje Acierta | Hasta 700                     | 3,18    | 1,16    | 10,85  | -Inf   |
| Puntaje Acierta | 700 - 800                     | 3,48    | 2,96    | 25,22  | -Inf   |
| Realiza Avances | Si                            | 15,08   | 23,06   | 45,34  | -Inf   |
| Realiza Compras | No                            | 0,00    | 0,00    | 10,50  | -Inf   |
| MCC Max Usado   | Plataformas Virtuales         | 0,09    | 0,03    | 8,01   | -Inf   |
| MCC Max Usado   | Avances                       | 3,56    | 3,31    | 27,57  | -Inf   |
| Estado Civil    | Sin Vinculo                   | 17,34   | 34,74   | 59,42  | -Inf   |

Tabla B-1.: Caracterización Cate  rias Grupo 1

| Variable             | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|----------------------|--------|------------------|--------------|----------------|-------------|
| Edad                 | 83.55  | 58.64            | 47.18        | 10.89          | 14.45       |
| Antigüedad Bancaria  | 83.19  | 17.19            | 10.25        | 8.00           | 8.79        |
| P.PL1                | 52.30  | 0.65             | 0.47         | 0.33           | 0.37        |
| D.UA                 | 48.34  | 593.74           | 448.20       | 254.48         | 317.06      |
| P.ALMM               | 47.89  | 0.24             | 0.15         | 0.24           | 0.20        |
| N.C                  | 23.50  | 37.74            | 27.41        | 55.37          | 46.29       |
| N.CUM                | 18.99  | 3.16             | 2.32         | 5.49           | 4.67        |
| V.C                  | 18.81  | 7125200.06       | 4905520.61   | 11850163.80    | 12429697.37 |
| P.COM                | 17.41  | 0.06             | 0.04         | 0.14           | 0.11        |
| P.PL13               | 9.12   | 0.14             | 0.12         | 0.21           | 0.19        |
| P.REST               | 8.67   | 0.03             | 0.02         | 0.07           | 0.07        |
| Total Ingreso SMMMLV | 6.14   | 7.66             | 6.61         | 9.32           | 17.95       |
| P.CO                 | 5.17   | 0.03             | 0.03         | 0.06           | 0.06        |
| P.PL36               | 1.30   | 0.08             | 0.07         | 0.16           | 0.16        |
| P.UM                 | -5.90  | 9.53             | 10.68        | 17.34          | 20.42       |
| P.TIEN               | -6.58  | 0.06             | 0.07         | 0.10           | 0.14        |
| V.A                  | -20.33 | 819589.65        | 1821652.21   | 3264955.47     | 5190293.42  |
| N.AUM                | -22.08 | 0.05             | 0.21         | 0.27           | 0.76        |
| P.PLAT               | -24.76 | 0.05             | 0.11         | 0.11           | 0.22        |
| N.A                  | -33.51 | 0.76             | 2.49         | 2.25           | 5.45        |
| D.UC                 | -36.64 | 59.29            | 136.13       | 89.20          | 220.86      |
| P.AV                 | -55.42 | 0.03             | 0.21         | 0.08           | 0.34        |
| P.PL3660             | -56.13 | 0.06             | 0.24         | 0.12           | 0.32        |

Tabla B-2.: Caracterización Numérica Grupo 1

## B.2. Grupo 2

| Variable        | Categoria                     | Cla/Mod | Mod/Cla | Global | v.test |
|-----------------|-------------------------------|---------|---------|--------|--------|
| Puntaje Acierta | 700 - 800                     | 51.47   | 48.56   | 25.22  | Inf    |
| Realiza Compras | Si                            | 29.87   | 100.00  | 89.50  | Inf    |
| MCC Max Usado   | Otras                         | 47.79   | 52.37   | 29.29  | Inf    |
| Estado Civil    | Sin Vinculo                   | 35.62   | 79.17   | 59.42  | Inf    |
| Nivel Academico | Profesional                   | 32.37   | 75.44   | 62.31  | 27.11  |
| Estrato         | Medio                         | 31.19   | 82.94   | 71.09  | 26.52  |
| Realiza Avances | No                            | 33.05   | 67.57   | 54.66  | 25.65  |
| MCC Max Usado   | Almacenes con y sin SuperM    | 38.12   | 33.97   | 23.82  | 22.78  |
| MCC Max Usado   | Servicios Publicos            | 47.96   | 10.89   | 6.07   | 18.73  |
| Estrato         | Bajo                          | 37.61   | 10.25   | 7.29   | 10.82  |
| Nivel Academico | Bachiller-Tecnicos-Tecnologos | 26.60   | 15.26   | 15.33  | -0.21  |
| Realiza Avances | Si                            | 19.12   | 32.43   | 45.34  | -25.65 |
| Puntaje Acierta | Superior 800                  | 20.54   | 45.71   | 59.49  | -27.32 |
| MCC Max Usado   | Tiendas de Ropa               | 0.15    | 0.03    | 5.24   | -29.22 |
| Nivel Academico | E.M.D.O                       | 11.12   | 9.30    | 22.35  | -32.79 |
| MCC Max Usado   | Plataformas Virtuales         | 0.09    | 0.03    | 8.01   | -36.65 |
| Realiza Compras | No                            | 0.00    | 0.00    | 10.50  | -Inf   |
| MCC Max Usado   | Avances                       | 2.63    | 2.72    | 27.57  | -Inf   |
| Estrato         | Alto                          | 8.42    | 6.81    | 21.63  | -Inf   |
| Puntaje Acierta | Hasta 700                     | 0.00    | 0.00    | 10.86  | -Inf   |
| Estado Civil    | Con Vinculo                   | 13.72   | 20.83   | 40.58  | -Inf   |

Tabla B-3.: Caracterización Cate  rias Cluster 2

| Variable             | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|----------------------|--------|------------------|--------------|----------------|-------------|
| Edad                 | -60.99 | 38.18            | 47.17        | 8.95           | 14.44       |
| Antigüedad Bancaria  | -58.31 | 5.01             | 10.24        | 4.48           | 8.78        |
| P.PL1                | 39.01  | 0.61             | 0.46         | 0.32           | 0.36        |
| D.UA                 | 26.47  | 533.87           | 448.19       | 289.24         | 317.06      |
| P.CO                 | 23.22  | 0.04             | 0.02         | 0.08           | 0.06        |
| N.C                  | 20.78  | 37.22            | 27.40        | 52.98          | 46.29       |
| N.CUM                | 19.17  | 3.23             | 2.32         | 5.42           | 4.66        |
| P.REST               | 16.05  | 0.03             | 0.02         | 0.08           | 0.06        |
| P.ALM                | 15.71  | 0.17             | 0.14         | 0.20           | 0.19        |
| P.COM                | 15.56  | 0.05             | 0.04         | 0.13           | 0.11        |
| P.PL13               | 10.44  | 0.14             | 0.12         | 0.20           | 0.19        |
| P.UM                 | 8.54   | 12.45            | 10.67        | 20.49          | 20.42       |
| V.C                  | 5.77   | 5638580.32       | 4905520.61   | 7706577.24     | 12429697.36 |
| P.PL36               | 1.07   | 0.07             | 0.07         | 0.16           | 0.16        |
| P.TIEN               | -3.16  | 0.06             | 0.07         | 0.09           | 0.14        |
| Total Ingreso SMMVLV | -8.70  | 5.01             | 6.61         | 5.48           | 17.95       |
| N.AUM                | -15.98 | 0.08             | 0.20         | 0.39           | 0.75        |
| P.PLAT               | -17.29 | 0.06             | 0.10         | 0.11           | 0.22        |
| V.A                  | -17.59 | 889492.36        | 1821652.20   | 2871171.09     | 5190293.41  |
| N.A                  | -23.63 | 1.17             | 2.49         | 2.72           | 5.44        |
| D.UC                 | -37.16 | 52.35            | 136.12       | 78.89          | 220.85      |
| P.PL3660             | -45.91 | 0.08             | 0.23         | 0.13           | 0.32        |
| P.AV                 | -49.84 | 0.03             | 0.20         | 0.07           | 0.33        |

Tabla B-4.: Caracterización Numérica Grupo 2

## B.3. Grupo 3

| Variable        | Categoria             | Cla/Mod | Mod/Cla | Global | v.test |
|-----------------|-----------------------|---------|---------|--------|--------|
| MCC Max Usado   | Plataformas Virtuales | 98.10   | 100.00  | 8.01   | Inf    |
| Realiza Avances | No                    | 12.14   | 84.47   | 54.66  | 29.91  |
| Realiza Compras | Si                    | 8.78    | 100.00  | 89.50  | 21.73  |
| Puntaje Acierta | Superior 800          | 9.30    | 70.44   | 59.49  | 10.75  |
| Estrato         | Alto                  | 10.78   | 29.66   | 21.63  | 8.92   |
| Nivel Academico | E.M.D.O               | 9.63    | 27.38   | 22.35  | 5.60   |
| Estado Civil    | Con Vinculo           | 8.81    | 45.48   | 40.58  | 4.70   |
| Nivel Academico | Profesional           | 7.91    | 62.75   | 62.31  | 0.42   |
| Estrato         | Bajo                  | 5.84    | 5.42    | 7.29   | -3.53  |
| Estado Civil    | Sin Vinculo           | 7.21    | 54.52   | 59.42  | -4.70  |
| Puntaje Acierta | 700 - 800             | 6.18    | 19.84   | 25.22  | -6.01  |
| Estrato         | Medio                 | 7.18    | 64.93   | 71.09  | -6.33  |
| Puntaje Acierta | Hasta 700             | 4.69    | 6.48    | 10.86  | -7.09  |
| Nivel Academico | B.T.T                 | 5.06    | 9.87    | 15.33  | -7.56  |
| MCC Max Usado   | Tiendas de Ropa       | 0.00    | 0.00    | 5.24   | -15.03 |
| MCC Max Usado   | Servicios Publicos    | 0.00    | 0.00    | 6.07   | -16.25 |
| Realiza Compras | No                    | 0.00    | 0.00    | 10.50  | -21.73 |
| Realiza Avances | Si                    | 2.69    | 15.53   | 45.34  | -29.91 |
| MCC Max Usado   | Almacenes y SuperM    | 0.00    | 0.00    | 23.82  | -34.23 |
| MCC Max Usado   | Avances               | 0.00    | 0.00    | 27.57  | -37.30 |
| MCC Max Usado   | Otros                 | 0.00    | 0.00    | 29.29  | -Inf   |

Tabla B-5.: Caracterización Cate gor as Grupo 3

| Variable             | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|----------------------|--------|------------------|--------------|----------------|-------------|
| P.PLAT               | 136.39 | 0.74             | 0.10         | 0.20           | 0.22        |
| P.PL3660             | 65.07  | 0.67             | 0.23         | 0.28           | 0.32        |
| D.UA                 | 28.60  | 639.67           | 448.19       | 215.43         | 317.06      |
| V.C                  | 6.21   | 6535608.67       | 4905520.61   | 32995367.40    | 12429697.36 |
| Antig edad Bancaria  | 5.70   | 11.30            | 10.24        | 8.55           | 8.78        |
| Total Ingreso SMMMLV | 4.12   | 8.17             | 6.61         | 12.00          | 17.95       |
| Edad                 | 0.84   | 47.43            | 47.17        | 13.84          | 14.44       |
| N.C                  | -0.13  | 27.27            | 27.40        | 42.20          | 46.29       |
| N.CUM                | -4.48  | 1.87             | 2.32         | 4.35           | 4.66        |
| P.UM                 | -10.17 | 6.29             | 10.67        | 16.46          | 20.42       |
| N.AUM                | -10.67 | 0.03             | 0.20         | 0.32           | 0.75        |
| V.A                  | -11.26 | 587082.84        | 1821652.20   | 3211965.53     | 5190293.41  |
| D.UC                 | -11.77 | 81.20            | 136.12       | 100.76         | 220.85      |
| P.REST               | -13.39 | 0.00             | 0.02         | 0.02           | 0.06        |
| P.COM                | -14.99 | 0.00             | 0.04         | 0.02           | 0.11        |
| P.PL36               | -15.09 | 0.02             | 0.07         | 0.05           | 0.16        |
| P.CO                 | -15.29 | 0.00             | 0.02         | 0.02           | 0.06        |
| N.A                  | -17.23 | 0.50             | 2.49         | 2.03           | 5.44        |
| P.TIEN               | -19.19 | 0.01             | 0.07         | 0.03           | 0.14        |
| P.PL13               | -21.64 | 0.03             | 0.12         | 0.07           | 0.19        |
| P.AV                 | -26.55 | 0.01             | 0.20         | 0.05           | 0.33        |
| P.ALM                | -28.82 | 0.02             | 0.14         | 0.05           | 0.19        |
| P.PL1                | -29.81 | 0.23             | 0.46         | 0.27           | 0.36        |

Tabla B-6.: Caracterizaci n N merica Grupo 3

## B.4. Grupo 4

| Variable        | Categoria             | Cla/Mod | Mod/Cla | Global | v.test |
|-----------------|-----------------------|---------|---------|--------|--------|
| MCC Max Usado   | Tiendas de Ropa       | 98.40   | 99.05   | 5.24   | Inf    |
| Realiza Avances | No                    | 8.10    | 85.03   | 54.66  | 24.53  |
| Realiza Compras | Si                    | 5.81    | 99.93   | 89.50  | 17.20  |
| Puntaje Acierta | Superior 800          | 5.69    | 65.01   | 59.49  | 4.30   |
| Nivel Academico | Profesional           | 5.53    | 66.25   | 62.31  | 3.11   |
| Estado Civil    | Con Vinculo           | 5.63    | 43.90   | 40.58  | 2.56   |
| Estrato         | Alto                  | 5.27    | 21.91   | 21.63  | 0.27   |
| Estrato         | Medio                 | 5.21    | 71.22   | 71.09  | 0.11   |
| Estrato         | Bajo                  | 4.90    | 6.87    | 7.29   | -0.61  |
| Nivel Academico | B.T.T                 | 4.98    | 14.68   | 15.33  | -0.68  |
| Puntaje Acierta | 700 - 800             | 4.69    | 22.72   | 25.22  | -2.21  |
| Estado Civil    | Sin Vinculo           | 4.91    | 56.10   | 59.42  | -2.56  |
| Nivel Academico | E.M.D.O               | 4.44    | 19.07   | 22.35  | -3.04  |
| Puntaje Acierta | Hasta 700             | 3.43    | 7.16    | 10.86  | -4.75  |
| MCC Max Usado   | Servicios Publicos    | 0.00    | 0.00    | 6.07   | -13.06 |
| MCC Max Usado   | Plataformas Virtuales | 0.05    | 0.07    | 8.01   | -14.81 |
| Realiza Compras | No                    | 0.04    | 0.07    | 10.50  | -17.20 |
| Realiza Avances | Si                    | 1.72    | 14.97   | 45.34  | -24.53 |
| MCC Max Usado   | Almacenes y SuperM    | 0.06    | 0.29    | 23.82  | -26.80 |
| MCC Max Usado   | Avances               | 0.04    | 0.22    | 27.57  | -29.49 |
| MCC Max Usado   | Otros                 | 0.06    | 0.37    | 29.29  | -30.30 |

Tabla B-7.: Caracterización Categrías Grupo 4

| Variable             | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|----------------------|--------|------------------|--------------|----------------|-------------|
| P.TIEN               | 100.55 | 0.44             | 0.07         | 0.29           | 0.14        |
| D.UA                 | 23.21  | 641.87           | 448.19       | 215.10         | 317.06      |
| P.PL13               | 20.06  | 0.22             | 0.12         | 0.25           | 0.19        |
| Total Ingreso SMMVLV | 11.77  | 12.17            | 6.61         | 69.81          | 17.95       |
| P.PL36               | 11.02  | 0.12             | 0.07         | 0.21           | 0.16        |
| P.PL1                | 2.66   | 0.49             | 0.46         | 0.35           | 0.36        |
| P.REST               | 0.81   | 0.02             | 0.02         | 0.06           | 0.06        |
| Antigüedad Bancaria  | 0.19   | 10.28            | 10.24        | 8.59           | 8.78        |
| P.CO                 | -0.19  | 0.02             | 0.02         | 0.06           | 0.06        |
| P.UM                 | -1.18  | 10.04            | 10.67        | 19.28          | 20.42       |
| V.C                  | -2.22  | 4176440.83       | 4905520.61   | 7149631.61     | 12429697.36 |
| Edad                 | -2.64  | 46.17            | 47.17        | 13.64          | 14.44       |
| N.CUM                | -3.98  | 1.83             | 2.32         | 2.86           | 4.66        |
| P.PLAT               | -4.90  | 0.07             | 0.10         | 0.14           | 0.22        |
| N.C                  | -7.93  | 17.74            | 27.40        | 24.99          | 46.29       |
| N.AUM                | -8.59  | 0.03             | 0.20         | 0.21           | 0.75        |
| V.A                  | -10.60 | 372893.76        | 1821652.20   | 1763983.33     | 5190293.41  |
| D.UC                 | -11.16 | 71.21            | 136.12       | 97.36          | 220.85      |
| P.COM                | -11.29 | 0.00             | 0.04         | 0.02           | 0.11        |
| N.A                  | -14.98 | 0.34             | 2.49         | 1.28           | 5.44        |
| P.ALM                | -15.48 | 0.06             | 0.14         | 0.08           | 0.19        |
| P.PL3660             | -18.45 | 0.07             | 0.23         | 0.14           | 0.32        |
| P.AV                 | -21.11 | 0.02             | 0.20         | 0.06           | 0.33        |

Tabla B-8.: Caracterización Numérica Grupo 4

## B.5. Grupo 5

| Variable        | Categoria             | Cla/Mod | Mod/Cla | Global | v.test |
|-----------------|-----------------------|---------|---------|--------|--------|
| Puntaje Acierta | Hasta 700             | 86.06   | 100.00  | 10.86  | Inf    |
| Realiza Avances | Si                    | 14.61   | 70.91   | 45.34  | 26.90  |
| Estado Civil    | Sin Vinculo           | 13.10   | 83.32   | 59.42  | 26.73  |
| MCC Max Usado   | Avances               | 17.29   | 51.02   | 27.57  | 25.94  |
| Estrato         | Medio                 | 10.99   | 83.65   | 71.09  | 15.16  |
| Nivel Academico | B.T.T                 | 13.98   | 22.95   | 15.33  | 10.45  |
| Realiza Compras | No                    | 14.95   | 16.80   | 10.50  | 10.03  |
| Nivel Academico | Profesional           | 9.97    | 66.48   | 62.31  | 4.50   |
| Estrato         | Bajo                  | 11.95   | 9.32    | 7.29   | 3.93   |
| MCC Max Usado   | Otros                 | 8.21    | 25.75   | 29.29  | -4.09  |
| MCC Max Usado   | Servicios Publicos    | 6.57    | 4.27    | 6.07   | -4.09  |
| MCC Max Usado   | Almacenes y SuperM    | 6.70    | 17.09   | 23.82  | -8.51  |
| Realiza Compras | Si                    | 8.69    | 83.20   | 89.50  | -10.03 |
| MCC Max Usado   | Tiendas de Ropa       | 0.80    | 0.45    | 5.24   | -14.03 |
| MCC Max Usado   | Plataformas Virtuales | 1.66    | 1.42    | 8.01   | -15.09 |
| Nivel Academico | E.M.D.O               | 4.42    | 10.58   | 22.35  | -15.85 |
| Estrato         | Alto                  | 3.04    | 7.04    | 21.63  | -20.55 |
| Estado Civil    | Con Vinculo           | 3.84    | 16.68   | 40.58  | -26.73 |
| Realiza Avances | No                    | 4.97    | 29.09   | 54.66  | -26.90 |
| Puntaje Acierta | Superior 800          | 0.00    | 0.00    | 59.49  | -Inf   |
| Puntaje Acierta | 700 - 800             | 0.00    | 0.00    | 25.22  | -Inf   |

Tabla B-9.: Caracterización Cate gor as Grupo 5

| Variable            | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|---------------------|--------|------------------|--------------|----------------|-------------|
| P.AV                | 25.58  | 0.37             | 0.20         | 0.38           | 0.33        |
| N.A                 | 20.42  | 4.62             | 2.49         | 6.79           | 5.44        |
| P.PL3660            | 16.71  | 0.33             | 0.23         | 0.34           | 0.32        |
| N.AUM               | 12.05  | 0.38             | 0.20         | 1.01           | 0.75        |
| D.UC                | 10.88  | 182.27           | 136.12       | 258.98         | 220.85      |
| P.UM                | 6.11   | 13.07            | 10.67        | 23.71          | 20.42       |
| V.A                 | 4.58   | 2278773.47       | 1821652.20   | 3835055.99     | 5190293.41  |
| P.CO                | 2.75   | 0.02             | 0.02         | 0.06           | 0.06        |
| P.REST              | 1.19   | 0.02             | 0.02         | 0.06           | 0.06        |
| P.PL13              | -0.76  | 0.11             | 0.12         | 0.17           | 0.19        |
| P.COM               | -1.21  | 0.03             | 0.04         | 0.09           | 0.11        |
| P.PL36              | -1.67  | 0.06             | 0.07         | 0.14           | 0.16        |
| Total Ingreso SMMLV | -5.52  | 4.70             | 6.61         | 6.24           | 17.95       |
| N.CUM               | -6.88  | 1.70             | 2.32         | 3.56           | 4.66        |
| P.ALM               | -7.79  | 0.11             | 0.14         | 0.16           | 0.19        |
| P.TIEN              | -8.38  | 0.05             | 0.07         | 0.08           | 0.14        |
| N.C                 | -8.68  | 19.68            | 27.40        | 33.88          | 46.29       |
| V.C                 | -9.22  | 2702766.25       | 4905520.61   | 4326149.78     | 12429697.36 |
| P.PLAT              | -12.62 | 0.05             | 0.10         | 0.11           | 0.22        |
| P.PL1               | -14.30 | 0.36             | 0.46         | 0.33           | 0.36        |
| D.UA                | -27.71 | 279.46           | 448.19       | 299.04         | 317.06      |
| Antigüedad Bancaria | -39.86 | 3.51             | 10.24        | 4.56           | 8.78        |
| Edad                | -43.28 | 35.16            | 47.17        | 10.34          | 14.44       |

Tabla B-10.: Caracterización N merica Grupo 5

## B.6. Grupo 6

| Variable        | Categoria             | Cla/Mod | Mod/Cla | Global | v.test |
|-----------------|-----------------------|---------|---------|--------|--------|
| Realiza Avances | Si                    | 46.77   | 100.00  | 45.34  |        |
| Realiza Compras | No                    | 85.01   | 42.09   | 10.50  | Inf    |
| MCC Max Usado   | Avances               | 76.48   | 99.41   | 27.57  | Inf    |
| Puntaje Acierta | 700 - 800             | 34.18   | 40.65   | 25.22  | Inf    |
| Nivel Academico | B.T.T                 | 28.58   | 20.67   | 15.33  | 12.09  |
| Estrato         | Medio                 | 22.46   | 75.30   | 71.09  | 7.91   |
| Estrato         | Bajo                  | 27.70   | 9.52    | 7.29   | 7.01   |
| Estado Civil    | Sin Vinculo           | 21.82   | 61.14   | 59.42  | 2.95   |
| Estado Civil    | Con Vinculo           | 20.31   | 38.86   | 40.58  | -2.95  |
| Nivel Academico | Profesional           | 20.25   | 59.49   | 62.31  | -4.89  |
| Nivel Academico | E.M.D.O               | 18.83   | 19.84   | 22.35  | -5.12  |
| Estrato         | Alto                  | 14.89   | 15.18   | 21.63  | -13.58 |
| Puntaje Acierta | Superior 800          | 17.59   | 49.35   | 59.49  | -17.27 |
| MCC Max Usado   | Tiendas de Ropa       | 0.00    | 0.00    | 5.24   | -25.89 |
| MCC Max Usado   | Servicios Publicos    | 0.13    | 0.04    | 6.07   | -27.54 |
| Puntaje Acierta | Hasta 700             | 2.63    | 1.34    | 10.86  | -30.54 |
| MCC Max Usado   | Plataformas Virtuales | 0.00    | 0.00    | 8.01   | -32.33 |
| Realiza Avances | No                    | 0.00    | 0.00    | 54.66  | -Inf   |
| Realiza Compras | Si                    | 13.72   | 57.91   | 89.50  | -Inf   |
| MCC Max Usado   | Otros                 | 0.21    | 0.29    | 29.29  | -Inf   |
| MCC Max Usado   | Almacenes y SuperM    | 0.24    | 0.27    | 23.82  | -Inf   |

Tabla B-11.: Caracterización Categoriías Grupo 6

| Variable             | v.test | Mean in category | Overall mean | sd in category | Overall sd  |
|----------------------|--------|------------------|--------------|----------------|-------------|
| P.AV                 | 126.62 | 0.71             | 0.20         | 0.28           | 0.33        |
| D.UC                 | 87.24  | 365.11           | 136.12       | 320.57         | 220.85      |
| N.A                  | 67.97  | 6.89             | 2.49         | 8.41           | 5.44        |
| P.PL3660             | 67.71  | 0.49             | 0.23         | 0.37           | 0.32        |
| V.A                  | 51.67  | 5009422.63       | 1821652.20   | 8765345.42     | 5190293.41  |
| N.AUM                | 45.09  | 0.61             | 0.20         | 1.29           | 0.75        |
| Edad                 | 4.39   | 47.93            | 47.17        | 14.25          | 14.44       |
| P.PL36               | 2.52   | 0.07             | 0.07         | 0.17           | 0.16        |
| P.UM                 | 0.32   | 10.75            | 10.67        | 23.68          | 20.42       |
| Total Ingreso SMMMLV | -2.62  | 6.05             | 6.61         | 8.88           | 17.95       |
| Antigüedad Bancaria  | -5.28  | 9.69             | 10.24        | 8.50           | 8.78        |
| P.PL13               | -17.61 | 0.08             | 0.12         | 0.16           | 0.19        |
| P.COM                | -19.43 | 0.01             | 0.04         | 0.04           | 0.11        |
| P.REST               | -19.54 | 0.00             | 0.02         | 0.03           | 0.06        |
| P.CO                 | -22.70 | 0.00             | 0.02         | 0.03           | 0.06        |
| V.C                  | -23.57 | 1423133.09       | 4905520.61   | 3192615.68     | 12429697.36 |
| P.TIEN               | -25.24 | 0.03             | 0.07         | 0.07           | 0.14        |
| P.PLAT               | -31.74 | 0.02             | 0.10         | 0.07           | 0.22        |
| N.CUM                | -31.96 | 0.54             | 2.32         | 1.79           | 4.66        |
| P.ALM                | -37.57 | 0.05             | 0.14         | 0.10           | 0.19        |
| N.C                  | -38.17 | 6.40             | 27.40        | 14.09          | 46.29       |
| P.PL1                | -72.31 | 0.15             | 0.46         | 0.22           | 0.36        |
| D.UA                 | -94.38 | 92.56            | 448.19       | 96.88          | 317.06      |

Tabla B-12.: Caracterización Numérica Grupo 6

# Bibliografía

- Abascal Fernández, E. and Landaluce Calvo, M. I. (2002). Análisis factorial múltiple como técnica de estudio de la estabilidad de los resultados de un análisis de componentes principales.
- Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews: computational statistics*, 5(2):149–179.
- Aldás Manzano, J. and Uriel Jiménez, E. (2017). *Análisis multivariante aplicado con R*. Ediciones Paraninfo, SA.
- Barbosa Ariza, M. A. and González Jiménez, D. M. (2018). Perfiles de tarjetahabientes para fortalecer estrategias de marketing en una entidad financiera.
- Becerra, A. M. O. (2010). Comparación del análisis factorial múltiple y el análisis de componentes principales para datos cualitativos (prinqual), en la construcción de índices.
- Carter, R. (2023). ¿Qué es el código de categoría de comerciante? Tu guía definitiva para 2023 - Plataformas de comercio electrónico.
- Castellanos, S. (2004). La segmentación, una estrategia clave para obtener mayor rentabilidad en los negocios.
- Castro Puche, L. and Díz Garcés, V. (2014). *Determinación del mercado de tarjetas de crédito en Colombia (2002-2012)*. Universidad de Cartagena.
- Continuum (2018). tarjetahabiente-cardholder. <https://glosario.continuum.cl/glosario/tarjetahabiente-cardholder>.
- Díaz-Monroy, L. (2002). *Estadística multivariada: Inferencia y métodos [Multivariate statistics: Inference and methods]*. Bogotá, Colombia: Universidad Nacional de Colombia.
- Du, Y., Kamakura, W. A., and Mela, C. F. (2010). Segmentation using rfm analysis and k-means clustering for a spanish bank. *Expert Systems with Applications*, 37(9):6222–6228.
- Escofier, B. and Pagès, J. (1992). *Análisis Factoriales Simples y Múltiples: Objetivos, Métodos e Interpretación*. Universidad del País Vasco.

- Francois, H., Josse, J., Le, S., and Mazet, J. (2013). Multivariate exploratory data analysis and data mining with r. Versión 1.25, Páginas 29–33.
- Gallardo, J. (2011). Métodos jerárquicos de análisis clúster: Capítulo 3. *Curso de diplomatura estadística teórico práctico, Universidad de Granada*.
- Garitano (2023). Segmentación de Clientes RFM. *Indaru*.
- Gil, S. (2016). Tarjeta de crédito. <https://economipedia.com/definiciones/tarjeta-de-credito.html>. Accessed: 2023-4-8.
- Gonzalez, V., Àvila, E., and Olaya, J. (2009). Análisis de los espacios de representación del statis y del afm en el estudio de movilidad biográfica en bogotá 1993. *Revista Colombiana de Estadística*, 32(1):1–15.
- Gurrea, M. (2000). Análisis de componentes principales. *Proyecto e-Math Financiado por la Secretaría de Estado de Educación y Universidades (MECD)*.
- Hernández Meneses, I. (2015). Caracterización de los municipios de colombia basada en variables socioeconómicas. *Universidad del Valle, Facultad de Ingeniería, Escuela de Estadística*.
- Investopedia (2023). What is credit scoring? purpose, factors, and role in lending. [https://www.investopedia.com/terms/c/credit\\_scoring.asp](https://www.investopedia.com/terms/c/credit_scoring.asp).
- Jiménez, A. M. L. (2004). Análisis de conglomerados (cluster analysis). Diapositivas de clase.
- Kachigan, S. K. (1982). *Multivariate statistical analysis: A conceptual introduction*. Radius Press, New York, NY.
- Kagan, J. (2023). Cash Advance: Definition, Types, and Impact on Credit Score. *Investopedia*.
- Kassambara, A. and Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.
- Lebart, L., Morineau, A., and Piron, M. (1994). *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris.
- Li, W., Wu, X., Sun, Y., and Zhang, Q. (2010). Credit card customer segmentation and target marketing based on data mining. In *2010 International Conference on Computational Intelligence and Security*, pages 73–76.
- Lê, S., Josse, J., and Husson, F. (2008). Factominer: A package for multivariate analysis.

- Muñoz, A. (2018). Análisis de correspondencias múltiples en presencia de datos faltantes: el principio de datos disponibles del algoritmo nipals (acmpdd). Master's thesis.
- Pagès, J. (2004). Análisis factorial múltiple: principales características y su aplicación a datos sensoriales. *Revista Colombiana de Estadística*, 27(1):1–26.
- Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw Hill, Interamericana de España.
- R5 (2023). Puntaje crediticio. <https://www.grupor5.com/aprende/finanzas/puntaje-crediticio>.
- Raj, S., Roy, S., Jana, S., Roy, S., Goto, T., and Sen, S. (2023). Customer segmentation using credit card data analysis. In *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 383–388. IEEE.
- Rodríguez, B. (2014). Origen y evolución histórica de las tarjetas de crédito. *gestiopolis*.
- Struyf, A., Hubert, M., and Rousseeuw, P. J. (1997). Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):1–30.
- Tripathi, A. (2018). Approaches to clustering in customer segmentation. *International Journal of Business and Management*, 13(6):1–10.
- Umuhoza, J., Ntirushwamaboko, G., Awuah, G., and Birir, D. (2020). Using unsupervised machine learning techniques for behavioral-based credit card users segmentation in africa. *Journal of Big Data*, 7(1):1–22.