

# Trabajo de Investigación - Minería de Datos

Elaborado por: Diego Gomez, Ana Rodrigo

Junio, 2024

## Contents

<b>1</b>	<b>Introducción</b>	<b>3</b>
<b>2</b>	<b>Objetivos</b>	<b>3</b>
2.1	Objetivo General . . . . .	3
2.2	Objetivos Específicos . . . . .	3
<b>3</b>	<b>Motivación</b>	<b>4</b>
<b>4</b>	<b>Marco Teórico</b>	<b>4</b>
4.1	TECNICAS DE SEGMENTACIÓN Y CLUSTERING . . . . .	4
4.2	Seguimiento de Ventas . . . . .	5
4.3	Concentración de Ingresos . . . . .	5
4.4	Importancia de la Gestión de Metas e Indicadores . . . . .	5
4.5	Importancia de selección de variables . . . . .	6
<b>5</b>	<b>Metodología</b>	<b>6</b>
5.1	Descripción de la base datos . . . . .	6
5.2	Análisis de Componentes Principales . . . . .	7
5.3	Análisis de clusters K-medias . . . . .	7
5.4	Tratamiento sobre la base datos . . . . .	7
<b>6</b>	<b>Resultados y Análisis</b>	<b>7</b>
6.1	Preparación . . . . .	7
6.2	Datos incompletos . . . . .	8
6.3	Datos atípicos . . . . .	10
6.4	Análisis de componentes principales . . . . .	12
6.5	CLUSTER . . . . .	14
<b>7</b>	<b>Conclusiones</b>	<b>18</b>

<b>8 Recomendaciones</b>	<b>19</b>
<b>9 Referencias</b>	<b>19</b>

# 1 Introducción

El crédito es una operación financiera que permite a una persona acceder a una cantidad de dinero, hasta un límite especificado, durante un período de tiempo determinado. El crédito es una forma de financiación flexible que permite acceder al dinero prestado según las necesidades (Pérez López, Moya Fernández y Trigo Sierra, 2012).

El seguro es el instrumento a través del cual las empresas o personas transfieren riesgos a un tercero, empresa aseguradora, y en caso de siniestro tienen la obligación de indemnizar total o parcialmente sus pérdidas, dependiendo del contexto.

Las entidades aseguradoras otorgan cobertura a riesgos que podrían afectar a bienes, patrimonio o a individuos. Como contraprestación de ello requieren del pago de un monto (prima) y en el caso que suceda un determinado evento que ocasione una pérdida (siniestro) la aseguradora pagará el monto acordado (indemnización), cuyo riesgo de ocurrencia es objeto de cobertura por el seguro. La actividad del mercado inicia con la necesidad que se crea de los clientes a partir de la percepción de un riesgo y para el cual contrata la póliza de seguro de una compañía (García Lomas, 2018)

En entidades financieras al momento de otorgar un crédito se realiza la afiliación a diferentes seguros lo cual hace que exista un vínculo cercano entre el financiamiento y acceder a un seguro.

Durante el año 2021, el mercado asegurador boliviano experimentó un desempeño interesante. Aunque las primas directas se mantuvieron casi al mismo nivel que en 2020, hubo un aumento significativo en los pagos por siniestros, superando en más del 55% los montos del año anterior. Esto tuvo un fuerte impacto en el índice de siniestralidad anual, que alcanzó un récord histórico del 80%. Sin embargo, en 2022, se observó una notable disminución, llegando al 53%. Asimismo, para el 2023 existió mejora en los indicadores financieros del mercado asegurador con la reducción de la tasa de siniestralidad e incremento de la producción.

Sin duda, la pandemia ha puesto de manifiesto la importancia crucial de los seguros en nuestra sociedad. Su función fundamental es preservar la integridad de las personas y proteger su patrimonio, especialmente en áreas afectadas por la crisis, como la salud, la vida y el desempleo. La necesidad de reducir la exposición al riesgo se ha vuelto más evidente que nunca, y esto podría beneficiar a largo plazo el modelo de negocio de las compañías de seguros y contribuir a su desarrollo.

Al respecto, se observa un mercado potencial, por lo que los créditos son una fuente de gestión de venta de seguros tanto de personas como patrimoniales.

## 2 Objetivos

### 2.1 Objetivo General

- Identificar los patrones de comportamientos mediante el análisis de k-medias de los clientes con créditos de seguros del Banco PK.

### 2.2 Objetivos Específicos

- Relizar la preparación de los datos para así contar con una base sin datos perdidos ni atípicos.
- Relizar el análisis de componentes a las variables numéricas.
- Determinar el número óptimo de clusters para la base de datos acorde a la comparación del uso de todas las variables o de las más representativas en función al análisis de componentes.
- Crear un dashboard que presente de manera clara y visual los clusters identificados.

### **3 Motivación**

El Banco PK, al contar con una cantidad importante de clientes con canales de alto impacto, pretende conocer el comportamiento de los clientes para poder incrementar los ingresos por la comercialización de seguros.

Se advierte que el volumen de la cartera del Banco PK, créditos PYME, microcréditos, vivienda y consumo dirigido a personas naturales con cobertura del seguro, tuvo crecimiento en las gestiones 2017-2020 y una reducción al cierre de 2021 como resultado de la contracción económica debido a la pandemia, ya que en esta etapa se otorgó el diferimiento y reprogramación de créditos, pero para la gestión 2022 y 2023 se observa una recuperación de la cartera en riesgo con un crecimiento menor en valor absoluto de la gestión 2020.

La cartera susceptible del seguro está conformada por créditos PYMES, Microcréditos, Hipotecarios, Productivos y de Consumo, los cuales de manera individual han tenido un crecimiento continuo, situación favorable para el desarrollo de nuevos productos de seguros adicionales tomando en cuenta el potencial de la cartera (Parrales Ramos, 2013).

Las personas naturales que tienen créditos representan un grupo susceptible para la venta de seguros. Cuando alguien adquiere un préstamo, su interés en proteger su inversión y su capacidad de pago aumenta. Por lo tanto, existe una oportunidad para ofrecerles productos de seguros que cubran riesgos como enfermedades, accidentes o pérdida de empleo.

Además debemos considerar que la cartera de seguros ya está constituida, lo que significa que hay una base de clientes potenciales. A medida que aumenta la colocación de créditos, también se incrementa la demanda de seguros. Esto se debe a que las personas buscan proteger sus activos y garantizar la continuidad de sus compromisos financieros en caso de imprevistos.

En el entorno actual, la gestión efectiva de metas e indicadores es fundamental para el éxito en la industria financiera y aseguradora. Identificar la clasificación de clientes para su gestión comercial es predominante con el objetivo de alcanzar las metas financieras, por lo que desarrollar un dashboard específico para la entidad financiera permitirá visualizar y analizar de manera eficiente los tipos de clientes relacionados con la comercialización de seguros. Esto facilitará la toma de decisiones informadas y la identificación de áreas de mejora (Bazan, 2015).

Se ha observado que los clientes de seguros son variables cada mes por lo que se pretende mantener una tasa de crecimiento constante de acuerdo con las características de los tipos de clientes y a qué tipo de créditos aplican para evaluar en siguientes periodos los tipos de seguros a comercializar por canal, tipo de crédito y cliente. En ese sentido, se pretende generar un dashboard para otorgar la información oportunamente al área ejecutiva para la toma de decisiones entre el banco y aseguradora con acceso para los principales ejecutivos del área comercial (Celi Yanangomez, 2023).

### **4 Marco Teórico**

La clasificación de clientes, a través del seguimiento de ventas y la concentración de ingresos son elementos cruciales para el éxito de cualquier organización comercial. Este marco teórico explora los fundamentos esenciales de estos conceptos y cómo su comprensión y aplicación pueden contribuir a una gestión empresarial efectiva.

#### **4.1 TECNICAS DE SEGMENTACIÓN Y CLUSTERING**

La segmentación de clientes consiste en un conjunto de acciones por medio de las cuales se seleccionan aspectos en común de la base de clientes y se los divide en diferentes grupos: cada uno de ellos es un segmento de clientes (Marín, 2006).

Sin embargo, clusterizar, y muchas veces se confunde o no se ve la diferencia con segmentación, es descubrir grupos de clientes similares, a través de variaciones similares dentro de cada grupo. Todo a través de modelos matemáticos (Celi Yanangomez, 2023). Básicamente la diferencia es que en la segmentación sólo agrupa en función de criterios de negocio, los cuales, una vez dejan de ser relevantes para definir ese grupo, pueden hacer menos efectiva esa segmentación. Con un cluster, al utilizar criterios matemáticos, está constantemente actualizando los grupos, haciéndolos mucho más efectivos de cara a las posteriores acciones que desarrolles con ellos. Hoy día, este método es utilizado sobre todo en la informática, el marketing, el mundo empresarial y el artístico, y es tremadamente efectivo.

Utilizar una segmentación a través de clusters, en una estrategia de fidelización e incentivación en cualquier marca, puede aportarte las siguientes ventajas: - Identificar patrones de comportamiento como, por ejemplo, qué compra, cuánto compra, cómo compra, qué factores influye en X situación, en qué momento consume o utiliza nuestro servicio, etc. - Conocer el perfil de cliente por patrones de consumo homogéneo: por niveles de gasto, CLV (Customer Time Value), localización, hábitos de compra, variables socio demográfico, edad, etc. - Crear y desarrollar campañas y promociones Ad-hoc y políticas comerciales acorde al comportamiento que reflejan los datos, mejorando la eficiencia de las mismas. - Priorizar clientes: focalizar acciones concretas de fidelización en clientes de mayor valor. - Captaciones de nuevos clientes (idénticos): desarrollar acciones de captación orientadas al target con las mismas características que el cluster. - Redirigir los flujos de clientes hacia clusters de mayor valor. Es una manera incluso de desarrollar técnicas de Up-selling. - Aumentar la retención: la predicción de clusters susceptibles al abandono permite desarrollar acciones enfocadas a reducir la tasa de abandono.

## 4.2 Seguimiento de Ventas

El seguimiento de ventas implica el monitoreo sistemático de todas las actividades relacionadas con el proceso de ventas de seguros. Desde la generación de leads hasta el cierre de negocios, cada etapa se analiza para evaluar el rendimiento y mejorar la eficacia de las estrategias de ventas. - Optimización del Proceso: El seguimiento de ventas permite identificar cuellos de botella y optimizar cada etapa del proceso de ventas para mejorar la eficiencia y aumentar la productividad en la venta de seguros. - Toma de Decisiones Basada en Datos: La recopilación y análisis de datos de ventas proporciona información valiosa para la toma de decisiones informada, lo que permite ajustar estrategias según sea necesario. - Mejora Continua: Al comprender el rendimiento histórico, las organizaciones pueden implementar mejoras continuas en sus estrategias y tácticas de ventas.

## 4.3 Concentración de Ingresos

La concentración de ingresos se refiere al grado en que los ingresos del Banco dependen de un pequeño número de clientes, productos o segmentos de mercado. Una alta concentración puede ser riesgosa, ya que la pérdida de un cliente clave podría tener un impacto significativo en los ingresos totales (Borja Tacuri, 2019).

- Diversificación de Clientes: Entender y gestionar la concentración de ingresos motiva a las empresas a diversificar su base de clientes, reduciendo la dependencia de fuentes únicas.
- Gestión de Riesgos: La concentración de ingresos puede aumentar la vulnerabilidad ante cambios en el mercado. La gestión proactiva de riesgos es esencial para mitigar posibles impactos adversos.
- Planificación Estratégica: Comprender la concentración de ingresos influye en la planificación estratégica. Las organizaciones pueden desarrollar estrategias para reducir riesgos y aprovechar oportunidades de crecimiento.

## 4.4 Importancia de la Gestión de Metas e Indicadores

La gestión de metas e indicadores proporciona una visión integral del rendimiento empresarial. Al establecer y monitorear metas específicas, la entidad financiera puede evaluar su éxito en la comercialización de seguros y ajustar estrategias según sea necesario.

- Toma de Decisiones Informada: El acceso a indicadores clave a través de un dashboard facilita la toma de decisiones informada. Los líderes de la entidad pueden identificar áreas de mejora, detectar tendencias y anticipar desviaciones en tiempo real, permitiendo una respuesta ágil a las condiciones del mercado.
- Indicadores de Rendimiento: El dashboard incluirá indicadores como la tasa de crecimiento de cartera, el cumplimiento de metas de ventas y otros relevantes para la comercialización de seguros.
- Desgloses Geográficos y Temporales: La capacidad de desglosar el rendimiento por región, sucursal o periodo temporal proporcionará insights valiosos para entender variaciones y patrones en la comercialización.
- Seguridad de Datos y Acceso: Se implementarán medidas de seguridad robustas para garantizar que solo los usuarios autorizados accedan a la información sensible.

## 4.5 Importancia de selección de variables

En el presente trabajo se seleccionarán variables según observaciones de su importancia en trabajos previos tales como : Marín (2006), Parrales Ramos (2013), Pérez López, Moya Fernández y Trigo Sierra (2012), Vergara-Romero (2011) y Bazan (2015).

El monto desembolsado es una variable crucial en el análisis de carteras de créditos, pues representa la cantidad de dinero prestada a los clientes. Su inclusión en el análisis de k-medias es fundamental, ya que los préstamos de diferentes montos pueden tener comportamientos de pago distintos. Además, el monto desembolsado puede estar relacionado con la capacidad de pago de los prestatarios. Por lo tanto, al agrupar clientes en segmentos según este atributo, podemos identificar patrones de riesgo y tomar decisiones informadas sobre la gestión de la cartera (Vergara-Romero, 2011).

La edad influye en comportamientos financieros como la capacidad de pago pueden. Por ejemplo, los clientes jóvenes pueden tener una mayor aversión al riesgo y ser más propensos a cumplir con sus obligaciones. Por otro lado, los clientes mayores pueden tener una mayor estabilidad financiera pero también pueden enfrentar limitaciones debido a su jubilación.

Al incluir la edad como variable en el análisis de k-medias, podemos explorar si existen grupos etarios con patrones de pago similares. Esto nos permitirá personalizar estrategias de gestión de riesgos y adaptar nuestras políticas de crédito según las necesidades de cada grupo.

El plazo de pago del crédito en días es el período durante el cual se espera que los prestatarios reembolsen el préstamo. Es un indicador importante de la capacidad de pago y la disciplina financiera de los clientes. Los plazos más cortos pueden estar asociados con tasas de incumplimiento más bajas, ya que los clientes tienen menos tiempo para enfrentar dificultades financieras (Borja Tacuri, 2019).

Al considerar esta variable en el análisis de k-medias, podemos identificar grupos de clientes con diferentes perfiles de plazo de pago. Por ejemplo, podríamos encontrar un grupo de prestatarios que tienden a pagar antes de la fecha límite y otro grupo que necesita extensiones de plazo. Esta información nos ayudará a diseñar estrategias de cobranza más efectivas y a evaluar la salud general de la cartera de créditos.

## 5 Metodología

### 5.1 Descripción de la base datos

La base de datos actual proviene de las transacciones de un mes del Banco PK respecto a ingresos no financieros (seguros). En la base de datos se cuenta con 42 variables y 211k registros de evaluación respecto a datos del cliente, del seguro y operación crediticia. Se detalla a continuación las variables más relevantes para la construcción de los indicadores:

- PRODUCTO: Indica el tipo de seguro adquirido del clientes, el mismo puede ser categorizado como Seguros Personales o seguros generales (patrimoniales).

- DESC\_SUCURSAL: Indica el departamento de la sucursal donde se hizo la transacción
- VALOR\_ASEGURADO: Muestra el monto monetario del valor del seguro.
- MONEDA: El tipo de moneda en el que se hizo la transacción esta puede ser bolivianos, dólares, u otro tipo de moneda.
- TASA\_PRIMA\_INTERES: La tasa prima de interés, es el porcentaje respecto del valor asegurado que el cliente debe pagar como parte del pago del seguro.
- FECHA\_NACIMIENTO: La fecha en la que nació el cliente, la variable nos ayudará a obtener la edad en años de los clientes
- MONTO\_DESEMBOLSADO: Es el valor monetario donde el cliente pidió un préstamo y está pendiente de devolución a la entidad financiera.
- OPERACION\_CREDITICIA: Es la razón o tipo operación crediticia donde el banco dio un préstamo a un cliente.
- ACT\_EXO\_OCUPACION: El tipo de actividad económica del cliente.

## 5.2 Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica ampliamente utilizada en el análisis de datos multivariantes. Su objetivo es explorar conjuntos de datos multidimensionales con variables cuantitativas. El ACP proyecta las observaciones desde un espacio  $p$ -dimensional (con  $p$  variables) a un espacio  $k$ -dimensional (donde  $k < p$ ), preservando la máxima información posible (medida por la varianza total del conjunto de datos) de las dimensiones originales.

## 5.3 Análisis de clusters K-medias

El Análisis de Clústeres de K-medias es una técnica que busca identificar grupos de casos similares en función de las características seleccionadas. Utiliza un algoritmo que puede manejar grandes conjuntos de datos. Sin embargo, requiere que el usuario especifique el número de clústeres. Además, se pueden definir centros iniciales para los clústeres si se dispone de esa información previa. Existen dos métodos disponibles para clasificar los casos: actualización iterativa de los centros de los clústeres o simplemente la clasificación. Los resultados incluyen la pertenencia a los clústeres, información sobre la distancia y los centros finales. También es posible etiquetar los resultados por casos utilizando una variable específica. Además, se pueden solicitar estadísticos F para los análisis de varianza. Aunque estos estadísticos son oportunistas (ya que el procedimiento intenta formar grupos que realmente difieran), su tamaño relativo proporciona información sobre la contribución de cada variable a la separación de los grupos.

## 5.4 Tratamiento sobre la base datos

Para el presente conjunto de datos se cuidó que las diferentes variables correspondan a un tipo de dato diferente según su naturaleza, como ser variables categóricas, numéricas y otras. Por otro lado, respecto a los datos faltantes, se observó que son una cantidad muy pequeña respecto del total, así se decidió omitirlas. Posteriormente, el tratamiento de las diferentes variables se vio directamente para la construcción de cada indicador, y previamente creando algunas variables como la región en base al departamento de origen de la transacción o la edad de los clientes en base a su fecha de nacimiento.

# 6 Resultados y Análisis

## 6.1 Preparación

La base de datos de clientes de créditos y seguros cuenta con 211.494 observaciones y 42 variables por lo que se realizó la preparación con la transformación de variables inicialmente para la creación de nuevas variables

como ser los días de cobertura del seguro (aplicando diferencia entre las fechas de inicio y fin de cobertura), asimismo, la edad se calculó a partir de la fecha de nacimiento y se realizó la conversión a variable numérica para proceder a aplicar la categorización por rangos de edades con la creación de una nueva

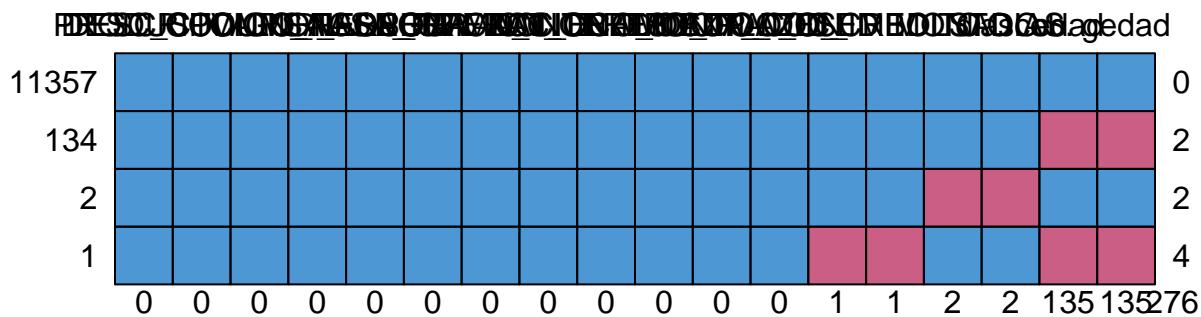
##Selección de variables

Posterior se realizó la selección de 18 variables que son relevantes para la información de los clientes, siendo que se tiene información de transacciones o garantías que no son relevantes para el objeto del estudio. Sin embargo, se crea una nueva base de datos solo con 18 variables de las 45 que se tenía, con la transformación al tipo de variable que corresponde cada atributo de la base de datos, ya sea char, num o factor.

## 6.2 Datos incompletos

Se hizo la evaluación de la base de datos para identificar datos incompletos y realizar su tratamiento, por lo que se muestra la matriz de patrones de datos faltantes y la frecuencia de cada patrón.

variable.



```
##      PRODUCTO DESC_SUCURSAL DESCRIPCION_AGENCIA CODIGO_PERSONA
## 211357      1             1                  1          1
## 134         1             1                  1          1
## 2           1             1                  1          1
## 1           1             1                  1          1
## 0           0             0                  0          0
##      VALOR_ASEGURADO MONEDA_OPERACION TASA_PRIMA_INTERES OPERACION_CREDITICIA
## 211357      1                 1                  1          1
## 134         1                 1                  1          1
```

```

## 2          1          1          1          1
## 1          1          1          1          1
## 0          0          0          0          0
##   NACIONALIDAD ACT_ECO_OCUPACION TIPO_DOC_ID EXT_DOC_ID MONTO_DESEMBOLSADO
## 211357      1          1          1          1          1
## 134        1          1          1          1          1
## 2          1          1          1          1          1
## 1          1          1          1          1          0
## 0          0          0          0          0          1
##   PLAZO_CREDITO.DIAS. MONTO diascob edad gedad
## 211357      1    1    1    1    1    0
## 134        1    1    1    0    0    2
## 2          1    0    0    1    1    2
## 1          0    1    1    0    0    4
## 0          1    2    2  135  135 276

```

La primera columna de la tabla indica el número de filas que tienen un patrón específico de datos faltantes:

- 211357 filas no tienen datos faltantes en ninguna variable (1 indica que el dato está presente).
- 134 filas tienen un patrón idéntico de presencia y ausencia de datos.
- 2 filas tienen otro patrón de presencia y ausencia de datos.
- 1 fila tiene otro patrón distinto de presencia y ausencia de datos.

Las filas de la matriz indican si los datos están presentes (1) o faltan (0) para cada variable. La última fila indica cuántos valores faltantes hay para cada variable: - 135 valores faltan en la variable edad. - 135 valores faltan en la variable rango de edad. - 276 valores faltan en la variable días de cobertura.

	pobs	influx	outflux	ainb	aout	fico
PRODUCTO	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
DESC_SUCURSAL	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
DESCRIPCION_AGENCIA	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
CODIGO_PERSONA	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
VALOR_ASEGURADO	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
MONTO	0.9999905	0.0000084	0.9855072	0.9411765	7.57e-05	0.0006383
MONEDA_OPERACION	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
TASA_PRIMA_INTERES	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
MONTO_DESEMBOLSADO	0.9999953	0.0000037	0.9855072	0.8235294	7.57e-05	0.0006430
PLAZO_CREDITO.DIAS.	0.9999953	0.0000037	0.9855072	0.8235294	7.57e-05	0.0006430
OPERACION_CREDITICIA	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
NACIONALIDAD	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
ACT_ECO_OCUPACION	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
TIPO_DOC_ID	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
EXT_DOC_ID	1.0000000	0.0000000	1.0000000	0.0000000	7.68e-05	0.0006478
edad	0.9993617	0.0005669	0.0144928	0.9403050	1.10e-06	0.0000095
gedad	0.9993617	0.0005669	0.0144928	0.9403050	1.10e-06	0.0000095
diascob	0.9999905	0.0000084	0.9855072	0.9411765	7.57e-05	0.0006383

La proporción de valores presentes en cada variable, valores cercanos a 1 indican que casi todos los datos están presentes. + influx: Proporción de valores faltantes para cada variable en relación con el total de valores faltantes. + Variables como PRIMA, CREDITO, diascred, edad, gedad, y diascob tienen valores muy bajos de influx, indicando que los valores faltantes en estas variables constituyen una pequeña fracción del total. + outflux: Proporción de datos presentes en una variable que coinciden con datos faltantes en

otras variables. La mayoría de las variables tienen un outflux de 1, lo que indica que los datos presentes en estas variables coinciden con los datos presentes en otras variables. La mayoría de las variables tienen muy pocos o ningún valor faltante. Algunas variables, como edad y gedad, tienen un número ligeramente mayor de valores faltantes. El influx y outflux indican que los valores faltantes están dispersos entre las observaciones y no se concentran en una sola variable. El análisis sugiere que la base de datos es bastante completa, con algunas variables que tienen un pequeño número de valores faltantes. Los patrones de datos faltantes parecen ser dispersos, sin concentrarse en una sola variable, por lo que se aplicará el método listwise con variables que sean con datos completos. Por lo que se genera una nueva base de datos con 211.357 de los 211.494 clientes (se observaron 137 casos 0.06%).

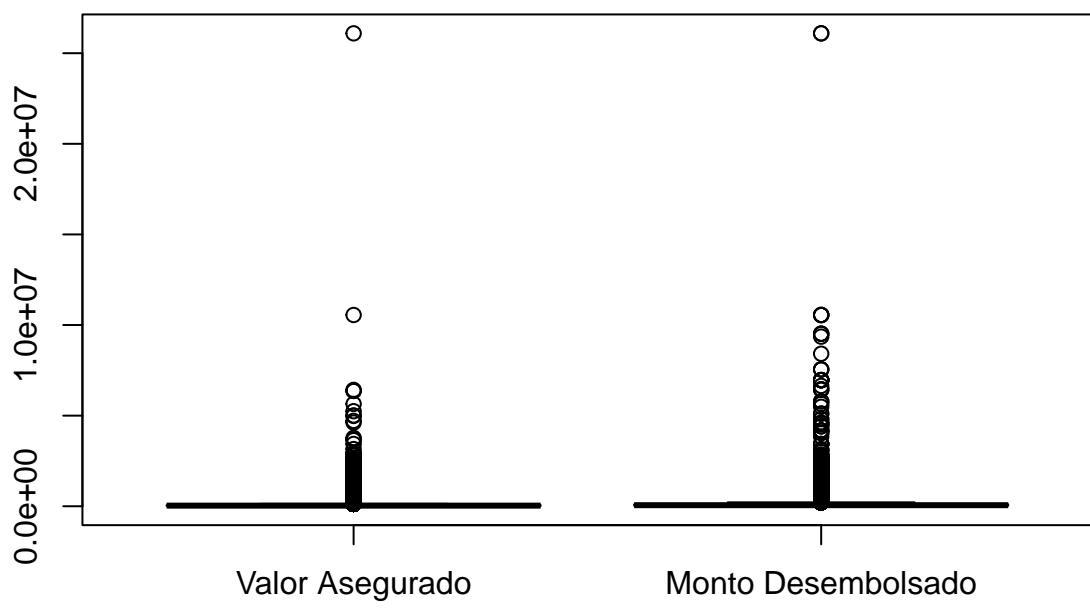
### 6.3 Datos atípicos

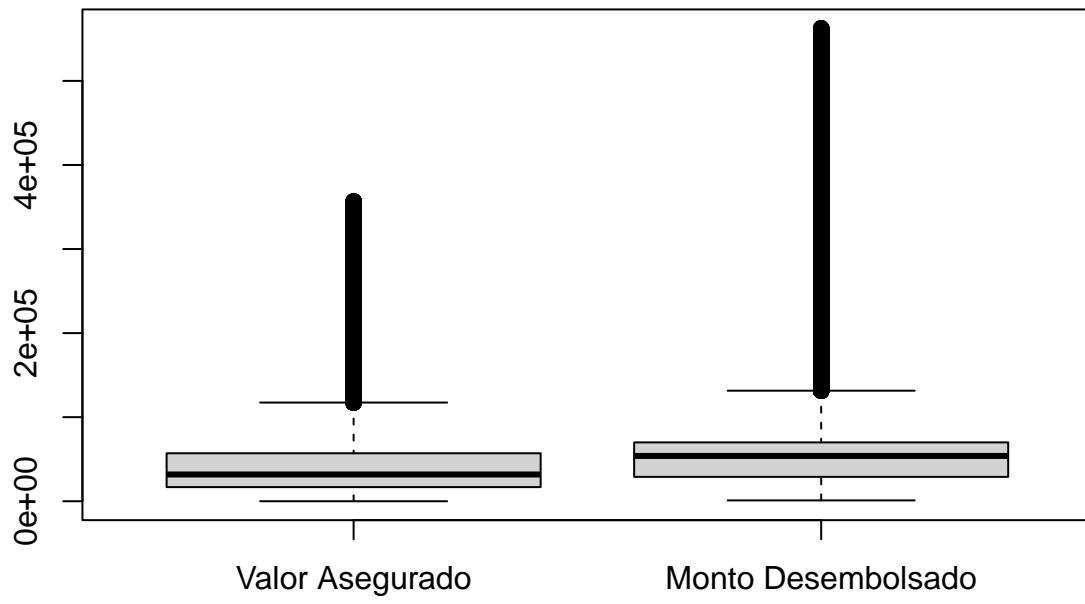
Estas son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Los datos atípicos son ocasionados por:

- Errores de procedimiento.
- Acontecimientos extraordinarios.
- Valores extremos. Por ejemplo, una muestra de datos del número de cigarrillos consumidos a diario contiene el valor 60 porque hay un fumador que fuma sesenta cigarrillos al día.
- Causas no conocidas.

Por lo que se recomienda que se los identifique y tratarlos de manera adecuada, generalmente excluyéndolos del análisis porque distorsionan los resultados de los análisis.

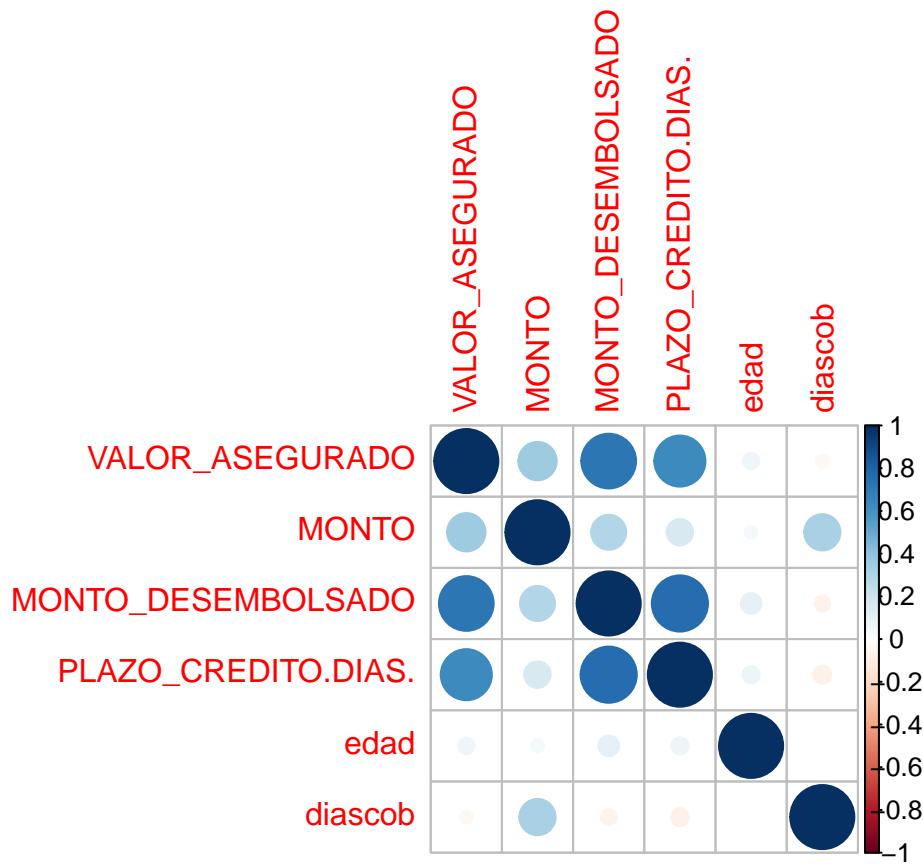
Es por ello que se recortó la base en un 5% para poder trabajar de mejor manera ya que los datos atípicos distorsionaban el análisis. Para ello se seleccionaron las variables ‘Monto desembolsado’ y ‘valor asegurado’ para su recorte.





#### 6.4 Análisis de componentes principales

Para identificar los principales componentes de la base de datos de clientes con créditos se realizó el análisis, con el gráfico de matriz de correlaciones visualizada mediante un “correlograma”. Se muestra la relación entre las variables numéricas de la base de datos.



Las variables numéricas consideradas en la matriz de correlación son: SALDO, PRIMA, CREDITO, diascred, edad, y diascob. El color de las celdas varía desde el azul oscuro (correlación positiva alta) hasta el rojo oscuro (correlación negativa alta). Azul oscuro indica una correlación positiva cercana a 1. Rojo oscuro indica una correlación negativa cercana a -1. Colores más claros indican correlaciones más cercanas a 0. Tamaño: El tamaño de los círculos también indica la magnitud de la correlación. Círculos más grandes indican una correlación más fuerte (positiva o negativa). Círculos más pequeños indican una correlación más débil. SALDO: Tiene una correlación positiva fuerte con PRIMA y CREDITO. PRIMA: Además de la fuerte correlación con SALDO, también muestra una correlación positiva notable con CREDITO. CREDITO: Correlaciona fuertemente con diascred. diascred: Tiene una fuerte correlación positiva con CREDITO. edad: Muestra correlaciones más débiles con las otras variables. diascob: Correlaciona fuertemente con diascred. Correlaciones Positivas Fuertes: - SALDO y PRIMA - SALDO y CREDITO - PRIMA y CREDITO - CREDITO y diascred - diascred y diascob

#### 6.4.1 Correlaciones Débiles o Inexistentes:

- Edad con otras variables muestra correlaciones débiles o casi inexistentes. La matriz de correlaciones nos indica que hay fuertes relaciones entre algunas de las variables financieras como SALDO, PRIMA, y CREDITO, así como entre CREDITO, diascred y diascob. Las correlaciones débiles de edad sugieren que la edad no tiene una relación fuerte con las demás variables en el conjunto de datos. Se procedió al análisis de componentes principales PCA sobre la matriz de correlaciones. La interpretación de los resultados implica descomponer la matriz de correlaciones y reconstruirla usando un subconjunto de los componentes principales.

El gráfico muestra los valores propios (eigenvalues) de la matriz de correlaciones en orden descendente. Este tipo de gráfico colabora a determinar el número de componentes principales a retener, por lo que se podría considerar aquellos valores superiores o iguales a 1, considerando los primeros tres componentes.

Los valores propios (m1\$values) y la proporción de varianza explicada por cada componente principal se presentan como sigue:

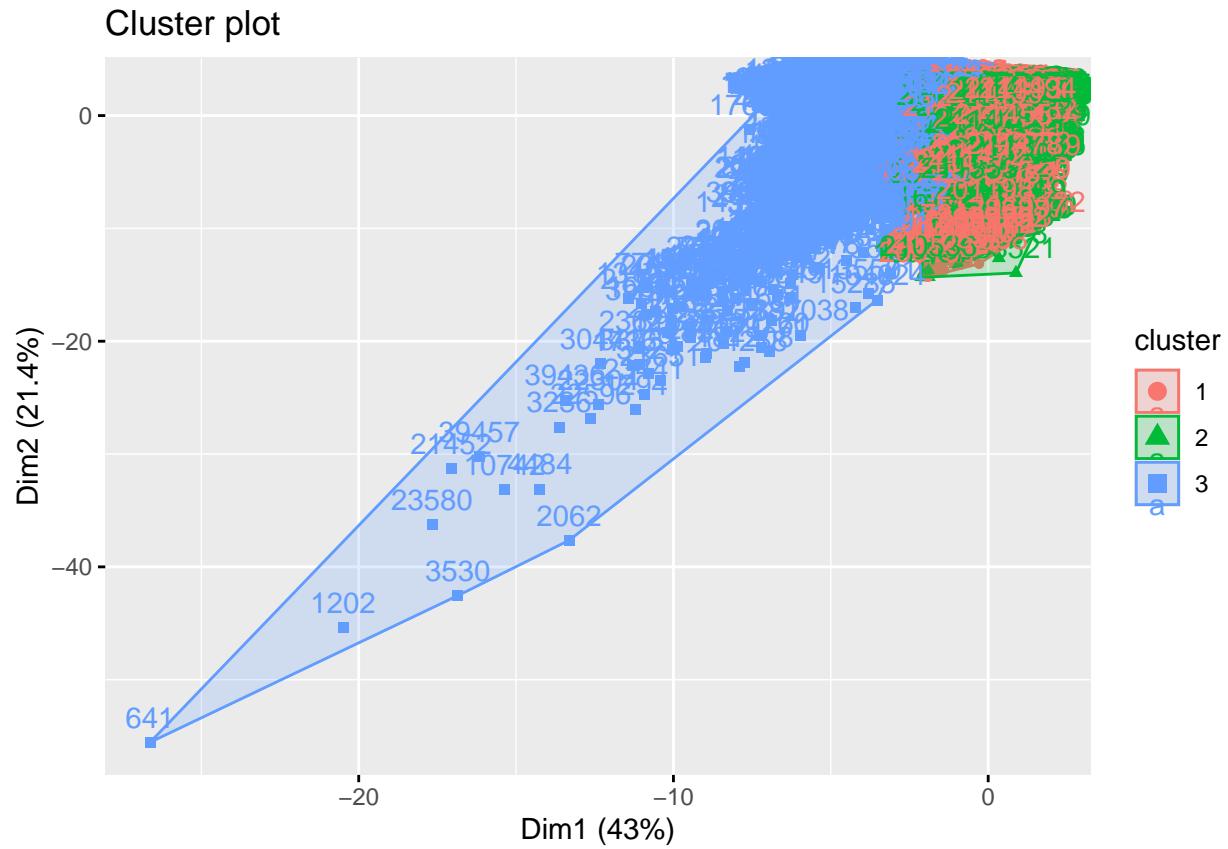
Primer componente principal: Explica el 42.37% de la varianza total. Segundo componente principal: Explica el 19.10% de la varianza total. Tercer componente principal: Explica el 16.47% de la varianza total. Cuarto componente principal: Explica el 12.06% de la varianza total. Quinto componente principal: Explica el 6.21% de la varianza total. Sexto componente principal: Explica el 3.79% de la varianza total. La varianza acumulada se calcula para determinar cuántos componentes principales son necesarios para explicar un cierto porcentaje de la varianza total: Primer componente principal: Explica el 42.37% de la varianza acumulada. Primeros dos componentes principales: Explican el 61.47% de la varianza acumulada. Primeros tres componentes principales: Explican el 77.94% de la varianza acumulada. Primeros cuatro componentes principales: Explican el 90.00% de la varianza acumulada. Primeros cinco componentes principales: Explican el 96.21% de la varianza acumulada. Todos los componentes principales: Explican el 100% de la varianza acumulada. La decisión sobre cuántos componentes principales retener puede basarse en el porcentaje de varianza acumulada. Un criterio común es retener suficientes componentes para explicar al menos el 70-90% de la varianza total. En este caso, los primeros tres componentes principales explican el 78.00% de la varianza acumulada, lo cual suele considerarse suficiente para la mayoría de los análisis. En el gráfico muestra un claro “codo” (punto de inflexión) después del primer componente principal, con una disminución más gradual en los valores propios después del tercer componente principal. Esto sugiere que los primeros tres componentes capturan la mayor parte de la información relevante en los datos.

## 6.5 CLUSTER

En el análisis de cluster por k-means se consideraron dos escenarios, donde en el primero se trabajan con todas las variables numéricas ‘valor asegurado’, ‘monto’, ‘monto desembolsado’, ‘plazo credito días’, ‘edad’, ‘dias cobrados’. En el segundo escenario se trabajan con variables seleccionadas en base a la teoría tales como ‘monto desembolsado’, ‘plazo credito días’ y ‘edad’.

Además según los resultados obtenidos en el análisis de componentes es que se realizará el análisis de k-means para 3 y 2 clusters, respectivamente.

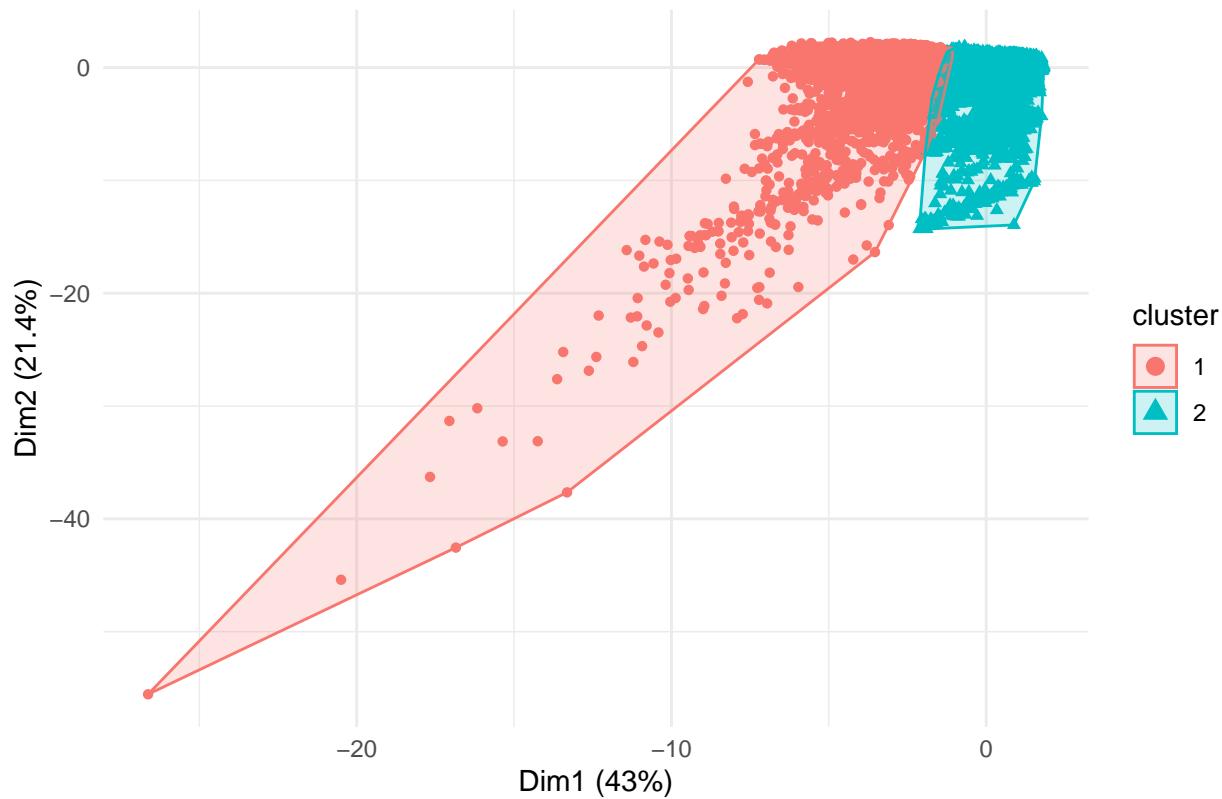
*Escenario 1- Todas las variables numéricas en 3 Clusters*



1. El gráfico muestra las dos primeras componentes principales (Dim1 y Dim2) que explican el 38.4% y 21.3% de la variabilidad total, respectivamente.
  2. Cluster 1 (Rojo): Los puntos rojos representan las observaciones asignadas al Cluster 1. Este cluster tiene una extensión bastante amplia a lo largo de Dim1 y Dim2.
  3. Cluster 2 (Verde): Los puntos verdes representan las observaciones asignadas al Cluster 2. Este cluster está más concentrado en el lado derecho del gráfico.
  4. Cluster 3 (Azul): Los puntos azules representan las observaciones asignadas al Cluster 3. Este cluster está más disperso en el área inferior derecha del gráfico.
  5. Se observa una cierta superposición entre los clusters 1 y 2, mientras que el Cluster 3 está más separado y bien definido.
  6. El Cluster 1 tiene una mayor dispersión, lo que indica una mayor variabilidad dentro de este grupo.

### *Escenario 1- Todas las variables numéricas en 2 Clusters*

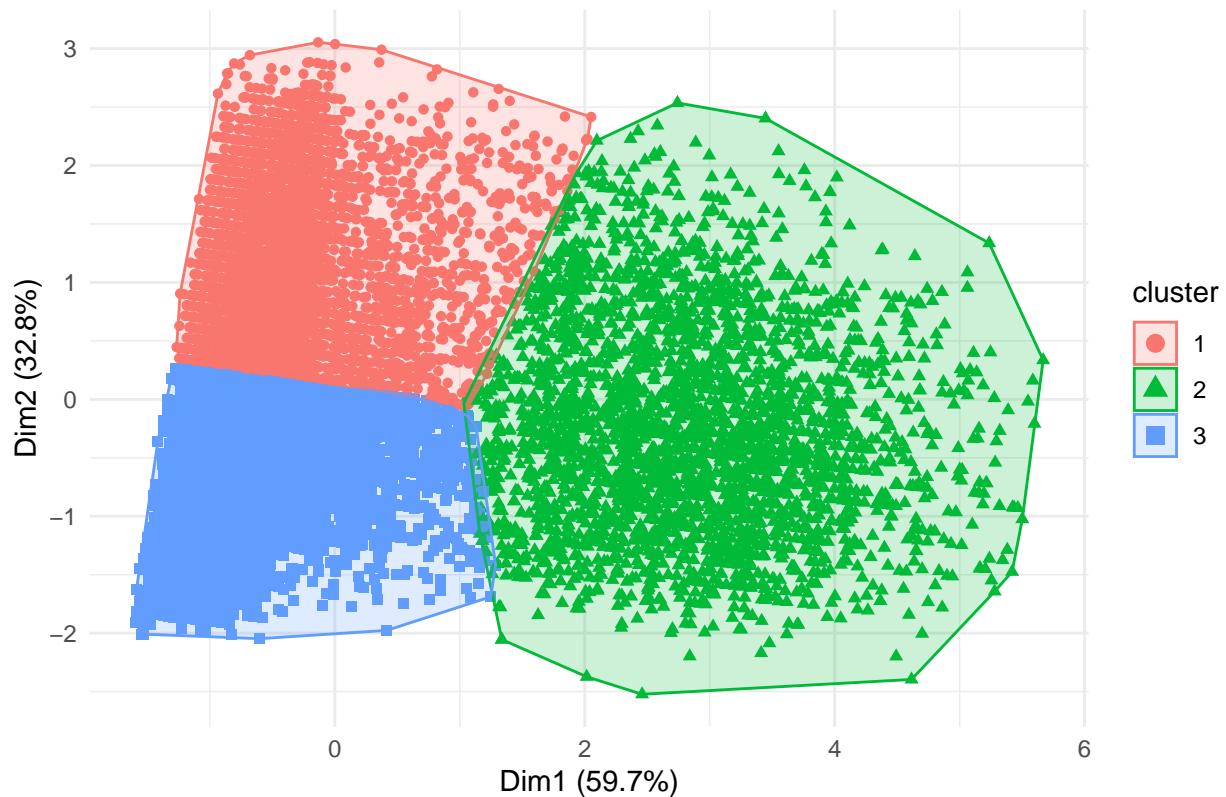
Cluster plot



1. Similar al gráfico anterior, muestra las dos primeras componentes principales con las mismas proporciones de variabilidad.
2. Cluster 1 (Rojo): Los puntos rojos representan las observaciones asignadas al Cluster 1. Este cluster abarca la mayor parte del gráfico, similar al Cluster 1 del análisis con 3 clusters.
3. Cluster 2 (Cian): Los puntos cian representan las observaciones asignadas al Cluster 2. Este cluster está más concentrado y ubicado principalmente en la parte derecha del gráfico.
4. Separación y Cohesión: La separación entre los dos clusters es más clara en comparación con el análisis de 3 clusters. El Cluster 1 es más amplio, mientras que el Cluster 2 es más compacto.
5. Distribución: La reducción de tres a dos clusters ha llevado a una fusión de los grupos, especialmente de los Clusters 2 y 3 del primer análisis en un solo Cluster 2 en este análisis.

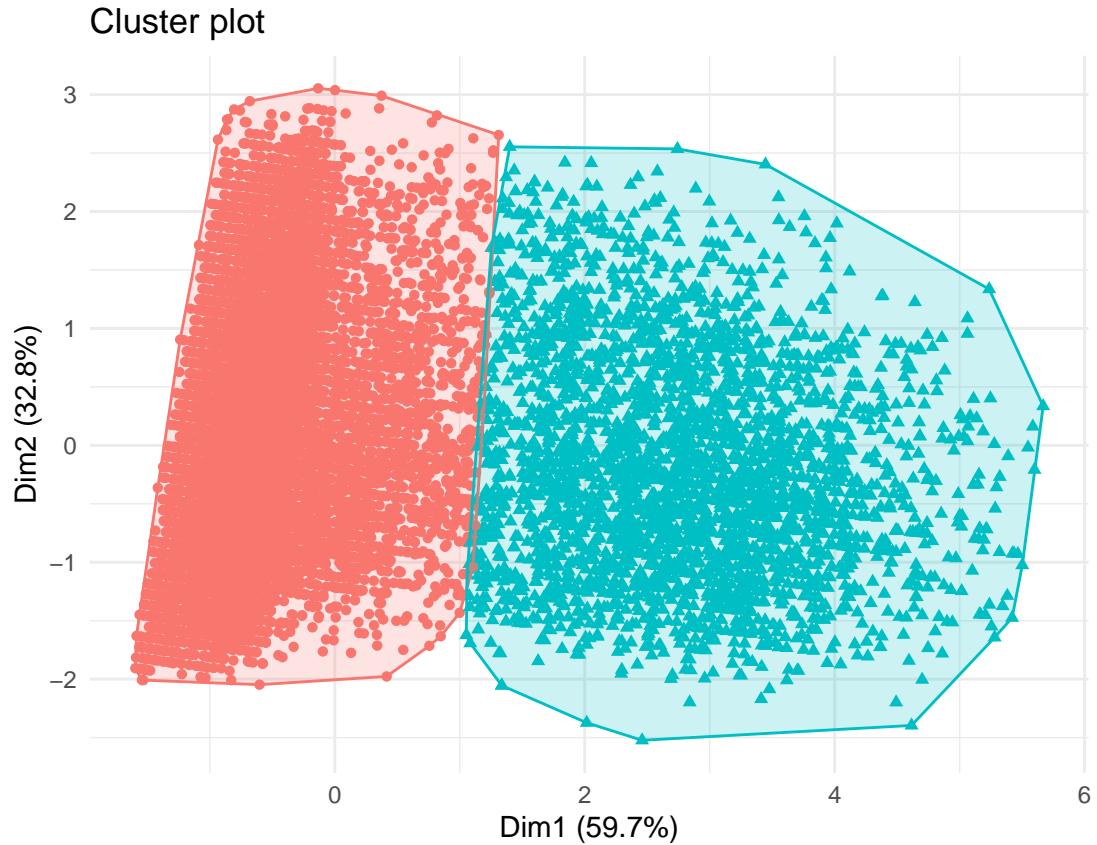
*Escenario 2- Variables numéricas seleccionadas en 3 Clusters*

Cluster plot



1. El gráfico muestra las dos primeras componentes principales (Dim1 y Dim2) que explican el 59.7% y 32.8.3% de la variabilidad total, respectivamente.
2. Cluster 1 (Rojo): Los puntos rojos representan las observaciones asignadas al Cluster 1. Este cluster tiene una extensión más cercana a la Dim1.
3. Cluster 2 (Verde): Los puntos verdes representan las observaciones asignadas al Cluster 2. Este cluster está más concentrado en el lado derecho del gráfico Dim1.
4. Cluster 3 (Azul): Los puntos azules representan las observaciones asignadas al Cluster 3. Este cluster está más disperso en el área inferior derecha del gráfico.
5. Se observa que los clusters si bien están separados, estos no se encuentran definidos.
6. El Cluster 2 tiene una mayor dispersión, lo que indica una mayor variabilidad dentro de este grupo.

*Escenario 2- Variables numéricas seleccionadas en 2 Clusters*



1. Similar al gráfico anterior, muestra las dos primeras componentes principales con las mismas proporciones de variabilidad. 2. Cluster 1 (Rojo): Este cluster está más concentrado y ubicado principalmente en la parte derecha de la Dim1. 3. Cluster 2 (Cian): Este cluster abarca la mayor parte del gráfico, donde se encuentra mayor dispersión de los datos y ubicación al lado izquierdo de la Dim1. 4. Separación y Cohesión: La separación entre los dos clusters es más clara en comparación con el análisis de 3 clusters. El Cluster 2 es más amplio, mientras que el Cluster 1 es más compacto. 5. Distribución: La reducción de tres a dos clusters ha llevado a una fusión de los grupos, especialmente de los Clusters 2 y 3 del primer análisis en un solo Cluster 2 en este análisis.

Comparación 1. El análisis con 3 clusters muestra una mayor segmentación en ambos escenarios, lo que puede ser útil si se busca una diferenciación más detallada dentro de los datos. 2. En el análisis con 3 clusters en ambos escenarios (pero con mayor énfasis en el primero), muestra que hay cierta superposición entre los clusters, especialmente entre los Clusters 1 y 2. 3. En el análisis con 2 clusters, la separación entre los clusters es más clara, lo que puede facilitar la interpretación y la toma de decisiones basadas en estos grupos. 4. Para una segmentación de clientes más detallada, el análisis con 3 clusters puede ser más apropiado. 5. Se debe considerar que en ambos escenarios con 2 o 3 clusters se puede encontrar segmentación de grupos, pero visualmente estos grupos se encuentran más dispersos.

## 7 Conclusiones

A continuación las principales conclusiones de la investigación:

- En el análisis de datos faltantes, la mayoría de las variables en el conjunto de datos tienen pocos o ningún dato faltante.
- Las variables con datos faltantes, como PRIMA, CREDITO, diascred, edad, gedad, y diascob, muestran que, aunque la cantidad de datos faltantes es pequeña.
- Se identificaron correlaciones significativas entre algunas variables, especialmente entre PRIMA y SALDO, y entre CREDITO y diascred.
- La variable edad no muestra una fuerte correlación con las demás variables.
- Los primeros componentes principales capturan la mayor parte de la variabilidad en los

datos, con los tres primeros componentes explicando aproximadamente el 78% de la variabilidad. - El análisis PCA ayudó a reducir la dimensionalidad del conjunto de datos. - La segmentación con k-means identificó 2 y 3 clusters distintos en el conjunto de datos. - El análisis con 3 clusters proporcionó una segmentación más detallada, mientras que el análisis con 2 clusters ofreció una clasificación más simplificada. - Se rechaza la hipótesis, pues esperábamos identificar patrones claros de comportamiento en la cartera de clientes, sin embargo, los resultados no mostraron agrupaciones claras en los datos. Por lo tanto, no podemos afirmar que el análisis de k-medias sea una herramienta efectiva para comprender el comportamiento de los clientes en este contexto específico.

## 8 Recomendaciones

- Se debe explorar más a fondo las relaciones entre variables que muestran correlaciones significativas para entender mejor las dinámicas subyacentes, simplificar el modelo y evitar multicolinealidad.
- Para modelo predictivos considerar el uso de las componentes principales en lugar de las variables originales.
- Continuar explorando componentes adicionales si se busca capturar más de la variabilidad en el conjunto de datos.
- Para ahondar en la segmentación, realizar una validación cruzada y ajustar los hiperparámetros del modelo de clustering para asegurar que la segmentación sea robusta y válida.
- Recomendamos explorar otras técnicas de análisis y considerar factores adicionales para obtener una visión más completa. Podemos sugerir un análisis de regresión lineal por ejemplo.

Las recomendaciones basadas en estos hallazgos ayudarán a mejorar la toma de decisiones.

## 9 Referencias

- Bazan, F. (2015). Aplicación de la técnica estadística clúster k-medias para la segmentación orientada a comprender las necesidades de financiamiento de clientes de una entidad financiera. (Tesis de doctorado). UNIVERSIDAD NACIONAL DE INGENIERÍA.
- Borja Tacuri, M. A. (2019). Factores de morosidad en la cartera de créditos en Caja Arequipa agencia El Tambo. Universidad Continental.
- Celi Yanangomez, G. M. (2023). Diseño de una estrategia de recuperación crediticia temprana para clientes Pequeñas Empresas de una institución financiera del Ecuador mediante los algoritmos k-medias y bosques aleatorios. (Tesis de licenciatura). Quito: EPN.
- García Lomas, V. A. (2018). Análisis de la cartera de créditos de la banca pública ecuatoriana (2008-2017). Revista científica UISRAEL, 5(3), 37-50. Universidad Tecnológica Israel.
- Marín, Z. I. C. (2006). La diversificación del riesgo en la cartera de créditos del sector financiero con base en la teoría de portafolios. Universidad EAFIT.
- Parrales Ramos, C. (2013). Análisis del índice de morosidad en la cartera de créditos del IECE-Guayaquil y propuesta de mecanismos de prevención de morosidad y técnicas eficientes de cobranzas. (Tesis de maestría).
- Pérez López, Á., Moya Fernández, A. J., & Trigo Sierra, E. (2012). Cuestiones prácticas de las ventas de carteras de créditos. Actualidad Jurídica, 33(3), 45-62. Dykinson.
- Vergara-Romero, A. (2011). Análisis de las carteras de créditos orientados a la Microempresa de los Bancos Privados del Ecuador 2009-2010 (Analysis of the Credit Portfolios Oriented to Microenterprises of the Private Banks of Ecuador 2009-2010). Repositorio Universidad de Guayaquil 2021.