

Mineria de Datos

Análisis Exploratorio de Datos y Predicción de Incumplimiento de Préstamos

Aina Llaneras Casas, Alejandro Arcas Alberti, Alessandro Natali Vilamú,
Berta Moyano Núñez, Blanca Romero Sainz, Iker Meneses Sales, Ismael
Argemí Fernández, Iván Martínez Yates, Marta Gomez de la Tia Privat, Mireia
Bohils Tenas, Mireia Bolívar Rubia, Oscar Arroyo Luque, Arnaut Goethals

GRUPO 1: Aina Llaneras, Blanca Romero, Iván Martínez

GRUPO 2: Alejandro Arcas Alberti, Alessandro Natali, Iker Meneses, Arnaut Goethals

GRUPO 3: Ismael Argemí, Mireia Bohils, Oscar Arroyo

GRUPO 4: Berta Moyano, Marta Gómez, Mireia Bolívar

14 de Noviembre del 2023

ÍNDICE

1. Definición del proyecto y asignación	1
1.1. Fuente de obtención de los datos	1
1.2. Descripción de los datos	1
1.3. Estructura e información de la matriz de datos	1
1.4. Plan de trabajo	6
1.5. Análisis de riesgos	7
2. Análisis descriptivo pre-preprocessing	7
2.1. Descriptiva univariante	7
2.2. Descriptiva bivariante	19
3. Preprocessing de los datos	29
3.1. Limpieza de datos y estandarización de formato	29
3.2. Detección y tratamiento de missings	32
3.3. Detección y tratamiento de outliers	36
3.4. Feature engineering	38
4. Análisis descriptivo post-preprocessing	41
4.1. Descriptiva univariante	41
4.2. Descriptiva bivariante	51
5. Análisis de Correspondencias Principales (ACP)	60
5.1. Interpretación de los planos factoriales	67
5.2. Representación de individuos	68
5.3. Representación de variables categóricas en primeros planos factoriales	69
6. Análisis de Correspondencias Múltiples (ACM)	78
6.1. Desarrollo del ACM	78
6.2. Gráfico de individuos y variables	81
6.3. Gráfico de individuos	82
6.4. Gráfico de variables	84
7. Clustering	90
7.1. K-Means	90
7.2. Clustering jerárquico	92
7.3. Profiling K-Means	96
7.4. Profiling clustering jerárquico	100
8. Clustering avanzado	110
8.1. K-Modes	110
8.2. DBSCAN	112
8.3. OPTICS	115
8.4. CURE	121
8.5. Fuzzy clustering	124
9. Reglas de asociación (Basket Market Analysis)	131

10. Modelos discriminantes	141
10.1. LDA y QDA	141
10.2. K-Nearest Neighbors	146
Anexo	149

Definición del proyecto y asignación

El objetivo principal de este trabajo es permitir a las instituciones financieras o analistas de riesgos realizar un análisis exploratorio de datos completo para evaluar la probabilidad de que un prestatario incumpla con sus obligaciones financieras. Para un mejor funcionamiento del equipo y una correcta distribución de tareas, se ha separado el conjunto de los integrantes en 4 subgrupos mencionados previamente, cada uno constando de 3 integrantes, para poder efectuar las tareas con mayor assertividad e independencia.

Fuente de obtención de los datos

Los datos se han extraído del repositorio de bases de datos Kaggle. El enlace de la página web es el siguiente: https://www.kaggle.com/datasets/gauravduttakiit/loan-defaulter?select=application_data.csv

Descripción de los datos

Esta base de datos está diseñada para abordar el desafío de identificar posibles incumplimientos de préstamos en un entorno empresarial real. El conjunto de datos contiene información relacionada con préstamos otorgados a diversos prestatarios, junto con detalles financieros y personales de los solicitantes.

Estructura e información de la matriz de datos

Filas (individuos)	Columnas (variables)	Nro. variables numéricas	Nro. variables categóricas	Nro. variables respuesta u objetivo
5000	15	7	8	1

VARIABLES EXPLICATIVAS

Nombre	Descripción	Tipo	Diccionario y dominio
CODE_GENDER	Género del cliente	Categórica	M—Male, F—Female
NAME_INCOME_TYPE	Tipo de ingresos	Categórica	1-Businessman, 2-Commercial associate, 3-Pensioner, 4-State servant, 5-Working
NAME_EDUCATION_TYPE	Nivel de estudios del cliente	Categórica	1-Academic degree, 2-Higher education, 3-Incomplete higher, 4-Lower secondary, 5-Secondary special
NAME_FAMILY_STATUS	Estado civil	Categórica	1-Married, 2-Single/not married, 3-Civil marriage, 4-Separated, 5-Widow

Nombre	Descripción	Tipo	Diccionario y dominio
OCCUPATION _TYPE	Actividad laboral	Categórica	1-Laborers, 2-Sales staff, 3-Core staff, 4-Managers, 5-Drivers, 6-Accountants, 7-Cleaning staff, 8- High skill tech staff, 9-HR staff, 10-IT staff, 11-Cooking staff, 12-Low-skill Laborers, 13-Medicine staff, 14-Private service staff, 15-Realty agents, 16-Security staff, 17-Secretaries, 18-Waiters/barmen staff

Nombre	Descripción	Tipo	Diccionario y dominio
ORGANIZATION_TYPE	Tipo de organización donde trabaja el cliente	Categórica	1-Advertising, 2-Agriculture, 3-Bank, 4-Business Entity Type 1, 5-Business Entity Type 2, 6-Business Entity Type 3, 7-Cleaning, 8-Construction, 9-Culture, 10-Electricity, 11-Emergency, 12-Government, 13-Hotel, 14-Housing, 15-Industry: type 1, 16-Industry: type 10, 17-Industry: type 11, 18-Industry: type 12, 19-Industry: type 13, 20-Industry: type 2, 21-Industry: type 3, 22-Industry: type 4, 23-Industry: type 5, 24-Industry: type 6, 25-Industry: type 7, 26-Industry: type 9, 27-Insurance, 28-Kindergarten, 29-Legal Services, 30-Medicine, 31-Military, 32-Mobile, 33-Other, 34-Police, 35-Postal, 36-Realtor, 37-Restaurant, 38-School, 39-Security, 40-Security Ministries, 41-Self-employed, 42-Services, 43-Telecom, 44-Trade: type 1, 45-Trade: type 2, 46-Trade: type 3, 47-Trade: type 4, 48-Trade: type 6, 49-Trade: type 7, 50-Transport: type 1, 51-Transport: type 2, 52-Transport: type 3, 53-Transport: type 4, 54-University, 55-XNA
REGION_RATING_CLIENT	Nuestra calificación de la región donde vive el cliente	Categórica	1, 2, 3
AMT_INCOME_TOTAL	Ingresos totales del cliente	Numérica	[29250, 2250000]
AMT_CREDIT	Importe de crédito del préstamo	Numérica	[45000, 3375000]
AMT_ANNUITY	Anualidad del préstamo	Numérica	[2673, 177827]

Nombre	Descripción	Tipo	Diccionario y dominio
DAY_S_BIRTH	Edad del cliente en número de días en el momento de pedir el préstamo	Numérica	[-25159, -7711]
OWN_CAR_AGE	Edad en años del coche del cliente	Numérica	[0, 65]
AMT_GOODS_PRICE	Para préstamos al consumo, es el precio de los bienes para los cuales se otorga el préstamo	Numérica	[45000, 3375000]
CNT_FAM_MEMBERS	Número de familiares del cliente	Numérica	[1, 8]

VARIABLE OUTPUT

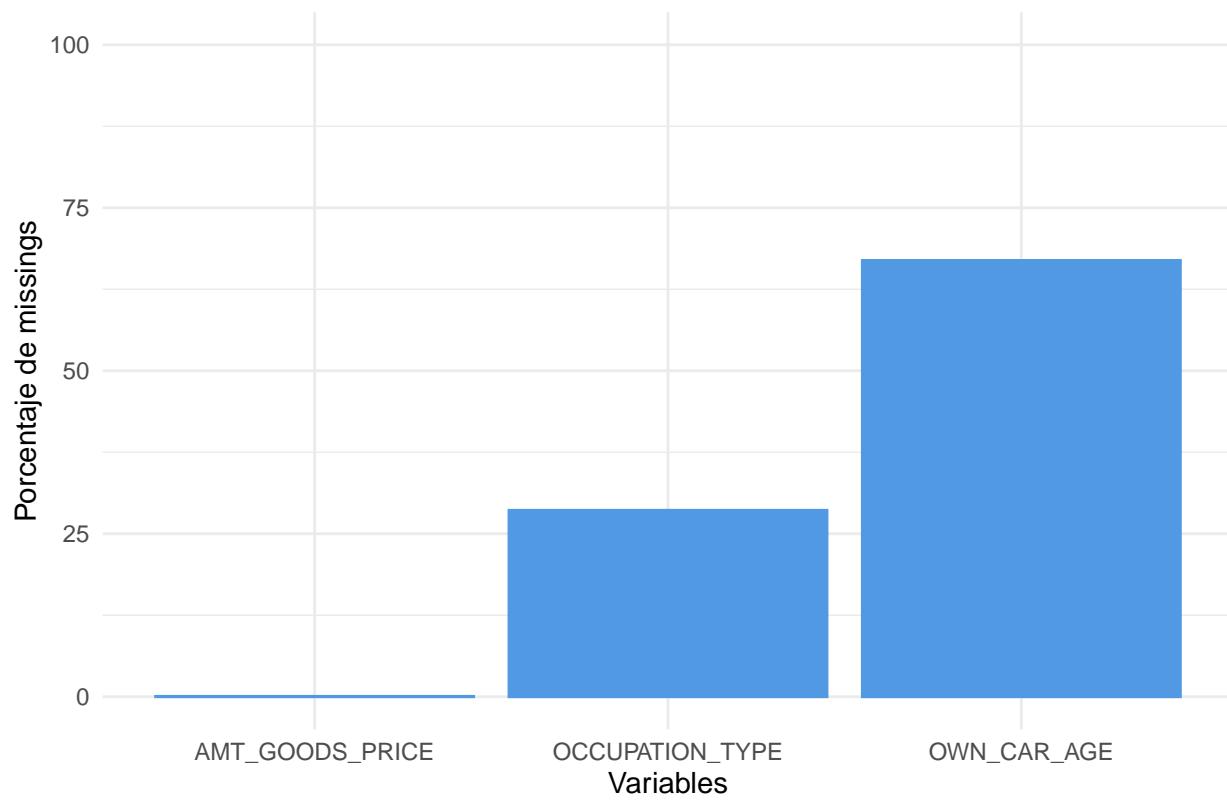
Nombre	Descripción	Tipo	Diccionario y dominio
Target	Target	Categórica	1 - Cliente con dificultades de pago: él/ella tuvo pagos atrasados de más de X días en al menos una de las primeras Y cuotas del préstamo en nuestra muestra. 0 - Todos los demás casos

VARIABLES MISSINGS

Nro. de casillas missings	Respeto del total de la matriz datos
4779	6.37 %

Porcentaje de missings por variable (tabla y histograma):

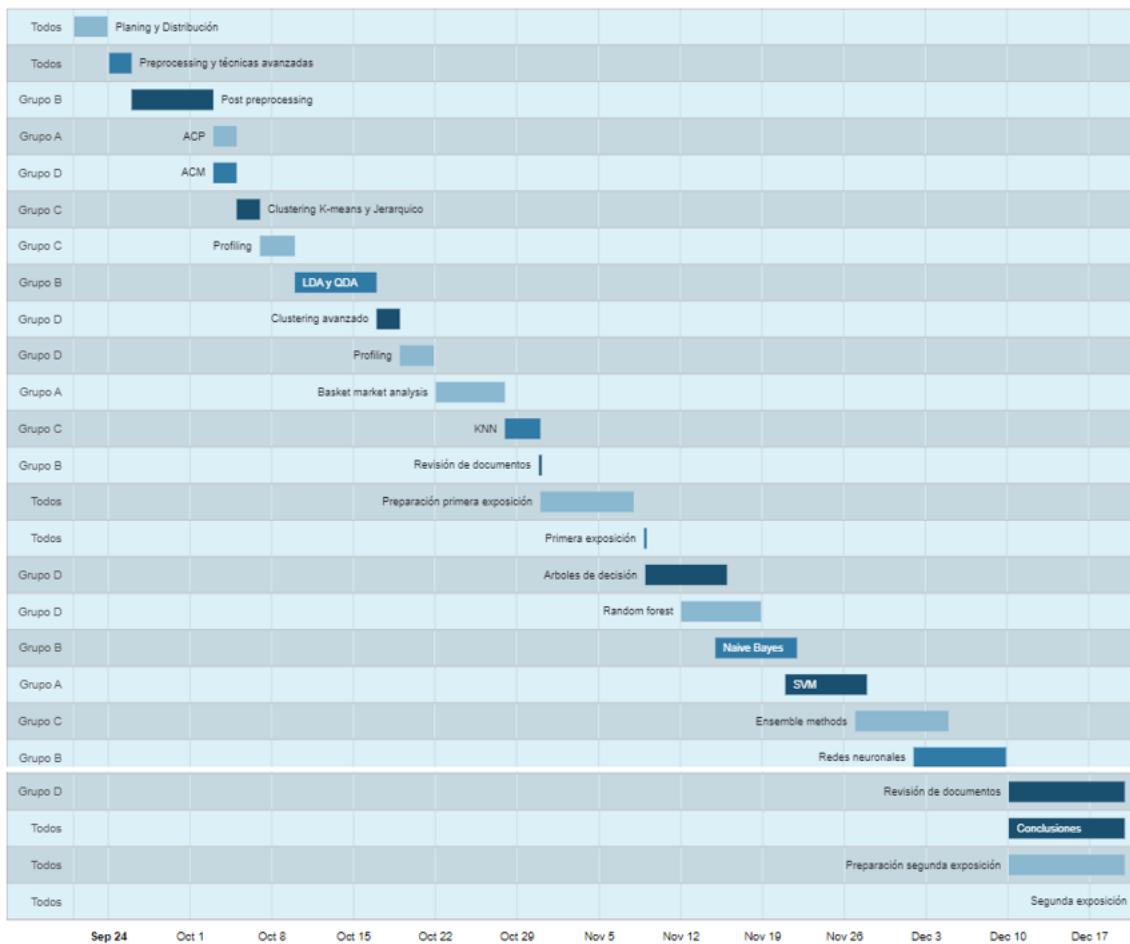
	Nro. de missings	Porcentaje de missings
OWN_CAR_AGE	3346	66.92 %
AMT_GOODS_PRICE	3	0.06 %
OCCUPATION_TYPE	1430	28.6 %

Gráfico de porcentaje de missings

Únicamente se han representado las variables que tienen algún valor faltante.

Plan de trabajo

Diagrama de Gantt



Análisis de riesgos

Se han identificado los siguientes riesgos que podrían afectar al correcto desarrollo del trabajo:

Possible problema	Probabilidad de suceso	Solución
Tarea crítica no finalizada a tiempo	Baja	Establecer una fecha límite previa para tener margen de maniobra
Falta y/o errores de comunicación entre los miembros del grupo	Alta	Canales de comunicación claros y efectivos y designar un líder por equipo
Error en una tarea inicial que impida la correcta evolución	Media	Tareas iniciales revisadas por miembros de otros grupos Asignar a dos grupos para que trabajen de forma simultánea
Ausencia temporal de algun membro del equipo	Alta	Un subgrupo dará soporte para la finalización de la tarea a tiempo Correcta explicación del avance realizado al integrante que ha faltado temporalmente
Ausencia permanente de algun miembro del equipo	Baja	Reasignación de los integrantes del subgrupo en otro y redistribución de las tareas.
Falta de conocimiento de tareas anteriores	Alta	Revisar todos los avances que se han realizado en cada uno de los grupos Asegurar que todos los miembros de cada grupo entiendan el proyecto
Falta de comprensión del proyecto	Baja	Asegurar que los miembros del grupo se reúnan regularmente
Dificultad a la hora de interpretar las conclusiones obtenidas	Media	Asegurar que todos los miembros entienden la totalidad de los resultados así como sus interpretaciones e implicaciones.

Análisis Univariante

Con la intención de realizar un buen análisis descriptivo univariante de los datos previo al pre-procesamiento se ha decidido integrar conjuntamente gráficos y tablas con resultados numéricos para lograr el mejor entendimiento de estos.

Análisis Univariante Numérico

Cuadro 7: Descripción Univariante Variables Numéricas

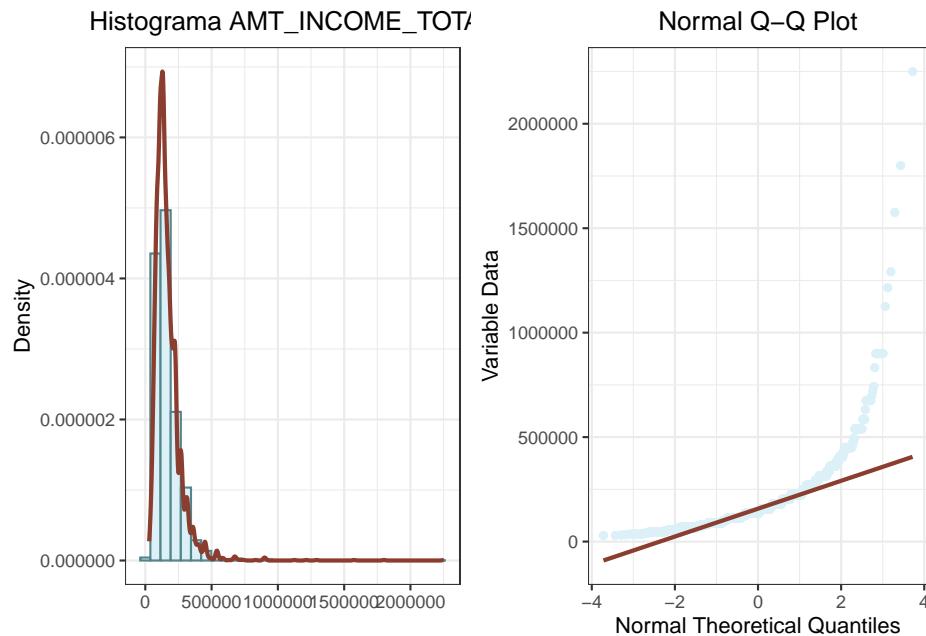
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
AMT_INCOME_TOTAL	1	5000	166848.84	102440.66	135000	153207.22	66717.00	29250	2250000.0	2220750.0	5.15	65.37	1448.73
AMT_CREDIT	2	5000	578795.63	382223.87	504000	530261.19	347595.57	45000	3375000.0	3330000.0	1.35	2.61	5405.46
AMT_ANNUITY	3	5000	26831.19	14163.79	24876	25425.87	12342.64	2673	177826.5	175153.5	1.67	7.98	200.31
DAYS_BIRTH	4	5000	-15586.63	4327.48	-15173	-15457.69	5225.42	-25159	-7711.0	17448.0	-0.22	-1.00	61.20
OWN_CAR_AGE	5	1654	12.81	12.42	10	10.71	7.41	0	65.0	65.0	2.53	7.80	0.31
AMT_GOODS_PRICE	6	4997	515795.79	351507.60	450000	468027.01	333585.00	45000	3375000.0	3330000.0	1.55	3.74	4972.56
CNT_FAM_MEMBERS	7	5000	2.17	0.93	2	2.06	1.48	1	8.0	7.0	0.91	1.24	0.01

Para comenzar, hemos creado una tabla que muestra varios estadísticos de todas las variables numéricas que hemos analizado. Además de los estadísticos más comunes, como la media o la desviación estándar, también hemos incluido otros estadísticos menos conocidos relacionados con la dispersión y centralización de los datos:

- **Trimmed mean:** Este es un estimador que calcula un estadístico para la variable al eliminar los valores más extremos de su distribución. En el caso de la trimmed mean', calcula la media de cada variable utilizando solo los datos que se encuentran en el intervalo [5 %, 95 %]. Al usar la trimmed mean, observamos que la variable **AMT_CREDIT** tiene una media similar a la mediana, lo que indica un alto grado de simetría.
- **Skew:** Este estadístico mide el grado de asimetría de la distribución. Toma valores positivos si la asimetría está hacia la derecha y negativos si está hacia la izquierda (es decir, si la media es menor que la mediana). Un alto grado de asimetría puede indicar la presencia de valores atípicos. Las variables **AMT_ANNUITY** y **AMT_GOODS_PRICE** muestran una asimetría positiva.
- **Kurtosis:** La curtosis es una medida que determina cuán concentrados están los valores de una variable alrededor del centro de la distribución de frecuencias. Un valor de 3 es considerado como el nivel central de curtosis. Una distribución mesocúrtica tiene un cociente de asimetría igual a 3, leptocúrticas por encima de 3 y las platicúrticas por debajo de 3. Las variables **AMT_ANNUITY**, **OWN_CAR_AGE** y sobre todo **AMT_INCOME_TOTAL** tienen coeficientes de curtosis muy elevados, lo que indica distribuciones con colas muy pesadas.
- **SE:** El error estándar es la desviación estándar de la distribución muestral de un estadístico muestral. Es decir, es la desviación típica dividida por la raíz cuadrada del tamaño de la muestra (n). Tanto las variables **AMT_CREDIT** como **AMT_GOODS_PRICE** muestran una variabilidad muy alta.

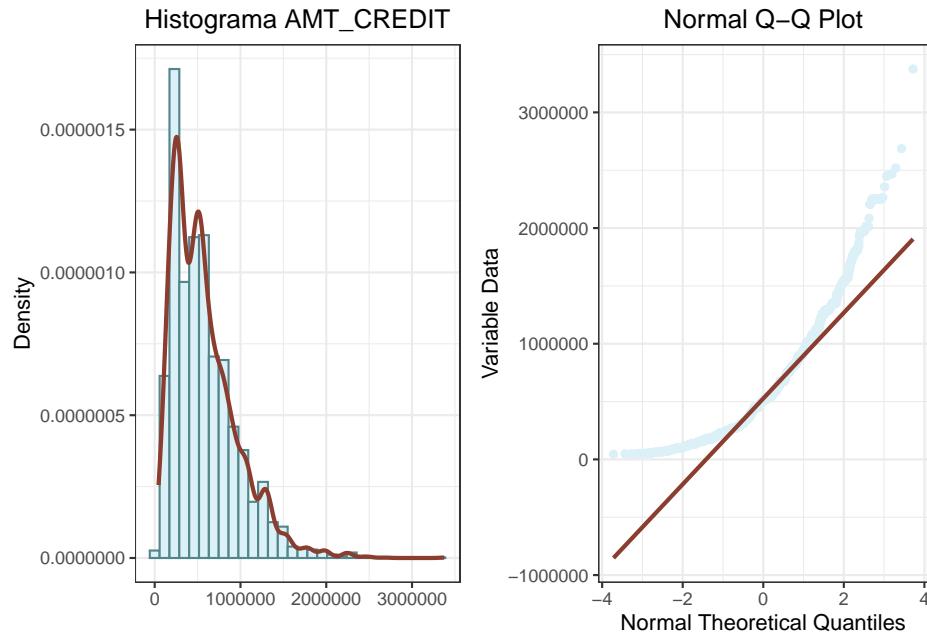
Así pues, tras hacer un análisis general, se procede a realizar un análisis más particular. Para ello, analizaremos cada variable una a una, de forma gráfica:

Figura 1: Análisis Gráfico Variable AMT INCOME TOTAL



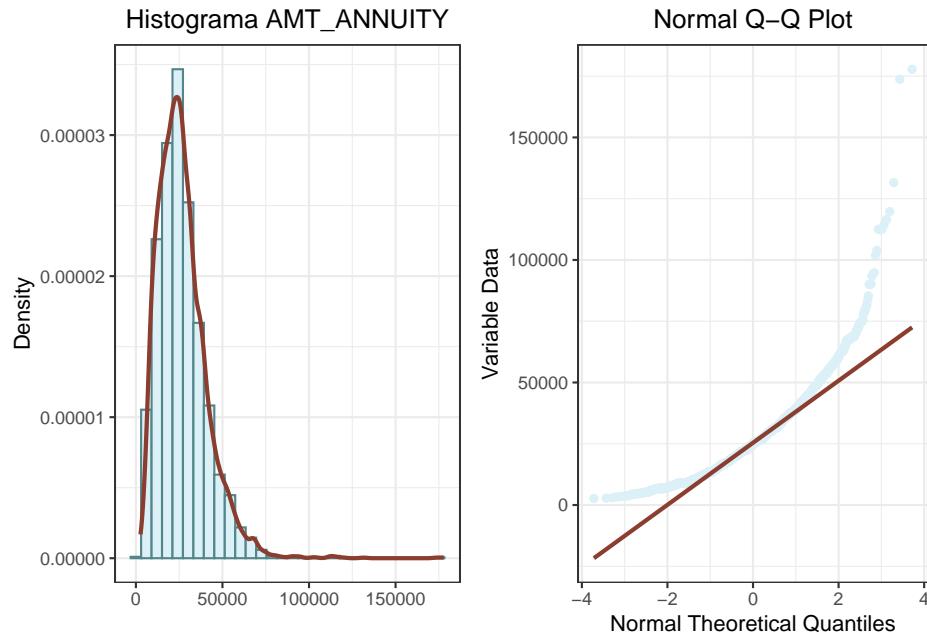
Estos gráficos muestran que los datos de la variable `AMT_INCOME_TOTAL` no siguen una distribución normal y parecen seguir una distribución exponencial. Esto tiene sentido, ya que la distribución de los ingresos totales de los individuos en una población generalmente no sigue una distribución normal. Además, al observar los resultados del test de normalidad Shapiro-Wilk, se confirma la hipótesis anterior sobre la no normalidad de los datos, ya que el p-valor obtenido es $1,0321299 \times 10^{-68}$.

Figura 2: Análisis Gráfico Variable AMT CREDIT



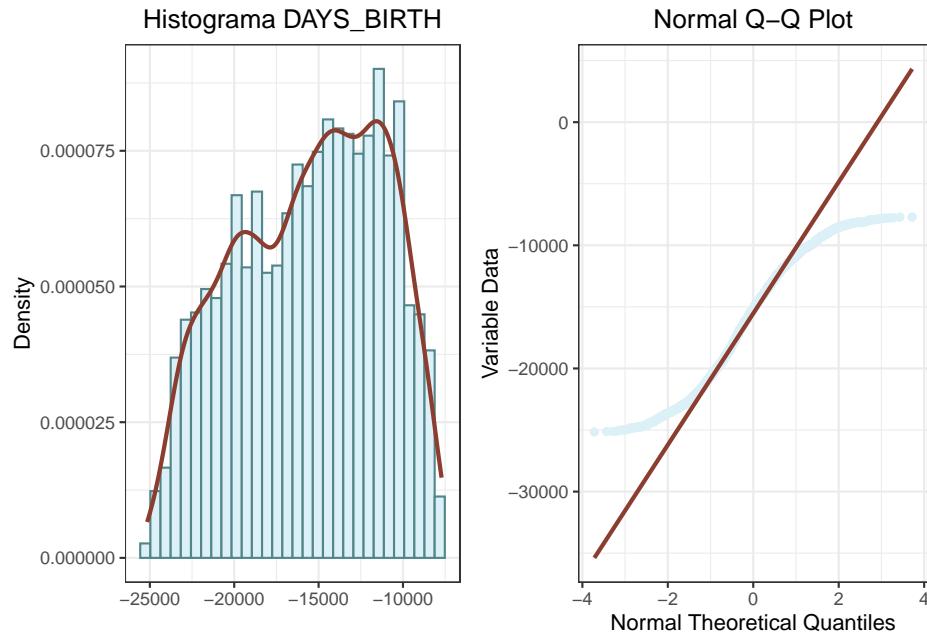
Al igual que en el caso anterior, la variable tampoco sigue una distribución normal, lo cual se confirma además por el test de Shapiro-Wilk con un p-valor de $3,1426414 \times 10^{-49}$. Parece que sigue una distribución exponencial.

Figura 3: Análisis Gráfico Variable AMT ANNUITY



Como en el caso anterior, la variable sigue aparente exponencial, con p-valor $5,171852 \times 10^{-48}$ del test de Shapiro Wilk.

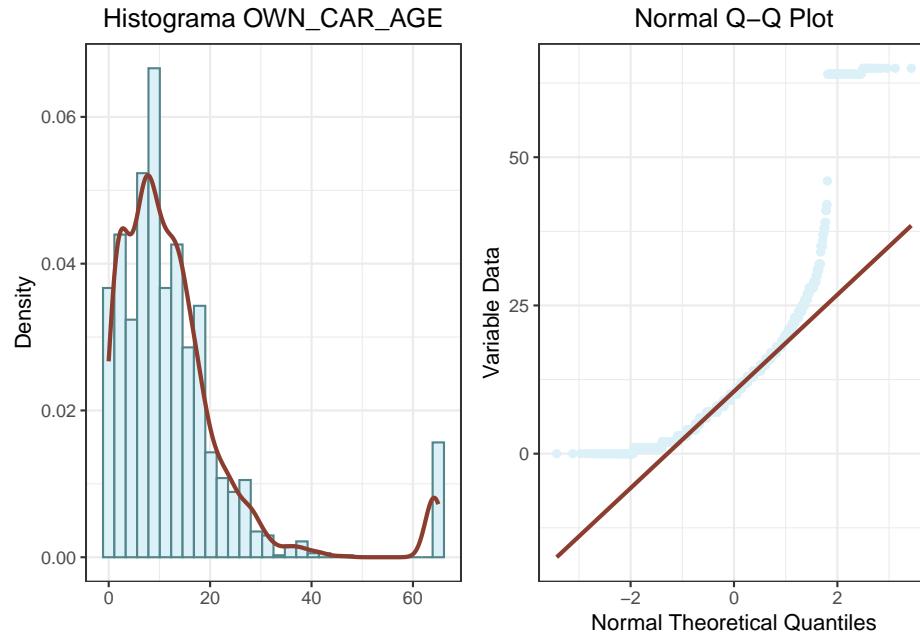
Figura 4: Análisis Gráfico Variable DAYS BIRTH



Como se puede apreciar en el histograma, la variable “Days Birth” presenta valores negativos. Esto se debe a que los datos indican la cantidad de días transcurridos desde el nacimiento del individuo hasta el momento

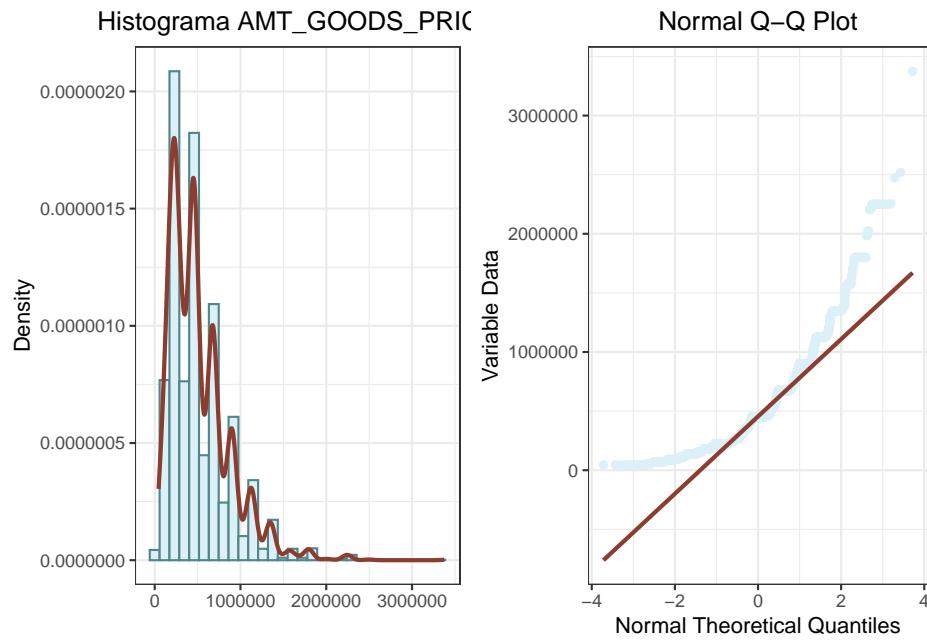
en que solicitó el crédito. Por lo tanto, es necesario transformar los datos para que sean positivos y modificar la variable de manera que represente las edades de los sujetos en años, lo que facilitará un mejor tratamiento y comprensión de los resultados.

Figura 5: Análisis Gráfico Variable OWN CAR AGE



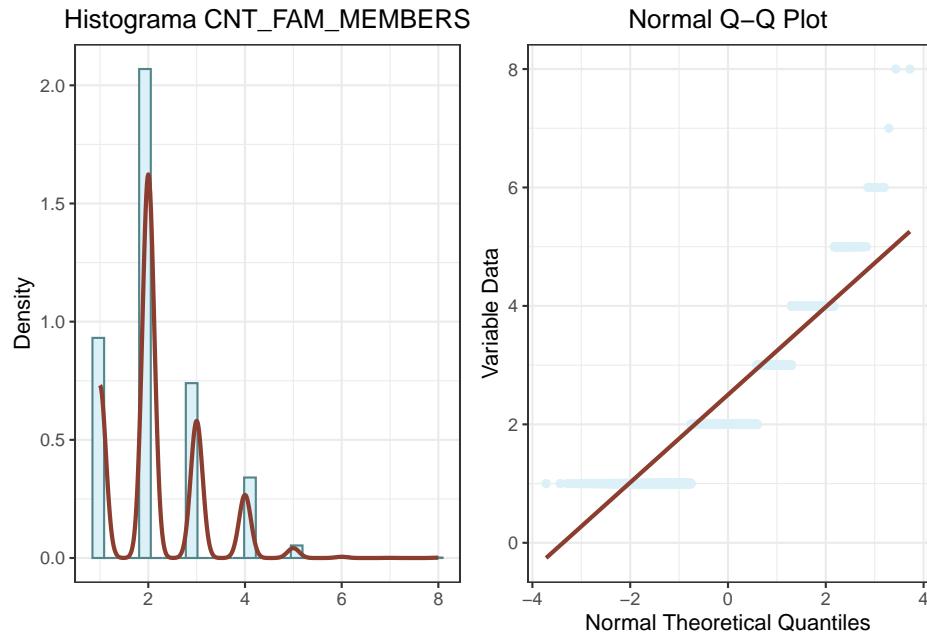
En la variable ‘Own car age’, también se observa que no sigue una distribución normal, como lo demuestra el test de Shapiro-Wilk con un p-valor de $3,2161311 \times 10^{-45}$. Se puede notar una alta concentración de datos alrededor de los 10 años, lo que muestra una estructura similar a una distribución exponencial. Por otro lado, también se observa una fuerte concentración de datos en los 60 años.

Figura 6: Análisis Gráfico Variable AMT GOODS PRICE



Al igual que en el caso anterior, la variable parece seguir una distribución exponencial, y la normalidad se rechaza con un p-valor de $6,9961429 \times 10^{-53}$. Aunque el Q-Q Plot y el histograma muestran una concentración de datos de forma periódica, una posible explicación podría ser que los bienes de alto costo tienden a tener precios redondeados o cantidades enteras en lugar de valores precisos. Por ejemplo, la moda podría ser 450000.

Figura 7: Análisis Gráfico Variable CNT FAM MEMBERS



La variable que representa el número de hijos, al ser discreta, no debe evaluarse como si siguiera una distribución normal. Aun así, es importante tener en cuenta que la mayoría de los clientes viven en pareja.

Análisis Univariante Categórico

Tras haber completado el análisis univariante numérico se procede a hacer el análisis categórico. En la siguiente tabla se presenta un resumen general sobre ellas:

Cuadro 8: Summary descriptives table

	[ALL] N=5000	N
CODE_GENDER:		
F	3098 (62.0 %)	5000
M	1902 (38.0 %)	
NAME_INCOME_TYPE:		
Businessman	1 (0.02 %)	5000
Commercial associate	1111 (22.2 %)	
Pensioner	763 (15.3 %)	
State servant	306 (6.12 %)	
Working	2819 (56.4 %)	
NAME_EDUCATION_TYPE:		
Academic degree	3 (0.06 %)	5000
Higher education	1018 (20.4 %)	
Incomplete higher	156 (3.12 %)	
Lower secondary	77 (1.54 %)	
Secondary / secondary special	3746 (74.9 %)	
NAME_FAMILY_STATUS:		
Civil marriage	546 (10.9 %)	5000
Married	3095 (61.9 %)	
Separated	320 (6.40 %)	
Single / not married	798 (16.0 %)	
Widow	241 (4.82 %)	
REGION_RATING_CLIENT:		
1	434 (8.68 %)	5000
2	3641 (72.8 %)	
3	925 (18.5 %)	
TARGET:		
0	2865 (57.3 %)	5000
1	2135 (42.7 %)	

Por lo tanto, en la tabla se presentan tanto la frecuencia absoluta como la frecuencia relativa de cada valor posible en cada variable categórica, ya sean dicotómicas o politómicas. Esto facilita la identificación de la moda de manera sencilla.

Una vez se ha realizado un resumen general, se ha procedido a analizar cada variable una a una:

```
data_f = select_if(data, is.factor)
p <- vector("list", length = ncol(data_f))
for (i in 1:ncol(data_f)) {
```

```

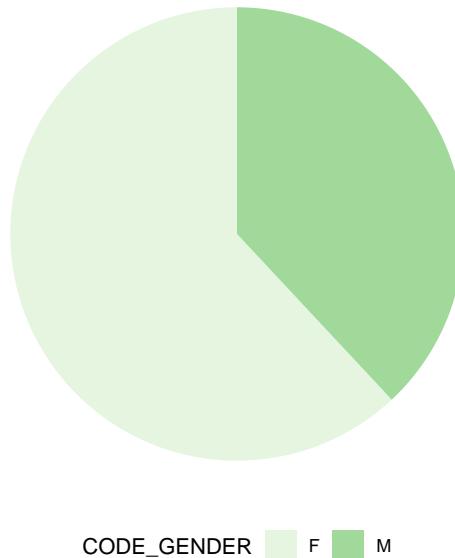
var <- names(data_f)[i]
freq_table <- table(data[[var]])
num_classes <- length(freq_table)

if (num_classes <= 4) {
  p[[i]] <- ggplot(data, aes(x = factor(1), fill = .data[[var]])) +
    geom_bar() +
    coord_polar(theta = "y") +
    labs(x = NULL, y = NULL, fill = var, title = var) +
    theme_void() +
    theme(legend.position = "bottom") +
    scale_fill_brewer(palette = "muted")
} else {
  p[[i]] <- ggplot(data, aes(x = .data[[var]])) +
    geom_bar(fill = "skyblue") +
    labs(x = var, y = "Frecuencia", title = var) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
}
}
}

```

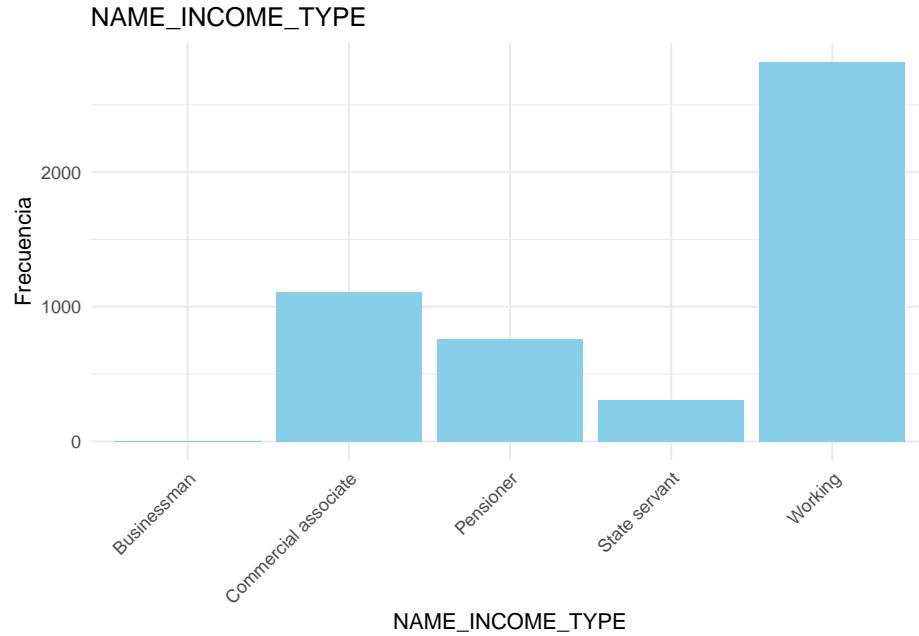
Figura 8: Pie Chart Variable CODE_GENDER

CODE_GENDER



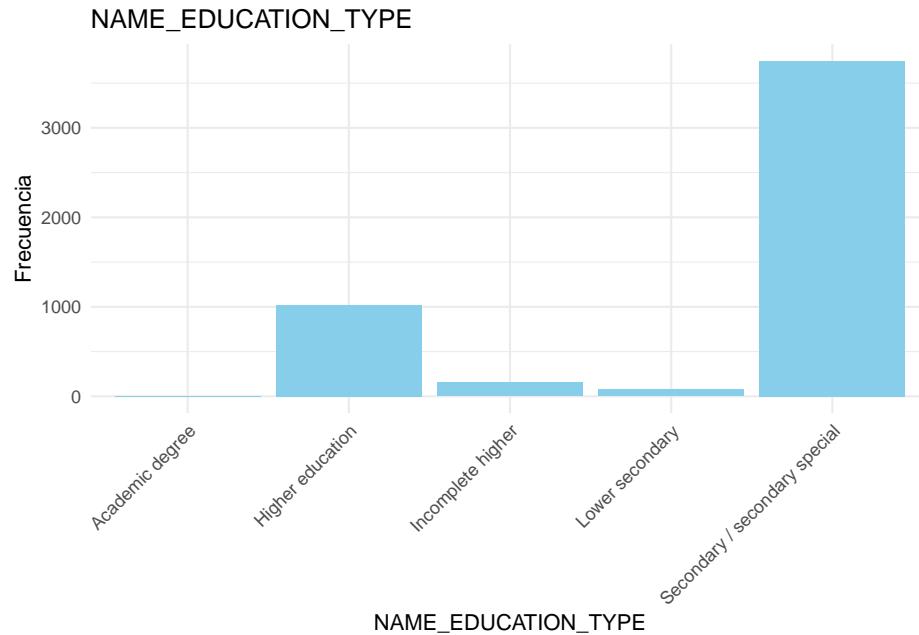
Para empezar, analizamos la variable referida al género. Como se puede apreciar, la gran mayoría de individuos de la base de datos son mujer, con un porcentaje del 61.96 %. Todo el resto de individuos son hombres.

Figura 9: Pie Chart Variable NAME INCOME TYPE



Seguidamente, analizamos la variable NAME_INCOME_TYPE. Gracias al gráfico superior, se puede apreciar que la gran mayoría de clientes son trabajadores, seguido de comerciales aunque bastante lejano. Únicamente disponemos de un empresario y un grupo numeroso de pensionistas.

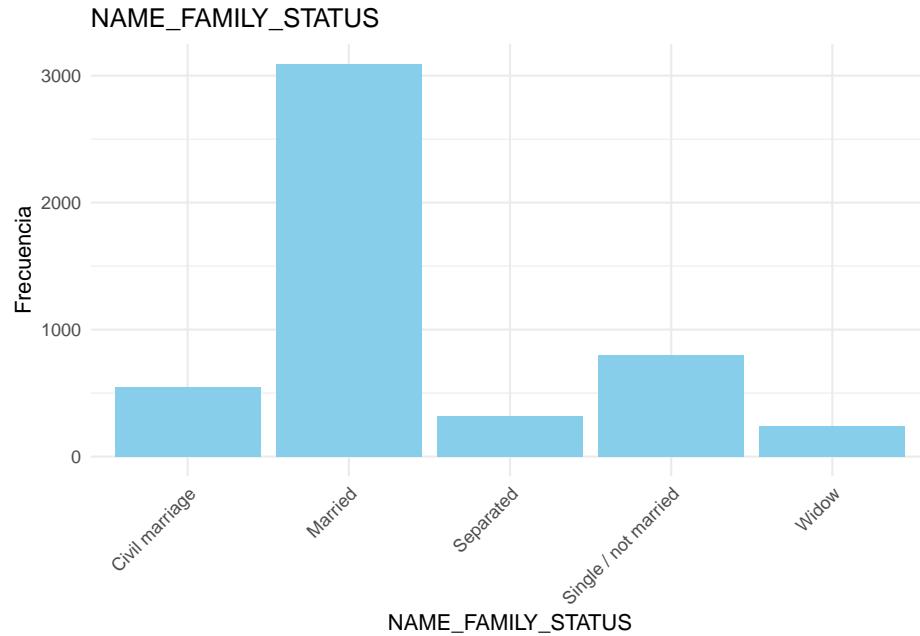
Figura 10: Pie Chart Variable NAME EDUCATION TYPE



Como podemos apreciar, la gran mayoría de los clientes tienen la secundaria como nivel educativo (74.92 %),

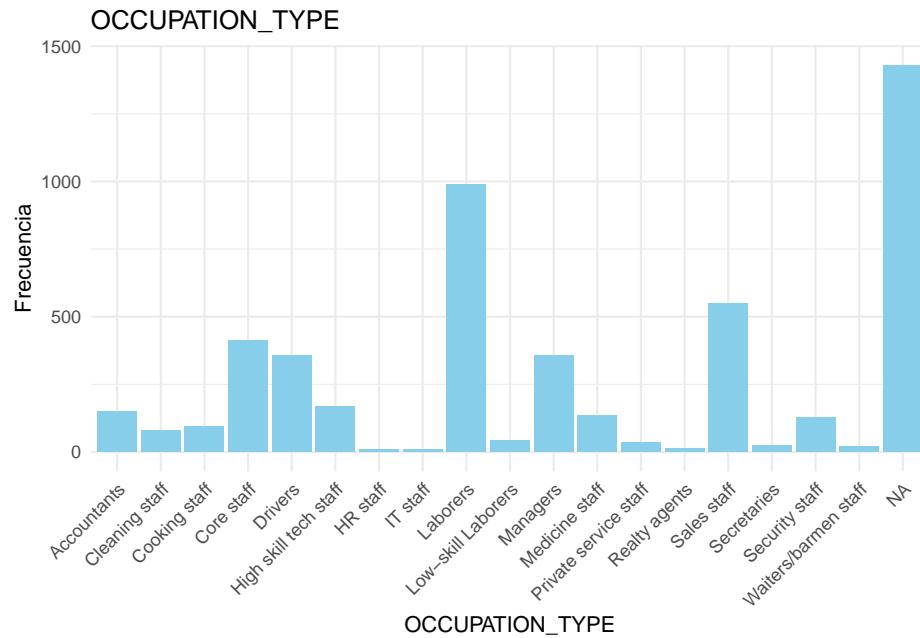
seguido de los universitarios (20.36 %). En general, se podría decir que hay pocos clientes con un nivel educativo bajo.

Figura 11: Pie Chart Variable NAME FAMILY STATUS



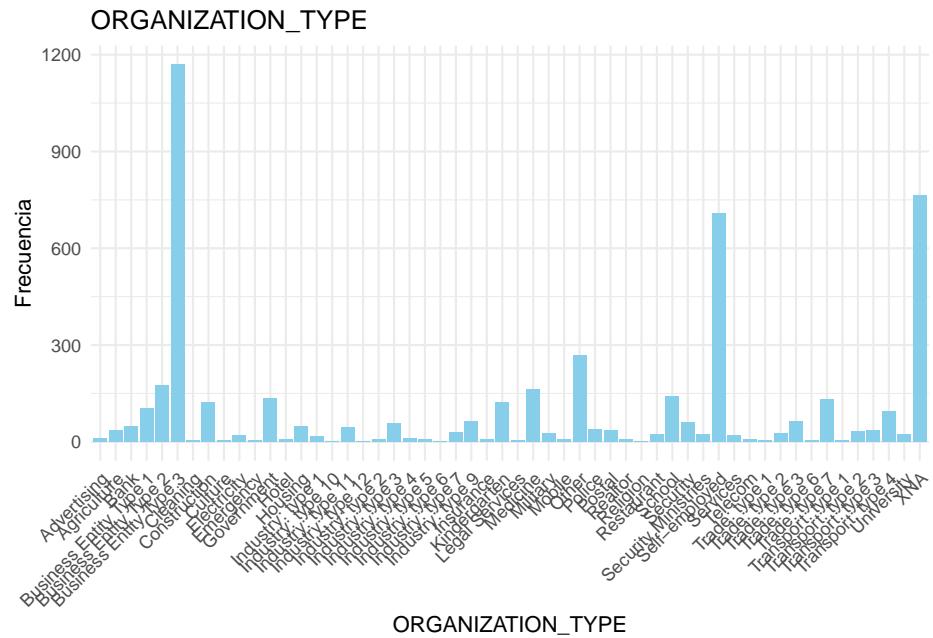
Si pasamos a hablar sobre el estado civil de los clientes, se puede apreciar que la gran mayoría están casados (61.9 %), seguido de los solteros o no casados (15.96 %). El resto de subgrupos es más minoritario.

Figura 12: Pie Chart Variable OCCUPATION TYPE



Seguidamente, si analizamos el tipo de puesto que ocupa cada cliente, vemos que la mayoría son trabajadores en empresas. Sin embargo, esta variable será necesario retocarla, ya que el hecho de que haya tantos NA complica el análisis en general.

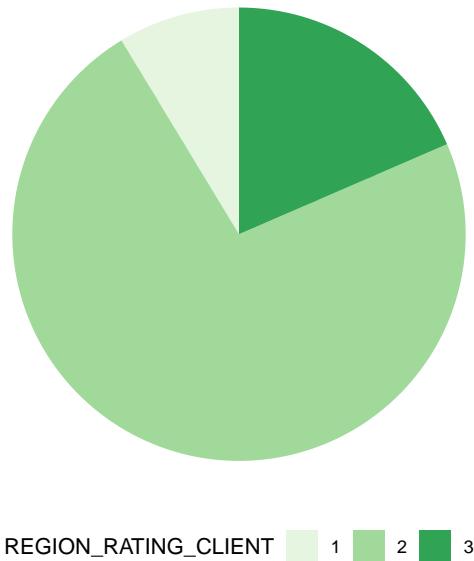
Figura 13: Pie Chart Variable ORGANIZATION TYPE



Sobre el tipo de empresa en el que trabajan los clientes, se puede apreciar que tenemos muchos tipos. Con este nivel de categorías es muy complicado trabajar, así que será necesario agrupar para poder hacer un análisis correcto.

Figura 14: Pie Chart Variable REGION RATING CLIENT

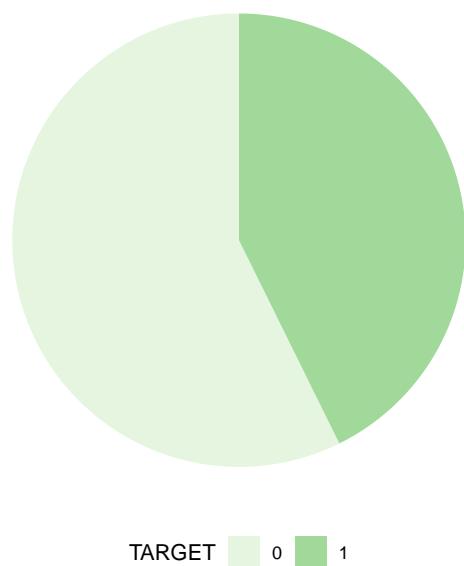
REGION_RATING_CLIENT



Acabando con las variables categóricas, pasamos a hablar de la variable `REGION_RATING_CLIENT`, la cual muestra el nivel de confianza que tiene la empresa sobre la región en la que vive el cliente. Así pues, en general, los clientes viven en áreas con un nivel de confianza medio, con un porcentaje de 72.82% que habitan en estas regiones. Respecto a las otras dos categorías, el 18.5% vive en áreas con mucha confianza y el 8.68%, en áreas con poca confianza.

Figura 15: Pie Chart Variable TARGET

TARGET

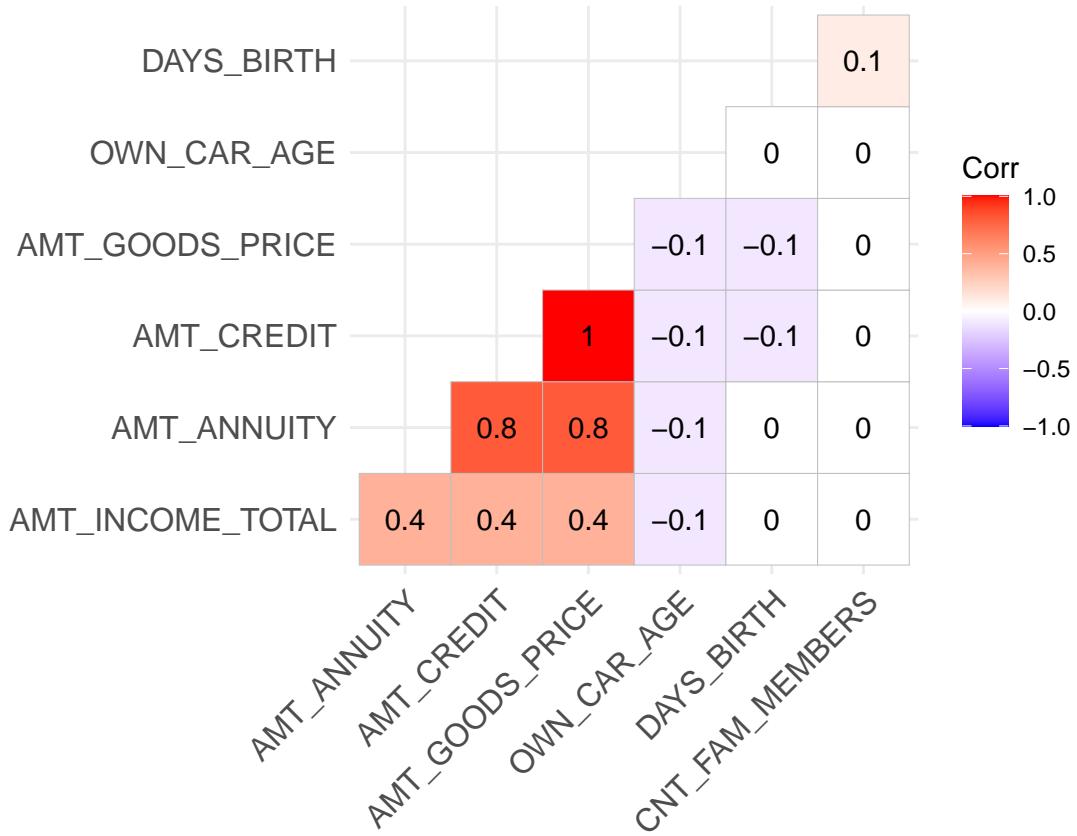


Por último, se analiza la variable respuesta de nuestra base de datos: TARGET. Como se puede apreciar, un 57.3% de los clientes no tienen problemas de solvencia, mientras que un 43.7% los podría presentar.

Análisis Bivariante Numérico

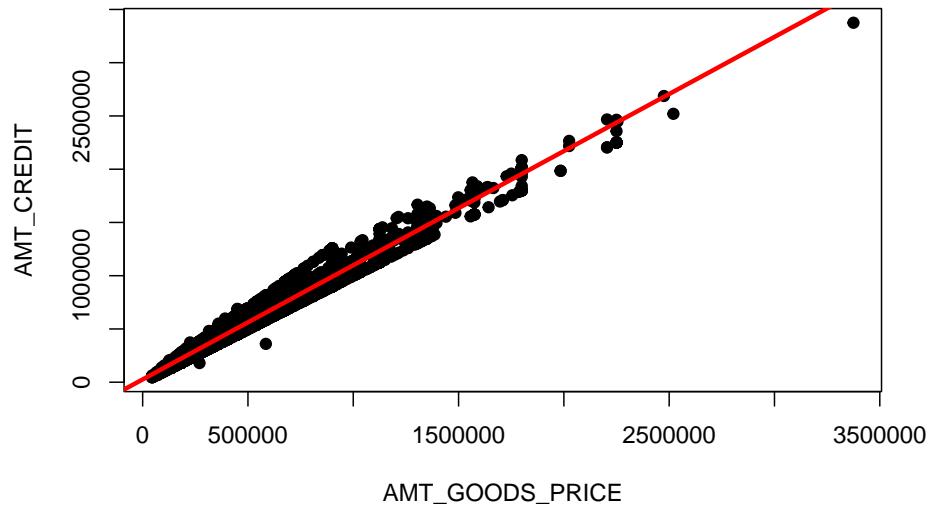
Con el propósito de identificar las relaciones más significativas entre las variables numéricas, se ha creado un gráfico de correlación utilizando la técnica de HeatMap. En este gráfico, los colores indican el grado de dependencia entre las variables numéricas. Cuanto más intenso sea el color, mayor será la relación, y se prestará una mayor atención a estas relaciones en nuestro análisis.

Figura 16: Matriz de Correlaciones para las Variables Numéricas



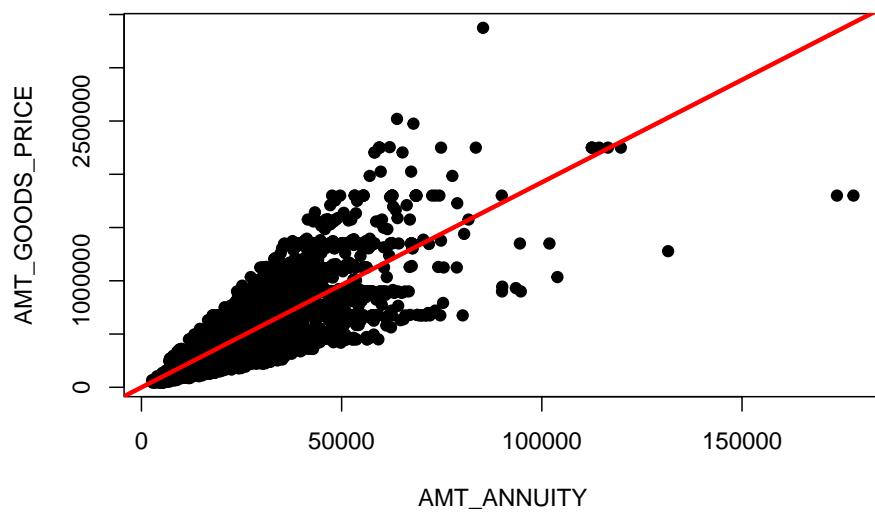
Tras analizar el gráfico, se destacan notables correlaciones entre las variables AMT_CREDIT y AMT_GOODS_PRICE. Esta correlación tiene sentido, ya que los prestamistas suelen otorgar créditos en función del valor del activo que el prestatario desea adquirir. En caso de impago, el prestamista retiene dicho activo como garantía. Además, se observa una alta correlación entre las variables AMT_ANNUITY y AMT_CREDIT. Esto se debe a que un mayor monto de crédito conlleva, de manera directa, una anualidad más elevada, especialmente cuando se busca un período de reembolso similar. También se aprecia una fuerte relación entre las variables AMT_GOODS_PRICE y AMT_ANNUITY, reflejando la conexión entre el crédito y el valor del activo.

Figura 17: Gráfico de dispersión AMT CREDIT vs AMT GOODS PRICE



En este gráfico se evidencia una fuerte correlación entre el valor del bien que el prestatario desea adquirir y la cantidad solicitada para el crédito. Es importante resaltar que los créditos de mayor cuantía muestran una correlación menor con el valor del bien, un aspecto que se explorará con mayor detalle en el transcurso del proyecto.

Figura 18: Gráfico de dispersión Gasto Total en Pescado vs Gasto Total en Fruta

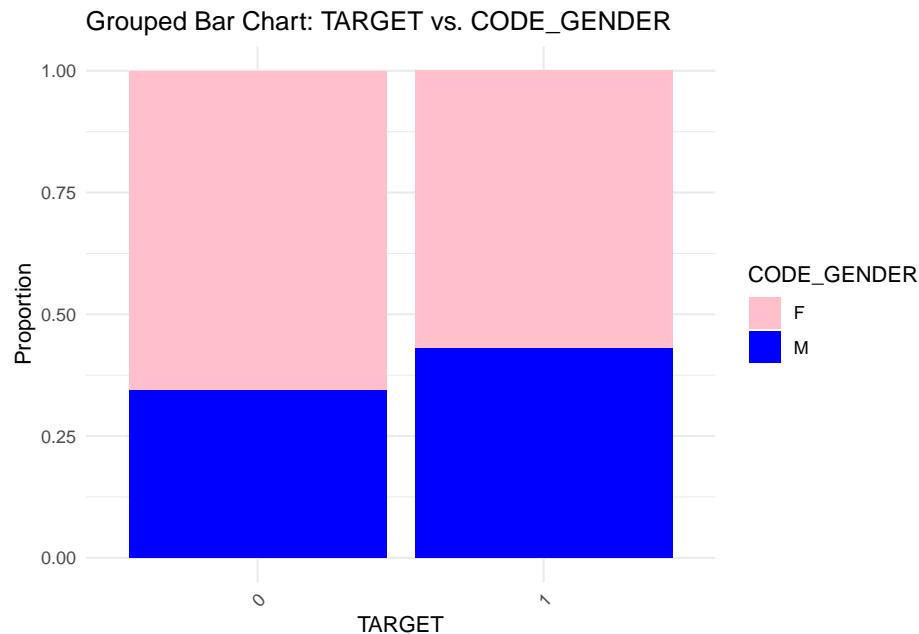


De manera similar, la correlación entre el valor de los bienes y la anualidad también es bastante alta. Es importante señalar que los clientes que posean una relación entre la anualidad y el valor del bien que compren (teniendo en cuenta que el precio del bien es igual al valor del préstamo) serán aquellos que deban destinar una proporción menor de sus ingresos al reembolso de la deuda.

Análisis Bivariante Categórico

Para concluir el análisis descriptivo antes de proceder al procesamiento de los datos, es necesario examinar la relación entre las variables categóricas y las numéricas. Para este propósito, utilizaremos la creación de varios boxplots, lo que nos permitirá presentar nuestras conclusiones de manera precisa y concisa.

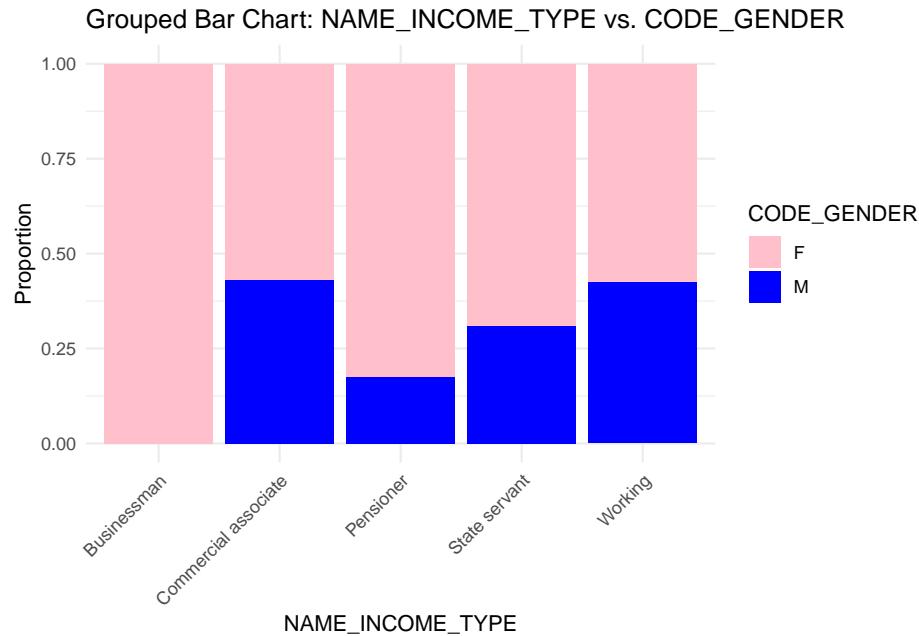
Figura 19: Stacked bar chart TARGET vs CODE_GENDER



En el primer gráfico, se observa que la mayoría de los sujetos son mujeres, sin importar cuál fue el resultado. Sin embargo, existe una diferencia en las proporciones. Menos del 56.96 % de los sujetos que tuvieron dificultades para pagar a tiempo son mujeres, en comparación con el 88.15 % de los individuos que no tuvieron problemas con el pago de sus deudas.

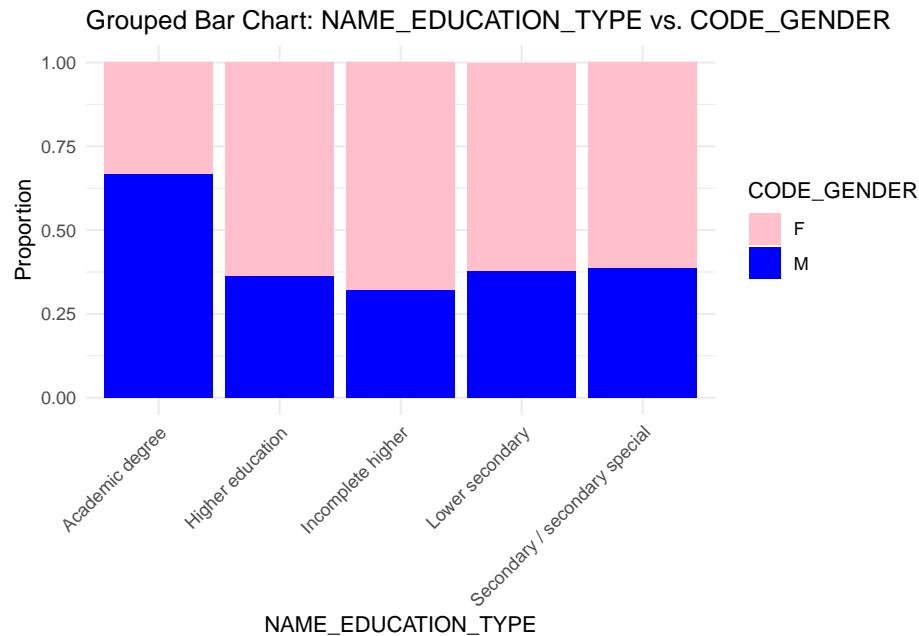
A pesar de que hay más mujeres en la base de datos, un análisis inicial de los datos revela que los hombres han tenido más dificultades para cumplir con los pagos en comparación con las mujeres.

Figura 20: Stacked bar chart NAME INCOME TYPE vs CODE GENDER



En este gráfico de barras apiladas, se compara la relación entre las variables ‘género’ y ‘tipo de ingreso’. Es evidente que la mayoría de los sujetos del estudio son mujeres, representando el 61.96 % de los datos. Como se observa en el gráfico, la mayoría de los pensionistas son mujeres, mientras que hay una proporción mayor de hombres en las categorías ‘Commercial associate’ o ‘Working’. Es importante mencionar que la categoría ‘Businessman’ no es relevante debido a que solo contiene un dato, y este corresponde a una mujer.

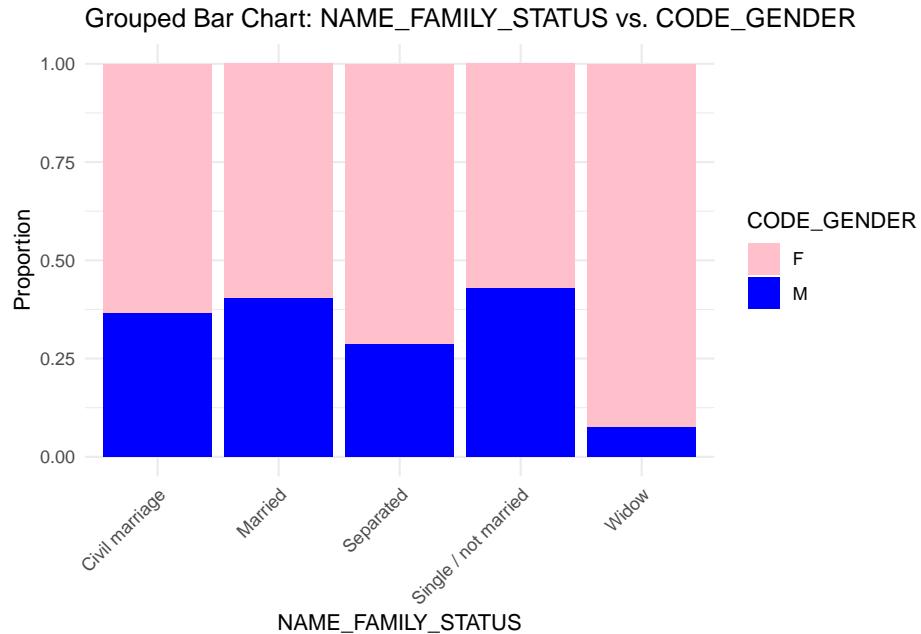
Figura 21: Stacked bar chart NAME EDUCATION TYPE vs CODE GENDER



En el tercer gráfico, se puede observar que la mayoría de las personas con estudios académicos son hombres, representando el 0.11 % de esta categoría. En contraste, las categorías “Higher education,” “Incomplete higher,” “Lower secondary,” y “Secondary/secondary special” están compuestas mayoritariamente por mujeres, con un porcentaje similar a la cantidad de datos de mujeres que hay en la base de datos.

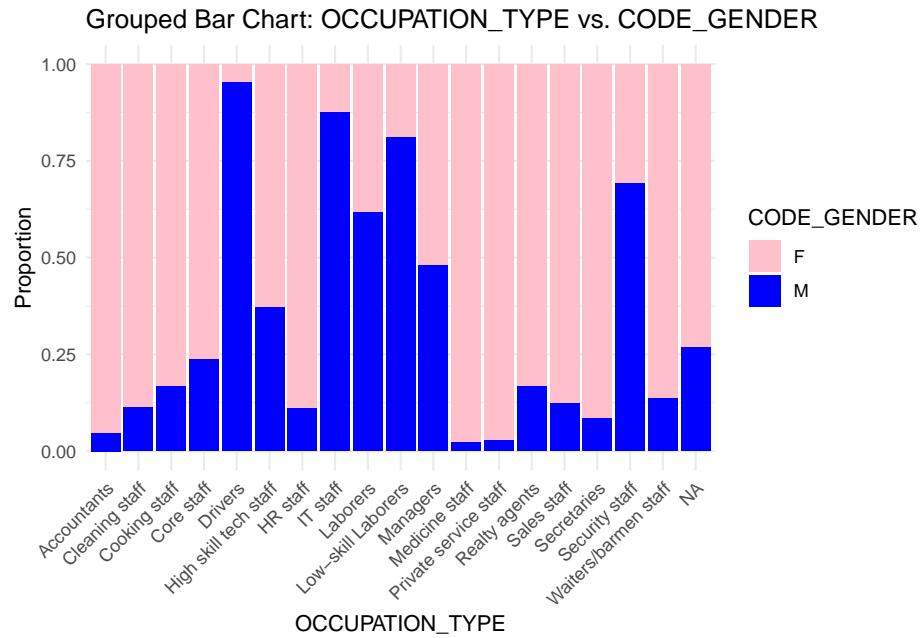
Cabe señalar que la clase “Academic degree” solo cuenta con tres sujetos, mientras que las categorías “Higher education” y “Secondary/secondary special” concentran el 95.28 % de los datos. Estos porcentajes se mantienen aproximadamente constantes en las diferentes categorías, lo que sugiere que la variable NAME_EDUCATION_TYPE no es visualmente relevante para poder entender mejor la estructura de los datos.

Figura 22: Stacked bar chart NAME FAMILY STATUS vs CODE GENDER



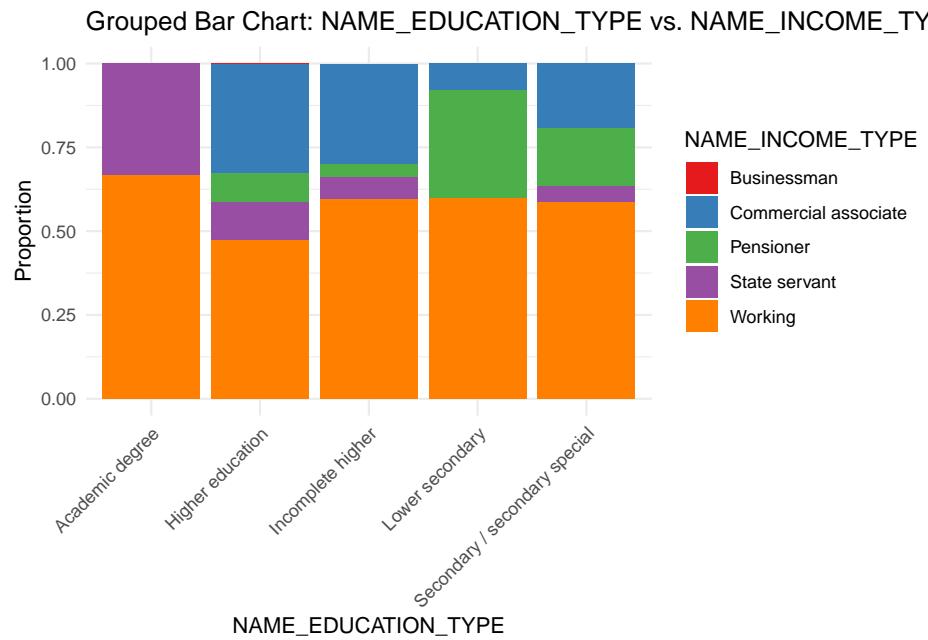
Este gráfico nos muestra la distribución del estatus civil con respecto al sexo de la persona. Se aprecia claramente como la gran mayoría de cónyugues supervivientes son mujeres, mientras que hay una menor desproporción en cuanto a la cantidad de personas solteras o no casadas. La clase con mayor frecuencia es la de casados, con un 61.9 % de mujeres, muy similar al porcentaje de mujeres respecto al total de los datos.

Figura 23: Stacked bar chart OCCUPATION_TYPE vs CODE_GENDER



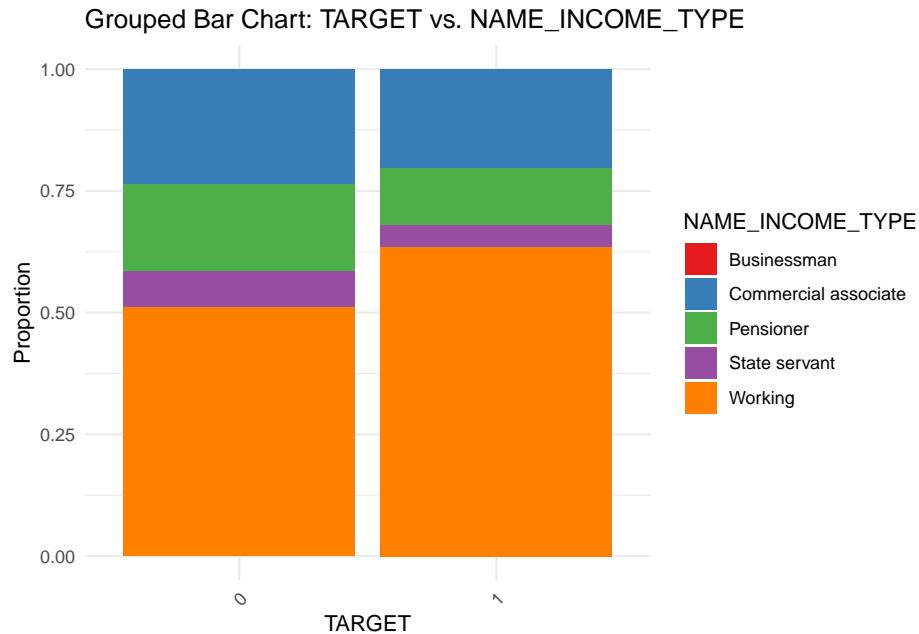
En este último gráfico respecto al género se observa la relación de esta categoría conjuntamente con el tipo de ocupación laboral. En un primer análisis visual se observa como las clases “Drivers”, “IT staff”, “Laborers” y “Security staff”, mientras que las mujeres predominan en la mayoría del resto de variables. Teniendo en consideración la frecuencia de los datos podemos determinar que el 70.6098843 % de los hombres son “Laborers” o “Drivers”. Por último cabe destacar que hay el 28.6 % de los datos son missing, por lo que se imputarán en el preprocessing, ya que no suponen una gran cantidad respecto al total de datos de la variable OCCUPATION_TYPE.

Figura 24: Grouped bar chart NAME INCOME TYPE vs NAME EDUCATION TYPE



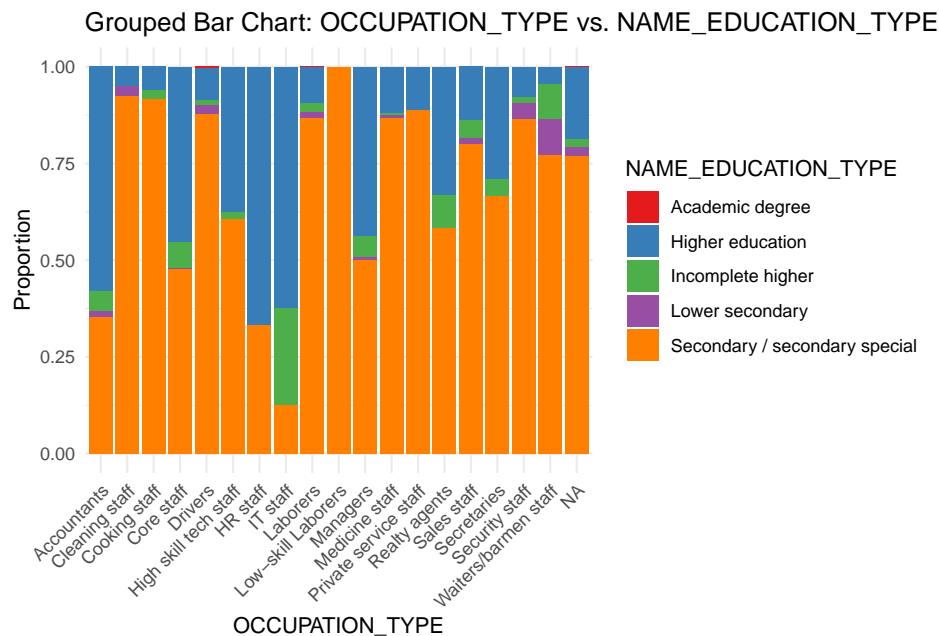
En este gráfico se analiza la relación entre el nivel de educación y el tipo de ingreso. Como se observa, la mayoría de los trabajadores en empleos del sector privado convencional presentan una diversidad de niveles educativos, mientras que aquellos con estudios académicos tienden a trabajar para el sector público. Vale la pena señalar que un porcentaje significativo de los cónyuges sobrevivientes tiene únicamente educación secundaria. Esto podría deberse al hecho de que estos trabajadores son de mayor edad y, en su momento, las oportunidades de acceder a educación superior eran limitadas.

Figura 25: Grouped bar chart NAME INCOME TYPE vs TARGET



En lo que respecta a la variable TARGET, se observa una disparidad en la capacidad de pago de los clientes en el sector privado, siendo los pensionistas y los comerciales quienes presentan proporcionalmente menos dificultades.

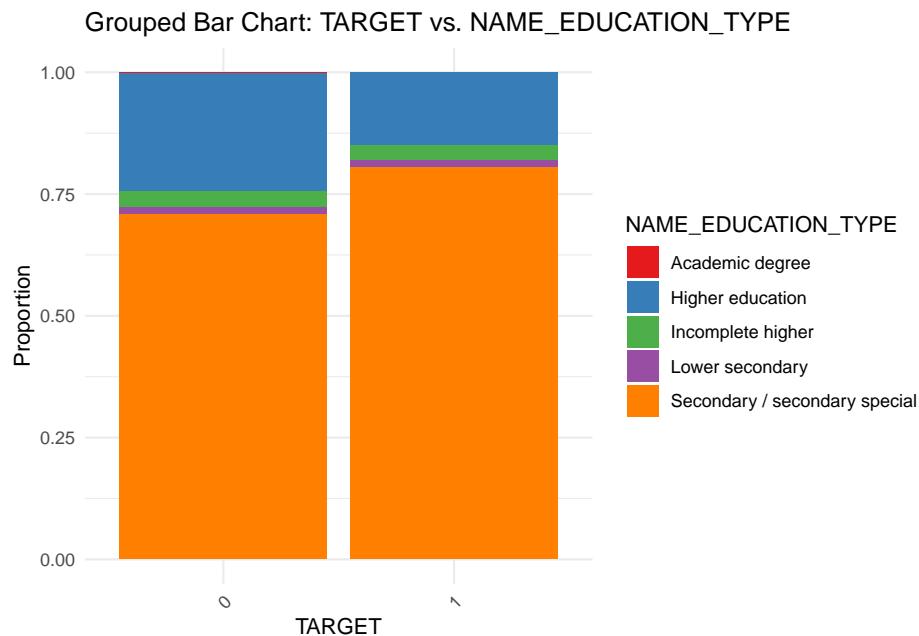
Figura 26: Grouped bar chart NAME EDUCATION TYPE vs OCCUPATION TYPE



En este gráfico se confirma la idea de que los trabajadores con niveles educativos más altos tienden a ocupar

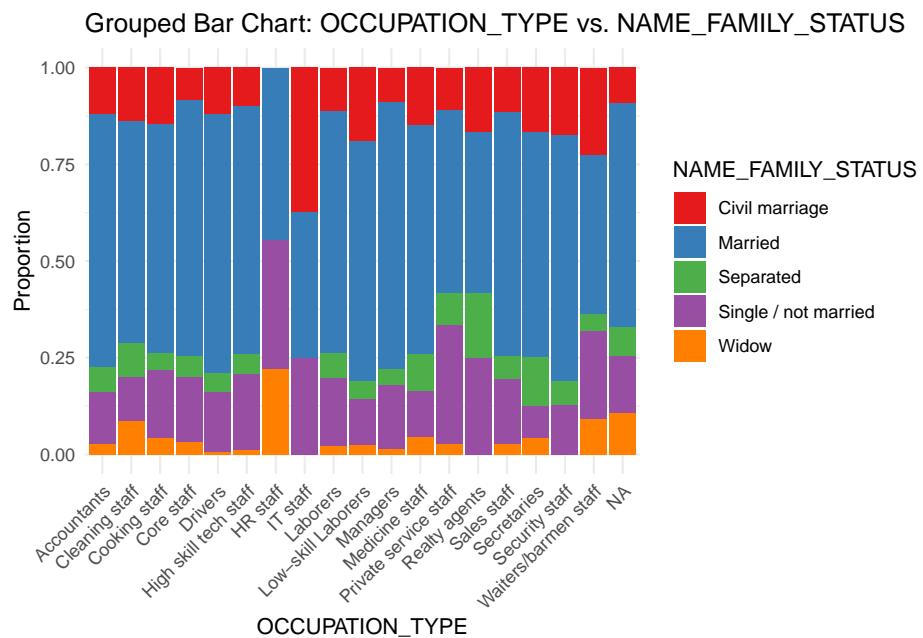
puestos de trabajo que requieren un mayor nivel de conocimientos técnicos, mientras que aquellos con niveles educativos más bajos suelen desempeñar empleos que demandan menos destrezas técnicas.

Figura 27: Grouped bar chart NAME EDUCATION TYPE vs TARGET



En lo que respecta al nivel de educación, es notable que aquellos trabajadores con un nivel educativo más bajo son quienes enfrentan mayores dificultades para cumplir con sus pagos de manera consistente.

Figura 28: Grouped bar chart NAME FAMILY STATUS vs OCCUPATION TYPE



En este gráfico se analiza la relación entre la ocupación de los individuos y el estado civil de ellos mismos. Como se observa, la mayoría de los trabajadores de cualquier sector están casados, muchos por la iglesia y unos pocos civilmente. Vale la pena señalar que un porcentaje significativo de los cónyuges sobrevivientes trabajan en recursos humanos. Esto podría deberse al hecho de que estos trabajadores son de mayor edad y, en su momento, las oportunidades de acceder este tipo de empleos eran más altas.

Seguidamente, se hará el preprocessing para corregir muchos de los problemas que se han presentado.

Preprocessing de los datos

Para realizar el preprocessamiento de los datos, será óptimo seguir los pasos propuestos por Karina Gibert con el objetivo de desarrollar correctamente el KDD y, así, obtener conclusiones óptimas a partir de nuestros datos.

Para ello, seguiremos 4 grandes bloques:

- Limpieza de datos y estandarización de formato
- Detección y tratamiento de missings
- Detección y tratamiento de outliers
- Feature Engineering

Limpieza de datos y estandarización de formato

Una vez hemos realizado la descriptiva preprocessing y hemos identificado el número de valores missing en nuestra base de datos, es óptimo analizar todas las variables una a una, así como algunas variables categóricas a las cuales se les puede reducir el número de categorías.

Para empezar, se puede apreciar que la variable `OCCUPATION_TYPE` tiene un total de 18 categorías:

Cuadro 9: Distribución inicial de la variable `OCCUPATION_TYPE`

Categoría	Frecuencia
	0
Accountants	150
Cleaning staff	80
Cooking staff	96
Core staff	412
Drivers	356
High skill tech staff	170
HR staff	9
IT staff	8
Laborers	987
Low-skill Laborers	42
Managers	355
Medicine staff	135
Private service staff	36
Realty agents	12
Sales staff	550
Secretaries	24
Security staff	126
Waiters/barmen staff	22
NA	1430

Una buena idea sería combinar algunas categorías con el objetivo de reducir el número de categorías y, además, aumentar el número de individuos por categoría. Seguidamente, se muestran los cambios realizados, donde se han agrupado todos los individuos en 5 categorías en función del capital humano empleado para su puesto:

- Low skill laborers: Engloba las categorías de “security staff”, “cooking staff”, “cleaning staff”, “drivers”, “low skill laborers”, “waiters staff”.
- Low-mid skill laborers: Engloba las categorías de “secretaries”, “private service staff” y “laborers”.
- Mid skill laborers: Engloba las categorías de “accountants”, “HR staff” y “sales staff”.
- Mid-high skill laborers: Engloba las categorías de “IT staff”, “realty agents” y “core staff”.
- High skill staff: Engloba las categorías de “high skill tech staff”, “managers” y “medicine staff”.

Cuadro 10: Distribución final de la variable OCCUPATION TYPE

Categoría	Frecuencia
High skill laborers	660
Low-mid skill laborers	1047
Low skill laborers	722
Mid-high skill laborers	432
Mid skill laborers	709
NA	1430

Este proceso lo repetiremos con la variable ORGANIZATION_TYPE:

Cuadro 11: Distribución inicial de la variable ORGANIZATION TYPE

Categoría	Frecuencia
Advertising	10
Agriculture	35
Bank	47
Business Entity Type 1	104
Business Entity Type 2	176
Business Entity Type 3	1169
Cleaning	4
Construction	124
Culture	4
Electricity	20
Emergency	5
Government	135
Hotel	9
Housing	49
Industry: type 1	18
Industry: type 10	1
Industry: type 11	45
Industry: type 12	1
Industry: type 2	9
Industry: type 3	56
Industry: type 4	12
Industry: type 5	8
Industry: type 6	3
Industry: type 7	28
Industry: type 9	63
Insurance	9
Kindergarten	121
Legal Services	5
Medicine	162
Military	27
Mobile	9
Other	269
Police	40
Postal	37
Realtor	8
Religion	2
Restaurant	24
School	142
Security	61
Security Ministries	24
Self-employed	708
Services	19
Telecom	8
Trade: type 1	5
Trade: type 2	26
Trade: type 3	63
Trade: type 6	6
Trade: type 7	133
Transport: type 1	5
Transport: type 2	34
Transport: type 3	35
Transport: type 4	96
University	24
XNA	763
NA	0

Como se puede apreciar, en este caso disponemos de muchísimas categorías, pero es de destacar la categoría XNA, la cual deberíamos sustituir a NA, para después poder imputarle algún valor. Así pues, se ha agrupado cada categoría profesional en función del sector al que se dedica el individuo. Así, la distribución final es la siguiente:

Cuadro 12: Distribución final de la variable ORGANIZATION TYPE

Categoría	Frecuencia
Business and bank	1505
Education	287
Industry and construction	368
Medicine	162
Other	390
Personal services	155
Public services	251
Self-employed	708
Trade and telecom	241
Transport	170

Ahora, esta variable pasa a tener 10 categorías, las cuales representan los diferentes sectores presentes en la economía presente hoy en día.

Así pues, el resto de variables tienen una uniformidad evidente: se puede apreciar cómo las variables categóricas presentan un número de categorías pequeño y, por parte de las variables numéricas, todas están expresadas en las mismas unidades, de forma que no habrá problemas con la manipulación de éstas.

Detección y tratamiento de missings

Para este apartado, trataremos de identificar aquellos valores desconocidos y valorar sobre su aleatoriedad para, posteriormente, imputar valores. Para empezar, es de destacar cómo hay 47 individuos con un coche de 64 años y 11 con un coche de 65. Si nos fijamos en la distribución de esta variable, es muy extraño que haya tantos individuos con valores atípicos, ya que el siguiente valor máximo es 46. Así, se potará por imputar valores nulos a estos individuos.

Seguidamente, pasaremos a imputar diferentes valores a aquellas variables donde hay observaciones sobre las cuales se desconocen sus valores reales. Este paso es necesario, ya que el hecho de disponer de valores desconocidos (también conocidos como NA) dificulta el análisis posterior de la variable.

Una vez hemos recategorizado todas aquellas variables que presentaban problemas, el número de NA por variables es el siguiente:

Cuadro 13: Missings por variable

Categoría	Frecuencia
AMT_INCOME_TOTAL	0
AMT_CREDIT	0
AMT_ANNUITY	0
DAYS_BIRTH	0
OWN_CAR_AGE	3404
AMT_GOODS_PRICE	3
CNT_FAM_MEMBERS	0
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	1430
ORGANIZATION_TYPE	763
REGION_RATING_CLIENT	0
TARGET	0

Una vez tenemos identificados todos los valores missing de nuestra base de datos, será necesario identificar si éstos son completamente aleatorios (MCAR), aleatorios (MAR), o no aleatorios (MNAR). Para ello, realizaremos el test de Little, el cual indica si los missings disponibles en la base de datos son fruto del azar o si siguen un patrón.

Para este test, diremos que los datos no siguen un patrón si no se rechaza hipótesis nula o, alternativamente, si no encuentra patrones entre los missings. Así pues, este es el resultado:

Cuadro 14: Test de Little

statistic	df	p.value	missing.patterns
2913.59628881402	79	0	7

Como se puede apreciar, el algoritmo ha detectado 7 patrones entre los valores missing, de forma que no se puede decir que hay un patrón aleatorio, de forma que calificaremos nuestros valores missing como MNAR.

Seguidamente, imputaremos los valores por los tres métodos de imputación conocido, pero antes de imputar los valores numéricos, será necesario pasar los NA a categoría `unknown`.

Seguidamente, toca imputar los NA disponibles en las variables numéricas de nuestros datos. Para ello, utilizaremos tres métodos distintos: kNN, MiMMi y MICE. Posteriormente, se comparará la imputación entre estos métodos y se seleccionará el método que resulte una distribución más parecida a la original antes de imputar.

Imputación por criterios estadísticos

En este caso, el objetivo será imputar en función de criterios estadísticos básicos. Para ello, se procederá a imputar valores en función de la media estadística o algún otro estadístico central de distribución.

Imputación por kNN

El algoritmo K-Nearest Neighbors (KNN), es un método de clasificación supervisada, que utiliza la proximidad para hacer clasificaciones o predicciones sobre un punto de datos desconocido. El algoritmo, utiliza

un hiperparámetro llamado “k”, que representa el número de vecinos más cercanos y el cual se ha obtenido mediante el cálculo de $k = \sqrt{n}$.

A continuación, se crean dos objetos: `fullVariables`, que corresponde a las variables que no presentan ningún dato faltante y `uncompleteVars`, que guarda las variables con missings.

Como se puede observar, se obtiene la imputación de los valores faltantes en el dataframe `df_knn` utilizando el algoritmo descrito previamente.

Imputación por MiMMi

La imputación por MiMMi se realiza utilizando un enfoque basado en clústeres y se utiliza la distancia de Gower como métrica de distancia para medir la similitud entre observaciones.

La función `uncompleteVar` se define para verificar si hay valores faltantes (representados como NA) en un vector dado.

La función `Mode` se define para calcular la moda de un vector. Esta función se utiliza más adelante para imputar valores faltantes en variables categóricas.

Se define la función MiMMi.

Se usa la función MiMMi y se obtienen los resultados imputados.

Imputación por MICE

Por último, se recurrirá a imputar a través del MICE como último método de imputación de valores numéricos. El MICE (Multiple Imputation by chained Equations) se basa en un método iterativo a partir del cual se resuelven ecuaciones consecutivamente con el objetivo de imputar valores de la forma más aproximada posible. Así pues, es momento de imputarlo:

Decisión del método de imputación elegido

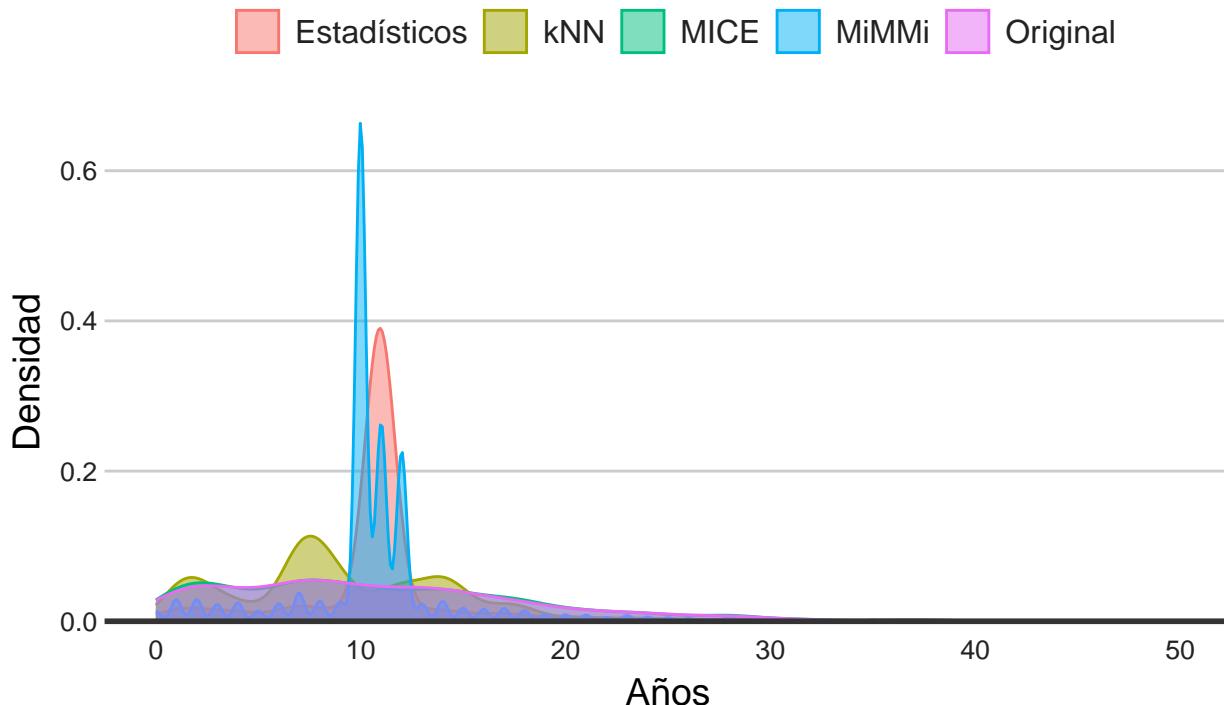
Llegados a este punto, en el momento de seleccionar el método de imputación elegido para el método de imputación final. En nuestro caso, como únicamente disponemos de dos variables numéricas con missings, podemos comparar la función de densidad de los datos originales contra los imputados por cada método. Así pues, vamos a mirar variable por variable:

OWN_CAR_AGE

Esta variable es la que presenta más valores no disponibles en nuestra base de datos, de forma que se acepta un mayor margen de error en cuanto a la imputación de valores se refiere. Así, la densidad resultante para cada método es la siguiente:

Distribución de la variable OWN_CAR_AGE

Por los 4 métodos de imputación



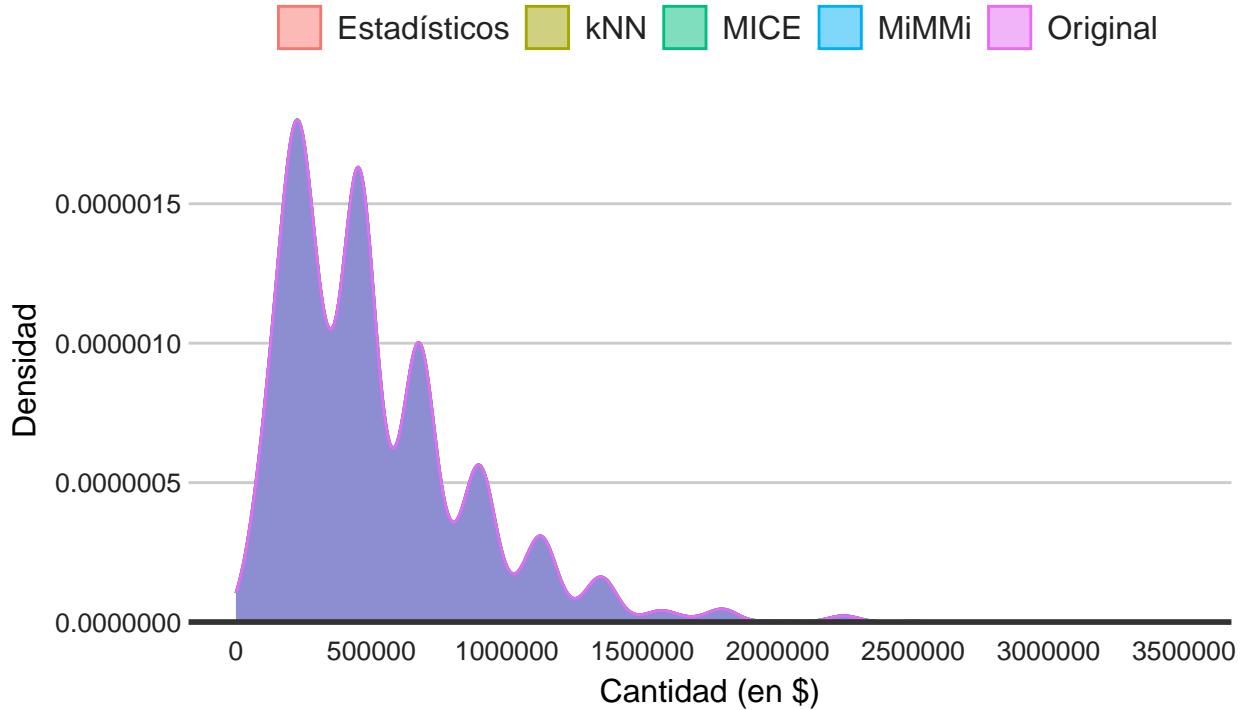
Como se puede apreciar, hay tres métodos de imputación que claramente se alejan mucho de la distribución inicial de los datos: criterios estadísticos, kNN y MiMMi. Así pues, se puede apreciar como el MICE es el algoritmo que aproxima la densidad de los datos a los originales, de forma que este será el método escogido.

AMT_GOODS_PRICE

Como se ha visto previamente en el descriptiva preprocessing, esta variable únicamente presentaba 3 NA, de forma que la densidad en todos los métodos será muy similar:

Distribución de la variable AMT_GOODS_PRIC

Por los 4 métodos de imputación



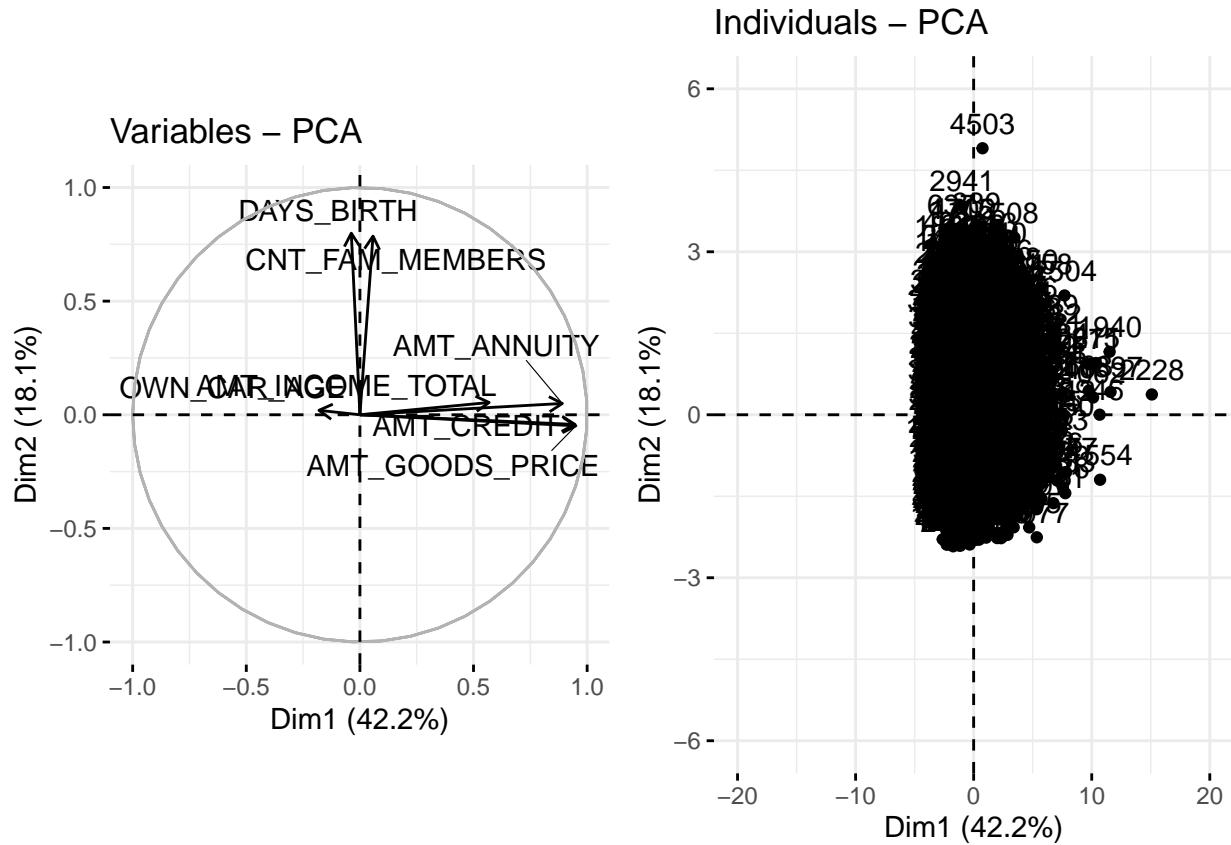
Como se puede apreciar, todos los métodos retornan una estimación similar de la densidad, por lo que se podría decir que es indiferente escoger un método en concreto. De esta forma, se decide usar el MICE como método de imputación final seleccionado.

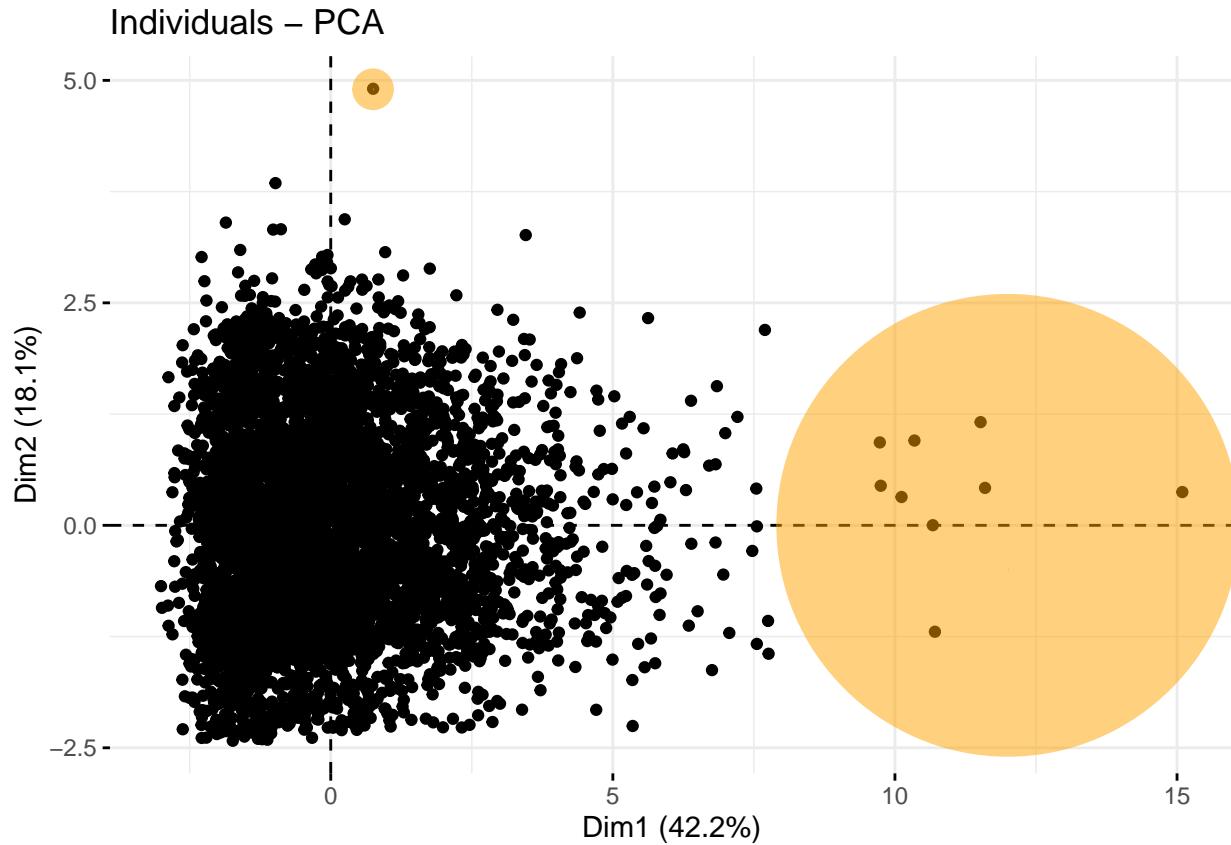
He aquí una tabla resumen sobre los resultados obtenidos acerca de cuál es el mejor criterio de imputación:

	OWN_CAR_AGE	AMT_GOODS_PRICE
Estadísticos	No	Yes
kNN	No	Yes
MICE	Yes	Yes
MiMMi	No	Yes

Detección y tratamiento de outliers

En este apartado se tratará de visualizar aquellas observaciones extremas y, además, discernir sobre si deben ser corregidas o no, dependiendo de la naturaleza de la variable. Para ello, se utilizarán métodos multivariantes, como el análisis de componentes principales (PCA). Así, se procede a representar la proyección de los individuos en los primeros planos factoriales para así observar cuáles se alejan del resto de puntos:





Procedemos a analizar estos individuos, empezando por el que destaca en la dimensión 2. Observamos que, en este caso, la variable que más destaca en este individuo es el número de miembros en su familia: 8. Pese a que este número sea muy elevado, es verosímil pensar que en una vivienda puedan vivir 8 personas, y más si en la base de datos únicamente hay 1 individuo que cumple esta característica. De esta forma, por tanto, este outlier se puede dejar en la base de datos sin sustituir.

Una vez hemos analizado este outlier, podemos pasar a analizar los que son valores extremos por la dimensión 1. Como se puede apreciar, el primer plano factorial viene dado por las variables referidas a cantidad de dinero de nuestra base de datos. Así pues, los outliers presentes son personas con unos ingresos muy altos y que, además, realizaron préstamos por una cantidad de dinero muy superior al que cobran. Así pues, se trata de personas ricas, las cuales existen en nuestra sociedad, de forma que se quedan en la base de datos tal y como aparece. Más adelante, se aplicará alguna transformación que pueda permitir corregir estos valores tan extremos.

Feature engineering

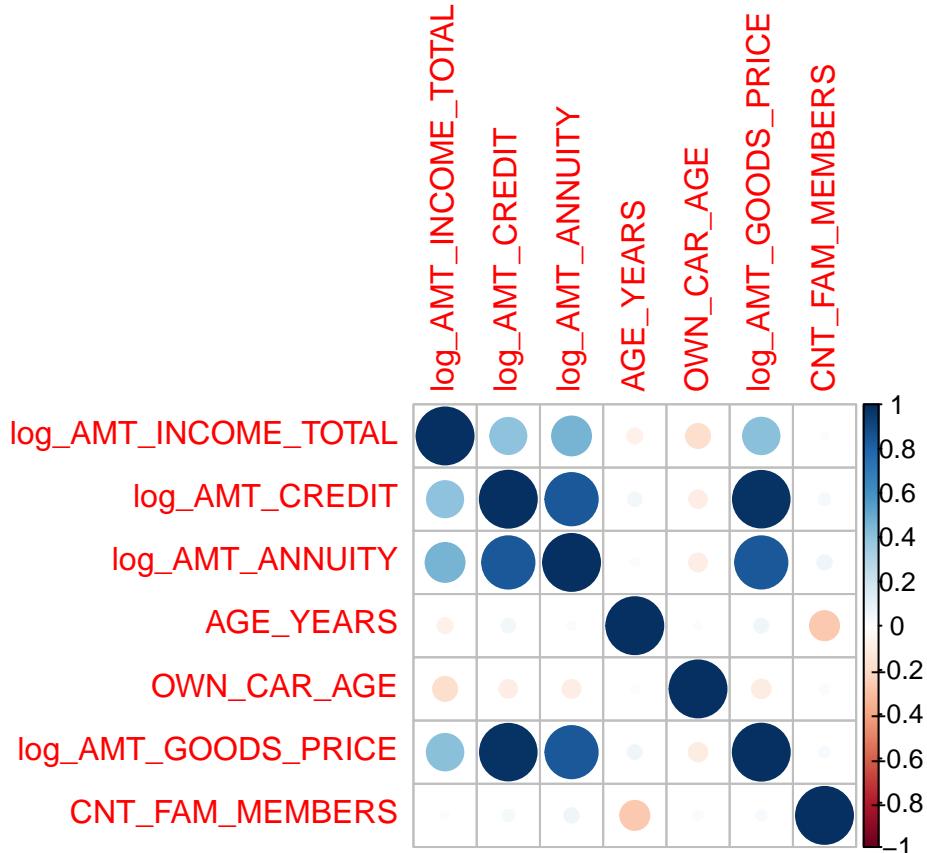
Por último, realizaremos la selección de variables final para nuestra base de datos, así como aplicar transformaciones correctas a nuestras variables para que cumplan algunas hipótesis, como normalidad o heteroscedasticidad. Para este apartado se hace una disección de cada variable una a una.

En primer lugar, se resolverán problemas relacionados con las variables numéricas. Como tenemos variables relacionadas con cantidades monetarias (salario, cantidad prestada...), tal vez sería mejor aplicar una transformación logarítmica:

Así pues, esta transformación debería resolver problemas relacionados con la normalidad de estas variables. Otro cambio a realizar es el respectivo a la variable DAYS_BIRTH, la cual muestra el número de días que lleva vivo el individuo. Sin embargo, el hecho de que esta variable esté en negativo y expresada en días (cuando normalmente se hace en años) hace que su interpretación sea complicada. De esta forma, se harán los cambios permanentes para encontrar la edad de los clientes, guardándola en una variable llamada AGE_YEARS.

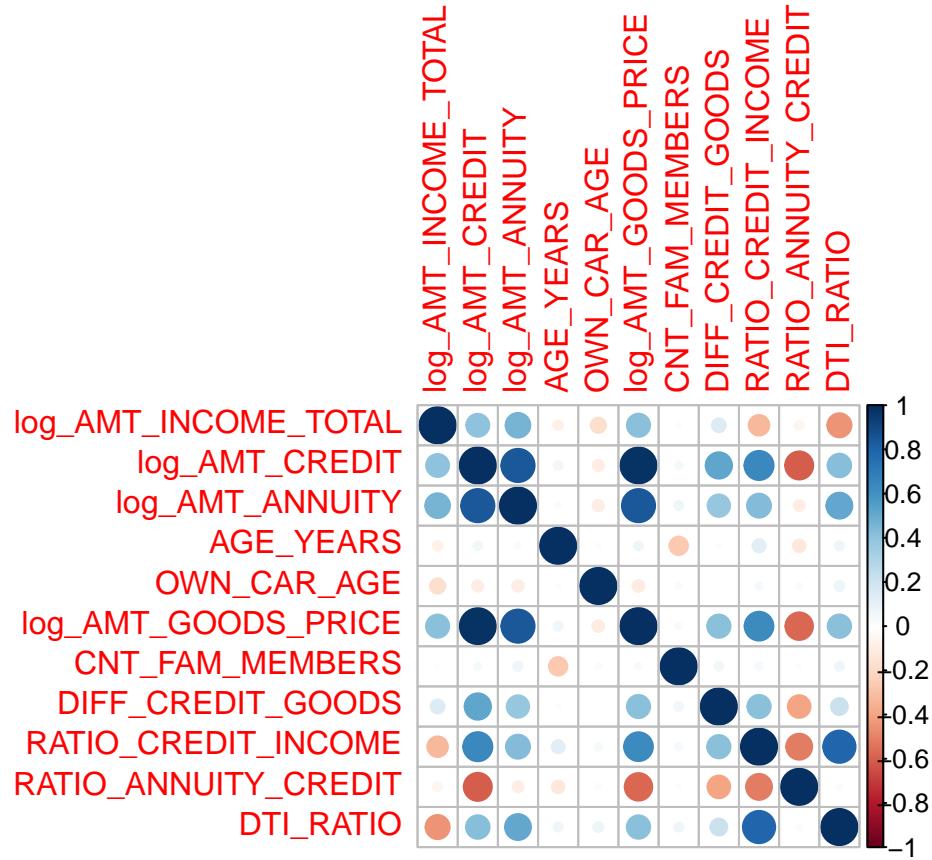
Ahora, vamos a unir aquellas variables ya preprocesadas con el objetivo de tener el dataset preparado para crear nuevas variables.

Antes de avanzar, haremos un correlograma para ver los pares de variables con un mayor coeficiente de correlación de Pearson:



Como se puede apreciar y como era de esperar, hay 3 variables que presentan una gran autocorrelación entre ellas: log_AMT_CREDIT, log_AMT_GOODS_PRICE y log_AMT_ANNUITY. de esta forma, sería ideal nuevas variables a partir de éstas con las cuales se pueda resolver este problema, ya que explican exactamente lo mismo. Para ello, será necesario basarse en la teoría económica y en qué se fijan las entidades de crédito para conceder préstamos. Así, el siguiente objetivo será crear ratios y variables que pretendan controlar y relacionar dinero prestado con capacidad del cliente para retornarlo:

- DIFF_CREDIT_GOODS: Diferencia entre el crédito pedido y el valor del bien para el que se quiere usar
- RATIO_CREDIT_INCOME: Ratio entre el crédito pedido y el salario anual del prestatario. También se puede contar como el número de años que se tarda en devolver el crédito
- RATIO_ANNUITY_CREDIT: Ratio entre la anuidad del préstamo y el crédito total solicitado
- DTI_RATIO: El DTI (Debt-to-income) ratio mide la capacidad del cliente para pagar la anuity de su préstamo en relación con sus ingresos



Se puede apreciar que, ahora, las nuevas variables creadas no presentan tanta correlación entre ellas como anteriormente había. Se puede apreciar, además, que las correlaciones entre las variables donde había problemas siguen teniéndolas y, como se aprecia en el PCA sencillo realizado antes, será necesario descartar alguna variable, ya que explican cosas similares en las mismas dimensiones. Así, en el PCA se deberá realizar el descarte adecuado de variables en función de su aportación al PCA resultante.

Análisis descriptivo post-preprocessing

Análisis Univariante

Con la intención de realizar un buen análisis descriptivo univariante de los datos después al pre-procesamiento se ha decidido integrar conjuntamente gráficos y tablas con resultados numéricos para lograr el mejor entendimiento de estos.

Análisis Univariante Numérico

Cuadro 16: Descripción Univariante Variables Numéricas

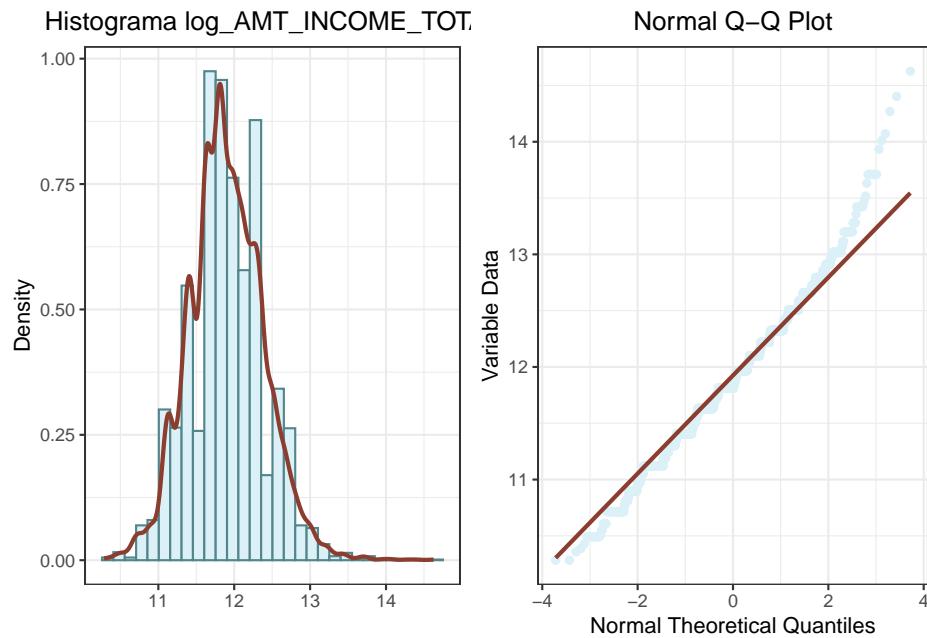
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
log_AMT_INCOME_TOTAL	1	5000	11.90	0.49	11.81	11.89	0.46	10.28	14.63	4.34	0.26	0.76	0.01
log_AMT_CREDIT	2	5000	13.05	0.69	13.13	13.07	0.75	10.71	15.03	4.32	-0.33	-0.20	0.01
log_AMT_ANNUITY	3	5000	10.06	0.54	10.12	10.08	0.52	7.89	12.09	4.20	-0.39	0.21	0.01
log_AMT_GOODS_PRICE	4	5000	12.93	0.69	13.02	12.95	0.76	10.71	15.03	4.32	-0.26	-0.17	0.01
AGE_YEARS	5	5000	42.20	11.85	41.00	41.84	14.83	21.00	68.00	47.00	0.22	-1.00	0.17
DIFF_CREDIT_GOODS	6	5000	63201.31	68968.70	47520.00	52961.87	70453.15	-225000.00	361746.00	586746.00	1.21	1.53	975.36
RATIO_CREDIT_INCOME	7	5000	3.90	2.64	3.20	3.53	2.03	0.12	33.97	33.85	1.89	7.85	0.04
RATIO_ANNUITY_CREDIT	8	5000	0.05	0.02	0.05	0.05	0.02	0.03	0.12	0.09	1.07	0.57	0.00
DTI_RATIO	9	5000	0.18	0.10	0.16	0.17	0.08	0.01	1.35	1.34	1.63	7.98	0.00

Como parte del análisis descriptivo en la fase del post preprocesamiento, se ha generado una tabla que presenta varios estadísticos de las variables numéricas. Estas estadísticas se han calculado después de aplicar las técnicas estadísticas necesarias para procesar adecuadamente los datos.

- Media truncada (Trimmed mean): Al igual que antes del preprocesamiento, la media truncada revela que la variable “Amt credit” tiene una media cercana a la mediana, lo que sugiere una alta simetría en esta variable.
- Asimetría (Skew): Después del procesamiento de datos, se observan cambios en la asimetría de algunas variables. Las variables “Diff_credit_goods,” “Ratio_credit_income,” “Ratio_annuity_credit,” y “DTI_ratio” muestran asimetría positiva, indicando que la mayoría de los valores se concentran a la izquierda de la media y la mediana.
- Curtosis (Kurtosis): Las variables “Ratio_credit_income” y “DTI_ratio” exhiben coeficientes de curtosis significativamente altos, lo que sugiere distribuciones con colas pesadas, es decir, son variables leptocúrticas con colas más puntiagudas que una distribución normal. Por otro lado, la variable “Age_years” tiene un coeficiente de curtosis negativo, lo que la clasifica como una distribución platicúrtica. Las demás variables muestran curtosis cercanas a 3, considerado el valor neutral que indica una distribución normal.
- Error estándar (SE): Todas las variables tienen desviaciones estándar pequeñas en relación a sus medias, excepto la variable “Diff_credit_goods,” lo que podría sugerir una gran diversidad de datos que no siguen una distribución gaussiana.

En la tabla, se aprecia que las variables han experimentado una normalización en el proceso de preprocesamiento. Sin embargo, algunas de las nuevas variables, en su mayoría ratios derivados de variables que ya no están en la base de datos postprocesada, presentan una variedad de distribuciones diferentes.

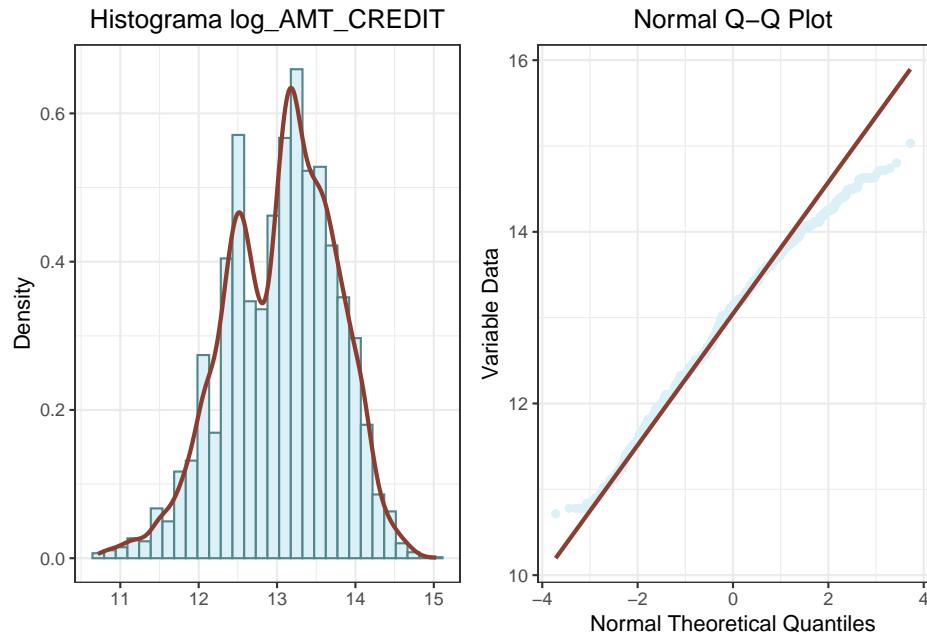
Figura 29: Análisis Gráfico Variable Year Birth



Como se observa en el análisis previo, la variable “Amt_income_total” no presenta una distribución gaussiana. Sin embargo, tras el proceso de eliminación e imputación de valores atípicos (outliers) y datos faltantes (NA), esta variable ha logrado una mayor similitud con una distribución normal en lugar de parecerse a una exponencial.

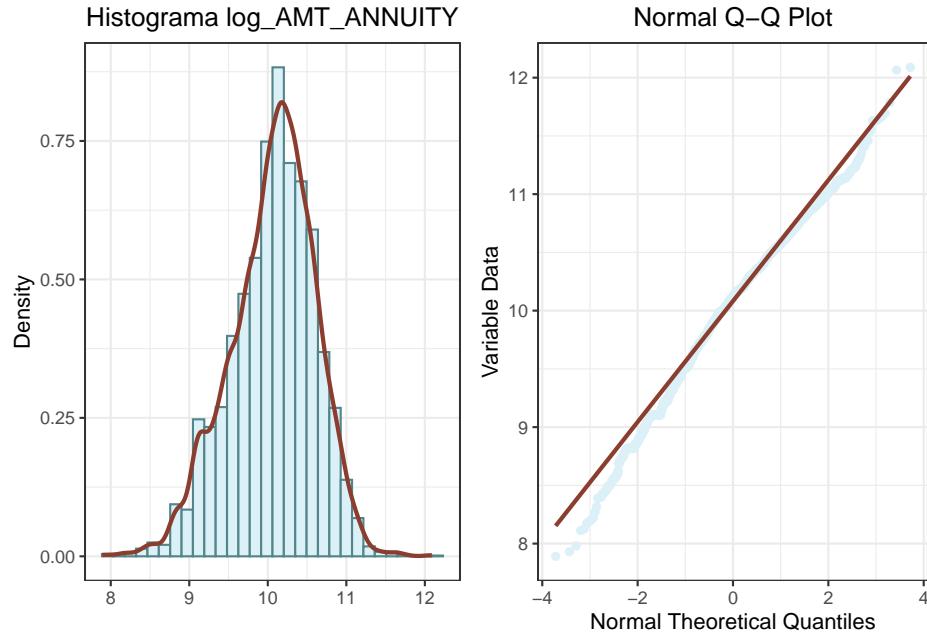
El gráfico Q-Q Plot muestra una notable mejora en la similitud de los cuantiles con los cuantiles teóricos, lo que sugiere una distribución más próxima a la normal. A pesar de este acercamiento visual a la normalidad, los resultados del test de normalidad “Shapiro-Wilk” confirman la hipótesis previa de que los datos no siguen una distribución normal, ya que el p-valor obtenido es ‘r s[[1]]’.

Figura 30: Análisis Gráfico Variable Income



La variable “Log_Amt_credit” presenta una transformación logarítmica realizada con el propósito de lograr una distribución más simétrica y una curtosis más próxima a la normalidad. Sin embargo, como indica el test de Shapiro-Wilk con un valor de ‘r s[[2]]’, esta variable aún no sigue una distribución normal.

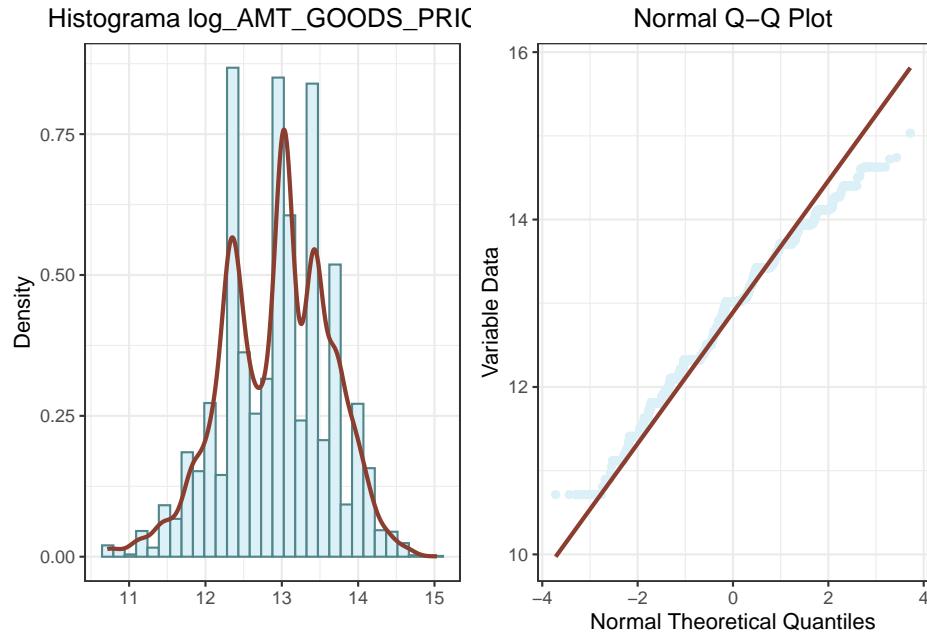
Figura 31: Análisis Gráfico Variable Year Birth



En este caso, se está analizando la variable “Amt_annuity”. A simple vista y según el gráfico Q-Q Plot,

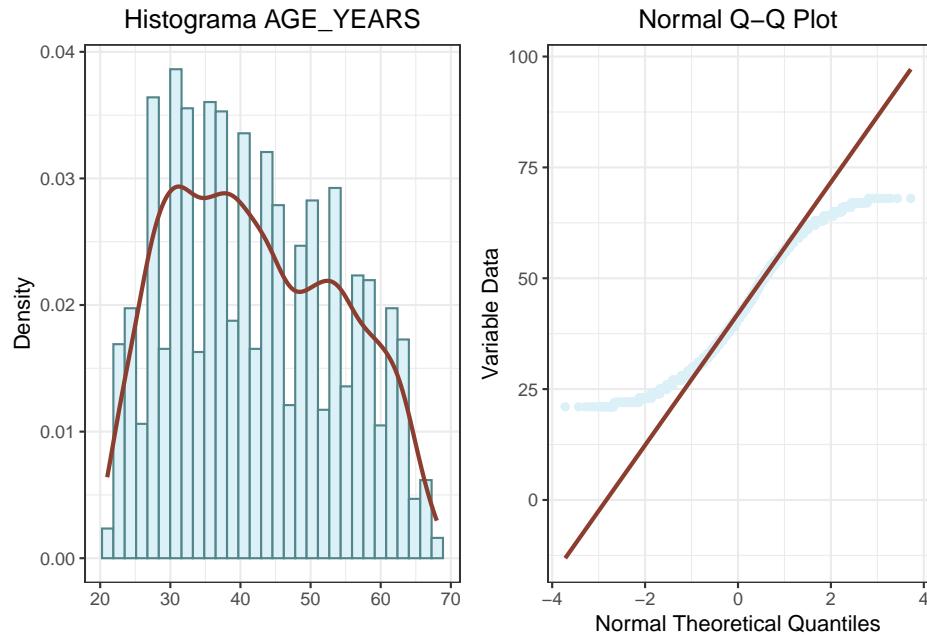
parece que esta variable sigue una distribución normal, en contraste con lo que se observó en el análisis descriptivo previo al procesamiento de datos. Sin embargo, el test de Shapiro-Wilk arroja un valor de ‘r s[[3]]’, indicando que la variable no sigue una distribución normal.

Figura 32: Análisis Gráfico Variable Year Birth



La variable “Amt_goods_price” ha sido transformada logarítmicamente. De igual forma que la variable anterior, los resultados del test de “Shapiro-Wilk” demuestran que esta variable no sigue una distribución gaussiana, teniendo un resultado del test de ‘r s[[4]]’.

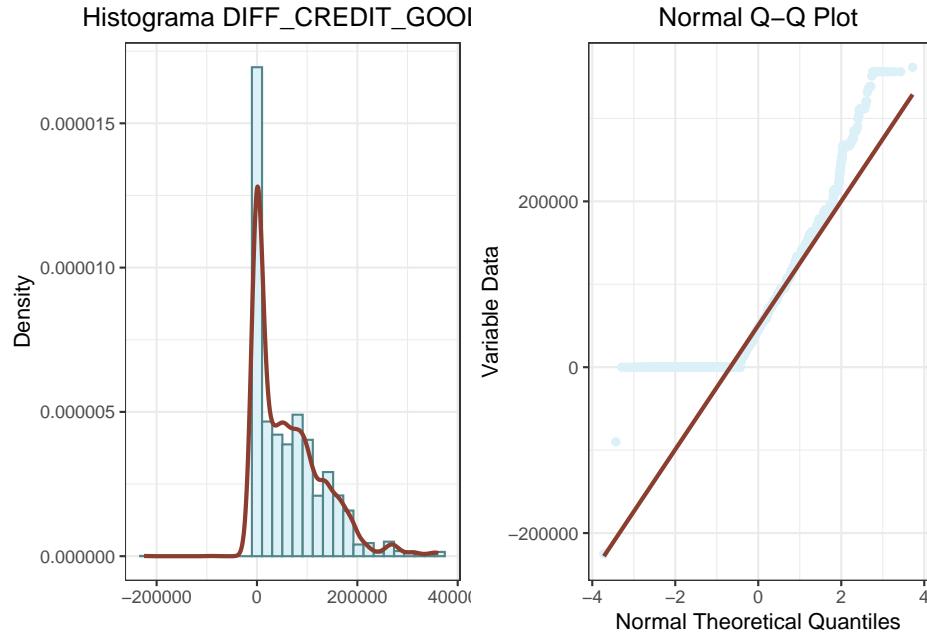
Figura 33: Análisis Gráfico post Variable AGE YEARS



La variable “age_years” no sigue una distribución normal debido a las restricciones naturales inherentes. Esta variable está limitada tanto inferiormente, ya que las personas solo pueden legalmente solicitar un crédito a partir de los 18 años, momento en el que su situación financiera suele ser menos sólida, como superiormente, dado que los créditos suelen ser a medio o largo plazo, lo que implica un crecimiento exponencial del riesgo crediticio relacionado con la edad.

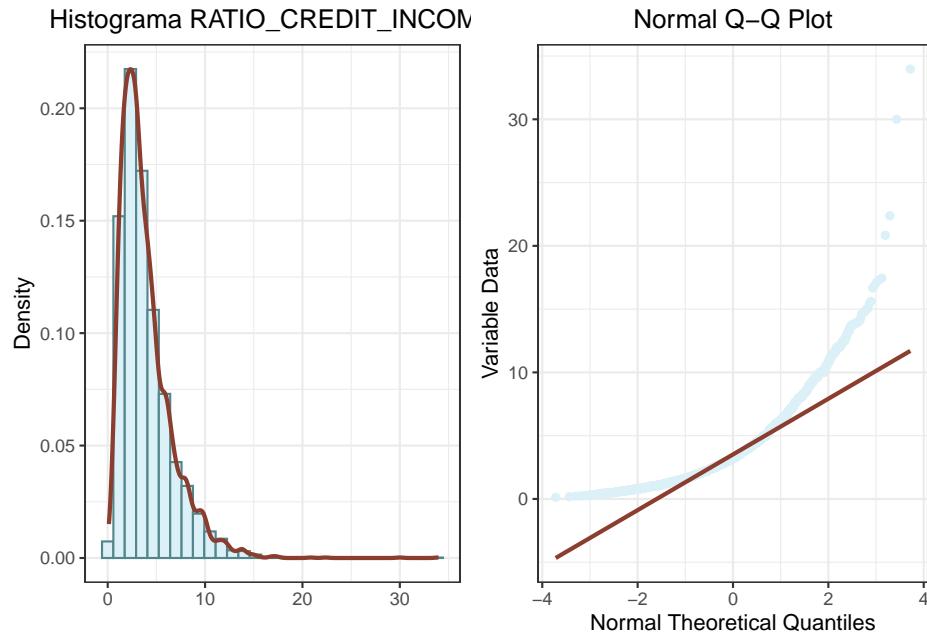
Las limitaciones legales y financieras imponen una clara sesgación en la distribución de edades de los solicitantes de crédito, lo que se refleja en la falta de normalidad en la variable “age_years”. Además, la calidad crediticia y el riesgo crediticio varían significativamente a lo largo de la vida de una persona, lo que también contribuye a la no conformidad con una distribución normal.

Figura 34: Análisis Gráfico Variable DIFF GOODS PRICE



La variable “Diff_credit_goods” presenta un valor mínimo de 0, dado que la diferencia mínima entre el monto del crédito obtenido y el valor del activo que se desea adquirir siempre es positiva. Por lo tanto, esta variable tiende a asemejarse más a una distribución exponencial que a una distribución normal. En este contexto, realizar un análisis gaussiano de esta variable resulta redundante debido a la naturaleza de los datos.

Figura 35: Análisis Gráfico Variable RATIO CREDIT INCOME



De manera similar a la variable anterior, el ratio entre el crédito concedido y el ingreso presenta una limitación en su valor mínimo de 0. Por lo tanto, no parece necesario llevar a cabo un análisis de normalidad de esta variable. La naturaleza de la variable, con un límite inferior en 0, hace que la asunción de normalidad sea poco relevante.

Análisis Univariante Categórico

Tras haber completado el análisis univariante numérico se procede a hacer el análisis categórico.

En la siguiente tabla se presenta un resumen general sobre ellas:

Cuadro 17: Summary descriptives table

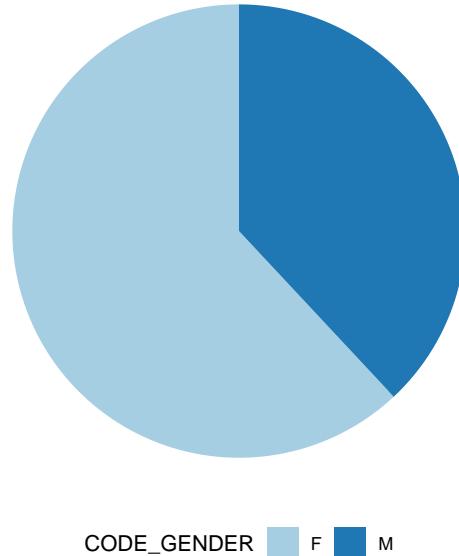
	[ALL]	N
	N=5000	
CODE_GENDER:		
F	3098 (62.0 %)	5000
M	1902 (38.0 %)	
NAME_INCOME_TYPE:		
Businessman	1 (0.02 %)	5000
Commercial associate	1111 (22.2 %)	
Pensioner	763 (15.3 %)	
State servant	306 (6.12 %)	
Working	2819 (56.4 %)	
NAME_EDUCATION_TYPE:		
Academic degree	3 (0.06 %)	5000
Higher education	1018 (20.4 %)	
Incomplete higher	156 (3.12 %)	
Lower secondary	77 (1.54 %)	
Secondary / secondary special	3746 (74.9 %)	
NAME_FAMILY_STATUS:		
Civil marriage	546 (10.9 %)	5000
Married	3095 (61.9 %)	
Separated	320 (6.40 %)	
Single / not married	798 (16.0 %)	
Widow	241 (4.82 %)	
OCCUPATION_TYPE:		
High skill laborers	660 (13.2 %)	5000
Low-mid skill laborers	1047 (20.9 %)	
Low skill laborers	722 (14.4 %)	
Mid-high skill laborers	432 (8.64 %)	
Mid skill laborers	709 (14.2 %)	
Unknown	1430 (28.6 %)	
REGION_RATING_CLIENT:		
1	434 (8.68 %)	5000
2	3641 (72.8 %)	
3	925 (18.5 %)	
TARGET:		
0	2865 (57.3 %)	5000
1	2135 (42.7 %)	

Por lo tanto, en la tabla se presentan tanto la frecuencia absoluta como la frecuencia relativa de cada valor posible en cada variable categórica, ya sean dicotómicas o politómicas. Esto facilita la identificación de la moda de manera sencilla.

Una vez se ha realizado un resumen general, se ha procedido a analizar cada variable una a una:

Figura 36: Pie Chart post Variable CODE GENDER

CODE_GENDER



Tal y como se aprecia en este gráfico pastel, la estructura de los datos en cuanto a la distribución del sexo no se ve modificada por el preprocessing.

Figura 37: Bar Chart post Variable NAME INCOME TYPE

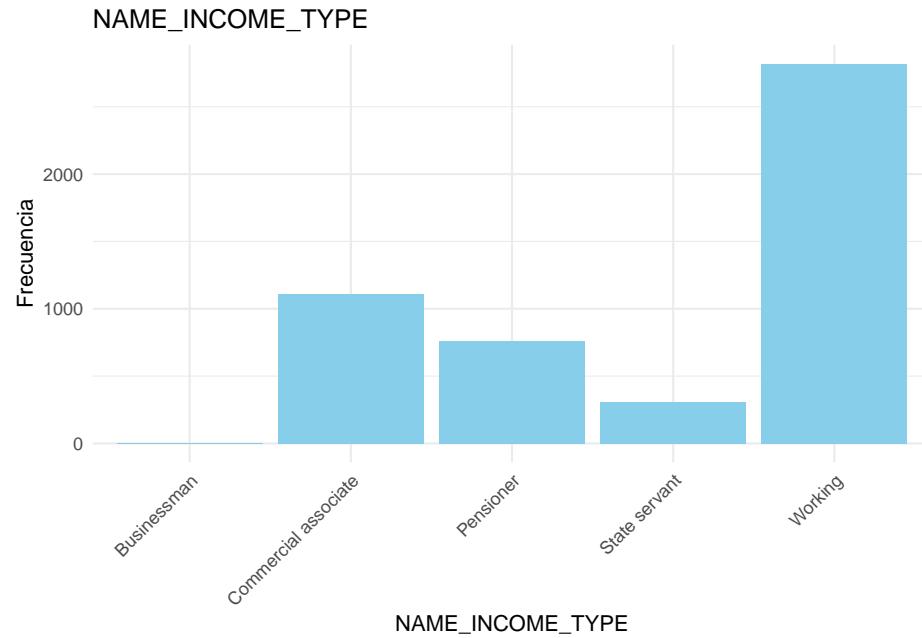


Figura 38: Bar Chart post Variable NAME EDUCATION TYPE

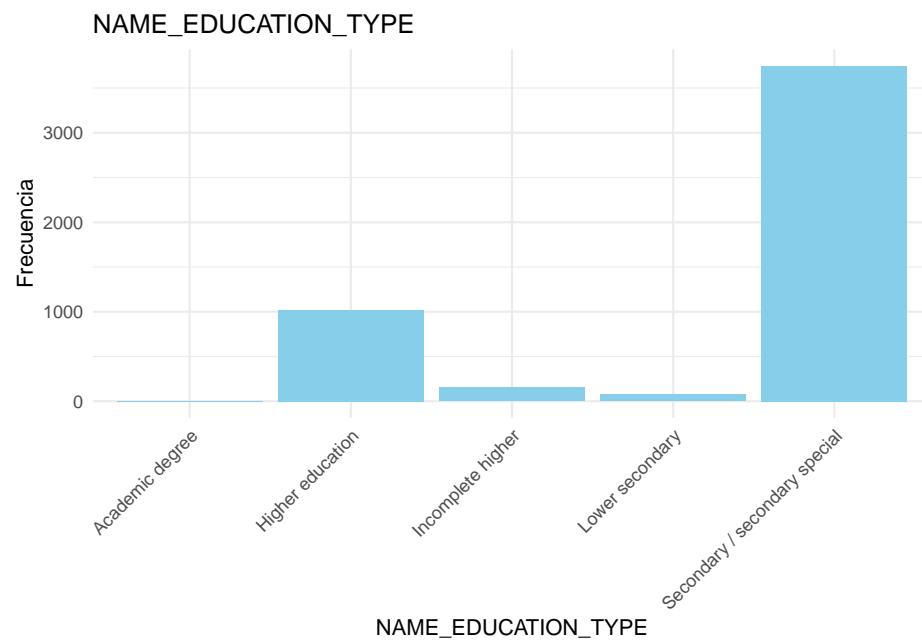
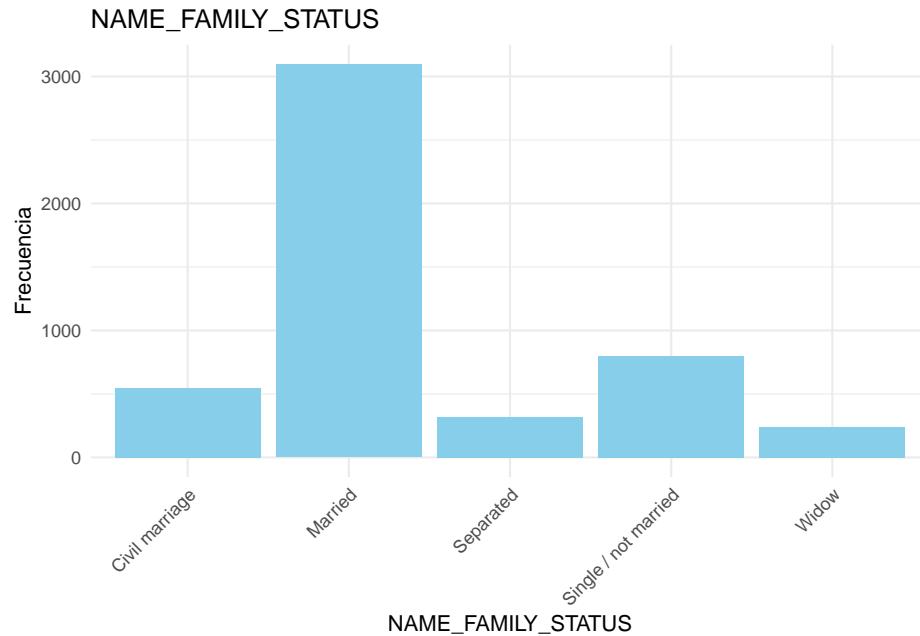
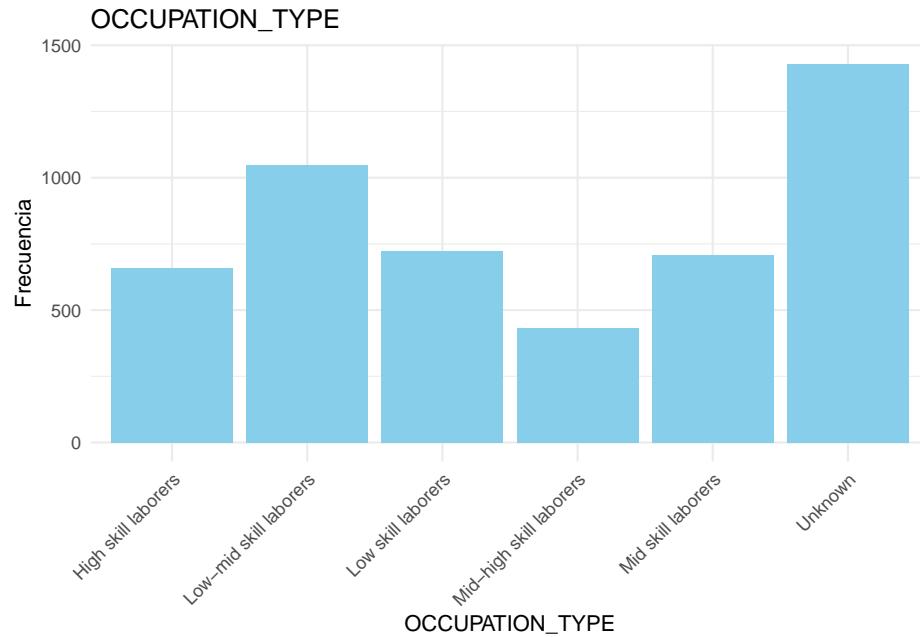


Figura 39: Bar Chart post Variable NAME FAMILY STATUS



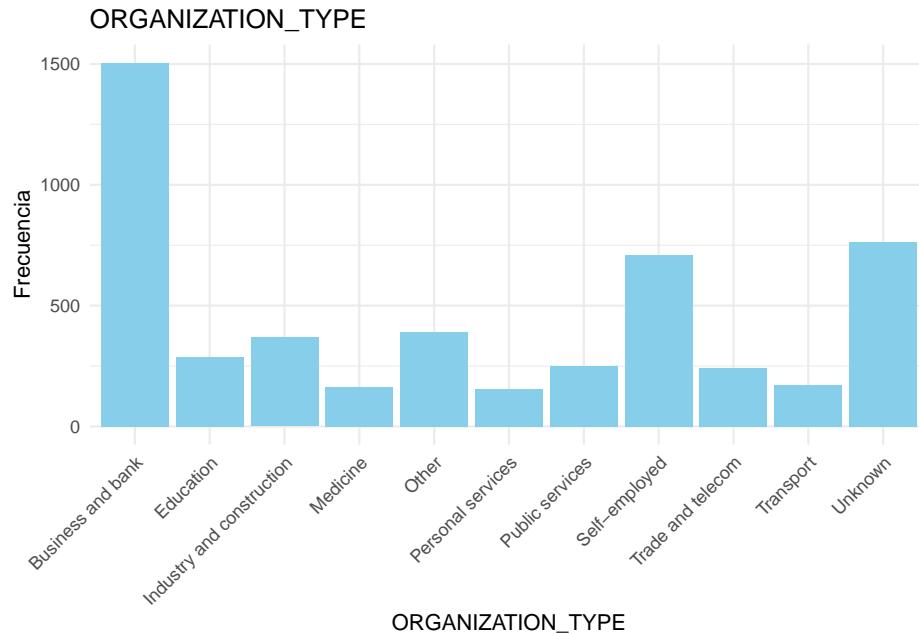
Del mismo modo, las variables NAME_INCOME_TYPE, NAME_EDUCATION_TYPE y NAME_EDUCATION_TYPE mantienen su estructura y patrones previos al preprocesamiento. Esto sugiere que, o bien había pocos valores atípicos o datos faltantes (NA), o que la imputación de datos se realizó de manera precisa. En consecuencia, la estructura se mantiene constante tanto antes como después del procesamiento.

Figura 40: Bar Chart post Variable OCCUPATION TYPE



En la variable `OCCUPATION_TYPE` se aprecia como se han reducido el número de categorias usando como criterio de agrupación el nivel de habilidades técnicas y nivel de responsabilidad de los distintos trabajos. Así, por ejemplo “High skill tech staff”, “Managers” y “Medicine staff” han sido consideradas “High skill laborers” debido a la gran responsabilidad y conocimiento requerido para desarrollar las tareas requeridas del trabajo.

Figura 41: Bar Chart post Variable ORGANIZATION TYPE

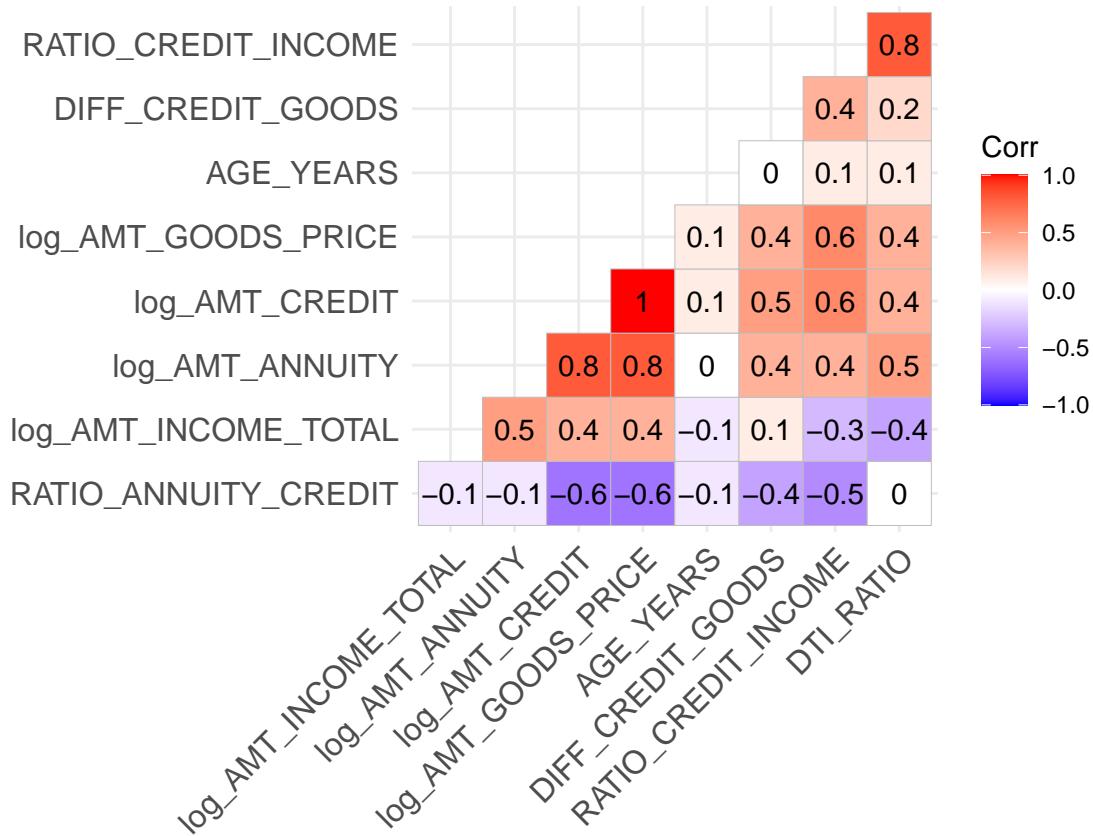


En la variable “Organization type,” se observan cambios en la distribución de los datos debido a la reorganización y a la imputación de valores atípicos y datos faltantes. Previo a la reagrupación, existian diferentes categóricas que podían hacer referencia a un mismo sector o grupo, por lo que al agruparlos aparece la categoría “Busiess and Bank” como la mas representativa. Además, ha surgido la categoría “unknown,” que incluye a los casos que no se han podido clasificar en ninguna de las otras categorías.

Análisis Bivariante Numérico

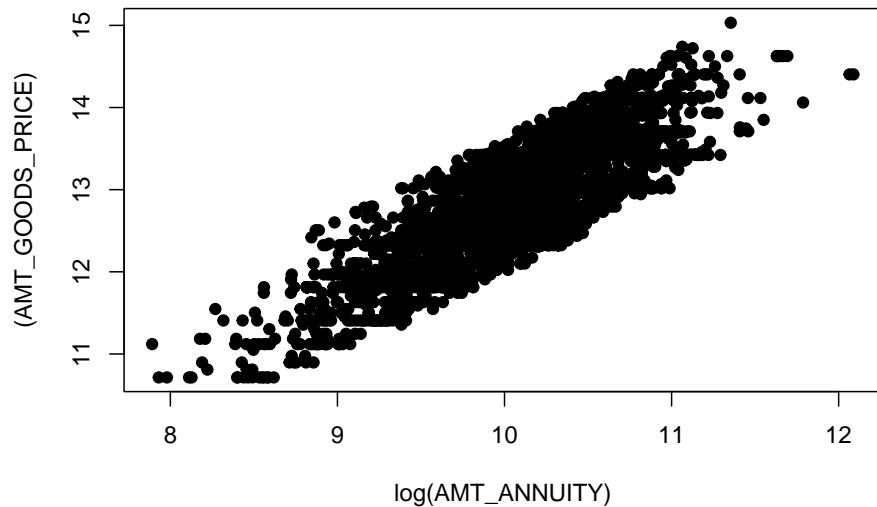
Con el propósito de identificar las relaciones más significativas entre las variables numéricas, se ha creado un gráfico de correlación utilizando la técnica de HeatMap. En este gráfico, los colores indican el grado de dependencia entre las variables numéricas. Cuanto más intenso sea el color, mayor será la relación, y se prestará una mayor atención a estas relaciones en nuestro análisis.

Figura 42: Matriz de Correlaciones post para las Variables Numéricas



Después de analizar el gráfico y considerar los cambios realizados en el procesamiento de los datos, se observa que las correlaciones entre las variables antes y después del procesamiento se mantienen constantes. Se destaca especialmente la relación entre la variable `AMT_CREDIT` y `AMT_GOODS_PRICE`, que se mantiene en 1, lo que indica que el valor del crédito otorgado es igual al valor del precio del bien. Además, se observa una correlación entre `AMT_ANNUITY` y `AMT_CREDIT`, así como entre `AMT_ANNUITY` y `AMT_GOODS_PRICE`. También se nota una correlación entre `RATIO_CREDIT_INCOME` y la variable `DTI_RATIO`. La alta correlación entre el ratio DTI y el ratio credit income se debe a que la primera variable representa la cantidad de deuda que se paga en cada período, es decir, la cuota mensual, mientras que el ratio credit income es la relación entre la cuota y el salario del prestatario.

Figura 43: Gráfico de dispersión Gasto Total en Pescado vs Gasto Total en Fruta

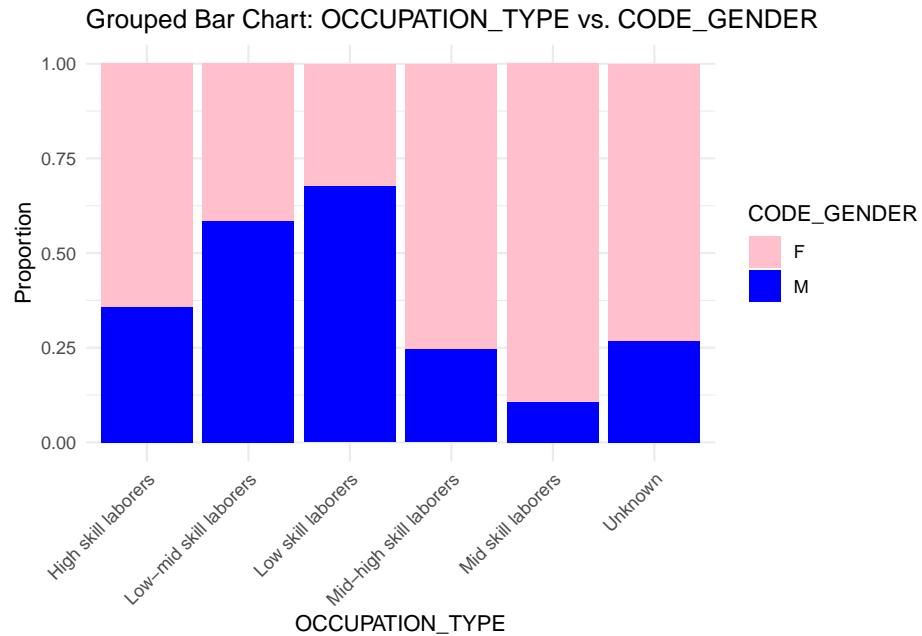


De manera similar, la estructura de los datos se ha mantenido constante, es decir, la correlación entre el valor de los bienes y la anualidad también es relativamente alta. Es importante señalar que los clientes que posean una relación entre la anualidad y el valor del bien que compren (teniendo en cuenta que el precio del bien es igual al valor del préstamo) serán aquellos que deban destinar una proporción menor de sus ingresos al reembolso de la deuda. Tal y como se aprecia en el gráfico, todos los datos están en una franja diagonal.

Análisis Bivariante Categórico-Descriptivo

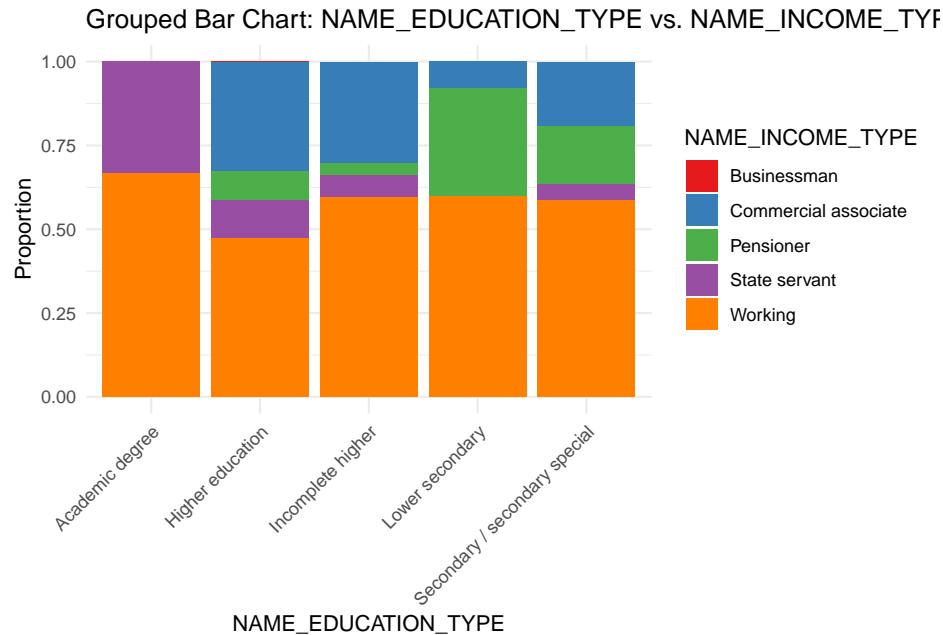
Para concluir el análisis descriptivo antes de proceder al procesamiento de los datos, es necesario examinar la relación entre las variables categóricas y las numéricas. Para este propósito, utilizaremos la creación de varios boxplots, lo que nos permitirá presentar nuestras conclusiones de manera precisa y concisa.

Figura 44: Gráfico comparación CODE GENDER vs OCCUPATION TYPE



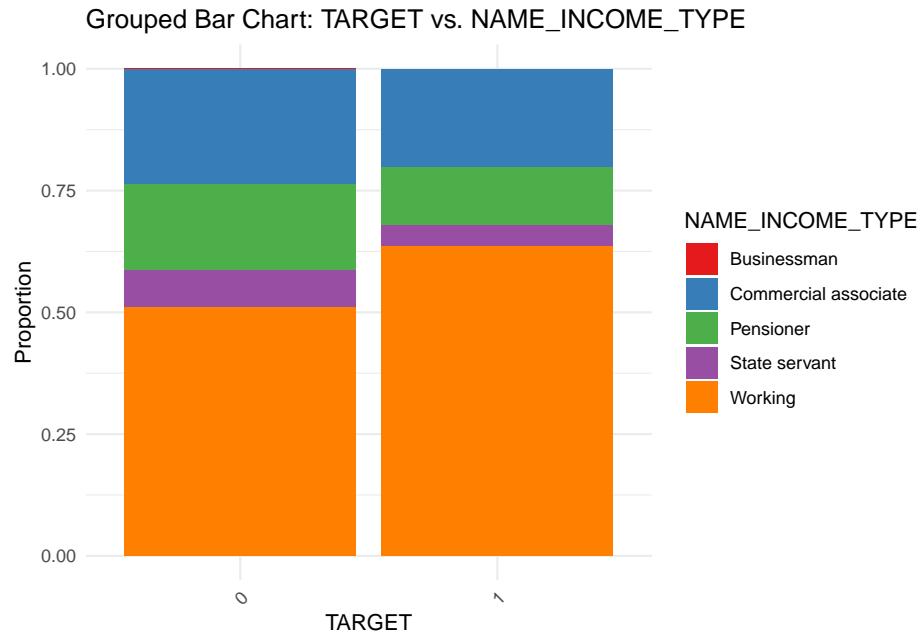
Debido al agrupamiento de las categorías en los distintos niveles de habilidades y responsabilidad, se aprecia como las mujeres tienden a tener puestos de trabajo mas demandantes en cuanto a estos dos atributos respecto a los hombres, completando casi con totalidad los trabajos de “Mid skill laborers y teniendo un peso altamente representativo en”Mid-high skill laborers” y “High skill laborers”.

Figura 45: Gráfico comparación NAME EDUCATION TYPE vs NAME INCOME TYPE



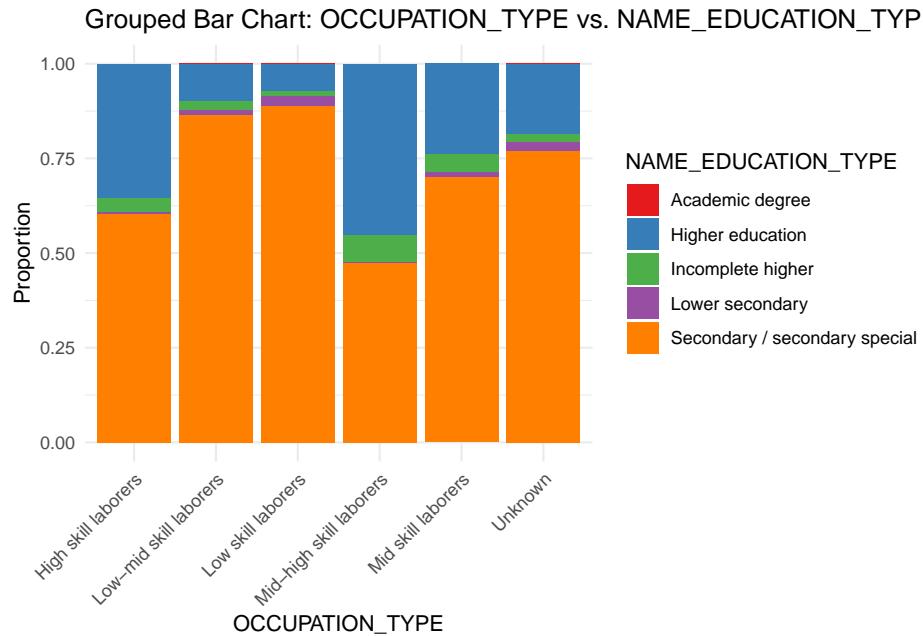
En este gráfico se analiza la relación entre el nivel de educación y el tipo de ingreso. Como se observa, la mayoría de los trabajadores en empleos del sector privado convencional presentan una diversidad de niveles educativos, mientras que aquellos con estudios académicos tienden a trabajar para el sector público. Vale la pena señalar que un porcentaje significativo de los cónyuges sobrevivientes tiene únicamente educación secundaria. Esto podría deberse al hecho de que estos trabajadores son de mayor edad y, en su momento, las oportunidades de acceder a educación superior eran limitadas.

Figura 46: Gráfico comparación TARGET vs NAME INCOME TYPE



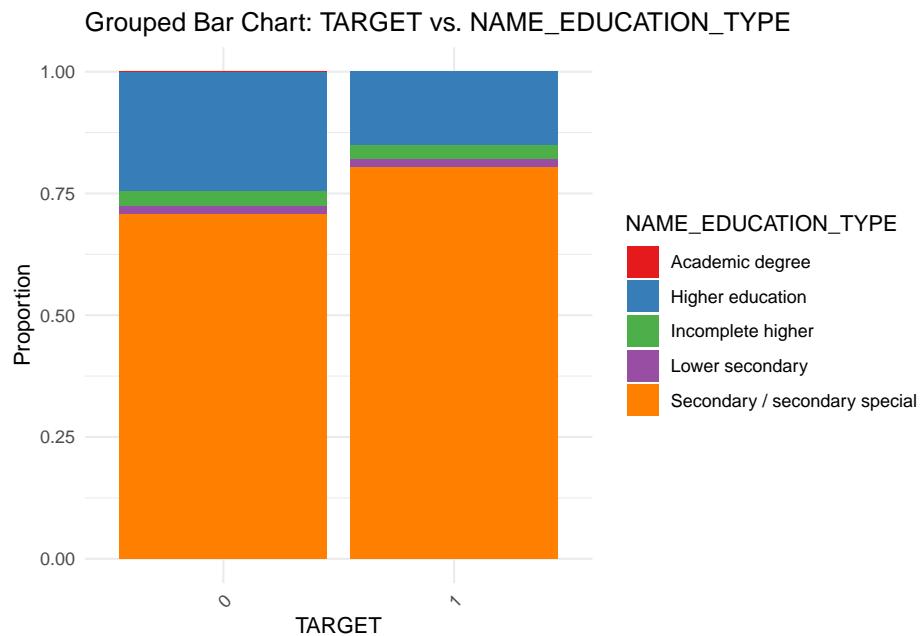
En lo que respecta a la variable TARGET, se observa una disparidad en la capacidad de pago de los clientes en el sector privado, siendo los pensionistas y los comerciales quienes presentan proporcionalmente menos dificultades.

Figura 47: Gráfico comparación post OCCUPATION_TYPE vs NAME EDUCATION TYPE



En este gráfico se confirma la idea de que los trabajadores con niveles educativos más altos tienden a ocupar puestos de trabajo que requieren un mayor nivel de conocimientos técnicos, mientras que aquellos con niveles educativos más bajos suelen desempeñar empleos que demandan menos destrezas técnicas.

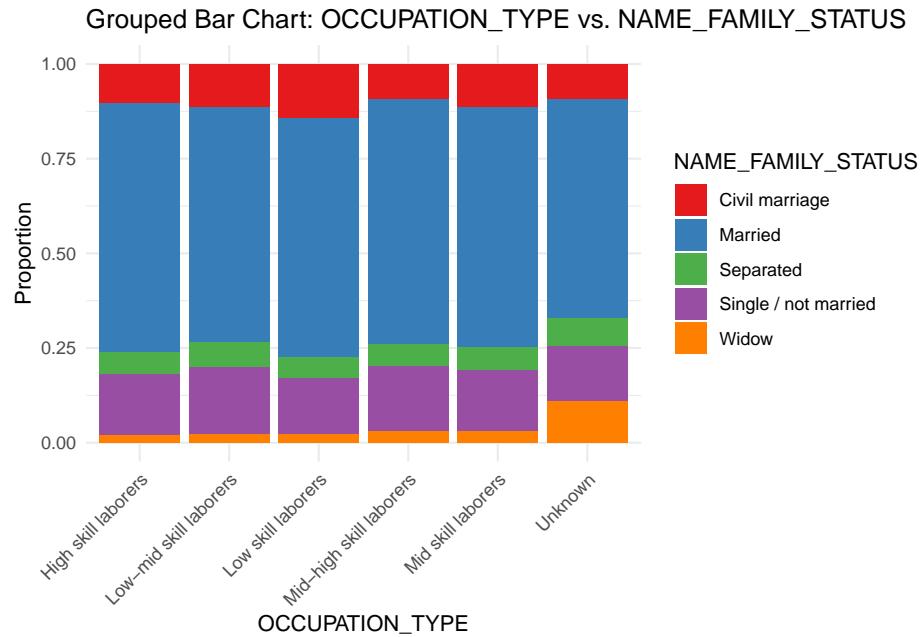
Figura 48: Gráfico comparación post TARGET vs NAME EDUCATION TYPE



En lo que respecta al nivel de educación, es notable que aquellos trabajadores con un nivel educativo más

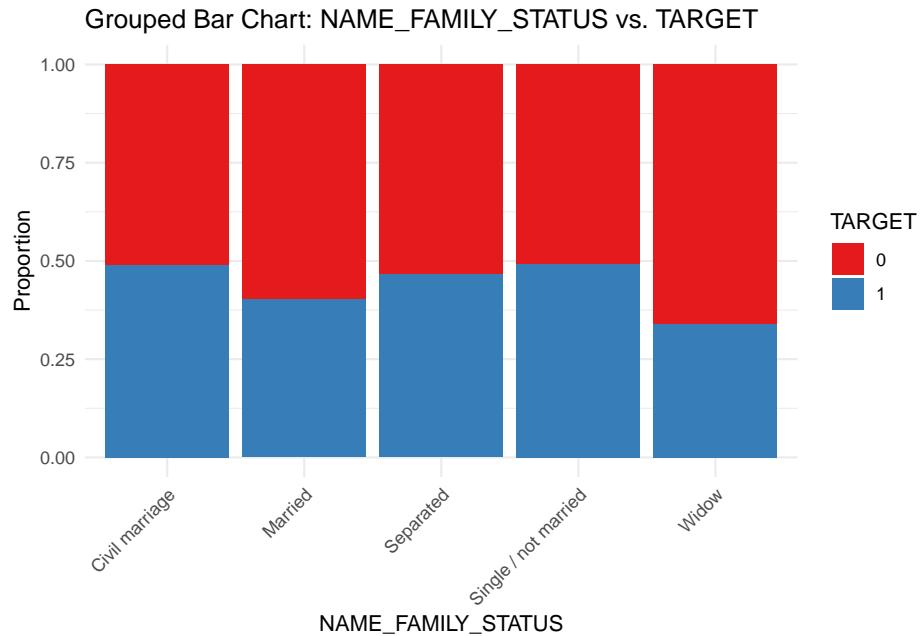
bajo son quienes enfrentan mayores dificultades para cumplir con sus pagos de manera consistente.

Figura 49: Gráfico comparación NAME FAMILY STATUS vs NAME OCCUPATION TYPE



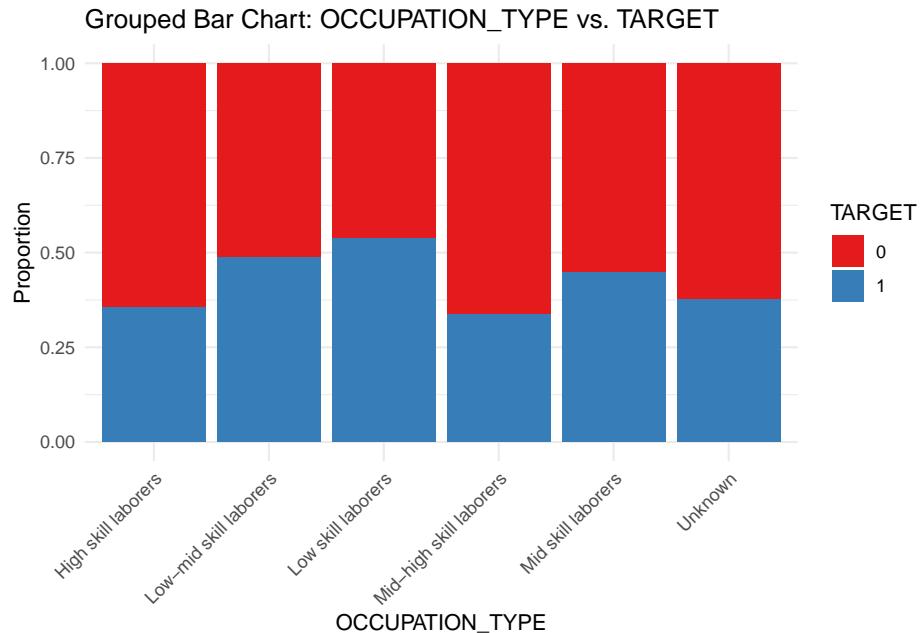
En este gráfico se analiza la relación entre la ocupación de los individuos y el estado civil de ellos mismos. Como se observa, la mayoría de los trabajadores de cualquier sector están casados, muchos por la iglesia y unos pocos civilmente. Vale la pena señalar que un porcentaje significativo de los cónyuges sobrevivientes trabajan en recursos humanos. Esto podría deberse al hecho de que estos trabajadores son de mayor edad y, en su momento, las oportunidades de acceder este tipo de empleos eran más altas.

Figura 50: Gráfico comparación NAME FAMILY STATUS vs TARGET



En este gráfico se confirma la idea de que dentro de los diferentes grupos de estados civiles no hay mucha diferencia con respecto al target. La única diferencia notable es que aquellas personas que han quedado viudas son quienes enfrentan menores dificultades para cumplir con sus pagos de manera consistente.

Figura 51: Gráfico comparación OCCUPATION_TYPE vs TARGET



En el segundo gráfico nos muestra los oficios de las personas con respecto al target. Se aprecia claramente

como la gran mayoría de trabajadores poco calificados tienen mas dificultades de pago respecto a los demás, mientras sorprende que los secretarios sean los que menos dificultades tengan en pagar proporcionalmente. Nos podemos basar hasta un cierto punto en este tipo de análisis ya que podria haber pocas personas dentro de un mismo grupo de trabajadores y muchas personas dentro de otro y esto dificultaria sacar conclusiones claras.

Análisis de componentes principales (ACP)

Cuadro 18: Clase de cada variable

CODE_GENDER	factor
NAME_INCOME_TYPE	factor
NAME_EDUCATION_TYPE	factor
NAME_FAMILY_STATUS	factor
OCCUPATION_TYPE	factor
ORGANIZATION_TYPE	factor
REGION_RATING_CLIENT	factor
TARGET	factor
AMT_INCOME_TOTAL	numeric
AMT_CREDIT	numeric
AMT_ANNUITY	numeric
DAYS_BIRTH	numeric
OWN_CAR_AGE	numeric
AMT_GOODS_PRICE	numeric
CNT_FAM_MEMBERS	numeric
log_AMT_INCOME_TOTAL	numeric
log_AMT_CREDIT	numeric
log_AMT_ANNUITY	numeric
log_AMT_GOODS_PRICE	numeric
AGE_YEARS	numeric
DIFF_CREDIT_GOODS	numeric
RATIO_CREDIT_INCOME	numeric
RATIO_ANNUITY_CREDIT	numeric
DTI_RATIO	numeric

Se observa que la base de datos tiene un total de 11 columnas numéricas. Por tanto, el análisis de componentes principales tendrá como máximo 11 componentes.

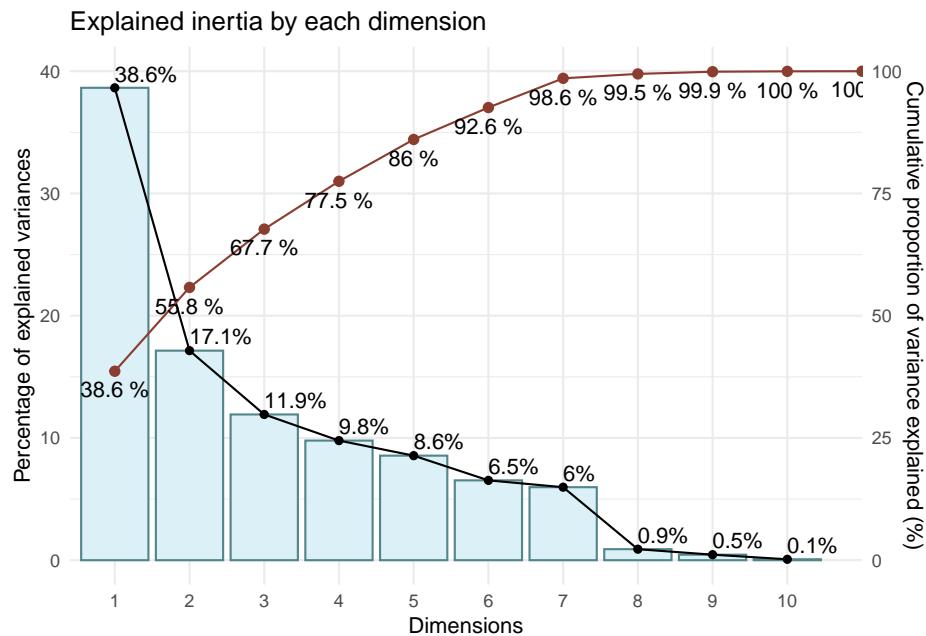
Selección de variables numéricas

Se proceden a eliminar, primeramente, aquellas variables para las cuales ya existe su transformación logarítmica. Esto se hace para no contar con variables que contengan la misma capacidad explicativa (y así evitar colinealidad). También se elimina la variable DAYS_BIRTH, ya que se cuenta con AGE_YEARS, que es una transformación de la inicial, debido a que DAYS_BIRTH no tenía una clara interpretación.

PCA

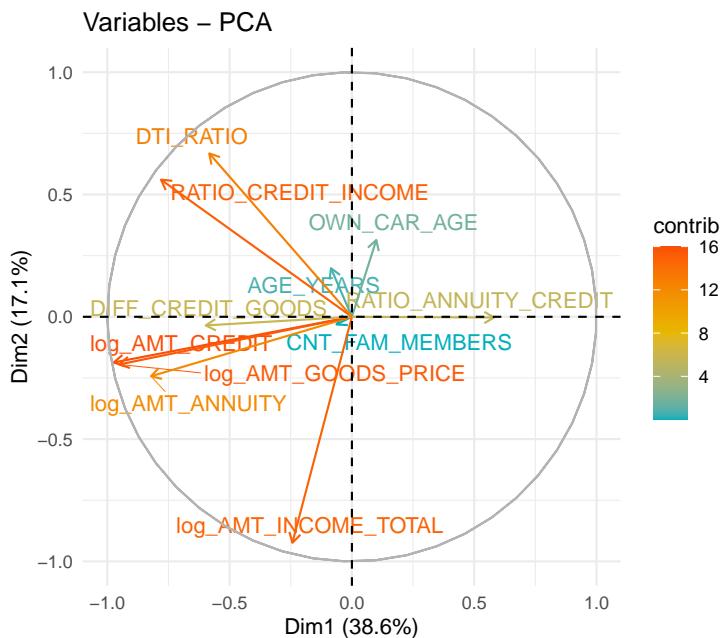
A partir de aquí, se procede con el análisis de componentes principales.

Figura 52: Porcentaje de inercia explicado por dimensión



Teniendo en cuenta que la inercia equivale a la proporción de la variabilidad de los datos, se sabe que con un 80 % de inercia se puede obtener casi toda la información o variabilidad de la base de datos original. Con ello, vemos que el 80 % de la inercia acumulada se logra con 5 planos factoriales, pero aún se pueden eliminar algunas variables.

Figura 53: Proyección de variables en los dos primeros planos factoriales



Observamos la tabla de rotaciones:

Cuadro 19: Correlación de cada variable con cada plano factorial

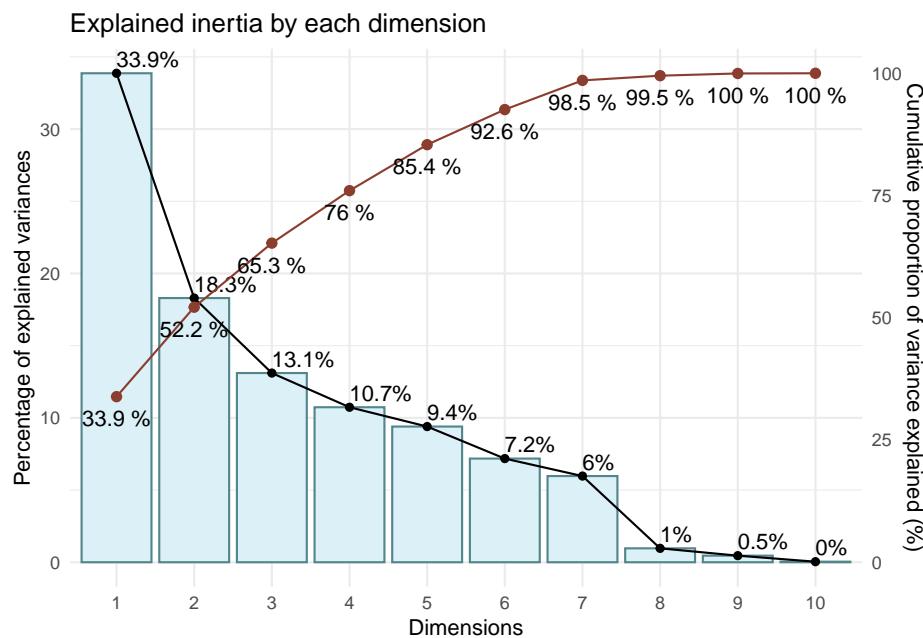
	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0483554	0.2287002	0.0530876	0.2413126	0.9271662
CNT_FAM_MEMBERS	-0.0305664	-0.0226378	0.6166566	0.3768438	-0.0869280
log_AMT_INCOME_TOTAL	-0.1180318	-0.6727750	-0.0175937	-0.0797217	0.1936175
log_AMT_CREDIT	-0.4721472	-0.1358924	-0.0277607	0.0061186	0.0394832
log_AMT_ANNUITY	-0.3983901	-0.1769402	0.2080437	-0.3959441	0.1684720
log_AMT_GOODS_PRICE	-0.4612147	-0.1418139	-0.0257791	-0.0460653	0.0317867
AGE_YEARS	-0.0415368	0.1447106	-0.6281559	-0.2150941	0.1247852
DIFF_CREDIT_GOODS	-0.2897238	-0.0256553	-0.0363394	0.3324768	0.0667128
RATIO_CREDIT_INCOME	-0.3784918	0.4081350	-0.0056133	0.0574366	-0.1162079
RATIO_ANNUITY_CREDIT	0.2819794	-0.0021710	0.3500729	-0.6037860	0.1736897
DTI_RATIO	-0.2828528	0.4863060	0.2310985	-0.3313722	-0.0259258

En el grafico vemos que las flechas de **log_AMT_GOODS_PRICE** y **log_AMT_CREDIT** se solapan entre ellas, eso quiere decir que las dos variables explican el mismo plano factorial. Vemos en la tabla de rotaciones que **log_AMT_CREDIT** contribuye más a explicar el primer plano factorial, y además las correlaciones entre cada una de las variables y cada dimensión son muy similares. Por esta razón eliminamos **log_AMT_GOODS_PRICE**.

Nos quedamos con una variable menos, por tanto tenemos 10 variables numéricas.

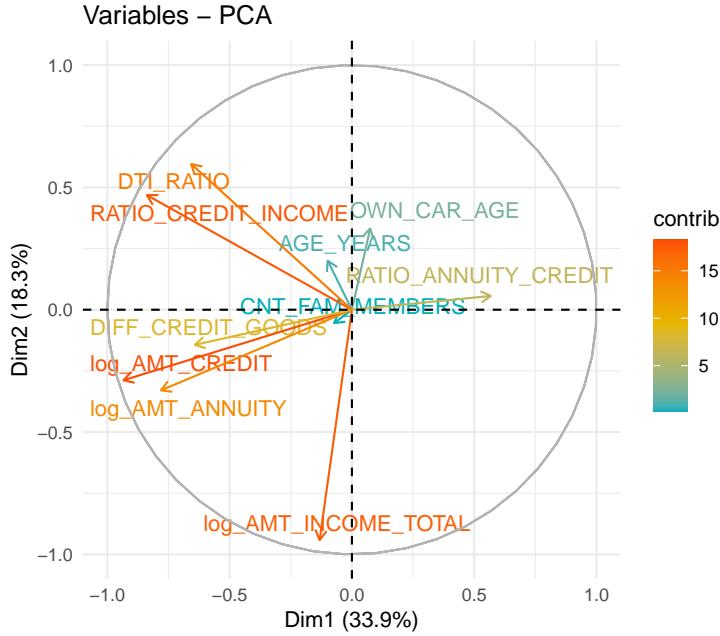
De vuelta, verificamos el porcentaje de inercia por cada componente principal y la acumulada:

Figura 54: Porcentaje de inercia explicado por dimensión



Como se puede ver, seguimos teniendo 5 dimensiones que acumulan el 80 % de la varianza.

Figura 55: Proyección de variables en los dos primeros planos factoriales



Vemos que las variables **CNT_FAM_MEMBERS**, **AGE_YEARS** y **OWN CAR AGE** no explican las dos primeras componentes pero si nos fijamos en la tabla de rotaciones vemos que sí tienen importancia a la hora de explicar las otras tres dimensiones:

Cuadro 20: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0406536	0.2454602	0.0602964	0.2636258	0.9169085
CNT_FAM_MEMBERS	-0.0394947	-0.0404725	0.6131245	0.3812205	-0.0952529
log_AMT_INCOME_TOTAL	-0.0717491	-0.6961680	-0.0308119	-0.0950202	0.2049254
log_AMT_CREDIT	-0.5073525	-0.2130431	-0.0401256	-0.0102255	0.0487471
log_AMT_ANNUITY	-0.4240602	-0.2436766	0.1986781	-0.4109116	0.1850255
AGE_YEARS	-0.0537890	0.1481145	-0.6256961	-0.2195745	0.1310781
DIFF_CREDIT_GOODS	-0.3487325	-0.1061162	-0.0544017	0.3010831	0.0894349
RATIO_CREDIT_INCOME	-0.4552798	0.3462113	-0.0081212	0.0515503	-0.1143317
RATIO_ANNUITY_CREDIT	0.3084733	0.0416441	0.3591344	-0.5960684	0.1815479
DTI_RATIO	-0.3570194	0.4403417	0.2343324	-0.3322223	-0.0197281

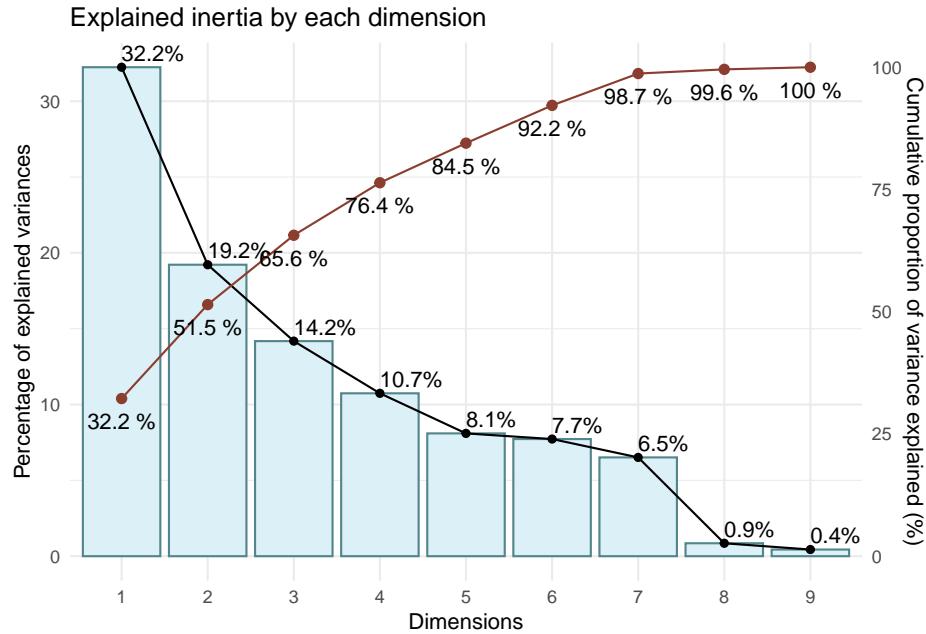
Por ejemplo, en el caso de **OWN CAR AGE** se puede ver en la tabla anterior que, se podría decir que no es la que mejor explica las primeras componentes, pero vemos que explica casi toda la componente 5.

Otra observación se podría hacer de las variables **log_AMT_CREDIT** y **log_AMT_ANNUITY**, donde se puede apreciar que tienen correlaciones similares con la primera y segunda dimensión. Teniendo en cuenta que esas dos primeras dimensiones (PC1 y PC2) son las más importantes, ya que acumulan la mayoría de la inercia (en total un 52.2 %), parece una decisión sensata eliminar una de ellas, en este caso **log_AMT_ANNUITY**.

Ahora conservamos 9 variables numéricas.

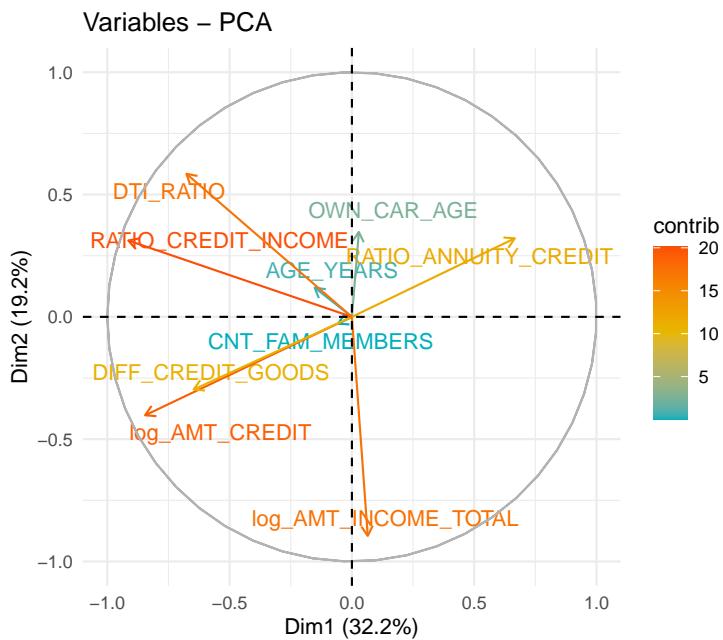
De forma igual que anteriormente, comprobamos el porcentaje de inercia para cada componente principal y la acumulada:

Figura 56: Porcentaje de inercia explicado por dimensión



Como se puede comprobar, las 5 dimensiones siguen siendo las necesarias para acumular el 80 % de la varianza.

Figura 57: Proyección de variables en los dos primeros planos factoriales



Observamos tambien la tabla de rotaciones para verificar si se puede eliminar alguna variable más:

Cuadro 21: Correlación de cada variable con cada plano factorial

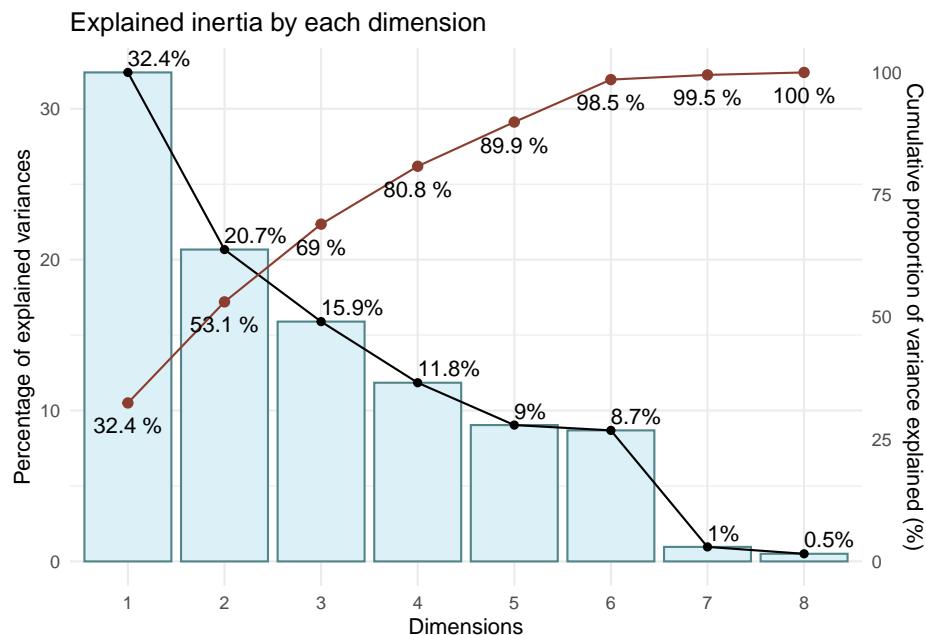
	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0167177	0.2638115	0.0414438	-0.9282134	-0.1613777
CNT_FAM_MEMBERS	-0.0330431	-0.0213960	0.7023098	-0.0655018	0.6542655
log_AMT_INCOME_TOTAL	0.0380475	-0.6808910	0.0089027	-0.0712942	-0.1829002
log_AMT_CREDIT	-0.4963663	-0.3058538	0.0179660	0.0098952	-0.1328703
AGE_YEARS	-0.0895071	0.0909410	-0.6812394	-0.0661836	0.5211075
DIFF_CREDIT_GOODS	-0.3797467	-0.2256788	0.0707076	-0.1959997	-0.1718557
RATIO_CREDIT_INCOME	-0.5366794	0.2372406	0.0189005	0.0813545	0.0015314
RATIO_ANNUITY_CREDIT	0.3908576	0.2440248	0.1423681	0.1802348	-0.3724869
DTI_RATIO	-0.3972254	0.4446957	0.1221834	0.2169086	-0.2344155

Si nos fijamos en el gráfico que incluye los dos primeros planos factoriales (PC1 y PC2), resulta fácil ver que **log_AMT_CREDIT** y **DIFF_CREDIT_GOODS** se solapan en su proyección, teniendo **log_AMT_CREDIT** más contribución dado que el vector es más largo. De aquí se entiende que las correlaciones de ambas variables en los dos primeros planos factoriales son muy similares, motivo por el cual solapan. En la tabla de correlaciones anterior se puede comprobar como efectivamente, estas correlaciones son similares. Incluso la correlación en ambas variables con la tercera dimensión (PC3) es baja, de forma parecida. Por tanto, se procede a eliminar aquella con menos contribución en PC1 y PC2, esta siendo **DIFF_CREDIT_GOODS**.

Ahora se conservan 8 variables numéricas.

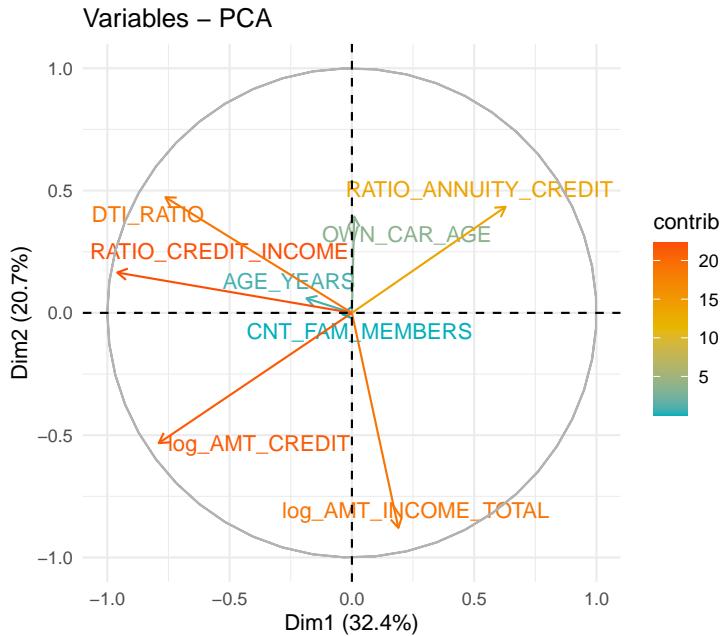
Se vuelven a ejecutar todos los pasos anteriores para volver a verificar si hace falta eliminar más variables:

Figura 58: Porcentaje de inercia explicado por dimensión



Se aprecia como la eliminación de **DIFF_CREDIT_GOODS** ha modificado el número de dimensiones necesarias para alcanzar el 80 % de inercia acumulada, pasando de 5 a 4 dimensiones.

Figura 59: Proyección de variables en los dos primeros planos factoriales



Cuadro 22: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4
OWN_CAR_AGE	0.0056870	0.3062623	-0.0030588	-0.9203547
CNT_FAM_MEMBERS	-0.0247561	-0.0041456	-0.7074073	-0.1177313
log_AMT_INCOME_TOTAL	0.1179538	-0.6836391	-0.0420849	-0.1218891
log_AMT_CREDIT	-0.4904066	-0.4139790	-0.0598471	-0.0681678
AGE_YEARS	-0.1154462	0.0462637	0.6823184	-0.0564692
RATIO_CREDIT_INCOME	-0.5958824	0.1282325	-0.0355111	0.0417412
RATIO_ANNUITY_CREDIT	0.3902375	0.3372441	-0.1098849	0.2629810
DTI_RATIO	-0.4735547	0.3675972	-0.1237687	0.2132899

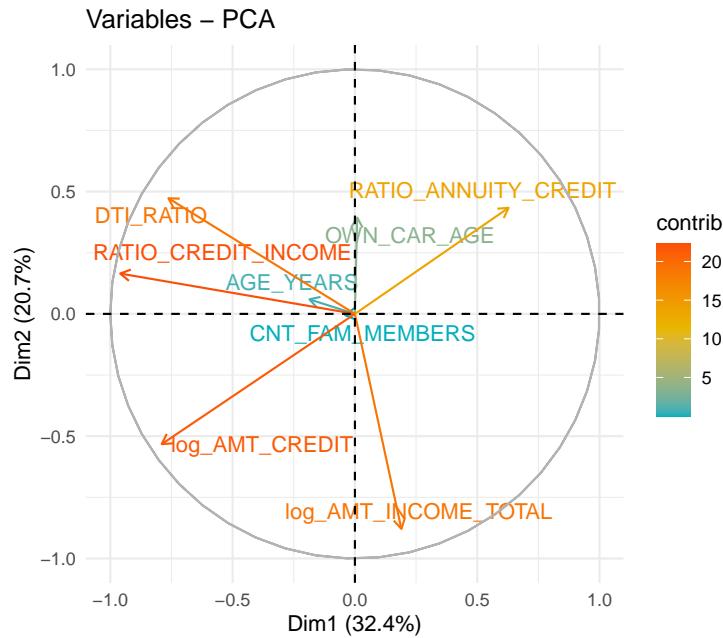
Comprobando el gráfico de las dos primeras dimensiones, y analizando las correlaciones, parece ser que ya no hace falta eliminar más variables. Por tanto, conservamos 8 variables numéricas.

Las variables eliminadas han sido: - **AMT_INCOME_TOTAL**, **AMT_CREDIT**, **AMT_ANNUITY**, **AMT_GOODS_PRICE**, todas ellas con motivo de que ya se había creado otra variable a partir de su transformación logarítmica. - **DAYS_BIRTH**, ya que la variable **AGE_YEARS** es una transformación de ella. - **log_AMT_GOODS_PRICE** - **log_AMT_ANNUITY** - **DIFF_CREDIT_GOODS**

Interpretación de planos factoriales

Para ayudar a dar nombre a las diferentes dimensiones, aparte de utilizar las herramientas gráficas, también podemos fijarnos en las correlaciones entre las variables y los componentes principales.

Figura 60: Proyección de variables en los dos primeros planos factoriales



Cuadro 23: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4
OWN_CAR_AGE	0.0056870	0.3062623	-0.0030588	-0.9203547
CNT_FAM_MEMBERS	-0.0247561	-0.0041456	-0.7074073	-0.1177313
log_AMT_INCOME_TOTAL	0.1179538	-0.6836391	-0.0420849	-0.1218891
log_AMT_CREDIT	-0.4904066	-0.4139790	-0.0598471	-0.0681678
AGE_YEARS	-0.1154462	0.0462637	0.6823184	-0.0564692
RATIO_CREDIT_INCOME	-0.5958824	0.1282325	-0.0355111	0.0417412
RATIO_ANNUITY_CREDIT	0.3902375	0.3372441	-0.1098849	0.2629810
DTI_RATIO	-0.4735547	0.3675972	-0.1237687	0.2132899

- **PC1:** Las variables más fuertemente correlacionadas con esta dimensión son **RATIO_CREDIT_INCOME**, **log_AMT_CREDIT** y **DTI_RATIO**, todas correlacionadas de forma negativa y en este respectivo orden. Con ello, podemos pensar que el primer plano factorial (**PC1**) tiene relación con “**Nivel monetario según préstamos**”. Puede entenderse que valores más elevados en la proyección sobre el primer plano factorial (**PC1**) indican individuos con unas diferencias menores entre el crédito pedido y lo que ingresan anualmente, y con préstamos más bajos a nivel monetario.
- **PC2:** Las variables con mayor correlación con la segunda dimensión, en orden decreciente, son **log_AMT_INCOME_TOTAL** con correlación negativa, y **log_AMT_CREDIT** con correlación negativa y **DTI_RATIO** con correlación positiva. Se puede intuir que los individuos con

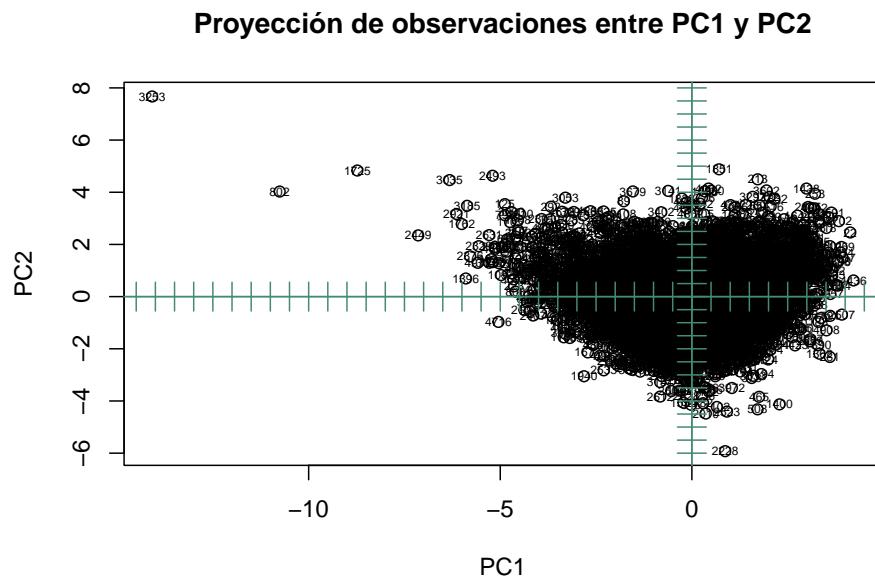
valores más altos en la proyección del **PC2** serán aquellos con unos ingresos totales menores y creditos concedidos menores. Por lo tanto, el segundo plano factorial (**PC2**) podría quedar definido por “**Nivel de ingresos según créditos**”

- **PC3:** Para este tercer plano factorial, las variables más significativas son **CNT_FAM_MEMBERS** de forma negativa y **AGE_YEARS** de forma positiva. Así pues, aquellos individuos que cumplen estas características son clientes con familias poco numerosas y mayores (si su año de nacimiento es un valor alto, significa que son más mayores, dado a la correlación positiva con la variable de edad). Podría decirse que el tercer plano factorial (**PC3**) representa la “**Edad y grandaría familiar**”.
 - **PC4:** Para el cuarto plano factorial, se puede ver que la variable con mayor contribución en gran diferencia a las demás es **OWN_CAR_AGE**, correlacionada de forma negativa. Es decir, los clientes con valores de proyección en PC4 más grandes serán aquellos con coches más nuevos. Por lo tanto, el cuarto plano factorial (**PC4**) podría recibir el nombre de “**Edad vehículo**”.

Representación de individuos

A continuación se representan los individuos sobre los dos primeros planos factoriales.

Figura 61: Proyección de individuos en los dos primeros planos factoriales

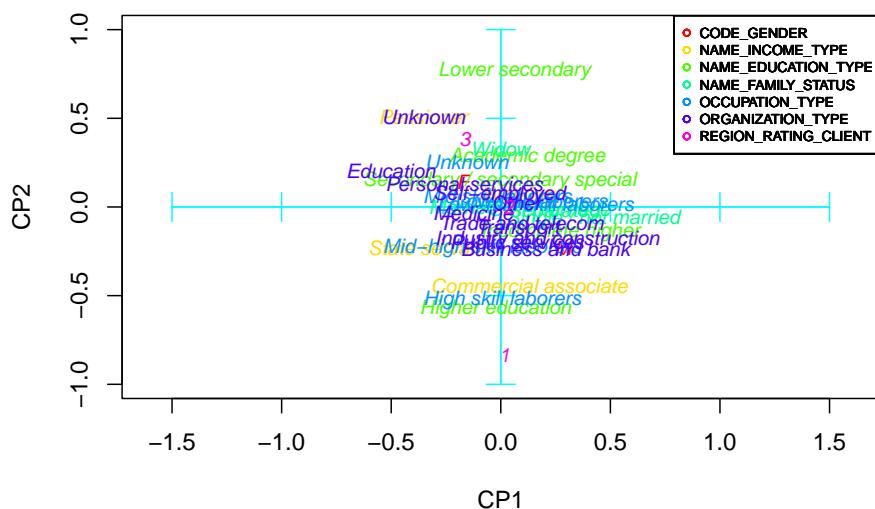


Como se puede observar, no se aprecian grupos diferenciados a partir de la proyección de los individuos. Hay una gran cantidad de estos que se concentran alrededor del origen de coordenadas, dando a entender que son individuos “ordinarios”. Sí se observan algunos puntos alejados de la nube principal, estos perteneciendo a la representación de algunos individuos con características más extrañas a las del conjunto central de individuos.

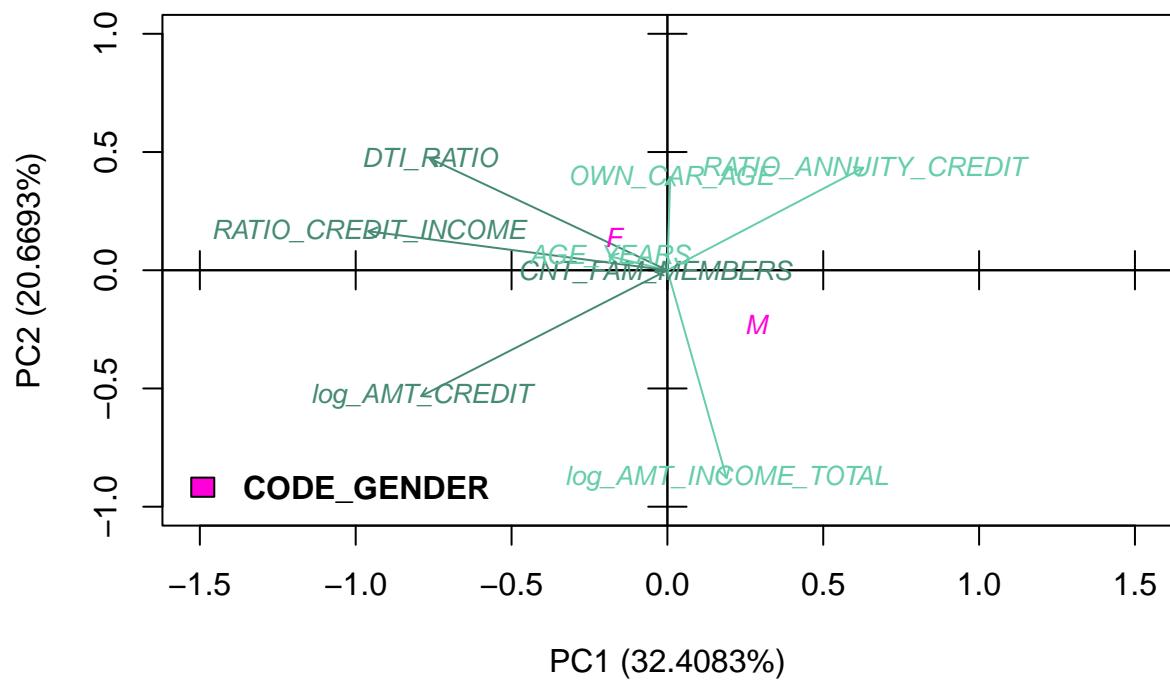
Representación de variables categóricas en primeros planos factoriales

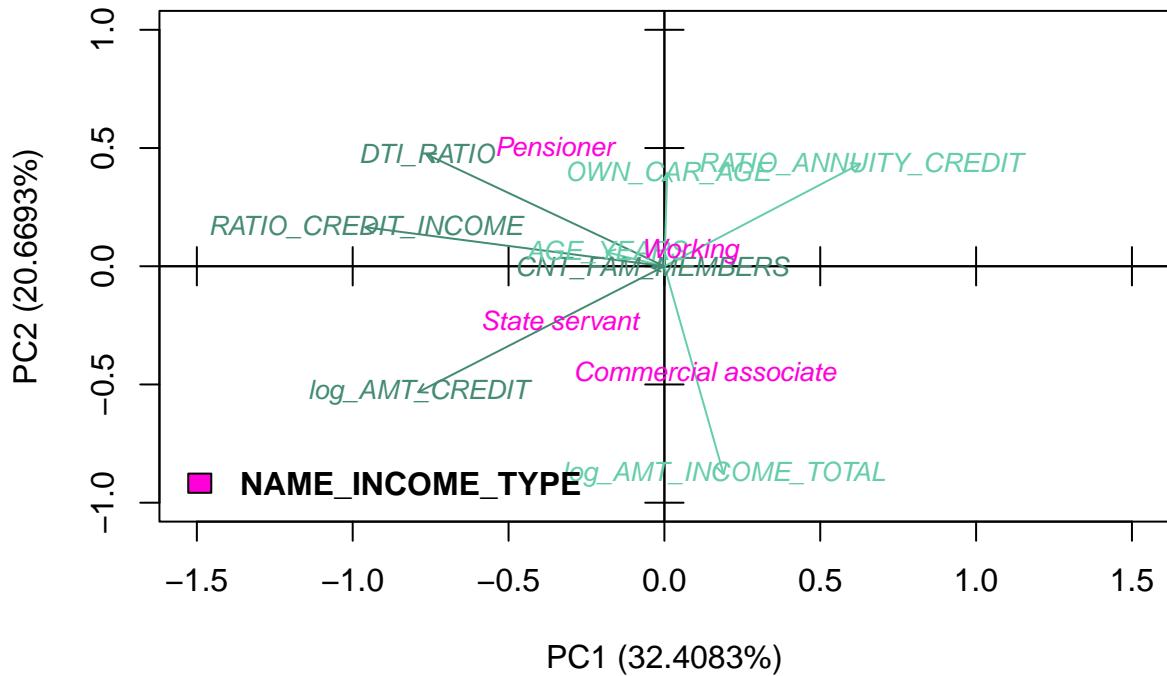
Una vez se han establecido los planos factoriales gracias a las variables numéricas, es necesario representar también las variables categóricas para así acabar de hacer un estudio completo usando toda la base de datos de la variable estudiada. De esta forma, se han representado los centroides de las coordenadas de cada nivel de cada variable categórica y se han obtenido los siguientes resultados:

Figura 62: Representación conjunta de variables categóricas en los dos primeros planos factoriales

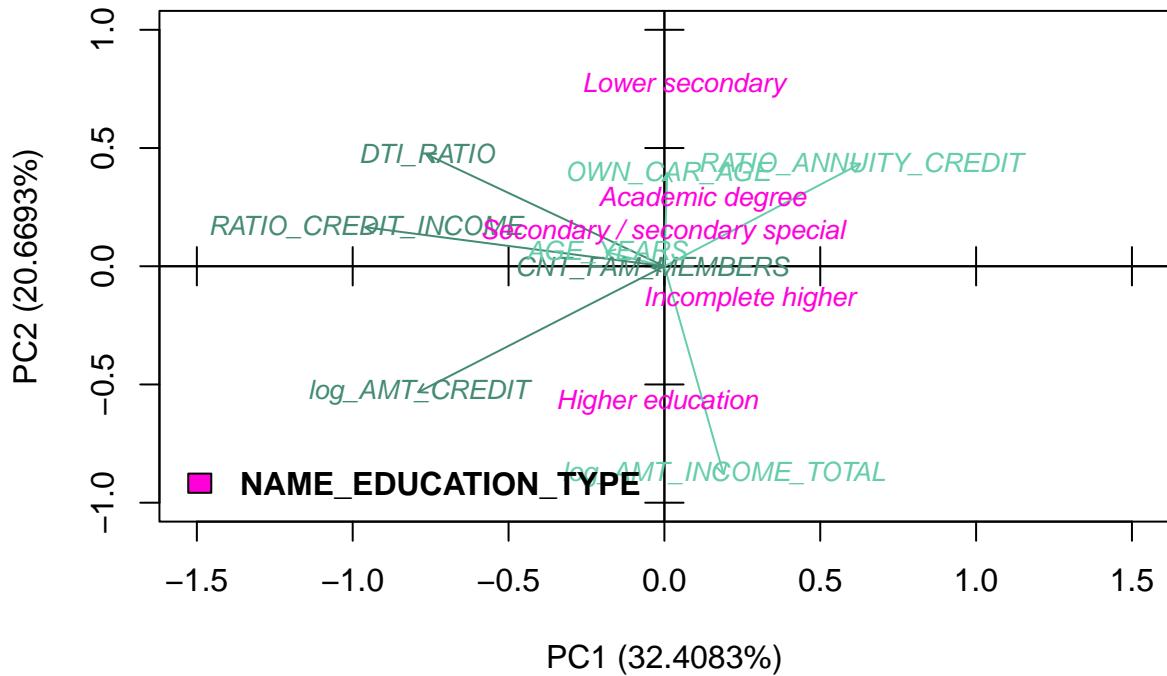


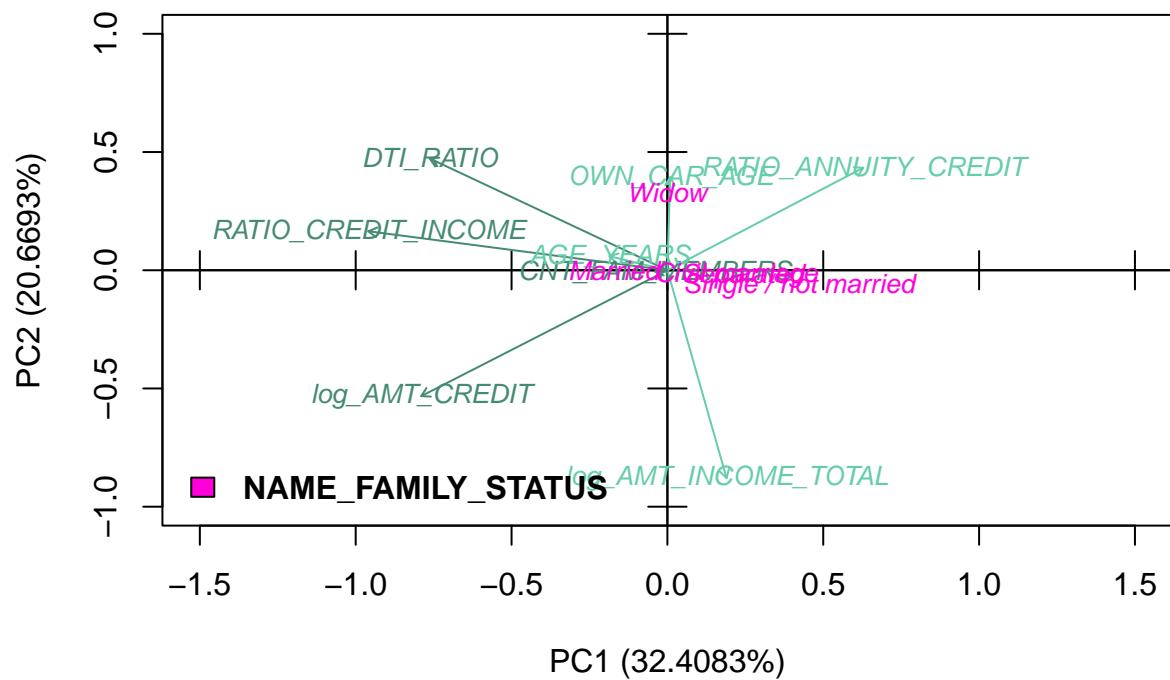
En este primer gráfico no se puede ver nada con claridad, por eso se ha decidido representar cada una de las variables categóricas en un gráfico distinto:

Proyecciones sobre el plano factorial de variables categóricas

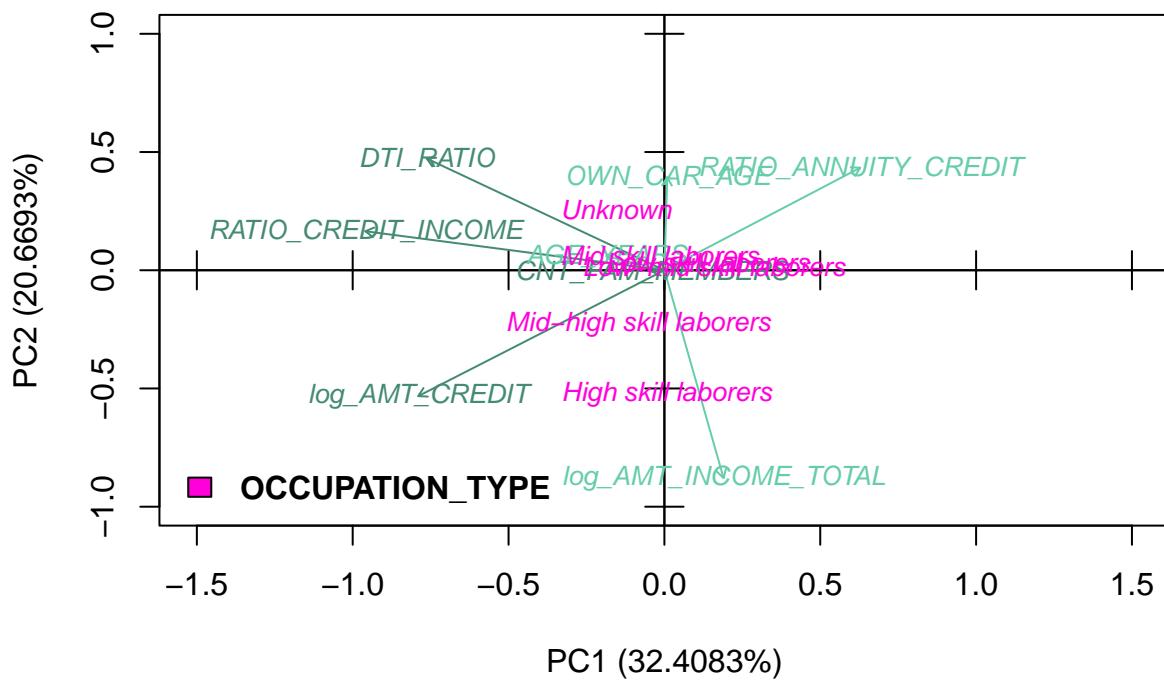
Proyecciones sobre el plano factorial de variables categóricas

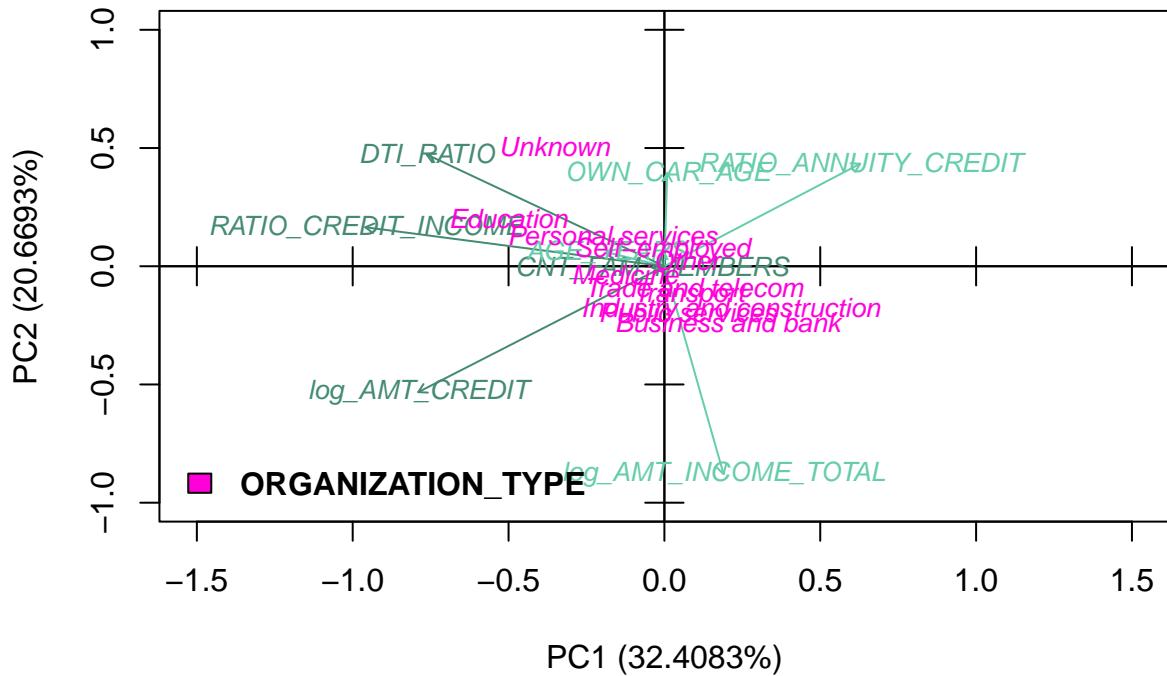
Proyecciones sobre el plano factorial de variables categóricas



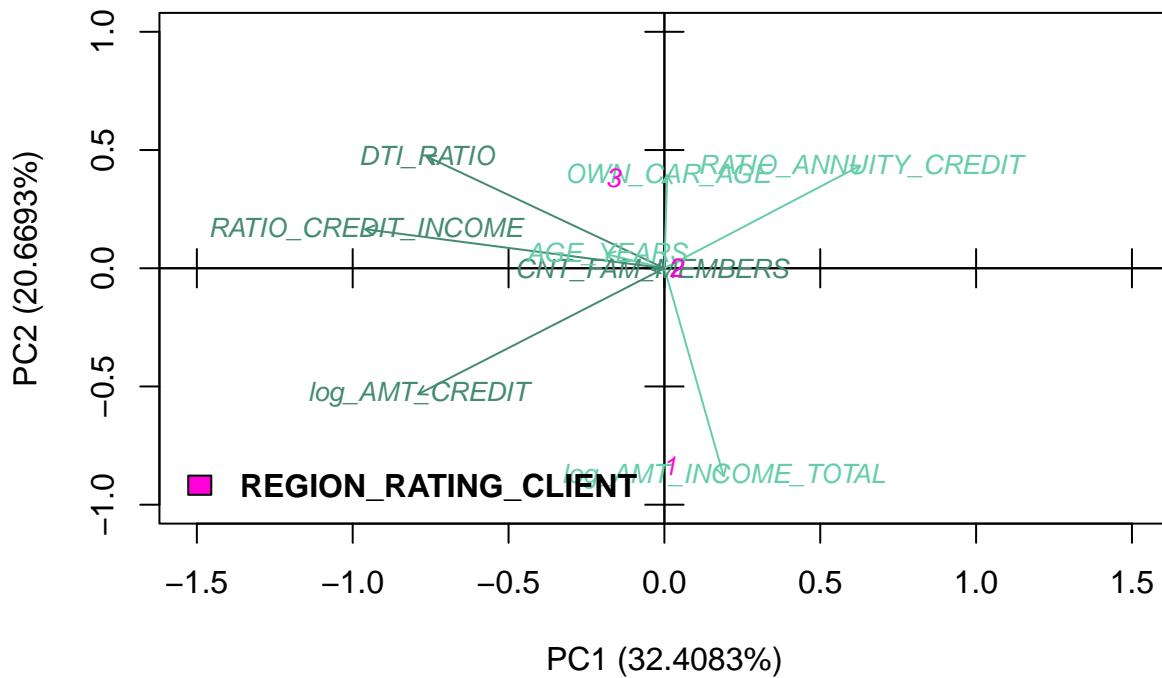
Proyecciones sobre el plano factorial de variables categóricas

Proyecciones sobre el plano factorial de variables categóricas



Proyecciones sobre el plano factorial de variables categóricas

Proyecciones sobre el plano factorial de variables categóricas



Algunos de los gráficos anteriores son interesantes de comentar. En el caso del gráfico que representa **NAME_EDUCATION_TYPE**, y de acuerdo con las descripciones establecidas de las dimensiones, se puede observar como los individuos con una educación “Lower secondary” son los que cuentan con unos ingresos totales menores y créditos concedidos de menor valor. Por otro lado, los individuos con una educación “Higher education” parecen ser los que piden crédito prestado de mayor valor monetario, y para los cuales sus ingresos totales son mayores. Uno de los motivos por los que se podría dar esto es por los préstamos solicitados para pagar la educación superior, y teniendo en cuenta que la base de datos es tomada en los Estados Unidos, se sabe que el precio de estos estudios es muy caro.

Observando el gráfico que incluye **CODE_GENDER**, se puede apreciar como los dos sexos presentan diferencias en la primera dimensión. De acuerdo con la explicación de la dimensión, los hombres son los que, en general, piden préstamos de menor valor monetario, y para los cuales la diferencia entre el crédito del préstamo y los ingresos anuales es menor. Es decir, que los hombres cuentan con menos años para pagar las deudas de los préstamos. Por el lado contrario, las mujeres presentan las características opuestas, préstamos más grandes y diferencias más significativas entre ingresos anuales y valor del crédito.

Para la variable **NAME_FAMILY_STATUS** se puede apreciar que los individuos que forman parte de la categoría “Widow” son también los que tienen menores ingresos y créditos concedidos de menor valor. En cambio, y fijando la atención en la variable **OCUPATION_TYPE**, se puede observar que las categorías de “Mid-high skill laborers” y “High skill laborers” representan lo contrario. Entre todos los tipos de ocupación, son los dos con mayores ingresos y mayores créditos concedidos, que se entiende como un fenómeno muy trivial.

En relación a la variable **REGION_RATING_CLIENT**, se aprecia la diferencia entre las tres puntuaciones de región; cuanto mejor es la “puntuación” de la región (1 siendo la mejor), mayores ingresos y más

crédito concedido tienen los clientes de dicha región.

Por último, en el gráfico que representa la variable **NAME_INCOME_TYPE**, se pueden analizar las dos dimensiones por separado. Primero, si se comprueba la primera dimensión, es interesante ver que tanto los pensionistas como los funcionarios tienden a pedir préstamos de mayor valor, y ambos presentan diferencias entre ingresos anuales y el valor de dicho préstamo solicitado. Segundo, si se observa en función de la segunda dimensión, se aprecia que los pensionistas son los que presentan unos ingresos totales menores, mientras que las personas con ingresos derivados de puestos de trabajo comerciales son las que tienen ingresos totales mayores.

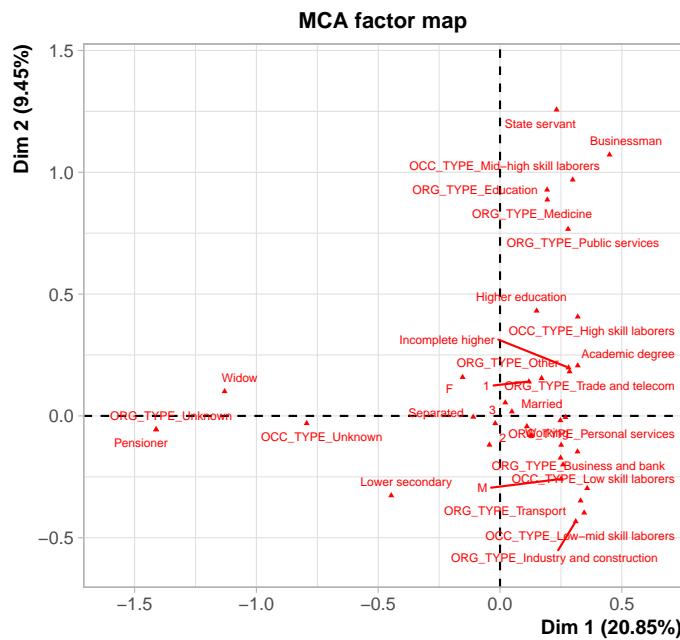
Análisis de correspondencias múltiples (ACM)

El Multiple Correspondance Analysis, ACM en adelante, es un método de análisis factorial para variables categóricas que permite analizar relaciones entre variables, así como reducir la dimensionalidad de la base de datos seleccionando sólo aquellas variables relevantes. Para realizarlo se deben escoger unas variables activas y otras de complementarias. En este caso, como el número de variables es relativamente bajo y se consideran todas relevantes, no se considerará ninguna variable complementaria. Además, se añadirán al análisis las variables numéricas como variables suplementarias, aunque solamente aquellas que se han considerado relevantes en el ACP.

Consideramos hacer una nueva codificación de las variables, reduciendo su longitud, de tal manera que los resultados obtenidos para el análisis se observen más claramente:

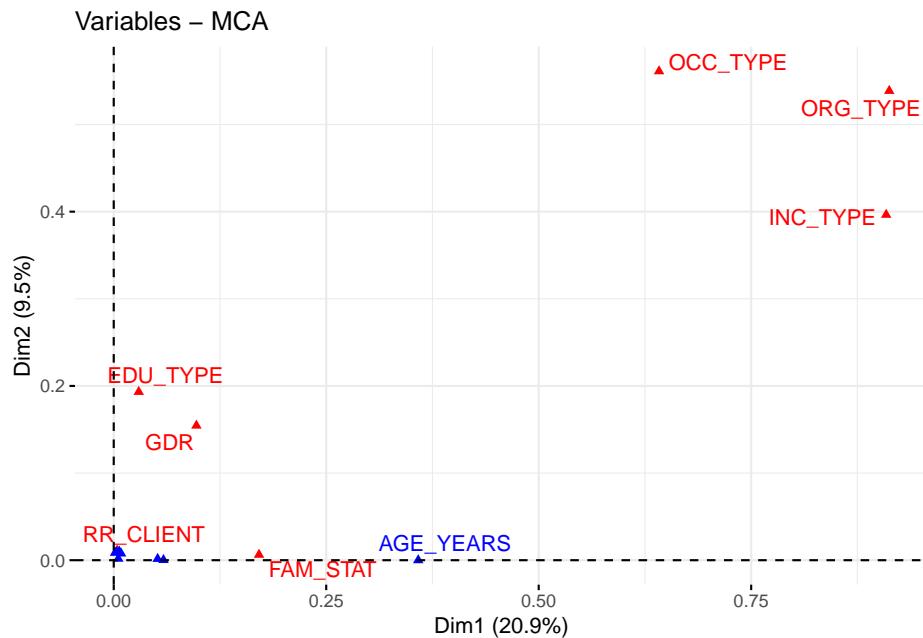
Desarrollo del ACM

Figura 63: Correlación entre Variables y Dimensiones Principales



En esta primera figura se representan las relaciones entre las modalidades de todas las variables categóricas con las dos primeras dimensiones del MCA. Se observa que la dimensión 1 se asocia con las variables que tienen relación con la edad, como la modalidad de Pensionista y Viudo. Se aprecia que la dimensión 2 se asocia a las modalidades relacionadas con la cualificación del trabajo del individuo, con una asociación positiva entre la dimensión y la cualificación del trabajador. Por tanto, se llamará a la dimensión 1 Edad, y a la dimensión 2 como cualificación del trabajador.

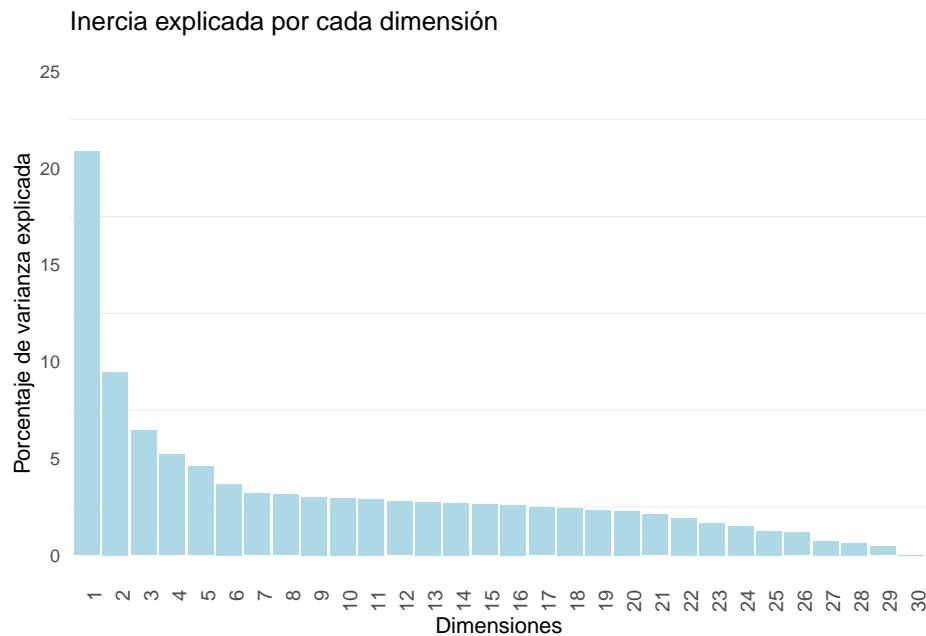
Figura 64: Variabilidad de las Variables Categóricas en las Dos Primeras Dimensiones



En el gráfico obtenido puede verse la variabilidad que expresan cada una de las variables categóricas en función de las dimensiones 1 y 2. Aquellas variables que estén más cerca del origen de coordenadas aportan muy poca información respecto a la variabilidad de los datos y, por tanto, son poco importantes. En cambio, aquellas variables más alejadas del centro aportan información más relevante.

Se representan gráficamente la inercia que explica cada una de las dimensiones generadas:

Figura 65: Inercia Explicada por cada Dimensión



Si una dimensión tiene una inercia baja, significa que todas las modalidades están muy cercanas al centro de gravedad y, en consecuencia, son muy similares. A medida que aumenta la inercia, va aumentando la distancia al centro de gravedad y, por tanto, se reduce la similitud.

Para poder estudiarlo más a fondo, se realiza la siguiente tabla en la que se puede observar para cada dimensión, su valor propio, el porcentaje de varianza (o inercia) explicada, y el porcentaje de varianza (o inercia) acumulada:

Cuadro 24: Varianza Explicada por cada Dimensión

	Valor propio	Porcentaje de la Varianza Acumulada	Porcentaje de Varianza
dim 1	0.16	20.85	20.85
dim 2	0.07	9.45	30.30
dim 3	0.05	6.45	36.76
dim 4	0.04	5.21	41.96
dim 5	0.03	4.59	46.56
dim 6	0.03	3.68	50.23
dim 7	0.02	3.19	53.42
dim 8	0.02	3.14	56.56
dim 9	0.02	3.00	59.57
dim 10	0.02	2.97	62.54
dim 11	0.02	2.91	65.45
dim 12	0.02	2.79	68.24
dim 13	0.02	2.74	70.98
dim 14	0.02	2.71	73.69
dim 15	0.02	2.63	76.32
dim 16	0.02	2.61	78.94
dim 17	0.02	2.48	81.41
dim 18	0.02	2.47	83.88
dim 19	0.02	2.33	86.21
dim 20	0.02	2.28	88.49
dim 21	0.02	2.12	90.61
dim 22	0.01	1.93	92.53
dim 23	0.01	1.68	94.21
dim 24	0.01	1.51	95.72
dim 25	0.01	1.25	96.97
dim 26	0.01	1.18	98.15
dim 27	0.01	0.74	98.90
dim 28	0.00	0.65	99.55
dim 29	0.00	0.45	100.00
dim 30	0.00	0.00	100.00

Tenemos un total de 31 dimensiones. La dimensión 1 destaca muy por encima del resto, explicando un 20.85 % de la variabilidad de los datos, seguida de la dimensión 2, explicando un 9.45 % de la variabilidad de los datos los datos. A partir de la dimensión 6, se ve que la gráfica se estabiliza bastante ya hasta la última dimensión.

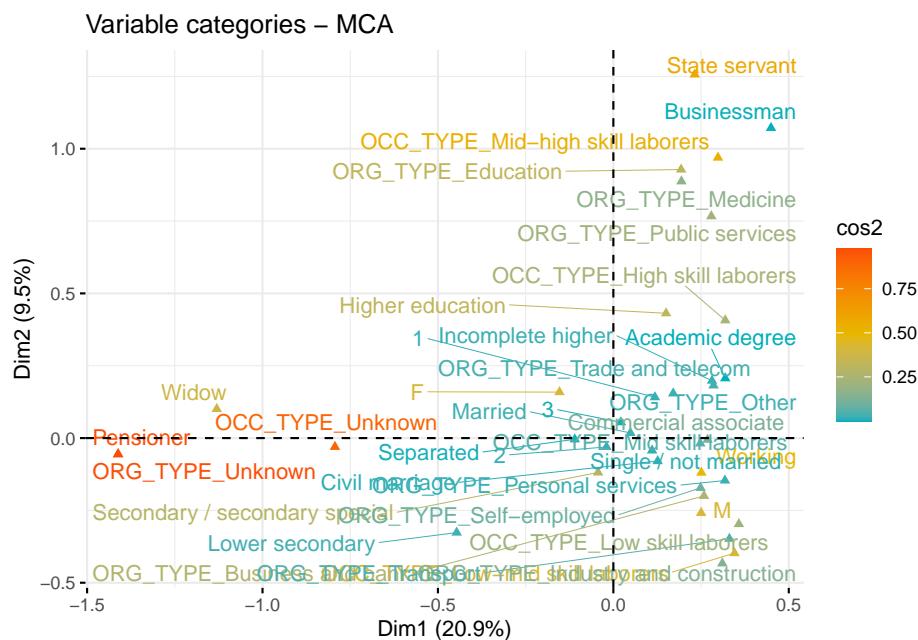
Por tanto, en total las dos primeras dimensiones ya explican un 30.3 % de la variabilidad de los datos, y se necesitan 17 dimensiones para llegar a tener una inercia acumulada por encima del 80 %.

Aunque las primeras dos dimensiones expliquen cerca del 30 % de la inercia, no todos los puntos se muestran igual de bien en las dos dimensiones. La calidad de la representación se llama coseno cuadrado (\cos^2), que mide el grado de asociación entre categorías de variables y un eje particular.

A continuación, se representa la calidad de las categorías a partir de ajustar los colores para cada punto proyectado, tomando como criterio el valor del coseno cuadrático (\cos^2). Si una categoría de variable está bien representada por dos dimensiones, la suma de \cos^2 es cercana a uno. Para algunos de los elementos de la fila, se requieren más de dos dimensiones para representar perfectamente los datos. Se considera lo siguiente:

- Las categorías de variables con valores bajos de cos2 se colorearán en “cian”.
 - Las categorías de variables con valores medios de cos2 se colorearán en “amarillo”.
 - Las categorías de variables con valores altos de cos2 se colorearán en “rojo”.

Figura 66: Calidad de las Variables Categóricas a partir del Coseno Cuadrático

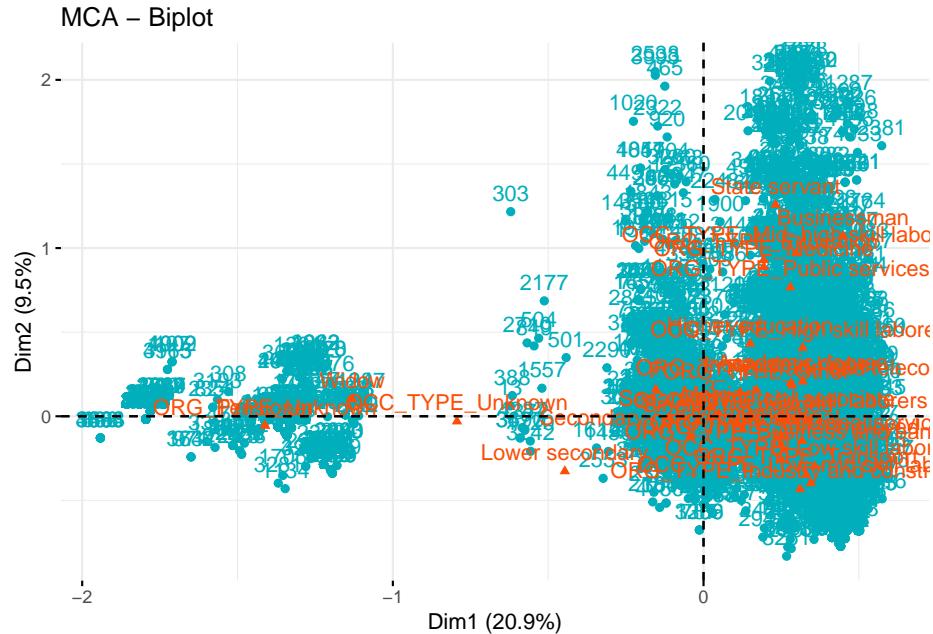


Salen muchas categorías que no están muy bien representadas por las dos primeras dimensiones. Esto implica que la posición de los puntos correspondientes en el diagrama de dispersión debe interpretarse con cierta cautela. Probablemente sea necesaria una solución de mayor dimensión. Aún así, se ha decidido no realizar el MCA de mayores dimensiones debido a la dificultad de representación gráfica. Por tanto, los resultados del MCA se analizarán con cautela, especialmente las variables poco representadas en las dos primeras dimensiones.

Gráfico de individuos y variables

Para una primera visualización de la relación entre las variables y las observaciones en un espacio reducido de dimensiones, acudimos al gráfico biplot. Este gráfico nos facilita la interpretación de la estructura de los datos al proporcionar una visualización que muestra la relación entre variables categóricas y observaciones.

Figura 67: Biplot de Individuos y Categorías

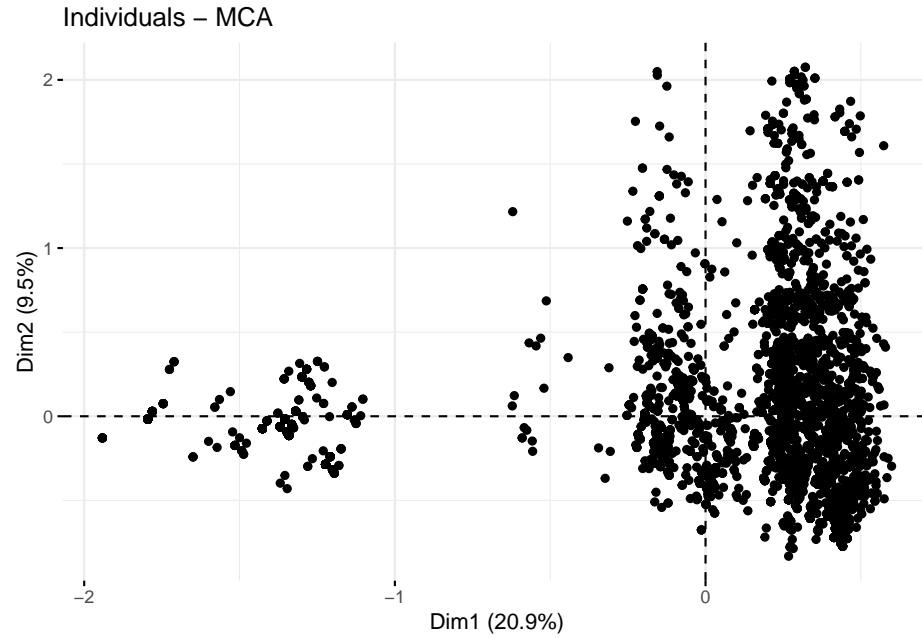


Con tal de poder analizar estos resultados de manera más precisa, se decide dividir este gráfico y analizar por una parte únicamente el gráfico de observaciones y por otra parte el gráfico de las variables categóricas.

Gráfico de individuos

Se representa gráficamente cómo se distribuyen los individuos en función de las dos primeras dimensiones que explican un 29.99 % de la variabilidad:

Figura 68: Gráfico de Individuos en las dos Primeras Dimensiones

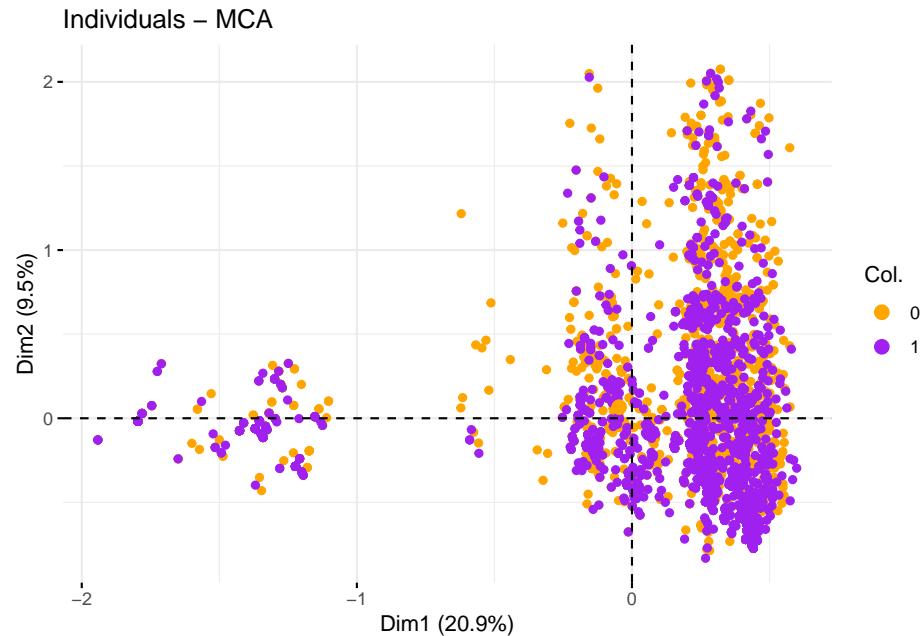


A simple vista, se aprecian varios grupos de individuos pero resulta difícil distinguir cuántos. Sin embargo, sí podríamos decir que los individuos se dividen en como mínimo 2 grupos. Para distinguir mejor las agrupaciones de individuos y su asociación con algunas modalidades se pasa a estudiar cada variable para observar si existe algún tipo de asociación entre ellas.

Gráfico de los individuos según variable TARGET

A continuación representamos los mismos individuos pero coloreandolos según la variable “target”, es decir, nuestra variable output, donde 1 indica aquel cliente con dificultades de pago, y 0 contrariamente:

Figura 69: Gráfico de Individuos según la Variable TARGET en las dos Primeras Dimensiones

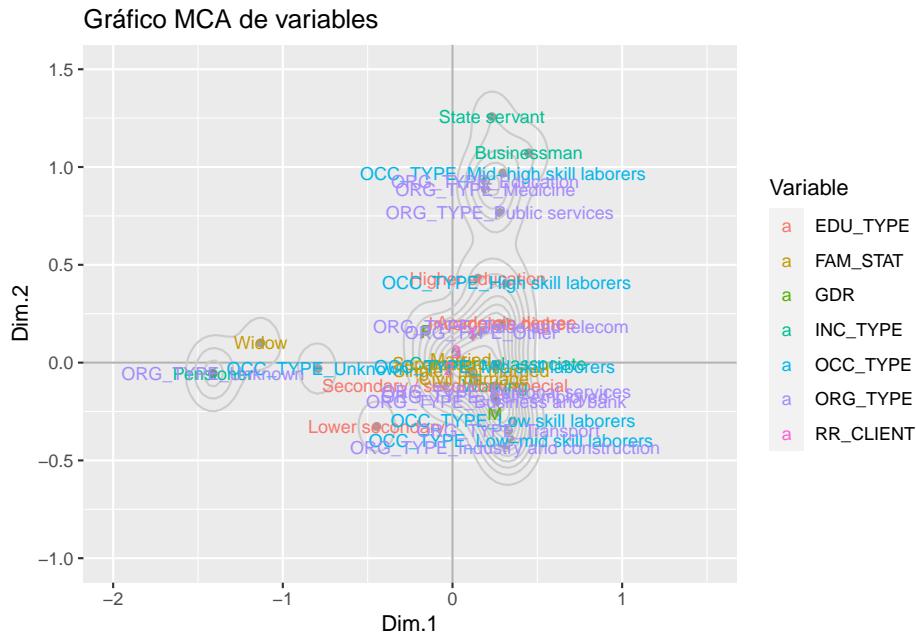


Observamos como diferenciando a los individuos según si tienen dificultades o no con el pago, no hay diferencias entre grupos de individuos, por lo tanto, podemos decir que no se ve ninguna asociación entre la variable TARGET y las modalidades de las variables representadas en estas dos dimensiones.

Gráfico de variables

Para tener una representación aún más clara sobre las variables y su asociación, a continuación se grafican estas variables con curvas de densidad para ver aquellas zonas donde hay una mayor concentración.

Figura 70: Representación de Variables en las Dimensiones del ACM



Como bien se ha comentado, la primera dimensión está asociada con las variables que tienen relación con la edad y la segunda dimensión se asocia a modalidades con la cualificación del trabajo del individuo.

Así pues, a partir de este gráfico de densidades, podemos ver como hay una relación muy destacada entre Widow y Pensioner en la dimensión 1. Esta correlación puede deberse a eventos de vida como la pérdida del cónyuge a una elevada edad y factores sociales, finalización laboral y por eso a una pensión por los años trabajados. La asociación de ambos términos con la población anciana y la edad es evidente.

En la dimensión 2 podemos ver una relación entre State Servant y Businessman con Mid/High Skill Laborers y con Education y Medicine. También podríamos ver relación con Higher education. La gente que trabaja en educación y/o mundo sanitario requieren un alto nivel de estudios y son trabajadores altamente cualificados. Al igual que podríamos asociarlo con los trabajadores en Servicios Públicos y Empresarios.

Aún y ver relación en este gráfico, hemos decidido realizar un gráfico de dispersión por cada una de las variables para ver si podíamos adquirir más información.

Gráficos de dispersión agrupado por cada variable

Con el objetivo de ver si las categorías son significativamente diferentes entre sí, se grafican gráficos de elipses alrededor de las categorías de cada una de las variables.

Se considerará que las categorías con elipses no superpuestas, es decir, separadas entre sí, son significativamente diferentes entre sí. Por el contrario, cuando las elipses se superponen, nos indican que hay una similitud o asociación entre categorías, es decir, no son significativamente diferentes entre ellas.

Figura 71: Gráfico de Elipses NAME EDUCATION TYPE

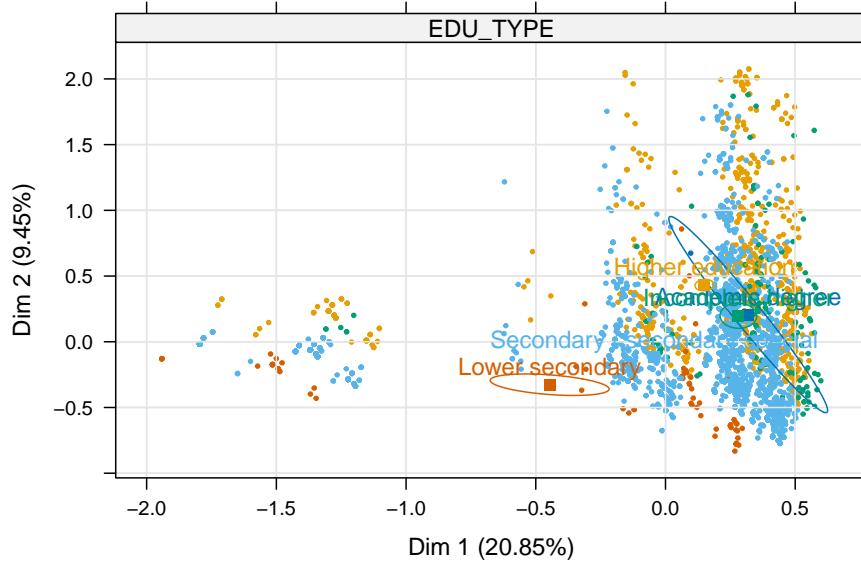


Figura 72: Gráfico de Elipses NAME FAMILY STATUS

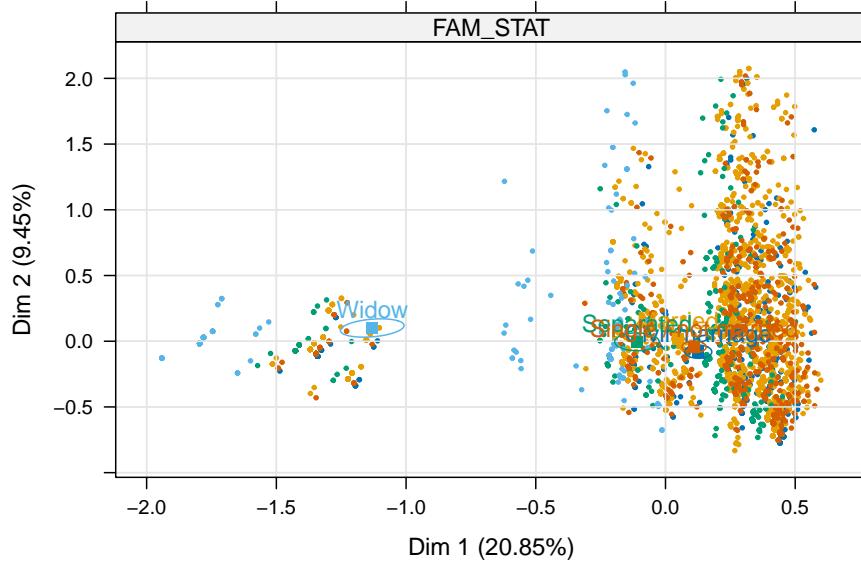


Figura 73: Gráfico de Elipses CODE GENDER

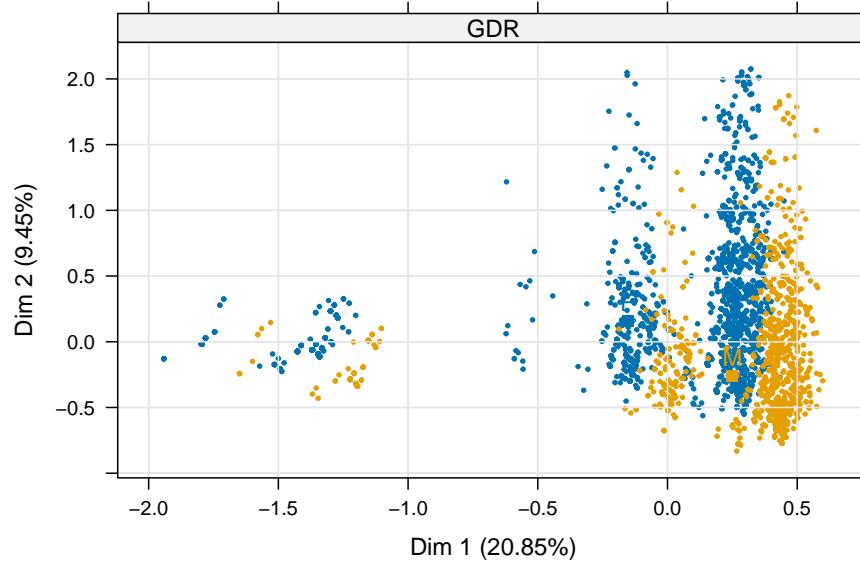


Figura 74: Gráfico de Elipses NAME INCOME TYPE

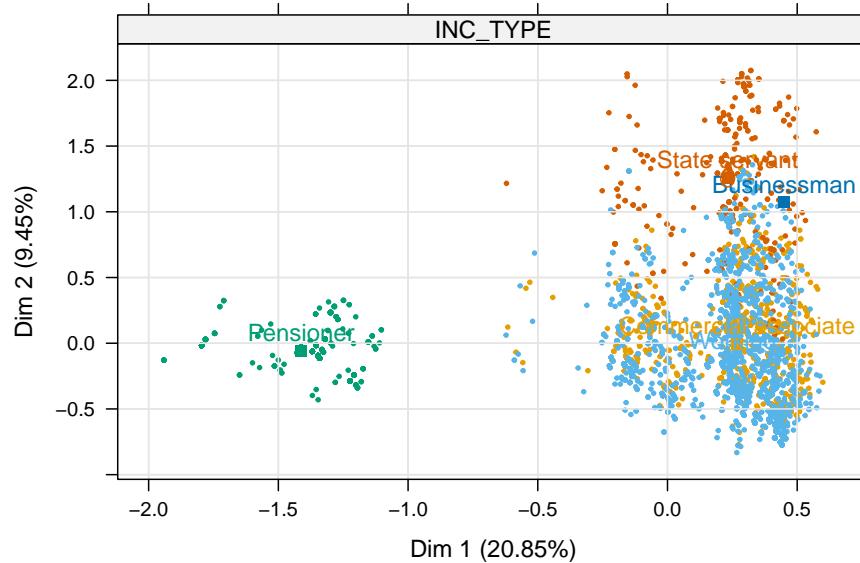


Figura 75: Gráfico de Elipses OCCUPATION TYPE

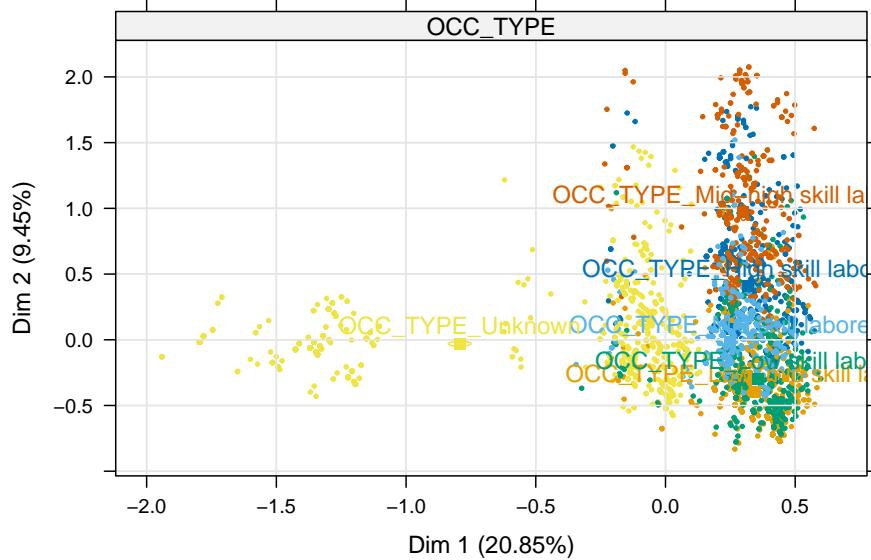


Figura 76: Gráfico de Elipses ORGANITATION TYPE

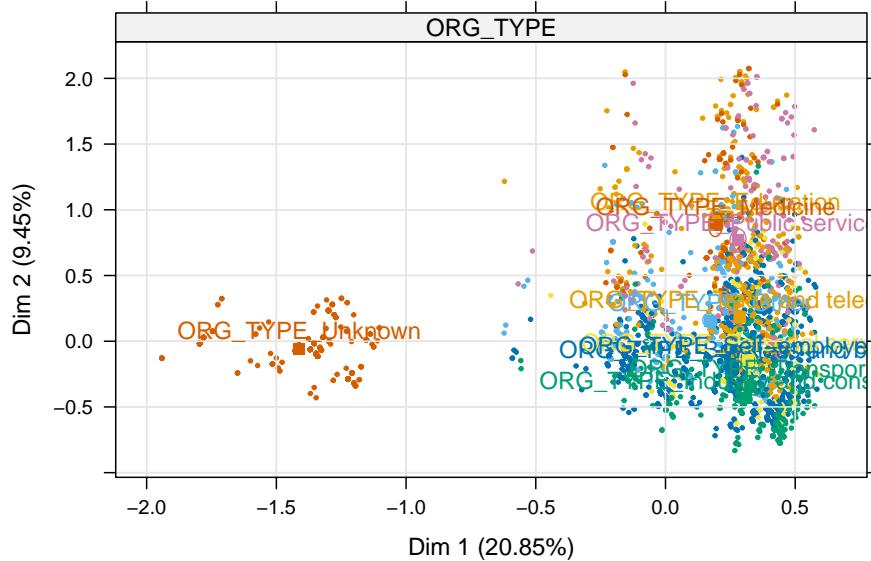
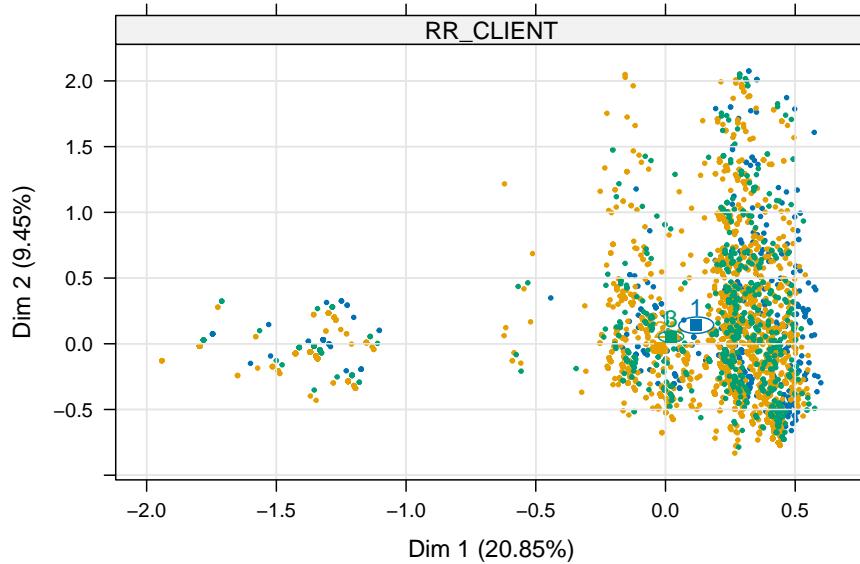


Figura 77: Gráfico de Elipses REGION RATING CLIENT



Al analizar cada variable de manera individual en las dos primeras dimensiones del ACM, se observa que hay muchos gráficos donde encontramos categorías superpuestas entre ellas, y que por tanto, no nos aportan información significativa.

Como se ha observado que en las dos primeras dimensiones únicamente se muestra aproximadamente el 30 % de la variabilidad, para poder estudiar las categorías y su asociación más a fondo, hemos probado de analizar tres dimensiones y no hi hay ningun gráfico significativo que nos permita extraer más información de la que ya hemos extraido con dos dimensiones. Por lo tanto, nos quedamos con los análisis sacados a partir de los gráficos de las dos primeras dimensiones.

Clustering

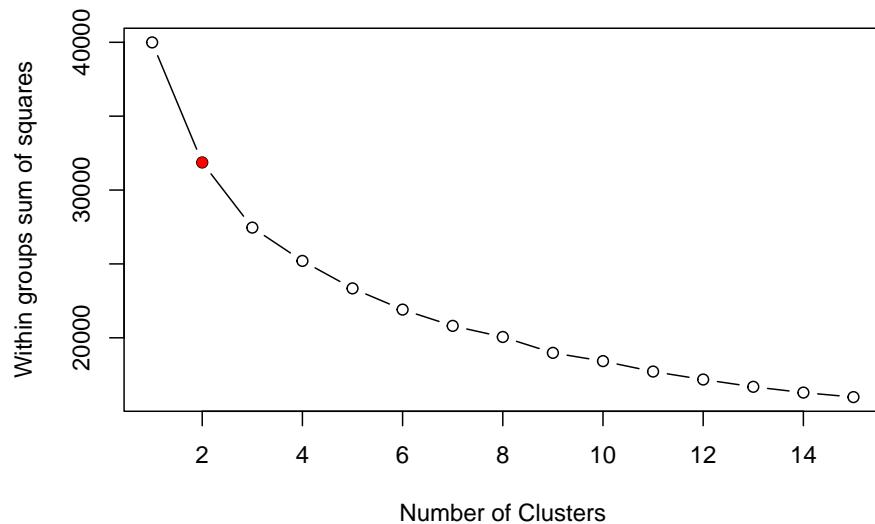
En esta sección, se emplearán diversos algoritmos de clasificación, específicamente el k-means y el Jerárquico. El propósito es asignar cada observación a un grupo correspondiente con el fin de llevar a cabo un perfilado. Este proceso implica etiquetar cada grupo con sus características más significativas, proporcionando así una descripción detallada y distintiva de cada perfil dentro de nuestros datos.

K-means

El algoritmo K-means solamente permitirá utilizar las variables numéricas. Por ello se separarán los datos numéricos de la base de datos preprocesada.

Antes de aplicar el propio algoritmo, se necesita seleccionar el número óptimo de clústeres. Para realizar esto, existen múltiples métodos, uno de ellos es el método del codo. Este consiste en aplicar el K-means para un rango de valores k y luego graficar la suma de los cuadrados de las distancias intraclúster en función de k . Para encontrar el óptimo con este método, sencillamente hace falta encontrar el “codo” del gráfico.

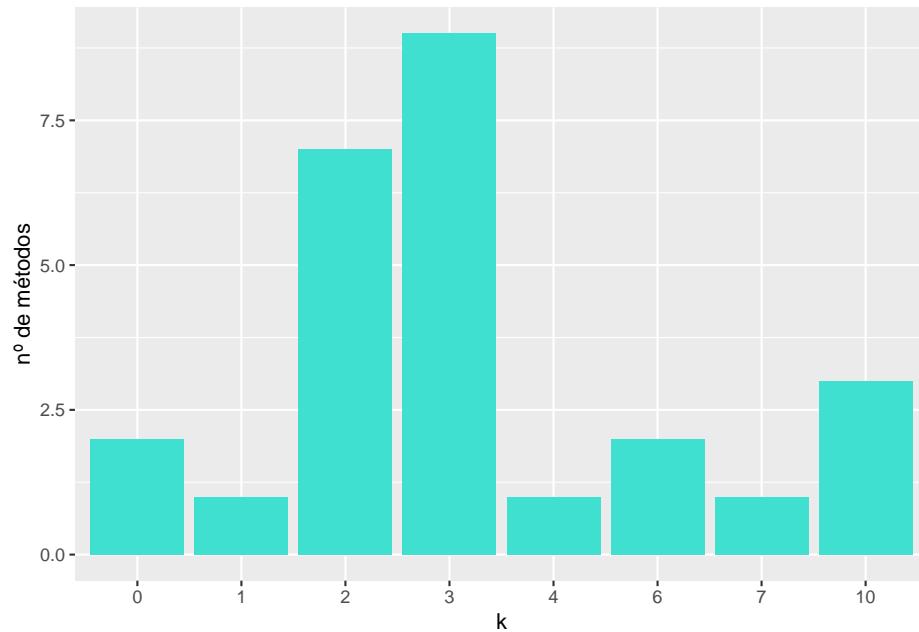
Figura 78: Método del codo



Como se puede apreciar en el gráfico, según el método del codo, el número óptimo de clústeres para el K-means de nuestra base de datos sería $k=2$.

Por lo que sigue, como existen muchos otros criterios para la selección de la k óptima, se usará la función NbClust, que permite aplicar una cantidad de 26 criterios para la selección de k , de esta manera se sabrá con mayor seguridad cuál es el óptimo real. Se grafican los resultados obtenidos por NbClust.

Figura 79: Número óptimo de clústers para el K-means

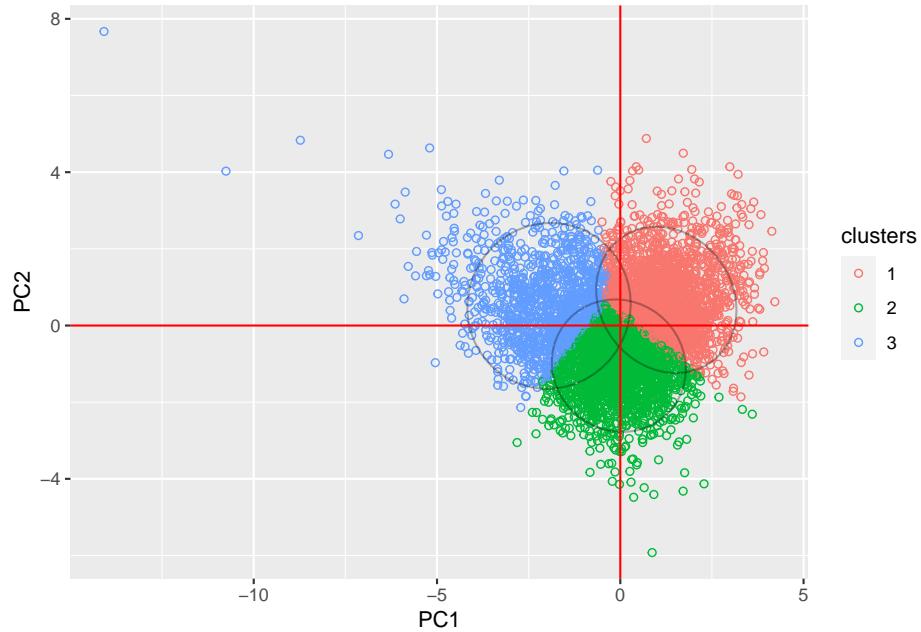


Como se puede apreciar en el histograma, tras haber utilizado todos los criterios, el número óptimo de clústeres que más métodos han escogido es $k=3$.

Como el óptimo se encuentra en $k=3$, el siguiente paso es realizar el K-means con esa k .

Después de aplicar el algoritmo y conseguir el grupo de cada individuo, se muestra el gráfico de los individuos pintados según su clase en el plano factorial de las dos primeras dimensiones del PCA, acompañado de cada una de las elipses de las clases.

Figura 80: Representación de las clases en PC1-PC2



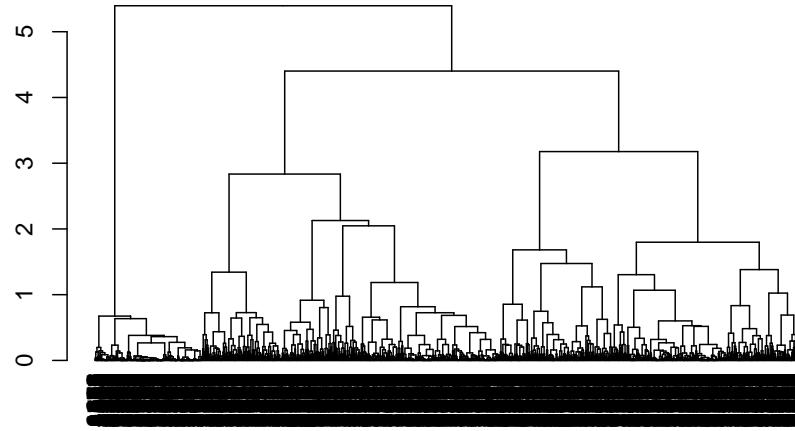
Ahora bien, como se puede ver en el gráfico, no se distinguen muy bien las tres clases, ya que están unas encima de las otras. Esto puede ser consecuencia de que la clasificación se ha hecho considerando solamente las variables numéricas, es por eso que es necesario realizar un clustering jerárquico.

Clustering jerárquico

En primera instancia, para realizar el clustering jerárquico se debe hacer primeramente un dendrograma con el método de Ward con la distancia de *Gower*². En el k-means solo se puede trabajar con las variables numéricas y en la base de datos hay variables tanto numéricas como cualitativas. La distancia de *Gower*² nos permitirá calcular las distancias tanto de las variables numéricas como de las categóricas.

Así, se calcula dicha distancia y se grafica un dendrograma:

Figura 81: Dendograma



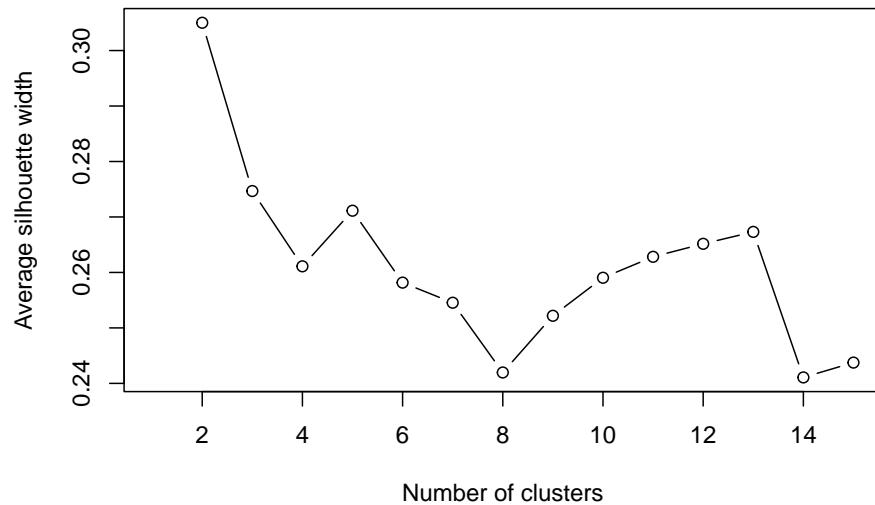
A primera vista se puede apreciar que el corte óptimo parece ser 2 clústeres. Esta cantidad de clústeres puede quedarse pequeña para los objetivos del trabajo. Consiguientemente, trataremos de tomar la decisión analíticamente, usando coeficientes que ayudan a decidir cuál es la mejor cantidad de clústeres.

Uno de ellos es el Coeficiente de Silhouette:

Los valores que retorna el Coeficiente de Silhouette van del 1 al -1. Generalmente, tomarán valores entre 1 y 0, siendo el 1 el mejor valor y 0 indicando la sobreposición de clústeres. Los valores negativos indicarían la asignación incorrecta de la muestra a los clústeres.

Lo que se hace es calcular el Coeficiente de Silhouette para diferentes cantidades de clúster y graficarlo, de manera que se cogerá el mayor valor como el número de clústeres según este criterio de Silhouette.

Figura 82: Silhouette



Como se puede ver, según el criterio de Silhouette, el número de clústeres óptimo es 2. No obstante, como existen muchos otros criterios, se usará -análogamente al K-means- la función NbClust, pero esta vez con la distancia de Gower, de este modo se consideran todas las variables.

Figura 83: Número óptimo de clústeres para el Jerárquico

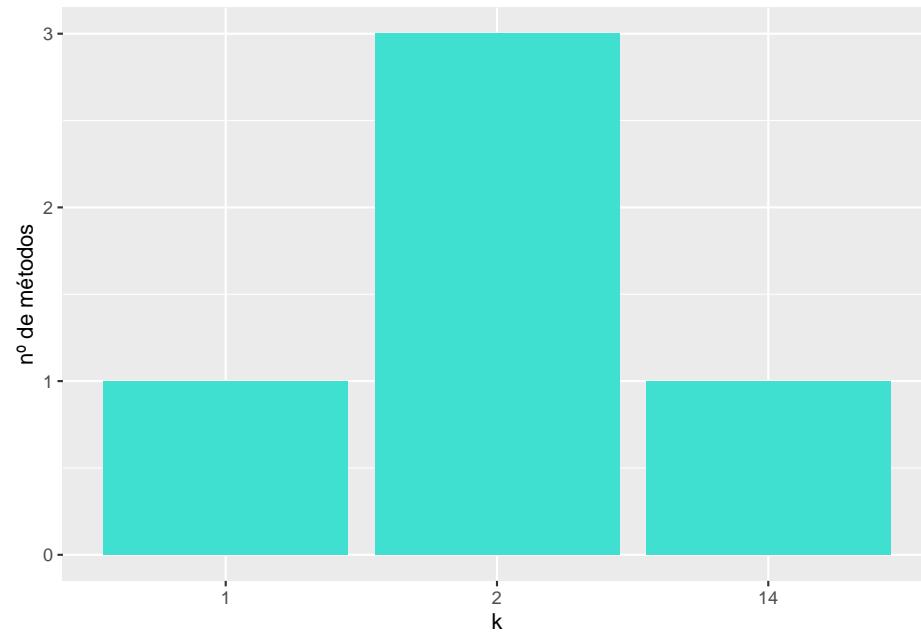
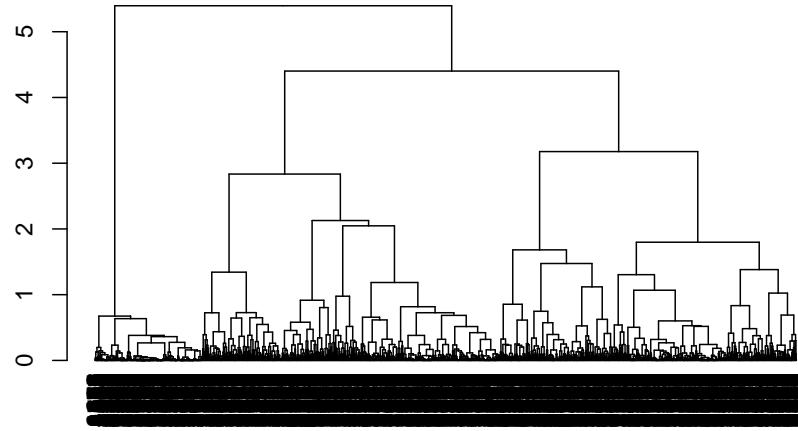
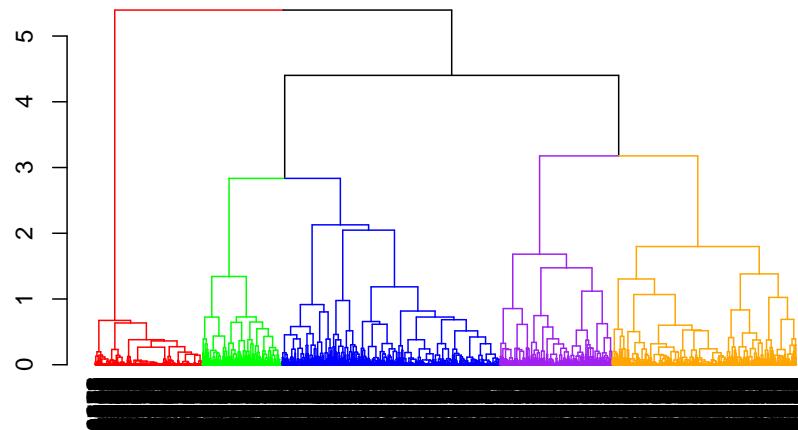


Figura 84: Dendograma



Con el dendrograma anterior, se confirma que el mejor corte (después de $k = 2$) es $k = 3$ y $k = 5$. Se divide el mismo dendrograma en $k = 5$ grupos:

Figura 85: Dendograma con la clasificación por clúster



Se escoge $k=5$ para hacer un perfilamiento de grupos detallado.

Profiling K-means

Con el objetivo de perfilar los grupos conseguidos mediante el algoritmo K-means primero veremos la significación de las variables para los grupos y después se graficarán para identificar las características definitorias de cada grupo.

A continuación se muestran los p-valores para evaluar la significación de cada variable. Primeramente de las variables categóricas y seguidamente las numéricas.

Cuadro 25: Significación de las categóricas

Variable	P_Value
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	0
ORGANIZATION_TYPE	0
REGION_RATING_CLIENT	0

Cuadro 26: Significación de las numéricas

OWN_CAR AGE	c(Cluster = 0.000114855796849977)
CNT_FAM_MEMBERS	c(Cluster = 4.27829589973067e-06)
log_AMT_INCOME_TOTAL	c(Cluster = 0.942503452038247)
log_AMT_CREDIT	c(Cluster = 0)
AGE_YEARS	c(Cluster = 1.546029214791e-21)
RATIO_CREDIT_INCOME	c(Cluster = 0)
RATIO_ANNUITY_CREDIT	c(Cluster = 0)
DTI_RATIO	c(Cluster = 0)

Elaboramos una tabla donde se indica con 1 si se considera variable significativa para el clúster y 0 en caso contrario.

Cuadro 27: Significancia de p-valores para variables numéricas:

	x
OWN_CAR AGE	1
CNT_FAM_MEMBERS	1
log_AMT_INCOME_TOTAL	0
log_AMT_CREDIT	1
AGE_YEARS	1
RATIO_CREDIT_INCOME	1
RATIO_ANNUITY_CREDIT	1
DTI_RATIO	1

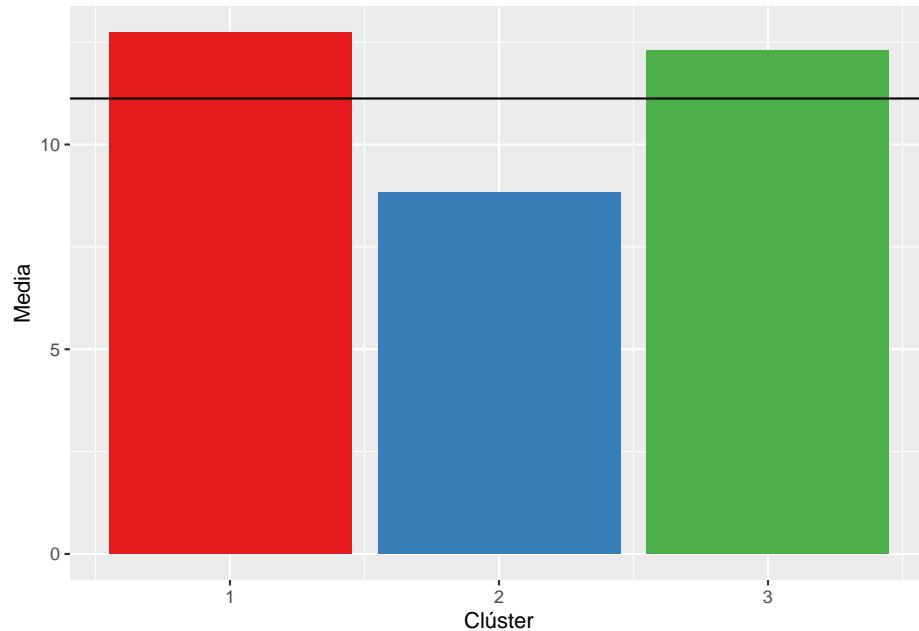
Cuadro 28: Significancia de p-valores para variables categóricas:

	x
CODE_GENDER	1
NAME_INCOME_TYPE	1
NAME_EDUCATION_TYPE	1
NAME_FAMILY_STATUS	1
OCCUPATION_TYPE	1
ORGANIZATION_TYPE	1
REGION_RATING_CLIENT	1

Vemos como solo nos descarta 1 variable, pero gráficamente muy pocas aportan información que muestren diferencias grandes entre clúster.

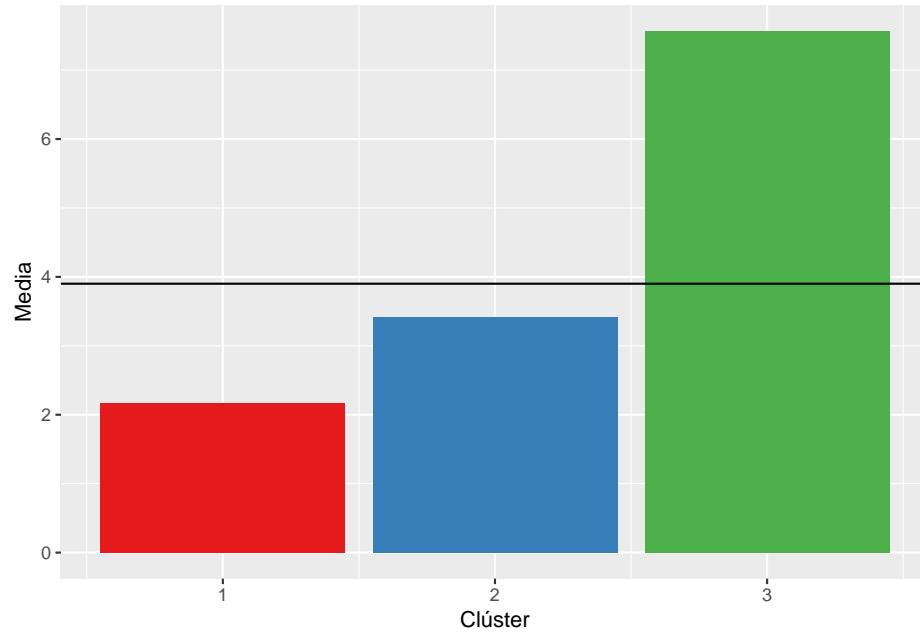
Se grafican las variables según clúster. Para las variables numéricas se mostrará la media grupal y la media global; para las variables categóricas se mostrarán las cantidades de cada nivel de la variable categórica por clúster.

Figura 86: Medias de la Edad en años del coche del cliente por clúster respecto la media global



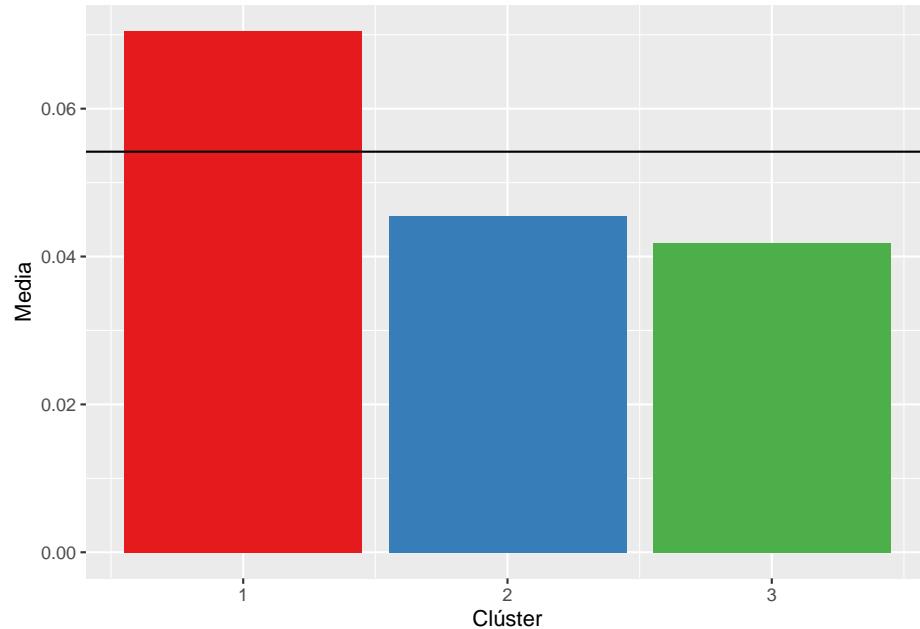
A partir del gráfico se puede observar como el clúster 2 es el que tiene los coches más nuevos.

Figura 87: Medias del Ratio del Importe del préstamo por clúster respecto la media global



Se ve como el clúster 1 es el que menos años tarda en devolver el préstamo, en concreto dos años. Por el contrario, el clúster 3 es el que más tarda en devolverlo, alrededor de 7 años.

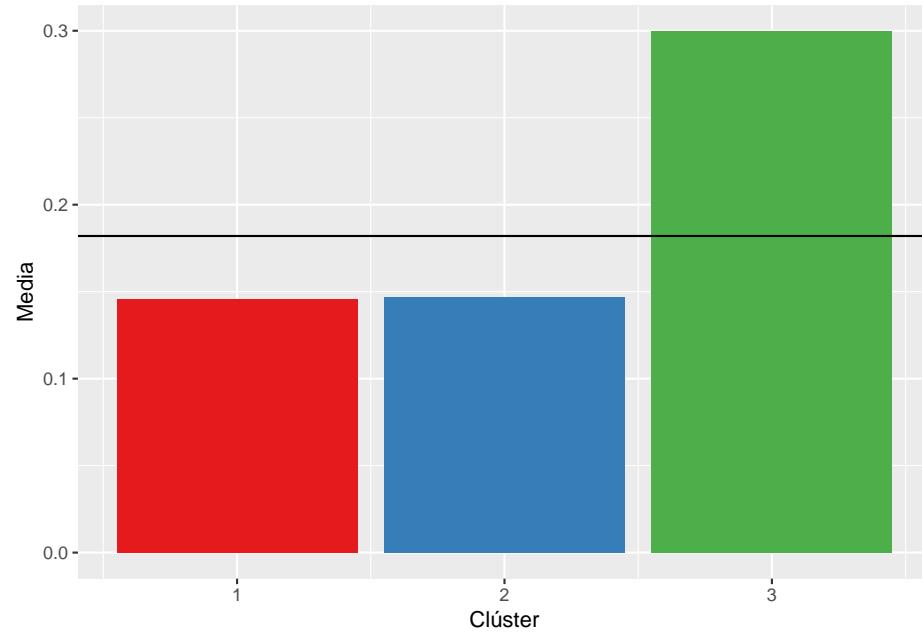
Figura 88: Medias del Ratio de la Anualidad del préstamo por clúster respecto la media global



Como se ha visto anteriormente, el clúster 1 es el que tiene una anualidad más alta, por otro lado el 3 es el que menos. Esto puede explicar el tiempo que se demoran en devolver el préstamo los individuos de cada

clúster.

Figura 89: Medias de la Capacidad de cliente para pagar la annuity con sus ingresos por clúster respecto la media global



Vemos como el clúster 3 tiene menos capacidad para pagar el préstamo.

Conclusiones

Clúster 1 se distingue por tener la anualidad más elevada y una menor demora en la devolución del préstamo.

Clúster 2 se caracteriza por incluir individuos con los coches más recientes.

Clúster 3 presenta una tendencia a tardar más en devolver el préstamo y exhibe una menor capacidad para hacerlo.

Es relevante notar que los patrones observados en los clústeres según las variables categóricas siguen la misma dinámica descrita. Esto podría deberse, en gran medida, a que el método k-means se centra exclusivamente en datos numéricos. Los gráficos asociados, aunque incluidos en el anexo, carecen de interpretación informativa directa sobre las características específicas de cada clúster.

Profiling Jerárquico

Con el objetivo de perfilar los grupos conseguidos mediante el algoritmo Jerárquico primero veremos la significación de las variables para los grupos y después se graficarán para identificar las características definitorias de cada grupo.

A continuación se muestran los p-valores para evaluar la significación de cada variable. Primeramente de las variables categóricas y seguidamente las numéricas.

Cuadro 29: Significación de las categóricas

Variable	P_Value
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	0
ORGANIZATION_TYPE	0
REGION_RATING_CLIENT	0

Cuadro 30: Significación de las numéricas

OWN_CAR AGE	c(Cluster = 3.29268185193439e-09)
CNT_FAM_MEMBERS	c(Cluster = 0.324319480327695)
log_AMT_INCOME_TOTAL	c(Cluster = 3.15217179279226e-59)
log_AMT_CREDIT	c(Cluster = 0.0242641506823843)
AGE_YEARS	c(Cluster = 1.35532523003757e-214)
RATIO_CREDIT_INCOME	c(Cluster = 8.70167189360433e-25)
RATIO_ANNUITY_CREDIT	c(Cluster = 2.4213002726531e-05)
DTI_RATIO	c(Cluster = 6.138172493865e-22)

Elaboramos una tabla donde se indica con 1 si se considera variable significativa para el clúster y 0 en caso contrario.

Cuadro 31: Significancia de p-valores para variables numéricas:

	x
OWN_CAR AGE	1
CNT_FAM_MEMBERS	0
log_AMT_INCOME_TOTAL	1
log_AMT_CREDIT	1
AGE_YEARS	1
RATIO_CREDIT_INCOME	1
RATIO_ANNUITY_CREDIT	1
DTI_RATIO	1

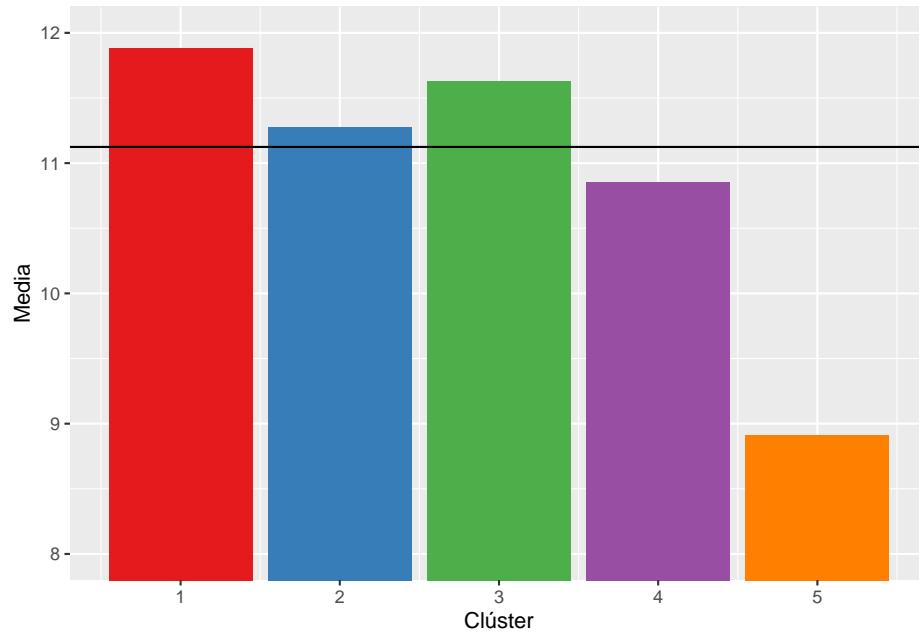
Cuadro 32: Significancia de p-valores para variables categóricas:

	x
CODE_GENDER	1
NAME_INCOME_TYPE	1
NAME_EDUCATION_TYPE	1
NAME_FAMILY_STATUS	1
OCCUPATION_TYPE	1
ORGANIZATION_TYPE	1
REGION_RATING_CLIENT	1

Podemos observar que solo se descarta una variable, pero al analizar gráficamente, podremos identificar qué variables son las que realmente aportan información significativa.

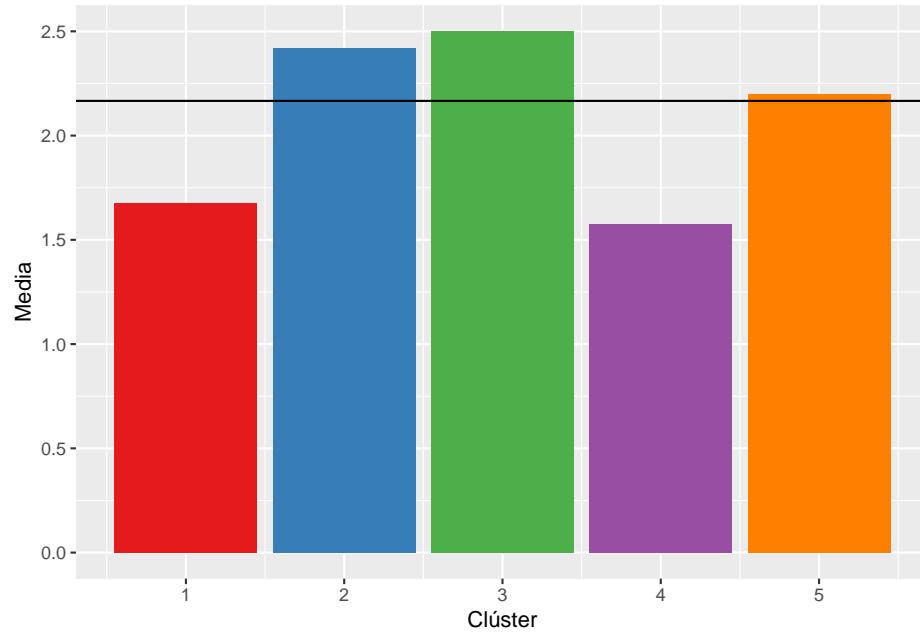
Se grafican las variables según clúster. Para las variables numéricas se mostrará la media grupal y la media global; para las variables categóricas se mostrarán las cantidades de cada nivel de la variable categórica por clúster.

Figura 90: Medias de la Edad en años del coche del cliente por clúster respecto la media global



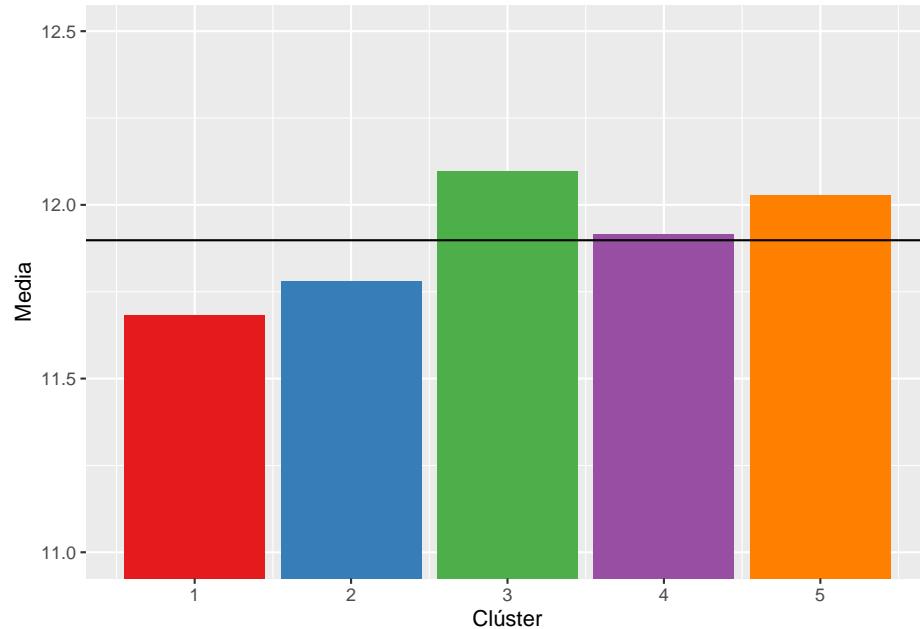
En lo que a la edad del coche para cada cliente respecta, vemos como el clúster 5 se caracteriza por tener una media de edad de coche mucho menor que los otros clústeres. Tiene los coches más nuevos, es decir, con menos años.

Figura 91: Medias del Número de familiares del cliente por clúster respecto la media global



Los clústeres 1 y 4 se caracterizan por tener el menor número de familiares. Por otro lado, el 2 y 3 tienen el mayor número de familiares.

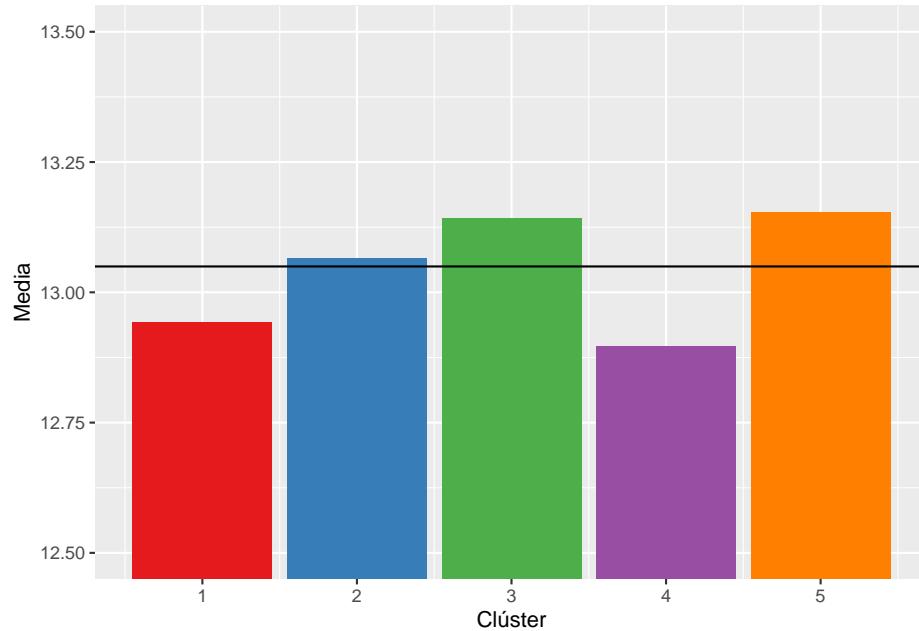
Figura 92: Medias del logaritmo de los Ingresos totales del cliente por clúster respecto la media global



A partir del gráfico, se observa que en los ingresos totales el clúster 1 se caracteriza por tener el menor número de ingresos y el clúster 3 el que mayor los tiene. No obstante, no hay diferencias muy grandes entre

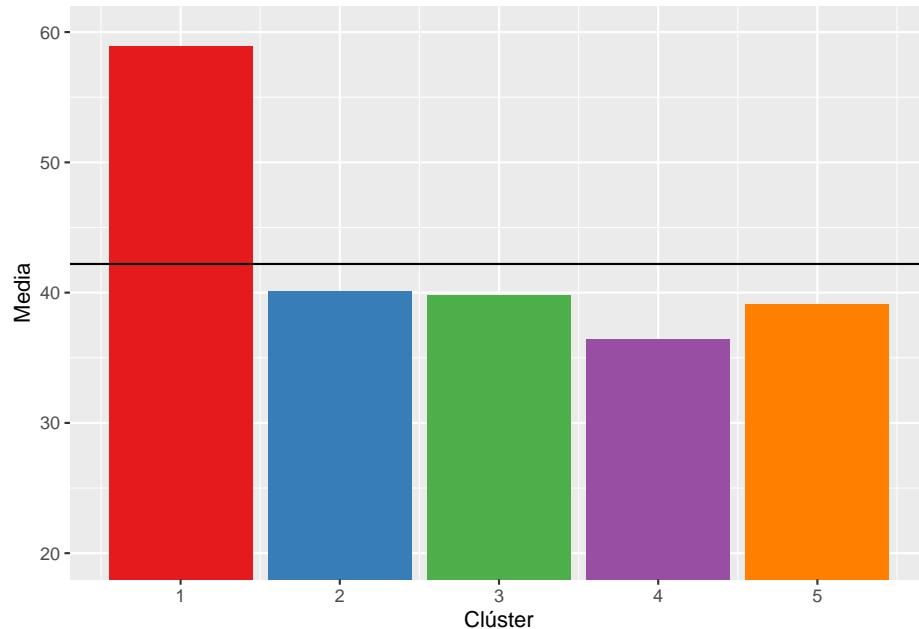
ellos debido a que es el logaritmo de estos ingresos y la interpretación queda afectada.

Figura 93: Medias del Importe de crédito del préstamo por clúster respecto la media global



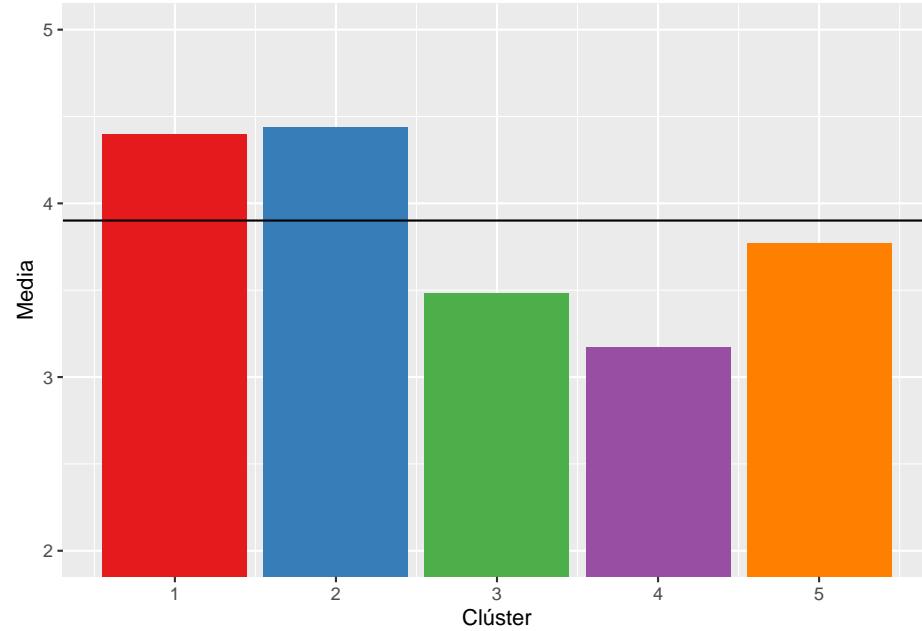
En el importe de crédito por préstamo se aprecia como el clúster 4 y 1 son los que más se diferencian, teniendo un importe menor. En cambio, el clúster 3 y 4 tienen un importe mayor que el resto.

Figura 94: Medias de la Edad por clúster respecto la media global



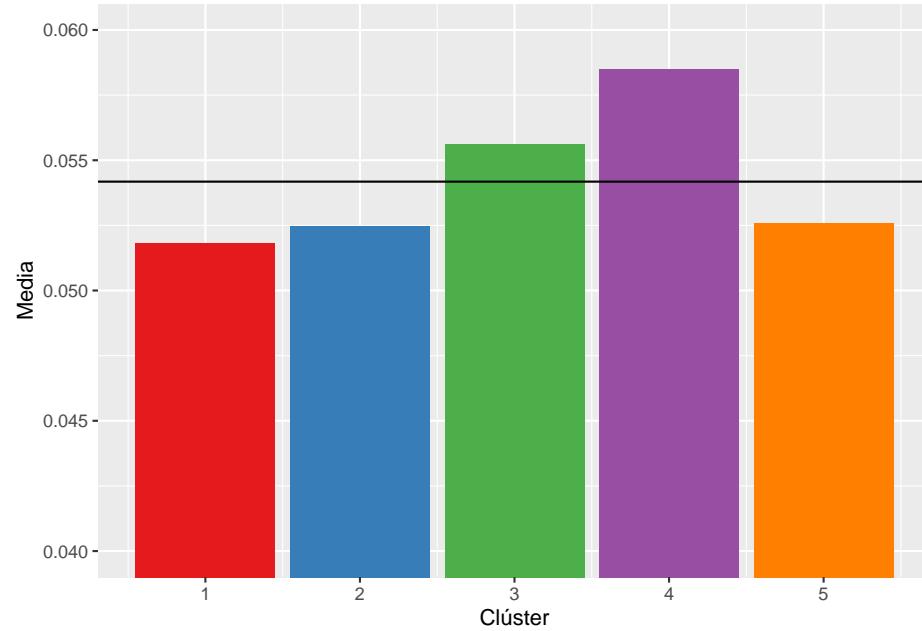
El clúster 1 se caracteriza por ser el grupo con individuos de más edad, en otras palabras, es el clúster con los individuos más mayores.

Figura 95: Medias del Ratio del Importe del préstamo por clúster respecto la media global



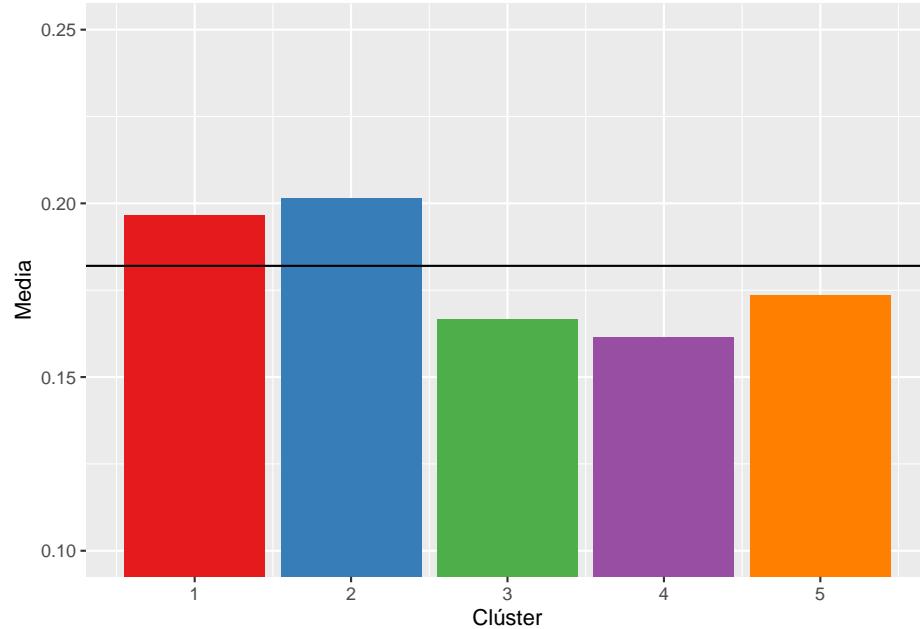
En el número de años que se tarda en devolver el crédito, se aprecia como los más rápidos son los sujetos del clúster 4. Contrariamente, el grupo 1 y 2 son los que más se demoran.

Figura 96: Medias del Ratio de la Anualidad del préstamo por clúster respecto la media global



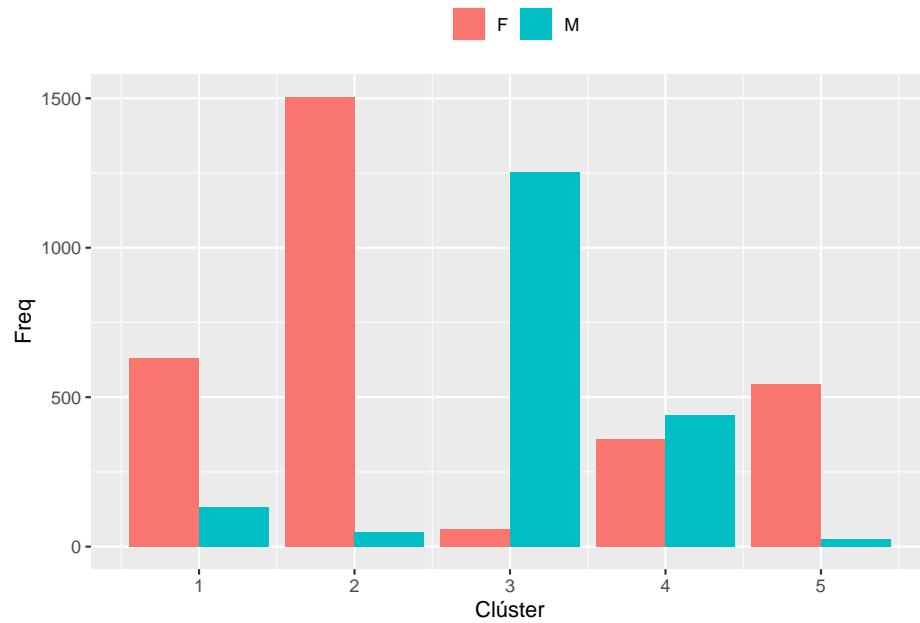
El clúster que se diferencia es el 4, con una Ratio entre la anualidad del préstamo y el crédito total solicitado.

Figura 97: Medias de la Capacidad de cliente para pagar la annuity con sus ingresos por clúster respecto la media global



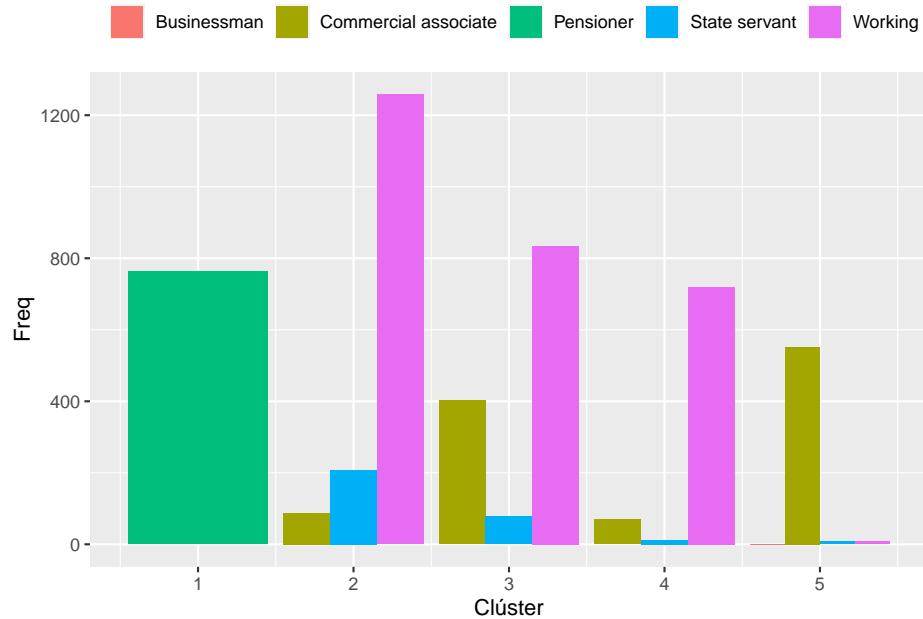
La ratio mide la capacidad del cliente para pagar la anualidad de su préstamo en relación con sus ingresos. Por ende, los menos capaces son los individuos del clúster 2.

Figura 98: Gráfico de la distribución del Género respecto el Clúster



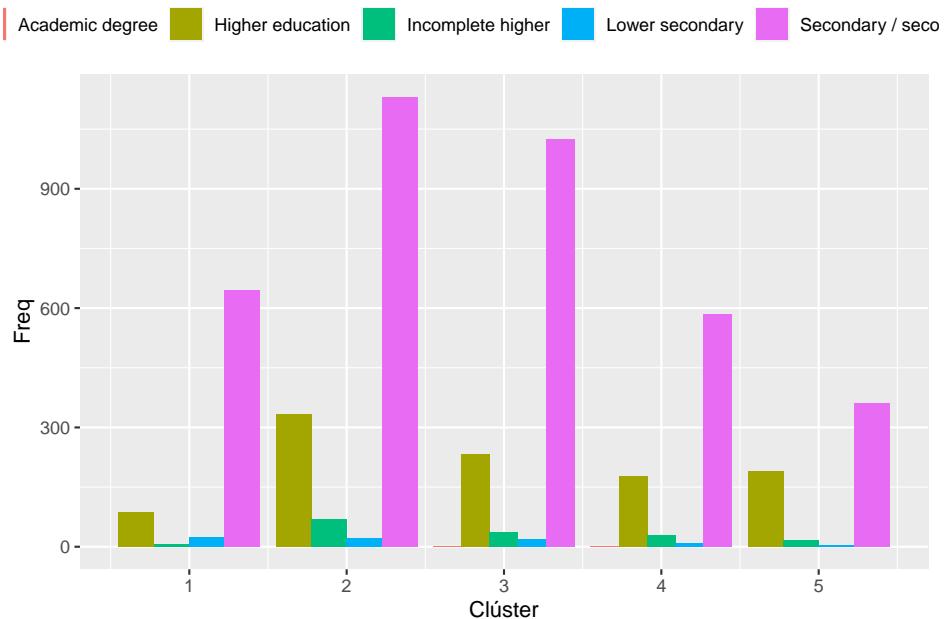
En la distribución del género de los individuos según el clúster, se observa como los grupos 1, 2 y 5 están formados por mujeres, el 3 por hombres y el 4 por hombres y mujeres a partes iguales.

Figura 99: Gráfico de la distribución del Tipo de ingresos respecto el Clúster



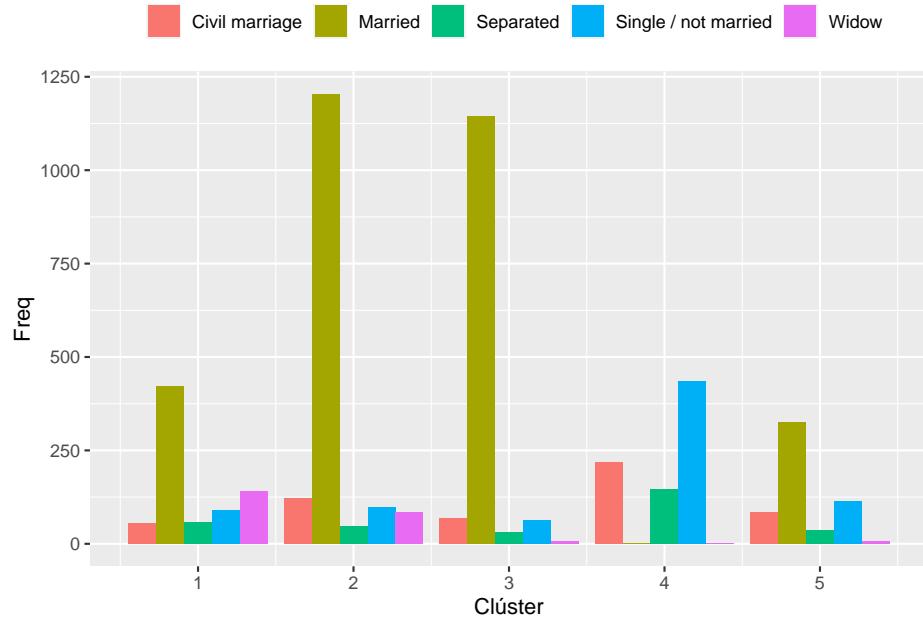
El primer grupo es el de los pensionistas y el quinto es el de los comerciantes asociantes.

Figura 100: Gráfico de la distribución del Nivel de estudios del cliente respecto el Clúster



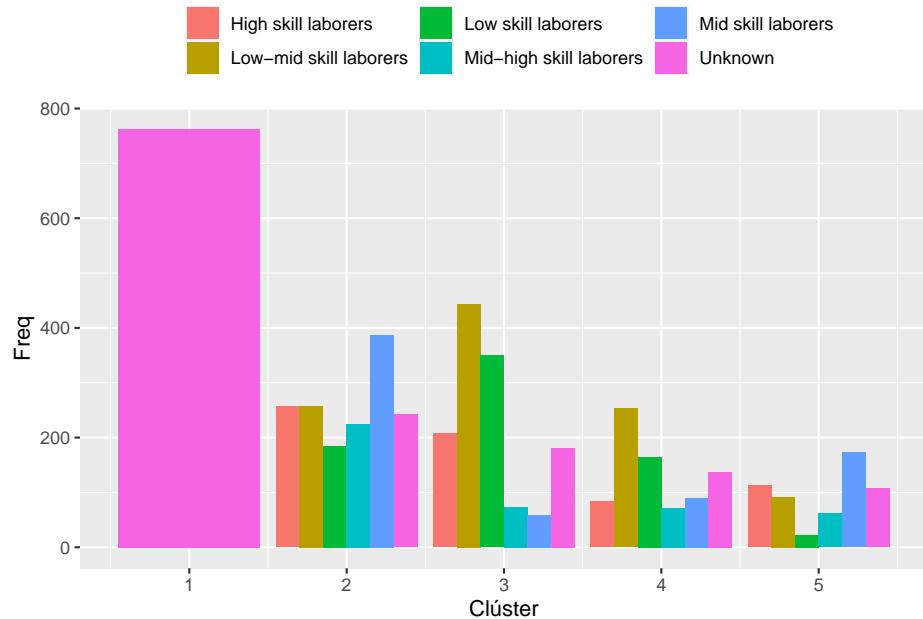
No hay diferencias con la distribución.

Figura 101: Gráfico de la distribución del Estado civil respecto el Clúster



En el Clúster 4 se incluye a personas que no están casadas o que tienen una unión civil.

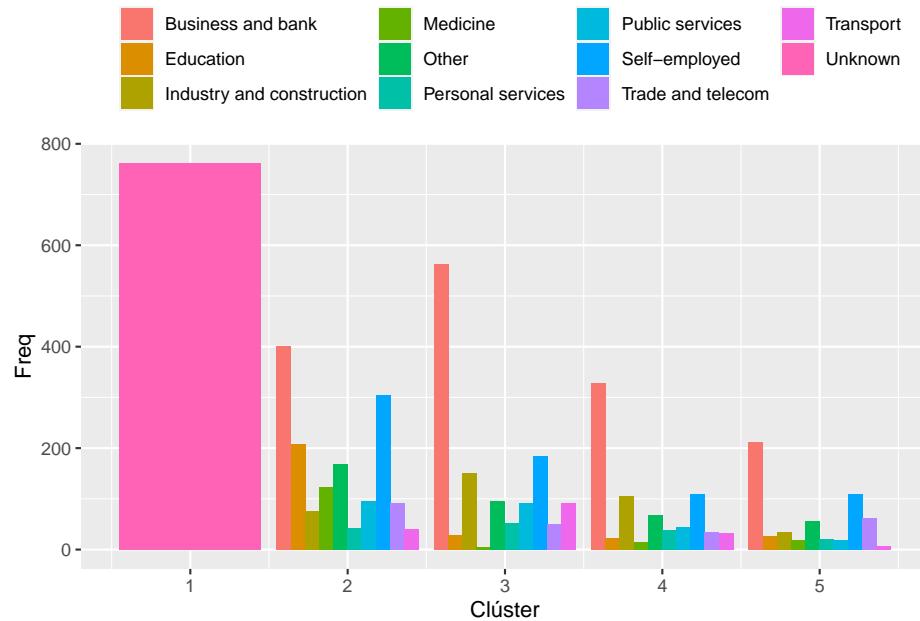
Figura 102: Gráfico de la distribución de la Actividad laboral respecto el Clúster



El grupo uno es el de los individuos que no trabajan, hecho que coincide con que también sea el grupo de

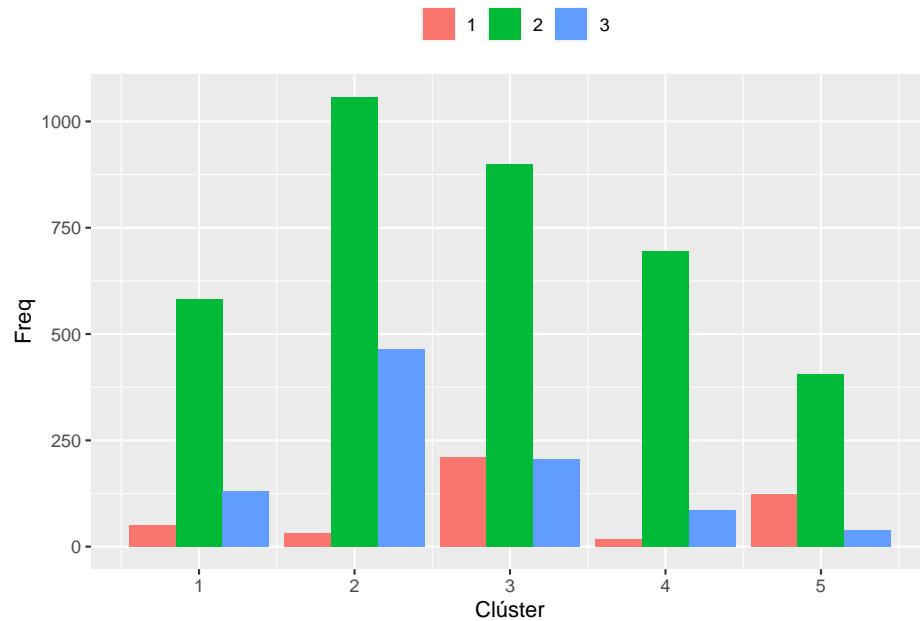
los pensionistas. Por otra parte, el grupo dos tiene más proporción de “mid skill workers”.

Figura 103: Gráfico de la distribución del Tipo de organización donde trabaja el cliente respecto el Clúster



Respecto al grupo uno no se puede saber a que tipo de organización pertenecen porque no trabajan, son los pensionistas. El resto sigue una distribución muy parecida aunque el grupo dos parece estar más dedicado a la educación que los otros.

Figura 104: Gráfico de la distribución de la Calificación de la región donde vive el cliente respecto el Clúster



No hay grandes diferencias en la distribución de la clasificación por región.

Conclusiones:

Clúster 1: Se caracteriza por individuos con pocos familiares y menor número de ingresos. Es el grupo de las mujeres de mayor edad pensionistas, las cuales tardan más en devolver el préstamo.

Clúster 2: Son mujeres, con una alta cantidad de familiares y mayor cantidad de importe del préstamo. Son los que peor capacidad de devolver el préstamo tienen, es decir, los que más se demoran en devolverlo. Proporcionalmente cuentan con más “mid skill workers”.

Clúster 3: Es el grupo de los hombres con mayor número de ingresos y con más cantidad de familiares.

Clúster 4: En este clúster están los individuos con menor importe de crédito por préstamo, ratio de anualidad más grande además de ser los más rápidos en devolver el crédito. Se caracteriza por estar compuesto en la misma proporción tanto de hombres como de mujeres, con pocos familiares. Los individuos están solteros o casados civilmente.

Clúster 5: Este último grupo está formado por las mujeres con mayor número de crédito por préstamo. Caracterizadas por tener coches más nuevos y ser “commercial associates”.

Comparación Profiling K-means y Jerárquico

Se destaca una diferencia en el número de clústers entre ambos métodos, con 3 clústers para K-means y 5 para el clústering jerárquico.

En el enfoque de clustering jerárquico, se logra obtener perfiles altamente específicos y fácilmente distinguibles para cada grupo, en contraste con los perfiles obtenidos a través de la metodología k-means.

En la metodología k-means, se observa que la explicación de los grupos se limita exclusivamente a variables numéricas, sin considerar ninguna variable categórica. Este enfoque numérico puro resulta en perfiles menos detallados, ya que no refleja la variabilidad explicada por las variables categóricas. Este aspecto contribuye a que los perfiles generados por k-means sean menos distintivos y caracterizados en comparación con los obtenidos mediante el clustering jerárquico.

K-MODES

El algoritmo K-MODES fue diseñado para agrupar grandes conjuntos de datos categóricos, y tiene como objetivo obtener las k modas que representan al conjunto. Permite extender el k-means, a partir del cálculo de una medida de disimilitud que permita comparar observaciones categóricas y la utilización de modas en lugar de medias para calcular los clusters.

El primer paso será seleccionar k número de modas. Queremos realizar un análisis de agrupamiento utilizando el algoritmo de K-Mode con 5 clusters. Como ya hemos detectado en el clustering jerárquico, con k = 5 conseguimos un mejor corte y un perfilamiento de grupos más detallado.

La manera de medir la distancia entre dos vectores de variables categóricas es la cantidad de valores que son diferentes en la misma variable entre clusters.

Función para ejecutar K-MODES

Cuadro 33: Obtención de los Parámetros de los Clústers

CODE_GENDER	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS
M	Working	Secondary / secondary special	Married
F	Commercial associate	Secondary / secondary special	Married
F	Working	Higher education	Married
M	Working	Secondary / secondary special	Married
F	Pensioner	Secondary / secondary special	Married

OCCUPATION_TYPE	ORGANIZATION_TYPE	REGION_RATING_CLIENT	TARGET
Unknown	Business and bank	2	1
Mid skill laborers	Business and bank	2	1
Mid-high skill laborers	Education	2	0
Low-mid skill laborers	Business and bank	2	0
Unknown	Unknown	2	0

Con esta tabla podemos ver estos 5 clusters. El primer cluster, y por lo tanto la clase mayoritaria, corresponde a hombres casados, que trabajan en empresas o bancos y educación secundaria o secundaria especial. Pertenecen al grupo de clientes sin dificultad de pago.

El segundo cluster está formado por mujeres casadas trabajadoras como asociadas en empresas o bancos con habilidades medianas y educación secundaria o secundaria especial. Pertenecen al grupo de clientes sin dificultad de pago.

En el tercer cluster tenemos a mujeres casadas, que trabajan en educación con habilidades medias o altas y educación alta. Pertenecen al grupo de clientes con dificultad de pago.

El cuarto está compuesto por hombres casados, que trabajan en empresas o bancos con habilidades bajas o medias y educación secundaria o secundaria especial. Pertenecen al grupo de clientes con dificultad de pago.

Y finalmente, el quinto y último cluster formado por mujeres casadas y pensionistas que han tenido educación secundaria o secundaria especial. Pertenecen al grupo de clientes con dificultad de pago.

Para poder calcular las distancias entre el primer y segundo cluster miramos la separación que existe usando las diferencias entre la primera fila y la segunda. Obtiene un valor de 3, ya que ni code_gender ni name_income_type ni occupation_type coinciden. Las dos variables que coincide en los cinco clusters son name_family_status y region_rating_client.

DBSCAN

El algoritmo DBSCAN es un método de clústering basado en densidad de aplicaciones con ruido. Este método permite agrupar los datos cuando estos presentan formas complejas, así como es un método robusto frente a la presencia de outliers. Para realizar el algoritmo DBSCAN se emplean sólo las variables numéricas normalizadas.

DBSCAN parte de dos parámetros que son: - Épsilon: distancia máxima a la que debe haber otra observación para ser considerar que cumple con el criterio de “estar cerca” - Mínimo de puntos: parámetro que controla la densidad mínima requerida para que un punto sea considerado un núcleo y se incluya en un grupo/clúster.

Cálculo de mínimo de puntos

Para calcularlo de manera empírica, diremos que el mínimo de puntos sea igual al 0.2% - 0.25% del total de los datos teniendo en cuenta que:

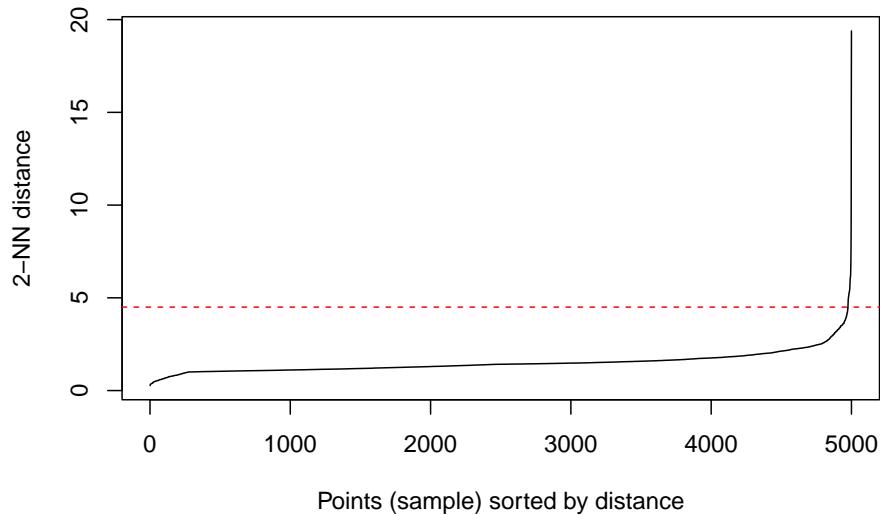
- El mínimo será de 2 para datos que sean muy pequeños
- El máximo será de 10 para datos con mucha información

Aplicando esto, se tiene que el número mínimo de puntos sería 11, pero lo limitamos a 10 en concordancia con la literatura.

Cálculo de épsilon

Se escogerá épsilon a partir del siguiente gráfico del codo, realizado con el método del k-NN. Como se han realizado otros métodos de clústering, se aplica el k-NN con el valor de K sacado de los métodos de clúster jerárquico, que concluyen que el número de clústers k óptimo es 2. Estas k-distancias se trazan en orden ascendente con el objetivo es determinar la “codo”, que corresponde al parámetro épsilon óptimo. A partir del siguiente gráfico del codo se puede observar el valor óptimo de épsilon.

Figura 105: Gráfico del Codo para el Valor Óptimo de Épsilon



El valor de épsilon se decide a partir de el corte en el máximo cambio de la pendiente. En el gráfico se observa que esto se da alrededor de épsilon = 4.5.

Resultado DBSCAN

Aplicamos el método DBSCAN con los valores extraídos: épsilon=4.5 y mínimo de puntos de 10.

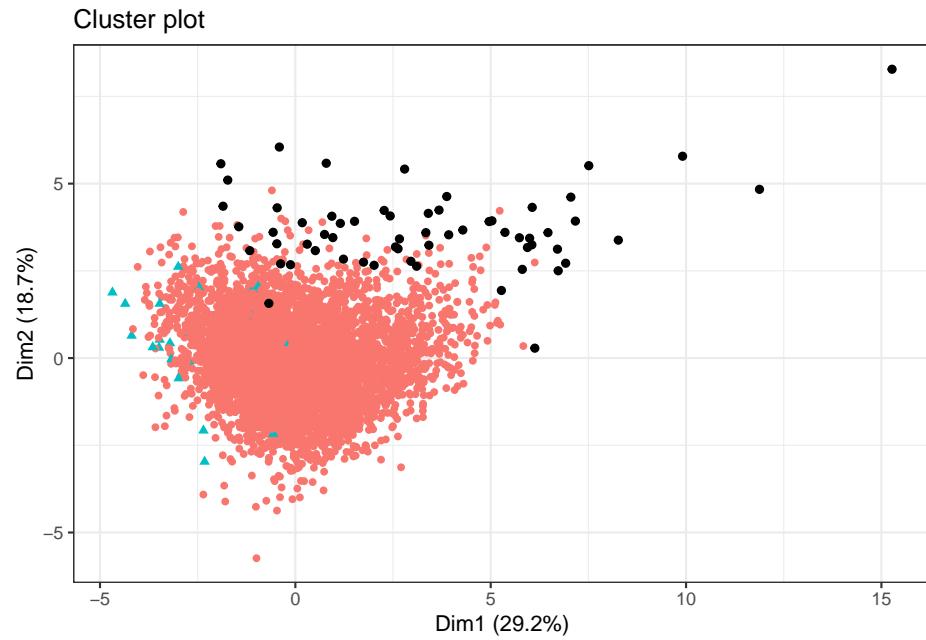
Cuadro 34: Resumen del número de puntos en cada clúster

Clúster	Frecuencia de puntos
0	61
1	4915
2	24

EL resultado del DBSCAN indica que agrupa los datos en 2 clústers, y devuelve 61 puntos como outliers.

Se presenta el gráfico de los clústers obtenidos con el DBSCAN:

Figura 106: Gráfico Clústers obtenidos con DBSCAN



Con estos resultados, ya se aprecia que el DBSCAN no realiza agrupaciones óptimas en estos datos, ya que la inmensa mayoría de datos se ubican en el primer clúster, y el segundo clúster contiene una proporción de datos ínfima. Esto puede ser debido a que el DBSCAN es un método que parte de las densidades, y en los datos que se agrupan con formas más simples y uniformes, como los que se tratan en este trabajo, puede no encontrar la solución óptima, como se considera en este caso. A esta misma conclusión se llegará también con el método OPTICS, ya que también es un método de clustering basado en densidades.

OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure), es otro algoritmo de clustering utilizado en minería de datos y análisis de datos para descubrir patrones y estructuras en conjuntos de datos; siendo su objetivo principal descubrir grupos de puntos que están densamente agrupados en el espacio de características. Fue propuesto como una mejora del algoritmo DBSCAN, dado que este tiene problemas con las fronteras.

El algoritmo OPTICS comienza identificando los puntos centrales (core points) en el conjunto de datos (llamado `minPts`) dentro de un radio específico (llamado `eps`). Dado que una de sus limitaciones es la elección adecuada de estos parámetros, ya que son cruciales para obtener resultados óptimos, a continuación se optimiza su búsqueda:

Búsqueda de los parámetros óptimos

Optimización de la búsqueda de parámetros para `épsilon` y `minPts` en Optics:

Primeramente, definimos los valores que se van a probar para `eps` y `minPts`, creando una cuadrícula de parámetros y, seguidamente, se establece el número de núcleos (cores) a utilizar para la optimización en paralelo, que se calcula automáticamente.

Función para ejecutar OPTICS con una combinación de parámetros y calcular el coeficiente de silueta:

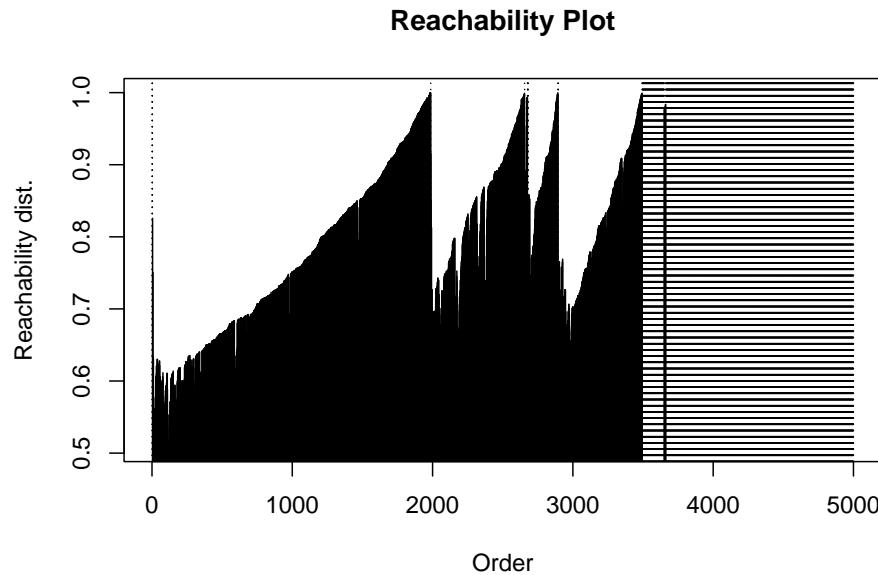
Cuadro 35: Obtención de los Parámetros Óptimos

	eps	minPts
	20	1

Como vemos, después del proceso iterativo, la combinación de resultados más óptima ha sido un radio (`eps`) de 1 con un mínimo de 10 puntos (`minPts`).

Así mismo, a continuación creamos el modelo OPTICS con los parámetros óptimos encontrados y observamos su reachability plot:

Figura 107: Reachability Plot



El gráfico de reachability (alcance) que acabamos de generar, es una herramienta para visualizar la estructura de clústeres identificados.

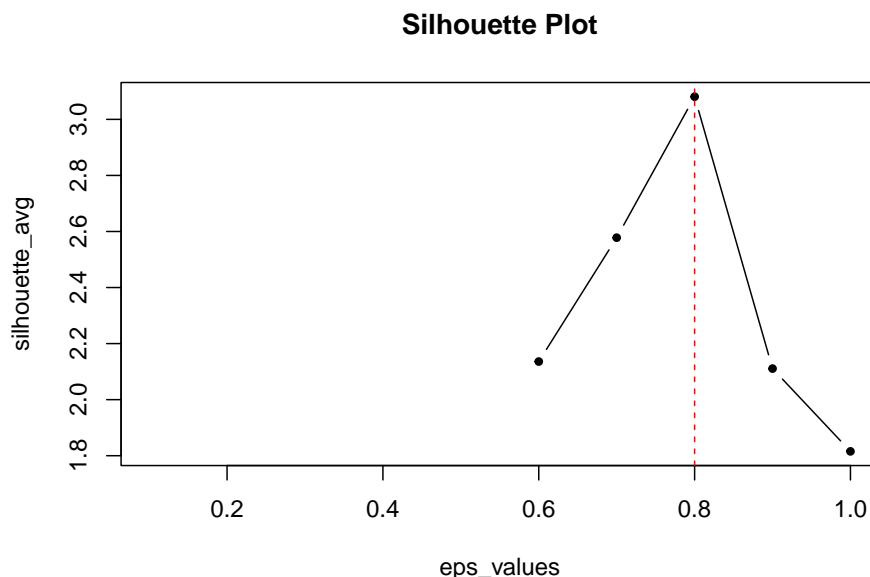
En los gráficos de reachability, cada punto representa un objeto de datos y la altura de la curva indica la distancia a la que se encuentra el objeto más cercano dentro del mismo clúster. Los valles en la curva indican la presencia de clústeres, ya que los puntos dentro de un mismo clúster tienden a estar más cerca entre sí que con puntos de otros clústeres.

Así pues, como se puede observar, a primera vista vemos como apriori podríamos clasificar nuestra base de datos en tres clústeres. Aun así, hay una gran parte de nuestros datos que no está bien representada (la parte derecha del gráfico).

Método de la silueta

Otra forma de encontrar los valores óptimos de los parámetros necesarios es a través del método de la silueta. En esta sección, se ejecutará OPTICS con diferentes valores de `eps` y se calculará la medida de silueta para cada valor. Luego, se graficará esta medida en función de ϵ y se identificará su valor óptimo.

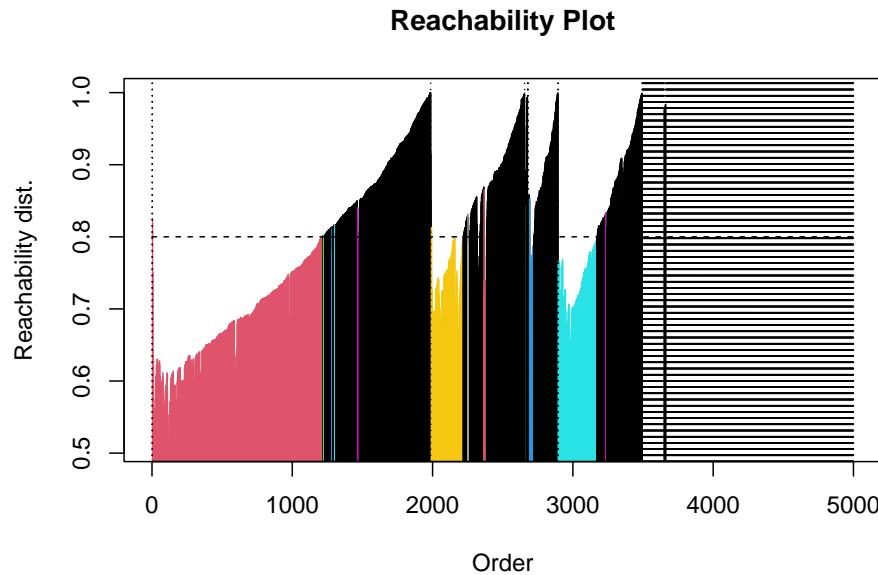
Figura 108: Gráfico del método de la silueta



Como se puede apreciar, al agregar una línea vertical en el valor óptimo de ϵ , vemos que se aconseja cortar en 0.8, valor que maximiza la silueta.

Por último, entramos en la etapa posterior al cálculo de la estructura de clústeres utilizando OPTICS. Esta última etapa, consiste en extraer y visualizar los resultados del clustering, donde a partir de la variable `opt_eps`, se determinará como se corta la curva de alcance para identificar los clústeres (diferenciados por colores).

Figura 109: Reachability Plot

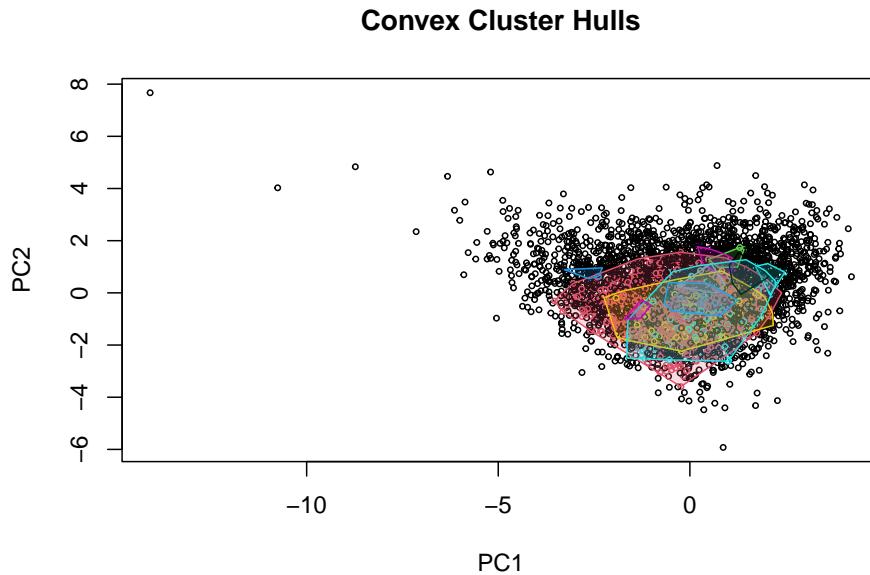


Con `plot(res)` se genera un gráfico que visualiza los clústeres obtenidos. Los puntos de datos se colorean de acuerdo con los clústeres a los que pertenecen, y los puntos que se consideran ruido se muestran en negro.

De igual manera que pasaba en el reachability plot anterior, hay una gran parte de nuestros datos que no sale bien representada. Además, al colorear los diferentes clústeres por colores, vemos que hay una gran parte de nuestros datos que se consideran ruido. Por otro lado, contrariamente a los resultados del primer reachability plot, en este se puede apreciar como nuestros datos podrían estar clasificados entre más grupos. No obstante, la presencia de tanto ruido y la parte no explicada nos podría estar informando de que nuestra base de datos no es adecuada para técnicas de clústring basadas en densidades.

Finalmente, visualizamos el gráfico con los grupos creados en forma de polígonos convexos. Estos polígonos nos ayudan a delimitar visualmente la extensión de cada clúster.

Figura 110: Polígonos de Convexidad para los Clústeres Identificados



Por un lado, el gráfico obtenido no nos permite extraer buenos resultados, dado que es una gran nube de puntos en donde la mayoría de los polígonos convexas se superponen entre sí.

Cuadro 36: Resumen del número de puntos en cada clúster

Clúster	Frecuencia de puntos
0	3153
1	1215
2	9
3	7
4	8
5	13
6	227
7	14
8	21
9	18
10	1
11	30
12	276
13	8

Por otro lado, la tabla obtenida nos resume el número de puntos en cada clúster. Así pues, observamos como aunque nos divide los datos en 13 clústeres, siendo el grupo 0 el dominante, indicando la presencia de 3153 outliers. En los 12 grupos siguientes, la mayoría de las observaciones se agrupan mayoritariamente en el primero, con 1215, seguidos por el clúster 6 con 227 y el 12 con 276. Los grupos restantes tienen muy pocas observaciones en cada uno de ellos.

Estos resultados nos indican que para nuestra base de datos, este tipo de clustering no es el más adecuado.

Conclusión

En conclusión, aunque las técnicas utilizadas nos hayan ayudado a encontrar unos buenos parámetros para poder agrupar nuestros datos de la manera más óptima, vemos como estos resultados nos ayudan a respaldar aún más el hecho de que nuestra base de datos no es válida para técnicas de clustering basados en densidad, posiblemente por no tener una distribución de densidad variable.

Las técnicas de clustering basadas en densidad asumen que los clústeres se forman en regiones de alta densidad de datos. Por lo tanto, si nuestros datos no tienen una distribución de densidad variable (puntos uniformemente distribuidos o clústeres sin una densidad significativamente mayor que el fondo), las técnicas de clustering basadas en densidad pueden no ser efectivas.

Así pues, ni DBSCAN ni OPTICS nos permiten extraer un buen análisis de nuestra base de datos, hay que recurrir, por ejemplo, a clustering jerárquico.

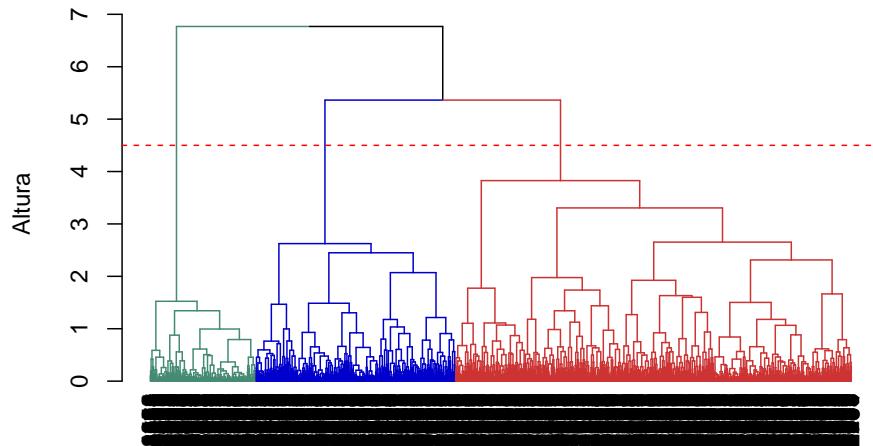
Algoritmo CURIE

Siguiendo con los algoritmos de clusterización para bases de datos grandes, es momento de realizar el CURIE. CURIE (Clustering Using REpresentatives) es un algoritmo de clustering para base de datos grandes en el cual se gestiona, inicialmente, una muestra de la base de datos a partir de la cual se realiza un clustering jerárquico (usando la distancia euclídea y el método de agregación simple) y se sacan un número pequeño de puntos (representantes) de cada cluster. Entonces, se acercan esos representantes hacia el centroide del cluster un 20 % y, a partir de estos representantes acercados, se busca cuál es el que se encuentra más cercano de cada punto de la base de datos restante. Finalmente, una vez se encuentra el representante más cercano a cada individuo, se asigna el individuo al cluster al que pertenece el representante.

En este caso, como la base de datos escogida dispone de datos numéricos y categóricos, se ha decidido modificar las reglas del CURIE y usar la distancia de Gower y el método de agregación de Ward en la construcción inicial del clustering. Así pues, realmente se podría afirmar que se está realizando un pseudoCURIE en este caso.

Inicialmente, para este caso, se ha decidido escoger una muestra significativa y grande para evitar problemas en la construcción de los clusters iniciales. Así, se ha usado una muestra de $n = 2000$ con el objetivo de realizar el primer cluster a partir del cual se elegirán los representantes. El dendograma resultante reporta la siguiente estructura:

Figura 111: Dendrograma inicial CURIE



Tras analizar los resultados, se puede apreciar que el número de clusters óptimo es $k = 3$. De esta forma, la partición inicial de la muestra en cada cluster se puede apreciar en la parte inferior:

Cuadro 37: Distribución inicial de individuos por cluster CURIE

Cluster	Observaciones
1	569
2	1129
3	302

Ahora, a partir de este clustering jerárquico inicial, se escogerán los representantes. Para ello, se busca aquellos puntos más alejados entre sí y, a la vez, más alejados del centroide de cada cluster. Para este paso, se han elegido exactamente 5 representantes por cluster. Una vez se tienen seleccionados, el siguiente paso es acercarlos al centro. En este caso, se ha decidido aproximarlos un 20% hacia el centroide del cluster al que pertenecen.

Por último, se analiza cada punto y se busca el representante más cercano. Una vez se tiene esa información, se le asigna al individuo el cluster al que pertenece el representante más cercano. Para este paso, se ha procedido a procesar los datos de 500 en 500, para así evitar problemas con la capacidad de gestión de datos del ordenador. Así, el resultado del clustering final se presenta en la tabla inferior:

Cuadro 38: Distribución Final de individuos por cluster CURE

1	1971
2	1993
3	1036

Profiling del CURE

A partir del CURE resultante, se analizan las características que diferencia cada cluster encontrado para así identificar las diferencias significativas más relevantes entre grupos. Aquí se muestran los gráficos de las medias por grupo para cada cluster resultante:

Antes de empezar a analizar cada variable, es importante destacar qué variables son significativas para diferenciar clusters. Para ello, se realizará un test de Chi-cuadrado para las variables categóricas y, por otro lado, un test F para las variables numéricas, a través de una tabla ANOVA:

Significación de las variables categóricas

En la siguiente tabla se puede apreciar cada variable con su p-valor asociado a la prueba de Chi-cuadrado correspondiente:

Cuadro 39: P-valor asociado a cada variable categórica

Variable	P_Value
CODE_GENDER	0.0000000
NAME_INCOME_TYPE	0.0000000
NAME_EDUCATION_TYPE	0.0000000
NAME_FAMILY_STATUS	0.0000000
OCCUPATION_TYPE	0.0000000
ORGANIZATION_TYPE	0.0000000
REGION_RATING_CLIENT	0.0000000
TARGET	0.0002074

Como se puede apreciar, en este caso, todas las variables son significativas, es decir, existen diferencias entre al menos un par de clusters. De esta forma, todas las variables categóricas serán tenidas en cuenta para el perfilamiento de los clusters.

Significación de las variables numéricas

Seguidamente, se seguirá el mismo procedimiento para las variables numéricas. Esta vez, sin embargo, se usarán los test F resultantes de la tabla ANOVA:

Cuadro 40: P-valor asociado a cada variable numérica

OWN_CAR_AGE	0.0134960
CNT_FAM_MEMBERS	0.0000000
log_AMT_INCOME_TOTAL	0.0000000
log_AMT_CREDIT	0.0000000
AGE_YEARS	0.0000000
RATIO_CREDIT_INCOME	0.0291747
RATIO_ANNUITY_CREDIT	0.1195032
DTI_RATIO	0.0770792

Nuevamente, como se puede apreciar, todas las variables son significativas, es decir, existen diferencias entre al menos un par de clusters. De esta forma, todas las variables categóricas serán tenidas en cuenta para el perfilamiento de los clusters.

Análisis del profiling

Una vez ya hemos presenciado que todas las variables se usarán en el proceso de profiling de cada cluster, se ha procedido a realizar gráficos para cada variable, para así ver las características de cada grupo. Todos los gráficos realizados para el análisis de cada cluster se hallan en los anexos finales del trabajo. Así pues, tras haber analizado atentamente el resultado ofrecido por el profiling realizado, se han extraído las siguientes conclusiones para cada cluster:

- Cluster 1: Este cluster está formado por la gente con un ratio credit/income menor, además de un DTI ratio más bajo que los otros dos clusters encontrados. Esto nos indica que son personas más responsables con sus cuentas y que piden crédito cuando su situación financiera es positiva, ya que tienden a endeudarse menos. Además, apreciando el análisis de las variables categóricas, se aprecia que este cluster presenta una mayor concentración de hombres (solteros en su mayoría) y dedicados principalmente al sector de la banca, en su mayoría como comerciales de este mismo sector. De esta forma, es lógico pensar que hagan una buena gestión de sus finanzas personales.
- Cluster 2: Este grupo se caracteriza principalmente por tener un ratio credit/income más alto y un ratio annuity/credit menor. Es decir, es gente que pide préstamos por una cantidad elevada en relación a sus ingresos, pero que generalmente deciden pagarlos a largo plazo. Esto hecho, además, va relacionado con que la edad del coche media sea la menor entre los tres grupos: tal vez una parte del préstamo solicitado se ha destinado al coche. Entrando en el análisis de las variables categóricas, se aprecia que en su mayoría son mujeres que ocupan trabajos de gran capital humano (state servant) y residen en lugares con un buen rating por la empresa.
- Cluster 3: Por último, en este cluster se sitúan aquellos ciudadanos con una edad superior, en su gran mayoría pensionistas. Además, poseen coches con más años que el resto y presentan núcleos familiares más reducidos. En este cluster se encuentran la gran mayoría de personas viudas y, en general, el nivel educativo que más predomina es la secundaria.

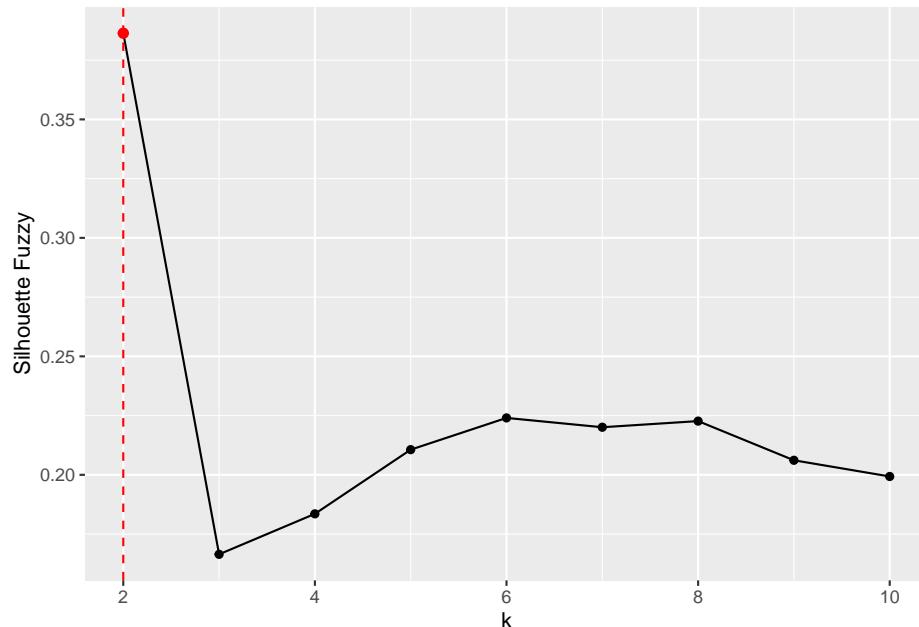
Fuzzy Clustering

La característica distintiva del Fuzzy Clustering en comparación con otros algoritmos radica en su capacidad para permitir que una observación pertenezca a más de una agrupación. En otras palabras, posibilita que los elementos o individuos tengan grados de pertenencia a varios grupos simultáneamente.

En el contexto de la ejecución del algoritmo, la función FKM() se encarga de realizar el Fuzzy Clustering para las K especificadas y, de manera automática, selecciona el valor óptimo de k.

La propia función FKM guarda los valores Silhouette fuzzy para cada k. Este es un índice específico del Fuzzy y cuanto mayor sea el índice del Silhouette mejor. Se grafican los valores SIL.F para cada k:

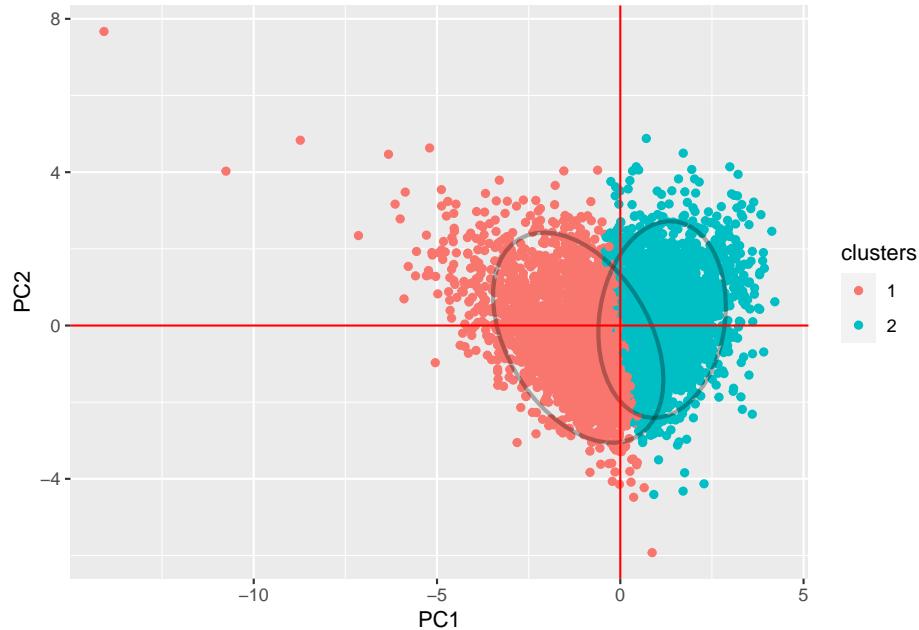
Figura 112: Índice de Silhouette Fuzzy



Como se puede ver en el gráfico, k=2 es la k óptima para el algoritmo fuzzy.

Acto seguido, visualizamos los clústeres en el gráfico factorial de las dos primeras dimensiones:

Figura 113: Representación de las clases en PC1-PC2



Predicción de un individuo aleatorio en el fuzzy

En esta sección, creamos un individuo hipotético y aplicamos el algoritmo Fuzzy C-means con el propósito de determinar a qué clúster se asignaría y con qué probabilidad. A continuación, se presenta la información detallada del individuo recién creado:

Individuo:

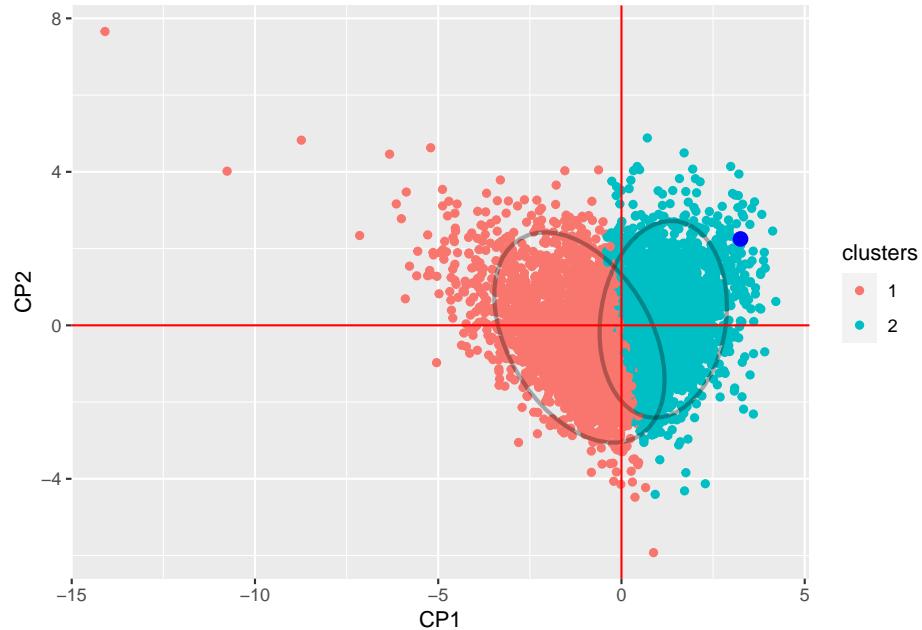
Sexo: Masculino Edad: 56 años Ingresos: \$100,000 Número de personas en la familia: 4 Años de posesión del automóvil: 7 Monto del crédito: \$12,000 Proporción de crédito: 0.12 Proporción de anualidad: 0.05 Proporción de deuda a ingresos: 0.15

Cuadro 41: Individuo generado

	5001
OWN_CAR_AGE	7.000000
CNT_FAM_MEMBERS	4.000000
log_AMT_INCOME_TOTAL	11.512925
log_AMT_CREDIT	9.392662
AGE_YEARS	56.000000
RATIO_CREDIT_INCOME	0.120000
RATIO_ANNUITY_CREDIT	0.050000
DTI_RATIO	0.150000

En este apartado, aplicamos el algoritmo Fuzzy C-means a la base de datos, incorporando el individuo recién añadido. Después de la ejecución del algoritmo, visualizamos los resultados mediante la representación gráfica en el plano factorial del PCA, resaltando el individuo en cuestión con un tono azul.

Figura 114: Representación de las clases en PC1-PC2 con el individuo añadido



```
##      Clus 1      Clus 2
## 0.4477844 0.5522156
```

Vemos como, a pesar de su ubicación distante de los puntos identificados en el Clúster 1, a este individuo se le asigna a este grupo debido a una alta probabilidad de pertenencia. Esto resalta la capacidad del algoritmo para reconocer patrones y adaptarse a la variabilidad de los datos.

Profiling Fuzzy

Con el objetivo de perfilar los grupos conseguidos mediante el algoritmo Fuzzy primero veremos la significación de las variables para los grupos y después se graficarán para identificar las características definitorias de cada grupo.

A continuación se muestran los p-valores para evaluar la significación de cada variable. Primeramente de las variables categóricas y seguidamente las numéricas.

Cuadro 42: Significación de las categóricas

Variable	P_Value
CODE_GENDER	0.0000000
NAME_INCOME_TYPE	0.0000000
NAME_EDUCATION_TYPE	0.0043287
NAME_FAMILY_STATUS	0.0000000
OCCUPATION_TYPE	0.0000134
ORGANIZATION_TYPE	0.0000000
REGION_RATING_CLIENT	0.0258849

Cuadro 43: Significación de las numéricas

OWN_CAR AGE	c(Cluster = 0.00165538345012159)
CNT_FAM_MEMBERS	c(Cluster = 0.866438547924501)
log_AMT_INCOME_TOTAL	c(Cluster = 0.826792734435659)
log_AMT_CREDIT	c(Cluster = 0)
AGE_YEARS	c(Cluster = 3.82997233028708e-48)
RATIO_CREDIT_INCOME	c(Cluster = 0)
RATIO_ANNUITY_CREDIT	c(Cluster = 0)
DTI_RATIO	c(Cluster = 1.88660547874547e-317)

Elaboramos una tabla donde se indica con 1 si se considera variable significativa para el clúster y 0 en caso contrario.

Cuadro 44: Significancia de p-valores para variables numéricas:

	x
OWN_CAR AGE	1
CNT_FAM_MEMBERS	0
log_AMT_INCOME_TOTAL	0
log_AMT_CREDIT	1
AGE_YEARS	1
RATIO_CREDIT_INCOME	1
RATIO_ANNUITY_CREDIT	1
DTI_RATIO	1

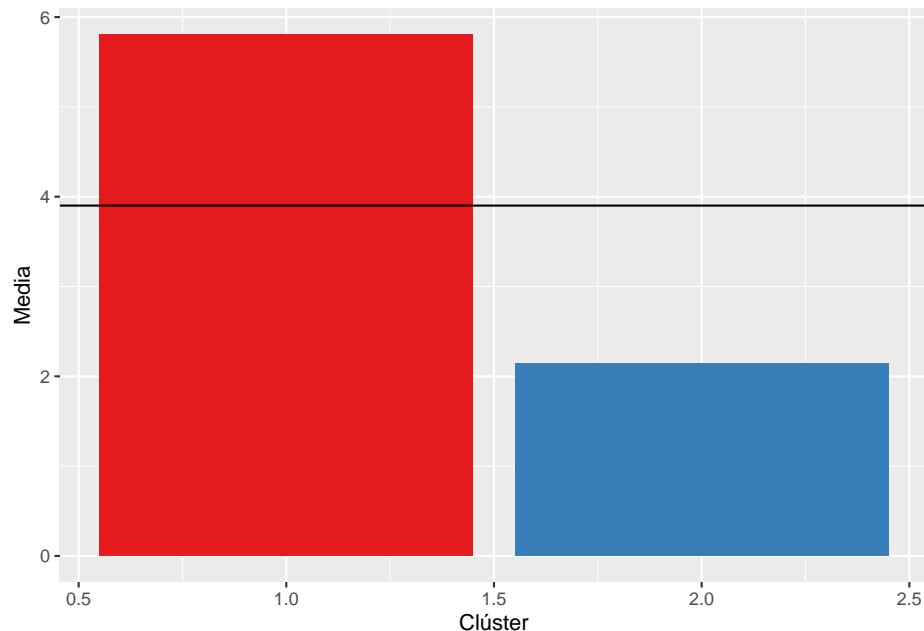
Cuadro 45: Significancia de p-valores para variables categóricas:

	x
CODE_GENDER	1
NAME_INCOME_TYPE	1
NAME_EDUCATION_TYPE	1
NAME_FAMILY_STATUS	1
OCCUPATION_TYPE	1
ORGANIZATION_TYPE	1
REGION_RATING_CLIENT	1

Vemos como solo nos descarta 2 variables, pero gráficamente muy pocas aportan información que muestren diferencias grandes entre clúster.

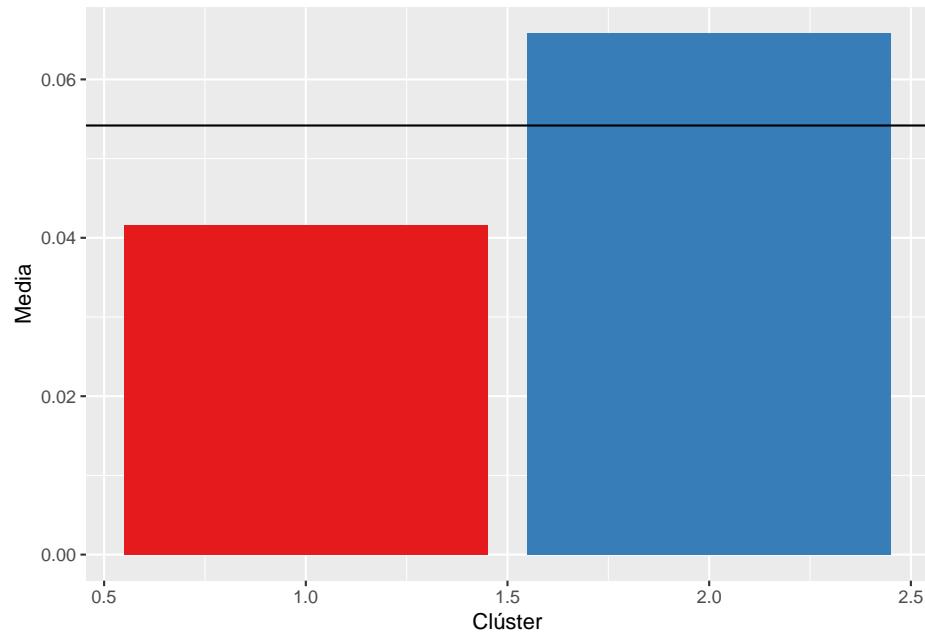
Se grafican las variables según clúster. Para las variables numéricas se mostrará la media grupal y la media global; para las variables categóricas se mostrarán las cantidades de cada nivel de la variable categórica por clúster.

Figura 115: Medias del Ratio del Importe del préstamo por clúster respecto la media global



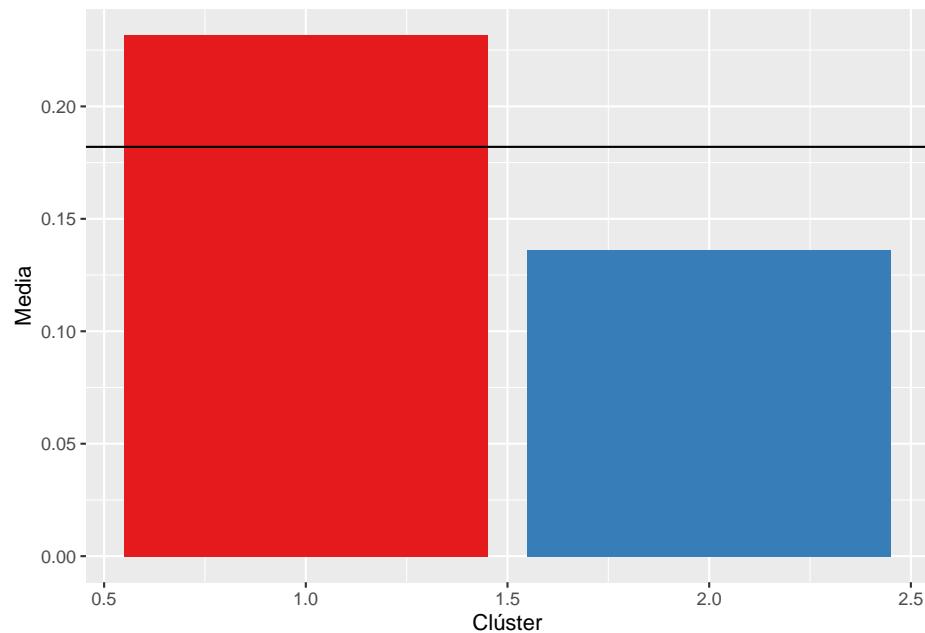
Observamos que el Clúster 1 presenta un período de reembolso más breve en comparación con el Clúster 2, en términos de la cantidad de años que los clientes tardan en devolver el préstamo.

Figura 116: Medias del Ratio de la Anualidad del préstamo por clúster respecto la media global



En el Clúster 1, se destaca que la proporción de anualidad con respecto al crédito es más alta en comparación con los otros grupos.

Figura 117: Medias de la Capacidad de cliente para pagar la annuity con sus ingresos por clúster respecto la media global



En el Clúster 1, se evidencia una mayor capacidad para reembolsar el préstamo en función de los ingresos

de los individuos.

Conclusiones:

La distinción fundamental entre nuestros clústeres reside en sus capacidades económicas. En particular, el Cluster 1 se caracteriza como el conjunto de clientes con la mejor capacidad para reembolsar el préstamo. Este grupo exhibe una proporción de anualidad más elevada y se destaca por devolver el préstamo de manera más rápida, con una media de 2 años.

Por otro lado, el Cluster 2 se identifica como el grupo de clientes con menor capacidad para reembolsar el préstamo, reflejándose en una proporción de anualidad más baja. Estos usuarios tienden a demorarse más en la devolución del préstamo, con una media de alrededor de 5-6 años.

Se adjuntan en el anexo otros gráficos relacionados con las variables analizadas, aunque no se han incluido en la presentación principal debido a su limitada relevancia informativa.

Reglas de asociación (Basket Market Analysis)

Transformamos las variables numéricas en categóricas aplicando la función `discretizeDF`.

Cabe resaltar que ahora la base de datos que se utilizará es “dcat” con las variables numéricas como categóricas.

Seguidamente, se transformará “dcat” en un data transactions para poder aplicar el Basket Market Analysis.

```
transactions in sparse format with
5000 transactions (rows) and
113 items (columns)
```

Con el siguiente summary, se puede ver con más detalle lo que se tiene:

```
transactions as itemMatrix in sparse format with
5000 rows (elements/itemsets/transactions) and
113 columns (items) and a density of 0.1415929
```

```
most frequent items:
NAME_EDUCATION_TYPE=Secondary / secondary special
3746
REGION_RATING_CLIENT=2
3641
CODE_GENDER=F
3098
NAME_FAMILY_STATUS=Married
3095
TARGET=0
2865
(Other)
63555
```

```
element (itemset/transaction) length distribution:
sizes
16
5000
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16	16	16	16	16	16

```
includes extended item information - examples:
labels           variables      levels
1               CODE_GENDER=F   CODE_GENDER      F
2               CODE_GENDER=M   CODE_GENDER      M
3 NAME_INCOME_TYPE=Businessman NAME_INCOME_TYPE Businessman
```

```
includes extended transaction information - examples:
transactionID
1                 1
2                 2
3                 3
```

Apriori

El primer paso consiste en especificar los parámetros:

El siguiente paso es crear las reglas de asociación:

Apriori

Parameter specification:

```
confidence minval smax arem aval originalSupport maxtime support minlen
      0.8     0.1     1 none FALSE           TRUE      5   0.002     1
maxlen target ext
      10    rules TRUE
```

Algorithmic control:

```
filter tree heap memopt load sort verbose
      0.1 TRUE TRUE FALSE TRUE     2    TRUE
```

Absolute minimum support count: 10

```
set item appearances ... [0 item(s)] done [0.00s].
set transactions ... [110 item(s), 5000 transaction(s)] done [0.00s].
sorting and recoding items ... [98 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [1.76s].
writing ... [2499418 rule(s)] done [0.55s].
creating S4 object ... done [1.13s].
```

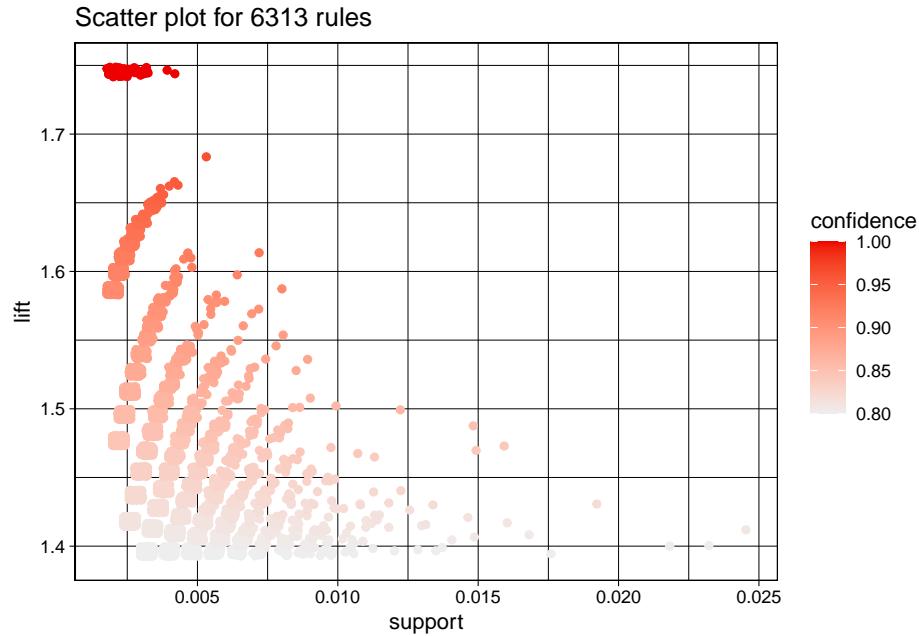
Dividimos las reglas de asociación obtenidas según lo consecuente que es la variable respuesta. La variable respuesta es TARGET, que toma valores de 1 o 0.

Se eliminan las reglas redundantes en ambos casos:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00000	0.01065	0.02847	0.02959	0.05815	0.06071

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.000e-09	4.957e-04	1.916e-03	2.117e-03	3.702e-03	4.951e-03

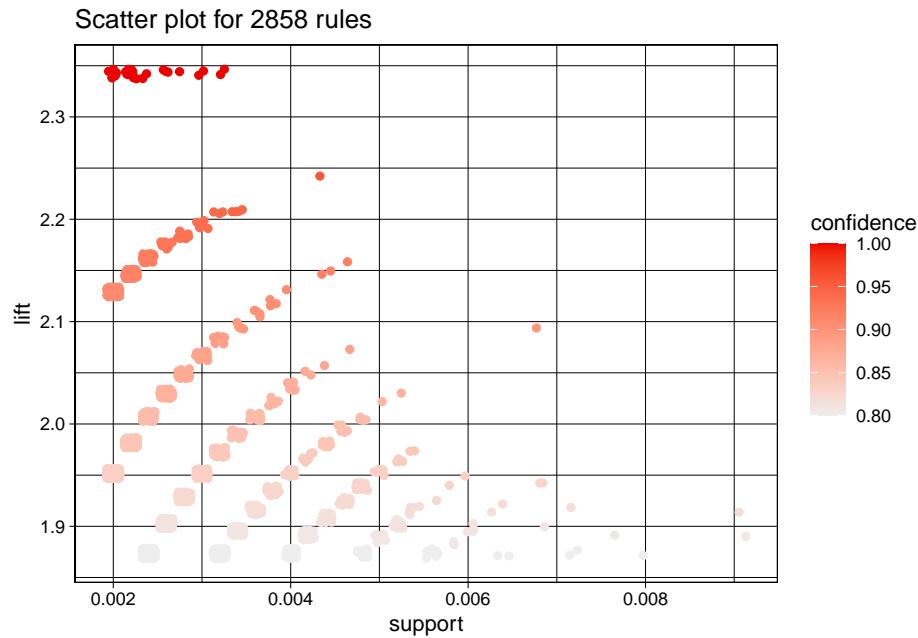
Figura 52: Scatter plot for 6313 rules



Como se puede ver, el primer gráfico muestra la matriz de puntos de las reglas de asociación filtrada respecto la métrica lift. La reglas de asociación de interés corresponden a los puntos con un color rojo de mayor intensidad (confianza que supere la mínima, 0.8) y se aprecia, estas reglas se sitúan en el gráfico con un soporte mayor al mínimo (0.002).

En el último gráfico se ve algo parecido, aquí las reglas de asociación que interesan corresponden a los puntos con una intensidad roja mayor y los puntos más grandes, que corresponderán a las reglas que tienen un soporte superior al mínimo (0.002).

Figura 53: Scatter plot for 2858 rules



Estos gráficos se interpretan de manera igual a los anteriores vistos.

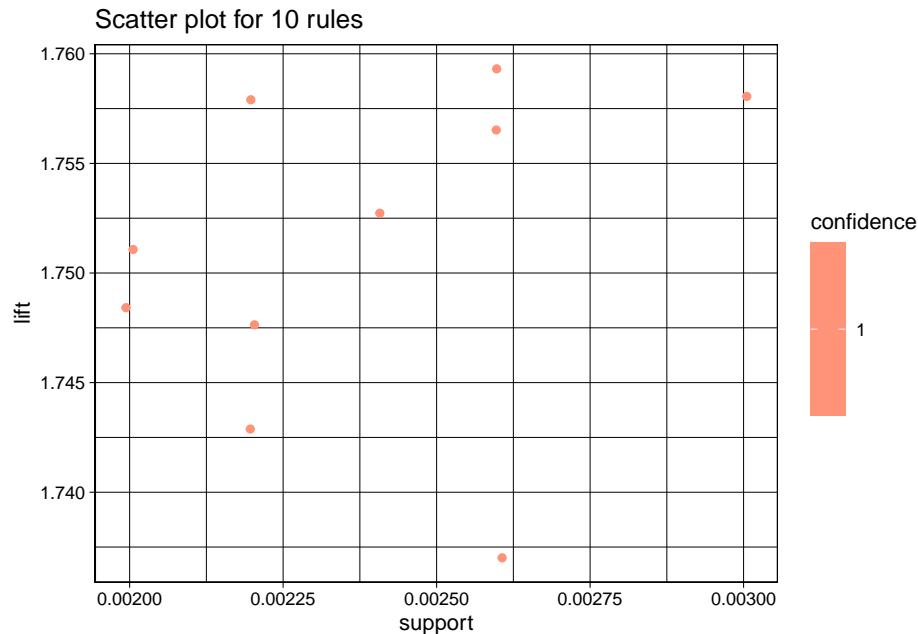
Con Target = 0 se obtienen 6313 reglas y con Tagret = 1 2858 reglas. Con la gran cantidad de reglas, la atención se centra en las 10 primeras reglas en cada caso con mayor lift.

Por tanto, se ven las 10 primeras reglas en cada caso con mayor lift, es decir, van ordenadas de forma decreciente siendo la primera la que tiene una mayor asociación encontrada con la variable respuesta, y se grafican en cada caso.

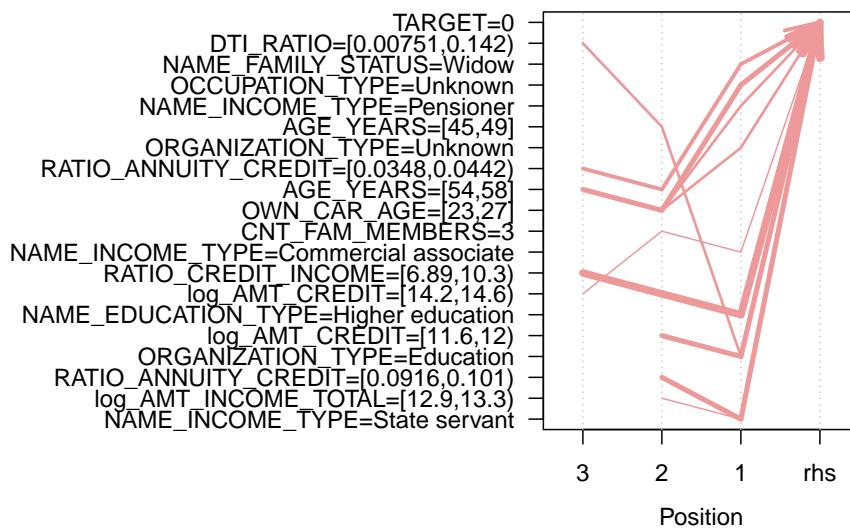
	lhs	rhs	support	confidence	coverage	lift	count
[1]	{NAME_INCOME_TYPE=State servant, log_AMT_INCOME_TOTAL=[12.9,13.3]}	=> {TARGET=0}	0.0020	1	0.0020	1.745201	10
[2]	{NAME_INCOME_TYPE=State servant, RATIO_ANNUITY_CREDIT=[0.0916,0.101)}	=> {TARGET=0}	0.0026	1	0.0026	1.745201	13
[3]	{ORGANIZATION_TYPE=Education, log_AMT_CREDIT=[11.6,12)}	=> {TARGET=0}	0.0026	1	0.0026	1.745201	13
[4]	{NAME_EDUCATION_TYPE=Higher education, log_AMT_CREDIT=[14.2,14.6), RATIO_CREDIT_INCOME=[6.89,10.3)}	=> {TARGET=0}	0.0030	1	0.0030	1.745201	15
[5]	{NAME_INCOME_TYPE=Commercial associate, CNT_FAM_MEMBERS=3, log_AMT_CREDIT=[14.2,14.6)}	=> {TARGET=0}	0.0020	1	0.0020	1.745201	10
[6]	{NAME_FAMILY_STATUS=Widow, AGE_YEARS=[54,58], RATIO_ANNUITY_CREDIT=[0.0348,0.0442)}	=> {TARGET=0}	0.0024	1	0.0024	1.745201	12
[7]	{ORGANIZATION_TYPE=Unknown, OWN_CAR_AGE=[23,27], AGE_YEARS=[54,58]}	=> {TARGET=0}	0.0022	1	0.0022	1.745201	11
[8]	{NAME_INCOME_TYPE=Pensioner,						

	OWN_CAR_AGE=[23,27], AGE_YEARS=[54,58]}	=> {TARGET=0} 0.0022	1	0.0022 1.745201	11
[9]	{OCCUPATION_TYPE=Unknown, OWN_CAR_AGE=[23,27], AGE_YEARS=[54,58]}	=> {TARGET=0} 0.0026	1	0.0026 1.745201	13
[10]	{ORGANIZATION_TYPE=Education, AGE_YEARS=[45,49], DTI_RATIO=[0.00751,0.142]}	=> {TARGET=0} 0.0022	1	0.0022 1.745201	11

Figura 54: Scatter plot for 10 rules

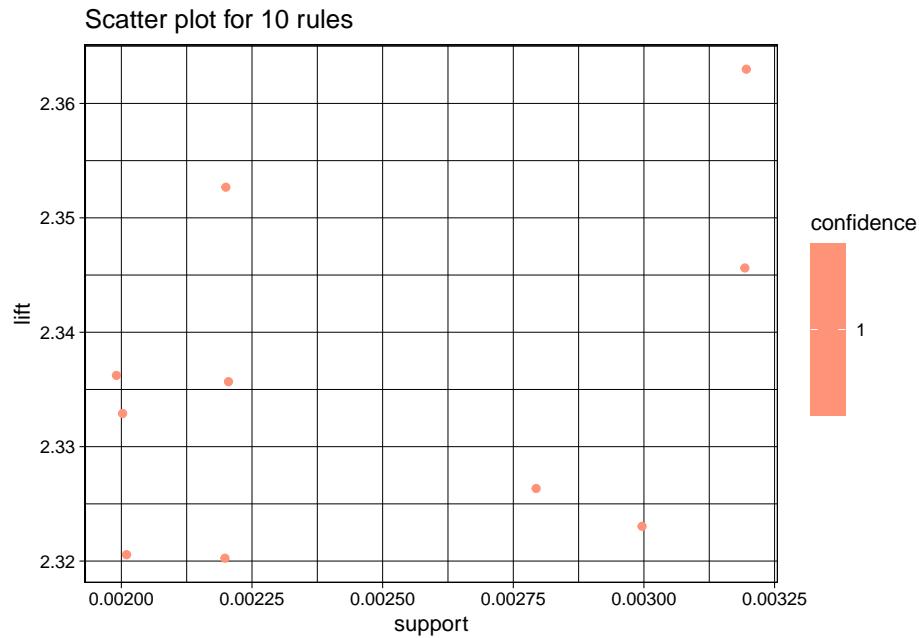


Parallel coordinates plot for 10 rules

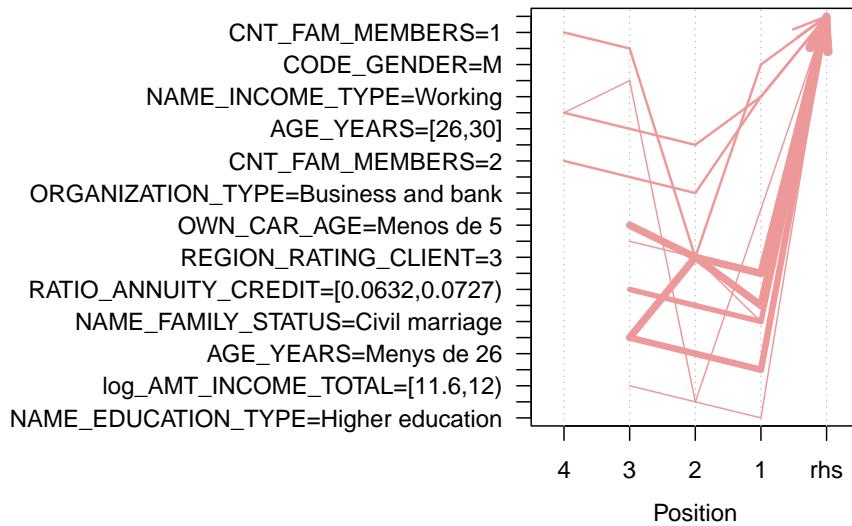


	lhs	rhs	support	confidence	coverage	lift	count
[1]	{NAME_EDUCATION_TYPE=Higher education, OWN_CAR_AGE=[23,27], log_AMT_INCOME_TOTAL=[11.6,12]}	=> {TARGET=1}	0.0020	1	0.0020	2.34192	10
[2]	{log_AMT_CREDIT=[12.4,12.9], AGE_YEARS=Menys de 26, RATIO_ANNUITY_CREDIT=[0.0727,0.0821]}	=> {TARGET=1}	0.0030	1	0.0030	2.34192	15
[3]	{NAME_FAMILY_STATUS=Civil marriage, OCCUPATION_TYPE=Low-mid skill laborers, RATIO_ANNUITY_CREDIT=[0.0632,0.0727]}	=> {TARGET=1}	0.0028	1	0.0028	2.34192	14
[4]	{OCCUPATION_TYPE=Low skill laborers, REGION_RATING_CLIENT=3, RATIO_ANNUITY_CREDIT=[0.0727,0.0821]}	=> {TARGET=1}	0.0032	1	0.0032	2.34192	16
[5]	{NAME_FAMILY_STATUS=Civil marriage, REGION_RATING_CLIENT=3, AGE_YEARS=[45,49]}	=> {TARGET=1}	0.0020	1	0.0020	2.34192	10
[6]	{OCCUPATION_TYPE=Low-mid skill laborers, REGION_RATING_CLIENT=3, OWN_CAR_AGE=Menos de 5}	=> {TARGET=1}	0.0032	1	0.0032	2.34192	16
[7]	{NAME_INCOME_TYPE=Working, ORGANIZATION_TYPE=Business and bank, OWN_CAR_AGE=[28,32], CNT_FAM_MEMBERS=2}	=> {TARGET=1}	0.0022	1	0.0022	2.34192	11
[8]	{NAME_INCOME_TYPE=Working, ORGANIZATION_TYPE=Trade and telecom, AGE_YEARS=[26,30], DTI_RATIO=[0.142,0.276]}	=> {TARGET=1}	0.0022	1	0.0022	2.34192	11
[9]	{OCCUPATION_TYPE=Unknown, OWN_CAR_AGE=[23,27], log_AMT_INCOME_TOTAL=[11.2,11.6], DTI_RATIO=[0.142,0.276]}	=> {TARGET=1}	0.0020	1	0.0020	2.34192	10
[10]	{CODE_GENDER=M, REGION_RATING_CLIENT=3, OWN_CAR_AGE=[19,22], CNT_FAM_MEMBERS=1}	=> {TARGET=1}	0.0022	1	0.0022	2.34192	11

Figura 55: Scatter plot for 10 rules



Parallel coordinates plot for 10 rules



ECLAT

Para este apartado, se crearán las reglas de asocioación con ECLAT.

Eclat

```
parameter specification:
```

```

tidLists support minlen maxlen           target ext
      FALSE    0.002       1      10 frequent itemsets TRUE

algorithmic control:
sparse sort verbose
    7   -2    TRUE

Absolute minimum support count: 10

create itemset ...
set transactions ...[110 item(s), 5000 transaction(s)] done [0.00s].
sorting and recoding items ... [98 item(s)] done [0.00s].
creating bit matrix ... [98 row(s), 5000 column(s)] done [0.00s].
writing ... [1992604 set(s)] done [0.66s].
Creating S4 object ... done [0.55s].
```

set of 51080 rules

```

rule length distribution (lhs + rhs):sizes
  2     3     4     5     6     7     8     9     10
  3   111   1756   8211  15442  14643   7758   2596    560

  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  2.00    6.00   7.00   6.55   7.00   10.00
```

summary of quality measures:

support	confidence	lift	itemset
Min. :0.002000	Min. :0.8000	Min. :1.396	Min. : 2
1st Qu.:0.002200	1st Qu.:0.8235	1st Qu.:1.437	1st Qu.: 271311
Median :0.002400	Median :0.8462	Median :1.477	Median : 670530
Mean :0.002764	Mean :0.8626	Mean :1.505	Mean : 767225
3rd Qu.:0.003000	3rd Qu.:0.9091	3rd Qu.:1.587	3rd Qu.:1289119
Max. :0.024600	Max. :1.0000	Max. :1.745	Max. :1987637

mining info:

```

data ntransactions support
  tr          5000    0.002
```

call

```

eclat(data = tr, parameter = list(support = soporte_minimo, minlen = 1, maxlen = tamanyo_conjunto))
confidence
  0.8
```

set of 10520 rules

```

rule length distribution (lhs + rhs):sizes
  3     4     5     6     7     8     9     10
  4   206   1324   3155  3432  1864   479    56

  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  3.000   6.000   7.000   6.672   7.000  10.000
```

```

summary of quality measures:
      support      confidence      lift      itemset
Min.   :0.002000   Min.   :0.8000   Min.   :1.874   Min.   : 694
1st Qu.:0.002000  1st Qu.:0.8235  1st Qu.:1.929  1st Qu.: 418570
Median :0.002400  Median :0.8462  Median :1.982  Median : 917806
Mean   :0.002532  Mean   :0.8607  Mean   :2.016  Mean   : 882066
3rd Qu.:0.002800 3rd Qu.:0.9091  3rd Qu.:2.129  3rd Qu.:1238898
Max.   :0.009200  Max.   :1.0000  Max.   :2.342  Max.   :1976927

mining info:
  data ntransactions support
  tr      5000     0.002
                                         call
eclat(data = tr, parameter = list(support = soporte_minimo, minlen = 1, maxlen = tamanyo_conjunto))
confidence
  0.8

      lhs                                rhs      support confidence      lift itemset
[1] {NAME_INCOME_TYPE=State servant,
     log_AMT_INCOME_TOTAL=[12.9,13.3]}    => {TARGET=0} 0.0020          1 1.745201  10346
[2] {NAME_EDUCATION_TYPE=Higher education,
     log_AMT_CREDIT=[14.2,14.6],
     RATIO_CREDIT_INCOME=[6.89,10.3]}      => {TARGET=0} 0.0030          1 1.745201  13160
[3] {NAME_INCOME_TYPE=Commercial associate,
     CNT_FAM_MEMBERS=3,
     log_AMT_CREDIT=[14.2,14.6]}          => {TARGET=0} 0.0020          1 1.745201  14725
[4] {NAME_INCOME_TYPE=Commercial associate,
     NAME_FAMILY_STATUS=Married,
     ORGANIZATION_TYPE=Personal services,
     RATIO_CREDIT_INCOME=[0.125,3.51]}     => {TARGET=0} 0.0030          1 1.745201  23611
[5] {CODE_GENDER=F,
     OCCUPATION_TYPE=Low-mid skill laborers,
     REGION_RATING_CLIENT=2,
     RATIO_ANNUITY_CREDIT=[0.101,0.111]}   => {TARGET=0} 0.0028          1 1.745201  62972
[6] {NAME_INCOME_TYPE=State servant,
     RATIO_ANNUITY_CREDIT=[0.0916,0.101]}   => {TARGET=0} 0.0026          1 1.745201  66662
[7] {NAME_FAMILY_STATUS=Widow,
     AGE_YEARS=[54,58],
     RATIO_ANNUITY_CREDIT=[0.0348,0.0442]}  => {TARGET=0} 0.0024          1 1.745201  95632
[8] {ORGANIZATION_TYPE=Education,
     log_AMT_CREDIT=[11.6,12]}             => {TARGET=0} 0.0026          1 1.745201  118213
[9] {NAME_INCOME_TYPE=Pensioner,
     OWN_CAR_AGE=[23,27],
     AGE_YEARS=[54,58]}                  => {TARGET=0} 0.0022          1 1.745201  132450
[10] {ORGANIZATION_TYPE=Unknown,
      OWN_CAR_AGE=[23,27],
      AGE_YEARS=[54,58]}                 => {TARGET=0} 0.0022          1 1.745201  132466

      lhs                                rhs      support confidence      lift itemset
[1] {NAME_INCOME_TYPE=Working,
     ORGANIZATION_TYPE=Business and bank,

```

	OWN_CAR_AGE=[28,32], CNT_FAM_MEMBERS=2}	=> {TARGET=1} 0.0022	1 2.34192 20786
[2]	{NAME_INCOME_TYPE=Working, ORGANIZATION_TYPE=Trade and telecom, AGE_YEARS=[26,30], DTI_RATIO=[0.142,0.276]}	=> {TARGET=1} 0.0022	1 2.34192 78683
[3]	{OCCUPATION_TYPE=Unknown, OWN_CAR_AGE=[23,27], log_AMT_INCOME_TOTAL=[11.2,11.6], DTI_RATIO=[0.142,0.276]}	=> {TARGET=1} 0.0020	1 2.34192 135824
[4]	{NAME_EDUCATION_TYPE=Higher education, OWN_CAR_AGE=[23,27], log_AMT_INCOME_TOTAL=[11.6,12]}	=> {TARGET=1} 0.0020	1 2.34192 137658
[5]	{CODE_GENDER=M, REGION_RATING_CLIENT=3, OWN_CAR_AGE=[19,22], CNT_FAM_MEMBERS=1}	=> {TARGET=1} 0.0022	1 2.34192 204659
[6]	{log_AMT_CREDIT=[12.4,12.9], AGE_YEARS=Menys de 26, RATIO_ANNUITY_CREDIT=[0.0727,0.0821]}	=> {TARGET=1} 0.0030	1 2.34192 213819
[7]	{NAME_FAMILY_STATUS=Single / not married, OCCUPATION_TYPE=Low skill laborers, log_AMT_INCOME_TOTAL=[11.6,12], AGE_YEARS=Menys de 26}	=> {TARGET=1} 0.0020	1 2.34192 217788
[8]	{NAME_FAMILY_STATUS=Separated, OCCUPATION_TYPE=Low-mid skill laborers, log_AMT_INCOME_TOTAL=[12,12.5], RATIO_CREDIT_INCOME=[0.125,3.51]}	=> {TARGET=1} 0.0024	1 2.34192 244609
[9]	{NAME_FAMILY_STATUS=Civil marriage, OCCUPATION_TYPE=Low-mid skill laborers, RATIO_ANNUITY_CREDIT=[0.0632,0.0727]}	=> {TARGET=1} 0.0028	1 2.34192 303233
[10]	{OCCUPATION_TYPE=Low-mid skill laborers, CNT_FAM_MEMBERS=2, RATIO_ANNUITY_CREDIT=[0.0632,0.0727], DTI_RATIO=[0.00751,0.142]}	=> {TARGET=1} 0.0026	1 2.34192 319187

Preprocessing de una base de datos real

Tras haber analizado la base de datos de forma estricta y haber aplicado diferentes métodos de clusterización, el siguiente paso será aplicar todo lo visto a una base de datos real. Como bien se sabe, en una base de datos de un banco es normal encontrar que la gran mayoría de clientes no sean morosos, ya que sino la salud de la que gozaría el banco sería pésima. Así pues, ahora se trabajará con una base de datos donde el número de clientes morosos no sea cercano al 50 %, sino más bien cercano al 0. Para ello, se ha seleccionado una base de datos desbalanceada.

Así pues, será necesario preprocessar la base de datos nuevamente. Para ello, será óptimo seguir los pasos propuestos por Karina Gibert con el objetivo de desarrollar correctamente el KDD y, así, obtener conclusiones óptimas a partir de nuestros datos.

Para ello, seguiremos 4 grandes bloques:

- Limpieza de datos y estandarización de formato
- Detección y tratamiento de missings
- Detección y tratamiento de outliers
- Feature Engineering

Limpieza de datos y estandarización de formato

Una vez se ha realizado la descriptiva preprocessing y se ha identificado el número de valores missing en nuestra base de datos, es óptimo analizar todas las variables una a una, así como algunas variables categóricas a las cuales se les puede reducir el número de categorías.

Para empezar, se puede apreciar que la variable OCCUPATION_TYPE tiene un total de 18 categorías:

Cuadro 46: Distribución inicial de la variable OCCUPATION TYPE

Categoría	Frecuencia
Accountants	163
Cleaning staff	80
Cooking staff	114
Core staff	440
Drivers	301
High skill tech staff	169
HR staff	7
IT staff	8
Laborers	863
Low-skill Laborers	29
Managers	380
Medicine staff	130
Private service staff	46
Realty agents	18
Sales staff	514
Secretaries	26
Security staff	112
Waiters/barmen staff	30
NA	1570

Una buena idea sería combinar algunas categorías con el objetivo de reducir el número de categorías y, además, aumentar el número de individuos por categoría. Seguidamente, se muestran los cambios realizados, donde se han agrupado todos los individuos en 5 categorías en función del capital humano empleado para su puesto:

- Low skill laborers: Engloba las categorías de “security staff”, “cooking staff”, “cleaning staff”, “drivers”, “low skill laborers”, “waiters staff”.
- Low-mid skill laborers: Engloba las categorías de “secretaries”, “private service staff” y “laborers”.
- Mid skill laborers: Engloba las categorías de “accountants”, “HR staff” y “sales staff”.
- Mid-high skill laborers: Engloba las categorías de “IT staff”, “realty agents” y “core staff”.
- High skill staff: Engloba las categorías de “high skill tech staff”, “managers” y “medicine staff”.

Cuadro 47: Distribución final de la variable OCCUPATION TYPE

Categoría	Frecuencia
High skill laborers	679
Low-mid skill laborers	935
Low skill laborers	666
Mid-high skill laborers	466
Mid skill laborers	684
NA	1570

Este proceso lo repetiremos con la variable ORGANIZATION_TYPE:

Cuadro 48: Distribución inicial de la variable ORGANIZATION TYPE

Categoría	Frecuencia
Advertising	8
Agriculture	46
Bank	33
Business Entity Type 1	91
Business Entity Type 2	153
Business Entity Type 3	1118
Cleaning	6
Construction	107
Culture	11
Electricity	8
Emergency	20
Government	152
Hotel	11
Housing	38
Industry: type 1	19
Industry: type 10	2
Industry: type 11	44
Industry: type 12	4
Industry: type 13	2
Industry: type 2	2
Industry: type 3	47
Industry: type 4	12
Industry: type 5	10
Industry: type 6	2
Industry: type 7	25
Industry: type 8	2
Industry: type 9	55
Insurance	7
Kindergarten	119
Legal Services	4
Medicine	191
Military	44
Mobile	4
Other	280
Police	50
Postal	32
Realtor	10
Religion	2
Restaurant	36
School	127
Security	49
Security Ministries	36
Self-employed	617
Services	28
Telecom	8
Trade: type 1	4
Trade: type 2	42
Trade: type 3	53
Trade: type 5	1
Trade: type 6	143
Trade: type 7	135
Transport: type 1	5
Transport: type 2	38
Transport: type 3	13
Transport: type 4	78
University	19

Como se puede apreciar, en este caso disponemos de muchísimas categorías, pero es de destacar la categoría XNA, la cual deberíamos sustituir a NA, para después poder imputarle algún valor. Así pues, se ha agrupado cada categoría profesional en función del sector al que se dedica el individuo. Así, la distribución final es la siguiente:

Cuadro 49: Distribución final de la variable ORGANIZATION TYPE

Categoría	Frecuencia
Business and bank	1402
Education	265
Industry and construction	333
Medicine	191
Other	419
Personal services	146
Public services	310
Self-employed	617
Trade and telecom	253
Transport	134

Ahora, esta variable pasa a tener 10 categorías, las cuales representan los diferentes sectores presentes en la economía presente hoy en día.

Así pues, el resto de variables tienen una uniformidad evidente: se puede apreciar cómo las variables categóricas presentan un número de categorías pequeño y, por parte de las variables numéricas, todas están expresadas en las mismas unidades, de forma que no habrá problemas con la manipulación de éstas.

Detección y tratamiento de missings

Para este apartado, trataremos de identificar aquellos valores desconocidos y valorar sobre su aleatoriedad para, posteriormente, imputar valores. Para empezar, es de destacar cómo hay 47 individuos con un coche de 64 años y 11 con un coche de 65. Si nos fijamos en la distribución de esta variable, es muy extraño que haya tantos individuos con valores atípicos, ya que el siguiente valor máximo es 46. Así, se potará por imputar valores nulos a estos individuos.

Seguidamente, pasaremos a imputar diferentes valores a aquellas variables donde hay observaciones sobre las cuales se desconocen sus valores reales. Este paso es necesario, ya que el hecho de disponer de valores desconocidos (también conocidos como NA) dificulta el análisis posterior de la variable.

Una vez hemos recategorizado todas aquellas variables que presentaban problemas, el número de NA por variables es el siguiente:

Cuadro 50: Missings por variable

Categoría	Frecuencia
AMT_INCOME_TOTAL	0
AMT_CREDIT	0
AMT_ANNUITY	0
DAYS_BIRTH	0
OWN_CAR_AGE	3363
AMT_GOODS_PRICE	2
CNT_FAM_MEMBERS	0
CODE_GENDER	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
OCCUPATION_TYPE	1570
ORGANIZATION_TYPE	930
REGION_RATING_CLIENT	0
TARGET	0

Una vez tenemos identificados todos los valores missing de nuestra base de datos, será necesario identificar si éstos son completamente aleatorios (MCAR), aleatorios (MAR), o no aleatorios (MNAR). Para ello, realizaremos el test de Little, el cual indica si los missings disponibles en la base de datos son fruto del azar o si siguen un patrón.

Para este test, diremos que los datos no siguen un patrón si no se rechaza hipótesis nula o, alternativamente, si no encuentra patrones entre los missings. Así pues, este es el resultado:

Cuadro 51: Test de Little

statistic	df	p.value	missing.patterns
3321.21225527829	79	0	7

Como se puede apreciar, el algoritmo ha detectado 7 patrones entre los valores missing, de forma que no se puede decir que hay un patrón aleatorio, de forma que calificaremos nuestros valores missing como MNAR.

Seguidamente, imputaremos los valores por los tres métodos de imputación conocido, pero antes de imputar los valores numéricos, será necesario pasar los NA a categoría `unknown`.

Seguidamente, toca imputar los NA disponibles en las variables numéricas de nuestros datos. Para ello, utilizaremos tres métodos distintos: kNN, MiMMi y MICE. Posteriormente, se comparará la imputación entre estos métodos y se seleccionará el método que resulte una distribución más parecida a la original antes de imputar.

Imputación por criterios estadísticos

En este caso, el objetivo será imputar en función de criterios estadísticos básicos. Para ello, se procederá a imputar valores en función de la media estadística o algún otro estadístico central de distribución.

Imputación por kNN

El algoritmo K-Nearest Neighbors (KNN), es un método de clasificación supervisada, que utiliza la proximidad para hacer clasificaciones o predicciones sobre un punto de datos desconocido. El algoritmo, utiliza

un hiperparámetro llamado “k”, que representa el número de vecinos más cercanos y el cual se ha obtenido mediante el cálculo de $k = \sqrt{n}$.

A continuación, se crean dos objetos: `fullVariables`, que corresponde a las variables que no presentan ningún dato faltante y `uncompleteVars`, que guarda las variables con missings.

Como se puede observar, se obtiene la imputación de los valores faltantes en el dataframe `df_knn` utilizando el algoritmo descrito previamente.

Imputación por MiMMi

La imputación por MiMMi se realiza utilizando un enfoque basado en clústeres y se utiliza la distancia de Gower como métrica de distancia para medir la similitud entre observaciones.

La función `uncompleteVar` se define para verificar si hay valores faltantes (representados como NA) en un vector dado.

La función `Mode` se define para calcular la moda de un vector. Esta función se utiliza más adelante para imputar valores faltantes en variables categóricas.

Se define la función MiMMi.

Se usa la función MiMMi y se obtienen los resultados imputados.

Imputación por MICE

Por último, se recurrirá a imputar a través del MICE como último método de imputación de valores numéricos. El MICE (Multiple Imputation by chained Equations) se basa en un método iterativo a partir del cual se resuelven ecuaciones consecutivamente con el objetivo de imputar valores de la forma más aproximada posible. Así pues, es momento de imputarlo:

Decisión del método de imputación elegido

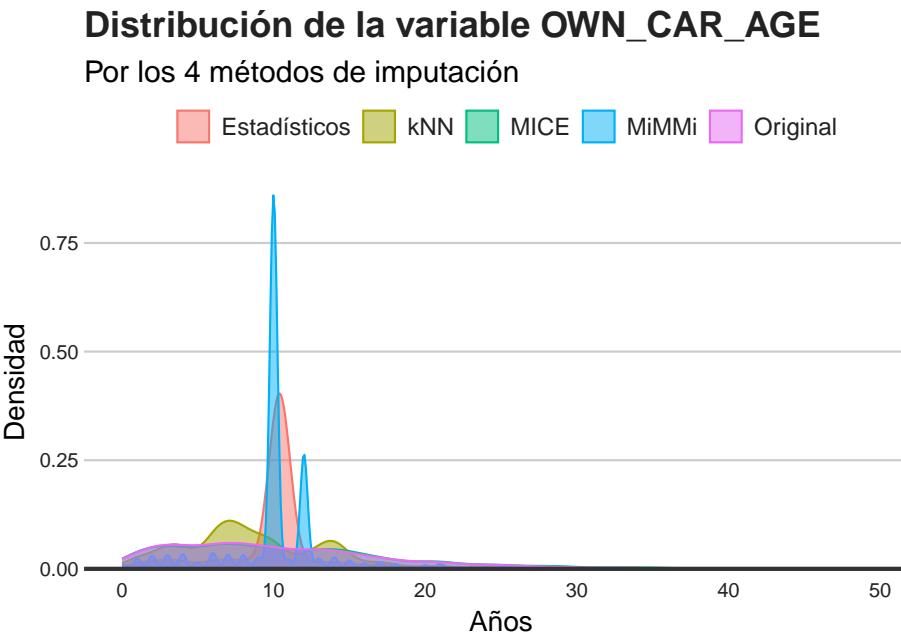
Llegados a este punto, en el momento de seleccionar el método de imputación elegido para el método de imputación final. En nuestro caso, como únicamente disponemos de dos variables numéricas con missings, podemos comparar la función de densidad de los datos originales contra los imputados por cada método. Así pues, vamos a mirar variable por variable:

`OWN_CAR_AGE`

Esta variable es la que presenta más valores no disponibles en nuestra base de datos, de forma que se acepta un mayor margen de error en cuanto a la imputación de valores se refiere. Así, la densidad resultante para cada método es la siguiente:

```
## Warning: Removed 3363 rows containing non-finite values ('stat_density()').
```

Figura 122: Gráfico comparación imputación OWN CAR AGE



Como se puede apreciar, hay tres métodos de imputación que claramente se alejan mucho de la distribución inicial de los datos: criterios estadísticos, kNN y MiMMi. Así pues, se puede apreciar como el MICE es el algoritmo que aproxima la densidad de los datos a los originales, de forma que este será el método escogido.

AMT_GOODS_PRICE

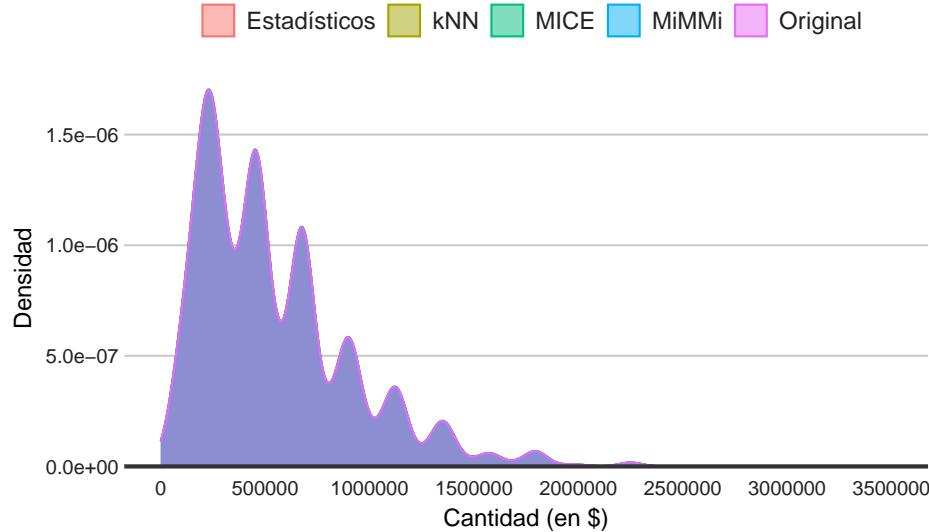
Como se ha visto previamente en el descriptiva preprocessing, esta variable únicamente presentaba 3 NA, de forma que la densidad en todos los métodos será muy similar:

```
## Warning: Removed 2 rows containing non-finite values ('stat_density()').
```

Figura 123: Gráfico comparación imputación AMT GOODS PRICE

Distribución de la variable AMT_GOODS_PRICE

Por los 4 métodos de imputación



Como se puede apreciar, todos los métodos retornan una estimación similar de la densidad, por lo que se podría decir que es indiferente escoger un método en concreto. De esta forma, se decide usar el MICE como método de imputación final seleccionado.

He aquí una tabla resumen sobre los resultados obtenidos acerca de cuál es el mejor criterio de imputación:

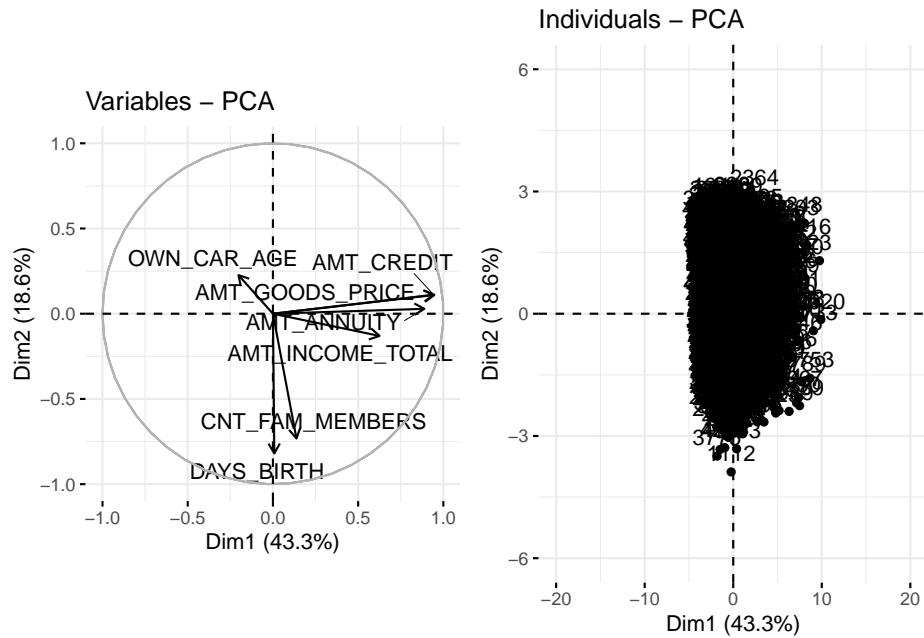
	OWN_CAR_AGE	AMT_GOODS_PRICE
Estadísticos	No	Yes
kNN	No	Yes
MICE	Yes	Yes
MiMMi	No	Yes

Detección y tratamiento de outliers

En este apartado se tratará de visualizar aquellas observaciones extremas y, además, discernir sobre si deben ser corregidas o no, dependiendo de la naturaleza de la variable. Para ello, se utilizarán métodos multivariantes, como el análisis de componentes principales (PCA). Así, se procede a representar la proyección de los individuos en los primeros planos factoriales para así observar cuáles se alejan del resto de puntos:

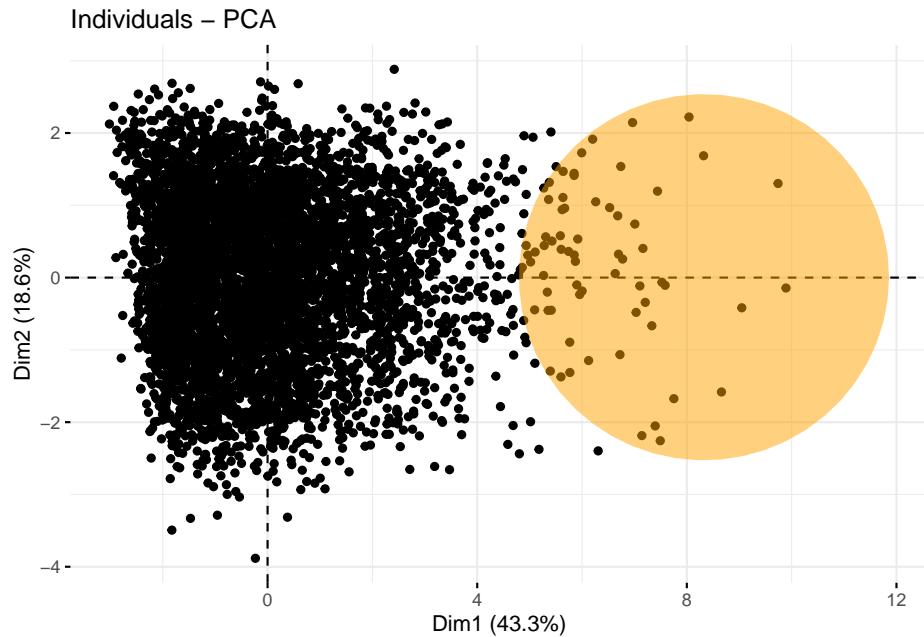
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Figura 124: Resultados PCA preliminar



Como se puede apreciar, la combinación de las dos primeras dimensiones del PCA acumulan un total del 60 % de la inercia total explicada, de forma que es un método de detección bastante fiable en nuestro caso. Identificamos, especialmente, un punto que sobresale del segundo plano factorial, mientras que podemos catalogar una decena de grupos realmente alejados del grupo en la primera dimensión:

Figura 125: Distribución individuos PCA



Tras haber realizado el PCA correspondiente, vemos claramente un grupo de outliers. Así pues, pasamos a analizar los que son valores extremos por la dimensión 1. Como se puede apreciar, el primer plano factorial viene dado por las variables referidas a cantidad de dinero de nuestra base de datos. Así pues, los outliers presentes son personas con unos ingresos muy altos y que, además, realizaron préstamos por una cantidad de dinero muy superior al que cobran. Así pues, se trata de personas ricas, las cuales existen en nuestra sociedad, de forma que se quedan en la base de datos tal y como aparece. Más adelante, se aplicará alguna transformación que pueda permitir corregir estos valores tan extremos.

Feature engineering

Por último, realizaremos la selección de variables final para nuestra base de datos, así como aplicar transformaciones correctas a nuestras variables para que cumplan algunas hipótesis, como normalidad o heteroscedasticidad. Para este apartado se hace una disección de cada variable una a una.

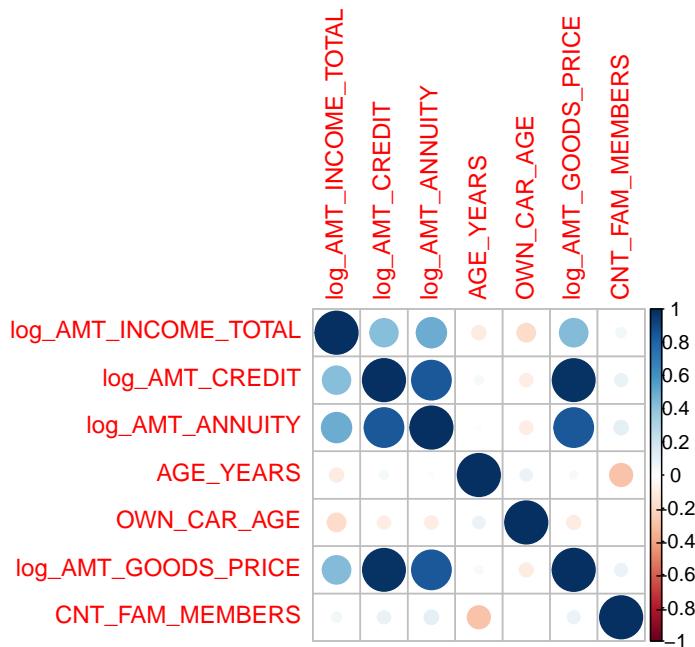
En primer lugar, se resolverán problemas relacionados con las variables numéricas. Como tenemos variables relacionadas con cantidades monetarias (salario, cantidad prestada...), tal vez sería mejor aplicar una transformación logarítmica:

Así pues, esta transformación debería resolver problemas relacionados con la normalidad de estas variables. Otro cambio a realizar es el respectivo a la variable `DAYS_BIRTH`, la cual muestra el número de días que lleva vivo el individuo. Sin embargo, el hecho de que esta variable esté en negativo y expresada en días (cuando normalmente se hace en años) hace que su interpretación sea complicada. De esta forma, se harán los cambios permanentes para encontrar la edad de los clientes, guardándola en una variable llamada `AGE_YEARS`.

Ahora, vamos a unir aquellas variables ya preprocesadas con el objetivo de tener el dataset preparado para crear nuevas variables.

Antes de avanzar, haremos un correlograma para ver los pares de variables con un mayor coeficiente de correlación de Pearson:

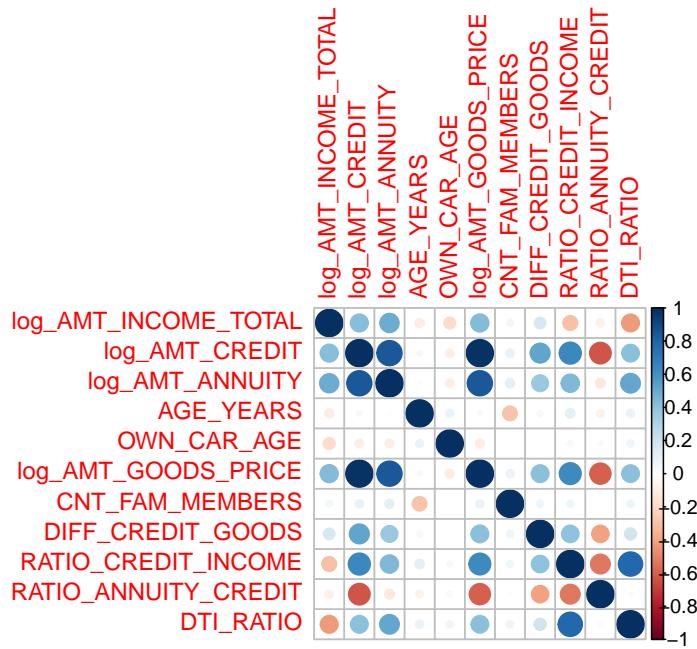
Figura 126: Matriz de correlaciones pre eliminación de variables



Como se puede apreciar y como era de esperar, hay 3 variables que presentan una gran autocorrelación entre ellas: `log_AMT_CREDIT`, `log_AMT_GOODS_PRICE` y `log_AMT_ANNUITY`. de esta forma, sería ideal nuevas variables a partir de éstas con las cuales se pueda resolver este problema, ya que explican exactamente lo mismo. Para ello, será necesario basarse en la teoría económica y en qué se fijan las entidades de crédito para conceder préstamos. Así, el siguiente objetivo será crear ratios y variables que pretendan controlar y relacionar dinero prestado con capacidad del cliente para retornarlo:

- `DIFF_CREDIT_GOODS`: Diferencia entre el crédito pedido y el valor del bien para el que se quiere usar
- `RATIO_CREDIT_INCOME`: Ratio entre el crédito pedido y el salario anual del prestatario. También se puede contar como el número de años que se tarda en devolver el crédito
- `RATIO_ANNUITY_CREDIT`: Ratio entre la anuidad del préstamo y el crédito total solicitado
- `DTI_RATIO`: El DTI (Debt-to-income) ratio mide la capacidad del cliente para pagar la anuity de su préstamo en relación con sus ingresos

Figura 127: Matriz de correlaciones post eliminación de variables

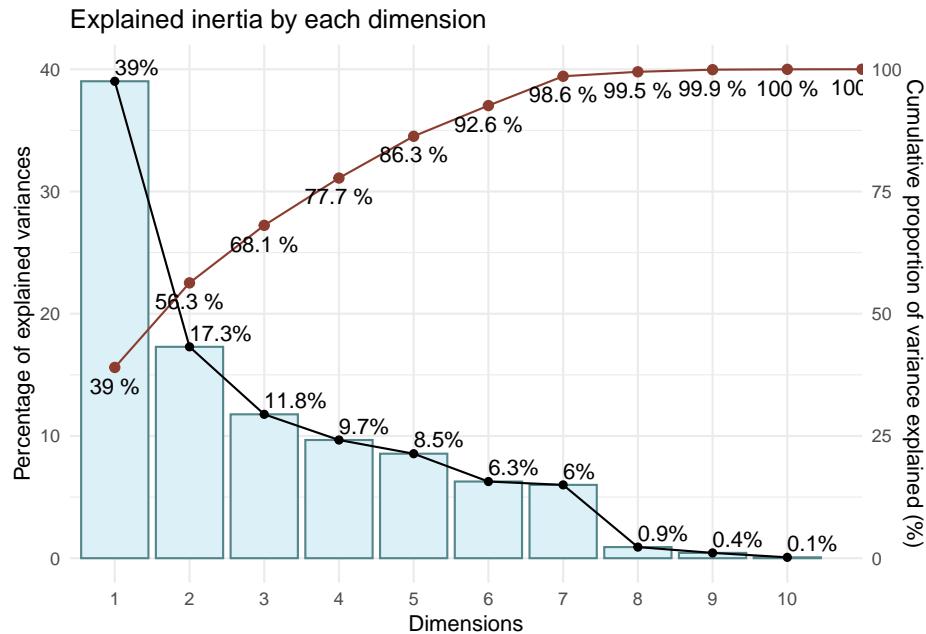


Se puede apreciar que, ahora, las nuevas variables creadas no presentan tanta correlación entre ellas como anteriormente había. Se puede apreciar, además, que las correlaciones entre las variables donde había problemas siguen teniéndolos y, como se aprecia en el PCA sencillo realizado antes, será necesario descartar alguna variable, ya que explican cosas similares en las mismas dimensiones. Así, en el PCA se deberá realizar el descarte adecuado de variables en función de su aportación al PCA resultante.

Eliminación de variables a través del PCA

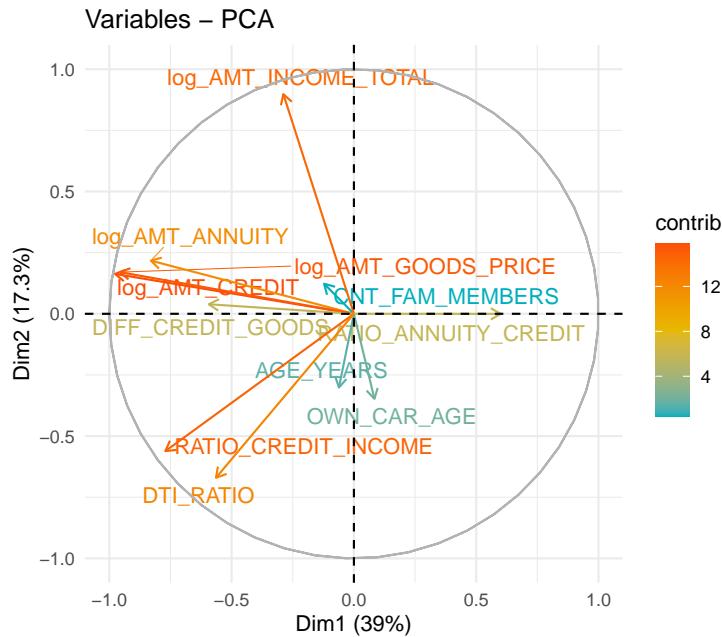
Se proceden a eliminar, primeramente, aquellas variables para las cuales ya existe su transformación logarítmica. Esto se hace para no contar con variables que contengan la misma capacidad explicativa (y así evitar colinealidad). También se elimina la variable `DAYS_BIRTH`, ya que se cuenta con `AGE_YEARS`, que es una transformación de la inicial, debido a que `DAYS_BIRTH` no tenía una clara interpretación.

Figura 128: Porcentaje de inercia explicado por dimensión



Teniendo en cuenta que la inercia equivale a la proporción de la variabilidad de los datos, se sabe que con un 80 % de inercia se puede obtener casi toda la información o variabilidad de la base de datos original. Con ello, vemos que el 80 % de la inercia acumulada se logra con 5 planos factoriales, pero aún se pueden eliminar algunas variables.

Figura 129: Proyección de variables en los dos primeros planos factoriales



Observamos la tabla de rotaciones:

Cuadro 53: Correlación de cada variable con cada plano factorial

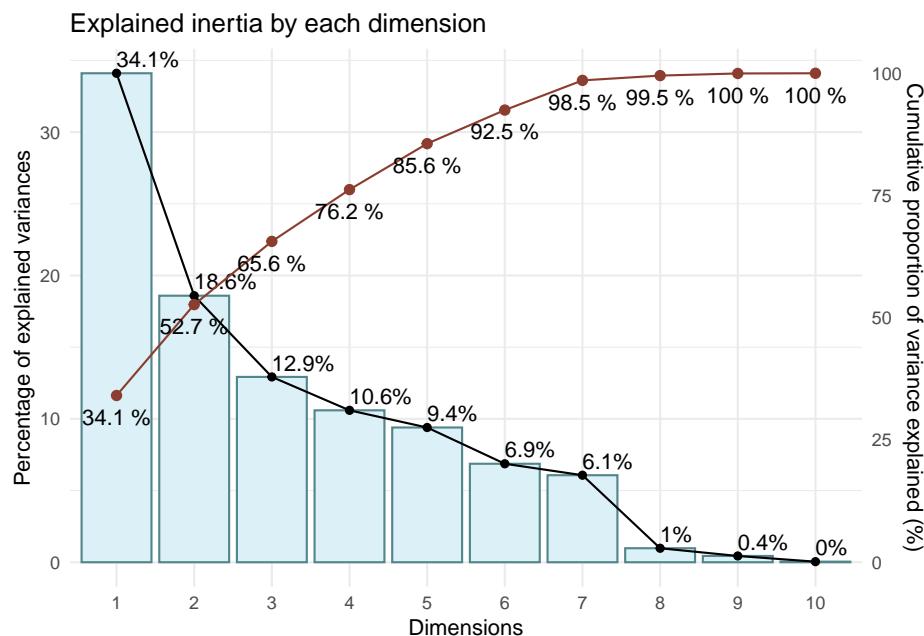
	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0404988	-0.2520416	0.0420325	0.1684744	-0.9283857
CNT_FAM_MEMBERS	-0.0587833	0.0890778	-0.6385913	0.3296125	-0.1048056
log_AMT_INCOME_TOTAL	-0.1392704	0.6511662	0.0917826	-0.1425163	-0.1993198
log_AMT_CREDIT	-0.4713797	0.1181949	0.0491501	-0.0075008	-0.0350003
log_AMT_ANNUITY	-0.4006526	0.1576017	-0.1522320	-0.4303298	-0.1562979
log_AMT_GOODS_PRICE	-0.4619578	0.1235471	0.0348434	-0.0554372	-0.0224585
AGE_YEARS	-0.0295574	-0.2184692	0.6087287	-0.2550239	-0.0856809
DIFF_CREDIT_GOODS	-0.2864311	0.0282989	0.1117310	0.3120808	-0.0804458
RATIO_CREDIT_INCOME	-0.3721363	-0.4069923	-0.0471240	0.0943421	0.1339615
RATIO_ANNUITY_CREDIT	0.2896051	-0.0002457	-0.3098813	-0.6173551	-0.1636010
DTI_RATIO	-0.2722293	-0.4860845	-0.2718906	-0.3191995	0.0520122

En el grafico vemos que las flechas de **log_AMT_GOODS_PRICE** y **log_AMT_CREDIT** se solapan entre ellas, eso quiere decir que las dos variables explican el mismo plano factorial. Vemos en la tabla de rotaciones que **log_AMT_CREDIT** contribuye más a explicar el primer plano factorial, y además las correlaciones entre cada una de las variables y cada dimensión son muy similares. Por esta razón eliminamos **log_AMT_GOODS_PRICE**.

Nos quedamos con una variable menos, por tanto tenemos 10 variables numéricas.

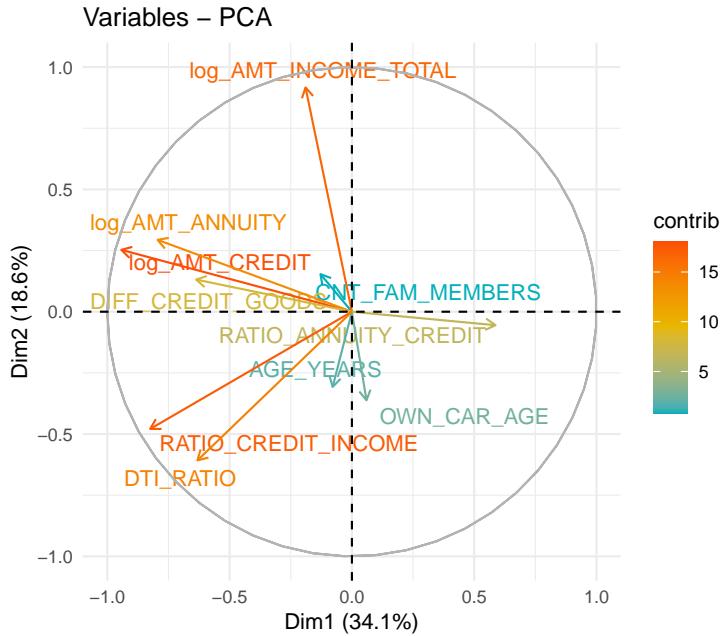
De vuelta, verificamos el porcentaje de inercia por cada componente principal y la acumulada:

Figura 130: Porcentaje de inercia explicado por dimensión



Como se puede ver, seguimos teniendo 5 dimensiones que acumulan el 80 % de la varianza.

Figura 131: Proyección de variables en los dos primeros planos factoriales



Vemos que las variables **CNT_FAM_MEMBERS**, **AGE_YEARS** y **OWN CAR AGE** no explican las dos primeras componentes pero si nos fijamos en la tabla de rotaciones vemos que sí tienen importancia a la hora de explicar las otras tres dimensiones:

Cuadro 54: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0322807	-0.2655285	0.0337229	0.1906453	-0.9215069
CNT_FAM_MEMBERS	-0.0695697	0.1126277	-0.6311899	0.3343099	-0.1006699
log_AMT_INCOME_TOTAL	-0.1026378	0.6723992	0.1061064	-0.1606319	-0.2082625
log_AMT_CREDIT	-0.5103846	0.1855628	0.0642747	-0.0265275	-0.0415701
log_AMT_ANNUITY	-0.4301674	0.2155338	-0.1420652	-0.4481163	-0.1695052
AGE_YEARS	-0.0426833	-0.2254254	0.6038354	-0.2627241	-0.0921654
DIFF_CREDIT_GOODS	-0.3447759	0.0963807	0.1357450	0.2733557	-0.0958677
RATIO_CREDIT_INCOME	-0.4465354	-0.3514036	-0.0419781	0.0869097	0.1330005
RATIO_ANNUITY_CREDIT	0.3178555	-0.0403997	-0.3230490	-0.6086835	-0.1711969
DTI_RATIO	-0.3419485	-0.4456210	-0.2751781	-0.3205941	0.0466332

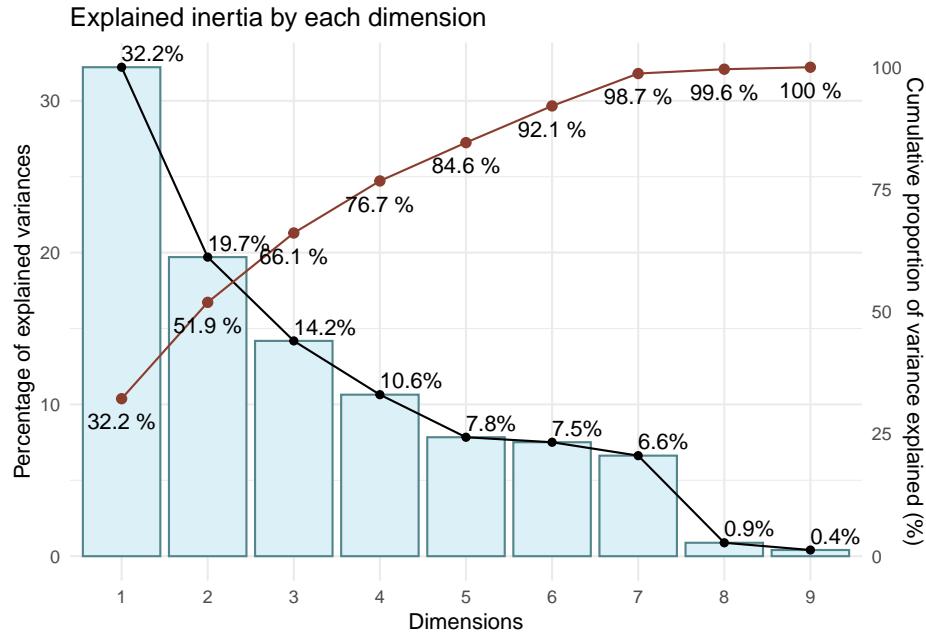
Por ejemplo, en el caso de **OWN CAR AGE** se puede ver en la tabla anterior que, se podría decir que no es la que mejor explica las primeras componentes, pero vemos que explica casi toda la componente 5.

Otra observación se podria hacer de las variables **log_AMT_CREDIT** y **log_AMT_ANNUITY**, donde se puede apreciar que tienen correlaciones similares con la primera y segunda dimensión. Teniendo en cuenta que esas dos primeras dimensiones (PC1 y PC2) son las más importantes, ya que acumulan la mayoría de la inercia (en total un 52.2 %), parece una decisión sensata eliminar una de ellas, en este caso **log_AMT_ANNUITY**.

Ahora conservamos 9 variables numéricas.

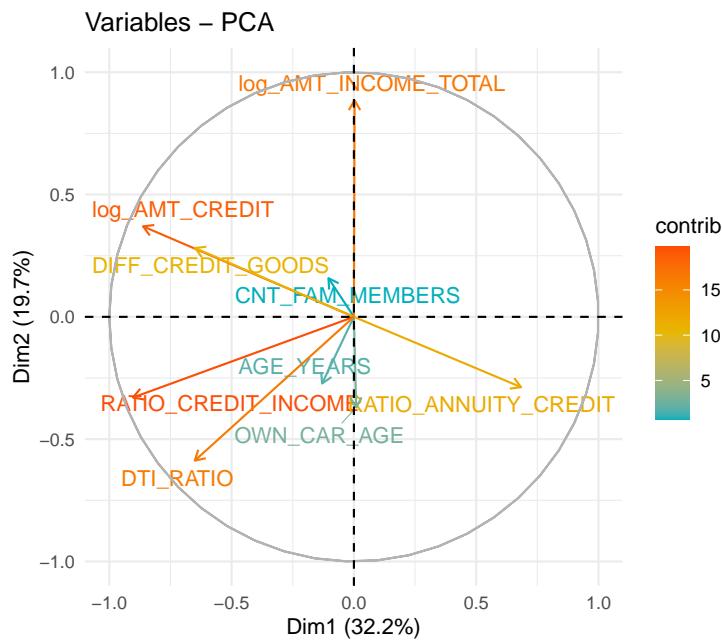
De forma igual que anteriormente, comprobamos el porcentaje de inercia para cada componente principal y la acumulada:

Figura 132: Porcentaje de inercia explicado por dimensión



Como se puede comprobar, las 5 dimensiones siguen siendo las necesarias para acumular el 80 % de la varianza.

Figura 133: Proyección de variables en los dos primeros planos factoriales



Observamos tambien la tabla de rotaciones para verificar si se puede eliminar alguna variable más:

Cuadro 55: Correlación de cada variable con cada plano factorial

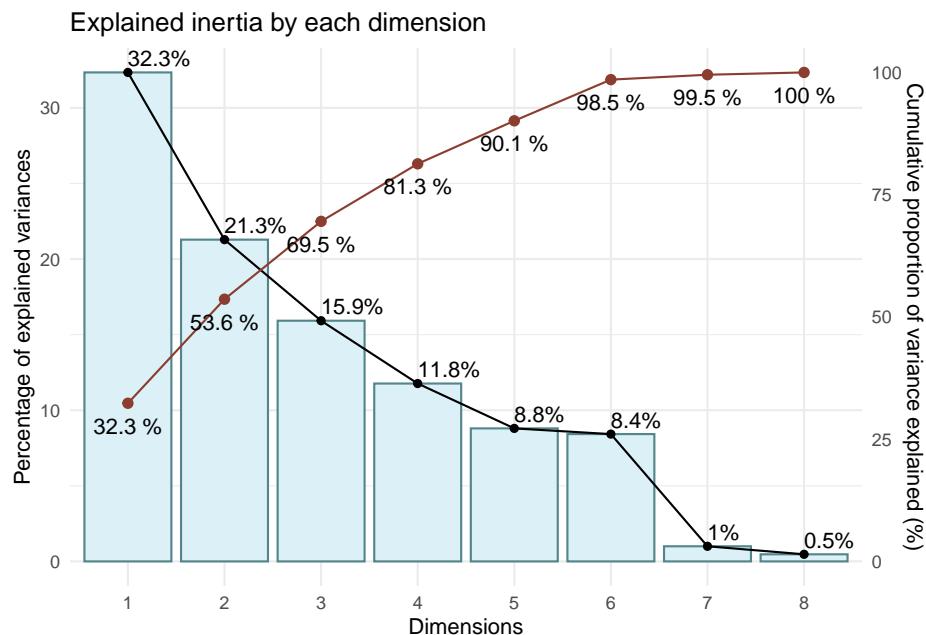
	PC1	PC2	PC3	PC4	PC5
OWN_CAR_AGE	0.0055603	-0.2780244	0.0531169	0.9223351	-0.0256378
CNT_FAM_MEMBERS	-0.0610978	0.1187701	-0.6903579	0.1973024	-0.3304238
log_AMT_INCOME_TOTAL	0.0015885	0.6650348	0.0959116	0.0773961	-0.3719778
log_AMT_CREDIT	-0.5063376	0.2777616	0.0304562	-0.0076639	-0.1507661
AGE_YEARS	-0.0763684	-0.2046021	0.6545383	-0.0052515	-0.5206031
DIFF_CREDIT_GOODS	-0.3814634	0.2113213	0.0565559	0.1870573	-0.0999994
RATIO_CREDIT_INCOME	-0.5305123	-0.2476189	-0.0692166	-0.0913992	0.0784052
RATIO_ANNUITY_CREDIT	0.4013462	-0.2158078	-0.1774102	-0.1340843	-0.6005254
DTI_RATIO	-0.3820828	-0.4413236	-0.2063341	-0.2072932	-0.2850602

Si nos fijamos en el gráfico que incluye los dos primeros planos factoriales (PC1 y PC2), resulta fácil ver que **log_AMT_CREDIT** y **DIFF_CREDIT_GOODS** se solapan en su proyección, teniendo **log_AMT_CREDIT** más contribución dado que el vector es más largo. De aquí se entiende que las correlaciones de ambas variables en los dos primeros planos factoriales son muy similares, motivo por el cual solapan. En la tabla de correlaciones anterior se puede comprobar como efectivamente, estas correlaciones son similares. Incluso la correlación en ambas variables con la tercera dimensión (PC3) es baja, de forma parecida. Por tanto, se procede a eliminar aquella con menos contribución en PC1 y PC2, esta siendo **DIFF_CREDIT_GOODS**.

Ahora se conservan 8 variables numéricas.

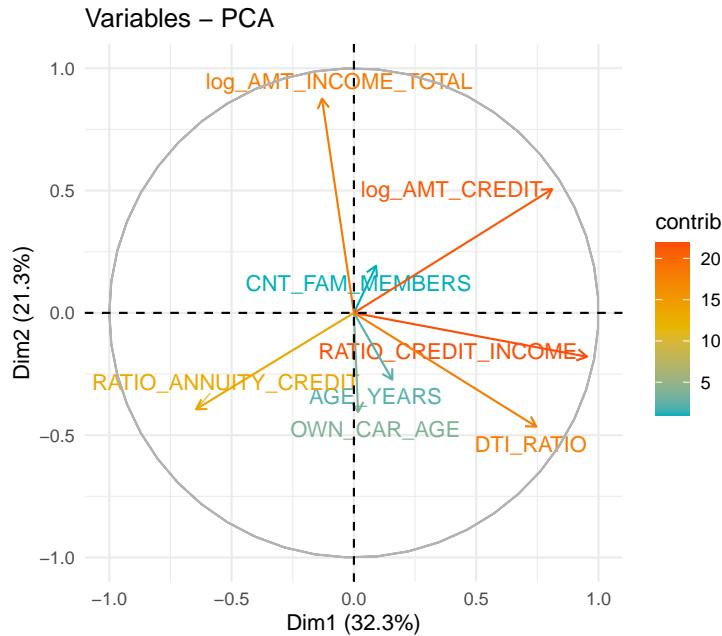
Se vuelven a ejecutar todos los pasos anteriores para volver a verificar si hace falta eliminar más variables:

Figura 134: Porcentaje de inercia explicado por dimensión



Se aprecia como la eliminación de **DIFF_CREDIT_GOODS** ha modificado el número de dimensiones necesarias para alcanzar el 80 % de inercia acumulada, pasando de 5 a 4 dimensiones.

Figura 135: Proyección de variables en los dos primeros planos factoriales



Cuadro 56: Correlación de cada variable con cada plano factorial

	PC1	PC2	PC3	PC4
OWN_CAR_AGE	0.0106237	-0.3106285	0.0266523	-0.9197928
CNT_FAM_MEMBERS	0.0562897	0.1475167	-0.6870283	-0.1995384
log_AMT_INCOME_TOTAL	-0.0811317	0.6713796	0.1197133	-0.1257007
log_AMT_CREDIT	0.5038620	0.3875551	0.0615613	-0.0648320
AGE_YEARS	0.0974651	-0.2092369	0.6506731	-0.0056776
RATIO_CREDIT_INCOME	0.5924242	-0.1374325	-0.0559061	0.0575498
RATIO_ANNUITY_CREDIT	-0.4012361	-0.3019369	-0.2026459	0.2169075
DTI_RATIO	0.4634685	-0.3563223	-0.2039469	0.2092098

Comprobando el gráfico de las dos primeras dimensiones, y analizando las correlaciones, parece ser que ya no hace falta eliminar más variables. Por tanto, conservamos 8 variables numéricas.

Las variables eliminadas han sido: - **AMT_INCOME_TOTAL**, **AMT_CREDIT**, **AMT_ANNUITY**, **AMT_GOODS_PRICE**, todas ellas con motivo de que ya se había creado otra variable a partir de su transformación logarítmica. - **DAYS_BIRTH**, ya que la variable **AGE_YEARS** es una transformación de ella. - **log_AMT_GOODS_PRICE** - **log_AMT_ANNUITY** - **DIFF_CREDIT_GOODS**

Balanceo de los datos

Por último, una vez ya se ha limpiado la base de datos y ésta es óptima para realizar un análisis de datos, es necesario balancearla. Para ello, entre todas las alternativas posibles que se han buscado, la opción que proporciona una base de datos balanceada más parecida a la base de datos original en cuanto a la estructura se refiere es la proporcionada por la función `SMOTE()`, del paquete `DMwR`. En este caso, esta función realiza el procedimiento SMOTE (Synthetic Minority Over-sampling TTechnique), el cual trata de crear datos sintéticos y artificiales en base a la clase minoritaria. Este algoritmo, en resumen, es una mejora al oversampling tradicional, de forma que se hace que existan problemas relacionados con el overfitting a la hora de aplicar modelos de machine learning. Así pues, en el siguiente apartado se realizará un análisis descriptivo de la nueva base de datos sobre la que se trabajará de aquí en adelante.

Análisis descriptivo de la nueva base de datos

Tras haber realizado el balanceo de la base de datos, toca hacer un análisis descriptivo de la misma para apreciar las diferencias que puedan existir con respecto a la que se utilizó anteriormente en el resto del trabajo. En la siguiente tabla se presenta un resumen general sobre la nueva base de datos balanceada y desbalanceada. Cabe mencionar que los cambios en cuanto a la estructura de la base de datos no es muy diferente, de forma que la descriptiva es muy similar a la que se puede encontrar en el principio del trabajo. Así pues, se comenzará realizando un análisis descriptivo muy general:

Cuadro 57: Tabla descriptiva datos desbalanceados

	[ALL]	N
	N=5000	
CODE_GENDER:		
F	3274 (65.5 %)	5000
M	1726 (34.5 %)	
NAME_INCOME_TYPE:		
Commercial associate	1159 (23.2 %)	5000
Pensioner	929 (18.6 %)	
State servant	371 (7.42 %)	
Student	1 (0.02 %)	
Unemployed	1 (0.02 %)	
Working	2539 (50.8 %)	
NAME_EDUCATION_TYPE:		
Academic degree	3 (0.06 %)	5000
Higher education	1254 (25.1 %)	
Incomplete higher	166 (3.32 %)	
Lower secondary	56 (1.12 %)	
Secondary / secondary special	3521 (70.4 %)	
NAME_FAMILY_STATUS:		
Civil marriage	472 (9.44 %)	5000
Married	3194 (63.9 %)	
Separated	320 (6.40 %)	
Single / not married	729 (14.6 %)	
Widow	285 (5.70 %)	
OCCUPATION_TYPE:		
High skill laborers	679 (13.6 %)	5000
Low-mid skill laborers	935 (18.7 %)	
Low skill laborers	666 (13.3 %)	
Mid-high skill laborers	466 (9.32 %)	
Mid skill laborers	684 (13.7 %)	
Unknown	1570 (31.4 %)	
REGION_RATING_CLIENT:		
1	520 (10.4 %)	5000
2	3679 (73.6 %)	
3	801 (16.0 %)	
TARGET:		
0	4599 (92.0 %)	5000
1	401 (8.02 %)	

Cuadro 58: Tabla descriptiva datos balanceados

	[ALL] N=2807	N
CODE_GENDER:		
F	1712 (61.0 %)	2807
M	1095 (39.0 %)	
NAME_INCOME_TYPE:		
Commercial associate	689 (24.5 %)	2807
Pensioner	431 (15.4 %)	
State servant	233 (8.30 %)	
Student	1 (0.04 %)	
Working	1453 (51.8 %)	
NAME_EDUCATION_TYPE:		
Higher education	745 (26.5 %)	2807
Incomplete higher	164 (5.84 %)	
Lower secondary	36 (1.28 %)	
Secondary / secondary special	1862 (66.3 %)	
NAME_FAMILY_STATUS:		
Civil marriage	323 (11.5 %)	2807
Married	1641 (58.5 %)	
Separated	207 (7.37 %)	
Single / not married	503 (17.9 %)	
Widow	133 (4.74 %)	
OCCUPATION_TYPE:		
High skill laborers	377 (13.4 %)	2807
Low-mid skill laborers	546 (19.5 %)	
Low skill laborers	399 (14.2 %)	
Mid-high skill laborers	267 (9.51 %)	
Mid skill laborers	411 (14.6 %)	
Unknown	807 (28.7 %)	
REGION_RATING_CLIENT:		
1	343 (12.2 %)	2807
2	1870 (66.6 %)	
3	594 (21.2 %)	
TARGET:		
0	1604 (57.1 %)	2807
1	1203 (42.9 %)	

Por lo tanto, en la tabla se presentan tanto la frecuencia absoluta como la frecuencia relativa de cada valor posible en cada variable categórica, ya sean dicotómicas o politómicas. Esto facilita la identificación de la moda de manera sencilla. Vemos que para los datos desbalanceados y balanceados las proporciones de datos no cambian de forma drástica, lo cual indica que los datos balanceados mantienen la forma con respecto a los originales.

Una vez se ha realizado un resumen general, se ha procedido a analizar cada variable una a una:

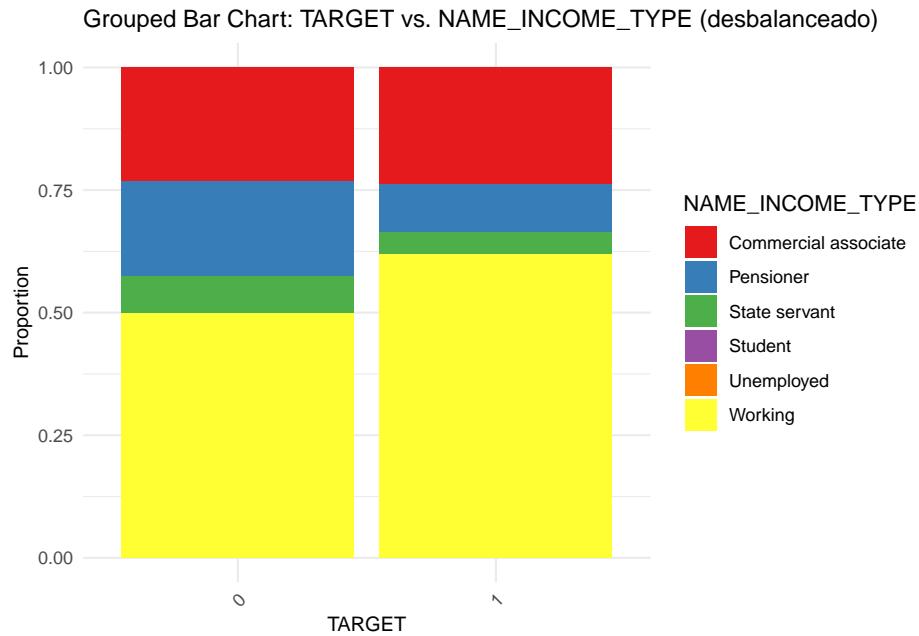
Figura 136: Pie charts de la variable target antes y después de balancear



Con respecto a la variable respuesta de nuestro estudio, se puede apreciar que originalmente se disponía de un porcentaje de morosos cercano al 8%, mientras que, análogamente, el porcentaje de no morosos es del 92%. Sin embargo, al balancear los datos, la proporción se mantiene en equilibrio: un 50% de los clientes son considerados morosos y un 50% no son morosos. Este hecho permite que a la hora de aplicar modelos clasificatorios sobre los datos, éstos sean capaces de detectar ambas clases al mismo tiempo.

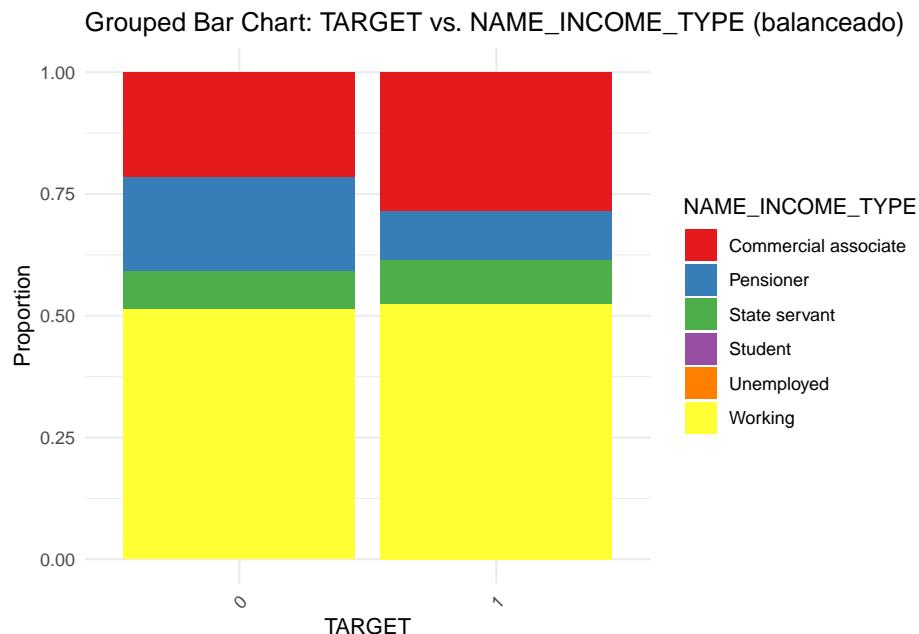
Seguidamente, se realizará un breve análisis bivariante. Como variable seleccionada para este caso, se intentará apreciar si hay diferencias entre los dos grupos de clientes y su cargo profesional:

Figura 137: Bar chart TARGET vs NAME INCOME TYPE antes de balancear



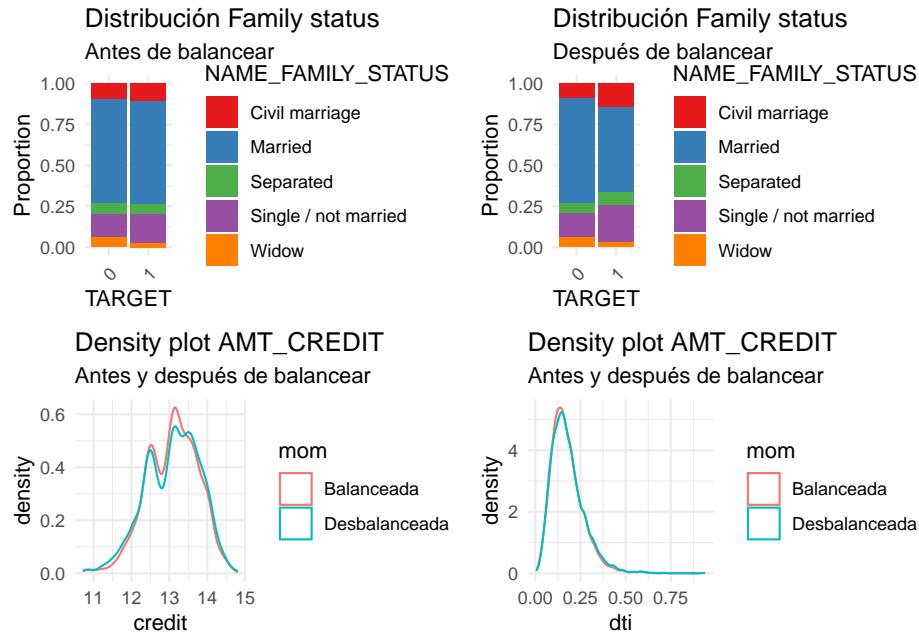
Así, en lo que respecta a la variable TARGET, se observa una disparidad en la capacidad de pago de los clientes en el sector privado, siendo los pensionistas y los comerciales quienes presentan proporcionalmente menos dificultades. Para comprobar si este patrón se mantiene al balancear la base de datos, se repite el procedimiento para los datos generados artificialmente:

Figura 138: Bar chart TARGET vs NAME INCOME TYPE después de balancear



Como se puede apreciar, parece que la distribución de esta variable respecto al target es muy similar al obtenido en los datos desbalanceados. Es por este motivo por el que se afirma que el balanceo está bien realizado, es decir, las diferencias son mínimas entre las dos bases de datos. Sin embargo, se analizará otro par de variables para ver si este hecho se mantiene:

Figura 139: Comparativa descriptiva bivariante antes y después de balancear



Como se puede apreciar, parece ser que no hay mucha diferencia entre los datos balanceados y no balanceados en cuanto a proporciones en variables categóricas se refiere. Además, ya se ha visto en el resumen inicial que las variables numéricas se mantienen bastante parecidas a lo largo del dominio de estas. Así pues, se reafirma que el balanceo de los datos se ha realizado de forma correcta.

Modelos discriminantes

A partir de este apartado, se usará nuestra base de datos con el objetivo de predecir la variable target a partir de los nuevos datos, en nuestro caso, el hecho de que un cliente se declare moroso. Para ello, se realizarán muchos modelos diferentes con el fin de predecir a cada uno de los clientes. Así pues, se comenzará por el más sencillo de todos: el LDA.

Como se preeverá, será necesario usar las dos bases de datos: la desbalanceada y la balanceada. Desde el grupo se es consciente que los resultados que se mostrarán contra la base de datos desbalanceada serán malos, ya que los modelos serán incapaces de detectar la clase minoritaria. Sin embargo, este paso es necesario para justificar que se balancea la base de datos. Así pues, se empezará con el modelo más sencillo, el Linear Discriminant Analysis (LDA).

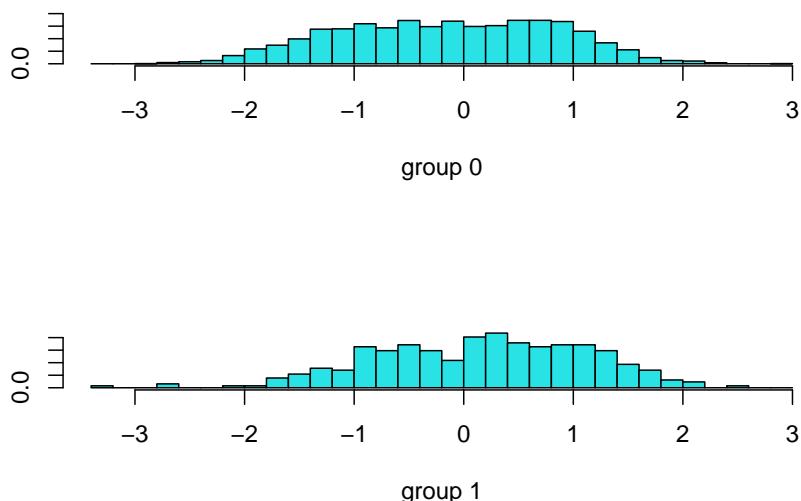
LDA (Linear Discriminant Analysis)

Para comenzar con los modelos discriminantes, se realizará en primer lugar un linear discriminant analysis (LDA) con el objetivo de intentar separar aquellos clientes que puedan tener dificultades de pago con aquellos solventes. Así pues, se procede a realizar dicho análisis discriminante.

Para ello, se recurrirá primero a un proceso de escalado de los datos a través de la función `scale()`, lo cual hará que todas las variables tengan un peso similar en la construcción del discriminante lineal. Una vez se ha realizado este proceso, el siguiente paso será realizar la partición de la base de datos disponible. Para ello, se realizará una partición clásica: el 80 % de los datos se destinarán a entrenar el modelo y el otro 20, a validar. Además, dentro de la partición del train se realizará un proceso 10-fold validation con el objetivo de reducir el overfitting y proporcionar un modelo robusto.

En el gráfico inferior se puede apreciar la proyección de cada observación sobre el discriminante:

Figura 140: Proyección de las observaciones sobre el discriminante para cada una de las clases LDA



Como se puede apreciar, los histogramas de las proyecciones se solapan entre ellos, lo cual da una idea que el

LDA no es el modelo que mejor discrimina entre las clases. Sin embargo, se realizará más adelante la matriz de confusión.

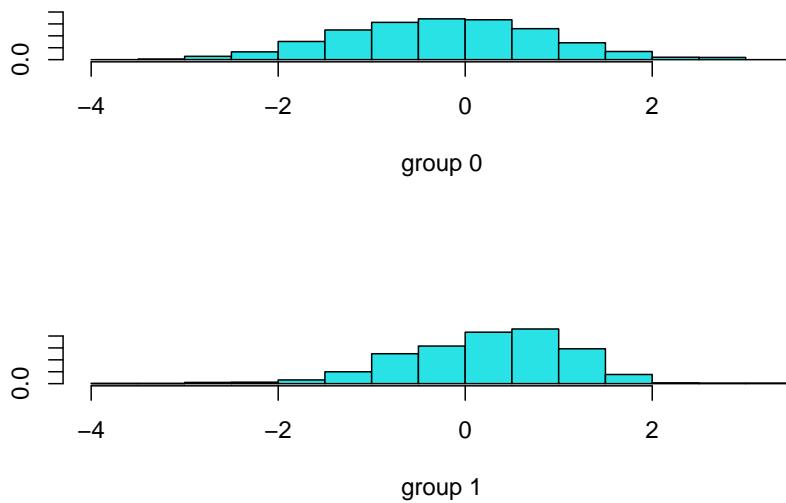
Antes de analizar los resultados obtenidos por el LDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 10-fold cross validation ha sido del 0.9197687, lo cual muestra unos resultados ciertamente pobres. Seguidamente, se ha validado el modelo contra el conjunto de validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 59: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	919	80
	Potencial moroso	0	0

Como se puede apreciar, los resultados son los esperados: al utilizar datos desbalanceados, el modelo no detecta bien la clase minoritaria, de forma que todas las predicciones de los datos llevaban a predecir todo como clientes no morosos. Así pues, se ha decidido aplicar este algoritmo a los datos ya balanceados (usando oversampling y undersampling a la vez):

Figura 141: Proyección de las observaciones sobre el discriminante para cada una de las clases LDA con datos balanceados



Cuadro 60: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	238	152
	Potencial moroso	82	88

Como se puede apreciar en los resultados de la matriz de confusión están bastante más balanceados. Apreciando los resultados obtenidos, se puede ver que la precisión obtenida por el modelo ha sido del 58.21 %, algo baja en comparación con ejemplos en otras áreas. Si desglosamos por sensibilidad y especificidad, vemos que los resultados en estos dos indicadores han sido de 36.67 % y una especificidad del 74.38 %. Así pues, el modelo ha sido capaz de detectar correctamente el 36.67 % de clientes potencialmente morosos, lo cual es un resultado muy malo para nuestro objetivo. Adicionalmente, el valor del F-score es de 0.6531, métrica que viene muy perjudicada por el problema a la hora de detectar morosos. Como otras métricas interesantes, se puede apreciar que el valor predictivo positivo es de 51.76 % y el valor predictivo negativo es de 61.03 %.

Si se testeó contra la base de datos original, se obtiene el siguiente resultado:

Cuadro 61: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	680	43
	Potencial moroso	239	37

Sin embargo, se sabe que el LDA puede presentar problemas en el momento en el que las variables no presentan normalidad o cuando las matrices de covarianzas son diferentes para cada grupo. Como ya se apreció en la descriptiva post-preprocessing, muchas de nuestras variables no presentaban normalidad, de forma que esto podría ser un problema de cara al uso del LDA. Es por eso por lo que se ha decidido realizar un QDA (Quadratic Discriminant Analysis) con el objetivo de corregir dichos problemas y mejorar la performance del LDA.

QDA (Quadratic Discriminant Analysis)

Así pues, repitiendo el procedimiento seguido anteriormente en el LDA, toca repetir los mismos pasos para este modelo. De esta forma, los resultados obtenidos son los siguientes:

Antes de analizar los resultados obtenidos por el QDA, cabe destacar que, durante el proceso de entrenamiento del modelo, el accuracy medio obtenido tras un proceso de 5-fold cross validation ha sido del 0.6511919, lo cual muestra unos resultados ciertamente pobres, pero mejores que LDA. Seguidamente, se ha validado el modelo contra el conjunto validación, con el cual se ha obtenido los siguientes resultados:

Cuadro 62: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	200	97
	Potencial moroso	120	143

Si se testeá contra los datos de validación desbalanceados, obtenemos:

Cuadro 63: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	516	36
	Potencial moroso	403	44

Como se puede apreciar, los resultados obtenidos son bastante mejores a los presentados en el discriminante lineal. De hecho, en este caso, la precisión ha sido del 56.0561 %, algo mejor que la del LDA. Sin embargo, si observamos sensibilidad y especificidad, apreciaremos que se ha obtenido una sensibilidad del 55 % (mucho mejor que LDA), pero una especificidad del 56.148 % (algo peor que el LDA). Si observamos otras métricas disponibles, apreciaremos una tasa de valores positivos predecidos de 9.8434 % y una tasa de valores negativos predecidos de 93.4783 %. Este hecho implica que al predecir una clase, la probabilidad de que ésta sea clasificada correctamente es de entorno al 60 %. Por último, podemos apreciar que el valor del F-score es de 0.5610198. Así pues, se podría decir que el modelo más útil entre estos dos es el LDA, ya que detecta de forma más consistentes el número de morosos.

En resumen, observando los resultados obtenidos, balanceando los datos se obtienen resultados más interesantes: el modelo es capaz de predecir e identificar las dos clases por igual. Sin embargo, se puede afirmar que los dos modelos discriminantes presentan resultados muy pobres: es probable que el hecho de añadir posteriormente las variables categóricas acabe de hacer que se mejore de forma clara los resultados conseguidos hasta ahora.

k-Nearest Neighbors (Base original/desbalanceada)

Con objeto de ajustar el modelo a nuestra base de datos para predecir la variable respuesta, se usará el método kNN. Para poder ajustar el modelo de manera óptima se sigue un proceso de preparación de los datos, donde se dividen en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba. El primer grupo que se utilizará para entrenar el modelo kNN estará compuesto por el 80 % de la base de datos original. Asimismo, el conjunto de prueba se empleará para evaluar el rendimiento y precisión del modelo.

La selección de un valor de K se considera un paso crucial, ya que K es un hiperparámetro en kNN que representa el número de vecinos más cercanos a considerar. Se recomienda realizar pruebas con diferentes valores de K y utilizar la validación cruzada para determinar el valor óptimo. La validación cruzada se usará dentro del conjunto de datos de entrenamiento para encontrar el valor óptimo de k.

En este caso, la realización de la Cross-validación se realiza a partir de folds. Esto consiste en dividir la base de datos perteneciente al entrenamiento en un número determinado de subgrupos aleatorios y ejecutar el algoritmo considerando como test un fold distinto en cada una de las iteraciones. En cada una de las iteraciones se calcula la precisión del algoritmo, para posteriormente calcular la media de estas precisiones.

El número de vecinos que haya tenido una media de las precisiones mayor, será el escogido.

Para el proceso de Cross-Validación se han fijado unos valores de k del 1 al 20. En cuanto al número de folds, se ha considerado oportuno utilizar una cantidad de 10 folds, lo que supone ejecutar el kNN 10 veces para cada uno de los valores de k propuestos. Esto hace que durante el proceso de Cross-Validación el kNN sea ejecutado una totalidad de 200 veces.

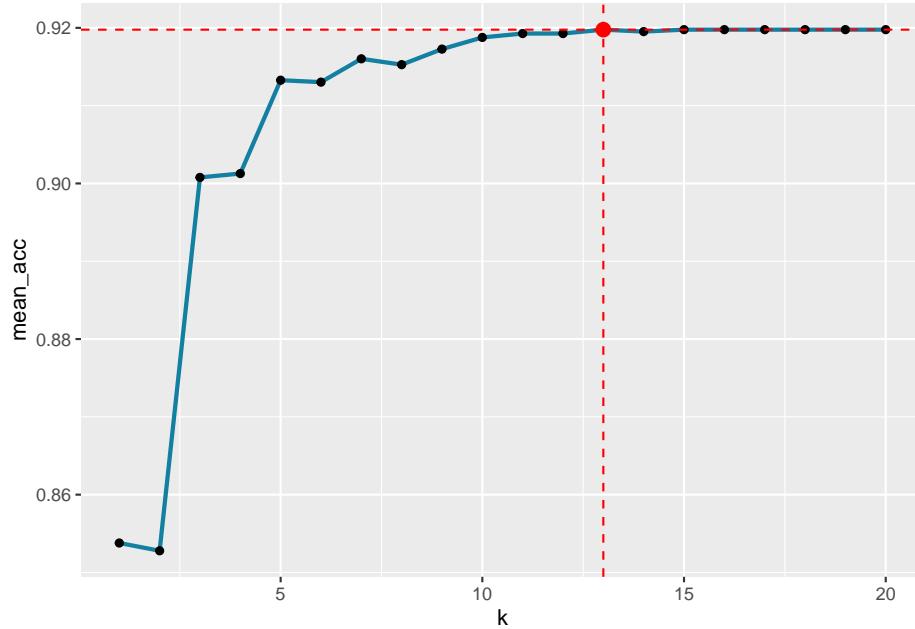
Acto seguido, se muestra una tabla en la que se recoge la media del accuracy para cada una de las k utilizadas en el proceso de Cross-Validación.

Cuadro 64: Medias de Accuracy para cada k en la CV

	k	mean_acc
k= 1	1	0.8537849
k= 2	2	0.8527874
k= 3	3	0.9007774
k= 4	4	0.9012756
k= 5	5	0.9132718
k= 6	6	0.9130231
k= 7	7	0.9160224
k= 8	8	0.9152718
k= 9	9	0.9172706
k= 10	10	0.9187706
k= 11	11	0.9192712
k= 12	12	0.9192706
k= 13	13	0.9197706
k= 14	14	0.9195206
k= 15	15	0.9197706
k= 16	16	0.9197706
k= 17	17	0.9197706
k= 18	18	0.9197706
k= 19	19	0.9197706
k= 20	20	0.9197706

Con el fin de facilitar la interpretación se reproduce un gráfico de la tabla anterior. En este se resalta la k con la que se ha conseguido un Accuracy más elevado y seguidamente se muestra su valor.

Figura 142: Media del Accuracy para cada k



Como se puede ver, la k que ha conseguido una Accuracy más elevada es la siguiente:

Cuadro 65: Accuracy de la k óptima en la CV

	k	mean acc
k= 13	13	0.9197706

Una vez terminado el proceso de Cross-Validación y habiendo encontrado la k óptima, el siguiente paso implica la implementación del algoritmo (con la k seleccionada) para predecir la categoría de la variable respuesta de los individuos del test mediante la información que proporcionan los individuos del train.

Una vez ejecutado el kNN se muestra en una tabla la matriz de confusión y se calcula la precisión con la que el algoritmo ha predicho la variable TARGET en la población del test.

Confusion Matrix and Statistics

```

    Reference
Prediction   0   1
      0 919  80
      1   0   0

Accuracy : 0.9199
95% CI  : (0.9013, 0.936)
No Information Rate : 0.9199
P-Value [Acc > NIR] : 0.5297

Kappa : 0

McNemar's Test P-Value : <2e-16

Sensitivity : 0.00000
Specificity : 1.00000
Pos Pred Value :      NaN
Neg Pred Value : 0.91992
Prevalence : 0.08008
Detection Rate : 0.00000
Detection Prevalence : 0.00000
Balanced Accuracy : 0.50000

'Positive' Class : 1

```

Recall
0

F1
NA

La anterior salida nos muestra la matriz de confusión junto con diversos estadísticos que tratan de explicar como de bien o mal ha predicho el algoritmo de kNN.

Resultados

De entre estos cabe destacar la Accuracy, que en este caso ha sido de 0.9199199, por lo que el algoritmo ha predicho correctamente el 91.991992 % de los individuos de Test.

La “Sensitivity” mide la proporción de individuos de TARGET=1 que han sido clasificados correctamente, que en este caso ha sido de 0.

Y finalmente, la “Specificity” mide la proporción de individuos de TARGET=0 que han sido clasificados correctamente, que ha dado 1

Conclusiones

El “modelo” muestra una alta especificidad para la clase 0, es decir, identifica correctamente la mayoría de clientes no morosos. Sin embargo, tiene un desempeño deficiente en la clase 1, sin predecir correctamente ninguna cliente moroso.

El estadístico Kappa y la sensibilidad para la clase 1 sugieren que el rendimiento del modelo es deficiente y posiblemente peor que el azar.

En una base de datos desbalanceada, la falta de equilibrio entre las clases nos lleva a tener una información limitada sobre los clientes morosos. Esto se ve eclipsado por la abrumadora cantidad de datos disponibles para los clientes no morosos. En consecuencia, el modelo KNN, al ser simplista (o “lazzy”), tiende a etiquetar a la mayoría de los clientes como no morosos. Esto resulta en predicciones poco precisas, ya que el modelo no logra distinguir claramente entre las dos clases.

Finalmente, la falta de valores para el recall de la clase 0 y un F1-Score como “NaN” sugieren que hay problemas en el rendimiento del modelo, especialmente en la identificación de clientes morosos.

Por lo tanto, consideramos que el desempeño de este algoritmo es deficiente para lograr el objetivo establecido de predecir correctamente a clientes morosos y no morosos.

k-Nearest Neighbors (Base balanceada)

Como se ha podido ver en el apartado anterior, el algoritmo kNN no funciona correctamente para la base de datos original, ya que la cantidad de individuos no morosos es muy superior a la cantidad de individuos morosos. Con el objetivo de ver una mejora en el funcionamiento del algoritmo, se aplicará a la base de datos una vez se ha realizado el balanceo.

El procedimiento a seguir es exactamente el mismo que con la base de datos original. Por lo que primero se realiza la Cross-Validación con el objetivo de detectar el número de vecinos próximos óptimo y se continuará con la validación del “modelo” en el conjunto de datos Test.

El hecho de usar la base de datos balanceada tiene como consecuencia unos resultados distintos, los cuales se espera que sean mejores que los obtenidos con la base de datos original, debido al desbalanceo que está presentaba.

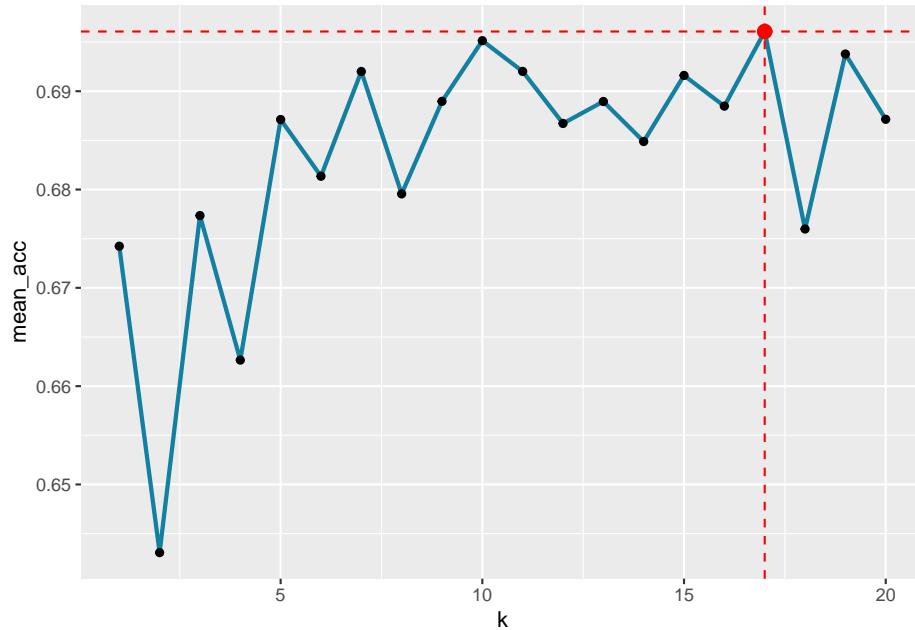
Acto seguido, se muestra una tabla en la que se recoge la media del accuracy para cada una de las k utilizadas en el proceso de Cross-Validación.

Cuadro 66: Medias de Accuracy para cada k en la CV

	k	mean_acc
k= 1	1	0.6742321
k= 2	2	0.6430682
k= 3	3	0.6773519
k= 4	4	0.6626495
k= 5	5	0.6871219
k= 6	6	0.6813527
k= 7	7	0.6920077
k= 8	8	0.6795564
k= 9	9	0.6889702
k= 10	10	0.6951423
k= 11	11	0.6920148
k= 12	12	0.6867209
k= 13	13	0.6889557
k= 14	14	0.6848838
k= 15	15	0.6916052
k= 16	16	0.6884821
k= 17	17	0.6960749
k= 18	18	0.6759828
k= 19	19	0.6937846
k= 20	20	0.6871456

Con el fin de facilitar la interpretación se reproduce un gráfico de la tabla anterior. En este se resalta la k con la que se ha conseguido un Accuracy más elevado y seguidamente se muestra su valor.

Figura 143: Media del Accuracy para cada k



Como se puede ver, la k que ha conseguido una Accuracy más elevada es la siguiente:

Cuadro 67: Accuracy de la k óptima en la CV

	k	mean_acc
k= 17	17	0.6960749

Una vez terminado el proceso de Cross-Validación y habiendo encontrado la k óptima, el siguiente paso implica la implementación del algoritmo (con la k seleccionada) para predecir la categoría de la variable respuesta de los individuos del test mediante la información que proporcionan los individuos del train.

Una vez ejecutado el kNN se muestra en una tabla la matriz de confusión y se calcula la precisión con la que el algoritmo ha predicho la variable TARGET en la población del test balanceado. Esto con el objetivo de comprobar si hay overfitting.

Confusion Matrix and Statistics

```

      Reference
Prediction   0   1
      0 259 129
      1  61 111

      Accuracy : 0.6607
      95% CI  : (0.6198, 0.6999)
No Information Rate : 0.5714
P-Value [Acc > NIR] : 9.63e-06

      Kappa : 0.2819

McNemar's Test P-Value : 1.17e-06

      Sensitivity : 0.4625
      Specificity : 0.8094
      Pos Pred Value : 0.6453
      Neg Pred Value : 0.6675
      Prevalence : 0.4286
      Detection Rate : 0.1982
      Detection Prevalence : 0.3071
      Balanced Accuracy : 0.6359

      'Positive' Class : 1
  
```

Recall
0.4625

F1
0.538835

Como se puede apreciar en la matriz de confusión, el Accuracy obtenido con el algoritmo kNN aplicado el test balanceado toma un valor parecido al obtenido durante la Cross-Validación. De esta manera, se comprueba que no hay overfitting.

A continuación, se aplica el algoritmo sobre el test original (desbalanceado).

Confusion Matrix and Statistics

```

      Reference
Prediction   0   1
      0 751 53
      1 168 27
  
```

```

Accuracy : 0.7788
95% CI : (0.7517, 0.8042)
No Information Rate : 0.9199
P-Value [Acc > NIR] : 1

Kappa : 0.0934

McNemar's Test P-Value : 1.741e-14

Sensitivity : 0.33750
Specificity : 0.81719
Pos Pred Value : 0.13846
Neg Pred Value : 0.93408
Prevalence : 0.08008
Detection Rate : 0.02703
Detection Prevalence : 0.19520
Balanced Accuracy : 0.57735

'Positive' Class : 1

```

Recall
0.3375

F1
0.1963636

La anterior salida nos muestra la matriz de confusión junto con diversos estadísticos que tratan de explicar como de bien o mal ha predicho el algoritmo de kNN.

De entre estos cabe destacar la Accuracy, que en este caso a sido de 0.7787788, por lo que el agloritmo ha predicho correctamente el 77.8778779 % de los individuos de Test (desbalanceado).

La “Sensitivity” mide la proporción de individuos de TARGET=1 que han sido clasificacdos correctamente, que en este caso ha sido de 0.3375.

Y finalmente la “Specificity” mide la proporción de individuos de TARGET=0 que han sido clasificados correctamente, que ha dado 0.8171926.

Conclusiones

Como se ha podido apreciar, el algoritmo kNN no genera overfitting, lo cual indica que encuentra un balance entre los datos y evitar el sobre ajuste. Al evitar el sobreajuste, es más probable que un modelo capte los patrones subyacentes en los datos y haga predicciones fiables sobre instancias nuevas y no vistas.

Si nos fijamos en la Accuracy esta toma un valor cercano al 0,78, lo que supone que el algoritmo predice bien el 78 % de los datos test. Si nos centramos en la Specificity, vemos que toma un valor cercano al 0,82, de manera que predice correctamente el 82 % de los clientes no morosos. Sin embargo, no pasa lo mismo con la Sensitivity, que da 0,34 aproximadamente, de manera que solamente se estaria prediciendo de manera correcta el 34 % de los clientes morosos.

Estos resultados vistos desde el punto de vista del banco son bastante pobres, ya que el principal objetivo del banco debería ser la correcta identificación de los clientes con altas posibilidades de impago.

Naive Bayes

A continuación se procede a aplicar el algoritmo de Naive Bayes a los datos balanceados para saber cual es el nivel de precisión de este algoritmo respecto a la cartera balanceada. Este algoritmo es un método de clasificación supervisada basado en el teorema de Bayes, que utiliza la probabilidad condicional para predecir la clase de un conjunto de datos en función de las características observadas. Este algoritmo se basa en el Teorema de Bayes:

$$P(A|B) = P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Donde:

- $P(A|B)$ es la probabilidad de A dado B.
- $P(B|A)$ es la probabilidad de B dado A.
- $P(A)$ y $P(B)$ son las probabilidades marginales de A y B, respectivamente.

Naive Bayes asume independencia condicional entre las variables predictoras, lo que significa que la presencia o ausencia de una característica no afecta la presencia o ausencia de otras características. Aun así, con el uso de kernels se permite relajar la suposición de independencia condicional, que es la suposición “ingenua” (naive) en Naive Bayes. Esto puede ser beneficioso cuando hay dependencias entre las variables predictoras.

En cuanto a la base de datos, es relevante mencionar que la naturaleza del negocio de una cartera de clientes es detectar correctamente los posibles morosos, es decir, se busca que haya pocos falsos negativos. Esto se debe a que un falso negativo es equivalente a un cliente que será moroso pero que el algoritmo no ha sido capaz de detectar que lo va a ser, hecho que altera negativamente la calidad de la cartera. Por otro lado, las consecuencias de los falsos positivos no es tan negativa, pues se estaría alarmaando de un cliente que no va a causar problemas. Además, en los tipos de algoritmos de predicción es muy recomendable hacer la validación cruzada, pues se consiguen unos resultados mas fiables al hacer la media de las metricas de 10 validaciones distintas, usando todos los datos en train y test en alguna de las particiones. Por eso, se empleará el cross validation con 10 particiones.

A continuación se procede a ejecutar el algoritmo e interpretar los resultados.

Cuadro 68: Matriz de confusión de Naive Bayes de datos desbalanceados

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	919	80
	Potencial moroso	0	0

En un primer análisis visual se aprecia como en la matriz de confusión resultante de los datos desbalanceados casi no hay sujetos como potenciales morosos, mientras clasifica al resto como no morosos. Cabe destacar que, en el proceso de entrenamiento, el accuracy obtenido por la CV ha sido del 91.98, resultado muy elevado.

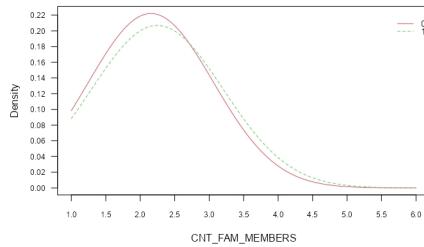
En este caso, es importante estudiar la especificidad del modelo, es decir, la capacidad del modelo de clasificar correctamente a los potenciales morosos. En este caso, dado que la base de datos esta desbalanceada, la sensitividad del modelo es 0 %. Por otro lado, valores como el accuracy o la especificidad no son relevantes, pues el accuracy mide la capacidad del modelo de clasificar correctamente los verdaderos positivos y los verdaderos negativos respecto al total, pero, dado que la especificidad es casi perfecta (capacidad del modelo de clasificar correctamente a los no morosos), siendo el 100 %, el accuracy resultante es muy elevado, pues al ser una base de datos desbalanceada casi todos los sujetos son no morosos. En el caso que los datos fueran balanceados se deberían estudiar los diferentes ratios que se ofrecen en la tabla, pero debido a que los datos son desbalanceados y que la naturaleza de los datos requieren de una correcta clasificación de las personas morosas.

Así pues, se comprueba como una base de datos desbalanceada no es eficiente para estudiar la predicción de la clasificación de los clientes y, por ende, se requiere de un balanceo de los datos para poder tener un algoritmo capaz de discriminar mejor ambos tipos de clientes.

Seguidamente se repite el proceso anterior, esta vez con los datos balanceados.

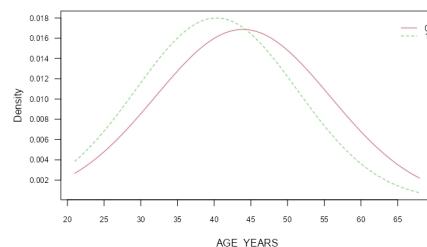
Inicialmente es necesario analizar algunos gráficos, resultado del algoritmo, para poder luego explicar correctamente y con mejor entendimiento de las características de ambos grupos de clientes y las clasificaciones de los clientes.

Figura 144: Gráfico de densidad de los clientes morosos y no morosos respecto al número de familiares



En el gráfico superior, se evidencia que la cantidad de familiares de los clientes morosos tiende a ser superior en comparación con aquellos clientes que no enfrentan dificultades de pago. Esto se refleja en la mayor concentración de la densidad de los clientes morosos en valores menores.

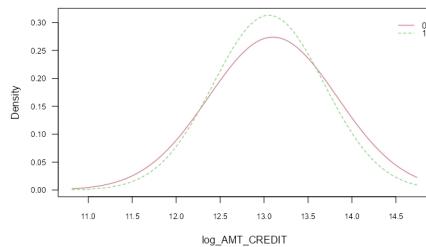
Figura 145: Gráfico de densidad de los clientes morosos y no morosos respecto a la edad del sujeto



En el gráfico superior se representa la distribución de las edades entre los clientes con y sin morosidad. Como se puede observar, los clientes no morosos tienden a tener edades más avanzadas en comparación con aquellos que enfrentan dificultades de pago, quienes generalmente son más jóvenes.

Este gráfico ilustra cómo la distribución del crédito de clientes, aquellos con y sin morosidad, exhibe similitudes con una distribución normal después de aplicar una transformación logarítmica. La distribución de los clientes con dificultades de pago muestra una mayor leptocurtosis, mientras que la de los clientes sin

Figura 146: Gráfico de densidad de los clientes morosos y no morosos respecto al crédito concedido



dificultades de pago es más plástica. Aunque este fenómeno puede tener relevancia estadística, por sí solo, carece de implicaciones económicas significativas. Sería necesario examinar este patrón junto con otras variables para identificar la relación entre variables.

A continuación se analizarán los resultados del algoritmo de Naive Bayes, esta vez con la base de datos balanceados.

Cuadro 69: Matriz de confusión de Naive Bayes de datos balanceados

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	206	91
	Potencial moroso	114	149

Si se prueba contra los datos desbalanceados, se obtienen los siguientes resultados:

Cuadro 70: Matriz de confusión de Naive Bayes de datos desbalanceados

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	575	40
	Potencial moroso	344	40

Se aprecia, a primera vista, como la matriz de confusión resultante de los datos balanceados es mucho más homogénea y parece tener una mejor capacidad predictiva. A diferencia de la base de datos desbalanceada, esta vez sí que es relevante estudiar más en profundidad más ratios.

En primer lugar, el accuracy cobra sentido cuando los datos están balanceados, pues no hay sobrerepresentación de ningún tipo de clientes. En este segundo algoritmo, el accuracy es del 61.56 %, un valor muy bueno considerando la naturaleza del negocio. Por otro lado, la especificidad es 62.57 %, es decir, la capacidad del modelo de detectar los verdaderos clientes que no van a tener problemas para realizar los pagos. Por lo tanto, el modelo es incapaz de clasificar correctamente a los no morosos. Por otro lado, la especificidad es 50 %, siendo ésta la capacidad de detectar los clientes que van a tener problemas para realizar los pagos. En cuanto a los valores correctamente predecidos, el 10.42 % de los valores positivos predecidos como positivos son correctamente positivos mientras que el 93.5 % de los valores considerados negativos son realmente negativos.

Los resultados proporcionados por este algoritmo son muy negativos, pues hay un gran descuadre entre la sensibilidad y la especificidad. Este hecho se debe a que el algoritmo detecta más unos que ceros siempre, debido probablemente a la realización de un oversampling en el balanceo de los datos. Este hecho es debido a que como naive bayes selecciona individuos similares para el cálculo de las probabilidades a posteriori, el hecho de generar muestras similares a las presentes hace que se sobrerepresente estos individuos. Así pues, estos resultados pueden no ser muy fiables debido a la propia forma de balancear los datos. Sin embargo, el hecho de que clasifique relativamente bien los individuos morosos hace que sea un algoritmo fiable de cara al objetivo del propio proyecto.

Support Vector Machine (SVM)

En este apartado se aborda el modelo predictivo llamado Support Vector Machine (SVM). Se trata de un algoritmo de machine learning supervisado, y usado para funciones de clasificación y regresión.

El objetivo de SVM es el de encontrar un hiperplano que mejor separe las clases sobre los datos de nuestra variable respuesta. Para encontrar este hiperplano óptimo, el cual separa bien nuestros datos y a la vez maximiza el margen (distancia entre hiperplano y puntos más cercanos a él de cada clase), en muchos casos hay que aumentar la dimensionalidad, llegando a dimensiones que no pueden representarse gráficamente, pero que sí permiten una correcta discriminación entre clases.

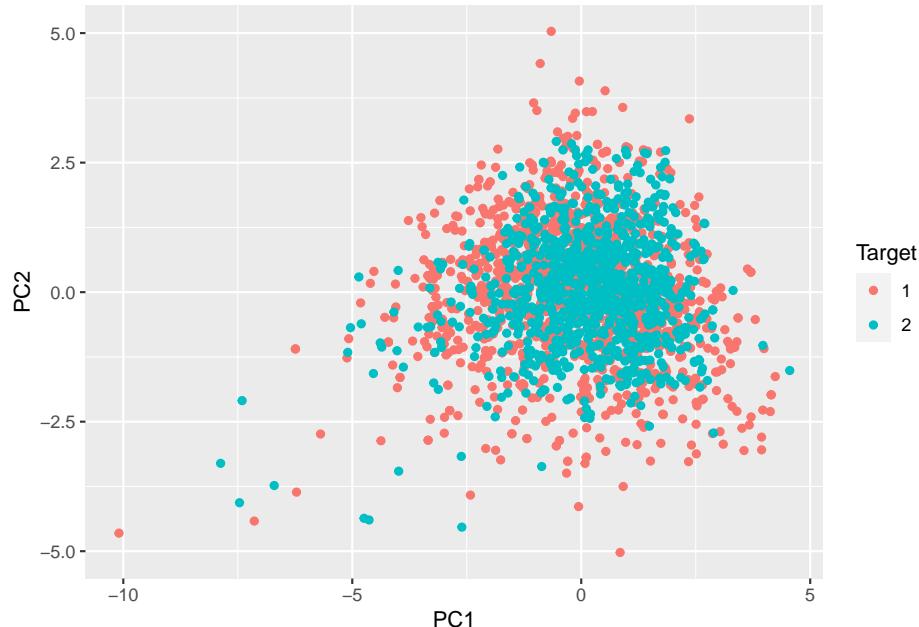
Obtener datos

Primero, se carga la base de datos balanceada, y se estandariza para evitar problemas derivados de la diferencia de escalas entre las variables. También se convierte la variable respuesta a factor, ya que se trata de un problema de clasificación.

Es trivial el hecho de que en este caso, los datos no pueden ser separados linealmente, dado el número de variables presentes, que son 16. Por esta razón, y dependiendo de la función kernel utilizada (y de sus parámetros), SVM utilizará espacios dimensionales transformados a partir del número de variables de la base de datos inicial. En otras palabras, el kernel define la manera como los datos se transforman en el nuevo espacio dimensional, y la dimensionalidad de dicho espacio resultante quedará determinada por los parámetros del kernel.

Para escoger el kernel se grafica las dos primeras dimensiones del PCA.

Figura 147: PC1 y PC2 respecto la variable Target



Como se puede ver, haciendo una representación del PCA de la primera dimensión y la segunda (PC1 y PC2) separando por colores la variable respuesta *TARGET*, siendo rojo (1) los clientes considerados como

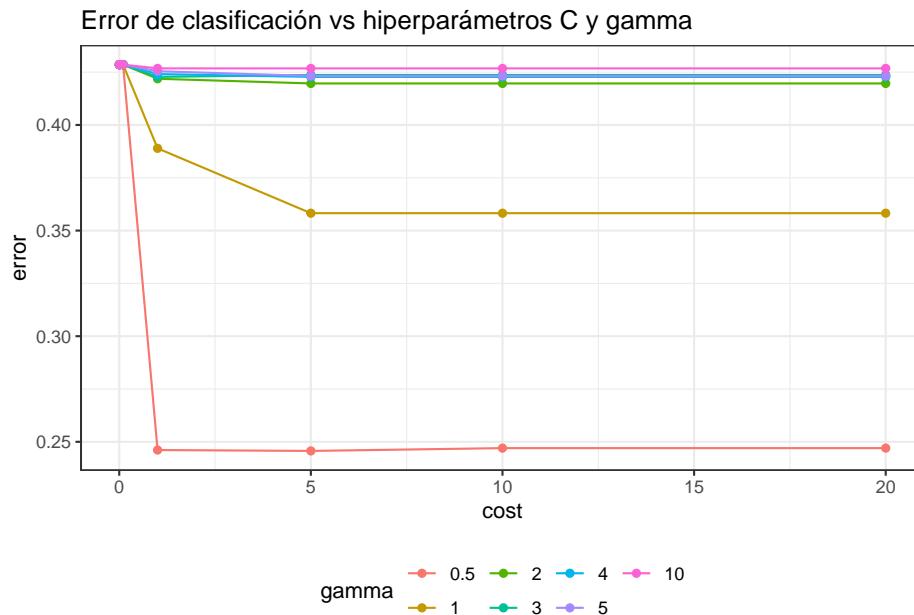
no morosos y azul (2) los potencialmente morosos, se puede concluir que parece que el grupo de no morosos se encuentra dentro del grupo de morosos formando prácticamente dos formas redondeadas en ambos casos. Al ver esta relación, se ha concluido que el kernel más adecuado en este caso era el radial.

Encontrar los valores de los hiperparámetros C (coste) y Gamma

Para poder determinar los valores óptimos de los hiperparámetros, se debe hacer una *validación cross-fold*. Una vez se encuentren estos valores óptimos, se ejecutará el SVM con ellos (es decir, el mejor modelo) para el conjunto de datos de validación. Esto nos permite obtener las métricas de rendimiento de este modelo sobre los datos, y poder compararlas con las de otros algoritmos de clasificación.

A continuación se usa la función *tune* para encontrar, dentro de una lista predefinida de valores para cada hiperparámetro, la combinación que resulte con el modelo con mejor rendimiento. En el gráfico siguiente se muestra una comparativa del error para cada combinación de hiperparámetros:

Figura 148: Representación del error en función de C y Gamma



Con el gráfico anterior se debería poder tener una idea, a nivel visual, del valor del hiperparámetro *gamma* que minimiza el error. En nuestro caso, la conclusión es muy clara, siendo el valor de 0.5 el que minimiza el error con mucha diferencia respecto a los demás valores. También se pueden obtener los valores óptimos con la siguiente instrucción seleccionando el objeto *best.parameters*, de la función *tune()*. El resultado es el siguiente:

Cuadro 71: Hiperparámetros óptimos

	cost	gamma
Valores	5	0.5

Por tanto, los valores óptimos para los dos hiperparámetros son los anteriores. A continuación se puede ver más información sobre el mejor modelo encontrado en el *cross-fold validation*.

Se puede apreciar como el número de vectores de soporte, aquellos puntos más cercanos al límite de decisión y que definen la posición del hiperplano, es elevado, contando con 2042.

Este fenómeno puede estar causado por varios motivos. Podría ser que la frontera de decisión entre las clases es inherentemente compleja, por lo que se necesitan muchos datos para representar de forma precisa la separación entre las dos clases. También podría haber ocurrido por *overfitting*, situación en la que el modelo tiene muy buen rendimiento con el conjunto de datos de entrenamiento, pero mal rendimiento con datos nuevos. El *overfitting* ocurre cuando el modelo es demasiado complejo en relación con la cantidad de datos disponibles para entrenarlo.

A pesar de esto, con el mejor modelo procedemos a hacer la predicción con la base de datos de validación, para obtener las métricas del rendimiento del modelo obtenido.

A continuación se muestra la matriz de confusión para evaluar dichas métricas y la capacidad predictiva del modelo. Hace falta aclarar que en esta matriz, el 0 representa a los “No morosos” y el 1 a los “Morosos”. En los comentarios subsecuentes, “negativo” es 0 y “positivo” es 1.

Cuadro 72: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	784	10
	Potencial moroso	135	70

Las matrices de confusión se utilizan comúnmente comúnmente en problemas de clasificación para evaluar el rendimiento de un modelo. En este caso, se trata con un modelo que clasifica clientes en dos categorías: no morosos (clase 0) y potencialmente morosos (clase 1). A continuación se interpretarán los elementos que componen esta matriz:

- **Verdaderos positivos (TP): 70.** Esto significa que 70 clientes fueron correctamente clasificados como morosos.
- **Verdaderos negativos (TN): 784.** Indica que 784 clientes fueron correctamente clasificados como no morosos.
- **Falsos positivos (FP): 135.** Representa a clientes que fueron incorrectamente clasificados como morosos cuando en realidad no lo eran.
- **Falsos negativos (FN): 10.** Muestra la cantidad de clientes que fueron incorrectamente clasificados como no morosos cuando en realidad sí lo eran .

Además, también vemos valores que nos indican la Sensibilidad , que es la proporción de positivos reales que se identificaron correctamente.

$$Sensitivity = \frac{TP}{TP+FN} = \frac{70}{70+10} \approx 0,8750$$

Tmbién la proporción de negativos reales que se identificaron correctamente como tales o especificidad.

$$Specificity = \frac{TN}{TP+FP} = \frac{70}{70+135} \approx 0,8531$$

Por último, destacar el valor de Accuracy o exactitud, que es la proporción de predicciones que el modelo clasificó correctamente.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{70+784}{70+784+10+135} \approx 0,8549$$

Para detectar si existe la presencia de overfitting en el modelo se debe comparar los valores de accuracy para el test y el train con los datos balanceados. El accuracy con los valores de training toma un valor de 0.7521, mientras que el accuracy con los valores de test toma el valor de 0.7286. Este hecho nos demuestra que, pese a realizar cross-validation, el modelo presenta un ligero overfitting sobre los datos train. Sin embargo, los resultados obtenidos son muy positivos para un modelo clasificadorio.

Árboles de Decisión

Siguiendo con los modelos predictivos, en este apartado se analizará el algoritmo de los Árboles de Decisión, CART en adelante, con el mismo propósito específico: la clasificación de clientes en categorías de riesgo crediticio. En particular, nos enfocaremos en discernir entre aquellos clientes que puedan tener dificultades de pago y aquellos que son financieramente solventes.

El algoritmo de Árboles de Decisión se revela como una herramienta particularmente poderosa en este contexto, ya que su capacidad para modelar relaciones complejas entre variables puede proporcionar insights para la toma de decisiones financieras. Exploraremos cómo el algoritmo selecciona de manera inteligente las variables más influyentes para segmentar eficientemente el conjunto de datos, permitiendo la identificación de patrones que podrían indicar riesgos financieros.

Algoritmo

En este contexto, la estructura de un Árbol de Decisión se modela de forma análoga a un proceso de decisiones estratégicas:

- Cada nodo interno del árbol representa una evaluación crítica sobre un atributo financiero específico. Estas evaluaciones sirven como puntos clave para discernir las distintas condiciones financieras de los clientes.
- Las ramas que se desprenden de cada nodo interno representan las diferentes trayectorias que un cliente puede seguir según el resultado de la evaluación realizada en ese nodo.
- Las hojas del árbol en el contexto financiero contienen la información crucial: la etiqueta o el valor predicho relacionado con la capacidad del cliente para afrontar compromisos financieros. Esto puede manifestarse como una clasificación de riesgo, como “solvente” o “en riesgo”, proporcionando una guía clara para las decisiones crediticias.

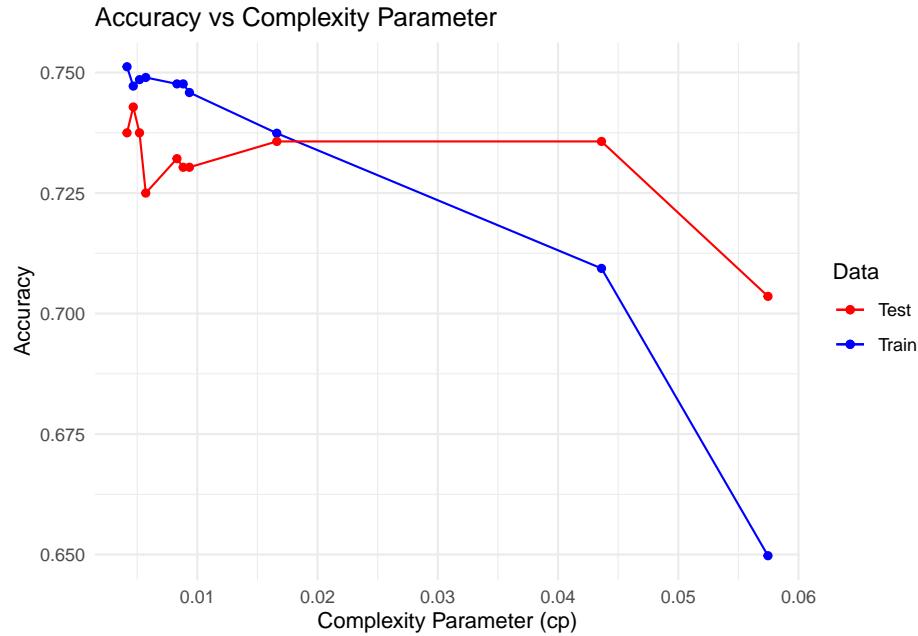
Así pues, a continuación se procede a realizar dicho análisis predictivo.

Desarrollo del CART

Para iniciar el desarrollo del modelo, el primer paso es encontrar el valor óptimo del complexity parameter, o parámetro de complejidad, que controla la cantidad de ramificaciones y nodos terminales en el árbol. Este parámetro juega un papel importante en la regularización del árbol, evitando que éste se vuelva demasiado complejo y se adapte demasiado a los datos de entrenamiento, lo que podría resultar en un sobreajuste del modelo.

Entonces, para encontrar este valor óptimo del parámetro de complejidad, se entrena el modelo con los datos balanceados de Train y se realiza un proceso de crosvalidación con 10 folds. Entonces calculamos el accuracy para cada 10 valores posibles del complexity parameter tanto para los datos train como test.

Figura 149: Evolución de la precisión (obtenida mediante validación cruzada) dependiendo del parámetro de complejidad



En el gráfico se observa como el primer valor del complexity parameter es el que reporta un mayor accuracy para el conjunto de datos de entrenamiento. No obstante, no únicamente buscamos el hiperparámetro que nos aporte un mayor accuracy, sino que también nos interesa encontrar un cp con el que además de maximizar el accuracy, evitemos overfitting (sobreajuste del modelo). Así pues, observamos como el segundo valor del complexity parameter nos da un valor que no se aleja mucho del accuracy óptimo y, además, nos evita en una gran medida un overfitting. Por lo tanto, concluimos que el cp óptimo para nuestro árbol de decisión final es 0.0046729.

Validación del modelo

Una vez ejecutado el modelo CART, con el objetivo de validar el modelo, se muestra en una tabla la matriz de confusión y se calcula la precisión con la que el algoritmo ha predicho la variable Target tanto en la población del Train como en la del Test, para observar si ha habido un sobreajuste o no.

Cuadro 73: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	295	120
	Potencial moroso	25	120

Cuadro 74: Medidas de Validación para el modelo CART

	Train	Test
Accuracy	0.7894971	0.7410714
Sensitivity	0.5711319	0.5000000
Specificity	0.9532710	0.9218750
Recall	0.5711319	0.5000000
F1	0.6993007	0.6233766
Precision	0.9016393	0.8275862

La anterior salida nos muestra la matriz de confusión junto con diversos estadísticos que tratan de explicar como de bien o mal ha predicho el algoritmo de CART. Así pues, como se observa, las medidas de validación son aproximadamente las mismas tanto para Train como para Test, por lo tanto, reafirmamos que no hay un sobreajuste en el modelo (como ya se había dicho anteriormente).

En este caso, la precisión ha sido del 0.7411 %, lo que indica que el algoritmo ha predicho correctamente el 74.1071 % de los individuos de Test. Esto indica que el modelo es capaz de clasificar correctamente a la mayoría de los clientes en categorías de riesgo crediticio.

La sensibilidad del modelo, que mide la capacidad de identificar clientes potencialmente morosos, es del 50 % en el conjunto de prueba. Esto sugiere que hay margen para mejorar en la identificación de clientes con dificultades de pago.

Por otro lado, el modelo muestra una alta especificidad, del 92.1875 %, indicando su habilidad para identificar clientes no morosos con precisión.

Si observamos otras métricas disponibles, apreciaremos como la precisión, que mide la exactitud de las predicciones positivas, es del 82.7586 % en el conjunto de prueba. Esto significa que cuando el modelo predice que un cliente es potencialmente moroso, es correcto en aproximadamente el 82.76 % de las veces. Por último, podemos apreciar como la puntuación F1, que equilibra precisión y recuperación, es del 62.3 % en el conjunto de prueba, indicandonos que el modelo logra un buen equilibrio entre la precisión de las predicciones positivas y la capacidad para recuperar casos positivos.

En resumen, el modelo muestra un buen rendimiento general, especialmente en términos de especificidad, pero hay margen para mejorar en la identificación de clientes potencialmente morosos, como lo sugiere la sensibilidad y la puntuación F1 en ambos conjuntos.

Prueba ácida

Cuadro 75: Matriz de confusión del conjunto de validación desbalanceado

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	856	68
	Potencial moroso	63	12

Cuadro 76: Medidas de Validación con el conjunto test desbalanceado para el modelo CART

	Train	Test_desbalanceado
Accuracy	0.7894971	0.8688689
Sensitivity	0.5711319	0.1500000
Specificity	0.9532710	0.9314472
Recall	0.5711319	0.1500000
F1	0.6993007	0.1548387
Precision	0.9016393	0.1600000

El accuracy del modelo en el conjunto de prueba desbalanceado ha sido del 0.8689 %, lo que indica que el 0.8689 % de las predicciones fueron buenas. Sin embargo, la exactitud puede ser engañosa en conjuntos de datos desbalanceados, donde la mayoría de las observaciones pertenecen a una clase particular. Por otra parte, la sensibilidad en el conjunto de prueba desbalanceado es bastante baja, solo del 15 %. Esto significa que el modelo tiene dificultades para identificar correctamente a los clientes morosos. La sensibilidad es especialmente crucial en situaciones financieras, ya que representa la capacidad del modelo para capturar la totalidad de los casos positivos (morosos) reales. En este caso, el bajo valor de sensibilidad indica que el modelo está dejando pasar un número significativo de clientes morosos sin detectarlos.

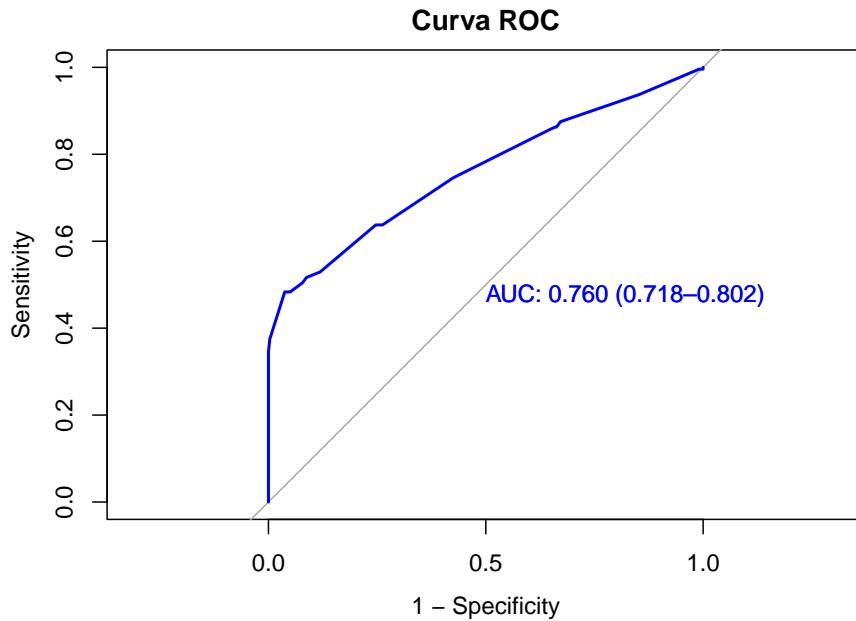
Así pues, la especificidad que mide la capacidad del modelo para identificar correctamente los casos negativos (clientes no morosos), es del 93.14 %. Esto sugiere que el modelo tiene un buen rendimiento al identificar a los clientes que no son morosos. Sin embargo, es importante destacar que la alta especificidad podría deberse al desbalance en los datos, ya que hay más clientes no morosos en el conjunto de prueba.

Finalmente, el valor de F1 es del 15.48 %, lo que refleja un equilibrio entre precisión y sensibilidad. Este valor relativamente bajo sugiere que hay margen de mejora en la capacidad del modelo para identificar clientes morosos sin comprometer demasiado la precisión. De la misma manera, en cuanto a la precisión (Precision), es del 16 %, lo que significa que de las instancias que el modelo predice como morosas, solo el 16 % son realmente morosas. Este valor puede ser bajo, indicando que el modelo podría estar generando demasiados falsos positivos.

Curva ROC

Para un análisis más profundo sobre la calidad de predicción del modelo, se representa la curva ROC y se interpreta su área bajo la curva (AUC).

Figura 150: Curva ROC

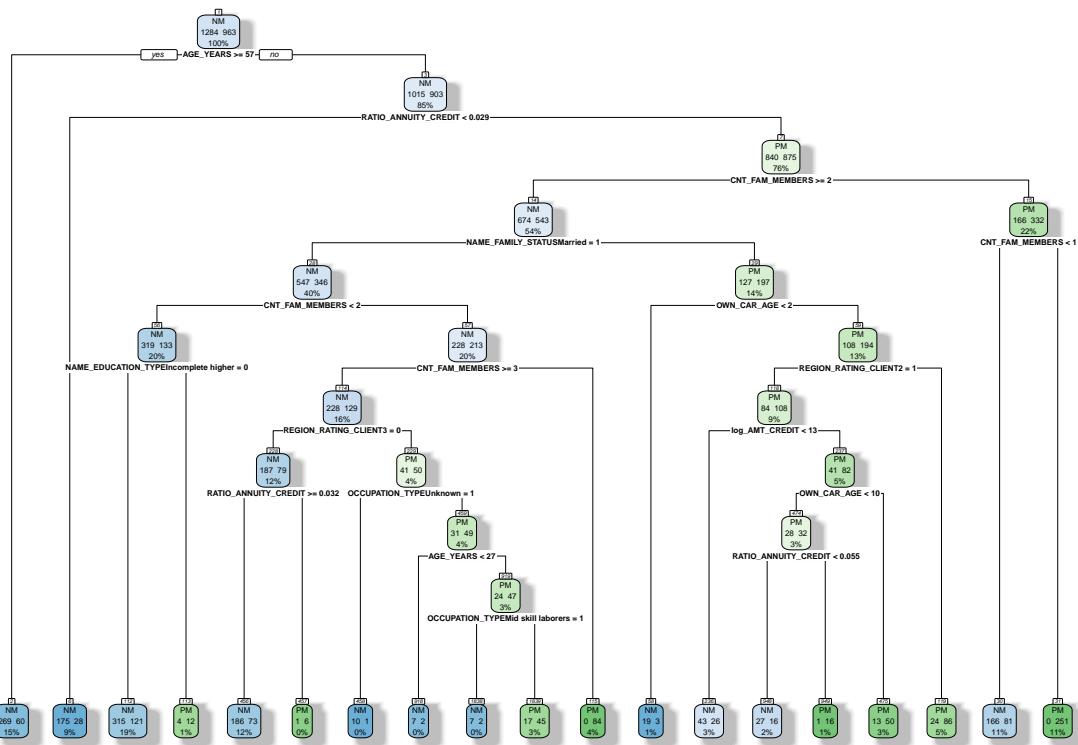


El AUC (Área Bajo la Curva) de 0.760 en la curva ROC sugiere que el modelo tiene una capacidad moderadamente buena para distinguir en la clasificación binaria entre morosos y no morosos. En otras palabras, el modelo es mejor que una clasificación aleatoria, es prometedor y sugiere que el modelo tiene un rendimiento decente en términos de discriminación. Sin embargo, hay un pequeño margen para mejorar.

Árbol de decisión

A continuación, se presenta el árbol de decisión final con el parámetro de complejidad óptimo elegido.

Figura 151: Árbol de clasificación de la variable TARGET, obtenido con la complejidad ‘óptima’



El árbol de decisión generado se inicia evaluando la edad del solicitante. Si la edad es mayor o igual a 57 años, el modelo tiende a clasificar al individuo como “No Moroso” con un accuracy del 85 %. Este primer nivel de decisión sugiere que la edad es un factor determinante en la predicción de la no morosidad.

Por otro lado, dentro de la categoría de clientes más jóvenes, el árbol se ramifica según el ratio anualidad/crédito (RATIO_ANNUITY_CREDIT). Aquellos con un ratio inferior a 0.0295, se los clasifica como no morosos con un accuracy del 91 %, sugiriendo que clientes con cargas de anualidad más bajas en comparación con su crédito son menos propensos a tener dificultades de pago.

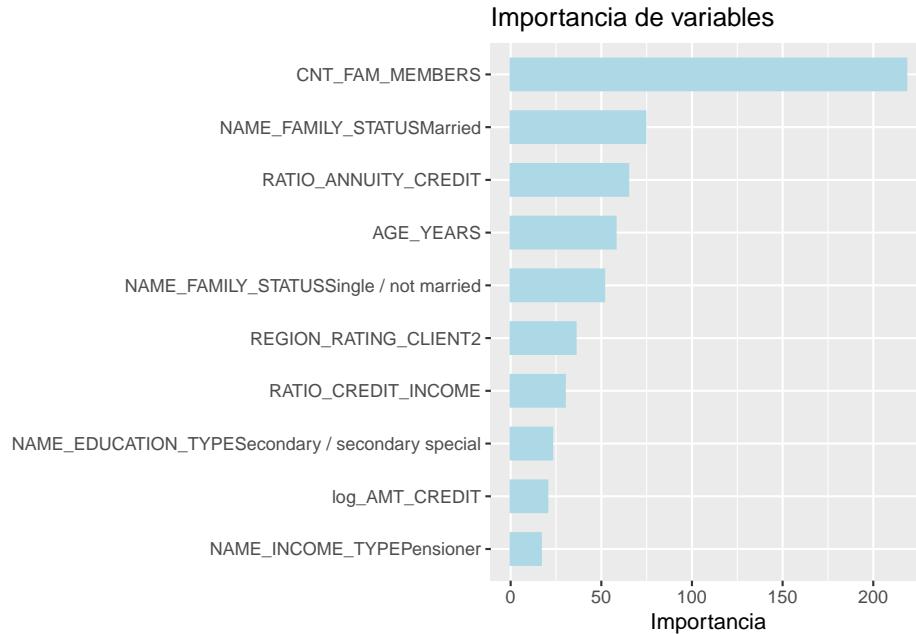
En contraste, para clientes con un ratio mayor o igual a 0.0295, factores adicionales como el estado civil, número de miembros familiares y educación influyen en la clasificación.

Clientes casados y con más de 2 miembros familiares tienden a tener una probabilidad de ser clasificados correctamente de no ser morosos (accuracy) de un 80 %. En casos específicos, como aquellos con menos de 2 miembros familiares y educación incompleta, la probabilidad de ser clasificados en clientes no morosos alcanza el 88 %.

La segmentación se profundiza aún más considerando variables como la región del cliente, la ocupación y la relación anualidad/crédito. En situaciones particulares, como ocupaciones desconocidas y ratios anualidad/crédito superiores a 0.03193, la probabilidad de morosidad se incrementa significativamente (71.81 %).

Este orden de variables en el árbol está determinado por la importancia relativa de cada variable en la tarea de clasificación. Las variables que ofrecen una mayor separación entre las clases son utilizadas en los niveles iniciales del árbol.

Figura 152: Importancia de las variables en CART



La variable “CNT_FAM_MEMBERS” (Número de miembros de la familia) es la característica más influyente en la clasificación del riesgo crediticio, con una importancia relativa del 218.19 %. Esto sugiere que la composición familiar tiene un impacto significativo en la capacidad de pago.

Por otro lado, el estado civil “Married” (Casado) y la relación entre la anualidad y el crédito (“RATIO_ANNUITY_CREDIT”) son también factores cruciales, con importancias del 74.26 % y 64.80 %, respectivamente. Estos indican que el estado civil y la relación entre la anualidad y el crédito desempeñan un papel fundamental en la toma de decisiones crediticias.

Además, la “Edad” (“AGE_YEARS”) del solicitante es otra variable clave, con una importancia del 57.85 %. Esto refuerza la conclusión de que la edad es un factor importante en la predicción de la no morosidad.

En resumen, las variables más influyentes, como el número de miembros de la familia, estado civil, relación anualidad/crédito y edad, resaltan la importancia de aspectos fundamentales en la evaluación del riesgo. Además, factores sociodemográficos como la ubicación geográfica y la educación juegan un papel crucial en la toma de decisiones crediticias.

Conclusiones

Como la sensibilidad obtenida ha sido mucho más baja que la especificidad, concluimos que el modelo tiene más dificultades para identificar los casos positivos reales (morosos) en comparación con su habilidad para identificar correctamente los casos negativos reales (no morosos). Este resultado no nos es beneficioso en la clasificación, ya que en este contexto quizás sea mejor detectar adecuadamente casos positivos (morosos), para así reducir el número de clientes morosos.

Así pues concluimos que el modelo CART proporciona una herramienta valiosa para la clasificación de riesgo crediticio, destacando la importancia de variables clave como los miembros de la familia, la edad y la relación entre la anualidad y el crédito. Para mejorar aún más, es recomendable explorar ajustes en la sensibilidad y considerar otras técnicas de modelado que puedan aportar mejoras específicas para el objetivo del problema.

Random Forest

El Random Forest, es un método que destaca por su capacidad para realizar predicciones precisas y robustas de conjuntos de datos. Utiliza un conjunto de árboles de decisión, cada uno entrenado en subconjuntos aleatorios de datos y utilizando subconjuntos aleatorios de características. Al combinar las predicciones de múltiples árboles, el Random Forest reduce el sobreajuste y mejora la generalización del modelo. Este enfoque de “ensamblado” hace que el algoritmo sea resistente al ruido y capaz de manejar conjuntos de datos grandes y complejos. Además, el Random Forest proporciona información sobre la importancia relativa de las características, lo que facilita la identificación de variables influyentes en el proceso de toma de decisiones.

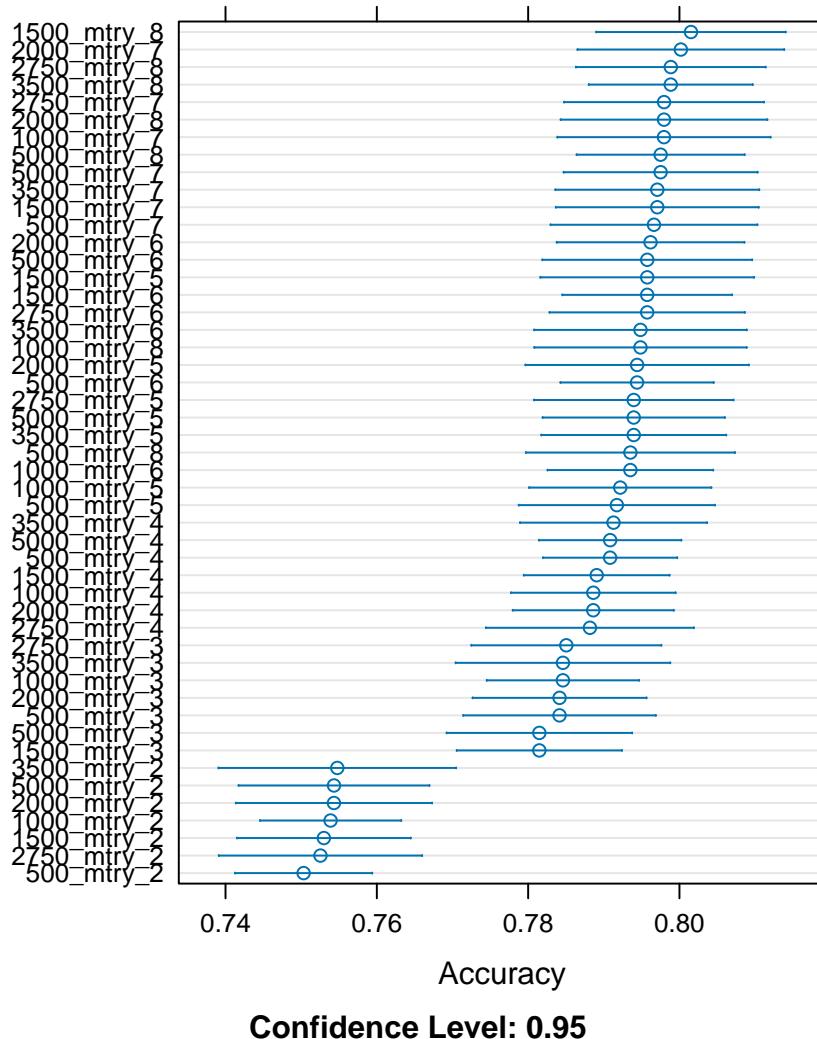
Selección de parámetros

Para llevar a cabo el Random Forest, se escogen sus dos parámetros principales:

1. **Número de árboles (ntree):** Este parámetro especifica la cantidad de árboles de decisión que se construirán en el bosque. Un mayor número de árboles generalmente mejora la estabilidad y la precisión del modelo, pero también aumenta el costo computacional. Sin embargo, hay un punto de rendimiento óptimo después del cual agregar más árboles puede no aportar mejoras significativas y puede conducir a un sobreajuste.
2. **Número de características seleccionadas en cada nodo (mtry):** Este parámetro indica cuántas características (variables) se deben considerar al hacer una división en cada nodo de un árbol. Una elección adecuada de mtry es crucial para el rendimiento del modelo. Valores bajos pueden llevar a la construcción de árboles más decorados y correlacionados, aumentando el riesgo de sobreajuste. Valores altos pueden hacer que los árboles sean más similares y reducir la diversidad del bosque.

Para encontrar los valores óptimos de ambos parámetros, se realizará un search grid para evaluar el accuracy del modelo para cada par de valores de los parámetros. Del parámetro ntrees se probarán los valores 1000, 1500, 2000, 2750, 3500 y 5000. No se probará un número mayor por el alto coste computacional que supone. Del mtry, se prueban los valores del 2 al 8, ya que el valor propuesto por la literatura sería 4. Para entrenar el modelo para cada par de variables se realiza un 10-fold cross validation.

Figura 153: Random Forest para pares de valores de parámetros Ntree y Mtry



Se prueban los pares de variables propuestos y en este gráfico, ordenado de mayor a menor accuracy, se observa que el mayor accuracy lo presenta modelo de Random Forest con ntree=1500 y mtry=8, con valores alrededor del 80 %.

Desarrollo y validación del modelo

A partir de la 10-folds cross validation del modelo escogido, en los datos de entrenamiento se obtiene un Recall de 'r sensitivity_train' y una Especificidad de 'r specificity_train'.

A continuación se presenta la matriz de confusión del conjunto de validación en el modelo de Random Forest entrenado.

Cuadro 77: Matriz de confusión del conjunto de validación

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	284	84
	Potencial moroso	36	156

Y a partir de las matrices de confusión de ambos conjuntos de datos train y test, se calculan las diferentes métricas de validación del modelo:

Cuadro 78: Medidas de Validación para el modelo Random Forest con conjunto test balanceado

	Train	Test
Accuracy	0.8067033	0.7857143
Sensitivity	0.8404118	0.6500000
Specificity	0.7891156	0.8875000
F1	0.8139564	0.7222222
Precision	0.6780893	0.8125000

El modelo Random Forest es más o menos exacto en las predicciones de la variable respuesta del conjunto de datos de validación, con un accuracy del ‘r MC\$overall[“Accuracy”]’ indica que el modelo acierta en casi el 79 % de las predicciones. Además, se observa que el modelo no presenta excesivo sobrajuste porque el Accuracy es aproximadamente igual para ambos conjuntos de datos train y test.

Analizando los valores de la Sensibilidad y Especificidad, con un valor de Sensibilidad del 65 % en el conjunto de prueba, el modelo tiene una capacidad moderada para identificar a los clientes morosos, lo que significa que el modelo está perdiendo algunas instancias de morosidad, mientras que con una Especificidad de casi el 90 % tiene una capacidad significativamente más alta para identificar los clientes no morosos, este valor minimiza los falsos positivos y ayuda a evitar que clientes no morosos sean clasificados incorrectamente como morosos.

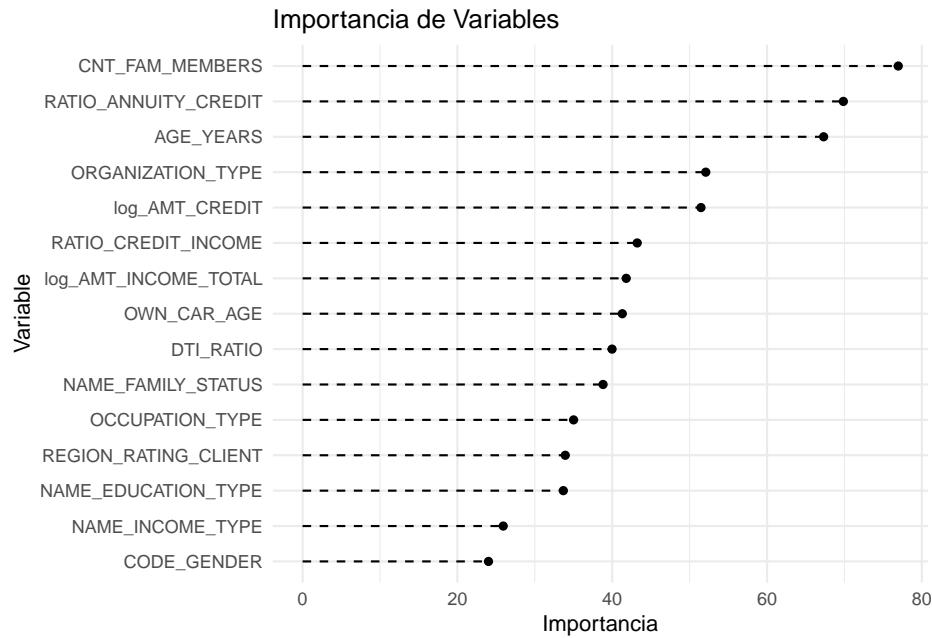
Sobre el F1, con un valor del 72.56 % en el conjunto de prueba, el F1 Score indica un equilibrio razonable entre precisión y sensibilidad. En problemas de morosidad, donde las consecuencias pueden ser significativas, es importante buscar un equilibrio entre identificar correctamente a los morosos y evitar clasificar incorrectamente a los no morosos.

Por último, analizando la Precisión, con un valor del 82.11 % en el conjunto de prueba indica que, cuando el modelo predice que un cliente es moroso, tiene una alta probabilidad de que sea correcto. Sin embargo, se debe considerar en conjunto con la sensibilidad para no pasar por alto clientes morosos.

Variables importantes

La importancia de las variables en Random Forest se basa en cuánto contribuyen al aumento de la homogeneidad de las clases cuando se utilizan para hacer divisiones en los árboles de decisión del bosque.

Figura 154: Importancia de las variables en Random Forest



Se consideraran las variables más importantes las primeras cuatro que se presentan en el gráfico. Se observa que la Ratio entre la anuidad del préstamo y el crédito total solicitado, el DTI (Debt-to-income) ratio (que mide la capacidad del cliente para pagar la annuity de su préstamo en relación con sus ingresos), la Ratio entre el crédito pedido y el salario anual del prestatario (también se puede contar como el número de años que se tarda en devolver el crédito) y el tipo de organización en la que trabajan, son las cuatro variables más importantes, y que por tanto influyen más en la predicción de morosidad de un cliente en el modelo de Random Forest.

Prueba ácida

Cuadro 79: Matriz de confusión del conjunto de validación desbalanceado

		Realidad	
		No moroso	Potencial moroso
Predicción	No moroso	819	13
	Potencial moroso	100	67

Cuadro 80: Medidas de Validación con el conjunto test desbalanceado para el modelo Random Forest

	Train	Test
Accuracy	0.8067033	0.8868869
Sensitivity	0.8404118	0.8375000
Specificity	0.7891156	0.8911861
F1	0.8139564	0.5425101
Precision	0.6780893	0.4011976

En el conjunto de entrenamiento, el modelo acierta alrededor del 80.67 % de las predicciones, mientras que en el conjunto de prueba, alcanza un 88.69 %. El accuracy es una métrica general que puede ser engañosa en conjuntos de datos desbalanceados, como en el caso de este conjunto de datos de validación, ya que puede estar dominada por la clase mayoritaria. La sensibilidad mide la proporción de casos positivos que el modelo identifica correctamente. En el conjunto de entrenamiento, el modelo identifica correctamente alrededor del 84.04 % de los casos positivos, mientras que en el conjunto de prueba, la sensibilidad es del 83.75 %, lo que indica que el modelo está capturando bien los casos positivos. El modelo también presenta una alta Especificidad para el conjunto de validación desbalanceado, con una tasa del 89,11 %, indicando una buena capacidad para predecir casos negativos.

En el conjunto de prueba, el F1 Score es del 54.25 %. Aunque esta puntuación es más baja que en el conjunto de entrenamiento, sigue siendo una métrica relevante para evaluar el rendimiento del modelo en un conjunto de datos desbalanceado. Un F1 Score más bajo en el conjunto de prueba sugiere que el modelo tiene dificultades para equilibrar precisión y recall en un entorno desbalanceado, y puede requerir ajustes para mejorar su rendimiento en la clasificación de la clase minoritaria. La precisión en el conjunto de prueba es del 40.12 %, lo que significa que el modelo tiene una proporción más baja de verdaderos positivos entre todas las instancias clasificadas como positivas, situación esperada en conjuntos desbalanceados.

En conclusión, el modelo Random Forest exhibe una sensibilidad excepcionalmente alta, indicando su habilidad sobresaliente para identificar correctamente los casos positivos. Esta capacidad para minimizar los falsos negativos, es decir, la tendencia del modelo a perder muy pocos casos positivos reales, es extremadamente valiosa, especialmente en el contexto financiero, donde la identificación precisa de la clase positiva es esencial. La baja incidencia de falsos negativos indica que nuestro modelo tiende a errar en el lado de la precaución, asegurándose de no perder casos positivos importantes. Sin embargo, es importante tener en cuenta que la alta sensibilidad va acompañada de un aumento en los falsos positivos, es decir, este modelo tiende a clasificar más instancias como positivas de lo necesario.

XGBoost

XGBoost es un modelo basado en árboles de decisión y es una mejora de otros métodos de ensamblaje como el Random Forest. El algoritmo utiliza varios métodos de optimización para mejorar la precisión y controlar el sobre ajuste.

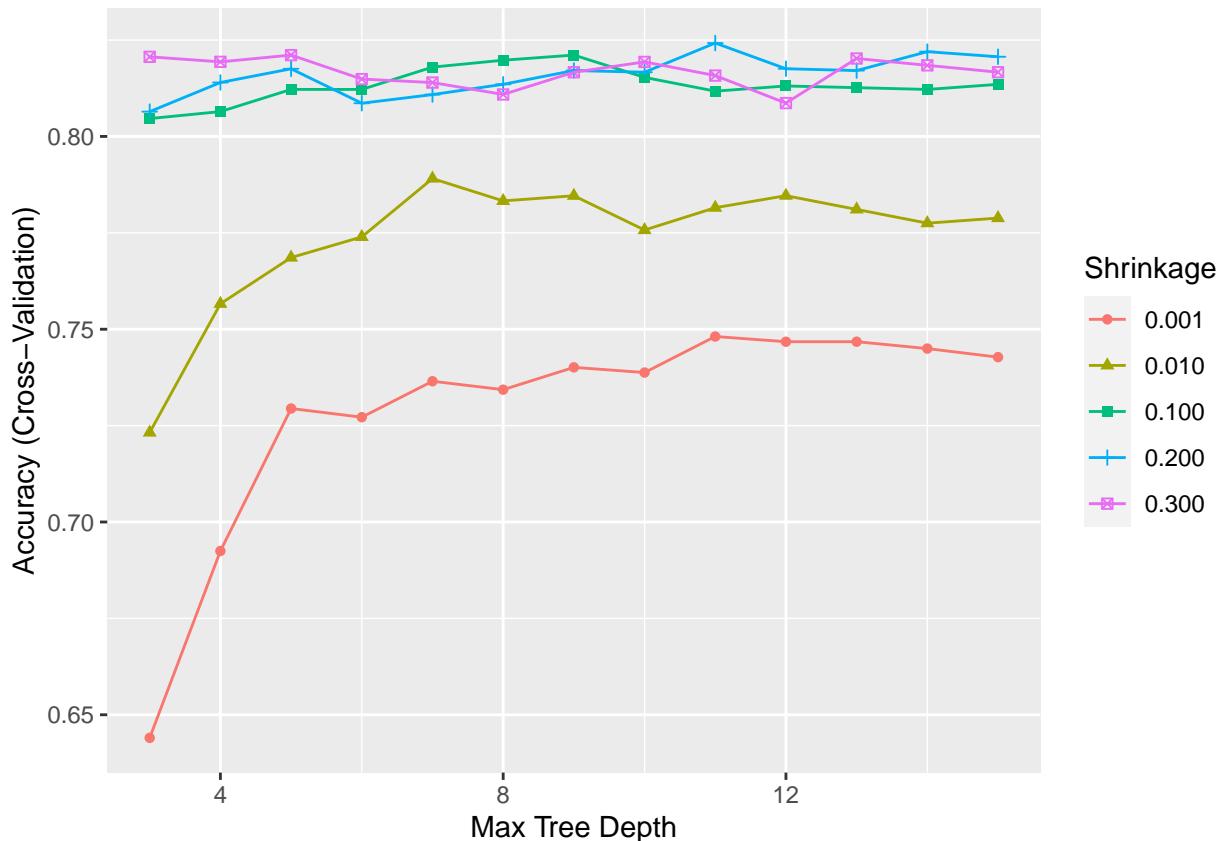
El proceso de XGBoost comienza con una predicción inicial y luego calcula los residuos, que son las diferencias entre las predicciones y los valores observados. Luego, crea un árbol de decisión con estos residuales y continúa este proceso, construyendo árboles secuenciales que aprenden de los errores del árbol anterior.

Este modelo se puede ajustar según ciertos parámetros. Para encontrar la mejor combinación de parámetros. Se llevará a cabo validación cruzada en el conjunto de entrenamiento con el fin de extraer los mejores hiperparámetros. Posteriormente, se aplicará el mejor modelo obtenido en nuestro conjunto de prueba para su validación.

En la validación cruzada usamos Grid Search. Se busca la mejor combinación de parámetros (como tasa de aprendizaje, número de rondas y profundidad del árbol) probando múltiples valores.

Esto implica entrenar y evaluar el modelo con cada combinación para encontrar la configuración que maximice el rendimiento.

Posteriormente, se grafica cómo varía el desempeño del modelo en función de estos valores, lo que ayudará a identificar la configuración óptima para obtener mejores resultados en otros conjuntos de datos.



Los valores óptimos de nuestros parámetros según la función podrían ser:

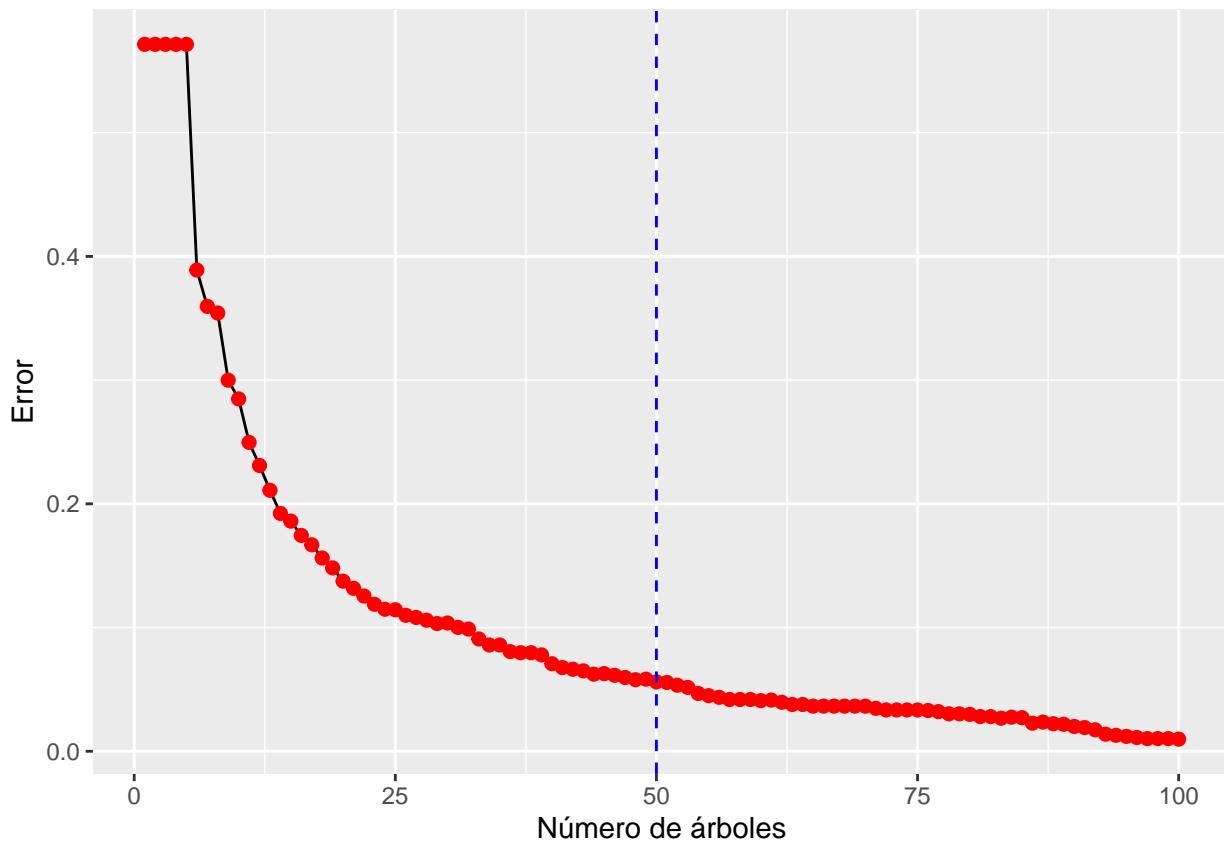
- Número de rondas de boosting 100

- Profundidad de árbol 11
- Tasa de aprendizaje 0.2

Aunque estamos obteniendo estos resultados, queremos estudiar más profundamente en qué valores fijar estos parámetros para mejorar la eficiencia y sencillez de nuestro árbol. Queremos observar cuándo convergen nuestros valores.

Así se propone:

La profundidad del árbol fijada en 7, ya que, según se puede observar en el gráfico anterior, la precisión parece estabilizarse a partir de esa profundidad. Por tanto, los niveles posteriores del árbol no aportarían información adicional que contribuyera a mejorar la precisión del modelo.



Por otro lado, en este gráfico se puede observar la evolución del error en función del número de iteraciones. Nos muestra como este valor tiende a estabilizarse a partir de la iteración 50. Así se fija el número de rondas en este valor.

Se escoge como valor de eta 0.1. Aun siendo un valor elevado que podría generar overfitting, queremos probar si obtenemos un buen modelo.

Para verificar que nuestro modelo no tiene overfitting, antes de aplicar la prueba ácida, usaremos un conjunto de Test balanceado con nuestro modelo.

Cuadro 81: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.7910714	0.91875	0.6208333

Comparando Accuracy, en train obtenemos un 0.8179768 y en test 0.7910714, lo que verifica que el modelo con estos parámetros no genera overfitting. Ahora aplicaremos la prueba ácida, con datos test desbalanceados.

Cuadro 82: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.8948949	0.9194777	0.6125

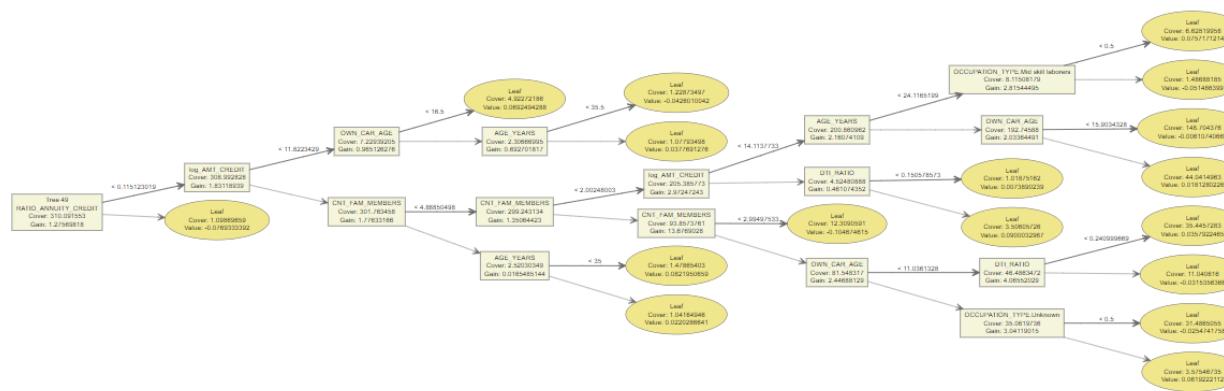
Resultados

Los resultados del modelo XGBoost muestran un rendimiento sólido en la clasificación, con una precisión en la predicción del 0.8948949 %.

Destaca la alta especificidad del modelo, alcanzando un 0.9194777, lo que indica su capacidad para identificar correctamente los verdaderos negativos. Además, la sensibilidad es del 0.6125, lo que nos muestra su eficacia en la identificación de casos positivos, aunque es mejorable.

Esta evaluación sugiere que el modelo no presenta overfitting, ya que se presentan valores similares en la precisión y otros.

Así, el árbol conseguido en el modelo es el siguiente:



La característica CNT_FAM_MEMBERS muestra un valor de Ganancia (Gain) de 1.77633166 al dividir los datos. Esto indica que contribuye significativamente a la reducción de la impureza o mejora la separación de los datos en comparación con otras características en este nivel específico del árbol.

Este patrón continúa a través de los niveles subsiguientes hasta alcanzar el séptimo nivel. Es importante destacar que la siguiente variable en orden de importancia es la edad.

Conclusiones:

Los resultados del modelo XGBoost revelan un rendimiento consistente en la tarea de clasificación, particularmente notorio por su alta especificidad. Consideramos que este modelo es eficaz para clasificar nuestros datos según nuestros objetivos, aunque nos interesa más una sensibilidad elevada que una alta especificidad. Una mayor sensibilidad implica tener más falsos negativos que falsos positivos, lo cual es crucial al considerar si un cliente será moroso o no.

Ensemble Híbrido

En esta sección del trabajo se implementa un método de ensamblaje híbrido.

El ensamblaje híbrido implica la combinación de diversas técnicas de ensamblaje en un único marco de trabajo con el fin de mejorar el rendimiento predictivo del modelo. Este enfoque se fundamenta en la premisa de que la combinación de múltiples modelos puede ofrecer predicciones más precisas y robustas que cualquier modelo individual.

El método de ensamblaje utilizado emplea un modelo lineal para combinar las predicciones de varios modelos, asignándoles pesos según su rendimiento, con el objetivo de obtener una predicción final ponderada.

Selección de los modelos para el ensamblaje

En la estrategia de ensamblaje híbrido que hemos implementado, se han seleccionado los tres mejores modelos de entre todos los evaluados a lo largo de nuestro estudio.

Nos hemos enfocado especialmente en tres aspectos clave para nuestra investigación: la precisión general del modelo (accuracy, en distintos conjuntos de prueba) y la sensibilidad. Estos tres modelos seleccionados como los mejores, en orden de desempeño, son Random Forest, SVM (Support Vector Machine) y XGBoost.

El proceso de selección se ha basado en la capacidad de estos modelos para generalizar patrones relevantes sin caer en el sobreajuste a los datos de entrenamiento. Además, nos hemos centrado en la sensibilidad, ya que es crucial para nuestro trabajo identificar correctamente los casos positivos, en este caso, los morosos.

Una vez finalizado el apartado dedicado al ensamblaje híbrido, se realizará una exposición más detallada sobre la selección de los mejores modelos. Esta sección se centrará en brindar una explicación exhaustiva de los modelos seleccionados y los motivos que respaldan su elección como los mejores dentro del estudio.

Resultados del ensamblaje En la implementación de los modelos en nuestro ensamblaje obtenemos los siguientes pesos para cada componente con su precisión (Accuracy) en el conjunto de entrenamiento:

Intercept: 2.3433 Random Forest: Peso de -1.8752. Support Vector Machine con kernel radial: Peso de 0.3761. Extreme Gradient Boosting Tree: Peso de -3.1035.

La precisión (Accuracy) obtenida en el conjunto de entrenamiento para este modelo ensamblado fue del 72.55 %.

Estos pesos nos indican de la contribución relativa de cada modelo al rendimiento general del ensamblaje, reflejando su influencia en la predicción del resultado final.

Interpretación de los coeficientes

Intercept en 2.34: Indica el valor base del modelo. En el contexto de un ensamblaje de modelos, este podría ser un término de ajuste o el valor inicial alrededor del cual se hacen las predicciones.

Random Forest con -1.8752: El peso negativo sugiere que el modelo de Random Forest está contribuyendo a disminuir la predicción final del ensamblaje. Puede ayudar en la corrección de la sobreestimación, y esto puede deberse a la interacción con el valor elevado del intercept.

Support Vector Machine con kernel radial en 0.3761: Este peso positivo indica que el modelo SVM está contribuyendo a aumentar la predicción final del ensamblaje, aunque su contribución es cercana a 0, lo que sugiere que este modelo tiene el menor impacto en el ensamblaje.

Extreme Gradient Boosting Tree con -3.1035: Este peso negativo, similar al de Random Forest, sugiere que el modelo XGBoost está contribuyendo a disminuir la predicción final del ensamblaje. Dado que este peso es más grande en magnitud que el de Random Forest, el modelo XGBoost tiene un mayor impacto en la corrección de la sobreestimación.

Validación

Ahora evaluaremos el modelo con un conjunto de prueba balanceado para evaluar si hay overfitting:

Cuadro 84: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.7839286	0.940625	0.575

Observamos que los valores de precisión (accuracy) en los conjuntos de datos de entrenamiento y prueba son bastante cercanos, lo que sugiere que no hay evidencia de sobreajuste (overfitting) por parte del modelo.

Asimismo, se observa una disparidad entre la sensibilidad y la especificidad del modelo, donde la última es considerablemente más alta que la sensibilidad. Esto puede indicar un desequilibrio en el rendimiento del modelo, con una mayor capacidad para predecir correctamente los casos negativos en comparación con los casos positivos.

Ahora hacemos la prueba ácida, con datos de prueba desbalanceados:

Cuadro 85: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.9219219	0.9564744	0.525

Vemos que el modelo mejora considerablemente su precisión en el conjunto de datos desbalanceado, lo cual podría deberse al desequilibrio entre sensibilidad y especificidad. En nuestro caso, la sensibilidad mide los verdaderos positivos, es decir, cuando un individuo moroso es clasificado correctamente como tal. Dado que nuestros datos están desbalanceados con muchos más no morosos que morosos, una especificidad tan alta hace que la mayoría de los datos, que son no morosos, se clasifiquen correctamente.

Estos resultados no son óptimos para nuestro estudio, ya que nuestro interés principal radica en clasificar de manera más eficiente a los individuos morosos que a los no morosos. La alta especificidad, si bien puede ser útil para identificar correctamente los casos negativos, no responde de manera satisfactoria a nuestro objetivo prioritario de identificar con precisión a los morosos.

El ensamblaje híbrido, en base a nuestros objetivos, no parece mejorar el rendimiento de nuestros modelos.