

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

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

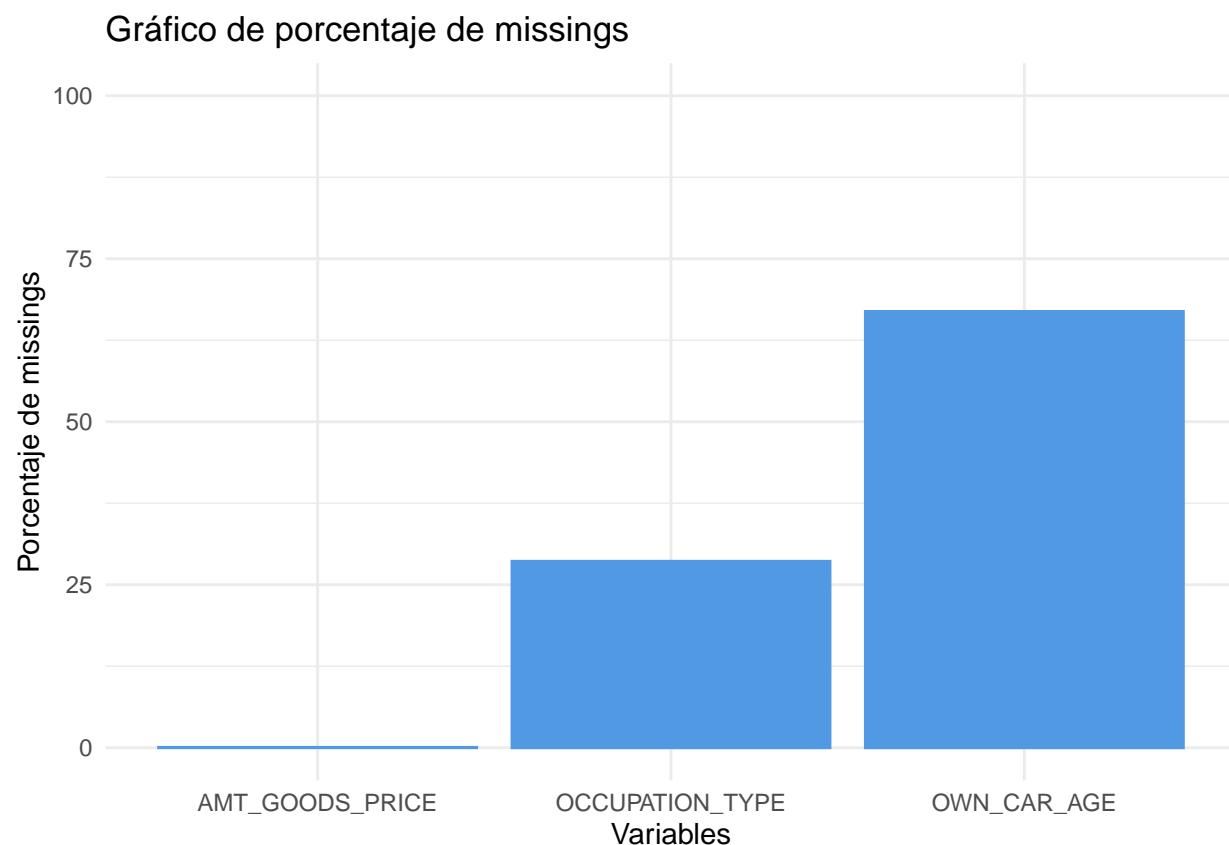
| Nombre | Descripción | Tipo | Diccionario y dominio |
|--------|-------------|------------|---|
| Target | Target | Categórica | 1 - Cliente con dificultades de pago: él/ella tuvo pagos atrasados de más de X días en al menos una de las primeras Y cuotas del préstamo en nuestra muestra. 0 - Todos los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

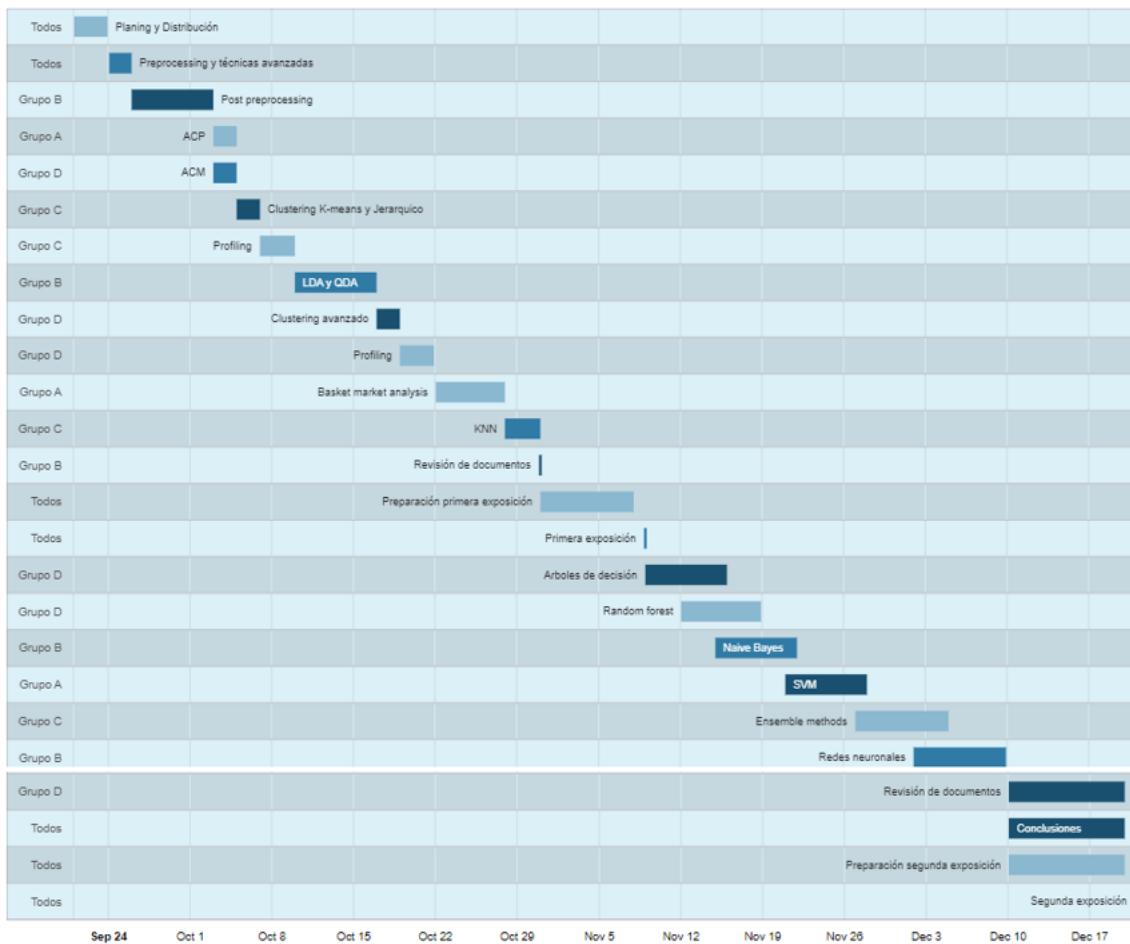
| | Nro. de missings | Porcentaje de missings |
|-----------------|------------------|------------------------|
| OWN_CAR_AGE | 3346 | 66.92 % |
| AMT_GOODS_PRICE | 3 | 0.06 % |
| OCCUPATION_TYPE | 1430 | 28.6 % |



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

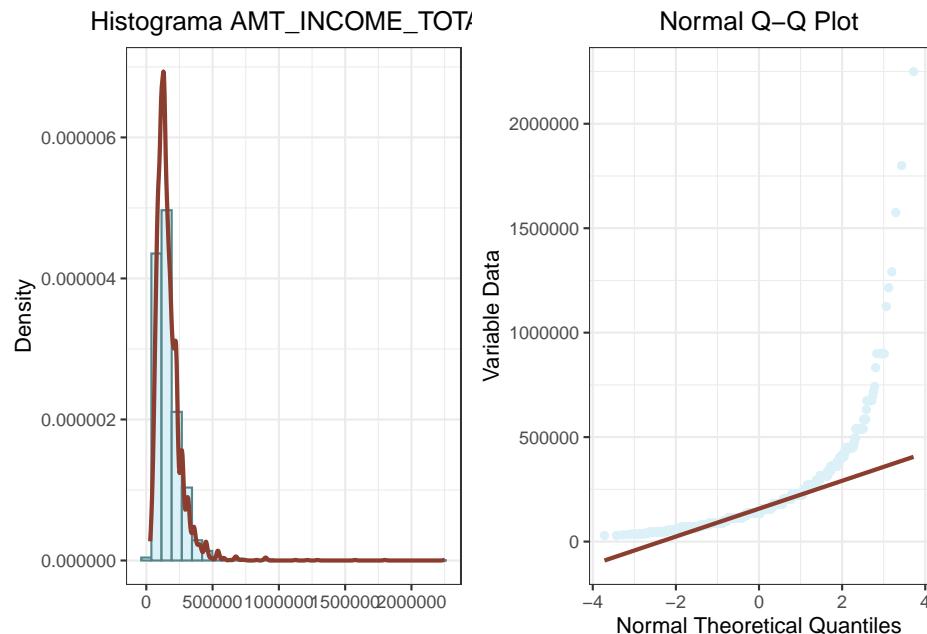
| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|------------------|------|------|-----------|-----------|--------|-----------|-----------|--------|-----------|-----------|-------|----------|---------|
| AMT_INCOME_TOTAL | 1 | 5000 | 166848.84 | 102440.66 | 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.

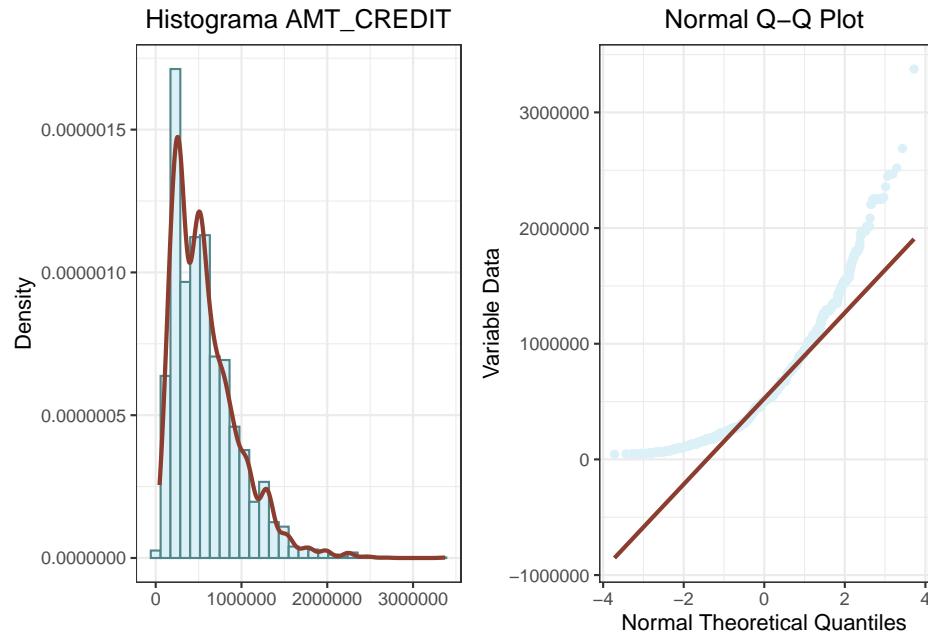
Así pues, tras hacer un análisis general, se procede a realizar un análisis más particular. Para ello, analizaremos cada variable una a una, de forma gráfica:

Figura 1: Análisis Gráfico Variable AMT INCOME TOTAL



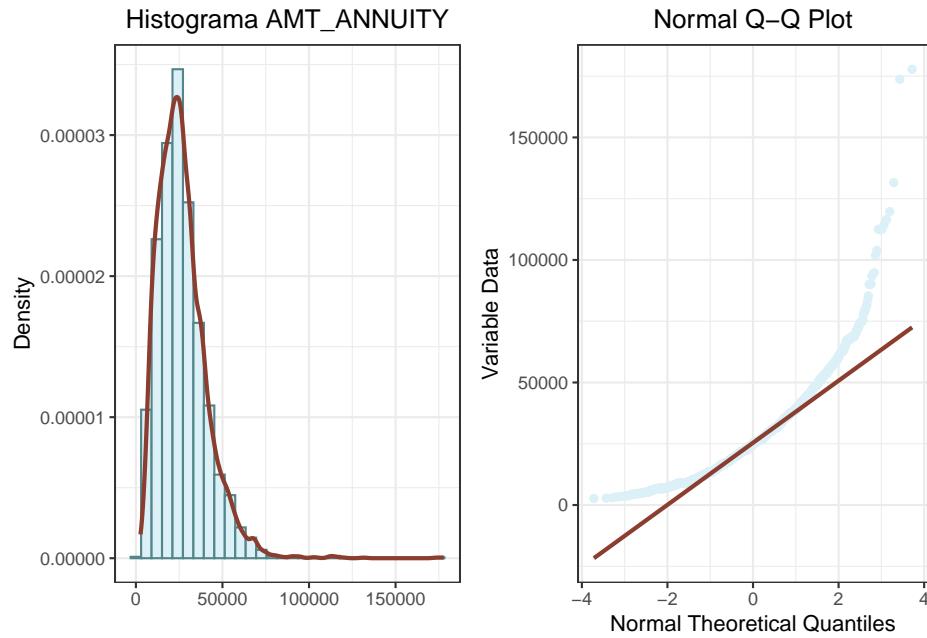
Estos gráficos muestran que los datos de la variable `AMT_INCOME_TOTAL` no siguen una distribución normal y parecen seguir una distribución exponencial. Esto tiene sentido, ya que la distribución de los ingresos totales de los individuos en una población generalmente no sigue una distribución normal. 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



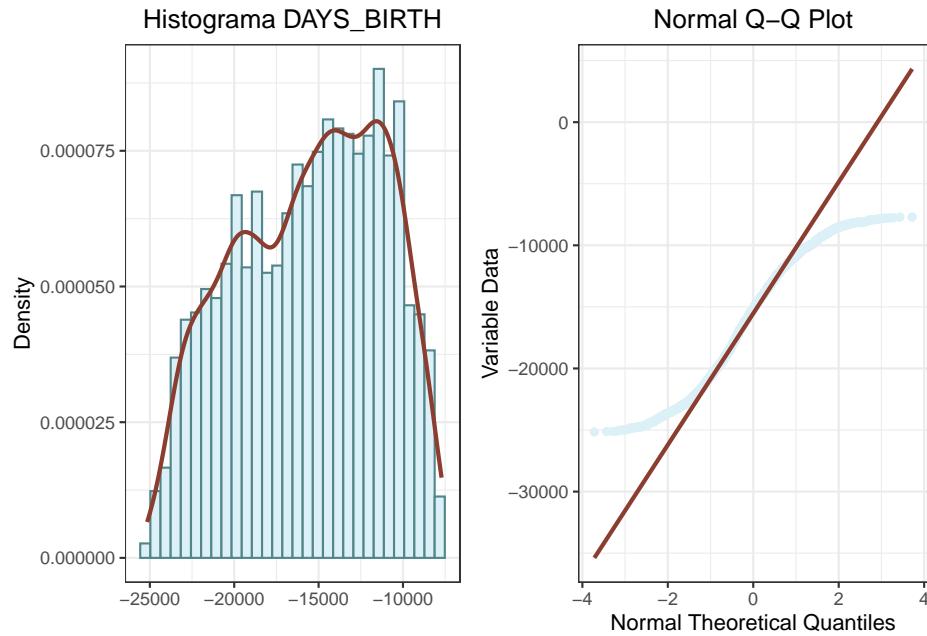
Al igual que en el caso anterior, la variable tampoco sigue una distribución normal, lo cual se confirma además por el test de Shapiro-Wilk con un p-valor de $3,1426414 \times 10^{-49}$. Parece que sigue una distribución exponencial.

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.

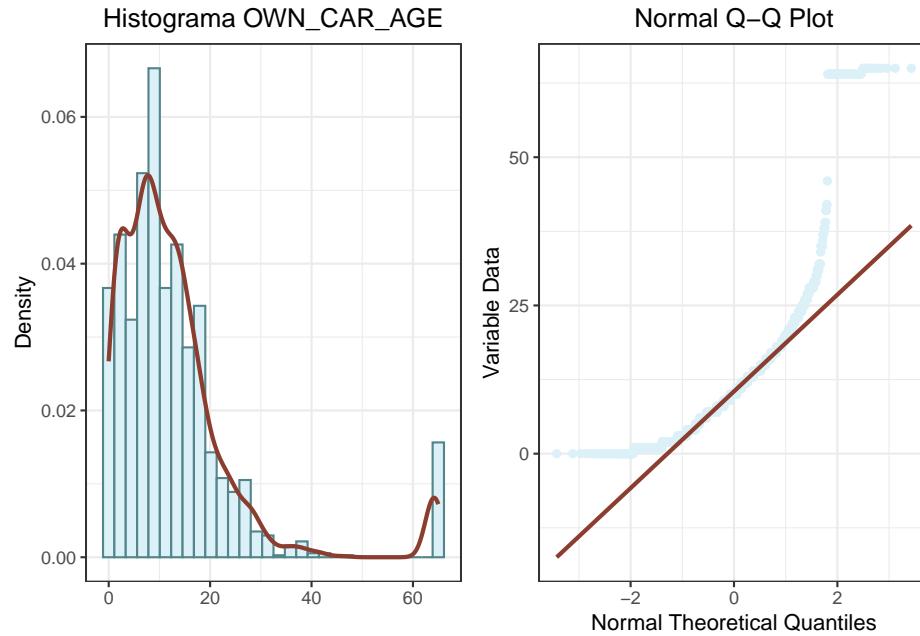
Figura 4: Análisis Gráfico Variable DAYS BIRTH



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

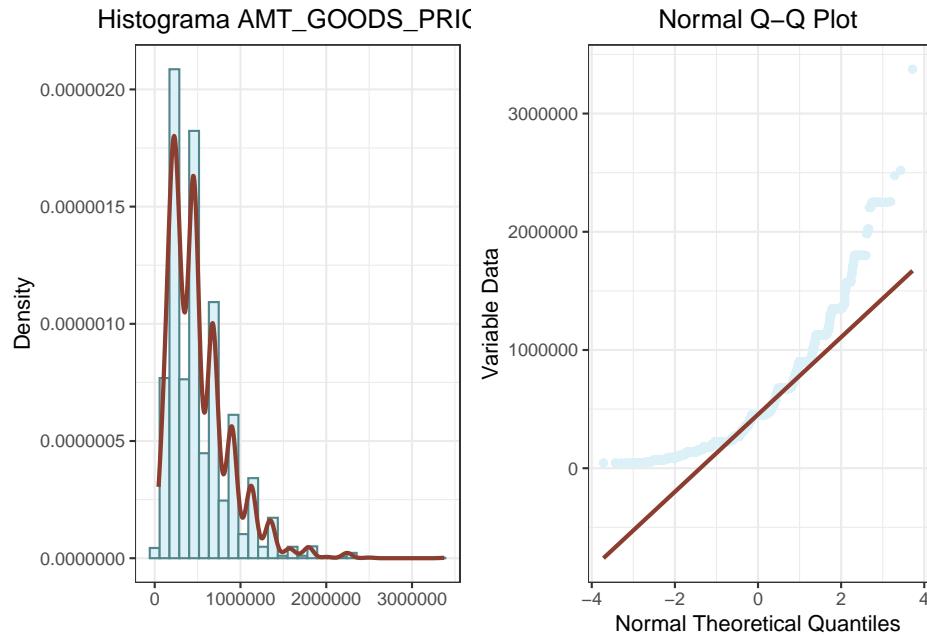
en que solicitó el crédito. Por lo tanto, es necesario transformar los datos para que sean positivos y modificar la variable de manera que represente las edades de los sujetos en años, lo que facilitará un mejor tratamiento y comprensión de los resultados.

Figura 5: Análisis Gráfico Variable OWN CAR AGE



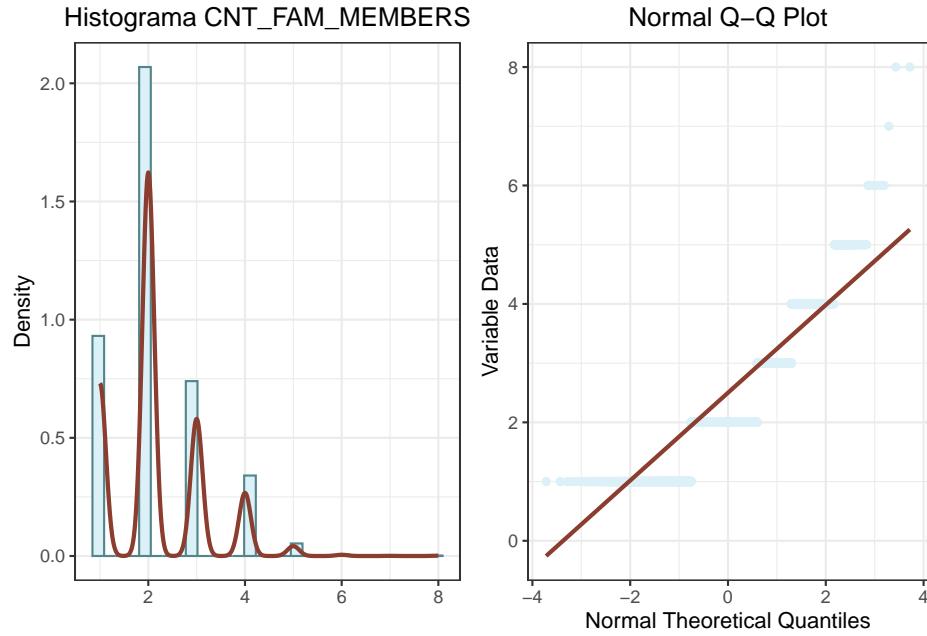
En la variable ‘Own car age’, también se observa que no sigue una distribución normal, como lo demuestra el test de Shapiro-Wilk con un p-valor de $3,2161311 \times 10^{-45}$. Se puede notar una alta concentración de datos alrededor de los 10 años, lo que muestra una estructura similar a una distribución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

```

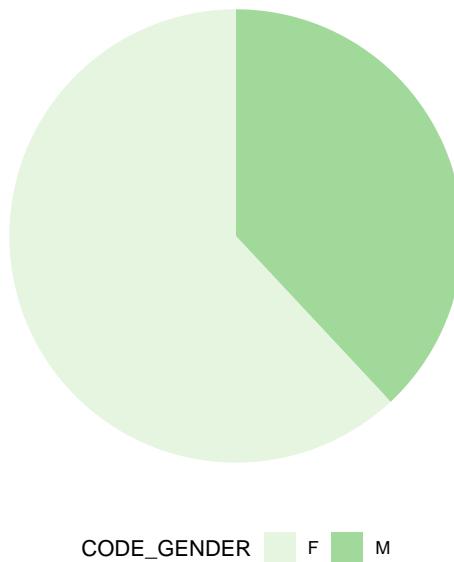
var <- names(data_f)[i]
freq_table <- table(data[[var]])
num_classes <- length(freq_table)

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))
}
}
}

```

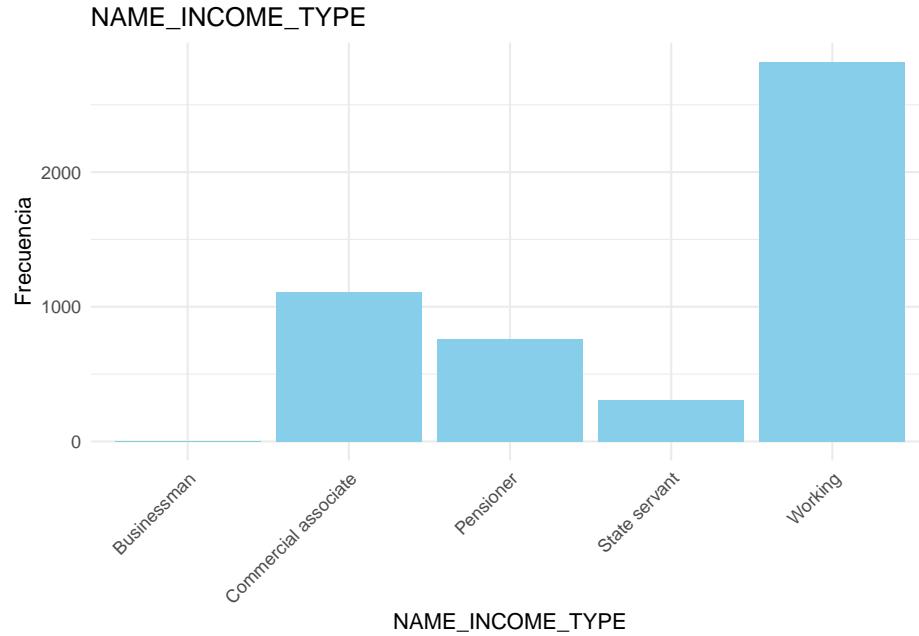
Figura 8: Pie Chart Variable CODE GENDER

CODE_GENDER



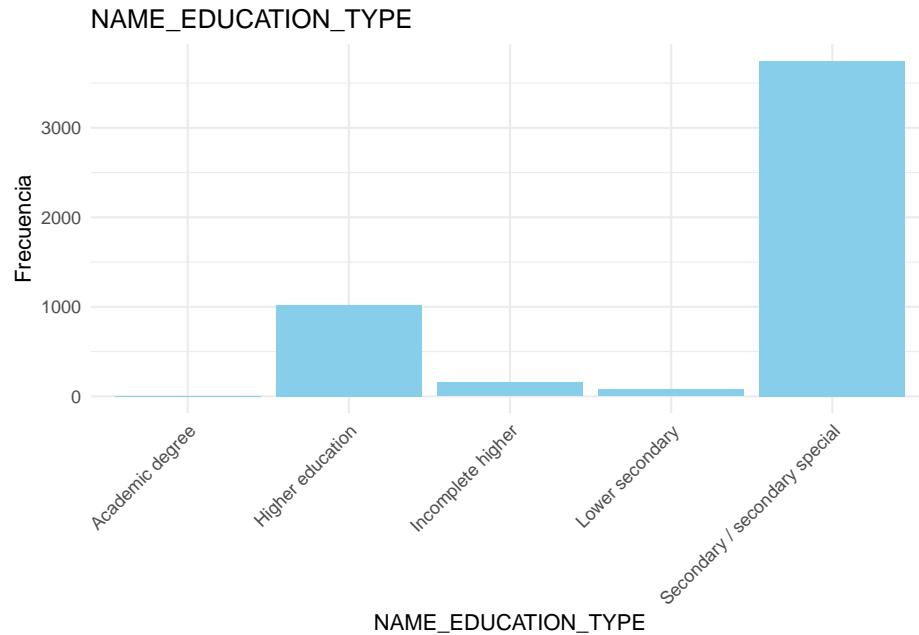
Para empezar, analizamos la variable referida al género. Como se puede apreciar, la gran mayoría de individuos de la base de datos son mujer, con un porcentaje del 61.96 %. Todo el resto de individuos son hombres.

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.

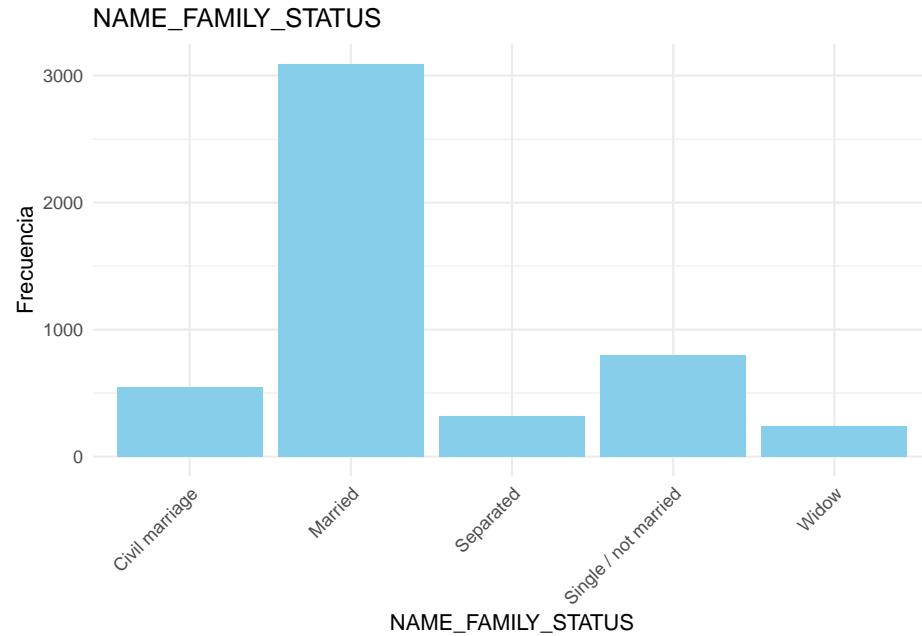
Figura 10: Pie Chart Variable NAME EDUCATION TYPE



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

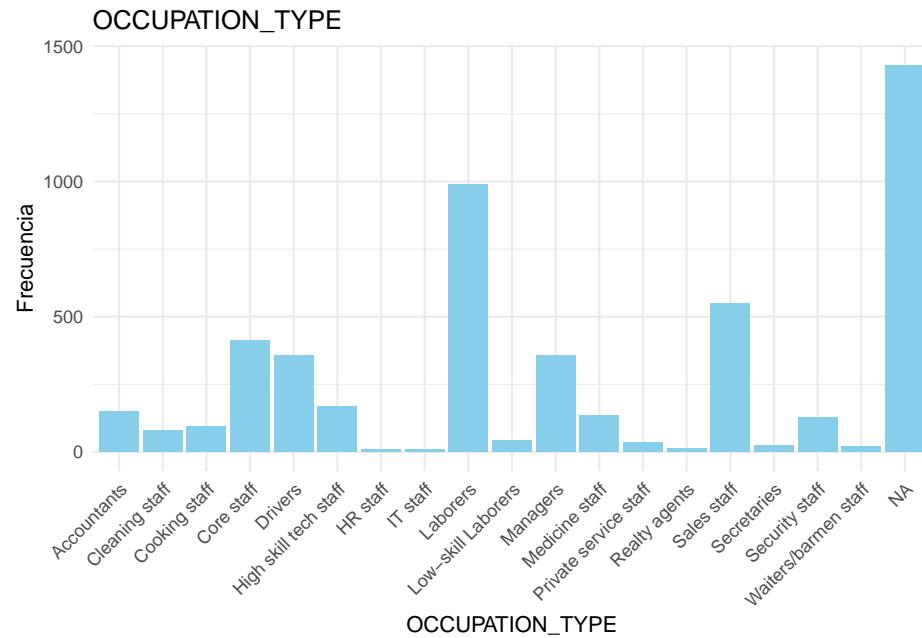
seguido de los universitarios (20.36 %). En general, se podría decir que hay pocos clientes con un nivel educativo bajo.

Figura 11: Pie Chart Variable NAME FAMILY STATUS



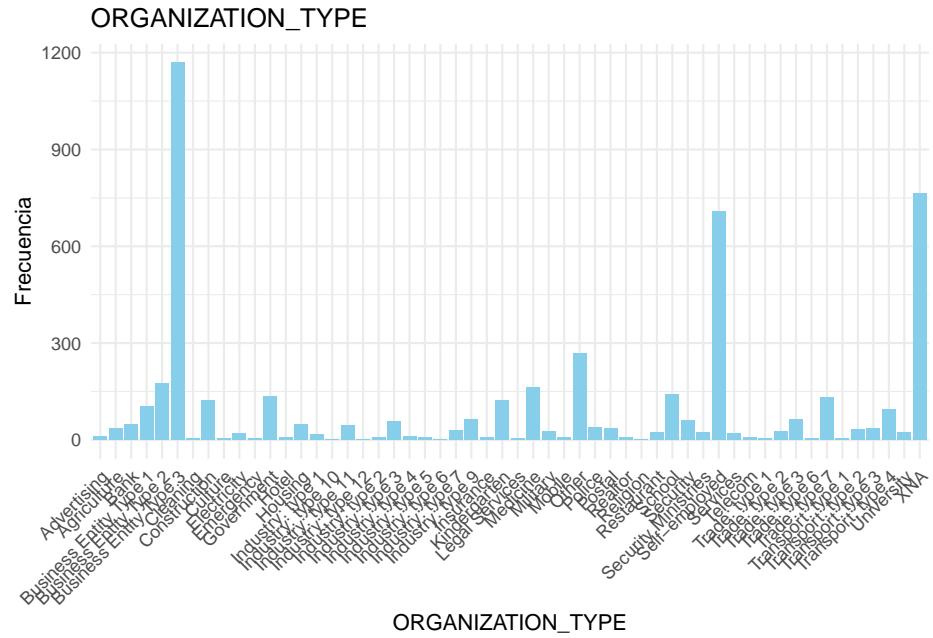
Si pasamos a hablar sobre el estado civil de los clientes, se puede apreciar que la gran mayoría están casados (61.9 %), seguido de los solteros o no casados (15.96 %). El resto de subgrupos es más minoritario.

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.

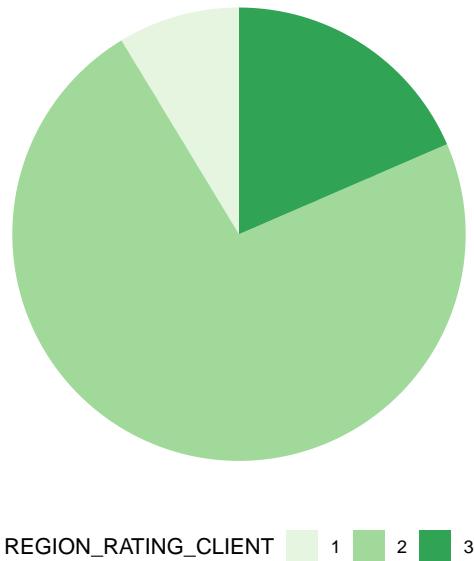
Figura 13: Pie Chart Variable ORGANIZATION TYPE



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

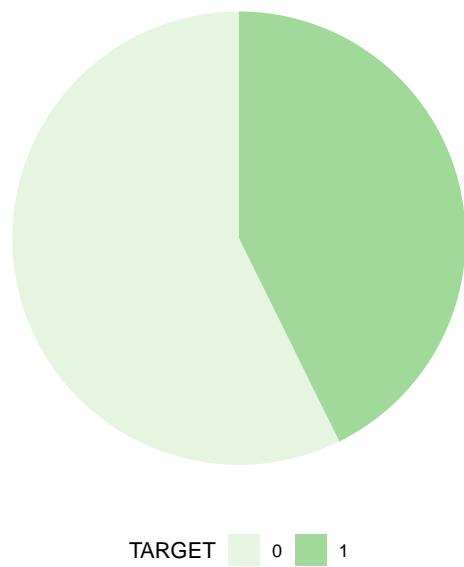
REGION_RATING_CLIENT



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

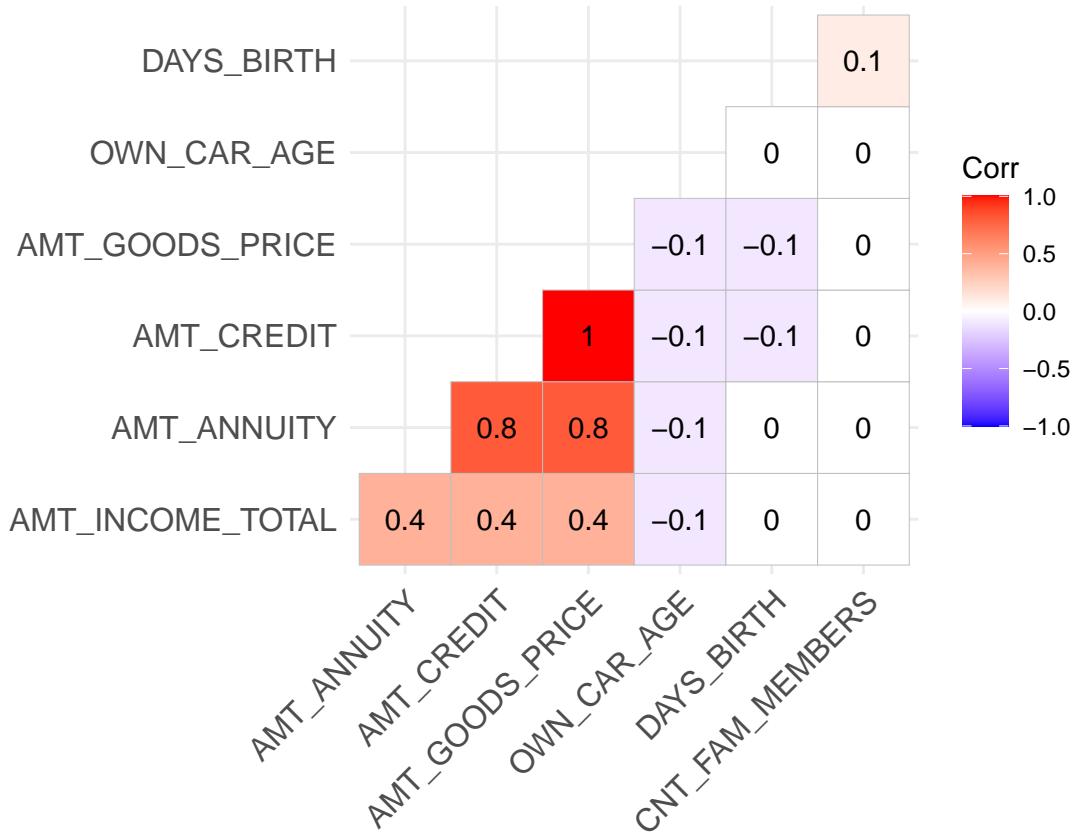


Por último, se analiza la variable respuesta de nuestra base de datos: TARGET. Como se puede apreciar, un 57.3% de los clientes no tienen problemas de solvencia, mientras que un 43.7% los podría presentar.

Análisis Bivariante Numérico

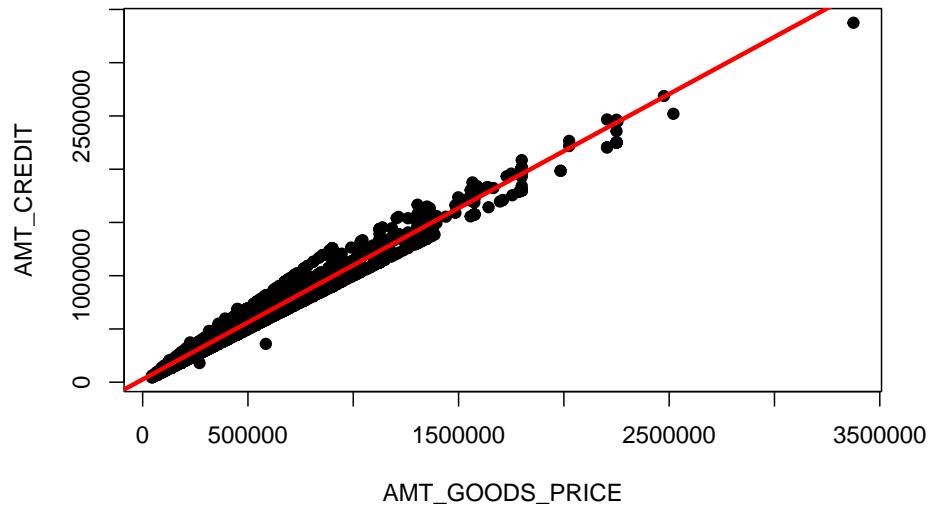
Con el propósito de identificar las relaciones más significativas entre las variables numéricas, se ha creado un gráfico de correlación utilizando la técnica de HeatMap. En este gráfico, los colores indican el grado de dependencia entre las variables numéricas. Cuanto más intenso sea el color, mayor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

Figura 16: Matriz de Correlaciones para las Variables Numéricas



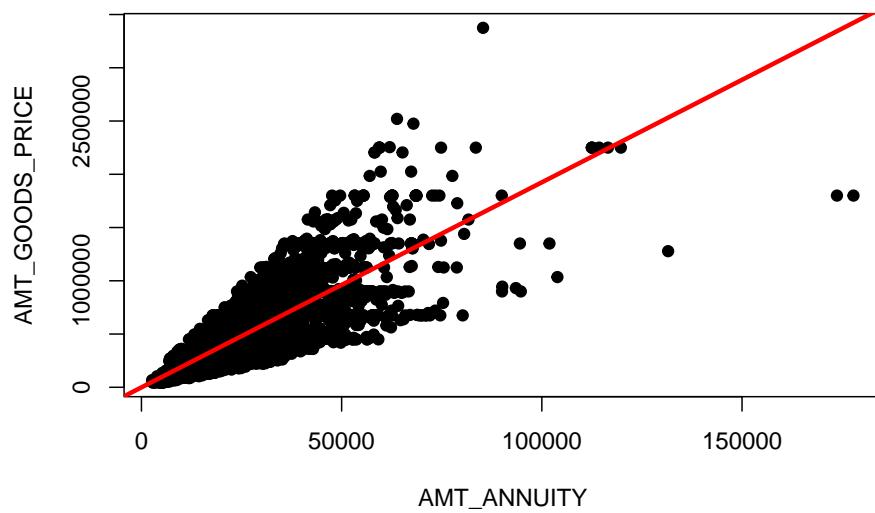
Tras analizar el gráfico, se destacan notables correlaciones entre las variables AMT_CREDIT y AMT_GOODS_PRICE. Esta correlación tiene sentido, ya que los prestamistas suelen otorgar créditos en función del valor del activo que el prestatario desea adquirir. En caso de impago, el prestamista retiene dicho 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

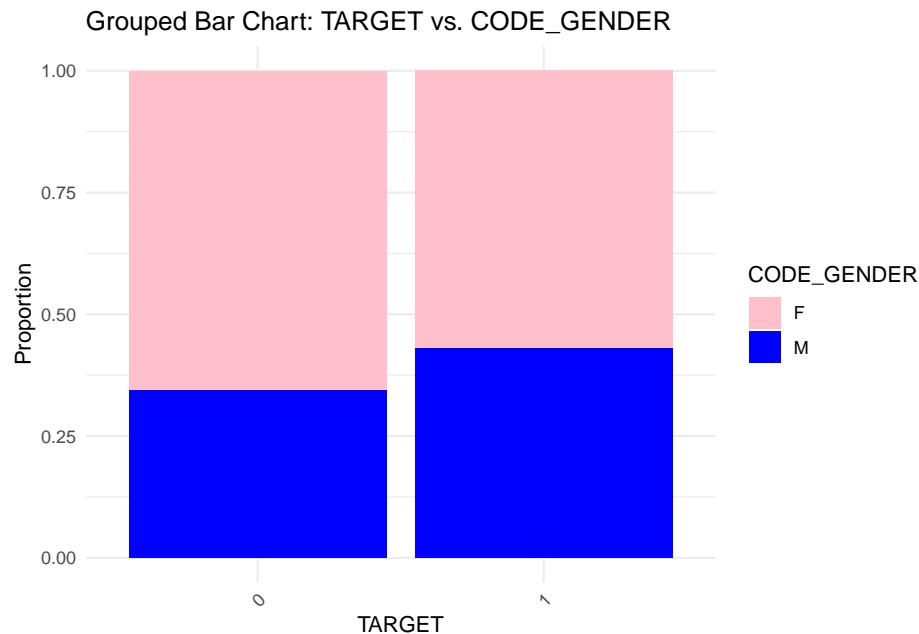


De manera similar, la correlación entre el valor de los bienes y la anualidad también es bastante alta. Es importante señalar que los clientes que posean una relación entre la anualidad y el valor del bien que compren (teniendo en cuenta que el precio del bien es igual al valor del préstamo) serán aquellos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

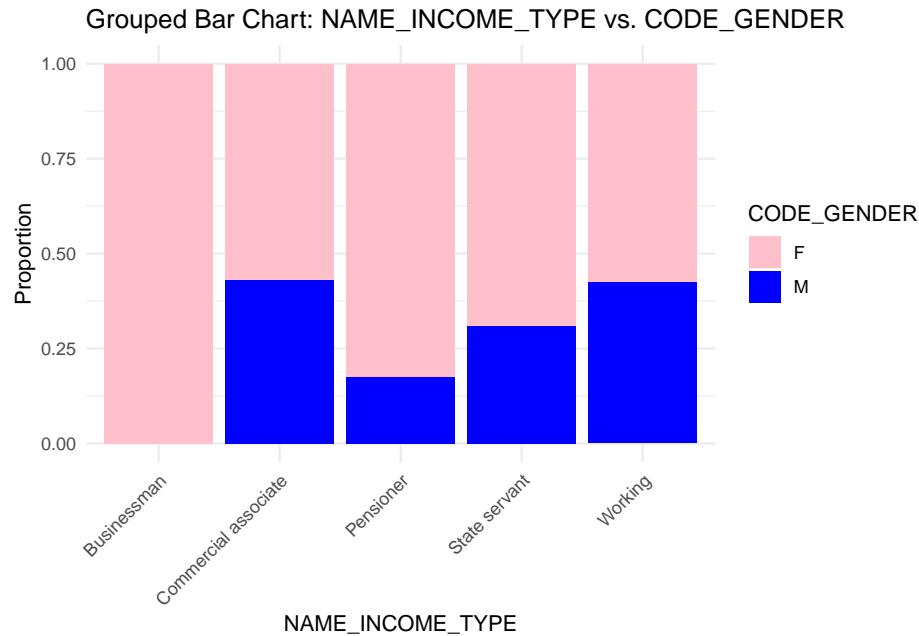
Figura 19: Stacked bar chart TARGET vs CODE_GENDER



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.

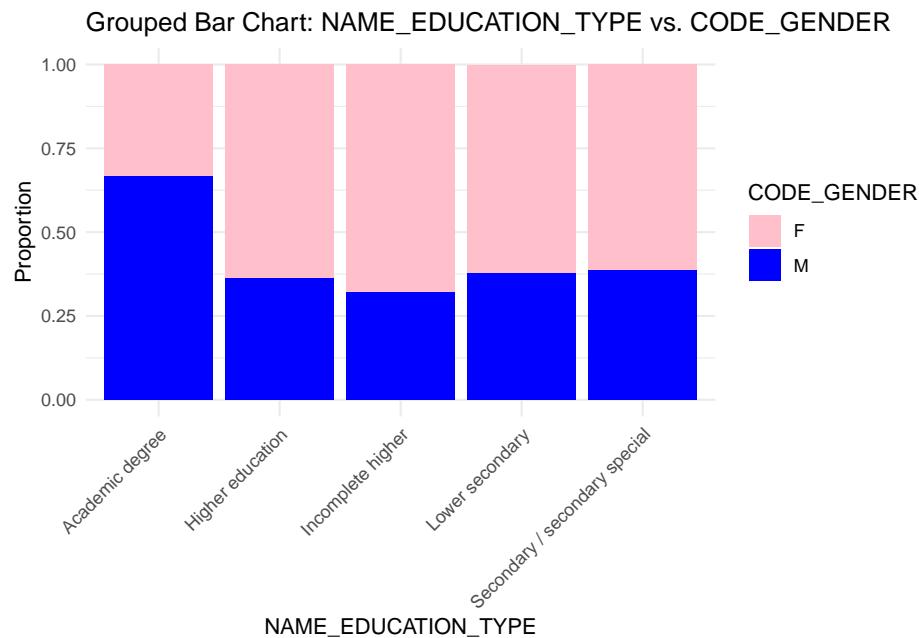
A pesar de que hay más mujeres en la base de datos, un análisis inicial de los datos revela que los hombres han tenido más dificultades para cumplir con los pagos en comparación con las mujeres.

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.

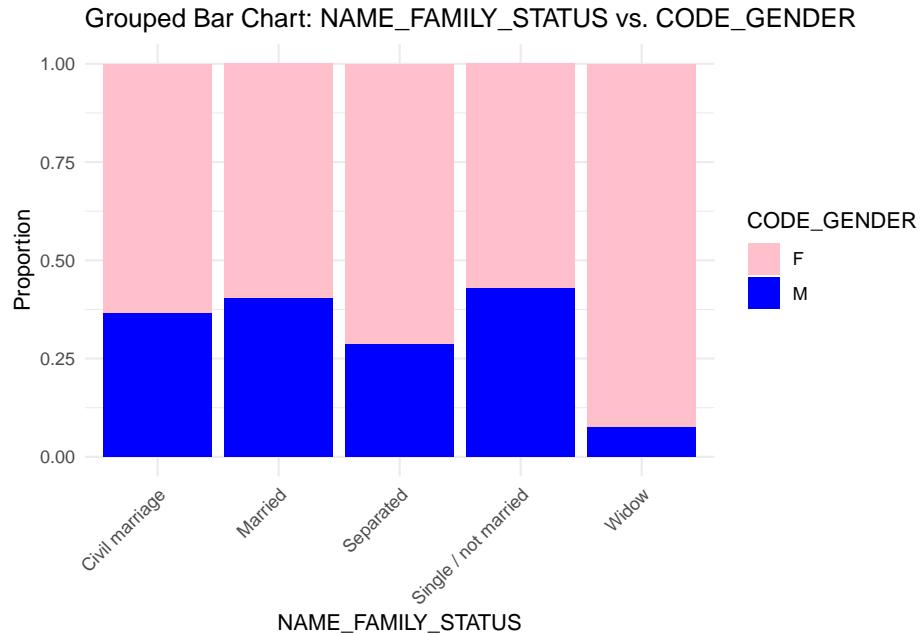
Figura 21: Stacked bar chart NAME EDUCATION TYPE vs CODE GENDER



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.

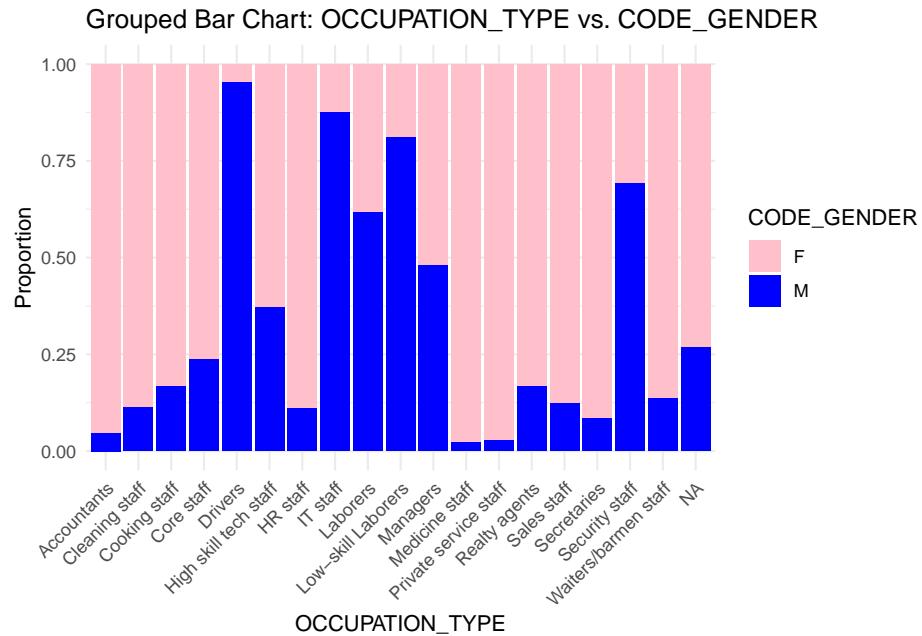
Cabe señalar que la clase “Academic degree” solo cuenta con tres sujetos, mientras que las categorías “Higher education” y “Secondary/secondary special” concentran el 95.28 % de los datos. Estos porcentajes se mantienen aproximadamente constantes en las diferentes categorías, lo que sugiere que la variable 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



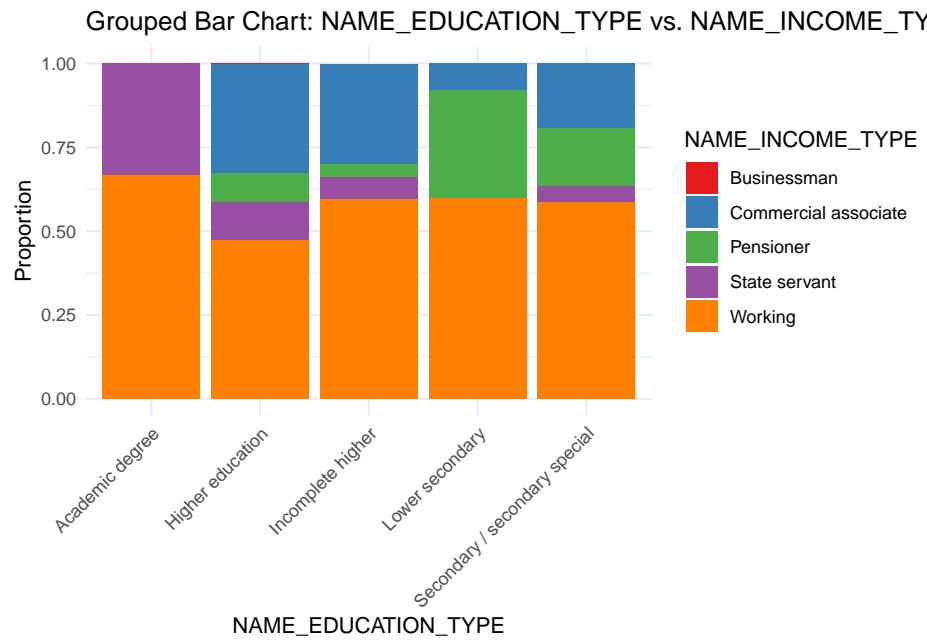
Este gráfico nos muestra la distribución del estatus civil con respecto al sexo de la persona. Se aprecia claramente como la gran mayoría de cónyugues supervivientes son mujeres, mientras que hay una menor desproporción en cuanto a la cantidad de personas solteras o no casadas. La clase con mayor frecuencia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



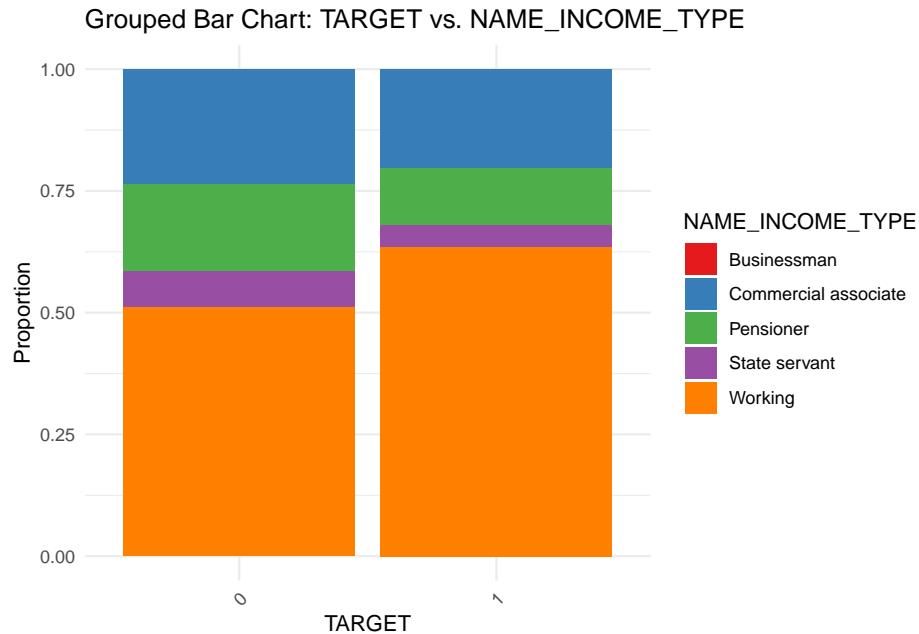
En este último gráfico respecto al género se observa la relación de esta categoría conjuntamente con el tipo de ocupación laboral. En un primer análisis visual se observa como las clases “Drivers”, “IT staff”, “Laborers” y “Security staff”, mientras que las mujeres predominan en la mayoría del resto 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



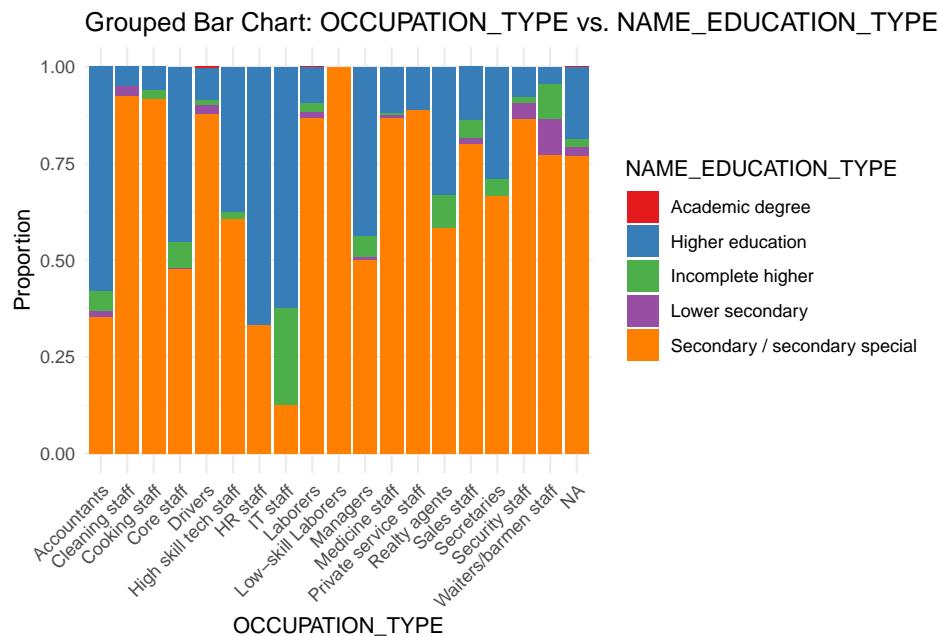
En este gráfico se analiza la relación entre el nivel de educación y el tipo de ingreso. Como se observa, la mayoría de los trabajadores en empleos del sector privado convencional presentan una diversidad de niveles educativos, mientras que aquellos con estudios académicos tienden a trabajar para el 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.

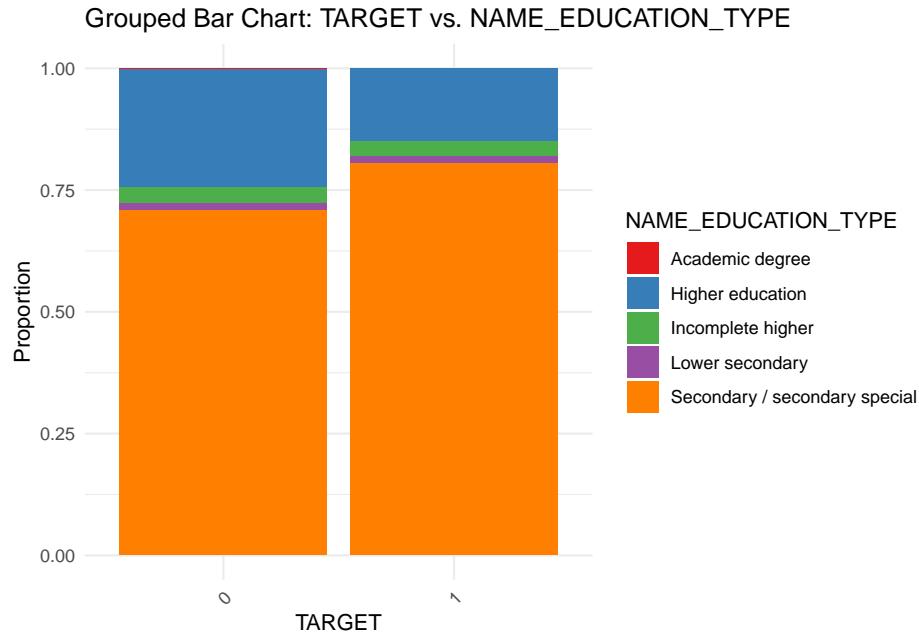
Figura 26: Grouped bar chart NAME EDUCATION TYPE vs OCCUPATION TYPE



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

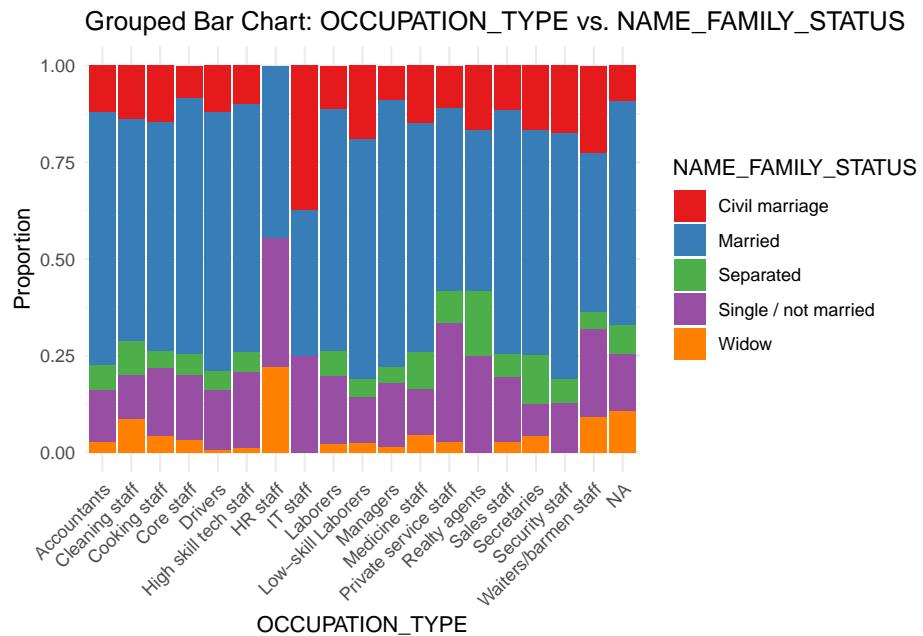
puestos de trabajo que requieren un mayor nivel de conocimientos técnicos, mientras que aquellos con niveles educativos más bajos suelen desempeñar empleos que demandan menos destrezas técnicas.

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

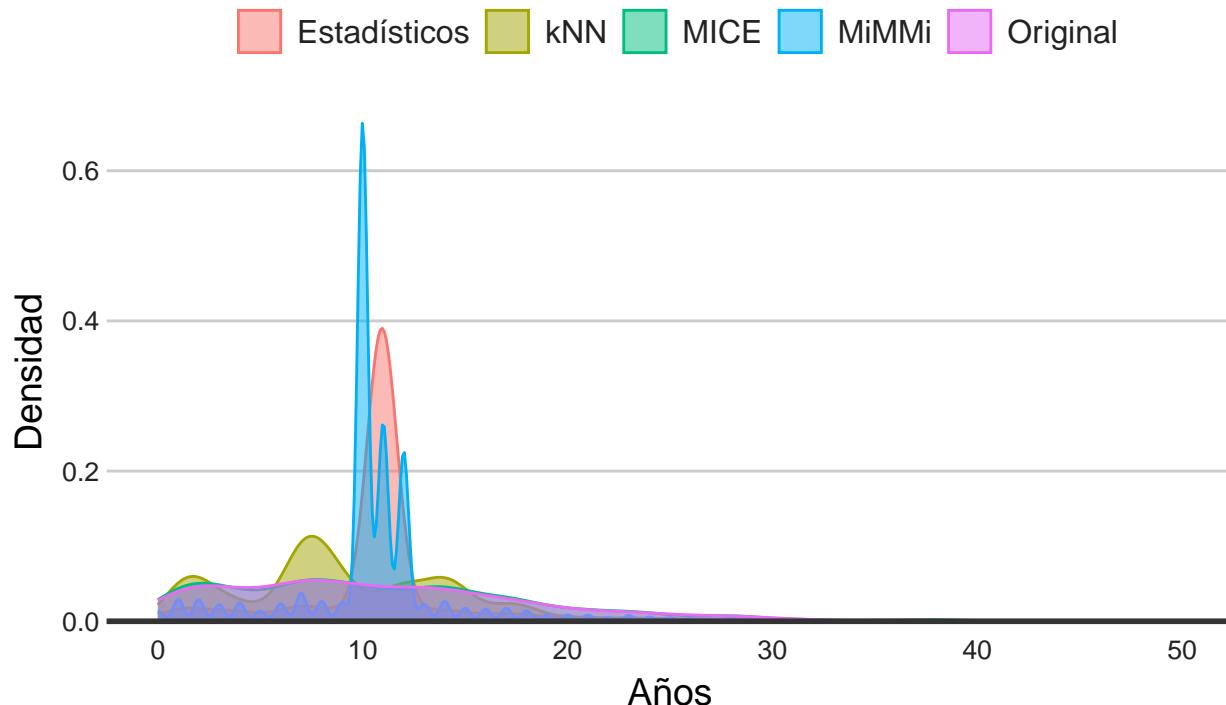
Llegados a este punto, en el momento de seleccionar el método de imputación elegido para el método de imputación final. En nuestro caso, como únicamente disponemos de dos variables numéricas con missings, podemos comparar la función de densidad de los datos originales contra los imputados por cada método.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



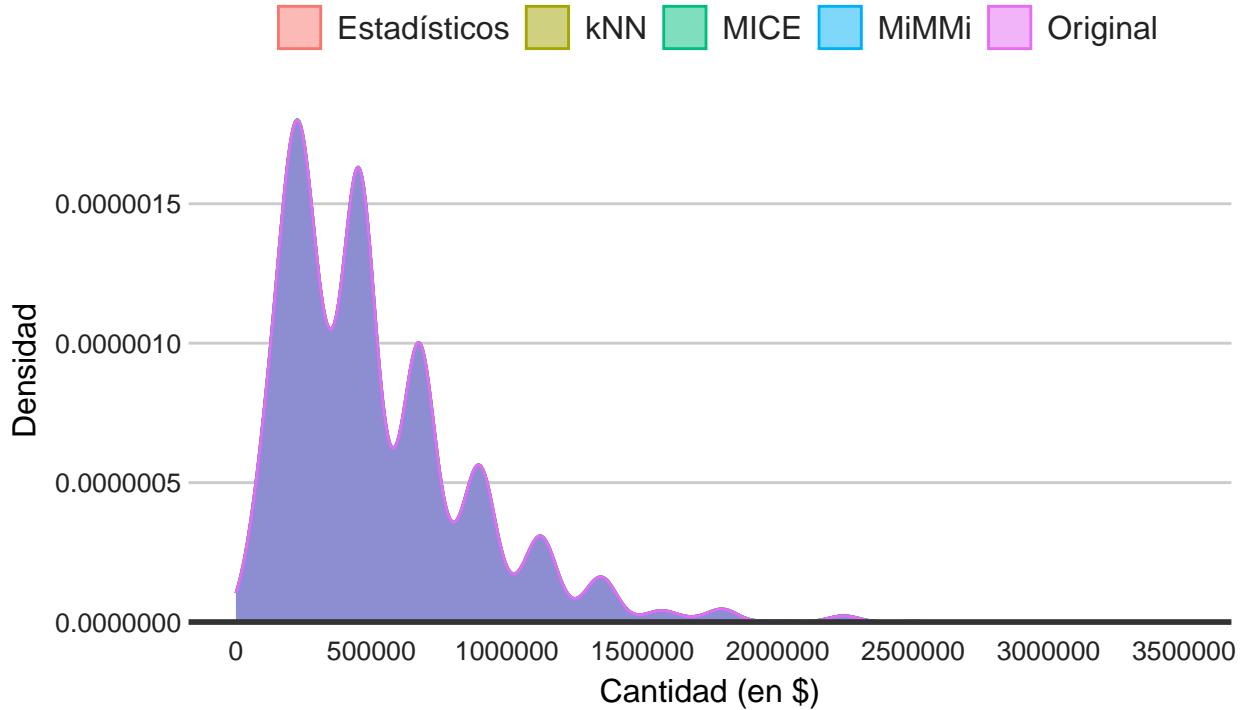
Como se puede apreciar, hay tres métodos de imputación que claramente se alejan mucho de la distribución inicial de los datos: criterios estadísticos, kNN y MiMMi. Así pues, se puede apreciar como el MICE es el algoritmo que aproxima la densidad de los datos a los originales, de forma que este será el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



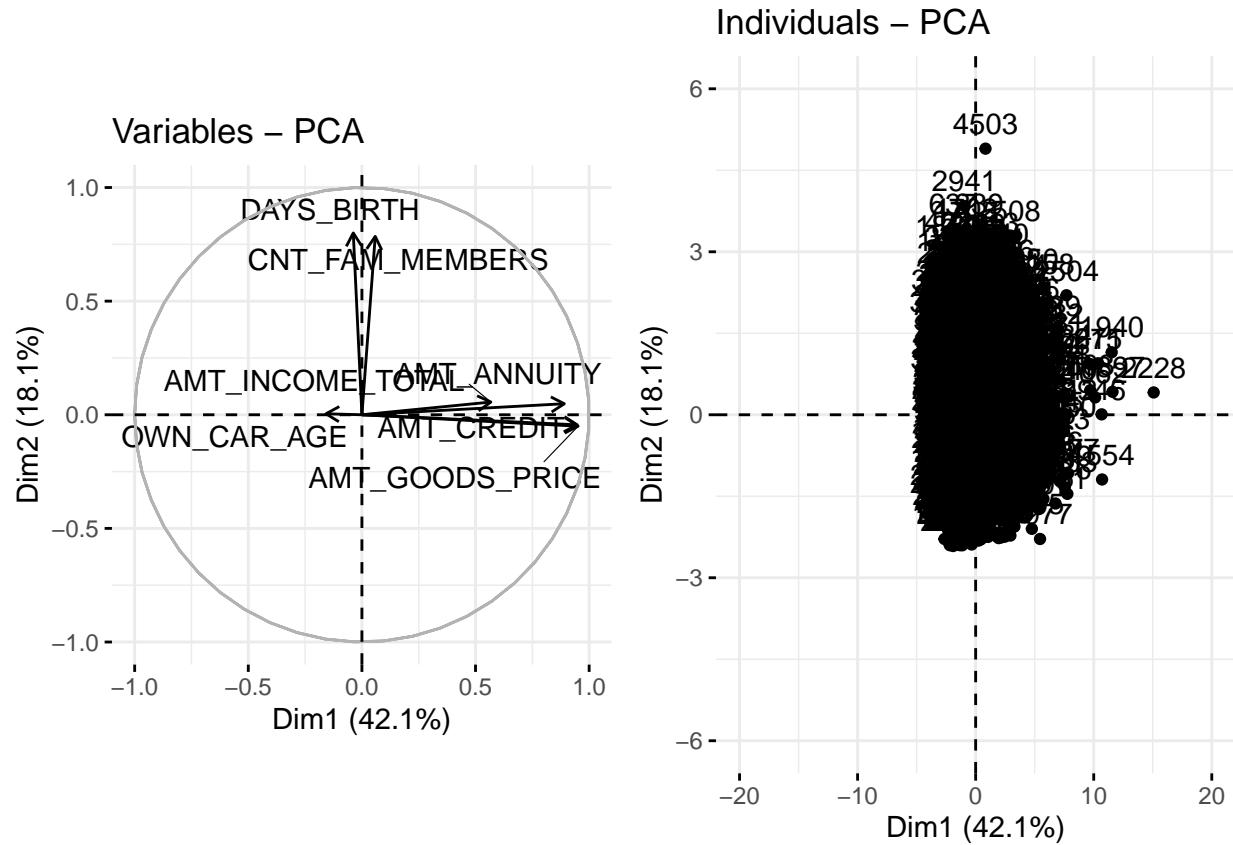
Como se puede apreciar, todos los métodos retornan una estimación similar de la densidad, por lo que se podría decir que es indiferente escoger un método en concreto. De esta forma, se decide usar el MICE como método de imputación final seleccionado.

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

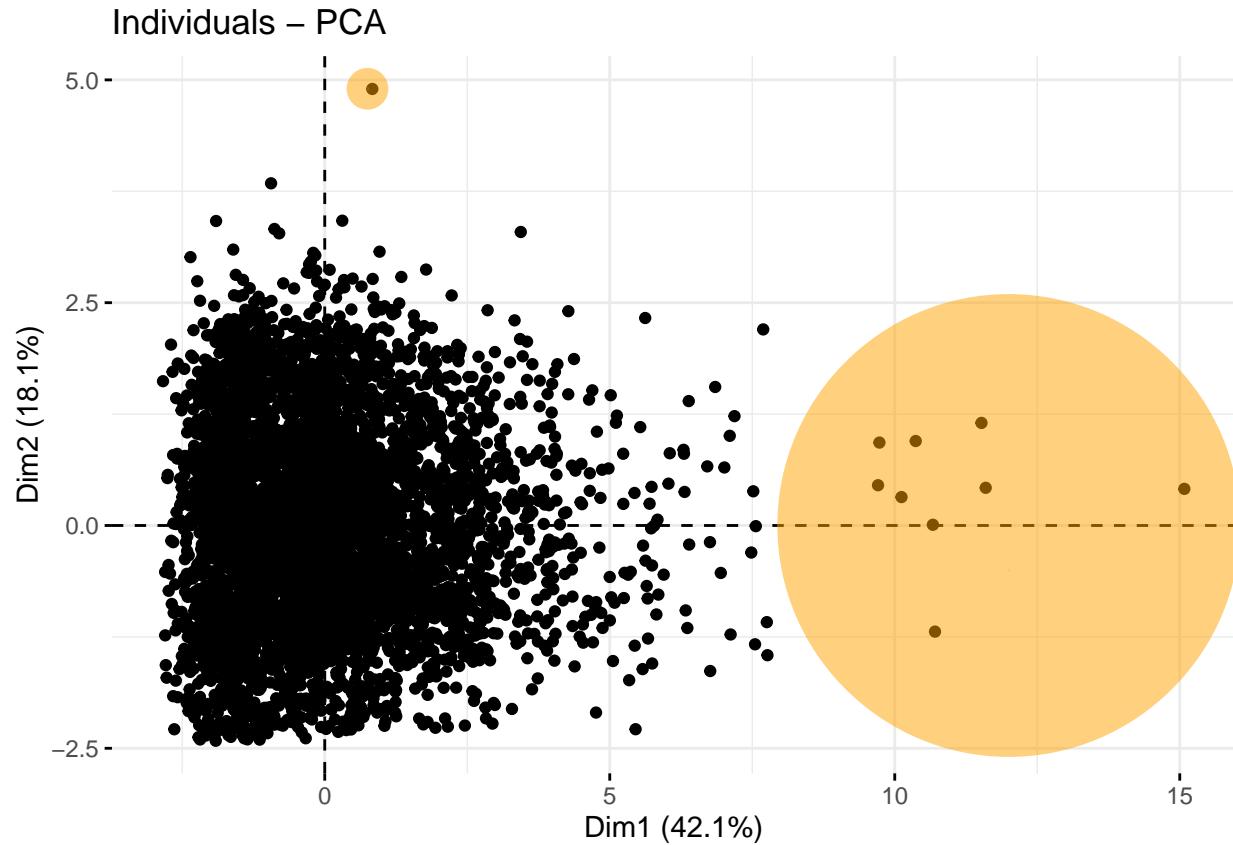
| | OWN_CAR_AGE | AMT_GOODS_PRICE |
|--------------|-------------|-----------------|
| Estadísticos | No | Yes |
| kNN | No | Yes |
| MICE | Yes | Yes |
| MiMMi | No | Yes |

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:



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:



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

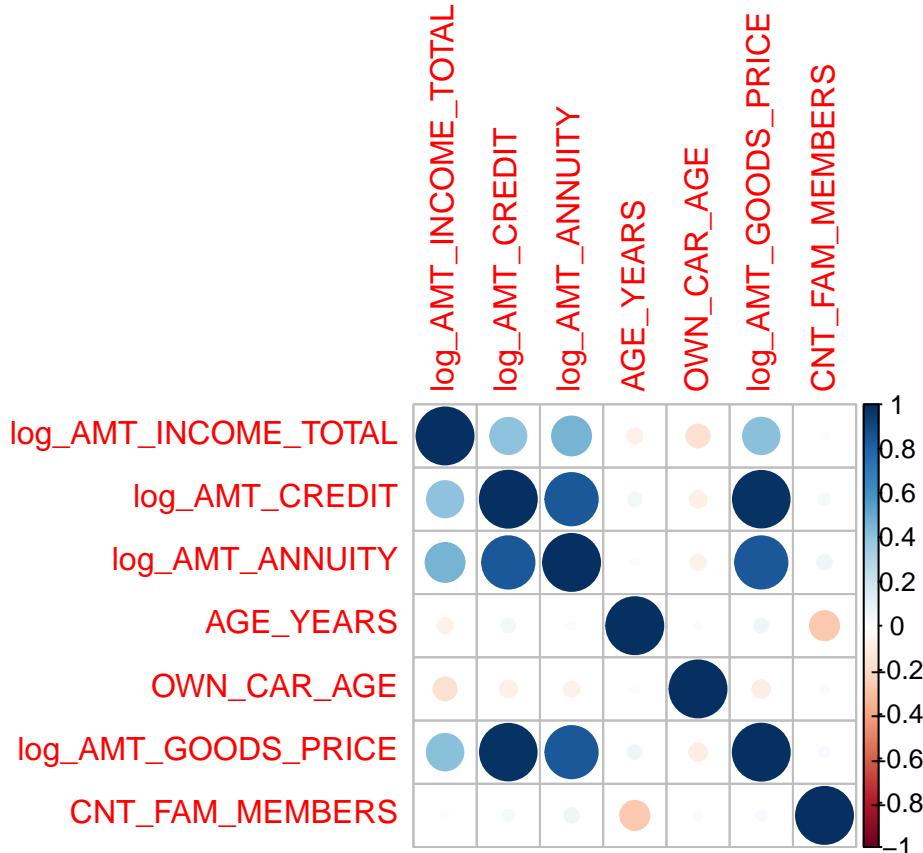
Por último, realizaremos la selección de variables final para nuestra base de datos, así como aplicar transformaciones correctas a nuestras variables para que cumplan algunas hipótesis, como normalidad o heteroscedasticidad. Para este apartado se hace una disección de cada variable una a una.

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.

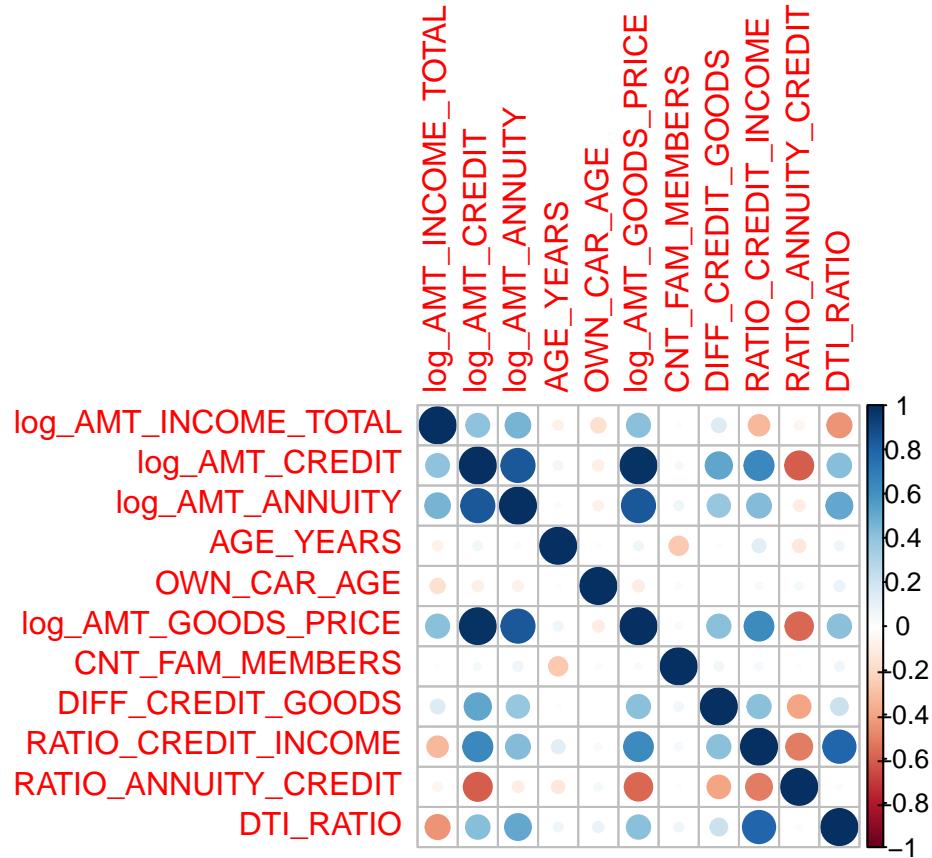
Ahora, vamos a unir aquellas variables ya preprocesadas con el objetivo de tener el dataset preparado para crear nuevas variables.

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

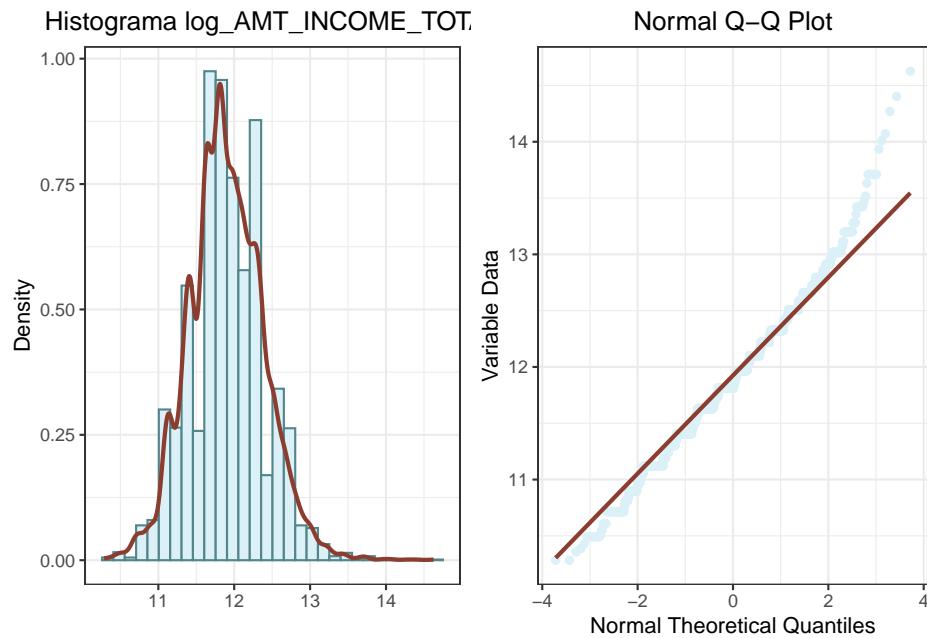
| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|----------------------|------|------|----------|----------|----------|----------|----------|------------|-----------|-----------|-------|----------|--------|
| log_AMT_INCOME_TOTAL | 1 | 5000 | 11.90 | 0.49 | 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 | 63203.11 | 68967.05 | 47520.00 | 52964.12 | 70453.15 | -225000.00 | 361746.00 | 586746.00 | 1.21 | 1.53 | 975.34 |
| 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.

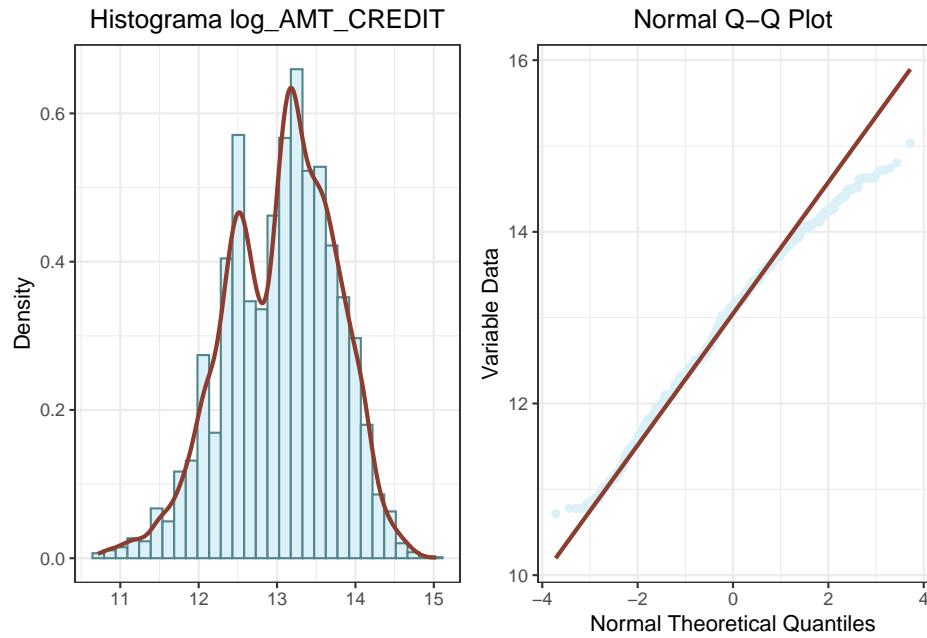
Figura 29: Análisis Gráfico Variable Year Birth



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.

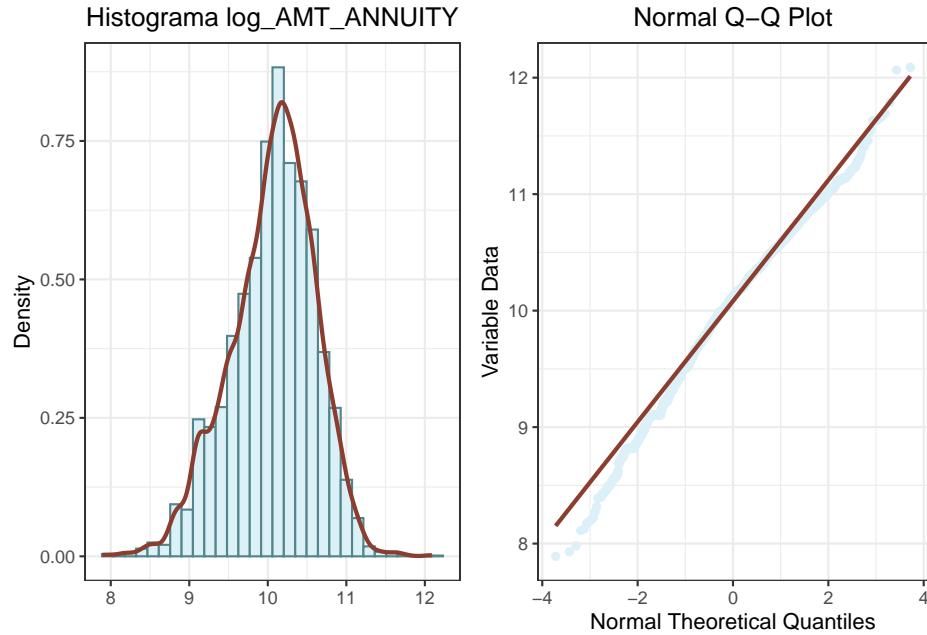
El gráfico Q-Q Plot muestra una notable mejora en la similitud de los cuantiles con los cuantiles teóricos, lo que sugiere una distribución más próxima a la normal. A pesar de este acercamiento visual a la normalidad, los resultados del test de normalidad “Shapiro-Wilk” confirman la hipótesis previa 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.

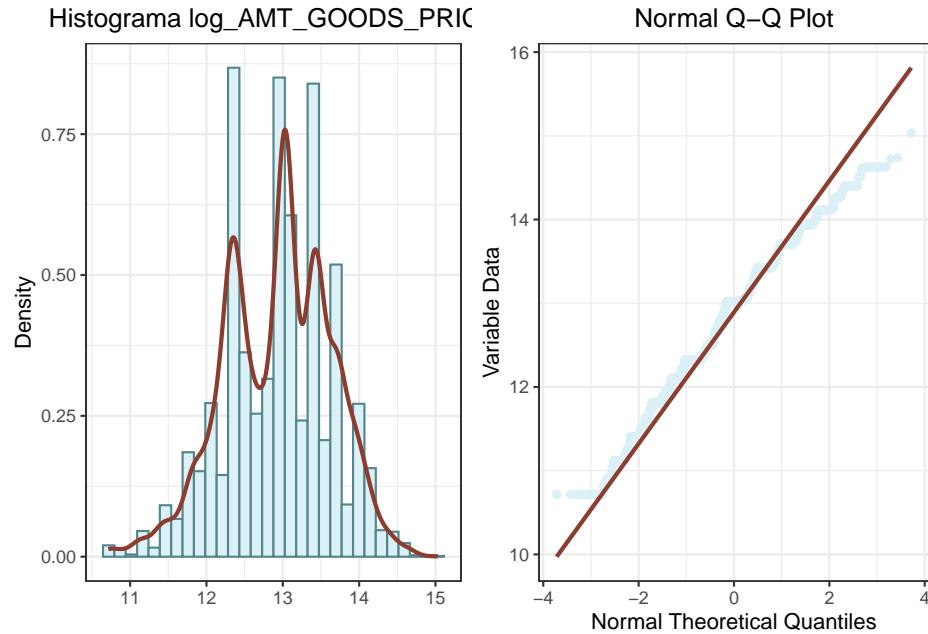
Figura 31: Análisis Gráfico Variable Year Birth



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

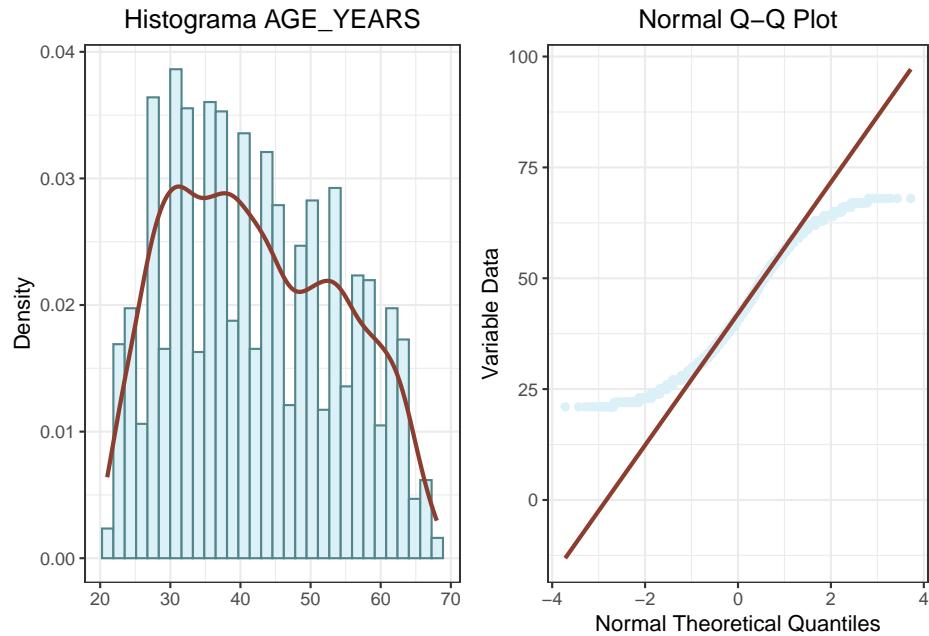
parece que esta variable sigue una distribución normal, en contraste con lo que se observó en el análisis descriptivo previo al procesamiento de datos. Sin embargo, el test de Shapiro-Wilk arroja un valor de ‘r s[[3]]’, indicando que la variable no sigue una distribución normal.

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]]’.

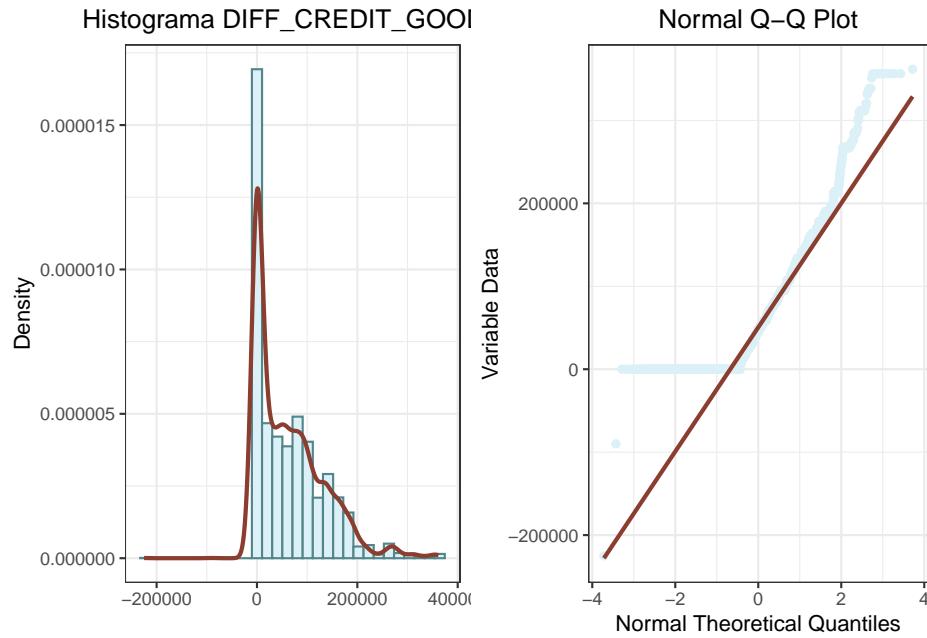
Figura 33: Análisis Gráfico post Variable AGE YEARS



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.

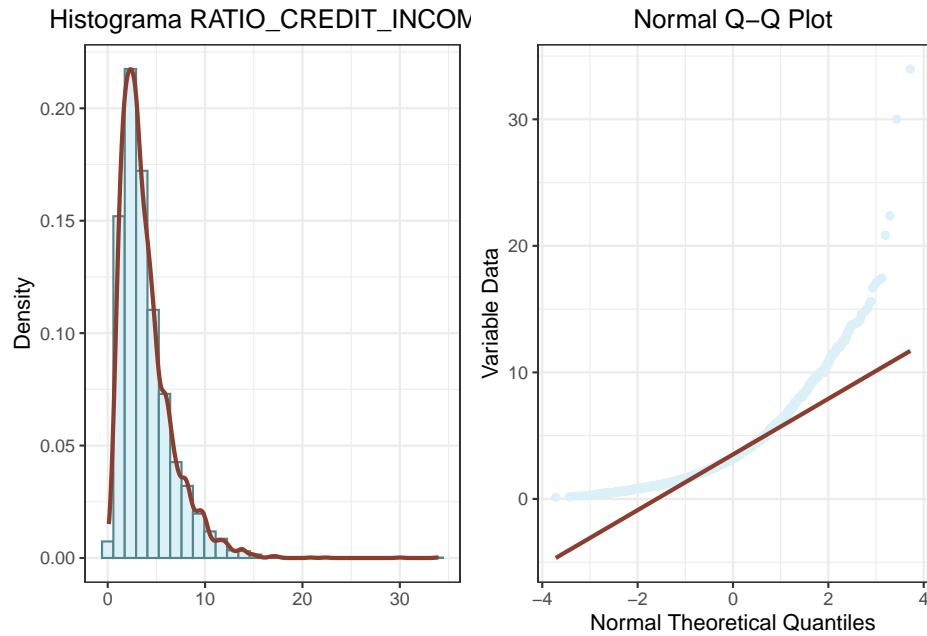
Las limitaciones legales y financieras imponen una clara sesgación en la distribución de edades de los solicitantes de crédito, lo que se refleja en la falta de normalidad en la variable “age_years”. Además, la calidad crediticia y el riesgo crediticio varían significativamente a lo largo de la vida 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

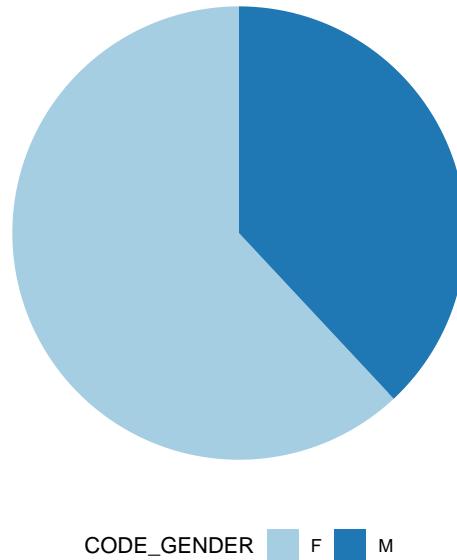
| | [ALL] | N |
|-------------------------------|---------------|------|
| | N=5000 | |
| CODE_GENDER: | | |
| F | 3098 (62.0 %) | 5000 |
| M | 1902 (38.0 %) | |
| NAME_INCOME_TYPE: | | |
| Businessman | 1 (0.02 %) | 5000 |
| Commercial associate | 1111 (22.2 %) | |
| Pensioner | 763 (15.3 %) | |
| State 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

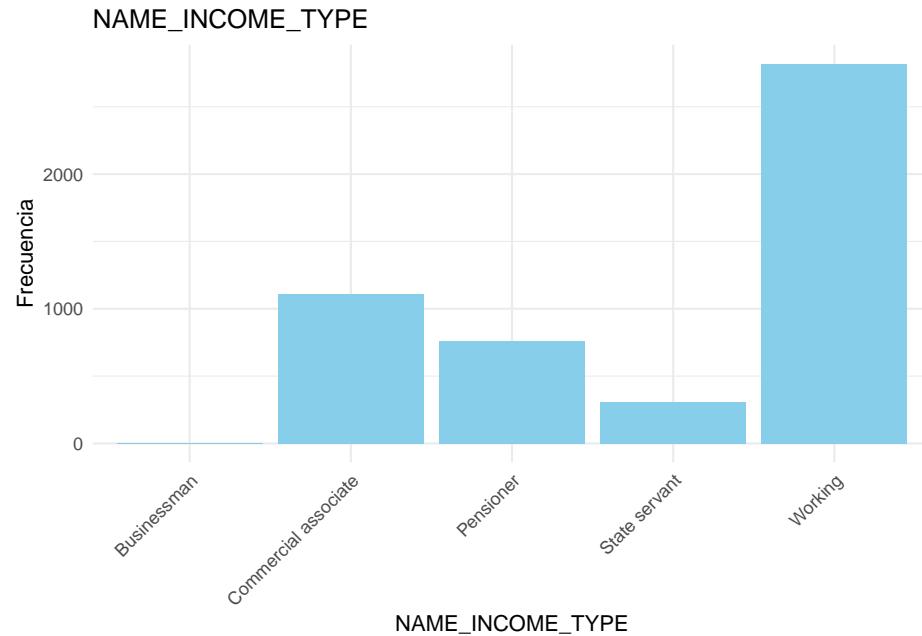


Figura 38: Bar Chart post Variable NAME EDUCATION TYPE

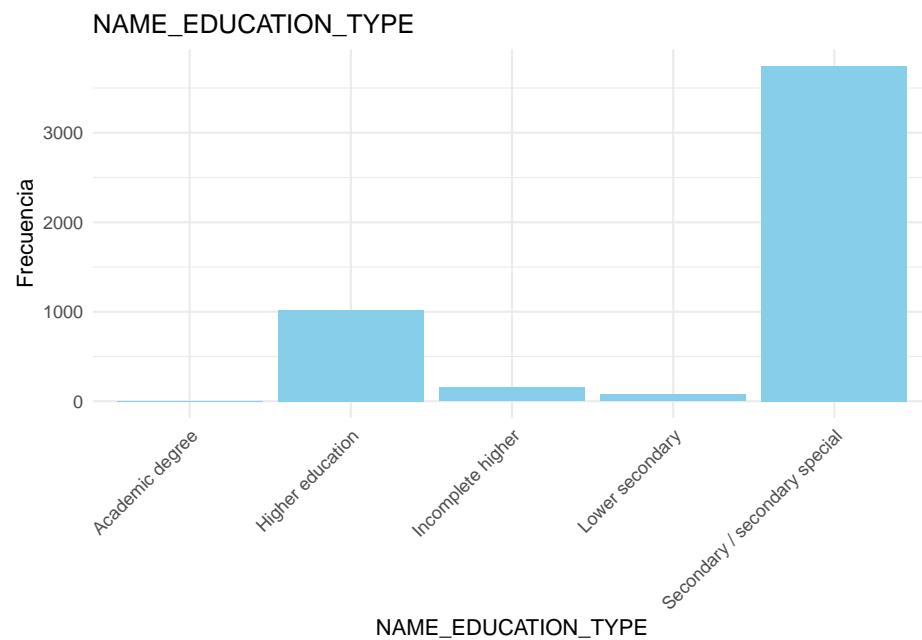
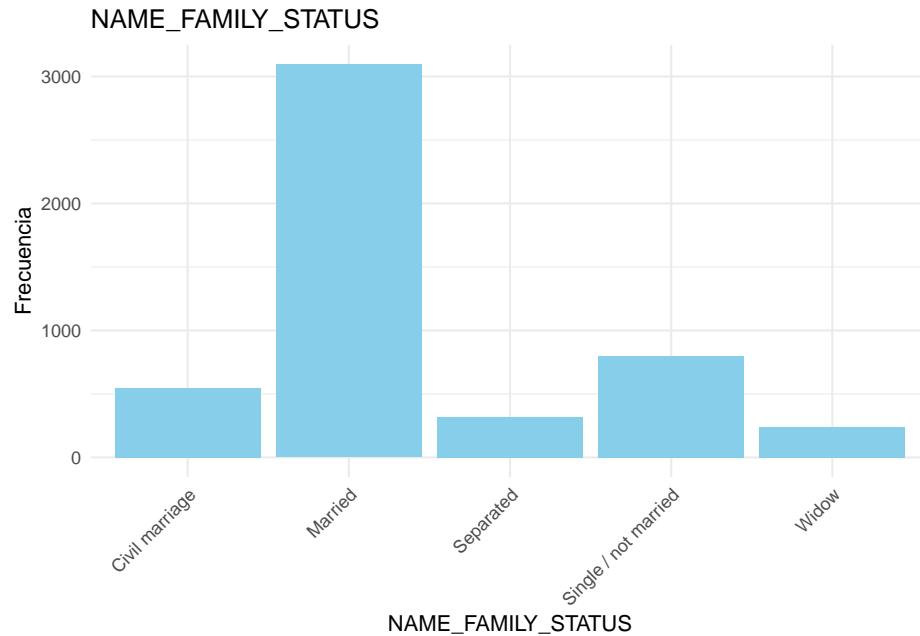
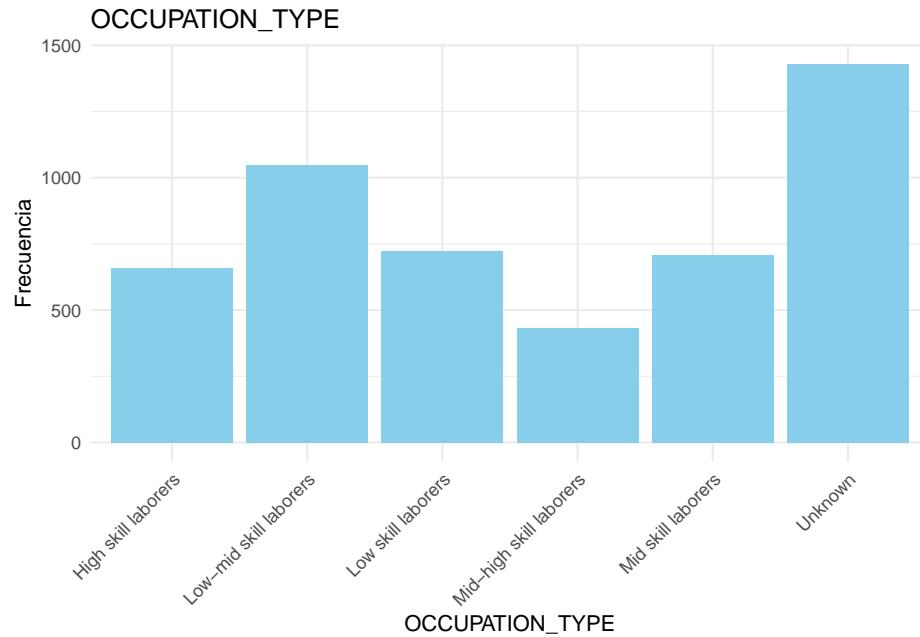


Figura 39: Bar Chart post Variable NAME FAMILY STATUS



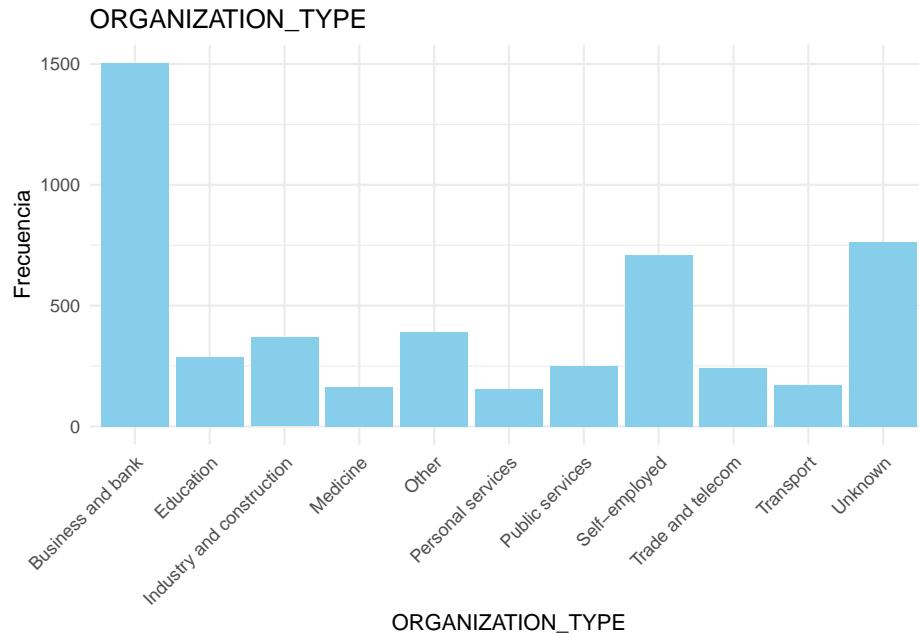
Del mismo modo, las variables NAME_INCOME_TYPE, NAME_EDUCATION_TYPE y NAME_EDUCATION_TYPE mantienen su estructura y patrones previos al preprocesamiento. Esto sugiere que, o bien había pocos valores atípicos o datos faltantes (NA), o que la imputación de datos se realizó de manera precisa. En consecuencia,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

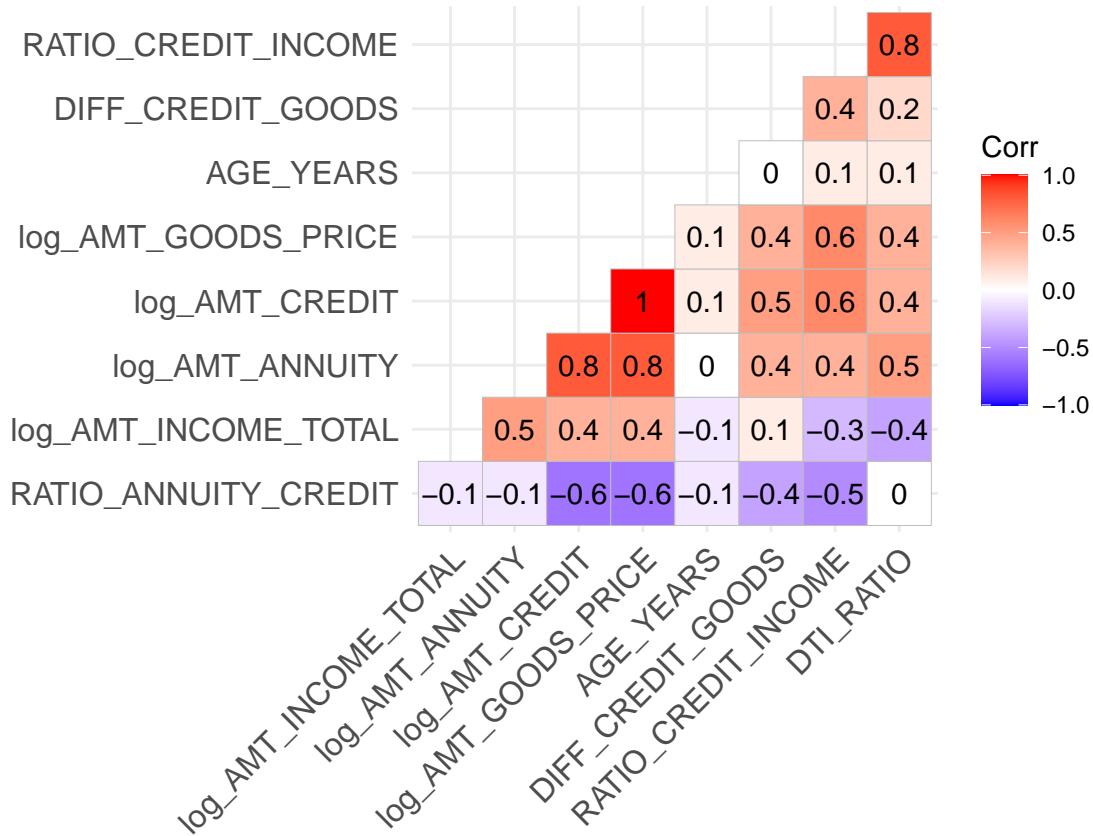


En la variable “Organization type,” se observan cambios en la distribución de los datos debido a la reorganización y a la imputación de valores atípicos y datos faltantes. Previo a la reagrupación, existian diferentes categóricas que podían hacer referencia a un mismo sector o grupo, por lo que al agruparlos 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

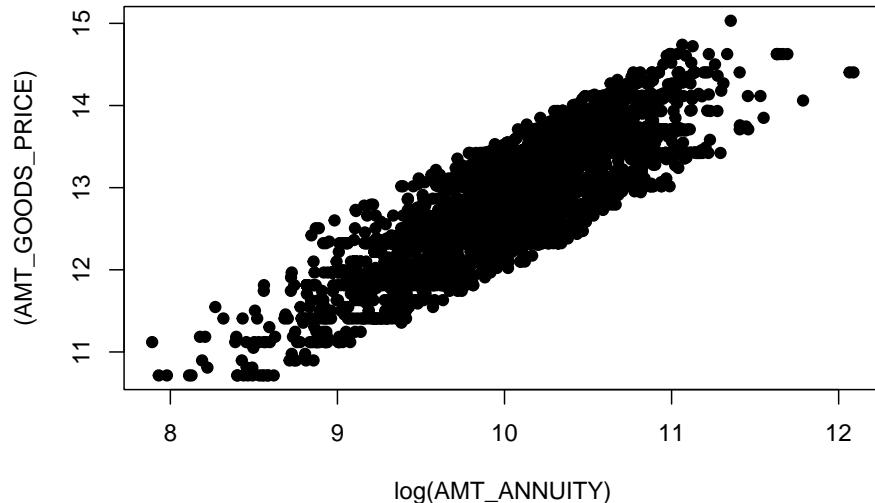
Con el propósito de identificar las relaciones más significativas entre las variables numéricas, se ha creado un gráfico de correlación utilizando la técnica de HeatMap. En este gráfico, los colores indican el grado de dependencia entre las variables numéricas. Cuanto más intenso sea el color, mayor 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

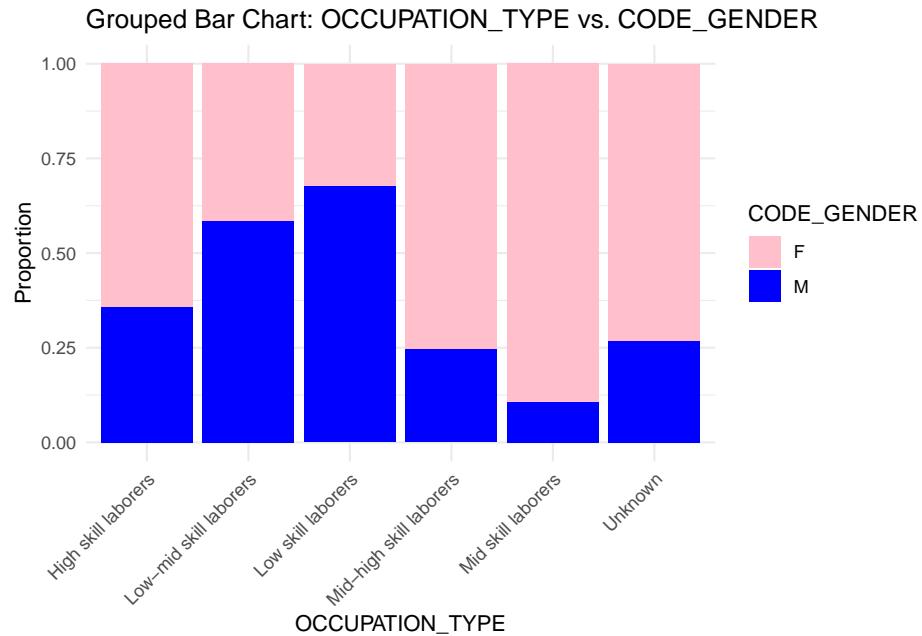


De manera similar, la estructura de los datos se ha mantenido constante, es decir, la correlación entre el valor de los bienes y la anualidad también es relativamente alta. Es importante señalar que los clientes que posean una relación entre la anualidad y el valor del bien que compren (teniendo en cuenta 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

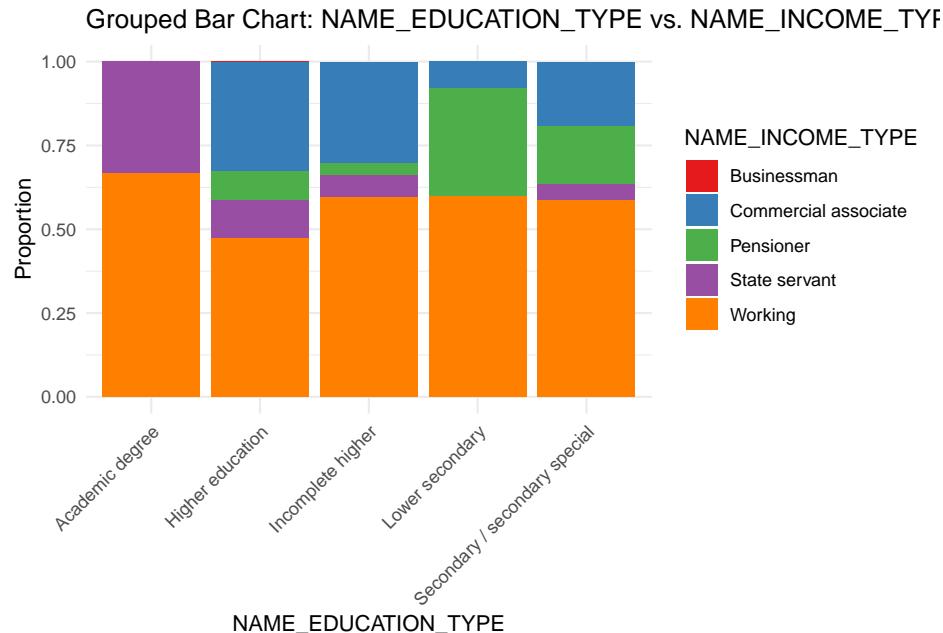
Para concluir el análisis descriptivo antes de proceder al procesamiento de los datos, es necesario examinar la relación entre las variables categóricas y las numéricas. Para este propósito, utilizaremos la creación de varios boxplots, lo que nos permitirá presentar nuestras conclusiones de manera precisa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

Figura 44: Gráfico comparación CODE GENDER vs OCCUPATION TYPE



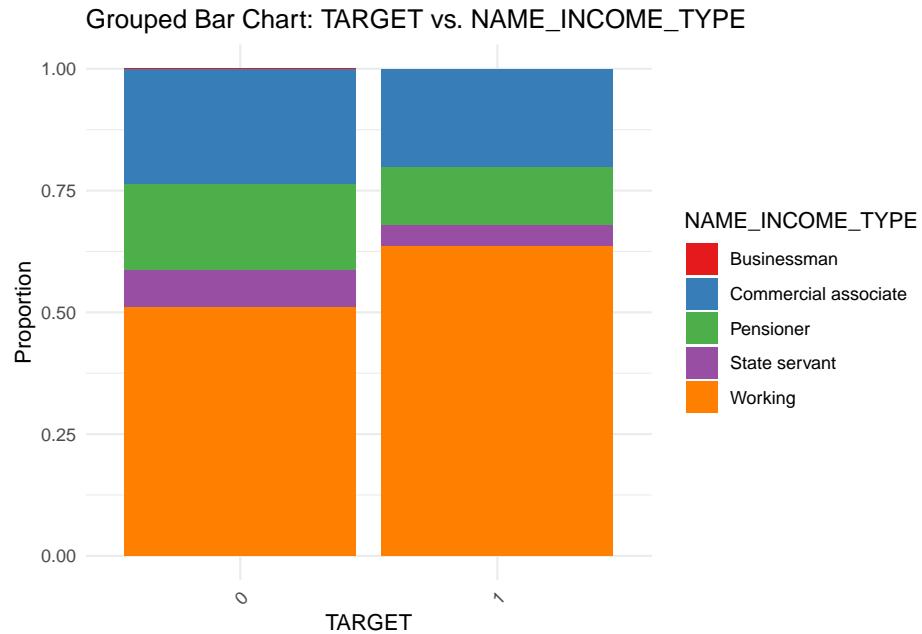
Debido al agrupamiento de las categorías en los distintos niveles de habilidades y responsabilidad, se aprecia como las mujeres tienden a tener puestos de trabajo mas demandantes en cuanto a estos dos atributos respecto a los hombres, completando casi con totalidad los trabajos de “Mid skill laborers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



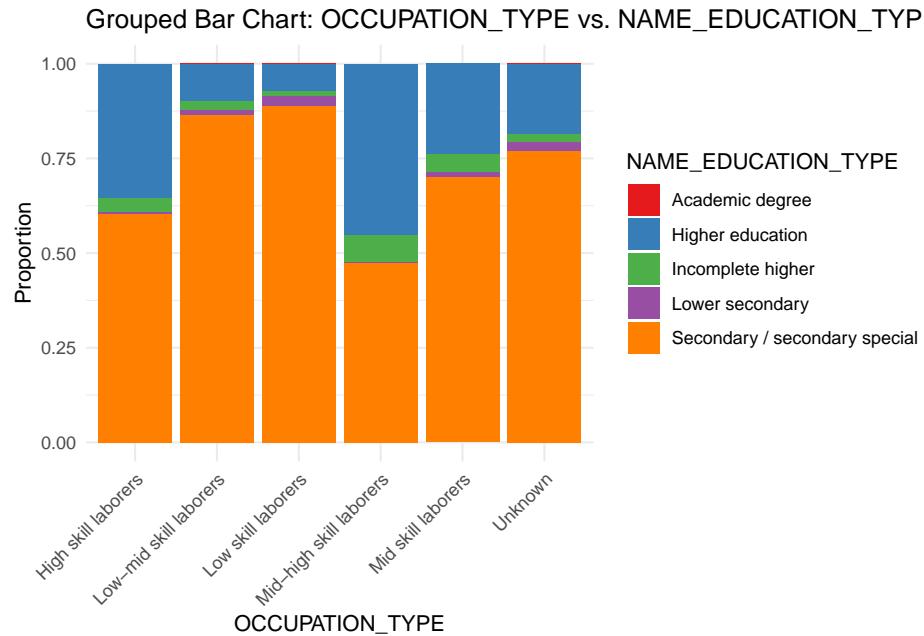
En este gráfico se analiza la relación entre el nivel de educación y el tipo de ingreso. Como se observa, la mayoría de los trabajadores en empleos del sector privado convencional presentan una diversidad de niveles educativos, mientras que aquellos con estudios académicos tienden a trabajar para el 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



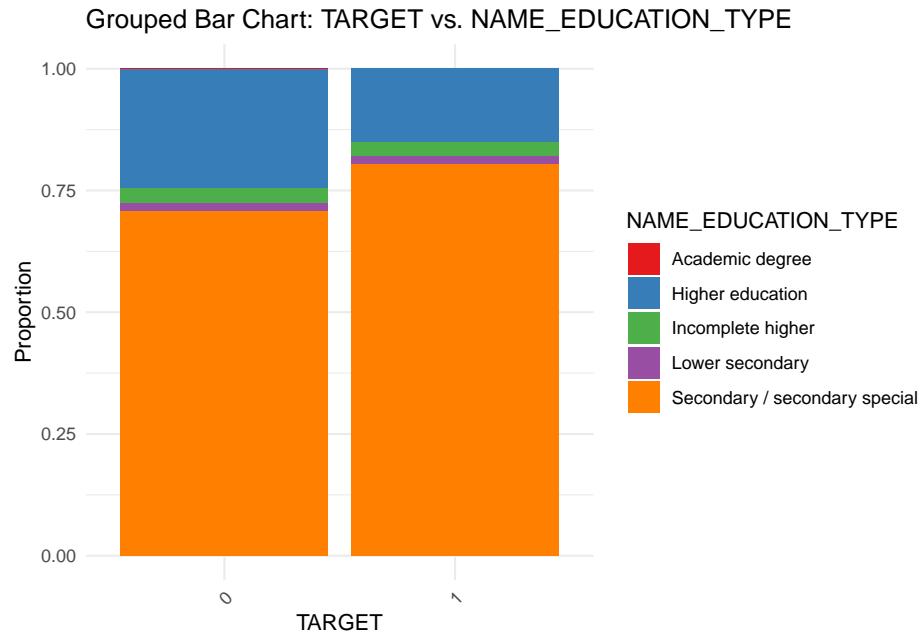
En lo que respecta a la variable TARGET, se observa una disparidad en la capacidad de pago de los clientes en el sector privado, siendo los pensionistas y los comerciales quienes presentan proporcionalmente menos dificultades.

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.

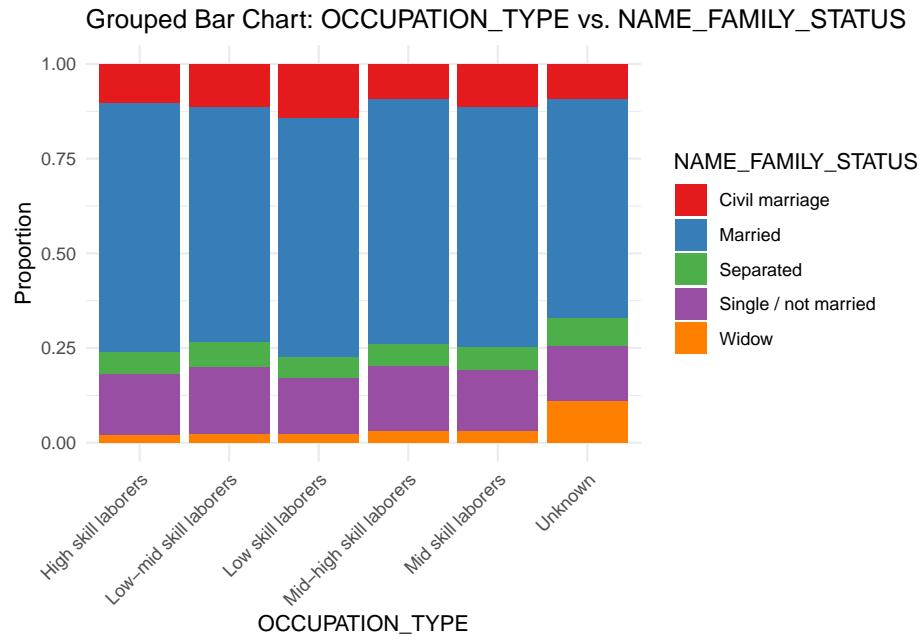
Figura 48: Gráfico comparación post TARGET vs NAME EDUCATION TYPE



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

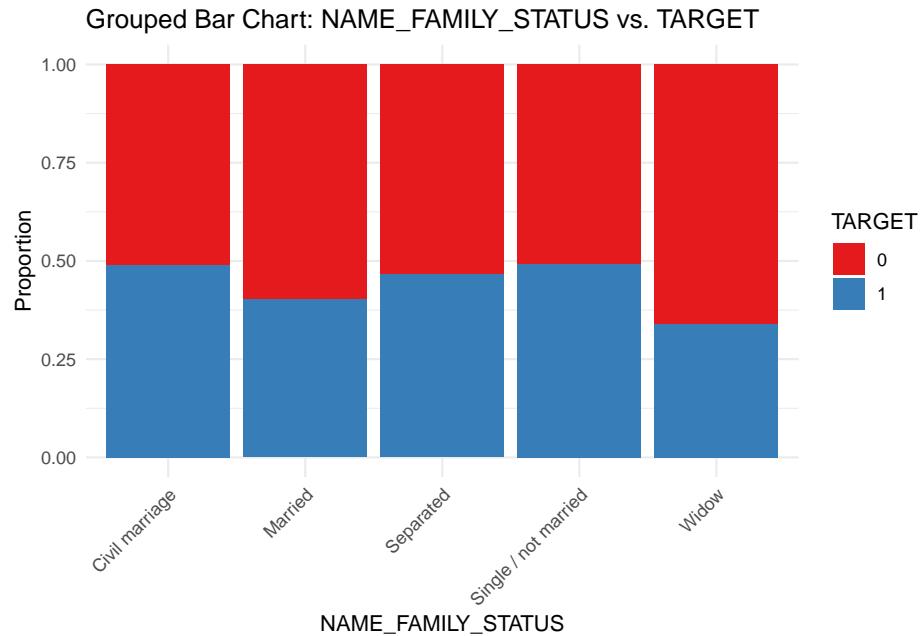
bajo son quienes enfrentan mayores dificultades para cumplir con sus pagos de manera consistente.

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



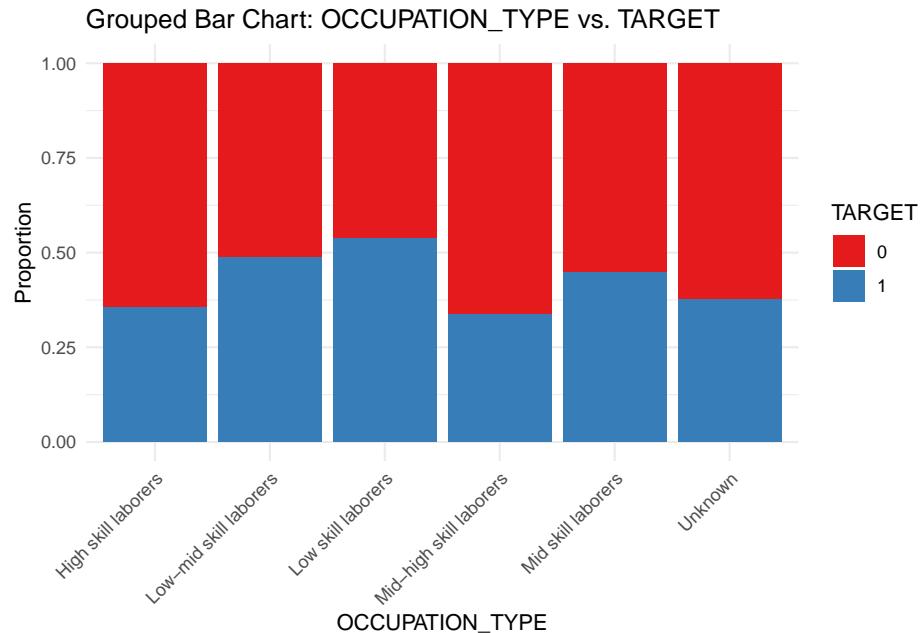
En este gráfico se analiza la relación entre la ocupación de los individuos y el estado civil de ellos mismos. Como se observa, la mayoría de los trabajadores de cualquier sector están casados, muchos por la iglesia y unos pocos civilmente. Vale la pena señalar que un porcentaje significativo de los 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.