Preprocessing

Iker Meneses Sales

2023-09-25

Para realizar el preprocesamiento 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:

Table 1: 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".

Table 2: 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`:

Table 3: 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:

Table 4: 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:

Table 5: 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:

Table 6: 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 algortimo, 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

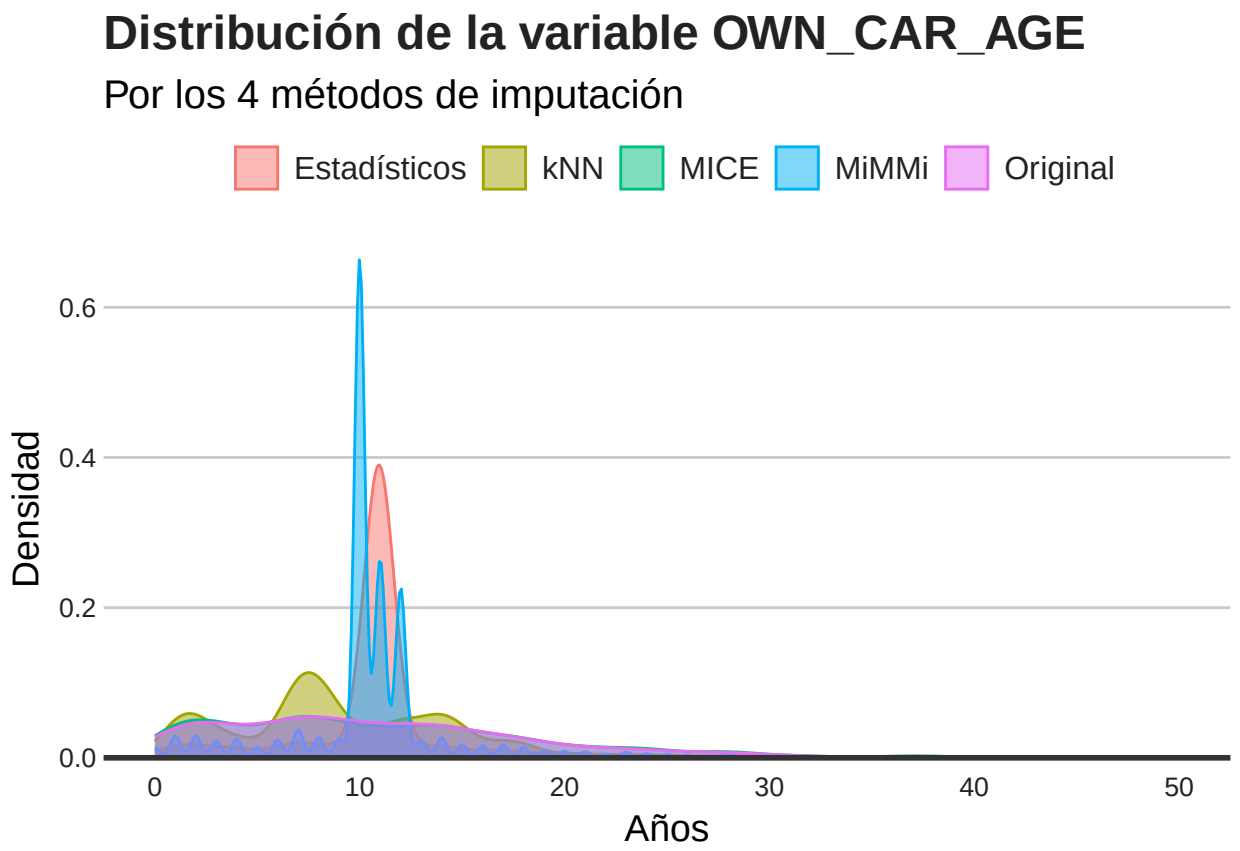
Por último, se recurrirá a imputar a través del MICE como último método de imputación de valores numéricos. El MICE (Multiple Imputation by chained Equations) se basa en un método iterativo a partir del cual se resuelven ecuaciones consecutivamente con el objetivo de imputar valores de la forma más aproximada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

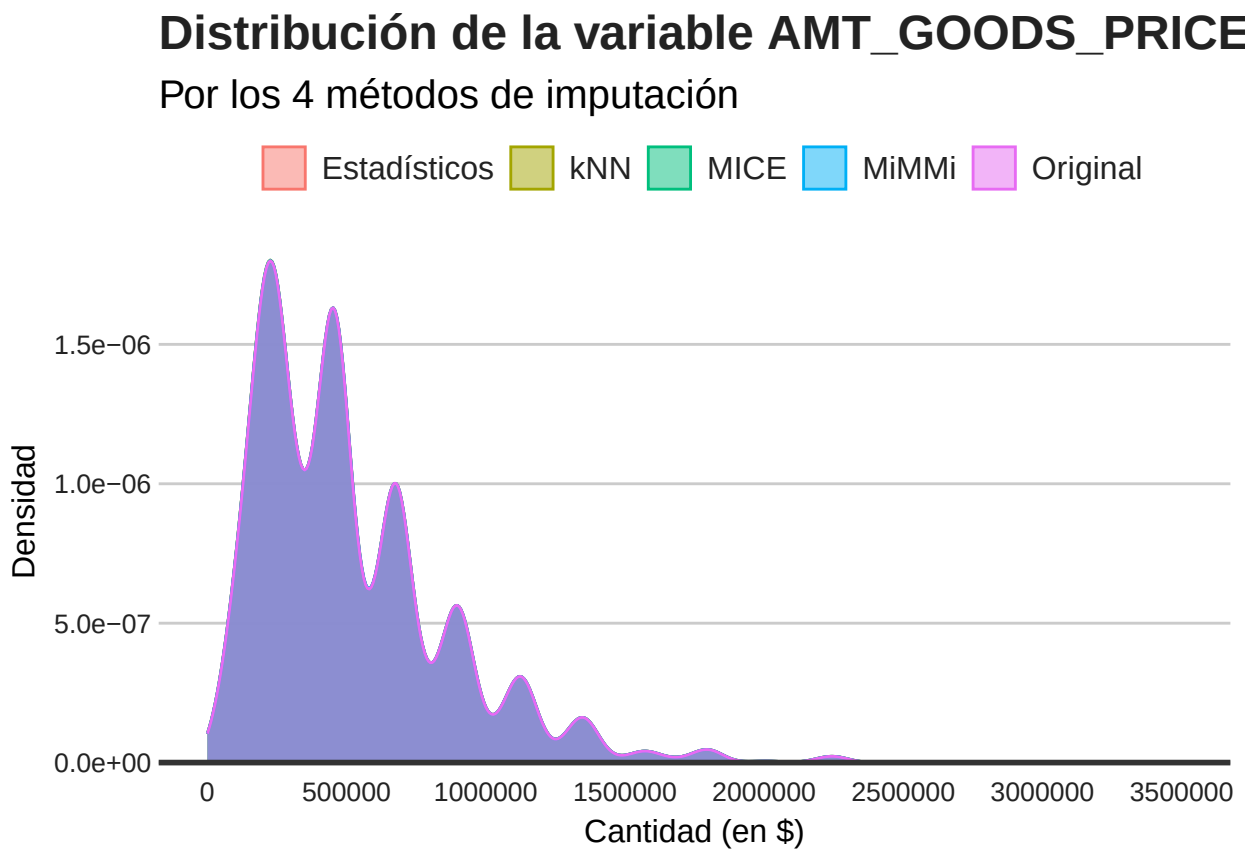
Esta variable es la que presenta más valores no disponibles en nuestra base de datos, de forma que se acepta un mayor margen de error en cuanto a la imputación de valores se refiere. Así, la densidad resultante para cada método es la siguiente:



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 al descriptiva preprocessing, esta variable únicamente presentaba 3 NA, de forma que la densidad en todos los métodos será muy similar:



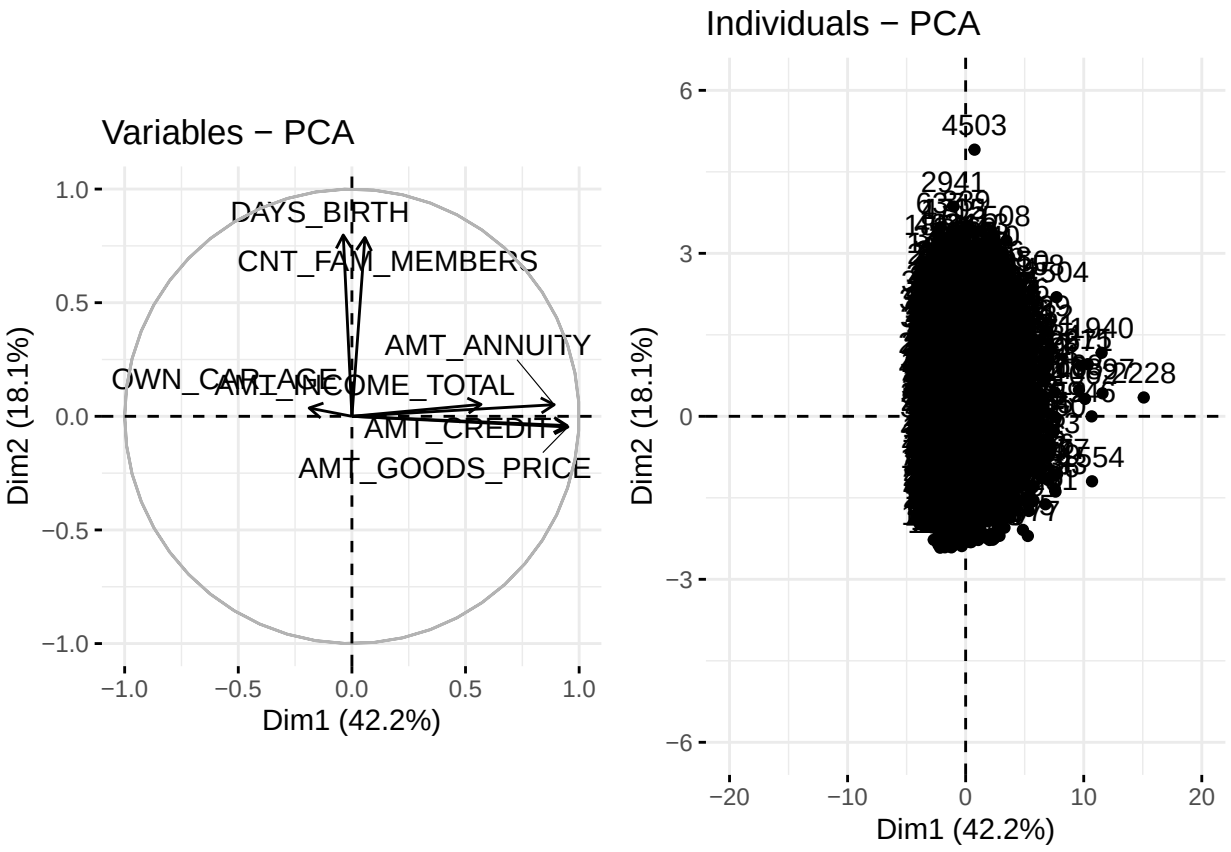
Como se puede apreciar, todos los métodos retornan una estimación similar de la densidad, por lo que se podría decir que es indiferente escoger un método en concreto. De esta forma, se decide usar el MICE como método de imputación final seleccionado.

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

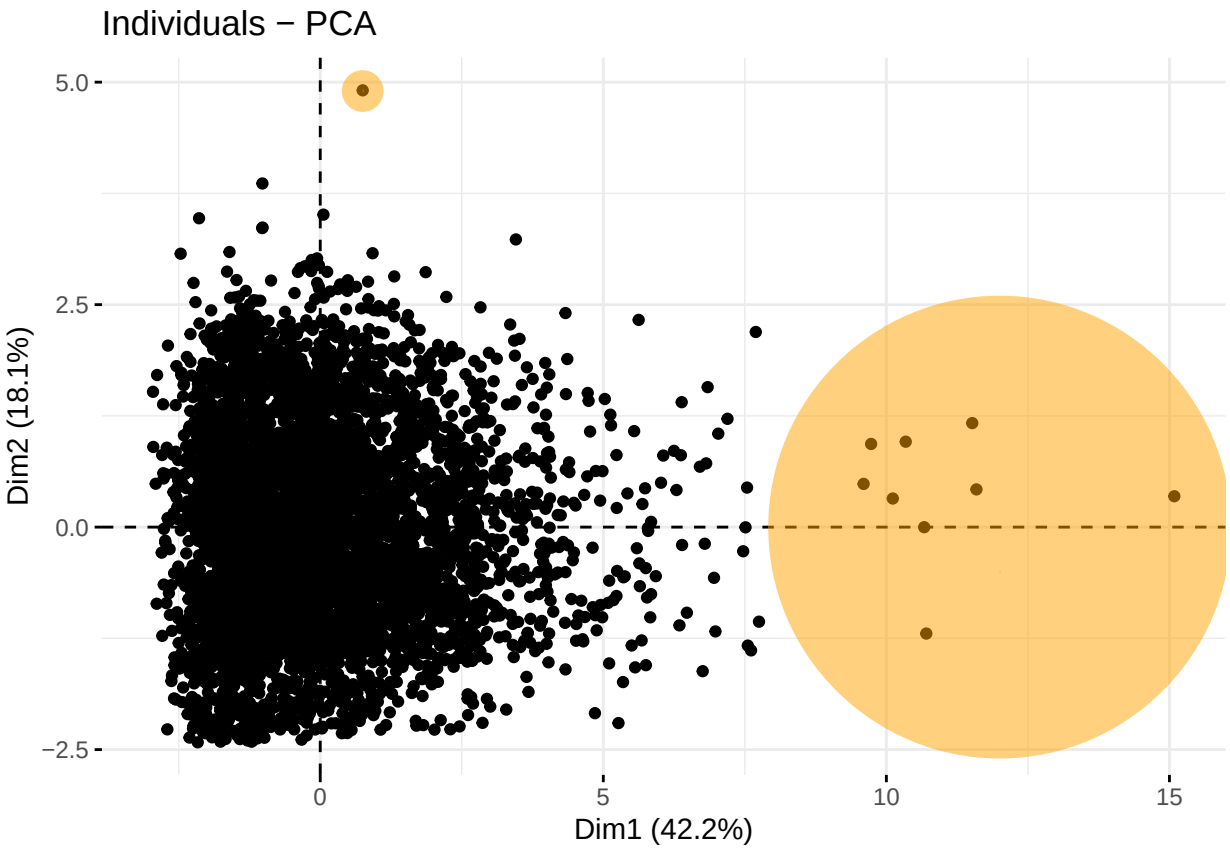
	OWN_CAR_AGE	AMT_GOODS_PRICE
Estadísticos	No	Yes
kNN	No	Yes
MICE	Yes	Yes
MiMMi	No	Yes

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 indiviuos, 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.

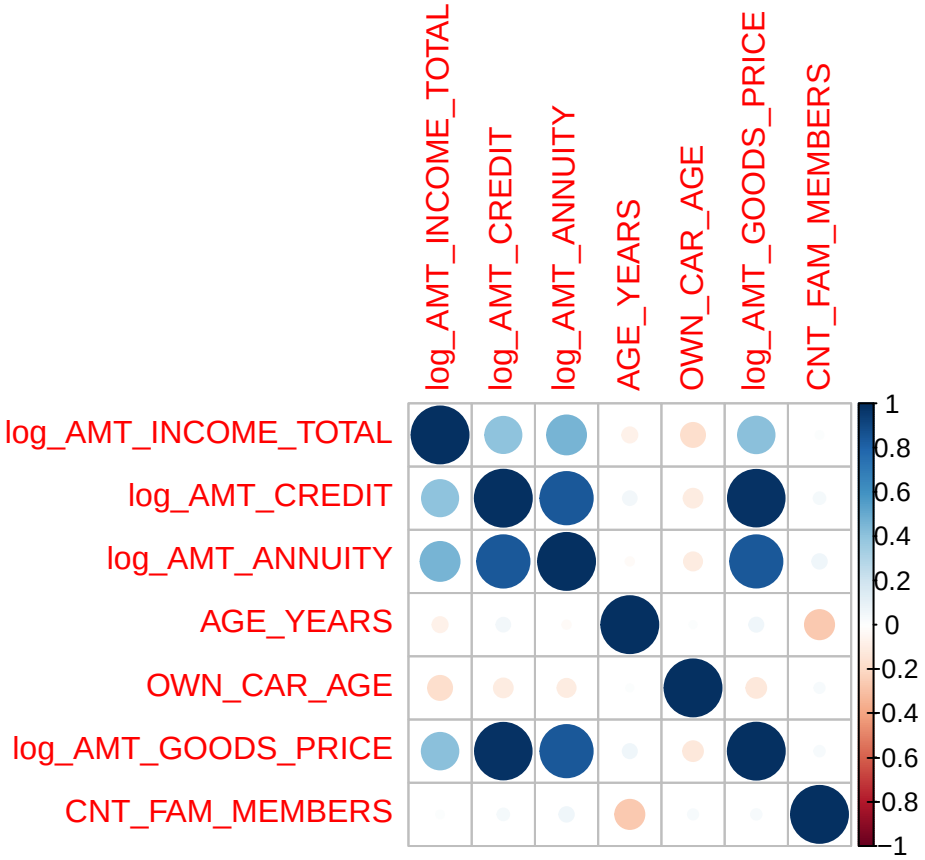
En primer lugar, se resolverán problemas relacionados con las variables numéricas. Como tenemos variables relacionadas con cantidades monetarias (salario, cantidad prestada...), tal vez sería mejor aplicar una transformación logarítmica:

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`.

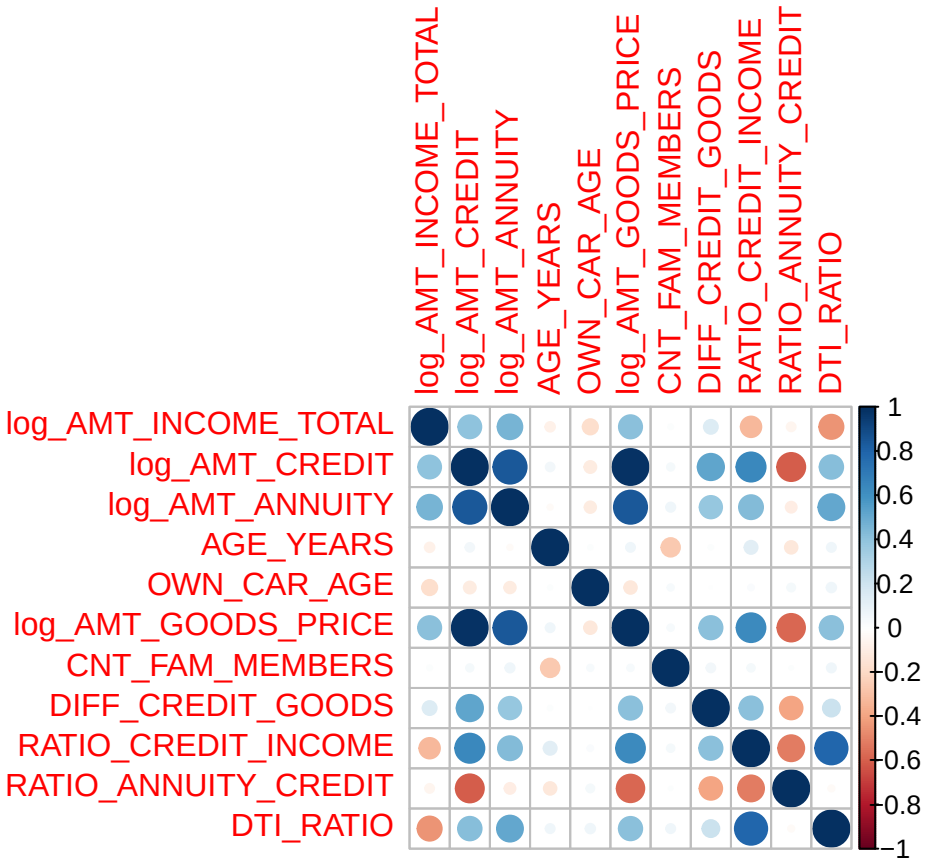
Ahora, vamos a unir aquellas variables ya preprocesadas con el objetivo de tener el dataset preparado para crear nuevas variables.

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 annuity 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 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.