

Trabajo Integrador

Introducción a la Ciencia de Datos

Ignacio Aguilera Martos

22 de Diciembre de 2019

Índice

1. Análisis Exploratorio de los Datos	2
1.1. Conjunto de Regresión	2
1.1.1. Estudio de los estadísticos	2
1.1.2. Estudio de la correlación de las variables	11
1.1.3. Valores perdidos	16
1.1.4. Outliers	16
1.1.5. Distribución de las variables	23
1.2. Conjunto de Clasificación	26
1.2.1. Estudio de los estadísticos	28
1.2.2. Estudio de la correlación de las variables	28
1.2.3. Valores perdidos	28
1.2.4. Outliers	28
1.2.5. Distribución de las variables	28

1. Análisis Exploratorio de los Datos

En esta primera sección vamos a hacer una exploración de los datos intentando extraer algunas conclusiones de los mismos y poder conocer más en profundidad la información que nos arrojan las variables. Esta sección va a estar dividida en dos subsecciones, una para el conjunto de regresión y otra para el conjunto de clasificación.

1.1. Conjunto de Regresión

El conjunto de datos del que dispongo para realizar regresión es el conjunto “treasury”. Vamos a analizar el conjunto de datos.

En primer lugar el conjunto de datos dispone de 16 variables numéricas como podemos observar con el siguiente código.

```
# En primer lugar vamos a visualizar el dataset de regresión.
library(foreign)
dataset_regresion<-read.arff("../DATOS/Datasets Regresion/treasury/treasury.dat")
dataset_regresion

# Ahora vamos a ver el número de variables y el tipo de cada una.
cat("El número de variables es: ", length(colnames(dataset_regresion)), "\n")
cat("El tipo de las variables es:\n")
for(i in 1:length(colnames(dataset_regresion))){
  cat("\t",colnames(dataset_regresion)[i], ": ", class(dataset_regresion[i][[1]]), "\n")
}
```

1Y-CDMaturityRate : numeric
30Y-CDMaturityRate : numeric
3M-Rate-AuctionAverage : numeric
3M-Rate-SecondaryMarket : numeric
3Y-CDMaturityRate : numeric
5Y-CDMaturityRate : numeric
bankCredit : numeric
currency : numeric
demandDeposits : numeric
federalFunds : numeric
moneyStock : numeric
checkableDeposits : numeric
loansLeases : numeric
savingsDeposits : numeric
tradeCurrencies : numeric
1MonthCDRate : numeric

Figura 1: Tipos de las variables y código para obtenerlos

Como podemos ver el conjunto dispone de 16 variables de las cuales todas son numéricas. Esto es lógico pues el conjunto está destinado para el problema de regresión y además nos facilita el trabajo.

Para poder seguir analizando el conjunto vamos a hacer un estudio pormenorizado de las variables del mismo en función a una serie de estadísticos básicos.

Estos estadísticos que voy a emplear son: media, mediana, moda, desviación típica, mínimo, máximo, curtosis y asimetría.

En este conjunto podemos dividir las variables en dos grupos. El primero de ellos corresponde a las variables de entrada que son las que nos van a permitir obtener resultados sobre la salida y el segundo grupo es la propia salida esperada del sistema.

Vamos a empezar a analizar en primer lugar la salida.

1.1.1. Estudio de los estadísticos

Salida

La variable de salida es la que tiene por nombre 1MonthCDRate. Los estadísticos que nos arroja esta variable son los siguientes:

```

1MonthCDRate :
  Media: 7.521945
  Mediana: 6.61
  Desviación típica: 3.377216
  Moda: 5.56
  Kurtosis: 1.733596
  Asimetría: 1.32818
  Mínimo: 3.02
  Máximo: 20.76

```

Figura 2: Estadísticos de la variable de salida.

Podemos observar que la media de la salida es aproximadamente de 7,5 y su desviación típica de 3,3, esto nos indica que la mayoría de los datos (el 95 % en caso de estar ante una distribución normal) van a estar dentro del intervalo $[0,9, 14, 1]$. Aún así podemos ver mediante el mínimo y el máximo que los datos se van a mover en el intervalo $[3,02, 20,76]$. Esto parece que nos va a indicar que deberíamos de tener una cola más pesada a la izquierda de la distribución y una cola más alargada a la derecha de la misma. Este hecho viene corroborado también por la mediana. Vemos que es 6,61 (menor que la media) con lo que nos está diciendo que el 50 % de los datos van a estar en el intervalo $[3,02, 6,61]$ y por tanto proporcionalmente esta cola será más pesada.

Otra forma que tenemos de comprobar lo que hemos dicho es mediante el coeficiente de asimetría. Al ser positivo nos está indicando que la distribución es asimétrica a la derecha, cuestión que ya sabemos y hemos razonado.

La curtosis nos está dando una indicación de cómo de puntiaguda o achatada es la distribución. En este caso el coeficiente es positivo, por lo que la distribución será más puntiaguda. Esto nos está indicando que vamos a tener una mayor concentración de datos entorno a la media.

Variable 1: 1Y-CMaturityRate

Los estadísticos que nos arroja esta variable son:

```

1Y-CMaturityRate :
  Media: 97.35363
  Mediana: 92.526
  Desviación típica: 14.47144
  Moda: 86.878
  Kurtosis: 0.2769317
  Asimetría: 1.10812
  Mínimo: 77.055
  Máximo: 142.645

```

Figura 3: Estadísticos de la variable 1 1Y_CMaturityRate

En primer lugar podemos observar que la media es de 97,35363 y la mediana es ligeramente inferior. Esto nos vuelve a indicar que vamos a tener una cola más corta a la izquierda y una más larga a la derecha, es decir, vamos a tener más valores a la derecha de la distribución. Podemos corroborar este hecho también comprobando el mínimo y el máximo. Como podemos

ver el máximo está mucho más alejado de la media que el mínimo por lo que podemos intuir que la distribución va a ser más alargada por ese lado.

De igual forma si miramos el coeficiente de asimetría vemos que es positivo lo que nos está indicando la asimetría de la cola derecha.

La curtosis es positiva pero no muy lejana del cero, por lo que tendrá la forma de una normal pero un poco más puntiaguda.

Variable 2: 30Y-CMortgageRate

Los estadísticos que nos arroja esta variable son:

```
30Y-CMortgageRate :  
  Media:  7.543937  
  Mediana:  6.71  
  Desviación típica:  3.105787  
  Moda:  5.59  
  Kurtosis:  0.4963418  
  Asimetría:  1.026759  
  Mínimo:  3.02  
  Máximo:  17.15
```

Figura 4: Estadísticos de la variable 2 30Y_CMortgageRate

Podemos ver que la media es 7,543937 y la mediana 6,71 lo que de nuevo nos hace sospechar que la cola derecha es más alargada. Esto lo podemos ver (y podemos decir ya que la cola va a ser bastante alargada) con el máximo y el mínimo. Sabemos que el 95 % de los datos debería de estar en el intervalo $[media - 2 * stdv, media + 2 * stdv]$ pero la realidad es que observamos que la cola derecha es más larga. Asimismo vemos que la moda es aún menor que la mediana con lo que corroboramos que la cola izquierda es más pesada y la derecha más alargada.

El coeficiente de asimetría nos termina de corroborar lo que estamos infiriendo pues al ser positivo nos indica una asimetría en la cola derecha.

En cuanto a la curtosis podemos ver que es más puntiaguda que una normal.

Variable 3: 3M-Rate-AuctionAverage

Los estadísticos que nos arroja esta variable son:

```

3M-Rate-AuctionAverage :
  Media: 10.40085
  Mediana: 9.9
  Desviación típica: 2.958872
  Moda: 10.4
  Kurtosis: -0.2401306
  Asimetría: 0.826309
  Mínimo: 6.49
  Máximo: 18.63

```

Figura 5: Estadísticos de la variable 3 3M_Rate_AuctionAverage

Esta variable tiene como media 10,40085 y mediana 9,9. En esta variable no vemos una diferencia tan grande por lo que a priori podemos pensar que no es muy asimétrica. La desviación típica es de 2,958872 por lo que (en caso de que fuese una normal) el 95 % de los datos se van a encontrar en el intervalo [4,483106, 16,318594]. Vemos que el mínimo es 6,49 con lo que podemos entender que la cola izquierda es más corta mientras que el máximo es 18,63 con lo que la cola derecha va a ser algo más alargada.

Si observamos los coeficiente de asimetría y curtosis podemos ver que el coeficiente de asimetría es positivo lo que nos indica que la distribución es asimétrica a la derecha y la curtosis es ligeramente negativa con lo que podemos decir que la distribución será algo más achatada que una distribución normal.

Variable 4: 3M-Rate-SecondaryMarket

Los estadísticos que nos arroja esta variable son:

```

3M-Rate-SecondaryMarket :
  Media: 6.85122
  Mediana: 5.81
  Desviación típica: 2.954287
  Moda: 5.12
  Kurtosis: 1.147822
  Asimetría: 1.206275
  Mínimo: 2.67
  Máximo: 16.75

```

Figura 6: Estadísticos de la variable 4 3M_Rate_SecondaryMarket

Podemos observar que el comportamiento de esta variable es el mismo que las anteriores y que va a tener una cola a la derecha más alargada. Esto es comprobable por todas las razones expuestas en las secciones anteriores.

Variable 5: 3Y-CMaturityRate

Los estadísticos que nos arroja esta variable son:

```

3Y-CMaturityRate :
  Media: 6.829342
  Mediana: 5.77
  Desviación típica: 2.942284
  Moda: 5
  Kurtosis: 1.14164
  Asimetría: 1.199487
  Mínimo: 2.69
  Máximo: 16.76

```

Figura 7: Estadísticos de la variable 5 3Y_CMaturityRate

De igual forma los estadísticos de esta variable nos están arrojando la misma información que para el resto de variables con una cola derecha muy alargada. Podemos corroborar esta información por lo mismo que hemos dicho antes (mediana, media, máximo y mínimo) además de coeficiente de asimetría que al ser positivo nos indica que es asimétrica a la derecha.

Por otro lado, al igual que en los casos previos, la curtosis nos indica que la distribución es más puntiaguda o apuntada que una distribución normal.

Variable 6: 5Y-CMaturityRate

Los estadísticos que nos arroja esta variable son:

```

5Y-CMaturityRate :
  Media: 8.117378
  Mediana: 7.44
  Desviación típica: 2.88388
  Moda: 6.53
  Kurtosis: -0.08284863
  Asimetría: 0.8626431
  Mínimo: 4.09
  Máximo: 16.47

```

Figura 8: Estadísticos de la variable 6 5Y_CMaturityRate

En esta variable observamos también una cola derecha más alargada pero podemos observar que aquí no es una cola tan alargada como en los casos anteriores.

Si comprobamos la media y la mediana observamos que la mediana es menor que la media por lo que ya nos está apuntando a la asimetría pero si vemos la desviación típica y calculamos el intervalo en el que deberían estar el 95 % de los datos ([2,349618, 13, 885138]) podemos percibir que la cola izquierda es más corta (pues el mínimo es 4,09) y la cola derecha más alargada pues el máximo llega hasta 16,47.

Podemos decir que la asimetría derecha se manifiesta pero con menor intensidad como podemos percibir por el coeficiente de asimetría, pues aunque es positivo, es menor que en otros casos.

En cuanto a la curtosis podemos ver que es ligeramente negativa pero muy cercana al cero por lo que podemos decir que la distribución esta ligerísimamente achatada con respecto a una normal aunque visualmente probablemente no pudiéramos distinguirlo.

Variable 7: bankCredit

Los estadísticos que nos arroja esta variable son:

```
bankCredit :  
  Media: 8.359104  
  Mediana: 7.76  
  Desviación típica: 2.766248  
  Moda: 6.63  
  Kurtosis: -0.2687357  
  Asimetría: 0.8139356  
  Mínimo: 4.17  
  Máximo: 16.13
```

Figura 9: Estadísticos de la variable 7 bankCredit

En este caso podemos ver que la mediana es más pequeña que la media por lo que podemos pensar de nuevo en que esta variable tiene una cola derecha más alargada. Si calculamos el intervalo en el que deberían de estar el 95 % de los datos ([2,826608, 13,8916]) podemos apreciar esa tendencia a una cola derecha más pesada.

Aún así, si comprobamos el coeficiente de asimetría, podemos ver que se manifiesta la asimetría derecha pero de forma ligera como en la variable anterior pues el coeficiente no es tan grande como en otras variables que ya hemos analizado.

En cuanto a la curtosis tenemos una curtosis negativa lo que nos indica que la distribución está algo más achatada que su correspondiente distribución normal.

Variable 8: currency

Los estadísticos que nos arroja esta variable son:

```
currency :  
  Media: 2639.677  
  Mediana: 2616.1  
  Desviación típica: 1010.521  
  Moda: 1287.7  
  Kurtosis: -0.9085405  
  Asimetría: 0.3116162  
  Mínimo: 1130.9  
  Máximo: 4809.2
```

Figura 10: Estadísticos de la variable 8 currency

En este caso la mediana también es algo más pequeña que la media, pero si comprobamos el rango de valores de la variable ([1130,9, 4809,2,]) observamos que la diferencia no es tan significativa por lo que no podemos decir tan rápido que tenemos una cola derecha más alargada. Como vemos la desviación típica es de 1010,521 por lo que el intervalo que debe contener el 95 % de los datos es [618,635, 4660, 719].

Con esta información podemos apuntar que debe existir una ligera asimetría en la cola derecha

porque el mínimo es mayor que el mínimo que nos da el intervalo del 95 % de los datos y el máximo es algo mayor que el máximo del intervalo del 95 %.

Para contrastar esta información tenemos el coeficiente de asimetría que, al ser positivo, nos dice que la distribución presenta asimetría derecha pero podemos ver que es mucho más pequeño que en el resto de variables con lo que la asimetría no es tan pronunciada.

En cuanto a la curtosis podemos ver que es negativa por lo que la distribución es más achatada que una distribución normal.

Variable 9: demandDeposits

Los estadísticos que nos arroja esta variable son:

```
demandDeposits :  
  Media: 256.8477  
  Mediana: 224.4  
  Desviación típica: 114.5754  
  Moda: 120.1  
  Kurtosis: -0.9246782  
  Asimetría: 0.5228948  
  Mínimo: 105.6  
  Máximo: 533
```

Figura 11: Estadísticos de la variable 9 demandDeposits

Como podemos observar tenemos que la mediana es menor que la media. Volvemos a tener una situación que nos lleva a pensar que tenemos una cola derecha más alargada que la izquierda.

Si calculamos el intervalo en el que debe estar el 95 % de los datos ([27,6969, 485,9985]) podemos observar que el máximo es algo más grande que el máximo de dicho intervalo e igualmente ocurre con el mínimo con lo que tenemos un desplazamiento de los datos hacia la derecha.

El hecho viene refrendado por el coeficiente de asimetría que al ser positivo nos indica dicha asimetría derecha.

En cuanto a la curtosis al tener una curtosis negativa estamos ante una distribución más achatada que en el caso de una normal.

Variable 10: federalFunds

Los estadísticos que nos arroja esta variable son:


```
federalFunds :  
  Media: 308.1154  
  Mediana: 287.7  
  Desviación típica: 59.80509  
  Moda: 277.9  
  Kurtosis: -1.40515  
  Asimetría: 0.3380899  
  Mínimo: 225.8  
  Máximo: 412.1
```

Figura 12: Estadísticos de la variable 10 federalFunds

Podemos observar en esta variable el mismo comportamiento que venimos destacando del resto. Tenemos que la mediana es más pequeña que la media lo que nos indica que la cola derecha de la distribución debe ser algo más alargada.

El intervalo en el que el 95 % de los datos debe caer es [188,50522, 427,72558]. Podemos ver que el máximo de este intervalo es ligeramente más grande que el máximo y el mínimo es ligeramente más grande que el del intervalo del 95 % por lo que podemos decir que si existe una asimetría derecha esta no es muy pronunciada.

Podemos contrastar lo que estamos afirmando por el coeficiente de asimetría que es ligeramente positivo, por lo que se confirma lo que estamos razonando de que tenemos una ligera asimetría derecha.

En cuanto a la curtosis tenemos que es negativa y grande por lo que será muy achatada con respecto a una normal de mismos parámetros.

Variable 11: moneyStock

Los estadísticos que nos arroja esta variable son:

```
moneyStock :  
  Media: 7.549495  
  Mediana: 6.64  
  Desviación típica: 3.538662  
  Moda: 5.45  
  Kurtosis: 1.779054  
  Asimetría: 1.341457  
  Mínimo: 2.86  
  Máximo: 20.06
```

Figura 13: Estadísticos de la variable 11 moneyStock

Podemos repetir exactamente el mismo razonamiento que hemos hecho anteriormente para argumentar la asimetría derecha, pero en este caso al comprobar el coeficiente de asimetría podemos ver que es mucho más pronunciada.

En cuanto a la curtosis tenemos el comportamiento opuesto al de la variable anterior teniendo que la distribución es significativamente más puntiaguda o apuntada que su normal de mismos parámetros.

Variable 12: moneyStock

Los estadísticos que nos arroja esta variable son:

```
checkableDeposits :  
  Media: 813.3304  
  Mediana: 796  
  Desviación típica: 258.6885  
  Moda: 1145.9  
  Kurtosis: -1.401232  
  Asimetría: -0.1719616  
  Mínimo: 381.1  
  Máximo: 1154.1
```

Figura 14: Estadísticos de la variable 12 checkableDeposits

Este es el primer ejemplo en el que podemos ver como la cola más alargada no va a ser la cola derecha sino la izquierda. En este caso, al igual que los anteriores, podemos ver que la mediana es menor que la media (aunque no mucho proporcionalmente al rango de valores que se toman). Esto podría indicarnos que la cola derecha continúa siendo algo más alargada que la izquierda, pero en este caso podemos apreciar que la moda (el valor más frecuente) es significativamente más grande que la media, por lo que no podemos intuir un comportamiento a priori. Tenemos información contradictoria, una parte nos dice que debemos tener la cola derecha más alargada y la otra nos está diciendo que la izquierda es la que debería ser más alargada.

Para poder estudiar esta situación compleja recurrimos al coeficiente de asimetría. Como podemos ver en este caso es negativo pero ligeramente. Tal y como podíamos pensar la situación está razonablemente equilibrada aunque mostrando una ligera asimetría izquierda.

En cuanto a la kurtosis tenemos que es negativa y grande, por lo que estamos ante una distribución mucho más achatada que una distribución normal de mismos parámetros.

Variable 13: savingsDeposits

Los estadísticos que nos arroja esta variable son:

```
savingsDeposits :  
  Media: 1959.122  
  Mediana: 2023.9  
  Desviación típica: 720.5311  
  Moda: 2112.7  
  Kurtosis: -0.8147294  
  Asimetría: 0.3023447  
  Mínimo: 868.1  
  Máximo: 3550.3
```

Figura 15: Estadísticos de la variable 13 savingsDeposits

El caso de esta variable es singular también. Podemos ver que la mediana es mayor que la media, la moda también por lo que podríamos pensar que la cola más alargada es la izquierda.

Por contra si miramos el intervalo mínimo y máximo podemos ver que lo más probable es que se extienda más la cola derecha pues el máximo es más grande proporcionalmente a la media que el mínimo.

Para contrastar la información miramos el coeficiente de asimetría y observamos que, aunque es pequeño, es positivo. Esto nos indica que la distribución es asimétrica derecha.

Si analizamos la curtosis podemos ver que es negativa lo que nos indica que la distribución es más achatada que una normal de mismos parámetros.

Variable 14: tradeCurrencies

Los estadísticos que nos arroja esta variable son:

```
tradeCurrencies :  
  Media: 954.6694  
  Mediana: 947.9  
  Desviación típica: 372.2925  
  Moda: 343.9  
  Kurtosis: -0.2597352  
  Asimetría: -0.1331363  
  Mínimo: 175.6  
  Máximo: 1758.1
```

Figura 16: Estadísticos de la variable 14 tradeCurrencies

En cuanto a esta variable podemos ver que la media es mayor (ligeramente) que la mediana y la moda es mucho menor que la media. Por contra tenemos que el mínimo es más significativo que el máximo en cuanto al intervalo de valores se refiere. En ese caso a priori no podemos decir nada.

Si miramos el coeficiente de asimetría podemos contrastar esta información pues es muy cercano a cero y en este caso ligeramente negativo por lo que si podemos decir algo es que es ligeramente asimétrica la distribución a la izquierda.

La curtosis es ligeramente negativa por lo que podemos decir que es un poco más achatada que su normal asociada.

1.1.2. Estudio de la correlación de las variables

En esta sección vamos a estudiar la correlación entre variables que no son de salida y de dichas variables con la de salida para intentar ver cuales van a ser las más relevantes para nuestro estudio o si hay alguna variable que podamos quitar.

En primer lugar vamos a hacer un estudio de la correlación entre las variables. Nuestro objetivo va a ser obtener aquellas que tienen alta correlación con otras variables, es decir, obtener aquellas variables que son explicadas por otras pues estas las podremos quitar.

Por supuesto todo este estudio habrá que contrastarlo sobre los resultados de la regresión, pero podemos hacer hipótesis previas al ajuste de los modelos.

El código que voy a emplear para el estudio es el siguiente:

```
obtainCorrelated<-function(varIndex, dataset, threshold=0.9){
  combinations<-combn(varIndex,2)

  correlations<-vector("numeric", dim(combinations)[2])

  for(i in 1:dim(combinations)[2]){
    pair<-combinations[,i]
    correlations[i]<-cor(dataset[,pair[1]], dataset[,pair[2]], method = c("pearson", "kendall", "spearman"))
  }

  for(i in 1:length(correlations)){
    if(abs(correlations[i])>threshold){
      cat("La pareja de variables: ", combinations[i,1], ", ", combinations[i,2], " tiene una correlación: ", correlations[i], "\n")
    }
  }
}
```

Figura 17: Código para el estudio de la correlación entre variables.

Para poder eliminar una variable de forma que estemos seguros de que no quitamos información debemos exigir que la correlación entre variables sea alta. En este caso como se puede ver en el código solamente vamos a mostrar las parejas de variables que presenten una correlación alta, en concreto que en valor absoluto sea mayor a 0,9.

Vamos a ver las parejas de variables que cumplen esta condición.

```
La pareja de variables: 2 , 3  tiene una correlación: 0.9367248
La pareja de variables: 2 , 4  tiene una correlación: 0.9864352
La pareja de variables: 2 , 5  tiene una correlación: 0.9874777
La pareja de variables: 2 , 6  tiene una correlación: 0.9849371
La pareja de variables: 2 , 7  tiene una correlación: 0.9668119
La pareja de variables: 2 , 11 tiene una correlación: 0.9692879
La pareja de variables: 3 , 6  tiene una correlación: 0.9717959
La pareja de variables: 3 , 7  tiene una correlación: 0.9806363
La pareja de variables: 3 , 12 tiene una correlación: -0.9201255
La pareja de variables: 4 , 5  tiene una correlación: 0.9977408
La pareja de variables: 4 , 6  tiene una correlación: 0.9527171
La pareja de variables: 4 , 7  tiene una correlación: 0.928719
La pareja de variables: 4 , 11 tiene una correlación: 0.9848922
La pareja de variables: 5 , 6  tiene una correlación: 0.9536684
La pareja de variables: 5 , 7  tiene una correlación: 0.9295766
La pareja de variables: 5 , 11 tiene una correlación: 0.9861269
La pareja de variables: 6 , 7  tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 9  tiene una correlación: 0.99126
La pareja de variables: 8 , 12 tiene una correlación: 0.9340624
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 9 , 10 tiene una correlación: 0.9059379
La pareja de variables: 9 , 12 tiene una correlación: 0.9252762
La pareja de variables: 9 , 14 tiene una correlación: 0.9826211
La pareja de variables: 9 , 15 tiene una correlación: 0.9319984
La pareja de variables: 10 , 12 tiene una correlación: 0.937445
La pareja de variables: 12 , 13 tiene una correlación: 0.9610653
La pareja de variables: 12 , 14 tiene una correlación: 0.9189515
La pareja de variables: 12 , 15 tiene una correlación: 0.9123665
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578
```

Figura 18: Primer filtrado de la correlación entre variables

Podemos observar claramente que la variable 2 tiene una correlación muy alta con otras varia-

bles, por lo que podemos eliminarla al ser explicada por las variables 3,4,5,6,7 y 11.

```

La pareja de variables: 3 , 6 tiene una correlación: 0.9717959
La pareja de variables: 3 , 7 tiene una correlación: 0.9806363
La pareja de variables: 3 , 12 tiene una correlación: -0.9201255
La pareja de variables: 4 , 5 tiene una correlación: 0.9977408
La pareja de variables: 4 , 6 tiene una correlación: 0.9527171
La pareja de variables: 4 , 7 tiene una correlación: 0.928719
La pareja de variables: 4 , 11 tiene una correlación: 0.9848922
La pareja de variables: 5 , 6 tiene una correlación: 0.9536684
La pareja de variables: 5 , 7 tiene una correlación: 0.9295766
La pareja de variables: 5 , 11 tiene una correlación: 0.9861269
La pareja de variables: 6 , 7 tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 9 tiene una correlación: 0.99126
La pareja de variables: 8 , 12 tiene una correlación: 0.9340624
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 9 , 10 tiene una correlación: 0.9059379
La pareja de variables: 9 , 12 tiene una correlación: 0.9252762
La pareja de variables: 9 , 14 tiene una correlación: 0.9826211
La pareja de variables: 9 , 15 tiene una correlación: 0.9319984
La pareja de variables: 10 , 12 tiene una correlación: 0.937445
La pareja de variables: 12 , 13 tiene una correlación: 0.9610653
La pareja de variables: 12 , 14 tiene una correlación: 0.9189515
La pareja de variables: 12 , 15 tiene una correlación: 0.9123665
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 19: Correlación entre variables después de eliminar la segunda.

Podemos ver a su vez que la variable 9 tiene una correlación muy alta con las variables 10, 12, 14 y 15 por lo que podemos eliminarla.

```

La pareja de variables: 3 , 6 tiene una correlación: 0.9717959
La pareja de variables: 3 , 7 tiene una correlación: 0.9806363
La pareja de variables: 3 , 12 tiene una correlación: -0.9201255
La pareja de variables: 4 , 5 tiene una correlación: 0.9977408
La pareja de variables: 4 , 6 tiene una correlación: 0.9527171
La pareja de variables: 4 , 7 tiene una correlación: 0.928719
La pareja de variables: 4 , 11 tiene una correlación: 0.9848922
La pareja de variables: 5 , 6 tiene una correlación: 0.9536684
La pareja de variables: 5 , 7 tiene una correlación: 0.9295766
La pareja de variables: 5 , 11 tiene una correlación: 0.9861269
La pareja de variables: 6 , 7 tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 12 tiene una correlación: 0.9340624
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 10 , 12 tiene una correlación: 0.937445
La pareja de variables: 12 , 13 tiene una correlación: 0.9610653
La pareja de variables: 12 , 14 tiene una correlación: 0.9189515
La pareja de variables: 12 , 15 tiene una correlación: 0.9123665
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 20: Correlación entre variables después de eliminar la segunda y la novena.

La variable 4 podemos ver que tiene alta correlación con las variables 5,6,7 y 11 por lo que podemos quitarla también.

```

La pareja de variables: 3 , 6 tiene una correlación: 0.9717959
La pareja de variables: 3 , 7 tiene una correlación: 0.9806363
La pareja de variables: 3 , 12 tiene una correlación: -0.9201255
La pareja de variables: 5 , 6 tiene una correlación: 0.9536684
La pareja de variables: 5 , 7 tiene una correlación: 0.9295766
La pareja de variables: 5 , 11 tiene una correlación: 0.9861269
La pareja de variables: 6 , 7 tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 12 tiene una correlación: 0.9340624
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 10 , 12 tiene una correlación: 0.937445
La pareja de variables: 12 , 13 tiene una correlación: 0.9610653
La pareja de variables: 12 , 14 tiene una correlación: 0.9189515
La pareja de variables: 12 , 15 tiene una correlación: 0.9123665
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 21: Correlación entre variables después de eliminar la segunda, la novena y la cuarta.

La tercera variable sigue manteniendo una alta correlación con las variables 6,7 y 12, por lo que podemos eliminarla.

```

La pareja de variables: 5 , 6 tiene una correlación: 0.9536684
La pareja de variables: 5 , 7 tiene una correlación: 0.9295766
La pareja de variables: 5 , 11 tiene una correlación: 0.9861269
La pareja de variables: 6 , 7 tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 12 tiene una correlación: 0.9340624
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 10 , 12 tiene una correlación: 0.937445
La pareja de variables: 12 , 13 tiene una correlación: 0.9610653
La pareja de variables: 12 , 14 tiene una correlación: 0.9189515
La pareja de variables: 12 , 15 tiene una correlación: 0.9123665
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 22: Correlación entre variables después de eliminar la segunda, la novena, la cuarta y la tercera.

Podemos ver que la variable 12 y 5 tienen una alta correlación con otras tres que además no se comparten por lo que podemos eliminar también ambas variables.

```

La pareja de variables: 6 , 7 tiene una correlación: 0.9955387
La pareja de variables: 6 , 11 tiene una correlación: 0.9300853
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 8 , 14 tiene una correlación: 0.9977185
La pareja de variables: 8 , 15 tiene una correlación: 0.9531331
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 23: Correlación entre variables después de eliminar la segunda, la novena, la cuarta, la tercera, la duodécima y la quinta.

Podemos observar que la variable 6 y 8 tienen alta correlación con otras dos variables no compartidas, por lo que podemos eliminar ambas.

```

La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 24: Correlación entre variables después de eliminar la segunda, la novena, la cuarta, la tercera, la duodécima, la quinta, la sexta y la octava.

Finalmente nos hemos quedado con dos parejas únicamente por lo que ya no está tan claro la eliminación de dichas variables. Vamos a parar por tanto el proceso de eliminación en este punto.

Veamos la correlación entre variables que nos queda con las que hemos decidido mantener.

```

La pareja de variables: 1 , 7 tiene una correlación: 0.6250722
La pareja de variables: 1 , 10 tiene una correlación: -0.6408934
La pareja de variables: 1 , 11 tiene una correlación: 0.4560397
La pareja de variables: 1 , 13 tiene una correlación: -0.7161397
La pareja de variables: 1 , 14 tiene una correlación: -0.5672793
La pareja de variables: 1 , 15 tiene una correlación: -0.4763829
La pareja de variables: 7 , 10 tiene una correlación: -0.8168239
La pareja de variables: 7 , 11 tiene una correlación: 0.904041
La pareja de variables: 7 , 13 tiene una correlación: -0.862284
La pareja de variables: 7 , 14 tiene una correlación: -0.8273161
La pareja de variables: 7 , 15 tiene una correlación: -0.8651072
La pareja de variables: 10 , 11 tiene una correlación: -0.7072023
La pareja de variables: 10 , 13 tiene una correlación: 0.8731687
La pareja de variables: 10 , 14 tiene una correlación: 0.8663811
La pareja de variables: 10 , 15 tiene una correlación: 0.8259606
La pareja de variables: 11 , 13 tiene una correlación: -0.8347681
La pareja de variables: 11 , 14 tiene una correlación: -0.7063516
La pareja de variables: 11 , 15 tiene una correlación: -0.8060014
La pareja de variables: 13 , 14 tiene una correlación: 0.7864516
La pareja de variables: 13 , 15 tiene una correlación: 0.8127696
La pareja de variables: 14 , 15 tiene una correlación: 0.9502578

```

Figura 25: Correlación entre variables después de hacer la limpieza de variables.

Finalmente por tanto nos hemos quedado con las variables 1,7,10,11,13,14 y 15.

Sobre estas variables merece la pena estudiar su correlación con la variable de salida para ver que seguimos teniendo variables que explican el comportamiento de la salida.

```

La correlación de la variable 1 con la salida es de: 0.4507399
La correlación de la variable 7 con la salida es de: 0.9106219
La correlación de la variable 10 con la salida es de: -0.7021663
La correlación de la variable 11 con la salida es de: 0.9946502
La correlación de la variable 13 con la salida es de: -0.8280948
La correlación de la variable 14 con la salida es de: -0.693953
La correlación de la variable 15 con la salida es de: -0.7931364

```

Figura 26: Correlación de las variables no eliminadas con la variable de salida.

Podemos ver que las variables 7, 11, 13 y 15 tienen una correlación en valor absoluto mayor a 0,75 con lo que estas variables nos van a resultar de mucho interés a la hora de realizar un modelo de regresión lineal.

Todas estas hipótesis serán contrastadas en la sección de regresión al ajustar los modelos.

1.1.3. Valores perdidos

En cuanto a los valores perdidos vamos a comprobar en primer lugar si tenemos valores NA.

```
> which(is.na(dataset_regresion))
integer(0)
```

Figura 27: Código y resultados para comprobar si tenemos valores perdidos.

Como podemos comprobar no tenemos ningún valor perdido.

1.1.4. Outliers

Vamos a comprobar si tenemos outliers en nuestro conjunto de datos.

En primer lugar vamos a ver un pairplot de las variables para ver si podemos distinguir algo visualmente.

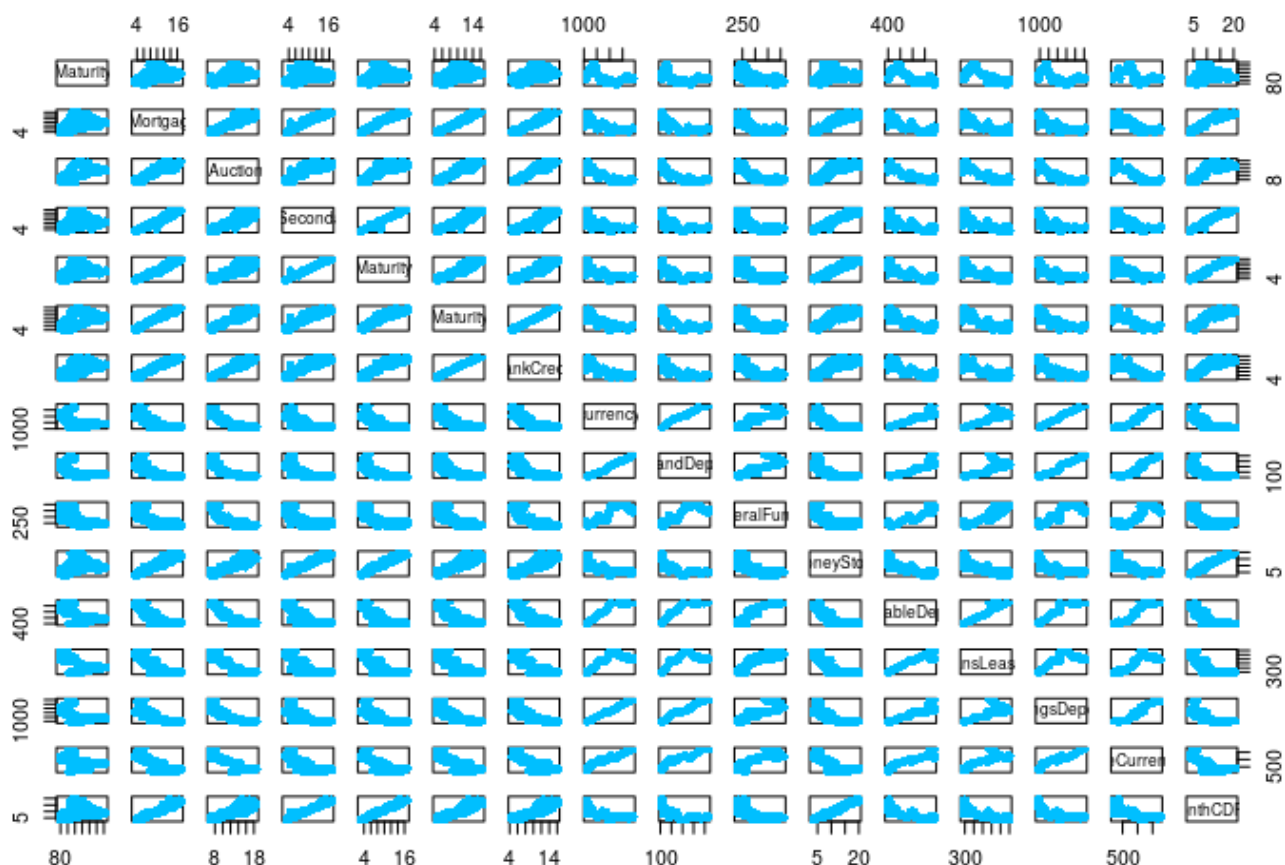


Figura 28: Scatterplot de todas las variables dos a dos

Visualmente con tantas variables no podemos destacar ninguna anomalía. Vamos a centrarnos solamente en las variables que hemos decidido quedarnos del estudio previo.

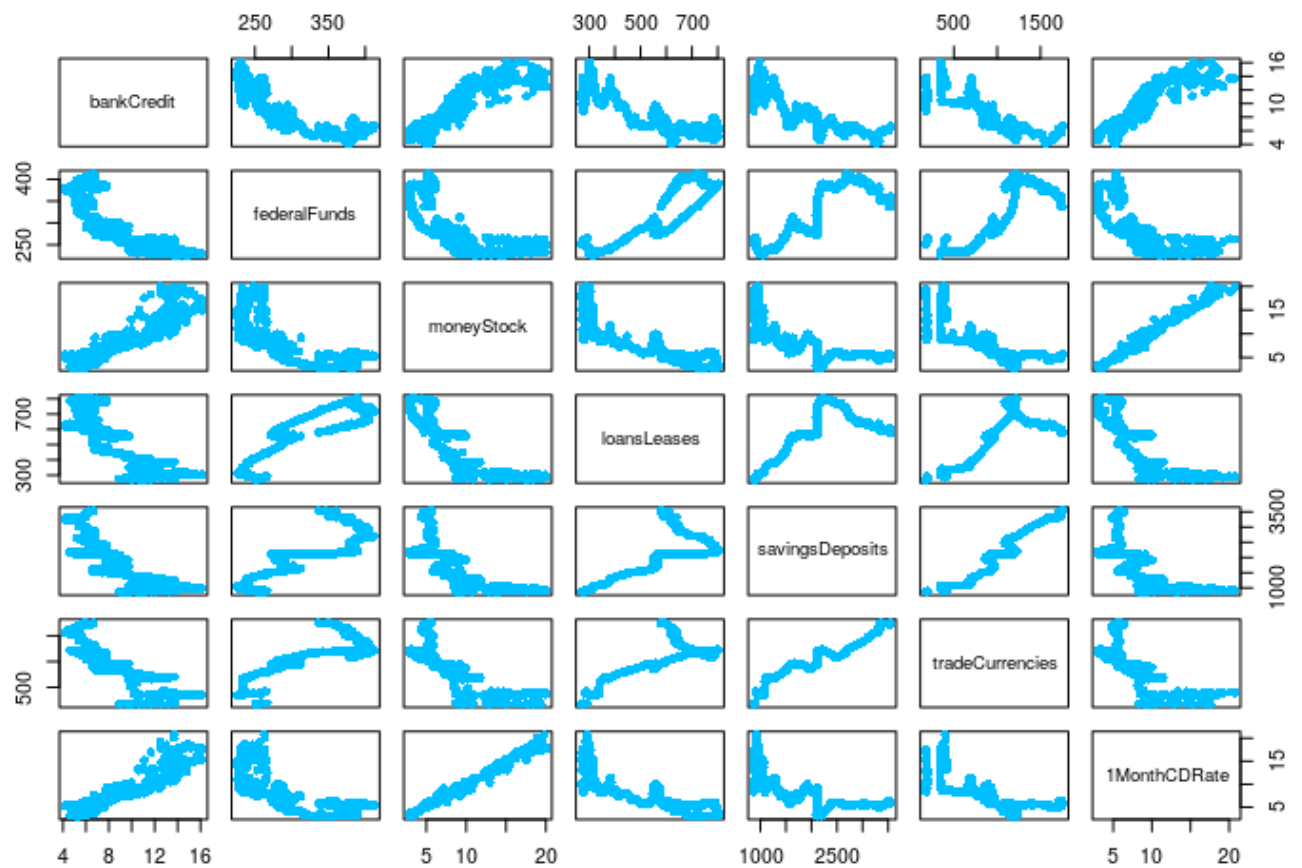


Figura 29: Scatterplot de las variables seleccionadas

Como podemos ver no podemos destacar ningún outlier en el conjunto de datos de forma visual.

De paso podemos ver que hemos hecho una selección de variables adecuada pues podemos observar una clara relación lineal con la salida.

Vamos a ver ahora un boxplot de todas las variables.

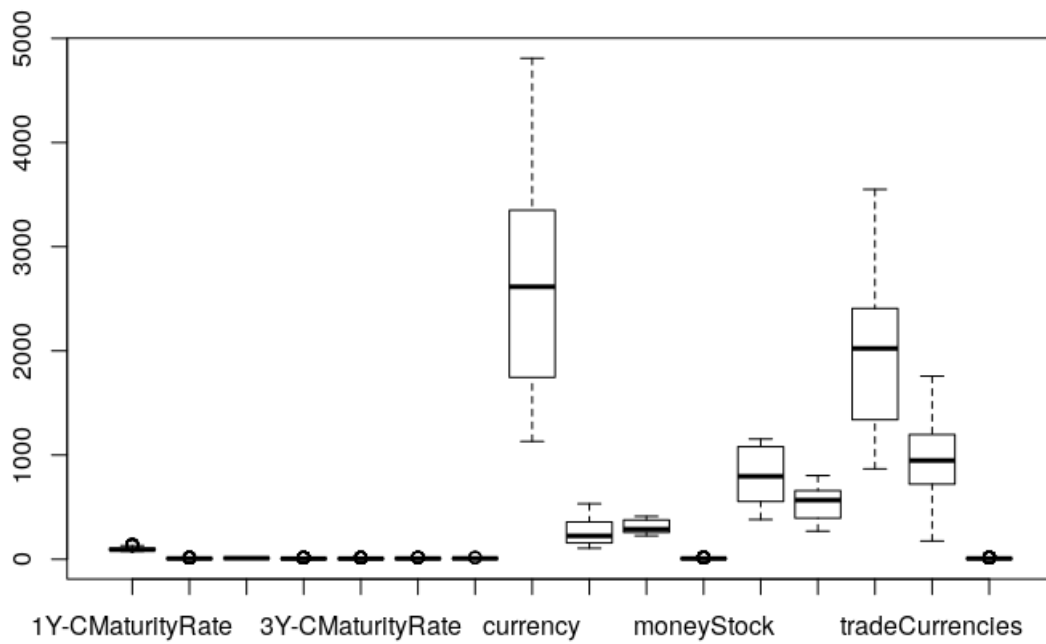


Figura 30: Boxplot de las todas las variables

Podemos ver que por la diferencia de escalas no todas las variables son visibles. En las que podemos observar de forma adecuada podemos ver que no hay outliers que nos destaquen fuera del rango intercuartil. Por tanto podemos eliminarlas del boxplot para reducir la escala y poder ver el resto de variables.

Para el siguiente boxplot vamos a quitar las variables 8, 9, 10, 12, 13, 14 y 15.

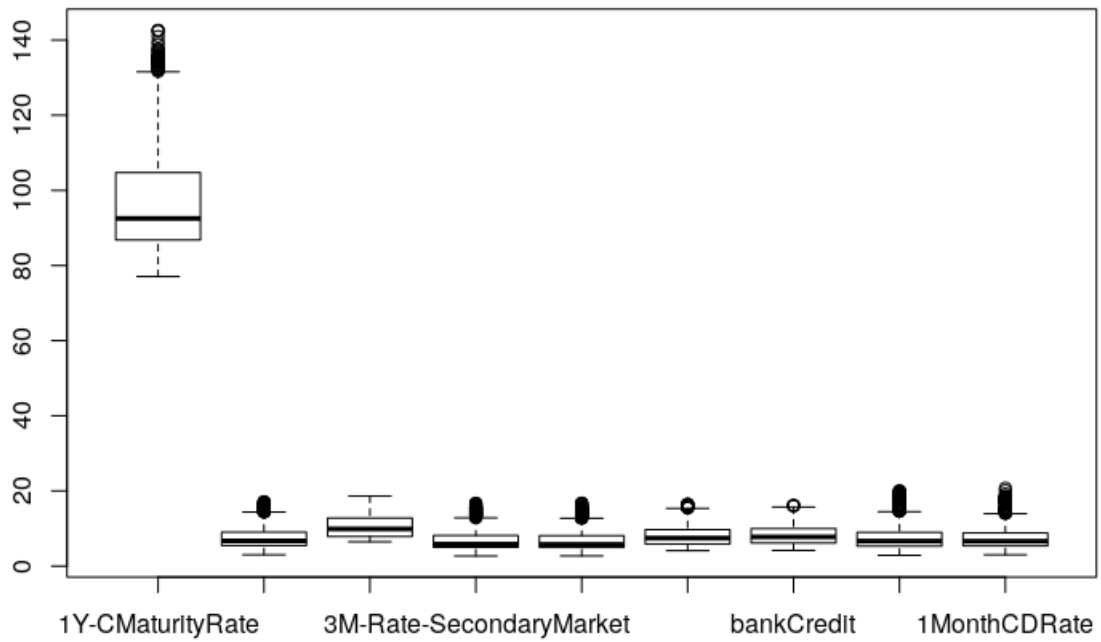


Figura 31: Boxplot quitando las variables 8, 9, 10, 12, 13, 14 y 15

Podemos observar que la primera variable tiene una escala mayor que el resto, por lo que para poder continuar la tendremos que quitar. Si observamos sus valores podemos ver que tenemos algunas anomalías por encima. Al tener tantos valores por encima no podemos decir de forma tan clara que estos valores son anómalos pues podría ser un comportamiento esperado de la variable y por tanto a priori no debemos eliminar dichos valores.

Eliminamos además la variable 1 para continuar con el estudio.

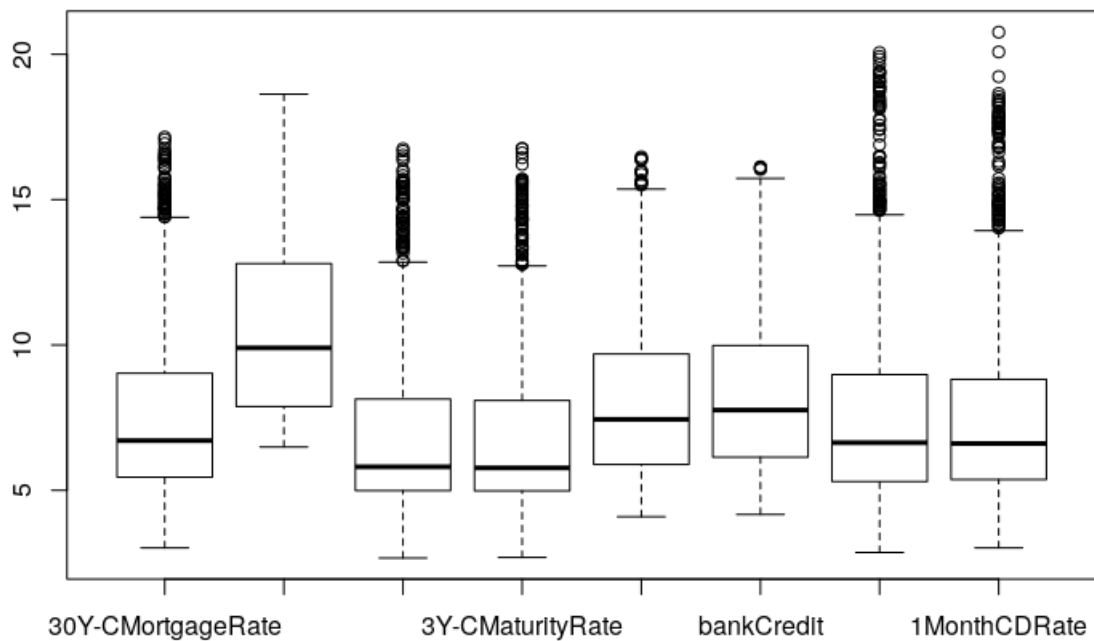


Figura 32: Boxplot quitando las variables 1, 8, 9, 10, 12, 13, 14 y 15

Podemos observar que en este último boxplot tenemos muchas más anomalías, de hecho, todas las variables poseen valores fuera del rango intercuartil menos para la segunda variable.

Podemos observar que la concentración de valores fuera de rango es muy grande por lo que no debemos eliminar dichos valores. Además este estudio es de todas las variables y no de las que hemos seleccionado para quedarnos. Veamos su boxplot.

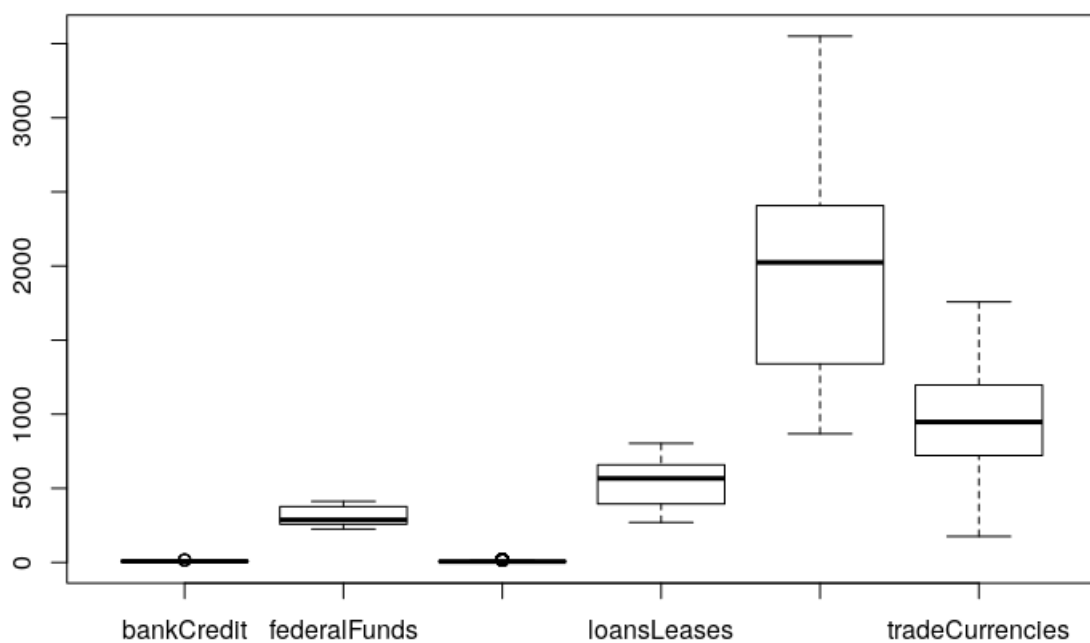


Figura 33: Boxplot manteniendo las variables seleccionadas.

Podemos ver que en las variables seleccionadas tenemos dos que poseen anomalías, vamos a estudiarlas en un boxplot aislado.

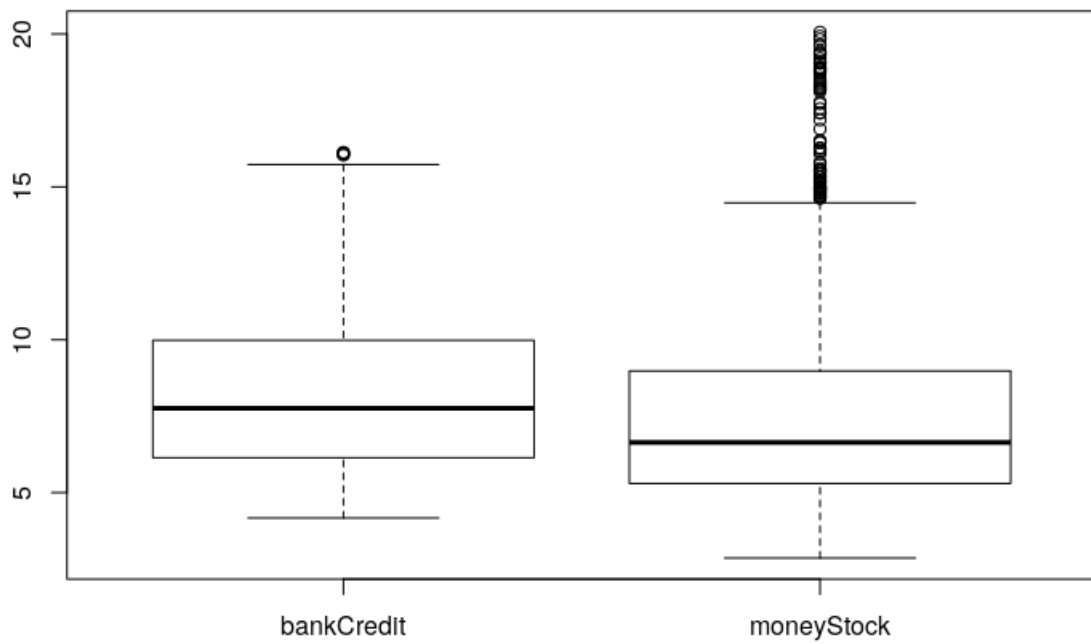


Figura 34: Boxplot de las variables que parecen tener anomalías.

Como podemos observar tenemos que la variable moneyStock tiene valores anómalos pero muy densos por lo que no debemos quitarlos. En el caso de bankCredit vamos a estudiar su scatterplot.

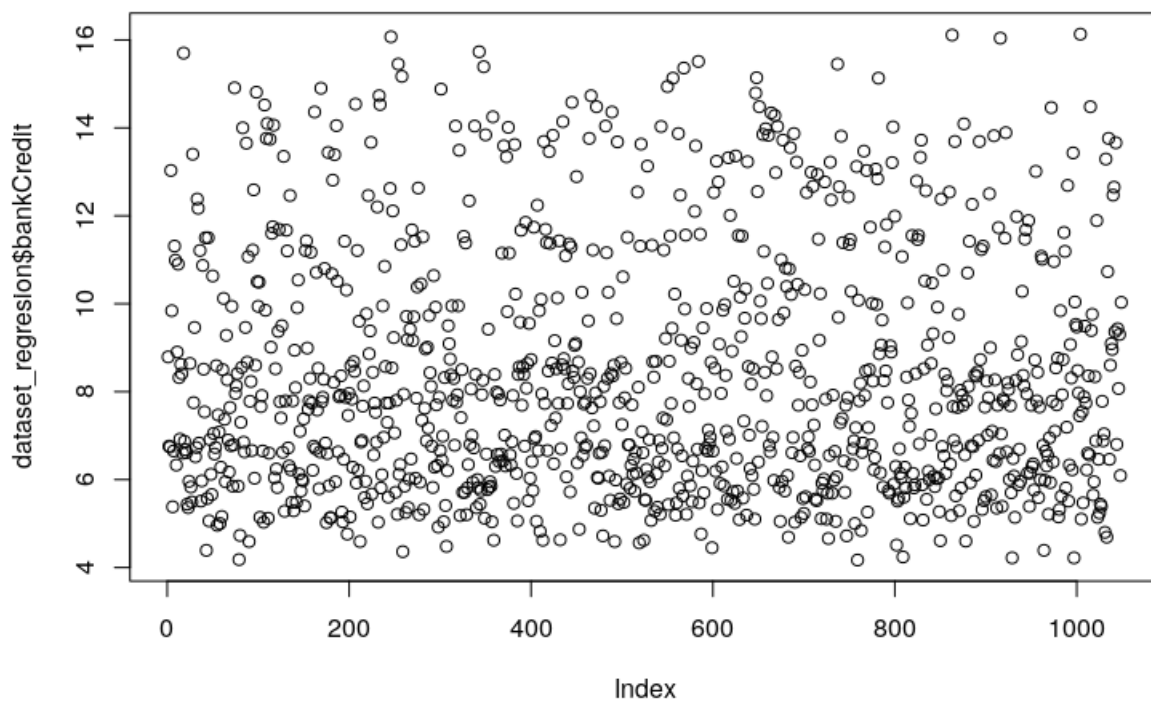


Figura 35: ScatterPlot de la variable bankCredit

Podemos ver que las anomalías que nos encontramos no son tal pues es un conjunto distribuido de forma casi uniforme.

1.1.5. Distribución de las variables

Vamos a ver en primer lugar unos histogramas de las variables.

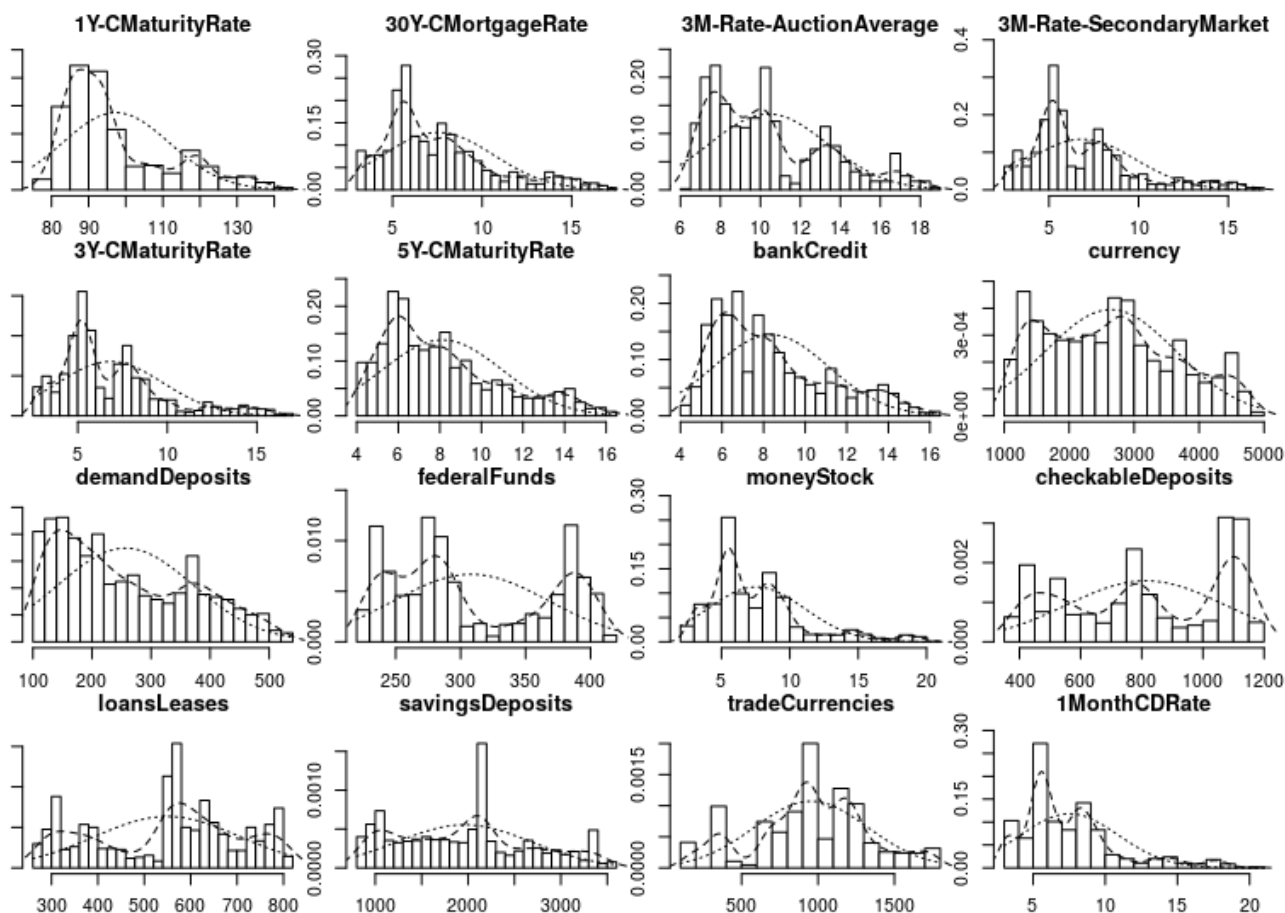


Figura 36: Histograma de todas las variables

Como podemos ver hemos acertado en el estudio previo de los estadísticos y podemos corroborar que la mayoría de variables tienen la cola derecha de su distribución más alargada.

Podemos ver que las distribuciones no se parecen para nada a una normal de forma visual, aunque para poder estar seguros vamos a hacer un test de normalidad.

Para esto vamos a utilizar el test de Wilcoxon. Veamos los resultados.

```
Test de normalidad para la variable 1
P-valor: 3.404098e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 37: Test de normalidad para la variable 1.

```
Test de normalidad para la variable 2
P-valor: 3.401871e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 38: Test de normalidad para la variable 2.

```
Test de normalidad para la variable 3
P-valor: 3.401848e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 39: Test de normalidad para la variable 3.

Test de normalidad para la variable 4
P-valor: 3.400497e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 40: Test de normalidad para la variable 4.

Test de normalidad para la variable 5
P-valor: 3.400346e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 41: Test de normalidad para la variable 5.

Test de normalidad para la variable 6
P-valor: 3.402266e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 42: Test de normalidad para la variable 6.

Test de normalidad para la variable 7
P-valor: 3.402179e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 43: Test de normalidad para la variable 7.

Test de normalidad para la variable 8
P-valor: 3.404102e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 44: Test de normalidad para la variable 8.

Test de normalidad para la variable 9
P-valor: 3.403861e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 45: Test de normalidad para la variable 9.

Test de normalidad para la variable 10
P-valor: 3.402683e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 46: Test de normalidad para la variable 10.

Test de normalidad para la variable 11
P-valor: 3.40101e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.

Figura 47: Test de normalidad para la variable 11.

```
Test de normalidad para la variable 12
P-valor: 3.403832e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 48: Test de normalidad para la variable 12.

```
Test de normalidad para la variable 13
P-valor: 3.403666e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 49: Test de normalidad para la variable 13.

```
Test de normalidad para la variable 14
P-valor: 3.404037e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 50: Test de normalidad para la variable 14.

```
Test de normalidad para la variable 15
P-valor: 3.403915e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 51: Test de normalidad para la variable 15.

```
Test de normalidad para la variable 16
P-valor: 3.395458e-173
Como es menor a 0.05 se rechaza la hipótesis nula, no sigue una normal.
```

Figura 52: Test de normalidad para la variable 16.

Como podemos ver por los resultados del test podemos rechazar en todos los casos la hipótesis nula por lo que podemos decir que ninguna de las variables sigue una distribución normal, tal y como hemos podido ver de forma gráfica.

1.2. Conjunto de Clasificación

El conjunto con el que vamos a atacar el problema de clasificación es el conjunto de datos “heart”. Vamos a analizar este conjunto de datos antes de abordar el problema.

El problema dispone de 14 variables, siendo la última la clase a la que pertenece cada instancia.

Vamos a ver los tipos de las variables.

```
Age : integer
Sex : integer
ChestPainType : integer
RestBloodPressure : integer
SerumCholestoral : integer
FastingBloodSugar : integer
ResElectrocardiographic : integer
MaxHeartRate : integer
ExerciseInduced : integer
Oldpeak : numeric
Slope : integer
MajorVessels : integer
Thal : integer
Class : integer
```

Figura 53: Tipos de las variables.

Podemos observar que todas las variables son de tipo integer, es decir, de tipo entero y que hay una variable de tipo numeric, o lo que es lo mismo, de tipo real.

Vamos a realizar el estudio del conjunto completo dividiéndolo en dos partes, la variable que nos indica la clase y el resto de variables.

1.2.1. Estudio de los estadísticos

Salida

Variable 1: 1Y-CMaturityRate

Variable 2: 30Y-CMortgageRate

Variable 3: 3M-Rate-AuctionAverage

Variable 4: 3M-Rate-SecondaryMarket

Variable 5: 3Y-CMaturityRate

Variable 6: 5Y-CMaturityRate

Variable 7: bankCredit

Variable 8: currency

Variable 9: demandDeposits

Variable 10: federalFunds

Variable 11: moneyStock

Variable 12: moneyStock

Variable 13: savingsDeposits

Variable 14: tradeCurrencies

1.2.2. Estudio de la correlación de las variables

1.2.3. Valores perdidos

1.2.4. Outliers

1.2.5. Distribución de las variables