Tema2_Ejercicio_01

Fran Camacho

2024-12-16

Tema2 - Ejercicio 1

Cargar datos del paquete ISLR

```
install.packages("ISLR")
```

```
## The following package(s) will be installed:
## - ISLR [1.4]
## These packages will be installed into "~/git/masterML/renv/library/linux-debian-bookworm/R-4.4/x86_64
##
## # Installing packages --------------------------------------------------------
## - Installing ISLR ...                          OK [linked from cache]
## Successfully installed 1 package in 9.5 milliseconds.
```

```
library("ISLR")
```

Elegimos el dataset Credit para examinar su contenido (estructura y resumen):

```
#Structure
str(Credit)
```

```
## 'data.frame':    400 obs. of  12 variables:
##  $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Income   : num  14.9 106 104.6 148.9 55.9 ...
##  $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
##  $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
##  $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
##  $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
##  $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
##  $ Gender   : Factor w/ 2 levels " Male","Female": 1 2 1 2 1 1 2 1 2 2 ...
##  $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
##  $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
##  $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
##  $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

Este data frame contiene 400 observaciones, con 12 variables cada una de ellas (en la documentación se mencionan 10000 y 4 variables ..) Eliminamos el ID, mostramos algunos registros para ver su aspecto, y finalmente el resumen estadístico con "summary".

```r
credit <- Credit[-1]

head(credit,10)
```

```
##     Income Limit Rating Cards Age Education Gender Student Married
## 1   14.891  3606    283     2  34        11   Male      No     Yes
## 2  106.025  6645    483     3  82        15 Female     Yes     Yes
## 3  104.593  7075    514     4  71        11   Male      No      No
## 4  148.924  9504    681     3  36        11 Female      No      No
## 5   55.882  4897    357     2  68        16   Male      No     Yes
## 6   80.180  8047    569     4  77        10   Male      No      No
## 7   20.996  3388    259     2  37        12 Female      No      No
## 8   71.408  7114    512     2  87         9   Male      No      No
## 9   15.125  3300    266     5  66        13 Female      No      No
## 10  71.061  6819    491     3  41        19 Female     Yes     Yes
##           Ethnicity Balance
## 1         Caucasian     333
## 2             Asian     903
## 3             Asian     580
## 4             Asian     964
## 5         Caucasian     331
## 6         Caucasian    1151
## 7  African American     203
## 8             Asian     872
## 9         Caucasian     279
## 10 African American    1350
```

```r
tail(credit,10)
```

```
##      Income Limit Rating Cards Age Education Gender Student Married
## 391 135.118 10578    747     3  81        15 Female      No     Yes
## 392  73.327  6555    472     2  43        15 Female      No      No
## 393  25.974  2308    196     2  24        10   Male      No      No
## 394  17.316  1335    138     2  65        13   Male      No      No
## 395  49.794  5758    410     4  40         8   Male      No      No
## 396  12.096  4100    307     3  32        13   Male      No     Yes
## 397  13.364  3838    296     5  65        17   Male      No      No
## 398  57.872  4171    321     5  67        12 Female      No     Yes
## 399  37.728  2525    192     1  44        13   Male      No     Yes
## 400  18.701  5524    415     5  64         7 Female      No      No
##            Ethnicity Balance
## 391            Asian    1393
## 392        Caucasian     721
## 393            Asian       0
## 394 African American       0
## 395        Caucasian     734
## 396        Caucasian     560
## 397 African American     480
## 398        Caucasian     138
## 399        Caucasian       0
## 400            Asian     966
```

```r
#Summary
summary(credit)
```

```
##      Income           Limit           Rating          Cards
##  Min.   : 10.35   Min.   :  855   Min.   : 93.0   Min.   :1.000
##  1st Qu.: 21.01   1st Qu.: 3088   1st Qu.:247.2   1st Qu.:2.000
##  Median : 33.12   Median : 4622   Median :344.0   Median :3.000
##  Mean   : 45.22   Mean   : 4736   Mean   :354.9   Mean   :2.958
##  3rd Qu.: 57.47   3rd Qu.: 5873   3rd Qu.:437.2   3rd Qu.:4.000
##  Max.   :186.63   Max.   :13913   Max.   :982.0   Max.   :9.000
##       Age          Education        Gender    Student   Married
##  Min.   :23.00   Min.   : 5.00    Male :193   No :360   No :155
##  1st Qu.:41.75   1st Qu.:11.00   Female:207   Yes: 40   Yes:245
##  Median :56.00   Median :14.00
##  Mean   :55.67   Mean   :13.45
##  3rd Qu.:70.00   3rd Qu.:16.00
##  Max.   :98.00   Max.   :20.00
##             Ethnicity      Balance
##  African American: 99   Min.   :   0.00
##  Asian           :102   1st Qu.:  68.75
##  Caucasian       :199   Median : 459.50
##                         Mean   : 520.01
##                         3rd Qu.: 863.00
##                         Max.   :1999.00
```

Todos los registros contienen todos los valores (no hay NAs, ni nulos). Aunque algunos balances están a 0 (según la documentación es un valor promedio). También me llama mucho la atención la columna "Etnia" ...

A la función "summary" se le puede pasar un vector con las variables a analizar, si se quiere poner el foco en algunas de ellas. Elijo tres que considero bastante relevantes en este dataset:

```r
#Summary of Income, Rating and Balance
summary(credit[c("Income", "Rating", "Balance")])
```

```
##      Income           Rating         Balance
##  Min.   : 10.35   Min.   : 93.0   Min.   :   0.00
##  1st Qu.: 21.01   1st Qu.:247.2   1st Qu.:  68.75
##  Median : 33.12   Median :344.0   Median : 459.50
##  Mean   : 45.22   Mean   :354.9   Mean   : 520.01
##  3rd Qu.: 57.47   3rd Qu.:437.2   3rd Qu.: 863.00
##  Max.   :186.63   Max.   :982.0   Max.   :1999.00
```

No me parece que la media y la mediana estén muy separadas para estas variables (la media es afectada mucho más por valores extremos, así que entiendo que no hay muchos valores de esa clase).

Cuantiles y rango intercuartil (IQR: interquartile range)

La diferencia entre el cuartil 1 (Q1) y el cuartil 3 (Q3) es conocido como "rango intercuartil", y es de interés porque representa una medida simple de la dispersión de los datos.

```r
#IQR Income
IQR(credit$Income)
```

```
## [1] 36.4635
```

```r
#IQR Rating
IQR(credit$Rating)
```

```
## [1] 190
```

```r
#IQR Balance
IQR(credit$Balance)
```

```
## [1] 794.25
```

Aquí se puede ver por ejemplo que el 50% de los ingresos están en el rango que va desde los 21.000$ del Q1 a los 57000 del Q3 (IQR=36.464$)

La función que nos devuelve los cuantiles es "quantile". Sin parámetros nos devuelve los mismos valores que la función "summary". Si por ejemplo queremos averiguar los ingresos del 10% de la población con menores y mayores ingresos, debemos llamar a la función quantile con estos parámetros:

```r
quantile(credit$Income, probs = c(0.1, 0.9))
```

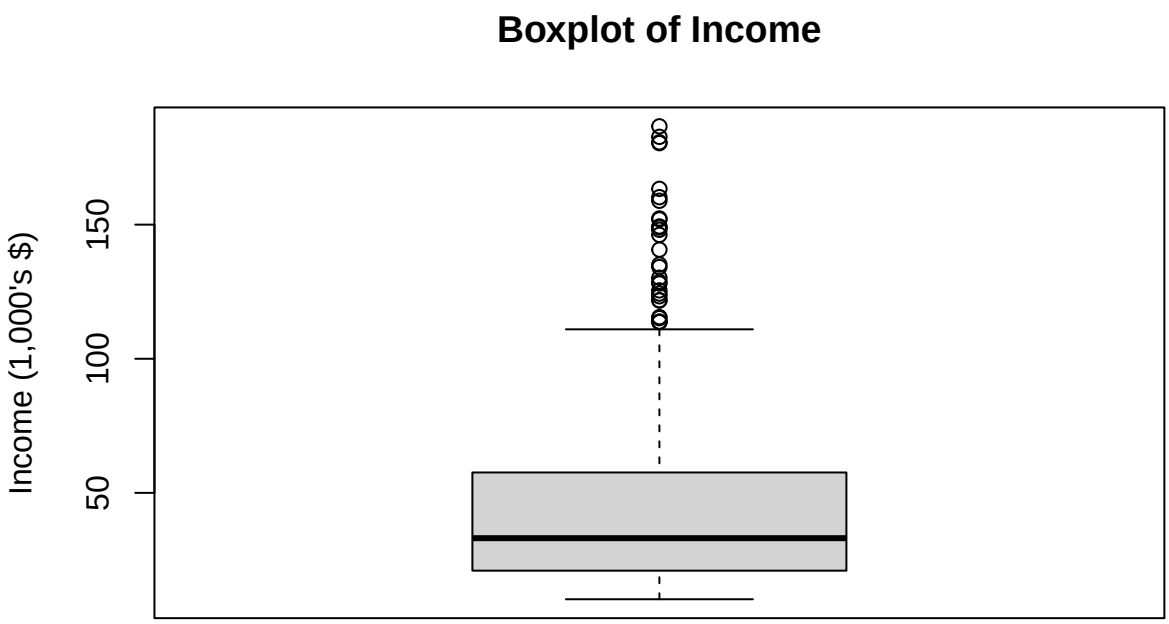
```
##     10%     90%
## 14.5834 92.4513
```

El 10% de la población tiene ingresos inferiores a 14.5583$, y el 10% con más ingresos gana más de 92.451$.

Visualización de variables numéricas con diagramas de cajas ("boxplots")

Los diagramas de caja permiten una visualización de los datos que puede ayudar a diagnosticar problemas que afecten al rendimiento de los algoritmos. Gracias a ellos se puede ver a simple vistala la mediana, el rango de los valores, los valores atípicos ... (también los 5 cuantiles obtenidos al llamar a la función "summary").

```r
#Boxplot for variable income
boxplot(credit$Income, main = "Boxplot of Income", ylab = "Income (1,000's $)")
```



Al analizar la variable ingresos con la ayuda de su diagrama de cajas, se puede ver que la mediana está más cerca del Q1 que del Q3, y que sí hay bastantes valores atípicos.

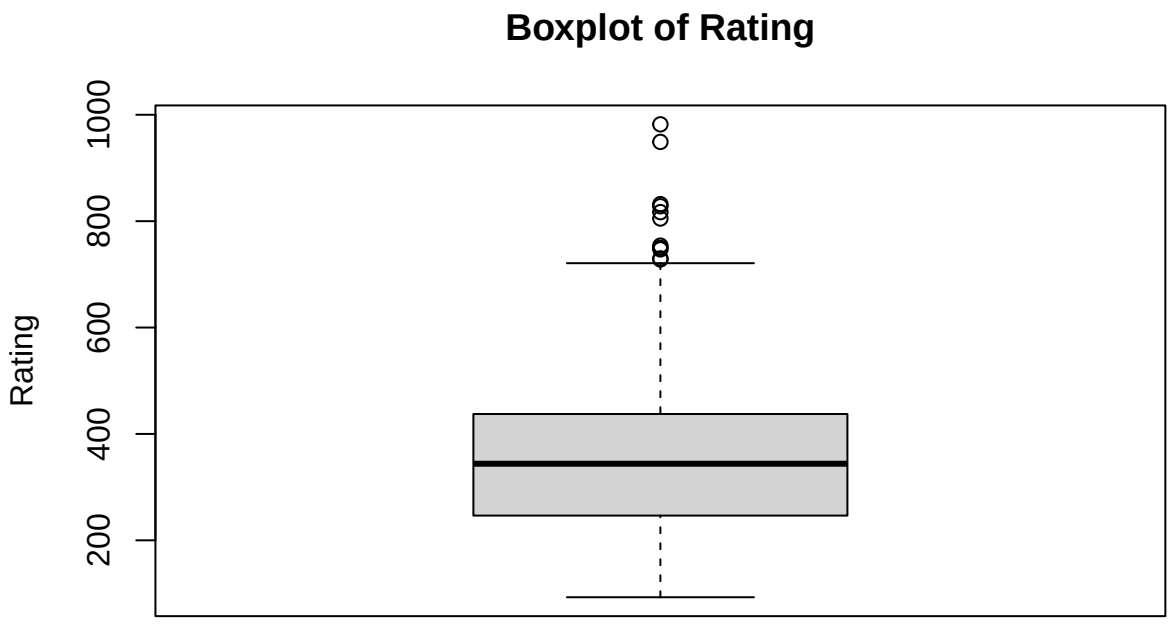
Estos valores atípicos (superiores) se suelen calcular de la siguiente manera:

Q3 + 1.5 * IQR 57470 + 1.5 * 36464 = 112.000

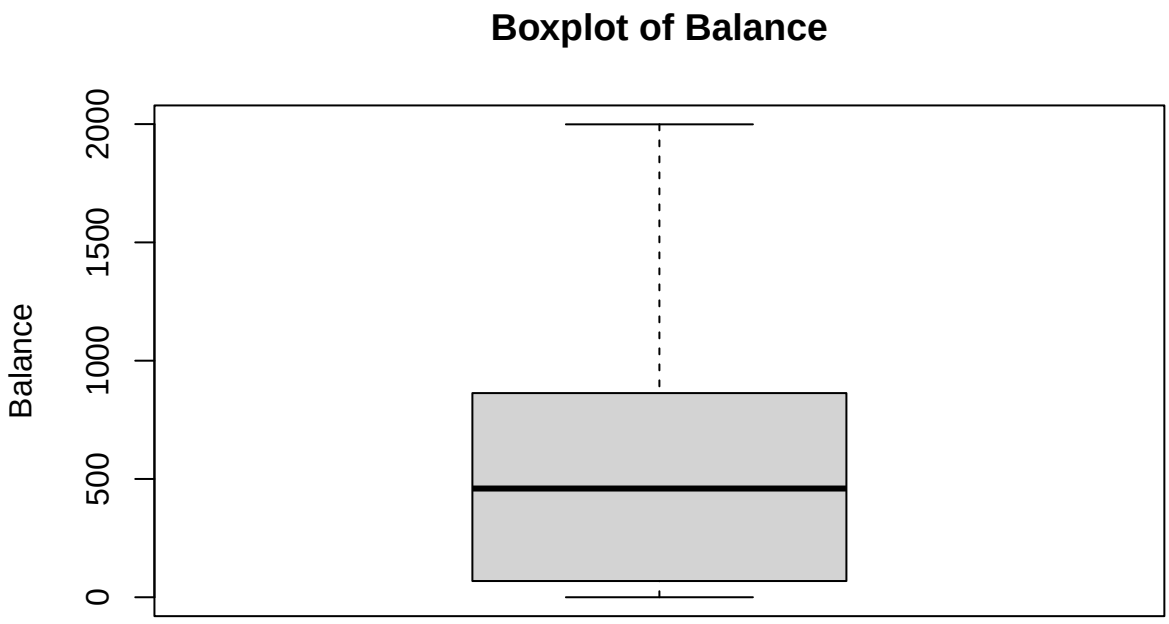
Todos los salarios por encima de la línea que marca los 112.000$, son considerados por tanto valores atípicos ("outliers")

Si analizamos las otras dos variables "rating" y balance medio en la cuenta:

```
#Boxplot for variables rating and balance
boxplot(credit$Rating, main = "Boxplot of Rating", ylab = "Rating")
```



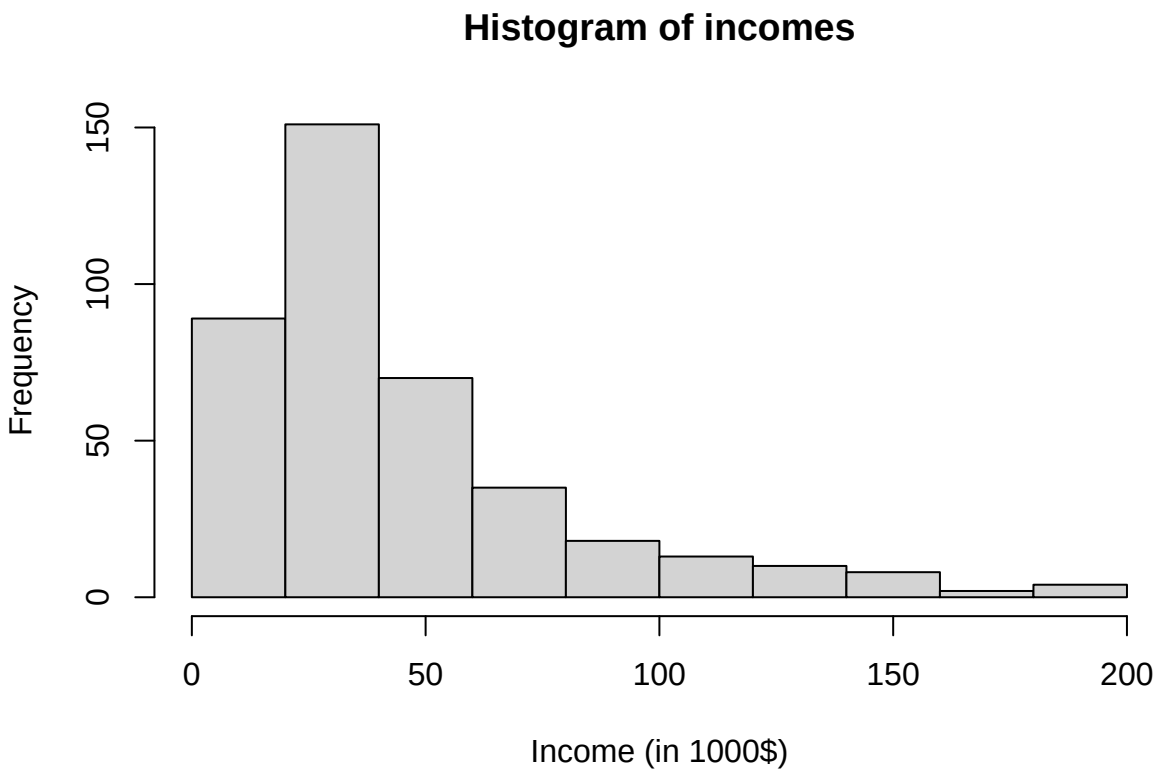

```r
boxplot(credit$Balance, main = "Boxplot of Balance", ylab = "Balance")
```



Vemos que el diagrama de cajas del rating es bastante parecido al de los ingresos, mientras que el del balance no muestra valores extremos (debido probablemente a que es un valor medio, y a que las tarjetas de crédito suelen tener unos límites estándares).

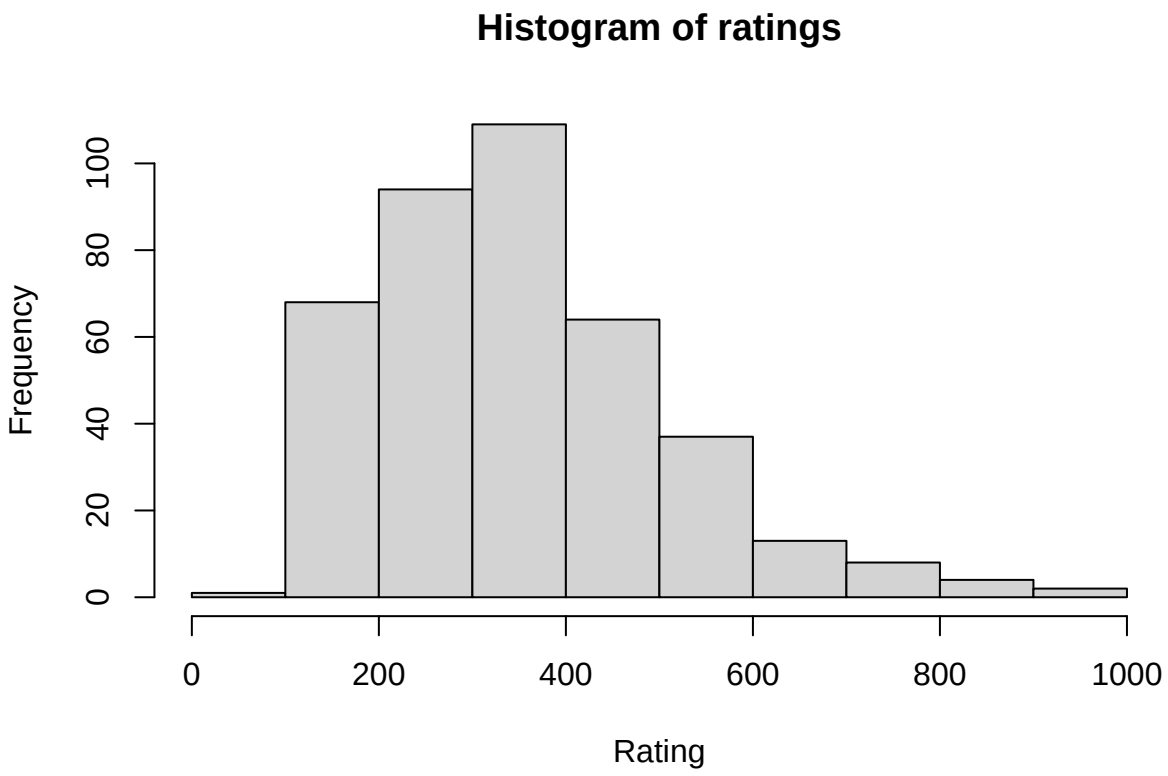
Visualización de variables numéricas con histogramas

```
hist(credit$Income, main = "Histogram of incomes", xlab = "Income (in 1000$)")
```

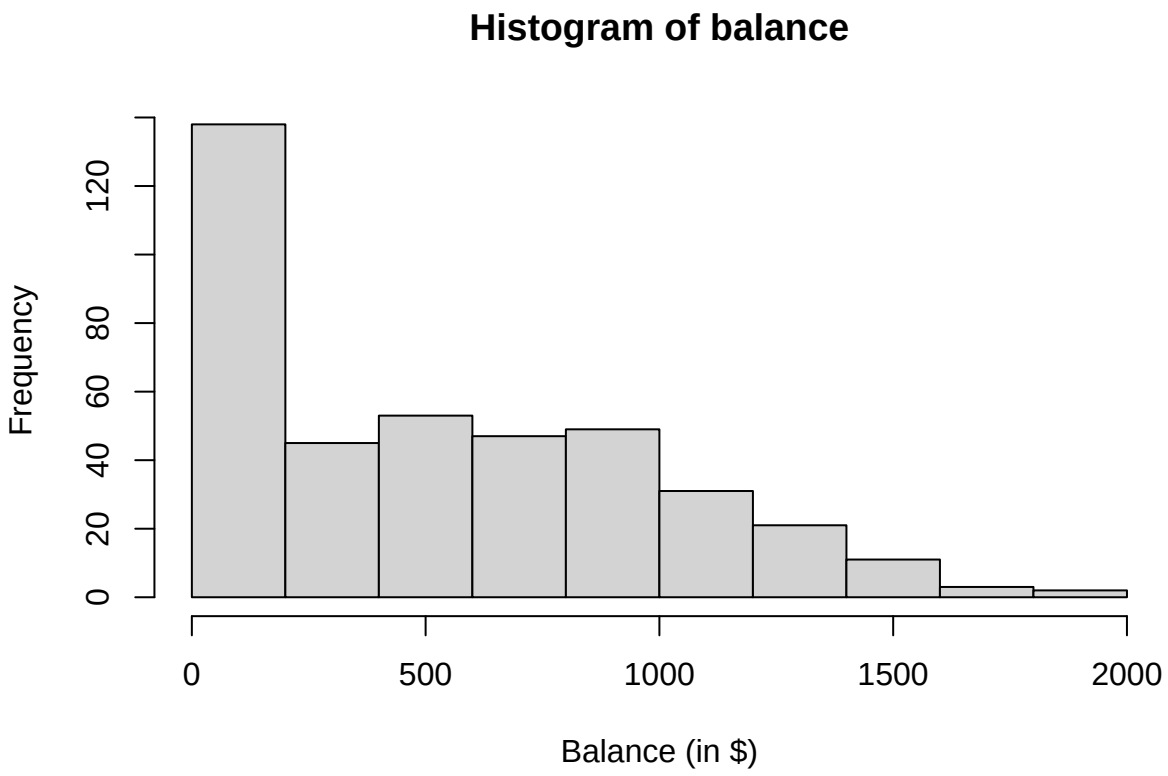


Se puede ver claramente como en los ingresos hay un sesgo hacia la derecha (los valores en el extremo derecho están más dispersos).

```
hist(credit$Rating, main = "Histogram of ratings", xlab = "Rating")
```




```
hist(credit$Balance, main = "Histogram of balance", xlab = "Balance (in $)")
```



También se da el sesgo a la derecha para el rating y el balance. Entiendo que las tres variables están muy relacionadas, lógicamente.

Varianza y desviación estándar

La varianza y la desviación estándard proporcionan información sobre la dispersión de una variable. Una varianza con un valor elevado indica una dispersión elevada. La desviación estándard indica la diferencia de cada valor con la media.

```
# variance and standard deviation of income
var(credit$Income)
```

```
## [1] 1242.159
```

```
sd(credit$Income)
```

```
## [1] 35.24427
```

```
# variance and standard deviation of rating
var(credit$Rating)
```

```
## [1] 23939.56
```

```r
sd(credit$Rating)
```

```
## [1] 154.7241
```

```r
# and balance
var(credit$Balance)
```

```
## [1] 211378.2
```

```r
sd(credit$Balance)
```

```
## [1] 459.7589
```

La varianza del balance, me parece completamente exagerada. Solo puedo deducir que se debe a la presencia de muchas personas que tienen una media de 0 en sus balances.

Exploración de variables categóricas

En el dataset Credit hay 4 variables categóricas (género, "etnia", estado civil, y estado "estudiantil" -la educación viene en años cursados, no por el grado/título obtenido, así que es numérica en este caso). Para analizar variables categóricas se suelen utilizar tablas en lugar de funciones estadísticas:

```r
# Gender
table(credit$Gender)
```

```
##
##   Male Female
##    193    207
```

```r
# Married
table(credit$Married)
```

```
##
##  No Yes
## 155 245
```

```r
# Ethnicity
table(credit$Ethnicity)
```

```
##
## African American            Asian        Caucasian
##               99              102              199
```

Si se quiere ver el porcentaje de observaciones de cada categoría:

```r
# Ethnicity
prop.table(table(credit$Ethnicity))*100
```

```
##
## African American            Asian        Caucasian
##            24.75            25.50            49.75
```