Explorando bases

Diego Alberto Baños Lopez
A01275100

2023-08-17

1. Leer el archivo de trabajo: datos de McDonald

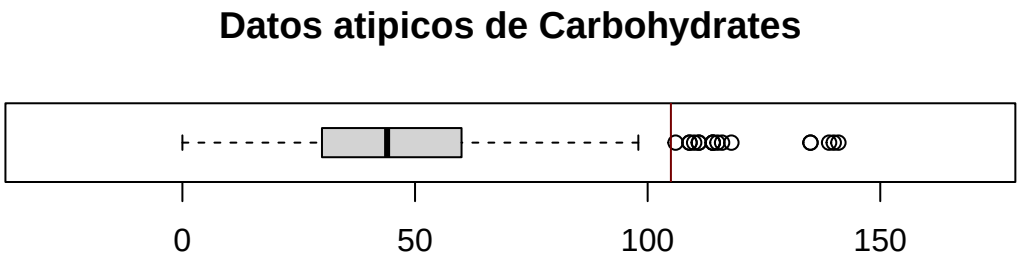
```
# Lectura del csv (En caso de replicar favor de cambiar setwd)
# setwd le dice a R en que carpeta debe de buscar el CSV
setwd("E:/Seagate_4tb/Documentos/Github_clone/ML-personal/ML_repo_personal/")
M <- read.csv("./Archivos_R/mc-donalds-menu-1.csv")
```

2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

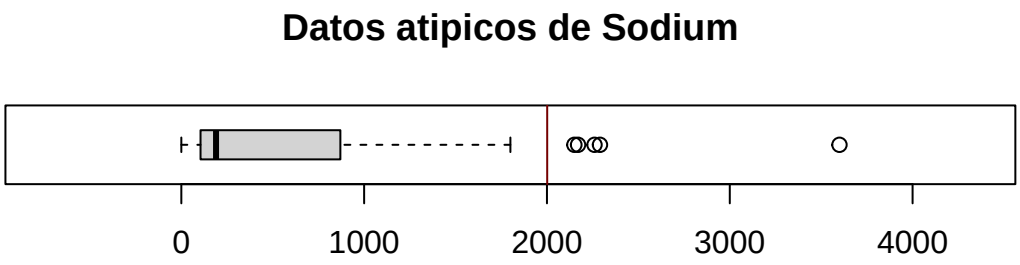
- Calorias
- Carbohidratos
- Proteinas
- Sodio
- Azucares (Sugars)

```
# Tenemos que considerar que las variables dentro del CSV estan en ingles
# Por lo tanto crearemos un arreglo de las variables que se nos piden que
# coincidan con el nombre que tienen las filas en el csv
variablesI <- c("Carbohydrates", "Sodium")

# Analisis y evaluacion de las variables
for (var in variablesI) {
  X <- M[[var]]
  q1 <- quantile(X, 0.25)
  q3 <- quantile(X, 0.75)
  ri <- IQR(X)
  ylim <- c(min(X) - ri, max(X) + ri)
  par(mfrow = c(2, 1))
  boxplot(X, horizontal = TRUE, main = paste("Datos atipicos de", var), ylim = ylim)
  abline(v = q3 + 1.5 * ri, col = "#750000")
  X1 <- M[M[[var]] < q3 + 1.5 * ri, var]
  print("#-----------------------------------------------------------------")
  cat("Resumen para", var, "sin datos atípicos:\n")
  print(summary(X1))
  cat("Resumen para", var, "original:\n")
```

```
  print(summary(X))
  print("-----------------------------------------------------------------#")
}
```




```
## [1] "#-----------------------------------------------------------------"
## Resumen para Carbohydrates sin datos atípicos:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00
## Resumen para Carbohydrates original:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00
## [1] "-----------------------------------------------------------------#"

## [1] "#-----------------------------------------------------------------"
## Resumen para Sodium sin datos atípicos:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    95.0   190.0   456.6   830.0  1800.0
## Resumen para Sodium original:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   107.5   190.0   495.8   865.0  3600.0
## [1] "-----------------------------------------------------------------#"
```



Como se puede observar en carbohidratos la media esta en 47.35 y en sodio la media es de 495.8, en donde la mediana de ambos siendo 43 y 190 respectivamente se peude denotar en los cuartiles grafiicados, y adeomas las graficas nos ayudan a apreciar valores anormales, aproximadamente en carbohidratos a partir de 120 y en sodio a partir de 2000

3.

Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase)

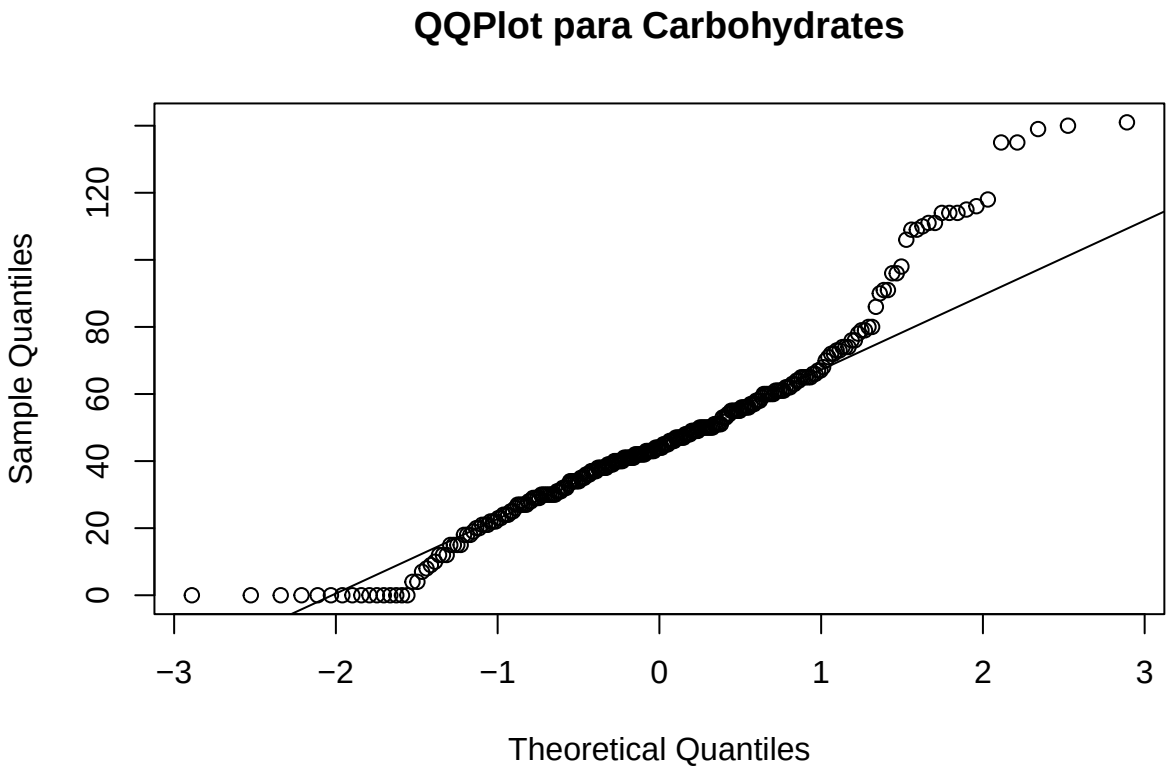
```
# Usando el metodo de Anderson-Darling por la cantidad de datos
# en las variables a analizar
library(nortest)
for (var in variablesI) {
  X <- M[[var]]
  print("#------------------------------------------------------------------")
  cat("Prueba de Anderson-Darling para", var, ":")
  print(ad.test(X))
  print("------------------------------------------------------------------#")
}
```

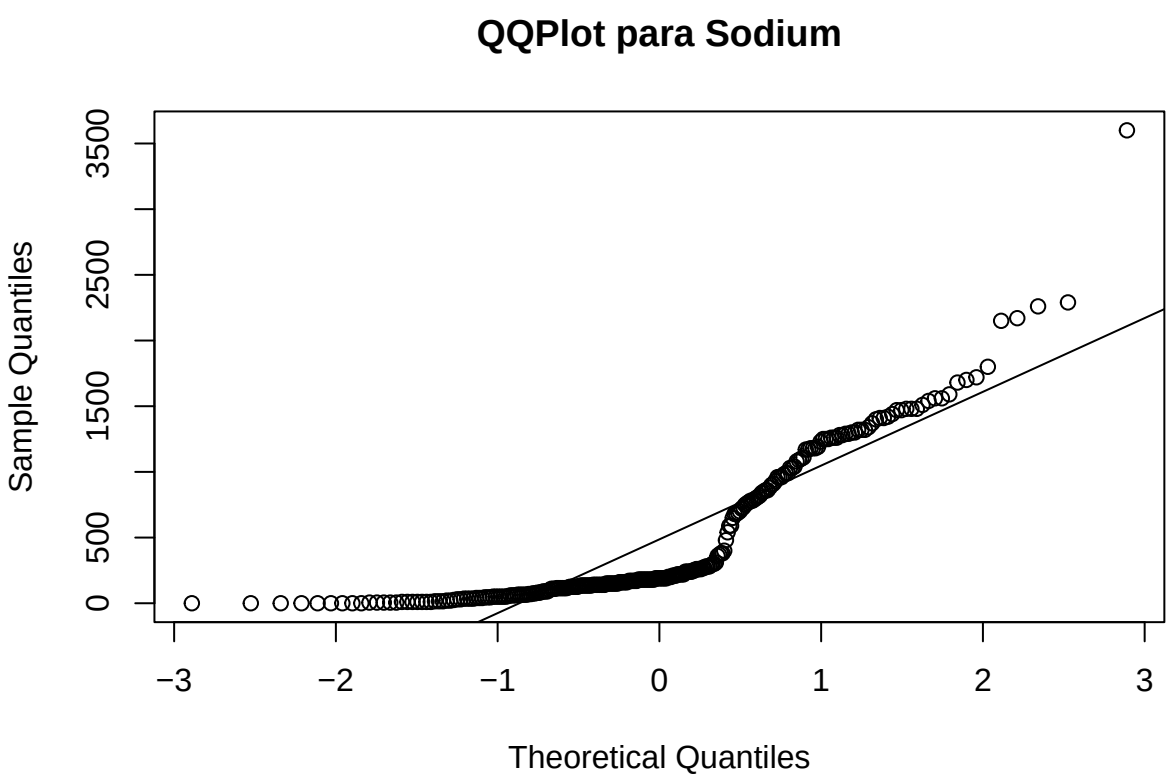
```
## [1] "#------------------------------------------------------------------"
```

```
## Prueba de Anderson-Darling para Carbohydrates :
##  Anderson-Darling normality test
##
## data:  X
## A = 4.1402, p-value = 2.547e-10
##
## [1] "------------------------------------------------------------------#"
## [1] "#------------------------------------------------------------------"
## Prueba de Anderson-Darling para Sodium :
##  Anderson-Darling normality test
##
## data:  X
## A = 21.406, p-value < 2.2e-16
##
## [1] "------------------------------------------------------------------#"
```

Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable

```r
for (var in variablesI) {
  X <- M[[var]]
  qqnorm(X, main = paste("QQPlot para", var))
  qqline(X)
}
```





Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable.

```
library(moments)
for (var in variablesI) {
  X <- M[[var]]
  print("#------------------------------------------------------------------")
  cat("Sesgo para", var)
  print(skewness(X))
  cat("Curtosis para", var)
  print(kurtosis(X))
  print("------------------------------------------------------------------#")
}
```

```
## [1] "#------------------------------------------------------------------"
## Sesgo para Carbohydrates[1] 0.9074253
## Curtosis para Carbohydrates[1] 4.357538
## [1] "------------------------------------------------------------------#"
## [1] "#------------------------------------------------------------------"
## Sesgo para Sodium[1] 1.535166
## Curtosis para Sodium[1] 5.796412
## [1] "------------------------------------------------------------------#"
```

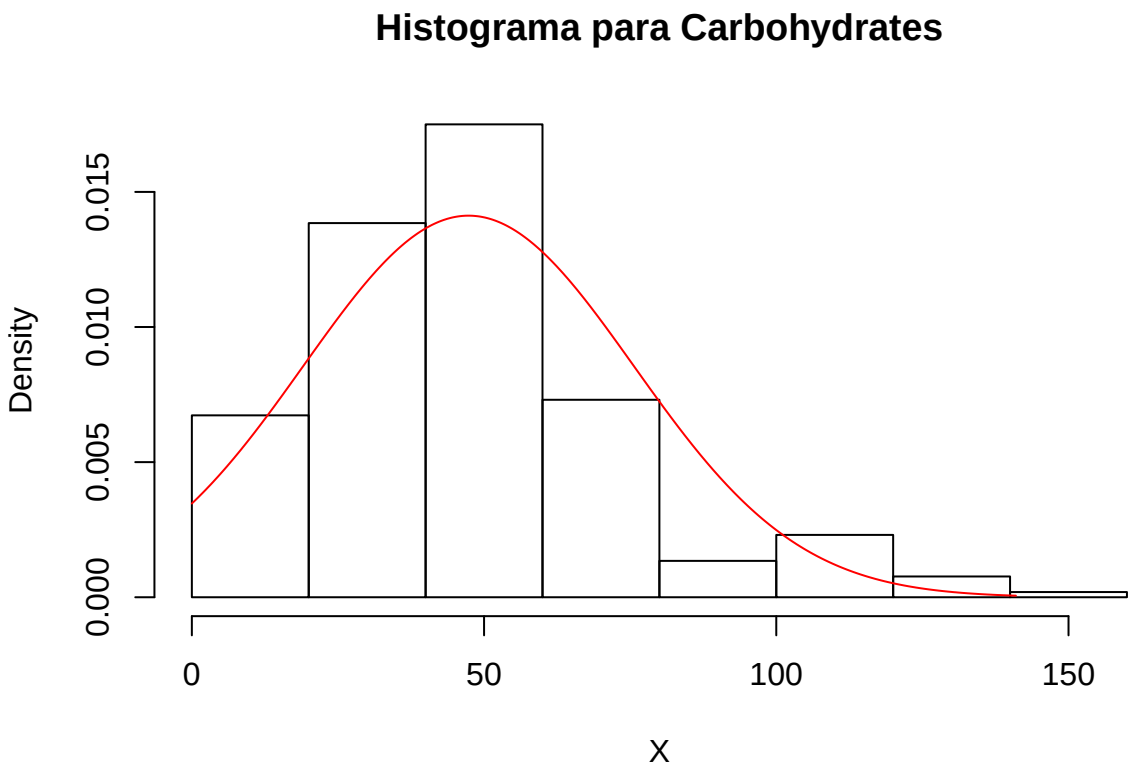
Compara las medidas de media, mediana y rango medio de cada variable.

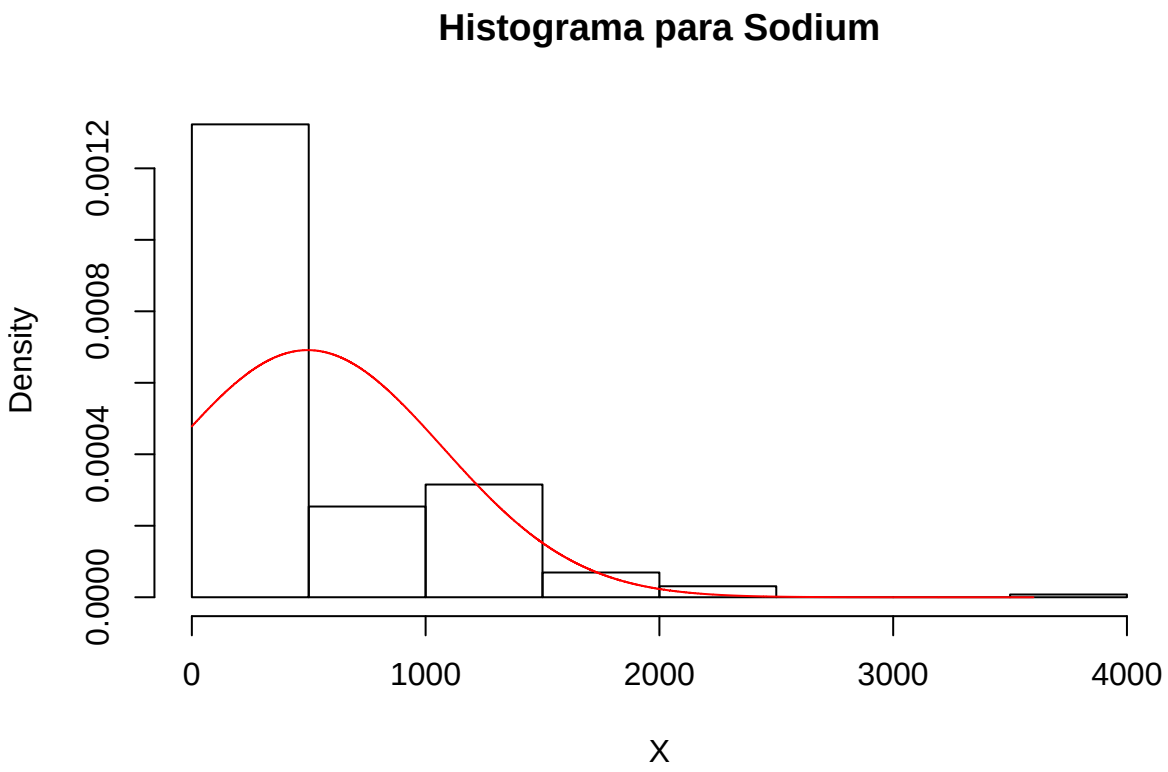
```r
for (var in variablesI) {
  X <- M[[var]]
  print("#------------------------------------------------------------------")
  cat("Media para", var, ":")
  print(mean(X))
  cat("Mediana para", var, ":")
  print(median(X))
  cat("Rango medio para", var, ":")
  print(mean(range(X)))
  print("------------------------------------------------------------------#")
}
```

```
## [1] "#------------------------------------------------------------------"
## Media para Carbohydrates :[1] 47.34615
## Mediana para Carbohydrates :[1] 44
## Rango medio para Carbohydrates :[1] 70.5
## [1] "------------------------------------------------------------------#"
## [1] "#------------------------------------------------------------------"
## Media para Sodium :[1] 495.75
## Mediana para Sodium :[1] 190
## Rango medio para Sodium :[1] 1800
## [1] "------------------------------------------------------------------#"
```

Realiza el histograma y su distribución teórica de probabilidad.

```r
for (var in variablesI) {
  X <- M[[var]]
  hist(X, prob = TRUE, col = 0, main = paste("Histograma para", var))
  x <- seq(min(X), max(X), 0.1)
  y <- dnorm(x, mean(X), sd(X))
  lines(x, y, col = "red")
}
```


Histograma para Carbohydrates
Density
0.000
0.005
0.010
0.015
0
50
100
150
X



Los datos arrojados en este documento para las variables carbohidratos y sodio demuetran lo siguiente: * En las pruebas Anderson-Darling su valor p son muy bajos ($p<0.05$) y sus valores A son demasiados altos, esto hace rechazar la hipótesis nula de que los datos siguen una distribución normal * En cuanto a sesgo en carbohidratos demuestra tener uno positivo hacia la derecha, haciendoce notar en la cola y en el pico que es mas pronunciado que una distribución normal * En cuanto a sesgo en Sodio demuestra tener un sesgo aun mas fuerte hacia la derecha, siendo mas notorio que en el caso de los carbohidrtatos * El analisis de las medidas, en donde se observa diferencias por los valores extremos, en las pruebas y en los coeficientes demuestran un sesgo y curtosis positivo, esto junto con lo dicho anteriormente refuerza la idea que estamos frente a datos que no siguen una distribución normal