4. Explorando bases

Juan Bernal

2024-08-13

1. Baja el archivo de trabajo: datos de McDonald

```
data=read.csv("mc-donalds-menu.csv") #leer la base de datos
```

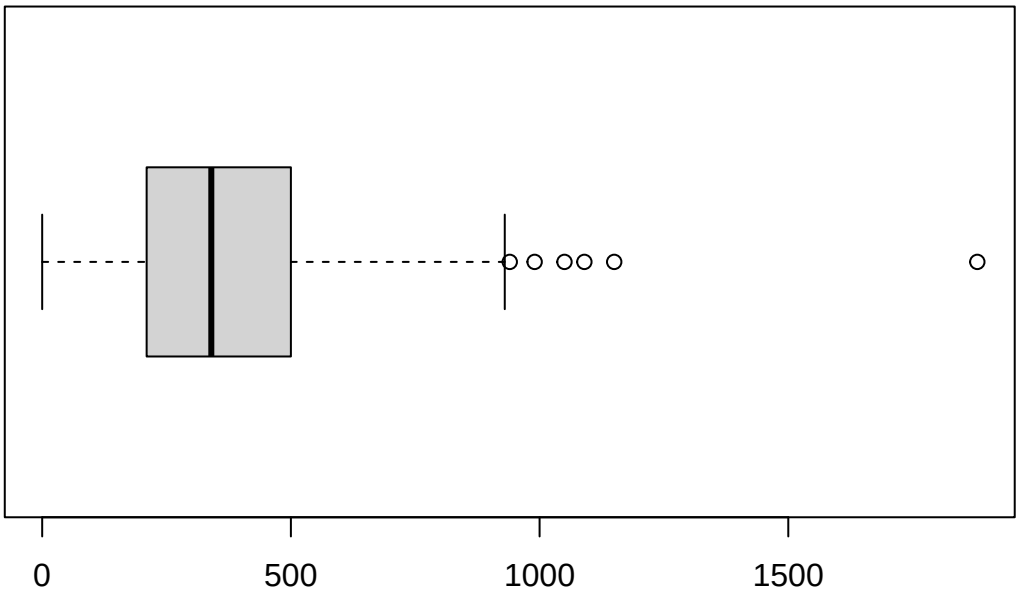
2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

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

3. Para analizar datos atípicos se te sugiere:

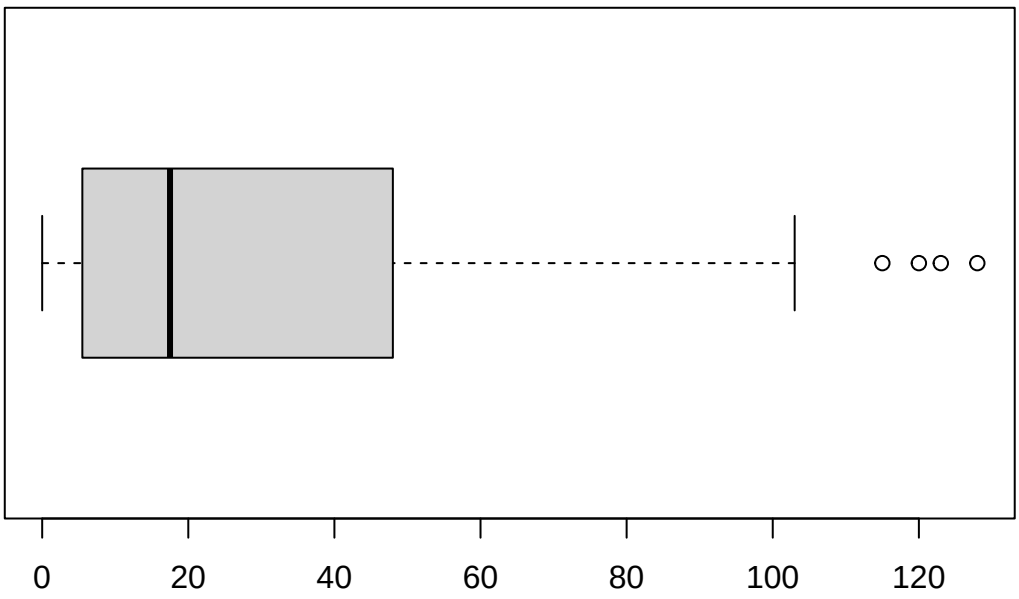
   Graficar el diagrama de caja y bigote

```
boxplot(data$Calories, horizontal = TRUE)
```



En esta boxplot de la variable "Calorías" se observan 5 datos atípicos y 1 dato extremo. Además, se nota un sesgo hacia la derecha, por lo que podemos inferir que la distribución no es normal.

```
boxplot(data$Sugars, horizontal = TRUE)
```



En esta boxplot de la variable "Azúcares" se observan 4 datos atípicos. Además, se nota un sesgo hacia la derecha, por lo que podemos inferir que la distribución no es normal.

Calcula el rango intercuartílico y los cuartíles

```
quantile(data$Calories,c(0.25)) #Cuartil 1 de Calorías
```

```
## 25%
## 210
```

```
quantile(data$Calories,c(0.5)) #Cuartil 2 de Calorías
```

```
## 50%
## 340
```

```
quantile(data$Calories,c(0.75)) #Cuartil 3 de Calorías
```

```
## 75%
## 500
```

```
print("Rango intercuartílico de Calorías")
```

```
## [1] "Rango intercuartílico de Calorías"
```

```
quantile(data$Calories,c(0.75)) - quantile(data$Calories,c(0.25)) #Rango intercuartílico de Calorías
```

```
## 75%
## 290
```

```
quantile(data$Sugars,c(0.25)) #Cuartil 1 de Azúcares
```

```
##  25%
## 5.75
```

```
quantile(data$Sugars,c(0.5)) #Cuartil 2 de AZúcares
```

```
##  50%
## 17.5
```

```
quantile(data$Sugars,c(0.75)) #Cuartil 3 de Azúcares
```

```
## 75%
```

```
##  48
```

```
print("Rango intercuartílico de Azúcares")
```

```
## [1] "Rango intercuartílico de Azúcares"
```

```
quantile(data$Sugars,c(0.75)) - quantile(data$Sugars,c(0.25)) #Rango intercuartílico de Azúcares
```

```
##   75%
## 42.25
```

Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo c

```
print("Cota superior de 1.5 rangos intercuartílicos de las Calorías")
```

```
## [1] "Cota superior de 1.5 rangos intercuartílicos de las Calorías"
```

```
quantile(data$Calories,c(0.75)) + 1.5*(quantile(data$Calories,c(0.75)) - quantile(data$Calories,c(0.25)
```

```
## 75%
## 935
```

```
print("Cota inferior de 1.5 rangos intercuartílicos de las Calorías")
```

```
## [1] "Cota inferior de 1.5 rangos intercuartílicos de las Calorías"
```

```
quantile(data$Calories,c(0.25)) - 1.5*(quantile(data$Calories,c(0.75)) - quantile(data$Calories,c(0.25)
```

```
##  25%
## -225
```

```
print("Los datos atípicos son")
```

```
## [1] "Los datos atípicos son"
```

```
data[data$Calories > 935, 4]
```

```
## [1] 1090 1150  990 1050  940 1880
```

```
data[data$Calories < -225, 4]
```

```
## integer(0)
```

```
print("Cota superior de 1.5 rangos intercuartílicos de los Azúcares")
```

```
## [1] "Cota superior de 1.5 rangos intercuartílicos de los Azúcares"
```

```
quantile(data$Sugars,c(0.75)) + 1.5*(quantile(data$Sugars,c(0.75)) - quantile(data$Sugars,c(0.25)))
```

```
##     75%
## 111.375
```

```
print("Cota inferior de 1.5 rangos intercuartílicos de los Azúcares")
```

```
## [1] "Cota inferior de 1.5 rangos intercuartílicos de los Azúcares"
```

```
quantile(data$Sugars,c(0.25)) - 1.5*(quantile(data$Sugars,c(0.75)) - quantile(data$Sugars,c(0.25)))
```

```
##     25%
## -57.625
```

```
print("Los datos atípicos son")
```

```
## [1] "Los datos atípicos son"
```

```r
data[data$Sugars > 111.375, 19]
```

```
## [1] 123 120 115 128
```

```r
data[data$Sugars < -57.625, 19]
```

```
## integer(0)
```

Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con

```r
print("Cota superior de 3 desviaciones estándar alrededor de la media en las Calorías")
```

```
## [1] "Cota superior de 3 desviaciones estándar alrededor de la media en las Calorías"
```

```r
3*sd(data$Calories)+mean(data$Calories)
```

```
## [1] 1089.079
```

```r
print("Cota inferior de 3 desviaciones estándar alrededor de la media en las Calorías")
```

```
## [1] "Cota inferior de 3 desviaciones estándar alrededor de la media en las Calorías"
```

```r
3*sd(data$Calories)-mean(data$Calories)
```

```
## [1] 352.5404
```

```r
print("Los datos atípicos son")
```

```
## [1] "Los datos atípicos son"
```

```r
data[data$Calories>1089.079, 4]
```

```
## [1] 1090 1150 1880
```

```r
data[data$Calories<352.5404, 4]
```

```
##   [1] 300 250 350 300 150 290 260 240 290 350 190 280 140 220 140 290 340 260
##  [19] 330 250 280 230 340 110  20  15 150 250 160 150  45 330 340 280 140 200
##  [37] 280 100   0   0   0   0 140 190 270 100   0   0   0   0 140 200 280 100
##  [55] 100 130  80 150 190 280   0   0   0   0   0 150 180 220 110   0   0   0
##  [73] 170 210 280 270 340 270 330 260 330 210 260 330 100 130 170 200 250 310
##  [91] 200 250 310 190 240 300 140 170 220 340 270 330 320 250 310 280 340 140
## [109] 190 270 130 180 260 130 180 250 120 170 240  80 120 160 290 350 240 290
## [127] 280 340 230 270 220 260 340 210 250 330 210 260 340 340
```

```r
print("Cota superior de 3 desviaciones estándar alrededor de la media en los Azúcares")
```

```
## [1] "Cota superior de 3 desviaciones estándar alrededor de la media en los Azúcares"
```

```r
3*sd(data$Sugars)+mean(data$Sugars)
```

```
## [1] 115.4625
```

```r
print("Cota inferior de 3 desviaciones estándar alrededor de la media en los Azúcares")
```

```
## [1] "Cota inferior de 3 desviaciones estándar alrededor de la media en los Azúcares"
```

```r
3*sd(data$Sugars)-mean(data$Sugars)
```

```
## [1] 56.61631
```

```r
print("Los datos atípicos son")
```

```
## [1] "Los datos atípicos son"
```

```
data[data$Sugars>115.4625, 19]
```

```
## [1] 123 120 128
```

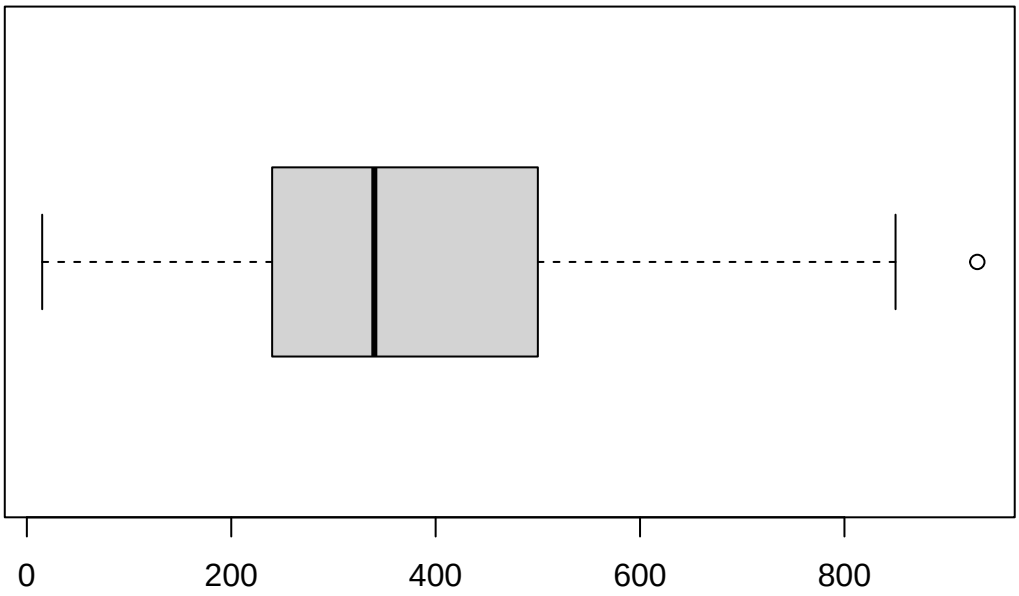
```
data[data$Sugars<56.6163, 19]
```

```
##   [1]  3  3  2  2  2  3  3  4  3  4  2  3  2  3  3  3  3  4  3 15 16 15 15 15  7
##  [26]  8  7  3  3  3  4 17 17 17 18 14 14  2  0 32 32 18  9 10 12 10  9 10  6  7
##  [51]  7 14  7  7  7  6 11 10  8 11  9 11  9 16 14  7  5  6  6  5  7  6  8  6 12
##  [76] 10 14 12  0  0  0  0  1  5  4  5  4  6 12 10  8  7  3  2  3  2  0  0  0  0
## [101]  2  3 23 13 15 13  6 48 43 45 39 55 28  0  0  0  0 35 51 26  0  0  0  0 37
## [126] 54 27 12 22 19 30 39  0  0  0  0  0 36 45 54 27  0  0  0 12 15 20 38 48 38
## [151] 47 36 45 56 12 15 20 13 16 21 39 48 38 48 37 46 56 13 16 21 42 53 43 53 40
## [176] 50 41 51 45 56 46 22 30 45 21 28 42 20 28 41 19 26 39  1  2  2 34 43 35 43
## [201] 33 41 33 41 44 54 44 54 46 56 43 51
```

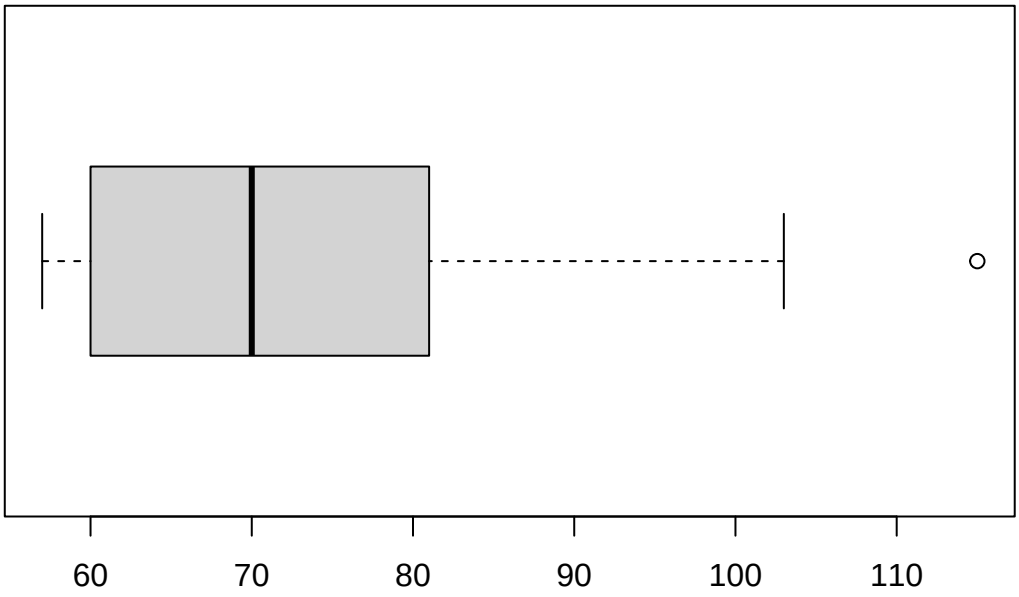
```
Toma una decisión de si conviene o no quitar los datos atípicos (para ello interpreta la variable en el
```

En el caso de las calorías solo conviene quitar los valores atípicos detectados con la cota de rangos intercuartílicos y los ceros, mientras que en el caso de los azúcares parece necesario quitar valores según las cotas de 3 desviaciones estándar alrededor de la media.

```
calories <- data[data$Calories > 0 & data$Calories <= 935, 4]
boxplot(calories, horizontal = TRUE)
```




```
sugar <- data[data$Sugars > 56.6163 & data$Sugars <= 115.4625, 19]
boxplot(sugar, horizontal = TRUE)
```



4. Para analizar normalidad se te sugiere:

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

```
library(nortest) ###REALIZA 10 PRUEBAS DE NORMALIDAD###

###Pruena de Lilliefors (Kolmogorov-Smirnov)###
lillie.test(calories)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  calories
## D = 0.079004, p-value = 0.001048
```

Dado que el p-value es menor al alpha estándar (0.05), se rechaza la hipótesis inicial H0, por lo que la distribución de la variable no es normal.

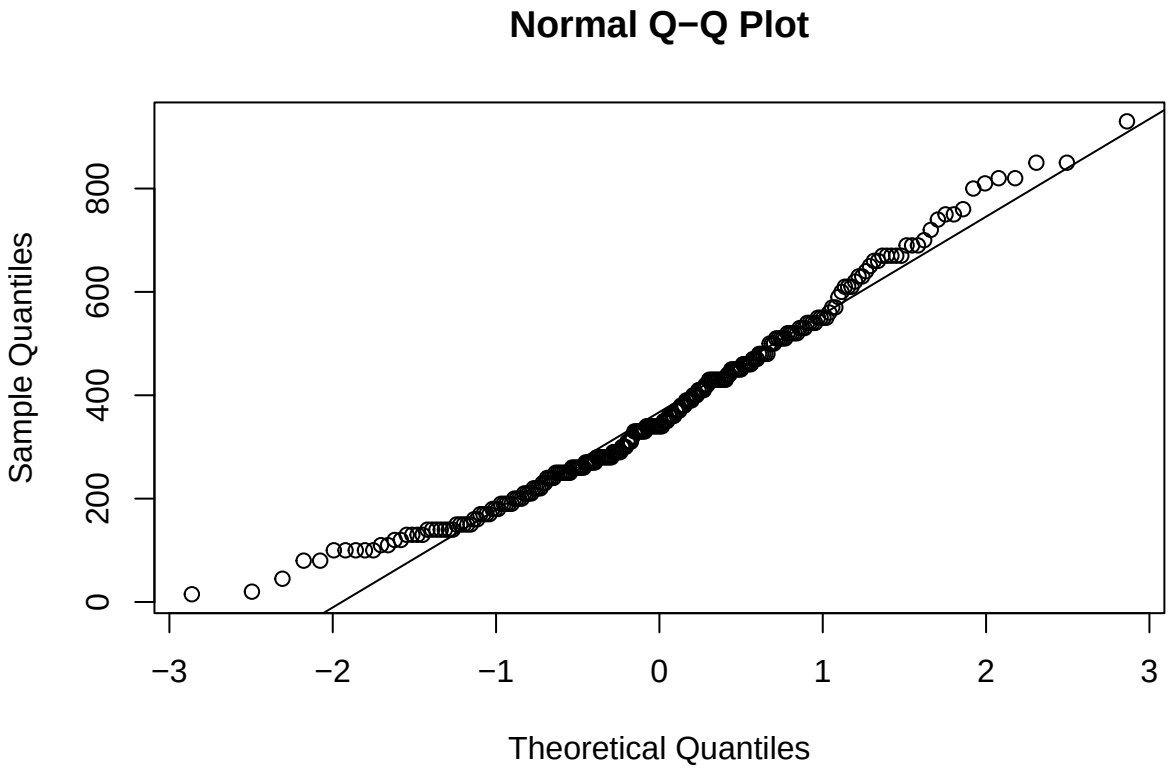
```
lillie.test(sugar)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  sugar
## D = 0.15719, p-value = 0.006989
```

Dado que el p-value es menor al alpha estándar (0.05), se rechaza la hipótesis inicial H0, por lo que la distribución de la variable no es normal.

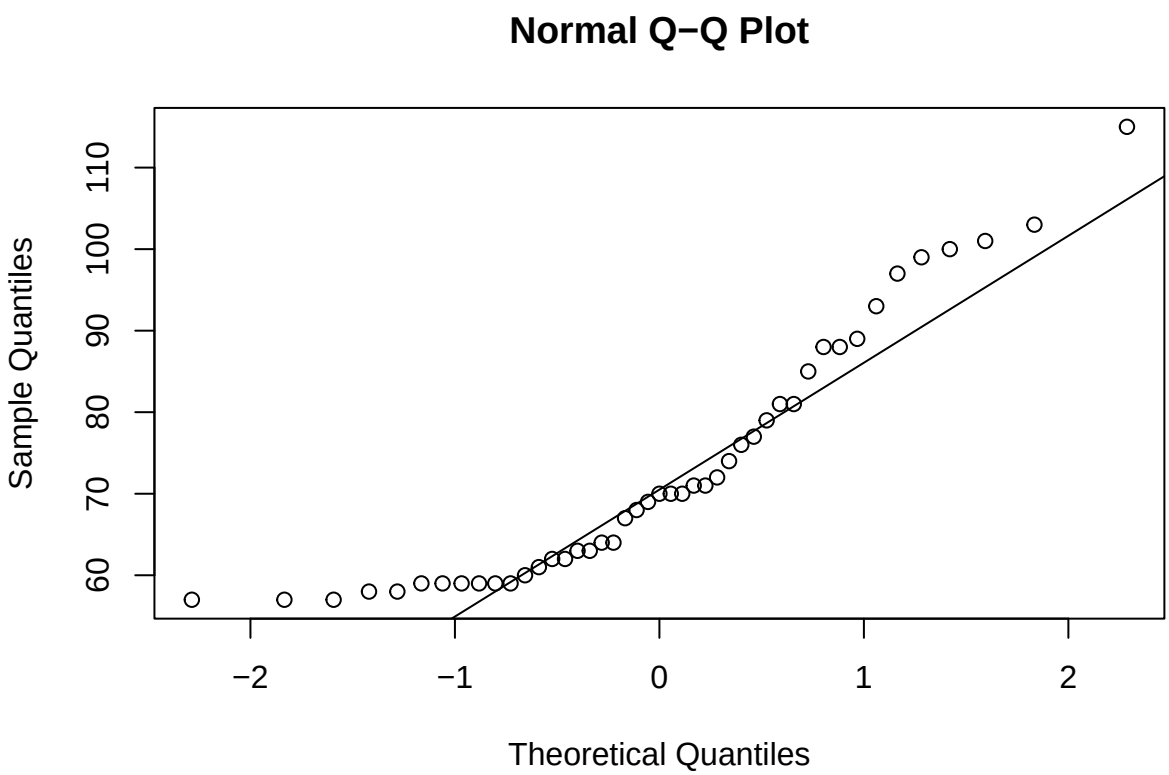
2. Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable

```
qqnorm(calories)
qqline(calories)
```



En la QQplot de las Calorías observamos que las colas tienden a una asimetría positiva, por lo que muestran un sesgo a la derecha.

```
qqnorm(sugar)
qqline(sugar)
```



La QQplot de los Azúcares muestra también asimetría positiva, por lo que también tiene un sesgo a la derecha.

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

```
library(moments)

skewness(calories) #Sesgo de Calorías
```

```
## [1] 0.5413006
```

```
kurtosis(calories) #Curtosis de Calorías
```

```
## [1] 2.790177
```

Las calorías ahora demuestran una curtosis cercana a 0 y un sesgo de aproximadamente 0.5, por lo que se muestra una distribución casi normal.

```
skewness(sugar) #Sesgo de Azúcares
```

```
## [1] 0.9240127
```

```
kurtosis(sugar) #Curtosis de Azúcares
```

```
## [1] 2.863312
```

Los azúcares demuestran una curtosis cercana a 0 y un sesgo de aproximadamente 1, por lo que se acercan a una distribución normal.

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

```
summary(calories) #Media y mediana de las Calorías
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0   240.0   340.0   372.5   495.0   930.0
```

```
print("Rango medio de las Calorías")
```

```
## [1] "Rango medio de las Calorías"
```

```
(max(calories) - min(calories))/2 #Rango medio de las Calorías
```

```
## [1] 457.5
```

```
summary(sugar) #Media y mediana de las Azúcares
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    57.0    60.0    70.0    73.2    81.0   115.0
```

```
print("Rango medio de las Azúcares")
```

```
## [1] "Rango medio de las Azúcares"
```

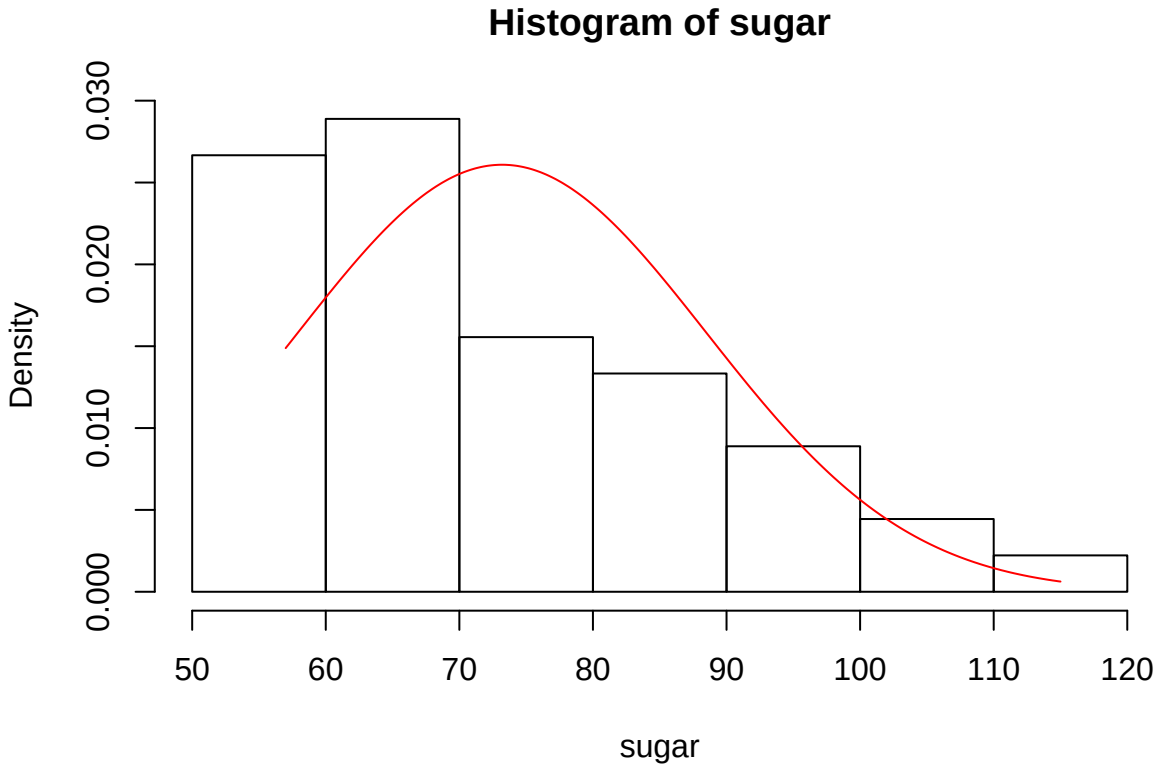
```
(max(sugar) - min(sugar))/2 #Rango medio de los Azúcares
```

```
## [1] 29
```

5. Realiza el histograma y su distribución teórica de probabilidad.

```
hist(calories,prob=TRUE,col=0)
x=seq(min(calories),max(calories),0.1)
y=dnorm(x,mean(calories),sd(calories))
lines(x,y,col="red")
```


```
hist(sugar,prob=TRUE,col=0)
x=seq(min(sugar),max(sugar),0.1)
y=dnorm(x,mean(sugar),sd(sugar))
lines(x,y,col="red")
```



6. Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos.

Aún y cuando la forma de los datos en los histogramas se sigue notando un sesgo hacia la derecha para las Calorías y los Azúcares, el corte de los datos atípicos ayudó a mejorar la curtosis y el sesgo que se tenía con los datos originales. Llevando la curtosis a datos aproximados a 0 y el sesgo menor a 1. Es de esta forma que las distribuciones son casi normales.