4. Explorando bases

Ricardo Salinas

2024-08-13

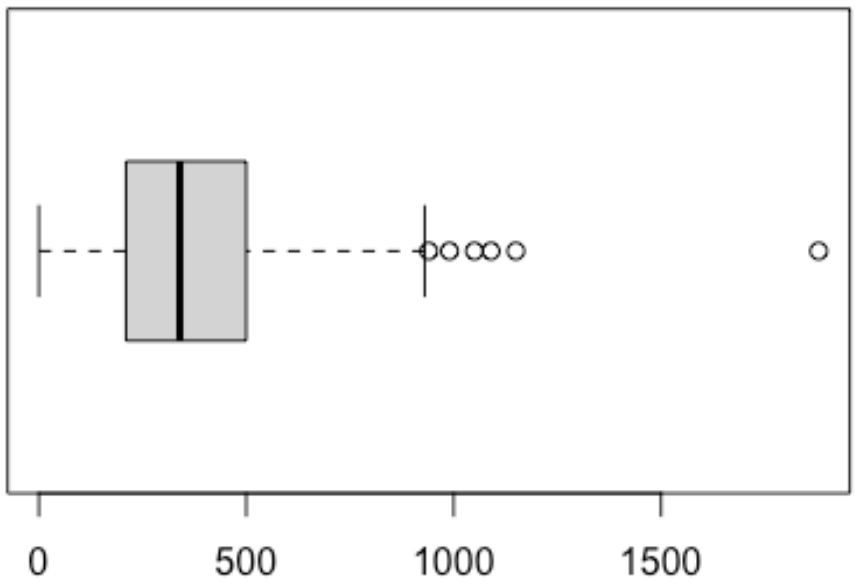
1. Baja el archivo de trabajo: datos de McDonald
2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

```
M = read.csv("Downloads/mcdonalds.csv") #leer la base de datos
M$Calories   #para llamar una variable, aunque también la puedes leer con
corchetes cuadrados M[renglón, columna]
```

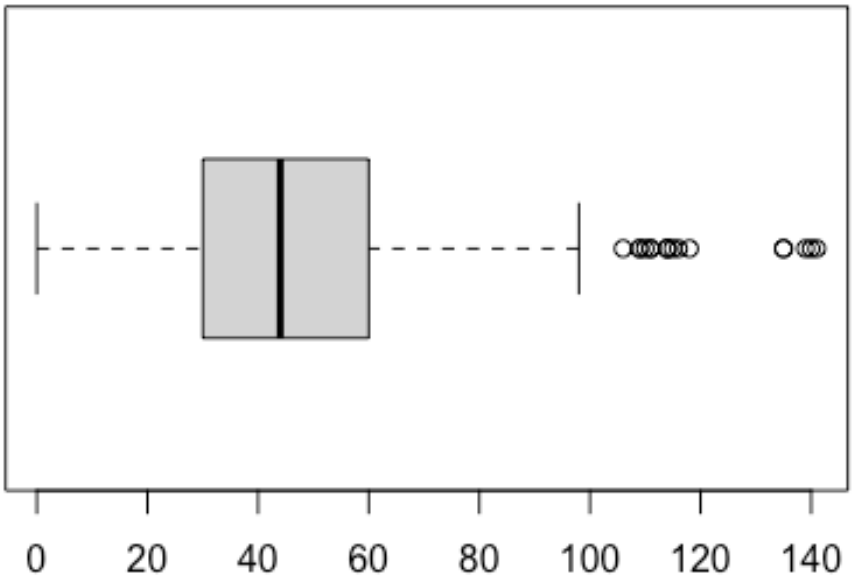
```
##   [1]  300  250  370  450  400  430  460  520  410  470  430  480  510
570  460
##  [16]  520  410  470  540  460  400  420  550  500  620  570  670  740
800  640
##  [31]  690 1090 1150  990 1050  350  520  300  150  460  290  260  530
520  600
##  [46]  610  540  750  240  290  430  720  380  440  430  430  500  510
350  670
##  [61]  510  610  450  750  590  430  360  480  430  360  630  480  610
450  670
##  [76]  520  540  380  190  280  470  940 1880  390  140  380  220  140
450  290
##  [91]  340  260  330  250  360  280  230  340  510  110   20   15  150
250  160
## [106]  150   45  330  340  280  140  200  280  100    0    0    0    0
140  190
## [121]  270  100    0    0    0    0  140  200  280  100  100  130   80
150  190
## [136]  280    0    0    0    0    0  150  180  220  110    0    0    0
170  210
## [151]  280  270  340  430  270  330  430  260  330  420  210  260  330
100  130
## [166]  170  200  250  310  200  250  310  190  240  300  140  170  220
340  410
## [181]  500  270  330  390  320  390  480  250  310  370  360  440  540
280  340
## [196]  400  140  190  270  130  180  260  130  180  250  120  170  240
80  120
## [211]  160  290  350  480  240  290  390  280  340  460  230  270  370
450  550
## [226]  670  450  550  670  530  630  760  220  260  340  210  250  330
210  260
## [241]  340  530  660  820  550  690  850  560  700  850  660  820  650
930  430
## [256]  510  690  340  810  410
```

3. Para analizar datos atípicos se te sugiere:

```
#Graficar el diagrama de caja y bigote
boxplot(M$Calories,horizontal=TRUE)
```




```
boxplot(M$Carbohydrates,horizontal=TRUE)
```




```r
#Calcula el rango intercuartílico y los cuartíles
ri=IQR(M$Calories)
print("El rango intercuartilico de Calorias")
```

```
## [1] "El rango intercuartilico de Calorias"
```

```r
print(ri)
```

```
## [1] 290
```

```r
ri2=IQR(M$Carbohydrates)
print("El rango intercuartilico de Carbohydrates")
```

```
## [1] "El rango intercuartilico de Carbohydrates"
```

```r
print(ri2)
```

```
## [1] 30
```

```r
print("El cuartil 1 de Calories es:")
```

```
## [1] "El cuartil 1 de Calories es:"
```

```r
q1calories=quantile(M$Calories, 0.25)
print(q1calories)
```

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

```
print("El cuartil 2 de Calories es:")
```

```
## [1] "El cuartil 2 de Calories es:"
```

```
q2calories=quantile(M$Calories, 0.50)
print(q2calories)
```

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

```
print("El cuartil 3 de Calories es:")
```

```
## [1] "El cuartil 3 de Calories es:"
```

```
q3calories=quantile(M$Calories, 0.75)
print(q3calories)
```

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

```
print("El cuartil 1 de Carbohydrates es:")
```

```
## [1] "El cuartil 1 de Carbohydrates es:"
```

```
q1carbs=quantile(M$Carbohydrates, 0.25)
print(q1carbs)
```

```
## 25%
##  30
```

```
print("El cuartil 2 de Carbohydrates es:")
```

```
## [1] "El cuartil 2 de Carbohydrates es:"
```

```
q2carbs=quantile(M$Carbohydrates, 0.50)
print(q2carbs)
```

```
## 50%
##  44
```

```
print("El cuartil 3 de Carbohydrates es:")
```

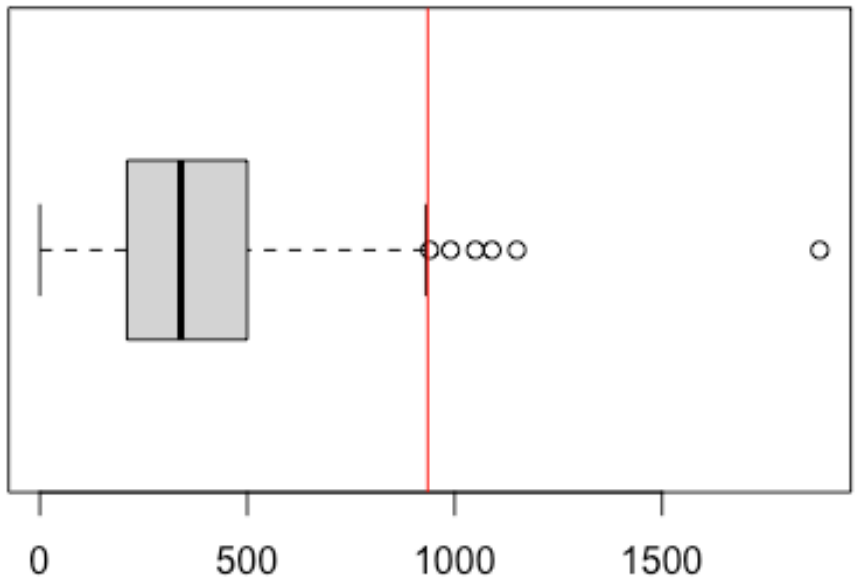
```
## [1] "El cuartil 3 de Carbohydrates es:"
```

```
q3carbs=quantile(M$Carbohydrates, 0.75)
print(q3carbs)
```

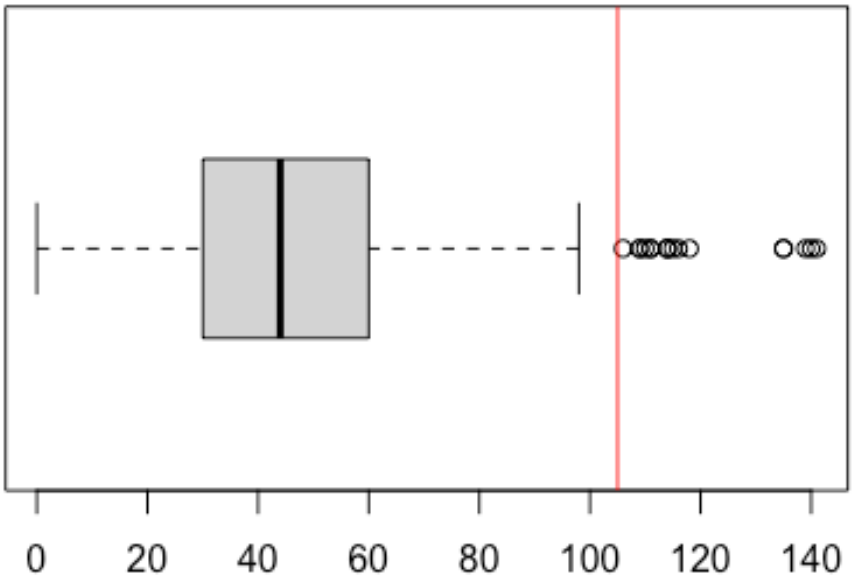
```
## 75%
##  60
```

```
#Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay
datos atípicos de acuerdo con este criterio?
```

```r
boxplot(M$Calories,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje Y
abline(v=q3calories+1.5*ri, col="red")  #linea vertical en el límite de los datos atípicos o extremos
```




```r
boxplot(M$Carbohydrates,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje
abline(v=q3carbs+1.5*ri2, col="red")  #linea vertical en el límite de los datos atípicos o extremos
```

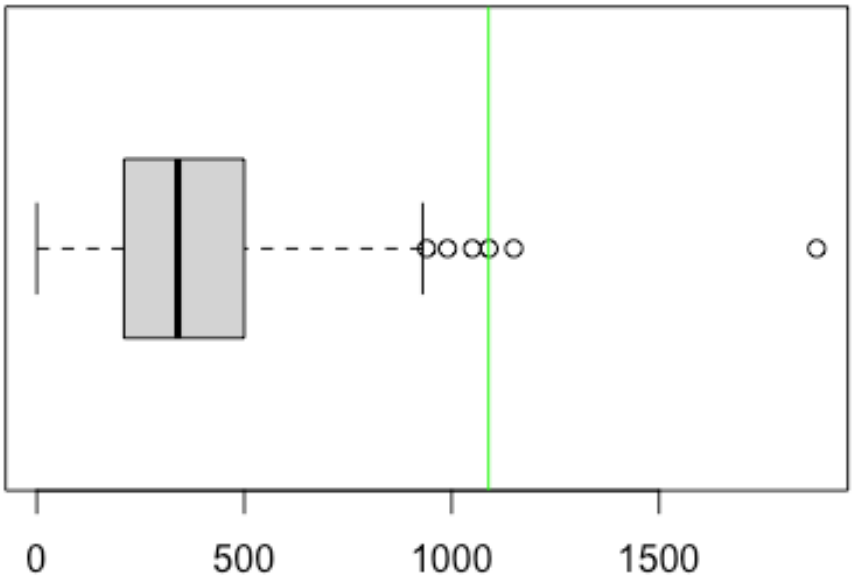



```
print("Ambas variables contienen datos atipicos, ya que estos sobrepasan la linea roja la cual delimita hasta donde se consigera un dato atipico")
```

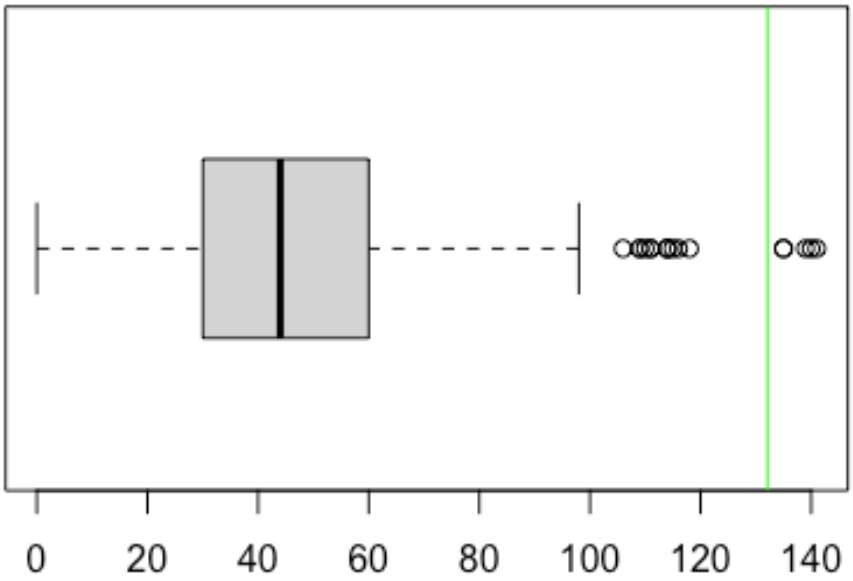
```
## [1] "Ambas variables contienen datos atipicos, ya que estos sobrepasan la linea roja la cual delimita hasta donde se consigera un dato atipico"
```

```
#Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio?

boxplot(M$Calories,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje
abline(v=mean(M$Calories)+3*sd(M$Calories), col="green")  #linea vertical en el límite de los datos atípicos o extremos
```




```r
boxplot(M$Carbohydrates,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje
abline(v=mean(M$Carbohydrates)+3*sd(M$Carbohydrates), col="green")  #linea vertical en el límite de los datos atípicos o extremos
```




```
print("Ambas graficas demuestran datos atipicos fuera del rango de las 3 desviaciones estandar")
```

```
## [1] "Ambas graficas demuestran datos atipicos fuera del rango de las 3 desviaciones estandar"
```

```
#Toma una decisión de si conviene o no quitar los datos atípicos (para ello interpreta la variable en el contexto del problema y determina si es necesario quitarlos o no quitarlos)
```

```
print("En mi opinion no deberiamos quitar los datos atipicos ya que aunque salgan del rango, son importantes porque el restaurante cuenta con muchos productos de gran importancia, los cuales pueden contar con una mayor cantidad calorica")
```

```
## [1] "En mi opinion no deberiamos quitar los datos atipicos ya que aunque salgan del rango, son importantes porque el restaurante cuenta con muchos productos de gran importancia, los cuales pueden contar con una mayor cantidad calorica"
```

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

```r
library(moments)
library(nortest)
ad.test(M$Calories)
```

```
##
##  Anderson-Darling normality test
##
## data:  M$Calories
## A = 2.5088, p-value = 2.369e-06
```

```r
ad.test(M$Total.Fat)
```

```
##
##  Anderson-Darling normality test
##
## data:  M$Total.Fat
## A = 6.7424, p-value < 2.2e-16
```

```r
shapiro.test(M$Calories)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  M$Calories
## W = 0.91902, p-value = 1.119e-10
```

```r
shapiro.test(M$Total.Fat)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  M$Total.Fat
## W = 0.83217, p-value = 4.389e-16
```

```r
cvm.test(M$Calories)
```

```
##
##  Cramer-von Mises normality test
##
## data:  M$Calories
## W = 0.38145, p-value = 4.102e-05
```

```r
cvm.test(M$Total.Fat)
```

```
##
##  Cramer-von Mises normality test
##
## data:  M$Total.Fat
## W = 0.93193, p-value = 2.855e-09
```

```r
lillie.test(M$Calories)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  M$Calories
## D = 0.073753, p-value = 0.001611
```

```
lillie.test(M$Total.Fat)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  M$Total.Fat
## D = 0.15935, p-value < 2.2e-16
```

```
jarque.test(M$Calories)
```

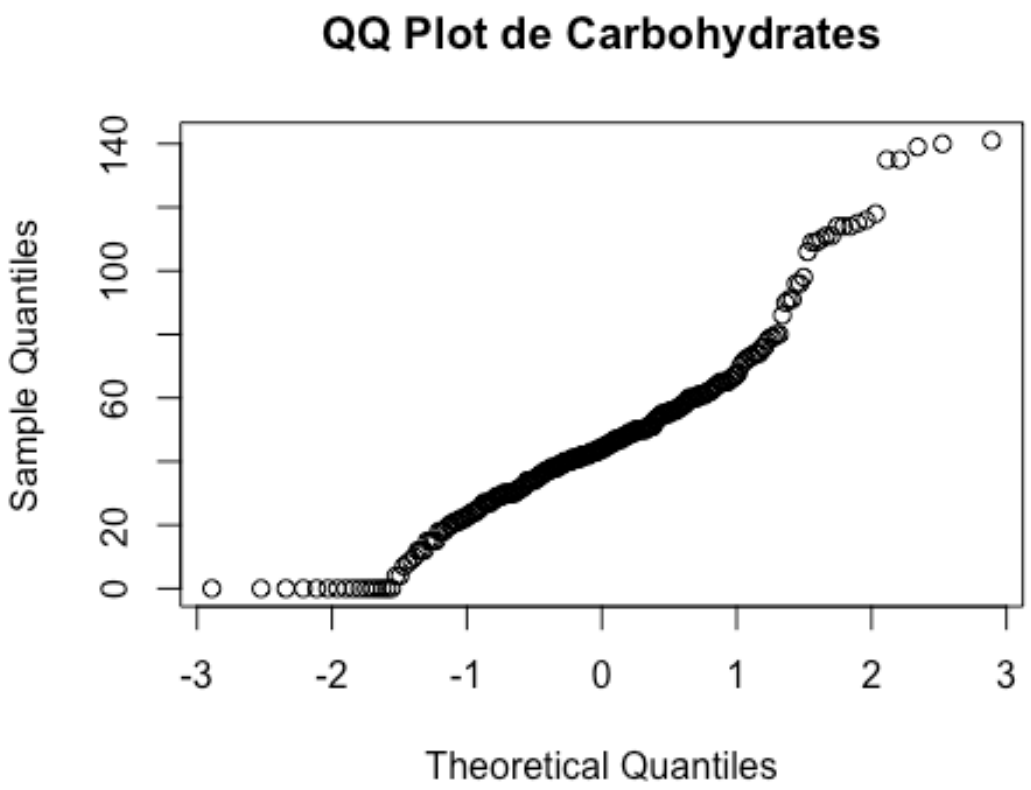
```
##
##  Jarque-Bera Normality Test
##
## data:  M$Calories
## JB = 435.62, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
jarque.test(M$Total.Fat)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  M$Total.Fat
## JB = 1382.7, p-value < 2.2e-16
## alternative hypothesis: greater
```

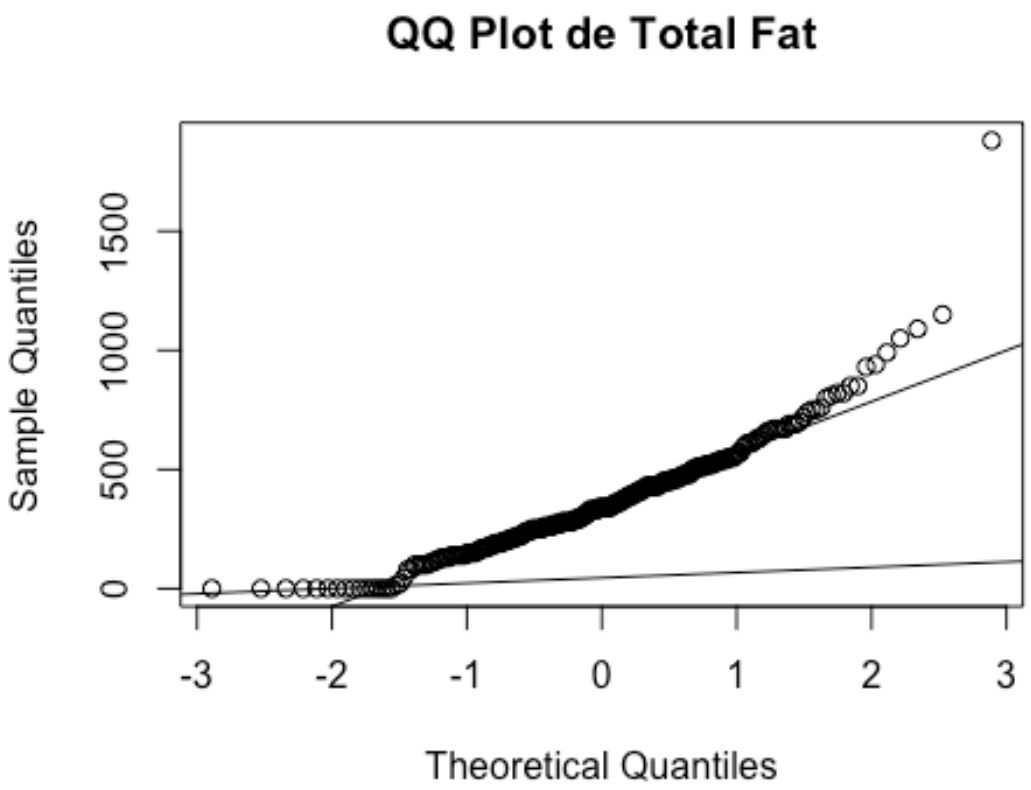
2.Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable

```
qqnorm(M$Carbohydrates, main = "QQ Plot de Carbohydrates")
```




```r
qqnorm(M$Calories, main = "QQ Plot de Total Fat")


qqline(M$Carbohydrates, main = "QQ Plot de Carbohydrates")
qqline(M$Calories, main = "QQ Plot de Total Fat")
```




```
print("En estas graficas podemos ver que la mayoria de los datos se llegan a concentrar dentro de un intervalo, ambas tambien demostrando tener datos un tanto atipicos siendo que estos estan separados de los clusters de datos.")
```

```
## [1] "En estas graficas podemos ver que la mayoria de los datos se llegan a concentrar dentro de un intervalo, ambas tambien demostrando tener datos un tanto atipicos siendo que estos estan separados de los clusters de datos."
```

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

```
library(moments)

print("El sesgo de Calories es:")
```

```
## [1] "El sesgo de Calories es:"
```

```
skewness(M$Calories)
```

```
## [1] 1.444105
```

```
print('La curtosis de Calories es:')
```

```
## [1] "La curtosis de Calories es:"
```

```
kurtosis(M$Calories)
```

```
## [1] 8.645274
```

```
print('El sesgo de Carbohydrates es:')
```

```
## [1] "El sesgo de Carbohydrates es:"
```

```
skewness(M$Carbohydrates)
```

```
## [1] 0.9074253
```

```
print('La curtosis de Carbohydrates es:')
```

```
## [1] "La curtosis de Carbohydrates es:"
```

```
kurtosis(M$Carbohydrates)
```

```
## [1] 4.357538
```

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

```
print('La media de Calorias es:')
```

```
## [1] "La media de Calorias es:"
```

```
mean(M$Calories)
```

```
## [1] 368.2692
```

```
print('La mediana de Calorias es:')
```

```
## [1] "La mediana de Calorias es:"
```

```
median(M$Calories)
```

```
## [1] 340
```

```
print('El rango de Calorias es:')
```

```
## [1] "El rango de Calorias es:"
```

```
range(M$Calories)
```

```
## [1]    0 1880
```

```
print('La media de Carbohydrates es:')
```

```
## [1] "La media de Carbohydrates es:"
```

```
mean(M$Carbohydrates)
```

```
## [1] 47.34615
```

```
print('La mediana de Carbohydrates es:')
```

```
## [1] "La mediana de Carbohydrates es:"
```

```r
median(M$Carbohydrates)
```

```
## [1] 44
```

```r
print('El rango de Carbohydrates es:')
```

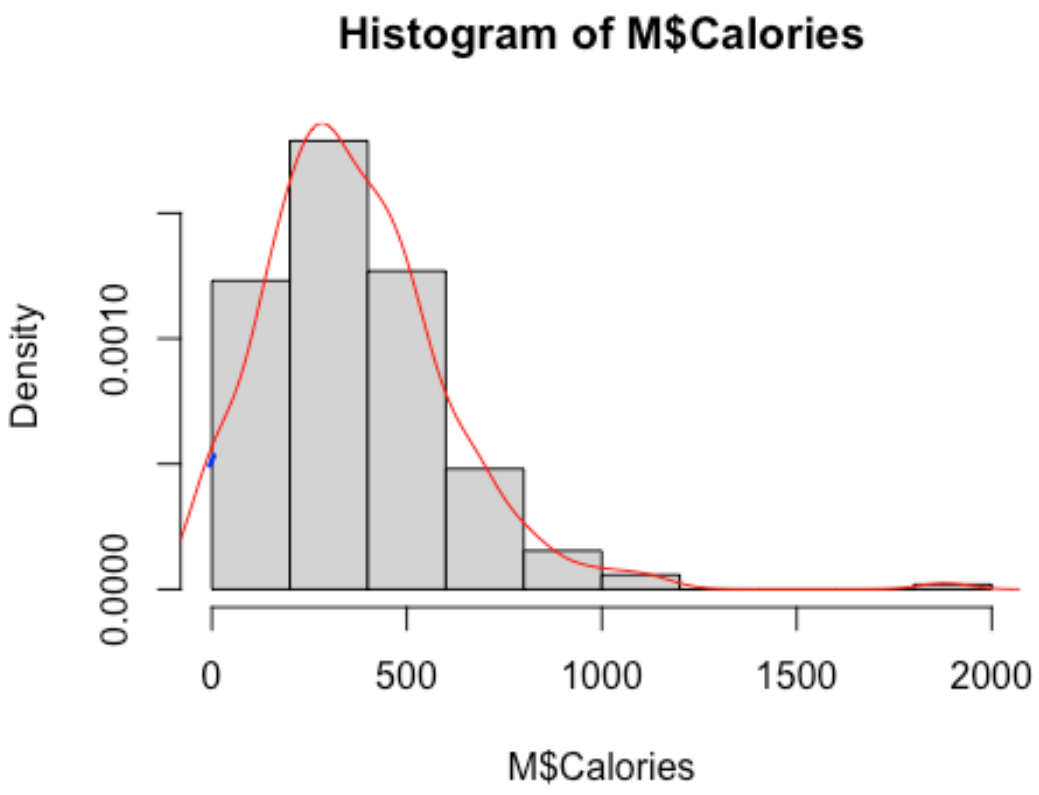
```
## [1] "El rango de Carbohydrates es:"
```

```r
range(M$Carbohydrates)
```

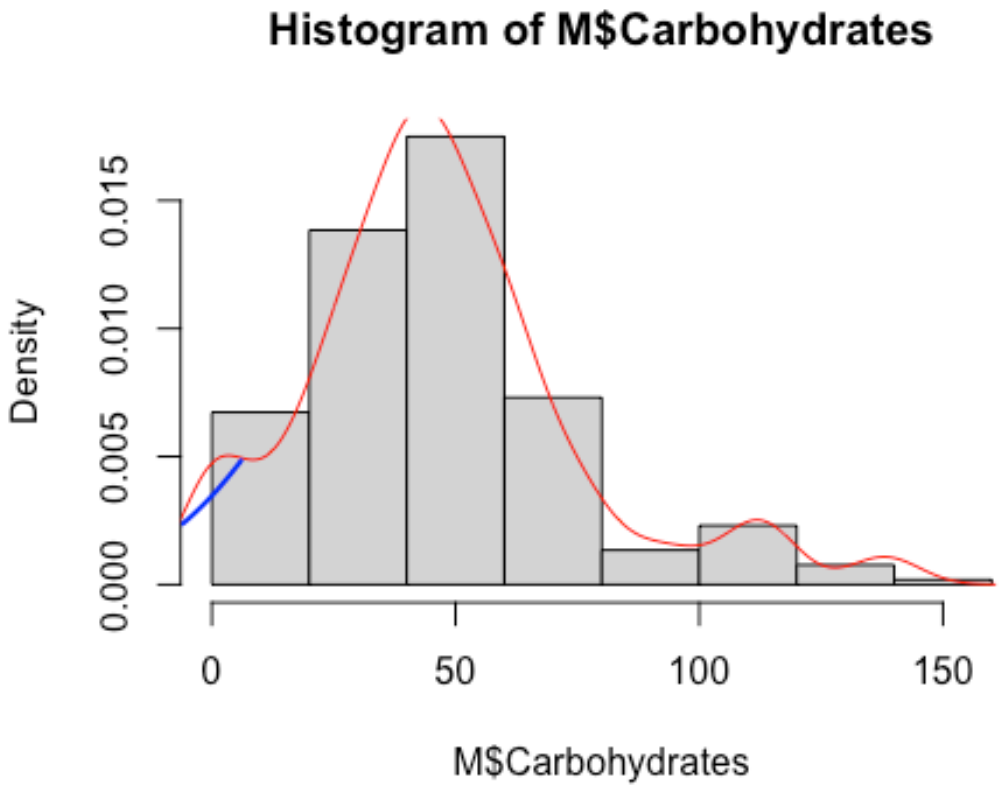
```
## [1]   0 141
```

Realiza el histograma y su distribución teórica de probabilidad

```r
hist(M$Calories,freq=FALSE)
lines(density(M$Calories),col="red")
curve(dnorm(x, mean = mean(M$Calories), sd = sd(M$Calories)), from = -6, to =
6, add = TRUE, col = "blue", lwd = 2)
```




```r
hist(M$Carbohydrates,freq=FALSE)
lines(density(M$Carbohydrates),col="red")
curve(dnorm(x, mean = mean(M$Carbohydrates), sd = sd(M$Carbohydrates)), from
= -6, to = 6, add = TRUE, col = "blue", lwd = 2)
```



Identifica cómo influyen los datos atípicos en la normalidad de los datos

```
print("Una gran cantidad de datos atipicos o una separacion muy extrema de los datos tipicos puede causar inconsistencias o una pobre asertividad al momento de hacer predicciones y estimaciones, pero tambien se pueden mantener datos atipicos los cuales no esten muy lejos de los datos tipicos cuando estos brindan informacion de valor para las estimaciones y predicciones")
```

```
## [1] "Una gran cantidad de datos atipicos o una separacion muy extrema de los datos tipicos puede causar inconsistencias o una pobre asertividad al momento de hacer predicciones y estimaciones, pero tambien se pueden mantener datos atipicos los cuales no esten muy lejos de los datos tipicos cuando estos brindan informacion de valor para las estimaciones y predicciones"
```