Actividad 2: Explorando bases

Samantha Daniela Guanipa Ugas A01703936

2023-08-14

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

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

1) Se sube y lee la base de datos de McDonald:

```
M=read.csv("mc-donalds-menu-act2.csv") #leer la base de datos
```

2) Se llamarán dos variables que fueron seleccionadas y se calcularán sus cuartiles:

a) *Sodio*

```
q1S = quantile(M$Sodium,0.25) #Se obtiene el primer cuartil de los datos de Sodio
q1S
```

```
##   25%
## 107.5
```

Observación: Se obtuvo que el 25% de los datos son menores a 107.5.

b) *Calorías*

```
q1C = quantile(M$Calories,0.25)
q1C
```

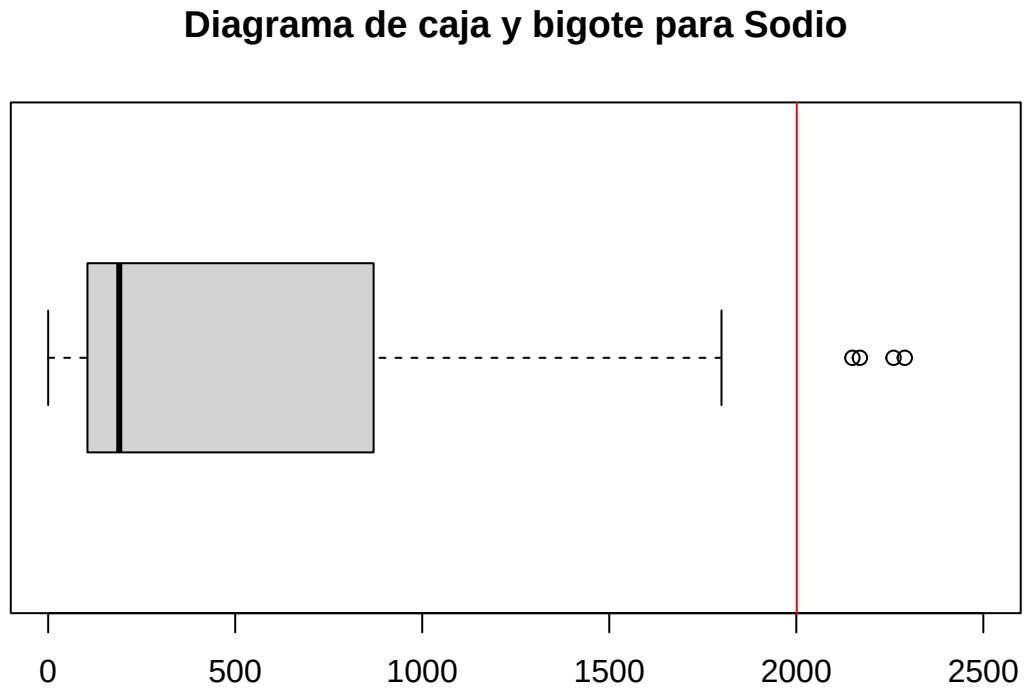
```
## 25%
## 210
```

Observación: Se obtuvo que el 25% de los datos son menores a 210.

3) Se creará el diagrama de caja y bigote para el análisis de:

a)*Sodio*

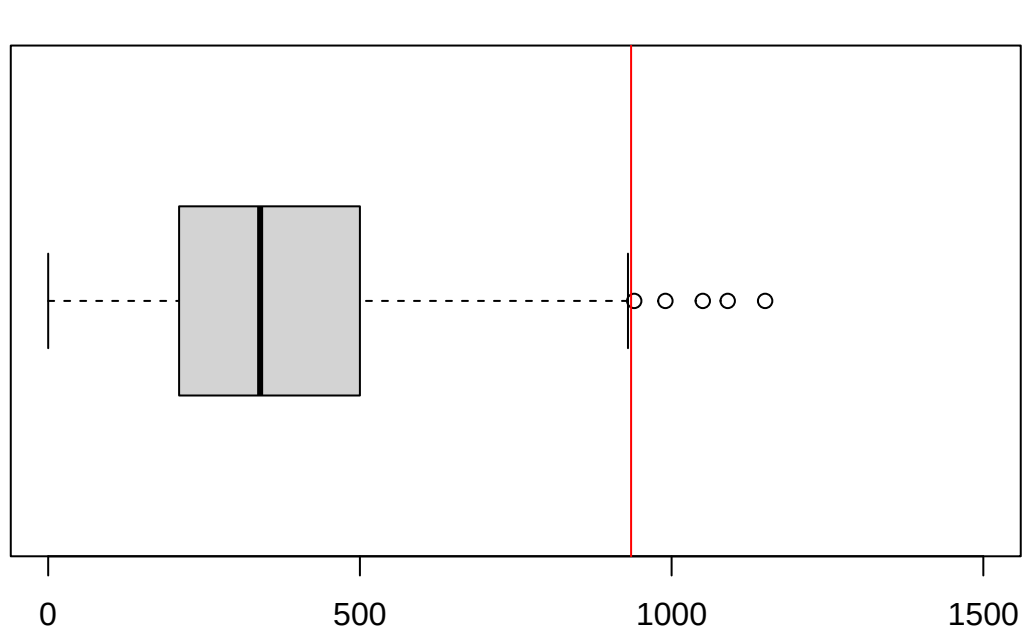
```
q3S = quantile(M$Sodium,0.75) #Se calcula el 3er cuartil
riS= q3S - q1S #Se calcula el rango intercuartilico de X
boxplot(M$Sodium,horizontal=TRUE,ylim=c(0,2500))
abline(v=q3S+1.5*riS,col="red")  #linea vertical en el límite de los datos atípicos o extremos
title(main = "Diagrama de caja y bigote para Sodio")
```



Observación: Se puede ver que el valor mínimo para el sodio de los alimentos en McDonald's es de 0, y su valor máximo está rondando los 1800. Por otra parte, los valores atípicos están entre los 2000 y los 2500. Además, la mediana está más cercana al quartil 1, por lo que se tiene una distribución asimétrica positiva.

b. Calorías

```
q3C = quantile(M$Calories,0.75) #Se calcula el 3er cuartil
riC= q3C - q1C #Se calcula el rango intercuartilico de X
boxplot(M$Calories,horizontal=TRUE,ylim=c(0,1500))
abline(v=q3C+1.5*riC,col="red")  #linea vertical en el límite de los datos atípicos o extremos
title(main = "Diagrama de caja y bigote para Calorías")
```



Observación: Se puede ver que el valor mínimo para las calorías de los alimentos en McDonald's es de 0, y su valor máximo está rondando los 800. Por otra parte, los valores atípicos están entre los 800 y los 1200. Además, la mediana está muy cerca del centro, por lo que se tiene una distribución simétrica.

4) Se calcula el rango intercuartilico de X para:

a) Sodio

```
XS = M[M$Sodium<q3S+1.5*riS,c("Sodium")]  #En la matriz M, quitar datos más allá de 3 rangos intercuart
print("Summary Sodium en un rango normal:")
```

```
## [1] "Summary Sodium en un rango normal:"
```

```
summary(XS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    95.0   190.0   456.6   830.0  1800.0
```

```
print("Summary Sodium:")
```

```
## [1] "Summary Sodium:"
```

```
summary(M$Sodium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   107.5   190.0   495.8   865.0  3600.0
```

Observación: Para este caso, en un rango dentro de lo normal, el valor máximo de sodio en un alimento es de 1800, pero si no se genera este filtrado, el valor máximo podrá alcanzar los 3600 para el sodio en un alimento.

b) Calorías

```
XC = M[M$Calories<q3C+1.5*riC,c("Calories")]  #En la matriz M, quitar datos más allá de 3 rangos interc
print("Summary Calories en un rango normal:")
```

```
## [1] "Summary Calories en un rango normal:"
```

```
summary(XC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   202.5   335.0   349.0   480.0   930.0
```

```
print("Summary Calories:")
```

```
## [1] "Summary Calories:"
```

```
summary(M$Calories)
```

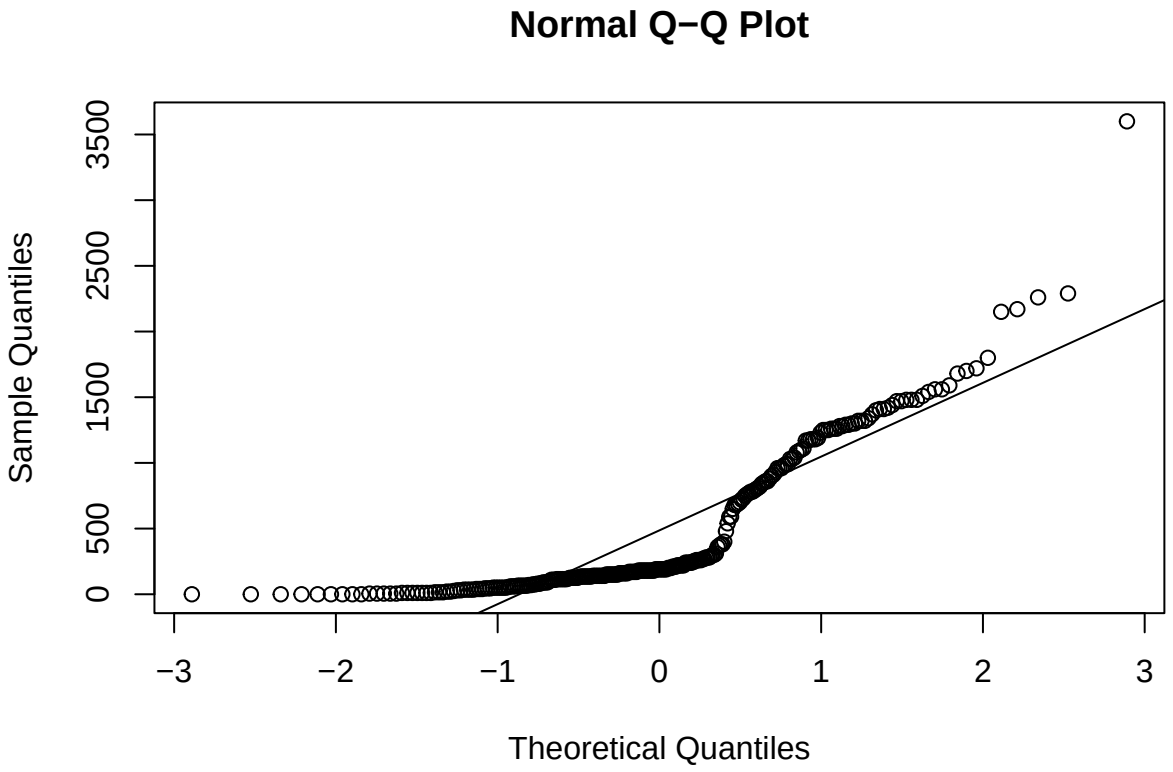
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   210.0   340.0   368.3   500.0  1880.0
```

Observación: Para este caso, en un rango dentro de lo normal, el valor máximo de calorías en un alimento es de 930, pero si no se genera este filtrado, el valor máximo podrá alcanzar los 1880 de calorías en un alimento.

5) Se generan los gráficos Q-Q plot para:

a) Sodio

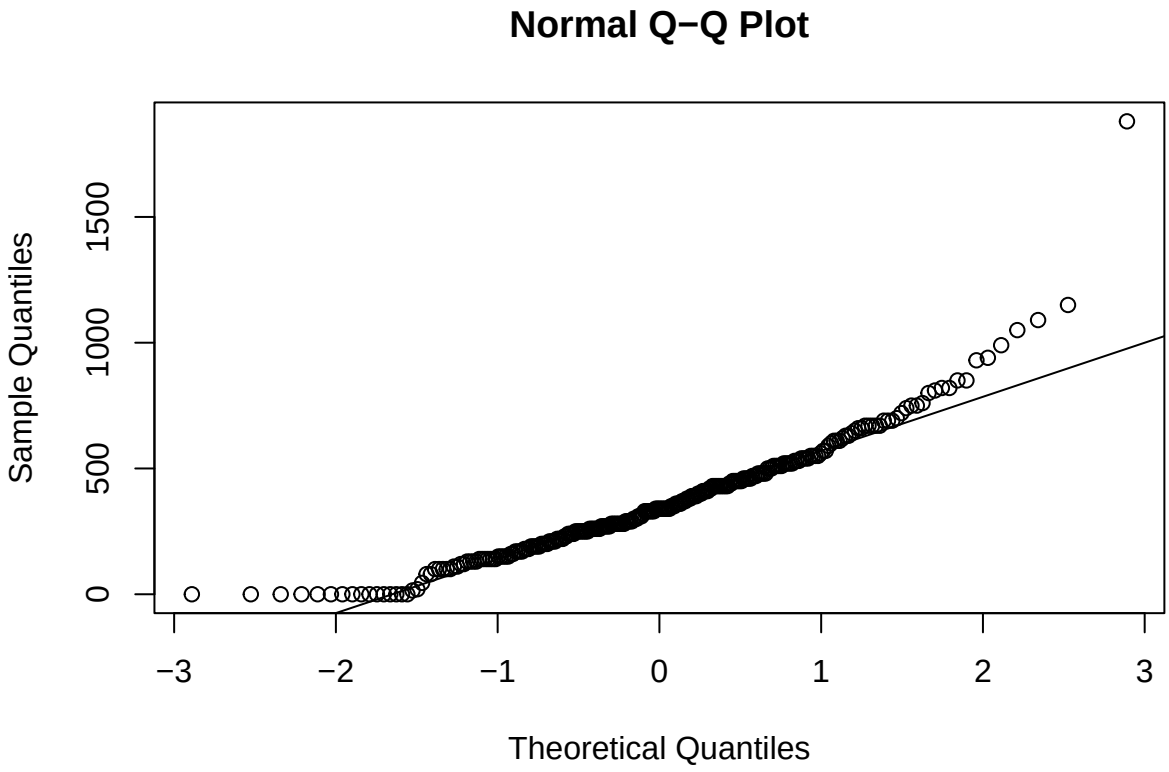
```
qqnorm(M$Sodium)
qqline(M$Sodium)
```



Observación: Como la mayoría de los datos realmente no siguen la línea central no se podría asumir normalidad. Por lo tanto, como se ve que incrementa de manera precipitada se asume que los datos están sesgados a la derecha.

b) Calorías

```
qqnorm(M$Calories)
qqline(M$Calories)
```

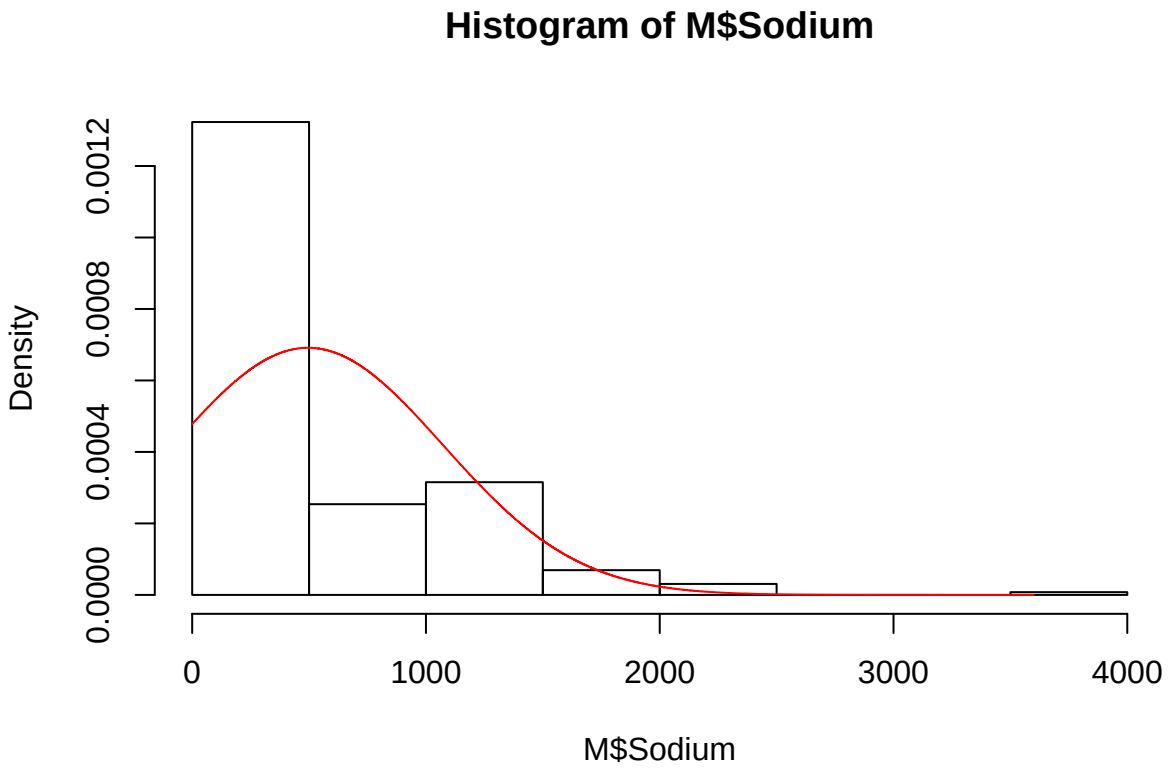


Observación: Como la mayoría de los datos siguen la línea del referencia se puede asumir normalidad.

6) Se generan los gráficos de densidad de probabilidad para:

a) Sodio

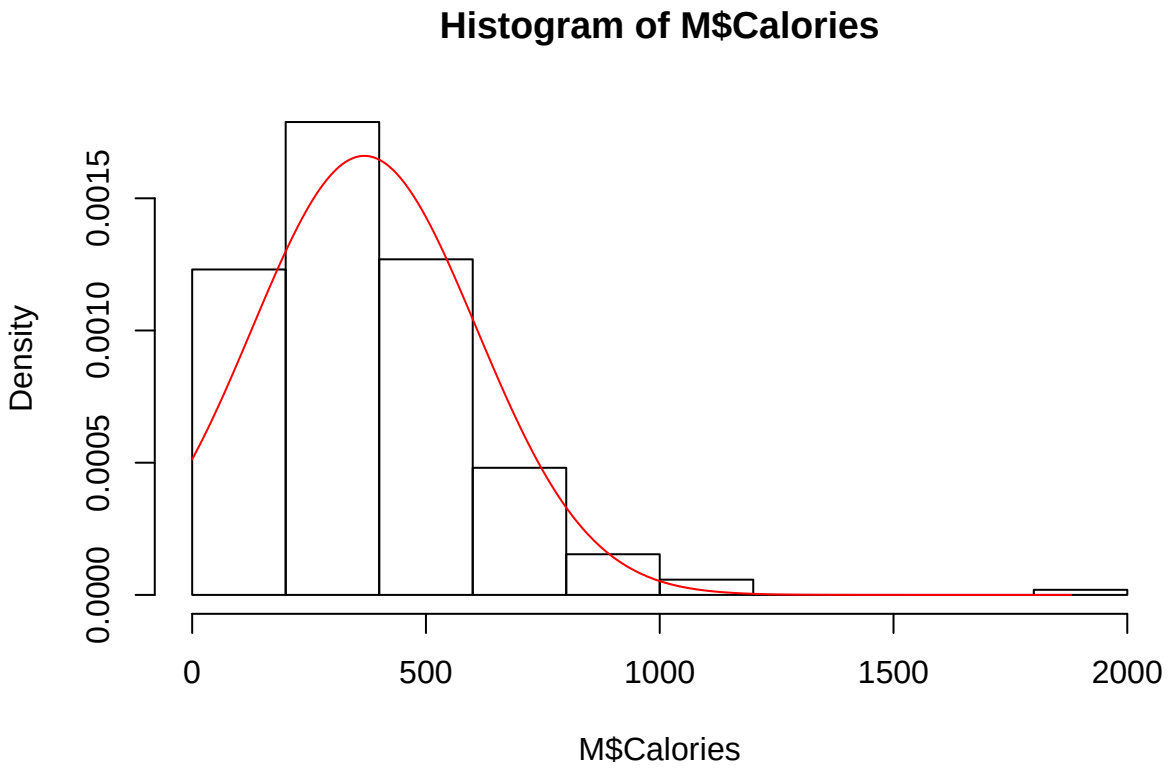
```
hist(M$Sodium,prob=TRUE,col=0)
x=seq(min(M$Sodium),max(M$Sodium),0.1)
y=dnorm(x,mean(M$Sodium),sd(M$Sodium))
lines(x,y,col="red")
```



Observación: Se observa que hay un pico en el valor de densidad de 0.0012 en el rango de sodio 0 a 500, por lo que se deduce que está sesgado a la derecha ya que las colas se extienden a la derecha.

b) Calorías

```
hist(M$Calories,prob=TRUE,col=0)
x1=seq(min(M$Calories),max(M$Calories),0.1)
y1=dnorm(x1,mean(M$Calories),sd(M$Calories))
lines(x1,y1,col="red")
```



Observación: Se observa que hay un pico en el valor de densidad de 0.04 en el rango de calorías de 0 a 0.5, por lo que se deduce que es una distribución asimétrica a la derecha.

7. Se exploran curtosis y sesgos para:

a) Sodio

```
library(moments)
skewness(M$Sodium)
```

```
## [1] 1.535166
```

```
library(e1071)
```

```
##
## Attaching package: 'e1071'
```

```
## The following objects are masked from 'package:moments':
##
##     kurtosis, moment, skewness
```

```
skewness(M$Sodium)
```

```
## [1] 1.526317
```

Observación: Se observa que hay una asimetría positiva. Esto representa que cola derecha de la distribución es más larga, por lo que es una distribución asimétrica a la derecha.

```
library(moments)
kurtosis(M$Sodium)
```

```
## [1] 2.75191
```

```
library(e1071)
kurtosis(M$Sodium)
```

```
## [1] 2.75191
```

Observación: Al obtener un valor mayor a 3 se deduce que la distribución tiene una curtosis "leptocúrtica", y que hay una alta concentración de datos alrededor de la media y son valores extremos.

b) Calorías

```
library(moments)
skewness(M$Calories)
```

```
## [1] 1.435782
```

```
library(e1071)
skewness(M$Calories)
```

```
## [1] 1.435782
```

Observación: Se observa que hay una asimetría positiva. Esto representa que la cola derecha de la distribución es más larga, por lo que es una distribución asimétrica a la derecha.

```
library(moments)
kurtosis(M$Calories)
```

```
## [1] 5.5789
```

```
library(e1071)
kurtosis(M$Calories)
```

```
## [1] 5.5789
```

Observación: Al obtener un valor mayor a 3 se deduce que la distribución tiene una curtosis "leptocúrtica", y que hay una alta concentración de datos alrededor de la media y son valores extremos.