Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad M1. Explorando bases

TC3006C.101 Inteligencia artificial avanzada para la ciencia de datos I

Profesores:

Ivan Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernandez

Antonio Carlos Bento

Frumencio Olivas Alvarez

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

17 de Agosto de 2023

Instrucciones

1. Baja el archivo de trabajo: datos de McDonald

In [ ]:
```
# Montar Google Drive
from google.colab import drive
drive.mount('/content/drive')
file_path = "/content/drive/MyDrive/Colab Notebooks/mc-donalds-menu-1.csv"
```

```
Mounted at /content/drive
```

In [ ]:
```
%load_ext rpy2.ipython
```

In [ ]:
```
%%R
# Cargamos las librerías necesarias
install.packages("moments")
install.packages("e1071")
library(moments)
library(e1071)
```

In [ ]:
```
# Pasamos la variable de la ruta de los datos a R
%Rpush file_path
```

In [ ]:
```
# Cargamos los datos CSV desde Google Drive
%%R
library(readr)

M <- read.csv(file_path)
```

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

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

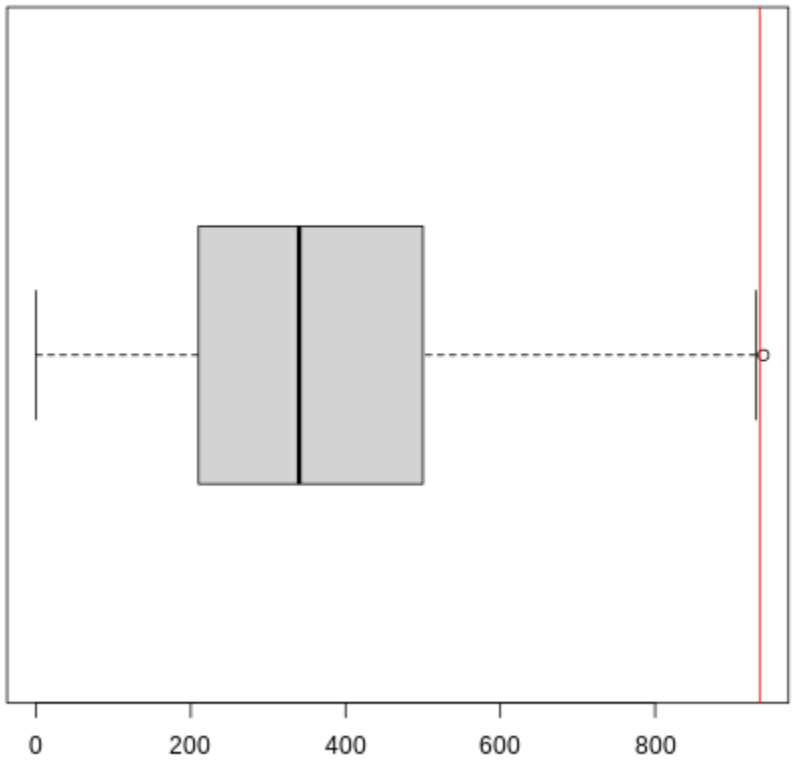
In [ ]:
```
%%R

# Elegimos las variables que queremos analizar
calories <- "Calories"
proteins <- "Protein"
```

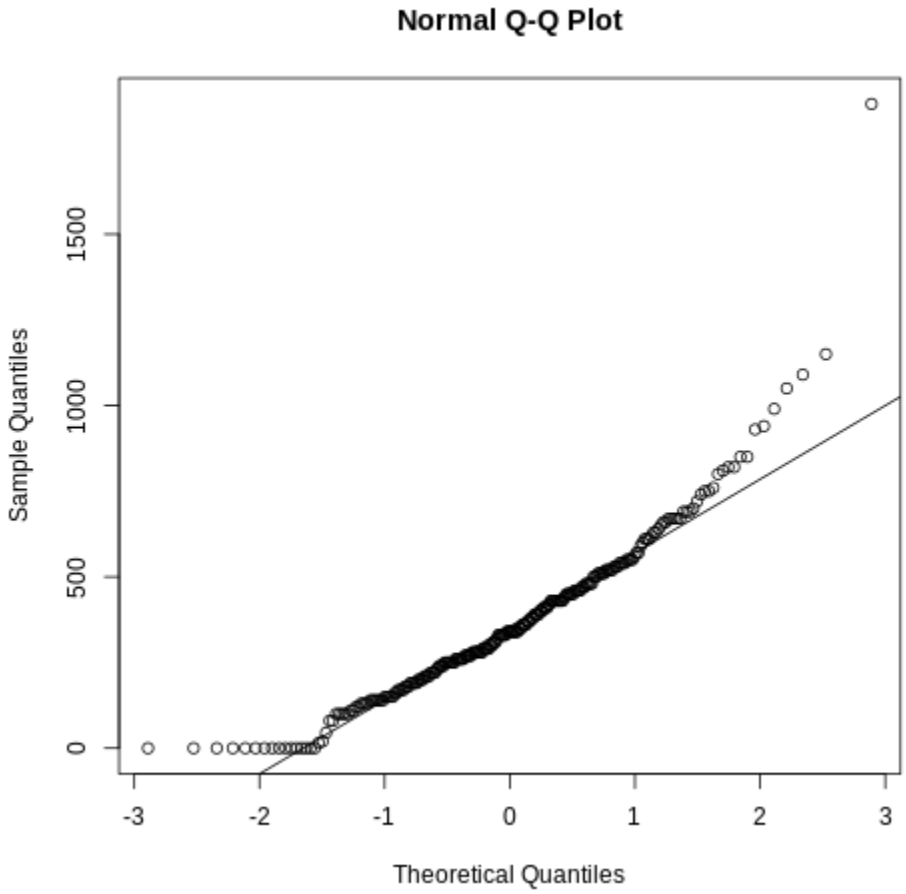
Calorías

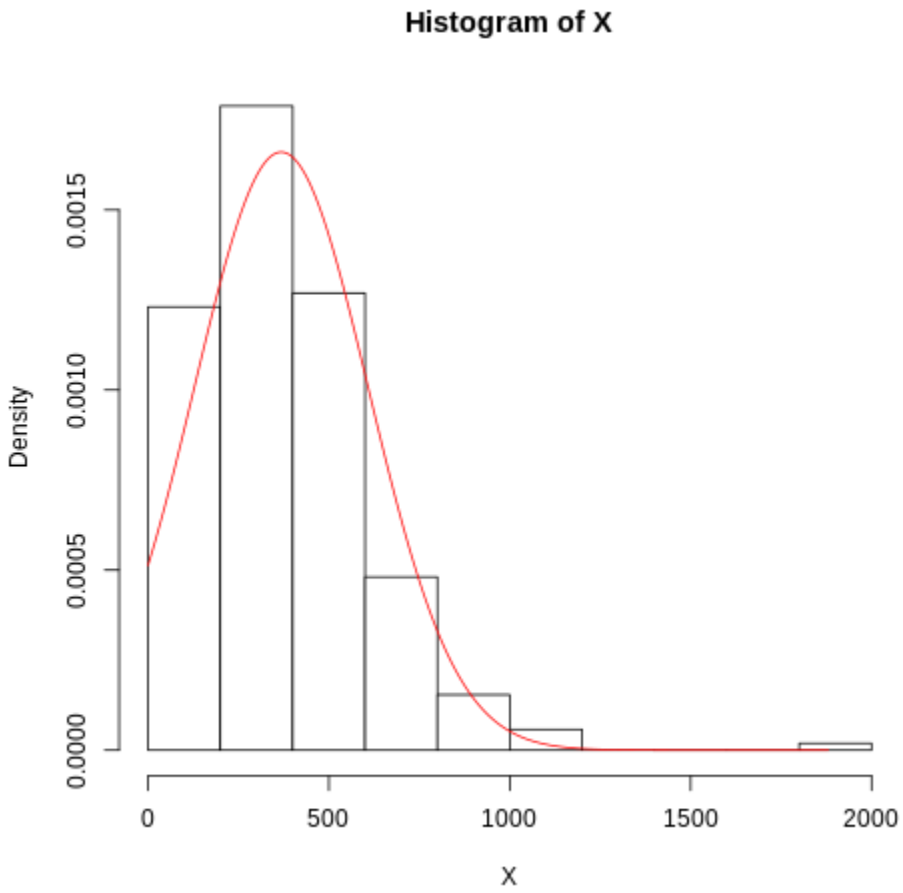
In [ ]:
```
%%R
```

```r
# Extraemos la variable de calorias de la data
X <- M[[calories]]

# Calculamos q1 y q3 para el análisis de los valores atípicos
q1 <- quantile(X, 0.25)
q3 <- quantile(X, 0.75)
ri <- q3 - q1

# Boxplot con valores atípicos marcados
boxplot(X, horizontal=TRUE, ylim=c(min(X), q3 + 1.5 * ri))
abline(v=q3 + 1.5 * ri, col="red")

# Removemos los valores atípicos
X1 <- X[X < q3 + 1.5 * ri]

# Mostramos la data con y sin los valores atípicos
cat("Resumen con datos atípicos:\n")
print(summary(X))
cat("\nResumen sin datos atípicos\n")
print(summary(X1))

# Calculamos skewness y kurtosis
cat("\nSesgo:", skewness(X), "\n")
cat("Curtosis:", kurtosis(X), "\n")

cat("\n\n")

# Generamos la QQ plot, la gráfica de la normal y la gráfica de densidad
qqnorm(X)
qqline(X)
hist(X, prob=TRUE, col=0)
x <- seq(min(X), max(X), 0.1)
y <- dnorm(x, mean(X), sd(X))
lines(x, y, col="red")
```

```
Resumen con datos atípicos:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   210.0   340.0   368.3   500.0  1880.0 

Resumen sin datos atípicos
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   202.5   335.0   349.0   480.0   930.0 

Sesgo: 1.435782
Curtosis: 5.5789
```


0
200
400
600
800


Normal Q-Q Plot
Sample Quantiles
0
500
1000
1500
-3
-2
-1
0
1
2
3
Theoretical Quantiles

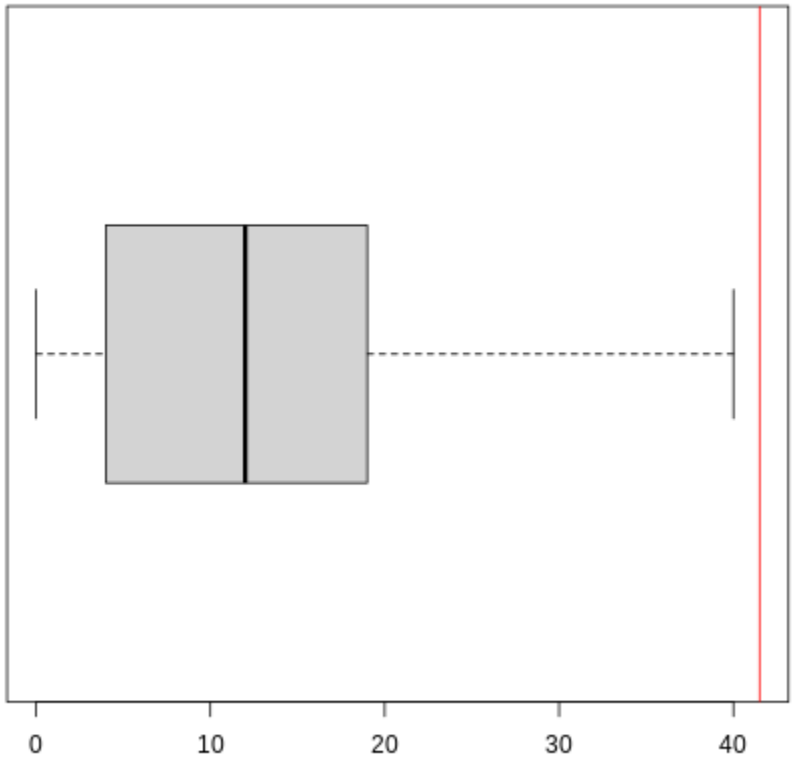


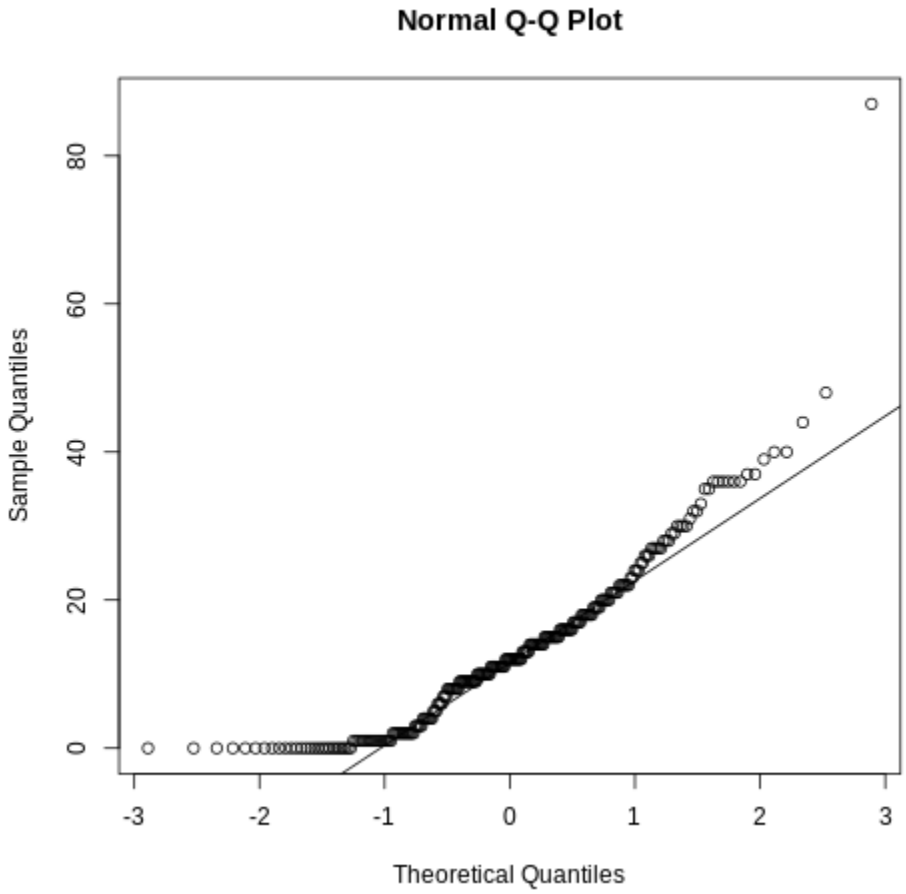
Proteínas

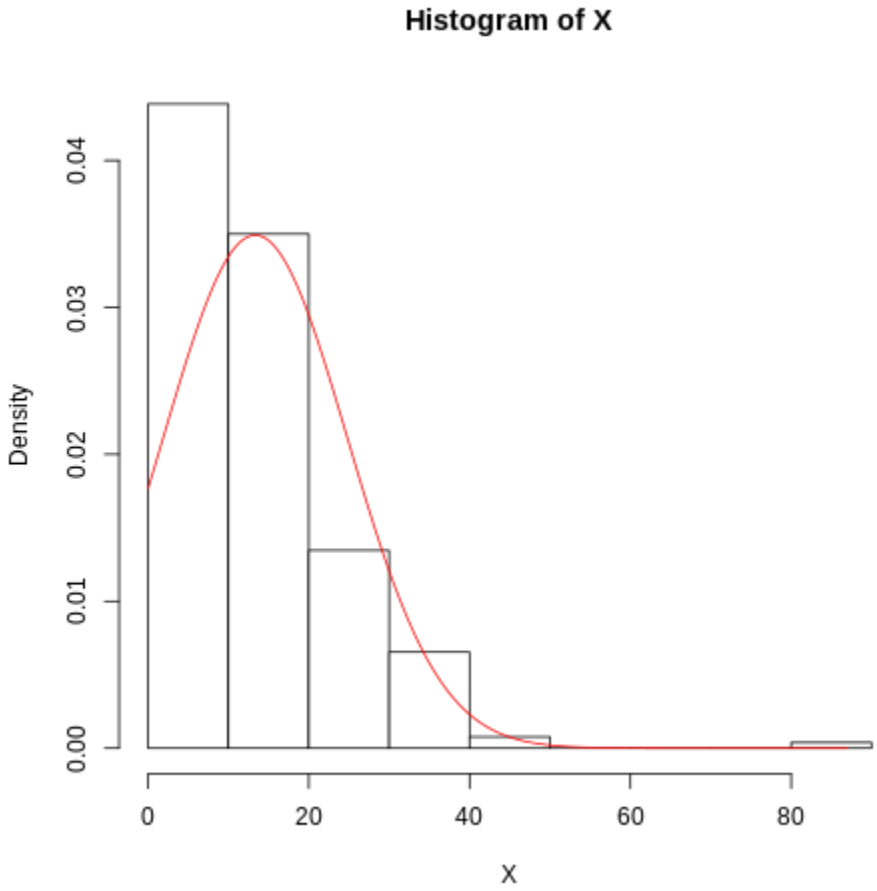
In [ ]:
```
%%R

# Extraemos la variable de calorias de la data
X <- M[[proteins]]

# Calculamos q1 y q3 para el análisis de los valores atípicos
q1 <- quantile(X, 0.25)
q3 <- quantile(X, 0.75)
ri <- q3 - q1

# Boxplot con valores atípicos marcados
boxplot(X, horizontal=TRUE, ylim=c(min(X), q3 + 1.5 * ri))
abline(v=q3 + 1.5 * ri, col="red")

# Removemos los valores atípicos
X1 <- X[X < q3 + 1.5 * ri]

# Mostramos la data con y sin los valores atípicos
cat("Resumen con datos atípicos:\n")
print(summary(X))
cat("\nResumen sin datos atípicos\n")
print(summary(X1))
```

```
# Calculamos skewness y kurtosis
cat("\nSesgo:", skewness(X), "\n")
cat("Curtosis:", kurtosis(X), "\n")

cat("\n\n")

# Generamos la QQ plot, la gráfica de la normal y la gráfica de densidad
qqnorm(X)
qqline(X)
hist(X, prob=TRUE, col=0)
x <- seq(min(X), max(X), 0.1)
y <- dnorm(x, mean(X), sd(X))
lines(x, y, col="red")
```

```
Resumen con datos atípicos:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    4.00   12.00   13.34   19.00   87.00 

Resumen sin datos atípicos
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0     4.0    12.0    12.8    18.0    40.0 

Sesgo: 1.561741
Curtosis: 5.7955
```




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
-3
-2
-1
0
1
2
3



3. Para analizar normalidad se te sugiere:

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

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

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

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

8. Realiza el histograma y su distribución teórica de probabilidad (sugerencia, adapta el código: * hist(datos,freq=FALSE) * lines(density(datos),col="red") * curve(dnorm(x,mean=mean(datos,sd=sd(datos)), from=-6, to=6, add=TRUE, col="blue",lwd=2)

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

4. Para leer los datos de un archivo usa las siguientes instrucciones de R:

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

5. Para explorar y quitar los datos atípicos, usa las siguientes instrucciones de R:

```
q1=quantile(X,0.25)  #Cuantil 1 de la variable X
ri= q3-q1   o   ri=IQR(X)    #Rango intercuartílico de X
par(mfrow=c(2,1)  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE,ylim=c(y1,y2))
abline(v=q3+1.5*ri,col="red")  #linea vertical en el límite de los
datos atípicos o extremos
X1= M[M$X<q3+1.5*ri,c("X")]  #En la matriz M, quitar datos más allá
de 3 rangos intercuartílicos arriba de q3 de la variable X
summary(X1)
summary(X)
```

6. Para realizar el gráfico de densidad de probabilidad y compararla con la de normalidad hipótetica, use los siguientes códigos:

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

7. Para explorar curtosis y sesgo:

```
library(moments): skewness(X) y kurtosis(X)
library e1071: skeness(X) y kurtosis(X)
```