Actividad 4. Explorando bases

Oscar Gutierrez

2024-08-13

Cargar dataset

```
M= read.csv("mc-donalds-menu.csv")
```

Seleccionar las variables

```
calorias= M$Calories
proteinas = M$Protein
```

Analisis de datos atipicos

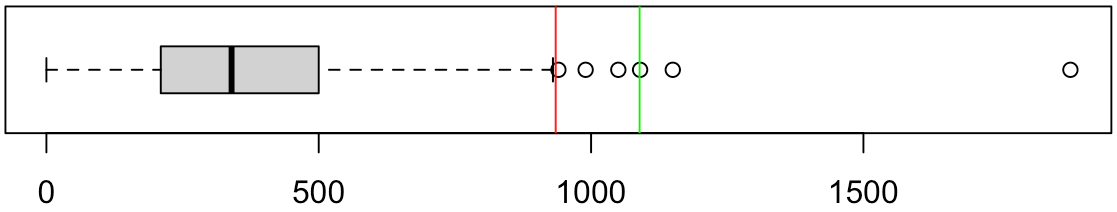
Calorias

```
q1=quantile(calorias, 0.25)
q3 = quantile(calorias, 0.75)
ri=IQR(calorias)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(calorias,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje Y
abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos atipicos o extremos
abline(v= mean(calorias)+ 3*sd(calorias), col="green") # linea vertical a 3 sd de la media
X1= M[M$calorias<q3+1.5*ri ]  #En la matriz M, quita datos más allá de 1.5 rangos interc uartílicos arriba de q3 de la variable X
summary(X1)
```

```
## < table of extent 0 x 0 >
```

```
summary(calorias)
```

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



En este caso, se puede observar que hay varios valores atipicos, sin embargo, es comun que haya ciertos alimentos que tengan un alto valor calorico por lo que no es encesario removerlos. Hay datos que estan despues de 3 desviaciones estandar de la media, por lo que estos serian valores extremos.

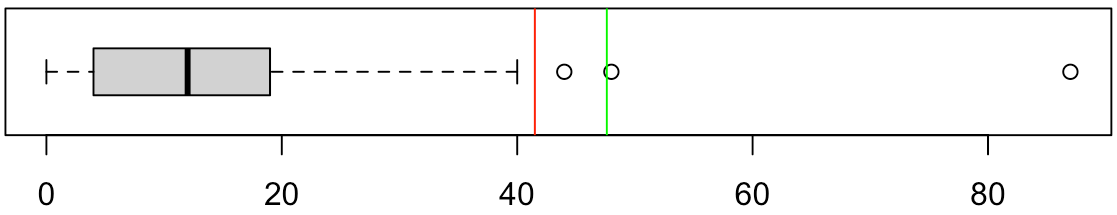
Proteinas

```
q1=quantile(proteinas, 0.25)
q3 = quantile(proteinas, 0.75)
ri=IQR(proteinas)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(proteinas,horizontal=TRUE)   #y1=min en la escala del eje Y, y2=máx en la escala del eje Y
abline(v=q3+1.5*ri, col="red")  #linea vertical en el límite de los datos atípicos o extremos
abline(v= mean(proteinas)+ 3*sd(proteinas), col="green") # linea vertical a 3 sd de la media
X1= M[M$proteinas<q3+1.5*ri ]  #En la matriz M, quita datos más allá de 1.5 rangos intercuartílicos arriba de q3 de la variable X
summary(X1)
```

```
## < table of extent 0 x 0 >
```

```
summary(proteinas)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    4.00   12.00   13.34   19.00   87.00
```



Al igual que con las calorias, no es necesario remover los valores atipicos. En este caso también hay valores que se encuentran a 3 desviaciones estandar de la media, estos serian considerados extremos.

Pruebas de normalidad

###Prueba Anderson Darling para las calorias h0: Los datos siguen una distribucion normal h1: Los datos no siguen una distribucion normal

```
library(nortest)
ad.test(calorias)
```

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

El valor p es muy pequeño, por lo cual los datos no siguen una distribución normal

Prueba Anderson Darling para proteinas

h0: Los datos siguen una distribucion normal h1: Los datos no siguen una distribucion normal

```
library(nortest)
ad.test(proteinas)
```

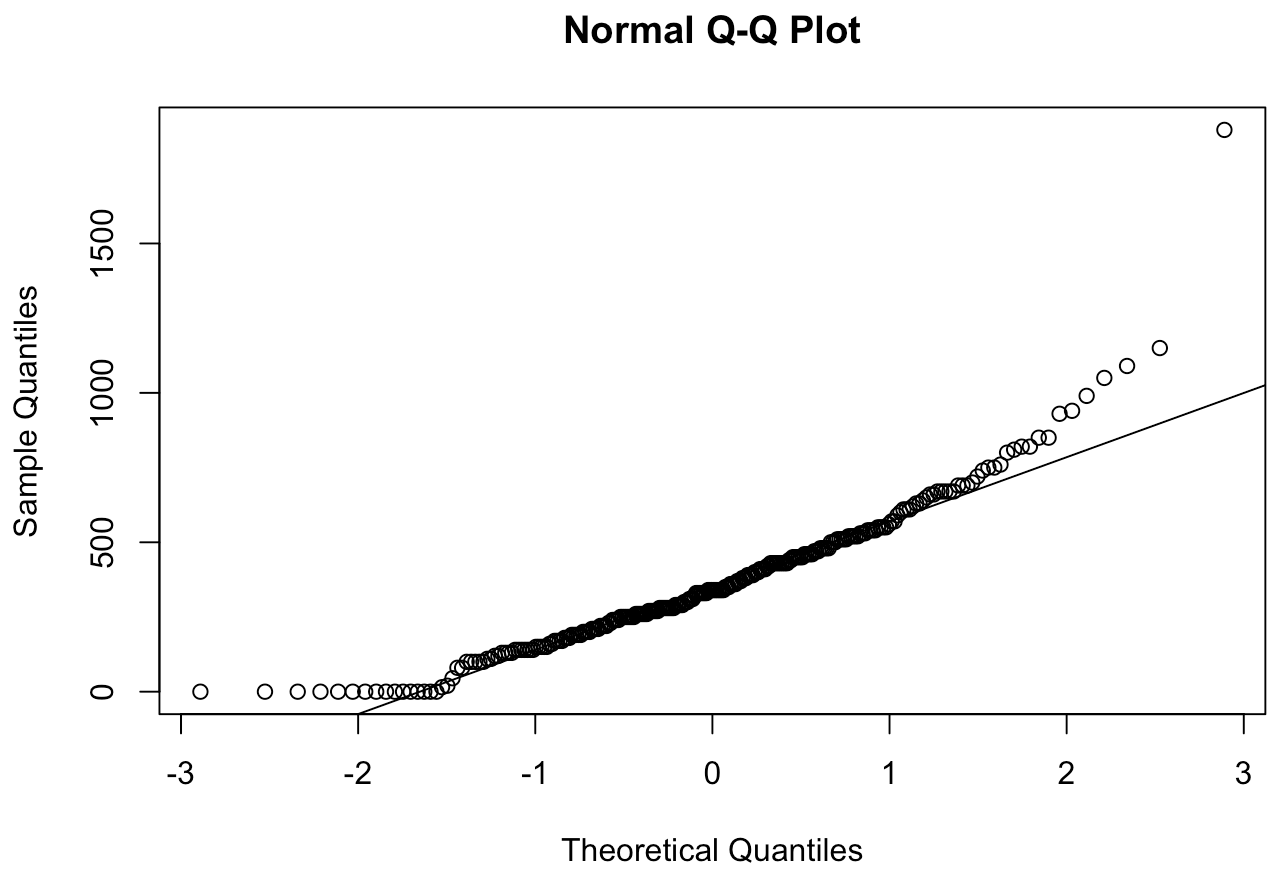
```
##
##  Anderson-Darling normality test
##
## data:  proteinas
## A = 4.7515, p-value = 8.515e-12
```

El valor p es muy pequeño, por lo cual las proteinas no siguen una distribución normal.

QQ plots

QQ plot para calorias

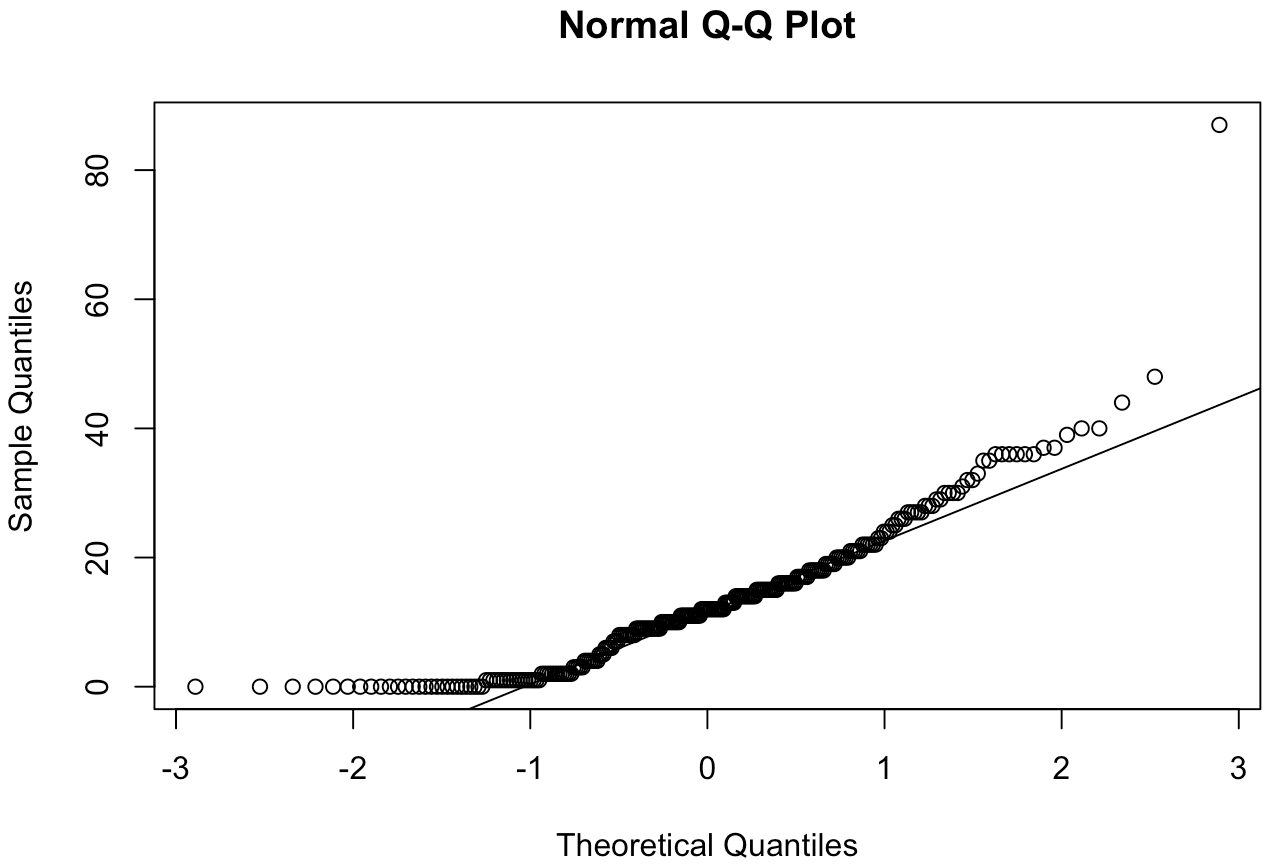
```
qqnorm(calorias)
qqline(calorias)
```



Los datos no siguen la normalidad en las colas.

###QQ plots para proteinas

```
qqnorm(proteinas)
qqline(proteinas)
```



Los datos no cuentan con distribución normal debido a que que no siguen los cuantiles teoricos.

Coeficientes de sesgo y curtosis

Coeficientes para calorias

```
library(moments)
cat("sesgo= ",skewness(calorias), "\ncurtosis=",kurtosis(calorias))
```

```
## sesgo=  1.444105
## curtosis= 8.645274
```

Coeficientes para proteinas

```
cat("sesgo= ",skewness(proteinas), "\ncurtosis=",kurtosis(proteinas))
```

```
## sesgo=  1.570794
## curtosis= 8.86355
```

Los coeficientes de sesgo son muy diferentes a 0 y los de curtosis son muy diferentes de 3, por lo que no concuerdan con una distribucion normal.

Calculo de media, mediana y rango medio

Calorias

```
cat("media=", mean(calorias), "\nmediana=",median(calorias), "\nrango medio=", (max(calo
rias)-min(calorias))/2)
```

```
## media= 368.2692
## mediana= 340
## rango medio= 940
```

Proteinas

```
cat("media=", mean(proteinas), "\nmediana=",median(proteinas), "\nrango medio=", (max(pr
oteinas)-min(proteinas))/2)
```

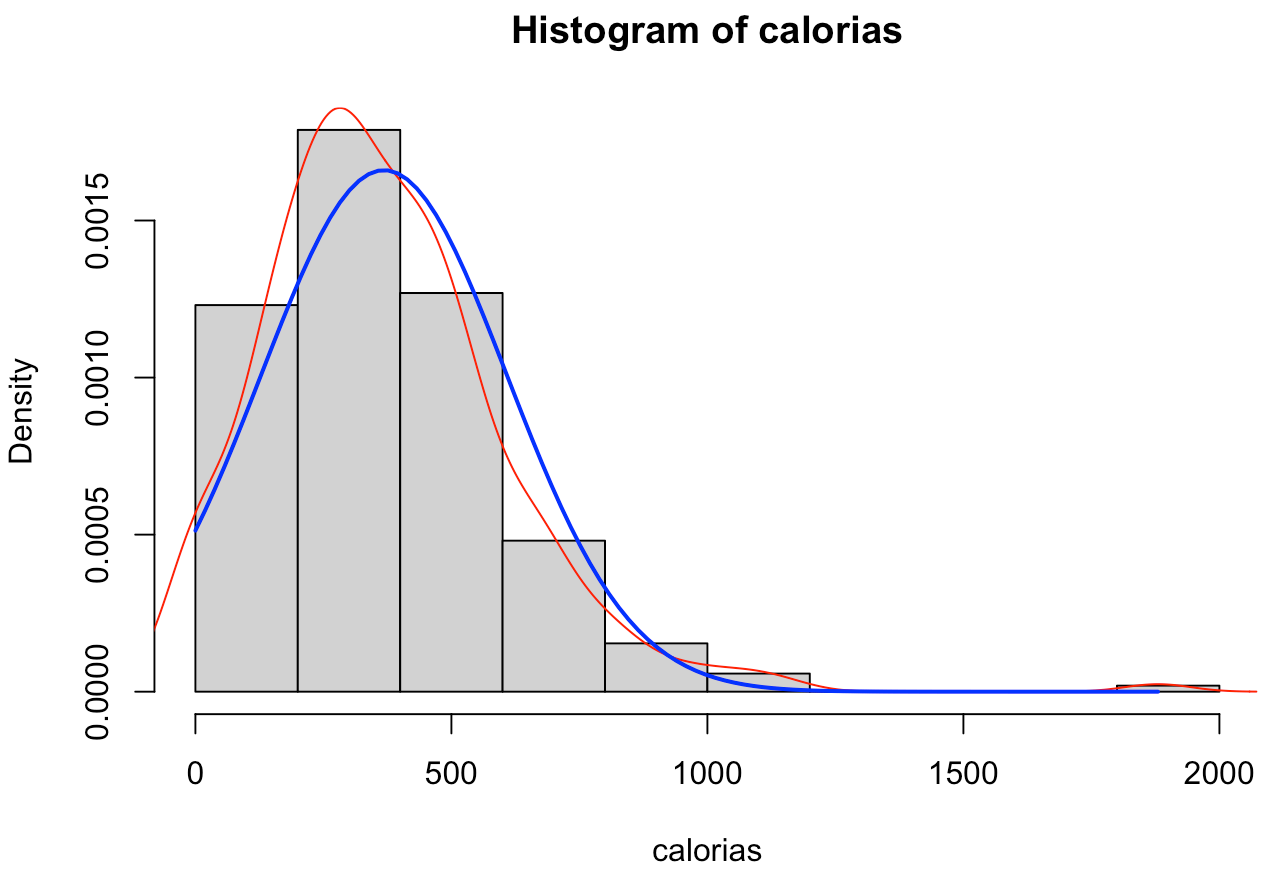
```
## media= 13.33846
## mediana= 12
## rango medio= 43.5
```

En una distribucion normal estos 3 valores deberian ser iguales.

Histograma y distribucion teorica

Grafica para calorias

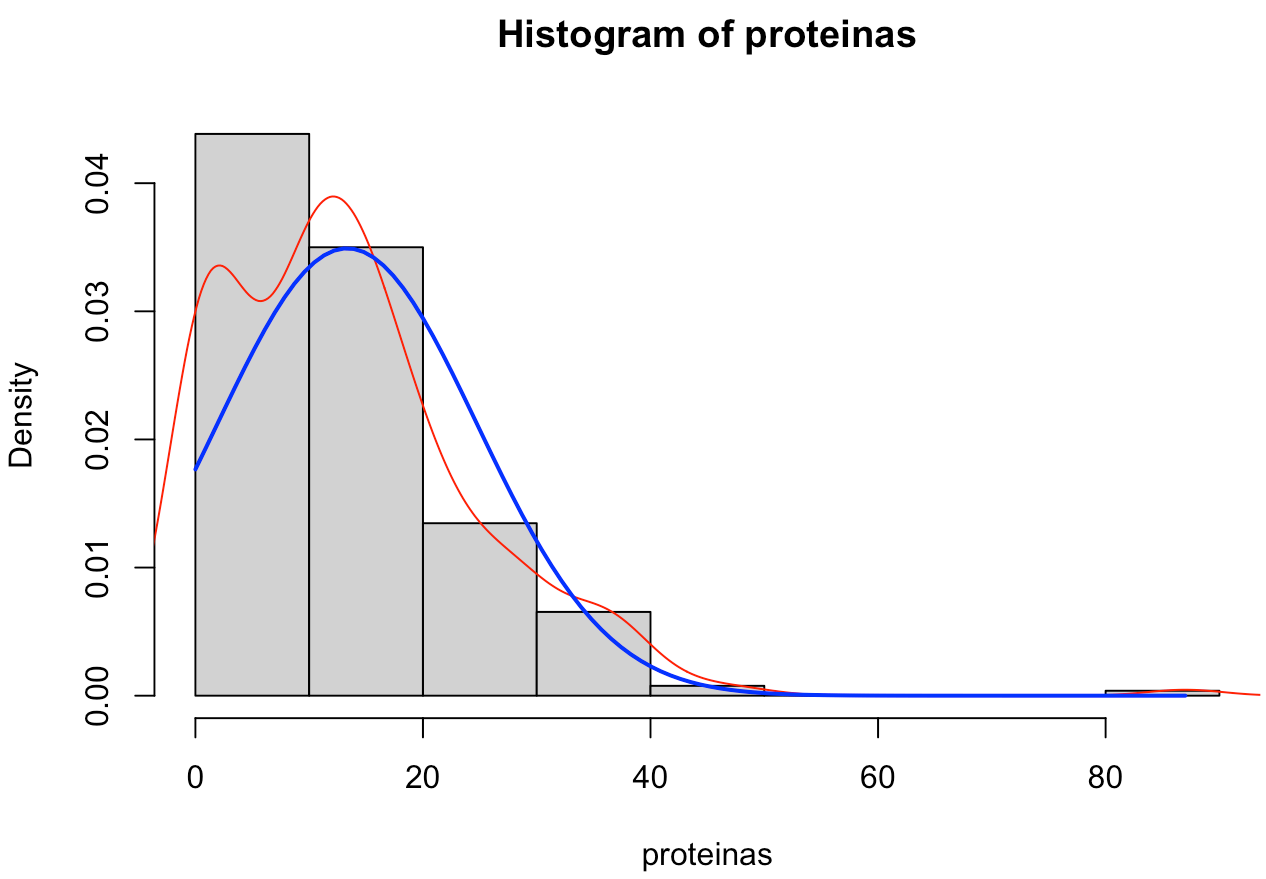
```
hist(calorias,freq=FALSE)
lines(density(calorias),col="red")
curve(dnorm(x,mean=mean(calorias),sd=sd(calorias)), from=min(calorias), to= max(caloria
s), add=TRUE, col="blue",lwd=2)
```



###

Grafica para proteinas

```
hist(proteinas,freq=FALSE)
lines(density(proteinas),col="red")
curve(dnorm(x,mean=mean(proteinas),sd=sd(proteinas)), from=min(proteinas), to= max(prote
inas), add=TRUE, col="blue",lwd=2)
```



Como se puede observar, las distribuciones de probabilidad no concuerdan con una distribucion normal teorica.

Los datos atipicos pueden influir en la normalidad debido a que pueden generar sesgo.