2. Explorando bases

Diego Rodríguez A00829925

2023-08-17

Lectura de datos y limpieza

Primero tenemos que leer la base de datos y revisar las variables que contiene.

```
M = read.csv("mc-donalds-menu-1.csv")
summary(M)
```

```
##    Category             Item           Serving.Size          Calories
##  Length:260         Length:260         Length:260         Min.   :   0.0
##  Class :character   Class :character   Class :character   1st Qu.: 210.0
##  Mode  :character   Mode  :character   Mode  :character   Median : 340.0
##                                                           Mean   : 368.3
##                                                           3rd Qu.: 500.0
##                                                           Max.   :1880.0
##  Calories.from.Fat   Total.Fat       Total.Fat....Daily.Value. Saturated.Fat
##  Min.   :   0.0    Min.   :  0.000   Min.   :  0.00            Min.   : 0.000
##  1st Qu.:  20.0    1st Qu.:  2.375   1st Qu.:  3.75            1st Qu.: 1.000
##  Median : 100.0    Median : 11.000   Median : 17.00            Median : 5.000
##  Mean   : 127.1    Mean   : 14.165   Mean   : 21.82            Mean   : 6.008
##  3rd Qu.: 200.0    3rd Qu.: 22.250   3rd Qu.: 35.00            3rd Qu.:10.000
##  Max.   :1060.0    Max.   :118.000   Max.   :182.00            Max.   :20.000
##  Saturated.Fat....Daily.Value.   Trans.Fat       Cholesterol
##  Min.   :  0.00                Min.   :0.0000   Min.   :  0.00
##  1st Qu.:  4.75                1st Qu.:0.0000   1st Qu.:  5.00
##  Median : 24.00                Median :0.0000   Median : 35.00
##  Mean   : 29.97                Mean   :0.2038   Mean   : 54.94
##  3rd Qu.: 48.00                3rd Qu.:0.0000   3rd Qu.: 65.00
##  Max.   :102.00                Max.   :2.5000   Max.   :575.00
##  Cholesterol....Daily.Value.     Sodium       Sodium....Daily.Value.
##  Min.   :  0.00              Min.   :   0.0   Min.   :  0.00
##  1st Qu.:  2.00              1st Qu.: 107.5   1st Qu.:  4.75
##  Median : 11.00              Median : 190.0   Median :  8.00
##  Mean   : 18.39              Mean   : 495.8   Mean   : 20.68
##  3rd Qu.: 21.25              3rd Qu.: 865.0   3rd Qu.: 36.25
##  Max.   :192.00              Max.   :3600.0   Max.   :150.00
##  Carbohydrates    Carbohydrates....Daily.Value. Dietary.Fiber
##  Min.   :  0.00   Min.   : 0.00                 Min.   :0.000
##  1st Qu.: 30.00   1st Qu.:10.00                 1st Qu.:0.000
##  Median : 44.00   Median :15.00                 Median :1.000
##  Mean   : 47.35   Mean   :15.78                 Mean   :1.631
##  3rd Qu.: 60.00   3rd Qu.:20.00                 3rd Qu.:3.000
```

```
##  Max.   :141.00   Max.   :47.00                Max.   :7.000
##  Dietary.Fiber....Daily.Value.     Sugars          Protein
##  Min.   : 0.000                Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 0.000                1st Qu.:  5.75   1st Qu.: 4.00
##  Median : 5.000                Median : 17.50   Median :12.00
##  Mean   : 6.531                Mean   : 29.42   Mean   :13.34
##  3rd Qu.:10.000                3rd Qu.: 48.00   3rd Qu.:19.00
##  Max.   :28.000                Max.   :128.00   Max.   :87.00
##  Vitamin.A....Daily.Value. Vitamin.C....Daily.Value. Calcium....Daily.Value.
##  Min.   :  0.00            Min.   :  0.000           Min.   : 0.00
##  1st Qu.:  2.00            1st Qu.:  0.000           1st Qu.: 6.00
##  Median :  8.00            Median :  0.000           Median :20.00
##  Mean   : 13.43            Mean   :  8.535           Mean   :20.97
##  3rd Qu.: 15.00            3rd Qu.:  4.000           3rd Qu.:30.00
##  Max.   :170.00            Max.   :240.000           Max.   :70.00
##  Iron....Daily.Value.
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 4.000
##  Mean   : 7.735
##  3rd Qu.:15.000
##  Max.   :40.000
```

Una vez ya conocemos todas sus variables y los tipos de datos que son, se seleccionarán 2 de ella. Para esta actividad se ha decidido trabajar con la variable Cholesterol y con Protein.

Valores nulos

Se revisaran si estas variables cuentan con valores nulos:

```
sum(is.na(M$Protein))
```

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

```
sum(is.na(M$Cholesterol))
```

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

Afortunadamente ninguna de ellas tiene valores faltantes. A continuación se revisarán datos atípicos y se decidirá qué hacer con ellos.

Valores atípicos

```
prot = M$Protein

q1 = quantile(prot, probs = 0.25)
q3 = quantile(prot, probs = 0.75)
print(q1)
```

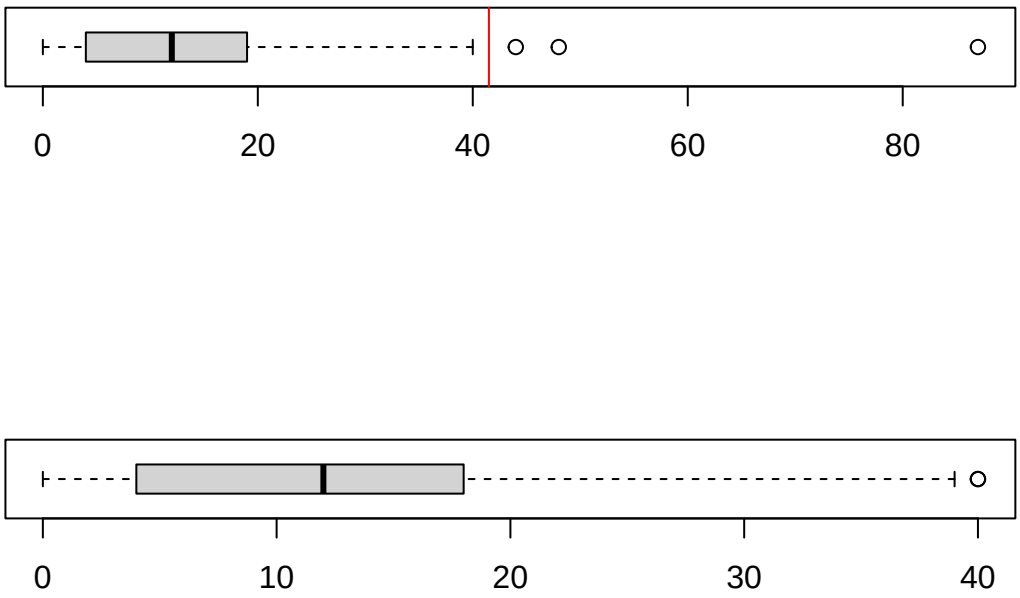
```
## 25%
##   4
```

```
print(q3)
```

```
## 75%
##  19
```

```
ri=IQR(prot)
par(mfrow=c(2,1))
boxplot(prot, horizontal=TRUE)
abline(v=q3+1.5*ri, col="red")


X1= M[M$Protein<q3+1.5*ri,c("Protein")]
boxplot(X1, horizontal=TRUE)
```




```
summary(X1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     4.0    12.0    12.8    18.0    40.0
```

En el primer gráfico se puede apreciar como existen algunos valores atípicos, pero estos son eliminados del dataset bajo la condición de no exceder el cuartil 3 más 1.5 veces el rango intercuartil. En el segundo boxplot se muestra la variable de proteina ya

```r
chol = M$Cholesterol

q1c = quantile(chol, probs = 0.25)
q3c = quantile(chol, probs = 0.75)
print(q1c)
```

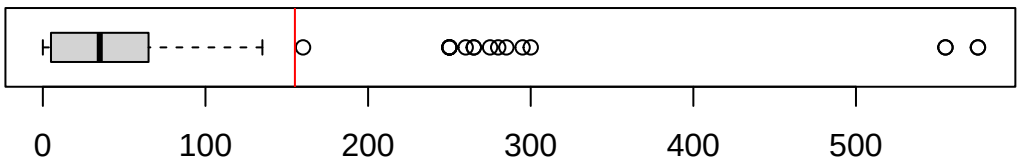
```
## 25%
##   5
```

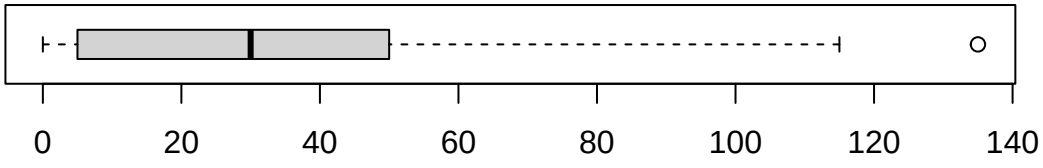
```r
print(q3c)
```

```
## 75%
##  65
```

```r
ric = IQR(chol)
par(mfrow = c(2,1))
boxplot(chol, horizontal=TRUE)
abline(v = q3c+1.5*ric, col="red")


X2= M[M$Cholesterol < q3c + 1.5*ric, c("Cholesterol")]
boxplot(X2, horizontal=TRUE)
```





```
summary(X2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    5.00   30.00   34.67   50.00  135.00
```

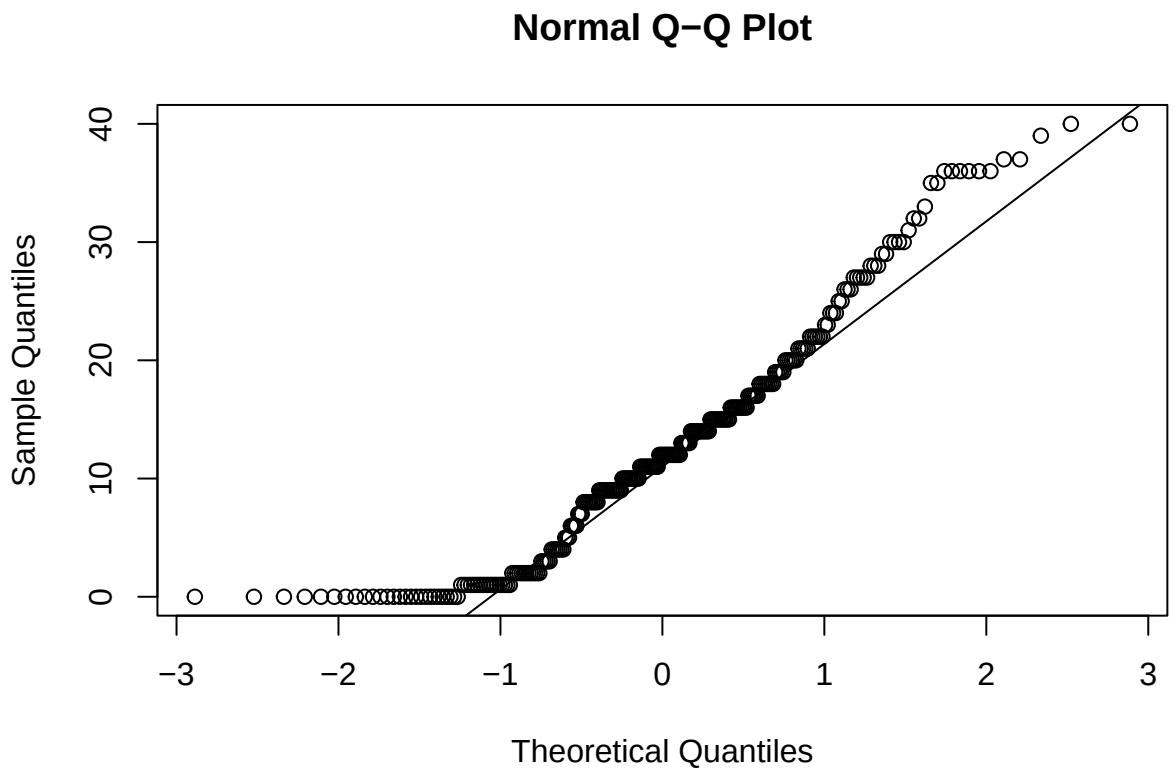
El mismo proceso es ejecutado en la variable de Colesterol donde se puede observar que existían más valores atípicos y también fueron removidos.

Análisis de normalidad

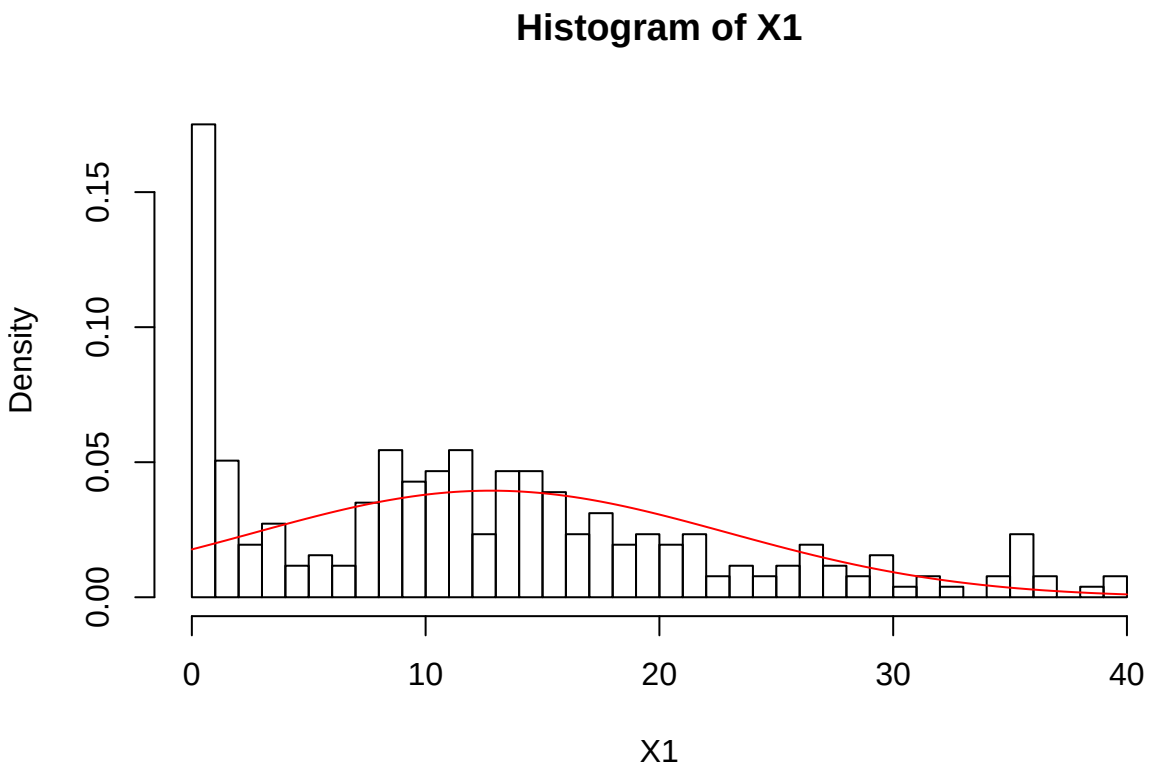
Ahora tenemos las dos variables limpias para probar su normalidad. Primero realizaremos un gráfico de densidad y las compararemos con la normalidad hipotética.

Análisis de normalidad de proteína

```
qqnorm(X1)
qqline(X1)
```



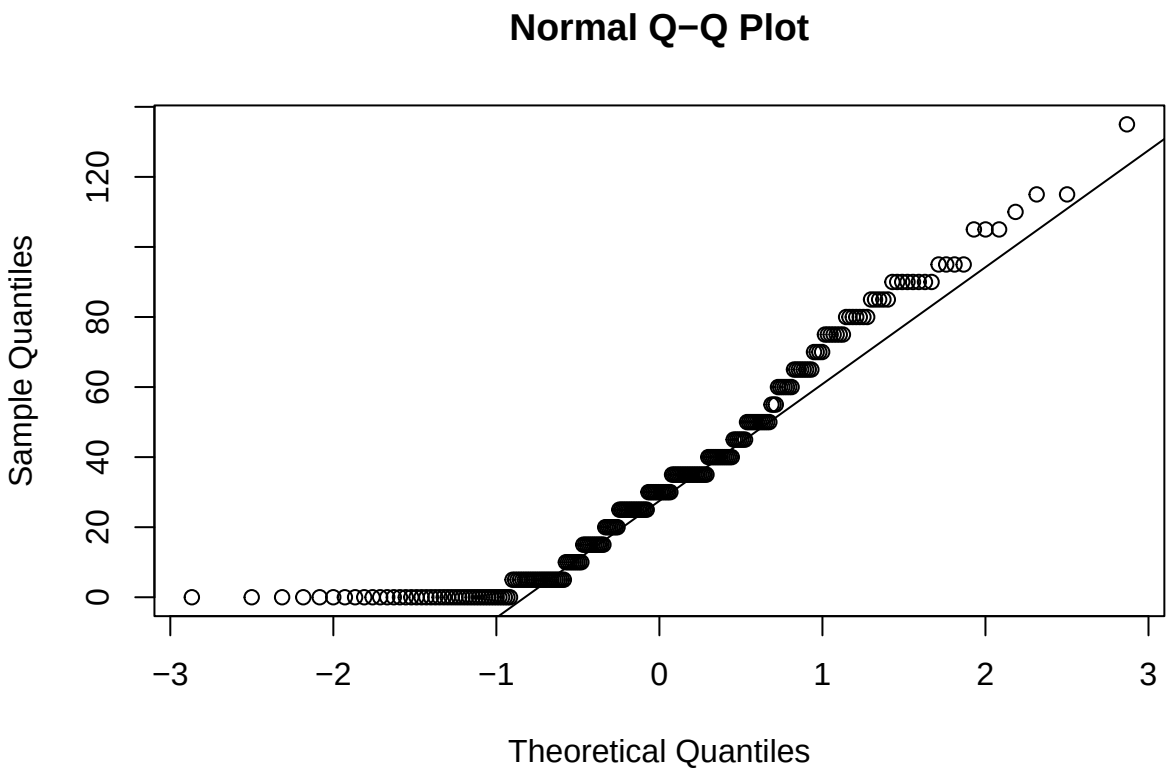

```
hist(X1, prob = TRUE, col = 0, breaks = 50)
x = seq(min(X1), max(X1), 0.1)
y = dnorm(x, mean(X1), sd(X1))
lines(x, y, col = "red")
```



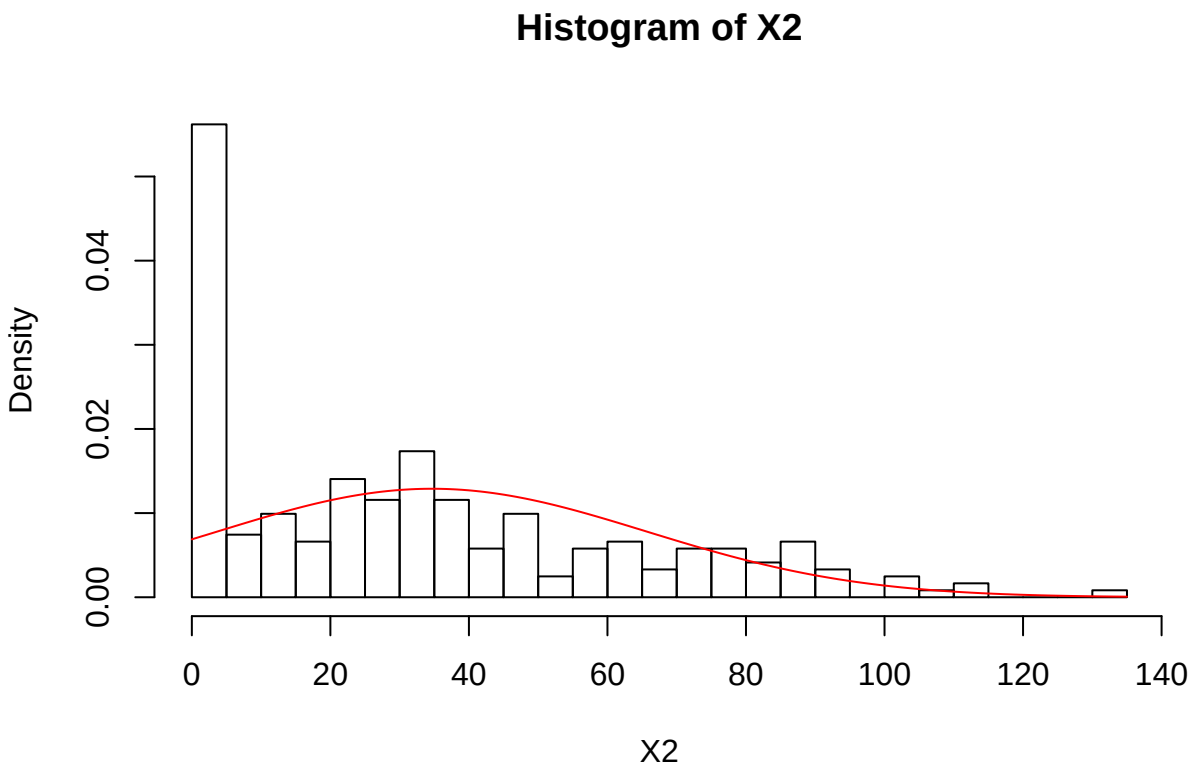
De ambas gráficas podemos apreciar que alrededor de la media los datos parecen comportarse de manera muy similar a la normal, sin embargo, cuando se observan las colas tanto la derecha como izquierda los datos se dispersan y se alejan de la normalidad.

Análisis de normalidad de Cholesterol

```
qqnorm(X2)
qqline(X2)
```




```
hist(X2, prob = TRUE, col = 0, breaks = 20)
x2 = seq(min(X2), max(X2), 0.1)
y2 = dnorm(x2, mean(X2), sd(X2))
lines(x2, y2, col = "red")
```



La variable de cholesterol tiene un comportamiento similar a la de proteína en donde al rededor de la media se asemeja a la normal, pero en las colas se distancia de ella. A pesar de ello, luce mejor que la de proteína.

Ahora calcualremos su curtosis y sesgo para determinar si las consideramos normales.

```
library(moments)

cat("Momentos de la Proteína \n")
```

```
## Momentos de la Proteína
```

```
cat("Sesgo de la proteína: ", skewness(X1), "\n")
```

```
## Sesgo de la proteína:  0.6742646
```

```
cat("Kurtosis de la proteína: ", kurtosis(X1), "\n", "\n")
```

```
## Kurtosis de la proteína:  2.83266
##
```

```
cat("Momentos del Cholesterol \n ")
```

```
## Momentos del Cholesterol
##
```

```r
cat("Sesgo del cholesterol: ", skewness(X2), "\n")
```

```
## Sesgo del cholesterol:  0.7525343
```

```r
cat("Kurtosis del cholesterol: ", kurtosis(X2), "\n")
```

```
## Kurtosis del cholesterol:  2.741518
```

El sesgo de una normal es 0 y la kurtosis es de 3, los valors obtenidos se acercan a estos pero no me convencen para declarar que las variables se comporten de manera normal. En lugar de ello, haremos una prueba shapiro aprovechando que contamos con más de 30 valores para ambas.

```r
cat("Shapiro test de la Proteína \n")
```

```
## Shapiro test de la Proteína
```

```r
shapiro.test(X1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X1
## W = 0.93388, p-value = 2.545e-09
```

```r
cat("Shapiro test del Cholesterol \n ")
```

```
## Shapiro test del Cholesterol
##
```

```r
shapiro.test(X2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X2
## W = 0.91049, p-value = 7.384e-11
```

En ambos casos, los valores p son extremadamente pequeños (en notación científica, con exponente negativo), lo que indica que hay evidencia significativa para rechazar la hipótesis nula de que los datos siguen una distribución normal. En otras palabras, los datos de ambas variables "Proteína" y "Cholesterol" no se ajustan a una distribución normal.

Dado que los valores p son mucho menores que un nivel de significancia típico (como 0.05 o 0.01), se rechazaría la hipótesis nula en ambos casos. Esto sugiere que las variables no pueden considerarse como provenientes de una distribución normal.