Explorando bases

AUTHOR
Alfredo García A00830952

Leyendo los datos

```
M = read.csv('mc-donalds-menu-1.csv')
head(M)
```

```
   Category                           Item   Serving.Size Calories
1 Breakfast                    Egg McMuffin 4.8 oz (136 g)      300
2 Breakfast                Egg White Delight 4.8 oz (135 g)      250
3 Breakfast                 Sausage McMuffin 3.9 oz (111 g)      370
4 Breakfast        Sausage McMuffin with Egg 5.7 oz (161 g)      450
5 Breakfast Sausage McMuffin with Egg Whites 5.7 oz (161 g)      400
6 Breakfast             Steak & Egg McMuffin 6.5 oz (185 g)      430
  Calories.from.Fat Total.Fat Total.Fat....Daily.Value. Saturated.Fat
1               120        13                        20             5
2                70         8                        12             3
3               200        23                        35             8
4               250        28                        43            10
5               210        23                        35             8
6               210        23                        36             9
  Saturated.Fat....Daily.Value. Trans.Fat Cholesterol
1                            25         0         260
2                            15         0          25
3                            42         0          45
4                            52         0         285
5                            42         0          50
6                            46         1         300
  Cholesterol....Daily.Value. Sodium Sodium....Daily.Value. Carbohydrates
1                          87    750                    31            31
2                           8    770                    32            30
3                          15    780                    33            29
4                          95    860                    36            30
5                          16    880                    37            30
6                         100    960                    40            31
  Carbohydrates....Daily.Value. Dietary.Fiber Dietary.Fiber....Daily.Value.
1                            10             4                            17
2                            10             4                            17
3                            10             4                            17
4                            10             4                            17
5                            10             4                            17
6                            10             4                            18
  Sugars Protein Vitamin.A....Daily.Value. Vitamin.C....Daily.Value.
1      3      17                        10                         0
```

```
2      3      18                       6                       0
3      2      14                       8                       0
4      2      21                      15                       0
5      2      21                       6                       0
6      3      26                      15                       2
  Calcium....Daily.Value. Iron....Daily.Value.
1                      25                    15
2                      25                     8
3                      25                    10
4                      30                    15
5                      25                    10
6                      30                    20
```

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

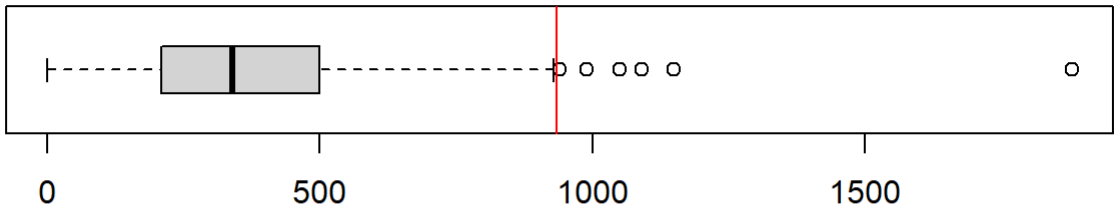
```
X = M$Calories

q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3 = quantile(X,0.75)
ri=IQR(X)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE)
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
summary(X1)
```

```
Length  Class   Mode
     0   NULL   NULL
```

```
summary(X)
```

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



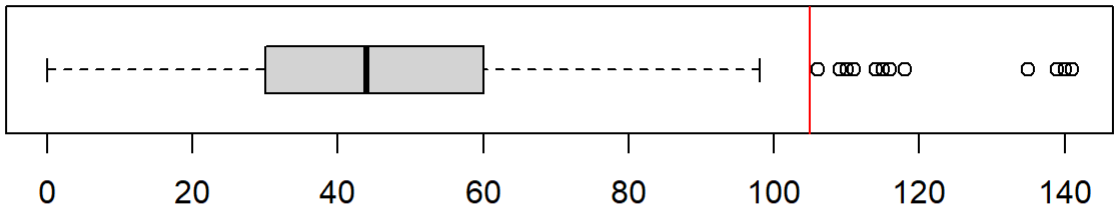

```r
X = M$Carbohydrates

q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3 = quantile(X,0.75)
ri=IQR(X)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE)
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
summary(X1)
```

```
Length  Class   Mode
     0   NULL   NULL
```

```r
summary(X)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   30.00   44.00   47.35   60.00  141.00
```



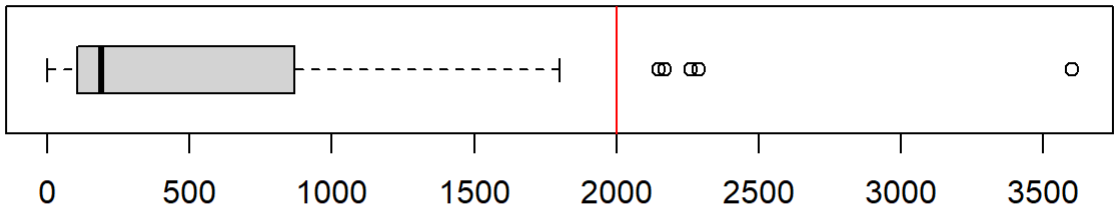

```r
X = M$Sodium

q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3 = quantile(X,0.75)
ri=IQR(X)    #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE)
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
summary(X1)
```

```
Length  Class   Mode
     0   NULL   NULL
```

```r
summary(X)
```

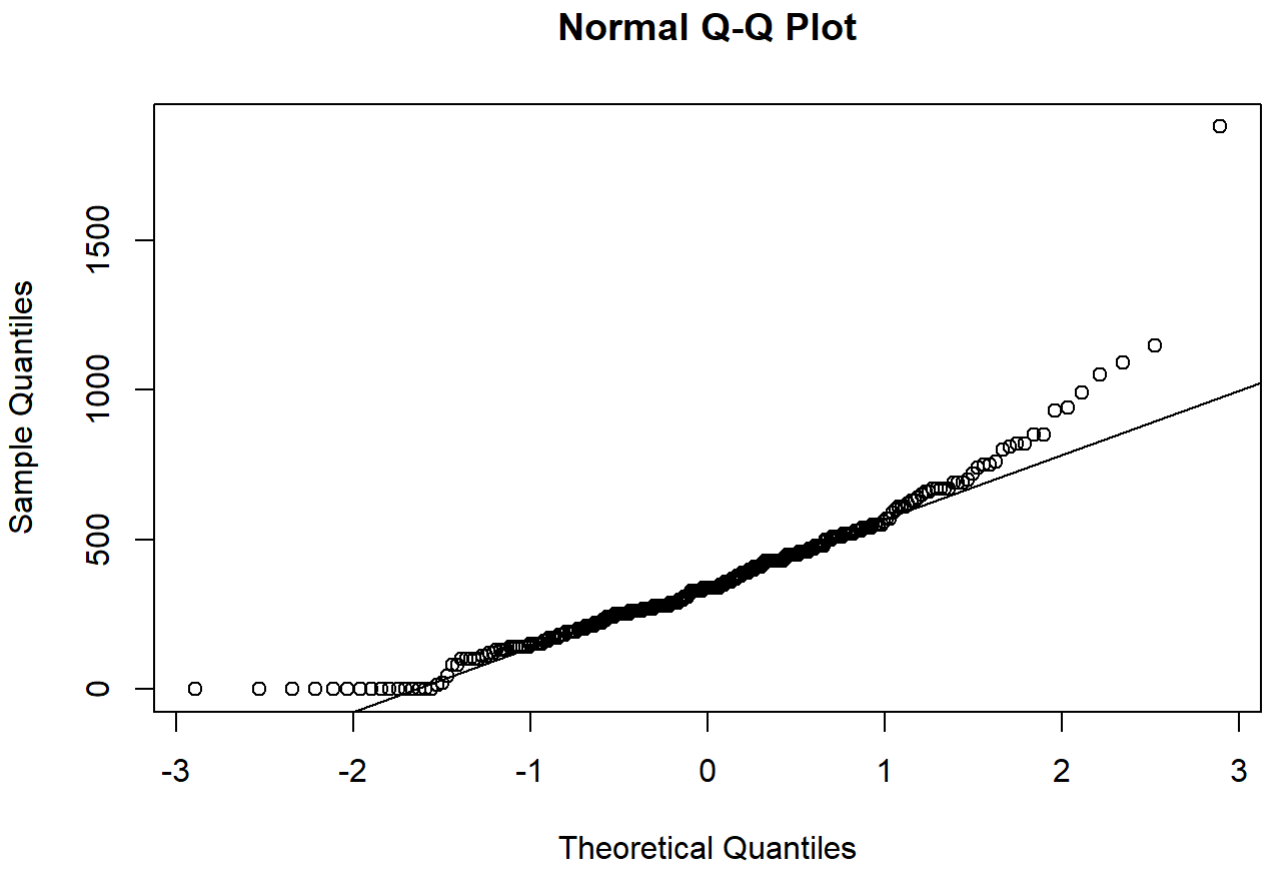
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   107.5   190.0   495.8   865.0  3600.0
```



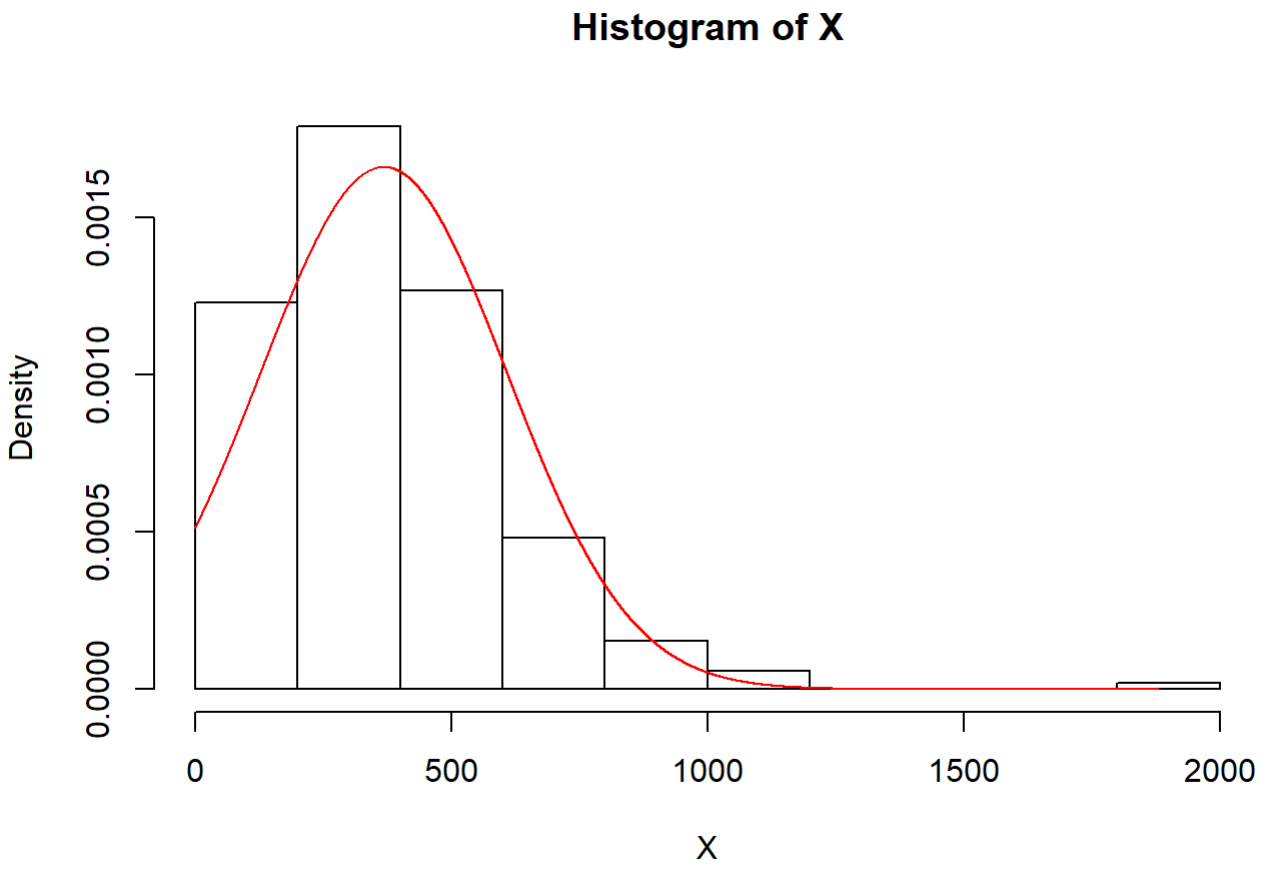
4. Para realizar el gráfico de densidad de probabilidad y compararla con la de normalidad hipótetica, use los siguientes códigos:

```
X = M$Calories

qqnorm(X)
qqline(X)
```

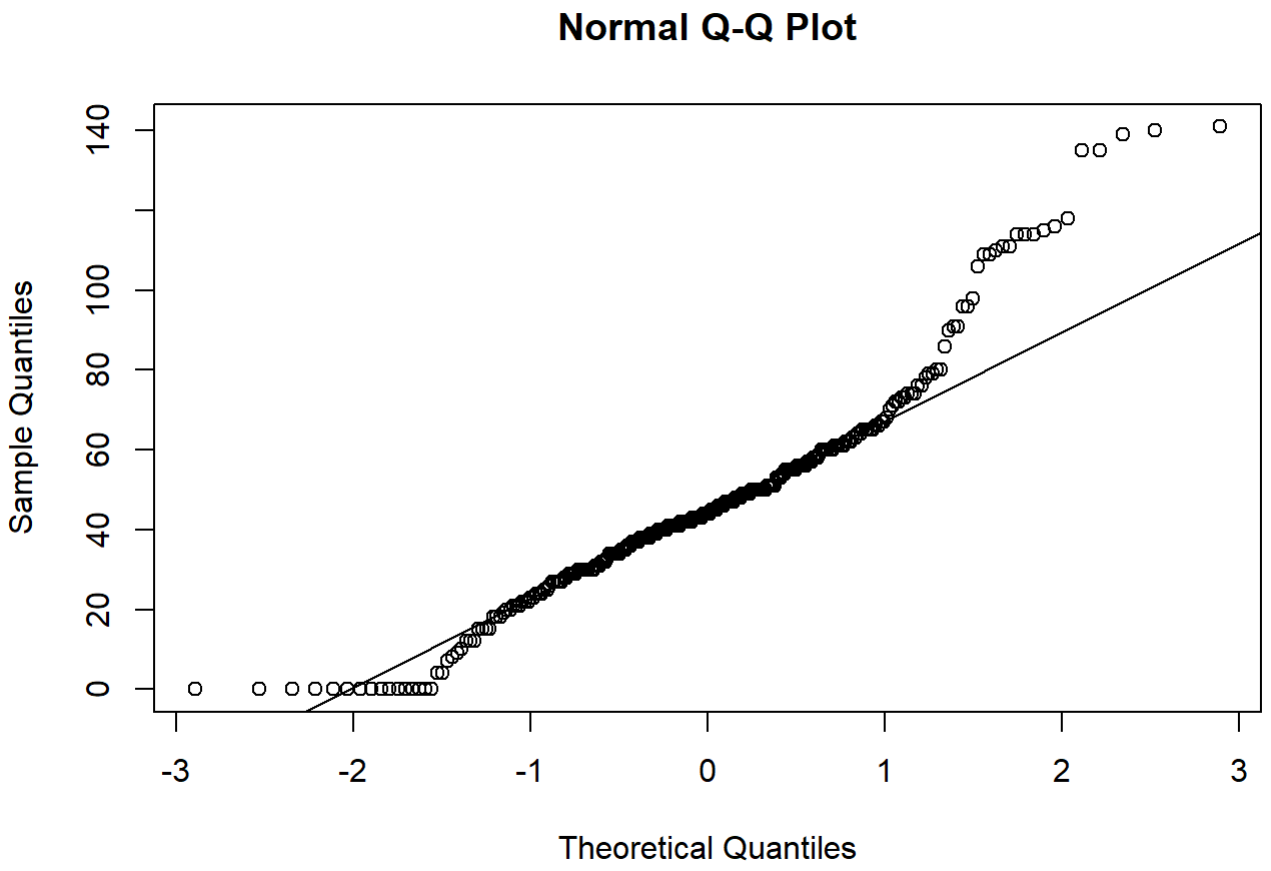



```r
hist(X,prob=TRUE,col=0)
x=seq(min(X),max(X),0.1)
y=dnorm(x,mean(X),sd(X))
lines(x,y,col="red")
```

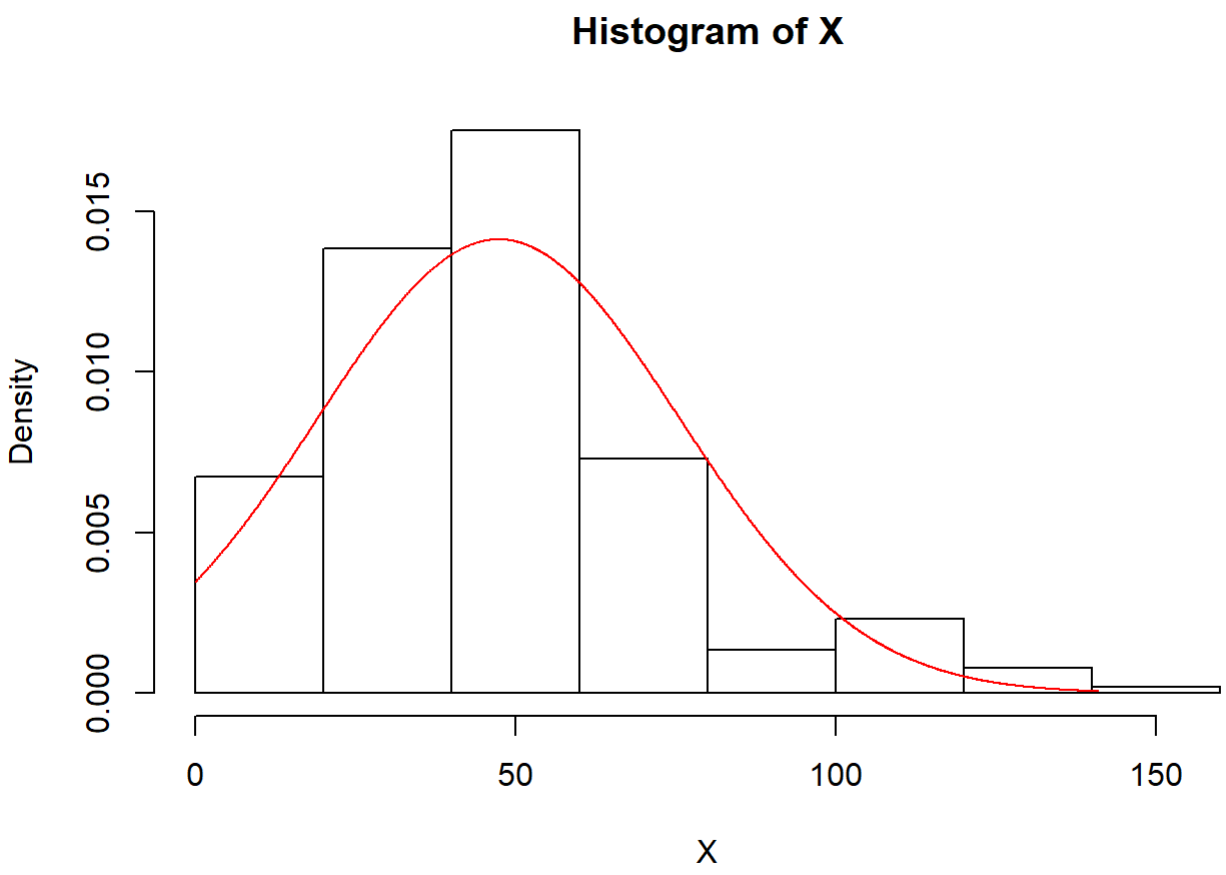



```
X = M$Carbohydrates

qqnorm(X)
qqline(X)
```

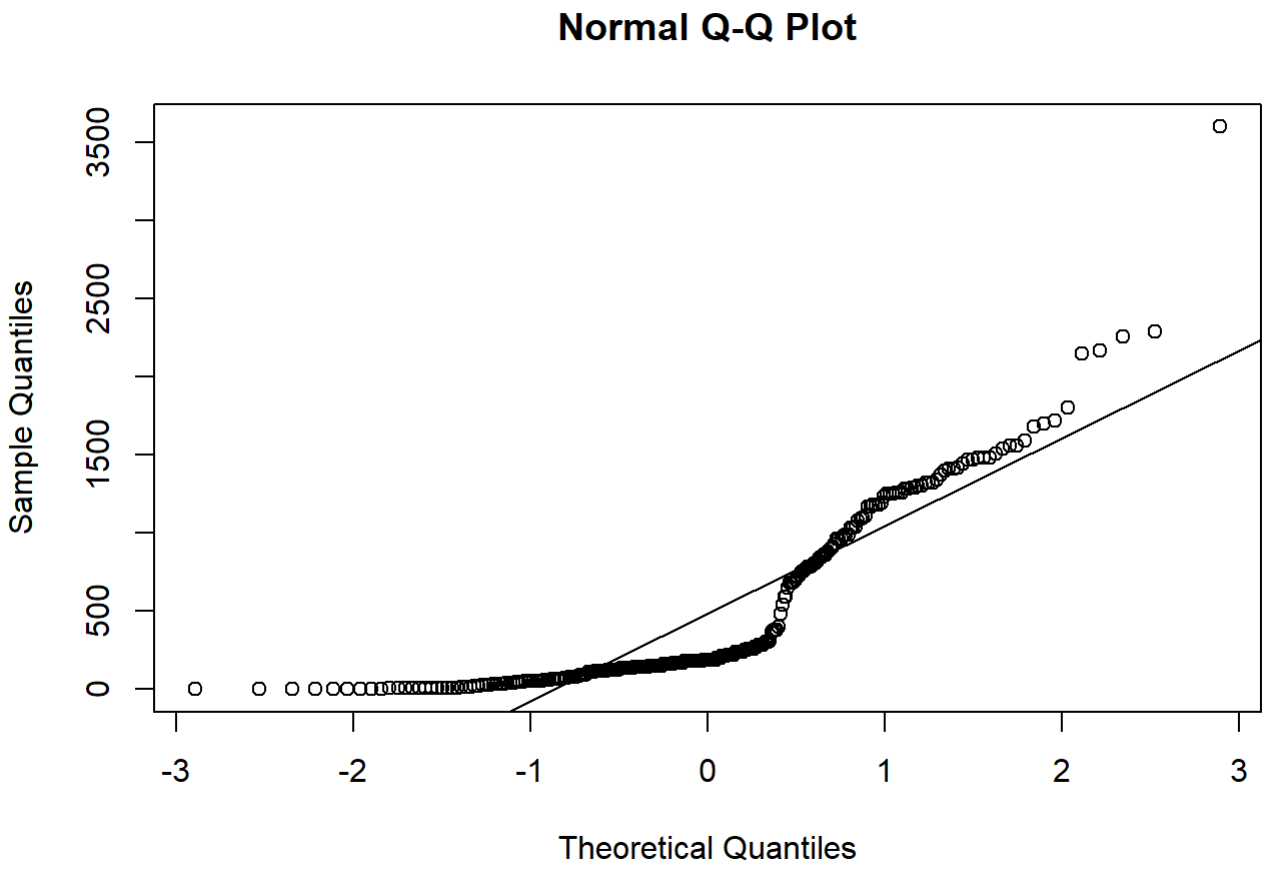



```r
hist(X,prob=TRUE,col=0)
x=seq(min(X),max(X),0.1)
y=dnorm(x,mean(X),sd(X))
lines(x,y,col="red")
```

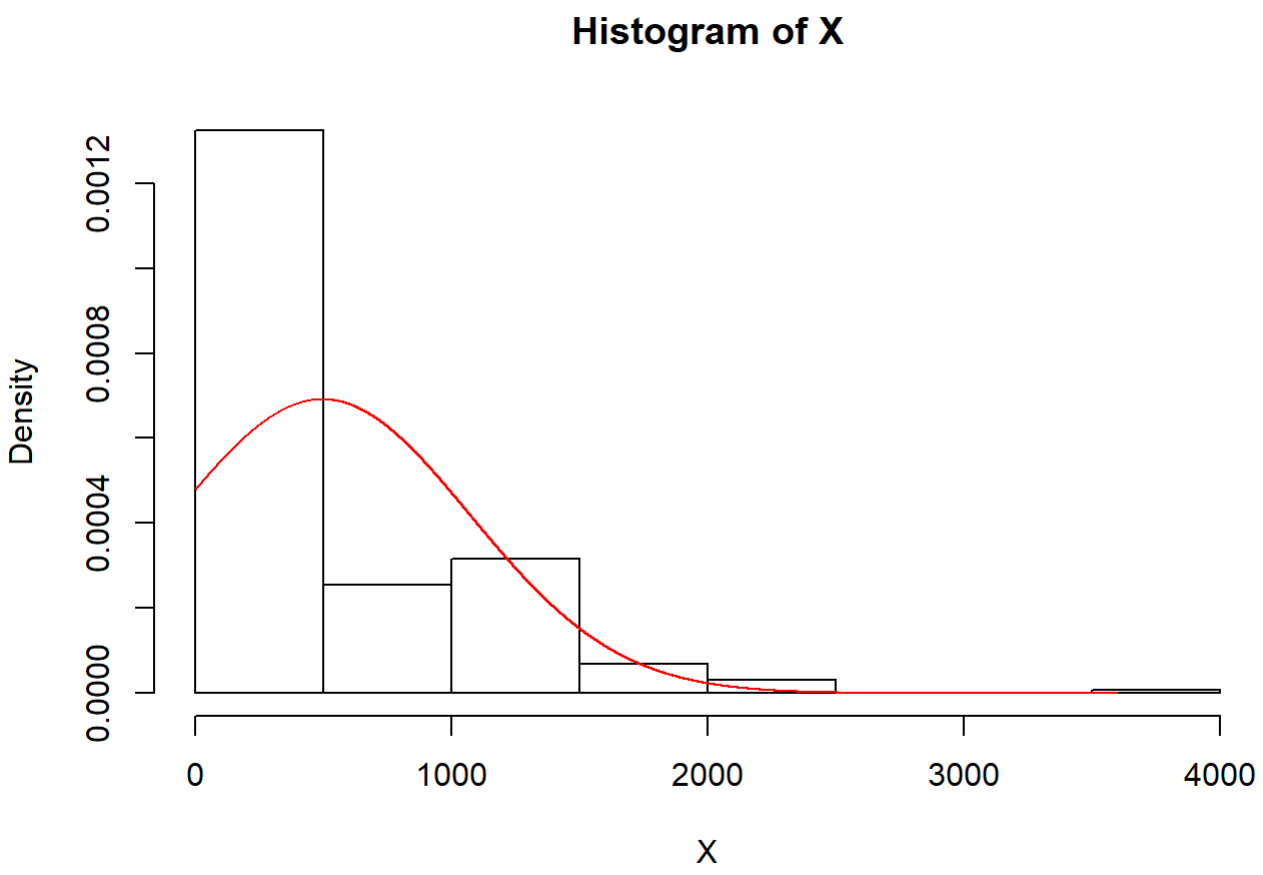



```
X = M$Sodium

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")
```



5. Para explorar curtosis y sesgo:

```
library(moments)
library(e1071)
```

```
Warning: package 'e1071' was built under R version 4.2.2


Attaching package: 'e1071'

The following objects are masked from 'package:moments':

    kurtosis, moment, skewness
```

```
X = M$Cholesterol
skewness(X)
```

```
[1] 3.755186
```

```
kurtosis(X)
```

```
[1] 16.87947
```

```
X = M$Carbohydrates
skewness(X)
```

```
[1] 0.9021952
```

```
kurtosis(X)
```

```
[1] 1.324083
```

```
X = M$Sodium
skewness(X)
```

```
[1] 1.526317
```

```
kurtosis(X)
```

```
[1] 2.75191
```

Realizar la prueba de Anderson - Darling

```
library(nortest)
```

```
X = M$Calories
ad.test(X)
```

```
    Anderson-Darling normality test

data:  X
A = 2.5088, p-value = 2.369e-06
```

```
X = M$Carbohydrates
ad.test(X)
```

```
    Anderson-Darling normality test

data:  X
A = 4.1402, p-value = 2.547e-10
```

```
X = M$Sodium
ad.test(X)
```

```
    Anderson-Darling normality test

data:  X
A = 21.406, p-value < 2.2e-16
```

Interpretación para verificar si las distribuciones son normales.

1. Calorías: Esta variable es la primera que analicé y me gusta centrarme mayormente en la parte de verificar los residuos en el QQ-plot y el grafico de frecuencias, en donde podemos ver claramente que se presenta un sesgo a la derecha en los datos, que luego pordemos confirmar con el coeficiente de sesgo que calcuamos el final, en donde vemos que se presenta una asimetria positiva siendo > 0 y con colas un poco pesadas quizá por esos datos atipicos cerca de 2000 los cuales estaría interesante ver la opción de removerlos, por lo tanto podemos decir que esta variable no sigue una distribución normal. Además vemos que el p-valor ≤ 0.05: Hay evidencia suficiente para rechazar la hipótesis nula. Los datos no siguen una distribución normal. 2.- Carbohidratos: Aquí nuevamente revisando el QQ-plot y la grafica de frecuencias, podemos ver claramente que no sigue una distrubución normal. se presenta un moderado sesgo a la derecha y tambien vemos que las colas de la QQ plot se desalinean del centro, luego en los coeficientes de sesgo y curtosis se ve mejor que calorias, sin embargo, con el coeficiente de sesgo que calcuamos el final, en donde vemos que se presenta una asimetria positiva siendo > 0 y el de kurtosis es cercano a 1, es decir, está puntiaguda con colas livianas. Además vemos que el p-valor ≤ 0.05: Hay evidencia suficiente para rechazar la hipótesis nula. Los datos no siguen una distribución normal
2. Sodio: con esta variable desde el momento en que vemos las primeras 2 graficas podemos descartar completamente las opciones de que los datos siguen una distribucion normal, podemos ver un marcado sesgo a la derecha aunado a que los residuos en la QQ plot están completamente fuera de la diagonal central, revisando el coeficiente de sesgo que calcuamos el final, podmeos ver que se presenta una asimetria positiva siendo > 0 y su coeficiente de curtosis es cercano al 3 por lo que se parece a una normal pero tambien cuenta con colas livianas. Además vemos que el p-valor ≤ 0.05: Hay evidencia suficiente para rechazar la hipótesis nula. Los datos no siguen una distribución normal.