# tarea_4_A01742161

Rogelio Lizárraga

2024-08-13

```
M=read.csv("mc-donalds-menu(1).csv") #leer la base de datos
M$Protein
```

```
##   [1] 17 18 14 21 21 26 19 19 20 20 11 11 18 18 18 18 17 17 25 19 20 11 20 21 30
##  [26] 30 33 28 28 26 26 36 36 35 35  8 15 12  1  6  5  5 24 30 37 37 29 48 12 15
##  [51] 24 39 22 27 22 22 22 24 28 36 40 32 36 36 40 21 14 22 20 14 32 36 27 30 27
##  [76] 31 23 27  9 13 22 44 87 15  9 25 29  6 23 27 14 16 14 16 15 16  2  4  6  1
## [101]  1  0  4  2  2  2  1  8  7  6  0  0  0  0  0  0  0  0  0  0  0  0  2  3  4
## [126]  1  0  0  0  0  8  9  0  2  3  4  0  0  0  0  0  1  1  1  0  0  0  0  9 11
## [151] 15  9 11 15  9 11 15  9 11 15  9 12 15 10 12 16 10 12 16 10 12 16 10 12 16
## [176] 10 12 16 10 13 16 11 13 17 10 12 16 10 13 17 11 14 17 12 14 19  1  1  2  1
## [201]  1  2  1  1  2  1  1  2  1  1  2  8  9 14  8 10 14  8  9 13  8 10 14  7  9
## [226] 11  7  9 11  8  9 12  2  3  4  3  4  5  2  3  4 11 14 18 12 15 18 12 15 19
## [251] 14 18 13 20  9 12 15  8 21 10
```

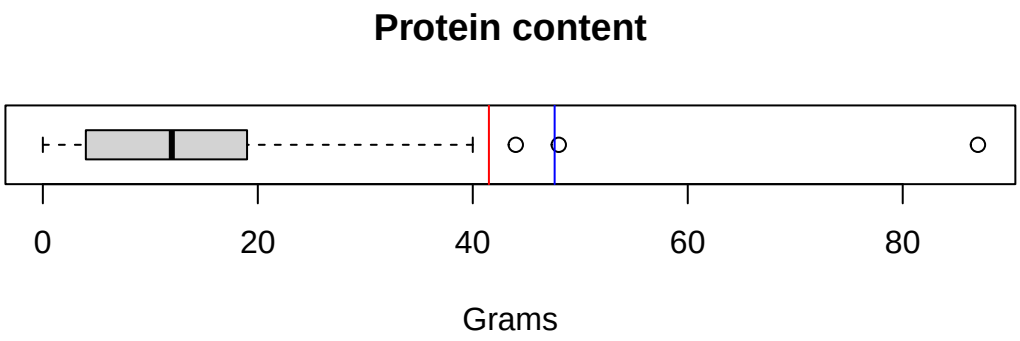
## Análisis de datos atípicos y normalidad de Protein y Calories

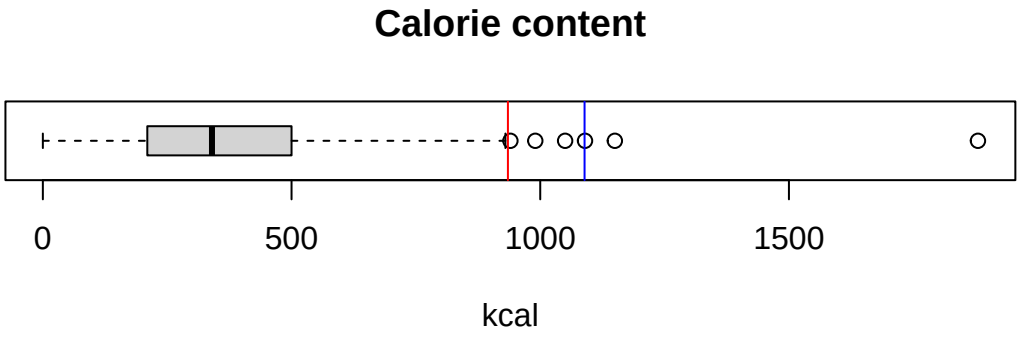
```
q1 = quantile(M$Protein, 0.25)  #Cuantil 1 de la variable X
q3 = quantile(M$Protein, 0.75)
ri= IQR(M$Protein)  #Rango intercuartílico de X

q1_cals = quantile(M$Calories, 0.25)
q3_cals = quantile (M$Calories, 0.75)
ri_cals = IQR(M$Calories)
par(mfrow=c(2,1))

g_caja = boxplot(M$Protein,horizontal=TRUE,ylim=c(min(M$Protein), max(M$Protein)), main = "Protein conte
abline(v=q3+1.5*ri,col="red")
abline(v= mean(M$Protein) + 3*sd(M$Protein),col="blue")

g_caja_2 = boxplot(M$Calories,horizontal=TRUE,ylim=c(min(M$Calories), max(M$Calories)), main = "Calorie
abline(v=q3_cals+1.5*ri_cals,col="red")
abline(v= mean(M$Calories) + 3*sd(M$Calories),col="blue")
```




Observando nuestras línea rojas (criterio de rangos intercuartílicos) y nuestras líneas azules ($\mu + 3\sigma$), hay datos que se encuentran fuera de los rangos comunes. Sin embargo, estos datos no se deben borrar, pues son alimentos del Menú de McDonalds, donde se parte de una población y una muestra. Al ser la población, estos datos no son erróneos y sí son correctos.

### Medidas de media, mediana y rango medio en Calories y Protein

```
cat(c('Media de Calories:', mean(M$Calories), '\n'))
```

```
## Media de Calories: 368.269230769231
```

```
cat(c('Mediana de Calories:', median(M$Calories), '\n'))
```

```
## Mediana de Calories: 340
```

```
cat(c('Rango medio de Calories:', ((min(M$Calories) + max(M$Calories)) /2), '\n'))
```

```
## Rango medio de Calories: 940
```

```
cat(c('Media de Protein:', mean(M$Protein), '\n'))
```

```
## Media de Protein: 13.3384615384615
```

```r
cat(c('Mediana de Protein:', median(M$Protein), '\n'))
```

```
## Mediana de Protein: 12
```

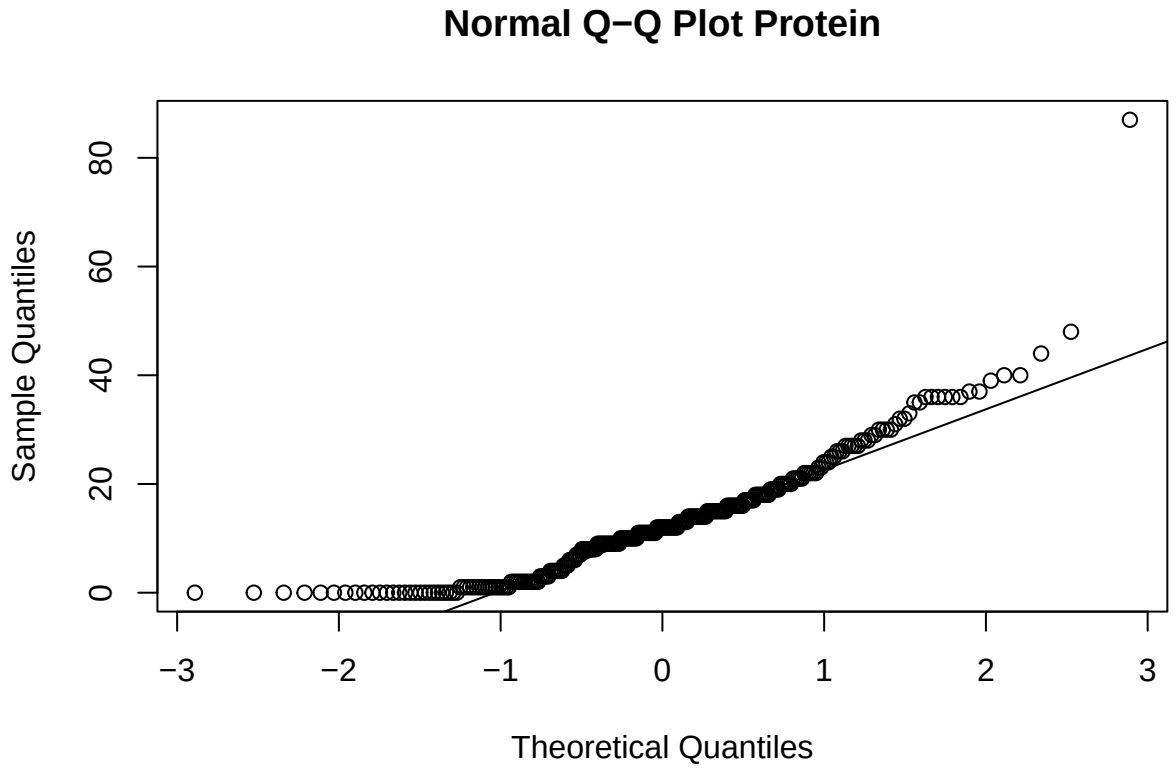
```r
cat(c('Rango medio de Protein:', ((min(M$Protein) + max(M$Protein)) /2), '\n'))
```

```
## Rango medio de Protein: 43.5
```

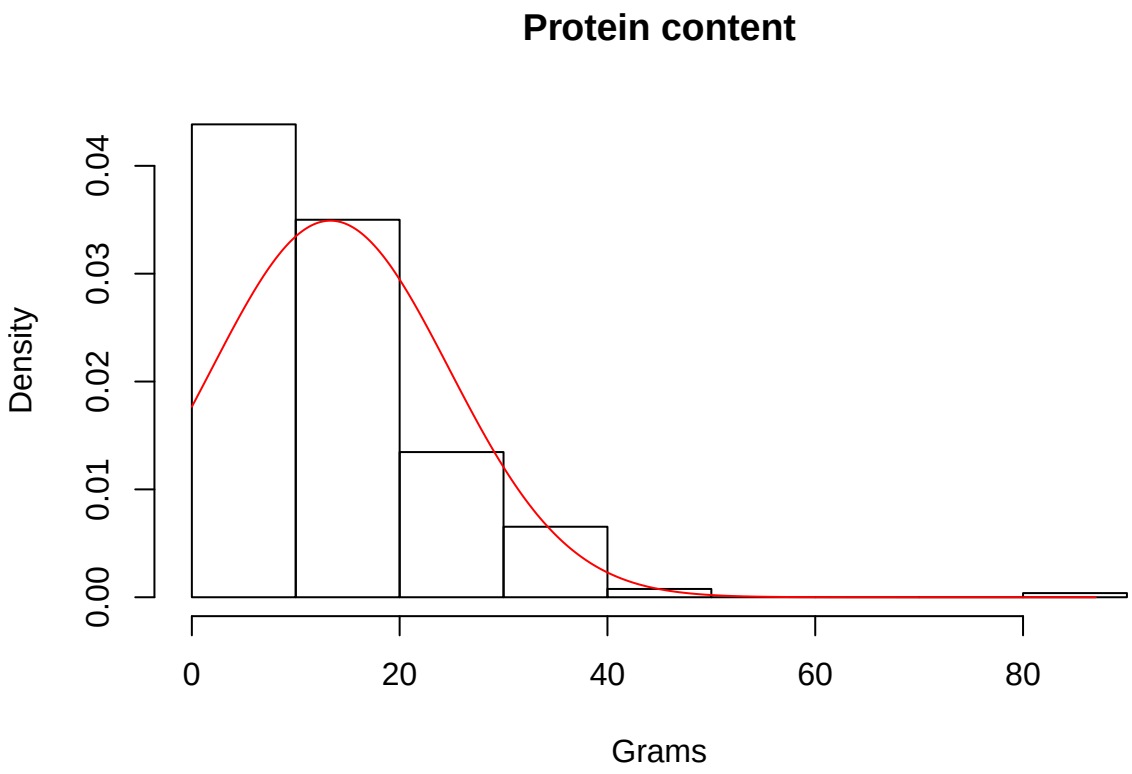
Observamos que las medias y las medianas no son muy lejanas. Sin embargo, vemos que el rango medio sí está muy alejado, debido a los datos atípicos que se desvían de la media.

Ahora, analizaremos normalidad, donde se incluyen todos los datos:

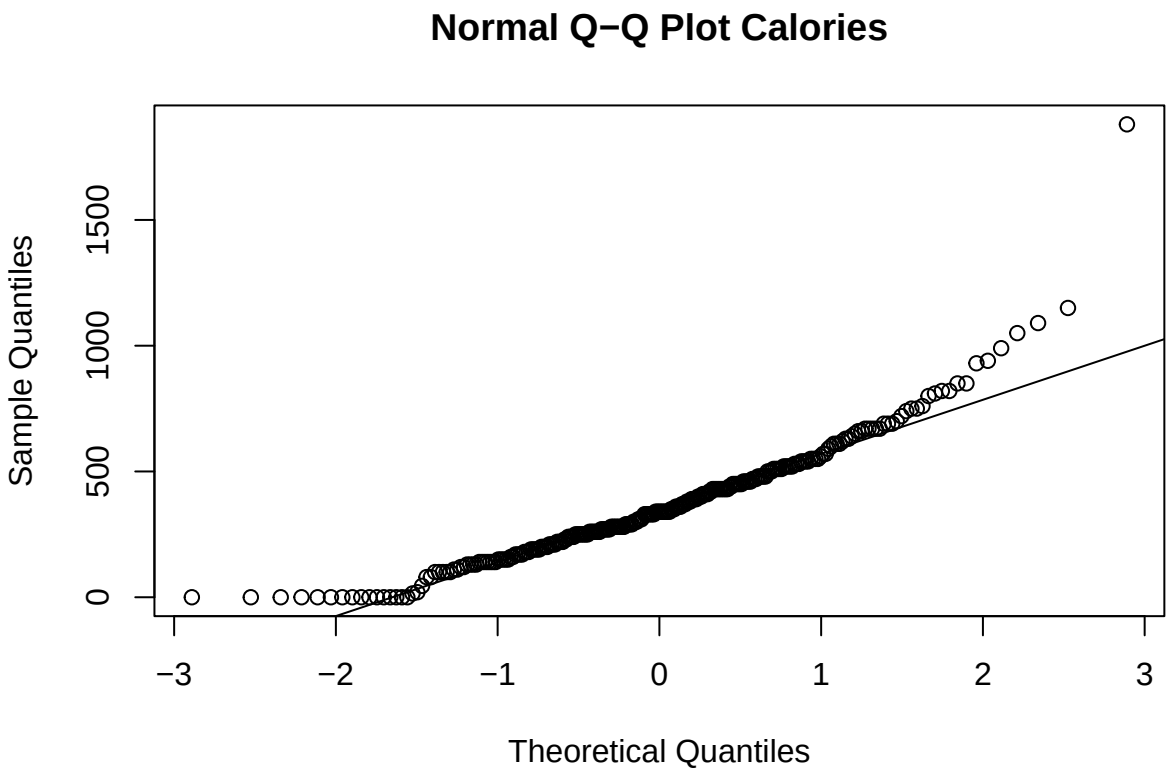
```r
qqnorm(M$Protein, main = 'Normal Q-Q Plot Protein')
qqline(M$Protein)
```



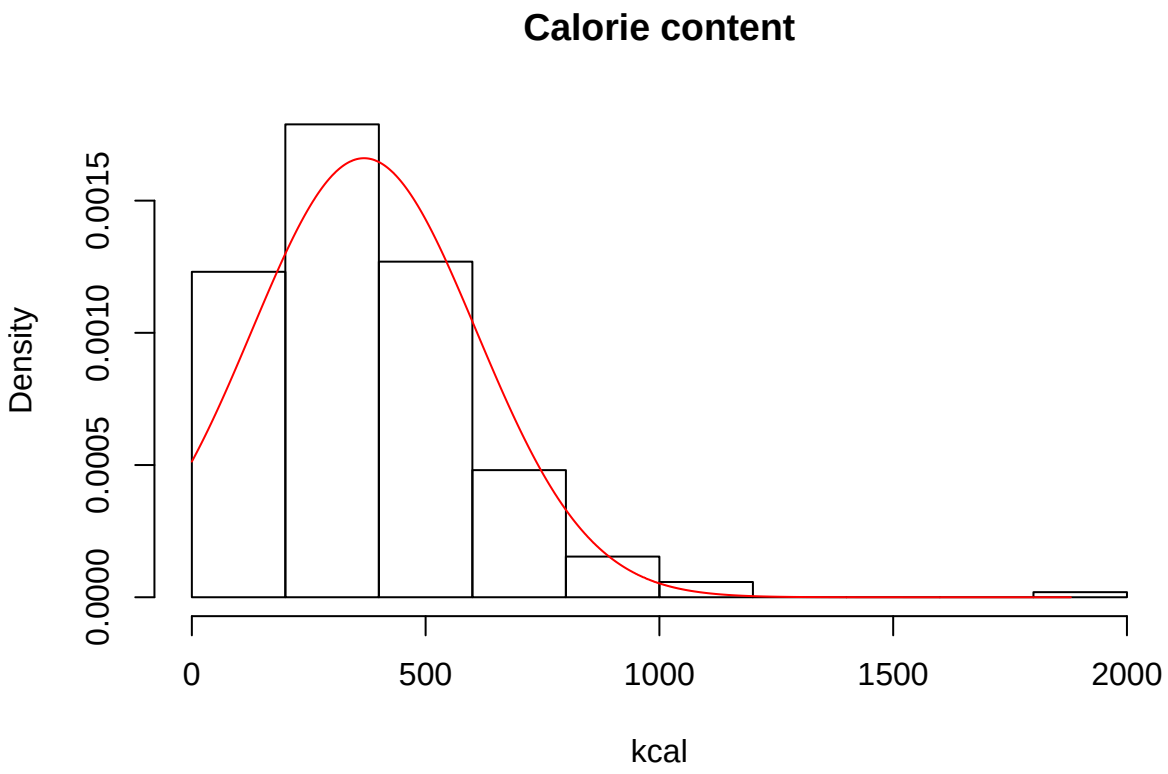

```r
hist(M$Protein,prob=TRUE,  main = "Protein content", xlab = "Grams",col=0)
x=seq(min(M$Protein),max(M$Protein),0.1)
y=dnorm(x,mean(M$Protein),sd(M$Protein))
lines(x,y,col="red")
```




```
qqnorm(M$Calories, main = 'Normal Q-Q Plot Calories')
qqline(M$Calories)
```




```r
hist(M$Calories,prob=TRUE,  main = "Calorie content", xlab = "kcal",col=0)
x=seq(min(M$Calories),max(M$Calories),0.1)
y=dnorm(x,mean(M$Calories),sd(M$Calories))
lines(x,y,col="red")
```




Observando los gráficos QQ, observamos que estos cuentan con colas muy largas, por lo que parece que los datos no se distribuyen como una normal. En los histogramas observamos algo similar, puesto que los datos tienen una cola izquierda muy larga y la mayoría de los datos deberían de estar en $(\mu \pm \sigma)$, pero observamos que no es así.

### Pruebas de hipótesis de Protein

$H_0$ = el conjunto de datos tiene una asimetría y una curtosis que coincide con una distribución normal. $H_1$ = el conjunto de datos tiene una asimetría y una curtosis que no coincide con una distribución normal

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

```
## [1] 1.570794
```

```
kurtosis(M$Protein)
```

```
## [1] 8.86355
```

```
jarque.test(M$Protein)
```

```
##
##  Jarque-Bera Normality Test
##
```

```
## data:  M$Protein
## JB = 479.38, p-value < 2.2e-16
## alternative hypothesis: greater
```

Como nuestro valor $p < 0.05$, se rechaza $H_0$, por lo que el conjunto de datos tiene una asimetría y una curtosis que NO coincide con una distribución normal. Además, observamos que tenemos un sesgo de 1.57 y una curtosis de 8.86, los cuales son valores demasiado elevados que NO coinciden con una distribución normal.

### Pruebas de hipótesis de Calories

$H_0$ = el conjunto de datos tiene una asimetría y una curtosis que coincide con una distribución normal. $H_1$ = el conjunto de datos tiene una asimetría y una curtosis que no coincide con una distribución normal

```
skewness(M$Calories)
```

```
## [1] 1.444105
```

```
kurtosis(M$Calories)
```

```
## [1] 8.645274
```

```
jarque.test(M$Calories)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  M$Calories
## JB = 435.62, p-value < 2.2e-16
## alternative hypothesis: greater
```

Como nuestro valor $p < 0.05$, se rechaza $H_0$, por lo que el conjunto de datos tiene una asimetría y una curtosis que NO coincide con una distribución normal. Además, observamos que tenemos un sesgo de 1.44 y una curtosis de 8.85, los cuales son valores demasiado elevados que NO coinciden con una distribución normal.

## Conclusión

Finalmente, podemos concluir que los conjuntos de datos tienen asimetrías y curtosis que NO coinciden con una distribución normal. Además, observando los gráficos Q-Q y los histogramas, observamos que los datos no se distribuyen como normales.

Por otro lado, se puede concluir que no es posible eliminar los datos atípicos, debido a la relevancia que tienen dentro de los datos, pues son parte de la población del menú de McDonalds.