2. Explorando bases

Datos del alumno

Luis Ángel Guzmán Iribe - A01741757

```
library(readr)
library(moments)
data <- read.csv("mc-donalds-menu-1.csv")
```

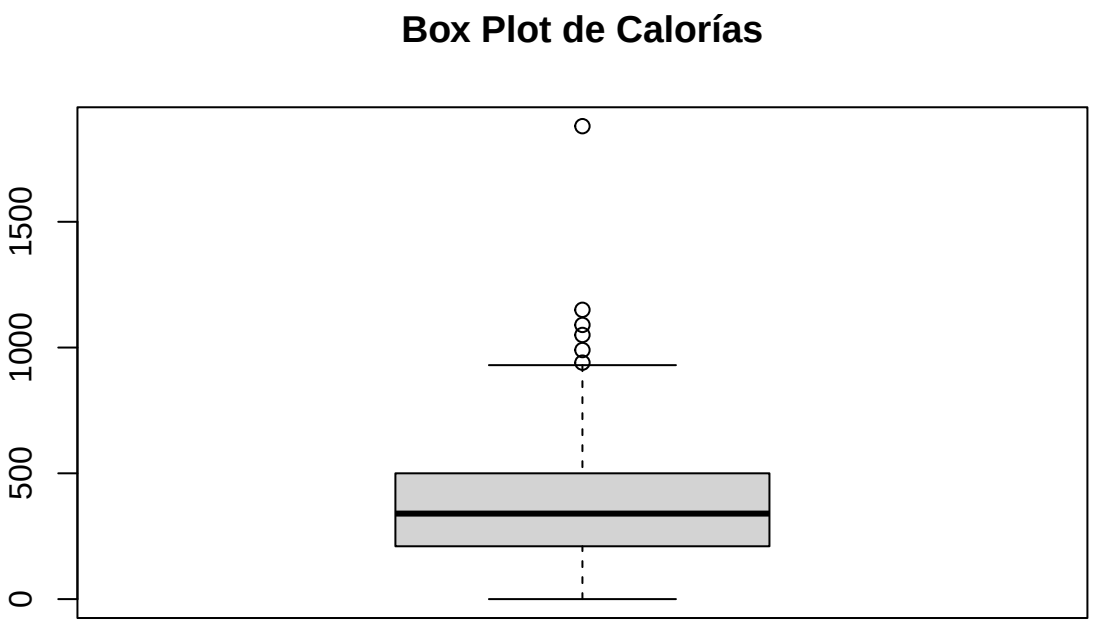
Analisis de calorías

Box plot

```
data_calories <- data$Calories

q1 <- quantile(data_calories, 0.25)
q3 <- quantile(data_calories, 0.75)
ri <- q3 - q1
mfrow=c(2,1)  #Matriz de gráficos de 2x1

boxplot(data_calories, main = "Box Plot de Calorías")
abline(v = q3 + 1.5 * ri, col = "red")
```



La gráfica de caja y bigote, de buenas a primeras, parace ser un buen indicador de normalidad en la distribución de los datos, aunque tendiendo más hacia una mayor dispersión de datos hacia el tercer cuartil, así como la existencia de algunos datos anómalos, los cuales se procesarán en el próximo pedazo de código.

Sesgo y curtosis

```
data_calories1 <- data[data$Calories < q3 + 1.5 * ri, c("Calories")]
summary(data_calories1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   202.5   335.0   349.0   480.0   930.0
```

```
summary(data_calories)
```

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

```
# Sesgo y coeficiente de curtosis
skewness_value <- skewness(data_calories1)
kurtosis_value <- kurtosis(data_calories1)

cat("Sesgo:", skewness_value, "\n")
```

```
## Sesgo: 0.3490549
```

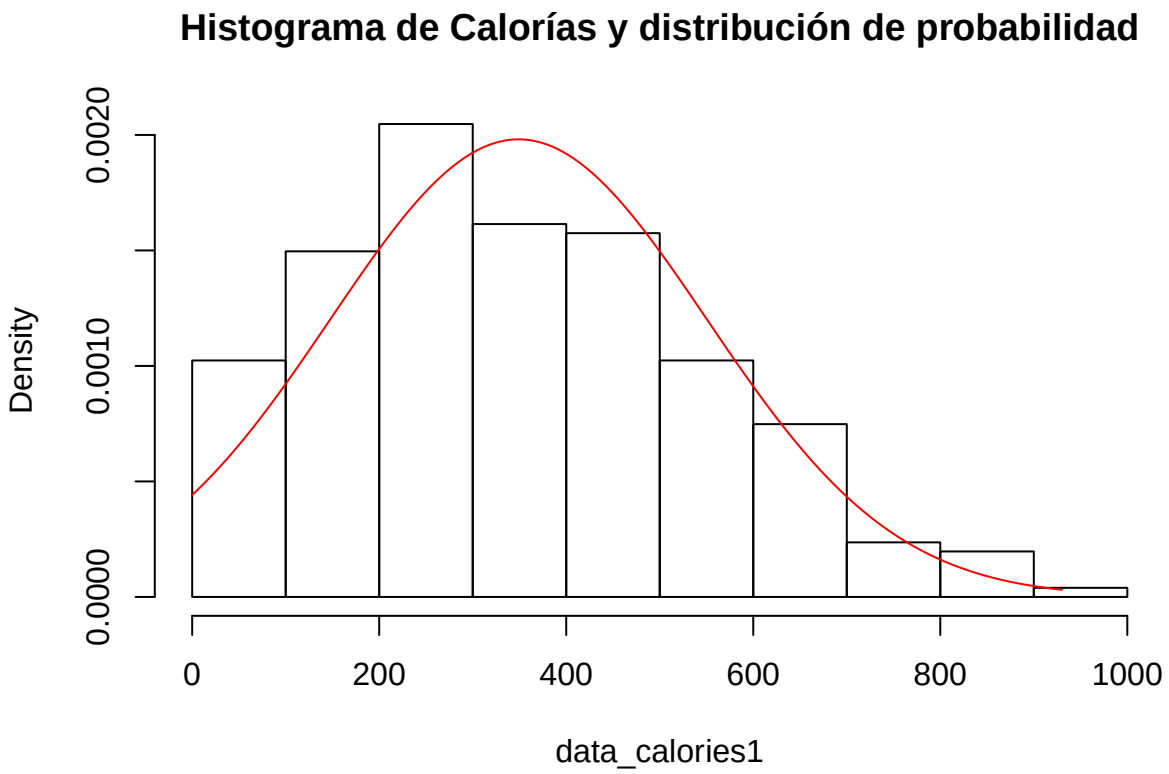
```
cat("Curtosis:", kurtosis_value, "\n")
```

```
## Curtosis: 2.716828
```

De los valores más relevantes que podemos extraer de esta sección, la curtosis, nos ayuda a comprender mejor el comportamiento de la distribución de datos. Un valor de 2.27 es indicativo de una distribución normal con ligeramente más peso en el centro, o lo que llamaríamos coloquialmente, más "picuda". Por otro lado, un valor de sesgo cercano a 0 es tambien un buen indicador de simetría en la curva, algo esperado de una distribución normal, en este caso, el sesgo ocurre hacia el lado derecho de la distribución, dado que tratamos con un valor positivo de 0.34.

Histograma y distribución de probabilidad

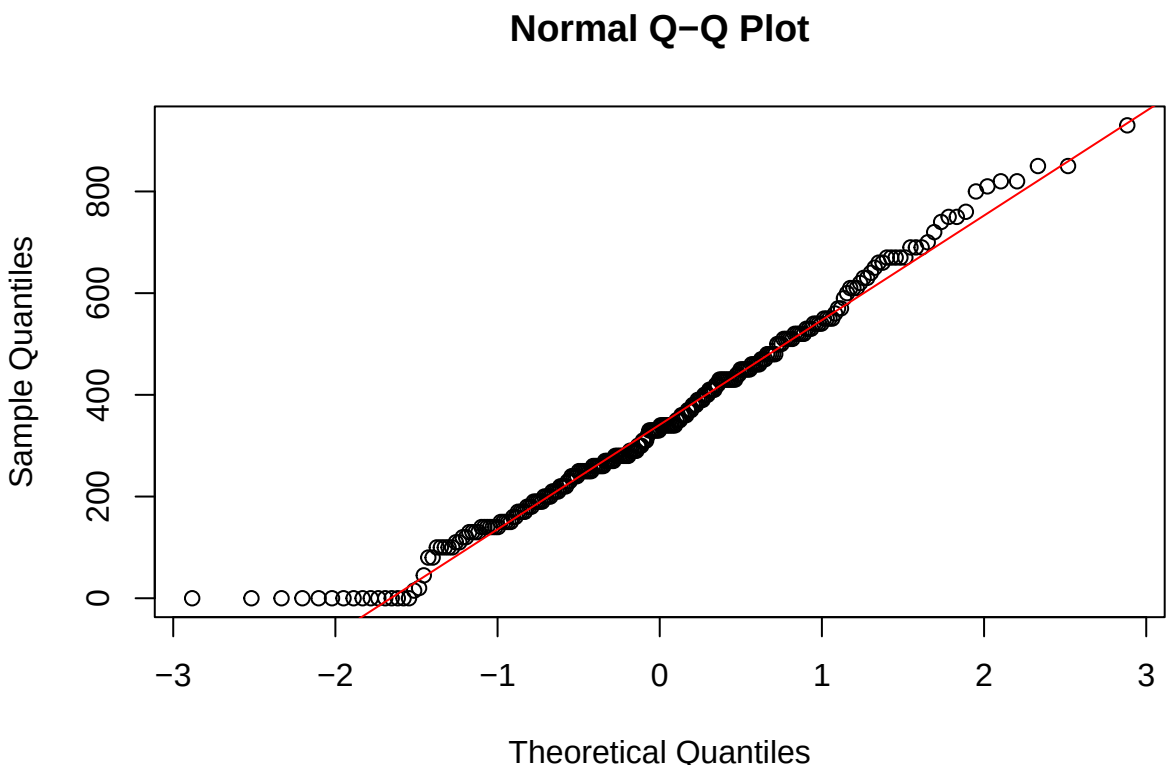
```
# Histograma y distribución teorica de probabilidad
hist(data_calories1, prob = TRUE, col = 0, main = "Histograma de Calorías y distribución de probabilidad
x <- seq(min(data_calories1), max(data_calories1), 0.1)
y <- dnorm(x, mean(data_calories1), sd(data_calories1))
lines(x, y, col = "red")
```



Como describía anteriormente, los datos parecen indicar a una distribución normal no perfecta, más concentrada en los centros y sesgada ligeramente hacia la derecha, pero en la tabla de histograma podemos apreciar como la predicción probabilistica se ajusta de manera satisfactoria a los datos proveidos, exceptuando por un mayor margen de error hacia la izquierda de la tabla (indicado anteriormente por el sesgo positivo) y la concentración central elevada (indicada por la curtosis menor a 3).

QQ plot

```
# QQ Plot
qqnorm(data_calories1)
qqline(data_calories1, col="red")
```



A grandes rasgos, podemos afirmar que los puntos se alinean de manera satisfactoria con la linea recta, lo que es de nuevo un buen indicador de normalidad, las desviaciones que observamos pueden ser explicadas con los puntos mencionados en las secciones anteriores, como un grado ligeramente negativo de curtosis, y un sesgo hacia valores mayores.

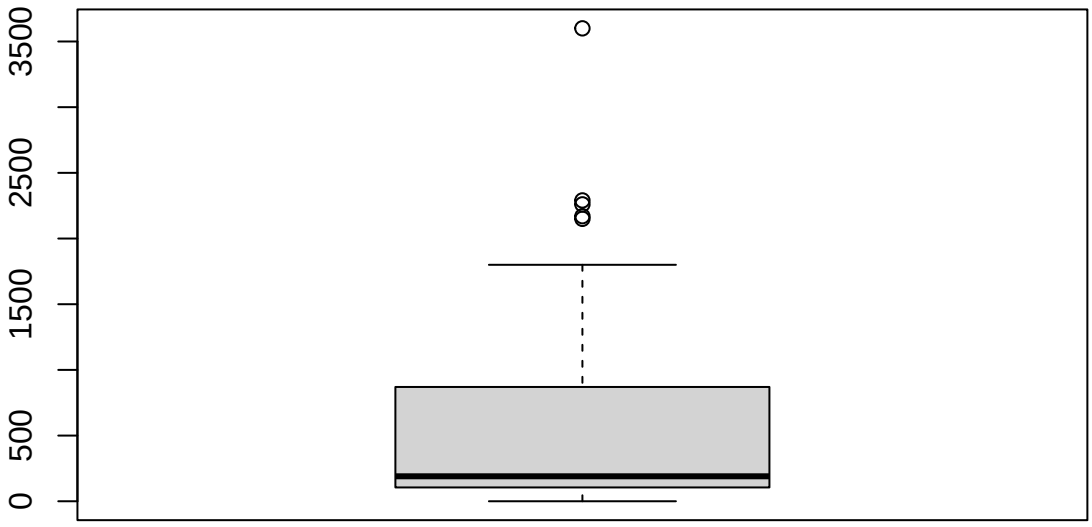
Analisis de Sodio

Box plot

```
data_sodium <- data$Sodium

q1 <- quantile(data_sodium, 0.25)
q3 <- quantile(data_sodium, 0.75)
ri <- q3 - q1
mfrow=c(2,1)  #Matriz de gráficos de 2x1

boxplot(data_sodium, main = "Box Plot de Sodio")
abline(v = q3 + 1.5 * ri, col = "red")
```



Este tipo de gráfica de caja y bigute, indica una muy fuerte concentración de datos entre los primeros 2 cuartiles de la gráfica, lo que indica un muy fuerte sesgo hacia los valores más bajos de la tabla, llevando de este modo a la sospecha inicial de que esta no se trata de una distribución normal.

Sesgo y curtosis

```
data_sodium1 <- data[data$Sodium < q3 + 1.5 * ri, c("Sodium")]
summary(data_sodium1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    95.0   190.0   456.6   830.0  1800.0
```

```
summary(data_sodium)
```

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

```
# Sesgo y coeficiente de curtosis
skewness_value <- skewness(data_sodium1)
kurtosis_value <- kurtosis(data_sodium1)

cat("Sesgo:", skewness_value, "\n")
```

```
## Sesgo: 1.034162
```

```
cat("Curtosis:", kurtosis_value, "\n")
```

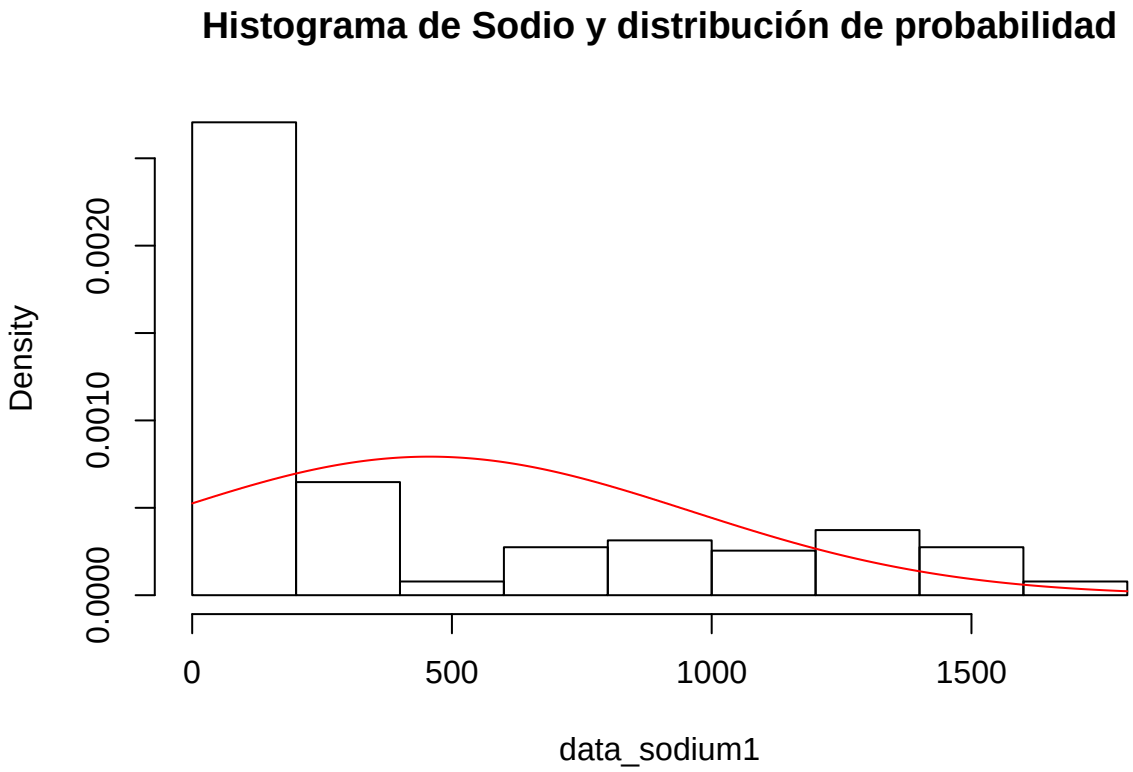
```
## Curtosis: 2.598528
```

El sesgo en esta ocación es, como mencionaba anteriormente, mucho más pronunciado que en las gráficas sobre el colesterol. Un sesgo mayor a 1 puede ser considerado como suficiente para afirmar que no estamos tratando con una distribución normal, y sería más apropiado buscar otras distribuciones estadisticas para analizar nuestro conjunto de datos.

Histograma y distribución de probabilidad

```
# Histograma y distribución teorica de probabilidad
hist(data_sodium1, prob = TRUE, col = 0, main = "Histograma de Sodio y distribución de probabilidad")
```

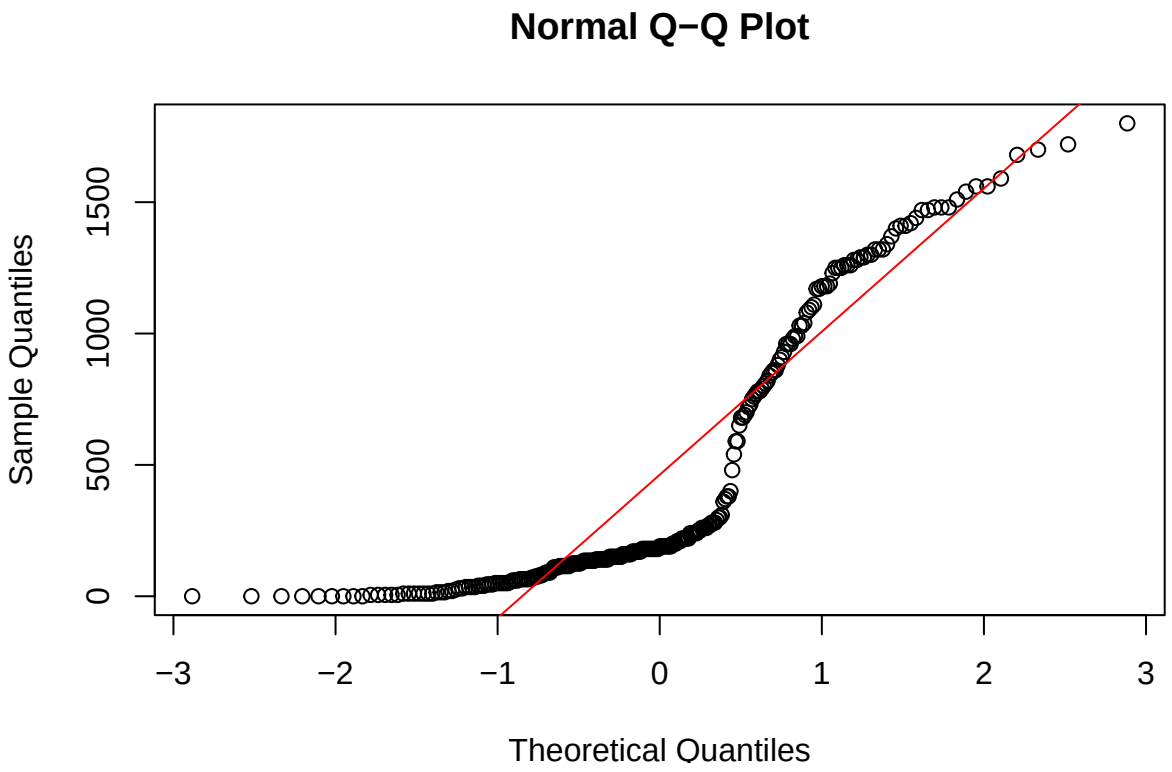
```r
x <- seq(min(data_sodium1), max(data_sodium1), 0.1)
y <- dnorm(x, mean(data_sodium1), sd(data_sodium1))
lines(x, y, col = "red")
```



El histograma viene a confirmar nuestras sospechas, podemos apreciar como la mayoría de los datos se encuentran en la primera barra del histograma, con un rápido declive y leve resugimiento casi al final de la tabla, en prácticamente todos los puntos de la tabla la linea de distribución falla en representar de manera adecuada el comportamiento de los datos.

QQ plot

```r
# QQ Plot
qqnorm(data_sodium1)
qqline(data_sodium1, col="red")
```



De igual modo, cuando en una distribución normal esperaríamos que los datos se alineen con la linea recta, en esta ocación encontramos que no se acerca en lo más minimo, si bien la linea se ajusta a la tendencia general de los datos, falla en reconocer los matices en prácticamente todos los puntos de la gráfica. Este patrón en forma de “s” podría ser indicativo de una desviación más compleja de curva normal, de nuevo, apuntando a que sería prudente provar con otras distribuciones estadisticas que representen más apropiadamente este conjunto de datos.