Tarea 4 - Explorando Bases

Héctor Hibran Tapia Fernández - A01661114

2024-08-13

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.2.3
```

1. Importamos base de datos

```
data <- read.csv("./mc-donalds-menu.csv", header = TRUE, sep = ",")
```

2. Analiza las siguientes variables en cuanto a sus datos atípicos y normalidad:

- Calorias
- Proteinas

```
summary(data$Calories)
```

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

```
cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
q1 <- quantile(data$Calories, 0.25)
q2 <- quantile(data$Calories, 0.50)
q3 <- quantile(data$Calories, 0.75)

iqr <- IQR(data$Calories)

# Límites para los datos atípicos
lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr

outliers_IQR <- data$Calories[data$Calories < lower_bound | data$Calories > upper_bound]
# Datos atípicos o outliers

cat("Cuartiles:\n")
```

```
## Cuartiles:
```

```
cat("Al 25%: ", q1, "\n")
```

```
## Al 25%:  210
```

```
cat("Al 50%: ", q2, "\n")
```

```
## Al 50%:  340
```

```
cat("Al 75%: ", q3, "\n")
```

```
## Al 75%:  500
```

```
cat("Rango Intercuartílico: ", iqr, "\n")
```

```
## Rango Intercuartílico:  290
```

```
sesgo <- skewness(data$Calories)
curtosis <- kurtosis(data$Calories)

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Coeficiente de sesgo: ", sesgo, "\n")
```

```
## Coeficiente de sesgo:  1.435782
```

```
cat("Coeficiente de curtosis: ", curtosis, "\n")
```

```
## Coeficiente de curtosis:  5.5789
```

```
media <- mean(data$Calories, na.rm = TRUE)
mediana <- median(data$Calories, na.rm = TRUE)
rango_medio <- (max(data$Calories, na.rm = TRUE) + min(data$Calories, na.rm = TRUE)) / 2

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Media: ", media, "\n")
```

```
## Media:  368.2692
```

```
cat("Mediana: ", mediana, "\n") # Es igual al q2
```

```
## Mediana:  340
```

```
cat("Rango Medio: ", rango_medio, "\n")
```

```
## Rango Medio:  940
```

```
sd_calories <- sd(data$Calories, na.rm = TRUE)

lower_bound_sd <- media - 3 * sd_calories
upper_bound_sd <- media + 3 * sd_calories

outliers_sd <- data$Calories[data$Calories < lower_bound_sd | data$Calories > upper_boun
d_sd]

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Outliers con la cota de 1.5 rangos intercuartílicos: ", outliers_IQR, "\n")
```

```
## Outliers con la cota de 1.5 rangos intercuartílicos:  1090 1150 990 1050 940 1880
```

```
cat("Outliers con la cota de 3 desviaciones estándar al rededor de la media: ", outliers
_sd, "\n")
```

```
## Outliers con la cota de 3 desviaciones estándar al rededor de la media:  1090 1150 18
80
```

Vamos por partes…

1.Tenemos primero que el valor más pequeño de la muestra es 0.0 y el valor más grande es de 1880.0. 2. Cuartiles: 210 valor debajo del cual se encuentra el 25% de los datos. Cuartiles: 340 valor debajo del cual se encuentra el 50% de los datos, o mediana. Cuartiles: 500 valor debajo del cual se encuentra el 75% de los datos. Rango Intercuartílico: Es la diferencia entre el tercer cuartil (500) y el primer cuartil (210).Mide la dispersión de la mitad central de los datos y es una medida de la variabilidad.

3. Coeficiente de sesgo: 1.435782 -> Como es positivo indica que la distribución está sesgada hacia la derecha (positiva). Los datos tienden a acumularse en el lado izquierdo con una cola más larga hacia la derecha. Coeficiente de curtosis: 5.5789 -> Indica que la distribución tiene colas más pesadas que una

distribución normal (kurtosis > 0). Es una distribución leptocúrtica.

4. Media: 368.2692 -> Indica el valor promedio de los datos. Mediana: 340 -> Muestra la mediana o el valor central de la variable ordenada. Rango Medio: 940 -> Se obtiene el punto medio entre los valores extremos de la distribución.

5. Outliers con la cota de 1.5 rangos intercuartílicos: 1090 1150 990 1050 940 1880 Outliers con la cota de 3 desviaciones estándar al rededor de la media: 1090 1150 1880

Para quitar los outliers dependerá mucho del contexto de problema que tengamos, se necesita analizar tal vez la correlación de estos y así poder decidir cuál de los dos contexto es mejor para nuestro problema.

```
summary(data$Protein)
```

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

```
cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
q1 <- quantile(data$Protein, 0.25)
q2 <- quantile(data$Protein, 0.50)
q3 <- quantile(data$Protein, 0.75)

iqr <- IQR(data$Protein)

# Límites para los datos atípicos
lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr

outliers_IQR <- data$Protein[data$Protein < lower_bound | data$Protein > upper_bound] # Datos atípicos o outliers

cat("Cuartiles:\n")
```

```
## Cuartiles:
```

```
cat("Al 25%: ", q1, "\n")
```

```
## Al 25%:  4
```

```
cat("Al 50%: ", q2, "\n")
```

```
## Al 50%:  12
```

```
cat("Al 75%: ", q3, "\n")
```

```
## Al 75%:  19
```

```
cat("Rango Intercuartílico: ", iqr, "\n")
```

```
## Rango Intercuartílico:  15
```

```
sesgo <- skewness(data$Protein)
curtosis <- kurtosis(data$Protein)

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Coeficiente de sesgo: ", sesgo, "\n")
```

```
## Coeficiente de sesgo:  1.561741
```

```
cat("Coeficiente de curtosis: ", curtosis, "\n")
```

```
## Coeficiente de curtosis:  5.7955
```

```
media <- mean(data$Protein, na.rm = TRUE)
mediana <- median(data$Protein, na.rm = TRUE)
rango_medio <- (max(data$Protein, na.rm = TRUE) + min(data$Protein, na.rm = TRUE)) / 2

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Media: ", media, "\n")
```

```
## Media:  13.33846
```

```
cat("Mediana: ", mediana, "\n") # Es igual al q2
```

```
## Mediana:  12
```

```
cat("Rango Medio: ", rango_medio, "\n")
```

```
## Rango Medio:  43.5
```

```
sd_Protein <- sd(data$Protein, na.rm = TRUE)

lower_bound_sd <- media - 3 * sd_Protein
upper_bound_sd <- media + 3 * sd_Protein

outliers_sd <- data$Protein[data$Protein < lower_bound_sd | data$Protein > upper_bound_s
d]

cat("-----------------------------------------------\n")
```

```
## -----------------------------------------------
```

```
cat("Outliers con la cota de 1.5 rangos intercuartílicos: ", outliers_IQR, "\n")
```

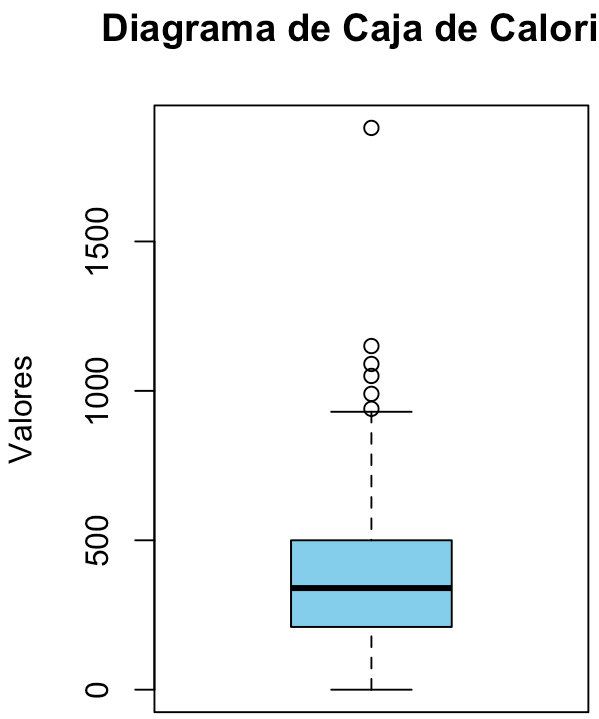
```
## Outliers con la cota de 1.5 rangos intercuartílicos:  48 44 87
```

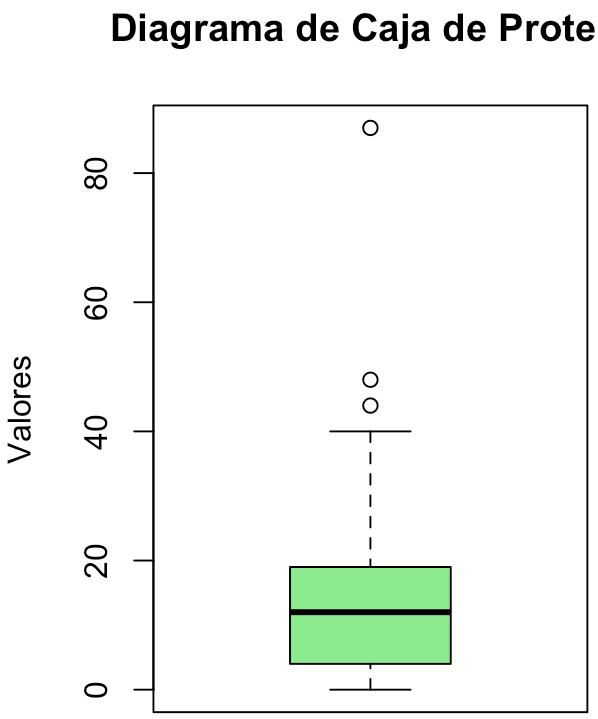
```
cat("Outliers con la cota de 3 desviaciones estándar al rededor de la media: ", outliers
_sd, "\n")
```

```
## Outliers con la cota de 3 desviaciones estándar al rededor de la media:  48 87
```

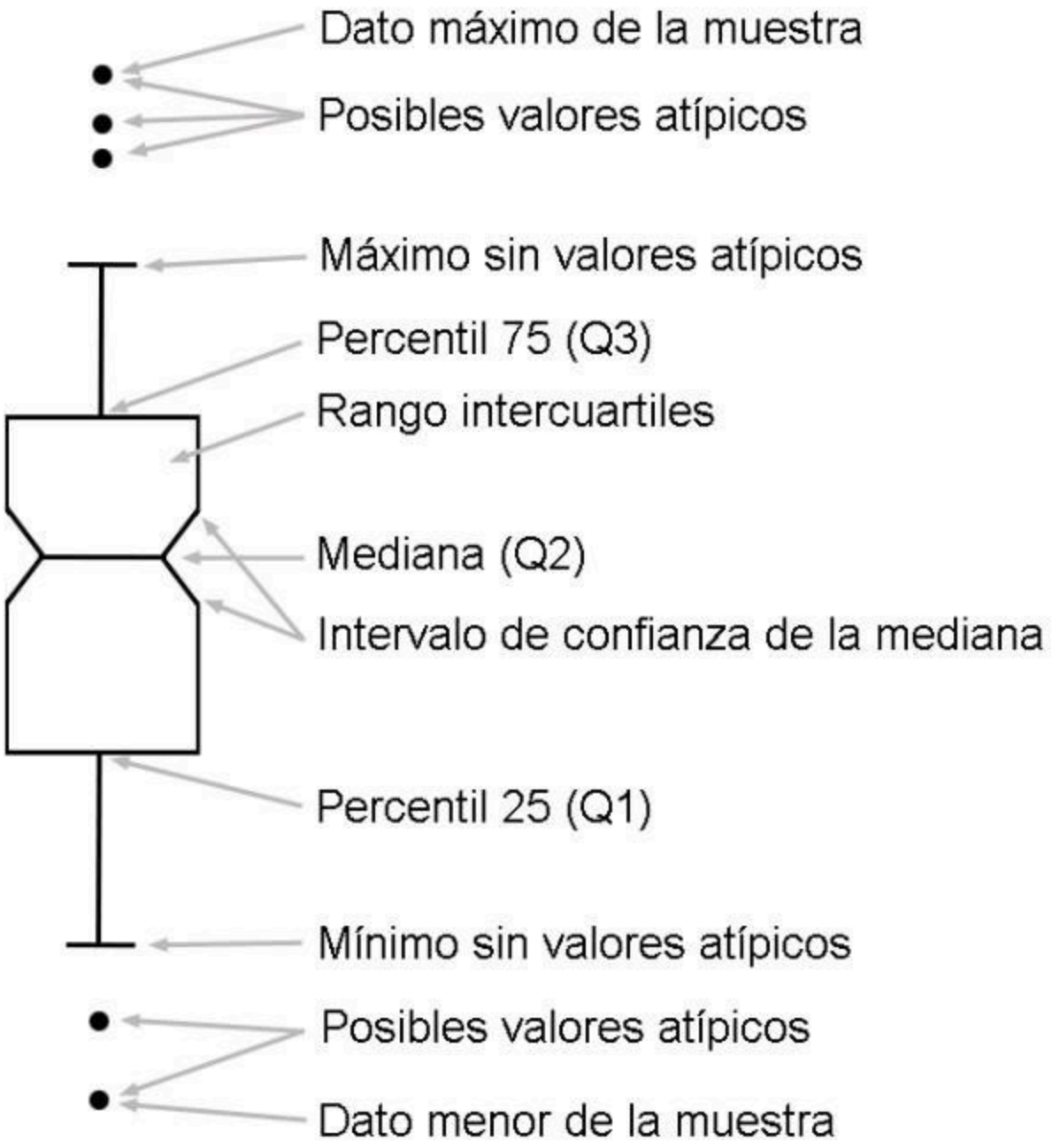
Se repite la misma lógica que se ocupó para Calorias.

```
par(mfrow = c(1, 2))  # 1 fila, 3 columnas
boxplot(data$Calories, main = "Diagrama de Caja de Calories", ylab = "Valores", col = "s
kyblue")
boxplot(data$Protein, main = "Diagrama de Caja de Protein", ylab = "Valores", col = "lig
htgreen")
```





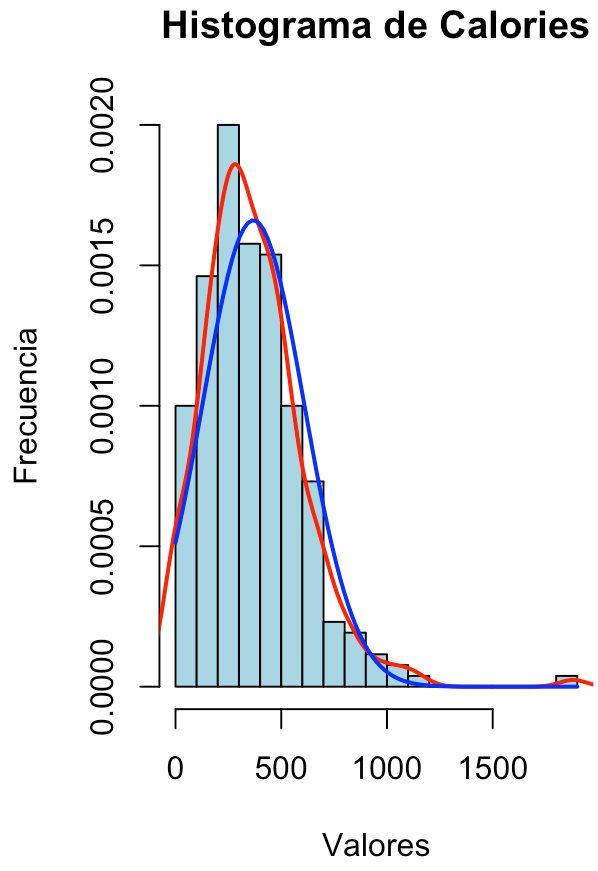

```
par(mfrow = c(1, 1))
```

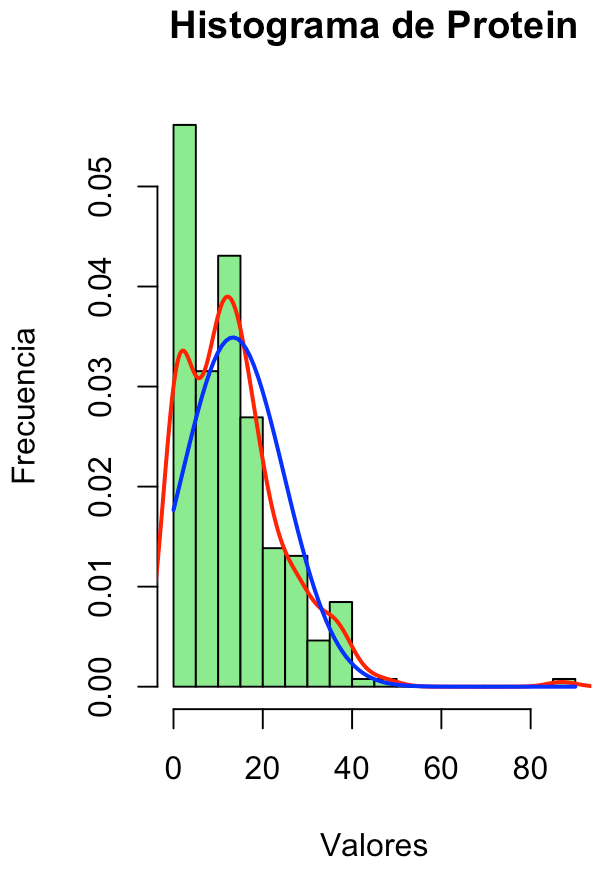


Descripción de la imagen

```
par(mfrow = c(1, 2))  # 1 fila, 2 columnas
hist(data$Calories, main = "Histograma de Calories", xlab = "Valores", ylab = "Frecuenci
a", col = "lightblue", breaks = 20, prob = TRUE)
calories_density <- density(data$Calories)
lines(calories_density, col = "red", lwd = 2)
curve(dnorm(x, mean = mean(data$Calories), sd = sd(data$Calories)), add = TRUE, col = "b
lue", lwd = 2)

hist(data$Protein, main = "Histograma de Protein", xlab = "Valores", ylab = "Frecuenci
a", col = "lightgreen", breaks = 20, prob = TRUE)
protein_density <- density(data$Protein)
lines(protein_density, col = "red", lwd = 2)
curve(dnorm(x, mean = mean(data$Protein), sd = sd(data$Protein)), add = TRUE, col = "blu
e", lwd = 2)
```



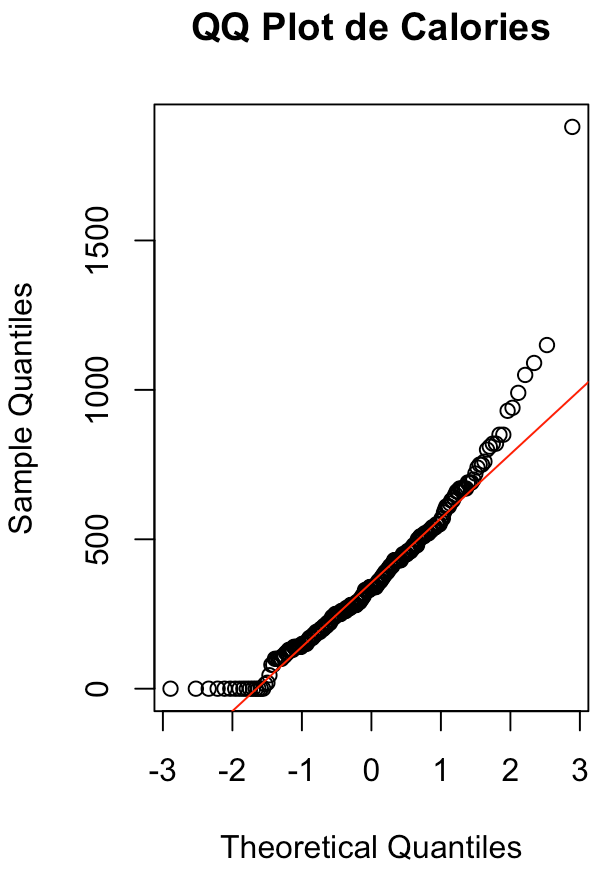


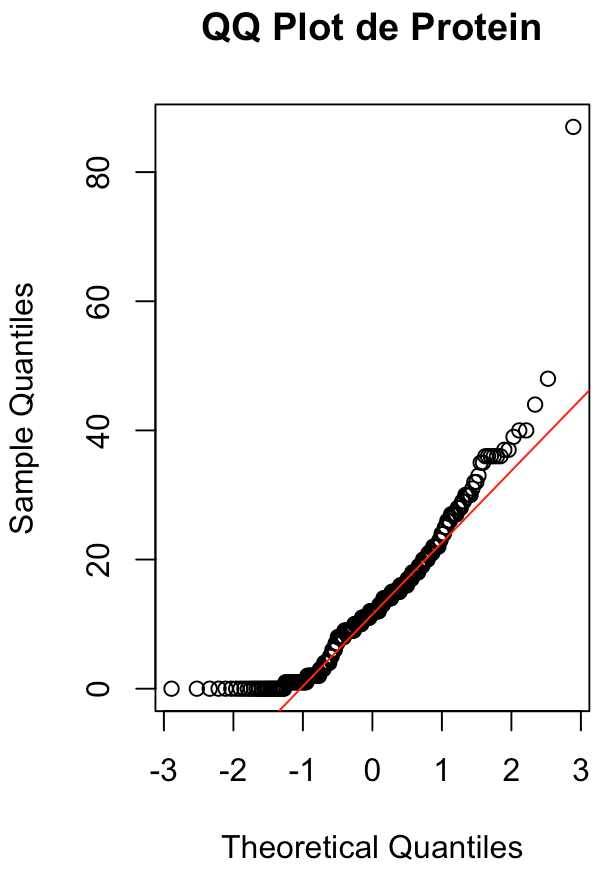

```
par(mfrow = c(1, 1))
```

Los histogramas muestran cómo se distribuyen los valores de calorías y proteínas, las curvas rojas y azules ayudan a visualizar la distribución de una manera más suave y continua.

```
par(mfrow = c(1, 2))  # 1 fila, 3 columnas
qqnorm(data$Calories, main = "QQ Plot de Calories")
qqline(data$Calories, col = "red")

qqnorm(data$Protein, main = "QQ Plot de Protein")
qqline(data$Protein, col = "red")
```






```
par(mfrow = c(1, 3))
```

Las qqplots las ocupamos para comparar las formas en la que los datos se distribuyen comparando con la de una muestra normal teoríca, en ambos casos se nota que en los valores medios de las distribucioness se comportan como una normal, pero en los extremos es donde se notan los outliers, o los valores atípicos de cada distribución

Para los test...

El test Shapiro-Wilks intenta rechazar la hipotesis nula a nuestro nivel de significanza..

La H_0 dice que los datos provienen de una distribución normal.

“Siendo la hipótesis nula que la muestra está distribuida normalmente, si el p-valor es menor a alfa (nivel de significancia) entonces la hipótesis nula es rechazada (se concluye que los datos no vienen de una distribución normal). Si el p-valor es mayor a alfa, se concluye que no se puede rechazar dicha hipótesis.”

Esta cool este vídeo: https://www.youtube.com/watch?v=eh9eYLBecWk (https://www.youtube.com/watch?v=eh9eYLBecWk)

```
shapiro.test(data$Calories)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Calories
## W = 0.91902, p-value = 1.119e-10
```

W = 0.91902: No es cercano a 1, lo que dice que podría haber desviación de la normalidad. p-value = 1.119e-10: Es mucho menor que el nivel de significancia común (0.05) por lo que se rechaza la hipótesis nula.

Lo que que los datos no siguen una distribución normal según el test de Shapiro-Wilk.

```
shapiro.test(data$Protein)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Protein
## W = 0.88573, p-value = 4.445e-13
```