Act4 Explorando Bases

Andrés Villarreal González

2024-08-13

Leyendo los datos

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

Seleccionamos variables

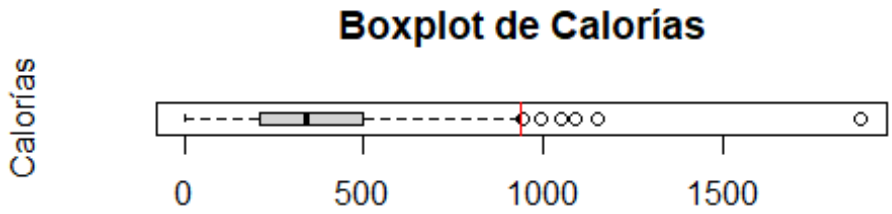
```
calorias = M$Calories
carbohidratos = M$Carbohydrates
```

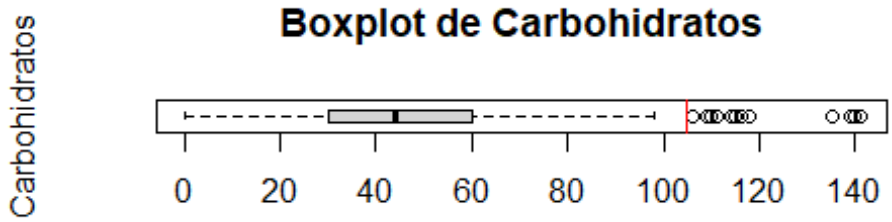
Calcula el rango intercuartílico y los cuartíles

```
# Cálculo del primer cuartil (q1) y tercer cuartil (q3) para Calorías y
Carbohidratos
q1_cal <- quantile(calorias, 0.25)
q3_cal <- quantile(calorias, 0.75)
iqr_cal <- q3_cal - q1_cal  # Rango intercuartílico para Calorías

q1_carb <- quantile(carbohidratos, 0.25)
q3_carb <- quantile(carbohidratos, 0.75)
iqr_carb <- q3_carb - q1_carb  # Rango intercuartílico para Carbohidratos
```

Gráficos de Boxplot

```
# Gráficos de boxplot con líneas en los límites de 1.5 rangos
intercuartílicos
par(mfrow=c(2, 1))  # Dividir la ventana gráfica en una matriz de 2x1

# Boxplot para Calorías
boxplot(calorias, horizontal=TRUE, ylim=c(min(calorias), max(calorias)),
        main = "Boxplot de Calorías", ylab = "Calorías")
abline(v=q3_cal + 1.5*iqr_cal, col="red")  # Línea límite superior en 1.5
IQR

# Boxplot para Carbohidratos
boxplot(carbohidratos, horizontal=TRUE, ylim=c(min(carbohidratos),
max(carbohidratos)),
        main = "Boxplot de Carbohidratos", ylab = "Carbohidratos")
abline(v=q3_carb + 1.5*iqr_carb, col="red")  # Línea límite superior en
1.5 IQR
```





Remover datos atípicos

```
# Remover los datos atípicos que están más allá de 1.5 rangos
intercuartílicos

calorias_sin_outliers <- M$Calories[M$Calories < q3_cal + 1.5*iqr_cal]
carbohidratos_sin_outliers <- M$Carbohydrates[M$Carbohydrates < q3_carb +
1.5*iqr_carb]
```

Resumenes de datos originales y sin outliers

```
# Resúmenes después de quitar los outliers
summary(calorias_sin_outliers)

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

summary(calorias)

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

summary(carbohidratos_sin_outliers)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00

summary(carbohidratos)
```

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

Prueba de normalidad Anderson Darling (datos originales)

```r
library(nortest)

# Prueba de Anderson-Darling para Calorías
ad.test(calorias)
```

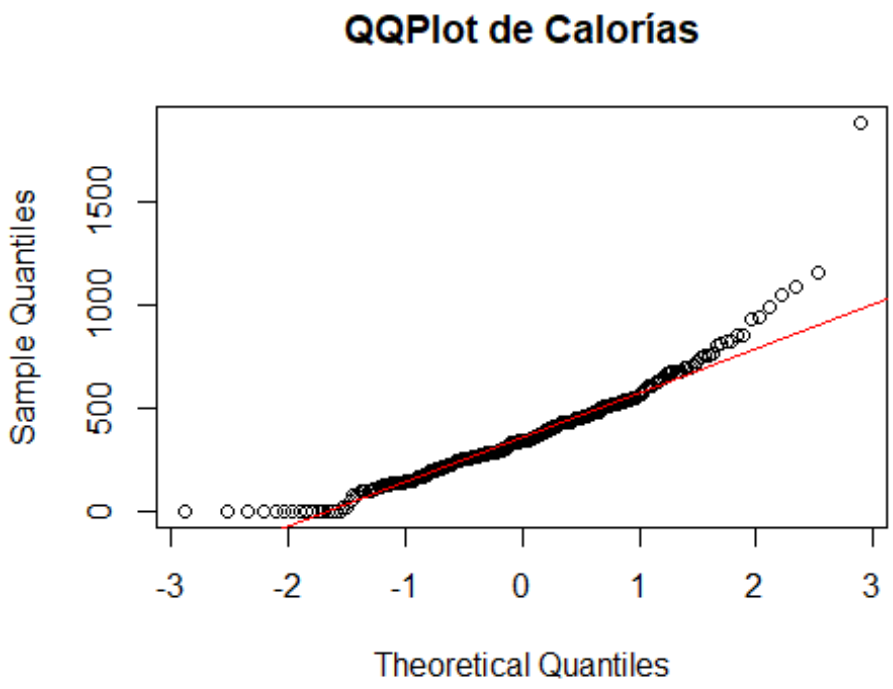
```
##
##  Anderson-Darling normality test
##
## data:  calorias
## A = 2.5088, p-value = 2.369e-06
```

```r
# Prueba de Anderson-Darling para Carbohidratos
ad.test(carbohidratos)
```

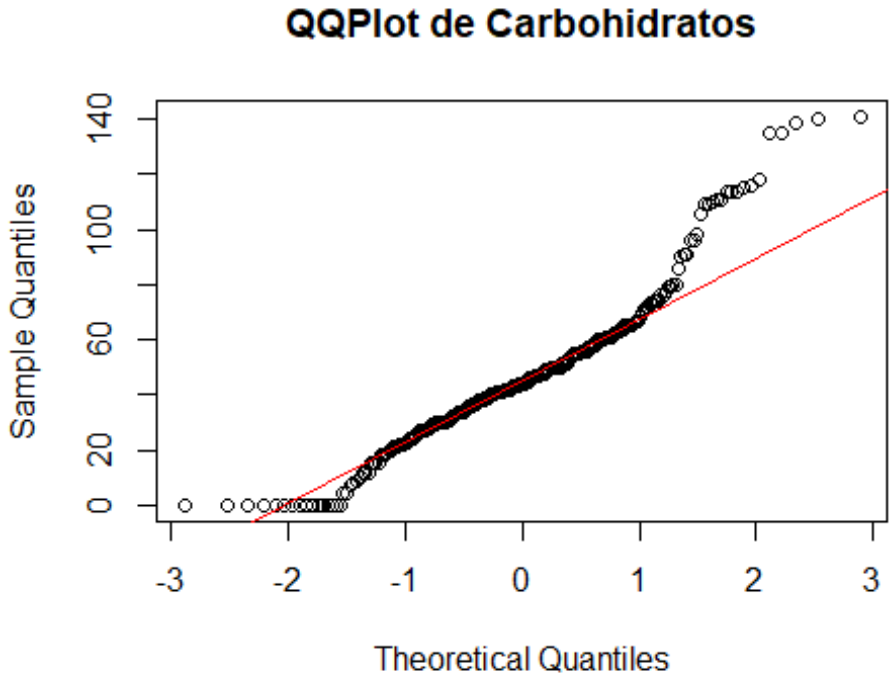
```
##
##  Anderson-Darling normality test
##
## data:  carbohidratos
## A = 4.1402, p-value = 2.547e-10
```

Graficos QQPlot (datos originales)

```r
# QQPlot para Calorías
qqnorm(calorias, main="QQPlot de Calorías")
qqline(calorias, col="red")
```




```r
# QQPlot para Carbohidratos
qqnorm(carbohidratos, main="QQPlot de Carbohidratos")
qqline(carbohidratos, col="red")
```



Sesgo y Curtosis

```r
library(moments)

# Calcular sesgo y curtosis para Calorías
sesgo_cal <- skewness(calorias)
curtosis_cal <- kurtosis(calorias)

# Calcular sesgo y curtosis para Carbohidratos
sesgo_carb <- skewness(carbohidratos)
curtosis_carb <- kurtosis(carbohidratos)

# Mostrar los resultados
cat("Sesgo y curtosis de Calorías:", sesgo_cal, curtosis_cal, "\n")
```

```
## Sesgo y curtosis de Calorías: 1.444105 8.645274
```

```r
cat("Sesgo y curtosis de Carbohidratos:", sesgo_carb, curtosis_carb,
"\n")
```

```
## Sesgo y curtosis de Carbohidratos: 0.9074253 4.357538
```

Para Calorías, debido a la alta curtosis y el sesgo, sería prudente considerar la exclusión de los datos atípicos, ya que estos valores extremos están afectando la distribución. En cambio, para Carbohidratos, la eliminación de datos atípicos puede no ser tan crítica, ya que la distribución es menos afectada por estos valores extremos.

Media, mediana y rango medio

```r
# Calorías
media_cal <- mean(calorias)
mediana_cal <- median(calorias)
rango_medio_cal <- (min(calorias) + max(calorias)) / 2

# Carbohidratos
media_carb <- mean(carbohidratos)
mediana_carb <- median(carbohidratos)
rango_medio_carb <- (min(carbohidratos) + max(carbohidratos)) / 2

# Mostrar las medidas
cat("Calorías - Media:", media_cal, "Mediana:", mediana_cal, "Rango
Medio:", rango_medio_cal, "\n")
```

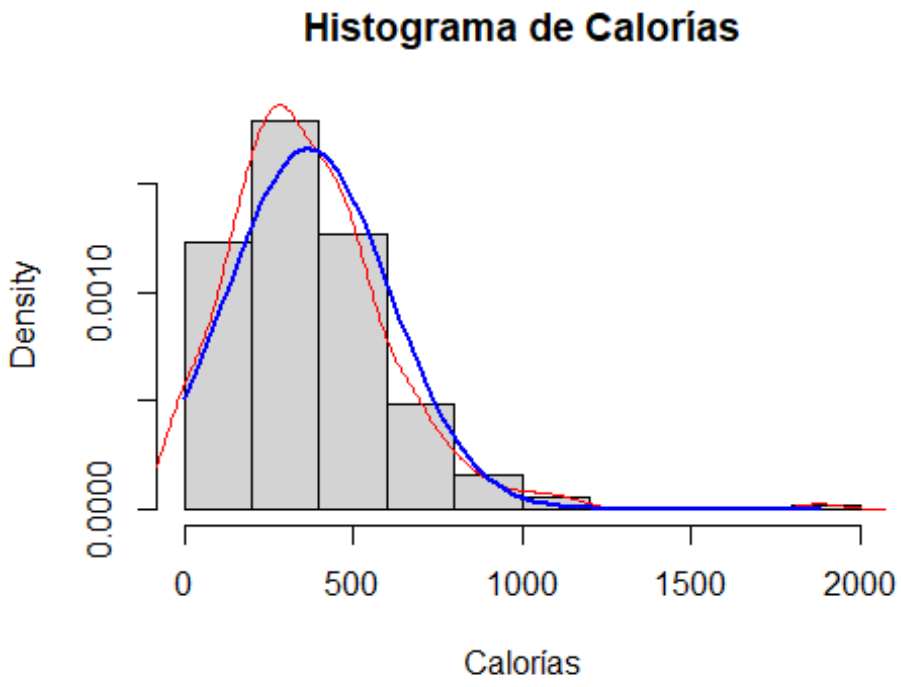
```
## Calorías - Media: 368.2692 Mediana: 340 Rango Medio: 940
```

```r
cat("Carbohidratos - Media:", media_carb, "Mediana:", mediana_carb,
"Rango Medio:", rango_medio_carb, "\n")
```

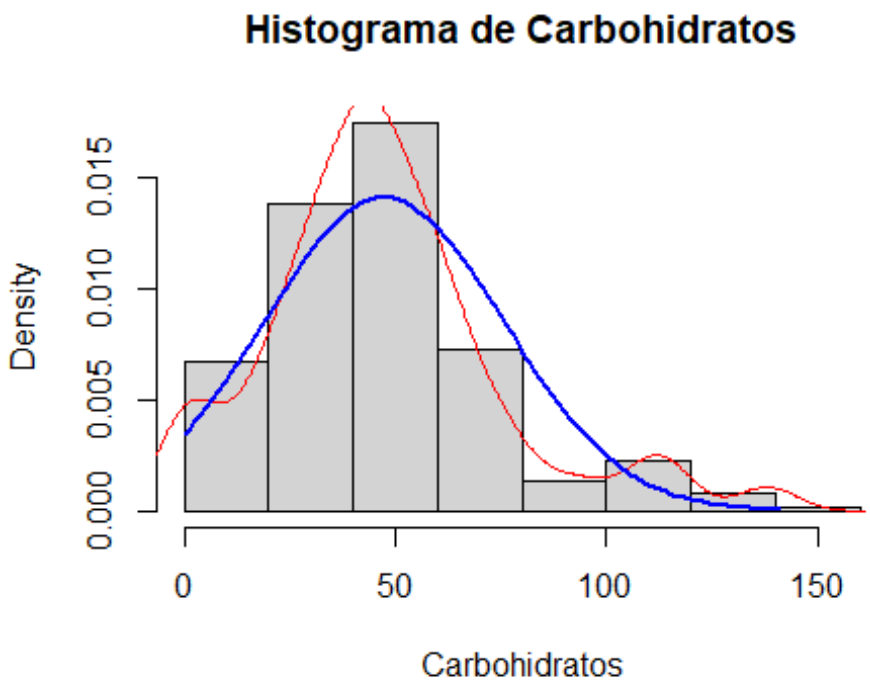
```
## Carbohidratos - Media: 47.34615 Mediana: 44 Rango Medio: 70.5
```

Histogrmas de densidad

```
# Histograma con densidad y curva normal para Calorías
hist(calorias, freq=FALSE, main="Histograma de Calorías",
xlab="Calorías")
lines(density(calorias), col="red")
curve(dnorm(x, mean=mean(calorias), sd=sd(calorias)),
      from=min(calorias), to=max(calorias), add=TRUE, col="blue", lwd=2)
```




```
# Histograma con densidad y curva normal para Carbohidratos
hist(carbohidratos, freq=FALSE, main="Histograma de Carbohidratos",
xlab="Carbohidratos")
lines(density(carbohidratos), col="red")
curve(dnorm(x, mean=mean(carbohidratos), sd=sd(carbohidratos)),
      from=min(carbohidratos), to=max(carbohidratos), add=TRUE,
col="blue", lwd=2)
```

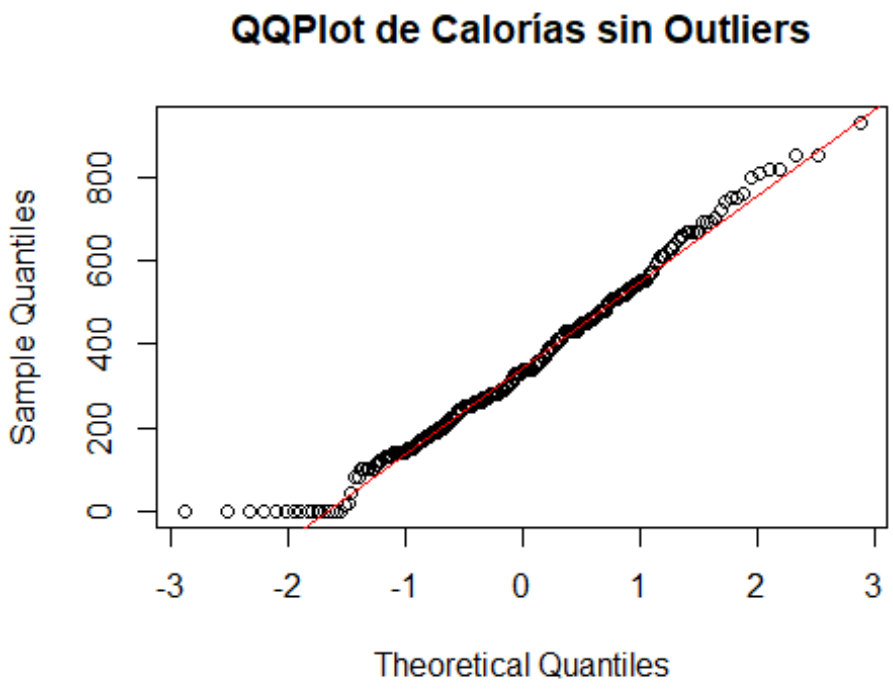


Pruebas de normalidad y QQPlot sin outliers

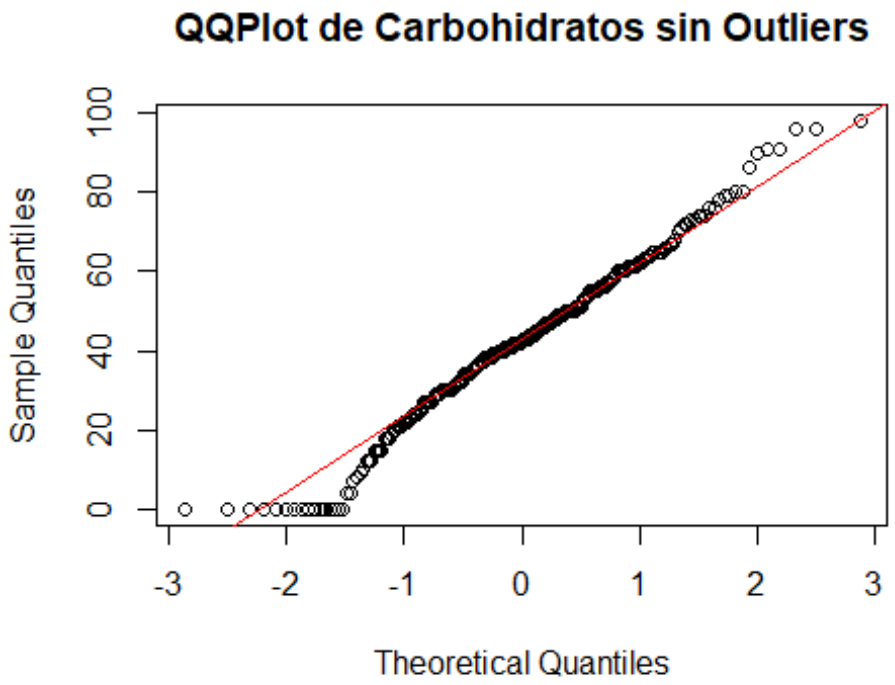
```
# Pruebas de normalidad sin outliers
ad.test(calorias_sin_outliers)
```

```
##
##  Anderson-Darling normality test
##
## data:  calorias_sin_outliers
## A = 0.89786, p-value = 0.02166
```

```
qqnorm(calorias_sin_outliers, main="QQPlot de Calorías sin Outliers")
qqline(calorias_sin_outliers, col="red")
```




```
# Pruebas de normalidad sin outliers
ad.test(carbohidratos_sin_outliers)

##
##  Anderson-Darling normality test
##
## data:  carbohidratos_sin_outliers
## A = 0.74917, p-value = 0.05048

qqnorm(carbohidratos_sin_outliers, main="QQPlot de Carbohidratos sin
Outliers")
qqline(carbohidratos_sin_outliers, col="red")
```



Podemos observar despues de realizar varias pruebas que ambas variables no cuentan con datos normales ya que al realizar la prueba de Anderson-Darling ambos valores p estan debajo del valor de 0.05. Despues de remover datos atipicos y volver a realizar la prueba de normalidad vemos que la variables de calorias sigue teniendo un valor debajo de 0.05 pero la variable de carbohidratos si cuenta con normalidad ya que tiene un valor p de 0.05048 que es mayor a 0.05. Los graficos QQPlot e histogramos tambien son utiles para poder identificar normalidad en los datos pero hacer las pruebas de normalidad siempre será mas exacto.