proba4_ExplorandoBases

Fernanda Pérez

2024-08-13

```
library(moments)
```

1. Baja el archivo de trabajo: datos de McDonaldDownload McDonald

```
M=read.csv("D:/Downloads/mc-donalds-menu.csv")
```

```
Cal=M$Calories
Car=M$Carbohydrates
```

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

Calorias Carbohidratos Proteinas Sodio Azucares (Sugars)

```
#Calorias
#Cal
q1_cal = quantile(Cal, 0.25)
q3_cal = quantile(Cal, 0.75)
ri_cal = IQR(Cal)

#Carbohidratos
#Car
q1_car = quantile(Car, 0.25)
q3_car = quantile(Car, 0.75)
ri_car = IQR(Car)
```

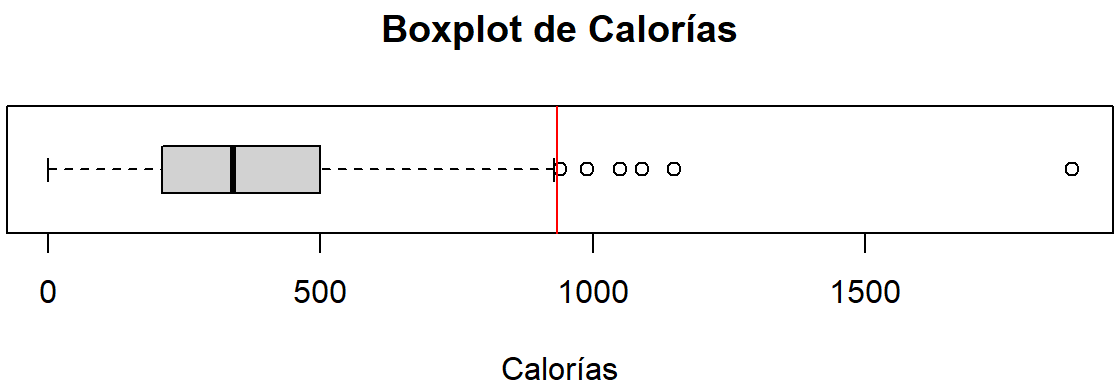
3. Para analizar datos atípicos se te sugiere:

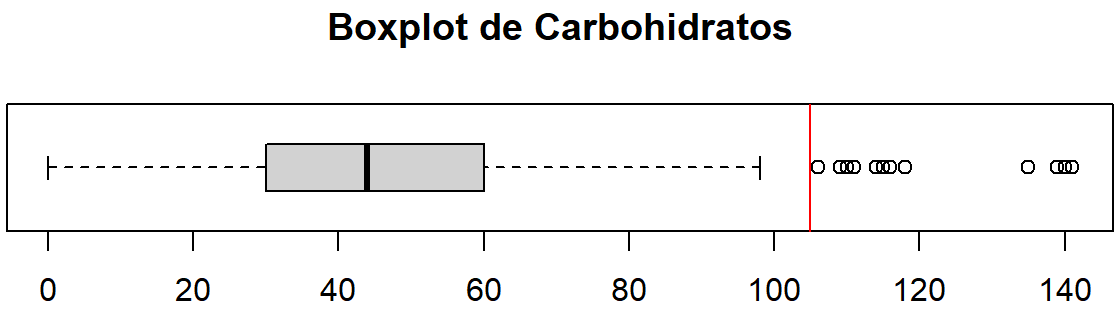
Graficar el diagrama de caja y bigote Calcula el rango intercuartílico y los cuartíles Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio? Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio? Toma una decisión de si conviene o no quitar los datos atípicos (para ello interpreta la variable en el contexto del problema y determina si es necesario quitarlos o no quitarlos)

4. Para analizar normalidad se te sugiere:

Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase) Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable. Compara las medidas de media, mediana y rango medio de cada variable. Realiza el histograma y su distribución teórica de probabilidad (sugerencia, adapta el código: hist(datos,freq=FALSE) lines(density(datos),col="red") curve(dnorm(x,mean=mean(datos,sd=sd(datos)), from=-6, to=6, add=TRUE, col="blue",lwd=2) Identifica cómo influyen los datos atípicos en la normalidad de los datos Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos

```
par(mfrow = c(2, 1))
boxplot(Cal, horizontal = TRUE, main = "Boxplot de Calorías", xlab = "Calorías")
abline(v = c(q1_cal - 1.5 * ri_cal, q3_cal + 1.5 * ri_cal), col = "red")

boxplot(Car, horizontal = TRUE, main = "Boxplot de Carbohidratos", xlab = "Carbohidratos")
abline(v = c(q1_car - 1.5 * ri_car, q3_car + 1.5 * ri_car), col = "red")
```






```
Cal1 = M[M$Calories >= (q1_cal - 1.5 * ri_cal) & M$Calories <= (q3_cal + 1.5 * ri_cal), "Calorie
s"]
Car1 = M[M$Carbohydrates >= (q1_car - 1.5 * ri_car) & M$Carbohydrates <= (q3_car + 1.5 * ri_ca
r), "Carbohydrates"]


library(moments)

# Calorías
shapiro.test(Cal)
```

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

```
skewness(Cal)
```

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

```
kurtosis(Cal)
```

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

```
# Carbohidratos
shapiro.test(Car)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Car
## W = 0.93666, p-value = 3.931e-09
```

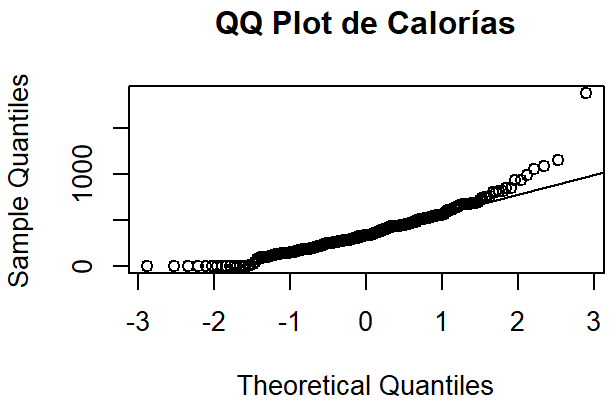
```
skewness(Car)
```

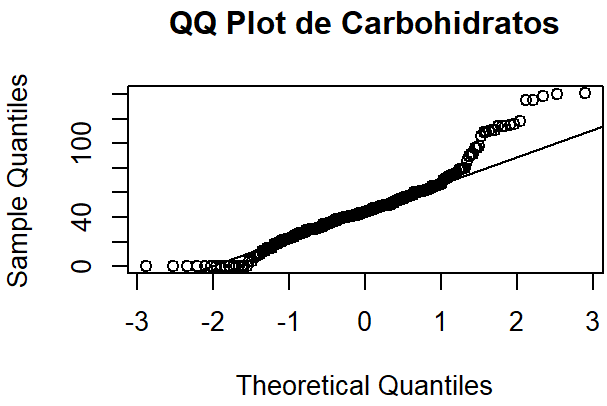
```
## [1] 0.9074253
```

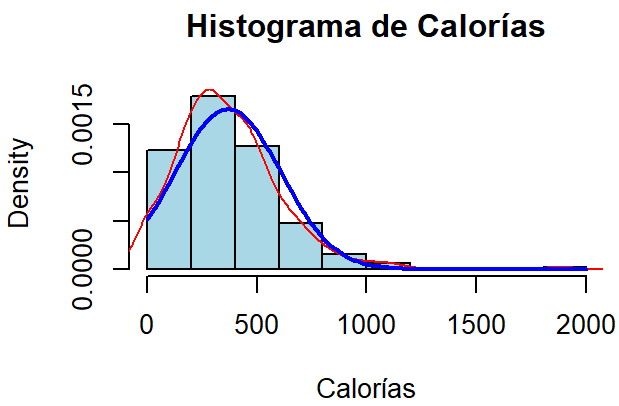
```
kurtosis(Car)
```

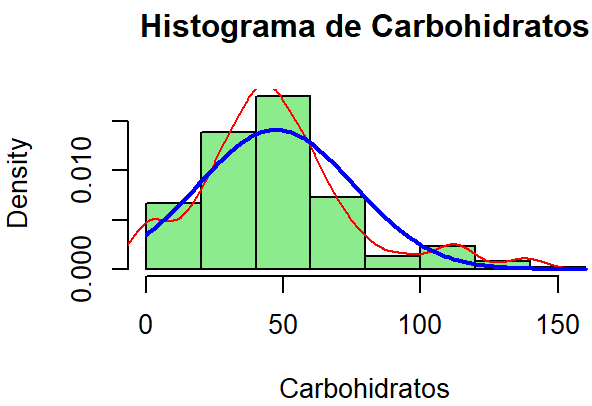
```
## [1] 4.357538
```

```
# QQPlot
par(mfrow = c(2, 2))
qqnorm(Cal, main = "QQ Plot de Calorías")
qqline(Cal)

qqnorm(Car, main = "QQ Plot de Carbohidratos")
qqline(Car)

# Histogramas
hist(Cal, prob = TRUE, main = "Histograma de Calorías", xlab = "Calorías", col = "lightblue")
lines(density(Cal), col = "red")
curve(dnorm(x, mean = mean(Cal), sd = sd(Cal)), add = TRUE, col = "blue", lwd = 2)

hist(Car, prob = TRUE, main = "Histograma de Carbohidratos", xlab = "Carbohidratos", col = "ligh
tgreen")
lines(density(Car), col = "red")
curve(dnorm(x, mean = mean(Car), sd = sd(Car)), add = TRUE, col = "blue", lwd = 2)
```










```
summary(Cal)
```

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

```
summary(Car)
```

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

```
shapiro.test(Cal1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cal1
## W = 0.98059, p-value = 0.001523
```

```
shapiro.test(Car1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Car1
## W = 0.98397, p-value = 0.007726
```

```
# Coeficiente de sesgo  Cal
sesgo_cal = skewness(Cal)
# Coeficiente de curtosis  Cal
curtosis_cal = kurtosis(Cal)

# Coeficiente de sesgo Car
sesgo_car = skewness(Car)
# Coeficiente de curtosis  Car
curtosis_car = kurtosis(Car)


print(paste("Sesgo de Calorías:", sesgo_cal))
```

```
## [1] "Sesgo de Calorías: 1.44410491051015"
```

```
print(paste("Curtosis de Calorías:", curtosis_cal))
```

```
## [1] "Curtosis de Calorías: 8.64527387047867"
```

```
print(paste("Sesgo de Carbohidratos:", sesgo_car))
```

```
## [1] "Sesgo de Carbohidratos: 0.907425282256168"
```

```
print(paste("Curtosis de Carbohidratos:", curtosis_car))
```

```
## [1] "Curtosis de Carbohidratos: 4.3575379316182"
```

```
media_cal = mean(Cal)
mediana_cal = median(Cal)
rango_medio_cal = (min(Cal) + max(Cal)) / 2

# resultados para Cal
print(paste("Media de Calorías:", media_cal))
```

```
## [1] "Media de Calorías: 368.269230769231"
```

```
print(paste("Mediana de Calorías:", mediana_cal))
```

```
## [1] "Mediana de Calorías: 340"
```

```
print(paste("Rango Medio de Calorías:", rango_medio_cal))
```

```
## [1] "Rango Medio de Calorías: 940"
```

```
media_car = mean(Car)
mediana_car = median(Car)
rango_medio_car = (min(Car) + max(Car)) / 2

#resultados para Car
print(paste("Media de Carbohidratos:", media_car))
```

```
## [1] "Media de Carbohidratos: 47.3461538461538"
```

```
print(paste("Mediana de Carbohidratos:", mediana_car))
```

```
## [1] "Mediana de Carbohidratos: 44"
```

```
print(paste("Rango Medio de Carbohidratos:", rango_medio_car))
```

```
## [1] "Rango Medio de Carbohidratos: 70.5"
```

3.

a. Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio?

- Calorias: Sí los hay ya que en el boxplot se muestran varios datos atípicos a la derecha del rango intercuartílico, lo cual nos indica que que hay elementos con valores de calorías significativamente altos en comparación con el resto.
- Carbohidratos: Si, también podemos observar datos atípicos en el boxplot.

b. Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio?

- Calorias: Al calcular la media y desviación estándar, si uso el criterio de 3 desviaciones estándar, algunos valores extremos superan este umbral, con lo cual sabemos que hay datos atípicos.
- Carbohidratos: Aquí también existen valores que se alejan más de 3 desviaciones estándar de la media con lo cual igual sabemos que hay datos atípicos.

c. Toma una decisión de si conviene o no quitar los datos atípicos (para ello interpreta la variable en el contexto del problema y determina si es necesario quitarlos o no quitarlos)

Quitarlos o no dependerá del contexto del análisis. Si los datos atípicos son errores de medición o no son representativos del grupo general de alimentos, considero que si vale la pena eliminarlos para obtener una distribución más normalizada y mejores resultados. Pero si estos valores atípicos si representan productos

relevantes si seria mejor dejarlos para que el análisis refleje un resultado más real.

4.

Segun la prueba de Shapiro-Wik las variables de calorias y carbohidratos no siguen una distribución normal ya que el p-value de ambas es muy bajo por cual se desvia significativamente de distribución la normal.

En los 2 QQ plots, los puntos se desvían de la línea de normalidad en los extremos, lo vemos más en las colas. Este comportamiento se da porque existen datos atípicos en la muestra, y estos datos atipicos distorsionan la alineación de los datos con una distribución normal teórica.

Histogramas:

- Calorías: El histograma de calorias nos muestra una distribución asimétrica con una cola larga. La distribución se aleja de la normalidad.
- Carbohidratos: El histograma de carbohidratos es un poco menos asimétrico que el de calorías, pero aún nos muestra una forma irregular con posibles datos atípicos hacia los extremos. Esto igualmente sugiere una desviación de la normalidad debido a la presencia de valores extremos.

4.4 Compara las medidas de media, mediana y rango medio de cada variable.

- Calorías: Sesgo (1.44): Valor positivo nos indica una distribución asimétrica hacia la derecha, con valores altos que estiran la cola. Curtosis (8.64): Al ser alta nos sugiere que tiene colas largas y picos pronunciados, indicando presencia de valores que salen del rango esperado. Media (368.27), Mediana (340), Rango Medio (940): Como la media es un poco mayor que la mediana, refleja el sesgo. Y como el rango medio es mucho mayor, nos indicaa valores extremos.
- Carbohidratos: Sesgo (0.91): Seguimos teniendo un sesgo positivo pero ahora es menor por lo cual tenemos una ligera asimetría hacia la derecha. Curtosis (4.36): Siendo moderadamente alta, nos indica que tambien hay valores que se salen del rango pero menos extremos. Media (47.35), Mediana (44), Rango Medio (70.5): La media es mayor que la mediana, inidicando una ligera asimetría con algunos valores altos.

Como las 2 variables presentan sesgo positivo y una curtosis alta,nos indica la presencia de valores atípicos que afectan la distribución y la normalidad de los datos.