Actividad 4: Explorando Bases

Daniela Jiménez Téllez

2024-08-13

Importación de librerías

```
library(nortest)
library(moments)
```

1. Baja el archivo de trabajo: datos de McDonald

```
data <- read.csv("mc-donalds-menu.csv")
```

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

- Calorias
- Carbohidratos
- Proteinas
- Sodio
- Azucares (Sugars)

En este trabajo se usarán las variables: “Carbohidratos” y “Azúcares”.

```
carbohidratos <- data$Carbohydrates
azucares <- data$Sugars
```

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

### ANÁLISIS CARBOHIDRATOS

## Análisis de datos atípicos

boxplot(carbohidratos, main = "Boxplot de Carbohidratos", col = "pink")

# Cuartiles y rango intercuartílico

Q1 <- quantile(carbohidratos, 0.25)
Q2 <- quantile(carbohidratos, 0.50)
Q3 <- quantile(carbohidratos, 0.75)
cat("Los cuartiles son:\n")
```

```
## Los cuartiles son:
```

```
print(Q1)
```

```
## 25%
##  30
```

```
print(Q2)
```

```
## 50%
##  44
```

```
print(Q3)
```

```
## 75%
##  60
```

```
ri <- IQR(carbohidratos)
cat("El rango intercuartílico es:\n")
```

```
## El rango intercuartílico es:
```

```
print(ri)
```

```
## [1] 30
```

```
# Cota de 1.5 rangos

cota_inf <- Q1 - 1.5 * ri
cota_sup <- Q3 + 1.5 * ri

valores_atp_15 <- carbohidratos[carbohidratos < cota_inf | carbohidratos > cota_sup]

cat("La cota de 1.5 rangos intercuartílicos para los valores atípicos es:\n")
```

```
## La cota de 1.5 rangos intercuartílicos para los valores atípicos es:
```

```
print(valores_atp_15)
```

```
##  [1] 111 116 110 115 118 111 109 135 114 140 114 141 109 135 139 106 114
```

```
# Cota de 3 desviaciones estándar

mean_carbo <- mean(carbohidratos)
sd_carbo <- sd(carbohidratos)
cota_inf_3std <- mean_carbo - 3 * sd_carbo
cota_sup_3std <- mean_carbo + 3 * sd_carbo
valores_atp_3 <- carbohidratos[carbohidratos < cota_inf_3std | carbohidratos > cota_sup_3std]

cat("La cota de 3 desviaciones estándar al rededor de la media para los valores atípicos es:\n")
```

```
## La cota de 3 desviaciones estándar al rededor de la media para los valores atípicos es:
```

```
print(valores_atp_3)
```

```
## [1] 135 140 141 135 139
```

```
## Análisis de normalidad

# Pruebas de normalidad univariada

shapiro_test <- shapiro.test(carbohidratos)
cat("Los resultados del Shapiro-Wilk test son:\n")
```

```
## Los resultados del Shapiro-Wilk test son:
```

```
print(shapiro_test)
```

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

```
# Gráficas (qqplot)

qqnorm(carbohidratos, main = "QQ Plot de Carbohidratos", col = "pink")
qqline(carbohidratos, col = "black")

# Coeficiente de sesgo y curtosis

sesgo_carbo <- skewness(carbohidratos)
curtosis_carbo <- kurtosis(carbohidratos)
cat("El coeficiente de sesgo es:\n")
```

```
## El coeficiente de sesgo es:
```

```
print(sesgo_carbo)
```

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

```
cat("El coeficiente de curtosis es:\n")
```

```
## El coeficiente de curtosis es:
```

```
print(curtosis_carbo)
```

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

```
# Media, mediana y rango medio

media_carbo <- mean(carbohidratos)
mediana_carbo <- median(carbohidratos)
rango_medio_carbo <- (min(carbohidratos) + max(carbohidratos)) / 2

cat("La media es:\n")
```

```
## La media es:
```

```
print(media_carbo)
```

```
## [1] 47.34615
```

```
cat("La mediana es:\n")
```

```
## La mediana es:
```

```
print(mediana_carbo)
```

```
## [1] 44
```

```
cat("El rango medio es:\n")
```

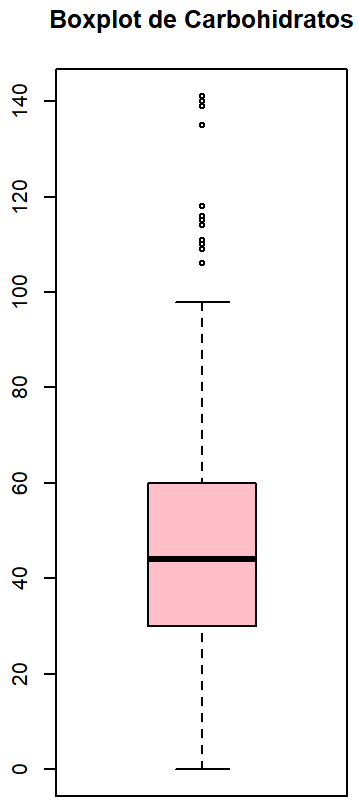
```
## El rango medio es:
```

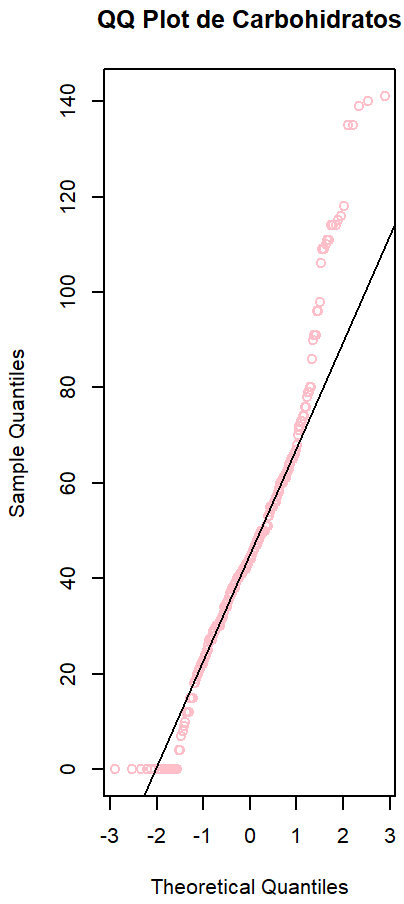
```
print(rango_medio_carbo)
```

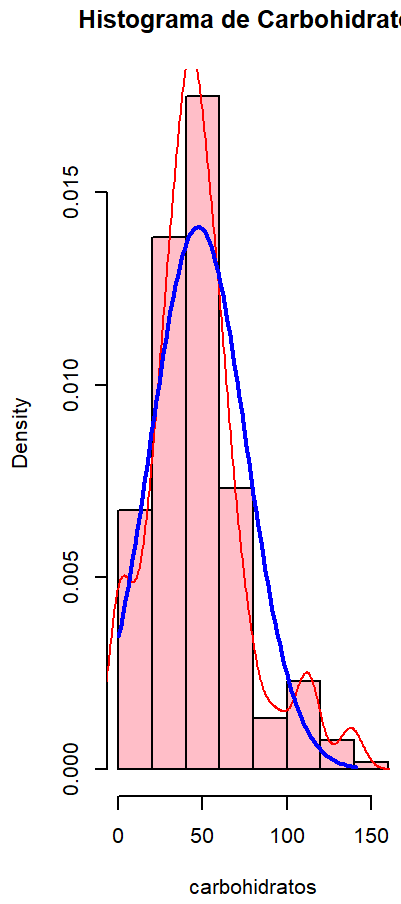
```
## [1] 70.5
```

```
# Histograma

hist(carbohidratos, freq = FALSE, main = "Histograma de Carbohidratos", col = "pink")
lines(density(carbohidratos),col = "red")
curve(dnorm(x, mean = mean(carbohidratos), sd = sd(carbohidratos)), from = min(carbohidratos), t
o = max(carbohidratos), add = TRUE, col = "blue", lwd = 2)
```







De acuerdo a los resultados anteriores, en cuanto al análisis de valores atípicos:

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

Sí hay datos atípicos. Como se puede observar, de las 260 filas hay 17 outliers, los cuales tienen un valor de 100+, cuando el rango intercuartílico es de 30 y la media de 47.35, lo que nos dice que están muy alejados del resto.

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

Sí hay. A comparación de la cota de 1.5 rangos, con el criterio de 3 desviaciones estándar podemos ver menos valores atípicos, de los cuales muchos coinciden. Igualmente, en el diagrama de boxplot se puede notar de una manera más visual.

Habiendo dicho esto, puedo decir que para saber si quitar o no los datos atípicos se necesitaría más contexto sobre el problema que se quiere analizar. Viendolo más por encima, dependiendo del criterio hay menos valores, los cuales no forman gran porcentaje del total de datos, así que no debería ser muy difícil parcharlos para así poder conservarlos.

Por otro lado, para el análisis de normalidad se tienen las siguientes interpretaciones:

1. Identifica cómo influyen los datos atípicos en la normalidad de los datos

La manera en la que influyen los datos atípicos en la normalidad de los datos es que si estos están muy dispersos de la línea de referencia (qqline) quiere decir que es posible que nuestros datos no se comporten como una normal. Igualmente, estos pueden provocar que los datos parezcan “menos normales” de como lo serían si no hubieran datos atípicos.

2. Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos

En la gráfica del QQ Plot podemos ver que los datos siguen la línea por la mayor parte de pasos; sin embargo, un poco antes del final se empiezan a dispersar. Esto nos indica que los datos cuentan con valores atípicos, lo que puede interferir en nuestra interpretación de la gráfica ya que no se podría saber si los datos siguen el comportamiento de una distribución normal o no.

Por otro lado, en cuanto al histograma se pueden ver que hay outliers en los datos. Igualmente, que la línea de densidad empírica y la línea de la distribución normal teórica no son iguales, y al final se empiezan a dispersar entre sí, lo que nos dice que nuestros datos no son normales, confirmando lo que vimos en el QQ Plot.

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

### ANÁLISIS AZUCARES

## Análisis de datos atípicos

boxplot(azucares, main = "Boxplot de Azúcares", col = "lightblue")

# Cuartiles y rango intercuartílico

Q1 <- quantile(azucares, 0.25)
Q2 <- quantile(azucares, 0.50)
Q3 <- quantile(azucares, 0.75)
cat("Los cuartiles son:\n")
```

```
## Los cuartiles son:
```

```
print(Q1)
```

```
##  25%
## 5.75
```

```
print(Q2)
```

```
##  50%
## 17.5
```

```
print(Q3)
```

```
## 75%
##  48
```

```
ri <- IQR(azucares)
cat("El rango intercuartílico es:\n")
```

```
## El rango intercuartílico es:
```

```
print(ri)
```

```
## [1] 42.25
```

```
# Cota de 1.5 rangos

cota_inf <- Q1 - 1.5 * ri
cota_sup <- Q3 + 1.5 * ri

valores_atp_15 <- azucares[azucares < cota_inf | azucares > cota_sup]

cat("La cota de 1.5 rangos intercuartílicos para los valores atípicos es:\n")
```

```
## La cota de 1.5 rangos intercuartílicos para los valores atípicos es:
```

```
print(valores_atp_15)
```

```
## [1] 123 120 115 128
```

```
# Cota de 3 desviaciones estándar

mean_az <- mean(azucares)
sd_az <- sd(azucares)
cota_inf_3std <- mean_az - 3 * sd_az
cota_sup_3std <- mean_az + 3 * sd_az
valores_atp_3 <- azucares[azucares < cota_inf_3std | azucares > cota_sup_3std]

cat("La cota de 3 desviaciones estándar al rededor de la media para los valores atípicos es:\n")
```

```
## La cota de 3 desviaciones estándar al rededor de la media para los valores atípicos es:
```

```
print(valores_atp_3)
```

```
## [1] 123 120 128
```

```
## Análisis de normalidad

shapiro_test <- shapiro.test(azucares)
cat("Los resultados del Shapiro-Wilk test son:\n")
```

```
## Los resultados del Shapiro-Wilk test son:
```

```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  azucares
## W = 0.87708, p-value = 1.269e-13
```

```
# Pruebas de normalidad univariada

# Gráficas (qqplot)

qqnorm(azucares, main = "QQ Plot de Azucares", col = "lightblue")
qqline(azucares, col = "black")

# Coeficiente de sesgo y curtosis

sesgo_az <- skewness(azucares)
curtosis_az <- kurtosis(azucares)
cat("El coeficiente de sesgo es:\n")
```

```
## El coeficiente de sesgo es:
```

```
print(sesgo_az)
```

```
## [1] 1.025977
```

```
cat("El coeficiente de curtosis es:\n")
```

```
## El coeficiente de curtosis es:
```

```
print(curtosis_az)
```

```
## [1] 3.487744
```

```
# Media, mediana y rango medio

media_az <- mean(azucares)
mediana_az <- median(azucares)
rango_medio_az <- (min(azucares) + max(azucares)) / 2

cat("La media es:\n")
```

```
## La media es:
```

```
print(media_az)
```

```
## [1] 29.42308
```

```
cat("La mediana es:\n")
```

```
## La mediana es:
```

```
print(mediana_az)
```

```
## [1] 17.5
```

```
cat("El rango medio es:\n")
```

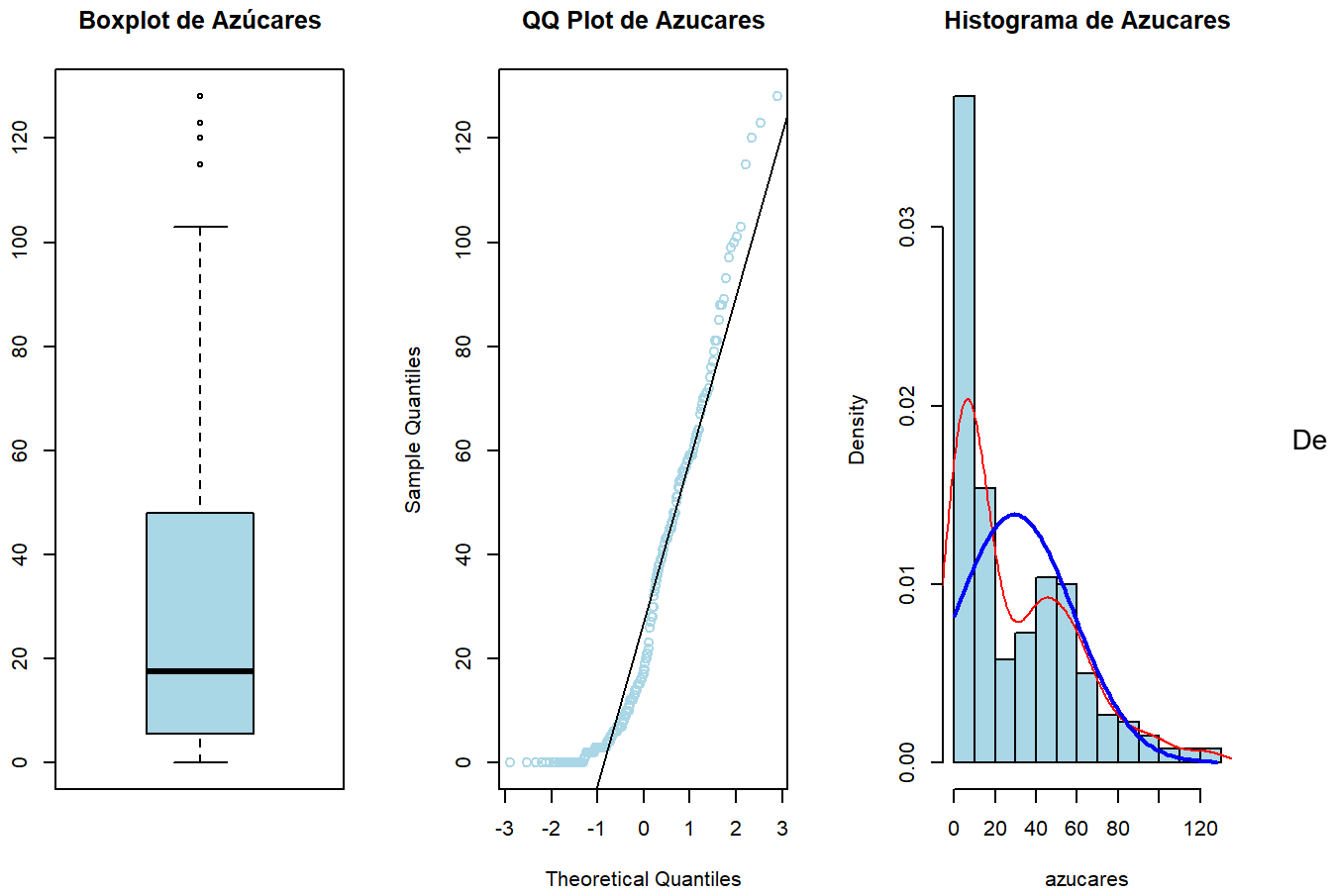
```
## El rango medio es:
```

```
print(rango_medio_az)
```

```
## [1] 64
```

```
# Histograma

hist(azucares, freq = FALSE, main = "Histograma de Azucares", col = "lightblue")
lines(density(azucares),col = "red")
curve(dnorm(x, mean = mean(azucares), sd = sd(azucares)), from = min(azucares), to = max(azucare
s), add = TRUE, col = "blue", lwd = 2)
```



De acuerdo a los resultados anteriores, en cuanto al análisis de valores atípicos:

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

> A pesar de que hay muy pocos, sí hay datos atípicos, los cuales son: 123, 120, 115, 128. Estos son muy altos a comparación del resto de los valores de la columna.

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

> Igual que en el caso anterior, la cota de 3 desviaciones estándar nos regresa menos valores atípicos. En este caso solo hubieron 3.

Habiendo dicho esto, puedo decir que para saber si quitar o no los datos atípicos se necesitaría más contexto sobre el problema que se quiere analizar. Viendolo más por encima, dependiendo del criterio hay menos valores, los cuales no forman gran porcentaje del total de datos, así que no debería ser muy difícil parcharlos para así poder conservarlos.

Por otro lado, para el análisis de normalidad se tienen las siguientes interpretaciones:

1. Identifica cómo influyen los datos atípicos en la normalidad de los datos

La manera en la que influyen los datos atípicos en la normalidad de los datos es que si estos están muy dispersos de la línea de referencia (qqline) quiere decir que es posible que nuestros datos no se comporten como una normal. Igualmente, estos pueden provocar que los datos parezcan “menos normales” de como lo serían si no hubieran datos atípicos.

2. Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos

En la gráfica de QQ Plot podemos observar que al prinipio los datos se comportan como una distribución normal, sin embargo, mientras va avanzando estos empiezan a dispersarse, que es lo que provocan los outliers. Se asume que si los datos estuvieran limpios y todos estuvieran dentro del mismo rango, estos se comportarían como una normal.

Por otro lado, en el histograma se puede ver que al igual que con Carbohidratos, la variable Azucares también tiene valores atípicos, y que nuestros datos no se comportan como una normal debido a que la línea de densidad empírica y la de la distribución normal teórica son diferentes entre sí.