Explorando bases

Luis Maximiliano López Ramírez

2024-08-13

```
# Especificar el nuevo directorio
nuevo_directorio <- "C:/Users/luism/Escritorio/Documetos_2/Actividades Concentración"

# Cambiar al nuevo directorio
setwd(nuevo_directorio)
```

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

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

Calorias Carbohidratos Proteinas Sodio Azucares (Sugars)

Las variables serán Calories y Carbohydrates

```
# Carga los datos desde el archivo datosRes.csv
datos <- read.csv("mc-donalds-menu.csv")
X_calories <- datos$Calories
X_carbohydrates <- datos$Carbohydrates
```

 # 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)

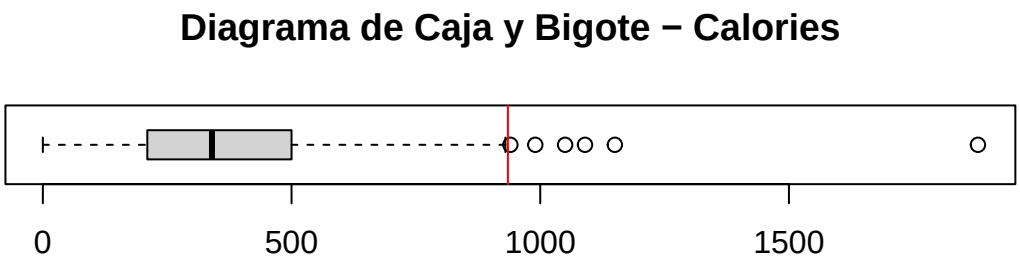
```
# Función para calcular y graficar
procesar_variable <- function(X, nombre) {
  # Cuantiles y rango intercuartílico
  q1 <- quantile(X, 0.25)
  q3 <- quantile(X, 0.75)
  ri <- q3 - q1

  # Imprimir cuartiles y rango intercuartílico
  cat("\n", nombre, "\n")
  cat("Cuartil 1 (Q1):", q1, "\n")
  cat("Cuartil 3 (Q3):", q3, "\n")
  cat("Rango Intercuartílico (RI):", ri, "\n")

  # Gráfico de caja y bigote con línea en el límite de los datos atípicos
  boxplot(X, horizontal = TRUE, ylim = c(min(X), max(X)), main = paste("Diagrama de Caja y Bigote -", n
  abline(v = q3 + 1.5 * ri, col = "red")
```

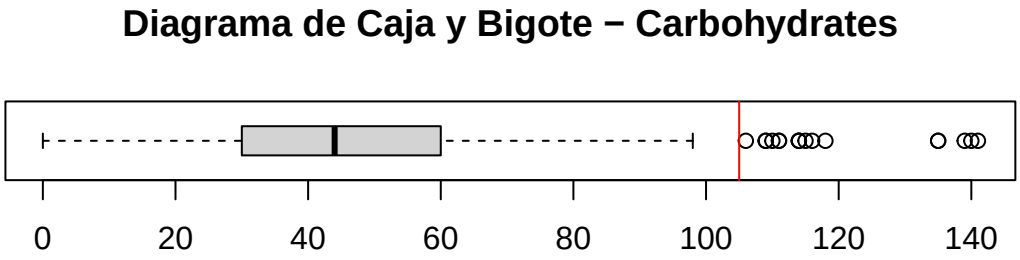
```r
  # Filtrar datos atípicos según 1.5 RI
  X_filtrado <- X[X < q3 + 1.5 * ri]

  # Identificar datos atípicos según 3 desviaciones estándar
  media <- mean(X)
  desviacion_estandar <- sd(X)
  lim_inf <- media - 3 * desviacion_estandar
  lim_sup <- media + 3 * desviacion_estandar
  atipicos_3sd <- X[X < lim_inf | X > lim_sup]

  cat("Cota Inferior (3SD):", lim_inf, "\n")
  cat("Cota Superior (3SD):", lim_sup, "\n")
  cat("Datos atípicos (3SD):", atipicos_3sd, "\n")
}

# Configuración de la matriz de gráficos 2x1
par(mfrow = c(2, 1))

# Procesar ambas variables
procesar_variable(X_calories, "Calories")
```

```
##
##  Calories
## Cuartil 1 (Q1): 210
## Cuartil 3 (Q3): 500
## Rango Intercuartílico (RI): 290

## Cota Inferior (3SD): -352.5404
## Cota Superior (3SD): 1089.079
## Datos atípicos (3SD): 1090 1150 1880
```

```r
procesar_variable(X_carbohydrates, "Carbohydrates")
```

```
##
##  Carbohydrates
## Cuartil 1 (Q1): 30
## Cuartil 3 (Q3): 60
## Rango Intercuartílico (RI): 30
```






```
## Cota Inferior (3SD): -37.41054
## Cota Superior (3SD): 132.1028
## Datos atípicos (3SD): 135 140 141 135 139
```

Si conviene quitarlos ya que puede afectar a procesos futuros además no haber tantos datos atípicos.

 # 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

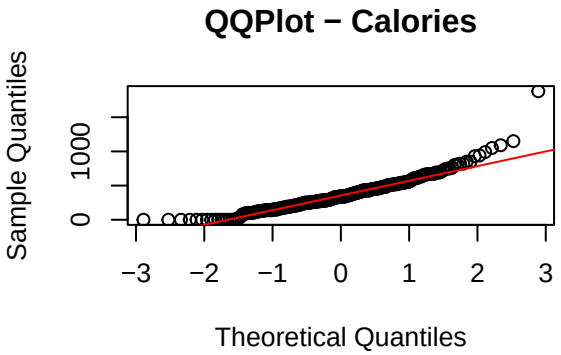
```r
# Cargar las librerías necesarias
library(nortest)  # Para la prueba de Anderson-Darling
library(moments)  # Para calcular sesgo y curtosis

# Función para realizar análisis de normalidad
analizar_variable <- function(X, nombre) {
  cat("\n--- Análisis para", nombre, "---\n")

  # 1. Prueba de normalidad (Anderson-Darling)
```

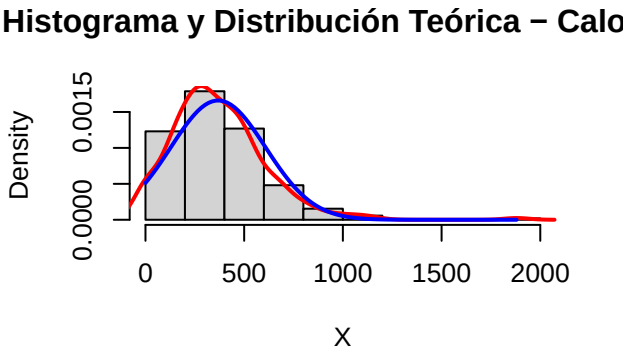
```r
  normalidad <- ad.test(X)
  cat("Prueba de Anderson-Darling:\n")
  print(normalidad)

  # 2. Gráfico QQPlot
  qqnorm(X, main = paste("QQPlot -", nombre))
  qqline(X, col = "red")

  # 3. Calcular coeficiente de sesgo y curtosis
  sesgo <- skewness(X)
  curtosis <- kurtosis(X)
  cat("Coeficiente de Sesgo:", sesgo, "\n")
  cat("Coeficiente de Curtosis:", curtosis, "\n")

  # 4. Comparar media, mediana y rango medio
  media <- mean(X)
  mediana <- median(X)
  rango_medio <- (max(X) + min(X)) / 2
  cat("Media:", media, "\n")
  cat("Mediana:", mediana, "\n")
  cat("Rango Medio:", rango_medio, "\n")

  # 5. Histograma y distribución teórica
  hist(X, freq = FALSE, main = paste("Histograma y Distribución Teórica -", nombre), col = "lightgray")
  lines(density(X), col = "red", lwd = 2)  # Distribución empírica
  curve(dnorm(x, mean = mean(X), sd = sd(X)),
        from = min(X), to = max(X), add = TRUE, col = "blue", lwd = 2)  # Distribución teórica

}

# Configuración de la matriz de gráficos 2x2
par(mfrow = c(2, 2))

# Analizar ambas variables
analizar_variable(X_calories, "Calories")
```

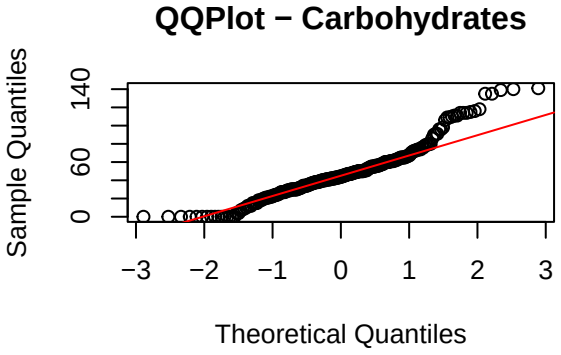
```
##
## --- Análisis para Calories ---
## Prueba de Anderson-Darling:
##
##  Anderson-Darling normality test
##
## data:  X
## A = 2.5088, p-value = 2.369e-06

## Coeficiente de Sesgo: 1.444105
## Coeficiente de Curtosis: 8.645274
## Media: 368.2692
## Mediana: 340
## Rango Medio: 940
```

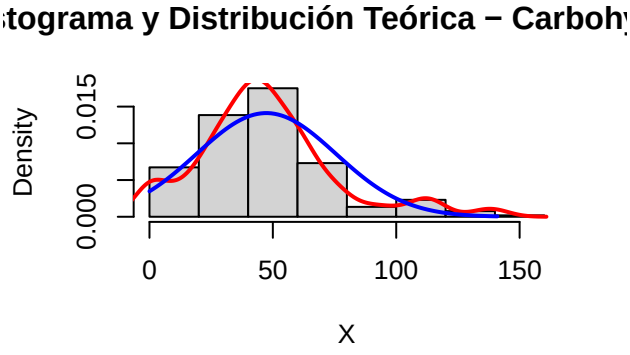
```r
analizar_variable(X_carbohydrates, "Carbohydrates")
```

```
##
## --- Análisis para Carbohydrates ---
## Prueba de Anderson-Darling:
##
##  Anderson-Darling normality test
##
## data:  X
## A = 4.1402, p-value = 2.547e-10

## Coeficiente de Sesgo: 0.9074253
## Coeficiente de Curtosis: 4.357538
## Media: 47.34615
## Mediana: 44
## Rango Medio: 70.5
```









Comentarios sobre gráficos y normalidad

La prueba de Anderson-Darling sugiere que la variable no sigue una distribución normal

Un sesgo positivo indica una distribución asimétrica a la derecha, mientras que un sesgo negativo indica lo contrario.

Una curtosis mayor a 3 sugiere una distribución leptocúrtica (picos más altos), mientras que una menor a 3 sugiere una distribución platicúrtica (picos más bajos).

La comparación de la media, mediana y rango medio puede indicar la simetría de la distribución. Si son similares, es un buen indicio de normalidad.

También se puede pbservar que los datos atípicos influyen en las colas de las gráficas hechas con QQPlot y QQline para ambas gráficas.