Explorando_Bases

Arturo

2023-08-30

Autor: Arturo Garza Campuzano

Matrícula: A00828096

Explorando Bases

Importar módulos

```
# Instalacion y carga de paquetes
if (!require(nortest) || !require(e1071) || !require(moments)) {
  install.packages("moments")
  install.packages("e1071")
  install.packages('nortest')
}
```

```
## Loading required package: nortest
```

```
## Loading required package: e1071
```

```
## Loading required package: moments
```

```
##
## Attaching package: 'moments'
```

```
## The following objects are masked from 'package:e1071':
##
##     kurtosis, moment, skewness
```

```
library(moments)
library(e1071)
library(nortest)
```

Cargar datos

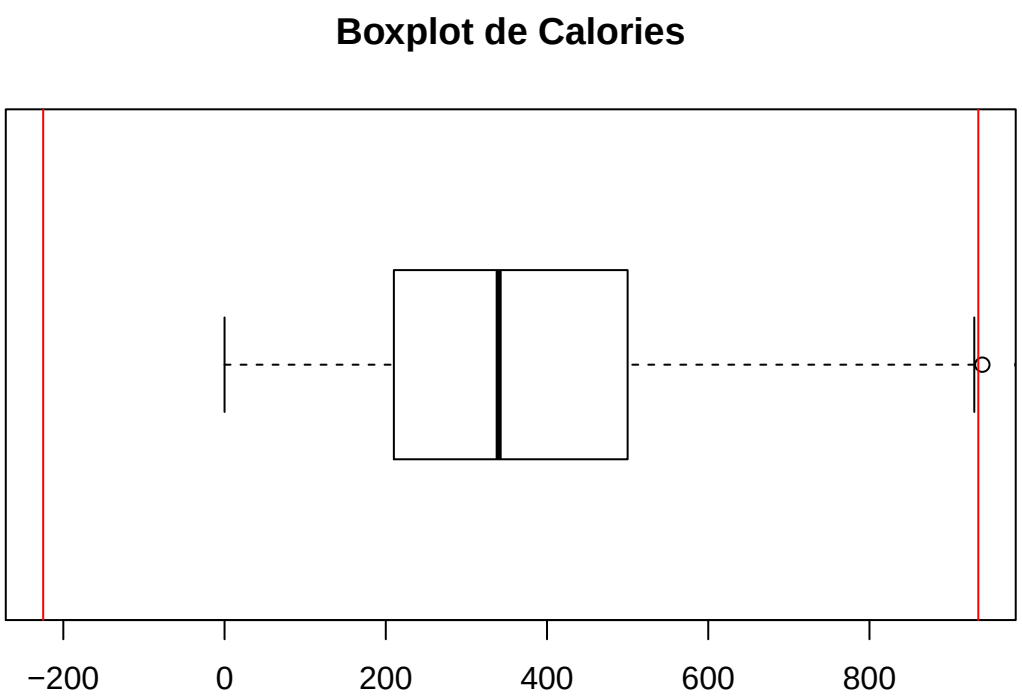
Se eligen las variables *Calories* y *Carbohydrates* para realizar un análisis en cuanto a sus **datos atípicos** y **normalidad**.

```
# Lectura del archivo (cambiar routeo si es necesario)
M = read.csv("mc_donalds_menu_1.csv")

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

Datos atípicos

Se exploran y eliminan los **datos atípicos** de las variables seleccionadas. En este caso, se considera un **dato atípico** aquel que esté por debajo de $Q1 - 1.5 \times IRQ$ o por encima de $Q3 + 1.5 \times IRQ$.

```
# Explorar y quitar los datos atípicos
datos_atipicos <- function(variable, variable_nombre) {
  q1 <- quantile(variable, 0.25)
  q3 <- quantile(variable, 0.75)
  ri <- q3 - q1

  # Boxplot
  boxplot(variable, horizontal=TRUE, main = paste("Boxplot de", variable_nombre), ylim=c(q1 - 1.5 * ri,
  abline(v=q1 - 1.5 * ri, col="red")
  abline(v=q3 + 1.5 * ri, col="red")

  cat("Resumen de",variable_nombre,"con datos atípicos\n")
  cat(capture.output(summary(variable)), sep = "\n")

  variable <- variable[variable <= q3 + 1.5 * ri]
  variable <- variable[variable >= q1 - 1.5 * ri]

  cat("Resumen de",variable_nombre,"sin datos atípicos\n")
  cat(capture.output(summary(variable)), sep = "\n")

  return(variable)
}

calorias = datos_atipicos(calorias, 'Calories')
```

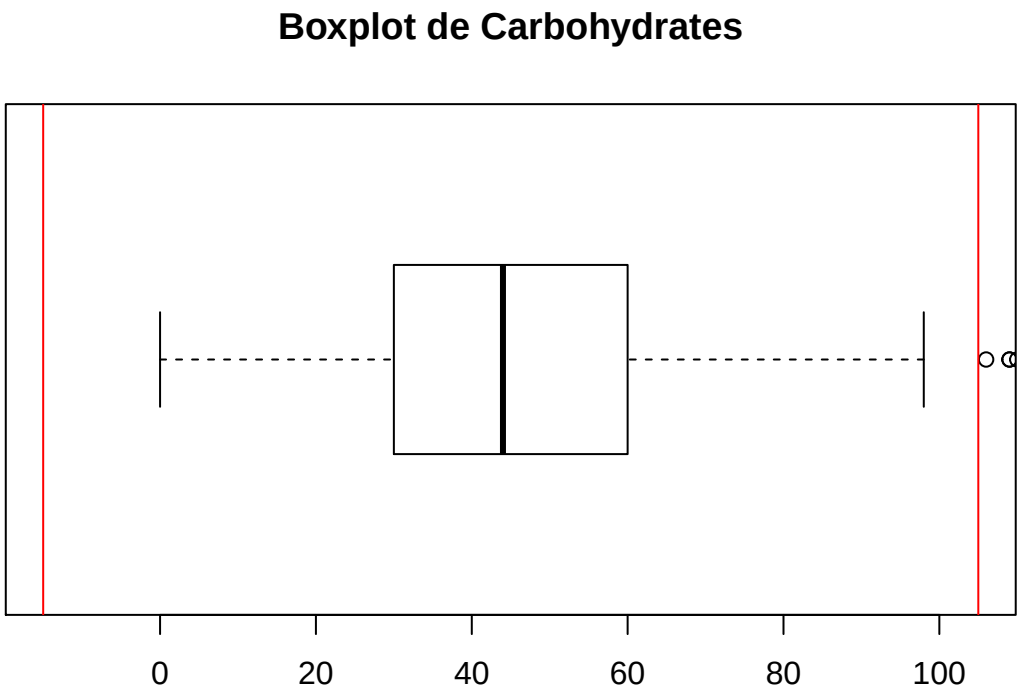



```
## Resumen de Calories con datos atípicos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   210.0   340.0   368.3   500.0  1880.0
```

```
## Resumen de Calories sin datos atípicos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   202.5   335.0   349.0   480.0   930.0
```

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

```
carbohidratos = datos_atipicos(carbohidratos, 'Carbohydrates')
```




```
## Resumen de Carbohydrates con datos atípicos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00
## Resumen de Carbohydrates sin datos atípicos
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   43.00   42.28   56.00   98.00
```

Después de realizar la limpieza sobre las variables seleccionadas el tamaño de la muestra cambia:

- Calories: cambió de 260 a 254.
- Carbohydrates: cambió de 260 a 243.

Análisis de normalidad

1. Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase).

Se elige la prueba de Anderson-Darling para realizar la prueba de normalidad univariada sobre las dos variables.

```
# Prueba de Anderson-Darling de normalidad
prueba_ad_normalidad <- function(variable, nombre_variable) {
  result <- ad.test(variable)

  cat(nombre_variable, "\n")
  cat("Estadística de Anderson-Darling:", result$statistic, "\n")
  cat("P-valor:", result$p.value, "\n")
```

```
  alpha <- 0.05
  if (result$p.value < alpha) {
    cat("La variable parece no seguir una distribución normal.\n")
  } else {
    cat("La variable parece seguir una distribución normal.\n")
  }
}

prueba_ad_normalidad(calorias, 'Calories')
```

```
## Calories
## Estadística de Anderson-Darling: 0.8978551
## P-valor: 0.0216579
## La variable parece no seguir una distribución normal.
```

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

```
prueba_ad_normalidad(carbohidratos, 'Carbohydrates')
```

```
## Carbohydrates
## Estadística de Anderson-Darling: 0.7491741
## P-valor: 0.05047919
## La variable parece seguir una distribución normal.
```

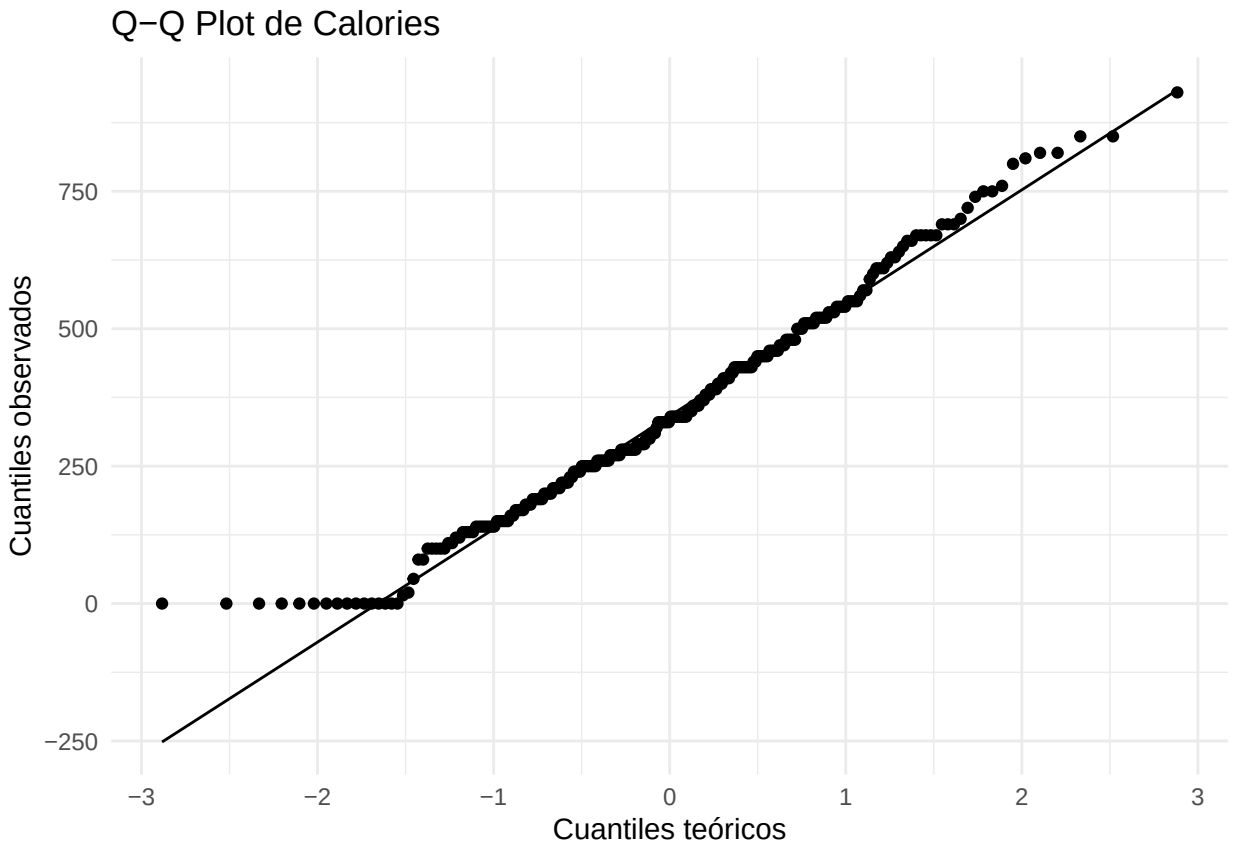
Como se puede observar, sólo la variable *Carbohydrates* parece seguir una distribución normal; el valor p supera al valor de alfa por muy poco.

2. Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable.

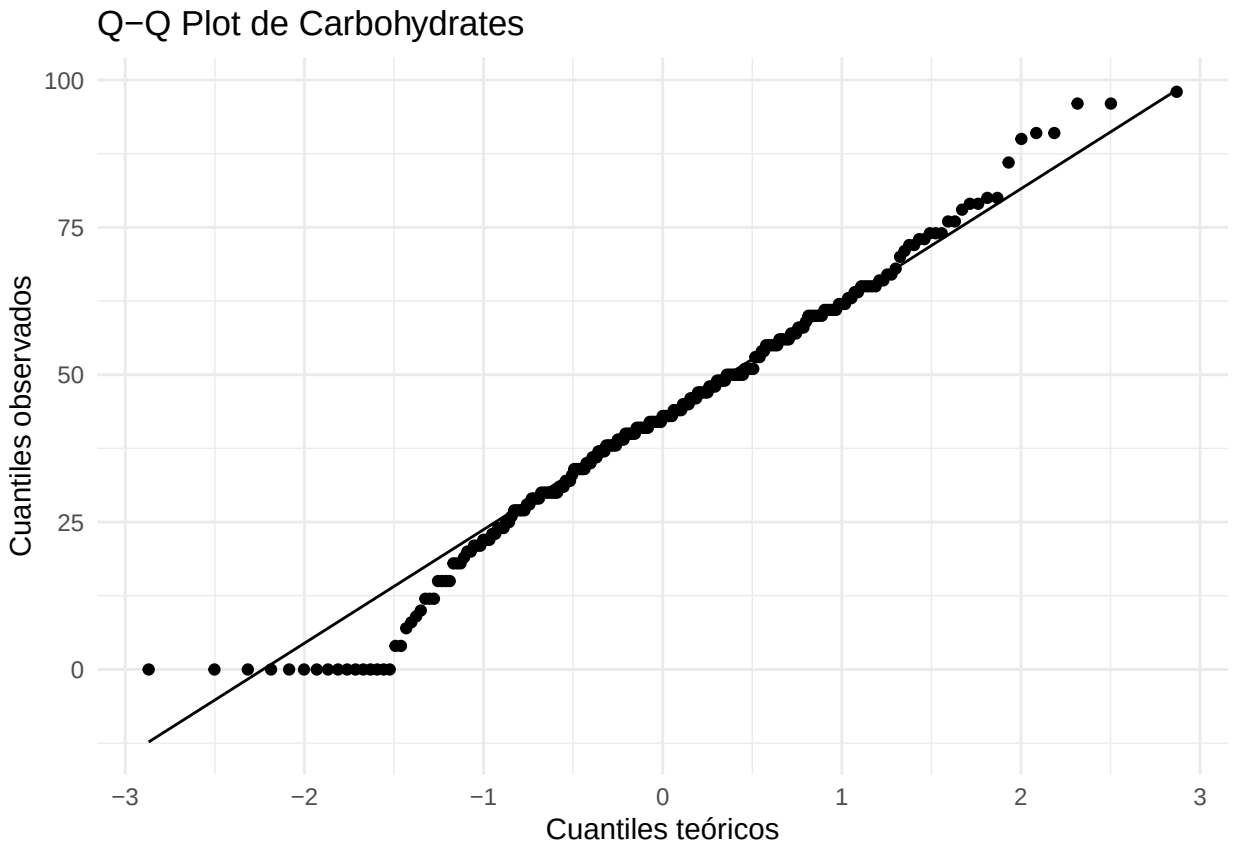
```
library(ggplot2)
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_unnamed' by
## 'rlang::check_dots_unnamed' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_used' by
## 'rlang::check_dots_used' when loading 'tibble'

## Warning: replacing previous import 'ellipsis::check_dots_empty' by
## 'rlang::check_dots_empty' when loading 'tibble'
```

```
grafica_datos <- function(variable, nombre_variable) {
  qq_data <- data.frame(Observed = variable)

  ggplot(qq_data, aes(sample = Observed)) +
    geom_qq() +
    geom_qq_line() +
    labs(title = paste("Q-Q Plot de", nombre_variable),
         x = "Cuantiles teóricos",
         y = "Cuantiles observados") +
    theme_minimal()
}

# Llamadas a la función
grafica_datos(calorias, 'Calories')
```




```
grafica_datos(carbohidratos, 'Carbohydrates')
```



Como se puede observar, parece ser que las distribuciones de ambas variables son con colas delgadas, lo cual indica que hay una baja curtosis y una distribución leptocúrtica.

3. Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable.

```
# Calcula sesgo y curtosis
coeficientes_momentos34 <- function(variable, variable_nombre){
  sesgo <- skewness(variable)
  kurtosis <- kurtosis(variable)

  cat(variable_nombre, "\n")
  cat("Sesgo:", sesgo, "\n")
  cat("Curtosis:", kurtosis, "\n")
}

coeficientes_momentos34(calorias, 'Calories')
```

```
## Calories
## Sesgo: 0.3490549
## Curtosis: 2.716828
```

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

```
coeficientes_momentos34(carbohidratos, 'Carbohydrates')
```

```
## Carbohydrates
## Sesgo: -0.02861759
## Curtosis: 2.931357
```

En cuanto a la variable *Calories* se puede observar que la distribución es moderadamente simétrica y ligeramente

leptocúrtica. Por otro lado, para la variable *Carbohydrates* se puede contemplar que la distribución es casi simétrica y mesocúrtica; estos cálculos coinciden con los resultados obtenidos de la prueba de normalidad.

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

```
calculo_medidas <- function(variable, variable_nombre){
  cat("Resumen de estadísticos para",variable_nombre,":\n")
  cat("Media:", mean(variable), "\n")
  cat("Mediana:", median(variable), "\n")
  cat("Rango medio:", (max(variable) + min(variable)) / 2, "\n")
}

calculo_medidas(calorias, 'Calories')
```

```
## Resumen de estadísticos para Calories :
## Media: 349.0157
## Mediana: 335
## Rango medio: 465
```

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

```
calculo_medidas(carbohidratos, 'Carbohydrates')
```

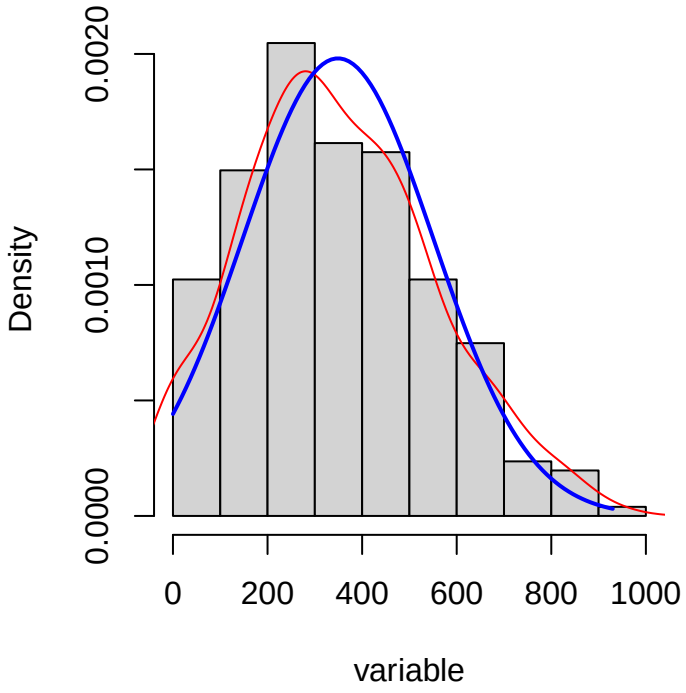
```
## Resumen de estadísticos para Carbohydrates :
## Media: 42.27572
## Mediana: 43
## Rango medio: 49
```

Hay una mayor diferencia entre los estadísticos de *Calories* que entre los estadísticos de *Carbohydrates*. Esto puede sugerir que *Carbohydrates* se distribuye de forma más normal que *Calories*.

5. Realiza el histograma y su distribución teórica de probabilidad.

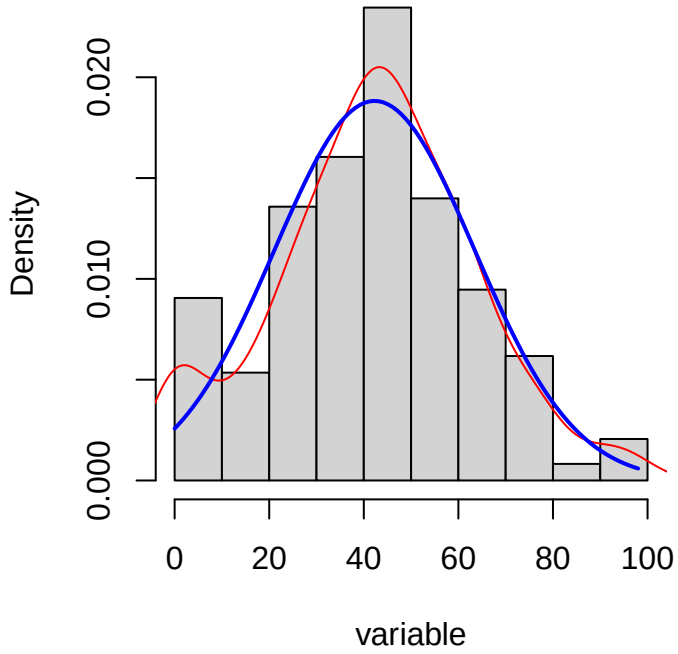
```
histograma_distribucion <- function(variable, variable_nombre){

  # Set graphical parameter pty to ensure correct scaling
  par(pty = "s")

  # Set up a 1x1 grid for the plots (adjust rows and columns as needed)
  par(mfrow = c(1, 1))

  # Graficar histograma
  hist(variable, freq = FALSE, main = paste("Histograma y Distribución Normal de", variable_nombre), co

  # Línea de densidad
  lines(density(variable), col = "red")

  # Curva de distribución normal
  curve(dnorm(x, mean = mean(variable), sd = sd(variable)),
        from = min(variable), to = max(variable),
        add = TRUE, col = "blue", lwd = 2)
}

histograma_distribucion(calorias, 'Calories')
```

Histograma y Distribución Normal de Calories




```
histograma_distribucion(carbohidratos, 'Carbohydrates')
```

Histograma y Distribución Normal de Carbohydrates



Comparando las distribuciones teóricas con las de la muestra se puede observar que hay una mayor similitud entre éstas para la variable *Calories* que para la variable *Carbohydrates*