4. Explorando bases

```
---
title: "4 Explorando bases"
author: "Ozner Leyva"
date: "2024-08-13"
output: word_document
---
```

```
{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

2. Para leer los datos de un archivo usa las siguientes instrucciones de R:

```
{r}
M=read.csv("C:/Users/ozner/Downloads/mc-donalds-menu.csv") #leer la base de datos
# Cargar las librerías necesarias
install.packages("moments")
library(moments)

# Función para calcular sesgo y curtosis
skewness <- function(x) {
  n <- length(x)
  (sum((x-mean(x))^3)/n)/(sum((x-mean(x))^2)/n)^(3/2)
}
kurtosis <- function(x) {
  n <- length(x)
  (sum((x-mean(x))^4)/n)/(sum((x-mean(x))^2)/n)^2
}

# 1. Análisis de datos atípicos
analizar_atipicos <- function(datos, nombre_variable) {
  cat("Análisis de datos atípicos para", nombre_variable, "\n\n")

  # Diagrama de caja y bigote
  boxplot(datos, main=paste("Boxplot de", nombre_variable), horizontal=TRUE)

  # Cálculo de cuartiles y rango intercuartílico
  q1 <- quantile(datos, 0.25)
  q3 <- quantile(datos, 0.75)
  ri <- IQR(datos)
  cat("Q1:", q1, "\n")
  cat("Q3:", q3, "\n")
  cat("Rango intercuartílico:", ri, "\n")
```

```r
  # Cota de 1.5 rangos intercuartílicos
  cota_15_ri_inf <- q1 - 1.5 * ri
  cota_15_ri_sup <- q3 + 1.5 * ri
  atipicos_15_ri <- sum(datos < cota_15_ri_inf | datos > cota_15_ri_sup)
  cat("Cota inferior (1.5 RI):", cota_15_ri_inf, "\n")
  cat("Cota superior (1.5 RI):", cota_15_ri_sup, "\n")
  cat("Datos atípicos (1.5 RI):", atipicos_15_ri, "\n")

  # Cota de 3 desviaciones estándar
  media <- mean(datos)
  desv_est <- sd(datos)
  cota_3_sd_inf <- media - 3 * desv_est
  cota_3_sd_sup <- media + 3 * desv_est
  atipicos_3_sd <- sum(datos < cota_3_sd_inf | datos > cota_3_sd_sup)
  cat("Cota inferior (3 SD):", cota_3_sd_inf, "\n")
  cat("Cota superior (3 SD):", cota_3_sd_sup, "\n")
  cat("Datos atípicos (3 SD):", atipicos_3_sd, "\n\n")
}

# 2. Análisis de normalidad
analizar_normalidad <- function(datos, nombre_variable) {
  cat("Análisis de normalidad para", nombre_variable, "\n\n")

  # Prueba de Shapiro-Wilk
  test_normalidad <- shapiro.test(datos)
  cat("Prueba de Shapiro-Wilk:\n")
  print(test_normalidad)
  cat("\n")

  # QQ-Plot
  qqnorm(datos, main=paste("QQ-Plot de", nombre_variable))
  qqline(datos)

  # Coeficientes de sesgo y curtosis
  sesgo <- skewness(datos)
  curtosis <- kurtosis(datos)
  cat("Coeficiente de sesgo:", sesgo, "\n")
  cat("Coeficiente de curtosis:", curtosis, "\n\n")
```

```r
  # Medidas de tendencia central
  cat("Media:", mean(datos), "\n")
  cat("Mediana:", median(datos), "\n")
  cat("Rango medio:", (max(datos) + min(datos)) / 2, "\n\n")

  # Histograma y distribución teórica
  hist(datos, freq=FALSE, main=paste("Histograma de", nombre_variable))
  lines(density(datos), col="red")
  curve(dnorm(x, mean=mean(datos), sd=sd(datos)),
        from=min(datos), to=max(datos), add=TRUE, col="blue", lwd=2)
}

# 3. Influencia de datos atípicos en la normalidad
analizar_influencia_atipicos <- function(datos, nombre_variable) {
  cat("Influencia de datos atípicos en la normalidad para", nombre_variable, "\n\n"
)

  q1 <- quantile(datos, 0.25)
  q3 <- quantile(datos, 0.75)
  ri <- IQR(datos)
  cota_15_ri_inf <- q1 - 1.5 * ri
  cota_15_ri_sup <- q3 + 1.5 * ri

  datos_sin_atipicos <- datos[datos >= cota_15_ri_inf & datos <= cota_15_ri_sup]
  test_normalidad_sin_atipicos <- shapiro.test(datos_sin_atipicos)
  cat("Prueba de Shapiro-Wilk sin datos atípicos:\n")
  print(test_normalidad_sin_atipicos)
  cat("\n")
}
```

```r
decision_atipicos <- function(datos, nombre_variable) {
  cat("Análisis para decidir sobre datos atípicos en", nombre_variable, "\n\n")

  # Calcular estadísticas con todos los datos
  media_original <- mean(datos)
  mediana_original <- median(datos)
  desv_est_original <- sd(datos)

  # Identificar datos atípicos (usando 1.5 * IQR)
  q1 <- quantile(datos, 0.25)
  q3 <- quantile(datos, 0.75)
  iqr <- q3 - q1
  limite_inferior <- q1 - 1.5 * iqr
  limite_superior <- q3 + 1.5 * iqr

  datos_sin_atipicos <- datos[datos >= limite_inferior & datos <= limite_superior]

  # Calcular estadísticas sin datos atípicos
  media_sin_atipicos <- mean(datos_sin_atipicos)
  mediana_sin_atipicos <- median(datos_sin_atipicos)
  desv_est_sin_atipicos <- sd(datos_sin_atipicos)

  # Calcular el porcentaje de datos atípicos
  porcentaje_atipicos <- (length(datos) - length(datos_sin_atipicos)) / length
(datos) * 100

  # Realizar pruebas de normalidad
  test_original <- shapiro.test(datos)
  test_sin_atipicos <- shapiro.test(datos_sin_atipicos)

  # Imprimir resultados
  cat("Estadísticas con todos los datos:\n")
  cat("Media:", media_original, "\n")
  cat("Mediana:", mediana_original, "\n")
  cat("Desviación estándar:", desv_est_original, "\n")
  cat("p-valor (prueba de normalidad):", test_original$p.value, "\n\n")
```

```
  cat("Estadísticas sin datos atípicos:\n")
  cat("Media:", media_sin_atipicos, "\n")
  cat("Mediana:", mediana_sin_atipicos, "\n")
  cat("Desviación estándar:", desv_est_sin_atipicos, "\n")
  cat("p-valor (prueba de normalidad):", test_sin_atipicos$p.value, "\n\n")

  cat("Porcentaje de datos atípicos:", round(porcentaje_atipicos, 2), "%\n\n")

  # Tomar una decisión basada en los resultados
  if (porcentaje_atipicos > 10) {
    cat("Recomendación: Mantener los datos atípicos.\n")
    cat("Razón: El porcentaje de datos atípicos es alto (>10%). Eliminarlos podría
resultar en una pérdida significativa de información.\n")
  } else if (abs(media_original - media_sin_atipicos) / media_original > 0.1) {
    cat("Recomendación: Mantener los datos atípicos.\n")
    cat("Razón: La eliminación de datos atípicos cambia significativamente la media
(>10% de diferencia).\n")
  } else if (test_original$p.value < 0.05 && test_sin_atipicos$p.value >= 0.05) {
    cat("Recomendación: Considerar la eliminación de datos atípicos para análisis
que requieran normalidad.\n")
    cat("Razón: La eliminación de datos atípicos mejora significativamente la
normalidad de los datos.\n")
  } else {
    cat("Recomendación: Mantener los datos atípicos.\n")
    cat("Razón: Los datos atípicos no parecen afectar significativamente las
estadísticas principales o la normalidad.\n")
  }
}
```

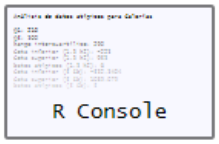
2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

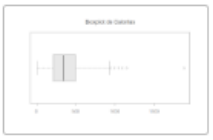
```
 Calorias
 Carbohidratos
 Proteinas
 Sodio
 Azucares (Sugars)
```

```{r}
cal=M$Calories
carb=M$Carbohydrates
```

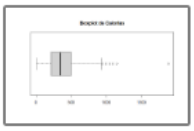
Calorias

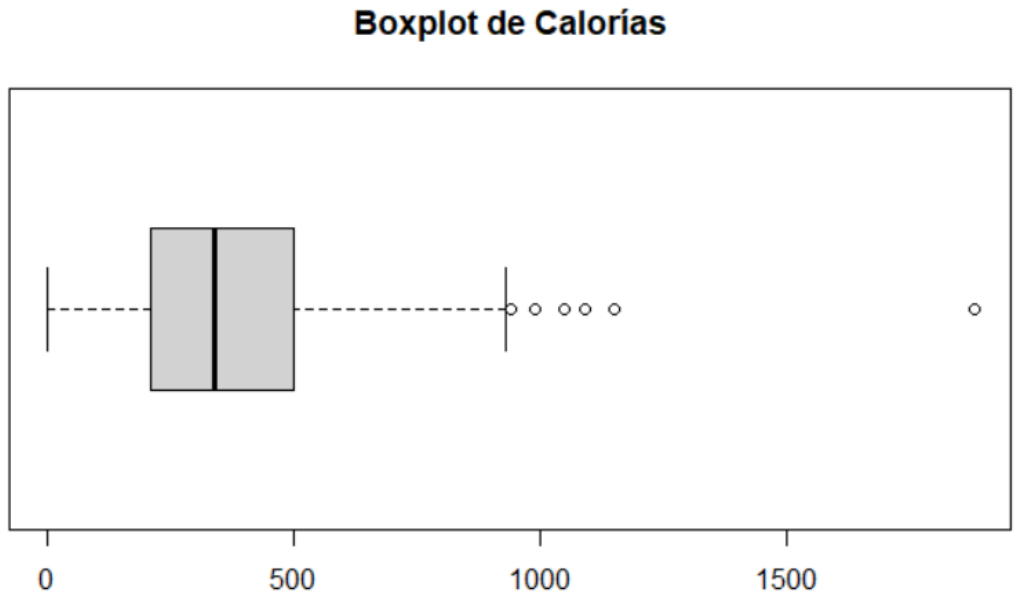
```{r}
analizar_atipicos(cal, "Calorías")
```



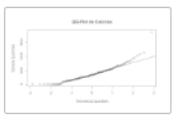
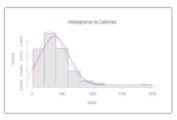


```
Análisis de datos atípicos para Calorías

Q1: 210
Q3: 500
Rango intercuartílico: 290
Cota inferior (1.5 RI): -225
Cota superior (1.5 RI): 935
Datos atípicos (1.5 RI): 6
Cota inferior (3 SD): -352.5404
Cota superior (3 SD): 1089.079
Datos atípicos (3 SD): 3
```

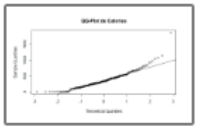


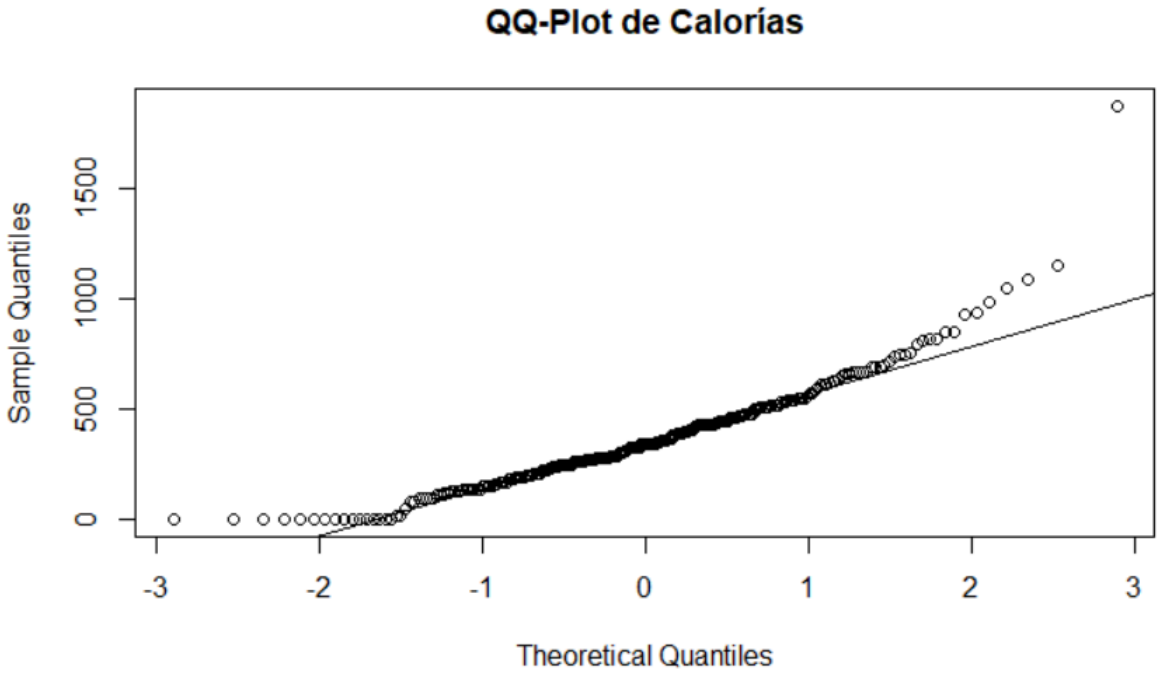


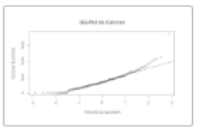
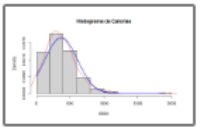
```{r}
analizar_normalidad(cal, "Calorías")
```

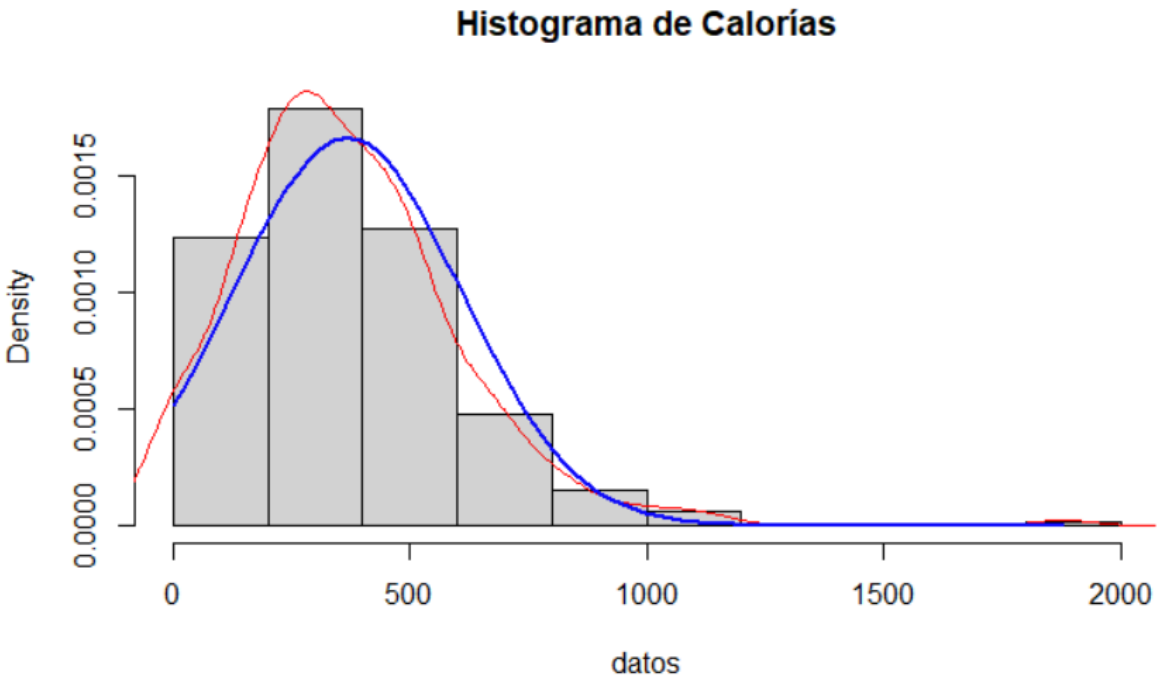



```
Análisis de normalidad para Calorías

Prueba de Shapiro-Wilk:

        Shapiro-Wilk normality test

data:  datos
W = 0.91902, p-value = 1.119e-10


Coeficiente de sesgo: 1.444105
Coeficiente de curtosis: 8.645274

Media: 368.2692
Mediana: 340
Rango medio: 940
```





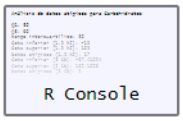
Interpretación de los datos para los datos atipicos

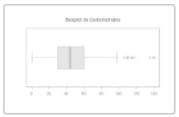
```{r}
decision_atipicos(cal, "Calorías")
```

```
Análisis para decidir sobre datos atípicos en Calorías

Estadísticas con todos los datos:
Media: 368.2692
Mediana: 340
Desviación estándar: 240.2699
p-valor (prueba de normalidad): 1.119028e-10

Estadísticas sin datos atípicos:
Media: 349.0157
Mediana: 335
Desviación estándar: 201.4013
p-valor (prueba de normalidad): 0.001522604

Porcentaje de datos atípicos: 2.31 %

Recomendación: Mantener los datos atípicos.
Razón: Los datos atípicos no parecen afectar significativamente las estadísticas principales o la normalidad.
```

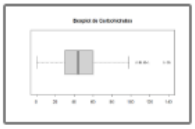
Carbohidratos

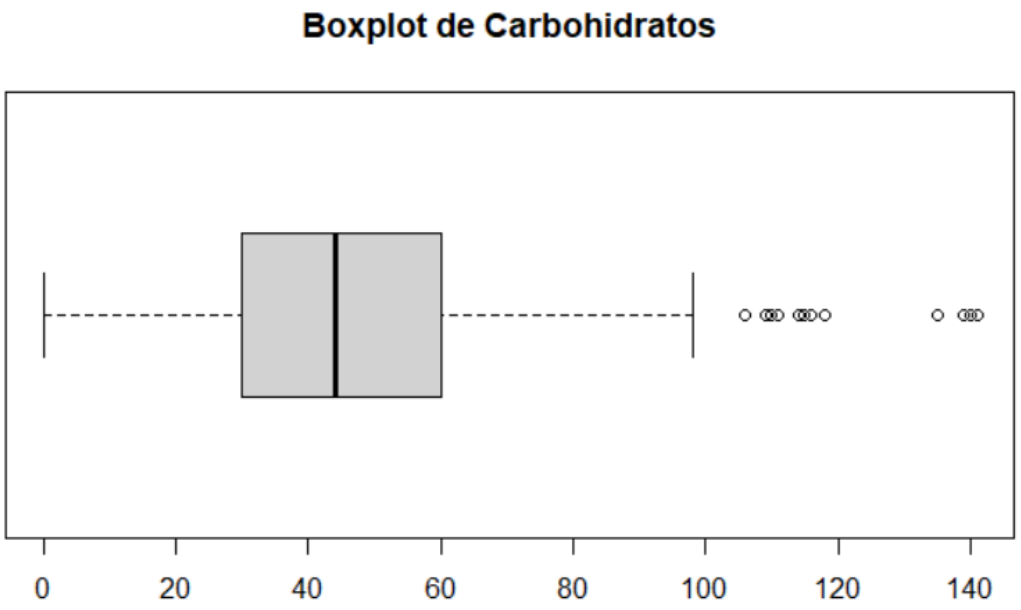
```{r}
analizar_atipicos(carb, "Carbohidratos")
```



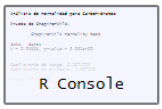


```
Análisis de datos atípicos para Carbohidratos

Q1: 30
Q3: 60
Rango intercuartílico: 30
Cota inferior (1.5 RI): -15
Cota superior (1.5 RI): 105
Datos atípicos (1.5 RI): 17
Cota inferior (3 SD): -37.41054
Cota superior (3 SD): 132.1028
Datos atípicos (3 SD): 5
```

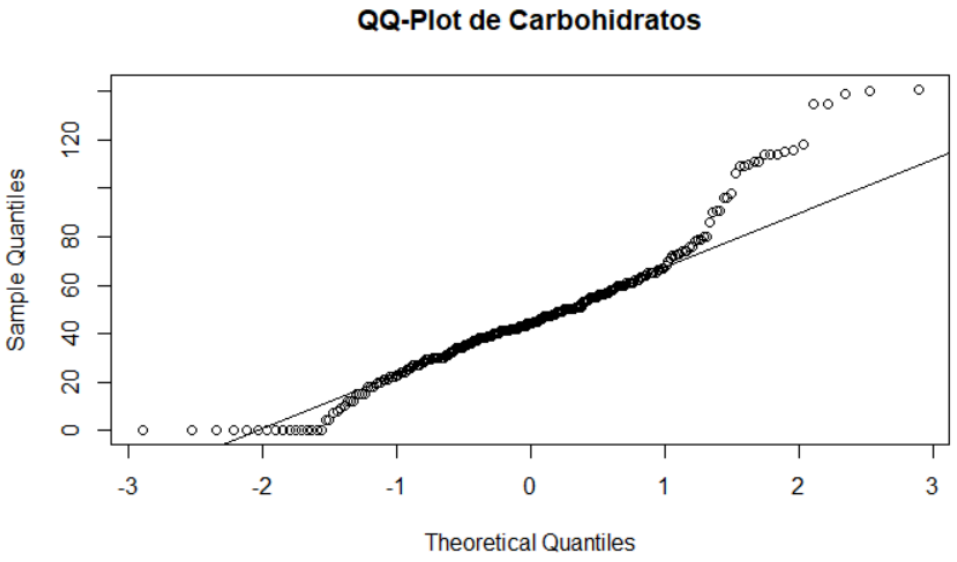


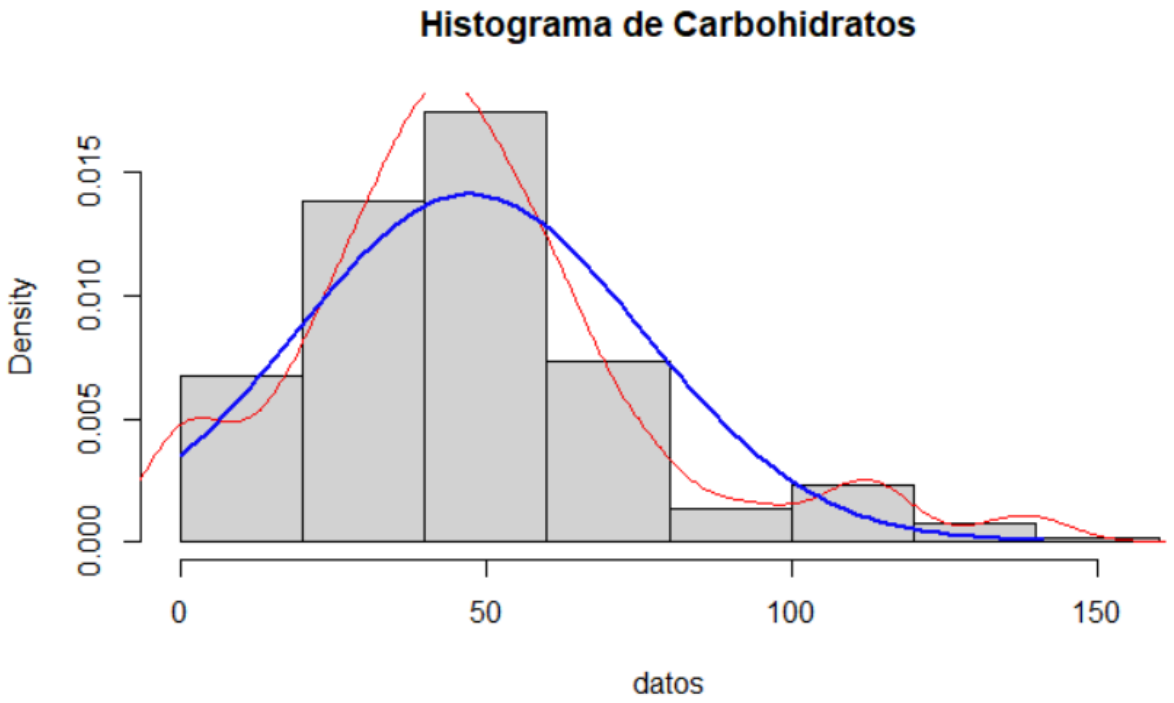


```r
analizar_normalidad(carb, "Carbohidratos")
```




```
Análisis de normalidad para Carbohidratos

Prueba de Shapiro-Wilk:

        Shapiro-Wilk normality test

data:  datos
W = 0.93666, p-value = 3.931e-09


Coeficiente de sesgo: 0.9074253
Coeficiente de curtosis: 4.357538

Media: 47.34615
Mediana: 44
Rango medio: 70.5
```



Interpretación de los datos para los datos atipicos

```{r}
decision_atipicos(carb, "Carbohidratos")
```

```
Análisis para decidir sobre datos atípicos en Carbohidratos

Estadísticas con todos los datos:
Media: 47.34615
Mediana: 44
Desviación estándar: 28.25223
p-valor (prueba de normalidad): 3.931191e-09

Estadísticas sin datos atípicos:
Media: 42.27572
Mediana: 43
Desviación estándar: 21.19162
p-valor (prueba de normalidad): 0.007725821

Porcentaje de datos atípicos: 6.54 %

Recomendación: Mantener los datos atípicos.
Razón: La eliminación de datos atípicos cambia significativamente la media (>10% de diferencia).
```