M1_AI1

Sofia Cantu

2024-08-20

1. Análisis descriptivo de la variable

Variable asignada: Sodium

1. Datos Atípicos

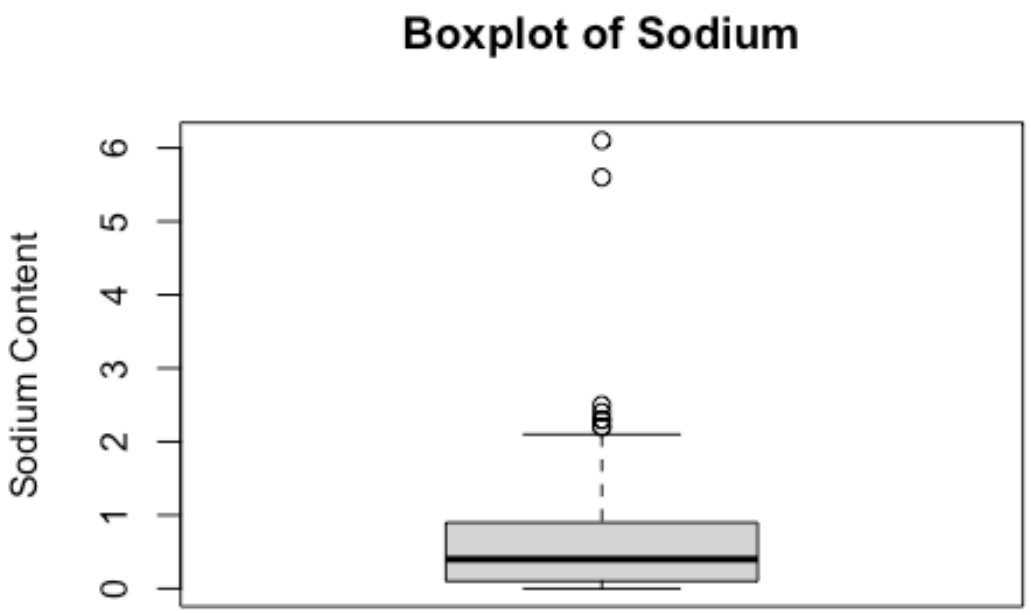
```
data = read.csv("~/Downloads/ArchivosCodigos/food_data_g.csv")
summary(data$Sodium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1000  0.4000  0.5732  0.9000  6.1000
```

```
M <- read.csv("~/Downloads/ArchivosCodigos/mc-donalds-menu.csv")
summary(M$Total.Fat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.375  11.000  14.165  22.250 118.000
```

```
boxplot(data$Sodium, main="Boxplot of Sodium", ylab="Sodium Content")
```




```r
IQR_value <- IQR(data$Sodium)
lower_bound <- quantile(data$Sodium, 0.25) - 1.5 * IQR_value
upper_bound <- quantile(data$Sodium, 0.75) + 1.5 * IQR_value

outliers <- data$Sodium[data$Sodium < lower_bound | data$Sodium >
upper_bound]
print("Valores atípicos basados en la regla IQR.:")
```

```
## [1] "Valores atípicos basados en la regla IQR.:"
```

```r
print(outliers)
```

```
## [1] 2.4 2.3 2.5 2.2 2.3 2.2 5.6 6.1
```

```r
mean_value <- mean(data$Sodium)
sd_value <- sd(data$Sodium)

lower_bound_sd <- mean_value - 3 * sd_value
upper_bound_sd <- mean_value + 3 * sd_value

outliers_sd <- data$Sodium[data$Sodium < lower_bound_sd | data$Sodium >
upper_bound_sd]
```

```r
print("Valores atípicos vasados en la regla de las 3 desviaciones estándar alpinkedor de la media:")
```

```
## [1] "Valores atípicos vasados en la regla de las 3 desviaciones estándar alpinkedor de la media:"
```

```r
print(outliers_sd)
```

```
## [1] 2.5 5.6 6.1
```

```r
lower_bound_extreme <- quantile(data$Sodium, 0.25) - 3 * IQR_value
upper_bound_extreme <- quantile(data$Sodium, 0.75) + 3 * IQR_value

extreme_values <- data$Sodium[data$Sodium < lower_bound_extreme | data$Sodium > upper_bound_extreme]

print("Valores atípicos vasados en la regla de 3 rangos intercuartílicos para datos extremos:")
```

```
## [1] "Valores atípicos vasados en la regla de 3 rangos intercuartílicos para datos extremos:"
```

```r
print(extreme_values)
```

```
## [1] 5.6 6.1
```

Interpreta los resultados

Para los valóres atipicos basados en la regla de 1.5 IQR, estan por fuera del rango esperado. De la lista “2.4, 2.3, 2.5, 2.2, 2.3, 2.2, 5.6, 6.1”, los valores más grandes como 5.6 y 6.1 son considerablemente mayores y podrían indicar una tendencia más extrema.

Para el caso de la regla de las 3 desviaciones estándar, los valores son “2.5, 5.6, 6.1”. Podemos ver que esta regla es más estricuta que la anterior al solo ver los valores más extremos. Podemos concuir que podrían ser casos excepcionales o errores de medición.

Por ultimo para los valores basados en la regal de 3 rangos intercuartílicos, es incluso m´pas extricta que la regla de las 3 Desviaciones Estándar al solo mostrar los datos 5.6 y 6.1.

Es bueno probar varios metodos, pero a ojo de buen cubero pudimos llegar a la misma concusion desde la primera lista de valores atípicos basados en la regla de 1.5 IQR, sin necesidad de hacerlo por 3 metodos.

En conclusion, hacerlo varias veces no esta de mas pero tampoco trajo aportes significativos.

2. Normalidad

```r
if (!require(nortest)) install.packages("nortest")
if (!require(tseries)) install.packages("tseries")
```

```
## Loading required package: tseries
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
if (!require(moments)) install.packages("moments")
```

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

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

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

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

```
data = read.csv("~/Downloads/ArchivosCodigos/food_data_g.csv")
summary(data$Sodium)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1000  0.4000  0.5732  0.9000  6.1000
```

```
datos <- data$Sodium
```

```
# Prueba de Anderson-Darling
ad_test <- ad.test(datos)
cat("Anderson-Darling Test:\n")
```

```
## Anderson-Darling Test:
```

```
cat("Statistic:", ad_test$statistic, "\n")
```

```
## Statistic: 24.82739
```

```
cat("P-value:", ad_test$p.value, "\n")
```

```
## P-value: 3.7e-24
```

```
# Prueba de Jarque-Bera
jb_test <- jarque.bera.test(datos)
cat("\nJarque-Bera Test:\n")
```

```
##
## Jarque-Bera Test:
```

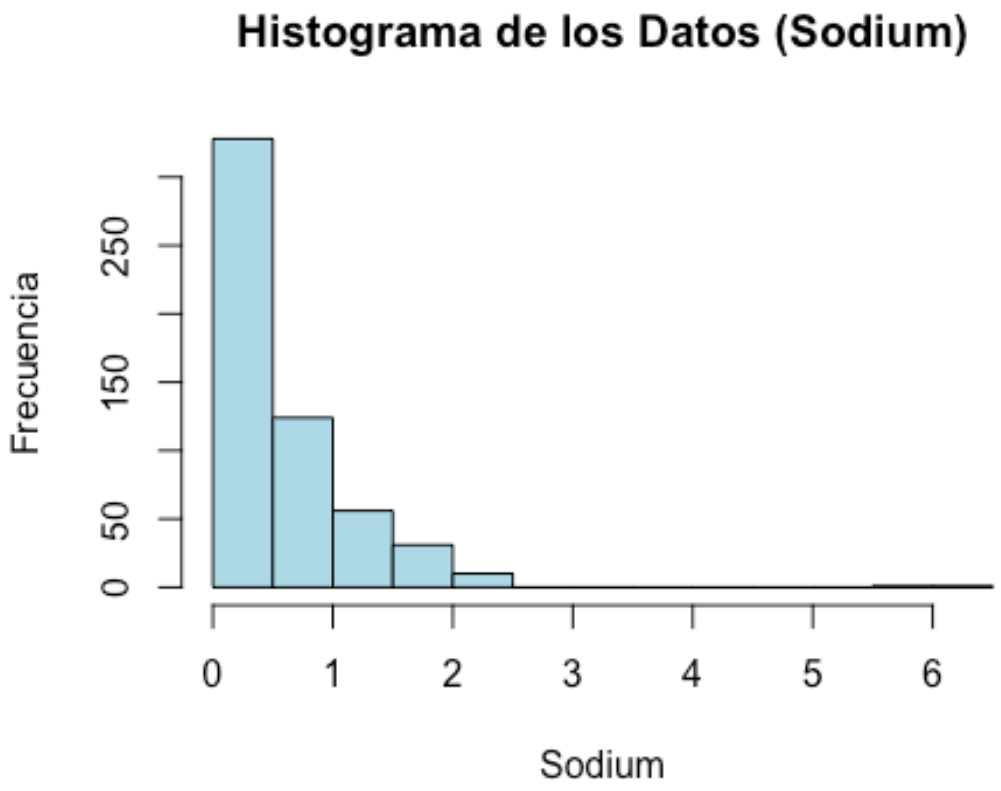
```
cat("Statistic:", jb_test$statistic, "\n")
```

```
## Statistic: 6834.182
```

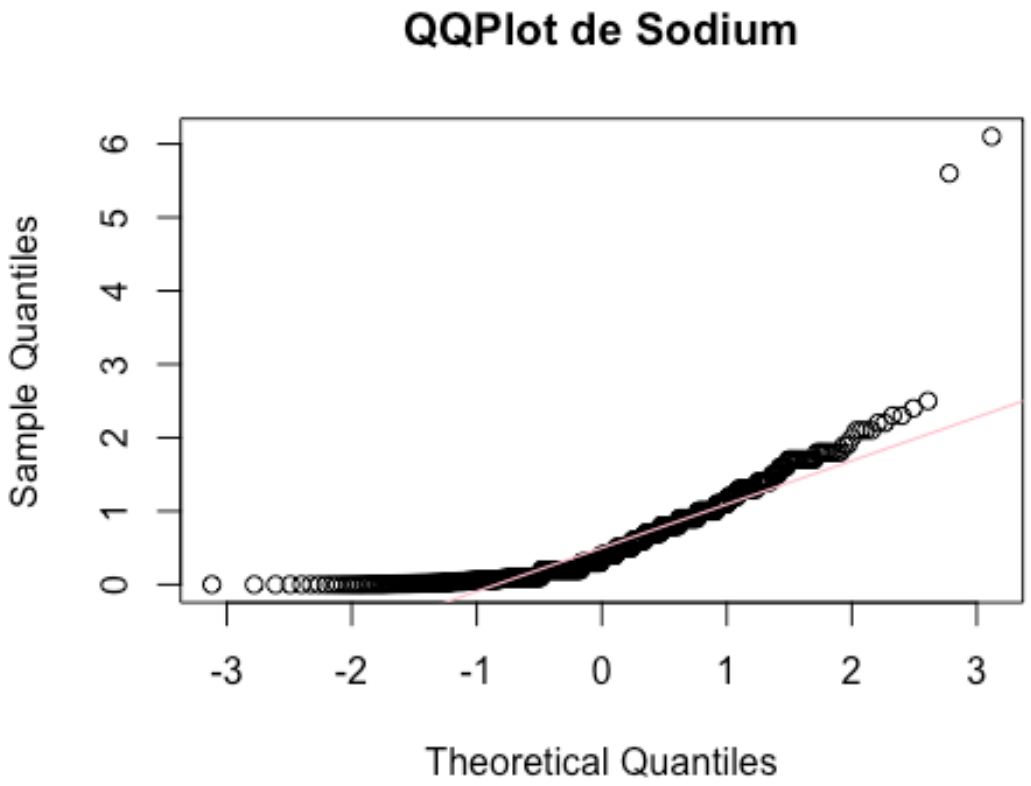
```
cat("P-value:", jb_test$p.value, "\n")
```

```
## P-value: 0

# Gráfico de los datos
hist(datos, main="Histograma de los Datos (Sodium)", xlab="Sodium",
ylab="Frecuencia", col="lightblue")
```




```
# QQPlot
qqnorm(datos, main="QQPlot de Sodium")
qqline(datos, col = "pink")
```




```r
# Calcular coeficiente de sesgo
sesgo <- skewness(datos, na.rm = TRUE)
cat("\nCoeficiente de Sesgo:", sesgo)

##
## Coeficiente de Sesgo: 2.735999

# Calcular coeficiente de curtosis
curtosis <- kurtosis(datos, na.rm = TRUE)
cat("\nCoeficiente de Curtosis:", curtosis)

##
## Coeficiente de Curtosis: 19.3626

# Calcular medidas
media <- mean(datos, na.rm = TRUE)
mediana <- median(datos, na.rm = TRUE)
rango_medio <- (min(datos, na.rm = TRUE) + max(datos, na.rm = TRUE)) / 2

cat("\nMedia:", media)

##
## Media: 0.5732051
```

```r
cat("\nMediana:", mediana)
```

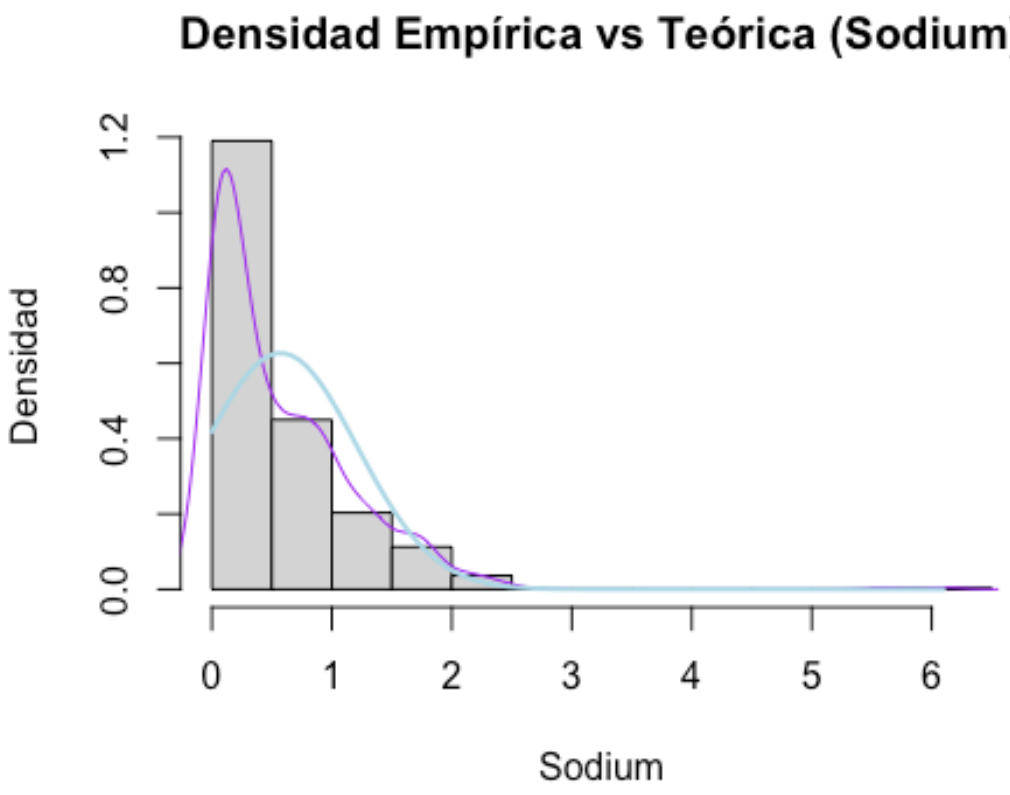
```
##
## Mediana: 0.4
```

```r
cat("\nRango Medio:", rango_medio)
```

```
##
## Rango Medio: 3.05
```

```r
# Gráfico de densidad empírica
hist(datos, freq=FALSE, main="Densidad Empírica vs Teórica (Sodium)",
xlab="Sodium", ylab="Densidad")
lines(density(datos, na.rm = TRUE), col="purple")

# Curva de densidad teórica bajo normalidad
curve(dnorm(x, mean=mean(datos, na.rm = TRUE), sd=sd(datos, na.rm = TRUE)),
      from=min(datos, na.rm = TRUE),
      to=max(datos, na.rm = TRUE),
      add=TRUE, col="lightblue", lwd=2)
```



####

Interpreta los resultados

Las pruebas de normalidad, Anderson-Darling y Jarque-Bera, ambos nos mostraron tener valores p-value bastante bajos, lo cual nos lleva a la concusion de que los datos de la variable de Sodio no siguen una distribucion normal.

Para los gráficos de los datos y QQPlot, el hisograma muestra una distibuciójn sesgada a la derecha. El QQPlot una desviación significativa de la línea de normalidad, en especial para los dartos extremos. La densidad empírica smostro una desviación considerable de la curva de densidad teórica, lo cual refuerza la observación de que los datos no siguen una distrbución normal y estan sesgados a la derecha.

Los valores atípicos nos confirmaron la presencia de que se desvían significativamente del centro de la distribución.

En concusión, los datos de la variable de Sodio no siguen una distribución normal, lo que se evidencia a través de las pruebas estadísticas, los gráficos y las medidas estadísticas.

2. Transformación a normalidad

Aplicar la Transformación de Box-Cox o Yeo-Johnson

```
library(nortest)
library(MASS)
library(e1071)
library(car)
library(tseries)
library(moments)


# Cargar los datos y seleccionar la variable 'Sodium'
data <- read.csv("~/Downloads/ArchivosCodigos/food_data_g.csv")
datos <- data$Sodium
datos <- datos + 1e-6

# Aplicar transformación Box-Cox
boxcox_result <- powerTransform(datos ~ 1, family="bcPower")
summary(boxcox_result)
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.2886        0.29       0.2514       0.3259
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                             LRT df       pval
## LR test, lambda = (0) 457.7722  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                             LRT df       pval
## LR test, lambda = (1) 722.9186  1 < 2.22e-16
```

```r
# Aplicar transformación Yeo-Johnson
yeojohnson_result <- powerTransform(datos ~ 1, family="yjPower")
summary(yeojohnson_result)
```

```
## yjPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -1.2375          -1      -1.5265      -0.9485
##
##  Likelihood ratio test that transformation parameter is equal to 0
##                            LRT df       pval
## LR test, lambda = (0) 85.6617  1 < 2.22e-16
```

```r
# Extraer lambda óptimo para cada método
lambda_boxcox <- boxcox_result$lambda
lambda_yeojohnson <- yeojohnson_result$lambda

# Transformar los datos con los lambdas encontrados
datos_boxcox <- bcPower(datos, lambda_boxcox)
datos_yeojohnson <- yjPower(datos, lambda_yeojohnson)
```

Formula general

Box-Cox: ##### y(λ)={ 1[y**λ-1/λ] si λ /=0 2[log(y)] si λ =0

Yeo-Johnson: ##### y(λ)={ 1[y**λ-1/λ] si λ /=0 2[log(y)] si λ = 0 3[-(y+1)**(2-λ)-1/λ/2-λ] si λ /=2 4[-log(y+1)] si λ = 2

Análisis de Normalidad de las Transformaciones

```r
# Medidas para los datos originales
summary(datos)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000001 0.100001 0.400001 0.573206 0.900001 6.100001
```

```r
cat("\nSesgo:", skewness(datos, na.rm = TRUE))
```

```
##
## Sesgo: 2.735999
```

```r
cat("\nCurtosis:", kurtosis(datos, na.rm = TRUE))
```

```
##
## Curtosis: 19.3626
```

```r
# Medidas para los datos transformados (Box-Cox)
summary(datos_boxcox)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.4002 -1.6821 -0.8051 -0.9280 -0.1038  2.3743
```

```r
cat("\nSesgo (Box-Cox):", skewness(datos_boxcox, na.rm = TRUE))
```

```
##
## Sesgo (Box-Cox): -0.1872574
```

```
cat("\nCurtosis (Box-Cox):", kurtosis(datos_boxcox, na.rm = TRUE))
```

```
##
## Curtosis (Box-Cox): 2.59613
```

```
# Medidas para los datos transformados (Yeo-Johnson)
summary(datos_yeojohnson)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000001 0.089905 0.275212 0.270173 0.442911 0.736633
```

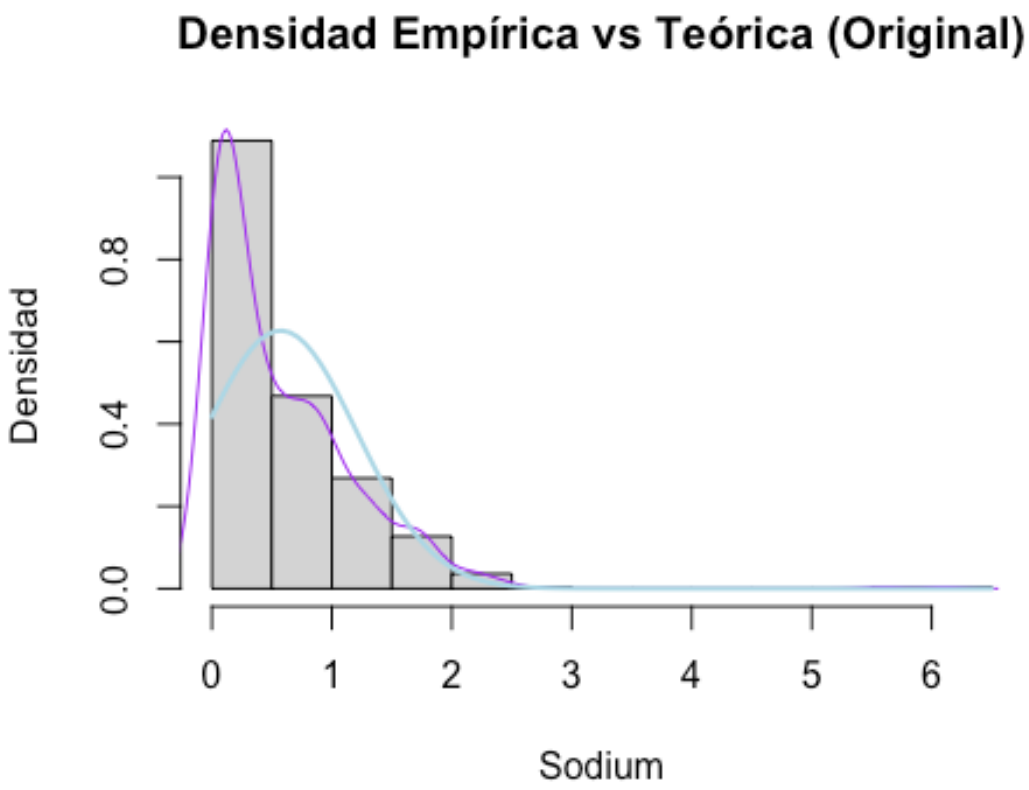
```
cat("\nSesgo (Yeo-Johnson):", skewness(datos_yeojohnson, na.rm = TRUE))
```

```
##
## Sesgo (Yeo-Johnson): 0.1809705
```

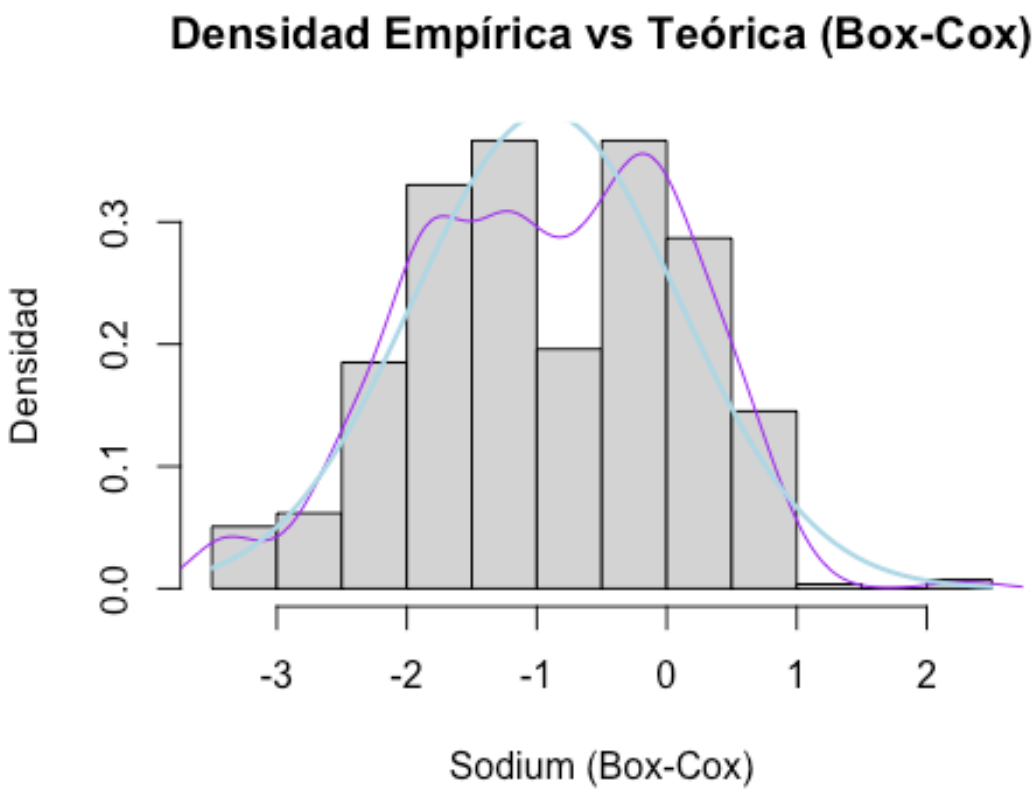
```
cat("\nCurtosis (Yeo-Johnson):", kurtosis(datos_yeojohnson, na.rm = TRUE))
```

```
##
## Curtosis (Yeo-Johnson): 1.711956
```

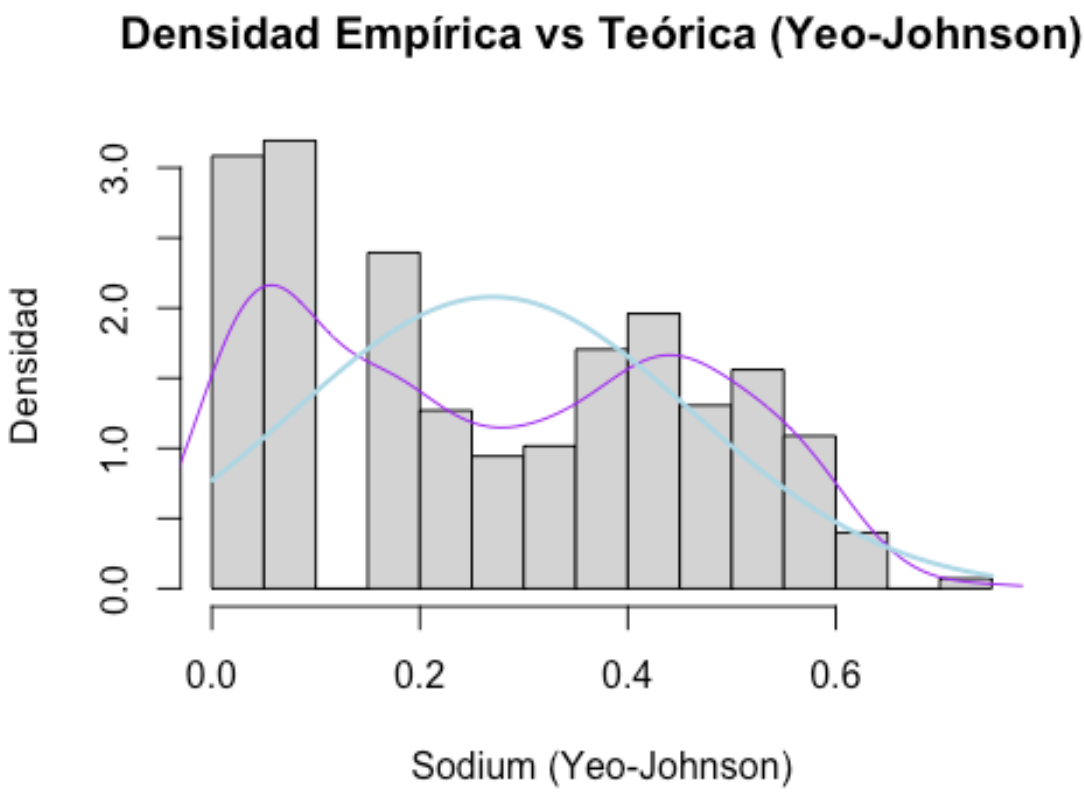
```
# Gráfico para los datos originales
hist(datos, freq=FALSE, main="Densidad Empírica vs Teórica (Original)",
xlab="Sodium", ylab="Densidad")
lines(density(datos, na.rm = TRUE), col="purple")
curve(dnorm(x, mean=mean(datos, na.rm = TRUE), sd=sd(datos, na.rm = TRUE)),
add=TRUE, col="lightblue", lwd=2)
```




```r
# Gráfico para los datos transformados (Box-Cox)
hist(datos_boxcox, freq=FALSE, main="Densidad Empírica vs Teórica (Box-Cox)",
xlab="Sodium (Box-Cox)", ylab="Densidad")
lines(density(datos_boxcox, na.rm = TRUE), col="purple")
curve(dnorm(x, mean=mean(datos_boxcox, na.rm = TRUE), sd=sd(datos_boxcox,
na.rm = TRUE)), add=TRUE, col="lightblue", lwd=2)
```




```r
# Gráfico para los datos transformados (Yeo-Johnson)
hist(datos_yeojohnson, freq=FALSE, main="Densidad Empírica vs Teórica (Yeo-Johnson)", xlab="Sodium (Yeo-Johnson)", ylab="Densidad")
lines(density(datos_yeojohnson, na.rm = TRUE), col="purple")
curve(dnorm(x, mean=mean(datos_yeojohnson, na.rm = TRUE), sd=sd(datos_yeojohnson, na.rm = TRUE)), add=TRUE, col="lightblue", lwd=2)
```




```r
# Prueba de Anderson-Darling y Jarque-Bera para datos originales
ad_test_original <- ad.test(datos)
jb_test_original <- jarque.bera.test(datos)

# Prueba de Anderson-Darling y Jarque-Bera para Box-Cox
ad_test_boxcox <- ad.test(datos_boxcox)
jb_test_boxcox <- jarque.bera.test(datos_boxcox)

# Prueba de Anderson-Darling y Jarque-Bera para Yeo-Johnson
ad_test_yeojohnson <- ad.test(datos_yeojohnson)
jb_test_yeojohnson <- jarque.bera.test(datos_yeojohnson)

# Mostrar resultados
cat("Original Data - Anderson-Darling:", ad_test_original$statistic, "P-
value:", ad_test_original$p.value, "\n")

## Original Data - Anderson-Darling: 24.82739 P-value: 3.7e-24

cat("Original Data - Jarque-Bera:", jb_test_original$statistic, "P-value:",
jb_test_original$p.value, "\n")

## Original Data - Jarque-Bera: 6834.182 P-value: 0
```

```r
cat("\nBox-Cox Transformed Data - Anderson-Darling:",
ad_test_boxcox$statistic, "P-value:", ad_test_boxcox$p.value, "\n")
```

```
##
## Box-Cox Transformed Data - Anderson-Darling: 3.517394 P-value: 8.494608e-
09
```

```r
cat("Box-Cox Transformed Data - Jarque-Bera:", jb_test_boxcox$statistic, "P-
value:", jb_test_boxcox$p.value, "\n")
```

```
## Box-Cox Transformed Data - Jarque-Bera: 6.964928 P-value: 0.0307316
```

```r
cat("\nYeo-Johnson Transformed Data - Anderson-Darling:",
ad_test_yeojohnson$statistic, "P-value:", ad_test_yeojohnson$p.value, "\n")
```

```
##
## Yeo-Johnson Transformed Data - Anderson-Darling: 12.80354 P-value: 3.7e-24
```

```r
cat("Yeo-Johnson Transformed Data - Jarque-Bera:",
jb_test_yeojohnson$statistic, "P-value:", jb_test_yeojohnson$p.value, "\n")
```

```
## Yeo-Johnson Transformed Data - Jarque-Bera: 41.09674 P-value: 1.191124e-09
```

```r
# Identificar y corregir anomalías
outliers <- datos[datos > 3*IQR(datos)]
datos_corregidos <- datos
datos_corregidos[datos > 3*IQR(datos)] <- NA # O puedes reemplazar con la
mediana o media

# Reaplicar la transformación si es necesario
datos_boxcox_corregidos <- bcPower(datos_corregidos, lambda_boxcox)
datos_yeojohnson_corregidos <- yjPower(datos_corregidos, lambda_yeojohnson)
```

Comentarios sobre la Normalidad de las Transformaciones Obtenidas

Para la comparación de las Medidas Estadísticas, los datos originales presentan un alto sesgo positivo (2.7360) y una curtosis bastante alta (19.3626), indicando una fuerte asimetría hacia la derecha y colas muy pesadas. En otras palabras, es una distribución no normal. Después de la transformación Box-Cox, el sesgo se reduce significativamente (-0.1873), acercándose a 0, y la curtosis se aproxima a la de una distribución normal (2.5961). Este proceso hizo la distribución más simétrica y menos apuntada. La transformación Yeo-Johnson también reduce el sesgo (0.1810) y la curtosis (1.7120), pero no tan cerca de los valores de una distribución normal como lo hace la transformación Box-Cox.

Como se puede ver en los graficos, el histograma muestra una fuerte asimetría con una gran acumulación de valores en la parte inferior y una cola extendida hacia la derecha. El histograma post-transformación Box-Cox muestra una distribución más simétrica, con los datos distribuidos más uniformemente alrededor de la media. Y al usar el Yeo, el histograma muestra una cierta mejora en la simetría, pero sigue siendo menos ajustado que el resultado de la transformación Box-Cox.

En concusion, la transformación Box-Cox es la mejor opción entre las dos evaluadas.