Actividad Integradora 1

Daniela Jiménez Téllez

2024-08-20

En este trabajo se utilizará el dataset Nutrición Mundial, del cual se analizará la variable Grasas Saturadas en específico.

Importación de librerías

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

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

Importación de datos

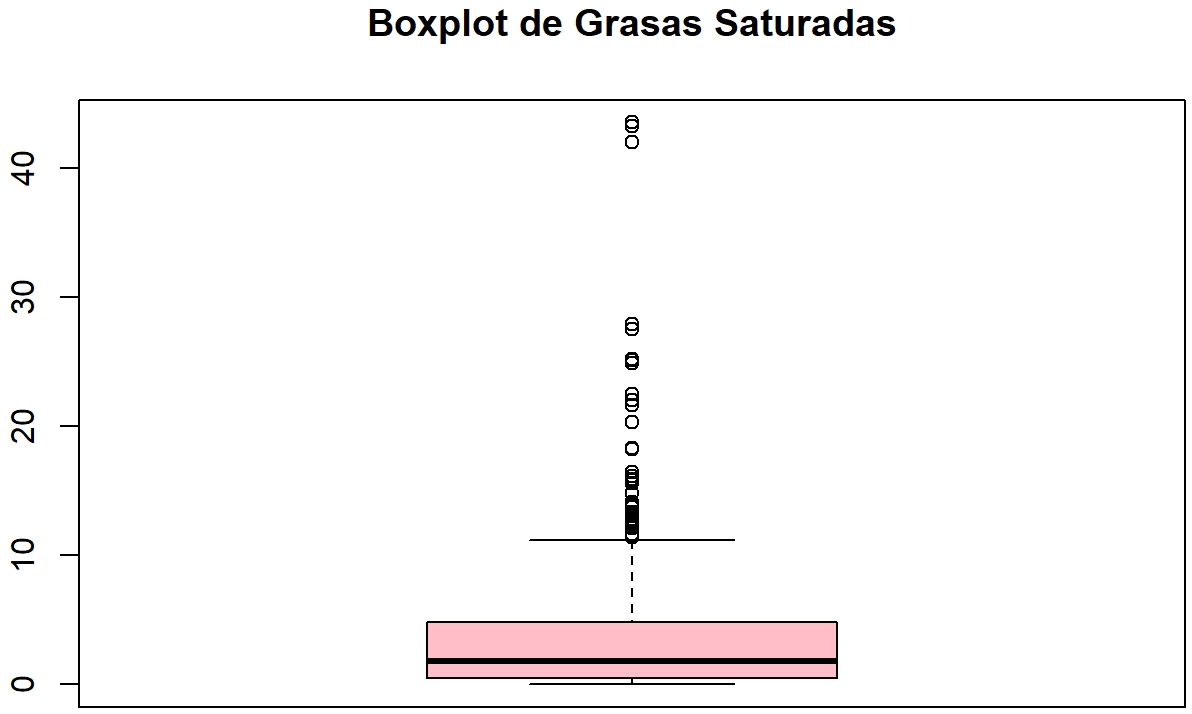
```
data <- read.csv("food_data_g.csv")
grasas_sat <- data$Saturated.Fats
```

Punto 1. Análisis descriptivo de la variable

> 1. Para analizar datos atípicos se te sugiere:

- Graficar el diagrama de caja y bigote.

```
boxplot(grasas_sat, main = "Boxplot de Grasas Saturadas", col = "pink")
```



- Calcula las principales medidas que te ayuden a identificar datos atípicos (utilizar summary te puede abreviar el cálculo): Cuartil 1, Cuartil 3, Media, Cuartil 3, Rango intercuartílico y Desviación estándar.

```
# Cuartiles y rangos intercuartílicos

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

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

```
print(Q1)
```

```
## 25%
## 0.5
```

```
print(Q2)
```

```
## 50%
## 1.8
```

```
print(Q3)
```

```
## 75%
## 4.8
```

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

```
## El rango intercuartílico es: 4.3
```

```
# Media y desviación estándar

mean_gs <- mean(grasas_sat)
sd_gs <- sd(grasas_sat)
cat("La media es:", mean_gs, "\n")
```

```
## La media es: 3.722715
```

```
cat("La desviación estándar es:", sd_gs, "\n")
```

```
## La desviación estándar es: 5.397021
```

- Identifica la cota de 1.5 rangos intercuartílicos para datos atípicos, ¿hay datos atípicos de acuerdo con este criterio? ¿cuántos son?

```
# Cota de 1.5 rangos intercuartílicos

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

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

cat("Los valores atípicos de acuerdo al criterio de la cota de 1.5 rangos intercuartílicos son:
", valores_atp_15)
```

```
## Los valores atípicos de acuerdo al criterio de la cota de 1.5 rangos intercuartílicos son:  2
2 43.5 20.3 12.8 16.4 16.1 13.3 24.9 25.2 15.8 27.5 13 22.5 25.1 43.2 11.5 11.4 12.2 14.1 11.4 1
1.6 18.2 12.4 14 11.4 12.6 14.8 13.7 15.6 11.6 12.5 15.9 21.6 27.9 13.9 42 18.3 21.6 12.3 12.4 1
2.1 12.4
```

- Identifica la cota de 3 desviaciones estándar alrededor de la media, ¿hay datos atípicos de acuerdo con este criterio? ¿cuántos son?

```
# Cota de 3 desviaciones estándar

cota_inf_3std <- mean_gs - 3 * sd_gs
cota_sup_3std <- mean_gs + 3 * sd_gs
valores_atp_3 <- grasas_sat[grasas_sat < cota_inf_3std | grasas_sat > cota_sup_3std]

cat("Los valores atípicos de acuerdo al criterio de la cota de 3 desviaciones estándar al rededo
r de la media son: ", valores_atp_3)
```

```
## Los valores atípicos de acuerdo al criterio de la cota de 3 desviaciones estándar al rededor
de la media son:  22 43.5 20.3 24.9 25.2 27.5 22.5 25.1 43.2 21.6 27.9 42 21.6
```

- Identifica la cota de 3 rangos intercuartílicos para datos extremos, ¿hay datos extremos de acuerdo con este criterio? ¿cuántos son?

```
cota_inf <- Q1 - 3 * ri
cota_sup <- Q3 + 3 * ri

extremos <- grasas_sat[grasas_sat < cota_inf | grasas_sat > cota_sup]

cat("Los datos extremos de acuerdo al criterio de la cota de 3 rangos intercuartílicos es:", ext
remos)
```

```
## Los datos extremos de acuerdo al criterio de la cota de 3 rangos intercuartílicos es: 22 43.5
20.3 24.9 25.2 27.5 22.5 25.1 43.2 18.2 21.6 27.9 42 18.3 21.6
```

- Interpreta los resultados obtenidos y argumenta sobre el comportamiento de los datos atípicos y extremos en la variable seleccionada.

Con los datos obtenidos podemos observar que hay valores atípicos tanto con el criterio de cota de 1.5 rangos intercuartílicos, 3 desviaciones estándar y 3 rangos intercuartílicos; sin embargo, también se puede ver que con el de 3 desviaciones estándar hay menos, ya que da más lugar para “error”. Con esto se puede concluir que los outliers están concentrados en valores muy altos y diferentes a la mayoría, y que se alejan demasiado de la media. Esto nos da a entender que la distribución se ve como si tuviera una cola larga a la derecha.

2. Para analizar normalidad se te sugiere:

- Realiza pruebas de normalidad univariada para la variable (utiliza las pruebas de Anderson-Darling y de Jarque Bera). No olvides incluir H0 y H1 para la prueba de normalidad.

```
# Anderson-Darling

and_dar <- ad.test(grasas_sat)
cat("Los resultados del Anderson-Darling test son: \n")
```

```
## Los resultados del Anderson-Darling test son:
```

```
print(and_dar)
```

```
##
##  Anderson-Darling normality test
##
## data:  grasas_sat
## A = 50.094, p-value < 2.2e-16
```

```
# Jarque-Bera

jarque_b <- jarque.bera.test(grasas_sat)
cat("Los resultados del Jarque Bera test son: \n")
```

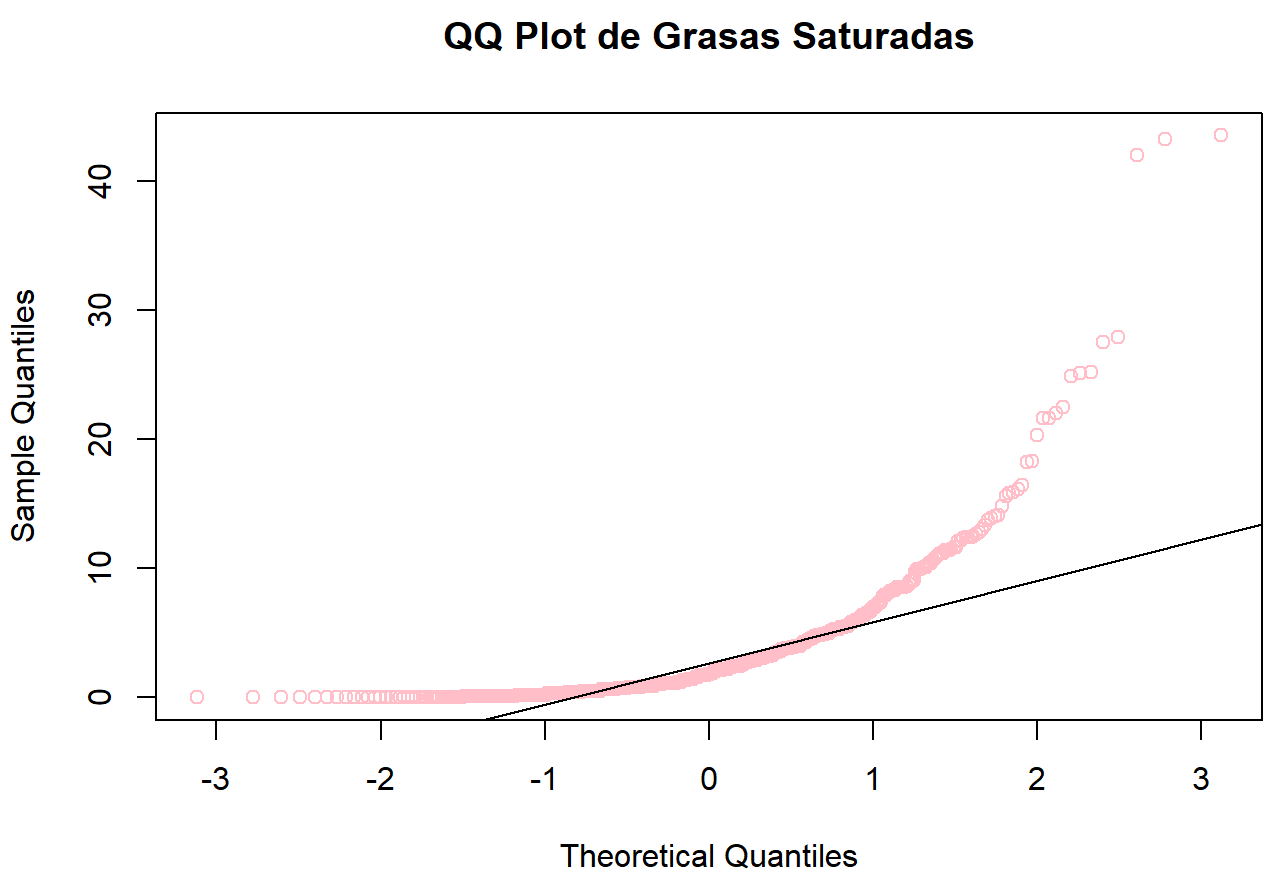
```
## Los resultados del Jarque Bera test son:
```

```
print(jarque_b)
```

```
##
##  Jarque Bera Test
##
## data:  grasas_sat
## X-squared = 7694.1, df = 2, p-value < 2.2e-16
```

- Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos)

```
qqnorm(grasas_sat, main = "QQ Plot de Grasas Saturadas", col = "pink")
qqline(grasas_sat, col = "black")
```



- Calcula el coeficiente de sesgo y el coeficiente de curtosis

```
sesgo_gs <- skewness(grasas_sat)
curtosis_gs <- kurtosis(grasas_sat)
cat("El coeficiente de sesgo es:", sesgo_gs, "\n")
```

```
## El coeficiente de sesgo es: 3.428631
```

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

```
## El coeficiente de curtosis es: 19.97384
```

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

```
mediana_gs <- median(grasas_sat)
rango_medio_gs <- (min(grasas_sat) + max(grasas_sat)) / 2

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

```
## La media es:  3.722715
```

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

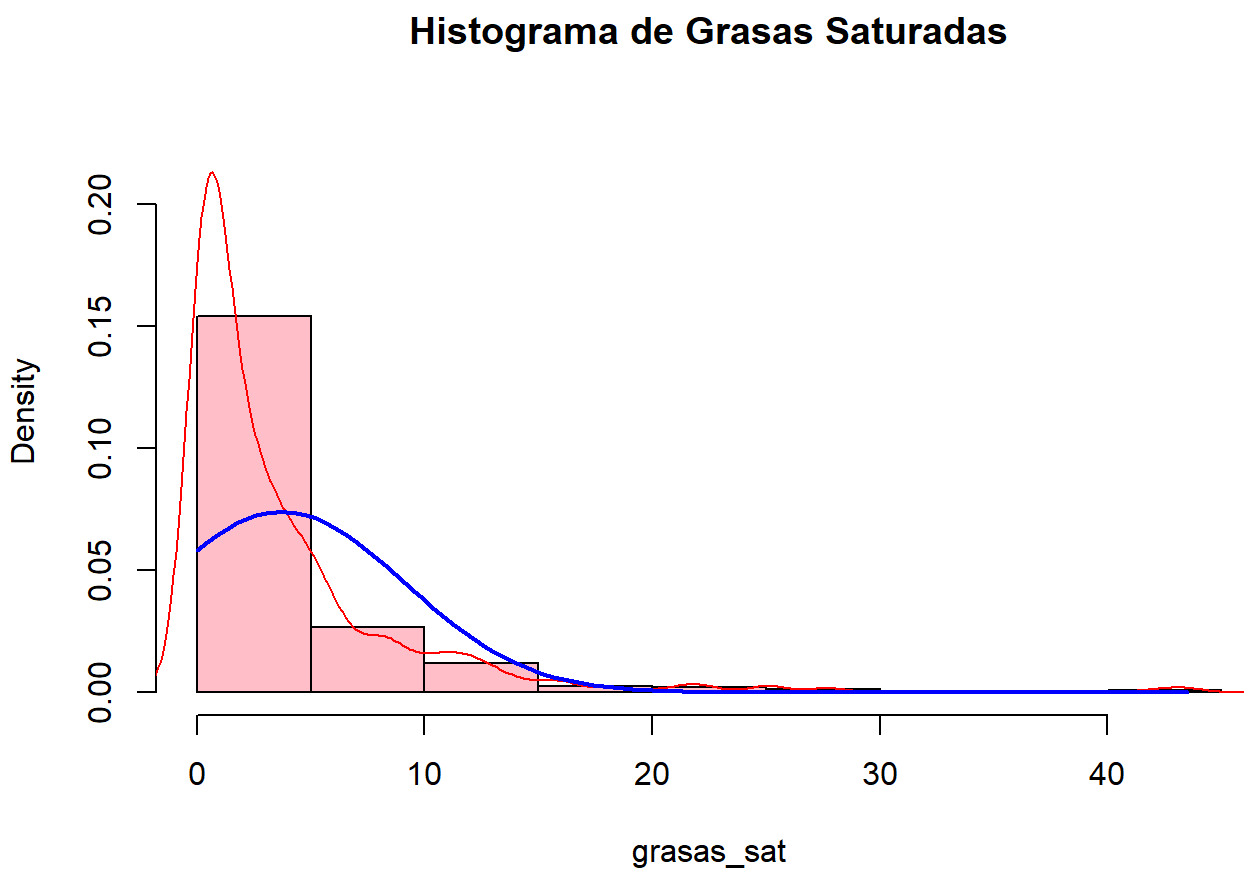
```
## La mediana es:  1.8
```

```
cat("El rango medio es: ", rango_medio_gs)
```

```
## El rango medio es:  21.75
```

- Realiza el gráfico de densidad empírica y teórica suponiendo normalidad en la variable. 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)

```
hist(grasas_sat, freq = FALSE, main = "Histograma de Grasas Saturadas", col = "pink", ylim = c
(0, 0.23))
lines(density(grasas_sat),col = "red")
curve(dnorm(x, mean = mean(grasas_sat), sd = sd(grasas_sat)), from = min(grasas_sat), to = max(g
rasas_sat), add = TRUE, col = "blue", lwd = 2)
```



- Interpreta los gráficos y los resultados obtenidos en cada punto con vías a indicar si hay normalidad de los datos
- Comenta las características encontradas:
  - Considera alejamientos de normalidad por simetría, curtosis
  - Comenta si hay aparente influencia de los datos atípicos en la normalidad de los datos

- Emite una conclusión sobre la normalidad de los datos. Se debe argumentar en términos de los 3 puntos analizados: las pruebas de normalidad, los gráficos y las medidas.

Con las pruebas hechas, podemos observar que la de Anderson-Darling nos dice que hay evidencia contra la hipótesis nula de normalidad. Igualmente, la prueba de Jarques Bera también rechazó la normalidad de los datos, por lo tanto, estos no se comportan como una distribución normal.

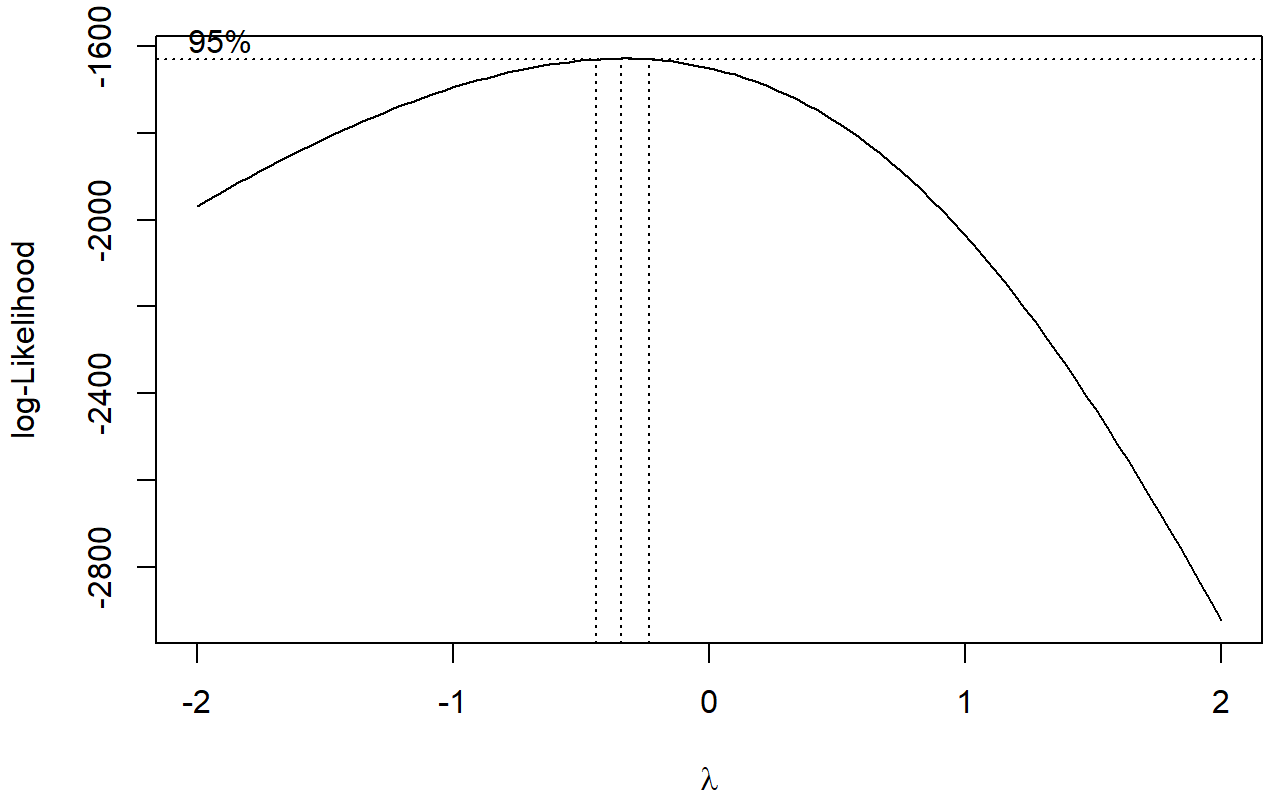
Por otro lado, en el QQPlot podemos ver cómo la línea de normalidad se desvía especialmente al final de la gráfica, lo que nos confirma que en efecto hay valores atípicos y que nuestros datos no se comportan de manera normal.

Finalmente, con los valores de curtosis y sesgo podemos decir que hay una asimetría positiva y que hay una cola muy larga, como se puede observar en el histograma, lo que una vez más nos dice que hay valores atípicos.

Punto 2. Transformación a normalidad

- Encuentra la mejor transformación de los datos para lograr normalidad. Puedes hacer uso de la transformación Box-Cox o de Yeo Johnson o el comando powerTransform para encontrar la mejor lambda para la transformación. Utiliza el modelo exacto y el aproximado de acuerdo con las sugerencias de Box y Cox para la transformación.

```
# Transformación de Box-Cox

boxcox_model <- boxcox(lm((grasas_sat + 1) ~ 1)) #
```



```
# Lambda
lambda <- boxcox_model$x[which.max(boxcox_model$y)]
print(lambda)
```

```
## [1] -0.3434343
```

- Escribe las ecuaciones de los modelos de transformación encontrados.

Para el modelo de BoxCox exacto se utilizó la ecuación

$$y(\lambda) = \frac{(X+1)^{\lambda} - 1}{\lambda}, \quad \text{donde } \lambda = -0.3434$$

Y para el exacto

$$y(x) = \sqrt{x+1}$$

```
bc_exacto <- ((grasas_sat + 1)^lambda - 1) / lambda
bc_aprox <- sqrt(grasas_sat + 1)
```

- Analiza la normalidad de las transformaciones obtenidas con los datos originales. Utiliza como argumento de normalidad:
    - Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.
    - Grafica las funciones de densidad empírica y teórica de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.
    - Realiza la prueba de normalidad de Anderson-Darling y de Jarque Bera para los datos transformados y los originales

```
# Medidas

min_exacto <- min(bc_exacto)
max_exacto <- max(bc_exacto)
media_exacto <- mean(bc_exacto)
mediana_exacto <- median(bc_exacto)
Q1_exacto <- quantile(bc_exacto, 0.25)
Q3_exacto <- quantile(bc_exacto, 0.75)
sesgo_exacto <- skewness(bc_exacto)
curt_exacto <- kurtosis(bc_exacto)

min_aprox <- min(bc_aprox)
max_aprox <- max(bc_aprox)
media_aprox <- mean(bc_aprox)
mediana_aprox <- median(bc_aprox)
Q1_aprox <- quantile(bc_aprox, 0.25)
Q3_aprox <- quantile(bc_aprox, 0.75)
sesgo_aprox <- skewness(bc_aprox)
curt_aprox <- kurtosis(bc_aprox)

cat("El mínimo del exacto es:", min_exacto, "\n")
```

```
## El mínimo del exacto es: 0
```

```
cat("El máximo del exacto es:", max_exacto, "\n")
```

```
## El máximo del exacto es: 2.12099
```

```
cat("La media del exacto es:", media_exacto, "\n")
```

```
## La media del exacto es: 0.8669697
```

```
cat("La mediana del exacto es:", mediana_exacto, "\n")
```

```
## La mediana del exacto es: 0.8672659
```

```
cat("El Q1 del exacto es:", Q1_exacto, "\n")
```

```
## El Q1 del exacto es: 0.3785005
```

```
cat("El Q3 del exacto es:", Q3_exacto, "\n")
```

```
## El Q3 del exacto es: 1.31967
```

```
cat("El sesgo del exacto es:", sesgo_exacto, "\n")
```

```
## El sesgo del exacto es: 0.09398812
```

```
cat("La kurtosis del exacto es:", curt_exacto, "\n\n")
```

```
## La kurtosis del exacto es: 1.922093
```

```
cat("El mínimo del aproximado es:", min_aprox, "\n")
```

```
## El mínimo del aproximado es: 1
```

```
cat("El máximo del aproximado es:", max_aprox, "\n")
```

```
## El máximo del aproximado es: 6.670832
```

```
cat("La media del aproximado es:", media_aprox, "\n")
```

```
## La media del aproximado es: 1.955075
```

```
cat("La mediana del aproximado es:", mediana_aprox, "\n")
```

```
## La mediana del aproximado es: 1.67332
```

```
cat("El Q1 del aproximado es:", Q1_aprox, "\n")
```

```
## El Q1 del aproximado es: 1.224745
```

```
cat("El Q3 del aproximado es:", Q3_aprox, "\n")
```

```
## El Q3 del aproximado es: 2.408319
```

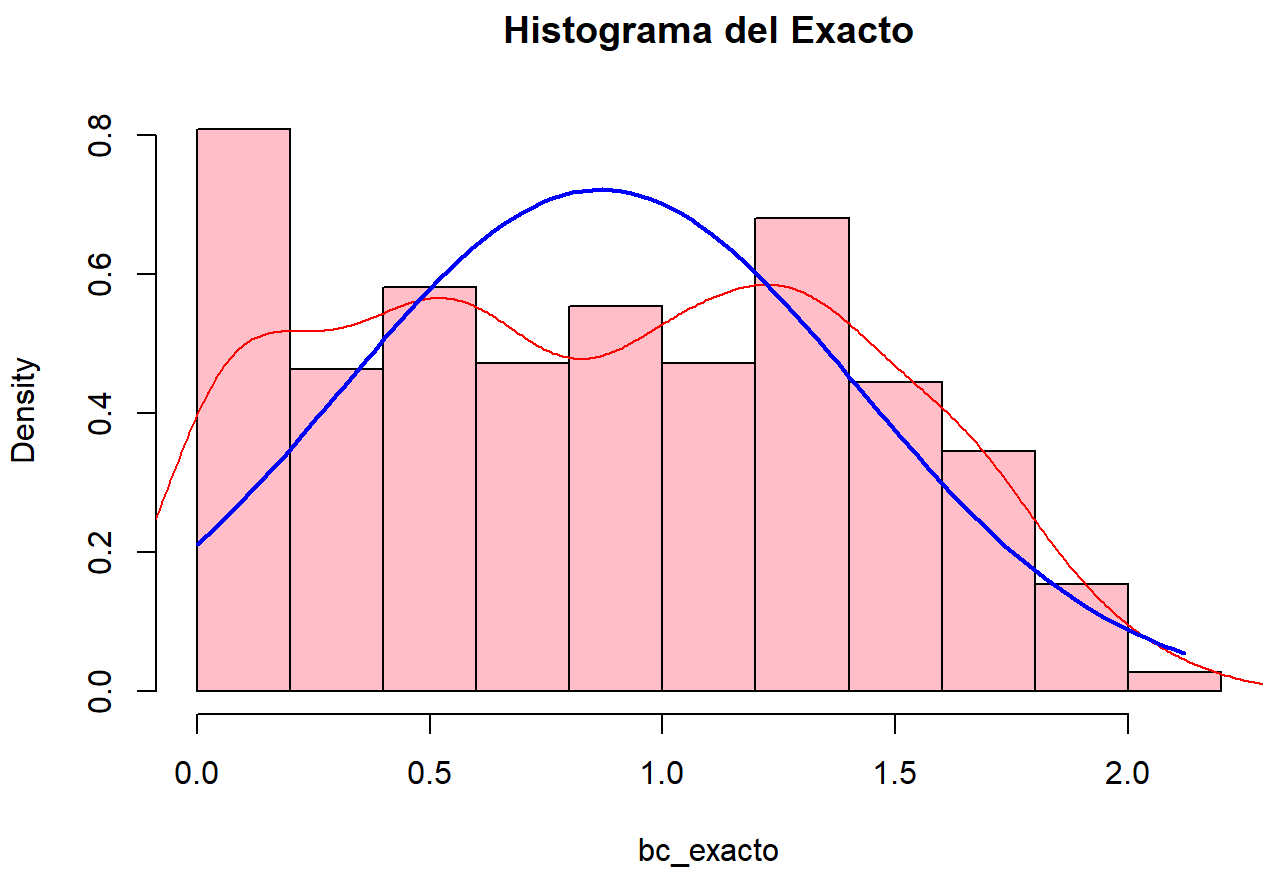
```
cat("El sesgo del aproximado es:", sesgo_aprox, "\n")
```

```
## El sesgo del aproximado es: 1.626347
```

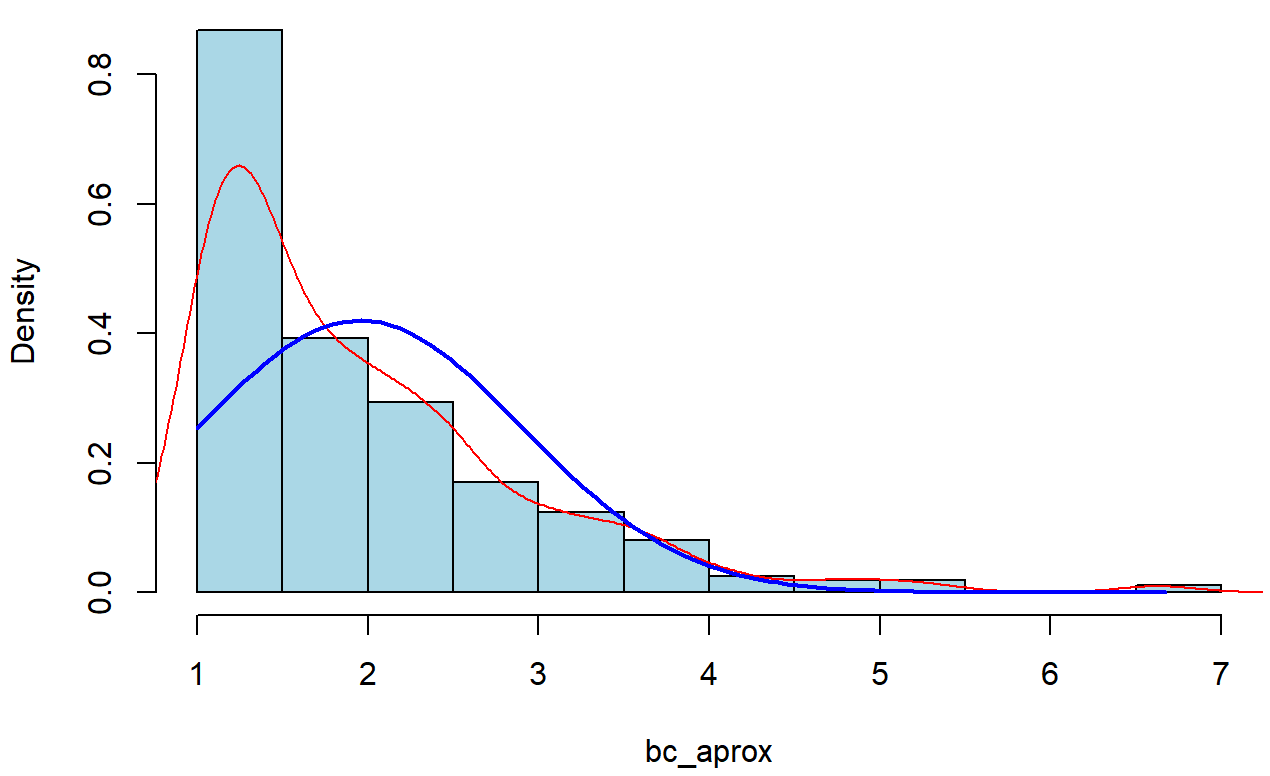
```
cat("La kurtosis del aproximado es:", curt_aprox, "\n")
```

```
## La kurtosis del aproximado es: 6.479268
```

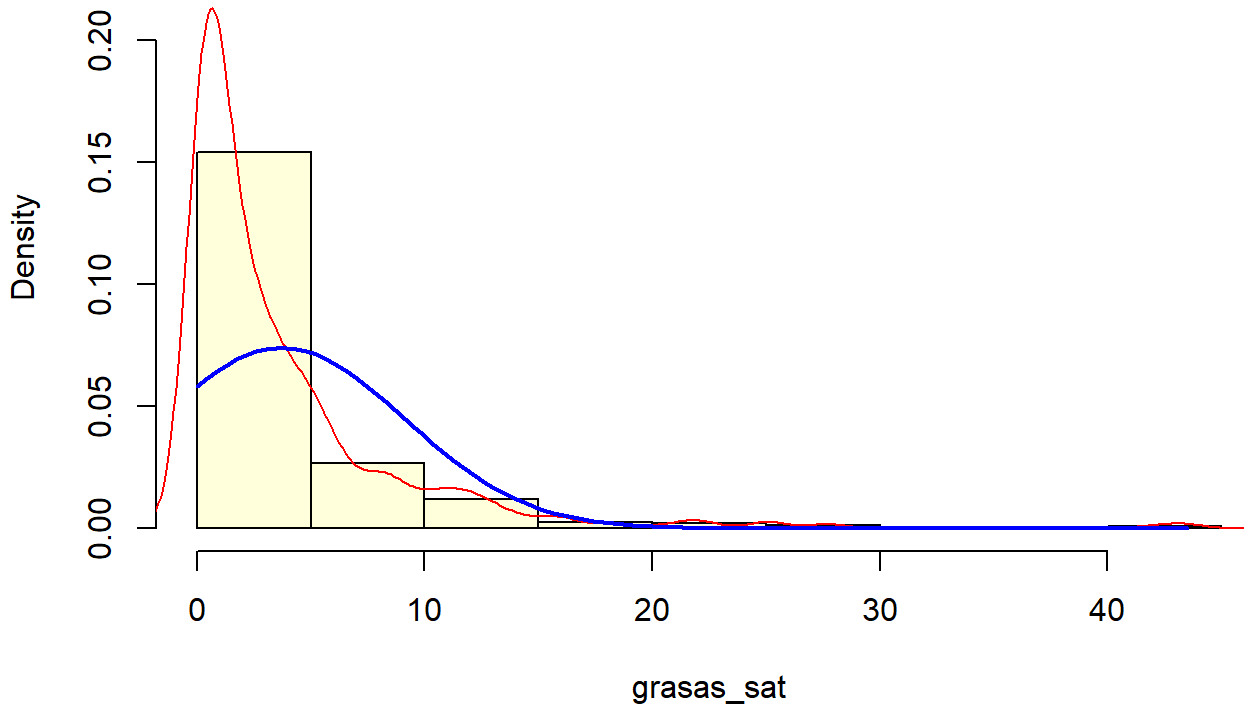
```
# Histogramas

hist(bc_exacto, freq = FALSE, main = "Histograma del Exacto", col = "pink")
lines(density(bc_exacto),col = "red")
curve(dnorm(x, mean = mean(bc_exacto), sd = sd(bc_exacto)), from = min(bc_exacto), to = max(bc_e
xacto), add = TRUE, col = "blue", lwd = 2)
```




```
hist(bc_aprox, freq = FALSE, main = "Histograma del Aproximado", col = "lightblue")
lines(density(bc_aprox),col = "red")
curve(dnorm(x, mean = mean(bc_aprox), sd = sd(bc_aprox)), from = min(bc_aprox), to = max(bc_apro
x), add = TRUE, col = "blue", lwd = 2)
```




```
hist(grasas_sat, freq = FALSE, main = "Histograma de Grasas Saturadas", col = "lightyellow", yli
m = c(0, 0.23))
lines(density(grasas_sat),col = "red")
curve(dnorm(x, mean = mean(grasas_sat), sd = sd(grasas_sat)), from = min(grasas_sat), to = max(g
rasas_sat), add = TRUE, col = "blue", lwd = 2)
```




```
# Prueba de Anderson - Darling y Jarque Bera

# Anderson-Darling

and_dar_exacto <- ad.test(bc_exacto)
cat("Los resultados del Anderson-Darling test para el exacto son: \n")
```

```
## Los resultados del Anderson-Darling test para el exacto son:
```

```
print(and_dar_exacto)
```

```
##
##  Anderson-Darling normality test
##
## data:  bc_exacto
## A = 5.4605, p-value = 1.762e-13
```

```
and_dar_aprox <- ad.test(bc_aprox)
cat("Los resultados del Anderson-Darling test para el aproximado son: \n")
```

```
## Los resultados del Anderson-Darling test para el aproximado son:
```

```
print(and_dar_aprox)
```

```
##
##  Anderson-Darling normality test
##
## data:  bc_aprox
## A = 21.111, p-value < 2.2e-16
```

```
# Jarque-Bera

jarque_b_exacto <- jarque.bera.test(bc_exacto)
cat("Los resultados del Jarque Bera test para el exacto son: \n")
```

```
## Los resultados del Jarque Bera test para el exacto son:
```

```
print(jarque_b_exacto)
```

```
##
##  Jarque Bera Test
##
## data:  bc_exacto
## X-squared = 27.486, df = 2, p-value = 1.075e-06
```

```
jarque_b_aprox<- jarque.bera.test(bc_aprox)
cat("Los resultados del Jarque Bera test son: \n")
```

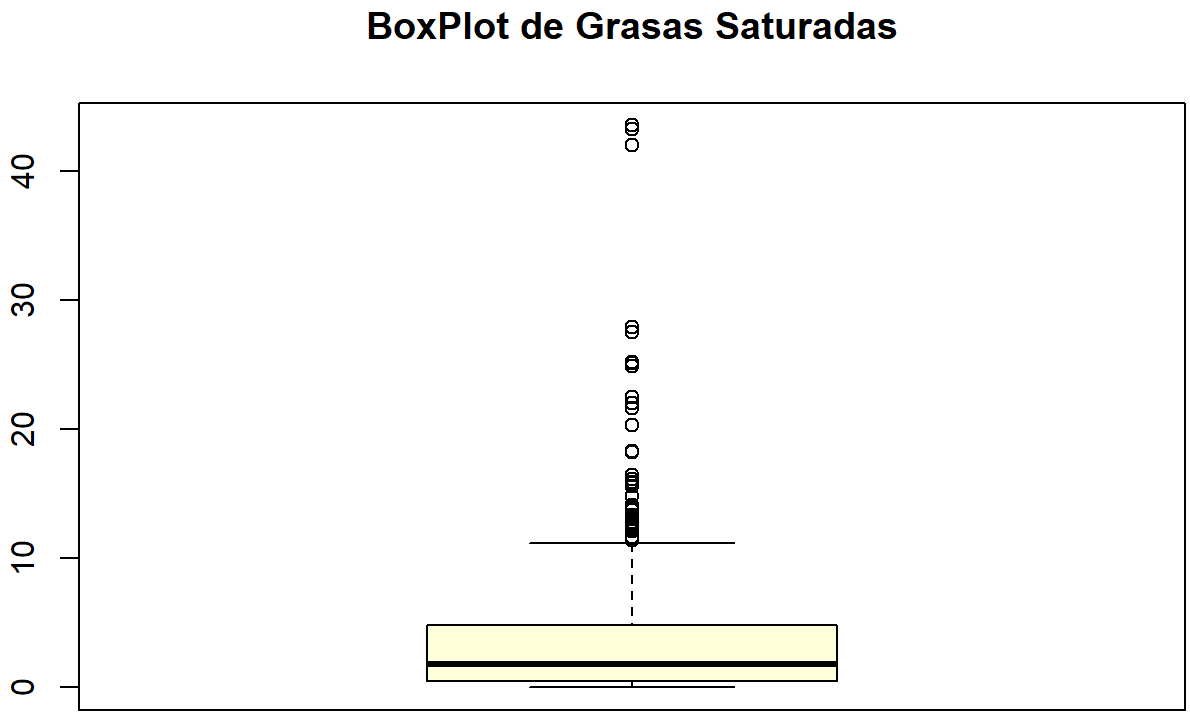
```
## Los resultados del Jarque Bera test son:
```

```
print(jarque_b_aprox)
```

```
##
##  Jarque Bera Test
##
## data:  bc_aprox
## X-squared = 520.82, df = 2, p-value < 2.2e-16
```

- Detecta anomalías y corrige tu base de datos (datos atípicos, ceros anámalos, etc).

```
boxplot(grasas_sat, col = "lightyellow", main = "BoxPlot de Grasas Saturadas")
```




```
Q1 <- quantile(grasas_sat, 0.25)
Q3 <- quantile(grasas_sat, 0.75)
IQR <- Q3 - Q1

outliers <- grasas_sat[grasas_sat < (Q1 - 1.5 * IQR) | grasas_sat > (Q3 + 1.5 * IQR)]
gs_clean <- grasas_sat[!(grasas_sat %in% outliers)]
```

- Comenta la normalidad de las transformaciones obtenidas. Utiliza como argumento de normalidad:
  - Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.
  - Obten el histograma de los 2 modelos obtenidos (exacto y aproximado) y de los datos originales.
  - Interpreta la prueba de normalidad de Anderson-Darling y Jarque Bera para los datos transformados y los originales
  - Indica posibilidades de motivos de alejamiento de normalidad (sesgo, curtosis, datos atípicos, etc)

Usando la transformación de Box-Cox Exacta pude reducir el valor de sesgo y de curtosis, lo que permite acercar los datos al comportamiento de una distribución normal. Igualmente, esto se ve reflejado en las medidas como son la media, mediana, etc. Por otro lado, la aproximada sí redujo los valores, pero no tanto como la exacta, así que es mejor la transformación de Box-Cox Exacta.

Finalmente, podemos observar que a pesar de esto, las pruebas de Anderson-Darling y Jarque Bera siguen rechazando la normalidad, pero se acercan más.

- Define la mejor transformación de los datos de acuerdo a las características de los modelos que encontraste. Toma en cuenta los criterios del inciso anterior para analizar normalidad y la economía del modelo.

Para terminar, puedo concluir que la transformación exacta de Box-Cox es la mejor ya que nos permite acercar más los datos a una distribución normal.