Actividad_Integradora1

Rodolfo Jesús Cruz Rebollar

2024-08-20

```
# Cargar los datos del archivo csv

food = read.csv("food.csv")

head(food)
```

```
##   X Unnamed..0                             food Caloric.Value  Fat
## 1 0          0                    cream cheese            51  5.0
## 2 1          1                neufchatel cheese           215 19.4
## 3 2          2 requeijao cremoso light catupiry            49  3.6
## 4 3          3                   ricotta cheese            30  2.0
## 5 4          4             cream cheese low fat            30  2.3
## 6 5          5            cream cheese fat free            19  0.2
##   Saturated.Fats Monounsaturated.Fats Polyunsaturated.Fats
Carbohydrates Sugars
## 1            2.9                1.300                0.200
0.8  0.500
## 2           10.9                4.900                0.800
3.1  2.700
## 3            2.3                0.900                0.000
0.9  3.400
## 4            1.3                0.500                0.002
1.5  0.091
## 5            1.4                0.600                0.042
1.2  0.900
## 6            0.1                0.091                0.075
1.4  1.000
##   Protein Dietary.Fiber Cholesterol Sodium Water Vitamin.A Vitamin.B1
## 1     0.9           0.0        14.6  0.016   7.6     0.200      0.033
## 2     7.8           0.0        62.9  0.300  53.6     0.200      0.099
## 3     0.8           0.1         0.0  0.000   0.0     0.000      0.000
## 4     1.5           0.0         9.8  0.017  14.7     0.075      0.019
## 5     1.2           0.0         8.1  0.046  10.0     0.016      0.080
## 6     2.8           0.0         2.2  0.100  12.9     0.063      0.020
##   Vitamin.B11 Vitamin.B12 Vitamin.B2 Vitamin.B3 Vitamin.B5 Vitamin.B6
Vitamin.C
## 1       0.064       0.092      0.097      0.084      0.052      0.096
0.004
## 2       0.079       0.090      0.100      0.200      0.500      0.078
0.000
## 3       0.000       0.000      0.000      0.000      0.000      0.000
0.000
```

```
## 4       0.079       0.091      0.027      0.041      0.016      0.007
0.006
## 5       0.062       0.049      0.026      0.080      0.100      0.003
0.000
## 6       0.089       0.092      0.021      0.025      0.200      0.038
0.000
##   Vitamin.D Vitamin.E Vitamin.K Calcium Copper  Iron Magnesium
Manganese
## 1     0.000     0.000     0.100   0.008 14.100 0.082     0.027
1.300
## 2     0.000     0.300     0.045  99.500  0.034 0.100     8.500
0.088
## 3     0.000     0.000     0.000   0.000  0.000 0.000     0.000
0.000
## 4     0.000     0.001     0.011   0.097 41.200 0.097     0.096
4.000
## 5     0.036     0.009     0.019  22.200  0.072 0.008     1.200
0.098
## 6     0.000     0.049     0.059  63.200  0.039 0.053     4.000
0.028
##   Phosphorus Potassium Selenium  Zinc Nutrition.Density
## 1      0.091      15.5   19.100 0.039             7.070
## 2    117.300     129.2    0.054 0.700           130.100
## 3      0.000       0.0    0.000 0.000             5.400
## 4      0.024      30.8   43.800 0.035             5.196
## 5     22.800      37.1    0.034 0.053            27.007
## 6     94.100      50.0    0.013 0.300            67.679
```

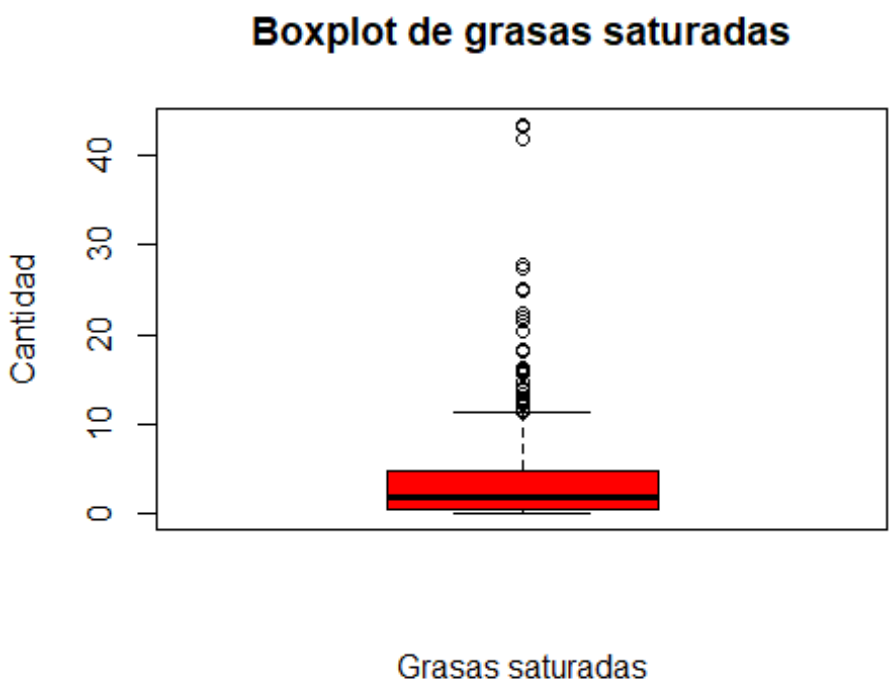
Punto 1. Análisis descriptivo de la variable grasas saturadas

```
# Colocar datos de grasas saturadas en una variable por separado

grasas_sat = food$Saturated.Fats

head(grasas_sat)

## [1]  2.9 10.9  2.3  1.3  1.4  0.1
```

Análisis de datos atípicos

Diagrama de caja y bigote

```
# Diagrama de caja y bigote de grasas saturadas

boxplot(grasas_sat, col = "red", main = "Boxplot de grasas saturadas",
        xlab = "Grasas saturadas", ylab = "Cantidad")
```




```
# Medidas para identificar datos atípicos

medidas = summary(grasas_sat)

medidas
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.500   1.800   3.723   4.800  43.500
```

Cota de 1.5 rangos intercuartílicos

```
# Calcular cota inferior

c_inf = medidas[2] - 1.5 * (medidas[5] - medidas[2])

# Calcular cota superior

c_sup = medidas[5] + 1.5 * (medidas[5] - medidas[2])

cat("La cota para 1.5 rangos intercuartílicos es: ", "\n", "Cota
inferior: ",
    c_inf, "\n", "Cota superior: ", c_sup)
```

```
## La cota para 1.5 rangos intercuartílicos es:
##  Cota inferior:  -5.95
##  Cota superior:  11.25
```

```r
# Buscar datos atipicos por el criterio de 1.5 rangos intercuartilicos

grasas_sat[(grasas_sat < c_inf) | (grasas_sat > c_sup)]
```

```
##  [1] 22.0 43.5 20.3 12.8 16.4 16.1 13.3 24.9 25.2 15.8 27.5 13.0 22.5
25.1 43.2
## [16] 11.5 11.4 12.2 14.1 11.4 11.6 18.2 12.4 14.0 11.4 12.6 14.8 13.7
15.6 11.6
## [31] 12.5 15.9 21.6 27.9 13.9 42.0 18.3 21.6 12.3 12.4 12.1 12.4
```

Como se observa, mediante este criterio hay en total 42 datos atípicos.

Cota de 3 desviaciones estandar alrededor de la media

```r
# Dado que los datos atípicos se ubican a más de 3 desviaciones estándar
de la
# media, se calcularán los valores de x correspondientes a z = -3 y z = 3

# Para z = -3

cota3_inf = -3 * sd(grasas_sat) + mean(grasas_sat)

# Para z = 3

cota3_sup = 3 * sd(grasas_sat) + mean(grasas_sat)

cat("La cota de 3 desviaciones estándar para datos atípicos es: ", "\n",
    "Cota inferior: ", cota3_inf, "\n", "Cota superior: ", cota3_sup)
```

```
## La cota de 3 desviaciones estándar para datos atípicos es:
##  Cota inferior:  -12.46835
##  Cota superior:  19.91378
```

```r
# buscar datos atipicos con el criterio de 3 desv estandar alrededor de
la media

grasas_sat[(grasas_sat < cota3_inf) | (grasas_sat > cota3_sup)]
```

```
##  [1] 22.0 43.5 20.3 24.9 25.2 27.5 22.5 25.1 43.2 21.6 27.9 42.0 21.6
```

Se observa que de acuerdo al criterio de 3 desviaciones estándar alrededor de la media, se encuentran un total de 13 datos atípicos.

Cota de 3 rangos intercuartilicos para datos extremos

```r
# Calcular la cota para buscar datos atipicos con el criterio de 3 rangos
intercuartilicos

# cora inferior

cota_infextremos = medidas[2] - 3 * (medidas[5] - medidas[2])
```

```r
# cota superior

cota_supextremos = medidas[5] + 3 * (medidas[5] - medidas[2])

cat("La cota de 3 rangos intercuartilicos para datos extremos es: ",
"\n",
    "Cota inferior: ", cota_infextremos, "\n", "Cota superior: ",
cota_supextremos)
```

```
## La cota de 3 rangos intercuartilicos para datos extremos es:
##  Cota inferior:  -12.4
##  Cota superior:  17.7
```

```r
# Buscar datos extremos mediante el criterio de 3 rangos intercuartilicos

grasas_sat[(grasas_sat < cota_infextremos) | (grasas_sat >
cota_supextremos)]
```

```
##  [1] 22.0 43.5 20.3 24.9 25.2 27.5 22.5 25.1 43.2 18.2 21.6 27.9 42.0
18.3 21.6
```

Se aprecia que en total se encuentran 15 datos extremos por medio del criterio de 3 rangos intercuartilicos.

Interpretación sobre los resultados y el comportamiento de datos atípicos/extremos

Después de obtener todos los resultados anteriores, se puede apreciar que en cuanto al comportamiento tanto de los datos atipicos como extremos, el criterio que logra identificar una mayor cantidad de los mismos es el de los 1.5 rangos intercuartílicos, teniendo este mismo una cota mínima positiva de 11.25 para clasificar a los datos encontrados como atípicos, por lo que al tener la cota menor de todos los métodos implementados, éste criterio logra identificar la mayor cantidad posible de datos atipicos ya que no se requiere que los datos sean muy grandes para clasificarlos como tal, esto a comparación del criterio de las 3 desviaciones estándar alrededor de la media, cuya cota mínima positiva para clasificar datos atípicos es de 19.91, lo cual permite que dicho criterio prácticamente logre detectar una cantidad significativamente menor de datos atípicos que los otros 2, además, en relación al comportamiento de los datos extremos y atípicos en lso 3 casos reportados, se evidencia que los datos en cuestión tienen una variación medianamente significativa entre ellos, lo cual a su vez ocasiona que algunos de ellos sean considerablemente grandes, motivo por el cual, pasan a considerarse como datos atípicos/extremos por los 3 criterios empleados, además de que también cabe mencionar que dichos datos atípicos/extremos se ubican en su totalidad en un rango de valores superior a 10 unidades de grasas saturadas, teniendo el mayor de ellos, un valor numérico que sobrepasa las 40 unidades de grasas saturadas, por lo cual, en resumen, los datos atípicos se ubican entre 10 y 40 unidades de grasas saturadas por lo que en pocas palabras una característica general de suc omportamiento radica en el hecho de quese

ubican todos entre el cuartil 3 y el valor máximo de todo el conjunto de datos (cuartil 4).

Análisis de normalidad

Prueba de hipótesis:

H_0: la variable grasas saturadas sigue una distribución normal.

H_1: la variable grasas saturadas no sigue una distribución normal.

Prueba de normalidad de Anderson Darling

```
# Importar librería para realizar tests de normalidad

library(nortest)

# Realizar test de normalidad de Anderson Darling para los datos de
grasas saturadas

ad.test(grasas_sat)
```

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

Prueba de normalidad de jarque Bera

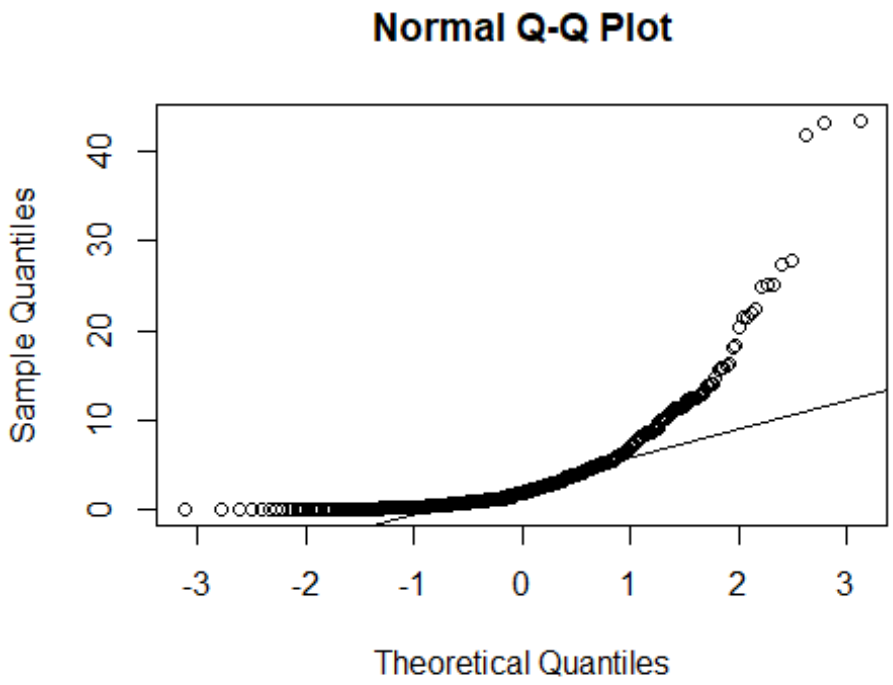
```
# Librería moments para realizar test de normalidad de jarque bera

library(moments)

# Aplicar test de normalidad de Jarque Bera sobre los datos de grasas
saturadas

jarque.test(grasas_sat)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  grasas_sat
## JB = 7694.1, p-value < 2.2e-16
## alternative hypothesis: greater
```

Graficar QQ plot para las grasas saturadas

```
# Graficar los percentiles para el gráfico QQ plot

qqnorm(grasas_sat)
```

```r
# Graficar la línea del gráfico de QQ plot

qqline(grasas_sat)
```



Coeficiente de sesgo y de curtosis

```r
# Calcular coeficiente de sesgo

cat("Sesgo: ", skewness(grasas_sat), "\n")

## Sesgo:  3.428631

# Calcular coeficiente de curtosis

cat("Curtosis: ", kurtosis(grasas_sat))

## Curtosis:  19.97384
```

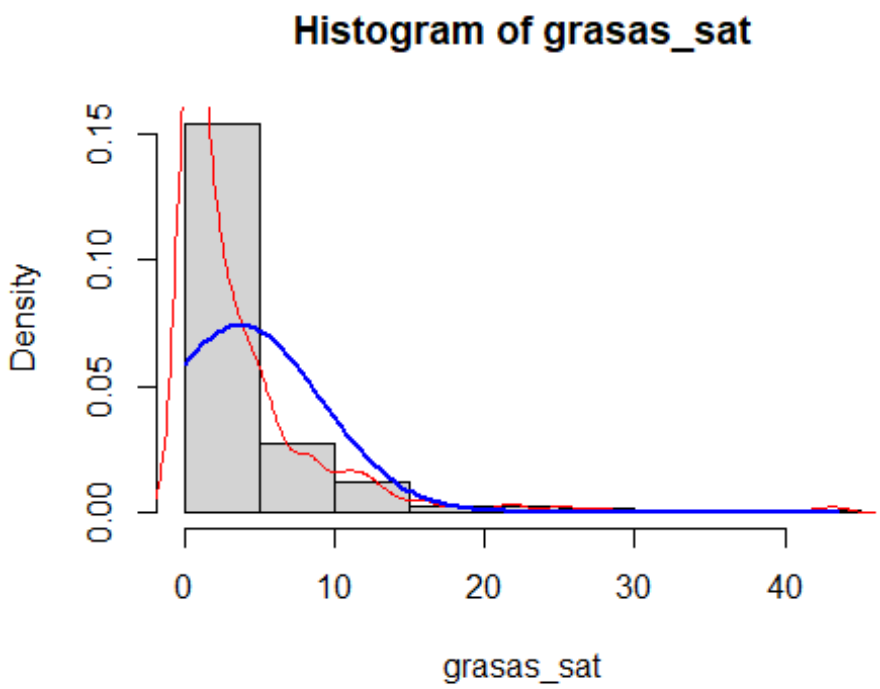
Media, mediana y rango medio de grasas saturadas

```r
# Media

cat("Media: ", mean(grasas_sat), "\n")

## Media:  3.722715
```

```r
# Mediana

cat("Mediana: ", median(grasas_sat), "\n")

## Mediana:  1.8

# Rango medio

cat("Rango medio: ", mean(c(min(grasas_sat), max(grasas_sat))))

## Rango medio:  21.75
```

Gráfico de densidad empírica y teórica

```r
# Histograma de los datos de grasas saturadas

hist(grasas_sat, freq=FALSE)

lines(density(grasas_sat), col="red") # distribución empírica

curve(dnorm(x, mean=mean(grasas_sat), sd=sd(grasas_sat)),
      from = 0, to = 43.5, add = TRUE, col = "blue", lwd = 2) #
distribución teórica
```



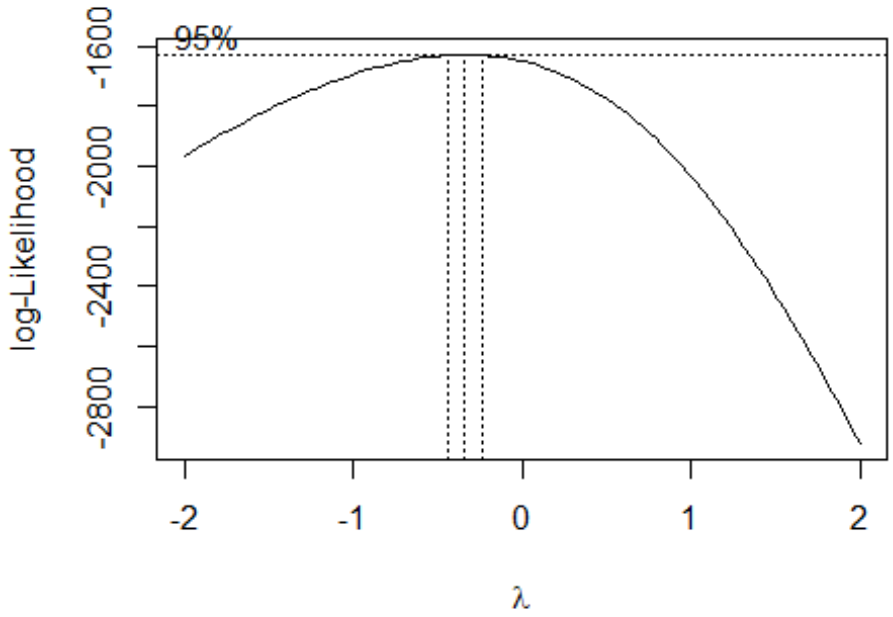
De acuerdo con los resultados obtenidos anteriormente, se puede apreciar que primeramente, en base a las pruebas de normalidad realizadas, la de Anderson Darling arroja un p valor inferior a 2.2e-16, mientras que la de Jarque Bera arroja un p

valor igualmente menor que 2.2e-16, motivo por el cual de acuerdo a ambas pruebas, se rechaza H_0, por lo que se afirma que los datos de la variable grasas saturadas no siguen una distribución normal. Además de lo anterior, también se aprecia que de acuerdo al QQ plot de los datos, en los extremos de la línea recta que indica el caso ideal de una distribución normal, los puntos graficados se alejan de forma gradual de la línea recta del gráfico, motivo por el cual el QQ plot también indica que no hay normalidad entre los datos de grasas saturadas, apoyando así la conclusión obtenida mediante los tests de normalidad previos. Por otro lado, al momento de calcular tanto el coeficiente de sesgo como de curtosis de los datos, se observa que la curtosis de 19.97 es bastante alta, mientras que al mismo tiempo, el valor del sesgo es de 3.42, lo cual también es considerablemente elevado, motivo por el cual, eso indica que la distribución de los datos en cuestión es bastante asimétrica, por lo que se encuentra muy lejos de ser una distribución normal, apoyando las conclusiones obtenidas tanto del qq plot como de los test de normalidad realizados. Adicionalmente, al momento de calcular algunas medidas estadísticas de los datos, tales como media, mediana y rango medio, es necesario considerar que para que la distribución de los datos sea normal o mesocúrtica, las tres medidas deben tener valores iguales, lo cual no sucede en el caso de los datos analizados, ya que la media de los mismos es de 3.72, la mediana es de 1.8 y el rango medio es de 21.75, por lo que se aprecia que las 3 medidas calculadas poseen valores significativamente lejanos entre sí, lo cual respalda la conclusión de que los datos analizados no siguen una distribución que sea normal. Finalmente, en relación al gráfico de la distribución empírica y teórica de los datos, se aprecia que la distribución de los datos dada por el histograma y por las curvas graficada en color rojo, se aleja bastante de la curva graficada en color azul que representa la distribución teórica de los datos (caso ideal si tuvieran distibución normal), motivo por el cual, se concluye a partir de ello que los datos no tienen una distribución normal, además dado que en el gráfico de comparación entre distribuciones se aprecia que la región derecha del gráfico no tiene casi datos, y es justamente la región donde se ubican los datos atípicos encontrados, por lo que aparentemente, los datos atípicos tienen una influencia casi nula en la normalidad de los datos en cuestión.

Punto 2. Transformación a normalidad

```r
# Sumar 1 a cada uno de los datos originales de grasas saturadas para
eliminar los valores 0

grasas_sat = grasas_sat + 1

# Verificar que se hayan eliminado los valores 0

grasas_sat[!(grasas_sat > 0)]

## numeric(0)

# Biblioteca MASS para realizar transformación de Box-Cox

library(MASS)
```

```r
# Realizar la transformación de Box-Cox para la variable grasas saturadas

boxcox_grasas = boxcox(grasas_sat ~ 1)
```




```r
# Calcular el valor del Lambda que maximiza la función de verosimilitud

lambda_optimo = boxcox_grasas$x[which.max(boxcox_grasas$y)]

boxcox_grasas
```

```
## $x
##   [1] -2.00000000 -1.95959596 -1.91919192 -1.87878788 -1.83838384 -
1.79797980
##   [7] -1.75757576 -1.71717172 -1.67676768 -1.63636364 -1.59595960 -
1.55555556
##  [13] -1.51515152 -1.47474747 -1.43434343 -1.39393939 -1.35353535 -
1.31313131
##  [19] -1.27272727 -1.23232323 -1.19191919 -1.15151515 -1.11111111 -
1.07070707
##  [25] -1.03030303 -0.98989899 -0.94949495 -0.90909091 -0.86868687 -
0.82828283
##  [31] -0.78787879 -0.74747475 -0.70707071 -0.66666667 -0.62626263 -
0.58585859
##  [37] -0.54545455 -0.50505051 -0.46464646 -0.42424242 -0.38383838 -
0.34343434
##  [43] -0.30303030 -0.26262626 -0.22222222 -0.18181818 -0.14141414 -
```

```
0.10101010
##  [49] -0.06060606 -0.02020202  0.02020202  0.06060606  0.10101010
0.14141414
##  [55]  0.18181818  0.22222222  0.26262626  0.30303030  0.34343434
0.38383838
##  [61]  0.42424242  0.46464646  0.50505051  0.54545455  0.58585859
0.62626263
##  [67]  0.66666667  0.70707071  0.74747475  0.78787879  0.82828283
0.86868687
##  [73]  0.90909091  0.94949495  0.98989899  1.03030303  1.07070707
1.11111111
##  [79]  1.15151515  1.19191919  1.23232323  1.27272727  1.31313131
1.35353535
##  [85]  1.39393939  1.43434343  1.47474747  1.51515152  1.55555556
1.59595960
##  [91]  1.63636364  1.67676768  1.71717172  1.75757576  1.79797980
1.83838384
##  [97]  1.87878788  1.91919192  1.95959596  2.00000000
##
## $y
##   [1] -1968.518 -1954.752 -1941.171 -1927.779 -1914.584 -1901.589 -
1888.799
##   [8] -1876.221 -1863.860 -1851.722 -1839.813 -1828.140 -1816.709 -
1805.526
##  [15] -1794.599 -1783.935 -1773.542 -1763.426 -1753.596 -1744.061 -
1734.828
##  [22] -1725.906 -1717.304 -1709.032 -1701.099 -1693.515 -1686.290 -
1679.435
##  [29] -1672.961 -1666.878 -1661.199 -1655.934 -1651.098 -1646.701 -
1642.757
##  [36] -1639.280 -1636.282 -1633.778 -1631.783 -1630.310 -1629.376 -
1628.995
##  [43] -1629.183 -1629.957 -1631.332 -1633.324 -1635.952 -1639.231 -
1643.178
##  [50] -1647.811 -1653.147 -1659.202 -1665.994 -1673.540 -1681.856 -
1690.959
##  [57] -1700.864 -1711.585 -1723.139 -1735.539 -1748.797 -1762.926 -
1777.936
##  [64] -1793.837 -1810.638 -1828.345 -1846.965 -1866.500 -1886.955 -
1908.328
##  [71] -1930.619 -1953.825 -1977.943 -2002.965 -2028.884 -2055.690 -
2083.373
##  [78] -2111.920 -2141.317 -2171.549 -2202.599 -2234.451 -2267.086 -
2300.486
##  [85] -2334.630 -2369.498 -2405.070 -2441.326 -2478.243 -2515.802 -
2553.981
##  [92] -2592.759 -2632.115 -2672.030 -2712.483 -2753.454 -2794.925 -
2836.877
##  [99] -2879.292 -2922.150
```

```r
# Desplegar el lambda óptimo para la transformacion de boxcox

cat("Lambda óptimo: ", lambda_optimo)
```

```
## Lambda óptimo:  -0.3434343
```

Dado que el lambda adecuado encontrado se aproxima más al valor de -0.5, por lo cual, se utilizará la expresión $1/sqrt(x)$ para realizar la transformación aproximada de Box Cox para los datos:

```r
# Realizar la transformación aproximada de Box-Cox utilizando la
expresión anterior

# Calcular el cociente de 1 sobre raíz de x de cada uno de los datos

bc_aprox = 1 / sqrt(grasas_sat)

# Mostrar datos transformados mediante el modelo aproximado

head(bc_aprox)
```

```
## [1] 0.5063697 0.2898855 0.5504819 0.6593805 0.6454972 0.9534626
```

Ahora para normalizar mediante el modelo exacto de Box Cox, se procederá a emplear la siguiente expresión matemática, dado que el valor de λ es distinto de 0:

$$f(x,\lambda) = \frac{x^{\lambda} - 1}{\lambda}$$

```r
# Transformación de BoxCox exacta para los datos de grasas saturadas

bc_exacto = (grasas_sat ^ lambda_optimo - 1) / lambda_optimo

# Mostrar datos normalizados con el modelo exacto de boxcox

head(bc_exacto)
```

```
## [1] 1.08717848 1.66789035 0.97943699 0.72437323 0.75611257 0.09376718
```

Ecuaciones de los modelos encontrados

Modelo aproximado:

$$1/sqrt(grasas)$$

Modelo exacto:

$$f(grasas,\lambda) = \frac{grasas^{-0.34} - 1}{-0.34}$$

Comentarios sobre la normalidad de las transformaciones obtenidas

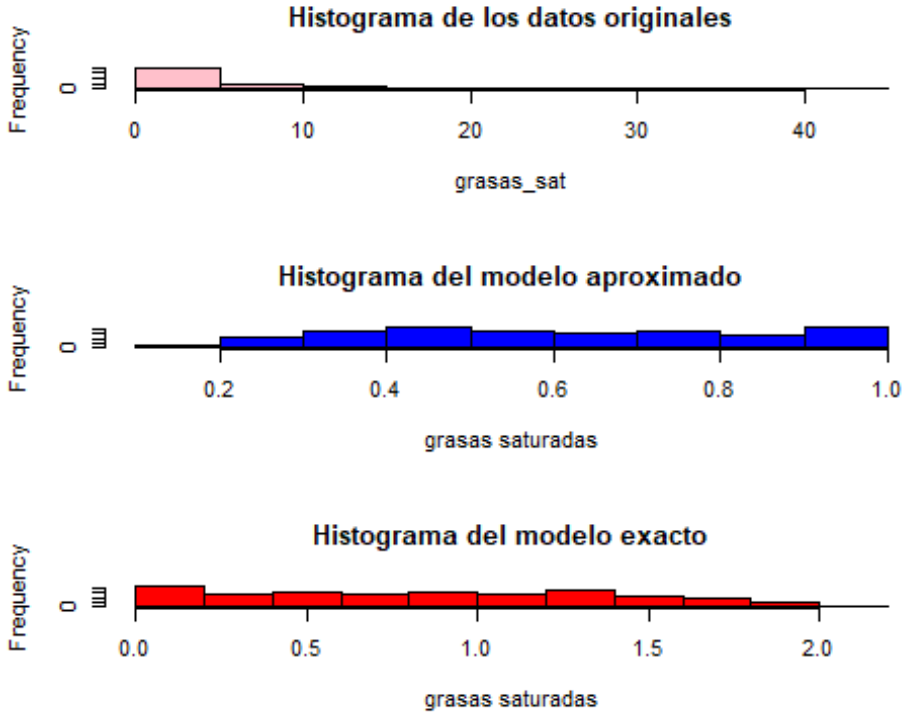
```r
# Medidas estadisticas de los datos transfomados con el modelo aproximado

summary(bc_aprox)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1499  0.4152  0.5976  0.6125  0.8165  1.0000
```

```r
# Medidas estadisticas de los datos transfomados con el modelo exacto

summary(bc_exacto)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3785  0.8673  0.8670  1.3197  2.1210
```

```r
# Graficar histograma de los datos normalizados (con ambos modelos) y los originales

par(mfrow = c(3, 1))

hist(grasas_sat, col = "pink", main = "Histograma de los datos originales")
hist(bc_aprox, col = "blue", main = "Histograma del modelo aproximado",
     xlab ="grasas saturadas")
hist(bc_exacto, col = "red", main = "Histograma del modelo exacto",
     xlab = "grasas saturadas")
```



Prueba de Jarque Bera y Anderson Darling para datos transformados

```r
# Prueba de anderson darling

ad.test(bc_aprox)

##
##  Anderson-Darling normality test
##
## data:  bc_aprox
## A = 6.3651, p-value = 1.221e-15

# prueba de jarque bera

jarque.test(bc_aprox)

##
##  Jarque-Bera Normality Test
##
## data:  bc_aprox
## JB = 31.179, p-value = 1.696e-07
## alternative hypothesis: greater

# anderson darling para modelo exacto

ad.test(bc_exacto)

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

# jarque bera para modelo aproximado

jarque.test(bc_exacto)

##
##  Jarque-Bera Normality Test
##
## data:  bc_exacto
## JB = 27.486, p-value = 1.075e-06
## alternative hypothesis: greater
```

De acuerdo a las pruebas de normalidad de los datos normalizados con ambos modelos de boxcox, se aprecia que el p valor obtenido en cada una de ellas es menor a 0.05, motivo por el cual se rechaza H_0 y se concluye que los datos en general no siguen una distribución normal.

Además, entre las posibilidades de alejamiento de la normal se encuentran el hecho de que los datos en cuestión tienen una alta curtosis además de un sesgo igualmente elevado, lo cual ocasiona principalmente que la distribución de los datos analizados tenga una cierta cantidad de sesgo ya sea positivo, o negativo, además de una

elevación vertical mayormente significativa debida al alto nivel de curtosis de los datos, motivos por los cuales, en términos generales, la transformación de Box Cox no resulta ser la más adecuada para que los datos analizados sigan una distribución normal.

Finalmente, a manera de conclusión de todo el presente análisis, se puede afirmar que de acuerdo a los datos modelos de normalización de boxcox empleados, el que más se acerca a normalizar los datos es el modelo de boxcox exacto, ya que se basa en una expresión claramente definida para normalizar los datos en cuestión a diferencia del método aproximado que se basa en varias posibles expresiones para realizar los mismo en base al valor óptimo que se obtenga del parámetro λ.