ActIntEs_A01571214_Lautaro_Coteja

A01571214 - Lautaro Coteja

2024-08-20

R Markdown

Actividad Integradora A01571214

Variable 4 (SODIO)

CARGAR DATOS

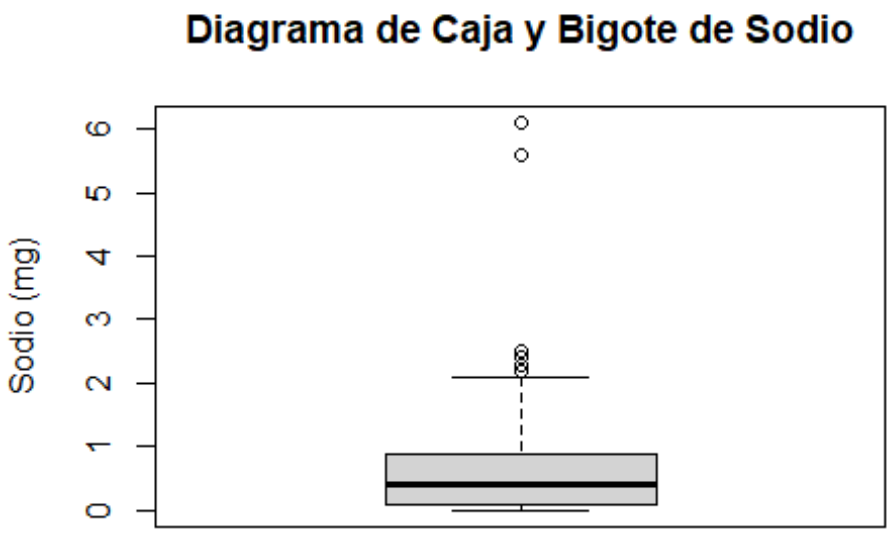
```
data = read.csv("C:/Users/lauta/Downloads/food_data_g.csv")
head(data)
```

```
##   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
```

1) Analizar Datos Atipicos

```
# Caja y Bigote
boxplot(data$Sodium, main = "Diagrama de Caja y Bigote de Sodio", ylab =
"Sodio (mg)")
```




```r
# Resumen
summary(data$Sodium)
```

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

```r
# Deviacion Estandar
sd_sodio = sd(data$Sodium)
sd_sodio
```

```
## [1] 0.6361261
```

```r
# Q1 y Q3
Q1 = quantile(data$Sodium, 0.25)
Q3 = quantile(data$Sodium, 0.75)

# IQR
IQR_sodio = Q3 - Q1
IQR_sodio
```

```
## 75%
## 0.8
```

```r
# Cota de 1.5 Rangos Intercuartilicos
lower_bound_1_5 = Q1 - 1.5 * IQR_sodio
upper_bound_1_5 = Q3 + 1.5 * IQR_sodio

# Contar los Datos Atípicos
```

```r
outliers_1_5 = data$Sodium[data$Sodium < lower_bound_1_5 | data$Sodium >
upper_bound_1_5]
num_outliers_1_5 = length(outliers_1_5)
num_outliers_1_5
```

```
## [1] 8
```

```r
# Cota de 3 Desviaciones Estandar alrededor de la media
mean_sodio = mean(data$Sodium)
lower_bound_3_sd = mean_sodio - 3 * sd_sodio
upper_bound_3_sd = mean_sodio + 3 * sd_sodio

#Contar los Datos Atipicos
outliers_3_sd = data$Sodium[data$Sodium < lower_bound_3_sd | data$Sodium >
upper_bound_3_sd]
num_outliers_3_sd = length(outliers_3_sd)
num_outliers_3_sd
```

```
## [1] 3
```

```r
# Cota de 3 Rangos Intercuartilicos para datos extremos y contar los datos
extremos
lower_bound_3_iqr = Q1 - 3 * IQR_sodio
upper_bound_3_iqr = Q3 + 3 * IQR_sodio

# Contar los Datos Atipicos
extreme_outliers = data$Sodium[data$Sodium < lower_bound_3_iqr | data$Sodium
> upper_bound_3_iqr]
num_extreme_outliers = length(extreme_outliers)
num_extreme_outliers
```

```
## [1] 2
```

Interpretacion de Resultados del Punto 1)

El diagrama de caja y bigote nos permite observar la distribucion de los valores de Sodium y ayuda a identificar visualmente los valoresa atipicos y extremos. Los Q1, Q3 y IQR ayudan con a notar la dispersion de nuestros datos. La SD y la media ayudan con la distribucion de los datos alrededor de la media. En cuanto a los datos atipicos, pudimos ver que si se han encontrado por lo que estos podrian estar afectando la distribucion de los datos, los mas significativos son los que estan mas alla de 3 SD, y los de 3 RI son muy raros. De acuerdo a los resultados podriamos concluir que no es tan preocupante ya que de los ultimos 2 hay pocos. Estos datos atipicos podrian indicar errores en los datos, valores inusuales, etc, y dependiendo de la cantidad se deberia decidir si eliminarlos, analizarlos, o transformarlos.

2

```r
#install.packages("nortest")
#install.packages("tseries")
#install.packages("moments")

library(nortest)
library(tseries)
```

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

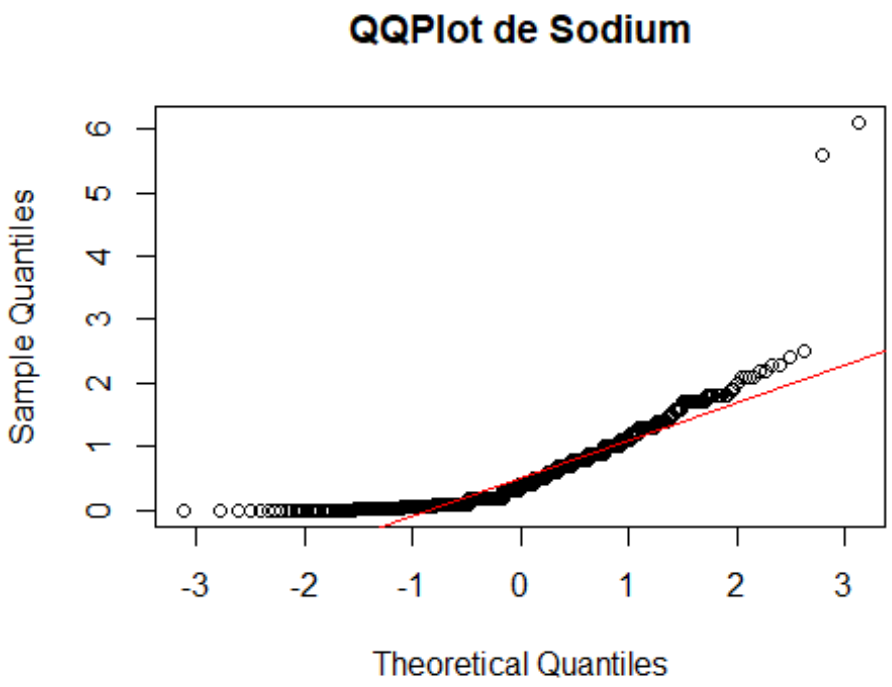
```r
library(moments)

# Anderson-Darling
ad_test = ad.test(data$Sodium)
ad_test
```

```
##
##  Anderson-Darling normality test
##
## data:  data$Sodium
## A = 24.827, p-value < 2.2e-16
```

```r
# Jarque-Bera
jb_test = jarque.bera.test(data$Sodium)
jb_test
```

```
##
##  Jarque Bera Test
##
## data:  data$Sodium
## X-squared = 6834.2, df = 2, p-value < 2.2e-16
```

```r
# QQPlot
qqnorm(data$Sodium, main = "QQPlot de Sodium")
qqline(data$Sodium, col = "red")
```




```r
# Sesgo
sesgo = skewness(data$Sodium)
sesgo
```

```
## [1] 2.735999
```

```r
# Curtosis
curtosis = kurtosis(data$Sodium)
curtosis
```

```
## [1] 19.3626
```

```r
# Medidas
media = mean(data$Sodium)
mediana = median(data$Sodium)
rango_medio = (min(data$Sodium) + max(data$Sodium)) / 2

media
```

```
## [1] 0.5732051
```

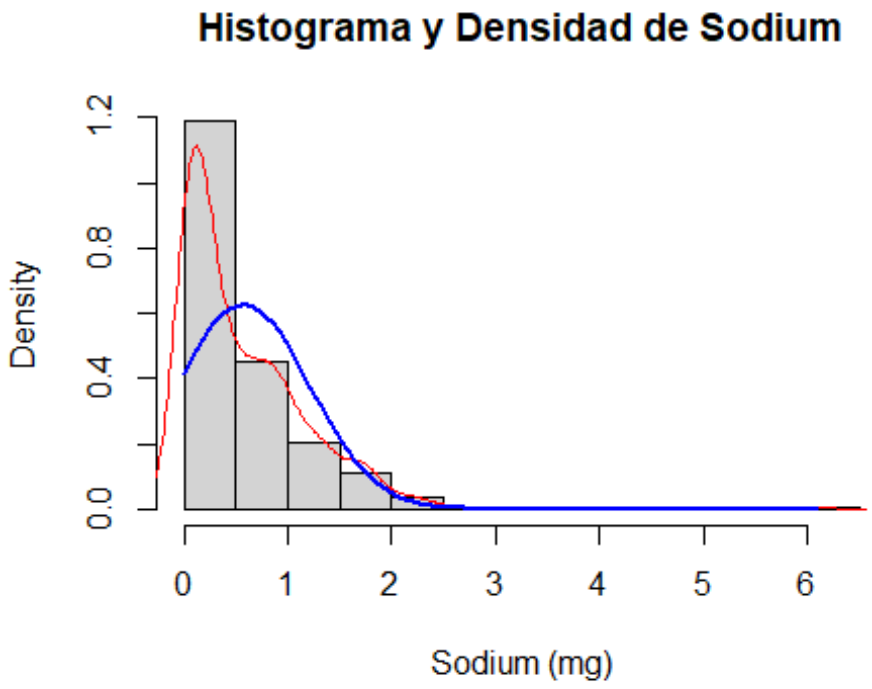
```r
mediana
```

```
## [1] 0.4
```

```r
rango_medio
```

```
## [1] 3.05
```

```r
# Grafico de Densidad Empirica y Teorica Suponiendo Normalidad
hist(data$Sodium, freq = FALSE, main = "Histograma y Densidad de Sodium",
xlab = "Sodium (mg)")
lines(density(data$Sodium), col = "red")  # Densidad empírica
curve(dnorm(x, mean = mean(data$Sodium), sd = sd(data$Sodium)),
      from = min(data$Sodium), to = max(data$Sodium),
      add = TRUE, col = "blue", lwd = 2)  # Densidad teórica bajo normalidad
```



Interpretacion de Resultados del Punto 2)

De acuerdo a las pruebas de normalidad, sabemos que si p es menor, se asume que los datos no siguen una distribucion normal, y si es mayor los datos teoricamente son normales. En cuanto al QQPlot, vemos que mas o menos siguen una linea recta, lo que indica una distribucion normal, aunque si los puntos se desvian de la linea, no sigue una distribucion normal al 100. Un sesgo cerca de 0 indica simetria, y si es negativo indica que es hacia la izquierda y positivo a la derecha, y el sesgo de Sodium salio en 2, por lo que no es totalmente simetrico. La curtosis de una distribucion normal suele ser 3 y aqui salio de 18 por lo que se asume lo mismo que antes. En cuanto a las medidas igual ayudan a analizar la distribucion. Como conclusion en base a todo esto, Sodium parece no tener una distribucion totalmente normal ni simetrica, y los datos atipicos podrian influir esto.

3

```r
#install.packages("MASS")
#install.packages("car")
#install.packages("nortest")
#install.packages("tseries")
#install.packages("bestNormalize")

library(MASS)
library(car)
```

```
## Cargando paquete requerido: carData
```

```r
library(nortest)
library(tseries)
library(bestNormalize)
```

```
##
## Adjuntando el paquete: 'bestNormalize'
```

```
## The following object is masked from 'package:MASS':
##
##     boxcox
```

```r
# Box-Cox y Yeo-Johnson
bn = bestNormalize(data$Sodium)
Sodium_transformed = bn$x.t

yj_transform = yeojohnson(data$Sodium)
Sodium_yj = yj_transform$x.t

# Aplicar Transformacion (Box-Cox)
bn$chosen_transform
```

```
## orderNorm Transformation with 551 nonmissing obs and ties
##  - 96 unique values
##  - Original quantiles:
##   0%  25%  50%  75% 100%
##  0.0  0.1  0.4  0.9  6.1
```

```r
# Aplicar Transformacion (Yeo-Johnson)
yj_transform$lambda
```

```
## [1] -1.237486
```

```r
# Power Transformation
pt = powerTransform(Sodium ~ 1, data = data, family = "yjPower")  # Cambiar
"yjPower" a "bcPower" para Box-Cox
summary(pt)
```

```
## 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) 85.66163  1 < 2.22e-16
```

```
best_lambda_pt = coef(pt, round=TRUE)
```

```
# Comparaciones
summary(data$Sodium)
```

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

```
skewness(data$Sodium)
```

```
## [1] 2.735999
```

```
kurtosis(data$Sodium)
```

```
## [1] 19.3626
```

```
summary(Sodium_transformed)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.235137 -0.589757  0.052340  0.002208  0.690280  3.118964
```

```
skewness(Sodium_transformed)
```

```
## [1] 0.04714618
```

```
kurtosis(Sodium_transformed)
```

```
## [1] 2.833263
```

```
summary(Sodium_yj)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.40848 -0.93979  0.02627  0.00000  0.90053  2.43178
```

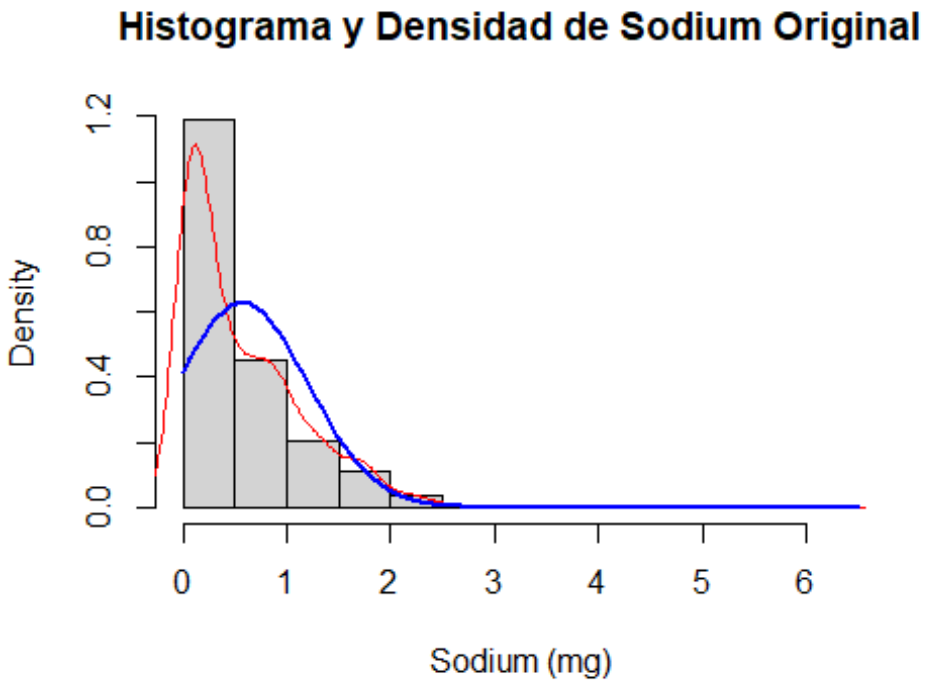
```
skewness(Sodium_yj)
```

```
## [1] 0.1809703
```

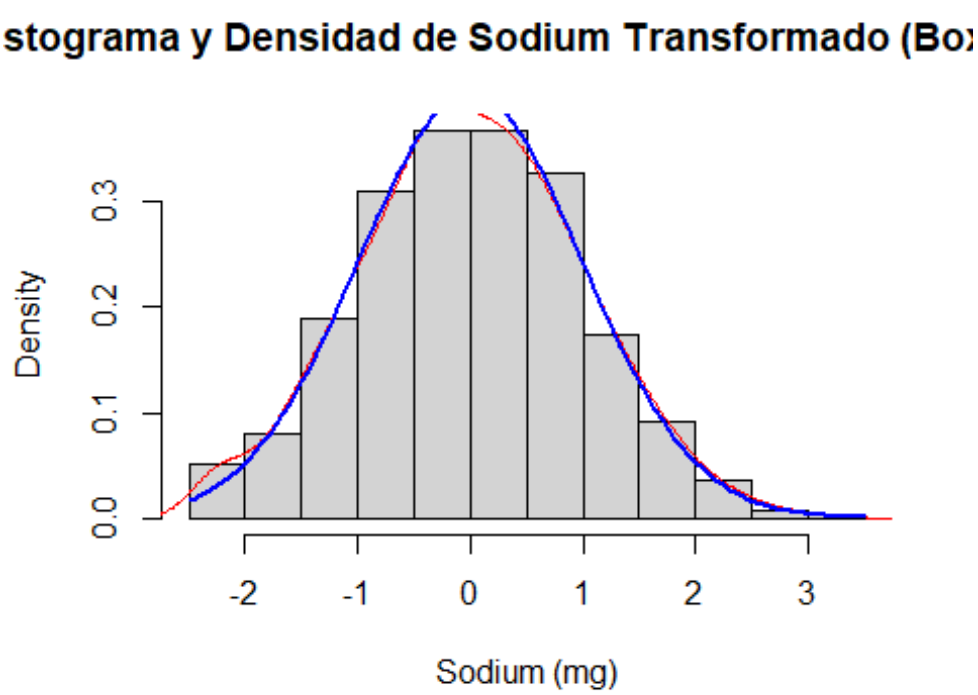
```
kurtosis(Sodium_yj)
```

```
## [1] 1.711956
```

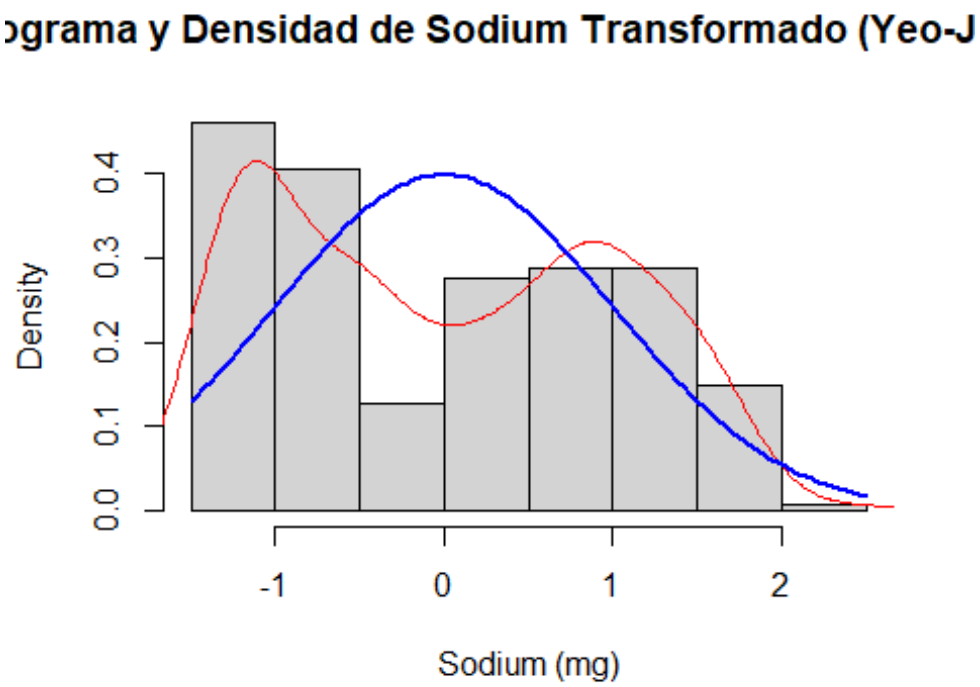
```
# Datos originales
hist(data$Sodium, freq = FALSE, main = "Histograma y Densidad de Sodium
Original", xlab = "Sodium (mg)")
lines(density(data$Sodium), col = "red")
curve(dnorm(x, mean = mean(data$Sodium), sd = sd(data$Sodium)), add = TRUE,
col = "blue", lwd = 2)
```




```r
# Transformación Box-Cox
hist(Sodium_transformed, freq = FALSE, main = "Histograma y Densidad de
Sodium Transformado (Box-Cox)", xlab = "Sodium (mg)")
lines(density(Sodium_transformed), col = "red")
curve(dnorm(x, mean = mean(Sodium_transformed), sd = sd(Sodium_transformed)),
add = TRUE, col = "blue", lwd = 2)
```




```r
# Transformación Yeo-Johnson
hist(Sodium_yj, freq = FALSE, main = "Histograma y Densidad de Sodium
Transformado (Yeo-Johnson)", xlab = "Sodium (mg)")
lines(density(Sodium_yj), col = "red")
curve(dnorm(x, mean = mean(Sodium_yj), sd = sd(Sodium_yj)), add = TRUE, col =
"blue", lwd = 2)
```




```
# Pruebas de normalidad para los datos originales
ad.test(data$Sodium)

##
##  Anderson-Darling normality test
##
## data:  data$Sodium
## A = 24.827, p-value < 2.2e-16

jarque.bera.test(data$Sodium)

##
##  Jarque Bera Test
##
## data:  data$Sodium
## X-squared = 6834.2, df = 2, p-value < 2.2e-16

# Pruebas de normalidad para los datos transformados con Box-Cox
ad.test(Sodium_transformed)

##
##  Anderson-Darling normality test
##
## data:  Sodium_transformed
## A = 0.75479, p-value = 0.04926

jarque.bera.test(Sodium_transformed)
```

```
##
##  Jarque Bera Test
##
## data:  Sodium_transformed
## X-squared = 0.84239, df = 2, p-value = 0.6563
```

```
# Pruebas de normalidad para los datos transformados con Yeo-Johnson
ad.test(Sodium_yj)
```

```
##
##  Anderson-Darling normality test
##
## data:  Sodium_yj
## A = 12.804, p-value < 2.2e-16
```

```
jarque.bera.test(Sodium_yj)
```

```
##
##  Jarque Bera Test
##
## data:  Sodium_yj
## X-squared = 41.097, df = 2, p-value = 1.191e-09
```

Interpretacion de Resultados del Punto 3) y Conclusion

Tras todas las transformaciones, se observo que la de box-cox fue la que tuvo una mejora significativa en la normalidad de Sodium. Las pruebas de Anderson-Darling y Yeo-Johnson indicaron una mayor normalidad despues de la transformacion. Igualmente el sesgo y curtosis mejoraron, indicando que los datos son mas normales despues de esta transformacion. Comparando todo lo mencionado pero con Yeo-Johnson, esta no tuvo mucho impacto en la normalidad de Sodium, por lo que se puede concluir que Box-Cox es la mejor opcion para la normalizacion de los datos de Sodium.