Actividad Integradora 1

Eliezer Cavazos

2024-08-20

Lectura Datos

```
library(nortest)
library(e1071)
library(moments)
```

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

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

```
library(MASS)
```

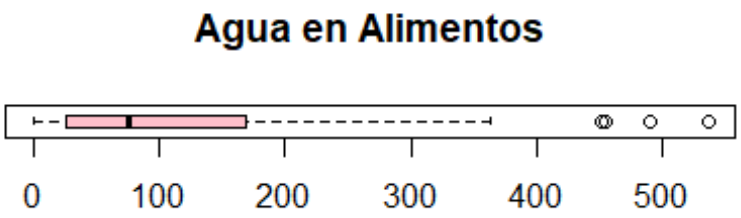
```
oNutricion =
read.csv("C:\\Users\\eliez\\OneDrive\\Desktop\\Clases\\food_data_g.csv")
#Leer la base de datos
```

```
oAgua = oNutricion$Water
```

1. Para analizar datos atípicos se te sugiere:

Graficar el diagrama de caja y bigote

```
par(mfrow=c(2,1))
boxplot(oAgua, horizontal = TRUE,col="pink", main="Agua en Alimentos")
```



Calcula las principales medidas que te ayuden a identificar datos atípicos (utilizar summary te puede abreviar el cálculo): Cuartil 1, Cuartil 2, Media, Cuartil 3, Rango intercuartílico y Desviación estándar

```
D0=ad.test(oAgua)

m0=round(c(as.numeric(summary(oAgua)),kurtosis(oAgua),skewness(oAgua),D0$p.va
lue),3)



m<-as.data.frame(rbind(m0))
row.names(m)=c("Original")
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","
Valor p")

m
```

```
##          Minimo   Q1 Mediana   Media     Q3 Máximo Curtosis Sesgo Valor p
## Original      0 25.9    76.7 101.659 169.05  535.8    4.411 1.084       0
```

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

```
y1 = min(oAgua)
y2 = max(oAgua)
```

```r
q1=quantile(oAgua,0.25)  #Cuantil 1 de la variable Agua
q2=quantile(oAgua,0.50)
q3 = quantile(oAgua, 0.75) # Cuartil 3
#ri= q3-q1   #o
ri=IQR(oAgua)    #Rango intercuartílico de Agua

rangoCuartil1 = q1+1.5*ri
rangoCuartil2 = q2+1.5*ri
rangoCuartil3 = q3+1.5*ri
```

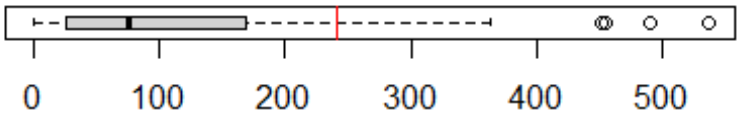
Cuartil 1

```r
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(oAgua,horizontal=TRUE)
abline(v=rangoCuartil1,col="red")  #linea vertical en el límite de los datos atípicos o extremos

oAguaClean= oAgua[oAgua<rangoCuartil1]  #En la matriz M, quitar datos más allá de 1.5 rangos intercuartílicos arriba de q3 de la variable Agua
summary(oAguaClean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   22.85   67.70   89.11  137.90  240.20
```

```r
summary(oAgua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```



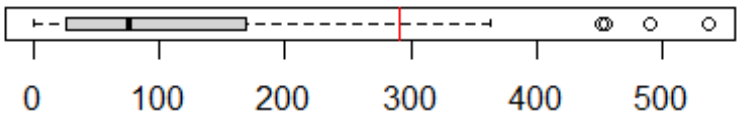
Cuartil 2

```r
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(oAgua,horizontal=TRUE)
abline(v=rangoCuartil2,col="red")  #linea vertical en el límite de los datos atípicos o extremos

oAguaClean= oAgua[oAgua<rangoCuartil2]  #En la matriz M, quitar datos más allá de 1.5 rangos intercuartílicos arriba de q3 de la variable Agua
summary(oAguaClean)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   23.80   73.10   95.56  158.90  289.70
```

```r
summary(oAgua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```



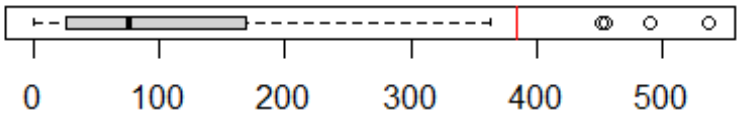
Cuartil 3

```r
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(oAgua,horizontal=TRUE)
abline(v=rangoCuartil3,col="red")  #linea vertical en el límite de los datos atípicos o extremos

oAguaClean= oAgua[oAgua<rangoCuartil3]  #En la matriz M, quitar datos más allá de 1.5 rangos intercuartílicos arriba de q3 de la variable Agua
summary(oAguaClean)
```

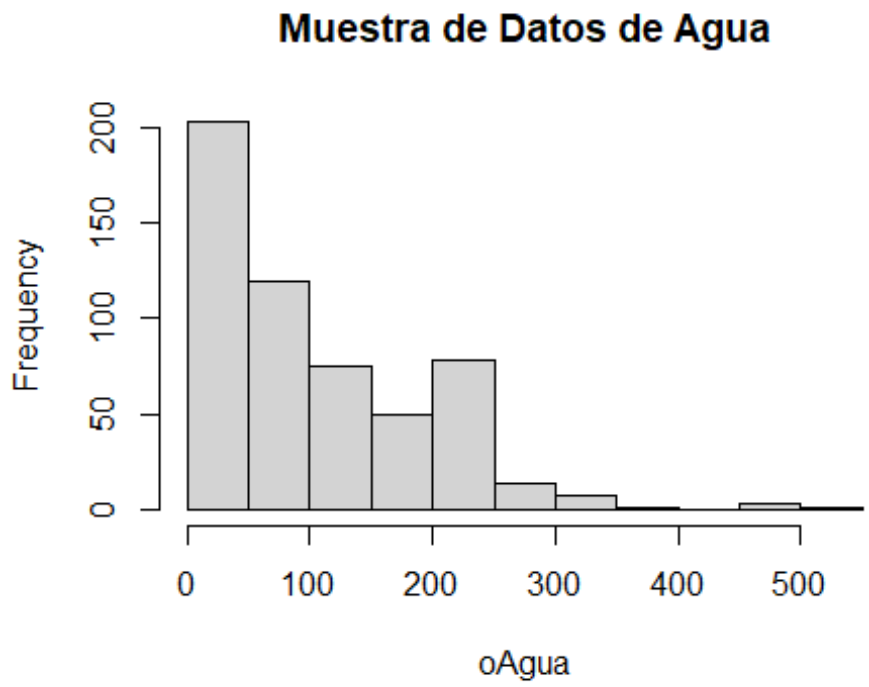
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   24.85   74.40   98.87  166.20  363.60
```

```
summary(oAgua)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    25.9    76.7   101.7   169.1   535.8
```




```
hist(oAgua, main = "Muestra de Datos de Agua")
```



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

```
media_Agua = mean(oAgua)
sd_Agua = sd(oAgua)

media_Agua

## [1] 101.6587

sd_Agua

## [1] 88.50171

min = media_Agua - (3 * sd_Agua)
max = media_Agua + (3 * sd_Agua)
min

## [1] -163.8464

max

## [1] 367.1638
```

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

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

Hay muchos datos atipicos menores y no se muestra viabilidad a que se distribuya de manera normal los datos

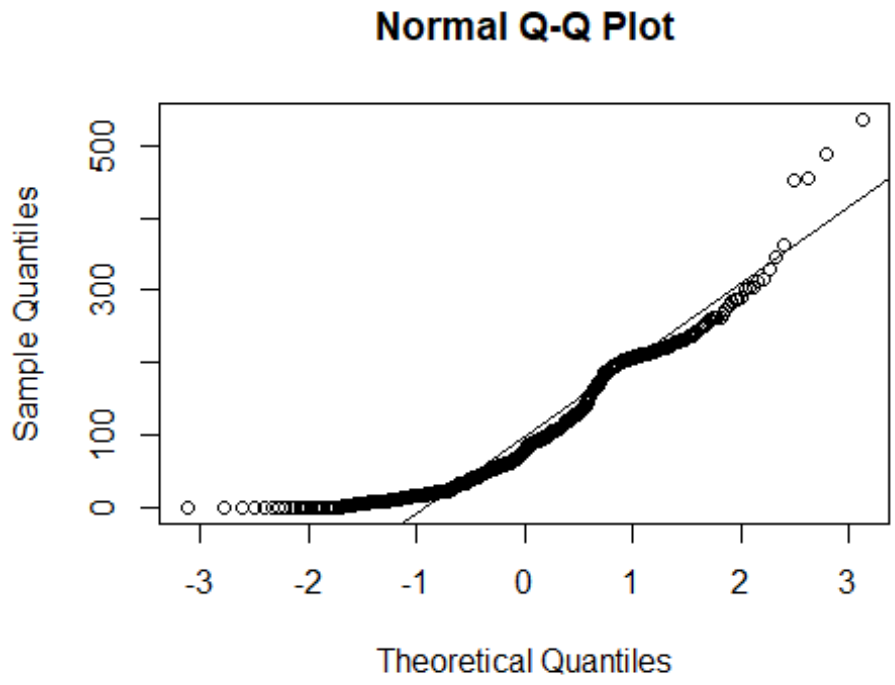
2. Para analizar normalidad se te sugiere:

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.

```
ad.test(oAgua)

##
##  Anderson-Darling normality test
##
## data:  oAgua
## A = 15.968, p-value < 2.2e-16

jarque.test(oAgua)

##
##  Jarque-Bera Normality Test
##
## data:  oAgua
## JB = 153.58, p-value < 2.2e-16
## alternative hypothesis: greater
```

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

```
qqnorm(oAgua)
qqline(oAgua)
```



Calcula el coeficiente de sesgo y el coeficiente de curtosis

```
sg_Agua = skewness(oAgua)
k_Agua = kurtosis(oAgua)

cat("Sesgo: ", sg_Agua)
```

```
## Sesgo:  1.083794
```

```
cat(" Curtosis: ", k_Agua)
```

```
##  Curtosis:  4.411058
```

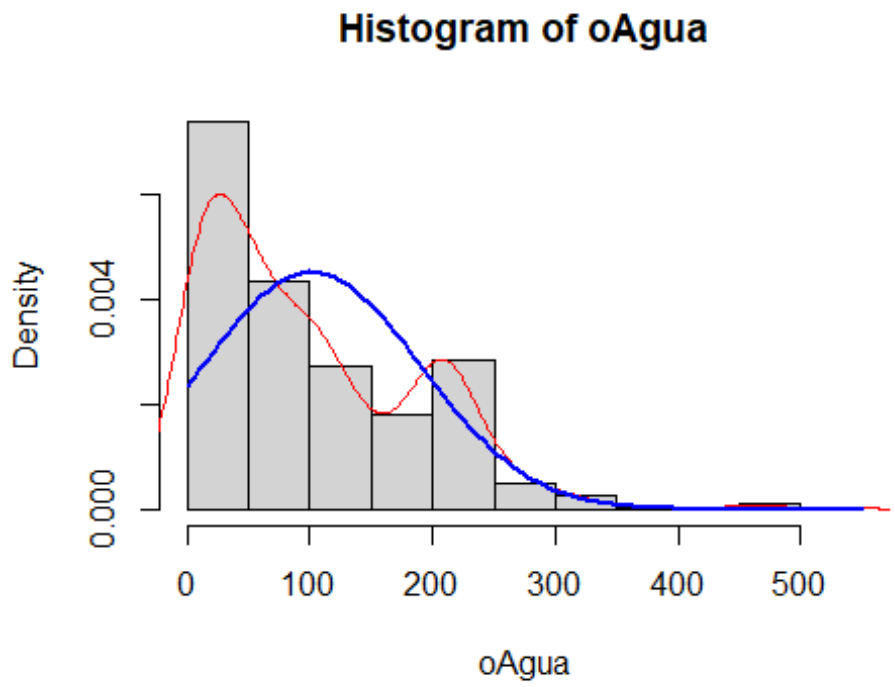
Compara las medidas de media, mediana y rango medio de cada variable

```
m
```

```
##          Minimo   Q1 Mediana   Media     Q3 Máximo Curtosis Sesgo Valor p
## Original      0 25.9    76.7 101.659 169.05  535.8    4.411 1.084       0
```

Realiza el gráfico de densidad empírica y teórica suponiendo normalidad en la variable. Adapta el código:

```
hist(oAgua,freq=FALSE)
lines(density(oAgua),col="red")
curve(dnorm(x,mean=media_Agua,sd=sd_Agua), 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

Los graficos muestran que los datos no se pueden normalizar ya que tiene un sesgo muy grande a la derecha

Comenta las características encontradas:

Considera alejamientos de normalidad por simetría, curtosis

La curtosis es muy grande por lo que se puede decir que la grafica no sigue una distribucion normal

##Comenta si hay aparente influencia de los datos atípicos en la normalidad de los datos Si hay una influencia de los datos atipicos en la normalidad de los datos y más de los datos atipicos iguales a 0 que los datos atipicos mayores al tercer cuartil ##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. Las pruebas de normalidad muestran que el valor de P es muy bajo por lo que se rechaza la hipotesis

Los Graficos muestran mucho sesgo a la derecha y el grafico de qqplot muestra tambien que la mayoria de datos atipicos se encuentran en los valores menores

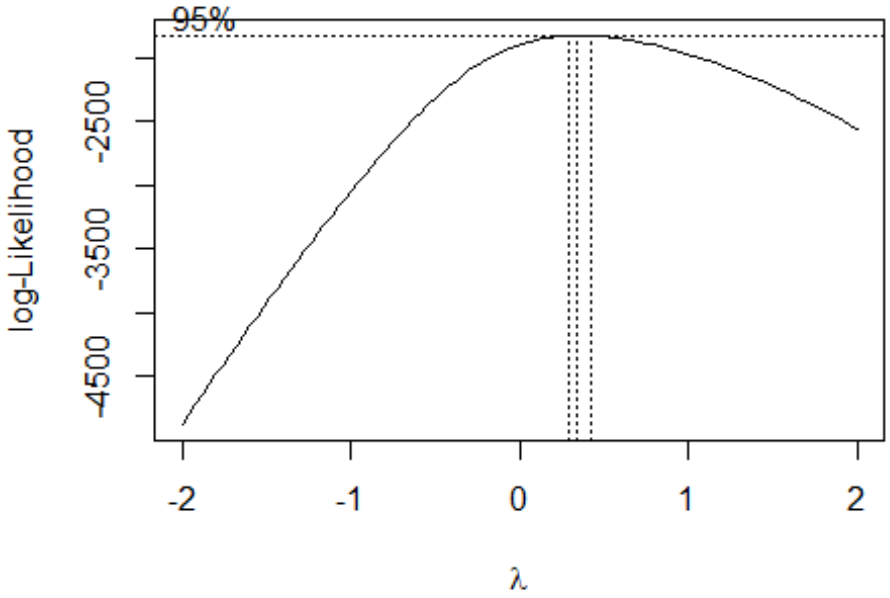
Las medidas de Sesgo y Curtosis demuestran que nuestos datos no son normales

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.

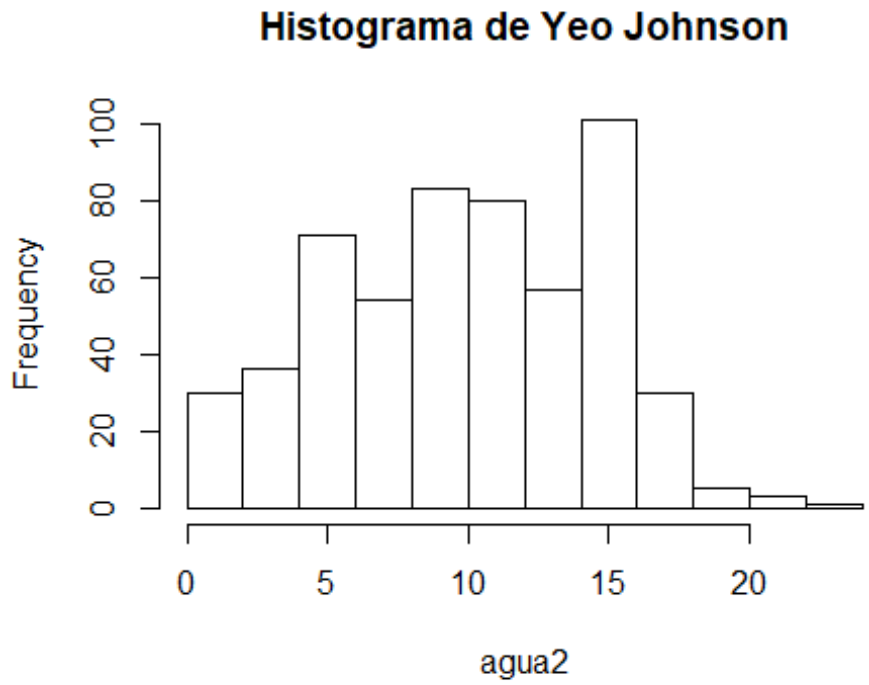
Obtenemos Lambda

```
bc<-boxcox((oAgua+1)~1)# agarra Agua
```

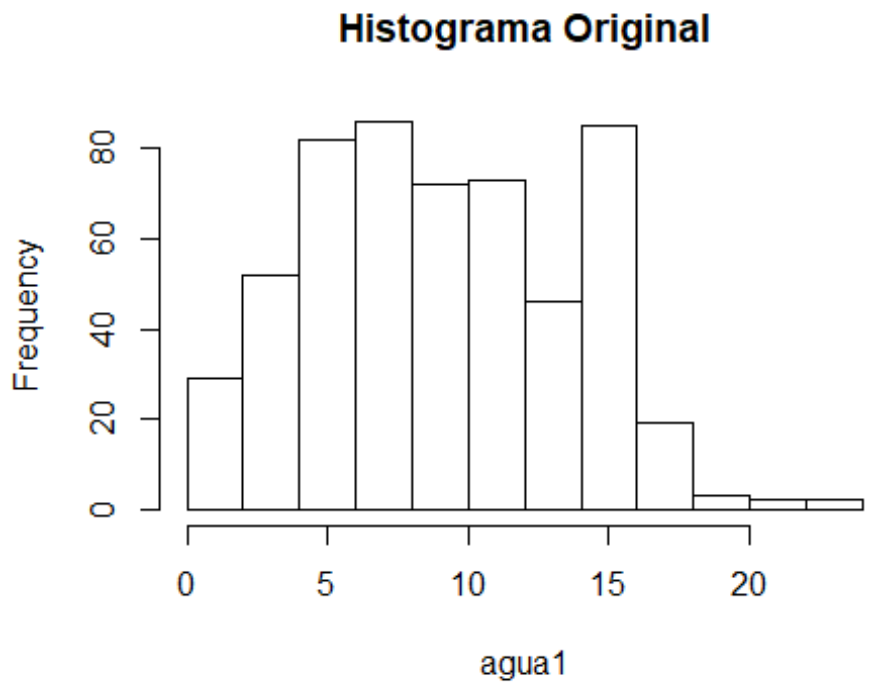



```
lambdaAgua=bc$x[which.max(bc$y)]

lambdaAgua #Lambda

## [1] 0.3434343

library(VGAM)

## Loading required package: stats4

## Loading required package: splines

agua1=sqrt(oAgua+1)
```

```r
agua2<- yeo.johnson(oAgua, lambda = lambdaAgua)
hist(agua2,col=0,main="Histograma de Yeo Johnson")
```




```r
hist(agua1,col=0,main="Histograma Original")
```



Escribe las ecuaciones de los modelos de transformación encontrados.

$\text{Agua1} = \sqrt{x + 1}$ $\text{Agua2} = \frac{(x+1)^{0.343434} - 1}{0.343434}$

3. Analiza la normalidad de las transformaciones obtenidas con los datos originales. Utiliza como argumento de normalidad:

3.1 Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

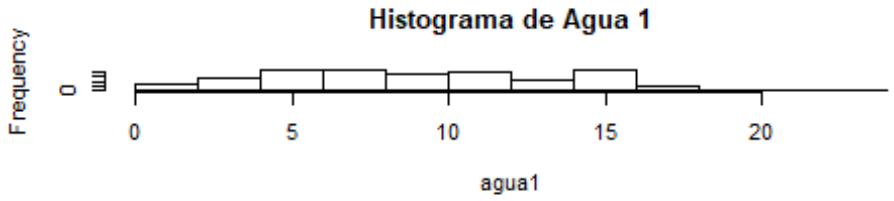
```
#agua2 = ((oAgua+1)^lambda-1)/lambdaSugar

D0=ad.test(oAgua)
D1=ad.test(agua1)
D2=ad.test(agua2)

m0=round(c(as.numeric(summary(oAgua)),kurtosis(oAgua),skewness(oAgua),D0$p.va
lue),3)
m1=round(c(as.numeric(summary(agua1)),kurtosis(agua1),skewness(agua1),D0$p.va
lue),3)
m2=round(c(as.numeric(summary(agua2)),kurtosis(agua2),skewness(agua2),D0$p.va
lue),3)

m<-as.data.frame(rbind(m0, m1, m2))
row.names(m)=c("Original", "Transformacion 1", "Transformacion 2")
```

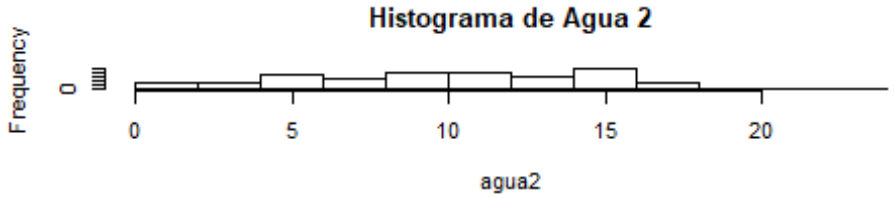
```r
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","
Valor p")

m
```

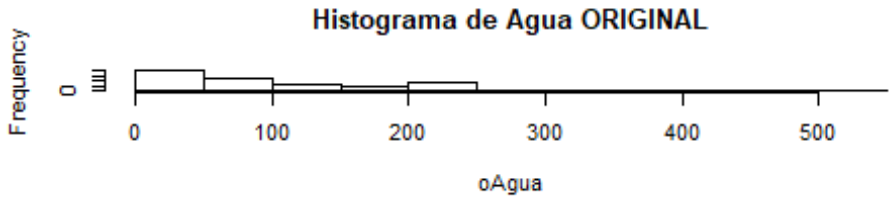
```
##                   Minimo     Q1 Mediana   Media      Q3  Máximo Curtosis
Sesgo
## Original               0 25.900  76.700 101.659 169.050 535.800    4.411
1.084
## Transformacion 1       1  5.187   8.815   9.057  13.040  23.169    2.242
0.154
## Transformacion 2       0  6.108  10.072   9.903  14.079  22.304    2.268
-0.179
##                   Valor p
## Original                0
## Transformacion 1        0
## Transformacion 2        0
```

3.2 Grafica las funciones de densidad empírica y teórica de los 2 modelos obtenidos (exacto y aproximado) y los datos originales.

```r
#agua1 = sqrt(oAguaClean+1)
#agua2 <- yeo.johnson(oAgua, lambda = lambdaAgua)

par(mfrow=c(3,1))
hist(agua1,col=0,main="Histograma de Agua 1")
hist(agua2,col=0,main="Histograma de Agua 2")
hist(oAgua,col=0,main="Histograma de Agua ORIGINAL")
```







3.3 Realiza la prueba de normalidad de Anderson-Darling y de Jarque Bera para los datos transformados y los originales

```
print("Original")
```

```
## [1] "Original"
```

```
jarque.test(oAgua)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  oAgua
## JB = 153.58, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
D0
```

```
##
##  Anderson-Darling normality test
##
## data:  oAgua
## A = 15.968, p-value < 2.2e-16
```

```
print("Transformacion 1")
```

```
## [1] "Transformacion 1"
```

```r
jarque.test(agua1)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  agua1
## JB = 15.364, p-value = 0.000461
## alternative hypothesis: greater
```

```
D1
```

```
##
##  Anderson-Darling normality test
##
## data:  agua1
## A = 4.0333, p-value = 4.785e-10
```

```r
print("Transformacion 2")
```

```
## [1] "Transformacion 2"
```

```r
jarque.test(agua2)
```

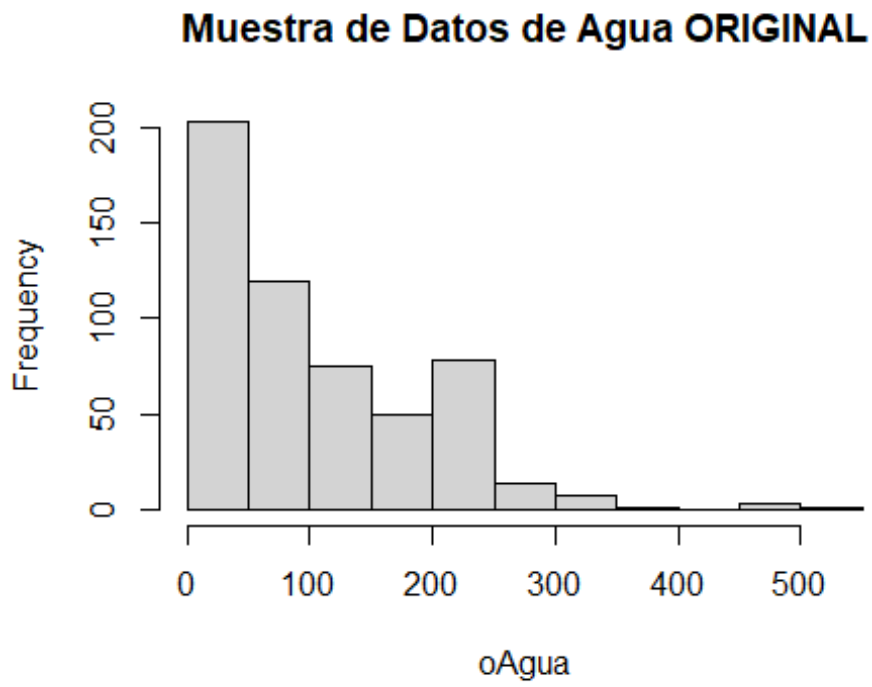
```
##
##  Jarque-Bera Normality Test
##
## data:  agua2
## JB = 15.262, p-value = 0.0004852
## alternative hypothesis: greater
```

```
D2
```

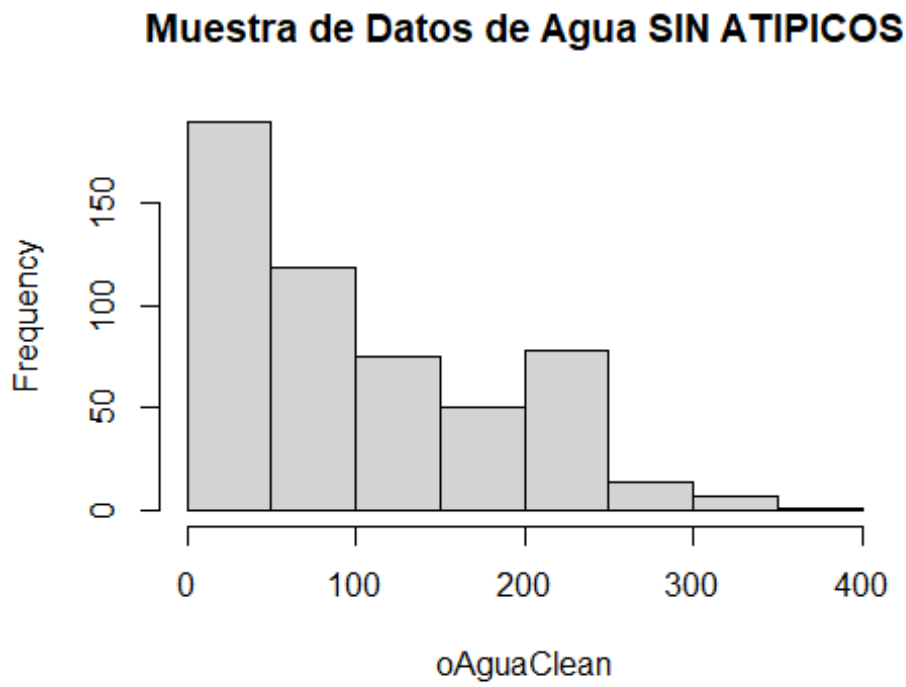
```
##
##  Anderson-Darling normality test
##
## data:  agua2
## A = 3.5229, p-value = 8.238e-09
```

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

```r
oAguaClean= oAgua[0<oAgua]#oAgua[oAgua<rangoCuartil3] # Se quitan datos atipicos más alla de 1.5 rangos intercuantilicos arriba de q3 de la variable Agua

oAguaClean = oAguaClean[oAguaClean<rangoCuartil3]


hist(oAgua, main = "Muestra de Datos de Agua ORIGINAL")
```




```r
hist(oAguaClean, main = "Muestra de Datos de Agua SIN ATIPICOS")
```



5. Comenta la normalidad de las transformaciones obtenidas. Utiliza como argumento de normalidad:

5.1 Compara las medidas: Mínimo, máximo, media, mediana, cuartil 1 y cuartil 3, sesgo y curtosis.

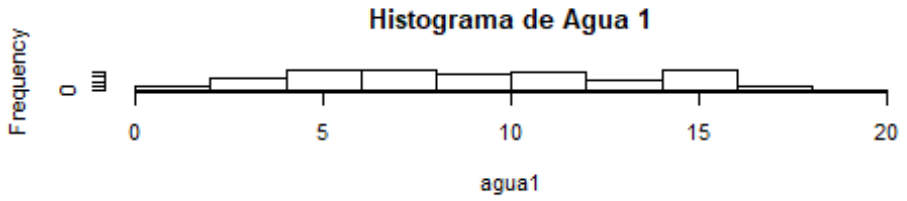
```
agua1 = sqrt(oAguaClean+1)
agua2 <- yeo.johnson(oAguaClean, lambda = lambdaAgua)

D0=ad.test(oAguaClean)
D1=ad.test(agua1)
D2=ad.test(agua2)

m0=round(c(as.numeric(summary(oAguaClean)),kurtosis(oAguaClean),skewness(oAguaClean),D0$p.value),3)
m1=round(c(as.numeric(summary(agua1)),kurtosis(agua1),skewness(agua1),D0$p.value),3)
m2=round(c(as.numeric(summary(agua2)),kurtosis(agua2),skewness(agua2),D0$p.value),3)

m<-as.data.frame(rbind(m0, m1, m2))
row.names(m)=c("Datos Limpios ORIGINAL", "Transformacion 1", "Transformacion 2")
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","Valor p")

m
```

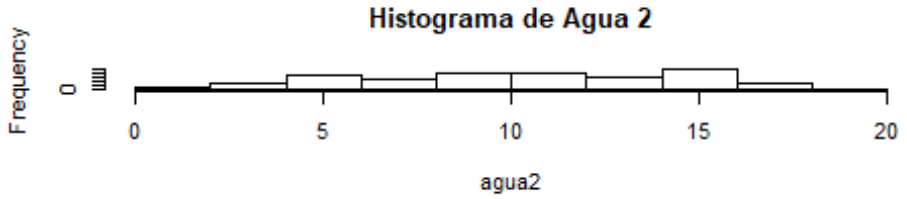
```
##                          Minimo     Q1 Mediana   Media      Q3  Máximo
Curtosis
## Datos Limpios ORIGINAL  0.043 30.700  81.400 101.280 169.775 363.600
2.447
## Transformacion 1        1.021  5.630   9.077   9.157  13.068  19.095
1.956
## Transformacion 2        0.042  6.631  10.336  10.058  14.104  19.167
2.100
##                          Sesgo Valor p
## Datos Limpios ORIGINAL  0.693       0
## Transformacion 1        0.073       0
## Transformacion 2       -0.190       0
```

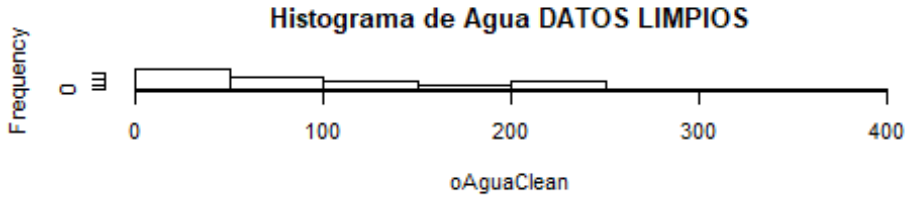
5.2 Obten el histograma de los 2 modelos obtenidos (exacto y aproximado) y de los datos originales.

```
#agua1 = sqrt(oAguaClean+1)
#agua2 <- yeo.johnson(oAgua, lambda = lambdaAgua)

par(mfrow=c(3,1))
hist(agua1,col=0,main="Histograma de Agua 1")
```

```r
hist(agua2,col=0,main="Histograma de Agua 2")
hist(oAguaClean,col=0,main="Histograma de Agua DATOS LIMPIOS")
```







5.3 Interpreta la prueba de normalidad de Anderson-Darling y Jarque Bera para los datos transformados y los originales

```r
print("----Original Datos Limpios----")
```

```
## [1] "----Original Datos Limpios----"
```

```r
jarque.test(oAgua)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  oAgua
## JB = 153.58, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
D0
```

```
##
##  Anderson-Darling normality test
##
## data:  oAguaClean
## A = 15.666, p-value < 2.2e-16
```

```r
print("----Transformacion 1----")
```

```
## [1] "----Transformacion 1----"
```

```
jarque.test(agua1)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  agua1
## JB = 24.734, p-value = 4.258e-06
## alternative hypothesis: greater
```

```
D1
```

```
##
##  Anderson-Darling normality test
##
## data:  agua1
## A = 4.9577, p-value = 2.829e-12
```

```
print("----Transformacion 2----")
```

```
## [1] "----Transformacion 2----"
```

```
jarque.test(agua2)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  agua2
## JB = 21.229, p-value = 2.456e-05
## alternative hypothesis: greater
```

```
D2
```

```
##
##  Anderson-Darling normality test
##
## data:  agua2
## A = 4.1988, p-value = 1.904e-10
```

5.4 Indica posibilidades de motivos de alejamiento de normalidad (sesgo, curtosis, datos atípicos, etc)

Los datos atipicos que estan afectando los datos son los datos 0 y los datos más alla de 1.5 rangos intercuantilicos arriba de q3 de la variable Agua

6 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.

La mejor transformacion que seria la Transformacion 1 porque es la transformacion con el Sesgo que más se acerca a 0, ademas de tener una menor Curtosis que define el numero de desviaciones estandar de la media.