Tarea 3

A01275465 Carol Arrieta Moreno

2023-08-21

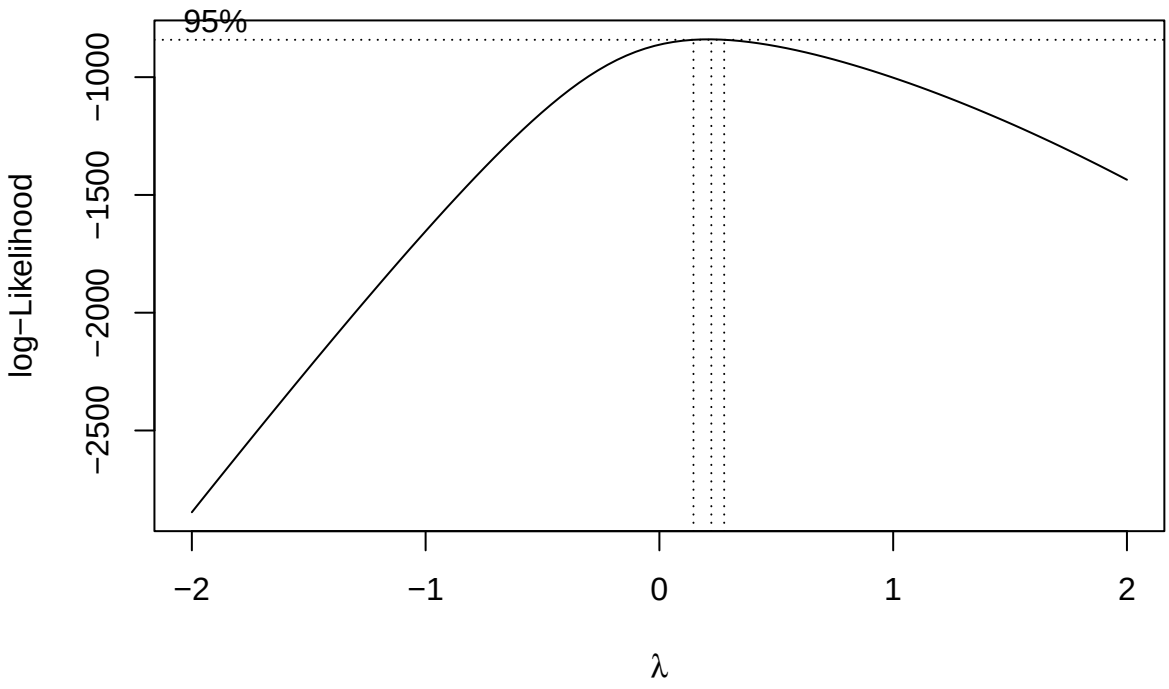
1. Utiliza la transformación Box-Cox. Utiliza el modelo exacto y el aproximado de acuerdo con las sugerencias de Box y Cox para la transformación

```
M=read.csv("mc-donalds-menu-1.csv") #leer la base de datos
```

2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos:

```
Sodio = M[13]
M = Sodio
```

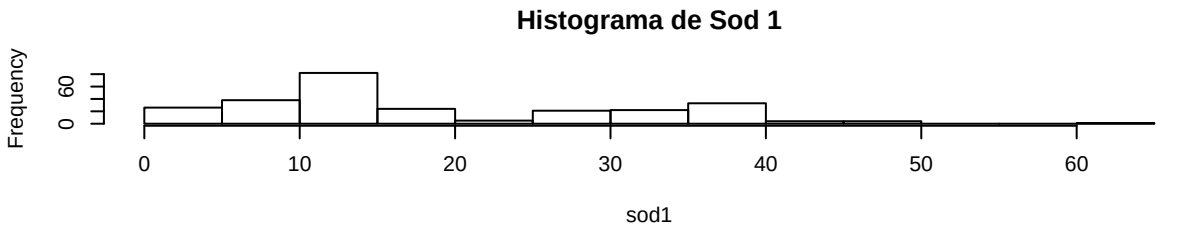
```
library(MASS)
bc<-boxcox((Sodio[ ,1]+1)~1)
```

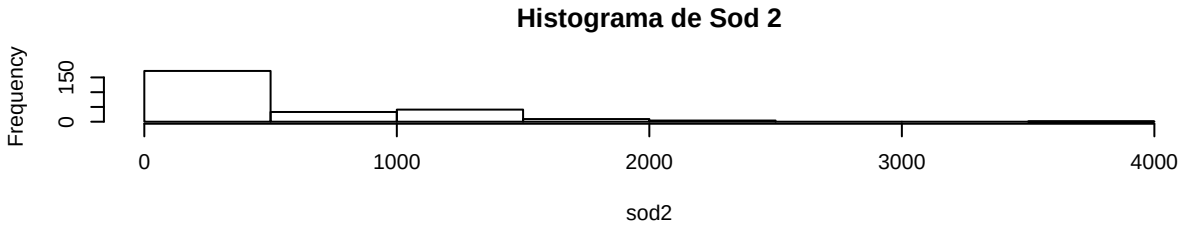


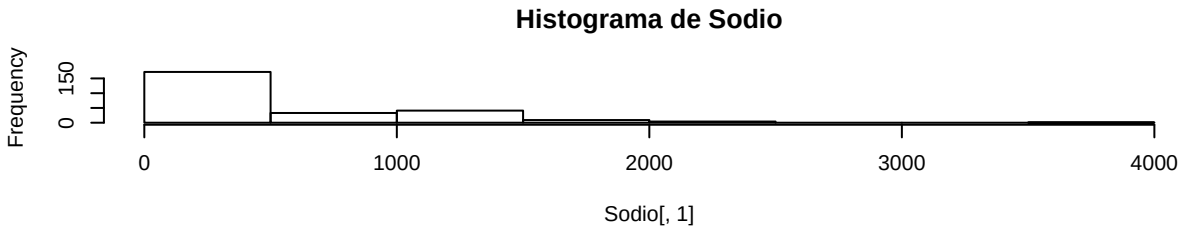

```
bc$x[which.max(bc$y)]
```

```
## [1] 0.2222222
```

```
sod1=sqrt(Sodio[ ,1]+1)
sod2=((Sodio[ ,1]+1)^1-1)/1
par(mfrow=c(3,1))
hist(sod1,col=0,main="Histograma de Sod 1")
hist(sod2,col=0,main="Histograma de Sod 2")
hist(Sodio[ ,1],col=0,main="Histograma de Sodio")
```








```
library(e1071)
summary(sod1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.41   13.82   18.60   29.43   60.01
```

```
print("Curtosis")
```

```
## [1] "Curtosis"
```

```
kurtosis(sod1)
```

```
## [1] -0.6100865
```

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```
skewness(sod1)
```

```
## [1] 0.6546664
```

```
library(e1071)
summary(sod2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   107.5   190.0   495.8   865.0  3600.0
```

```
print("Curtosis")
```

```
## [1] "Curtosis"
```

```
kurtosis(sod2)
```

```
## [1] 2.75191
```

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```
skewness(sod2)
```

```
## [1] 1.526317
```

```
library(e1071)
summary(Sodio[['Sodium']])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   107.5   190.0   495.8   865.0  3600.0
```

```
print("Curtosis")
```

```
## [1] "Curtosis"
```

```
kurtosis(Sodio[['Sodium']])
```

```
## [1] 2.75191
```

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```
skewness(Sodio[['Sodium']])
```

```
## [1] 1.526317
```

```
library(nortest)
D=ad.test(sod1)
D$p.value
```

```
## [1] 2.539143e-23
```

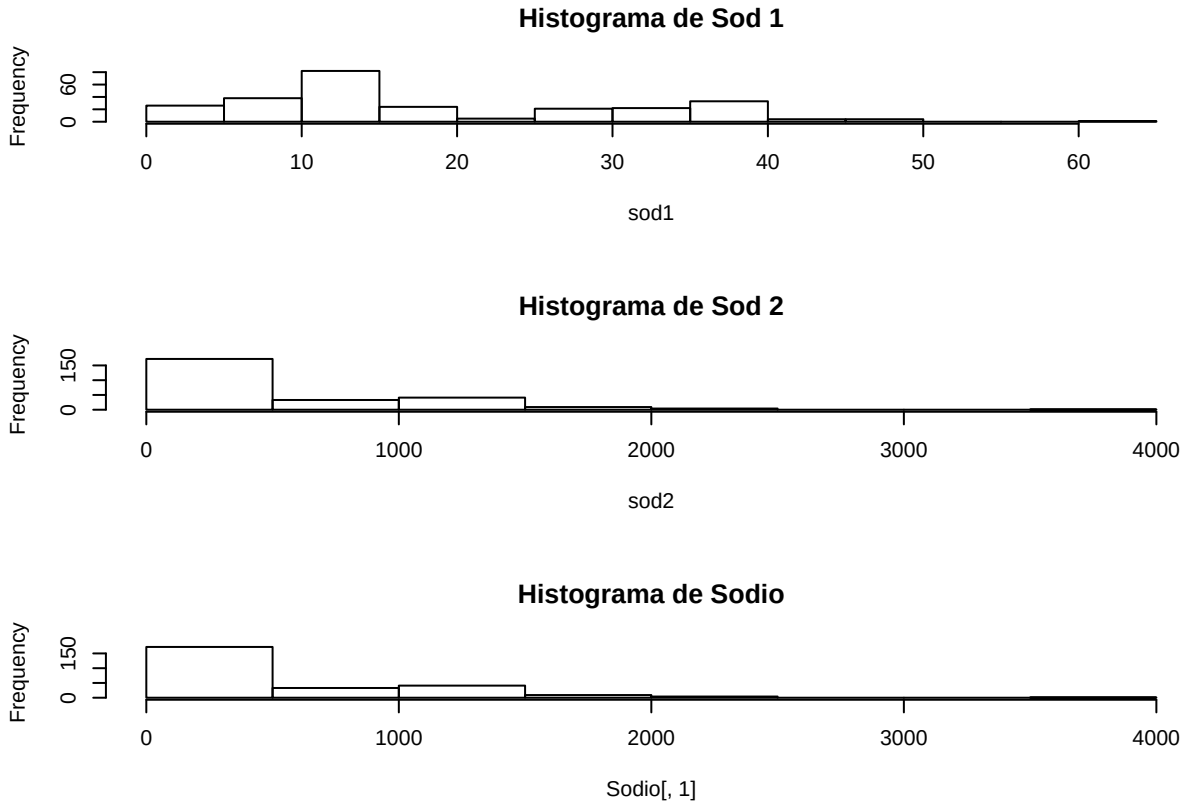
```
library(nortest)
D=ad.test(sod2)
D$p.value
```

```
## [1] 3.7e-24
```

```
library(nortest)
D=ad.test(Sodio[['Sodium']])
D$p.value
```

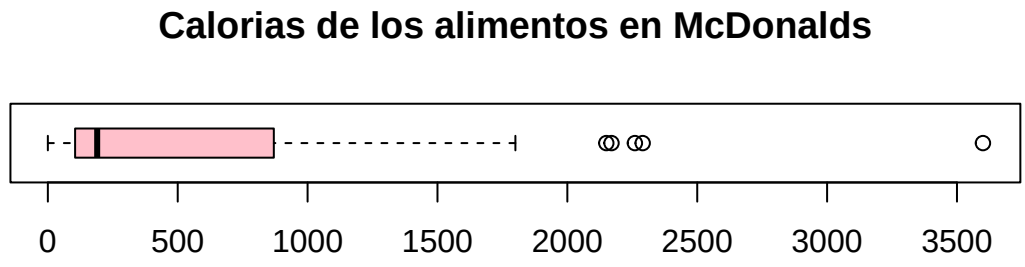
```
## [1] 3.7e-24
```

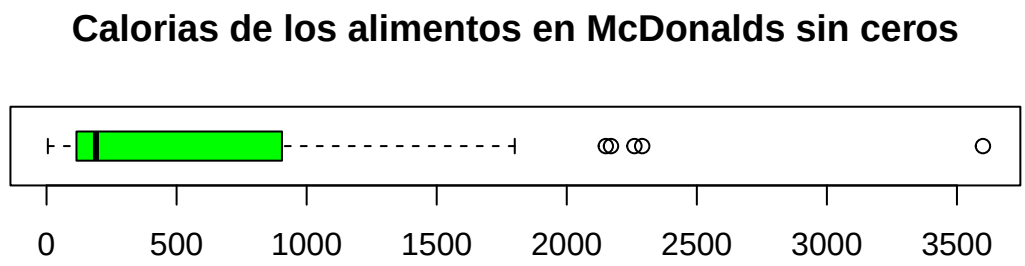
```
sod1=sqrt(Sodio[ ,1]+1)
sod2=((Sodio[ ,1]+1)^1-1)/1
par(mfrow=c(3,1))
hist(sod1,col=0,main="Histograma de Sod 1")
hist(sod2,col=0,main="Histograma de Sod 2")
hist(Sodio[ ,1],col=0,main="Histograma de Sodio")
```



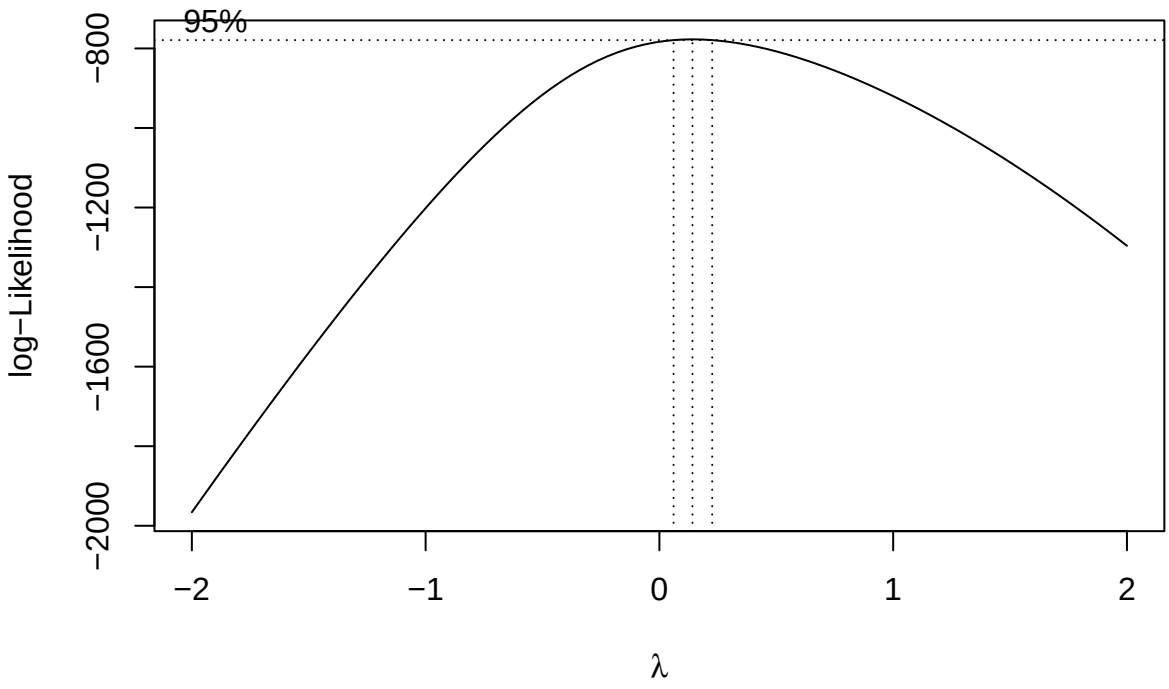
```
Sodio=subset(M,M[['Sodium']]>0)

par(mfrow=c(2,1))
boxplot(M[['Sodium']], horizontal = TRUE, col="pink", main="Calorias de los alimentos en McDonalds")
boxplot(Sodio[['Sodium']], horizontal = TRUE,col="green", main="Calorias de los alimentos en McDonalds s
```





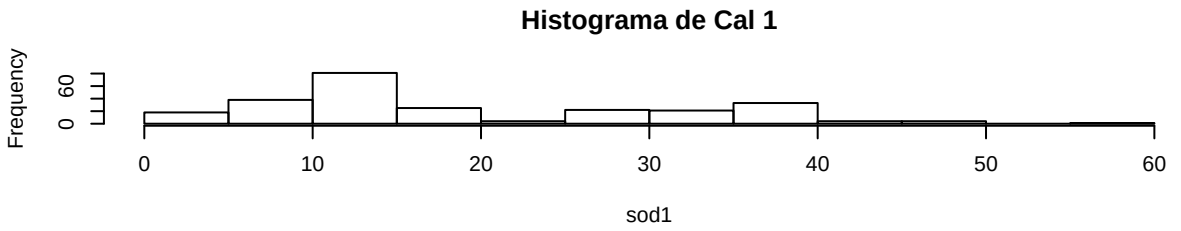

```
library(MASS)
Sodio=subset(Sodio,Sodio[['Sodium']]>0)
bc<-boxcox((Sodio[ ,1])~1)
```

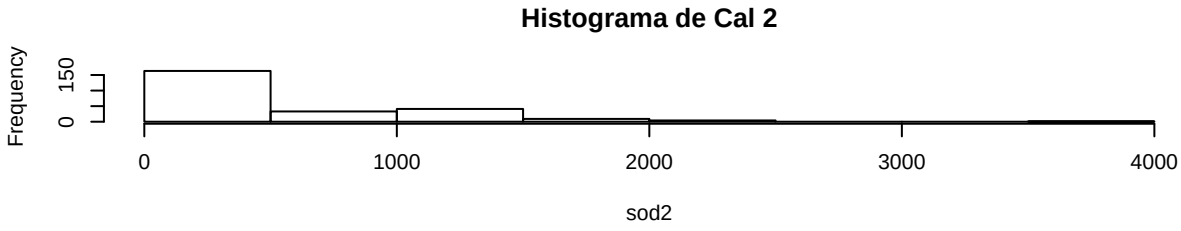


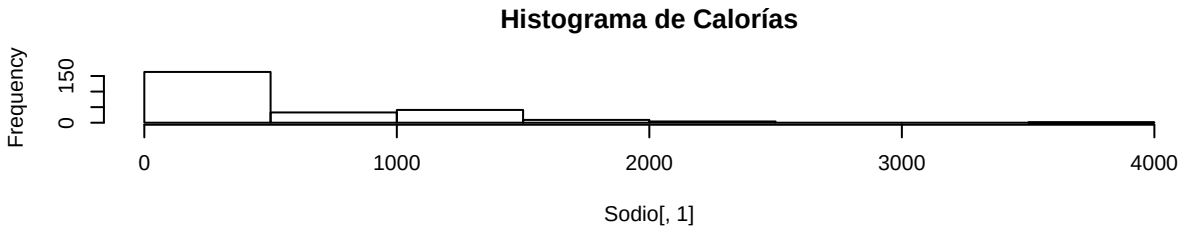

```
bc$x[which.max(bc$y)]
```

```
## [1] 0.1414141
```

```
sod1=sqrt(Sodio[ ,1])
sod2=(Sodio[ ,1]^1-1)/1
par(mfrow=c(3,1))
hist(sod1,col=0,main="Histograma de Cal 1")
hist(sod2,col=0,main="Histograma de Cal 2")
hist(Sodio[ ,1],col=0,main="Histograma de Calorías")
```








```
library(nortest)
D0=ad.test(Sodio[,1])
D1=ad.test(sod1)
D2=ad.test(sod2)
```

```
library(e1071)
m0=round(c(as.numeric(summary(Sodio[ ,1])),kurtosis(Sodio[ ,1]),skewness(Sodio[ ,1]),D0$p.value),3)
m1=round(c(as.numeric(summary(sod1)),kurtosis(sod1),skewness(sod1),D1$p.value),3)
m2=round(c(as.numeric(summary(sod2)),kurtosis(sod2),skewness(sod2),D2$p.value),3)
```

```
m<-as.data.frame(rbind(m0,m1,m2))
row.names(m)=c("Original","Primer modelo","Segundo Modelo")
names(m)=c("Minimo","Q1","Mediana","Media","Q3","Máximo","Curtosis","Sesgo","Valor p")
m
```

```
##                Minimo      Q1 Mediana   Media      Q3 Máximo Curtosis Sesgo
## Original        5.000 115.000 190.000 513.526 905.000   3600    2.651 1.492
## Primer modelo   2.236  10.724  13.784  19.184  30.083     60   -0.636 0.672
## Segundo Modelo  4.000 114.000 189.000 512.526 904.000   3599    2.651 1.492
##                Valor p
## Original             0
## Primer modelo        0
## Segundo Modelo       0
```

```
library(VGAM)
```

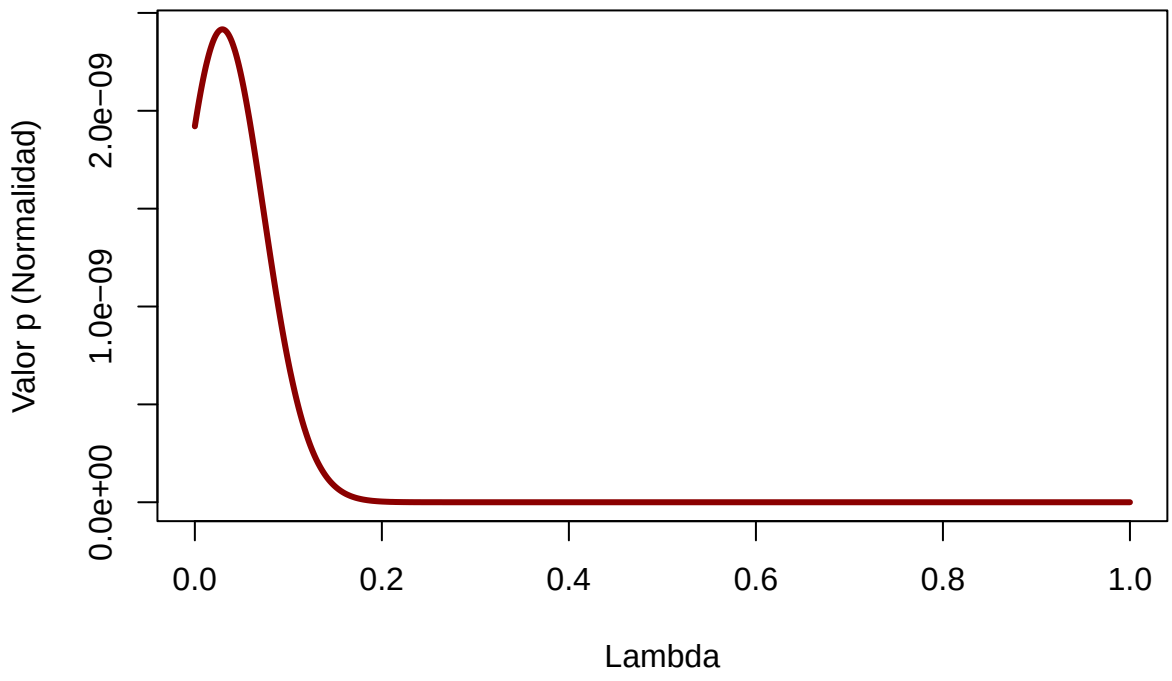
```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
sod3<- yeo.johnson(Sodio[ ,1], lambda = 1)
```

```
library(VGAM)
lp <- seq(0,1,0.001) # Valores de lambda propuestos
nlp <- length(lp)
n=length(Sodio[ ,1])
D <- matrix(as.numeric(NA),ncol=2,nrow=nlp)
d <-NA
for (i in 1:nlp){
d= yeo.johnson(Sodio[ ,1], lambda = lp[i])
p=ad.test(d)
D[i,]=c(lp[i],p$p.value)}
```

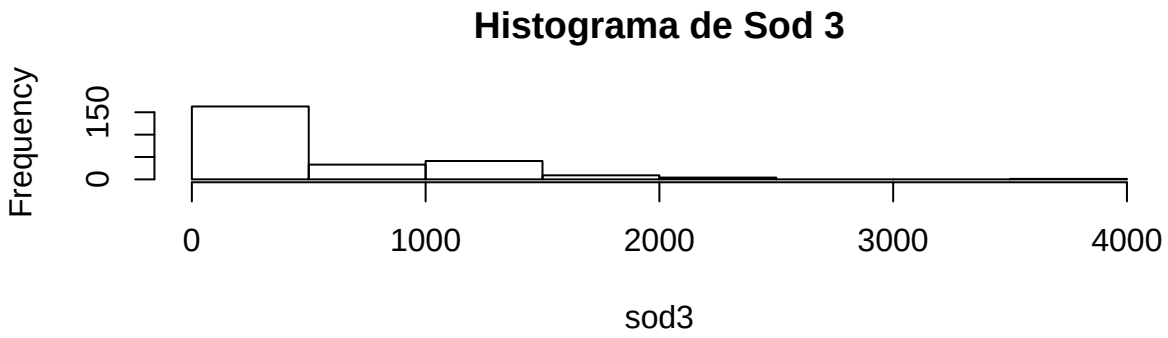
```
N=as.data.frame(D)
plot(N[['V1']],N[['V2']],
type="l",col="darkred",lwd=3,
xlab="Lambda",
ylab="Valor p (Normalidad)")
```

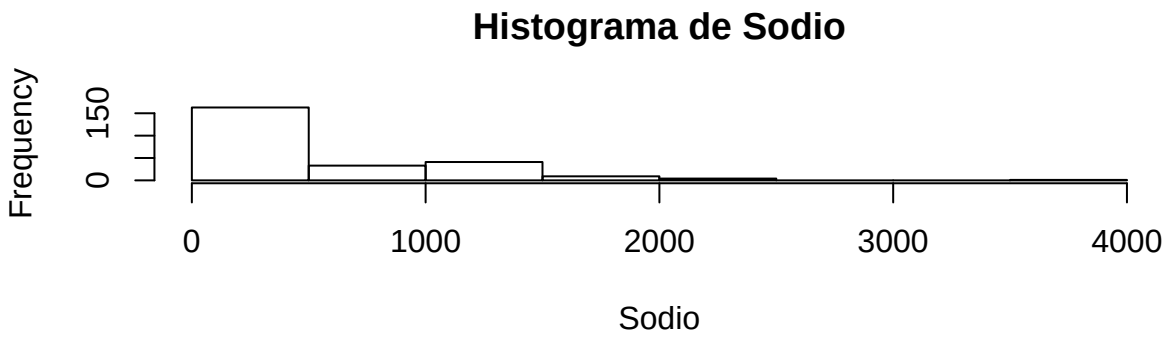


```r
G=data.frame(subset(N,N$'V2'==max(N$'V2')))
G
```

```
##       V1           V2
## 30 0.029 2.416069e-09
```

```r
par(mfrow=c(2,1))
hist(sod3,col=0,main="Histograma de Sod 3")
hist(Sodio[,1],col=0,main="Histograma de Sodio",xlab="Sodio")
```






```r
library(e1071)
summary(sod3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0   115.0   190.0   513.5   905.0  3600.0
```

```r
print("Curtosis")
```

```
## [1] "Curtosis"
```

```r
kurtosis(sod3)
```

```
## [1] 2.650539
```

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```
skewness(sod3)
```

```
## [1] 1.491921
```

```
library(e1071)
summary(Sodio[['Sodium']])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0   115.0   190.0   513.5   905.0  3600.0
```

```
print("Curtosis")
```

```
## [1] "Curtosis"
```

```
kurtosis(Sodio[['Sodium']])
```

```
## [1] 2.650539
```

```
print("Sesgo")
```

```
## [1] "Sesgo"
```

```
skewness(Sodio[['Sodium']])
```

```
## [1] 1.491921
```

Ventajas y desventajes de los modelos de Box Cox y de Yeo Johnson.

Box Cox no admite valores nulos ni negativos, mientas que Yeo Johnson si los permite. Box Cox es un modelo más simple de entender a diferencia de Yeo Johnson. Box Cox es más limitado, funciona mejor con valores mayores a 0, mientas que Yeo Johnson al utilizar una fórmula diferente es más flexible.

Diferencias entre la transformación y el escalamiento de los datos

El escalamiento de los datos es para poder facilitar la comprención de las variables, conservando la distribución de los datos. Mientras que la transofmración buscar modificar la distribución de los datos, por lo que se puede llegar a un comportamiento completamente diferente al original.