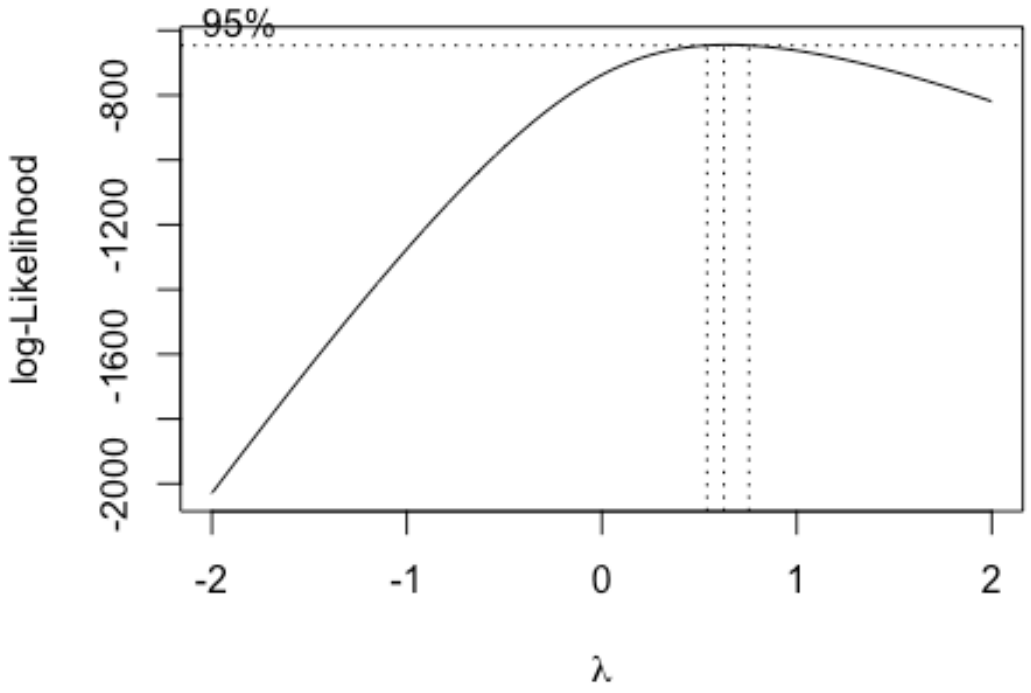
5. Transformaciones

Ricardo Salinas

2024-08-14

Selecciona una variable, que no sea Calorías, y encuentra la mejor transformación de datos posible para que la variable seleccionada se comporte como una distribución Normal. Realiza:

```
M = read.csv("mcdonalds.csv") #Leer la base de datos
M1 = M$Carbohydrates

library(MASS)

#Box Cox
bc <- boxcox((M1 + 1) ~ 1)
```



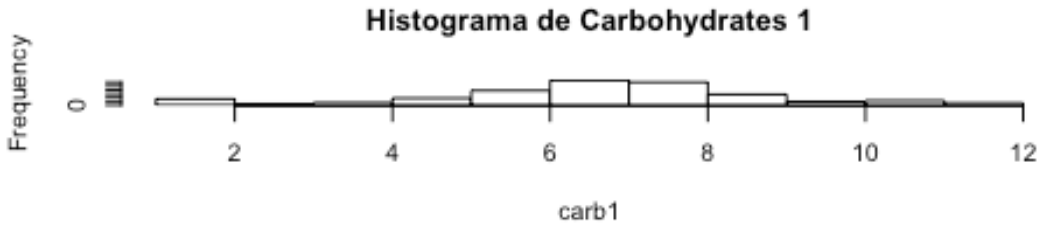

```
l=bc$x[which.max(bc$y)]
print("El valor optimo de lambda es:")
```

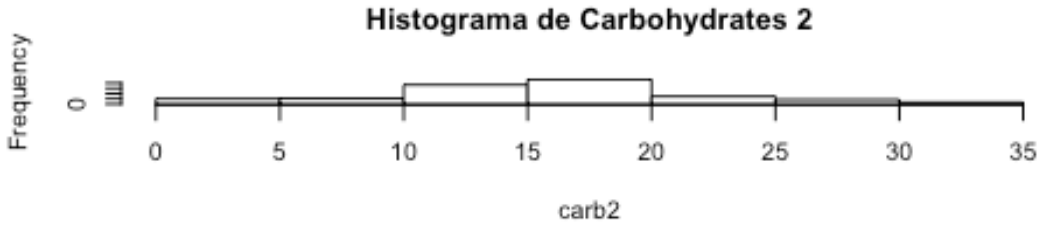
```
## [1] "El valor optimo de lambda es:"
```

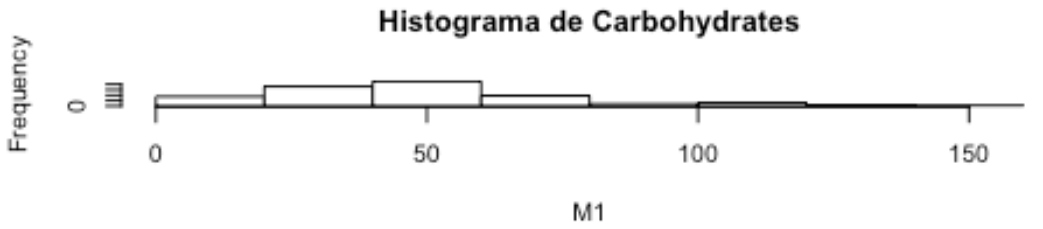
```
print(l)
```

```
## [1] 0.6262626
```

```
#Histogramas
carb1=sqrt((M1 + 1) + 1)
carb2=((M1+1)^l-1)/l
par(mfrow=c(3,1))
hist(carb1,col=0,main="Histograma de Carbohydrates 1")
hist(carb2,col=0,main="Histograma de Carbohydrates 2")
hist(M1,col=0,main="Histograma de Carbohydrates")
```








```
library(e1071)
summary(M1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   30.00   44.00   47.35   60.00  141.00
```

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

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

```
kurtosis(M1)
```

```
## [1] 1.324083
```

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

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

```
skewness(M1)
```

```
## [1] 0.9021952
```

```
library(nortest)
D=ad.test(M1)
D$p.value
```

```
## [1] 2.546548e-10
```

Escribe las ecuaciones de los modelos encontrados.

x1 = $\sqrt{x+1}$

x2 = $\frac{x^{.63}-1}{.63}$

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

```
library(e1071)


summary(carb1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.414   5.657   6.782   6.681   7.874  11.958
```

```
print("Curtosis Carb1")
```

```
## [1] "Curtosis Carb1"
```

```
kurtosis(carb1)
```

```
## [1] 0.7685457
```

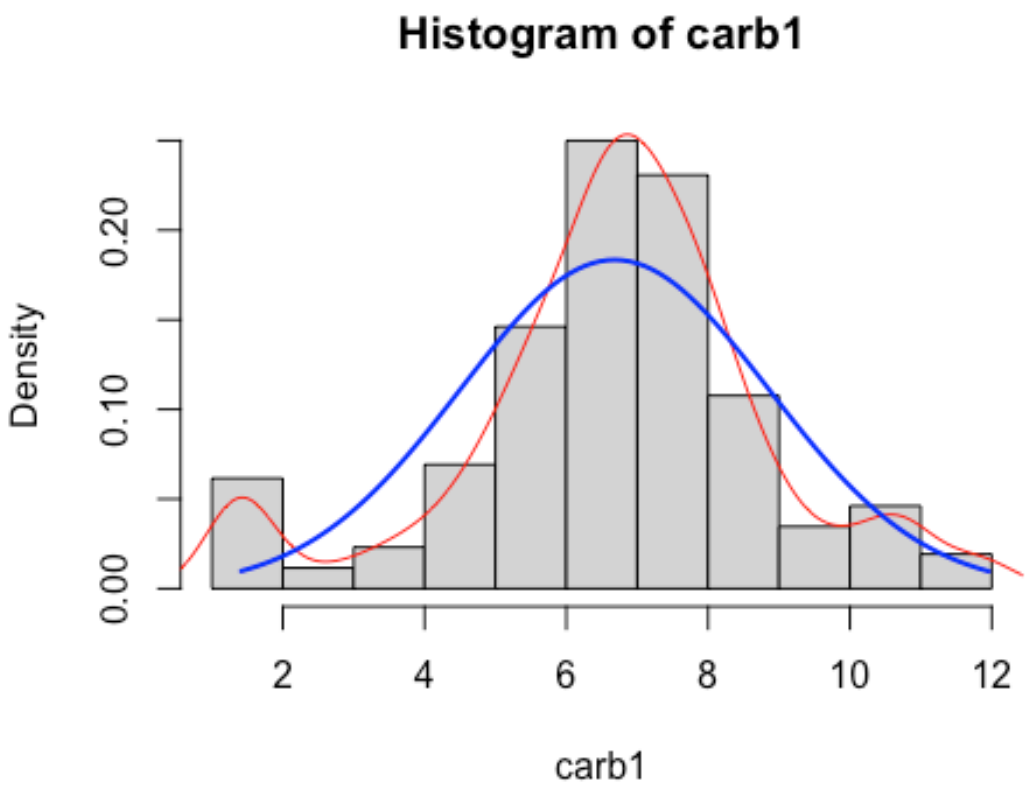
```
print("Sesgo Carb1")
```

```
## [1] "Sesgo Carb1"
```

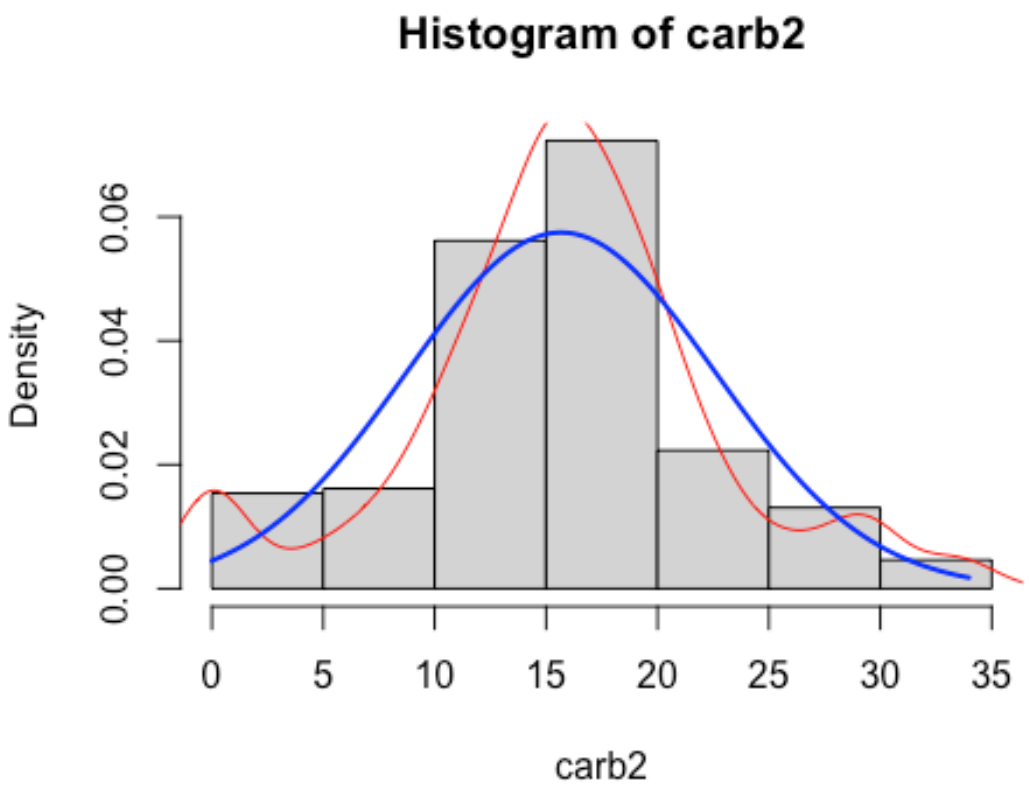
```
skewness(carb1)
```

```
## [1] -0.3931435
```

```
hist(carb1,freq=FALSE)
lines(density(carb1),col="red")
curve(dnorm(x, mean = mean(carb1), sd = sd(carb1)), from = 1.414, to = 11.958
, add = TRUE, col = "blue", lwd = 2)
```




```
hist(carb2,freq=FALSE)
lines(density(carb2),col="red")
curve(dnorm(x, mean = mean(carb2), sd = sd(carb2)), from = 0, to = 33.98, add
= TRUE, col = "blue", lwd = 2)
```




```
summary(carb2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.12   15.72   15.67   19.36   33.98
```

```
print("Curtosis Carb2")
```

```
## [1] "Curtosis Carb2"
```

```
kurtosis(carb2)
```

```
## [1] 0.6381974
```

```
print("Sesgo Carb2")
```

```
## [1] "Sesgo Carb2"
```

```
skewness(carb2)
```

```
## [1] -0.08250202
```

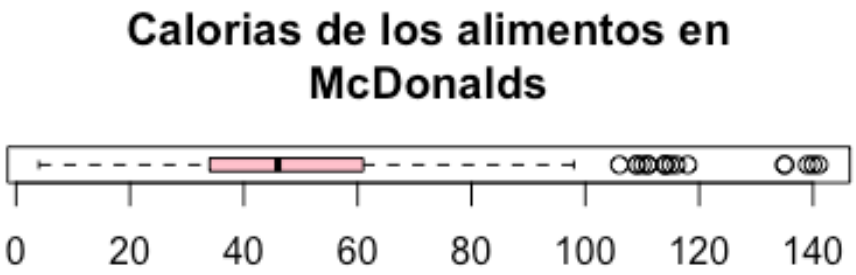
```
library(nortest)
library(moments)
```

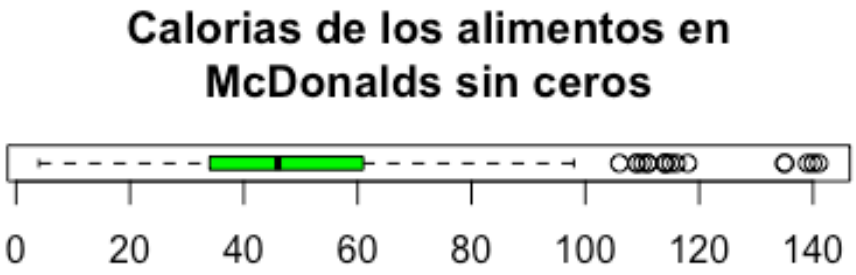
```
##
## Attaching package: 'moments'

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

ad.test(M1)

##
##  Anderson-Darling normality test
##
## data:  M1
## A = 4.1402, p-value = 2.547e-10

jarque.test(M1)

##
##  Jarque-Bera Normality Test
##
## data:  M1
## JB = 55.646, p-value = 8.251e-13
## alternative hypothesis: greater

ad.test(carb1)

##
##  Anderson-Darling normality test
##
## data:  carb1
## A = 3.9283, p-value = 8.301e-10

jarque.test(carb1)

##
##  Jarque-Bera Normality Test
##
## data:  carb1
## JB = 13.669, p-value = 0.001076
## alternative hypothesis: greater

ad.test(carb2)

##
##  Anderson-Darling normality test
##
## data:  carb2
## A = 3.1076, p-value = 8.182e-08

jarque.test(carb2)

##
##  Jarque-Bera Normality Test
```

```
##
## data:  carb2
## JB = 5.1086, p-value = 0.07775
## alternative hypothesis: greater
```

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

```
M2=subset(M1,M1>0)

par(mfrow=c(2,1))
boxplot(M2, horizontal = TRUE,col="pink", main="Calorias de los alimentos en
McDonalds")
boxplot(M2, horizontal = TRUE,col="green", main="Calorias de los alimentos en
McDonalds sin ceros")
```

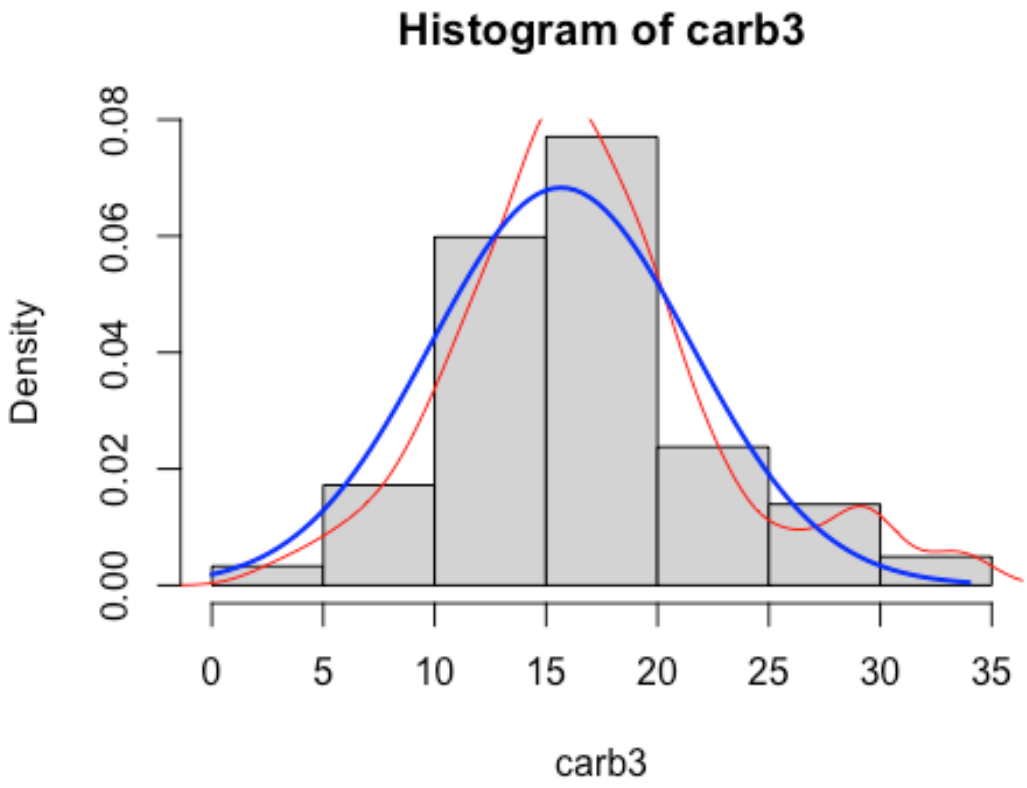




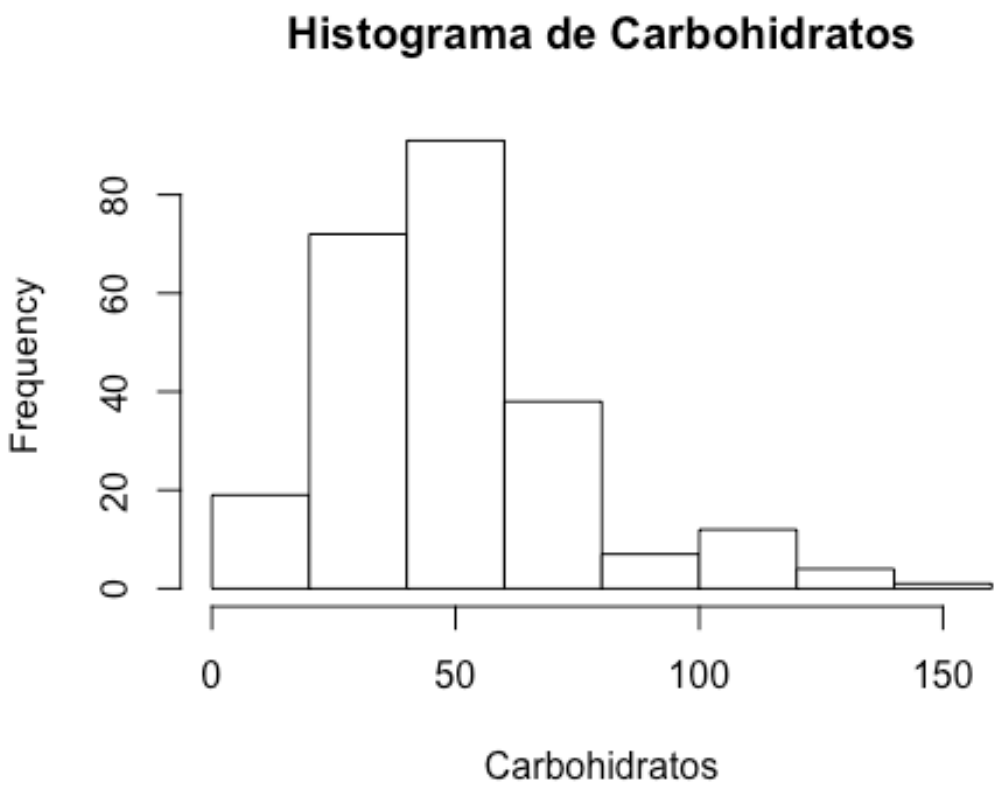
Utiliza la transformación de Yeo Johnson y encuentra el valor de lambda que maximiza el valor p de la prueba de normalidad que hayas utilizado (Anderson-Darling o Jarque Bera).

```
library(VGAM)

## Loading required package: stats4

## Loading required package: splines
```

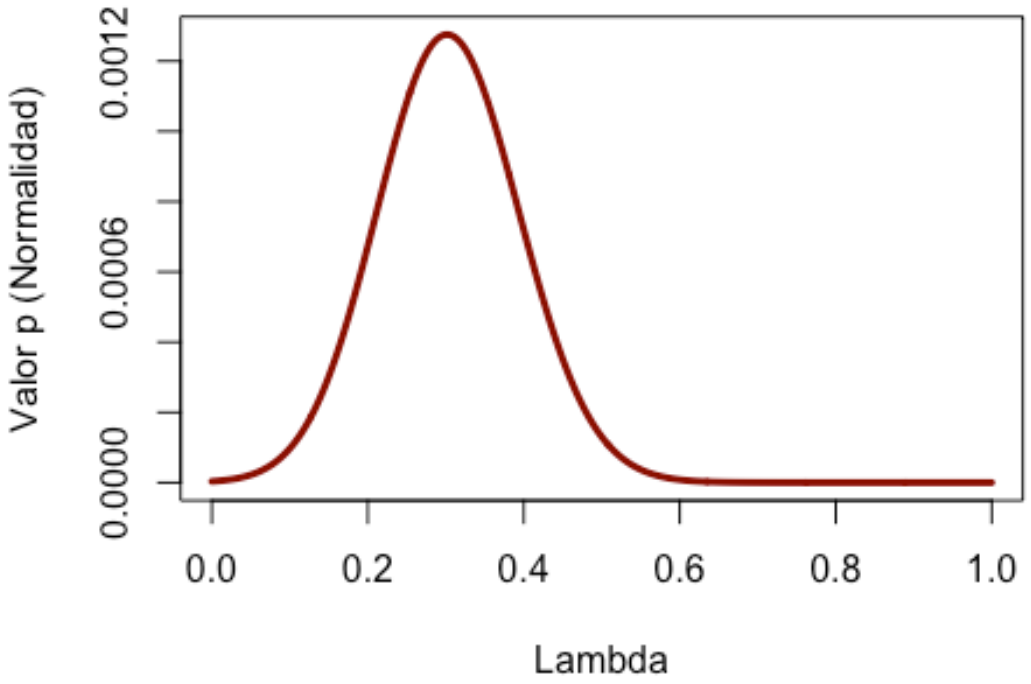
```r
carb3<- yeo.johnson(M2, lambda = 1)

hist(carb3,freq=FALSE)
lines(density(carb3),col="red")
curve(dnorm(x, mean = mean(carb2), sd = sd(carb3)), from = 0, to = 33.98, add
= TRUE, col = "blue", lwd = 2)
```




```r
hist(M2,col=0,main="Histograma de Carbohidratos",xlab="Carbohidratos")
```




```r
library(VGAM)
lp <- seq(0,1,0.001) # Valores de lambda propuestos
nlp <- length(lp)
n=length(M2)
D <- matrix(as.numeric(NA),ncol=2,nrow=nlp)
d <-NA
for (i in 1:nlp){
d= yeo.johnson(M2, lambda = lp[i])
p=ad.test(d)
D[i,]=c(lp[i],p$p.value)}

N=as.data.frame(D)
colnames(N) = c("Lambda", "Valor-p")
plot(N$Lambda,N$`Valor-p`, type="l",
col="darkred", lwd=3,
xlab="Lambda",
ylab="Valor p (Normalidad)")
```




```
G=data.frame(subset(N,N$`Valor-p`==max(N$`Valor-p`)))
print(G)

##     Lambda     Valor.p
## 303  0.302 0.001275547
```

Escribe la ecuación del modelo encontrado.

x1 = $\sqrt{x+1}$

x2 = $\frac{x^{.302}-1}{.302}$

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

```
summary(carb3)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.778  13.202  16.203  16.696  19.575  33.978

print("Curtosis Carb3")

## [1] "Curtosis Carb3"
```

```r
kurtosis(carb3)
```

```
## [1] 3.763806
```

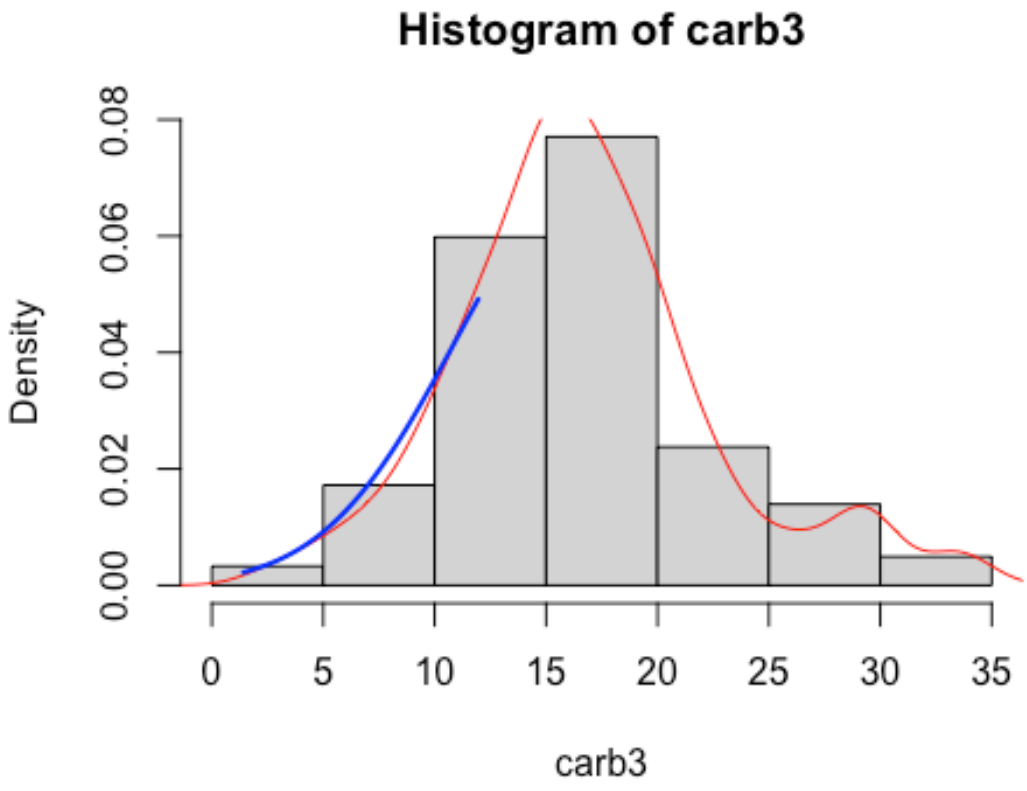
```r
print("Sesgo Carb3")
```

```
## [1] "Sesgo Carb3"
```

```r
skewness(carb3)
```

```
## [1] 0.5912871
```

```r
hist(carb3,freq=FALSE)
lines(density(carb3),col="red")
curve(dnorm(x, mean = mean(carb3), sd = sd(carb3)), from = 1.414, to = 11.958
, add = TRUE, col = "blue", lwd = 2)
```




```r
library(nortest)
ad.test(carb3)
```

```
##
##  Anderson-Darling normality test
##
## data:  carb3
## A = 2.4495, p-value = 3.3e-06
```

```
jarque.test(carb3)

##
##  Jarque-Bera Normality Test
##
## data:  carb3
## JB = 20.149, p-value = 4.214e-05
## alternative hypothesis: greater
```

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.

Box cox aunque no acepta datos en 0, es en mi opinion la mas util gracias a su facilidad y a que sus formulas matematicas son facilmente representadas.

Concluye sobre las ventajas y desventajes de los modelos de Box Cox y de Yeo Johnson.

Box Cox es muy facil de usar y permite flexibilidad con las variables, tambien se representa por formulas simples de entender, pero una desventaja que no te en este trabajo fue que no acepta valores en 0, lo cual puede llegar a afectar otras estmiaciones que no consideren los valores en 0 como inconsistencias o innecesarios. Yeo Johnson si permite datos en 0, lo cual permite un uso conveniente, pero como desventaja es que cuenta con una formula mas extensa, en si estas formulas son un tanto opuestas, y depende de los datos que tenga el usuario para poder decidir cual es la mas conveniente, dependiendo principalmente de si tiene datos en 0 que sean de importancia.

Analiza las diferencias entre la transformación y el escalamiento de los datos:

Escribe al menos 3 diferencias entre lo que es la transformación y el escalamiento de los datos

Estos dos procesos son muy diferentes aunque ambos llegan a manipular los datos en su propio modo, la principal diferencia entre estas dos es que los valores no se ven modificados en el escalamiento, solo cambia su rango, en cambio con la transformacion las variables cambian para poder pertenecer a un modelo, otra diferencia principal es que el escalamiento simplifica el analisis de los datos, mientras la transformacion puede llegar a dificultarlo, y l ultima diferencia es que dependiendo de los metodos usados, la transformacion puede llegar a no ser reversible, mientras que el escalamiento es facilmente reversible.

Indica cuándo es necesario utilizar cada uno

El escalamiento es mucho mas amistoso ya que puede ser muy util para normalizar los datos y darles una misma importancia a estos, mientras la transformacion puede ser usada principalmente cuando necesitas modificar datos para que se ajusten a un modelo.