Actividad 5. Transformaciones

Oscar Gutierrez

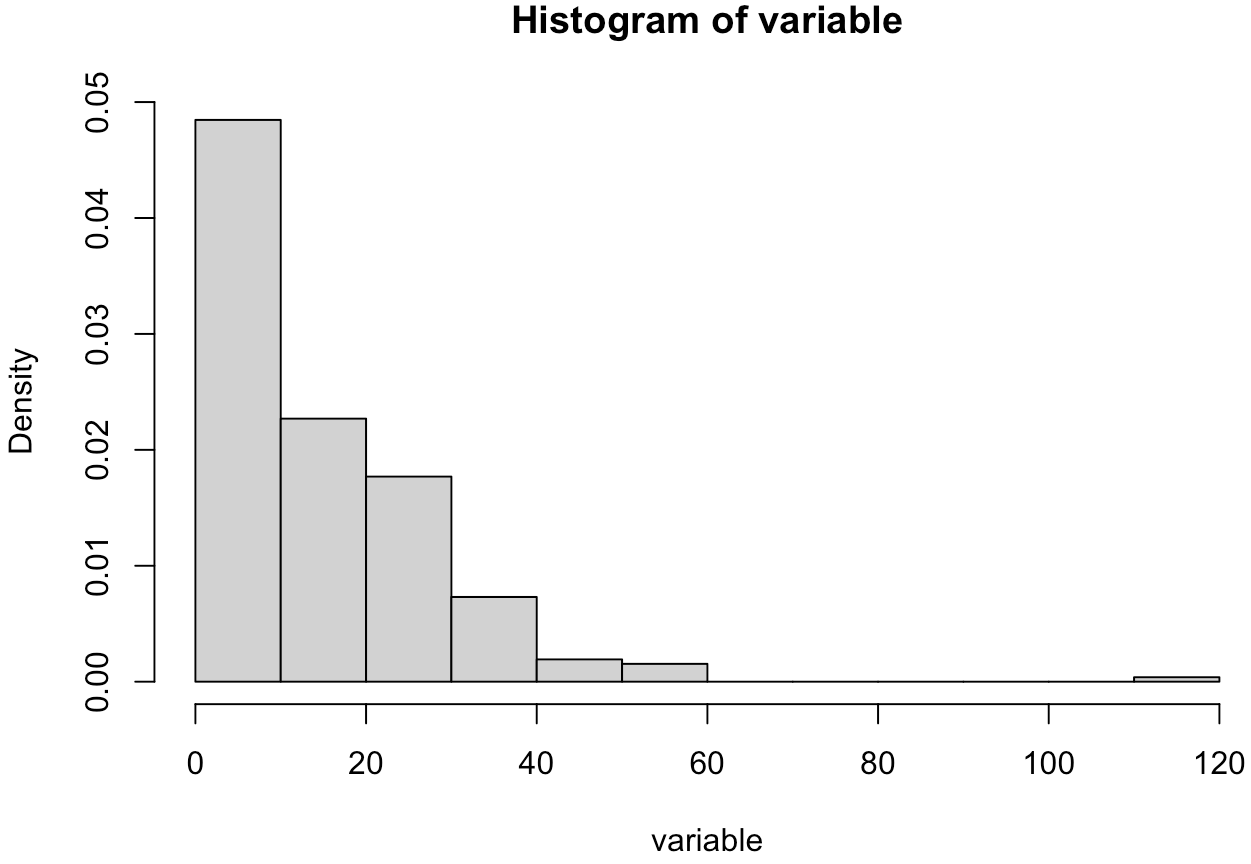
2024-08-14

Cargar dataset y seleccionar variable

```
M= read.csv("mc-donalds-menu.csv")
```

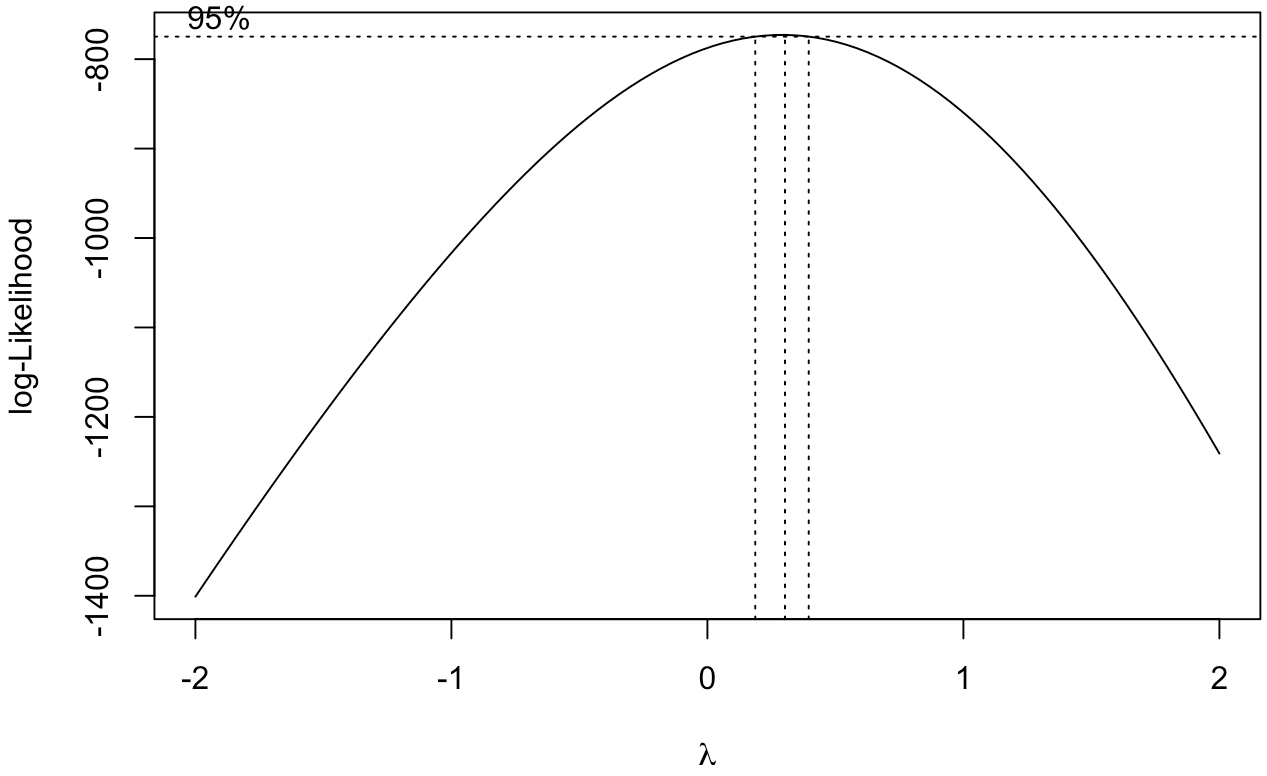
Seleccionar la variable

```
variable = M$Total.Fat
hist(variable,freq=FALSE)
```



Box-Cox

```
library(MASS)
bc<-boxcox((variable+1)~1)
```




```
l=bc$x[which.max(bc$y)]
cat('Lambda=',l)
```

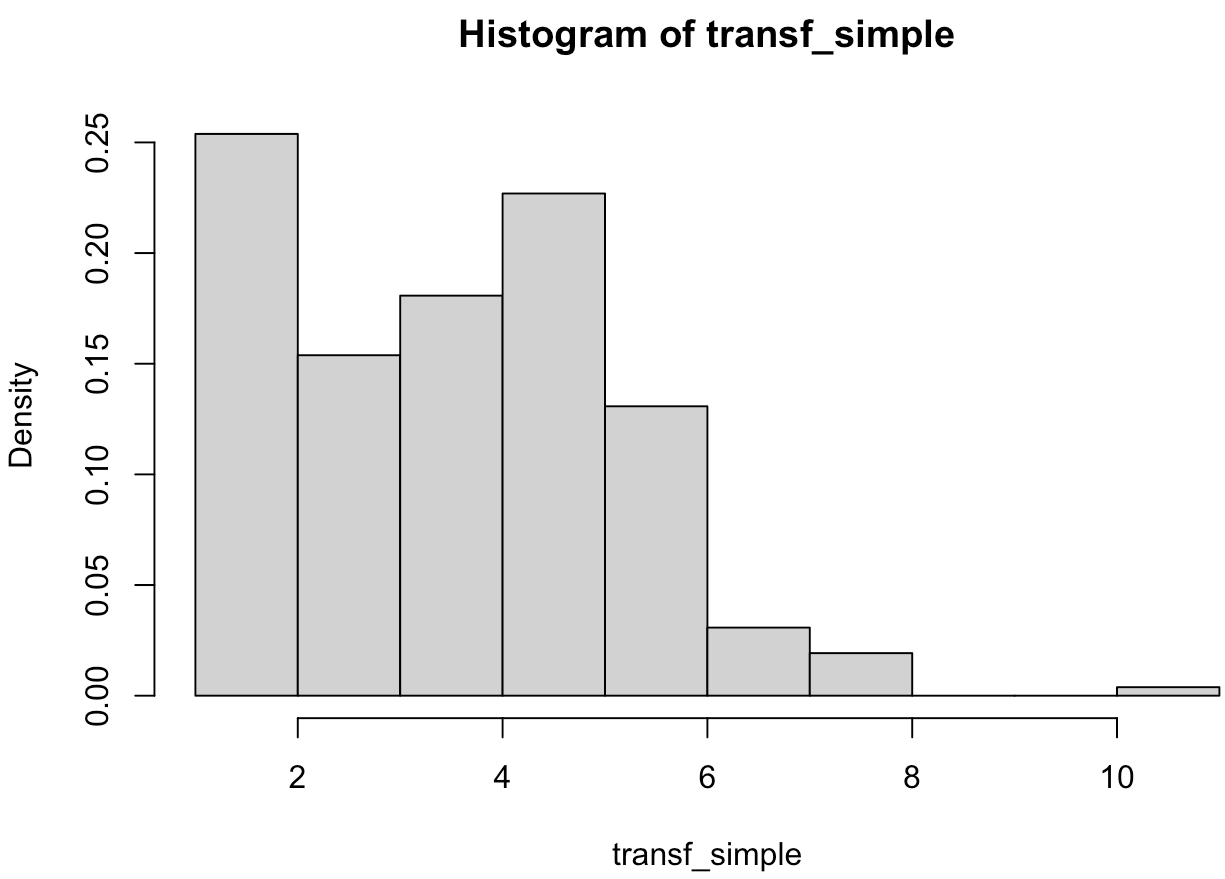
```
## Lambda= 0.3030303
```

Histogramas con transformaciones

Transformacion simple

La ecuacion para la transformacion simple es $\sqrt{x+1}$

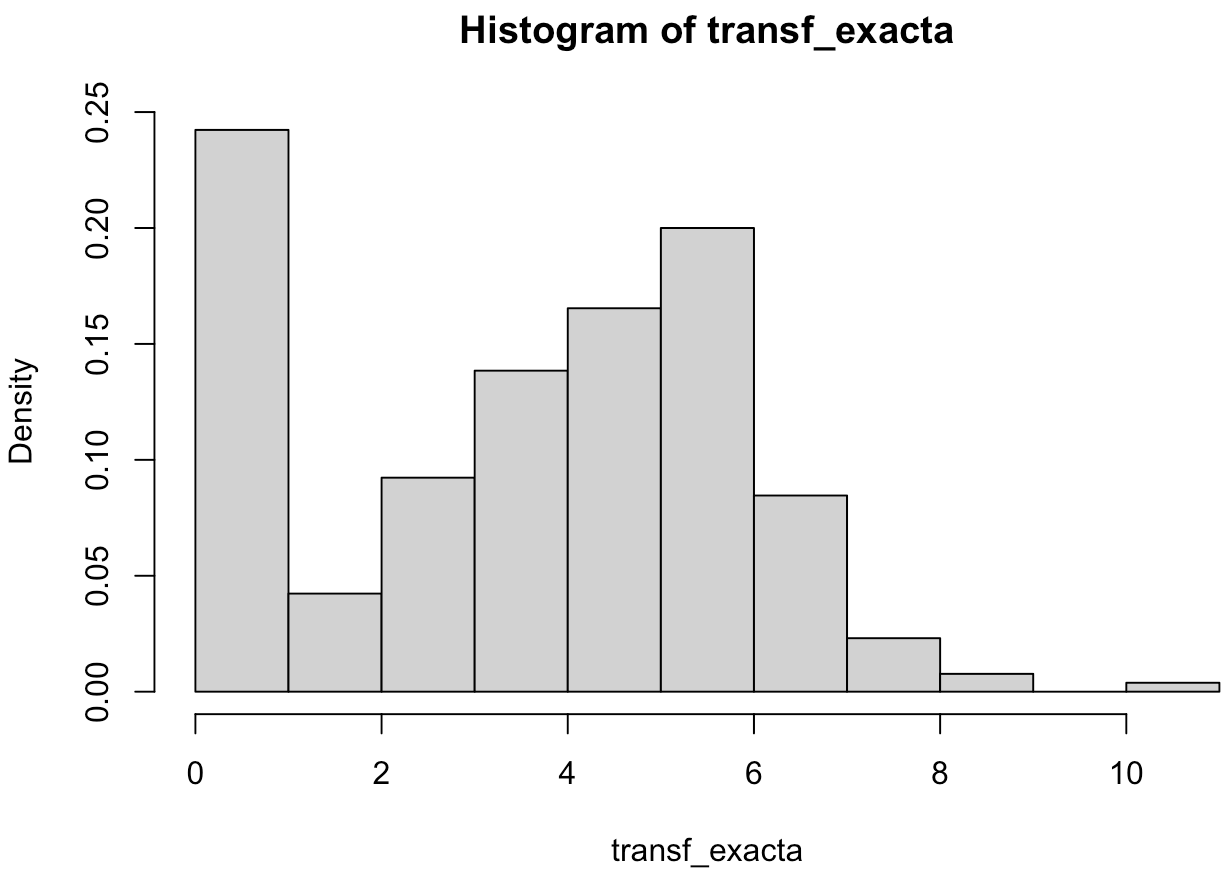
```
transf_simple = sqrt(variable +1 )
hist(transf_simple,freq=FALSE)
```



Transformacion exacta

La ecuacion para la transformacion exacta es $\frac{x^{\lambda}+1}{\lambda}$

```
transf_exacta = ((variable +1 )^l - 1)/l
hist(transf_exacta,freq=FALSE)
```



###

Resultados

```
library(nortest)
library(moments)
D0=ad.test(variable)
D1=ad.test(transf_simple)
D2=ad.test(transf_exacta)

m0=round(c(as.numeric(summary(variable)),kurtosis(variable),skewness(variable),D0$p.valu
e),3)
m1=round(c(as.numeric(summary(transf_simple)),kurtosis(transf_simple),skewness(transf_si
mple),D1$p.value),3)
m2=round(c(as.numeric(summary(transf_exacta)),kurtosis(transf_exacta),skewness(transf_ex
acta),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            0 2.375  11.000 14.165 22.250 118.000   13.455  2.140
## Primer modelo       1 1.836   3.464  3.450  4.822  10.909    2.942  0.310
## Segundo Modelo      0 1.469   3.707  3.433  5.262  10.743    2.165 -0.116
##                Valor p
## Original             0
## Primer modelo        0
## Segundo Modelo       0
```

Valores atípicos

Se remueven los 49 objetos sin grasas, es una ensalada y el resto son bebidas.

```
variable2 = variable[variable > 0]
```

Transformacion Yeo-Johnson

```
library(VGAM)
```

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

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

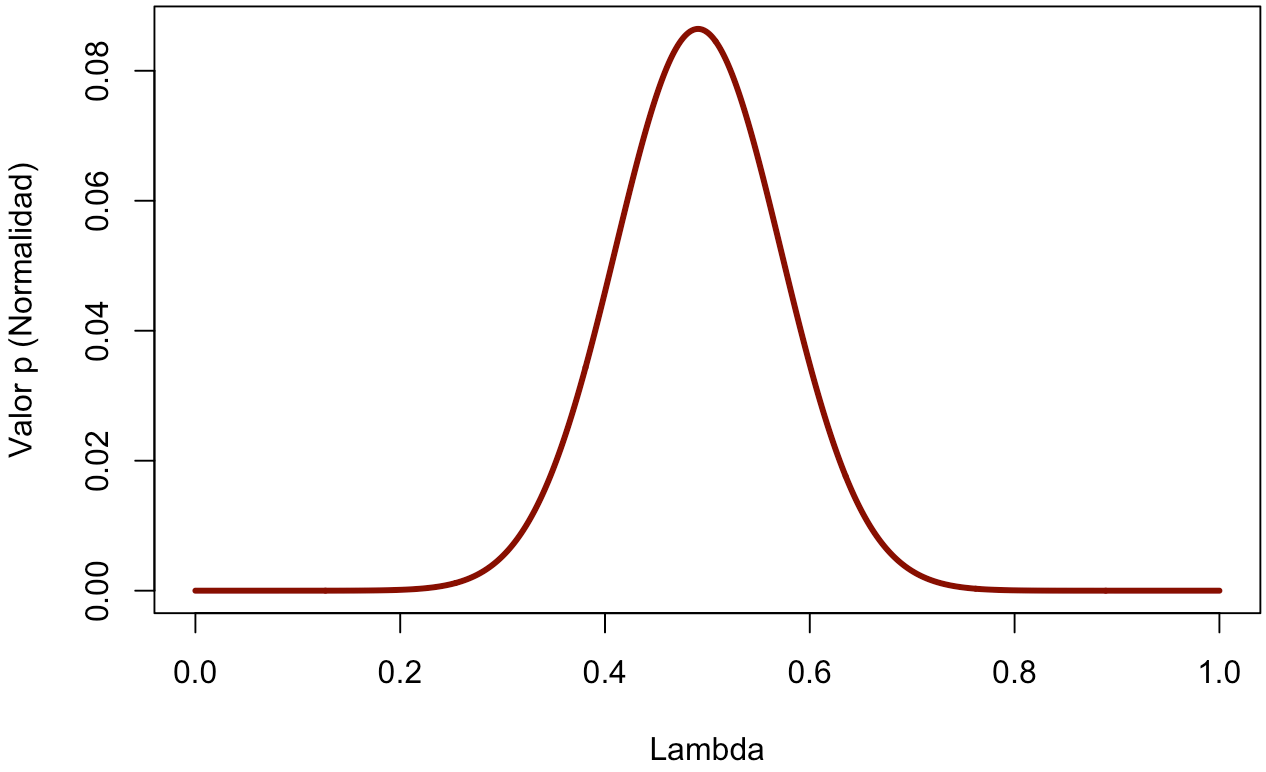
```
lp <- seq(0,1,0.001) # Valores de lambda propuestos
nlp <- length(lp)
n=length(variable2)
D <- matrix(as.numeric(NA), ncol=2, nrow=nlp)
d <- NA

for (i in 1:nlp) {
  d = yeo.johnson(variable2, lambda = lp[i])
  p = ad.test(d)
  D[i,] = c(lp[i], p$p.value)
}

# Convert matrix to data frame and name the columns
N <- as.data.frame(D)
colnames(N) <- c("Lambda", "Valor-p")

# Remove any rows with NA or infinite values
N <- N[is.finite(N$`Lambda`) & is.finite(N$`Valor-p`), ]

# Now, plot the data
plot(N$Lambda, N$`Valor-p`, type="l", col="darkred", lwd=3, xlab="Lambda", ylab="Valor p
(Normalidad)")
```

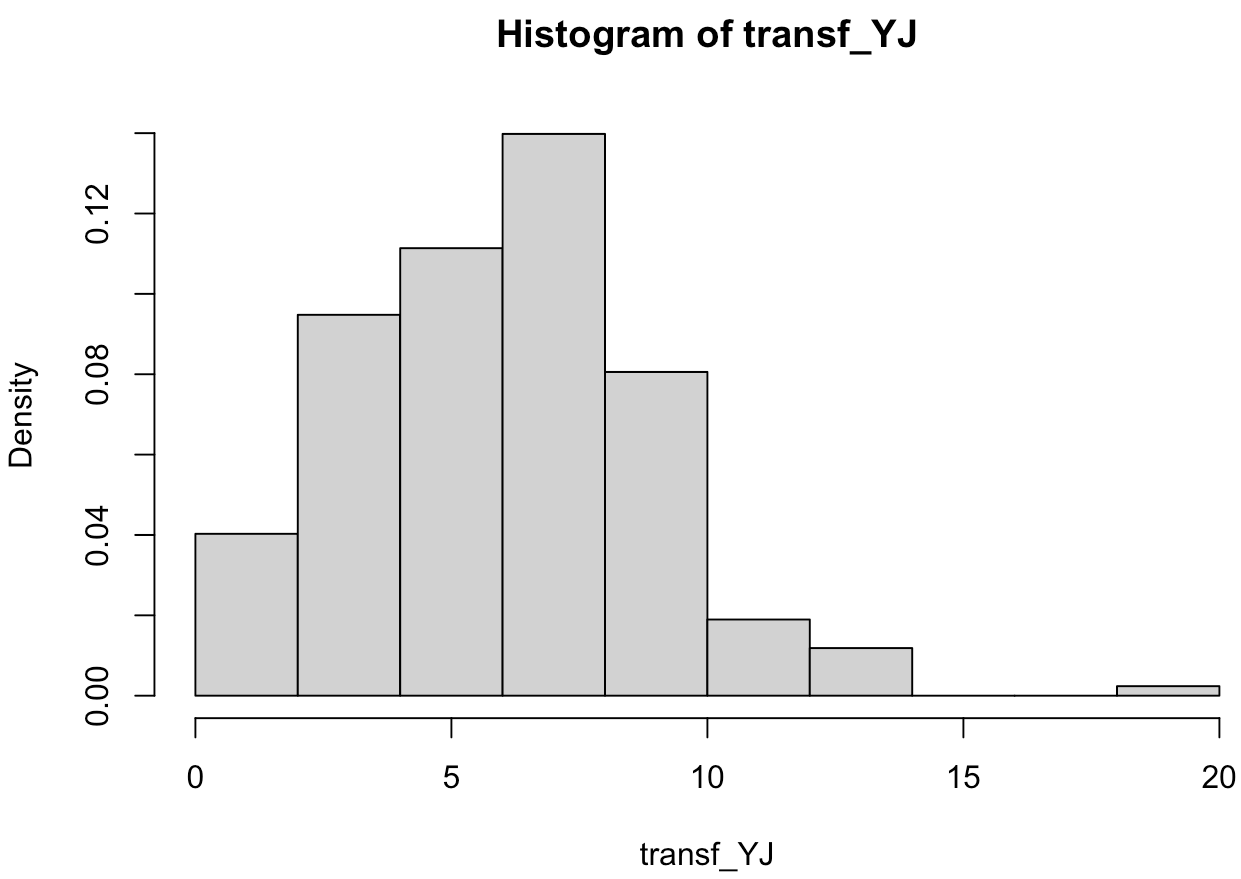



```
G=data.frame(subset(N,N$`Valor-p`==max(N$`Valor-p`)))
G
```

```
##     Lambda    Valor.p
## 492  0.491 0.08644269
```

Histograma

```
l_YJ = 0.491
transf_YJ = ((variable2 +1 )^l_YJ - 1)/l_YJ
hist(transf_YJ,freq=FALSE)
```



La ecuación encontrada es $\frac{x^{0.302}+1}{0.302}$.

```
D0=ad.test(variable)
D1=ad.test(transf_simple)
D2=ad.test(transf_YJ)

m0=round(c(as.numeric(summary(variable)),kurtosis(variable),skewness(variable),D0$p.valu
e),3)
m1=round(c(as.numeric(summary(transf_simple)),kurtosis(transf_simple),skewness(transf_si
mple),D1$p.value),3)
m2=round(c(as.numeric(summary(transf_YJ)),kurtosis(transf_YJ),skewness(transf_YJ),D2$p.v
alue),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        0.000 2.375  11.000 14.165 22.250 118.000   13.455  2.14
## Primer modelo   1.000 1.836   3.464  3.450  4.822  10.909    2.942  0.31
## Segundo Modelo  0.449 3.954   6.149  5.937  7.758  19.245    4.099  0.42
##                Valor p
## Original         0.000
## Primer modelo    0.000
## Segundo Modelo   0.086
```

El sesgo y la curtosis de la transformacion de Yeo-Johnson es mejor que las transformaciones anteriores, además de que los valores de media y mediana están muy cerca uno de otro, lo cual se busca para asumir normalidad, el valor p obtenido es apenas suficiente para aceptar H_0.

Ventajas y desventajas de Box Cox y Yeo Johnson

Considerando H_0 : La variable se distribuye normalmente y
H_1 : La variable no se distribuye normalmente Utilizando la transformación de Yeo-Johnson se logró obtener un valor de p suficientemente grande para considerar normalidad.

La ventaja que tiene Yeo Johnson sobre Box Cox es que puede ser aplicada a numeros negativos y positivos, mientras que Box Cox solo funciona con numeros positivos.

Diferencias entre transformación y escalamiento

1. La transformación afecta la distribución de los datos a diferencia del escalamiento, esto quiere decir que puede cambiar la media, varianza y mediana.
2. Las tranformaciones son útiles cuando se requiere tener de una distribución en específico, como para el caso de regresiones lineales.
3. Un ejemplo en el que se puede utilizar el esacalamiento es cuando se desea estandarizar una distribución, como es el caso de la Z~N(0,1), una transformación no lograría esto.