

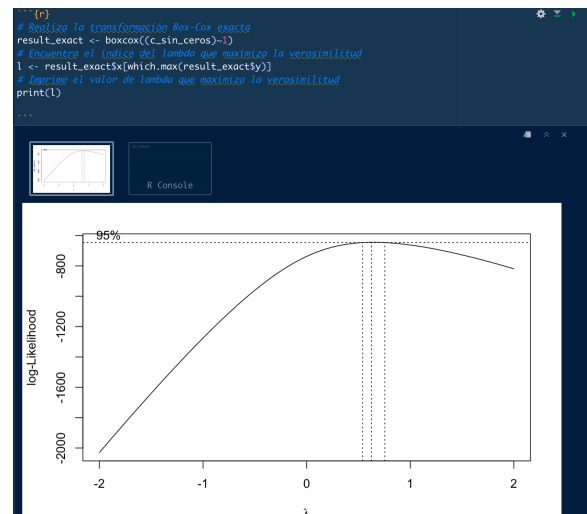
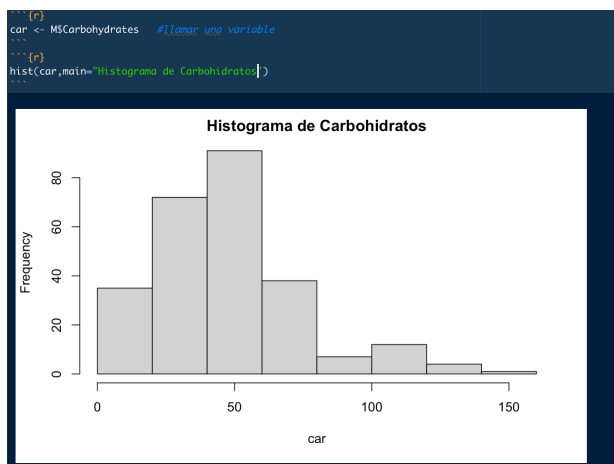
Jorge Javier Sosa Briseño

A01749489

August 21, 2023

Transformaciones.

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.



2.- Ecuaciones

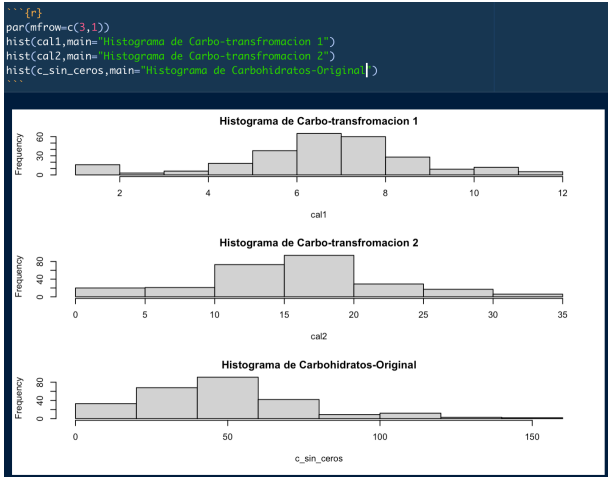
$\text{cal1} = \sqrt{(c_sin_ceros)+1}$ <- Primera transformación

$\text{cal2} = ((c_sin_ceros)^{\lambda}-1)/\lambda$ <- Segunda transformación

3.- Analiza la normalidad de las transformaciones obtenidas con los datos originales.

Utiliza como argumento de normalidad:

Analizamos cada una de las variables, las cuales son sin ceros , con la primera transformación y la segunda.



```

[[r]
summary(c_sin_ceros)
print("Kurtosis")
kurtosis(c_sin_ceros)
print("Sesgo")
skewness(c_sin_ceros)
D=ad.test(c_sin_ceros)
print("P")
D$p.value

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   31.00   45.00   48.35   61.00   142.00
[1] "Kurtosis"
[1] 1.324083
[1] "Sesgo"
[1] 0.5021952
[1] "p"
[1] 2.546548e-10

```

```

[[r]
summary(cal1)
print("Kurtosis")
kurtosis(cal1)
print("Sesgo")
skewness(cal1)
D=ad.test(cal1)
print("P")
D$p.value

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.414   5.657   6.782   6.681   7.874   11.958
[1] "Kurtosis"
[1] 0.7685457
[1] "Sesgo"
[1] -0.3931435
[1] "p"
[1] 8.30067e-10

```

```

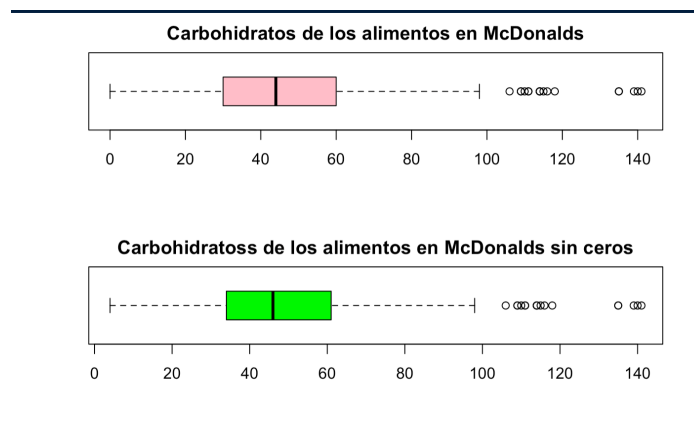
[[r]
summary(cal2)
print("Kurtosis")
kurtosis(cal2)
print("Sesgo")
skewness(cal2)
print("P")
D=ad.test(cal2)
D$p.value

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00   12.12   15.72   15.67   19.36   33.98
[1] "Kurtosis"
[1] 0.6381974
[1] "Sesgo"
[1] -0.08250202
[1] "p"
[1] 8.1823e-08

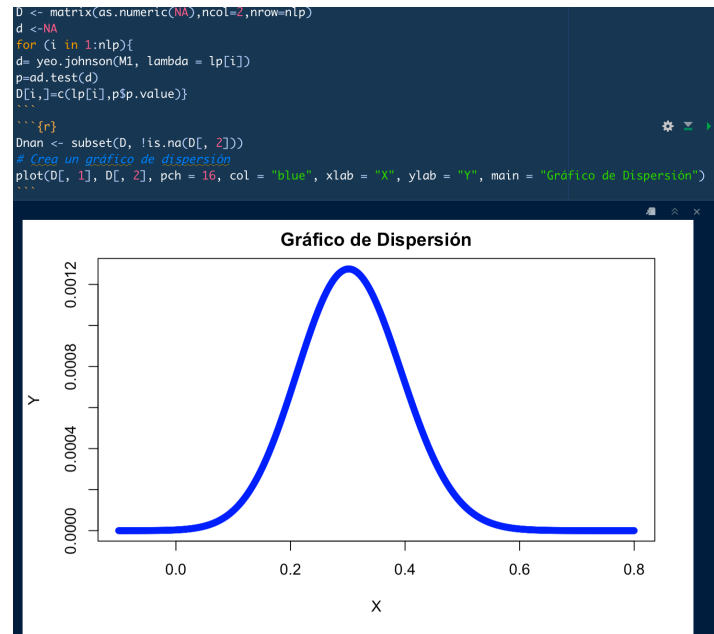
```

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

Aqui comparamos como cambian los datos con y sin ceros.



5.- 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).



6.- Escribe la ecuación del modelo encontrado.

$cal3 = ((M1)^{l-1})/l <-$ Transformación

$cal3_{yj} = yeo.johnson(M1, lambda = l1) <-$

Transformación

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

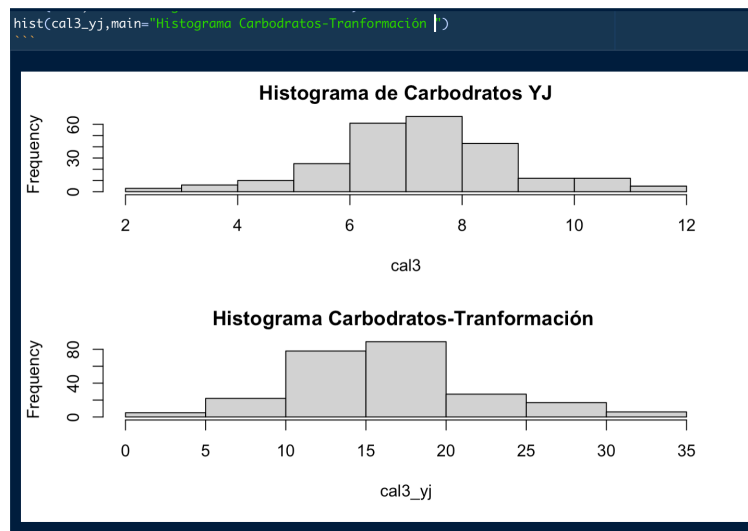
```
max_y = which.max(D[, 2])
l1 = D[max_y, 1]

[1] 0.302

library(VGAM)
library(stats4)
library(splines)
cal3 <- yeo.johnson(M1, lambda = l1)
summary(cal3)
print("Curtosis")
kurtosis(cal3)
print("Sesgo")
skewness(cal3)
D = ad.test(cal3)
print("P")
D$p.value
print("Lambda")
l1
[1] 0.302
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
"Curtosis"	2.072	6.378	7.280	7.282	8.205	11.479

```
[1] "Curtosis"
[1] 0.6878867
[1] "Sesgo"
[1] -0.06490649
[1] "p"
[1] 0.001275547
[1] "Lambda"
[1] 0.302
```



8.- Define la mejor transformación de los datos de acuerdo a las características de los modelos que encuentre.

La mejor transformación fue la de Yeo Johnson ya que se obtuvo una lambda de 0.302 un valor muy alto y mas cercano a uno a comparación con las primeras transformaciones donde lambda era muy pequeño.

9.- Concluye sobre las ventajas y desventajas de los modelos de Box Cox y de Yeo Johnson.

Modelo de Transformación de Box-Cox:

Ventajas:

- **Solo datos positivos:** Debe de ser aplicado solo a datos positivos, ya que incluye una transformación logarítmica.
- **Parametrización:** Introduce un único parámetro (λ) para ajustar la función que esta transformando.

Desventajas:

- **Sensibilidad al valor cero:** La aplicabilidad se limita so los datos contienen ceros, el valor de λ no se puede calcular.
- **No siempre mejora la normalidad:** Puede no funcionar bien con datos que ya son aproximadamente normales.

Modelo de Transformación de Yeo-Johnson:

Ventajas:

- **Flexibilidad:** Puede aplicarse a una gama más amplia de datos, incluyendo aquellos con valores negativos y ceros.
- **Complejidad:** Al poder usar valores positivos negativos y positivos es ligeramente más complejo que el modelo de Box-Cox.
- **Parámetros adicionales:** Introduce un parámetro adicional (λ) para manejar la transformación de datos negativos.

10.- 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

Diferencias entre Transformación y Escalamiento:

- Objetivo:
 - **Transformación:** Una transformación matemática implica como su nombre lo dice, transformas la distribución de los datos para que cumplan las suposiciones estadísticas, como la normalidad en este caso.
 - **Escalamiento:** El escalamiento se refiere a ajustar los valores de los datos para que estén en una escala específica que nos permita entender mejor.
- Tipo de Operación:
 - **Transformación:** Involucra operaciones matemáticas que provienen de un espacio métrico y cumplen con ciertas axiomas, éstos se aplican sobre los datos, como logaritmos, raíces cuadradas o exponenciales.
 - **Escalamiento:** Se aplica una función para ajustar y cambiar la escala, pero no modifica los datos.
- Efecto en la Distribución:
 - **Transformación:** Convierte totalmente la distribución original, cambiando los datos sesgados en simétricos y modificando la variabilidad.
 - **Escalamiento:** No afecta intrínsecamente a los datos solo provoca un cambio de longitud en casa uno de los ejes.

Cuándo utilizar Transformación o Escalamiento:

- **Transformación:**
 - Se hace uso de éstas cuando los datos no siguen una distribución normal y se requiere de una para estudiar adecuadamente las suposiciones estadísticas.
 - Con el fin de reducir la variabilidad
- **Escalamiento:**
 - Al momento de trabajar con algoritmos que son sensibles a la escala de las características, como regresión lineal, k-means o métodos similares a gradiente descendiente.
 - En algunos algoritmos donde se utilizan medidas de distancia o similitud, como k-nearest neighbors o métodos de clustering.
 - Cuando quieres evitar que ciertas características se sobrepongan gracias a sus valores más grandes.