Tarea #: 1
Tema: Exploración de datos
Fecha entrega: 11:59 pm Marzo 06 de 2023
Objetivo: Utilizar conceptos estadísticos para entender la relación entre las variables de una base de datos. Adicionalmente, utilizar python como herramienta de exploración de datos y validación de hipótesis.
Entrega: Crear un repositorio en su github personal. Dentro del proyecto debe existir una carpeta llamada tarea 1, dentro debe tener una carpeta doc con este documento incluyendo todas las respuestas y los gráficos. Adicionalmente, debe existir una carpeta src con el código del notebook utilizado. Debe adicionar la cuenta jdramirez como colaborador del proyecto y enviar un email antes de q se termine el dia indicando el commit desea le sea calificado.

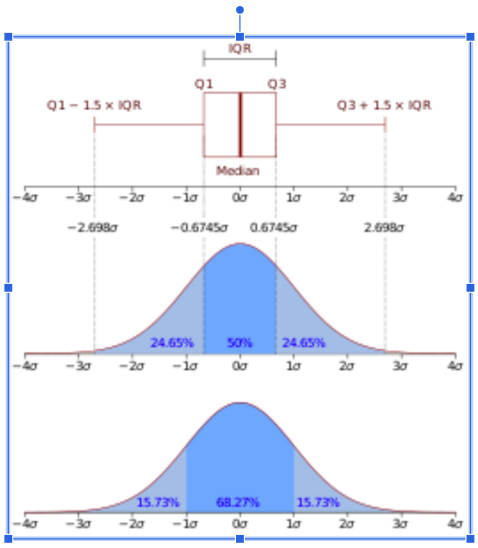
1. Utilizas el siguiente set de datos para calcular paso por paso (mostrar procedimiento y fórmulas):

Id	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

Tabla:1

1.1. ¿Cuál es la media, mediana y desviación estándar?, y la moda y los valores repeticiones de la moda para los datos categóricos.

1.2. Dibujar un boxplot a mano. Utilizando los datos de la tabla 1 y las siguientes proporciones.



1.3. Cual es la covarianza entre las 2 variables X1, X2

$$Cov(x,y) = \frac{\sum (x_i - \overline{x}) * (y_i - \overline{y})}{N}$$

1.4. Cuál es la correlación entre la variable x1 y x2 (Calcularla a mano). Correlación puede ser escrita también como:

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

1.5. Explica la relación entre covarianza y correlación.
1.6. Calcule el resultado del algoritmo K-means sobre este set de datos. Vamos a crear 2 grupos, es decir, k=2 (2 clusters).

2. Utilizando el dataset del proyecto data/CARS.csv crear:
   2.1. Distribución de cada variables:
      2.1.1. Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.
      2.1.2. Para las variables numéricas crear histogramas. Listar los modelos de carros que están más lejos de 4 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.
   2.2. Gráfico de la relación de cada variable con respecto a MPG_City:
      2.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico
      2.2.2. Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico
   2.3. Matriz de correlación.
      2.3.1. Cree la matriz de correlación, cuales son las variables más importantes para explicar la variabilidad de MPG_City. Explique por qué el coeficiente es negativo o positivo.
      2.3.2. Cree la matriz de correlación nuevamente removiendo todas los modelos de carro que fueron catalogados como un outlier. (Puede utilizar .query('Model in["MDX","TSX 4dr"]'). Existe alguna variación en la correlación.

7.1

Media:

Formula: $\frac{\sum_{i=1}^{n} x_i}{n}$

Procedimiento:

$\bar{x}_1 = \frac{1+7+0+5+6+4}{6}$

$\bar{x}_2 = \frac{4+3+4+7+2+0}{6}$

Resultado:

$x_1 = \frac{17}{6}$

$x_2 = \frac{14}{6}$

Mediana?

Formula:

Para datos ordenados la mediana es el valor central

Procedimiento:

$X_1 = [0, 1, 1, 4, 5, 6]$ la mediana es 2.5

$X_2 = [0, 1, 2, 3, 4, 4]$ la mediana es 2.5

(Promedio Valores centrales)

Resultado:

$X_1 = 4$

$X_2 = 2.5$

Desviación estandar

Formula: $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$

Procedimiento:

$a = \frac{77}{6}$

$S(x_1) = \sqrt{\frac{(7-a)^2 + (7-a)^2 + (6-a)^2 + (5-a)^2 + (6-a)^2 + (4-a)^2}{6}}$

$b = \frac{74}{6}$

$S(x_2) = \sqrt{\frac{(4-b)^2 + (3-b)^2 + (4-b)^2 + (7-b)^2 + (2-b)^2 + (6-b)^2}{6}}$

Resultado

$S(x_1) \approx 1.471$

$S(x_2) \approx 1.825$

Moda

Formula:

Valor que aparece con mas frecuencia

Procedimiento y resultado

Moda para X1 = 7

Moda para X2 = 4

boxplots X1

1. ordenar los datos

X1 : [0, 1, 1, 4, 5, 6]

2. Cuartiles

Q1 = 1 (Valor Posición 2)

Q2 = 2.5 (mediana)

Q3 = 5

IQR = Q3 - Q1 = 5 - 1 = 4

Linite inferior = Q1 - 1.5 x IQR

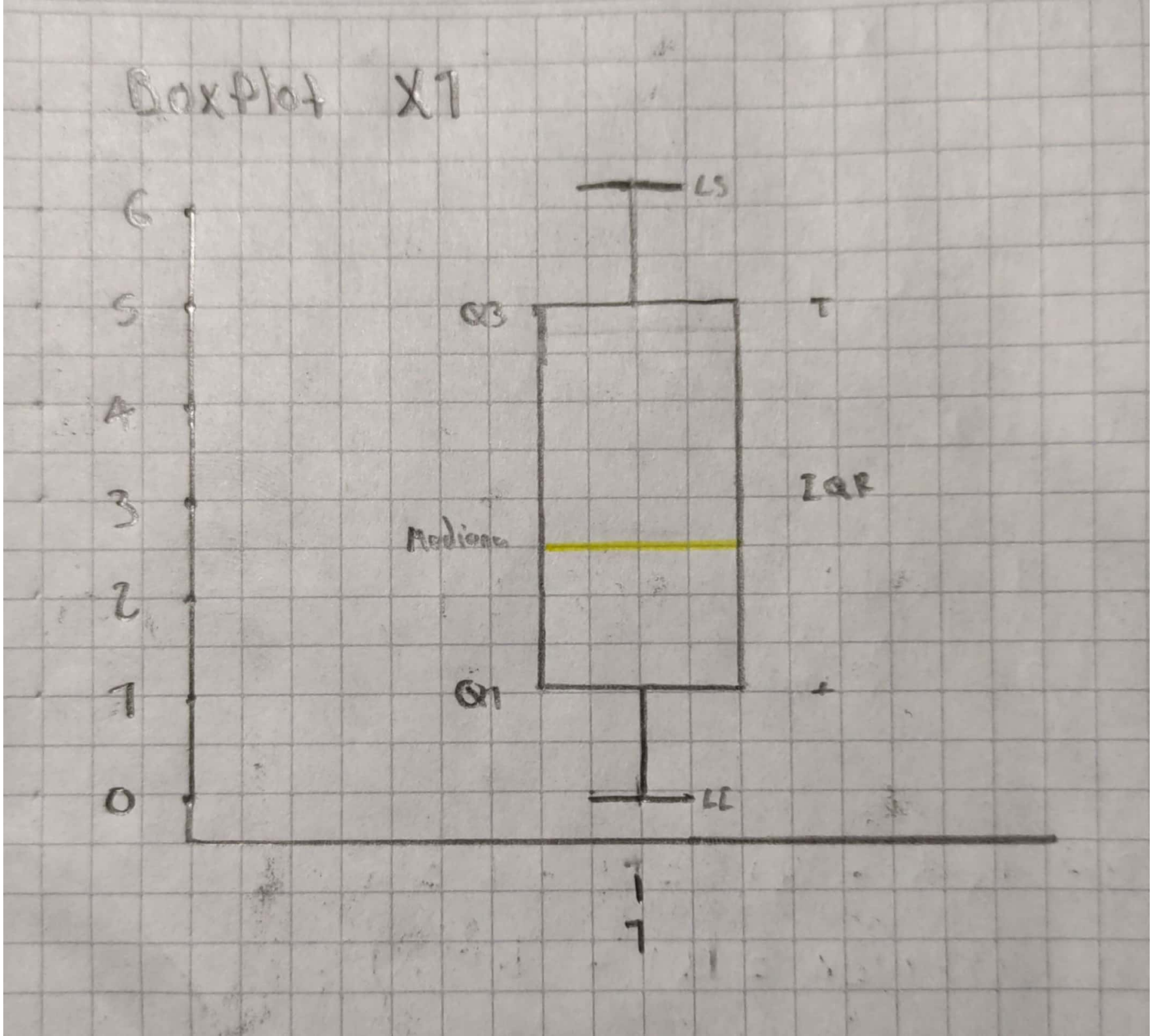
= 1 - 1.5 x 4

= -5.5

Linite superior = Q3 + 1.5 x IQR

= 5 + 1.5 x 4

= 11.5


Boxplot XT
6
5
4
3
2
1
0
LS
Q3
T
IQR
Mediana
Q1
LI

boxPlots X2

7. ordenar datos

$X2 = [0, 7, 2, 3, 4, 4]$

$Q1 = X2_2 = 7$

$Q2 = 2.5$ (mediana)

$Q3 = X_{5.5} = 4$

$IQR = Q3 - Q1 = 4 - 7 = 3$

Limite inferior: $Q1 - 1.5 \times IQR$

$= 7 - 7.5 \times 3$

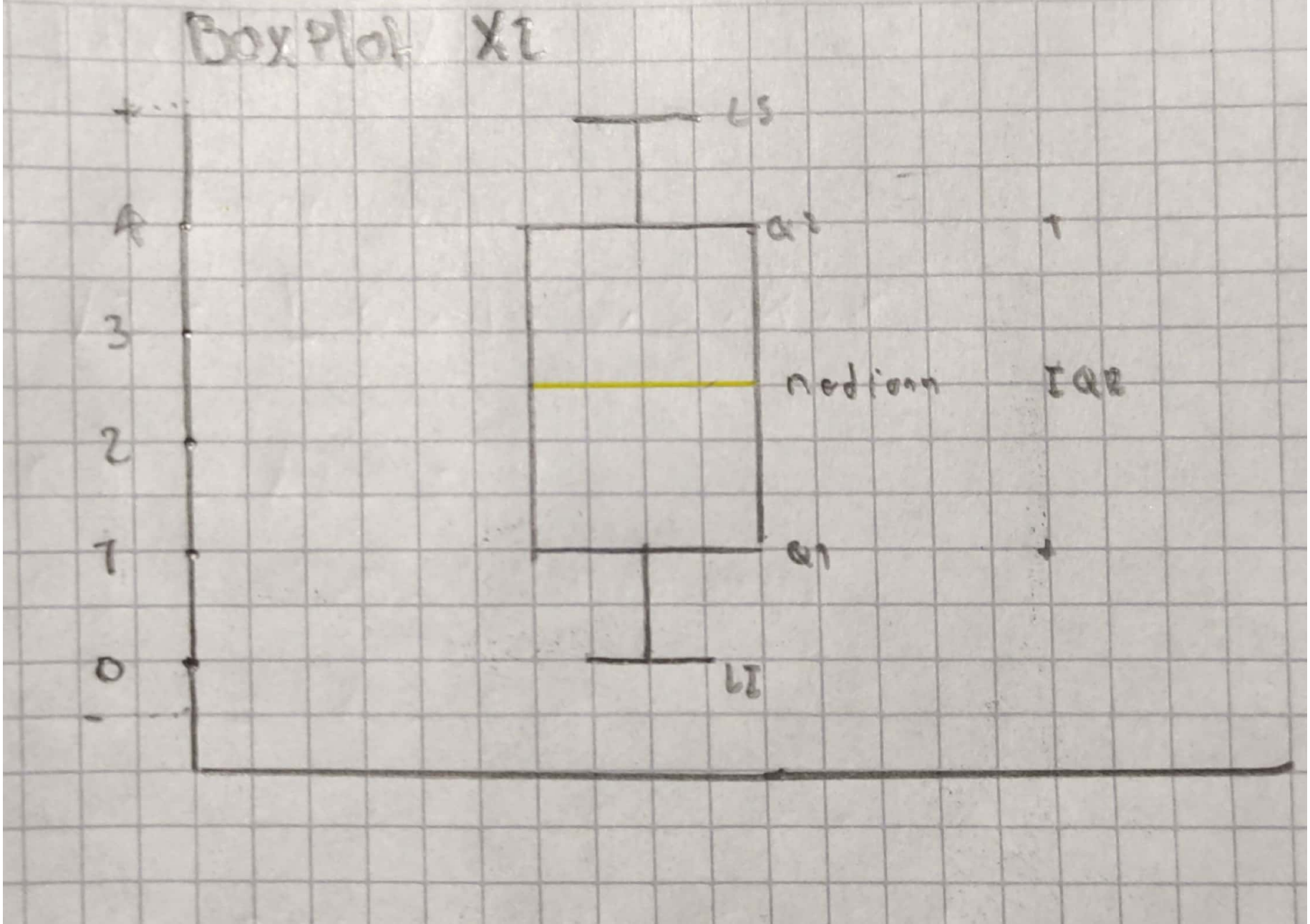
$= -3.5$

Limite superior: $Q3 + 1.5 \times IQR$

$4 + 7.5 \times 3$

$= 8.5$

DD MM



7-3

Covarianza entre X_1 y X_2

Formula

$$Cov(X_1, X_2) = \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{n}$$

Procedimiento

$\bar{X}_1 = \frac{42}{6}$ (media) $= a$

$\bar{X}_2 = \frac{24}{6}$ (media) $= b$

$Cov(X_1, X_2) =$

$$= \frac{(6-a)(4-b)+(7-a)(3-b)+(7-a)(4-b)+(4-a)(7-b)+(5-a)(7-b)+(5-a)(6-b)}{6}$$

$Cov(X_1, X_2) \approx -2.33$

Correlación X_1 y X_2

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{S(X_1)\,S(X_2)}$$

Procedimiento

$Cov(X_1, X_2) \approx -2.33$ (Calculo anteriormente)

$S(X_1) \approx 7.471$ (Calculado ")

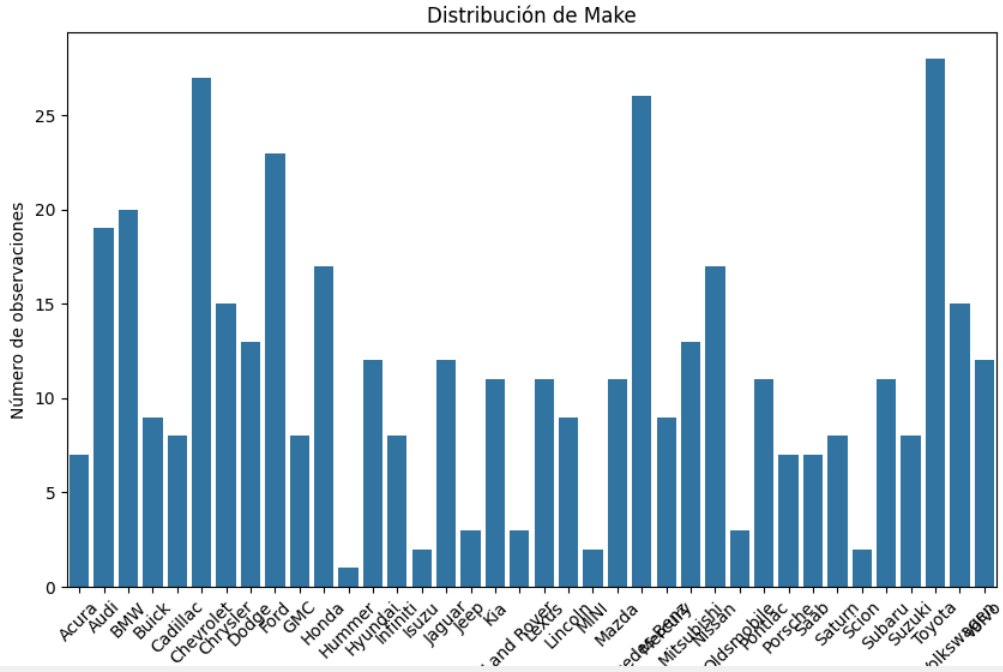
$S(X_2) \approx 7.825$ (Calculado ")

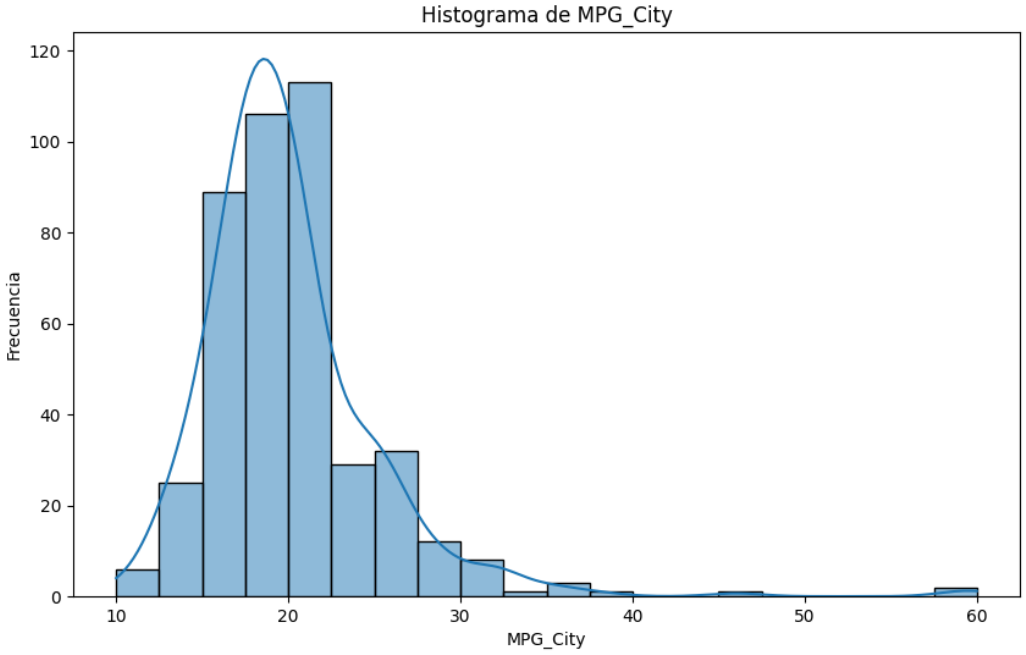
$$Corr(X_1, X_2) \approx \frac{-2.33}{(7.471)(7.825)}$$

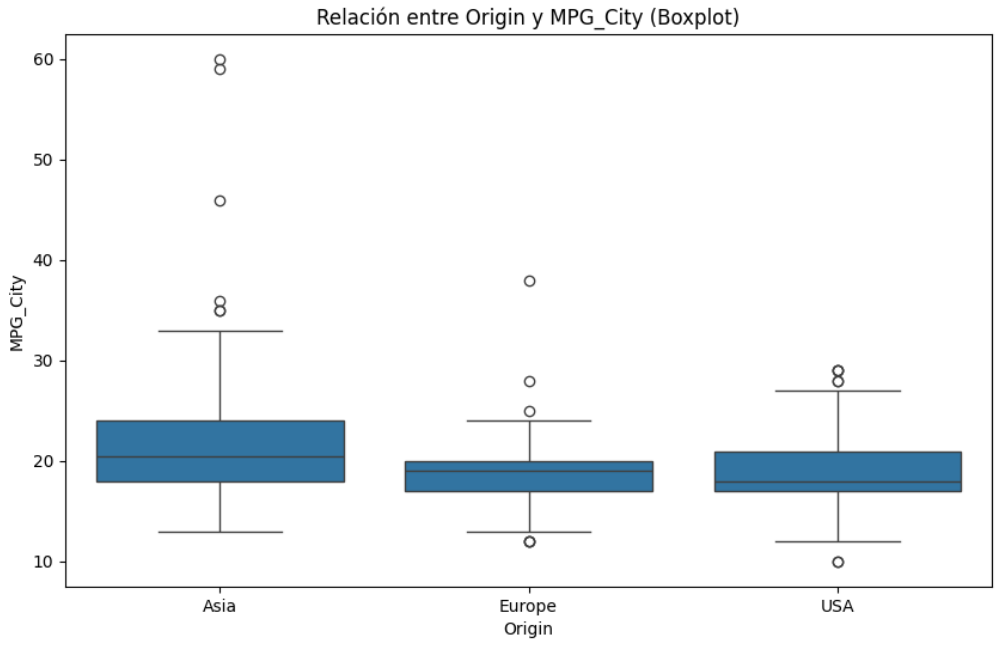
$$Corr(X_1, X_2) \approx -0.82$$

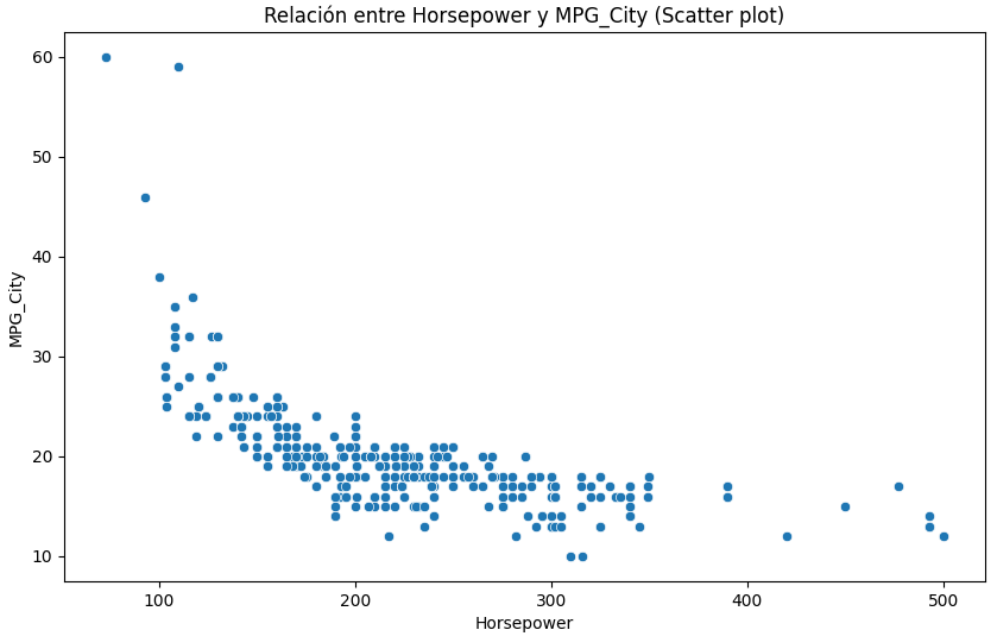
1.5. Explica la relación entre covarianza y correlación.

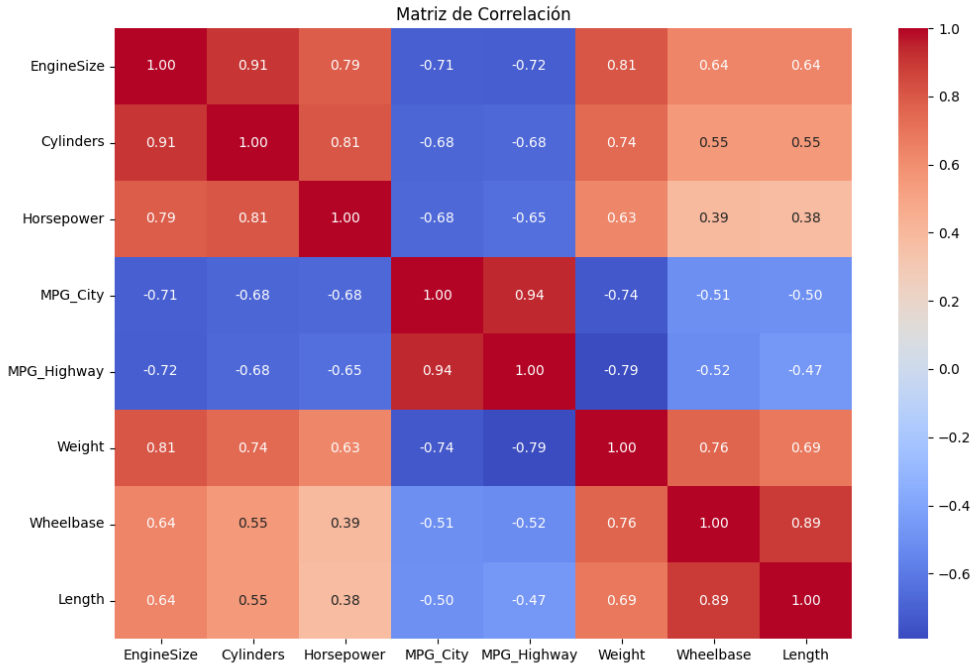
La covarianza proporciona información sobre la dirección de la relación, mientras que la correlación además cuantifica la fuerza y normaliza la medida en una escala más fácil de interpretar. Una correlación cercana a 1 o -1 indica una relación más fuerte, mientras que valores cercanos a 0 indican una relación más débil.

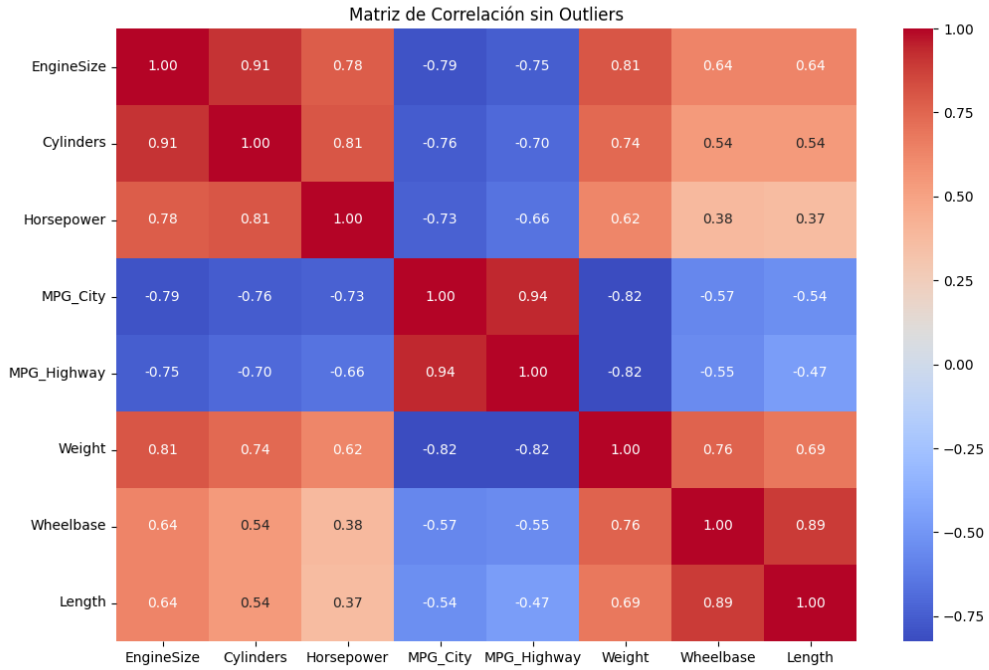

Distribución de Make
Número de observaciones
0
5
10
15
20
25
Acura
Audi
BMW
Buick
Cadillac
Chevrolet
Chrysler
Dodge
Ford
GMC
Honda
Hummer
Hyundai
Infiniti
Isuzu
Jaguar
Jeep
Kia
Land Rover
Lexus
Lincoln
MINI
Mazda
Mitsubishi
Nissan
Pontiac
Porsche
Saab
Saturn
Scion
Subaru
Suzuki
Toyota


Histograma de MPG_City
Frecuencia
0
20
40
60
80
100
120
10
20
30
40
50
60
MPG_City


Relación entre Origin y MPG_City (Boxplot)
MPG_City
60
50
40
30
20
10
Asia
Europe
USA
Origin


Relación entre Horsepower y MPG_City (Scatter plot)
MPG_City
60
50
40
30
20
10
100
200
300
400
500
Horsepower


Matriz de Correlación
EngineSize 1.00 0.91 0.79 -0.71 -0.72 0.81 0.64 0.64
Cylinders 0.91 1.00 0.81 -0.68 -0.68 0.74 0.55 0.55
Horsepower 0.79 0.81 1.00 -0.68 -0.65 0.63 0.39 0.38
MPG_City -0.71 -0.68 -0.68 1.00 0.94 -0.74 -0.51 -0.50
MPG_Highway -0.72 -0.68 -0.65 0.94 1.00 -0.79 -0.52 -0.47
Weight 0.81 0.74 0.63 -0.74 -0.79 1.00 0.76 0.69
Wheelbase 0.64 0.55 0.39 -0.51 -0.52 0.76 1.00 0.89
Length 0.64 0.55 0.38 -0.50 -0.47 0.69 0.89 1.00
EngineSize Cylinders Horsepower MPG_City MPG_Highway Weight Wheelbase Length
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6


Matriz de Correlación sin Outliers
EngineSize 1.00 0.91 0.78 -0.79 -0.75 0.81 0.64 0.64
Cylinders 0.91 1.00 0.81 -0.76 -0.70 0.74 0.54 0.54
Horsepower 0.78 0.81 1.00 -0.73 -0.66 0.62 0.38 0.37
MPG_City -0.79 -0.76 -0.73 1.00 0.94 -0.82 -0.57 -0.54
MPG_Highway -0.75 -0.70 -0.66 0.94 1.00 -0.82 -0.55 -0.47
Weight 0.81 0.74 0.62 -0.82 -0.82 1.00 0.76 0.69
Wheelbase 0.64 0.54 0.38 -0.57 -0.55 0.76 1.00 0.89
Length 0.64 0.54 0.37 -0.54 -0.47 0.69 0.89 1.00
EngineSize Cylinders Horsepower MPG_City MPG_Highway Weight Wheelbase Length
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

```
  import pandas as pd
Outliers:
                                    Model  MPG_City
149  Civic Hybrid 4dr manual (gas/electric)        46
150              Insight 2dr (gas/electric)        60
373                Prius 4dr (gas/electric)        59
Estadístico de prueba: 0.8078395264093217, Valor p: 3.38966959244819 45e-22
La distribución no es normal.
Outliers:
                                    Model  MPG_City
149  Civic Hybrid 4dr manual (gas/electric)        46
150              Insight 2dr (gas/electric)        60
373                Prius 4dr (gas/electric)        59
Estadístico de prueba: 0.8078395264093217, Valor p: 3.38966959244819 45e-22
La distribución no es normal.
Variables más importantes:
MPG_Highway    0.941021
Weight         0.737966
EngineSize     0.709471
Cylinders      0.684402
Horsepower     0.676699
Wheelbase      0.507284
Length         0.501526
Name: MPG_City, dtype: float64
```

Por: Juan Pablo Toro Hurtado

CC: 1001477295

Correo: Juanpatoro2011@gmail.com