Juan Diego Valencia Sánchez

Universidad católica del oriente

Facultad de ingenierías

Estadística III

2023

|  |  |  |
| --- | --- | --- |
| id | X1 | X2 |
| 1 | 1 | 4 |
| 2 | 1 | 3 |
| 3 | 0 | 4 |
| 4 | 5 | 1 |
| 5 | 6 | 2 |
| 6 | 4 | 0 |

**1.1** ¿Cuál es la media, mediana y desviación estándar?, y la moda y los

valores repeticiones de la moda para los datos categóricos.

Media(x1) = (1+1+0+5+6+4) /6 = 2.83

Media(x2) = (4+3+4+1+2+0) /6 = 2.3

Mediana(x1) = 0,1,1,4,5,6 = (1+4) /2 = 2.5

Mediana (x2) = 0, 1, 2, 3, 4, 4 = 2, 3 = (2 + 3) / 2 = 2.5

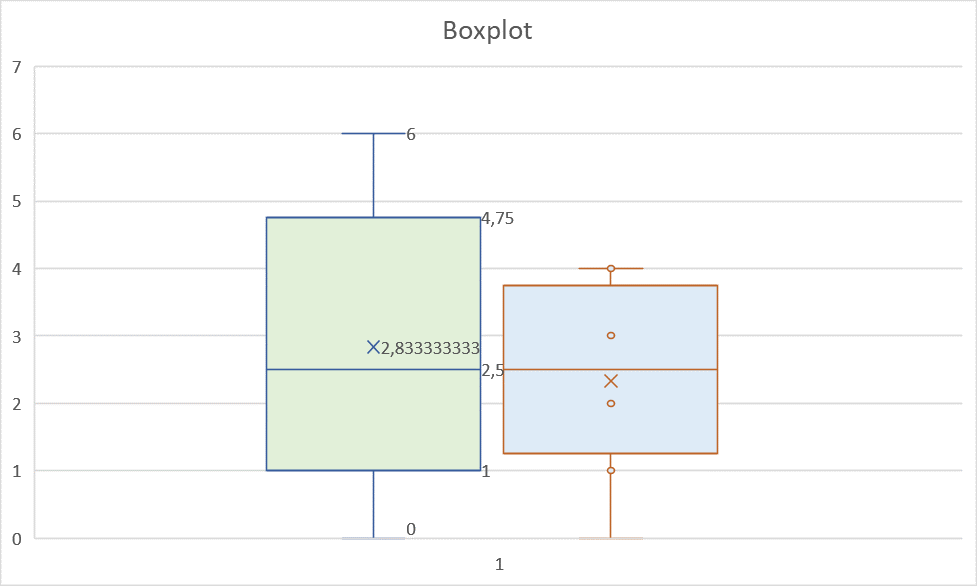
Desviación estándar(x1) =



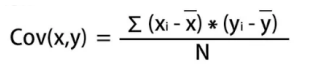
Desviación estándar (x2

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

siguientes proporciones.



**1.3** Cuál es la covarianza entre las 2 variables X1, X2



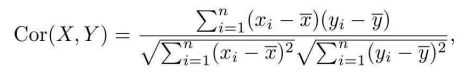
X1’ = (1+1+0+5+6+4) /6 = 2.83

X2‘= (4+3+4+1+2+0) /6 = 2.3

Cov (x,y) = ((1-2.83)\* (4-2.3)+ (1-2.83)\* (3-2.3)+ (0-2.83)\* (4-2.3)+(5-2.83)\* (1-2.3)+(6-2.83)\* (2-2.3)+ (4-2.83)\* (0-2.3))/6 = -2.611

**1.4** Cuál es la correlación entre la variable x1 y x2 (Calcularla a mano).

Correlación puede ser escrita también como:





**1.5** Explica la relación entre covarianza y correlación.

La covarianza de -2.6 indica que hay una relación negativa fuerte entre las dos variables y la correlación de -0.77 indica que hay una correlación negativa fuerte entre las dos variables. Ya que ambas medidas indican una fuerte relación negativa entre las dos variables, podemos concluir que cuando una variable aumenta, la otra disminuye y viceversa,

**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).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **X1** | **X2** | **X1-Mean** | **X2-Mean** |
| 2 | 0 | 4 | -2,833333333 | 1,666666667 |
| 5 | 5 | 1 | 2,166666667 | -1,333333333 |
| 1 | 1 | 4 | -1,833333333 | 1,666666667 |
| 3 | 1 | 3 | -1,833333333 | 0,666666667 |
| 4 | 4 | 0 | 1,166666667 | -2,333333333 |
| 6 | 6 | 2 | 3,166666667 | -0,333333333 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Random** | **X1** | **X2** | **Iteración 1** | **Iteracion 2** |
| 0 | -0,333333333 | 0,166666667 | 8,5 | 12,0625 |
| 0 | -0,333333333 | 0,166666667 | 8,5 | 5,5625 |
| 1 | 0,166666667 | -0,083333333 | 4,5 | 7,0625 |
| 1 | 0,166666667 | -0,083333333 | 2,5 | 4,5625 |
| 1 | 0,166666667 | -0,083333333 | 8,5 | 6,0625 |
| 1 | 0,166666667 | -0,083333333 | 12,5 | 9,0625 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **New Label** | **X1** | **X2** | **Iteración 1** | **Iteracion 2** | **New Label** |
| 0 | -2,166666667 | 1,333333333 | 0,555555556 | 34 | 0 |
| 1 | 2,166666667 | -1,333333333 | 25,88888889 | 4,93038E-32 | 1 |
| 0 | -2,166666667 | 1,333333333 | 0,222222222 | 25 | 0 |
| 0 | -2,166666667 | 1,333333333 | 0,555555556 | 20 | 0 |
| 1 | 2,166666667 | -1,333333333 | 24,55555556 | 2 | 1 |
| 1 | 2,166666667 | -1,333333333 | 31,22222222 | 2 | 1 |

2.1. Distribución de cada variable:

2.1.1. Para las variables categóricas un gráfico de barras. Categoría

numero de observaciones.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

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.

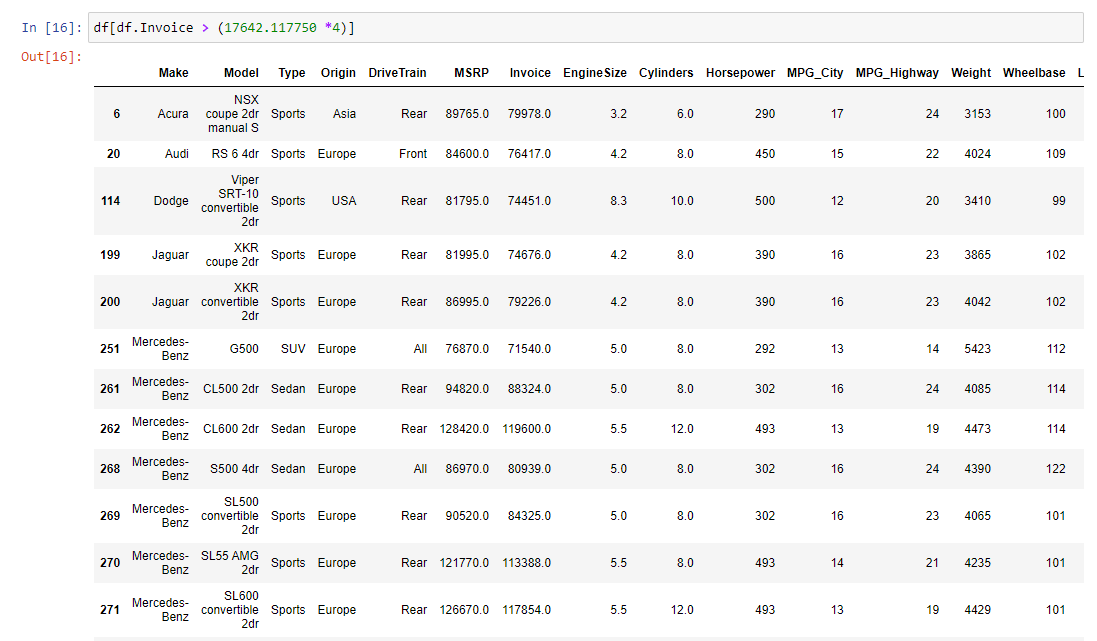
Gráfico, Histograma

Descripción generada automáticamente



Gráfico, Histograma

Descripción generada automáticamente



Texto

Descripción generada automáticamente

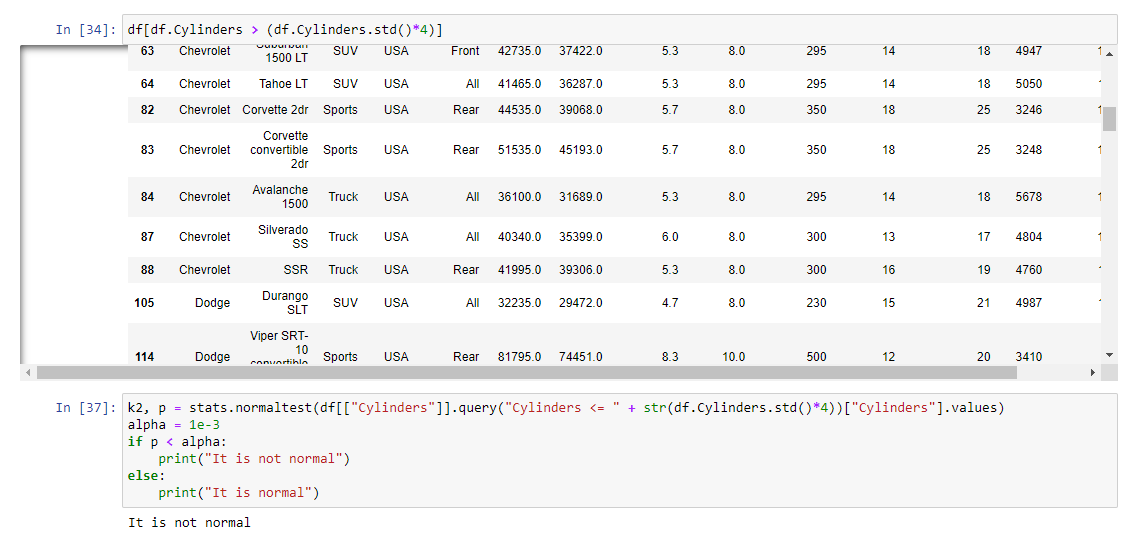
Gráfico, Histograma

Descripción generada automáticamente



Gráfico, Gráfico de líneas

Descripción generada automáticamente



Gráfico, Histograma

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

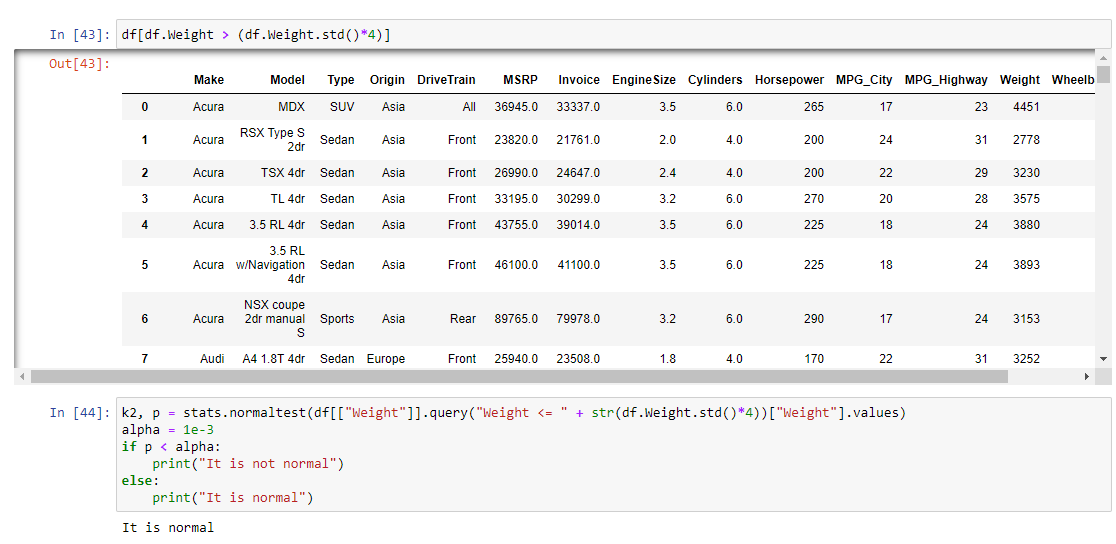
Gráfico, Histograma

Descripción generada automáticamente



Gráfico, Histograma

Descripción generada automáticamente



Gráfico, Histograma

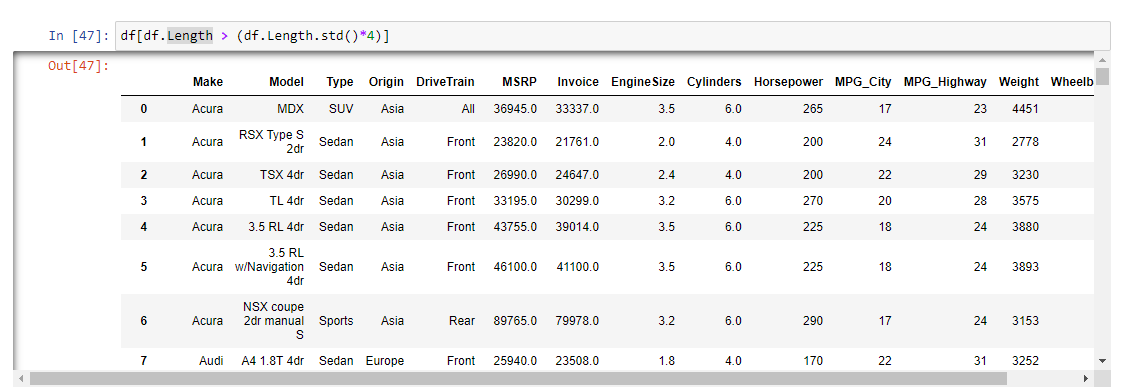
Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

Gráfico, Histograma

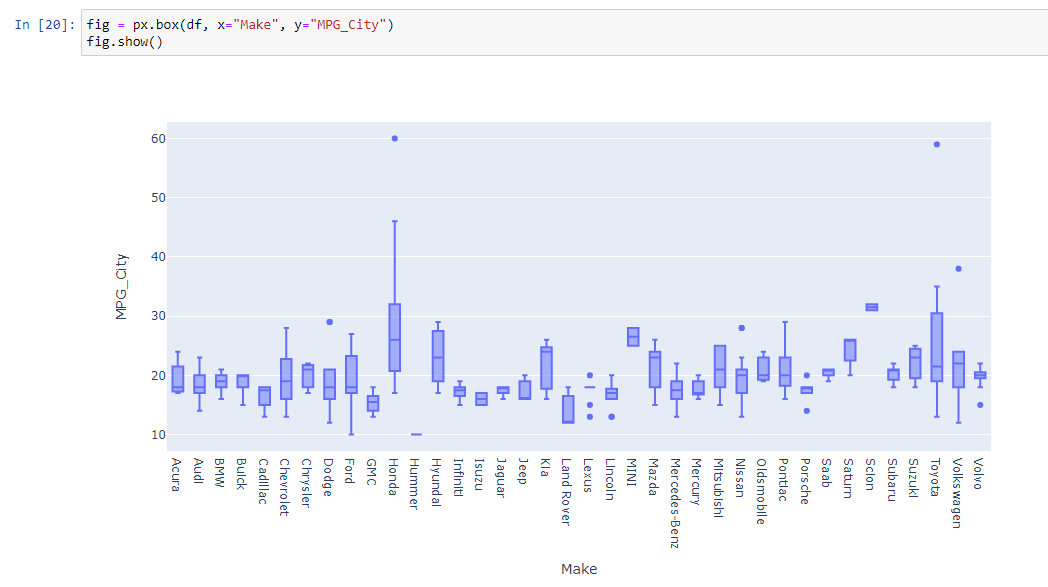
Descripción generada automáticamente



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



Interfaz de usuario gráfica

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

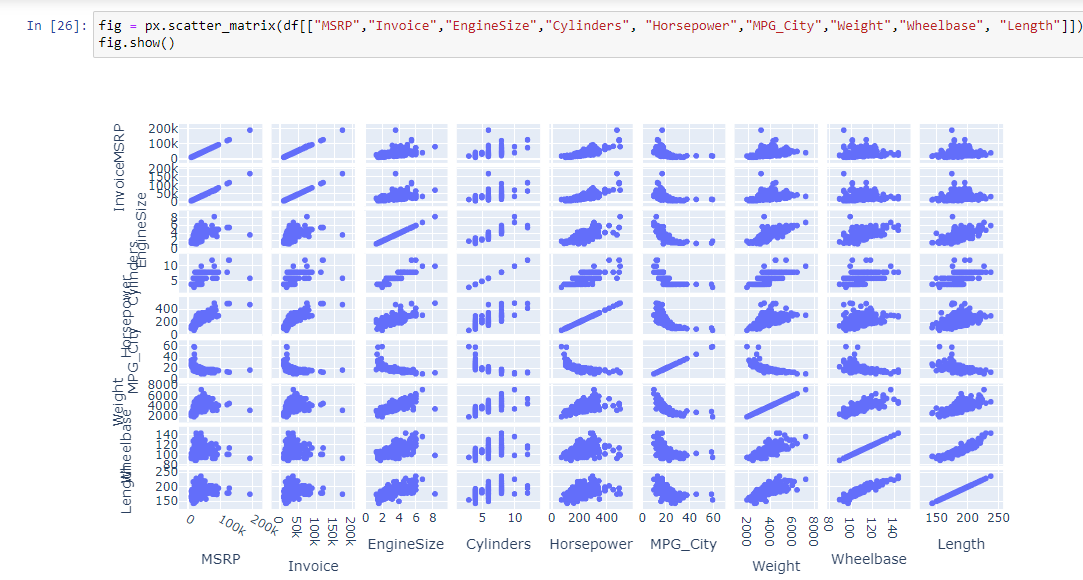
Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

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, cuáles son las variables más

importantes para explicar la variabilidad de MPG\_City. Explique por

qué el coeficiente es negativo o positivo.

Según los coeficientes de correlación presentados, las variables más importantes para explicar la variabilidad de MPG\_City son: MPG\_Highway, Weight, Cylinders, Horsepower, y EngineSize.MPG\_Highway tiene un coeficiente de correlación positivo muy alto (0.9410205), lo que indica que los vehículos que tienen una mayor eficiencia de combustible en carretera también tienden a tener una mayor eficiencia de combustible en la ciudad.Weight, Cylinders, Horsepower y EngineSize tienen coeficientes de correlación negativos (de -0.73966 a -0.7094711), lo que indica que a medida que estas variables aumentan, la eficiencia de combustible disminuye.El coeficiente negativo indica que hay una relación inversa entre la variable en cuestión y la eficiencia de combustible. Por ejemplo, a medida que aumenta el peso del vehículo, es probable que su eficiencia de combustible disminuya. Del mismo modo, a medida que aumenta la potencia del motor, es probable que la eficiencia de combustible disminuya debido a que se necesita más combustible para alimentar un motor más potente.