ESTADISTICA 3

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 |

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

17 / 6 = 2.83

Media X2

4 + 3 + 4 + 1 + 2 +0 = 14

14/6 = 2.33

Mediana X1

0 , 1 , 1 , 4 , 5 , 6

1+ 4 = 5

5/2 = 2.5

Mediana X2

0 , 1 , 2 , 3 , 4, 4

2+3 = 5

5/2 = 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.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

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 = 2.33

Covarianza (X1, X2) = ((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:

Correlación (X1, X2) = ((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)) / (√((1 − 2.83)ʌ2 + (1 − 2.83)ʌ2 + (0 − 2.83)ʌ2+ (5 − 2.83)ʌ2+ (6 − 2.83)ʌ2++ (4 − 2.83) ʌ2)\* √((4 − 2.3) ʌ2 + (3 − 2.3) ʌ2 + (4 − 2.3) ʌ2+ (1 − 2.3) ʌ2+ (0 − 2.3)ʌ2+(2 − 2.3)ʌ2))) = -0.77

1.5. Explica la relación entre covarianza y correlación.

correlación = covarianza / (desviación estándar de la variable 1 x desviación estándar de la variable 2)

En este caso, si sustituimos los valores conocidos, obtenemos:

correlación = -2.611 / (2.48 x 1.63) = -0.775

El valor de la correlación calculado (-0.775) es muy cercano al valor de la correlación que se nos dio (-0.77). Esto sugiere que la correlación dada es consistente con los valores de covarianza y desviaciones estándar proporcionados. Además, al tener en cuenta las desviaciones estándar de ambas variables, podemos decir que hay una fuerte correlación negativa entre las dos variables, lo que sugiere que cuando una variable aumenta, la otra tiende a disminuir y viceversa, y que la magnitud de esta relación es relativamente alta.

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 | -2,8333333 | 1,6666667 |
| 5 | 5 | 1 | 2,1666667 | -1,3333333 |
| 1 | 1 | 4 | -1,8333333 | 1,6666667 |
| 3 | 1 | 3 | -1,8333333 | 0,6666667 |
| 4 | 4 | 0 | 1,1666667 | -2,3333333 |
| 6 | 6 | 2 | 3,1666667 | -0,3333333 |

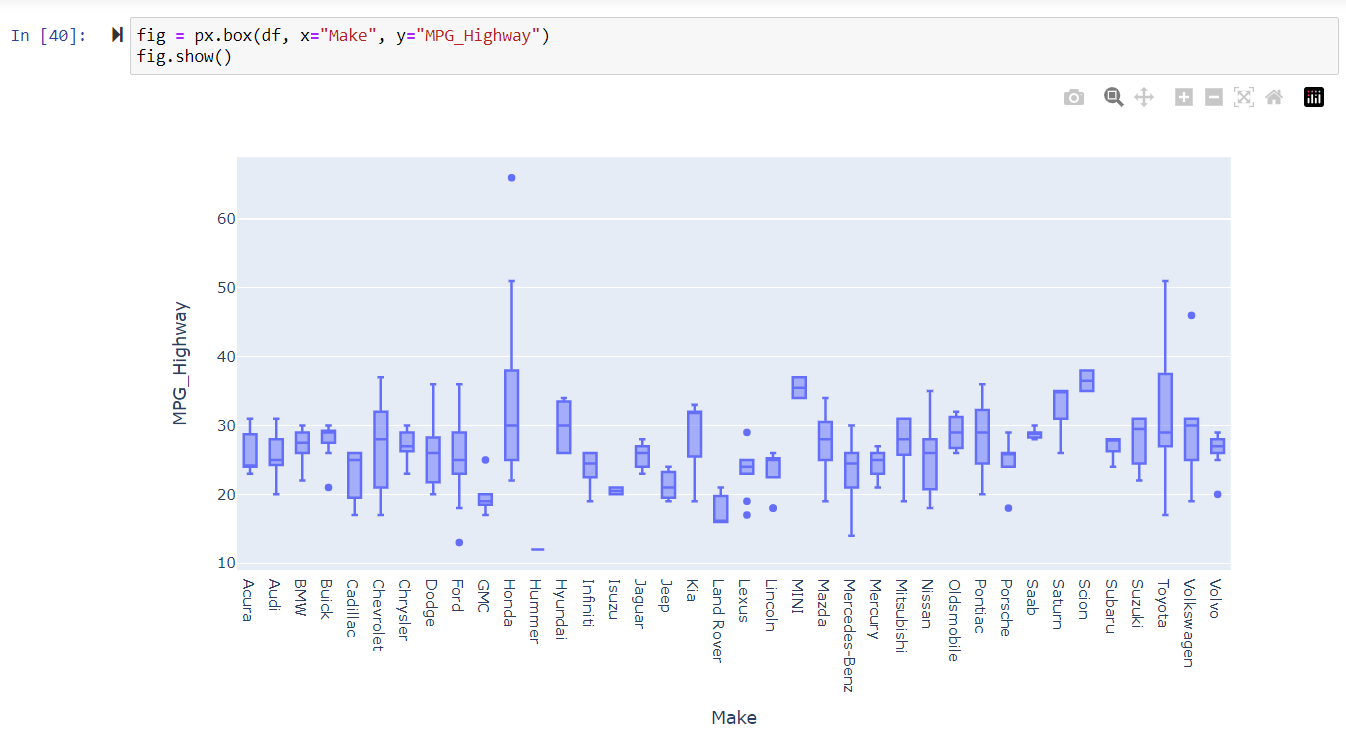
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Aleatorio** | **X1** | **X2** | **Iteración 1** | **Iteracion 2** |
| 0 | -0,3333333 | 0,1666667 | 8,5 | 12,0625 |
| 0 | -0,3333333 | 0,1666667 | 8,5 | 5,5625 |
| 1 | 0,1666667 | -0,0833333 | 4,5 | 7,0625 |
| 1 | 0,1666667 | -0,0833333 | 2,5 | 4,5625 |
| 1 | 0,1666667 | -0,0833333 | 8,5 | 6,0625 |
| 1 | 0,1666667 | -0,0833333 | 12,5 | 9,0625 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **New Label** | **X1** | **X2** | **Iteración 1** | **Iteracion 2** | **New Label** |
| 0 | -2,1666667 | 1,3333333 | 0,5555556 | 34 | 0 |
| 1 | 2,1666667 | -1,3333333 | 25,888889 | 4,93E-32 | 1 |
| 0 | -2,1666667 | 1,3333333 | 0,2222222 | 25 | 0 |
| 0 | -2,1666667 | 1,3333333 | 0,5555556 | 20 | 0 |
| 1 | 2,1666667 | -1,3333333 | 24,555556 | 2 | 1 |
| 1 | 2,1666667 | -1,3333333 | 31,222222 | 2 | 1 |

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.



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

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

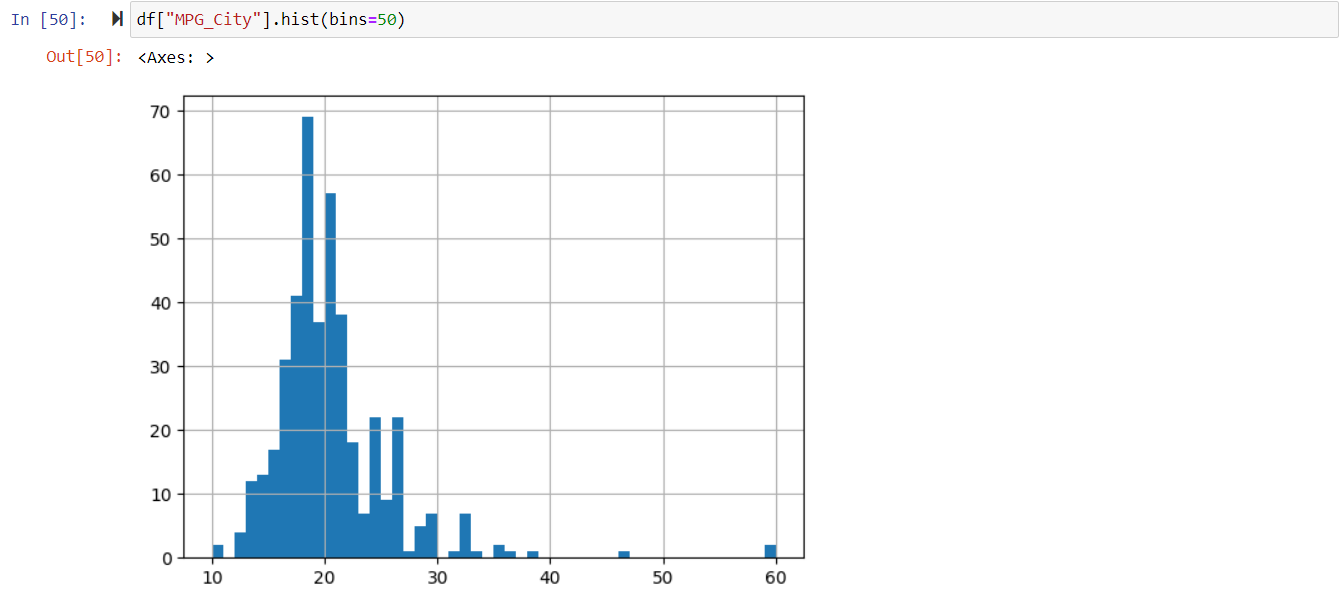
Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente



Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

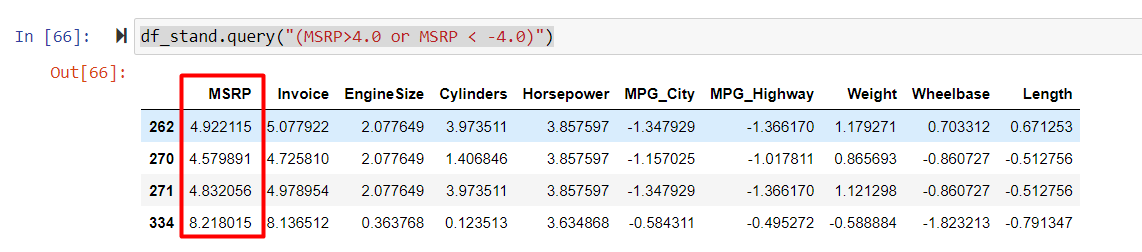
Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

Los outliers que encontramos son:

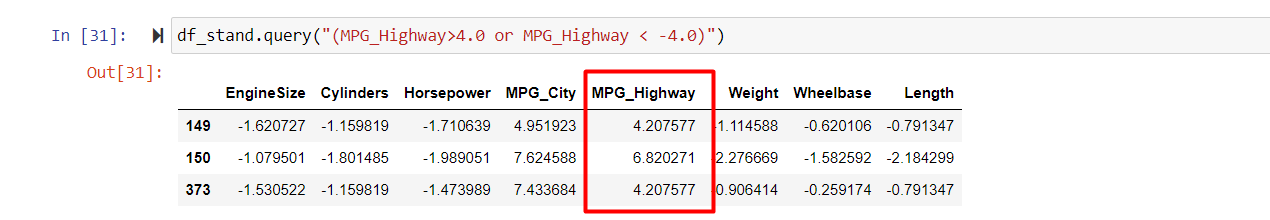


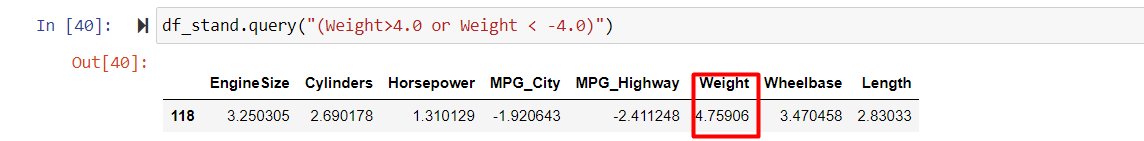
Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Interfaz de usuario gráfica, Texto, Aplicación

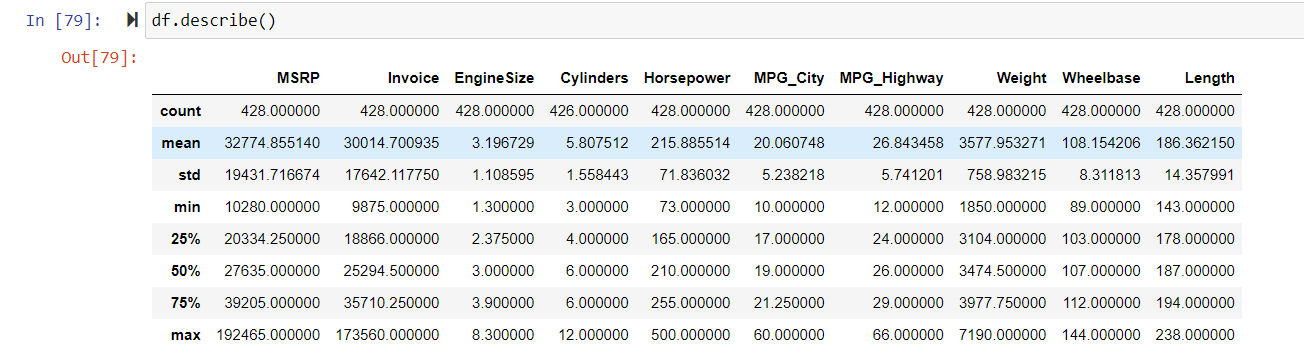
Descripción generada automáticamente





Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente



2.2.1

Cada una de las marcas están representadas de acuerdo a sus datos relacionados, cada uno cuenta con una media que esta representada, y podemos interpretar que la los puntos mas dispersos de los otros son los que son diferentes a los demás.

Imagen que contiene Gráfico

Descripción generada automáticamente

Los datos de los modelos son muy pocos para realizar una grafica mas pequeña

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

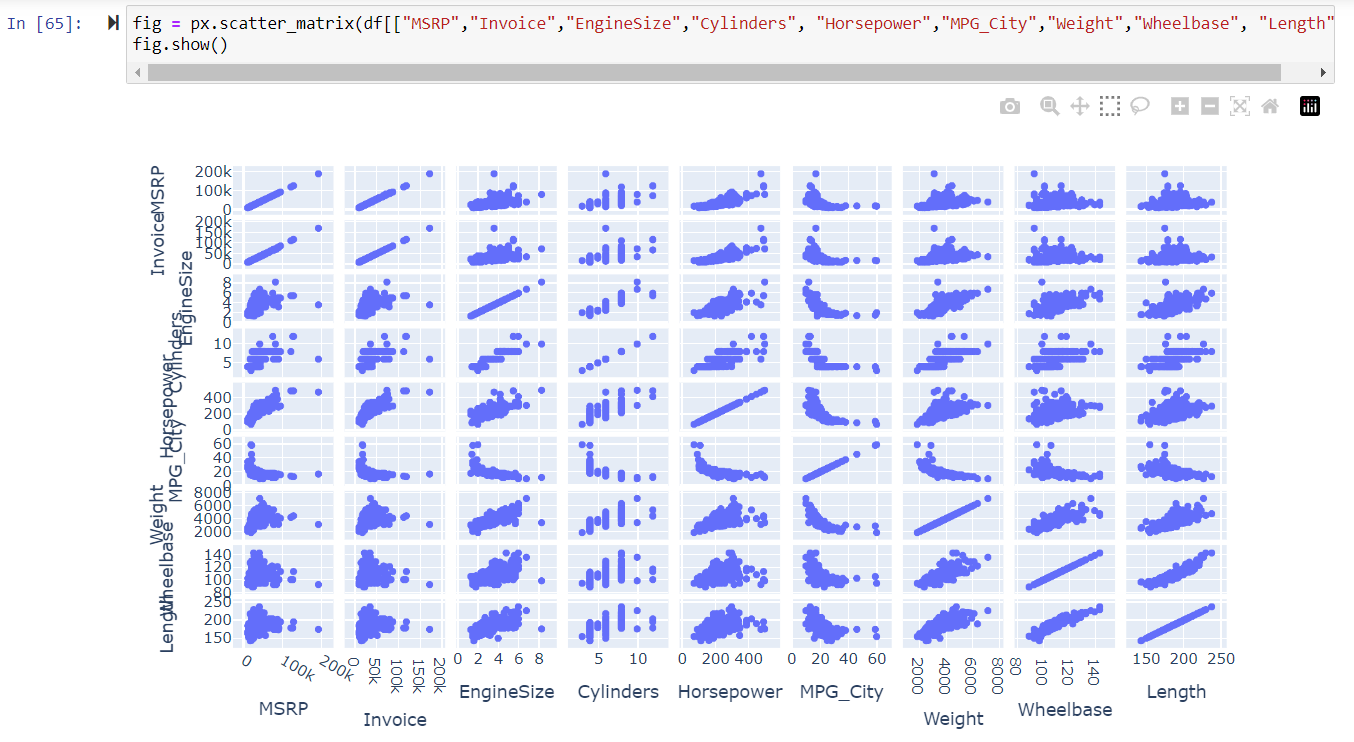
Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

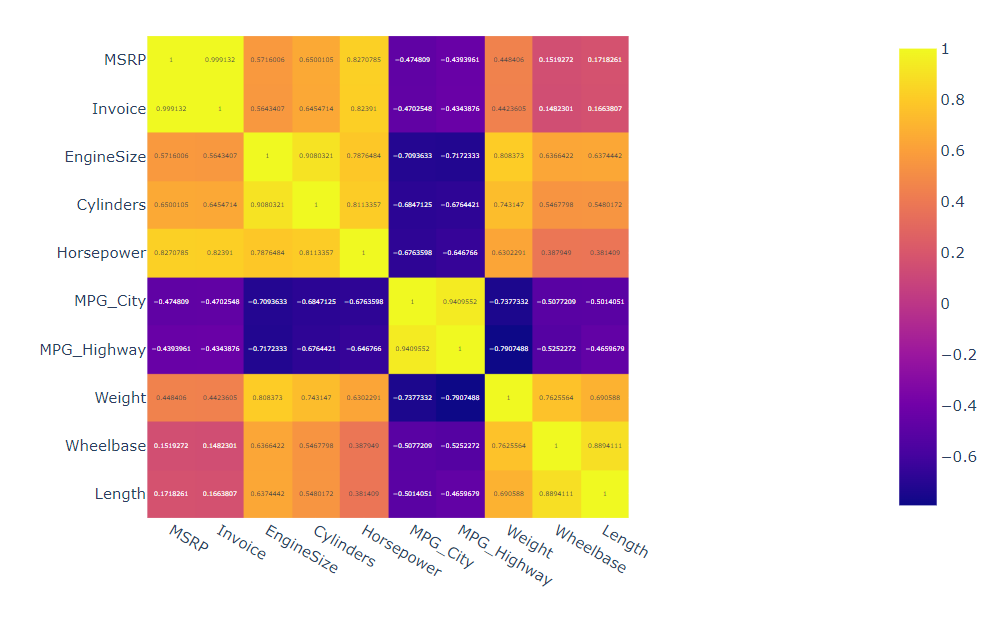
Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

2.2.2 Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico



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



Según la matriz de correlación proporcionada, las variables más importantes para explicar la variabilidad de MPG\_City son: MPG\_Highway, Weight, Cylinders, Horsepower, EngineSize, Wheelbase, MSRP, Invoice y Length, en ese orden.Las variables con coeficientes de correlación negativos, como MSRP, Invoice, EngineSize, Cylinders, Horsepower, Weight, Wheelbase y Length, indican que a medida que aumentan estas variables, se espera que el MPG\_City disminuya.

Esto significa que estas variables están inversamente relacionadas con el MPG\_City, lo que sugiere que los vehículos con motores más grandes, más cilindros, mayor peso y tamaño, y mayor potencia tienen una menor eficiencia en el consumo de combustible.

Por otro lado, la variable MPG\_Highway tiene un coeficiente de correlación positivo, lo que indica que a medida que aumenta la eficiencia del combustible en carretera (MPG\_Highway), también se espera que aumente la eficiencia del combustible en ciudad (MPG\_City). Esto sugiere que los vehículos con mejor rendimiento de combustible en carretera también tienen un mejor rendimiento de combustible en ciudad.En resumen, la matriz de correlación sugiere que la eficiencia del combustible en ciudad (MPG\_City) está influenciada por una variedad de factores, como el tamaño del motor, el número de cilindros, la potencia del motor, el peso, el tamaño y la eficiencia del combustible en carretera.

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.

