Tarea #: 1
Tema: Exploración de datos
Fecha entrega: 12:00 am Marzo 12 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.

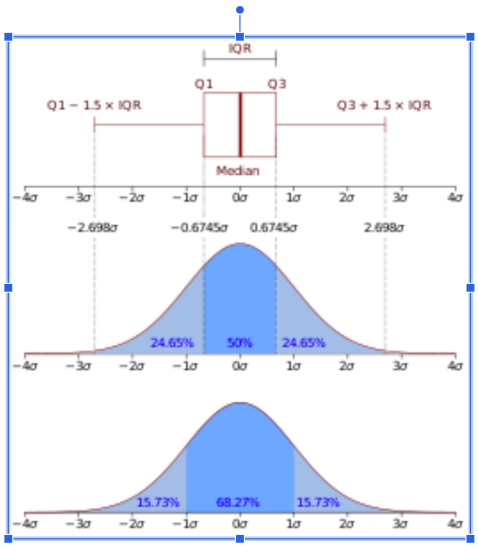
1. Utilizas el siguiente set de datos para calcular paso por paso (mostrar procedimiento y fórmulas):

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.