**Curso:** Estadistica 3
**Profesor:** Jose Daniel Ramirez Soto

**Tarea #**: 1
**Tema**: Exploración de datos Y PCA
**Fecha entrega**: 11:59 pm 11 de Marzo de 2024
**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 ):

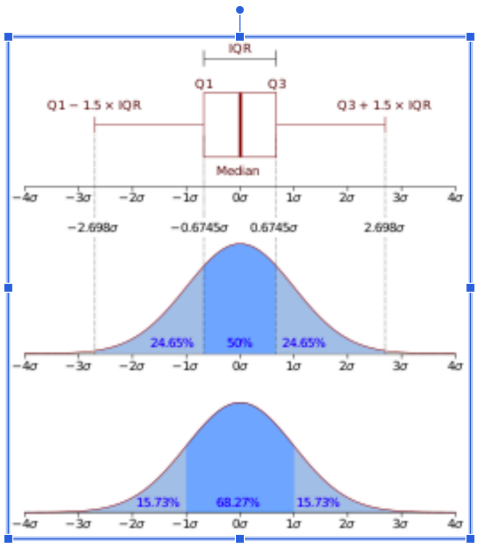
| DEPARTAENTO | PIB MILLONES(X1) | Poblacion(X_2) | PIB_percapita en MILLONES(x3) |
|---|---|---|---|
| Amazonas | 1067855.672 | 76589 | 13.94267678 |
| Antioquia | 212514957.4 | 6407102 | 33.16865524 |
| Arauca | 8548114.653 | 262174 | 32.60473828 |
| Atlántico | 63764770.77 | 2535517 | 25.1486268 |
| Bogotá D.C. | 357258620.8 | 7412566 | 48.19634938 |
| Bolívar | 51404352.37 | 2070110 | 24.83170091 |
| Boyacá | 38858162.12 | 1217376 | 31.91960588 |
| Caldas | 23953112.45 | 998255 | 23.9949837 |
| Caquetá | 5461366.78 | 401849 | 13.59059443 |
| Casanare | 23660657.37 | 420504 | 56.26737766 |
| Cauca | 25758151.71 | 1464488 | 17.58850309 |
| Cesar | 37523918.98 | 1200574 | 31.25498218 |
| Chocó | 6001844.915 | 534826 | 11.2220515 |
| Córdoba | 24991953.76 | 1784783 | 14.00279685 |
| Cundinamarca | 91945942.28 | 2919060 | 31.49847632 |
| Guainía | 497704.0127 | 48114 | 10.34426597 |
| Guaviare | 1123857.696 | 82767 | 13.57857232 |
| Huila | 24011616.06 | 1100386 | 21.82108466 |

| | | | |
|---|---|---|---|
| La Guajira | 22262575.88 | 880560 | 25.28229295 |
| Magdalena | 19738417.36 | 1341746 | 14.710994 |
| Meta | 58439500.07 | 1039722 | 56.20685151 |
| Nariño | 21775426.15 | 1630592 | 13.35430699 |
| Norte de Santander | 23056874.23 | 1491689 | 15.45689097 |
| Putumayo | 5616558.269 | 348182 | 16.13109888 |
| Quindío | 11941644.16 | 539904 | 22.11808795 |
| Risaralda | 23786362.42 | 943401 | 25.21341659 |
| San Andrés, Providencia y Santa Catalina (Archipiélago) | 2125410.333 | 61280 | 34.68358898 |
| Santander | 92276678.16 | 2184837 | 42.23504003 |
| Sucre | 11516270.76 | 904863 | 12.7270877 |
| Tolima | 30438180.15 | 1330187 | 22.8826324 |
| Valle del Cauca | 139863153.5 | 4475886 | 31.2481492 |
| Vaupés | 381851.6785 | 40797 | 9.359797989 |
| Vichada | 956576.6785 | 107808 | 8.872965629 |

Tabla tomada del DANE https://www.dane.gov.co/files/operaciones/PIB/departamental/anex-PIBDep-TotalDepartamento-2022pr.xlsx.

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 a mano como lo hicimos en excel. Vamos a crear 4 grupos, es decir, k=4 ( clusters).

1.7. Calcula el resultado de un dedograma utilizando la distancia maxima en un excel

2. PCA. Utilizar los datos de la tabla 1, para calcular PCA y reducir la dimensionalidad de 2 dimensiones a 1. Para este ejercicio se debe utilizar las variables X1, y X2 y crear un vector con una sola dimensión.
   2.1. Cual es la matriz de covarianza
   2.2. Cuales son los eigenvalues
   2.3. Cuál es la varianza explicada por el eigenvalue.
   2.4. Cual es el valor del eigenvector
   2.5. Cuál es la matriz proyectada.
   2.6. Cual es el error o diferencia entre la matriz proyectada

3. PCA (20%)

Cargar el data set de caras que está en la carpeta datos de la tarea 2 (ver notebook https://github.com/jdramirez/UCO_ML_AI/blob/master/src/notebook/PCA.ipynb):

1. Calcular la mean face. Que es la cara con el promedio de los pixeles y visualizarla.
2. Centrar los datos, utilizar PCA. ¿Cuántos componentes se deben utilizar para mantener el 90% de las características?. Crear una tabla para mostrar las primeras 5 caras utilizando, la mean face + los datos reconstruidos utilizando la primera componente, después con 3 componentes, después con las primeras 20 componentes, después con las componentes que explican el 95% de la varianza y por último con el numero de componentes que tiene el 99% de la varianza. ¿Qué se puede concluir de los resultados?

| Cara original | MeanFace + 1 comp | MeanFace + 3 comp | MeanFace + 20 comp | MeanFace + 95% comp | MeanFace + 99% comp |
|---|---|---|---|---|---|
| 1 | | | | | |
| 2 | | | | | |
| 3 | | | | | |
| 4 | | | | | |

4. Utilizando el dataset del proyecto data/CARS.csv crear: **Utilizar la librería de plotly**.
   - 4.1. Distribución de cada variables:
     - 4.1.1. Para las variables categóricas un gráfico de barras. Categoría numero de observaciones.
     - 4.1.2. Para las variables numéricas crear histogramas. Listar los modelos de carros que están más lejos de 5 estándares de desviación, y serían considerados outliers. Hacer test de si es una distribución normal o no.
   - 4.2. Gráfico de la relación de cada variable con respecto a MPG_City:
     - 4.2.1. Variables categóricas debes crear un boxplot. Explique cómo interpreta el gráfico
     - 4.2.2. Variables numéricas vas a crear un scatter plot.Explique cómo interpreta el gráfico
   - 4.3. Matriz de correlación.
     - 4.3.1. 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.
     - 4.3.2. Cree las dummy variables para todas las variables categóricas y genere la matriz de correlación nuevamente. ¿Cuál es el valor de variable categórica con mayor correlación?
     - 4.3.3. 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.