DMG: Assignment 1

Total Marks: 10

Instructions :

1. The assignment is to be attempted in groups of 2 students.
2. Programming Language: Python + (Excel / Tableau)

   Each task should be performed using Python as well as Excel or Tableau.
3. For Plagiarism, institute policy will be followed.
4. You need to submit the Code files in .ipynb format and excel/tableau sheets. The submission should be made in the classroom in a zip folder with the name 'A1_RollNumber1_RollNumber2.zip'.

   One member should submit on google classroom while another member can mark turn in without the attachment.
5. In case of doubts, please comment in the classroom.

This assignment is mandatory.
In this assignment, you will identify a dataset of interest and perform exploratory data analysis.

Dataset:

A sample dataset is given here, but you are free to choose any dataset that picks your interest from the below given links and from the next assignment onwards, you should be working on the same dataset.

- https://archive.ics.uci.edu/
- https://data.gov.in/

Exploratory Data Analysis (EDA): (8 marks)

1. Choose a real-world dataset from a domain of your interest. Describe the dataset's attributes, size, and source. Perform basic summary statistics on the dataset, and identify any potential challenges. **(1 marks)**
2. Create histograms, box plots, and scatter plots to visualise the distribution of numerical attributes and visually validate the observed patterns. **(2 marks)**
3. Explore relationships between pairs of attributes by computing the correlation matrix for this dataset using the formula for the cosine between centered attribute vectors. Output which attribute pairs are i) the most correlated, ii) the most anti-correlated, and iii) the least correlated. **(1 marks)**
4. Investigate any outliers or anomalies and discuss their potential impact on analysis. **(1 marks)**
5. Compute the mean vector μ for the data matrix, and then compute the total variance. **(1 marks)**
6. Calculate the sample covariance matrix, denoted as Σ, by evaluating the inner products among the attributes of the data matrix that has been centered. Subsequently, determine the sample covariance matrix by summing the outer products of the centered data points. **(2 marks)**

Dimensionality Reduction: (2 marks)

1. Apply dimensionality reduction techniques like Principal Component Analysis (PCA) or t-SNE to visualise high-dimensional data in lower dimensions.**(1 marks)**
2. Interpret the results of the dimensionality reduction and discuss whether it helped reveal any patterns or clusters. **(1 marks)**