C4 - Unsupervised Learning
PGP AI/ML
Assignment 1 - K-means clustering

Statement:

By conducting a Customer Personality Analysis, businesses can refine their products based on the preferences of specific customer segments. Rather than allocating resources to market a new product to the entire customer database, companies can identify the segments most likely to be interested in the product. Subsequently, targeted marketing efforts can be directed toward those particular segments, optimizing resource utilization and increasing the likelihood of successful product adoption.

Given is a dataset about shopping behavior and details of customers and following features are present in the dataset:

Details of Features are as below:

- **Id:** Unique identifier for each individual in the dataset.
- **Year_Birth:** The birth year of the individual.
- **Education:** The highest level of education attained by the individual.
- **Marital_Status:** The marital status of the individual.
- **Income:** The annual income of the individual.
- **Kidhome:** The number of young children in the household.
- **Teenhome:** The number of teenagers in the household.
- **Dt_Customer:** The date when the customer was first enrolled or became a part of the company's database.
- **Recency:** The number of days since the last purchase or interaction.
- **MntWines:** The amount spent on wines.
- **MntFruits:** The amount spent on fruits.
- **MntMeatProducts:** The amount spent on meat products.
- **MntFishProducts:** The amount spent on fish products.
- **MntSweetProducts:** The amount spent on sweet products.
- **MntGoldProds:** The amount spent on gold products.
- **NumDealsPurchases:** The number of purchases made with a discount or as part of a deal.
- **NumWebPurchases:** The number of purchases made through the company's website.
- **NumCatalogPurchases:** The number of purchases made through catalogs.
- **NumStorePurchases:** The number of purchases made in physical stores.
- **NumWebVisitsMonth:** The number of visits to the company's website in a month.
- **AcceptedCmp3:** Binary indicator (1 or 0) whether the individual accepted the third marketing campaign.

- **AcceptedCmp4:** Binary indicator (1 or 0) whether the individual accepted the fourth marketing campaign.
- **AcceptedCmp5:** Binary indicator (1 or 0) whether the individual accepted the fifth marketing campaign.
- **AcceptedCmp1:** Binary indicator (1 or 0) whether the individual accepted the first marketing campaign.
- **AcceptedCmp2:** Binary indicator (1 or 0) whether the individual accepted the second marketing campaign.
- **Complain:** Binary indicator (1 or 0) whether the individual has made a complaint.
- **Z_CostContact:** A constant cost associated with contacting a customer.
- **Z_Revenue:** A constant revenue associated with a successful campaign response.
- **Response:** Binary indicator (1 or 0) whether the individual responded to the marketing campaign.

Your task is to do customer segmentation on this dataset using the K-Means algorithm.
It is imperative that you be thorough with the feature engineering before clustering - dropping redundant features, meaningful feature transformations and creation, handling noise in categorical features (many string values could indicate the same basic information), dimensionality reduction for clustering, etc.

Tasks:

1. **Exploratory data analysis - Do not do feature engineering in this section** [2 marks]
   a. Null values, feature distribution plots, numerical features statistics, etc
2. **Feature engineering** - [2 marks]
   a. Take the insights and observations from section 1 and address them through feature engineering in this section
3. **Clustering** - [2+1 marks]
   a. Apply K-means on reduced dataset (after dimensionality reduction). This involves finding the best K using the elbow method. Justification is important
4. **Model fitting** - [0.5 + 0.5 mark]
   a. Cluster the data using k value found in part 3 and report number of points in each cluster
5. **Post training analysis** - [2 marks]
   a. Pertaining to the problem statement described above, do visual analysis through plots which might contain useful insights based on the cluster labels you have now. For example, Income Vs Spend scatterplot coloured by cluster number.
   b. Add your takeaways after each such plot

Submission Instructions:

Submit an HTML file (with outputs of cells printed) and your code in IPYNB format.