## **Project 8: Customer Segmentation using Data Science**

**Phase 2**

**Short Explanation:**

Customer segmentation is the process of dividing a company's target audience or customer base into distinct groups based on shared characteristics, behaviors, or demographics. This helps businesses better understand and cater to the specific needs and preferences of different customer segments, allowing for more effective marketing, product development, and customer engagement strategies. By tailoring their efforts to the unique attributes of each segment, businesses can enhance customer satisfaction, loyalty, and overall profitability.

**Dataset:**

i)We got dataset from Kaggle.

ii)The dataset link: <https://www.kaggle.com/datasets/akram24/mall-customers>.

iii)It is a mall customer dataset. It contains 5 columns and 200 rows. It have details

about 200

iv)customers, they are customer ID,genre(gender),age,Annual income and their

spending score.

**Columns:**

1.Customer ID: A unique identifier for each customer.

2.Age: Age of the customer.

3.Gender: Customer's gender (male, female, other).

4.Income: Customer's annual income.

5.Spending score: Spending score, also known as customer spending score or

purchase score, is a numerical value that represents a customer's spending behavior

or propensity to make purchases. It is commonly used in customer segmentation as

one of the key variables to group customers based on their purchasing patterns and

preferences.

**Libraries to be used:**

1.Numpy.

2.Pandas.

3.Matplotlib.

4.Seaborn.

5.sklearn.cluster import kmeans.

**Way to download libraries:**

To perform customer segmentation, we typically use data analysis and machine

learning libraries in programming languages like Python or R. Here's how you can

download and install the necessary libraries for customer segmentation using Python:

**1.Install Python:** If you don't already have Python installed, you can download it from the official Python website (<https://www.python.org/downloads/>) and follow the installation instructions for your operating system

**2.Package Management:** Python has package managers that make it easy to install and manage libraries. The two most commonly used package managers are pip and conda.

* + **pip (for Python packages):**

1.Python 2: pip is typically included with Python 2.

2.Python 3: pip may need to be installed separately. You can install it by running this command in your terminal or command prompt:

python -m ensurepip --default-pip

* + p

**3.Install Data Science Libraries:**

You'll need several data science and machine learning libraries for customer segmentation. You can install these libraries using pip or conda. Open a terminal or command prompt and use the following commands:

pip install numpy pandas scikit-learn matplotlib seaborn

Here's a brief explanation of what these libraries are for:

* **numpy:** Numerical computing library, often used for data manipulation.
* **pandas:** Data manipulation and analysis library, great for working with tabular data.
* **scikit-learn:** Machine learning library that includes various algorithms for clustering and segmentation.
* **matplotlib:** Data visualization library.
* **seaborn:** Data visualization library built on top of matplotlib

**How to train and test data:**

1. **Data Preparation:**
   1. Collect and clean the customer data, ensuring it's in a format suitable for analysis.
   2. Handle missing values, outliers, and data inconsistencies.
   3. Normalize or standardize numeric features if necessary.
   4. Encode categorical variables if needed (e.g., one-hot encoding).
2. **Data Splitting:**
   1. Split your dataset into two subsets: one for training and one for testing.
   2. Common splitting ratios are 70-80% for training and 20-30% for testing. Adjust as needed based on our dataset size.
3. **Feature Selection or Engineering:**
   1. Choose relevant features for customer segmentation. This step may involve domain knowledge or feature selection techniques.
   2. Create new features if they could enhance segmentation (e.g., calculating RFM scores).
4. **Model Selection:**
   1. Decide on the segmentation algorithm(s) you want to use. We are going to use k-means clustering.
   2. Choose appropriate hyper parameters for the chosen algorithm(s).
5. **Training:**
   1. Use the training dataset to fit your segmentation model(s). We're using k-means clustering, we'd fit the k-means algorithm to our training data.
6. **Testing:**
   1. Apply the trained model to the testing dataset to segment the customers in the test set.
   2. Evaluate the performance of our segmentation model. Common evaluation metrics include silhouette score, Davies-Bouldin index, or domain-specific metrics.
7. **Visualization and Interpretation:**
   1. Visualize the segmented customer groups to gain insights.
   2. Interpret the characteristics of each segment and give them meaningful labels (e.g., "High-Value Customers," "Low-Activity Customers").
   3. We use scatter plots.
8. **Application:**
   1. Apply the customer segments to your business strategies, such as targeted marketing, product recommendations, or customer retention efforts.

* Here's a simplified Python example using scikit-learn for k-means clustering:

# Load and pre-process your customer data

from sklearn.cluster import KMeans  
 from sklearn.model\_selection import train\_test\_split  
 import pandas as pd  
  
 # Load and pre-process your customer data  
 data = pd.read\_csv('customer\_data.csv')  
 # Data cleaning, feature selection, and preprocessing go here...  
  
 # Split the data into training and testing sets  
 X\_train, X\_test = train\_test\_split(data, test\_size=0.2, random\_state=42)  
  
 # Choose the number of clusters (k) for k-means  
 k = 3  
  
 # Initialize and train the k-means model on the training data  
 kmeans = KMeans(n\_clusters=k, random\_state=42)  
 kmeans.fit(X\_train)  
  
 # Apply the trained model to the testing data to obtain segment labels  
 test\_segment\_labels = kmeans.predict(X\_test)  
  
 # Evaluate the performance of your segmentation (if applicable)  
 # Visualize and interpret the results  
  
 # Apply the segments to your business strategies

* Remember that customer segmentation is an iterative process, and you may need to fine-tune your model, adjust features, or try different algorithms to achieve the best results for your specific business goals.

**Accuracy check:**

Accuracy in customer segmentation is typically assessed using metrics like silhouette score, Davies-Bouldin index, or other clustering evaluation metrics. Scatter plots themselves are not used to directly measure accuracy, but they can be a valuable tool for visualizing and interpreting our segmentation results.

