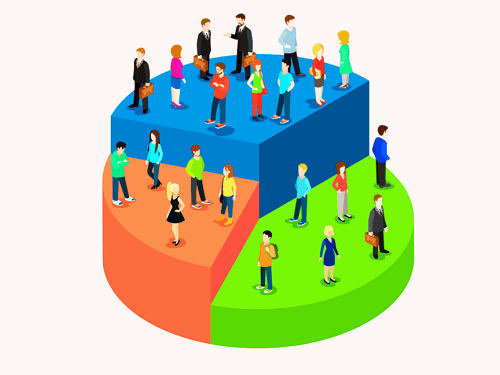
IBM-APPLIED DATA SCIENCE

GROUP-3

Phase 3: Development Part 1

PROJECT: Customer Segmentation Using Data Science



**INTRODUCTION:**

### In Phase 3, "Development Part 1," you will start building your project by loading and preprocessing the dataset. Here are the steps you should follow with examples for each step:

**Step 1: Data Collection:**

**Data Source**: Identify the source of your customer data. It can come from various sources, such as CRM systems, sales databases, or online surveys.

**Example:** Let us say you are working with a fictional e-commerce company, and your customer data is stored in a SQL database.

**Python Code:**

import pandas as pd

import sqlite3

**# Connect to the database and retrieve the data**

conn = sqlite3.connect('customer\_data.db')

query = 'SELECT \* FROM customers'

df = pd.read\_sql\_query(query, conn)

conn.close()

**Step 2: Data Cleaning:**

**Handle Missing Values:** Identify and handle missing data points.

**Example**: If there are missing values in the "age" column, you can impute them with the median age.

**Python Code:**

median\_age = df['age'].median()

df['age'].fillna(median\_age, inplace=True)

**Remove Duplicates:** Check for and remove duplicate entries if necessary.

**Example:** Remove duplicates based on the 'customer\_id' column.

**Python Code:**

df = df.drop\_duplicates(subset='customer\_id')

**Step 3: Feature Engineering:**

- **Create Relevant Features:** Create new features that can be useful for segmentation, such as customer lifetime value (CLV), purchase frequency, and recency.

**Example**: Calculate CLV by summing the total purchase amounts for each customer over a specified period.

**Python Code:**

df['CLV'] = df.groupby('customer\_id')['purchase\_amount'].transform('sum')

```

**Step 4: Data Transformation:**

**Standardize or Normalize**: Depending on your analysis method (e.g., k-means clustering), you may need to standardize or normalize your features to have similar scales.

**Example**: Use the `StandardScaler` to standardize your features.

**Python Code:**

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

df[['age', 'CLV']] = scaler.fit\_transform(df[['age', 'CLV']])

```

**Step 5: Exploratory Data Analysis (EDA):**

**Data Visualization**: Perform data visualization to gain insights into the data distribution and relationships between variables.

**Example**: Create histograms and scatter plots to visualize the distribution of age and CLV.

**Python Code:**

import matplotlib.pyplot as plt

plt.hist(df['age'], bins=20)

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.title('Age Distribution')

plt.show()

```

**Step 6: Save the Processed Data:**

- Save the preprocessed data for future analysis.

**Example:** Save the cleaned and transformed dataset to a CSV file.

**Python Code:**

df.to\_csv('preprocessed\_customer\_data.csv', index=False)

Summary:

### This is just the initial part of developing a customer segmentation project. The next steps typically involve selecting the appropriate segmentation technique (e.g., k-means clustering, hierarchical clustering), determining the optimal number of segments, and performing the actual segmentation. The insights gained from this analysis can then be used to tailor marketing strategies, enhance customer experience, and make data-driven business decisions.