Customer Segmentation Using Data Science

PHASE 4: DEVELOPMENT PART-2

Introduction

This document provides a step-by-step walkthrough of a customer segmentation project using Python, pandas, scikit-learn, and Matplotlib. The goal of this project is to cluster customers based on their Annual Income and Spending Score to gain insights into customer behavior and preferences.

Dataset

The dataset used in this project is named 'Mall_Customer.csv' and contains the following columns:

1) CustomerID: Unique identifier for each customer
2) Genre: Gender of the customer (Male or Female)
3) Age: Age of the customer
4) Annual Income (k$): The annual income of the customer in thousands of dollars
5) Spending Score (1-100): A score representing how much a customer spends, with 1 being the lowest and 100 being the highest.

Project Implementation

1. Data Loading

The first step is to load the dataset into a pandas DataFrame.

Python

```
import pandas as pd

data = pd.read_csv('Mall_Customers.csv')
```

2. Data Exploration

Explore the dataset by displaying the first few rows to understand its structure.

Python

```
print(data.head())
```

3. Feature Selection

For this customer segmentation project, we'll use the 'Annual Income' and 'Spending Score' columns as the relevant features for clustering.

Python

```
X = data[['Annual Income (k$)', 'Spending Score (1-100)']]
```

4. Data Preprocessing

Standardize the features to ensure that both 'Annual Income' and 'Spending Score' have the same scale.

Python

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

5. Determine the Optimal Number of Clusters

To determine the optimal number of clusters for K-Means, we use the Elbow Method. This involves running K-Means with different numbers of clusters and calculating the Within-Cluster Sum of Squares (WCSS) for each.

Python

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

wcss = []  # Within-Cluster Sum of Squares
for i in range(1, 11):
  kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10,
random_state=0)
  kmeans.fit(X_scaled)
  wcss.append(kmeans.inertia_)
```

```
# Plot the Elbow Method
plt.plot(range(1, 11), wcss)
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```

Based on the Elbow Method, an appropriate number of clusters is selected (e.g., 5).

6. Cluster Customers

Apply K-Means clustering with the selected number of clusters.

Python

```
num_clusters = 5

kmeans = KMeans(n_clusters=num_clusters, init='k-means++', max_iter=300,
n_init=10, random_state=0)
y_kmeans = kmeans.fit_predict(X_scaled)
```

7. Add Cluster Labels to the Dataset

Add the cluster labels to the original dataset.

Python

```
data['Cluster'] = y_kmeans
```

8. Visualize the Clusters

Visualize the clusters by creating a scatter plot of 'Annual Income' and 'Spending Score'.

Python

```
plt.scatter(X_scaled[y_kmeans == 0, 0], X_scaled[y_kmeans == 0, 1], s=100,
c='red', label='Cluster 1')
plt.scatter(X_scaled[y_kmeans == 1, 0], X_scaled[y_kmeans == 1, 1], s=100,
c='blue', label 'Cluster 2')
# Repeat for other clusters...
```

```
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=300,
c='yellow', label='Centroids')
plt.title('Customer Segmentation')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```

9. Analyze Each Cluster

Examine each cluster's statistics to understand customer segments.

Python

```
for cluster_num in range(num_clusters):
  cluster_data = data[data['Cluster'] == cluster_num]
  print(f'Cluster {cluster_num} Statistics:')
  print(cluster_data.describe())
```

10. Save the Clustered Dataset

You can save or export the clustered dataset for further analysis or marketing strategies.

Python

```
data.to_csv('Mall_Customer.csv', index=False)
```

Conclusion

In this project, we successfully segmented customers into different clusters based on their annual income and spending scores. This information can be used for targeted marketing and improving customer services. Further analysis and strategies can be developed based on the insights gained from this customer segmentation.