CUSTOMER SEGMENTATION FOR RETAIL STORE

Project Manager: Abhimanyu Kumar

Submission to: CipherSchools

Start Date: 13/07/2024

Completion Date: 17/07/2024

Introduction

Objective and Use Case

Objective: The primary objective of this project is to analyze customer data from a retail store and segment customers into distinct groups based on their purchasing behavior. By identifying these segments, the retail store can tailor its marketing strategies to better meet the needs of each customer group, ultimately enhancing customer satisfaction and boosting sales.

Use Case: Customer segmentation is a crucial aspect of customer relationship management (CRM) and marketing strategies. In the context of a retail store, understanding different customer segments allows the store to:

- **Develop Targeted Marketing Campaigns:** Tailor promotions and advertisements to specific customer groups based on their purchasing habits and preferences.
- **Personalize Customer Experiences:** Offer personalized recommendations and services to improve customer satisfaction and loyalty.
- **Optimize Product Offerings:** Adjust inventory and product offerings to align with the preferences of different customer segments.
- **Increase Customer Retention:** Implement strategies to retain high-value customers and reduce churn rates.
- **Enhance Sales and Revenue:** Identify opportunities for cross-selling and up-selling to maximize sales and revenue.

By leveraging customer segmentation, the retail store can implement more effective marketing strategies, improve operational efficiency, and ultimately achieve a competitive advantage in the market.

Overview of the Dataset

Dataset Description: The dataset used in this project is the "Mall Customers" dataset, which provides information about customers from a mall. The dataset contains demographic and behavioral attributes of the customers, which can be used to perform segmentation. The dataset includes the following columns:

1. **CustomerID:** Unique identifier for each customer.
2. **Gender:** Gender of the customer (Male/Female).
3. **Age:** Age of the customer.
4. **Annual Income (k$):** Annual income of the customer in thousands of dollars.
5. **Spending Score (1-100):** Spending score assigned by the mall based on customer behavior and spending nature (1 being lowest and 100 being highest).

Dataset Snapshot:

CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79

Attributes:

- **CustomerID:** A numerical identifier unique to each customer.
- **Gender:** Categorical variable indicating the customer's gender.
- **Age:** Numerical variable indicating the customer's age.
- **Annual Income (k$):** Numerical variable indicating the customer's annual income in thousands of dollars.
- **Spending Score (1-100):** Numerical variable indicating the spending score, a metric assigned by the mall based on customer spending behavior.

Purpose of the Dataset: The dataset is used to perform customer segmentation analysis. By examining the demographic and behavioral attributes of the customers, we aim to identify distinct groups of customers who exhibit similar purchasing behaviors. These insights will enable the retail store to develop targeted marketing strategies and enhance overall customer satisfaction.

Data Source: The dataset is publicly available and can be downloaded from Kaggle at the following link: Mall Customers Dataset on Kaggle.

Data Analysis and Segmentation: In this project, we will:

1. **Clean the data** to handle missing values and ensure consistency.
2. **Perform Exploratory Data Analysis (EDA)** to understand the distribution and relationships within the data.
3. **Apply K-Means Clustering** to segment customers into distinct groups.
4. **Visualize the results** using Matplotlib and Power BI to gain actionable insights.

Data Collection:

The dataset was imported using the Pandas library in Python. Here's a brief overview of the dataset structure:

```python
# Importing the Data
import pandas as pd
Data=pd.read_csv("Mall_Customers.csv")
print(Data)
```

Data Cleaning:

The dataset was remarkably clean, with no missing values. Therefore, imputation was unnecessary. We confirmed the datatypes were appropriate for our analysis and no transformations were required for this specific dataset.

```python
import pandas as pd
Data=pd.read_csv("Mall_Customers.csv")
# renaming the Genre to Gender
Data.rename(columns={"Genre":"Gender"},inplace=True)
# print(Data)
# checking the missing values
missing_values=Data.isnull().sum()
print(missing_values)
```

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

Handling the outliers

```
import matplotlib.pyplot as plt
plt.figure(figsize=(14,7))
# plotting a box plot for the Age and Annual Income (k$) coloumn
plt.subplot(1,2,1)
plt.boxplot(Data["Age"])
plt.title("Age Outliers using boxplot")

plt.subplot(1,2,2)
plt.boxplot(Data["Annual Income (k$)"])
plt.title("Income Outliers using boxplot")
plt.show()
```

```
# now capping the outliers using IQR method
for col in ["Age","Annual Income (k$)"]:
    # Calculate IQR
    Q1 = Data[col].quantile(0.25)
    Q3 = Data[col].quantile(0.75)
    IQR = Q3 - Q1
    IQR=Q3-Q1
    lower_bound=Q1-(1.5*IQR)
    upper_bound=Q3+(1.5*IQR)
    Data[col]=Data[col].apply(lambda x: upper_bound if(x>upper_bound) else(lower_bound if(x<lower_bound) else x) )
# print(Data)
```

Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process where an analyst explores the data to understand its underlying structure, extract important variables, detect outliers and anomalies, and test underlying assumptions.

```
# Summary
Summary=Data.describe()
# print(Summary)

#visualization using Matplotlib
# ploting a bar_graph of income_by_gender
income_by_gender=Data.groupby("Gender")["Annual Income (k$)"].mean()
# reindexing as Male,Female
income_by_gender=income_by_gender.reindex(["Male","Female"])

income_by_gender.plot(kind="bar",color=["green","pink"])
plt.xlabel("Gender")
plt.ylabel("Average Income")
plt.title("Average Income by Gender")
plt.show()
```

```
# creating a histogram of Spending by Age below 40 and above 40
Data_below_40=Data[Data["Age"]<40]
Data_above_40=Data[Data["Age"]>=40]
plt.figure(figsize=(12, 6))
# Histogram for ages below 40
plt.subplot(1, 2, 1)
plt.hist(Data_below_40['Spending Score (1-100)'], bins=5, color='blue', alpha=0.7)
plt.title('Spending Score of people having Ages Below 40')
plt.xlabel('Spending Score')
plt.ylabel('Frequency')
# Histogram for ages 40 and above
plt.subplot(1, 2, 2)
plt.hist(Data_above_40['Spending Score (1-100)'], bins=5, color='green')
plt.title('Spending Score of people having Ages 40 and Above')
plt.xlabel('Spending Score')
plt.ylabel('Frequency')
plt.show()
```

Customer Segmentation:

Customer segmentation is the process of dividing a customer base into distinct groups that share similar characteristics. This allows businesses to tailor their marketing strategies, products, and services to better meet the needs of each segment.

```
### Feature Selection ###
#Variance Thresholding
# removing the columns which have less variance
df=Data.copy()
# changing male to 0 and female to 1
df['Gender'] = df['Gender'].map({'Male': 0, 'Female': 1})
from sklearn.feature_selection import VarianceThreshold
selector=VarianceThreshold(0.1)
Data_Vfiltered=pd.DataFrame(selector.fit_transform(df),columns=df.columns[selector.get_support()])
if(Data_Vfiltered.shape==df.shape):
    print("No columns having less variance")
else:
    print("the data have less variance features")
```

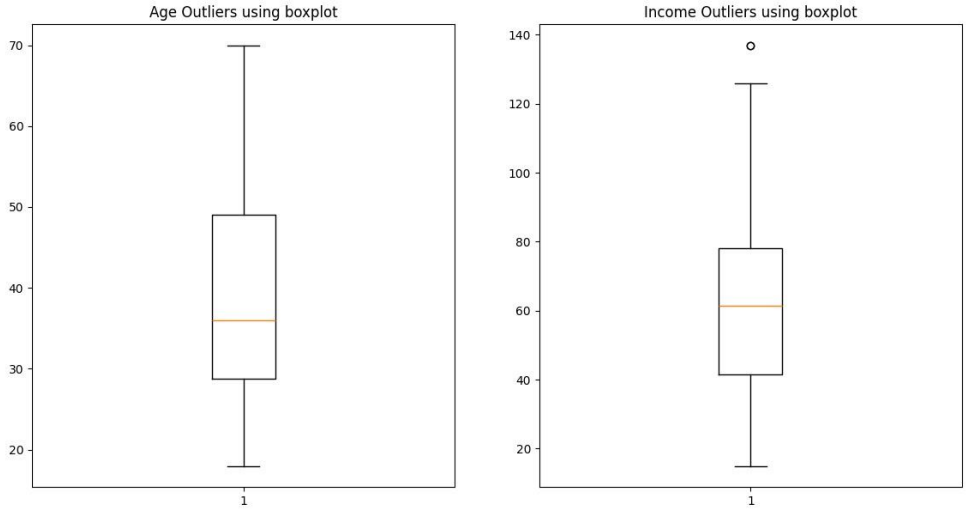
```
### K-Means Clustering ###
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

#selecting features for the clustering
features=Data[["Age","Annual Income (k$)",'Spending Score (1-100)']]

#standarizing the features
scaler=StandardScaler()
scaled_features=scaler.fit_transform(features)
# print(scaled_features)
```

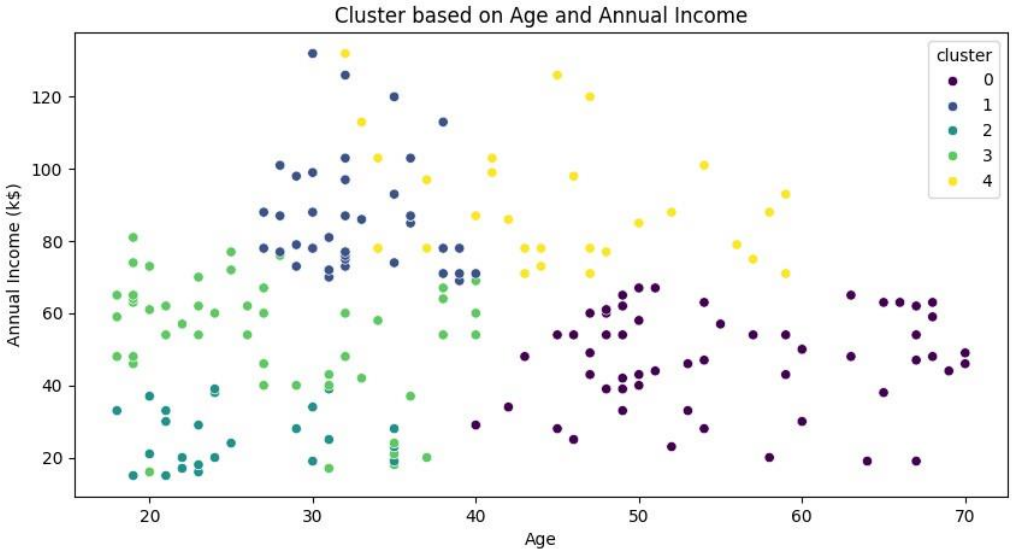
```
# Elbow method to find optimal k
inertia=[]
for i in range(1,11):
    kmeans=KMeans(n_clusters=i,random_state=42)
    kmeans.fit(scaled_features)
    inertia.append(kmeans.inertia_)
# plot the elbow curve
plt.figure(figsize=(10,5))
plt.plot(range(1,11),inertia,marker="o")
plt.title('Elbow Method for Optimal k')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
# plt.xticks(range(1, 11))
plt.show()
```

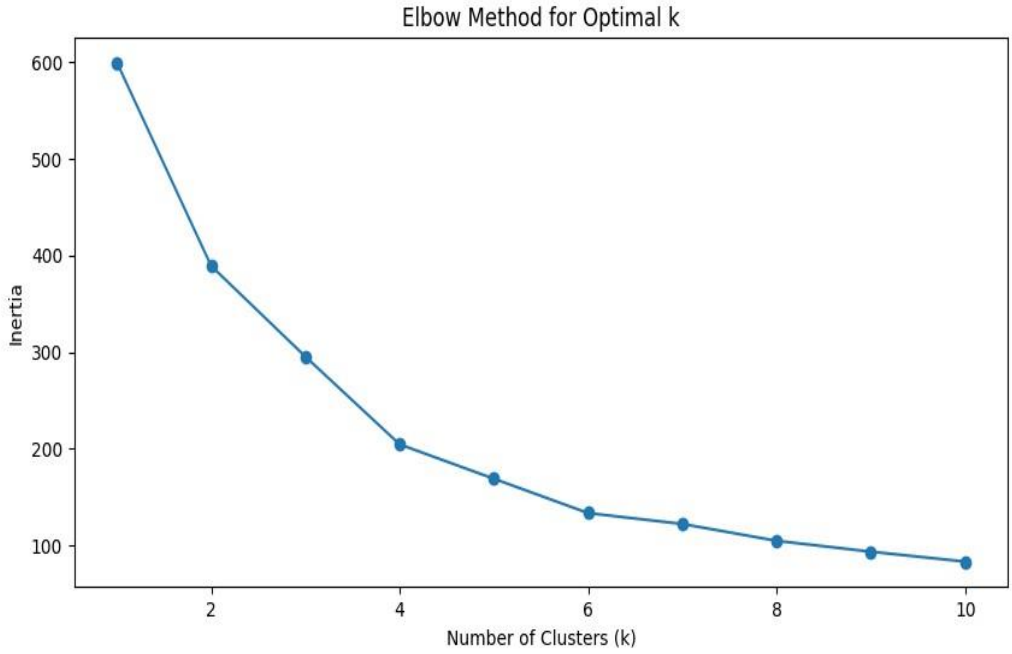
Visualization with python:

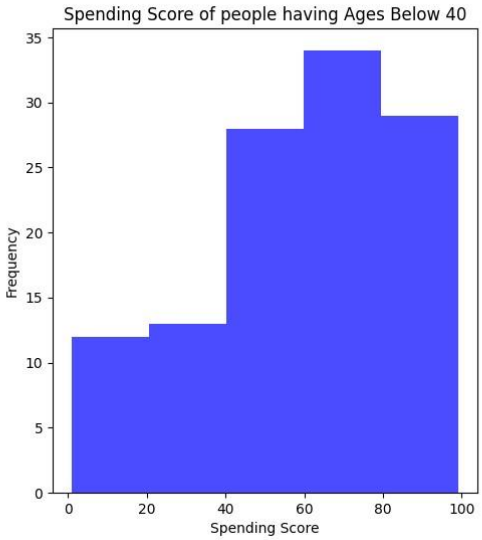
Visualizing data effectively is a crucial part of customer segmentation. In Python, libraries like Matplotlib, Seaborn, and Plotly can be used to create insightful visualizations. Here's a step-bystep guide on how to visualize customer segmentation using Python:

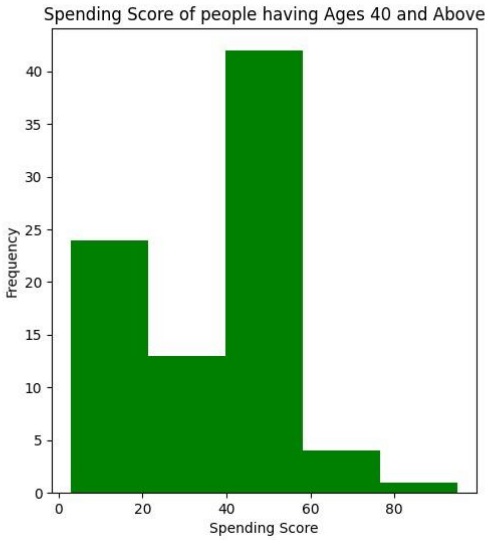
```python
import matplotlib.pyplot as plt
plt.figure(figsize=(14,7))
# plotting a box plot for the Age and Annual Income (k$) coloumn
plt.subplot(1,2,1)
plt.boxplot(Data["Age"])
plt.title("Age Outliers using boxplot")

plt.subplot(1,2,2)
plt.boxplot(Data["Annual Income (k$)"])
plt.title("Income Outliers using boxplot")
plt.show()
```

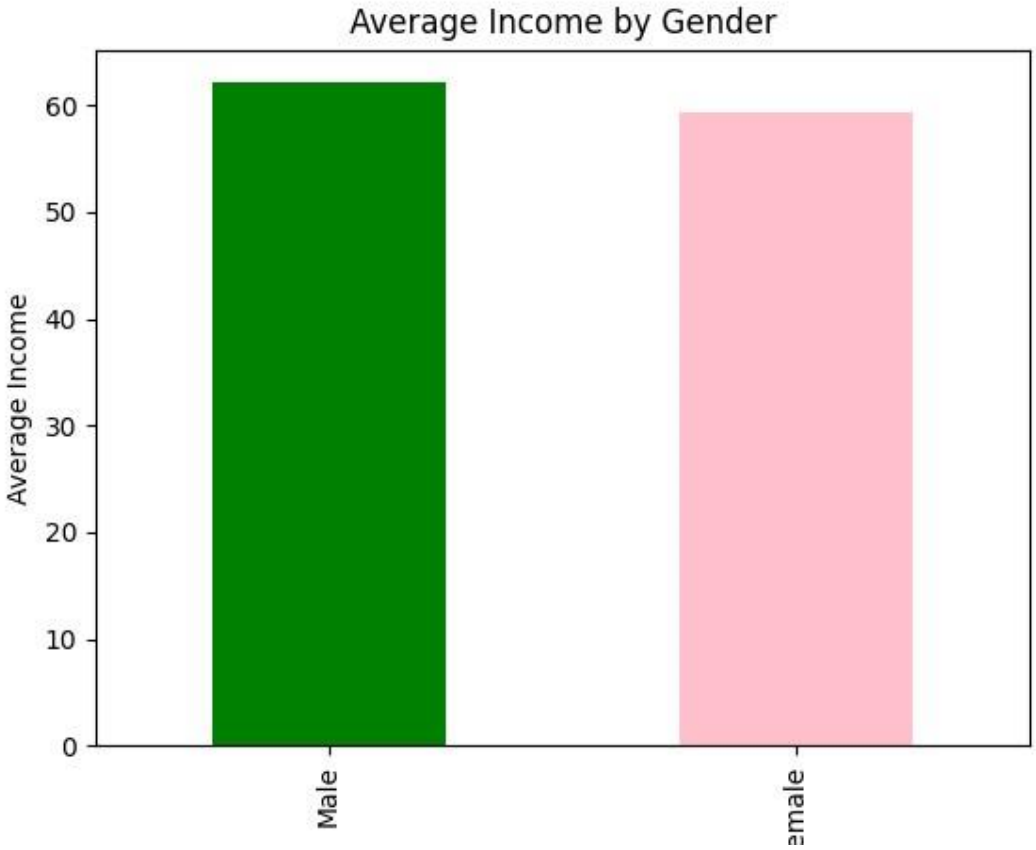



Cluster based on Age and Annual Income
cluster
0
1
2
3
4
Annual Income (k$)
120
100
80
60
40
20
20
30
40
50
60
70
Age


Elbow Method for Optimal k
Inertia
600
500
400
300
200
100
2
4
6
8
10
Number of Clusters (k)



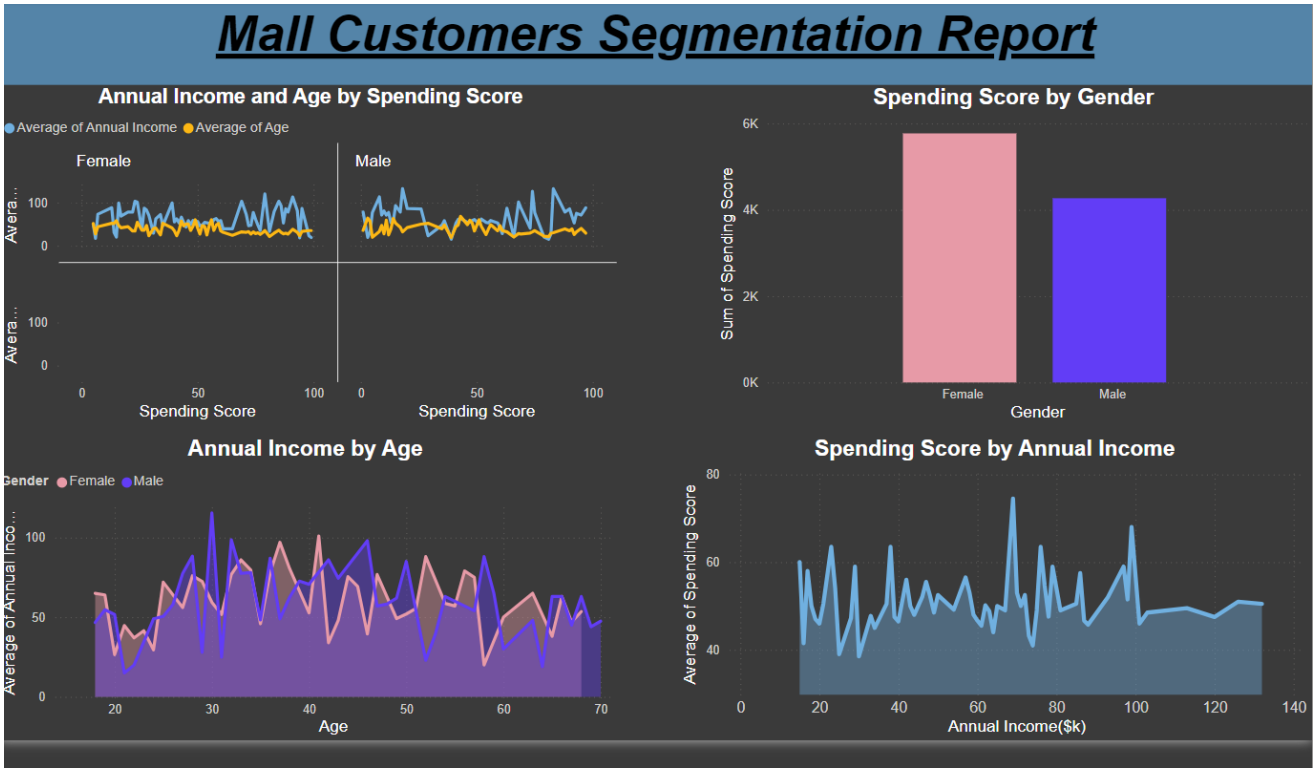




Visualization with Power BI:

Power BI is a business analytics tool that allows users to visualize and share insights from their data. By connecting to various data sources, transforming and modeling data, and using a range of visualizations like charts and maps, users can create interactive reports and dashboards to support informed decision-making.

The cleaned dataset was imported into Power BI. We then created interactive dashboards to showcase the customer segments and their characteristics.



Conclusion of the Project

Objective and Use Case:

The objective of this project was to segment customers based on their purchasing behavior using a dataset from a mall. This segmentation helps in understanding customer groups better, which can guide targeted marketing strategies, personalized services, and better business decisions.

Data Collection and Cleaning:

The dataset was imported and cleaned to ensure accuracy and consistency: •

The 'Genre' column was renamed to 'Gender'.

- Missing values were checked, and none were found, allowing us to proceed without data imputation.
- Outliers were identified using box plots, particularly in the 'Annual Income (k$)' column.

Exploratory Data Analysis (EDA):

EDA involved descriptive statistics and visualization to understand the distribution and relationships within the data:

- Histograms and scatter plots revealed the distribution of key variables such as Age, Annual Income, and Spending Score.
- Insights from these visualizations helped in understanding the data structure and preparing it for clustering.

Customer Segmentation:

Using the K-Means clustering algorithm, customers were segmented into distinct groups:

- The optimal number of clusters was determined using the Elbow Method.
- The K-Means algorithm was applied to segment the customers into 5 clusters.
- Each cluster was analyzed to understand its characteristics, such as age, income, and spending behavior.

Visualization:

Visualizations were created to illustrate the customer segments:

- Scatter plots and pair plots showed the distribution of clusters.
- Centroid visualization helped in understanding the central tendencies of each cluster.
- Optional 3D plots provided interactive visual insights.

Insights and Recommendations:

- **Cluster 0**: Comprised of young customers with high income and high spending scores. Marketing strategies could focus on premium products and personalized experiences for this segment.
- **Cluster 1**: Included older customers with moderate income and spending scores. This group might prefer value-for-money products and services.
- **Cluster 2**: Consisted of customers with lower income and spending scores. Budgetfriendly options and discounts would appeal to this segment.
- **Cluster 3**: Contained young customers with moderate income and high spending scores. Engaging marketing campaigns targeting tech-savvy individuals would be effective.

- **Cluster 4**: Comprised of older customers with high income and moderate spending scores. Personalized services and premium product offerings could attract this segment.

Conclusion:

This project successfully segmented mall customers into distinct groups using K-Means clustering. The insights gained from the analysis can guide targeted marketing strategies and personalized services, ultimately improving customer satisfaction and business performance. Future work could involve incorporating additional features, such as purchase history or online behavior, to refine the segmentation further.