Customer Segmentation Using K-Means Clustering

Name: G.V. Savita Bhargavi

Date: 2025-09-08

Role: Data Science Student / Analyst

1. Introduction

Every business wants to understand its customers better. Not all customers are the same – some spend a lot, some are cautious, and some fall somewhere in the middle. By grouping customers with similar behavior, businesses can create targeted marketing strategies, improve sales, and provide better experiences.

In this project, we use the Mall Customers dataset, which contains details about customers' Age, Annual Income, and Spending Score. The goal is to segment the customers into groups that help the mall understand who their high spenders, cautious spenders, and average customers are.

2. Objective

1. Segment customers based on their spending habits, income, and age.
2. Identify key customer groups for targeted marketing.
3. Provide actionable business insights for improving sales and customer engagement.

3. Dataset Overview

Column Name	Description
CustomerID	Unique ID of the customer
Gender	Male/Female
Age	Age of the customer
Annual Income	Annual income in thousands of dollars
Spending Score	Score assigned by the mall based on spending habits (1–100)

For clustering, we focus on Age, Annual Income, and Spending Score.

4. Methodology

Step 1: Data Preprocessing
- Removed unnecessary columns (CustomerID, Gender) to focus on clustering features.
- Checked for missing values → none found.
- Standardized features using StandardScaler to normalize Age, Income, and Spending Score.

Step 2: Choosing Number of Clusters
- Used Elbow Method to determine optimal k (number of clusters).
- Plotted inertia (WCSS) for k = 1 to 10.
- Optimal clusters found at k = 5.

Step 3: Applying K-Means
- Initialized K-Means with n_clusters = 5.
- Assigned each customer to a cluster (0 to 4).
- Added cluster labels to the dataset.

Step 4: Visualization
- Scatter plot of Annual Income vs Spending Score, colored by cluster.

5. Python Code & Implementation

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

# Load the customer data
customer_df = pd.read_csv(r"C:\Users\HP\OneDrive\task4\Mall_Customers.csv")

# Display first few rows and dataset info
print(customer_df.head())
print(customer_df.info())
print(customer_df.isnull().sum())

# Select features for clustering
features = ['Age', 'Annual Income (k$)', 'Spending Score (1-100)']
X = customer_df[features]

# Standardize the features
scaler = StandardScaler()
```

```
X_scaled = scaler.fit_transform(X)

# Elbow Method to find optimal number of clusters
inertia = []
for k in range(1, 11):
   kmeans = KMeans(n_clusters=k, random_state=42)
   kmeans.fit(X_scaled)
   inertia.append(kmeans.inertia_)

plt.figure(figsize=(8,6))
plt.plot(range(1, 11), inertia, marker='o')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.show()

# Train K-Means with optimal clusters
k = 5
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X_scaled)

# Assign clusters
customer_df['Cluster'] = kmeans.labels_

# Cluster profiling
cluster_profile = customer_df.groupby('Cluster')[features].mean()
print(cluster_profile)

# Scatter plot of clusters (Income vs Spending Score)
plt.figure(figsize=(12,8))
sns.scatterplot(x='Annual Income (k$)', y='Spending Score (1-100)',
            hue='Cluster', data=customer_df, palette='viridis', s=100)
plt.title('Customer Segmentation based on Income and Spending Score')
plt.show()

# Save results to CSV
customer_df.to_csv("customers_with_clusters.csv", index=False)
cluster_profile.to_csv("cluster_profile.csv")
```

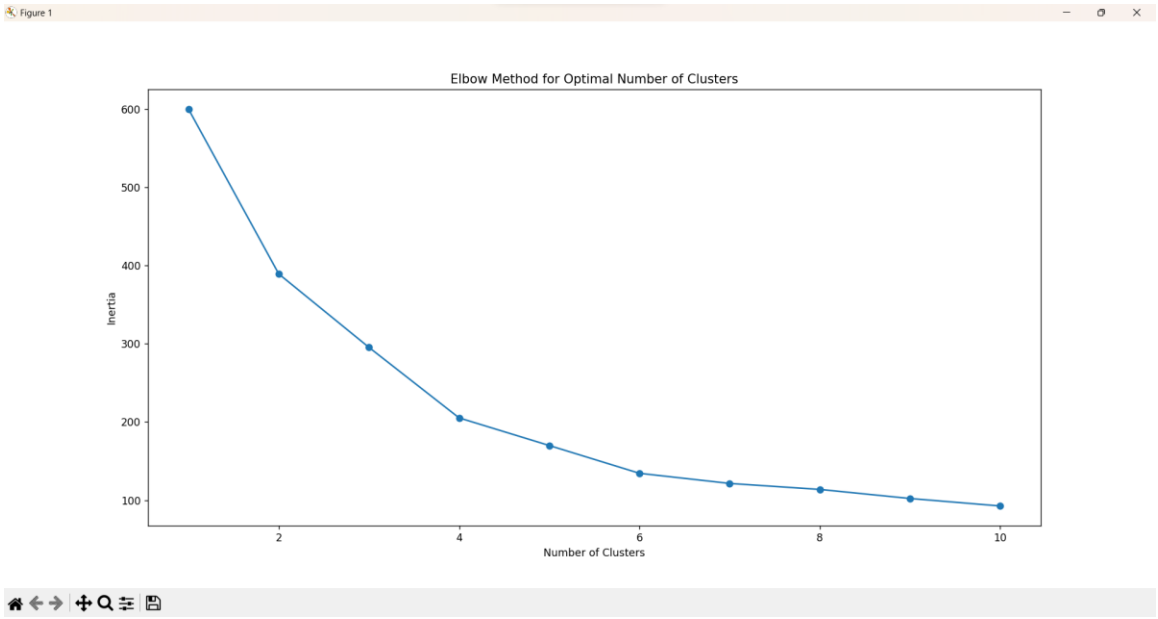
6. Results & Observations

Cluster Summary

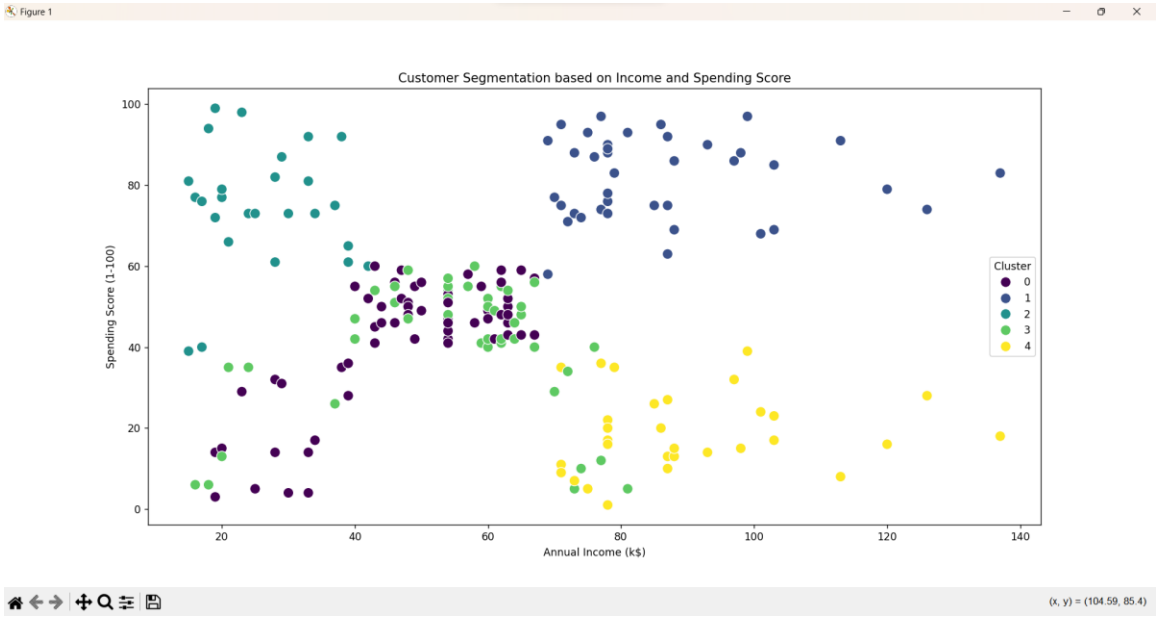
| Cluster | Characteristics | Business Insight |

0 (Purple)	Medium income & medium spending	Average customers → maintain with regular deals
1 (Dark Blue)	High income & high spending	VIPs / Premium customers → reward loyalty, exclusive offers
2 (Teal)	Low income & high spending	Young, enthusiastic spenders → target with trendy marketing & discounts
3 (Green)	Medium income & low spending	Careful spenders → low priority, awareness campaigns
4 (Yellow)	High income & low spending	Wealthy but cautious → upsell luxury/premium products

Key Observations:
1. VIPs (Cluster 1) spend a lot and are very valuable → focus on loyalty programs and exclusive perks.
2. Young spenders (Cluster 2) may be students or trend-seekers → target with discounts and new products.
3. Cautious wealthy (Cluster 4) have money but spend little → opportunity for upselling premium products.
4. Average customers (Cluster 0) are stable → maintain engagement with regular offers.
5. Low spenders (Cluster 3) are difficult to convert → use awareness campaigns but not a priority.



Visualization Placeholder



7. Conclusion

This project shows how K-Means clustering can segment customers into actionable groups, providing valuable business insights:

- High-value customers can be retained through loyalty programs.
- Potential spenders can be engaged through targeted campaigns.
- Average customers should be maintained with regular deals.
- Careful spenders are a lower priority but can be influenced with awareness campaigns.

Overall, customer segmentation helps businesses understand behavior, improve marketing, and optimize sales strategies.

8. data set used:

https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?resource=download