Customer Segmentation

Internship Task Submission

Submitted by:

Alisha Rafique

Submitted to:

Elevvo.tech

Date:

05th September, 2025

1. Introduction

Customer segmentation is a fundamental marketing strategy that involves partitioning a customer base into groups of individuals that are similar in specific ways, such as demographics, purchasing behavior, or income. This enables businesses to target specific groups with tailored strategies, thereby improving marketing efficiency, customer retention, and overall profitability.

This project utilizes the Mall Customers Dataset from Kaggle to perform customer segmentation based on two key attributes: **Annual Income** and **Spending Score**. The core objective is to apply unsupervised machine learning techniques, specifically clustering algorithms, to identify distinct customer segments and derive actionable business insights.

2. Dataset and Methodology

2.1 Dataset

The Mall Customer Dataset contains information about 200 customers, with the following features:

- CustomerID: Unique identifier for each customer.
- Gender: Gender of the customer.
- Age: Age of the customer.
- Annual Income (k$): Annual income of the customer in thousands of dollars.
- Spending Score (1-100): A score assigned by the mall based on customer behavior and spending nature.

For this clustering task, the features Annual Income and Spending Score were selected.

2.2 Methodology

The project was executed in the following stages:

- **Data Preprocessing & Exploration:** The dataset was cleaned and explored to understand the distributions of the key features.
- **Feature Scaling:** The data was standardized using StandardScaler to ensure both features contributed equally to the distance calculations in the clustering algorithm.
- **Determining Optimal Clusters:** The Elbow Method was used to identify the optimal number of clusters (k) for the K-Means algorithm.
- **Model Implementation:** The K-Means algorithm was applied with the optimal k.
- **Bonus - Alternative Algorithm:** The DBSCAN algorithm was implemented for comparison.
- **Cluster Analysis & Visualization:** The resulting clusters were analyzed and visualized using 2D scatter plots.
- **Insight Generation:** Each cluster was profiled based on its average income and spending to generate business recommendations.

2.3 Tools & Libraries

To successfully implement the project, the following tools and libraries were used:

- Programming Language: Python
- Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- Environment: Jupyter Notebook

3. Analysis and Results

3.1. Data Preprocessing and Visual Exploration

The initial data exploration involved visualizing the distribution of the data before clustering. The scatter plot below shows the raw relationship between Annual Income and Spending Score, revealing natural groupings that the clustering algorithm can uncover. Figure 1 shows the initial scatter plot and distribution of the data

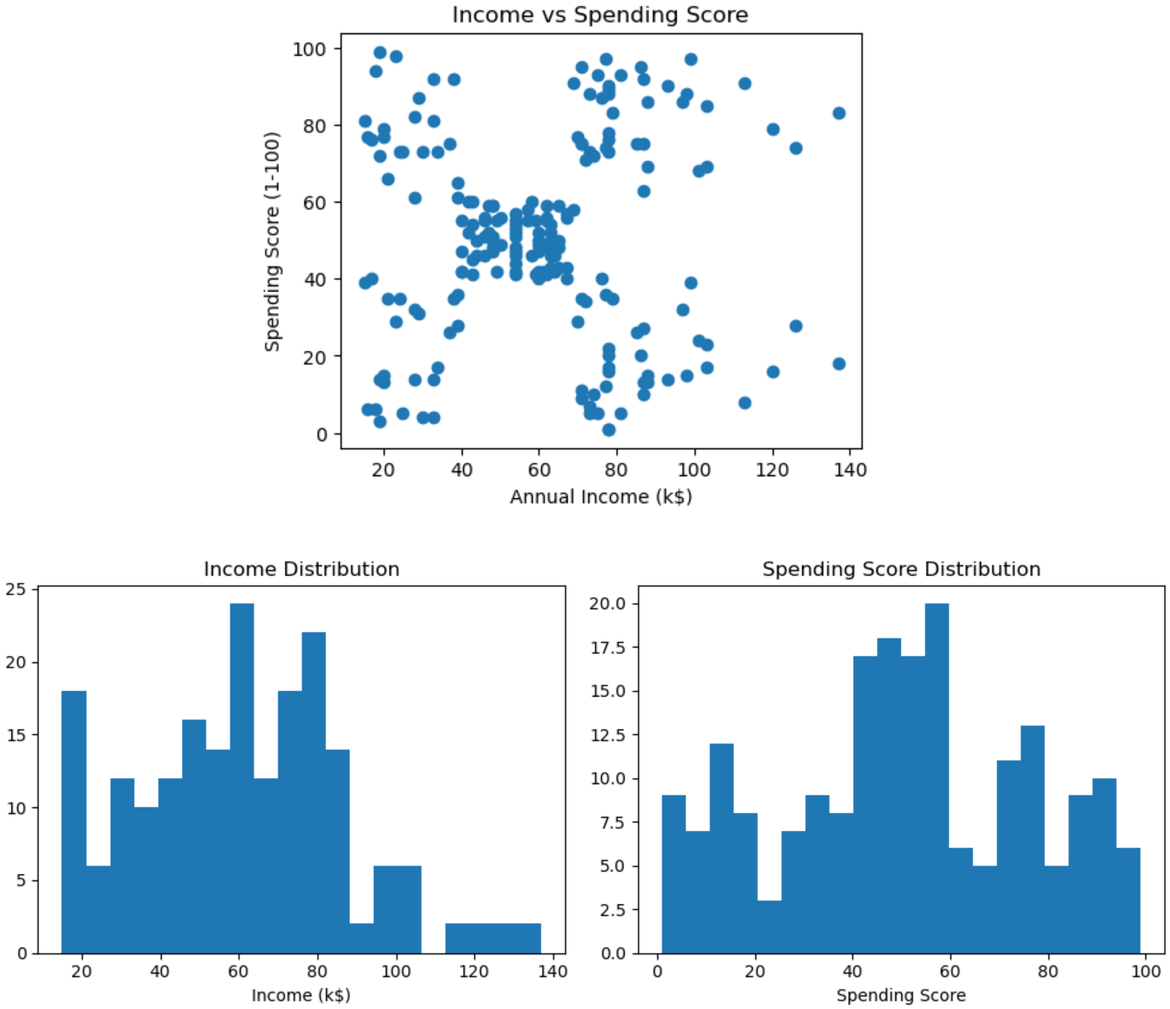

Figure 1: Initial Data Exploration

3.2. Feature Scaling

Since K-Means is a distance-based algorithm, the features Annual Income and Spending Score were standardized to a common scale. This prevents the variable with a larger range (Income) from dominating the distance calculation and ensures both features contribute equally to the model.

Code Snippet:

python

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

3.3. Determining the Optimal Number of Clusters (Elbow Method)

The Elbow Method plots the Within-Cluster-Sum-of-Squares (WCSS) against the number of clusters. The "elbow" point, where the rate of decrease in WCSS sharply changes, indicates the optimal number of clusters. For this dataset, the optimal value was determined to be k=5.

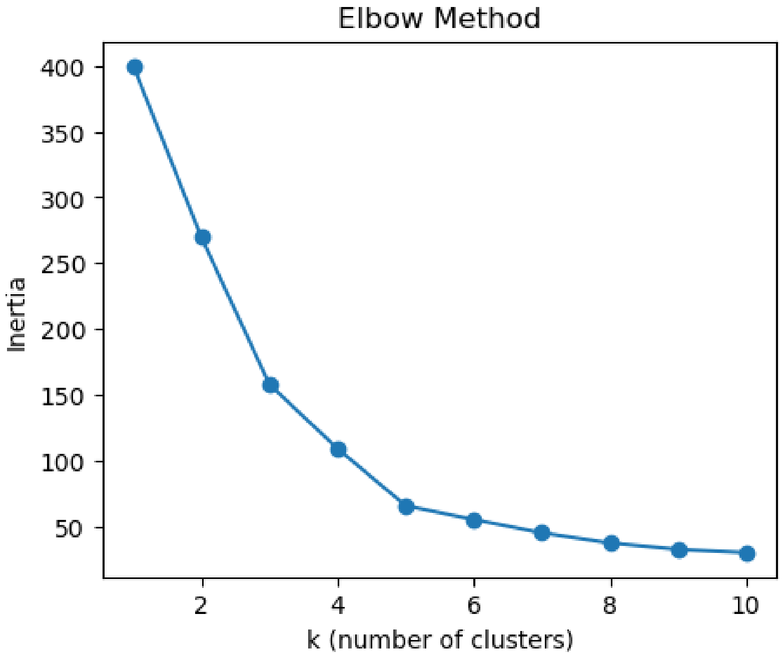

Figure 2: Elbow Method Graph

3.4. K-Means Clustering & Visualization

Using k=5, the K-Means algorithm was applied. The resulting clusters were visualized on a 2D plot, clearly showing five distinct customer segments.

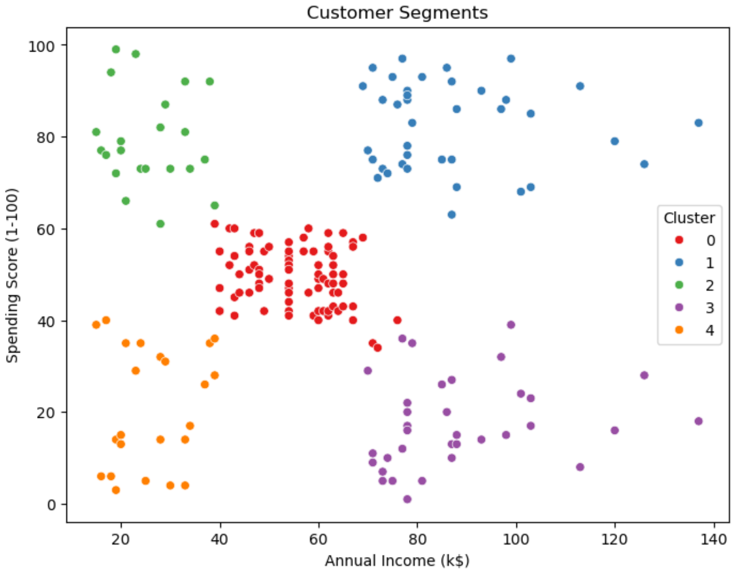

Figure 3: Final K-Means Clusters (Visualized on Original Data)

3.5. DBSCAN Clustering

The DBSCAN algorithm was applied as an alternative to K-Means. It identified a different cluster structure, including noise points (outliers). This provides an alternative perspective on how customers can be grouped.

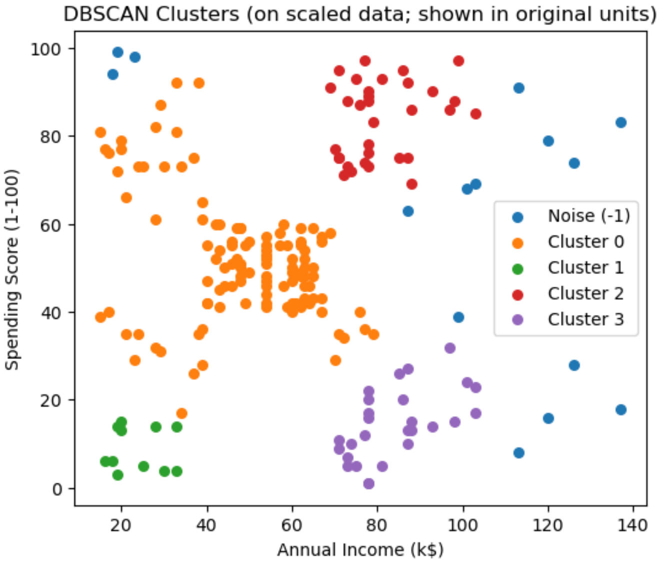

Figure 4: DBSCAN Clustering

3.6. Cluster Analysis and Average Spending

The clusters were analyzed to understand their characteristics. The table below summarizes the average annual income and spending score for each segment, forming the basis for our business insights. Summary of the average characteristics and size for each customer segment identified by K-Means clustering.

Table 1: Cluster Profile Analysis

Cluster	Avg. Income (k$)	Avg. Spending Score	Customers
0	55.296296	49.518519	81
1	86.538462	82.128205	39
2	25.727273	79.363636	22
3	88.200000	17.114286	35
4	26.304348	20.913043	23

4. Interpretation of Clusters & Business Insights

The clustering algorithm divided the customers into 5 distinct clusters, each representing a group of customers with similar behavior and characteristics. The average metrics for each cluster are shown in Table 1. Below is the interpretation of the results:

1. **Cluster 0 - Standard Customers:** This is the largest segment (81 customers) with medium annual income (~$55k) and a medium spending score (~50). They represent the reliable, average customer base. Strategy: Engage with broad marketing campaigns and cross-selling opportunities.
2. **Cluster 1 - Target Customers (VIPs):** This highly valuable segment (39 customers) has high income (~$87k) and a high spending score (~82). They are the ideal customers who should be retained. Strategy: Prioritize them with premium services, exclusive offers, and loyalty programs.
3. **Cluster 2 - Carefree Spenders:** This segment (22 customers) has low income (~$26k) but a very high spending score (~79). They are highly motivated buyers who spend a significant portion of their income. Strategy: Target with promotional deals, discounts, and value-oriented packages to maintain their engagement.
4. **Cluster 3 - Prudent High-Income:** This segment (35 customers) has the highest income (~$88k) but a low spending score (~17). They have high potential but are currently not engaged. Strategy: Use personalized marketing and high-quality product recommendations to convert their financial capacity into spending.

5. **Cluster 4 - Budget-Conscious:** This segment (23 customers) has low income (~$26k) and a low spending score (~21). They are the least profitable segment. Strategy: Focus on low-cost, high-volume marketing strategies if any, but do not allocate significant premium resources.

Conclusion

This project successfully demonstrated the application of unsupervised learning for customer segmentation. By applying the K-Means algorithm, customers were segmented into five distinct groups based on their income and spending behavior. The analysis provides clear, actionable insights that can directly inform targeted marketing strategies, resource allocation, and customer relationship management. The addition of DBSCAN offers an alternative clustering perspective, enriching the analysis. This work confirms that data-driven segmentation is a powerful tool for converting raw customer data into a strategic business asset.