Clustering Analysis of Mall Customers

1 Tasks & Procedure

This report follows the steps provided in the assignment to perform clustering analysis on the Mall Customers dataset. The steps include data exploration, preprocessing, and applying three clustering algorithms: K-Means, Hierarchical, and DBSCAN.

2 Part 1: Data Exploration and Preprocessing

2.1 Load the Data

The dataset `Mall_Customers.csv` is loaded into a pandas DataFrame.

2.2 Explore the Dataset

- Display the first few rows using `.head()`.
- Check the summary including data types and non-null values using `.info()`.
- Generate descriptive statistics using `.describe()`.

2.3 Data Selection

We focus on two features: **Annual Income (k$)** and **Spending Score (1-100)**. These columns are extracted to a new DataFrame for clustering.

2.4 Initial Visualization

The scatter plot below shows the distribution of customers based on their Annual Income and Spending Score:

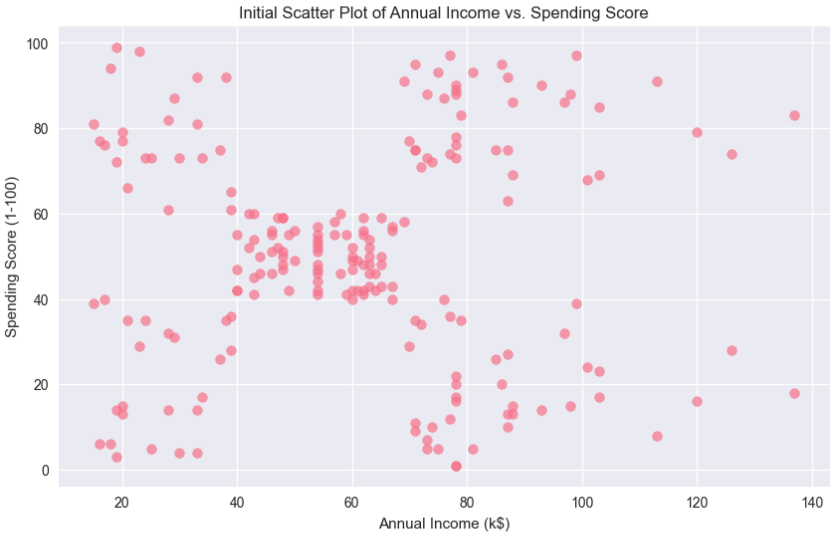

Figure 1: Scatter plot of Annual Income vs Spending Score

3 Part 2: K-Means Clustering

3.1 Finding the Optimal Number of Clusters (k)

The Elbow Method is used to identify the optimal number of clusters:

- Iterate over $k = 1$ to 10 and calculate WCSS (Within-Cluster Sum of Squares).
- Plot WCSS vs number of clusters to locate the "elbow" point.

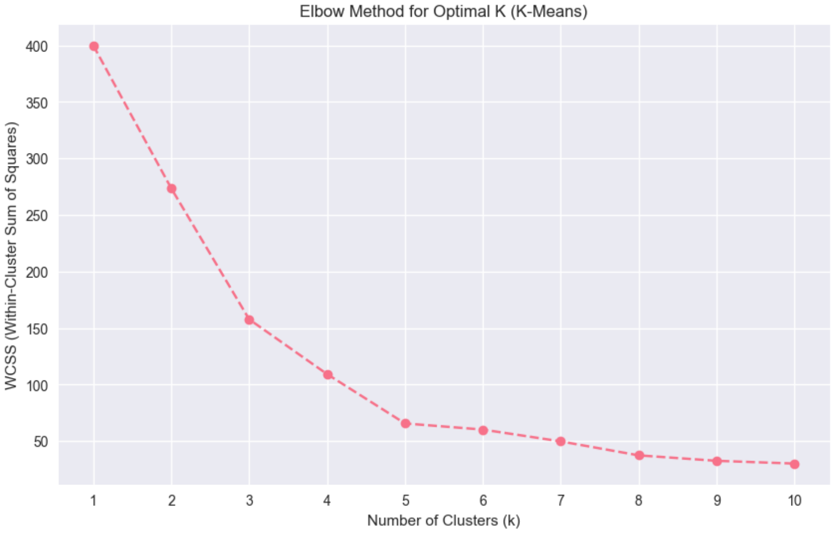

Figure 2: Elbow Method plot to determine optimal k

3.2 Applying K-Means

- Fit the K-Means model with the chosen number of clusters.
- Obtain cluster labels for each data point.

3.3 Visualize K-Means Results

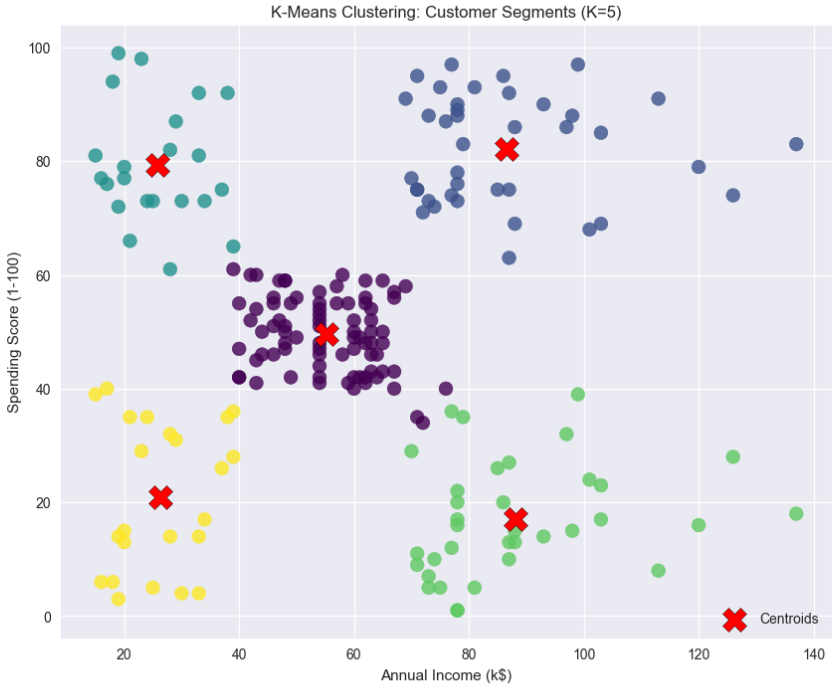

Figure 3: K-Means clustering results with cluster centroids

4 Part 3: Agglomerative Hierarchical Clustering

4.1 Creating a Dendrogram

- Generate dendrogram using `scipy.cluster.hierarchy` with Ward linkage.
- Determine optimal number of clusters by observing the dendrogram.

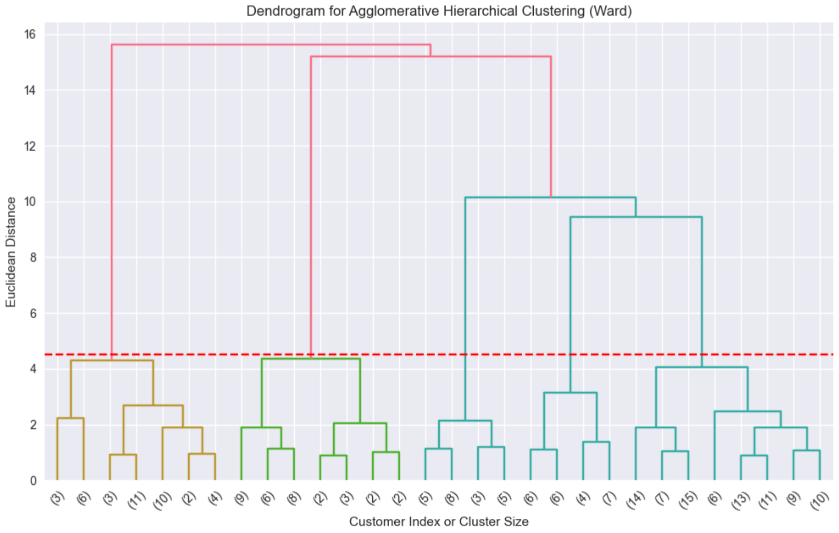

Figure 4: Dendrogram for Hierarchical Clustering

4.2 Applying Hierarchical Clustering

- Fit `AgglomerativeClustering` with the optimal number of clusters.
- Obtain cluster labels for each data point.

4.3 Visualize Hierarchical Clustering Results

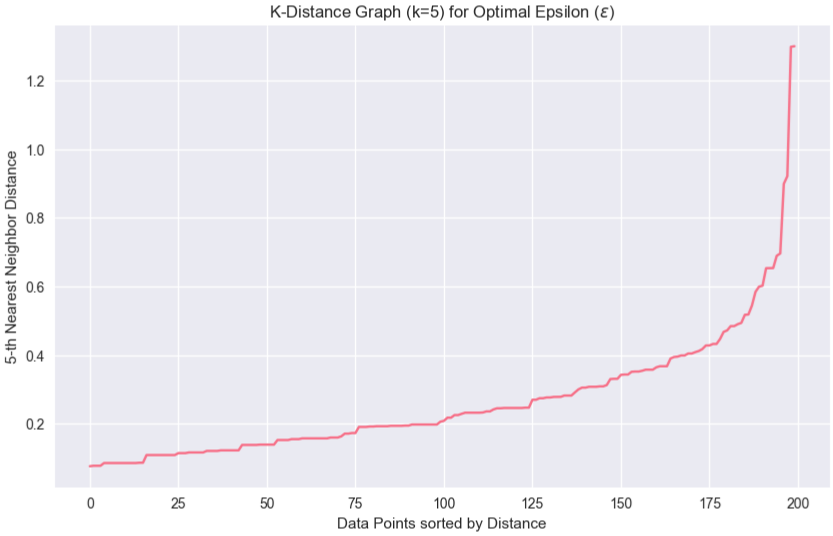

Figure 5: Hierarchical Clustering results

5 Part 4: DBSCAN Clustering

5.1 Applying DBSCAN

- DBSCAN requires parameters: `eps` and `min_samples`.
- Experiment with different values; for example, `eps=5`, `min_samples=5`.
- Fit DBSCAN and obtain cluster labels. Noise points are labeled as -1.

5.2 Visualize DBSCAN Results

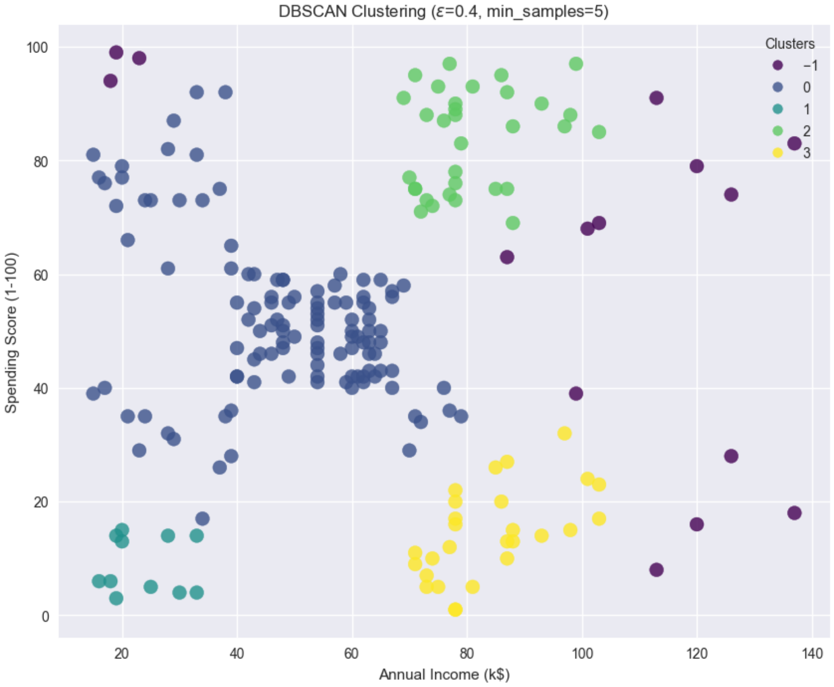

Figure 6: DBSCAN clustering results, with noise points in black

6 Part 5: Analysis and Questions

6.1 Optimal Clusters

- K-Means optimal clusters: `k=?` (determined from Elbow Method)
- Hierarchical Clustering optimal clusters: `?` (determined from dendrogram)

6.2 Cluster Comparison

Discuss similarities and differences between the clusters produced by the three algorithms.

6.3 DBSCAN Performance

Comment on DBSCAN's identification of clusters and noise points, and comparison with K-Means and Hierarchical Clustering.

6.4 Algorithm Suitability

Explain which algorithm is most suitable for this dataset and why, considering cluster shapes and density.

6.5 Real-World Application

Provide a hypothetical marketing scenario using the identified customer segments. For example, targeting high-income but low-spending customers with personalized offers.