ADSC4710

MACHINE LEARNING

Thompson Rivers University

Winter-2025

THOMPSON RIVERS UNIVERSITY

Customer Clustering: A Dual Approach Using Hierarchical and K-Means Methods

Submitted To:
Prof. Minoli

Submitted By:
T00728369 - Huynh Hiep Tran (Alex)
T00736529 - Viswateja Adothi
T00736533 - Akansha bhargavi
T00729053 -Khadiza Tannee

Table of contents

1. Abstract

This project focuses on customer segmentation using unsupervised machine learning techniques to identify distinct customer groups based on their behavior and demographic attributes. Customer segmentation plays a crucial role in personalized marketing, customer relationship management, and strategic business decision-making. We utilized the Mall Customers dataset from Kaggle, which includes demographic and behavioral information such as age, gender, annual income, and spending score.

The objective of this project was to apply and compare two popular clustering algorithms—K-Means Clustering and Hierarchical Clustering—to uncover meaningful patterns in customer behavior. The dataset underwent preprocessing, including column renaming, feature selection, and standardization. The Elbow Method and Silhouette Score were used to determine the optimal number of clusters, identified as five for both algorithms.These insights can support businesses in developing targeted marketing strategies, enhancing customer engagement, and improving resource allocation.

2. Introduction

In today's data-driven business environment, understanding customer behavior has become a critical factor in driving business success, enhancing customer satisfaction, and improving product or service personalization. As markets grow increasingly competitive, organizations must go beyond generic marketing approaches and instead focus on data-informed customer segmentation strategies that allow them to target the right audience with the right message.

Customer segmentation is the process of dividing a broad consumer or business market into sub-groups of customers based on shared characteristics such as demographics, purchasing habits, or behavioral traits. Traditional segmentation methods often fall short in capturing hidden patterns in complex datasets. This is where unsupervised machine learning techniques, particularly clustering algorithms, provide a powerful solution.

In this project, we aim to explore customer segmentation using the Mall Customers dataset from Kaggle, which includes customer demographics and spending behavior. Our primary objective is to apply and compare K-Means and Hierarchical Clustering algorithms to identify natural customer groupings. The effectiveness of these methods is evaluated using visualizations and Silhouette Score, offering actionable insights for data-driven marketing and strategic decision-making.

3. Dataset & Preprocessing

a. Dataset

The dataset used in this project is the Mall Customers Dataset, sourced from Kaggle. It contains a total of 200 observations, each representing a unique customer who is part of a mall's customer base. This dataset is widely used for clustering and segmentation tasks due to its simplicity, well-structured format, and relevance to real-world retail scenarios.

The dataset includes the following five attributes:

- CustomerID – A unique identifier assigned to each customer.
- Gender – The gender of the customer (Male/Female).
- Age – The age of the customer in years.
- Annual Income (k$) – The customer's annual income in thousands of dollars.
- Spending Score (1–100) – A score assigned by the mall, based on customer behavior and spending patterns. A higher score indicates higher spending and engagement.

This dataset provides a good balance of demographic (age, gender, income) and behavioral (spending score) features, making it suitable for customer segmentation using clustering algorithms.

It is a clean dataset with no missing values, making it ideal for direct application of machine learning techniques without intensive preprocessing. The relatively small size allows for fast experimentation with various clustering methods.

Dataset Source:

https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

b. Preprocessing Steps:

1. Loading and Inspecting the Data:

- The dataset was imported using pandas, and preliminary inspection was conducted using .head(), .info(), and .describe() to understand the structure and quality of the data.
- A check for missing values confirmed that the dataset was clean and complete, requiring no imputation.

2. Column Renaming:

- To maintain consistency and simplify column referencing, two columns were renamed:
  - Annual Income (k$) → Annual_income
  - Spending Score (1–100) → Spending_score

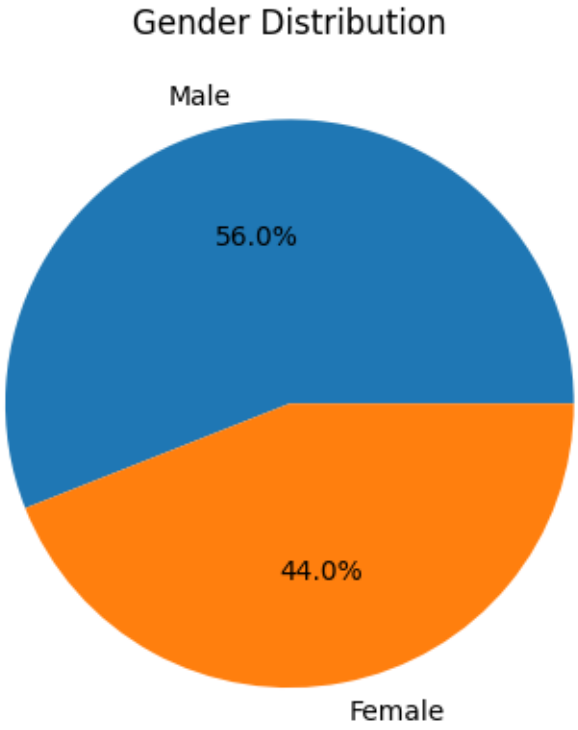
3.Feature Selection:

- The CustomerID column was removed, as it functions solely as a unique identifier and holds no analytical value in the clustering process.
- The Gender column, being categorical and not encoded in this version, was excluded from clustering analysis.
- The final selected features were:
  - Age
  - Annual_income
  - Spending_score

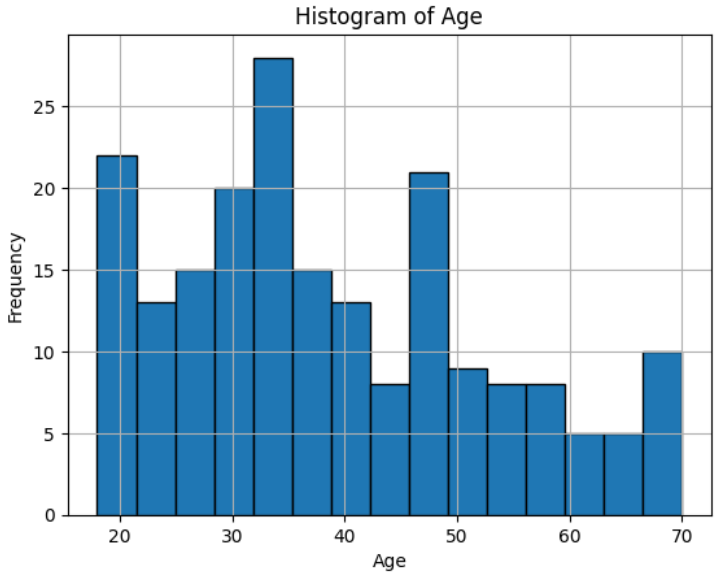
4. Feature Scaling
   - Numerical features were standardized using StandardScaler from Scikit-learn to ensure all features contribute equally to distance-based clustering.
   - This step helps improve the effectiveness of algorithms like K-Means and Hierarchical Clustering, which are sensitive to feature scales.

c. EDA

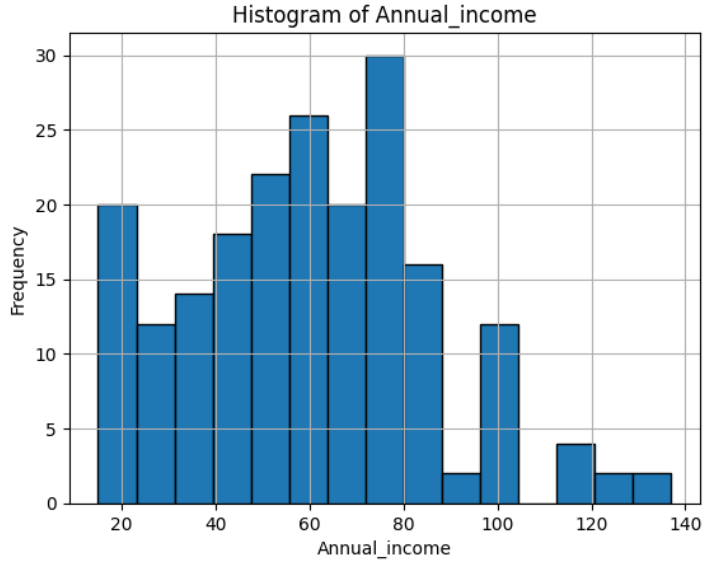
1. Gender distribution
    - The dataset consists of **56% males** and **44% females**, indicating a slightly higher representation of males. The distribution is relatively balanced.



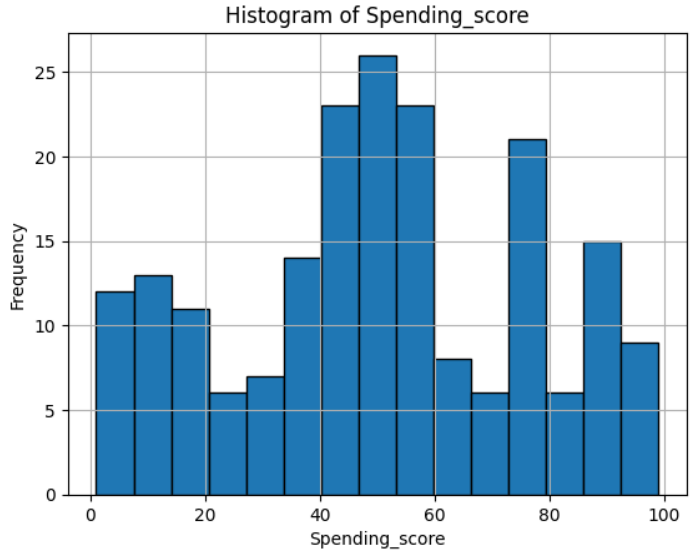
2. Histogram of Age
    - Below histogram displays the distribution of customer ages, revealing a roughly uniform spread with peaks in the early 30s and late 40s, and a tail extending to the late 60s.



3. Histogram of Annual_income
    - This histogram shows the distribution of annual incomes, revealing a right-skewed pattern with a concentration in the $60,000 to $80,000 range and a long tail towards higher incomes.



4. Histogram of Spending_score
    - This histogram displays the distribution of spending scores, revealing a multi-modal pattern with peaks around 10-20, 40-60, and 70-80, suggesting distinct customer segments based on spending behavior.

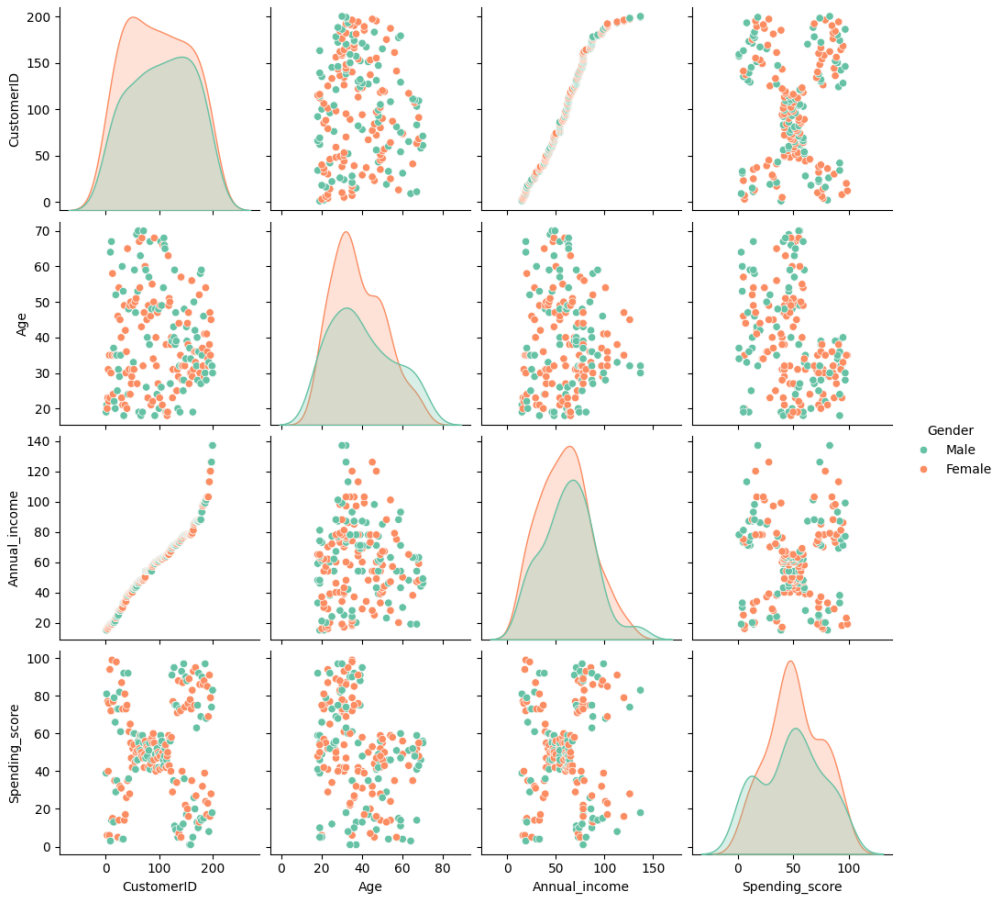


5. Pairplot:
    - Below pairplot was generated to visualize the relationships between CustomerID, Age, Annual Income, and Spending Score, stratified by Gender. The diagonal plots depict the distribution of each individual variable, while the off-diagonal plots illustrate the pairwise relationships.

Key observations:

- **CustomerID:** Appears uniformly distributed, as expected.
- **Age:** Shows a relatively normal distribution with a slight skew towards younger ages.
- **Annual Income:** Exhibits a strong linear relationship with CustomerID, suggesting a potential ordering effect.
- **Spending Score:** Displays a more complex distribution, with differences observed between male and female spending patterns.

This visualization provides a comprehensive overview of the data's structure and potential relationships between variables, highlighting differences related to gender."



6. Correlation matrix

The correlation matrix heatmap illustrates the linear relationships between Age, Annual Income, and Spending Score.

- **Age** shows a moderate negative correlation (-0.33) with **Spending Score**, suggesting older customers tend to have lower spending scores.
- **Age** and **Annual Income**, as well as **Annual Income** and **Spending Score**, exhibit negligible linear correlations (approximately 0).

This indicates that while age has some influence on spending score, annual income shows little to no linear relationship with either age or spending score.

4. Methodology

To achieve meaningful customer segmentation, we applied and compared two widely used unsupervised learning techniques: K-Means Clustering and Hierarchical Clustering. Both models were implemented using Python in Jupyter Notebooks, leveraging libraries such as Scikit-learn, Pandas, Seaborn, and Matplotlib for modeling, preprocessing, and visualization:

a. K-Means Clustering

K-Means is a centroid-based clustering algorithm that partitions the dataset into a predefined number of clusters (k). To determine the optimal value for k, we used the Elbow Method, which plots the Within-Cluster Sum of Squares (WCSS) for different values of k and identifies the point where the rate of decrease sharply changes. Based on this, k = 5 was selected.

We used the k-means++ initialization strategy to improve convergence and reduce the chances of suboptimal solutions. The clustering performance was evaluated using the Silhouette Score, which measures the compactness and separation of clusters. Additionally, we used Principal Component Analysis (PCA) to visualize clusters in 3D space.

b. Hierarchical Clustering

Hierarchical Clustering is a connectivity-based approach that does not require a predefined number of clusters. We experimented with multiple linkage methods—single, complete, average, and ward—and selected average linkage as the final method based on Silhouette Score performance.

A dendrogram was used to visualize the clustering hierarchy and identify a suitable number of clusters (5). PCA was again used for 3D visualization of the resulting clusters.

5. Results & Analysis

To evaluate the performance of both clustering algorithms, we used the Silhouette Score, which quantifies how well each data point fits within its assigned cluster relative to other clusters. A higher score indicates better-defined, more cohesive clusters.

Clustering Method	Number of Clusters	Silhouette Score
K-Means	5	0.42
Hierarchical (Average)	5	0.4096
Hierarchical (Complete)	5	0.4000
Hierarchical (Ward)	5	0.3900
Hierarchical (Single)	5	0.0030

K-Means Clustering Insights:

- Achieved the highest Silhouette Score (0.42) among all methods, indicating strong intra-cluster cohesion and inter-cluster separation.
- Computationally efficient and well-suited for larger datasets.
- PCA-based 3D plots showed visually distinct clusters with minimal overlap.
- Effective for quick and scalable customer segmentation.

Hierarchical Clustering Insights:

- All four linkage methods were evaluated.
- Average linkage provided the best clustering performance among them, with a Silhouette Score of 0.4096.
- Dendrogram visualization allowed intuitive exploration of cluster structures.
- Suitable for smaller datasets and detailed cluster analysis.

Cluster Profiling:

The clusters revealed distinct consumer patterns:

- **Cluster 0**: High income, high spenders
- **Cluster 1**: Young with moderate income and high spending
- **Cluster 2**: Low income, low spending
- **Cluster 3**: Middle-income, balanced spending
- **Cluster 4**: Older customers, varied spending patterns

These insights enable businesses to craft data-driven marketing strategies, focus loyalty programs, and optimize service offerings for specific customer segments.

6. Discussion & Challenges

Key Findings:

This project demonstrated that both K-Means and Hierarchical Clustering can effectively segment mall customers into meaningful groups based on their age, income, and spending behavior. Each algorithm successfully identified five distinct clusters, representing customer types with unique patterns in purchasing power and spending activity.

K-Means Clustering emerged as the best-performing method based on the Silhouette Score (0.42). It offered faster computation and clear separation among clusters, making it suitable for real-time or large-scale applications. In contrast, Hierarchical Clustering, while computationally heavier, provided an interpretable hierarchical structure, allowing detailed analysis through dendrograms.

Furthermore, the model selection process using Silhouette Score was particularly insightful in Hierarchical Clustering. We compared four linkage methods and found that average linkage achieved the best performance, reinforcing the importance of evaluating different configurations even within the same algorithm.

Challenges:

1. Determining the Optimal Number of Clusters:
   Though we used the Elbow Method and Silhouette Score, choosing the right number of clusters still involves subjective interpretation. Balancing between model complexity and interpretability was an ongoing decision.
2. Selecting the Best Linkage in Hierarchical Clustering:
   Initially, Ward linkage was assumed to be best, but empirical testing

showed average linkage had superior clustering performance. This required additional experimentation and analysis.

3. Feature Limitations:
   The dataset lacked transactional data, customer preferences, or location-based features. This limited our ability to cluster based on deeper behavioral patterns or lifecycle stages.

Suggested Improvements:

- Include Categorical Features:
  Encoding Gender using one-hot or label encoding could add another behavioral dimension to the clustering.
- Explore Other Clustering Algorithms:
  Methods like DBSCAN or Gaussian Mixture Models (GMM) could be explored to detect non-linear and probabilistic patterns.
- Enrich the Dataset:
  Adding features such as purchase history, visit frequency, or product categories would enable more robust segmentation.
- Cluster Validation:
  Use additional evaluation metrics like Calinski-Harabasz Index or Davies-Bouldin Index for multi-metric validation.

7. Conclusion & Future Work

This project successfully demonstrated the value of unsupervised learning techniques for performing customer segmentation. By applying and comparing K-Means Clustering and Hierarchical Clustering, we were able to uncover meaningful customer groups based on demographic and behavioral data. Each cluster revealed distinct consumer patterns, such as high-income high-spenders or young moderate-income individuals with elevated spending scores, providing businesses with valuable insights to inform personalized marketing strategies, targeted promotions, and resource allocation.

Among the clustering methods tested, K-Means proved to be the most efficient and scalable, achieving the highest Silhouette Score (0.42). Meanwhile, Hierarchical Clustering offered more interpretability through dendrograms and benefited from careful linkage selection, where average linkage outperformed other methods.

Despite the dataset being relatively small (200 samples), K-Means Clustering outperformed all hierarchical linkage methods, including average linkage, which had the highest score among the hierarchical options. This may seem counterintuitive, as hierarchical clustering is often recommended for smaller datasets due to its interpretability and precision. However, after standardizing the data and selecting three well-separated numerical features, the underlying structure of the data favored K-Means, which is optimized for forming compact, spherical clusters. This result reinforces the idea that algorithm selection should be guided by data characteristics and empirical performance metrics, such as the Silhouette Score, rather than general assumptions about dataset size or complexity.

While the clustering was successful, there remains room for improvement. Future work may include incorporating richer datasets with features such as purchase history, geographic location, and customer loyalty metrics to refine segmentation. Additionally, exploring alternative clustering techniques like DBSCAN (for density-based clustering) or Gaussian Mixture Models (GMM) (for soft clustering) could yield more nuanced and flexible segmentation strategies, especially when dealing with more complex or noisy datasets.

8. GitHub Repository Link

https://github.com/viswatejaadothi/Machine-Learning-Project.git

9. Contributions Section

Our team worked cohesively to analyze and interpret the dataset findings. Each member contributed equally to modeling, interpretation, and documentation tasks, ensuring a comprehensive and balanced approach to the analysis.

10. References

- Kaggle: Customer Segmentation Dataset
https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
- Scikit-learn Clustering User Guide
https://scikit-learn.org/stable/modules/clustering.html
- Seaborn Documentation
https://seaborn.pydata.org/
- Matplotlib Documentation
https://matplotlib.org/
-Scikit-learn: Silhouette Score
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
-Scikit-learn: Agglomerative Clustering
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
- Scikit-learn: KMeans
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- PCA: Principal Component Analysis
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html