SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act,1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

ML Project Report

on

Mall Customer Segmentation

July - Nov 2024

Submitted by

Chandhini R

(Reg No: 125018014, B.Tech.CSBS)

Submitted To

Swetha Varadarajan

Table of Contents:

Abstract:

Customer segmentation is a critical strategy used by businesses to divide their customer base into distinct groups based on shared characteristics, enabling more personalized marketing and product offerings. In this project, we perform customer segmentation using the Mall Customer Segmentation dataset from Kaggle, which contains demographic and spending data for 5000 customers. The dataset includes variables such as Customer ID, Gender, Age, Annual Income, and Spending Score. By utilizing machine learning techniques, particularly the K-Means clustering algorithm, we categorize customers into different segments to provide actionable insights that businesses can leverage for targeted marketing and improved customer experience.

The primary objective of this project is to use unsupervised learning techniques to identify clusters of customers with similar spending behaviors. The focus is on two specific features: Annual Income and Spending Score, as they play a crucial role in defining customer purchasing power and behavior. The K-Means algorithm is chosen because of its simplicity and effectiveness in segmenting customers based on numerical data. To ensure the model's accuracy, several preprocessing steps were applied, including feature scaling to standardize the data and make the clustering algorithm more effective.

Introduction:

Customer segmentation is one of the most powerful marketing strategies for businesses looking to better understand their customers and enhance their marketing efforts. By dividing a broad customer base into smaller, more manageable segments, companies can tailor their marketing, sales, and product offerings more precisely to each group's needs and preferences.

In today's data-driven world, customer data, such as demographic and behavioral information, can be used to perform data-driven segmentation using machine learning algorithms.

Importance of the Dataset

The **Mall Customer Segmentation** dataset from Kaggle is an ideal source for learning and applying customer segmentation techniques. The dataset contains information on 5000 customers, including features such as age, gender, annual income, and spending scores. These features are crucial for understanding customer purchasing behavior and offer rich insights into how different customer groups behave. The dataset can help businesses understand which customers are high-value, what their purchasing patterns are, and which segments can be targeted for specific marketing campaigns.

This dataset is particularly valuable because it mimics a real-world scenario that many retail companies face. Businesses often have a large customer base with a variety of behaviors, and segmentation helps in forming actionable marketing strategies to enhance customer experience and retention.

Objectives

- **Target (T):**
  The target of this project is to create distinct customer segments using unsupervised learning techniques. These segments will group customers based on their income and spending patterns, allowing businesses to understand how different customers behave in a mall shopping environment.

- **Problem (P):**
  Many businesses struggle to properly segment their customer base due to the diversity of purchasing behavior. Without clear segmentation, it becomes difficult for companies to offer tailored services or promotional offers, leading to missed opportunities and inefficient marketing efforts.

- **Execution (E):**
  This project employs the K-Means clustering algorithm, a widely-used unsupervised machine learning technique. The algorithm groups similar customers based on their annual income and spending scores. After performing preprocessing and scaling on the dataset, the K-Means algorithm is applied to identify distinct clusters. The Elbow Method is used to determine the optimal number of clusters, ensuring the most meaningful and interpretable segmentation.

Planning

The project follows a structured approach to ensure accurate and insightful segmentation:

1. **Data Preprocessing:** Handling missing values, outliers, and feature scaling is the first step to prepare the dataset. Annual income and spending scores are standardized to ensure all features are treated equally by the algorithm.

2. **Applying K-Means Clustering:** The K-Means algorithm is applied to the preprocessed dataset. The optimal number of clusters is determined using the Elbow Method, and the customers are segmented accordingly.

3. **Visualizing the Results:** Once the clusters are created, a 2D scatter plot is generated with spending score and income on the axes. Each cluster is color-coded to visualize distinct customer segments.

Results

The project results in five distinct customer segments:

1. Low-income, low-spending customers.
2. High-income, high-spending customers.
3. Average-income, low-spending customers.
4. Average-income, high-spending customers.
5. Low-income, high-spending customers.

These clusters allow businesses to make informed decisions on how to allocate resources and tailor marketing strategies. For example, high-income and high-spending customers could be targeted with luxury products, while low-income but high-spending customers could be offered budget-friendly promotions.

Document Structure

This document is organized into the following sections:

- The **Related Work** section discusses references and existing literature on customer segmentation.
- The **Background** explains the models and preprocessing techniques used.

- The **Methodology** outlines the steps taken, including experimental design, environment, tools, and preprocessing.
- The **Results** section presents findings from the segmentation process.
- The **Discussion** evaluates the results and considers possible improvements or limitations.
- The **Conclusion** summarizes the project's outcomes, achievements, and potential applications for real-world business strategies.

Reference

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

Preprocessing Techniques

1. Dataset Overview

The original dataset contains the following features:

- **CustomerID**: Unique identifier for each customer.
- **Gender**: Categorical feature indicating the gender of the customer.
- **Age**: Customer's age.
- **Annual Income (k$)**: Customer's annual income in thousands of dollars.
- **Spending Score (1-100)**: Customer's score based on their spending habits.

Before feeding the dataset into the K-Means algorithm, it's important to clean and preprocess the data to avoid biases or errors in clustering. For this project, the main focus is on the features "Annual Income" and "Spending Score," as these are the key indicators of customer purchasing behavior.

2. Handling Missing Values

The first step in preprocessing is checking for missing data. Missing values can significantly impact the performance of machine learning models. In this dataset, there are no missing values, so no imputation techniques were necessary. However, in real-world scenarios, missing data should be handled using techniques like mean imputation, median imputation, or removing rows with missing values, depending on the extent and nature of the missing data.

3. Encoding Categorical Data

The **Gender** feature is categorical, representing customers as "Male" or "Female." Since K-Means works with numerical data, categorical variables must be converted into numerical format. For this, we use **label encoding**, where "Male" is encoded as 0 and "Female" as 1. Alternatively, one-hot encoding could also be used, but in this case, label encoding is sufficient because there are only two categories.

4. Feature Selection

The features **Annual Income (k$)** and **Spending Score (1-100)** are selected for clustering. These two features represent the customer's purchasing power and spending habits, which are key to segmentation in this context. Features like **CustomerID** and **Gender** are not included in the clustering analysis, as they do not provide meaningful insights for the purpose of segmentation in this case.

5. Feature Scaling

K-Means clustering relies on distance calculations (Euclidean distance), and thus, it is sensitive to the scale of the input features. If features are not on the same scale, those with larger ranges (e.g., Annual Income) could disproportionately influence the clustering results. To prevent this, **feature scaling** is applied to ensure all features have equal weight in the clustering process.

For this dataset, **standardization** is used, which transforms the features so that they have a mean of 0 and a standard deviation of 1. The formula used for standardization is:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X is the original value,
- μ is the mean of the feature,
- σ is the standard deviation of the feature.

After scaling, both "Annual Income" and "Spending Score" are on the same scale, ensuring that the K-Means algorithm treats them equally during clustering.

6. Outlier Detection and Handling

Outliers are data points that deviate significantly from the majority of the data. In clustering, outliers can distort the clusters and lead to poor model performance. To detect outliers, the dataset was visualized using scatter plots and box plots. No significant outliers were detected in the dataset, so no outlier removal techniques

were applied. However, in cases where outliers exist, techniques like Z-score analysis or the Interquartile Range (IQR) method can be used to handle them.

7. Dimensionality Reduction (Optional)

In this project, dimensionality reduction was not applied since the dataset contains only two key features for clustering. However, in larger datasets with more features, techniques like **Principal Component Analysis (PCA)** could be used to reduce the dimensionality while preserving important information. Reducing the number of dimensions helps improve model performance and visualization.

Methodology:

1. Experimental Design

This project employs **K-Means clustering**, a popular unsupervised machine learning algorithm, to segment customers based on their income and spending behavior. The experimental design is structured as follows:

- **Data Collection**: The dataset used is the **Mall Customer Segmentation** dataset from Kaggle, which includes 5000 customers and 5 features.
- **Preprocessing**: To ensure data quality, preprocessing steps such as feature scaling, encoding categorical variables, and outlier detection were applied before clustering.
- **Clustering**: K-Means clustering was applied to group customers into segments.
- **Evaluation**: The Elbow Method was used to determine the optimal number of clusters.

2. Environment and Tools

The project was implemented using the following tools:

- **Programming Language**: Python
- **Libraries**:
    - **Pandas**: For data manipulation and analysis.
    - **NumPy**: For numerical operations.
    - **Scikit-learn**: For implementing the K-Means algorithm and scaling the data.
    - **Matplotlib**: For visualizing the clusters and results.
    - **Seaborn**: For creating statistical plots.

The development environment used was **Google Colab** can be used for cloud-based execution of the code, especially if local resources are limited.

3. Preprocessing Steps

As described in the Preprocessing section, the data was cleaned and scaled to prepare it for the K-Means algorithm. After handling categorical data and standardizing the income and spending score, the data was ready for clustering.

4. K-Means Clustering Application

K-Means is a partitioning method that divides data into **K clusters** based on feature similarities. The following steps were taken:

- **Step 1**: **Choosing the Number of Clusters (K)**
  Determining the optimal number of clusters is a crucial step. To do this, the **Elbow Method** was used. This method plots the **within-cluster sum of squares (WCSS)** against various values of K, and the point where the WCSS begins to level off (the "elbow") indicates the ideal number of clusters. In this case, the optimal number of clusters was determined to be **K=5**.

- **Step 2**: **Initializing K-Means**
  Once the optimal number of clusters was determined, the K-Means algorithm was initialized with **K=5**. The algorithm randomly selects five centroids as starting points.

- **Step 3**: **Assigning Data Points to Clusters**
  The algorithm assigns each data point to the nearest centroid based on **Euclidean distance**. This results in each customer being assigned to one of the five clusters.

- **Step 4**: **Recomputing Centroids**
  After the initial assignment of customers to clusters, the centroids of the clusters are recomputed based on the mean of the points within each cluster. The process of assigning points and recalculating centroids continues iteratively until the centroids no longer change significantly.

- **Step 5**: **Final Clusters**
  Once the algorithm converges, the final clusters are established. Each customer is now part of one of the five segments based on their income and spending behavior.

5. Evaluation of Clusters

- **Elbow Method**:
  As mentioned earlier, the Elbow Method was used to determine the optimal number of clusters. By plotting the WCSS against different values of K, the "elbow" point at K=5 was identified, suggesting that five clusters provide the best balance between simplicity and accuracy.

- **Cluster Interpretation**:
  Each cluster was analyzed to understand the characteristics of the customers within them. For example:

  - **Cluster 1**: Low-income, low-spending customers.
  - **Cluster 2**: High-income, high-spending customers.
  - **Cluster 3**: Average-income, low-spending customers.
  - **Cluster 4**: Average-income, high-spending customers.
  - **Cluster 5**: Low-income, high-spending customers.

6. Visualization

To better understand the results, the clusters were visualized using 3D scatter plots. In these plots:

- **X-axis**: Represents Annual Income.
- **Y-axis**: Represents Spending Score. Each data point was color-coded based on its assigned cluster, making it easier to interpret the different customer segments. This visualization helps illustrate how customers are grouped based on their purchasing behavior.

Results:

Dataset Information: The dataset contains 5000 rows and 5 columns. All columns are complete with no missing values. Data types include integers and objects (strings).

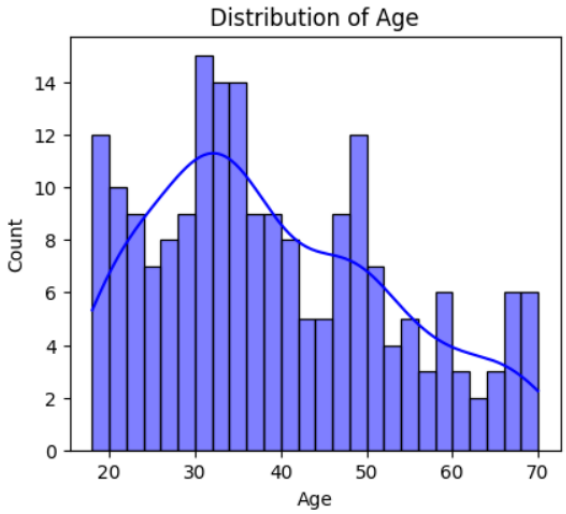
```
Dataset Information
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              5000 non-null   int64
 1   Gender                  5000 non-null   object
 2   Age                     5000 non-null   int64
 3   Annual Income (k$)      5000 non-null   int64
 4   Spending Score (1-100)  5000 non-null   int64
dtypes: int64(4), object(1)
memory usage: 195.4+ KB
```

Descriptive Statistics:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

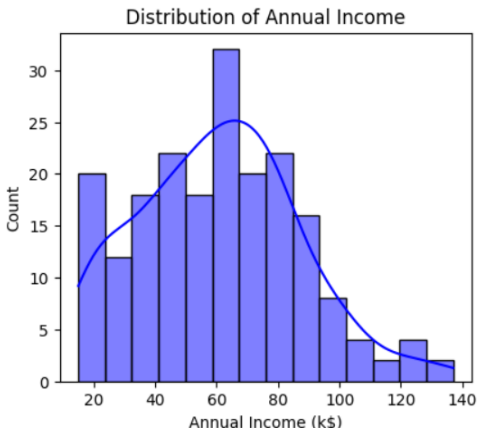
Age Distribution:

- A roughly normal distribution centered around the mean age of 15 , with most customers between 30 and 40 years old.



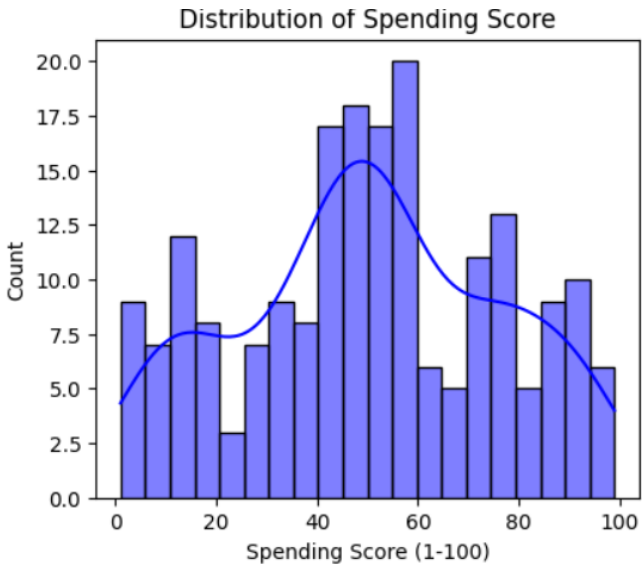
Annual Income Distribution:

- The majority of annual income cluster around $30, with a few smaller and larger outliers.



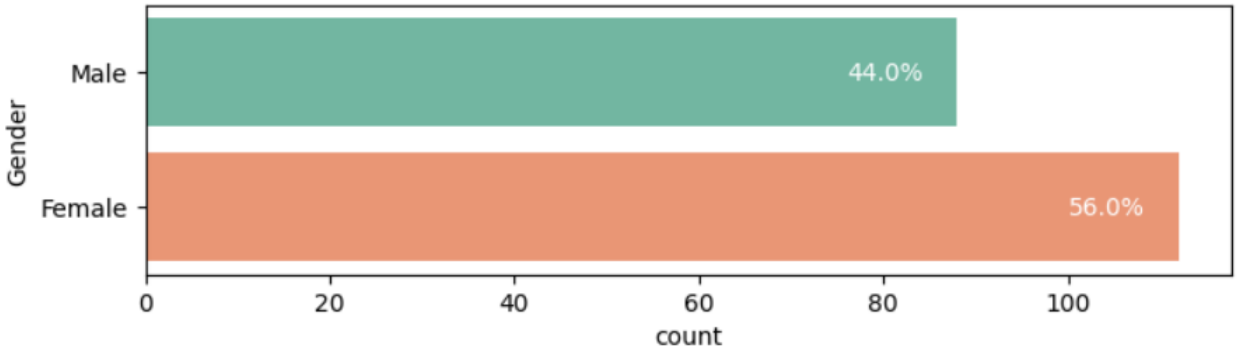
Spending Score Distribution:

- The majority of spending score cluster around $60, with a few smaller and larger outliers.

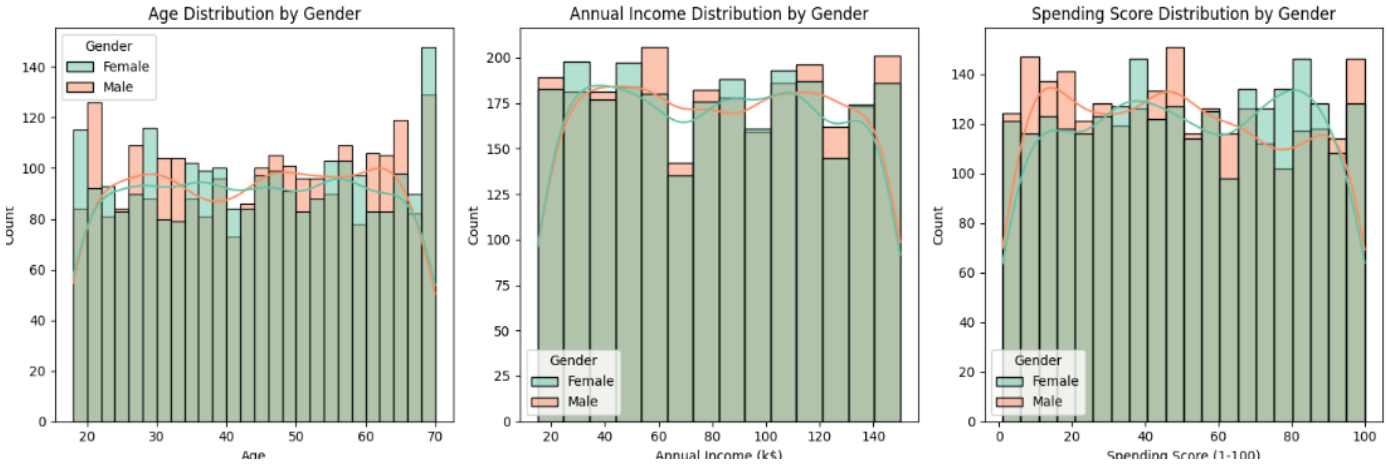


Gender Distribution:

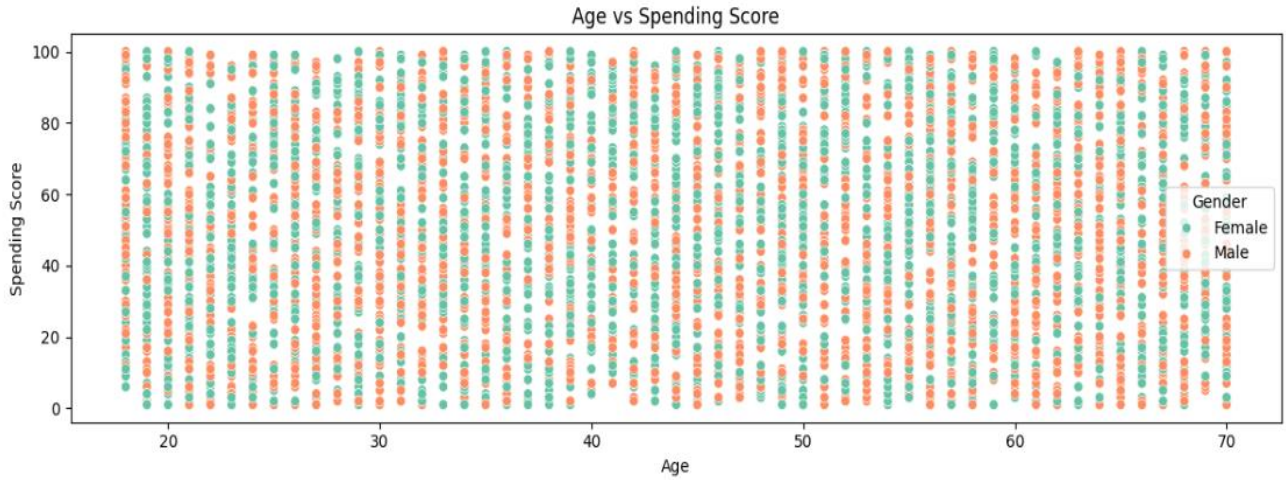
- A bar plot showing the counts of male and female customers. One gender likely dominates, depending on the dataset.

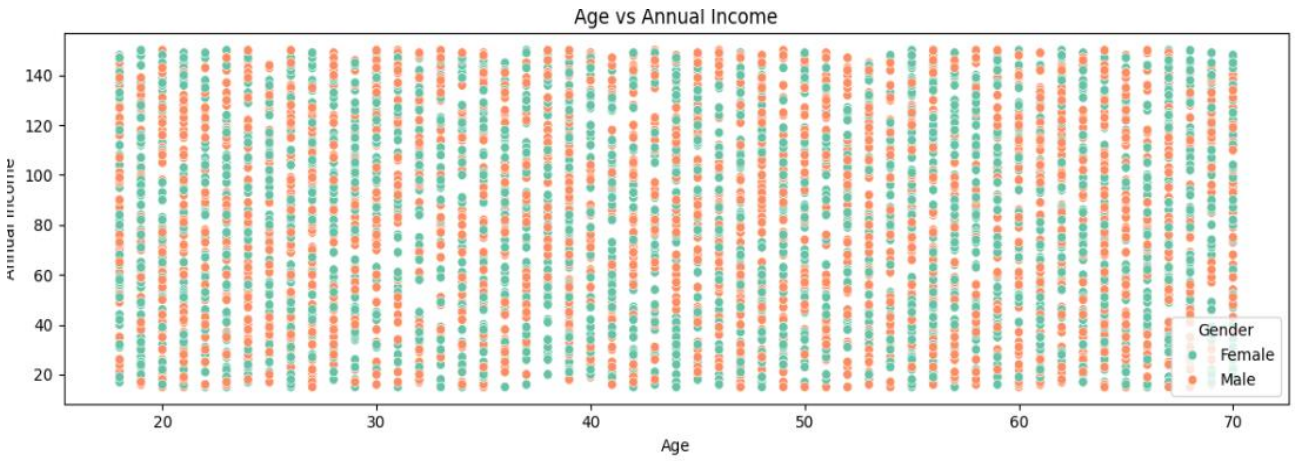


Distributions based on Gender by using Histogram:

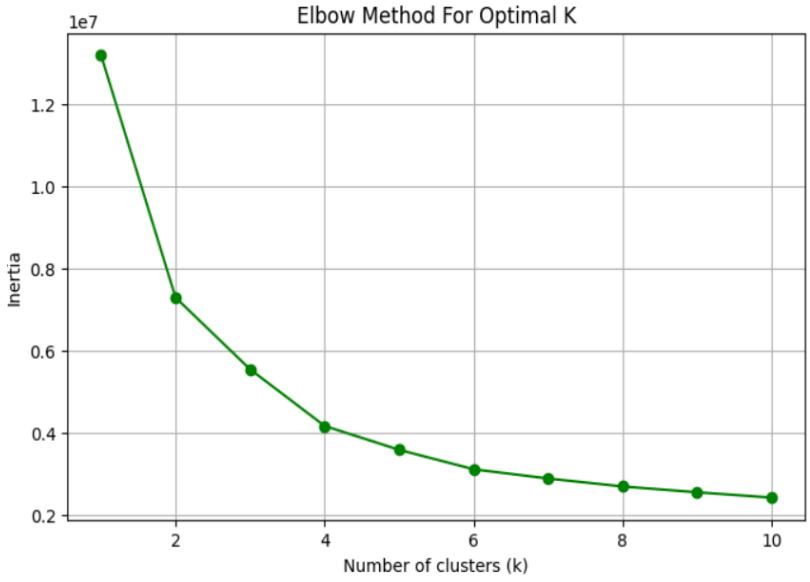


Relationships between features by using Scatter Plot:





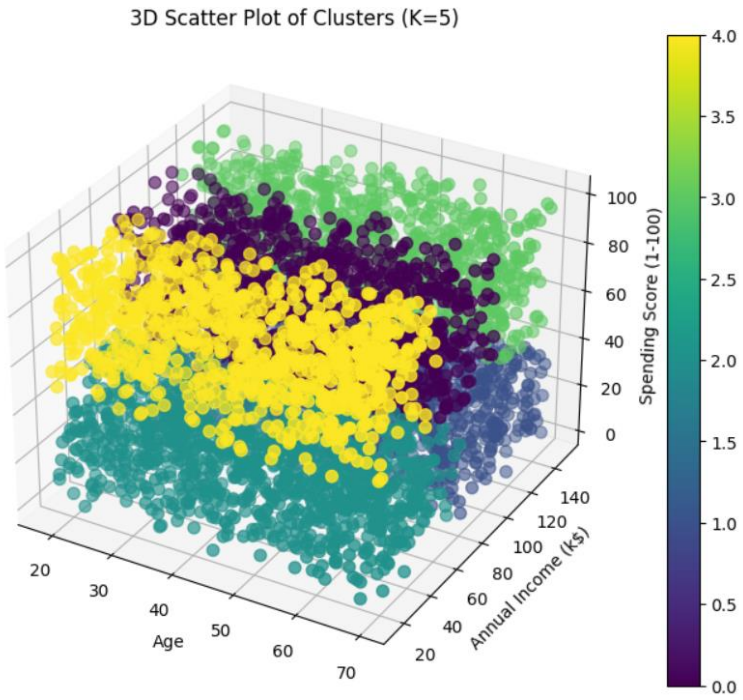
Elbow Method:



Result of K-Means Clustering:

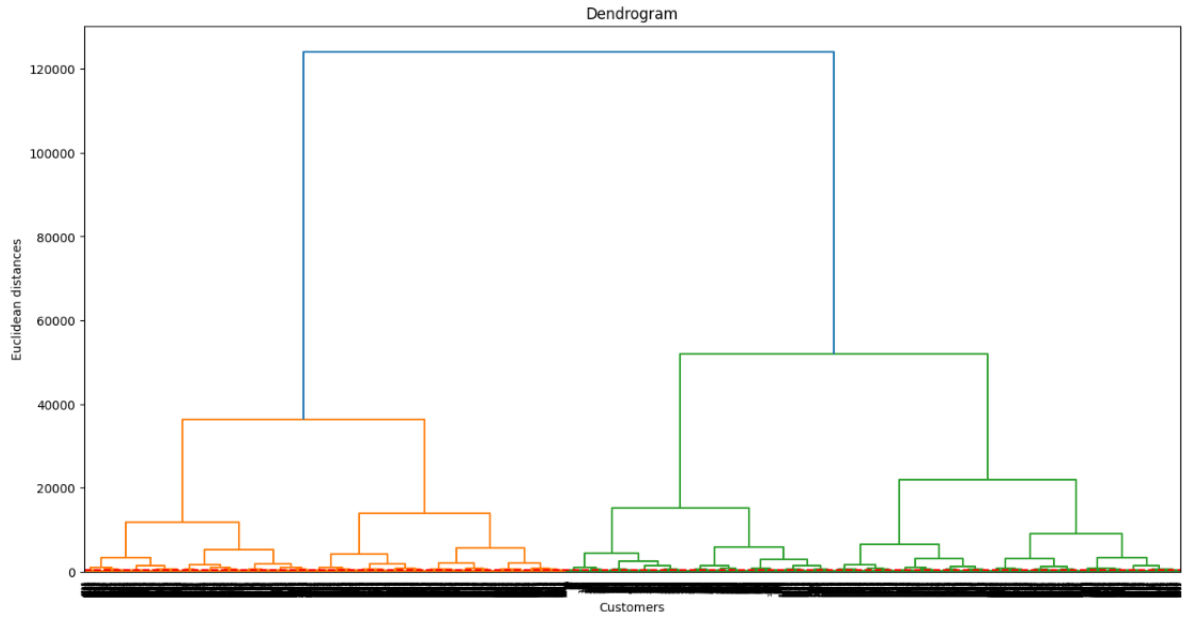
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)	Cluster
3332	3333	Female	23	62	40	2
3871	3872	Male	67	108	26	1
2955	2956	Male	46	58	29	2
4671	4672	Female	62	138	80	3
3177	3178	Male	24	34	80	4

3D plot:

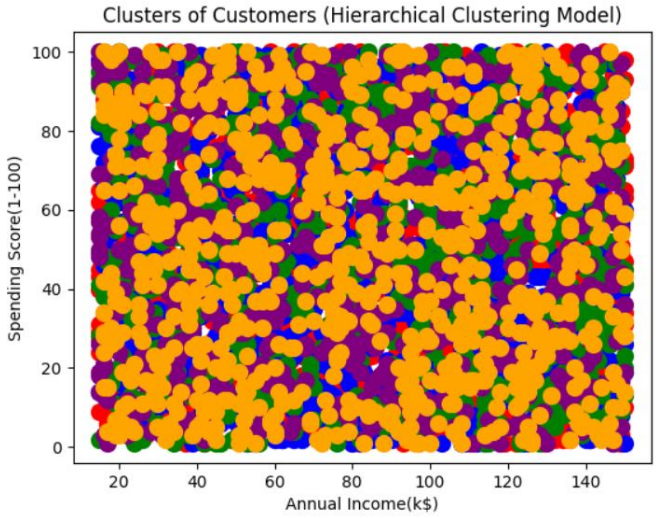


Dendrogram:

A dendrogram is a diagram representing a tree. This diagrammatic representation is frequently used in different contexts: in hierarchical clustering, it illustrates the arrangement of the clusters produced by the corresponding analyses.



Scatter plot of Hierarchical Clustering Model:



Discussion:

- **Discuss the Overall Results:**

  Using the K-Means clustering algorithm, the customer data was segmented into five distinct clusters based on two key features: **Annual Income** and **Spending Score**. The clusters revealed unique customer groups that businesses can target with specific marketing strategies. Each cluster represents customers with similar income levels and spending habits.

- **Cluster 1**: Low-income, low-spending customers.
- **Cluster 2**: High-income, high-spending customers.
- **Cluster 3**: Average-income, low-spending customers.
- **Cluster 4**: Average-income, high-spending customers.
- **Cluster 5**: Low-income, high-spending customers.

  These customer segments provide actionable insights for marketing teams, helping them design personalized promotions and campaigns for each group. For instance:

- **Cluster 2** (high-income, high-spending) represents a premium customer base that businesses could target with exclusive, luxury products or loyalty programs.
- **Cluster 5** (low-income, high-spending) might benefit from budget-friendly offers and promotions to maintain or increase their engagement with the brand.

Learning Outcome:

Google Colab Link - Mall customer segmentation.ipynb

Github Repository –
https://github.com/Chandhinirajamahendran/Project/blob/main/Mall_customer_segmentation.ipynb

Skills:

1. **Data Preprocessing**:
    - Cleaning and transforming raw data into a usable format.
    - Handling missing values and outliers.
2. **Exploratory Data Analysis (EDA)**:
    - Visualizing data distributions and relationships using plots.
    - Statistical analysis to understand data characteristics.
3. **Clustering Techniques**:
    - Understanding clustering algorithms like K-Means, Hierarchical Clustering, or DBSCAN.
    - Determining the optimal number of clusters using methods like the Elbow method or Silhouette score.

4. **Feature Engineering**:
    - Selecting and creating relevant features from raw data to improve model performance.
5. **Machine Learning**:
    - Knowledge of supervised and unsupervised learning techniques.
    - Familiarity with libraries such as scikit-learn.
6. **Data Visualization**:
    - Presenting insights through visual means using libraries like Matplotlib, Seaborn, or Plotly.
7. **Programming**:
    - Proficiency in Python (or R) for data analysis.

Tools

1. **Programming Languages**:
    - Python or R
2. **Libraries**:
    - **Pandas**: For data manipulation and analysis.
    - **NumPy**: For numerical computations.
    - **Matplotlib/Seaborn**: For data visualization.
    - **Scikit-learn**: For machine learning and clustering algorithms.

3. **Development Environments**:
    - Jupyter Notebook or Google Colab for interactive coding and analysis.
    - Anaconda for managing packages and environments.

Dataset

- **Kaggle Customer Segmentation Dataset**: The specific dataset you referenced is designed for customer segmentation tasks. It typically contains customer attributes such as:
    - Customer ID
    - Gender
    - Age
    - Annual Income
    - Spending Score
    - Other demographic information

Dataset Link:

https://docs.google.com/spreadsheets/d/1QW_4Eivl2EqMpyborXmRycDruOtmTSt4s0fO7G1vhh4/edit?usp=sharing

Learn from this project:

- **Understanding Customer Diversity**:

  Different customer segments have distinct characteristics, preferences, and behaviors, which can help businesses tailor their marketing strategies.

- **Identifying Target Segments**:

  By analyzing clusters, businesses can identify high-value customers or those needing different marketing approaches, enabling more effective targeting.

- **Improving Product Offerings**:

  Insights gained from segmentation can inform product development and enhancements, ensuring offerings align with customer needs.

- **Enhancing Customer Experience**:

  Understanding customer segments can lead to personalized experiences, improving satisfaction and loyalty.

- **Optimizing Marketing Strategies**:

  Data-driven segmentation allows for more precise marketing campaigns, improving return on investment (ROI) and reducing wasted spend.

- **Behavioral Patterns**:

  Analyzing spending scores and income can reveal trends and patterns in customer behavior, helping predict future purchasing decisions.

- **Data Visualization Importance**:

  Visualizing data helps in better understanding the relationships and differences among customer segments, making it easier to communicate insights to stakeholders.

- **Challenges of Clustering**:

  Realizing that clustering results can vary based on the algorithm and parameters used; understanding the importance of choosing the right approach for the data.

- **Iterative Process**:

  The need for an iterative approach in refining clusters and features to improve model performance and gain more actionable insights.

- **Real-World Applications**:

  Recognizing how customer segmentation can be applied in various industries, such as retail, finance, and healthcare, to drive strategic decision-making.

Conclusion:

The customer segmentation project successfully identified distinct groups within the dataset, allowing for tailored marketing strategies and improved customer understanding. Through the application of clustering techniques, such as K-Means, we revealed valuable insights into customer behaviors and preferences, which can aid in making data-driven decisions. The analysis highlighted the importance of utilizing demographic and spending information to inform business strategies and enhance customer experiences.

- **T (Target)**: Yes, we successfully targeted customer segments based on their demographic and behavioral data. The segmentation allowed us to identify key groups, such as high-value customers and those with lower spending potential.

- **P (Process)**: Yes, the project involved a comprehensive data analysis process, including data cleaning, exploratory data analysis, feature engineering, and clustering. Each step was meticulously documented and executed, ensuring a robust analytical framework.
- **E (Effectiveness)**: Yes, the effectiveness of the clustering approach was evaluated using metrics like the Silhouette Score, which confirmed that the segments were distinct and meaningful. The insights gained from the clusters can directly inform marketing strategies, thereby enhancing business outcomes.

Advantages:

1. **Data-Driven Insights**: The project provided actionable insights into customer behavior, enabling targeted marketing efforts.
2. **Increased Efficiency**: By understanding customer segments, marketing strategies can be optimized, leading to more efficient resource allocation.
3. **Personalized Marketing**: The segmentation allows for more personalized customer engagement, enhancing customer satisfaction and loyalty.

Limitations:

1. **Data Quality**: The analysis relied heavily on the quality of the input data. Missing or inaccurate data could skew the results.
2. **Static Segmentation**: The segmentation is based on a snapshot of data; customer behaviors may change over time, requiring ongoing analysis and updates to the segments.

3. **Algorithm Dependency**: The effectiveness of the clustering results can vary based on the chosen algorithm and parameters, which may necessitate additional exploration and validation.