Project Documentation: Customer Segmentation for a Retail Store

**Project Manager**

**Sanja Ranjan**

**Start Date**

[Insert Start Date]

**End Date**

[Insert End Date]

# 1. Objectives

The objective of this project is to segment customers into distinct groups based on their purchasing behavior. By understanding different customer segments, the retail store can improve targeted marketing strategies, enhance customer satisfaction, and increase sales.

## 1.1 Scope

Data Cleaning

Exploratory Data Analysis (EDA)

Customer Segmentation using K-Means Clustering

Visualization using Matplotlib and Power BI

## 1.2 Deliverables

Customer segmentation insights

Visualization reports

Recommendations for targeted marketing strategies

# 2. Business Requirements Document (BRD)

## 2.1 Business Problem

The retail store lacks a clear understanding of different customer profiles, leading to untargeted and inefficient marketing strategies.

## 2.2 Business Objectives

Improve customer satisfaction

Increase sales through targeted marketing

## 2.3 Functional Requirements

Data analysis

Clustering customers into segments

Visualization of customer segments

## 2.4 Non-functional Requirements

Performance: Ensure the solution is efficient and scalable.

Usability: The solution should be easy to use and interpret.

# 3. Technical Requirements Document (TRD)

**3.1 Data Sources**

Mall Customers dataset

## 3.2 Technologies

Python

Google Collab Notebook

Matplotlib

Seaborn

Scikit-learn

Power BI

## 3.3 Architecture

Data preprocessing

Exploratory Data Analysis (EDA)

Clustering

Visualization

## 3.4 Data Flow

Import data

Clean data

Analyze data

Segment customers

Visualize results

# 4. Project Plan

## 4.1 Tasks

Data Collection

Data Cleaning

Exploratory Data Analysis (EDA)

Clustering using K-Means

Visualization using Power BI

Documentation

## 4.2 Risks

Data quality issues

Algorithm performance

Visualization limitations

# 5. Methodology

## 5.1 Data Collection

The data was collected from the retail store's customer database. The dataset includes the following columns:

CustomerID

Genre (Gender)

Age

Annual Income (k$)

Spending Score (1-100)

## 5.2 Data Cleaning

Data cleaning involved handling missing values, converting categorical variables to numerical, and ensuring consistency in the data.

**Python code:**

*import pandas as pd* ***#Read data*** *path = "/content/drive/MyDrive/dataset/Mall\_Customers.csv" dataset = pd.read\_csv(path) dataset.head(20)* ***# Check dimensions*** *print("Dataset shape:", dataset.shape)*

***# Display first few rows*** *print(dataset.head(5))* ***# Check data types*** *print(dataset.dtypes)*

***# Summary statistics***

*print(dataset.describe())*

Output:

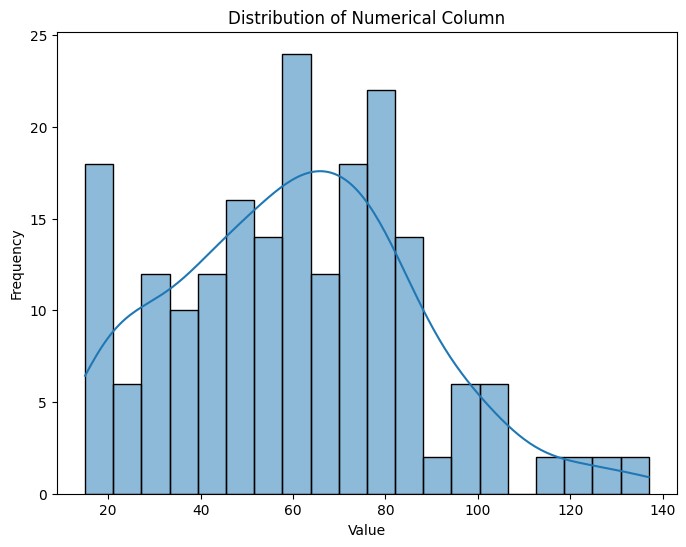
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **CustomerID** | **Genre** | **Age** | **Annual Income (k$)** | **Spending Score (1-100)** |
| **0** | 1 | Male | 19 | 15 | 39 |
| **1** | 2 | Male | 21 | 15 | 81 |
| **2** | 3 | Female | 20 | 16 | 6 |
| **3** | 4 | Female | 23 | 16 | 77 |
| **4** | 5 | Female | 31 | 17 | 40 |

5.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution of data and identify patterns.

1. Histogram of a numerical column

*import seaborn as sns import matplotlib.pyplot as plt plt.figure(figsize=(8, 6)) sns.histplot(dataset['Annual Income (k$)'], bins=20, kde=True) plt.title('Distribution of Numerical Column') plt.xlabel('Value') plt.ylabel('Frequency') plt.show()*

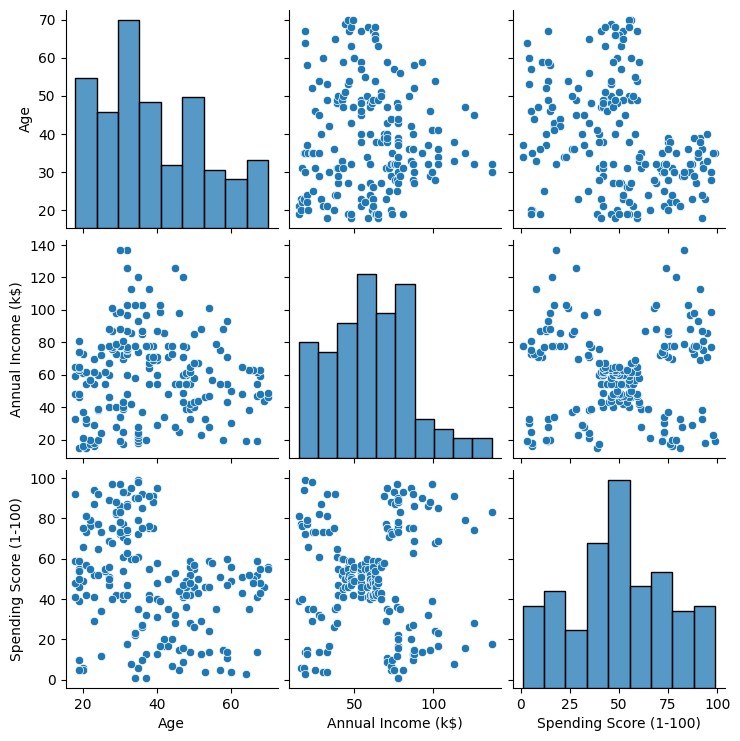
Output:

# 2. Pair plot to visualize relationships between numerical variables

**Python code:**

*sns.pairplot(dataset[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']]) plt.show()*

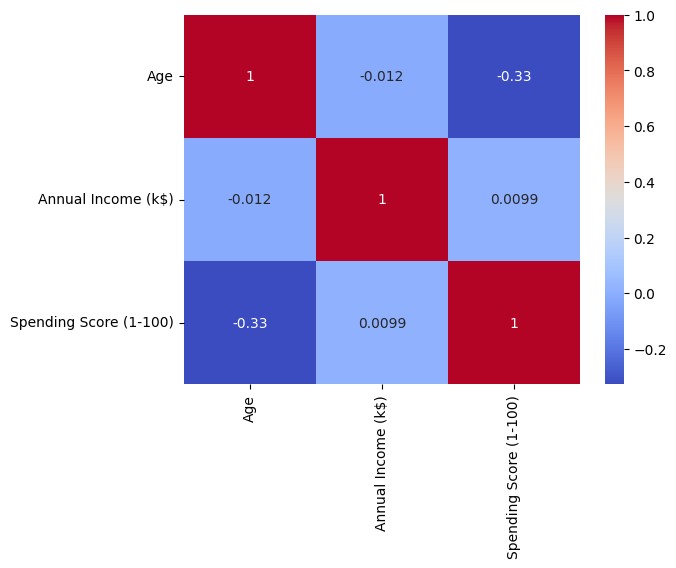
Output:



3. Correlation Matrix Python code:

*correlation\_matrix = dataset[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']].corr() print(correlation\_matrix) sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm') plt.show()*

Output:



5.4 Clustering Using K-Means

K-Means clustering was used to segment customers into distinct groups based on their purchasing behavior.

Python code:

*from sklearn.preprocessing import StandardScaler from sklearn.cluster import KMeans*

*# Plot the Elbow Graph # Select features for clustering*

*X = dataset[['Age', 'Annual Income (k$)', 'Spending Score (1-100)']]*

*# Standardize the features scaler = StandardScaler()*

*X\_scaled = scaler.fit\_transform(X)*

#KMeans Clustering

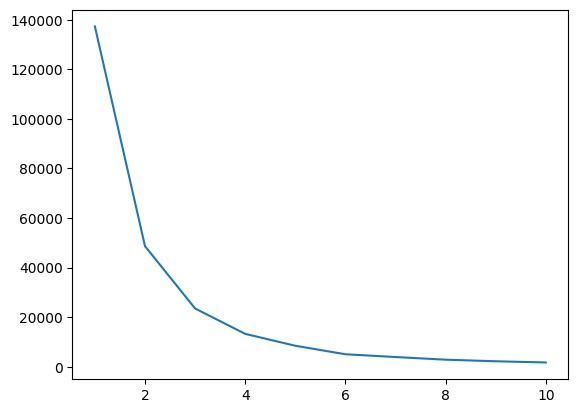
# Determine the optimal number of clusters using the elbow method

*inertia\_scores=[] for i in range(1,11):*

*kmeans=KMeans(n\_clusters=i) kmeans.fit(dataset[['Annual Income (k$)']]) inertia\_scores.append(kmeans.inertia\_) inertia\_scores*

*plt.plot(range(1,11),inertia\_scores)*

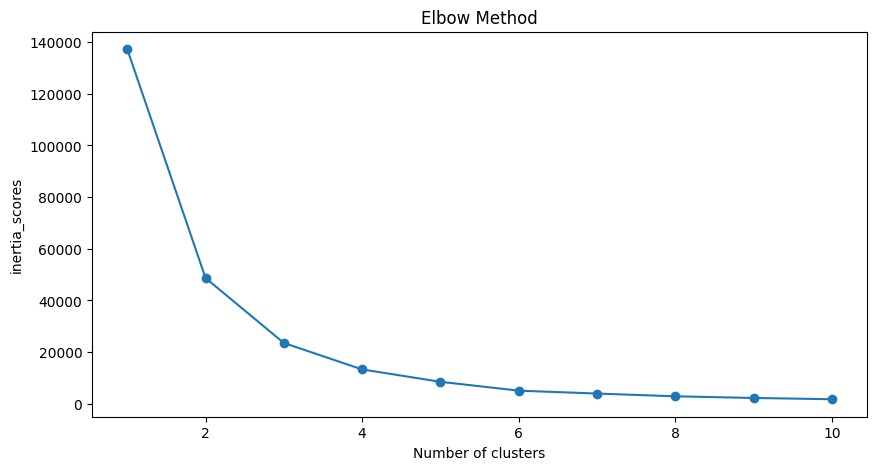
Output:



**Plot the Elbow Method:**

*Python code:*

*plt.figure(figsize=(10, 5)) plt.plot(range(1, 11), inertia\_scores, marker='o') plt.title('Elbow Method') plt.xlabel('Number of clusters') plt.ylabel('inertia\_scores') plt.show()*

Output: ***# Choose the optimal number of clusters (for example, 5)*** *optimal\_clusters = 5 kmeans = KMeans(n\_clusters=optimal\_clusters, init='k-means++', max\_iter=300, n\_init=10, random\_state=0) dataset['Cluster'] = kmeans.fit\_predict(X\_scaled) # Export the dataset with cluster labels to a CSV file dataset.to\_csv('customer\_segmentation.csv', index=False)*

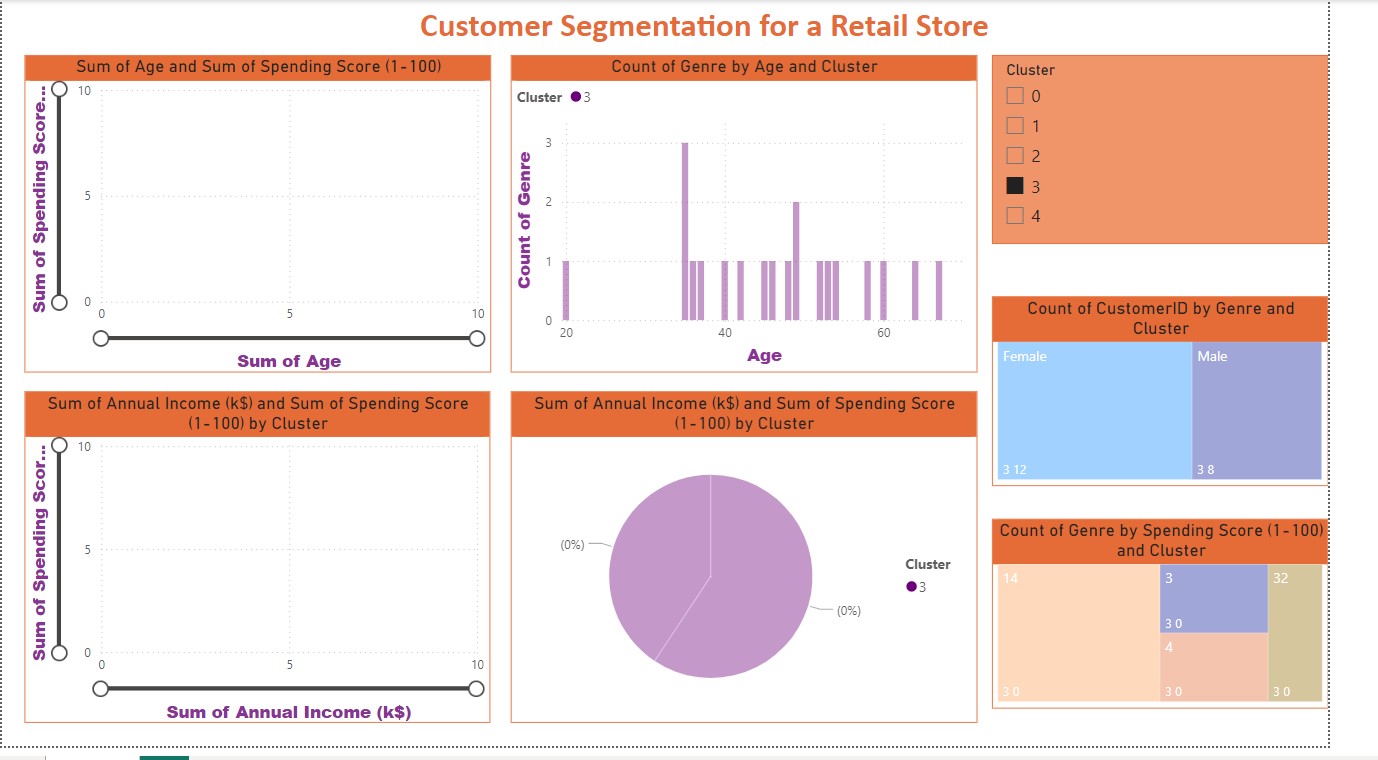
## 5.5 Visualization Using Power BI

The segmented data was visualized using Power BI to provide actionable insights. Below are the detailed steps and types of visualizations used.

**Load Data into Power BI:**

Import the Mall\_Customers\_with\_Clusters.csv file into Power BI.

**Visualization:**



**Detailed Explanation of VisualizationsScatter Plot:**

This visualization helps to identify how clusters are distributed based on annual income and spending score. It provides a clear view of customer segments and their spending patterns.

**Clustered Column Chart:**

This chart compares the count of customers across different genres within each cluster. It helps to understand the distribution of clusters by gender.

**Slicers:**

Slicers are used for dynamic filtering, allowing users to interactively explore different clusters and demographics. This makes the dashboard more user-friendly and insightful.

**Pie Chart:**

The pie chart shows the proportion of each cluster within the entire dataset, helping to understand the relative size of each customer segment.**Line Chart:**

The line chart analyzes trends over age within different clusters, showing how spending scores change with age for each cluster.

**Treemap:**

The treemap provides a hierarchical view of clusters within genres, highlighting the distribution and size of clusters in different genres.

# Conclusion

By implementing the above steps, we successfully segmented the customers into distinct groups based on their purchasing behavior. These insights can be used to tailor marketing strategies, improve customer satisfaction, and increase sales. The visualizations in Power BI provide a comprehensive view of customer segments, enabling data-driven decision-making.