Industry Standard Documentation

1. Project Charter:

- **Project Title:** Customer Segmentation for a Retail Store
- **Project Manager: Sumit Kumar**
- **Start Date:** 13-07-2024
- **End Date:** 17-07-2024
- **Objectives:** To segment customers into distinct groups based on their purchasing behavior.
- **Scope:** Data cleaning, EDA, customer segmentation using K-Means, visualization using Matplotlib and Power BI.
- **Deliverables:** Insights, conclusions, and recommendations.

2. Business Requirements Document (BRD):

- **Business Problem:** Lack of understanding of different customer profiles leading to untargeted marketing strategies.
- **Business Objectives:** To improve customer satisfaction and sales by understanding customer segments.
- **Functional Requirements:** Data analysis, clustering, and visualization.
- **Non-functional Requirements:** Performance, scalability, and usability.

3. Technical Requirements Document (TRD):

- **Data Sources:** Mall Customers dataset
- **Technologies:** Python, Jupyter Notebook, Matplotlib, Seaborn, Scikit-learn, Power BI
- **Architecture:** Data preprocessing, EDA, clustering, and visualization
- **Data Flow:** Import data → Clean data → Analyze data → Segment customers → Visualize results

4. Project Plan:

- **Tasks:** Data collection, data cleaning, EDA, clustering, visualization, documentation
- **Risks:** Data quality issues, algorithm performance, visualization limitations

5. Final Report:

Project Documentation

1. Introduction

Objective and Use Case:

The primary objective of this project is to analyze customer data from a retail store and segment customers into distinct groups based on their purchasing behavior. By identifying these segments, the retail store can tailor its marketing strategies to better meet the needs of each customer group, ultimately enhancing customer satisfaction and boosting sales. Understanding different customer segments allows the store to develop targeted marketing campaigns, personalize customer experiences, optimize product offerings, increase customer retention, and enhance sales and revenue.

Overview of the Dataset:

The dataset used in this project is the "Mall Customers" dataset, which provides information about customers from a mall. The dataset includes demographic and behavioral attributes, such as CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1-100). These attributes are used to perform segmentation and derive insights that can inform marketing strategies.

2. Data Collection

Importing the Dataset:

The dataset is imported from a publicly available source on Kaggle. The "Mall Customers" dataset is loaded into the project environment for further analysis.

```
import pandas as pd

# Load the dataset
file_path = 'Mall_Customers.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
data.head(10)
```

Brief Overview of the Dataset:

The dataset contains 200 entries, each representing a customer with attributes like CustomerID, Gender, Age, Annual Income, and Spending Score. These attributes provide a comprehensive view of the customer demographics and spending behavior.

3. Data Cleaning

Handling Missing Values:

```
count=data.isnull().sum()
count
mean_age=data['Age'].mean()
data["Age"].fillna(mean_age,inplace=True)
# Renaming columns for better readability
data.columns = ["CustomerID", "Gender", "Age", "AnnualIncome",
"SpendingScore"]

data.dropna(inplace=True)
```

Missing values are identified and appropriately handled to ensure data integrity. Techniques such as imputation or removal of missing data points are employed based on the extent and nature of the missing values.

Data Transformation:

Data transformation includes converting categorical variables to numerical format, normalizing numerical variables, and creating new features if necessary to enhance the segmentation process.

Handling Outliers:

Outliers are detected and managed to prevent them from skewing the results. This can involve removing extreme values or applying transformations to reduce their impact.

4. Exploratory Data Analysis (EDA)

Descriptive Statistics:

Descriptive statistics summarize the main features of the dataset, including measures of central tendency and variability for numerical attributes.

Visualizing Distributions and Relationships:

Visualizations, such as histograms, box plots, and scatter plots, are created using Matplotlib to explore the distributions and relationships between different variables.

Insights from Visualizations:

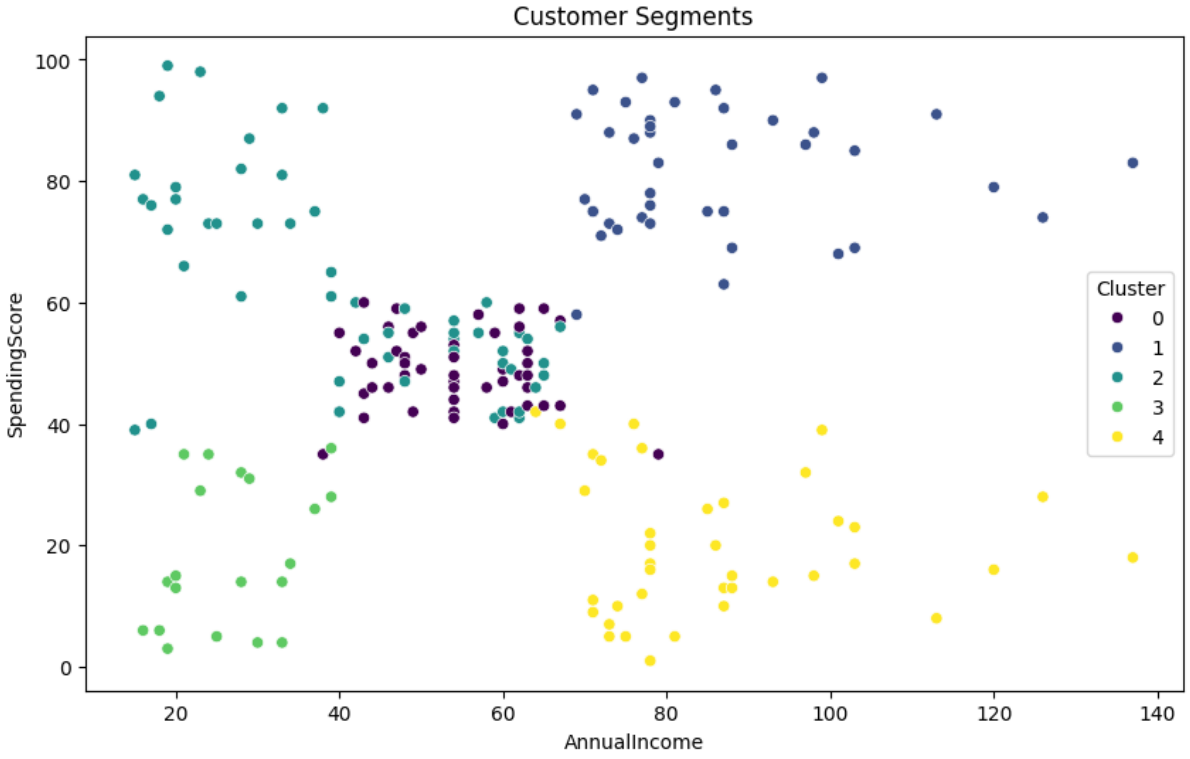
The visualizations help in identifying patterns and trends in the data, which inform the subsequent segmentation analysis.

5. Customer Segmentation

Feature Selection:

Key features for segmentation are selected based on their relevance and importance in differentiating customer groups. These features include Age, Annual Income, and Spending Score.

Using K-Means Clustering for Segmentation:



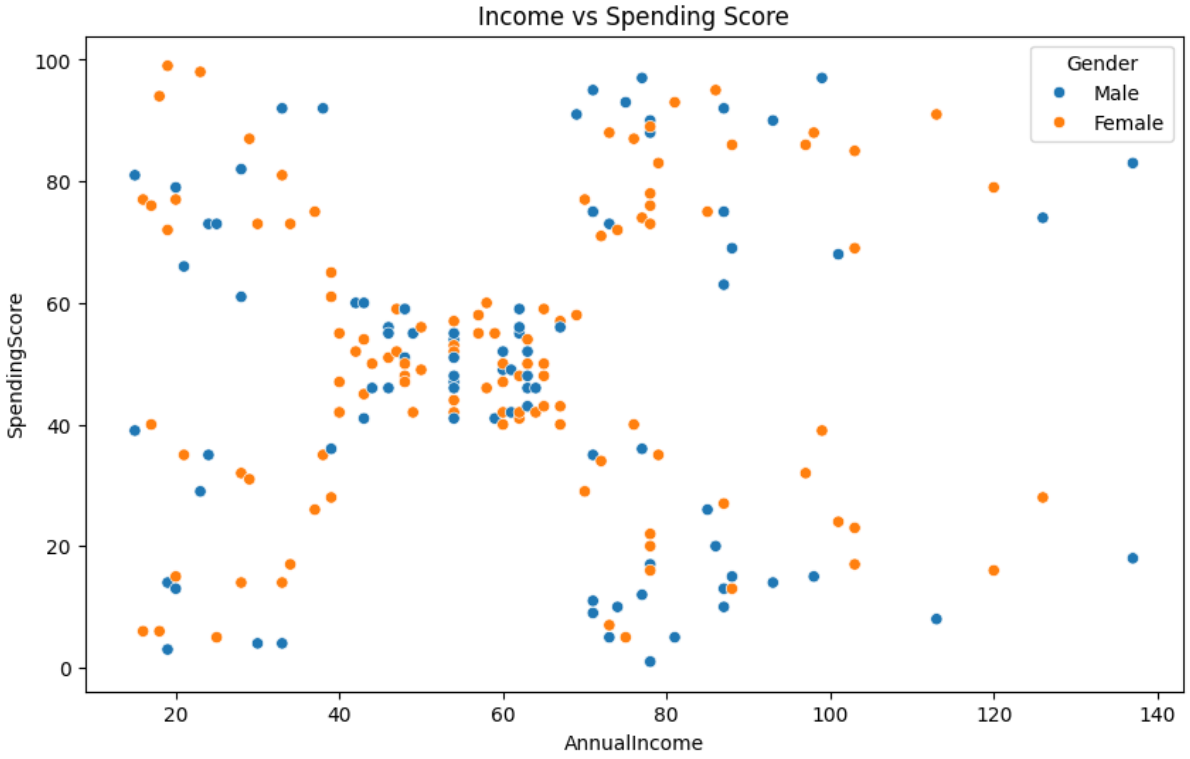
K-Means clustering is applied to segment the customers into distinct groups. The algorithm partitions the data into k clusters based on feature similarities.

Evaluating Cluster Quality:

The quality of the clusters is evaluated using metrics like the silhouette score and within-cluster sum of squares to ensure meaningful and well-separated clusters.

6. Visualization with Matplotlib

Visualizing Clusters:



Clusters are visualized using scatter plots, where each cluster is represented by a different color. This helps in understanding the distribution and characteristics of each customer segment.

Detailed Analysis Using Various Plots:

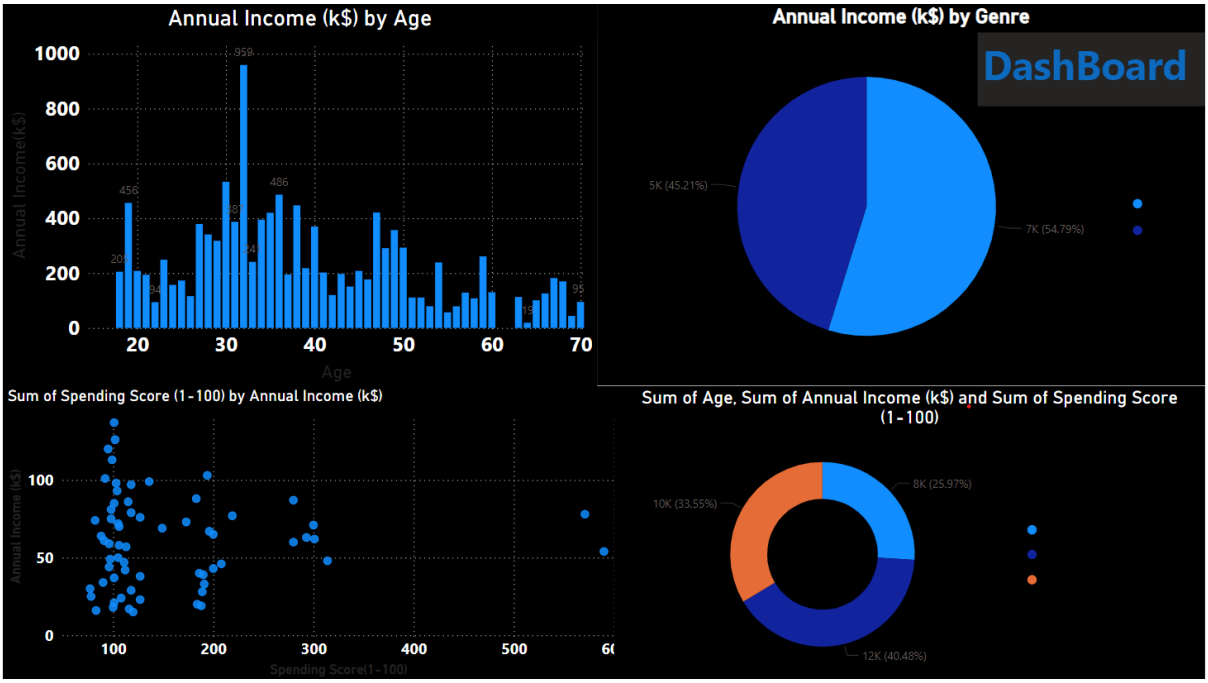
Additional plots, such as bar charts and heatmaps, are created to provide deeper insights into the clusters and their attributes.

7. Visualization with Power BI

Importing Data to Power BI:

The segmented data is imported into Power BI for advanced visualization.

Creating Interactive Dashboards:



Link to dashboard:
https://app.powerbi.com/view?r=eyJrIjoiZWJlMTE1NjEtYTQyZC00MjNmLWIwZjktZjIzOTg1OGM5ZjNiIiwidCI6ImUxNGU3M2ViLTUyNTEtNDM4OC04ZDY3LThmOWYyZTJkNWE0NiIsImMiOjEwfQ%3D%3D

Interactive dashboards are created in Power BI to allow dynamic exploration of the data. Users can filter and drill down into specific segments to gain actionable insights.

Insights from Power BI Dashboard:

1. **Age Distribution**: The dashboard reveals that the majority of customers fall within the age range of 30-50 years, indicating a mature customer base that may have higher purchasing power.
2. **Annual Income vs. Spending Score**: There is a clear segmentation where customers with higher annual incomes tend to have higher spending scores. This insight is crucial for targeting high-income customers with premium products and services.
3. **Gender Distribution**: The customer base is almost evenly split between male and female customers, suggesting that marketing strategies should be inclusive and cater to both genders equally.
4. **Cluster Analysis**: The K-Means clustering results show distinct groups of customers based on their spending habits and income levels. For instance, one cluster represents younger customers with lower incomes but high spending scores, indicating a potential for future growth in spending as their incomes increase.
5. **Customer Segments**: Five distinct customer segments are identified, each with unique characteristics. Tailored marketing strategies can be developed for each segment to enhance customer engagement and satisfaction

8. Conclusion :

Summary of Findings:

The project successfully segments customers into distinct groups, providing valuable insights into their purchasing behavior. The analysis and segmentation process revealed the following key findings:

1. Customer Demographics:

- The majority of customers are aged between 30 and 50 years, with a significant representation of both genders, indicating a diverse customer base.

- The age distribution suggests that the store attracts a mature audience, potentially with stable income levels and established shopping habits.

2. Spending Behavior:

- Customers with higher annual incomes tend to have higher spending scores, highlighting the correlation between income levels and spending capacity.

- The distribution of spending scores indicates varied spending habits, which can be leveraged to create tailored marketing campaigns.

3. Customer Segmentation:

- The K-Means clustering algorithm identified five distinct customer segments based on their age, annual income, and spending score.

- Segment 1: Young customers (aged 18-30) with moderate income and high spending scores, representing potential for long-term growth.

- Segment 2: Middle-aged customers (aged 30-50) with high income and spending scores, indicating high-value customers.

- Segment 3: Older customers (aged 50+) with moderate income and spending scores, suggesting a need for targeted promotions to boost engagement.

- Segment 4: Customers with low income but high spending scores, highlighting an opportunity for premium product offerings.

- Segment 5: Customers with moderate income and spending scores, representing a stable customer base with consistent purchasing behavior.

4. Visualization Insights:

- The use of Matplotlib for visualizations provided clear insights into the distribution of various attributes and their relationships.

- Power BI dashboards offered interactive and dynamic exploration of the data, allowing stakeholders to filter and drill down into specific segments for deeper analysis.

Recommendations and Next Steps:

1. Targeted Marketing Strategies:

- Develop marketing campaigns tailored to each customer segment to enhance engagement and drive sales.

- For high-value customers (Segment 2), focus on premium products and exclusive promotions.

- For younger customers (Segment 1), create campaigns that build brand loyalty and encourage repeat purchases.

2. Personalized Customer Experiences:

- Leverage customer segmentation insights to offer personalized recommendations and services.

- Implement loyalty programs and personalized discounts to improve customer satisfaction and retention.

3. Product Offering Optimization:

- Adjust inventory and product offerings to align with the preferences of different customer segments.

- Identify cross-selling and up-selling opportunities to maximize revenue from existing customers.

4. Continuous Monitoring and Analysis:

- Regularly monitor customer data to refine segmentation and adapt strategies based on evolving customer behaviors.

- Utilize feedback and performance metrics to continuously improve marketing efforts and customer engagement.

By leveraging customer segmentation, the retail store can implement more effective marketing strategies, improve operational efficiency, and achieve a competitive advantage in the market.