**Technical Requirements Document (TRD)**

**Data Sources**

**Mall Customers Dataset:**

* **Explanation:** The primary dataset for this project contains details about customers from a mall, including:
  + **CustomerID:** Unique identifier for each customer.
  + **Gender:** Gender of the customer.
  + **Age:** Age of the customer.
  + **Annual Income (k$):** Annual income of the customer in thousands of dollars.
  + **Spending Score (1-100):** Score assigned by the mall based on customer behavior and spending patterns. This data is crucial for understanding customer demographics and behaviors, which will be used for segmentation.

**Technologies**

**Python:**

* **Explanation:** Python is chosen for its simplicity, readability, and extensive library support. It is widely used in data analysis and machine learning projects.

**Jupyter Notebook:**

* **Explanation:** Jupyter Notebook provides an interactive environment where Python code can be written and executed in cells. It is ideal for exploratory data analysis, visualizations, and sharing insights through interactive notebooks.

**Matplotlib:**

* **Explanation:** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. It will be used to plot data distributions, relationships, and cluster results.

**Seaborn:**

* **Explanation:** Seaborn is built on top of Matplotlib and provides a high-level interface for drawing attractive and informative statistical graphics. It will be used for advanced data visualization to uncover patterns and insights.

**Scikit-learn:**

* **Explanation:** Scikit-learn is a powerful machine learning library in Python that provides simple and efficient tools for data mining and data analysis. It will be used for implementing the K-Means clustering algorithm to segment customers.

**Power BI:**

* **Explanation:** Power BI is a business analytics service by Microsoft that provides interactive visualizations and business intelligence capabilities. It will be used to create interactive dashboards to present the segmented customer data to stakeholders.

**Architecture**

**Data Preprocessing:**

* **Explanation:** This step involves cleaning and preparing the data for analysis. Tasks include handling missing values, encoding categorical variables, and normalizing numerical data to ensure it is suitable for clustering.

**Exploratory Data Analysis (EDA):**

* **Explanation:** EDA involves examining the data through statistical summaries and visualizations to understand its structure, distributions, and relationships. This step helps in identifying patterns and informing the feature selection for clustering.

**Clustering:**

* **Explanation:** Using the selected features (e.g., Age, Annual Income, Spending Score), the K-Means clustering algorithm will be applied to segment customers into distinct groups. The optimal number of clusters will be determined using methods like the Elbow Method.

**Visualization:**

* **Explanation:** Visualization is key to interpreting and presenting the results. Matplotlib and Seaborn will be used to create visualizations of the clusters, while Power BI will be used for interactive dashboards to share insights with stakeholders.

**Data Flow**

**Import Data:**

* **Explanation:** Load the Mall Customers dataset into the analysis environment using Python libraries like Pandas.

**Clean Data:**

* **Explanation:** Perform data cleaning tasks such as handling missing values, encoding categorical variables, and normalizing numerical features to prepare the data for analysis.

**Analyze Data:**

* **Explanation:** Conduct exploratory data analysis to understand the data distributions and relationships between variables. This step involves generating descriptive statistics and visualizations.

**Segment Customers:**

* **Explanation:** Apply the K-Means clustering algorithm to segment customers into distinct groups based on their features. Evaluate the quality of the clusters and interpret the results.

**Visualize Results:**

* **Explanation:** Create visualizations using Matplotlib and Seaborn to illustrate the clusters and their characteristics. Develop interactive dashboards in Power BI to present the segmented data and insights to stakeholders.

By following this structured approach and using these technologies, the project aims to achieve a comprehensive understanding of customer segments, which can then be leveraged to improve marketing strategies and overall business performance.