Technical Requirements Document (TRD)

Project Name	Customer Segmentation for retail store
Date Submitted	18-07-2024
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.

Data Sources:

- **Mall Customers Dataset:**
  - This dataset, named **Mall_Customers.csv**, contains detailed information about customers who visit the retail store. Key attributes include customer ID, gender, age, annual income, and spending score, among others. The data can be sourced from the store's internal systems or a publicly available dataset representative of mall customers. The dataset serves as the foundation for the customer segmentation analysis, providing the necessary attributes to understand customer behavior and preferences.

Technologies:

- **Python:**
  - Used as the primary programming language for data manipulation, analysis, and machine learning. Python's versatility and extensive libraries make it ideal for this project.
- **Google Colab:**
  - An online platform for running Jupyter notebooks, Google Colab provides a collaborative environment with access to powerful computational resources, facilitating efficient data analysis and model development.
- **Jupyter Notebook:**
  - An interactive environment for conducting and documenting data analysis and modeling steps. Jupyter Notebooks allow for easy visualization of data and results, making the analysis process more transparent and reproducible.
- **Pandas:**
  - A powerful data manipulation library in Python, Pandas is used for data cleaning, preprocessing, and analysis. It provides flexible data structures such as DataFrames, which are essential for handling structured data.
- **NumPy:**
  - A fundamental package for scientific computing in Python, NumPy is used for numerical operations and array manipulations, providing support for large multi-dimensional arrays and matrices.
- **Matplotlib:**

    - A plotting library for creating static, animated, and interactive visualizations in Python. Matplotlib is used to generate detailed plots that help in understanding the data distribution and relationships.
- **Seaborn:**
    - A statistical data visualization library based on Matplotlib, Seaborn is used for making complex plots easier. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Scikit-learn:**
    - A machine learning library in Python that provides simple and efficient tools for data mining and data analysis. It includes the K-Means clustering algorithm used for segmenting customers into distinct groups.
- **Power BI:**
    - A business analytics tool used to create interactive dashboards and reports. Power BI allows users to visualize and share insights from the customer segmentation analysis, making it accessible to stakeholders.

Architecture:

The project architecture follows a structured approach to handle data preprocessing, exploratory data analysis (EDA), customer segmentation, and visualization. The key components of the architecture include:

1. **Data Ingestion:**
    - Import data from various sources (e.g., CSV files, databases) into a Jupyter Notebook or Google Colab for analysis.
2. **Data Preprocessing:**
    - Clean the data by handling missing values, removing duplicates, and correcting any inconsistencies to ensure data quality.
    - Normalize or standardize numerical features to ensure they are on a comparable scale, which is essential for effective clustering.
    - Encode categorical variables to numerical formats if necessary to prepare the data for analysis.
3. **Exploratory Data Analysis (EDA):**
    - Perform descriptive statistics to understand the distribution and central tendencies of the data.
    - Use visualizations (e.g., histograms, box plots, scatter plots) to identify patterns, correlations, and anomalies within the data.
    - Formulate initial hypotheses about potential customer segments based on EDA findings.
4. **Customer Segmentation:**
    - Apply clustering algorithms, particularly K-Means, to group customers into distinct segments based on their purchasing behavior and other attributes.
    - Evaluate the optimal number of clusters using methods such as the elbow method or silhouette score to ensure meaningful and actionable segments.
    - Interpret the characteristics of each segment to understand the unique traits and behaviors of customers within each group.

5. **Visualization:**
    - Create static visualizations using Matplotlib and Seaborn to represent the characteristics of each customer segment. This includes cluster profiles, segment distributions, and key differentiators.
    - Develop interactive dashboards in Power BI to enable dynamic exploration of customer segments. These dashboards should allow users to filter and drill down into specific segments to gain deeper insights.

Data Flow:

The data flow diagram outlines the sequential steps taken to transform raw customer data into actionable insights through segmentation and visualization:

1. **Import Data:**
    - Load the customer dataset (Mall_Customers.csv) into the analysis environment (e.g., Jupyter Notebook, Google Colab).
2. **Clean Data:**
    - Preprocess the data to ensure quality and consistency, including handling missing values, removing duplicates, and normalizing numerical features.
3. **Analyze Data:**
    - Perform EDA to gain initial insights and understand the underlying patterns in the data. Use visualizations to explore data distributions and relationships.
4. **Segment Customers:**
    - Apply clustering algorithms, specifically K-Means, to segment the customers into distinct groups based on their purchasing behavior and other relevant attributes.
    - Validate the clustering results and interpret the characteristics of each segment.
5. **Visualize Results:**
    - Create visual representations of the customer segments using Matplotlib and Seaborn.
    - Develop interactive dashboards in Power BI to allow stakeholders to dynamically explore the segmentation results and derive actionable insights.

By leveraging these technologies and following a structured architecture, the project aims to deliver comprehensive customer segmentation insights that can drive targeted marketing strategies, improve customer satisfaction, and enhance sales performance.