**Technical Requirements Document**

**1. Data Sources**

* **Mall Customers Dataset: This dataset contains detailed information about customers who visit the retail store. Key attributes may include customer ID, gender, age, annual income, and spending score, among others. The data is sourced from the store's internal systems or a publicly available dataset representative of mall customers behavior.**

**2. Technologies**

* **Python: Used as the primary programming language for data manipulation, analysis, and machine learning.**
* **Jupyter Notebook: An interactive environment for conducting and documenting data analysis and modeling steps.**
* **Matplotlib: A plotting library for creating static, animated, and interactive visualizations in Python.**
* **Seaborn: A statistical data visualization library based on Matplotlib, used for making complex plots easier.**
* **Scikit-learn: A machine learning library in Python that provides simple and efficient tools for data mining and data analysis, particularly for clustering algorithms like K-Means.**
* **Power BI: A business analytics tool used to create interactive dashboards and reports, allowing users to visualize and share insights from the customer segmentation analysis.**

**3. Architecture**

**3.1 Data Preprocessing**

* **Handle Missing Values: Identify and handle any missing values in the dataset.**
* **Remove Duplicates: Ensure there are no duplicate records in the dataset.**
* **Encode Categorical Variables: Convert categorical variables (e.g., Genre) into numerical format.**
* **Scale Numerical Features: Standardize or normalize numerical features to prepare them for clustering algorithms.**

**3.2 Exploratory Data Analysis (EDA)**

* **Distribution Analysis: Analyze the distribution of each feature (e.g., histograms, box plots).**
* **Correlation Analysis: Identify correlations between variables using techniques such as correlation matrices.**
* **Outlier Detection: Detect and analyze outliers and anomalies in the data.**

**3.3 Clustering**

* **Unsupervised Learning Algorithms: Apply clustering algorithms such as K-Means and DBSCAN to segment customers.**
* **Optimal Number of Clusters: Determine the optimal number of clusters using methods like the elbow method and silhouette analysis.**

**3.4 Visualization**

* **Informative Plots and Charts: Create visualizations to represent customer segments (e.g., scatter plots, cluster heatmaps).**
* **Integration with Power BI: Integrate visualizations into a Power BI dashboard for interactive exploration and presentation.**

**4. Data Flow**

1. **Import Data: Load the Mall Customers dataset into a Jupyter Notebook environment.**
2. **Clean Data:**
   * **Handle missing values.**
   * **Remove duplicate records.**
   * **Encode categorical variables.**
   * **Scale numerical features.**
3. **Analyze Data:**
   * **Perform exploratory data analysis (EDA) to understand the distribution of features.**
   * **Identify correlations between variables.**
4. **Segment Customers:**
   * **Apply unsupervised learning algorithms (e.g., K-Means, DBSCAN) to segment the customers into distinct groups.**
   * **Determine the optimal number of clusters.**
5. **Visualize Results:**
   * **Create informative plots and charts using Matplotlib and Seaborn.**
   * **Integrate the visualizations into a Power BI dashboard.**

# **Project Plan**

|  |  |
| --- | --- |
| **Project Name** | Customer Segmentation for retail store |
| **Date Submitted** | 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. |

**Tasks:**

1. **Data Collection:**
   * Identify and gather the necessary data sources, including the **Mall\_Customers.csv** dataset.
   * Import **Mall\_Customers.csv** into the analysis environment
   * Ensure the data is available in a format suitable for analysis, such as CSV or database export.
   * Verify all 200 records are present.
   * Confirm the following columns are available: CustomerID, Genre, Age, Annual Income (k$), Spending Score (1-100)
   * Example:

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)

1. **Data Cleaning:**
   * Handle missing values by imputing or removing them based on the context and relevance.
   * Verify data types (e.g., ensure Age and Annual Income are numeric)
   * Detect and remove duplicate records to maintain data integrity.
   * Correct any inconsistencies in the data, such as outliers or incorrect data entries.
   * Normalize or standardize numerical features to ensure they are on a comparable scale.
   * Example:

count = data.isnull().sum()

mean\_age = data['Age'].mean()

data["Age"].fillna(mean\_age, inplace=True)

data.head(10)

# Renaming columns for better readability

data.columns = ["CustomerID", "Gender", "Age", "AnnualIncome", "SpendingScore"]

data

mode\_gender = data['Gender'].mode()[0]

data.dropna(inplace=True)

data["Gender"].fillna(mode\_gender, inplace=True)

data.head(20)

count = data.isnull().sum()

# Data transformation (e.g., encoding categorical variables)

data['Gender'] = data['Gender'].map({'Male': 0, 'Female': 1})

count = data.isnull().sum()

data

1. **Exploratory Data Analysis (EDA):**
   * Conduct descriptive statistical analysis to summarize the main features of the dataset.
   * Use visualizations (e.g., histograms, scatter plots, box plots) to explore data distributions and relationships.
   * Identify key trends, patterns, and anomalies within the data.
   * Formulate hypotheses and potential segments based on initial findings.
   * Example:

data.describe()

1. **Clustering:**
   * Select appropriate clustering algorithms (e.g., K-Means) for customer segmentation.
   * K-means Clustering:
     + Determine optimal number of clusters using elbow method and silhouette analysis.
     + Apply K-means algorithm with optimal k
     + Evaluate cluster quality using silhouette score
   * Perform clustering on the dataset to group customers into distinct segments.
   * Validate the clustering results to ensure meaningful and actionable segments.
   * Example:

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler

# Feature selection

features = data[['Age', 'AnnualIncome', 'SpendingScore']]

# Standardizing the features

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(features)

# Applying K-Means clustering

kmeans = KMeans(n\_clusters=5, random\_state=42)

data['Cluster'] = kmeans.fit\_predict(scaled\_features)

# Evaluating cluster quality

import matplotlib.pyplot as plt

import seaborn as sns

plt.figure(figsize=(10, 6))

sns.scatterplot(data=data, x='AnnualIncome', y='SpendingScore', hue='Cluster', palette='viridis')

plt.title('Customer Segments')

plt.show()

1. **Visualization:**
   * Create visualizations using Matplotlib and Seaborn to represent the characteristics of each customer segment.
   * Develop interactive dashboards in Power BI to allow stakeholders to explore segmentation results dynamically.
   * Ensure visualizations are clear, insightful, and tailored to the needs of the audience.
   * Example:

# Visualizing distributions

plt.figure(figsize=(10, 6))

sns.histplot(data['Age'], bins=30, kde=True)

plt.title('Age Distribution')

plt.show()

plt.figure(figsize=(10, 6))

sns.histplot(data['AnnualIncome'], bins=30, kde=True)

plt.title('Annual Income Distribution')

plt.show()

plt.figure(figsize=(10, 6))

sns.histplot(data['SpendingScore'], bins=30, kde=True)

plt.title('Spending Score Distribution')

plt.show()

# Visualizing relationships

plt.figure(figsize=(10, 6))

sns.scatterplot(data=data, x='AnnualIncome', y='SpendingScore', hue='Gender')

plt.title('Income vs Spending Score')

plt.show()

1. **Documentation:**

* Technical Report:
  + Detailed methodology
  + Code documentation
  + Statistical analysis results
  + Cluster evaluation metrics
* Executive Summary:
  + Key findings and insights
  + Visualizations of customer segments
  + Actionable recommendations
* Document all steps taken during the data collection, cleaning, analysis, and clustering processes.
* Provide recommendations based on the segmentation results and suggest future steps for the retail store.

**Timeline:**

**Day 1:** **Project setup and data import**

* Data Collection: Identify and gather data sources.
* Initial Data Exploration: Conduct preliminary analysis to understand data structure.
* Data Cleaning: Handle missing values, remove duplicates, and correct inconsistencies.
* Normalization/Standardization: Prepare numerical features for analysis.

**Day 2:** **Complete EDA**

* Exploratory Data Analysis (EDA): Perform descriptive statistics and initial visualizations.
* Prepare data for clustering (normalization, encoding).
* Hypothesis Formulation: Identify potential customer segments.
* Clustering: Apply K-Means and determine the optimal number of clusters.
* Validate Clustering: Assess the meaningfulness of the clusters.

**Day 3:** **Create EDA visualizations**

* Visualization: Create static visualizations using Matplotlib and Seaborn.
* Dashboard Development: Develop interactive Power BI dashboards.
* Documentation: Compile a comprehensive report with insights, visualizations, and recommendations.
* Review and Finalize: Revise and finalize documentation for presentation.

1. **Resources:**

* **Datasets:**
  + Mall customers dataset (primary data source).
* **Software and Tools:**
  + Python (for data manipulation and analysis).
  + Google Colab (for interactive data analysis).
  + Jupyter Notebook (for interactive data analysis).
  + Matplotlib and Seaborn (for static visualizations).
  + Scikit-learn (for clustering algorithms).
  + Power BI (for interactive dashboards).
* **Hardware**:
  + Computers with minimum 16GB RAM, quad-core processor
  + Cloud computing resources (e.g., AWS, GCP) for scalability if needed
* **Human Resources:**
  + Data Analyst/Scientist: Responsible for data cleaning, EDA, and clustering.
  + Visualization Specialist: Responsible for creating visualizations and dashboards.
  + Project Manager: Overseeing the project timeline, tasks, and deliverables.
* **Other Resources:**
  + Industry reports on retail customer behavior
  + Company historical data for context (if available)
  + Documentation and training materials for Power BI users.

1. **Risks:**

* **Data Quality Issues**:
  + Risk: Incomplete or inaccurate data can lead to misleading analysis and segmentation results.
  + Mitigation:
    - Thorough data cleaning process with documentation
    - Regular data quality checks throughout the analysis
    - Consider multiple imputation techniques for missing data
* **Algorithm Performance**:
  + Risk: The clustering algorithm may not perform well with the given data, leading to suboptimal segments.
  + Mitigation:
    - Test multiple clustering algorithms (K-means, hierarchical, DBSCAN)
    - Use various evaluation metrics (silhouette score, Calinski-Harabasz index)
  + Conduct sensitivity analysis on algorithm parameters
* **Visualization Limitations**:
  + Risk: Static visualizations may not effectively convey insights to stakeholders.
  + Mitigation:
    - Develop interactive and user-friendly dashboards in Power BI for better engagement and understanding.
* **Time Constraints:**
  + Risk: Complex analysis may exceed the planned timeline
  + Mitigation:
    - Build buffer time into each phase of the project
    - Prioritize core analyses and mark advanced techniques as optional
    - Hold weekly progress reviews to identify and address delays early

These visualizations will help communicate the findings effectively to both technical and non-technical stakeholders. The combination of static and interactive visualizations will allow for both quick insights and in-depth exploration of the customer segments.

By addressing these tasks, timeline, resources, and potential risks, the project plan ensures a structured and comprehensive approach to achieving the objectives of customer segmentation for the retail store.