Project Plan

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.

Tasks:

1. **Data Collection:**
   - Identify and gather the necessary data sources, including the **Mall_Customers.csv** dataset.
   - Ensure the data is available in a format suitable for analysis, such as CSV or database export.
   - 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)
```

2. **Data Cleaning:**
   - Handle missing values by imputing or removing them based on the context and relevance.
   - 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
```

3. **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()
```

4. **Clustering:**
   - Select appropriate clustering algorithms (e.g., K-Means) for customer segmentation.
   - Determine the optimal number of clusters using methods like the elbow method or 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()
```

5. **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()
```

6. **Documentation:**
    - Compile a comprehensive report detailing the entire project, including objectives, methodology, results, and conclusions.
    - Document all steps taken during the data collection, cleaning, analysis, and clustering processes.
    - Include visualizations, charts, and insights derived from the analysis.
    - Provide recommendations based on the segmentation results and suggest future steps for the retail store.

Timeline:

Day 1:

- 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:

- Exploratory Data Analysis (EDA): Perform descriptive statistics and initial visualizations.
- 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:

- 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.

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).
- **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:**
  - Access to a high-performance computing environment (if dealing with large datasets).
  - Documentation and training materials for Power BI users.

Risks:

- **Data Quality Issues:**
  - Incomplete or inaccurate data can lead to misleading analysis and segmentation results.
  - **Mitigation:** Conduct thorough data cleaning and validation before analysis.
- **Algorithm Performance:**
  - The clustering algorithm may not perform well with the given data, leading to suboptimal segments.
  - **Mitigation:** Experiment with different clustering techniques and parameter tuning to achieve the best results.
- **Visualization Limitations:**

- Static visualizations may not effectively convey insights to stakeholders.
- **Mitigation:** Develop interactive and user-friendly dashboards in Power BI for better engagement and understanding.

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.