

Project Plan

Project Name	Customer Segmentation for retail store
Date Submitted	22-07-2024
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:

```
1 import pandas as pd
2
3 # Load the dataset
4 file_path = 'Mall_Customers.csv'
5 data = pd.read_csv(file_path)
6
7 # Display the first few rows of the dataset
8 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:

```
1 count = data.isnull().sum()
2 mean_age = data['Age'].mean()
3 data["Age"].fillna(mean_age, inplace=True)
4 data.head(10)
5
6 # Renaming columns for better readability
7 data.columns = ["CustomerID", "Gender", "Age", "AnnualIncome", "SpendingScore"]
8 data
9
10 mode_gender = data['Gender'].mode()[0]
11 data.dropna(inplace=True)
12 data["Gender"].fillna(mode_gender, inplace=True)
13
14 data.head(20)
15 count = data.isnull().sum()
16
17 # Data transformation (e.g., encoding categorical variables)
18 data['Gender'] = data['Gender'].map({'Male': 0, 'Female': 1})
```

```
19 count = data.isnull().sum()
20 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:

```
1 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:

```
1 from sklearn.cluster import KMeans
2 from sklearn.preprocessing import StandardScaler
3
4 # Feature selection
5 features = data[['Age', 'AnnualIncome', 'SpendingScore']]
6
7 # Standardizing the features
8 scaler = StandardScaler()
9 scaled_features = scaler.fit_transform(features)
10
11 # Applying K-Means clustering
12 kmeans = KMeans(n_clusters=5, random_state=42)
13 data['Cluster'] = kmeans.fit_predict(scaled_features)
14
15 # Evaluating cluster quality
16 import matplotlib.pyplot as plt
17 import seaborn as sns
18
19 plt.figure(figsize=(10, 6))
20 sns.scatterplot(data=data, x='AnnualIncome', y='SpendingScore', hue='Cluster', palette='viridis')
21 plt.title('Customer Segments')
22 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:

```
1 # Visualizing distributions
2 plt.figure(figsize=(10, 6))
3 sns.histplot(data['Age'], bins=30, kde=True)
4 plt.title('Age Distribution')
5 plt.show()
6
7 plt.figure(figsize=(10, 6))
```

```

8 sns.histplot(data['AnnualIncome'], bins=30, kde=True)
9 plt.title('Annual Income Distribution')
10 plt.show()
11
12 plt.figure(figsize=(10, 6))
13 sns.histplot(data['SpendingScore'], bins=30, kde=True)
14 plt.title('Spending Score Distribution')
15 plt.show()
16
17 # Visualizing relationships
18 plt.figure(figsize=(10, 6))
19 sns.scatterplot(data=data, x='AnnualIncome', y='SpendingScore', hue='Gender')
20 plt.title('Income vs Spending Score')
21 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.