Customer Segmentation Project Report

1.Introduction:

This project aims to perform customer segmentation using the KMeans clustering algorithm on a dataset containing information about customers. The dataset includes features such as age, income, gender, marital status, occupation, and settlement size. The goal is to group customers based on age and income to identify distinct segments within the customer base.

2.Data Exploration and Cleaning:

Loaded the dataset and conducted initial exploration.

Checked the dataset's shape and dropped the 'ID' column.

Explored summary statistics, data types, missing values, and duplicates.

Utilized visualizations to gain insights into the distribution of age, income, gender, marital status, and occupation.

3.Data Visualization:

Plotted histograms to visualize the distribution of age and income.

Used count plots to display the distribution of gender and marital status.

Created a scatter plot to explore the relationship between age and income.

Utilized box plots to understand income distribution across different occupations.

Constructed a correlation heatmap to examine correlations between numerical features.

4. KMeans Clustering Model:

Applied KMeans clustering to segment customers based on age and income.

Utilized a pipeline for preprocessing and model building.

Defined a parameter grid for hyperparameter tuning using GridSearchCV.

Evaluated the model using Silhouette Score and Calinski-Harabasz Index.

Achieved the best model from the pipeline using GridSearchCV.

5. Model Evaluation:

Evaluated the model using Silhouette Score and Calinski-Harabasz Index to assess clustering quality.

6.Results and Visualizations:

Visualized the identified clusters and centroids on a scatter plot.

Presented the clusters in different colors and marked centroids with black stars.

7.Conclusion:

Successfully implemented customer segmentation using the KMeans clustering algorithm.

Identified distinct customer segments based on age and income.

Evaluated the model using silhouette and calinski-harabasz indices to measure clustering quality.