1. Background

The dataset used in this analysis is called the '**Mall Customer Segmentation Data**' obtained from Kaggle. This dataset includes features such as 'Customer ID','Gender''Age', 'Annual Income ', and 'Spending Score (1-100)' of customers visiting a particular mall. The goal of this analysis is to perform clustering on the dataset based on the features like 'Gender''Age', 'Annual Income ', and 'Spending Score (1-100)' to group similar customers in the same clusters .

2. Methods

The first step in the analysis is to load the data and perform some initial exploratory data analysis. This includes printing the first 10 rows of the dataset and a summary of the dataset.

Next, the 'Gender' column is transformed into two binary columns 'Gender_Female' and 'Gender_Male' using one-hot encoding. The dataset is then checked for missing values.The columns 'Age', 'Annual Income', and 'spending Score' only had missing values.

The selected columns(Columns with missing values) for the analysis are the second to fourth columns of the dataset. Boxplots are created for these columns to identify any outliers. The result shows that there are no outliers in any of the columns except in 'Annual Income'.

Missing values of the columns with no outliers i.e, 'Age' and 'Spending Score' are filled with the mean of the column and missing values for the 'Annual Income' column have been filled with the median of the column.

The features selected for the clustering analysis are 'Age', 'Annual Income', and 'Spending Score (1-100)', 'Gender_Female', and 'Gender_Male'. These features are then scaled using the StandardScaler.

The Elbow Method is used to determine the optimal number of clusters for the KMeans algorithm. The Within-Cluster-Sum-of-Squares (WCSS) is calculated for different numbers of clusters ranging from 1 to 10, and the 'elbow point' in the plot of WCSS versus the number of clusters is identified as the optimal number of clusters. The optimal number of clusters comes out to be 4.

Finally, the KMeans algorithm is applied with the optimal number of clusters(K = 4), and the cluster labels are added to the original dataset.

3. Results

The optimal number of clusters identified by the Elbow Method is printed and Cluster Labels are produced

4. Conclusions

The clustering analysis successfully identifies groups within the data. The visualizations provide insight into the characteristics of these clusters in terms of the selected features.