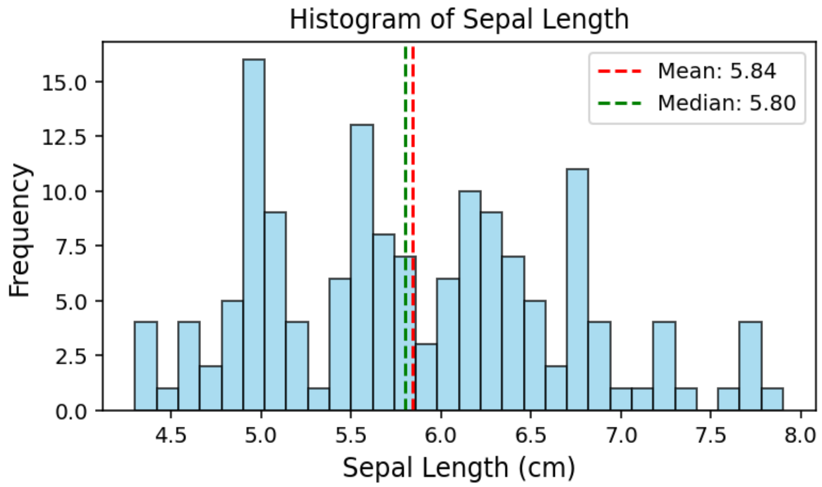
ANALYSIS OF IRIS DATASET

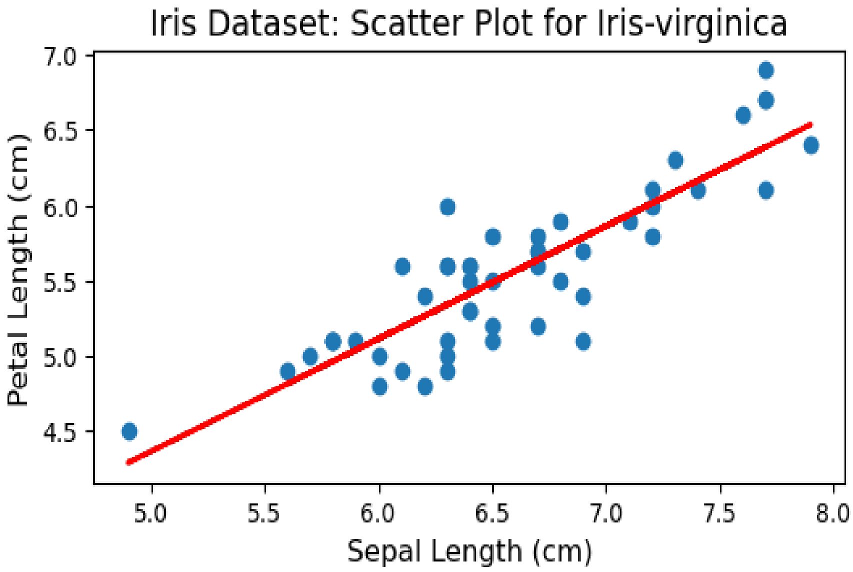
Student ID: 23082951

GitHub Link: https://github.com/Sanmaria21/Clustering-and-Fitting

Introduction: This dataset is taken from Kaggle. Iris setosa, Iris versicolor, and Iris virginica are the 3 species used in this dataset. There are150 observations of the iris and each observation consist of 4 features sepal length, petal length, sepal width, and petal width.



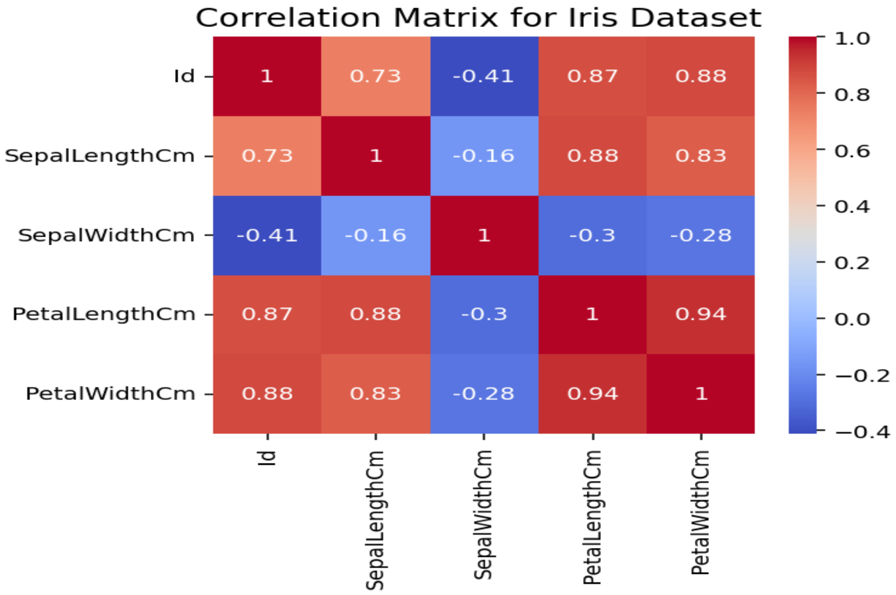
Histogram Plot: This plot illustrates the distribution of the sepal length of the iris dataset. It is evident from the graph that the distribution is close to normal with a slight positive skewness. The mean and median value of the sepal length is **5.84**cm and **5.80**cm. The standard deviation value is **0. 83**cm. The skewness value is **0.31** and the kurtosis value is **–0.55.**



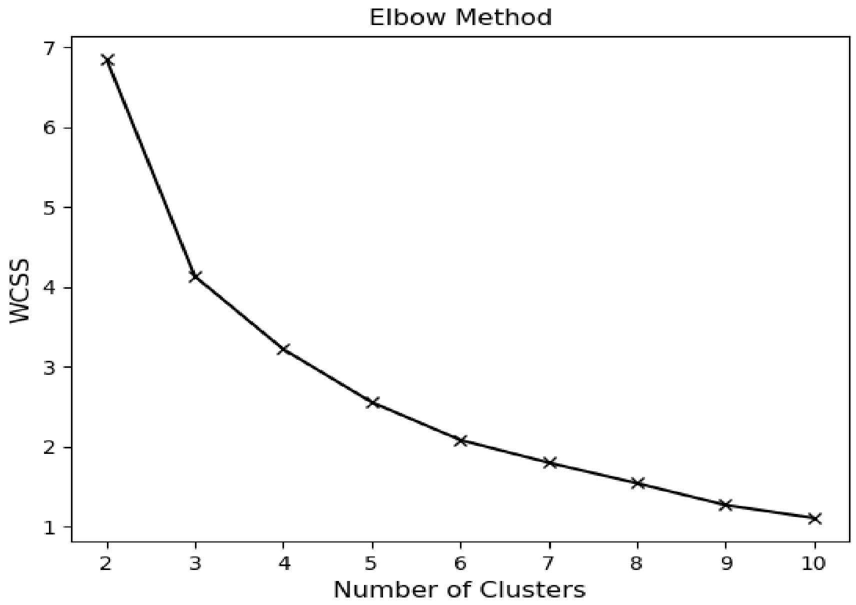
Scatter Plot- Cluster with Line Fitting: The bonding between sepal length and petal length of Iris-virginica is demonstrated in this graph.It indicates a strong positive correlation as the data points are clustered around the line fitting. A linear relationship between the two features is observable in the graph as both increases simultaneously.

Correlation Matrix of Iris Dataset:

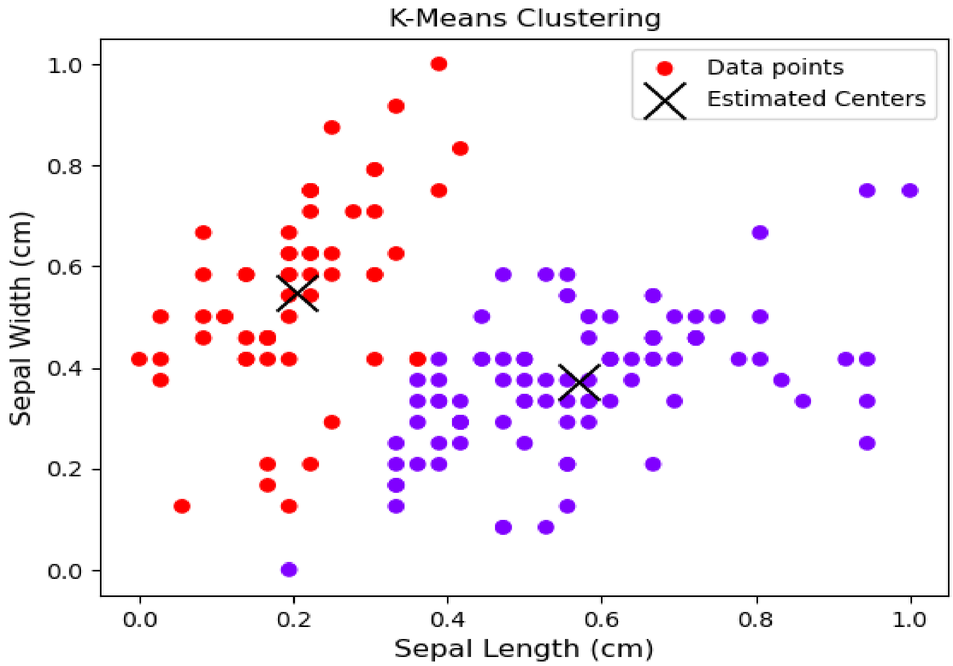
The given correlation matrix demonstrates a strong positive connection between petal length and petal width (**0.94**), and between sepal length and both petal width (**0.88**) and petal length (**0.87**). A moderate positive relationship of value **0.73** is observed between sepal length and sepal width. The heatmap shows a negative value with petal length of **–0.41** and petal width of **–0.3** illustrating a weak relationship between them



Elbow Method and K-means Clustering:



This graph illustrates the best number of clusters in K-mean clustering by applying elbow method. The number of clusters increases when the WCSS decreases. A slowdown in rate of decrease is observed after a certain point as result an "elbow" shape is formed. In this figure, the elbow appears to be around 3 or 4 clusters. This means that a good balance between separation of the data and compactness can be provided using 3 or 4 clusters.



K-mean clustering:

The graph represents the analysis of K-means clustering applied to iris dataset by visualising two features (sepal length and sepal width). The two different colors (blue and red) the data points that are divided into two distinct clusters. The centers of each cluster (estimated centers) are denoted by the symbol " X". It is clear from the graph that the two clusters are separated well, indicating that the K-means algorithm has accurately identified two distinct groups within the data.

Conclusion:

This report highlights the clustering and statistical properties of the data by proper analysis of the iris dataset. From the histogram, the distribution of sepal length is observed along with all its major moments: mean, median, standard deviation, skewness and kurtosis. From the scatter plot,a strong linear relationship is observed between sepal length and petal length of the Iris-virginica species. The heatmap demonstrates a correlation between various features used in the dataset.

To determine the best number of clusters, the elbow method is applied. The dataset was successfully divided into two distinct groups showing the accuracy of K-mean algorithm in identifying the natural patterns in data. These outcomes indicates that these clustering techniques and statistical analysis can be utilized for interpreting and exploring complex dataset.