Name: Adnan Naseem Student ID: 23099739

Clustering and Fitting

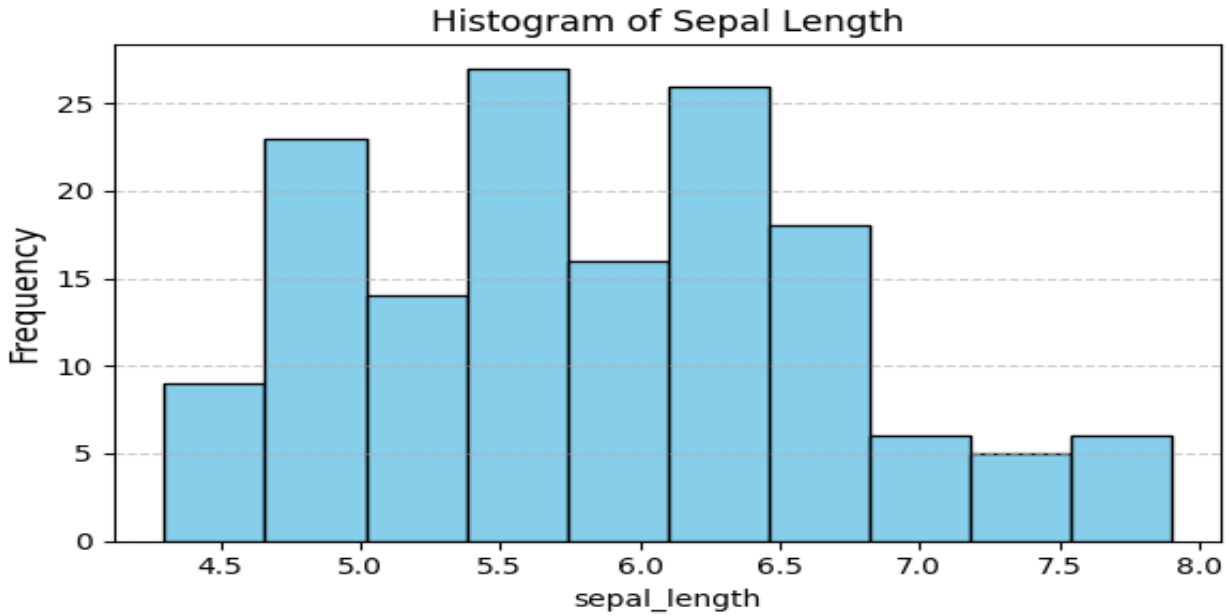
https://github.com/aadnon822/Clustering-and-Fitting/blob/main/Clustering%20and%20Fitting_23099739.ipynb

Introduction

I am using a famous dataset of flower species, named iris dataset. This dataset has 150 samples with 4 variables. All these belong to just 3 species of flowers. I will apply clustering and regression fitting techniques to find out different patterns and any relationships among these variables.

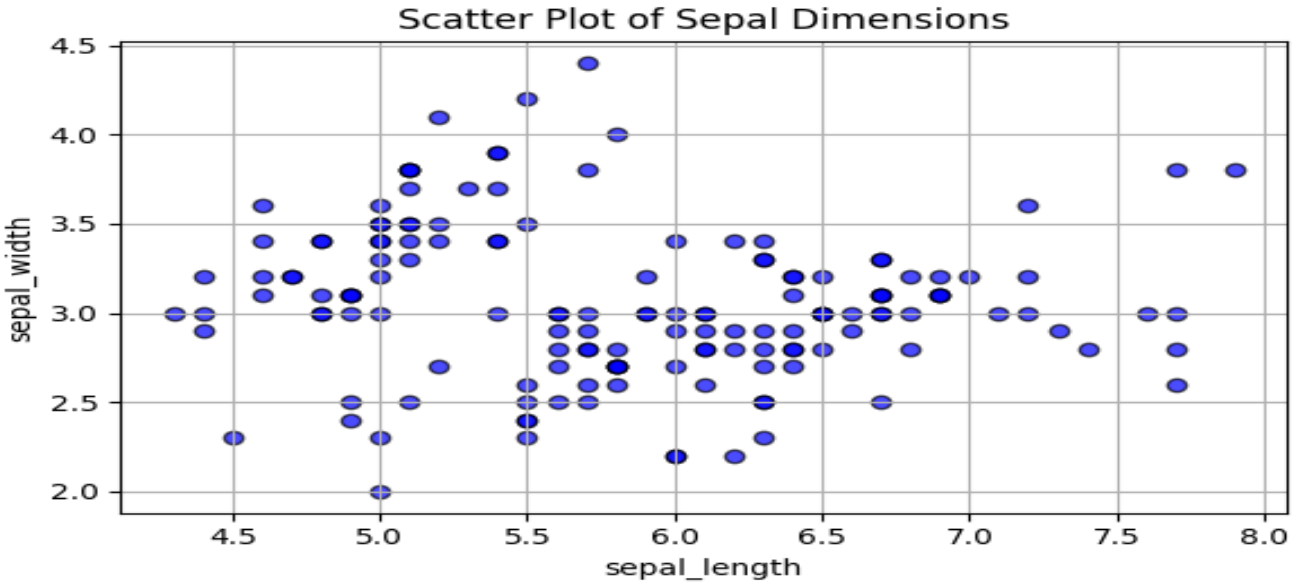
Histogram (Sepal_Length):

The first graph is histogram to explore the distribution of length of sepals. The graph shows that utmost values reside around 5-6 cm which make it approximately normal distribution. The Skewness: 0.31 and Kurtosis: -0.57 also confirm adequate peak and negligible asymmetry.



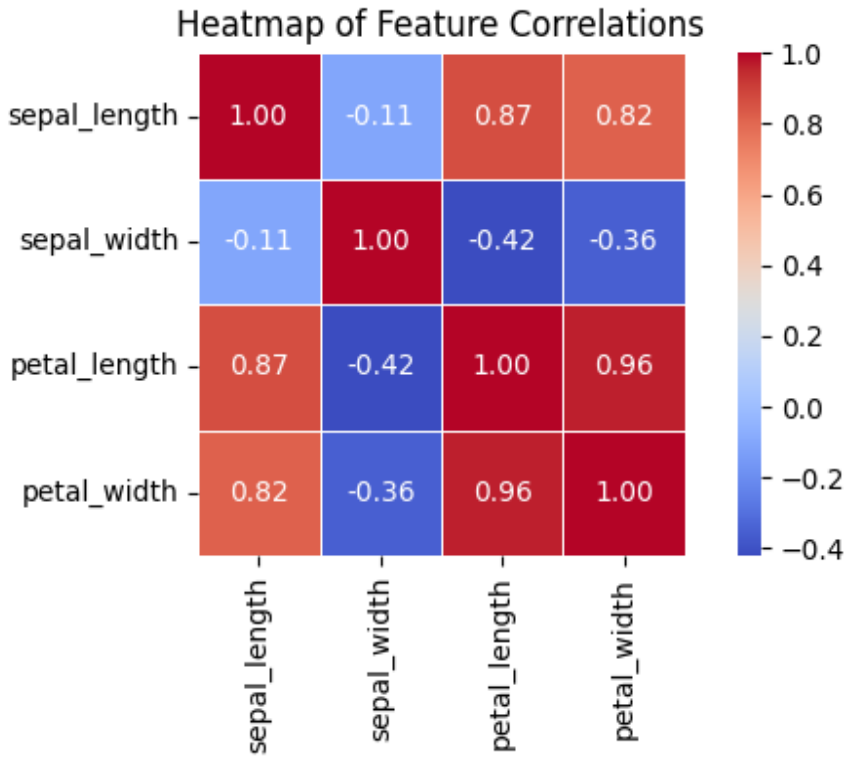
Scatter Plot (Impressions vs. Total Conversion):

This is a relational graph to show the relationship between sepal_length and sepal_width. There is no clear correlation however; there are clear clustering patterns which provides a way to explore further. The groups or clusters similar to the classification of species, which shows a conceivable separation based on these features.



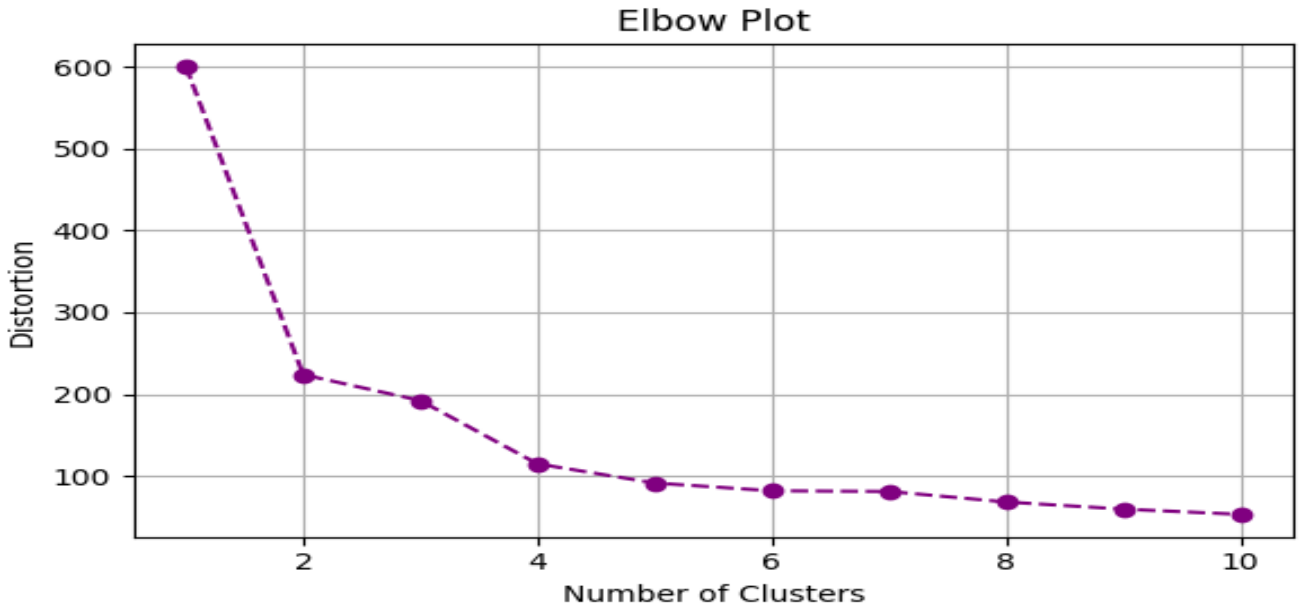
Heatmap:

This is the heatmap to show the correlation matrix of numerical variables, showing important relationship. It can be observed that petals are showing strong +ive correlation whereas; the sepals are showing weak correlation. This awareness leads us to feature selection for clustering and apply regression techniques.



Elbow Plot:

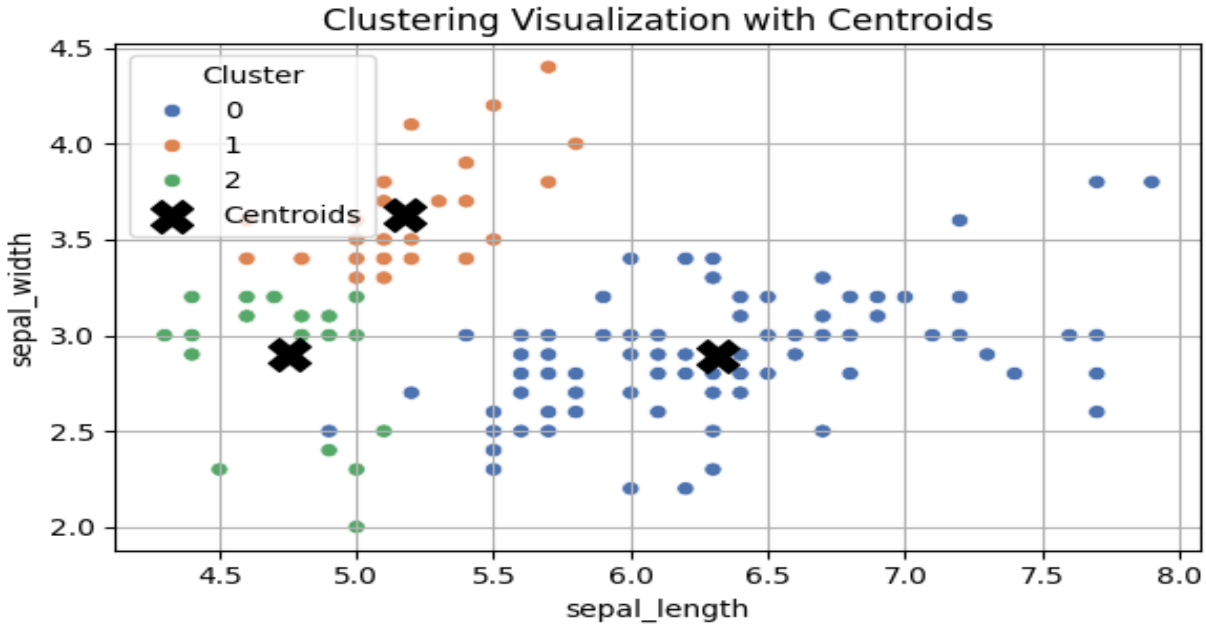
The purpose of this plot is to understand the optimal number of cluster for this dataset. Using the Elbow Method, distortions were gauged for one to ten clusters, and the elbow indicated 3 number of clusters which is aligned to number of species of flowers.



Name: Adnan Naseem Student ID: 23099739

K-Means Clustering:

The dataset is dissection into three groups with the help of K-Means clustering across sepal length and sepal width. Centroids are also placed on each cluster for better visibility and understanding. The clusters are also supporting with the classification of flower species. It further validates that the model is effective.



Linear Regression:

I have used the linear regression to model the sepal dimensions of the flower species. It is observed a weak positive correlation between sepal dimensions. The predictive analysis is also completed on sepal_length=8.00 and sepal_width=2.93 emphasized in red. Although, there is not a strong correlation between sepal dimensions however, this model provides a foundational insight of these variables.