**Data Analysis and Clustering Report**  
**GitHub Repository:** [Insert link to your repository]

**Introduction**

This report explores the Iris dataset, applying clustering and regression techniques to uncover patterns and relationships among the features. The analysis includes visualizations such as histograms, scatter plots, a correlation matrix heatmap, and an elbow plot. We also assess the performance of the k-means clustering algorithm and linear regression model.

**Analysis and Visualization**

**Scatter Plot: Sepal Length vs. Petal Length**

A scatter plot highlights the relationship between sepal\_length and petal\_length. The strong positive trend indicates that flowers with larger sepals tend to have longer petals, suggesting a high correlation.

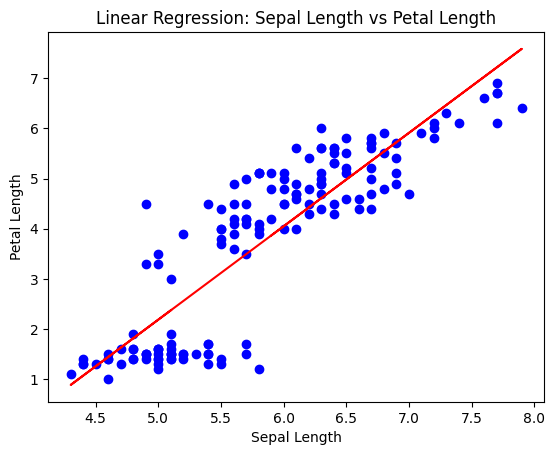


Figure 1Scatter plot illustrating the relationship between Sepal Length and Petal Length.

**Correlation with Matrix Heatmap**

The correlation matrix heatmap provides insights into relationships between all numerical features. High positive correlations are observed between petal\_length and petal\_width (0.96), as well as between sepal\_length and petal\_length (0.87). Conversely, sepal\_width has weaker correlations with the other features.

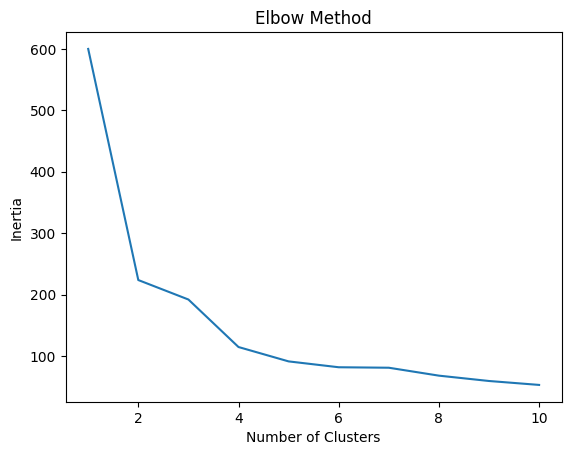
A screenshot of a graph

Description automatically generated

Figure 2Heatmap of the correlation matrix between numerical features.

**K-means Clustering**

Using the elbow method, we determined the optimal number of clusters for k-means clustering to be three. The silhouette score of **0.478** indicates moderate clustering performance. The clustering visually separates the data into distinct groups based on sepal\_length and petal\_length.

A diagram of a clustering graph

Description automatically generated

*Caption: Elbow plot showing the optimal number of clusters based on inertia.*

*Caption: Scatter plot of k-means clustering results, showing three distinct clusters.*

**Linear Regression**

Linear regression was applied to model the relationship between sepal\_length (independent variable) and petal\_length (dependent variable

**Box Plot of Sepal Length by Species**

The box plot shows the distribution of sepal\_length for each species. While there is some overlap, the distributions for Setosa, Versicolor, and Virginica exhibit distinct median values and ranges, aiding species differentiation.

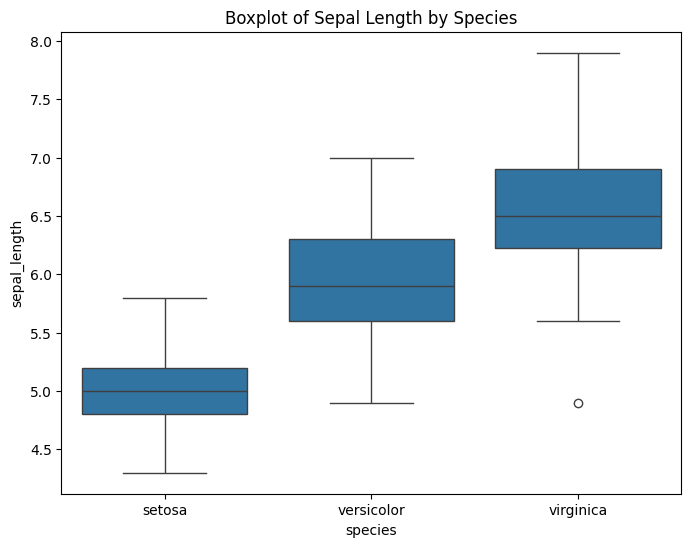


Figure 3Box plot of Sepal Length grouped by Species.

**Conclusions**

1. **Distribution Insights**: Sepal\_length and petal\_length show strong relationships and clear trends in the dataset.
2. **Clustering Results**: K-means clustering effectively groups data into three clusters, consistent with the three species of Iris.
3. **Correlation Analysis**: Strong correlations between petal-related features highlight the interdependence of flower dimensions.
4. **Linear Regression**: A strong linear relationship exists between sepal\_length and petal\_length, suggesting predictability between these features.

This analysis demonstrates a structured approach to data exploration, clustering, and regression, providing insights into the Iris dataset’s features and relationships.