IRIS Dataset - Visualization and Analysis Report

Analytics and Systems of Big Data

Thallapally Nimisha
CS22B1082
B.Tech in Computer Science and Engineering
IIITDM Kancheepuram

Contents

Dataset Description

The IRIS dataset contains measurements of 150 iris flowers from three species: *Setosa*, *Versicolor*, and *Virginica*. Each sample includes:

- Sepal Length (cm)
- Sepal Width (cm)
- Petal Length (cm)
- Petal Width (cm)
- Species (target class)

Libraries and Packages Used

- `pandas` – for data manipulation and analysis
- `numpy` – for numerical operations
- `matplotlib.pyplot` – for static plotting
- `seaborn` – for statistical data visualization
- `plotly.express` – for interactive charts (e.g., treemaps)

1 Q1: Visualization Techniques

A subset of the IRIS dataset attributes were used to create the following plots using Python and Matplotlib.

1.1 Bar Chart - Mean Sepal Length per Species

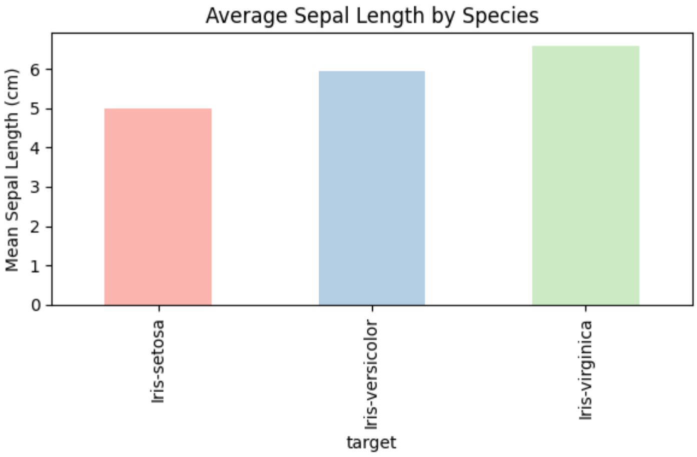
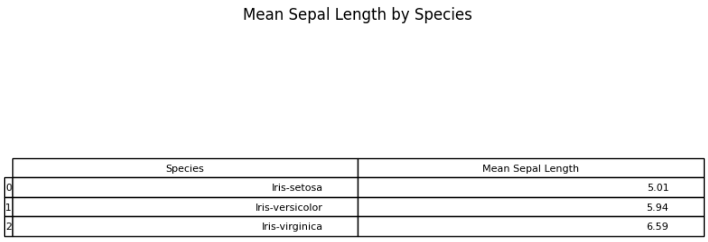


Figure 1: Average Sepal Length by Species



Mean Sepal Length by Species

	Species	Mean Sepal Length
0	Iris-setosa	5.01
1	Iris-versicolor	5.94
2	Iris-virginica	6.59

Figure 2: Average Sepal Length by Species

1.2 Pie Chart - Species Distribution

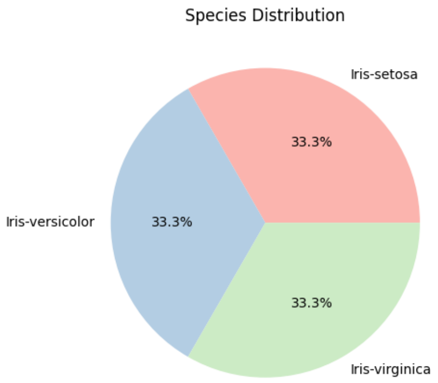
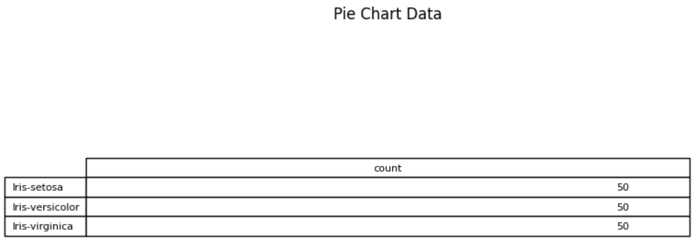


Figure 3: Pie Chart of Species Distribution



Pie Chart Data

	count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

Figure 4: Species Distribution

1.3 Doughnut Chart - Species Distribution

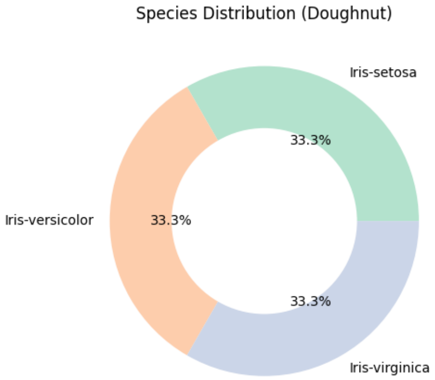


Figure 5: Doughnut Chart of Species Distribution

1.4 Pareto Chart

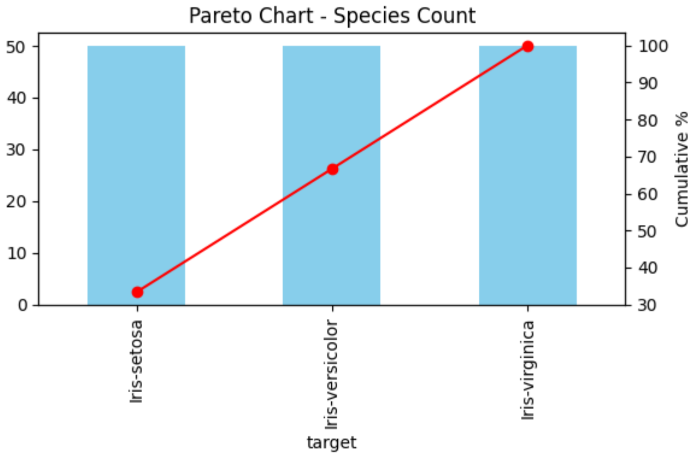
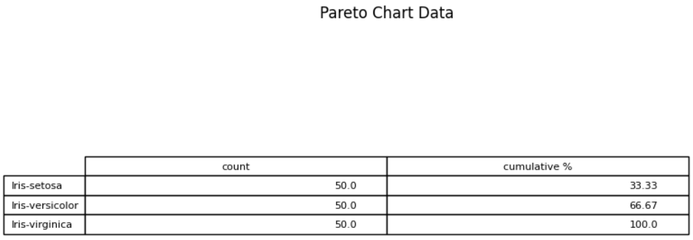


Figure 6: Pareto Chart of Species Count



Pareto Chart Data

	count	cumulative %
Iris-setosa	50.0	33.33
Iris-versicolor	50.0	66.67
Iris-virginica	50.0	100.0

Figure 7: Pareto Chart - Species Count

1.5 Scatter Plot - Sepal Length vs Width

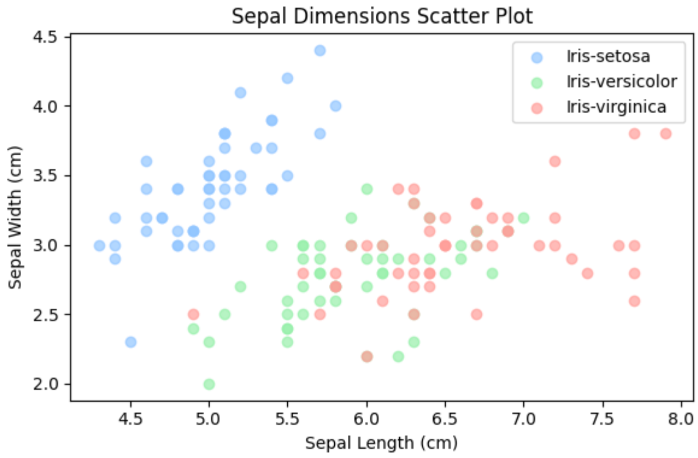


Figure 8: Scatter Plot: Sepal Length vs Sepal Width

1.6 Line Chart - Feature Means by Species

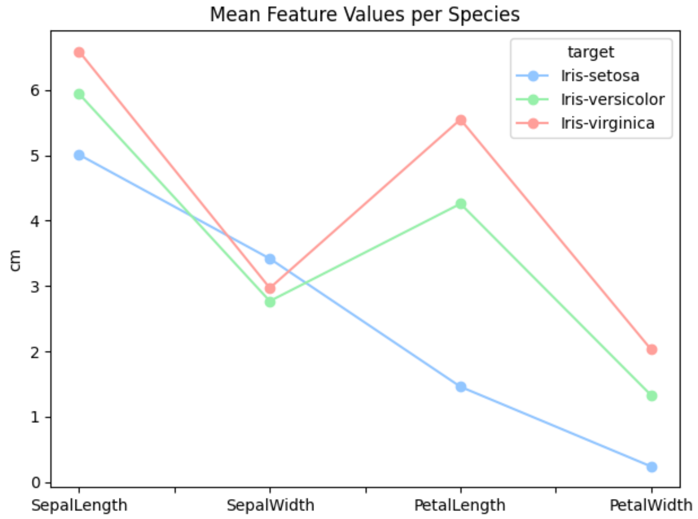
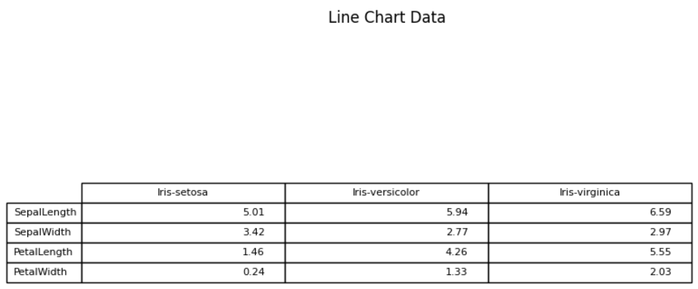


Figure 9: Line Chart: Mean Feature Values per Species



Line Chart Data

	Iris-setosa	Iris-versicolor	Iris-virginica
SepalLength	5.01	5.94	6.59
SepalWidth	3.42	2.77	2.97
PetalLength	1.46	4.26	5.55
PetalWidth	0.24	1.33	2.03

Figure 10: Mean feature value

1.7 Radar Chart

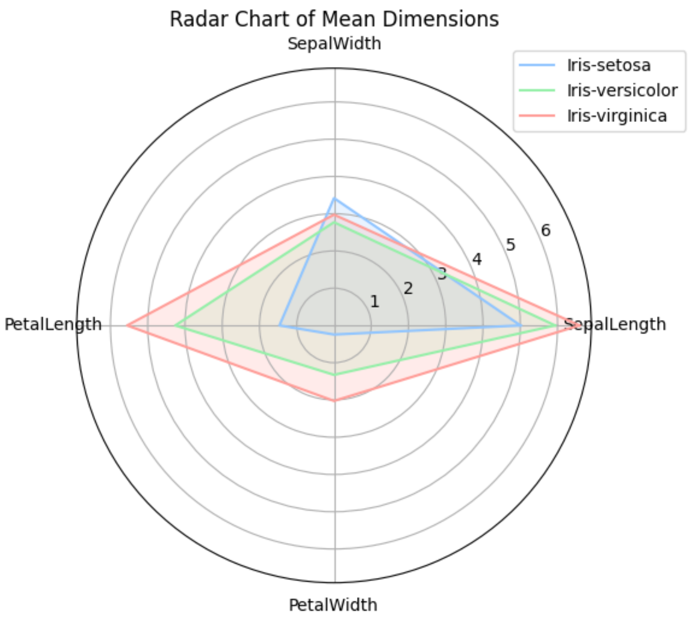
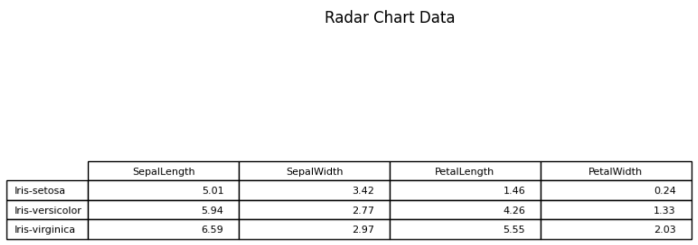


Figure 11: Radar Chart: Average Dimensions per Species



Radar Chart Data

	SepalLength	SepalWidth	PetalLength	PetalWidth
Iris-setosa	5.01	3.42	1.46	0.24
Iris-versicolor	5.94	2.77	4.26	1.33
Iris-virginica	6.59	2.97	5.55	2.03

Figure 12: Average Dimensions per Speciess

1.8 Area Chart

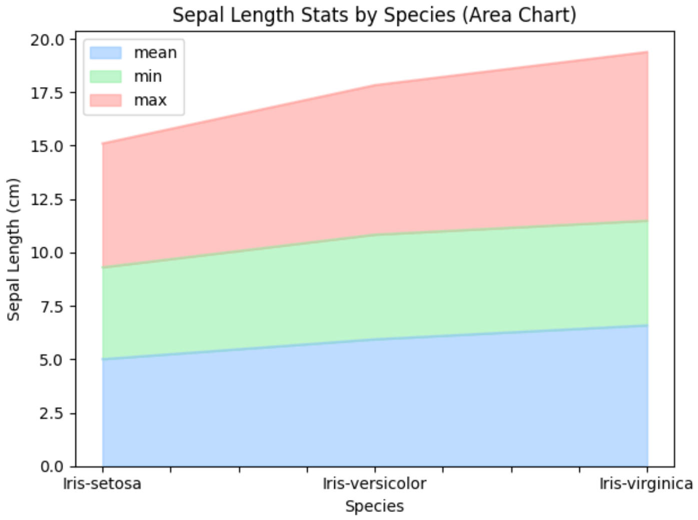
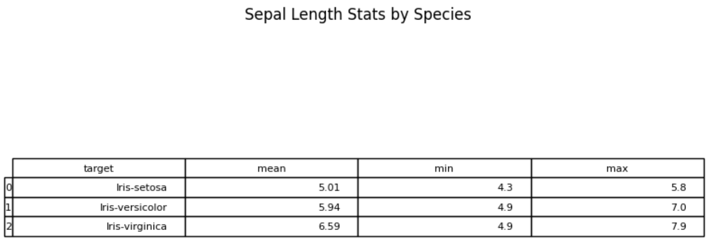


Figure 13: Area Chart: Sepal Length Statistics



Sepal Length Stats by Species

	target	mean	min	max
0	Iris-setosa	5.01	4.3	5.8
1	Iris-versicolor	5.94	4.9	7.0
2	Iris-virginica	6.59	4.9	7.9

Figure 14: Sepal Length Statistics

1.9 Histogram

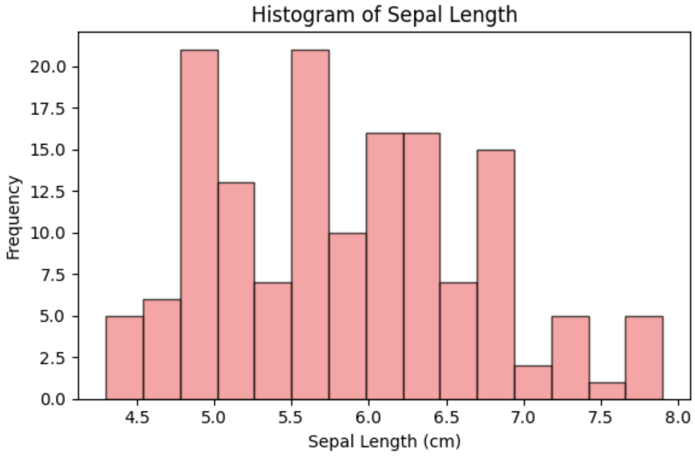
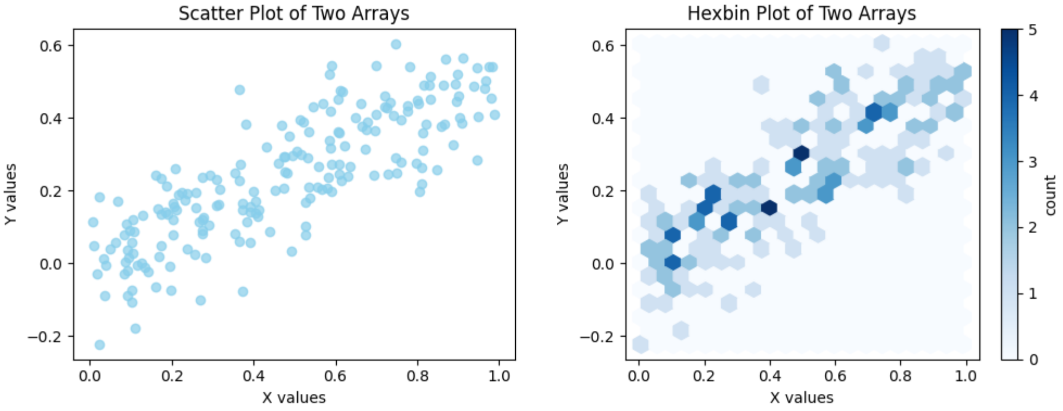


Figure 15: Histogram: Sepal Length Distribution

2 Q2: Visualizing Two Numeric Arrays and IRIS Subsets

2.1 Visualization of Two Random Arrays



(a) Scatter Plot

(b) Hexbin Plot

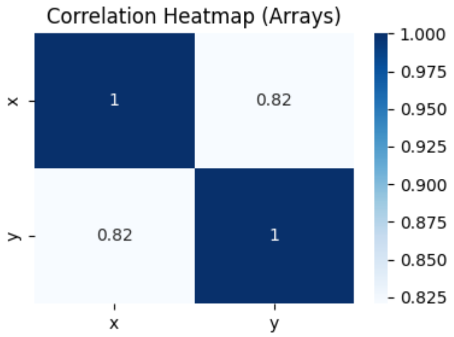
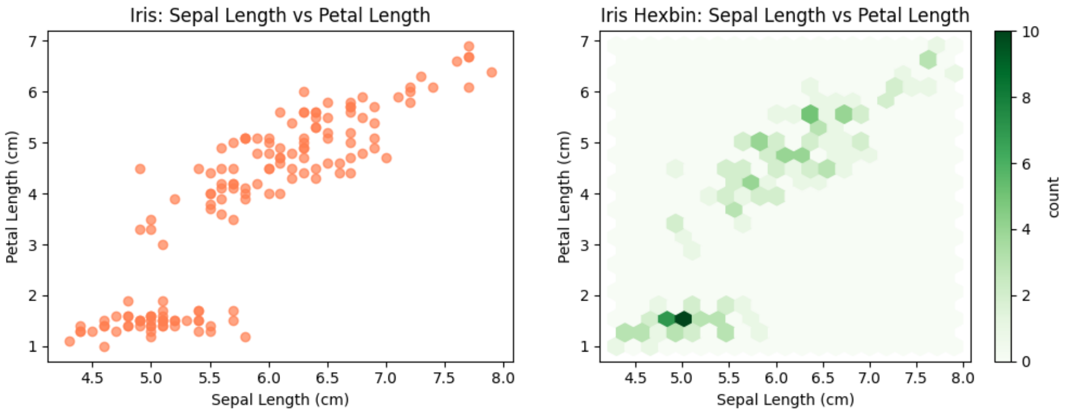


Figure 17: Correlation Heatmap of X and Y

2.2 IRIS Dataset Subset (Sepal Length vs Petal Length)



(a) Scatter Plot

(b) Hexbin Plot

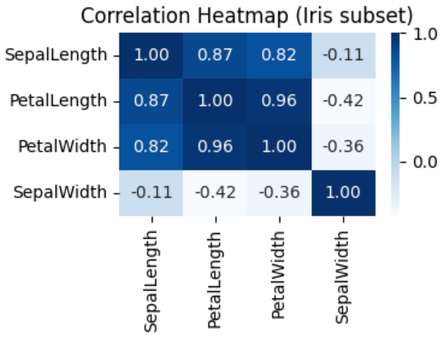


Figure 19: Correlation Heatmap (Iris Subset)

3 Q3: Correlogram on IRIS Dataset

3.1 Heatmap of Feature Correlations

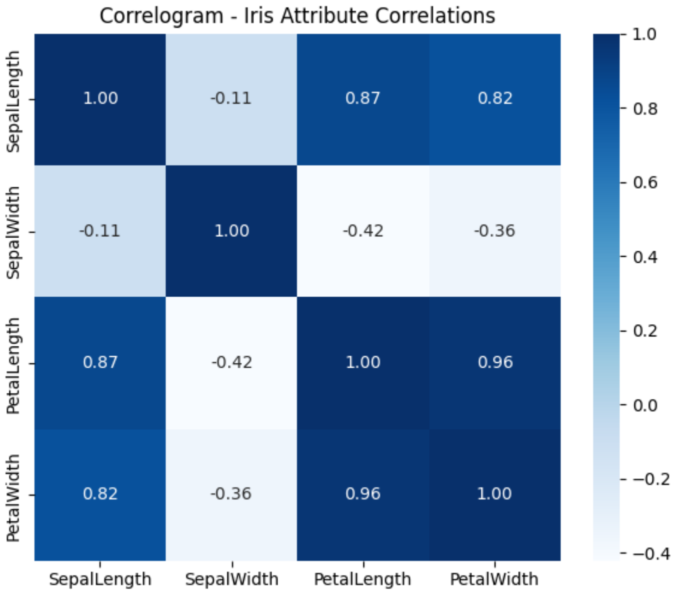


Figure 20: Correlogram: Feature Correlation Heatmap

3.2 Pairplot for Pairwise Relationships

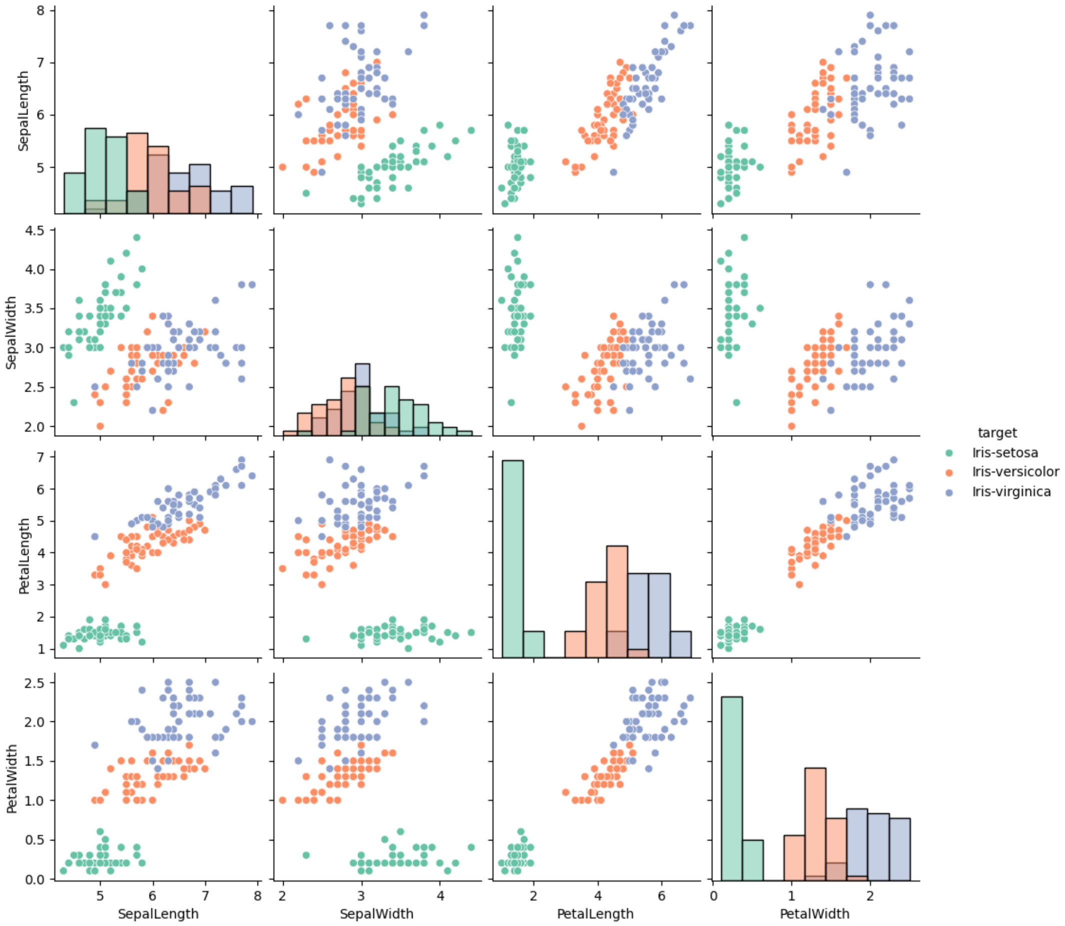


Figure 21: Correlogram: Pairplot of Features by Species

Inferences from Correlogram

- **Petal Length** and **Petal Width** are highly correlated (correlation ≈ 0.96).
- **Sepal Length** moderately correlates with **Petal Length** (≈ 0.87).
- **Sepal Width** has a weak or negative correlation with the other features.
- From the pairplot:
  - *Setosa* is easily separable based on petal features.
  - *Versicolor* and *Virginica* have overlapping clusters but show gradual separation.
- Petal-based features are stronger indicators for classification.

4 Q4: Hierarchical Visualization using TreeMap

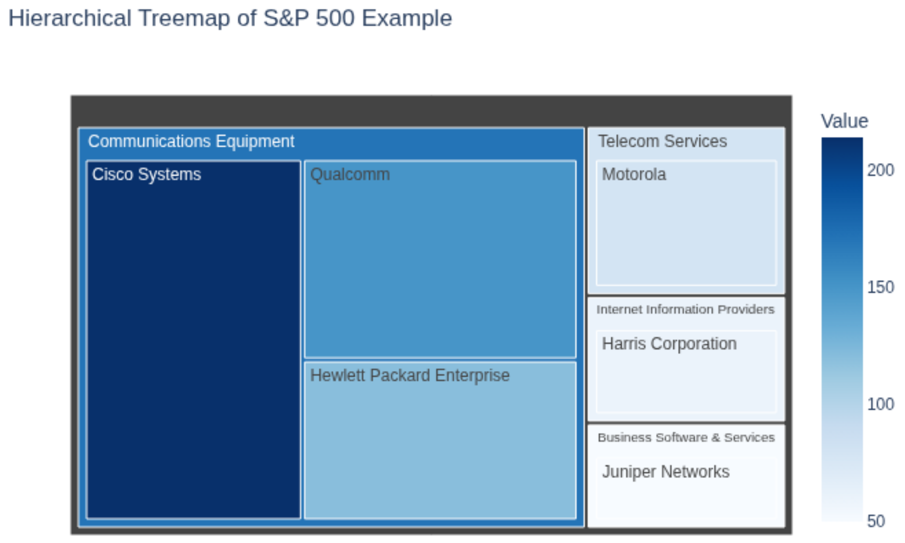


Figure 22: Treemap of S&P 500 Example Hierarchical Data

Treemap Insights

- Displays parent-child relationships in hierarchical data.
- Helpful for identifying the largest subcomponents (e.g., Cisco in Communications Equipment).
- Visualization created using `plotly.express`.

Conclusion

In this report, multiple data visualization techniques were explored using Python. The IRIS dataset provided insights into separability of species and feature importance. Advanced visualizations such as heatmaps, pairplots, and treemaps helped understand correlations and hierarchical structure.

Appendix: Source Code and Data

All source code is provided in the assignment submission along with the dataset.
Note: All outputs are also included in the PDF submission as required.