**DATA ANALYST INTERN PROJECT PHASE – 1**

**Name: Sonam Singh**

**Project: 1**

**Topic: Iris Dataset Basic Analysis**

**Company: Nexus Info**

**Date: 20/08/2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| SR. NO. | INDEX | PG. NO. |
| 1 | **Introduction** | **3-4** |
| 2 | **Data Exploration** | **5-6** |
| 3 | **Exploratory Data Analysis (EDA)** | **7-12** |
| 4 | **Data Visualization** | **12-15** |
| 5 | **Conclusion** | **16** |

**INTRODUCTION**

Iris Dataset is considered as the ‘Hello World’ of data science. It contains five columns namely – Petal Length, Petal Width, Sepal Length, Sepal Width, and Species type. The sample size includes 50 instances for each of the three species of Iris plants – Setosa, Versicolor and Verginica. The dataset records various measurements of these Iris flowers, making it ideal for beginners in data analysis. The researchers have measured various features of the different iris flowers and recorded them digitally.

**OBJECTIVES OF THE PROJECT**

The main purpose of this project is –

1. To perform the Exploratory Data Analysis (EDA) on the Iris dataset using Python, to obtain the basic analysis, visualize key statistics and distribution to gain insights into the dataset.
2. To create data visualizations in Tableau to represent the pattern observed during EDA.
3. And to explore correlations, patterns, and trends within the dataset to understand the relationships between different features.

**DATA METHODOLOGY**

The following steps outline the methodology adopted to analyze the Iris dataset –

1. **Data Loading:** The Iris dataset was loaded in Python only. The dataset was converted into a Data Frame for easier manipulation and analysis.
2. **Data Exploration:** Basic descriptive statistics were calculated to understand the central tendencies, dispersion, and shape of the dataset’s distribution; and the dataset was checked for any missing values or inconsistencies that could affect the analysis.
3. **Exploratory Data Analysis:** Various plots such as histograms, box plots, scatter plots and a correlation matrix were generated to visualize the distributions and identify any relationships between the numerical features in the dataset.
4. **Data Visualization with Tableau:** The processed data from Python was imported into Tableau to create interactive dashboards. Visualizations such as scatter plots, histograms, and box plots were recreated to facilitate a deeper understanding of the data.
5. **Documentation:** Each step of the analysis was documented, including the Python code, Tableau visualizations, and insights derived from the data. This documentation serves as a comprehensive guide to the project’s workflow and findings.

**DATA OVERVIEW**

The Iris dataset consists of 150 observations of Iris flowers, with the following features –

1. Sepal Length (cm): It is one of the primary measurements used to classify the species of the Iris plant.
2. Sepal Width (cm): It complements the sepal length in providing distinguishing characteristics of the different Iris species.
3. Petal Length (cm): Petal measurements are crucial in differentiating between species, as they tend to vary more significantly than sepal measurements.
4. Petal Width (cm): This feature, along with petal length, plays a significant role in identifying the Iris species.
5. Species: The target variable, representing the type of Iris plant. It has three categories Setosa, Versicolor, and Virginica, encoded in 0, 1, and 2 respectively.

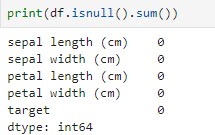
**DATA EXPLORATION**

**LOADING THE DATASET**

The Iris dataset was accessed by using the ‘scikit-learn’ library in Python. This dataset is widely used for pattern recognition and machine learning. The dataset is well-structured and free of missing values, making it ideal for analysis. The codes used for loading and downloading the dataset are given in the Appendix, at the end of this project. The data frame was also created by using ‘DataFrame(,)’ method for easier manipulation.

**CHECKING FOR MISSING VALUES**

To ensure the integrity of the analysis, we first checked for missing values using the ‘isnull().sum()’ method in Python. This method provides a count of missing values in each feature. Fortunately, the iris dataset contained no missing values, allowing us to proceed with the analysis without the need for data imputation. The output for the same obtained from Python is –



Figure

**SUMMARY STATISTICS**

After confirming that the dataset contains no missing values, the next step was to generate summary statistics for each numerical feature in the dataset. This was done using the ‘describe()’ method in Python, which provides key statistics such as the mean, standard deviation, minimum, and maximum values, as well as, the 25th, 50th(median), and 75th percentiles.

The summary statistics provide insights into the overall distribution of the data, helping to identify any potential outliers or skewness in the data. The summary statistics output was then loaded in an excel file which is as below –

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1 - SUMMARY STATISTICS** | | | | | |
|  | **sepal length (cm)** | **sepal width (cm)** | **petal length (cm)** | **petal width (cm)** | **target** |
| **count** | 150 | 150 | 150 | 150 | 150 |
| **mean** | 5.843333333 | 3.057333333 | 3.758 | 1.199333333 | 1 |
| **std** | 0.828066128 | 0.435866285 | 1.765298233 | 0.762237669 | 0.819232 |
| **min** | 4.3 | 2 | 1 | 0.1 | 0 |
| **25%** | 5.1 | 2.8 | 1.6 | 0.3 | 0 |
| **50%** | 5.8 | 3 | 4.35 | 1.3 | 1 |
| **75%** | 6.4 | 3.3 | 5.1 | 1.8 | 2 |
| **max** | 7.9 | 4.4 | 6.9 | 2.5 | 2 |

From the above table, we can interpret that –

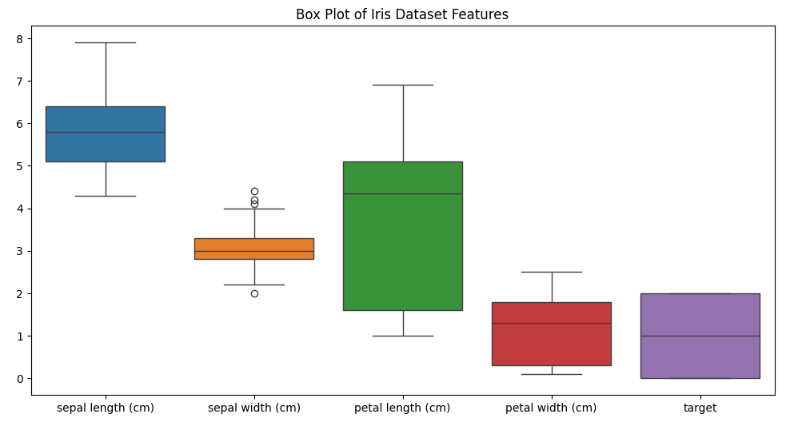
* **Sepal Length (cm):** The mean sepal length is approximately 5.84 cm, with a standard deviation of 0.83 cm, indicating a moderate spread around the mean. The data appears to be fairly symmetric, as the median (50th percentile) of 5.8 cm is close to the mean.
* **Sepal Width (cm):** The mean sepal width is 3.05 cm with a standard deviation of 0.43 cm. The distribution is narrower compared to sepal length, with values ranging from 2 cm to 4.4 cm. The lower 25th percentile of 2.8 cm suggests that a significant number of observations have smaller sepal widths.
* **Petal Length (cm):** It has a higher variability, with a mean of 3.76 cm and a standard deviation of 1.77 cm. The wide range from 1 cm to 6.9 cm indicates distinct differences among the species, which may be a strong predictor of species classification.
* **Petal Width (cm):** The petal width has the lowest mean (1.2 cm) but a significant spread, with values ranging from 0.1 cm to 2.5 cm. The 50th percentile of 1.3 cm shows that the distribution is skewed towards smaller values.
* **Target:** The target variable indicates the species, with uniform distribution.

**EXPLORATORY DATA ANALYSIS (EDA)**

Exploratory Data Analysis (EDA) is a crucial step in any data analysis project. It involves summarizing the main characteristics of the dataset, often with visual methods, to understand the data’s underlying structure, detect anomalies and discover patterns. For this project, EDA was performed on the Iris dataset to gain insights into the distribution and relationships of the various features.

**BOX PLOT**

The box plot provides a summary of the distribution of each feature in the Iris dataset. It shows the medium, quartiles, and potential outliers for each feature.



Figure

From the above figure, we interpret that -

* **Sepal Length:** The box plot for sepal length indicates that the central 50% of the data (the IQR) is relatively narrow, suggesting that most species have sepal length clustered around the mean. The sepal length values range from approximately 4.3 to 7.9 cm. The box plot is fairly symmetric, indicating a normal distribution with no outliers.
* **Sepal Width:** The sepal width has a tighter IQR. The presence of outliers below the lower quartile suggests that some species have sepal widths that are unusually narrow, indicating potential anomalies or a species with significantly different morphological characteristics.
* **Petal Length:** Petal length has a much wider IQR, reflecting greater variability across species. The median petal length is approximately 4.35 cm, but the values span a large range from 1 to 6.9 cm, which suggests significant differences between species. The distribution is positively skewed and the data is evenly spread without extreme outliers.
* **Petal Width:** Petal width also shows a wide range of values, with the median 1.3 cm. the distribution is somewhat positively skewed, indicating that more species have narrower petals with fewer having very wide petals.
* **Target:** This plot indicates that the target variable which represents the species is well distributed, with no significant outliers.

**CORRELATION MATRIX**

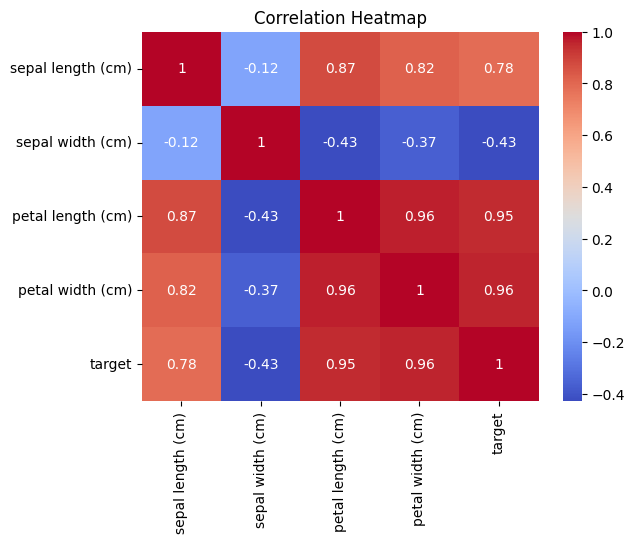


Figure 3

The heatmap was obtained by using ‘.heatmap()’ method in Python. The correlation heatmap shows the correlation between the different features in the Iris dataset. From figure 3, we can interpret that –

* **Positive Correlation:**
* The correlation coefficient of 0.96 indicates a very strong positive relationship between petal length and petal width. This indicates that if the petal length increases, the petal width also increases and vice versa. Their strong relationship with the target variable indicates that both are important features for determining the species.
* The target variable shows a positive correlation of 0.95 with petal length and a highly positive correlation of 0.96 with petal width. These strong correlations show that the petal dimensions are highly indicative of the species.
* **Negative Correlation:** The sepal width shows a negative correlation of -0.43 and -0.37 with petal length and petal width respectively. These negative correlations indicates that species with wider sepals tend to have shorter and narrower petals. Although these correlations are not extremely strong, they suggest an inverse relationship that might help in differentiating certain species when combined with other features.
* **Low/No Correlation:** The Sepal Length shows low to moderate correlations with other features, which suggests that it might not be as crucial in distinguishing between species compared to petal dimensions. However, its moderate correlation with the target variable (0.78), indicates that it still holds some importance, especially when used in conjunction with other features.

**PAIR PLOTS**

The pair plot provides a scatterplot matrix of all feature pairs, colored by the target variable (species type). It also provides the density plots of the Iris features.

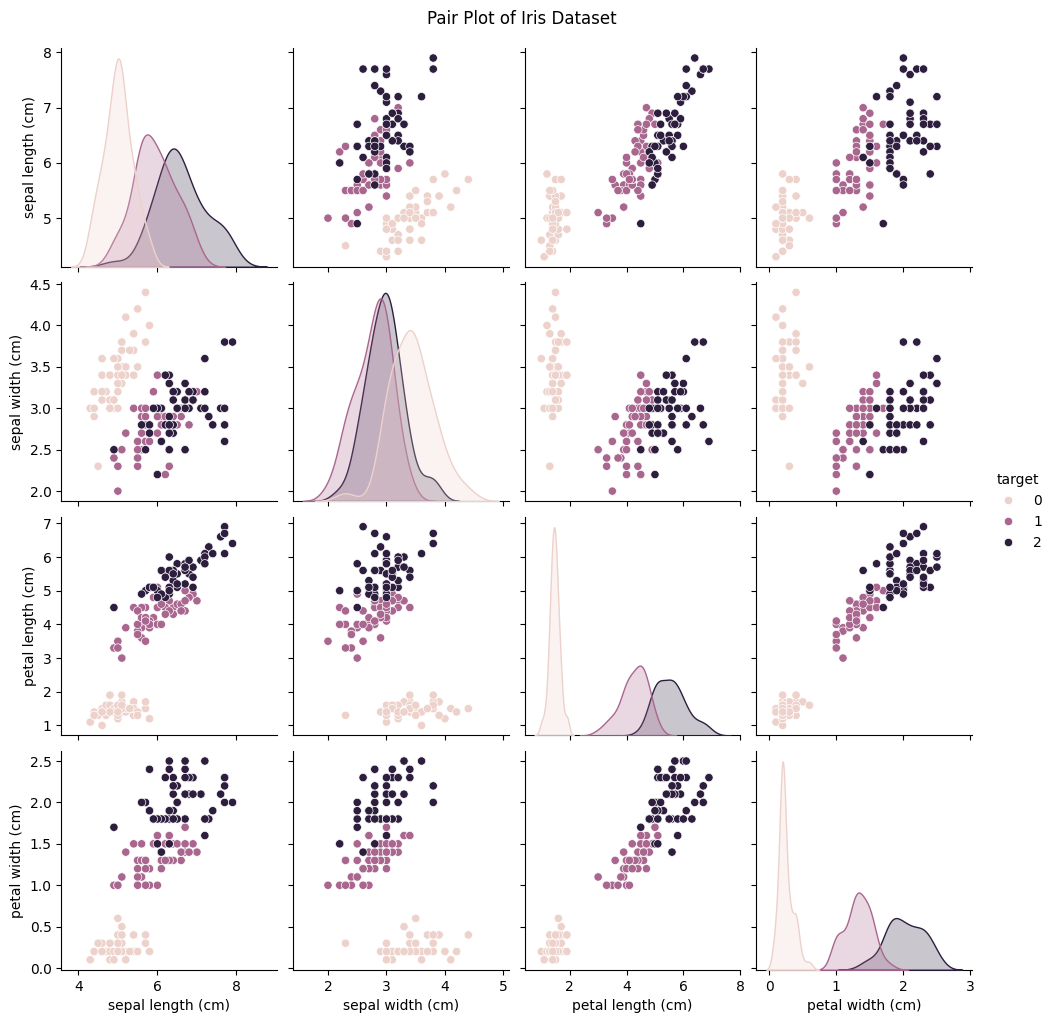


Figure 4

From the above figure, we interpret that-

* **Species Separation:** The pair plot provides a detailed view of how different features interact with each other and how well they separate the species. For instance, in the petal length vs. petal width scatterplots, distinct clusters are observed for each species, indicating that these two features together are excellent at distinguishing between species.
* **Overlapping Species:** Some overlap is observed in the plots involving sepal width, especially when compared with sepal length. This suggests that sepal width alone may not be a strong differentiator of species, and its predictive power may be limited unless used in combination with other features.
* **Diagonal Plots (KDE):** The diagonal plots show the kernel density estimation (KDE) for each feature, providing a smoothed version of the histogram. The KDE for petal length shows clear peaks for each species, reinforcing the observation that petal length is a strong feature for classification. The KDE paths for sepal width, however, show more overlap, indicating that this may not be as effective in separating the species.

**HISTOGRAMS**

The histograms provide a view of the frequency distribution of each feature in the Iris dataset.

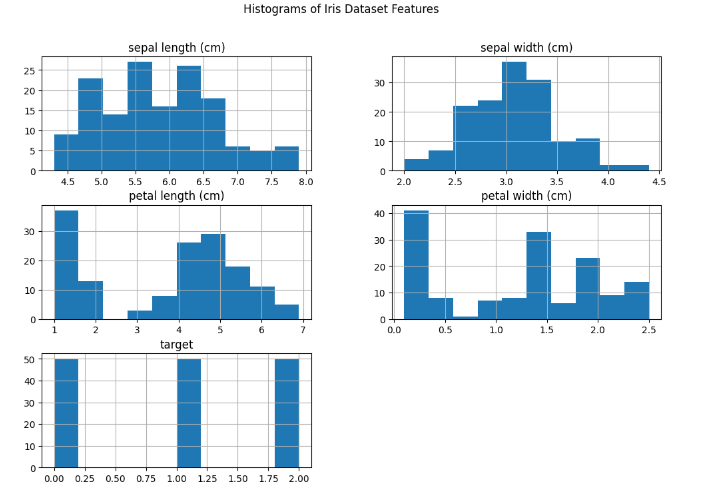


Figure 5

From the above figure, we can interpret that-

* **Sepal Length & Sepal Width:** The histogram of both the features exhibit a fairly normal distribution with multiple peaks, indicating more variability among the species. The presence of several peaks suggests that certain species may have distinct sepal length and sepal width that set them apart from others.
* **Petal Length & Petal Width:** Both petal length and petal width histograms show more distinct bimodal and multimodal distributions. These features are highly variable between species, which is evident from the clear separation in the histograms. The concentration of lower values suggests that one or more species have significantly shorter and narrower petals, while other species have much larger petals.
* **Target:** The histogram for the target variable shows that the dataset is well-balanced with an equal number of samples for each species. This balance is important for building a robust classification model, as it ensures that the model will not be biased towards any particular species due to an uneven distribution of data.

**DATA VISUALIZATION**

In this section, we explore the Iris dataset through a series of visualizations created using Tableau. Data visualization plays a pivotal role in revealing patterns, trends, and insights that may not be immediately apparent from the raw data alone. These visualizations not only aid in exploratory data analysis (EDA) but also serve as a foundation for more advanced data modeling and decision-making processes.

**PAIRWISE SCATTER PLOT**

* The scatter plot of Sepal features in fig. 6, shows some separation between species, but the correlation is weaker compared to petal features and the different color shades demonstrate different species (0, 1, 2) of the target variable, going from light shade to dark.

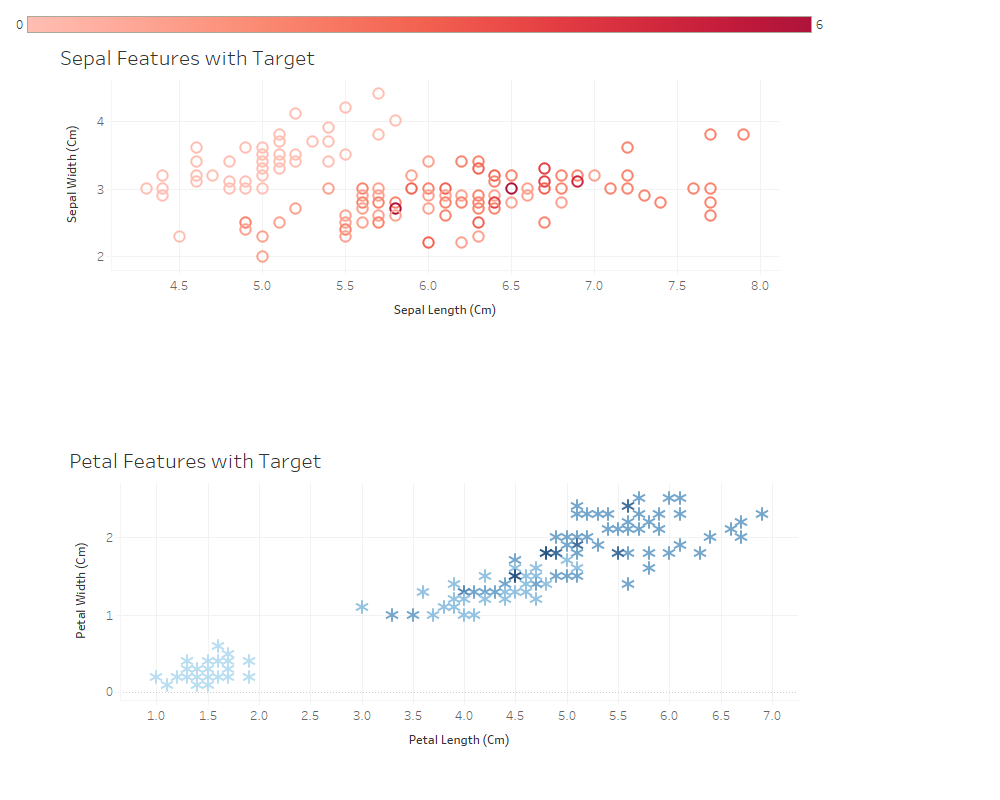


Figure 6

* The scatter plot for petal features exhibits a strong linear correlation, indicating that these features are more effective in distinguishing between species.

**BAR CHART OF AGGREGATED VALUES & HISTOGRAMS OF IRIS FEATURES**

* The bar chart of aggregated values of Iris features in fig.7, summarizes the total values of each feature, showing that Sepal length and Petal length have the highest sums, indicating larger average measurements.

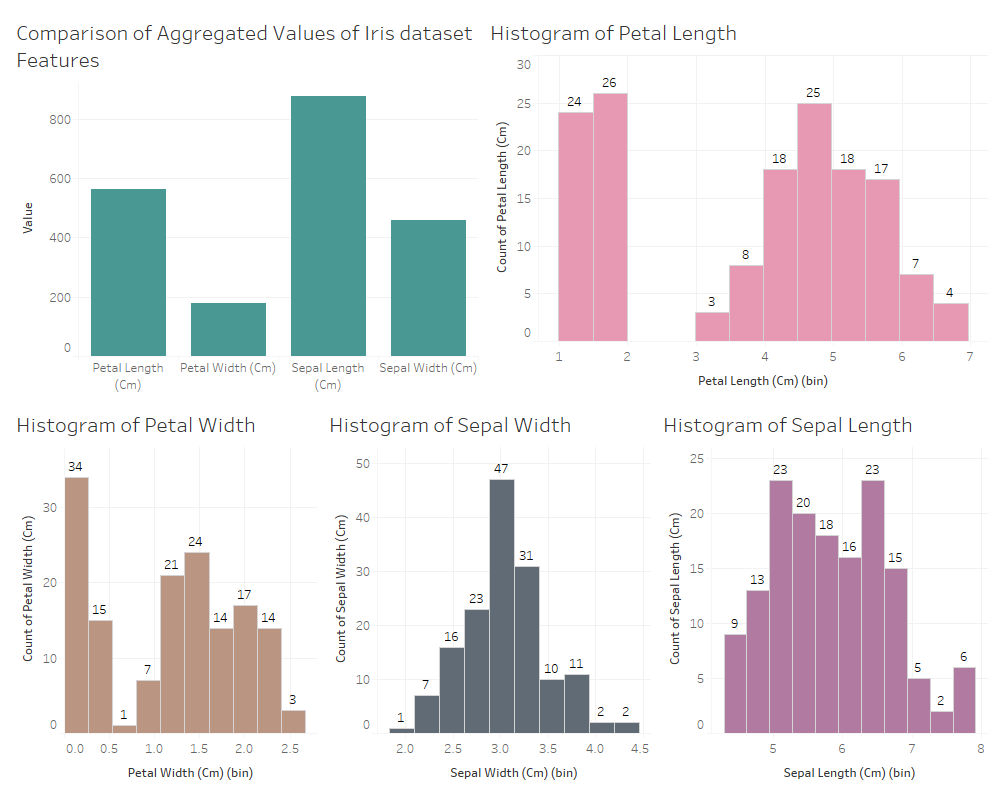


Figure 7

* Histograms reveal the distribution of measurements. Sepal features show a more uniform distribution, while Petal features have more distinct peaks, indicating variations in species characteristics.

**SUMMARY OF MORE ANALYSIS**

* The box plots in fig.8, highlight the spread and central tendency of each Iris feature, with potential outliers identified, especially in Petal Width.
* The scatter plot matrix or the heatmap matrix visually demonstrates the correlation between features. Petal Length & Petal Width show a strong positive correlation, whereas Sepal features have weaker relationships.

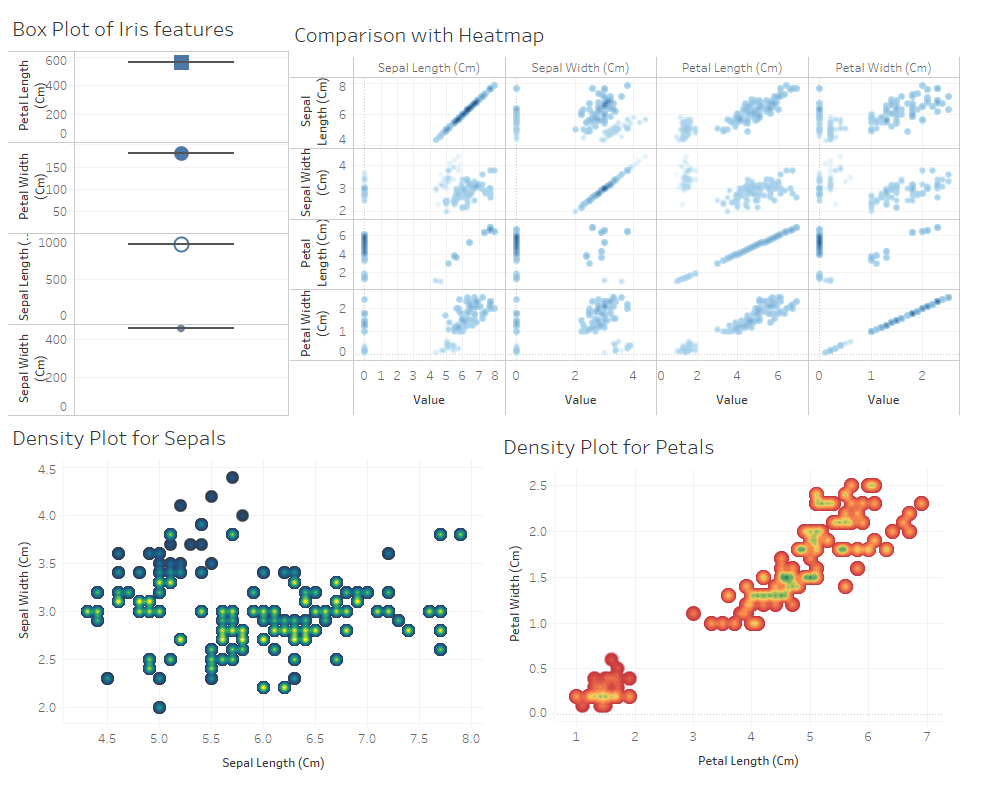


Figure 8

* The density plots for Sepals and Petals show the concentration of data points in the feature space. We can see that Petal features form tighter clusters, emphasizing their importance in species classification compared to Sepal features.

**CONCLUSION**

In this project, we performed a comprehensive Exploratory Data Analysis (EDA) on the Iris dataset using Python and visualized the findings using Tableau. Through various analyses and visualizations, we gained valuable insights into the characteristics within the dataset.

The analysis revealed that the Petal Length and Petal Width features are highly effective in distinguishing between the three species of Iris. The correlation matrix & scatter plots confirmed strong relationship between these features. On the other hand Sepal features were found to be less discriminative for classification.

The Tableau visualizations further emphasized these findings, illustrating clear patterns and distributions within the data. The plots demonstrated that while Sepal dimensions overlap significantly across species, Petal dimensions provide distinct separation.

While the analysis provided valuable insights, it is important to acknowledge the limitation that the analysis focused primarily on feature correlations and distributions, and did not delve into more sophisticated machine learning techniques that could further enhance species classification.

Overall, the Iris dataset proved to be an excellent example of how feature selection and visualization can lead to clear and actionable insights in a multivariate dataset. The combination of Python and Tableau for visualization has allowed for a robust and interactive exploration of the data, ultimately leading to a deeper understanding of the relationships within the Iris dataset. The techniques and methodologies applied here are foundational for any data analysis process and can be extended to more complex datasets and real-world applications.