Data Analysis on Iris Dataset

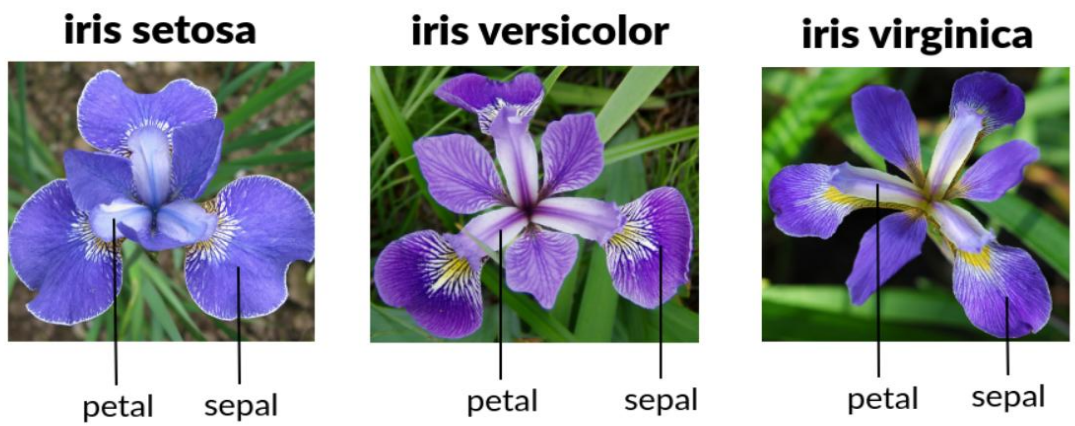
- Vaibhav Saini

- EDA Analysis
- Tableau Analysis

Table of Contents

Introduction

- The Iris dataset was introduced by biologist and statistician Ronald A. Fisher in 1936.
- The dataset includes three species of iris flowers: **setosa, versicolor**, and **virginica**.



- Four features are measured for each flower -
  * **Sepal length**
  * **Sepal width**
  * **Petal length**
  * **Petal width**

- There are 150 observations, with 50 samples from each species.

Methodology

- **Basic Exploratory Data Analysis (EDA) with Python:**

  *Use Python to perform basic EDA on the Iris dataset such as data understanding, cleaning of data and some statistical insight of data.

  *Visualize key statistics and distributions to gain insights into the dataset by using the libraries such as matplot and seaborn to get the visualizations such as histograms,correlation matrix,box plot,etc.

- **Data Visualization with Tableau:**

  *Create visualizations in Tableau to represent the patterns observed during EDA.

  *Explore correlations, patterns,central tendencies and trends within the Iris dataset using plots such as histogram, scatter plots,box plot,bubble plot,etc.

Basic EDA with Python

Step-1: Importing neccesary libraries for analysis such as matplot,seaborn,numpy and pandas.

Step-2: Loading data inside the enviourment through Iris.csv file .

Step-3: Undesrtanding the type of data the file is having.

Step-4: Analyzing the data for it's size , null entries, no. of variable or fields and also the missing values.

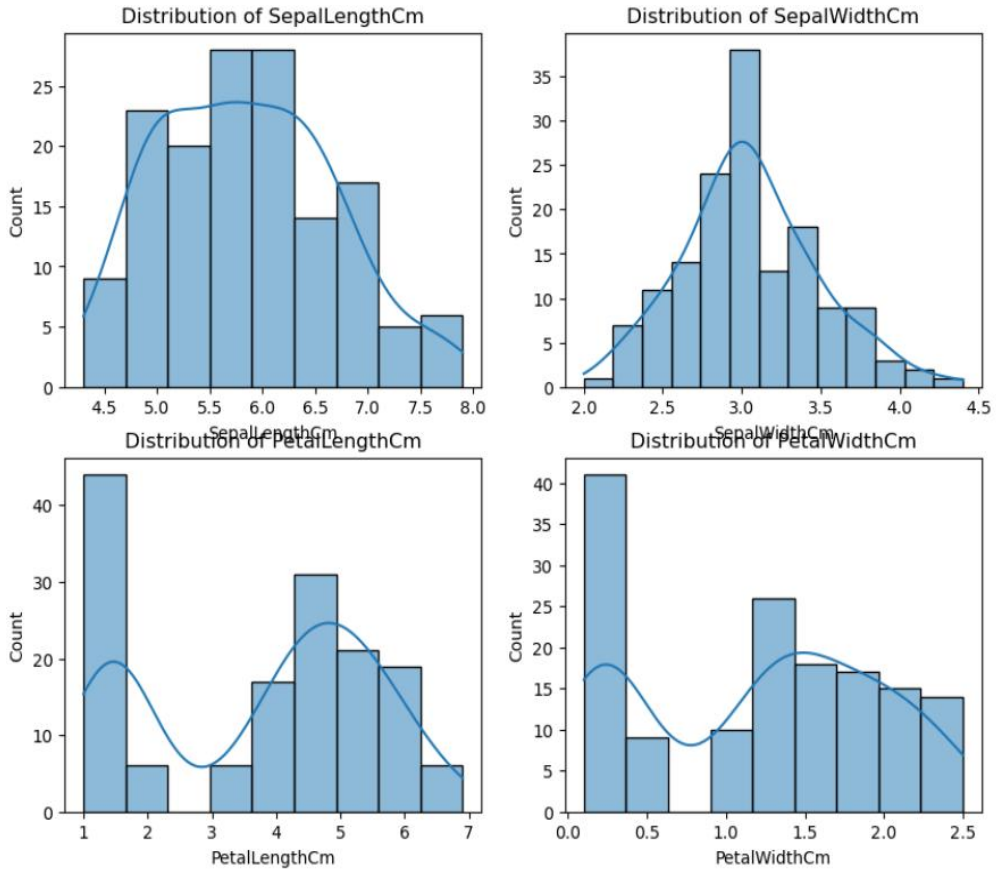
Step-5: Preparing the data by removing or replacing the missing or repititive rows or columns if any but in this case it is zero.

Step-6: Getting insights about the statistical measures if dataset such as mean, standard deviation ,etc.

Step-7: Using the matplot and seaborn library to get the visulization graph for getting the clear insight of the data.

- **Histogram:**

Given plot describe the distribution of SepalLength, SepalWidth,PetalLength and PetalWidth in the given dataset.



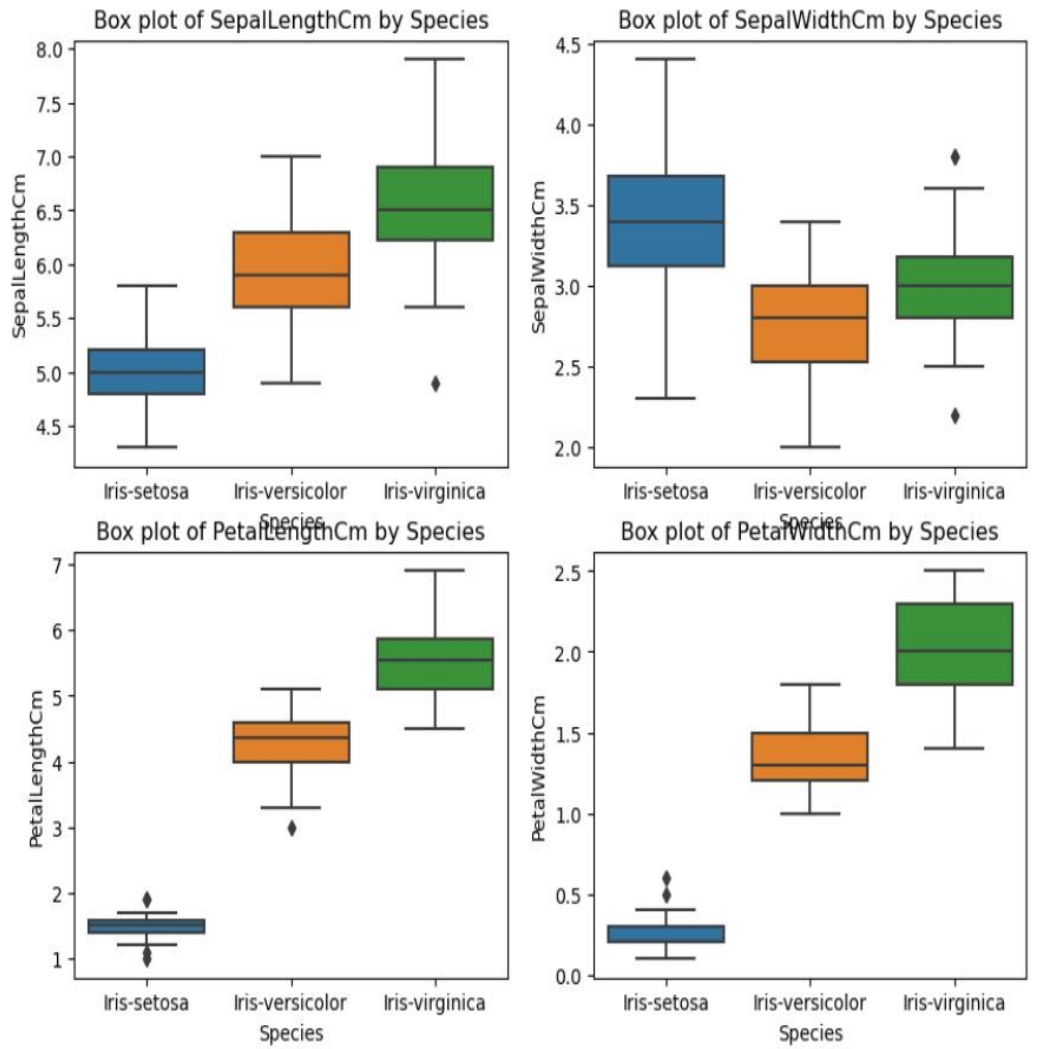
Insights:

❖ Most of the flowers has-

***Petal Length(in range of [1-2cm])**
***Petal Width(in 0-0.5cm)**
***Sepal Length(in 5.5-6.5cm)**
***Sepal Widht(in 2.8-3.2cm)**

- **Box Plot:**

Given plot is showing the distribution of different species across the scale of sepal and petal ,length or width.Also giving the statistical insight of dataset such as ,Median,Interquartile range(IQR) and even some outliners shown by dot .



Insights:

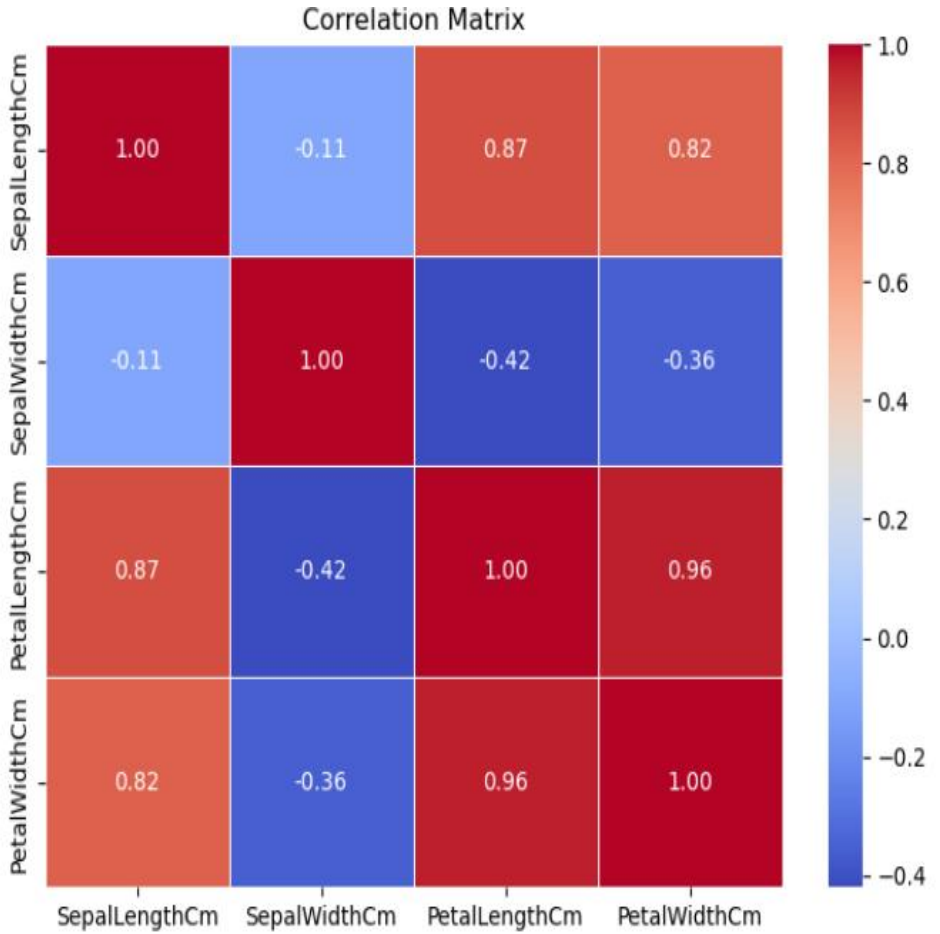
- ❑ Setosa has fewer features and is less dispersed.
- ❑ Versicolor is dispersed evenly and has average characteristics.
- ❑ Virginica has a huge variety of qualities and characteristics and is widely dispersed.
- ❑ Each plot clearly shows the mean/median values for several characteristics (sepal length & width, petal length & width)**[Refer Dashboard 4]**

Correlation Matrix:

The following is used to depict the relation of one characterstic of flower on another characterstic,i.e here it is showing the relation between petal lenght, petal width,sepal length and sepal width. And the number in cells is showing the factor of dependency.

Insight:

- ❑ Sepal Length and Width traits are marginally correlated with each other.
- ❑ There is a strong correlation between column for petal length and width.



Data Visualization with Tableau

Histogram shown in fig1 is showing how the flowers are equally distributed in all 3 species i.e,

- Iris-setosa: 50 flowers
- Iris-versicolor: 50 flowers
- Iris-virginca:50 flowers

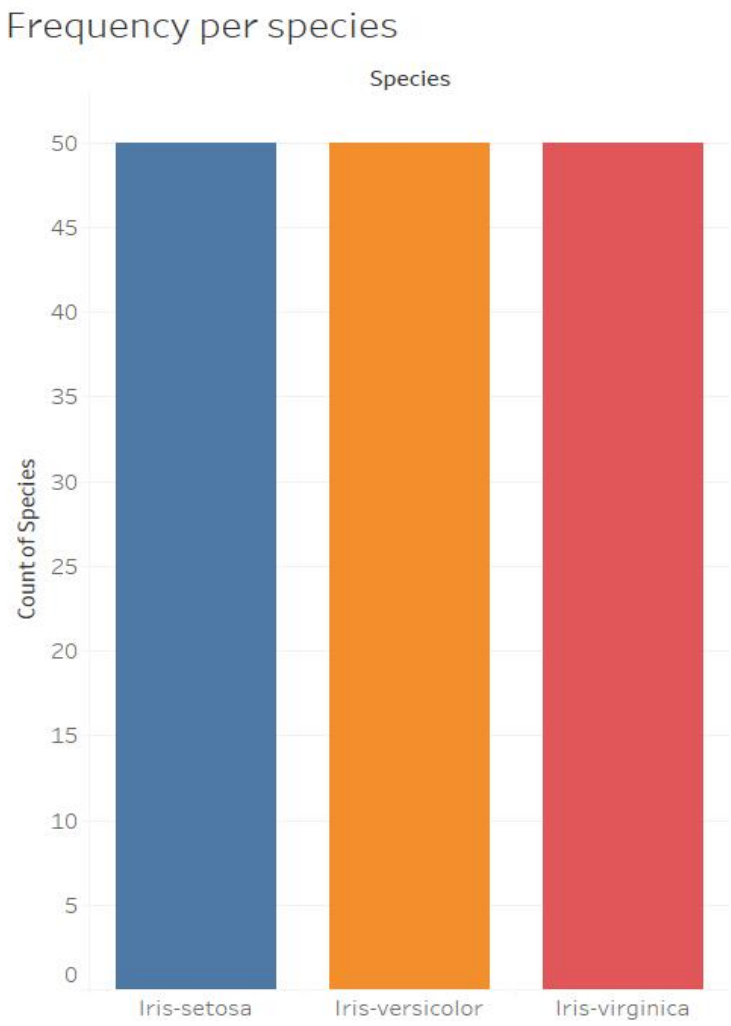
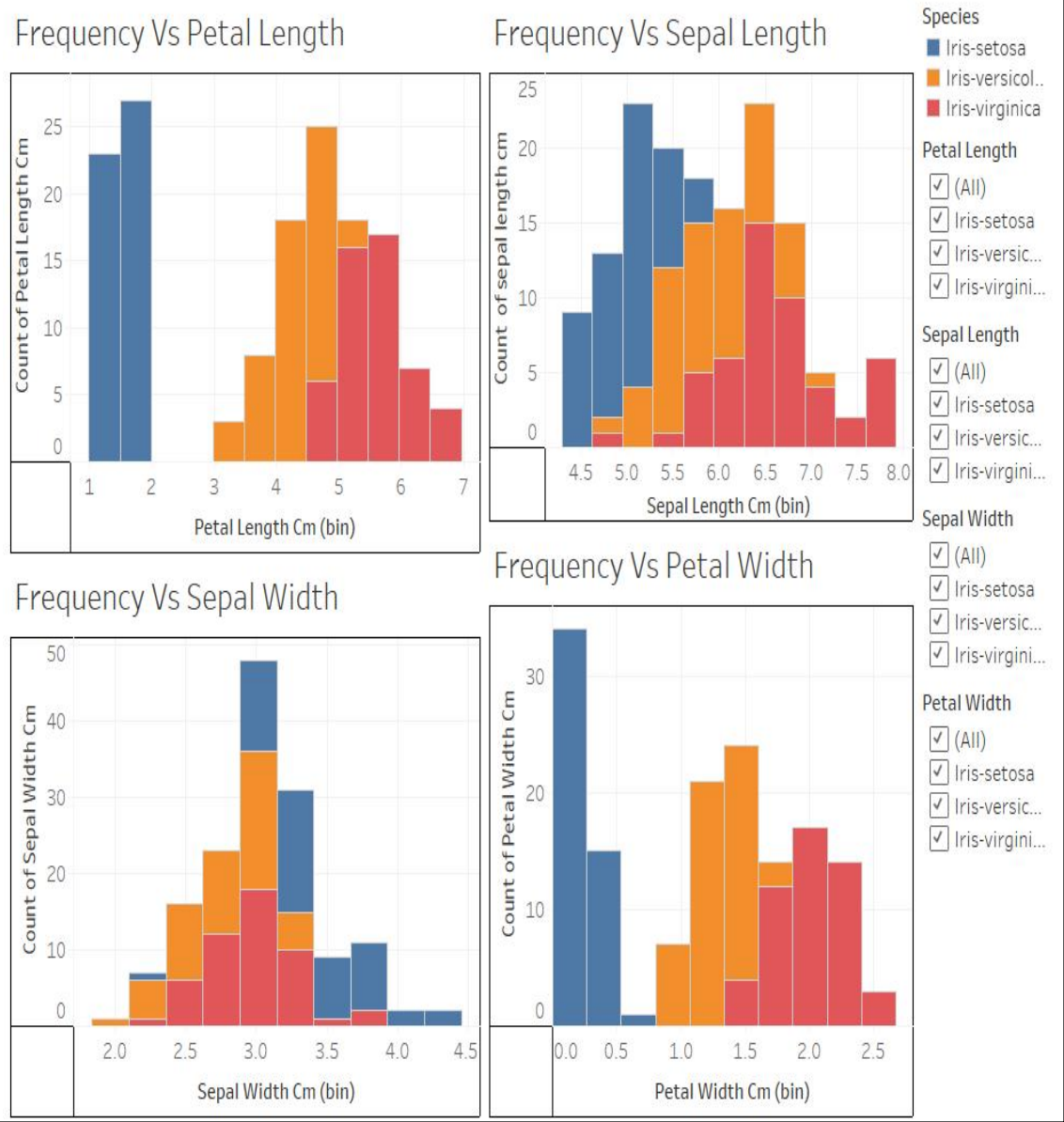


fig1

Given **histogram** are showing frequency vs characterstic lenght and width of sepal and petal seperately:

- It is showing that order of average :
  Petal Lenght: Virginica>Versicolor>Setosa
  Petal Width: Virginica>Versicolor>Setosa
  Sepal Lenght:Virginica>Versicolor>Setosa
  Sepal Width: Setosa>Virginica>Versicolor

- For **Iris-Setosa most common range** for:
  Petal Length: 1-2 cm
  Petal Width: 0-0.8 cm
  Sepal Length: 0-6 cm
  Sepal Width: 3-4.5 cm



- For **Iris-Versicolor most common range** for:
  Petal Length: 3-3.5 cm
  Petal Width: 0.8-1.8 cm
  Sepal Length: 4.5-7 cm
  Sepal Width: 1.8-3.4 cm

- For **Iris-Vriginica most common range** for:
  Petal Length: 4.5-7 cm
  Petal Width: 1.4-2.6 cm
  Sepal Length: 5.5-8 cm
  Sepal Width: 0-4 cm

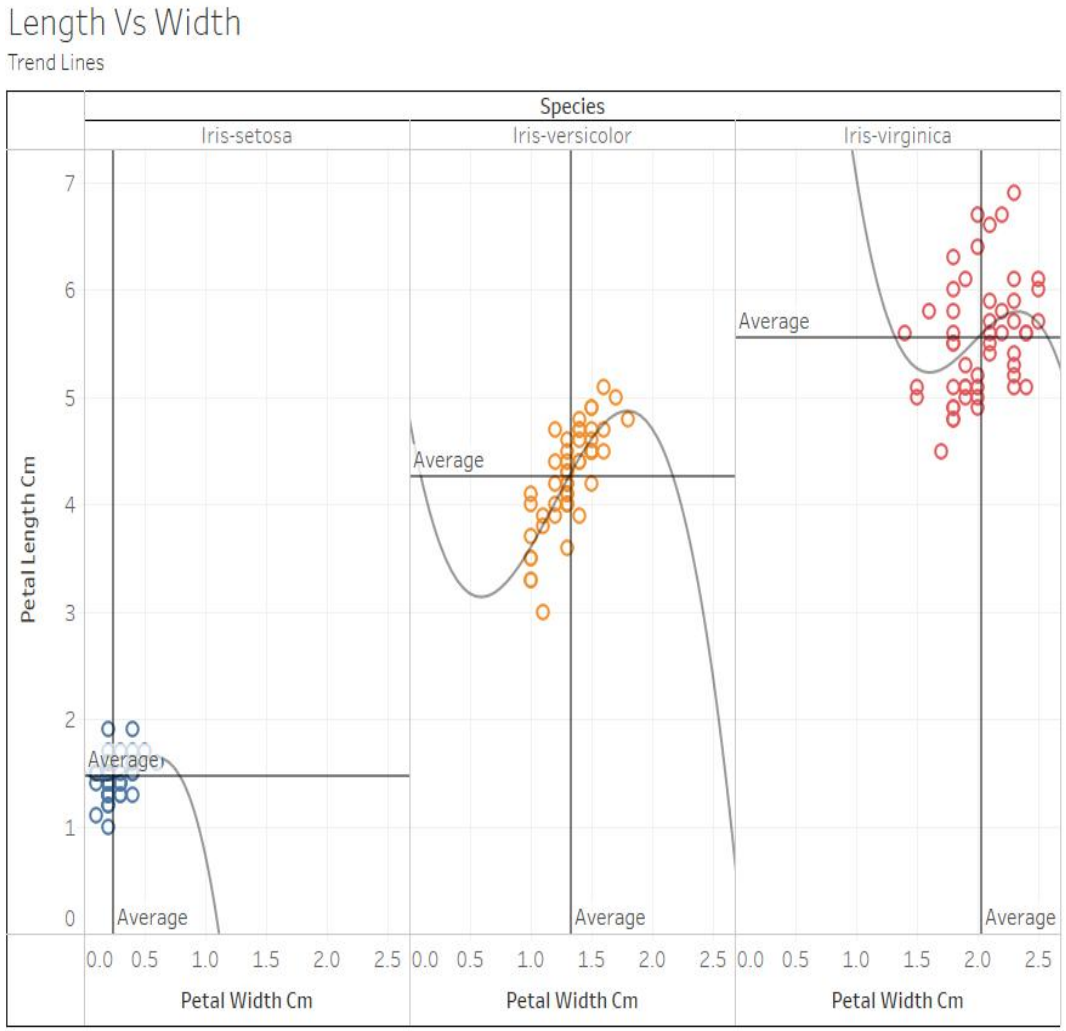
Hence, we can get to know to about the species of flower by considering the range of length and width of flower's petal and sepal.

- Given **scatter plot** is between is between Petal Length and Width of each species of flower seperately showing how the change in width will effect the length of petal and vice versa.

- It is also showing the trend line predicting the data more than just known and how the dataset is behaving based on length and width charecterstics.

- It also determine the average value of length and width also:
  Avg. Petal Length of:
  Iris-setosa: 1.464 cm
  Iris-Versicolor: 4.260 cm
  Iris-virginica:5.552 cm

- **Avg. Petal Width of:**
  Iris-setosa: 0.244 cm
  Iris-Versicolor: 1.326 cm
  Iris-virginica:2.026 cm

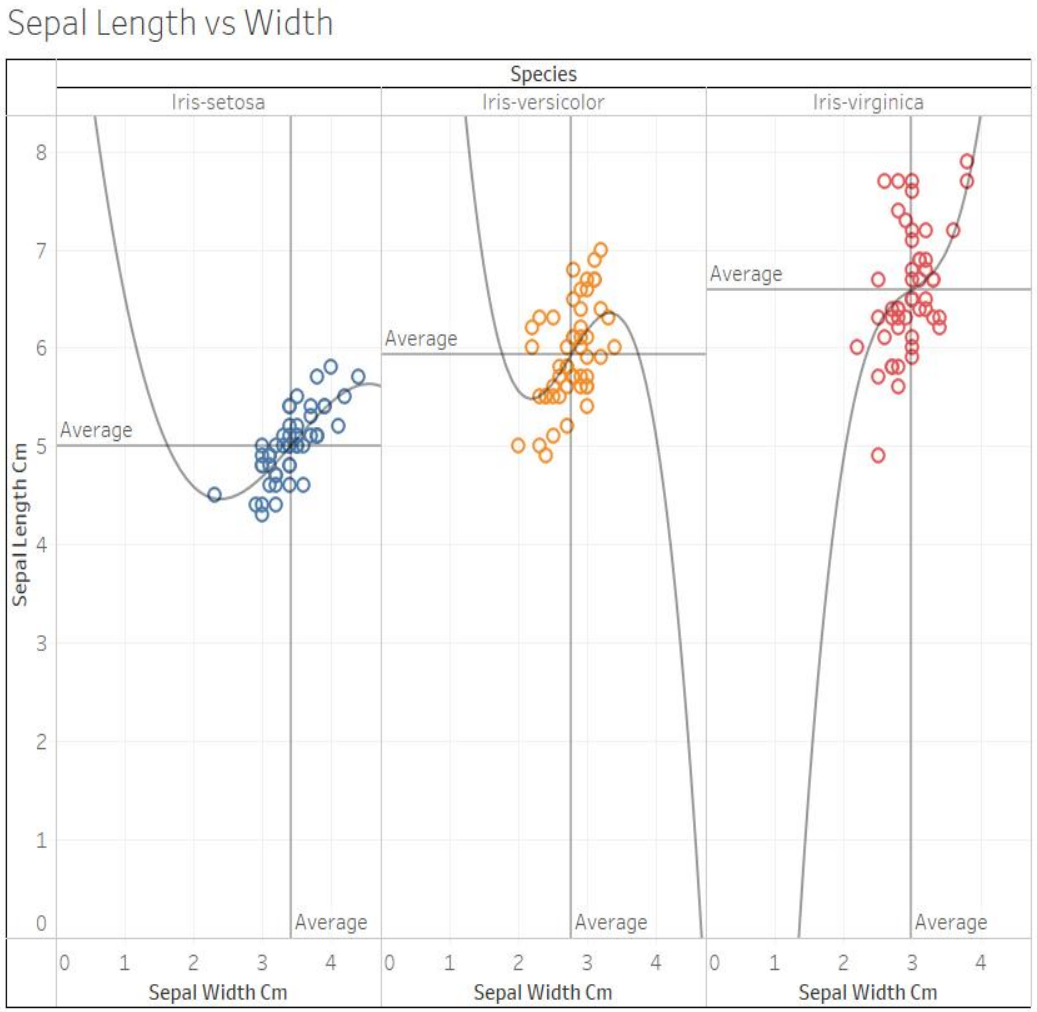


- Given **scatter plot** is between is between Sepal Length and Width of each species of flower seperately showing how the change in width will effect the length of sepal and vice versa.

- Similar to previous plot, it is also showing the trend line predicting the data more than just known and how the dataset is behaving based on length and width charecterstics.

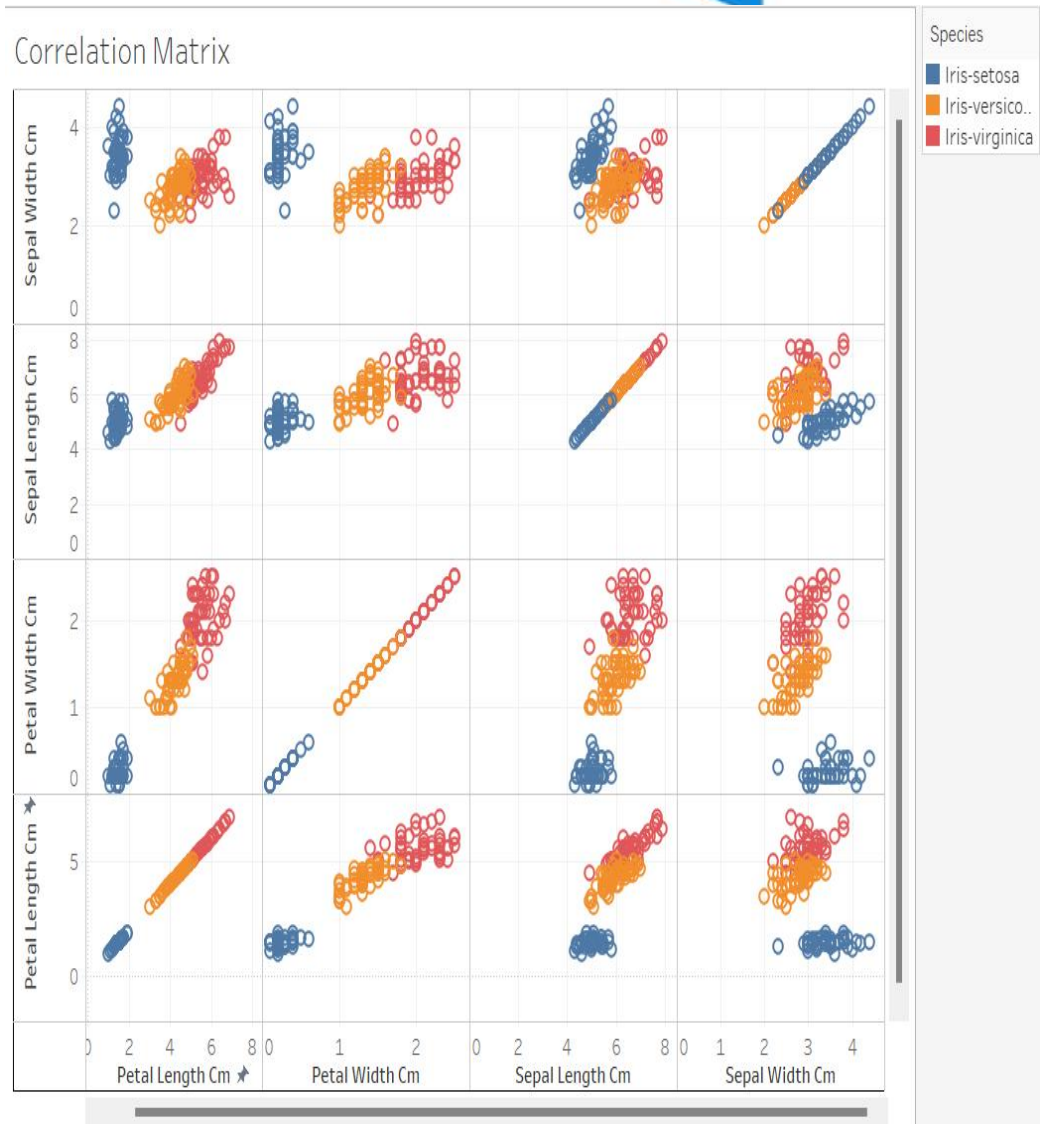
- It also determine the average value of length and width also:
  Avg. Sepal Length of:
  Iris-setosa: 5.006 cm
  Iris-Versicolor: 5.936 cm
  Iris-virginica:6.588 cm

- **Avg. Sepal Width of:**
  Iris-setosa: 3.418 cm
  Iris-Versicolor: 2.770 cm
  Iris-virginica:2.974 cm

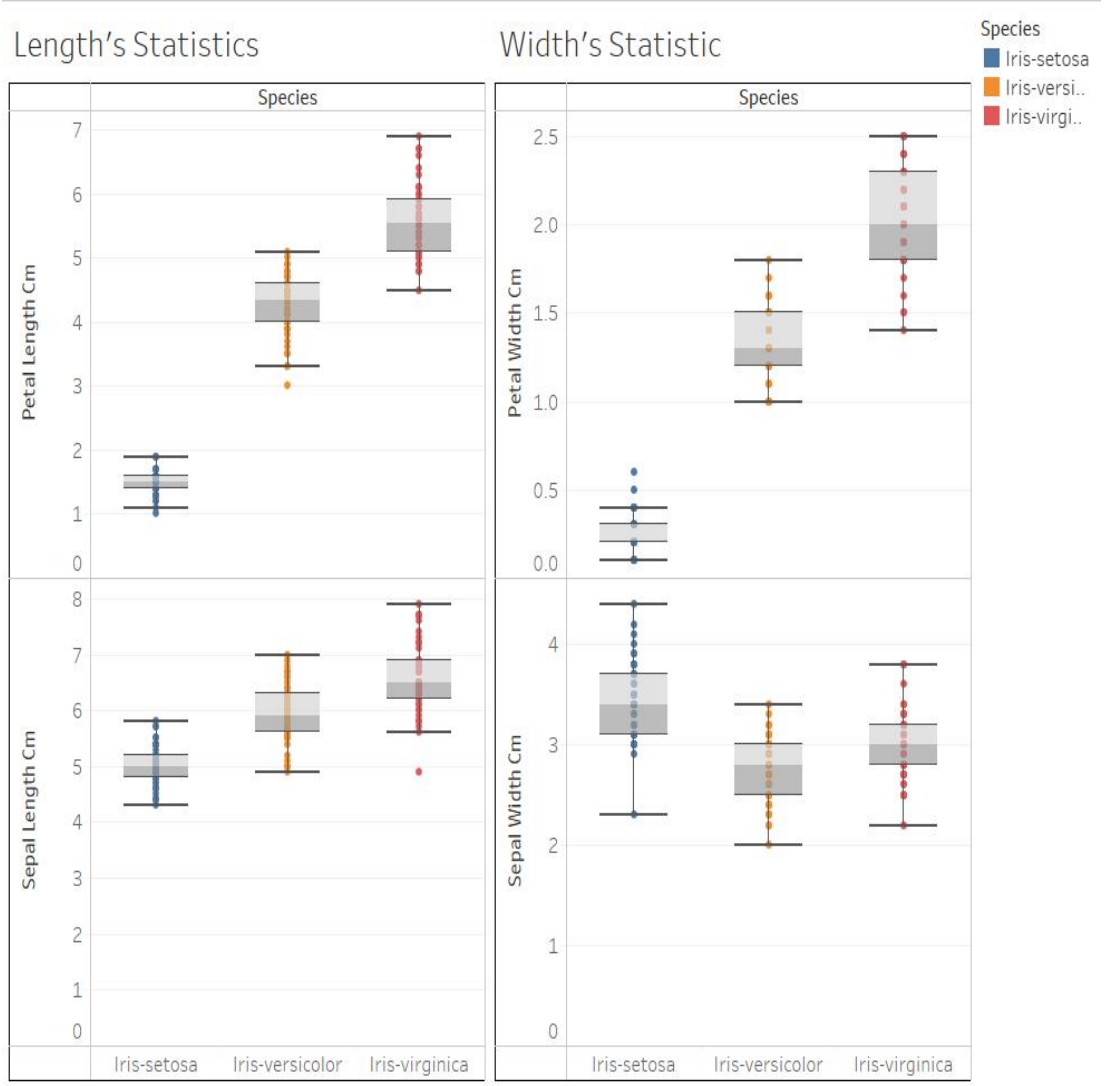


This shows the plot of **correlation matrix** showing the dependency of petal length, petal width , sepal length and sepal width on different all the 3 species.

- Setosa has short petal width and length.
- Setosa has a wide sepal and a short sepal length.
- Versicolor has standard petal width and length.
- Virginica has long and wide petals.
- Versicolor's sepal dimensions have average values.
- Virginica has a narrow breadth but a long sepal length.



- **Boxplots** provide a summary of the distribution of each numeric feature for each species.
- They show the median, quartiles, and potential outliers.
- From the boxplots, we can see that some features (e.g., petal length and petal width) have significant differences in their distributions among the three species, making them valuable for classification.



Statistical data we achieved from box plot are-

- **Iris-setosa:**

- For **petal length-**

  Median:1.5 cm
  IQR:1.4 - 1.6 cm
  Range:1.1 - 1.9 cm
  Outliners:1 cm

- For **petal width-**

  Median:0.2 cm
  IQR:0.2 - 0.3 cm
  Range:0.1 - 0.4 cm
  Outliners:0.6,0.5 cm

- For **sepal length-**

  Median:5 cm
  IQR:4.8 - 5.2 cm
  Range:4.3 - 5.8 cm
  Outliners:none

- For **sepal width-**

  Median:3.4 cm
  IQR:3.1- 3.7 cm
  Range:2.3- 4.4 cm
  Outliners:none

- **Iris-versicolor:**

- For **petal length-**

  Median:4.35 cm
  IQR:4 - 4.6 cm
  Range:3.3 - 5.1 cm
  Outliners:3 cm

- For **petal width-**

  Median:1.3 cm
  IQR:1.2 - 1.5 cm
  Range:1 - 1.8 cm
  Outliners:none

- For **sepal length-**

  Median:5.9 cm
  IQR:5.6 - 6.3 cm
  Range:4.9 - 7 cm
  Outliners:none

- For **sepal width-**

  Median:2.8 cm
  IQR:2.5- 3 cm
  Range:2- 3.4 cm
  Outliners:none

- **Iris- verginica:**

- For **petal length-**

  Median:5.55 cm
  IQR:5.1 - 5.9 cm
  Range:4.5 - 6.9 cm
  Outliners:none

- For **petal width-**

  Median:2 cm
  IQR:1.8 - 2.3 cm
  Range:1.4 - 2.5 cm
  Outliners:none

- For **sepal length-**

  Median:6.5 cm
  IQR:6.2 - 6.9 cm
  Range:5.6 - 7.9 cm
  Outliners:4.9 cm

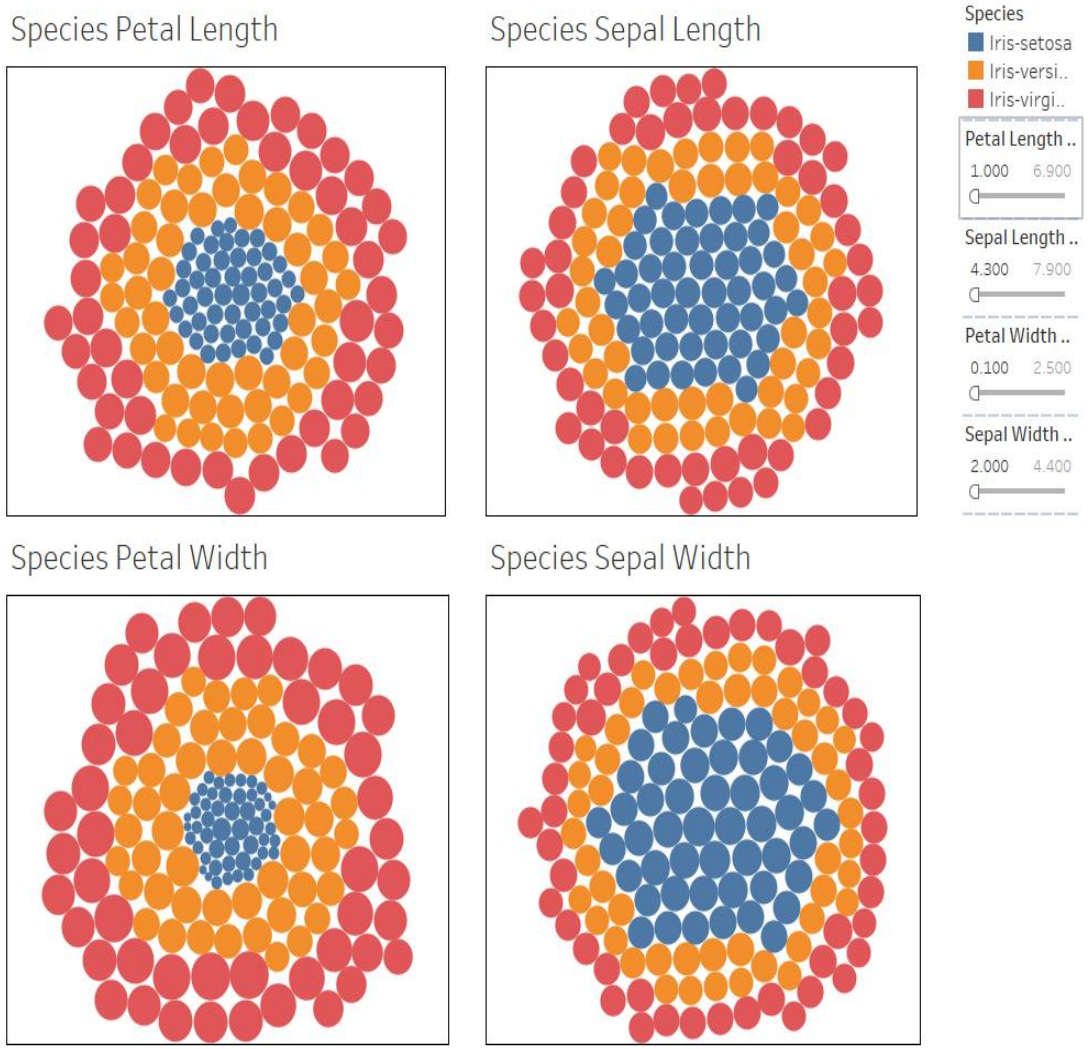
- For **sepal width-**

  Median:3 cm
  IQR:2.8- 3.2 cm
  Range:2.2- 3.8 cm
  Outliners:none

Packed Bubble Plot:

It simply shows the distribution of petal and sepal length and width value for each specie.

Smaller radius circles are showing the less length or width compare to circles with large radius.



Conclusion

Overall, the EDA and visualization of the Iris dataset reveals that features like petal length and petal width are highly discriminative for distinguishing between the three Iris species

THANK YOU