Import modules:

```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot  as plt
```

- ***pandas*** *- used to perform data manipulation and analysis.*
- ***NumPy*** *- used to perform a wide variety of mathematical operations on arrays.*
- ***matplotlib*** *- used for data visualization and graphical plotting.*
- ***seaborn*** *- built on top of matplotlib with similar functionalities.*

Dataset Information

The Iris flower data set or Fisher's Iris data set is one of the most famous multivariate data set used for testing various Machine Learning Algorithms.

IRIS FLOWER

2.characteristics of this dataset.

Loading the dataset:

```
ir = pd.read_csv(r"D:\internship\Iris.csv")
```

```
ir
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

Let us see the statistical information of the attributes.

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667
std	43.445368	0.828066	0.433594	1.764420	0.763161
min	1.000000	4.300000	2.000000	1.000000	0.100000
25%	38.250000	5.100000	2.800000	1.600000	0.300000
50%	75.500000	5.800000	3.000000	4.350000	1.300000
75%	112.750000	6.400000	3.300000	5.100000	1.800000
max	150.000000	7.900000	4.400000	6.900000	2.500000

Let us see the data type information of the attributes

```
ir.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Id              150 non-null    int64  
 1   SepalLengthCm   150 non-null    float64
 2   SepalWidthCm    150 non-null    float64
 3   PetalLengthCm   150 non-null    float64
 4   PetalWidthCm    150 non-null    float64
 5   Species         150 non-null    object 
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

Checking the balance:

```
ir["Species"].value_counts()
```

```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
```

Checking null values:

NO null value inside this dataset

```
features_with_na = [features for features in ir.columns if ir[features].isnull().sum()>1]
for ft in features_with_na:
    print(ft,np.round(ir[ft].isnull().mean()*100,4), '% missing values ')
```

```
ir.isnull().sum()
```

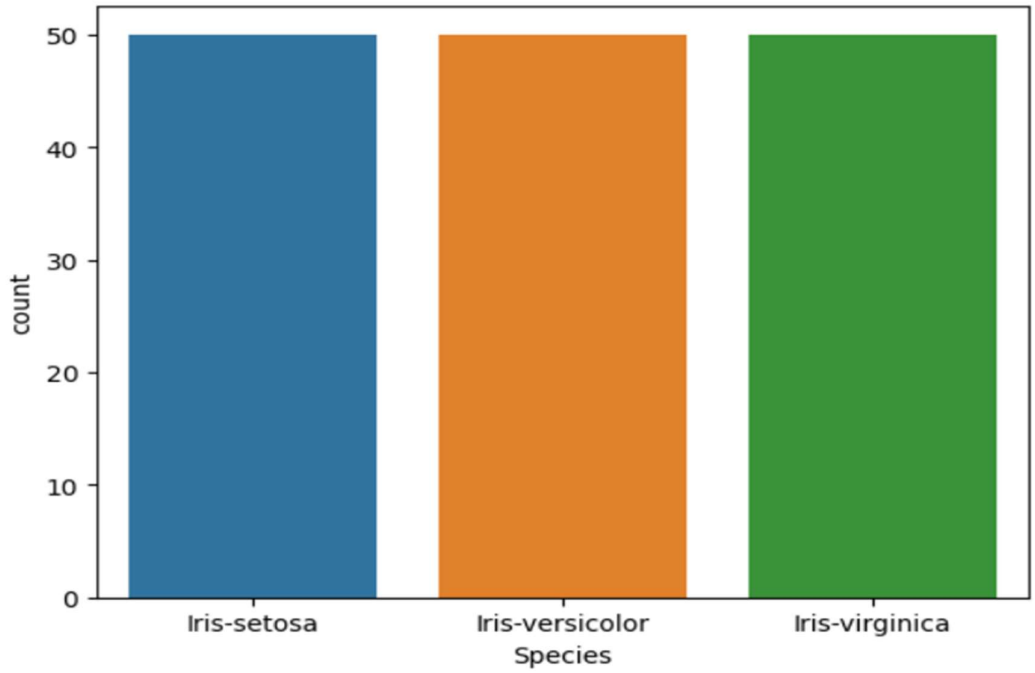
```
Id              0
SepalLengthCm   0
SepalWidthCm    0
PetalLengthCm   0
PetalWidthCm    0
Species         0
dtype: int64
```

EDA

Species count:

```
sns.countplot(x = "Species",data =ir)
```

```
<Axes: xlabel='Species', ylabel='count'>
```



Data Insight:

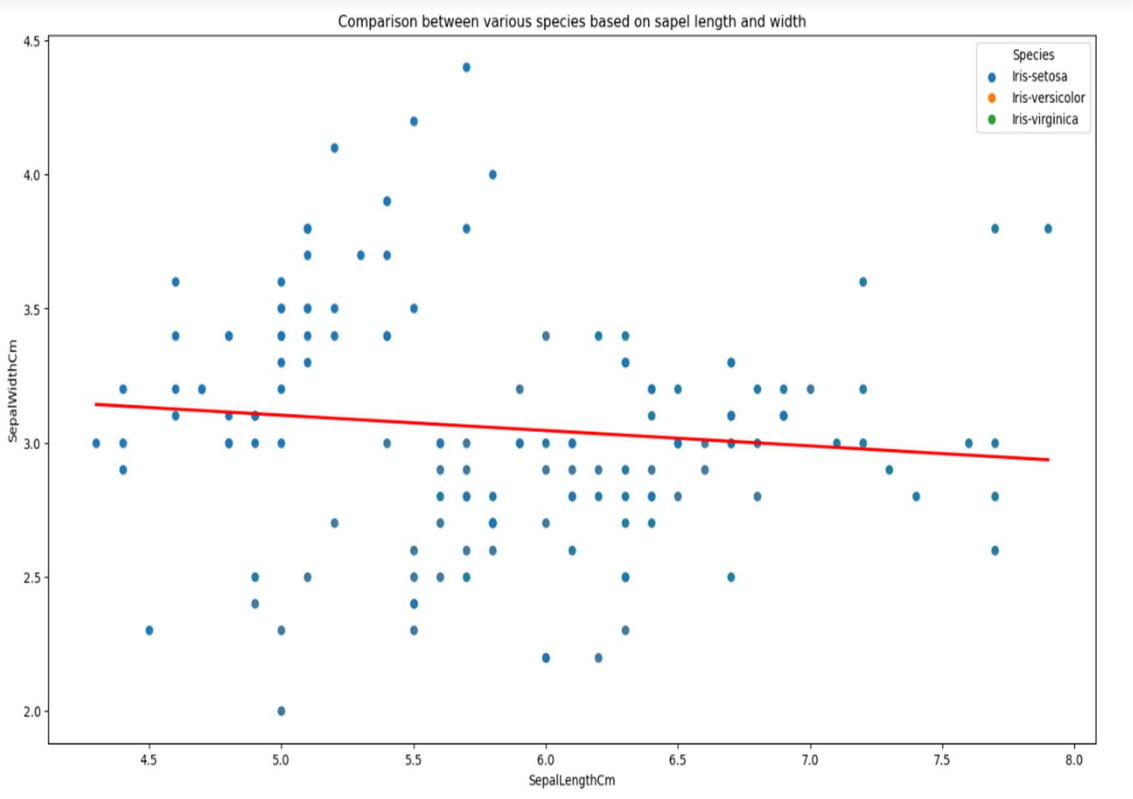
•This further visualizes that species are well balanced.

•Each species (Iris virginica, setosa, versicolor) has 50 as it's count.

Iris Flower Species

Uni-variate Analysis:

Comparison between various species based on sepal length and width:



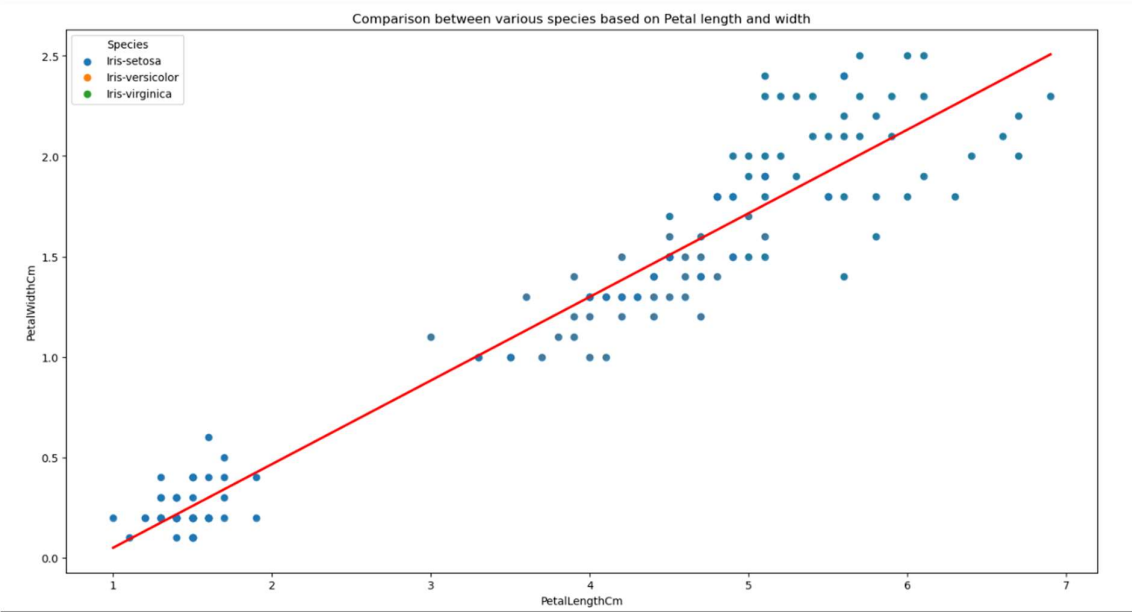
Data Insights:

1)Iris Setosa species has smaller sepal length but higher width.

2)Versicolor lies in almost middle for length as well as width

3)Virginica has larger sepal lengths and smaller sepal widths.

Comparison between various species based on petal length and width



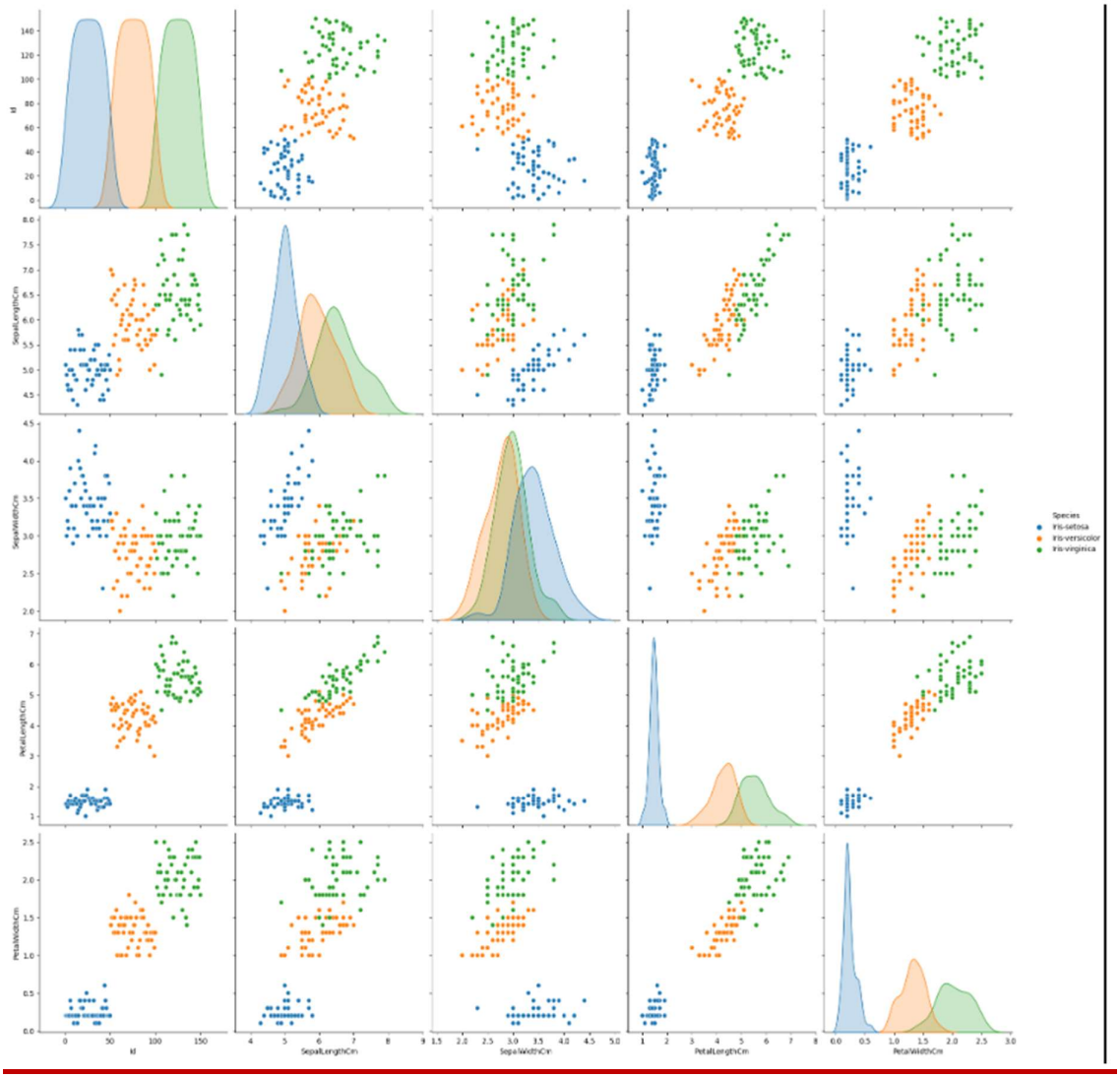
Data Insights:

1)Setosa species have the smallest petal length as well as petal width

2)Versicolor species have average petal length and petal width

3)Virginica species have the highest petal length as well as petal width

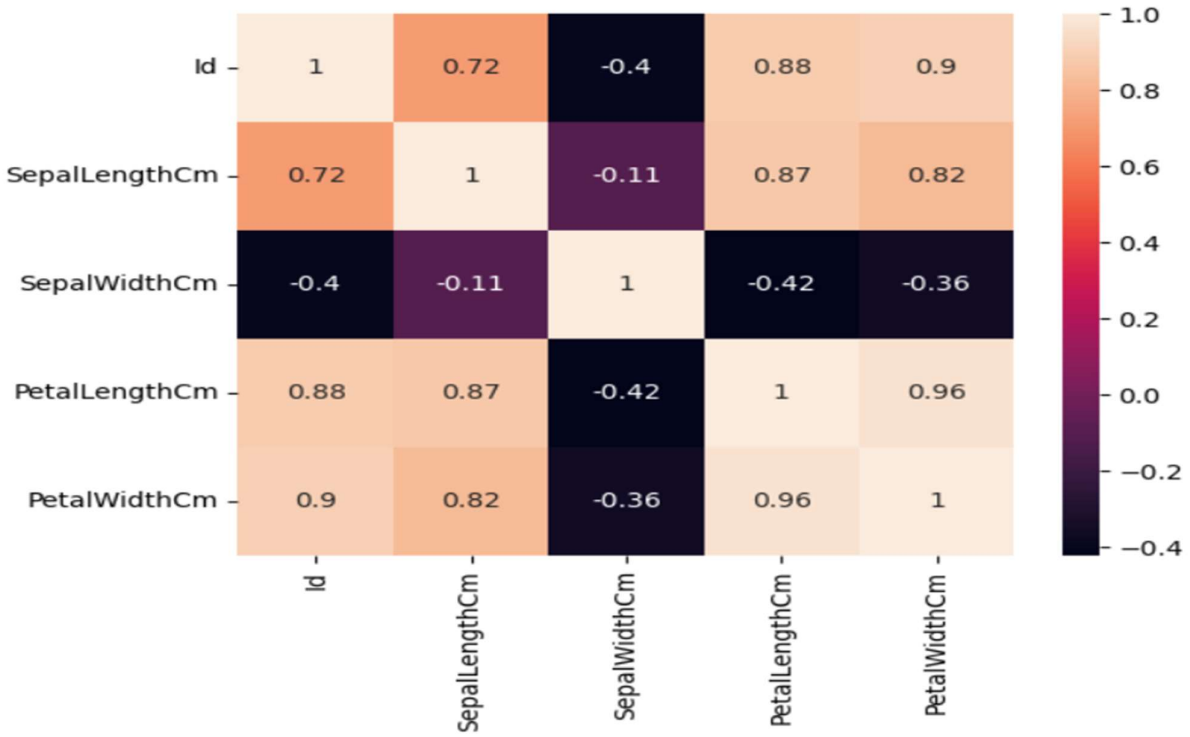
Bi-variate Analysis:



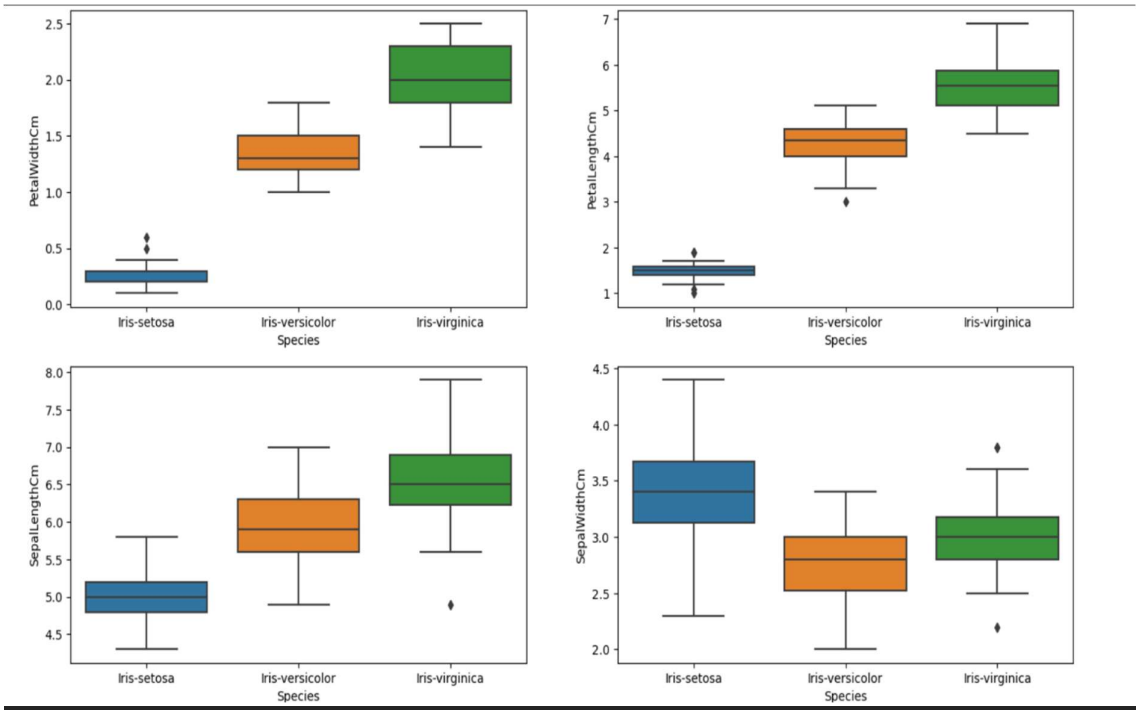
Data Insights:

1. High co relation between petal length and width columns.
2. Setosa has both low petal length and width
3. Versicolor has both average petal length and width
4. Virginica has both high petal length and width.
5. Sepal width for setosa is high and length is low.
6. Versicolor have average values for for sepal dimensions.
7. Virginica has small width but large sepal length

Checking Correlation:



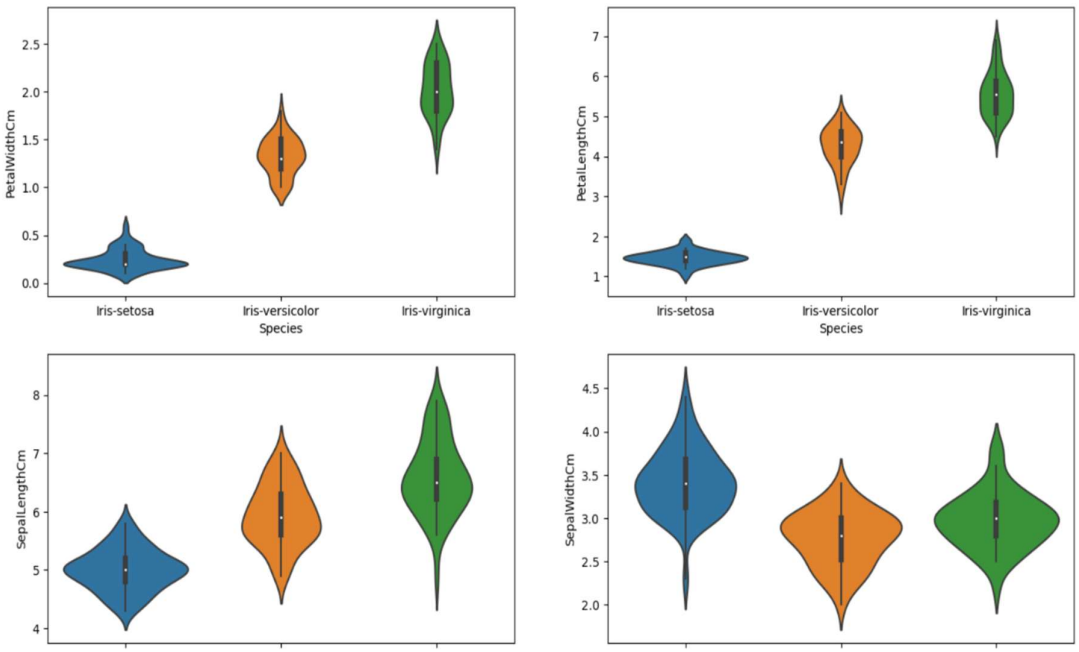
Box plots to know about distribution:



Data Insights:

1. Setosa is having smaller feature and less distributed
2. Versicolor is distributed in a average manner and average features
3. Virginica is highly distributed with large no .of values and features
4. Clearly the mean/ median values are being shown by each plots for various features(sepal length & width, petal length & width)

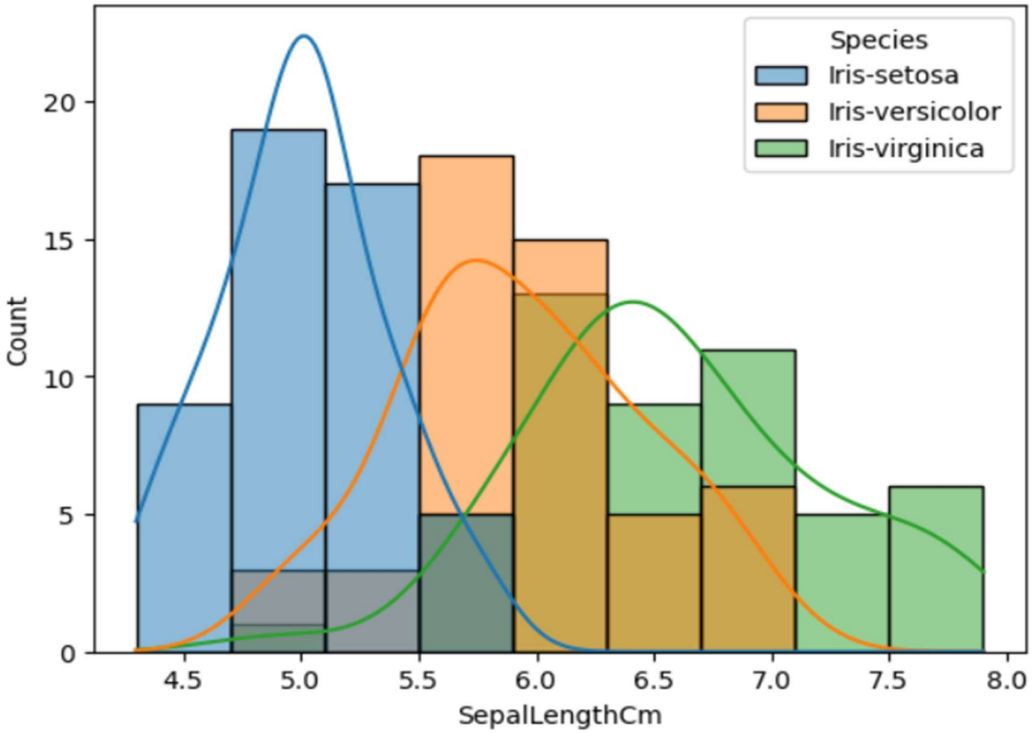
Violin Plot for checking distribution

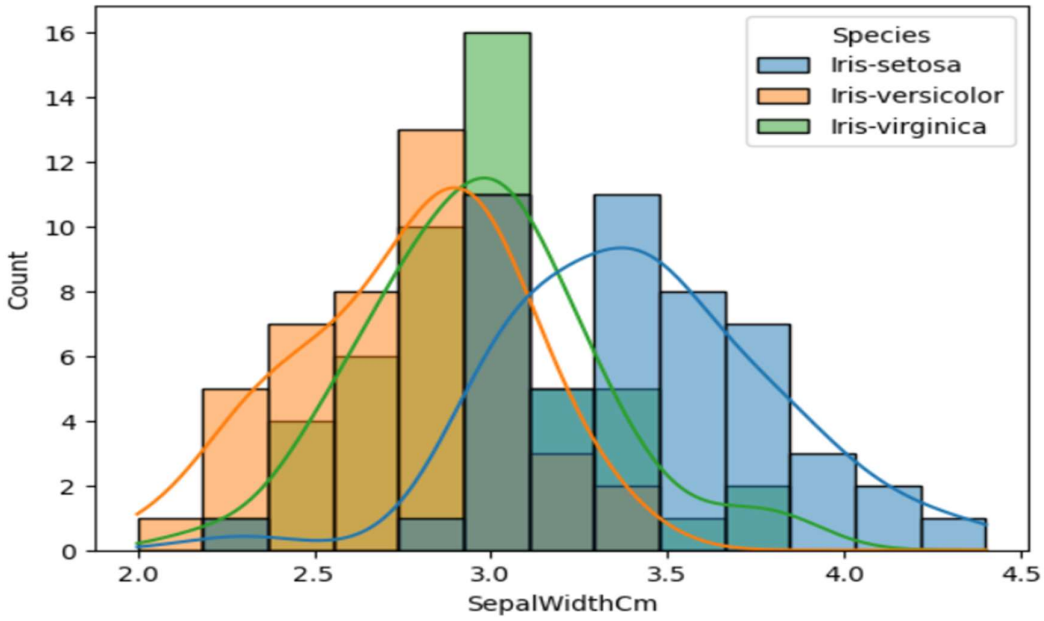


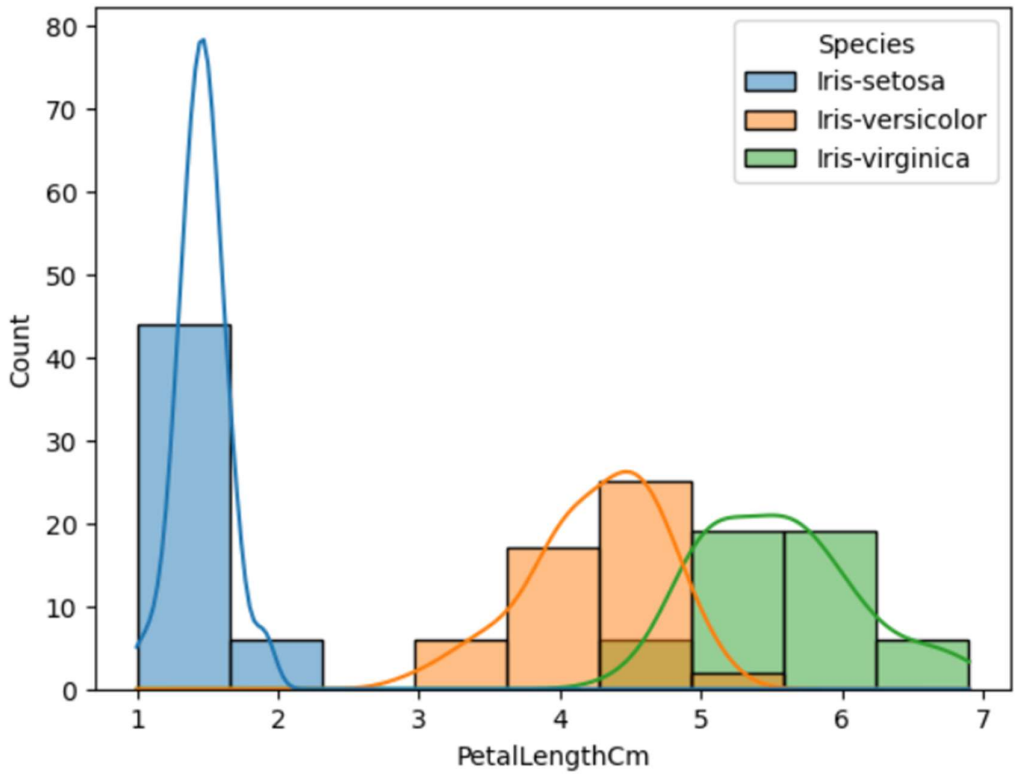
- **Data Insights:**
    1. Setosa is having less distribution and density in case of petal length & width
    2. Versicolor is distributed in a average manner and average features in case of petal length & width
    3. Virginica is highly distributed with large no .of values and features in case of sepal length & width

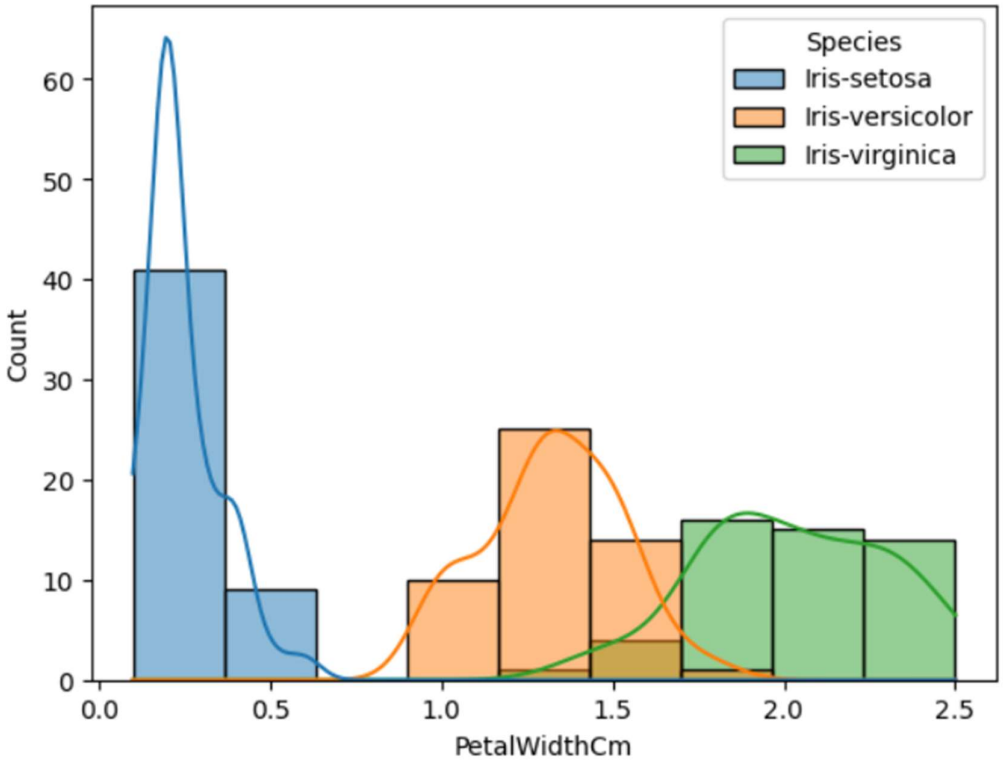
4. High density values are depicting the mean/median values, for example: Iris Setosa has highest density at 5.0 cm (sepal length feature) which is also the median value(5.0) as per the table.

Plotting the Histogram & Probability Density Function (PDF)






Species
Iris-setosa
Iris-versicolor
Iris-virginica
Count
PetalLengthCm
0
10
20
30
40
50
60
70
80
1
2
3
4
5
6
7


Species
Iris-setosa
Iris-versicolor
Iris-virginica
Count
PetalWidthCm
0
10
20
30
40
50
60
0.0
0.5
1.0
1.5
2.0
2.5

Data Insights:

1. Plot 1 shows that there is a significant amount of overlap between the species on sepal length, so it is not an effective Classification feature

2. Plot 2 shows that there is even higher overlap between the species on sepal width, so it is not an effective Classification feature

3. Plot 3 shows that petal length is a good Classification feature as it clearly separates the species . The overlap is extremely less (between Versicolor and Virginica) , Setosa is well separated from the rest two

4. Just like Plot 3, Plot 4 also shows that petal width is a good Classification feature . The overlap is significantly less (between Versicolor and Virginica) , Setosa is well separated from the rest two