**Project 1 – Iris Dataset Basic Analysis**

**1.Dataset Overview :**

This Dataset represents the dimensions of the species of iris flower ,It consists of 150 samples of iris flowers, each belonging to one of three species: Setosa, Versicolor, and Virginica. For each sample, four features are measured: length and width of the sepals and petals.

**2.Column Descriptions**

- **Id** : Unique identifier for each row

- **SepalLengthCm** : Represents the length of the sepal in centimeters.

- **SepalWidthCm** : Represents the width of the sepal in centimeters.

- **PetalLengthCm** : Represents the length of the petal in centimeters.

- **PetalWidthCm** : Represents the width of the petal in centimeters.

- **Species** : Specifies the species of each flower as : "Iris-setosa" or "Iris-versicolor" or "Iris-virginica".

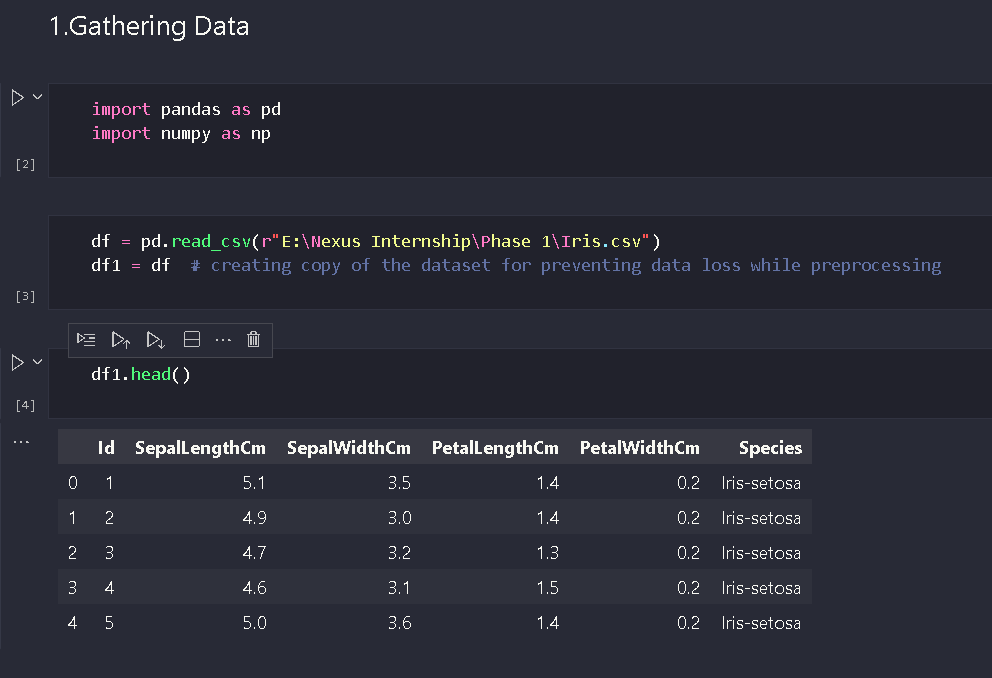
**3.Mannual Analysis Conclusion**

By overviewing the dataset manually there are no any visible issues so, lets check the data programatically for any issues further.

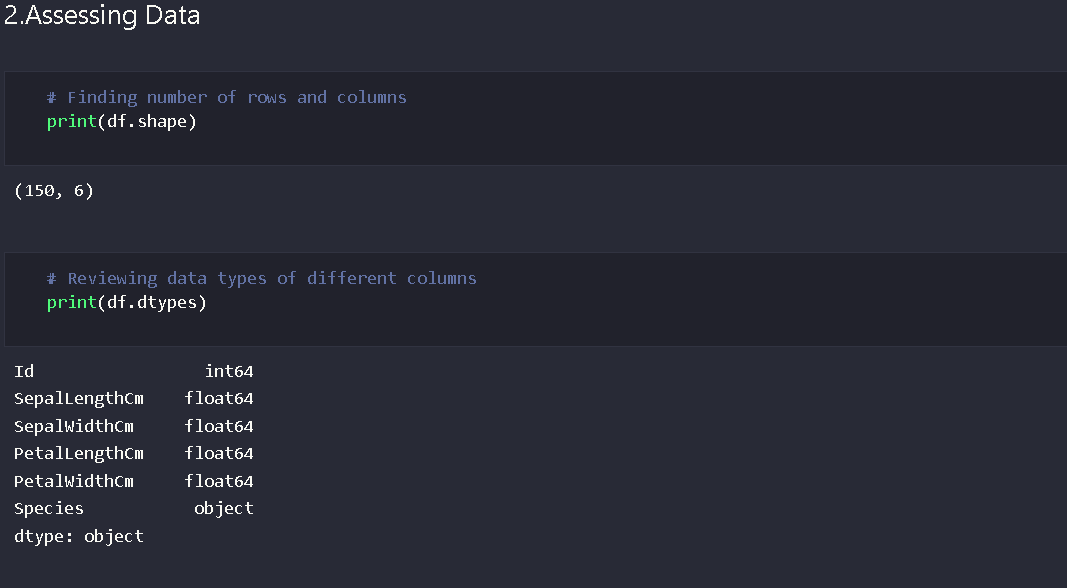
**4.Data Wrangling Processes:**

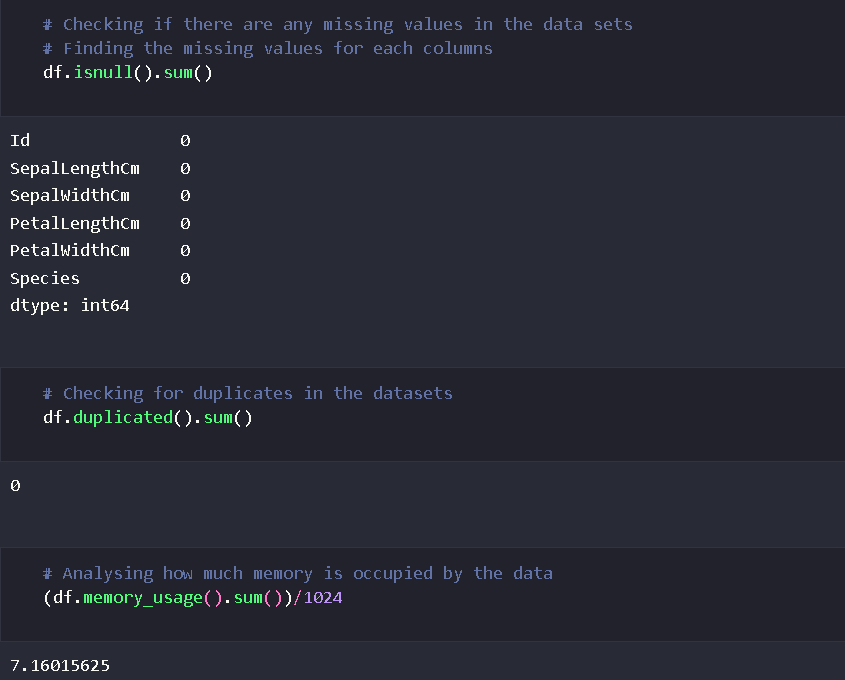
**4.1.Gathering the data from data source :**

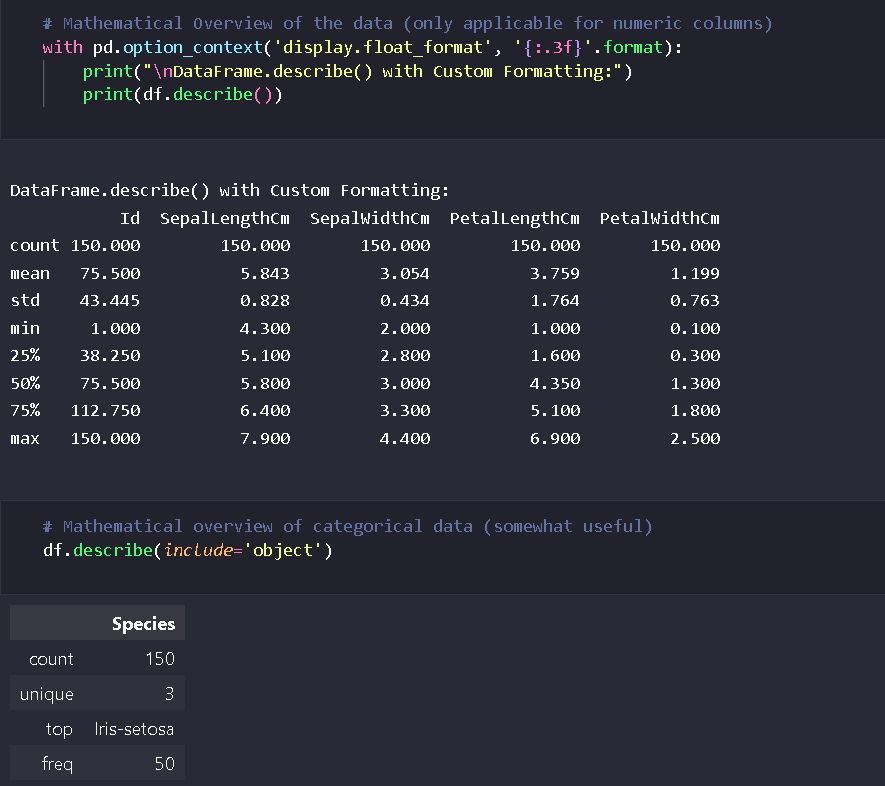
(Here the data source is my local machine)

****

**4.2. Assessing the data :**



****

****

**4.3. Cleaning the data :**

There are no any duplicates,null,missing and invalid values in the dataset and the dataset is already clean.

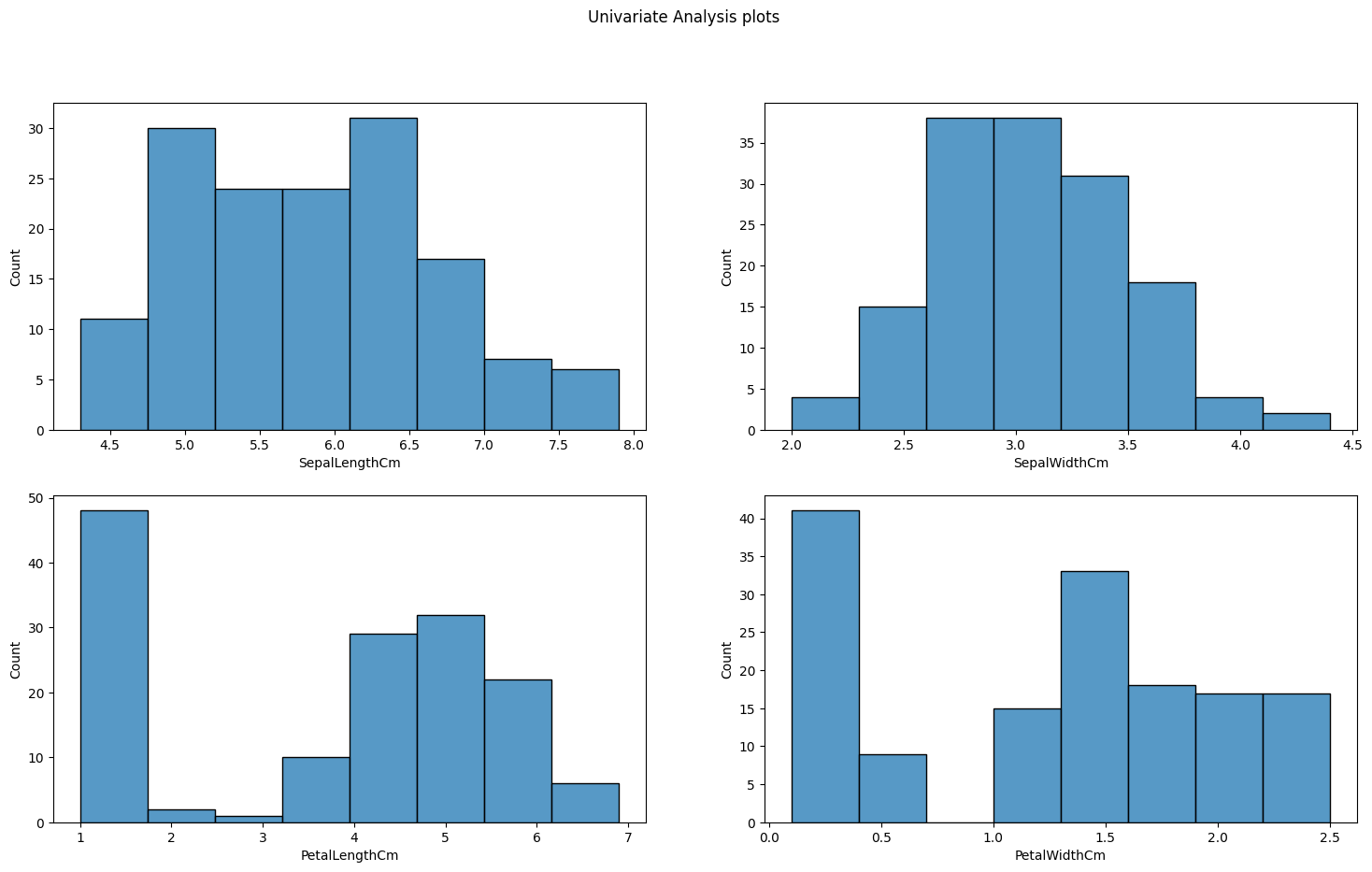
**Conclusion after data wrangling :**

So far , there are no any issues in the dataset we will get to know more about the data quality by performing further analysis and visualizations so, at this stage the data is already clean as there are no any missing,duplicate or messy values hence we will skip data cleaning

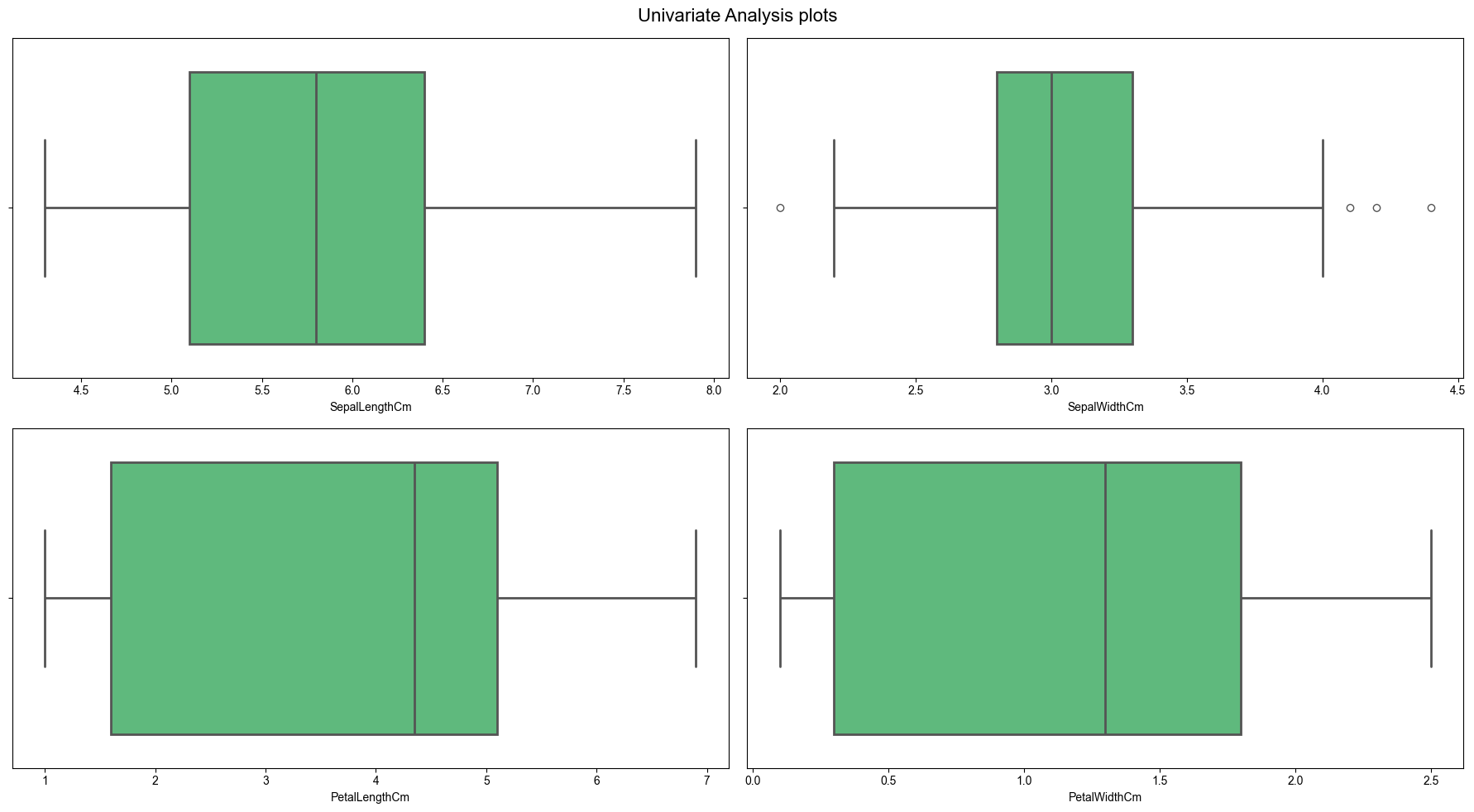
**5. Univariate Analysis :**

**5.1. Univariate analysis on numerical columns :**

1.Values count for each attribute :



2.Distribution of each attribute :



**Median Values for above distribution :**

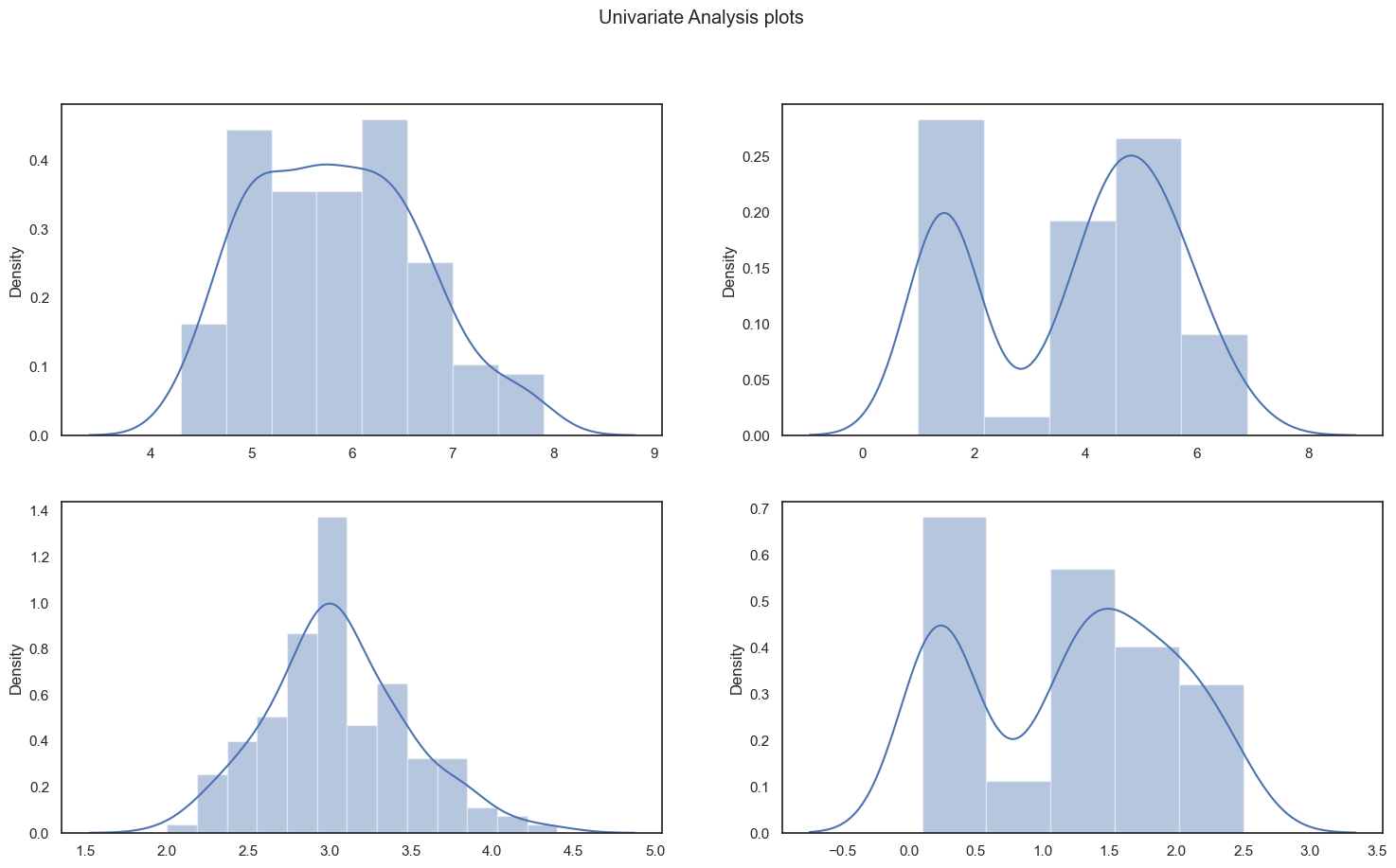
-The median value of SepalLengthCm is : 5.8

-The median value of SepalWidthCm is : 3.0

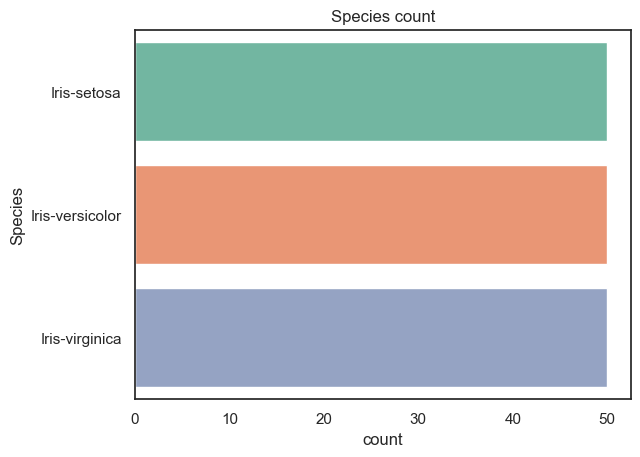
-The median value of PetalLengthCm is : 4.35

-The median value of PetalWidthCm is : 1.3

3.Density of datapoints for each attribute :



**5.1. Univariate analysis on categorical columns :**



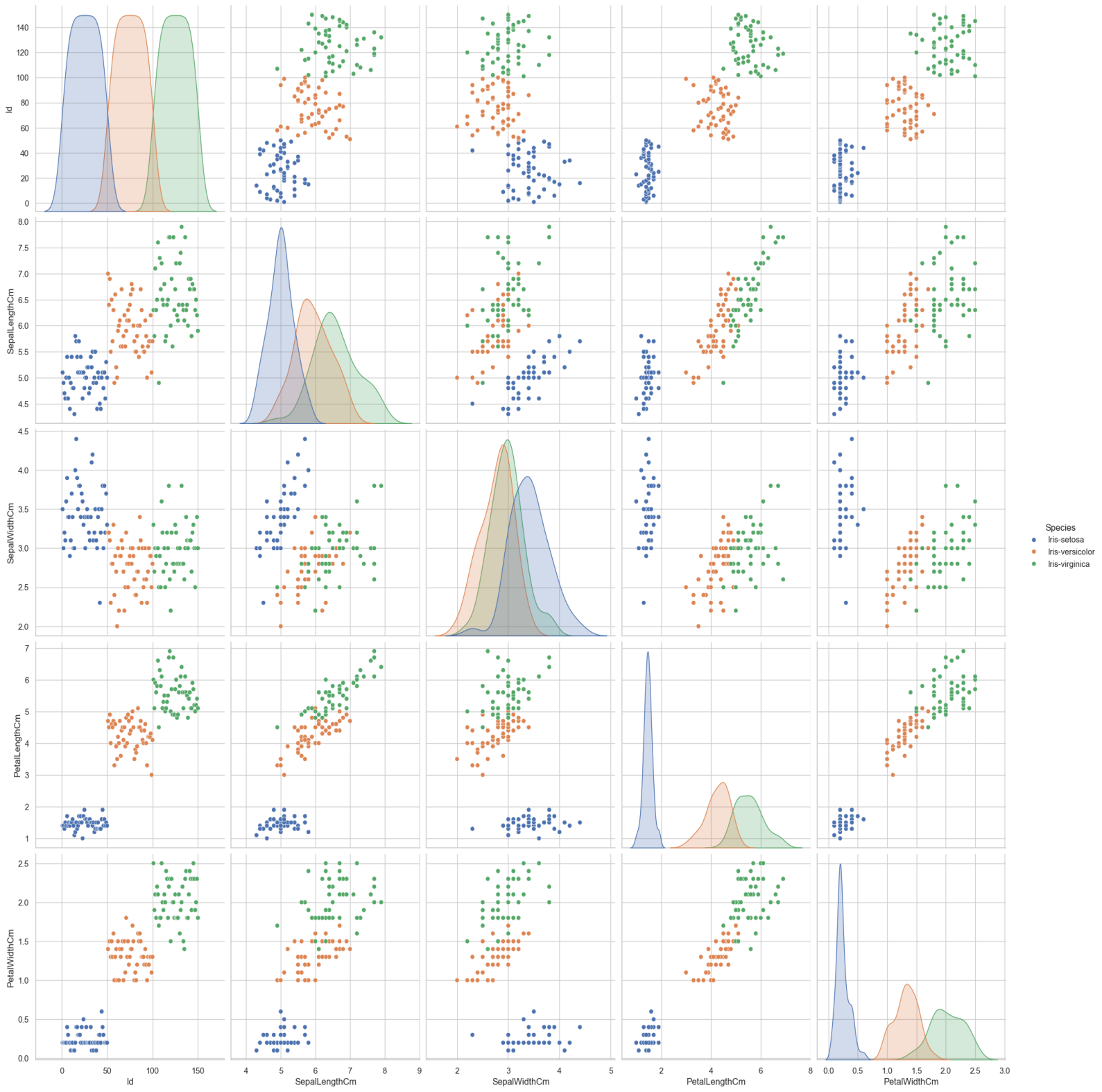
- This summarize that each species data is equaly distributed i.e. 50 entries for each species

- This visualization specifies that the species distribution is well balanced

- Each species has 50 count

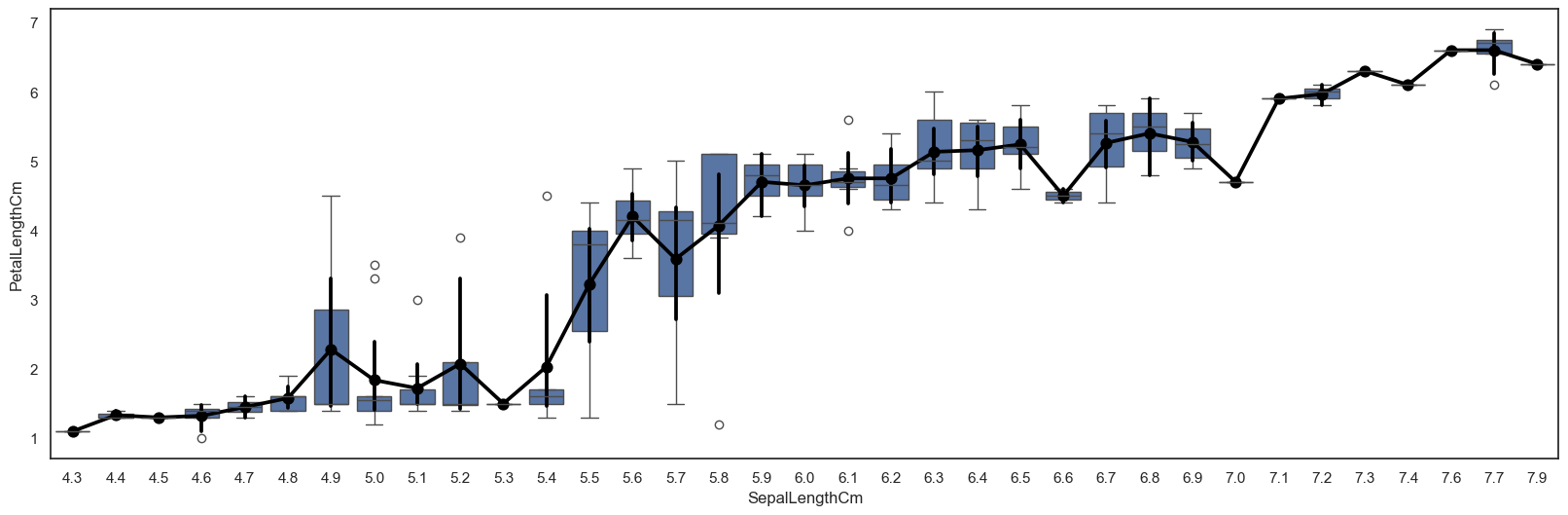
**5. Bivariate Analysis :**

(Rough pairplots for overall understanding the patterns and relations among the attributes)

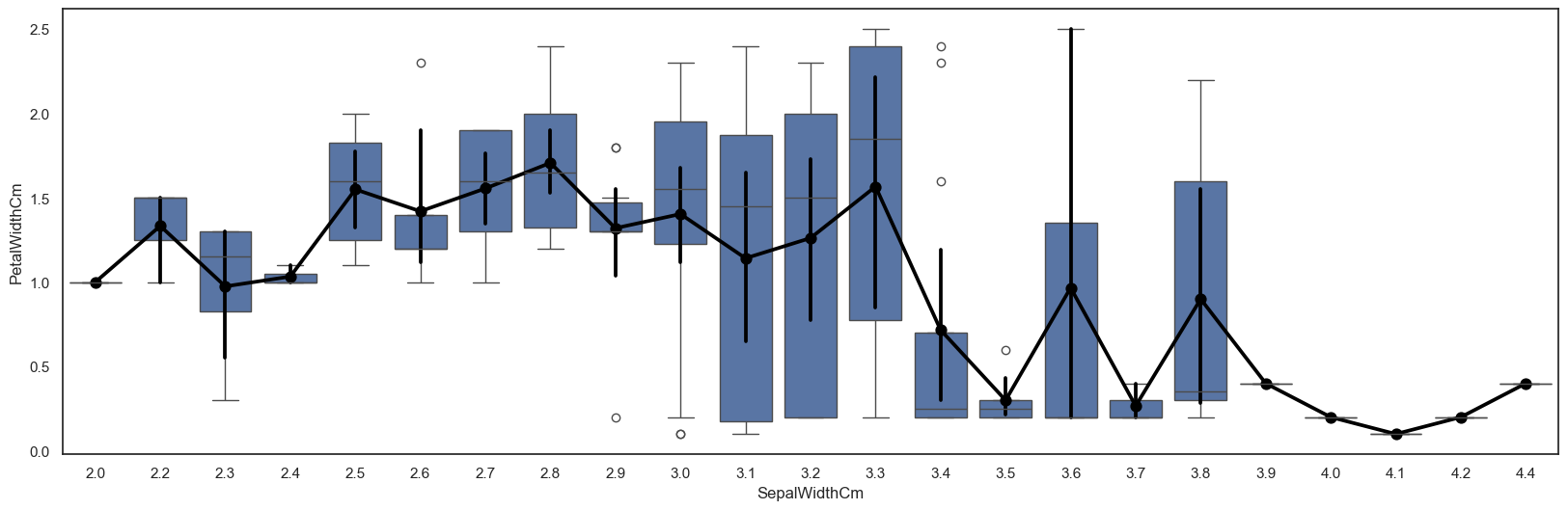
****

**5.1. Finding the Central Tendencies between most relating attributes :**

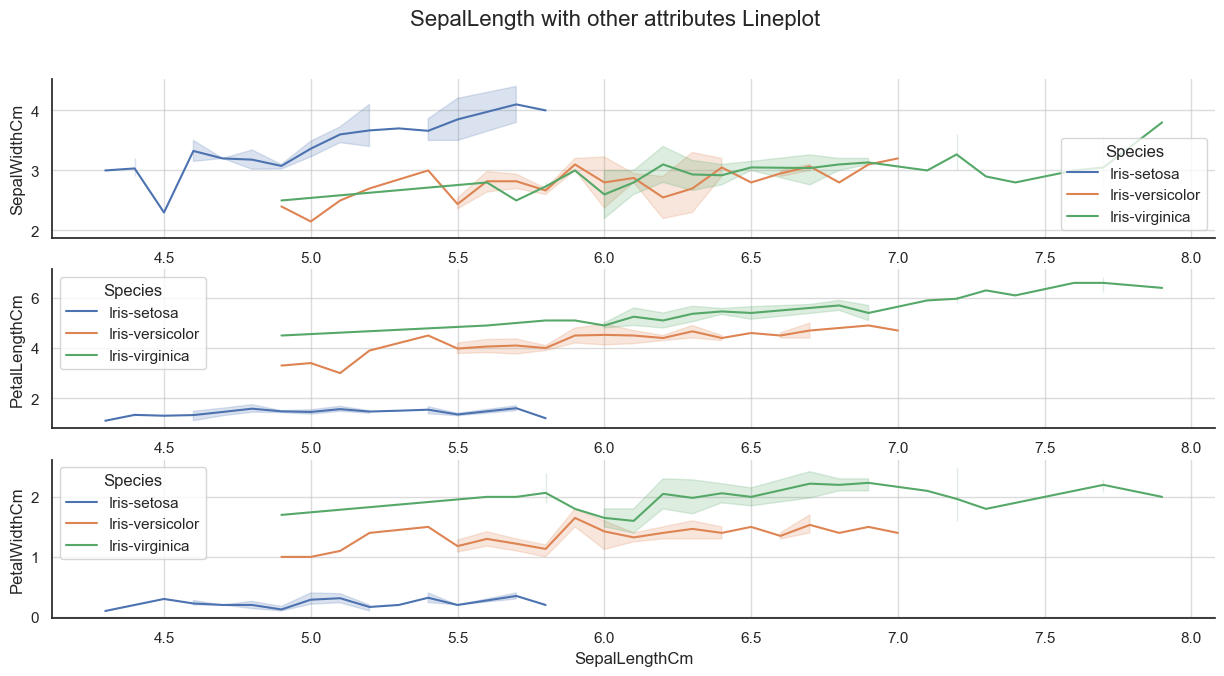
Differentiating between central tendencies of SepalLength with PetalLength –

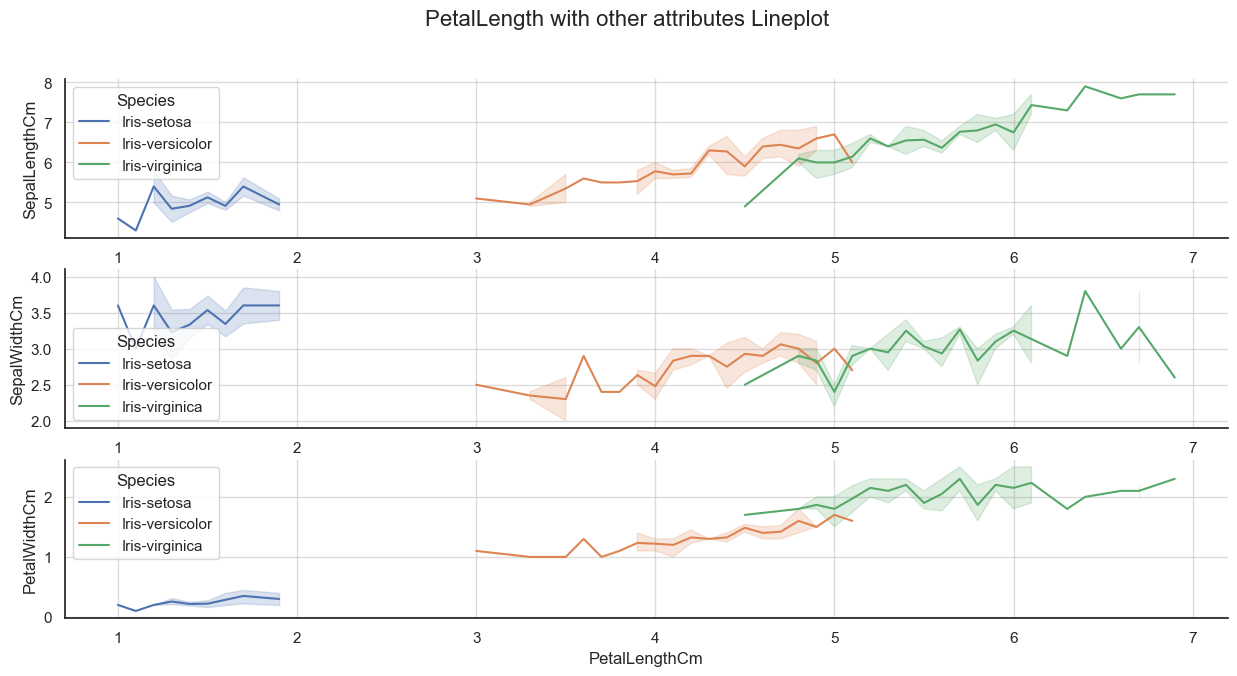


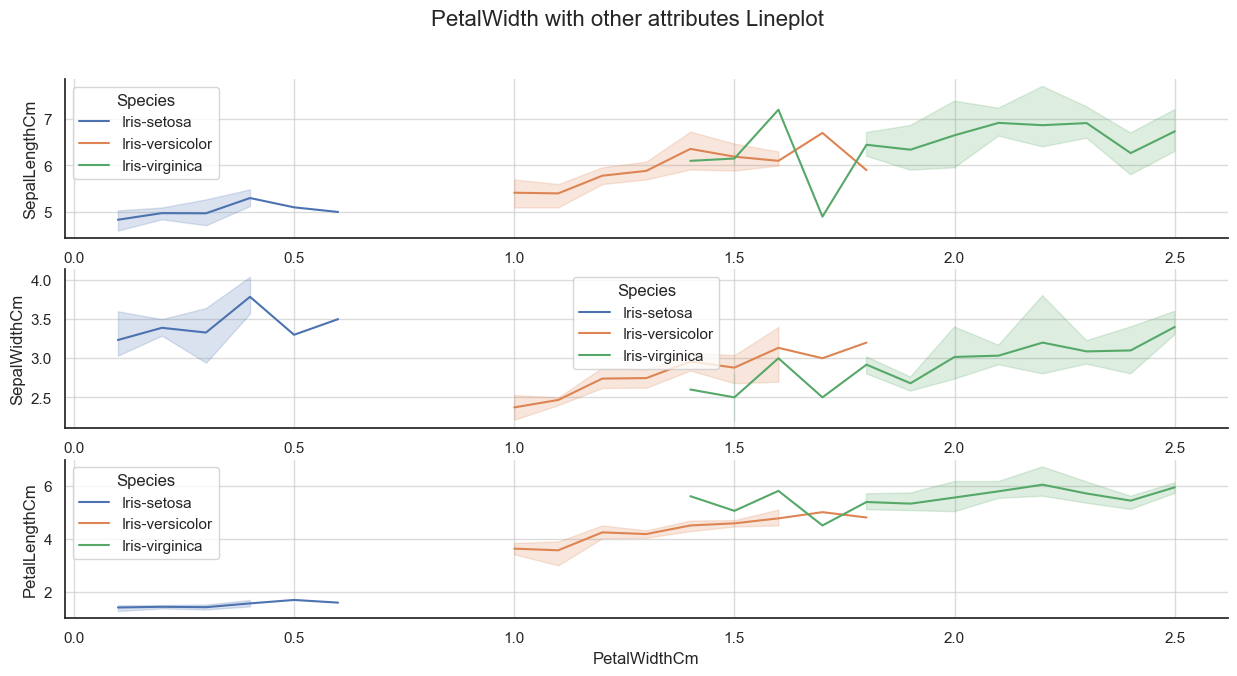
Differentiating between central tendencies of SepalWidth with PetalWidth –



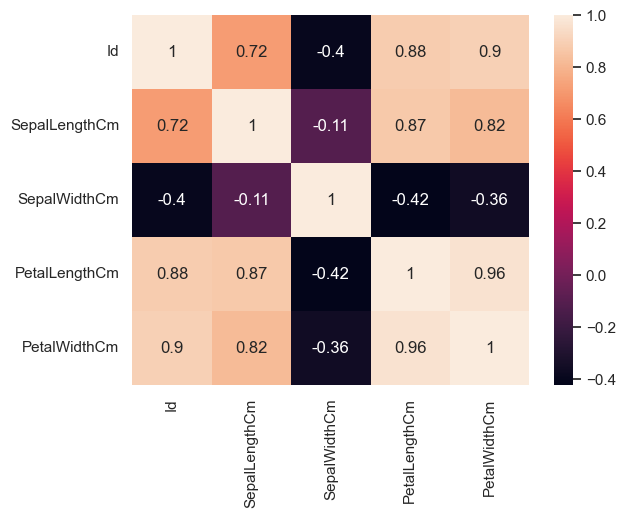
**6.Multivariate Analysis :**





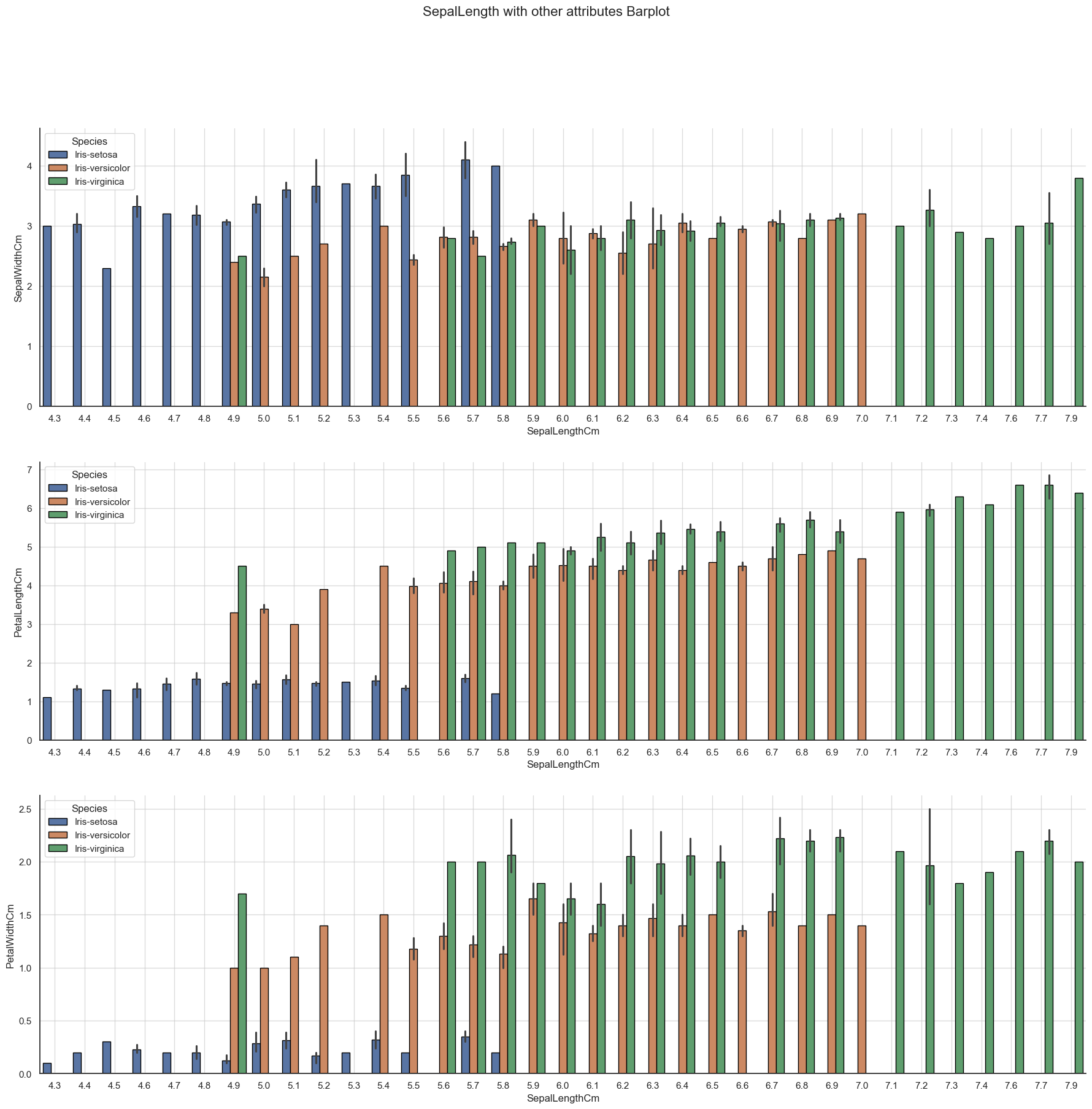


**6.1.Correlation Between the attributes :**



(ignore the id attribute)

(Random barplot for understanding the sepal length relation with other attributes)



**Insights gained after visualizing and analyzing the dataset:**

1. The dimensions of sepal and petal vary among different species of iris.
2. Specifically iris versicolor and iris virginica have larger petals compared to iris setosa.
3. Sepals are usually larger and broader than petals.
4. There are correlations between the features such as, petal length and petal width tend to be positively correlated, meaning as one increases, the other tends to increase as well. Similarly, sepal length and petal length might also be correlated.
5. The size of these flower may be approximately in following order :

Iris Setosa < Iris Versicolor < Iris Virginica.

1. Almost all the species are highly correlated hence there is a higher chance that the dimensions will change if the values of one attribute changes.