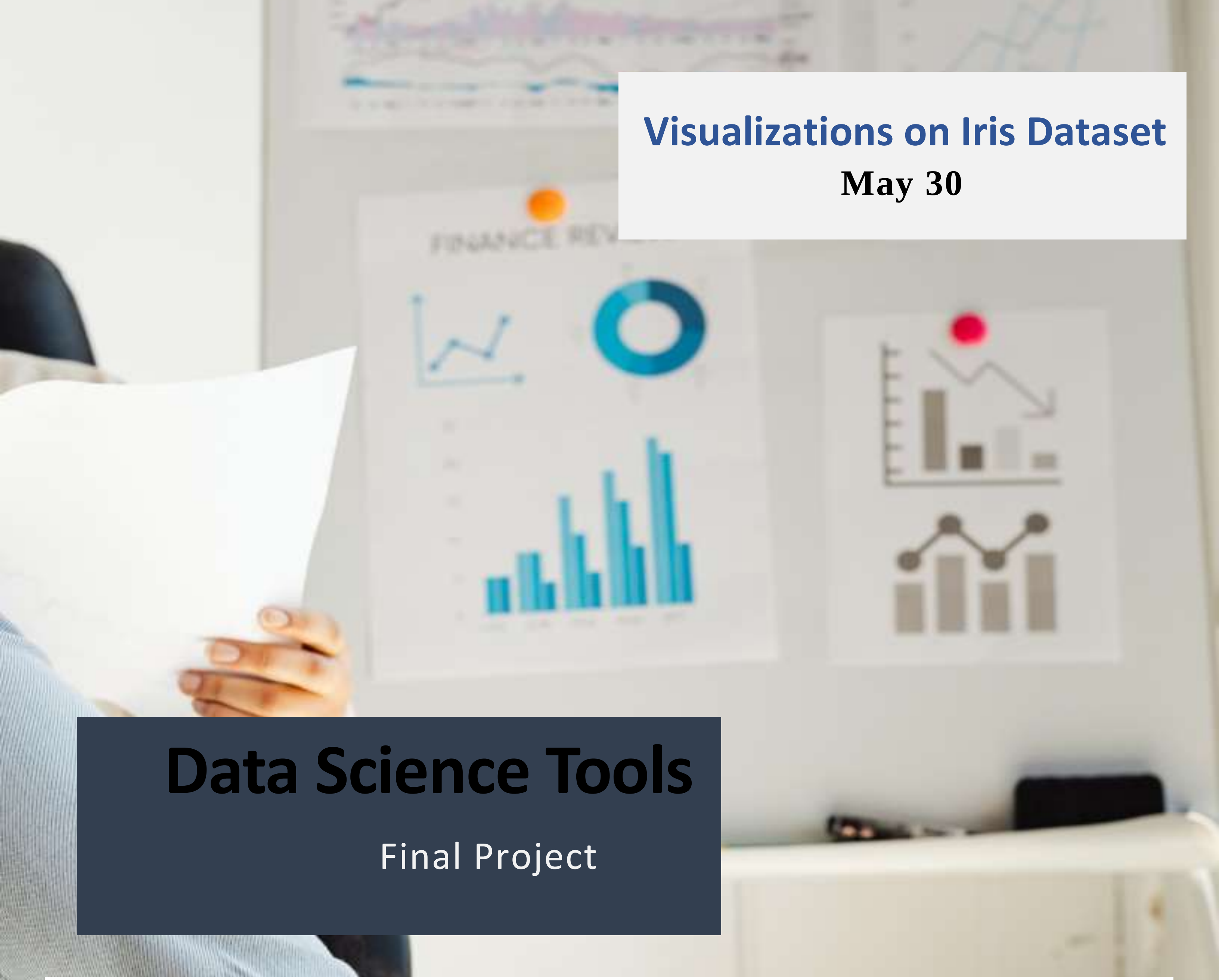


Visualizations on Iris Dataset

May 30

Data Science Tools

Final Project

Abstract

In this project we are going to apply some data visualization techniques on Iris dataset using variety of data visualization tools not only Python and MATLAB but also, R, Tableau and Power BI.

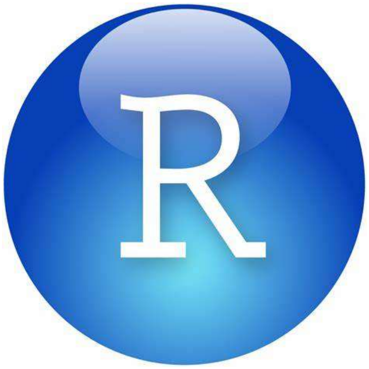
And finally, we will comment on each data visualization tool, which one we liked the most and with one is the most useful and which one has the most diversity in plots, shapes and ideas.

Data Visualization is one of the best ways in communicating results about the datasets.

Participants Information

Name	ID
Amr Mamdouh Gaber Ibrahim Shaltuot	20201497739
Abdelrahman Ashraf Saaed Abd El-Aziz	20201376979
Karim Hussam Al-Din El-Sayed Mohamed	20201446854
Abdelrahman Mohamed Ali Mohamed	20201446699

Our Tools



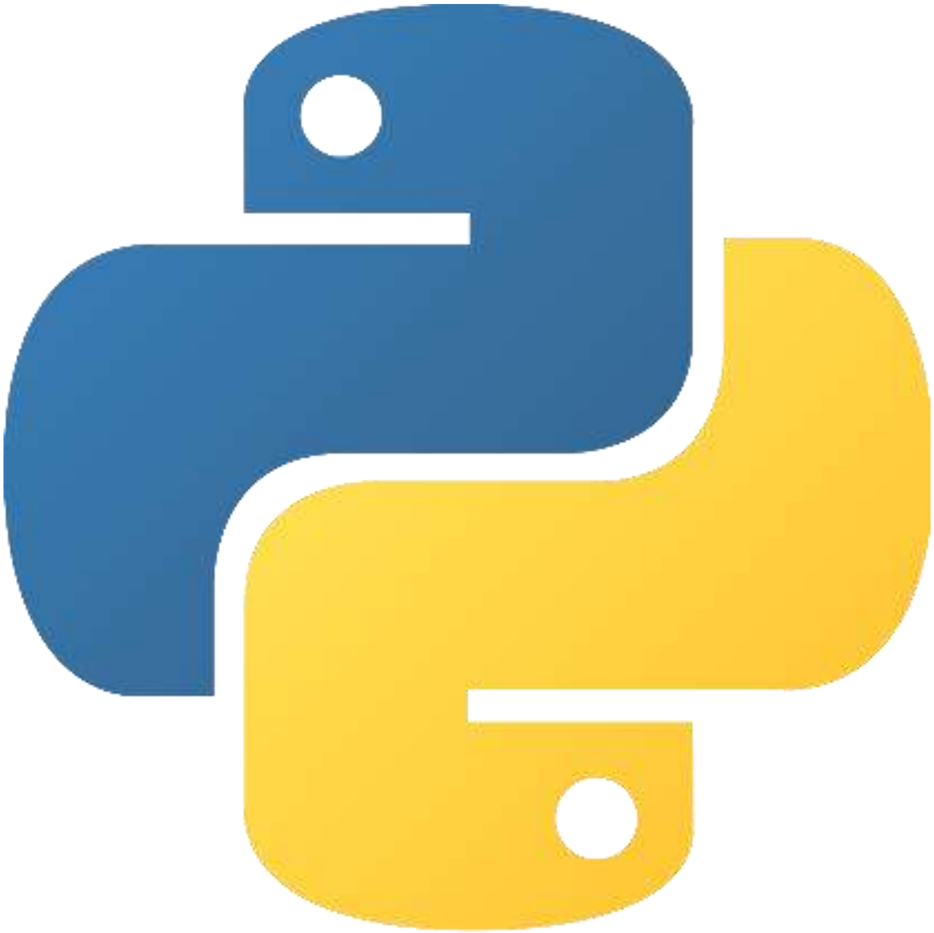
Explanation of the iris dataset attributes and details

The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These measures were used to create a linear discriminant model to classify the species. The dataset is often used in data mining, classification and clustering examples and to test algorithms.

Attribute	***Count***
Sepal Length in Cm	*150*
Sepal Width in Cm	*150*
Petal Length in Cm	*150*
Petal Width in Cm	*150*
Species	*3*

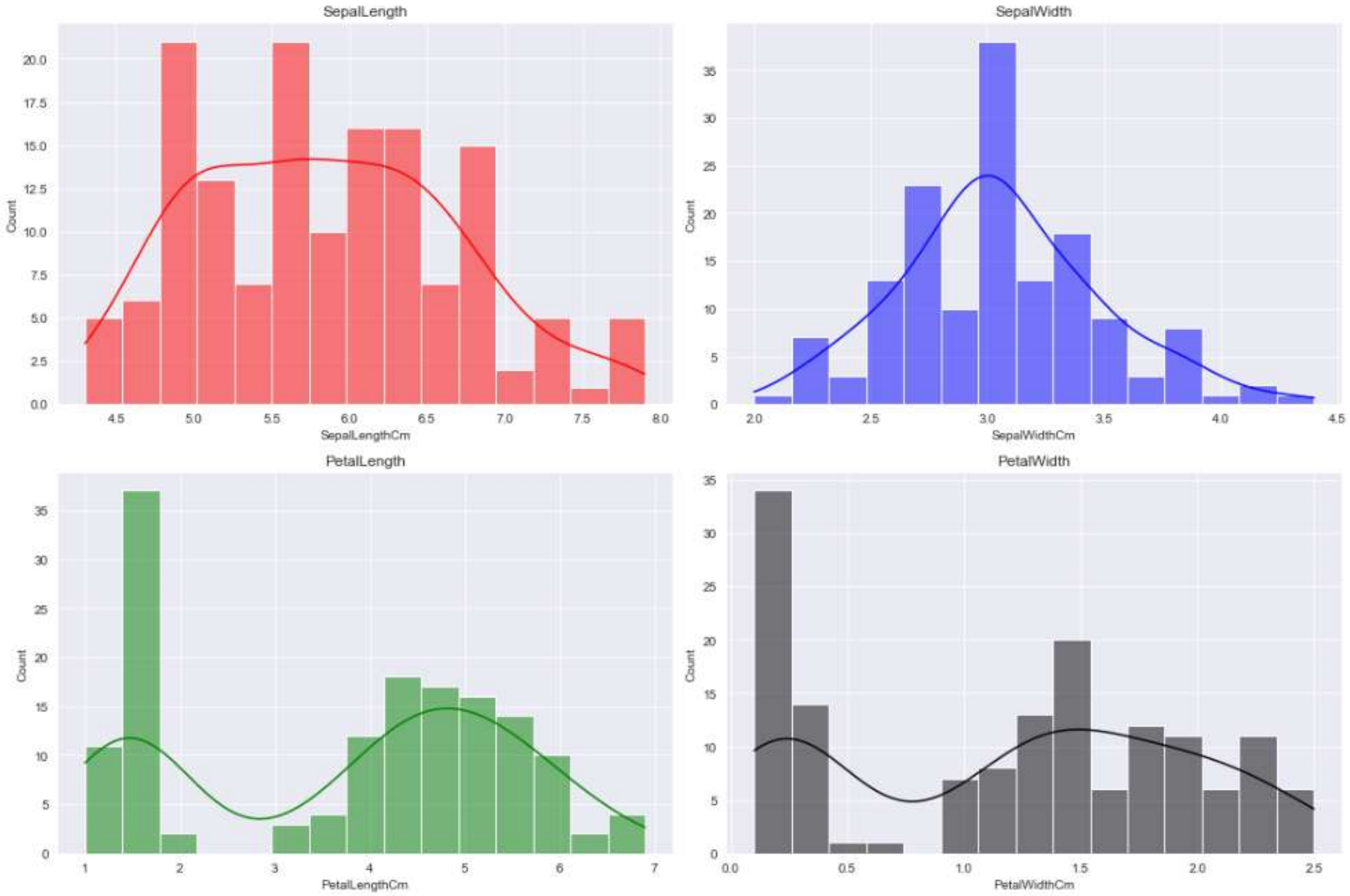
Iris dataset is the Hello World for the Data Science, so if you have started your career in Data Science and Machine Learning you will be practicing basic ML algorithms on this famous dataset. Iris dataset contains five columns such as Petal Length, Petal Width, Sepal Length, Sepal Width and Species Type. Iris is a flowering plant, the researchers have measured various features of the different iris flowers and recorded digitally.

Dataset Characteristics	*Multivariate*	***Number of Instances***	*150*	***Area***	*Life*
Attribute Characteristics	*Real*	***Number of Attributes***	*4*	***Data Donated***	*1988-07-01*
Associated Tasks	*Classification*	***Missing values?***	*No*	***Number of Web Hits***	*4689377*



Python

1. Histogram for Target data

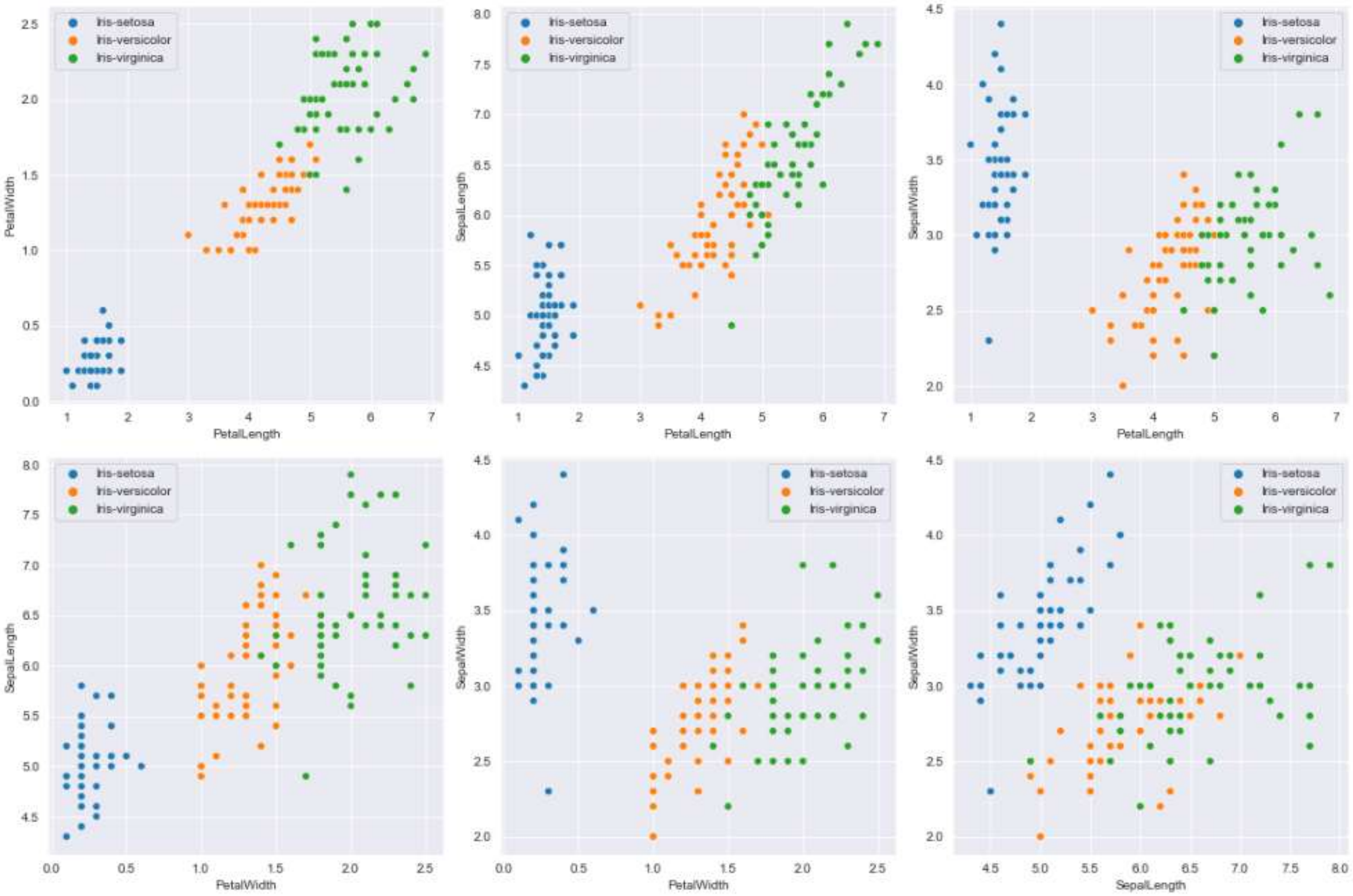


Conclusion

Sepal Width feature follows normal distribution.

While **other** features tend to be right skewed more.

2. Scatter Plot for every 2 attributes

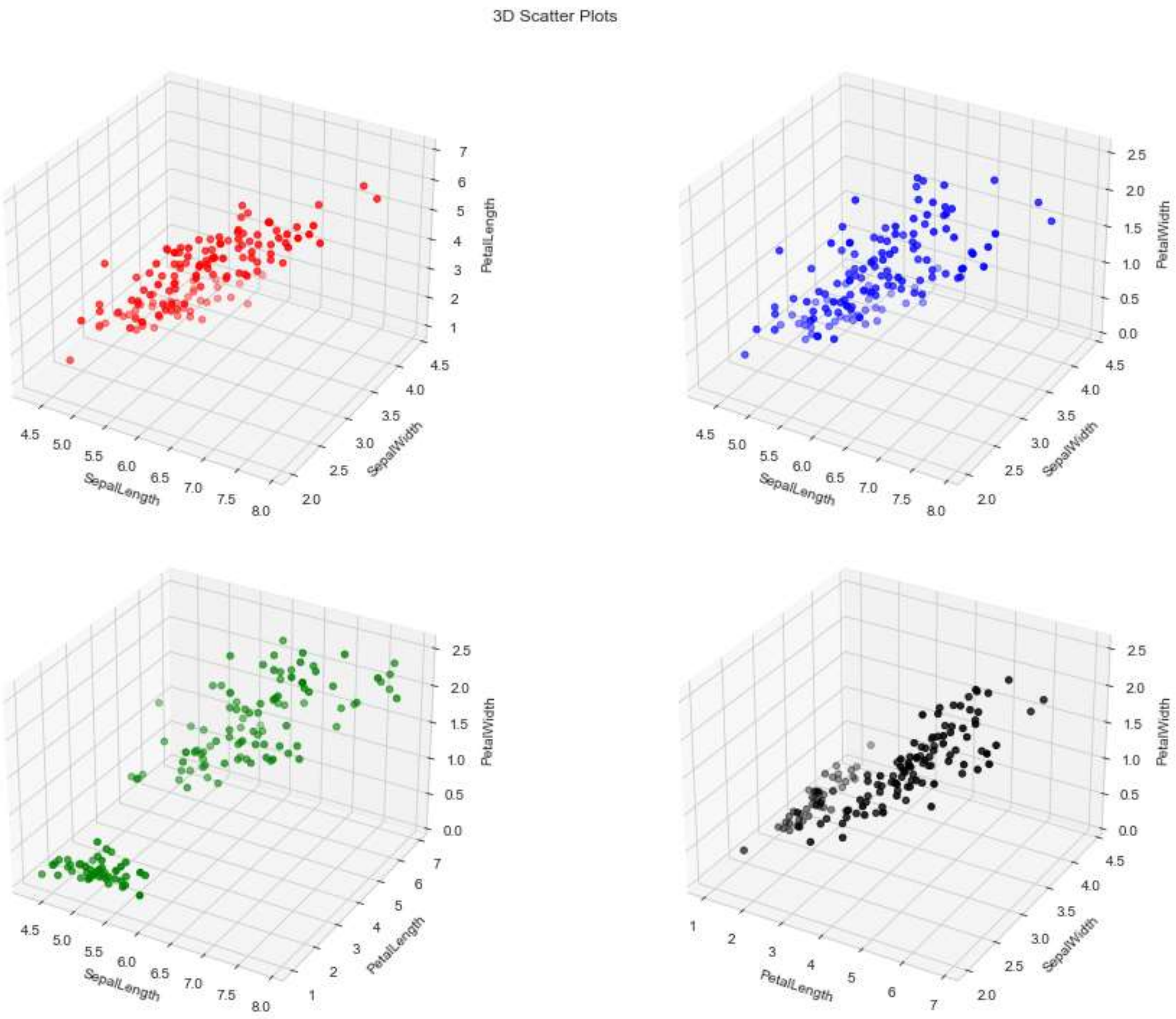


Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.

So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

3. 3D Scatter Plot for every 3 attributes



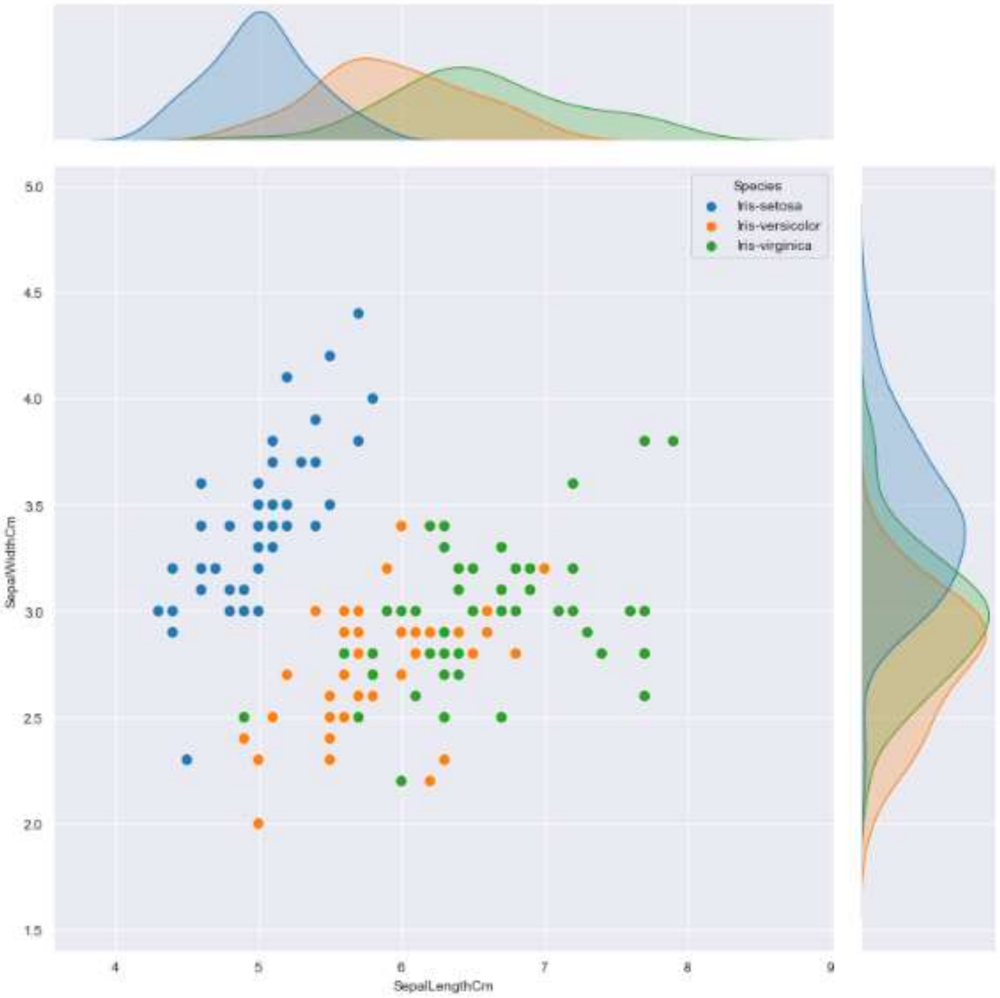
Conclusion

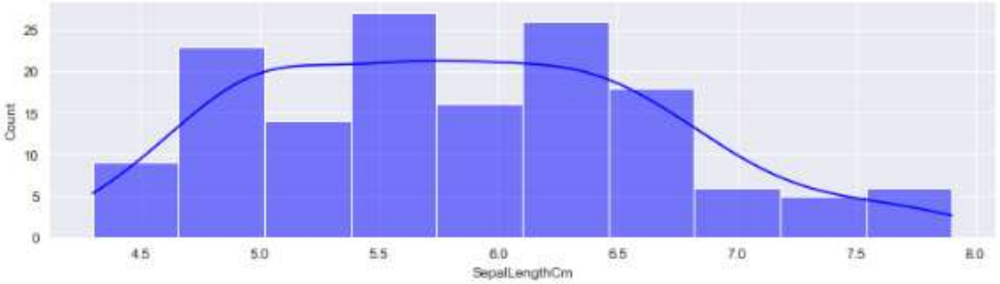
As it seems in the scatter plots data are already been classified and separated into clusters.

It's not colored but, the clusters can also be recognized.

So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

4. Jointplot and KDE plot, Histogram

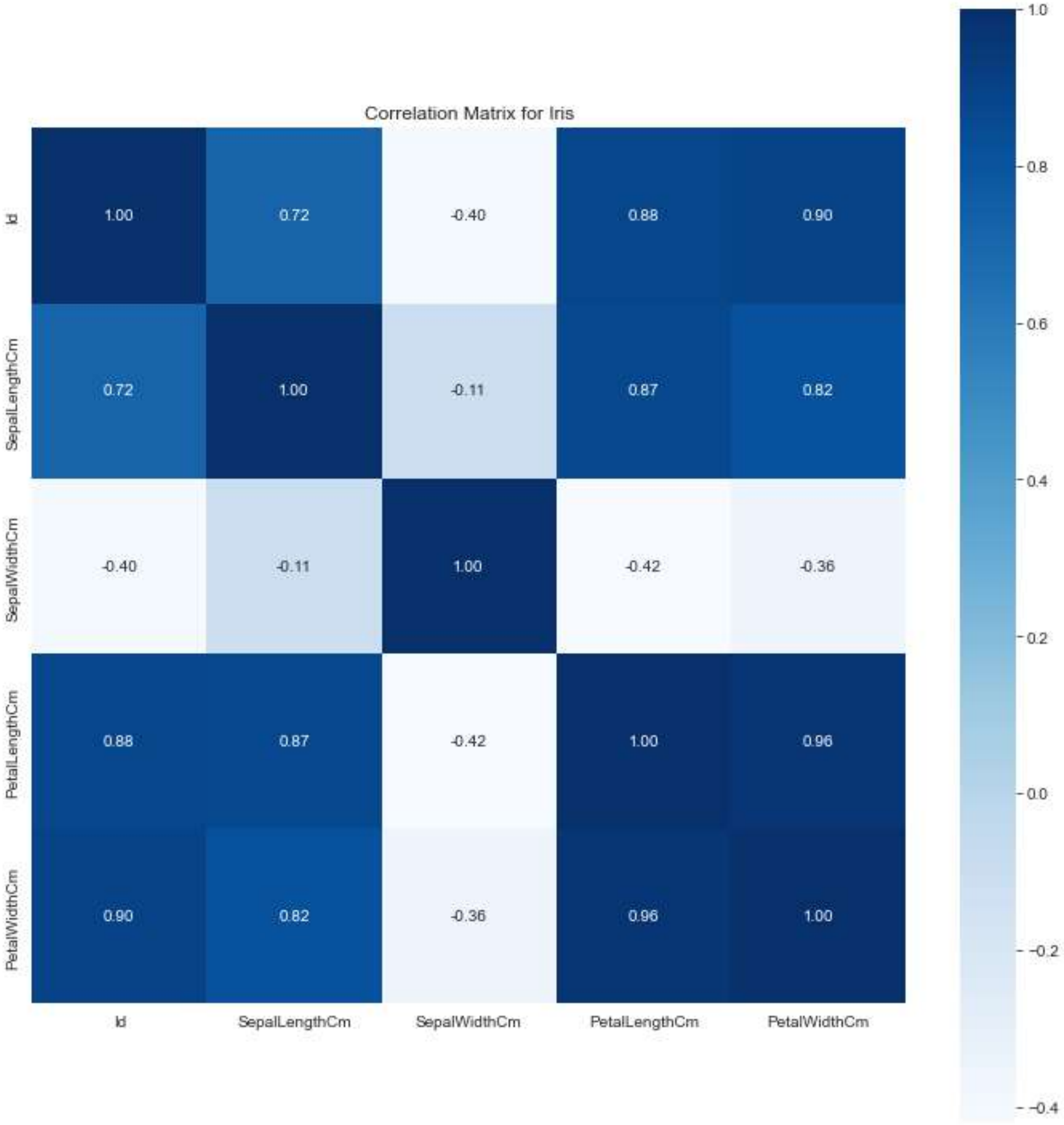




Conclusion

Well, it was kind of easy to plot such view using **sns.jointplot** method in *seaborn*

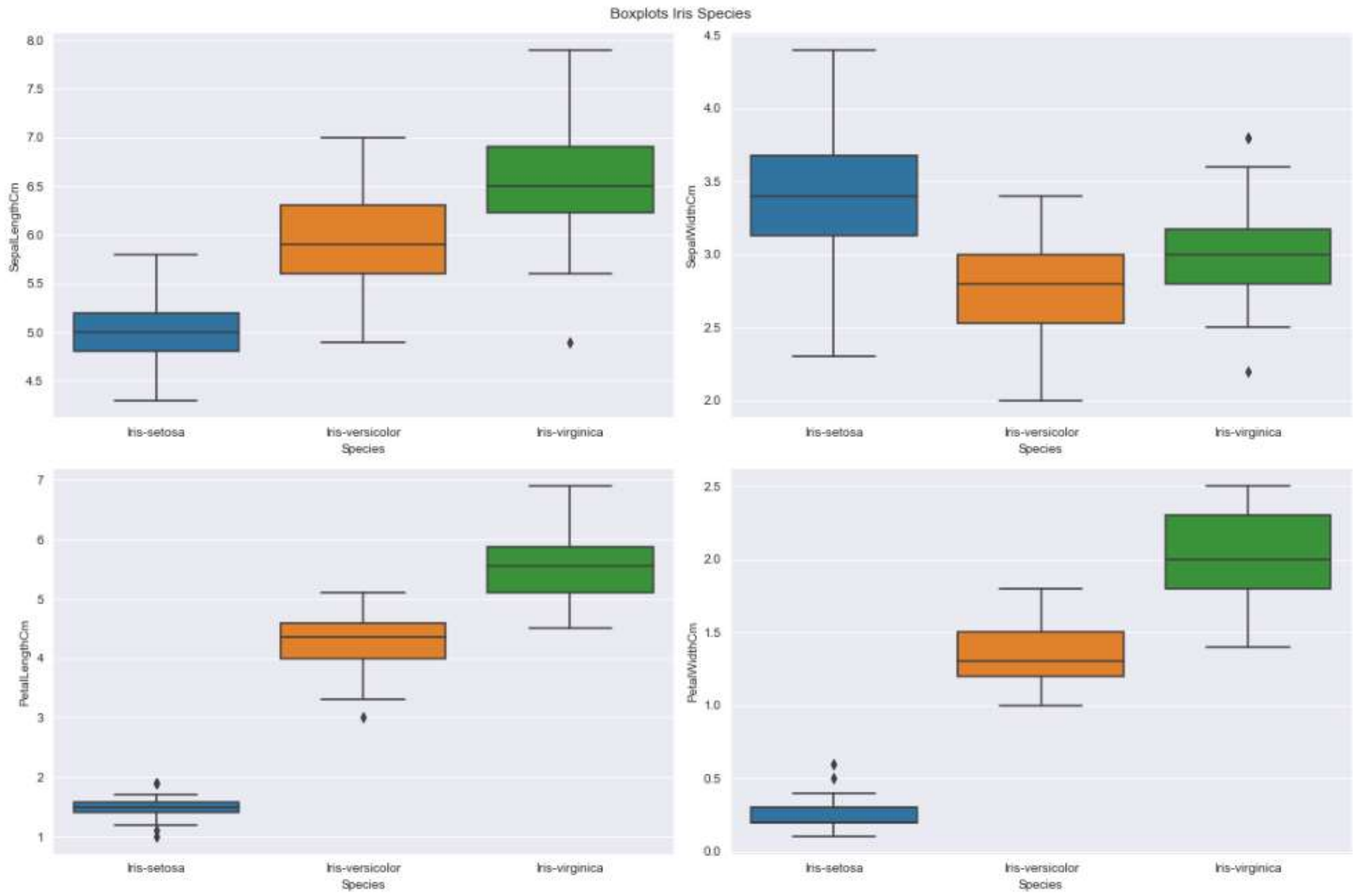
5. Correlation Matrix



Conclusion

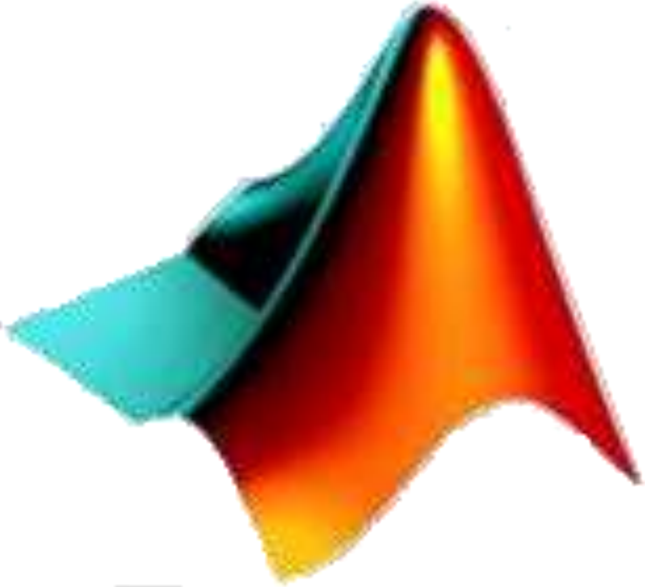
As it seems, the all the features are **highly positively correlated** except for the ***SepalWidth*** attribute with all the features is **negatively correlated** and almost there is **no correlation** value between most of them.

6. Boxplots for each feature



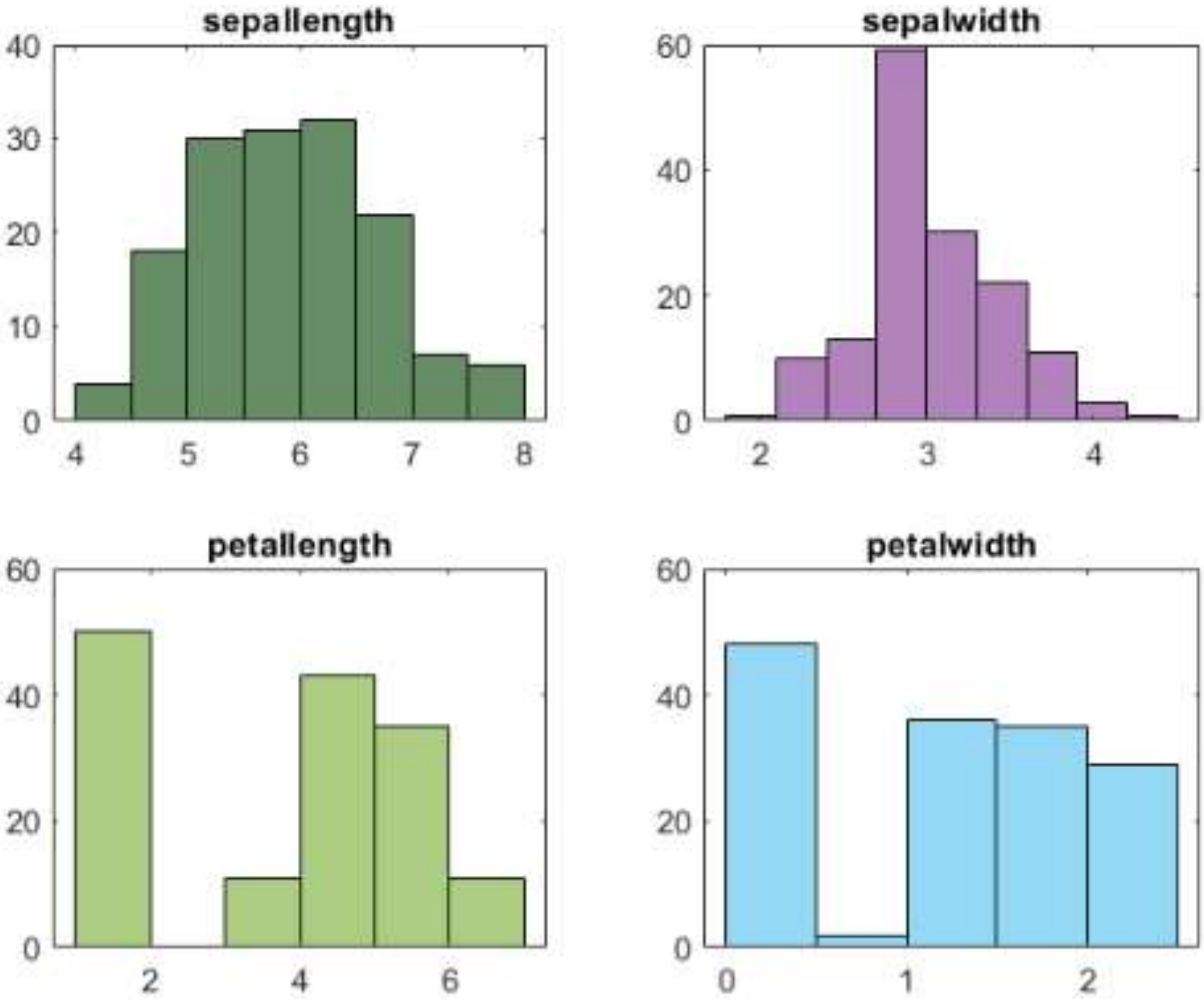
Conclusion

Seems as an overall that there are **no outliers** even one or two maybe, and the average lengths of **Sentosa's SepalWidth are tall** but for **other features are short**. For the **Versicolor** it has an **average height** for all features except the **SepalWidth feature it's the smallest**. and finally, the **tallest Specie is Virginica** for all features but, for the **SepalWidth it has an average height**.



MATLAB

1. Histogram for Target data

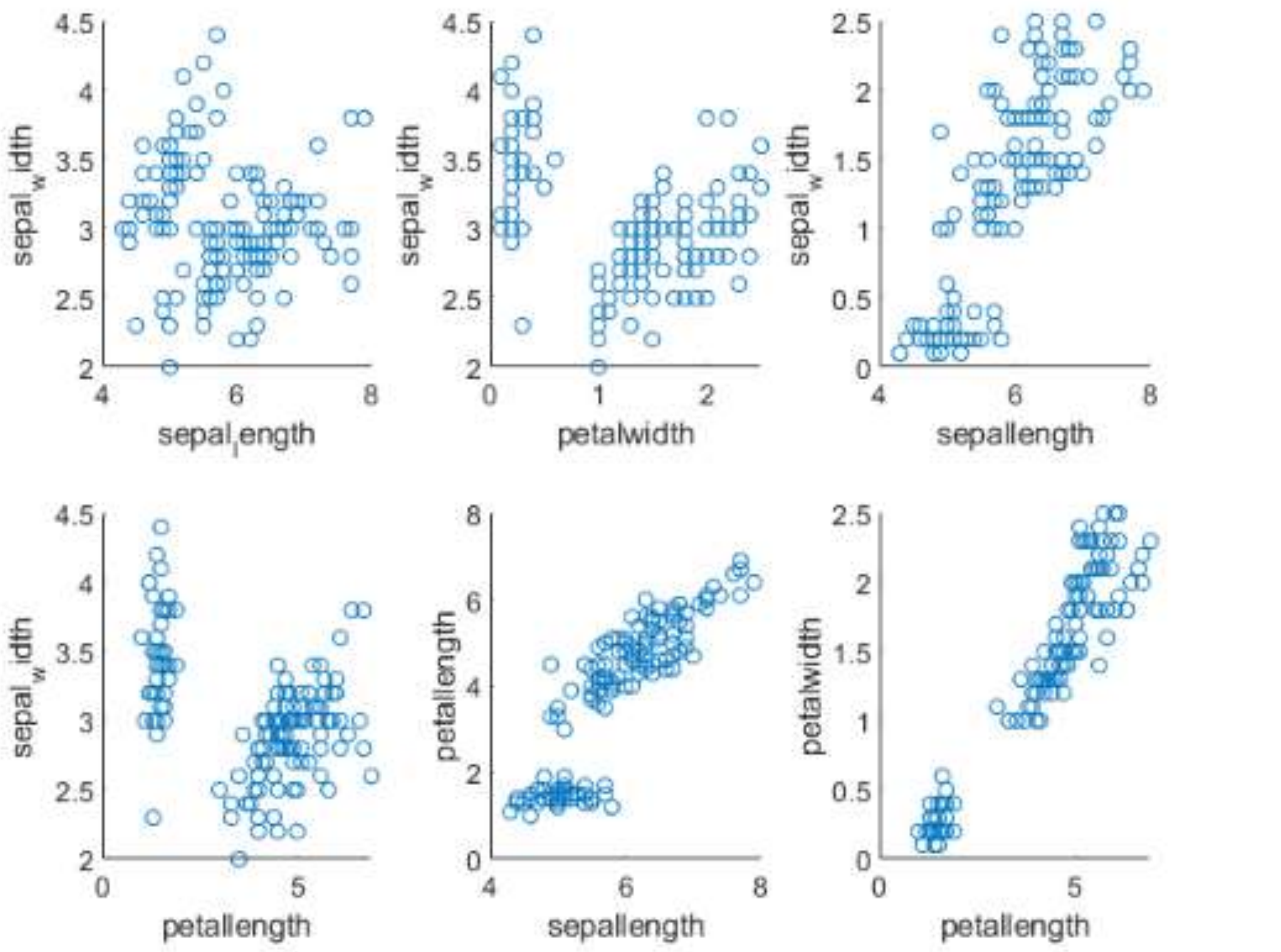


Conclusion

Sepal Width feature follows normal distribution.

While **other** features tend to be right skewed more.

2. Scatter Plot for every 2 attributes

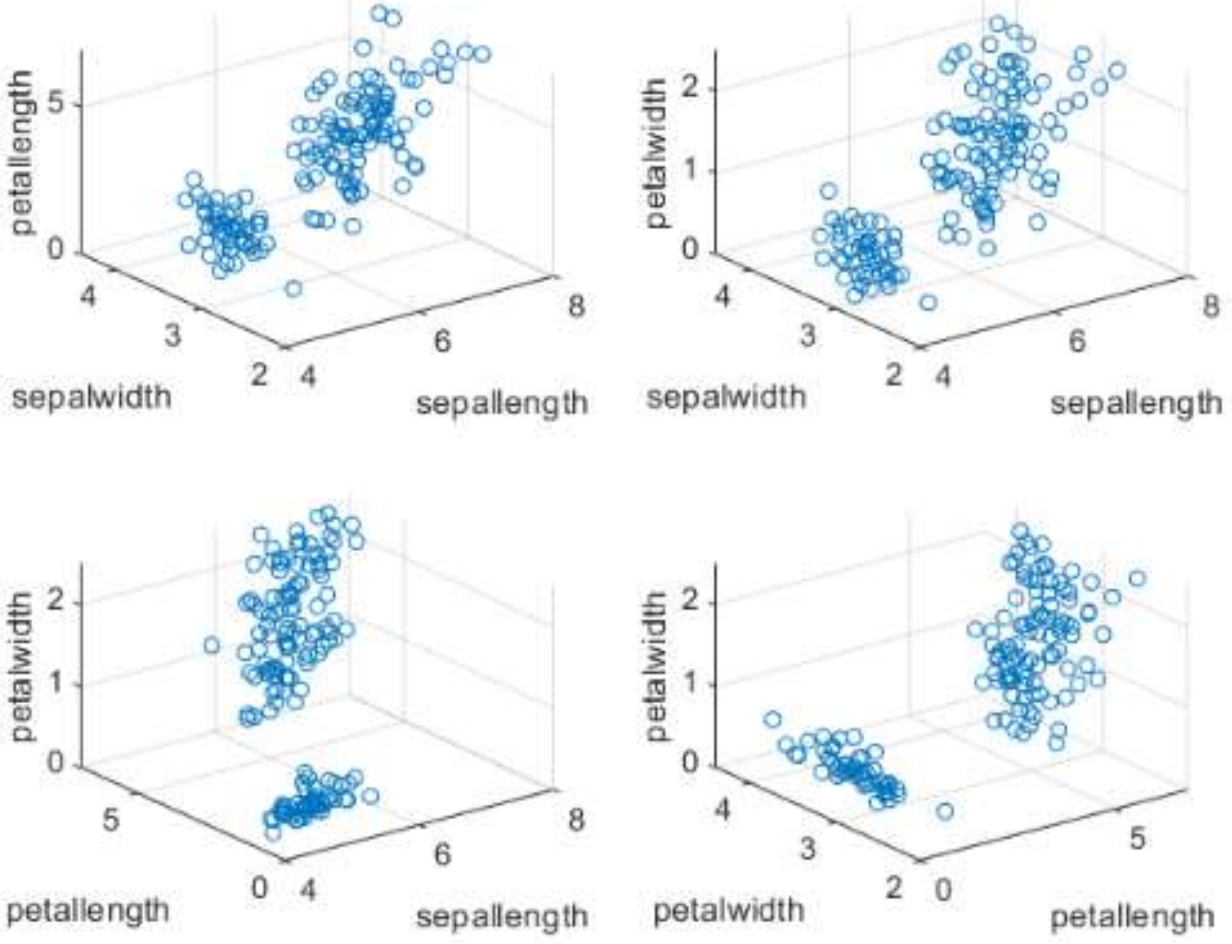


Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.

So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

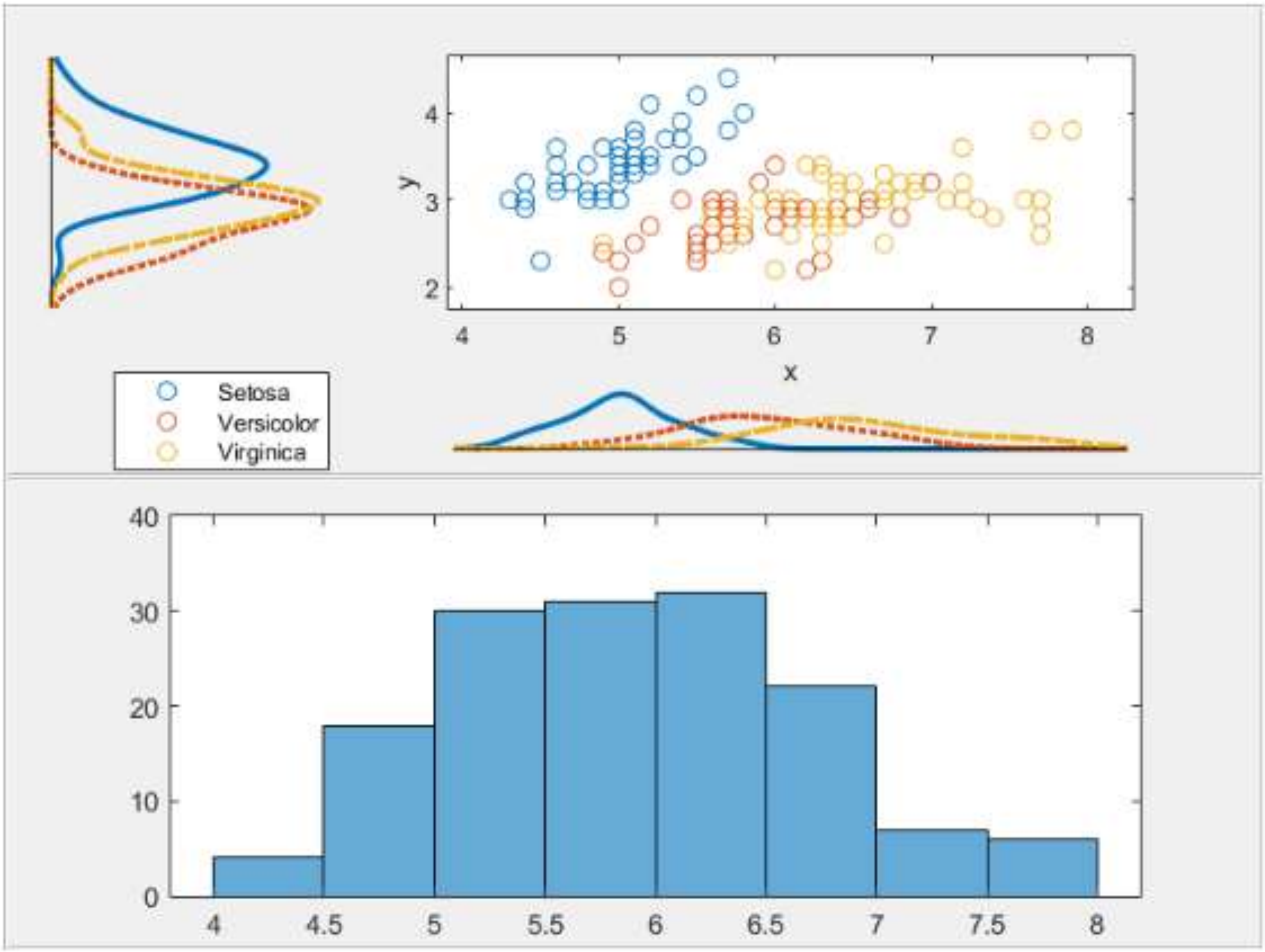
3. 3D Scatter Plot for every 3 attributes



Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
It's not colored but, the clusters can also be recognized.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

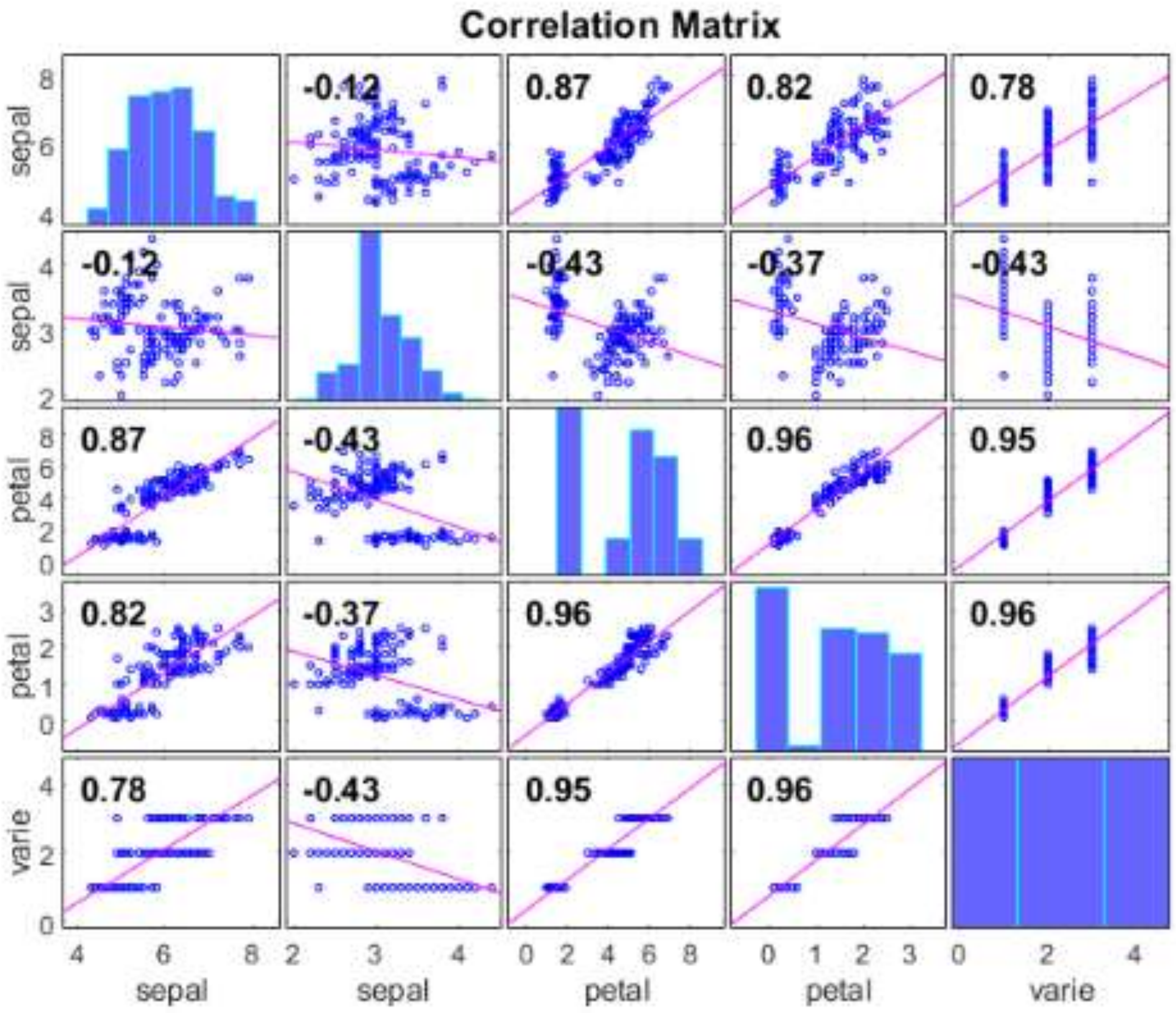
4. Jointplot and KDE plot, Histogram



Conclusion

Well, it was kind of easy to plot such view using **unipanel** and **scatterhist** plot in MATLAB.

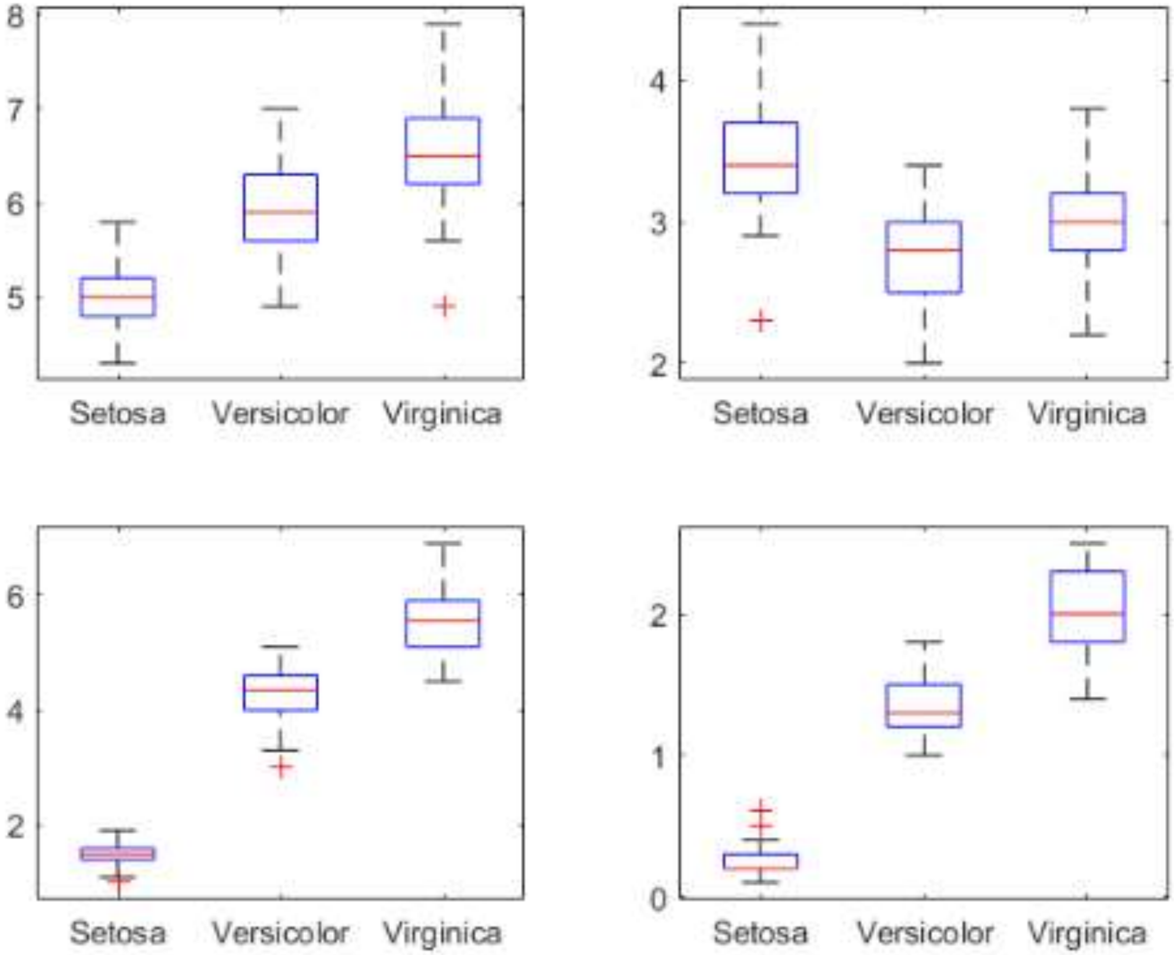
5. Correlation Matrix



Conclusion

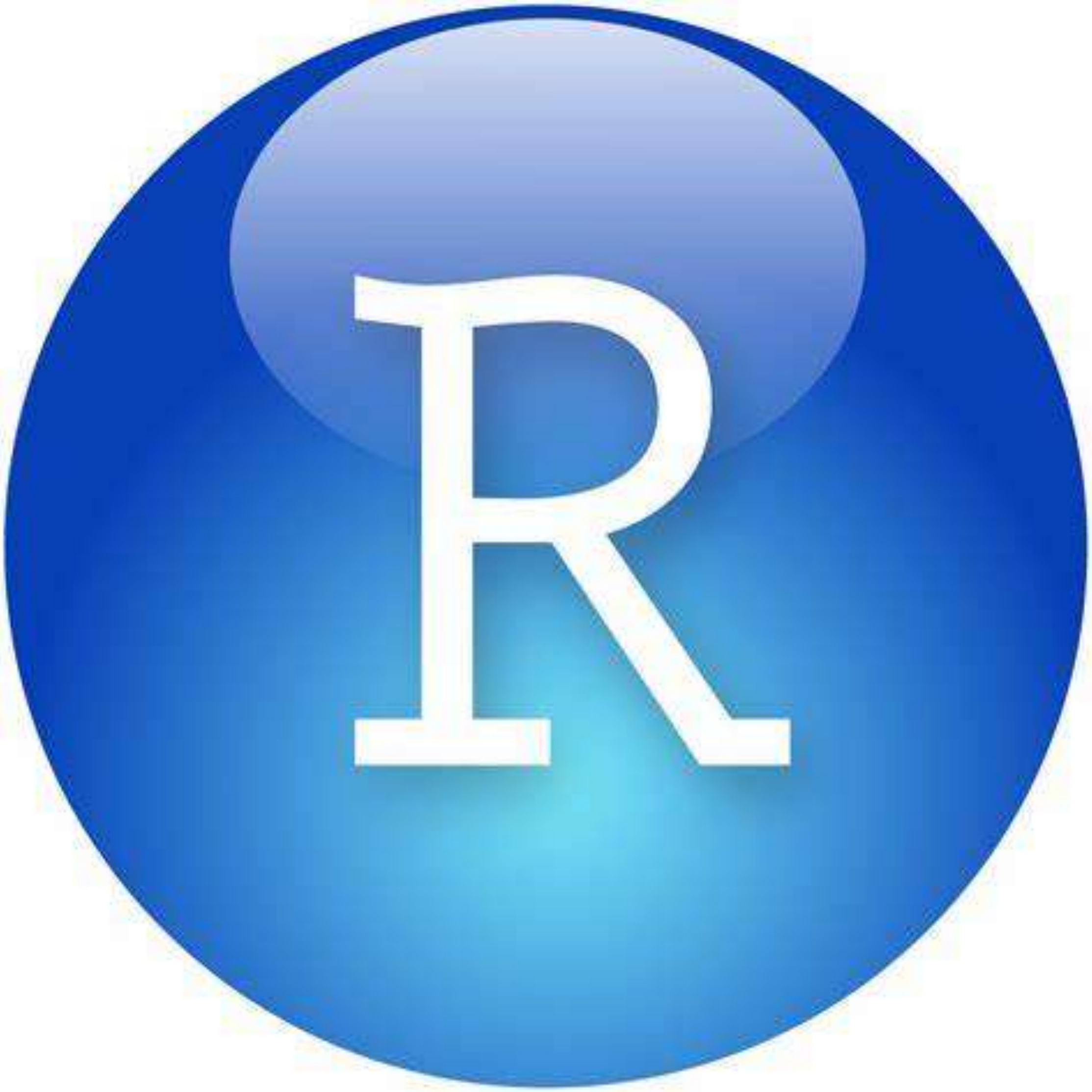
As it seems, the all the features are **highly positively correlated** except for the ***SepalWidth*** attribute with all the features is **negatively correlated** and almost there is **no correlation** value between most of them.

6. Boxplot for each feature



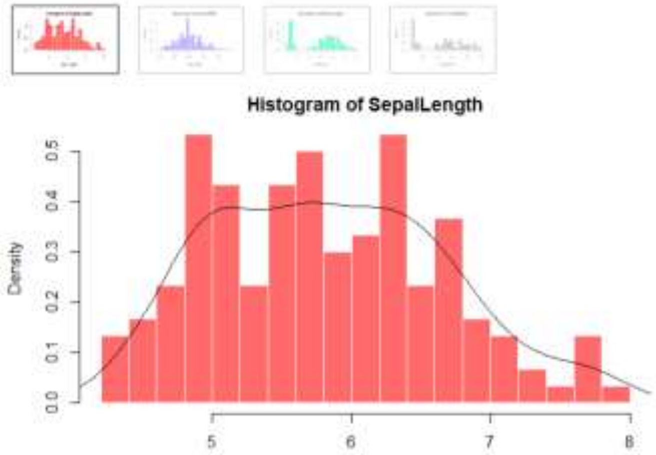
Conclusion

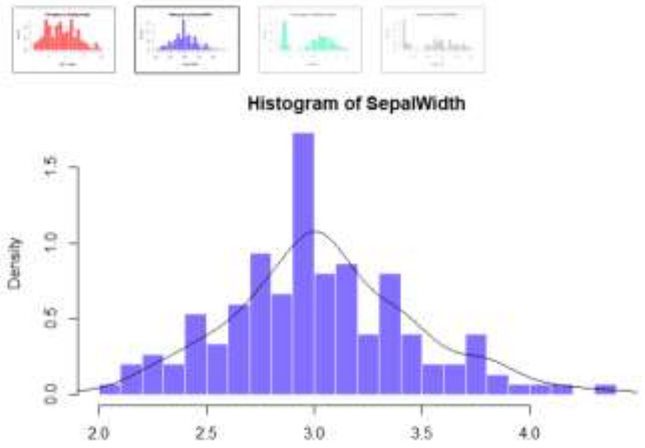
Seems as an overall that there are **no outliers** even one or two maybe, and the average lengths of **Sentosa's SepalWidth are tall** but for **other features are short**. For the **Versicolor** it has an **average height** for all features except the **SepalWidth feature it's the smallest**. and finally, the **tallest Specie is Virginica** for all features but, for the **SepalWidth it has an average height**.

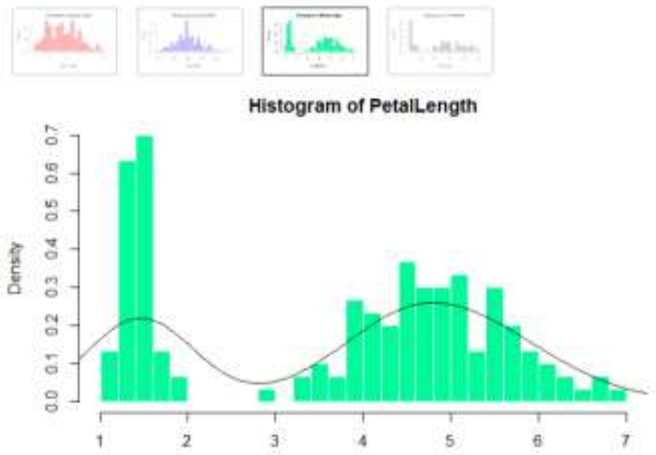


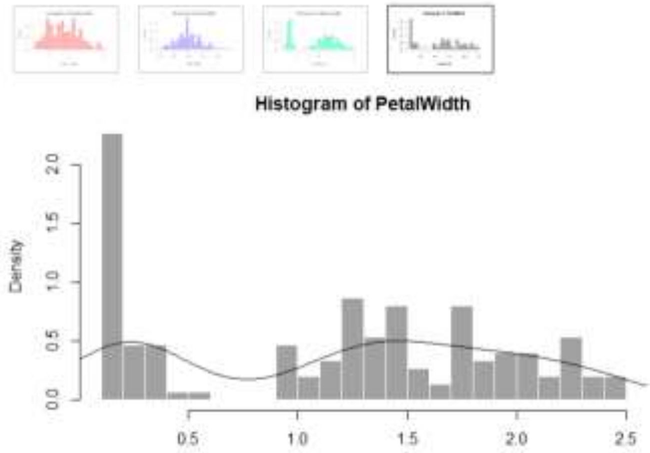
R Studio

1. Histogram for Target data







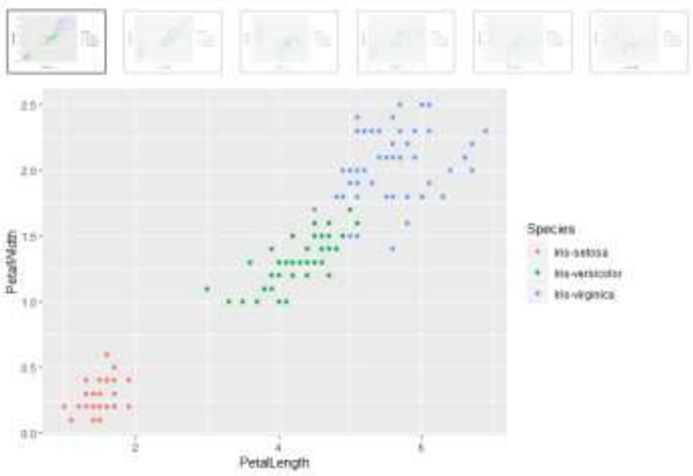


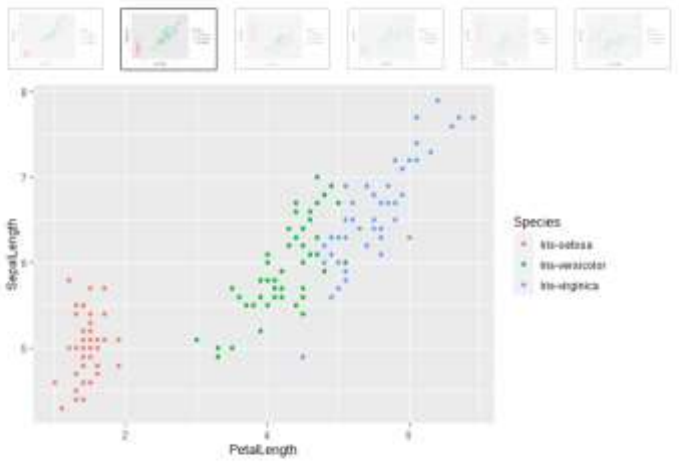
Conclusion

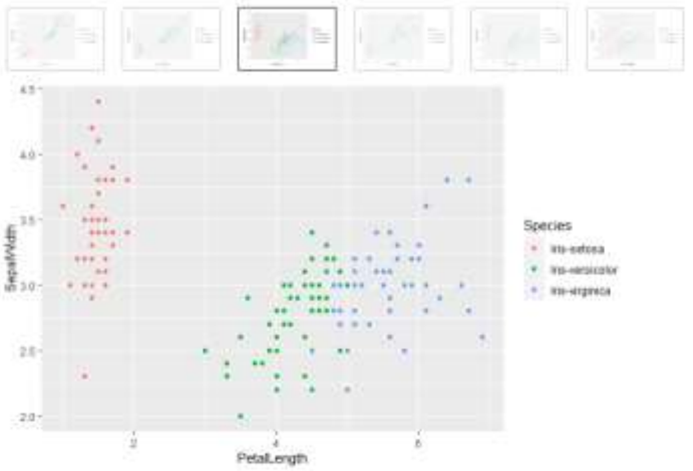
Sepal Width feature follows normal distribution.

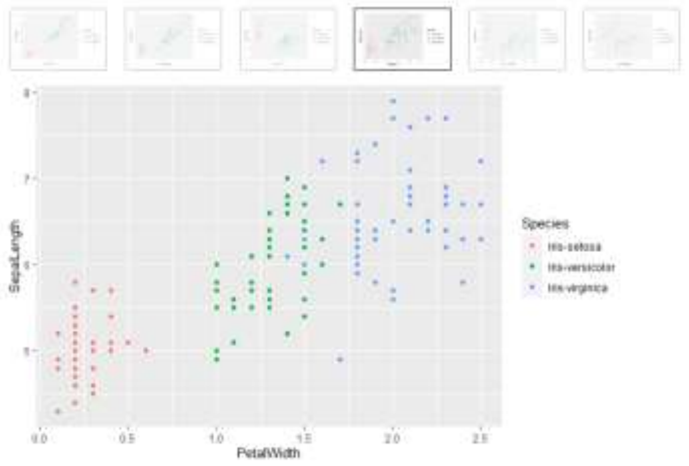
While **other** features tend to be right skewed more.

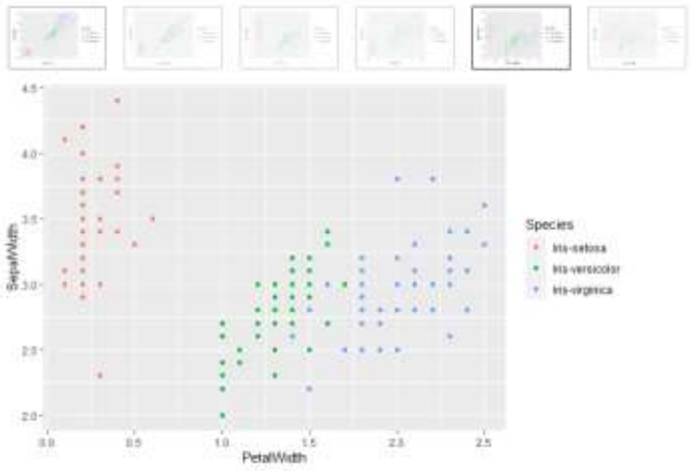
2. Scatter Plot for every 2 attributes

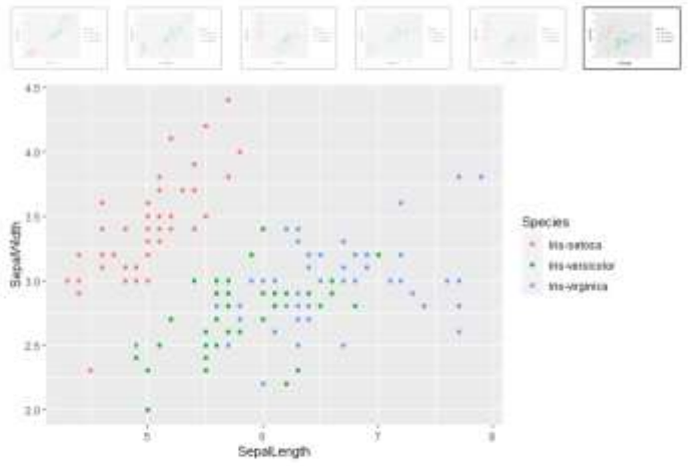










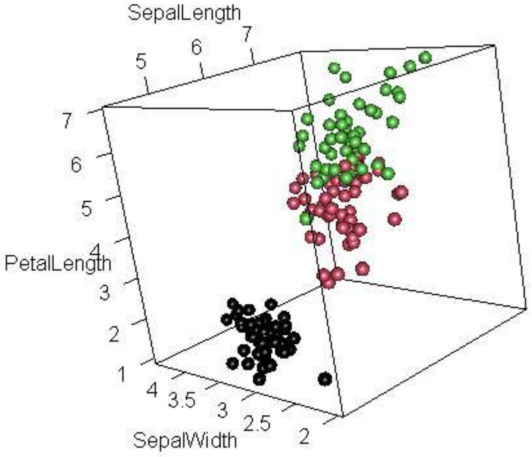


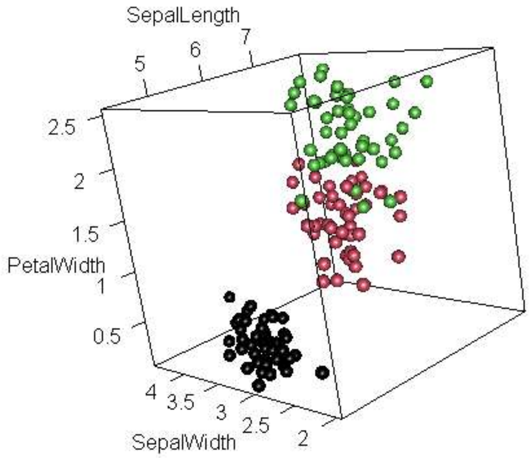
Conclusion

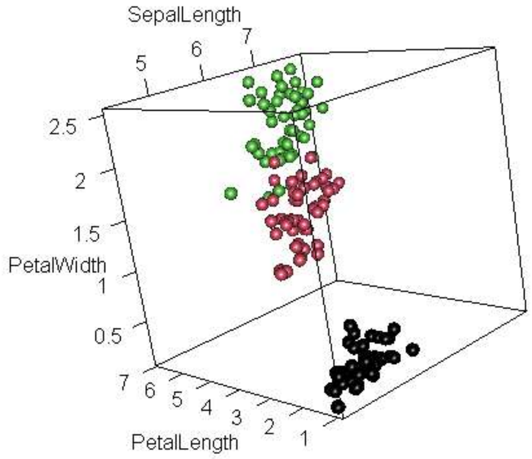
As it seems in the scatter plots data are already been classified and separated into clusters.

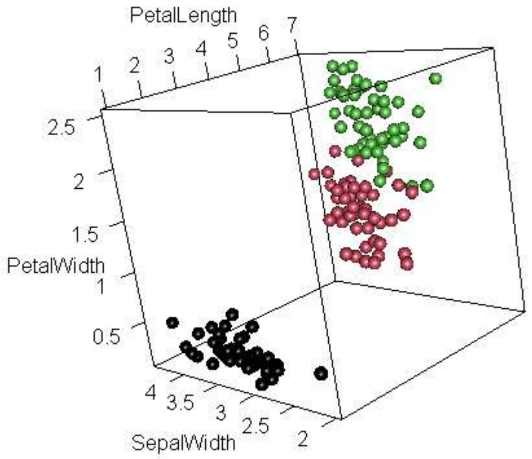
So as it seems the **Setosa Specie** is the *smallest* one then,
the **Versicolor** and after that the **Virginica**.

3. 3D Scatter Plot for every 3 attributes





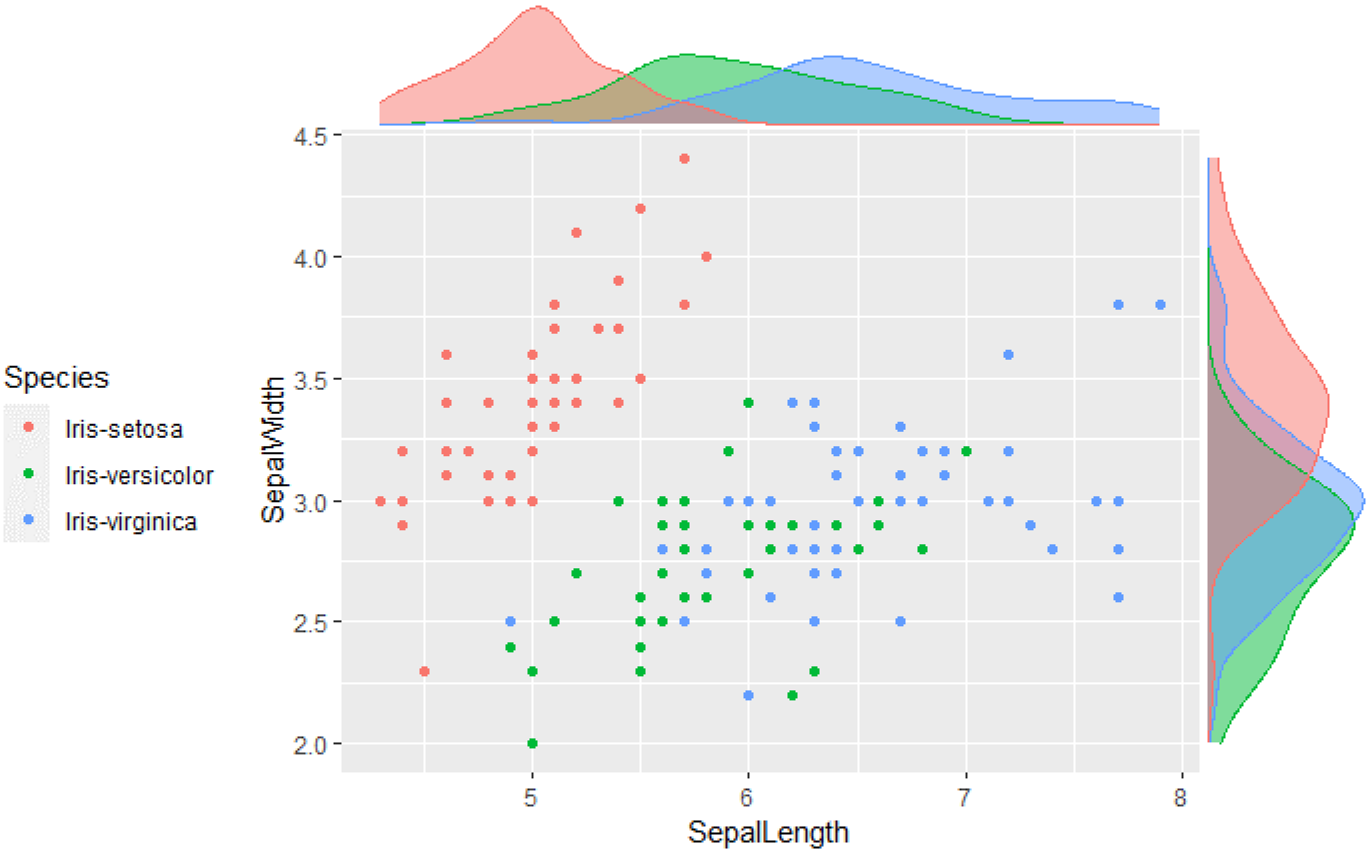


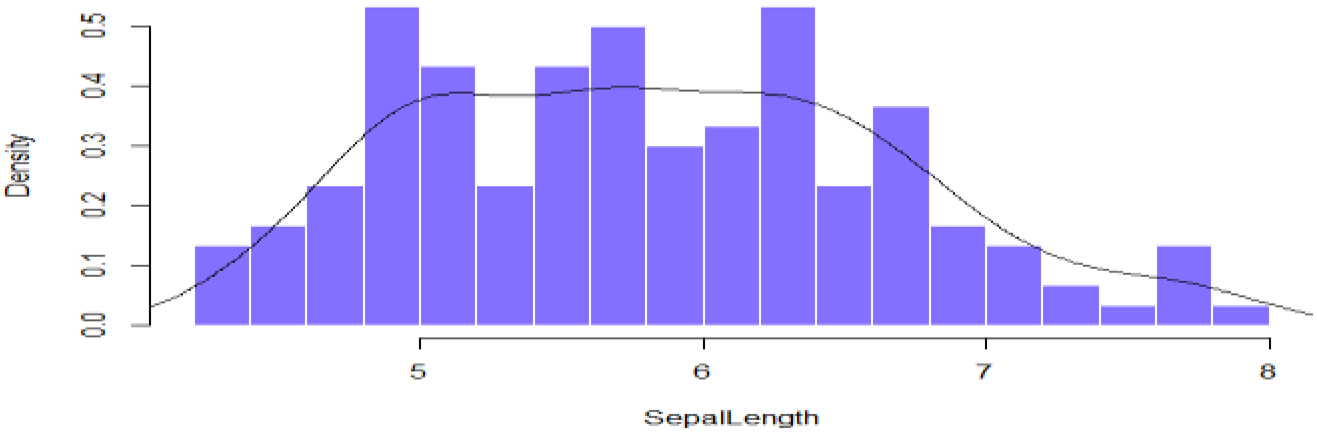


Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
It's not colored but, the clusters can also be recognized.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

4. Jointplot and KDE plot, Histogram

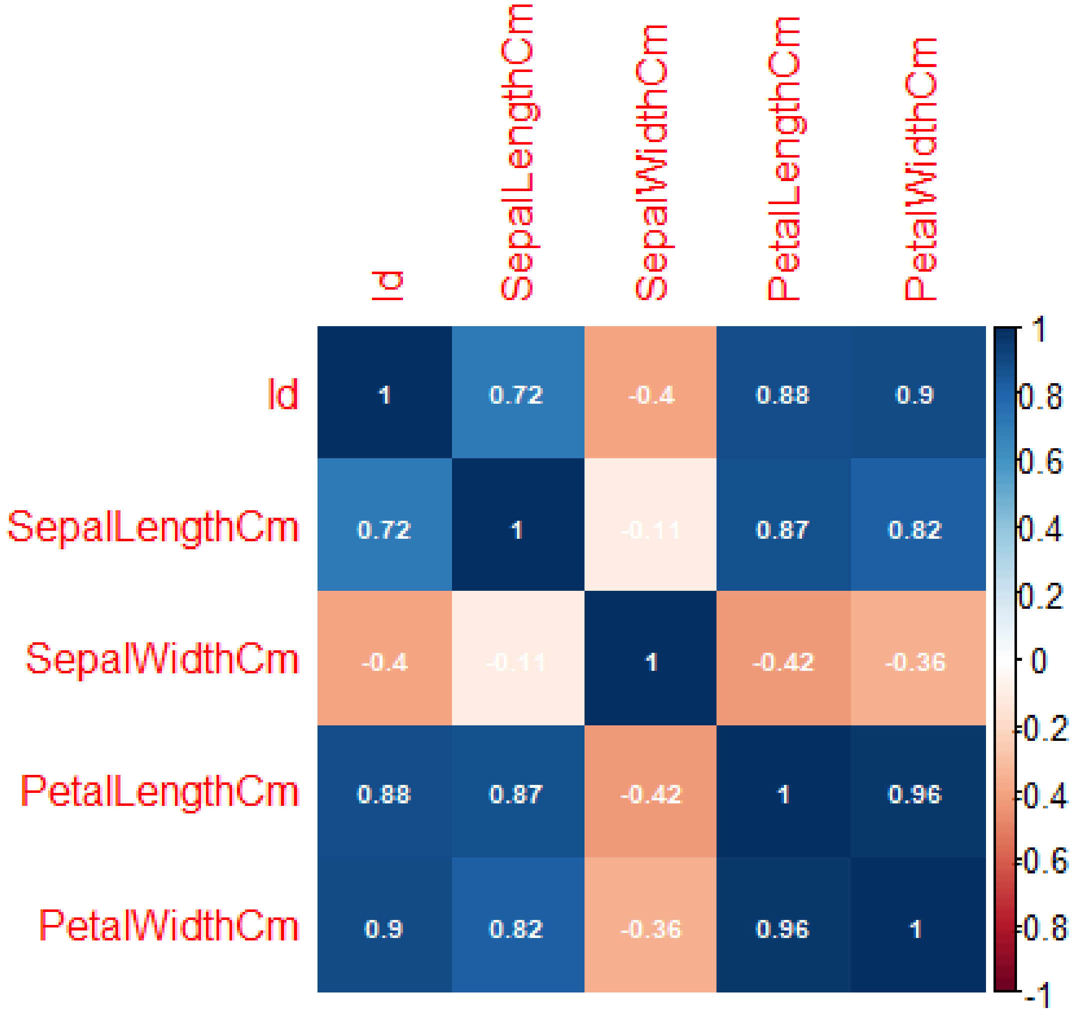




Conclusion

Well, it was kind of easy to plot such view using **unipanel** and **scatterhist** plot in R.

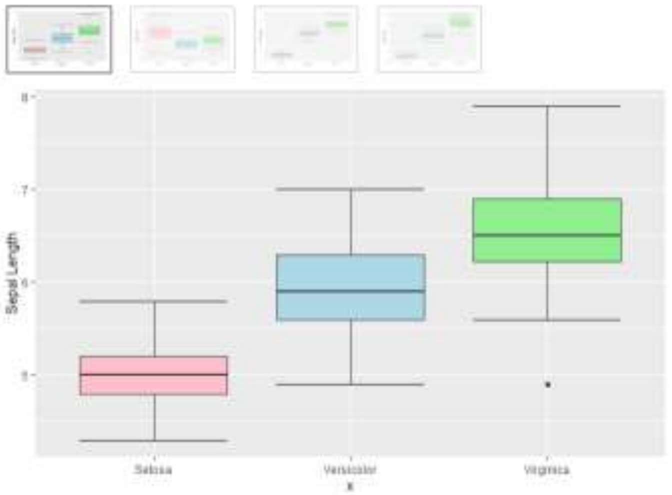
5. Correlation Matrix

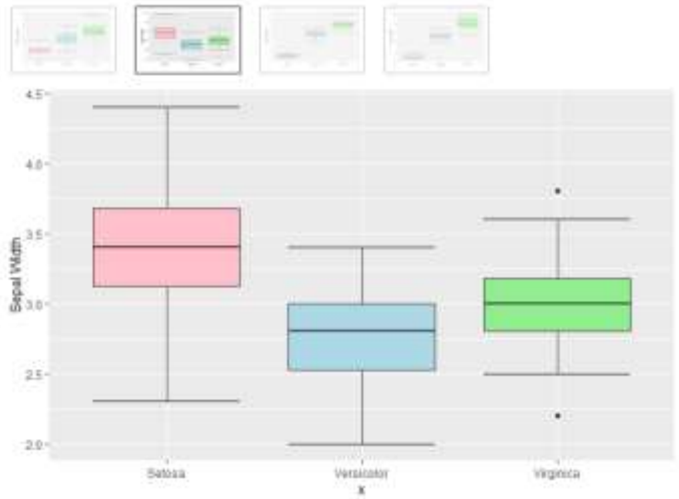


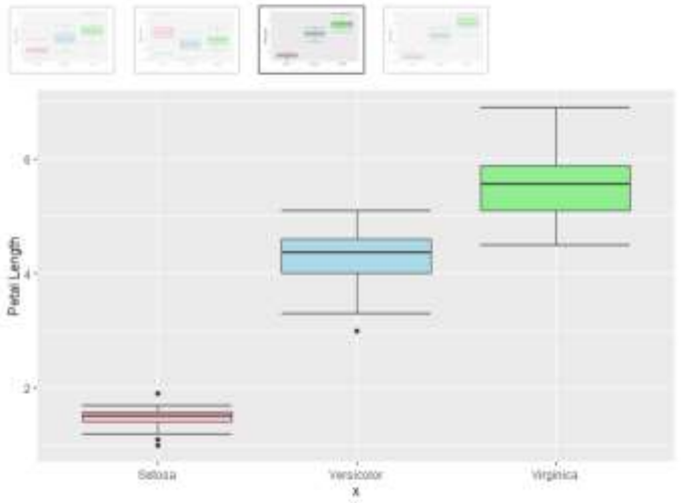
Conclusion

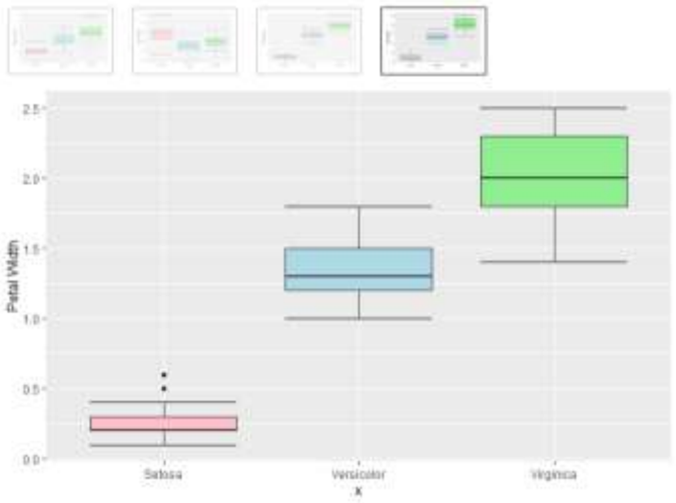
As it seems, the all the features are **highly positively correlated** except for the ***SepalWidth*** attribute with all the features is **negatively correlated** and almost there is **no correlation** value between most of them.

6. Boxplots for each feature







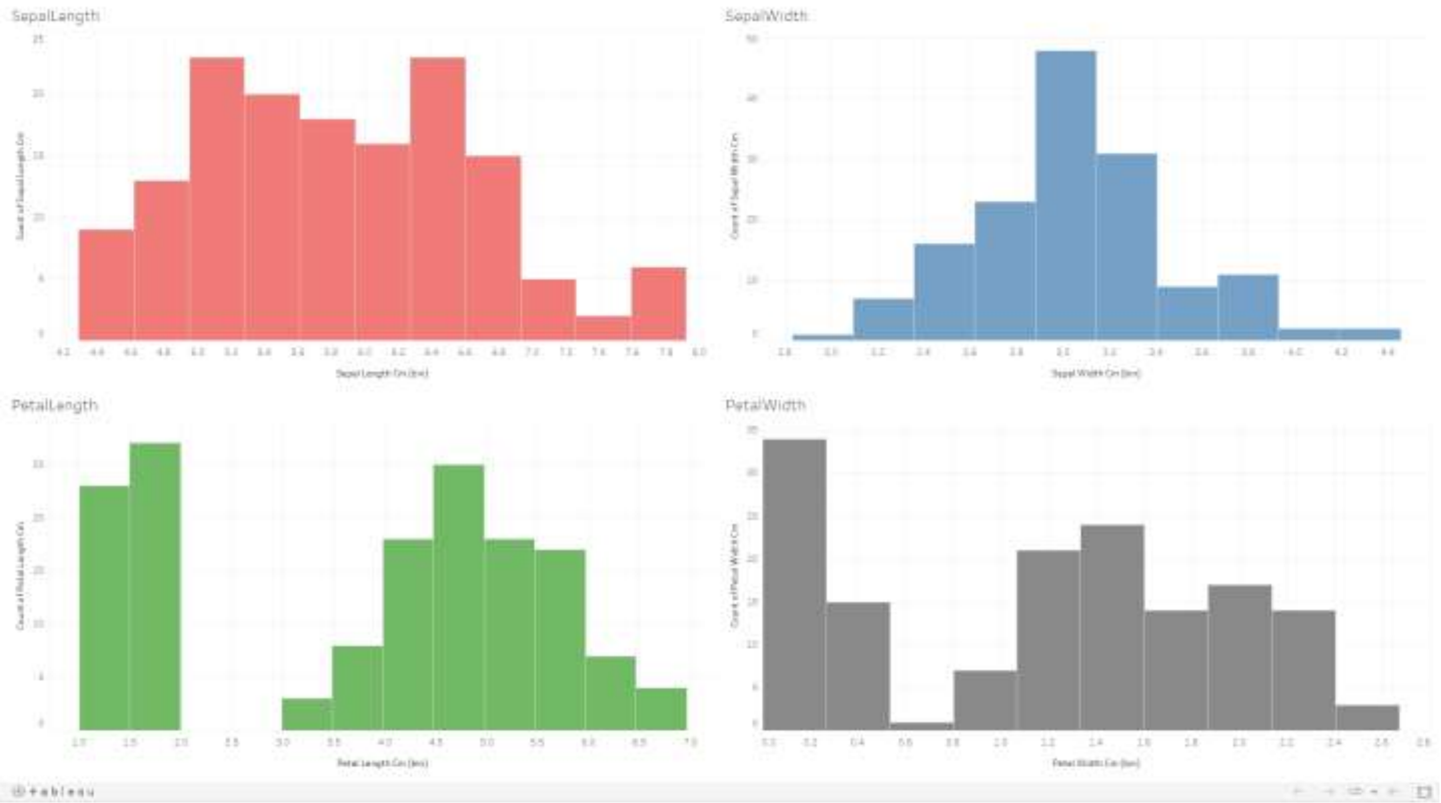


Conclusion

Seems as an overall that there are **no outliers** even one or two maybe, and the average lengths of **Sentosa's SepalWidth are tall** but for **other features are short**. For the **Versicolor** it has an **average height** for all features except the **SepalWidth feature it's the smallest**. and finally, the **tallest Specie is Virginica** for all features but, for the **SepalWidth** it has an **average height**.

tableau®
SOFTWARE

1. Histogram for Target data

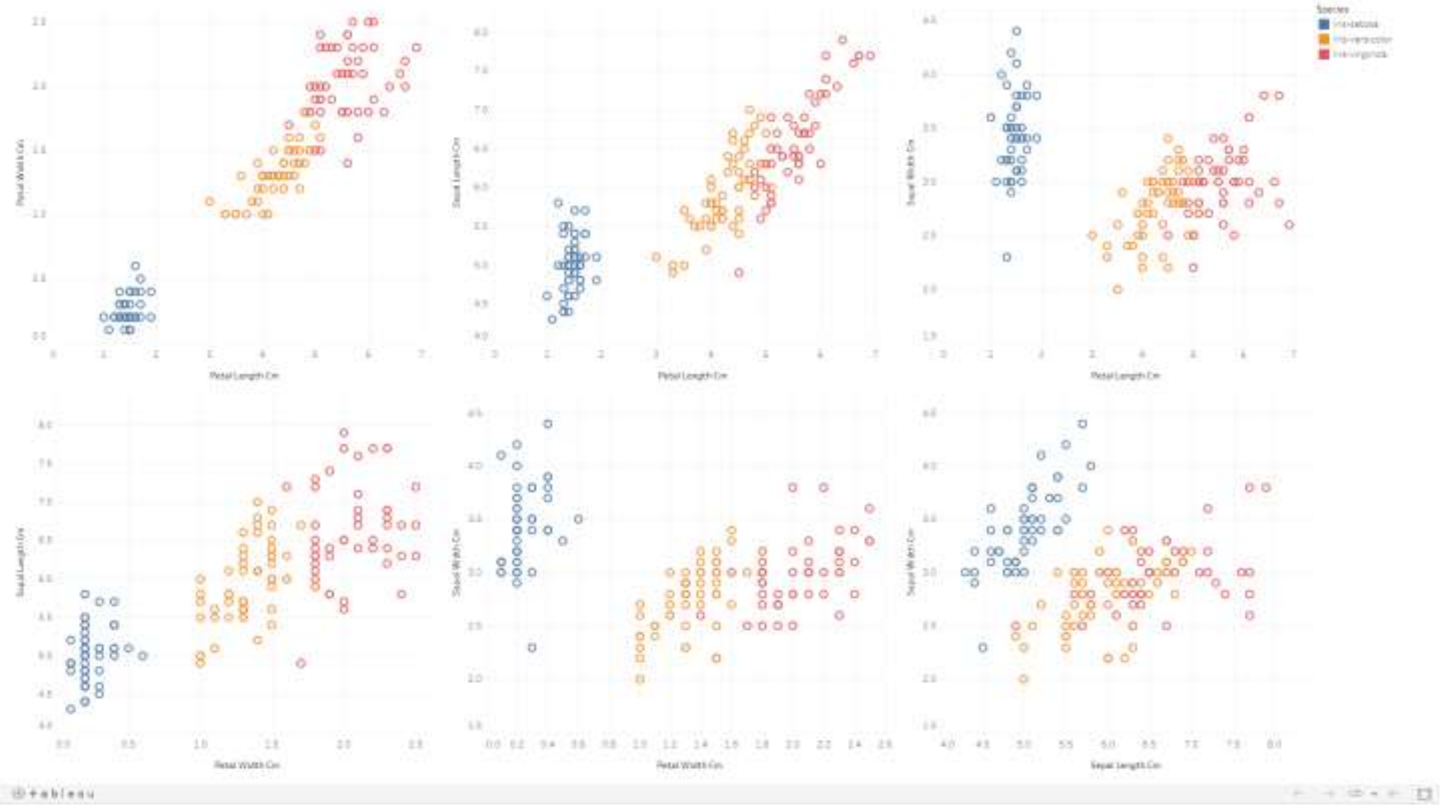


Conclusion

Sepal Width feature follows normal distribution.

While **other** features tend to be right skewed more.

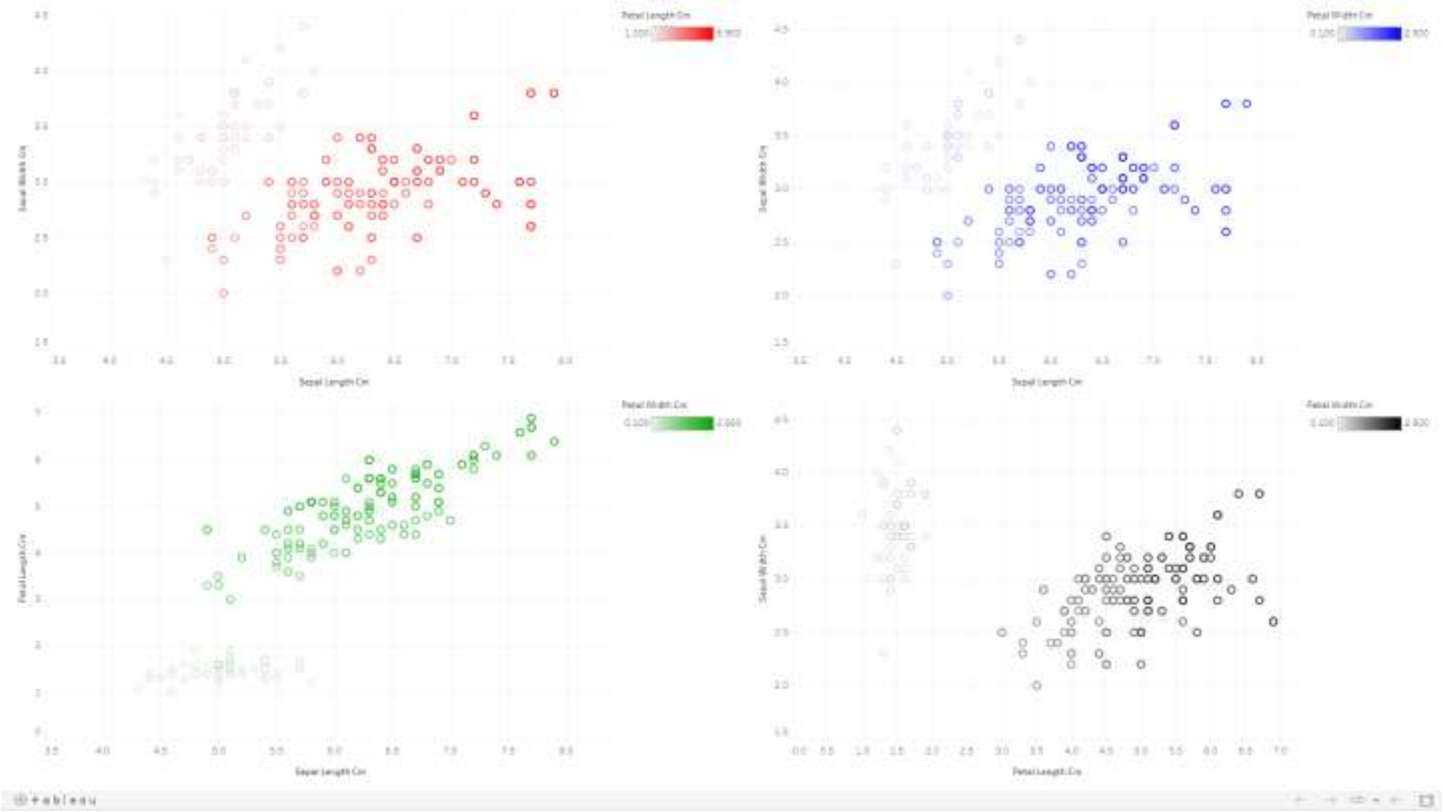
2. Scatter Plot for every 2 attributes



Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

3. 3D Scatter Plot for every 3 attributes

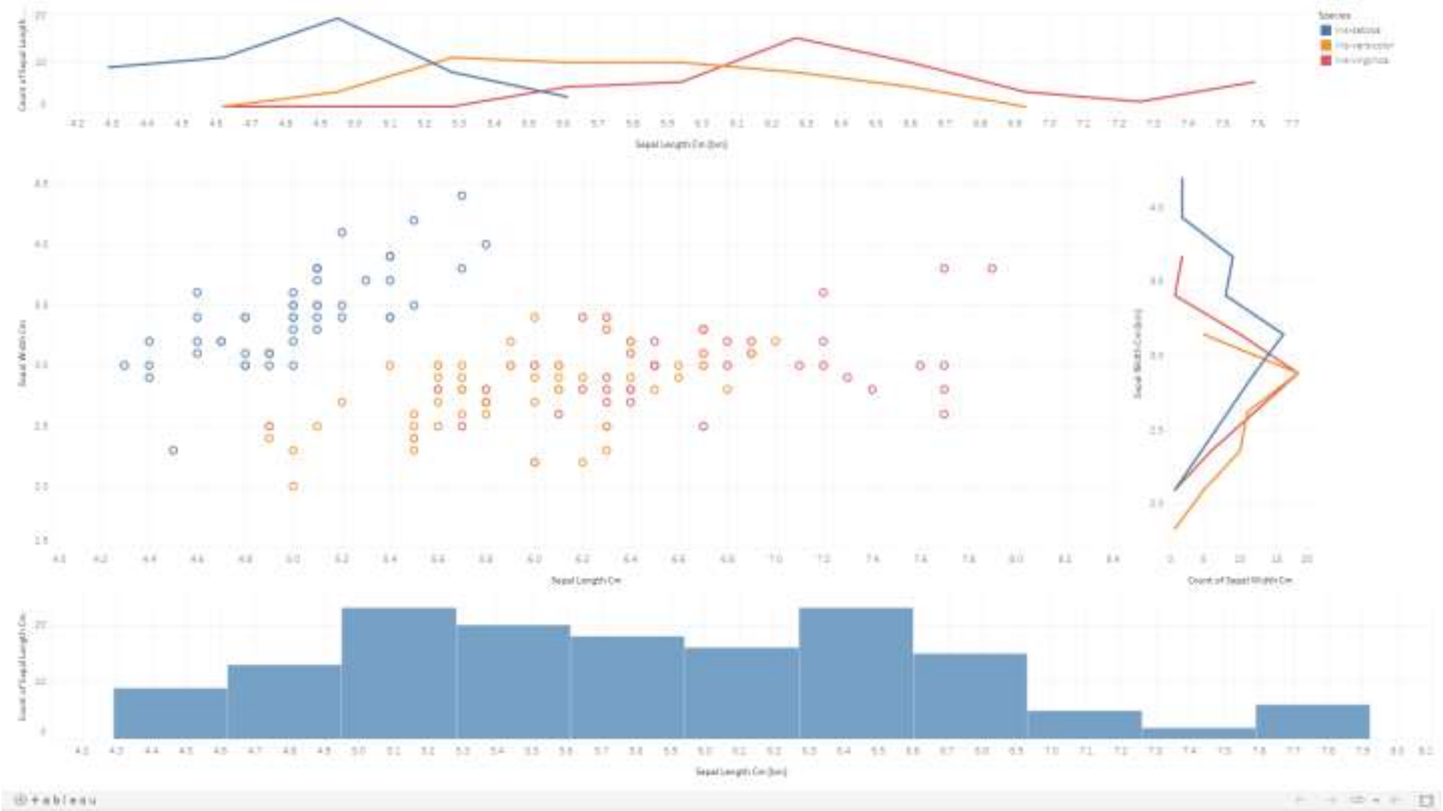


Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
It's not colored but, the clusters can also be recognized.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

Also, there is no 3D Scatter plot in Tableau so we used the color as a third dimension for the data.

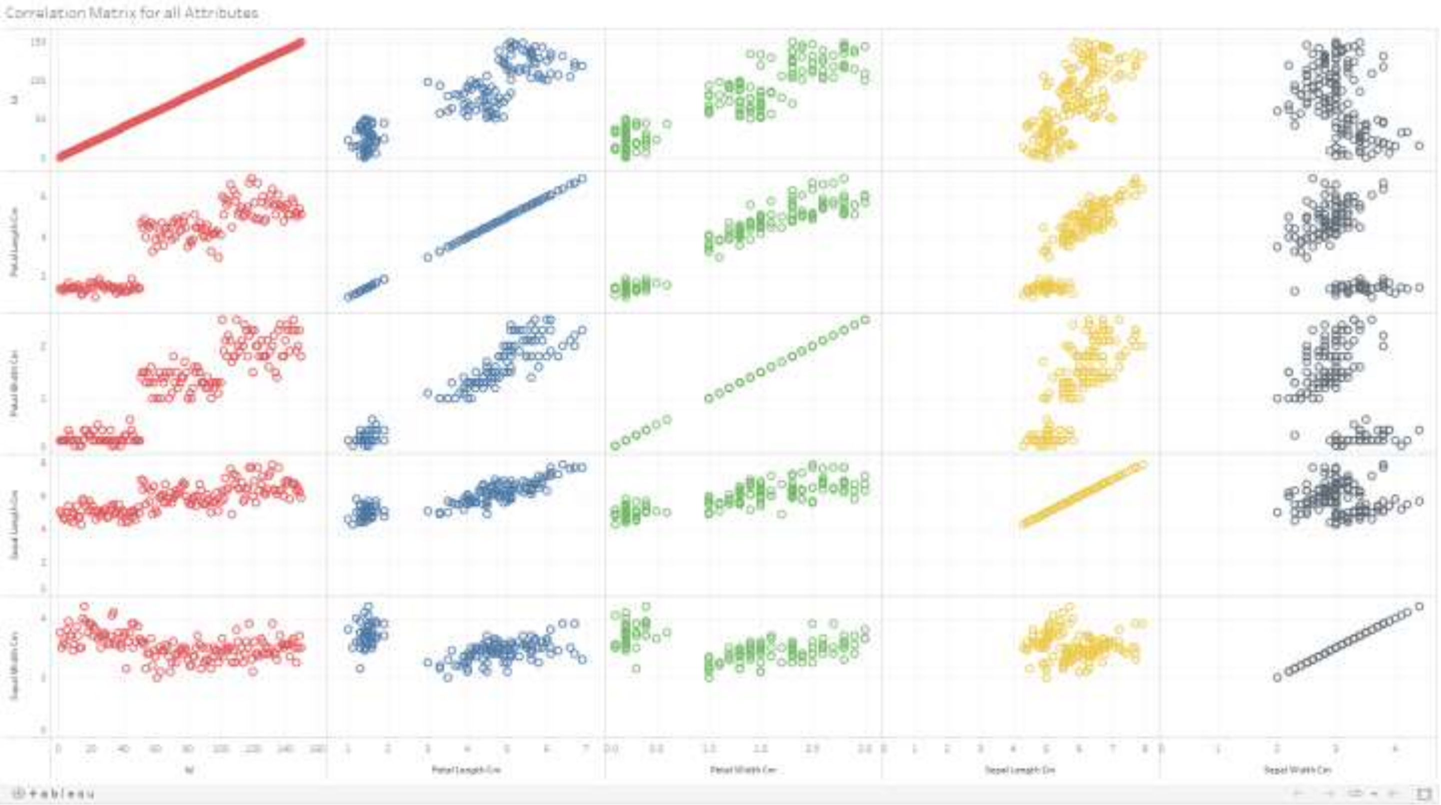
4. Jointplot and KDE plot, Histogram



Conclusion

Well, it was kind of easy to plot such view using **2-line plots** and **Scatter plot** and **Histogram** in Tableau.

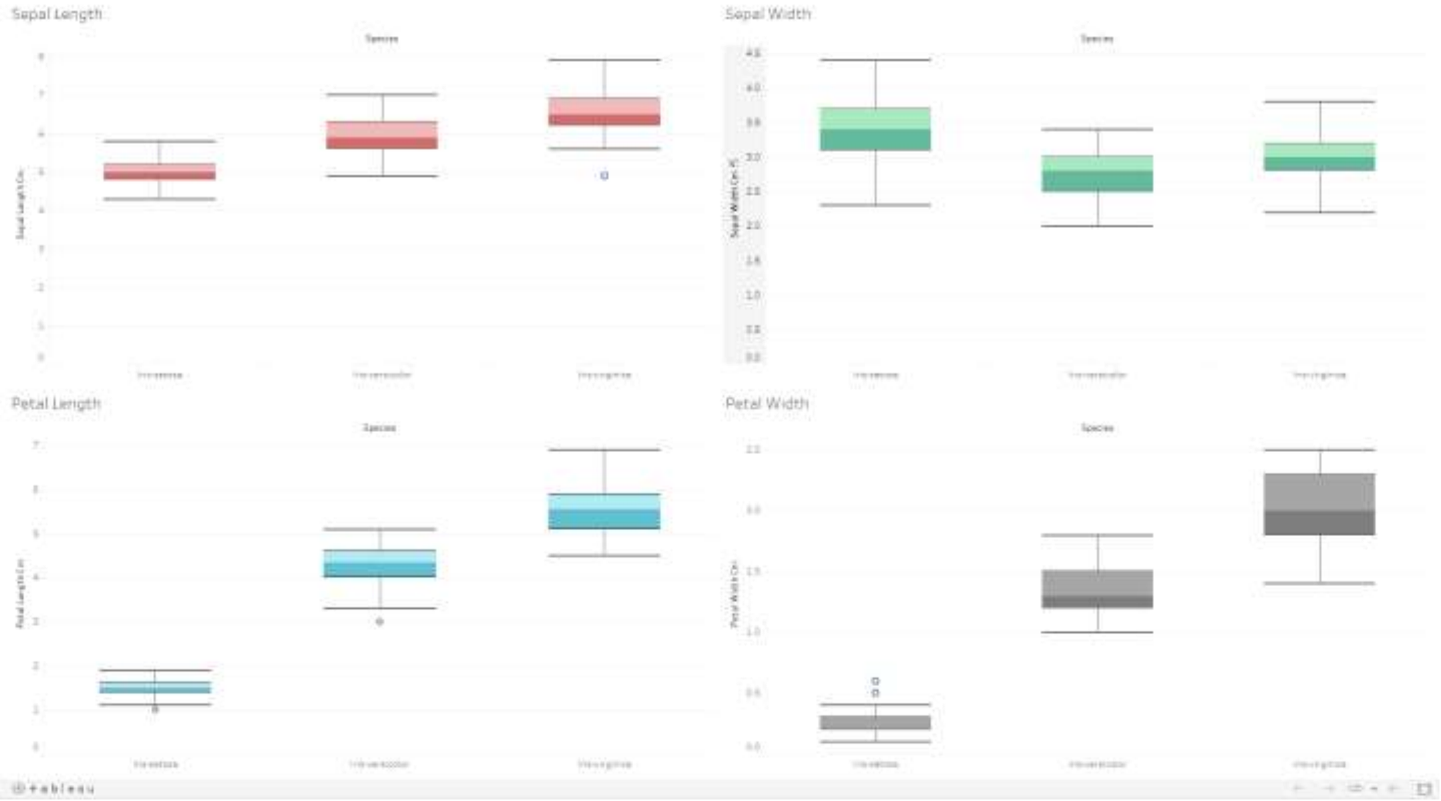
5. Correlation Matrix



Conclusion

As it seems, the all the features are **highly positively correlated** except for the ***SepalWidth*** attribute with all the features is **negatively correlated** and almost there is **no correlation** value between most of them.

6. Boxplots for each feature

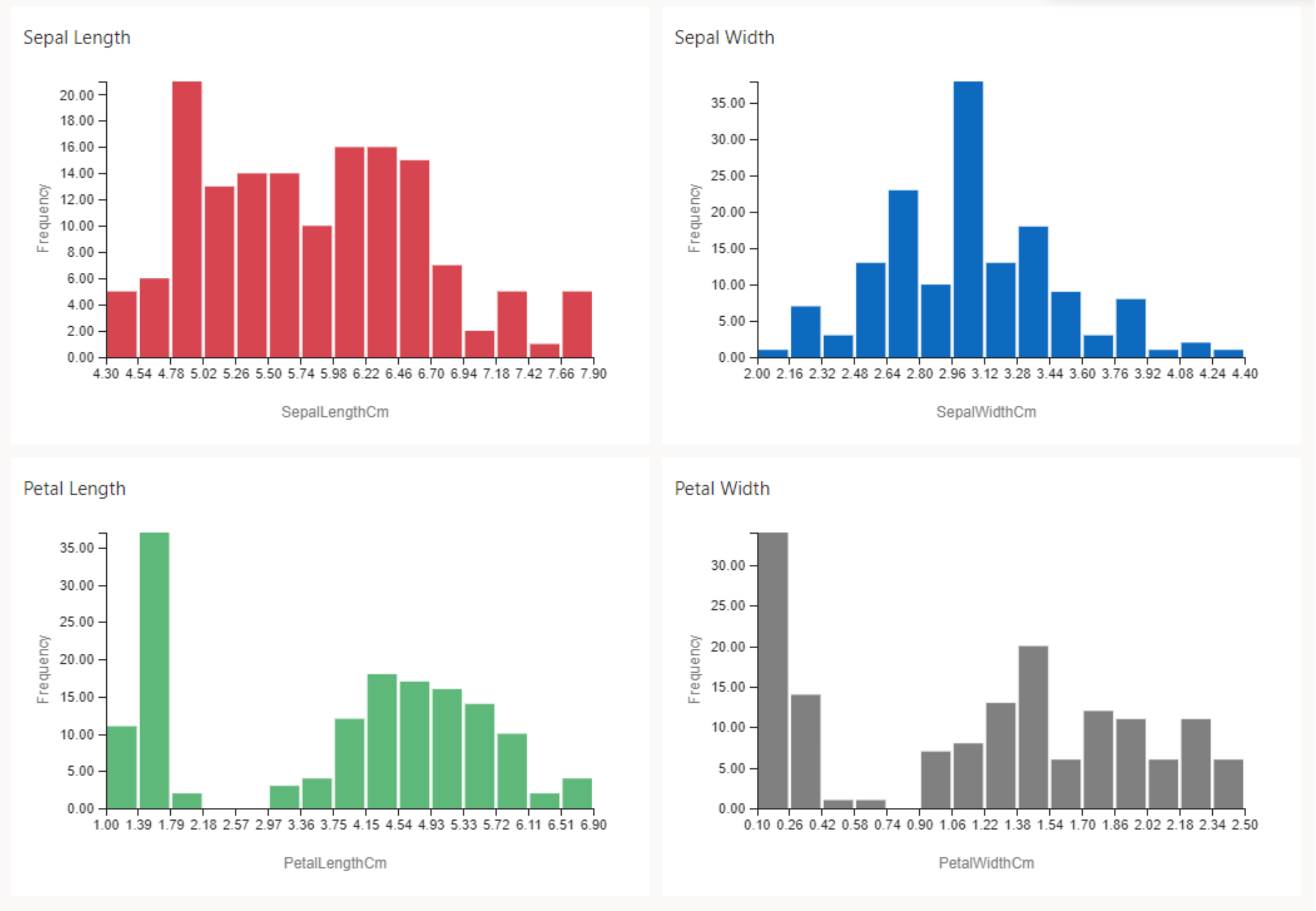


Conclusion

Seems as an overall that there are **no outliers** even one or two maybe, and the average lengths of **Sentosa's SepalWidth are tall** but for **other features are short**. For the **Versicolor** it has an **average height** for all features except the **SepalWidth feature it's the smallest**. and finally, the **tallest Specie is Virginica** for all features but, for the **SepalWidth** it has an **average height**.

Power BI

1. Histogram for Target data

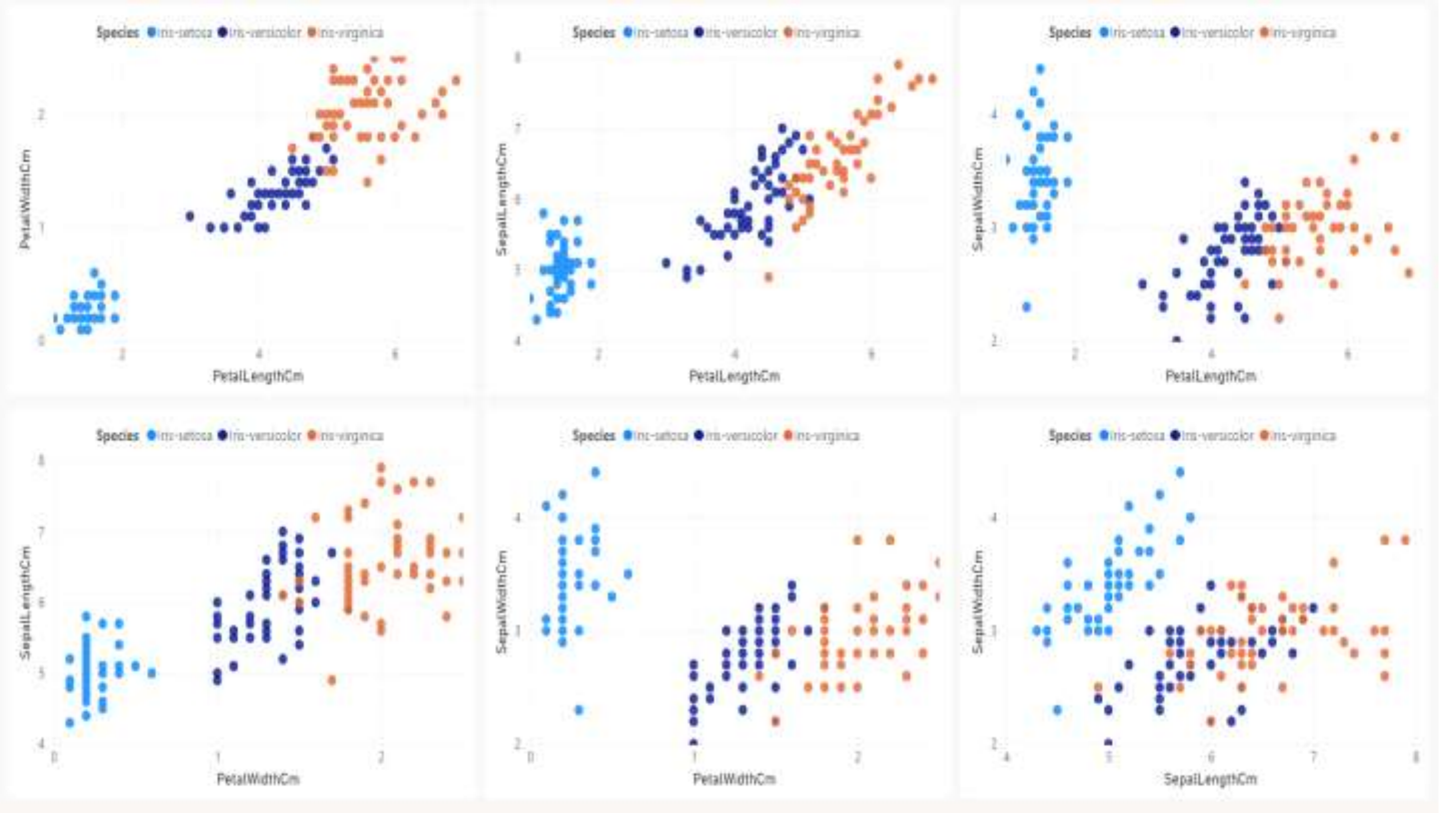


Conclusion

Sepal Width feature follows normal distribution.

While **other** features tend to be right skewed more.

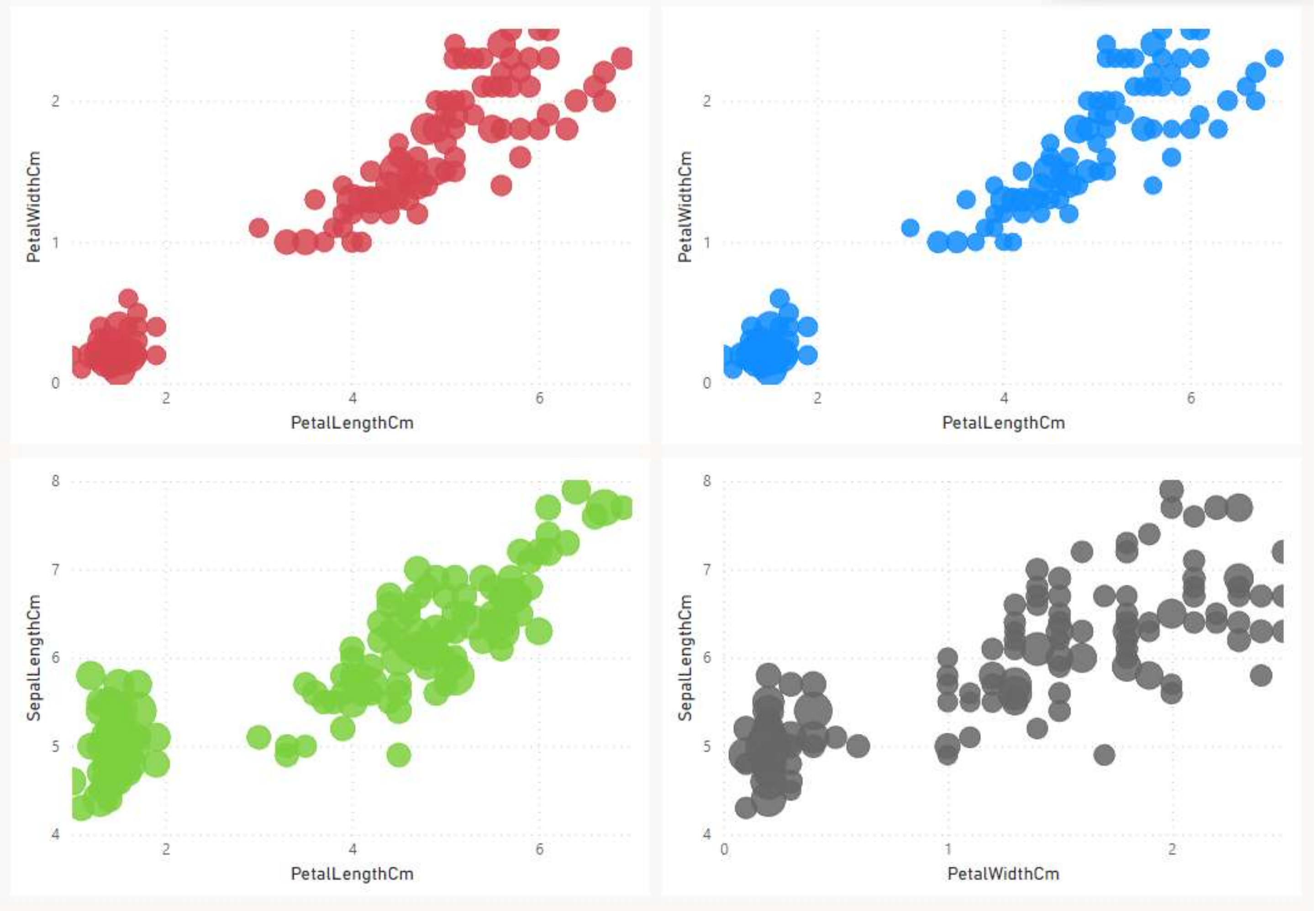
2. Scatter Plot for every 2 attributes



Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

3. 3D Scatter Plot for every 3 attributes

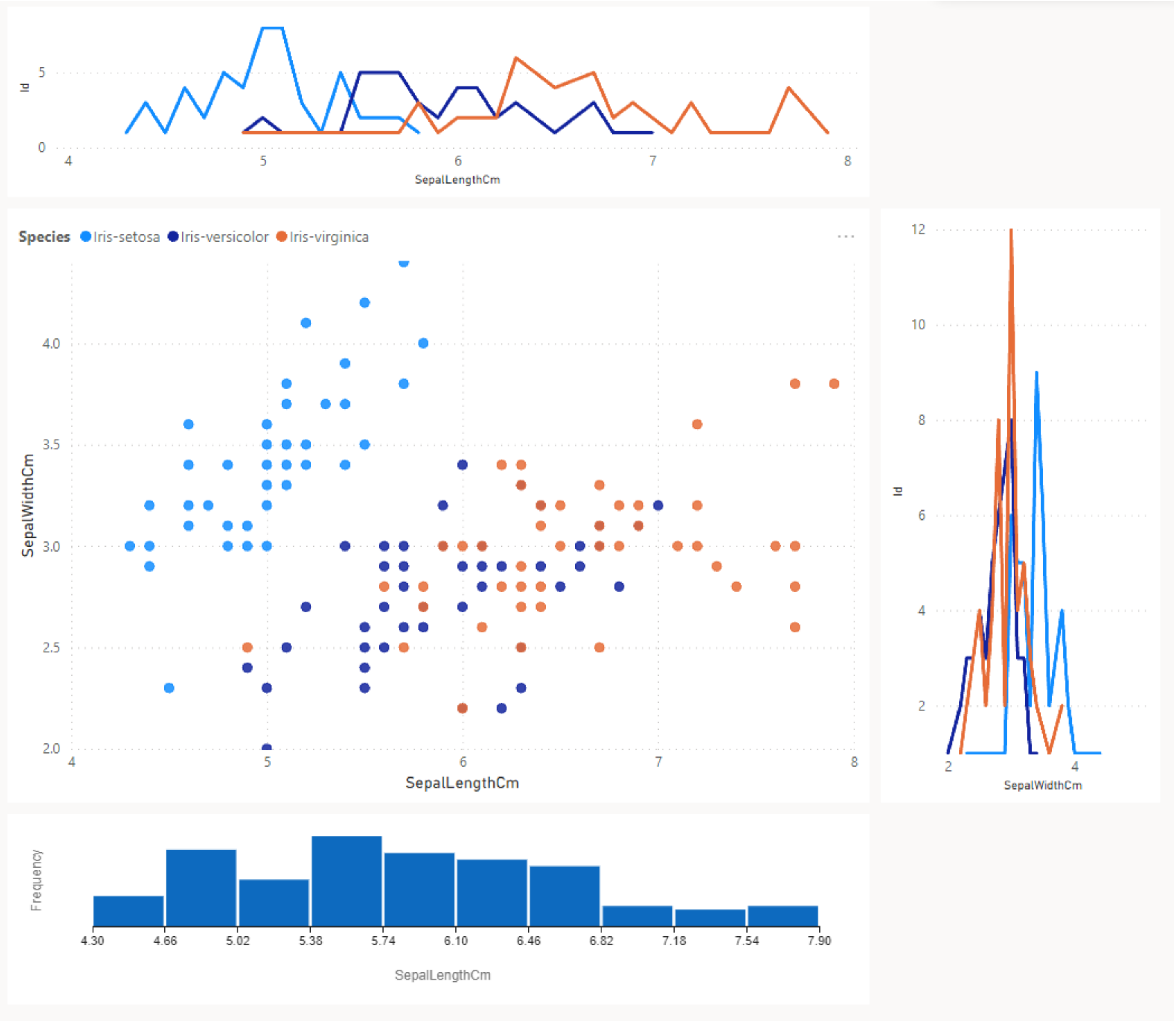


Conclusion

As it seems in the scatter plots data are already been classified and separated into clusters.
It's not colored but, the clusters can also be recognized.
So as it seems the **Setosa Specie** is the *smallest* one then, the **Versicolor** and after that the **Virginica**.

Also, there is no 3D Scatter plot in Power BI so we used the size as a third dimension for the data.

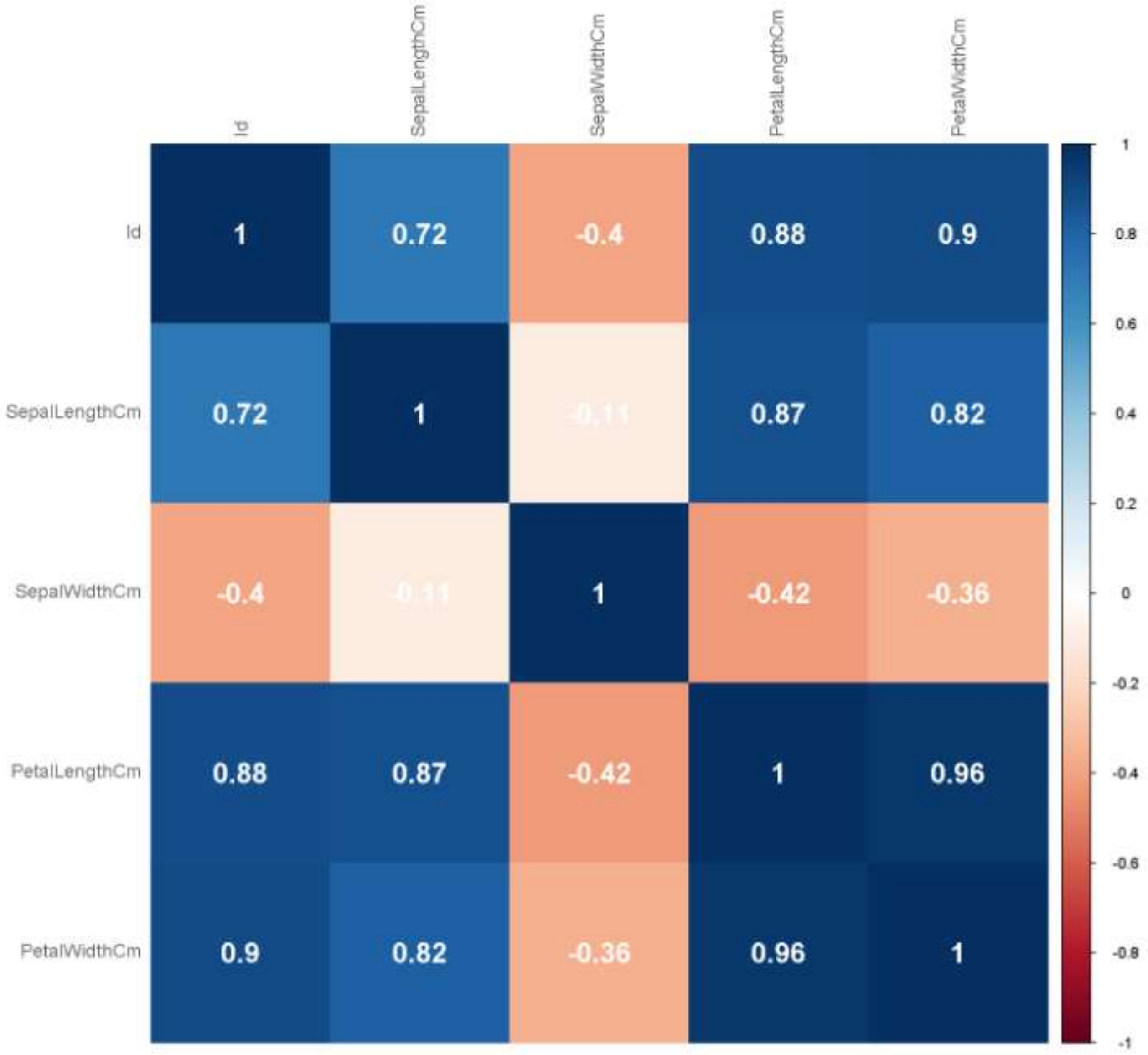
4. Jointplot and KDE plot, Histogram



Conclusion

Well, it was kind of easy to plot such view using **2-line plots** and **Scatter plot** and **Histogram** in Power BI.

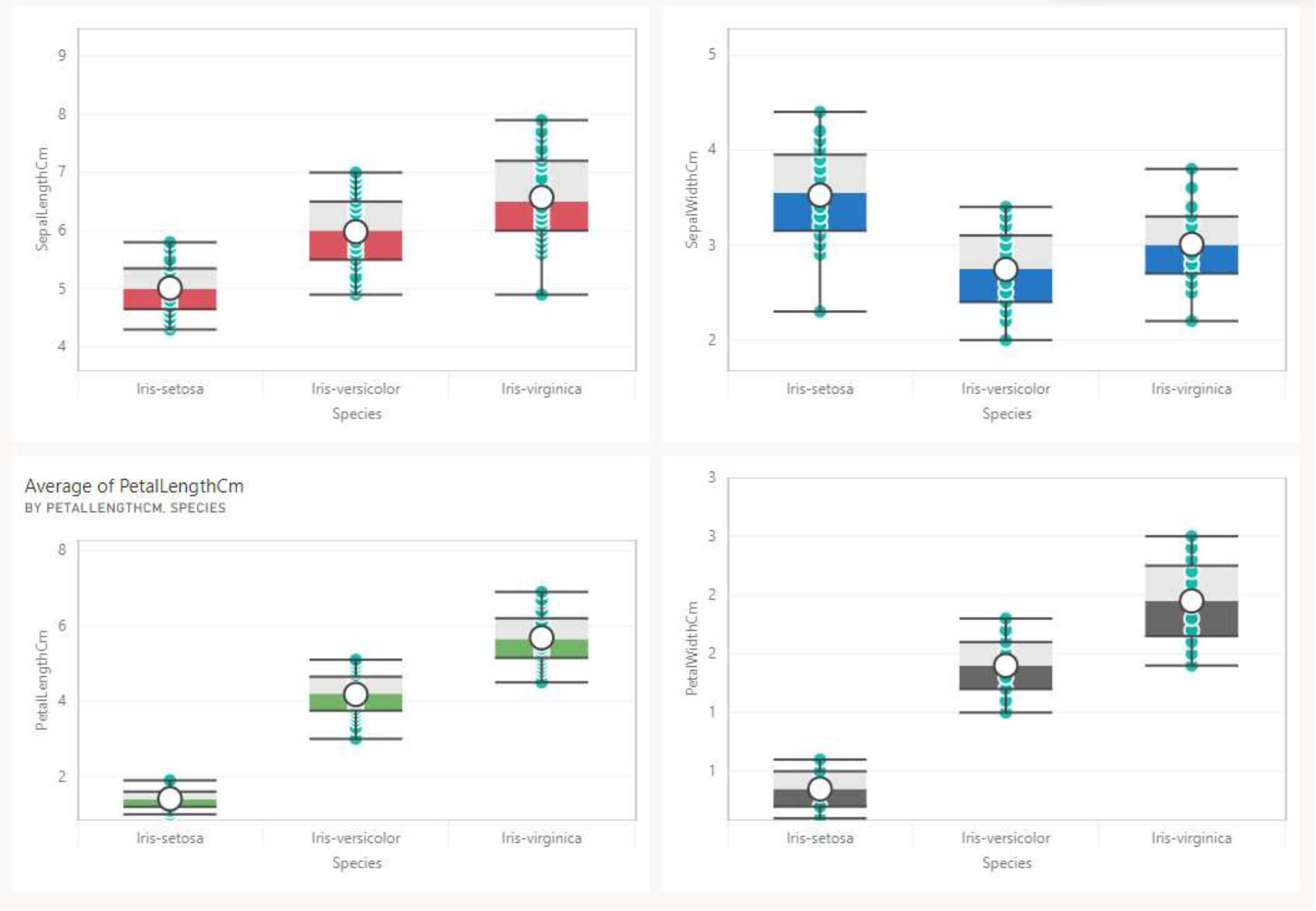
5. Correlation Matrix



Conclusion

As it seems, the all the features are **highly positively correlated** except for the ***SepalWidth*** attribute with all the features is **negatively correlated** and almost there is **no correlation** value between most of them.

6. Boxplots for each feature



Conclusion

Seems as an overall that there are **no outliers** even one or two maybe, and the average lengths of **Sentosa's SepalWidth are tall** but for **other features are short**. For the **Versicolor** it has an **average height** for all features except the **SepalWidth feature it's the smallest**. and finally, the **tallest Specie is Virginica** for all features but, for the **SepalWidth** it has an **average height**.

Comparison between all of the tools

From the projects point of view, first for python, since python is an easy programming language it facilitated the work pretty much also it's very efficient.

Unlike MATLAB cause it's not the easiest in usage specially for the live code part also it's not that super-efficient tool and it doesn't include everything.

But, for the R programming language it's a bit of both cause yeah, it's a bit annoying language to be used but also, it's efficient and contains a lot of Visualizations and different types of plots to be used describing your data.

Finally for the 2 most popular data visualization tools Tableau and Power BI

Tableau and Power Bi are very effective tools, easy to be used and with great outputs to be communicated on but for example, both of them doesn't contain 3D Scatter plots as these plots aren't the most efficient way to show a 3 dimensioned numerical data types.

Also, for example in Power BI you can't easily put color as a third dimensioned attribute unlike Tableau.

Same for tableau where it's not easy to plot a donut chart for example but in Power BI it's very easy to do though.

So as a final answer, **what tool did we prefer the most?**

It's **Python,** cause regardless all of that but working on a programming language from scratch not only gives you more fun but also, it provides the best experience as you can do whatever you want using a programming language as long as you are super familiar with it unlike the already available applications … well, most of them at least.

Thank ❤

You ❤