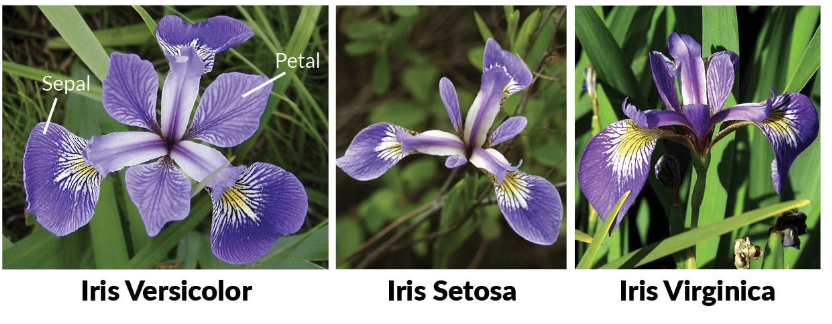
**K-Means Clustering of Iris Dataset**

The Iris flower data set, initially introduced by the British statistician and biologist Ronald Fisher in 1936, serves as a multivariate dataset exemplifying linear discriminant analysis. It is alternatively referred to as Fisher's Iris data set or Anderson's Iris data set, named after its collector Edgar Anderson. This dataset's historical context and its subsequent controversy in the realm of statistical education have led to discussions about its ongoing relevance.

Comprising 50 samples from each of three Iris species (Iris setosa, Iris virginica, and Iris versicolor), the dataset encompasses measurements of four attributes from each sample: sepal and petal length and width, all recorded in centimeters. Fisher harnessed the amalgamation of these attributes to develop a linear discriminant model, enabling the differentiation of the distinct Iris species from one another.



This study is for clustering Iris Dataset used Kmeans

[Features Information:](https://archive.ics.uci.edu/ml/datasets/iris)

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: -- Iris Setosa -- Iris Versicolour -- Iris Virginica

# ****Data Visualization****



(all the Variety Label) contain an equal amount of unique values of Setosa/Versicolor and Virginica

# 

# Comparing Sepal Length and Sepal Width

# 

# 

The plot shown above indicates that

* Setosa has smaller sepal lengths but larger sepal widths.
* Versicolor lies in the middle of the other two species in terms of sepal length and width
* Virginica has larger sepal lengths but smaller sepal widths.

# Petal Length and Petal Width Distribution

# 

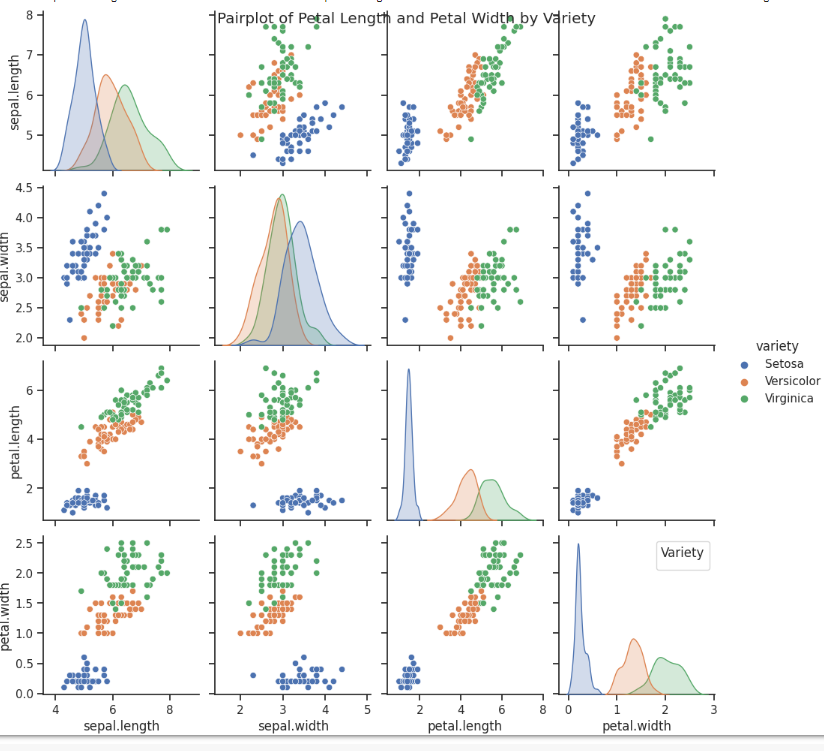
# 

The plot shown above indicates that

* Setosa has smaller petal lengths and widths.
* Versicolor lies in the middle of the other two species in terms of petal length and width
* Virginica has the largest of petal lengths and widths.

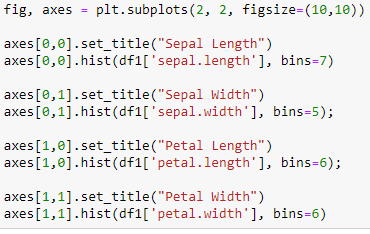
**Features Correlation**

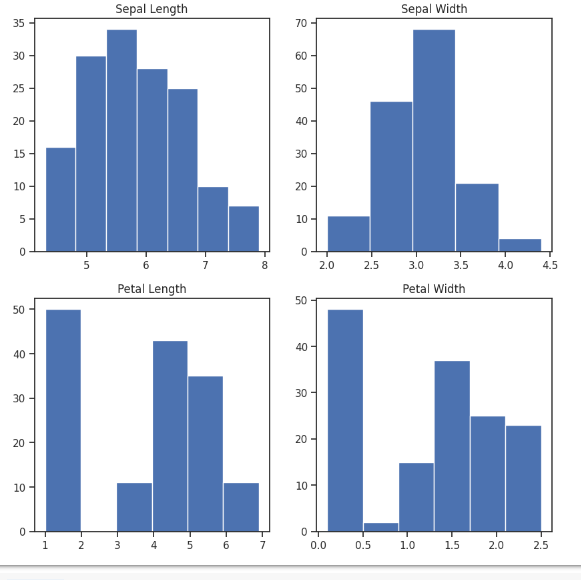




The plot shows that Setosa has the smallest of petals widths and lengths. It also has the smallest sepal length but larger sepal widths

**Histograms**

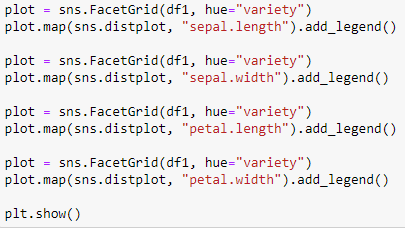


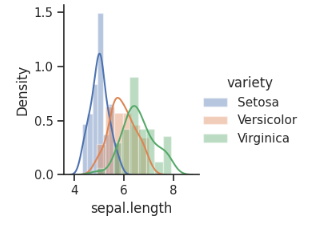
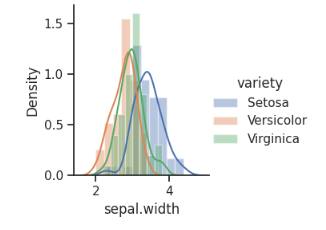


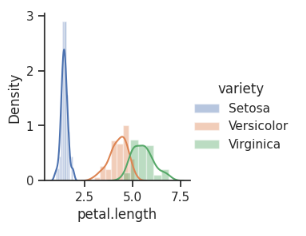
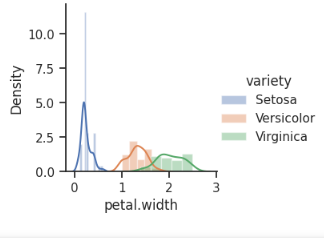
according to the plot above., we can see that –

* The highest frequency of the sepal length is between 30 and 35 which is between 5.5 and 6 cm
* The highest frequency of the sepal Width is around 68 which is between 3.0 and 3.5 cm
* The highest frequency of the petal length is around 49 which is between 1 and 2
* The highest frequency of the petal width is between 40 and 50 which is between 0.0 and 0.5

**Histograms with Distplot Plot**



From the Graph, we can observe:

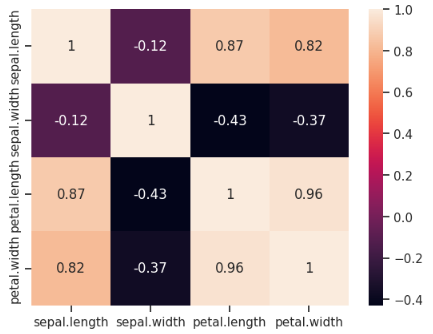
* In the case of Sepal Length, there is a significant overlapping in Sepal length
* In the case of Sepal Width also, there is a sigmificant amount of overlapping.
* In the case of Petal Length, there is a hardly of overlapping.
* In the case of Petal Width also, there is a hardly of overlapping.

Petal Length and Petal Width are good features to be used in Clustering

### Heatmaps



Heatmaps shows a correlation between all numerical variables in the dataset



From the above graph, we can see that –

* Petal width and petal length have high correlations.
* Petal length and sepal Length have good correlations.
* Petal Width and Sepal length have good correlations

# ****BOX Plot****

# 

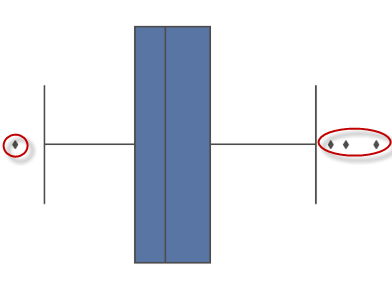
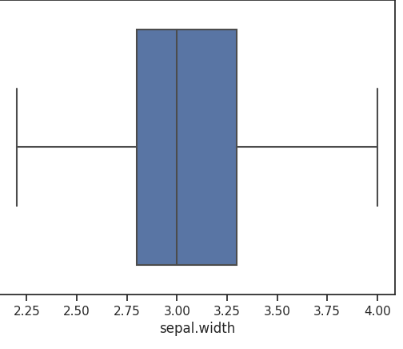
# 

The "Setosa" species exhibits the smallest features with relatively low distribution and some outliers, while the "Versicolor" species displays average features, and the "Virginica" species showcases the largest features among them.

# ****Handling Outliers****

# 

**Before After**

# ****Using Elbow method to determine optional cluster number****

# 

# 

# As can be seen from the above graph, best number for clustering is 3

# ****Implementing K-Means Clustering****

# 

# 