**Experiment No:2**

**Aim: Data Wrangling II Create an “Academic performance” dataset of students and perform the following operations using Python. 1. Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them. 2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them. 3. Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution. Reason and document your approach properly.**

**Detailed Explanation of Exploratory Data analysis using Iris Dataset:**

For Complete code please visit: <https://github.com/Naidu-Bhavya/Exploratory-Data-Analysis-on-Iris-Dataset>

**What Is Exploratory Data Analysis?**

1) Exploratory data analysis is a task of analyzing the data using simple tools from statistics , some plotting tools , linear algebra.

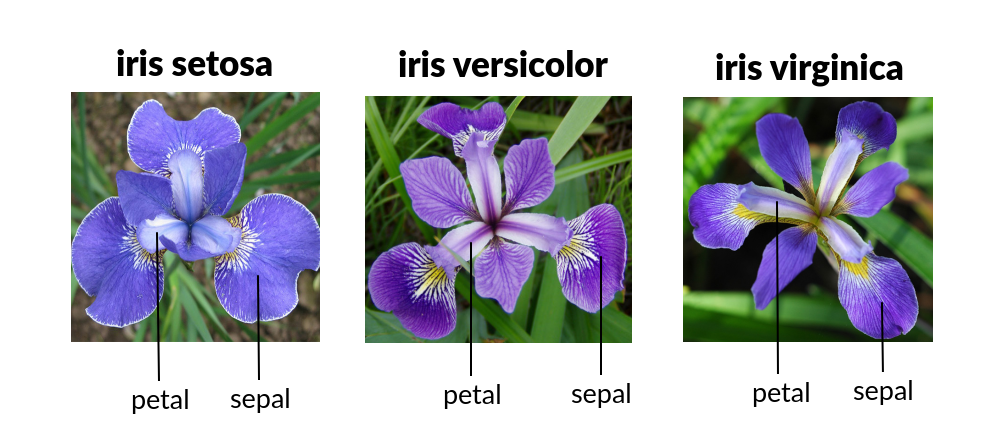
2) Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling of your data. By doing this you can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Pre-processing step or move on to modeling.

3) Once Exploratory Data Analysis is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling.

**Importance of EDA:** Many Data Scientists will be in a hurry to get to the machine learning stage, some either entirely skip exploratory process or do a very minimal job. This is a mistake with many implications, including generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, and using resources inefficiently because of realizing only after generating models that perhaps the data is skewed, or has outliers, or has too many missing values, or finding that some values are inconsistent.

**EDA on Iris Dataset:**

**Let’s see Iris Flower dataset**

**About the dataset:** Iris dataset is taken from <https://www.kaggle.com/arshid/iris-flower-dataset>

1) The dataset contains 4 features and 3 classes

2) Number of Instances: 150

3) Number of Attributes: 4 (including the class atribute)

4) Attribute Information: a) sepal\_length (numerical) b) sepal\_width (numerical) c) petal\_length (numerical) d) petal\_width (numerical)

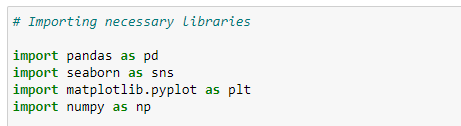
5) Missing Attribute Values: None

Now we are going to do EDA with the programming language named Python.

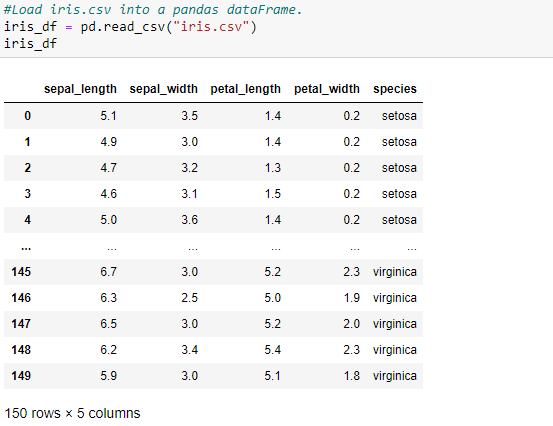
Python has a massive amount of library to apply the various types of operations on data to find out the best results.

**Objective:** Classify a new flower as belonging to one of the 3 classes given the 4 features.

**Importing some Libraries.**

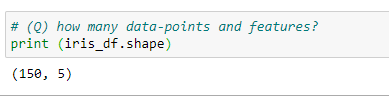


The next step is to load data. if data is in CSV format then use these lines of code to load data into a variable named Data.



1. **Analyzing the Data (High level statistics of the dataset):**

**Dimensionality of the dataset:**

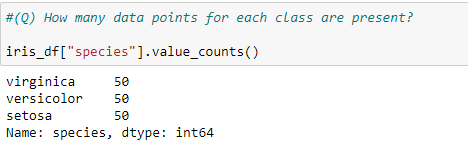


The shape of the data is (150, 4).

**Column names or feature names:**

https://miro.medium.com/max/882/1*dOQBGS3vBGCc--GYPH9AeQ.png

Index([‘sepal\_length’, ‘sepal\_width’, ‘petal\_length’, ‘petal\_width’, ‘species’], dtype = ‘object’ *)***How many points of each class:**



As we can observe all three classes are equally distributed in terms of the number of counts of each class. Here we can understand a very interesting concept called balanced and imbalanced dataset.

**Balanced Dataset**

Let say if we have 10000 rows and it has 2 classes. and class A has 4000 points(point = row) and class B has 6000 points and this dataset is the balanced dataset. (Examples: (5000, 5000), (5500, 4500), (6000, 4000))

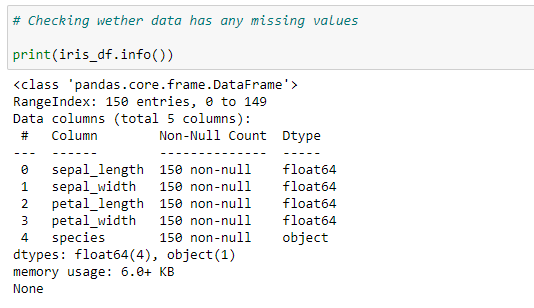
**Imbalanced Dataset**

Let say if we have 10000 rows and it has 2 classes. and class A has 2000 points and class B has 8000 points and this dataset is the imbalanced dataset. (Examples: (1000, 9000), (1500, 8500), (2000, 8000)).

Our aim is to get a balanced dataset because if data is imbalanced then our machine learning algorithms could be biased and inaccurate. there are various techniques to handle the case of an imbalanced dataset for you can go to this[link](https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/)

From the above explanation, you may get some idea about the balanced and imbalanced dataset.

**Some Basic information about the dataset:**



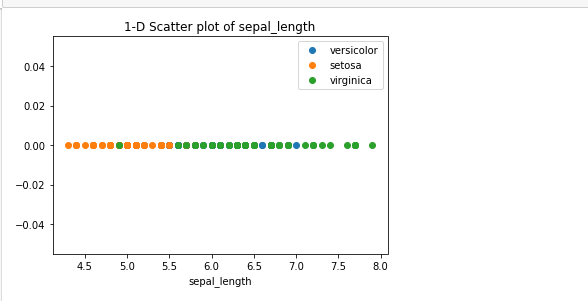
**Observation:**

Now we can say that there is not any data point is missing in any feature. And all first 4 features are of float64 type which is used in numpy and pandas. And indexes are from 0 to 149 for 150 entries. And the last column is of type object which is used for the class labels. Memory used by this data frame around 6 KB.

**2. Univaraite analysis:** we will do univaraite analysis to understand which features are useful for classification of class label ‘species’.

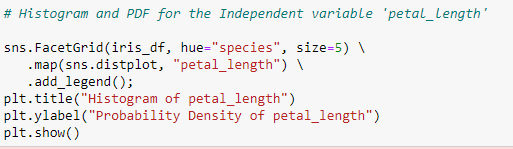
**1.1 1-D Scatter plots**

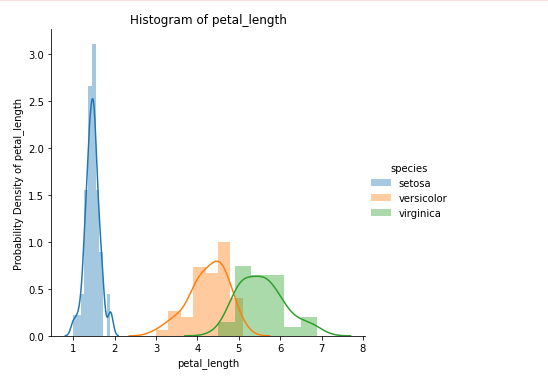




**Observation:** The disadvantage of this 1D scatter plot is a lot of overlap between Versicolor and Verginica and we can not say anything about it.

**1.2 Histogram and PDF:** If you wish to have both the histogram and densities in the same plot, the seaborn package (imported as sns) allows you to do that via the distplot(). Since seaborn is built on top of matplotlib, you can use the sns and plt one after the other.



**Observation:** Its X-axis tells us all setosa flowers are having a petal length between 1 and 1.8. And Versicolor having petal lengths between 3 and 5.2 and Virginica have petal length between 4.5 and 6.9. Its Y-axis tells us the count of the flower at this x value. or how often they come at this value. And Setosa is fully separated from the other two classes but Versicolor and Virginica are not fully separated they have some overlap of some data points.

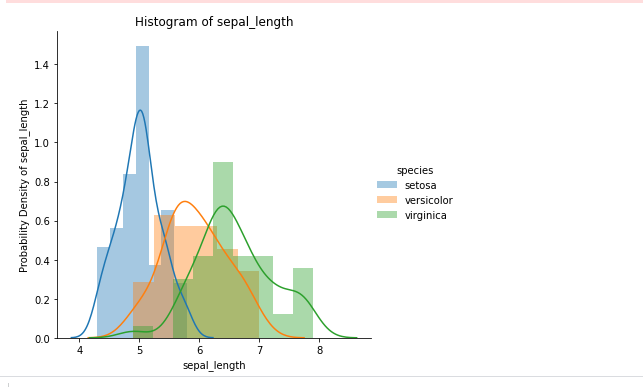
At x = 5 there is a high probability to get Virginica rather than Versicolor because of the height of the Virginica histogram if larger than Versicolor.

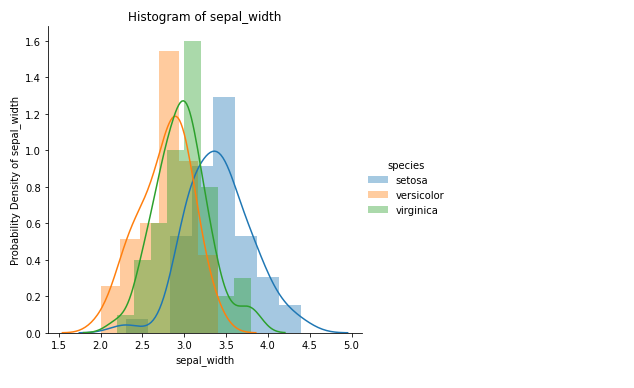
And this smooth curve is called PDF (Probability Density Function). it is a smooth histogram.

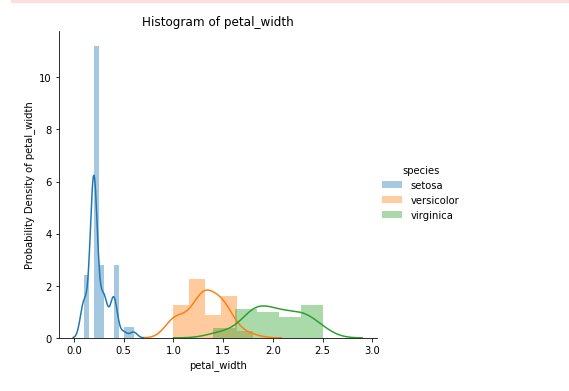
**PDF(Probability Density Function):** The PDF is the density of probability rather than the probability mass. The concept is very similar to mass density in physics. **OR**

Probability density function (PDF) is a statistical expression that defines a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) for a continuous [random variable](https://www.investopedia.com/terms/r/random-variable.asp) as opposed to a discrete random variable. When the PDF is graphically portrayed, the area under the curve will indicate the interval in which the variable will fall. The total area in this interval of the graph equals the probability of a continuous random variable occurring.

For other features:







**Observation:** Petal length will do a great job to classify all classes of flowers but we can say that petal length is slightly better than the petal width because petal width is also doing a better job.

1) PDFs and Histograms drawn considering ‘petal\_width’ feature is overlapped for versicolor and virginica. But the PDF of setosa is well seperated from for versicolor and virginica.

2) From the above PDFs we can say petal\_width of setosa lies below 0.5 units.

3) We can say petal\_width of virginica is more when compared to setosa and versicolor.

We can seperate/classify setosa flower perfectly based on ‘petal\_width’ alone. We can conclude that the features ‘petal\_length’ and ‘petal\_width’ are very useful to classify flowers (setosa, versicolor and virginica). We will proceed our further analysis using CDF to get more results

**Conclusion from Histogram and PDFs of all independent variables:**

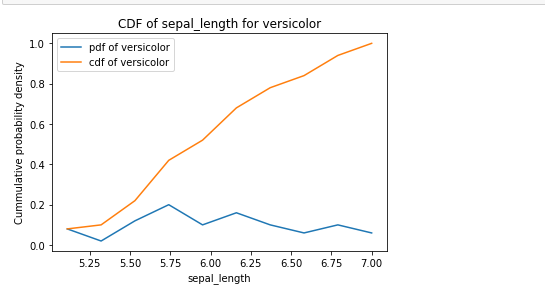
1) petal\_length of setosa lies between 1 and 2 units and petal\_length of virginica is more when compared to setosa and versicolor.

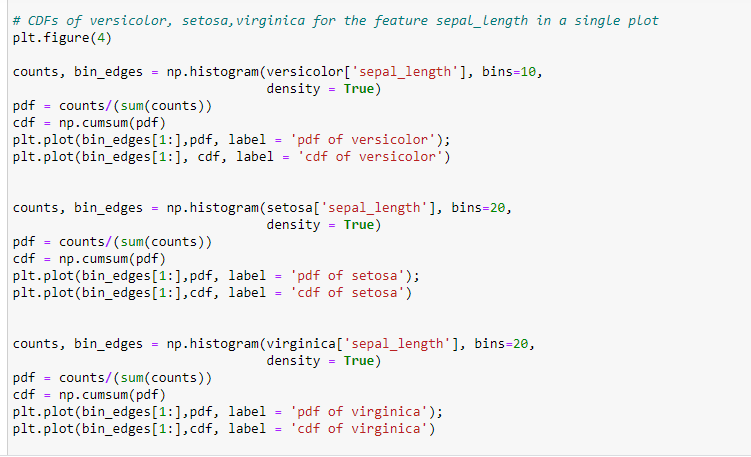
2) petal\_width of setosa lies below 0.5 units and petal\_width of virginica is more when compared to setosa and versicolor.

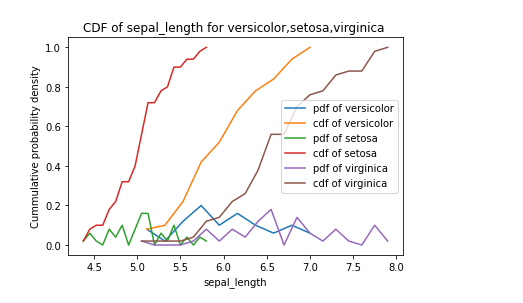
**1.3 CDF (Cumulative distribution function):**

In probability theory and statistics, the cumulative distribution function (**CDF**) of a real-valued random variable X, or just distribution function of X, evaluated at x, is the probability that X will take a value less than or equal to x.





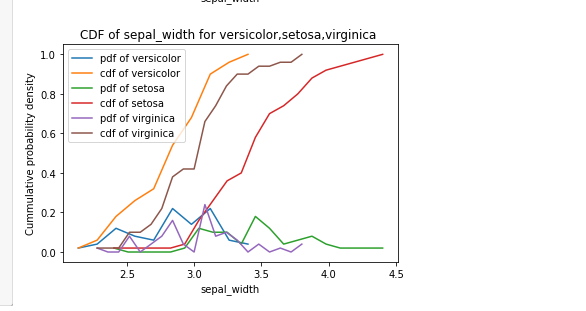


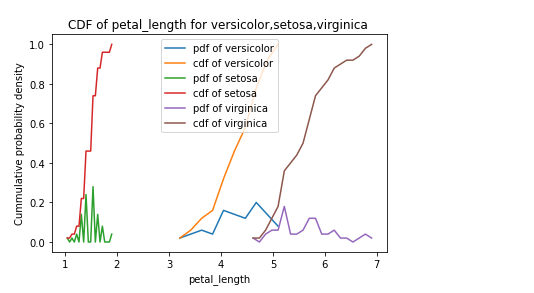


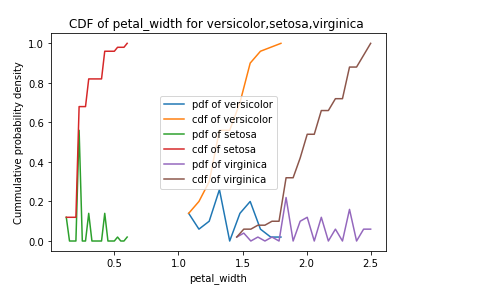
**Observation:**

From CDF we can observe that 100% of the setosa flowers have sepal\_length less than 5.5. We can say if the sepal\_length is <= 5.5 units, then the probability of the flower being setosa is high

For other features:







**Conclusion from above CDFs:** 1) For setosa flower , petal\_length will be lessthan or equal to 2 units and petal\_width will be lessthan 0.6 units.

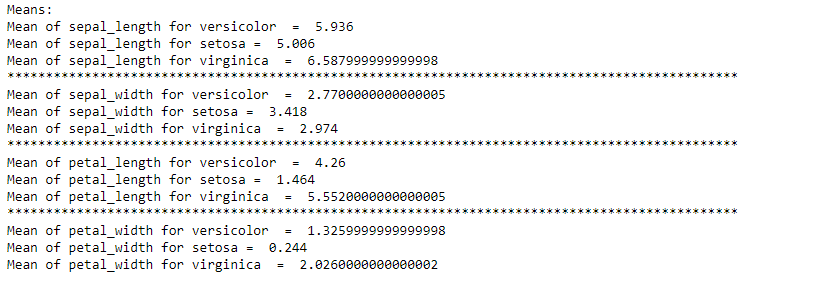
2) virginica has more petal\_length,petal\_width when compared to setosa,versicolor.

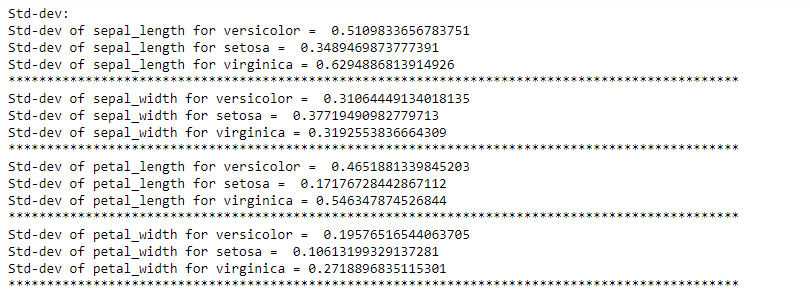
**2. Statistical analysis using Mean, Median, STD, Quantiles and Percentiles:**

**2.1 Mean and Std-dev**

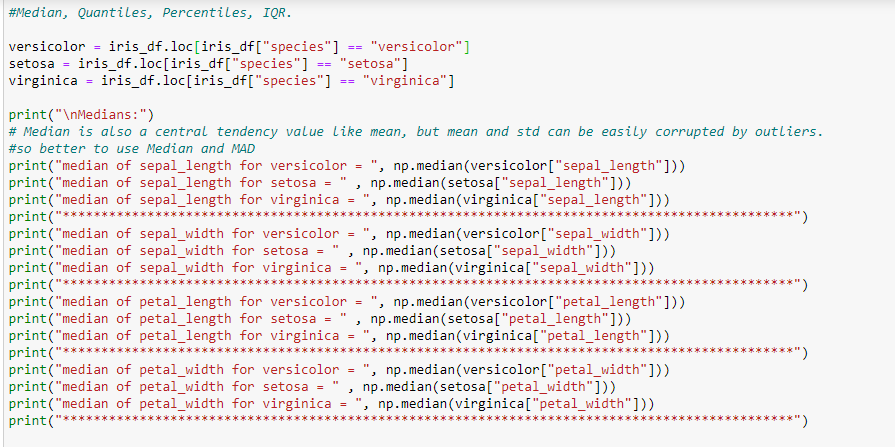




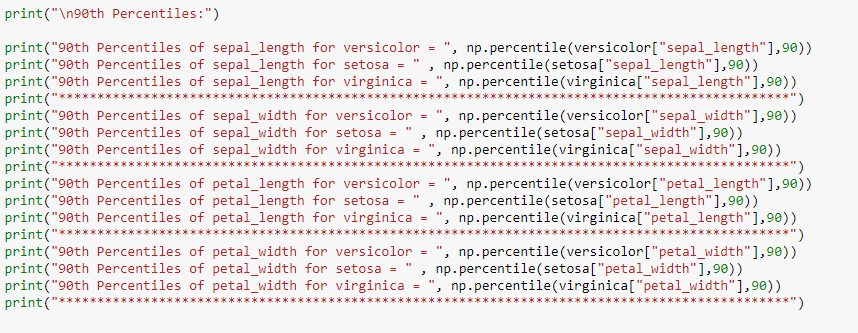


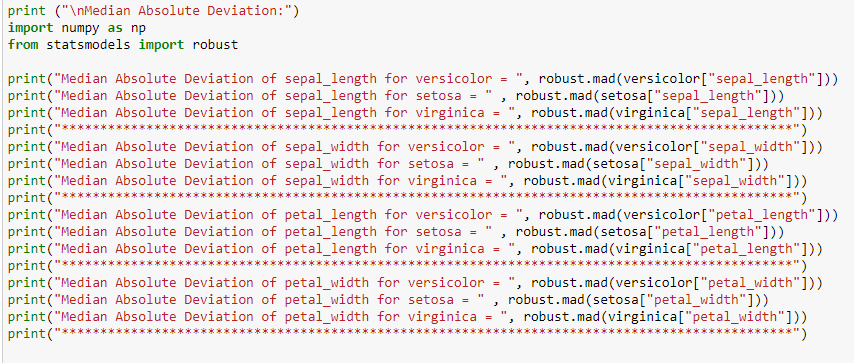


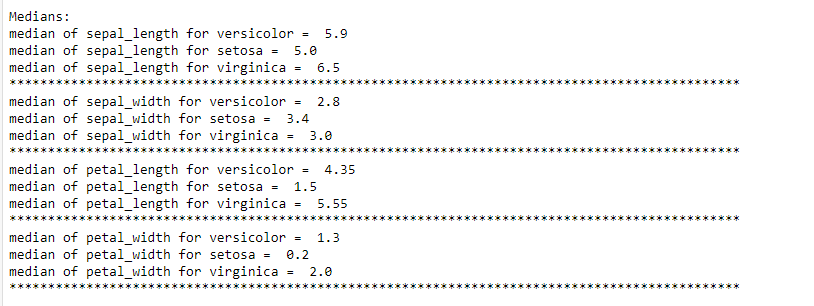
**Observations from mean and STD:** We know that mean is the average value , it is a central tendency value . Variance is the spread around mean 1) From above statistical data we can observe that setosa flower has small petal\_length (central tendency at 1.4 units with variance of 0.17 units) and petal width (central tendency at 0.24 units with variance of 0.10 units). 2) We can observe that sepal\_length , sepal\_width , petal\_length , petal\_width are more for virginica.

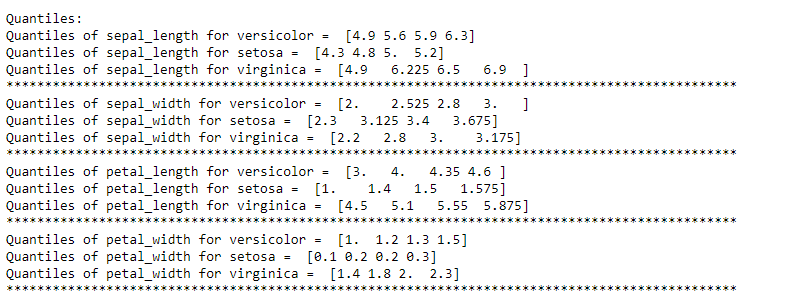
**2.2 Median, Percentile, Quantile, IQR, MAD:**

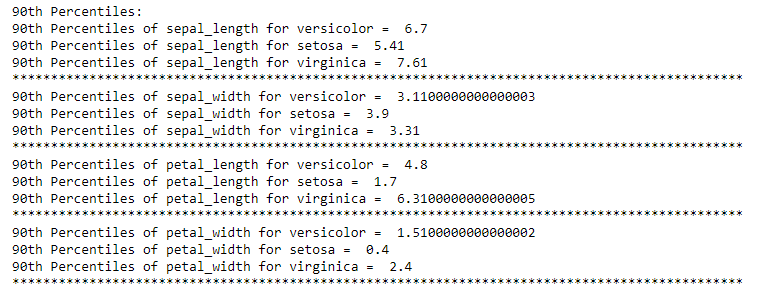


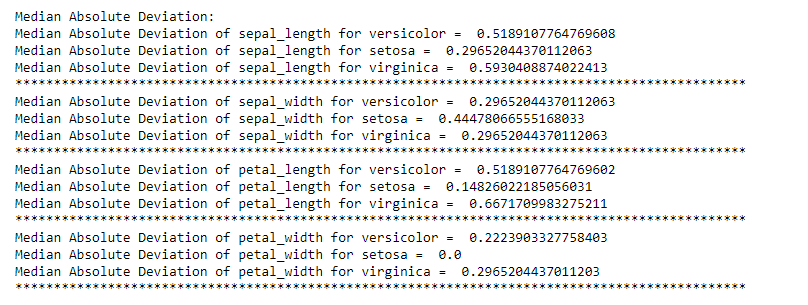










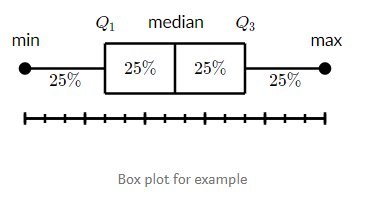


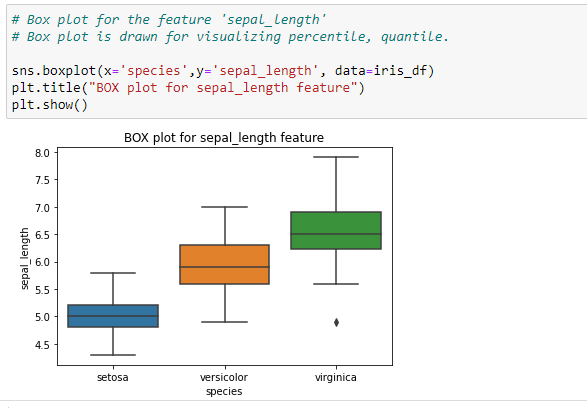
**Observations from Median, Quantiles, Percentiles, IQR:** Mean can be easily impacted/corrupted by outliers. So We go for Median: 1) From median values we can observe that setosa flower has less petal\_length and petal\_width. 2) 90% of the setosa flowers have petal\_length < 1.7 units and petal\_length < 0.4. 3) We can observe that sepal\_length , sepal\_width , petal\_length , petal\_width are more for virginica.

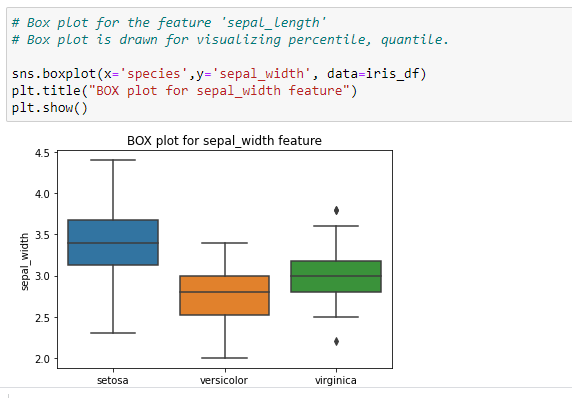
**3. Univariate analysis with BOX plots and Violin plots:**

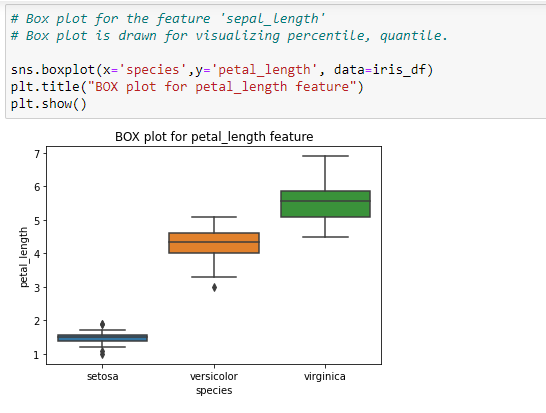
**3.1 Box plot and Whiskers**

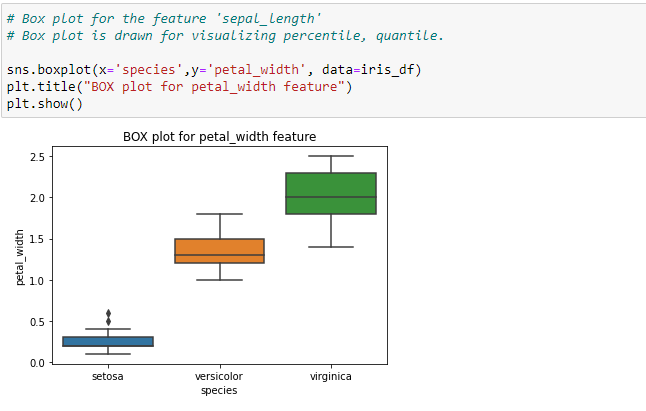
A box and whisker plot — also called a box plot — displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum. In a box plot, we draw a box from the first quartile to the third quartile. A vertical line goes through the box at the median. The whiskers go from each quartile to the minimum or maximum.









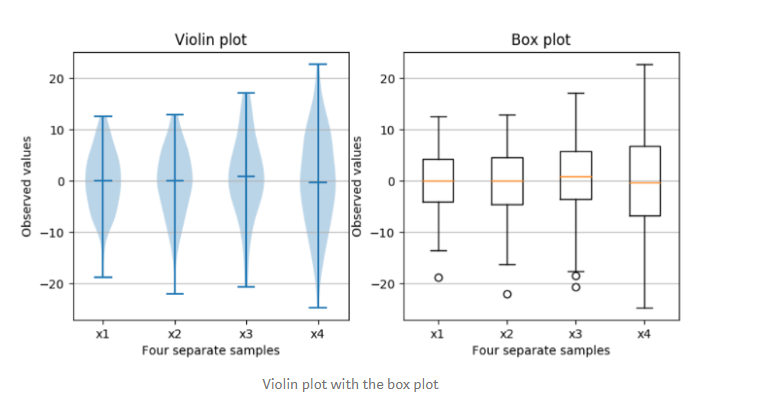


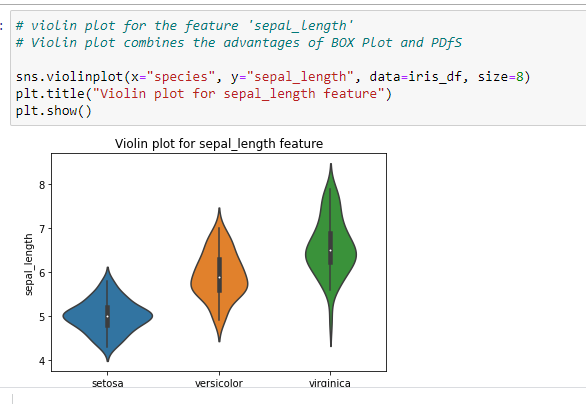
**Observation from Box plot:** We know that with in the BOX 25th percentile to 75th percentile values lies. Means 50% of the points lies with in the box. 1) From the Box plots we can observe that box plots of features (‘sepal\_length’,’sepal\_width’, ‘petal\_length’,’petal\_width’) for 3 different classes (versicolor, setosa, virginica) are not overlaped completely. It means all four features are important in classification task. 2) If we observe Box plots of ‘petal\_length’ and ‘petal\_width’ features we can say using these two features we can sepearte setosa from versicolor and virginica as Box plots of ‘petal\_length’ and ‘petal\_width’ features of setosa doesnot overlap with versicolor and virginica.

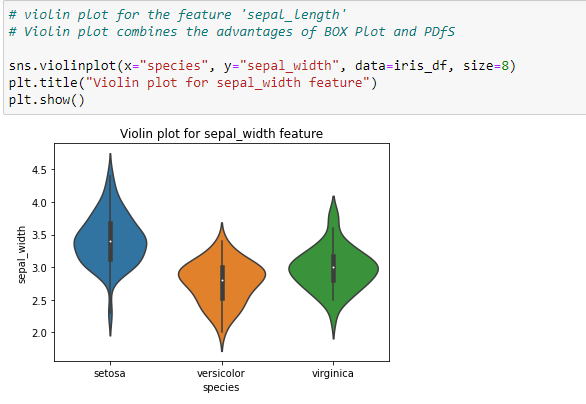
**3.2 Violin plots:**

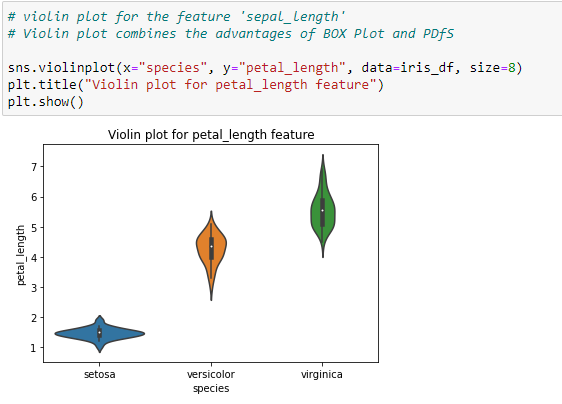
**What is a violin plot?**

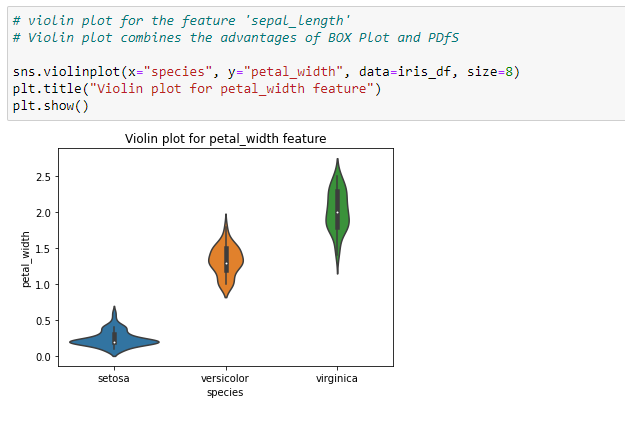
A violin plot is a method of plotting numeric data. It is similar to a box plot, with the addition of a rotated kernel density plot on each side. Violin plots are similar to box plots, except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.









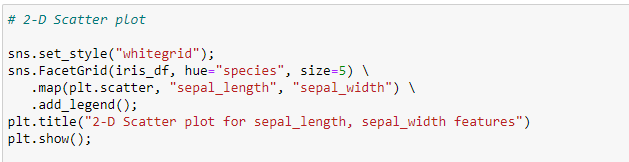


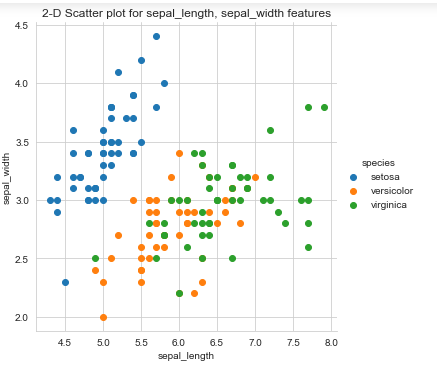
**Observations from violin plot:**

Violin plot combines the advantages of BOX Plot and PDfS: From the Violin plots we can observe that box plots of features (‘sepal\_length’,’sepal\_width’, ‘petal\_length’,’petal\_width’) for 3 different classes (versicolor, setosa, virginica) are not overlaped completely. It means all four features are important in classification task.

**4. Bi-Variate analysis / Multivariate analysis:**

**4.1 2-D Scatter Plots: Scatter plot:** A scatter plot is a set of points plotted on horizontal and vertical axes. Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables). If no correlation exists between the variables, the points appear randomly scattered on the coordinate plane. If a large correlation exists, the points concentrate near a straight line.

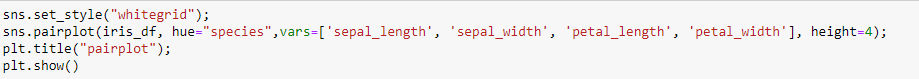


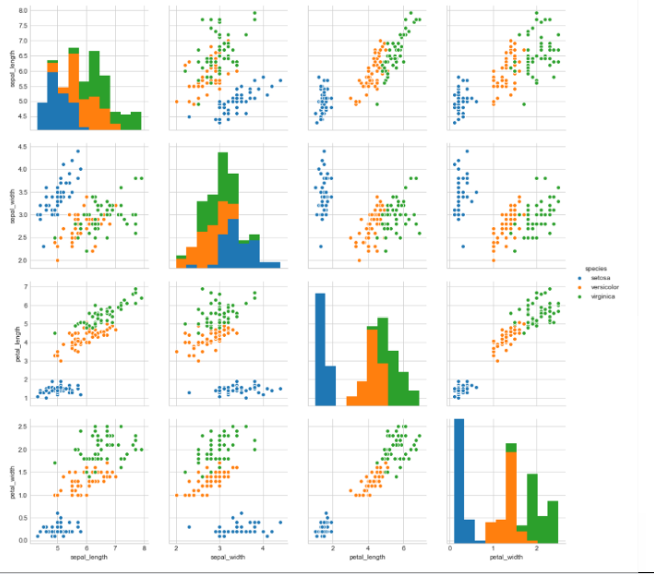


**Observation from 2-D Scatter plot:** The 2-D scatter plot drawn above is almost overlapped for versicolor and virginica. However we can interpret the following points from the plot.

1. We can separate setosa from versicolor and virginica using sepal\_length and sepal\_width features

**4.2 Pair-plot (To observe all 2-D Scatter Plot):** Objective here is to find which two independent variables are better in classifying the flowers





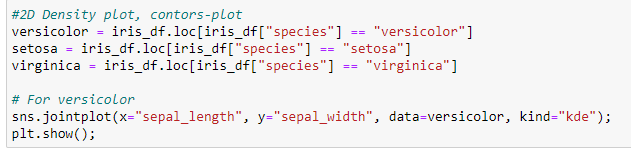
**Reading the pairplot:** Here we will get 4c2 = 6 plots (4 Independent variables selecting 2 at a time). In pair plot we won’t consider principle diagonal plots. Plots above the diagonal and below the diagonal are just mirror images of each other.

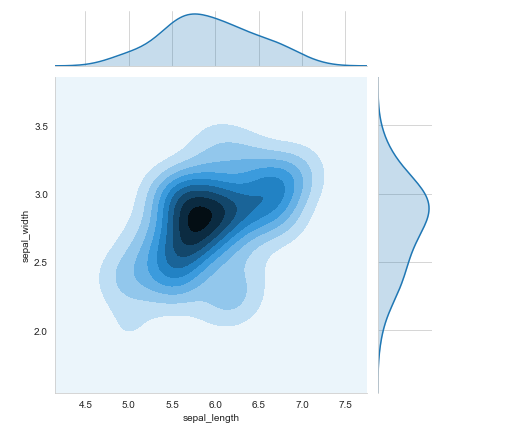
**Observations from pair plot:**

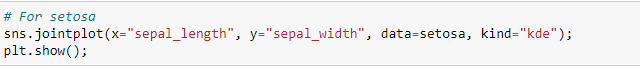
1) From pair plot , we observe 2-D scatter plots, as described below.

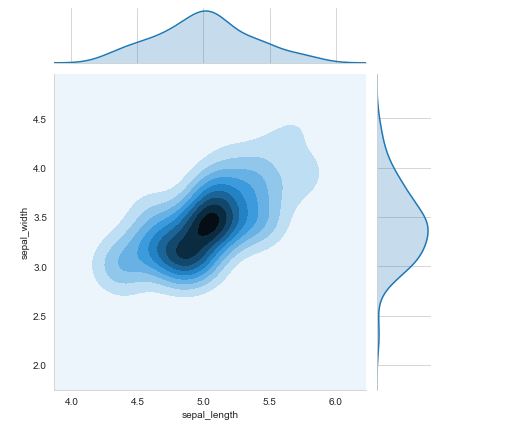
1) From the plots between ('petal\_length', 'petal\_width') , ('sepal\_width', 'petal\_width'),   
 ('sepal\_length','petal\_width') we can observe that setosa is well seperated from versicolor and virginica. 2) From the plots between ('petal\_length', 'petal\_width') we can seperate versicolor and virginica eventhouh there slight overlap.

**5. Multivariate probability density plot or Contour Plot:** Contour plot is the density plot in 2-D considering two features at a time. The more denser area will be darker in colour and less denser area is lighter in colour. We can imagine Contour plot like hill coming out of the screen

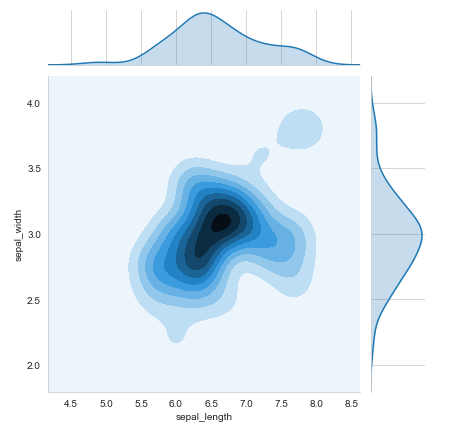


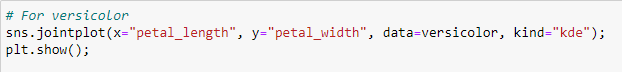


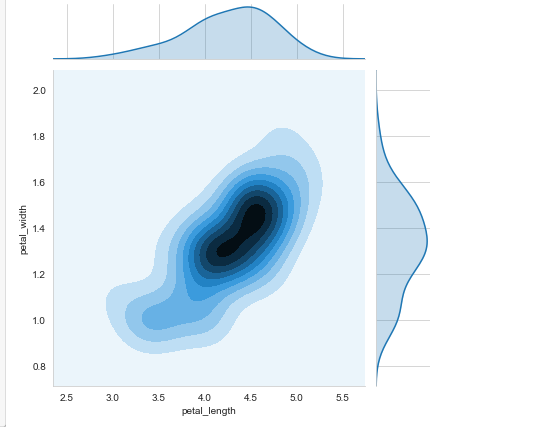












**Observations from contour plots:** From the contour plot drawn between petal\_width and petal\_length of setos flower we can observe that most of the setosa flowers will have petal\_length between 1.4 to 1.6 units and petal width in the range 0.1 to 0.2 units.

**Conclusion from above analysis:** All four features ‘sepal\_length’, ‘sepal\_width’, ‘petal\_length’, ‘petal\_width’ are useful in classifying the species. 2) Using ‘petal\_length’, ‘petal\_width’ features we can perfectly sepearte setosa from versicolor,virginica. 3) Using ‘petal\_length’, ‘petal\_width’ features we can seperate versicolor and virginica to some extent. 4) Most of the virginica flowers have high values for all the four features ‘sepal\_length’, ‘sepal\_width’, ‘petal\_length’, ‘petal\_width’.  **References:**

[https://www.appliedaicourse.com](https://www.appliedaicourse.com/)