10. Data Visualization III

**Aim :**

Download the Iris flower dataset or any other dataset into a DataFrame. (e.g., https://archive.ics.uci.edu/ml/datasets/Iris ). Scan the dataset and give the inference as:

1. List down the features and their types (e.g., numeric, nominal) available in the dataset.

2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

3. Create a box plot for each feature in the dataset.

4. Compare distributions and identify outliers.

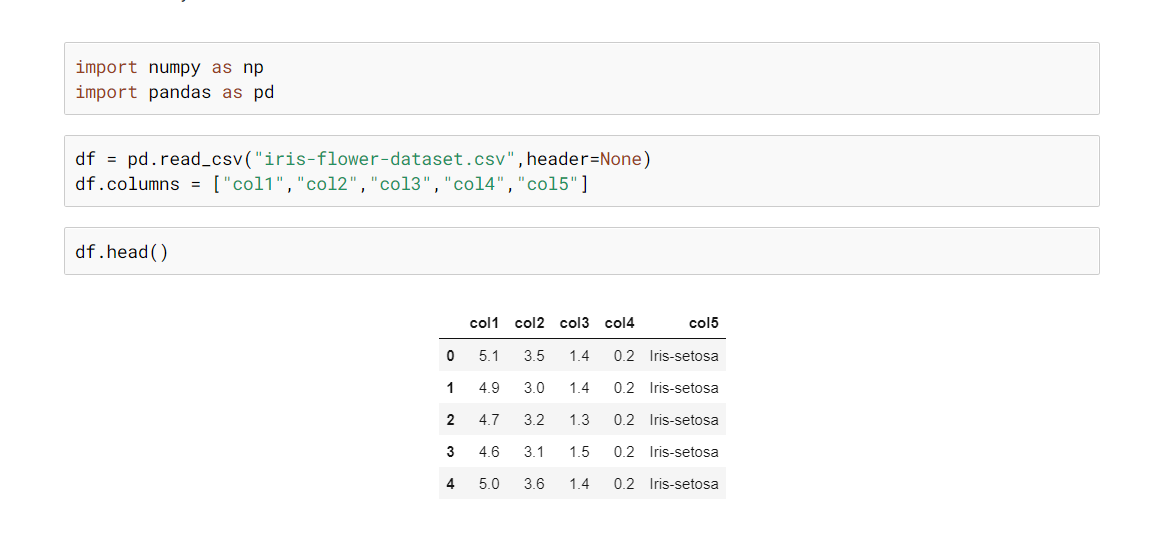
**Objective :**

Studying the Iris Dataset and listing down the inferences.

**What is Data Analysis?**

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

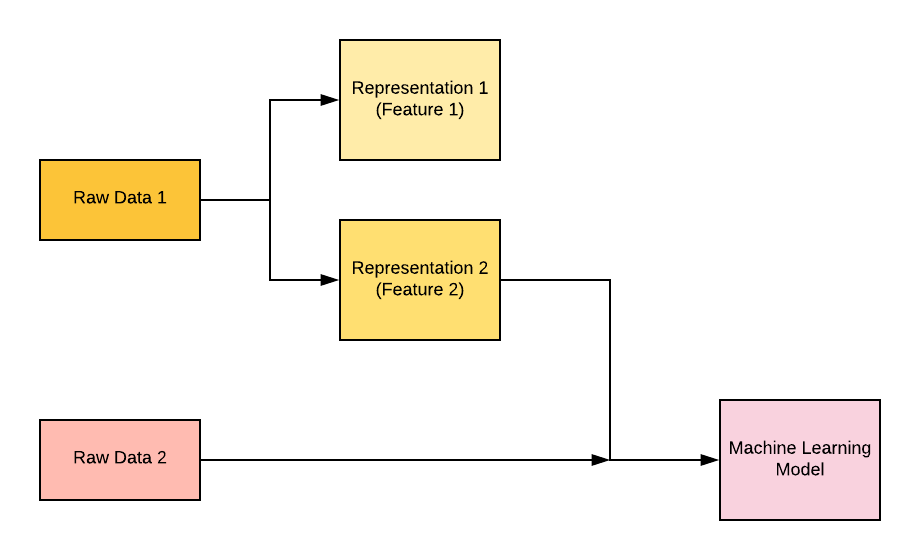
**Load libraries & data**



**What are features ?**

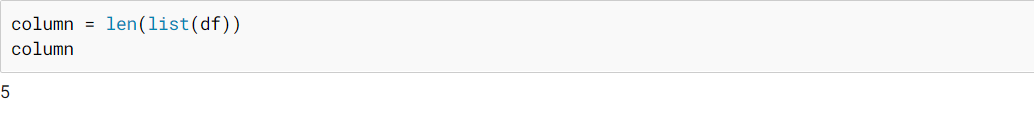
Features are nothing but the independent variables in machine learning models. What is required to be learned in any specific machine learning problem is a set of these features (independent variables), coefficients of these features, and parameters for coming up with appropriate functions or models (also termed as hyperparameters).

Features can be in the form of raw datathat is very straightforward and can be derived from real-life as it is. However, not all problems can be solved using raw data or data in its original form. Many times, they need to be represented or encoded in different forms.

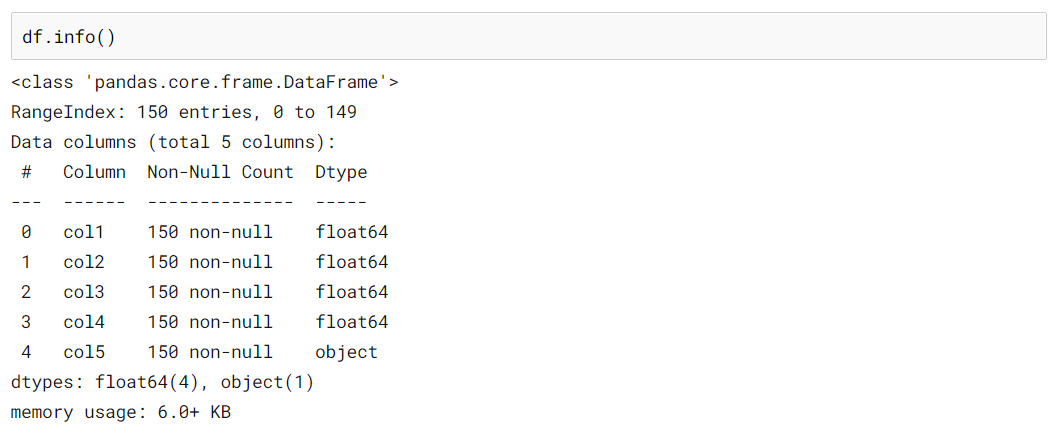


**Fig. Features – Key to Machine Learning**

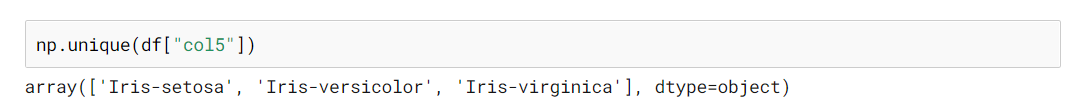
1. List down the features and their types (e.g., numeric, nominal) available in the dataset.



Dataset has 5 columns indicating 5 features about the data.



Dataset contains 4 numerical columns & 1 object column.



1. Create a histogram for each feature in the dataset to illustrate the feature distributions.

**Histogram :**

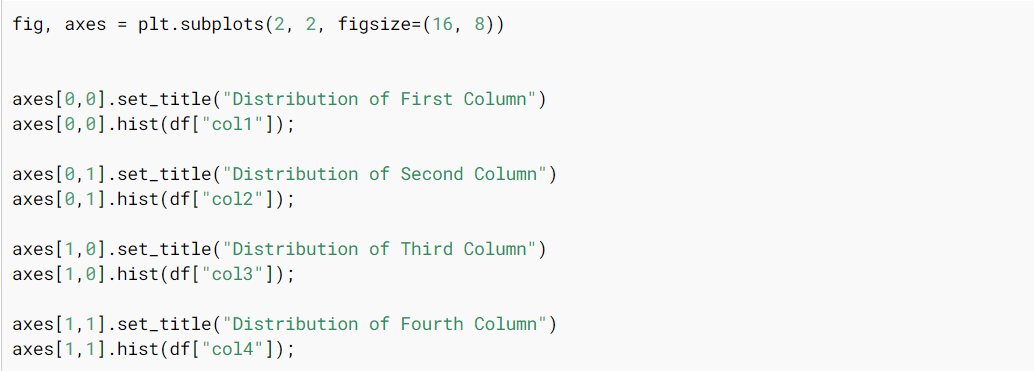
Histograms group the data in bins and is the fastest way to get idea about the distribution of each attribute in dataset. The following are some of the characteristics of histograms –

* It provides us a count of the number of observations in each bin created for visualization.
* From the shape of the bin, we can easily observe the distribution i.e. weather it is Gaussian, skewed or exponential.
* Histograms also help us to see possible outliers.

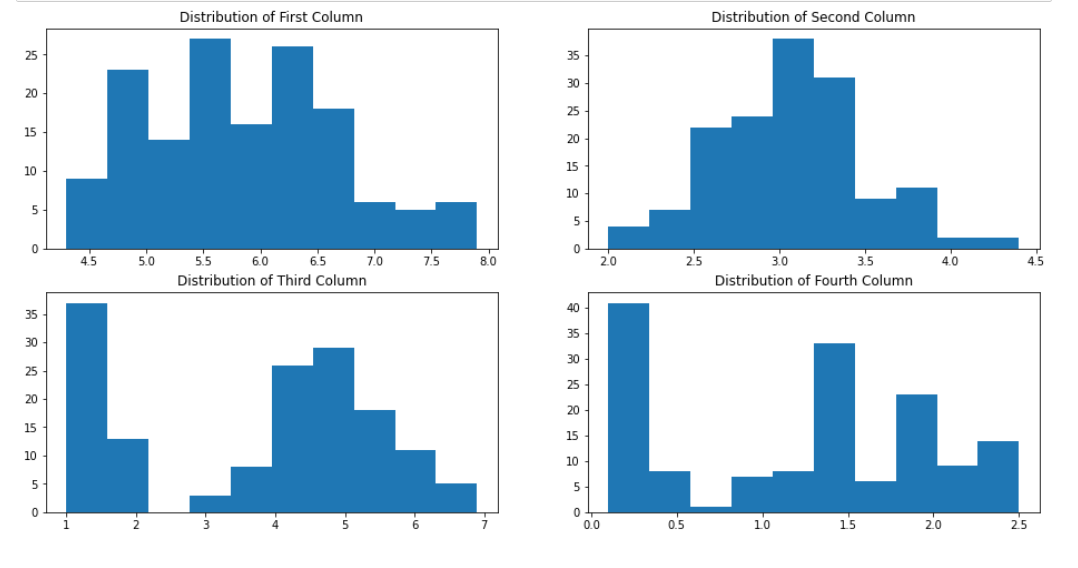
Import visualization libraries seaborn , matplotlib.



Code :



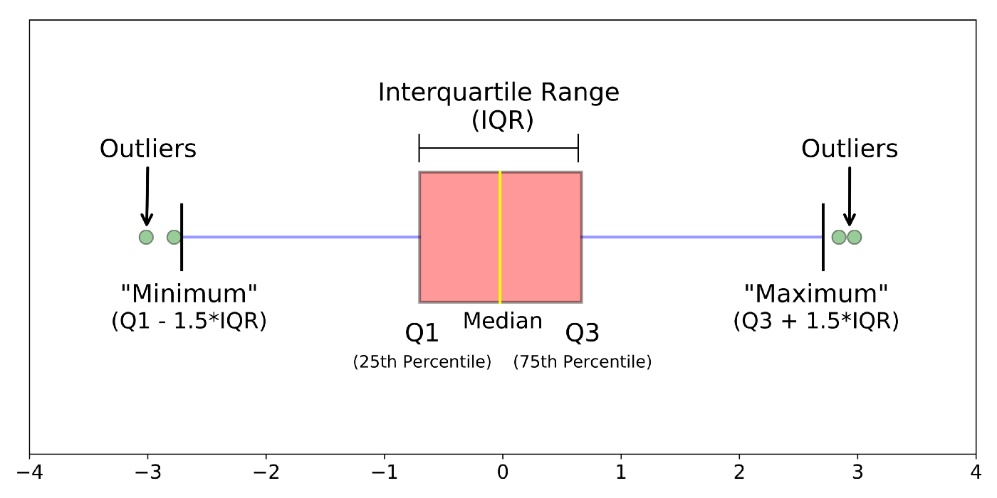
Output :

It

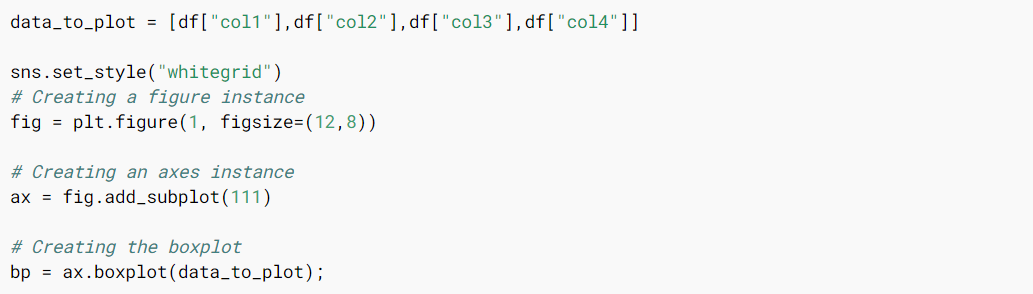
3. Create a box plot for each feature in the dataset.

What is Boxplot?

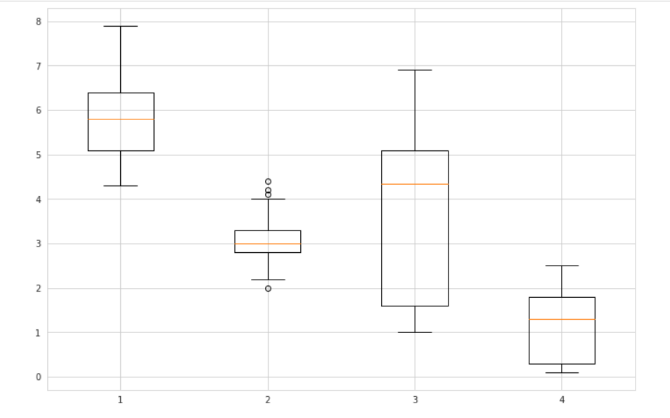
A boxplot is a standardized way of displaying the distribution of data based on a five number summary (“minimum”, first quartile (Q1), median, third quartile (Q3), and “maximum”). It can tell you about your outliers and what their values are.



Code :



Output :



4. Compare distributions and identify outliers.

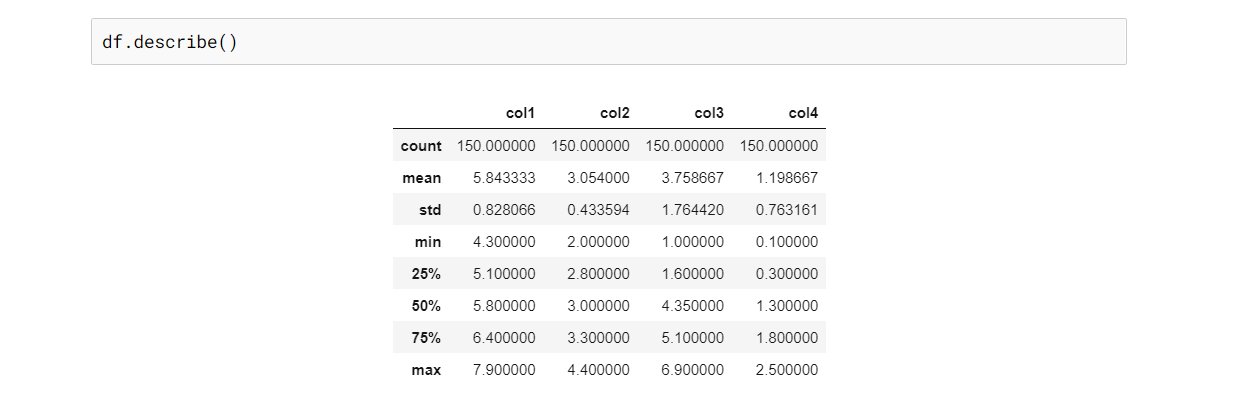
What is Distribution ?

A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.

Why are Distribution important in Machine Learning?

In Machine Learning, data satisfying Normal Distribution is beneficial for model building. It makes math easier. Models like LDA, Gaussian Naive Bayes, Logistic Regression, Linear Regression, etc., are explicitly calculated from the assumption that the distribution is a bivariate or multivariate normal.

Summary statistics for each feature available in the dataset.



What is Outlier ?

An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.

For identifying outliers.

First of all Box plot is a data visualization plotting function. It shows the min, max, median, first quartile, and third quartile. All of the things will be explained briefly. All of the property of box plot can be accessed by *dataframe.column\_name.describe()* function.

Explanation of the different parts of the box plot.

The maximum and the minimum is the max and min value of the data-set. 50 percentile is the median of the data-set. The first quartile is the median of the data between the min to 50% and the third quartile is the median of the data between 50% to max. The outliers will be the values that are out of the (1.5\*interquartile range) from the 25 or 75 percentile.

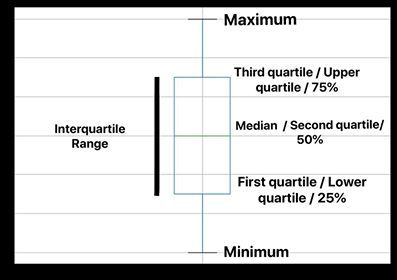


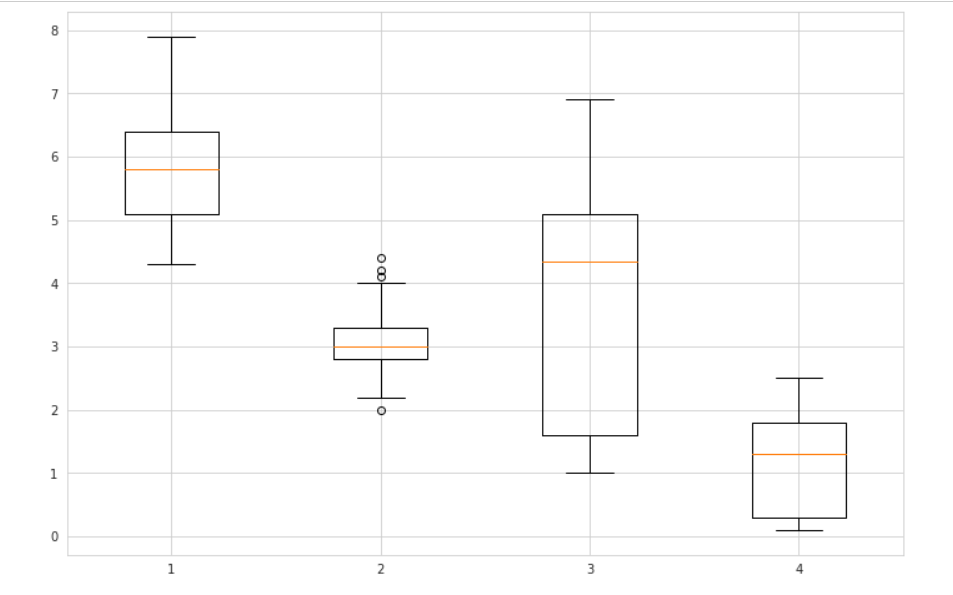
Fig. Features of Boxplot

**Methods of finding the values**

* Use the median to divide the ordered data set into two halves.  
  1) If there is an odd number of data points in the original ordered data set, do not include the median (the central value in the ordered list) in either half.  
  2) If there is an even number of data points in the original ordered data set, split this data set exactly in half.
* The lower quartile value is the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.
* An extreme value is considered to be an outlier if it is at least 1.5 interquartile ranges below the first quartile, or at least 1.5 interquartile ranges above the third quartile.

The box plot seem useful to detect outliers but it has several other uses too. Box plots take up less space and are therefore particularly useful for comparing distributions between several groups or sets of data. It is a direct representation of the Probability Density Function which indicates the distribution of data.

Box plot of Iris data set



If we observe closely. for the box 2, interquartile distance is roughly around 0.75 hence the values lying beyond this range of (third quartile + interquartile distance) i.e. roughly around 4.05 will be considered as outliers. Similarly outliers with other boxplots can be found.

Conclusion :

We studied the iris dataset & plotted box plot for identifying outliers.

Questions :

1. What is Data Analysis?
2. What are outliers?
3. Explain different parts of boxplot?
4. What are different types of distribution?