Univariate Analysis

In Data Analysis, it is important to understand the nature of each variable and also relationship between two or more variables to get an insight which could help us provide a solution to the problem, improvement to the existing process, bring profits or avoid losses in terms of Business and much more… There are three types of analysis based on the number of variables that are considered for analysis. They are **Univariate analysis, Bivariate Analysis and Multivariate Analysis.**

Univariate Analysis

When there is only one variable that is involved in the analysis, then it is called Univariate Analysis. Example include height of students in a particular class, weight of people in a particular gym. This analysis is used usually to understand the spread of data between its maximum and minimum limits. The variable that is selected for analysis is initially categorised into continuous or discrete variable. There are two methods of doing univariate analysis - Summary Statistics and Charts.

Univariate Analysis using Summary Statistics

In pandas, there are certain functions and methods that are used to represent the nature, location, distribution and dispersion of the data.

1. **The info method -** This method displays all the columns, their count and data type of each of the variable. These are used for both discrete and continuous variables
2. **The describe method -** This method displays all the columns that has continuous values, their count, mean, standard deviation and also 5 point statistics (Min, 25%, 50%, 75%, Max)

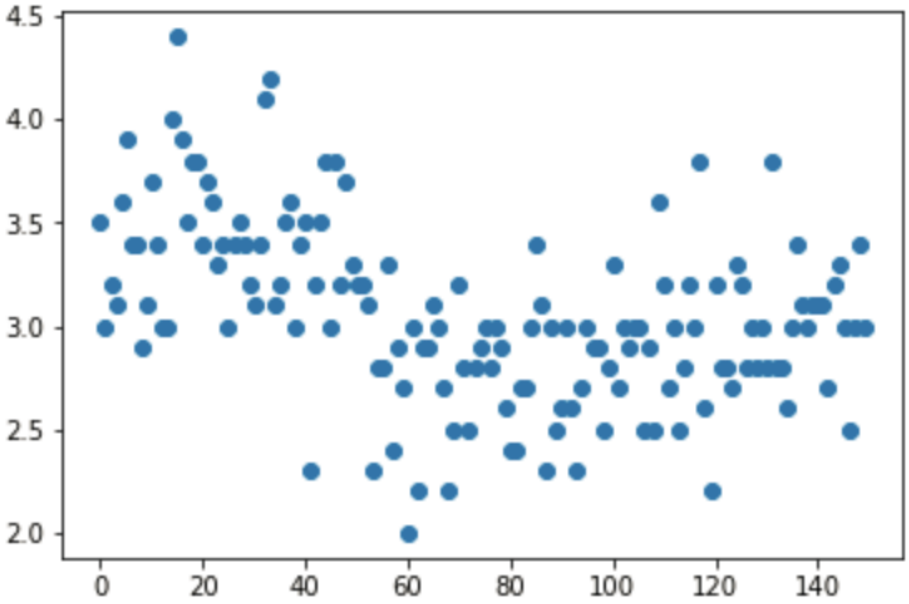
Univariate Analysis using charts

Based upon this characterisation, following are the different visualisation methods used in univariate analysis to help us understand the location of the a particular observation in the data variable and also its distribution and dispersion.

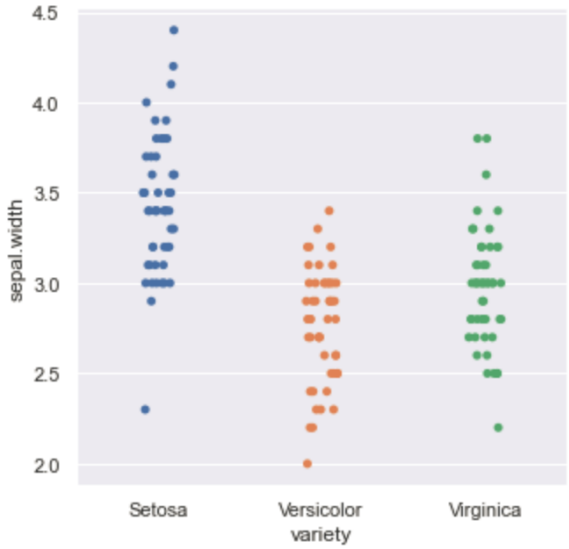
The classification further goes on like uni-variate enumerative plots (which gives the observation of data and its distribution - scatter plot, strip plot, swarm plot) and uni-variate summary plots (which gives the location, distribution and dispersion more

precisely than an enumerative plot - histogram, density plot, box plot, rug plot, dist plot, violin plot, bar charts)

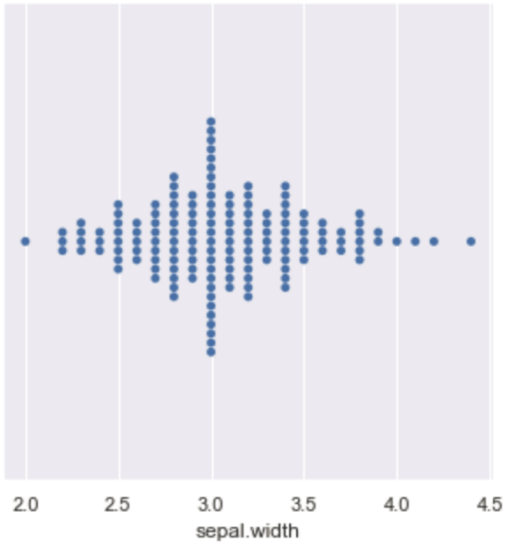
1. **Univariate Scatter plot -** Usually scatter plot is used to compare between two variables, but here the variable under analysis is plotted against the index of the dataset.



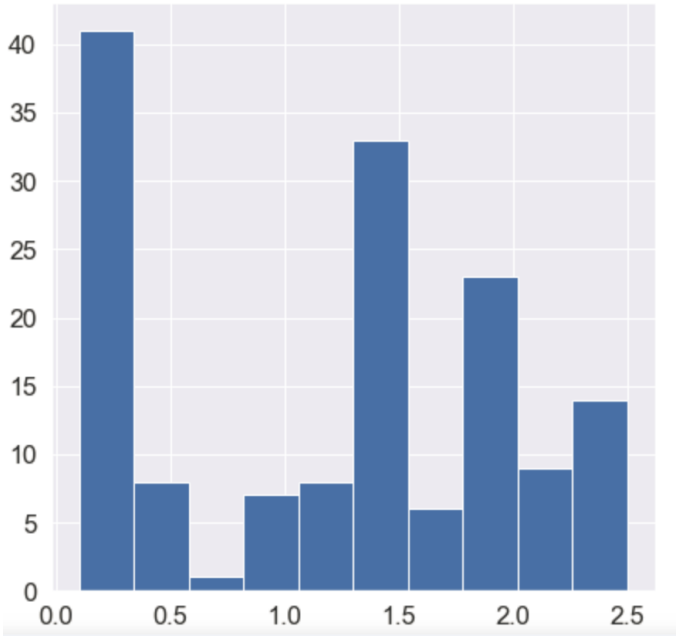
2. **Strip plot -** This is basically similar to scatter plot, except that it differentiates between categories and creates a own scatter within a category.



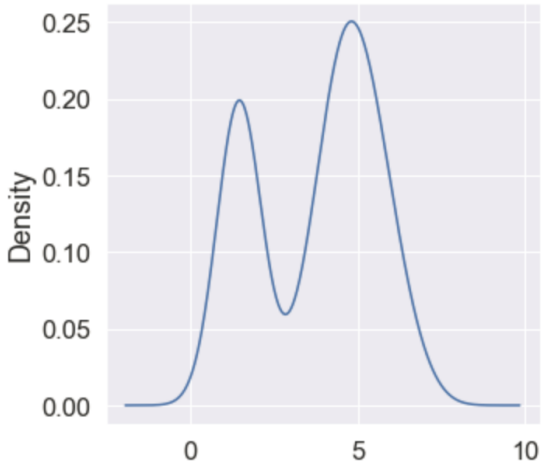
3. **Swarm plot -** Swarm plot is similar to strip plot, but it avoids the overlap of continuous values and provides better visualisation.



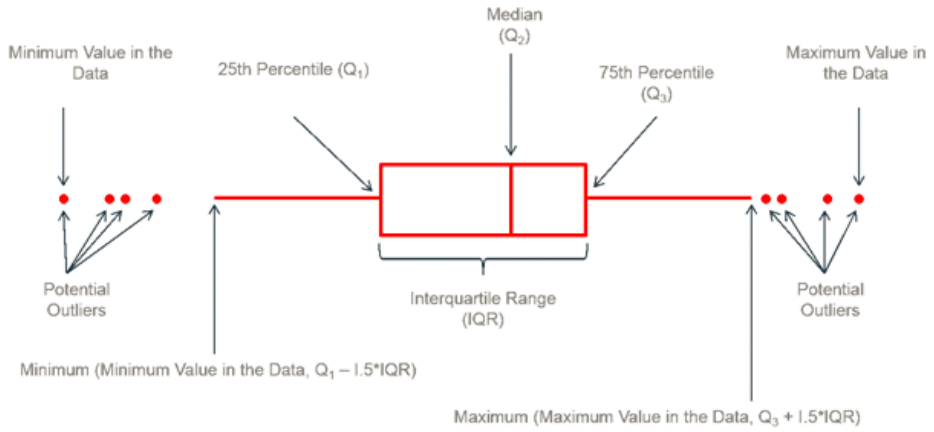
4. **Histograms -** Histograms are bar charts which has the class intervals in the x-axis and frequency of data points in the dataset falling in this interval in y-axis.



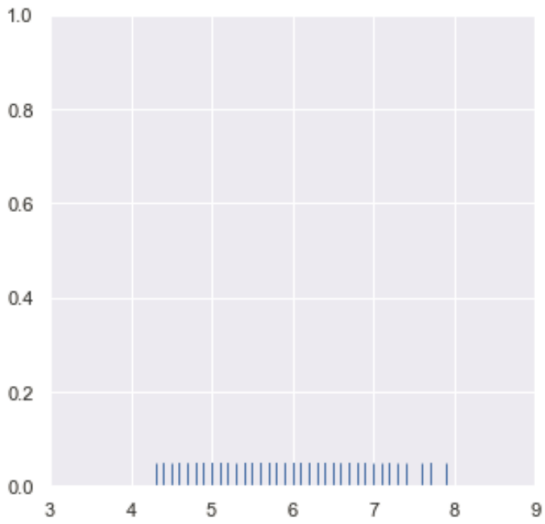
5. **Density plots -** Density plots are smoother versions of histograms, instead of bar charts the data points are represented by a curve.



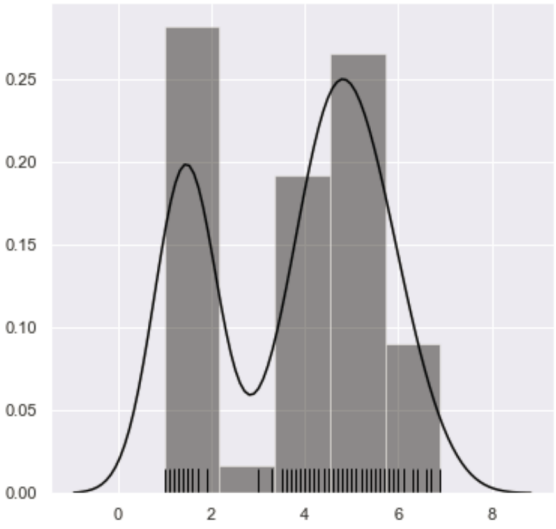
6. **Box plots -** Box plots are used to provide the distribution of data in 5-number summary format. The box plots are used for continuous variables.



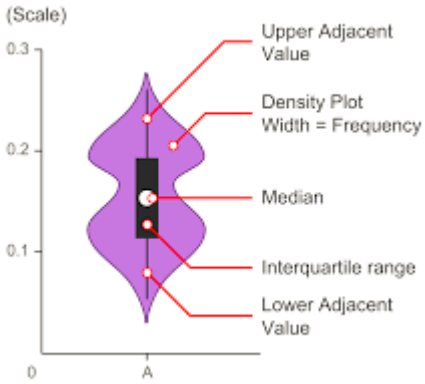
7. **Rug Plot -** Rug plots are similar to bar charts, except that they are of equal heights and are represented in a continuous manner instead of intervals or binning in bar plots. The frequency of the values depend upon the thickness of the bar. When these charts are used with other charts then they provide great meaning.



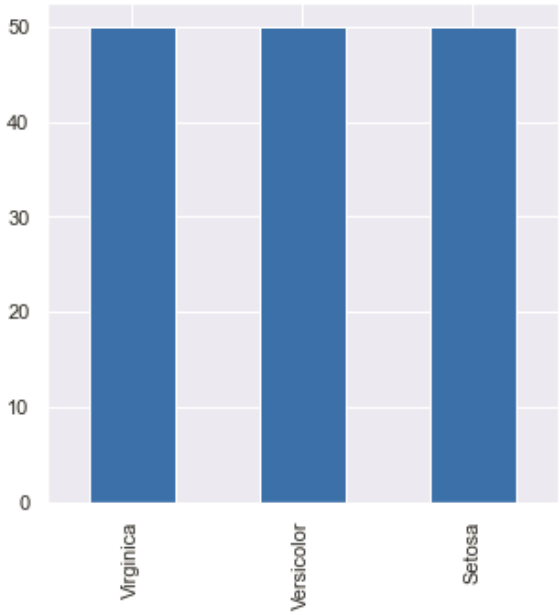
8. **Dist plots -** These plots combine the hist plot in matplotlib with kde plot and the rug plot.



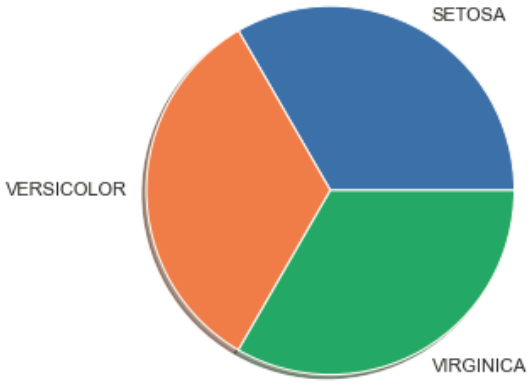
9. **Violin plots -** It is similar to box plot, except that it has a rotated kernel density plot to both the sides of the center line.



10. **Bar Charts -** These charts are used extensively for categorical variable to understand the frequency of a particular class in the data.



11. **Pie Charts -** These are also used for categorical variables. It differs from the bar chart in the sense that it shows the presence of a particular percentage of a class in the variable.


SETOSA
VERSICOLOR
VIRGINICA