# **EXPLORATORY DATA ANALYSIS(EDA)**

Exploratory data analysis is a simple classification technique usually done by visual methods. It is an approach to [analysing data sets](https://www.edureka.co/blog/football-world-cup-best-xi-analysis-using-python/) to summarize their main characteristics. when you are trying to build a [machine learning model](https://www.edureka.co/blog/logistic-regression-in-python/) you need to be pretty sure whether your data is making sense or not.

EDA is a task of analysing data using simple tools from statistics, simple plotting tools.

What is the need of EDA?

Every machine learning problem solving starts with EDA. It is probably the most important part of a machine learning project. With the growing market, the size of data is also growing. It becomes harder for companies to make decision without proper analysing it.

Here data scientist plays their part. With the use of charts and certain graphs, one can make sense out of the data and check whether there is any relationship or not.

Various plots are used to determine any conclusions. This helps the company to make a firm and profitable decisions. Once **Exploratory Data Analysis**is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modelling.

How can we perform EDA?

There are a lot of tools where one can perform EDA. Programming languages used are mainly R and python. Tools like Tableau, IBM Cognos etc are used which are often known as Business intelligence tools (BI).

Talking about Python, we use certain libraries like NumPy, Pandas, Matplotlib and seaborn for EDA.

Talking about performing EDA, it is like an art for data scientist. It shows how creative we can be or should be while dealing with EDA. The more creative we become with data more insights we can visualise. So, while performing EDA, always ask the right question, be more creative towards data and understand the pattern thoroughly.

One should always have some basic information related to the problem domain. For example, here we will deal with Haberman Dataset which is regarding cancer. So some medical terms or knowledge is required when tackling this Dataset.

Some methods and plots are distinguished as: -

* Univariate analysis
* Bivariate analysis
* Multivariate analysis

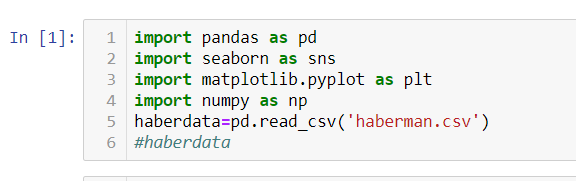
Here are the common graphs used while performing EDA:

* Scatter Plot
* Pair plots
* Histogram
* Box plots
* Violin Plots
* Contour plots

Let’s head over to an actual example on how to perform EDA. Here we are using a simple dataset which is the [Haberman Dataset](https://www.kaggle.com/gilsousa/habermans-survival-data-set).

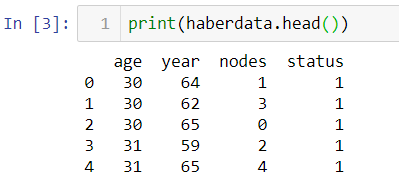
I am attaching my code in [github](https://github.com/Vihaanshah29/Haberman_dataset_EDA) for reference.

1. Import the necessary libraries.

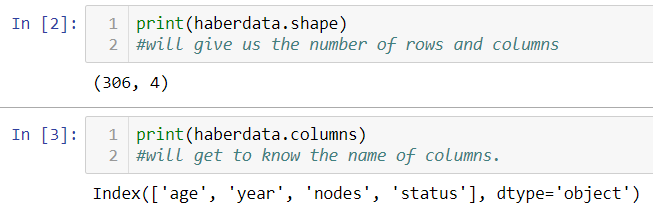
For performing EDA, we need to import some libraries.

We imported the libraries and with the help of pandas, we read the csv file of the Dataset and stored into variable haberdata.

1. Display few rows of dataset

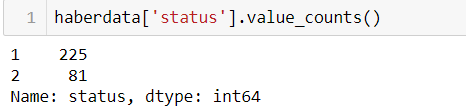


1. Understand the shape and column names of the data



* Age: - It represents the age of the patients undergone the surgery. It ranges from 30 to 83. (Can be determine by the code: haberdata['age']
* Year- Year in which the patients had the operation. It ranges from 1958–1969.
* Nodes:- A lymph node, or lymph gland is a kidney-shaped organ of the lymphatic system, and the adaptive immune system.(reference:-<https://en.wikipedia.org/wiki/Lymph_node>)
* Status: - Denoted by 1 and 2. 1 means the patient survived 5 years or longer and 2 means the patient died within 5 year.

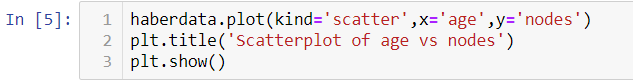
1. Understanding what our Class Label will be



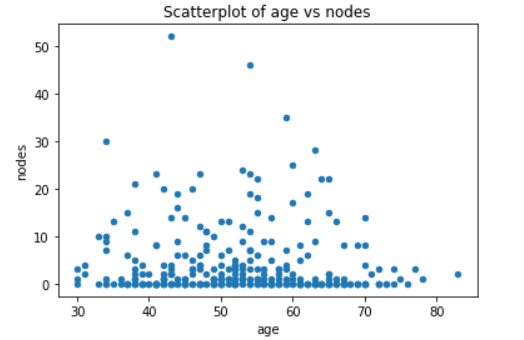
From the above code we can conclude that out of 306 patients (that we got from the code 'haberdata.shape'), 225 patients survived 5 years or longer and sadly 81 patients died within 5 years.

**Status will be the class label.**

1. **Jumping over to plots which will give us more details about this data.**
2. **Scatter plot: - It is a type of plot which will be in a scatter format. It is mainly between 2 features. Here we will plot nodes Vs age and see if there is any linearity.**

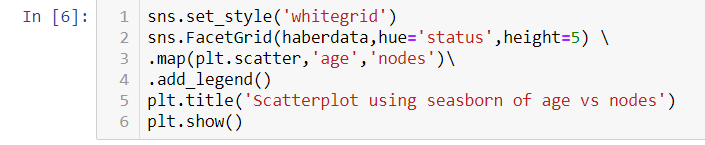
****

Output:

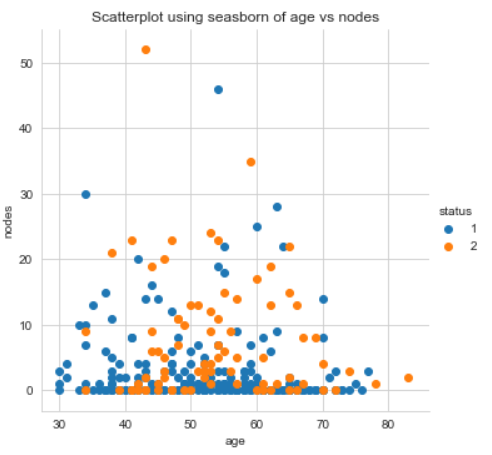


The above plot does not give a better understanding as the colour is same and a lot of points are overlapping.

1. As matplotlib did not give a clear output, we will seaborn module to find whether we can get some insights.



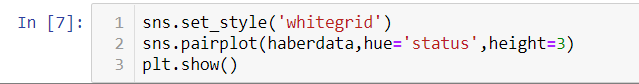
Output:



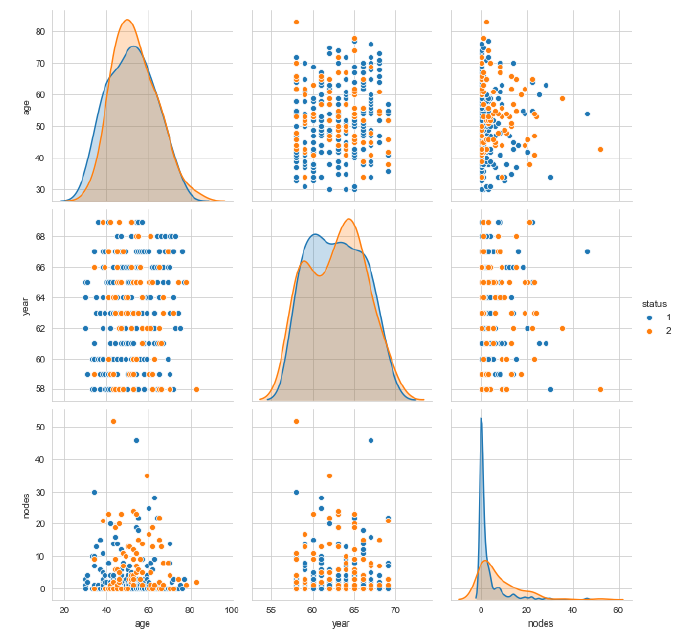
Here we used seaborn for 2-D scatter plotting. Here blue and orange dots represent the survival status of the patients. blue represents the patient survived 5 years or longer and orange dot represents the patient died within 5 year.

As we can see, we cannot point out good and solid information out of this plot. Here we used 3 features for plotting.

1. Pair plots:- Used to see the behaviour of all the features present in the dataset also we get to see the [PDF](https://en.wikipedia.org/wiki/Probability_density_function) representation.



Output:

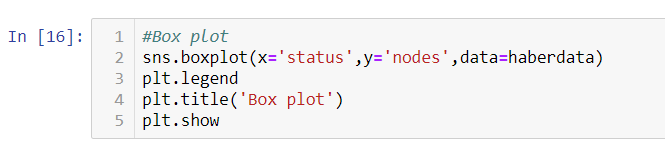


From the above plot we can get some interesting facts. We can say that plot 6 is readable compared to the other two but certainly we cannot make any concrete observations based on this graph.

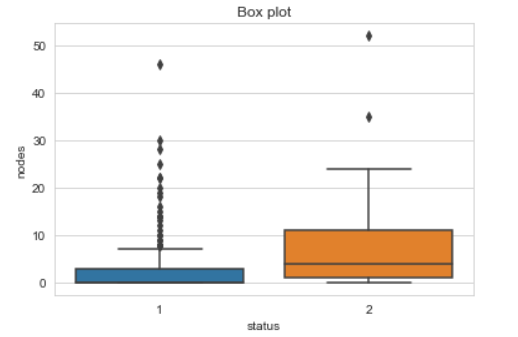
The plot 4, plot 7 and plot 8 are the inverted plots of plot 2, plot 3 and plot 6 respectively.

I have used seaborn as it has good presentation and is also fast.

1. Box-Plots: Box plots tell us the percentile plotting which other plots cant tell easily. It also helps in detection of outliers.

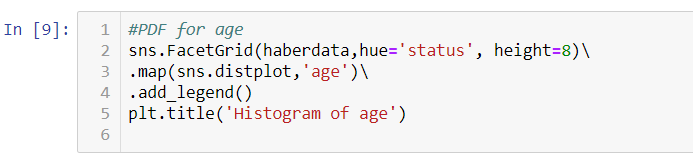


Output:

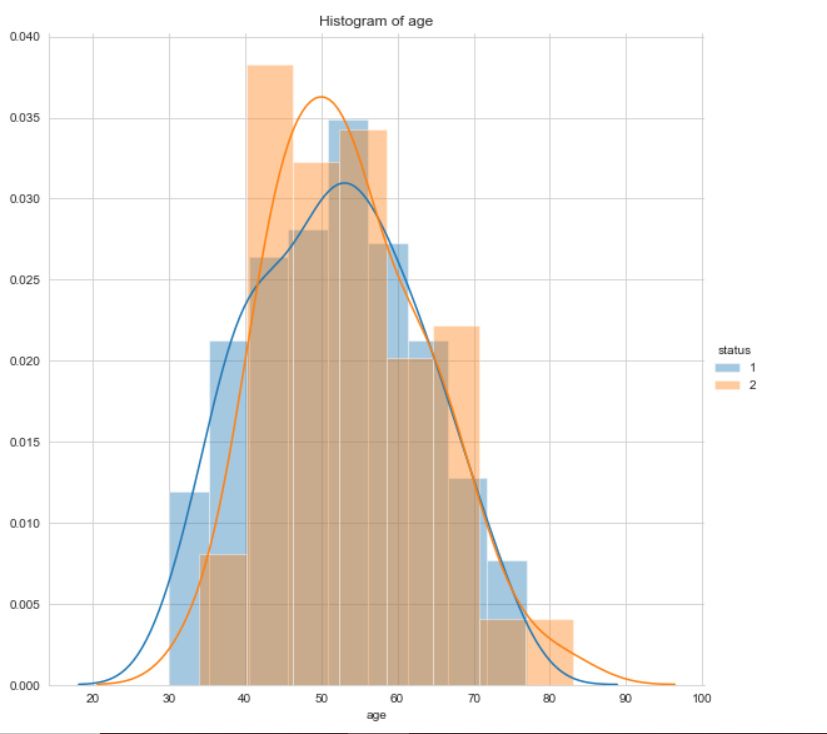


This is still not pleasing to our eyes. Here we can know about the quartile range and the outliers situation.

1. Histogram: Histogram plots are used to depict the distribution of any continuous variable. These types of plots are very popular in statistical analysis.



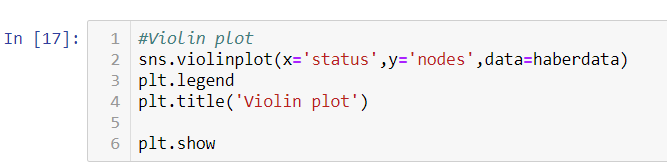
Output:



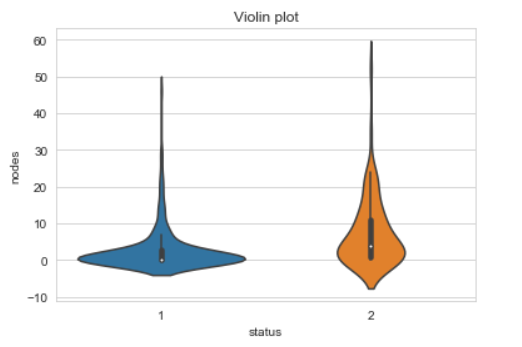
The line represents the PDF which is the Probability Density Function.

By plotting Histogram and PDF of different features we might get good insights regarding the Dataset.

1. **Violin plots-** It is a extension of box plots in this the kernel density plot is also plotted with box plots.

****

Output:



Some observation can be carried out likecategory 1, we can see at 0 nodes it is highly elevated and the whiskers ranges from 0 to 8 .

Conclusion:

These are some basic plots used in EDA. It is always important to read and understand what the plot is saying. It is never good to skip EDA for a machine learning project.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Please refer my code in [github](https://github.com/Vihaanshah29/Haberman_dataset_EDA) for reference.