**Problem Definition:**

The **Titianic dataset** consists of around 891 which shows the number of people were on the ship when it started its final journey before getting hit by a glacier in the mid of the sea.

Here our aim is to identify the number of people survived during that mishap.we have many sample of dataset in kaggle for self analysis even you can go for it.

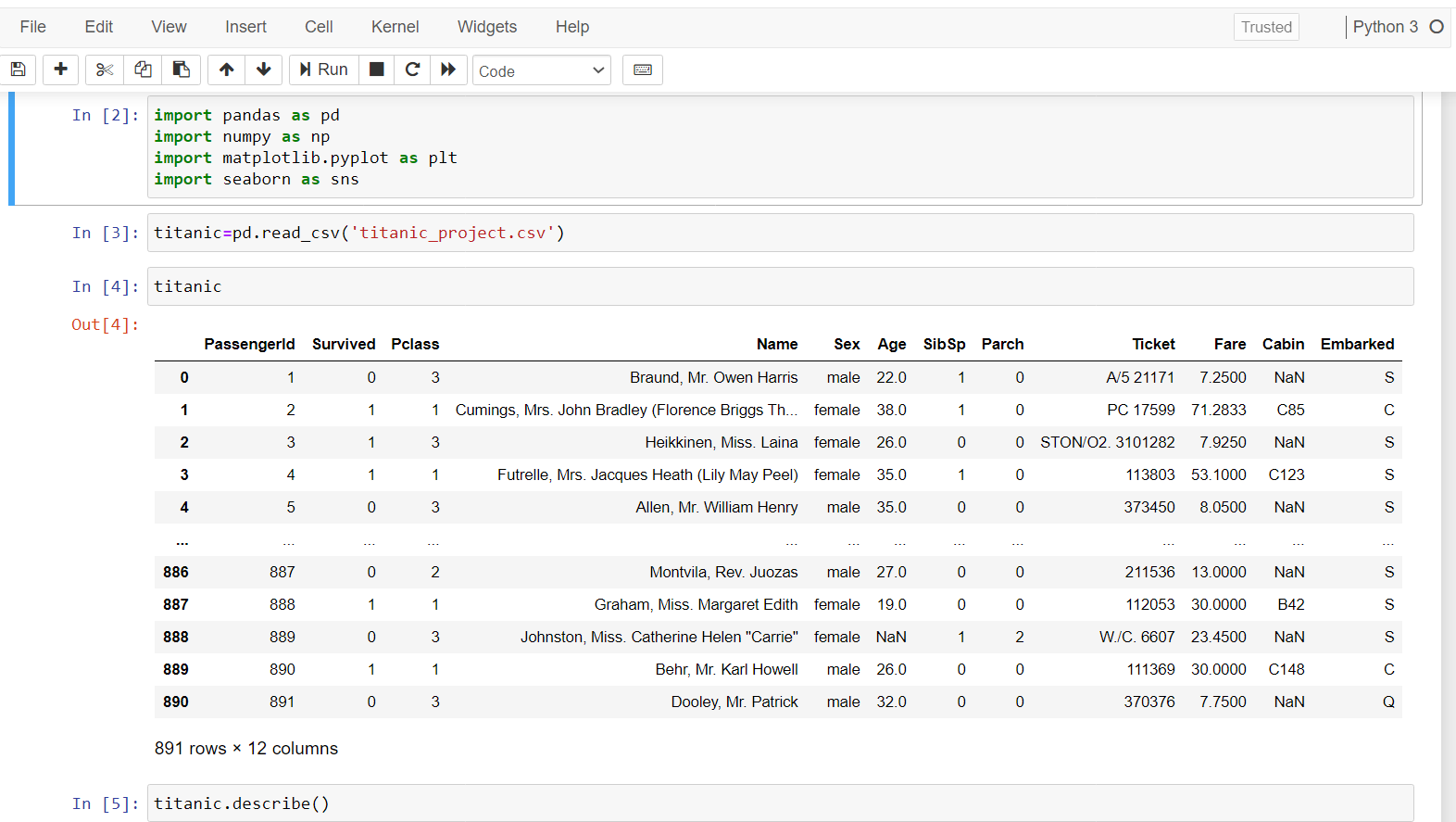
**Data Analysis**

For Analysis i am using **python jupyter notebook**, As it user friendly and can handle the large datsets with ease. As per my knowledge it is one of the best open source tool which is available in market for free of cost.Started my career in data science couple of months back with **DataTrained institute**, I have really enjoyed my journey till now, Will be sharing you all my experience .

Here we begin **Exploring the Data Analysis** partGo to the kaggle or github repository , Download the required dataset , Save it using “**.csv**” file so that it will be easy to load in jupyter notebook.

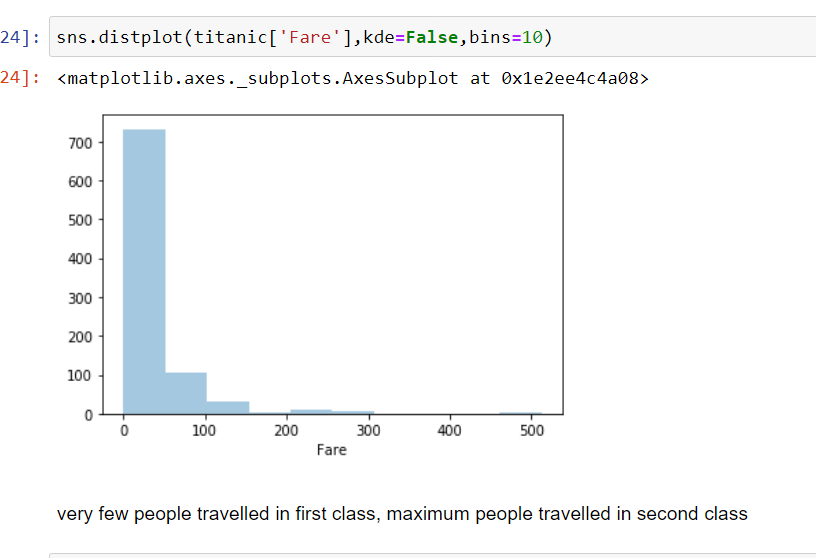
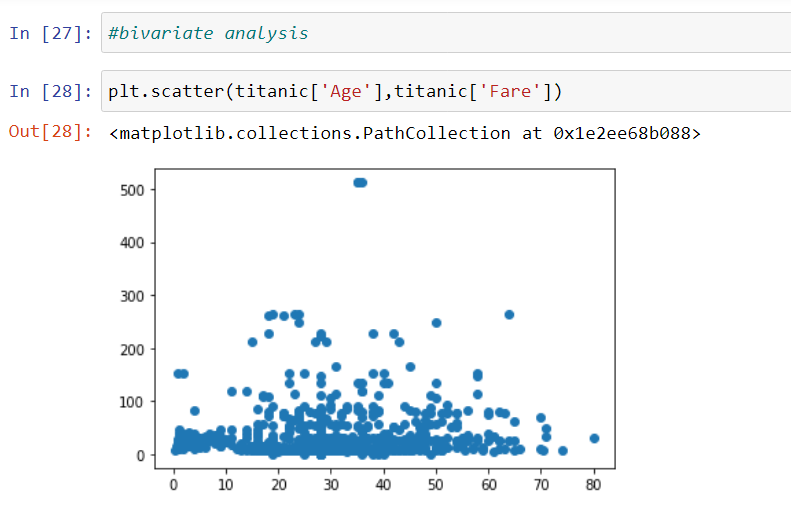
Now we are ready to use the jupyter notebook.

Using the function **read\_csv** load the data and start exploring it. Using **Head** and **Tail** we can see the top and bottom 5 rows of the dataset.



I am having describe and info function for the statistic analysis and then we will go for **Uni variate** and **Bi Variate** analysis.

**Uni variate analysis** : we will be using only one input variable for analysis here are the few Screenshots for the reference.

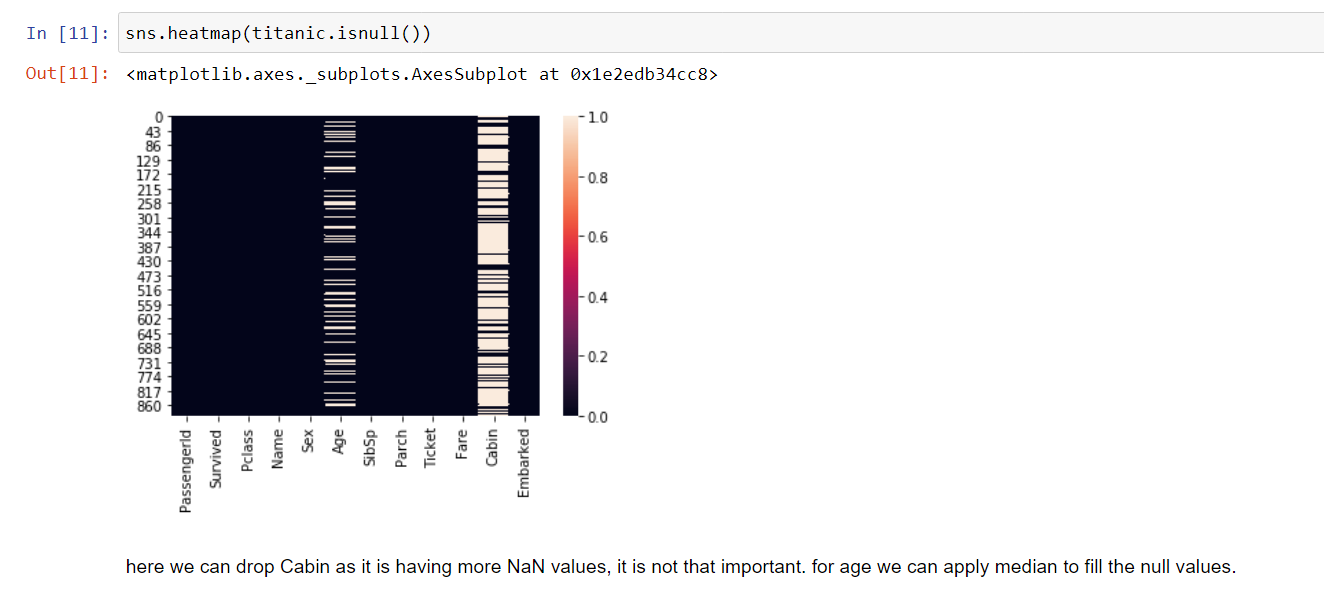
**Bi variate** analysis is used for two input variables analysis, check with the below Screenshot for more information

After these two we analysis we will get idea how our data is distributed and we can take many insights from it to make a good model.

Checking for columns and null values present in the data.if we have null values then we need to fill it up using **“most frequent” or “mean”, “median” or “mode”** based on the dataset.

**Countplot** will help you out in giving the proper info of how many males or females were present or else the number of people survived and died.it is useful as it gives a valuable information with a small line of code.

Using **heatmap** we can check the null values in visualization.



We have **box plot** to detect the outliers and it is very much useful in showing the data.

For data such as age, marks we use **Distplot**, it clearly shows how the data is distributed and we can treat the data if it is **right skewed** or **left skewed** using the skewness methods available .

From the data analysis it is clear that we are having outliers hence we are treating them using **Z-score** method. In this method we have a range of -3 to +3 , apart from these all the other values >3 will be removed as an outliers.

After this method we can check the number of rows removed using this method if it has more than 10% of data loss then we should got with **IQR ( Inter Quartile Range )** or else some other outliers removal method but in this dataset it is iblt 5.6% so we can proceed ahead with Zscore.

**EDA Concluding remarks**

From the EDA we can conclude that the data shown above is perfectly analyzed and hence we now have a good Idea on the Exploratory data analysis part before moving it to the Model Building. With all the graphs above we can get an idea of how effective an EDA should be to make sure you get a proper analysis on the current dataset.

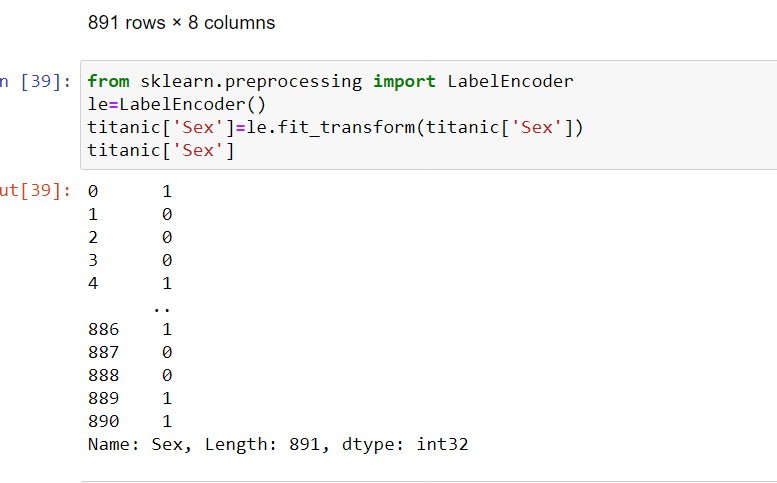
**Pre-Processing Pipelines**

Here we are supposed to check data for duplicates , **Standard Scaler** format and cleaning the data and feature engineering all comes into preprocessing pipeline methods

By using **Standard Scaler** we are bringing or scaling the data into one format, so that it can be passed into a **ML model**.

**Feature Engineering** has a big part to play as the name itself suggest that we need to do the engineering with the data, hence in few datasets we will be getting time series, hours and minutes, numerical categorical data combined in these type of scenarios we are supposed to split into two columns after that reatin the valid data and delete the invalid data using **drop function.**

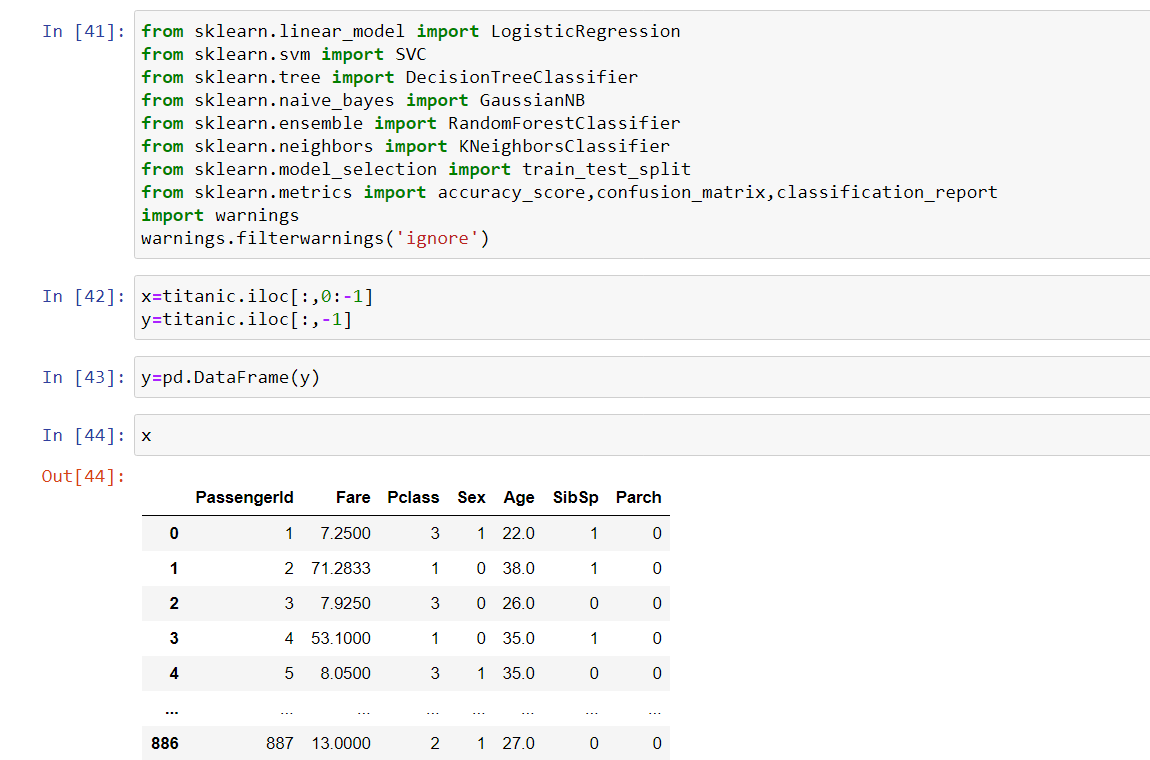
Converting the data of dtypes of object to int or float using Label Encoder.



It converts the data from categorical to numeric in the form of 0,1,2 based on the items available in dataset, so that the machine can learn according to that .

**Building machine learning models:**

we have scaled the data properly and now using models we will train our data.Now need to separate the data using x and y variables, x contains variables and y contains target variable.



Can do these using Loc. Iloc or else drop method. Based upon our resources.

Once the x and y variables are assigned then need to import **traintest split** , **logistic regression**, **GaussianNB** **classifier**,**Decision classifier** etc etc, which ever you feel that would fetch you a good result you can try all those methods.

Once you have done , using metrics we have to calculate the accuracy of the model using **Classification report, accuracy score,confusion matrix,** Now using Gridsearchcv find out the best paramaters for the model.

Once done with this you can use ensemble technique such as RandomForestClassifier to boost up the accuracy of the model.

Once done with all these now we have to save the model using Joblib or pickle, these two are the external libraries used for saving the model.



Now you can call the saved model any time in future for testing the data and using load you can load and start prediction.

**Conclusion:**

Here in Titanic project we got the accuracy score of 82.8% which is very good

And hence by comparing both the test data and predict data , we can compare the result ,as it is a machine we can expect some minute errors. Thus it is a very good model for the prediction of dataset.