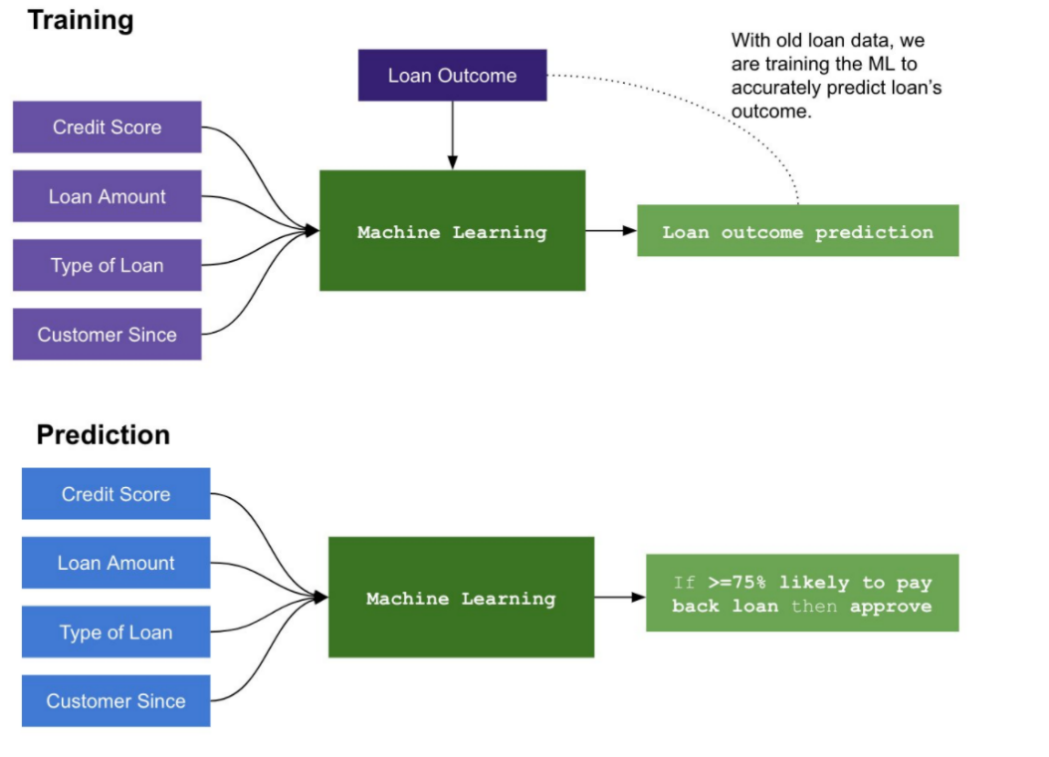
### **Prediction of Loan Approval using Machine Learning algorithms**



**Abstract: -** Loan approval is a very important process for banking organizations. Loan-taking from banks and financial institutes has become a very common phenomenon. Every day a large number of people are applying and making applications for loans. But all applications for a loan are not approved by the bank because most people are not eligible for a loan it depends on people's credit score and transections history to the bank or any other financial company. Every year we have seen a large of cases in which people do not repay loan amounts to the banks due to which they suffer huge losses. Recovery of loans is a major contributing parameter in the financial statements of a bank. The bank and financial institute with making a loan approval are immense. This project is to gather loan data from multiple data sources and use various machine learning algorithms on this data to extract important information. So that the model can be used by the organizations in making the right decision to approve or reject the loan request of the customers. In this paper, we have applied various machine algorithms like logistic (LR), Decision Tree (DT), and Random Forest (RF) are applied to predict the loan approval of customers.



1. **Introduction: -**

Now a day’s people rely on bank loans to fulfil their needs. The rate of loan applications increases at a very fast speed in recent years. Almost every bank's core business is loan distribution. All most of the bank's assets and profit are directly derived from the loans distributed by the bank. Bank’s plays a vital role in a market economy and their successful growth. The primary goal in the banking industry is to place their loan in safe hands. The success or failure of an organization largely depends on the industry’s fundamental and ability to evaluate credit risk. Before giving the credit loan to borrowers, the bank decides whether the borrower is bad (defaulter) or good (non-defaulter), which means the bank checks his credit score and transaction history. The prediction of borrower status i.e., in the future borrower will be defaulter or non-defaulter is a challenging task for any organization or bank. The loan defaulter prediction is a binary classification problem Loan amount; costumers history governs his creditability for receiving the loan. The problem is to classify borrowers as defaulters or non-defaulter. However, developing such a model is a very complicated task due to the increase in demand for loans. This work includes the construction of different machine learning models Credit Risk assessment is a crucial issue faced by Banks nowadays which helps them to evaluate if a loan application can be a defaulter or non-defaulter. The machine learning models and algorithms provide the help to predict the genuine borrower and any organization for loan approval.

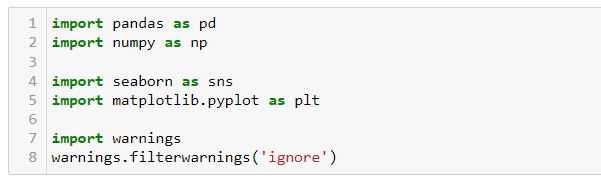
**Problem definition**

 Here we aim to automate the loan eligibility process based on the data provided by the customer while filling up the application. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. We will be using all attributes of the applicant in the machine learning model to predict or evaluate the genuine applicant which is eligible for the loan. The machine learning model predicts the all-relevant factor of the applicant which is analysed by all banks and financial institutes.

**Data Analysis**

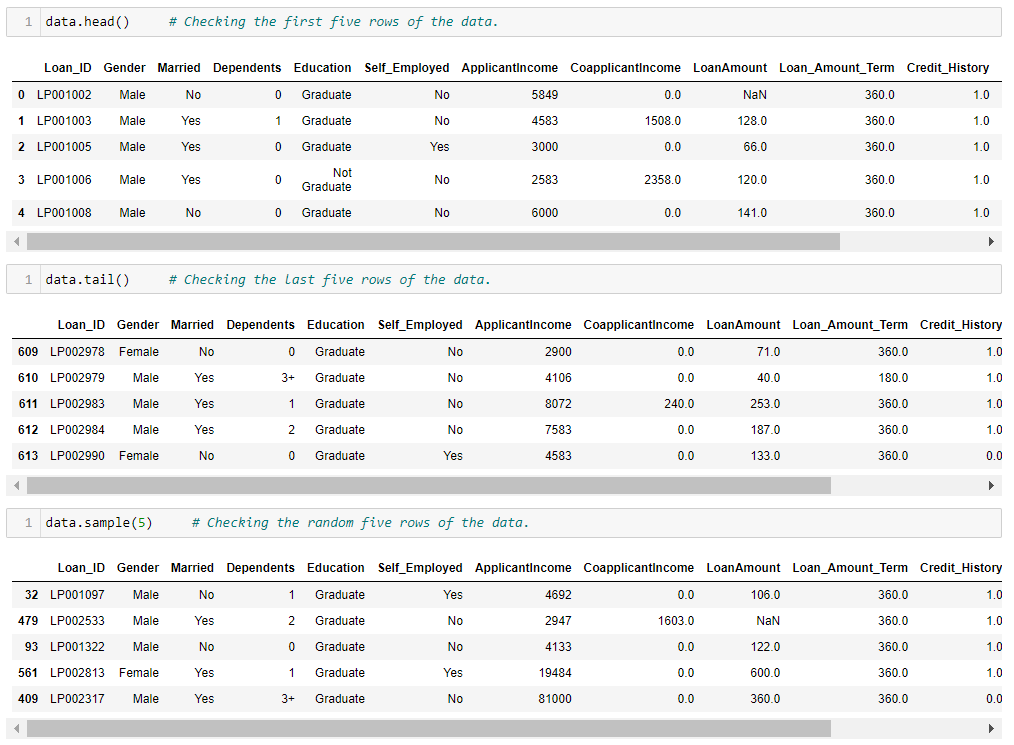
In this section we will analyse and visualize our data using tools available in python for analysis like pandas’ data frame for data processing, NumPy for linear algebra and Matplotlib, Seaborn for visualization. Let’s start with analysis by import all the require library for data analysis.

Importing all the required libraries:



Here we use the Pandas for performing the data [processing task and NumPy for linear algebra and Matplotlib and Seaborn for data visualization and find the inside of the data.

Now our data set is loaded into our data variable, let’s check our data looks like.



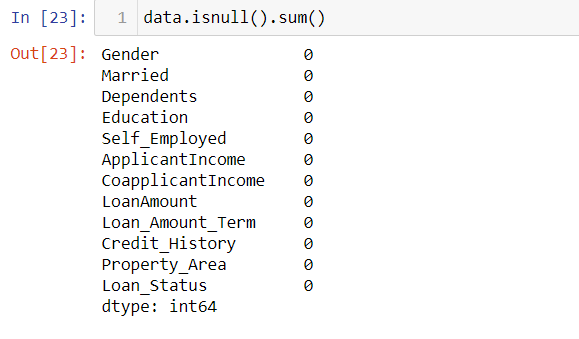
Dataset have total 614 entries.  
It is having 614 Rows and Total 13 columns.

### **Dataset Description:**

Loan ID :- Unique Loan ID  
Gender :- Male/ Female  
Married :- Applicant married (Y/N)  
Dependents :- Number of dependents  
Education :- Applicant Education (Graduate/ Under Graduate)  
Self Employed :- Self-employed (Y/N)  
Applicant Income :- Applicant income  
Co-applicant Income :- Co-applicant income  
Loan Amount :- Loan amount in thousands  
Loan Amount Term :- Term of loan in months  
Credit History :- credit history meets guidelines  
Property Area :- Urban/ Semi Urban/ Rural  
Loan Status :- Loan approved (Y/N)

The above all factor we have used in machine learning model to predict the eligibility of customer for loan. From sample we can see we have columns like Gender, Married, Education, Applicant income, Co-Applicant income, Loan Amount and Loan Status etc. Loan Status is our target column and all other columns are feature columns. We have only two types of values in Loan Status ‘Yes’ and ‘No’, which indicates it is classification problem. We can see Loan ID is unique identification number which we can drop during analysis.

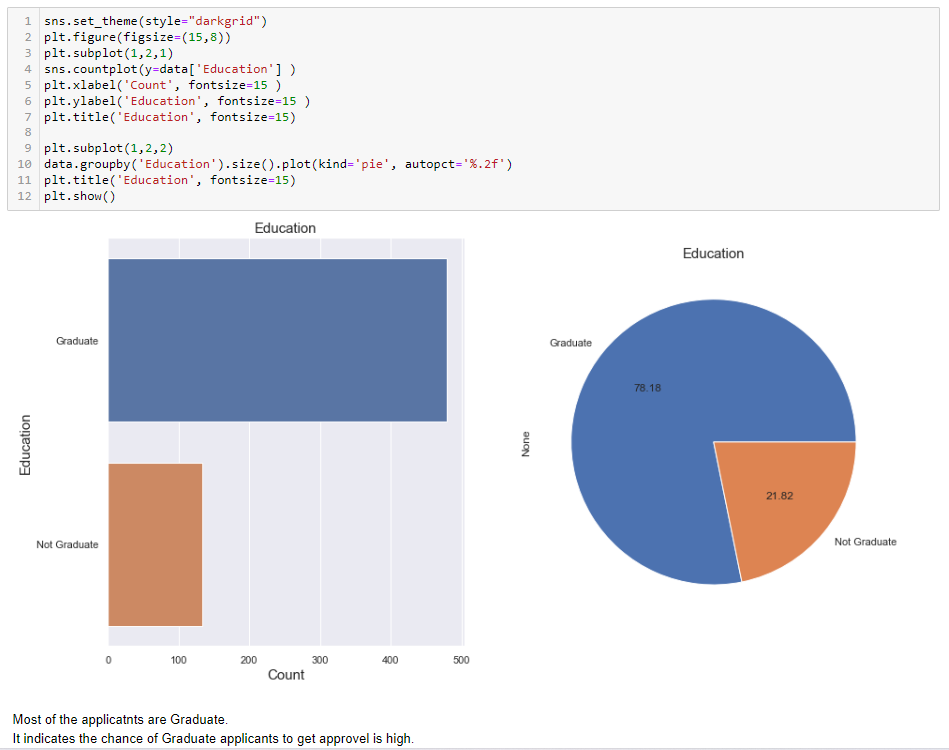
We can see that after filling the all-null value the data set is completed and data to not have any null value. And after that we did Exploratory data analysis.



**Exploratory Data Analysis**

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights and representing the data in the graphical format. Here we have analyzed the various factor which is required for the loan Process in every bank and financial institute.

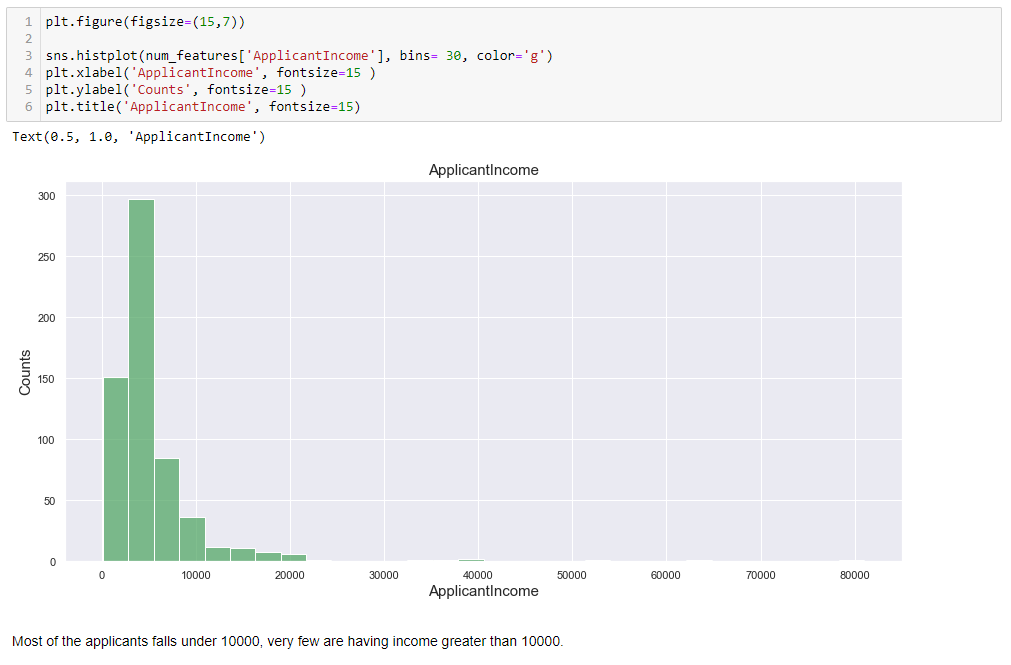
**Education:** Education is a very important role play for loan distribution. From the data, we have analyzed that most of the applicants are educated. It indicates that’s the chance of loan approval is high if those applicants are educated. The below graph shows the data of the educated and non-educated applicant.



**Self-employed:** From the data, we have analyse that the chance to get load approval is high for applicants who are working professionals or those who is in jobs. The below graph shows the data of the self-employed applicant.



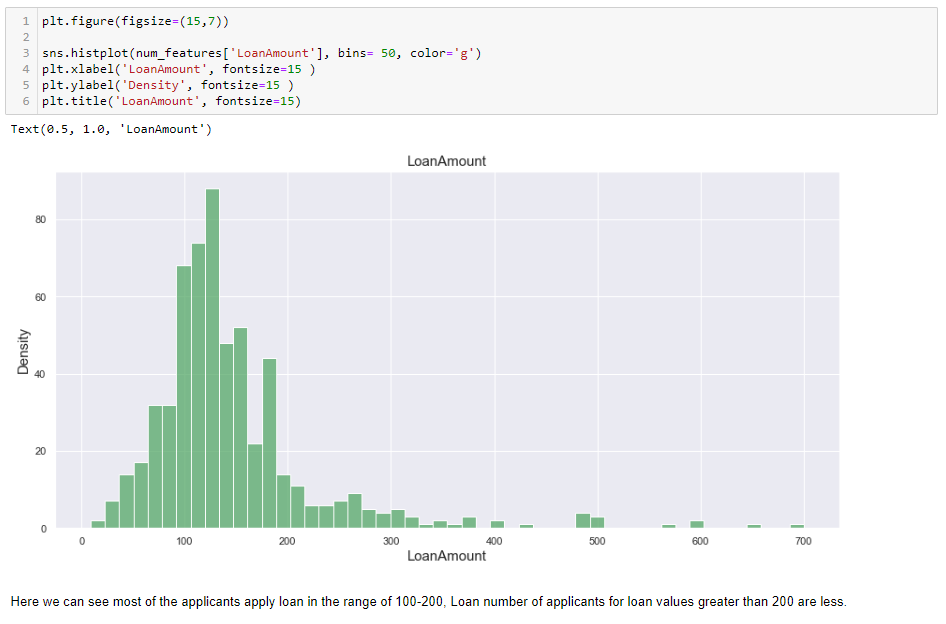
**Applicant income:** From the data we have analyzed that the Most of the applicants fall under 10000, very few are having income greater than 10000. The below graph shows the data of applicant income.



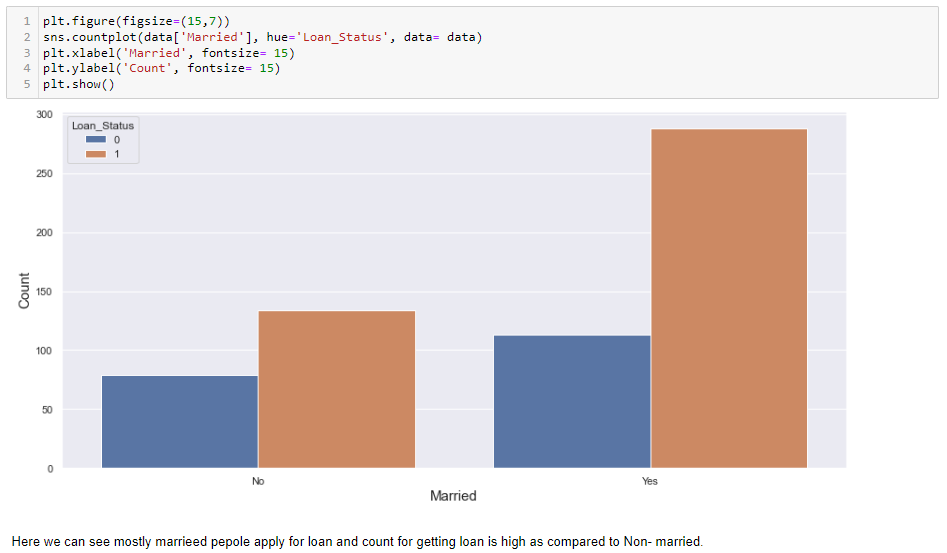
**Credit history:** The credit history of the applicant is the most important factor of any type of loan process. If the applicant has a credit score as per the loan requirement. those applicant loan approval chance is very high.  From the data, we have analyzed that the chance of those applicants is high whose credit history and record are good. The below graph shows the credit history of the applicant.



**Loan Amount:** From the data we have analyzed that most of the applicants apply for a loan in the range of 100-200, Loan number of applicants for loan values greater than 200 are less. The below graph shows the loan amount range of applicants.



**Married:** From the data, we have analyzed the most married people applying for the loan, and the count for getting a loan is high as compared to non-married. The below graph shows the data of married and non-married applicants who are applying for the loan.



## EDA Summary

Gender: For male applicant the chance of loan approval is high.  
Married: Mostly the male applicant is applying for loan so the chance of married applicants is high of loan approval.  
Dependents: From the data we analyse when higher the number of dependent people on applicant then lower the chance of loan approval.

Education: Chances for approval increase with higher education.  
Self Employed: From the data we have analyse that the chance to get load approval is high for applicant who are working professional or those are in jobs.

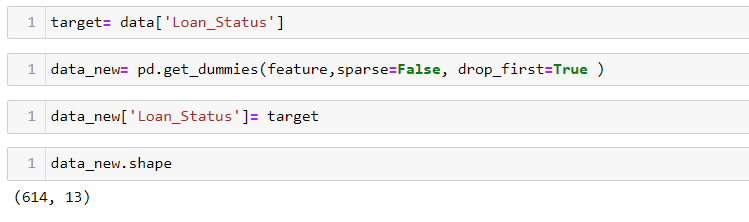
Co-applicant Income: Co-applicant Income increase the chance of loan approval increases. Loan Amount: For higher loan amount, the chances of approval are low but it also depends on the income of applicant.  
Loan Amount Term: For specific term the chances are high, but it is not linear it also depends on the income of applicant.

Credit History: Applicants having good credit history have high chance of loan approval.

Property Area: Semi Urban Area have high chance of loan approval.

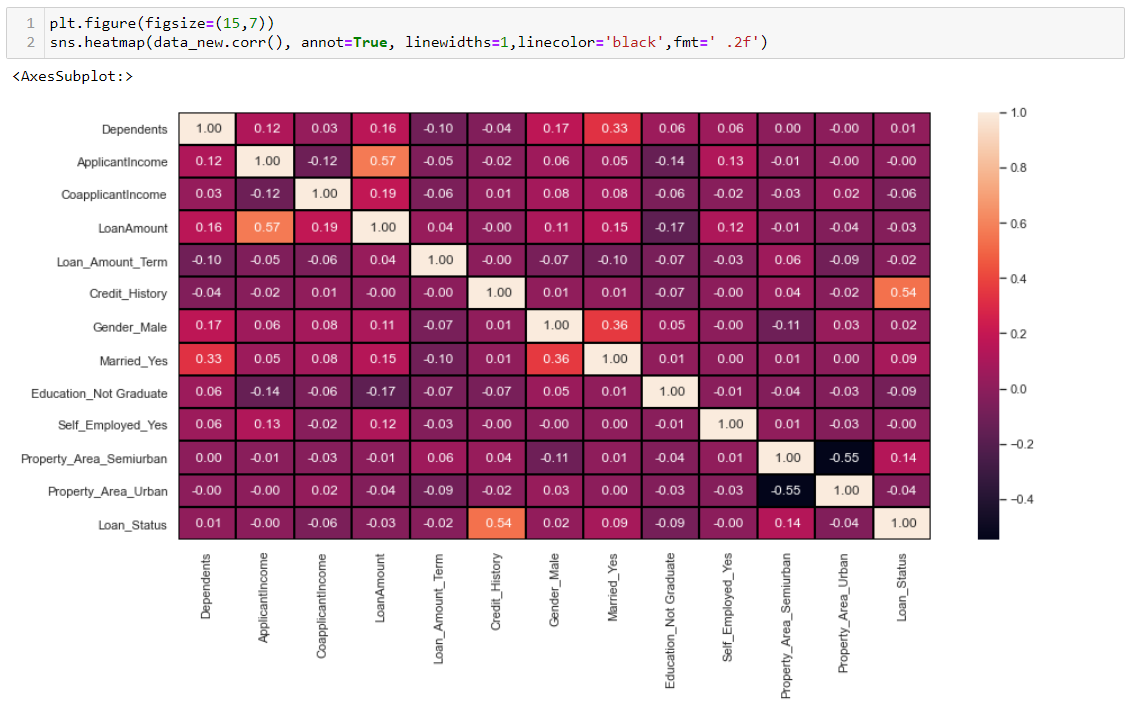
### **Data Pre-processing**

In this stage, we will prepare our data for model building. we will change the categorical data into numerical data using encoders Then we will check the outliers and skewness for continuous columns. At the scaling of data etc.



**Let’s check the correlation:**

Correlation itself is a mathematical technique to examine a relationship between two quantitative variables from the heatmap the data is represent the correlation of every attribute to each other and the main use of heatmap is to compare the relation of all the attributes with the Loan Status (target). The below heatmap represent the correlation of all attributes to each other.



From the heatmap the data shows the attribute like credit history, property area and married are shows the positive correlation with Loan-Status and all others attributes are having negative correlation with Loan-Status.

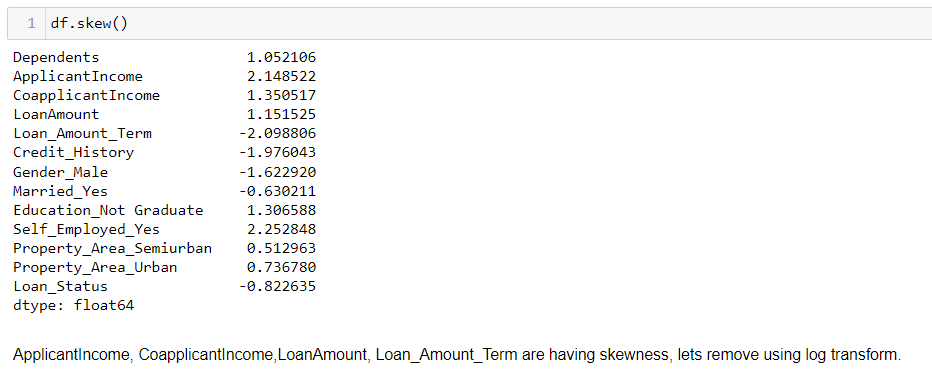
**Outliers: -** Outlier Analysis is a process that involves identifying the anomalous observation in the dataset. Outliers are caused due to incorrect entry or computational errors. our data is present the outliers, we have found some attributes having the outliers present. The attribute like applicant income, co-applicant income, Loan amount, and Loan amount Term are present the outliers.



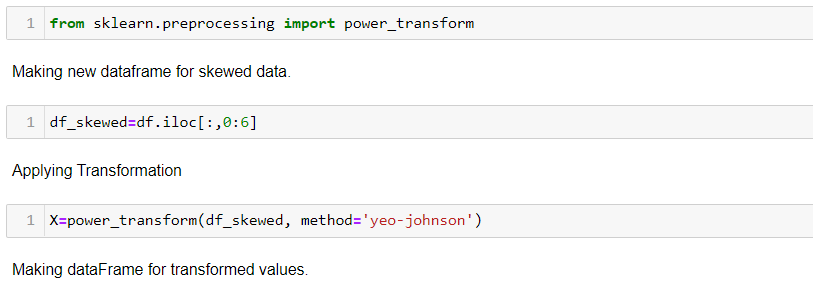
The above all graph have outliers, for further analysis we have to remove the all outliers from the data set and after that we have to check the skewness of data set.

**Skewness**

Skewness is a measure of the symmetry of a distribution. The below represent the skewness of dataset.



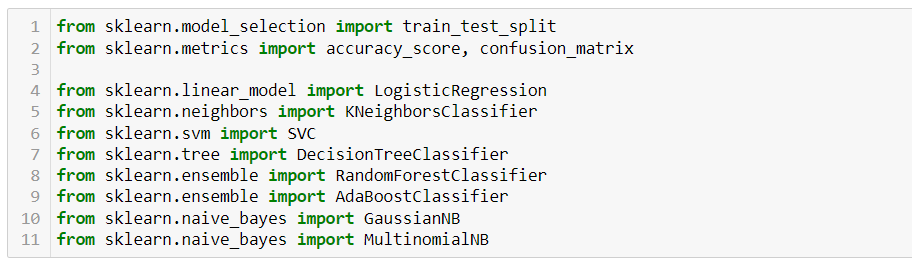
From the data we have observed that the Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term are having skewness, for further analysis of data, firstly have to remove the skewness of data so we have used the log transform method to remove the skewness. the below command is used to remove the skewness of data.



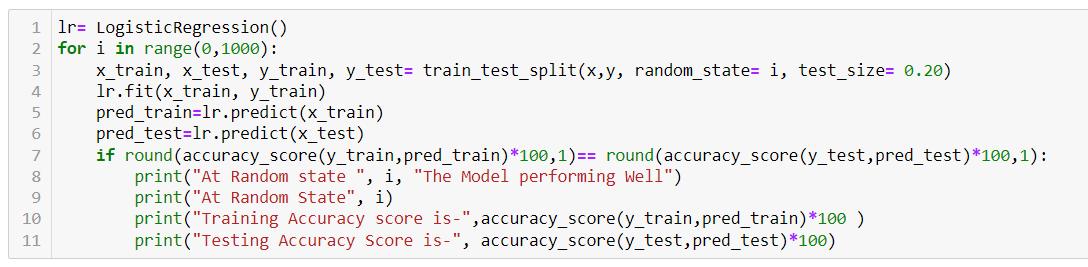
# Building Machine Learning Models:

Our Target column is having two kinds of values either yes or no which we encoded using a label encoder. So, it's classification-based problem. In the model development stage, we will start with the basic classification model and end up with complex models like Decision Tree, Random Forest, and AdaBoost classifier. We will check the accuracy and cross-validation for each model and select the best model which will be having a lesser difference in their cross Val score and accuracy score.

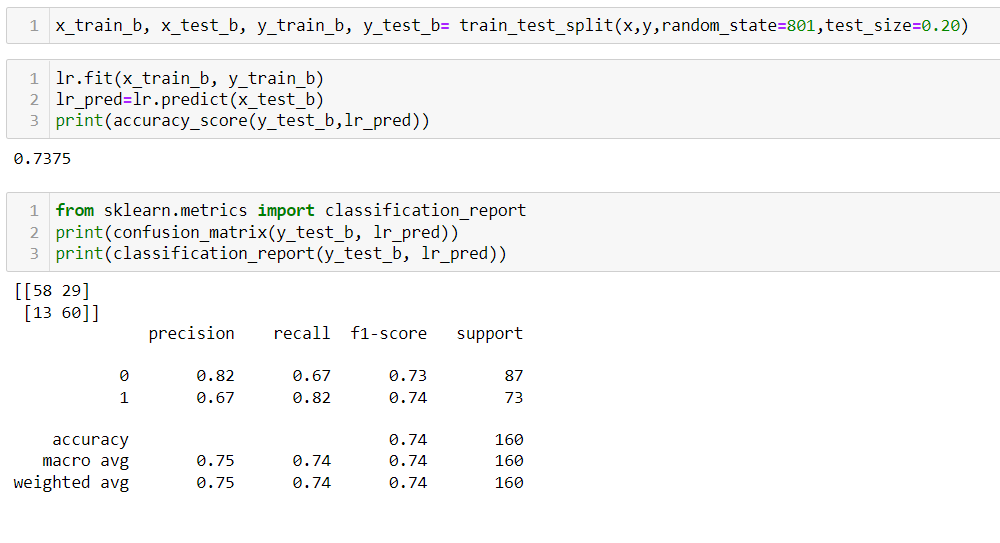
**Importing all the models: -**



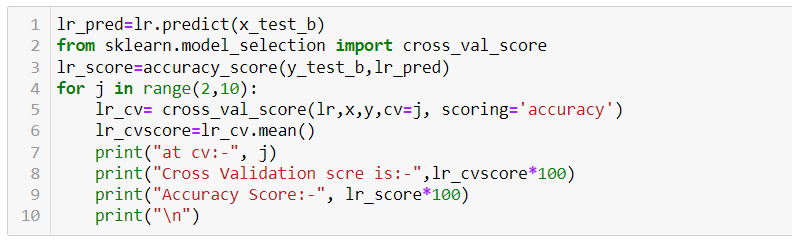
**Choosing best random state: -**



At Random State 801  
Training Accuracy score is- 75.0  
Testing Accuracy Score is- 75.0



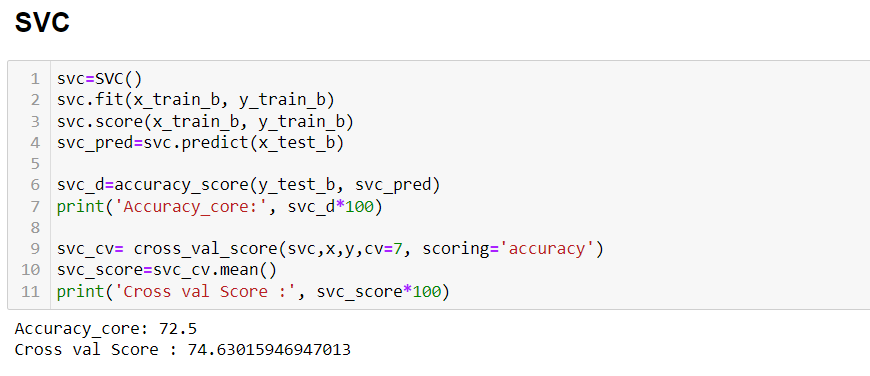
For linear model we have accuracy score 73.75 percent. We will check Cross validation score.



At CV= 7, Cross Val score 73.24062368310156

and Accuracy Score: - 73.75

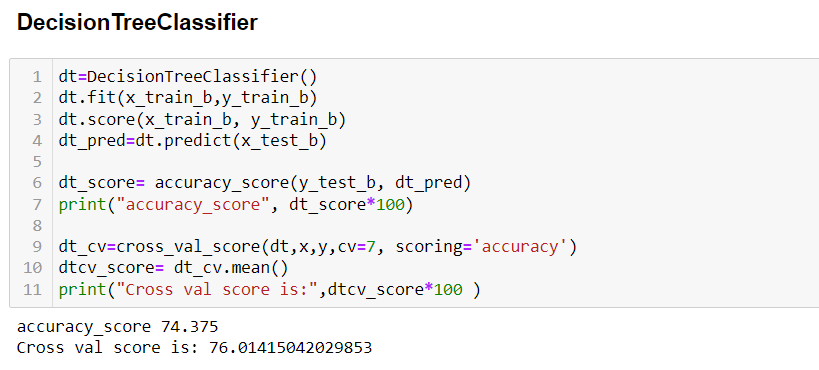
At cv-7 Cross Val score and Accuracy Score are having difference, Model is not performing well. But still need improvement, Let’s try another model



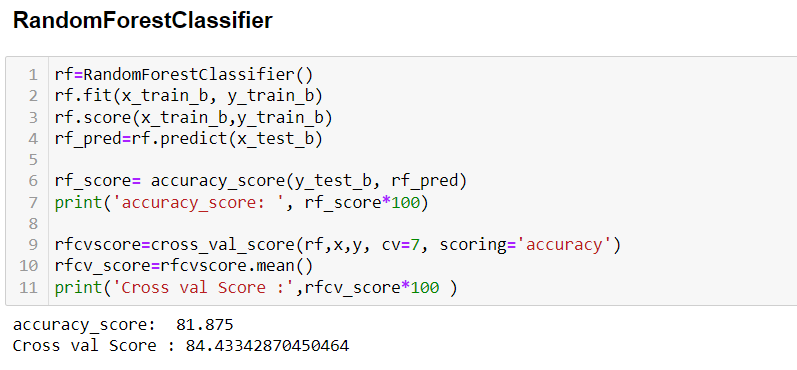
**support vector classification:** From the support vector classification model, we have got the accuracy core is 72.5 and cross Val score is74.63 the difference is less as compare to logistic regression. Let’s try the other model.



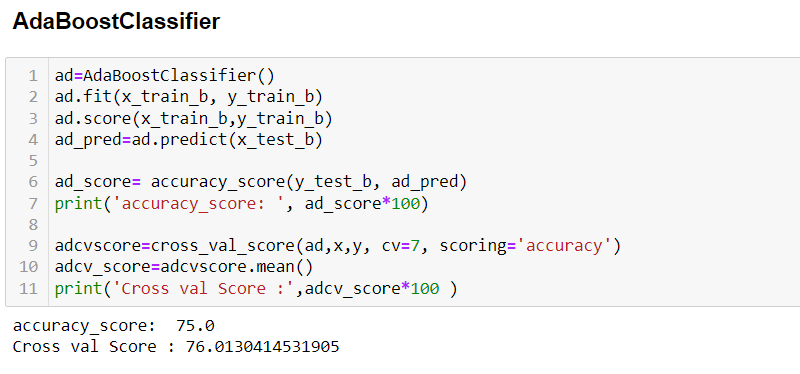
**KNeighborsClassifier**: From the KNeighborsClassifier model, we have got the accuracy core is 70.625 and cross Val score is 76.00 again, the difference is less. Let’s try the other model.



**Decision Tree Classifier**: From the Decision Tree Classifier model, we have got the accuracy core is 74.375 and cross Val score is 76.014 the Decision Tree Classifier model shows the major difference between the accuracy score and Cross Val score so that the model performing is bad as compare the other model.



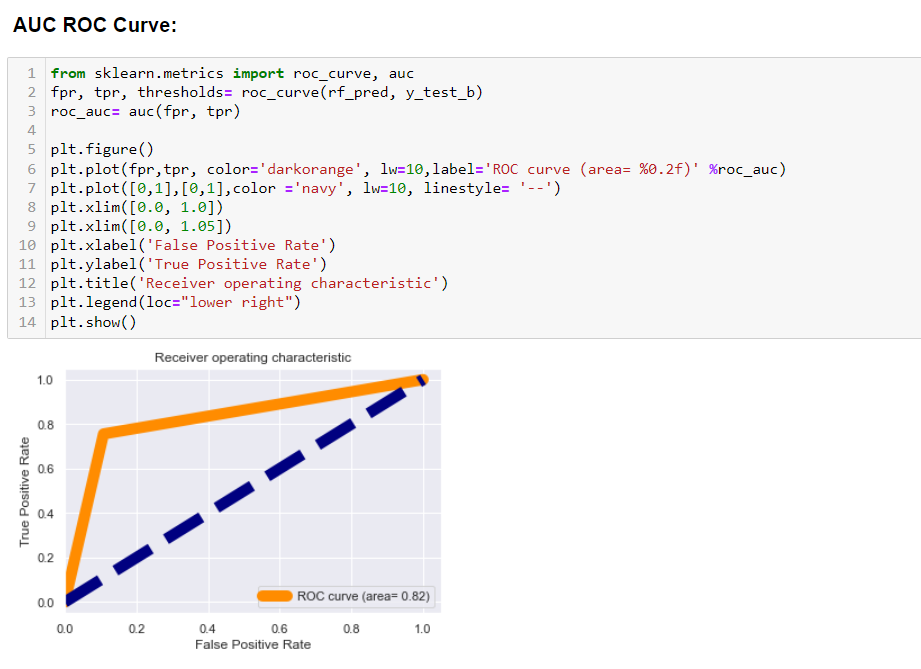
**Random Forest Classifier**: From the Random Forest Classifier model, we have got the accuracy core is 81.875 and cross Val score is 84.433 Both the accuracy score and cross Val score is having very less difference and the model perform is good as compare to other model



**AdaBoostClassifier**: From the AdaBoostClassifier model, we have got the accuracy core is 75.00 and cross Val score is 76013 . the AdaBoostClassifier model is performing very bad. Still the Random Forest Classifier is best.

# Concluding Remarks

This paper applied a machine learning model in the prediction of loan approval. We applied different machine learning algorithms to the dataset to determine which algorithms or models are the best fit for studying bank credit datasets. We also determined the most important features that influence the creditworthiness of customers. After performing all models, we have found that the Random Forest Classifier model performed well with an accuracy score of 81.875 and a cross Val score of 84.433. before saving the best model, we have drawn the AOC-ROC curve.



**Saving best model:** 

We have saved best model and at the end we have represent the both original and predict value shown in the table.