**INTRODUCTION**

The banking system has an important role in the modern economy.

Loans are the core business of banks (in short it’s a money spin). The profits generated comes mainly from loan’s interest. Banks provide loans after doing proper background verification of the customer applying for loan.

To fulfil monetary needs people apply for loans from banks. So banks is like a support to all the needy people out there.

The interest and fees from loans are a primary source of revenue for many banks, as well as some retailers through the use of credit facilities and credit cards.

Now lots of startup companies has came which gives loan with less paperwork involved and verification. So there is a stiff competition in this business.

**PROBLEM STATEMENT**

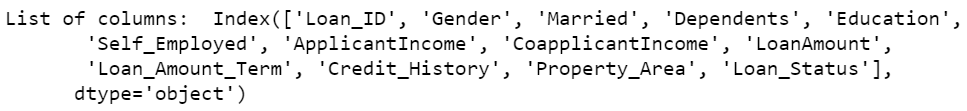
We have to build a model that can predict whether the loan of an applicant will be approved or not on the basis of the details provided in the dataset.

This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

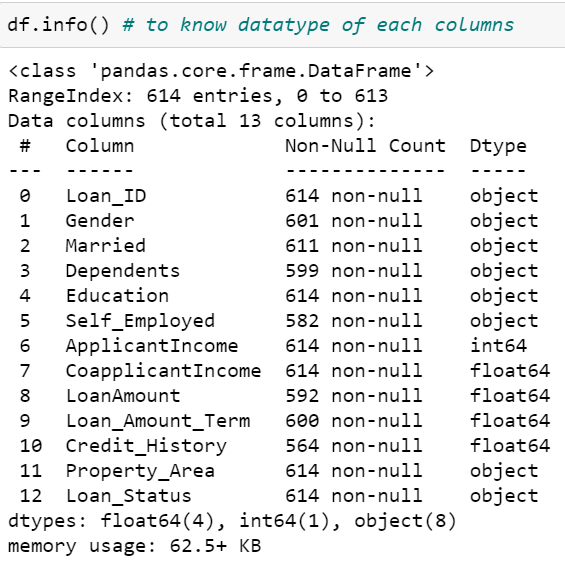
**DATA ANALYSIS**

The dataset provided contains 614 rows and 13 columns.  
 It has 13 attributes (12 features and 1 target variable).

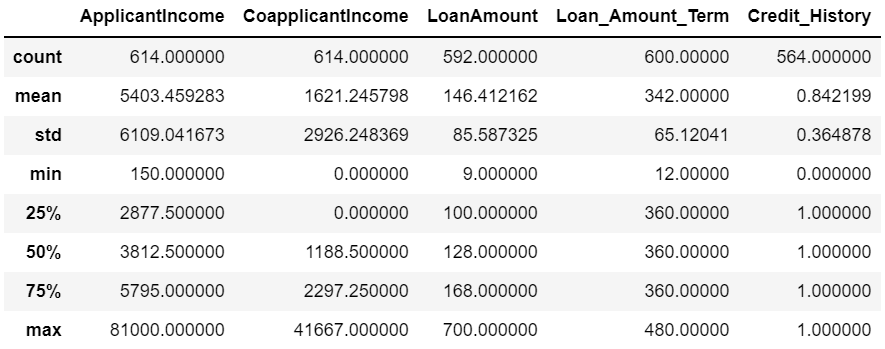
C:\Users\Nitish\OneDrive\Desktop\loan1\Screenshot 2021-09-02 112351.png



Datatypes of each columns

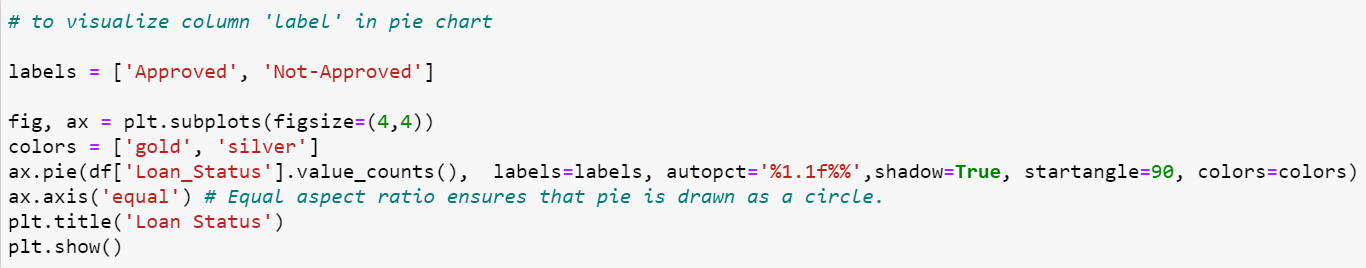


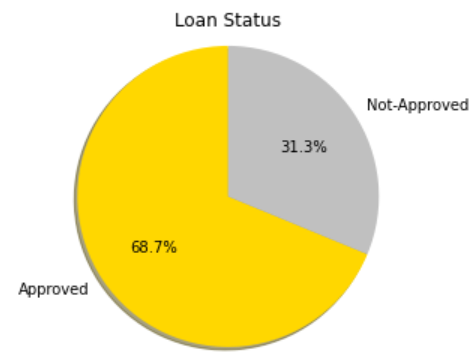
Using df.describe() function to get high understanding of dataset or to get overview/stats of the dataset



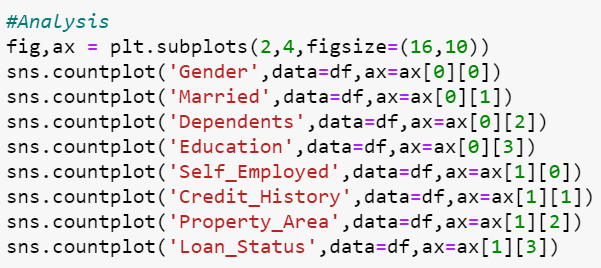
* The count is not same throughout, so there will be missing values.
* The Applicant Income and Co-Applicant Income standard deviation value is more than their mean value, so data of those two are messed up.

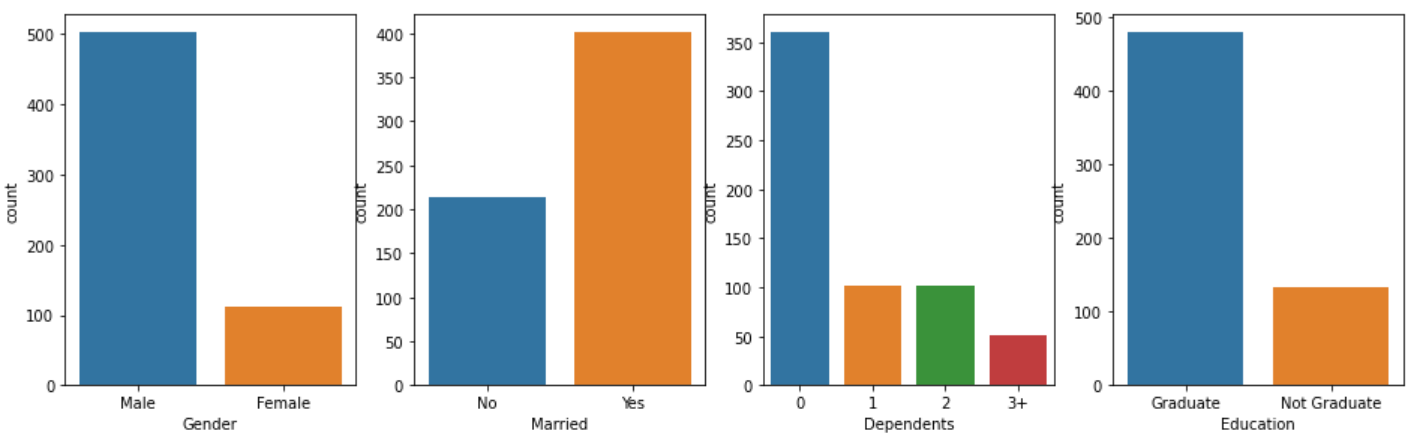
**EXPLORATORY DATA ANALYSIS (EDA)**

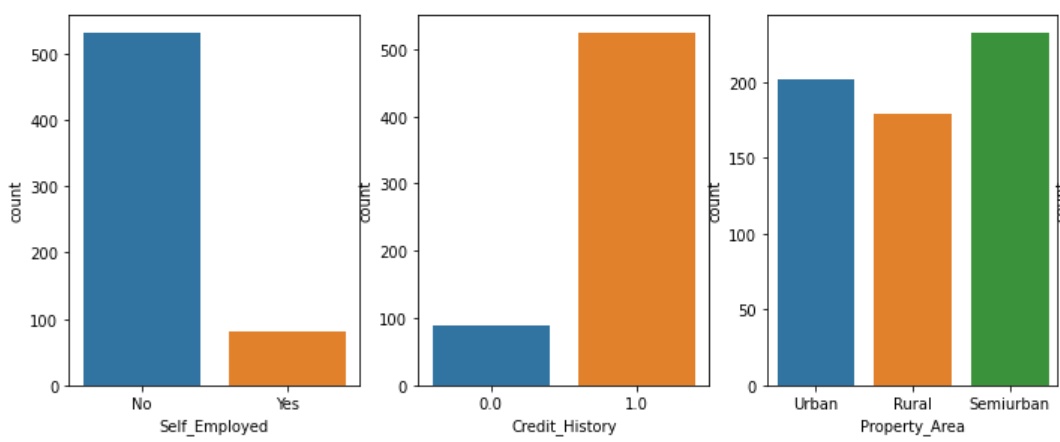
****



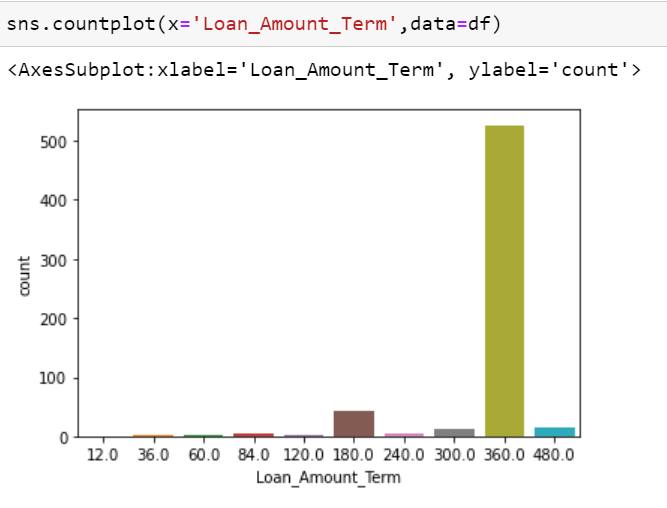
Its’s a imbalanced dataset as we can see that the customer whose loan is approved is 68.7% whereas customers whose loan is not approved is 31.3%.



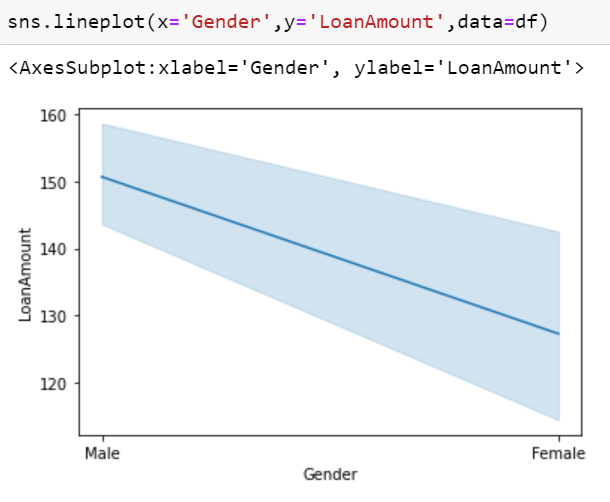




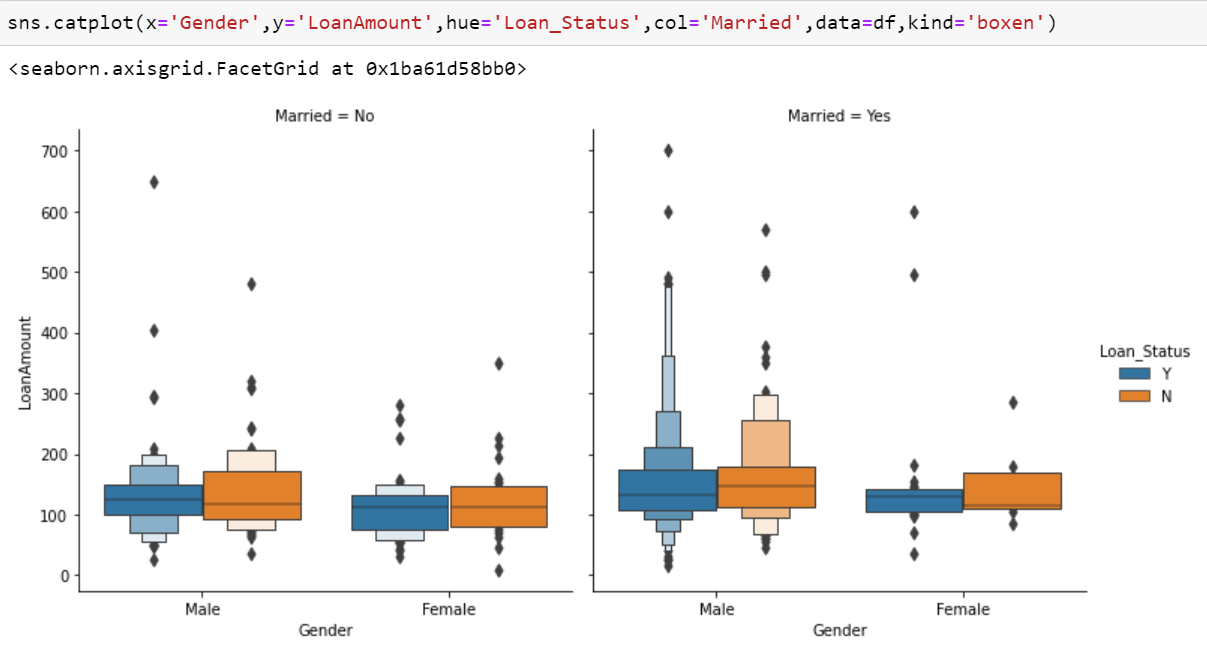
* Number of male applicants applying for loan is more than female
* Married applicants are approximately double in number to those Unmarried applicants applying for loan
* Dependants with number zero(0) is maximum
* Count of graduate applicants is more than non-graduate applicants
* Number of Self employed is less than that of Non-Self employed
* Credit History of maximum applicants is present
* Maximum Applicant properties are in Semi-urban area



* Lots of applicants have applied for loan for 360 months of loan duration.



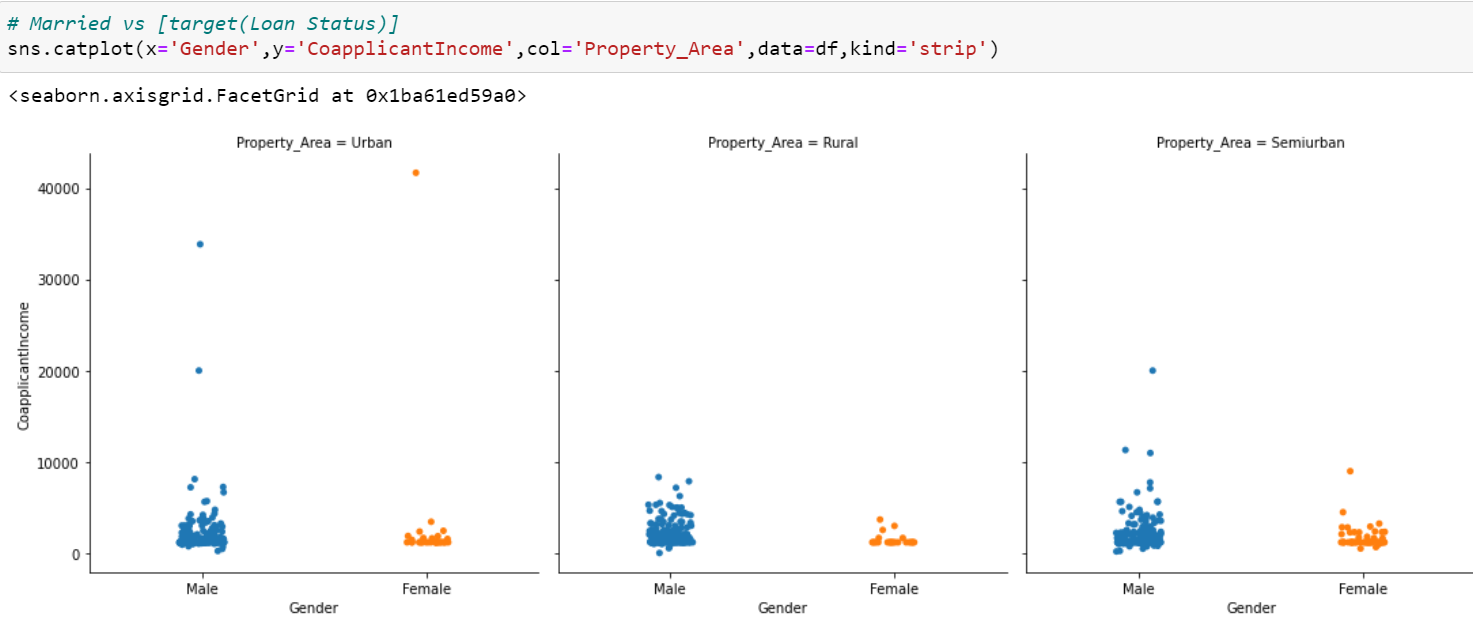
* Male Applicants applying for loan amount is higher than females



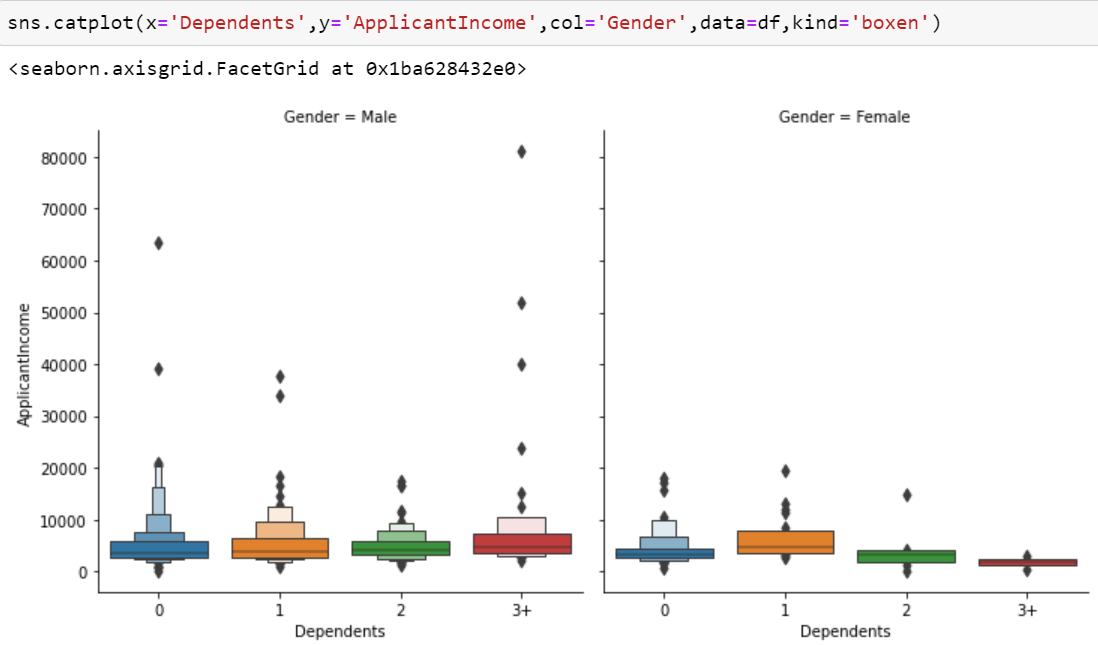
* Applicants those are married have requested for loan more than non-married



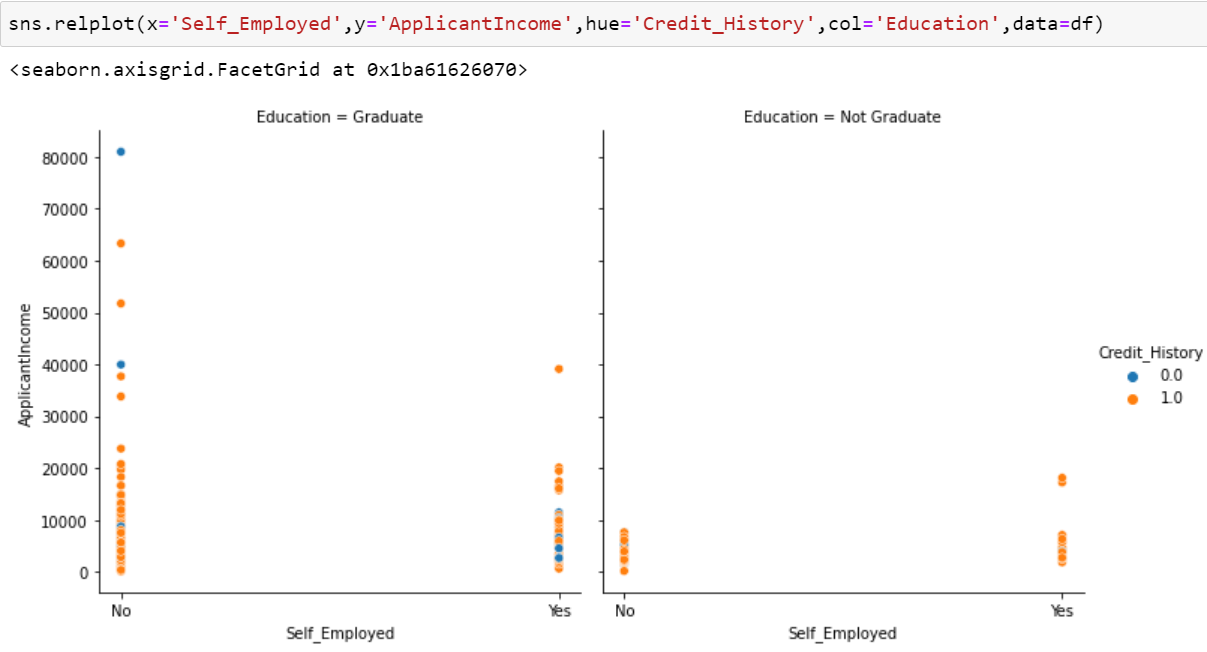
* From semi urban area, maximum number of male and female applicants have requested for loan amount



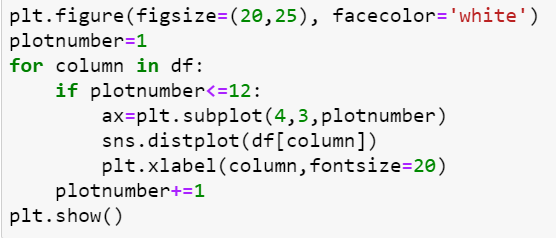
* Male are having co-applicant whose income is more than female co-applicant in all three property areas

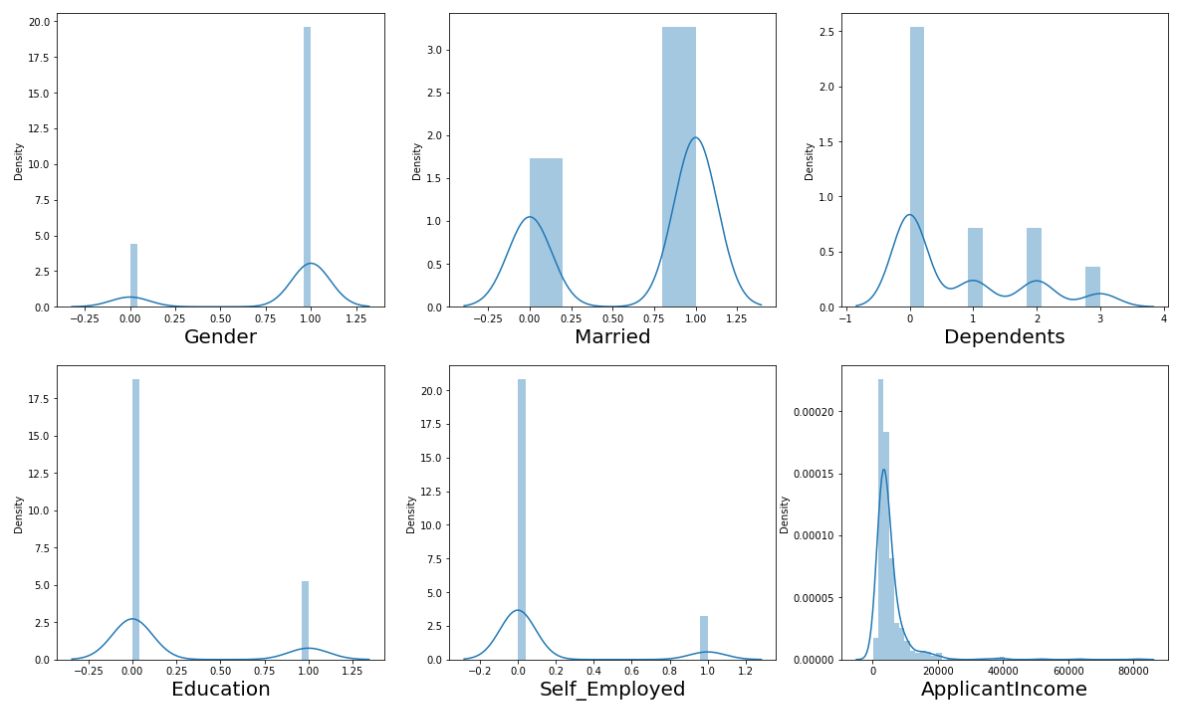


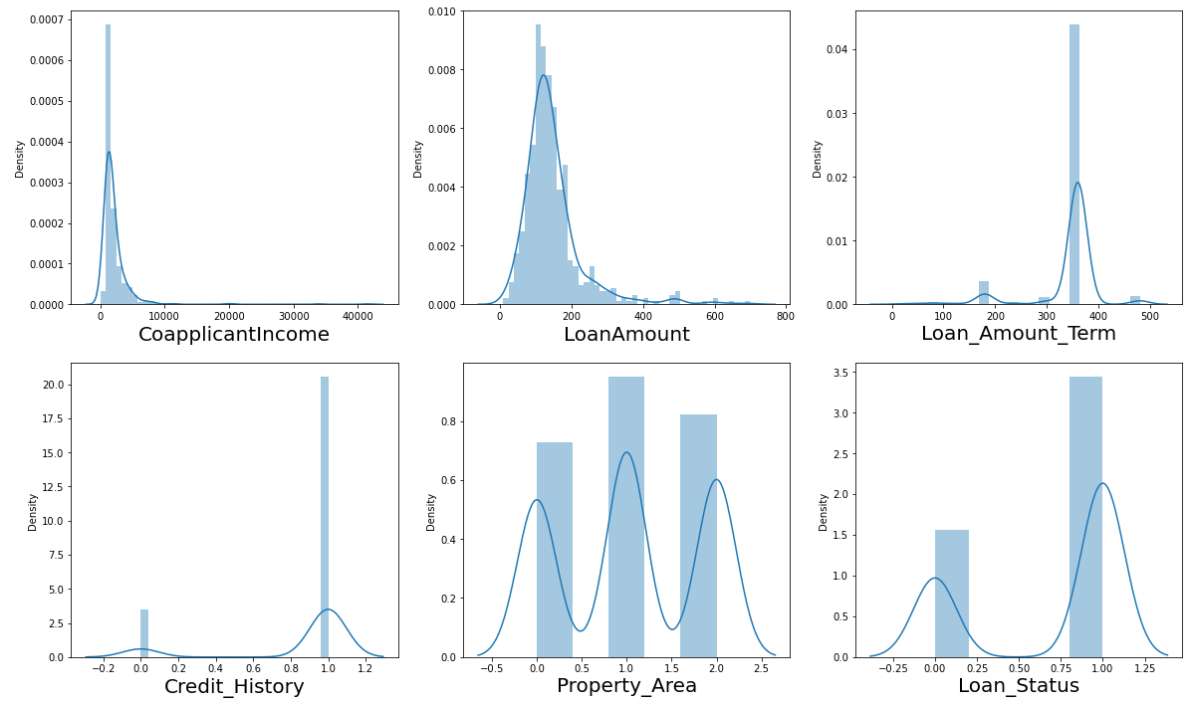
* Male have dependents much more than females



* Applicants those who are not self employed but graduate has good credit history over self employed applicants but not graduate.



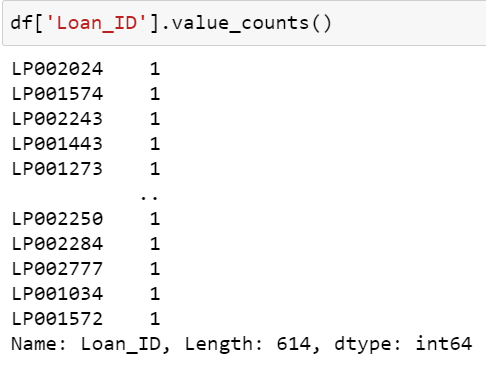




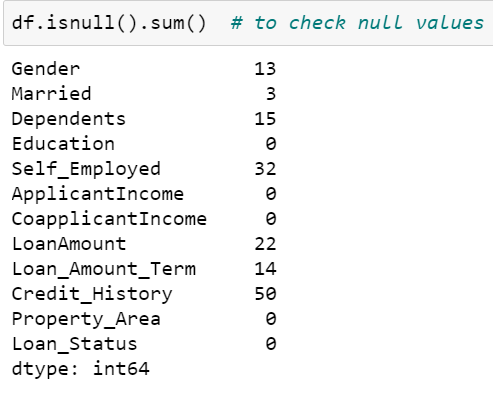
* Skewness can be seen for CoapplicantIncome, LoanAmount and ApplicantIncome.

**PREPROCESSING PIPELINE**

* Dropping ‘Loan\_ID’ column because it is just containing identifier which is not useful for prediction.

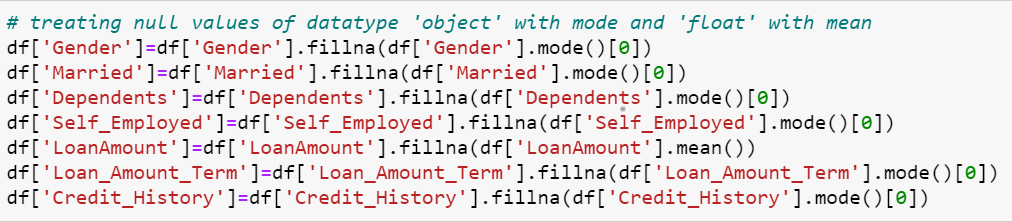


* Checking null values in the dataset.

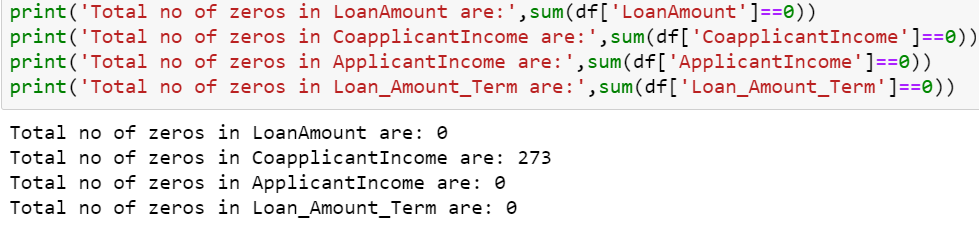


There are null values present.

* Treating null values

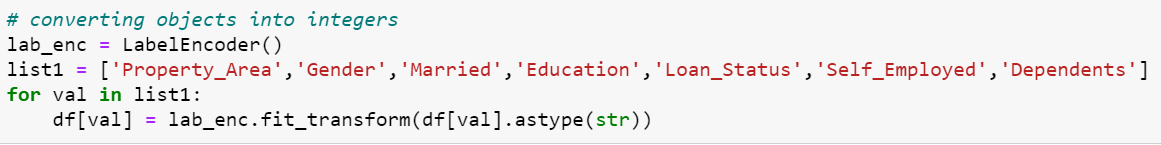


* Counting sum of zero’s present

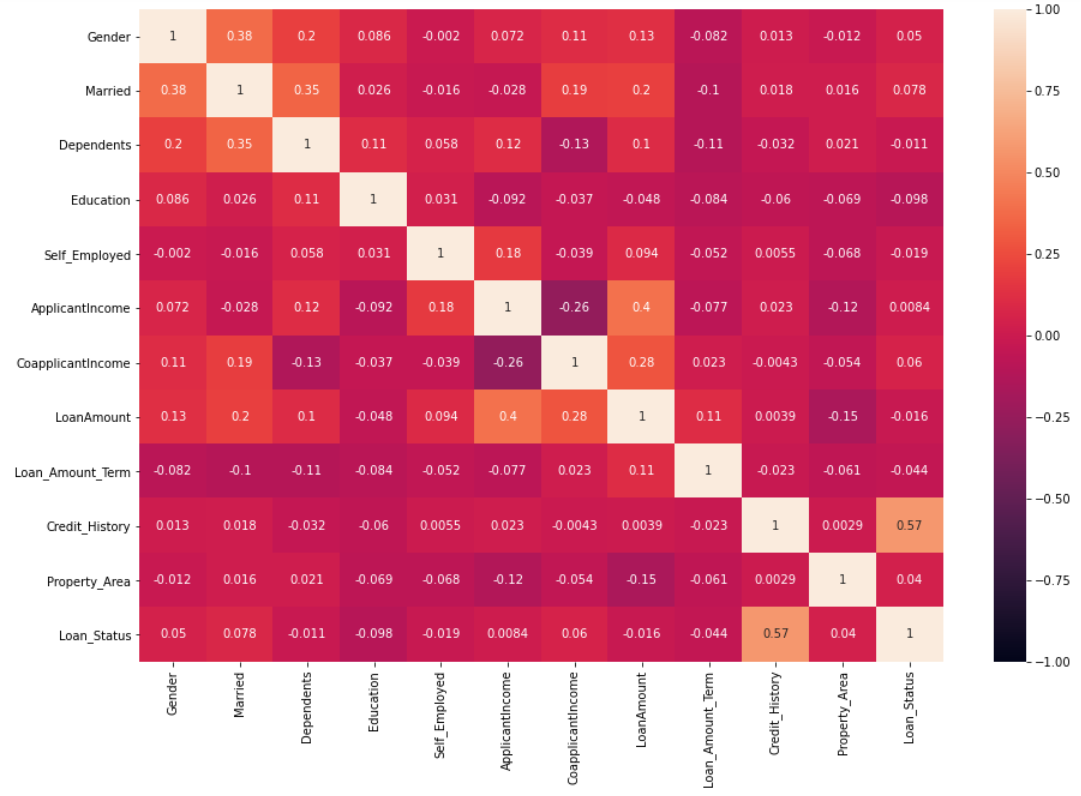
In Co Applicant Income column there are 273 values which are zero(0) which should not be the case if applicant is requesting for loan with the help of his co-applicant income.

Hence treating this error by applying median value.

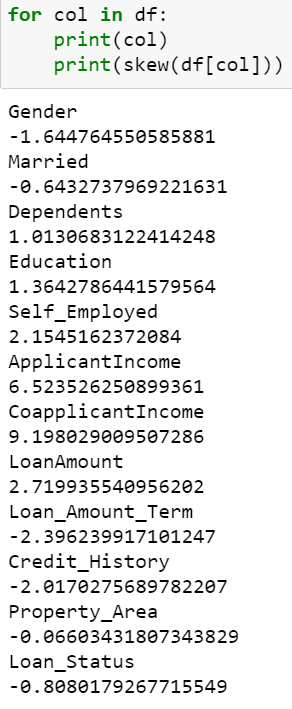
* Using Label Encoder to convert objects into integers/machine language



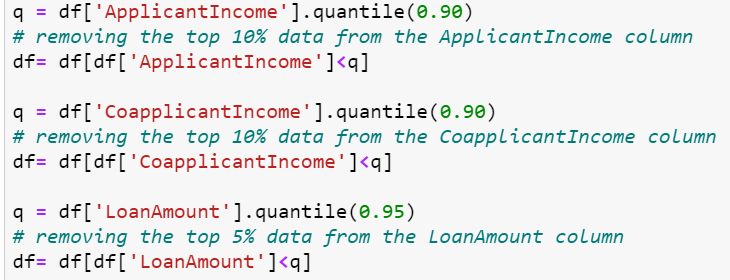
* Plotting heatmap to find correlation between features and target variable (Loan\_Status).



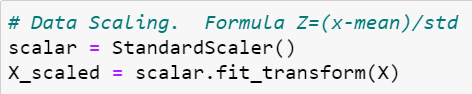
* Only credit history has good co-relation with loan status.
* Then checking for skewness for features.



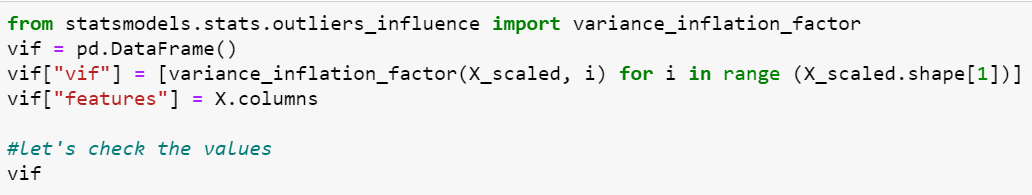
* Applying quantile method to treat skewness

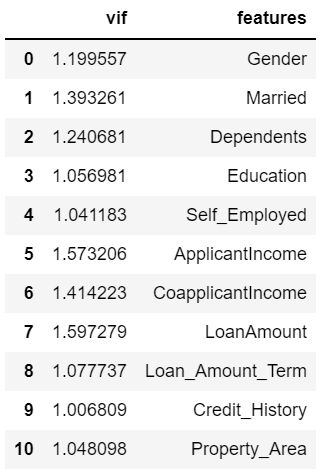


* Splitting target variable (Loan\_Status) to variable y and rest all features to variable x.
* Then doing data scaling to normalize the range of independent variables or features of data.



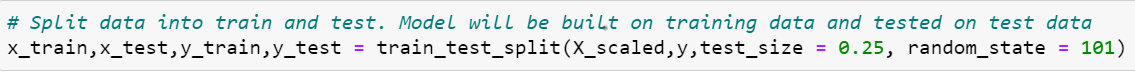
* Using variance inflation factor to test the effect of multiple variables/features on target variable (Loan\_Status).





No Multicollinearity problem exists.

* At last splitting data into train and test where model will be built on training data and tested on test data. Sample Data was divided 25% for test and rest 75% for train for model building.

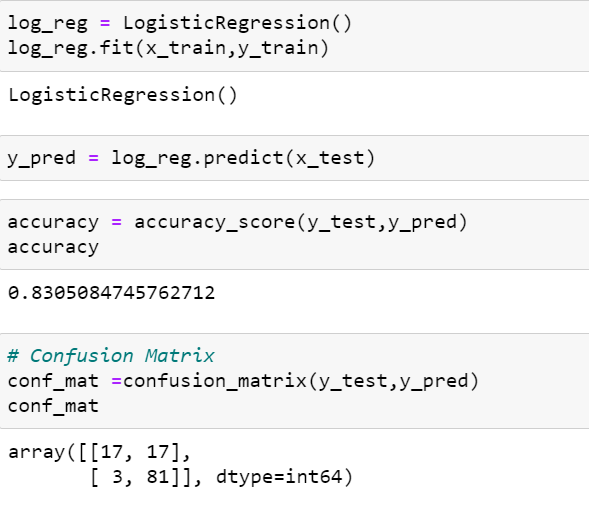


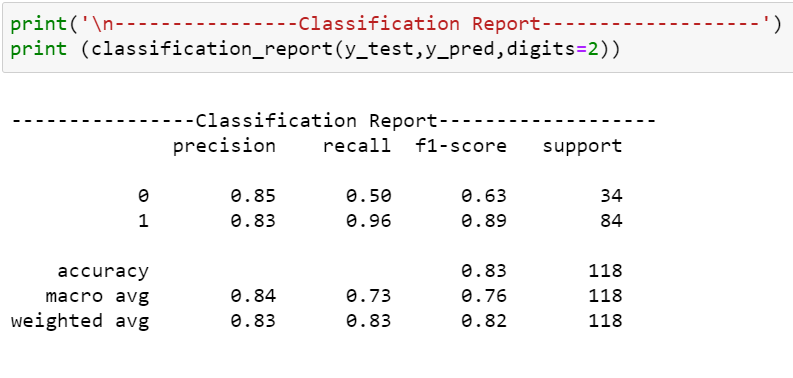
**BUILDING MACHINE LEARNING MODELS**

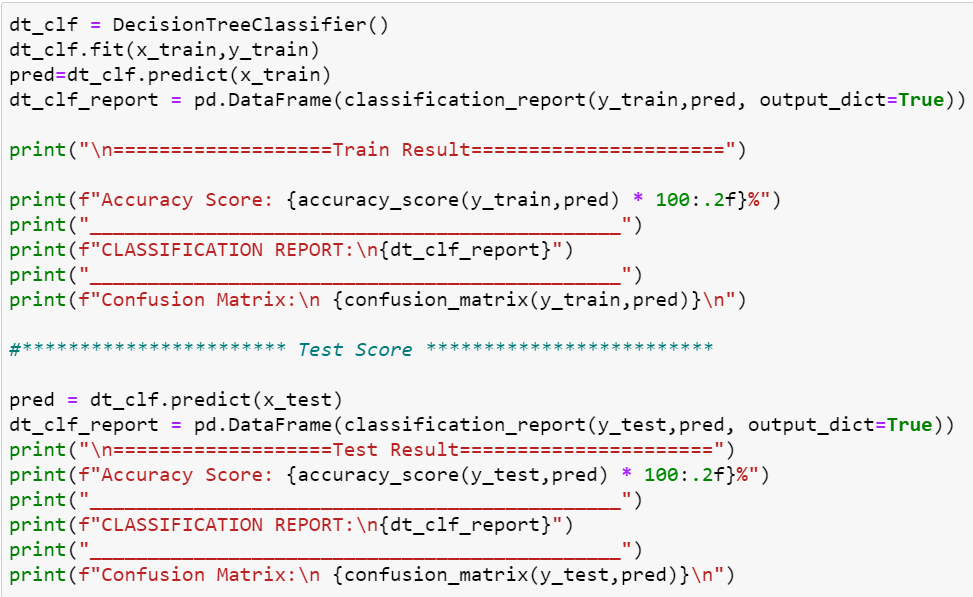
Different algorithms applied for prediction of loan application status of customers applied for loan.

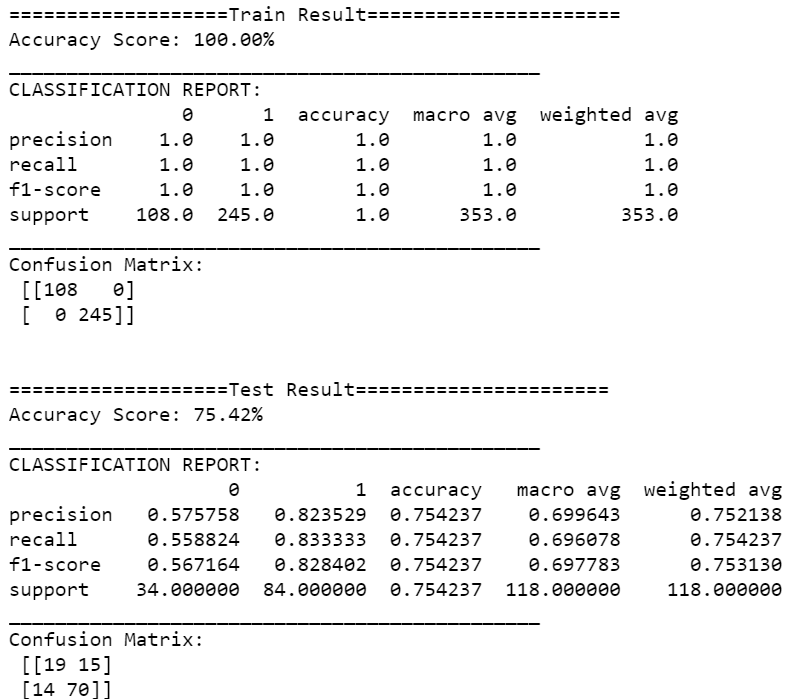
* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* K Nearest Neighbors (KNN) Classifier
* Support Vector Classifier
* Gradient Boosting Classifier
* Stochastic Gradient Descent Classifier

LOGISTIC REGRESSION





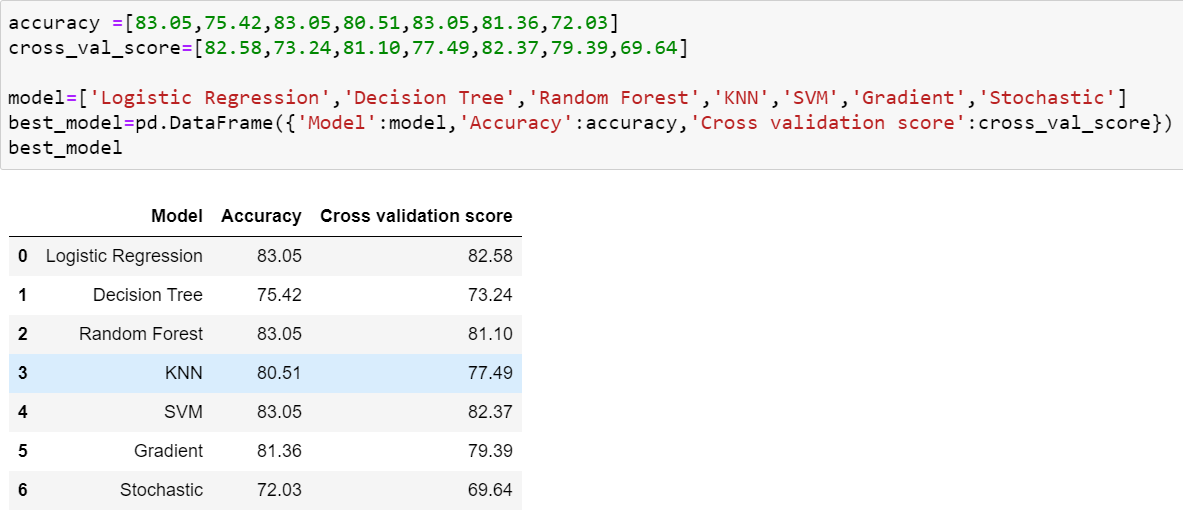
DECISION TREE CLASSIFIER****



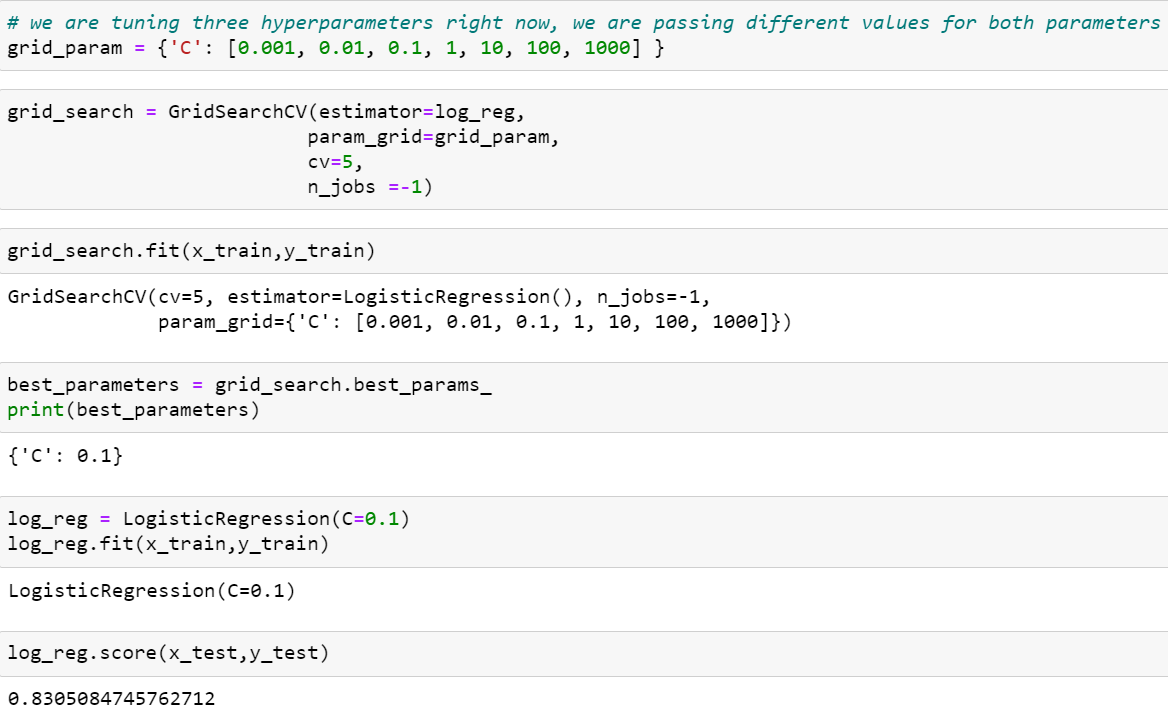
* Similarly done for rest all models like Random Forest,

KNN, SVC, Gradient Boosting and Stochastic Gradient Descent.

* Created a dataframe to observe accuracy and CV score for all models in a proper frame.



* Logistic Regression is our best model because the difference between its accuracy and CV score is least among other models.
* Now we will apply hyper parameter tuning on our best model (Logistic Regression) using GridSearchCV.

****

* Getting same accuracy after doing hyper parameter tuning.

**CONCLUDING REMARKS**

* The dataset consists of male customers mostly than female.
* Almost all applicants have favoured loan amount term of 360 months.
* Lots of loan proposal has come from married applicants.
* From semi-urban area, maximum applicants have applied for loan.
* Graduate candidates applying for loan has good credit score.
* Credit history is the most important factor for approval of loan, as it was also showing good correlation with target variable (Loan Status) observed in heat map.