**Loan Approval Prediction Machine Learning**

# Introduction

In this article we are going to solve the Loan Approval Prediction. This is a classificvation problem in whoch we need to classify whether the loan will be approved or not. classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

# Table of content

### Understanding the Problem Statement

1. About the dataset
2. Load essential python libraries
3. Load Trainning /Test datasets
4. Data Preprocessing
5. Exploratory Data Analysis(EDA)
6. Feature Engeneering
7. Build Machine Learning Model
8. Make predictions on the test dataset
9. Prepare submission file
10. Conclusion

This project deals all kinds of home loans. They have a presence across all urban, semi0urban, and rural area. The customer first applies for a home loan and after that, the company validates the customer elegibility for the loan

The company wants to automate the loan eligibility process based on customer details provided while filling out online application forms, These details are Gender, Marital Status, Education, number of dependents, Income,Loan Amount,Credit History and others.

To automate this process, they have provided a dataset to identify the customer segments that are eligible for loan amounts so that they can specifically target these customers.

The problem statement is given below and also download the dataset.

### Problem Statement:

Loan Application Status Prediction Problem Statement: This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

Independent Variables:

* Loan\_ID
* Gender
* Married
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* Loan\_Amount
* Loan\_Amount\_Term
* Credit History
* Property\_Area

Dependent Variable (Target Variable):

* Loan\_Status

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

As mentioned above this is a Binary classification probelm in which we need to predict our target label which is "Loan\_Status"

Loan status can have two values: Yes or No

Yes: if the loan is approved

No: if the loan is not approved

So using the dataset we will train our model and try to predict our target column that is "LoanStatus".

**About The Data Set**

1

Variable               Description

2

Loan\_ID                 Unique\_Id

3

Gender                 Male/Female

4

Married                 Applicant Married(Y/N)

5

Dependents             Number of Dependents

6

Education               Applicant Education(Graduate/Under Graduate)

7

Self\_Employed           Self\_Employed(Y/N)

8

ApplicantIncome         ApplicantIncome

9

CoapplicantIncome       Co ApplicantIncome

10

Loan\_amount             LoanAmount in Thousands

11

Loan\_Amount\_Term       Term of Loan in Months

12

CreditHistory           Credit History meets the guidelines

13

Property\_Area           Urban/Semi Urban/Rural

14

Loan\_sTatus             (Target) Loan Approved(y/N)

**Import Essential Libraries**

# 

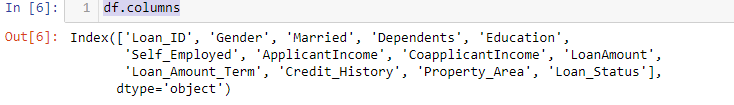
# load the dataset

# t



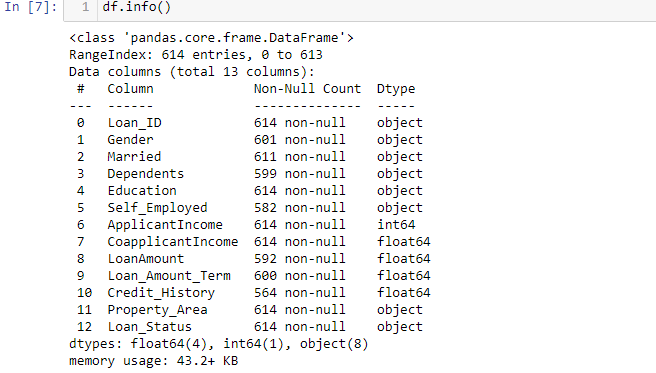


There are 614 rows and 13 columns in the dataset



**Categorical Columns**: Gender (Male/Female),Married (Yes/No),Number of Dependents (Possible values:0,1,2,3+), Education(Graduate/Under Graduate),Self-Employed(No/Yes), CreditHistory(Yes/No),PropertyArea(Rural/Urban/Semi-Urban) and Loan Status(Y/N)(i.e Target Variable)

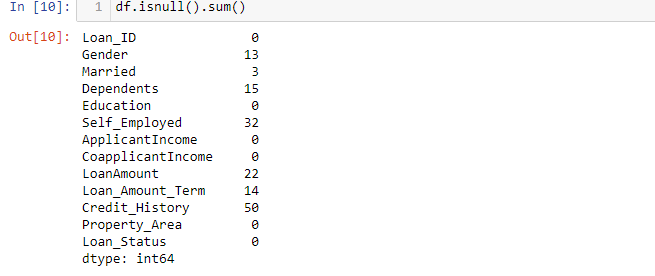
**Numerical Columns**: LoanID,Applicant Income, Co=Applicant Income, Loan Amount, and Loan amount term



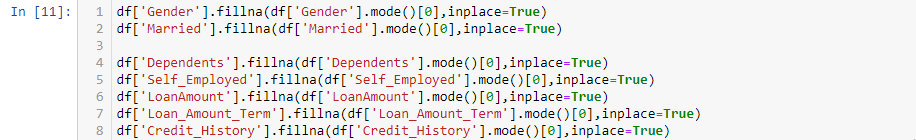
The dataset consist of 8 features are of object type, and 4 features are of float type and 1 is of type integer. Our target variable Loan\_Status is of type object

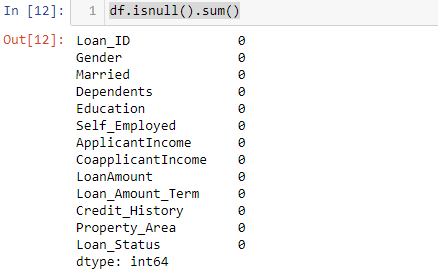
# EDA

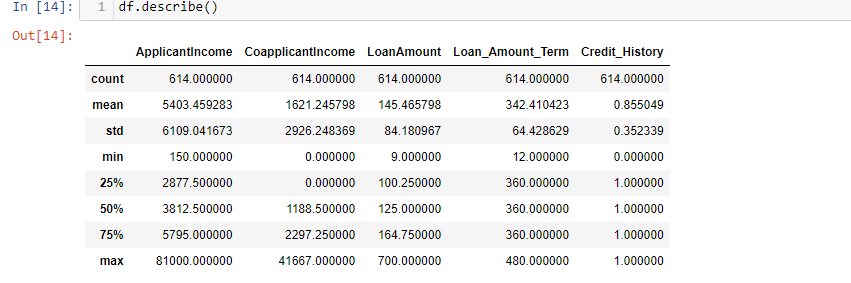
#Identifing missing values



The dataset consist of null values in the columns Gender,Married,Dependents,self\_Employed,LoanAmount,Loan\_Amount\_Term,Credit\_History. so we have to fill those null values For numerical\_data we fill with mean/median For categeorical\_data we fill with mode of that perticular column





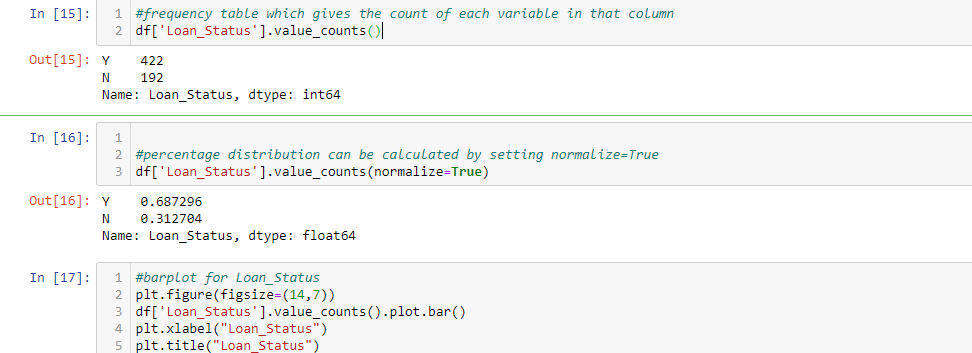
Describe funtion shows statistical data of all the features.count tells the no.of rows in each column,and min,max values of the columns,mean and Standarddeviation of the columns values,and the quartiles information. there is a lagre gap between 75% and max columns for ApplicantIncome,coapplicantIncome,LoanAmount,may be some outliers present in the data.

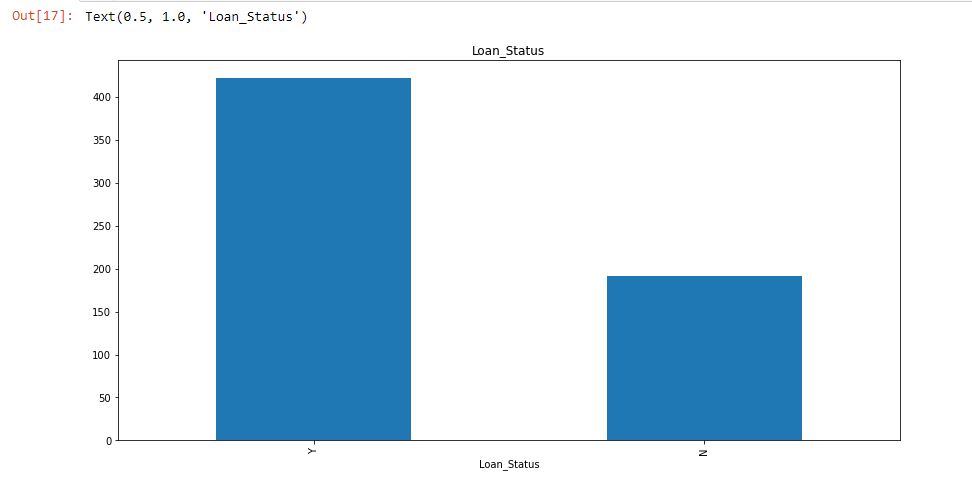
# Data Visualization

# Univariate Analysis

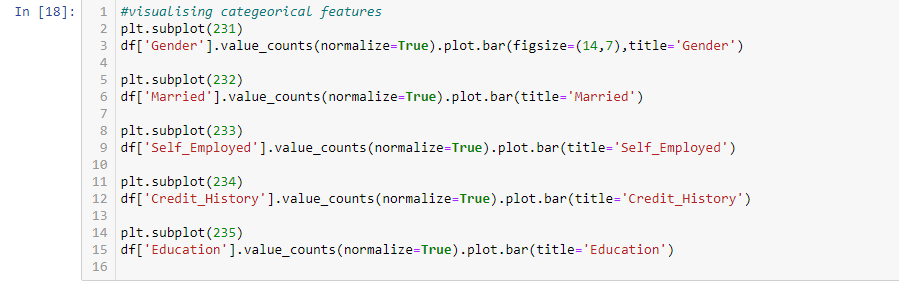
## Independent Variable(categeorical)

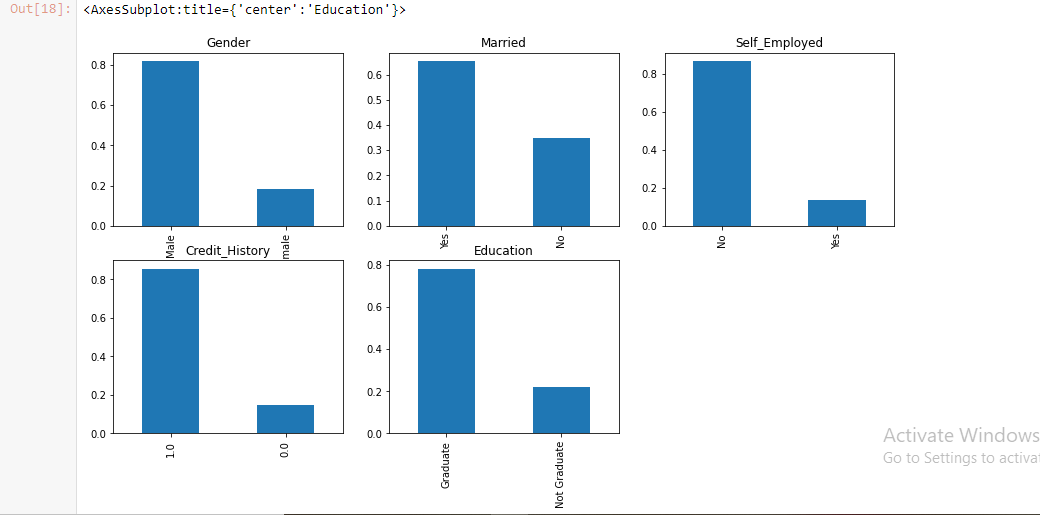
Univariate Analysis is when we use each variable individually.For categeorical data we use barplot or frequency table which will calculate each categeory in a perticular variable.





422 members got yes(loan approval) and 19 members got No





From the above bargraphs we can observe that

80% males are applied for loan

60% people are married

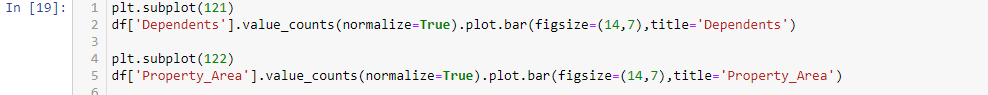
80% are self\_employed

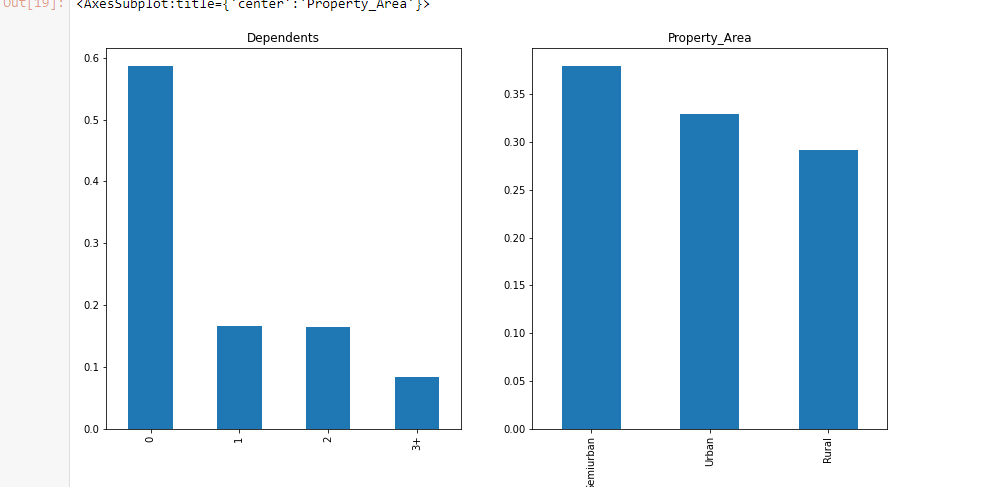
80% are having credit history

75% are Graduates

## Independent Variable(Ordinal)

variables in categeorical some variables are have some order(Dependents,Property\_Area)

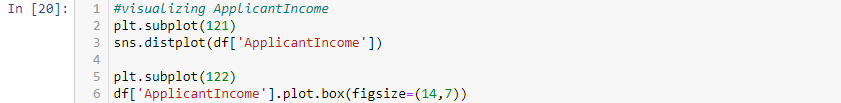


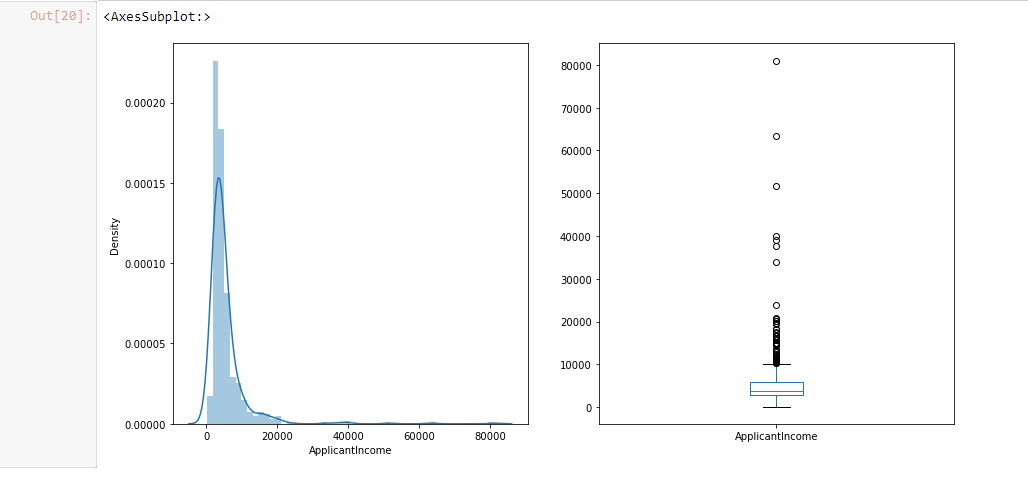


From the above graph we can observe that morethan half of the applicants are not having dependents,and most of the people are from semiurban area

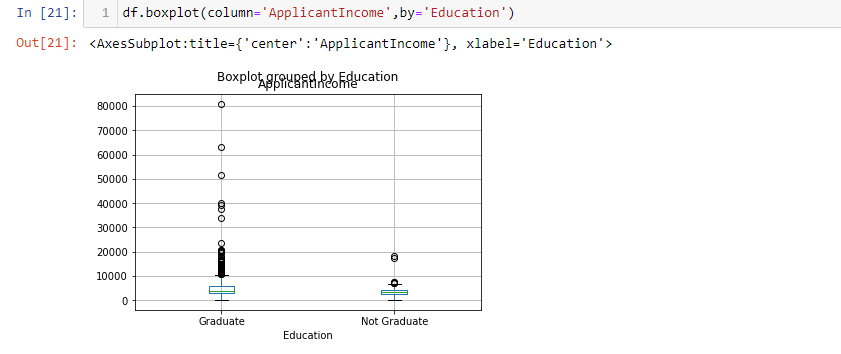
# Independent variable(Numerical)

The features 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount','Loan\_Amount\_Term' are having numerical values

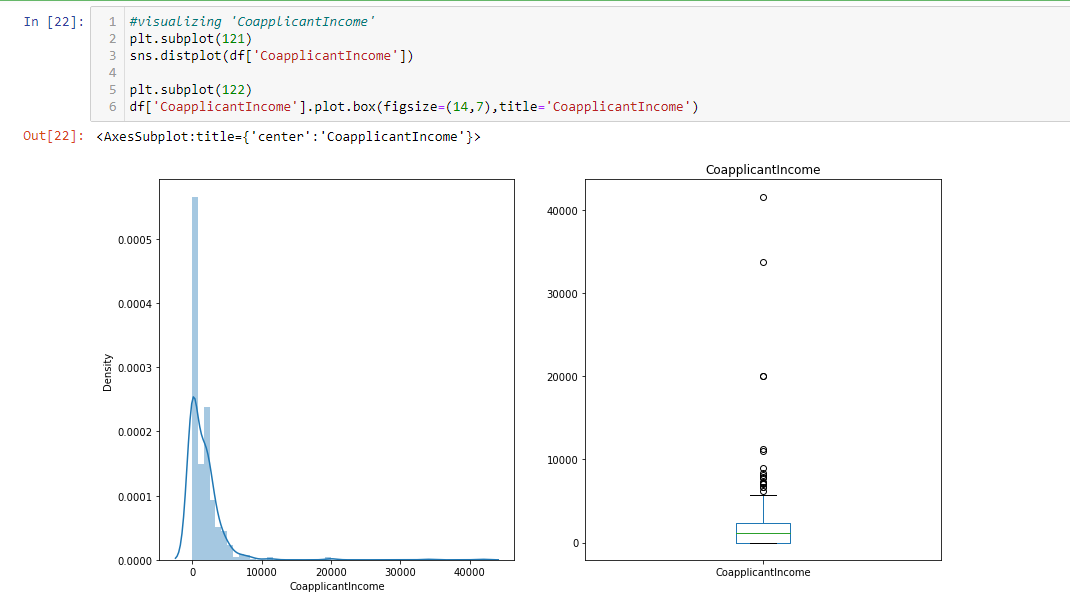




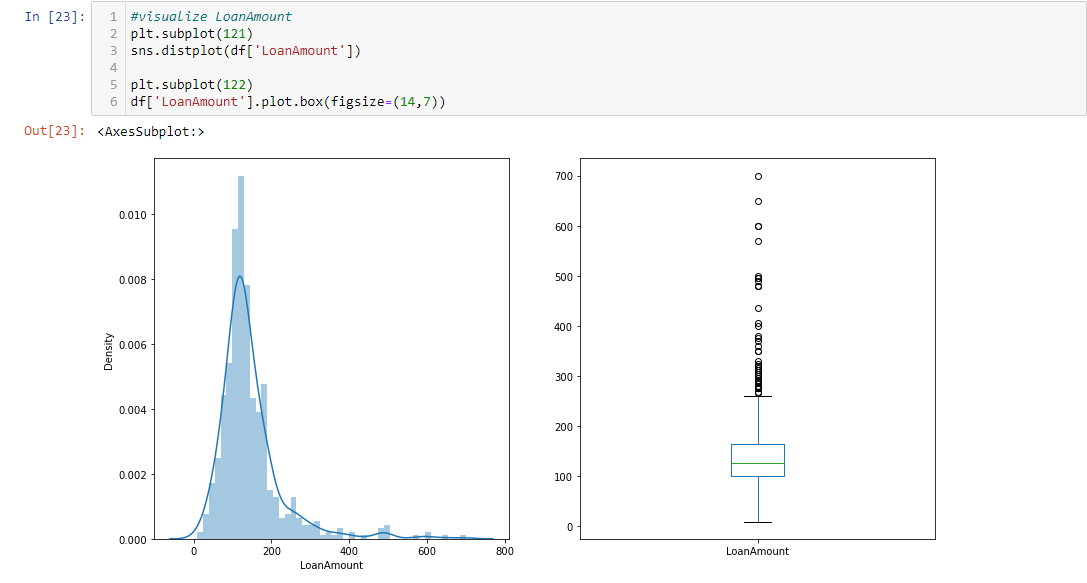
From the above graphs ApplicantIncome is rightskewed and therre are so many outliers present in the data we have to handle them in later to perform the model better



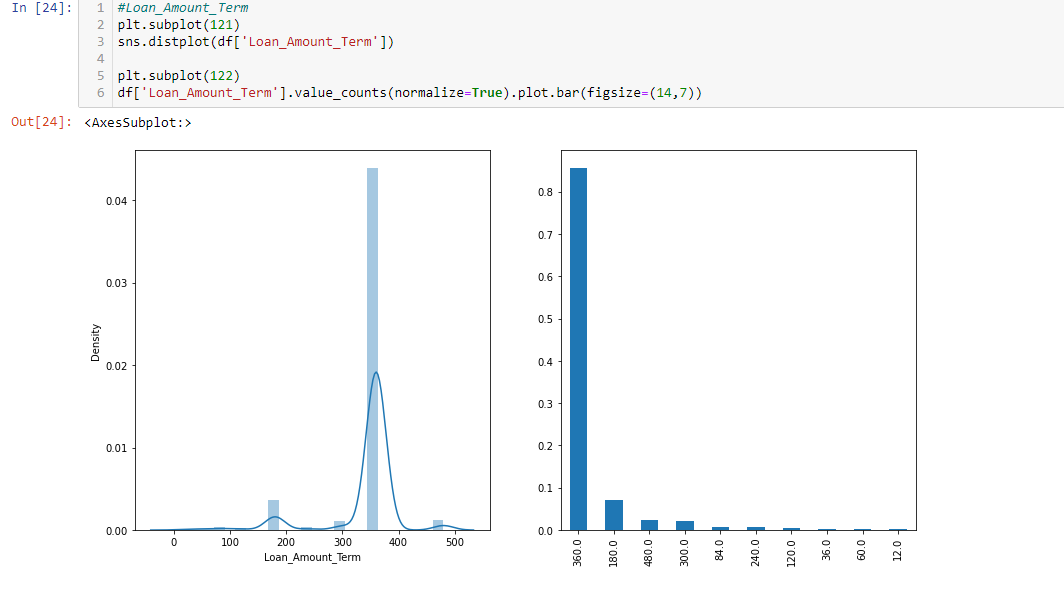
There is high income for Graduates may be that is present in the outliers



CoapplicantIncome is not normally distributed and outliers also present in the data



LoanAmount is normally distributed and slightly right skewed and there are outliers present in the data

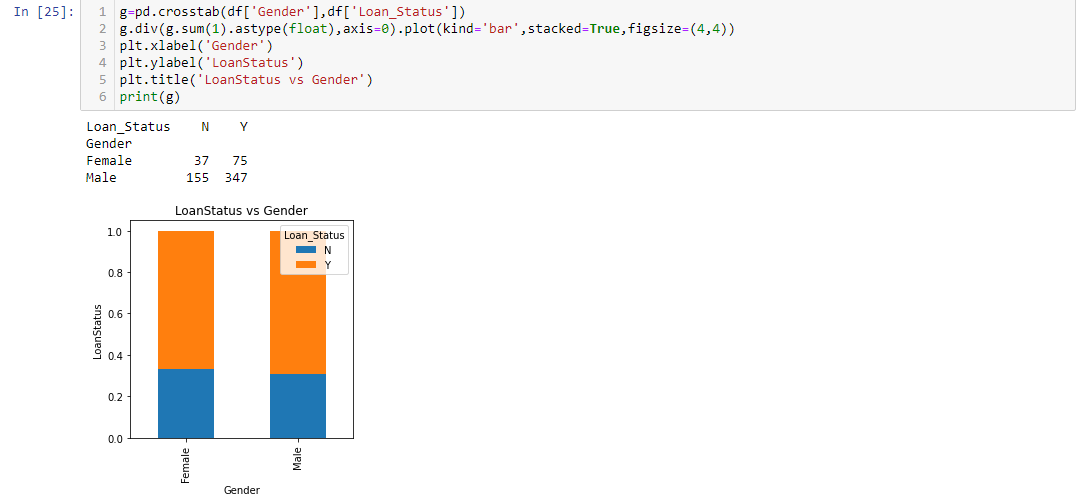


Most of the people are choosing the Loan\_amount\_Term as 360 months or 30 years of period and it is not normally skewed

# Bivariate Analysis

After exploring univariate Analysis we now analyze those features with target variable

## Categeorical Independent variables vs Target Variable



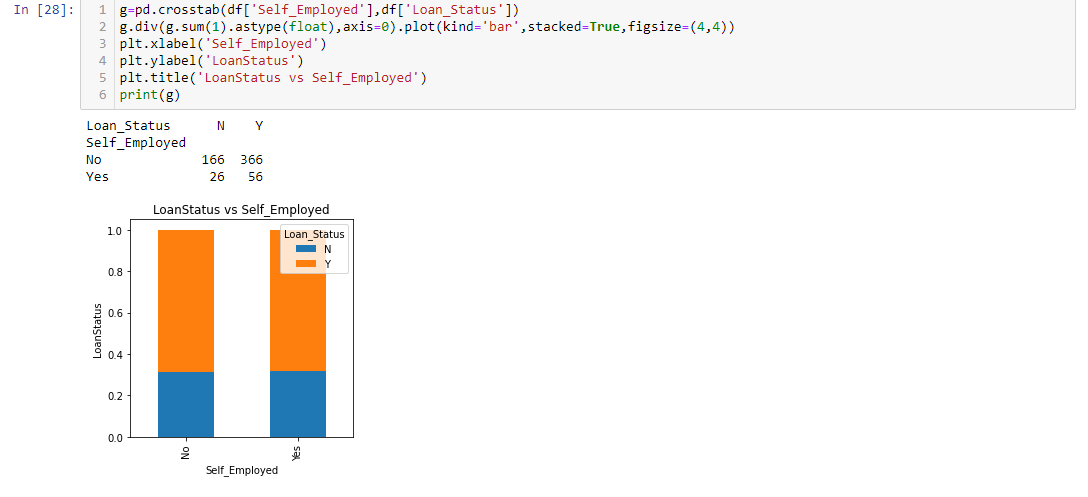
males LoanStatus is slightly highly accepted than female



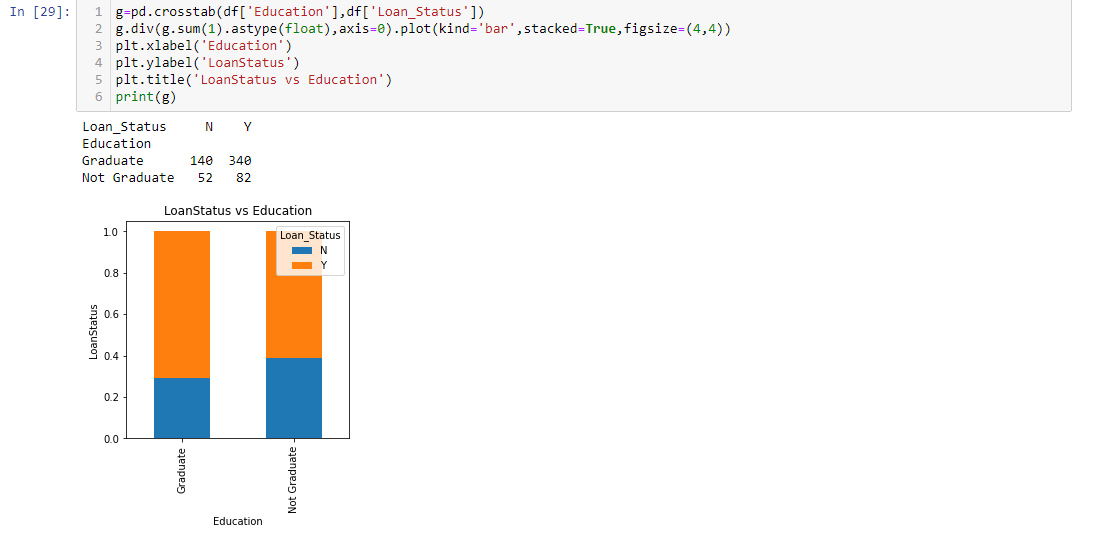
Married Applicants are accepted more for loanapproval



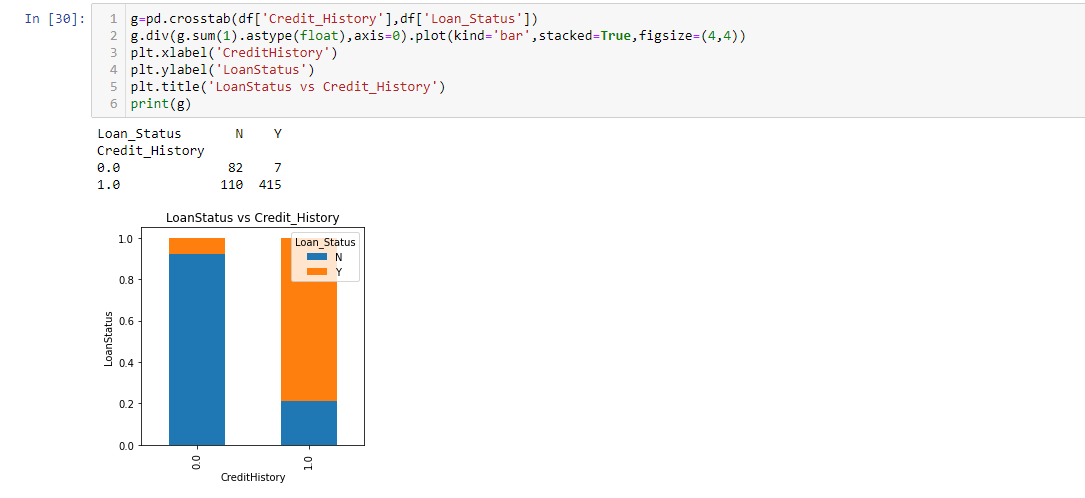
dependents with 1 and 3+ having same loan approval rates



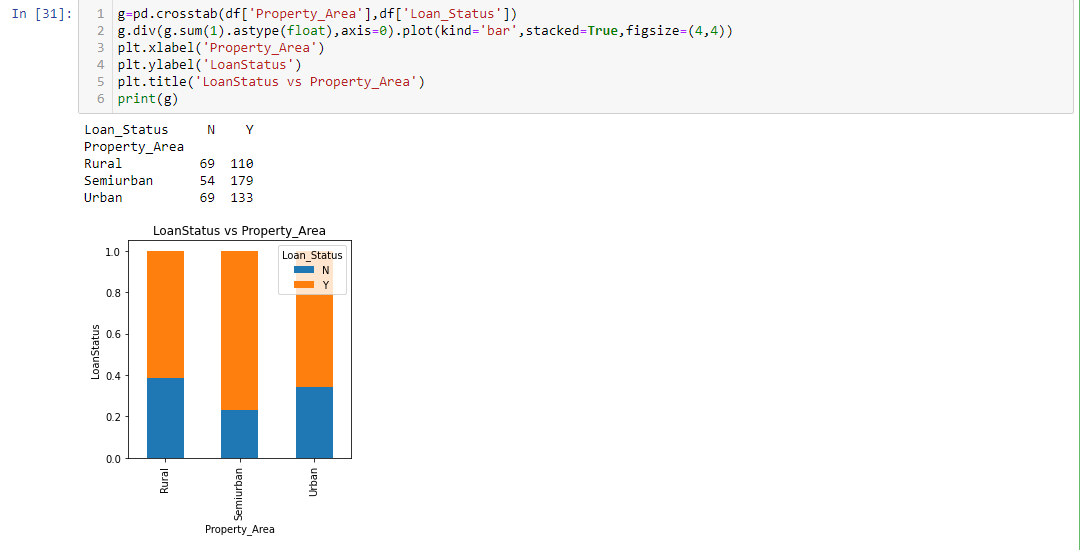
There is same loan approval ratio for self\_Employed



Graduates got high loan approval than Not-Graduates



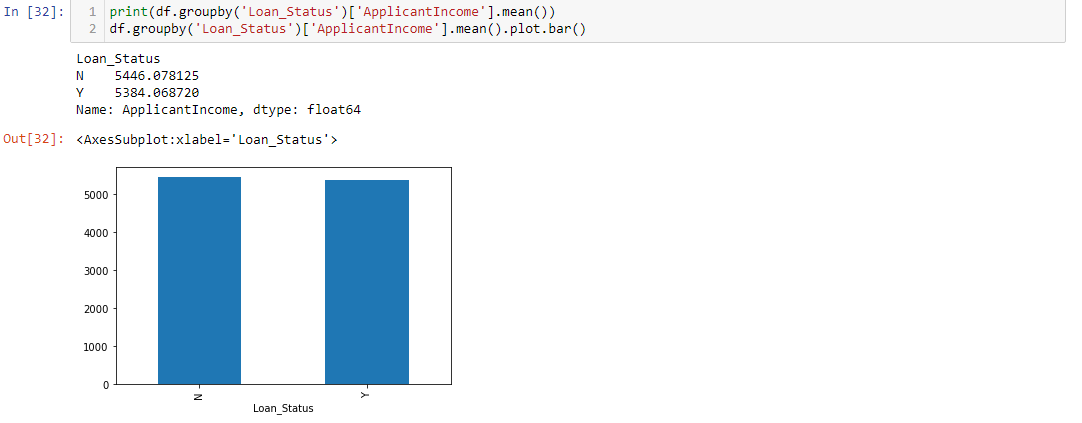
Credit\_History with having 1 got approved for loan



people of semi urban got loan approved

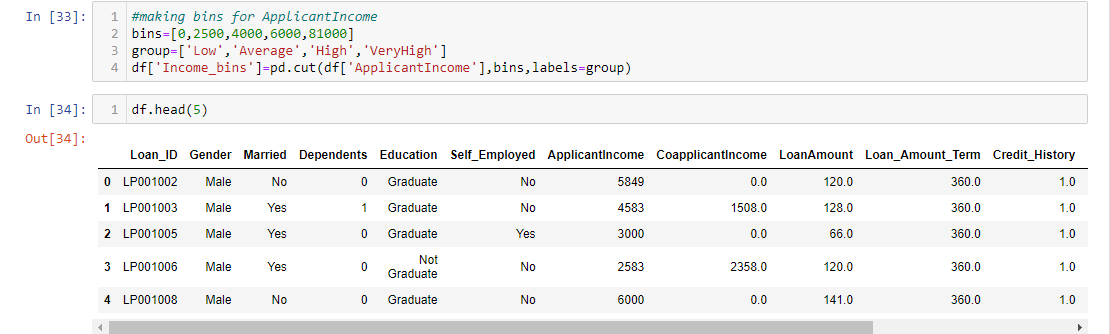
### Visualize NumericalVaribles Bivariate Analysis

we will try to find mean income of the people who got loan approved and not approved



There is no significant difference between LoanApproval for Appliicants income,so ,we make bins for ApplicantIncome values and analyse LoanStatus

# Feature Engineering





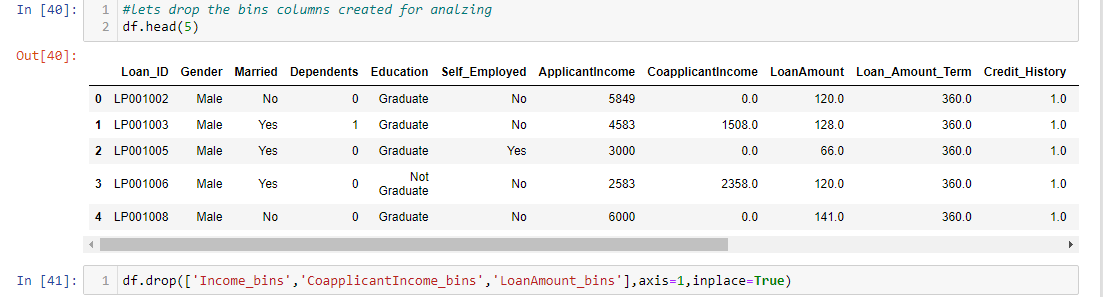
ApplicantIncome does not affect the loan Approval



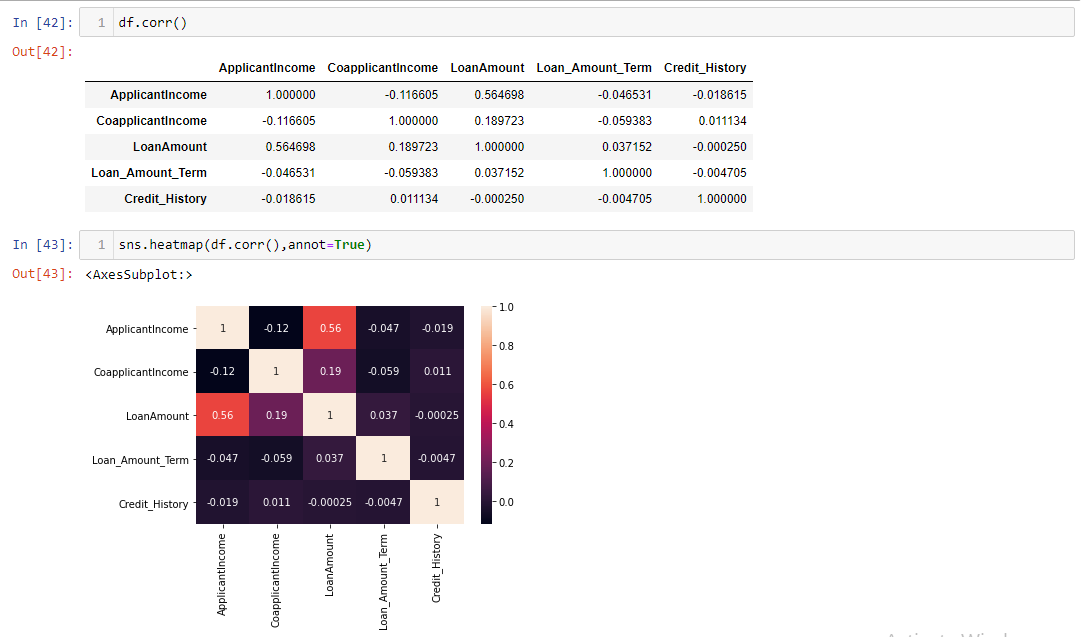
As we can observe from the above graph that low CoApplicantIncome got approved loan than the Average and High.But this is not right.May be most ofthe applicants dont have coapplicants.



proportion of Approved loans is high for Low and Average LoanAmount than Higher LoanAmount,i.e chance of LoanApproval is high when the LoanAmount is less



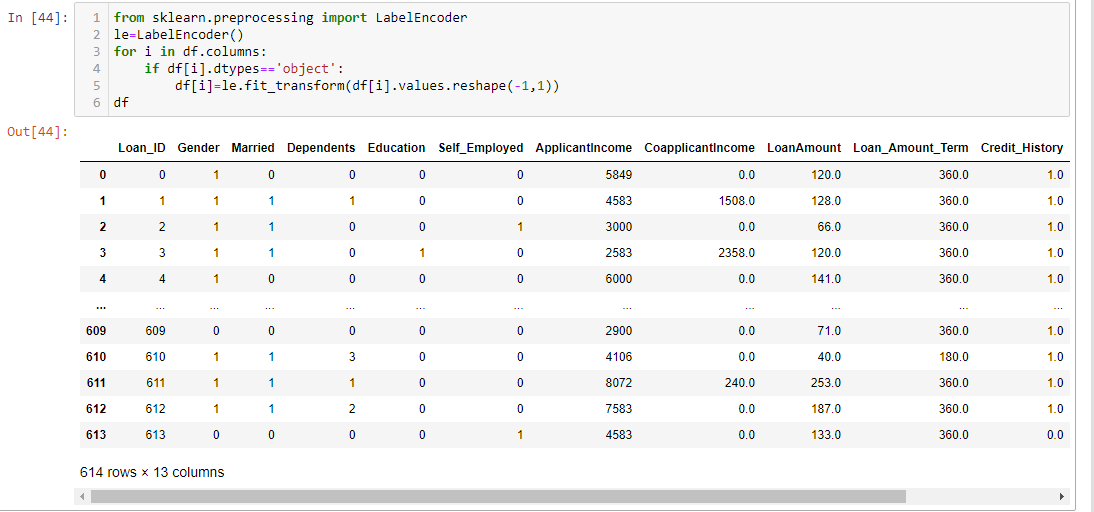
# Correlation

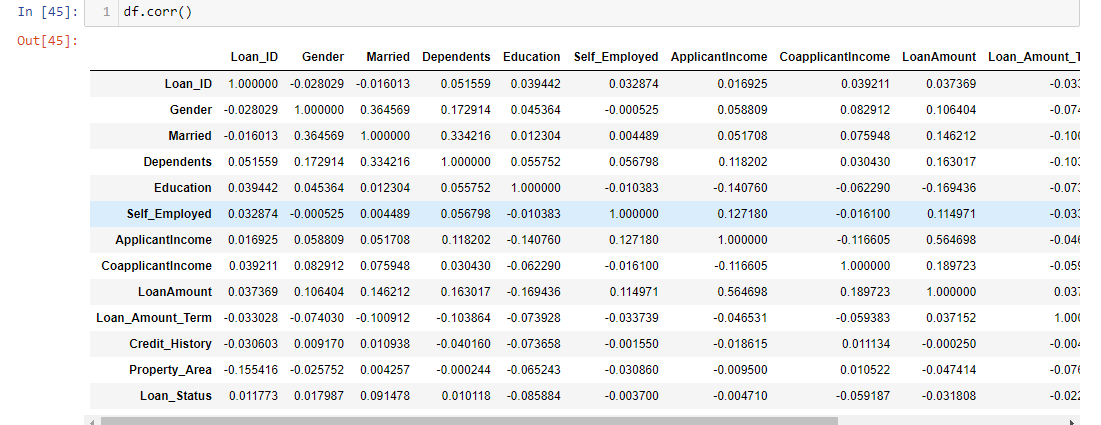


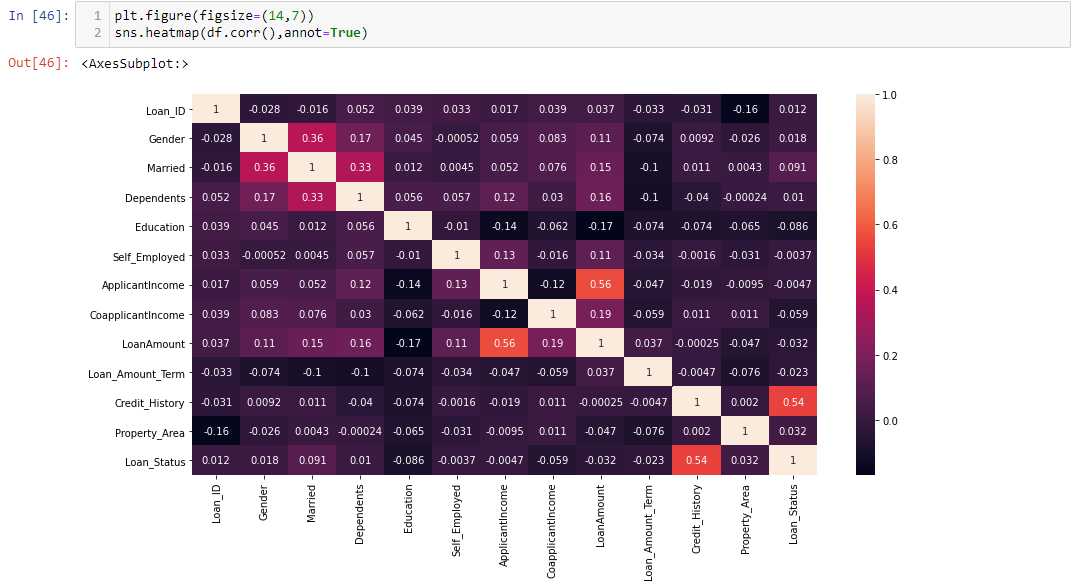
LoanAmount is correlaterd with ApplicantIncome with 56% LoanAmount is correlaterd with CoapplicantIncome with 19%

# Encoding Technique

# Our data consist of categeorical data so, we need to convert into numerical by using LabelEncoder technique







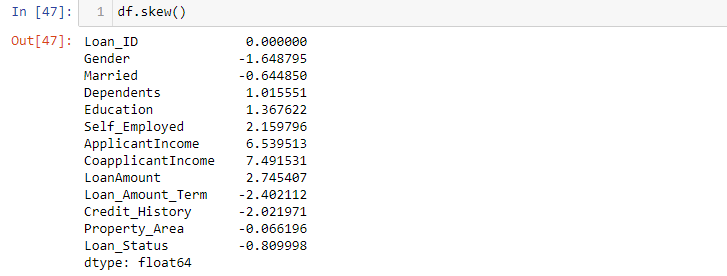
Credit\_History is 54% correlated with Loan\_Status

Married is 33% correlated with dependents and 36% correlated with Gender

ApplicantIncome and Loanamount are correlated with each other with 56%

All the other features are less correlated or negatively correlated with target variable

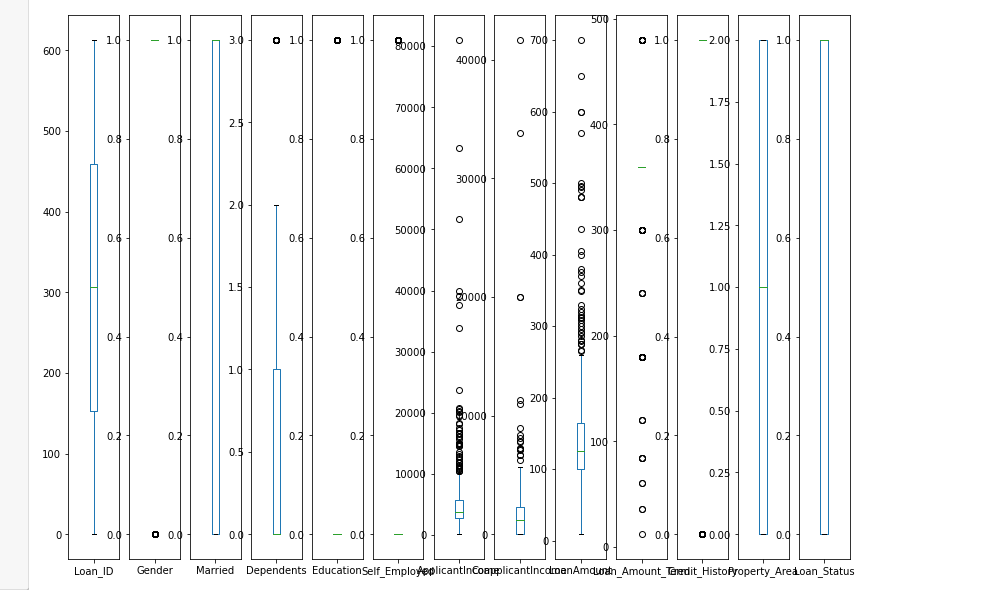
# Skewness Checking



Most of the features are not under the threshold value of skewness i.e+/-0.5

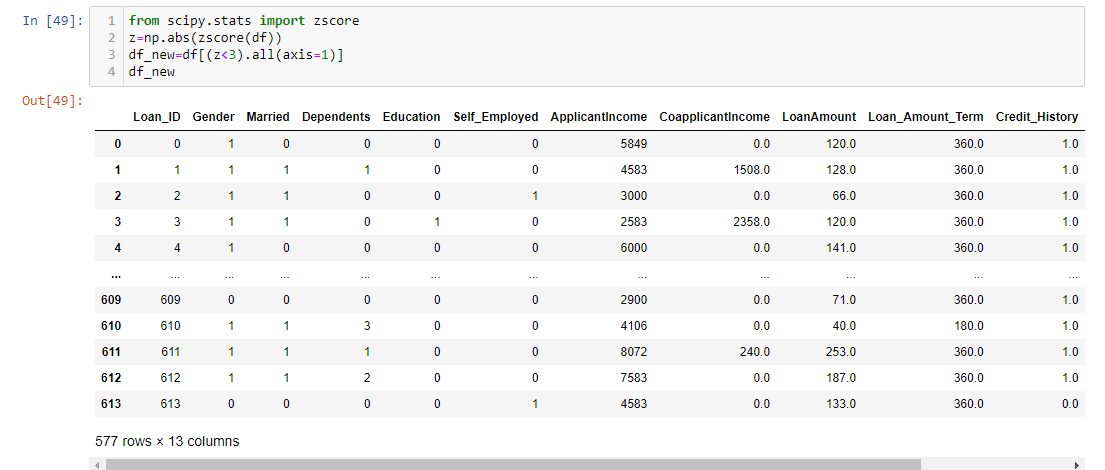
# Outliers Checking

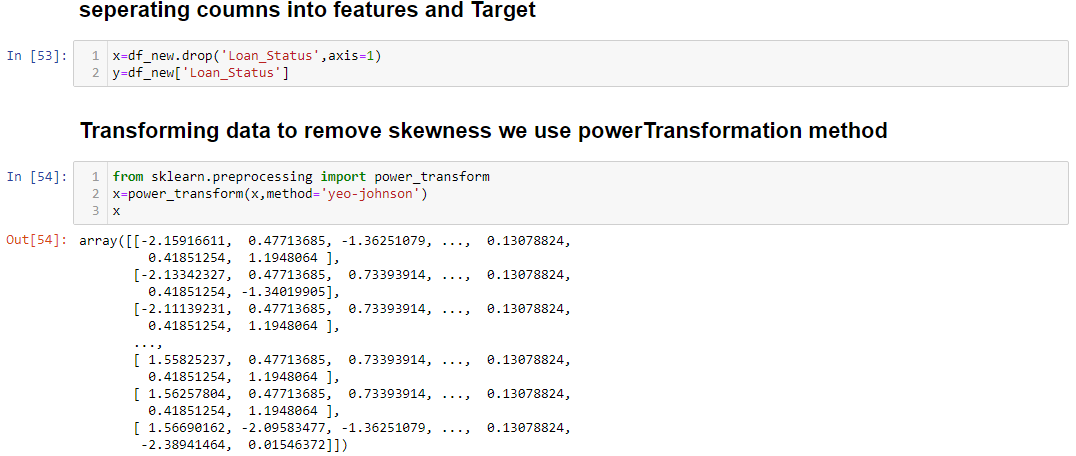
# 

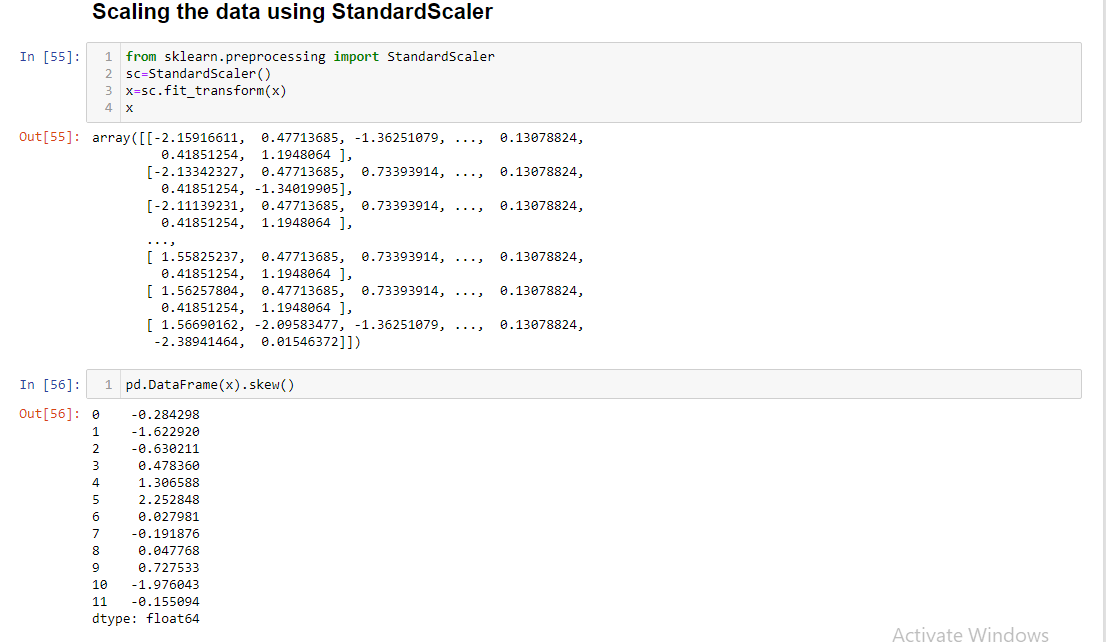


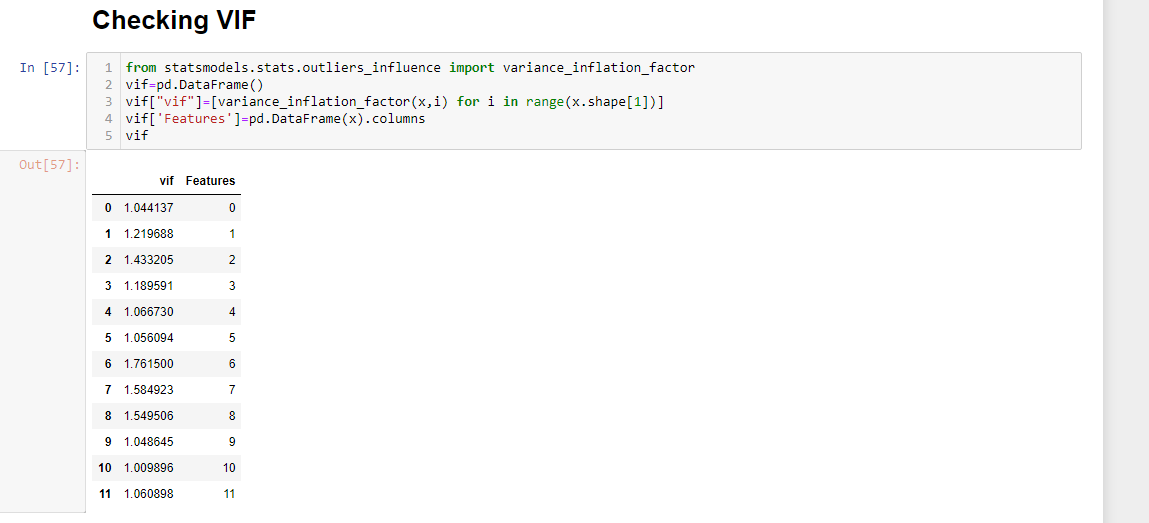
ApplicantIncome,CoapplicantIncome,LoanAmount,LoanAmount\_Term having outliers,we have to handle it

# Removing Outliers







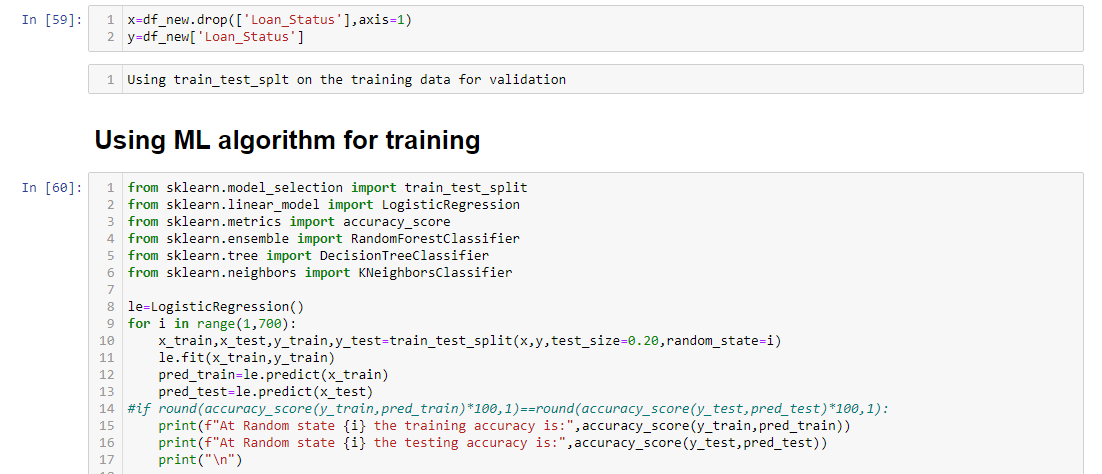


All the features are lessthan the cutoff value of vif i.e<5

# Model Building

since our target variable is bivariate so,we use the classification model

#seperating the independent variables and target variable



At Random state 10 the training accuracy is: 0.8112798264642083

At Random state 10 the testing accuracy is: 0.853448275862069

At Random state 11 the training accuracy is: 0.8286334056399133

At Random state 11 the testing accuracy is: 0.7844827586206896

At Random state 12 the training accuracy is: 0.7939262472885033

At Random state 12 the testing accuracy is: 0.9137931034482759

At Random state 13 the training accuracy is: 0.8264642082429501

At Random state 13 the testing accuracy is: 0.7758620689655172

At Random state 14 the training accuracy is: 0.841648590021692

At Random state 14 the testing accuracy is: 0.7241379310344828

At Random state 15 the training accuracy is: 0.8177874186550976

At Random state 15 the testing accuracy is: 0.8275862068965517

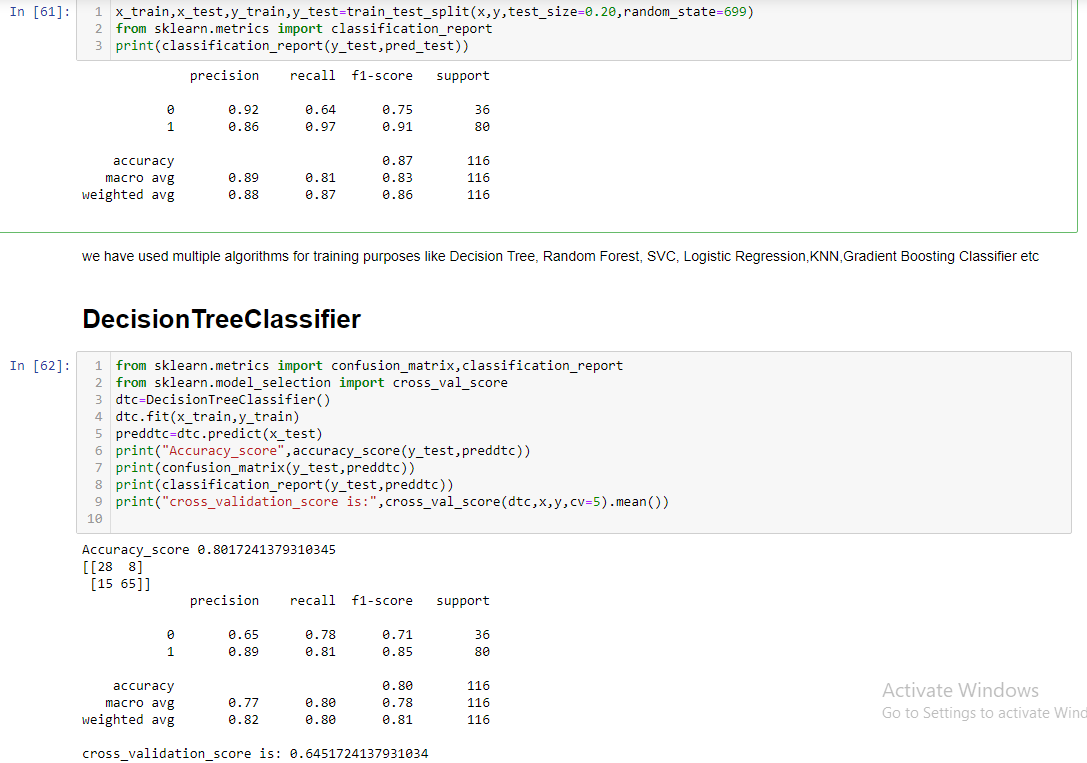
At Random state 16 the training accuracy is: 0.824295010845987

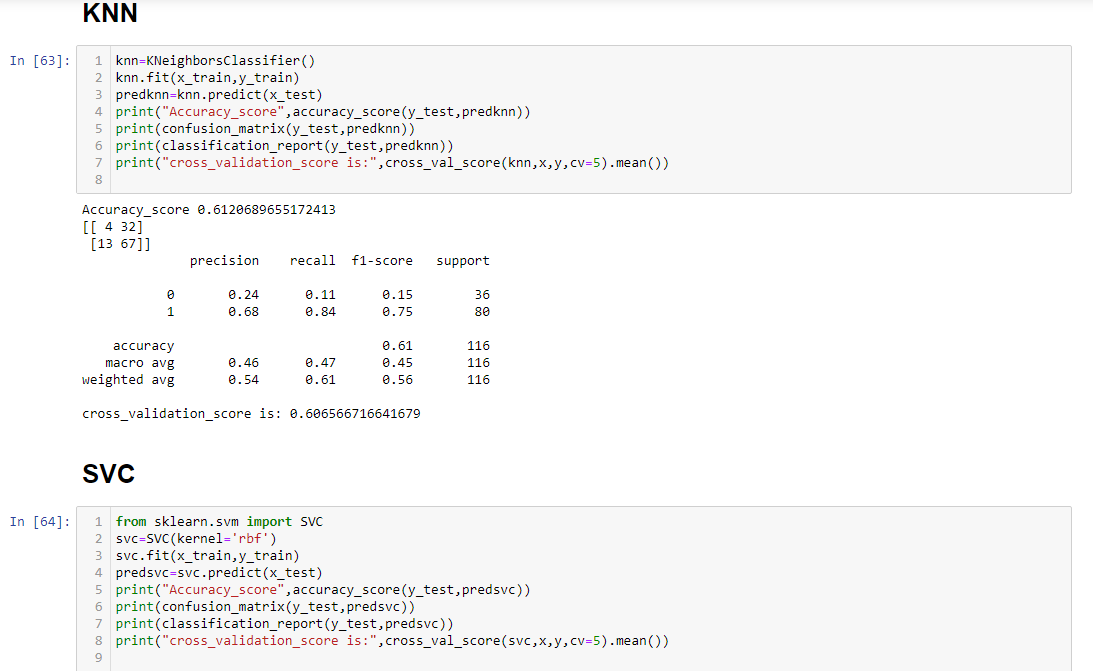
At Random state 16 the testing accuracy is: 0.7844827586206896

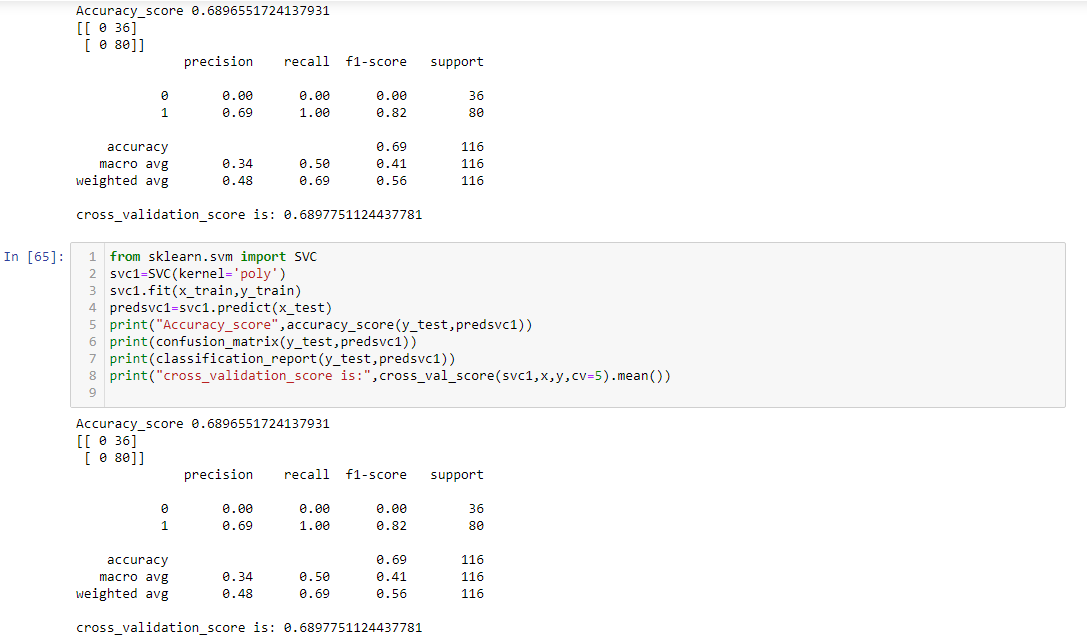
At Random state 17 the training accuracy is: 0.8112798264642083

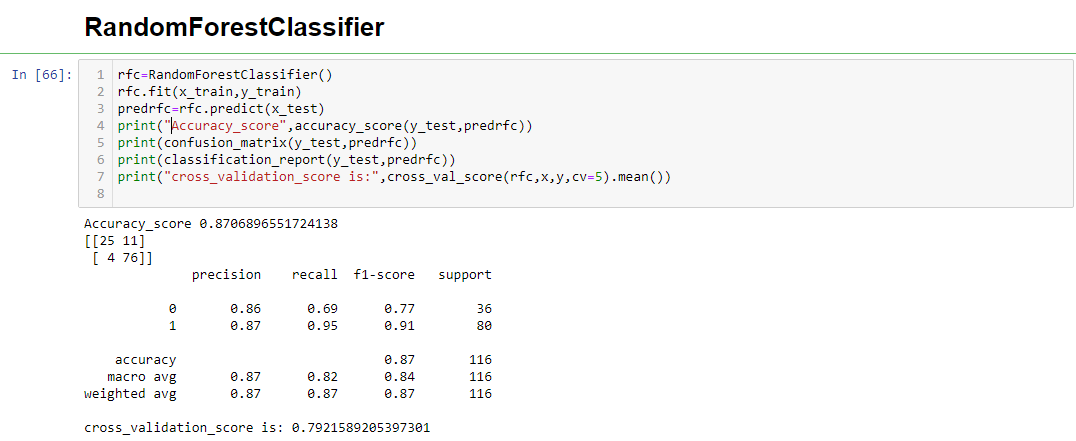
At Random state 17 the testing accuracy is: 0.7758620689655172

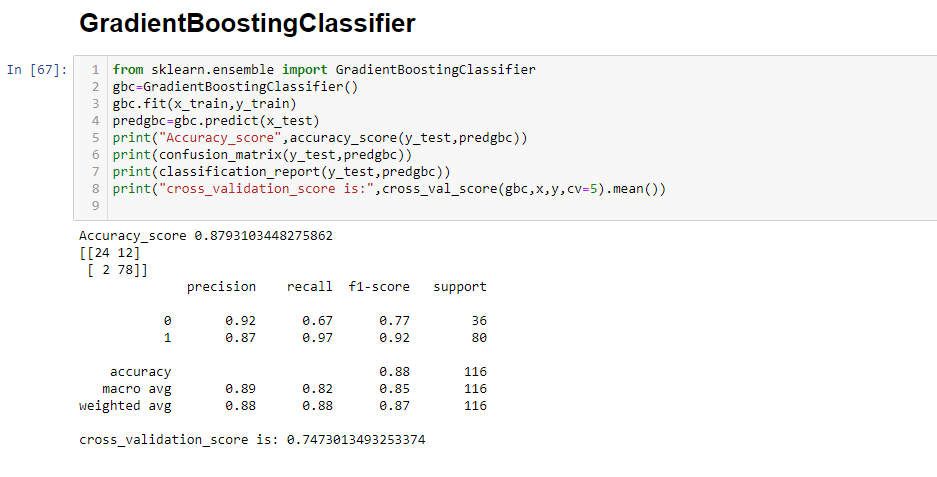
we have a (80:20) split on the training data







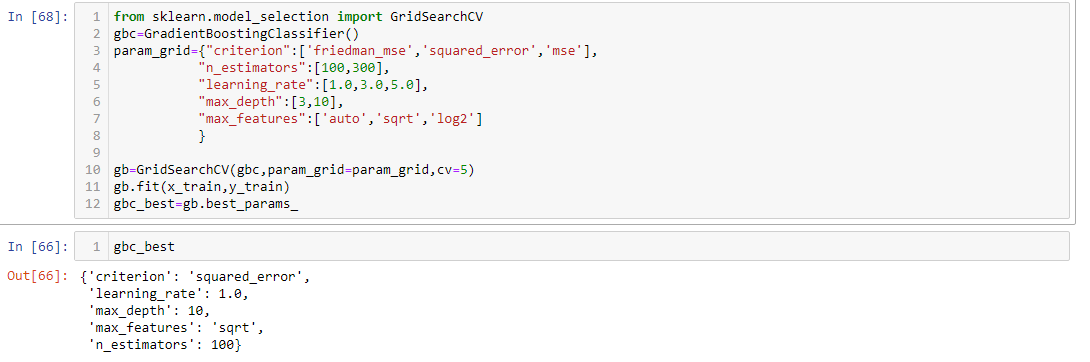




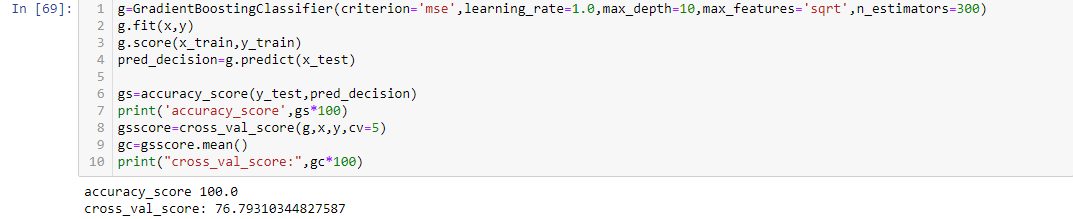
Among all the algorithms I got Gradient Boosting Classifier is getting the Highest Accuracy score i.e 87.93% with cross validation score 74.73%

After getting high accuracy\_score i tried fine-tuning it to improve my accuracyscore using GridSearchCV.

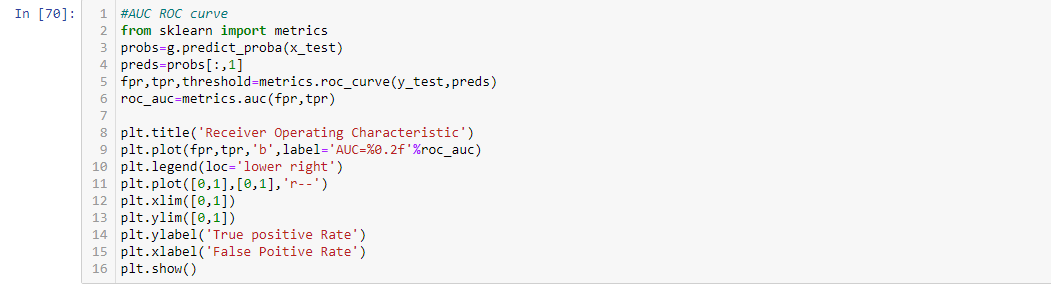
# HyperParameterTuning

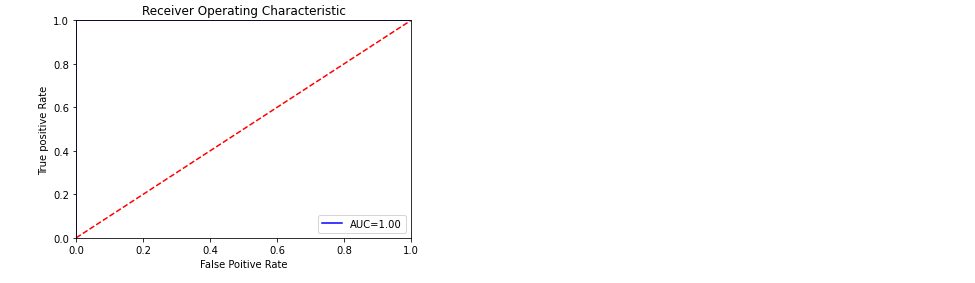


The best parameters I got after Hyperparameter tuning were:{'criterion': 'squared\_error', 'learning\_rate': 1.0, 'max\_depth': 10, 'max\_features': 'sqrt', 'n\_estimators': 100}

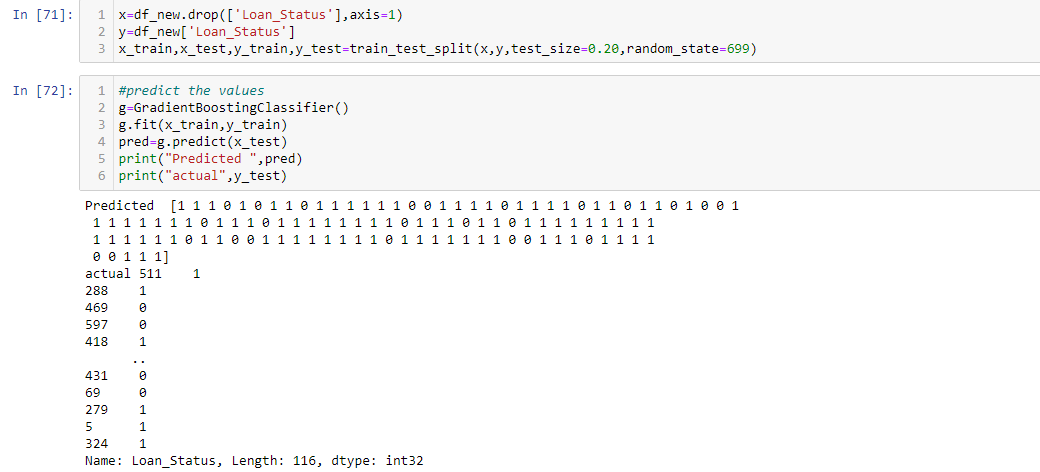


# AUC ROC curve





# splitting the data to Test





### Conclusion: we are getting GradientBoostingClassifier model accuracy score as 100% and cross\_val\_score as 76.7, so,we accept this model

# Saving the model

