# Loan Approval Prediction Machine Learning

# Introduction

Finance Companies biggest business is granting monitoring help as a term of LOAN. But as granting loan to any individual is considered as biggest risk of the business, hence all finance companies have very vast and different process to consider eligibility of the individual or company, that consist of Age, legal background, capability to repay the amount, etc. After all eligibility validation, still Finance Company doesn’t have assurance about individual will be able to repay the amount without difficulties.

However, checking eligibility for the individual is a very hectic and time consuming process. Hence here we are going to see how we can predict if a loan can be approved by considering historical data. This is type of classification problem to say if loan is approved or not. We can also refer it as an example for Predictive Modeling problem, where class label is predicted for given example of input of data.

# Problem Definition

Finance companies are vastly present across the all domain which provides many types of loans that consist of Home Loan, Personal Loan, Vehicle Loan, etc. As soon as customer applies for loan, a first step of process is to check if customer is eligible for the loan (asked amount). The company wants to automate the loan eligibility process on real-time based once customer provides details while filling out online process form. Available data set is as below

## Data-Set

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique ID |
| Gender | Male/Female |
| Married | Applicant Married (Y/N) |
| Dependents | Number of Dependant |
| Education | Applicant Education Qualification (Graduate/ Under Graduate) |
| Self\_Employed | Self-Employed (Y/N) |
| ApplicantIncome | Applicant Monthly Income |
| CoapplicantIncome | Co-applicant Monthly Income |
| LoanAmount | Loan Amount in Thousands |
| Loan\_Amount\_Term | Term of Loan in months |
| Credit\_History | Credit History meets guidelines |
| Property\_Area | Urban/Rural/Semi-Urban |

# Data Analysis

**1. Import Libraries** – First we will import important libraries.

**import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

**from** scipy.stats **import** norm

**from** sklearn.preprocessing **import** StandardScaler

**from** scipy **import** stats

**import** warnings

warnings**.**filterwarnings('ignore')

**%matplotlib** inline

**2. Reading Data –** By using pandas we import dataset.

*#creating Dataframe object*

df **=** pd**.**read\_csv('loan\_pridiction.csv')

df.head()

**3**. **Columns of dataset -**

df**.**columns

**output**- Index(['Loan\_ID', 'Gender', 'Married', 'Dependents', 'Education', 'Self\_Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount','Loan\_Amount\_Term', 'Credit\_History', 'Property\_Area' 'Loan\_Status'],

dtype='object')

**4. Size of dataset**-

df.shape

**Output**- (614,13)

**5. Data Types of Columns-**

df**.**dtype

**Output-**

Loan\_ID object

Gender object

Married object

Dependents object

Education object

Self\_Employed object

ApplicantIncome int64

CoapplicantIncome float64

LoanAmount float64

Loan\_Amount\_Term float64

Credit\_History float64

Property\_Area object

Loan\_Status object

dtype: object

There are 8 Catagorical columns and 5 Numeric columns.

**6. Null Values-** Finding null or missing values.

df**.**isnull()**.**sum(axis**=**0)

Loan\_ID 0

Gender 13

Married 3

Dependents 15

Education 0

Self\_Employed 32

ApplicantIncome 0

CoapplicantIncome 0

LoanAmount 22

Loan\_Amount\_Term 14

Credit\_History 50

Property\_Area 0

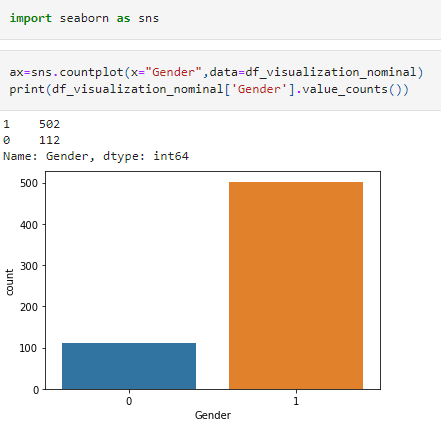
Loan\_Status 0

dtype: int64

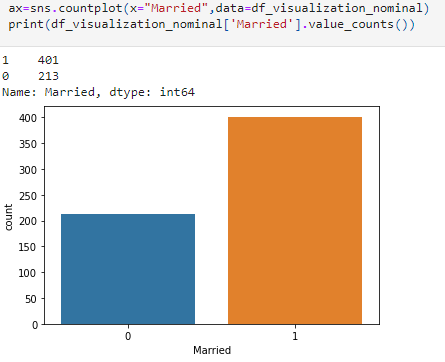
There are null values present in column -Gender,Dependents,Self\_Employed,LoanAmount,Loan\_Amount\_Term,Credit\_History. To handle this we find mean and Mode method.

EDA(Exploratory Data Analysis)

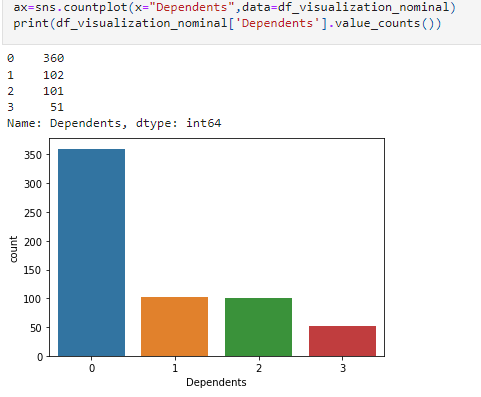
# 1. Visualization of Data-



From the above observation the total number of male(0) is 112 and female is 502.



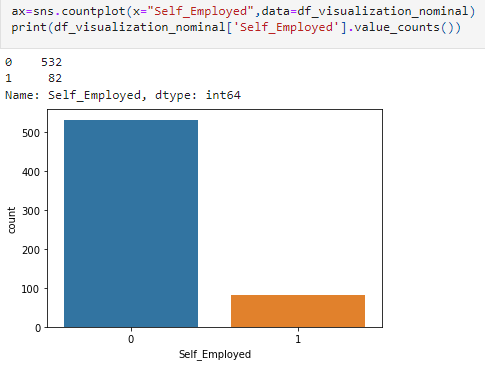
From the above observation the total number of Married person is 213 and Unmarried is 401.



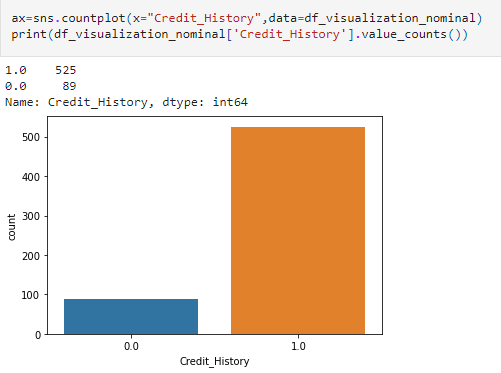
In total data 360 Persons has no dependency.



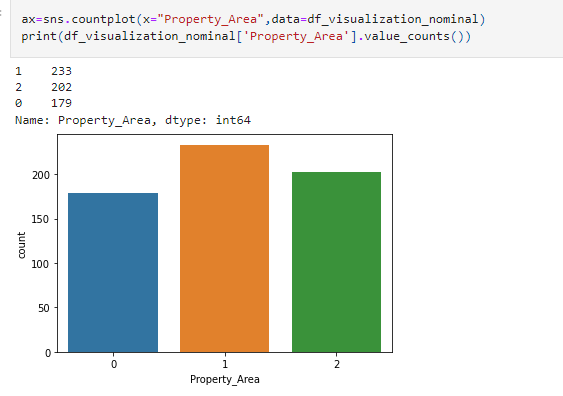
480 persons are Graduate and 134 are Non Graduate.



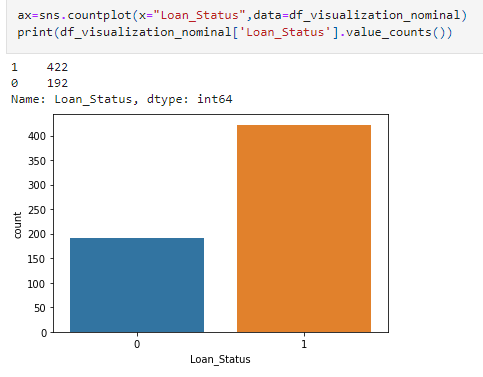
532 persons are not self employed and only 82 are employed.Credit\_History', 'Property\_Area', 'Loan\_Status'



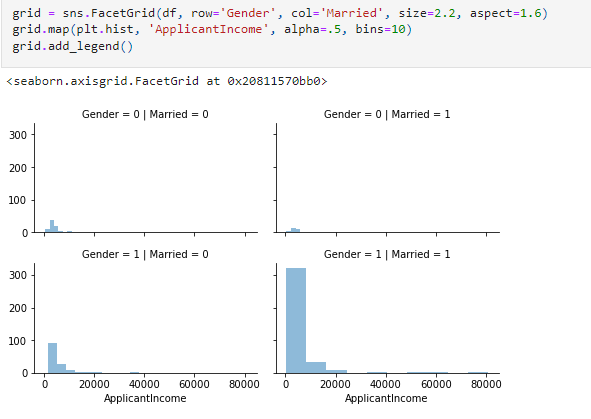
525 Person has credit history .



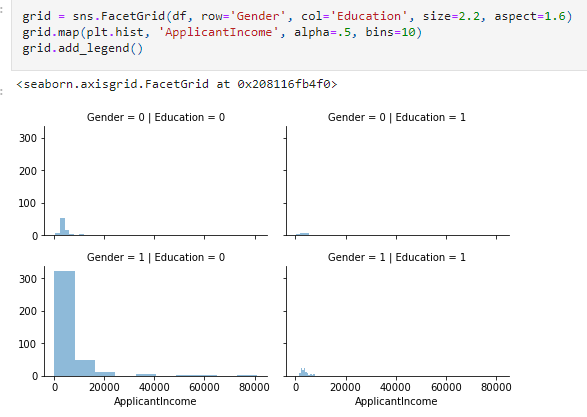
People from rural area are 179, from semirural is 233 and from urban is 202.



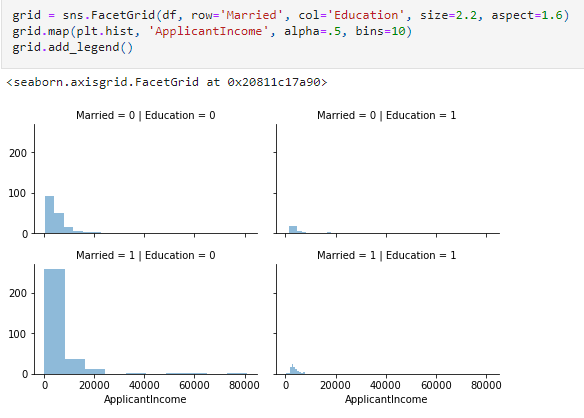
Persons whom loan are given is 422 and not given is 192.

****

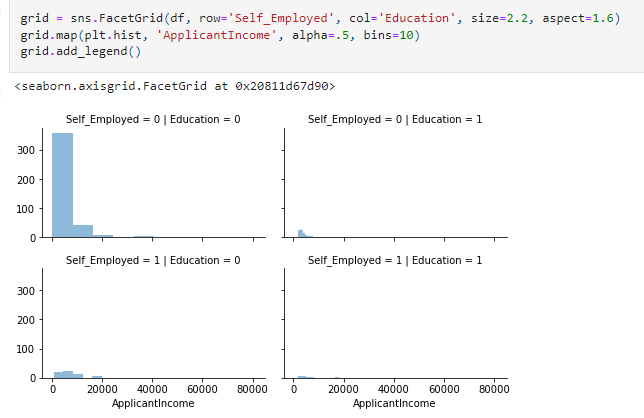
From this Most of the Person Who applied for loan are Male and Merried.



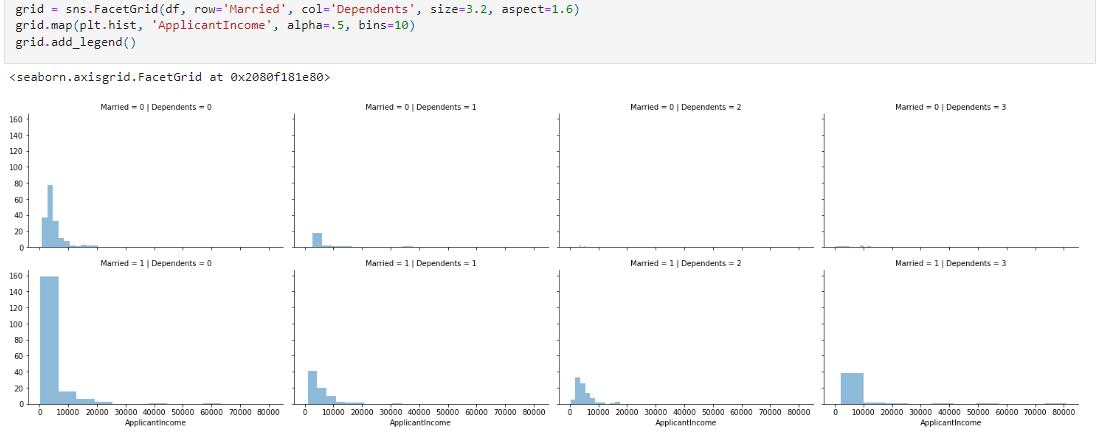
Most of the Person who applied for loan are Male and Educated.



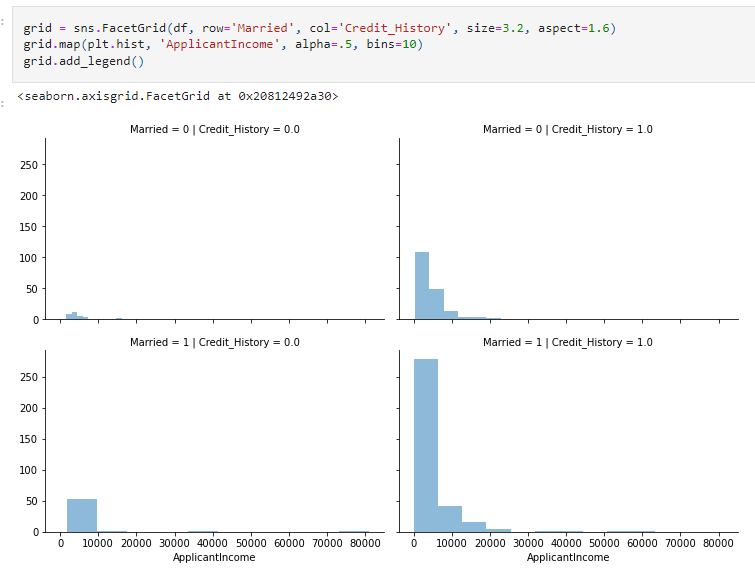
Most of the Person who applied for Loan are Married and Educated.



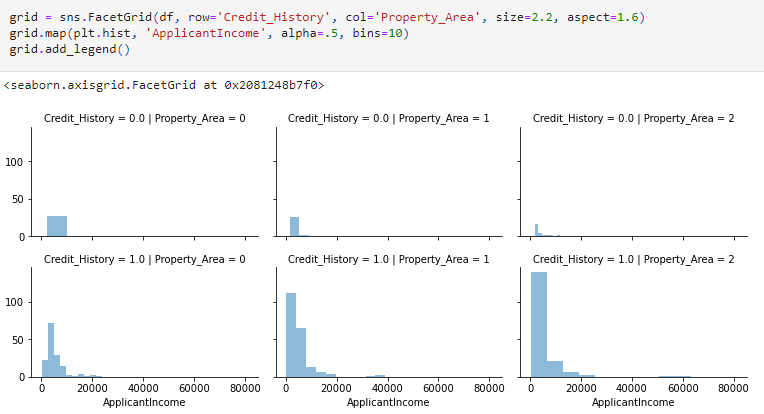
Most of the person who applied for income are not Self Employed but Educated.



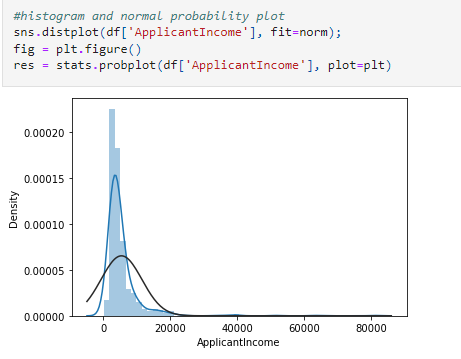
Most of the person who applied for Loan are merried and mostly have no dependents or some have 1,2,3 dependents.

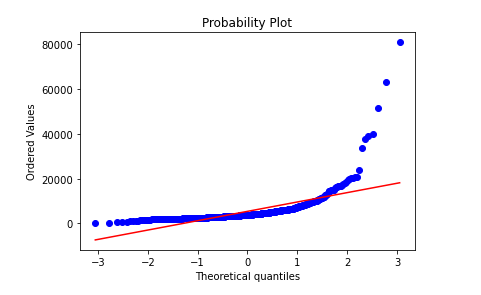


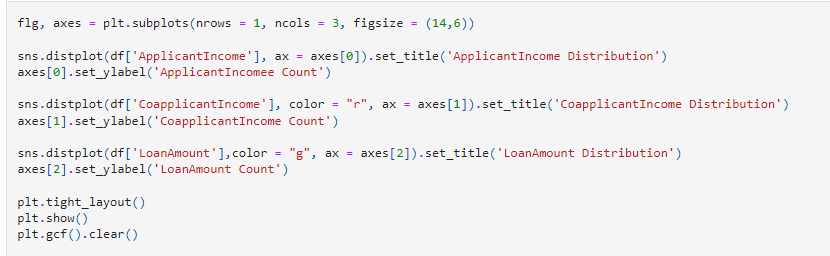
Most of the person who applied for loan are Merried and has credit history.

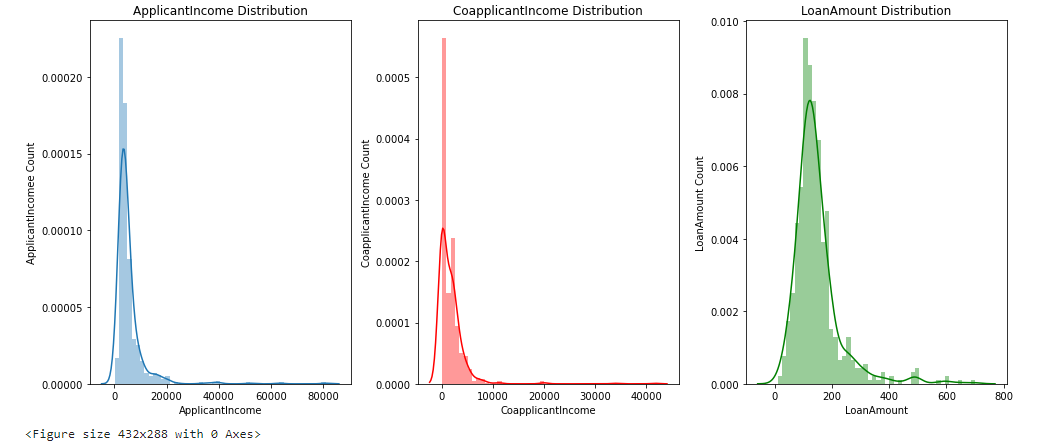


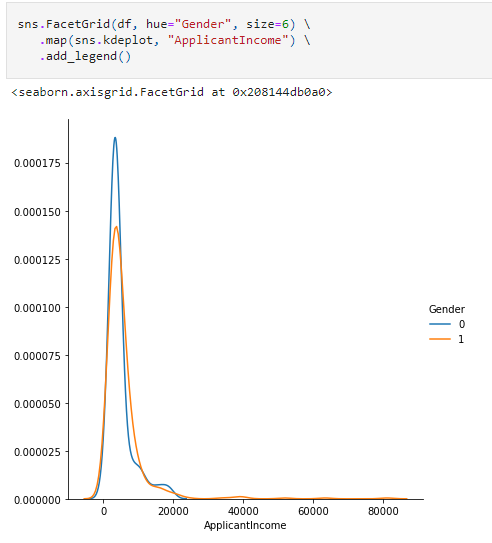
Most of the Person has Credit history and Property area

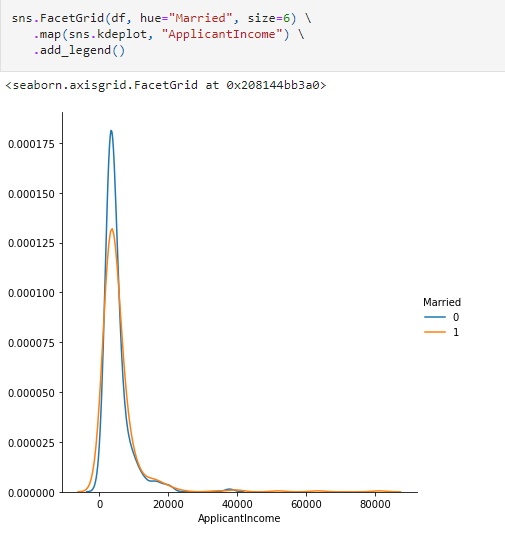


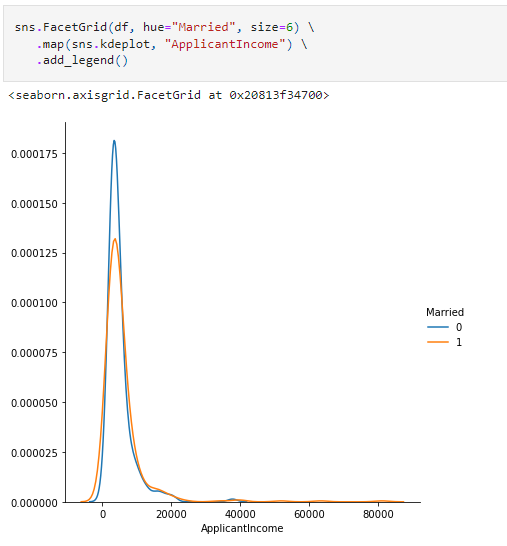


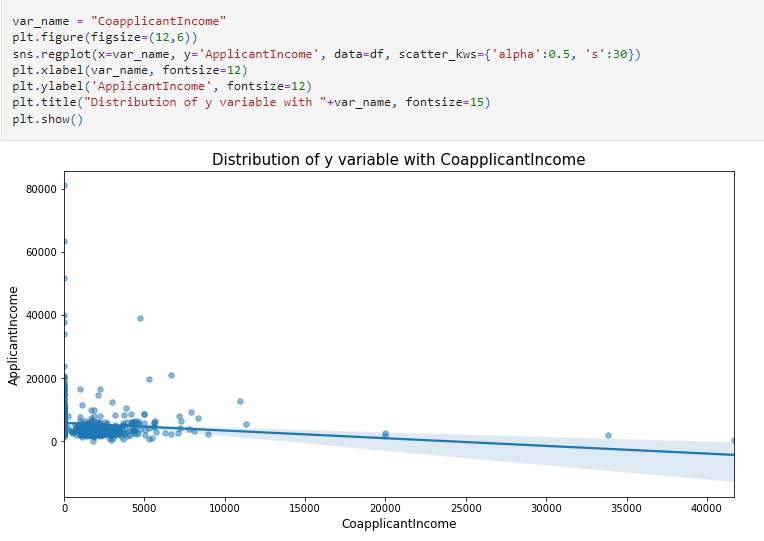


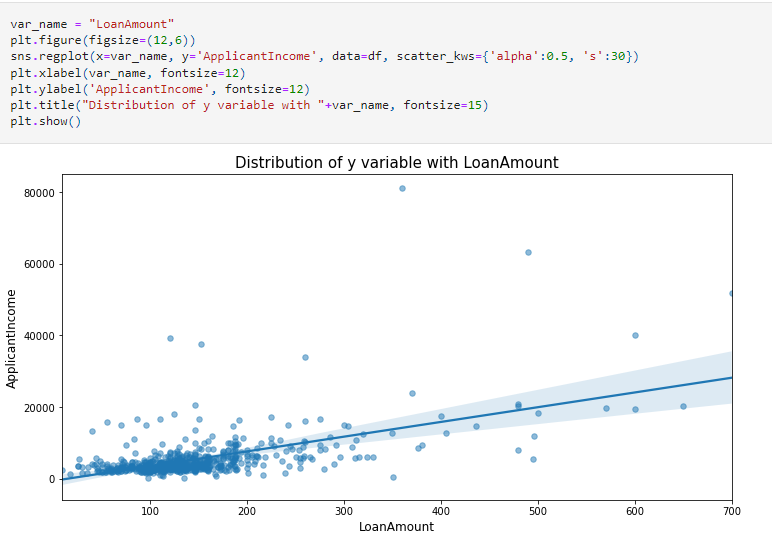












**Remarks-**

The loan of 422(around 69%) people out of 614 were approved.

More Loans are approved Vs Rejected

Count of Male applicants is more than Female

Count of Married applicant is more than Non-married

Count of graduate is more than non-Graduate

Count of self-employed is less than that of Non-Self-employed

Maximum properties are located in Semi urban areas

Credit History is present for many applicants

The count of applicants with several dependents=0 is maximum.

80% of applicants in the dataset are male.

Around 65% of the applicants in the dataset are married.

Around 15% of applicants in the dataset are self-employed.

Around 85% of applicants have repaid their doubts.

Most of the applicants don't have any dependents.

Around 80% of the applicants are Graduate.

Applicants with high incomes should have more chances of loan approval.

Applicants who have repaid their previous debts should have higher chances of loan approval.

Loan approval should also depend on the loan amount. If the loan amount is less, the chances of loan approval should be high.

Lesser the amount to be paid monthly to repay the loan, the higher the chances of loan approval.

The proportion of married applicants is higher for approved loans.

Distribution of applicants with 1 or 3+ dependents is similar across both the categories of Loan\_Status.

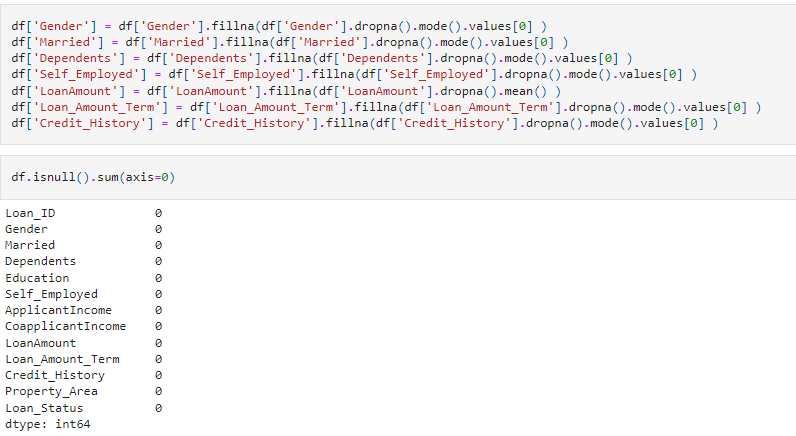
There is nothing significant we can infer from Self\_Employed vs Loan\_Status plot.

It seems people with a credit history as 1 are more likely to get their loans approved.

The proportion of loans getting approved in the semi-urban area is higher as compared to that in rural or urban areas.

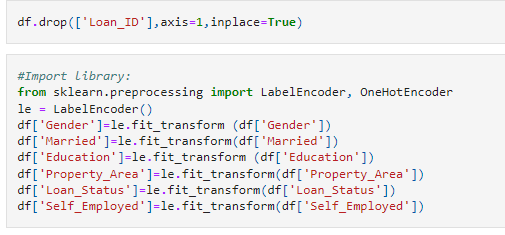
## Pre-processing Pipeline

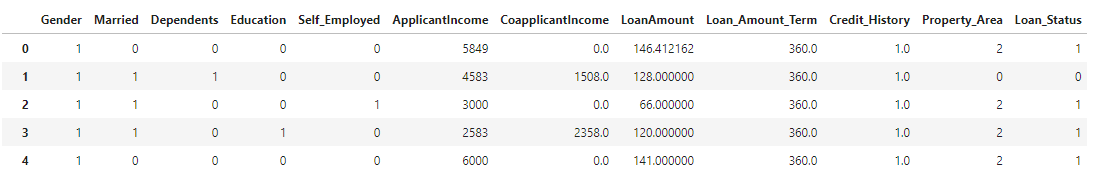
**1. Filling Null Values-**

****

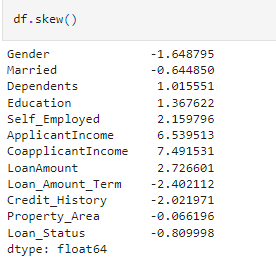
Now there is no null values present.

**2. Now we will replace the catagorical data of sex and Embarked column into numeric data-**





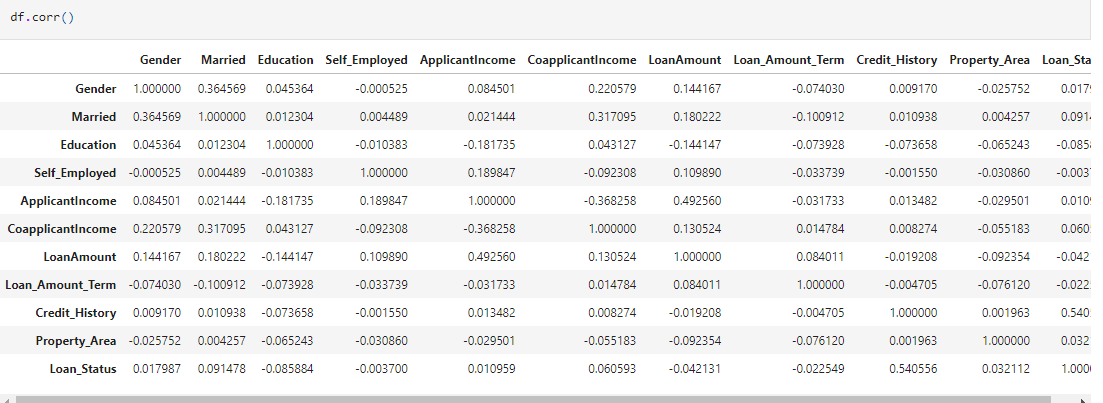
# 3. Skewness-



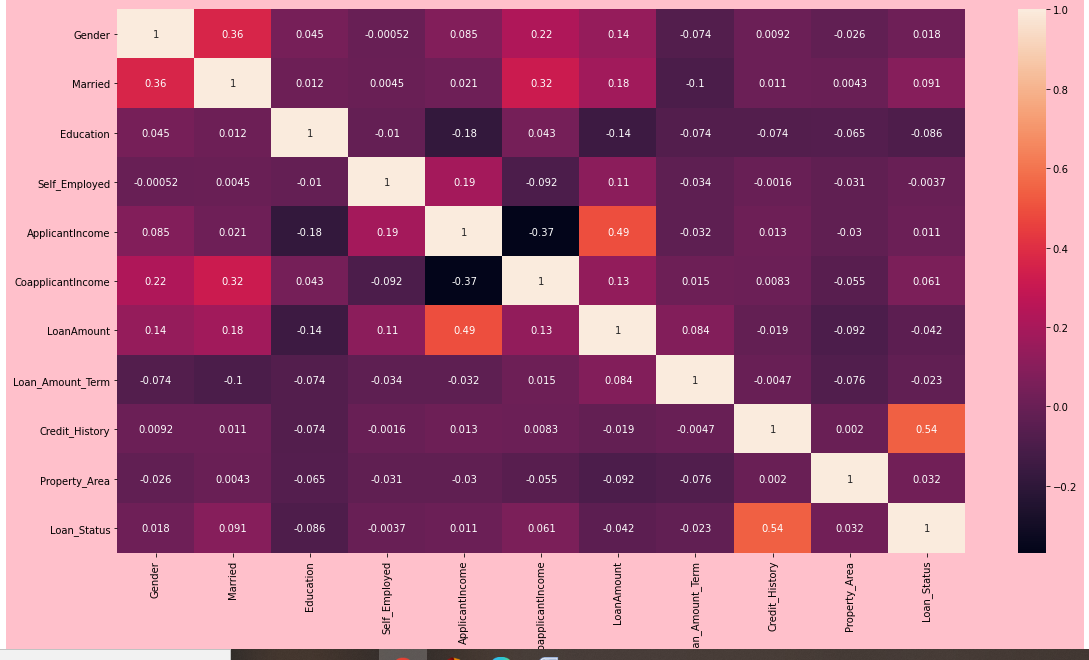
We are first calculating the skew value and some of the column skew value are far from zero. -The best skew value for normally distributes is very close to zero, so we are using “log1p” method to make the skew value near to zero.



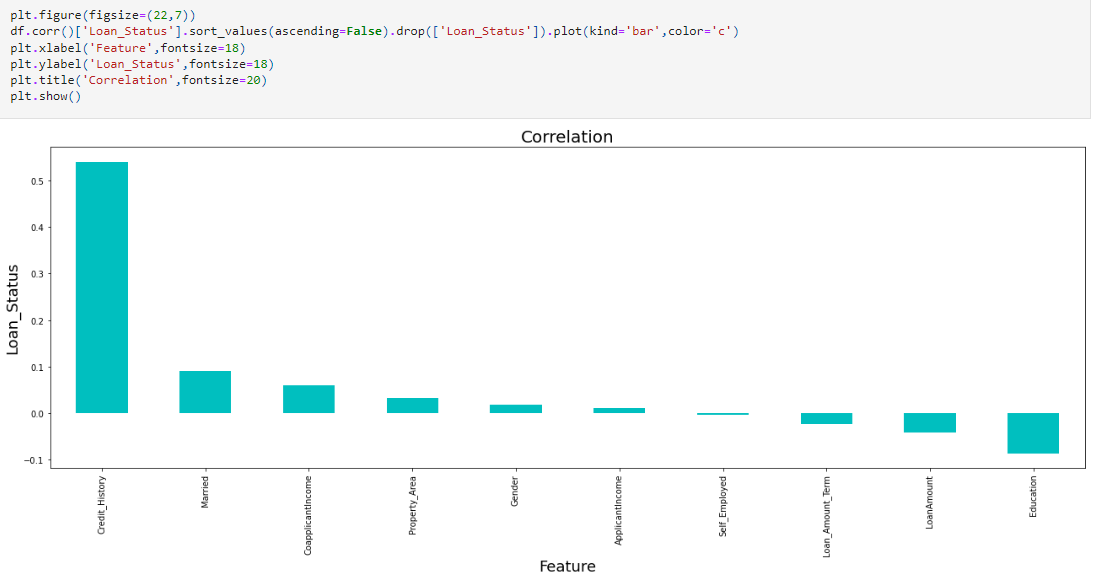
# 4. Correlation



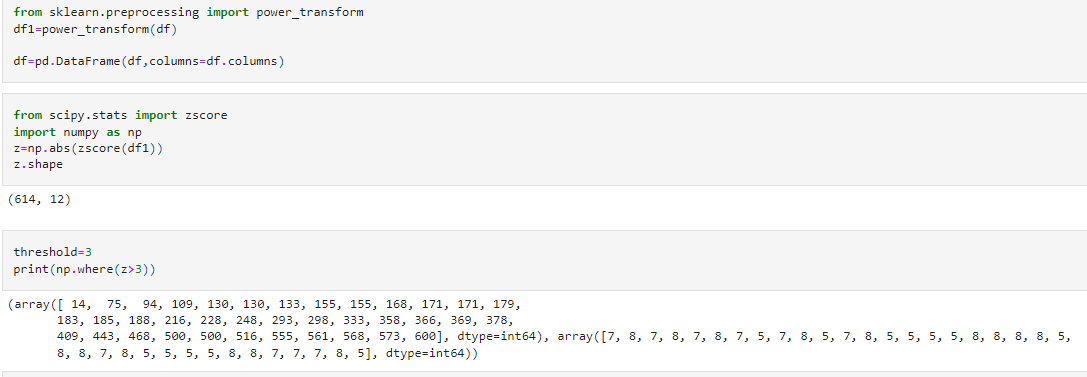
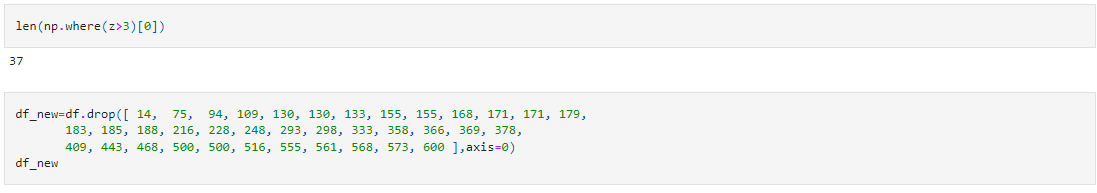


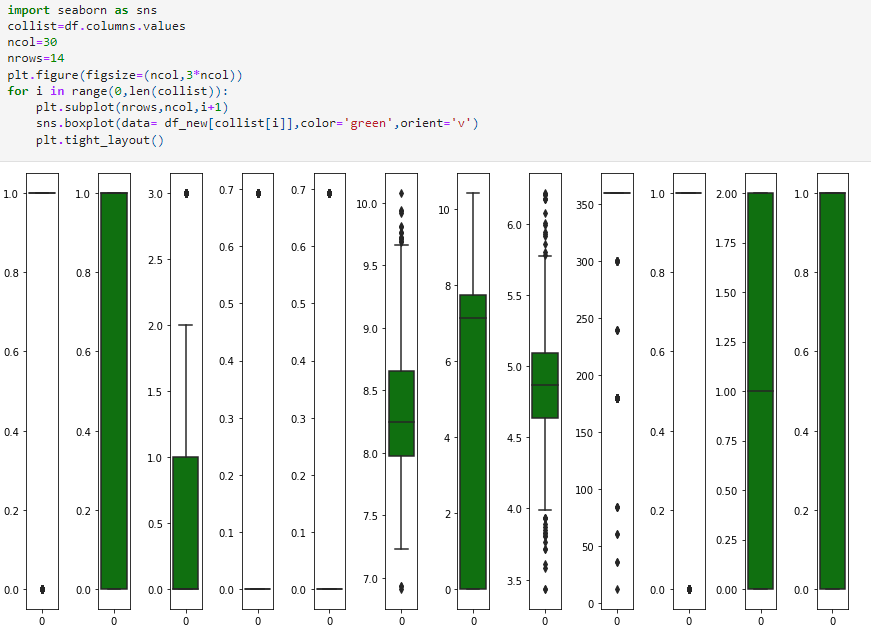


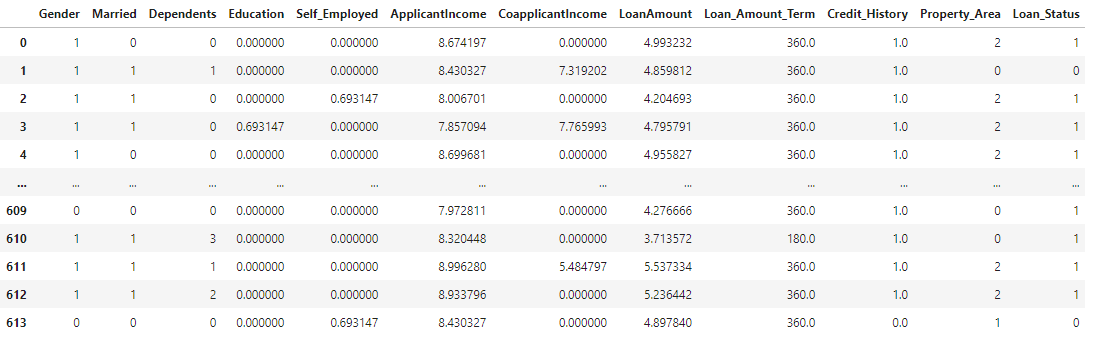
By Correlation we will find the relation between different columns to target column. Credit History and Loan Status column are highly correlated. Education and Loan Status column are not correlated with each other.

By this we can say that Credit\_History, married, CoapplicantIncome, PropertyArea, Gender, ApplicantIncome and Self\_Emplyed are in positive relation and Loan\_Amount\_Term, LoanAmount and Education are in negative relation.

**5. Outliers-**

****





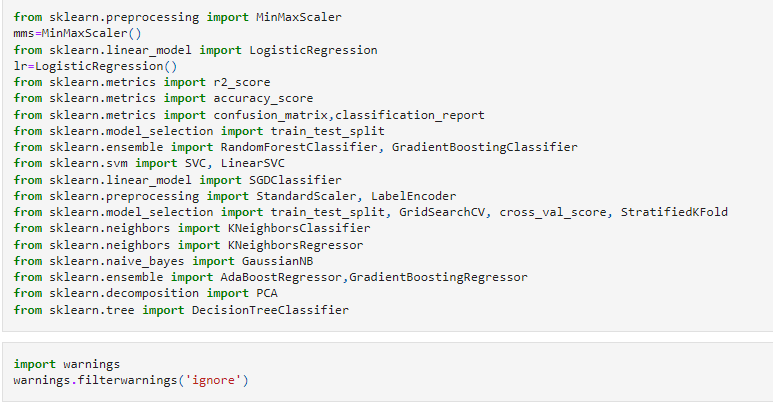
Very Few Outliers are present in LoanAmount and ApplicantIncome. So we can proceed now.

## Building machine Learning Models

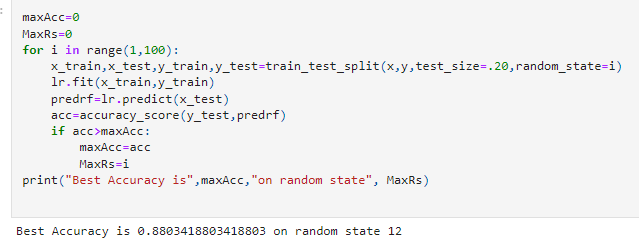
**1. Assign values to x and y-**



**2. Importing Libraries**-



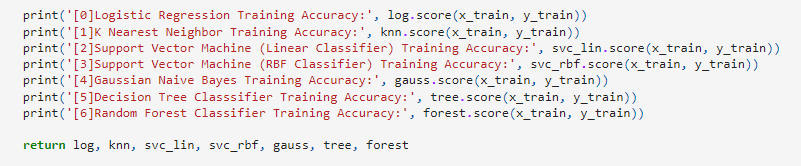
**3. Finding Best Random State-**

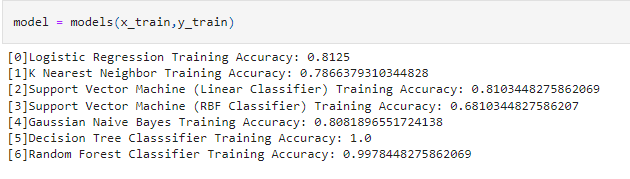
****

We saw that we get 88% accuracy at random state 12.

**4. Training the Model-**

****

****

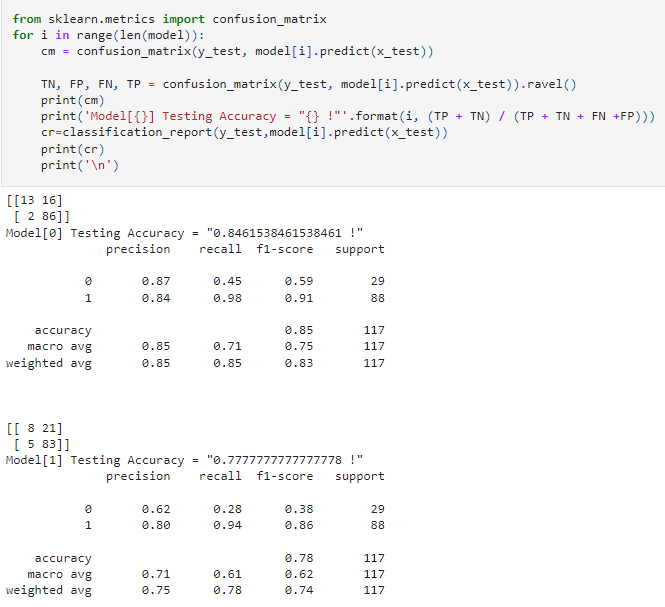
****

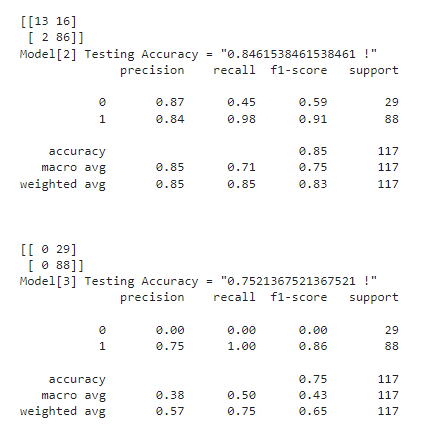
We train our model with different regression technique like Logistic Regression, K Nearest Neighbor , Support

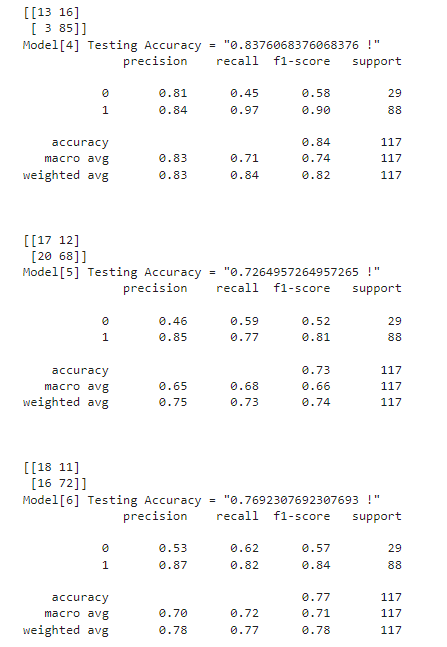
Vector Machine , Gaussian Naïve Bayes < decision Tree , Random Forest.

After that we will test our Model so that we can find out which model is giving best accuracy.

**5. Testing the model-** Confusion Matrix

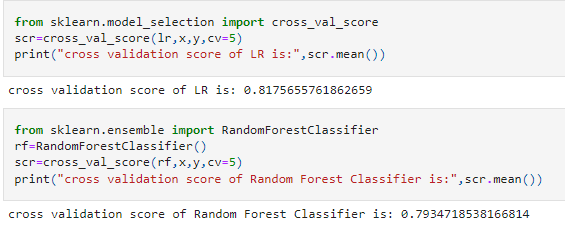
****

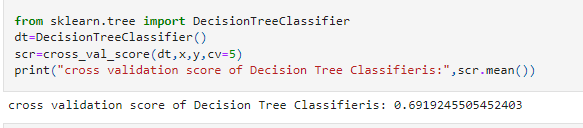


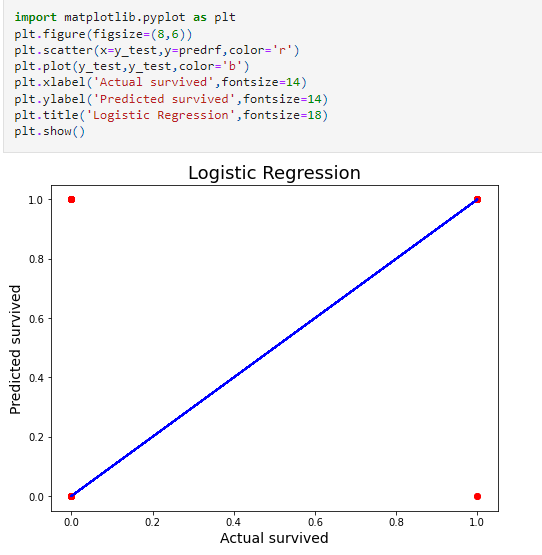
****

When we test with different regressor we find accuracy between 72 to 84. After that we will do Cross Validation for validate our model is good or not.

**6. Cross Validation –**

****



****

The Cross Validation score of Decision tree is 69% and random Forest is 80%. Which shows that our model is good.

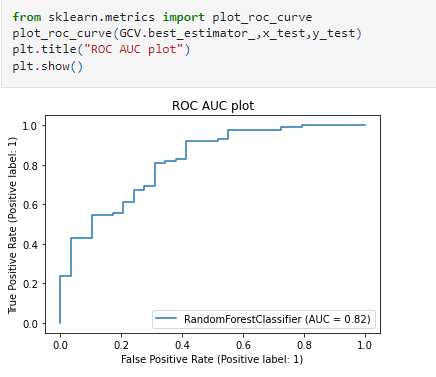
**7. Hyper Tunning Parameter-**

After getting an accuracy of 82.7% I tried fine-tuning it to improve my accuracy score using GridSearchCV.

****

After fine-tuning the Random Forest model the accuracy score improved from 79% to 84.61%

**8. AUC-ROC Curve-**

****

The ROC AUC plot is 82% .

## Concluding Remarks

**Conclusion-**

After the Final Submission of test data, my accuracy score was 84%. Feature engineering helped me increase my accuracy. Random Forest Model worked better than all other Ensemble models. Cross Validation Score was also 80%. Which shows that my Model is quite good. AUC Curve is also 82% fit. I hope my model works well.