**Loan Application Status Prediction**

Loan application status prediction is for banks or financial organisation who lent the loan to people on various basis. Banks or financial organisation have large number of loan applicants and it’s been very difficult for them deciding which applicant is capable of repaying the loan and who can be defaulter. In this dataset we have loan applicants with their various info about them and we have to build a model predicting the applicant can be defaulter or not.

**Problem Definition:**

Loan application status prediction dataset includes details of applicants who have applied for loan. As it depends on numerous factors on which the applicant is married, employment status, the amount of loan applicant is applying for, previous credit history, property belonging to applicant etc. Such cases are often classification problems.

This dataset includes details like:

**(Independent variable)**

- Loan ID

- Gender

- Married

- Dependents

- Education

- Self Employed

- Applicant Income

- Co-applicant Income

- Loan Amount

- Loan Amount Term

- Credit History

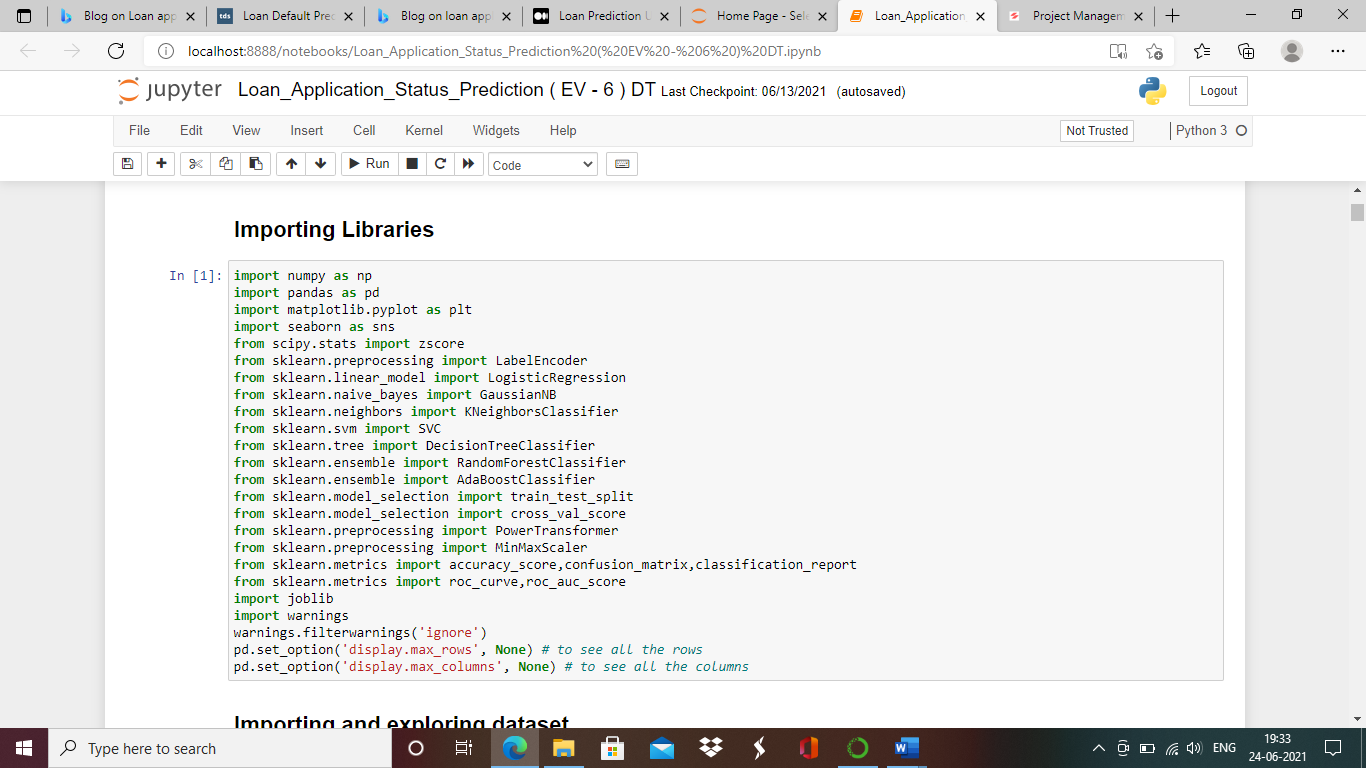
- Property Area

**(Target Variable)**

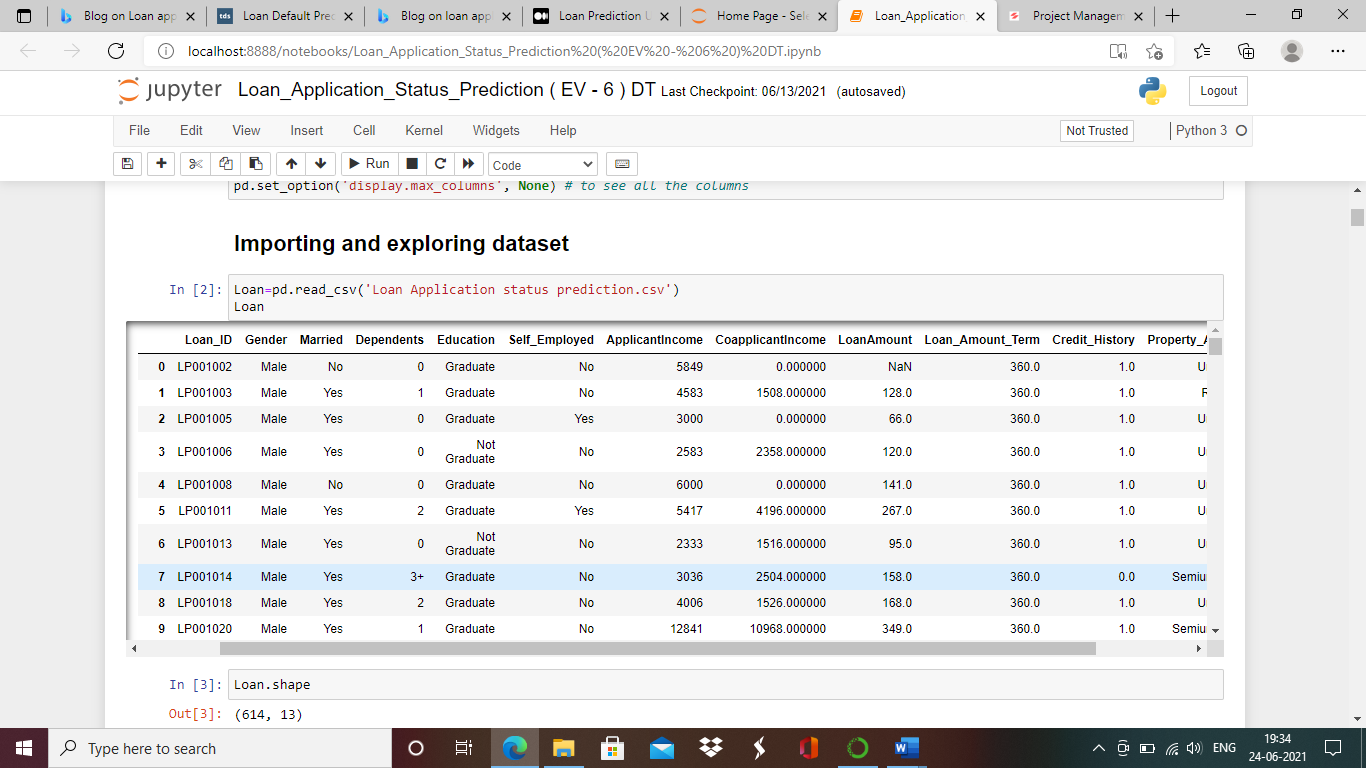
-Loan Status

**Data Analysing**

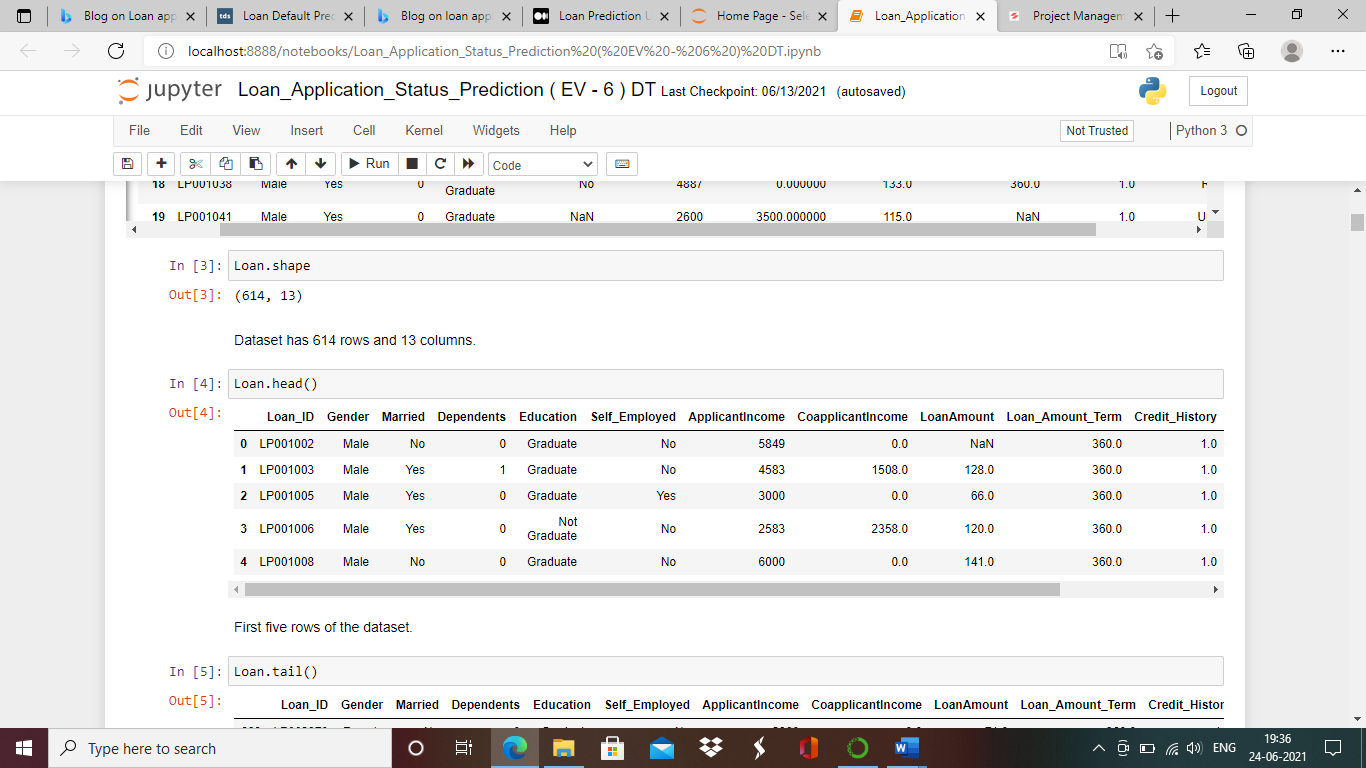
Initially we will call import all the libraries required to run this dataset.



Now we call the data set for analysing .

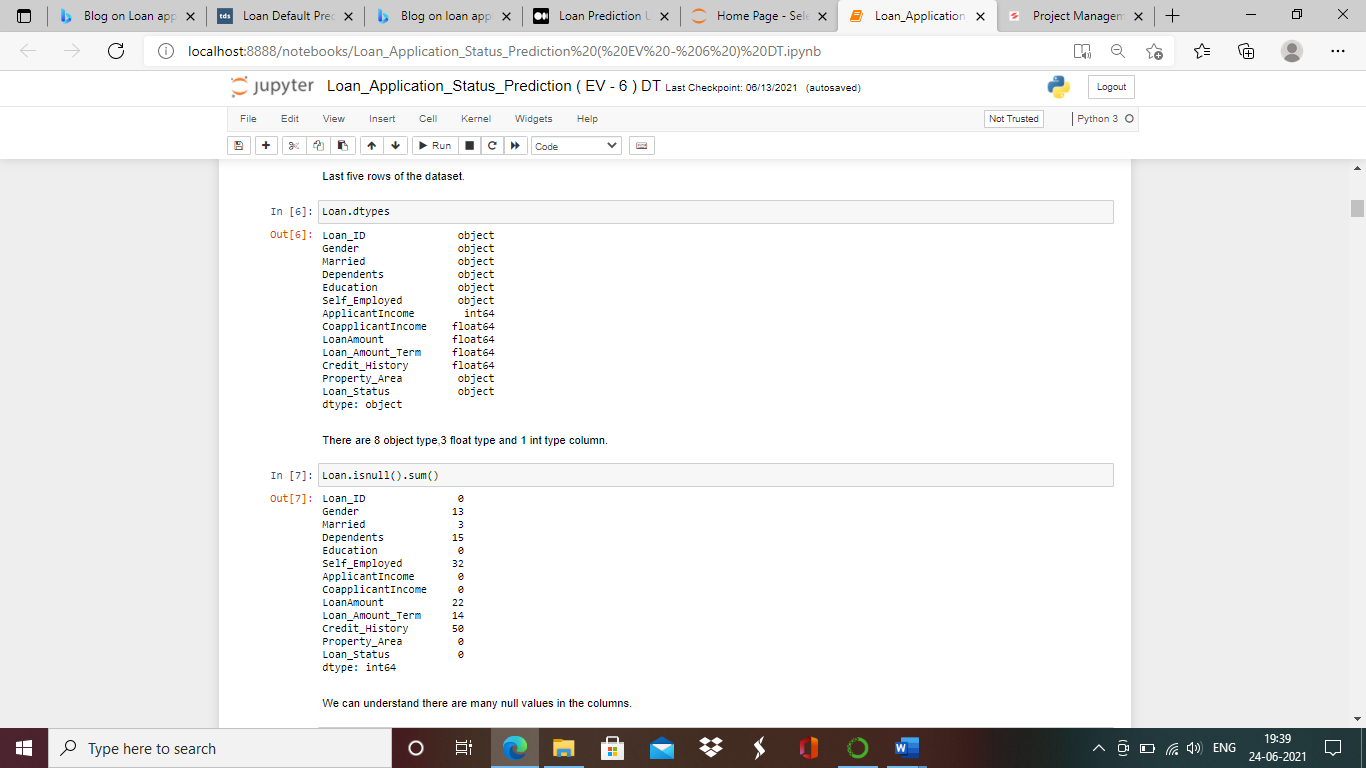


We will call this dataset as Loan for futher operation ,lets check the number of columns and rows in this dataset.



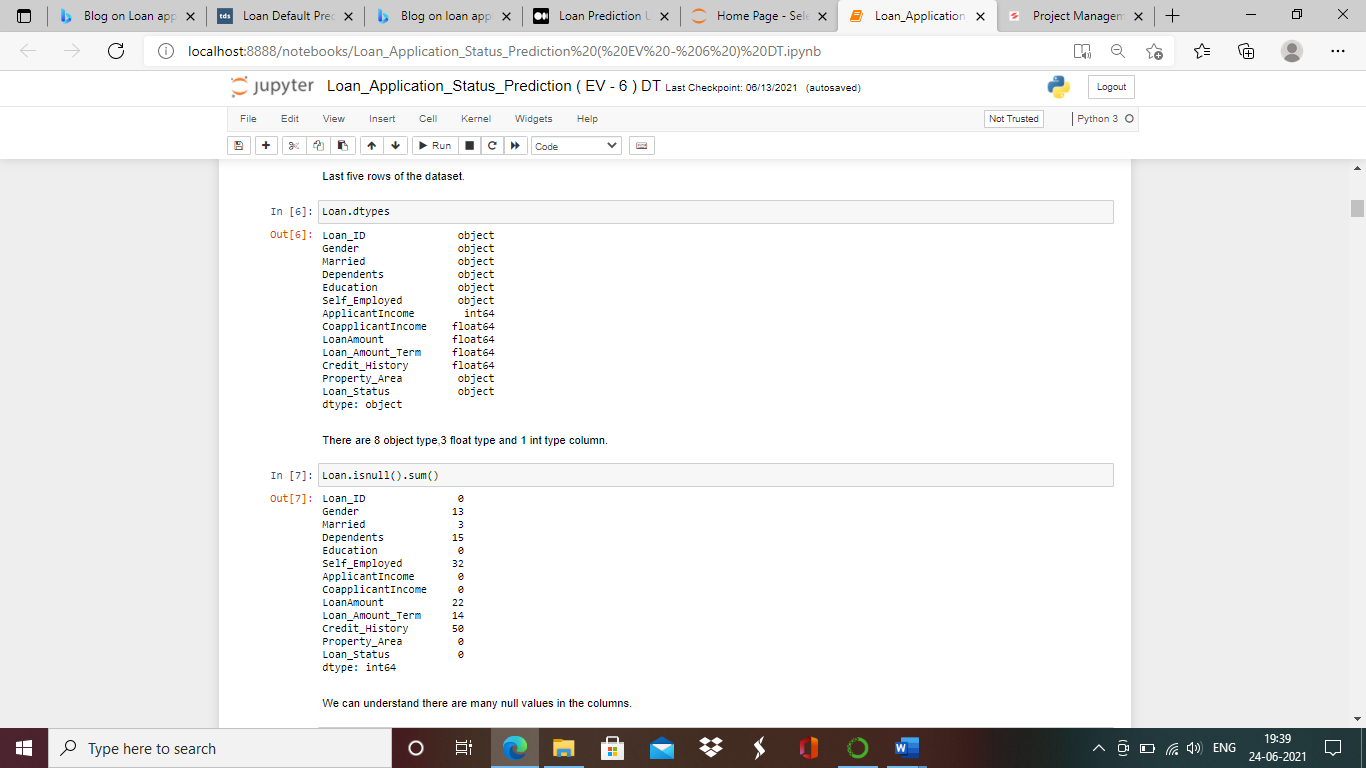
There are 614 rows and 13 columns in loan dataset ,means there are 614 applicants.

We will check the data type of the columns .



Most of the columns in dataset are categorized and object type. Coaplicant income,Loan Amount,Loan Amount term,Credit history are in float and only Applicant income is integer type.

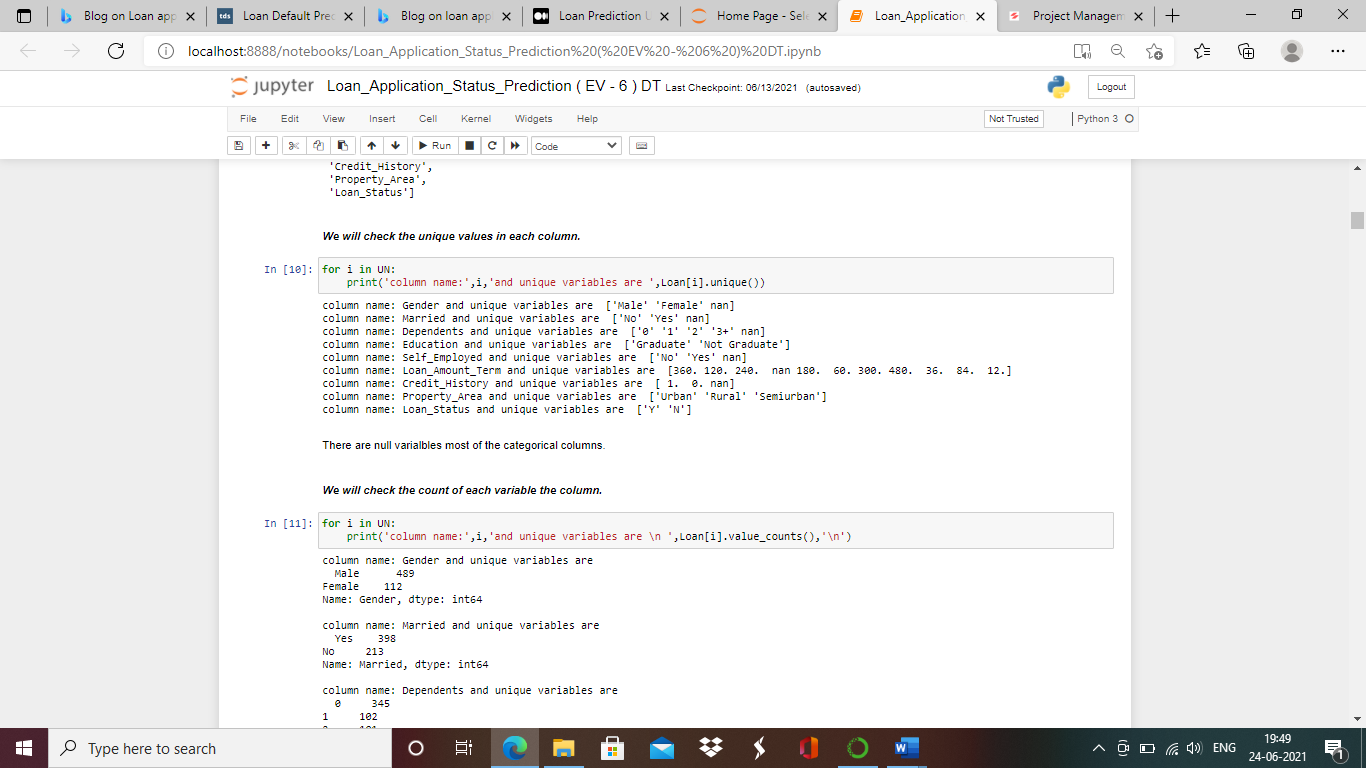
Check if Null values are present in dataset.



Looking at categorised column there are 13 null vlaues in Gender column,3 null values in married column,15 in dependents column,32 in self employed column,14 in loan amount term and 50 in credit history column.

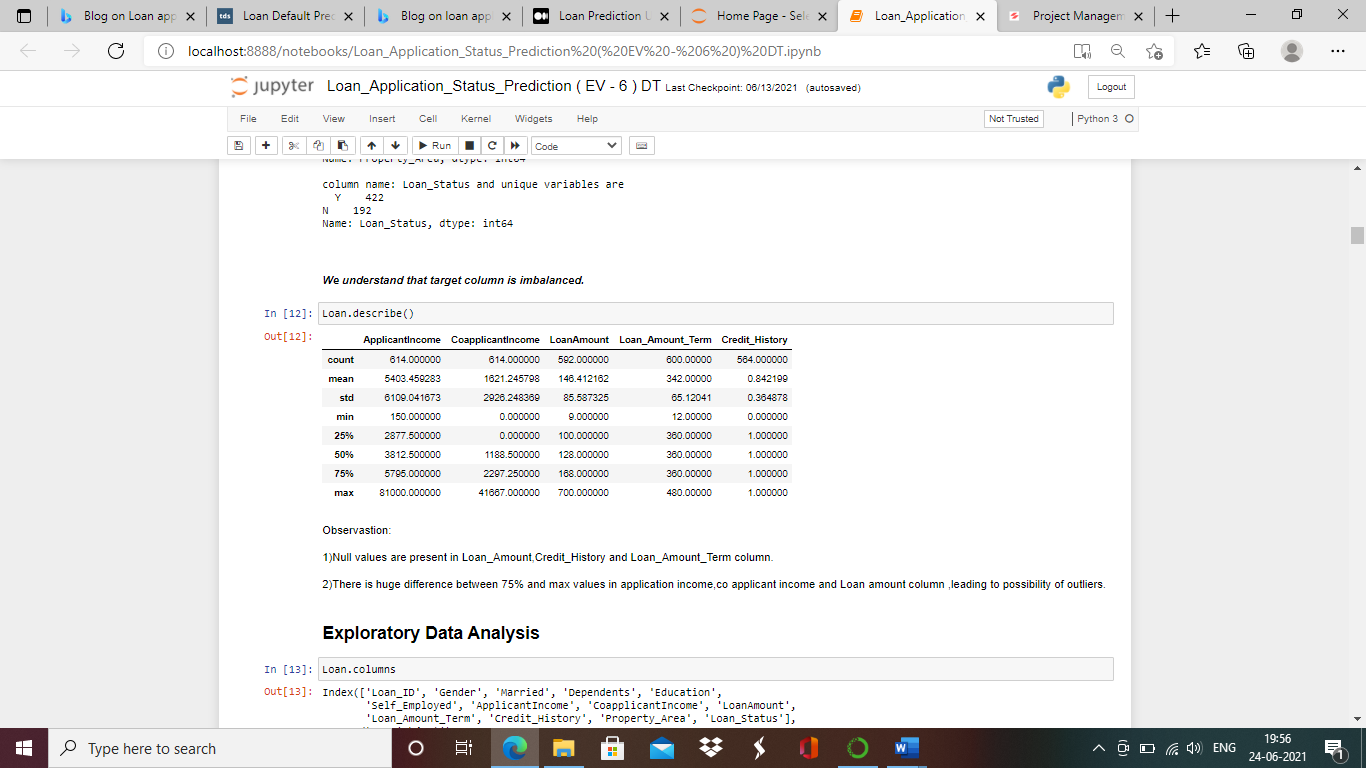
In continours variable columns we can see null values only in Loan amount column.

**Check Unique variables in categorised columns.**



We can understand from above the unique variables in column,columns like property area defines whether the applicants property is in rural,urban or semiurban are.Self employed states whether applicant is doing service or is self employed.Credit history makes us understand whether applicant has an credit history in past ( 1: Yes and 0: No ).

**Lets get statistical understanding of Loan dataset for numerical columns**



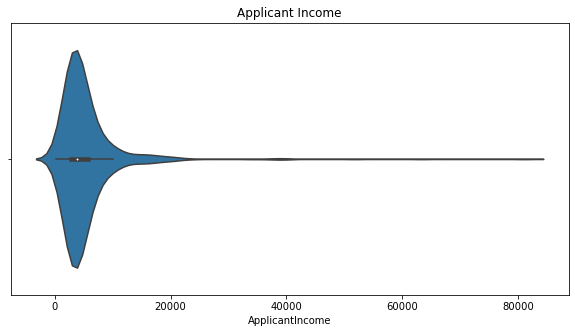
Observations:

1)Null values are present in Loan Amount, Credit History and Loan Amount Term column.

2)There is huge difference between 75% and max values in application income, co applicant income and Loan amount column, leading to possibility of outliers

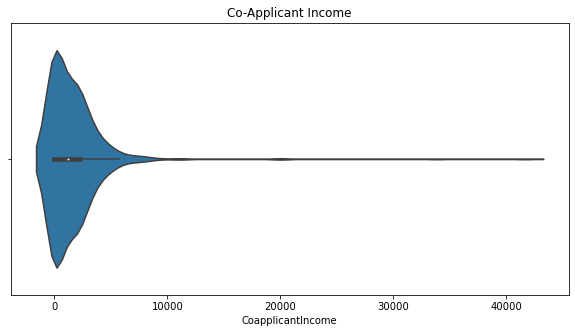
**Exploratory Data Analysis**

**Univariate Analysis**

****

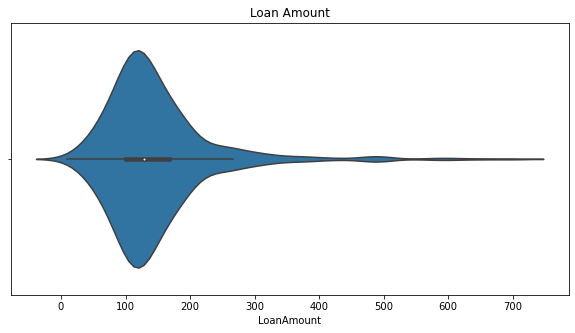
Observation:

There are more people with applicant income o to 10000



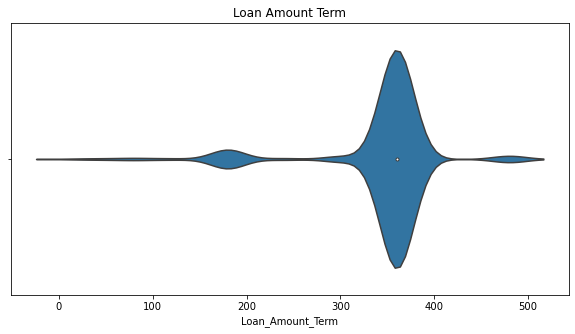
Observation:

Co-applicant income ranges majorily in range of 0 to 5000.



Observation:

Majority of people are looking loan from 50 to 200.

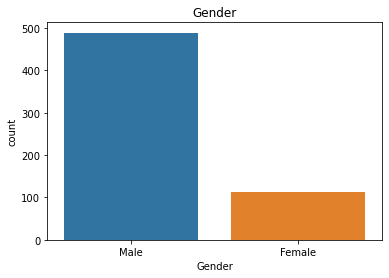


Observation:

1)Loan amount term is majorly from 300 to 400.

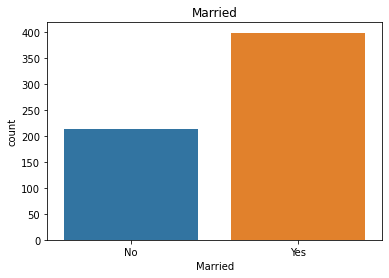
2)Few people are also applying loan amount term from 150 to 200.

3)Very few applicant are also looking for loan amount term from 450 to 500.



Observation:

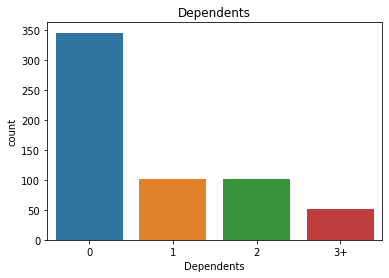
Male candidates are more than females for loan application



Observstion:

Married candidates are more compared to unmarried for loan application

.

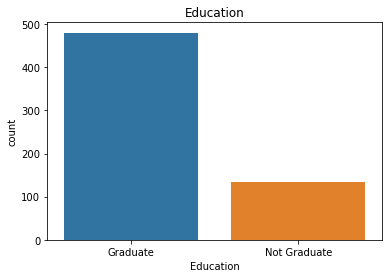


Observastion:

1)Almost 350 applicant have no dependents.

2)Applicant with 1 and 2 dependents are almost same around 100.

3)Few applicant have 3+ dependents.



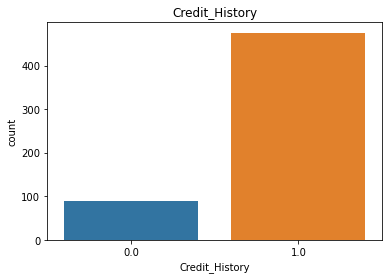
Observation:

1)Close to 500 applicants are graduate and very few approx:120 are not graduate.



Observation:

1)Around 500 applicants aren't self-employed reaming approx. 100 applicants are self-employed.



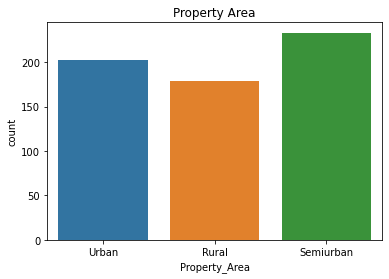
Observation:

1)Applicants with credit atleast once are in majority of numbers.



Observation:

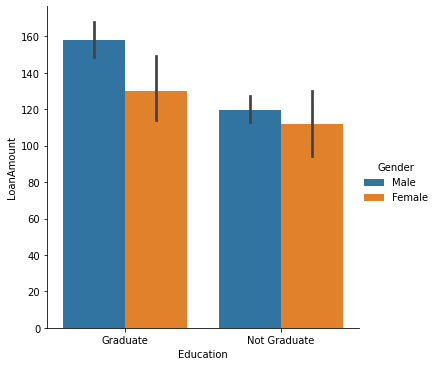
1)Loan application for most of the applicant has been approved and few have been rejected.



Observation:

1)Most applicants have property in semiurban area, approx. (200) applicants have property in urban.

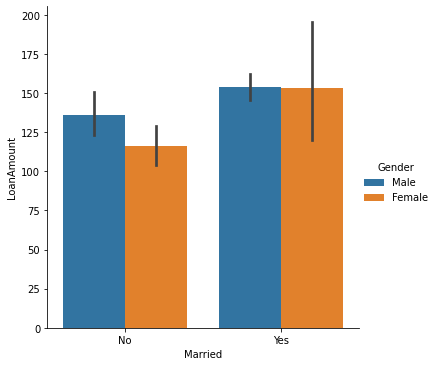
**Bivariate Analysis**



Observation:

1)In application of higher loan amount, male applicants are more in both graduate and non-graduate list.

2)Female applicants apply for lesser loan amount compared to male applicants, irrespective of qualification.



Observation:

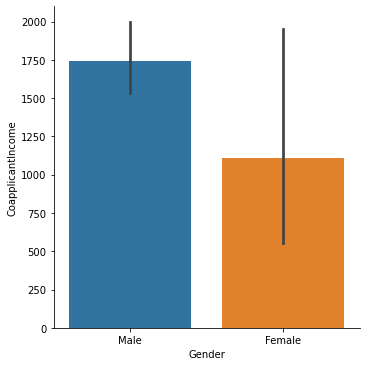
1)Married applicants apply for higher loan amount compared to unmarried.

2)In Unmarried category male applicant apply for higher loan amount.



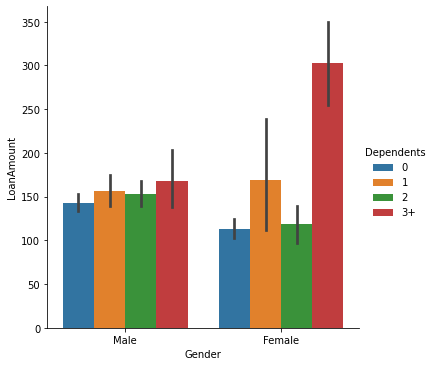
Observation:

1)Male applicants income is higher than female applicant.



Observation:

In co-applicant’s male have higher income than females.

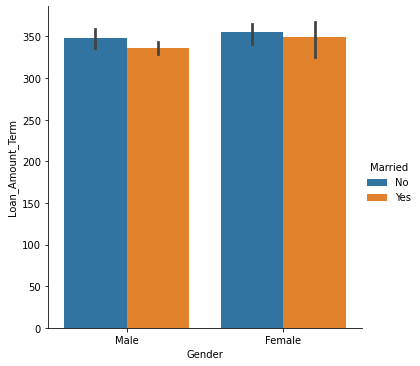


Observation:

1)Females and makes with 3+ dependents have applied for higher loan amount.

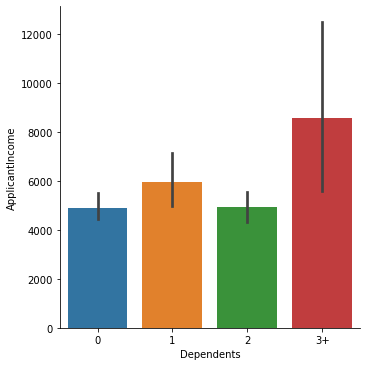
2)Males and females with 1 or 2 dependents have applied for almost similar loan amount.

3)Males with no dependents have applied for more loan amount compared to females with no dependents.



Observation:

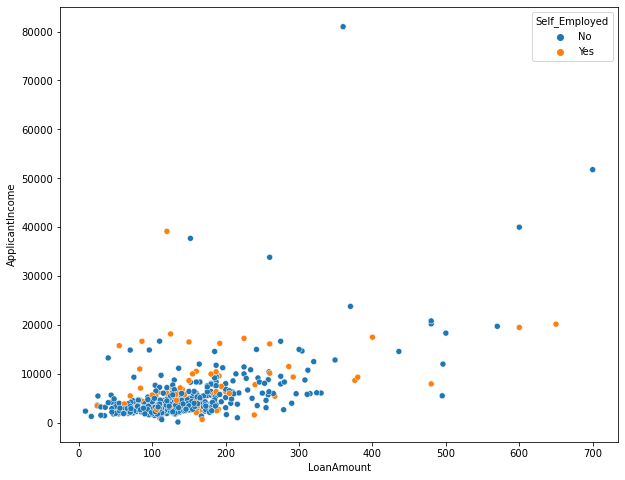
Loan amount term is almost similar in males and females. Unmarried males tend for bit higher loan amount term.



Observation:

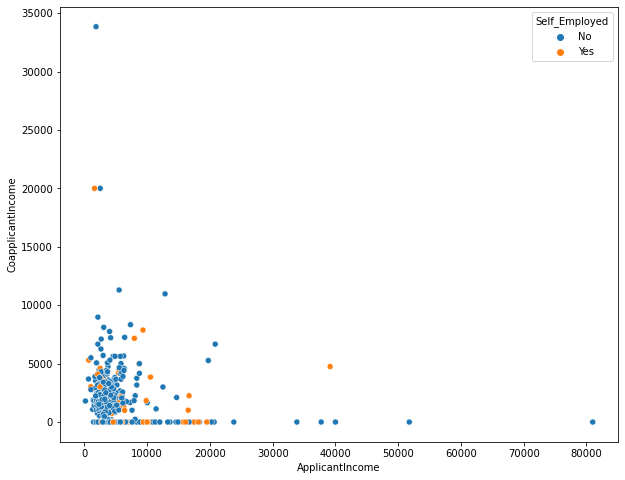
Applicant with 3+ dependents have highest income.

Applicant with 0 or 2 dependents have similar income.



Observation:

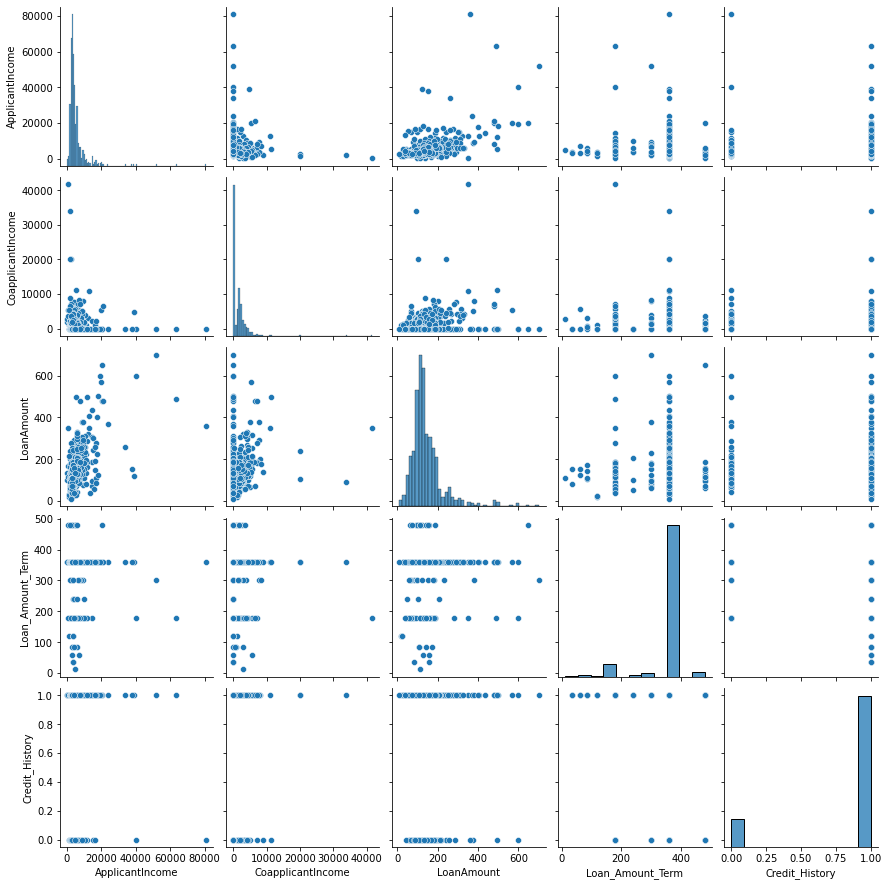
Loan amount and applicant is positively correlated and not self-employed applicants are more



Applicant income and co-applicant income are positively correlated.

There are more non-self-employed applicants

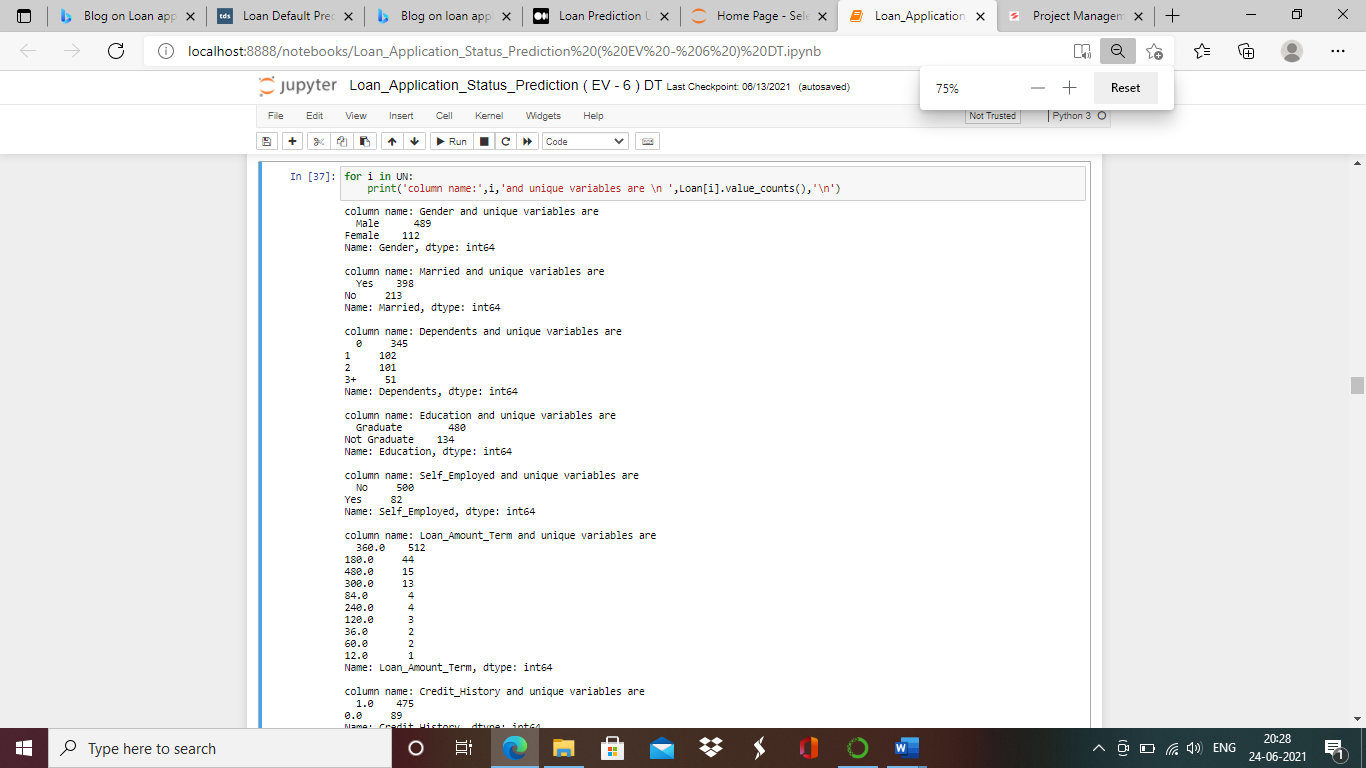
**Multivariate Analysis**



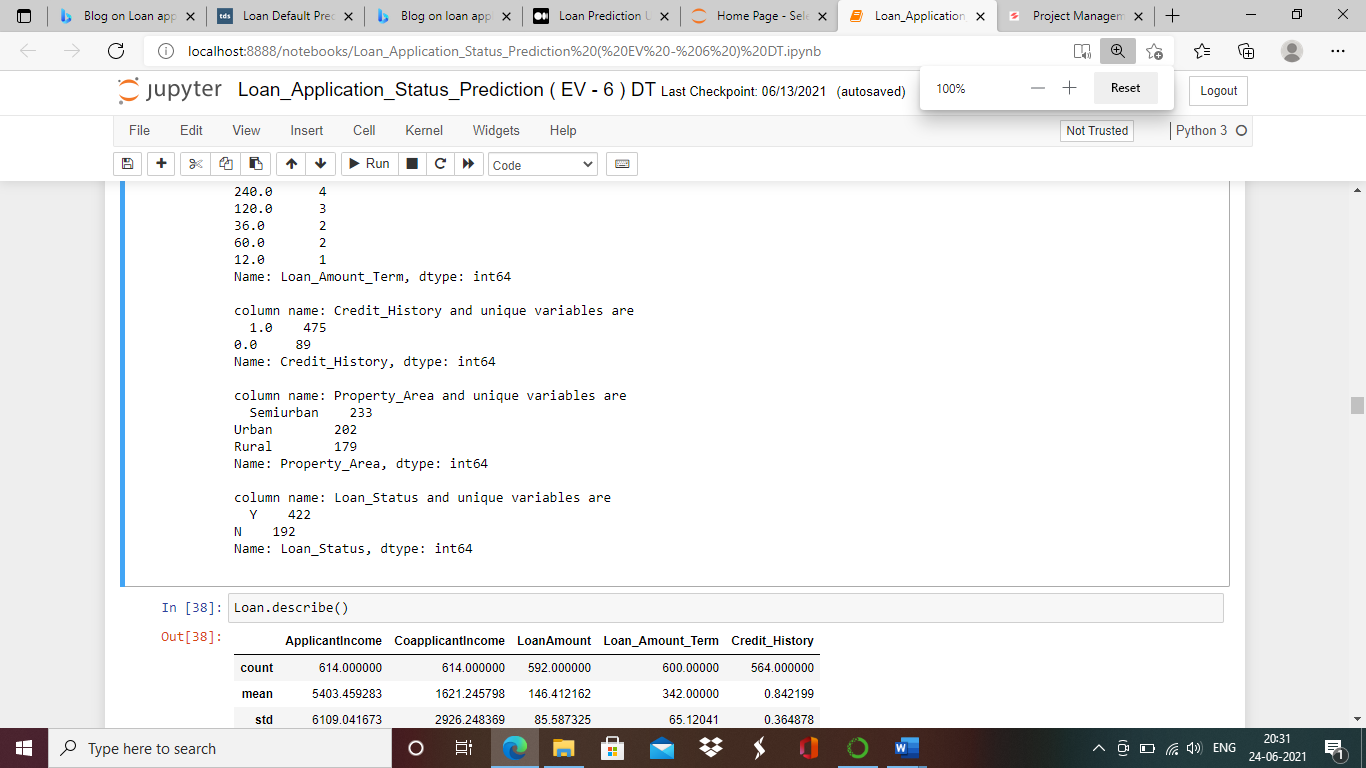
**Pre-Processing Pipeline**

EDA process has helped us to understand the data much better. Now, let’s treat the missing values in dataset.

Earlier we had checked the unique variables in categorized columns, now lets check the count of these variables



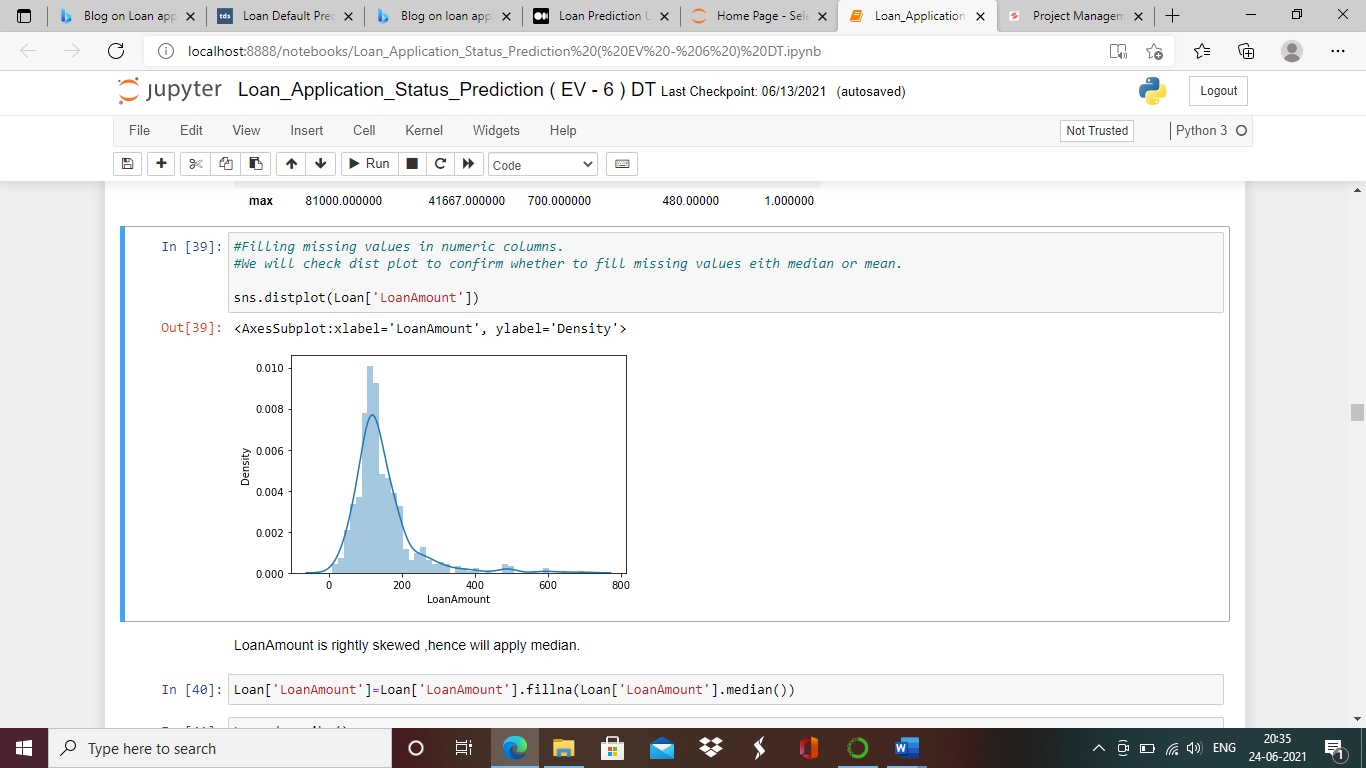
Above we can understand the count of unique variables in Gender,Married,Dependents,Education,Self Employed and Laon Amount Term columns.



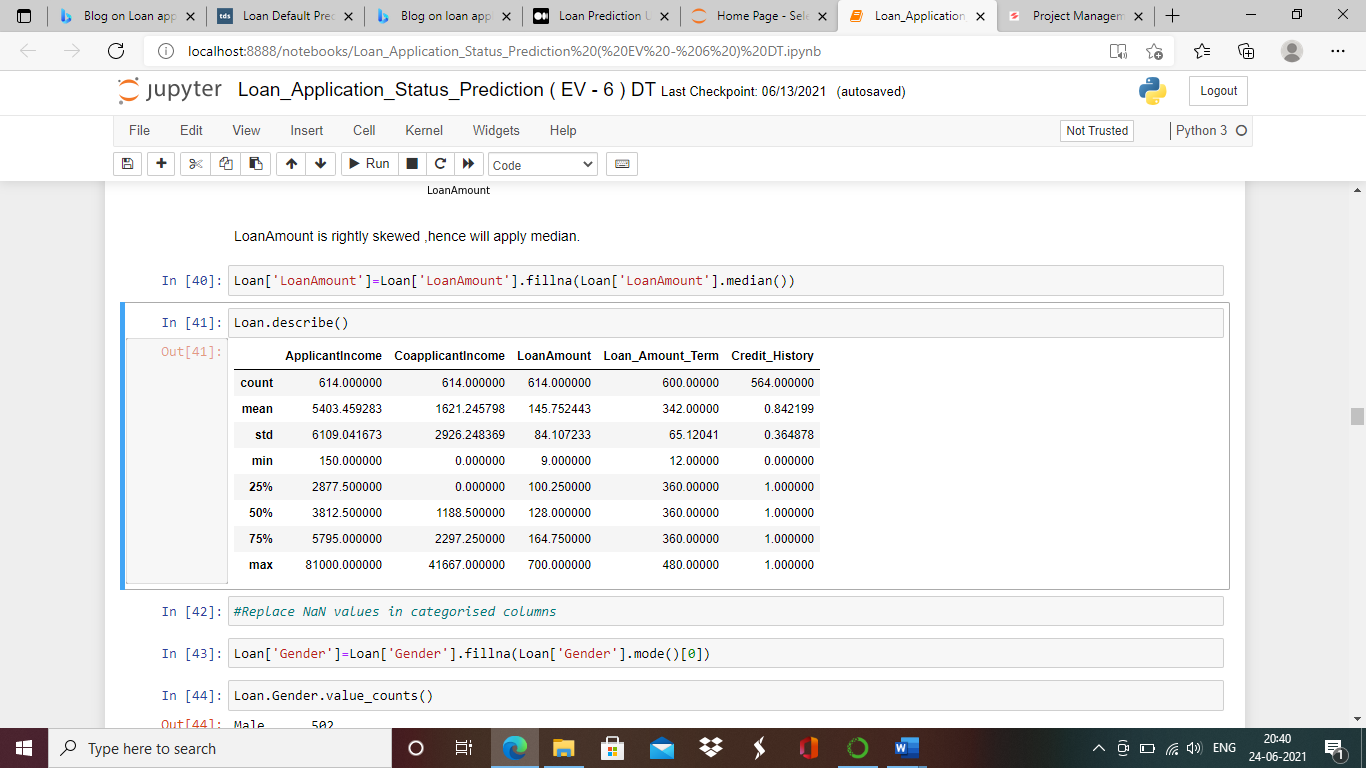
Loan status is our target/label column and can understand target column is imbalanced.We will correct the imbalancing later.

Treating Missing Values of Loan Amount:

We will check dist plot to confirm whether to fill missing values with median or mean value.

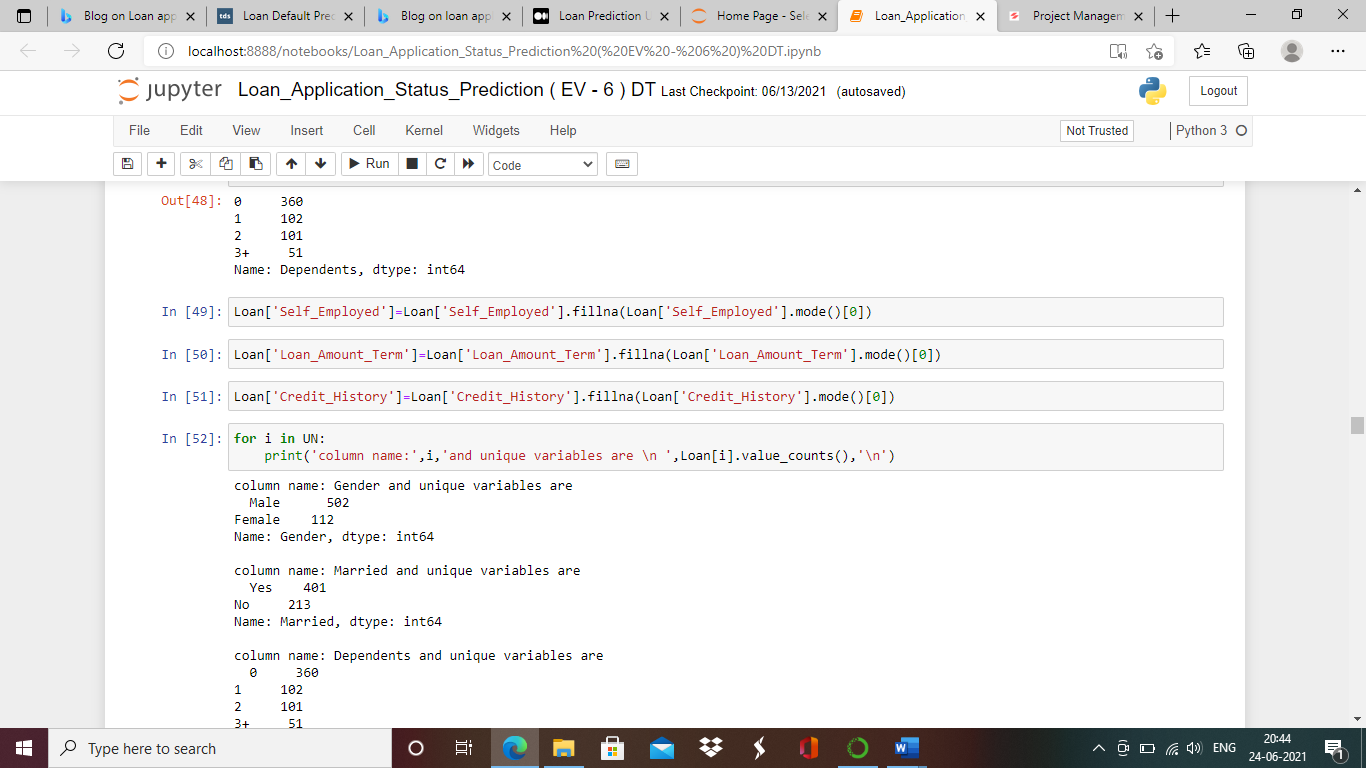


Loan Amount is rightly skewed ,hence will apply median.

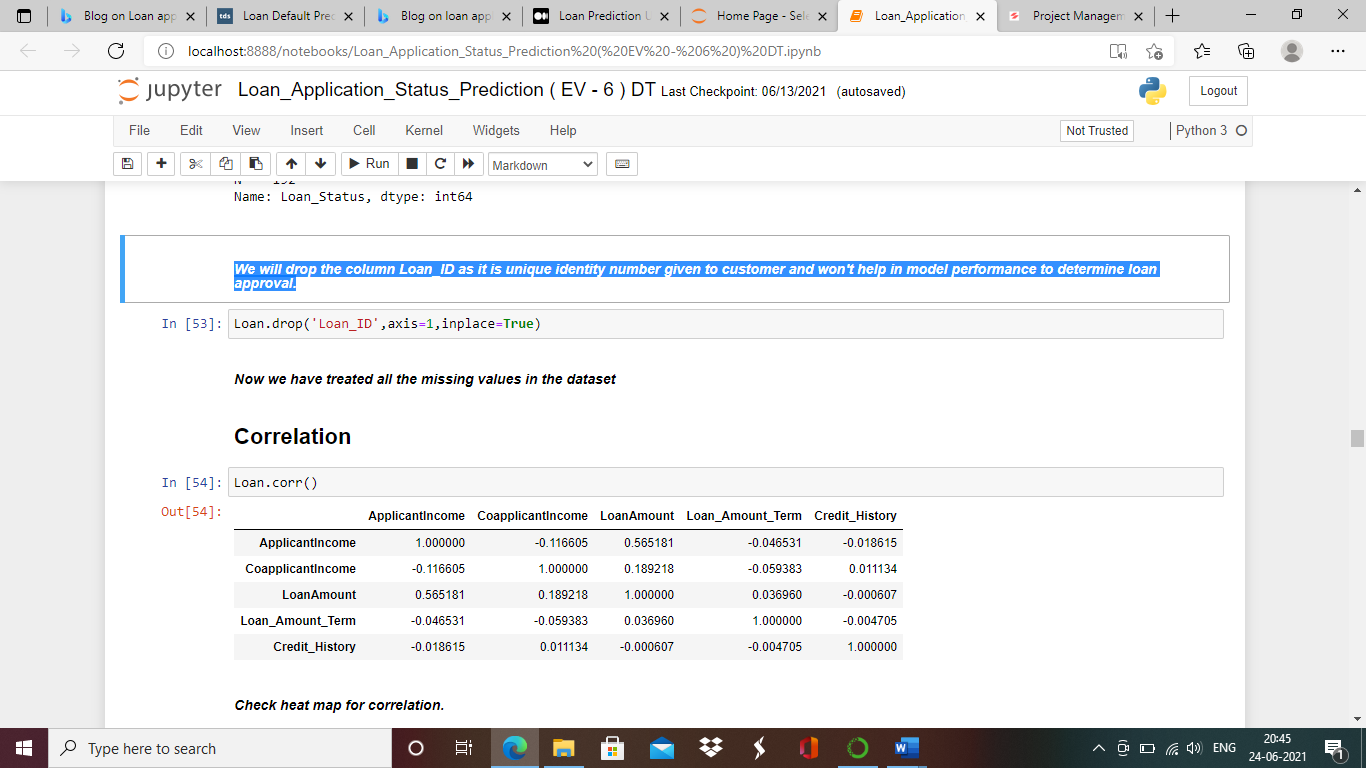


Treating Missing Values of Gender ,Married , Dependents, Self Employed, Loan Amount Term, Credit History:





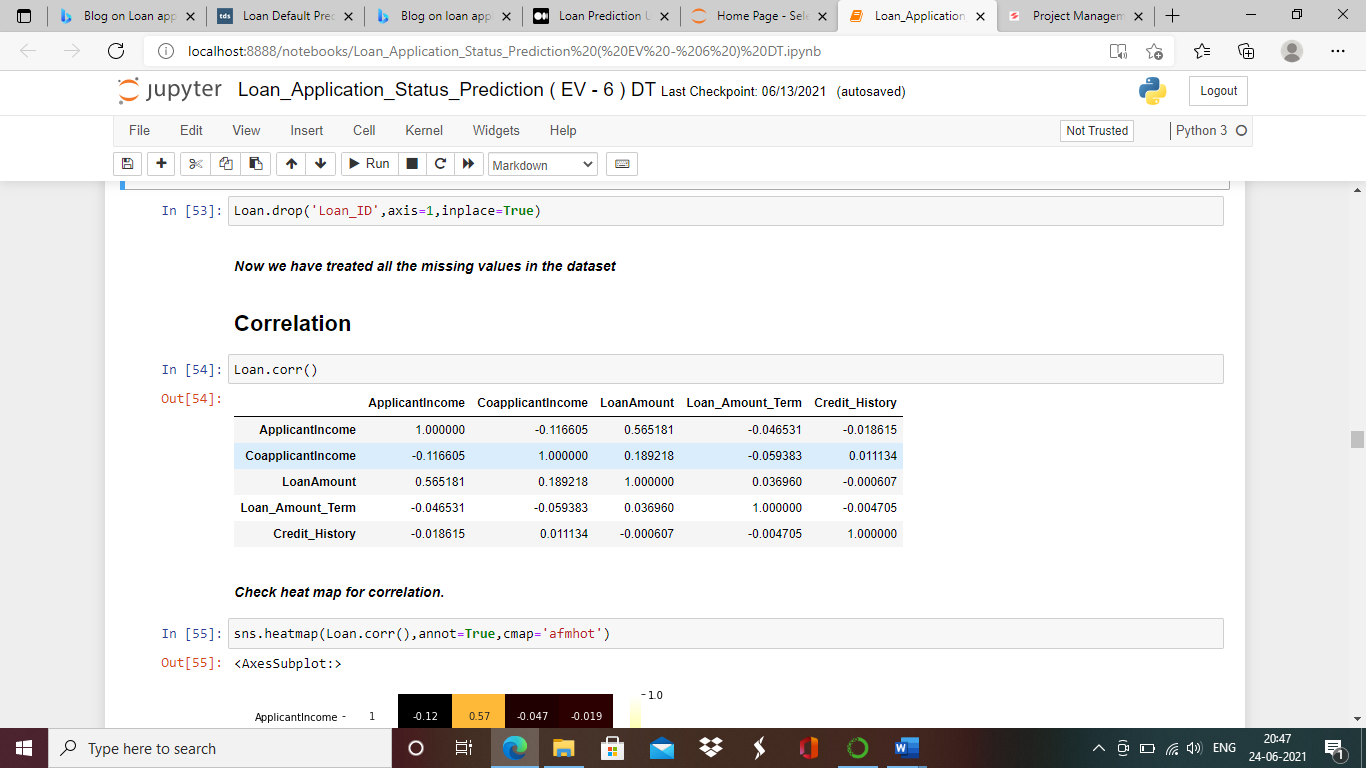
We will drop the column Loan\_ID as it is unique identity number given to customer and won't help in model performance to determine loan approval***.***



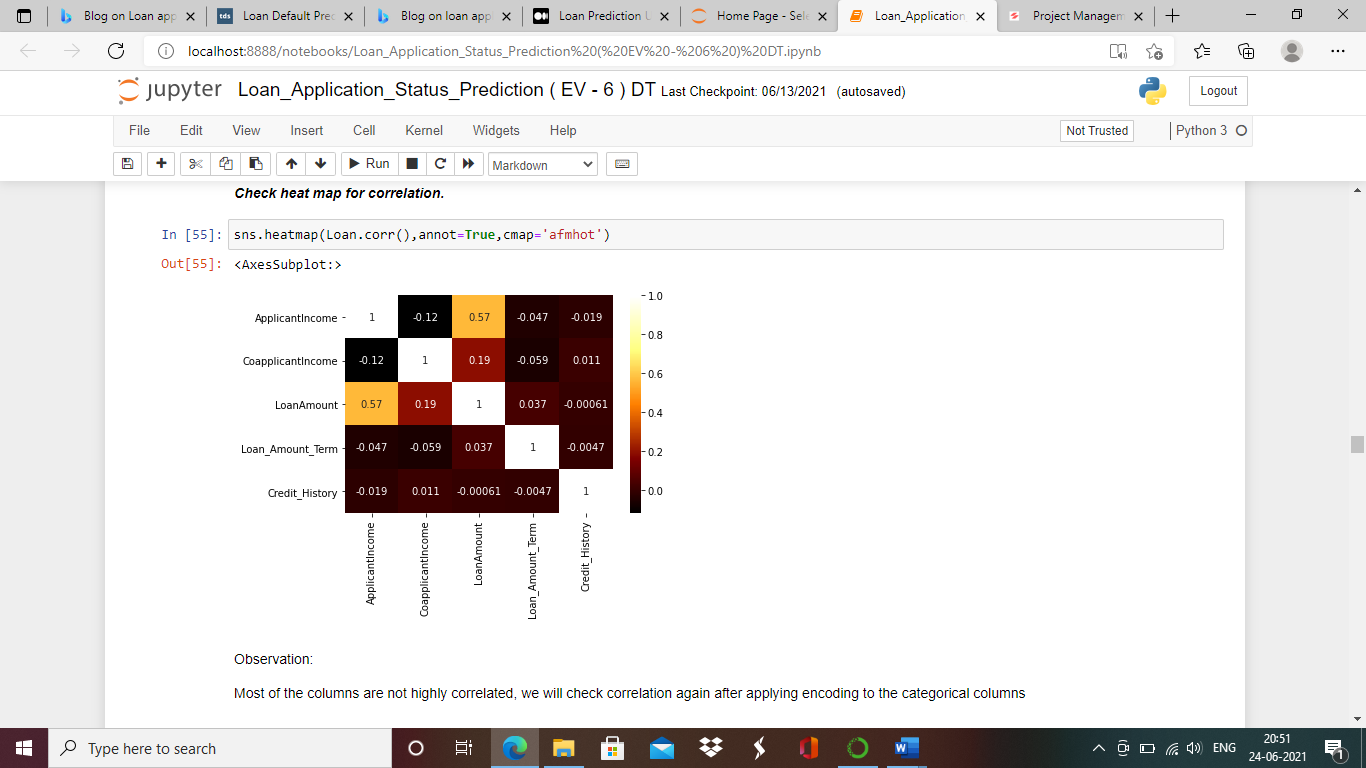
We have now treated all the missing values from loan dataset.

**We will check correlation of the columns**

Below is table justifiying the correlation of the columns.



For better understanding of the correlation between columns we will plot the correlation with help of heatmap.

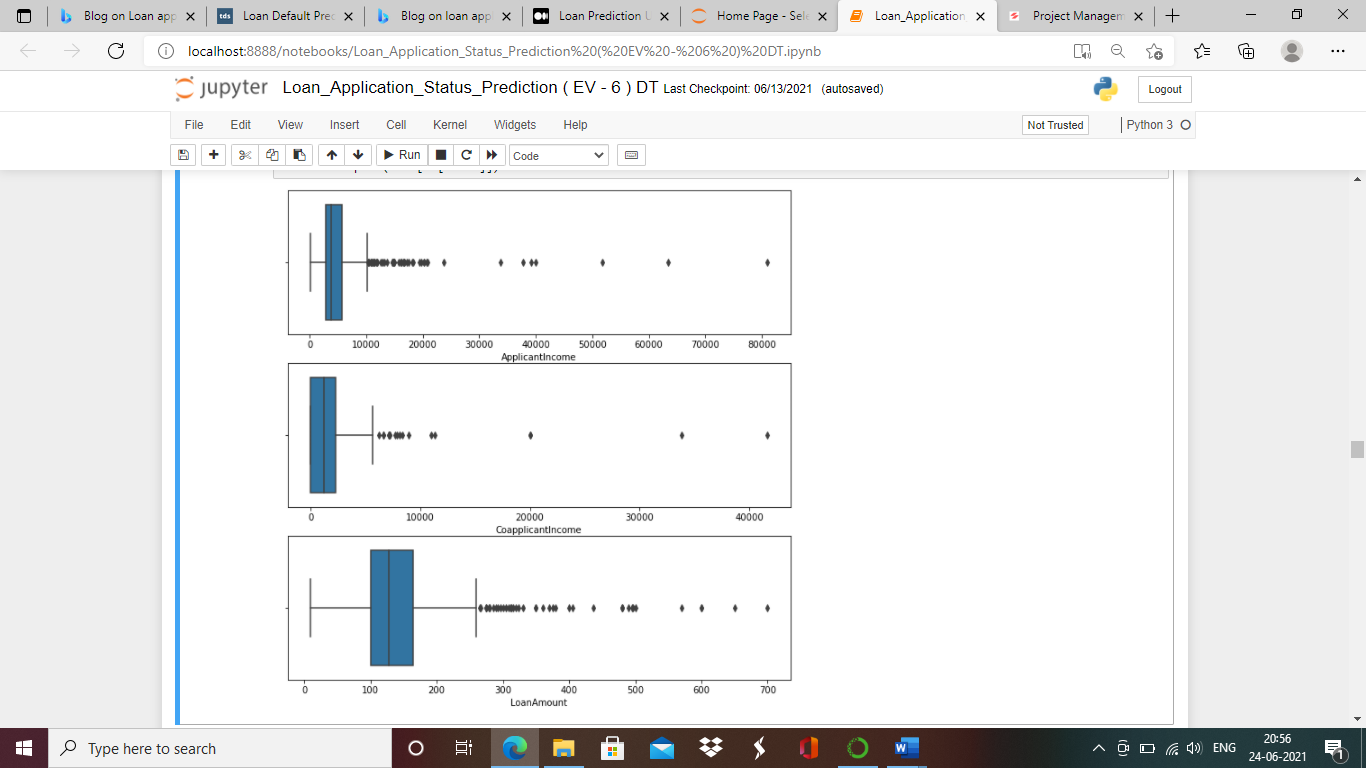


Observation:

Most of the columns are not highly correlated, we will check correlation again after applying encoding to the categorical columns

**Outliers**

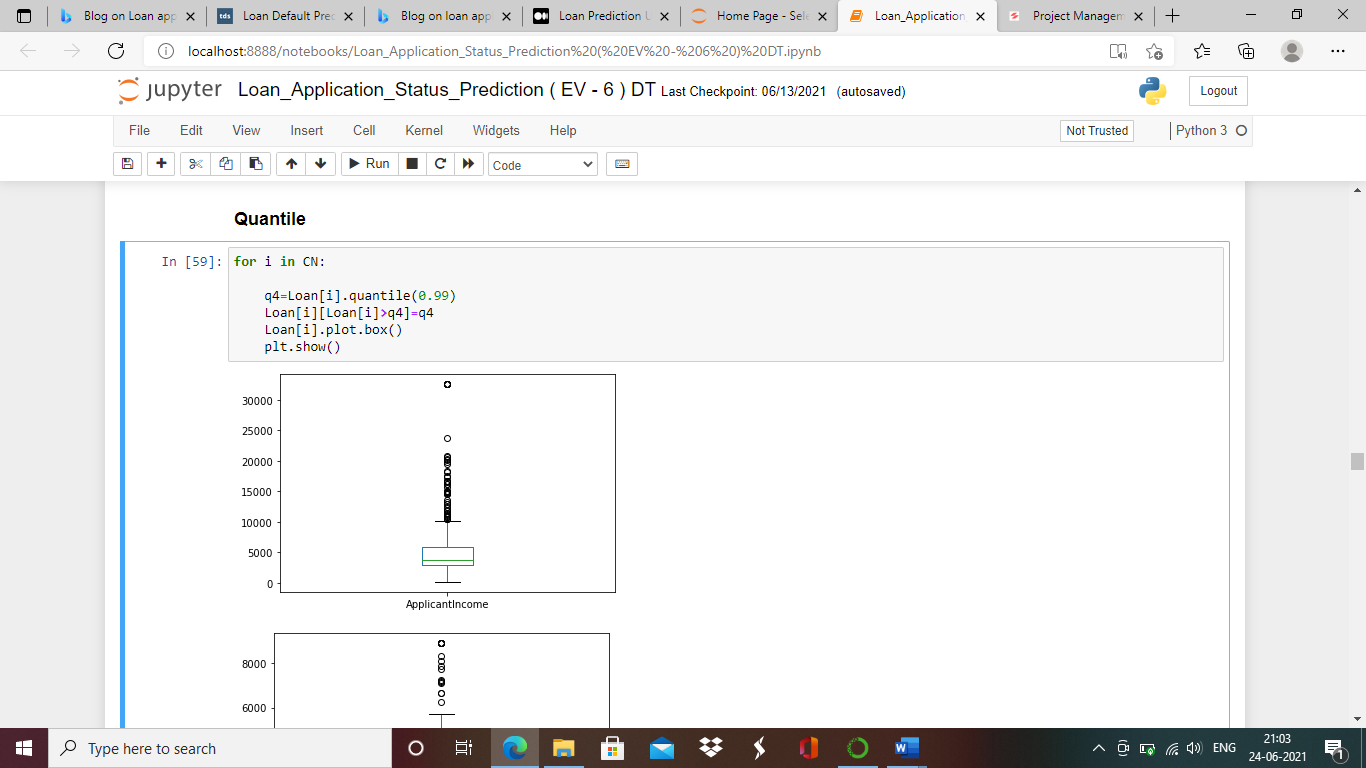
Let’s treat the outliers in loan dataset. Outliers are to be treated only from numeric and continuous type of columns .Applicant Income, Loan amount and Co applicant income are the numeric columns.

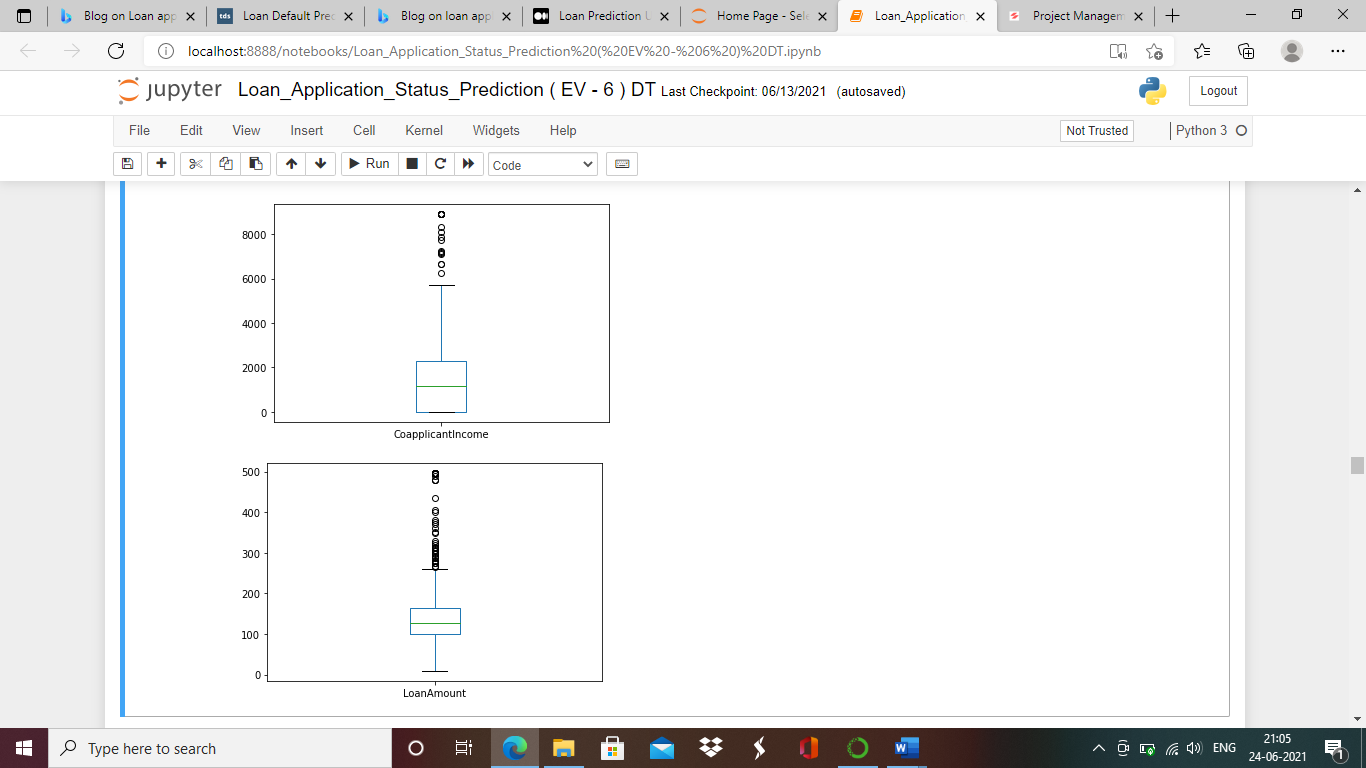


We can see huge amount of outliers are present in above columns, we will apply soft caping to treat them individually by quantile method.

### **Quantile**

Treating the q4 set of the columns with soft capping method.





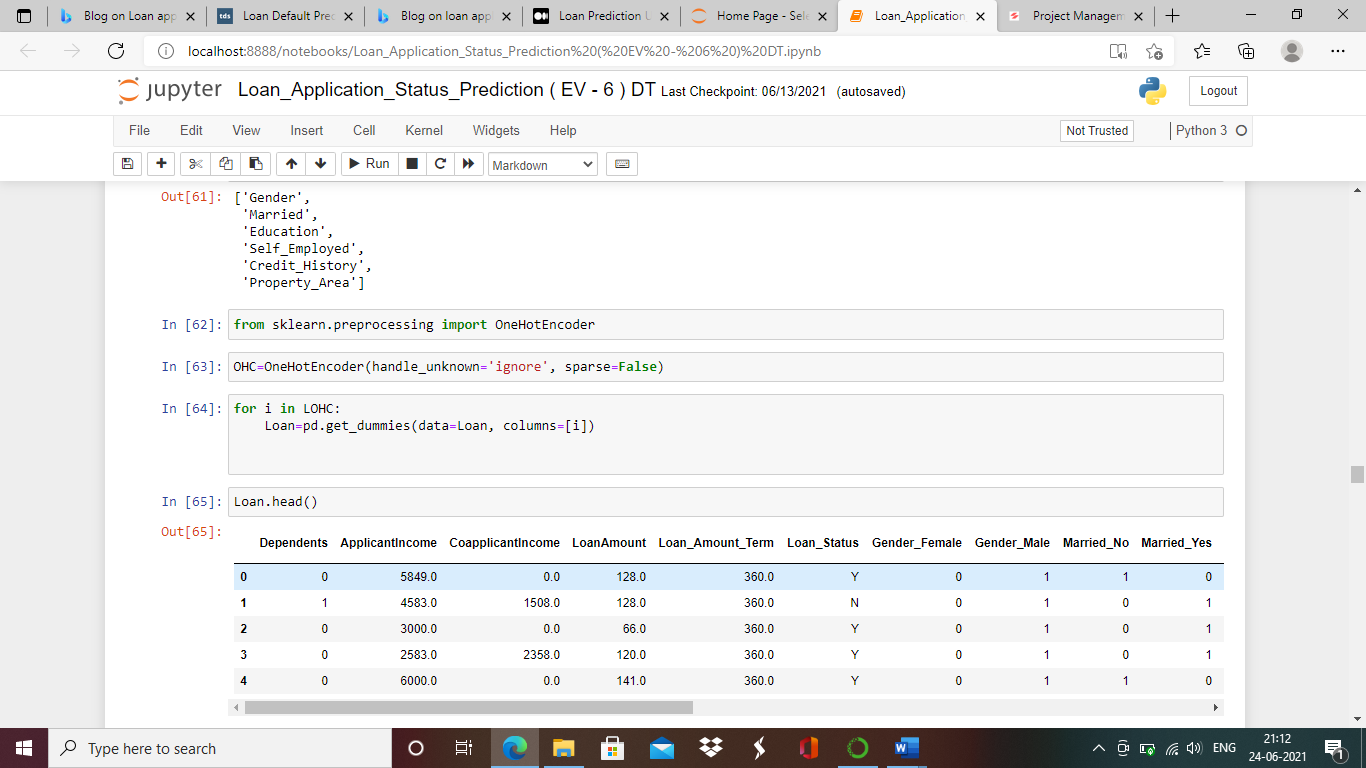
From above plots we can conclude most of the outliers are treated from required columns.

**Encoding**

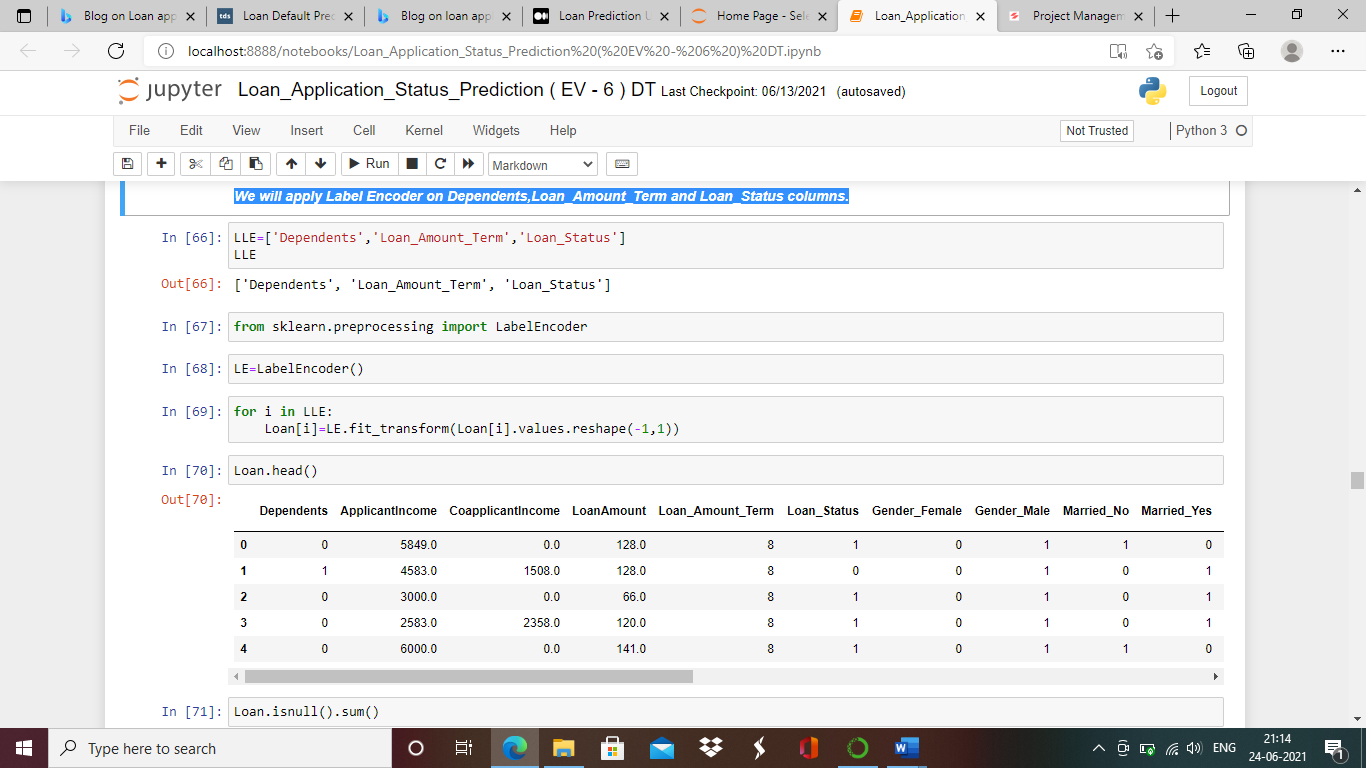
All object data type of columns to be converted from categorical variable to numerical numbers.

We will perform that with encoding method.

We will apply One hot encoder on Gender,Married,Education,Self\_Employed,Credit History,Property\_Area columns.



We will apply Label Encoder on Dependents,Loan\_Amount\_Term and Loan\_Status columns.

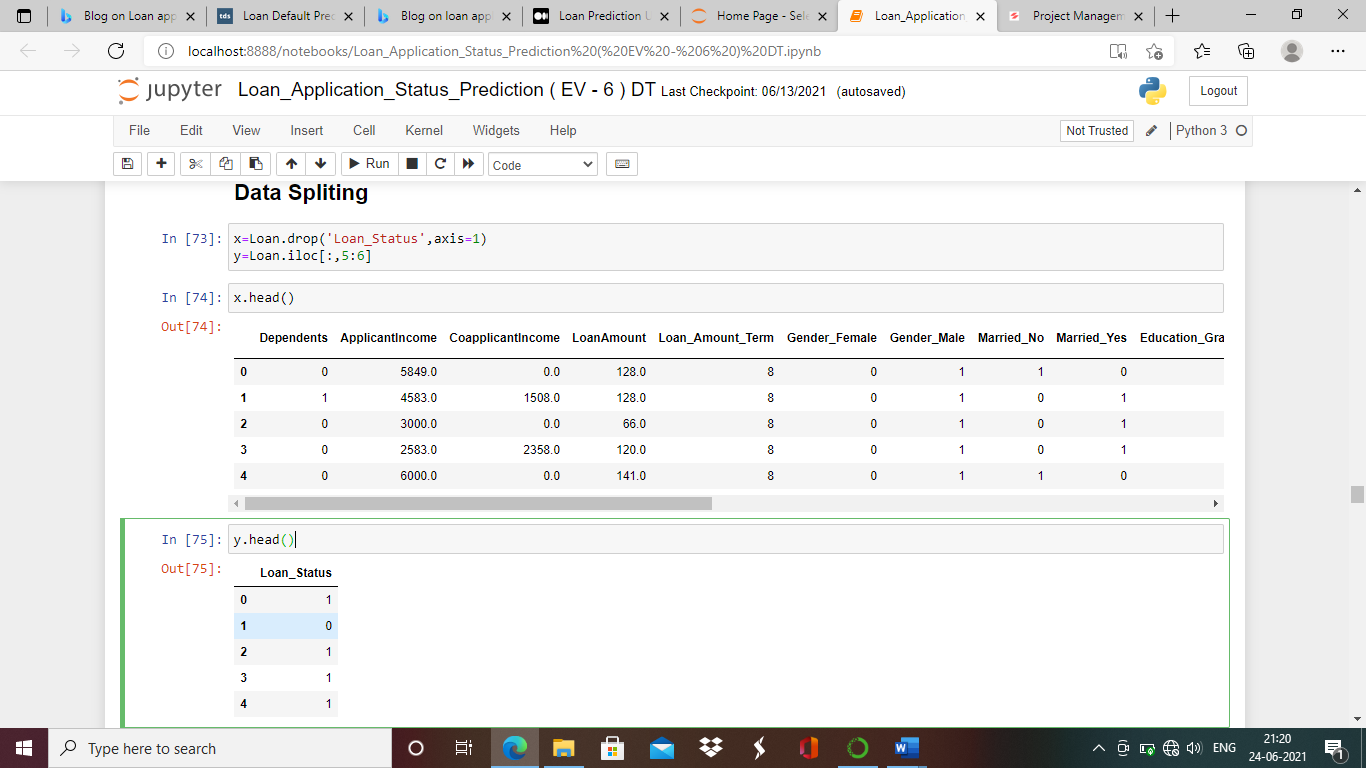


We have treated all the missing values form dataset,removed outliers and performed encoding an all categorized column.

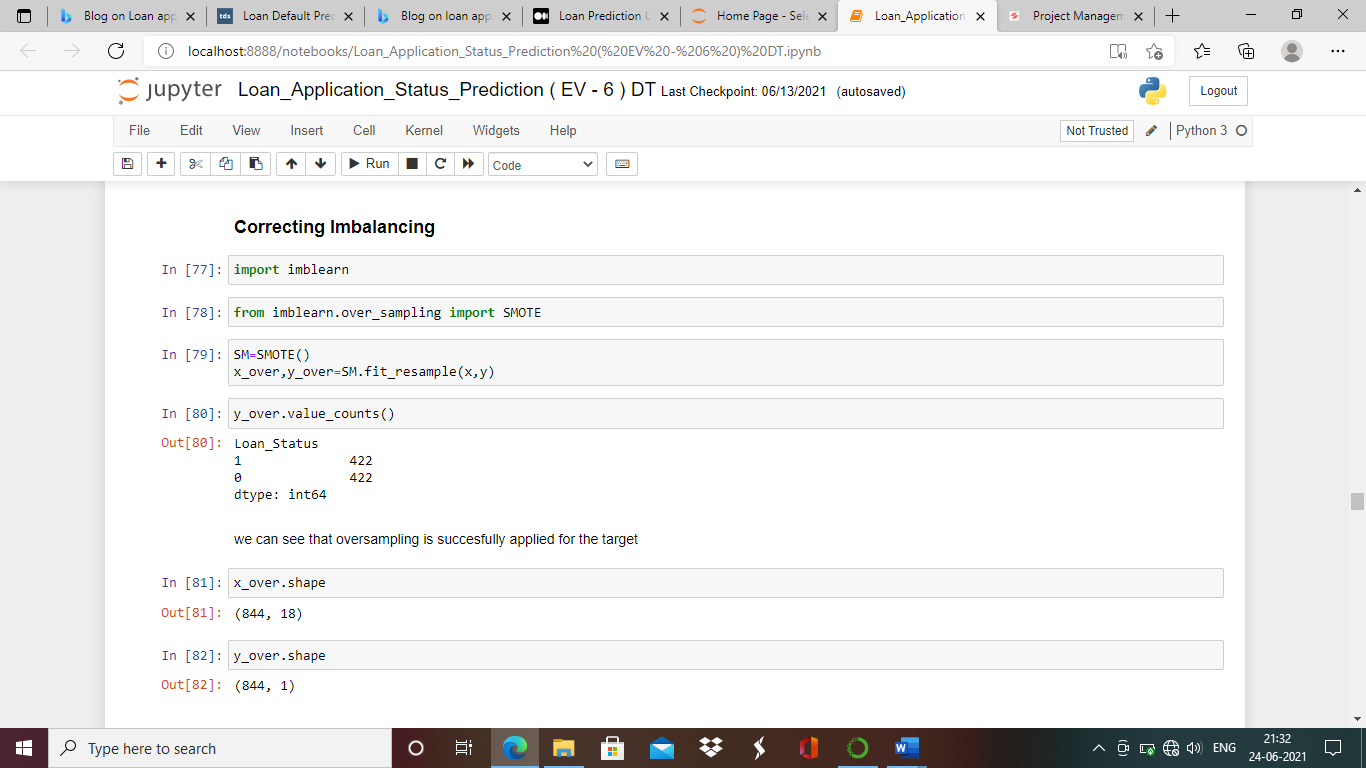
**Data Splitting :**

We will split the data in X and Y.

X set contains all the independent columns and Y set is dependent column.



We had seen earlier while treating misising values our Y set is imbalanced, and we need to correct this imbalancing for better performance of models.We can correct this imbalacing with oversampling method or under sampling .In this dataset we need to apply oversampling as Y set has 70% Non defaulter and 30 % defaulters, also rows in this dataset are very limited and under sampling will cause to lossing of the data .Which inturn can lead to impacting the model performance.



After appying over sampling our X and Y set size rows have increased from 614 to 844.

**Skewness**

Skewness measures the deviation of a random variable’s given distribution from the normal distribution, which is symmetrical on both sides.Models don’t run good when data is skewed and the reason behind this is that the tapering ends or the tail region of the skewed data distributions are basically the outliers in the data and it is known to us that outliers can severely damage the performance of a statistical model the best example of this being regression-models which show very bad results when trained over skewed data.Hence, before running X set the we should check the skewness of each numerical column and not categorized column.

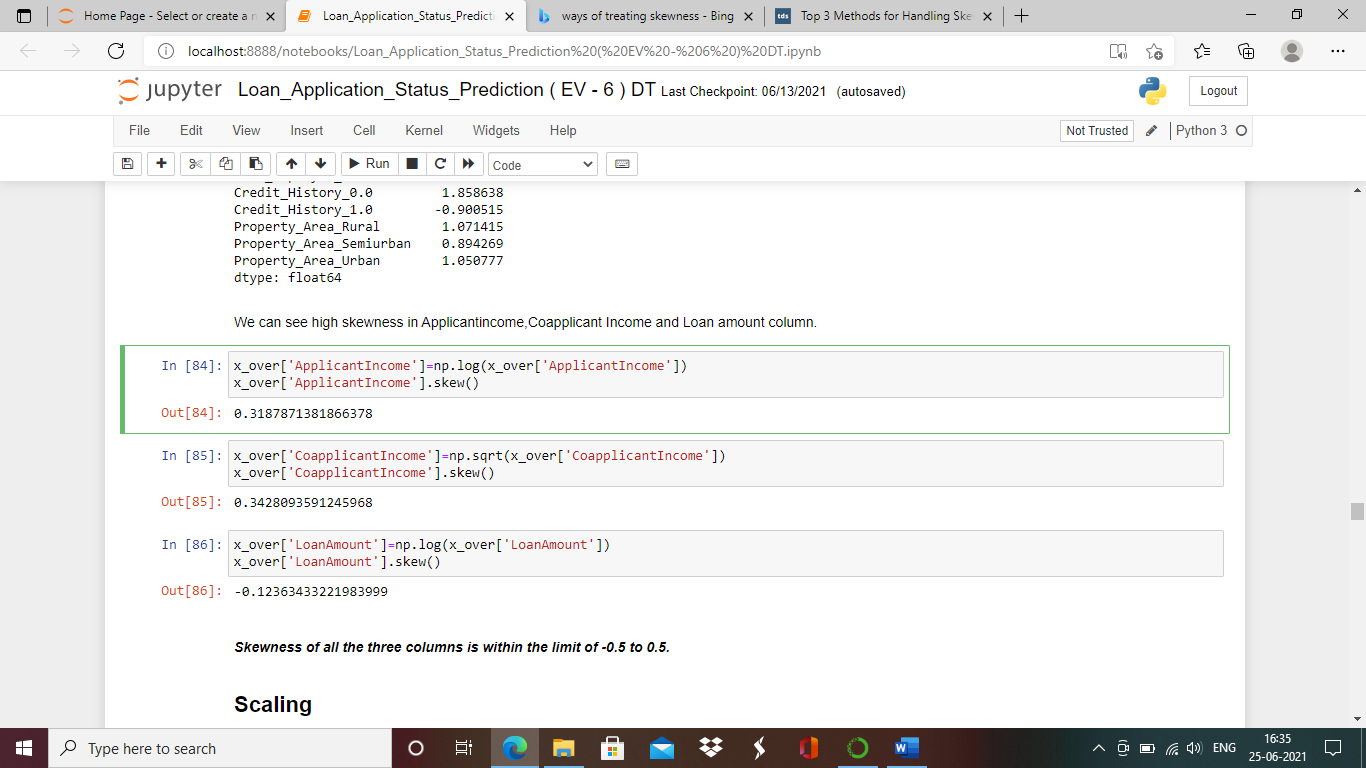
Skew value od column should aleways be between -0.5 to 0.5.

We will check the skewness of each column now.



Skewness is more than the limit in Applicant income, Co-applicant Income and Loan amount column.

To treat the skewness of column there are various methods like log transform, square root transform and box-cox transform. First, we will try correcting the skewness with log transform method for all three columns and if we still observe the skewness than we can use square root transform and at last box cox method.



Using log tranform we have corrected skewness of Applicant Income and Loan Amount.

Using square root transform we have corrected skewness of Co applicant Income.

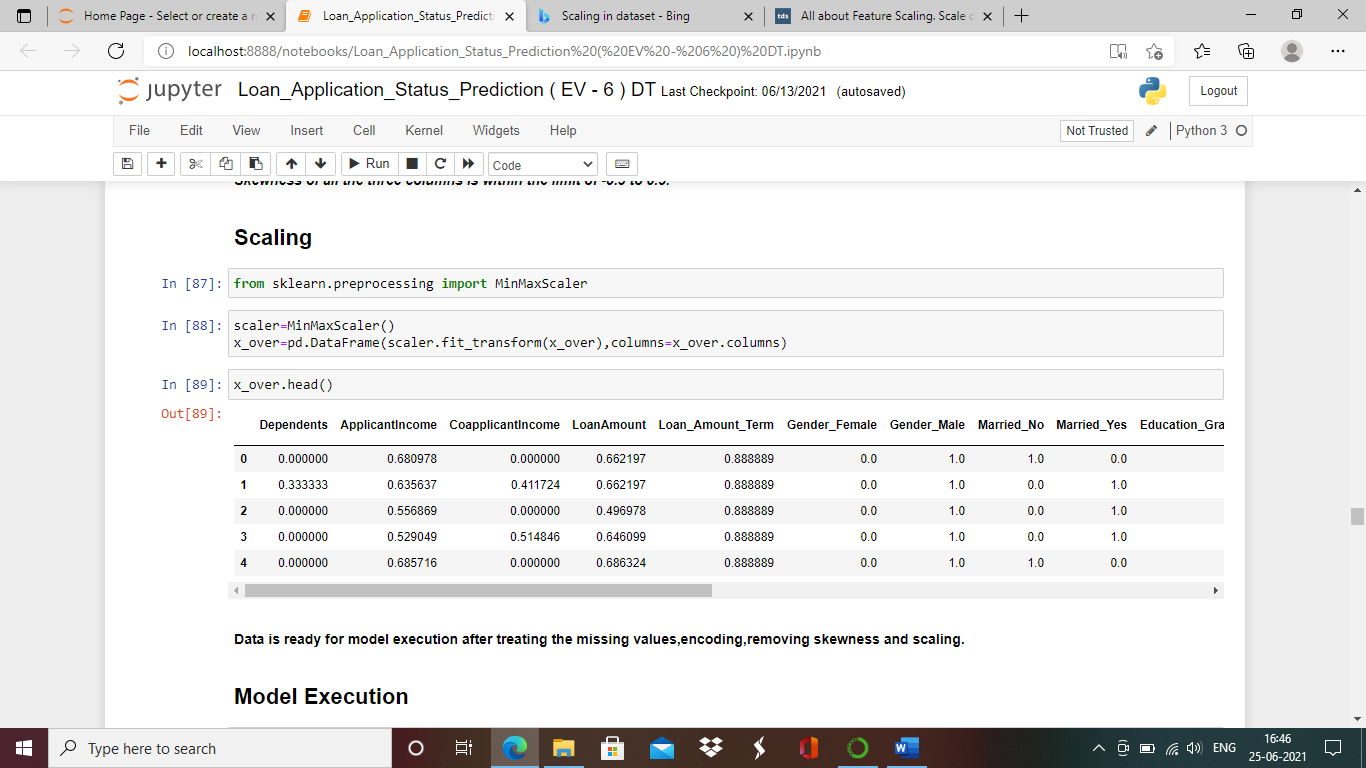
Now skewness of all the coulmns is with in the limt.

**Scaling**

Scaling is the final treatment we need to do before executing the data in model for prediciotn.

Scaling is performed when there is vast difference in range of data within column.There are various methods we can perform feature scaling. (i.e Minmax scaler ,standard scaler,Robust scaler etc.

We are using Min max scaler for data set as standard deviation is small.



Data is ready for model execution after treating the missing values, encoding, removing skewness and scaling.

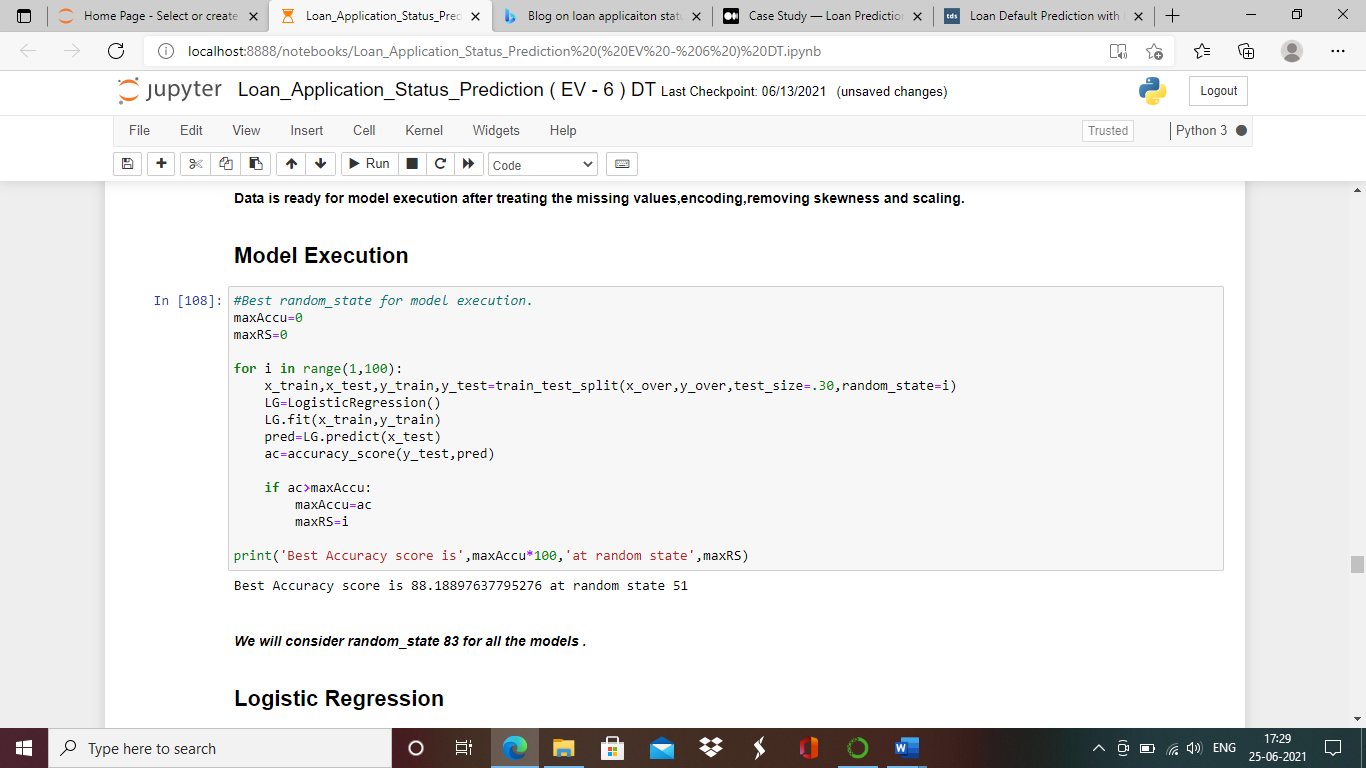
**Building Machine Learning Models**

For model execution we have to split the data in train and test data which plays a very critical role in model performance, as imbalanced fitting of test and train data can impact model performance. We should always try keeping training data close to 70% for better performance.

This being a classification problem we can use various models like logistic regression, Decision Tree Classifier, Random Forest classifier, Ada boost classifier, XgBoost classifier etc.

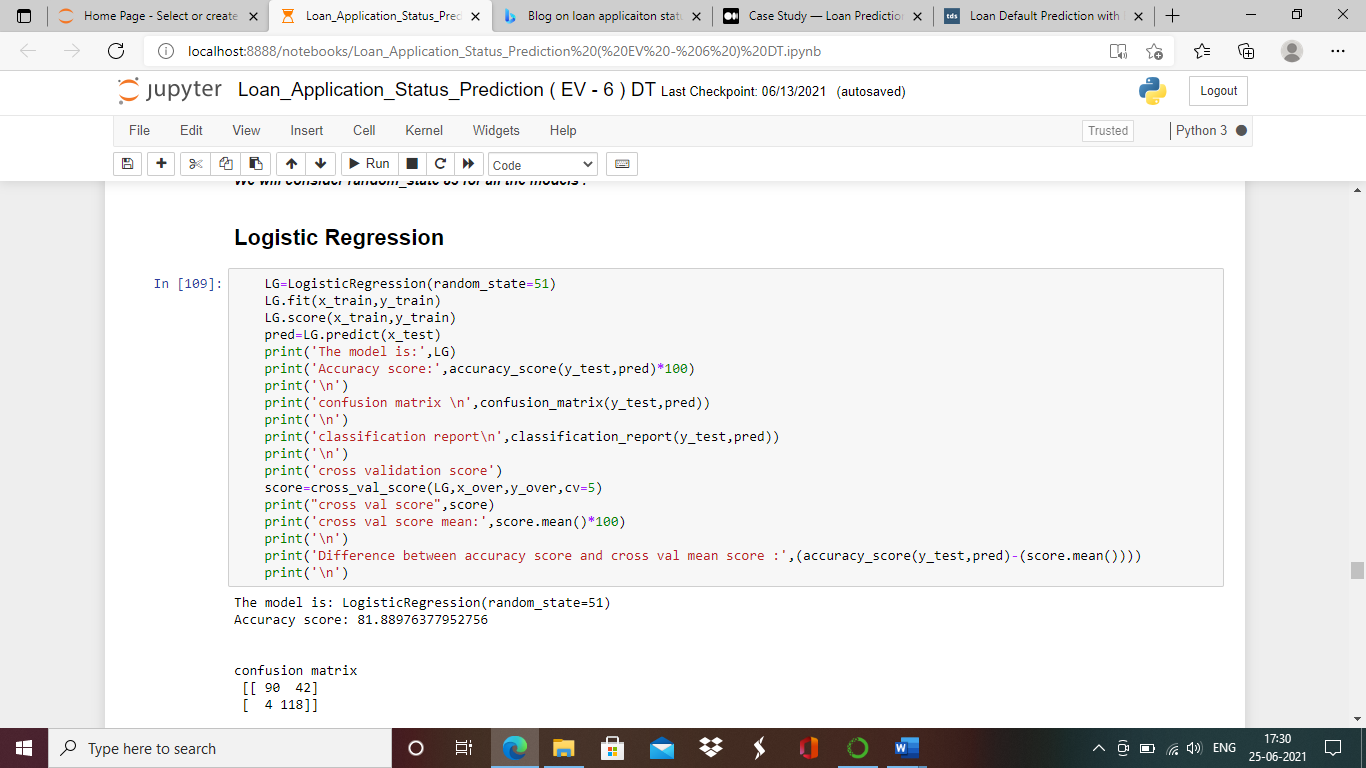
I have tried running all the above mentioned model and also have also applied hyper parameter tunning to improve the performance of each model. After running al modle I have to come to conclusion at logistic regression model was best performing model ,we will check the performance and details.

Initially we will check the best random state for model performance using logistic regression.



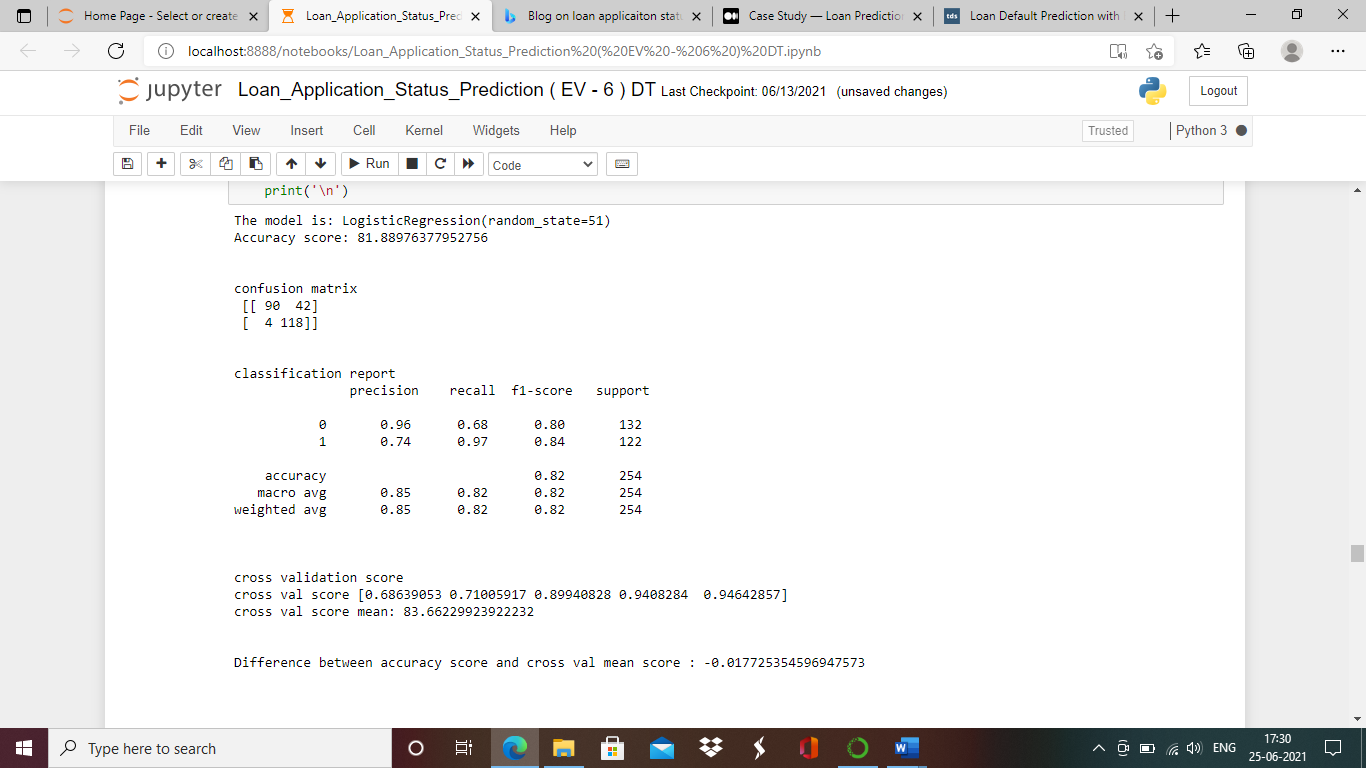
Best accuracy score we are getting is 88.58% at random state=45.

Logistic Regression.

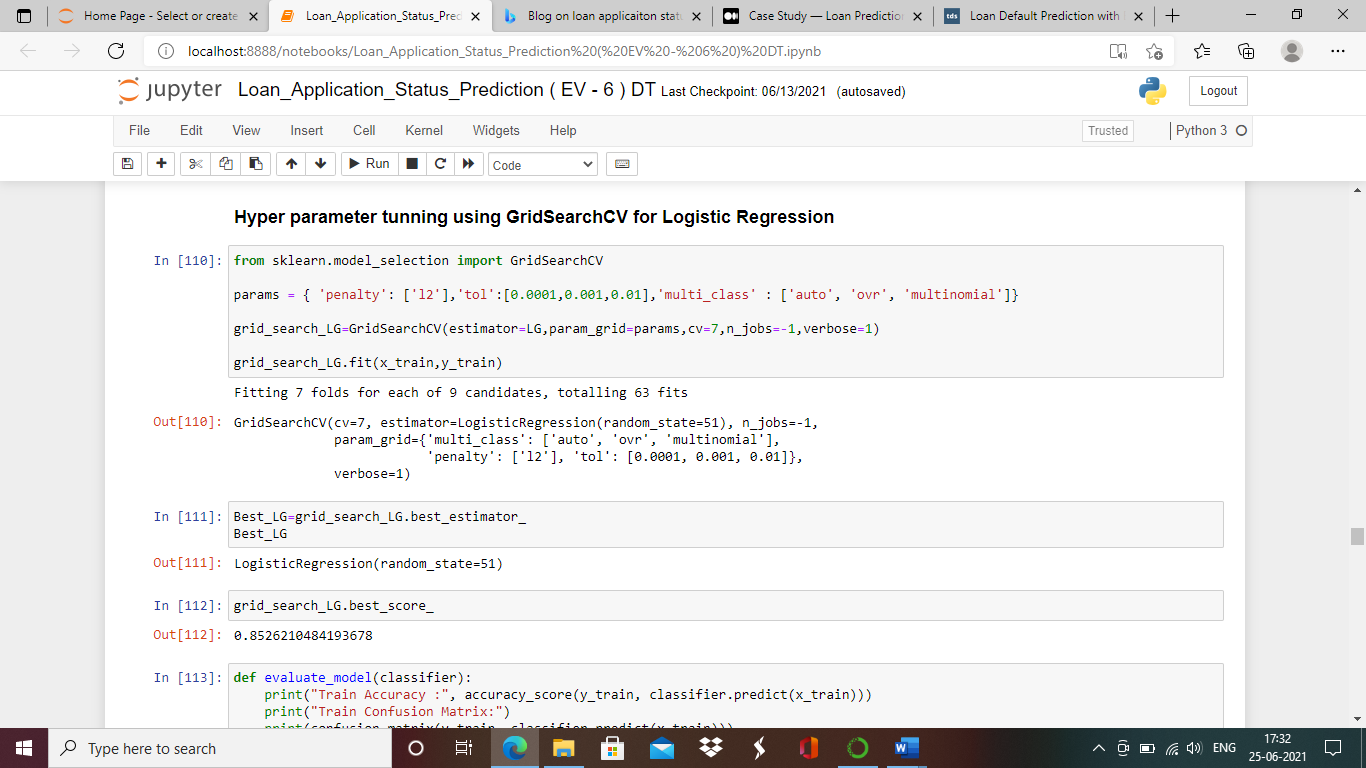


Above is the code for logistic regression model to check accuracy score of test model, confusion matrix and classification report .We will also check cross validation score of this model at different cross folds.

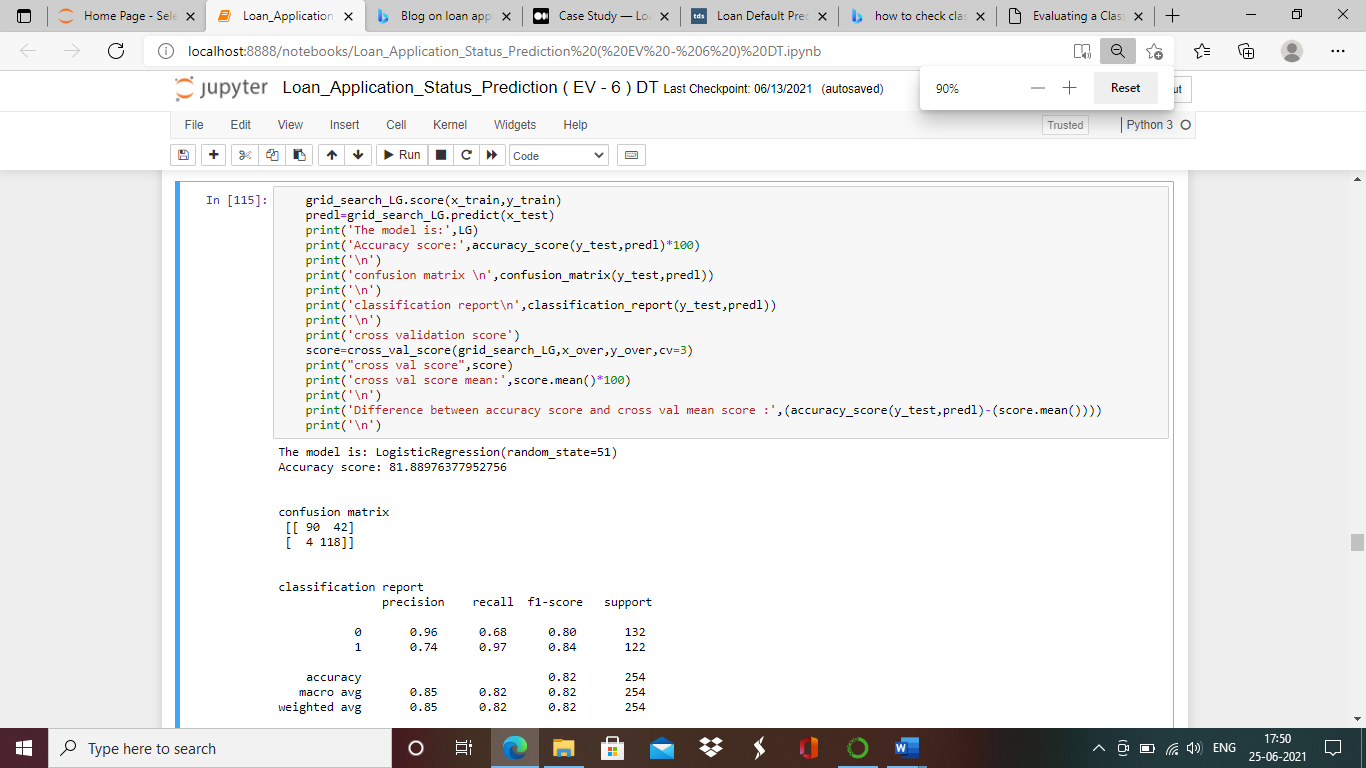
Finally, we check the difference between accuracy score and mean cross validation score, least the difference between them better is the model performance.



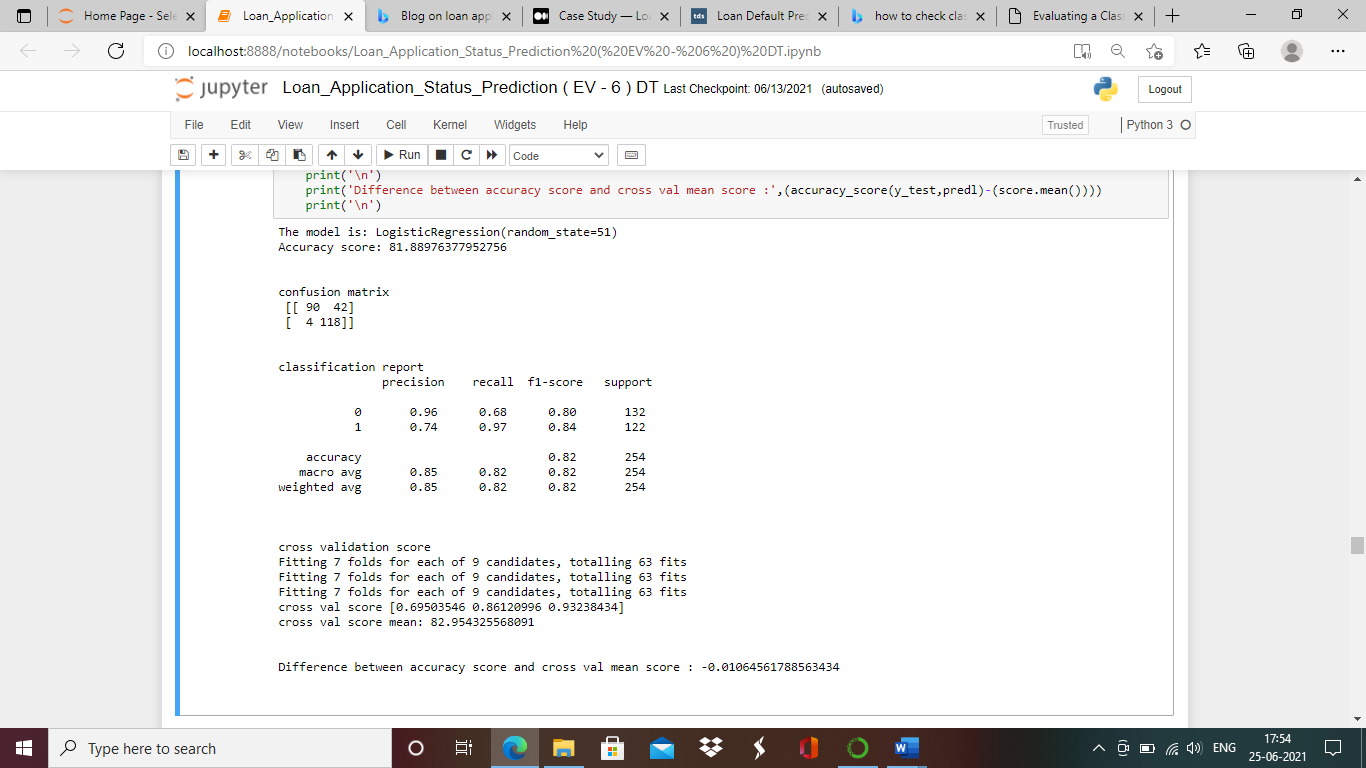
Let’s apply hyper parameter tunning using grid search cv for logistic regression



Code for logistic regression after applying hyper parameter tunning

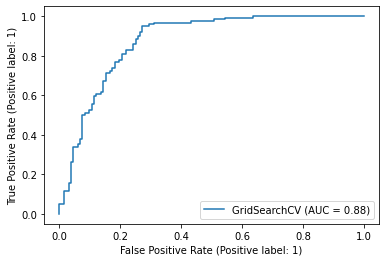


After running the above code we can conclude that



Logistic regression is giving us really good accuracy for test data of 81.88% at 3 cross folds also F1 score is good of .80.

AUC ROC CURVE (AREA UNDER CURVE)



AUC score is also good, concluding better prediction of values.

**Concluding Remarks**

In Loan application status prediction dataset, we have introduced the whole pipeline of an end-to-end machine learning. We have described the dataset, and explored all the variables in the dataset. Checked if there are any missing values. Exploratory data analysis is performed using univariate, bivariate and multivariate columns for better understanding of data.

Data featuring is performed by treating the missing the values, checking correlation between columns, encoding all the alphabetical columns to numerical. Outliers are cleared in dataset with quantile method and made ready to run the model. We have also found the label column was imbalanced which is corrected with oversampling method. We have run various models to find the best model and found loan application status prediction has used logistic regression model as best model to predict the defaulter. This model has achieved the accuracy score of 81.88% for test dataset after applying hyper parameter tunning. F1 score is also observed good of 0.80.

**Thank you!!!!**