**Loan Application Status Prediction**

**Problem Statement:**

There is a company named Dream Housing Finance that deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. However doing this manually takes a lot of time. Hence it wants to automate the loan eligibility process (real time) based on customer information

So the final thing is to identify the factors/ customer segments that are eligible for taking loan. How will the company benefit if we give the customer segments is the immediate question that arises. The solution is ….Banks would give loans to only those customers that are eligible so that they can be assured of getting the money back. Hence the more accurate we are in predicting the eligible customers the more beneficial it would be for the Dream Housing Finance Company.

**Type of Problem:**

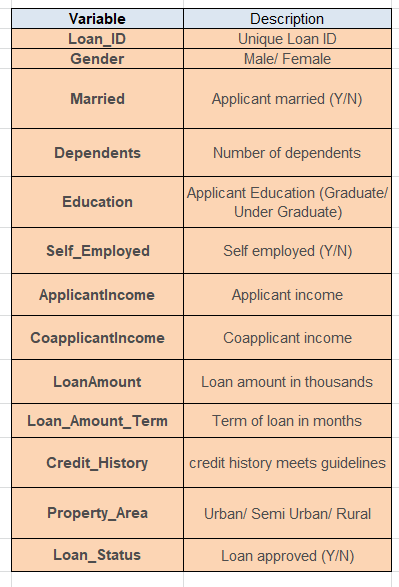
The above problem is a clear classification problem as we need to classify whether the Loan Status is yes or no. So this can be solved by any of the classification techniques like.

1. Logistic Regression .
2. Decision Tree Algorithm.
3. Random Forest Technique.

I have mentioned only few. We will be dealing with each of techniques later in this blog.

**Description about the Data Columns:**

It’s very useful to know about the data columns before getting in to the actual problem for avoiding confusion at a later state. Now let us understand the data columns (that has been already given by the company itself) first so that we will get a glance.



There are altogether 13 columns in our data set. Of them Loan Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not.

Now let us look in to the each variable and can make some assumptions. (It’s just assumptions right, there is no harm in just assuming few statements)

Loan ID -> As the name suggests each person should have a unique loan ID.

Gender -> In general it is male or female. No offence for not including the third gender.

Married -> Applicant who is married is represented by Y and not married is represented as N. The information regarding whether the applicant who is married is divorced or not has not been provided. So we don’t need to worry regarding all these.

Dependents -> the number of people dependent on the applicant who has taken loan has been provided.

Education -> It is either non -graduate or graduate. The assumption I can make is “The probability of clearing the loan amount would be higher if the applicant is a graduate”.

Self-employed -> As the name suggests Self Employed means, he/she is employed for himself/herself only. So freelancer or having a own business might come in this category. An applicant who is self-employed is represented by Y and the one who is not is represented by N.

Applicant Income -> Applicant Income suggests the income by Applicant. So the general assumption that i can make would be “The one who earns more have a high probability of clearing loan amount and would be highly eligible for loan ”

Co Applicant income -> this represents the income of co-applicant. I can also assume that “If co applicant income is higher, the probability of being eligible would be higher “

Loan Amount -> This amount represents the loan amount in thousands. One assumption I can make is that “If Loan amount is higher, the probability of repaying would be lesser and vice versa”

Loan\_Amount\_Term -> This represents the number of months required to repay the loan.

Credit\_History -> When I googled it, I got this information. A credit history is a record of a borrower’s responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.

Property\_Area -> The area where they belong to is my general assumption as nothing more is told. Here it can be three types. Urban or Semi Urban or Rural

Loan\_Status -> If the applicant is eligible for loan it’s yes represented by Y else it’s no represented by N.

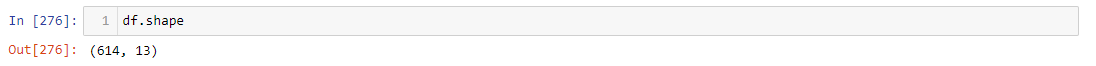
**Data Analysis:**

Now let me walk through the code. Firstly I just imported the necessary packages like pandas, numpy, seaborn etc. so that i can carry the necessary operations further.

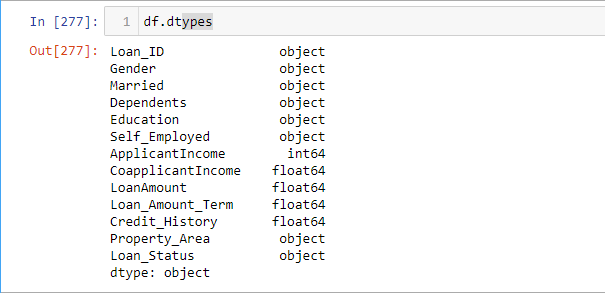


Now I am going to upload or read the files/data-sets using pandas. For this we used read\_csv.



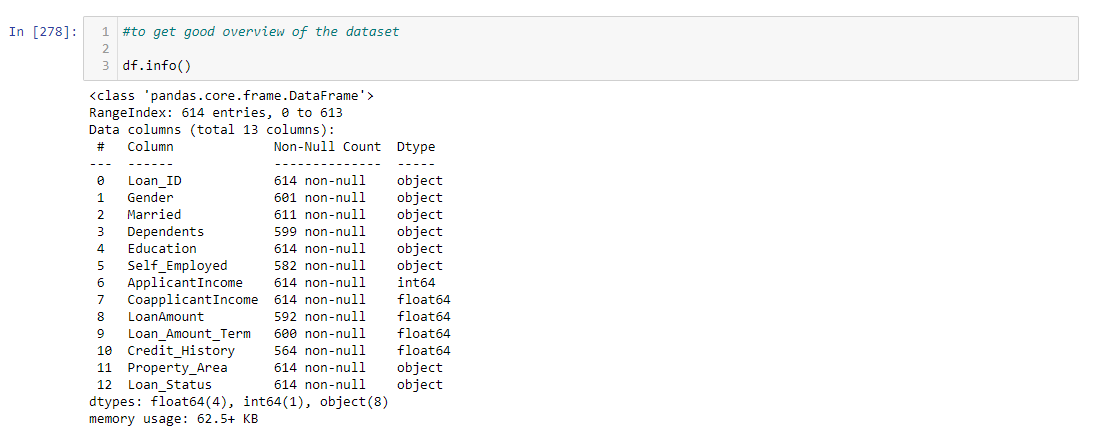
This Dataset has 614 rows and 13 columns.

**Now checking the Datatypes:**



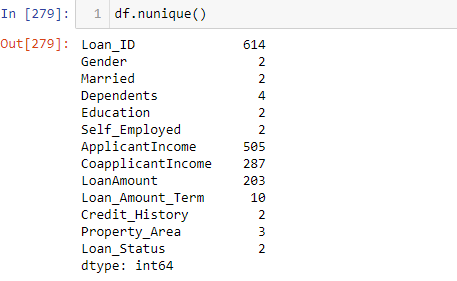
This Dataset contain Integer and object Datatype.

**Checking all Variables Datatype using df.info()**



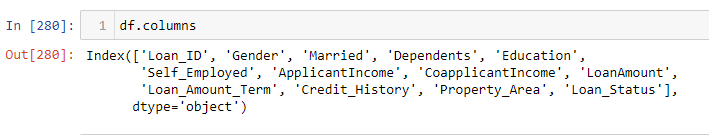
This gives the information about the dataset which includes indexing type, non-null values and memory usage. There are 8 object and 5 numerical columns in Data.

**Now let us check the unique values present in the Dataset:**



This shows that there are many null values present in the Dataset.

**Now checking the number of Columns present in the Dataset**:



These are the columns present in the Dataset.

**Now let us analyse the data using single variable:**

*#checking the value counts of each columns*

2

**for** i **in** df.columns:

3

print(df[i].value\_counts())

4

print('\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*')

LP001002 1

LP002328 1

LP002305 1

LP002308 1

LP002314 1

..

LP001692 1

LP001693 1

LP001698 1

LP001699 1

LP002990 1

Name: Loan\_ID, Length: 614, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Male 489

Female 112

Name: Gender, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Yes 398

No 213

Name: Married, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

0 345

1 102

2 101

3+ 51

Name: Dependents, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Graduate 480

Not Graduate 134

Name: Education, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

No 500

Yes 82

Name: Self\_Employed, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

2500 9

4583 6

6000 6

2600 6

3333 5

..

3244 1

4408 1

3917 1

3992 1

7583 1

Name: ApplicantIncome, Length: 505, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

0.0 273

2500.0 5

2083.0 5

1666.0 5

2250.0 3

...

2791.0 1

1010.0 1

1695.0 1

2598.0 1

240.0 1

Name: CoapplicantIncome, Length: 287, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

120.0 20

110.0 17

100.0 15

160.0 12

187.0 12

..

240.0 1

214.0 1

59.0 1

166.0 1

253.0 1

Name: LoanAmount, Length: 203, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

360.0 512

180.0 44

480.0 15

300.0 13

240.0 4

84.0 4

120.0 3

60.0 2

36.0 2

12.0 1

Name: Loan\_Amount\_Term, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1.0 475

0.0 89

Name: Credit\_History, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Semiurban 233

Urban 202

Rural 179

Name: Property\_Area, dtype: int64

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Y 422

N 192

Name: Loan\_Status, dtype: int64

**Conclusions**: (Through Single Variable Analysis)

1. We can see that approximately 81% are Male and 19% are female.
2. Percentage of applicants with no dependents is higher.
3. There are more number of graduates than non-graduates.
4. Semi Urban people is slightly higher than urban people among the applicants.
5. Larger Percentage of people have a good credit history.
6. The percentage of people that the loan has been approved has been higher rather than the percentage of applicant for which the loan has been declined.
7. **Exploratory Data Analysis:**

Well don’t get to worry about the fancy names like exploratory data analysis and all. By looking at the columns description in the above paragraph, we can make many assumptions like r loan approved and less data in declines loan, will have to deal with problem in data imbalance.

**Exploratory Data Analysis:**

Well don’t get to worry about the fancy names like exploratory data analysis and all. By looking at the columns description in the above paragraph, we can make many assumptions like:

1. The one whose salary is more can have a greater chance of loan approval.
2. The one who is graduate has a better chance of loan approval.
3. Married people would have a upper hand than unmarried people for loan approval.
4. The applicant who has less number of dependents have a high probability for loan approval.
5. The lesser the loan amount the higher the chance for getting loan.

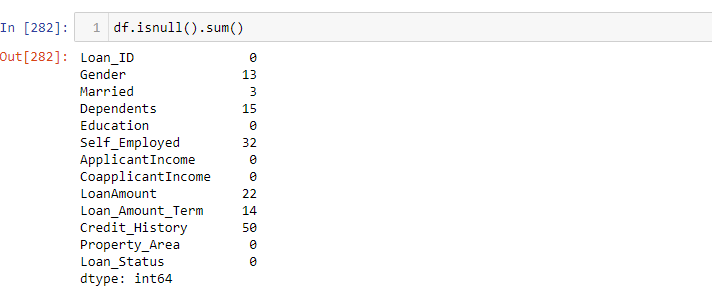
**Why are we doing EDA?**

We just to do EDA. Then there is no necessary for going through Like these there are many more we can assume. But one basic question you may get it …”Why are we doing all these? Why can’t we do directly modelling the data instead of knowing all these…..” Well in some cases we can easily come to conclusion if next models.

**DATA CLEANING AND STRUCTURING:**

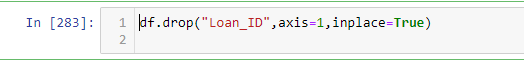
Before we go for modelling the data, we have to check whether the data is cleaned or not. And after cleaning part, we have to structure the Data. For cleaning part, first I have to check whether there exists any missing values. For that I am using the code snippet isnull().sum()

**Checking the Null values present in the Dataset:**



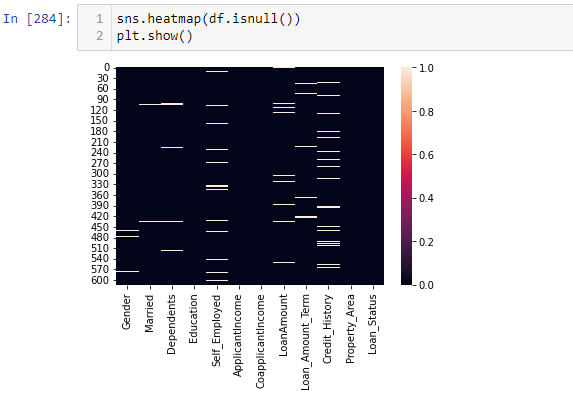
The above code suggests that there are 13 missing values in Gender, 3 in Married, 15 in Dependents, 32 in Self\_Employed, 22 in Loan Amount, 14 in Loan\_Amount\_Term and 50 in Credit History.

**The column Loan\_ID is the unique ID given to the applicants, is has no significance in the prediction so let’s drop this column:**



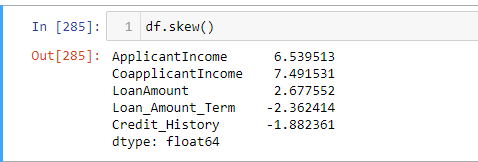
Hence, the column Loan\_ID is dropped.

**Visualizing null Values using Heat map:**



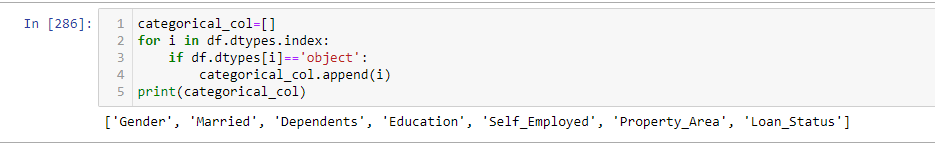
The white lines in the Heatmap represents the missing values in Dataset

**Checking the Skewness in Dataset:**



This shows the Skewness in present in each columns.

**Checking the Categorical Columns:**



The above info shows the categorical columns present in the Dataset.

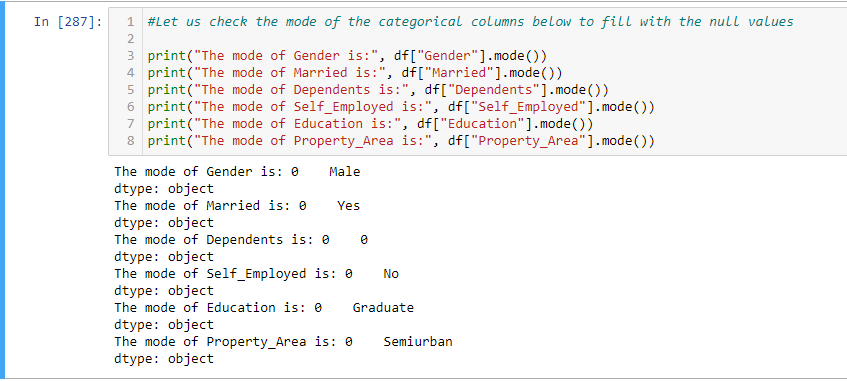
Except the Loan Amount and Loan\_Amount\_Term everything else which is missing is of type categorical. Hence we can replace the missing values by mode of that particular column. Before getting in to the code, I would like to say few things about mean, median and mode.

Mean is nothing but the average value whereas median is nothing but the central value and mode the most occurring value. Replacing the categorical variable by mode makes some sense.

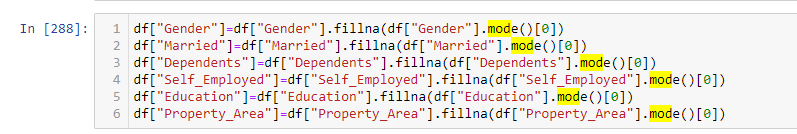
For example if we take the above case, 398 are married, 213 are not married and 3 are missing. So As married people are higher in number we are considering the missing values as married. This may be right or wrong. But the probability of them being married is higher. Hence I replaced the missing values by Married.

For categorical values this is fine. But what do we do for continuous variables. Should we replace by mean or by median.

**Treating Null values using Imputation Techniques:**

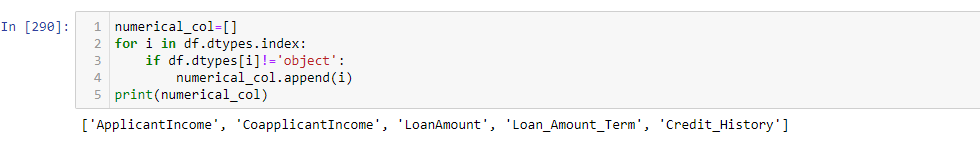


These are modes of the columns which contain null values. These are the values which are highly repeated in the columns. The missing values will be replaced by their respective mode values.

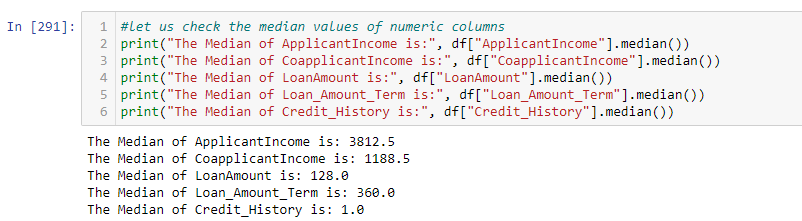


Hence, the missing values is been replaced with their respective mode values.

**Now checking the Numeric Columns present in the Dataset**:

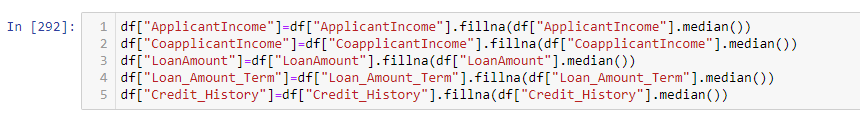


**Checking the Median values of Numeric Columns:**



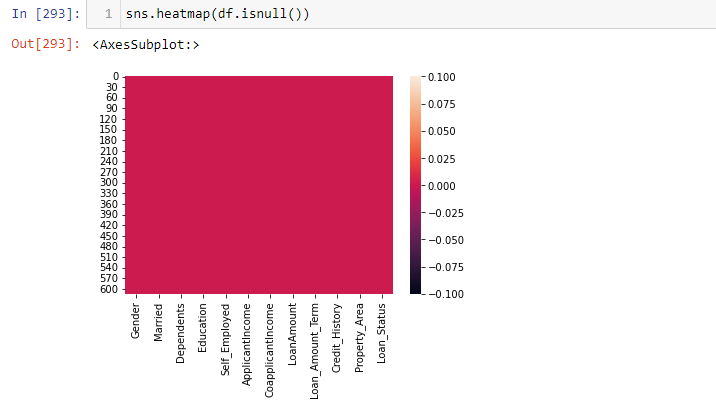
The above are the median values of numerical columns.

**Replacing the Numerical columns with their respective Median Values:**



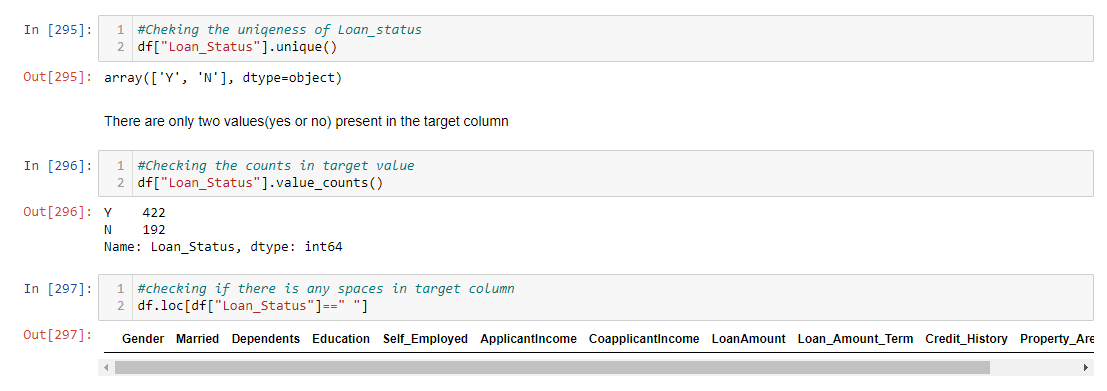
Hence, the above numerical columns is been replaced with their respective median values**.**

**Checking the null values after replacing with their respective mode and median values using Heatmap:**



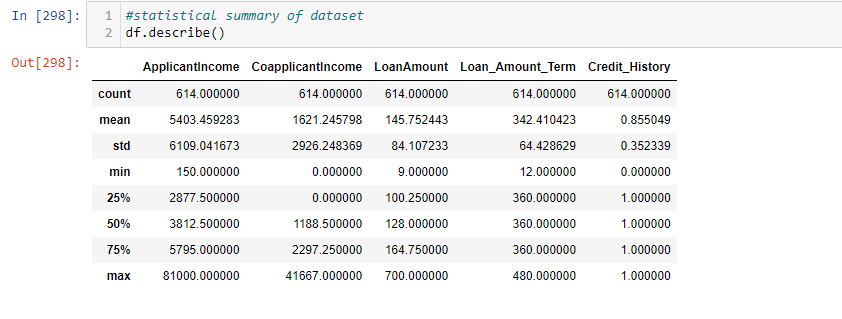
We can see that there are no null values present in Dataset.

**Checking the uniqueness, counts and spaces in the Target Variable Loan\_Status:**



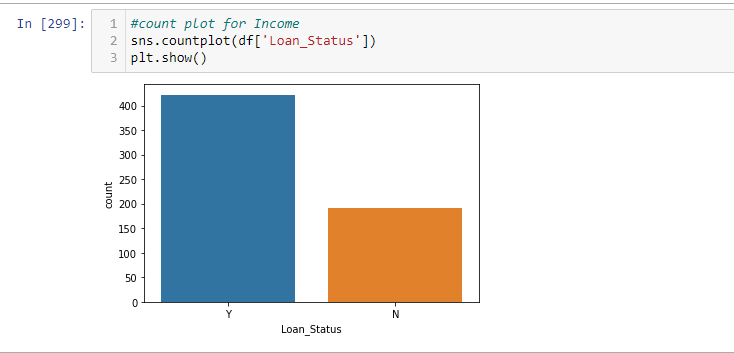
The target variable has two unique values yes and no. This variable has 422=Yes values and 192=No values. Also, we can see that this target variable has no spaces

**Statistical Summary of Dataset:**



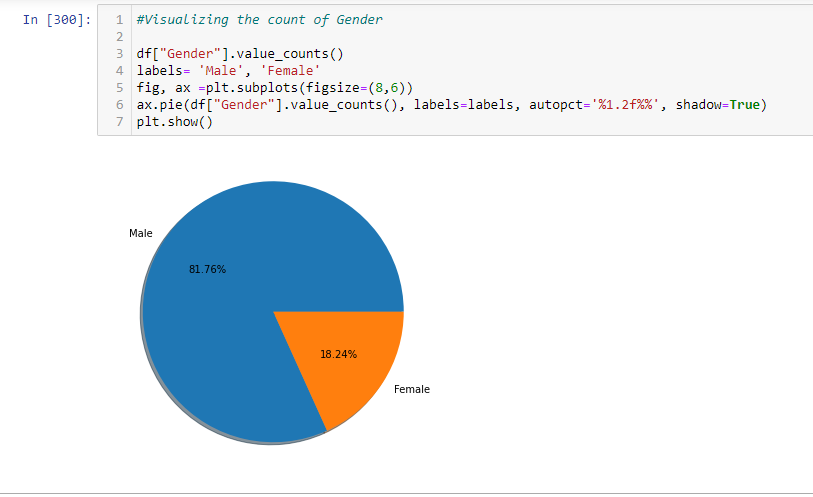
This gives the description of Dataset. The mean is more than median in all the columns except Loan\_Amount\_Term, it means that they are skewed to right. There is huge difference in max and 75% percentile which means there are outliers present in dataset.

**Data Visualization: Univariate Analysis**:



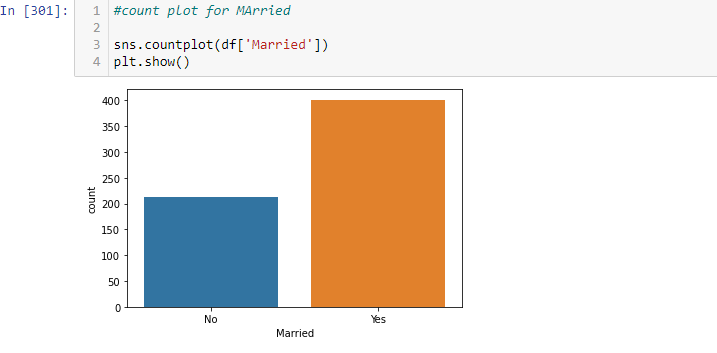
From the above graph we can conclude that more number of loan is approved and less has got denied.

**Visualizing the count of Gender:**



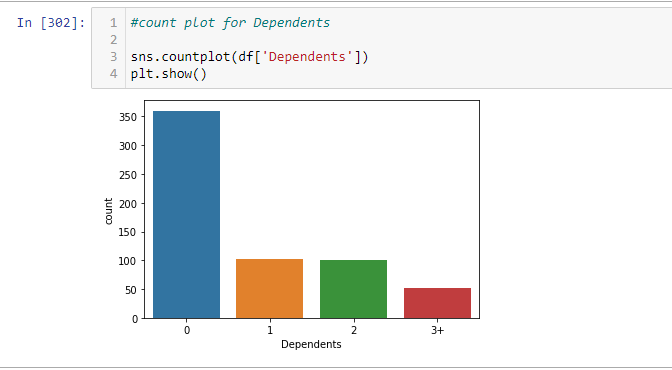
There are more number of Male Applicant (81.76%) for loan than Female (18.24%) Applicants.

**Countplot for Married Column:**



The number of Married applicant who applied for loan is higher than the unmarried applicants.

**Countplot for column Dependents:**

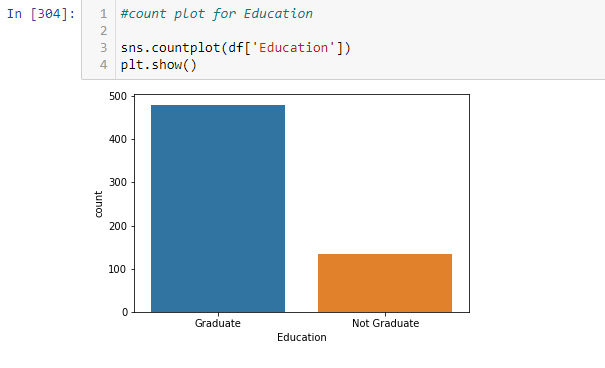


The applicant who have 0 dependent have high counts and the applicant having more than 3 dependent counts are very less.

**Countplot for column Self\_Employed:**

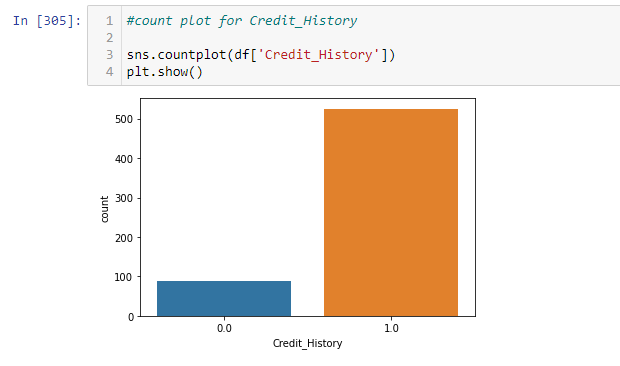
 Most of the loan applicant are not self employed.

**Countplot for Education:**

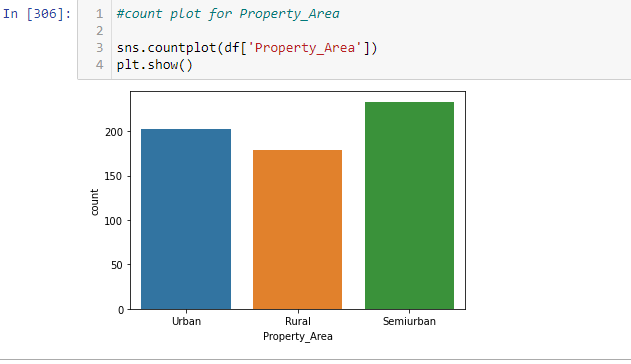


More number of people applied for loan are Graduates followed by not Graduates.

**Countplot for Credit\_History:**

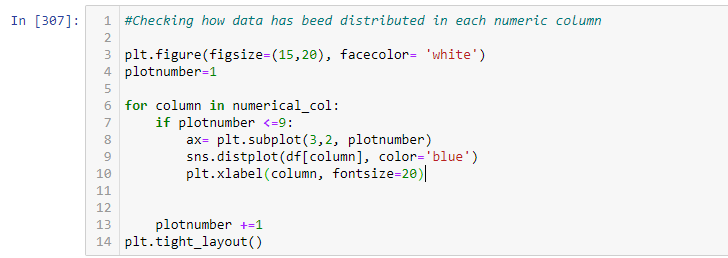
Most of the applicant who have credit history 1 are high in numbers.

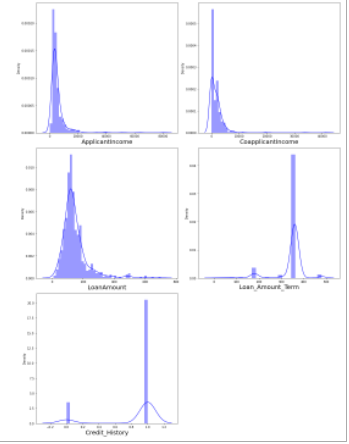
**Countplot for Property\_Area:**



Applicant having property in Semiurban area has more chance of getting loan approved.

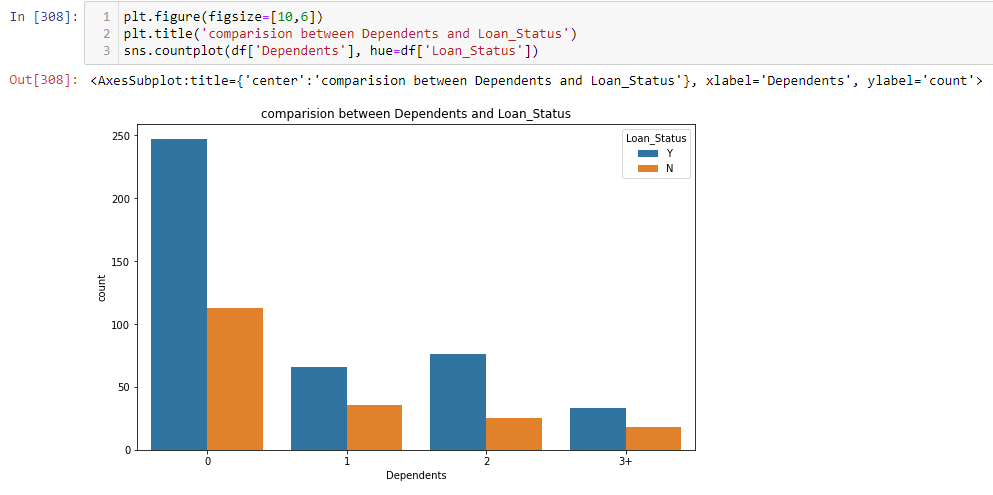
**Checking Data distribution in Numeric Columns:**





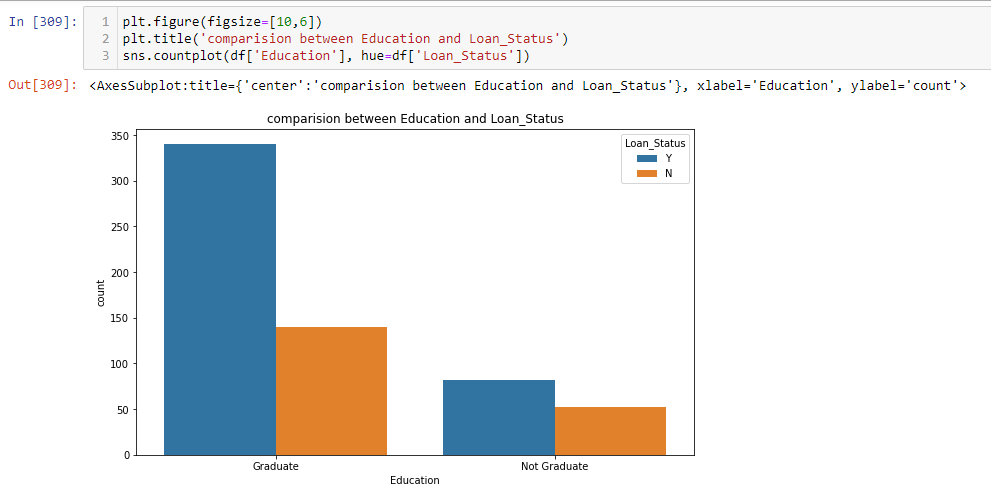
All the numeric columns are skewed.

**Bivariate Analysis:**



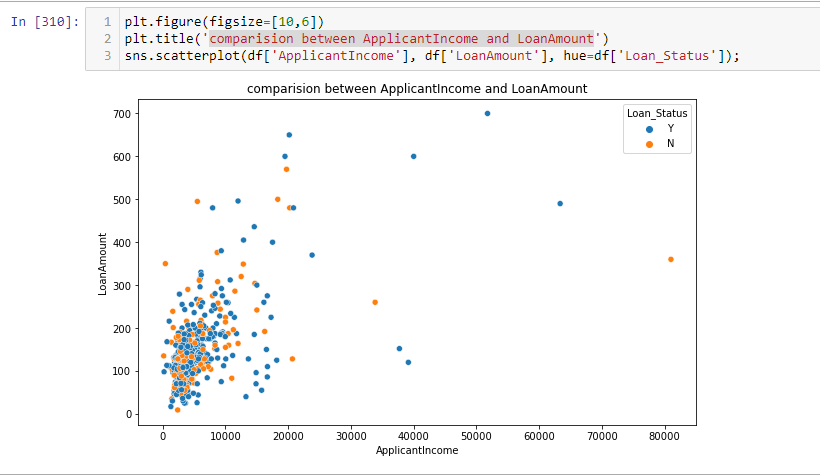
The count of 0 dependent is high followed by more number of dependents.

**Comparision between Education and Loan\_Status:**



The applicant who are graduate have tendency of getting loan than who are not.

**Comparision between ApplicantIncome and LoanAmount:**



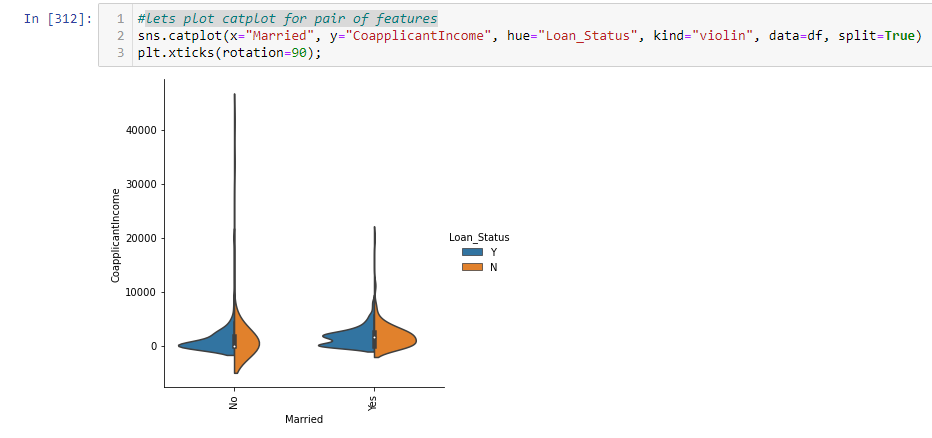
There is high density of points in the range of 0 to 2000 for applicants followed by others.

**Comparision between features using cat plot:**



Applicants who are not self employed has more chances of getting loan approved.

**Lets plot catplot for pair of features:**



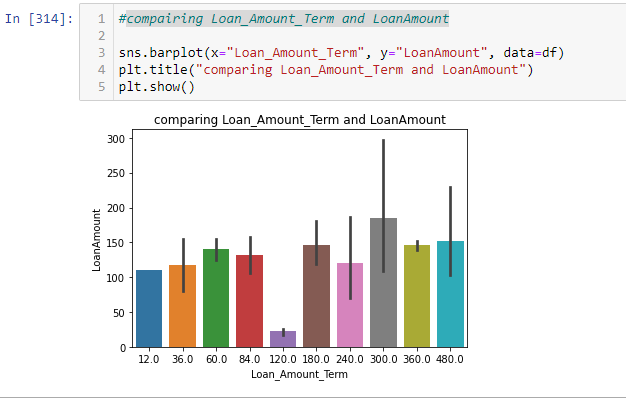
Married people has more chances of getting loan approved.

**Visualizing the property area of the applicants**:



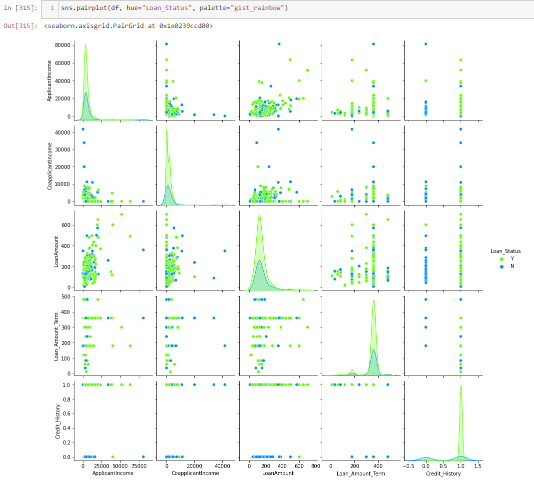
Most of the people from semiurban area are applying for loan followed by other areas.

**Comparing Loan\_Amount\_Term and LoanAmount:**



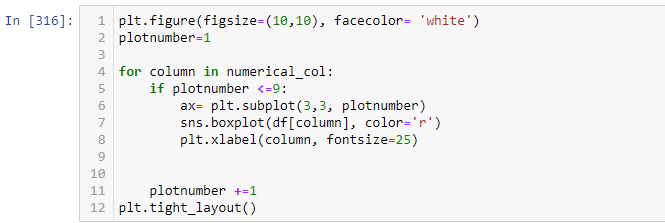
The loan amount term 300 is high with loan amout compare to other.

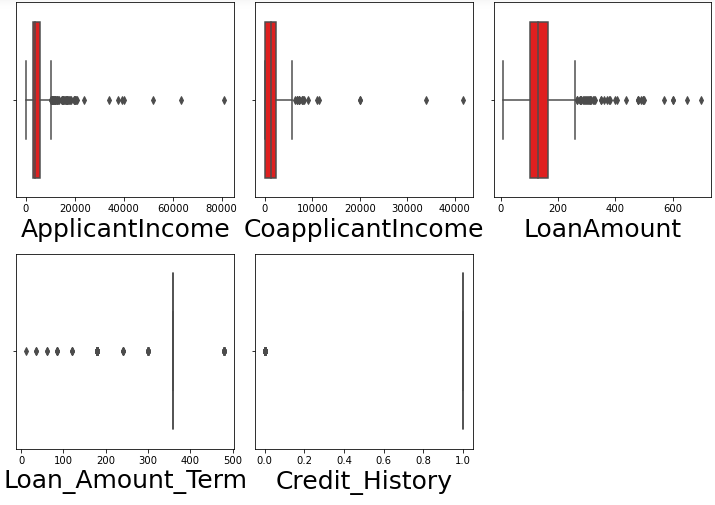
**Multivariate Analysis:**



Above is the pairplot for having Loan\_Status as target. There are some extreme outliers present in the dataset.

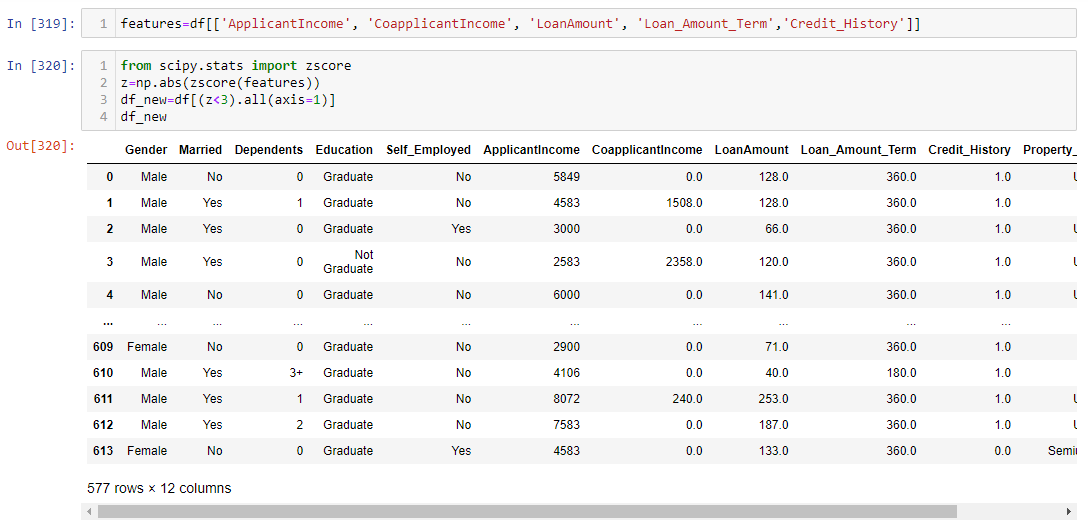
**Identifying the Outliers:**





since credit history and loan amount term are categorical column no need to remove the outliers. In all other we can remove the outliers.

**Outliers Handling:**



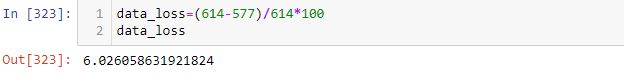
This is the dataframe after removing the outliers.

**Checking shape of Dataset before and after removing outliers:**



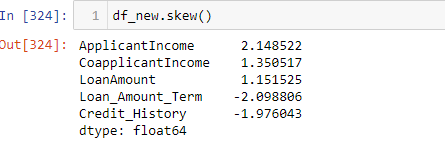
Before removing the outliers we had 614 rows and 12 columns and after removing outliers we have 577 rows and 12 columns.

**Checking the Dataloss:**



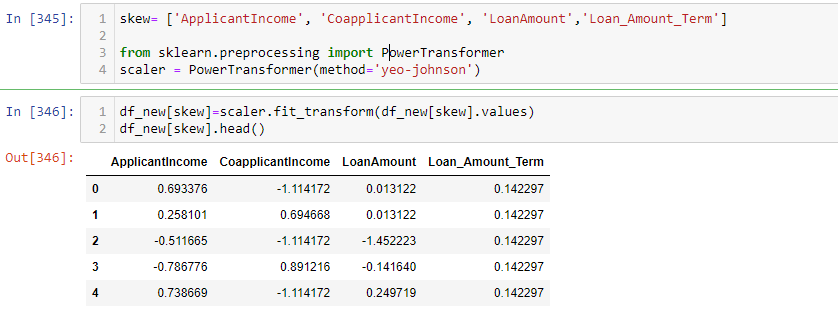
using Zscore i have data loss of 6% which is less than 10.

**Checking the Skewness:**

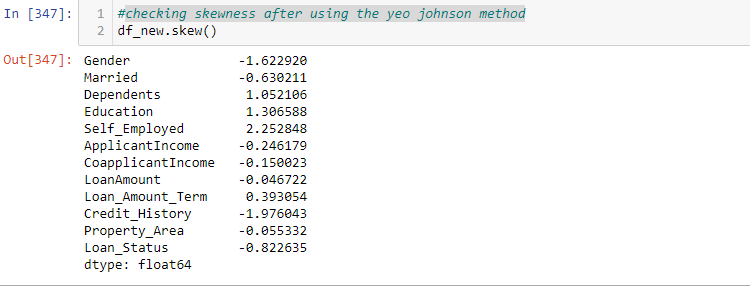


The skewness is present in all the columns.

Skewness more than+/- 0.5 will we treated but not treating the object and Target column, most of the columns are skewed but we have to check most of them are categorical features as well they must be imbalance that’s why showing as skewness.

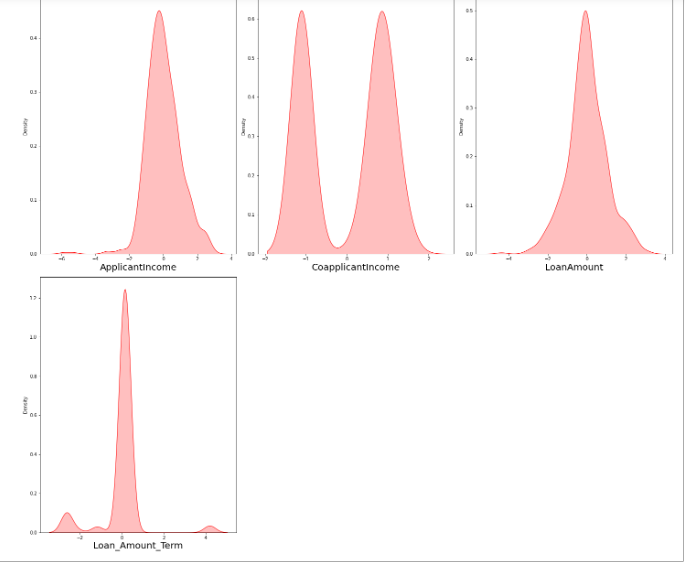


**Checking skewness after using the Yeo Johnson method:**



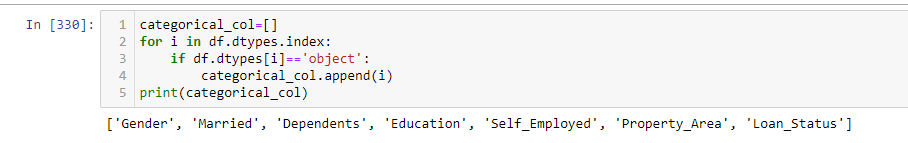
**After removing the skewness let us check the data distribution in each column:**





The data is almost normal also we have removed the skewness that we can notice in the above plot.

**Checking Categorical Columns:**

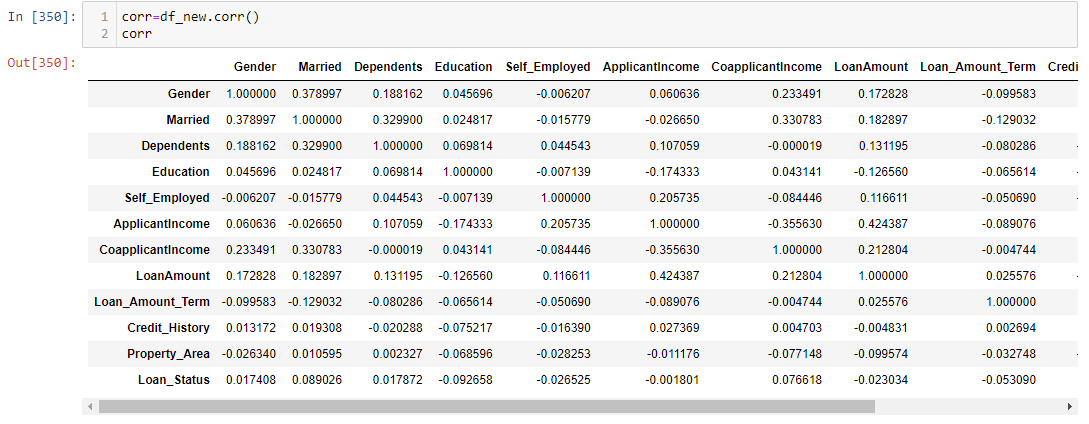


**Encoding the Categorical columns using Label Encoding:**



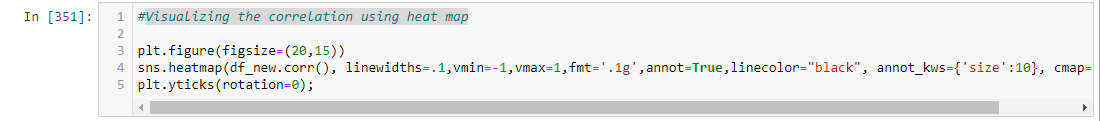
Hence, the categorical columns have been converted into numeric column using label encoder.

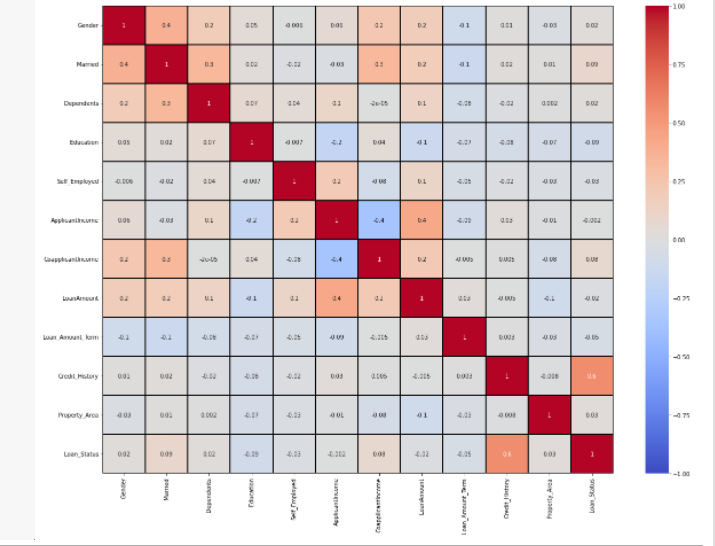
**Checking the correlation between variables:**



This gives the correlation between the dependent and independent variable. We can visualize this by plotting heatmap.

**Visualizing the correlation using heat map:**





There is no multicollinearity issue.

Credit history is highly correlated with target column.

Other features has less correlation with target column.

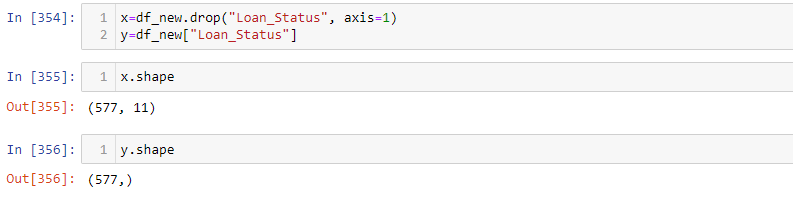
Dark shades are highly correlated and light shades are less correlated.

**Visualizing the correlation between label and features using bar plot:**

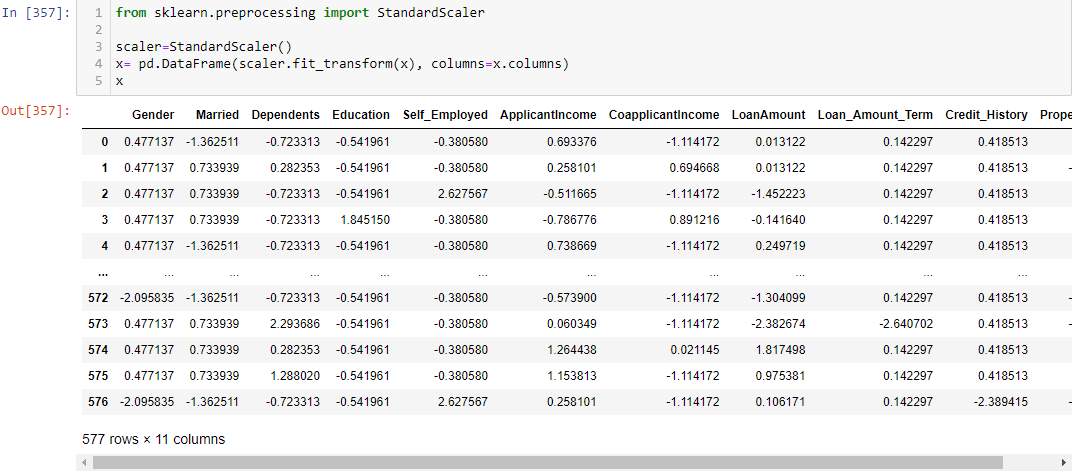


The column Applicant Income has very less correlation with label so we can drop it.

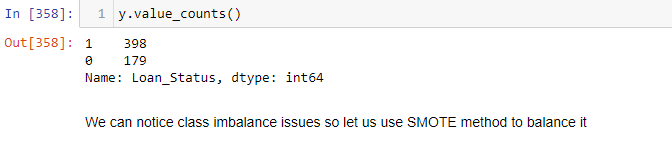
**As the data cleaning and data structuring are done, we will be going to our next section which is nothing but Model Building. Separating the features and label variables into x and y:**



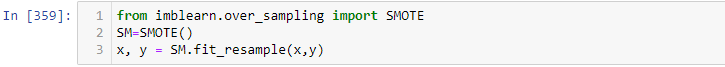
**Scaling the data using Standard Scaler:**



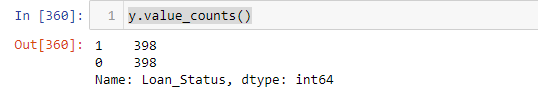
**Checking the value count of Target Variable:**



**Balancing the target variable using Oversampling:**

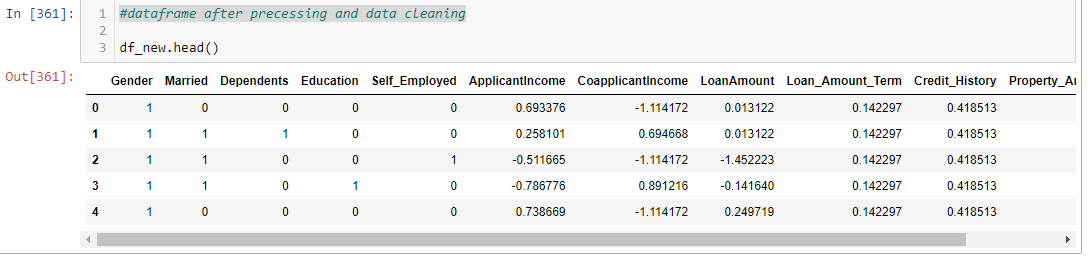


**Checking the value count of Target Variable after using Oversampling:**



Hence, the data is now balanced.

**Dataframe after processing and data cleaning:**

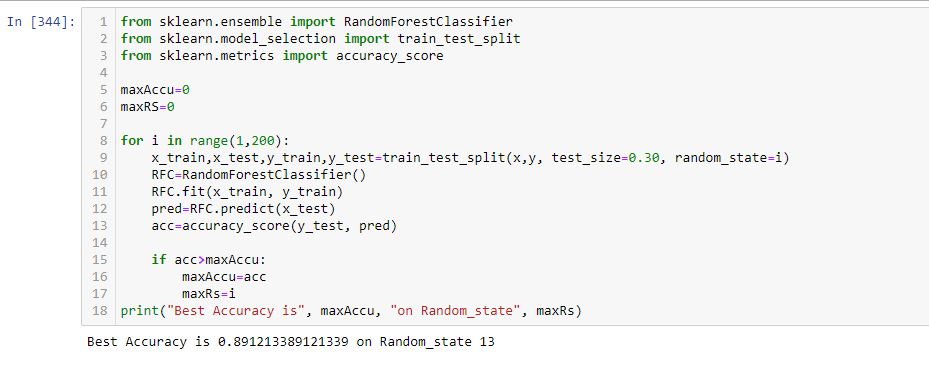


**Building Machine Learning Models**:

**SAMPLING TECHNIQUES AND NEED FOR THAT**:

There are many sampling techniques like Random Sampling, Stratified Sampling etc. The major purpose is to improve the accuracy which can be obtained by hiding some portion of train data and running the model so that on an average the one that gives higher accuracy can be taken for test data.

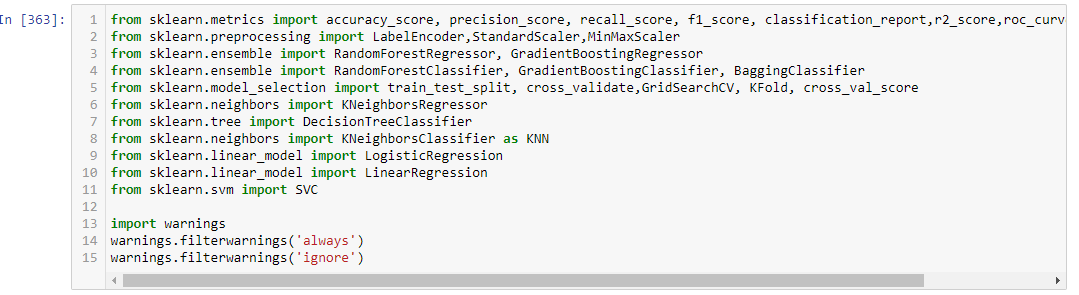
**Finding the best Random State and accuracy:**



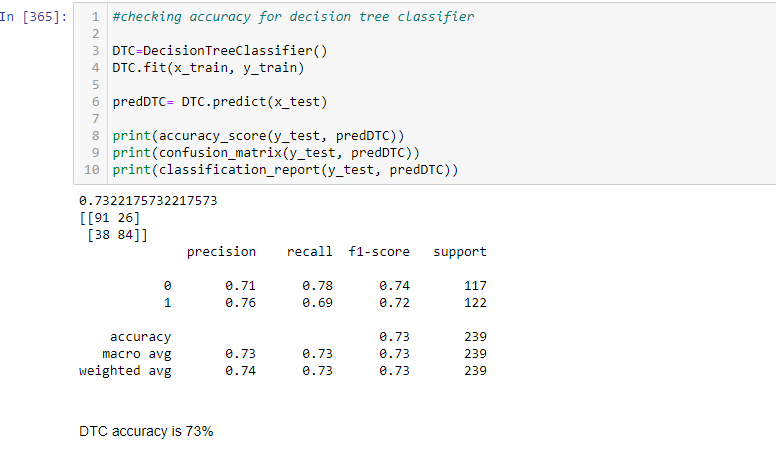
**Splitting the Data into Train and Test:**



**Importing important Libraries:**

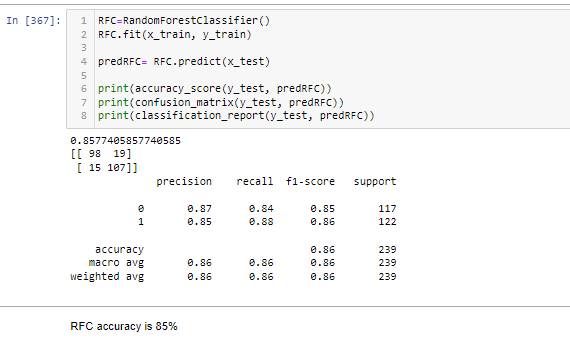


**Decision Tree Classifier:**

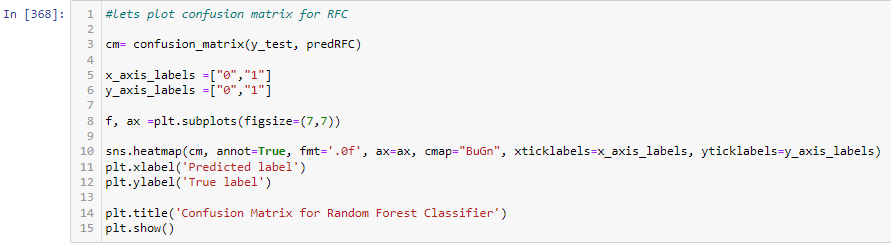
**Let’s plot confusion matrix for DTC:**

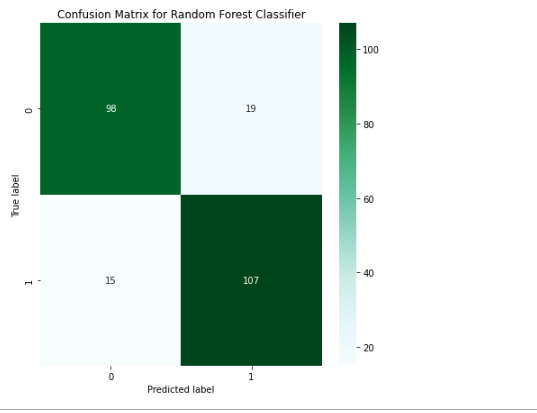


**Random Forest Classifier:**

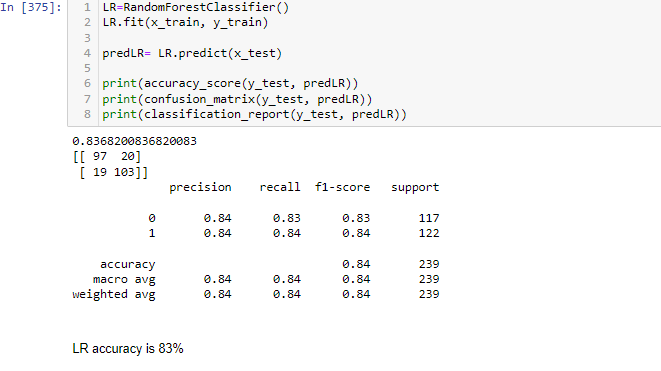


**Lets plot confusion matrix for RFC:**





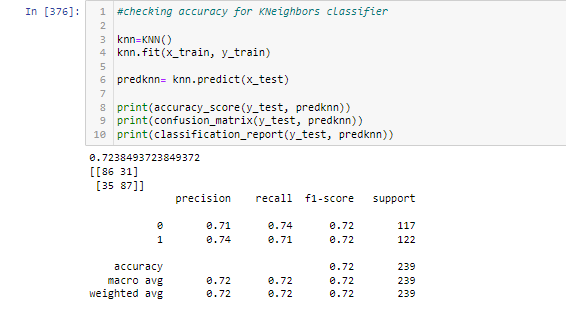
**Logistic Regression Classifier:**



**Lets plot confusion matrix for Logistic Regression:**



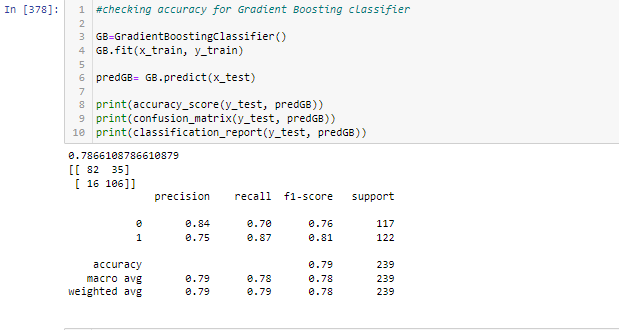
**KNeighbors Classifier:**



**Lets plot confusion matrix for KNN:**



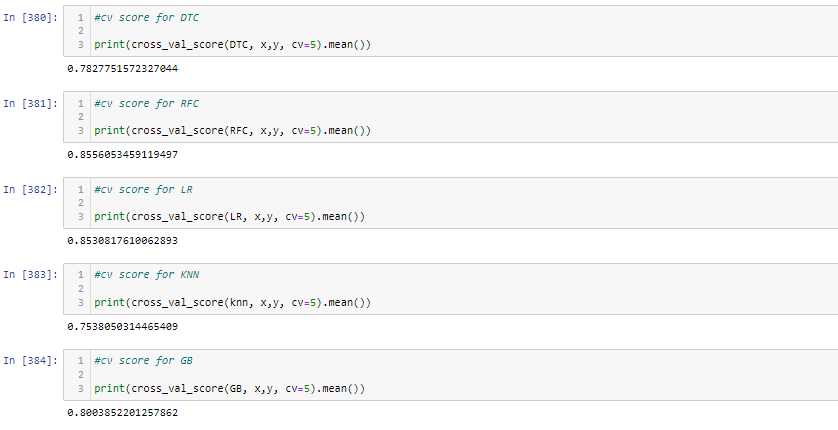
**Gradient Boosting Classifier:**



**Lets plot confusion matrix for GB:**



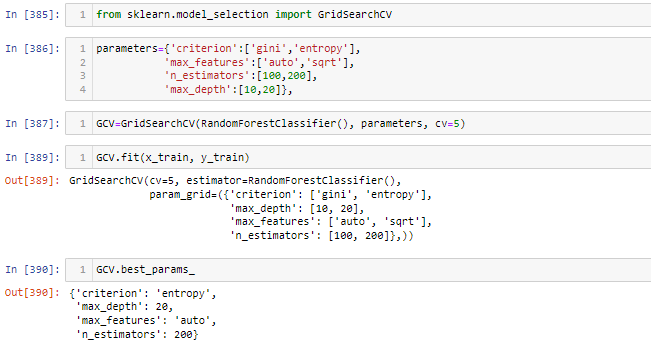
**Checking the Cross Validation Score:**



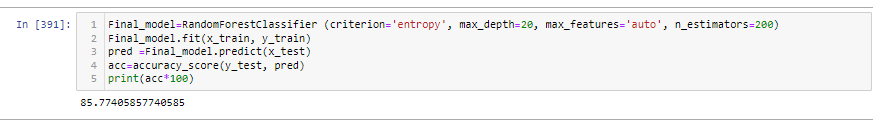
From the difference between accuracy\_score and CV score we can conclude that Random Forest Classifier is our best fitting model.

# HYPER PARAMETER TUNING:

To enhance the accuracy tuning the model. I have used the **GridSearchCV** grid for tuning the model so we can get the best parameters.

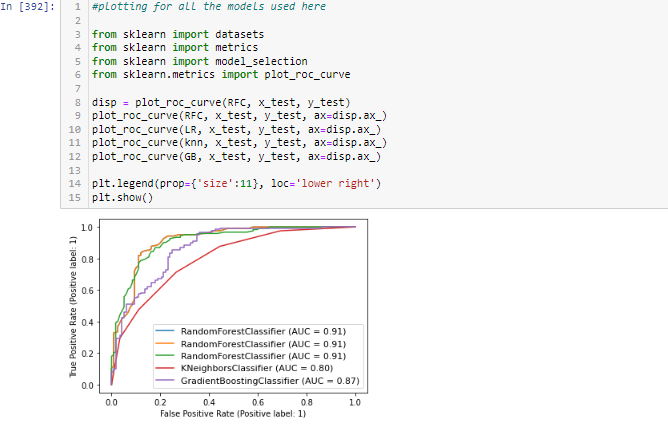


After tuning the model we got the best parameters for model. Lets train the model with best parameters.

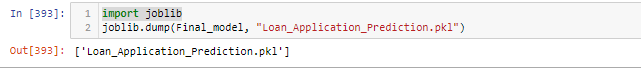


As we can see average accuracy is 85.77%. I have tried with other parameters as well but this the best I got.

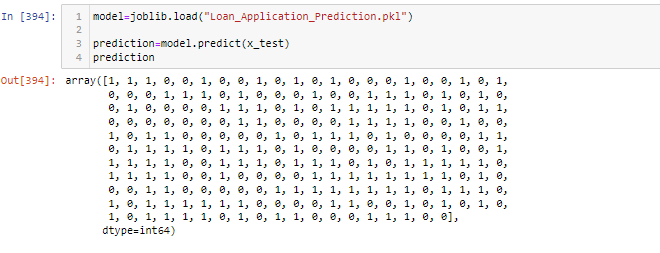
**ROC AUC Curve:**

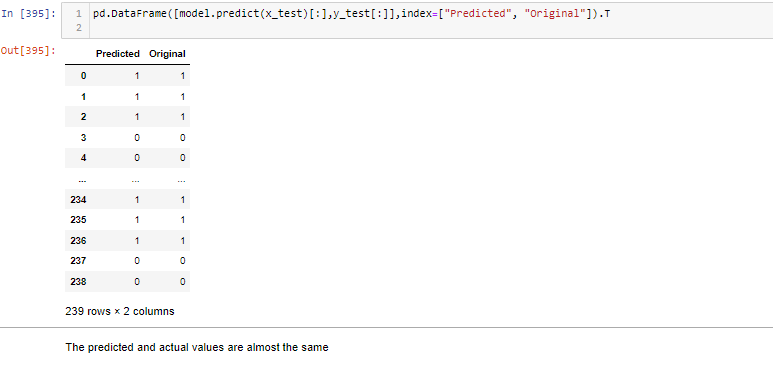


**Saving the Model:**



**Predicting the Saved model:**





***CONCLUSION*:**

Key Findings and Conclusions of the Study:

**CONCLUSION:**

Key Findings and Conclusions of the Study:

* 1. There are high chances of approval for Graduate people comparing not graduate
  2. There is high chances of Loan Approved when you have credit history, people those are not having any credit history mostly getting not approved
  3. People from Semiurban area are having high chances to get their loan approved comparing people from other area.

There are high chances for loan approval when you taking loan for less.