**Loan Application Status Prediction Using Machine Learning**



**Declaration**

I declare that this dissertation entitled “Loan Approval Prediction Using Machine Learning Techniques” is the result of my own work and analysis.

**Author - Deepak Kumar**

**Publish Date – April 2,2022**

**B-Tech in Information Technology**

**Email ID -** [swrrocky17@gmail.com](mailto:swrrocky17@gmail.com)

**Contents**:

Abstract …………………..………………………………………………….……………..1

Data Description…….……………………………………………………………………2

Data Preparation …………………………………………………………………………3

Importing modules………………………………….………………………3.1

Handle null values ………………………………………………………....3.2

Data Duplicates……………………………………………………………….3.3

Visualization and data cleaning………….……………………………3.4

Data Exploration and Transformation………………………………3.5

Split dataset…………………………………………………………………….3.6

Normalization………………………………………………………………….3.7

Feature Importance………………………………………….…………….3.8

Model Building ………………………………………………………………………….….4

Evaluation………………………………………………………………………...……………5

Save model …………………………………………………………………………………….6

End Notes……………………………………………………………………………………….7

1. **Abstract**

Loan is the essential product of banks and other financial institutions. As a big number of people go to banks to borrow money for different activities, the number of customers have increased and some banks expect to earn a lot of money as a result of interest paid on loans. However, loans are associated with risk of defaulting, i.e. the possibility that some borrowers may not be able to pay back their loans. Thus, high levels of non-performing loans can be a source of instability of the banking sector and lead to bankruptcy. One of the important steps for banks to decide if a loan has to be authorized is to ensure that the candidate to borrow has the capacity of paying back the loan in the proposed terms. The advancement of technology like machine learning, computer science and other science is playing an important role by supporting banks to predict the probability of defaulting for a given customer based on his past behavior. With past behavior, it means the historical record of the customers which banks keep with themselves.

In my research, I have taken the data of an unknown bank to predict the loan status of an applicant. I would recommend financial institutions to use machine learning techniques because it saves money and time for both sides. The finding shows that Credit History is the most important feature while checking the approval of Loan status. Credit History includes information on your present and previous credit accounts such as loans, mortgages, credit cards, payment history, account balances and the duration of each account being active or open. If the credit History is good, then it is 1 or else it is 0.

Acquiring loans for different purposes 2 such as the home loan, education loan, car loan, business loans etc., has become part of our day-to-day life from different financial institutions like credit unions and banks. However, many people are unable to determine the total amount of credit that they can afford to pay back. Analysis of creditworthiness is one of the most important for banks and other financial institutions to stay working in the highly competitive market and for their profitability. They must set clear and defined criteria for lending. These criteria must be sufficient and adequate to provide the required information about the structure of credit, borrowers, and mode of payment.

**2. Data Description**

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan\_Status in train dataset and 12 independent variables in test dataset. The Loan\_ID, Gender, Married, Education, Self\_Employed, Property\_Area, Loan\_Status are all categorical. Below is the brief description of all data.

* **Gender** - It is nothing but the sex of the applicant
* **Married** - It gives us information if the applicant is married, single or divorced
* **Dependents** - While applying for the Loan, Banks check how many dependents are there with the applicant.
* **Education** - It describes how much educated is the applicant.
* **Self\_Employed** - It gives us information if the applicant is self-employed or does any job.
* **Applicant Income** - It is the income of the applicant applying the loan.
* **Co-applicant Income** - It is the income of the co-applicant applying the loan. Generally, all bank checks the total income of the family applying the loan.
* **Loan Amount** - It is the amount of the loan sanctioned to the applicant.
* **Loan Amount Term** - It is the amount of the loan issues for the number of the months(In Short Loan period)
* **Credit History** - Credit History includes information on your present and previous credit accounts such as loans, mortgages, credit cards, payment history, account balances and the duration of each account being active or open. If the credit History is good, then it is 1 or else it is 0.
* **Property Area** - In which area does the property belongs, also is one of the factor for the bank to issue loan

**3.Data Preparation**

**3.1 Importing modules**



Python Libraries

The models are implemented using Python 3.7 with listed libraries:

Pandas is a Python package to work with structured and time series data. The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling, whereas the **NumPy** module works with the numerical data.

**Matplotlib** is well connected with Numpy and Pandas and acts as a graphics package for data visualization in python. Pyplot provides similar features and syntax as in MATLAB. Therefore, MATLAB users can easily study it.

**Seaborn** is a python library for building graphs to visualize data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data

**Sklearn** python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions.

Import the data from **Github** location.



The data in Github is in csv format, we are using Dataframe to read the csv file.

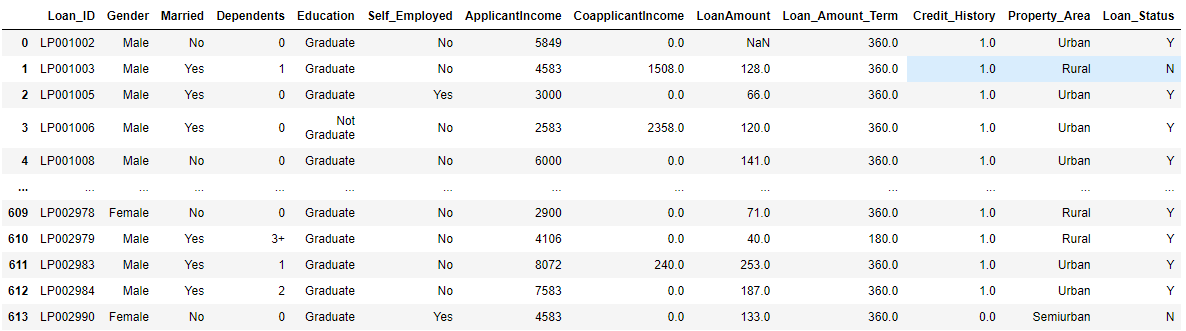
**pandas.DataFrame**

**class pandas.DataFrame(data=None, index=None, columns=None, dtype=None, copy=None)**[**[source]**](https://github.com/pandas-dev/pandas/blob/v1.4.2/pandas/core/frame.py#L459-L10970)

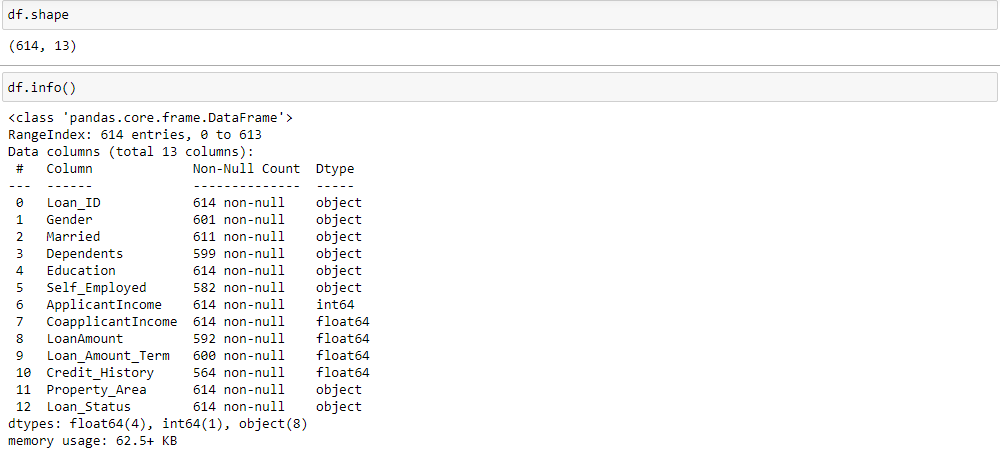
Two-dimensional, size-mutable, potentially heterogeneous tabular data.

Data structure also contains labeled axes (rows and columns). Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary pandas data structure.

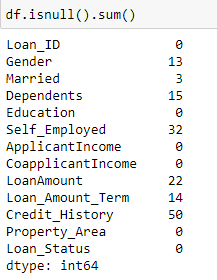
**Output**.



Shape and data info

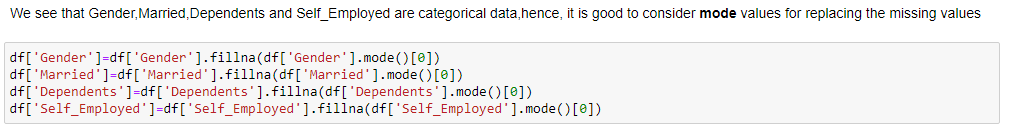


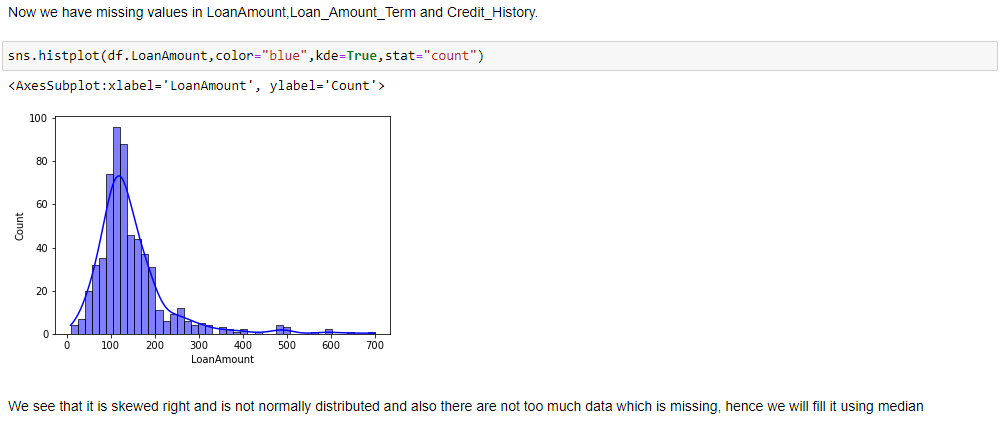
**3.2 Handle null values.**



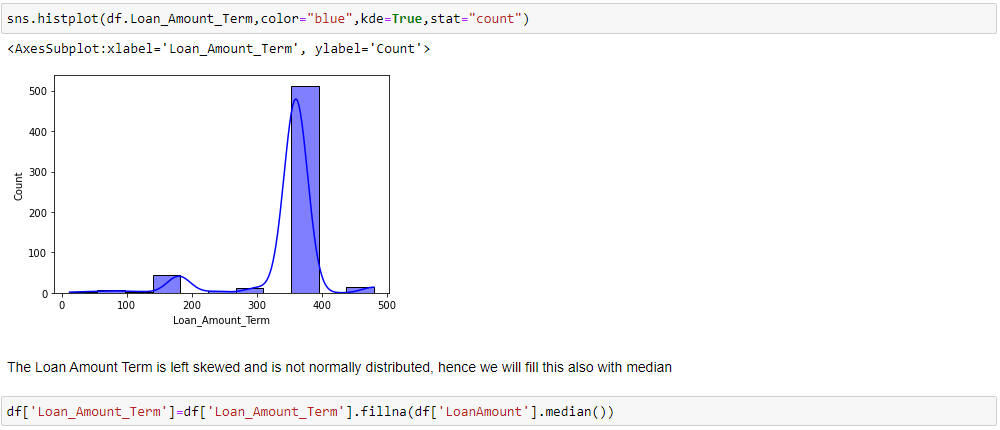
We see null values in Gender, Married, Dependents, Self\_Employed, Loan\_Amount, Loan\_Amount\_Term, Credit\_History.

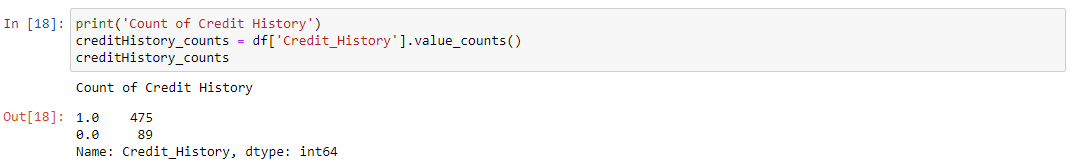
* We know that imputing missing data with **mean** values can only be done with **numerical** data.
* Another technique is **median** imputation in which the missing values are replaced with the median value of the entire feature column. When the data is skewed, it is good to consider using the median value for replacing the missing values. We must note that imputing missing data with median value can only be done with **numerical** data.
* Yet another technique is **mode** imputation in which the missing values are replaced with the mode value or most frequent value of the entire feature column. When the data is skewed, it is good to consider using mode values for replacing the missing values. We must note that imputing missing data with mode values can be done with **numerical** and **categorical** data.

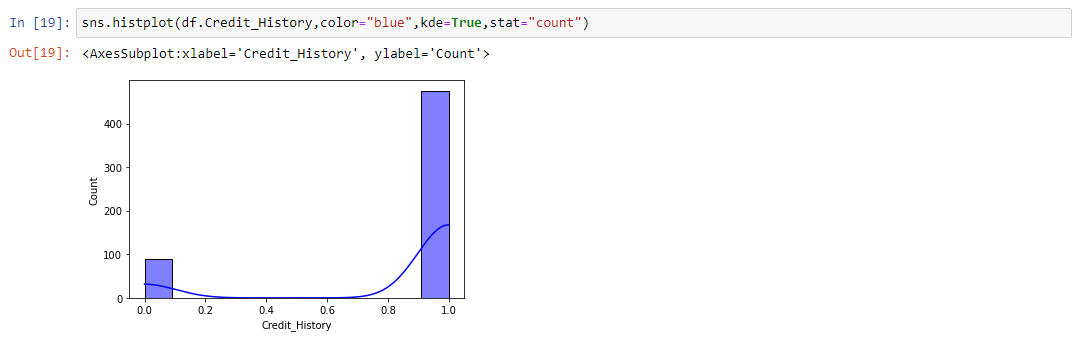




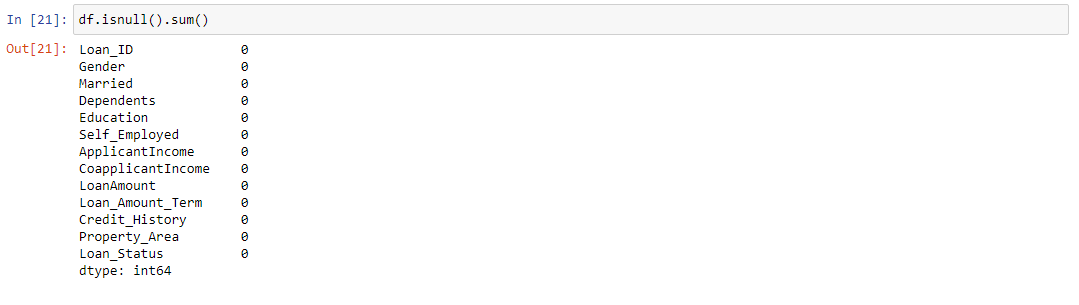








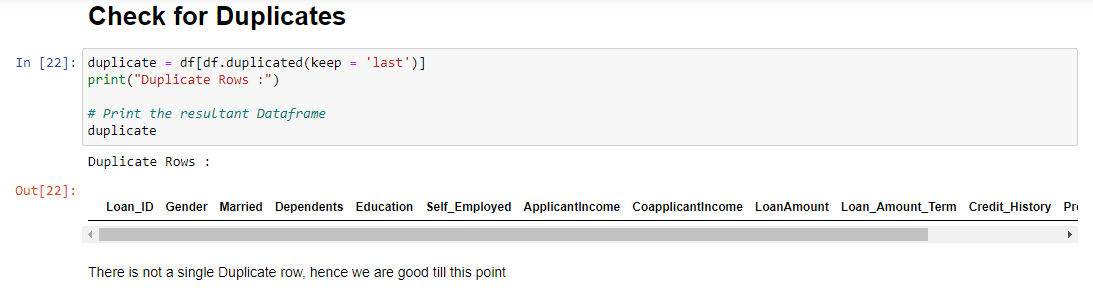
Now check for null values.



We don’t see any null values now.

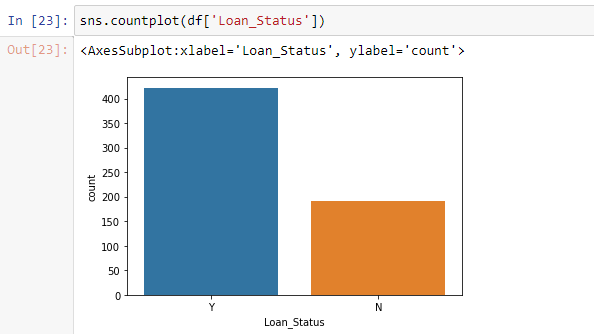
**3.3 Delete duplicates**

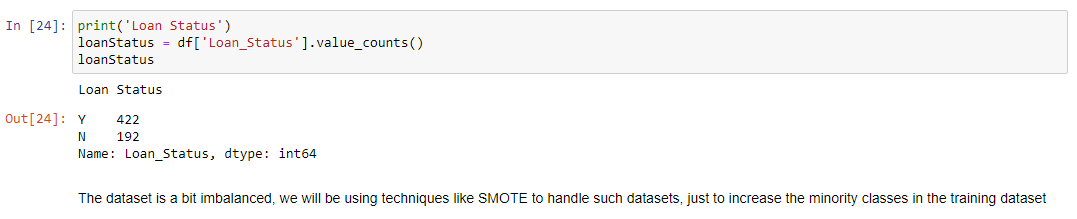
In data cleaning we check for duplicate and also if few columns are not required, we will drop them.



**3.4 Visualization and Data cleaning.**

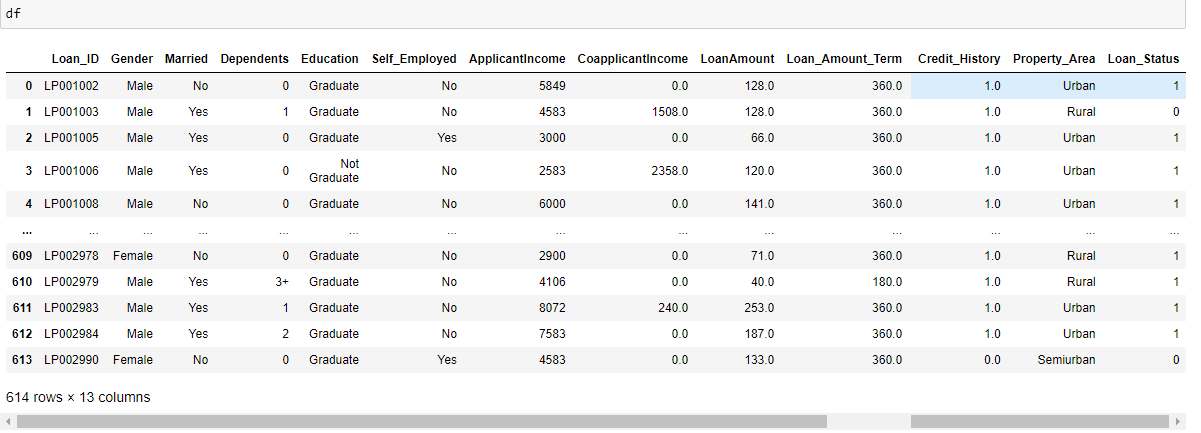
Virtualization is nothing but graphic representation of data that is used to find useful information.

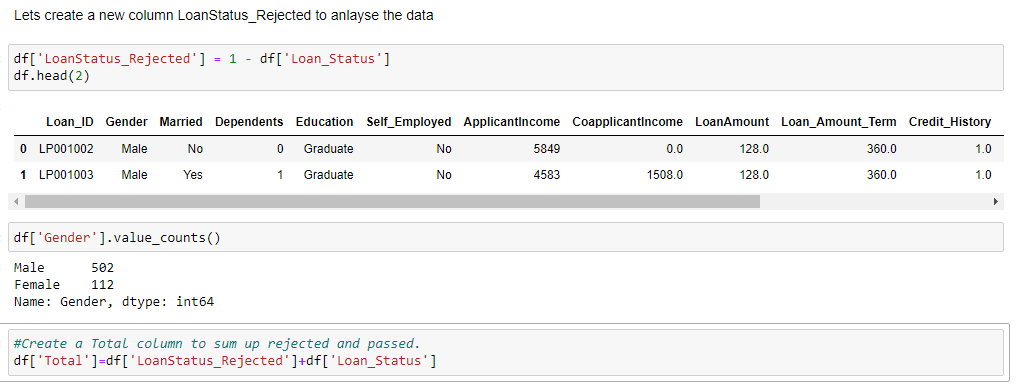


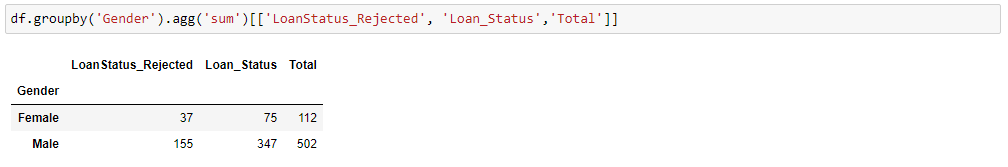


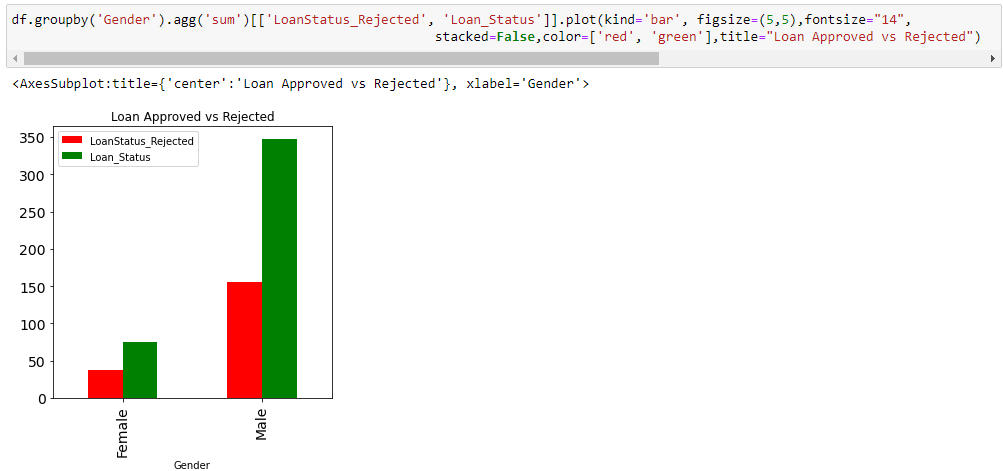


When we check, we see that Loan\_status is converted into 1 and 0, where 1 means Y and 0 means N



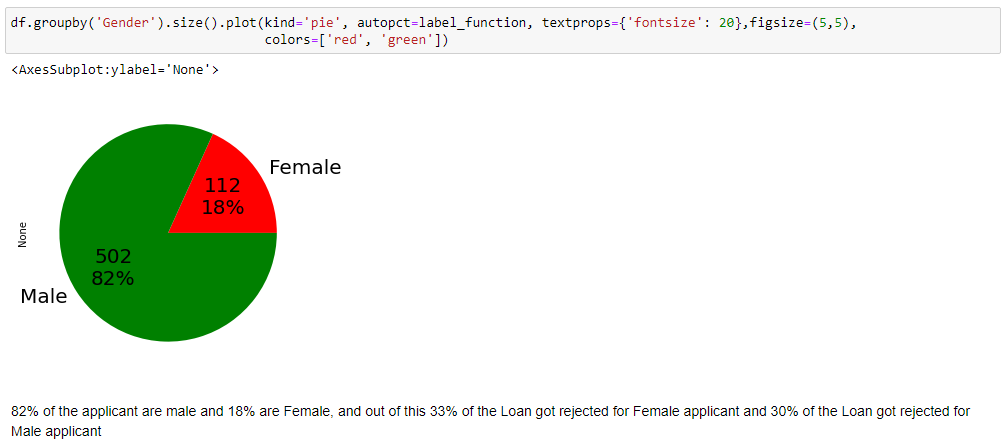


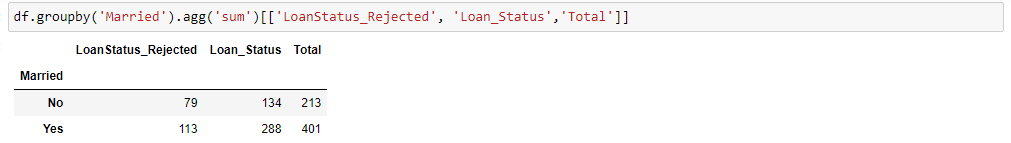


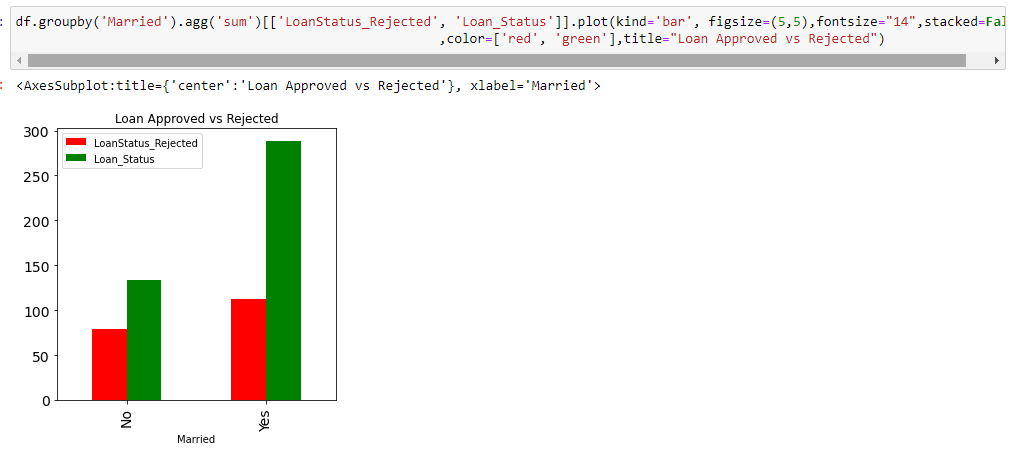


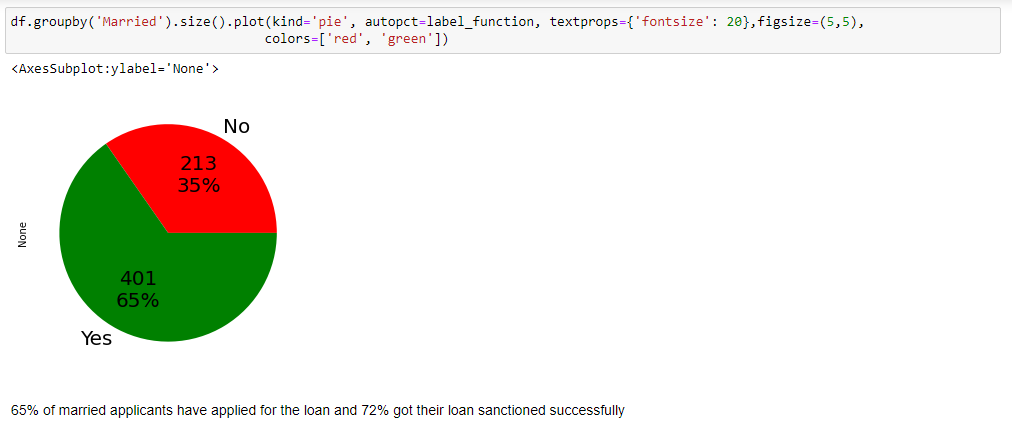
Let’s do the further analysis on Gender column using Pie chart. For that we defined a function which we will use at multiple areas.

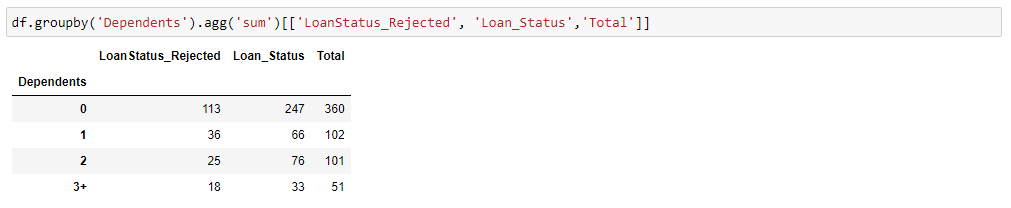


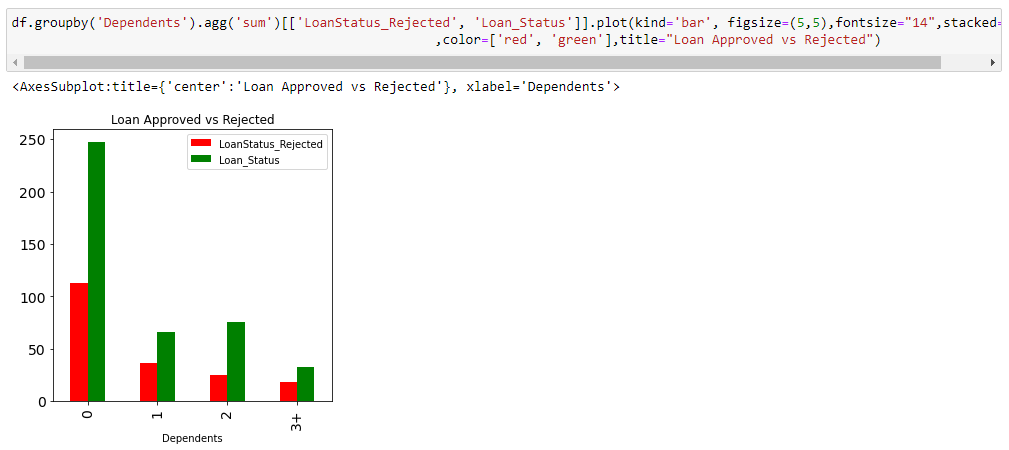


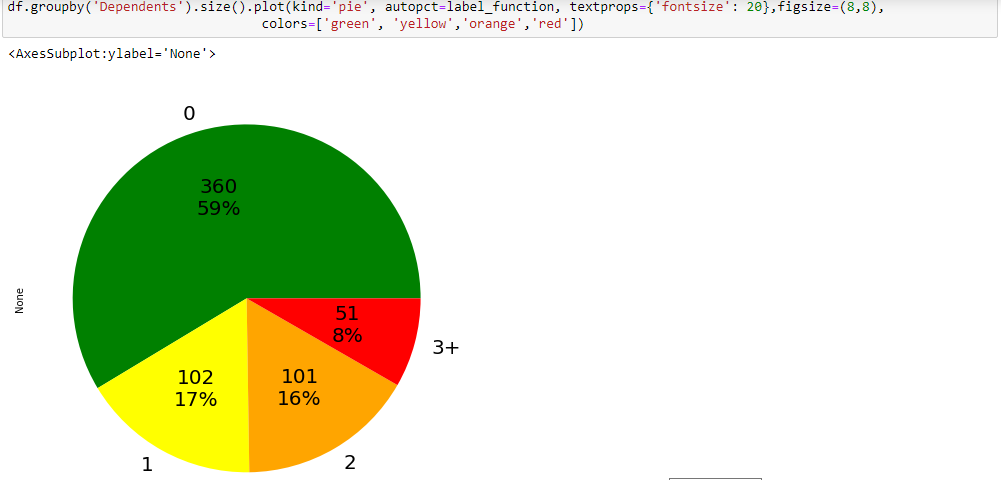


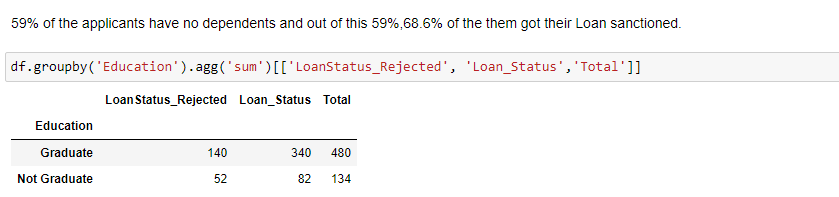


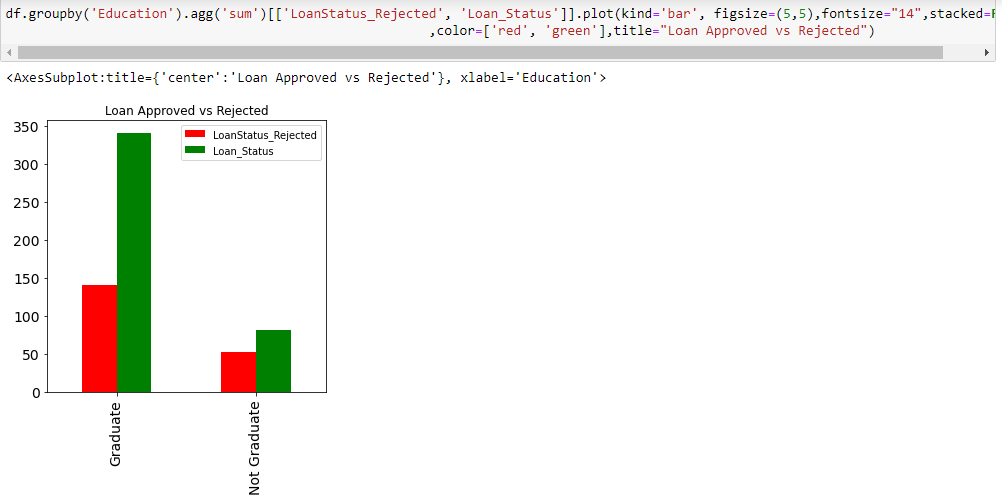




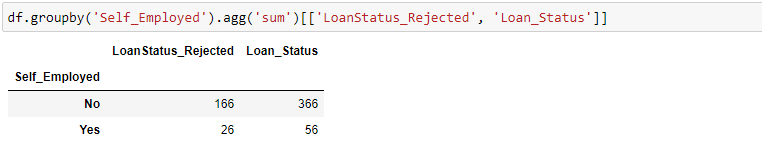


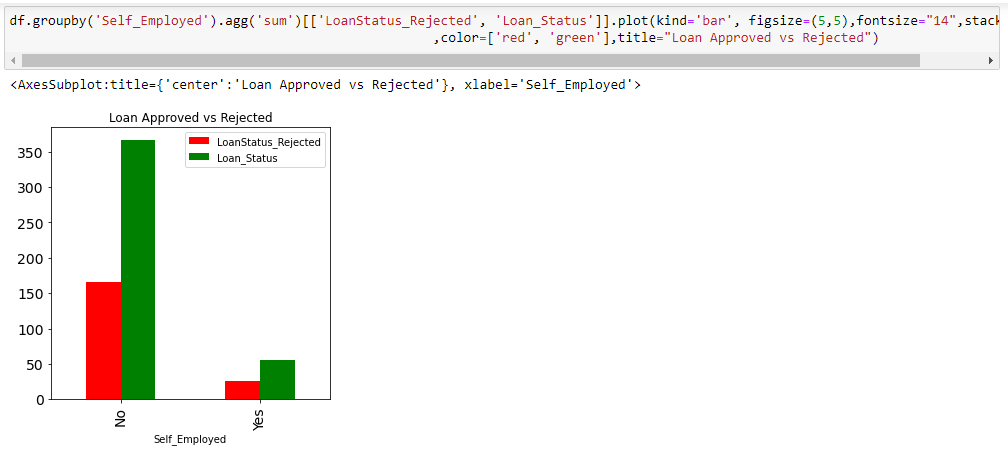


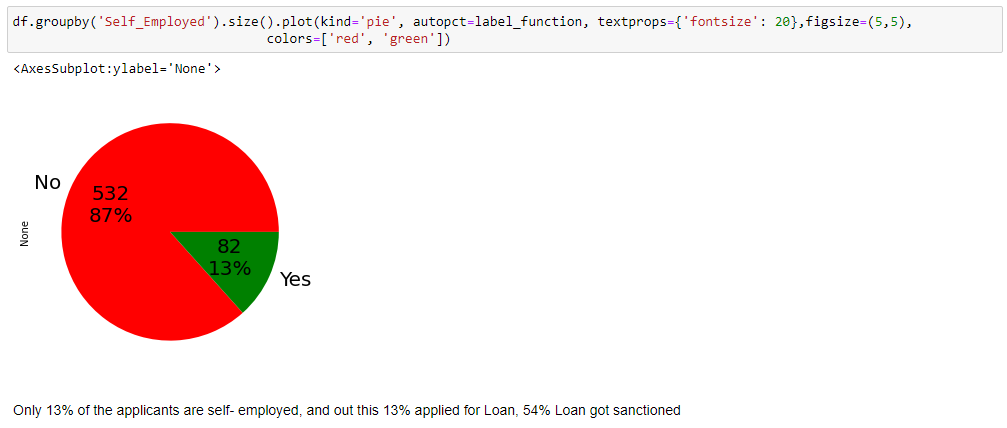


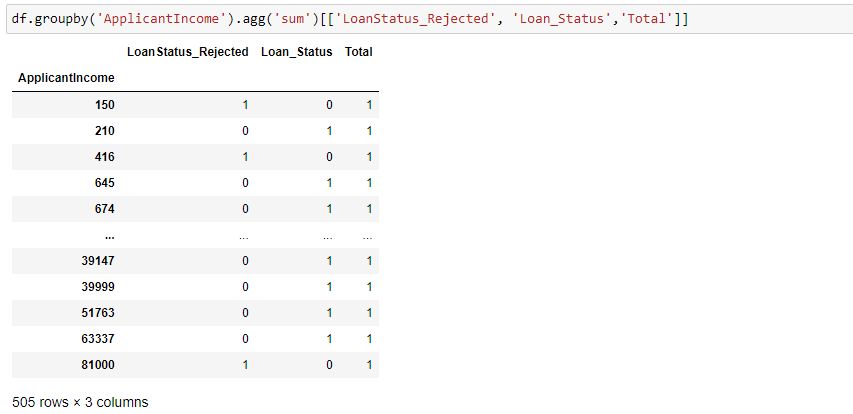






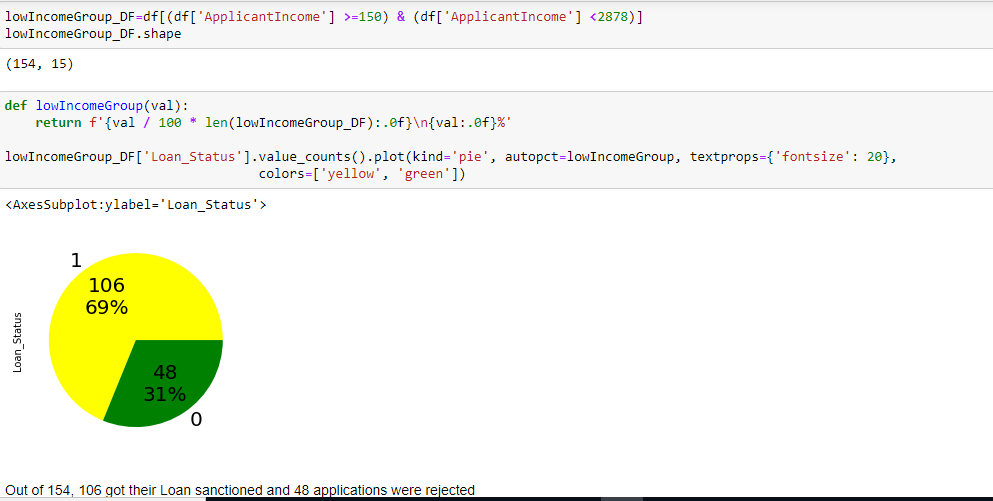






Most of the Applicant Income lies between 2877 to 5795, and co-applicant income is between 0 to 2297, probably most of them are housewife's. There are few outliers are in the income of Applicant and co-applicant.

Now we are dividing the Applicant Income in the range of 150 to 2878 as low income, 2878 to 3813 as medium Income group, 3813 to 5795 as High Income Group and more than 5795 as Rich for further data analysis.



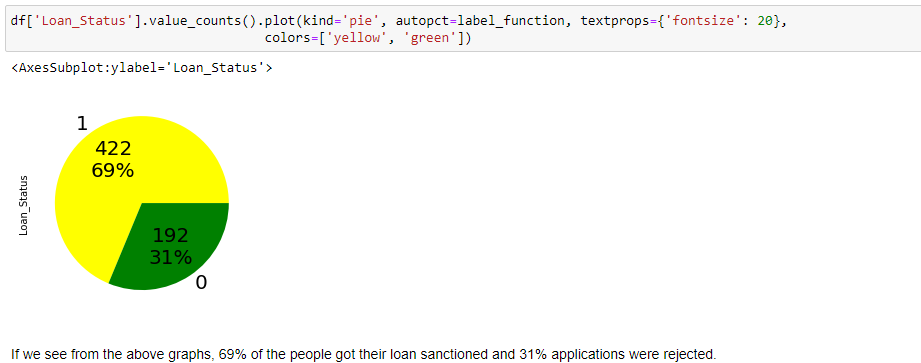


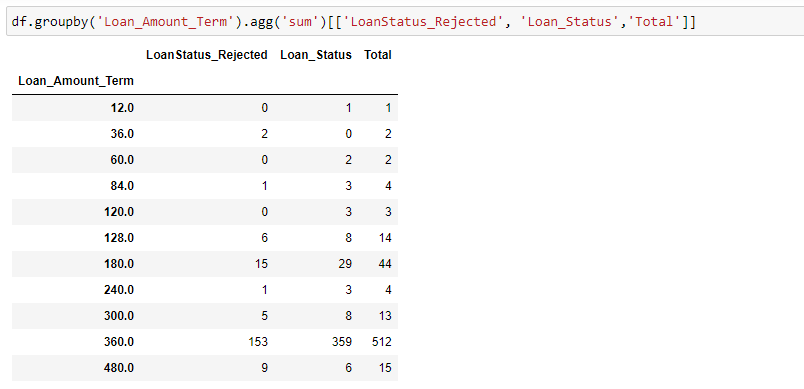
Out of 154 in medium earning range, 106 got their Loan sanctioned and 48 applications were rejected.



Out of 152 in High Income group, 105 got their Loan sanctioned and 47 applications were rejected

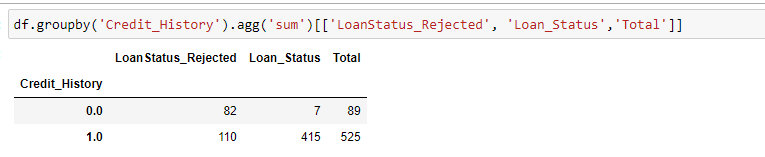


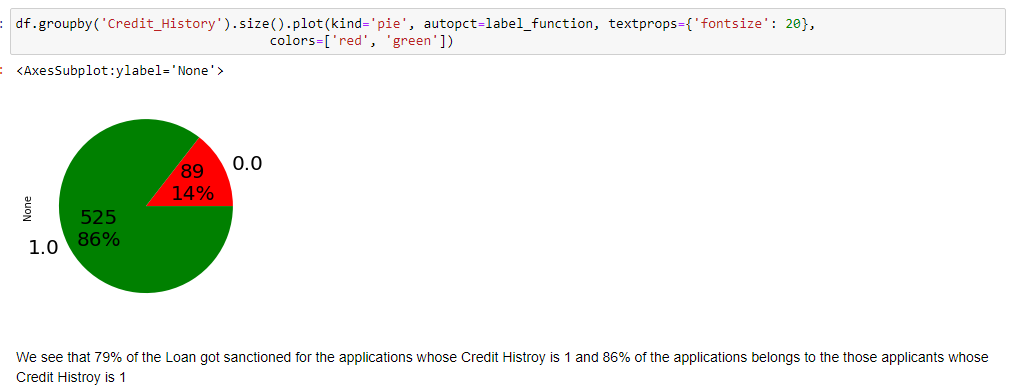


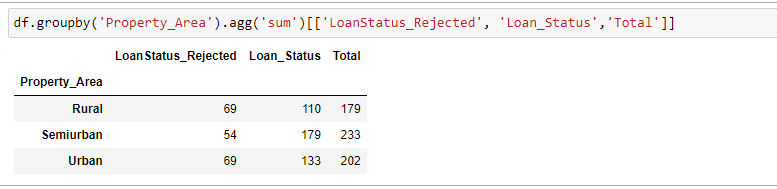




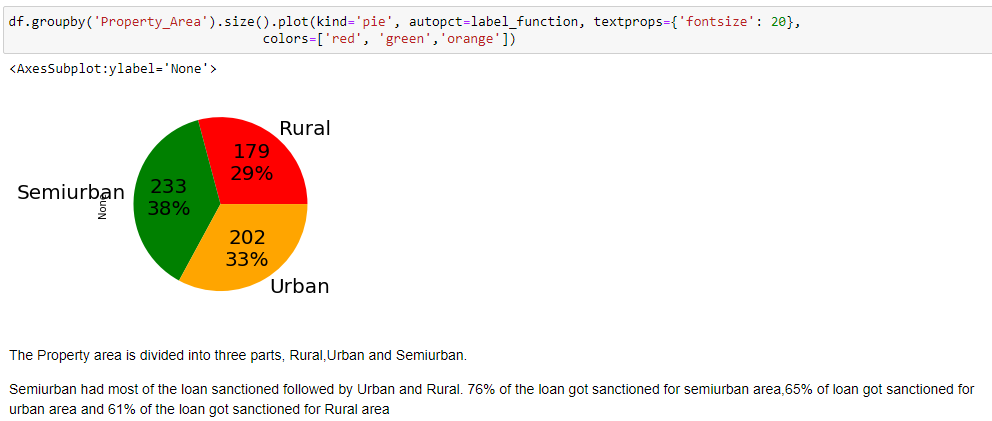
Maximum people applied for Loan term of 360 months i.e. 30 years

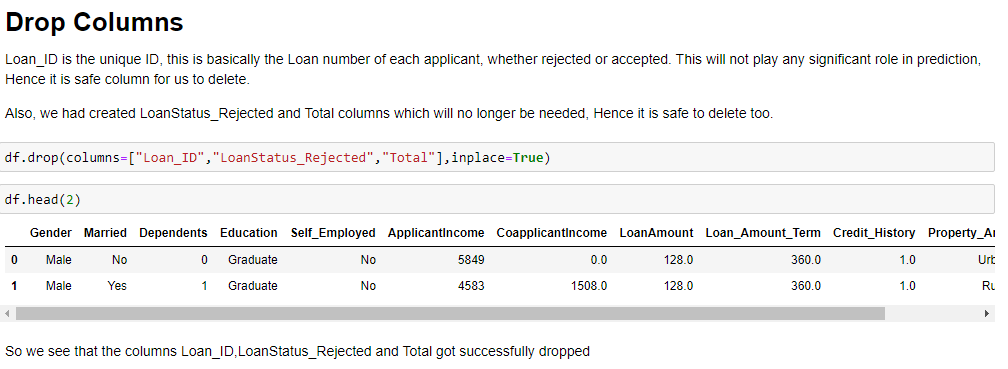










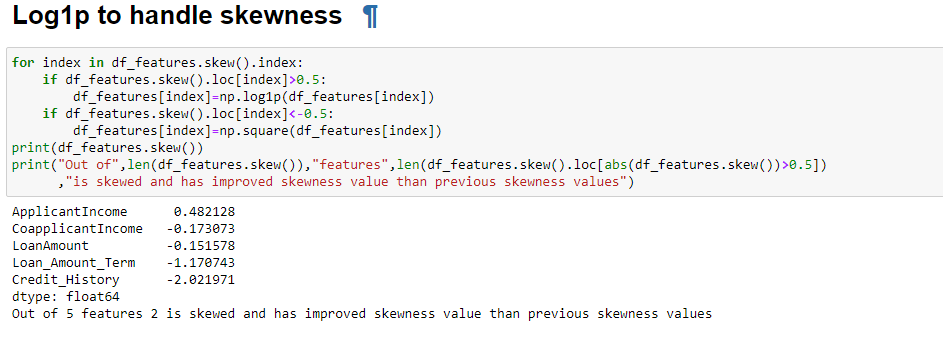


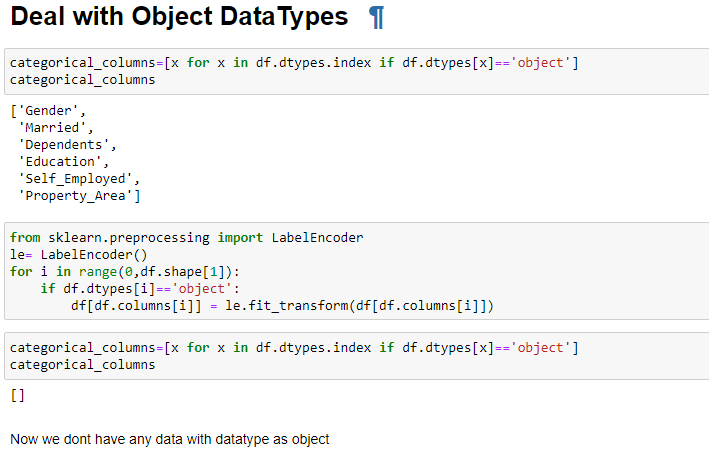
**3.5 Data Exploration and Transformation**

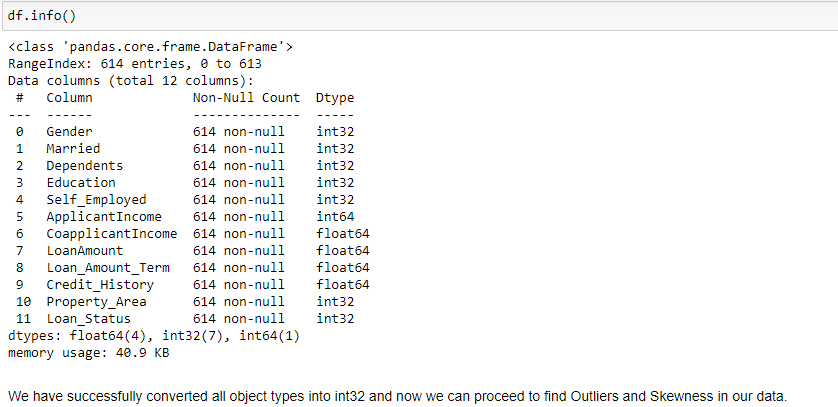


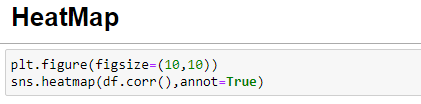
# Log Transform

Log transformation is most likely the first thing you should do to remove skewness from the predictor. It can be easily done via *Numpy*, just by calling the **log()** function on the desired column. We can then just as easily check for skew:

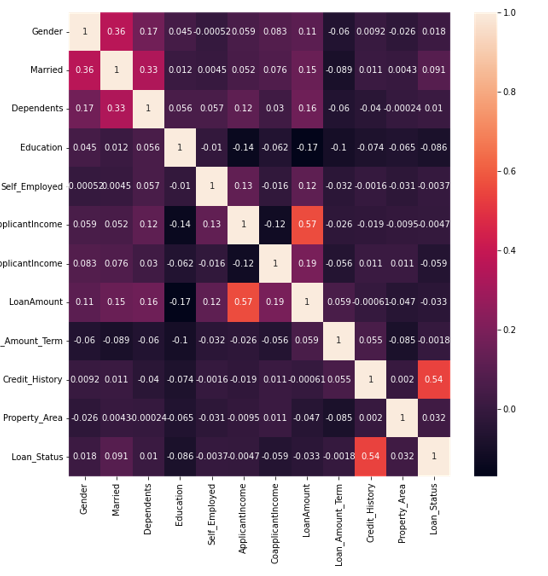








A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information. Heat maps are used in many areas such as defense, marketing and understanding consumer behavior.



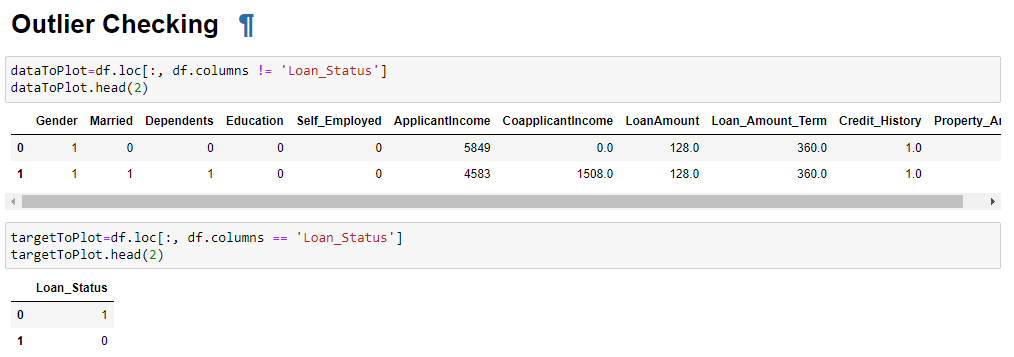
We notice that

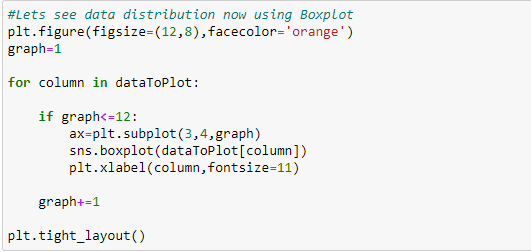
* Applicant Income and Loan Amount are bit correlated.
* Credit History and Loan\_Status are correlated to some extent

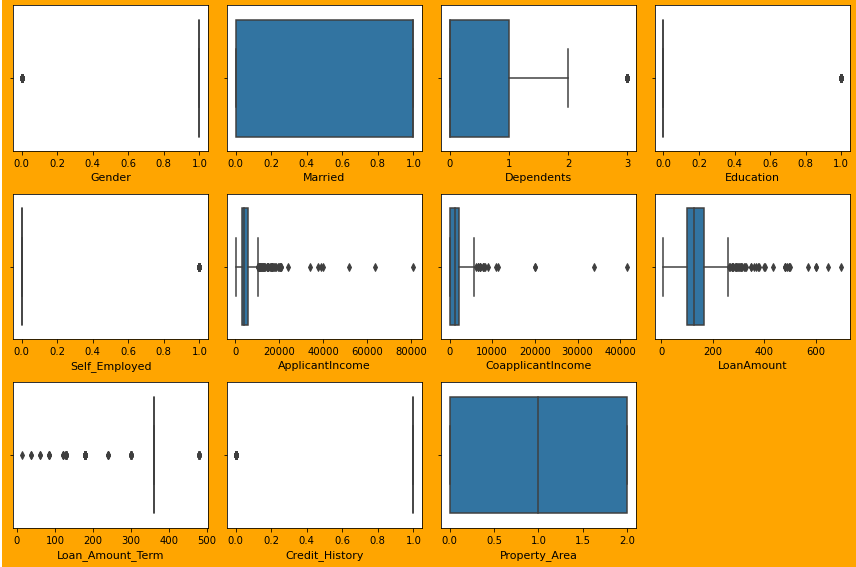
**Outliers**

An **outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

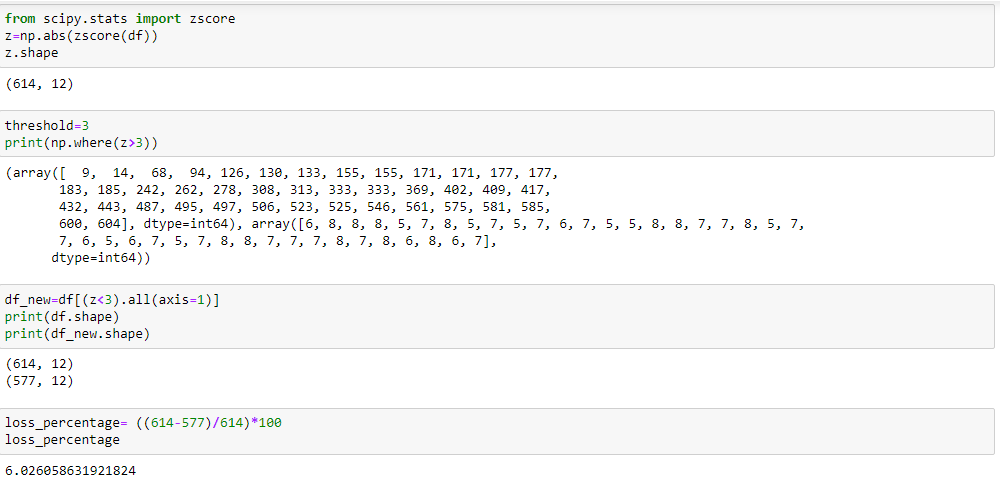
**Why outlier analysis?**  
Most data mining methods discard outliers noise or exceptions, however, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring one and hence, the outlier analysis becomes important in such case.







We Applicant Income, Co-applicant Income, Loan Amount have outliers. We will use zscore method to check if we can remove the outliers or not.

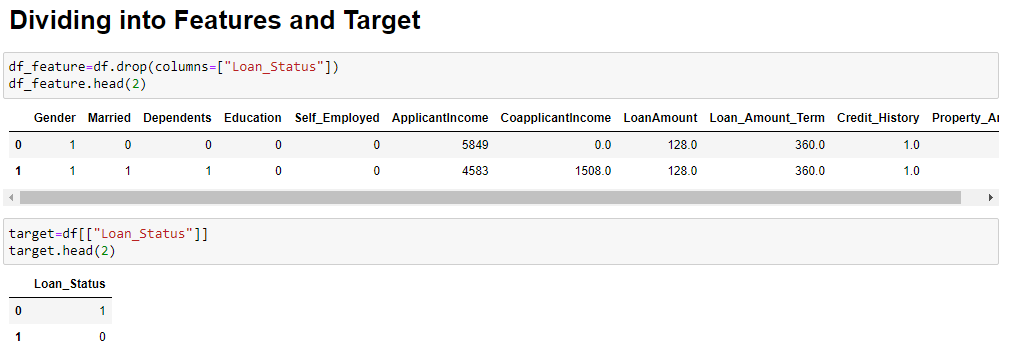


Loss Percentage is 6%, hence we will delete those data.



**3.6 Split Dataset.**

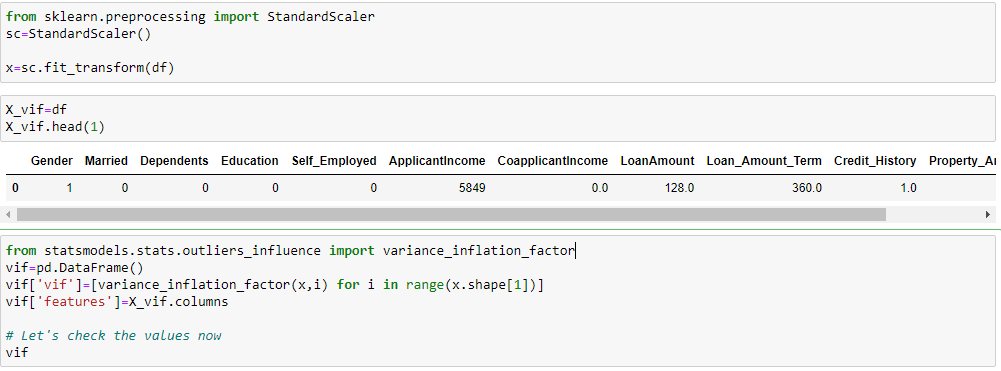
We will perform split dataset into features and target for train and test purpose.

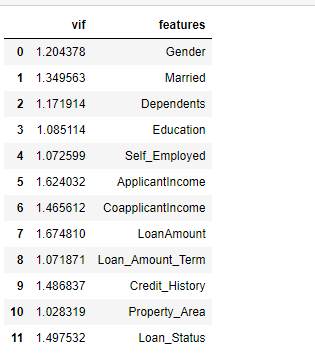


**3.7 Normalization**

**StandardScaler** follows **Standard Normal Distribution (SND)**. Therefore, it makes *mean = 0* and scales the data to unit variance.

We will also check Variance inflation factor (VIF). Variance inflation factor is a measure of the amount of [multicollinearity](https://www.investopedia.com/terms/m/multicollinearity.asp) in a set of multiple [regression](https://www.investopedia.com/terms/r/regression.asp) variables.

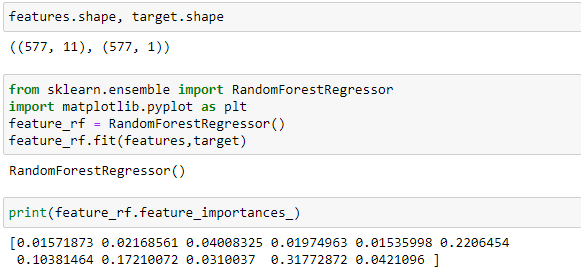


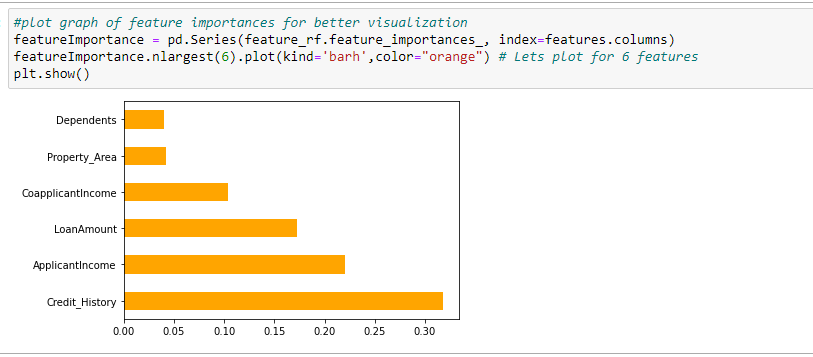


**3.8 Feature Importance**

One of the great quality of random forest is that they make it very easy to measure the relative importance of each feature. Sklearn measure a features importance by looking at how much the tree nodes, that use that feature, reduce impurity on average (across all trees in the forest). It computes this score automatically for each feature after training and scales the results so that the sum of all importance is equal to 1. We will access this below.

Top six Feature importance are represented below graphically.

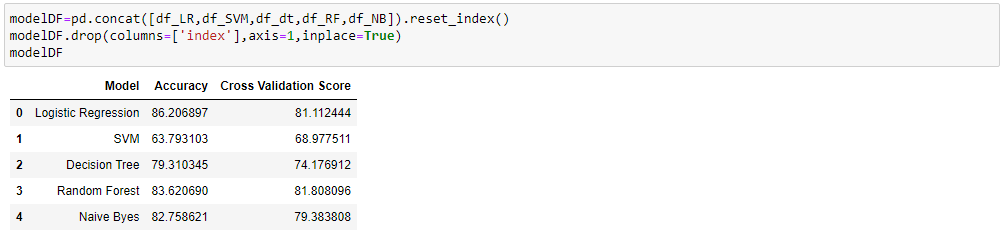




We see that Credit\_History is the most important feature.

**4.Model Building.**

I built 5 model and checked the Accuracy and cross validation score. Below is the output of all the five models.



Also, we will check the roc\_auc score of each graph.

The **Receiver Operator Characteristic (ROC)** curve is an evaluation metric for binary classification problems. It is a probability curve that plots the **TPR**against **FPR**at various threshold values and essentially **separates the ‘signal’ from the ‘noise’**. The **Area Under the Curve (AUC)**is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

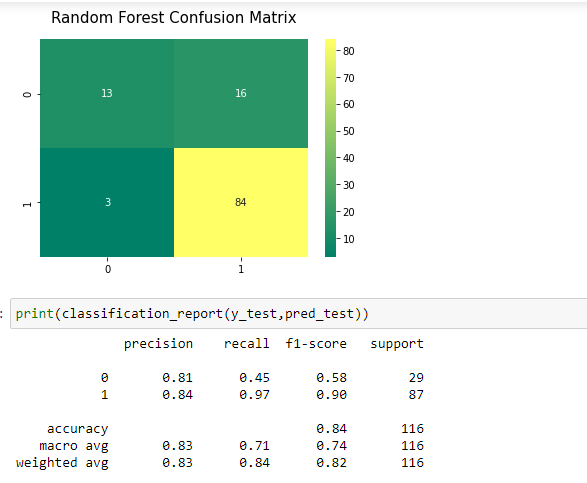
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

Let’s check the confusion matrix and classification report.

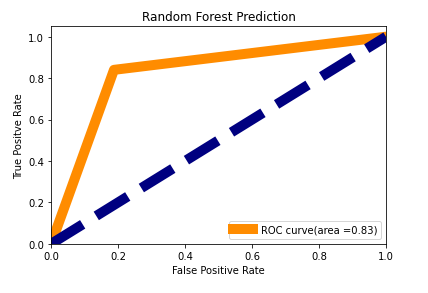
A **confusion matrix** is a technique for summarizing the performance of a classification algorithm.

**Classification report** provides a better understanding of the overall performance of our trained model.

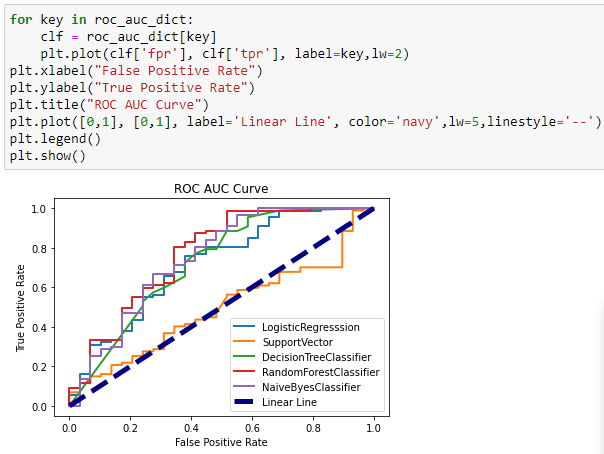
|  |  |
| --- | --- |
| **Metrics** | **Definitions** |
| * Precision | Precision is defined as the ratio of true positives to the sum of true and false positives |
| * Recall | Recall is defined as the ratio of true positives to the sum of true positives and false negatives. |
| * F1-Score | The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is |
| * Support | Support is the number of actual occurrences of the class in the dataset. It doesn’t vary between models, it just diagnoses the performance evaluation process. |



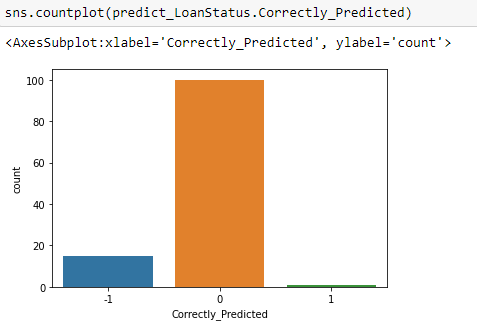
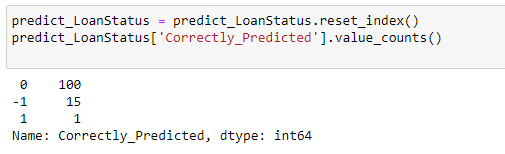
Let’s plot accuracy score graph for Random Forest classifier and observe.



Let’s plot ROC AUC curve of all the models built and anlyse.



**5.Evaluation.**



The above graphs show us the comparison between the True Positive and False Positive rates and we see that Random Forest Classifier is performing best. Also the least difference between Accuracy Score and Cross Validation Score is for Random Forest Classifier, so we will select this model for our prediction. Also F1 Score is 0.84, recall is 0.71 which is pretty good.

85.47% is the accuracy of correctness and we see that model is behaving well.

**6.Save the model.**

We will save the model at last.



**7. Conclusion.**

This is one of the interesting articles that I have written because it was on today’s current top technology machine learning, but I was used basic language to explain this article so that one can’t find it difficult to understand.

* We did Exploratory Data Analysis on the features of this dataset and saw how each feature is distributed.
* We analyzed each variable to check if data is cleaned and normally distributed.
* We cleaned the data and removed NA values
* We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.
* We calculated correlation between independent variables and found that applicant income and loan amount have significant relation.
* We created dummy variables for constructing the model and extra Loan\_Rejected variable for Visualization.
* We constructed models taking different variables into account and found through odds ratio that credit history is creating the most impact on loan giving decision
* We tested the data and got the accuracy of 85.47 %