Loan Application Status Prediction

In this article, I’ll be going through the process of building a machine learning model to predict the loan application status and it will take you through every step in detail and helps you to understand the whole machine learning model building process.



Introduction

Individuals all around the world in some way depend on banks to lend them loans for various reasons to help them overcome their financial constraints and achieve some personal goals. Due to the everchanging economy and ever-increasing competition in the financial world, the activity of taking a loan has become inevitable. Also, small-scale to large-scale banking firms depend on the activity of lending out loans to earn profits for managing their affairs, and function smoothly at times of financial constraints. A loan is the major source of income for the banking sector as well as the biggest source of financial risk for banks. Large portions of a bank’s assets directly come from the interests earned on loans given. Though lending loans is quite beneficial for both parties, the activity does carry great risks. These risks represent the inability of a borrower to pay back the loan by the designated time which was decided mutually by both the lender and the borrower and it is referred to as ‘Credit Risk’. For that, it is highly necessary to assess the clients’ credit suitability before authorizing a loan.

While in the past, banks used to hire highly professional individuals whose sole purpose was to evaluate applicants and after close review decide and tell whether a candidate was eligible for receiving a loan. The worthiness of a candidate for loan approval or rejection was based on a numerical score called ‘Credit Score’. Generally, the credit score helps the authorities to compute the probability of borrowers repaying the loan by the designated time based on their credit history or payment history along with their background.

The process of credit scoring required experts alongside statistical algorithms to accurately predict the creditworthiness of an applicant. However, quite recently, the researchers and the banking authorities have opted for training classifiers based on various machine learning and deep learning algorithms to automatically predict the credit score of an applicant based on their credit history and other historical data and make the process of selecting the eligible candidates a lot easier before the loan is approved.

Dataset Description

The Data set provided by Data Trained Education for Education Training purpose

Problem Statement

 This dataset includes details of applicants who have applied for loan. The dataset includes details like credit history, loan amount, their income, dependents etc.

Independent Variables:

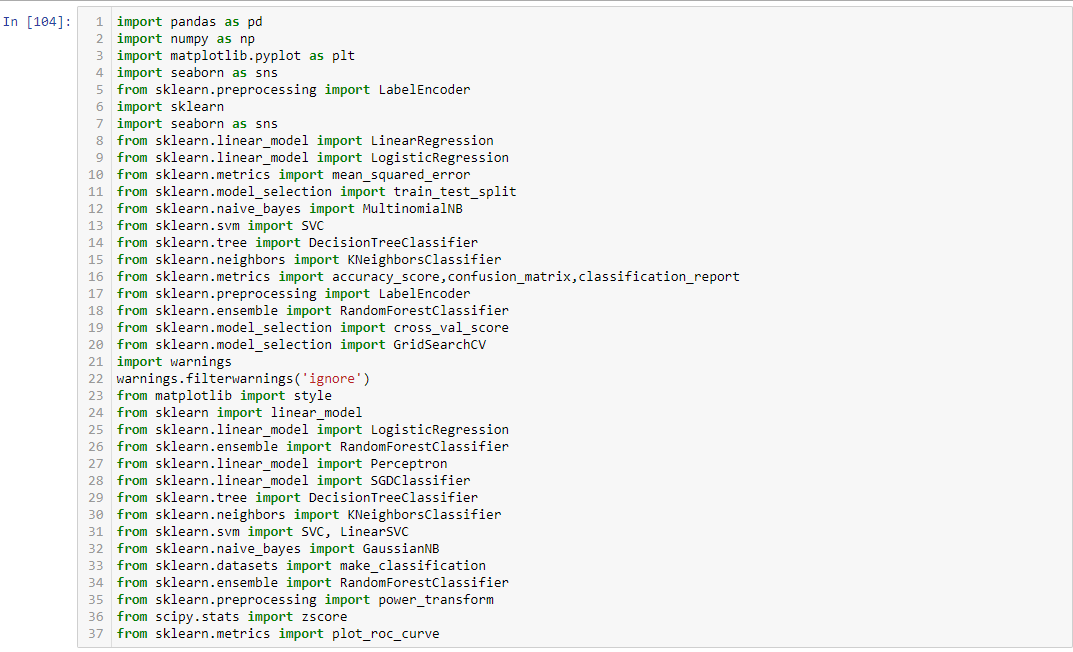
* Loan\_ID
* Gender
* Married
* Dependents
* Education
* Self\_Employed
* ApplicantIncome
* CoapplicantIncome
* Loan\_Amount
* Loan\_Amount\_Term
* Credit History
* Property\_Area

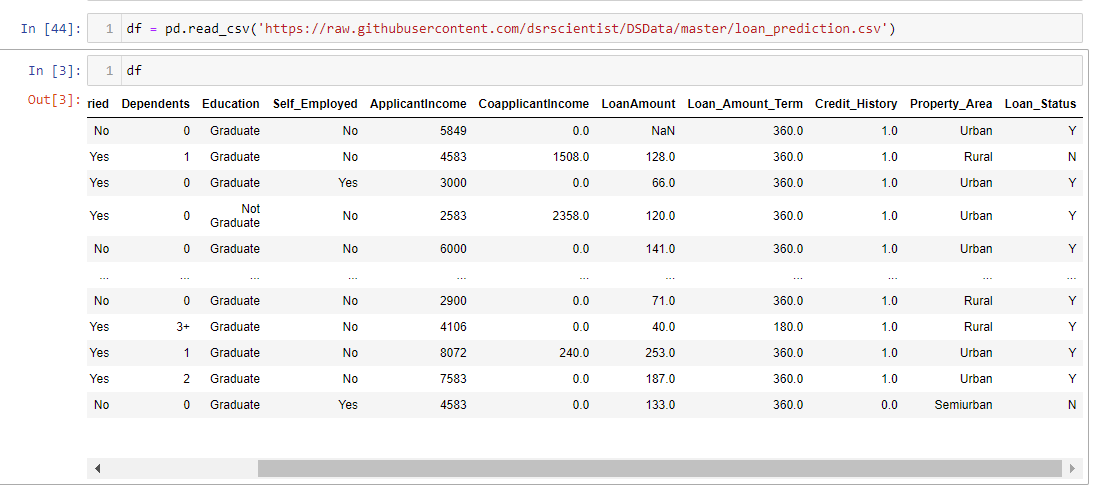
Dependent Variable (Target Variable):

* Loan\_Status

Data Analysis

The process of cleaning, transforming, and extracting data to discover useful information for business decision-making is called data analysis.

**Importing necessary libraries**

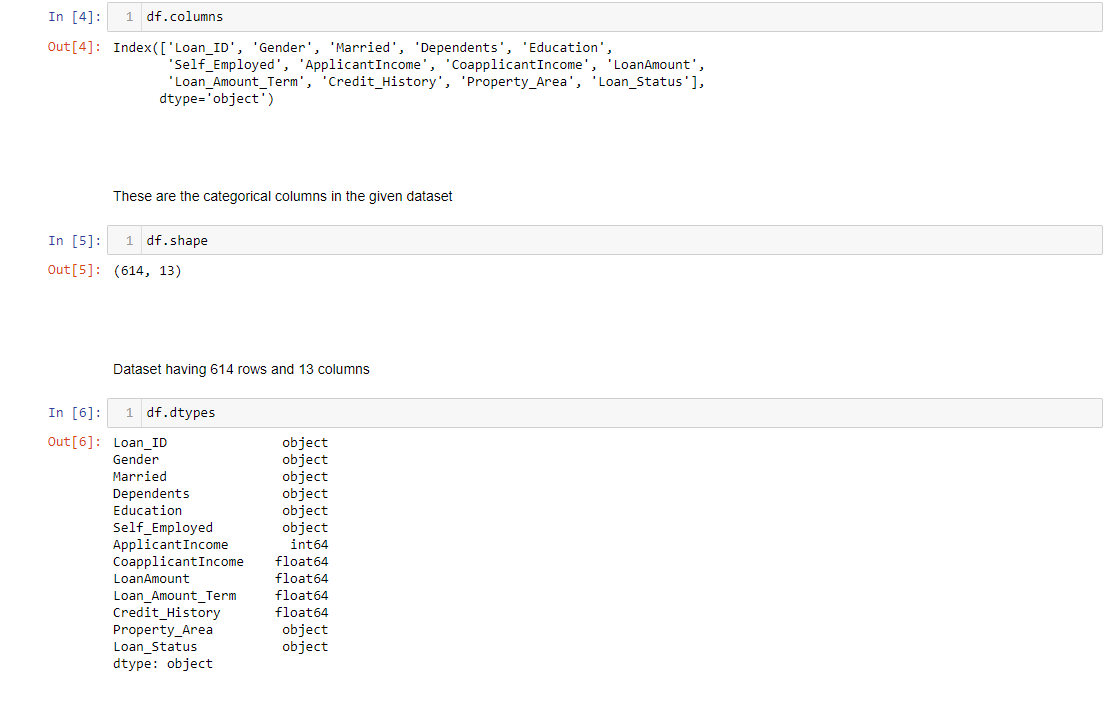
Importing the Dataset

In this dataset **"Loan Status"** is our target variable which has two classes. So this is a **"Classification type"** problem.

Exploratory Data Analysis (EDA)

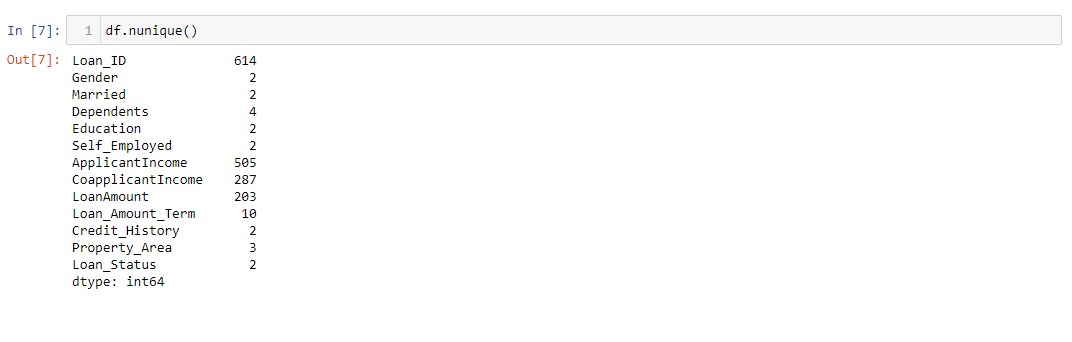
Exploratory Data Analysis (EDA) is an approach to analyzing the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations

Now we are doing EDA with our data set

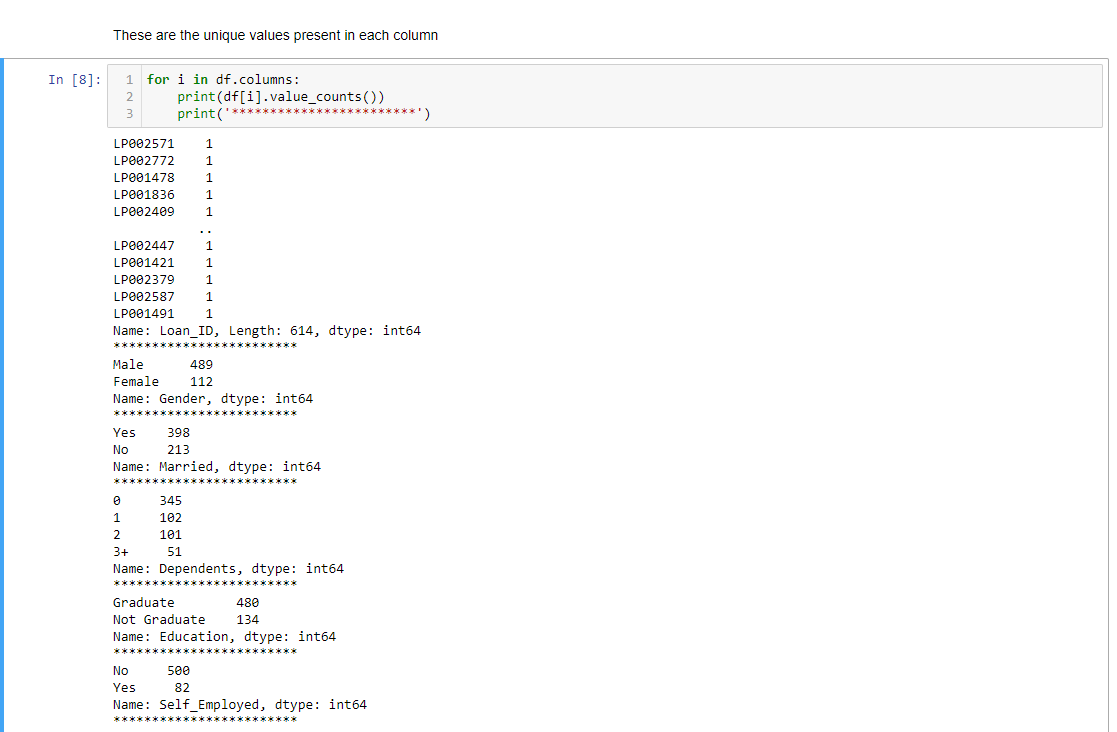


* I run df. columns to get the information about columns in our data set. In our dataset has these coloums Loan ID, Gender, Married, Dependents, Education, Self Employed, Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area, Loan Status
* Then I run the df.shape which gives the number of rows and columns present in the dataset and our dataset has 614 rows and 13 columns.
* After df.shape i run df.dtypes which gives the types of dataset present in the dataset and in our dataset we have integer, float, and objective data types

Checking the unique values in the dataset- df.nunique()

The unique() function is used to find the unique elements of an array. Returns the sorted unique elements of an array.

The number of unique values present in each column.

Checking the value count in the dataset -

Pandas**“df.value\_counts()”** function returns object containing counts of unique values. The resulting object will be in descending order so that the first element is the most frequently-occurring element. Excludes NA values by default.

After Value Counts I dropped the column Loan\_ID is the unique ID given to the applicants also it has no significance in the prediction so l dropped this column

Drop Function :-

The drop() method removes the specified row or column. By specifying the column axis ( axis='columns' ), the drop() method removes the specified column. By specifying the row axis ( axis='index' ), the drop() method removes the specified row.

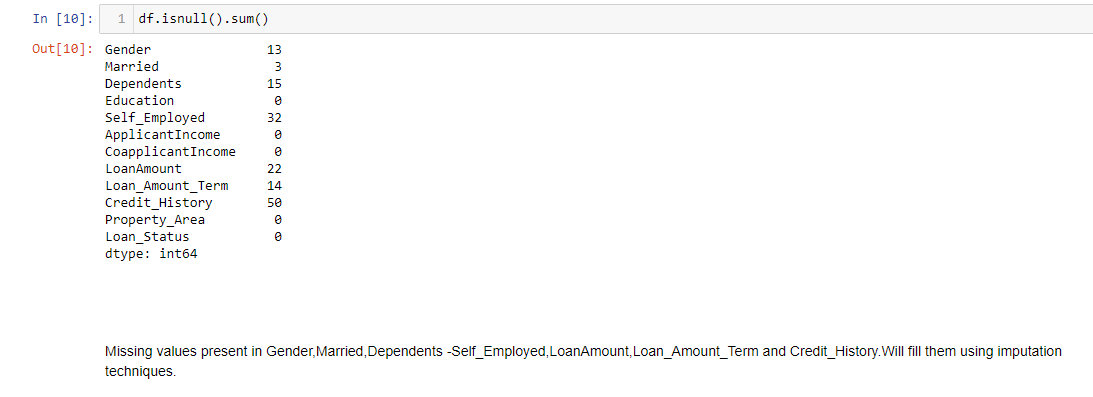
Let’s check if there any null values- df.isnull.sum()

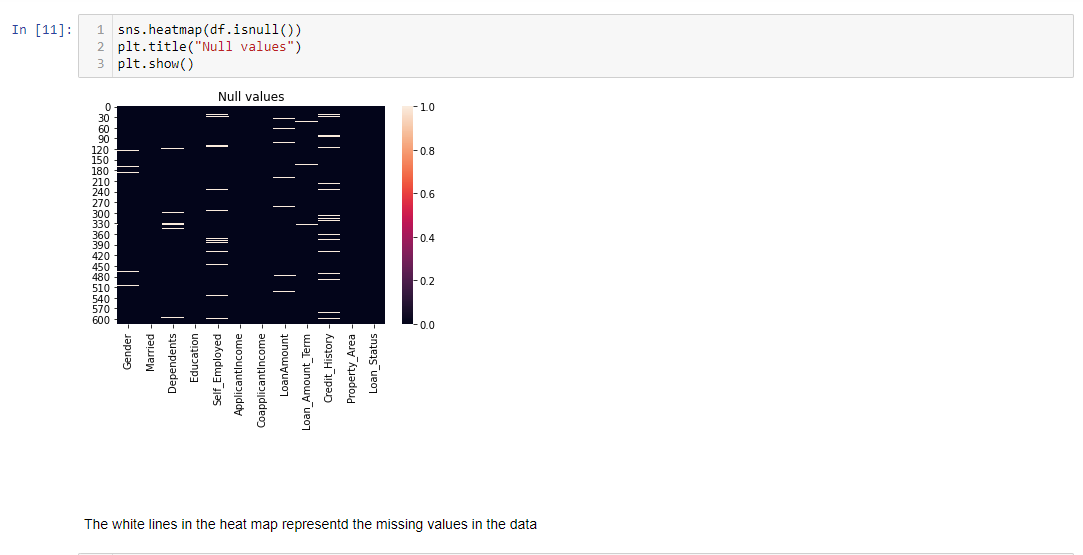
A null value in a relational database is used when the value in a column is unknown or missing. A null is neither an empty string (for character or datetime data types) nor a zero value (for numeric data types)

### Principles of NULL values:

* Setting a NULL value is appropriate when the actual value is unknown, or when a value would not be meaningful.
* A NULL value is not equivalent to a value of ZERO if the data type is a number and is not equivalent to spaces if the data type is character.
* A NULL value can be inserted into columns of any data type.
* A NULL value will evaluate NULL in any expression.
* Suppose if any column has a NULL value, then UNIQUE, FOREIGN key, CHECK constraints will ignore by SQL.

There are null values present in the dataset and we can observe them in the heat map visualization .





Removing Null Values

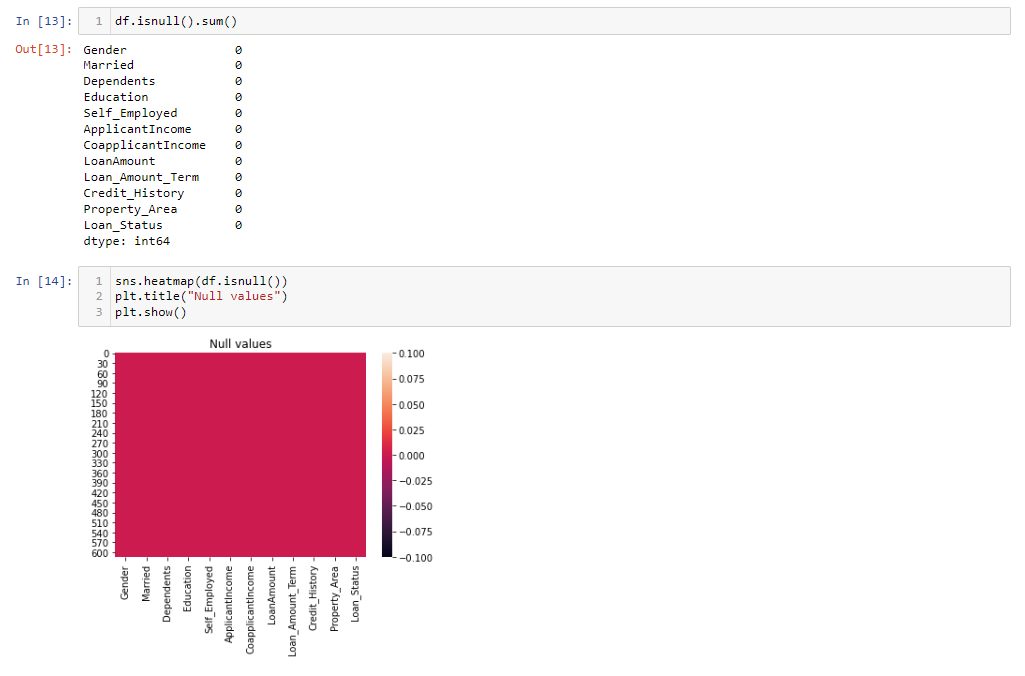
The values fill by fillna method

Fillna :-

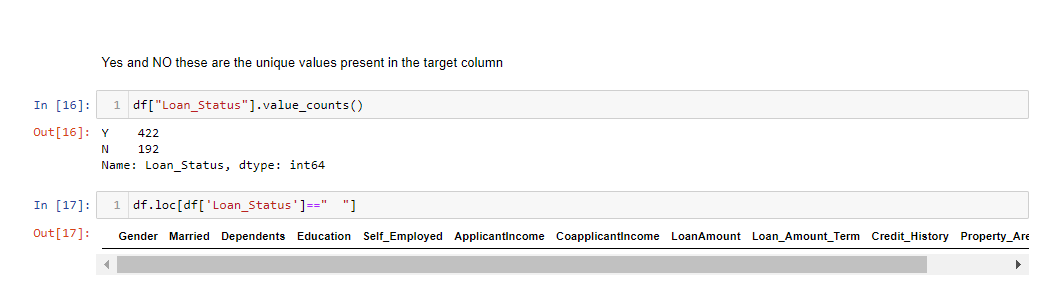
* The fillna() method replaces the NULL values with a specified value.
* The fillna() method returns a new Data frame object unless the inplace parameter is set to True, in that case the fillna() method does the replacing in the original Data frame instead.

Mode :-

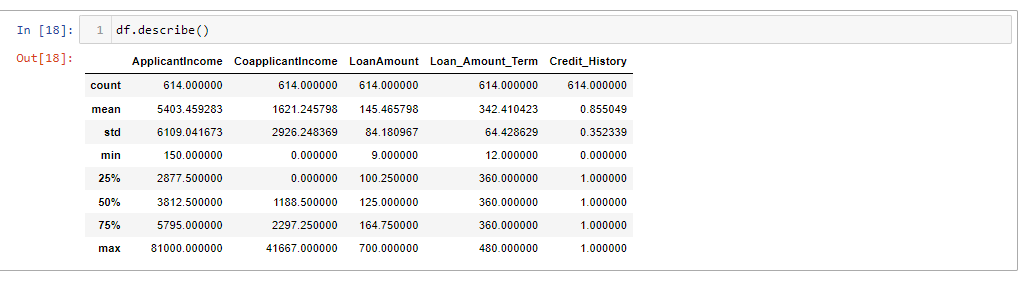
**mode imputation**in which the missing values are replaced with the **mode value** or **most frequent**value of the entire feature column. When the data is skewed, it is good to consider using mode values for replacing the missing values.

Checking Again For Null Values

Now there are no null values in our data set

Let’s check the uniqueness and value counts for target variable column

We can see here, that we have a data imbalance issue in the target variable column. I will balance this later with the appropriate method and there are no spaces in the dataset.

Describing the dataset – df.describe()

The describe() method returns description of the data in the DataFrame.If the DataFrame contains numerical data, the description contains these information for each column:count - The number of not-empty values.  
mean - The average (mean) value.  
std - The standard deviation.  
min - the minimum value.  
25% - The 25% percentile\*.  
50% - The 50% percentile\*.  
75% - The 75% percentile\*.  
max - the maximum value.

Observation :-

a). The counts of all the columns are same which means there are no null values present in the dataset.

b). The mean value is greater than the median(50%) in Applicant Income, Co applicant Income, Loan Amount which means they are skewed to right.

c). There is a huge difference in max and 75% percentile which means there are outliers present in the dataset.

Data Visualization

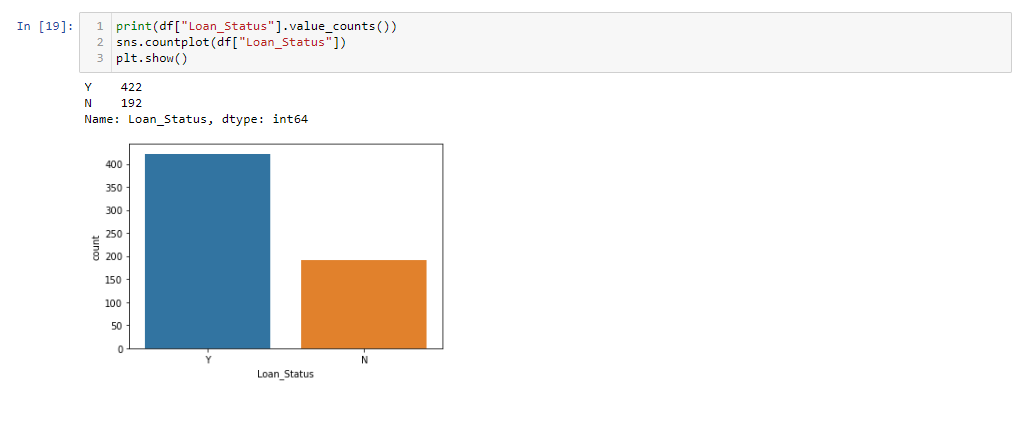
Univariate Analysis

Plotting categorical columns

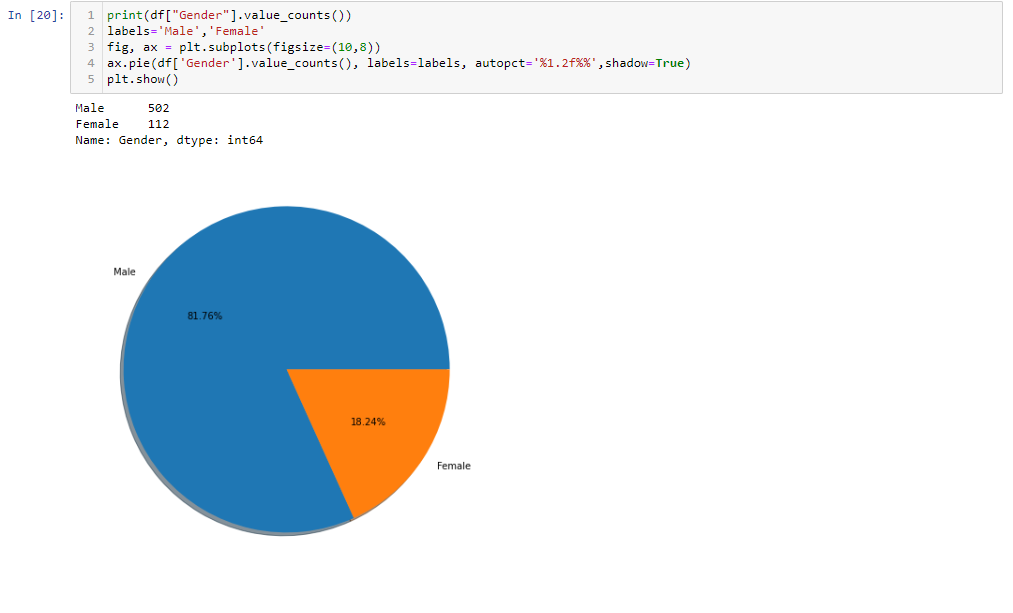
Python offers several plotting libraries, namely Matplotlib, Seaborn and many other such data visualization packages with different features for creating informative, customized, and appealing plots to present data in the most simple and effective way.

Matplotlib and Seaborn

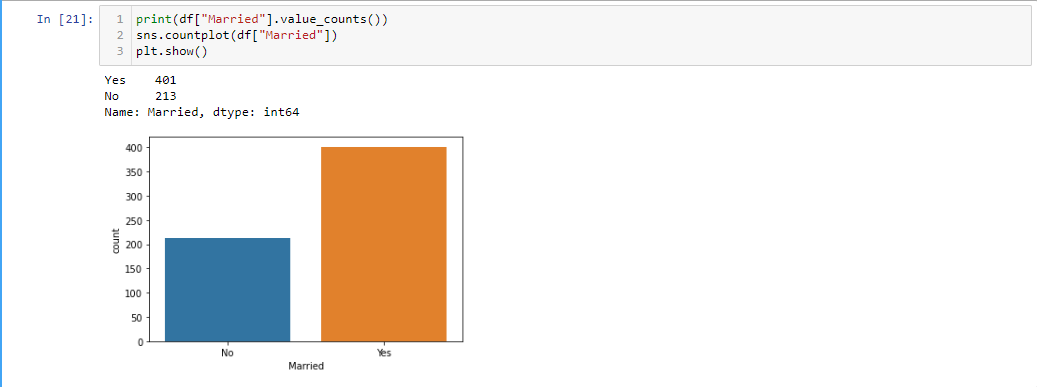
Matplotlib and Seaborn are python libraries that are used for data visualization. They have inbuilt modules for plotting different graphs. While Matplotlib is used to embed graphs into applications, Seaborn is primarily used for statistical graphs.

******

From the above graph we can conclude that more number of loan has been approved that is Y=422 and N=192 has got denied



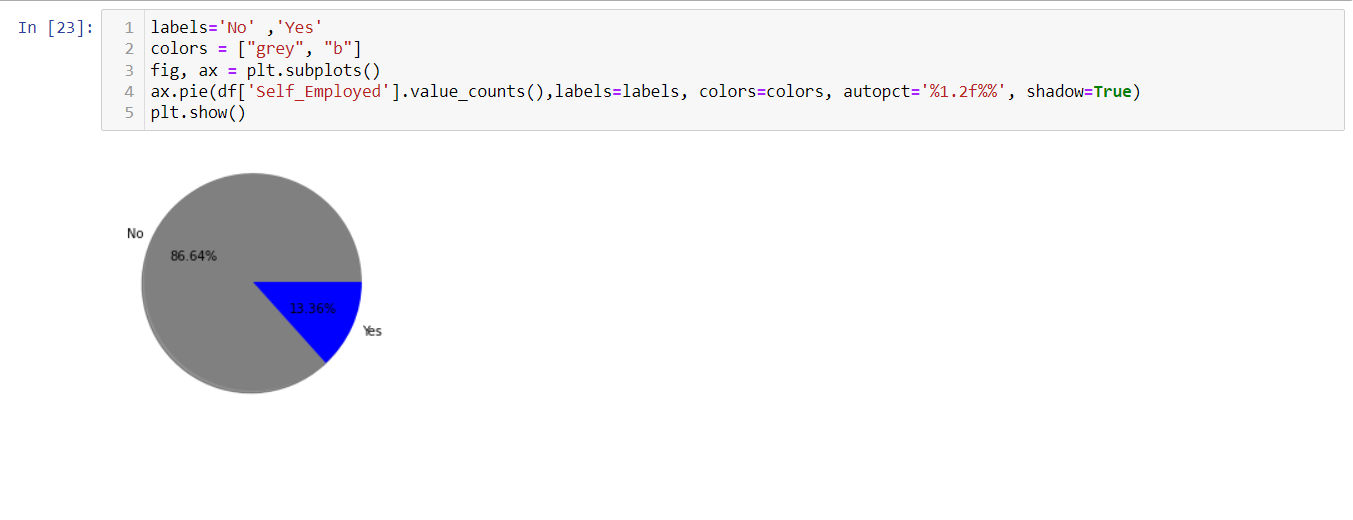
There are more number of Male applicants applying for loan than Female applicants. There are about 81% of the Male candidates and only 18% of Female candidates are applying for the loan



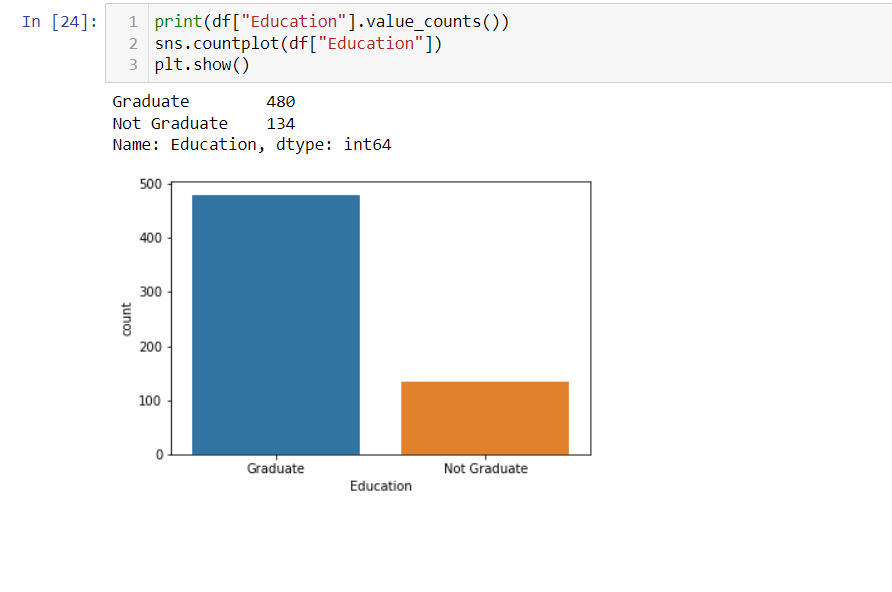
The number of Married applicants who are applied for loan is higher than the Unmarried applicants

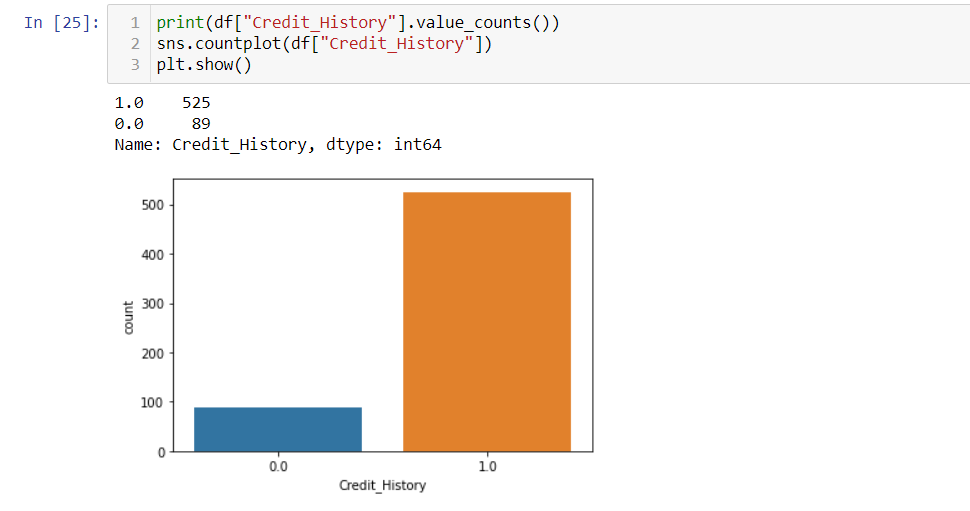


The applicants who have 0 dependents have high counts and the applicants having more than 3 dependents counts are very less.

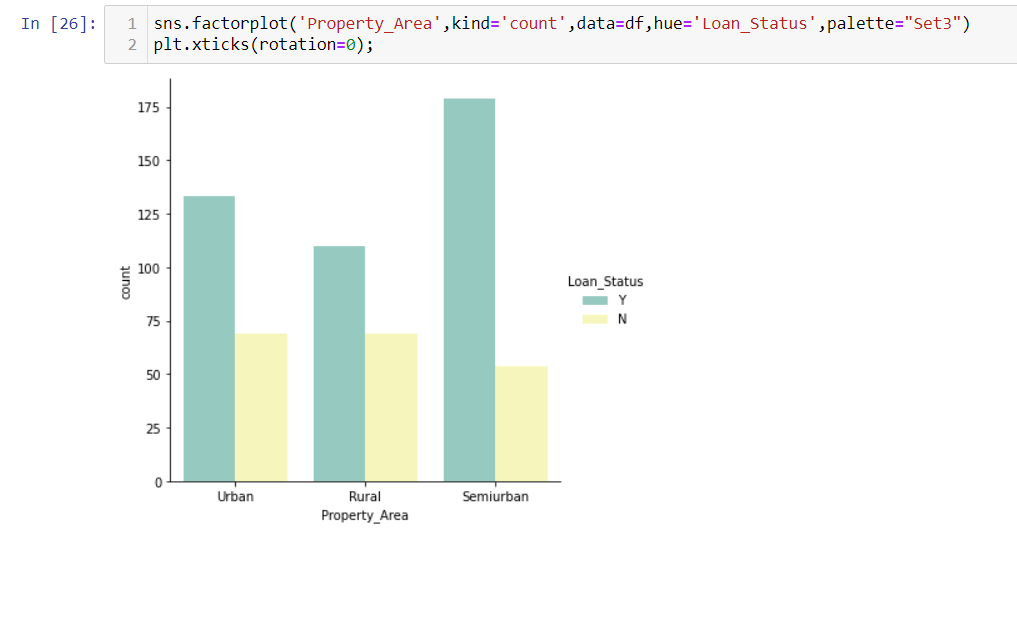


Most of the loan applicants are not self employed

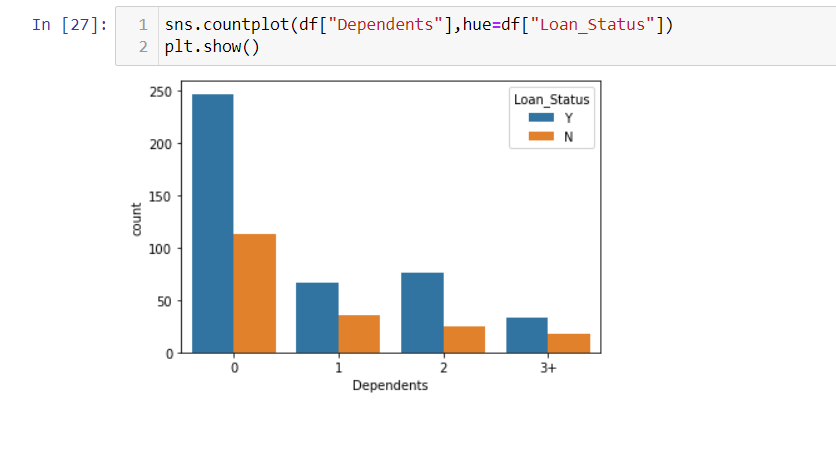
More number of people applied for loan are Graduates and few applicants are Not Graduates



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



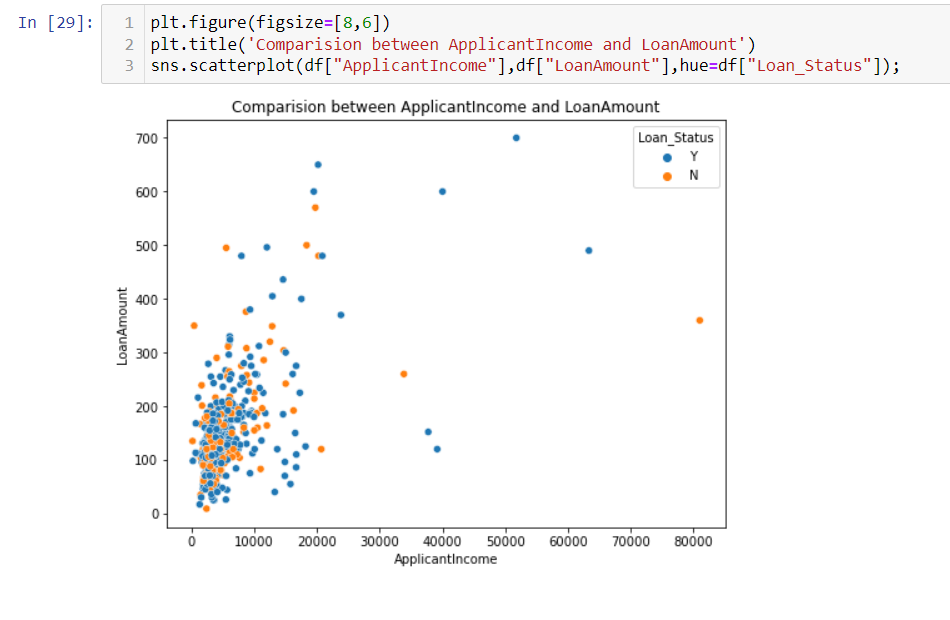
Applicants having property in semiurban area has more chance of getting loan approved

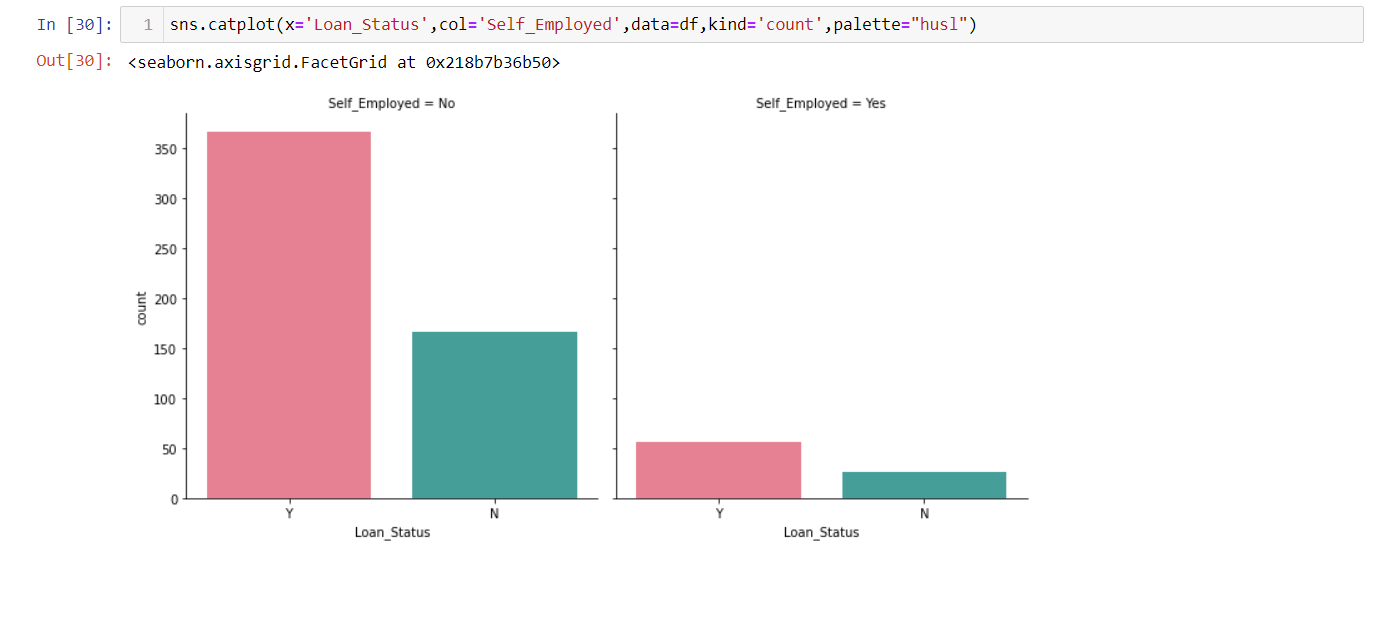
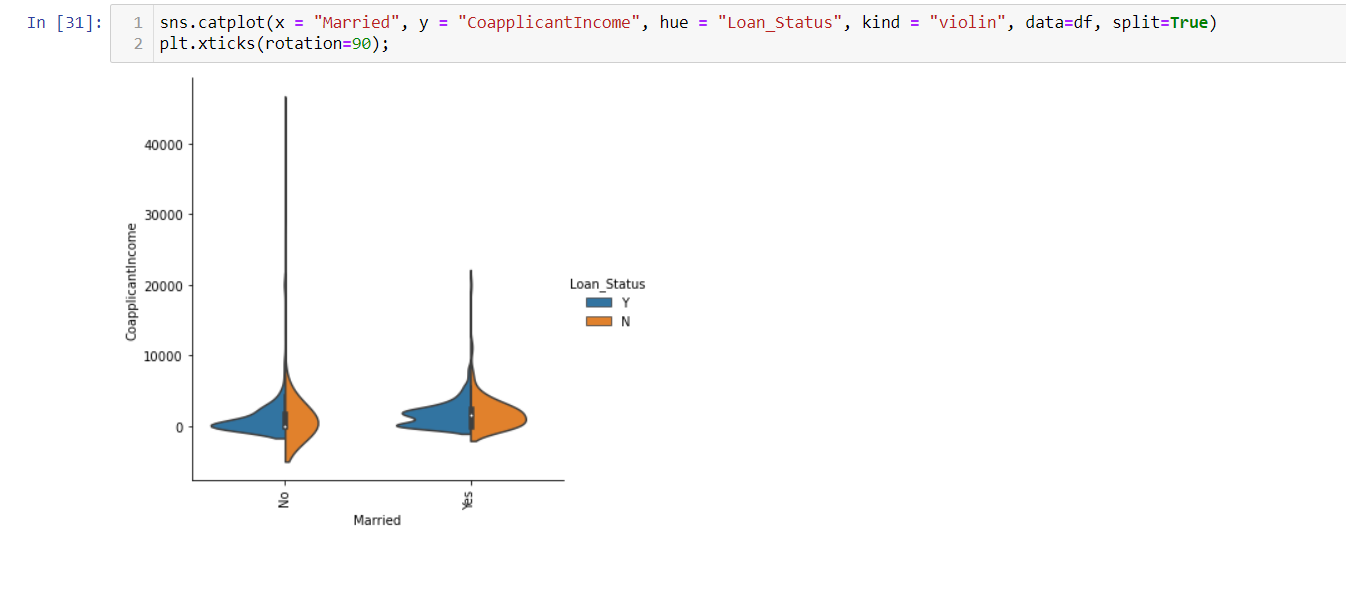


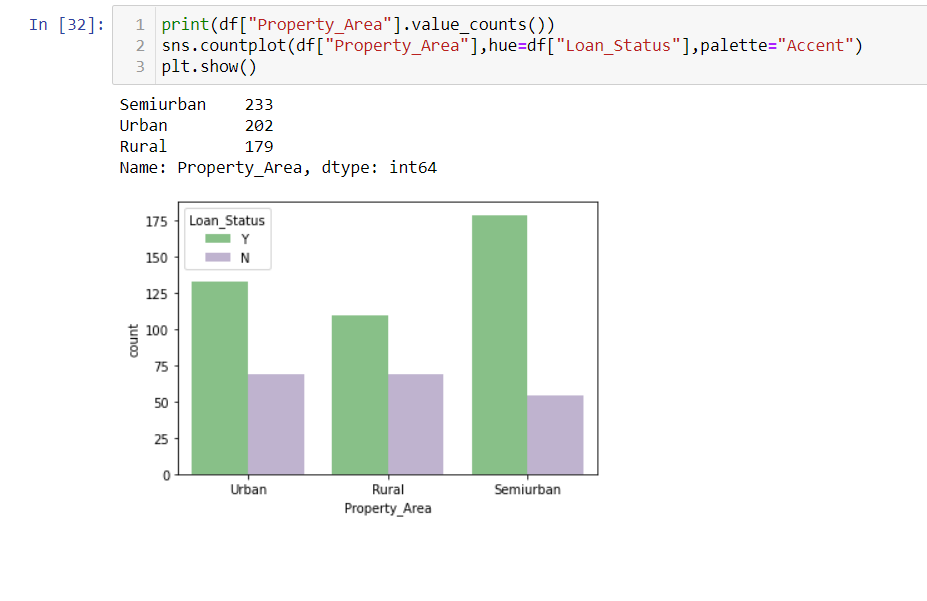
The count of 0 dependents is high which means most of the applicants have no dependents.The applicants who have dependents 0 are more likely to get their loan approved

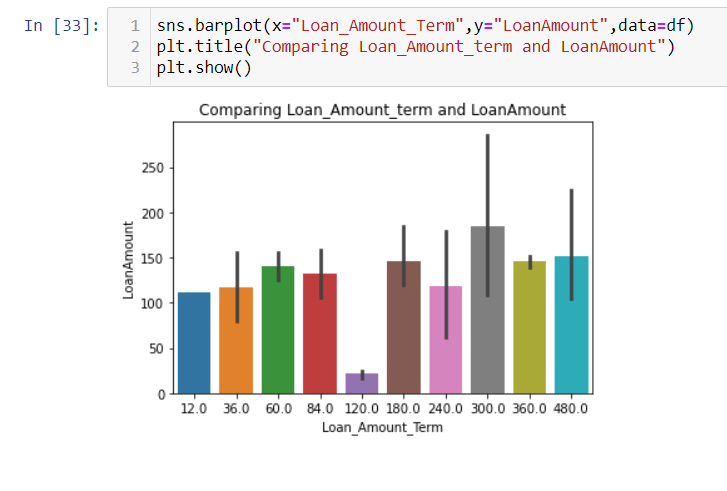


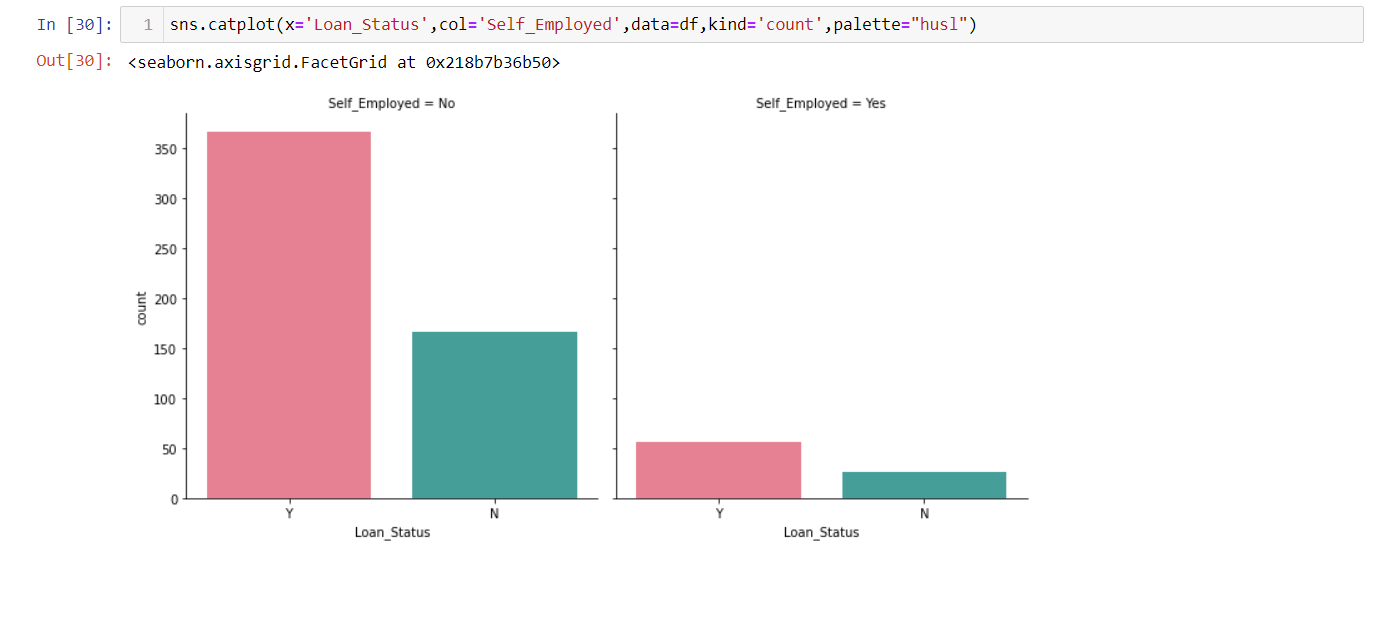
Most of the applicants who are applying for loan are graduated and only few are not graduated. Also the applicants who are graduated have tendency of getting loans than who are not



There is a high density of points in the range of 0-2000 for ApplicantIncome, and 0-300 for loan amount which means if Applicants income is in the range of 0-2000 then the loan amount will be approved in the range 0-300

Married people has more chance of getting loan approved

Most of the applicants from the Semiurban are applying for loan followed by Urban area. Also they have more chance of getting their loan approval

The loan amount term 300.0 is high with loan amount compared to others

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

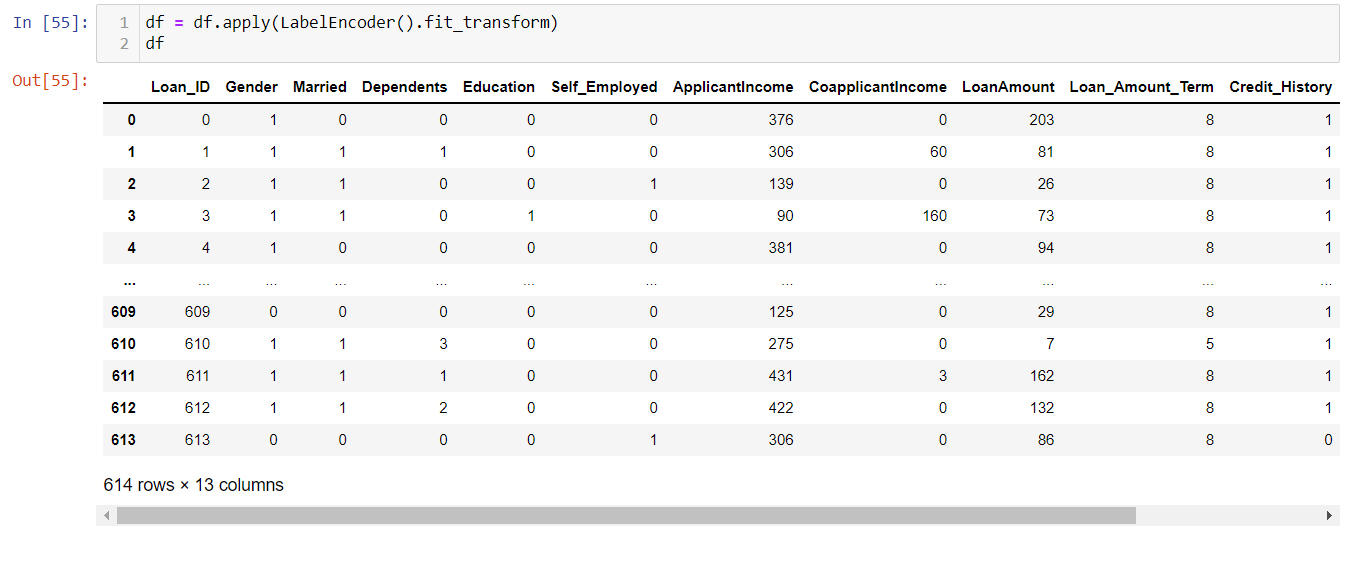


Above is the pair plot for having Laon\_Status as target.There are some extreme outliers present in the dataset

# Label encoding

Label Encoding in Python can be achieved using Sklearn Library. Sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values. LabelEncoder encode labels with a value between 0 and n\_classes-1 where n is the number of distinct labels.

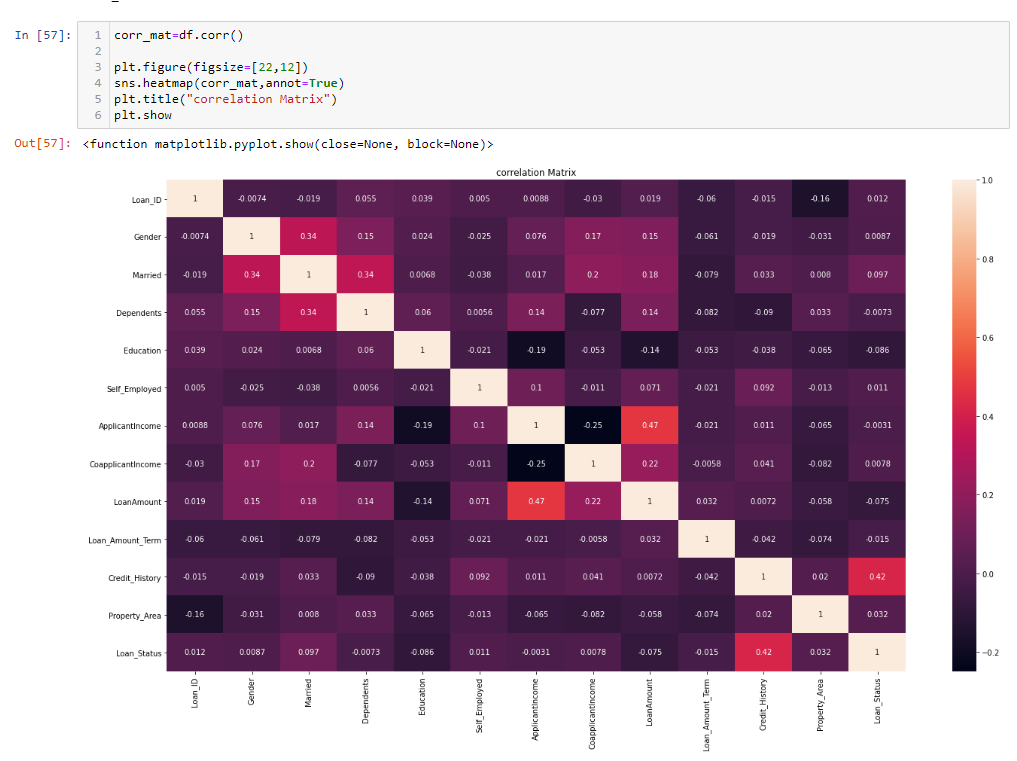
Use to convert Labels into numerical form so machine learning algorithms can decide in a better way how those labels must be operated.



Here I have converted categorical column into numerical form using Label encoder

# Correlation between the target variable and independent variables using HEAT map

 Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

****

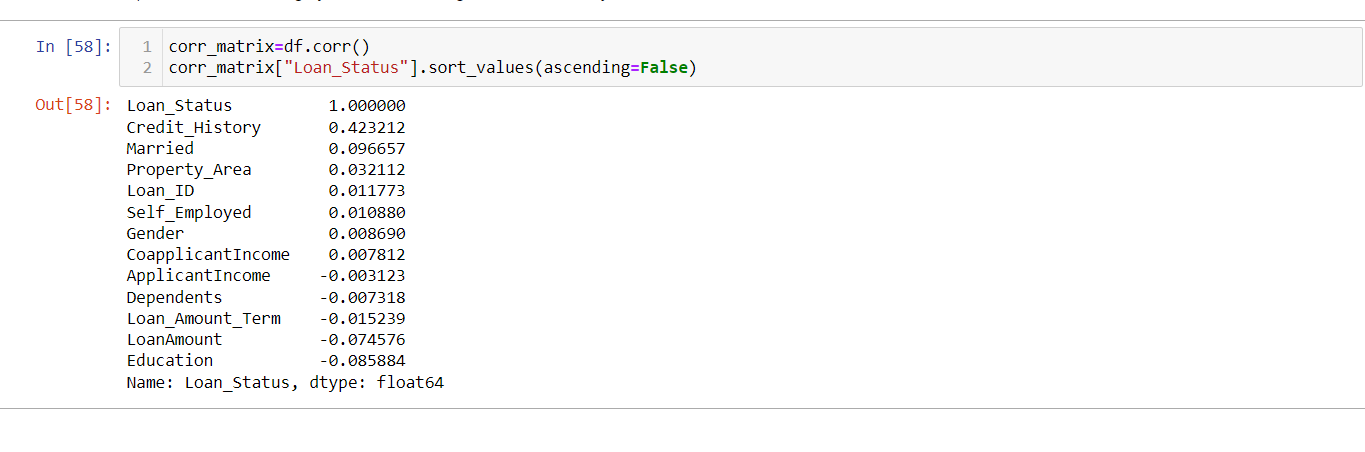
a.)The target column Loan\_Status is highly positively correlated with the feature Credit\_History.

b).The other features have very less correlation with the target column.

c).Also we can notice there is no multicollinearity issue in the features. Features have moderate level of correlation with each other.

d).ApplicantIncome and Gender is very less correlated with the target.

e).Dark shades are highly correlated and light shades are very less correlated.

A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. Key decisions to be made when creating a correlation matrix include choice of correlation statistic, coding of the variables, treatment of missing data, and presentation.

Finding Skewness

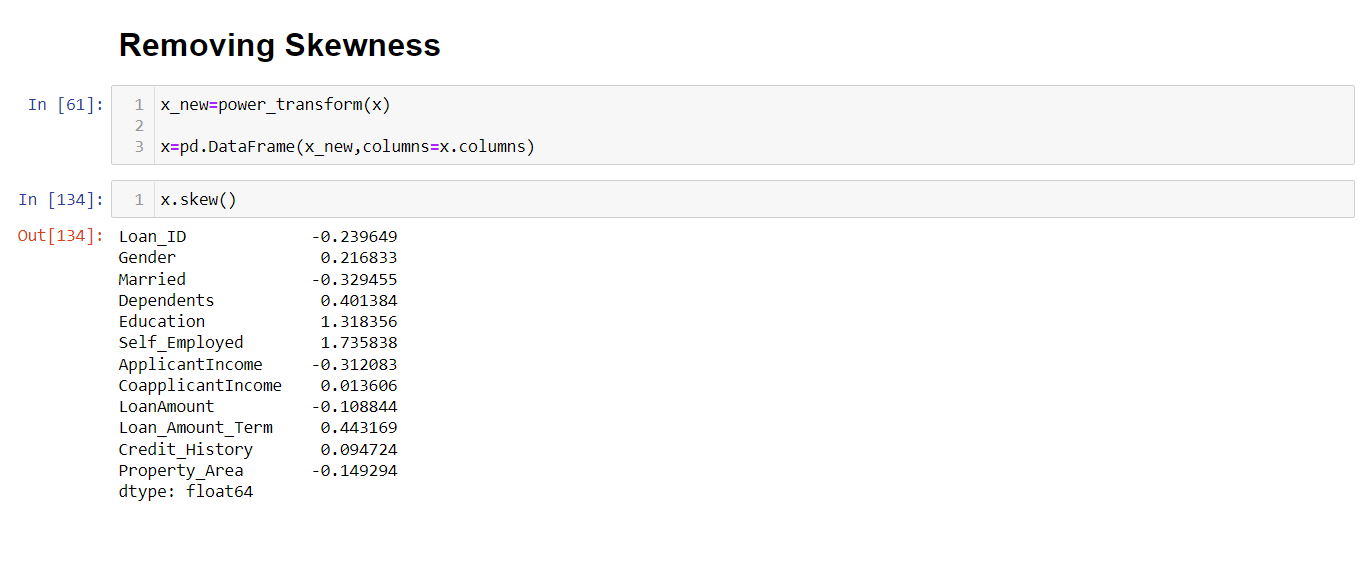
Skewness is a measure of asymmetry of a distribution. Another measure that describes the shape of a distribution is kurtosis. In a normal distribution, the mean divides the curve

Now we are find skew in our data set

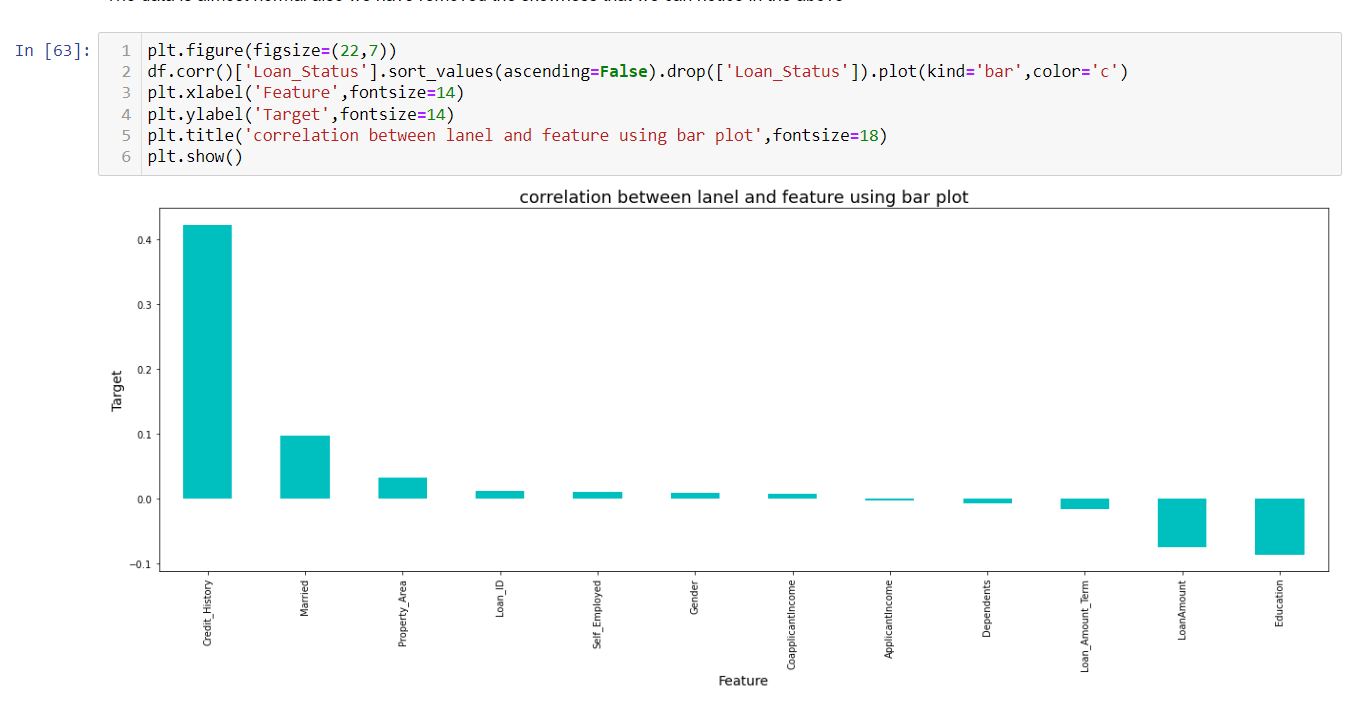
The skewness present in gender, married, dependents, education, self employed, co applicant income, loan amount term,

Removing Skewness

A skewness value greater than 1 or less than -1 indicates a highly skewed distribution. A value between 0.5 and 1 or -0.5 and -1 is moderately skewed. A value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.

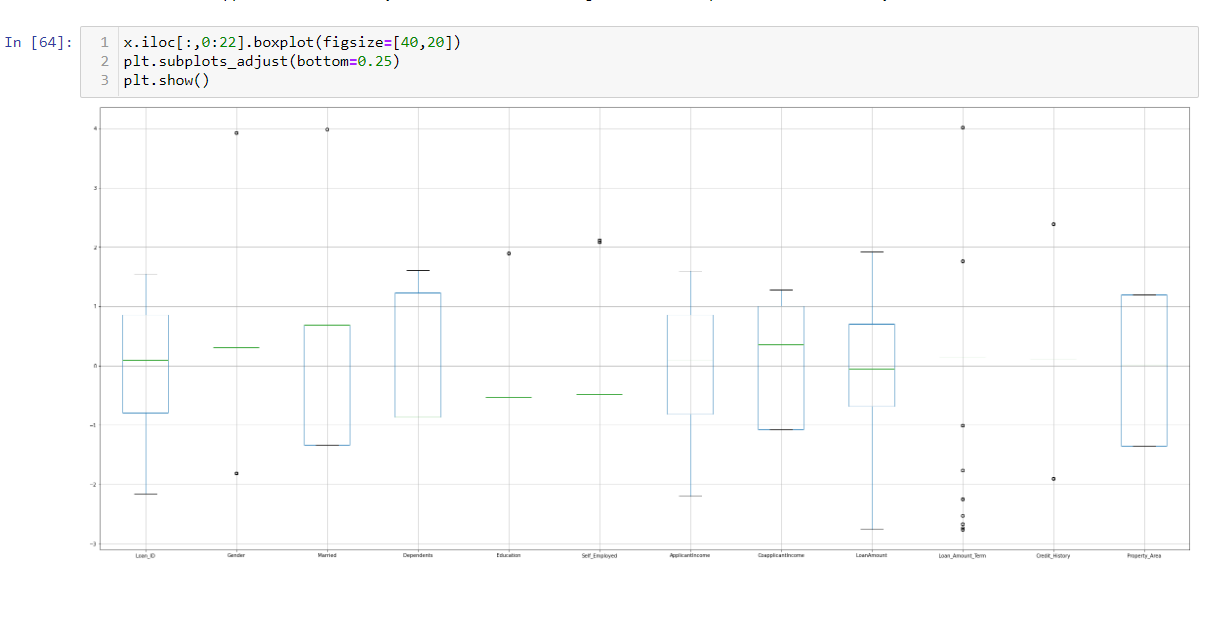
Now we are removing skewness from our data set

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



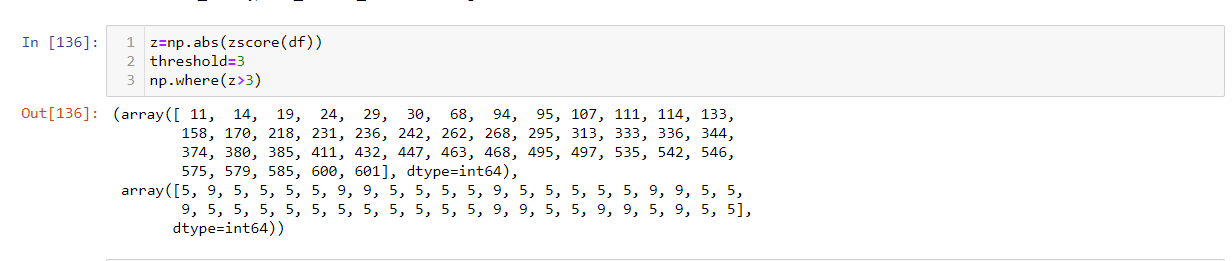
Outlier

 An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

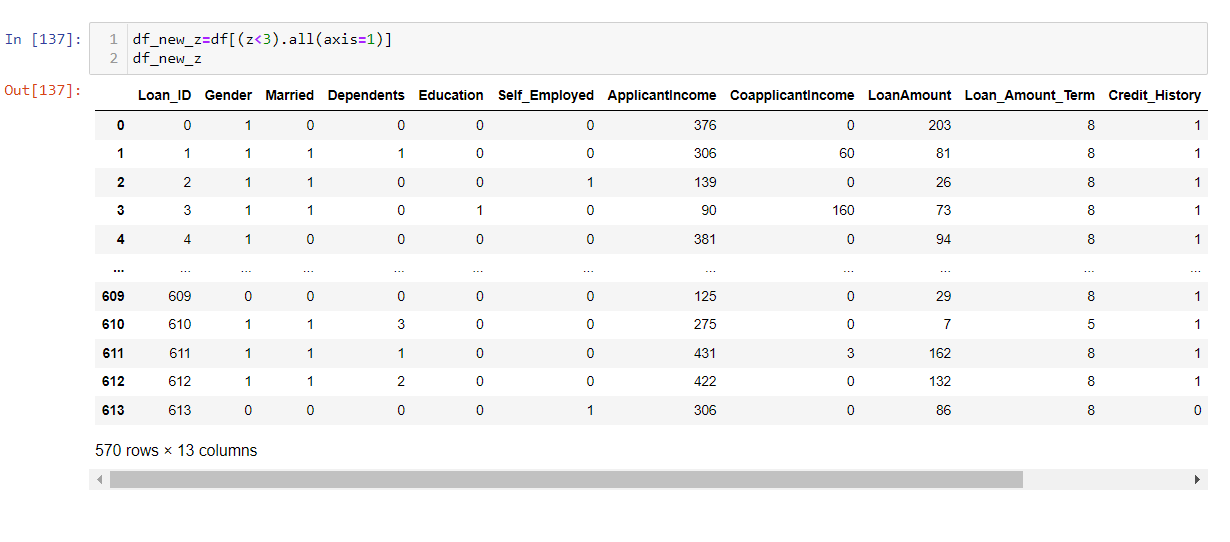
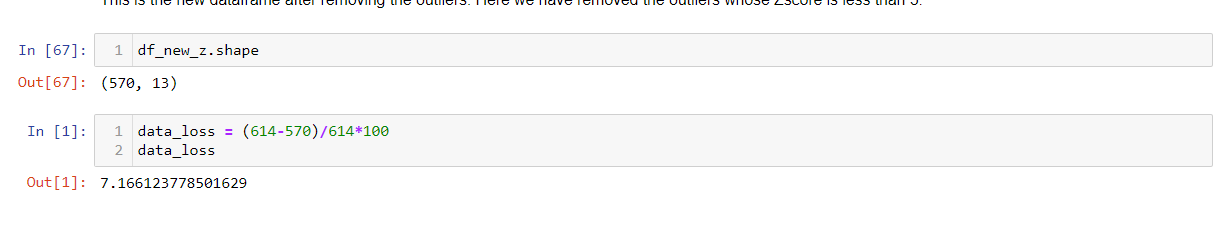
An outlier is an observation that is numerically distant from the rest of the data. When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.

Method of removing outlier

A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of [standard deviations](https://www.investopedia.com/terms/s/standarddeviation.asp) from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

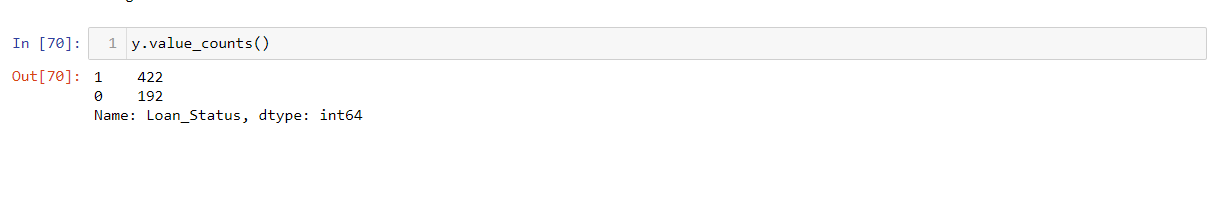
Now put the zscore in our data set for removing outlier

This is the new dataframe after removing the outliers. Here we have removed the outliers whose Zscore is less than 3.



Now we have a data loss report after removing outliers on how much data we have a loss

Using Zscore I have 7.16612% data loss. Which is less than 10%

Now we are checking whether our data is balanced or not if the data is not balanced ten we are balanced our data.

Imbalanced Data

A classification data set with skewed class proportions is called[**imbalanced**](https://developers.google.com/machine-learning/glossary#class_imbalanced_data_set)**.** Classes thatmake up a large proportion of the data set are called[**majority classes**](https://developers.google.com/machine-learning/glossary#majority_class)**.** Those that make upa smaller proportion are[**minority classes**](https://developers.google.com/machine-learning/glossary#minority_class)**.**

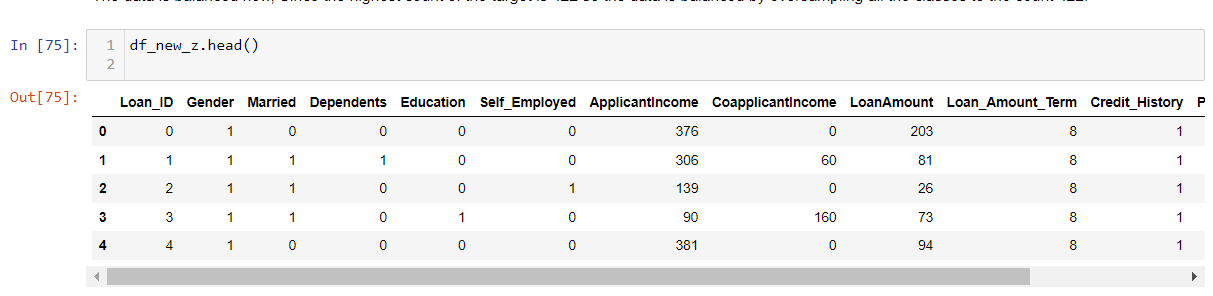
Now we are using SMOTE technique for balanced the data

SMOTE

SMOTE (**synthetic minority oversampling technique**) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances.

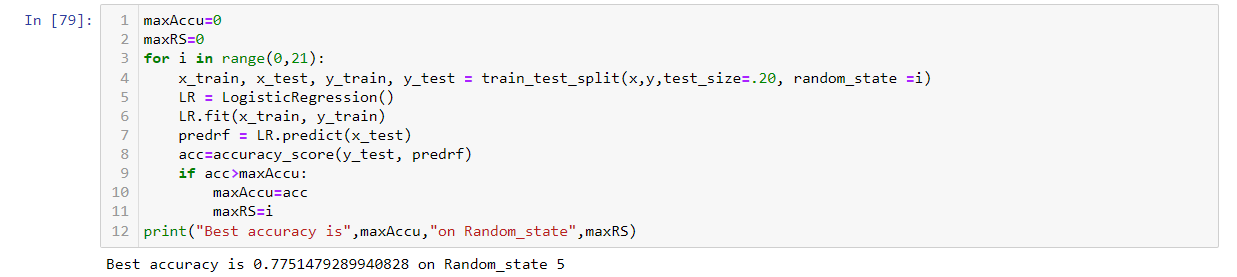


The data is balanced now, Since the highest count of the target is 422 so the data is balanced by oversampling all the classes to the count 422.



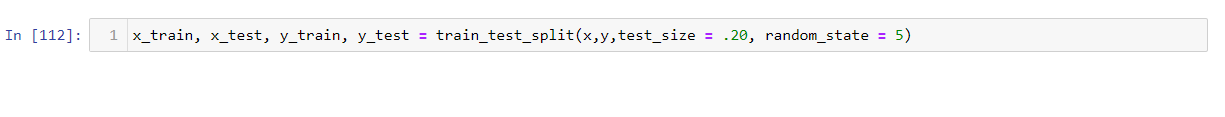
Random State , Max and Accuracy

* Randomness is a big part of machine learning. Randomness is used as a tool or a feature in preparing data and in learning algorithms that map input data to output data in order to make predictions.
* MAX will return the largest value in a given list of arguments. From a given set of numeric values, it will return the highest value. Unlike MAXA function, the MAX function will count numbers but ignore empty cells, text, the logical values TRUE and FALSE, and text values.
* Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

Now we are checking the best accuracy for our model

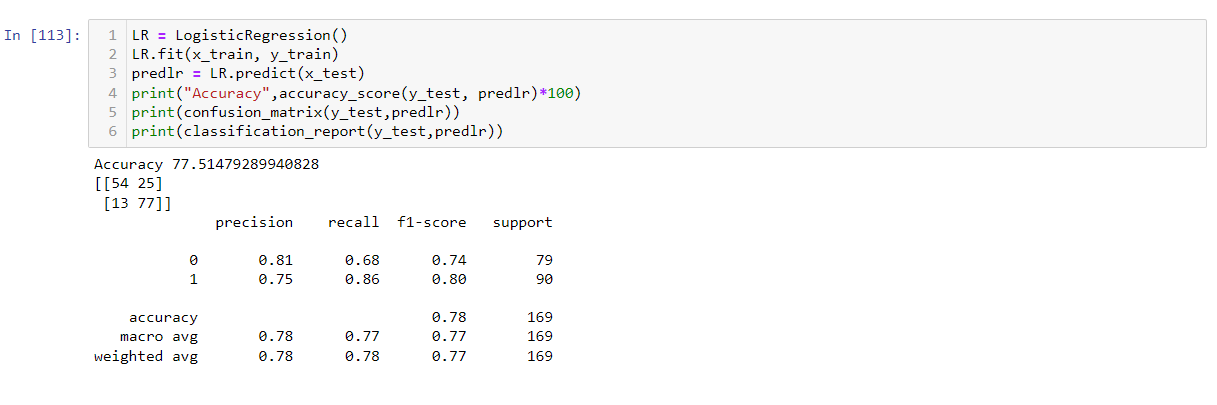
Splitting data into training and testing

The simplest way to split the modelling dataset into training and testing sets is to assign 2/3 data points to the former and the remaining one-third to the latte**r**. Therefore, we train the model using the training set and then apply the model to the test set. In this way, we can evaluate the performance of our model.

We have created a new train test split using Random State

Logistic Regression

It is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis can help you predict the likelihood of an event happening or a choice being made.



The accuracy using Logistic Regression Classifier is 77.514%

Confusion Matrix

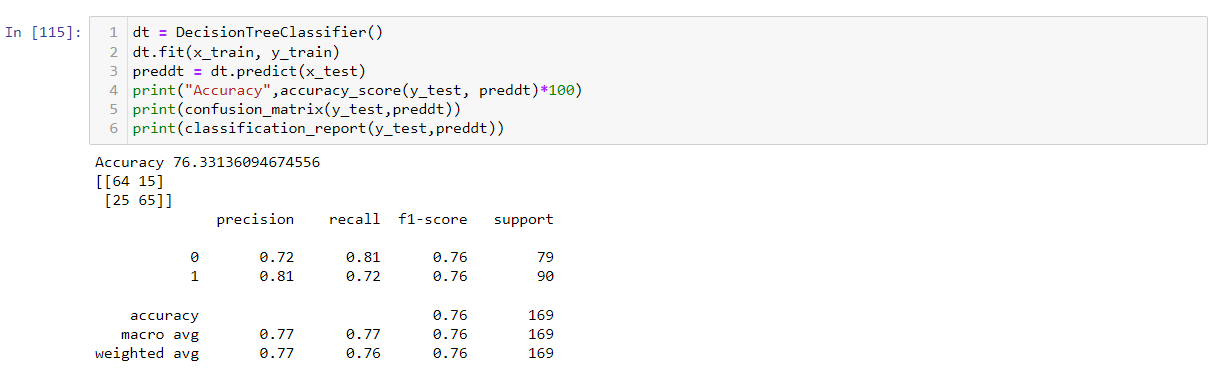
A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



Confusion matrix for logistic regression classifier

Decision Tree Classifier

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute.

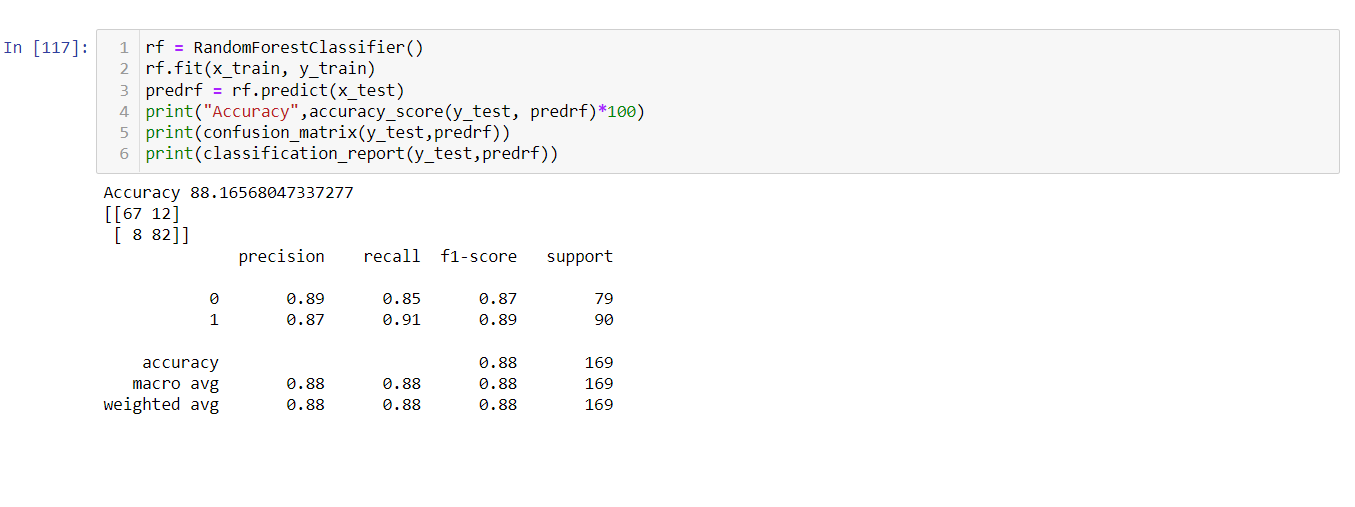




Dcision Tree Classifier giving the 76.33% of accuracy

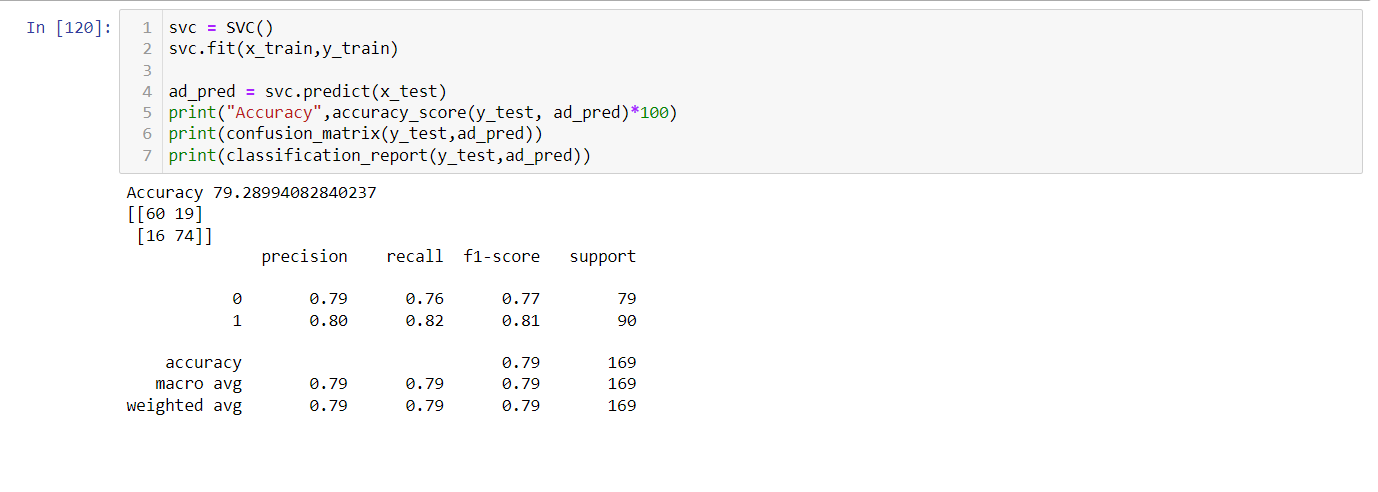
Random Forest Classifier

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.



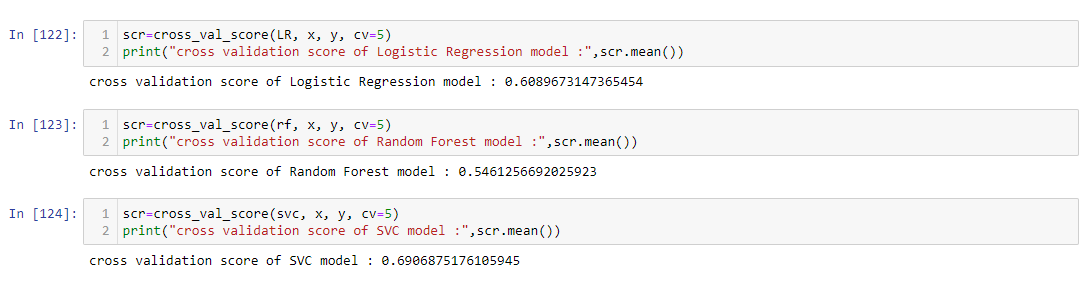
Random Forest Classifier giving the 88.16% of accuracy

SVC

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

SVC giving the 79.28% of accuracy

***Checking the Cross Validation Score***

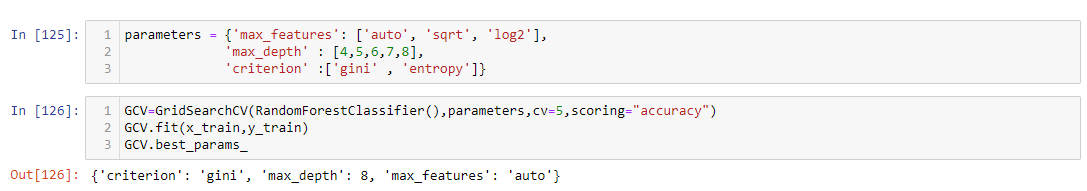
Cross-validation is a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, i.e, failing to generalize a pattern.

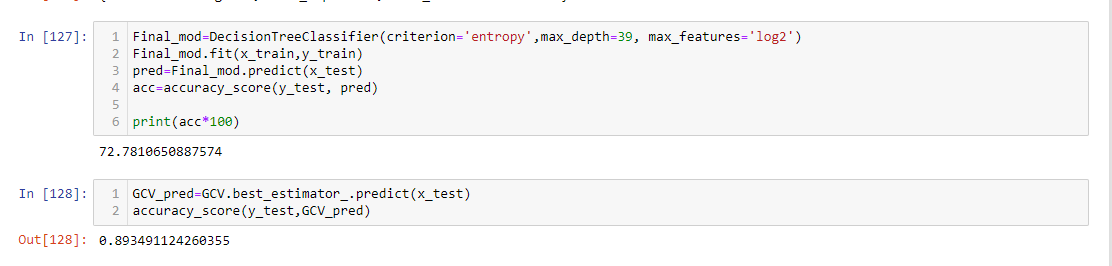
**Difference between Accuracy and Cross validation Score**

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy** | **CV Score** | **Difference** |
| Random Forest Classifier | 88.16 | 54.61 | 33.55 |
| Logistic Regression | 77.51 | 60.89 | 16.62 |
| SVC Classifier | 79.28 | 69.06 | 10.22 |

Hyper Parameter Tuning (Using GridSearchCV)

In GridSearchCV approach, machine learning is evaluated for a range of hyperparameter values. This approach is called “GridSearchCV” , because it searches for best set of hyperparameters from a grid of hyperparameters values.

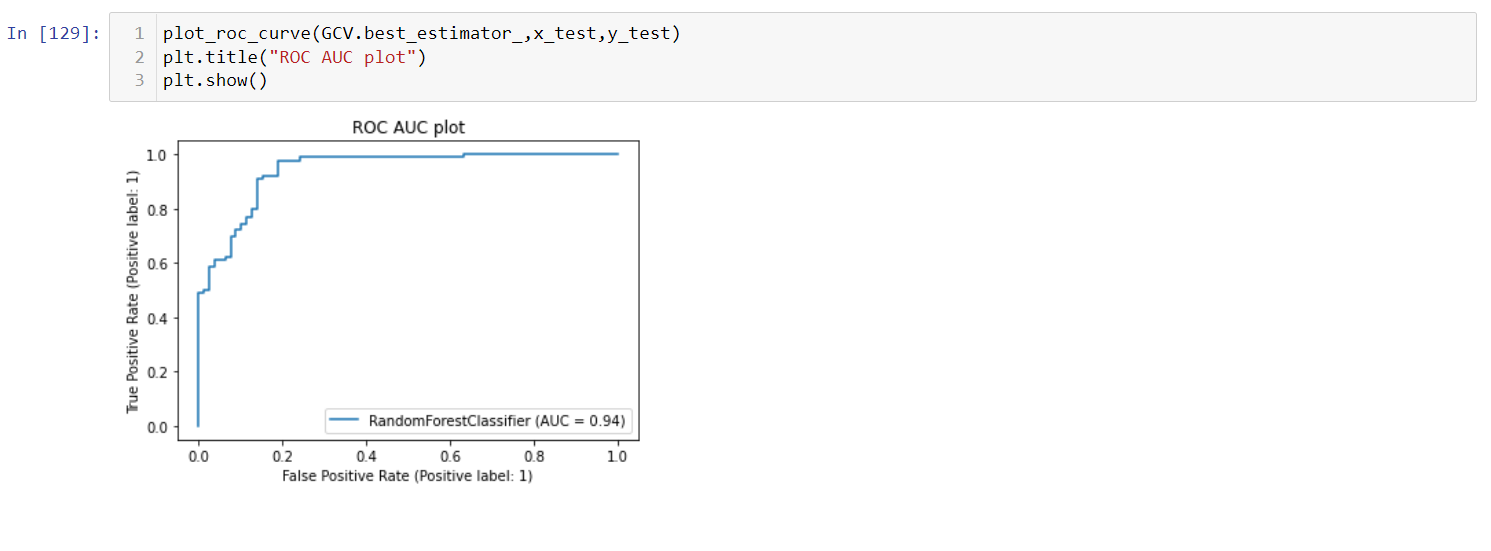




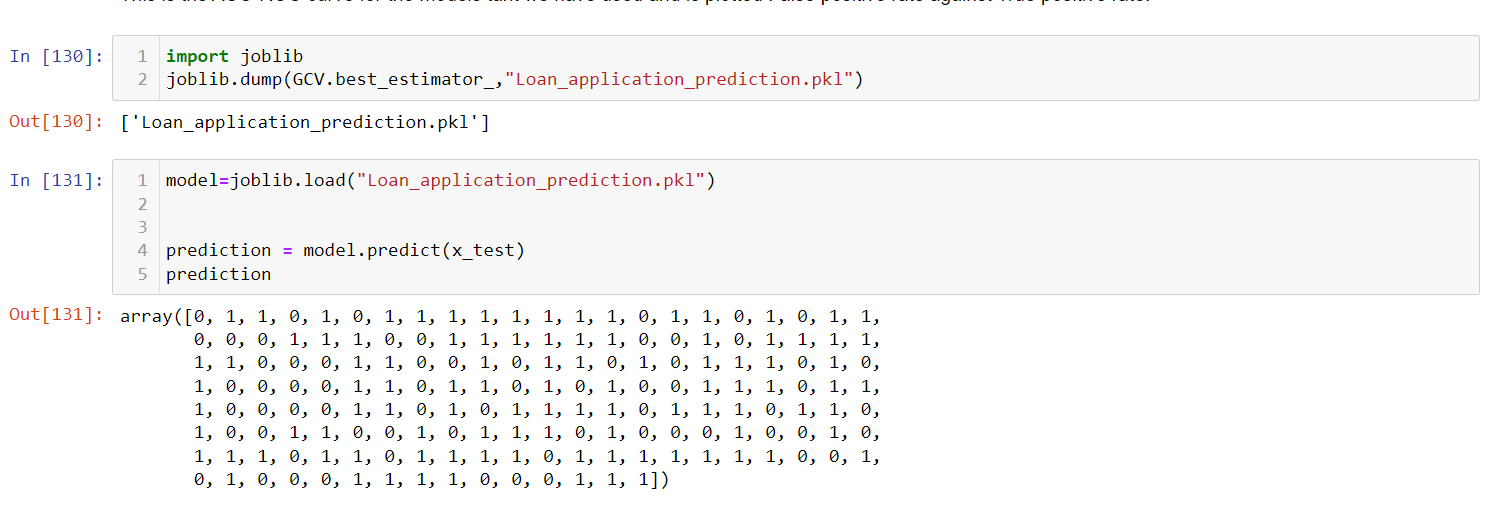
AUC-ROC Curve

It is a graph that shows the performance of a classification model at all possible thresholds. The curve is plotted between two parameters

* True Positive Rate
* False Positive Rate

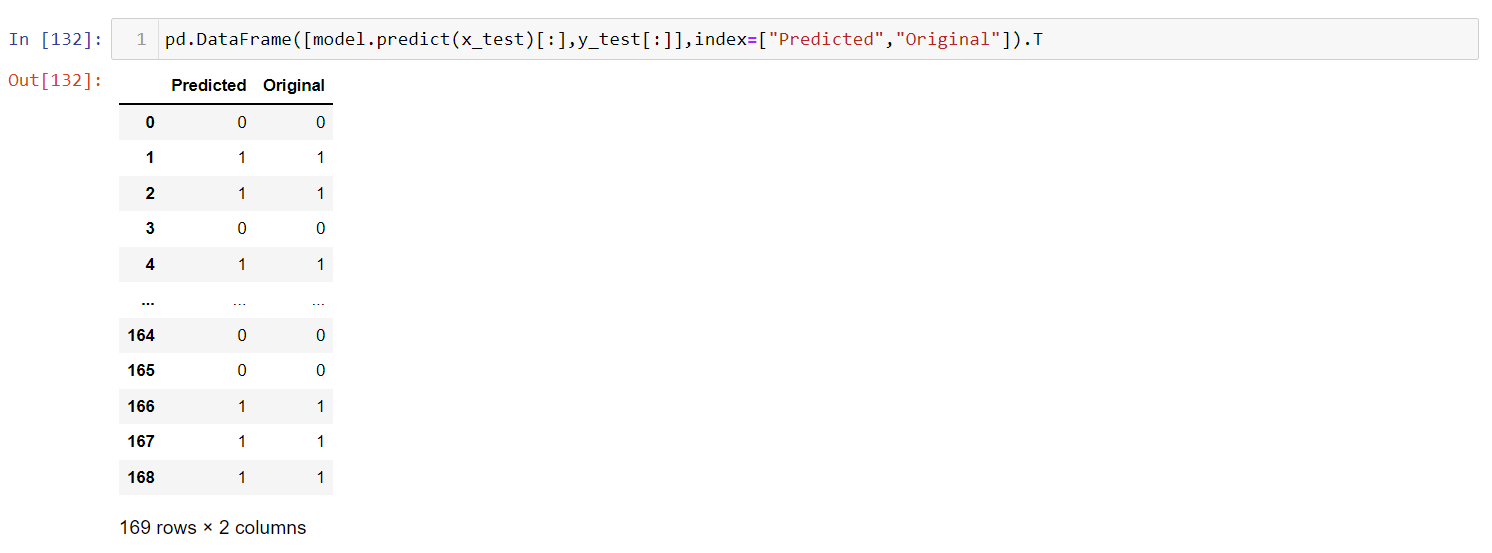


Saving Model & Predicting the save model



We have loaded the saved model to get predicted values.

Let’s compare the predicted and actual values.



We can see from the above observation, Predicted and the Original values are matching each other, which means model performance is good.

Conclusion Mark

I introduced the whole pipeline of an end-to-end machine learning model in a loan-approved prediction, with a loan approval dataset. I described the loan approval dataset and the relationships between each table. Steps and codes were demonstrated on how to import the dataset into MySQL database and then connect to Python and convert processed records into Pandas DataFrame. Features were extracted and transformed into an array, ready for feeding into machine learning models. As the last step, I fit data into four model , evaluated the model performance, and generated the list of top 5 features that play roles in predicting loan approval.

This machine learning pipeline is just a gentle touch of the one application that could be used with the loan approval dataset. It could go deeper since there is more useful information hidden in the intricate relationship among tables; it could also go wider since it can be extended to other applications . In this project we have gone through the feature engineering which is the most important thing to get the better performance models, we have removed the outliers, skewness and also handled the categorical columns by encoding the data, scaled the data, handled the data imbalance and at last, we built the different classification models to predict the attrition and perform the hyper parameter tuning to improve the model accuracy by using different parameters.

With the help of above techniques, our model is able to give the good performance and it can help in the organization to understand the attrition and overcome from the issue.