

**INTRODUCTION**

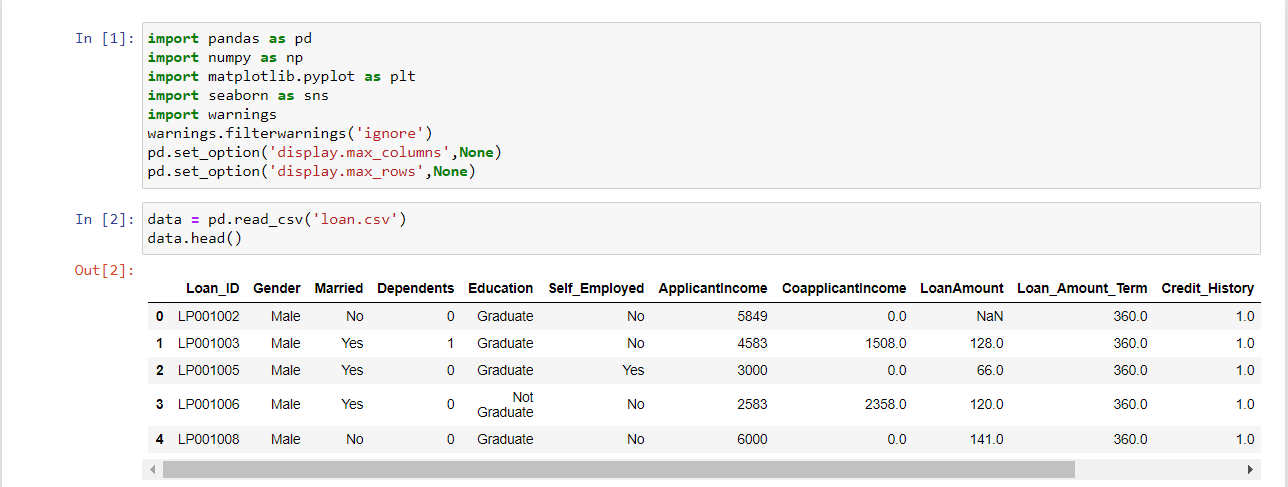
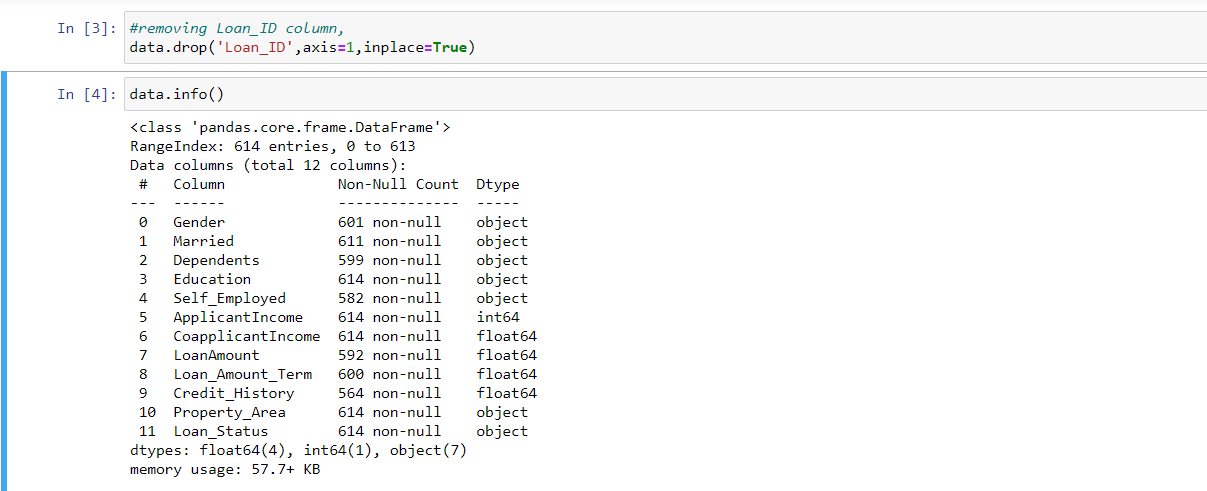
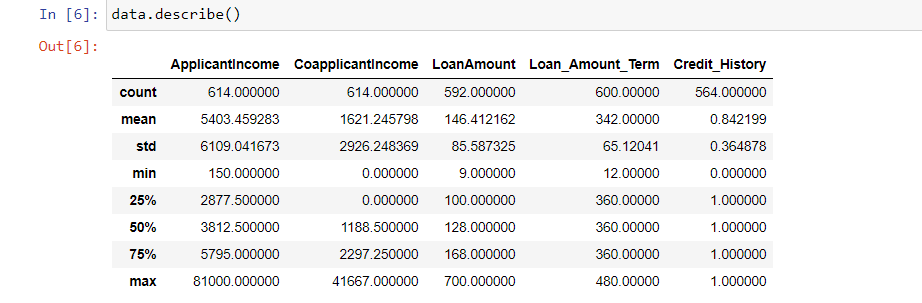
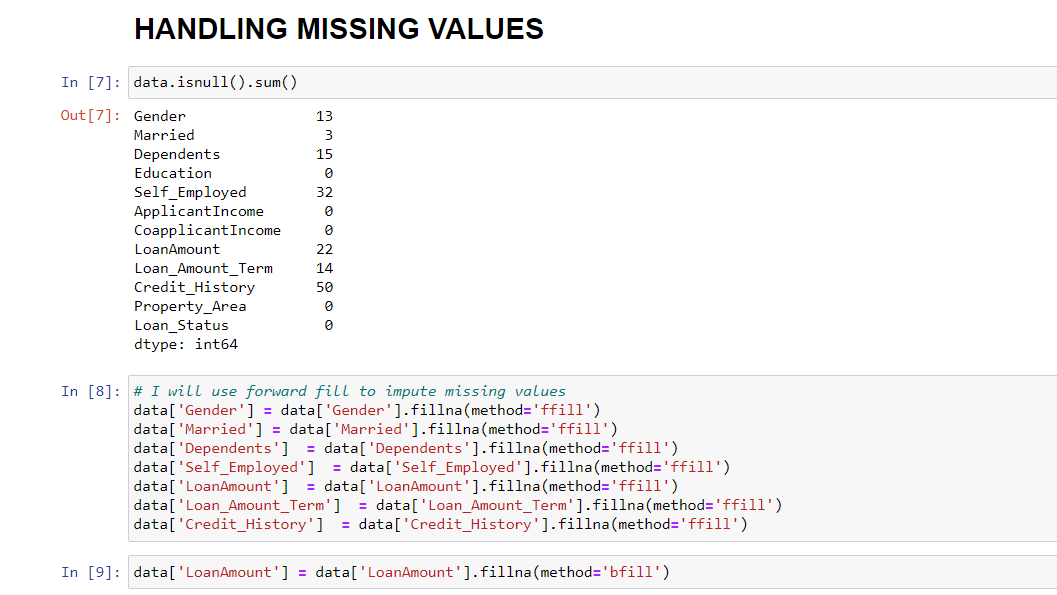
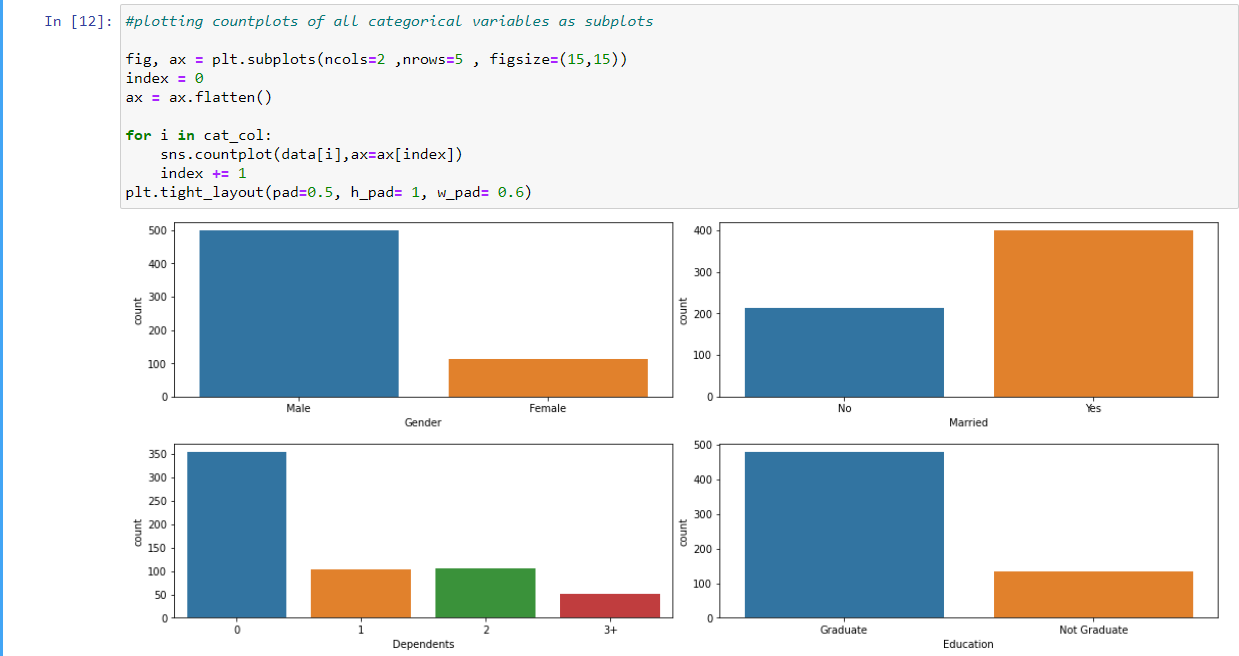
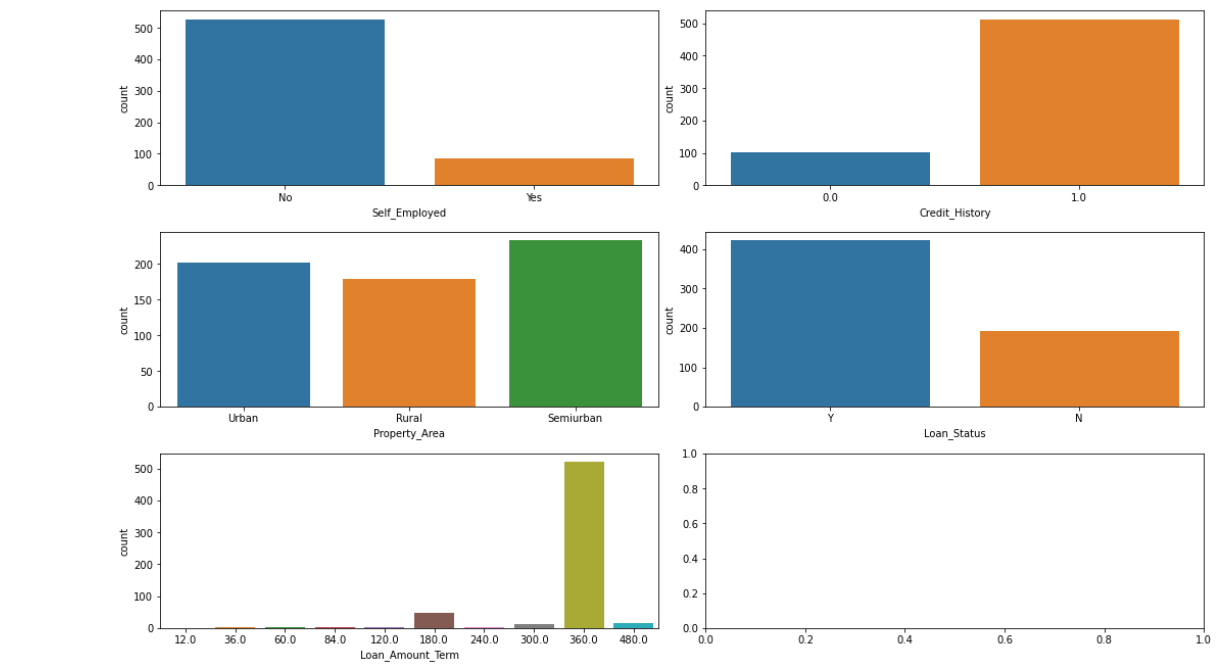
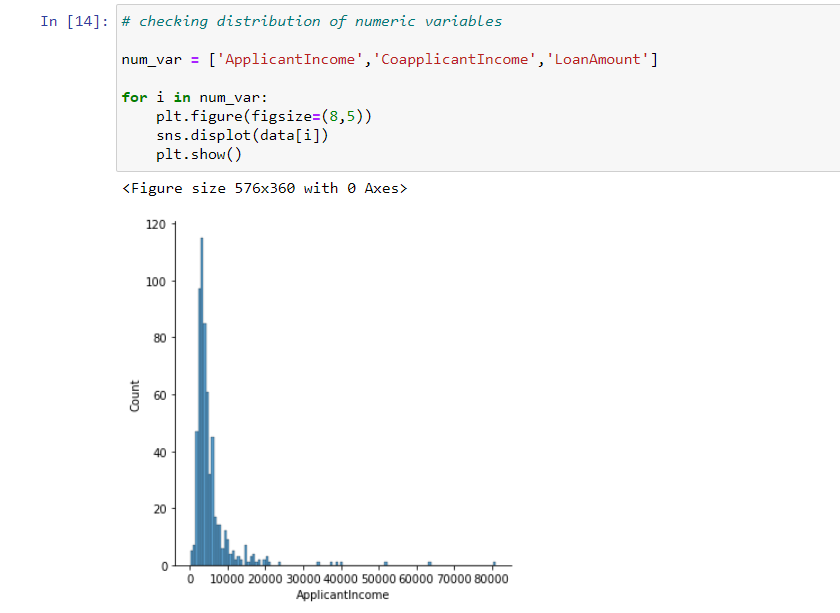
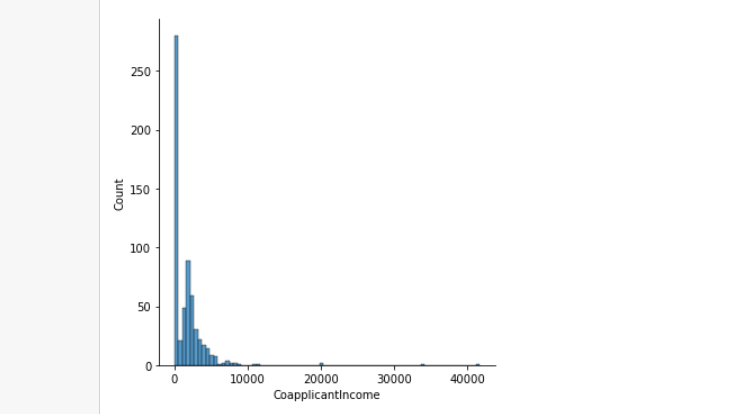
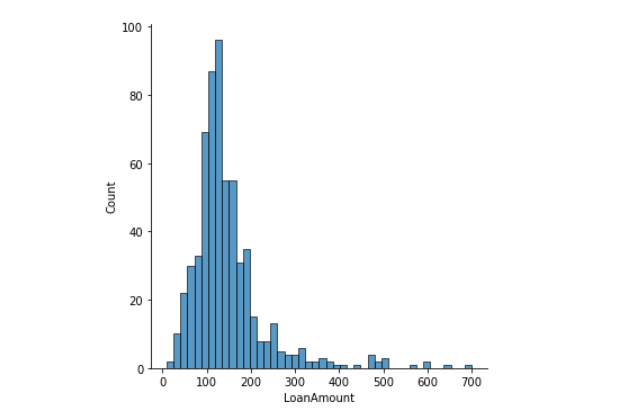
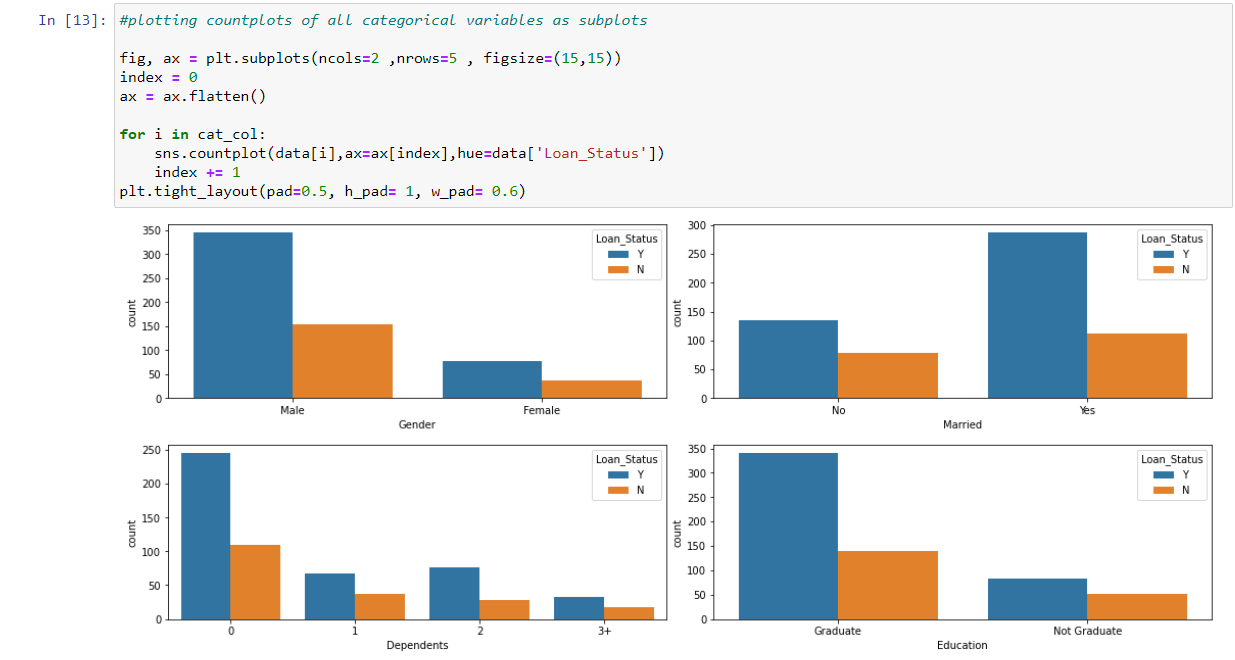
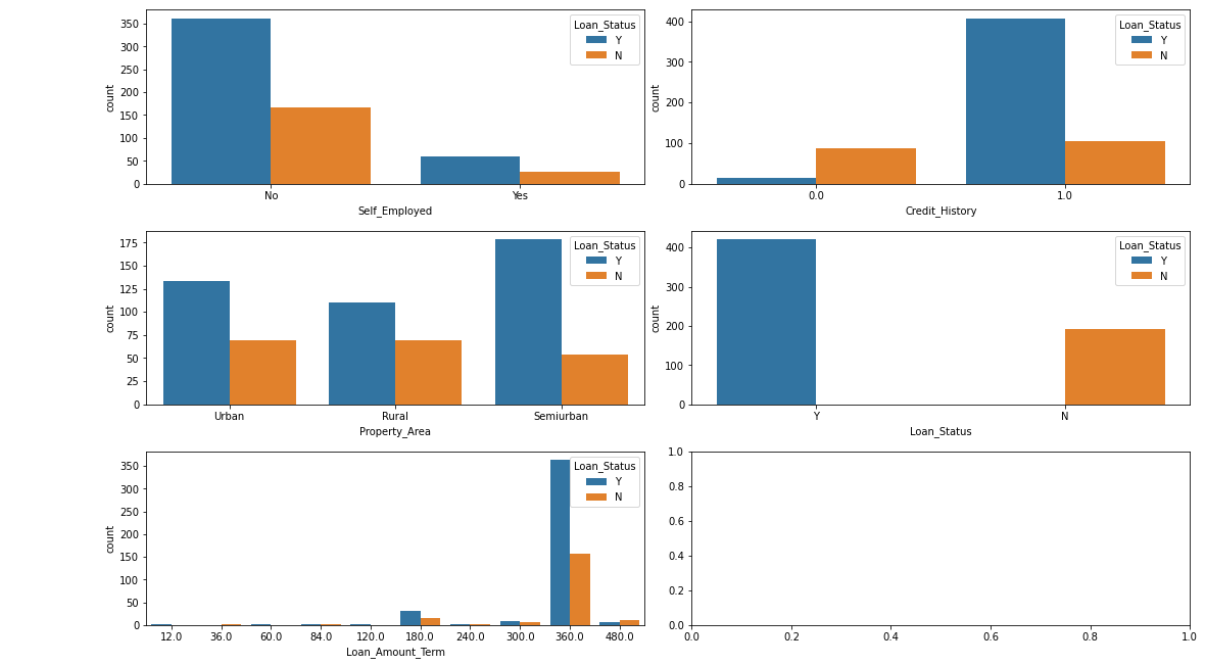
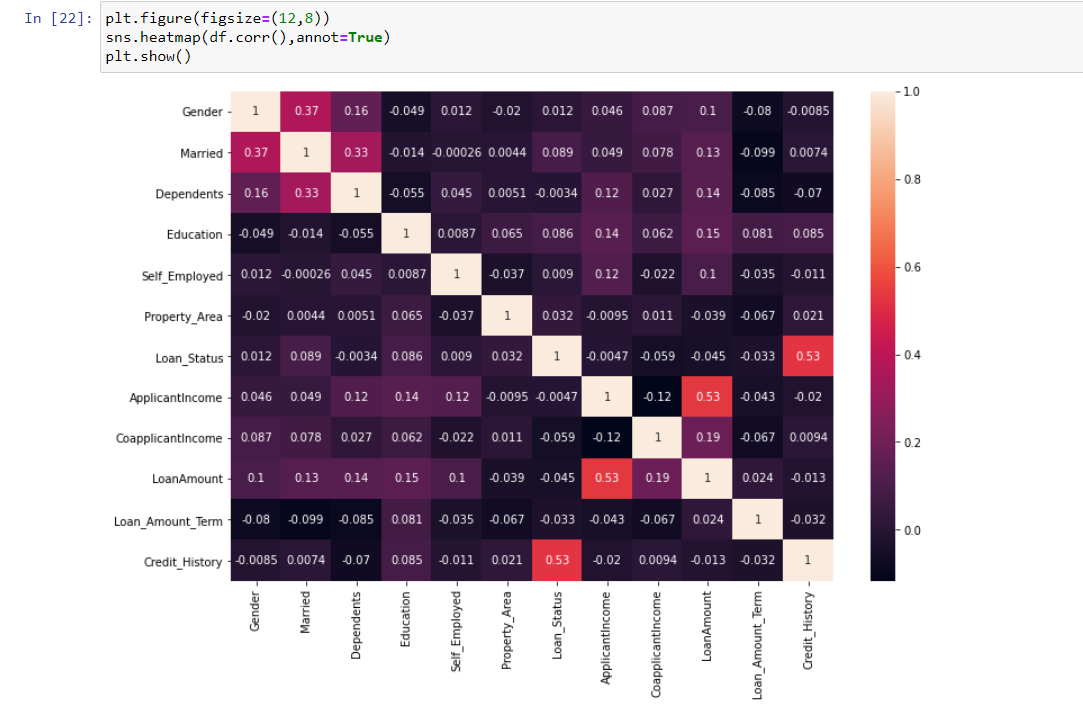
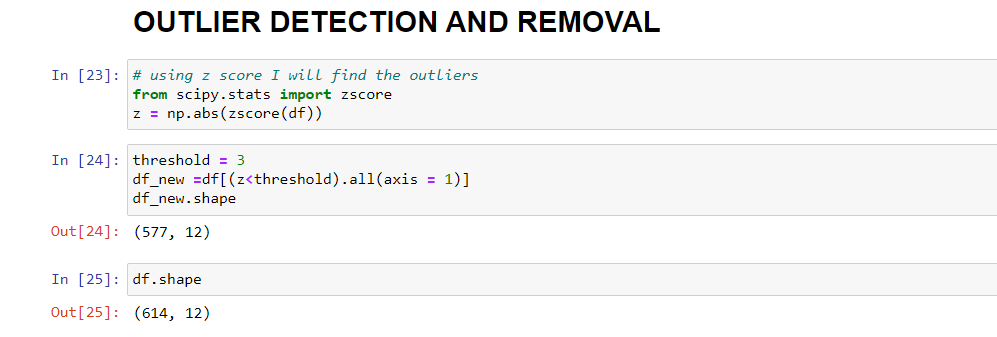
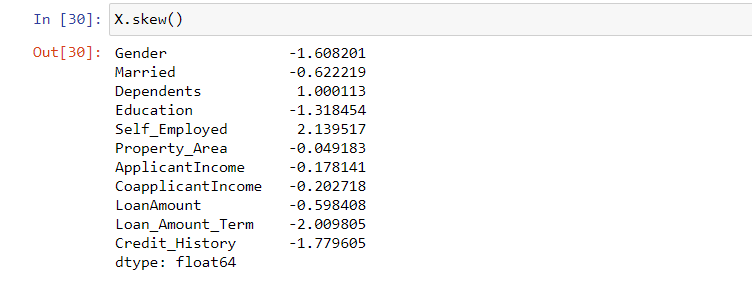
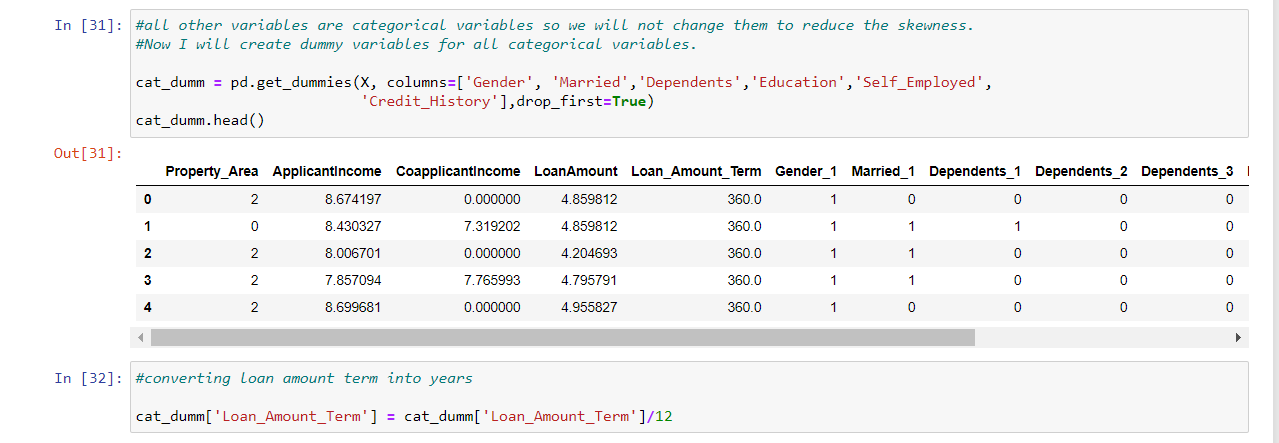
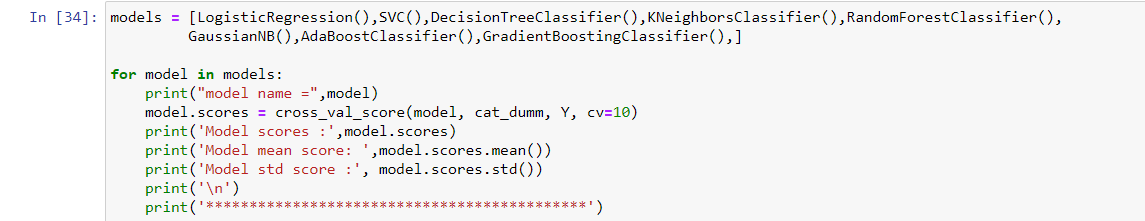
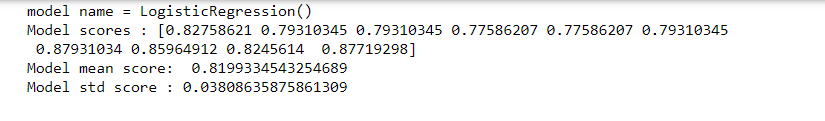
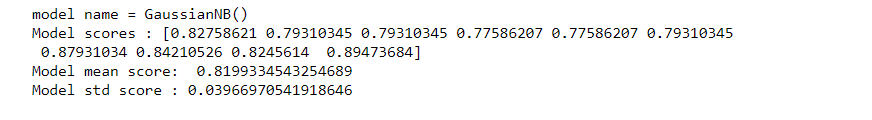
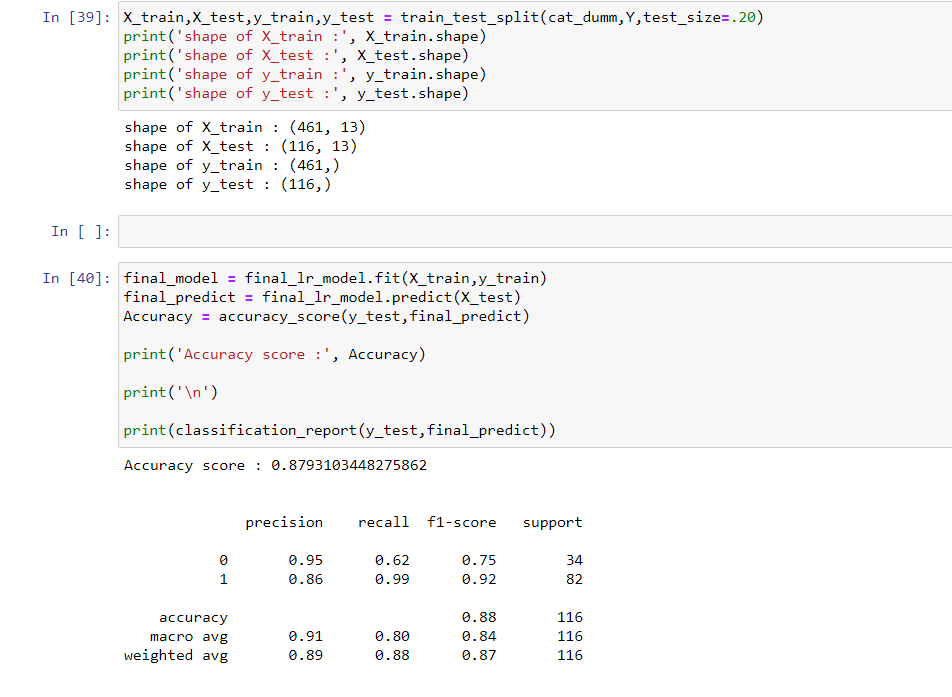
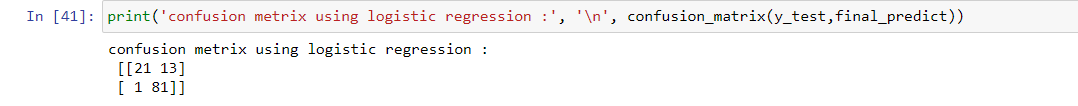
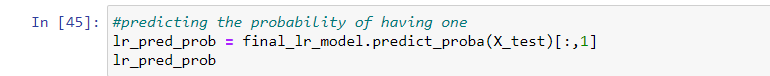
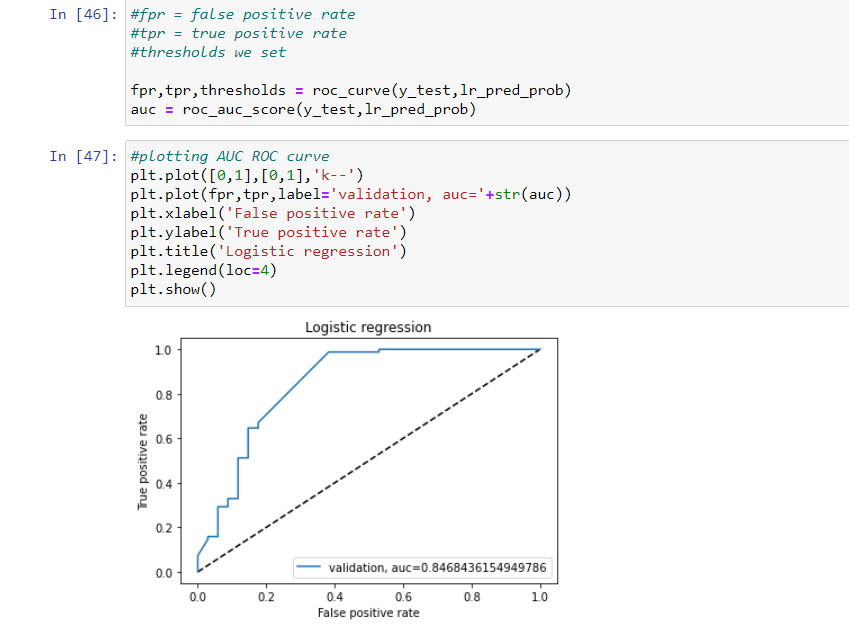
Banking sectors earns most of their profit from the interest, they charge on the amount lent to any individual, firm or company. That amount is known as loan.  
Different banks have different rate of interest depending upon the amount and term of the loan.  
Approval of loan is based on many factors like income, purpose, credit history and etc.  
There is a lot of verification process done before loan approval.   
The banks or any financial institution wants to make sure that the individual, firm or anyone should be able to repay the loan amount with the interest within time.

There are many types of loan provided by the banks and other financial institution.

* Home Loan.
* Personal Loan.
* Car Loan.
* Credit Card.
* Small Business Loan.

In this project work, we will analyze the given data and make a model which will helps us in decision making whether the applicant should be approved for loan or not.  
We will be using Jupyter notebook for our project work.

**NOW LET’S TAKE YOU TO THE JOURNEY OF DATA ANALYSIS OF THE SAME.**

1. **PROBLEM STATEMENT**  
     
   Our problem statement here is that we are required to predict our target variable, our target variable is loan status. It is a binary classification. The data set can be downloaded from [here](https://www.kaggle.com/granjithkumar/loan-approval-data-set). I will be analyzing what are the main factors that affect the loan status and based on our analyzing we will predict if the Loan status is yes or no.
2. **DATA ANALYSIS**  
     
   We will be loading important libraries data set in Jupyter notebook  
     
   Here, we have removed the Loan\_ID column from our data set and then left with 12 columns, initially we had 13 columns and 614 observations. And we have 9 categorical variables and 3 continuous variables.  
     
   Here, Loan amount term and credit history columns are described as a float type data, however after checking we found that they are categorical variable. So later we will be converting them into an object type data for graphical representation of our data.  
     
     
   Above image shows the statistical distribution numeric variables of our data set.  
   Applicant income shows that the minimum income of an applicant is 150 and maximum is 81000. Here we can see that the difference between min-max and quantile is very huge and the data is right skewed, there are chances of outliers present in the data.  
     
   Coapplicant income shows that minimum coapplicant income is 0 and maximum is 41667, here again we see a right skewed distribution and chances of outliers present in the data.  
     
   Loan amount is ranging from 9 to 700, and here again we see a right skewed distribution and presence of outliers.  
     
   We will later convert these columns using log transformation to reduce their skewness and then pass it to our model building.
3. **EDA**  
     
   **1. HANDELLING MISSNG VALUES**  
     
   In our data set we have many missing values in many columns as shown above, instead of removing the observations having missing values it’s better to impute them, here I have used forward fill method for imputing the missing values, and then used backward fill method for Loan amount as there was missing value in the 1st observation of that particular column.  
     
     
   **2. UNIVARIATE ANALYSIS**  
     
     
   We have 1st plotted the categorical variables.  
   From the above graphs we got the following insights:  
     
   - Count of male's applicant is more than of females.  
   - Number of Married applicants are higher.  
   - With 0 dependents the number of applicants is higher whereas at 3+ dependents the   
    the number of applicants is the least.  
   - Graduate people are more likely to go for loan, and they are more likely to get   
    approved for the loan.  
   - Applicant having credit history are more likely to apply for loan and the chances  
    of loan getting approved are higher when there is good credit history.   
    There are more applicants having credit history.  
   - Maximum number of applicants are of those whose property is in semiurban area  
    and least of those having property at rural area.  
   - In our data set we have more applicants whose loan status is Yes.  
   - Loan amount terms, here we can see that the count for loan application is more  
    for 360 months or 30 years, as this allows more time for the applicant to repay  
    the amount to the bank.  
     
   Now We will be creating distplot for our numerical variables.  
     
     
     
   - Applicant income shows a right skewed distribution of data, at 2500 the frequency is the highest, income of most of the applicants ranges from 150 to 10000, maximum income is at 81000, and mean income is 5403.459283.  
     
   - Co-applicant income also shows a right skewed distribution, most of the income of   
    co-applicant is 0, the reason could be that the applicant is not married or the partner is not working. Mean income of co-applicant is 1621.245798.  
     
   - Loan amount ranges from 9 to 700, from 50 to 150 there are more applicants, avg loan amount is approx. 146.   
     
   We can see that our numeric data is rightly skewed, uneven distributed we will later reduce the skewness using log transformation.  
     
   **3. BIVARIATE ANALYSIS**  
     
     
     
   - Approx. 69% of males got approved for loan where approx. 67% of females got approved for loan.  
     
   - Approx. 66% of married females got approved for loan, where approx. 69% of married males got approved for loan.  
     
   - Loan approved for 69% of applicants having 0 dependents, 64% having 1 dependent, 76.7% having 2 dependents and 62.5% having 3+ dependents.  
     
   - 70.9% of the graduates got approved for the loan, 61.65% not graduate got approved for the loan.  
     
   - Only 8% loan approved when credit history was 0, whereas 79.6% loan approved where credit history was one, it clearly states that credit history is a very imp factor for loan approval.  
     
   - 61.71% loan approved for property in Rural Area, 77.09% loan approved for property area in semiurban, 65.82% loan approved for property area in urban. Semiurban area has the highest loan approval.  
     
   - Loan Term Amount here we can see that most of the applications for loan are for 30 years (360 months / 12) where 70.04% are approved for loan.  
     
   **4. CORRELATION MATRIX**  
     
   As we have categorical variables also, so 1st we will convert the categorical variables into numeric one. Here we have mapped the categorical variable’s values with numeric values.  
   We have assigned the values of Gender (Male = 1, Female = 0), Married (Yes = 1, No = 0), Dependents (0 = 0, 1 = 1, 2 = 2, 3+ = 3), Education (Graduate = 1, Not Graduate = 0), Self Employed (Yes = 1, No = 0), Property Area (Rural = 0, Semiurban = 1, Urban = 2), Loan Status (Y = 1, N = 0). After mapping the values, we created a correlation matrix of our data set.  
     
   the most correlated variables are (Applicant income - Loan Amount), (Credit history - Loan status) and (Loan amount - coapplicant income)  
     
   **5. OUTLIER DETECTION**  
     
   We used SciPy.stats to import z-score library to deal with outliers. To get detailed info about z-score click [here](https://www.statisticshowto.com/probability-and-statistics/z-score/).  
     
   After removing the outliers, we are left with 577 observations.  
     
   **6. SCALING AND LOG TRANSFORMATION**  
     
      
   Here, we have splatted the data into inputs and output variables, then replacing the numeric column with their log transformation to reduce the skewness.  
   After that the last step of data EDA is to create dummy variables for our categorical variables. We created dummy variables of our input data and converted the loan amount term column into years instead of months (divided by 12).  
   
4. **MODEL BUILDING**  
     
   Using scikit-learn we will be testing different classification model on our data set, this is a Binary classification prediction.  
   Testing cross validation on different classification model to check which model is giving us the better result.  
     
   After doing cross validation of different models, we found that Logistic regression and GaussianNB, these 2 models are giving us the better accuracy.  
   With both models we are getting a mean score of 81.99%.  
     
     
     
     
   Now, by using gridsearchCV to get the best parameters for our Logistic Regression model.  
     
   We 1st created a parameter object as a dictionary and then applied gridsearchCV for hyper parameter tuning of our Logistic Regression model.  
   After getting the best parameters we set our model as final\_lr\_model.  
     
     
   for final model building using logistic regression we 1st splitted the data into train and test using scikit learn, our training data is having 461 observations and the test data is having 116 observations.  
   The accuracy score we got from Logistic regression is 87.93% which is good.  
     
   Let's also check the confusion matrix of our model:  
     
   A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.  
     
   here we see that 21 observations are True Positive, and 81 observations are True Negative, whereas 13 observations were actually (N) but model predicted them as (Y) and 1 observation was (Y) but model predicted it as (N). Here (Y) and (N) refers to our target variable Loan\_status (Yes or NO).  
     
   **AUC-ROC CURVE.**  
     
   AUC-ROC curve is used to check the model performance graphically.  
   ROC is a probability curve and AUC shows the degree or measure of separability. It tells how well the model is capable of distinguishing between different classes. Higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.  
     
     
     
   1st we checked the probability of our target variable getting 1. Then finding the TPR (True Positive Rate) and FPR (False Positive Rate) using roc\_curve and then using matplotlib we plotted the AUC-ROC curve.  
   Here we are getting an AUC score of 84.6% which shows that the model performance is good, not best but good.  
     
   At the end we saved our predicted file into .csv format and using joblib library saved the model.  
     
     
   We have also used GaussianNB for our model building, you can check the complete python code from [here](https://github.com/sonirl/LOAN-APPROVAL/blob/main/Untitled.ipynb).
5. **CONCLUSION AND REMARKS**  
     
   We found that the main factors that affect the Loan approval status is the loan term, Loan amount and the credit history.  
   Any financial institution will only approve a loan to anyone who will be able to repay the full amount with in time. We have created a model to classify the applicants if the loan should be approved or not.   
     
   Here we can also add some features or details like interest rate, purpose of loan, we can also build a model where total income would be applicant income + coapplicant income.  
   You can try these and check if the model performance is increasing or not.  
     
   Here we come to an end of our project discussion.  
     
   Hope the reading was helpful, thanks.  
     
   NEVER STOP LEARNING!