**BLOG/ARTICLE**

**SESSION-2020-21**



PROJECT SELECTED: -LOAN APPLICATION STATUS PREDICATION

SUBMITTED BY: - TEJENDRA SONI

SUBMITTED TO: - DATATRAINED

BATCH NO: - 1834

**DECLARATION: -**

I would like to show my gratitude towards data trained & the team for providing us the 17 different projects of classification & regression & pushing me through the challenges, Also I’m thankful to the support team for helping me out for my doubts.

**USED-METHODS: -**

1. JUPYTER NOTEBOOK
2. PANDAS
3. NUMPY
4. MATPLOTLIB.PYPLOT
5. SEABORN
6. LABEL ENCODER
7. STANDARD SCLAER
8. TRAIN TEST SPLIT
9. MODELS

Before understanding the dataset let’s first understand the problem statement that our dataset contains

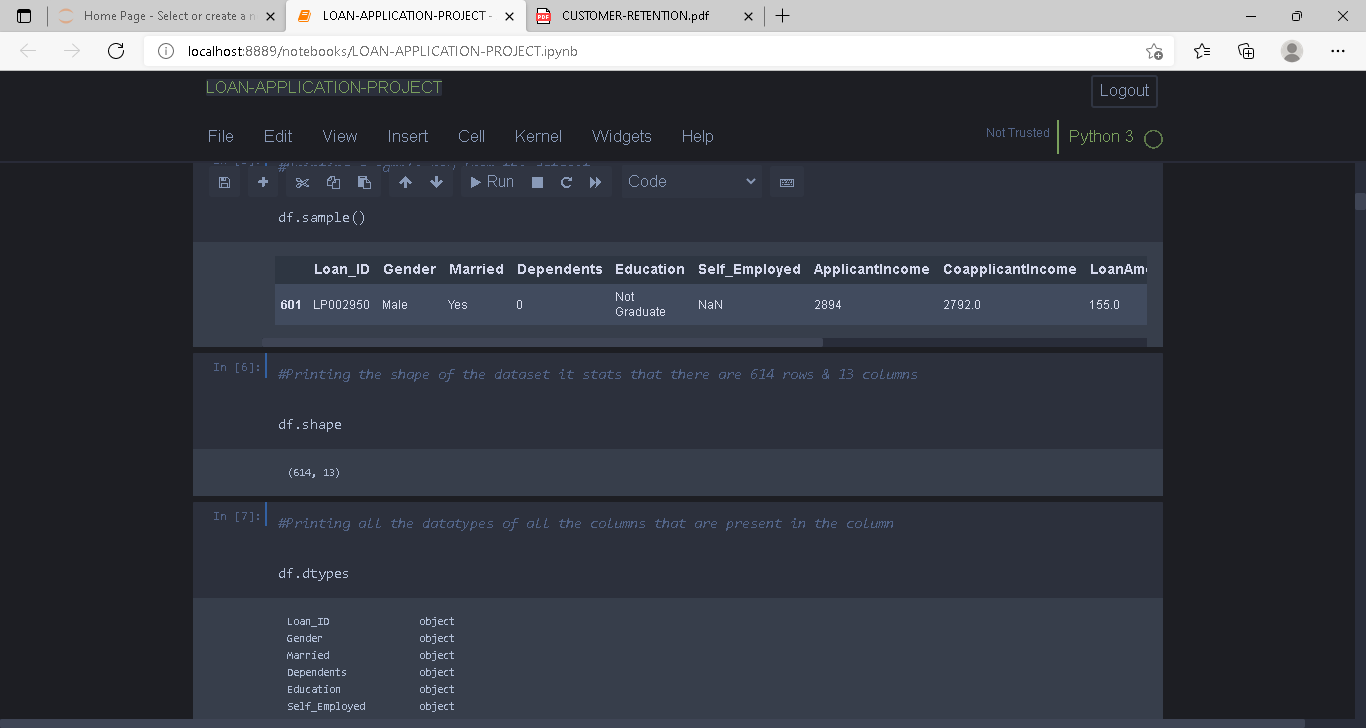
**PROBLEM DEFINITION: -**

This dataset includes details of different individual applicants who have applied for loan. The dataset includes various variables depending on that we will predict that can an individual **loan will get approved or not.**

In the given problem we will try to build a model that will predict that the individual loan will get approved or not on the basis of different given variable like Loan amount, Credit score etc.

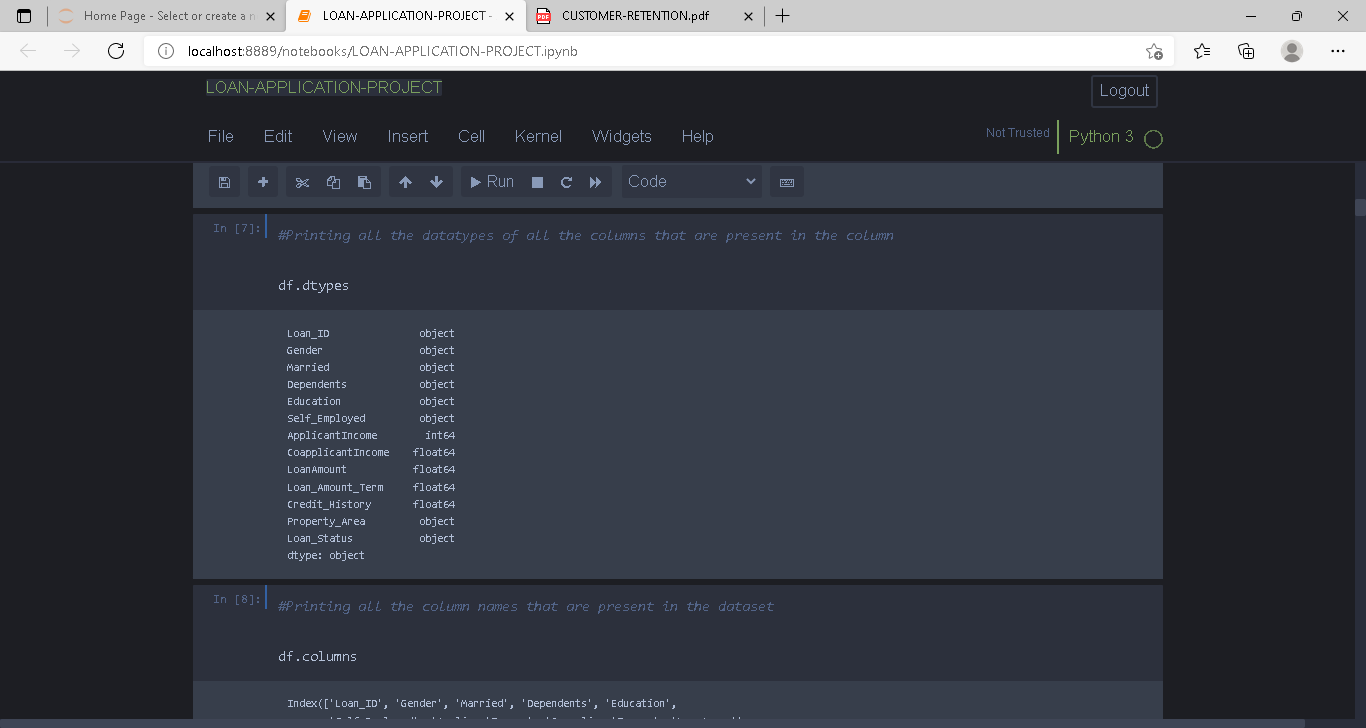
**DATA ANALYSIS: -**

Firstly, we will look into the dataset that what is the shape of our dataset.



Our dataset shows that there is total 614 rows & 13 columns

Now we will be checking the datatypes of all the columns that are present in our dataset.

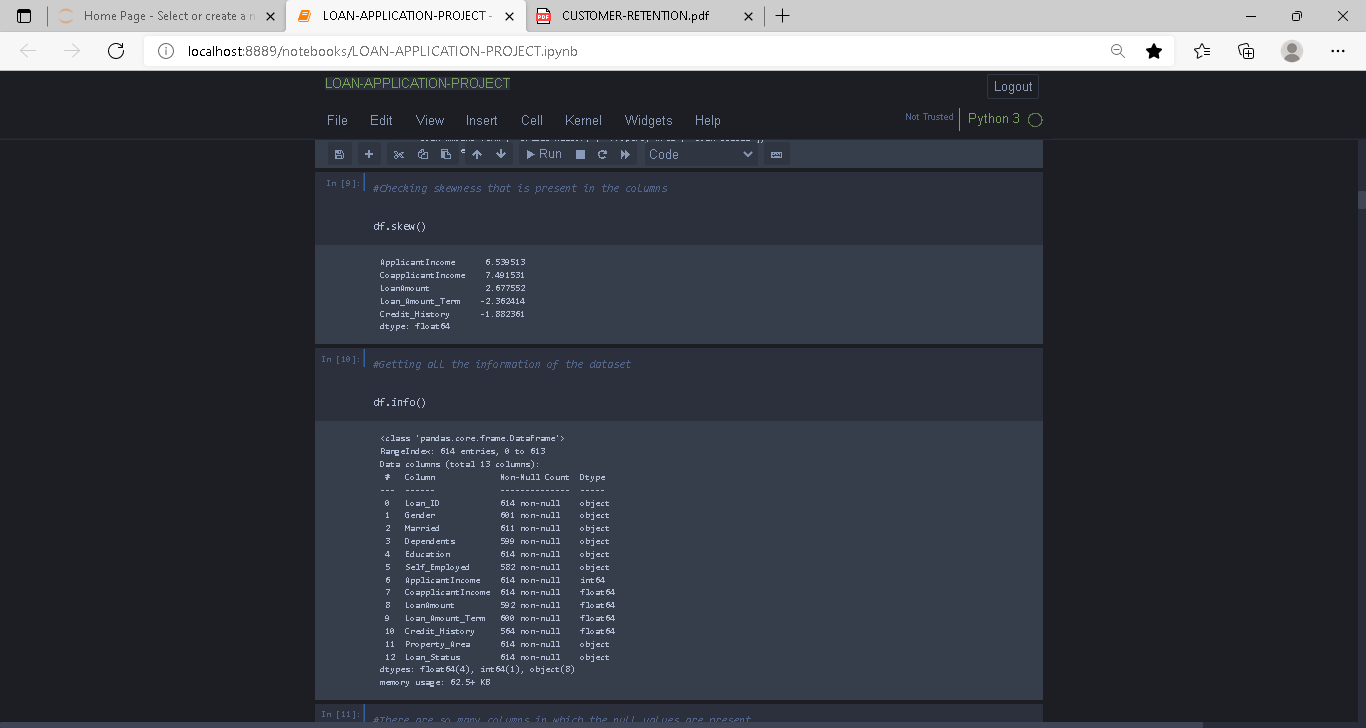


The greens denote object type, red denotes integer type & light blue denotes float type

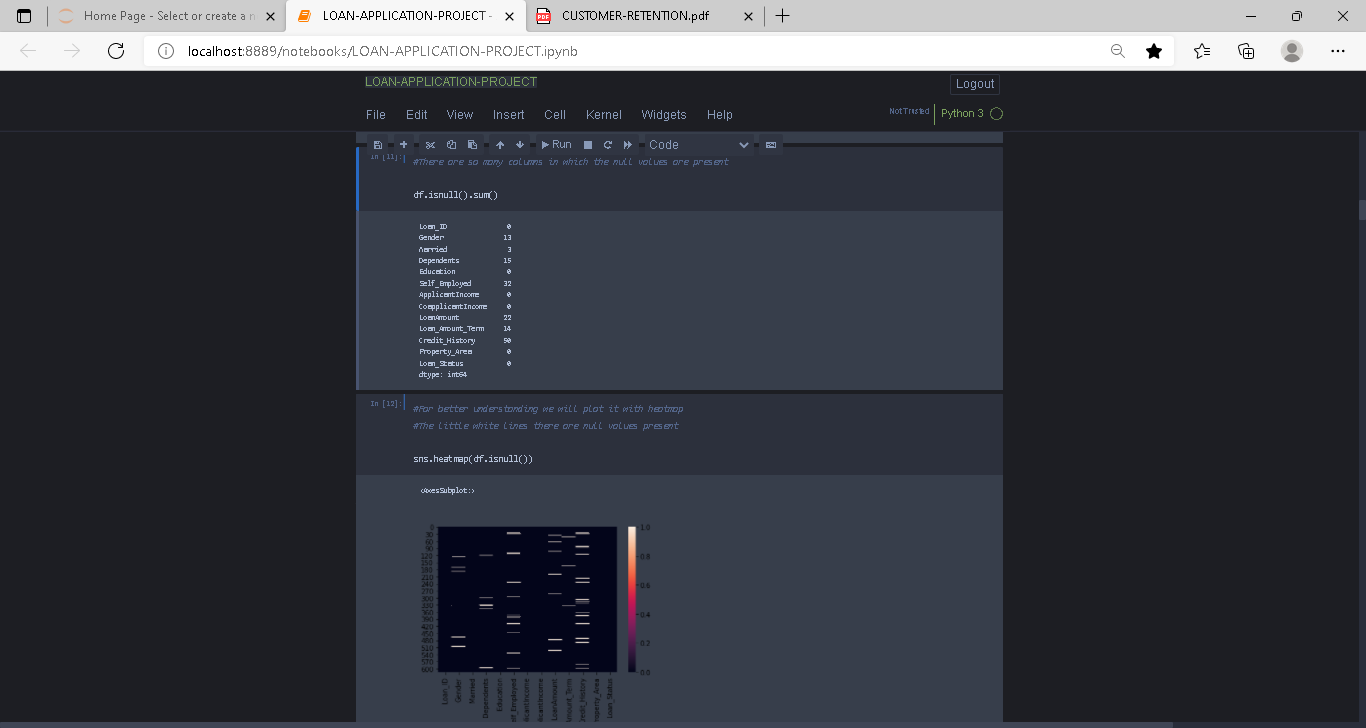
Loan\_id ,Gender, Married, Dependents, Education, Self\_employed, Property\_Area, Loan\_status are the variables that belongs to object type of variable.

Only Application Income is the variable is of Integer type of datatype

And rest of the variables are of float type

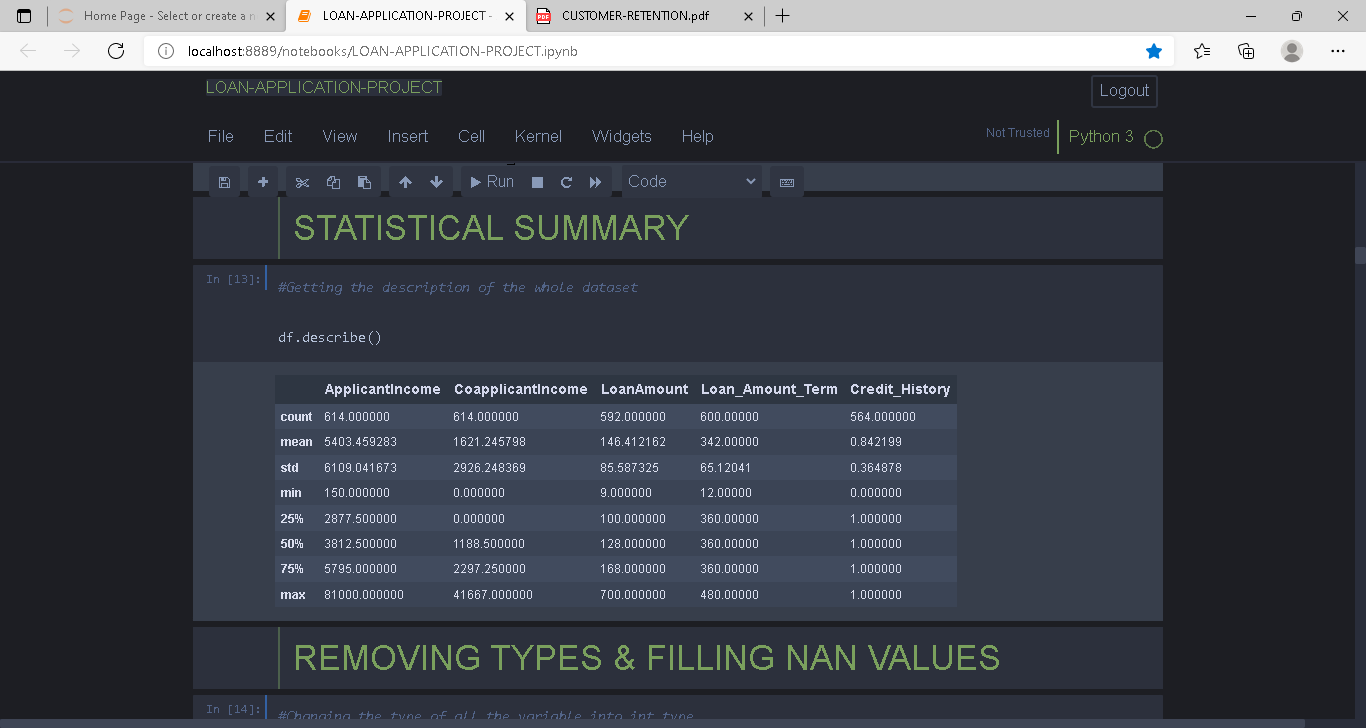


Onto the very next step we are checking the skewness & getting the whole information of the dataset.



There are some null values present in the columns we can see it on the above shown graph & also for better understanding we are plotting the null values with heatmap

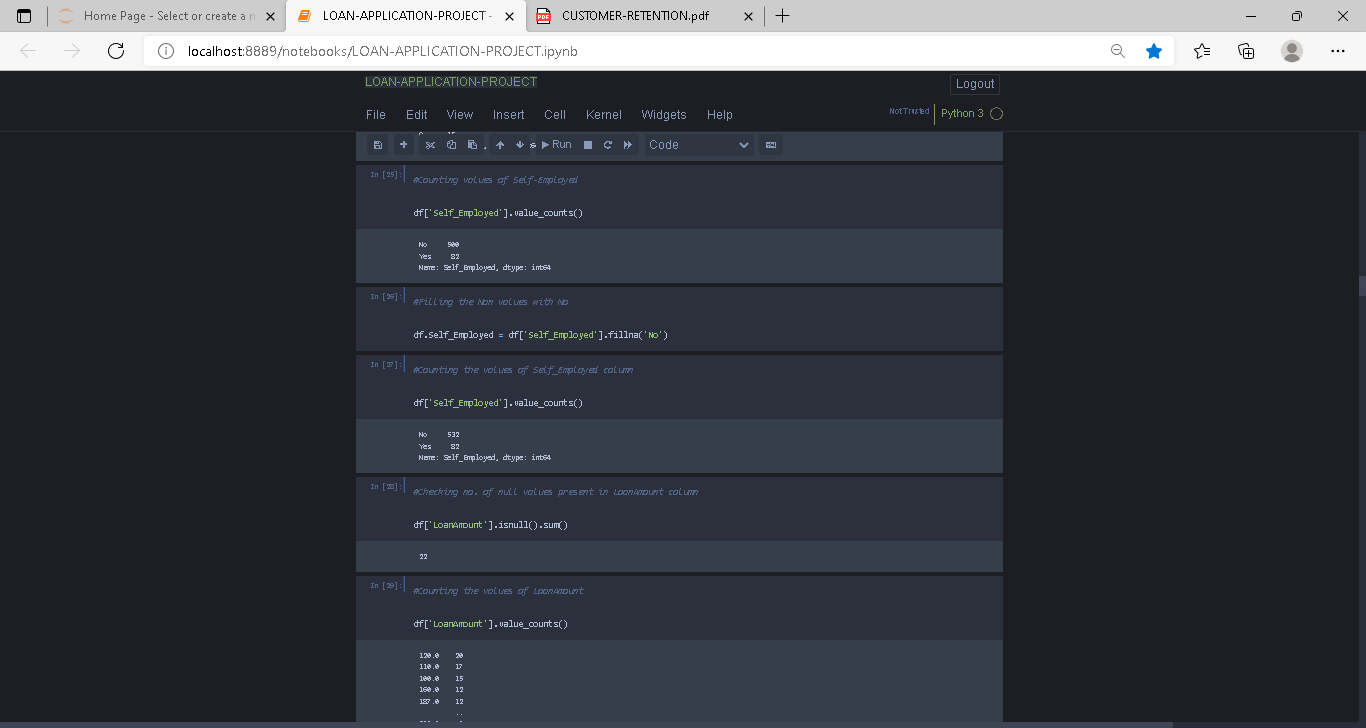
The **white lines** that are shown they are denoting the null values present in the columns.

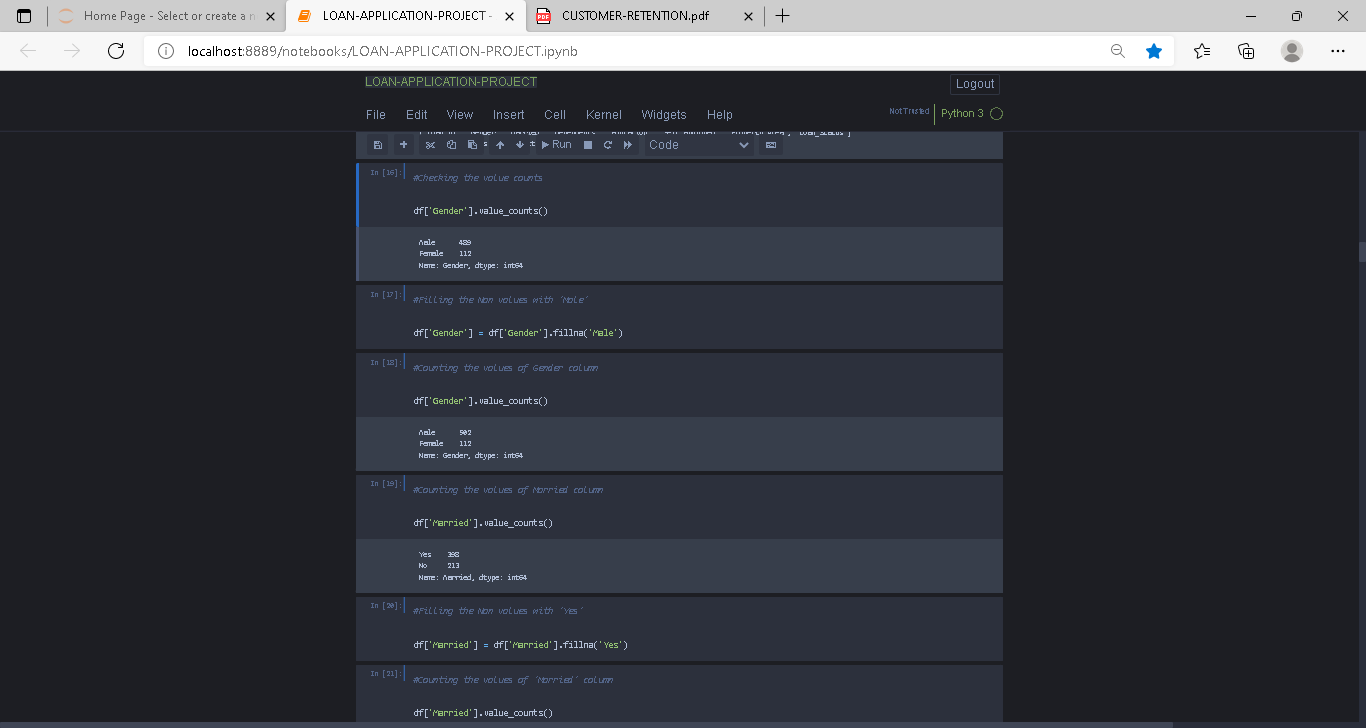


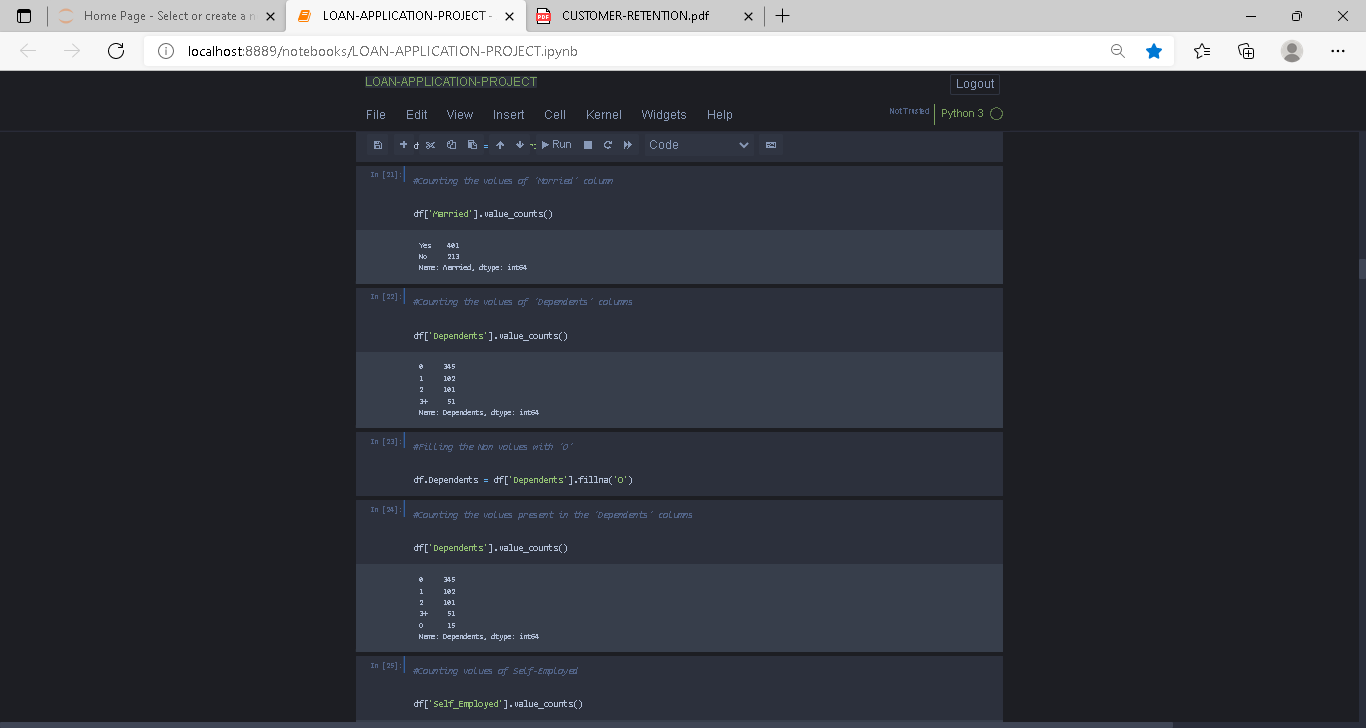
Here we are getting the whole statistical summary of the dataset i.e., here we are getting the **Mean, Standard Deviation, Minimum, Maximum, 25%, 50% & 75% of the columns**

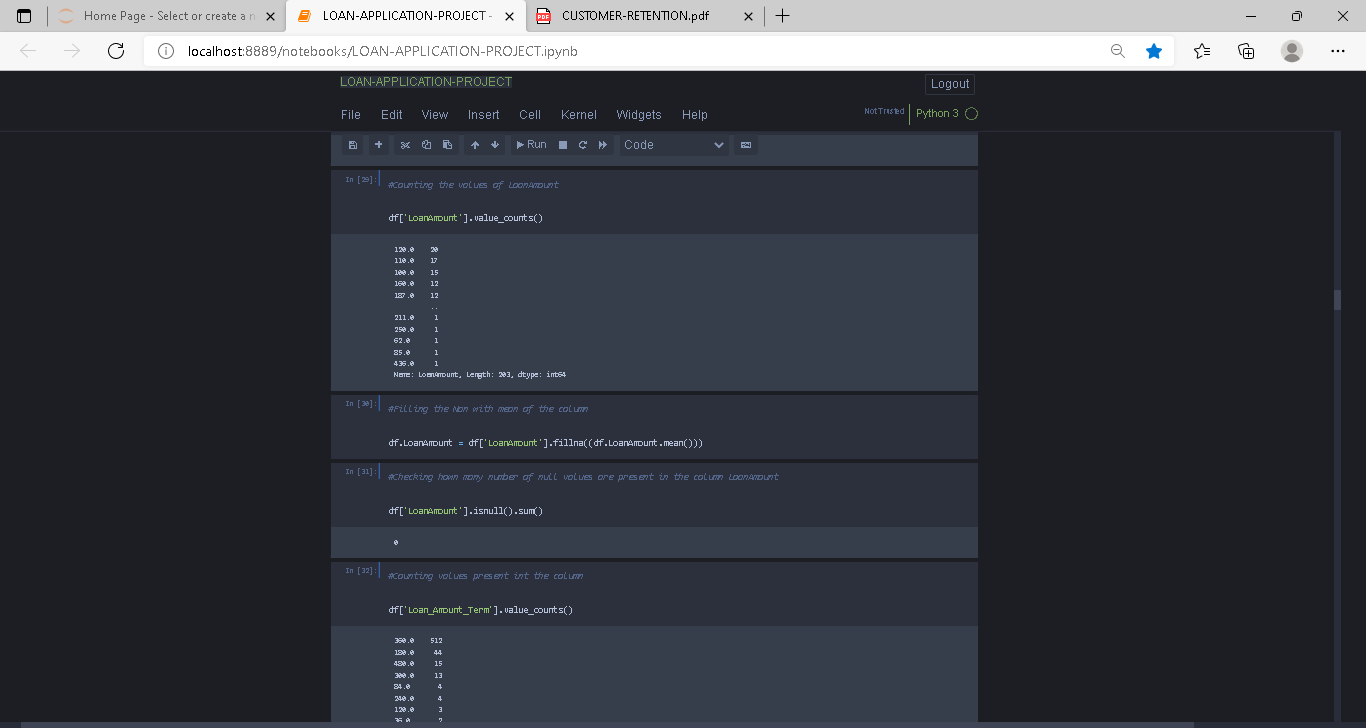
For getting ready for the analysis part, we need to process our data in such a way so that we could have a cleaned & sorted data for the analysis part & for that we will: -

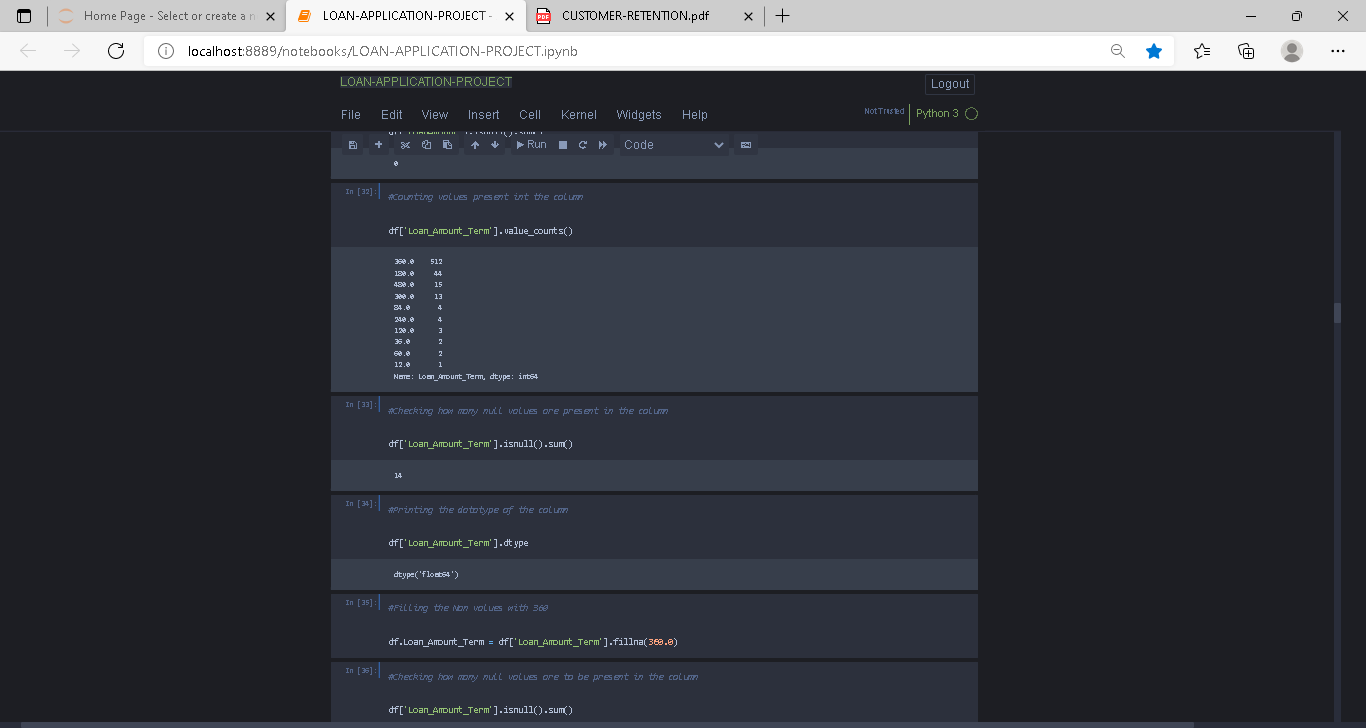
1. **Fill the Nan values** that are present in the columns
2. Encoding the columns
3. Dropping the columns that we don’t need

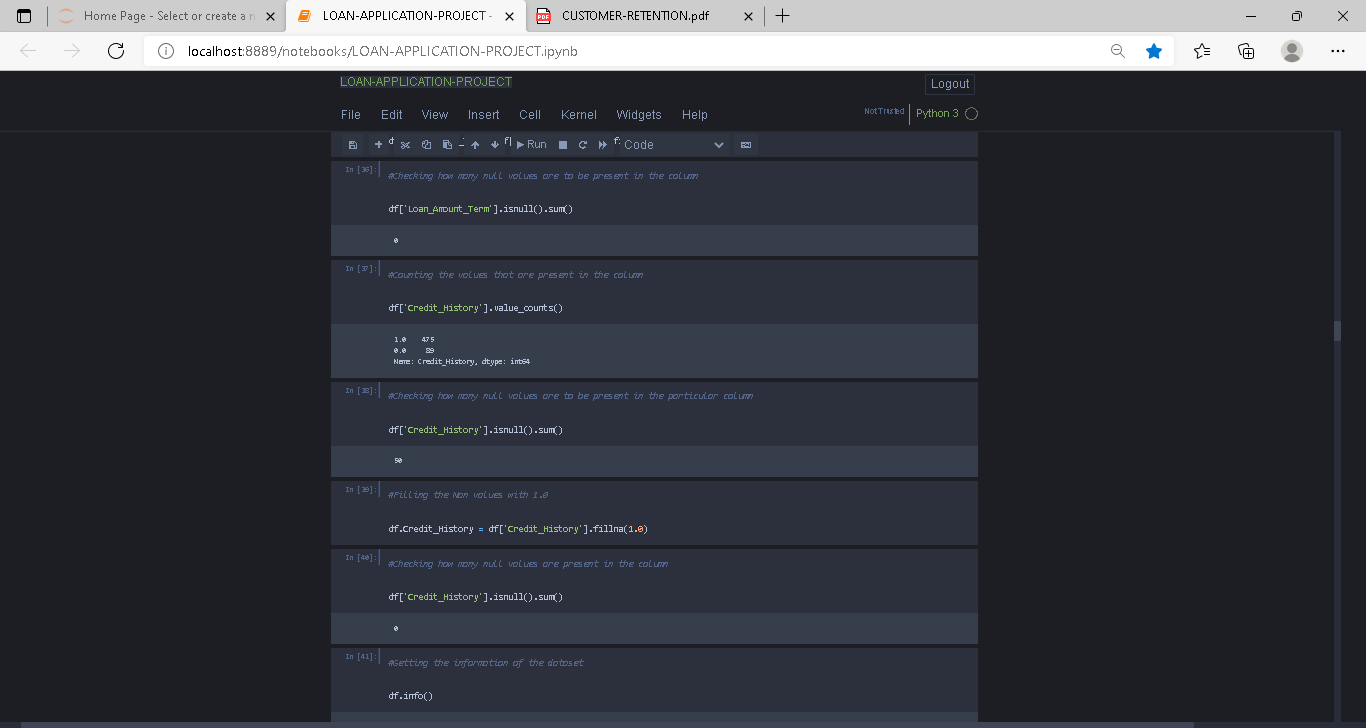






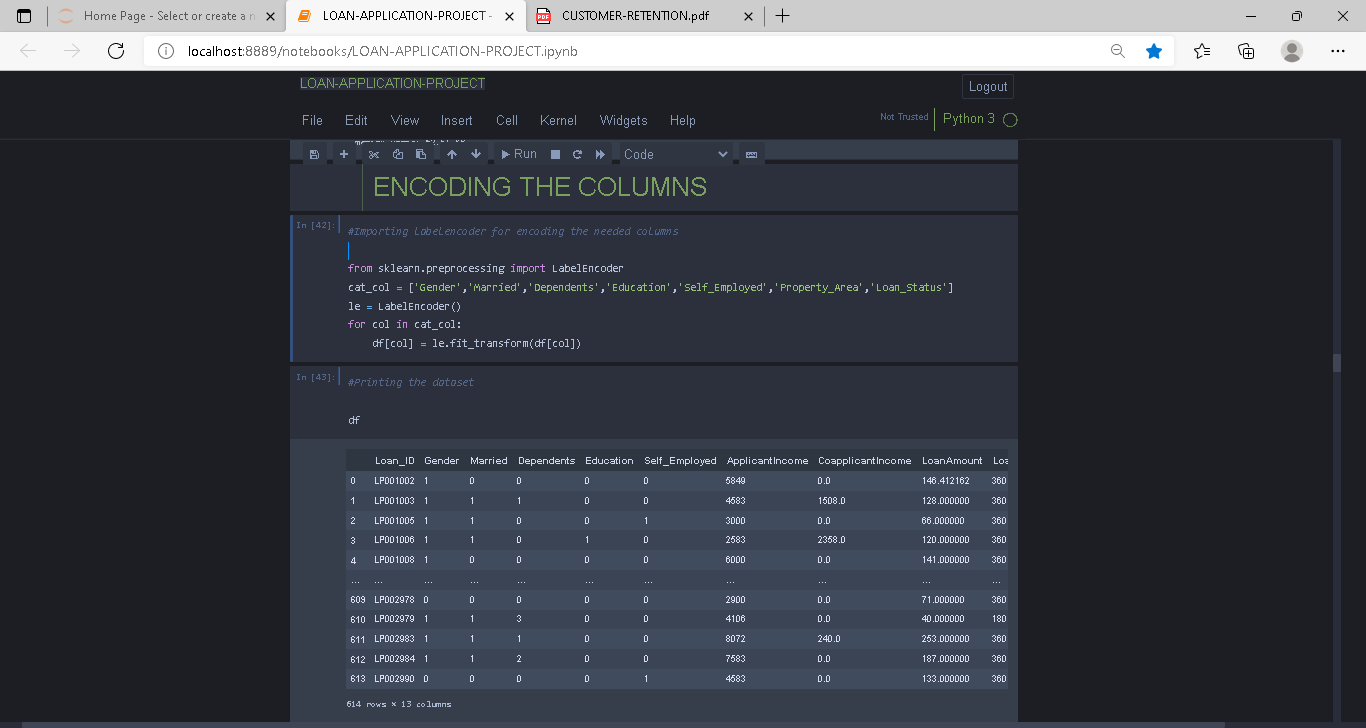




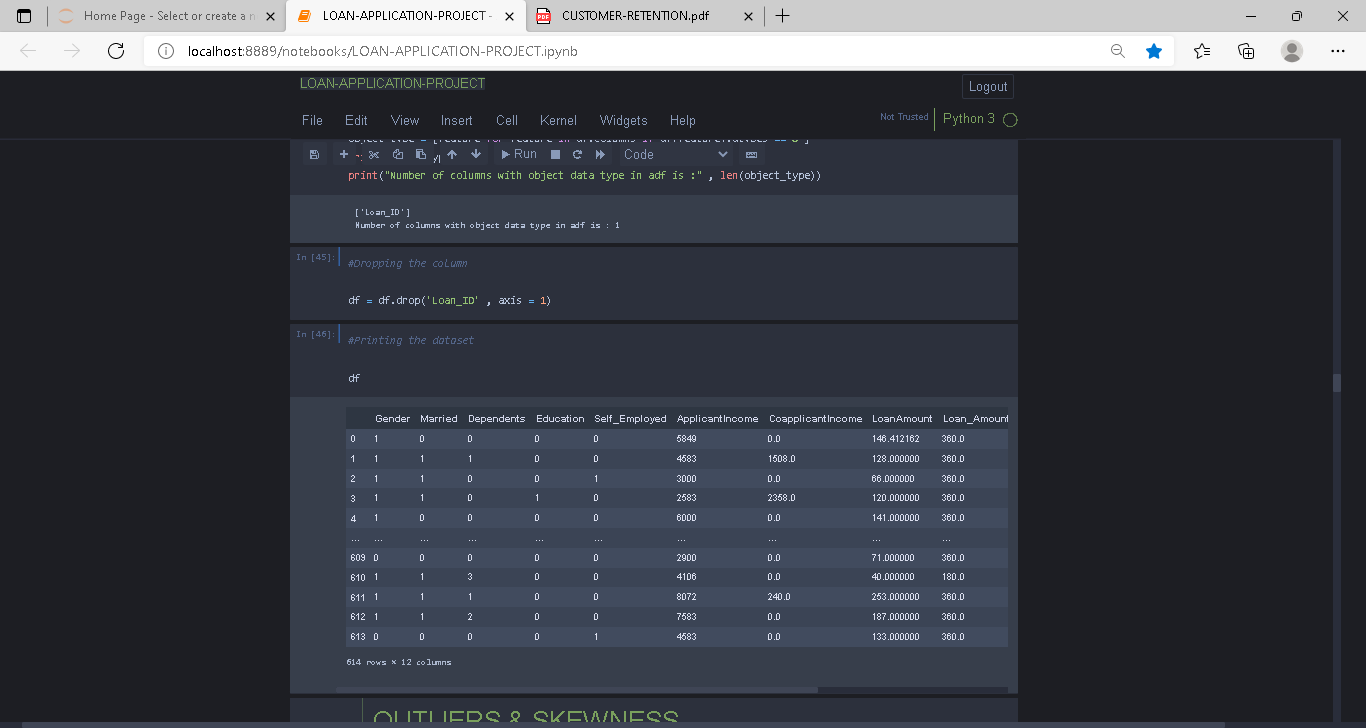


In the above 6 pictures we are just completing our first step i.e., to fill the Nan values of the column with their existing values only.

Further moving to the 2 step i.e., to encode our columns for performing analysis on the data



At the 3 step we are dropping Loan\_id column for getting a clean & sorted data



Now when we have a clean & processed data, we can perform Exploratory data analysis on that so without wasting our time let’s get into it.

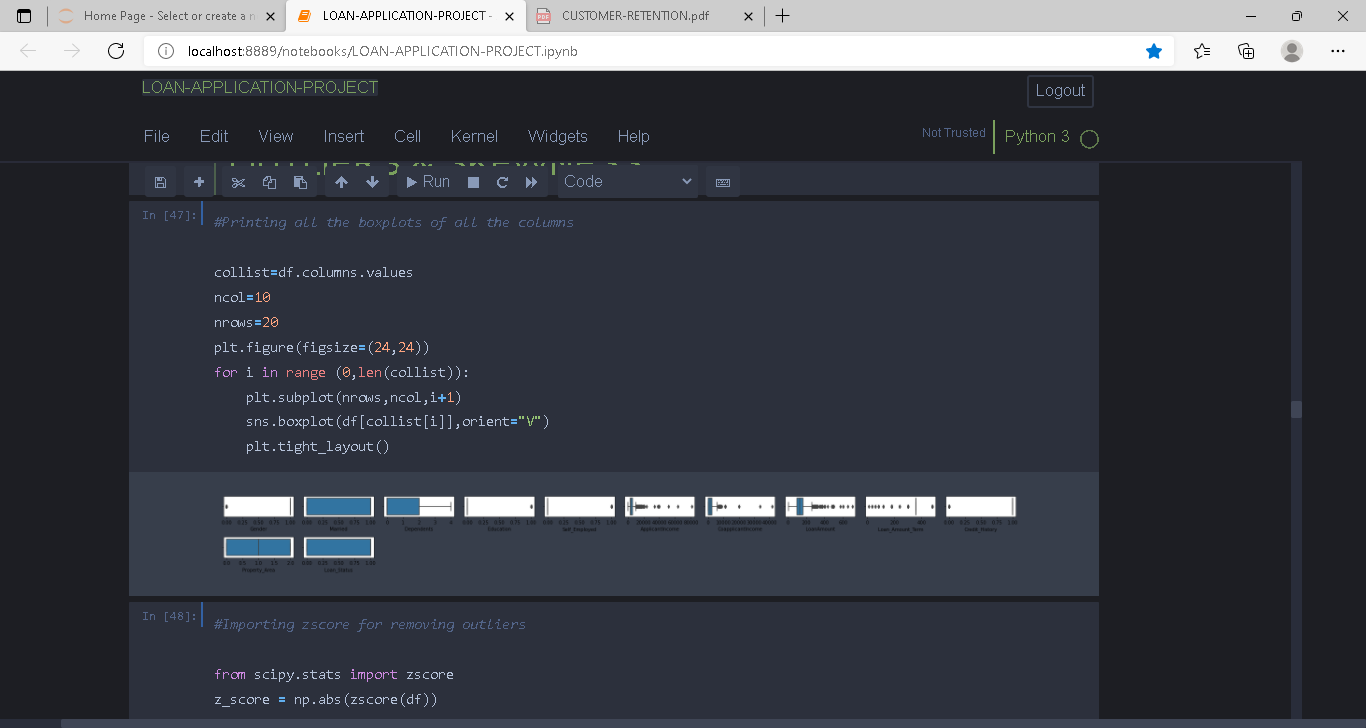
**EDA: -**

We will here see the exploratory data analysis of the data Here we will see different more methods from which we can gain more insights of the data we will do: -

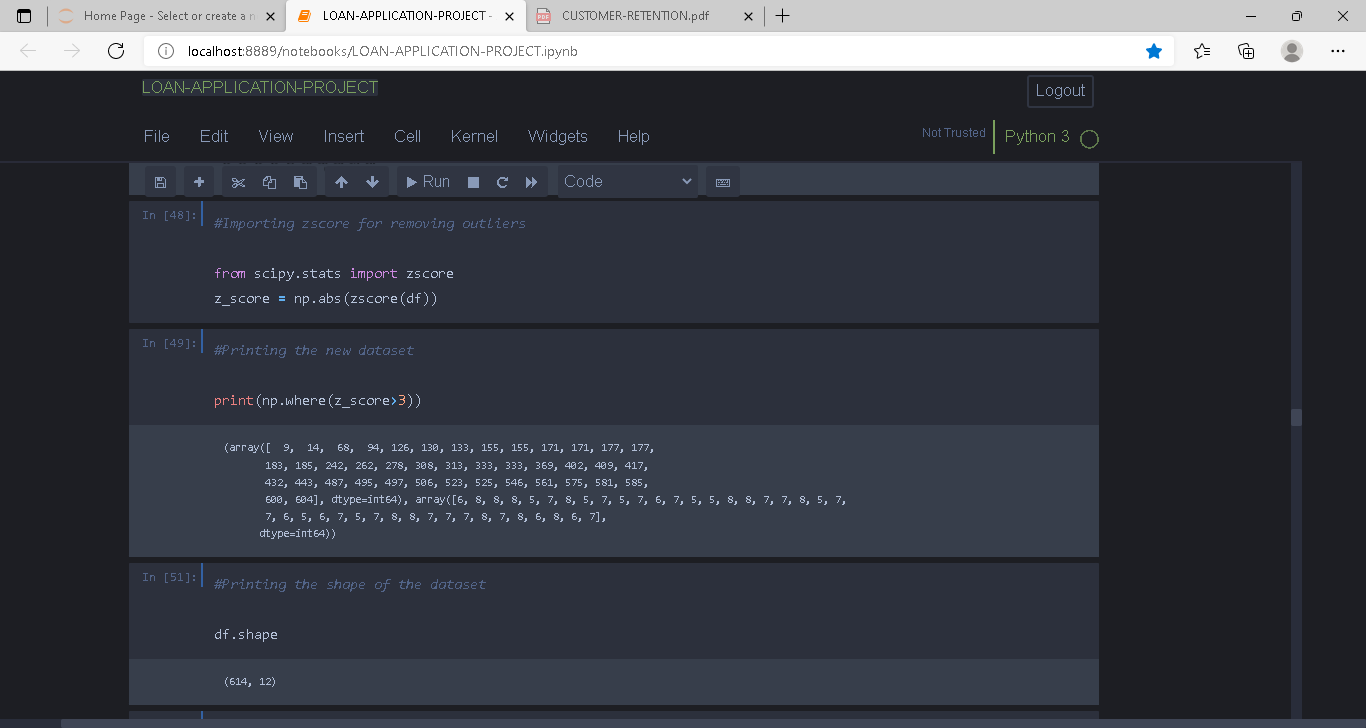
* 1. **Remove Outliers**
  2. Print the **Correlation** of all the columns
  3. We will plot the **distribution plot of all the columns for determining Skewness**
  4. Lastly, we will **Remove Skewness with power transformation method**

So, let’s get started with EDA

Firstly, we will be removing **Outliers & then Skewness**

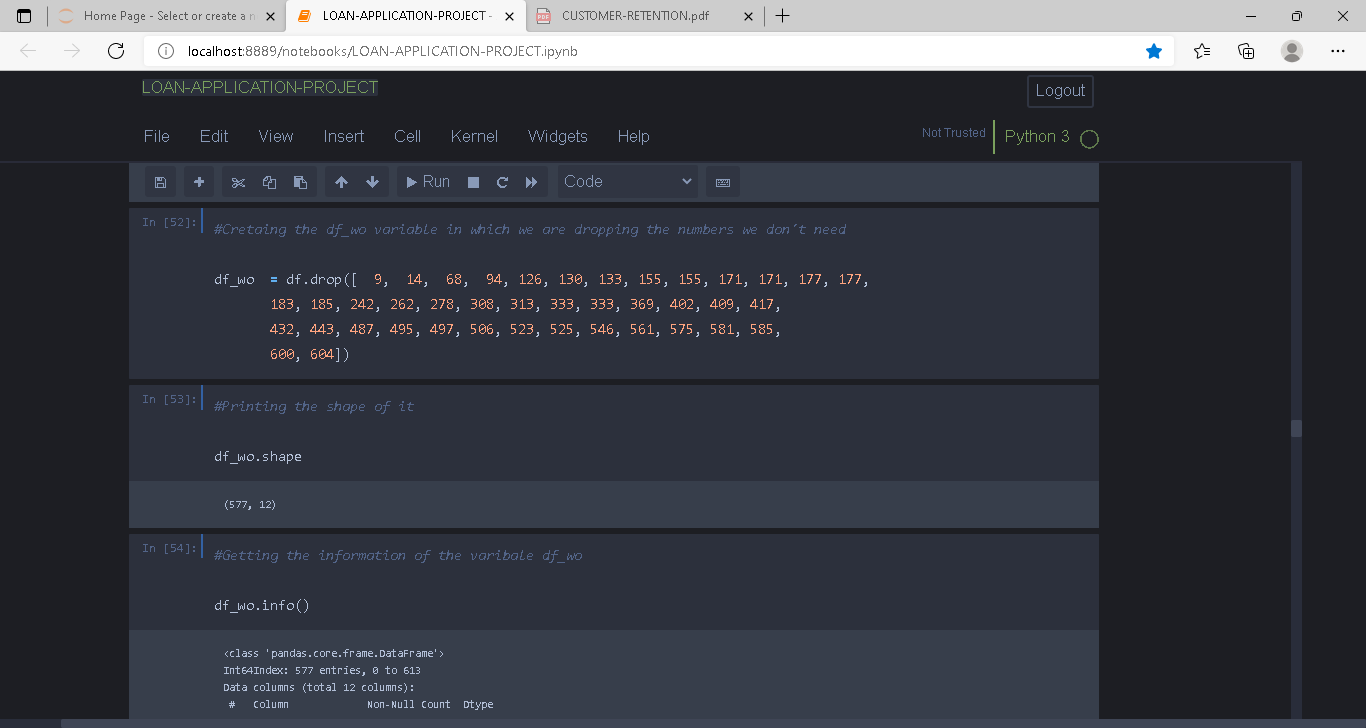


Here we are firstly plotting the boxplots of each column & then from **SciPy. Stats** we are importing zscore for **removing outliers.**

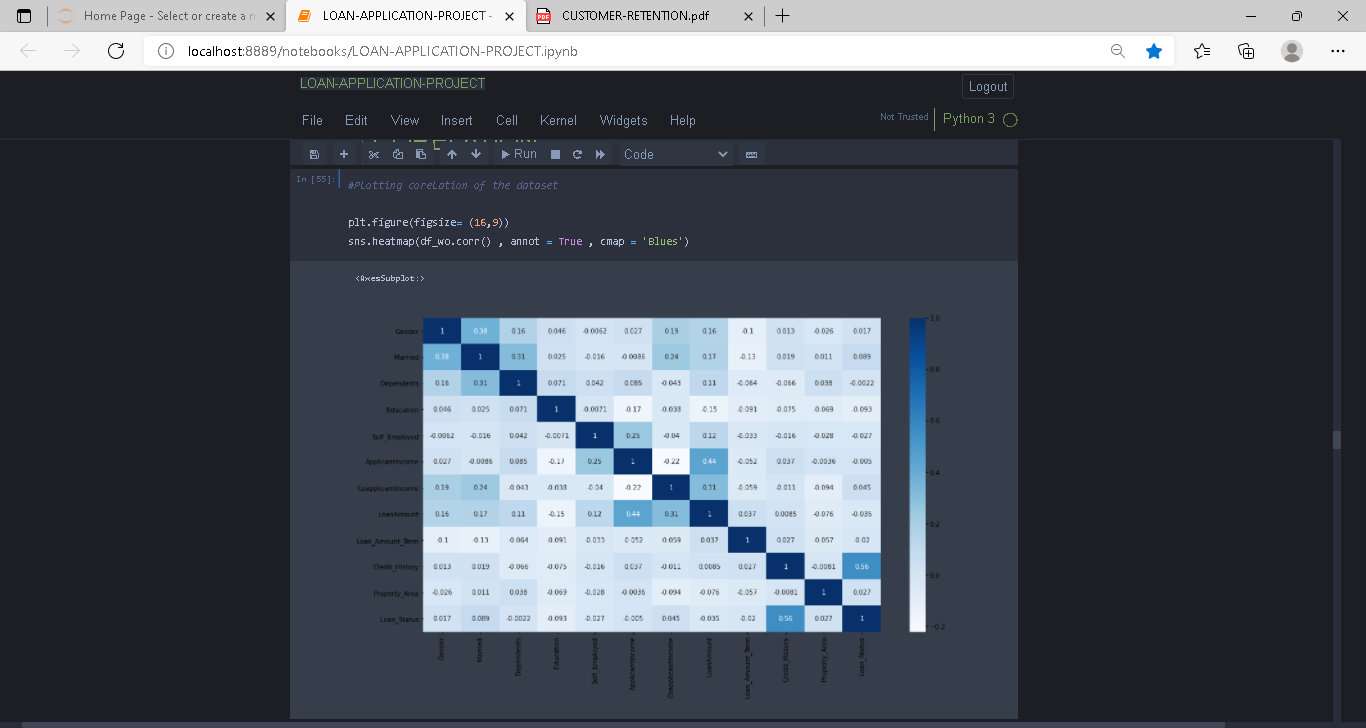


On the above the above image, we can see that with the help of zscore we are removing outliers and printing the shape of the dataset.

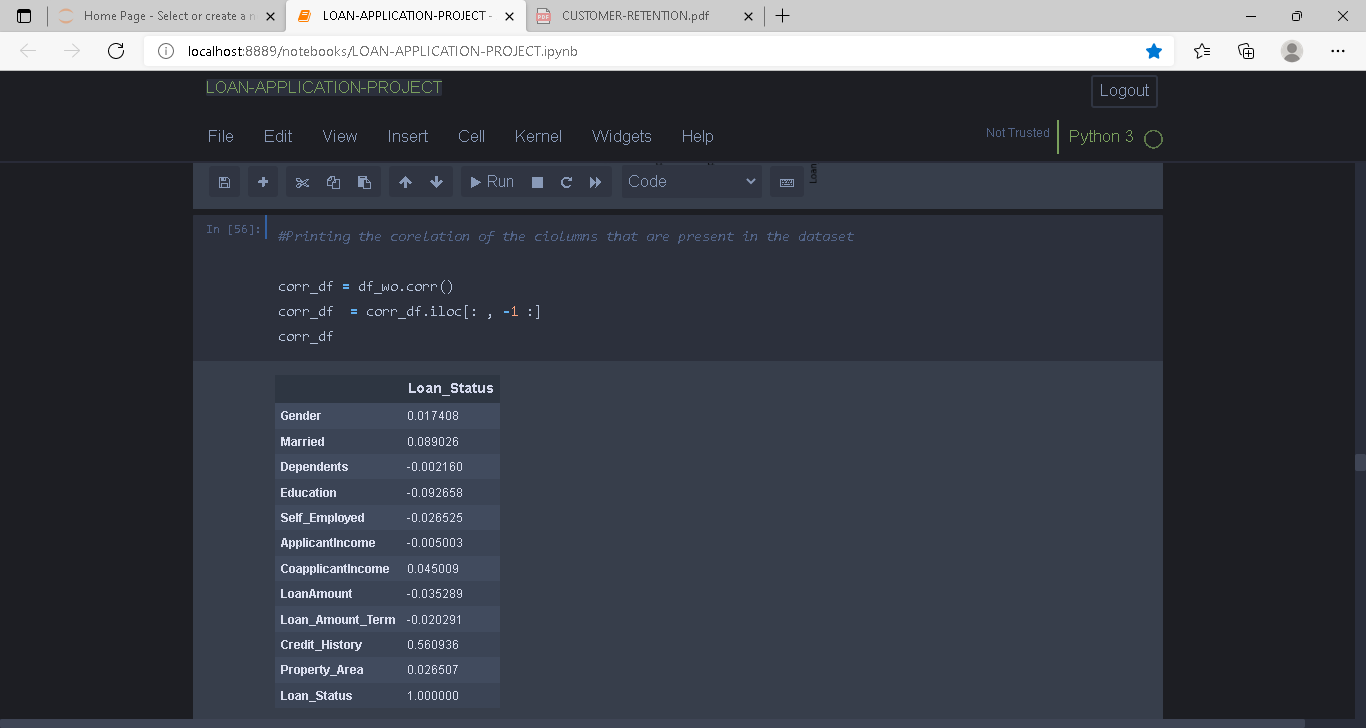
After that we are creating a new data frame with no outlier & printing the shape of the new data frame after removing the outliers & also printing the information of the new dataset



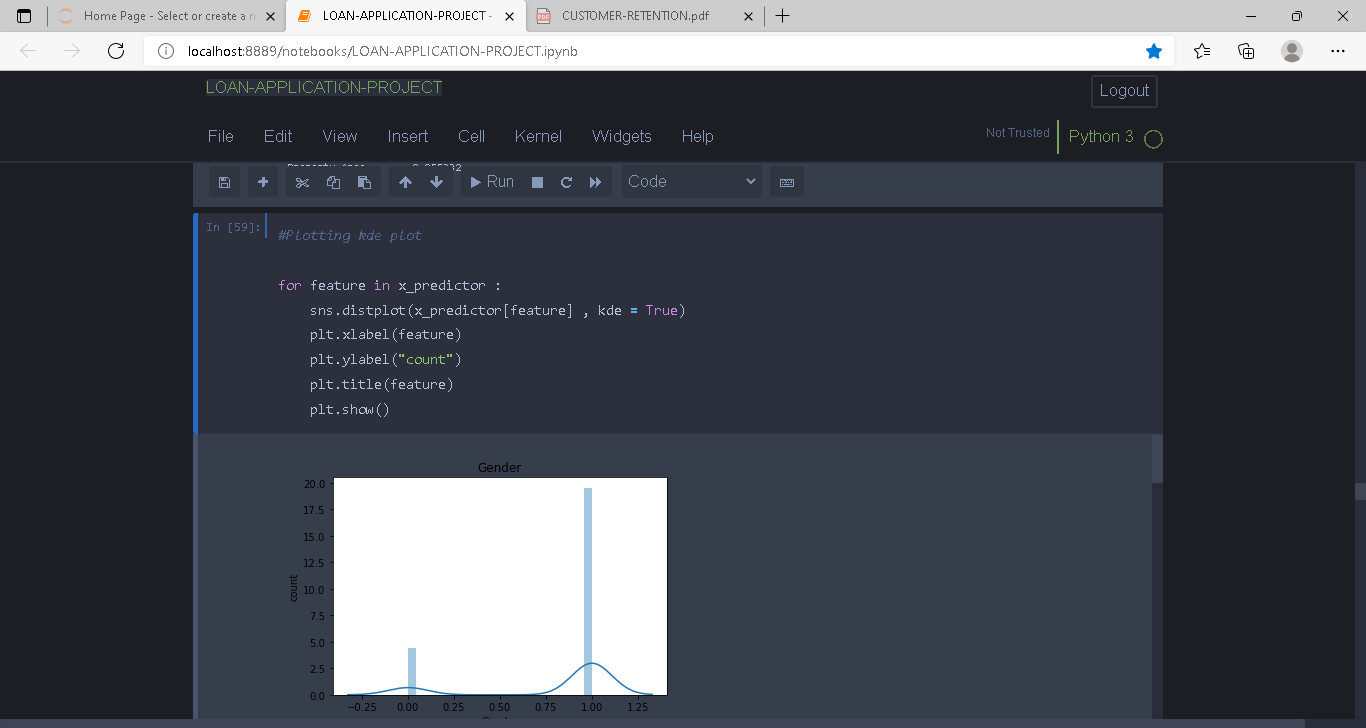
Plotting the **correlation** of all the columns



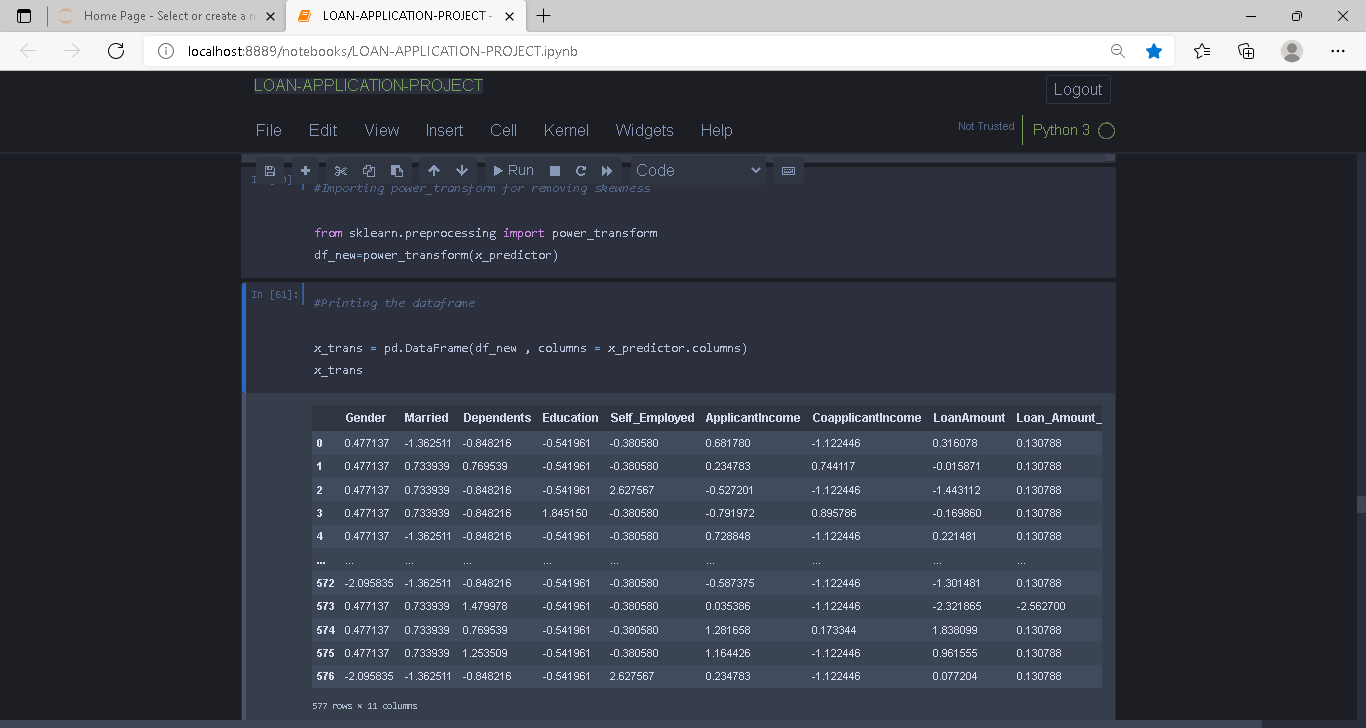
Here we have plotted all the columns with respect to each other & all the darker shades here are highly correlated



**Plotting the kde plot** for all the columns showing the distribution plot so that we can **determine the skewness present in each column**

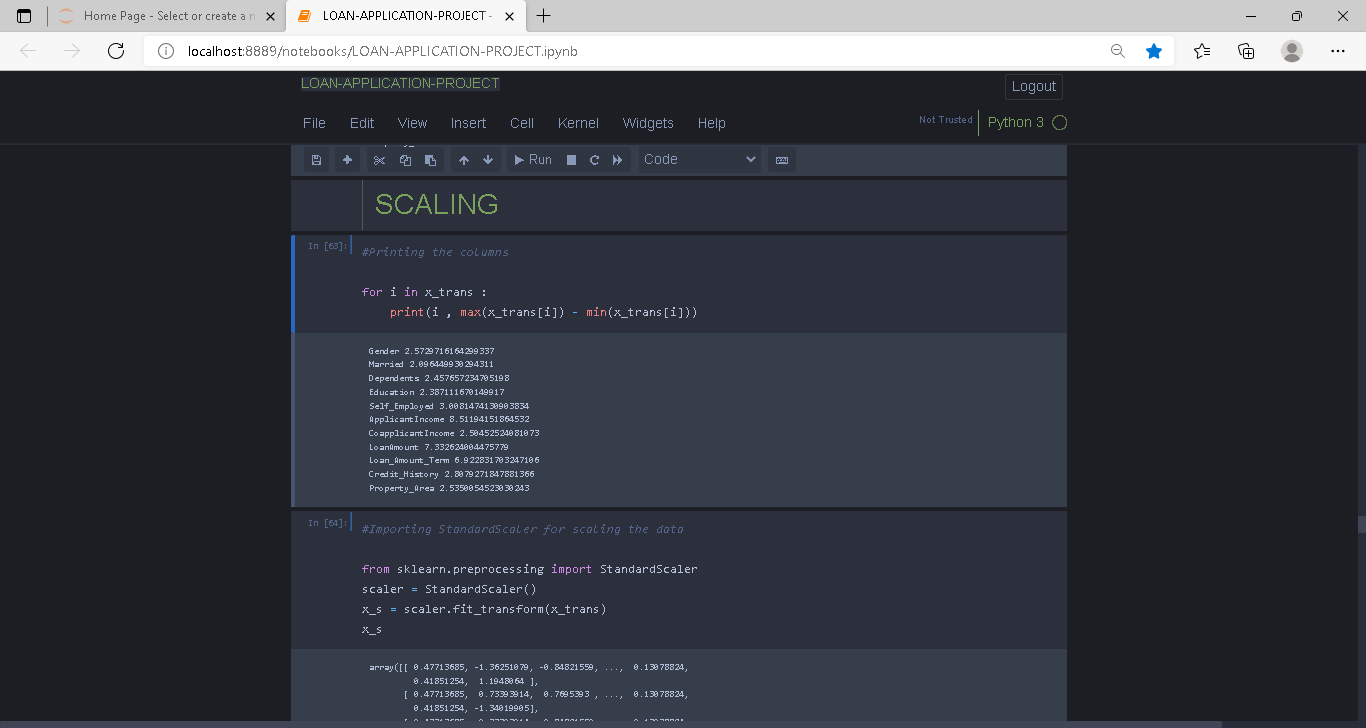


Using for loop we have printed the distribution plot of each column almost all the columns are not perfectly distributed that is why we were getting the skewness so now we will remove it with the help of **power transformation technique**.



**PREPROCESSING PIPELINE: -**

Here we will be using various pre-processing techniques from which we will preprocess the data



Here we are scaling our data to make it ready for the analysis part

After scaling we will divide our data for **Training & Testing phase**

After differentiating the x & y we will now build the models for prediction

**MODELS: -**

Here we are building 6 models & then we will be selecting the best possible model among them.

The models that we have used are: -

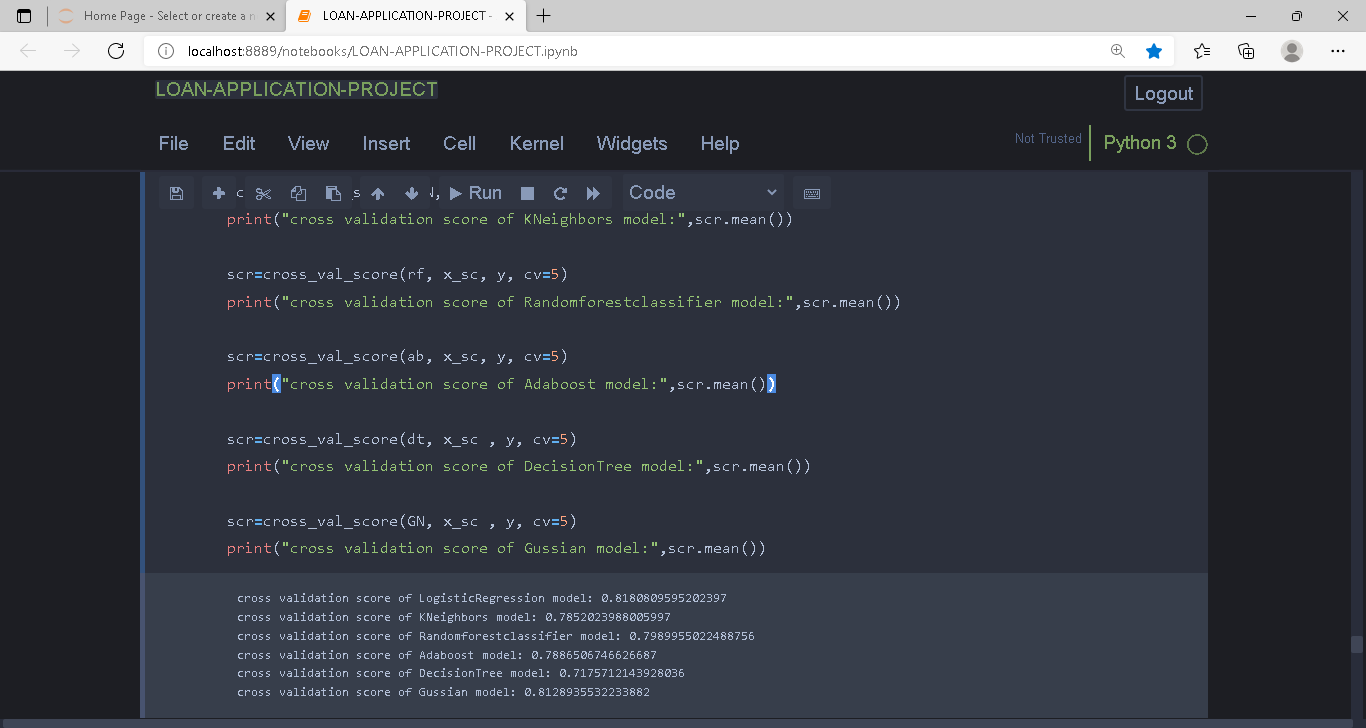
1. Logistic Regression
2. Kneighbors-Classifier
3. Random-Forest-Classifier
4. Ada-Boost-Classifier
5. Decision-Tree-Classifier
6. Gaussian-NB

|  |  |
| --- | --- |
| MODEL | ACCURACY |
| LR | 82.38 |
| KNEIGHBORS | 84.61 |
| RANDOM-FOREST | 1.0 |
| ADA-BOOST | 86.84 |
| DECISION-TREE | 1.0 |
| GAUSSIAN-NB | 82.63 |

The above table shows all the accuracy scores of each model that we have build

We can see that **Random-Forest** & **Decision-Tree** are the models that giving us 100% of accuracy

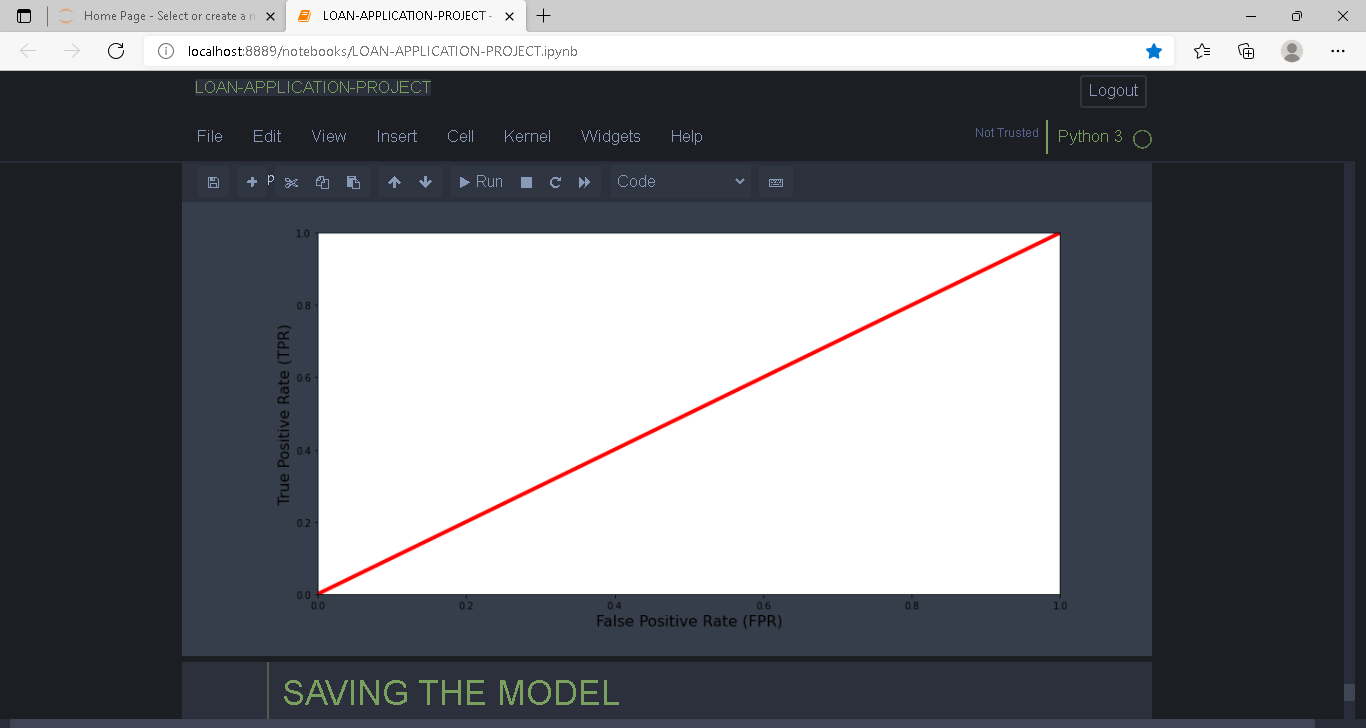
No need worry now we will perform the cross validation & we will find out which model will be the best for our dataset



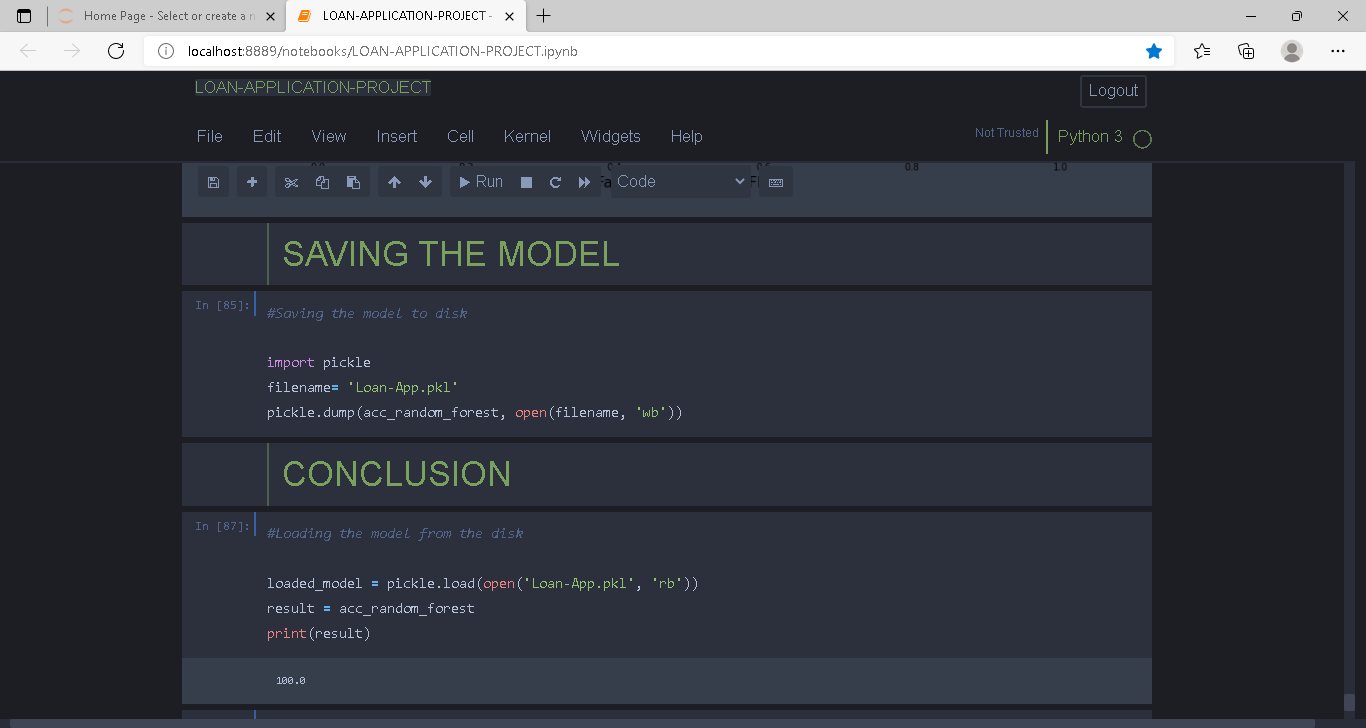
As we can see in the above image **Random-Forest** is giving us the highest cross-validation score so we will select **Random-Forest Model.**

After cross validating the model, we will send the model for hyper-parameter tunning

Now we will plot the AUC-ROC curve of the model



After this we will dump the model & will save it.



**CONCLUSION: -**

Finally, we have built a model with the 100% accuracy score that predicts that an individual who is applying for a loan that will get approved or not.

The answer to our problem statement is the individuals those who are applying for a loan will definitely get the approval

