**REPORT ON ELIGIBILITY OF LOAN PREDICTION**

**BY**

**SALISU IBRAHIM**

**DATA SCIENTIST APPLICANT INTERN**

|  |  |  |
| --- | --- | --- |
| S. NO | Topic | Page No. |
| 1 | Introduction | 3-4 |
| 2. | EDA and Business Implications | 4-9 |
| 3. | Data Cleaning and Pre=processing | 10-11 |
| 4. | Modeling Building | 11-13 |

1. **Introduction**

In this report, I'm going to predictions customers are eligible for the loan and check whether what are the missing criteria to know why customer not getting loan to make their own house.

**Problems we are to trying to solve**

* we are trying to know exact issue for loan application
* to find the ways for the applicant to apply the loan

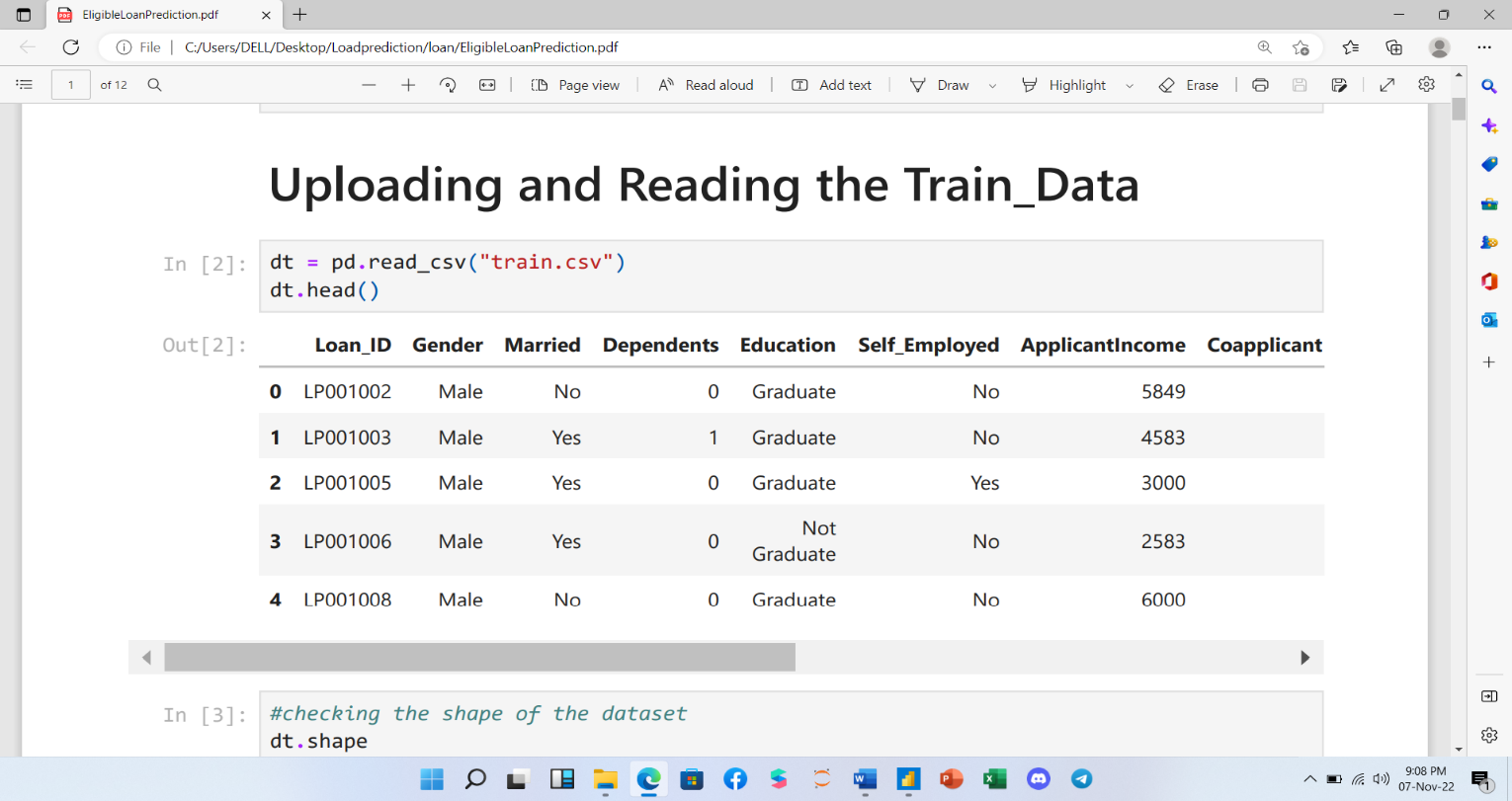
**Main Goal**

* To create an easier way for the applicant
* Find other and relevant information needed to the applicant before trying to apply for the loan
* Develop a model that can predict the eligibility of the application loan.

**About the dataset**

The dataset we have consisted of records of different applicant ID, region, loan term, loan status etc. the dataset is so useful which help by understanding the structure of the applicant, this will give us more highlight to make an analysis base on the given information on the dataset to use machine learning algorithm for developing a model to predict the eligibility of the applicant loan.

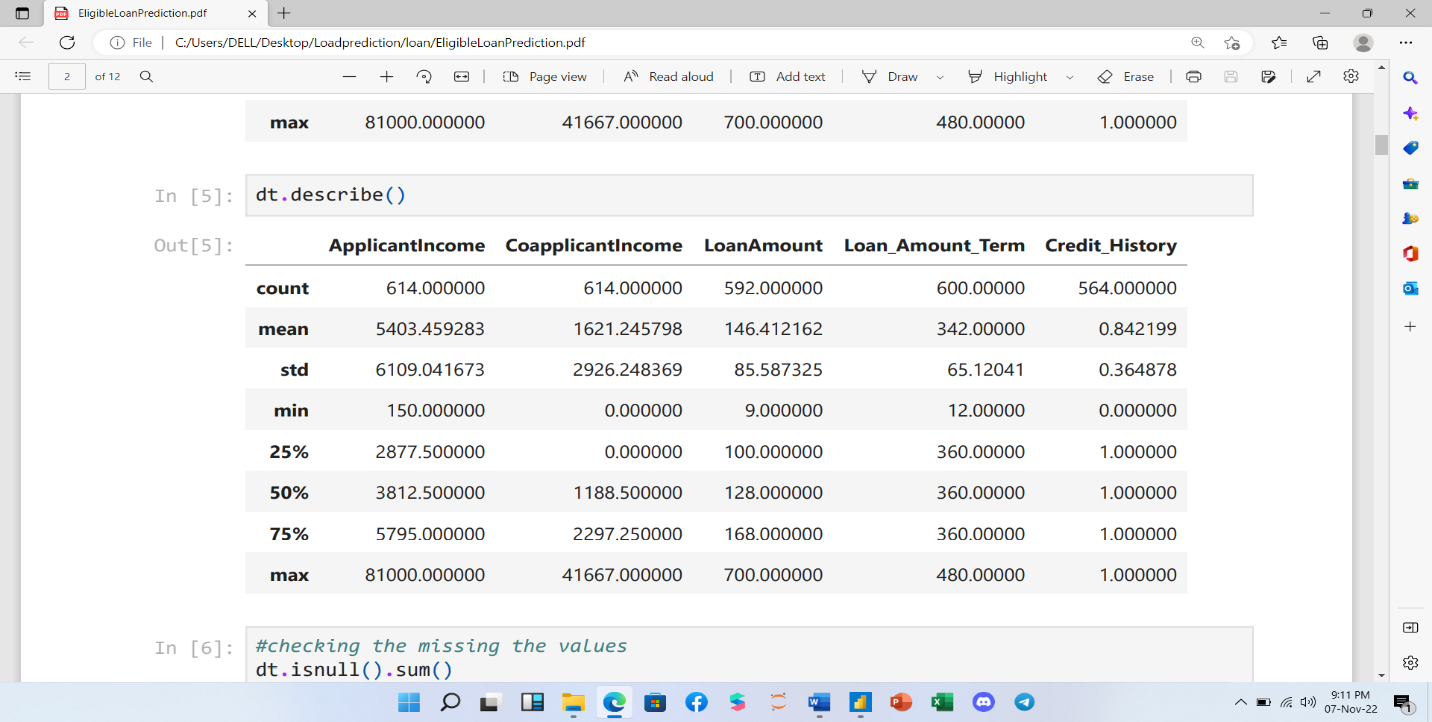
Below is how the dataset look like,



1. **EDA and Business Implication**

Exploratory data analysis (EDA) where we explore our data and grab insights from it. EDA helps us in getting knowledge in form of various plots and diagrams where we can easily understand the data and its features

**Analysis of the Data**

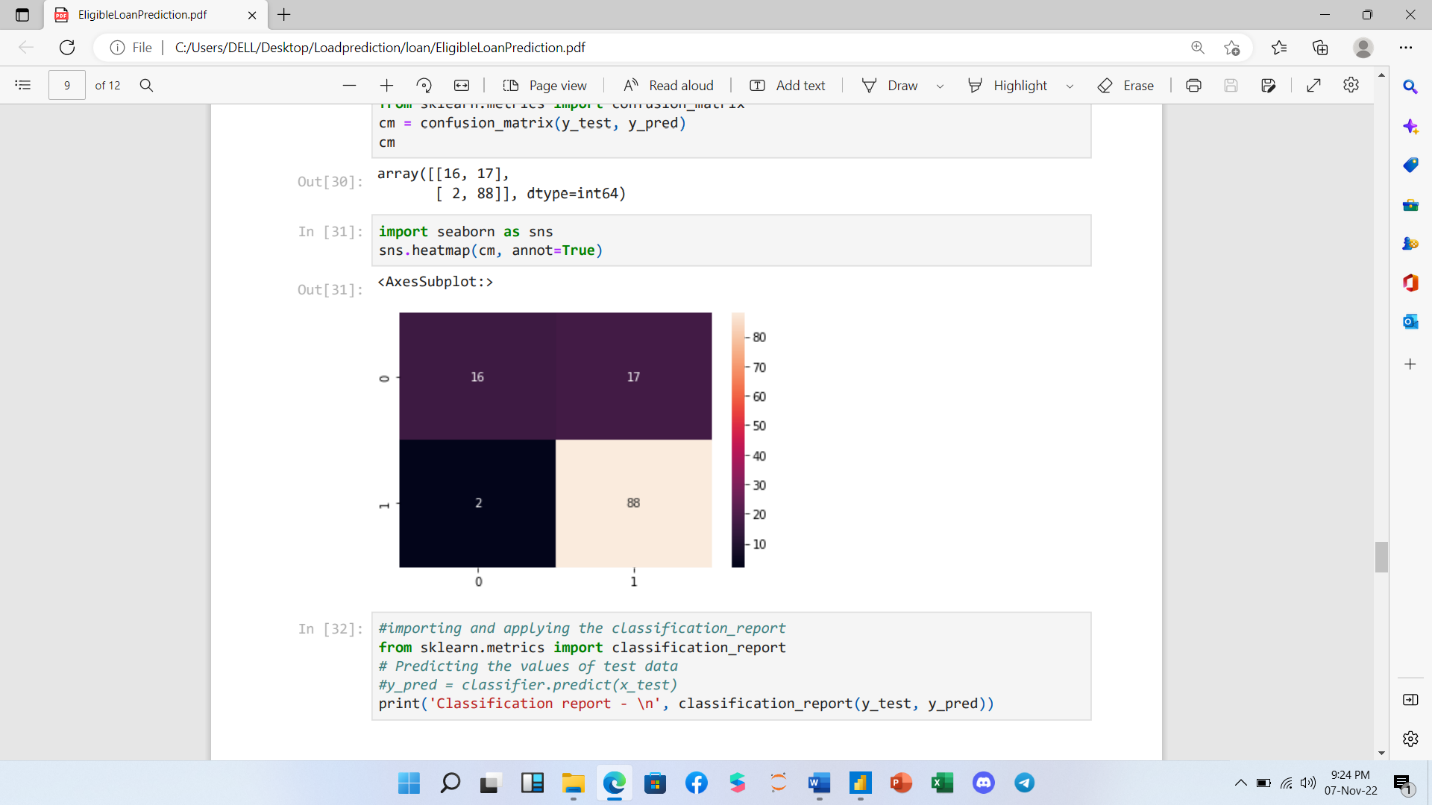
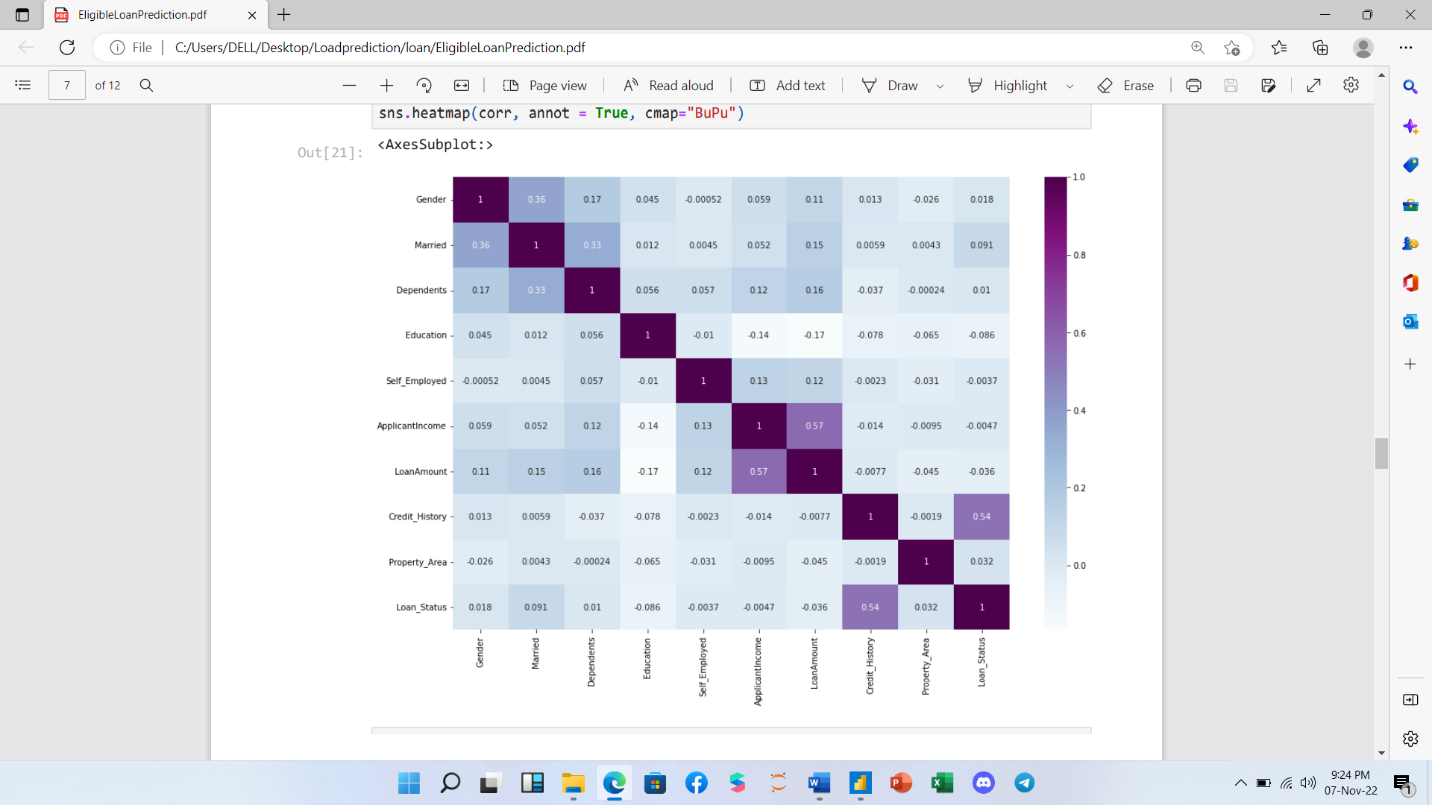
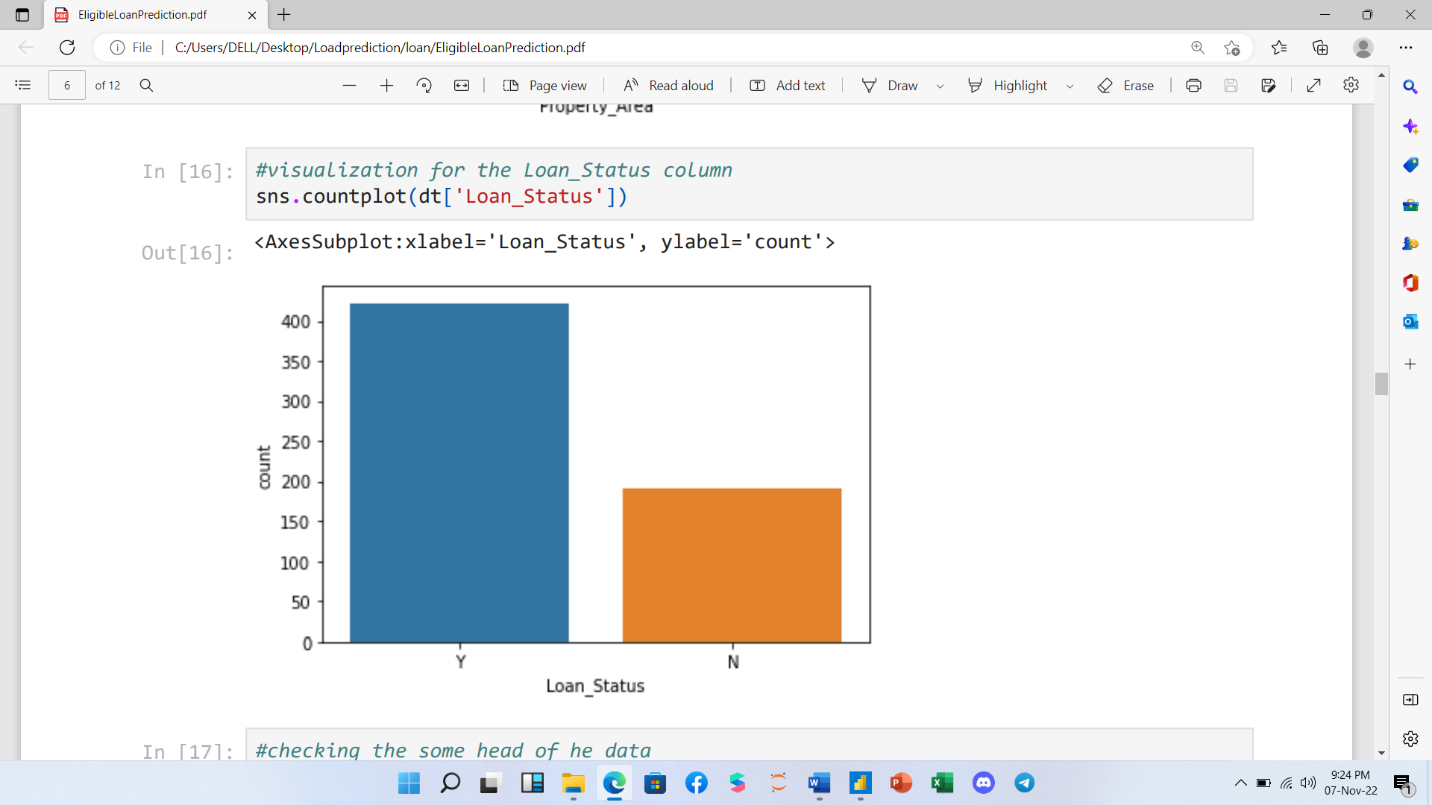
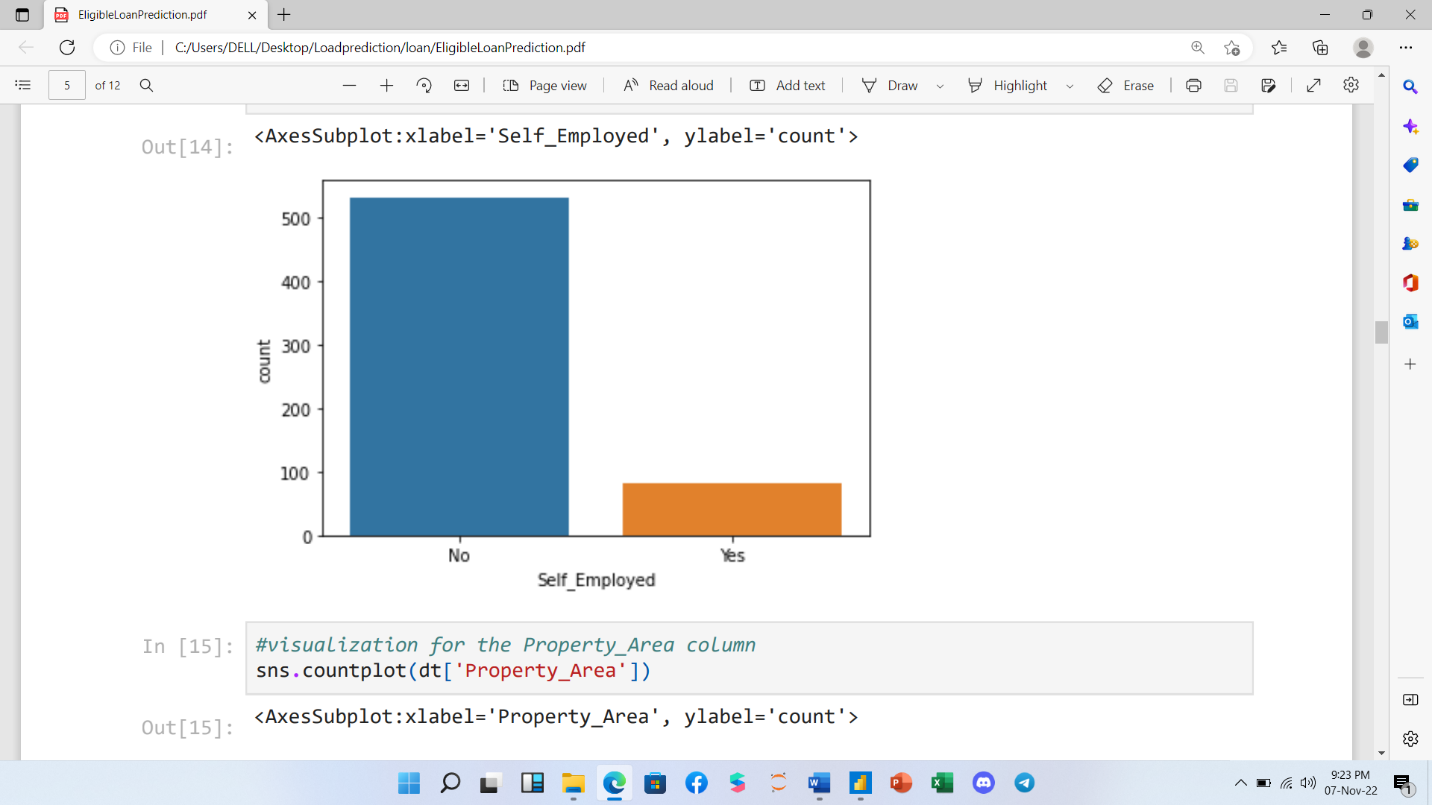
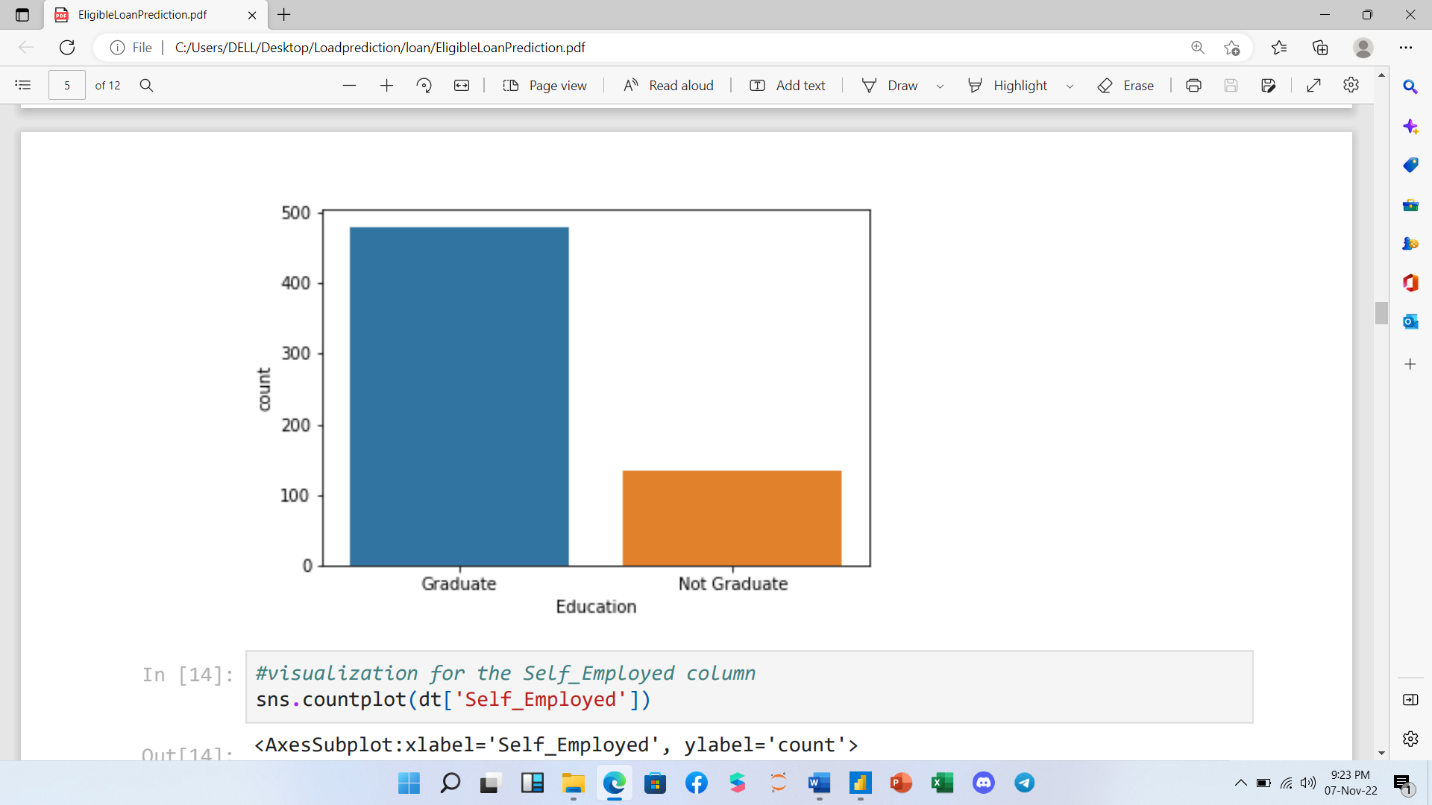
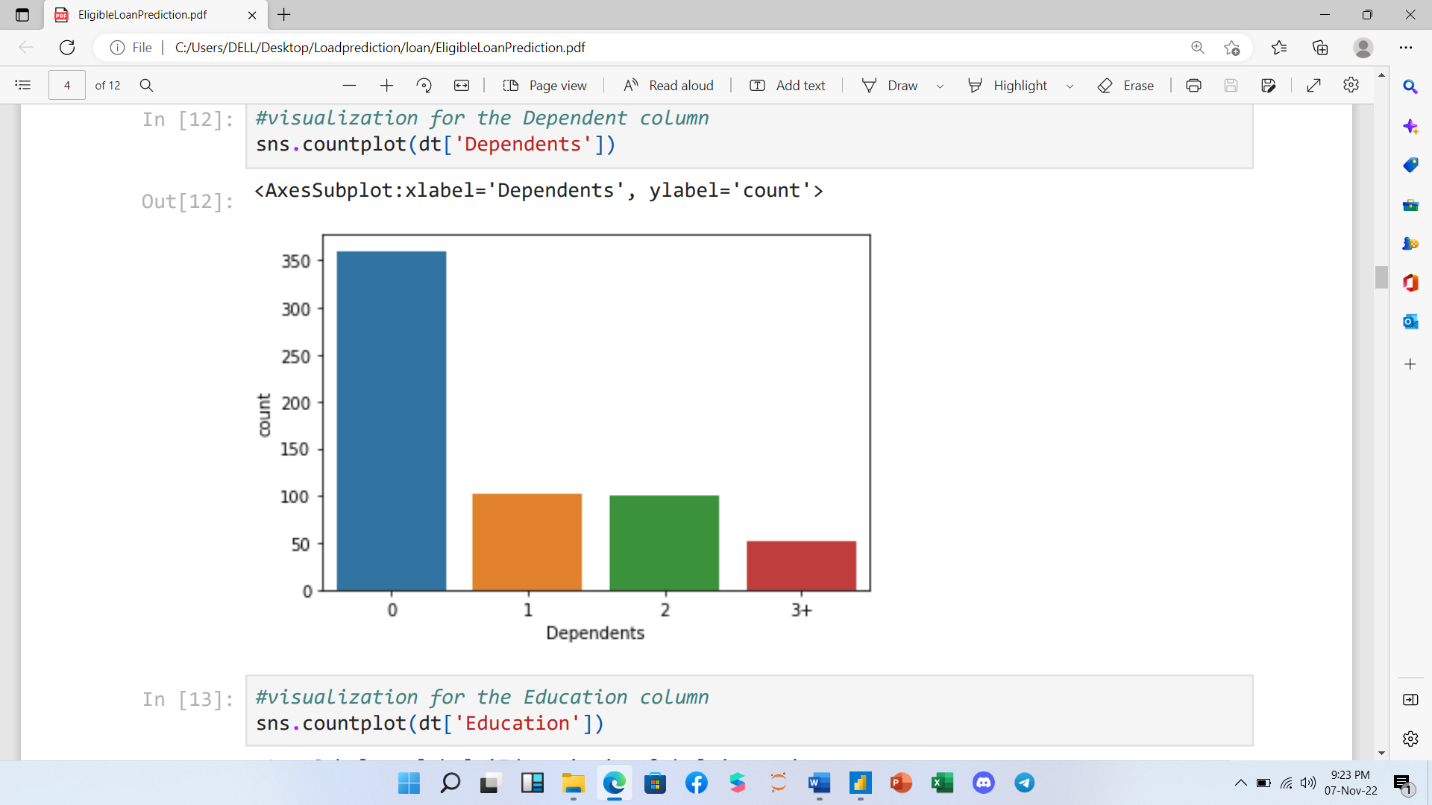
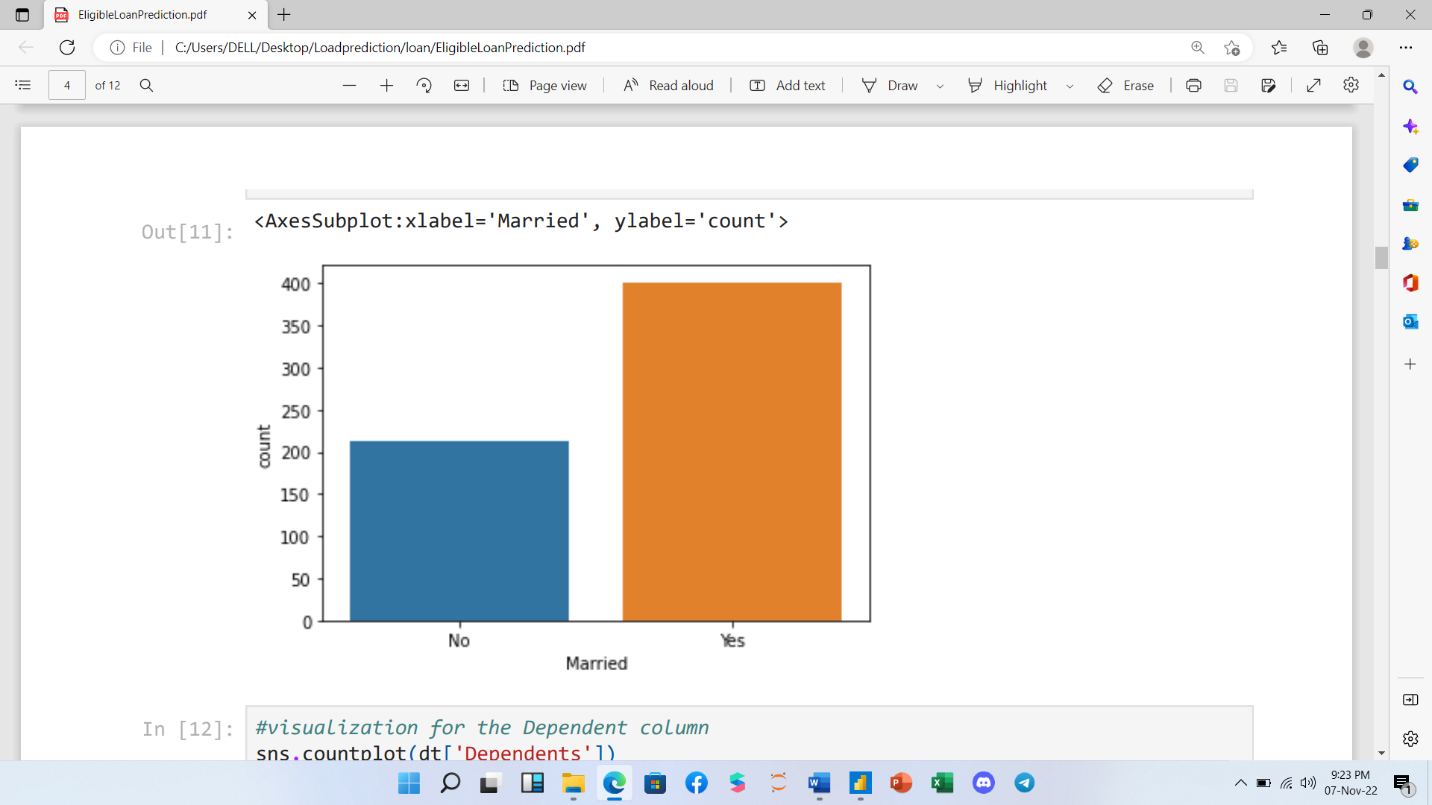


**Observation**

**Loan amount**: Loan amount is higher than the credit history for the first row first column which is 593 and credit history is 564

**Applicant Income**: Applicant income is 614 and loan amount is 592 which is eligible to get a loan

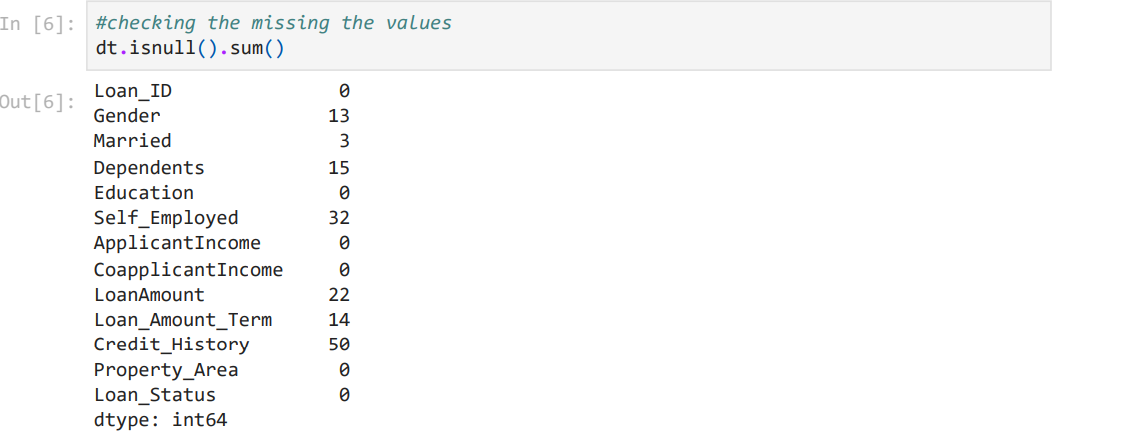
Below is the EDA visualization for the data



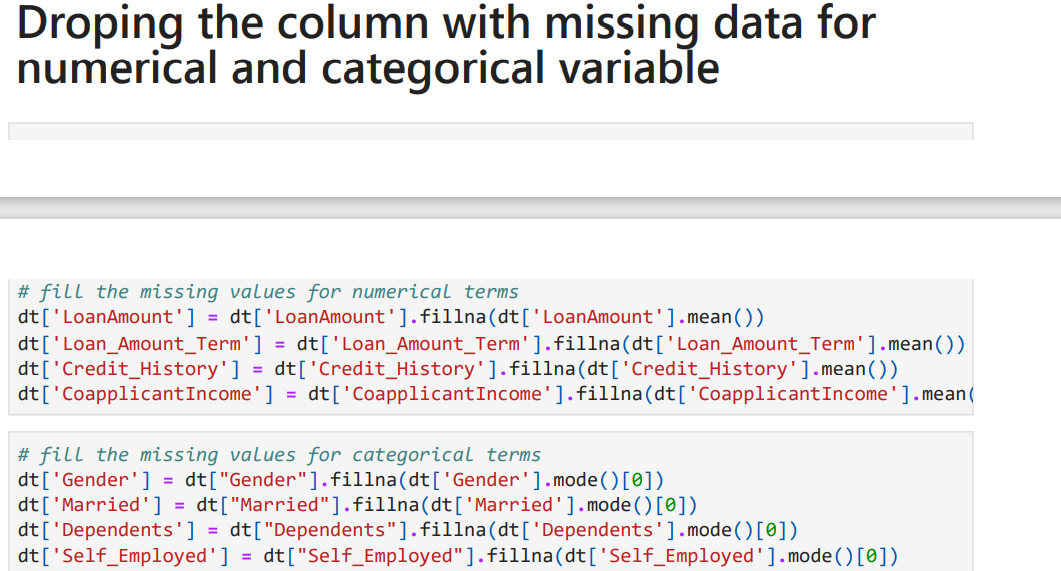
1. **Data Cleaning and Pre-Processing**

Data cleaning is the most important part of any data analysis which will help us to identify the area where we have the missing values in order to replace or remove unwanted column from our data so that to have good analysis and model prediction.

The method/approach used for identifying and treating the missing values and outlier treatment:-



To identify any missing values in our data set we have used **Pandas** pre built function **isnull()** to detect any missing values in our datasets. From the picture we can see that the column credit history has the highest number of missing values which is (50) followed by self employed with 32 missing values and married column with lowest missing values. The next step is how to handle those missing values either to manipulate it or even delete the column if we don’t want it.



Above are the picture for the step we followed for filling the missing values from our data.

1. **Modeling Building**

**Model Selection and Why?**

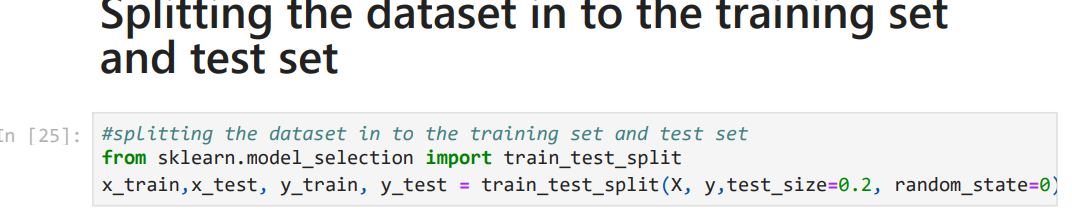
After cleaning and processing the data then comes the modeling part which includes building Machine Learning models, let’s first understand in brief what Machine Learning is?

Machine Learning is a technique that analyzes past data and tries to extract meaningful insights and patterns from them which can be further used to perform predictions in future. For example, classifying whether a tumor is benign or malignant, predicting stock prices, etc. One such application which we’re using right here is predicting house prices. Before making predictions first we need to build a model and train it using past data.

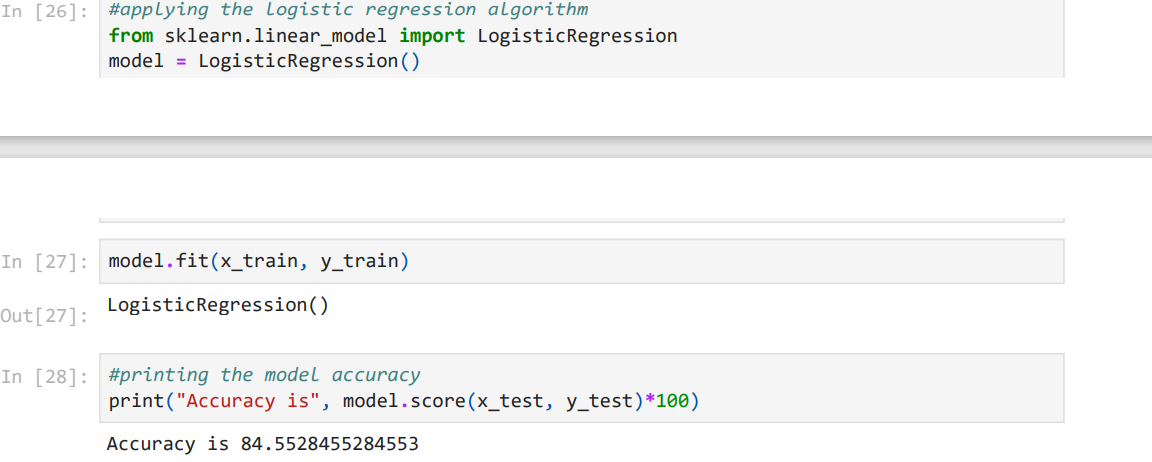
First, we need to separate the dataset into two parts: features (property attributes) and labels (prices) which is the required format for any model to be trained on.

Then the data needs to be split into 3 sets

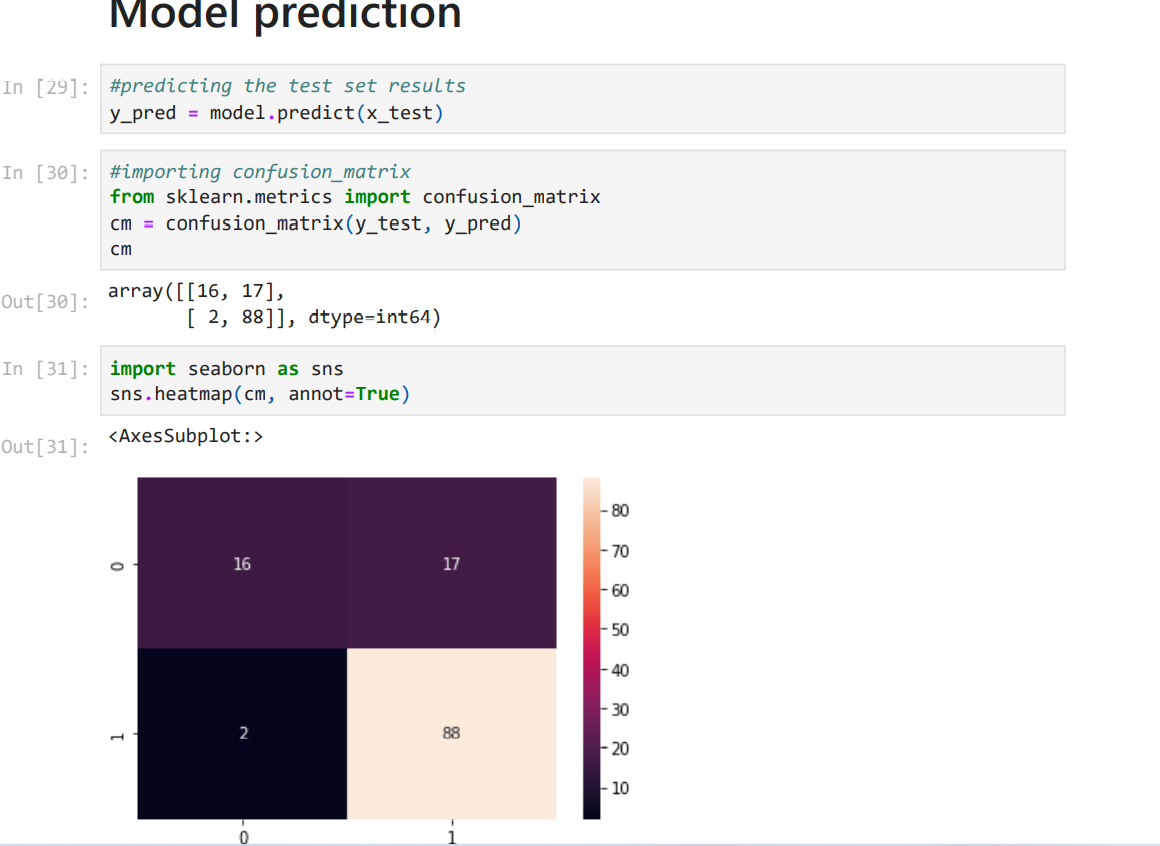
1. Training set - This will be the part of the dataset which the model will be using to train itself, the size should be at least 60-70% of the total data we’ve.
2. Validation set - This set is used for validating our model’s performance for a different set of hyperparameters. After taking out the train set, the remaining set can be split into validation and test set.
3. Testing set - To evaluate how the model is performing on the unseen data on which the model will be doing future predictions on, test set is used. It helps to understand how much error is there between actual and predicted values.



Then we’re to try applying different algorithm for training our model, then using the validation set we can determine which model to keep for making final predictions.



**Model Prediction**

****

**Final Words:**

* By comparison the logistic regression performs more than the 2 algorithm we used in this model
* We also understand that the graduate applicant are more request for loan than non-graduate.