****

**CAPSTONE PROJECT**

**INTEREST RATE PREDICTION**



**MENTOR -: Anjana Agrawal**

**SUBMITTED BY-:**

**Prem Chavan,Rupal Nikum,Atharva Shastri, Prakhar Choudhary,Ashish Bikkad,Chaitanya Thipse**

**DATASET INFORMATION**

The dataset had 164309 Rows and 14 columns in excel format. The data comprises of different features pertaining to various factors of every customer applying for loan.

**Data Dictionary**

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| Loan\_ID | A unique id for the loan. |
| Loan\_Amount\_Requested | The listed amount of the loan applied for by the borrower. |
| Length\_Employed | Employment length in years |
| Home\_Owner | The home ownership status provided by the borrower during registration. Values are: Rent, Own, Mortgage, Other. |
| Annual\_Income | The annual income provided by the borrower during registration. |
| Income\_Verified | Indicates if income was verified, not verified, or if the income source was verified |
| Purpose\_Of\_Loan | A category provided by the borrower for the loan request. |
| Debt\_To\_Income | A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income. |
| Inquiries\_Last\_6Mo | The number of inquiries by creditors during the past 6 months. |
| Months\_Since\_Deliquency | The number of months since the borrower's last delinquency. |
| Number\_Open\_Accounts | The number of open credit lines in the borrower's credit file. |
| Total\_Accounts | The total number of credit lines currently in the borrower's credit file |
| Gender | Gender |
| Interest\_Rate | Target Variable: Interest Rate category (1/2/3) of the loan application |

- Variable categorization (count of numeric and categorical)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Numerical** | **Categorical** |  | **Missing Values** | |
|  |  |  |  | |
| Loan\_Amount\_Requested | Home\_Owner |  | Length\_Employed | 7371 |
| Length\_Employed | Income\_Verified |  | Home\_Owner | 25349 |
| Annual\_Income | Purpose\_Of\_Loan |  | Annual\_Income | 25102 |
| Debt\_To\_Income | Gender |  | Months\_Since\_Deliquency | 88379 |
| Inquiries\_Last\_6Mo | Interest\_Rate |  |  |  |
| Months\_Since\_Deliquency |  |  |  | |
| Number\_Open\_Accounts |  |  |  | |
| Total\_Accounts |  |  |  | |
| **Total 8 Features** | **Total 5 Features** |  | **6.36% of total Data values** | |

**DATA CLEANING**

While conducting EDA the following discrepancies were found in the dataset:

1. Loan\_ID feature Is a unique id column that does don’t provide any type of insight not would help the machine learning model in prediction.
2. Loan\_Amount\_Requested is a feature that should be numerical but is object because of the commas present between the digits.
3. Length\_Employed also has special characters and strings instead of numerical values.

**FEATURE ENGINEERING**

Creating new features that might help our model predict more accurately

1. Accounts closed to Open accounts Ratio: Combining Total accounts and no. of open accounts to form a new feature which tells the no of accounts that have been closed by a client.
2. Assets or Liability: Categorise loans according to their purpose by segregating them according to whether the purpose of loan will help earn money in the future or not.
3. Financial Growth score: A new category that combines annual income and employment length of client, giving an idea of his/her financial growth over the years. (we will abstain from introducing this column as our annual income feature has lot of missing values, which will be imputed by us and this new feature might be biased towards our imputed value)

**MISSING VALUE IMPUTATION**

The dataset has a total of 6.36 % missing values. Which is not a suitable number to drop them, which will result in data loss. Hence we use standard null value imputation technique of fillna.

1. Home ownership null values replaced with new value created ‘NoHouse’.

2. Employment length null values impute with median.

### 3. Annual income null values filled we built the LinearRegression model and predict the Null value and imputed them

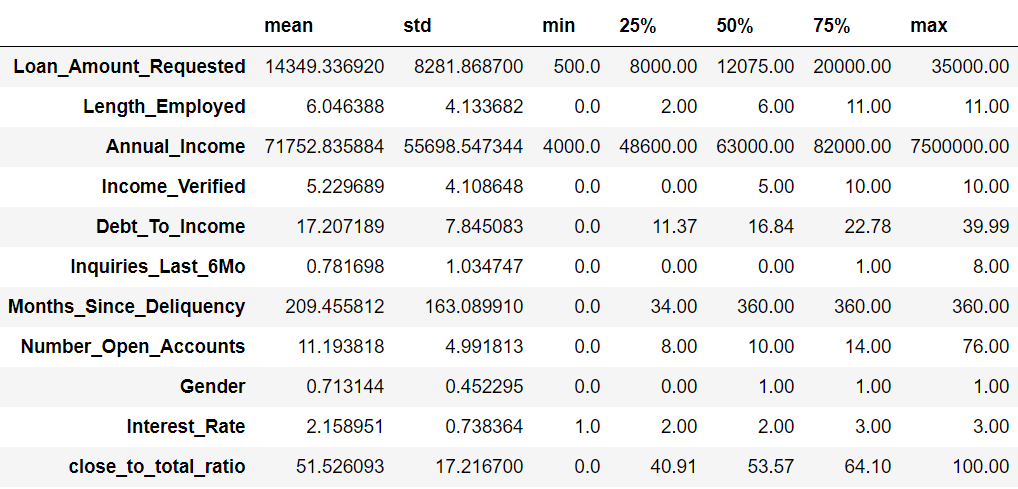
4. Months since Delinquency: as we have 53.55% Null value in this so we drop this variable

**ENCODING**

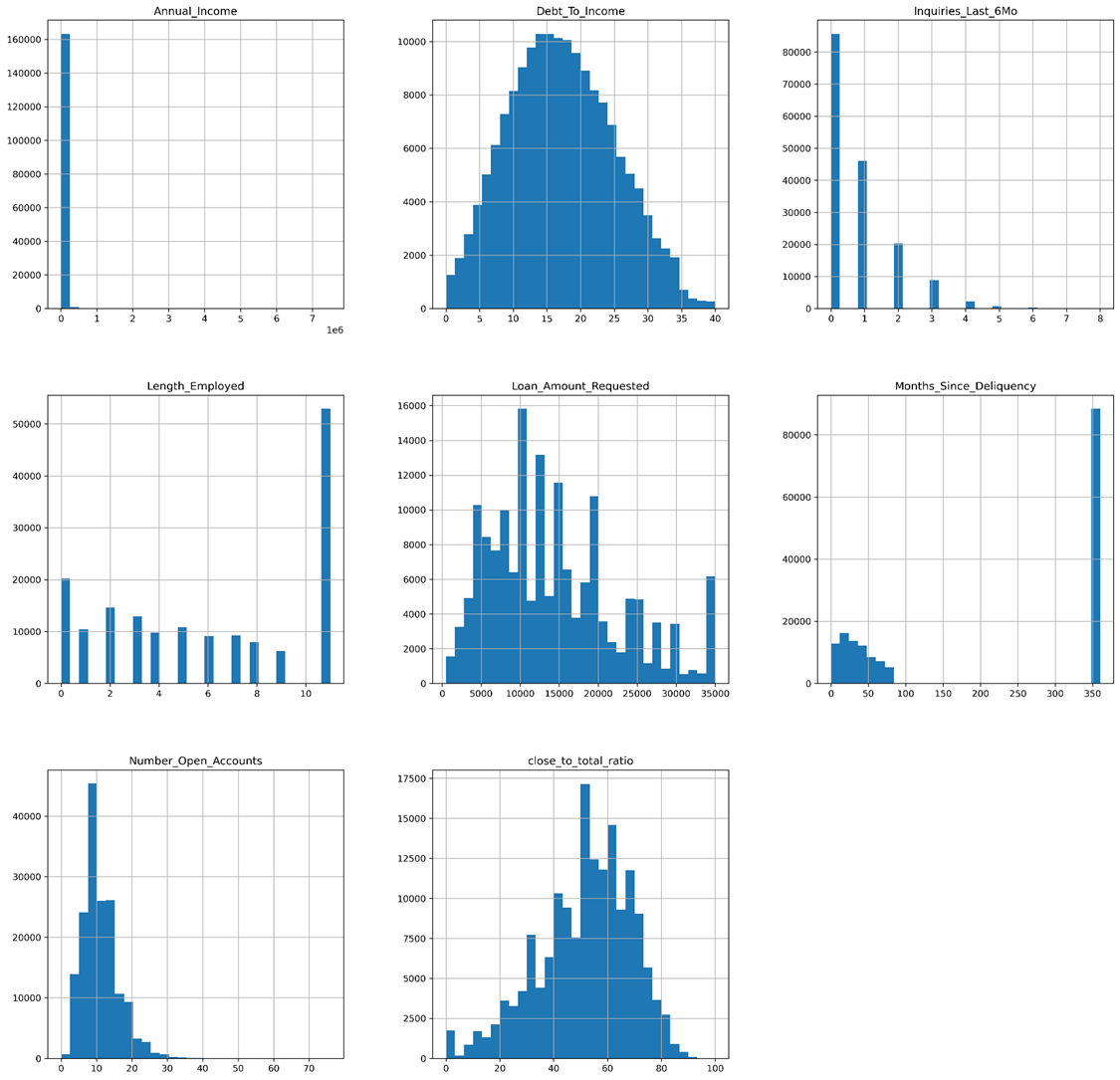
**One Hot Encoding:** The columns Assets\_Liability, Home\_Owner, Purpose\_Of\_Loan and Gender will be one hot encoded using the pd.get\_dummies.

**DATA DISTRIBUTION**

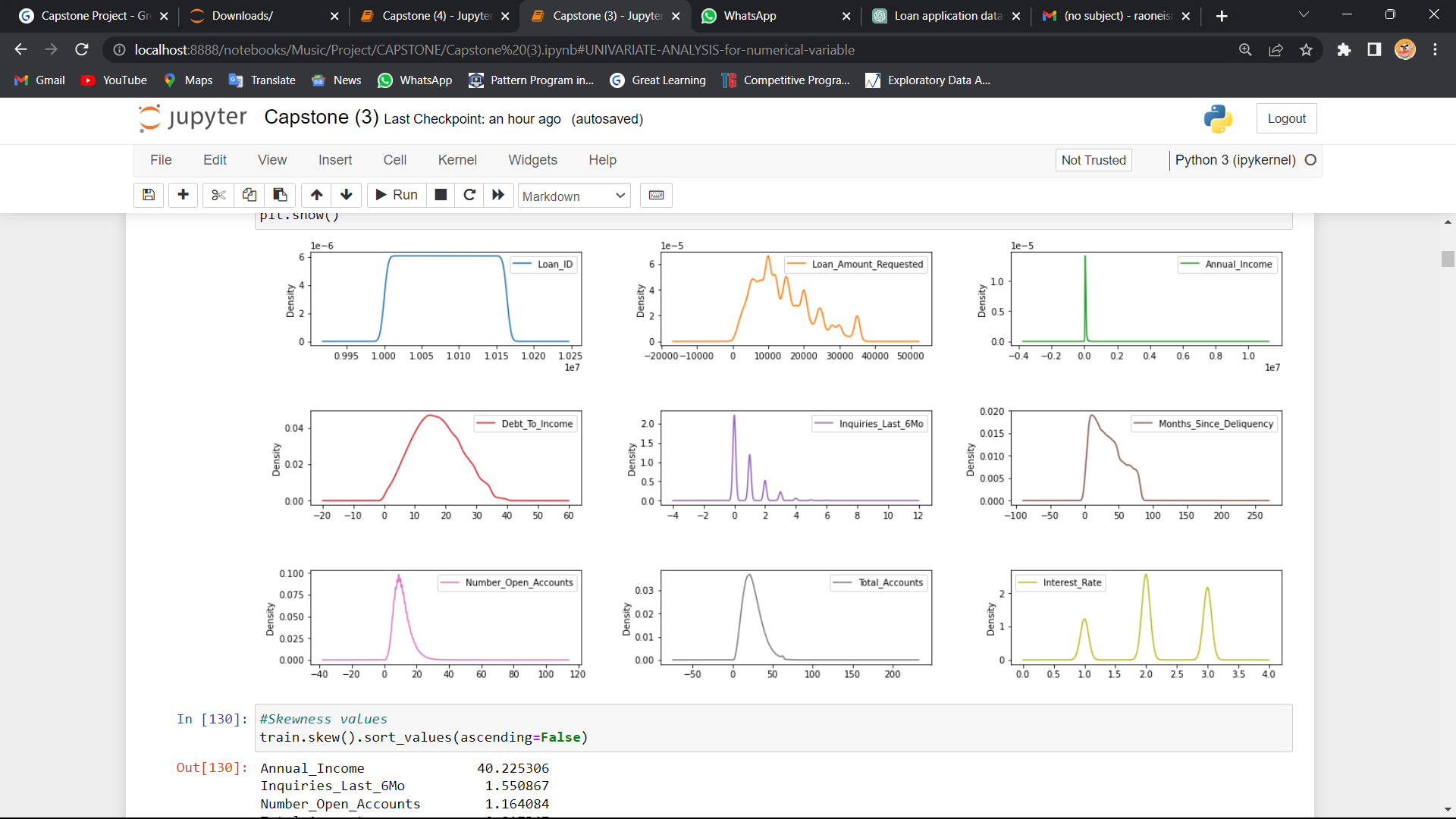
Using Data.describe() to create a 5 point summary of the data to get a better understanding of the numerical features in the dataset.



Using histograms plot in python to visualise data distribution.

****

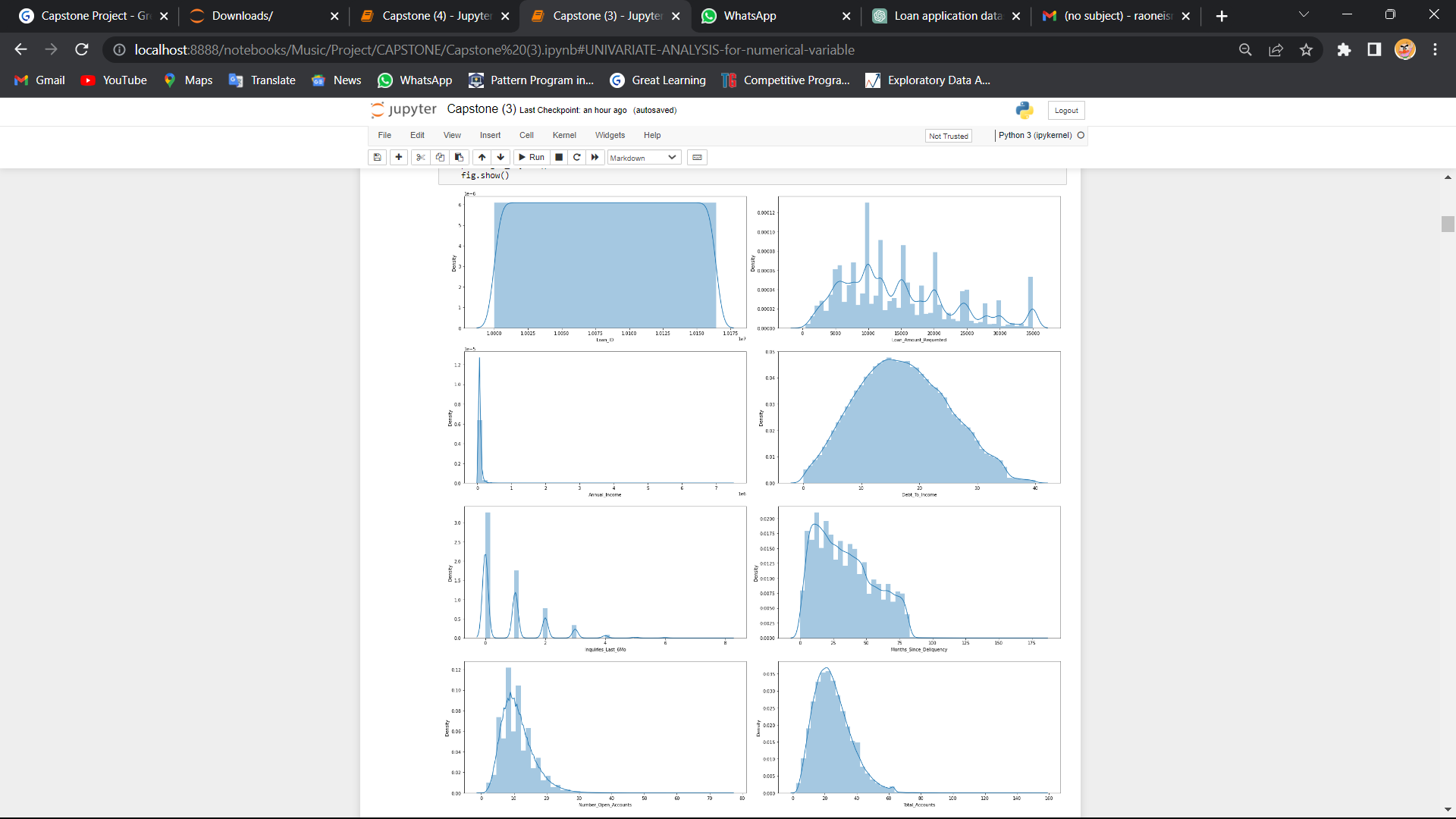
# UNIVARIATE ANALYSIS

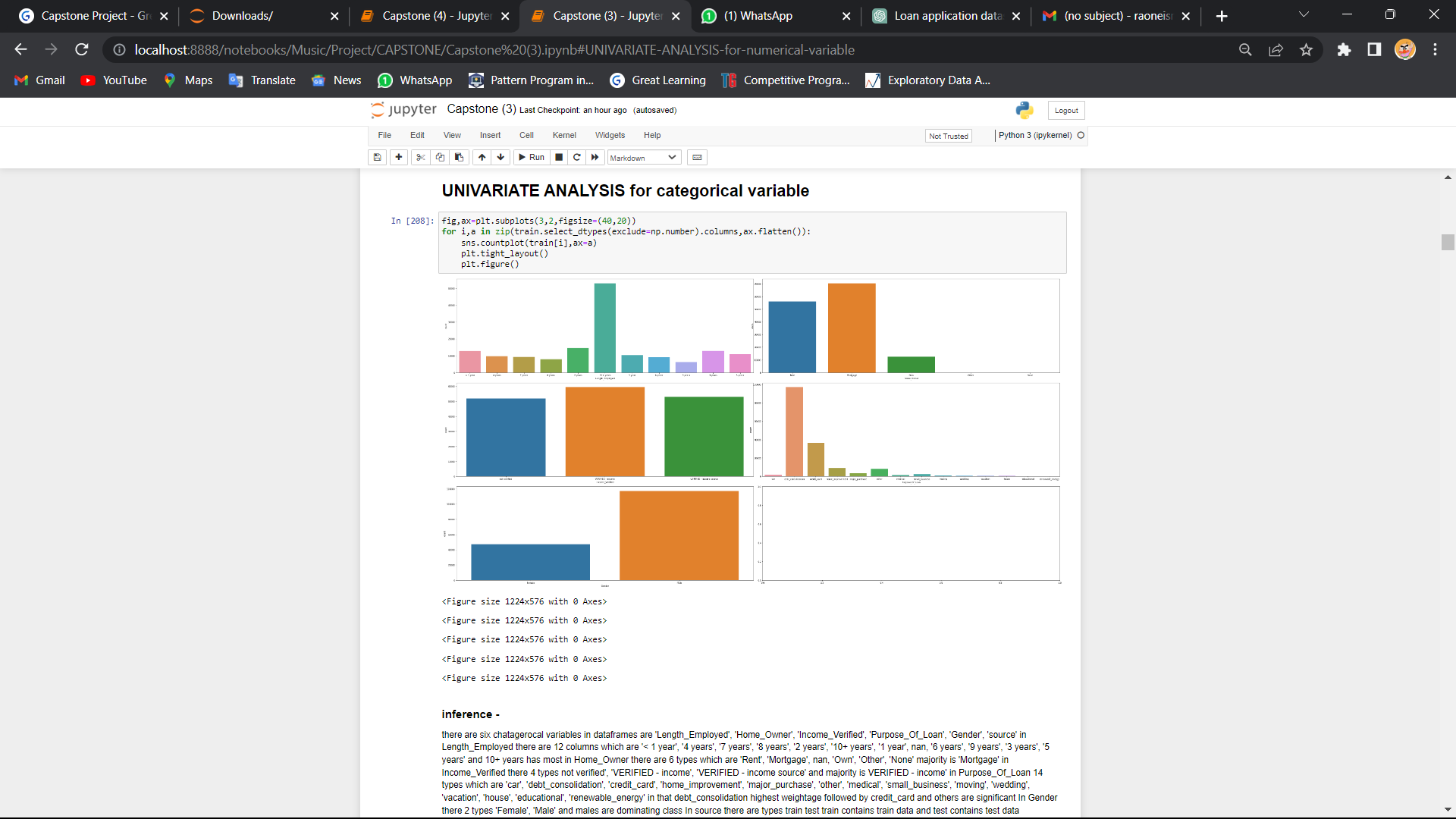
****

### Inference -

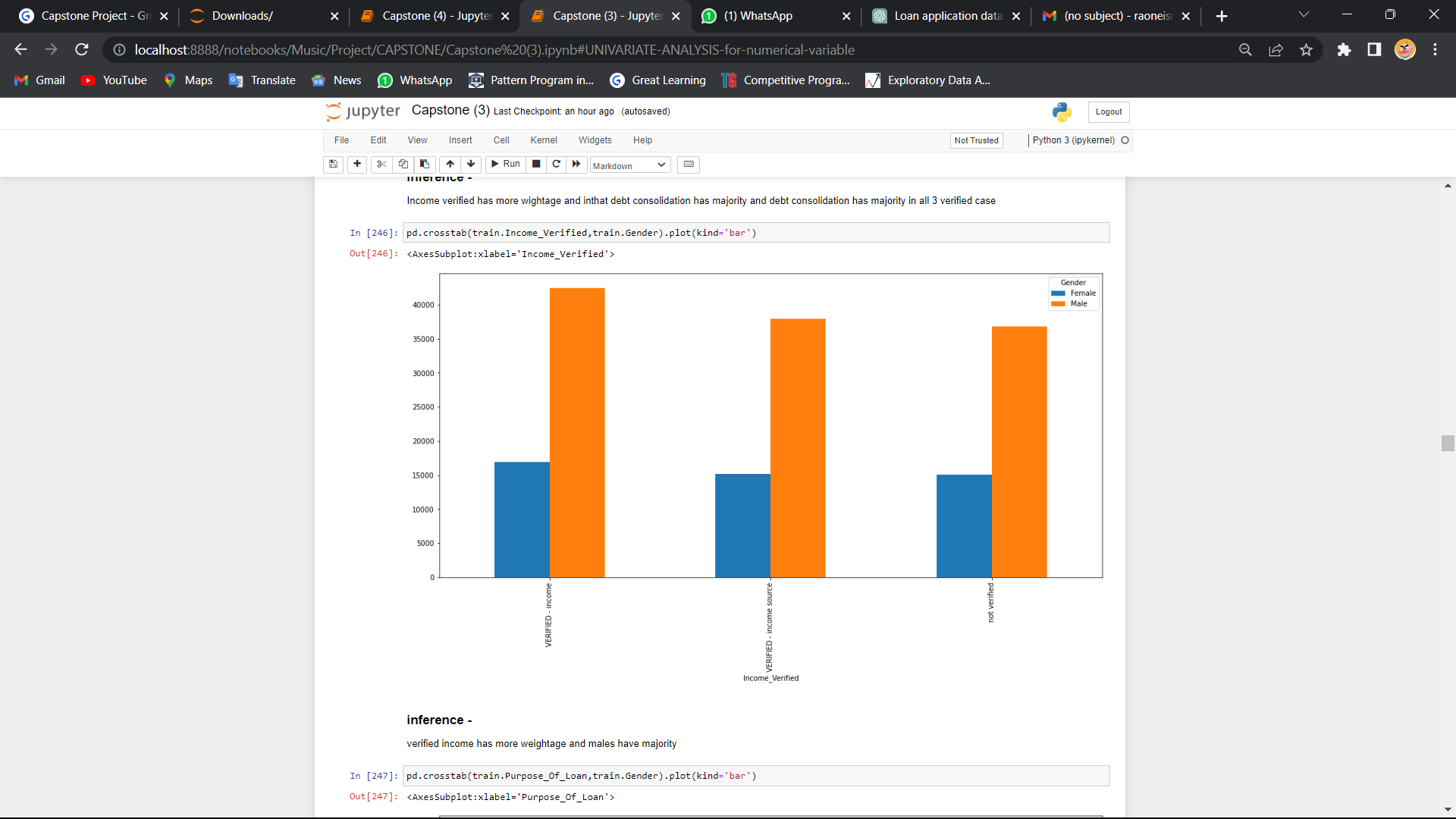
Based on the skewness values , it seems like 'Annual\_Income' has the highest skewness value of 40.225306, indicating a highly asymmetric distribution. 'Inquiries\_Last\_6Mo' and 'Number\_Open\_Accounts' also have high skewness values, indicating that their distributions may be skewed to the right.

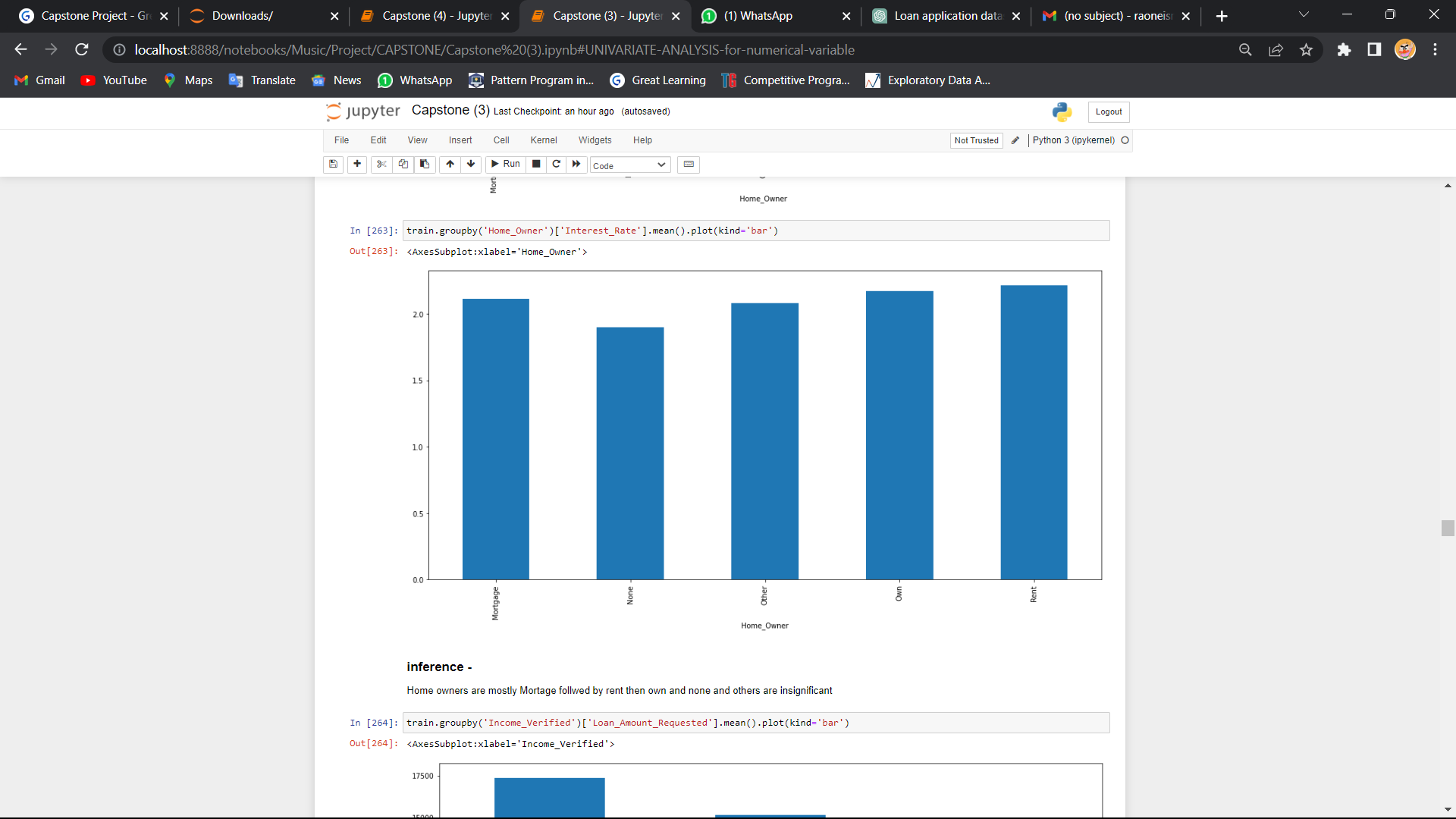
**Distriubution of Loan requested:-**

****

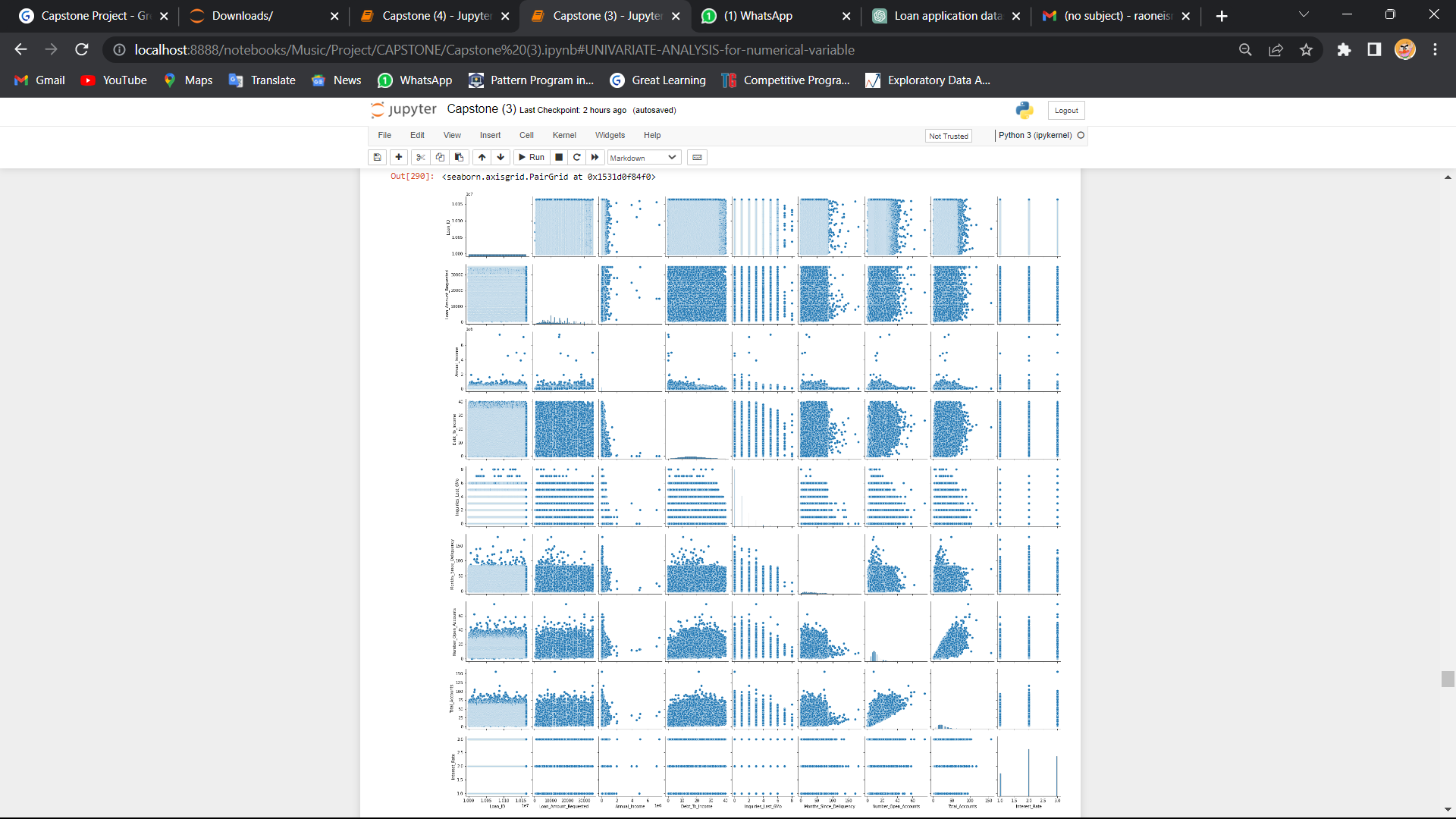
****

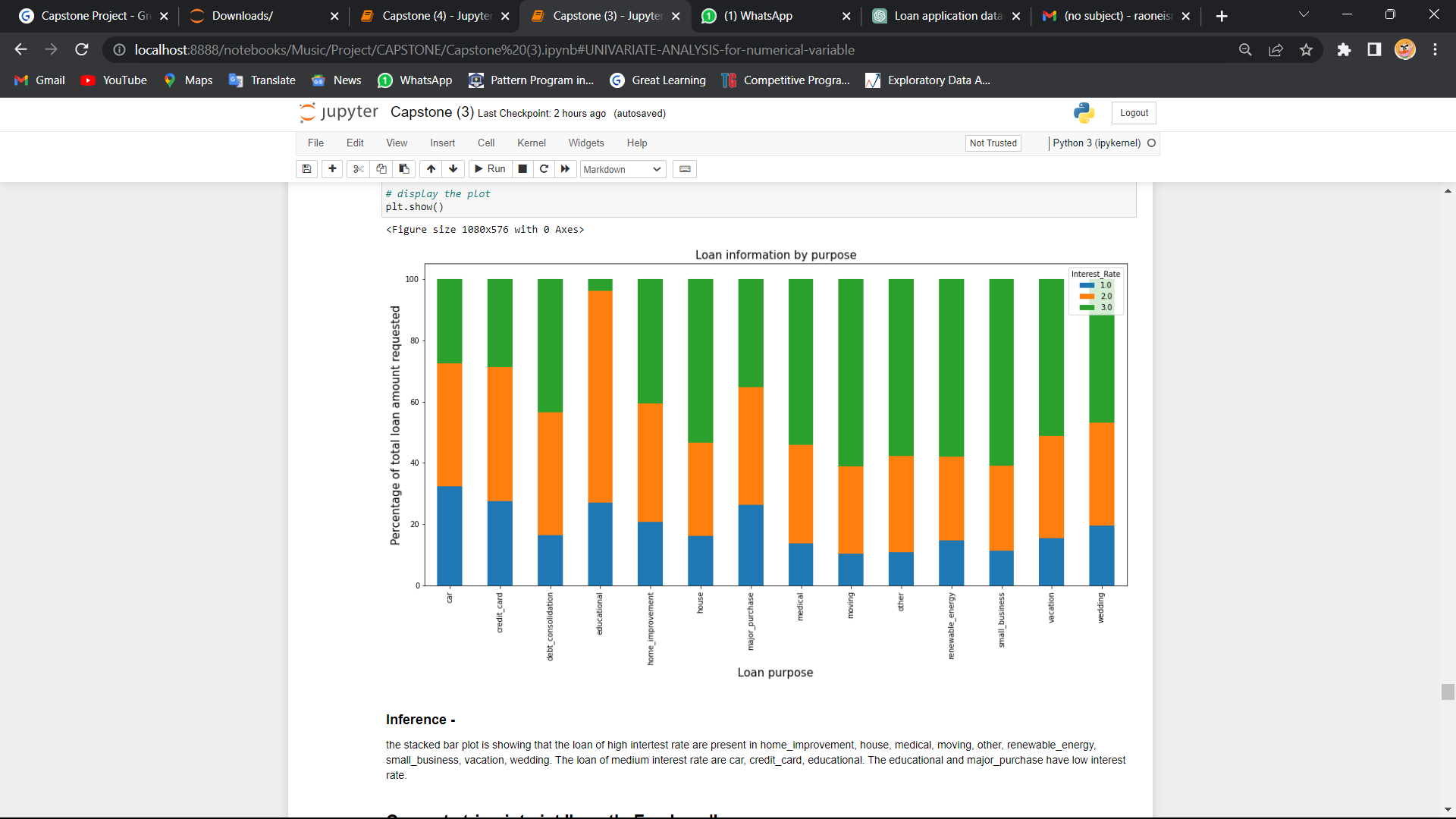
# Bivariate analysis

****

****

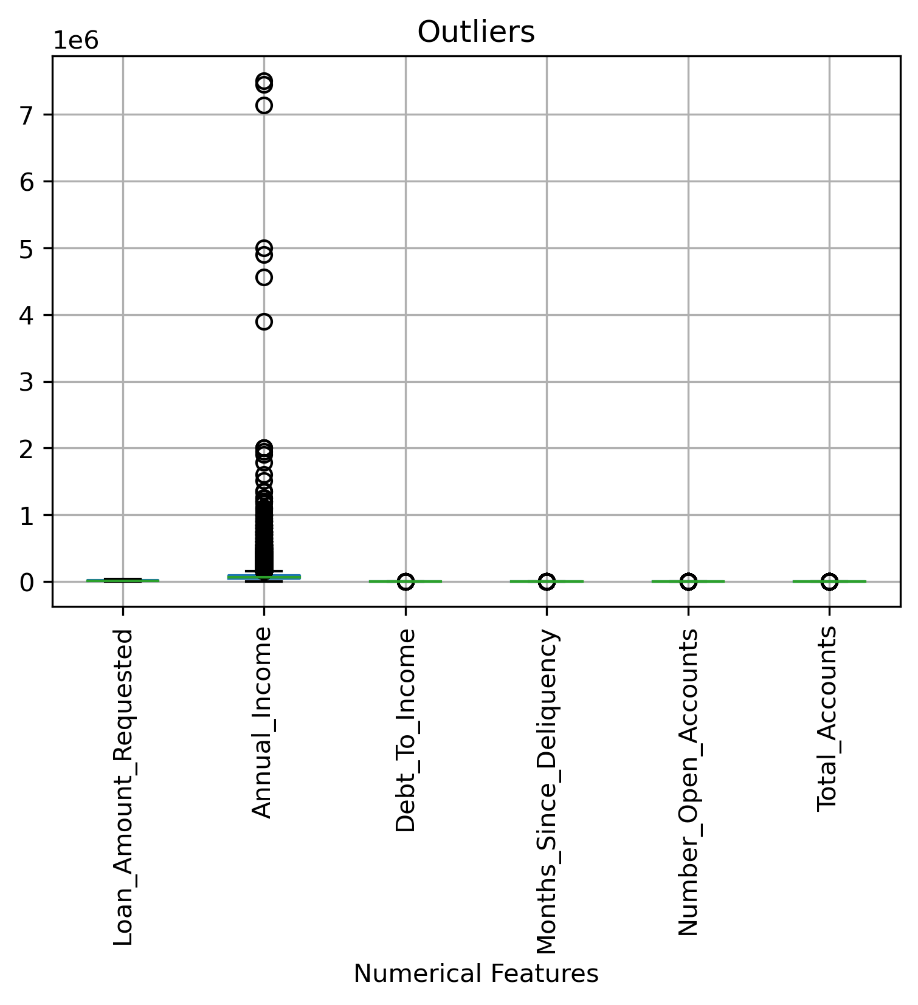
# Multivariate analysis

****

****

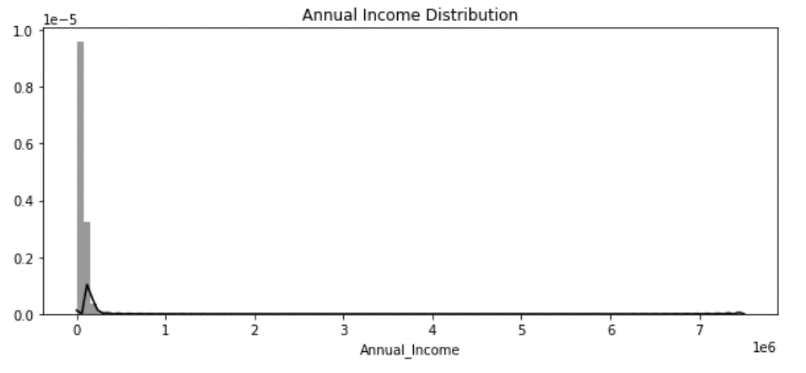
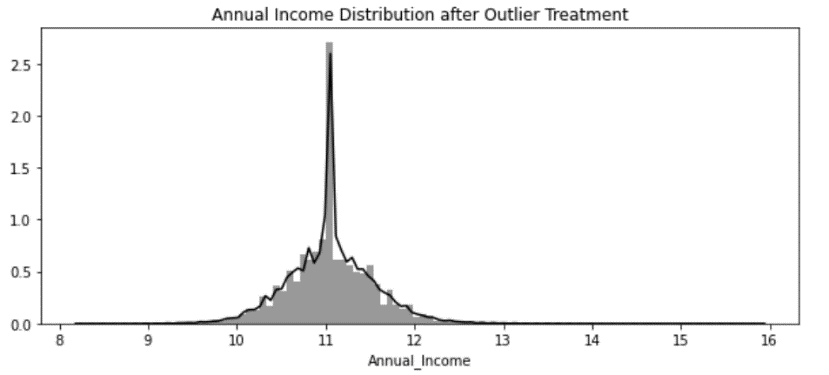
**OUTLIERS AND TREATMENT**

We use boxplots to visualise outliers present in the data.



Apparently the ‘Annual Income’ feature is having many outliers.

Since the outliers are increasing the range of the data and the data is skewed, we perform log transform to treat outliers.

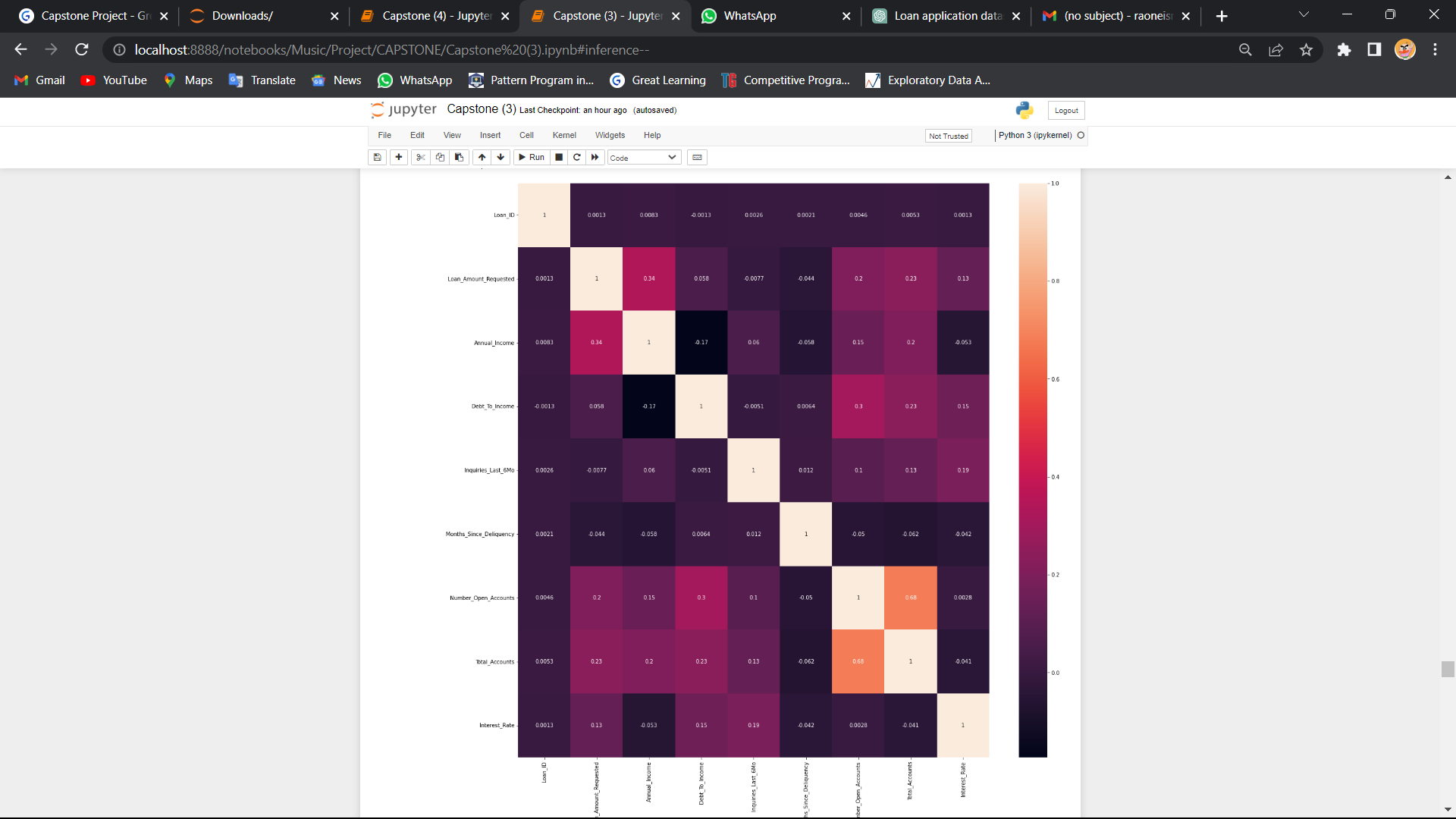
 

Before After

The histograms clearly show that now the data has less skewness and outliers as compared to earlier.

**FEATURE CORRELATION**

We use Pearson correlation to find correlation among features and plot them on a heatmap in Seaborn.

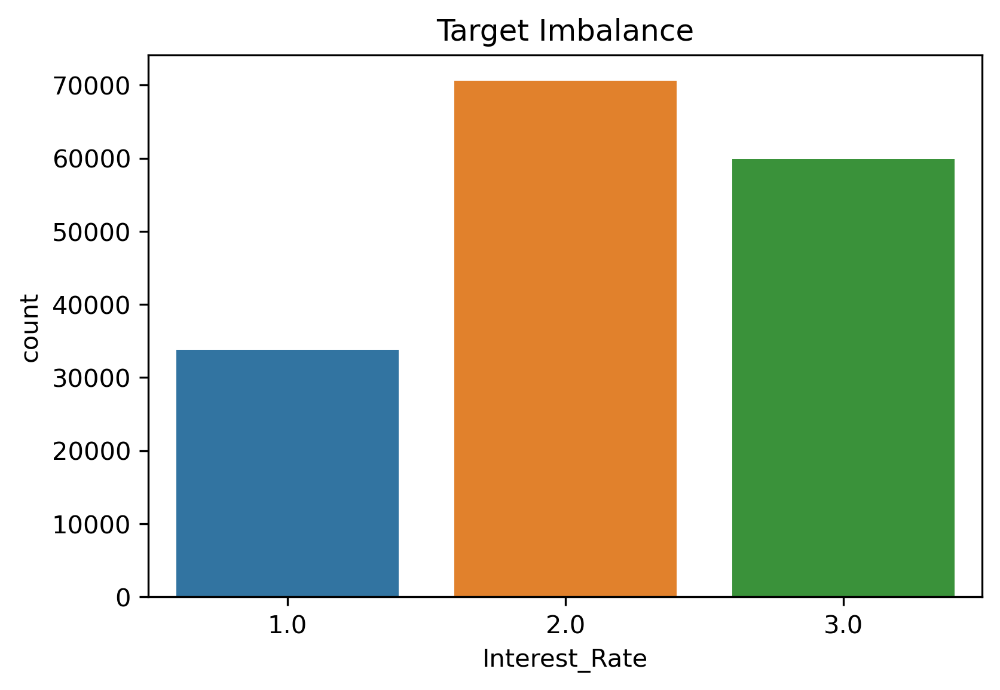


### inference -

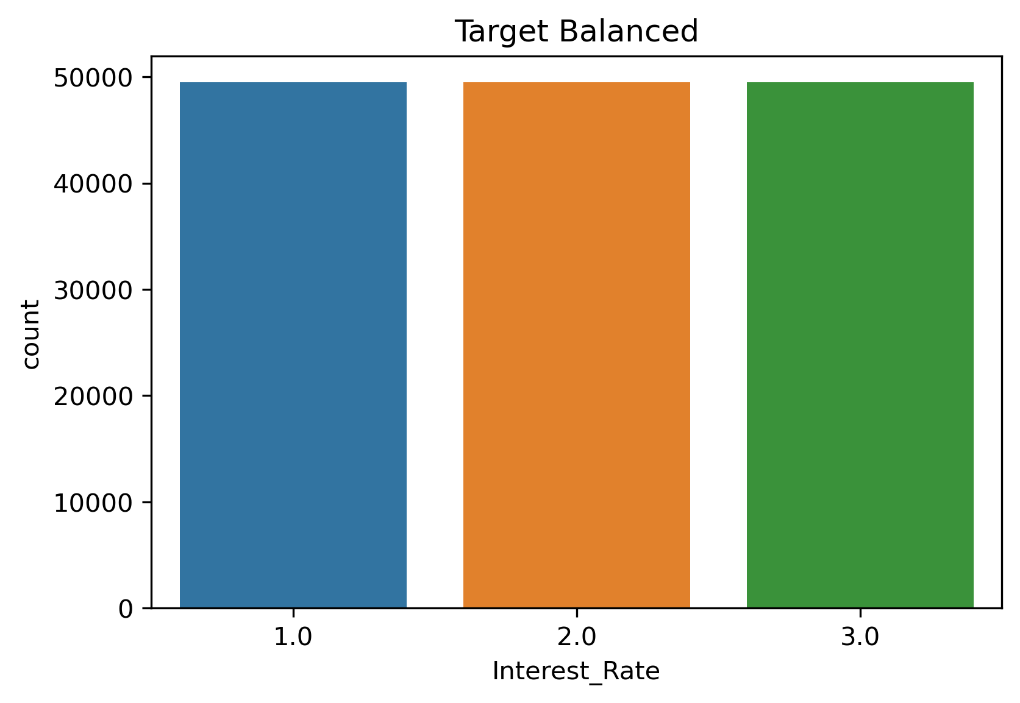
Here Loan ammount requested and annual income has highest corelation so we can infer person who has high annual income he can go for high loan ammount while loan ammount and inqiries last six month has least correlation

**BALANCING OF DATA**

During our initial EDA it was evident that our Target is imbalanced.



Because of the cons related to simple undersampling and oversampling techniques we will use SMOTE technique to balance the target. Smote creates artificial data points near to minorities to balance the data.



Now it is evident that that our target is balanced and ready for training on models.

**BASE MODEL**

Since ours is a classification problem at first we will fit a base model to getter a rough idea of predictions.

Here we will use Logistic Regression algorithm with ‘multinomial’ argument under the multiclass parameter as we have more than two classes in the target.

The report is as follows:

**Classification Report**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target Class | **precision** | **recall** | **f1-score** | **support** |
|  |  |  |  |  |
| 1 | 0.55 | 0.57 | 0.56 | 21102 |
| 2 | 0.43 | 0.42 | 0.42 | 21308 |
| 3 | 0.48 | 0.48 | 0.48 | 21112 |
|  |  |  |  |  |
| **accuracy** |  |  | 0.49 | 63522 |
| **macro avg** | 0.49 | 0.49 | 0.49 | 63522 |
| **weighted avg** | 0.49 | 0.49 | 0.49 | 63522 |

Our model gave an overall accuracy of 49%.