greatlearning

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

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

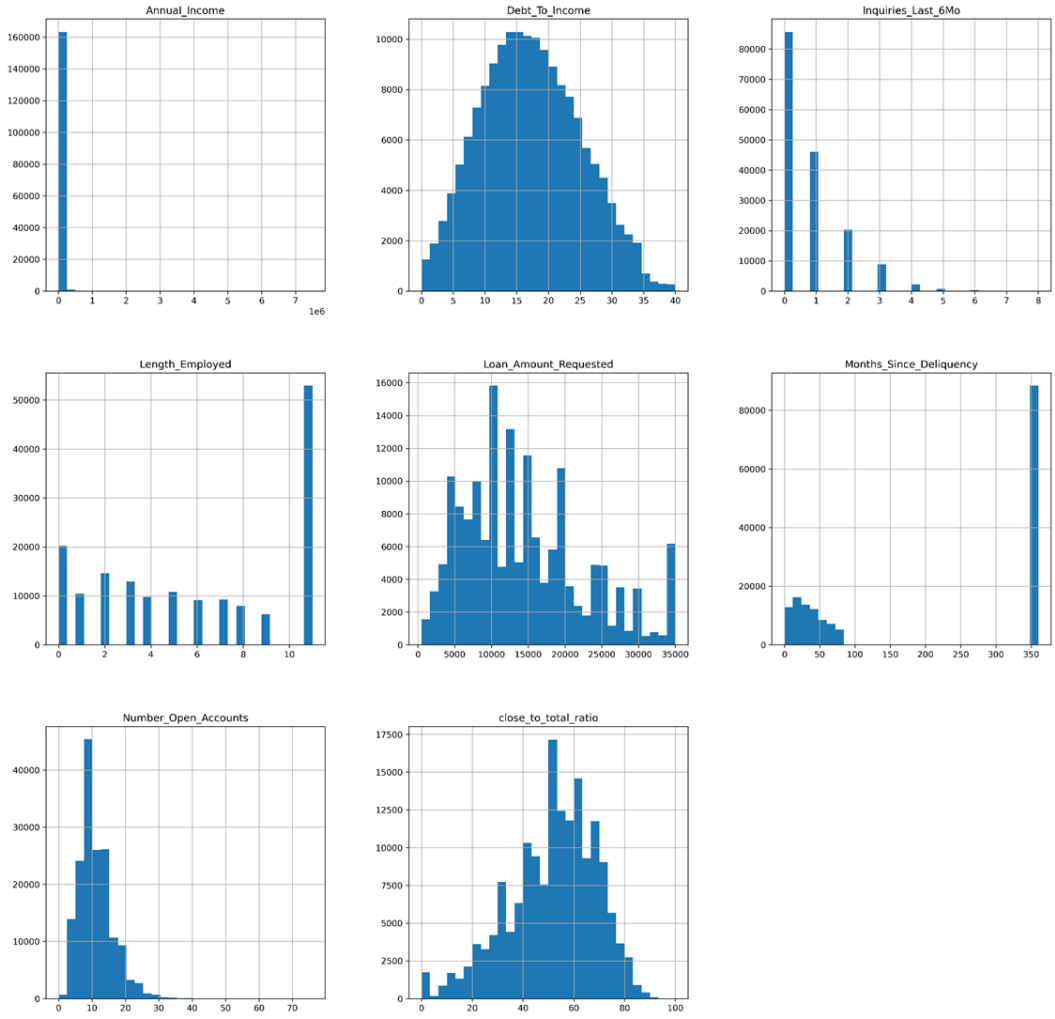
One Hot Encoding: The columns Assets_Liability, Home_Owner, Purpose_Of_Loan and Gender will be one hot encoded using the pd.get_dummies.

DATA DISTRIBUTION

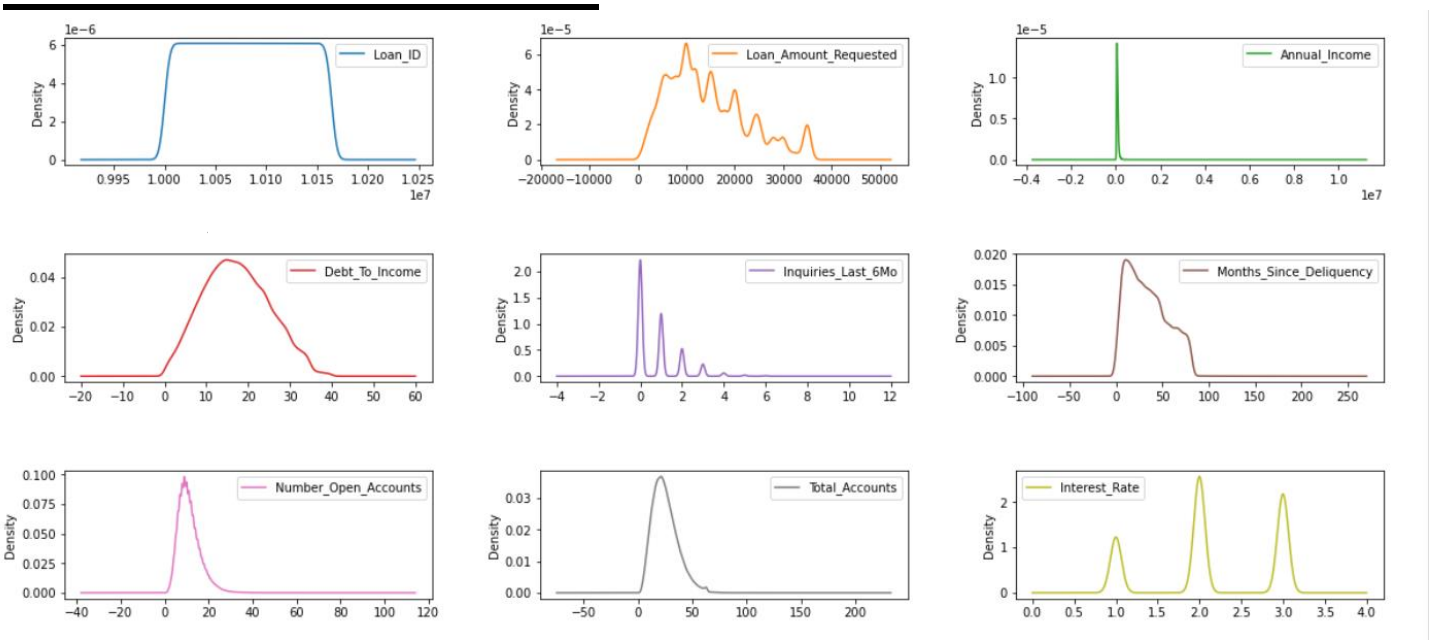
Using Data.describe() to create a 5 point summary of the data to get a better understanding of the numerical features in the dataset.

	mean	std	min	25%	50%	75%	max
Loan_Amount_Requested	14349.336920	8281.868700	500.0	8000.00	12075.00	20000.00	35000.00
Length_Employed	6.046388	4.133682	0.0	2.00	6.00	11.00	11.00
Annual_Income	71752.835884	55698.547344	4000.0	48600.00	63000.00	82000.00	7500000.00
Income_Verified	5.229689	4.108648	0.0	0.00	5.00	10.00	10.00
Debt_To_Income	17.207189	7.845083	0.0	11.37	16.84	22.78	39.99
Inquiries_Last_6Mo	0.781698	1.034747	0.0	0.00	0.00	1.00	8.00
Months_Since_Deliquency	209.455812	163.089910	0.0	34.00	360.00	360.00	360.00
Number_Open_Accounts	11.193818	4.991813	0.0	8.00	10.00	14.00	76.00
Gender	0.713144	0.452295	0.0	0.00	1.00	1.00	1.00
Interest_Rate	2.158951	0.738364	1.0	2.00	2.00	3.00	3.00
close_to_total_ratio	51.526093	17.216700	0.0	40.91	53.57	64.10	100.00

Using histograms plot in python to visualise data distribution.



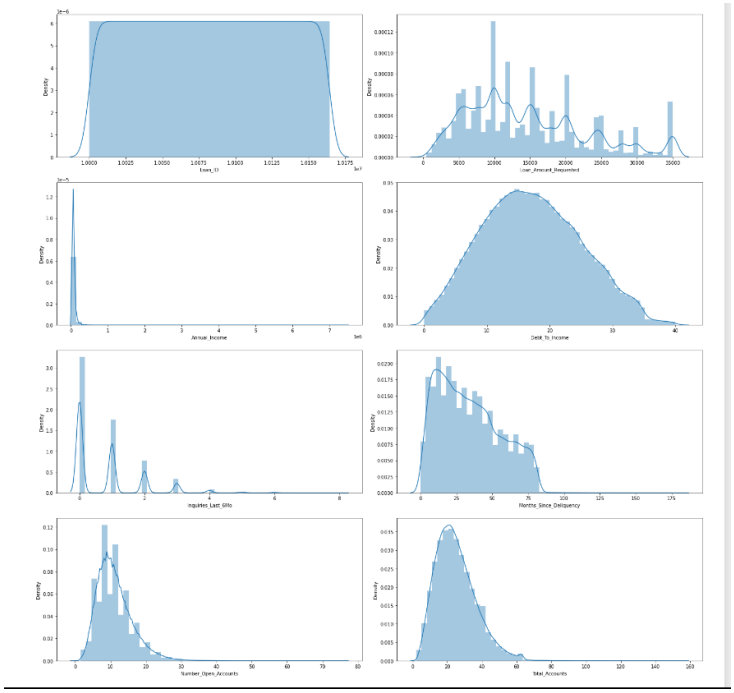
UNIVARIATE ANALYSIS



Inference -

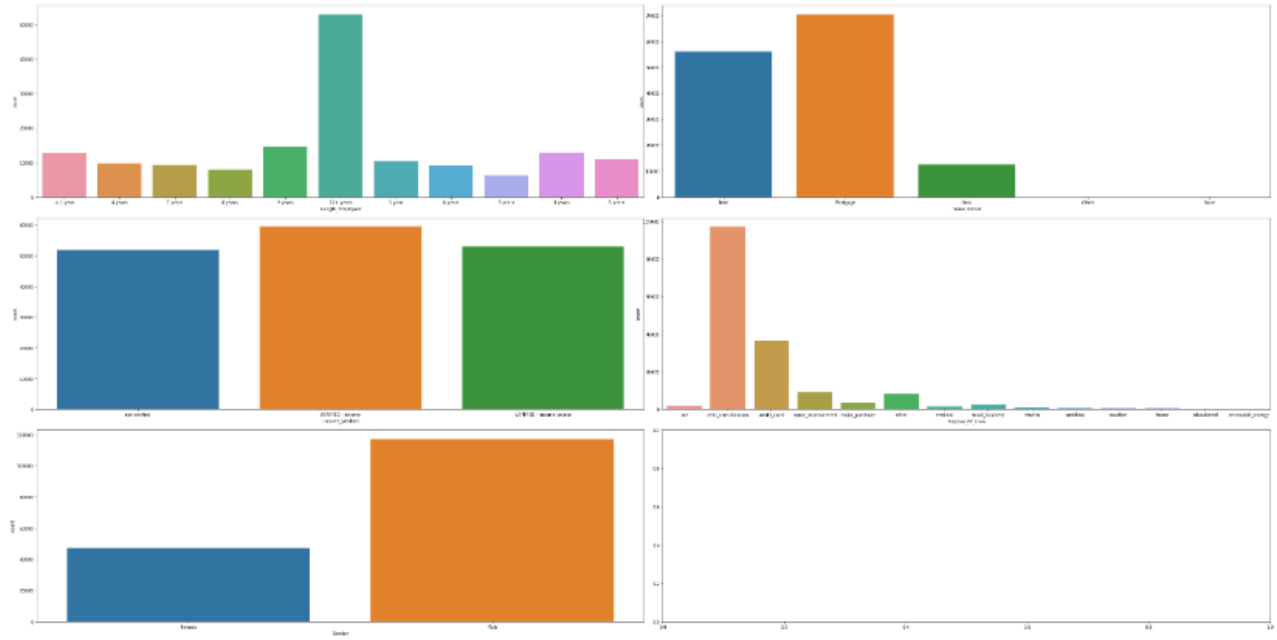
Based on the skewness values , it seems like 'Annual_Income' has the highest skewness value of 40.225306, indicating a highly asymmetric distribution. 'Inquiries_Last_6Mo' and 'Number_Open_Accounts' also have high skewness values, indicating that their distributions may be skewed to the right.

Distriubution of Loan requested:-



UNIVARIATE ANALYSIS for categorical variable

In [208]:
```
fig,ax=plt.subplots(3,2,figsize=(40,20))
for i,a in zip(train.select_dtypes(exclude=np.number).columns,ax.flatten()):
    sns.countplot(train[i],ax=a)
    plt.tight_layout()
    plt.figure()
```



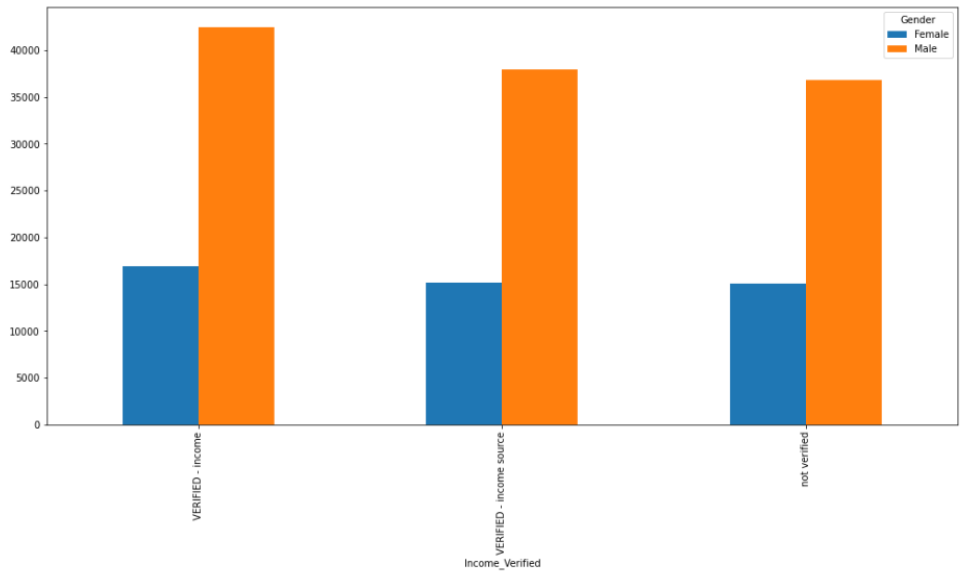
```
<Figure size 1224x576 with 0 Axes>
<Figure size 1224x576 with 0 Axes>
<Figure size 1224x576 with 0 Axes>
<Figure size 1224x576 with 0 Axes>
<Figure size 1224x576 with 0 Axes>
```

inference -

there are six chatagerocal variables in dataframes are 'Length_Employed', 'Home_Owner', 'Income_Verified', 'Purpose_Of_Loan', 'Gender', 'source' in Length_Employed there are 12 columns which are '< 1 year', '4 years', '7 years', '8 years', '2 years', '10+ years', '1 year', nan, '6 years', '9 years', '3 years', '5 years' and 10+ years has most in Home_Owner there are 6 types which are 'Rent', 'Mortgage', nan, 'Own', 'Other', 'None' majority is 'Mortgage' in Income_Verified there 4 types not verified', 'VERIFIED - income', 'VERIFIED - income source' and majority is VERIFIED - income' in Purpose_Of_Loan 14 types which are 'car', 'debt_consolidation', 'credit_card', 'home_improvement', 'major_purchase', 'other', 'medical', 'small_business', 'moving', 'wedding', 'vacation', 'house', 'educational', 'renewable_energy' in that debt_consolidation highest weightage followed by credit_card and others are significant In Gender there 2 types 'Female', 'Male' and males are dominating class In source there are types train test train contains train data and test contains test data

Bivariate analysis

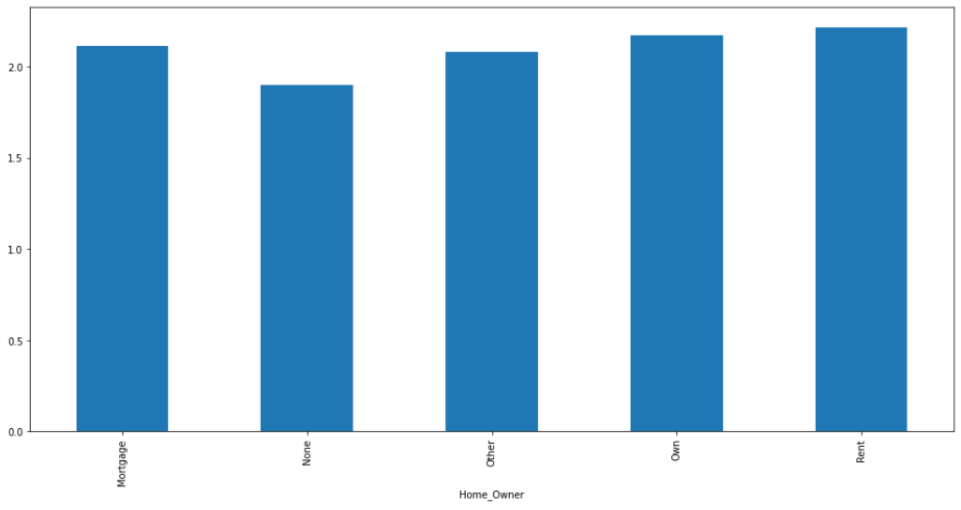
```
In [246]: pd.crosstab(train.Income_Verified,train.Gender).plot(kind='bar')
Out[246]: <AxesSubplot:xlabel='Income_Verified'>
```



inference -

verified income has more weightage and males have majority

```
In [263]: train.groupby('Home_Owner')['Interest_Rate'].mean().plot(kind='bar')
Out[263]: <AxesSubplot:xlabel='Home_Owner'>
```

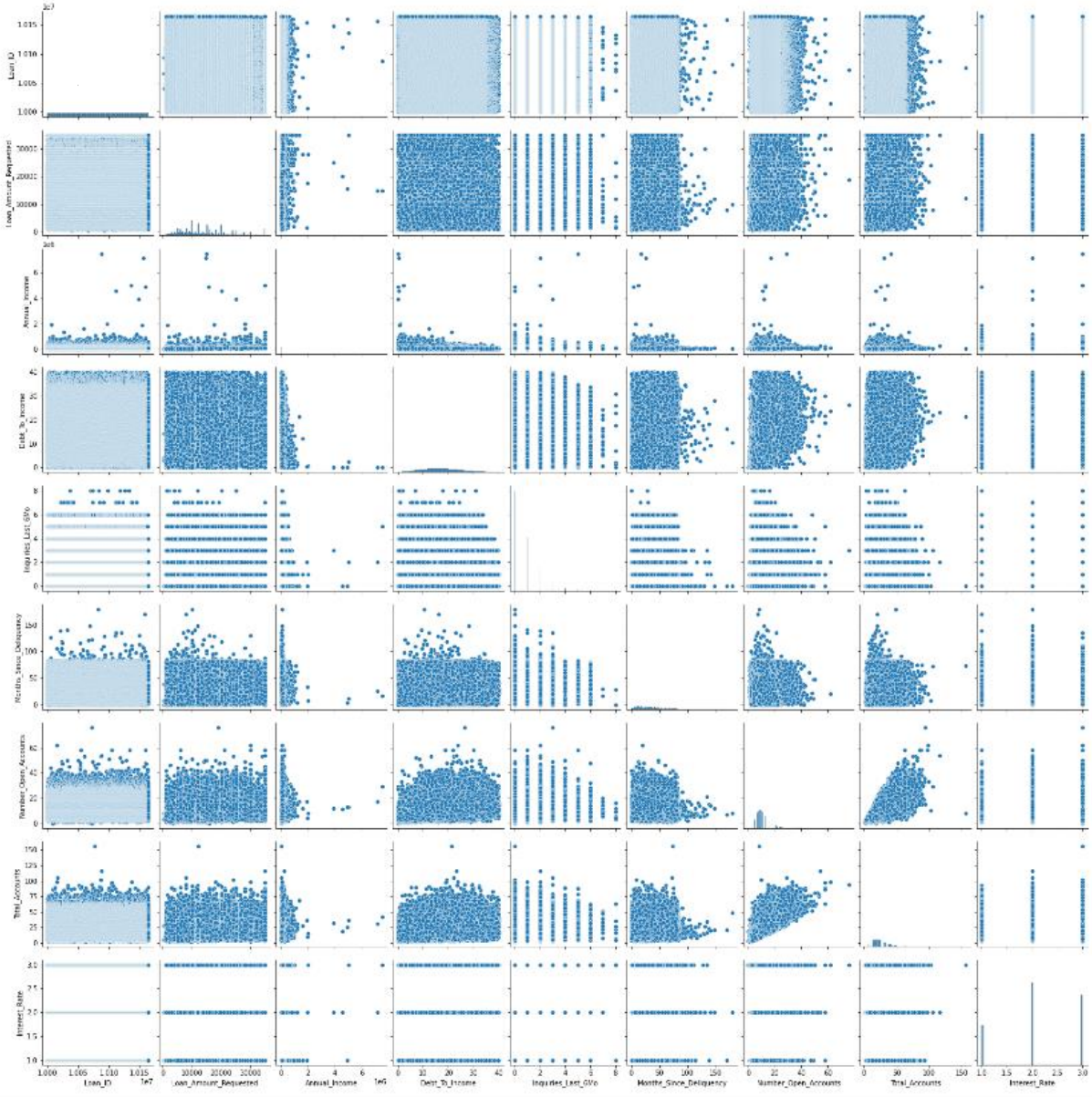


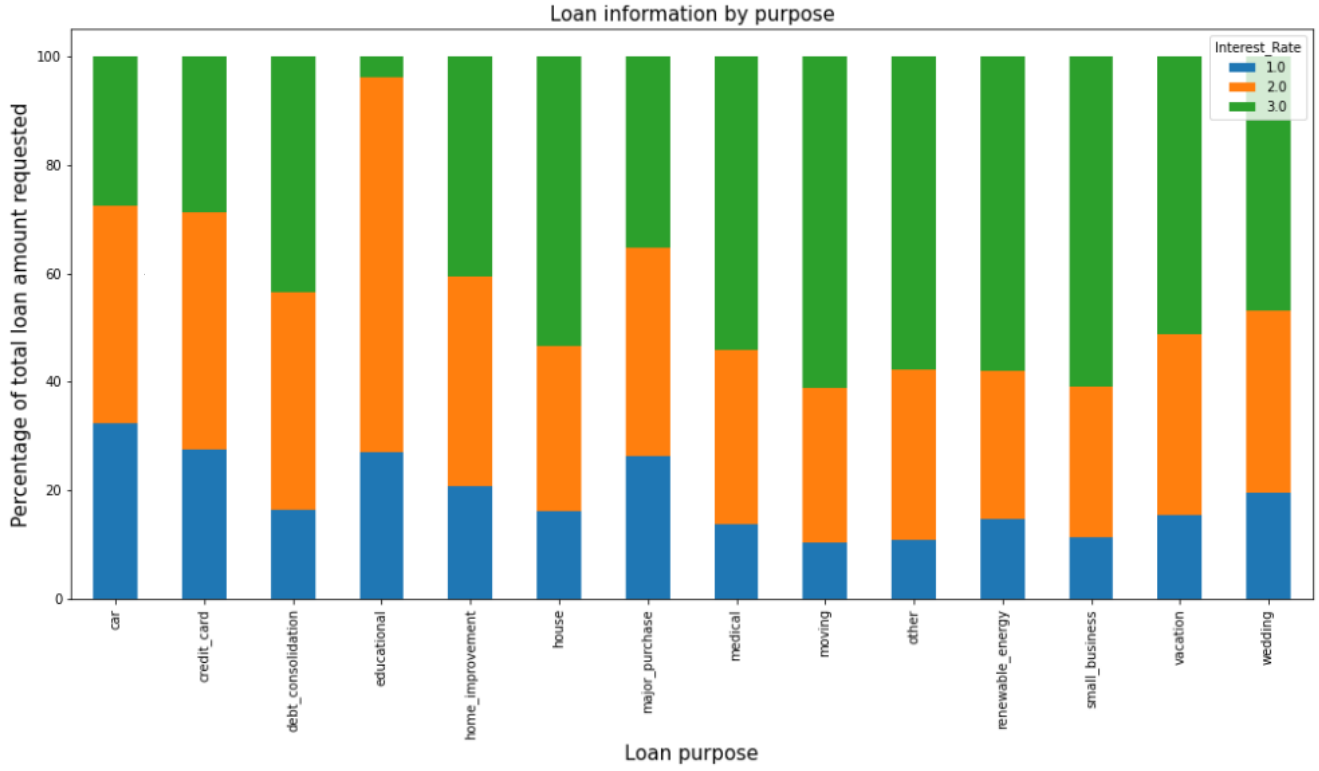
inference -

Home owners are mostly Mortage follwed by rent then own and none and others are insignificant

Multivariate analysis

```
Out[290]: <seaborn.axisgrid.PairGrid at 0x1531d0f84f0>
```



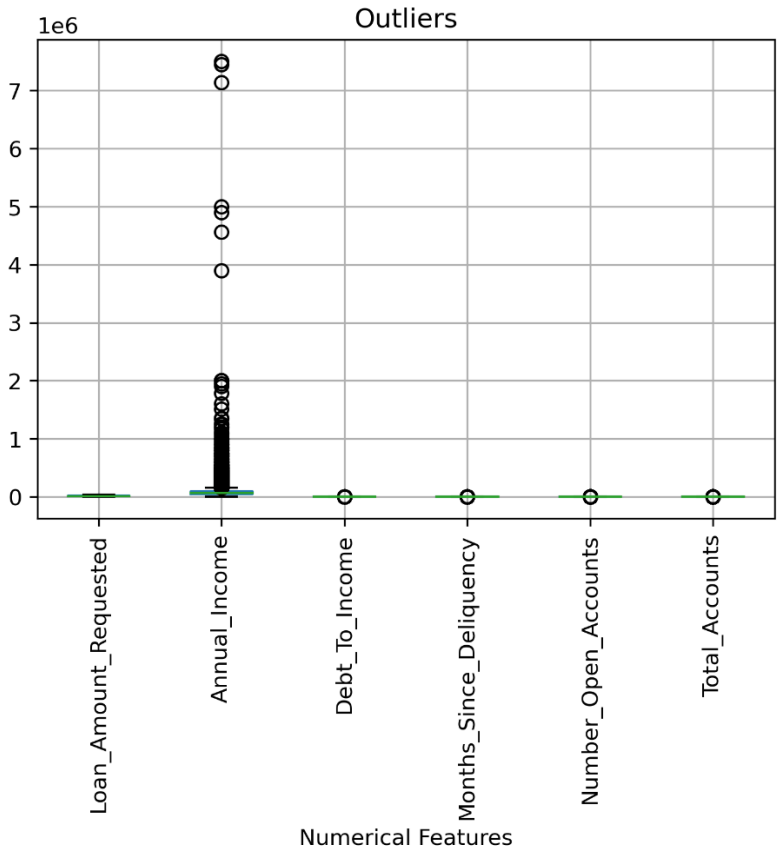


Inference -

the stacked bar plot is showing that the loan of high intertest rate are present in home_improvement, house, medical, moving, other, renewable_energy, small_business, vacation, wedding. The loan of medium interest rate are car, credit_card, educational. The educational and major_purchase have low interest rate.

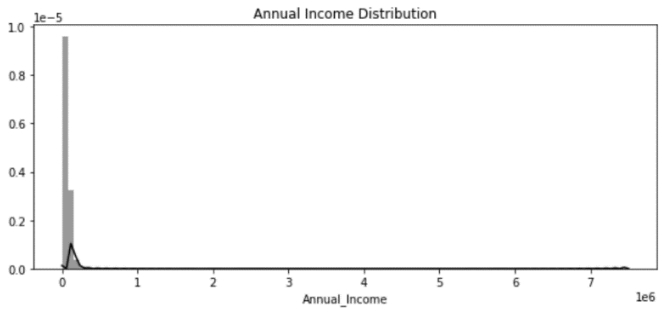
OUTLIERS AND TREATMENT

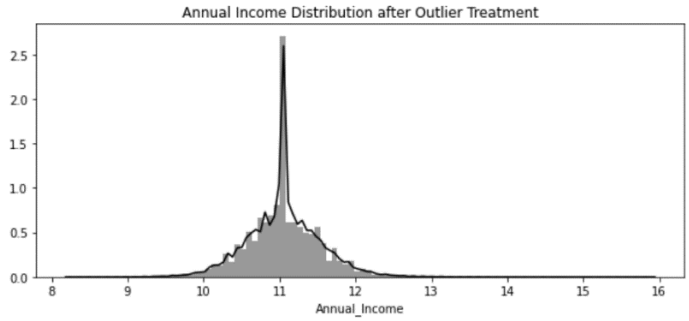
We use boxplots to visualise outliers present in the data.



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.



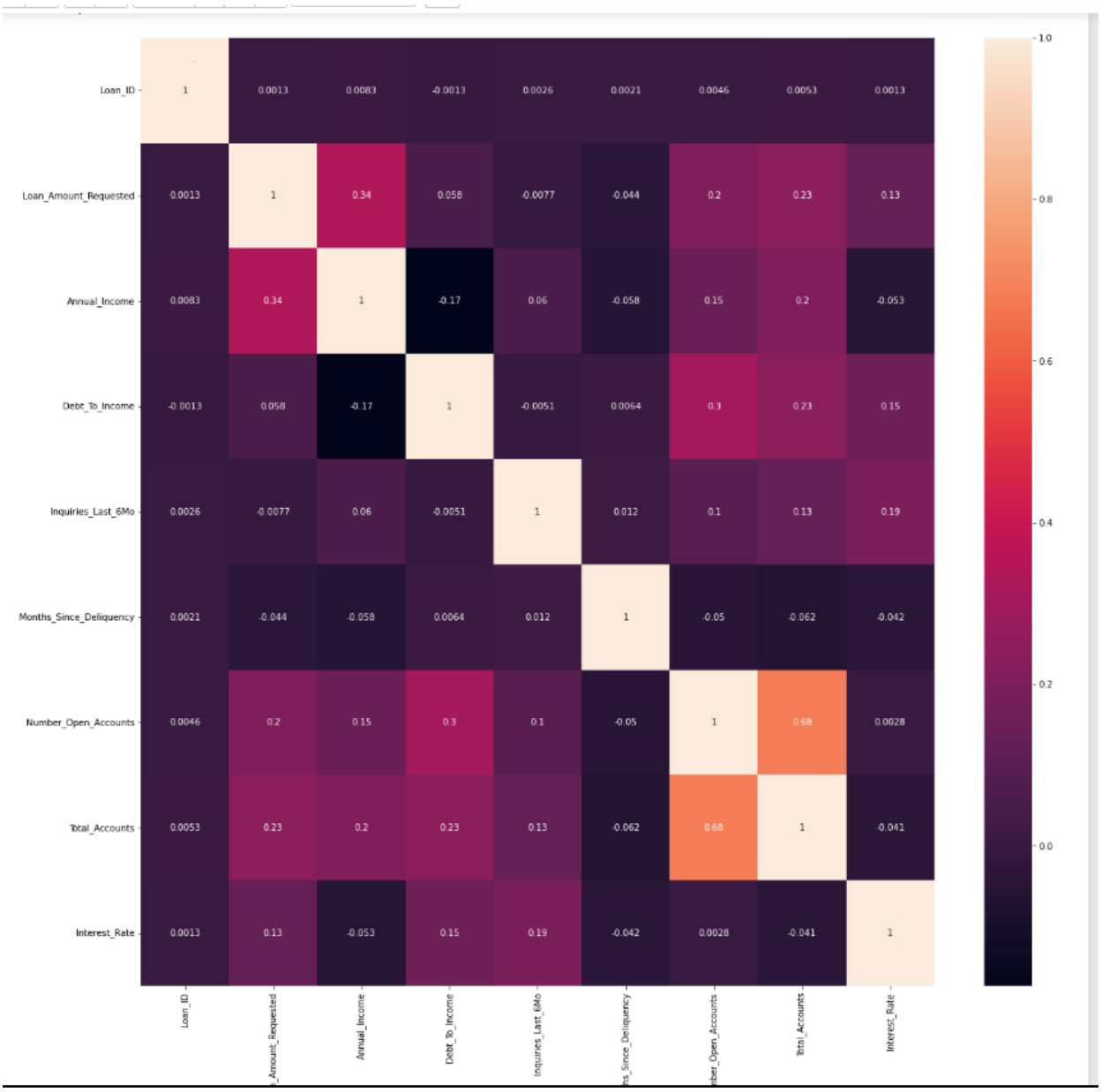


Before After

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.

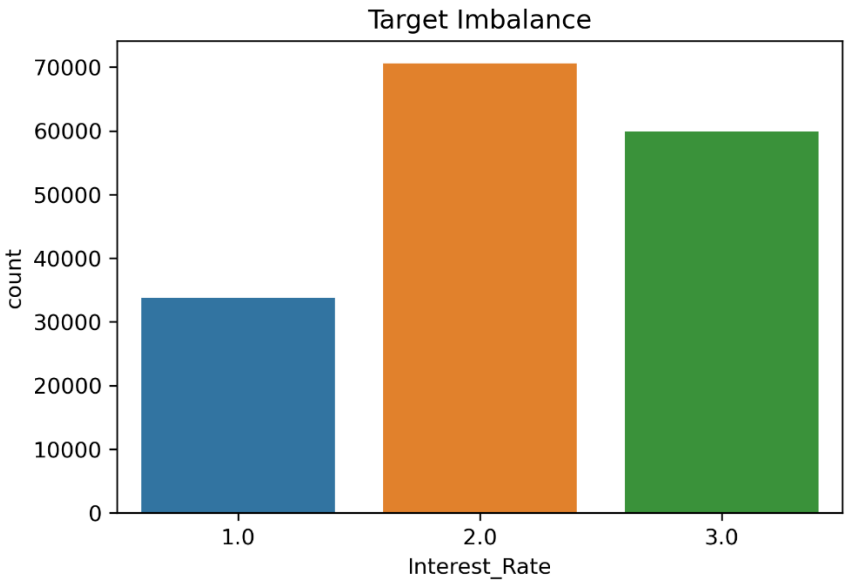


inference -

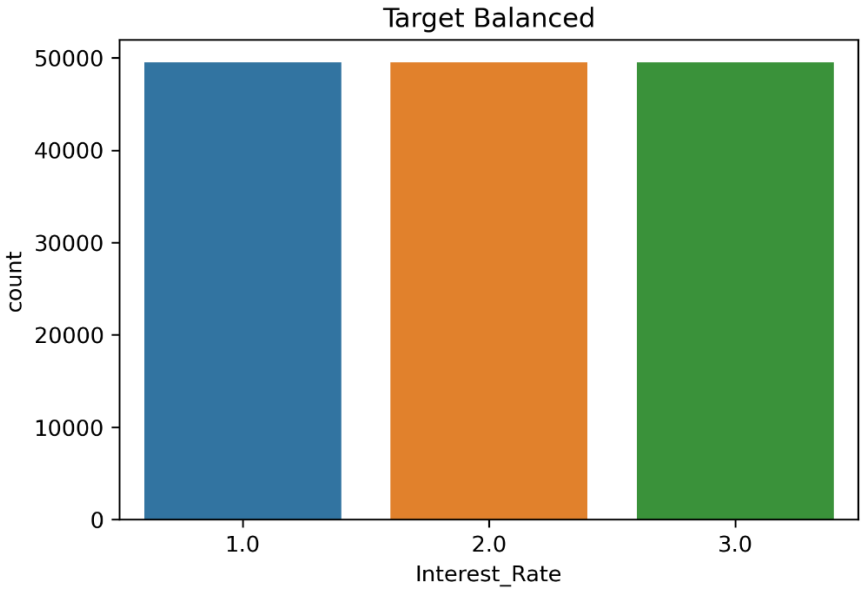
Here Loan ammount requested and annual income has highest corelation so we can infer person who has high annual income he can go for high loan ammount while loan ammount and inqiries last six month has least correlation

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%.