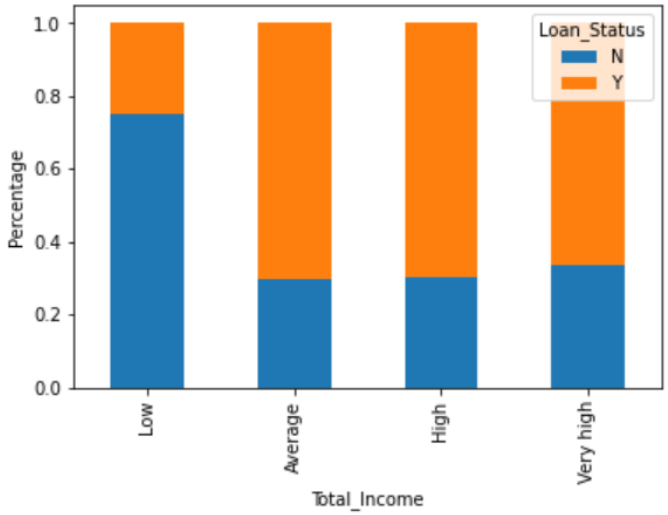
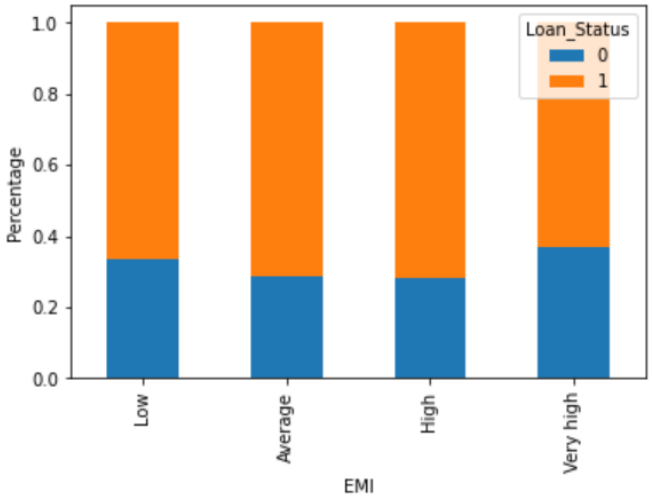
Feature Engineering

We will create the following new features based on the business logic
Total Income – We found that when co-applicants' income is less the chances of loan approval are high. But this does not seem right. The possible reason behind this may be that most of the applicants do not have any co-applicant so the co-applicant income for such applicants is taken as 0. The 0s in Co-applicant column can be interpreted as absence of coapplicant in the loan application. Hence, we can create a new feature Total_Income **combining applicant and coapplicant's income.**



[Where, low=(0-500000), Average=(500001-100000), High=(1000001-2000000) , Very high=(2000000-19225000)]

EMI- EMI is the monthly amount to be paid by the applicant to repay the loan. Idea behind creation of this feature is that, many applicants have chosen Loan amount term 360 months irrespective of their Loan amount. To categorise applicant based on their loan amount and loan term we need this feature. people who have high EMI's might find it difficult to pay back the loan. Assuming the loan amount as sum of principal and interest on loan, we can calculate the EMI by taking the ratio of loan amount with respect to loan amount term.



[Where low =(0-500), Average =(501-750), High=(750-1000), Very high =(1000-20000)]

This shows no significant impact of EMI feature on target variable. Same can be checked by **correlation heat map**.

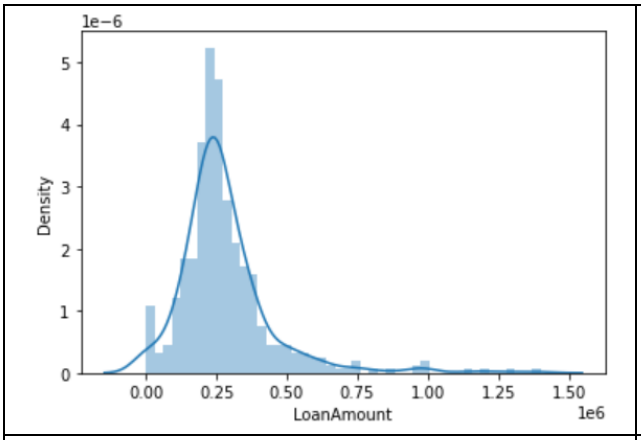
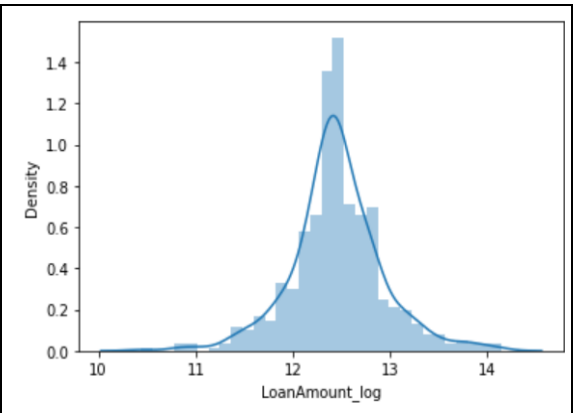
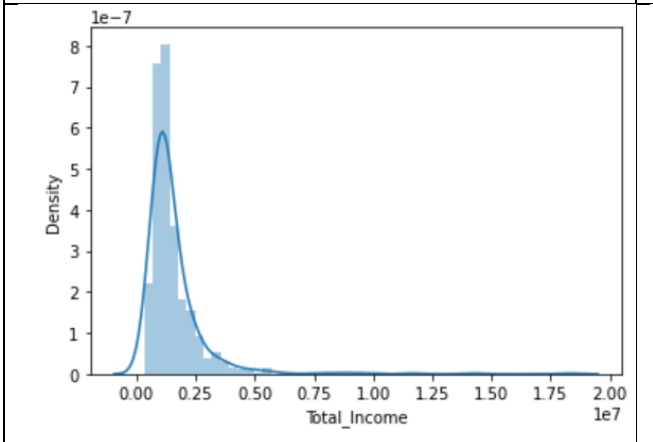
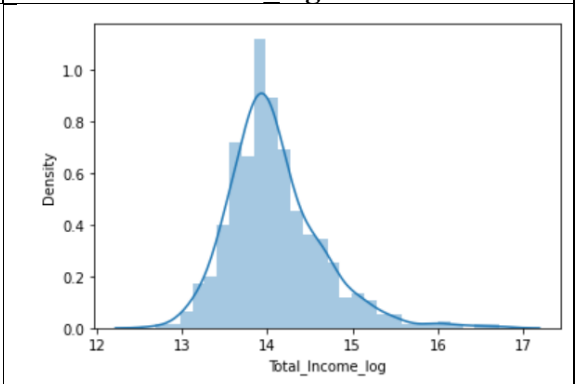
Missing Value Imputation

- There are very fewer missing values in **Gender, Married, Dependents, Credit_History and Self_Employed** features so we can fill them **using the mode** of the features.
- We have seen that in **Loan_Amount_Term** variable, the value of 360 is repeating the most. So, we will replace the missing values in this variable **using the mode** of this variable.
- **Loan Amount** cannot be 0. This may be assumed as missing value. We will use the median to fill the null values as earlier we saw that the loan amount has outliers

Outlier Treatment

- We found in univariate analysis that **LoanAmount** contains outliers so we must treat them as the presence of outliers affects the distribution of the data**. Log transformation will reduce the impact of outliers** and change a right skewed distribution to normal distribution
- Similar log transformation must be applied to the newly created feature **Total_Income** as the data distribution is right skewed

Data Distribution

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>LoanAmount Distribution</td><td>LoanAmount_log Distribution</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Total_Income Distribution</td><td>Total_Income_log Distribution</td></tr>
</table>

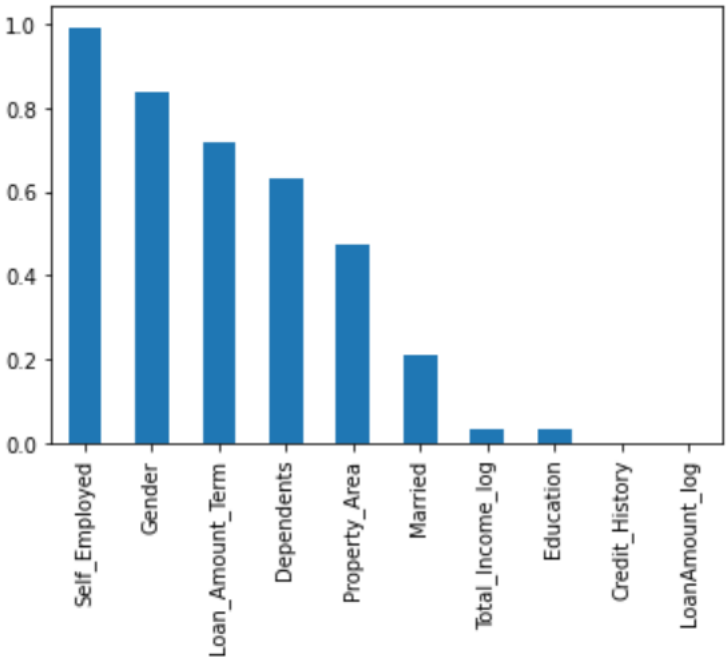
Feature Selection

I. Numerical Feature:

- The correlation (heatmap) suggested that the feature '**Loan_Id**' has no impact on loan approval. Hence it should be dropped.
- Earlier we found that the newly created feature **EMI** also has not much impact. Hence it must be dropped.
- Let us now drop the variables **'ApplicantIncome', 'CoapplicantIncome'** which we used to create Total_Income. Reason for doing this is, the correlation between those old features and the new feature will be very high and logistic regression assumes that the variables are not highly correlated.

II. Categorical Feature:

- From the **chi2 test**, we have found the **p-values** of Self_Employed', 'Gender', 'Loan_Amount_Term', and 'Dependents' features **greater than 0.5** which implies these features have less influence on Loan prediction. Hence these features can be dropped.



After Label Encoding and dropping the unnecessary features, our dataset left with the following features

```
Data columns (total 7 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Married           521 non-null    int32
 1   Education         521 non-null    int32
 2   Credit_History    521 non-null    int64
 3   Property_Area     521 non-null    int32
 4   Loan_Status       521 non-null    int64
 5   LoanAmount_log    521 non-null    int64
 6   Total_Income_log  521 non-null    int64
dtypes: int32(3), int64(4)
```

Submitted by: Subhransu Sekhar Mallick