**MARKET RESEARCH PHASE**

**1) Probable business impact of each independent feature on the target:**

**Loan\_Id**: This feature is not expected to have any significant impact on the target variable.

**Gender**: This feature is expected to have little to very less impact on the target variable.

**Married**: This feature will have an impact on the target variable as married individuals may have better financial stability and therefore better chances of loan approval.

**Dependents**: The number of dependents is an important feature as it reflects the applicant’s financial responsibilities. In general, higher the number of dependents of the applicant, the more will be the applicant’s financial burden which will have impact on his income and ability to repay the loan.

**Education**: Education will have a positive impact on target feature as certain levels of education may be associated with better financial stability and therefore better chances of loan approval. In general, a post graduate/ technical graduate employee is given the highest rating in the credit scoring appraisal form in banks.

**Self-employed**: This is an important feature which helps bank to determine the approval of loan as self-employed individuals may have less stable income, affecting loan approval as compared to regular and stable salaried individuals.

**Applicant Income**: Major feature for loan approval is applicant’s income. A higher income may be associated with better financial stability and therefore better chances of loan approval.

**CoApplicantIncome:** It may impact as additional income from a co-applicant may improve the ability to repay the loan and therefore better chances of loan approval.

**Loan Amount**: This feature reflects the financial risk to the lender and will have an impact on the loan approval.

**Loan\_Amount\_Term**: The loan term/ repayment period plays an important role in determining the EMI of the loan. This feature will depend on the type of loan required by the applicant. In general, a longer loan term is better in cases of Housing or Property loans where banks mortgage the property and keeps it as security. The reason is that even if the applicant defaults in such loans after a certain period of time, the value of the property will never depreciate, thereby decreasing the risk of the lender. So, a longer repayment tenure is preferred in such loans as it reduces the EMI making it easy for approval of the loan for the applicant. However, in case of vehicle loans a shorter repayment time is preferred by the banks as the value of vehicle depreciates thereby increasing the risk of the lender with increase in time.

**Credit\_History**: This feature will have impact significant impact on the target variable. It reflects the applicant’s past financial lending behaviour. A good credit history credit history increases chances of the loan being approved. CIBIL score and EXPERIAN score are used to determine the credit history of the applicant.

**Property\_Area**:. The property area can impact loan approval by influencing factors such as property value, employment opportunities, risk assessment, and repayment capacity. In general, urban areas are given higher preference due to higher property values, more job opportunities as compared to rural areas, which can increase the financial stability and repayment capacity of urban applicants.

**2) Ways in which the organization can benefit from analysing the data:**

* **Reduce time and effort of employees**: Automating the loan approval process using machine learning can help reduce the time and effort spent by bank employees on manual processing of loan applications.
* **Improve loan approval accuracy**: A machine learning model can help identify the appropriate candidature for loan approval, reducing the risk of approving bad loans.
* **Improve customer experience**: Faster and more accurate loan approvals can result in a better customer experience, leading to increased customer loyalty and retention.
* **Increase profitability**: Better loan approvals can help increase profitability by reducing bad debt, minimizing defaults, and improving loan repayment rates.

**3) Missing features that can help with the analysis based on business logic:**

* **Age of applicant**: Age of the applicant can have impact on the loan approval status as age can be associated with financial stability and affect the ability to repay the loan.
* **Purpose of the Loan**: The purpose of the loan or the type of loan required is also an important feature which can be taken into consideration. This feature will reflect the interest rate of the loan as a particular loan category will in general have a particular interest rate. Also if this feature is known other important features can be extracted out based on the type of loan. For example, in case of cas credits loan banks take into consideration the CA/ CL feature which in cases of term loans is not applicable and other parameter like DSCR is taken while evaluating a term loan proposal.
* **EMI/ NMI Ratio**: Equated Monthly Instalment divided by the Net Monthly Income is one of the important features which banks consider while approving a loan. A lower EMI/ NMI is preferred while approving a loan. Generally, most banks set a threshold for this parameter to 0.5 or 0.6.
* **Net Worth/ Loan amount**: Higher Net worth reflects the stability of the applicant. Net Worth/ Loan Amount may affect the target variable as a higher ratio of this feature will have positive impact on the target variable.
* **Loyalty**: Loyalty means the date or time since the applicant has opened and has been operating his accounts with the bank. In other words it is the measure of the timeline since the applicant has opened account in the bank, The longer the time and better the transactions in the applicant’s account, the better is the chance of loan approval.
* **Loan to Toatal Income Ratio**: The loan to income ratio is a measure of the borrower's ability to repay the loan based on their income. A higher loan to income ratio indicates that the borrower may have difficulty repaying the loan, and hence, a higher chance of defaulting on the loan.Therefore, the loan to income ratio is likely to have a significant impact on the target variable, which is the loan status.

**4) Best way to collect data for the suggested features:**

* **Age of applicant**: This information can be collected from official identification documents such as passport, driving license, or birth certificate. It can also be collected from bank’s server as bank’s are required to enter date of birth in their software while opening of accounts.
* **Purpose of the loan**: This information can be collected through a questionnaire or an interview with the applicant. If purpose of the loan is known, the interest rate of the loan will also be known.
* **EMI/ NMI Ratio**: EMI can be calculated if the interest rate of the loan is known along with loan amount and loan tenure. NMI is basically the income of the applicant which is present with us.
* **Net Worth/ Loan amount**: Net Worth of the applicant can be calculated by subtracting the total asset with the total liabilities of the applicant. Both asset and liabilities of the applicant can be assessed through a questionnaire or an interview with the applicant and later physically verifying it through pre sanction inspection by bank officials. Loan amount is already present with us. Loan amount feature can be dropped if this feature is implemented.
* **Loyalty**: The date when the customer has opened account with the bank can be easily extracted from bank’s server.
* **Loan to Toatal Income Ratio:** This feature can simply be obtained by dividing loan amount with the sum of applicant and co applicant income

**MILESTONE 2**

DATA ANALYTICS PHASE

Hypothesis Testing was performed to predict the impact of independent features on the target variable.

* Chi- Square test was conducted to predict the impact of Categorical Independent features on the target variable.
* T- test was conducted to predict the impact of Continuous Independent features on the target variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sl No.** | **Hypothesis** | **Test Name** | **Test Value** | **p-value** | **Conclusion** |
| 1 | Credit\_History vs Loan\_Status | Chi-square | 138.219174 | 6.53E-32 | Significant association between groups |
| 2 | Property\_Area vs Loan\_Status | Chi-square | 15.929602 | 3.47E-04 | Significant association between groups |
| 3 | Education vs Loan\_Status | Chi-square | 5.412089 | 2.00E-02 | Significant association between groups |
| 4 | Married vs Loan\_Status | Chi-square | 3.851837 | 4.97E-02 | Significant association between groups |
| 5 | Dependents vs Loan\_Status | Chi-square | 6.145781 | 1.05E-01 | No significant association between groups |
| 6 | Loan\_Amount\_Term vs Loan\_Status | Chi-square | 4.744131 | 4.52E-01 | No significant association between groups |
| 7 | ApplicantIncome vs Loan\_Status | t-statistic | -0.50764 | 6.12E-01 | No significant association between groups |
| 8 | LoanAmount vs Loan\_Status | t-statistic | -0.49899 | 6.18E-01 | No significant association between groups |
| 9 | Gender vs Loan\_Status | Chi-square | 0.158829 | 6.90E-01 | No significant association between groups |
| 10 | CoapplicantIncome vs Loan\_Status | t-statistic | -0.131229 | 8.96E-01 | No significant association between groups |
| 11 | Self\_Employed vs Loan\_Status | Chi-square | 0 | 1.00E+00 | No significant association between groups |

Detailed Conclusions from the above table:

Hypothesis 1: Gender vs Loan\_Status

Based on the Chi-square statistic of 0.15882917019977388, with 1 degree of freedom and a p-value of 0.6902367665572007, we fail to reject the null hypothesis. This suggests that there is no significant association between gender and loan status. In other words, gender does not have a significant impact on loan approval or rejection.

Hypothesis 2: Married vs Loan\_Status

With a Chi-square statistic of 3.851837310804866, 1 degree of freedom, and a p-value of 0.04969152864604629, we reject the null hypothesis. This indicates that there is a significant association between marital status (married or not) and loan status. Marital status may play a role in loan approval or rejection decisions.

Hypothesis 3: Dependents vs Loan\_Status

The Chi-square statistic of 6.145781110794439, with 3 degrees of freedom, and a p-value of 0.10472909943994584, does not provide enough evidence to reject the null hypothesis. This suggests that there is no significant association between the number of dependents and loan status. The number of dependents does not seem to impact loan approval or rejection decisions.

Hypothesis 4: Education vs Loan\_Status

Based on the Chi-square statistic of 5.412089360791851, 1 degree of freedom, and a p-value of 0.019997766888094164, we reject the null hypothesis. This indicates that there is a significant association between education level (graduate or not) and loan status. Education level may play a role in loan approval or rejection decisions.

Hypothesis 5: Self\_Employed vs Loan\_Status

The Chi-square statistic of 0.0, with 1 degree of freedom, and a p-value of 1.0, does not provide enough evidence to reject the null hypothesis. This suggests that there is no significant association between self-employment status and loan status. Being self-employed or not does not seem to impact loan approval or rejection decisions.

Hypothesis 6: Credit\_History vs Loan\_Status

With a Chi-square statistic of 138.21917360772596, 1 degree of freedom, and a very low p-value of 6.526142464733716e-32, we reject the null hypothesis. This indicates that there is a significant association between credit history (having a credit history or not) and loan status. Credit history plays a significant role in loan approval or rejection decisions.

Hypothesis 7: Property\_Area vs Loan\_Status

Based on the Chi-square statistic of 15.929602448472624, with 2 degrees of freedom, and a p-value of 0.00034748077291321715, we reject the null hypothesis. This suggests that there is a significant association between property area (urban, semiurban, or rural) and loan status. Property area may play a role in loan approval or rejection decisions.

Hypothesis 8: ApplicantIncome vs Loan\_Status

The t-statistic of -0.5076395514921876, with a p-value of 0.6119218009409668, does not provide enough evidence to reject the null hypothesis. This indicates that there is no significant difference in applicant income between approved and rejected loan applications. Applicant income does not seem to impact loan approval or rejection decisions.

Hypothesis 9: CoapplicantIncome vs Loan\_Status

The t-statistic of -0.13122884631313592, with a p-value of 0.8956450341804382, does not provide enough evidence to reject the null hypothesis. This indicates that there is no significant difference in coapplicant income between approved and rejected loan applications. Coapplicant income does not seem to impact loan approval or rejection decisions.

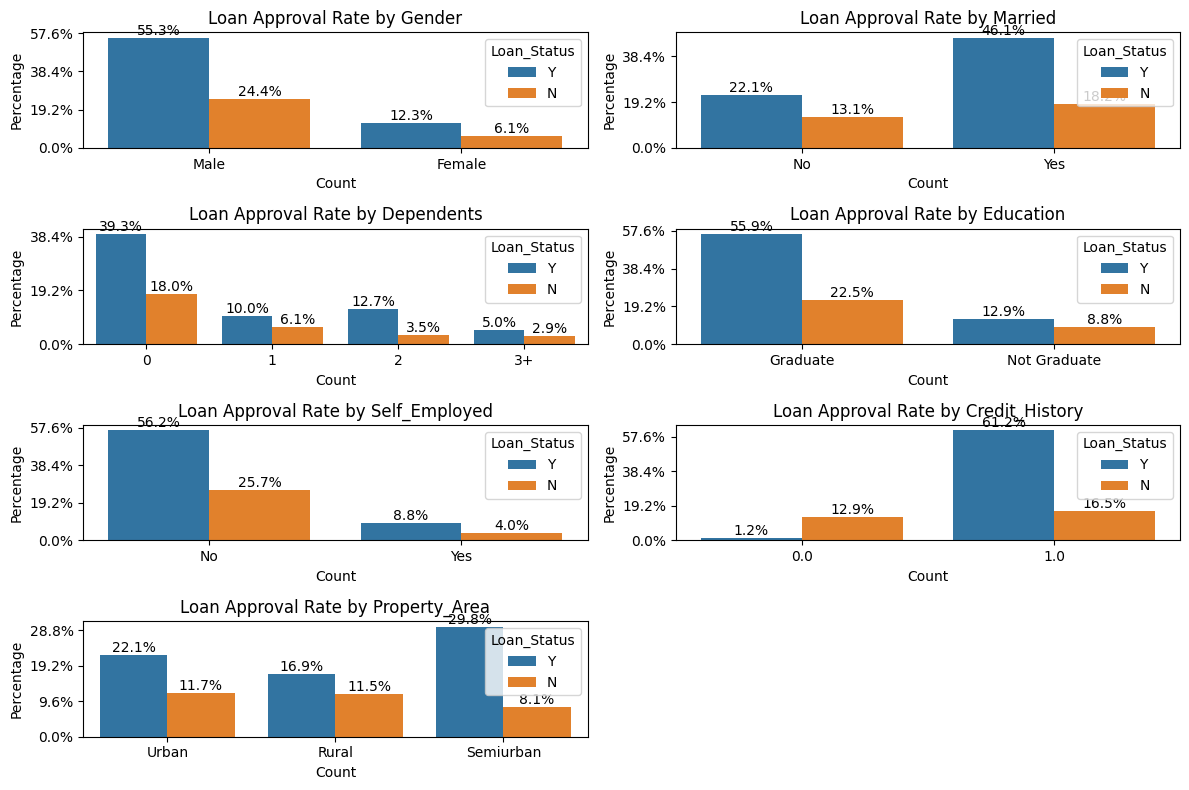
Hypothesis 10: LoanAmount vs Loan\_Status

The t-statistic of -0.49898976161448045, with a p-value of 0.6180740485664924, does not provide enough evidence to reject the null hypothesis. This suggests that there is no significant difference in loan amount between approved and rejected loan applications. Loan amount does not seem to impact loan approval or rejection decisions.

Hypothesis 11: Loan\_Amount\_Term vs Loan\_Status

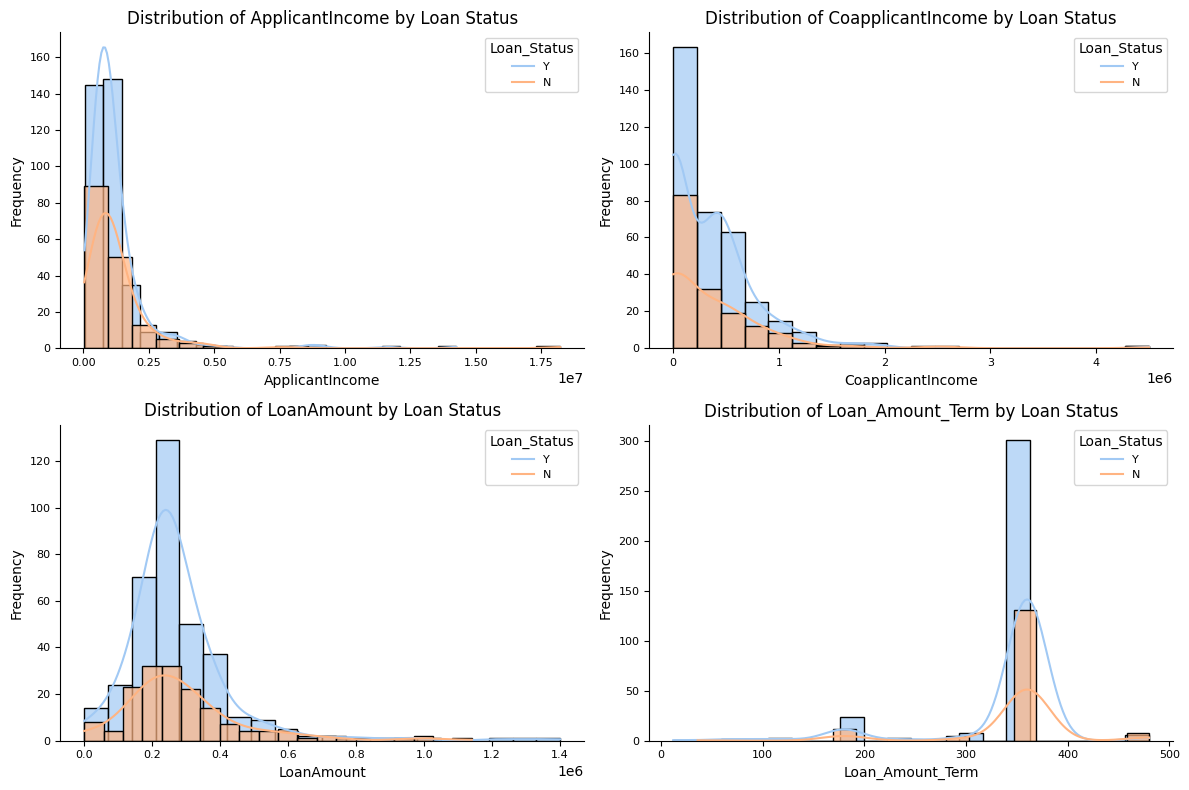
The Chi-square statistic of 4.744131257058896, with 5 degrees of freedom, and a p-value of 0.45200112296476167, does not provide enough evidence to reject the null hypothesis. This indicates that there is no significant association between loan amount term (in months) and loan status. Loan amount term does not seem to impact loan approval or rejection decisions.

In conclusion, based on the statistical tests performed, the results suggest that **marital status (Hypothesis 2), education level (Hypothesis 4), credit history (Hypothesis 6), and property area (Hypothesis 7)** are significantly associated with loan status. However, gender (Hypothesis 1), number of dependents (Hypothesis 3), self-employment status (Hypothesis 5), applicant income (Hypothesis 8), coapplicant income (Hypothesis 9), loan amount (Hypothesis 10), and loan amount term (Hypothesis 11) do not seem to have a significant impact on loan approval or rejection decisions.

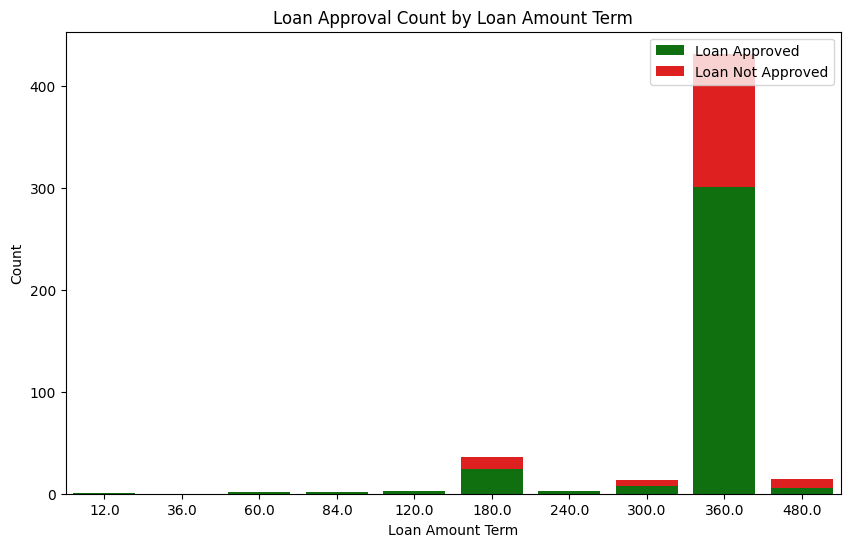


Folowing insights were drwan from the above bar charts:

* More number of loan applications are for male as compared to female. However ratio of loan approved to loan rejected for male is almost similar that of loan approved to loan rejected for females. So gender should not have a significant impact on the loan approved status as suggested in Milestone 1.
* More number of loan applications are inclined towards customers who are married. Also loan approval chance of married customers are higher as compared to non married as suggested in Milestone 1.
* The dependant feature suggests that more number of loan applications are for customers who do not have any dependants, which indicates that bank’s prefer aloan applications for customers who do not have dependents.
* As suggested in Milestone 1, the loan application of garduate customers are considerably higher as compared to non graduate counterparts. Also the chances of loan approval for graduate customers are higher as compared to non graduate customers.
* For Self Employed feature the bar graph plotted shows that financial institution in general prefers non self employed customers. This is because banks are more inclined towards salaried customers who have a stable source of income. The same hypothesis was given in Milestone 1 as well.
* The credit history seems to be one of the most important feature. Clearly more applications and more loans are approved for customers having credit history. This may be because the company looks at the credit history of applicants and prefers applications of customers having a good credit history. Even hypothesis testing suggested that credit history has significant impact on the target variable.
* The loacation of the property as per the data set doesn’t highlight significant conclusion. The graph seems to be almost equally distributed for urban, semi urbam and rural. However, more loan application are foccused on semi urban and then urban as compared to rural locations. The same trend is observed for ratio of loan approval to loan rejection with semi urban being the highest and rural being the lowest.



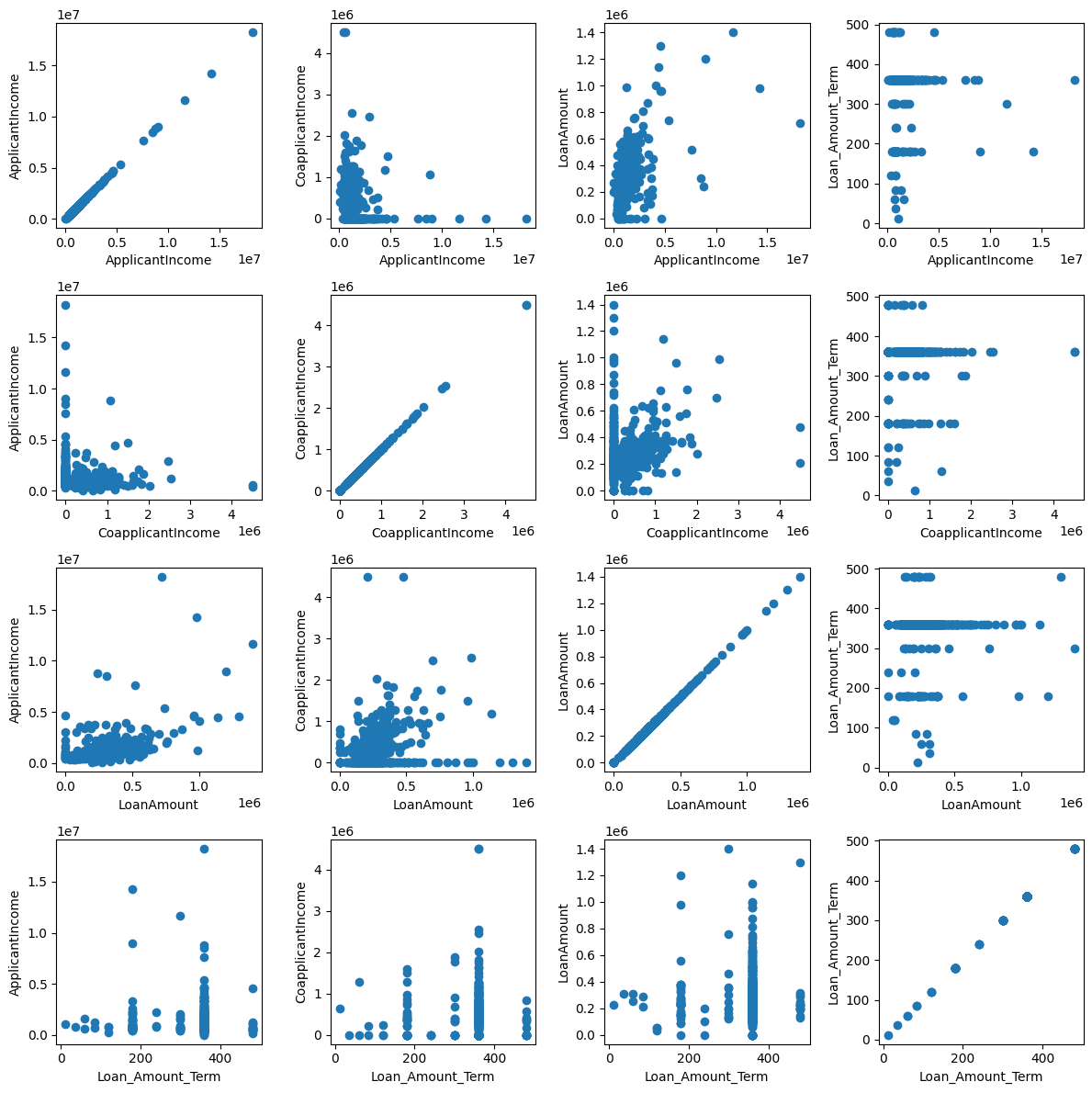
For continuous independent features histograms were plotted to observe the trend and the following results were obtained:

1. The company focusses applicants mainly from the lower income group . The income histogram by loan status graph is highly positively skewed indicating the that number of applicants , applicants whose loans were approved , applicants whose loan were rejected is primarily from the lower income region; with a few applicants having higher incomes that are pulling the distribution towards the right. This is not in terms with the affect suggested in Milestone 1 where it was suggested that higher income individuals may have higher chances of getting the loan.
2. The co applicant income also shows similar trend as that of applicant income. Lower Co applicant income applicants are more focussed by the company.
3. The curve of loan amount is again positively skewed but better and less skewed from the above two features. The graph indicates that the majority of loan amounts are clustered in the region between the lower end and middle of the distribution, with a few larger loan amounts extending towards the higher end, both for loan approved and rejected.
4. For loan term, since the values present under this feature were limited, a stag bar char chart was further created to find a more proper insight apart from the histogram. A very important result was found out from the graph of this feature. The graph clearly shows that majority of the applicants chosen were for higher loan terms like 180, 300, 360 and 400, with 360 months being huge in number. This also suggests that the type of loan have a good chance of property or housing loan as such higher terms are usually given to housing loans.

Correlation and Heat Map

To understand the correlation of the featues further aheat map was created which indicates that loan amount is having a very high correlation with applicant income column. This may be because majority of the loan amounts and applicants income are concentrated in the lower region. Similar conclusion may be drawn for co- applicant income and loan amount which is alos having a good correlation value.





Scatter Plots:

Scatter plots were also plotted to understand the hypothesis drawn from heat map, and also to study the effect of each continuous independent feature on each other.

Results revealed that majority of the applicants are having lower income with few people having a higher income; Majority of the co applicants’ income are also concentrated in the lower region, The cluster of loan amount is also concentrated in the lower end of the region with few exceptions