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.

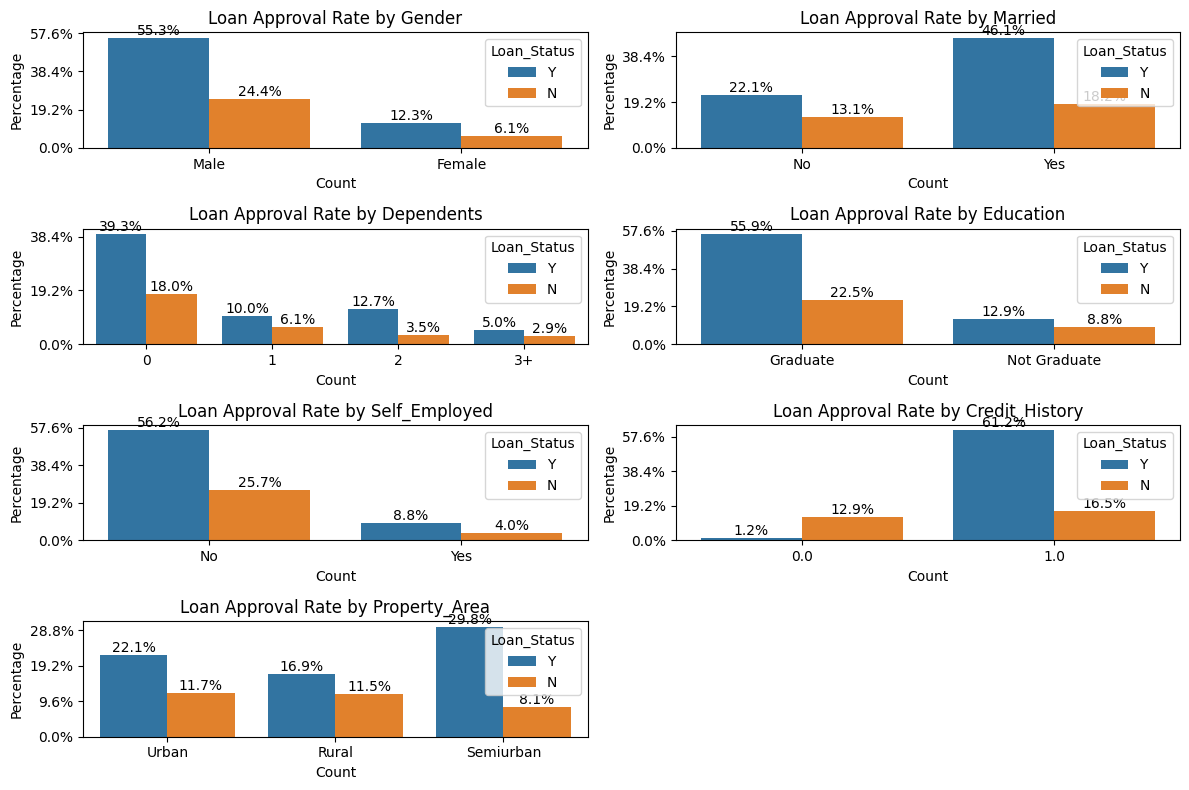
Hypothesis 10: LoanAmount vs Loan\_Status

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

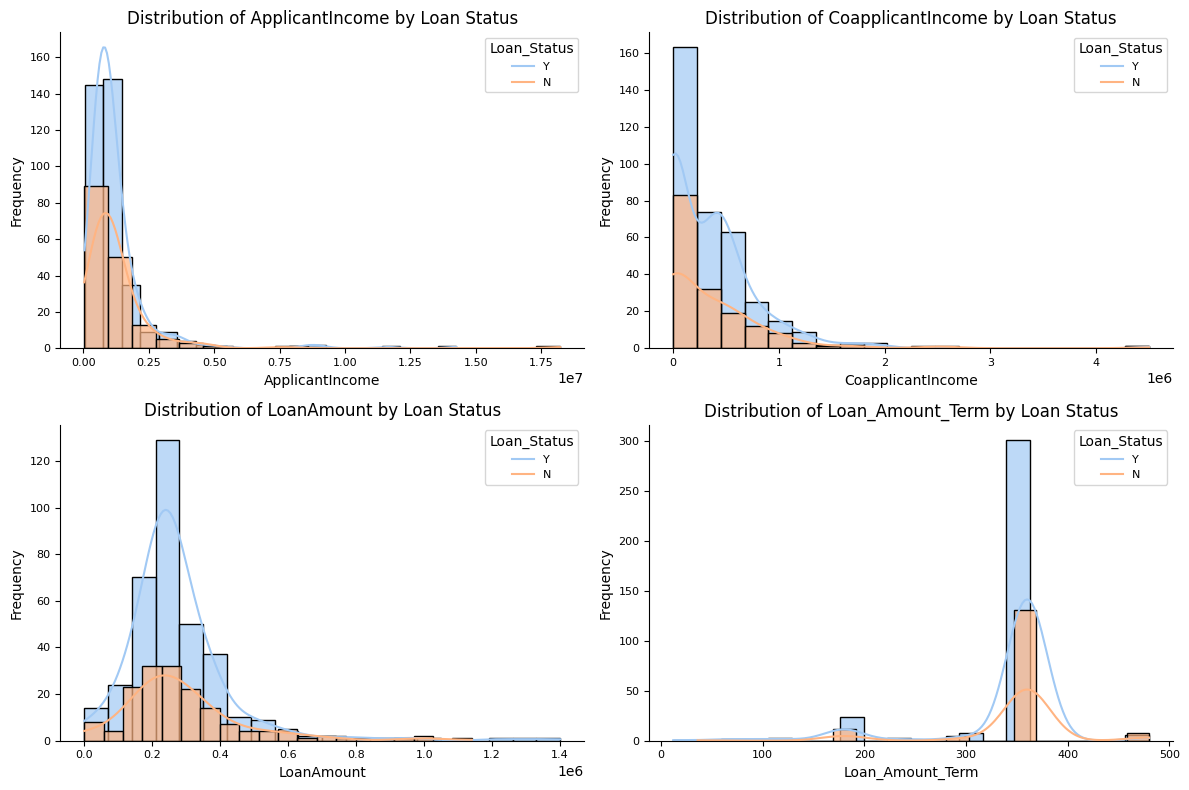
The Chi-square statistic of 4.744131257058896, with 5 degrees of freedom, and a p-value of 0.45200112296476167, does not provide enough evidence to reject the null hypothesis. This indicates that there is no significant association between loan amount term (in months) and loan status. Loan amount term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

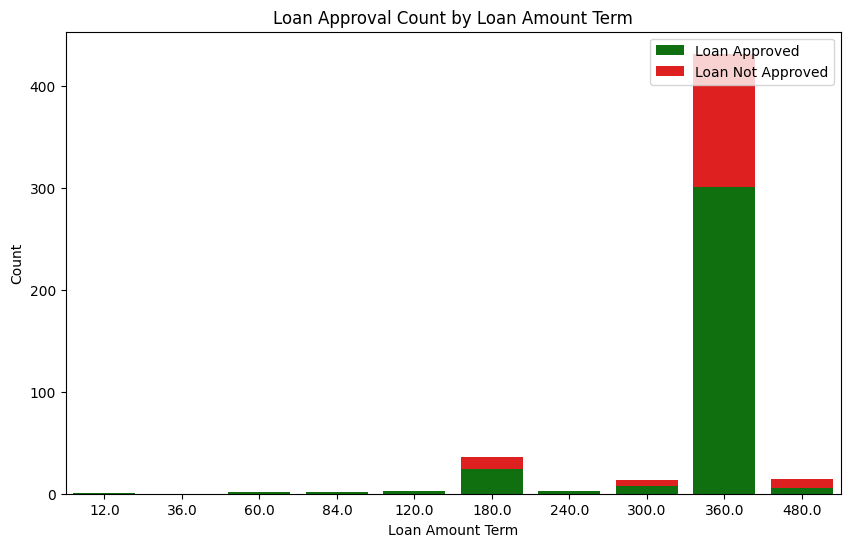


Folowing insights were drwan from the above bar charts:

* 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.



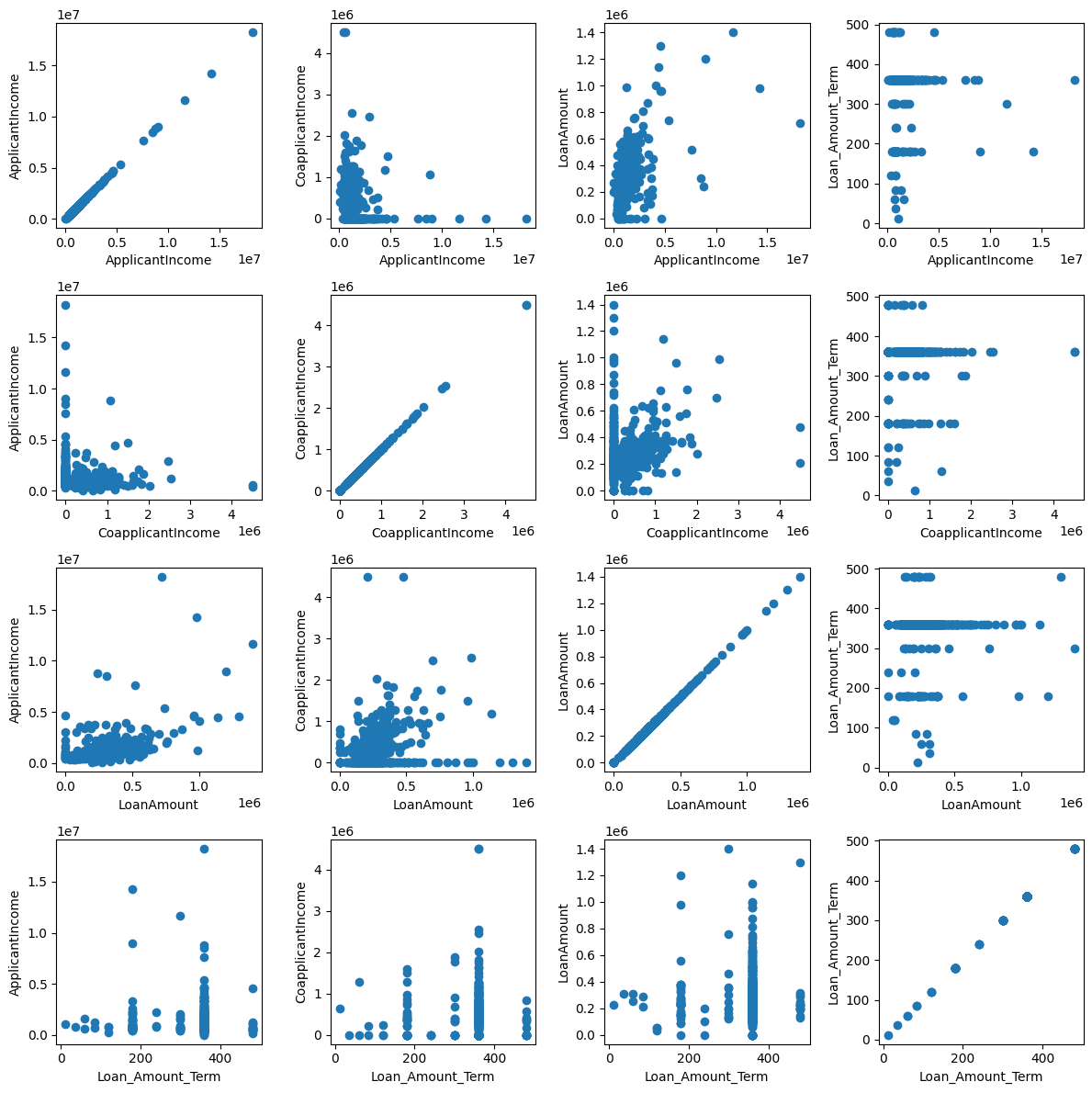
For continuous independent features histograms were plotted to observe the trend and the following results were obtained:

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.