Lending Club Case Study

PRABUSHANKAR CHINNASAMY

PRIYA BHALA

Problem Statement

- Risk assessment: Identify the factors or combination of factors that contribute to loan defaulting. From the given data set, analyse and conclude the set of features that will help raise a red-flag for a potential defaulter. They can be denied a loan, or lessen the fund or increase the interest rate or all/any of the above as decided by the bank
- Good customer identification: Apply the learning from risk assessment study to identify good customers who are very unlikely to default.

Approach

Data Preparation:

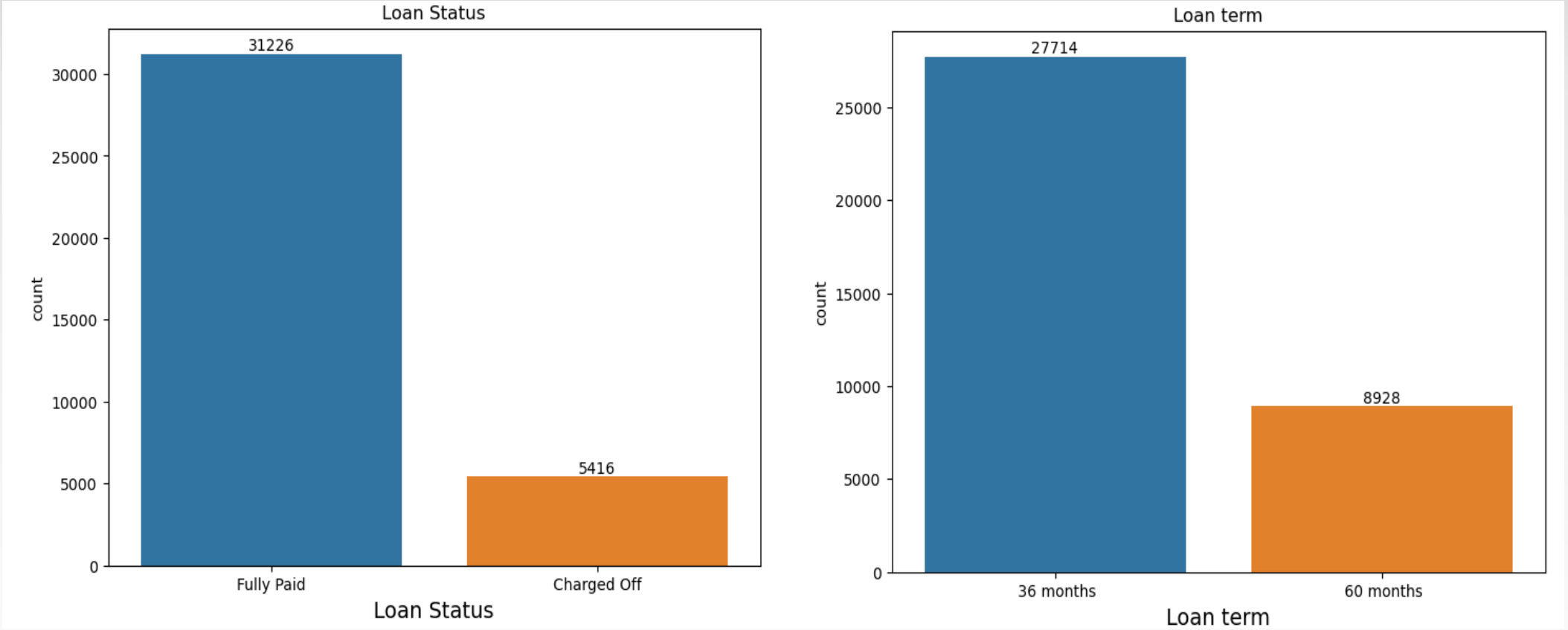
- Null value treatments
  - Remove the columns with all-null, redundant or irrelevant values
  - Fill NA with relevant aggregation method (mode for categorical and median for numerical)
- Check for duplicates and outliers
- Remove irrelevant columns and derive new columns
- Standardize Units

Analyze: start with each feature, combine explore them till some reasonable inferences show up

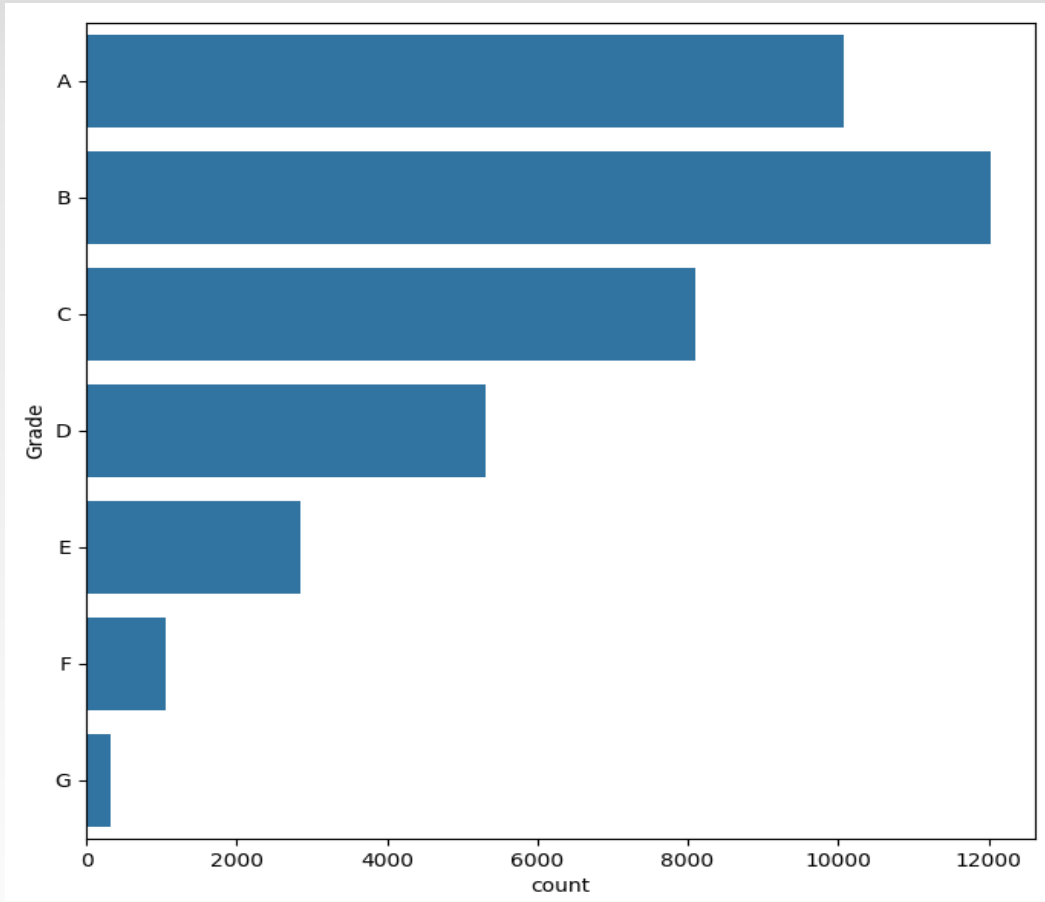
- Univariate – one variable at a time. Mostly look for the distribution and the statistical aspect of the data
- Bivariate– Mix and match to see how one feature affects the other and helps with problem statement.

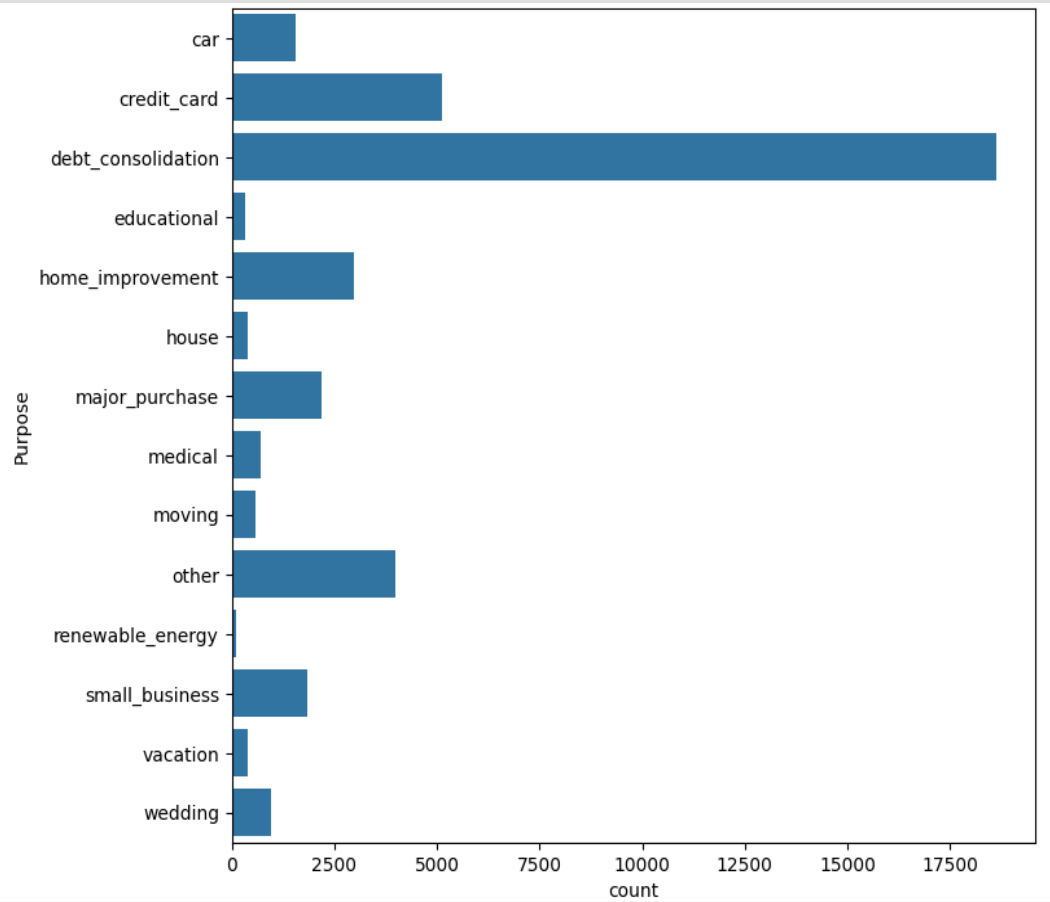
Conclude: List down the learnings and findings from this data analysis.

Univariate Analysis

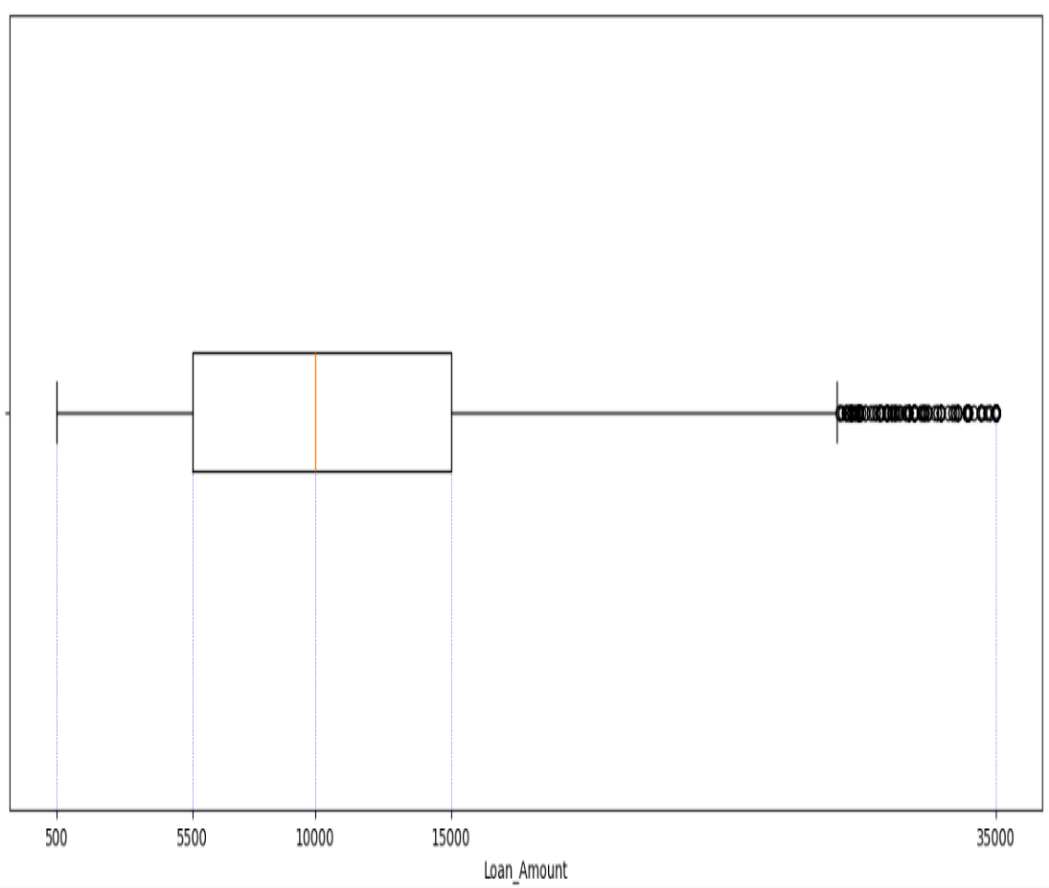


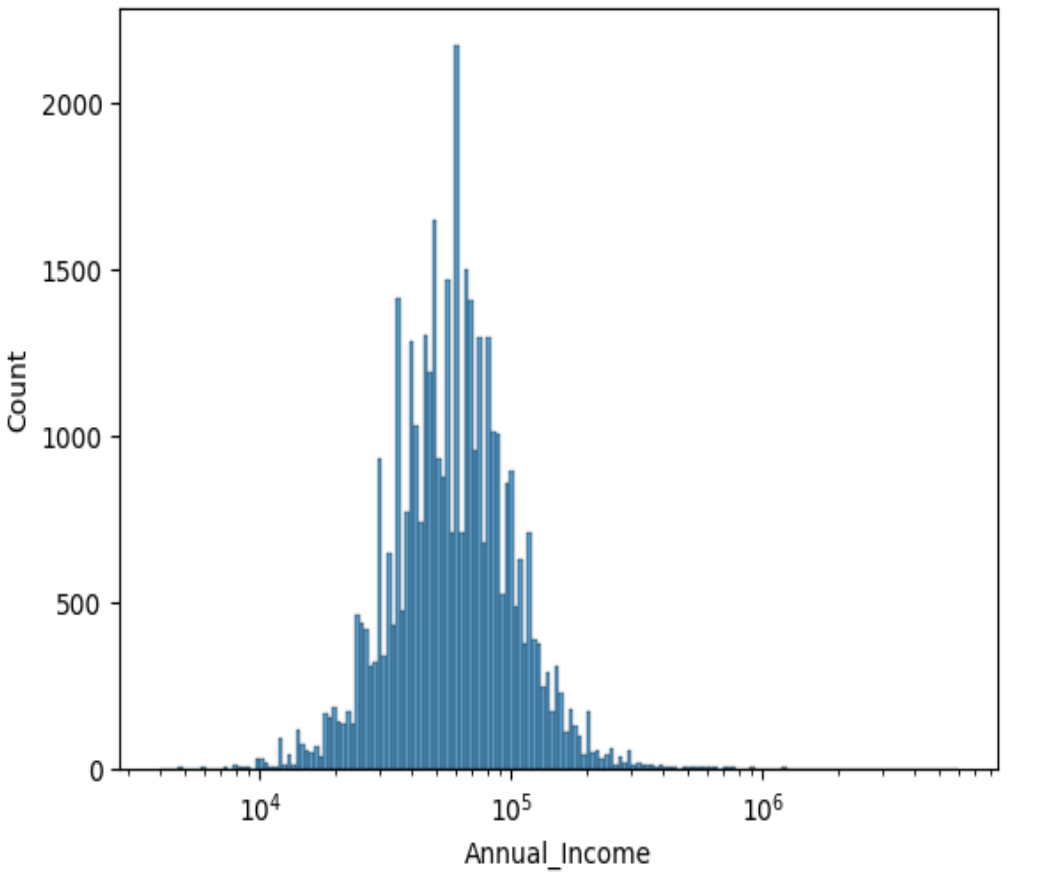
Univariate Analysis





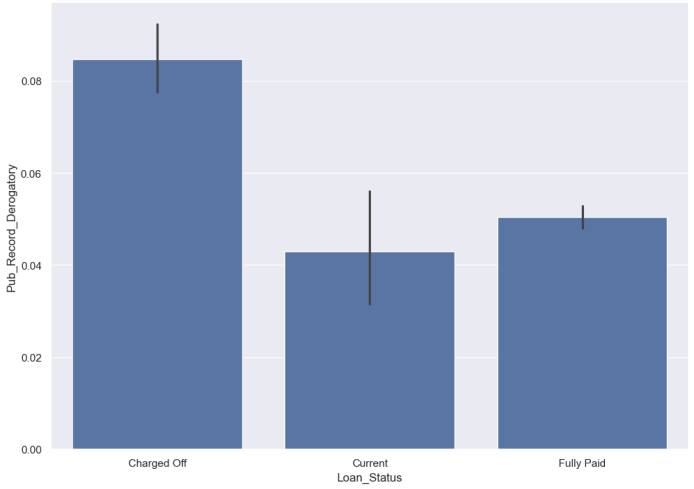
Univariate Analysis

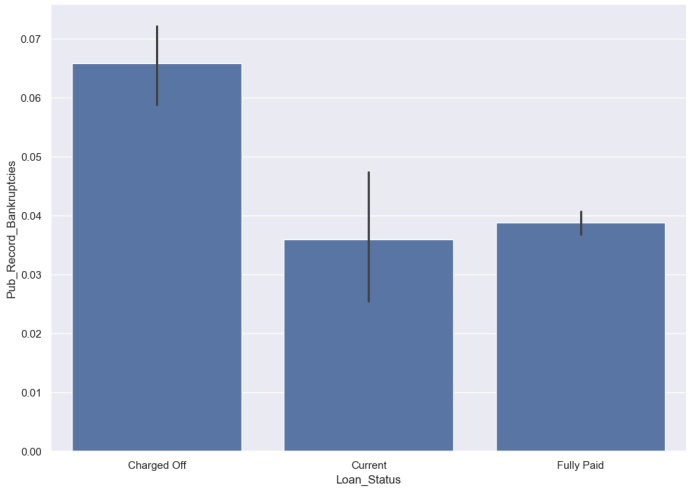


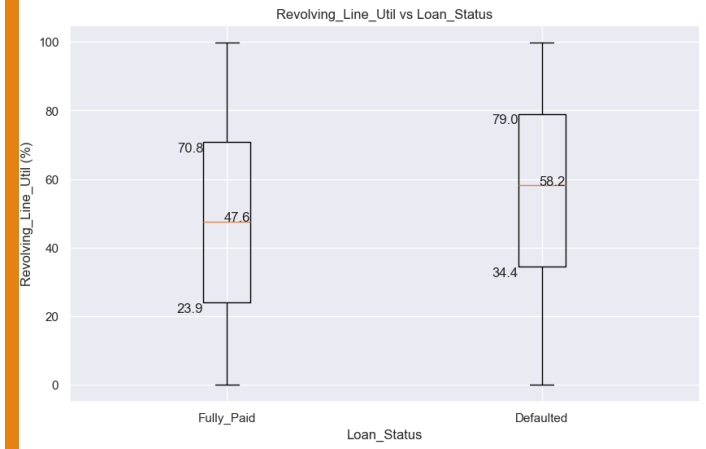


Univariate Analysis Observation

- There are more ‘Fully Paid’ and very less ‘Charged off’ values for *loan_status* column which makes this dataset imbalanced.
- From the bar chart, it is clear that there preferred loan *term* is 36 months as against 60 months
- Customers with *grade* B, A and C are high in number (in the same order).
- Customers quote ‘debt_consolidation’ followed by ‘credit_card’ as primary purpose for availing the loan
- On an average, most customers borrowed in the range of $5,500 and $15,000
- There are outliers present in the *loan_amount* column
- There are a large number of outliers in the *annual_income* with highest being $6 Million and lowest being $4,000

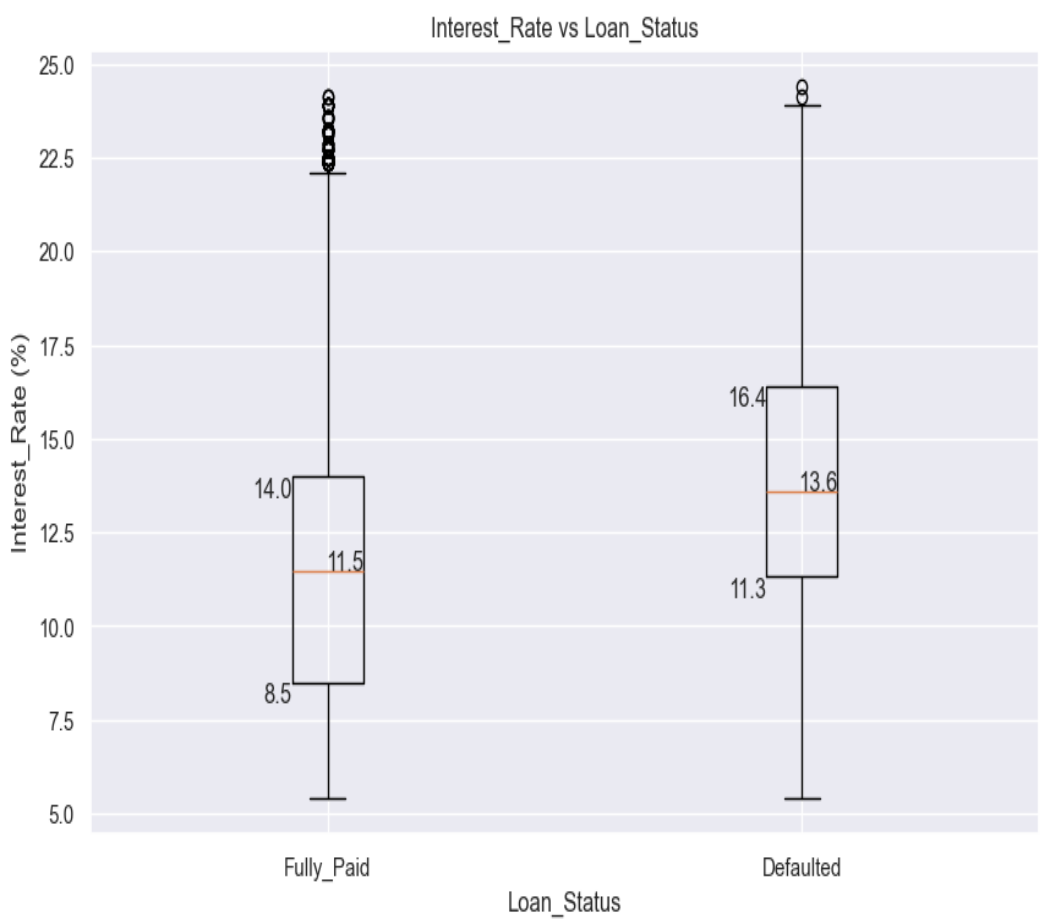


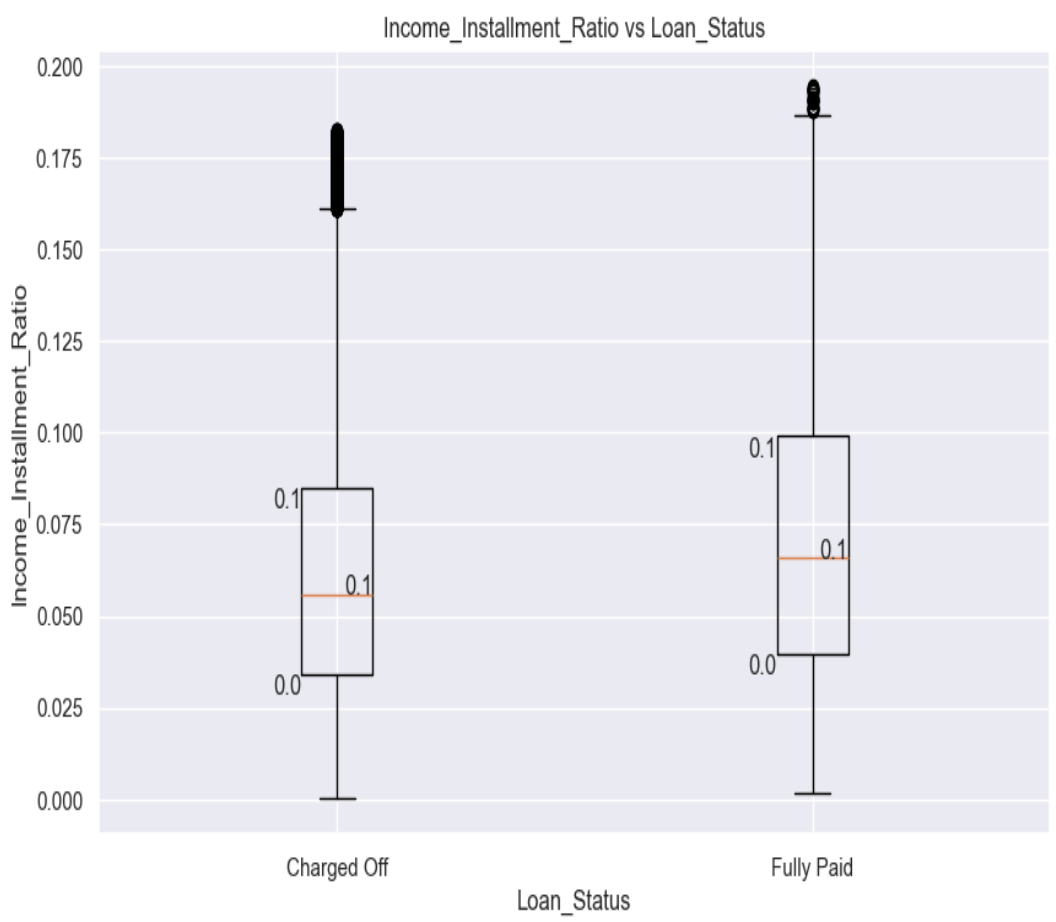




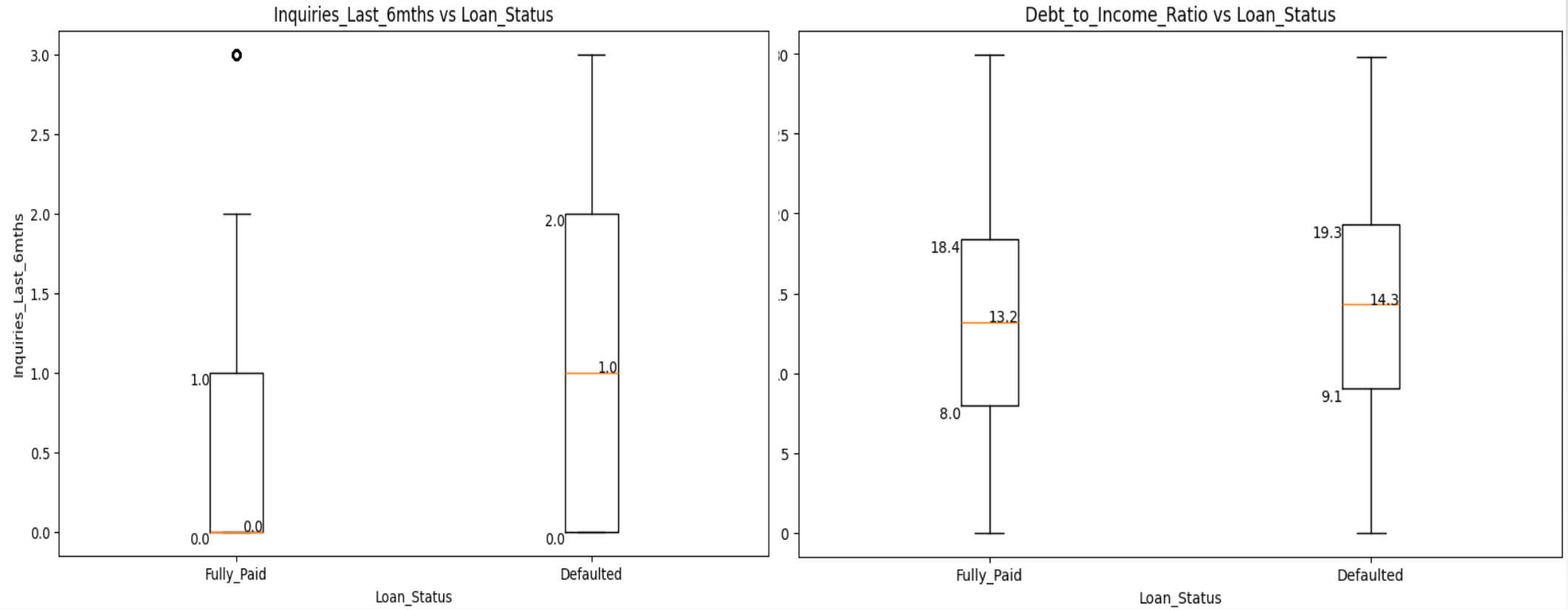
Bivariate Analysis

Bivariate Analysis





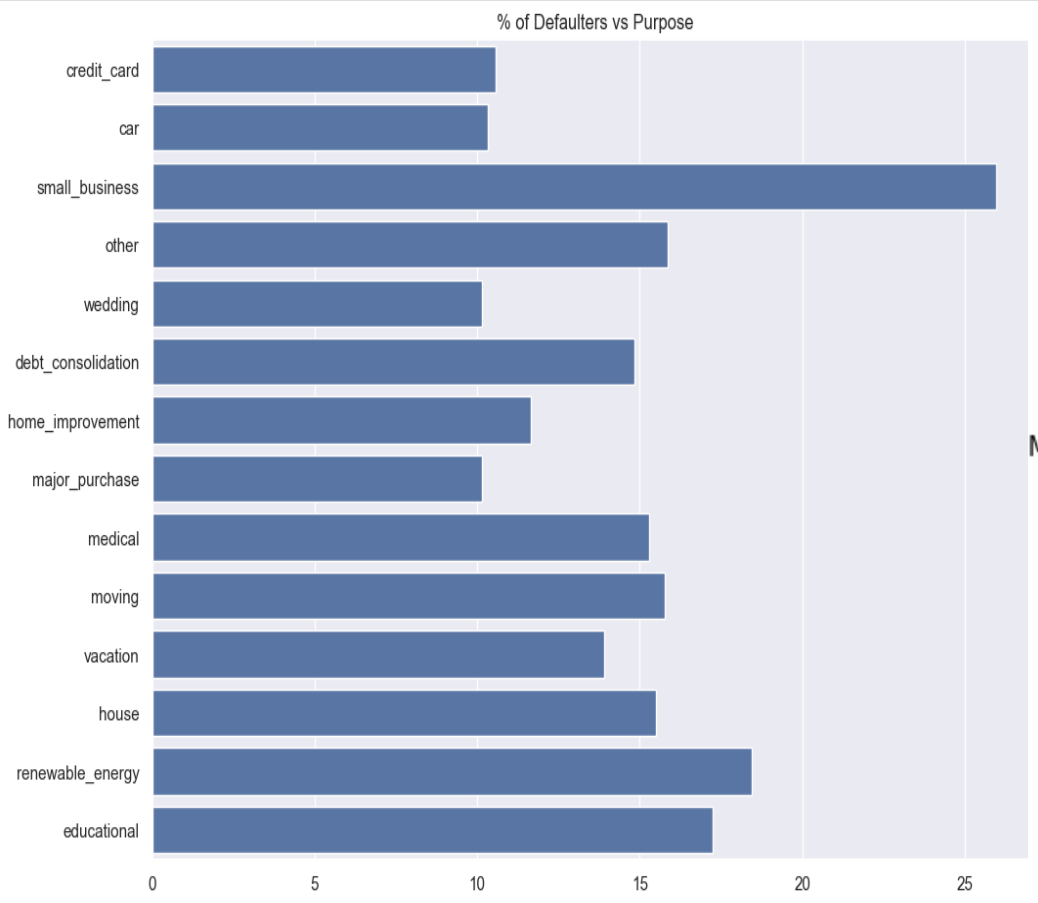
Bivariate Analysis

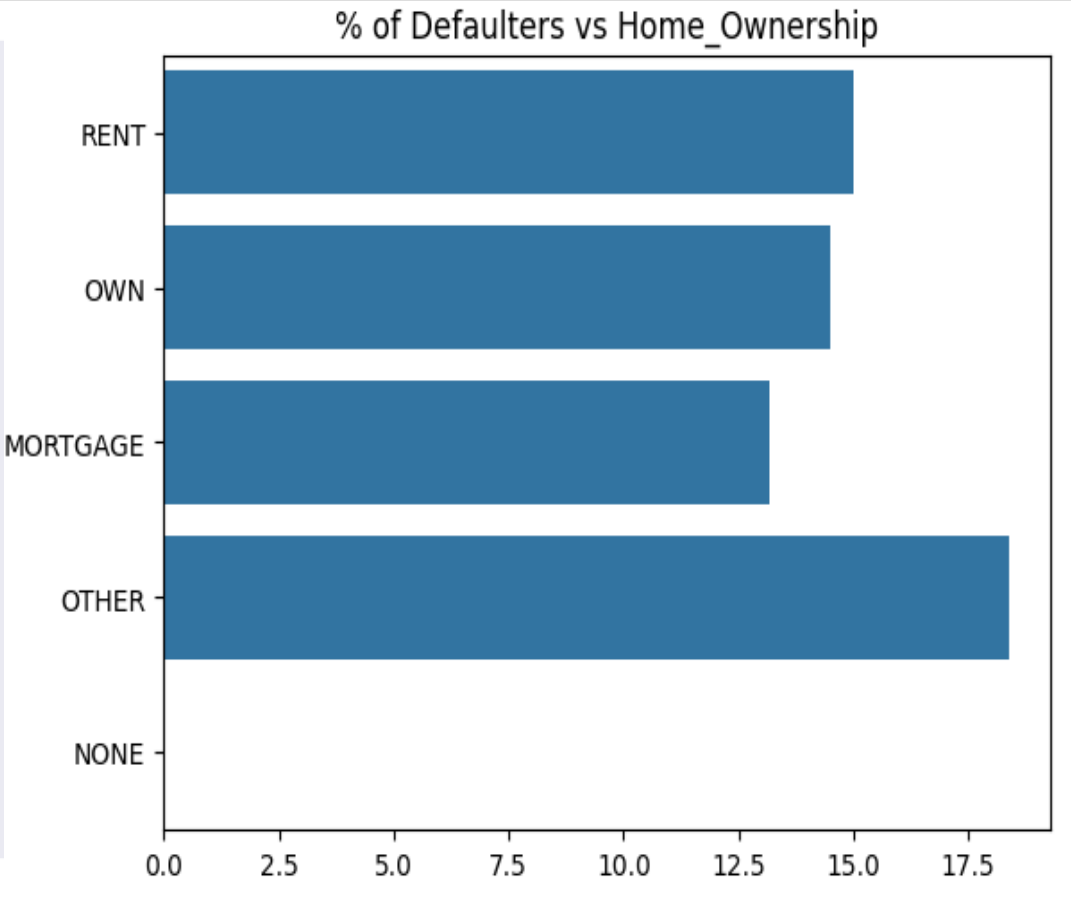


Bivariate Analysis Inference

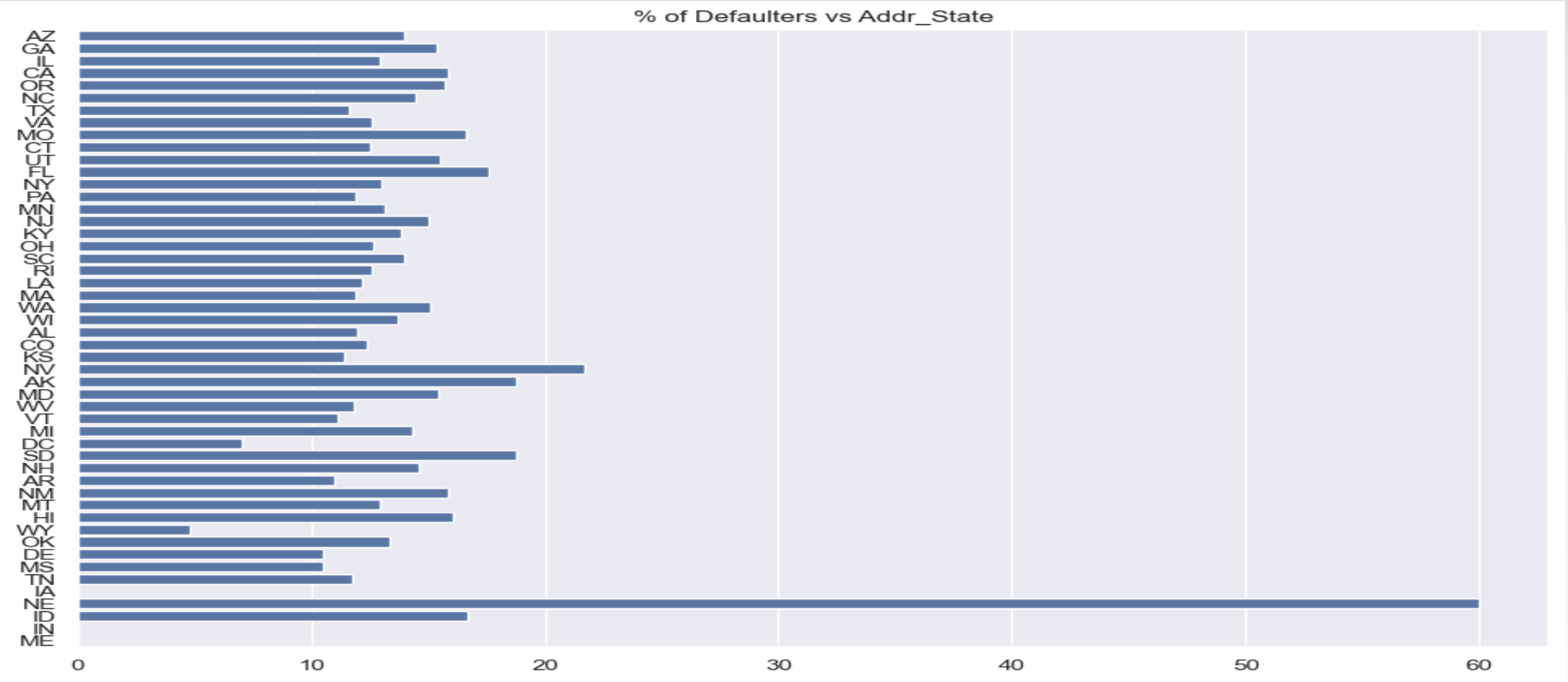
- Interest rates for the defaulters are high compared to the non-defaulters.
- 6% of the Defaulters have at least a record of bankruptcy against only 3% of the fully paid customers.
- 8% of the Defaulters have at least a record of derogatory against only 4% of the fully paid customers.
- It is observed that people who are not verified are paying loans compared to verified
- Debt_consolidation is highest among the purpose why people are taking loan followed by credit card and they tent to default more as well
- Most number of defaulters carry a revolving_line_util of 34% to 79%. This is against a recommended 30%
- Installment-to-Monthly-Income ratio(derived column) is high for defaulters.
- The defaulters have made more number of inquiries with various credit institutions before successfully availing a loan in this bank.

Segmented Analysis





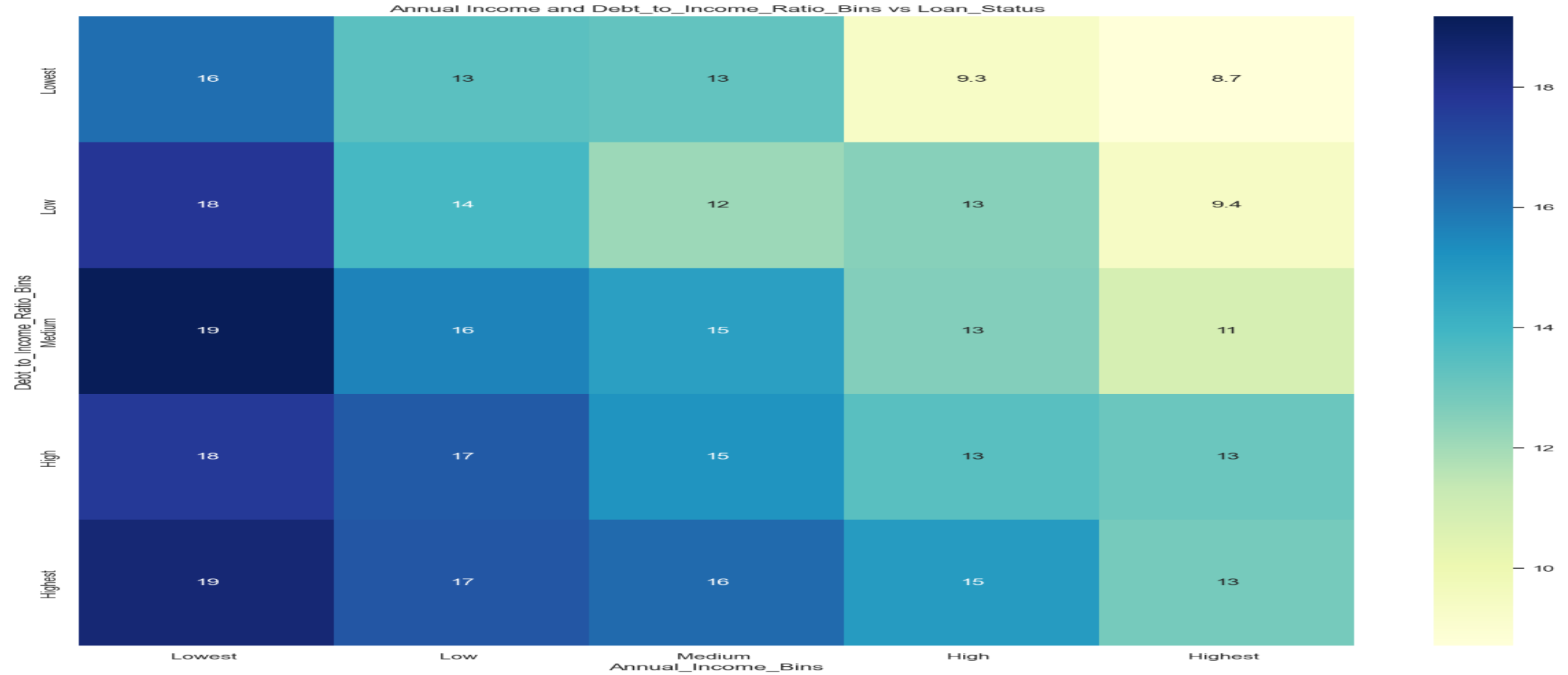
Segmented Analysis



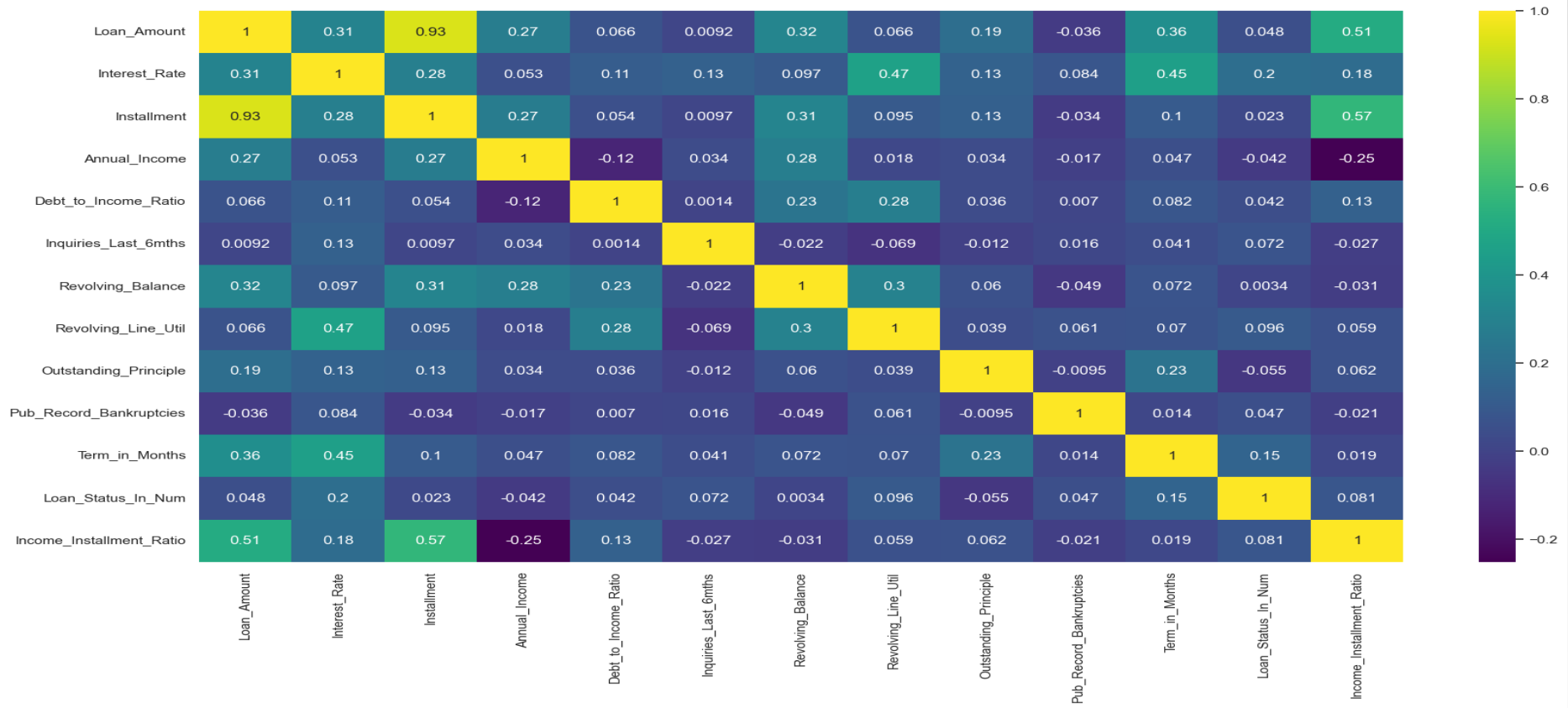
Segmented Analysis Inference

- The data above shows that most defaulters hold at least a public record of Derogatory
- It also shows that they hold at least 2 records of Bankruptcy
- Customers who tend to have borrowed for 'Small Business' top the list of defaulters
- This list is followed by Renewable energy and Education
- Though few customers from Nevada, there seems to be more defaulters, at a distance followed by Nevada and Arkansas at close range.

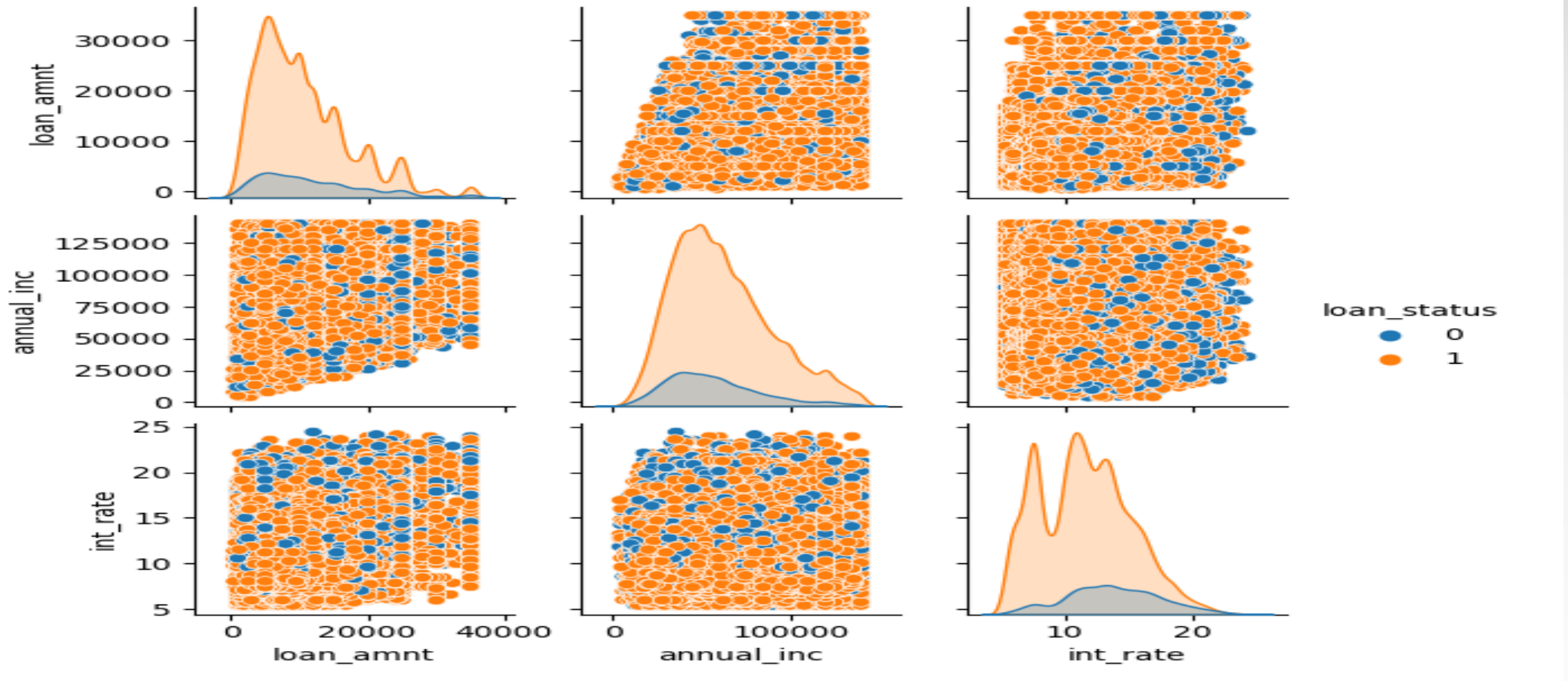
Multivariate Analysis



Multivariate Analysis



Multivariate Analysis



Multivariate Analysis Inference

- Few independent variables are strongly correlated to each other
- *annual_income* is negatively correlating with Debt-to-Income ratio
- Interest rate is correlating with *loan_status*. More the *interest_rate* more the defaulters.
- There is a correlation between *pub_record_bankruptcies* and *pub_record_derogatory*
- The high debt-to-income ratio combined with Low Income range contribute to more defaulters Highest income range and lowest debt-to-income ratio makes better candidates for good loans.

Conclusion

1. Public record of bankruptcy or derogatory is a very big indication of the potential defaulter.
2. Most number of defaulters carry a revolving line utility of 34% or above. This is against a recommended 30%.
3. The annual income of the fully paid are slightly higher than the defaulters. Low income paired with higher interest rates is a deterring factor for the non-repayment.
4. The defaulters have made more number of inquiries with various credit institutions and were not entertained. This is an indication that those banks see a red flag in these customers.
5. Though the numbers are less, people with 'Other' type of home ownership tend to default more.
6. Customers who tend to have borrowed for 'Small Business' top the list of defaulters. This is followed by Renewable energy and Education.
7. Though few customers from Nevada, there seems to be more defaulters, at a distance followed by Nebraska and Arkansas at close range.
8. The high Debt-to-income ratio combined with Low Income range contribute to more defaulters
9. Interest rate of the defaulters are higher than their fully-paid counter parts. This could be as a result of loss mitigation towards potential defaulting.
10. Highest income range and lowest debt-to-income ratio makes better candidates for good loans.