GRAMENER CASE STUDY

SUBMISSION

Group Member Names:

1. Pradeep Meriga
2. Parthasarathy Vijayan
3. Feta Joseph
4. Rama Kumar Jayanti

Data Exploration

Overview:

Loan dataset contains complete loan transactions issued, including loan requested amount, sanctioned amount with interest rate and monthly instalment amount etc…, for the years between 2007 – 2011 and its corresponding Borrower information in terms of Job, Salary, Credit report, Bankruptcies filing numbers etc..., with respect to the zip code and the state the borrower belongs to.

Loan dataset has 39717 observations of 111 variables.

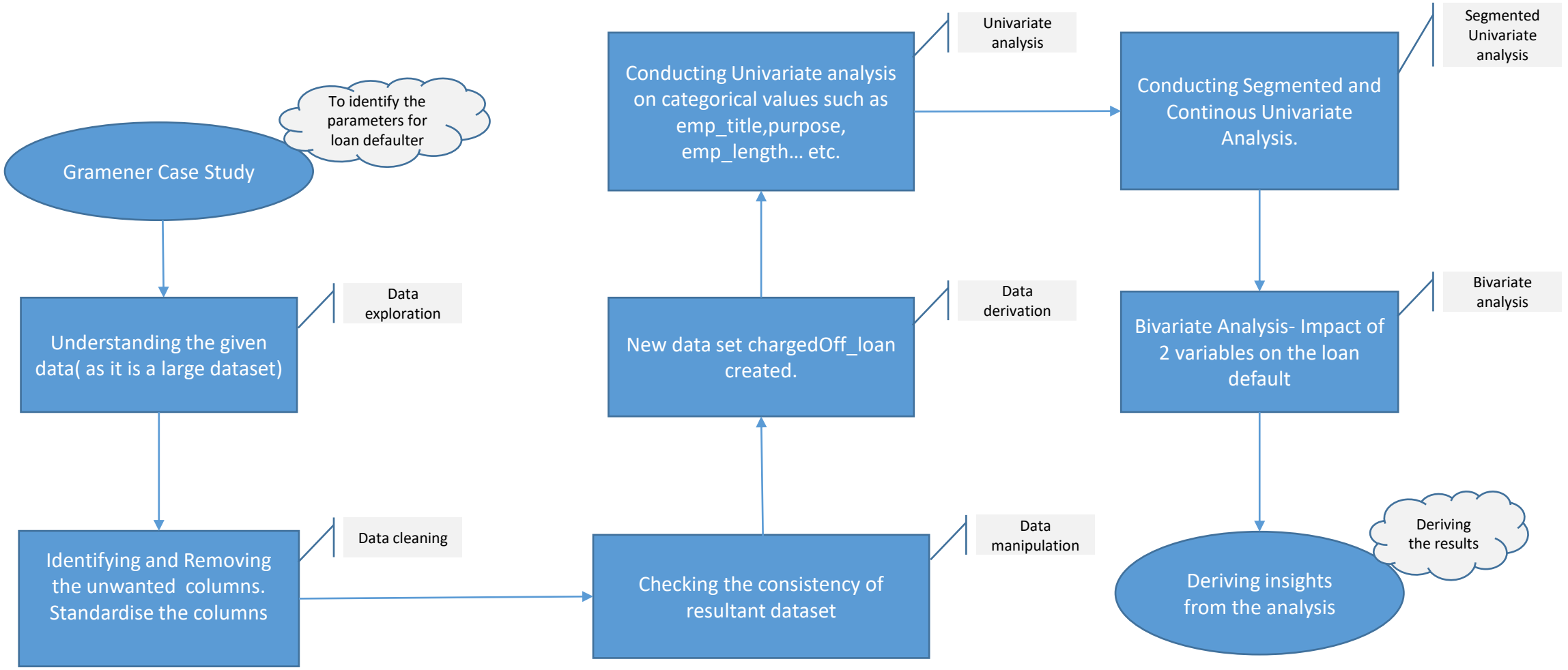
Problem Statement:

To provide insights to the online banking company with the insights of the parameters/variables which would influence the risk of loan defaults (financial loss).

Analysis:

For the Charged Off status, check the pattern of the loan request details and borrowers details to find the variables which could influence the high risk of default scenario.

Process Solving Methodology



Data Cleaning and Manipulation

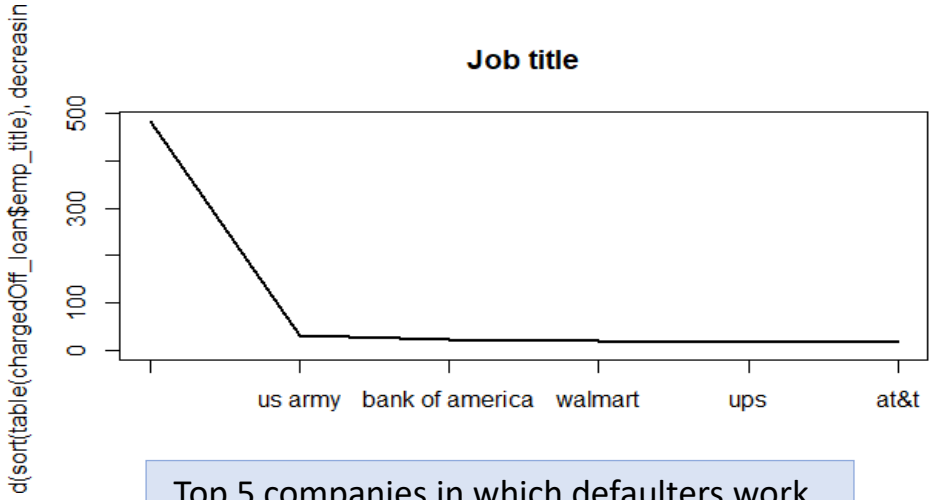
Data Quality Issue:

- Fixed columns which contains unique values
- Removed columns which are not needed for EDA
- Renamed few column names to have consistency
- Standardised the column values by such as removing the term “months” ,”%”
- Given date format are converted to the standard format
- Different cases of words have been standardised to lower case

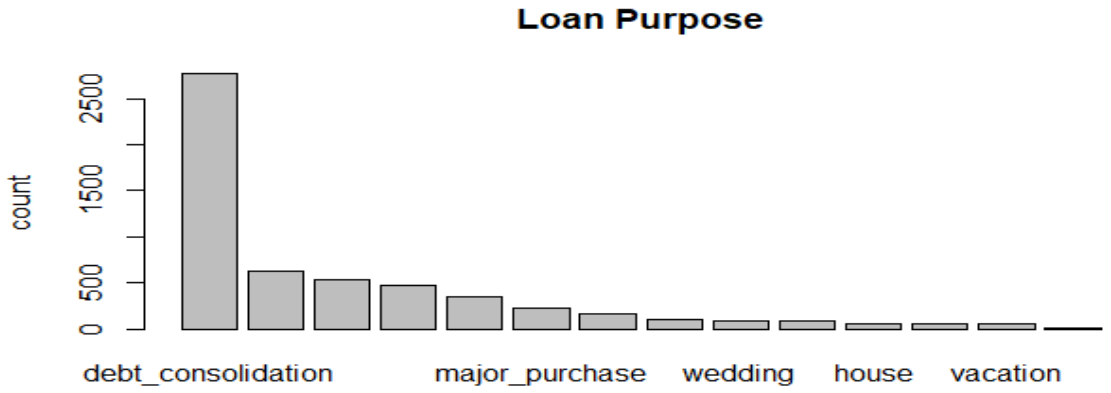
Data derived for Analysis:

New dataset **chargedOff_loan** contains only Charged Off status row items

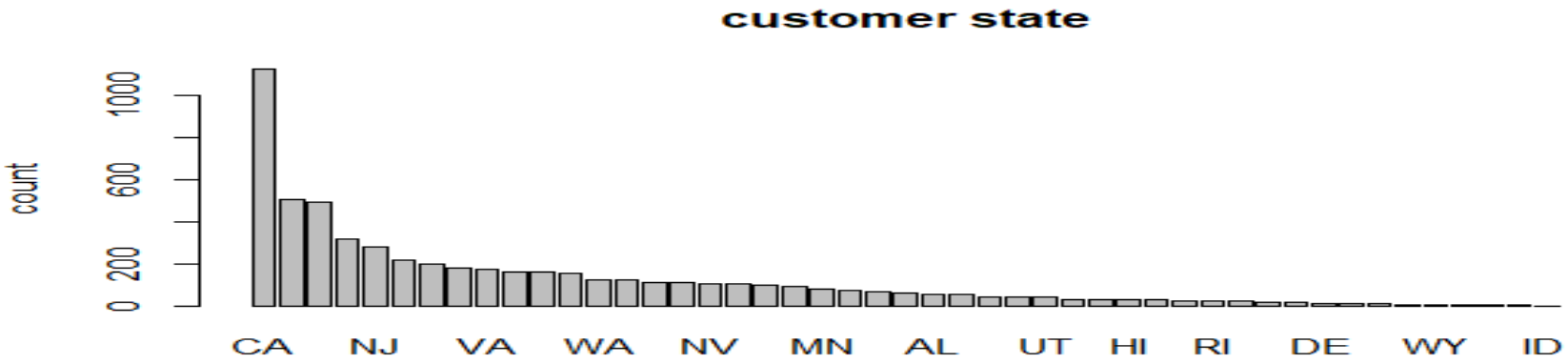
Univariate Analysis – Unordered Categorical



Top 5 companies in which defaulters work



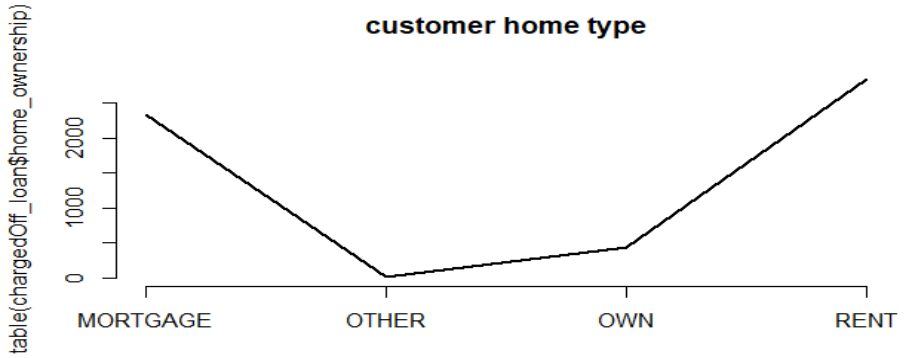
Debt_consolidation purpse clearly shows the higher trend of loan default



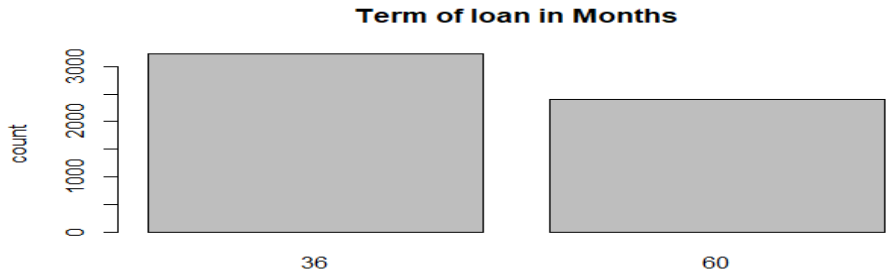
Customers from California state has higher chances of loan default.

Univariate Analysis – Ordered Categorical

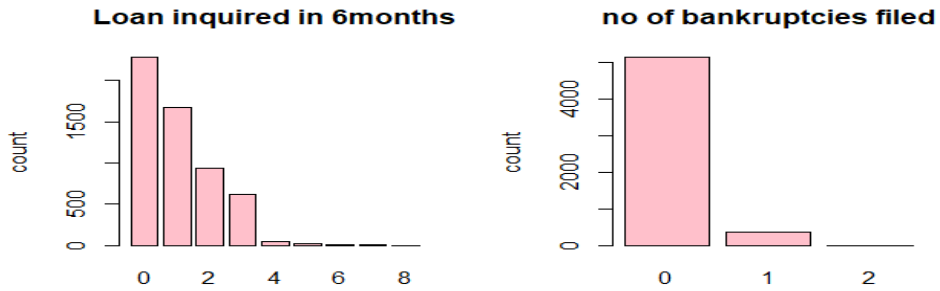
Customers with 10+ years are the highest defaulters



Customers who are in Rent house are the highest defaulters, followed up by customers house which are in Mortgage.

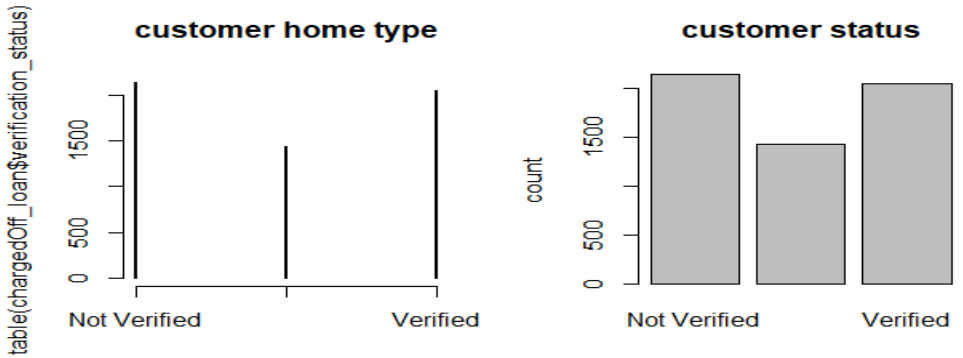


Lower the tenure, higher the risk.
36 months has higher loan default rate

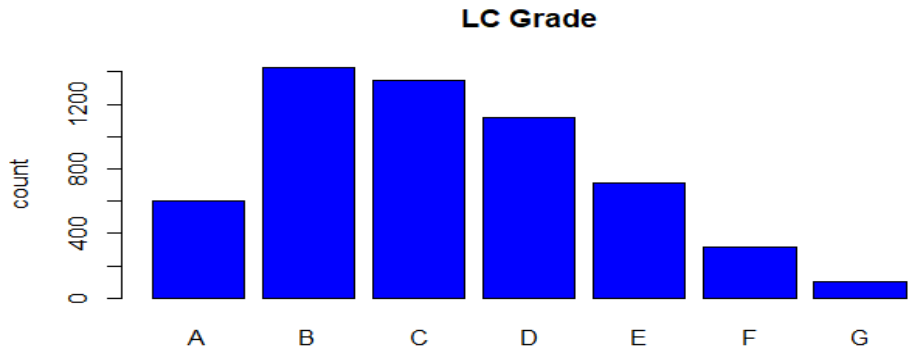


Customers having >1 should be flagged as high risk customers

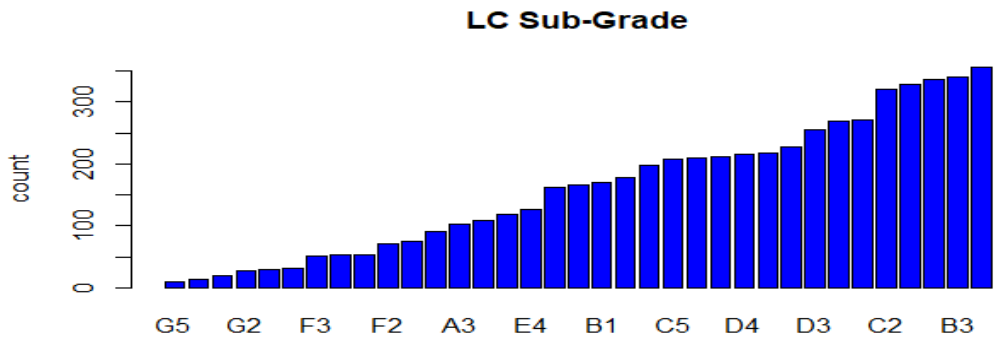
Univariate Analysis – Contd.



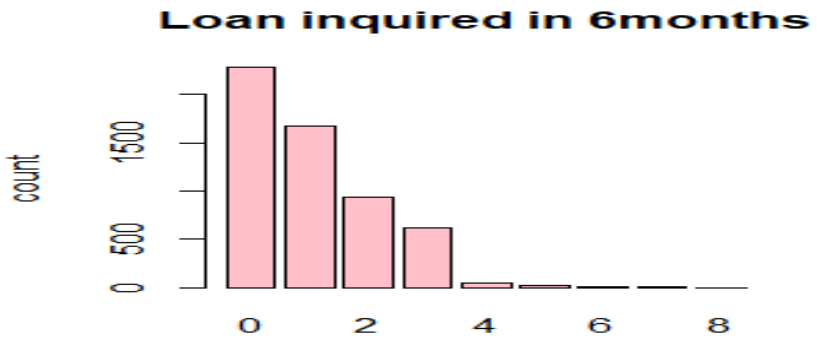
Almost both verified and not verified are equal, cannot infer any result



B grade has the top risk , followed by C grade

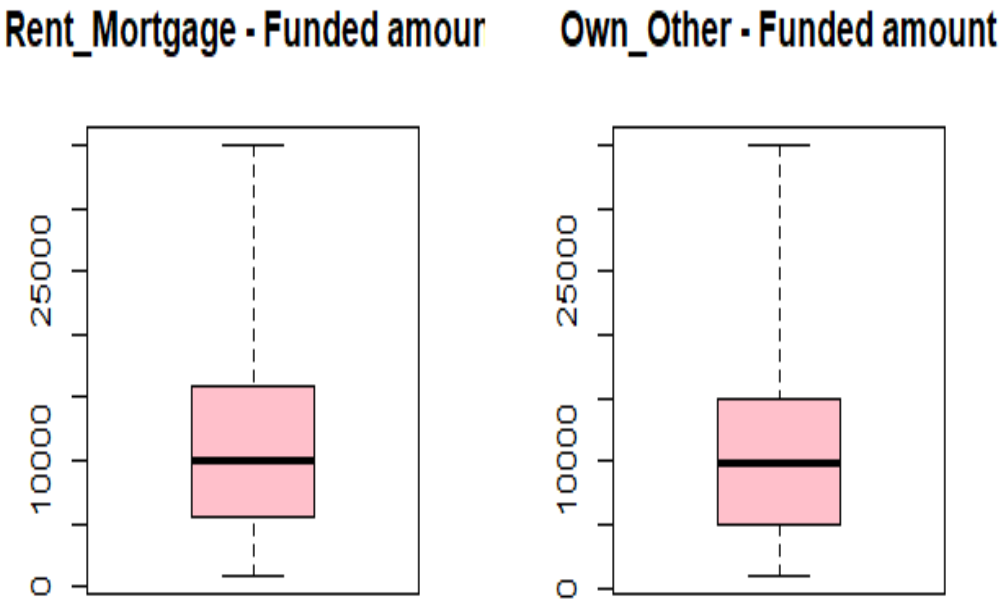


B5 grade has the top risk, followed by B3 grade

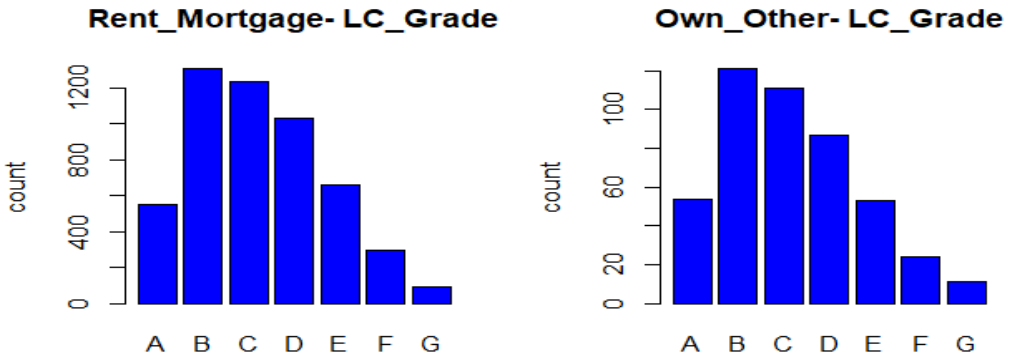


Customers with more inquiry numbers are high risk, inq_last_6mths > 3 can be flagged as risk

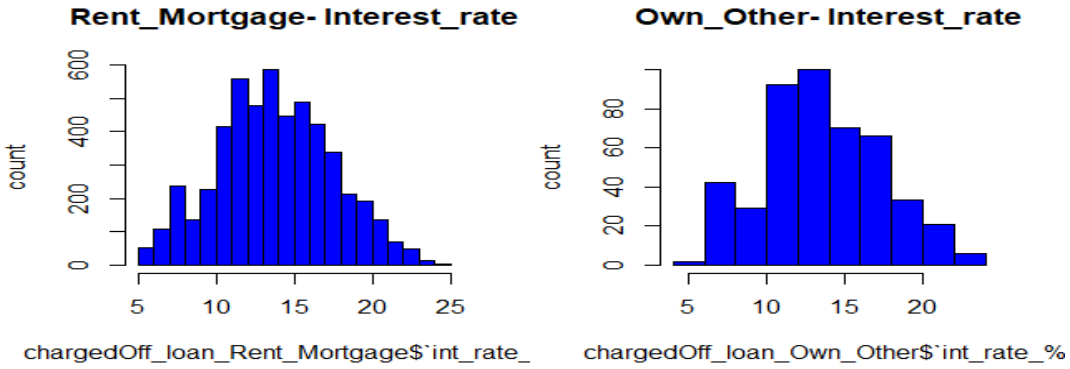
Segmented Univariate Analysis



Risk increases with increase in the funded amount, Rent_Mortgage has way too high of funded amount.

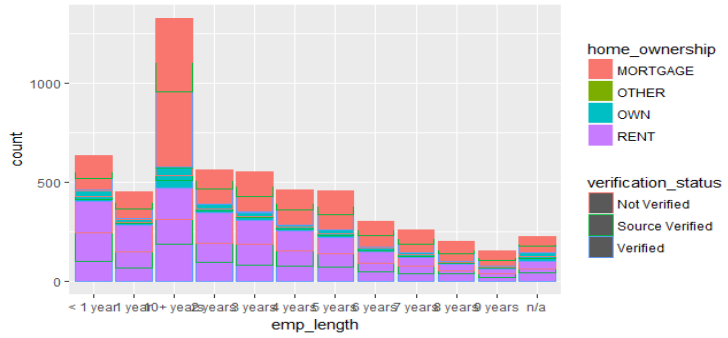


Contribution of risky LC grade B is more in Rent Mortgage

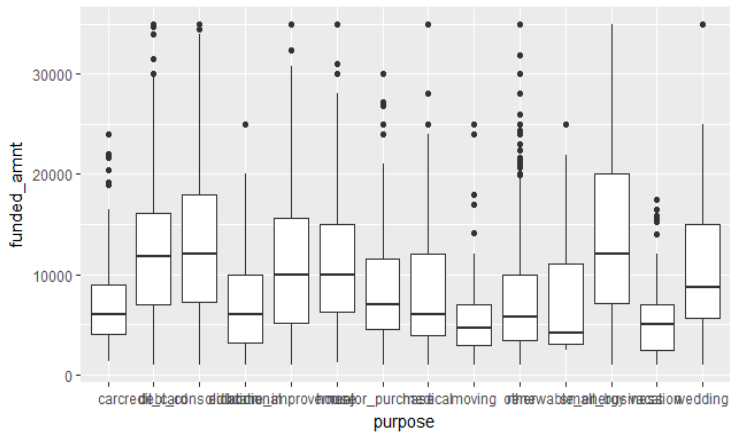


Contribution of frequency is more in Rent_Mortgage on the high interest rate, which is a clean risk for loan default

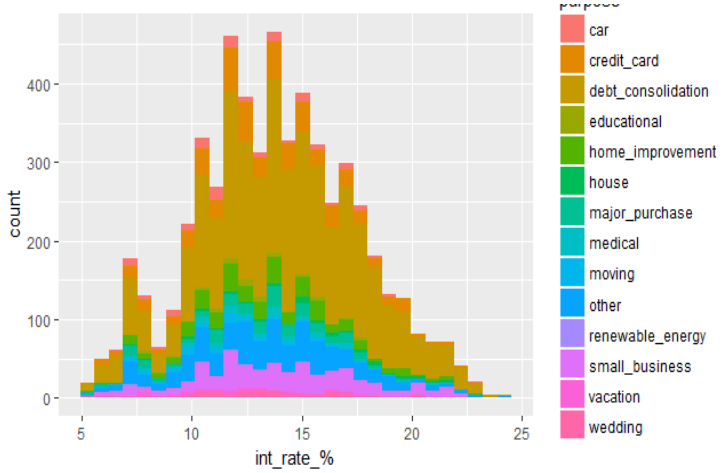
Bivariate Analysis



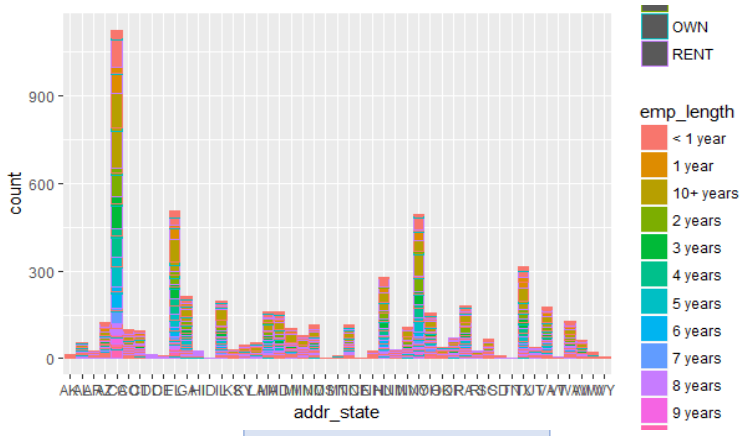
Mortgage and Rent home ownership are more loan defaulters



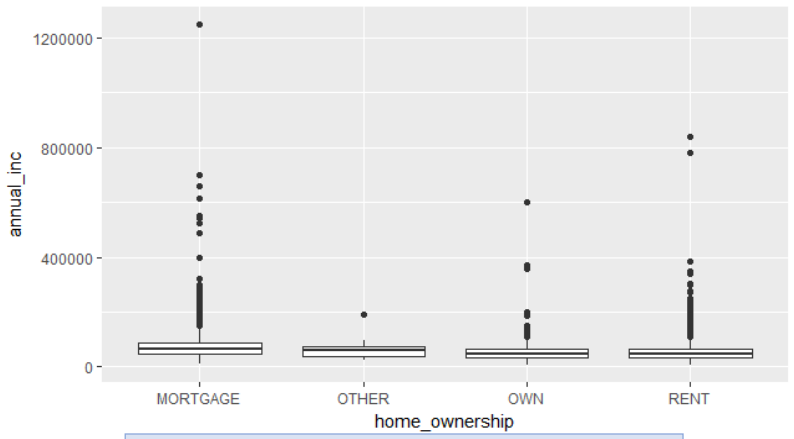
Debt_consolidation purpose tops the loan default rate



Higher the interest rate, Higher the risk



CA state heads most



Higher chance of loan defaulters if annual_inc is < 53000

Conclusion

The **Risk to default a loan** would be higher for a customer having the below combinations of variables:

- addr_state == CA
- emp_length == 10+
- home_ownership == RENT | MORTGAGE
- annual_inc < 53000
- purpose == debt_consolidation

Suggestion to the online loan lender company:

In order to keep the credit loss low or minimal, the company has to avoid lending a loan or to keep the interest rate higher if the borrower is from CA state, having 10+ years of experience and drawing less than 53000 annual income who is residing in Rented or Mortgaged home and requesting loan for debit consolidation purpose.

The combinations of all the variables said above could be the best case for loan rejection and combination of 2 or 3 variables could be the best case for loans with higher interest rates, which can be defined solid with the Business/Stakeholder's input

Note: this is the first cut analysis output and there could be other combinations as well which could be derived after modeling and further analysis.