A Case Study on Risk Analysis of Loan Applicants using Exploratory Data Analysis

Presented By:

Group Name: Data Coders

1. Siva Thota
2. Bhavani Bhavineni
3. Ujwal Kiran Bhargava
4. Vijayaraghavan

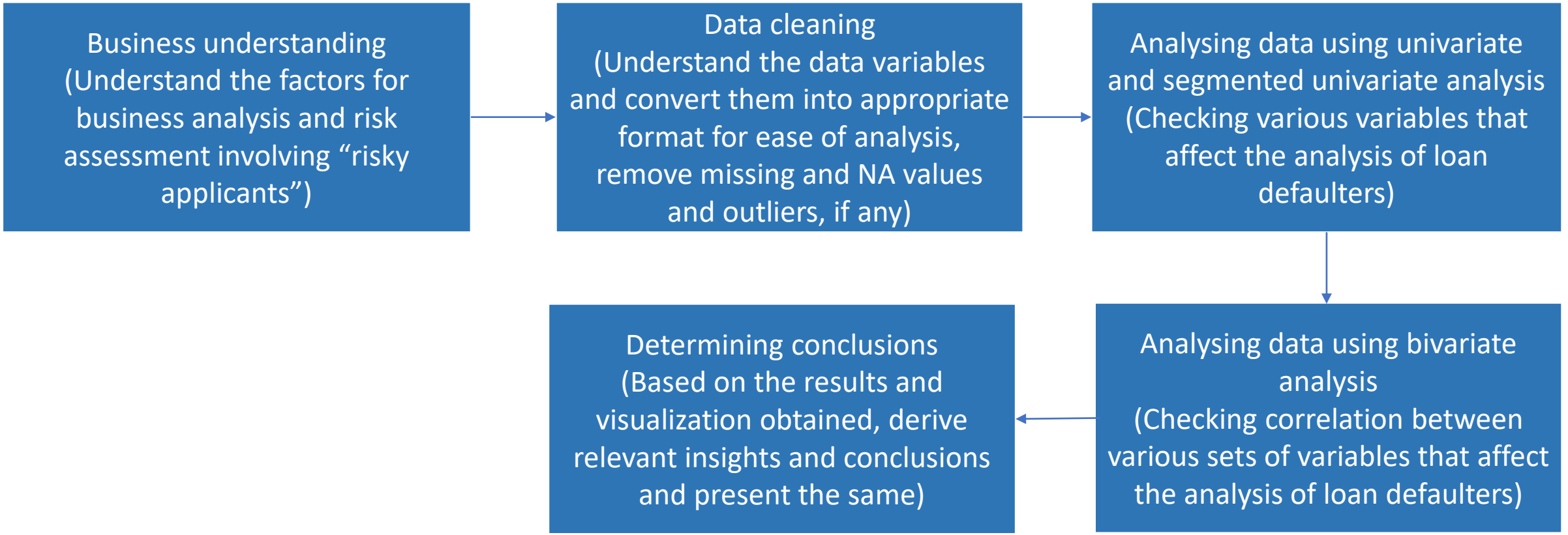
Introduction, Background and Problem Statement

- The company is in the consumer finance sector and it specialises in lending various types of loans to urban customers.
- When an applicant applies for a loan, it has to take a decision based on approving the same or not based on the applicant's background and profile.
- Two **types of risks** are associated with the bank's decision:
  - If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
  - If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

Introduction, Background and Problem Statement

- The main objective of the case study is to identify patterns that indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- This can be done by identifying risky applicants (loan defaulters labelled as “charged off”) so as to reduce the credit loss.
- The following steps are involved in this activity:
  - Understand the input .csv file and perform Exploratory Data Analysis (EDA) in R on the same to identify driver variables (driving factors) that strongly indicate loan default and ‘risky applicants’.
  - Perform Univariate and Bivariate Analysis to analyse the driver variables for risk assessment.
  - Present the observations and conclusions based on the above using visualisation.

Problem Solving Methodology



Understanding Data and Assumptions Used in the Case Study

- From the given .csv file, there are 3 types of data used for analysis in this case study:
  - Data based on customer information and demographics such as emp_title, emp_length, home_ownership, zip_code and addr_state
  - Data based on loan characteristics such as loan_amnt, funded_amnt, installment, loan_status and int_rate
  - Data based on customer behaviour such as last_pymnt_amnt, application_type, open_acc, total_pymnt and delinq_amnt
- However, customer behaviour data is neglected as it cannot be collected at the time of application.

Understanding Data and Assumptions Used in the Case Study

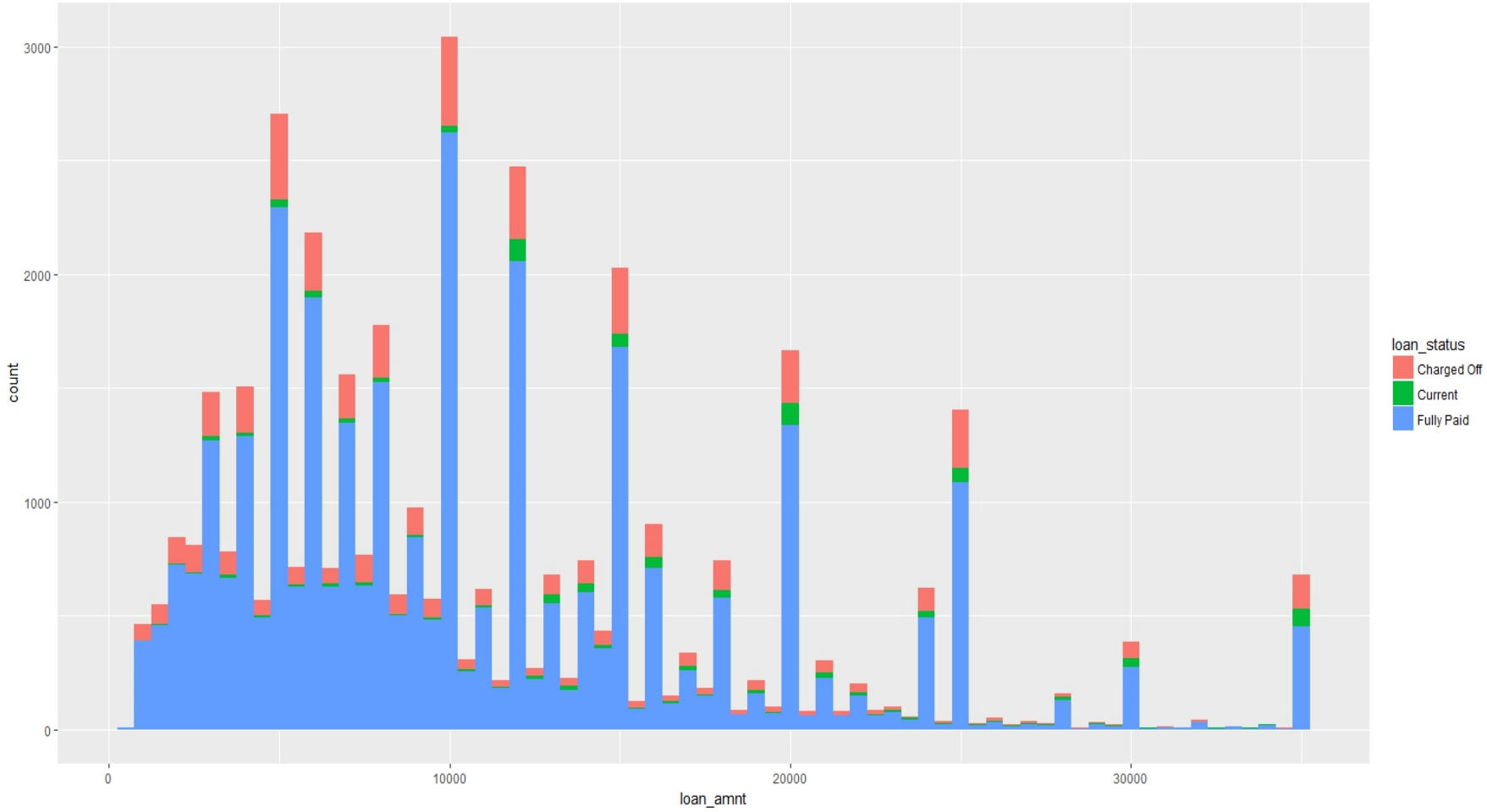
- Int_rate variable is converted into numeric type since it is in percentage.
- The columns having all NA’s or 0’s as row value are discarded.
- The columns having only one unique row value are also discarded.
- All character type variables are converted into factor type for ease of analysis.
- All month-year type variables are converted into appropriate data format for ease of analysis.

Derived Metrics Used in the Case Study

The following are the three major derived metrics used in this case study:

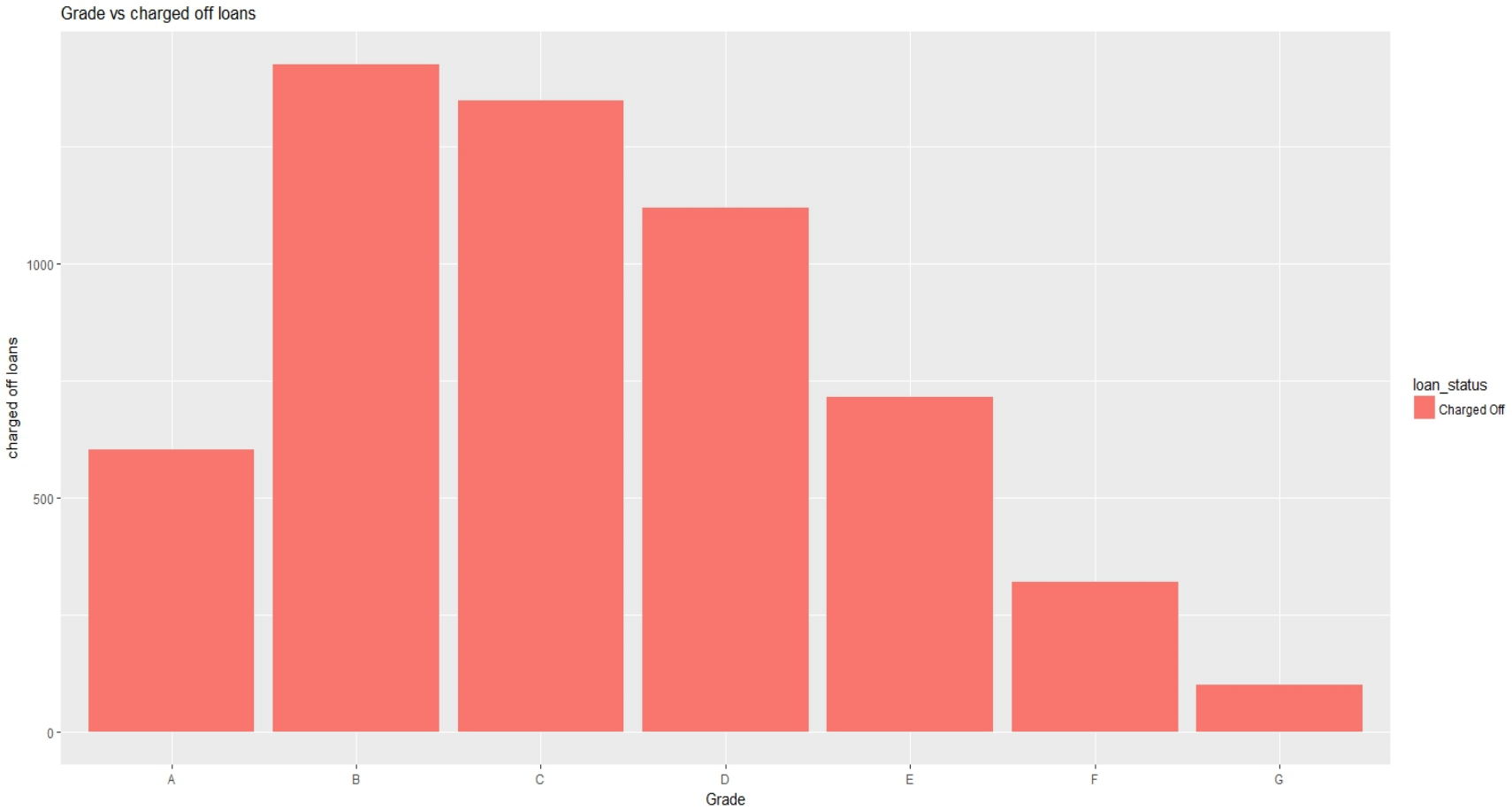
- loan_defaulters – derived from loan, used to indicate loans that are defaulted (i.e. having loan status as “Charged off”)
- annual_inc_status – derived from annual_inc, used to indicate high, medium and low types of annual income
- intrate_status- derived from int_rate, used to indicate high, medium, extreme and low types of interest rates

Segmented Univariate Analysis w.r.t Loan Amount



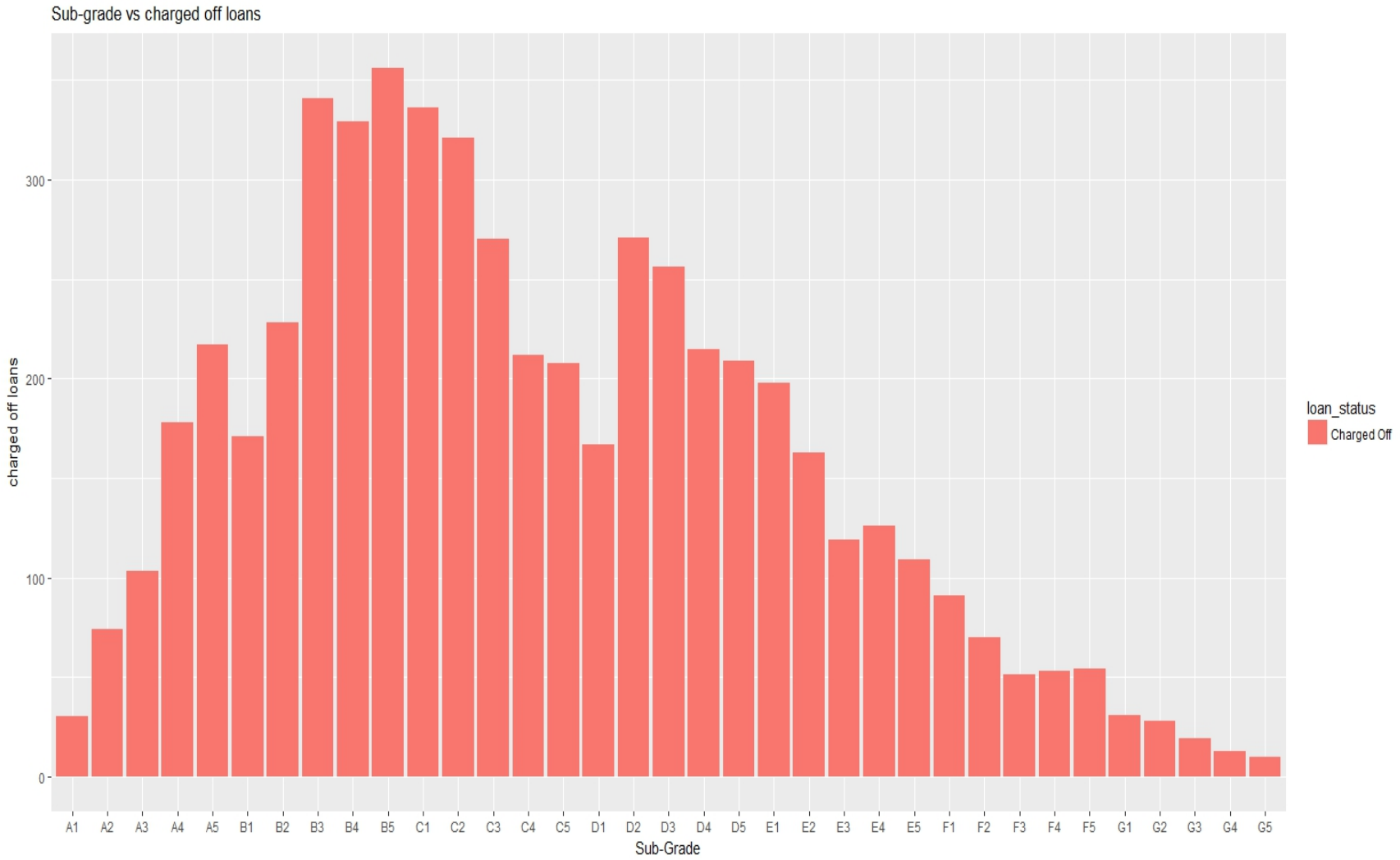
From the analysis, it is clear that loan amount 10000 has the most number of loans.

Segmented Univariate Analysis – 1



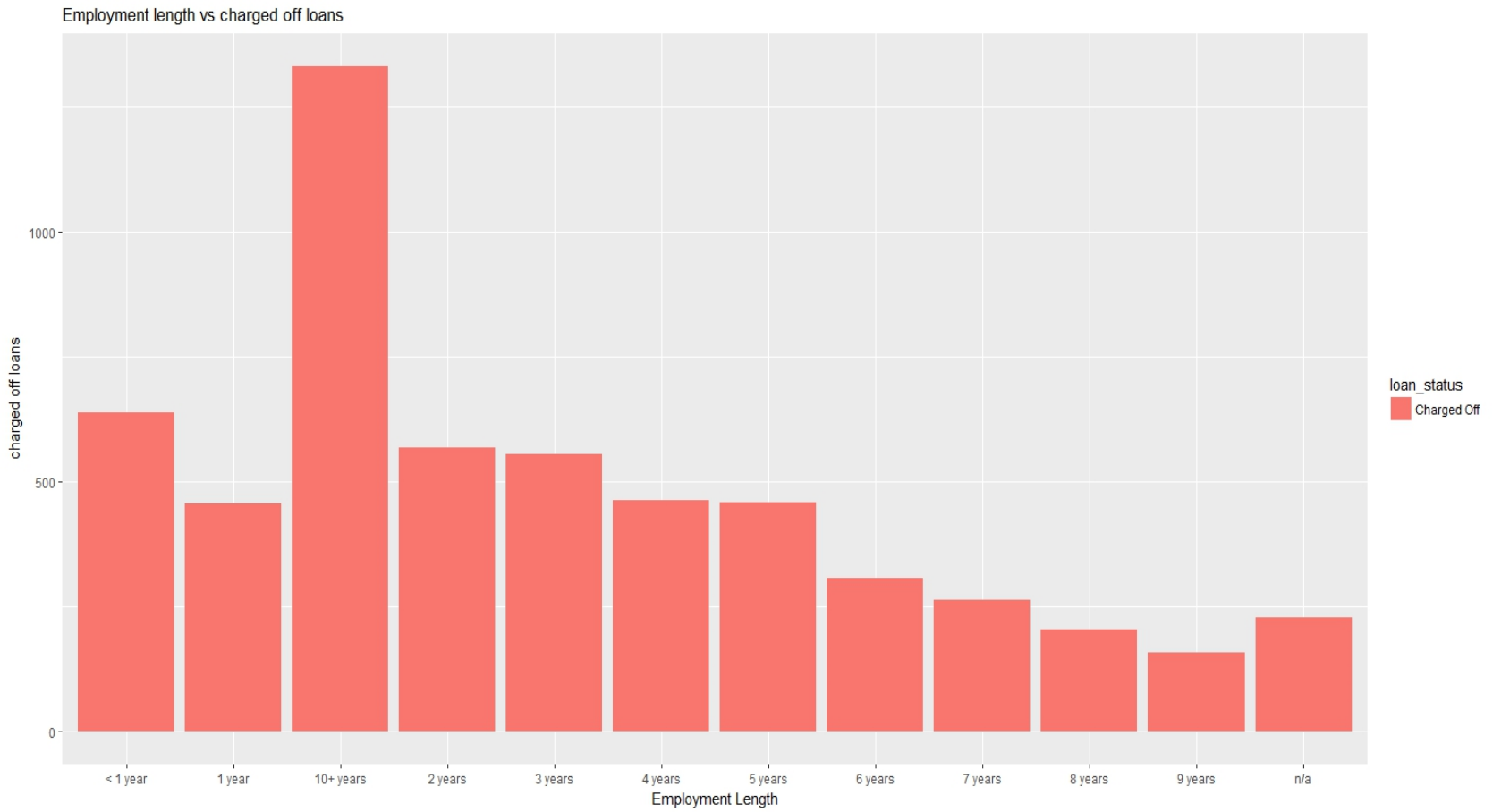
From the analysis, it is clear that Grades B, C and D have the most number of charged off loans.

Segmented Univariate Analysis – 2



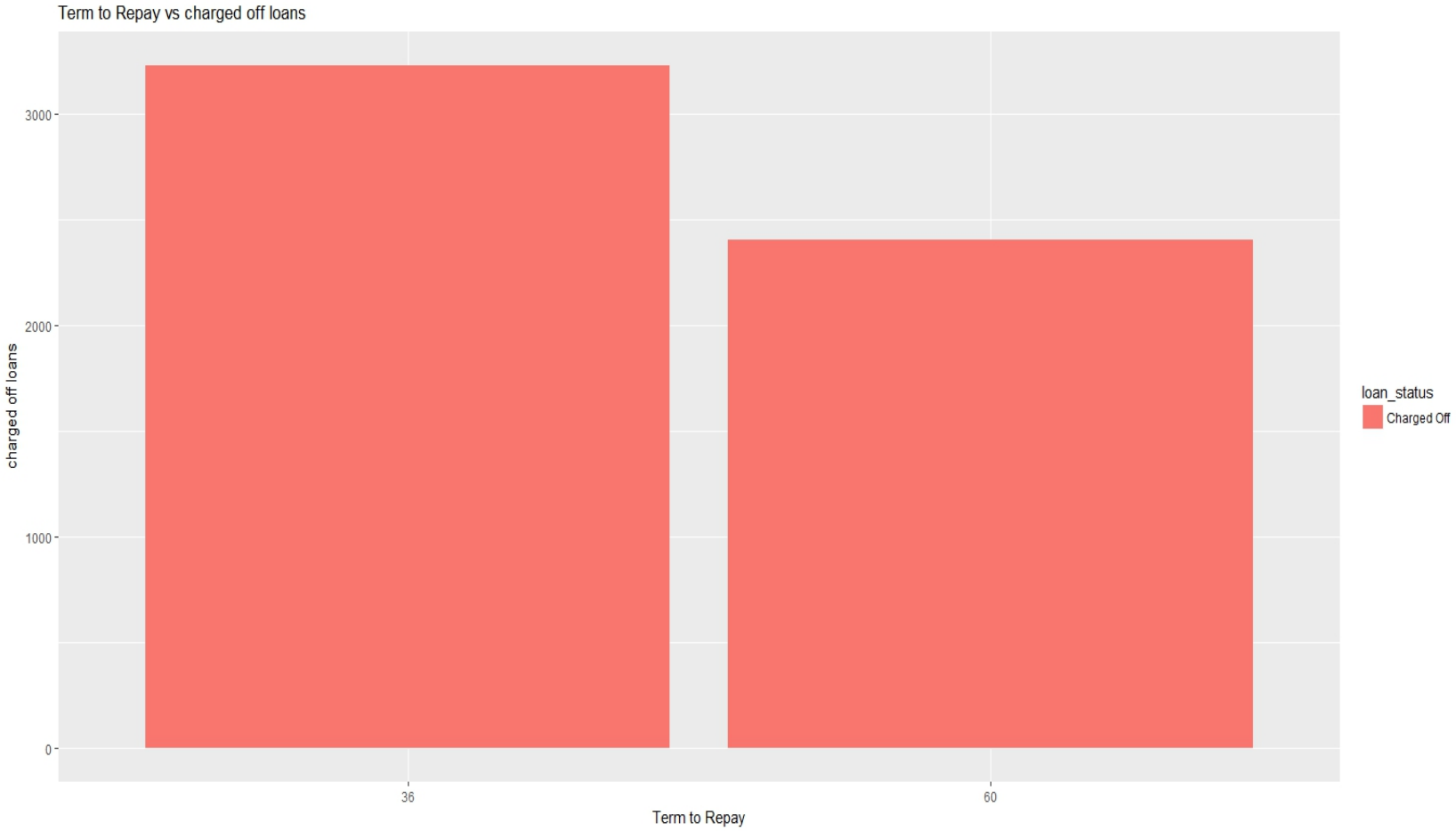
From the analysis, it is clear that sub-grade B5 has the most number of charged off loans.

Segmented Univariate Analysis – 3



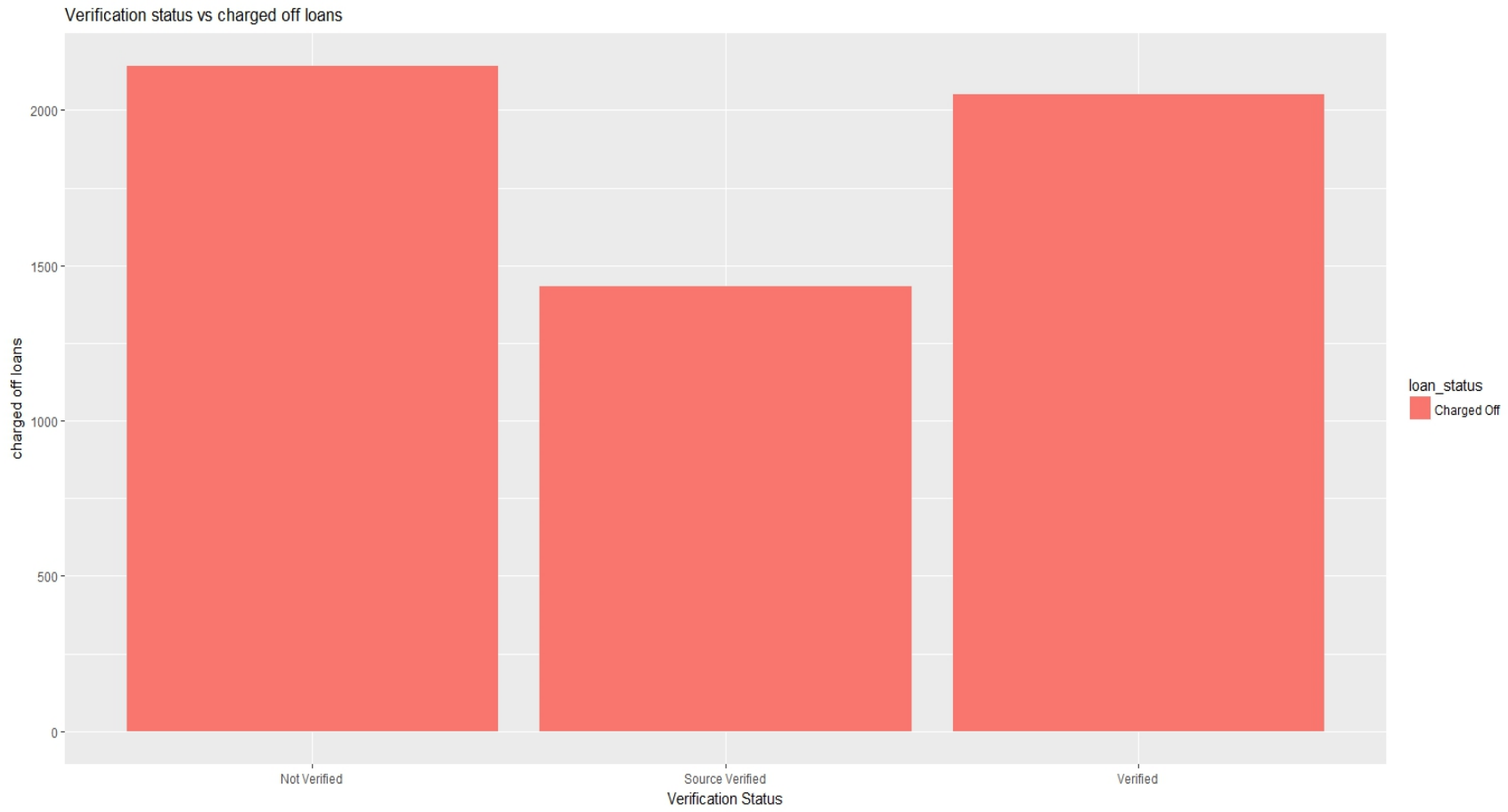
From the analysis, it is clear that loans applied by people employed for 10+ years have the highest number of charged-off loans.

Segmented Univariate Analysis – 4



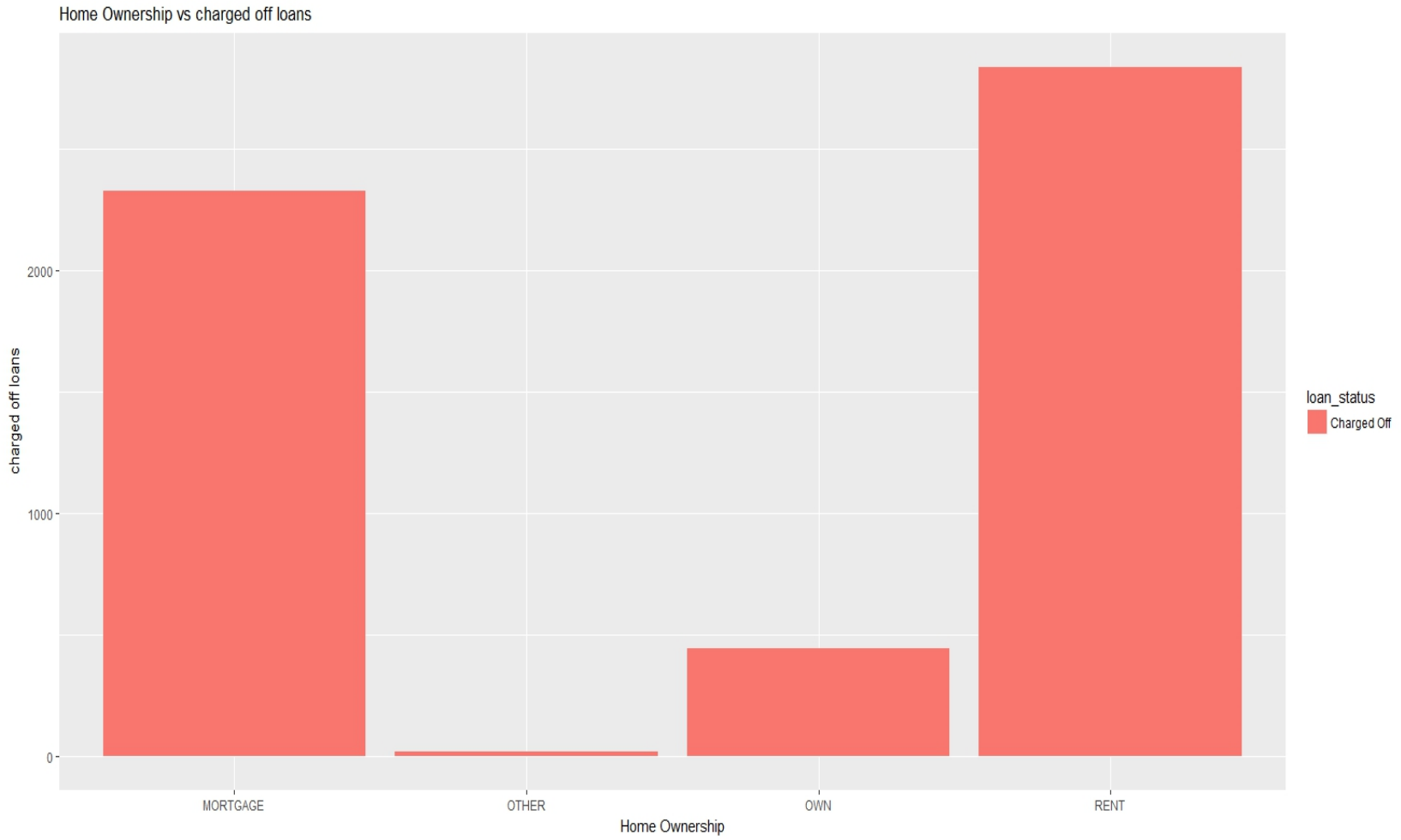
From the analysis, it is clear that loans having term of 36 months have the highest number of charged-off loans.

Segmented Univariate Analysis – 5



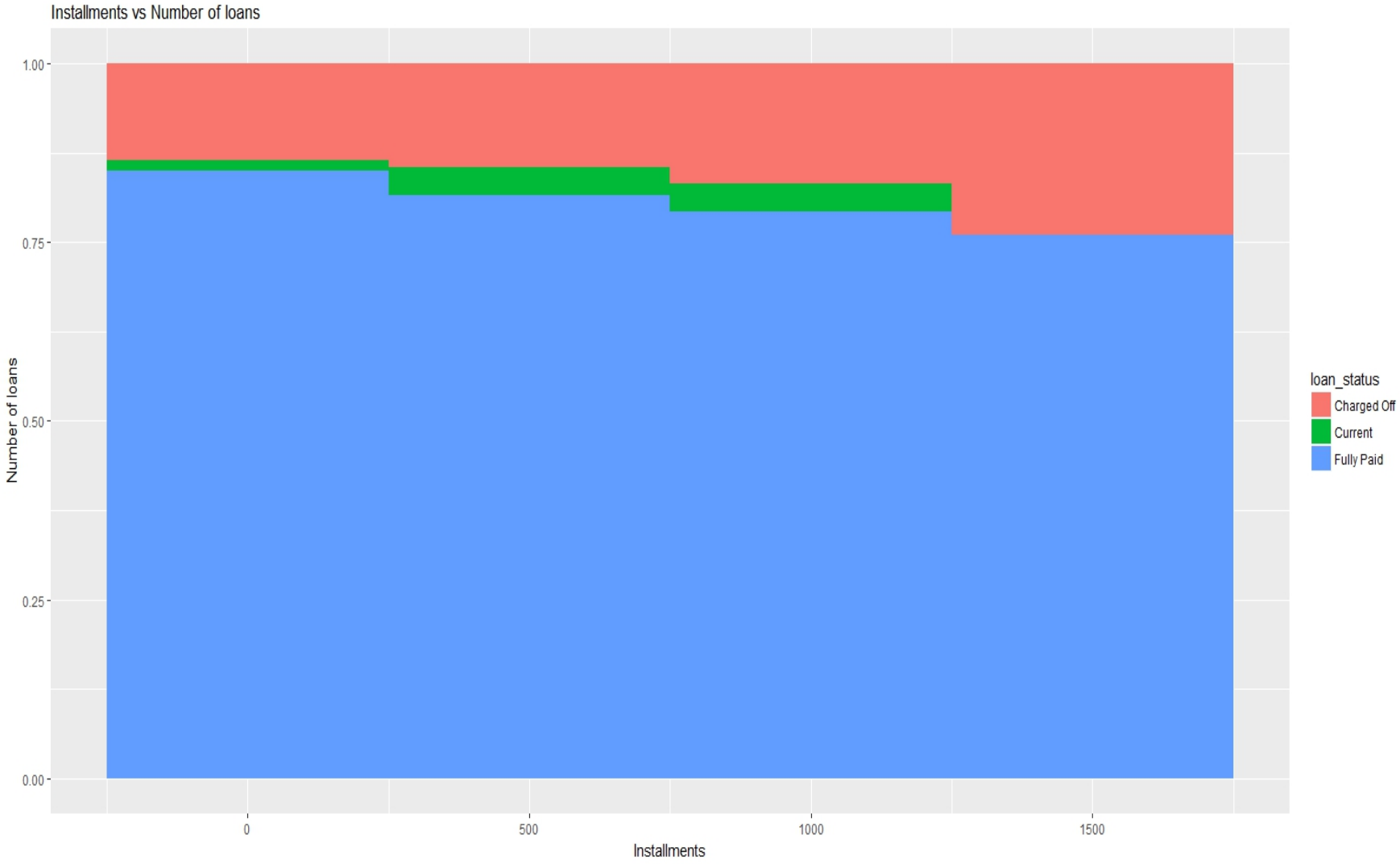
From the analysis, it is clear that non-verified type of loan applicants has the highest number of charged-off loans.

Segmented Univariate Analysis – 6



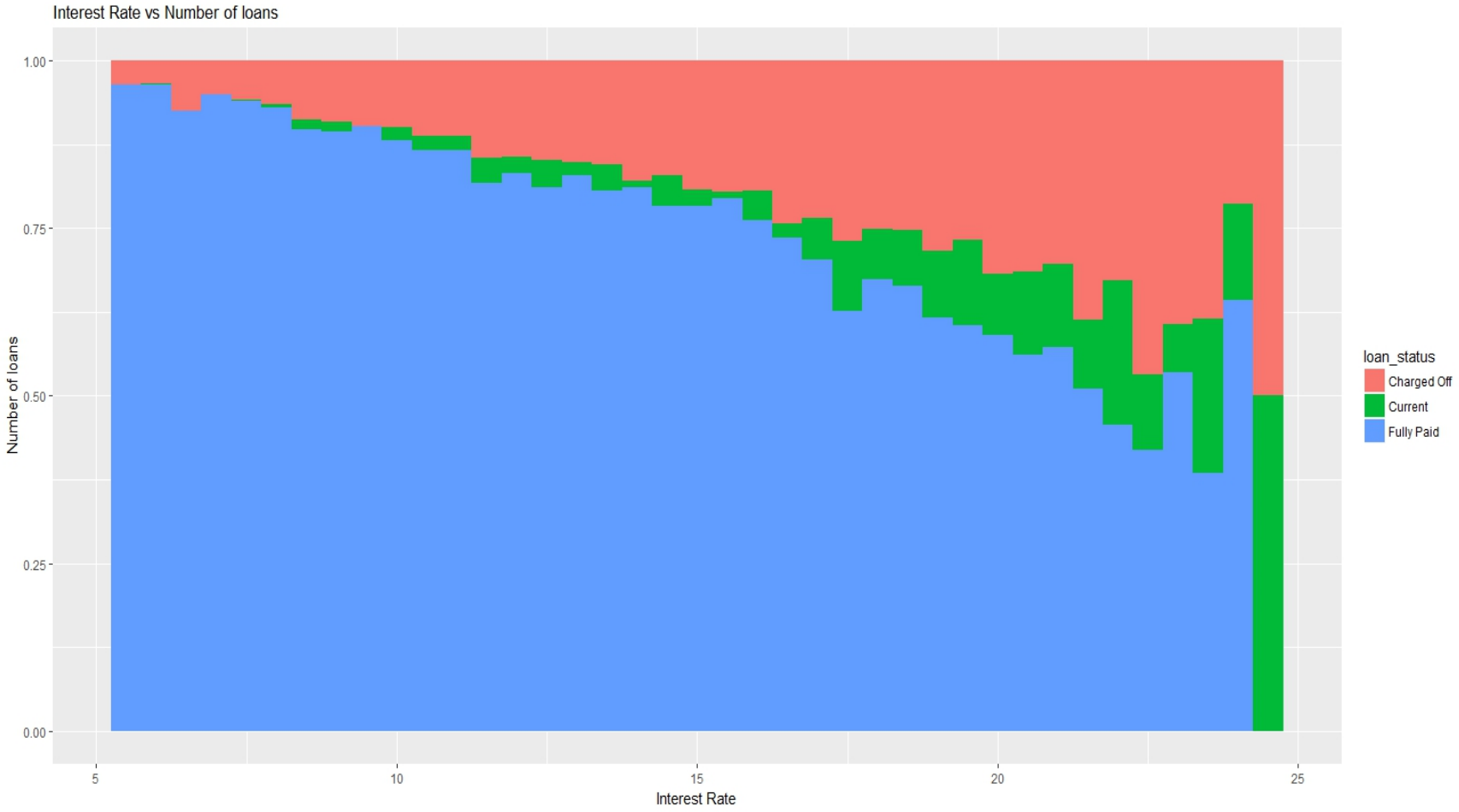
From the analysis, it is clear that mortgaged and rented ownership types have the highest number of charged-off loans.

Univariate Analysis – 1



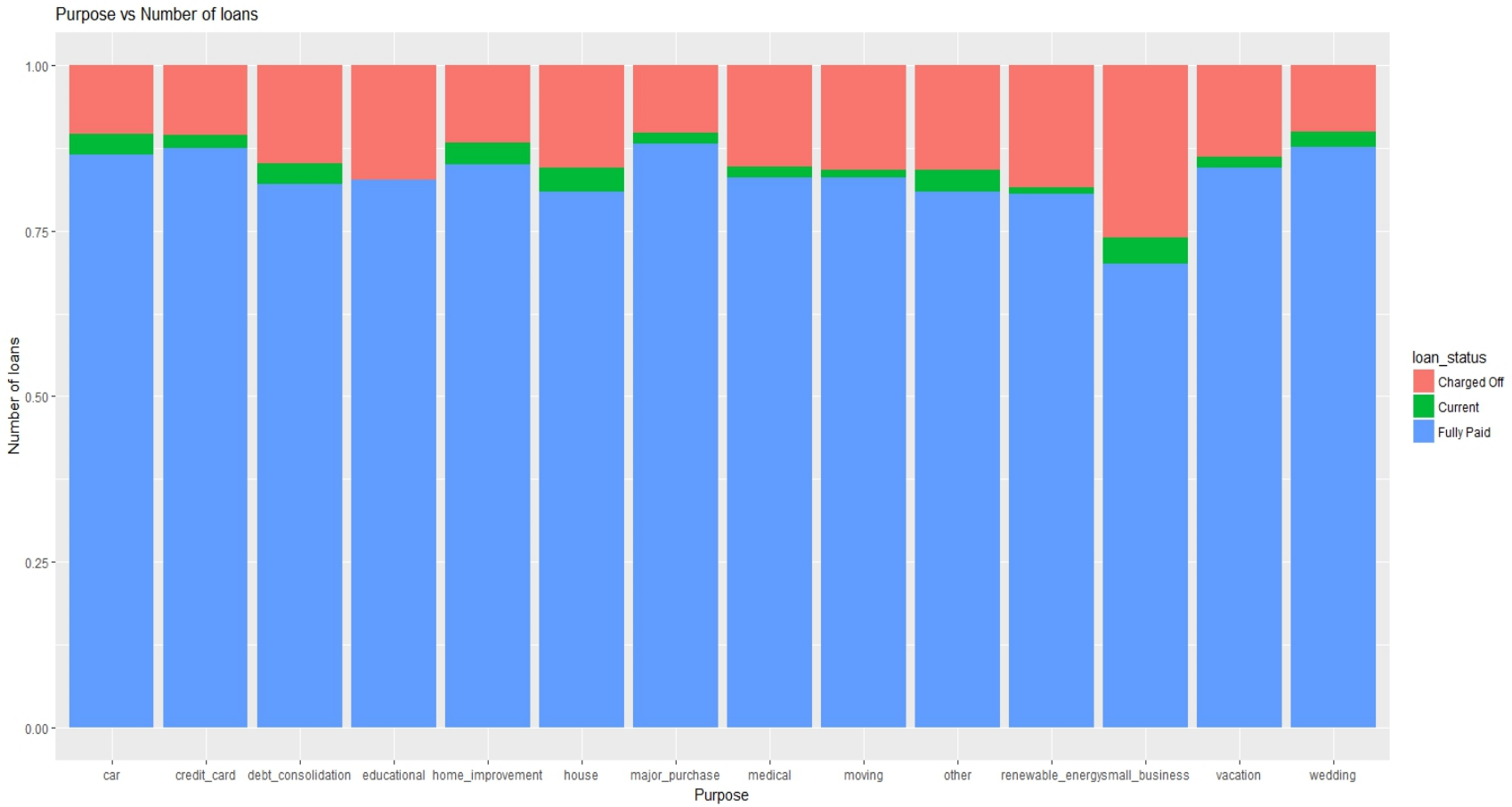
From the analysis, it is clear that installments of 1500 have the most number of charged-off loans.

Univariate Analysis – 2



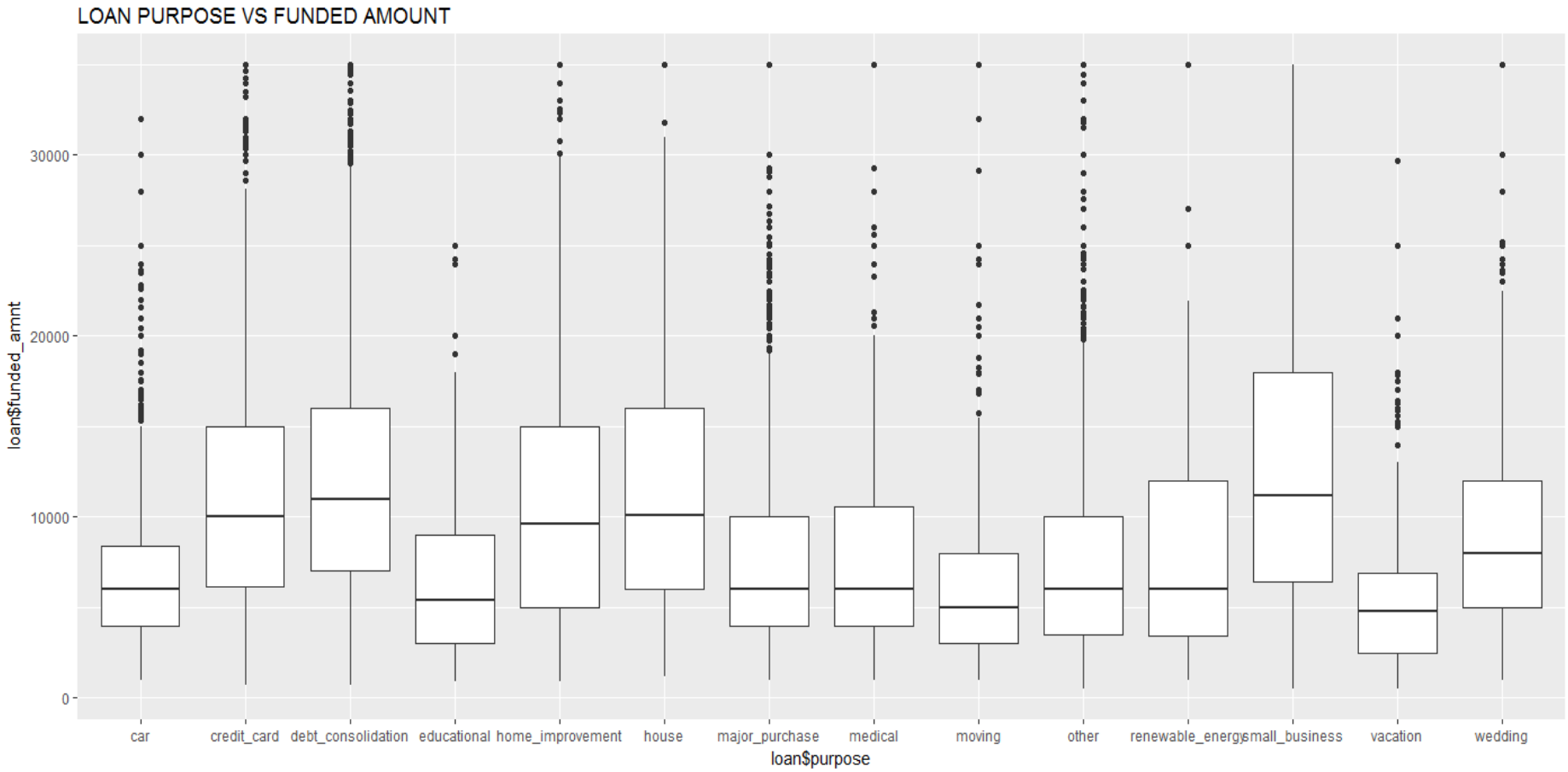
From the analysis, it is clear that higher interest rates lead to more number of charged-off loans.

Univariate Analysis – 3



From the analysis, it is clear that loans applied for small business purpose has the highest percentage of charged-off loans.

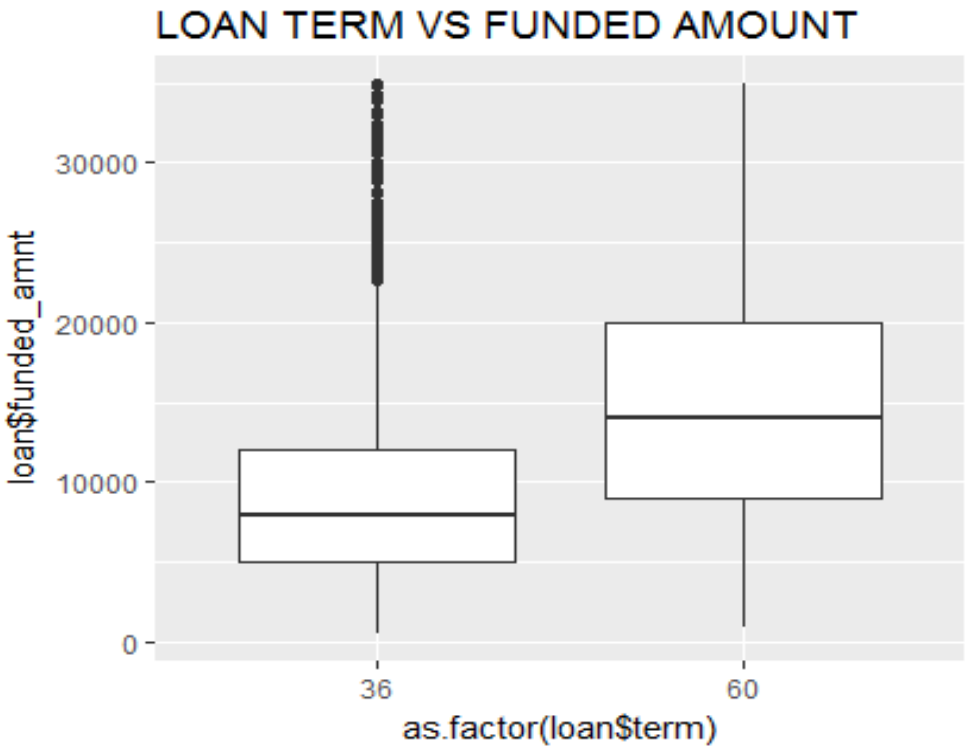
Bivariate Analysis – 1



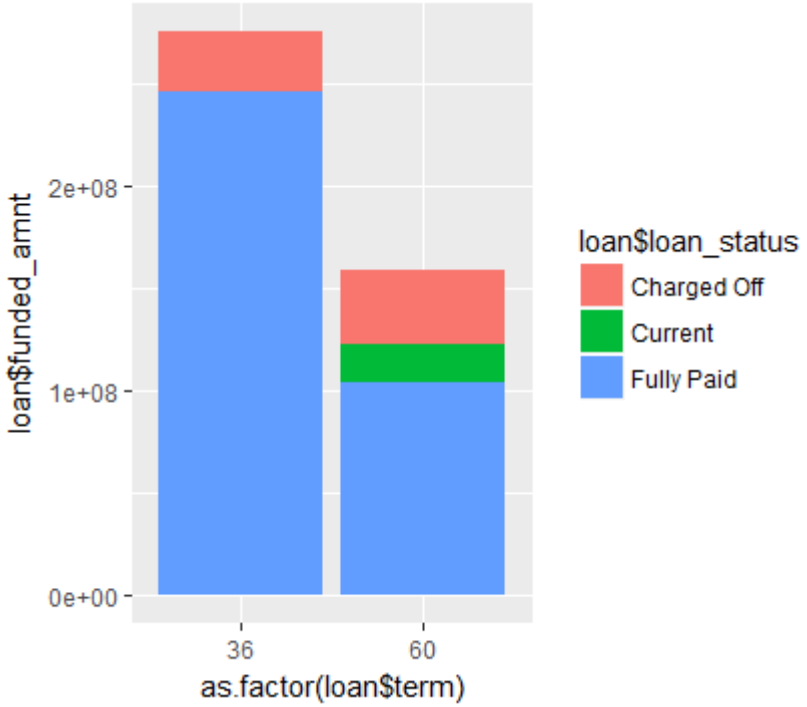
From the analysis, it is clear that these five major products got high funded amount:

1. Debt Consolidation
2. Small Business
3. Home Improvement
4. Credit Card
5. House

Bivariate Analysis – 2

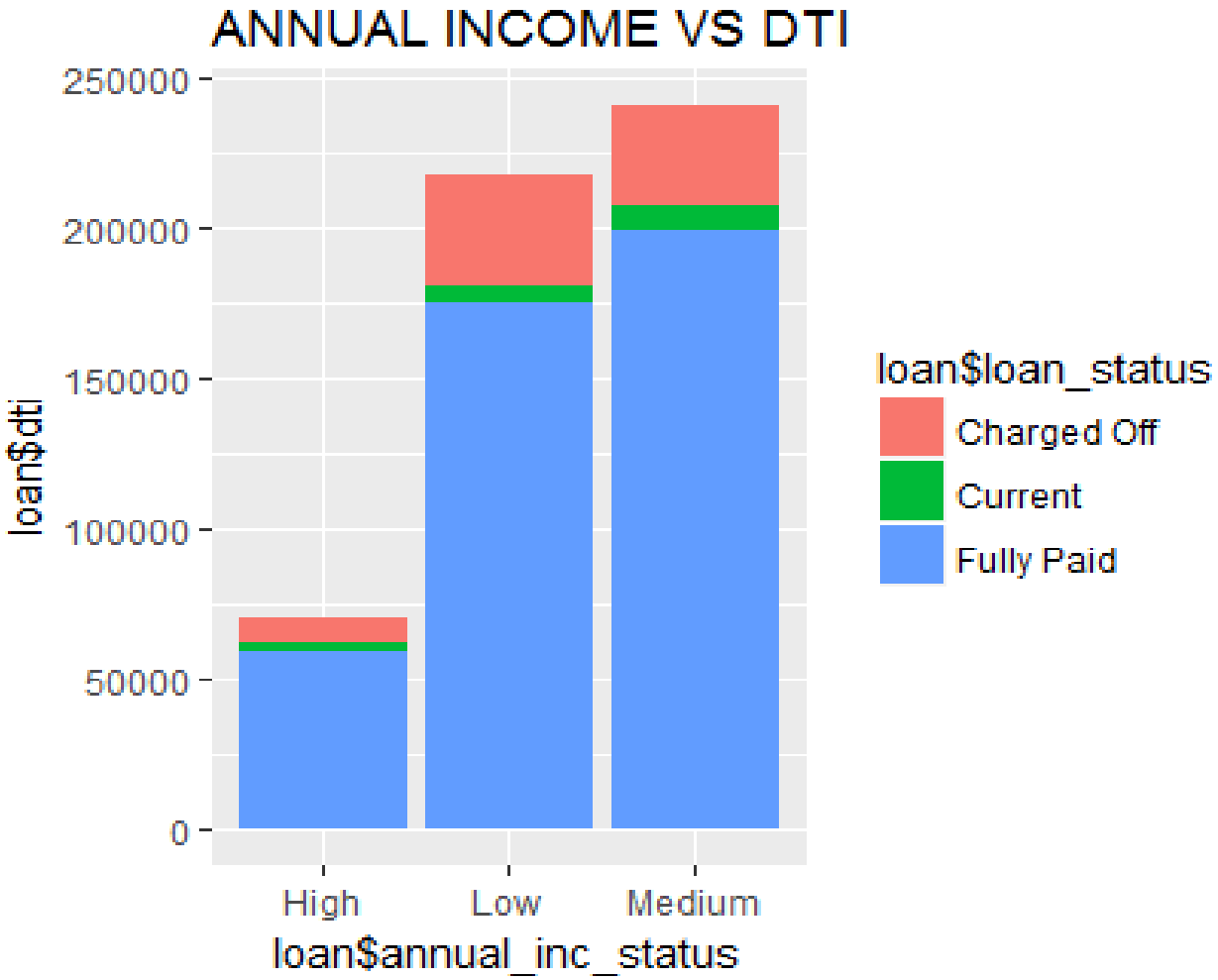


From the analysis, it is clear that major loans belong to the 60 months term.



From the analysis, comparatively 60 months term loans had a higher default rate.

Bivariate Analysis – 3



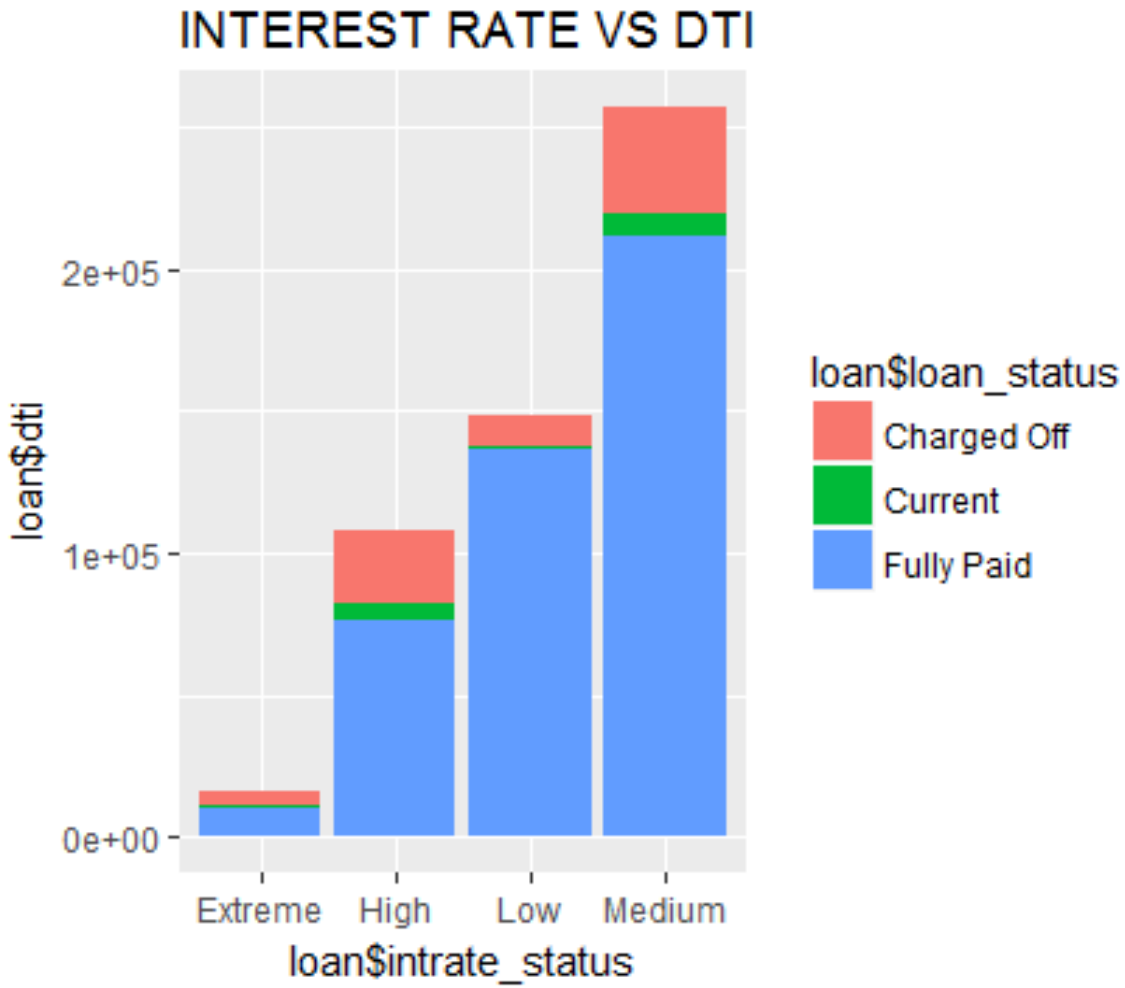
From this analysis,

- High-income earners have low DTI and low charged-off rate
- Low- and Medium-income earners with high DTI rate have high possibility of Loan Default Rate

Annual Income

- LOW (<= 50,000)
- MEDIUM (>50,000 & <= 1,00,000)
- HIGH (>1,00,000)

Bivariate Analysis – 4



From this analysis, it is clear that high and medium interest rates with high DTI rate have a higher chance of having charged-off loans

Interest Rate

- LOW (<= 10)
- MEDIUM (>10 & <=15)
- HIGH (>15 & <= 20)
- EXTREME (>20)

Conclusions and Recommendations

- The five major driving variables (main indicators of loan default) are:
  - grade (grade)
  - sub_grade (sub-grade)
  - emp_length (employment length)
  - term (term to repay loan)
  - verification_status (loan verification status)
  - home_ownership (home ownership type)
- Apart from these, there are a lot of other driver variables such as annual income (annual_inc), interest rates (int_rate), installment, loan purpose (purpose), annual_inc_status and intrate_status.
- For fewer charged-off loans and to reduce credit loss, the following measures can be taken:
  - Interest rates on loan and the loan amount should be reduced to ensure fewer loan defaults.
  - The loans should have fewer installments.

Correlation Matrix Based on Various Data Set Variables