

EDA CASE STUDY

SUBMISSION

Group Name:

1. Ashish Garg
2. Brijesh Nair
3. Tarak Taranekar
4. Karthik Kannan

Business Overview and Objectives

Business Overview

- The company is the largest online loan marketplace that offers borrowers easy access to lower interest rates for loans through a fast online interface

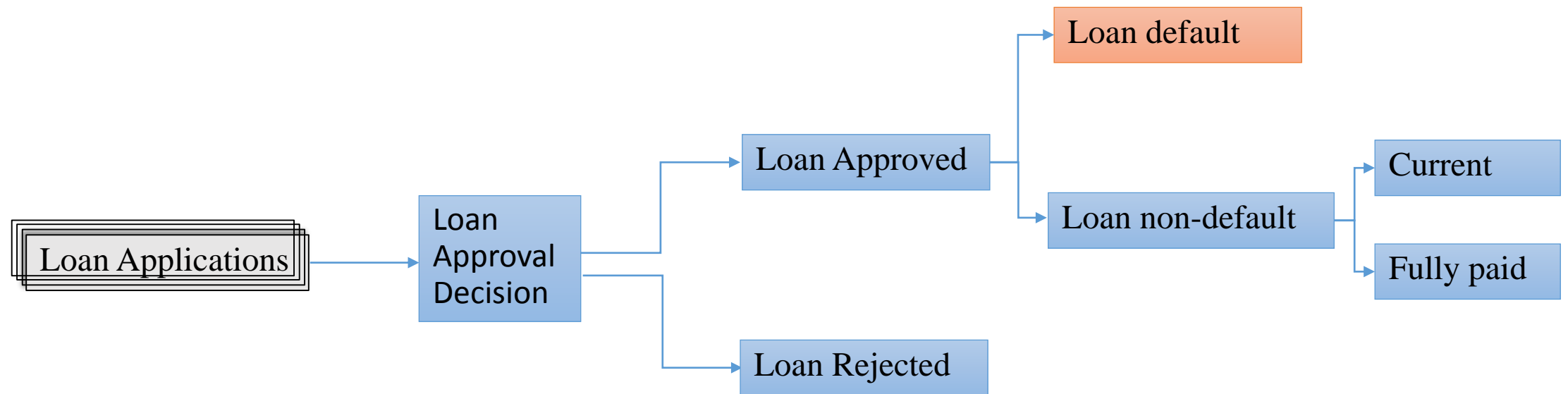
Business Objective

- Understand the **driving factors** that are strong indicators of loan default, thereby **segregating risky** loan applicants

Key constraints

- Rejecting risky customers with repaying capacity could result in business loss to the company
- Accepting risky customers with no intention to repay could result in financial loss to the company

Key Deciding factors



Data Cleaning Methods Used

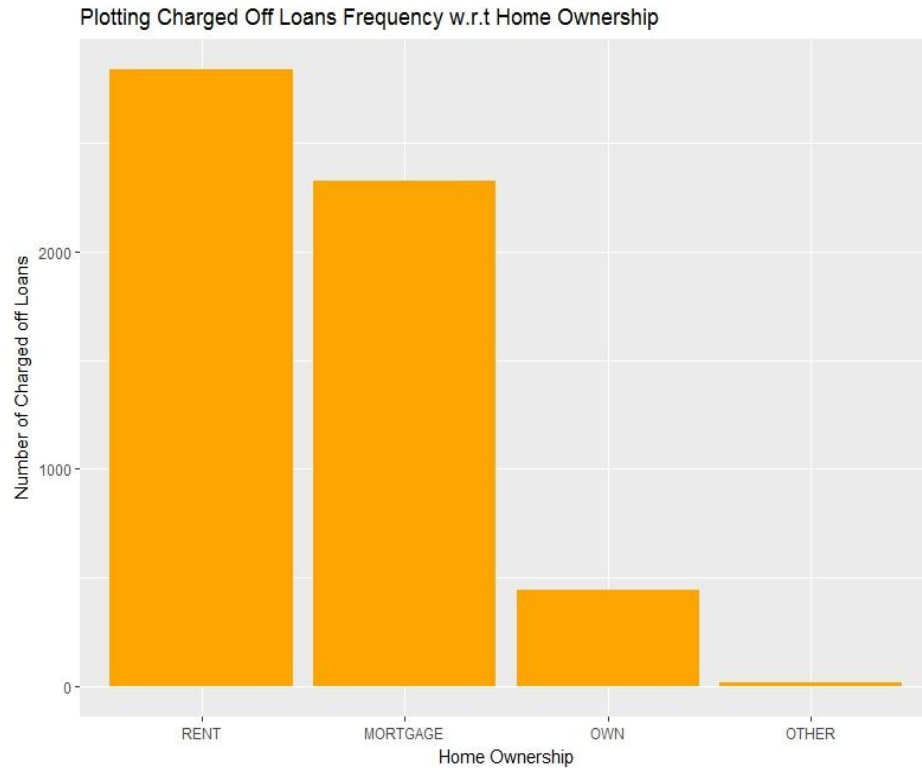
- Removal of blank ("") cells with NA
- Removal of cells with space characters (" ") with NA
- Check for duplicate IDs
- Removal of columns having only NA for all rows
- Check and remove columns having same values for all the rows
- Check and remove % symbols from int_rate, revol_util columns
- Remove 'months' from loan 'term' column
- Remove '+' and 'years' from emp_length column
- Update emp_length column with years having < 1 year as '0'
- Replacing values in emp_length column with uniform representation ('n/a' to 'NA')
- Standardizing verification_status field to 'Verified'
- Imputed credit line years earlier than 1970 to "<1970"

Consumer and Data Attributes

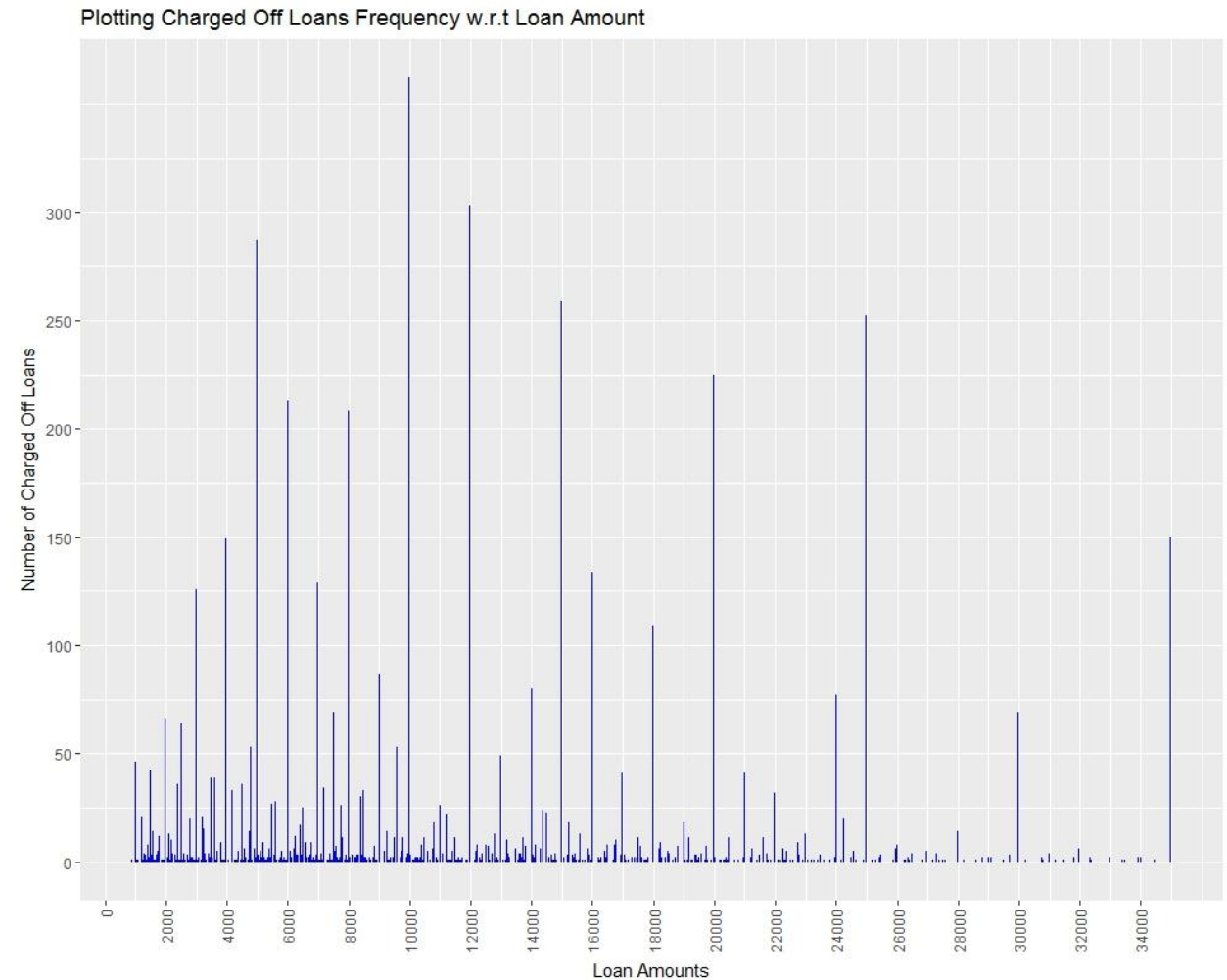
Business-driven attributes considered for the segmentation analysis

- **Consumer Attributes:**
 - Home Ownership Status (Categorical)
 - Duration of employment (Ordinal)
 - Annual Salary (Interval)
 - State of domicile (Categorical)
 - Credit line year (Ordinal)
 - Loan Grade (Ordinal)
- **Loan Attributes:**
 - Loan Amount (Interval)
 - Loan Purpose (Categorical)
 - Loan Rate (Interval)

Univariate Analysis



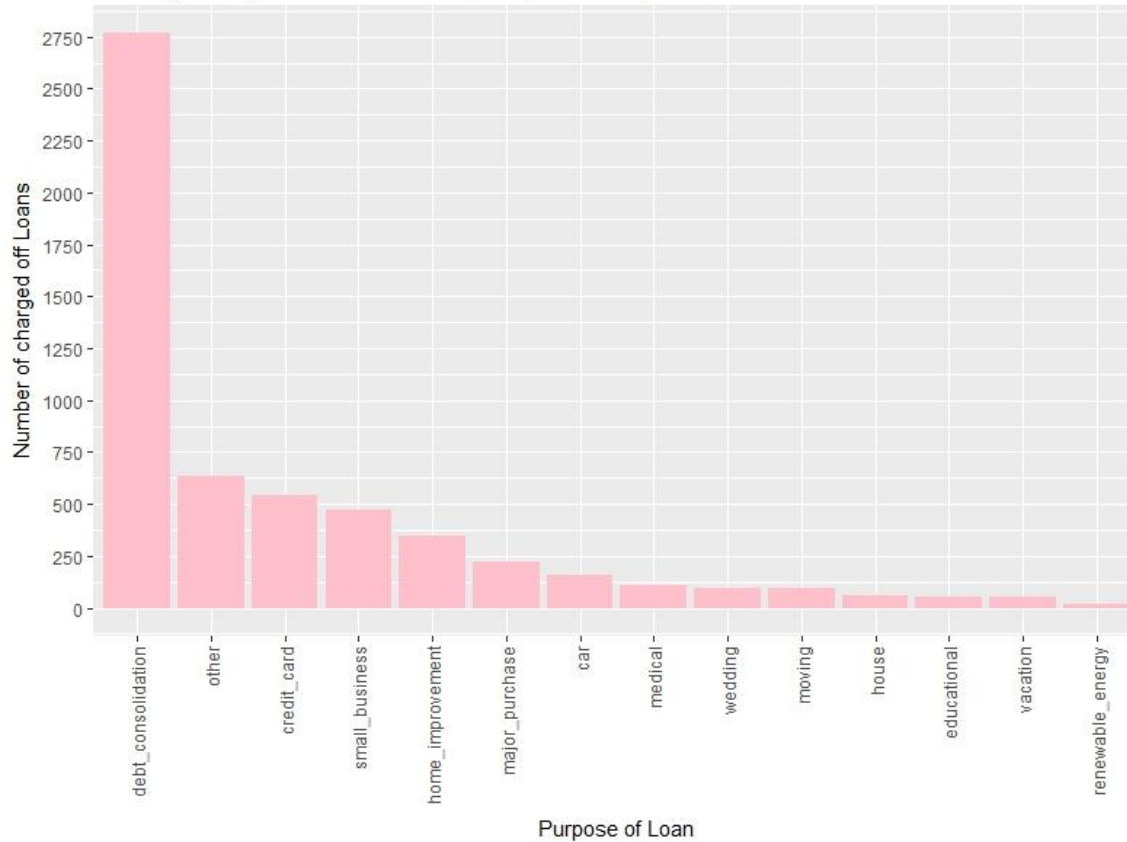
- It shows that a large number of loans taken by 'Rent' and 'Mortgage' customers have been defaulters



- It shows that customers who take loan in round figures are more prone to be defaulters

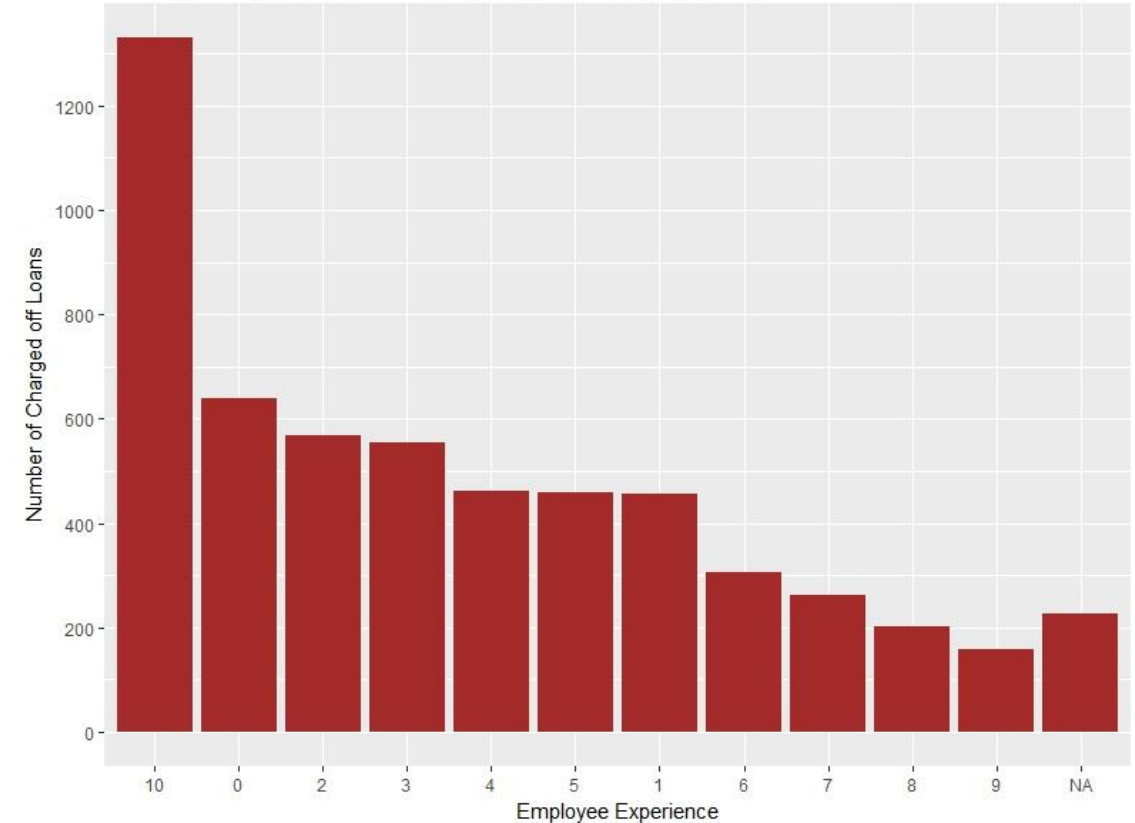
Univariate Analysis

Plotting Charged Off Loans Frequency w.r.t Purpose



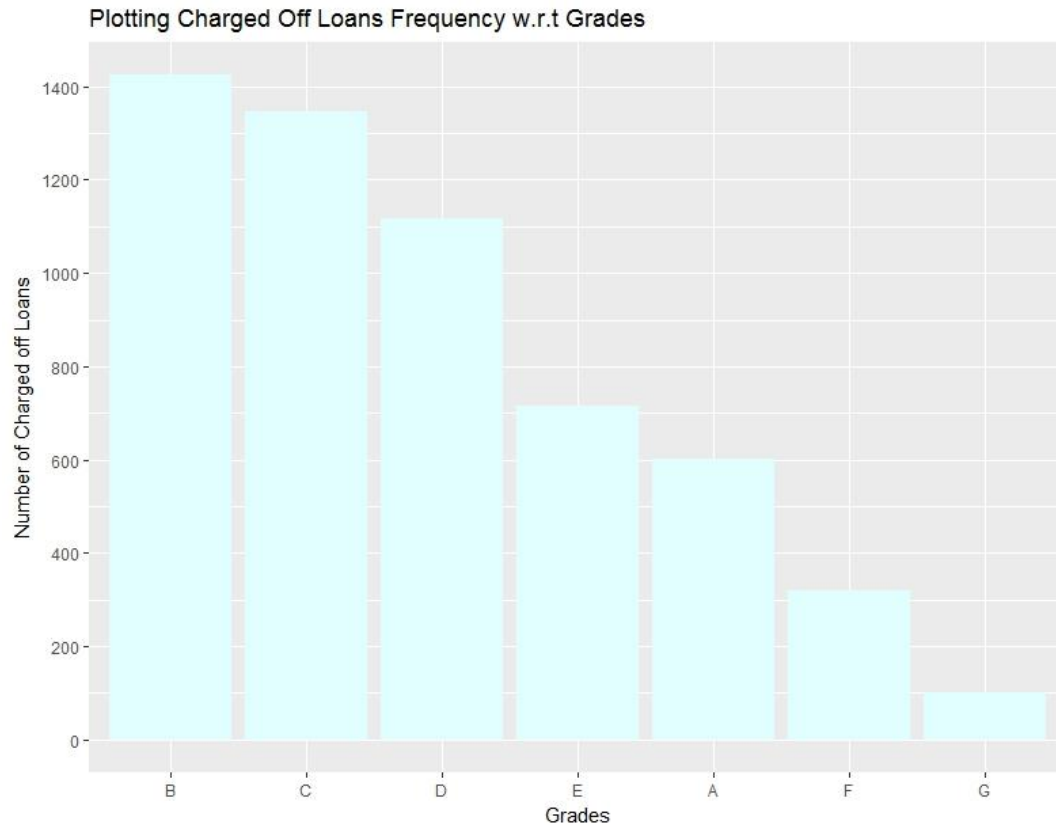
- It also shows that customers who take loan for 'debt_consolidation' as purpose are major defaulters

Plotting Charged Off Loans Frequency w.r.t Employee Exp

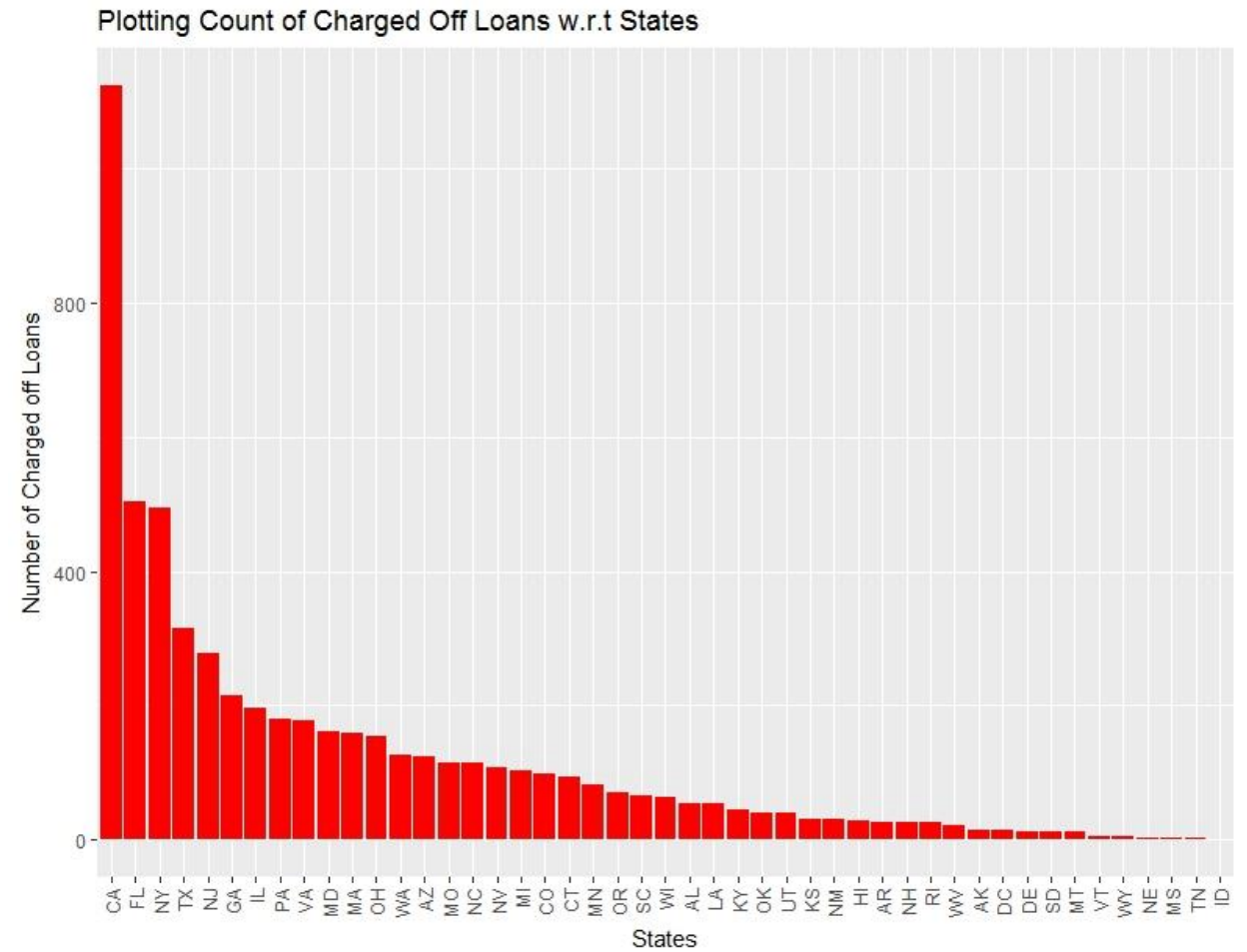


- The plot shows that customers with more than 10 years of experience are the major defaulters

Univariate Analysis



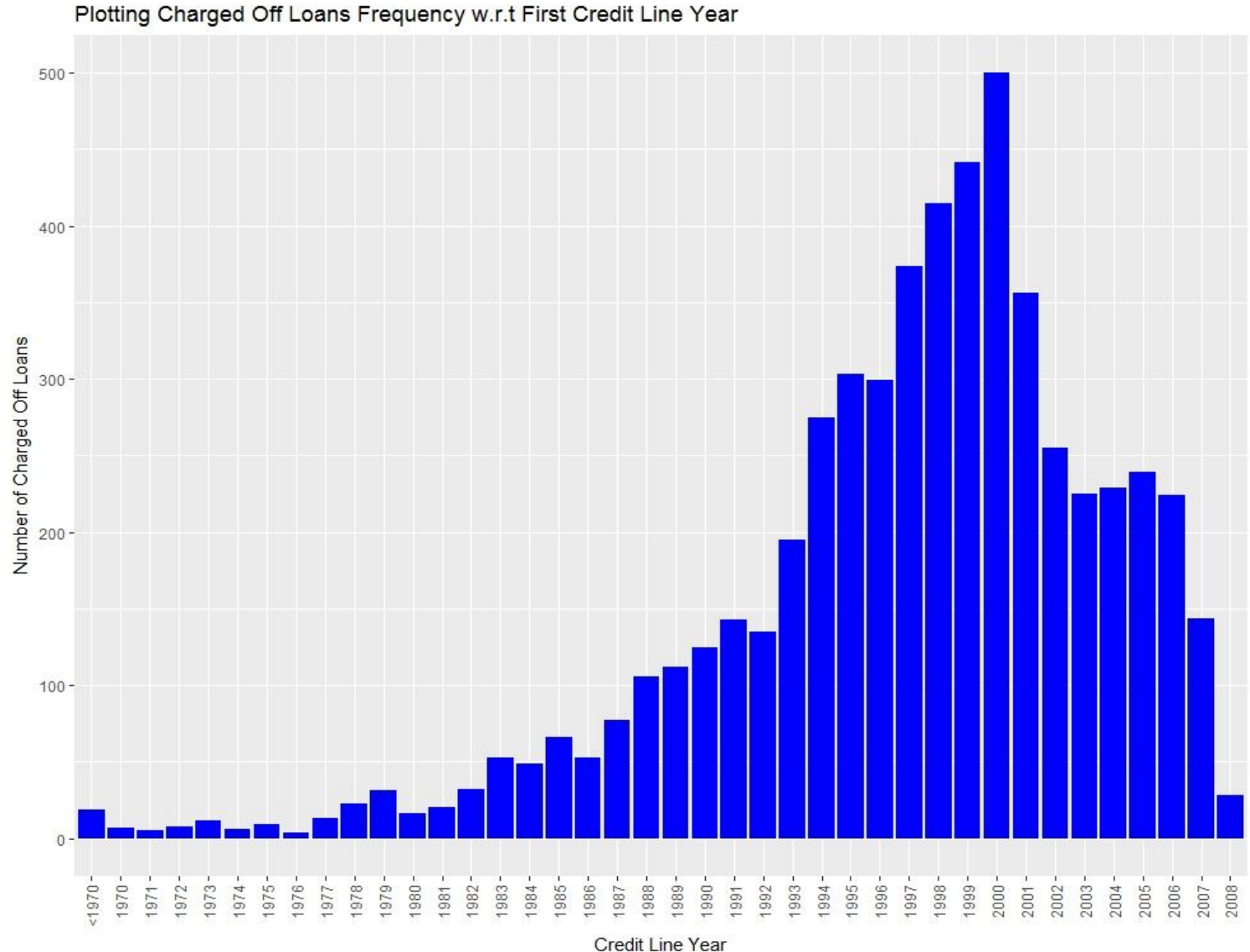
- The plot shows that customers with B, C, D grades are more prone to default



- It shows that CA, FL, NY are the top 3 states with maximum defaulters

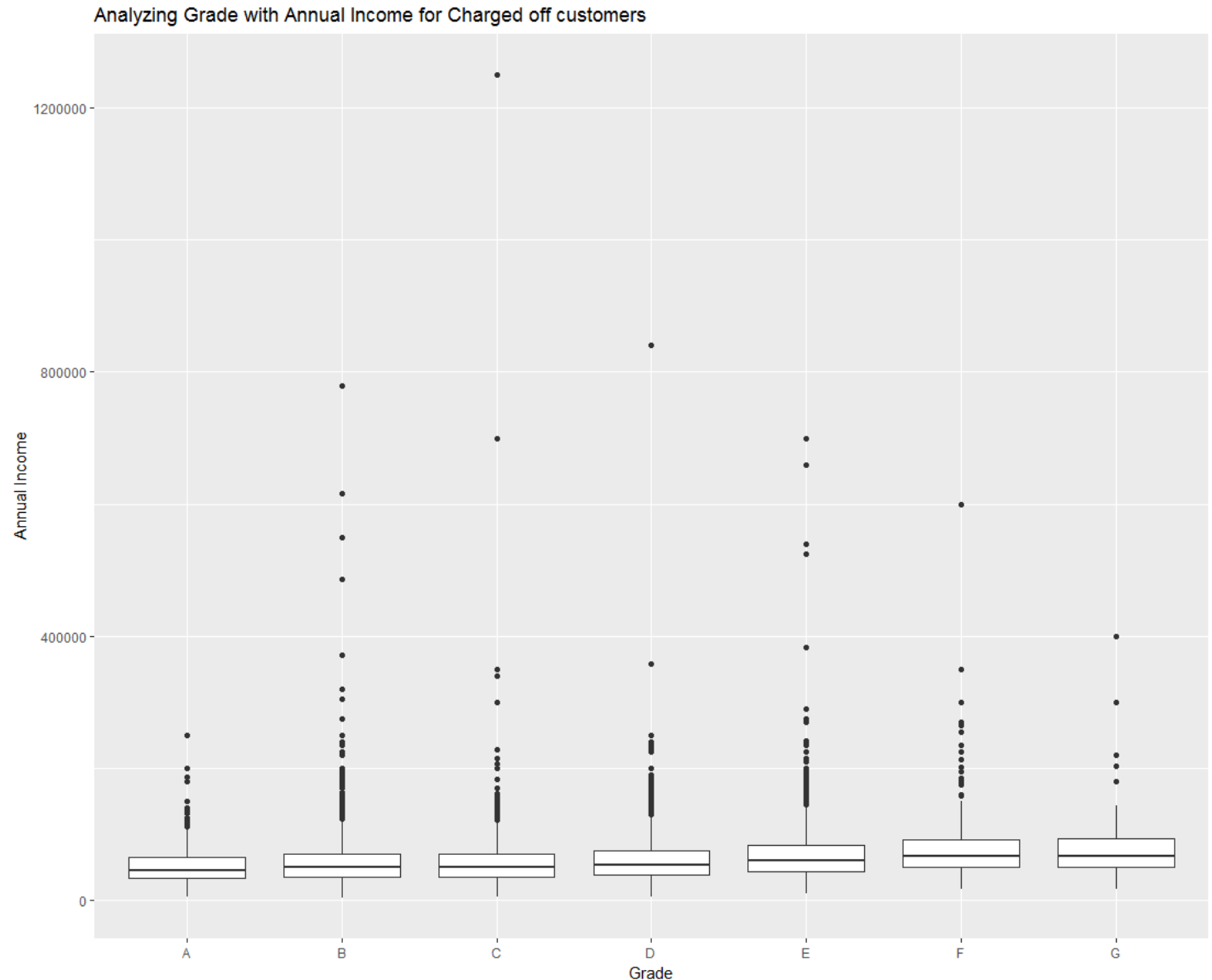
Univariate Analysis

- It can be observed that there is a spike in charged off loans for customers having first credit line year between 1994 to 2001
- For the subsequent years, number of charged off loans have dropped down

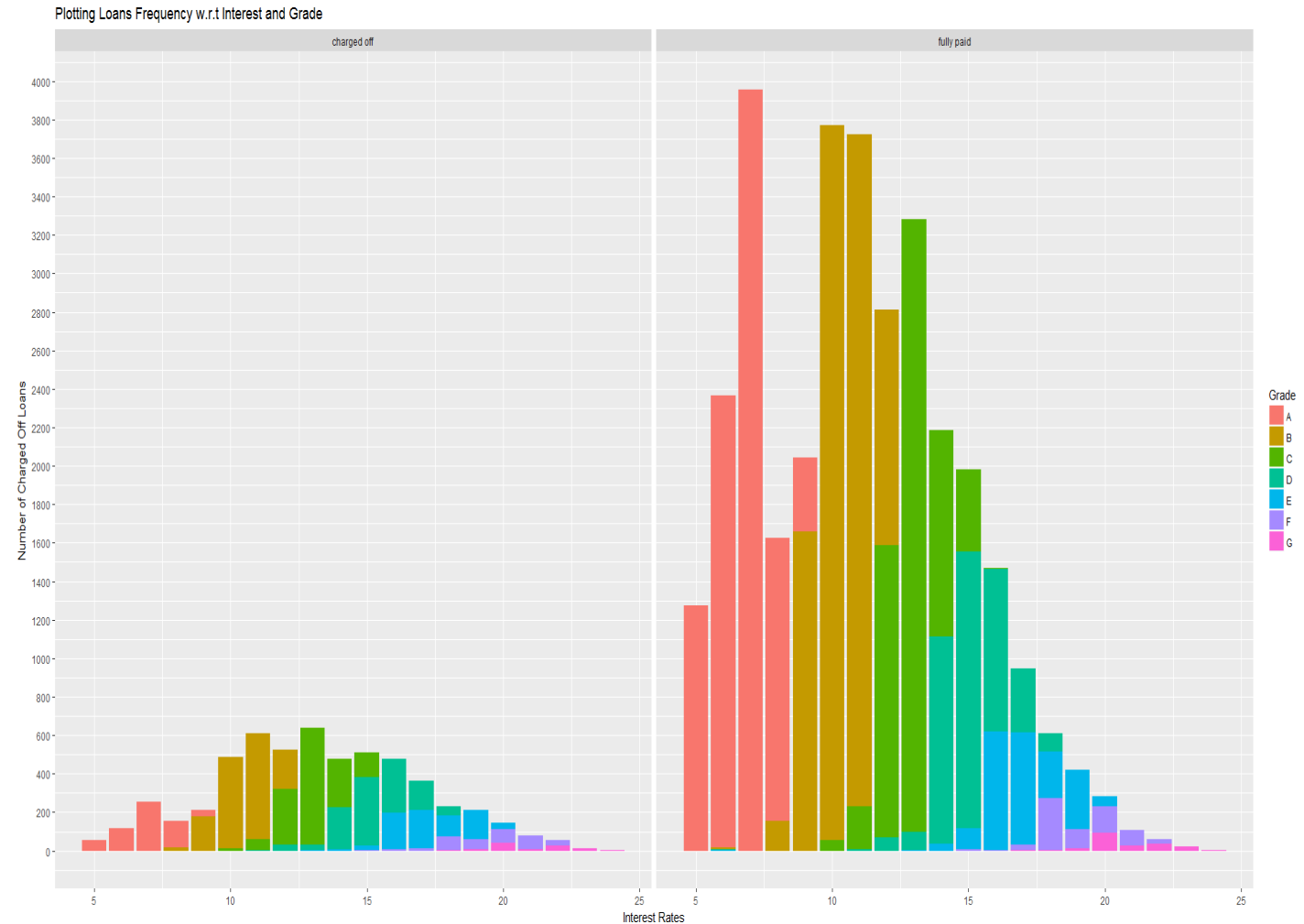
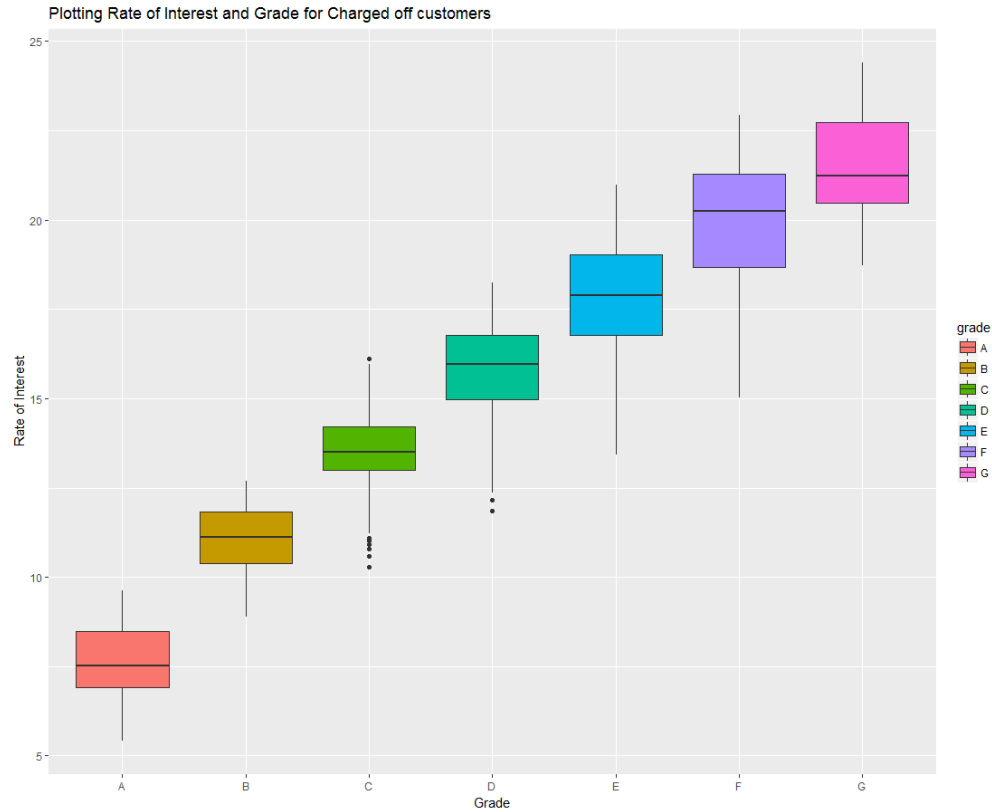


Univariate Analysis

- Annual Income of the customers who have taken loan is plotted in box plot
- There are outliers present in the annual income for customers
- 99 percentile of customers have annual income less than \$206k
- Only 57 customers have annual income more than \$206k and maximum income is \$1.25 Million



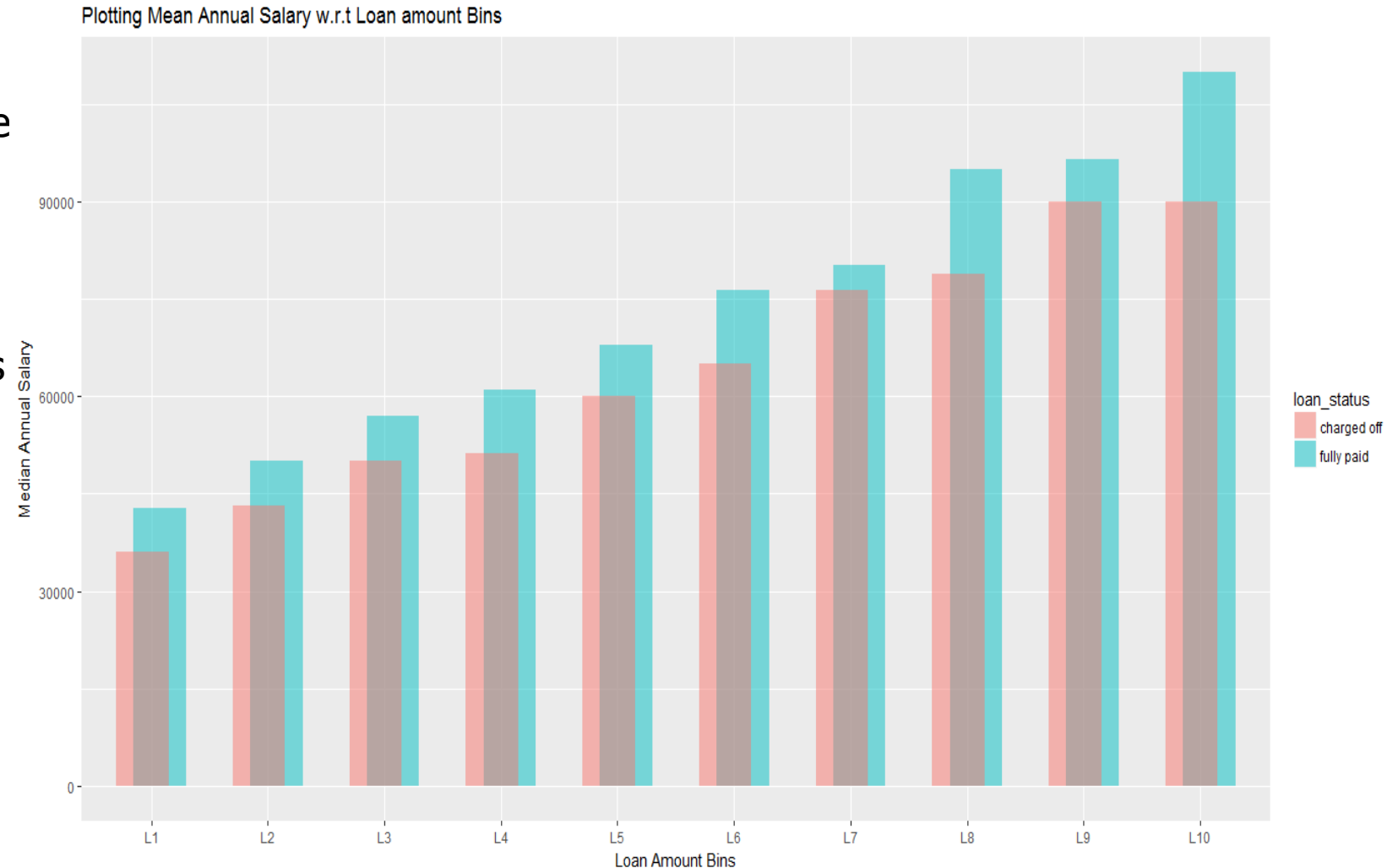
Multivariate Analysis



- It shows there is an increase in loan interest rate with change in 'grade'
- It shows there are higher defaulters amongst grades B, C and D

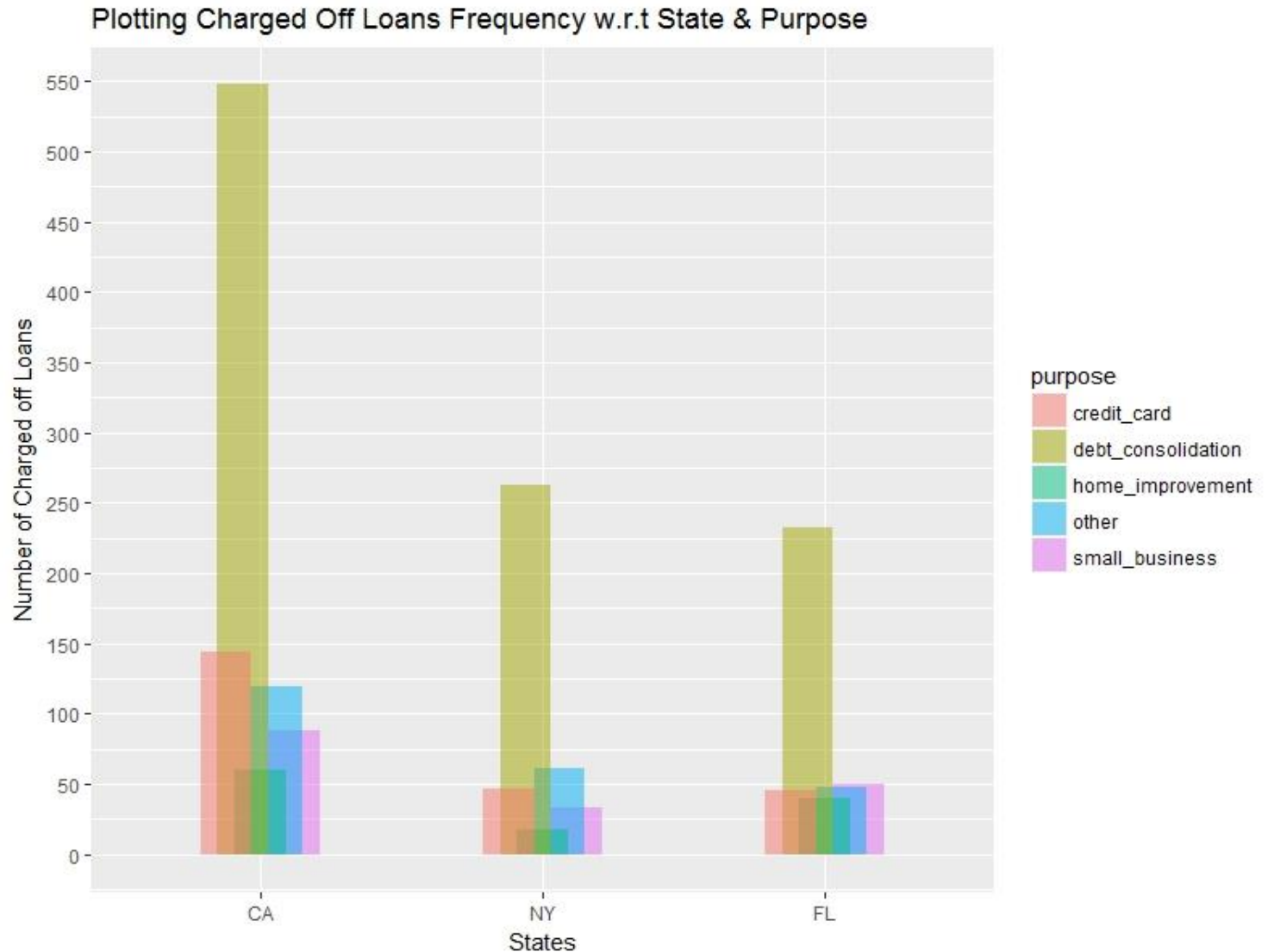
Multivariate Analysis

- Bins are created based on the loan amounts which are displayed on x axis
- Median salaries are found for the different loan amount bins and are plotted on y axis
- For higher loan amounts, difference between median income for charged off customers and full paid customers increases.



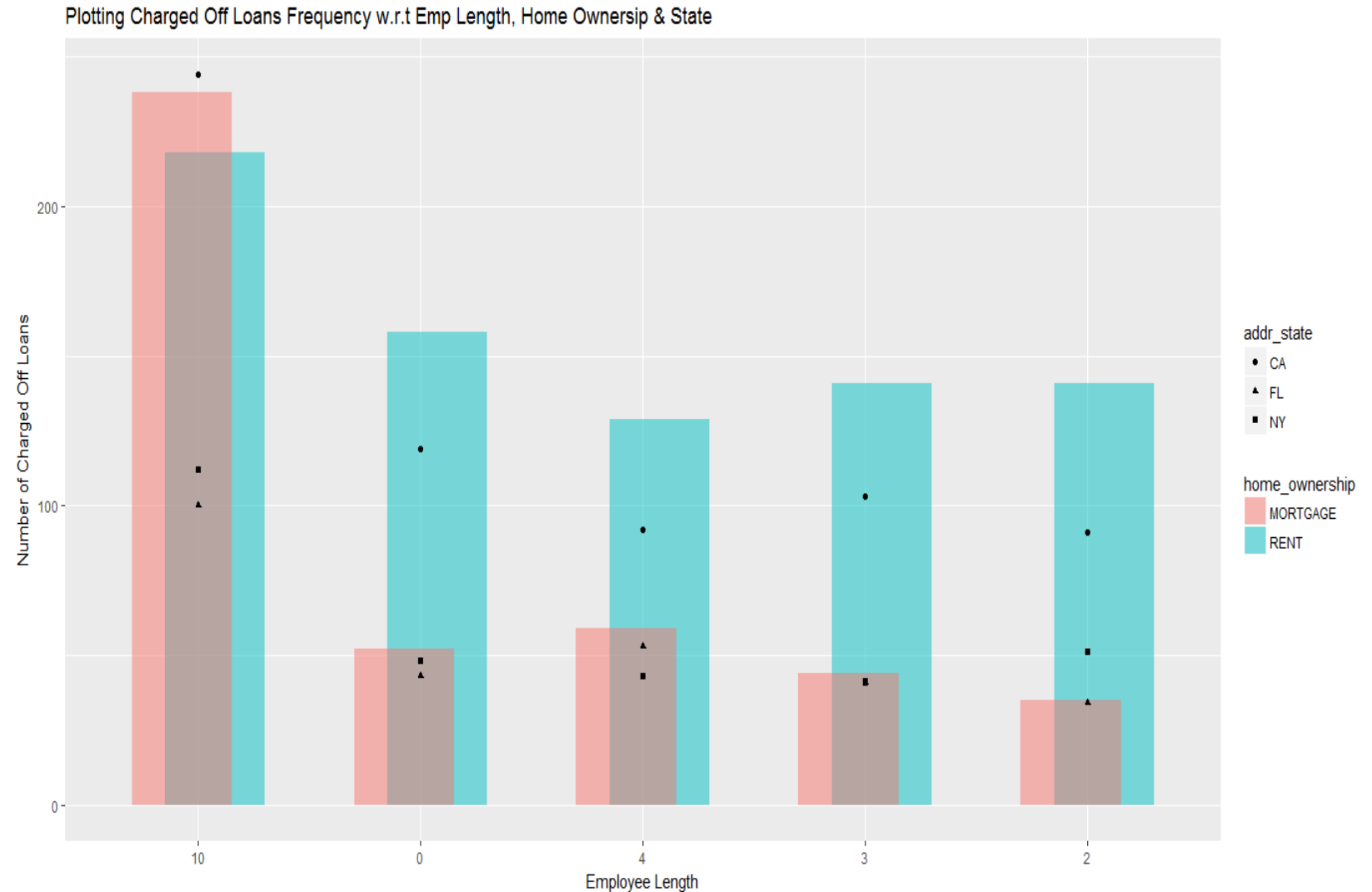
Multivariate Analysis

- Customers with loan purpose as 'debt consolidation' are highest amongst defaulters
- Customers belonging to CA, NY and FL states are major defaulters
- For these 3 states, 'debt consolidation' is the purpose of loans having the highest number of defaulters



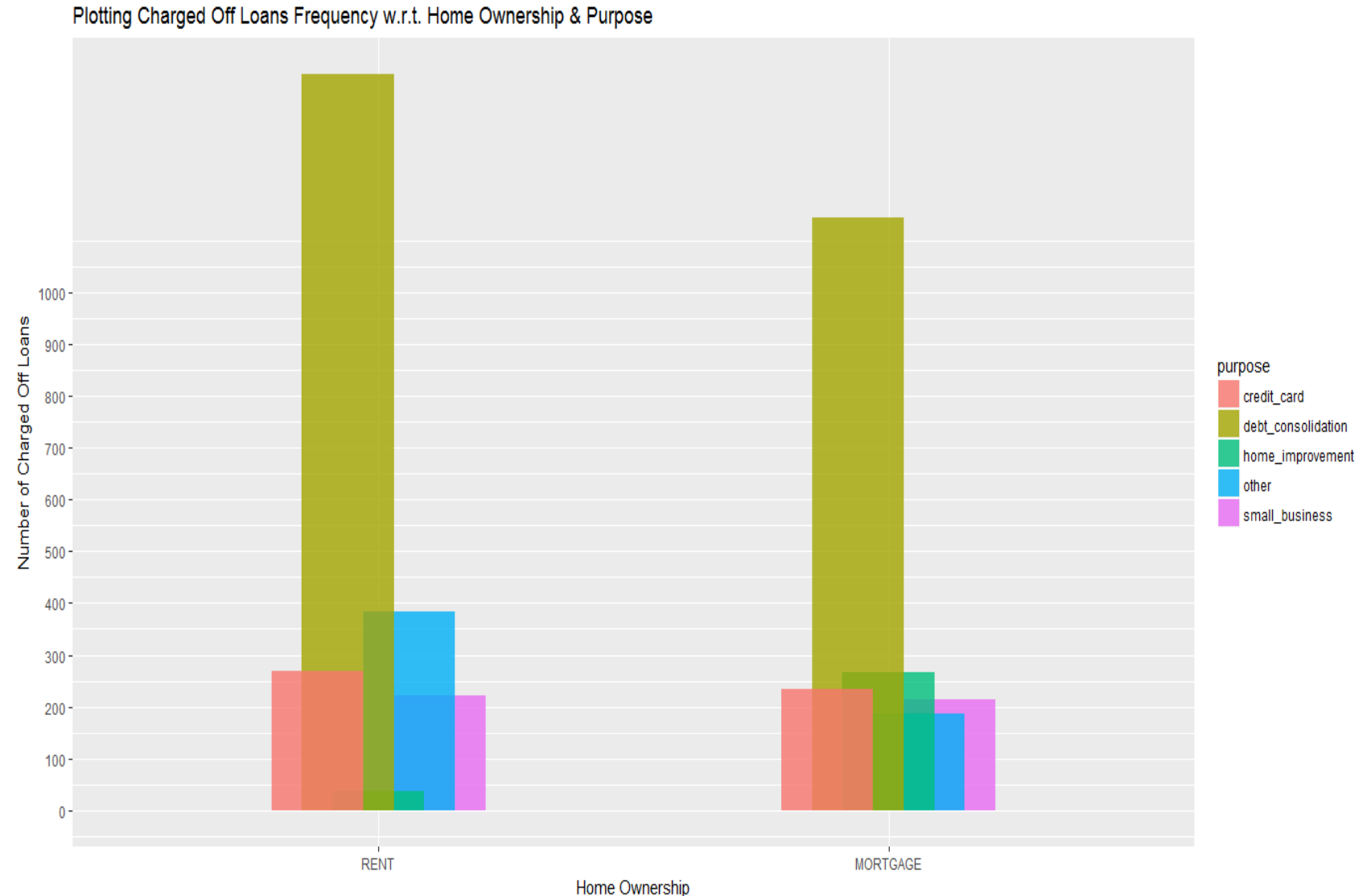
Multivariate Analysis

- Comparing employment duration with Home Ownership gives insight that customers with lower experience living on Rent tend to default more
- Conversely customers with 10+ years of experience having mortgage on their homes tend to default more
- Graph also shows number of defaulters in all three states with respect to employment duration



Multivariate Analysis

- Comparing Home Ownership and Purpose gives insight that customers with 'Rent' and 'Mortgage' home types, who have taken loan for 'debt_consolidation' are more prone to be defaulters



Conclusions

The data plots strongly indicate loan applications with more attributes meeting the below criteria increases the risk of default:

- ❖ Loan purpose as 'Debt consolidation'
- ❖ Loan Grade as 'B', 'C' or 'D'
- ❖ Home Ownership as 'Rent' or 'Mortgage'
- ❖ Employment duration with '10 years or more'
- ❖ Loan amount having round figures
- ❖ Customers having first credit line year between 1994 to 2001

Additional data points for decision making:

- In order to balance out between probable business and credit losses, the company may choose to increase the interest rates or reduce the approved loan amount
- As states of CA, NY and FL have highest loan applications and 'charged off' loans, the company needs to ensure applications from these states are carefully monitored for above attributes to avoid credit loss