Exploratory Data Analysis

Consumer Finance Company

Prepared By:
Atul Mehra
Surabhi Gupta
Kali K

Business Objective

Lending Various type of Loan are associated with some risks. We at a consumer finance company have to make a decision for loan approval based on the applicant's profile.

Two types of risks are associated with the decision:

1) If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

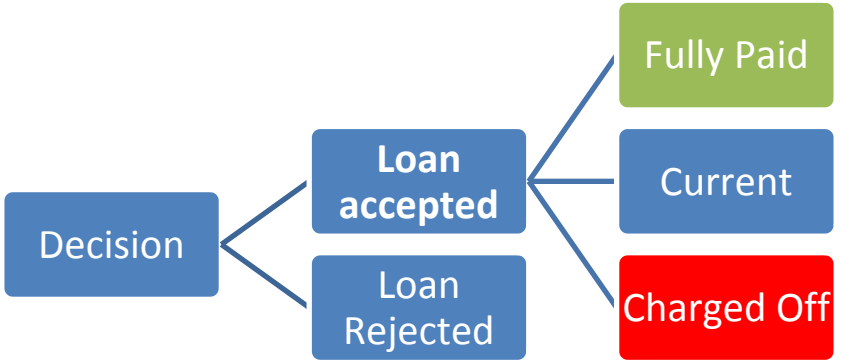
2) If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Business Understanding

- Making Decision on Loan Applications.
- Two Type of Decision can be taken:

a) **Loan accepted**

b) **Loan Rejected**



The Accepted Loan Application are at three stages:

1) **Fully Paid:** Applicant has fully paid the loan (the principal and the interest rate).
2) **Current**: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
3) **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted** on the loan.

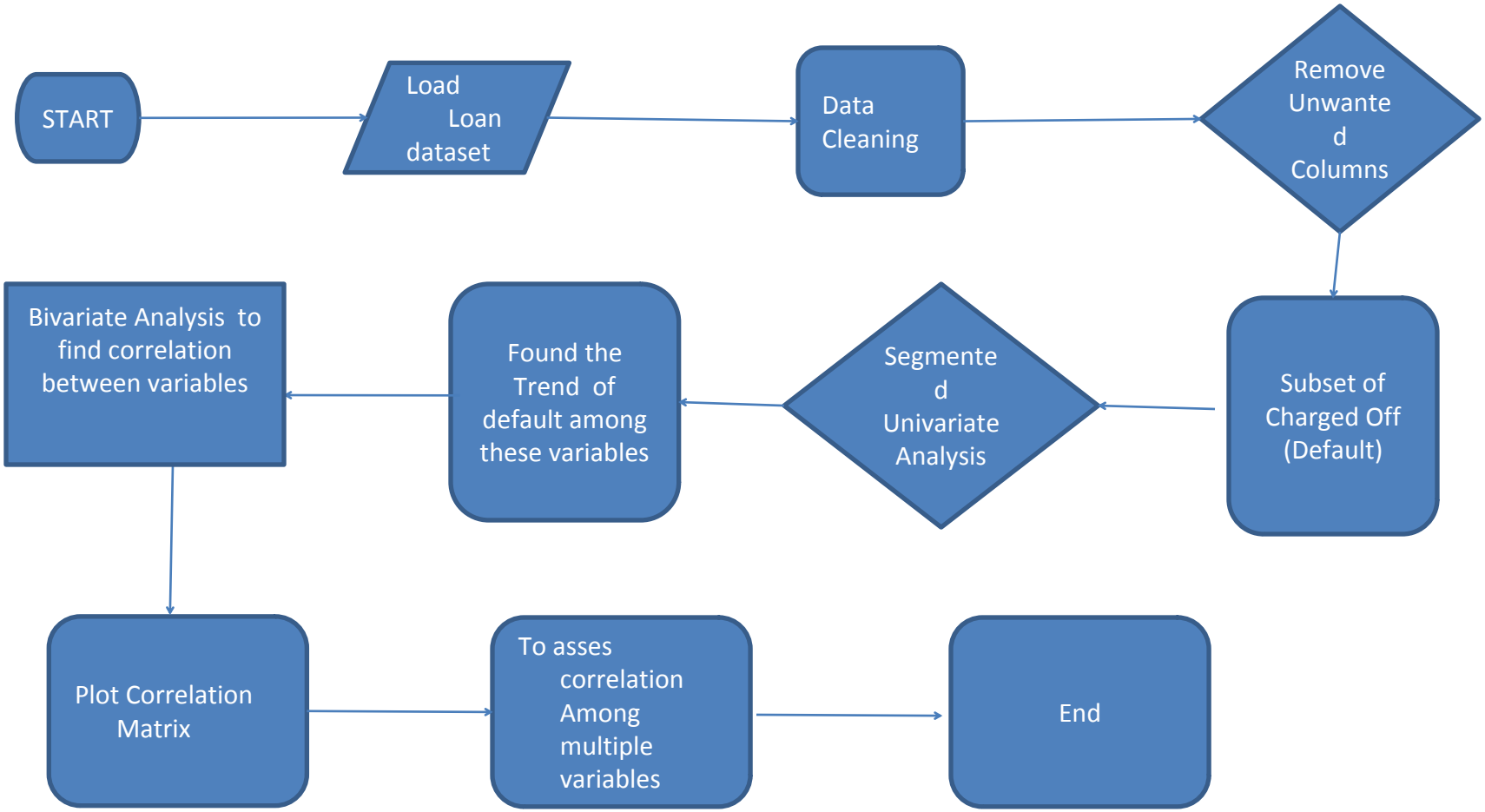
Business Understanding…contd....

- Aim is to identify the possible defaulters considering the “Loan” Dataset provided to us, which in turn help the company to cut down the Credit Loss.

Strategy

- Our Strategy for the Identification of defaulter would be to understand the **driving factors (or driver variables)** behind loan defaults. The company can utilize this knowledge for its portfolio and risk assessment.

Problem solving methodology



Data Understanding

- Loan dataset contains 111 variables.
- Loan dataset contains 39,717 Observations.
- "id" "member_id" "loan_amnt" "funded_amnt" "funded_amnt_inv" "term" "int_rate" "installment" "grade" "sub_grade" "emp_title" "emp_length" "home_ownership" "annual_inc" "verification_status" "issue_d" "loan_status" "pymnt_plan" "url" "desc" "purpose" etc.
- At first look on the dataset, it contains columns having "NA" , "0" values.
- The columns like "loan_amnt" "funded_amnt" "funded_amnt_inv" are seems to have similar values, which are having high correlation. Thus, two of them can be removed or keep out of analysis. This is because all these serve approximately same purpose.

Predictor(Input/feature) and Target(output) variables.

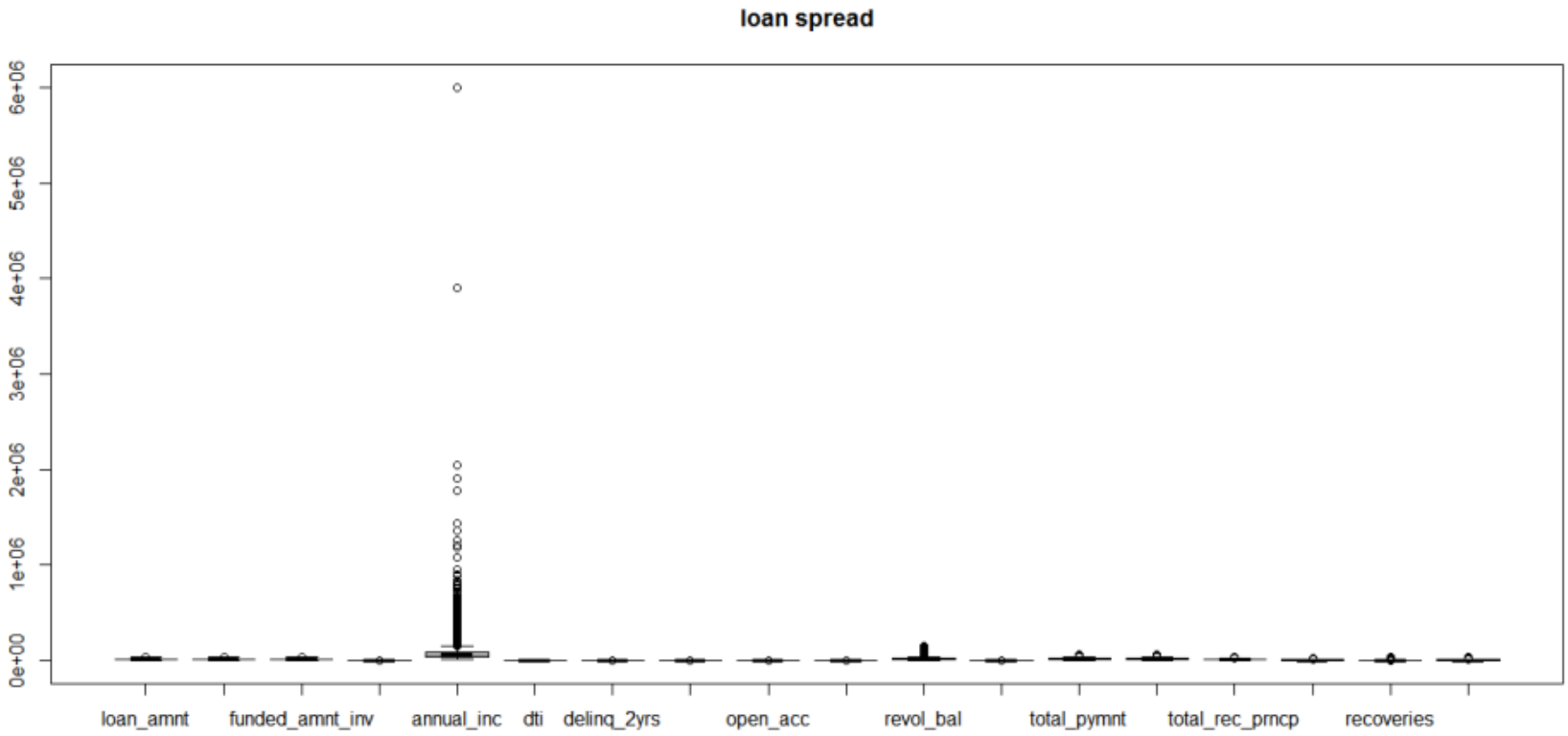
1. **Target: “**loan_status**”**

2. **Predictor**:

   "funded_amnt" "funded_amnt_inv“ "term" "int_rate“ "installment" "grade" "sub_grade“ "emp_title" "emp_length“ "annual_inc" "verification_status" "issue_d" "pymnt_plan" "url" "desc" "purpose" "title"

   "dti“ "delinq_2yrs" "earliest_cr_line" "inq_last_6mths“ "open_acc" "pub_rec" "revol_bal“ "revol_util" "total_acc" "initial_list_status“ "out_prncp" "out_prncp_inv" "total_pymnt“ "total_pymnt_inv" "total_rec_prncp" "total_rec_int“ "total_rec_late_fee" "recoveries" "collection_recovery_fee“ "last_pymnt_d" "last_pymnt_amnt" "last_credit_pull_d“ "collections_12_mths_ex_med" "policy_code" "application_type" "acc_now_delinq" "chargeoff_within_12_mths" "delinq_amnt“ "pub_rec_bankruptcies" "tax_liens"

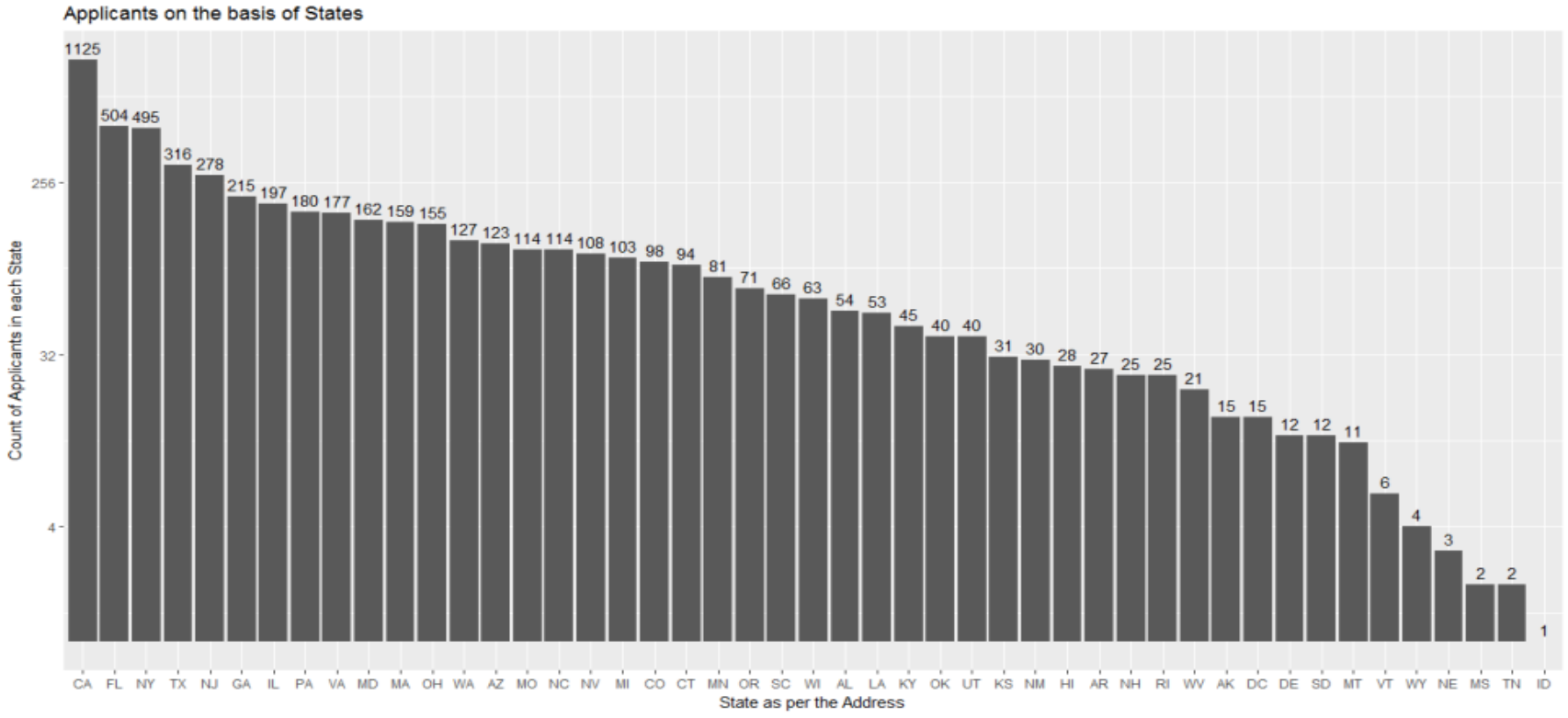
Box Plot Depicting spread of Variables



Univariate Analysis

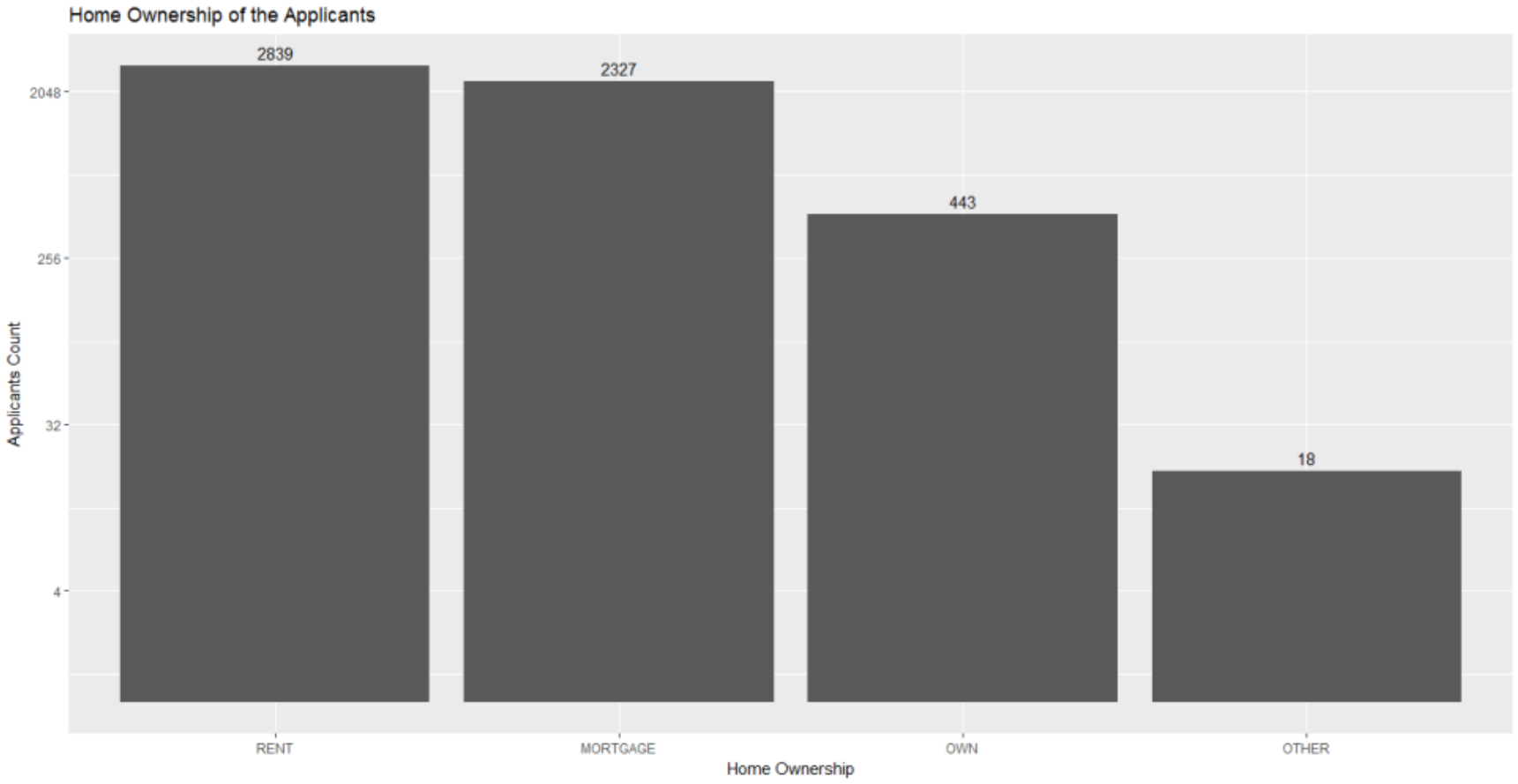
Variables related to customer demographic:

Address state : From plot below it can be seen that the loan applicant from State "CA".



Variables related to customer demographic:

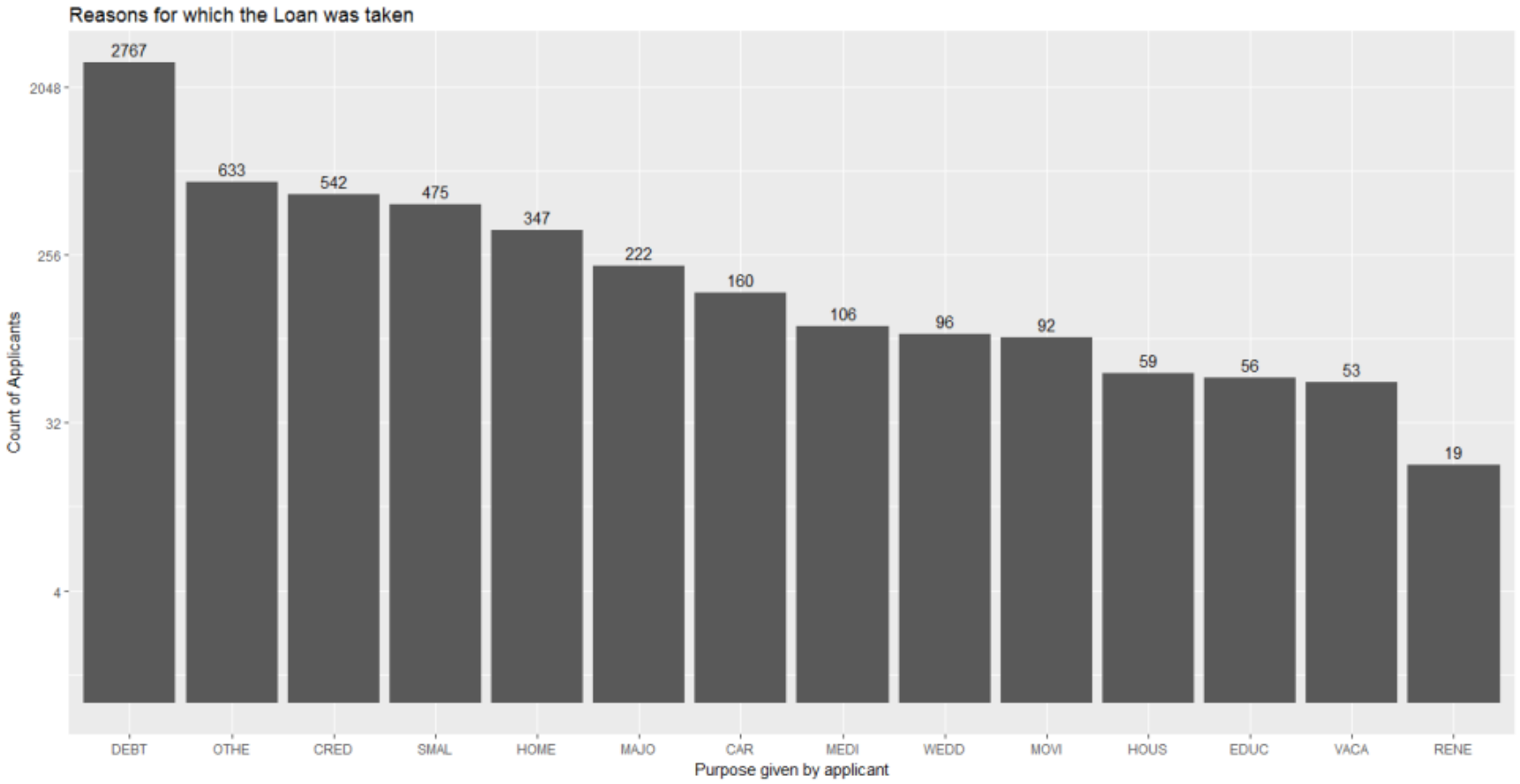
home_ownership : Applicant living on Rent are more prone to do default.



Variables related to customer information

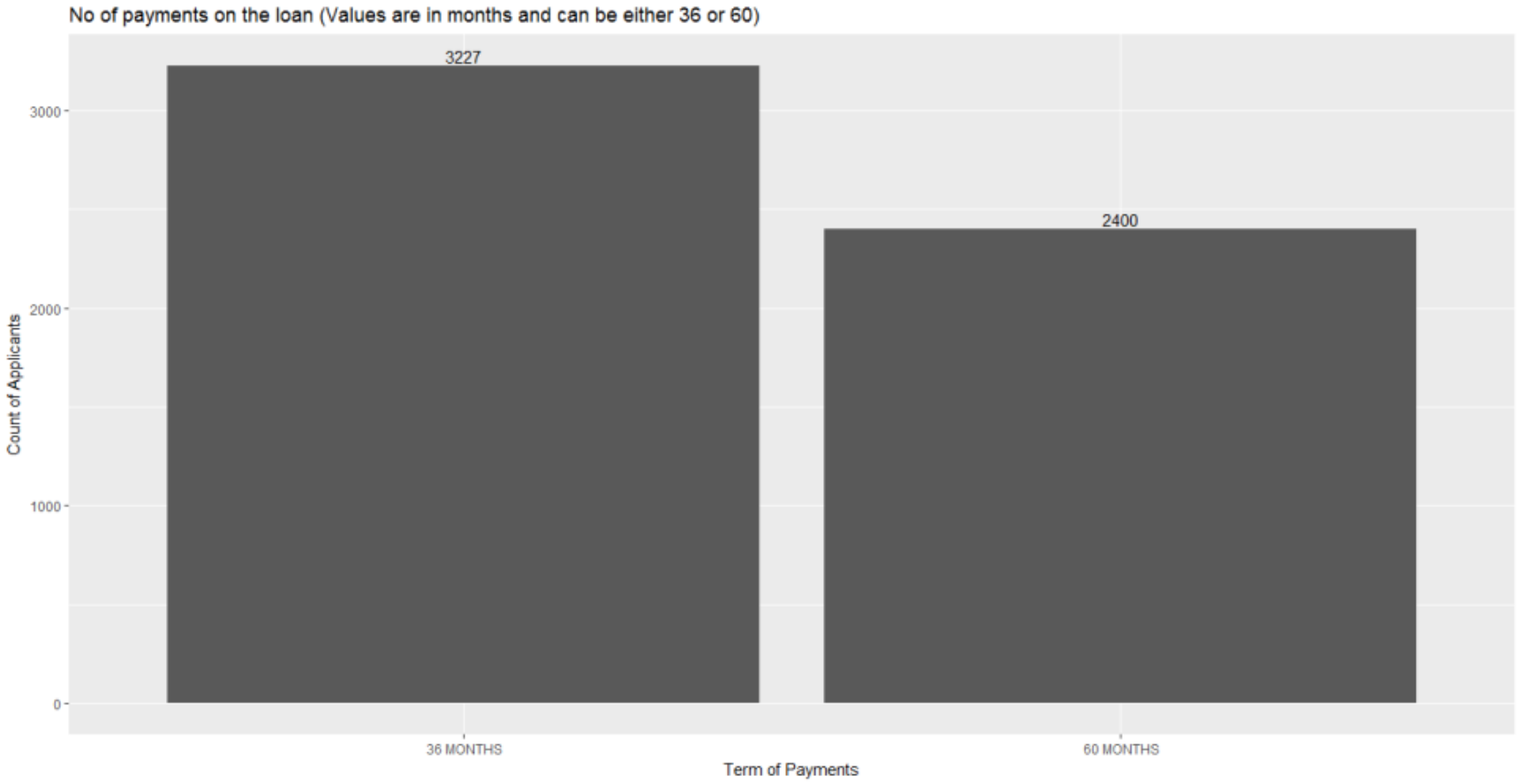
purpose :

Applicants whose purpose of taking loan is debt consolidation are more prone to do default



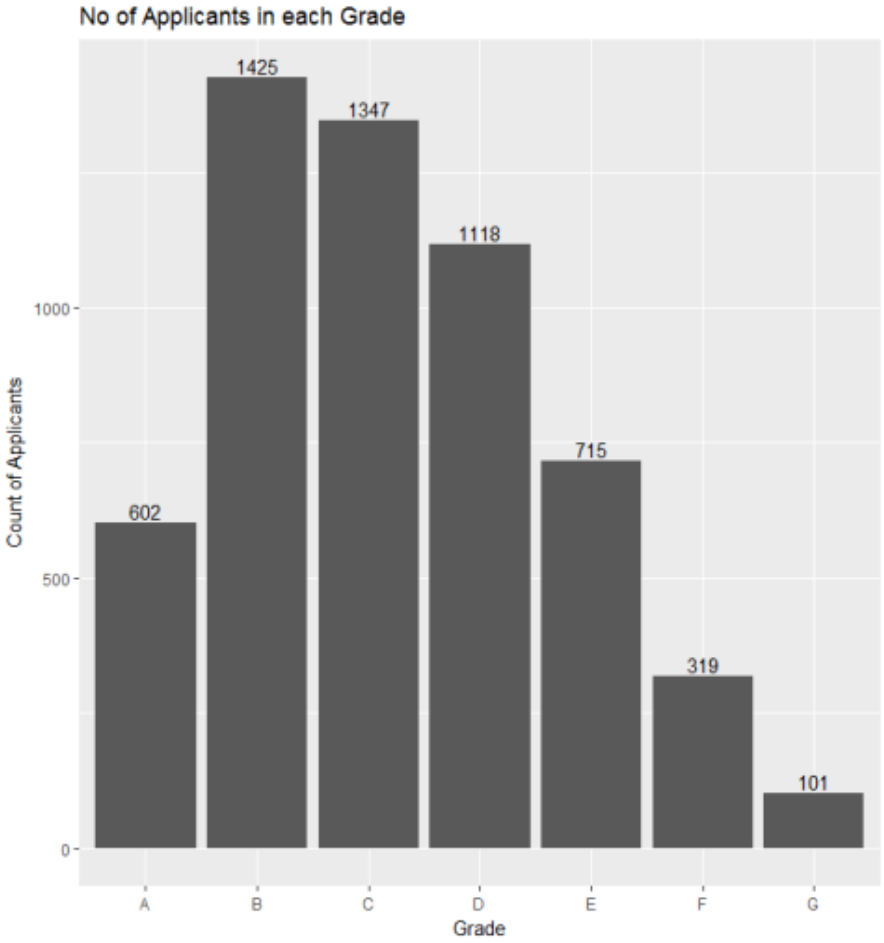
Variables related to customer information

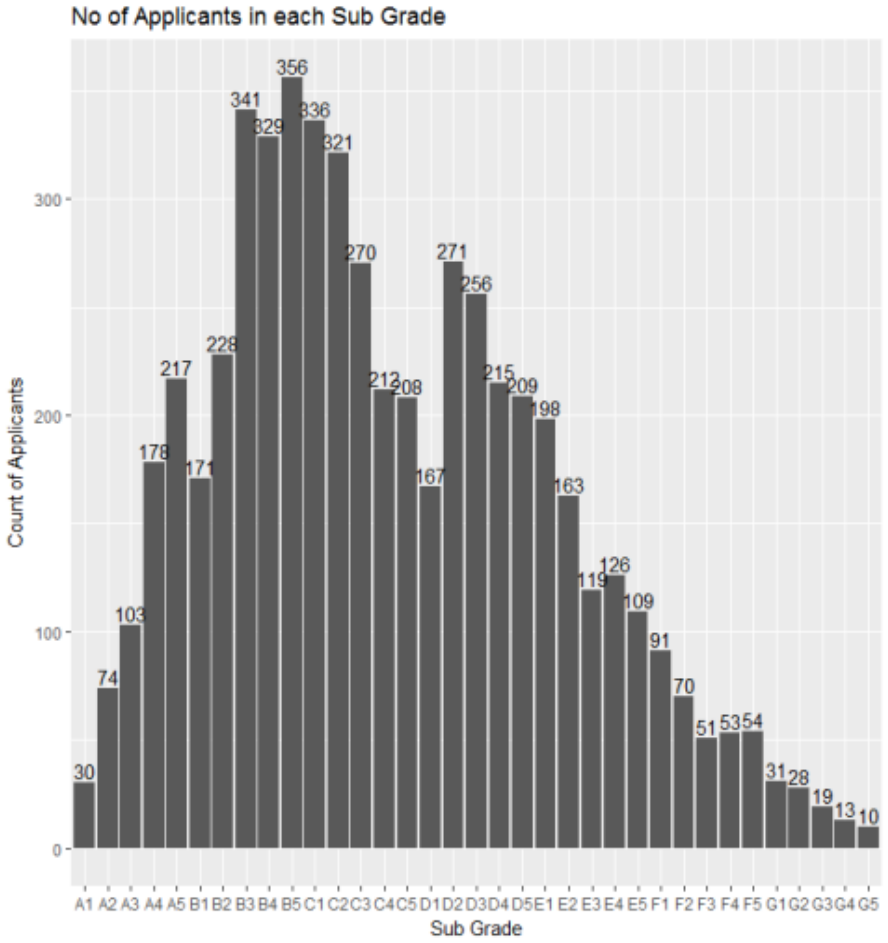
Term : From plot below it can be seen that the loan applicant having term loan for 36 Month are more tend to do default.



Grades & Subgrades

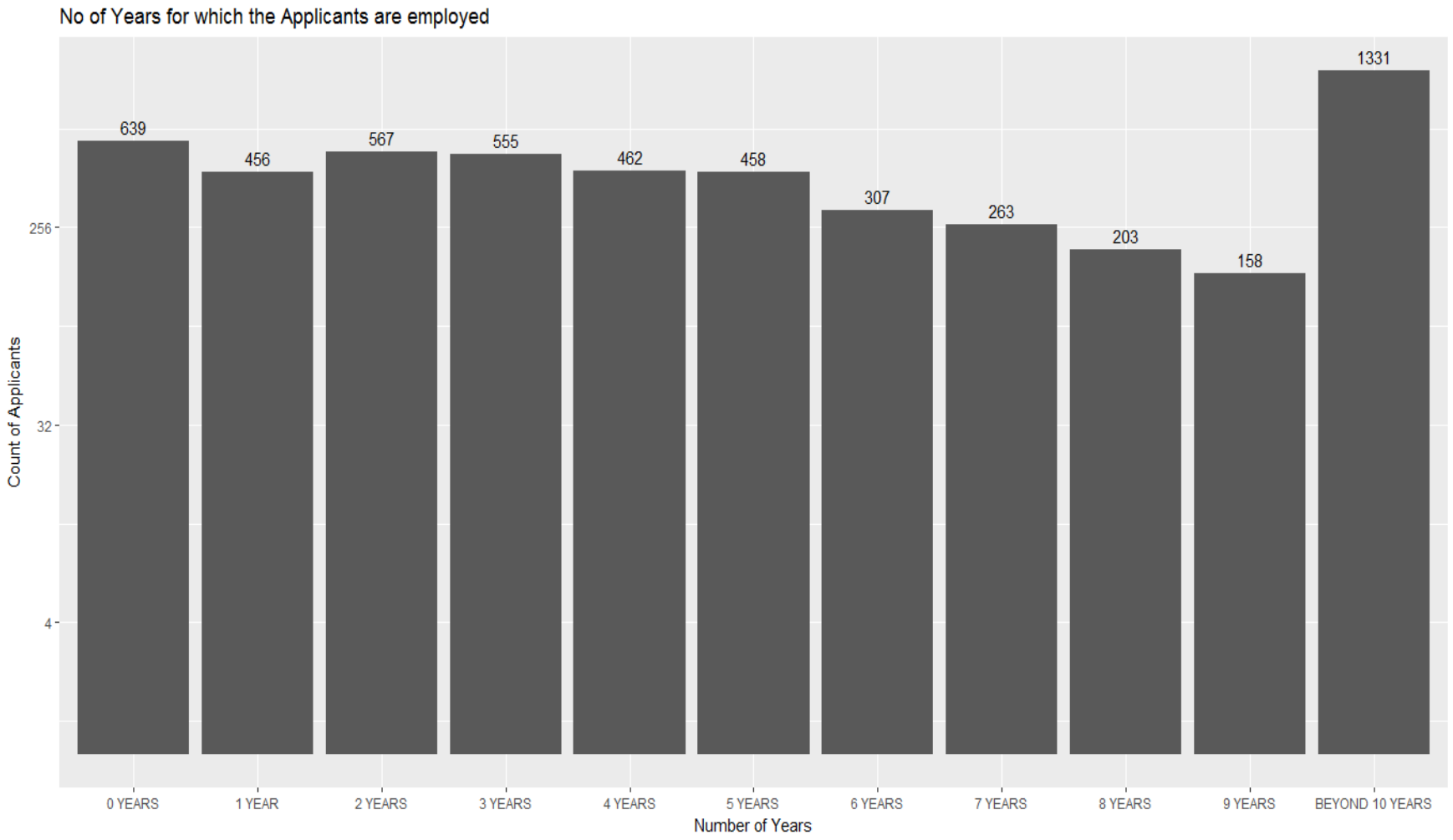
B & C grades are showing most of the Defaulted loans. B3,B4,B5,C1,C2 sub-categories are showing most of the Defaulted loans.





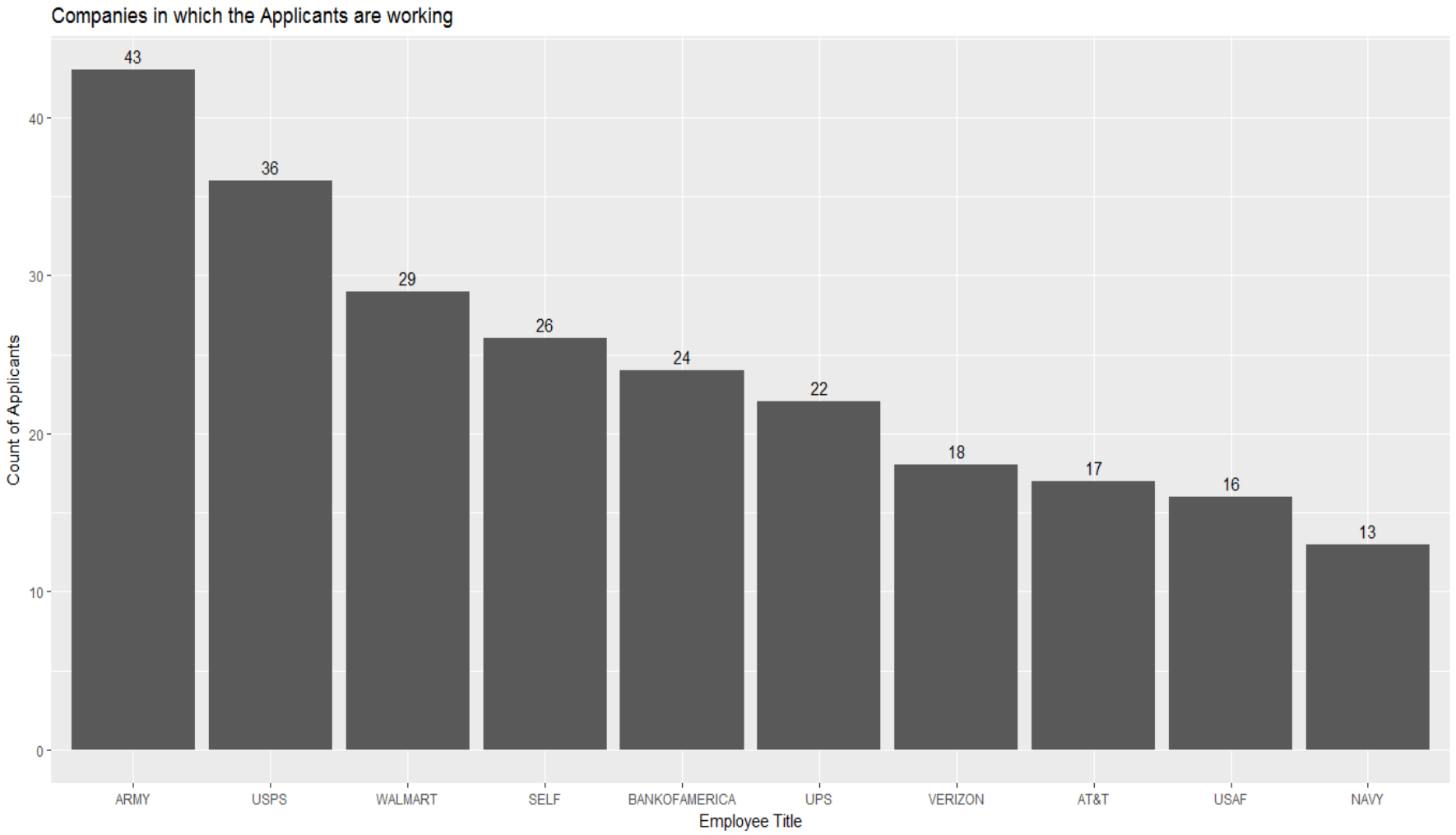
emp_length

Applicants having 10+ years of experience tend to default more



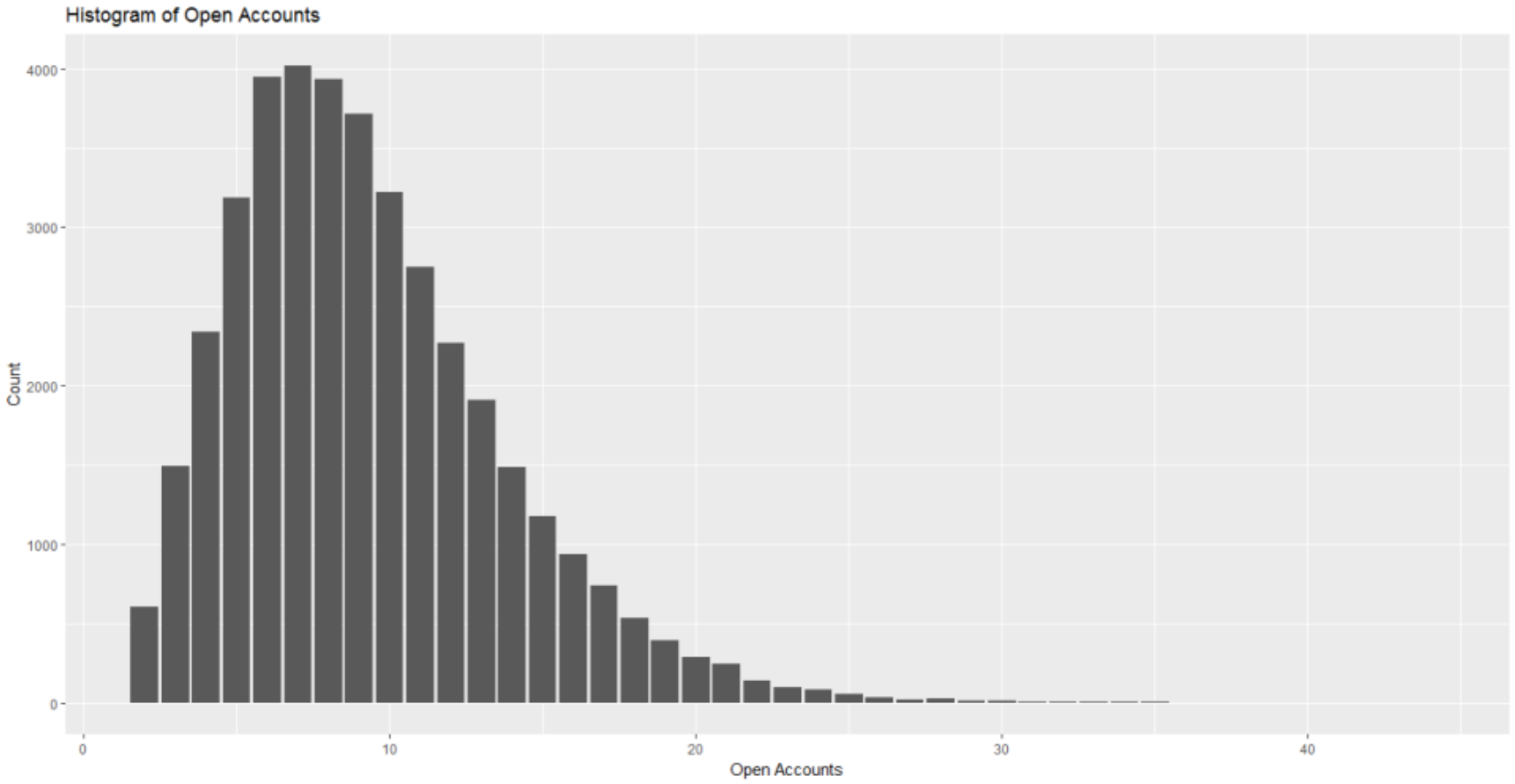
emp_title

Applicants from Army profession are showing more tendency to default.



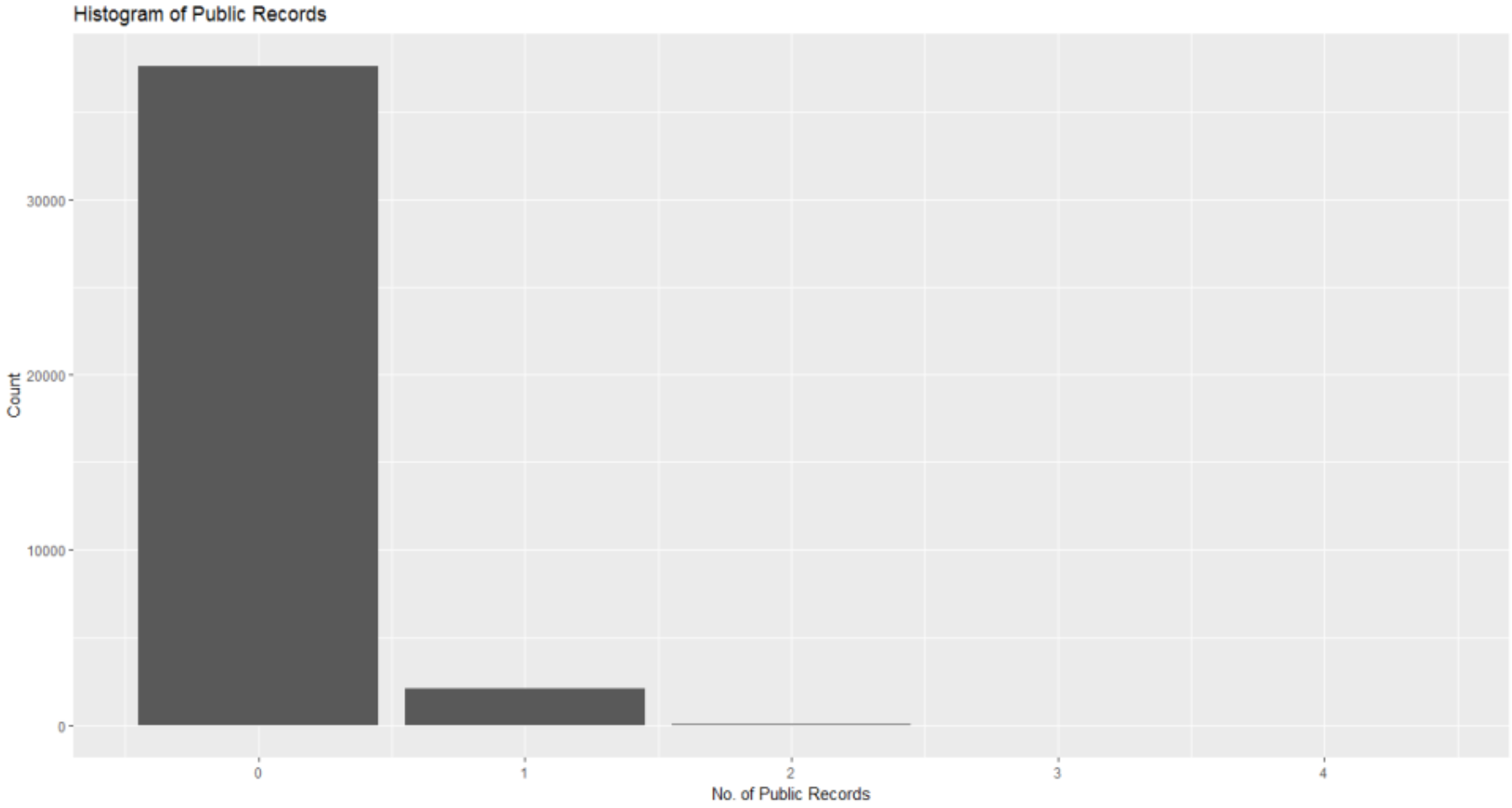
open_acc

Those having open account (or current credit) ranging from 6-9 are majorly involve in defaults



pub_rec

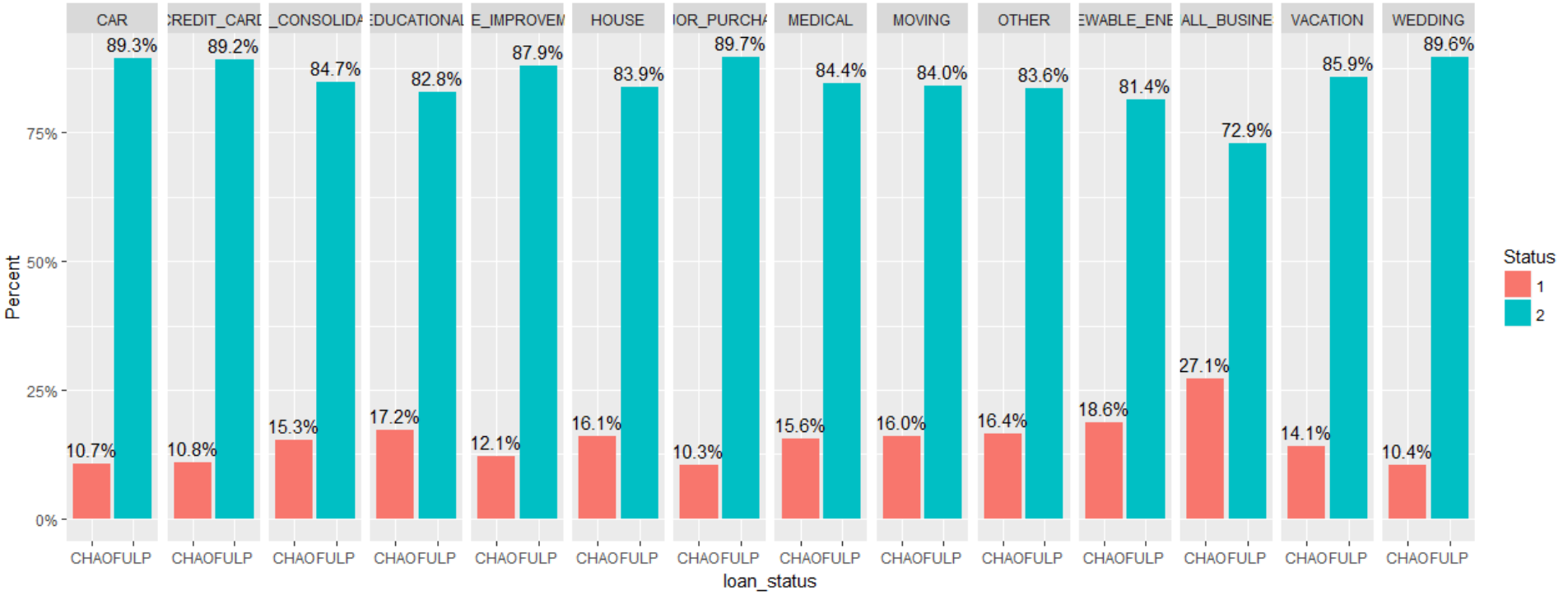
Those having no derogatory record (showing a critical or disrespectful attitude) top the list of defaulters it means there is nothing to do with their public attitude & default.



Bivariate Analysis

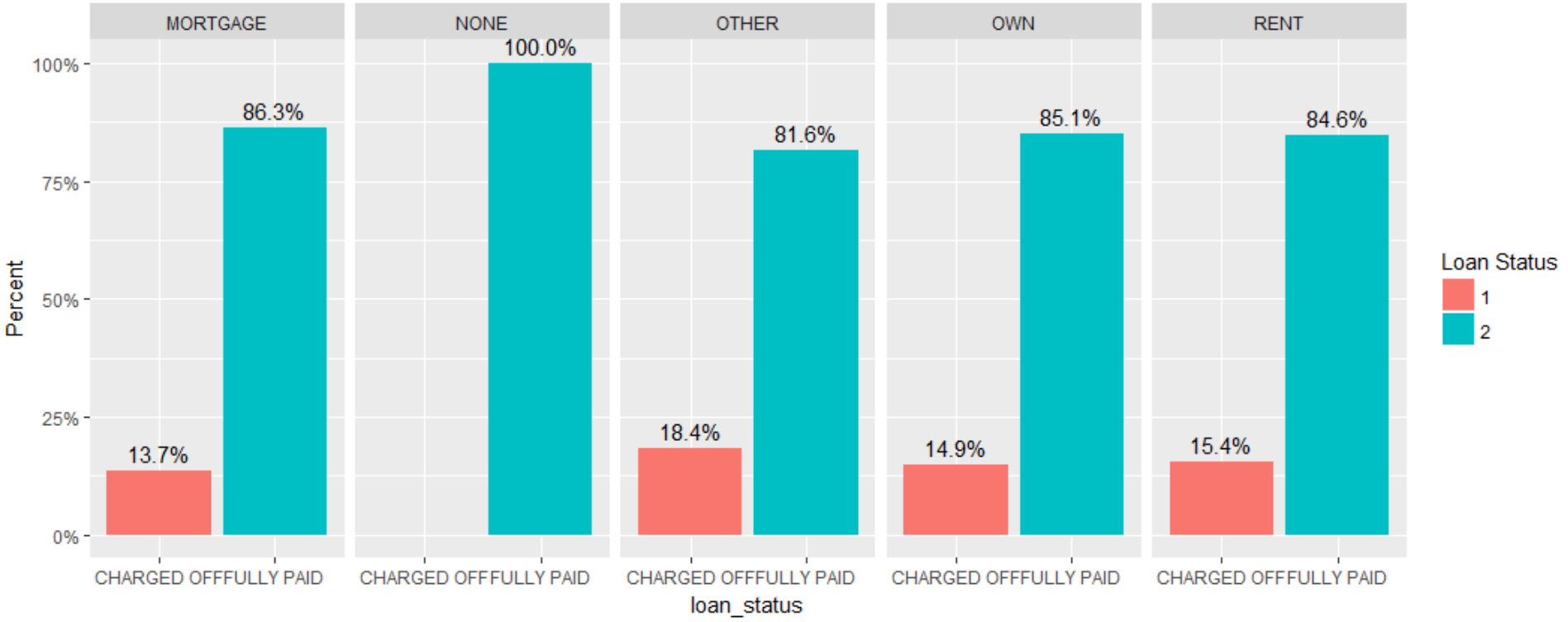
Purpose & Loan Status

Looking in terms of percentage small business creating more default



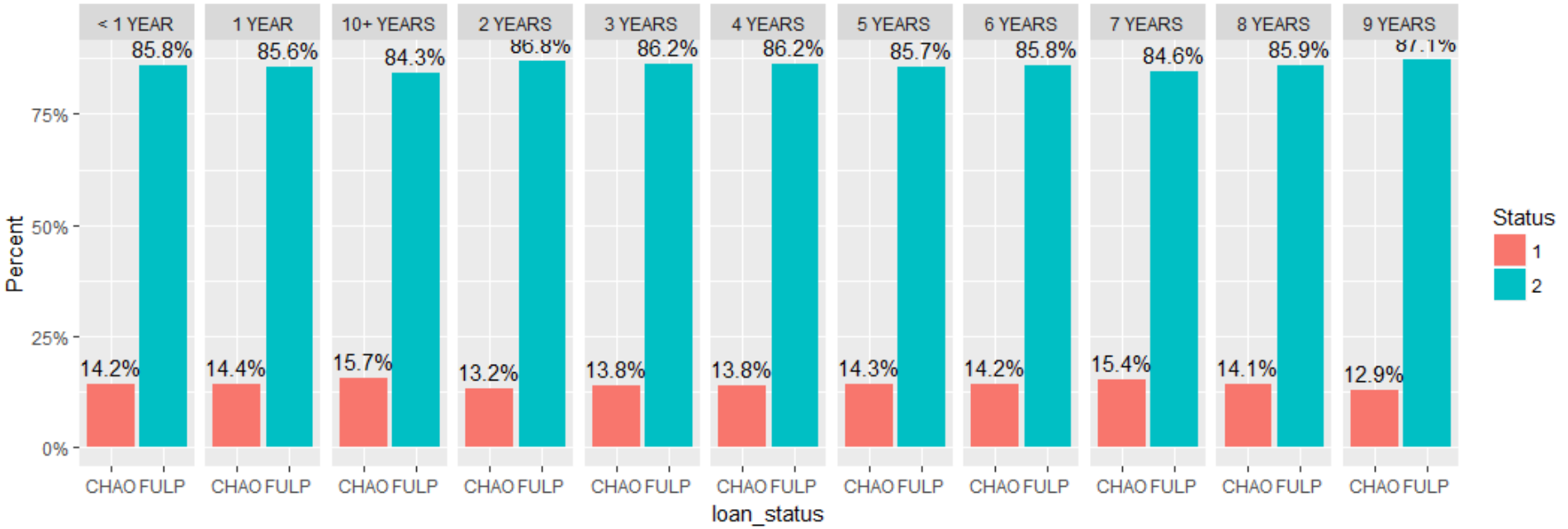
Home ownership &Loan status

Rent Residents more prone to do default



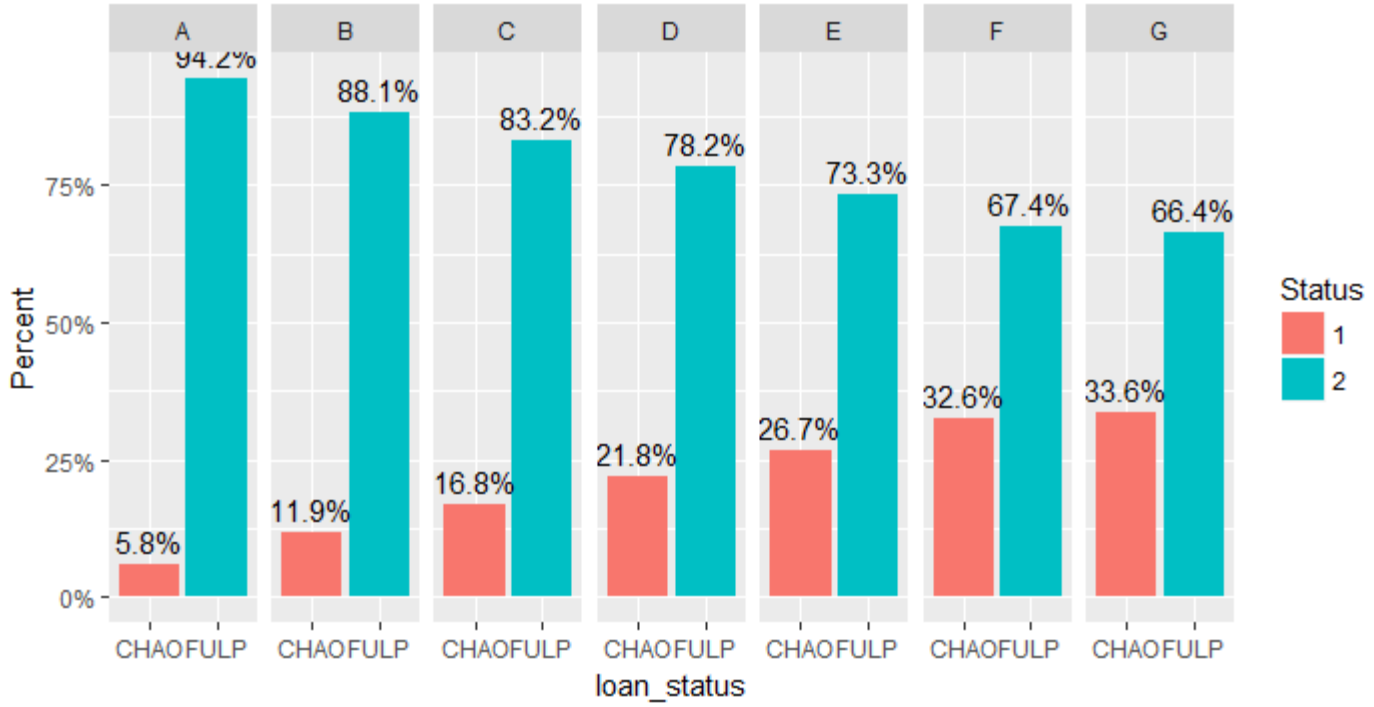
Emp_length & Loan status

Applicants having 10+ years of experience tend to default more



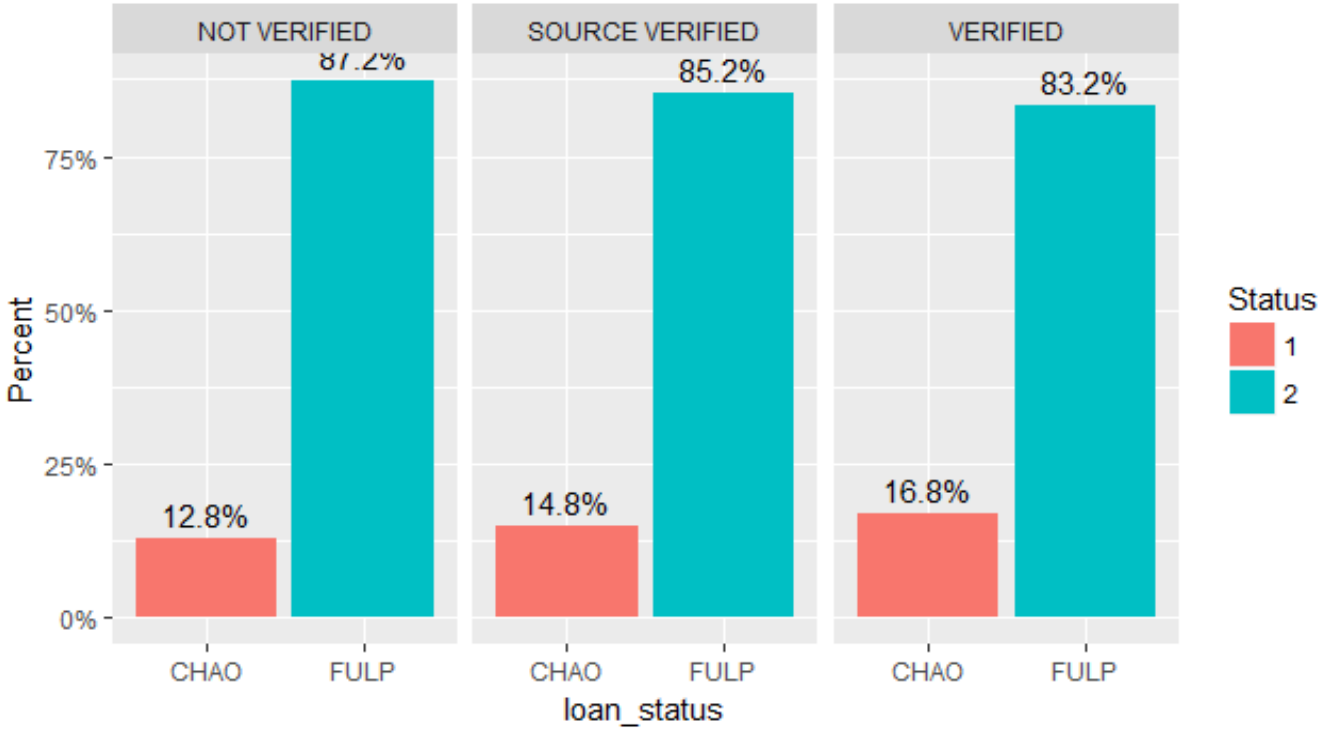
Grade & Loan Status

G grade showing more percent wise default



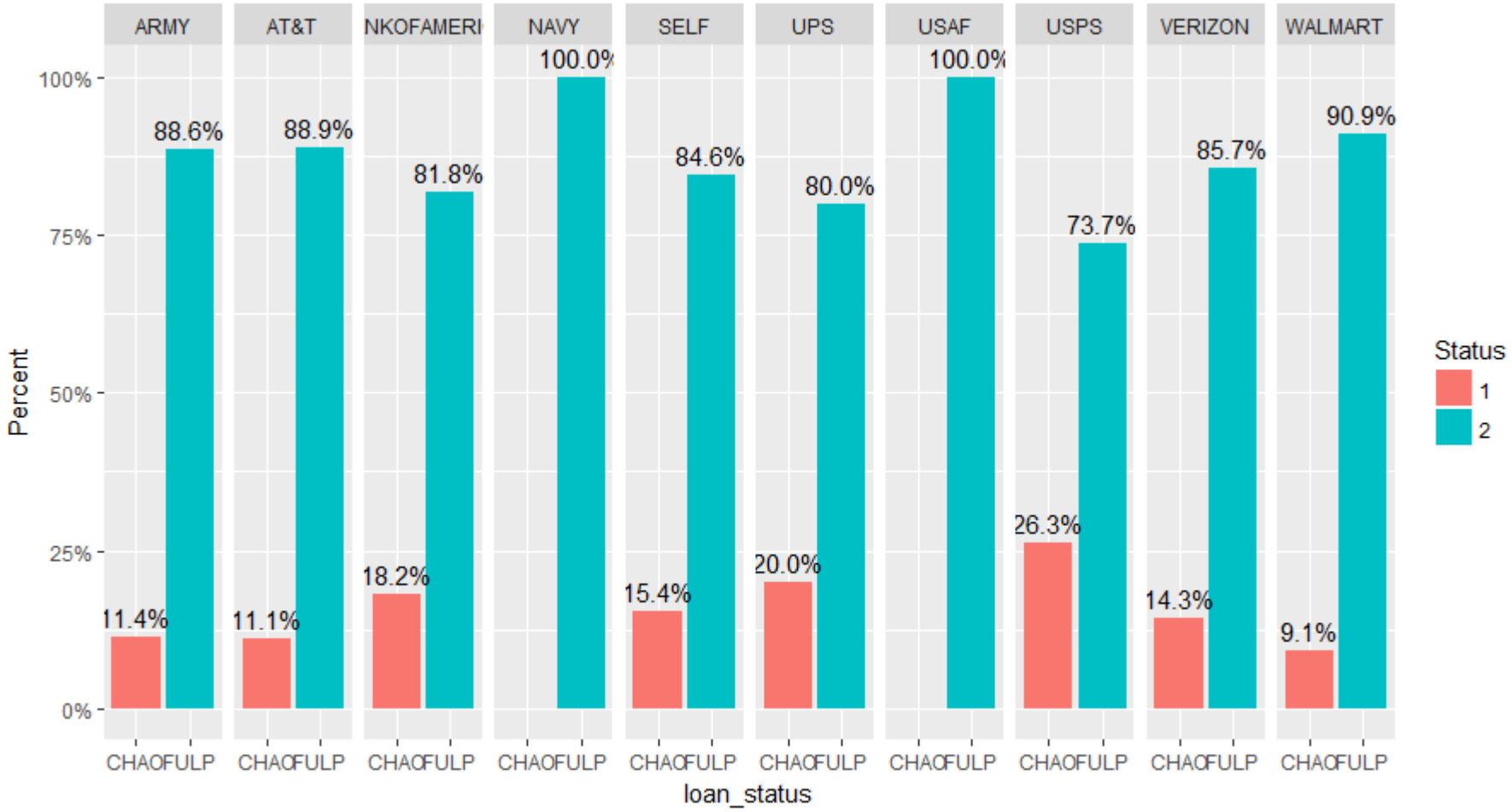
Verification Status & Loan Status

Verified Income status applicants are more tend to do defaults

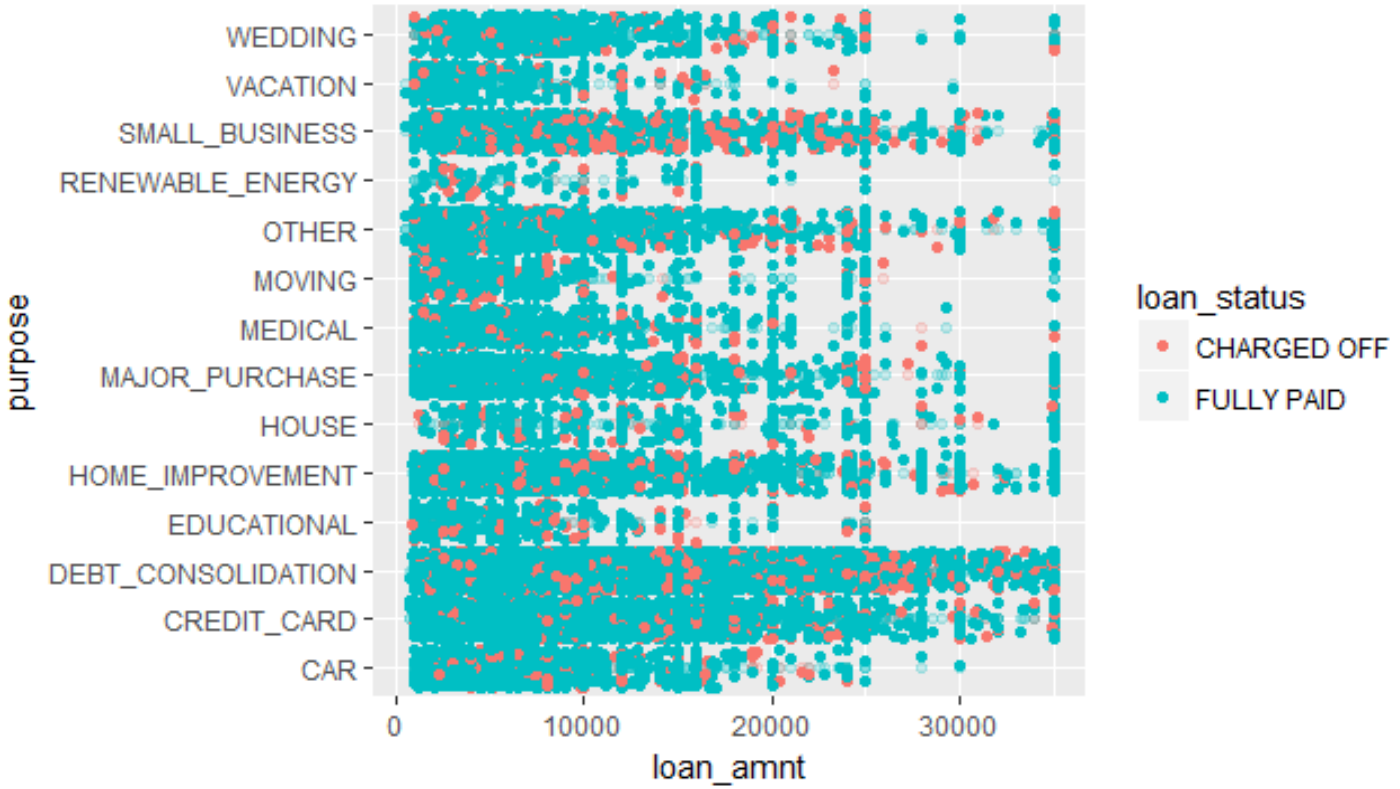


emp_title

Employers working in USPS are more likely to default



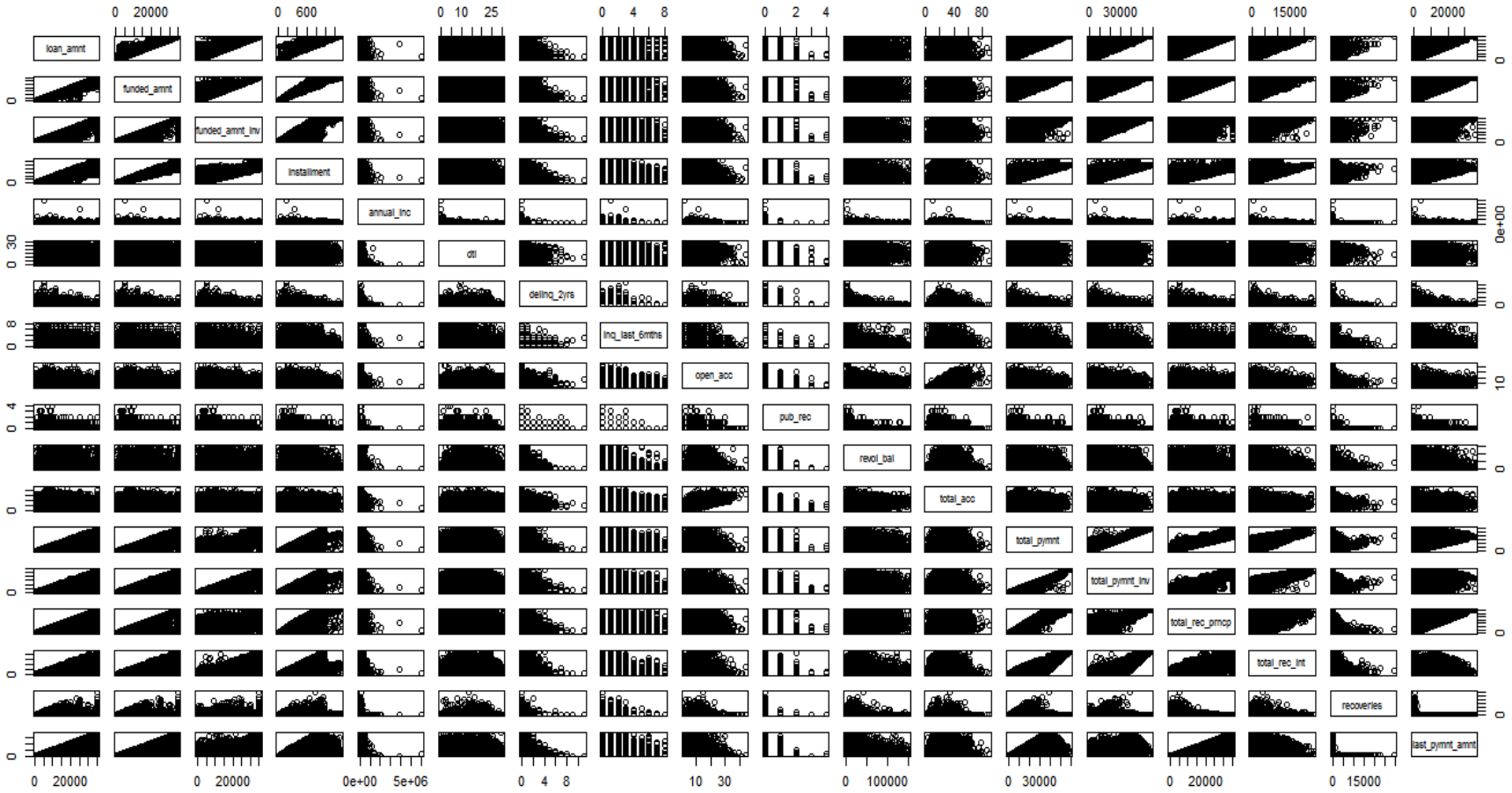
Purpose : Density of Red spot i.e. Charged off cases is more in Small Business & debt consolidations.



Correlation Among numerical variables

From Correlation it can be seen that the "loan_amnt" "funded_amnt" "funded_amnt_inv“ have high correlation . So only one variable can serve the purpose of analysis among the three.

Similarly, "total_pymnt" "total_pymnt_inv" "total_rec_prncp" have high correlation. All will behave in similar manner.

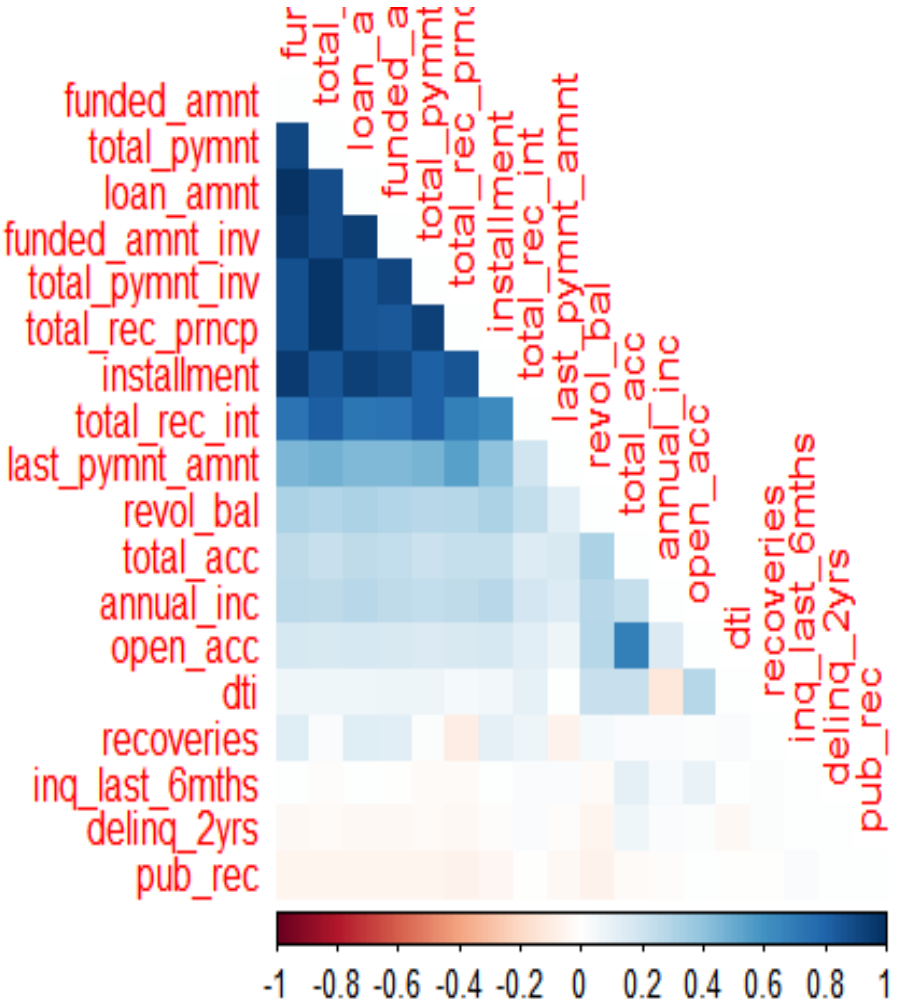


Correlation Matrix

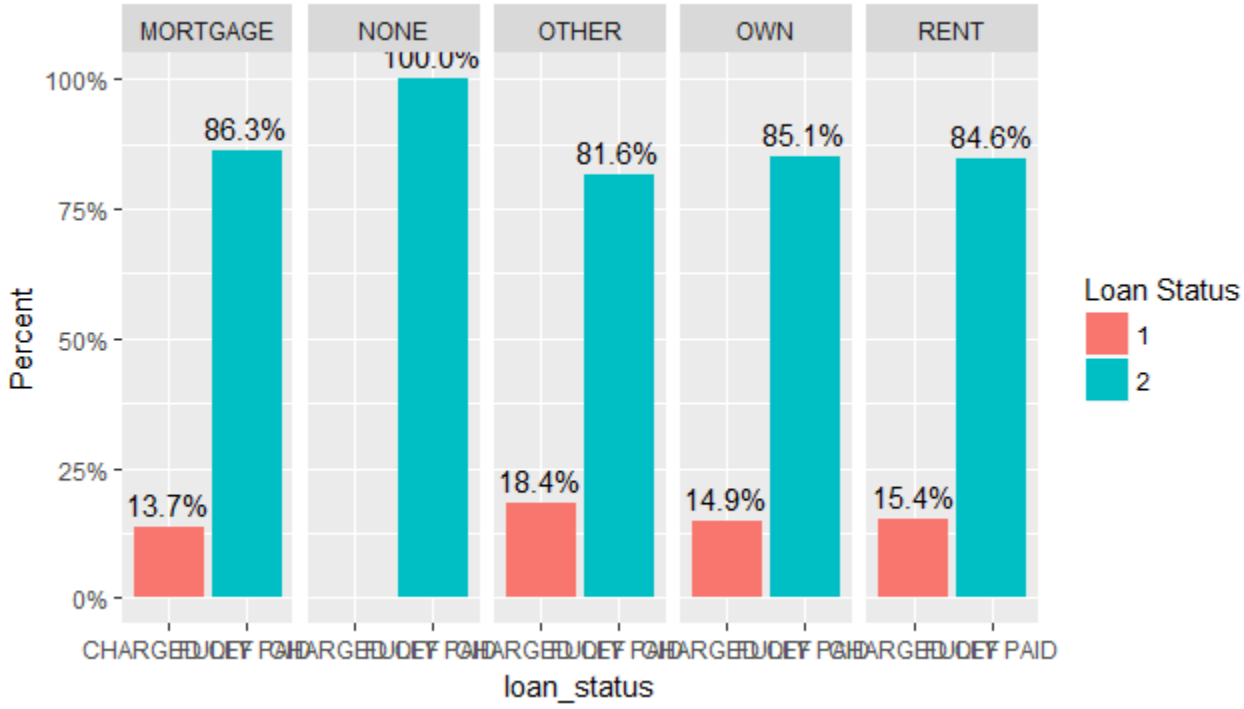
1. The Dark Blue color specifies High Correlation. White color signifies slight negative or no correlation between variables.

2. From Matrix , We infer that "loan_amnt" "funded_amnt" "funded_amnt_inv“ have high correlation among each other as well as with "total_pymnt" "total_pymnt_inv" "total_rec_prncp“.

3. dti & Annual Income has slight negative correlation. As Annual Income rises dti decreases.

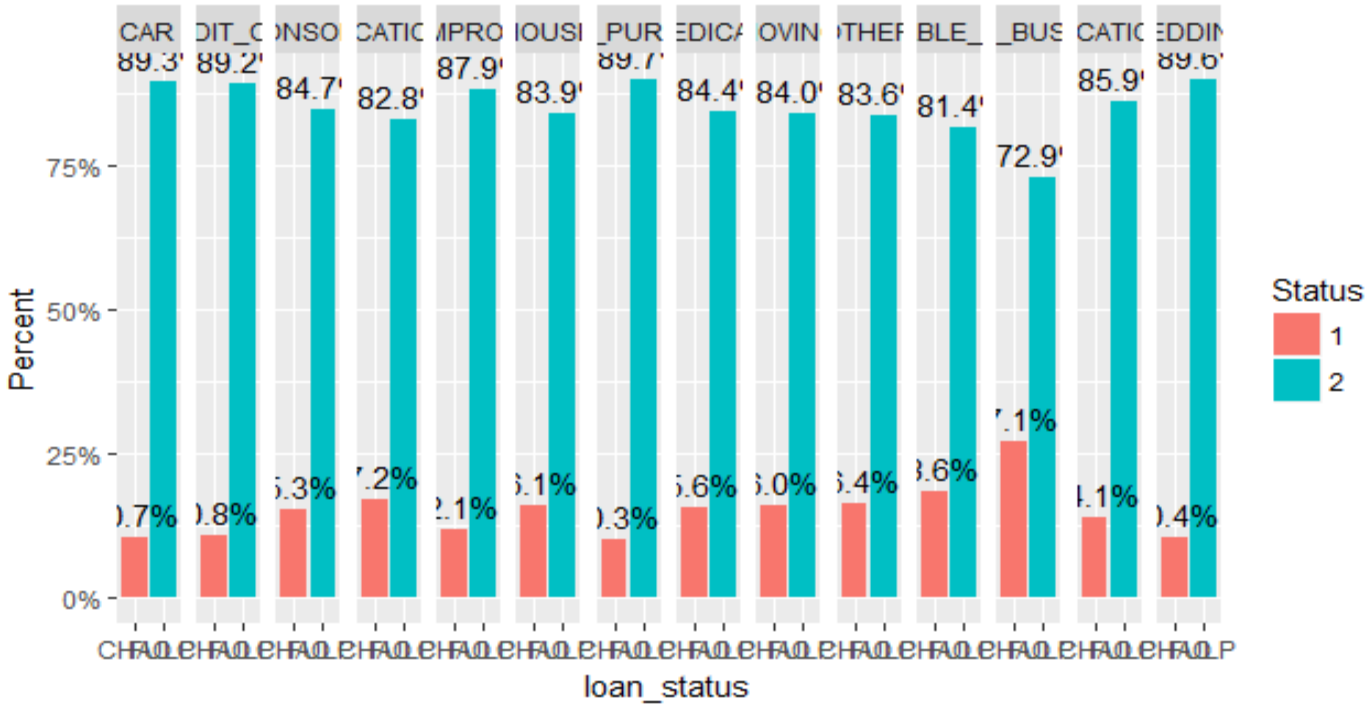


Conclusion: Top 5 Observations



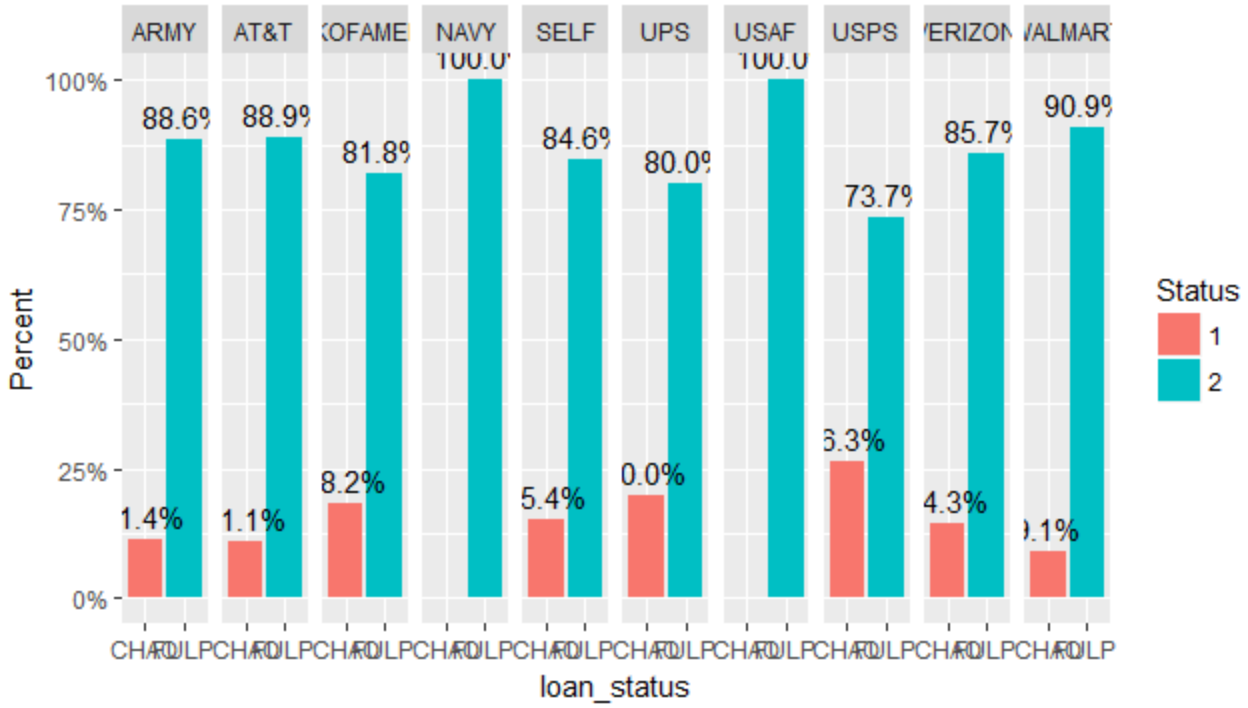
<u>Insight 1</u>: Most defaulting borrowers have “OTHERS” as ownership status.

Conclusion: Top 5 Observations



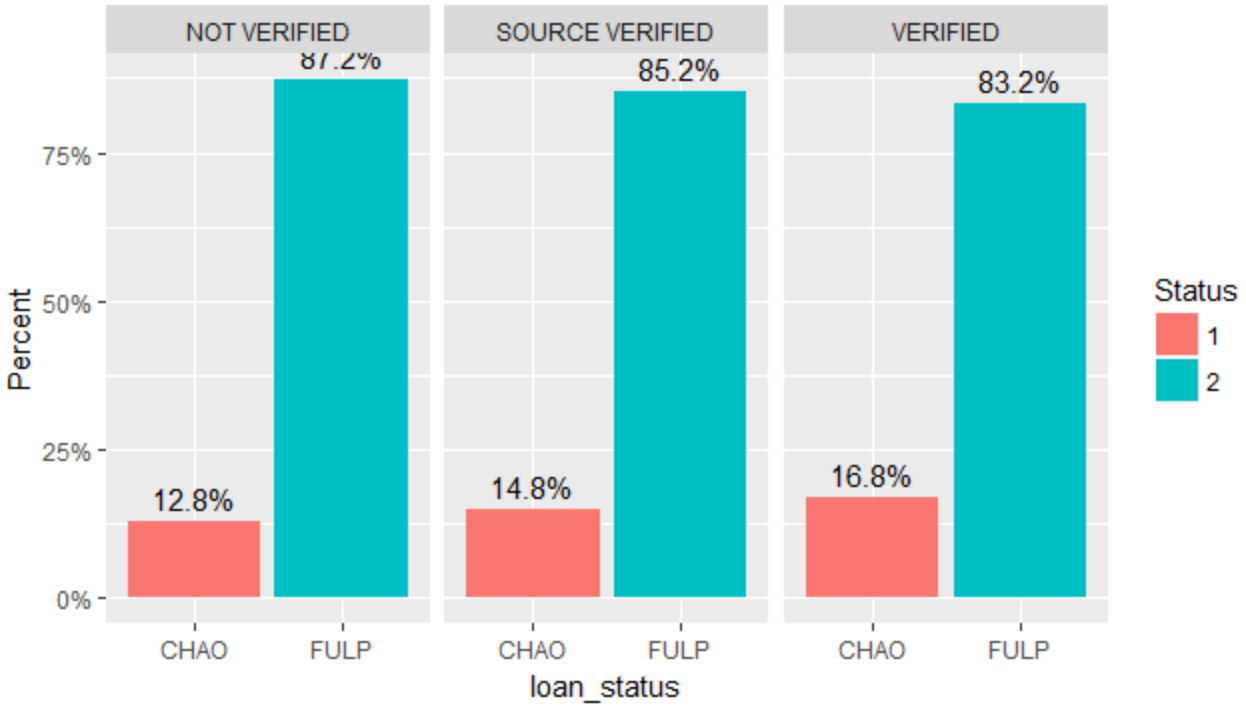
<u>Insight 2</u>: Most defaulting borrows mention purpose as SMALL BUSINESS

Conclusion: Top 5 Observations



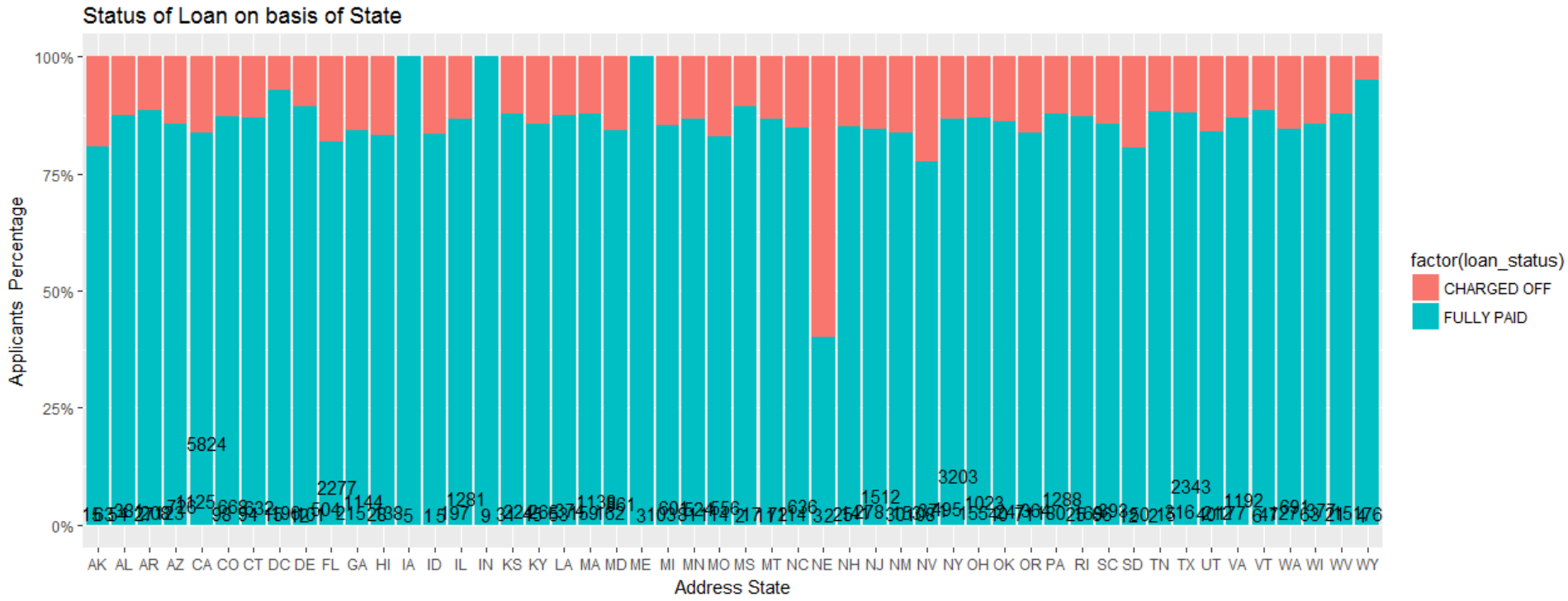
Insight 3: Most defaulting borrowers' are employed with USPS.

Conclusion: Top 5 Observations



Insight 4: Even though, LC has verified borrowers and borrowers sources, the percentage of defaulters for verified is higher than “Not Verified” cases.

Conclusion: Top 5 Observations



<u>Insight 5</u>: Most defaulting borrowers' are from NEVADA.