EDA Case study

Harish Kalyanaraman

Deepak Kumar

Vijay Kumar

Nikhil D Mehta

EDA Case study-Problem statement

This case study aims at analysing the driving factors behind loan default. The dataset has been analysed to identify the risk parameters associated with giving a loan.

The analysis has been done using EDA approach.

EDA-Approach

The approach taken to arrive at insights involve the following steps

- Data exploration
- Data cleaning
- Univariate and Segmented univariate analysis
- Bivariate analysis
- Identifying key insights based on the analysis

Data exploration

There are 111 columns available in the dataset. The main broad categories of variables are

- Variables related to customer demographic
- Variables related to customer information
- Variables related to customer payment behaviour

Since we are looking at insights on the driving factors leading to customer default the third category of variables don’t hold any relevance here.

Data cleaning

- We have looked at all the columns in the dataset and there have been more than 50 columns with NA values.
- We have removed columns where we have found just one value.
- We have removed percentage from interest rate and revolving utilization column for avoiding calculation related issues.
- After further data exploration we have removed other columns like the ones given below which we felt do not add value.
  - URL
  - Member id
  - Employee title

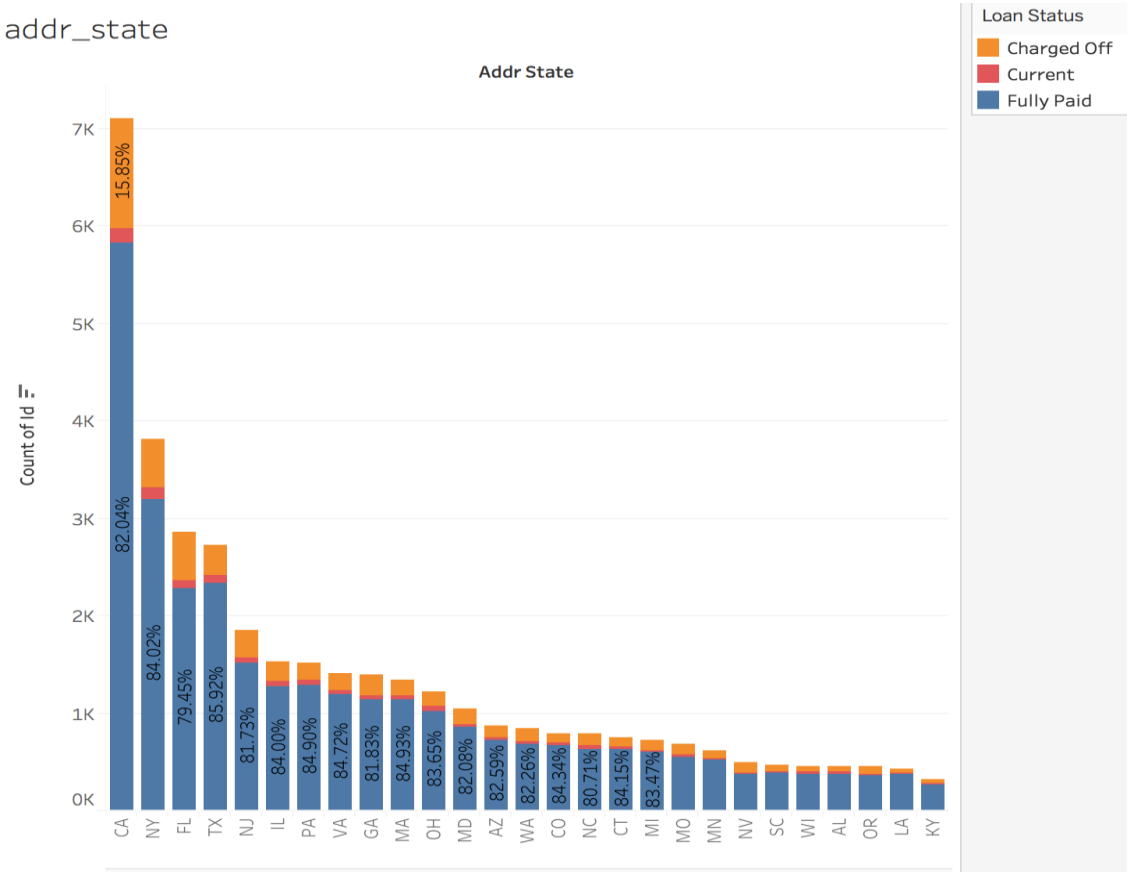
Derived metrics

We have derived the following metrics from the existing dataset.

- Annual income category – The conditions are mentioned below
  - Very low: <30K
  - Moderate: 30k-80k
  - High:>80k
- Percent_installment_monthly_income – Installment/Monthly income * 100

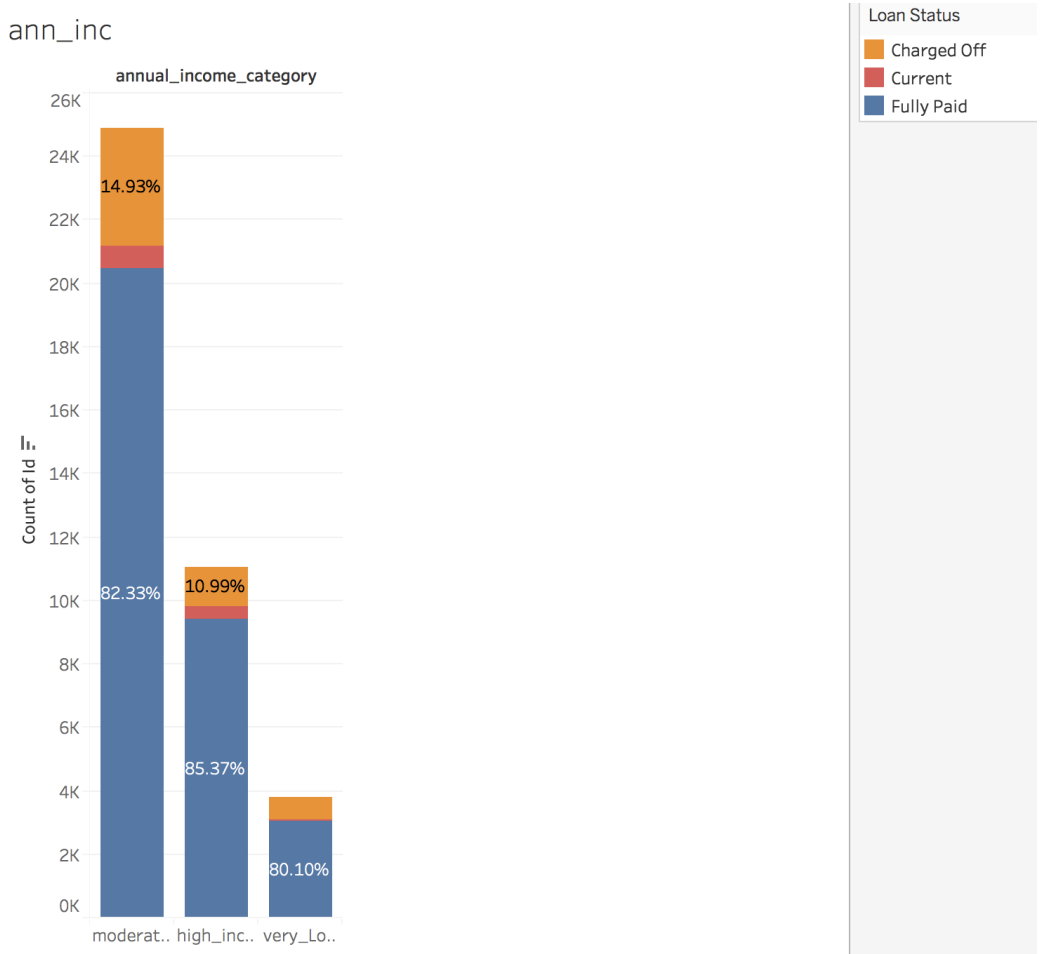
Univariate analysis- Percentage of users defaulting loan in various states

This plot shows that the percentage of loans being charged off is highest in the state of CA



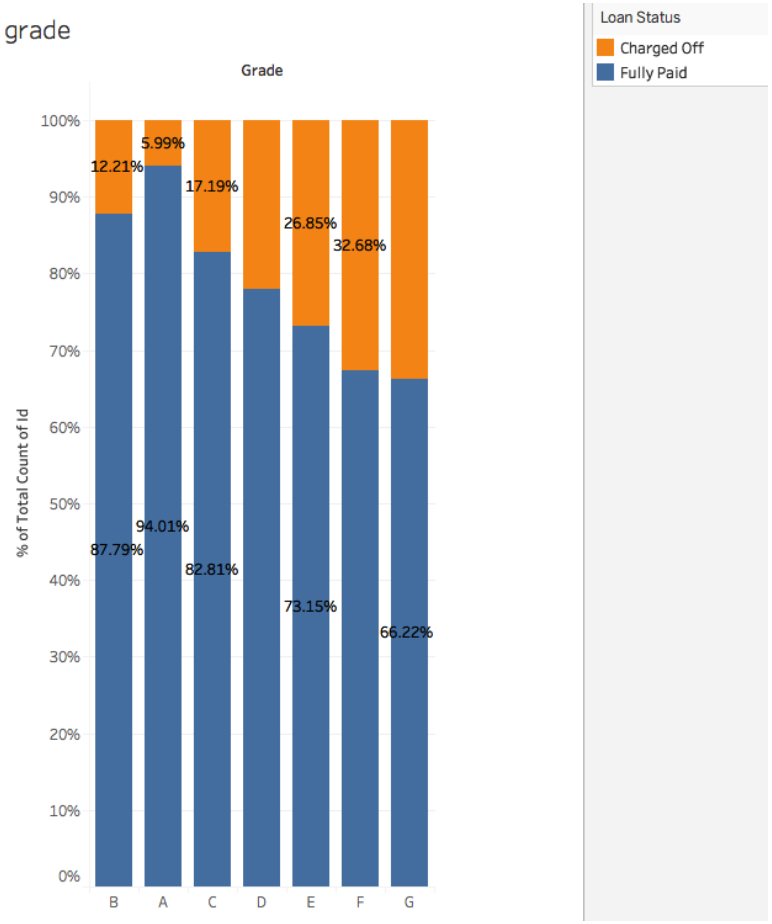
Univariate analysis- Percentage of users defaulting loan based on income category

This plot shows that as the income increases the percentage of loans that are being charged off reduces



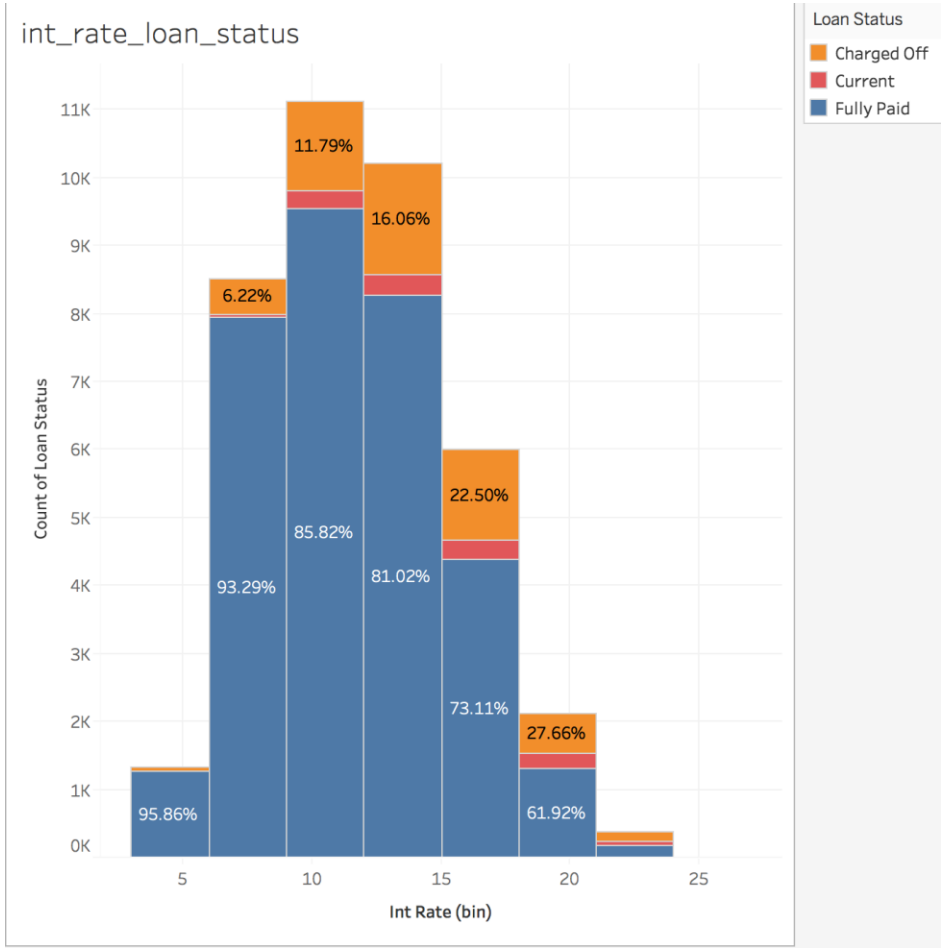
Univariate analysis- Percentage of users defaulting loan based on grade

This plot shows that as the grade becomes lower the percentage of users defaulting loans becomes higher.



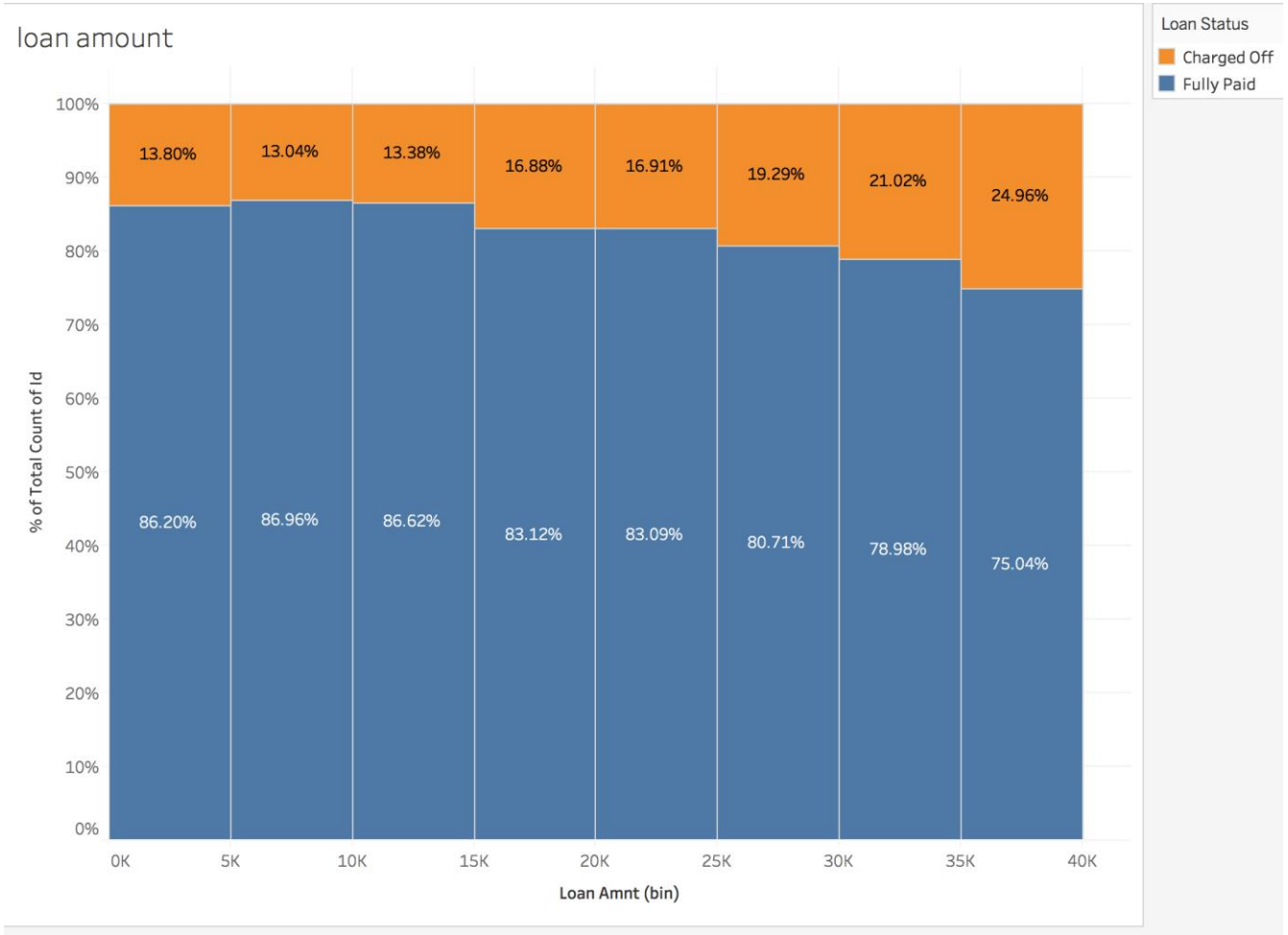
Univariate analysis- Percentage of users defaulting loan based on interest rates

This plot shows that as the interest rate becomes higher the percentage of users defaulting loans becomes higher.



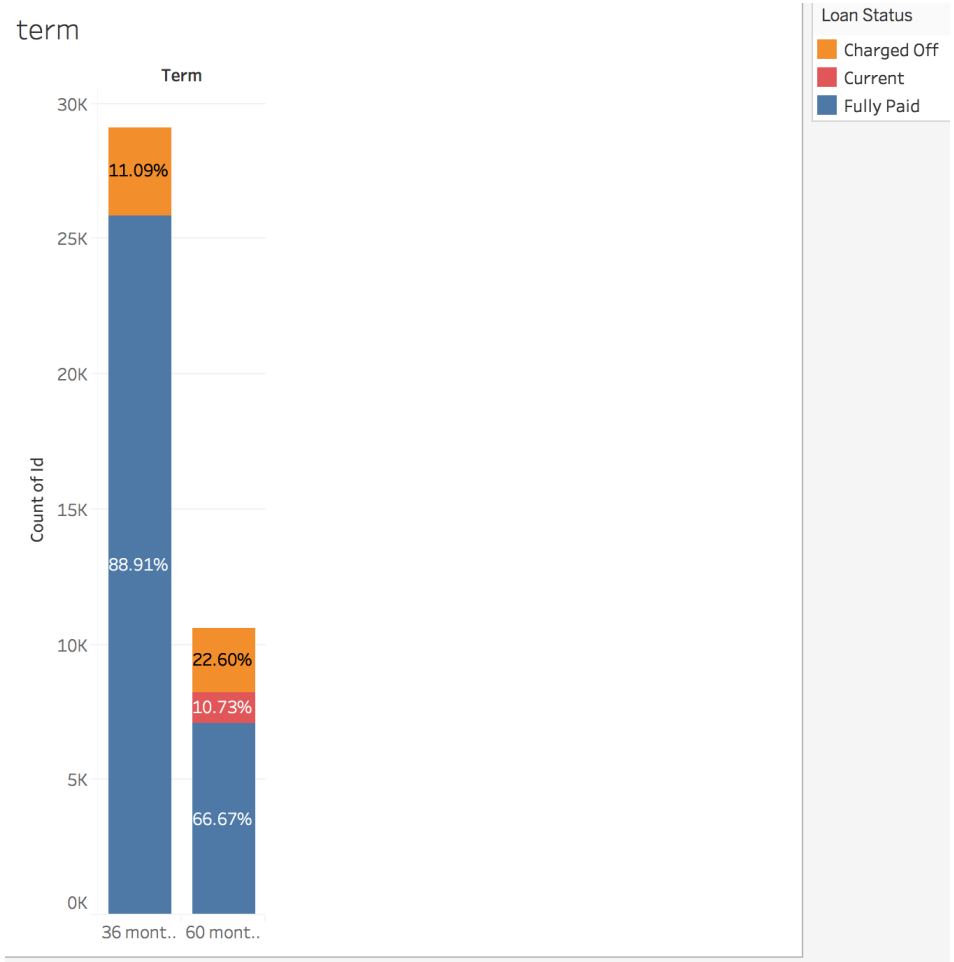
Univariate analysis- Percentage of users defaulting loan based on loan amount

This plot shows that as the loan amount is increasing the percentage of users defaulting also increases.



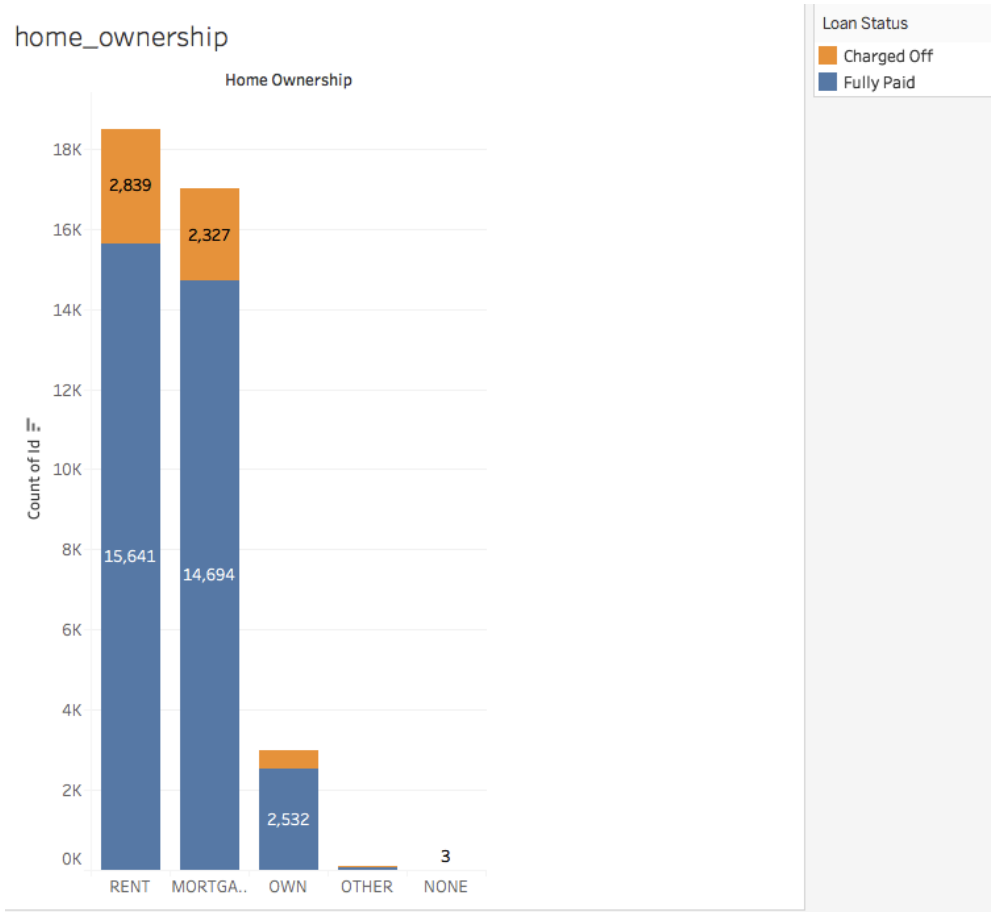
Univariate analysis- Percentage of users defaulting loan based on term

This plot shows that as the term is increasing the percentage of users defaulting also increases.



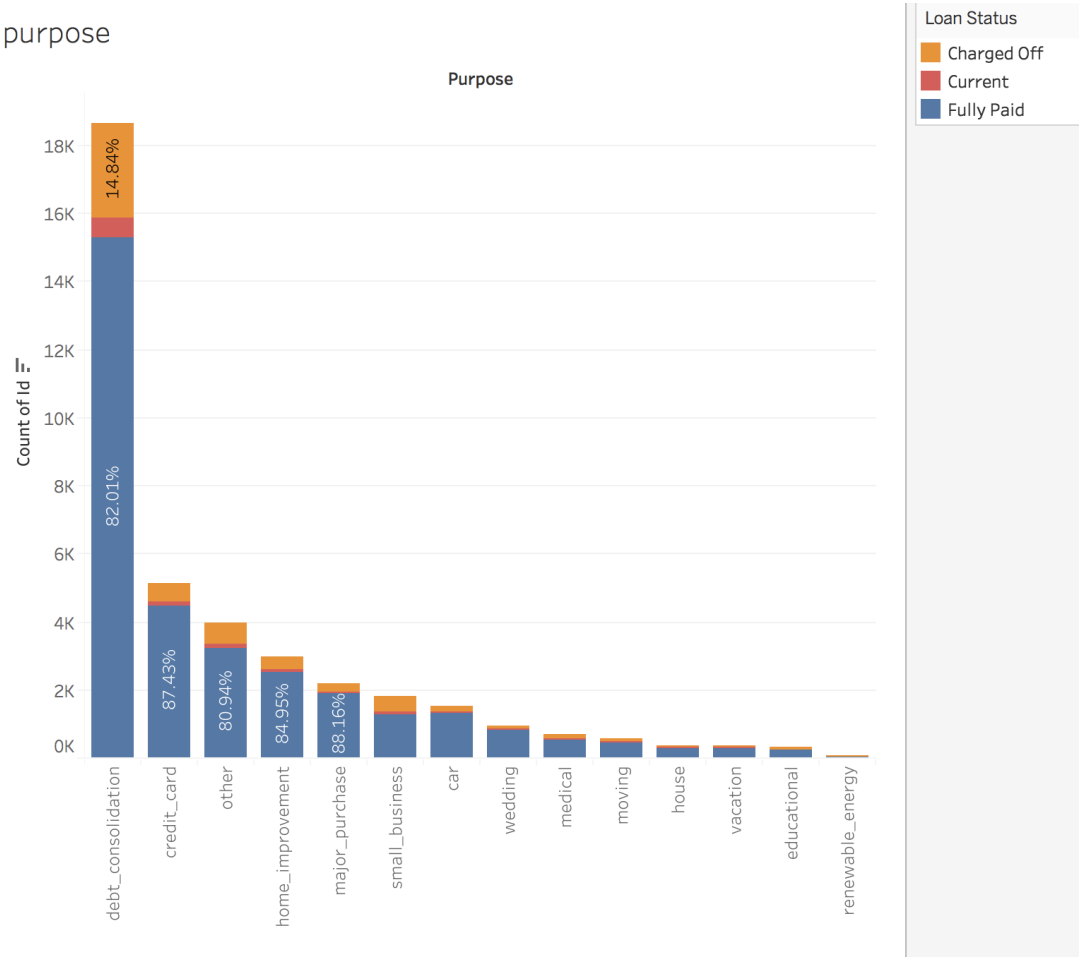
Univariate analysis- Count of users defaulting loan based on home ownership

This plot shows that most of the people taking loan are either having house on mortgage or in rent and the corresponding number of defaulters are maximum in these two categories.



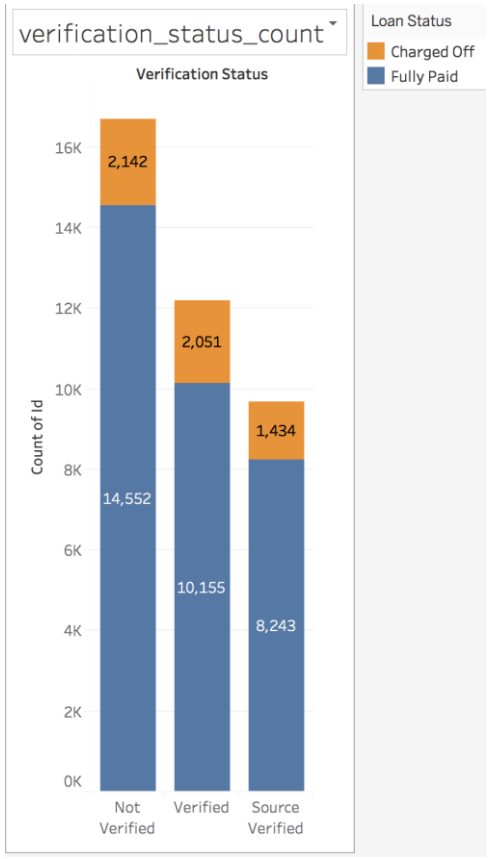
Univariate analysis- Percentage of users defaulting loan based on purpose

This plot shows that maximum loaan was provided for debt consolidation and this where there were maximum defaulters.



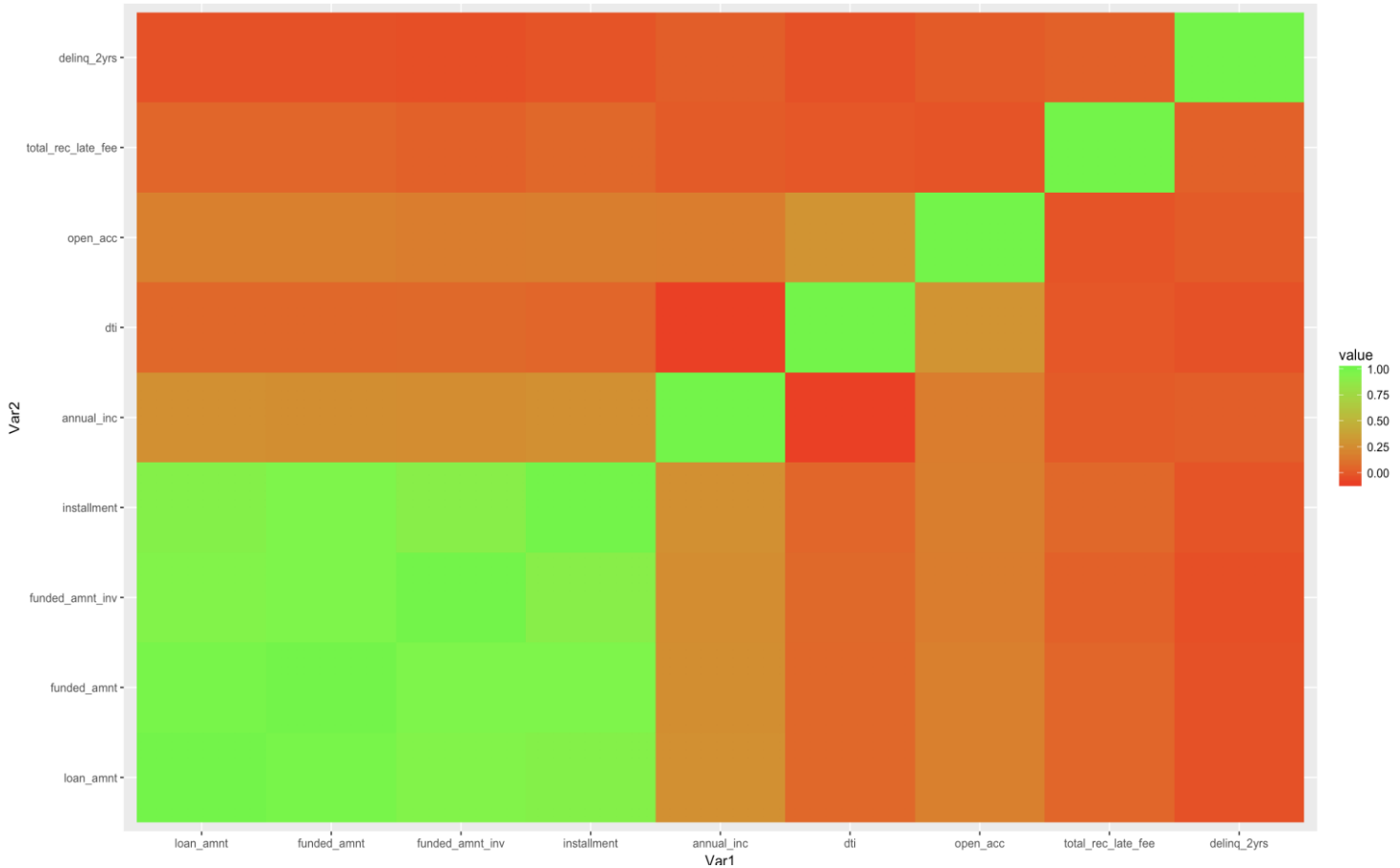
Univariate analysis-Breakdown of users based on verification

This plot shows that there were close to 16000 loans given without verification and there were around 2000 loans who have defaulted among them.



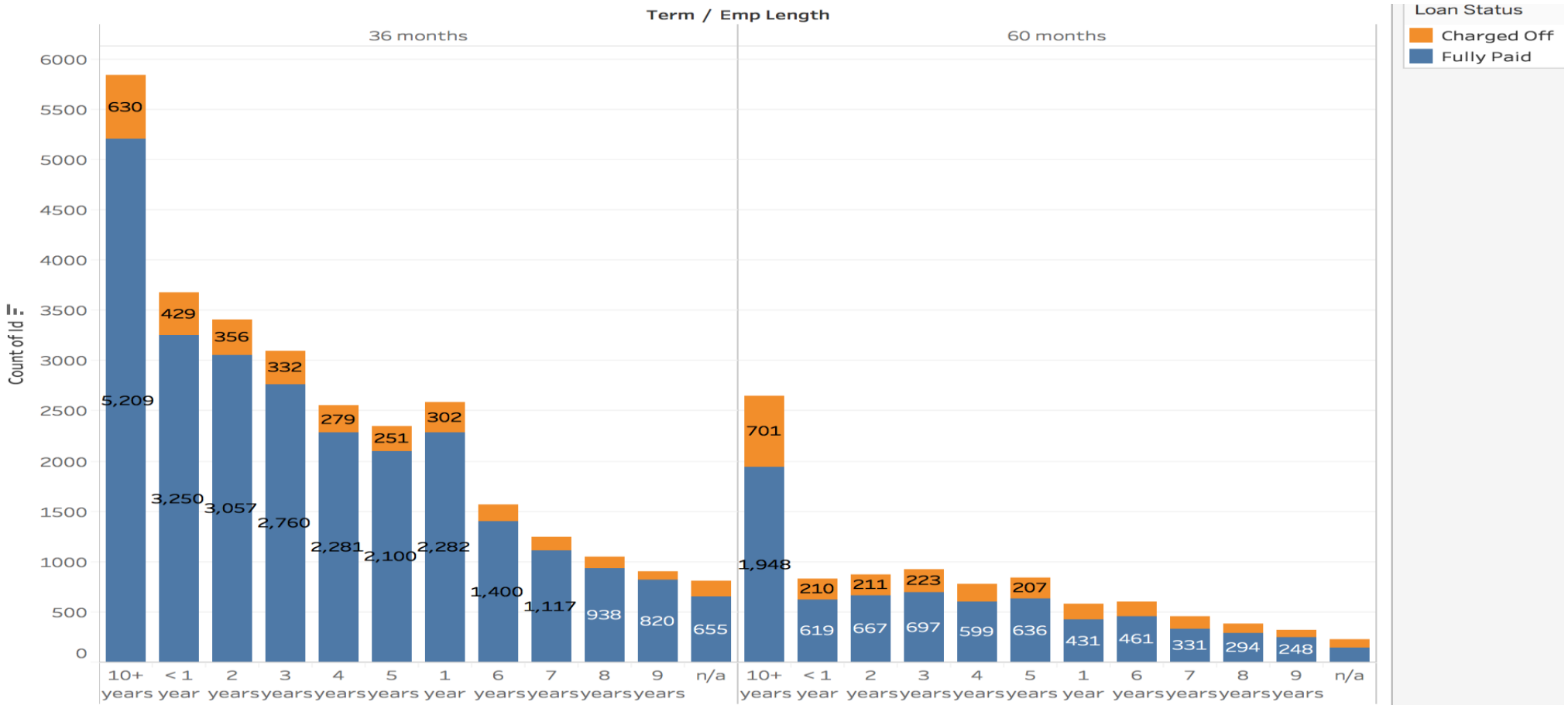
Bivariate analysis-Determining correlation among variables

The heatmap shows a positive correlation between loan_amt,funded_amt and funded_amt_inv.
All the other variables do not show any significant correlation.



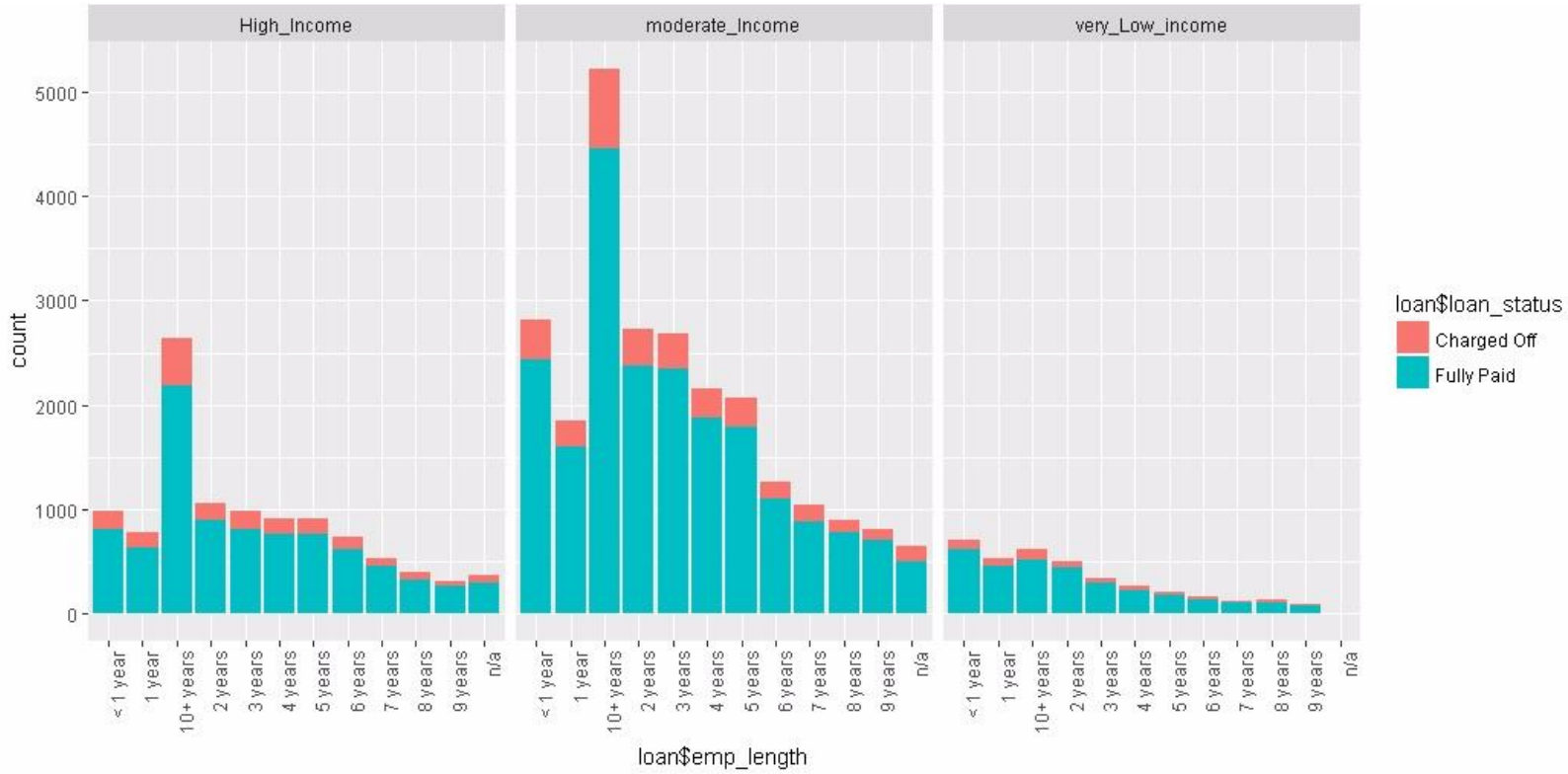
Bivariate analysis-Total experience and Term vs Status of loan

The analysis of term and total experience shows that people with more than 10 years tend to pay fully as well as default on the loans whose proportion is more for 5 year term.



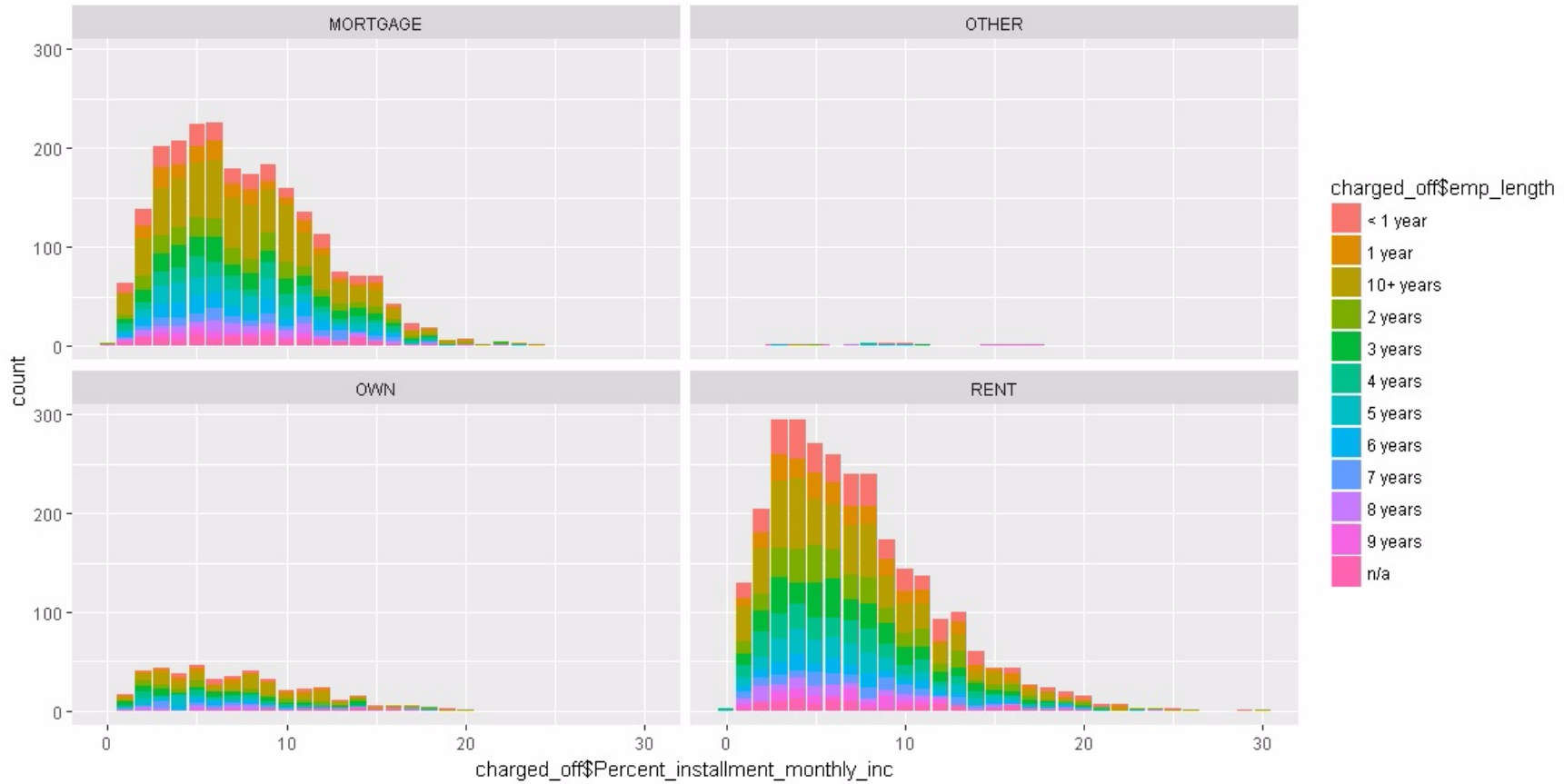
Bivariate analysis-Total experience and Term vs Status of loan

The total employment along with income category shows that people in the moderate income category of between 30-80k take maximum loans and people with 10+ experience have the highest charged off percentage.



Bivariate analysis-Percentage of monthly installment to income for loan defaulters with different home status

The analysis for installment to income ratio shows a number of people having debt ratio of 5-7% as problematic .It decreases after 7% .This does not provide clarity on threshold debt ratio for analysis.



Insights gained from EDA

Here are the recommendations based on the EDA analysis done.

- A large proportion of loans given out to people does not have proper background checks carried out. It should be enforced for all the transactions.
- The maximum defaulters are people earning between 30k-80k.Their repaying capacity has to be checked and based on it the loans need to be granted.
- As the credit grade goes down the possibility of defaulting also goes increases. The people with lower credits also tend to pay higher interest. Since this is a fine balance of making more profit and also losing money due to defaulting these customers need to be verified thoroughly and loans of greater amounts and tenure can be avoided.
- Maximum number of people who are on rent and mortgage tend to take loans and their possibility of defaulting is also the highest.
- The customers getting loans for the purpose of debt consolidation are the maximum among all categories and hence the corresponding number of defaults is also great compared to other categories.

Insights gained from EDA(Contd)

- The people with more than 10 years of experience take the most number of loans and also have the highest rate of defaults.

Based on all these insights the key categories for us to check for loan defaults are the credit grade, purpose of loan, home status and years of employment. These need to be carefully scanned before a decision is arrived at whether loan needs to be granted or not.