Lending Club Loan Analysis

Dhruti Contractor

Dhanilal MS

Data Understanding

- Lending Club (“LC”) is the world’s largest peer-to-peer online lending platform. It reduces the cost of lending and borrowing for individuals with advanced data analytics. The function of peer-to-peer companies is to match people who have money with people who want to borrow money.
- Investors are presented with a list of borrowers, along with their assigned risk assessment grades, and they have the opportunity to choose which borrowers they will fund, and the percentage of funding that they will cover.
- The business problem is to comprehensively assess these borrowers in order to make a smart business decision, by identifying new borrowers that would likely default on their loans

Data Description

- Lending Club provided us with 4 years of historical data (2007-2011). This dataset contained information pertaining to the borrower's past credit history and Lending Club loan information. The total dataset consist of 112 variables for 39717 records to conduct analysis.

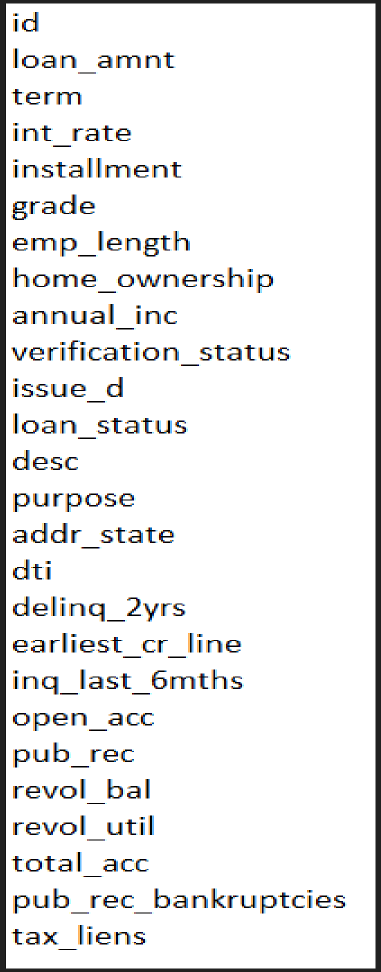
- Variables present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement.

- We only require the variables that had a direct or indirect response to a borrower's potential to default. To achieve this, we prepared the data by choosing select variables that would best fit this criteria.

Data Preparation and Processing

Prior to data mining model analysis, the data was reviewed, cleaned and prepared as follows:

- Out of the 115 variables, we choose only 23 variables including the variable with final loan status result.
- These columns have no value or have high amount of missing values or have constant values.(pymnt plan and application type
- Removed columns that had identical relationships to the analysis in question (E.g. funded_amnt and funded_amnt_inv as they are always the same as loan_amt)
- Leaving variables with lack of its information value. E.g. url, member id
- Replacing the NA values



id
loan_amnt
term
int_rate
installment
grade
emp_length
home_ownership
annual_inc
verification_status
issue_d
loan_status
desc
purpose
addr_state
dti
delinq_2yrs
earliest_cr_line
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
pub_rec_bankruptcies
tax_liens

Data Tranformation

- Transformed continuous variables to range of values to enhance interpretation of results (E.g. int_rate, Annual_income, credit_history_years, revol_util and loan Status)
- The loan status is evidently the most important variable for our purpose. It used to describe what is the current status of a loan. There are three possible loan statuses. We will focus on the records with status ‘Charged-Off’ and ‘Current’ and that leaves us with 6766 records.
- The variable emp length describes how long has been a borrower employedbefore asking for a loan. The values of emp length, such as 1 year, 2 years and 10+ years, make this variable categorical
- The variable annual income is grouped in to different range to better analysis

Exploratory Data Analysis

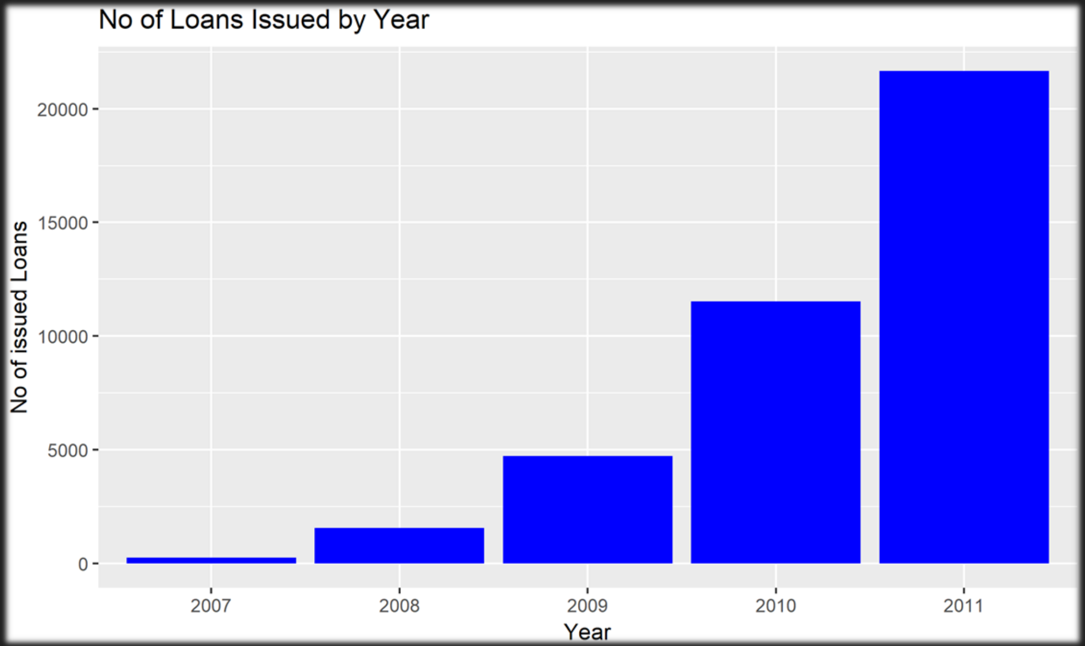
Six trends are investigated:

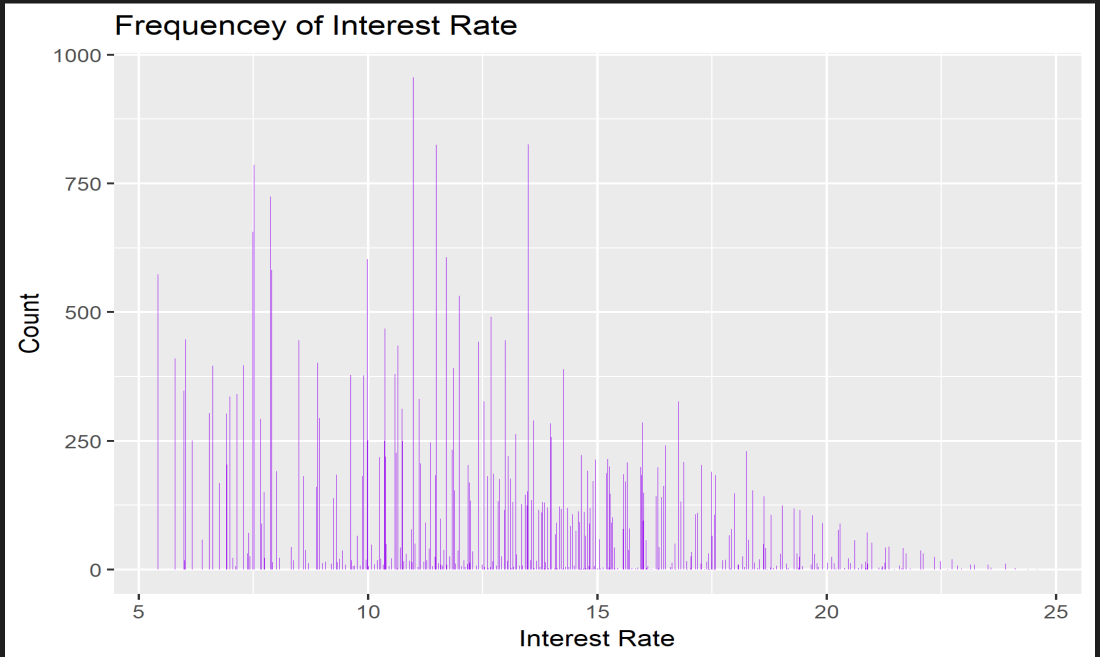
- The growth of issued loans, both in terms of dollars and volume
- The geographical distribution of loans
- The purposes for loans
- The interest rate changes over time
- The interest rate distribution over different grades
- The grades distribution over different home ownership
- The relation between annual income and its verification status

Growth of issued loans

Lending Club was launched in 2007, and its business has grown significantly since the 2009. From then, monthly loan amount and volume have

Rate of interest for majority of the loans approved in betwwen 10% to 15%

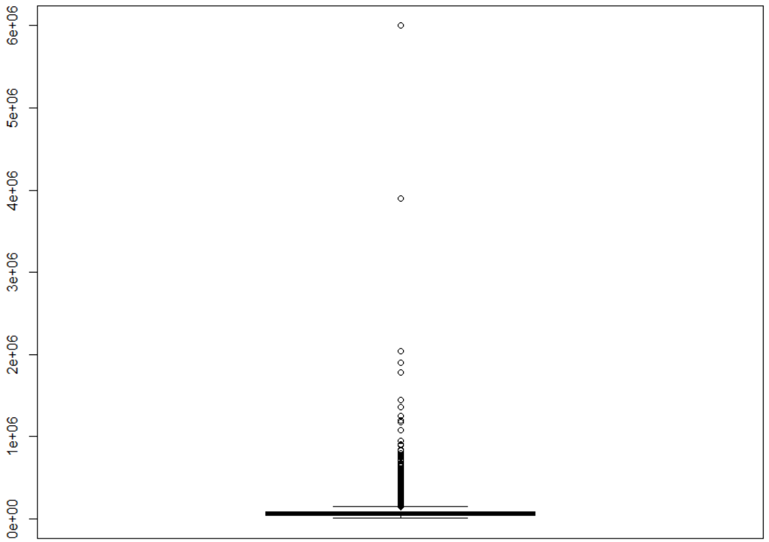




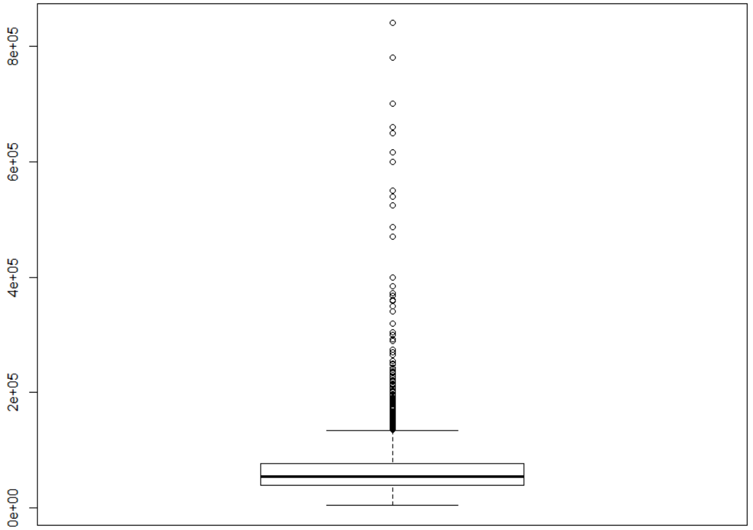
Identifying Outliers

- The data consist of 14 records with annual income more than $1M
- These records will skew the dataset and hence will eb treated ad outliers
- The boxplots displays the effect of outliers

With Outliers

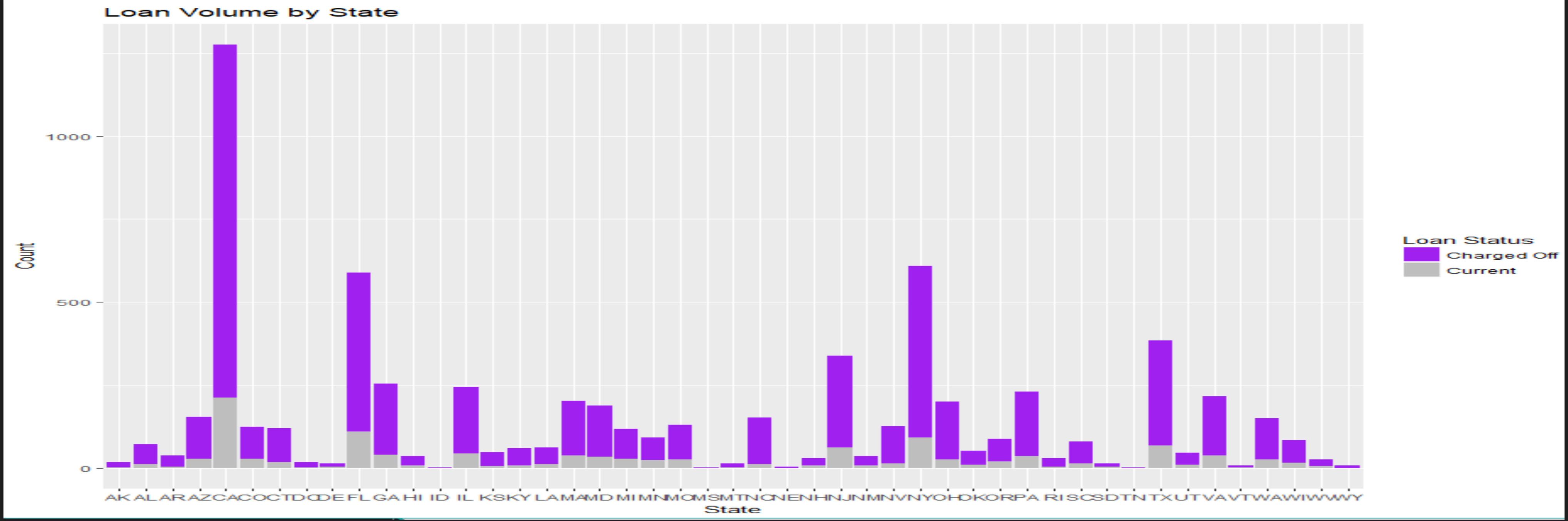


Excluding Outliers



Geographical Distribution

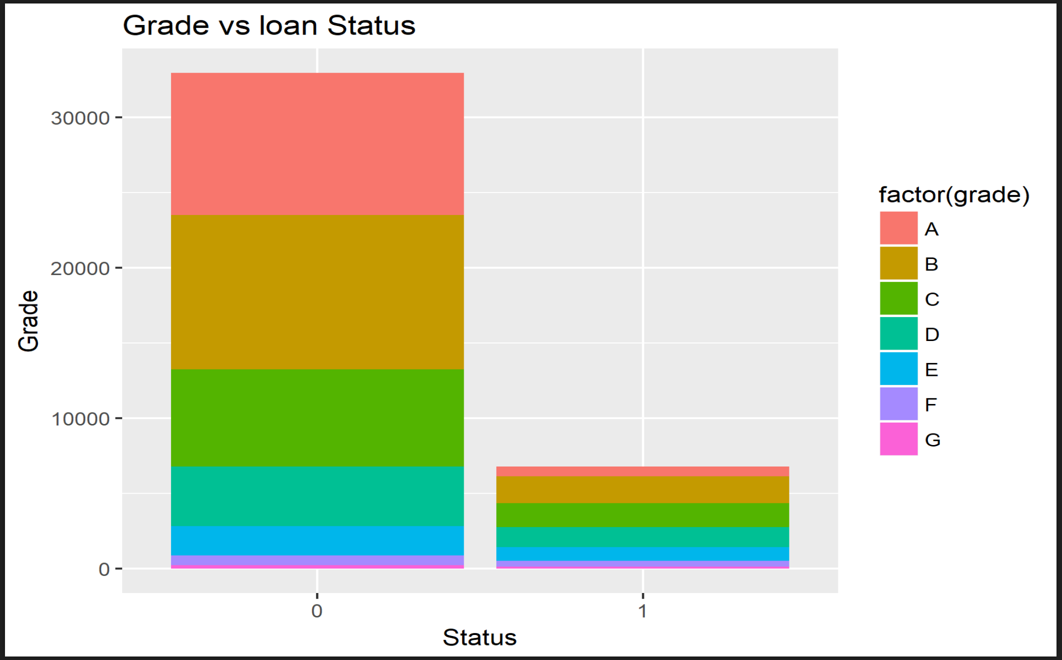
From a geographical perspective, California, Texas, New York, New jersey,Florida, and Illinois have the largest dollar amounts and volumes of loans. California is the location of Lending Club’s headquarters, so it is reasonable this state has more business and more number od defaulters as well. As for Texas, New York,New Jersey Florida, and Illinois, their high volume and amount of loans may be related to Lending Club’s promotion activities.

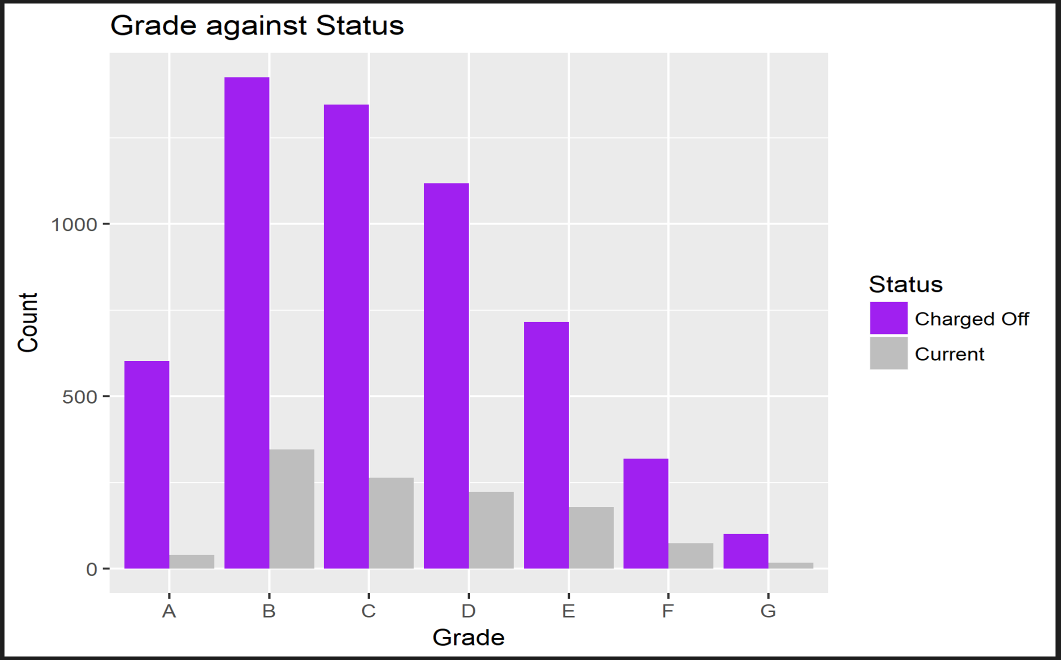


Grade and Default

The proportion of low-grade loans such as grades A, B, C,and D become significantly larger and larger when the loan status moves from the best status, ‘current & Fully paid’, to the worst status, ‘Charged off & Default’.

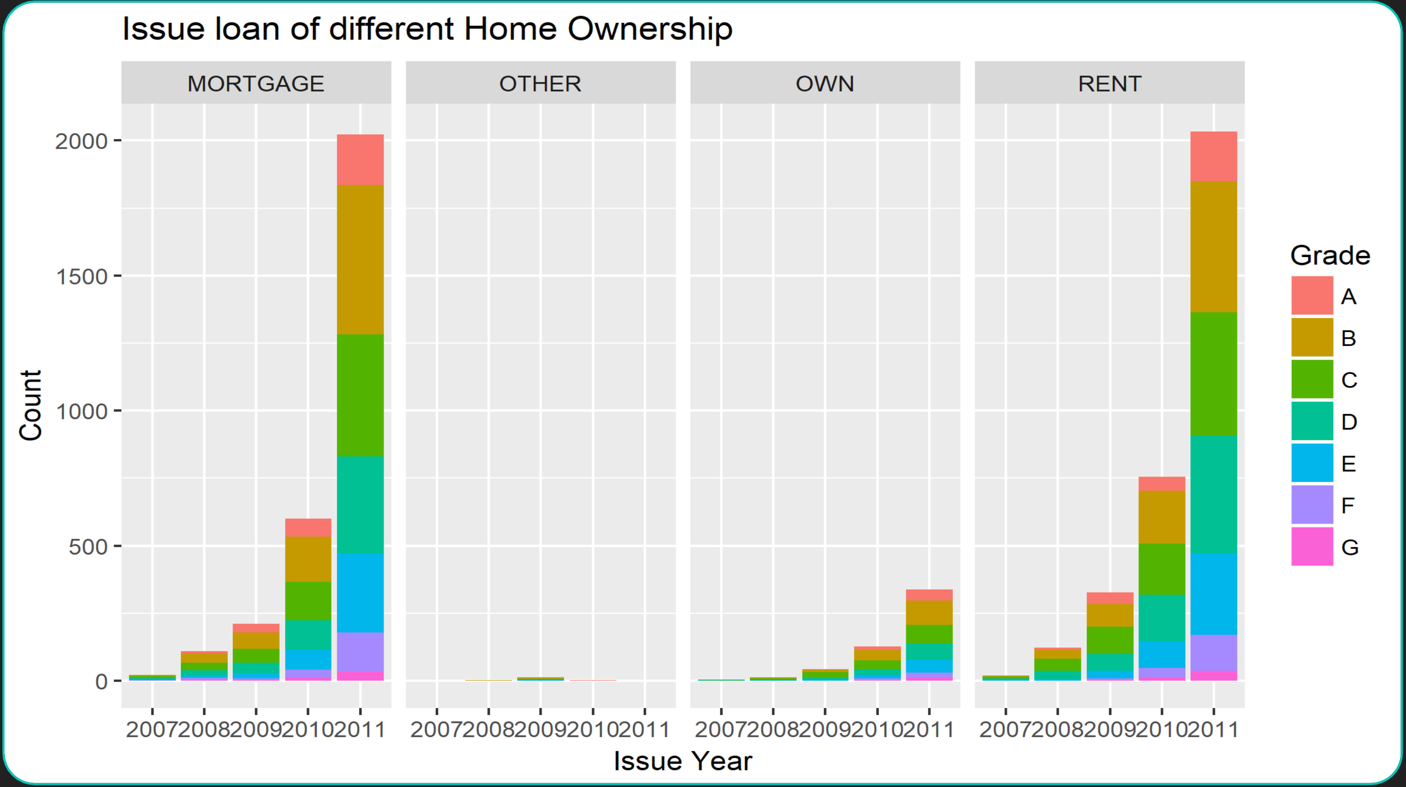
Based on below graphs, the long tail in right side of the distribution is shorter and shorter, and the top two grades change from B, C to C, D. This is not an exact examination of the efficiency of grades, but it still provides information about risk of loans in different grades.





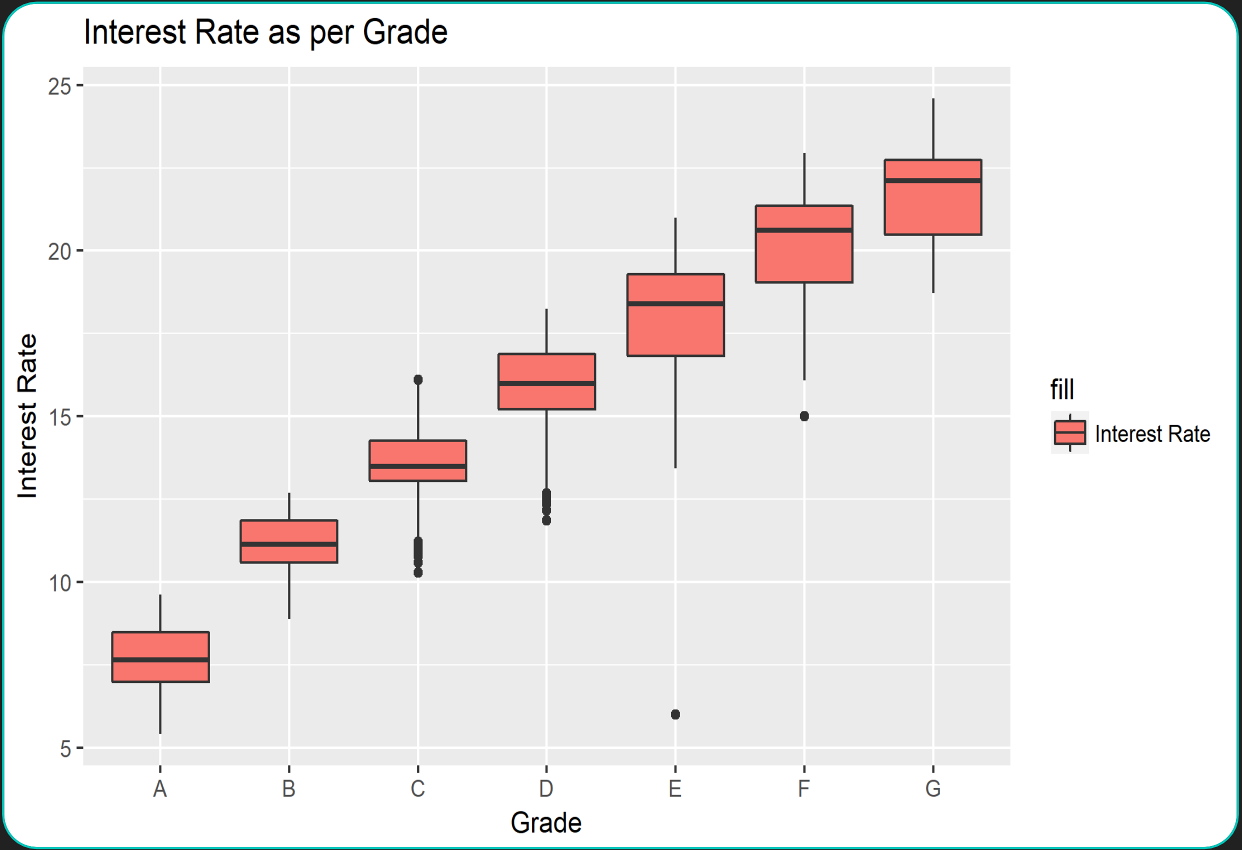
Grade and Home Ownership

- People in 'MORTGAGE' and 'RENT' have much more demands of borrowing money than people in 'OWN' based on the bar chart. That's because people who own a house usually have better financial situation than others.
- Also, there is a substantial increase in the not of loans from 2009 to 2011.
- Defaulters seem to have loans with grade B C and D more than the remaining


Issue loan of different Home Ownership

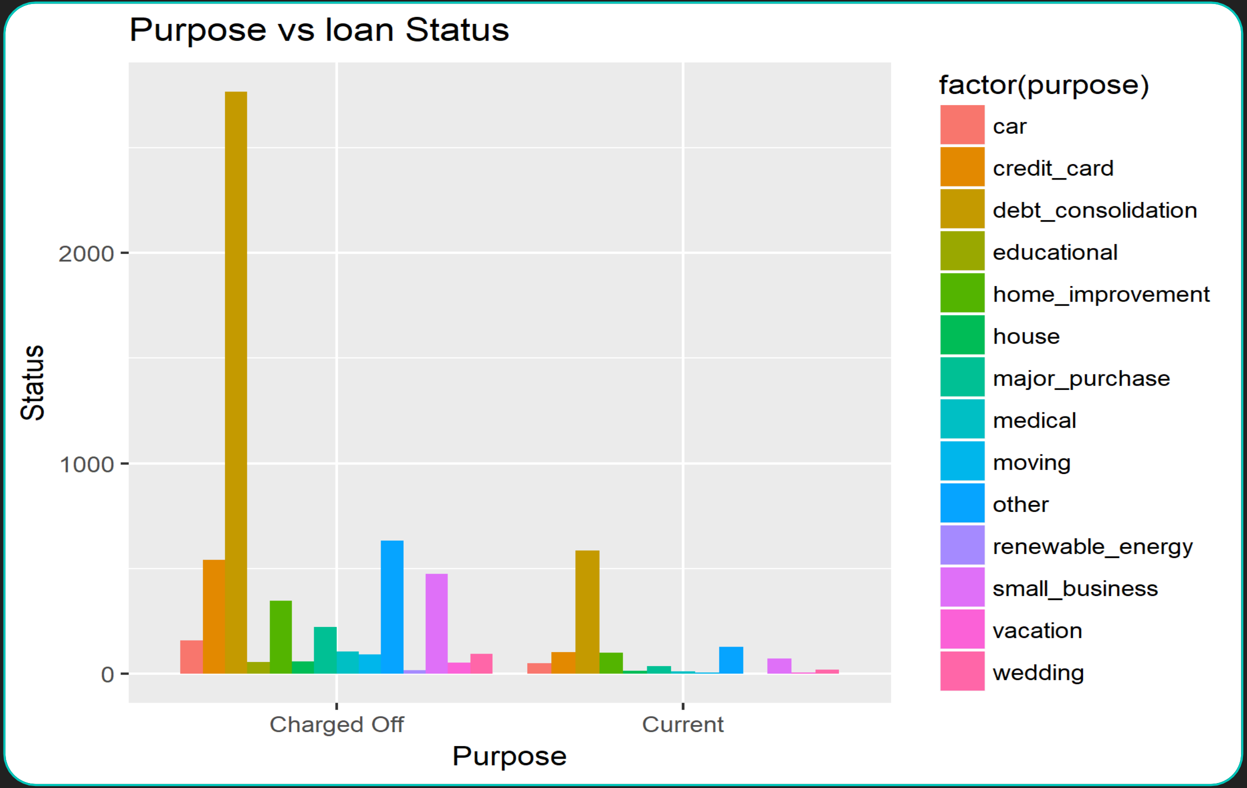
Grade and Interest Rate

- Interest rates of loans with different grades behave in a significant trend through years. The interest rates have been increasing for low-grade loans such as grades D, E, F, G.
- Also, we see that disparities of interest rates become larger and larger, and finally interest rate intervals among different grades are almost equal.
- This change is significant proof that LC has become more and more proficient in the evaluation of loans’ risk and debt management.



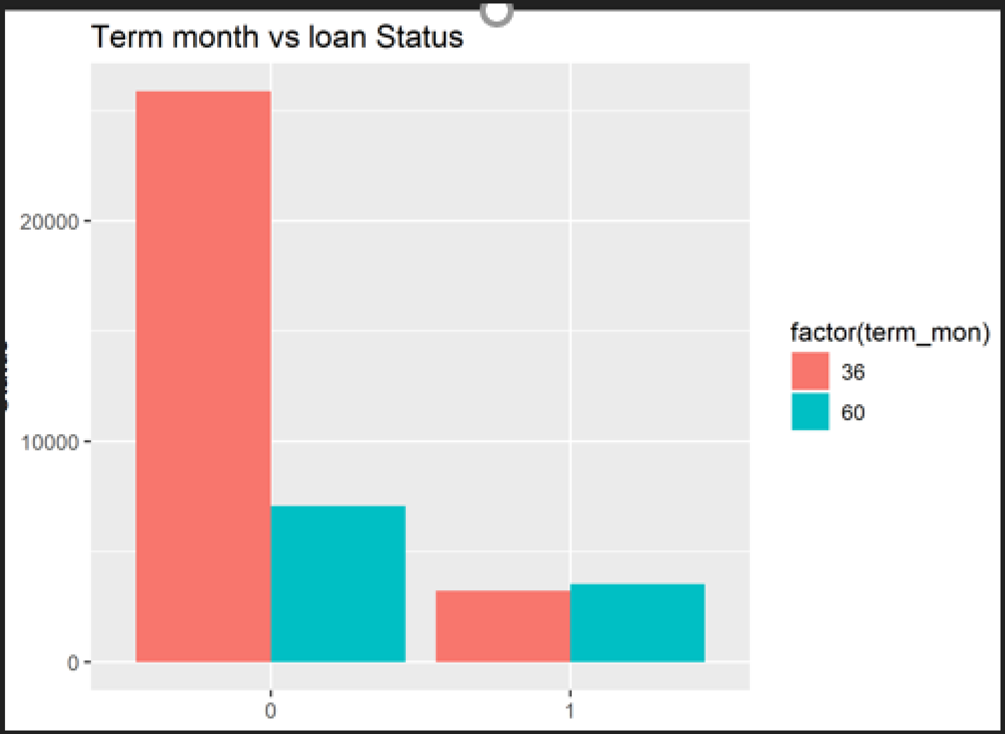
Purpose of loans

- Debt consolidation is the most common reason for borrowing. The greatest advantage of peer-to-peer lending is the low cost. Loans issued by LC usually charge lower interest rates compared with money provided by traditional banks. Most consumers choose to consolidate debt to enjoy lower borrowing costs.
- Debt consolidation and credit card are the most popular reasons for borrowing.
- Loans for credit card,house and small business usually have lower average amount than other purposes.

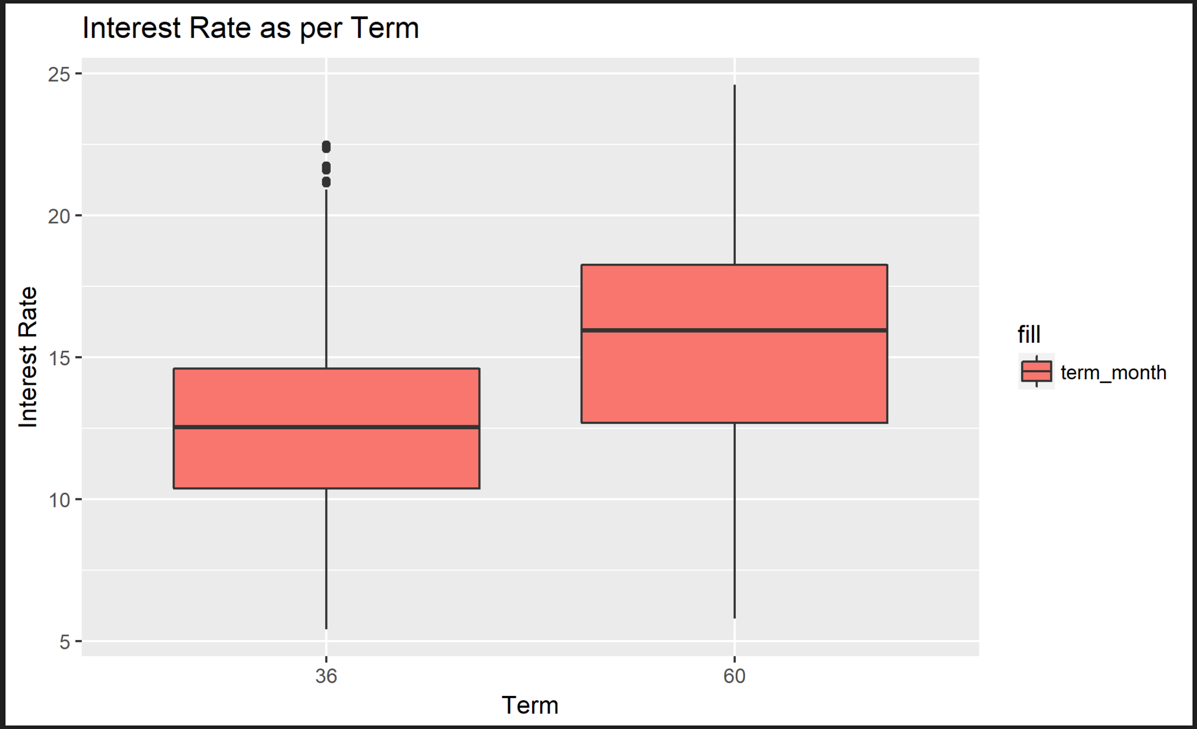


Interest rate and Loan length

Defaulters are significantly more for term of 36 months than 60 months

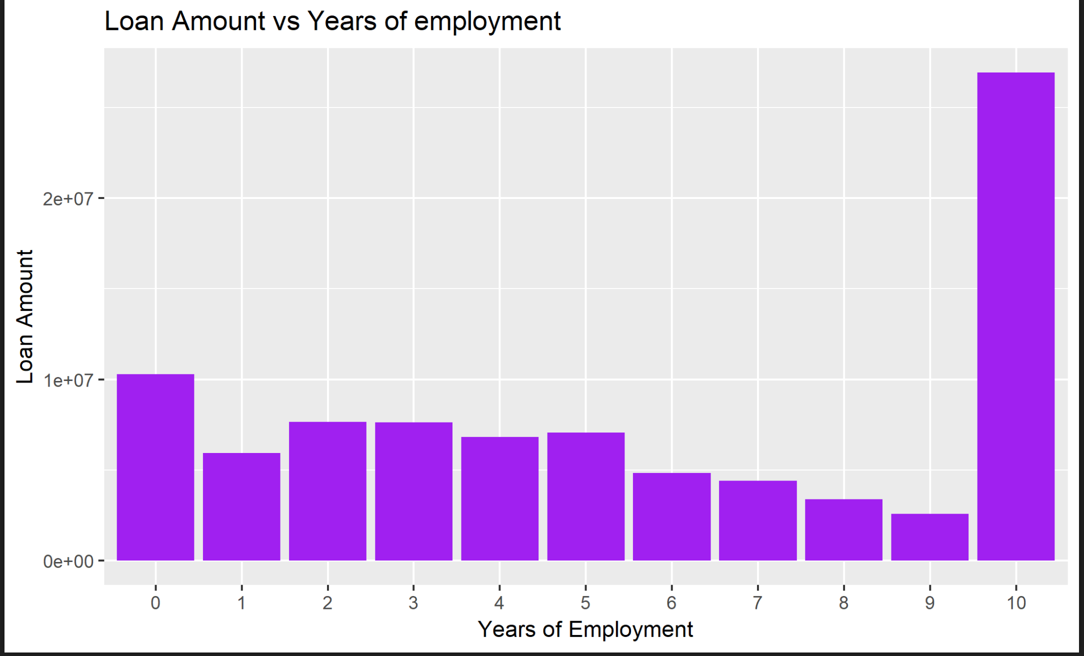


Interest rates increases with the term

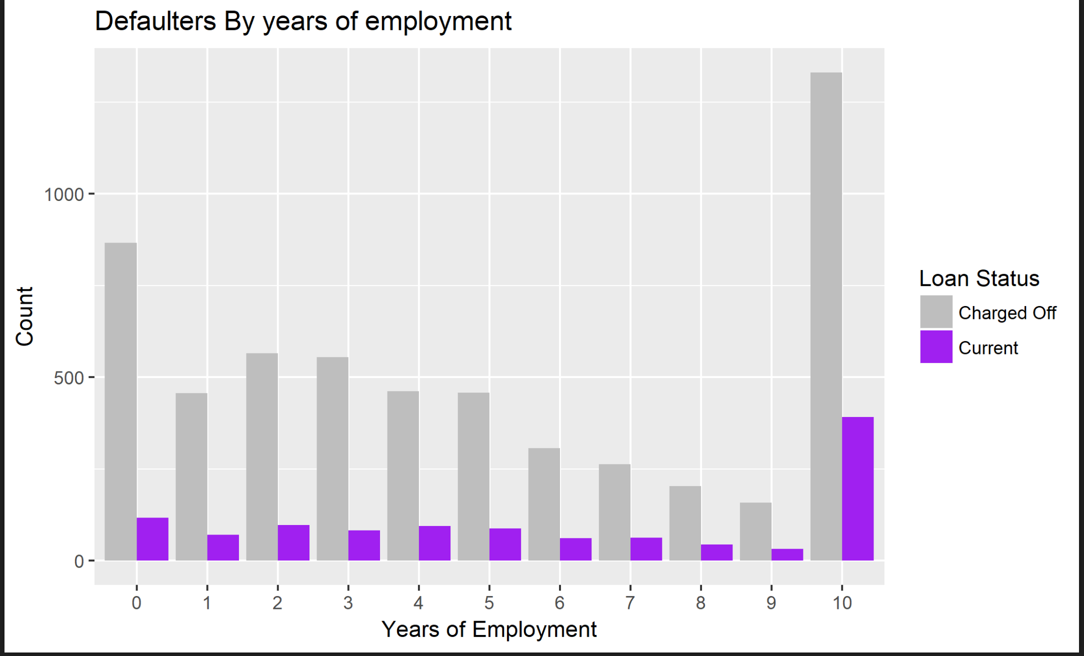


Years of Employment

As seen below, people with 10+ years of employment have loans with more amount

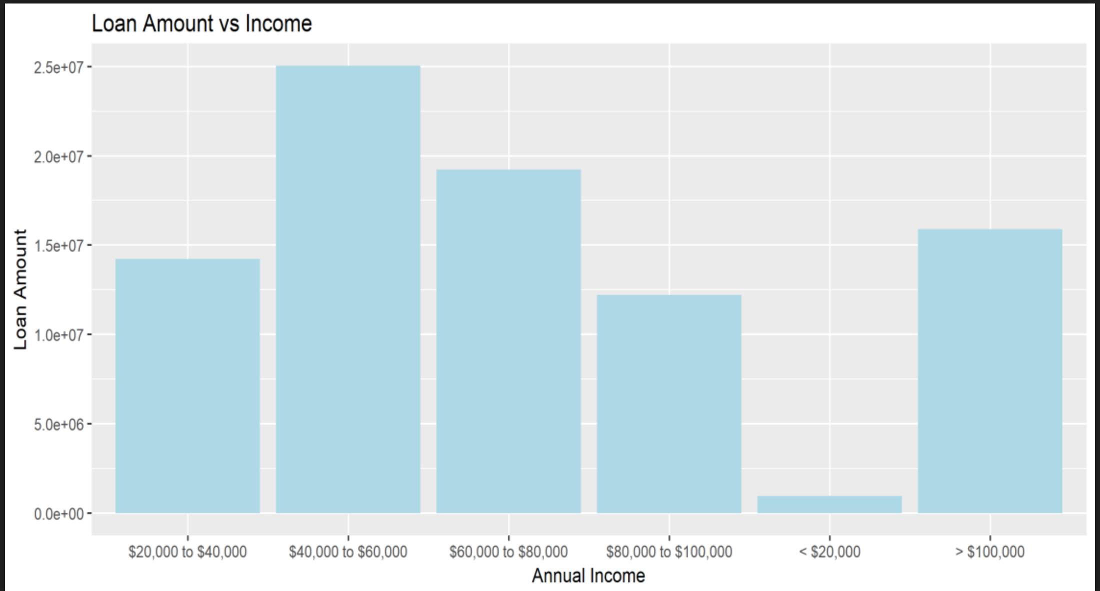

Loan Amount vs Years of employment

Also, majority of the defaulters are with 10+ years of employment followed by less than 1 year

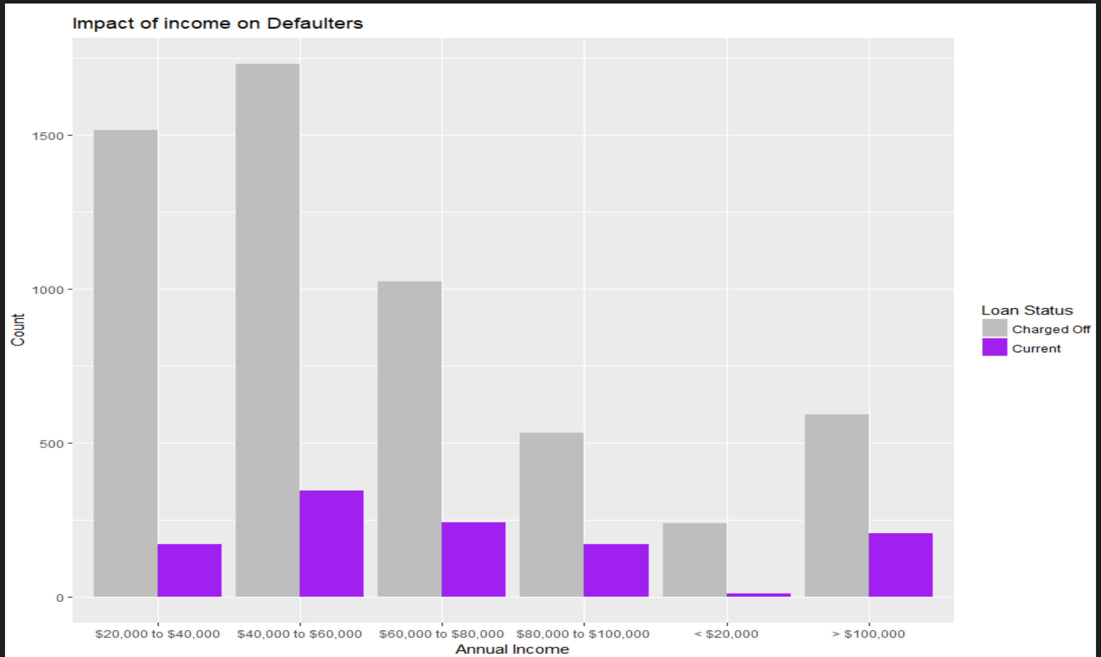

Defaulters By years of employment

Income and Defaulters

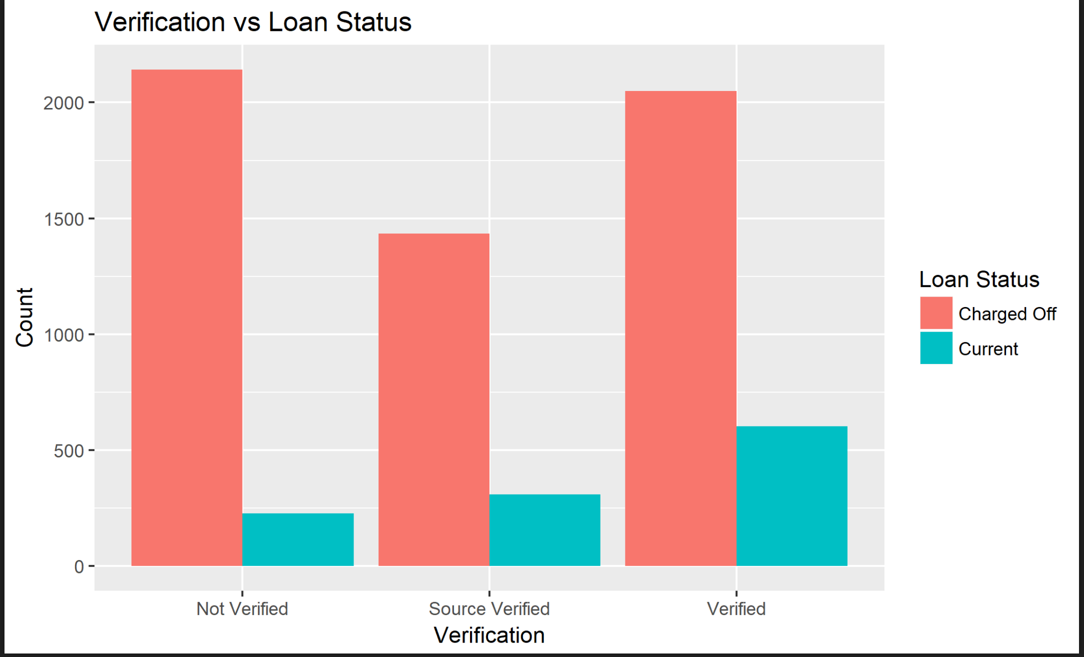
People with income between $40,000 & $60,000 tend to take more loans and then followed by $60,000 & $80,000

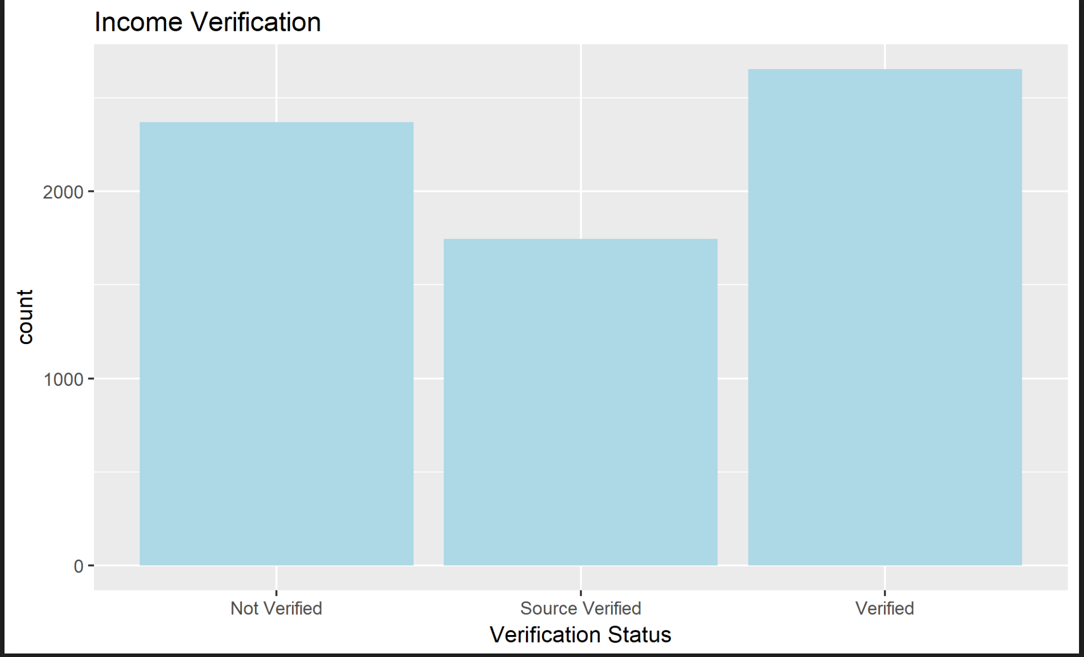


There are more number of defaulters with income between $40,000 & $60,000 followed by $20,000& $40,000



Verification Status





Recommendations

- The interest rate which we receive depends on the various factors like Homeownership, Purpose of loan, Team length of loan, loan amount requested, Annual income, Employee length, Issue month, Previous bankrupcies and Debt to income ratio. If a person is wanting to get a good interest rate then he need to focus on above factors before applying for a lending club loan.
- The project uses visualization to analyze LendingClub's loan applicants and extends to an application for future loss estimation. I find that the trait of applicants usually exhibit quite different default probabilities, especially the increase in rate of interest ith lower gardes. In addition, average interest rates differs quite a lot across states and time, and serve as a good indicator of the application pool of the borrowers.
- Lastly, the expected loss for the outstanding loans at time being is relatively much higher in California, Texas, New York, and Florida, that more resources should be allotted to loan recollection and screening for new applications in these states.
- Make sure to verify the income before approval