Lending Club Case Study

BY ABHINISH AND PAVAN

Business Scenario

- We are working for a consumer finance company which specializes in lending various types of loans to urban customers.
- When the company receives a loan application, the company has to make a decision for a loan approval based on the applicant's profile.
- The main objective here is to help the company determine the driving factors that indicate whether a borrower is going to default on a loan or pay it off successfully
- Based on the inferences made, the company can then make data driven decisions regarding whether to approve a prospective borrower’s loan application or reject it.
- Through this, the company will be able to reduce the losses made due to defaulting on a loan and increase business by seeing that the loans get paid off.

The Approach taken to solve the problem

The following steps were taken to solve the task at hand :-

1. Data Cleaning:
    a. In this step, we will be removing all attributes that have most of its values as null,
    b. We will also remove those attributes that have no significance in the data analysis as well as those variables that do not provide any information gain.
2. Univariate Analysis:
    a. In this step, we will be finding insights on individual variables through visualizations.
3. Bivariate Analysis:
    a. With this step, we will explore how one variable can influence other variables and gain further insights from their behaviour.

Through these steps we will be able to find the variables that have a driving influence that determines whether an applicant should be considered for a loan or not.

Data Cleaning - Removal of Null Values

1. When we take a look at the dataset, we see that many of the columns have more than 50% of its values are null.
2. The first step is the remove all these null values as they don't have any data to provide and will not help in the analysis in any way.
3. We had 111 columns representing different attributes and on removing them we have only 54 columns remaining.
4. This meant that 57 columns had more than 50% of its values as null.

Data Cleaning - Converting Continuous variables in String format

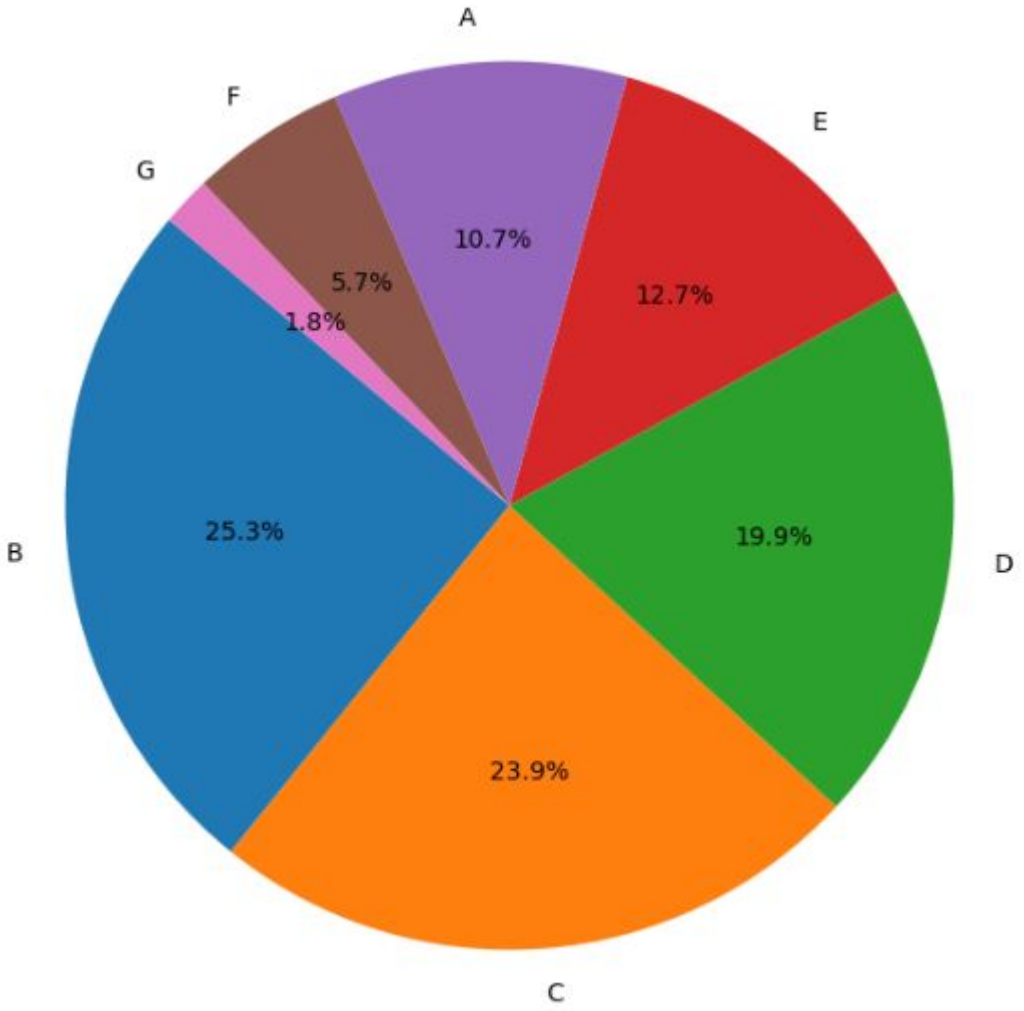
1. The next thing that is to be done is to convert a couple of attributes from a formatted string into continuous values.
2. These variables are the Rate of Interest which is represented by ‘int_rate’ and Revolving Line Utilization Rate which is represented by ‘revol_util’.
3. We proceed with their conversion by
    a. Removing the percentage sign that is at the end of the string
    b. And then converting the truncated string to floating value that contains precision upto two digits after the decimal point
4. Some of the values in the ‘revol_util’ column had null values.
    a. I imputed the values for this column by using the value 0 which is the most frequent number i.e. mode imputation method

Data Cleaning - Remaining Steps

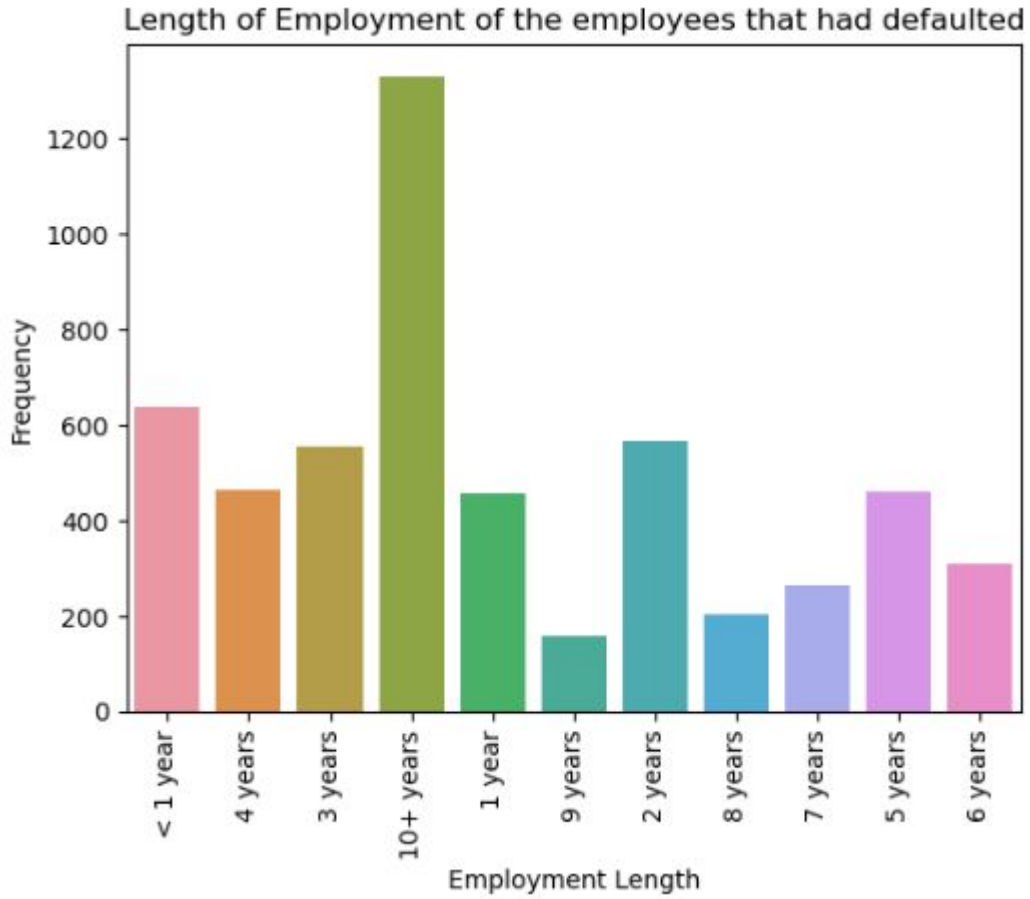
1. We will now remove all of the variables that have only a single value for all the entries.
2. There is no point in keeping them as they provide no information gain, after all, irrespective of whether a particular applicant is a defaulter, or a non-defaulter these variables will have a constant value which brings in no influence, so it's best to remove them.
3. We will also be removing all the rows that have the 'loan_status' value as 'Current' as they will play no role in the analysis. We are looking at either defaulters or those that have paid off the loans
4. We are going to remove 'id' and 'member_id' as they have no role in the analysis that is to be made. So there is no sense in keeping them.

LC Assigned Grades for Loans That Defaulted

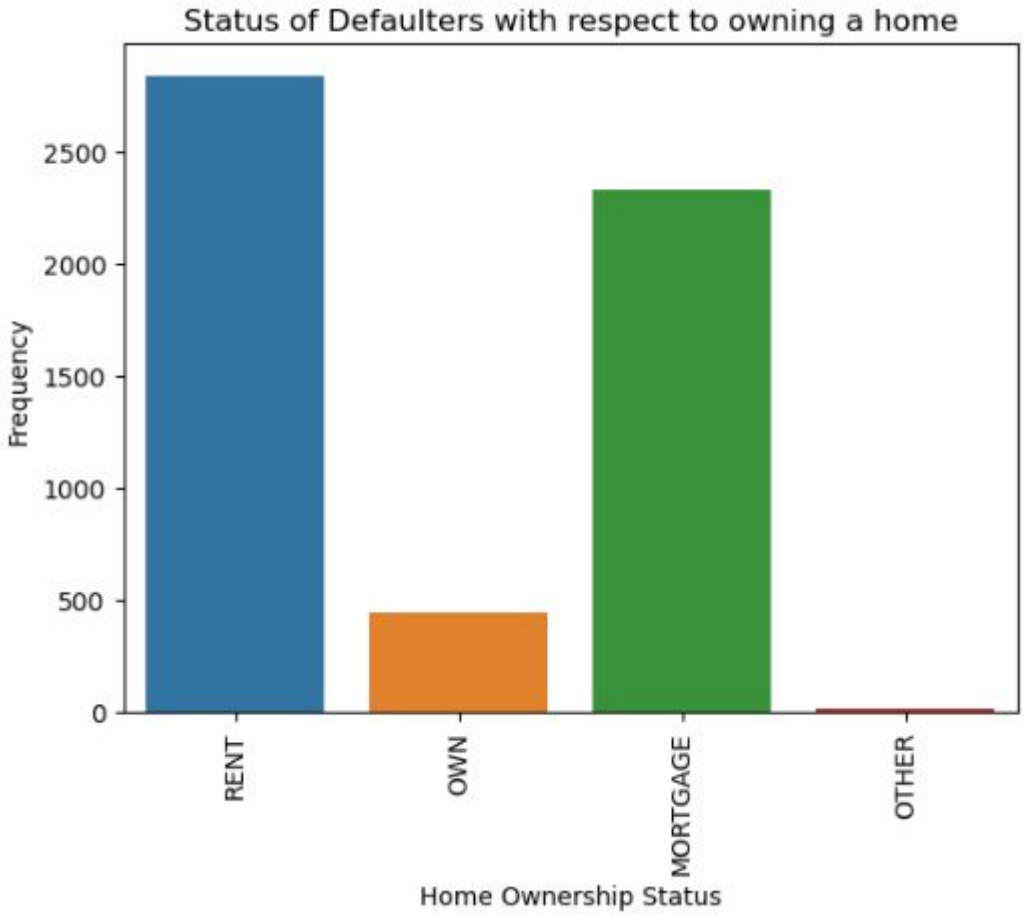
Most of the loans that defaulted were either of type 'B', 'C' or 'D' with these three grades making up 69% of the loans that defaulted



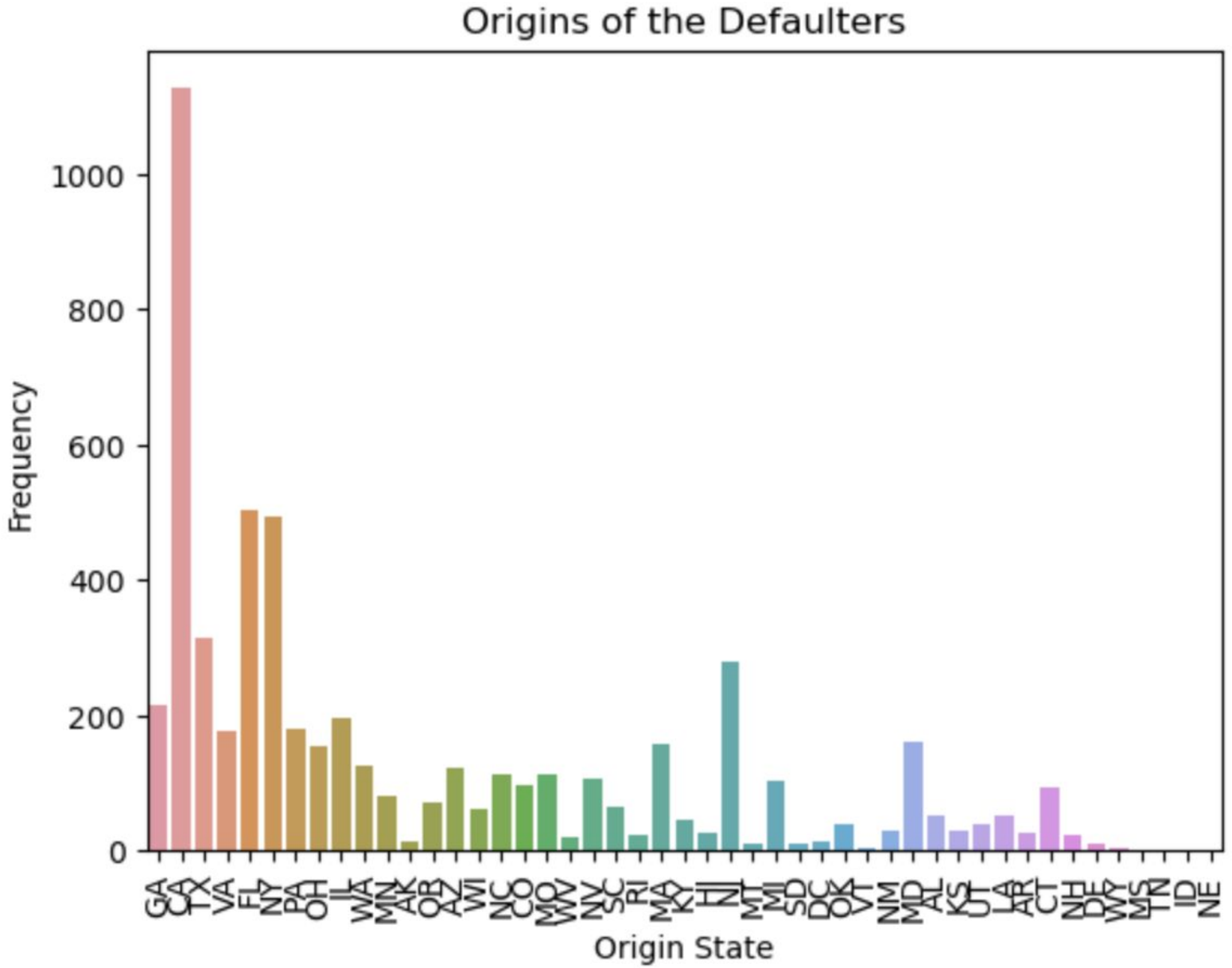
From the histogram we can see that the highest number of defaulters are those that had worked for 10 years or more. Those with 6-9 years of experience had the lowest number of defaulters



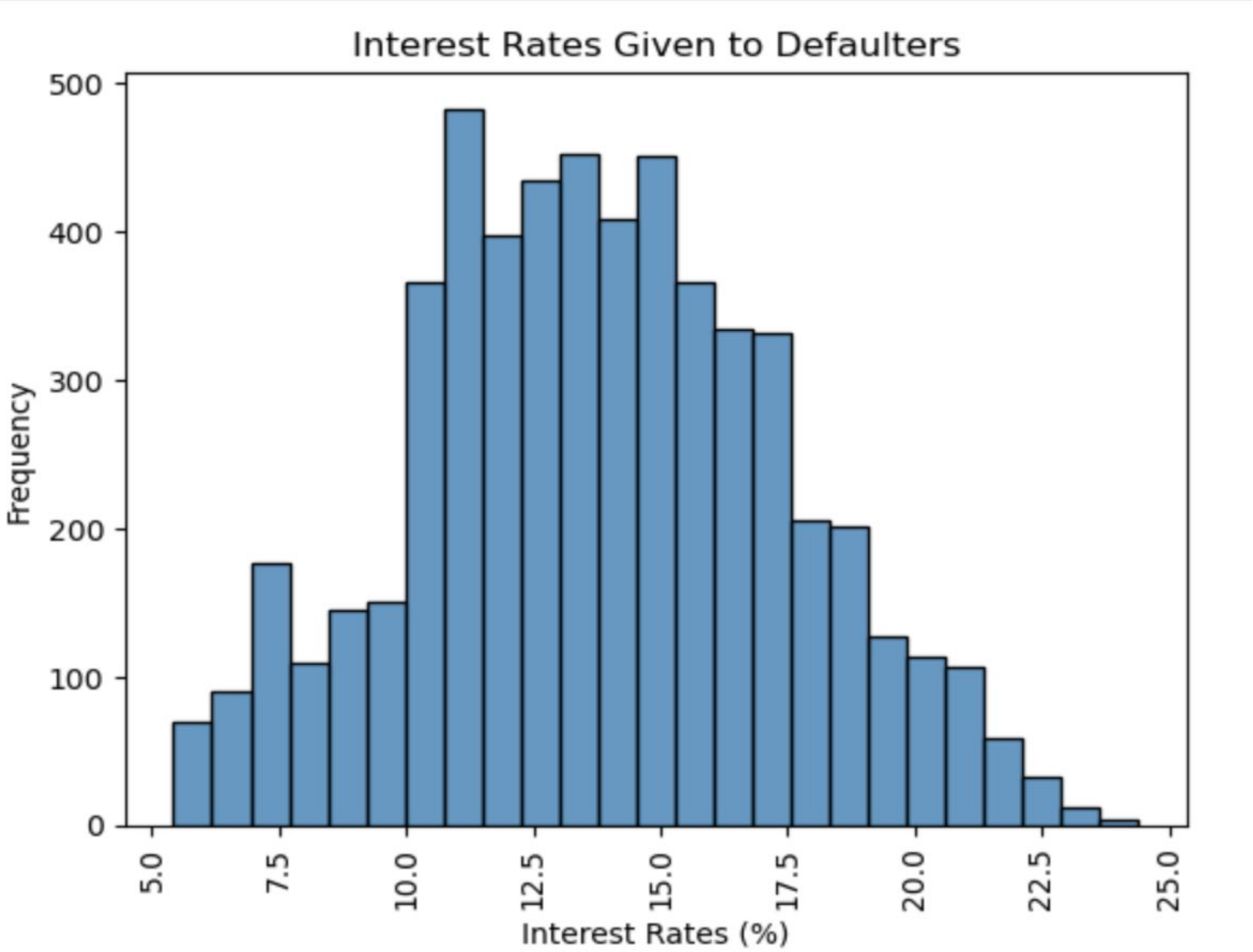
From the count plot we can see that a majority of defaulters are either paying rent or have a mortgage to pay



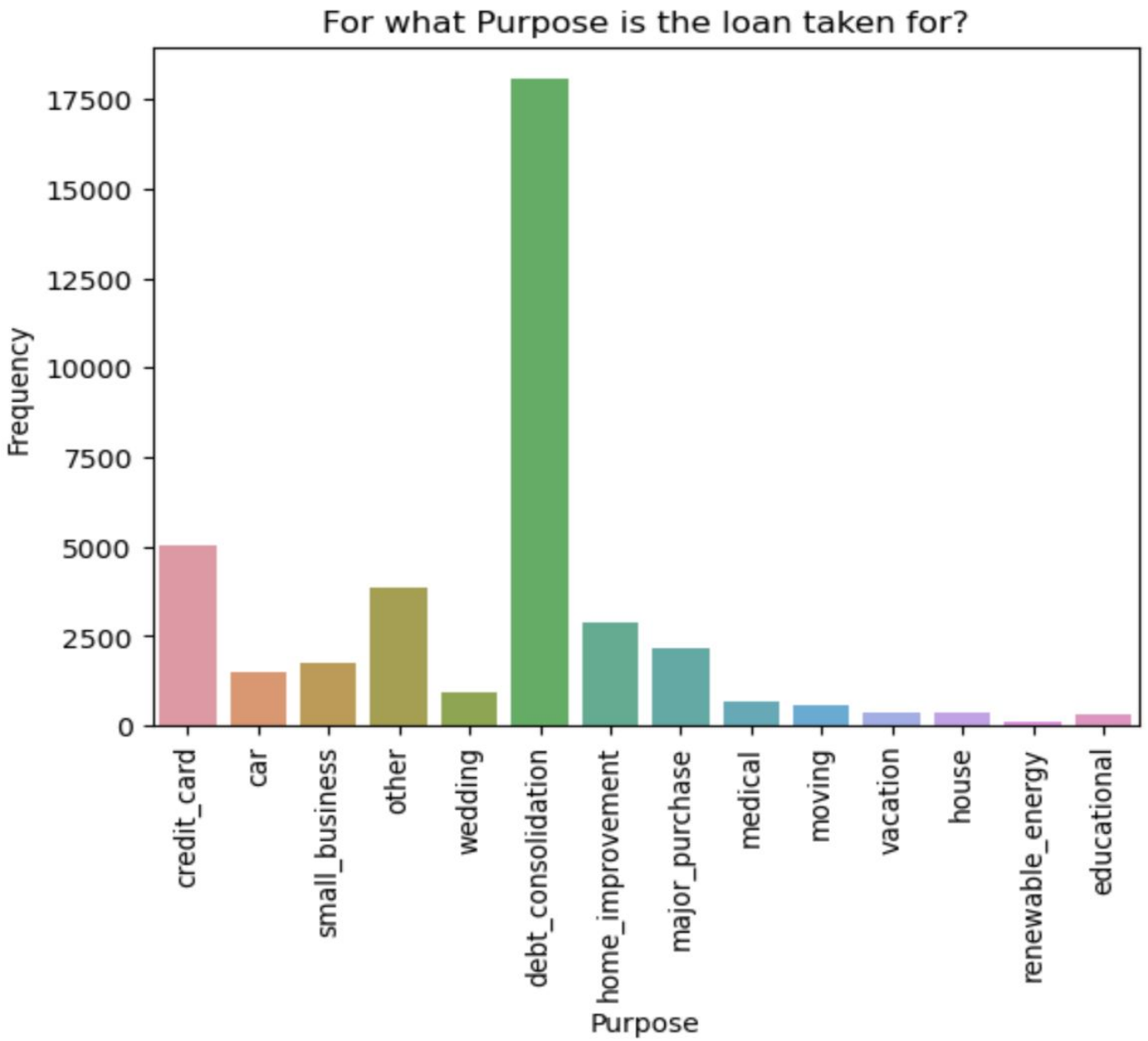
From the below histogram we see that a major portion of the loan defaulters were from the states of 'CA', 'FL' and 'NY' with 'CA' having the highest number of defaulters



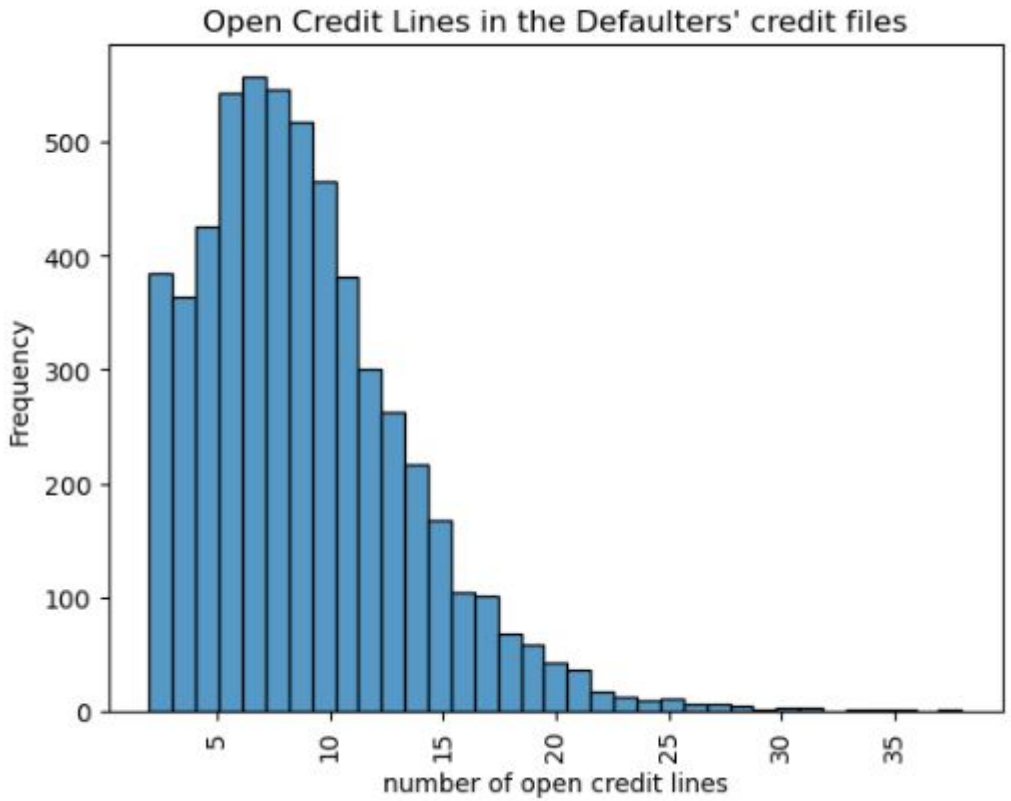
A Majority of the Defaulters had interest rates between 10% to 17.5%



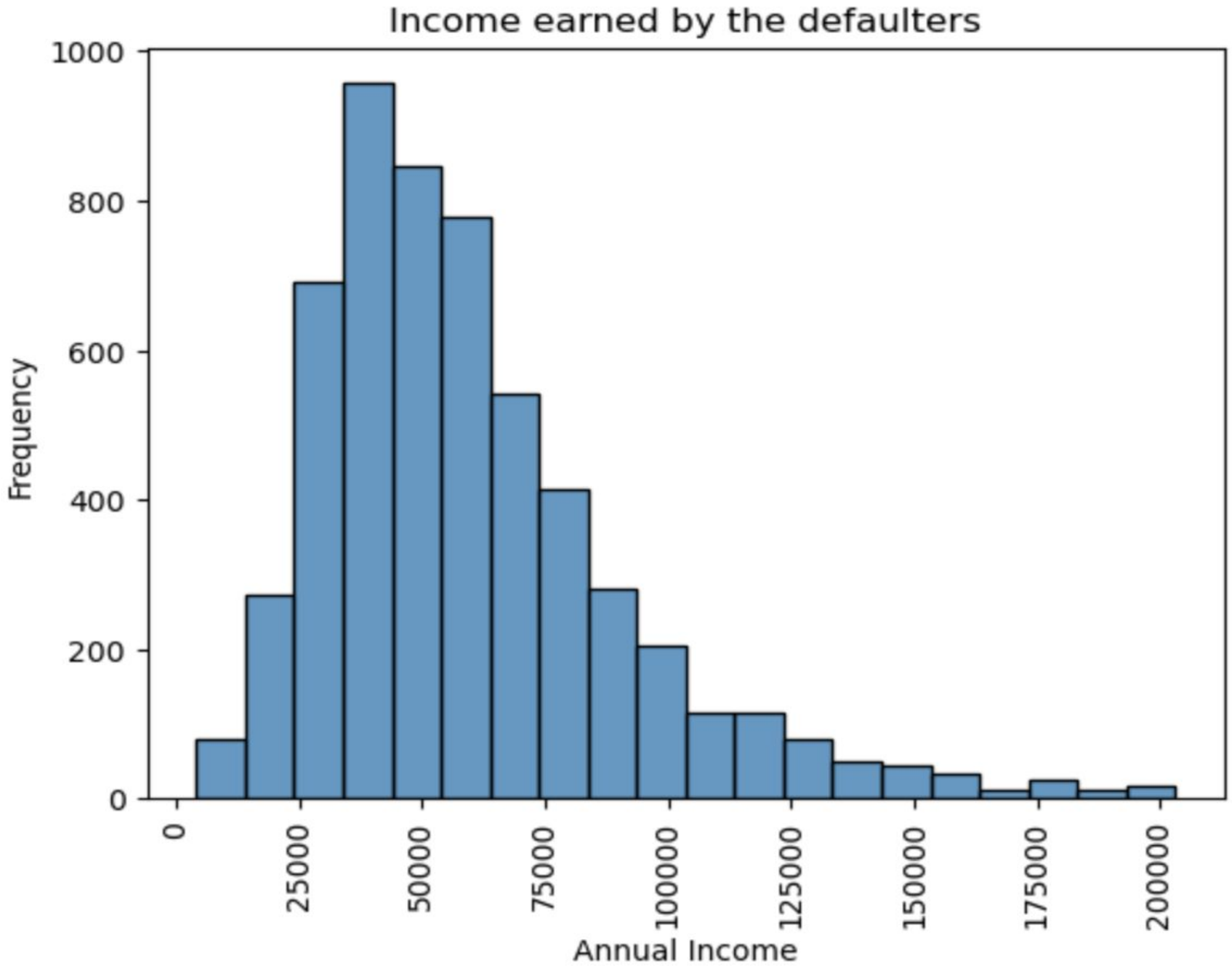
Most people end up taking loans for the purpose of debt consolidation



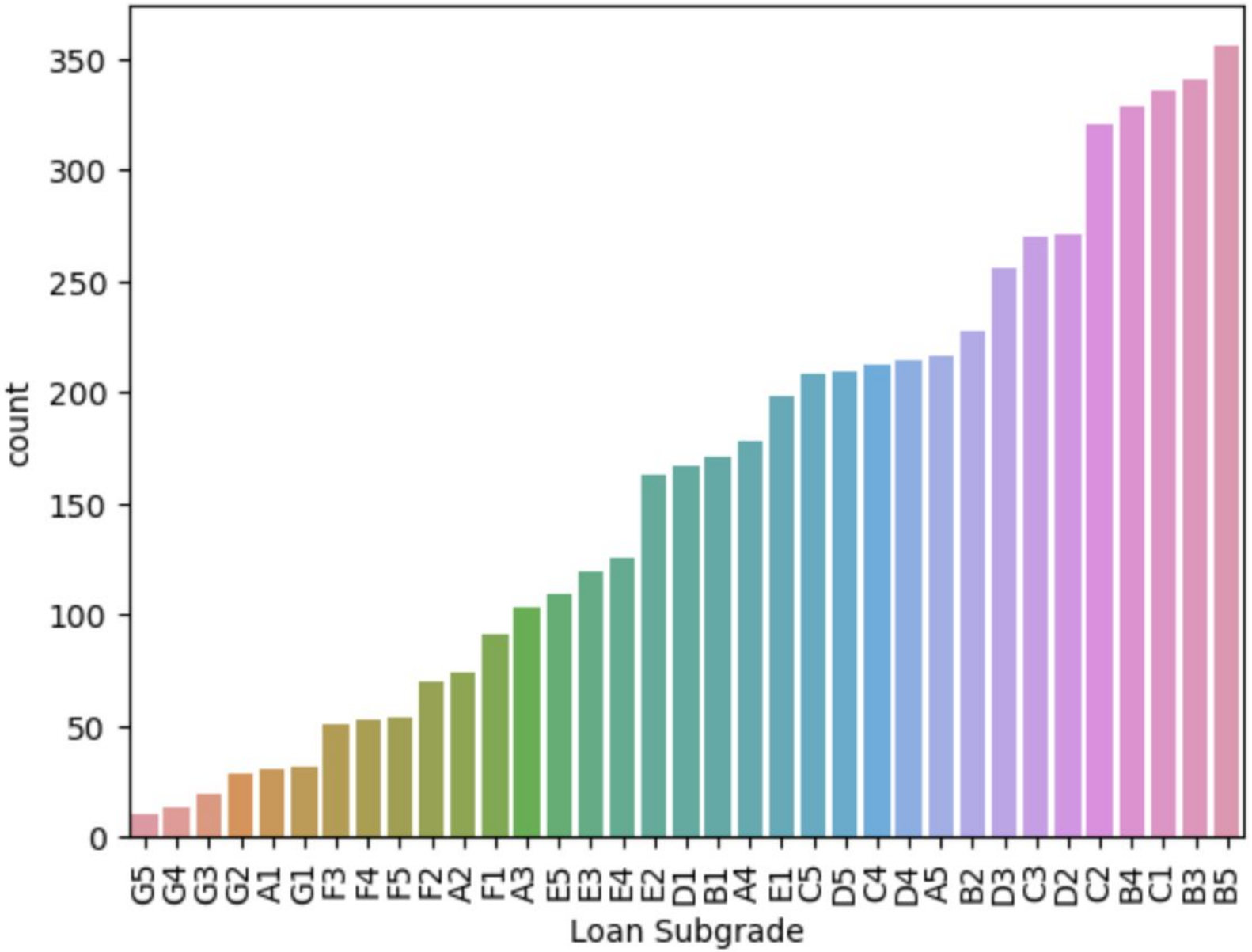
Majority of the defaulters had between 5 to 11 open credit lines in their credit files



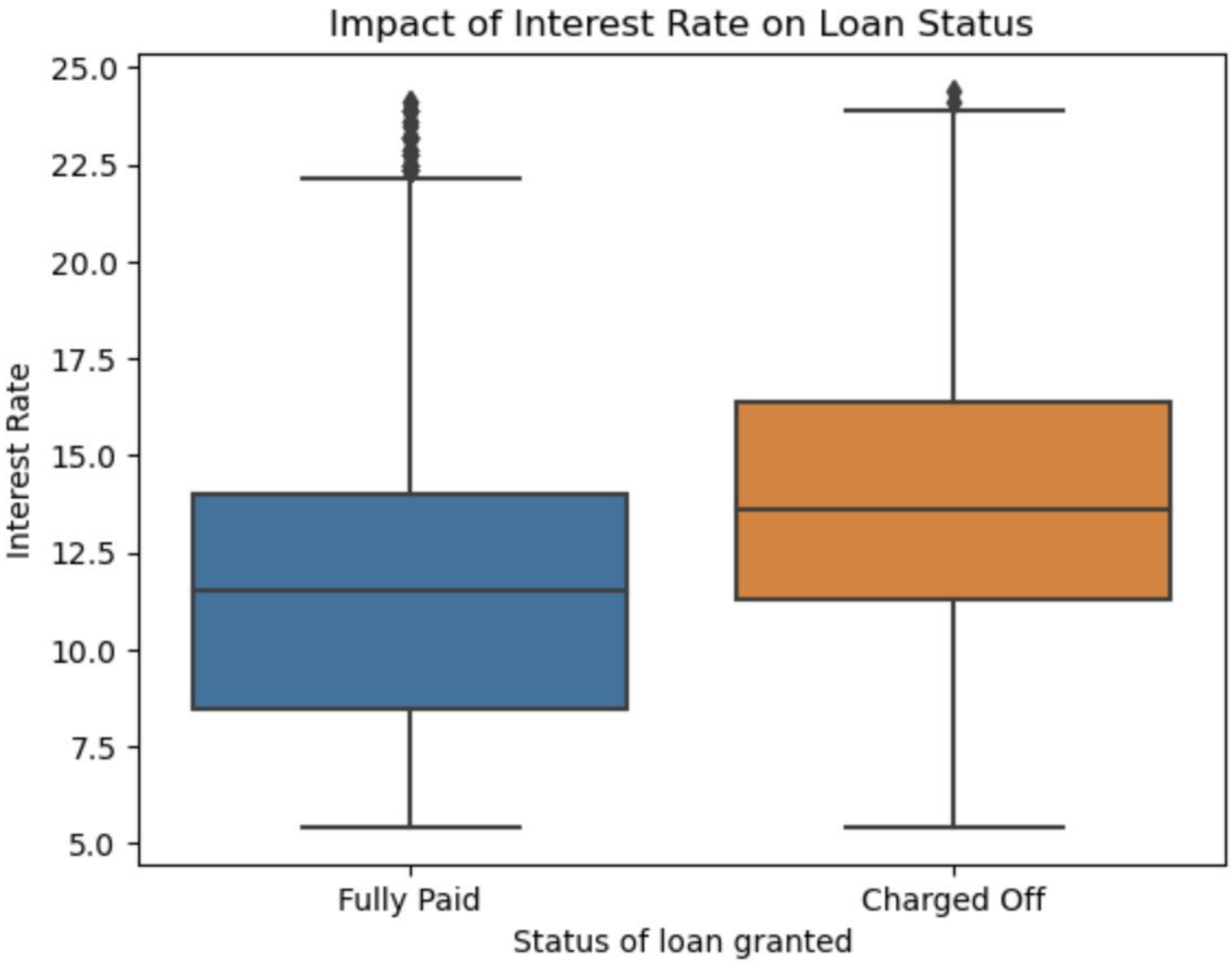
Most of the defaulters earn between 25000 and 75000



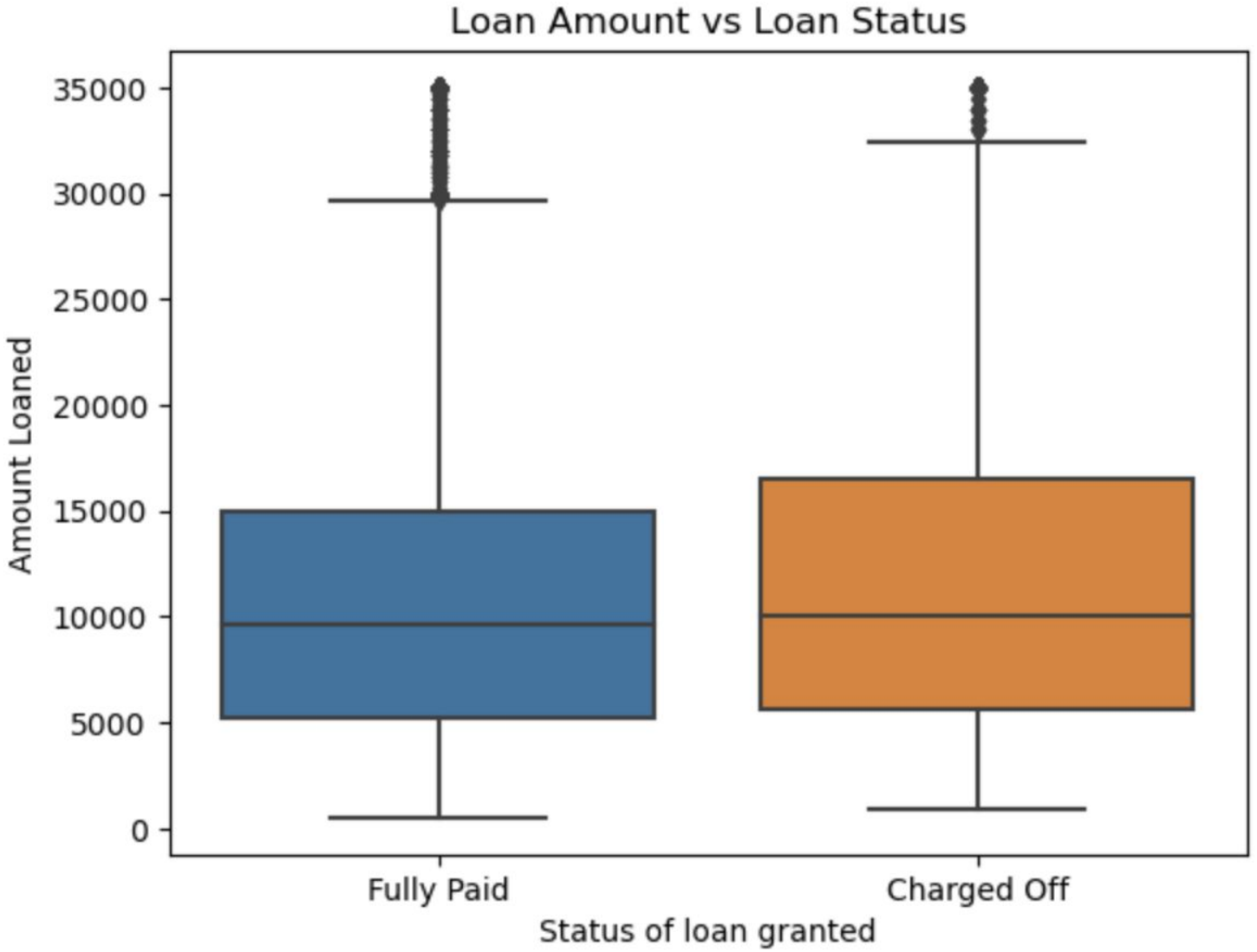
Loan Subgrades having values 'B4', 'C1', 'B3' and 'B5' have the highest number of defaulters



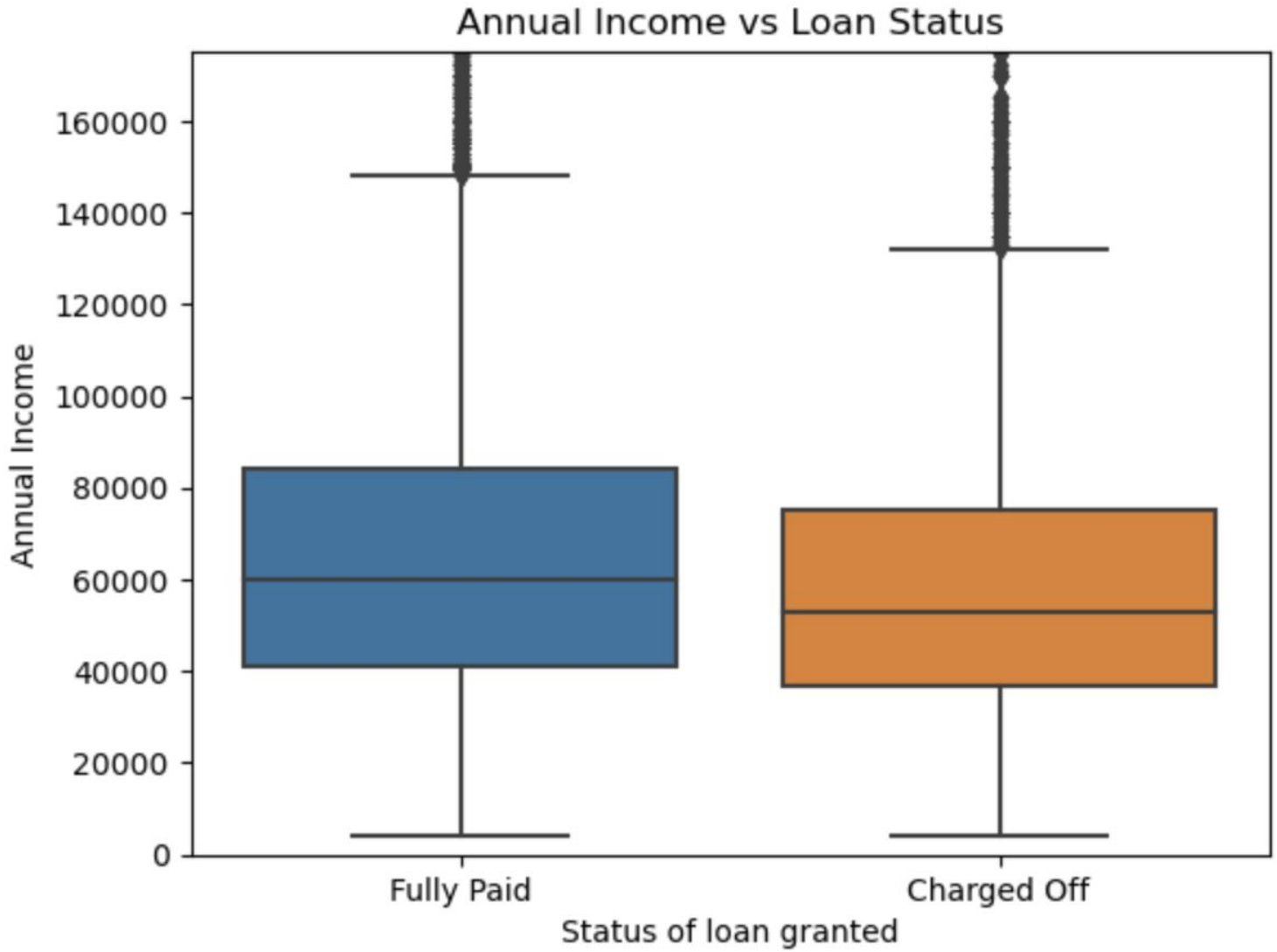
From this comparison, we see that the loans that defaulted had higher interest rates than those that were paid off. This is seen in the boxplot below which shows that the median rate for the loans that defaulted were greater than those that were fully paid



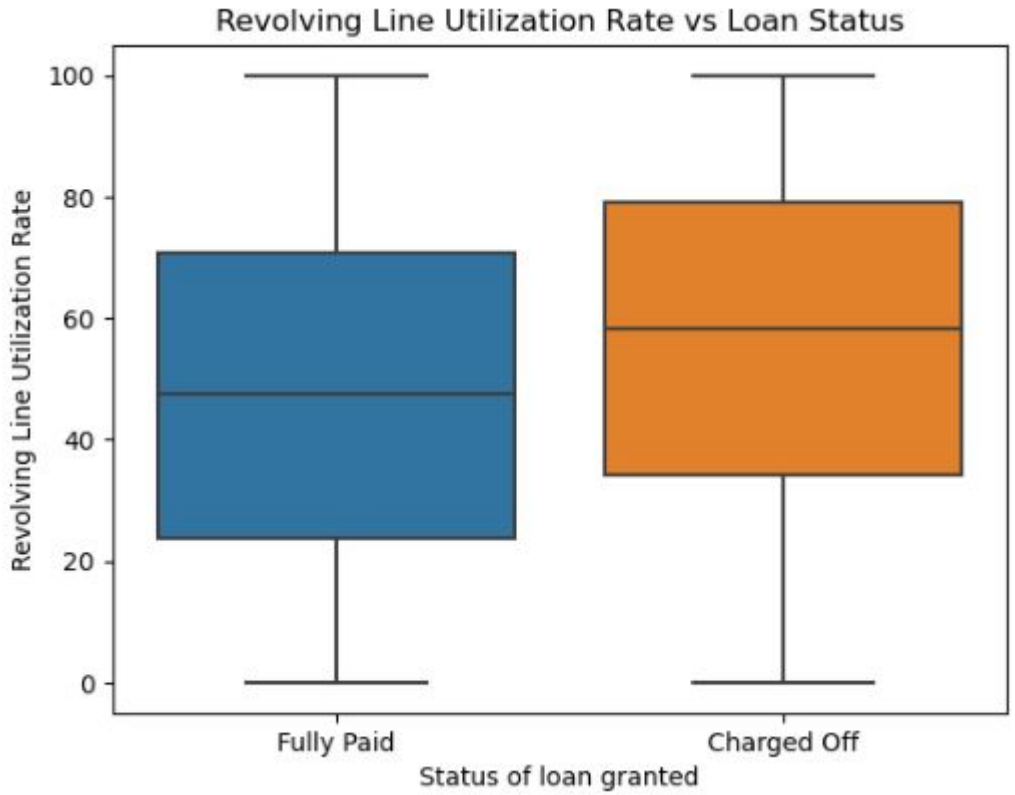
This difference in the interest rate is interesting, considering that there is not much of a difference with respect to the amount loaned to the borrowers as seen below, even if the loan amount for those that have defaulted are slightly higher



However, we see that non-defaulters have a higher income than those that have defaulted. This can definitely be a factor

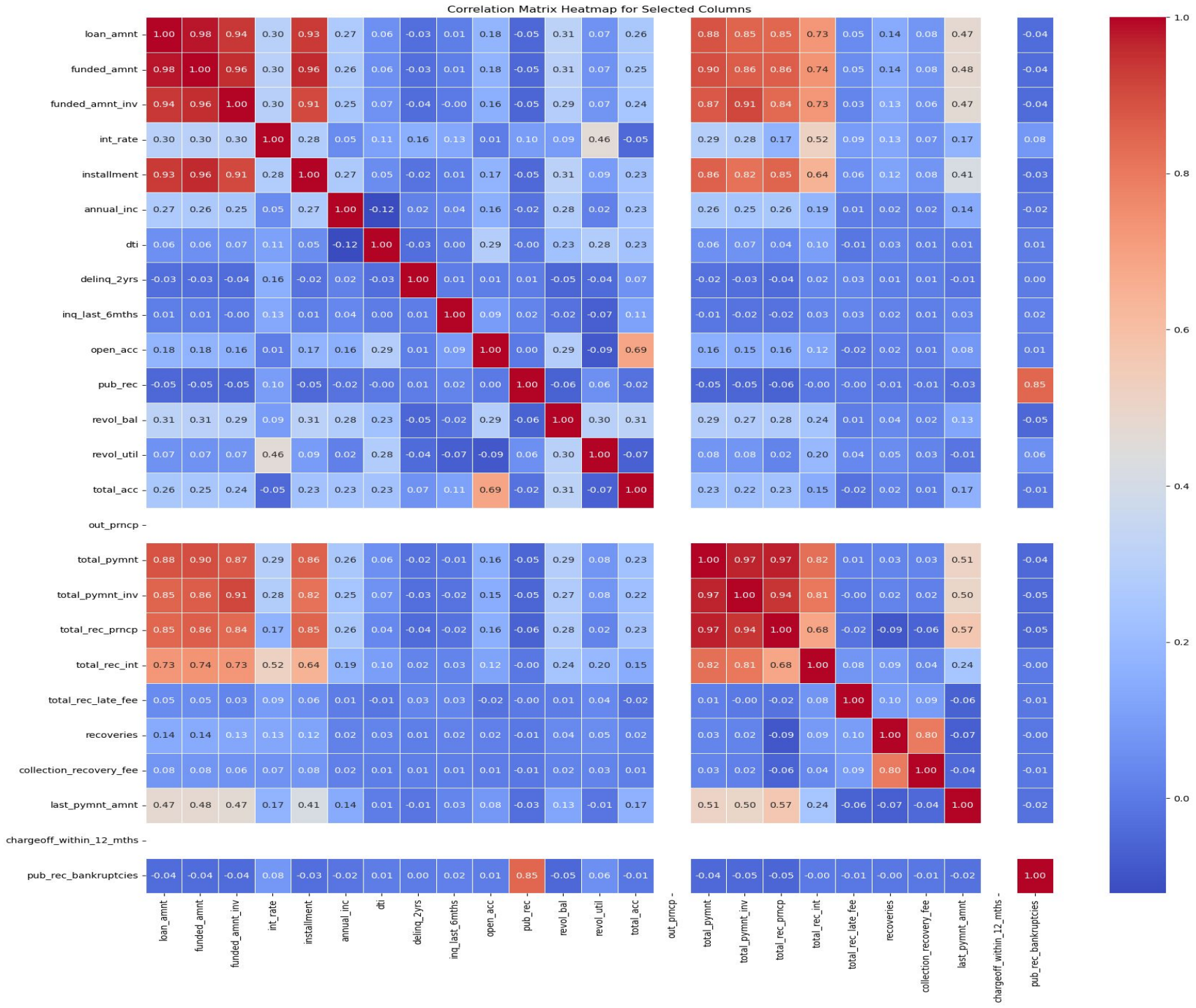


From this comparison, we see that the borrowers that defaulted had a higher amount of credit being used relative to all available revolving credit. This is seen in the boxplot which shows that the median Revolving Line Utilization Rate for the loans that defaulted were greater than those that were fully paid

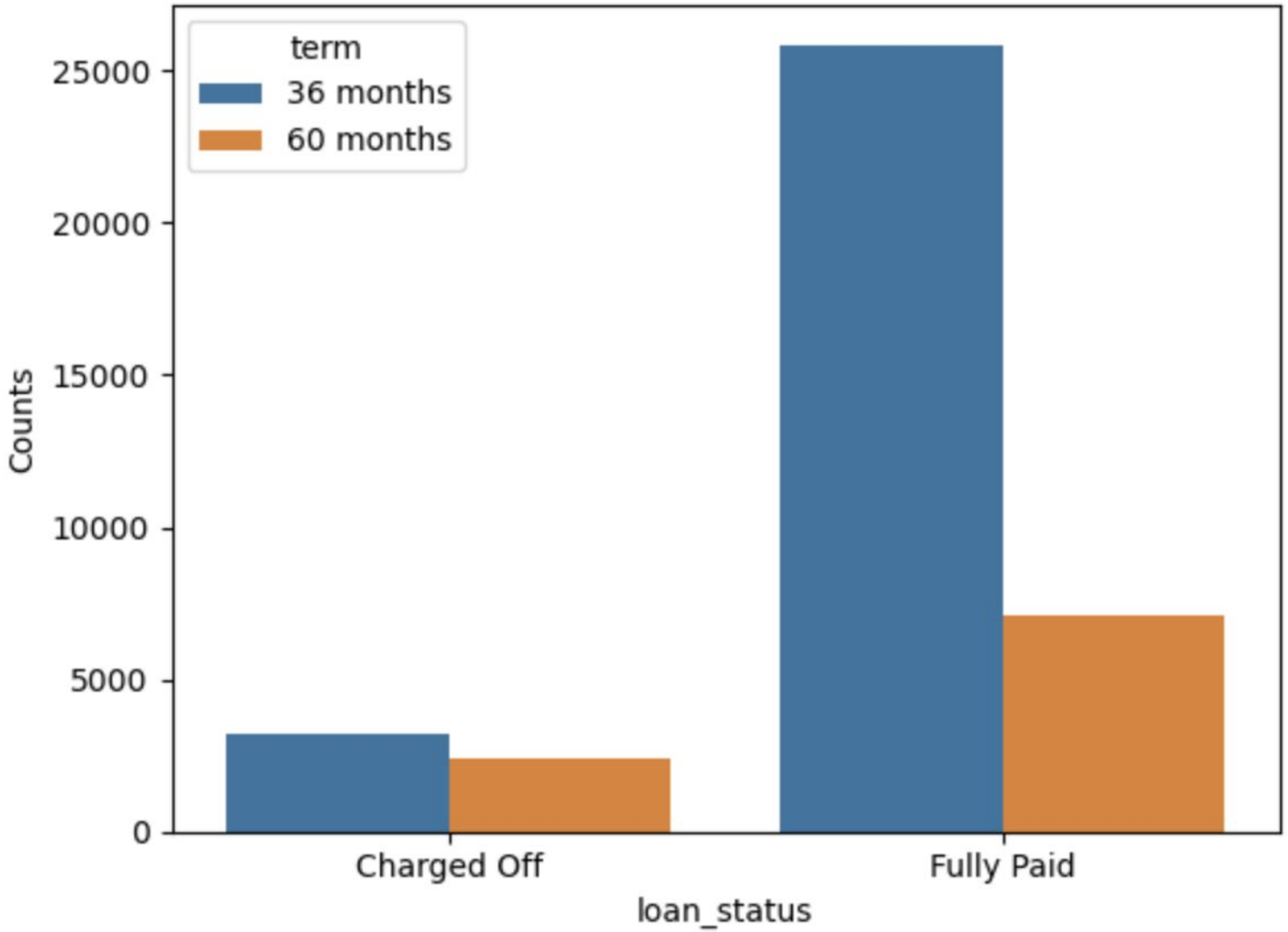


We see that installment is directly proportional to the loan amount, funded amount, and funded amount inverse

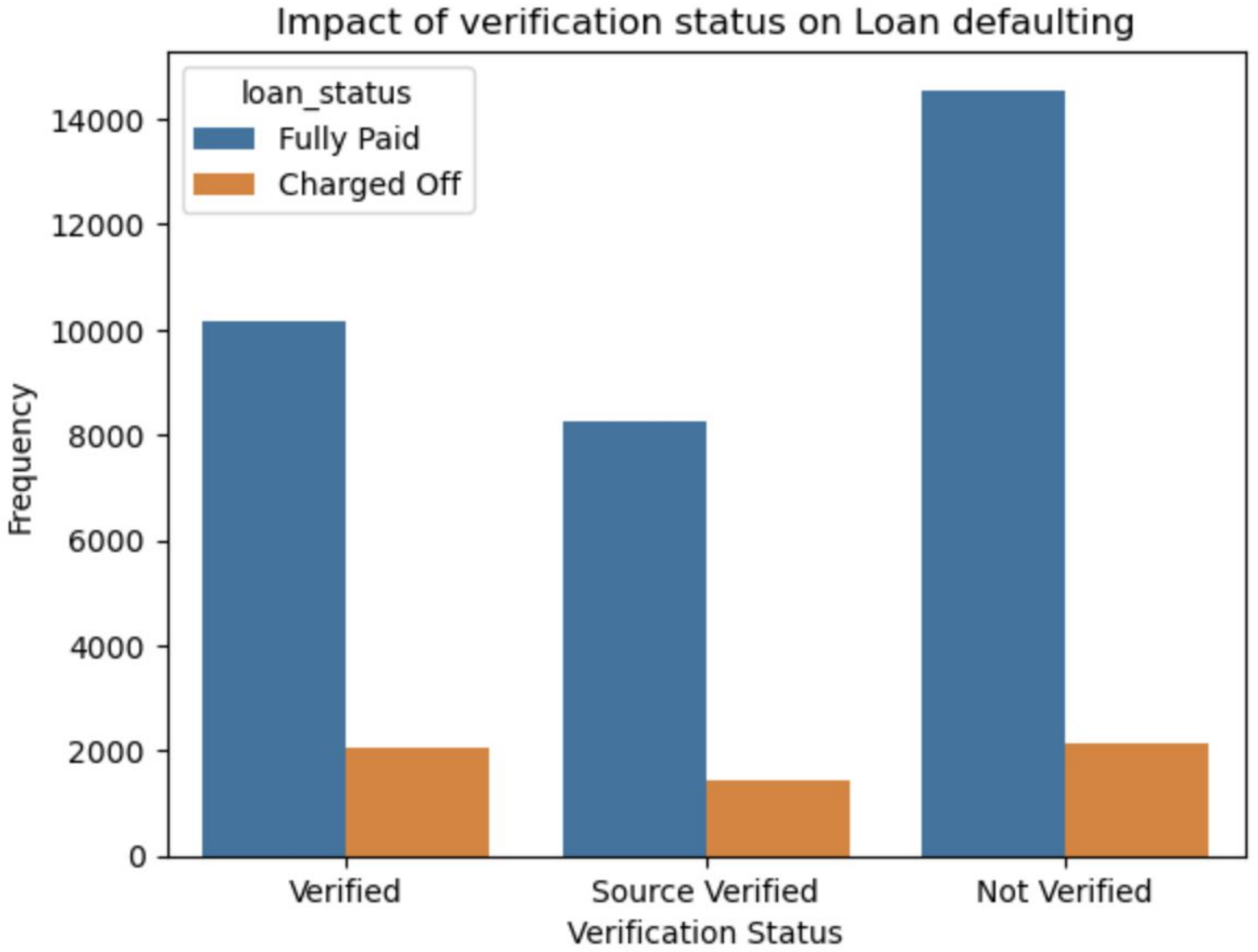
We also see a strong correlation between number of derogatory public records and pubublicly recorded bankruptcies



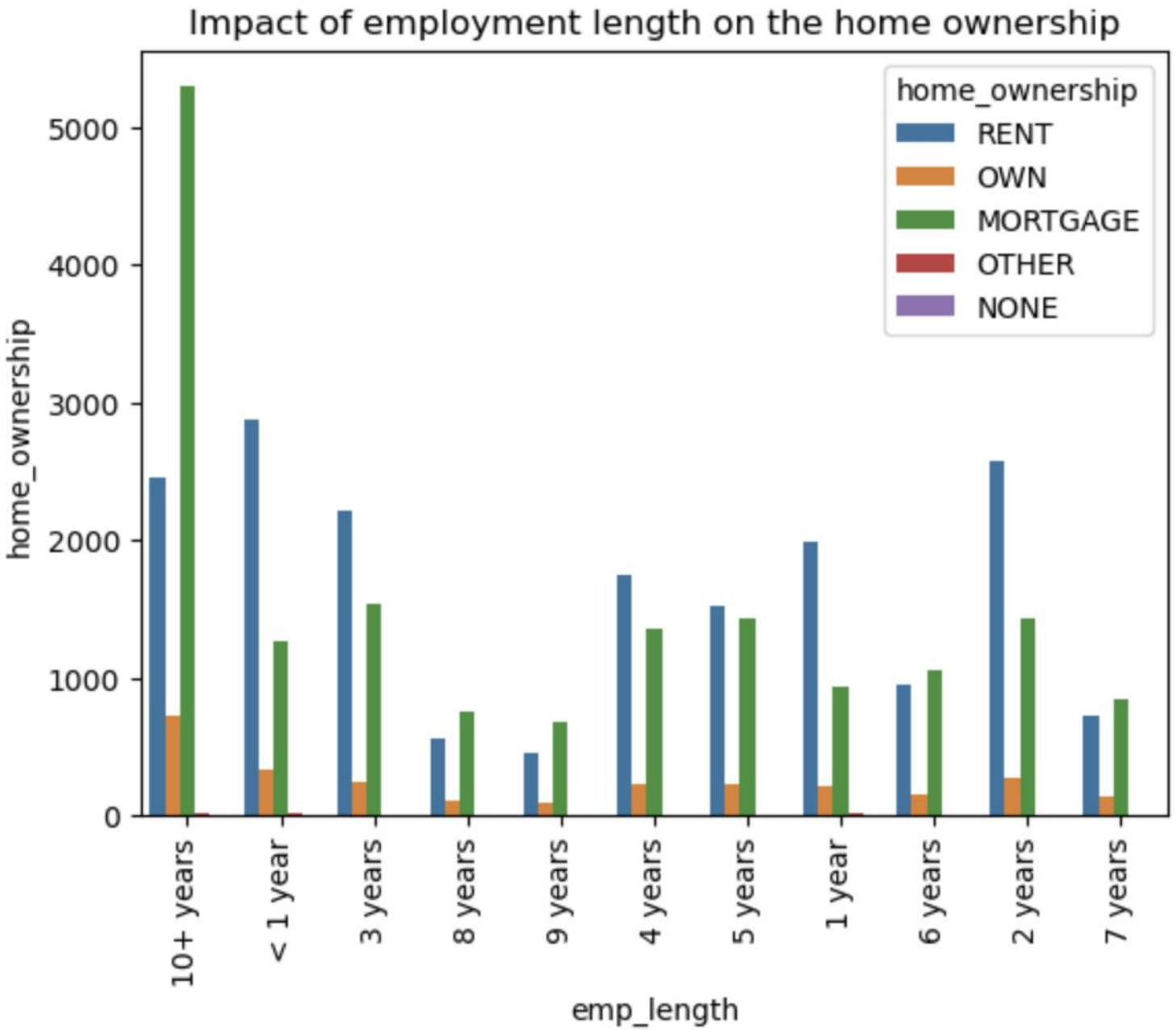
Based on the plot below, there is not much of a difference in loan terms for defaulters, but a majority of the lendees who have paid off their loans have opted for the shorter term of 36 months



Interestingly, Verification status does not make much of an impact on the defaulting of a loan



Borrowers who have been employed for at least 10 years have a highly prone to having a mortgage



Factors that play a major role in determining whether a borrower can be prone to defaulting on a loan

- Interest Rate (int_rate)
- Employment Length (emp_length)
- Term of the loan (term)
- Revolving Line Utilization Rate (revol_util)
- Annual Income (annual_inc)
- Home Ownership (home_ownership)

Recommendations

In order to prevent defaulting on the loans, it is best to set up a term of 36 months

Considering that defaulted loans had higher interest rates, it is best to reduce them in order to ensure that it is easier to pay off the loans.

Considering that those that have defaulted have lower incomes compared to those that have paid off their loans, we could relax the loan criteria for those with lower incomes

We must pay special attention to those that have worked for at least 10 years or those who are in the starting of their careers.

We must carefully consider those who are paying rent or having a mortgage.

Employees who have been employed for at least 10 years are prone to having a mortgage. We must pay more attention to this

We must be a little cautious around those that have a higher credit utilization ratio (revol_util)

Be cautious when considering a loan with the grades B,C and D

Borrowers who earn between 25000 and 75000 have the highest possibility of defaulting

Many Borrowers who had their status as verified or source verified still ended up defaulting, stricter checks must be implemented to ensure that the sources must be truly verified.