Gramener

A Data Science Company

International Institute of Information Technology Bangalore

Gramener Case Study

Statistics and Exploratory Data Analytics

UpGrad

Contributors:

- Preethi Doravari
- Arifa Kovuru
- Sahil Bansal
- Aditya Chopra

Introduction

- Solving this case study will give you an idea about how real business problems are solved using EDA.

- In this case study, apart from applying the techniques you have learnt in EDA, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Objectives

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

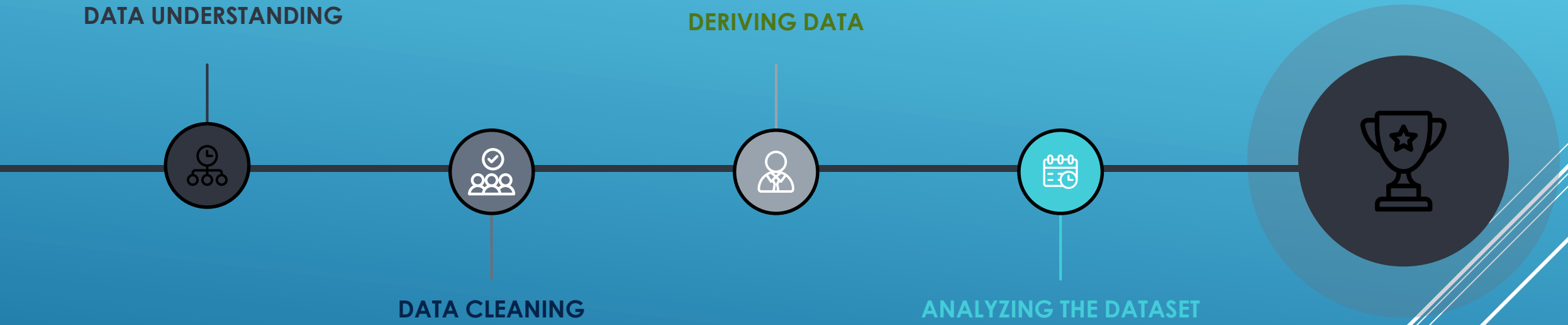
- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

Business Objectives Contd.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Problem Solving Methodology



DATA UNDERSTANDING

- Going through the whole dataset
- Understanding the meaning of the columns given
- Looking at the dataset both using Python as well as Excel
- Creating a list of columns that can be derived further into other columns to get useful insights

DATA CLEANING

- Finding out the null percentages of the columns
- Removing the columns with 100% null values
- Removing other columns with less than 100% null values but not required
- Removing rows with null values to get a clean dataset
- Converting the columns to the required datatype

DERIVING DATA

Creating new columns out of existing columns to get better insights

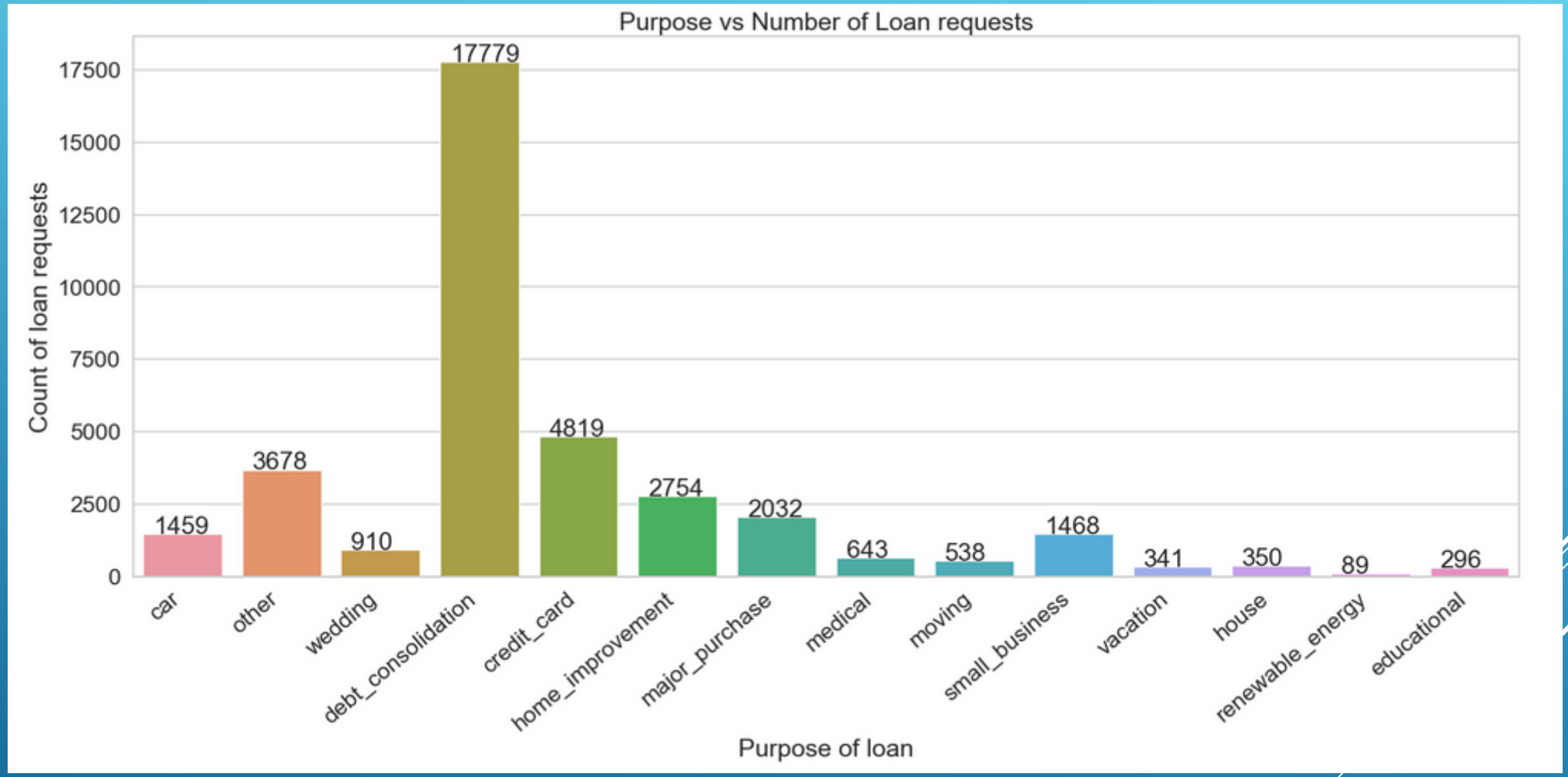
ANALYZING THE DATASET

- Performing Univariate analysis on the columns
- Performing Bivariate analysis
- Jotting down the insights that we get from the analysis
- Removing the outliers if found just like in case of Annual Income

DISPLAYING THE RESULTS

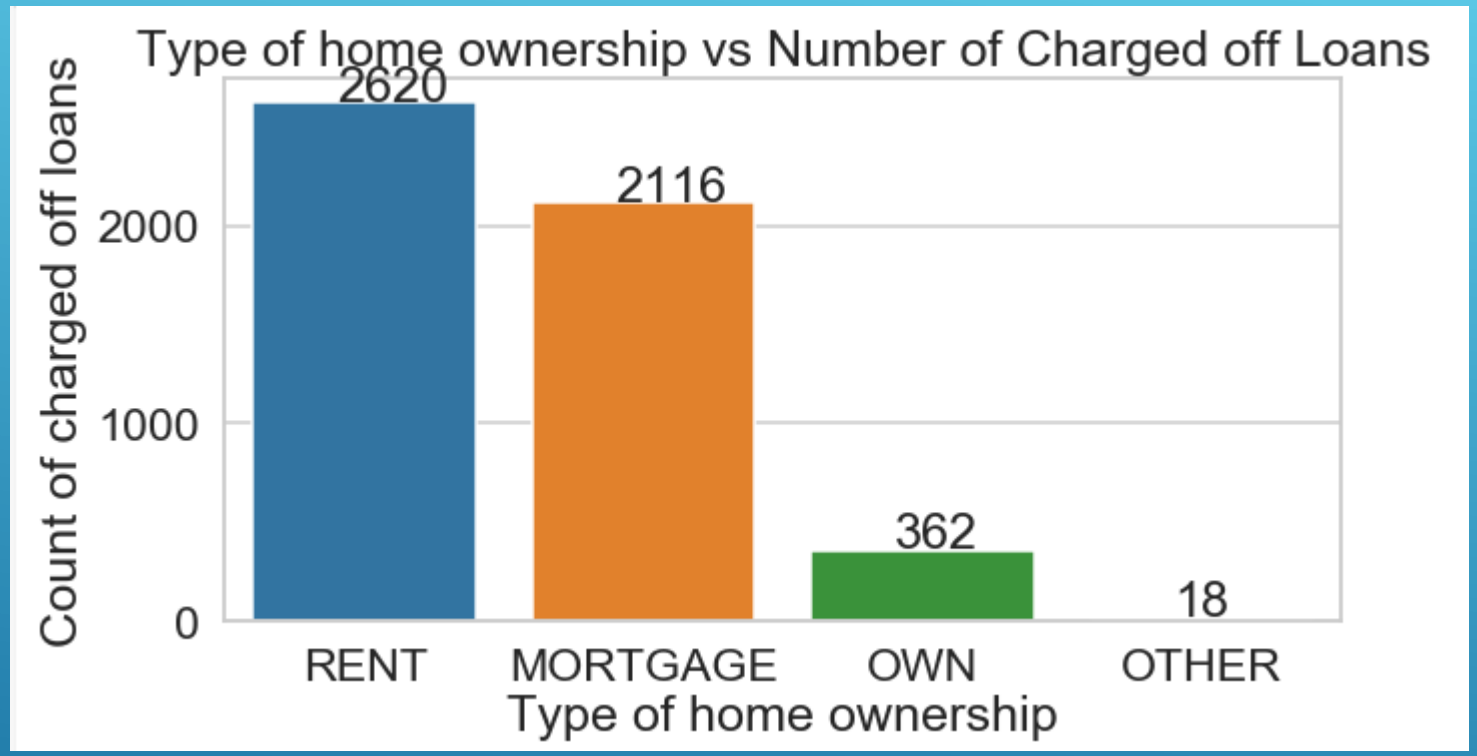
- Creating the final presentation to display the analysis
- Also, describing the factors that may be used in order to filter out defaulters

Univariate Analysis



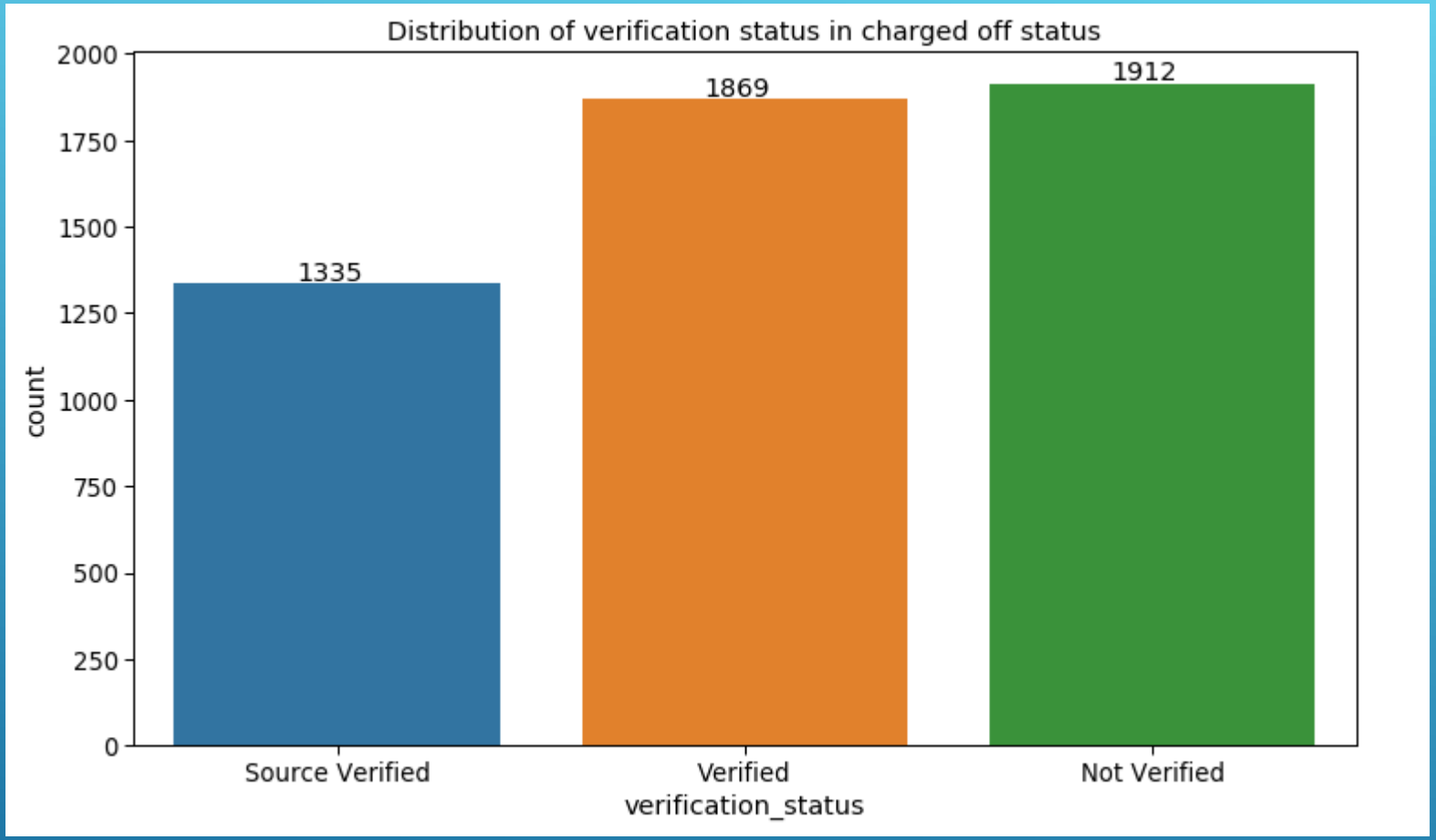
The highest number of requests are received for the purpose of 'debt_consolidation' as seen above

Univariate Analysis

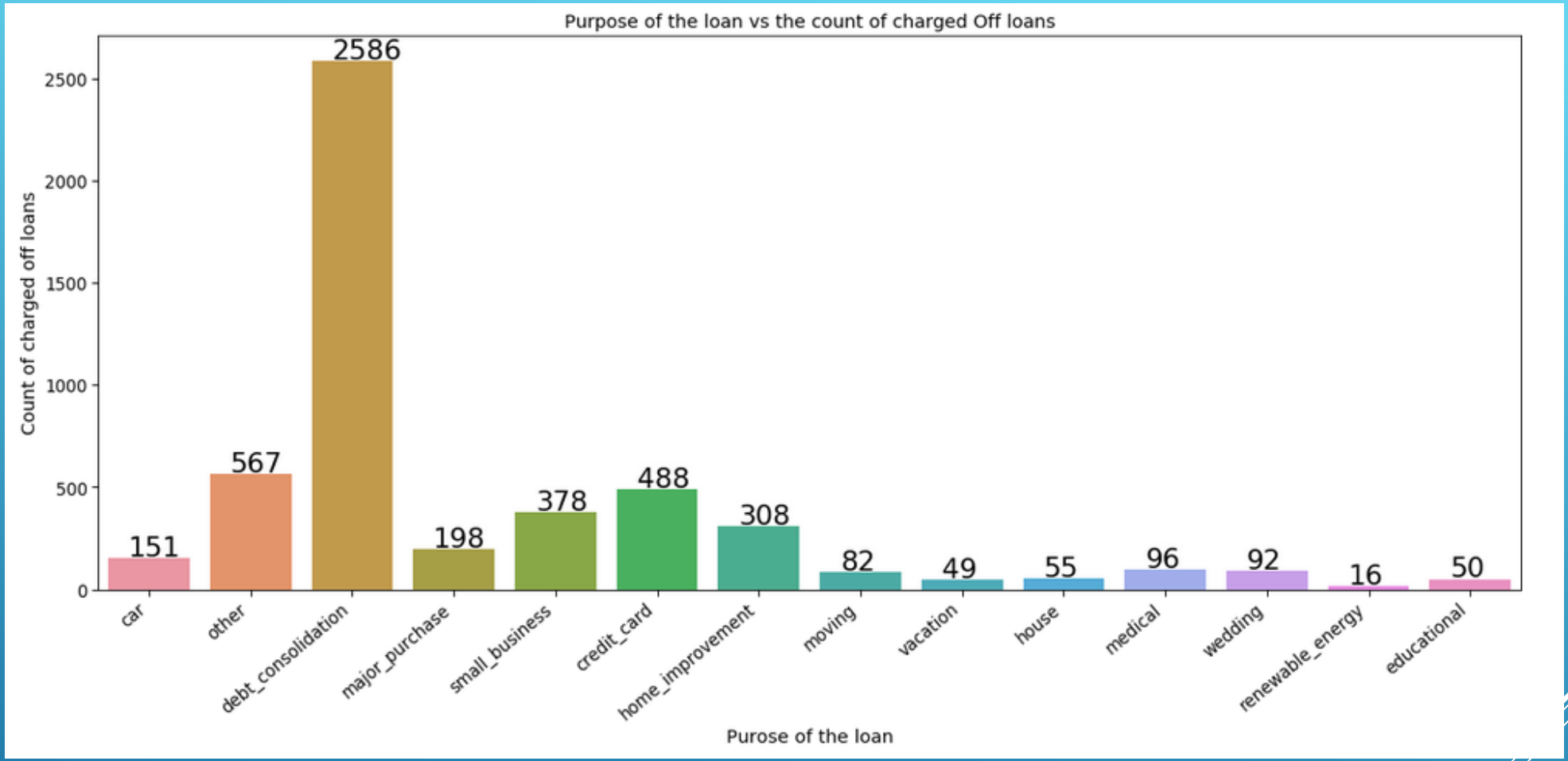


From the above plot we see that loan requesters who have 'rent' as the home ownership are the most likely to default. Moreover, we can rely on the persons who OWN the house rathern than who have 'RENT' or 'MORTGAGE' as their 'home_ownership' status

Univariate Analysis

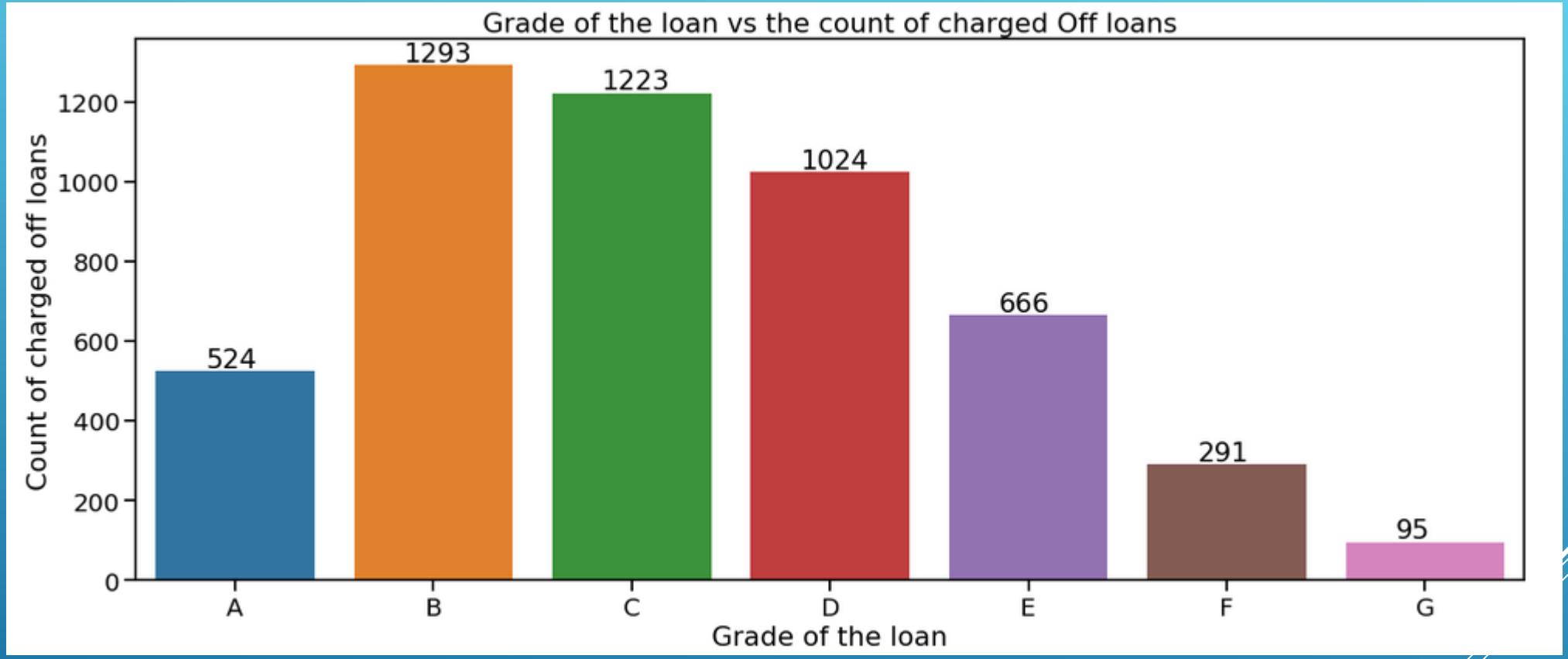


Not verified and Verified verification status shows higher count of charged off loans than the source verified ones. Thus, the banks should do the source verification to minimise the credit loss.



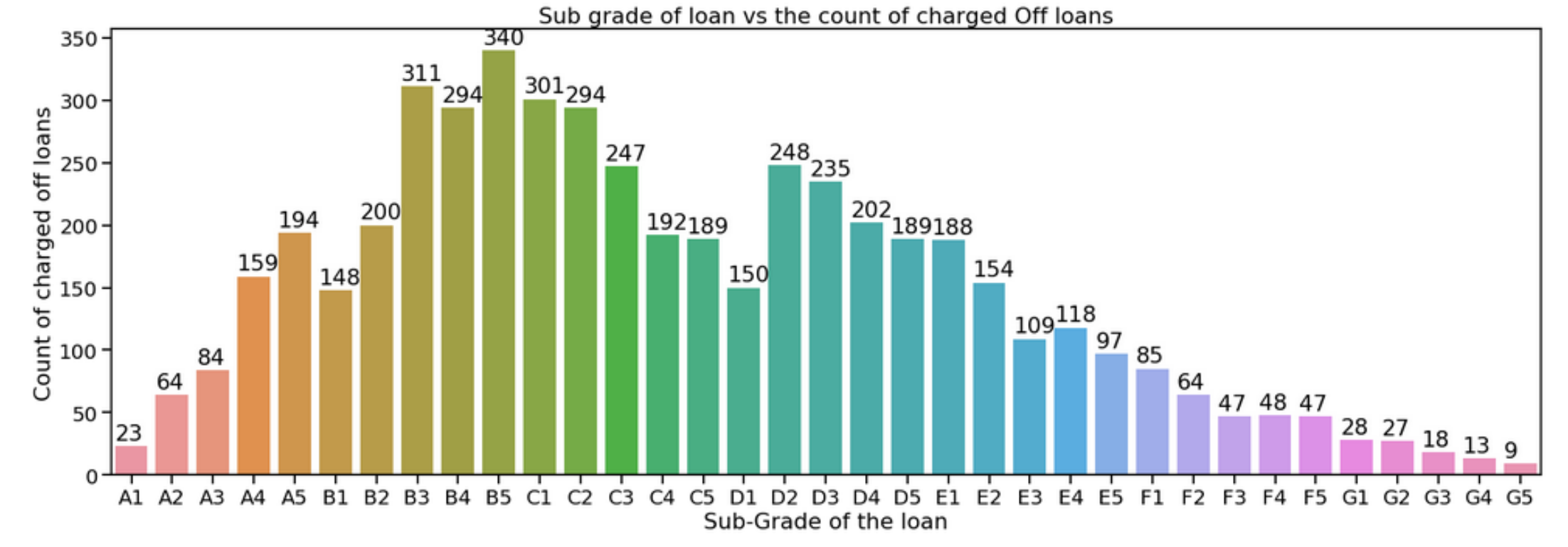
- From the above plot we see that the purpose 'debt_consolidation' has the highest count of charged off loans. So, one should minimise the loan sanctions for this purpose.
- Moreover, the second highest charged off loans are for the purpose 'other' which means that the banks should define some more categories or ask the person if he/she is taking the loan mentioning purpose as 'others' to minimise the risk.

Univariate Analysis



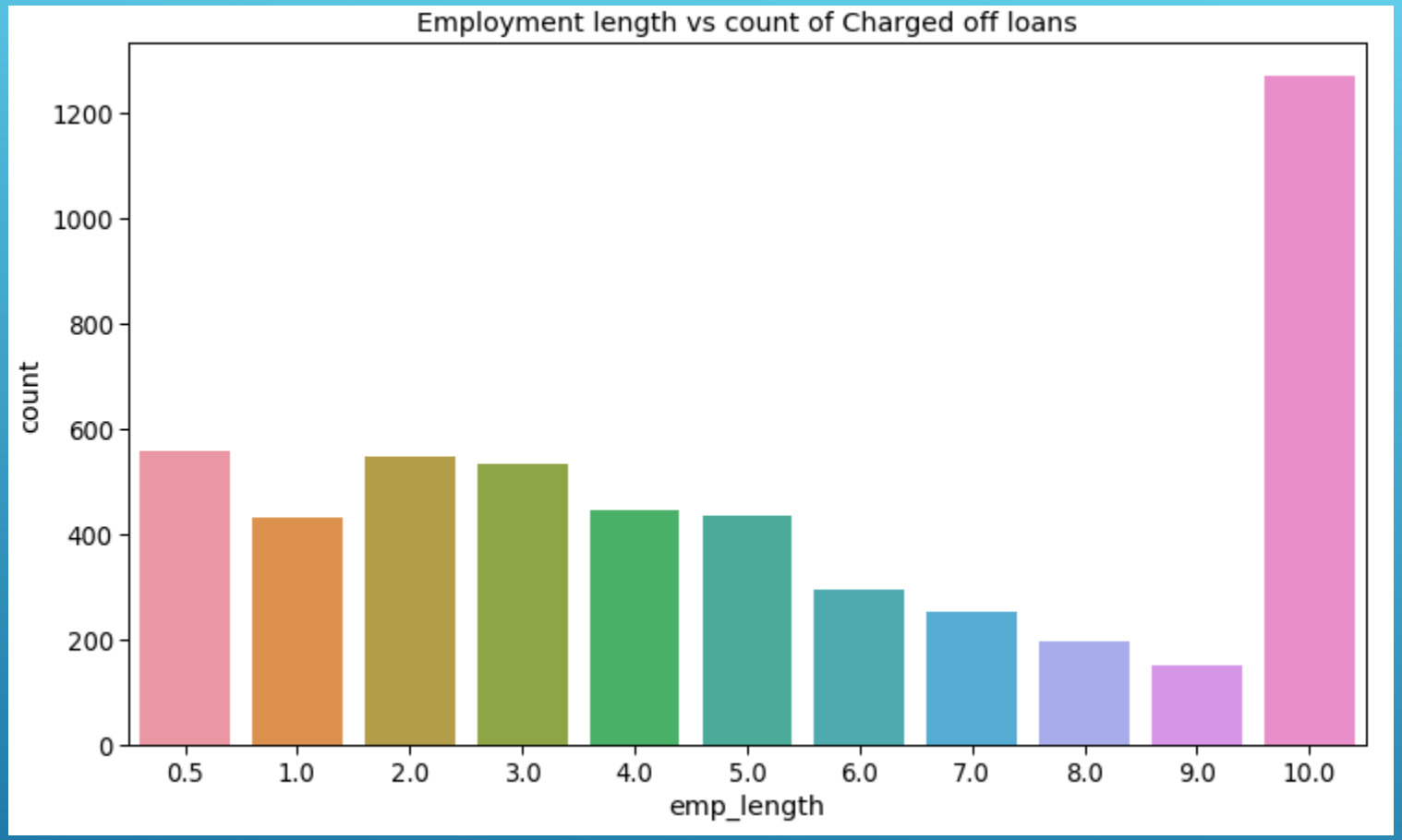
From the above plot, its evident that the charged off loans are highest in grade 'B' and this number decrease as we move on to higher grades i.e. C to G

Univariate Analysis



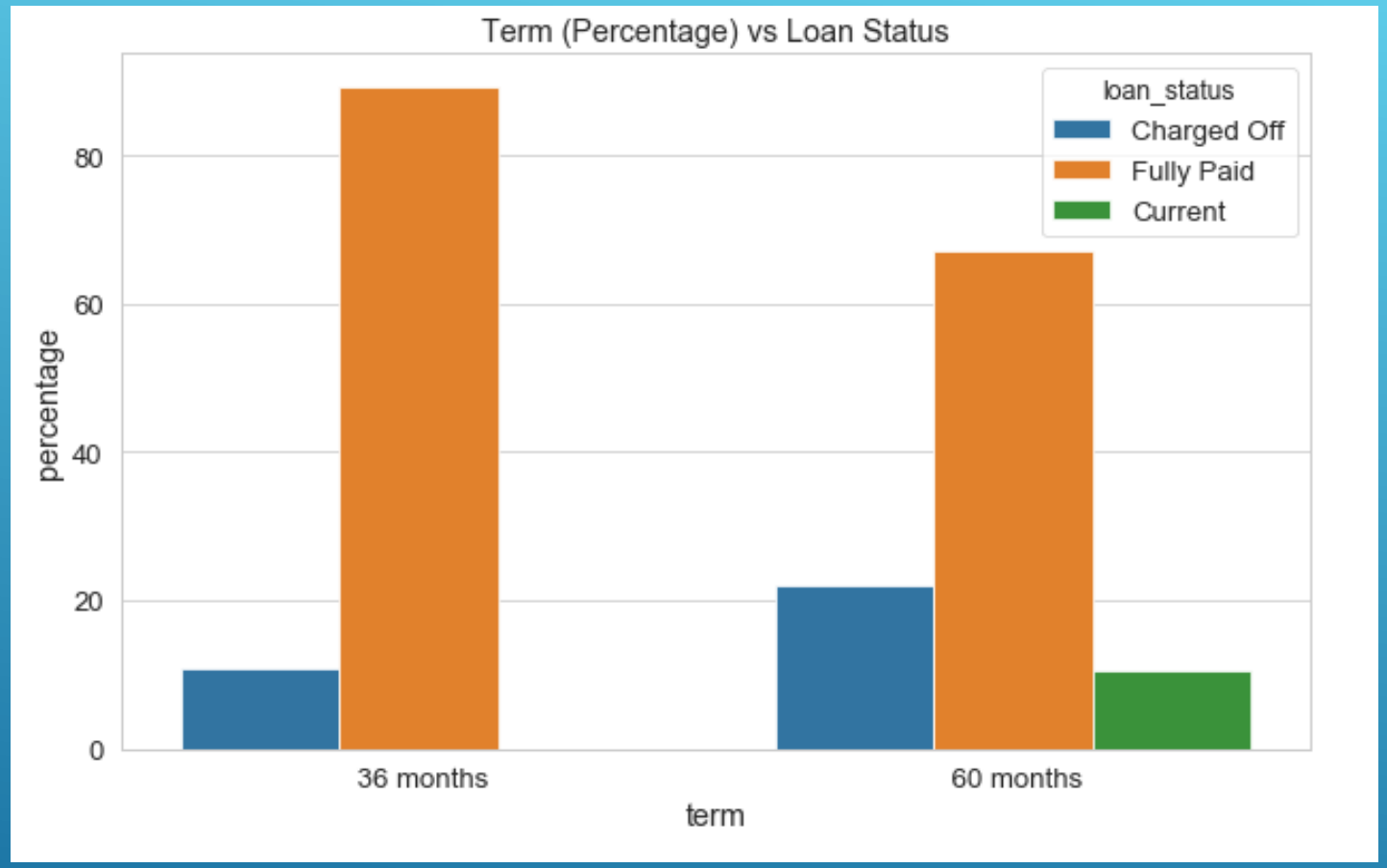
From the above plot, we see that the states where the loans have been charged off are mostly from CA, NY, FL, TX, NJ, GA

Univariate Analysis



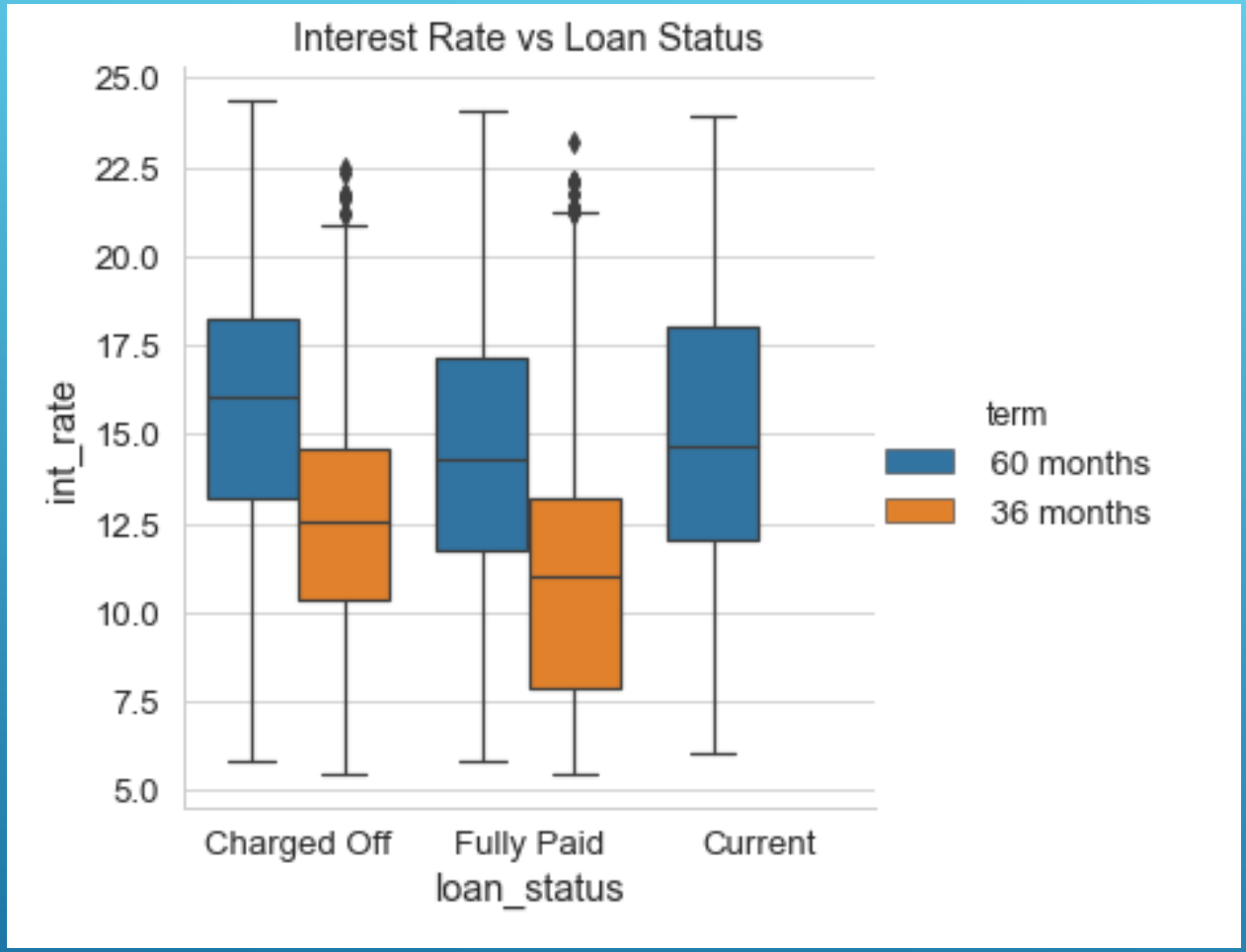
Interestingly 10+ i.e. (10 here) years experienced members have 2 times more charged off records than freshers, however members with 4-9 years of job experience showed relatively 2-6 times less charged off record when compared to 10+years experienced. So, the 10+ experiences are most likely to default as per the above plot.

Bivariate Analysis



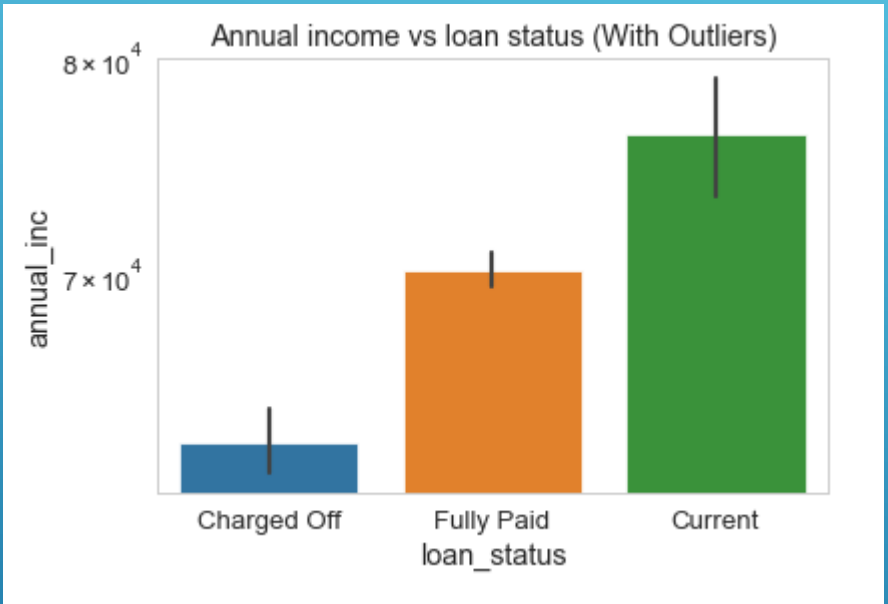
From the above graph it is evident that borrowers with 60 months term are likely to default as compared to 36 months, whereas fully paid members are high in 36 months.

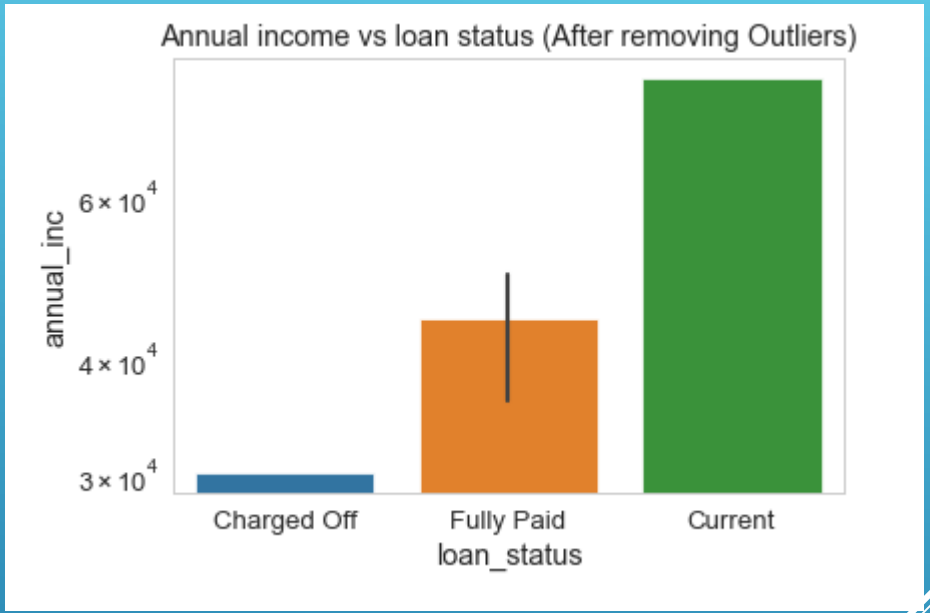
Bivariate Analysis



- As seen from the above plot, we see that we have more fully paid loans for lower interest rates and lower term (period)
- Also, the borrower is more likely to default with higher interest rates and higher term

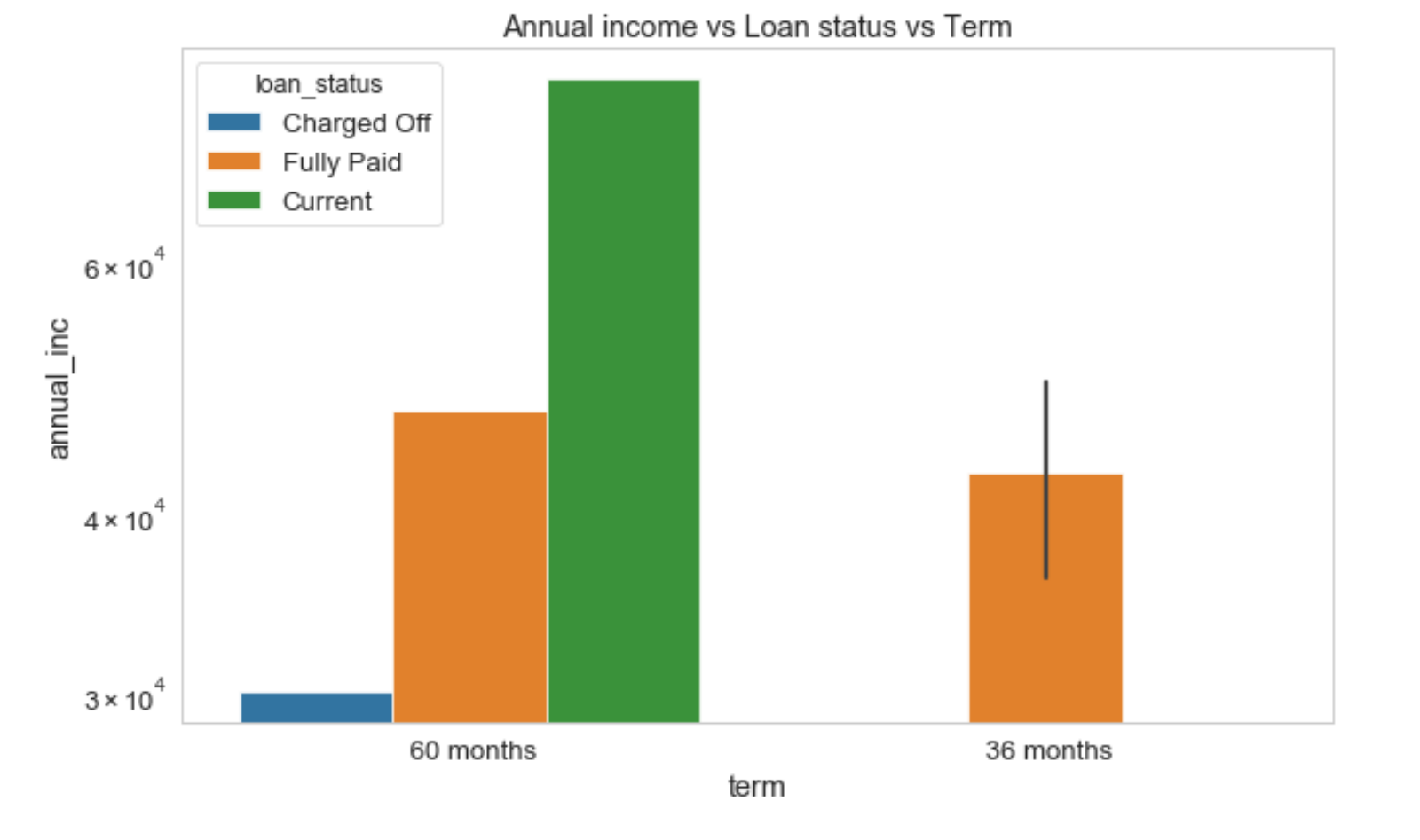
Bivariate Analysis





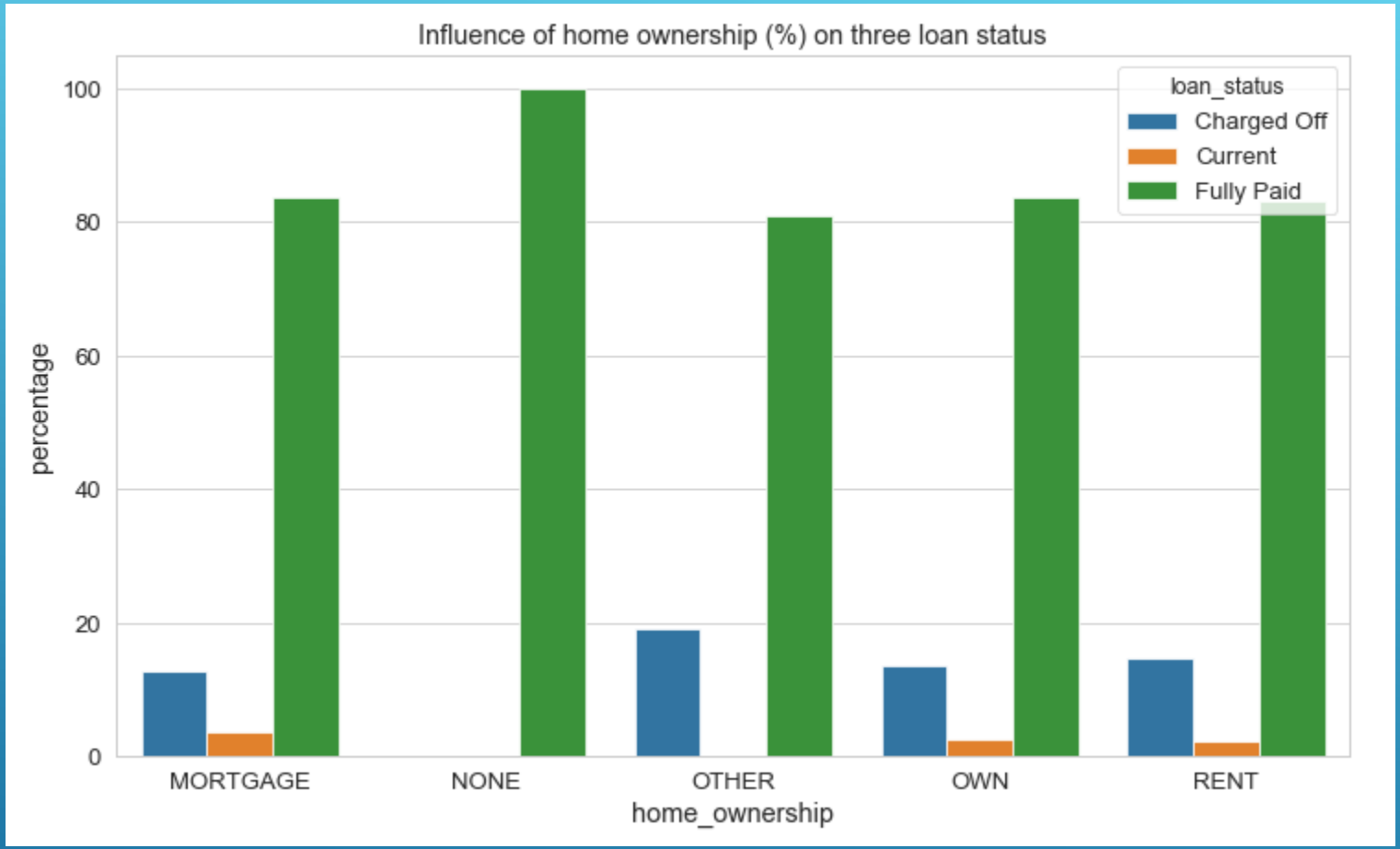
From the above graph, it is evident that mean annual income of borrowers who are charged off from the loans is less when compared to fully paid or current status ones.

Bivariate Analysis



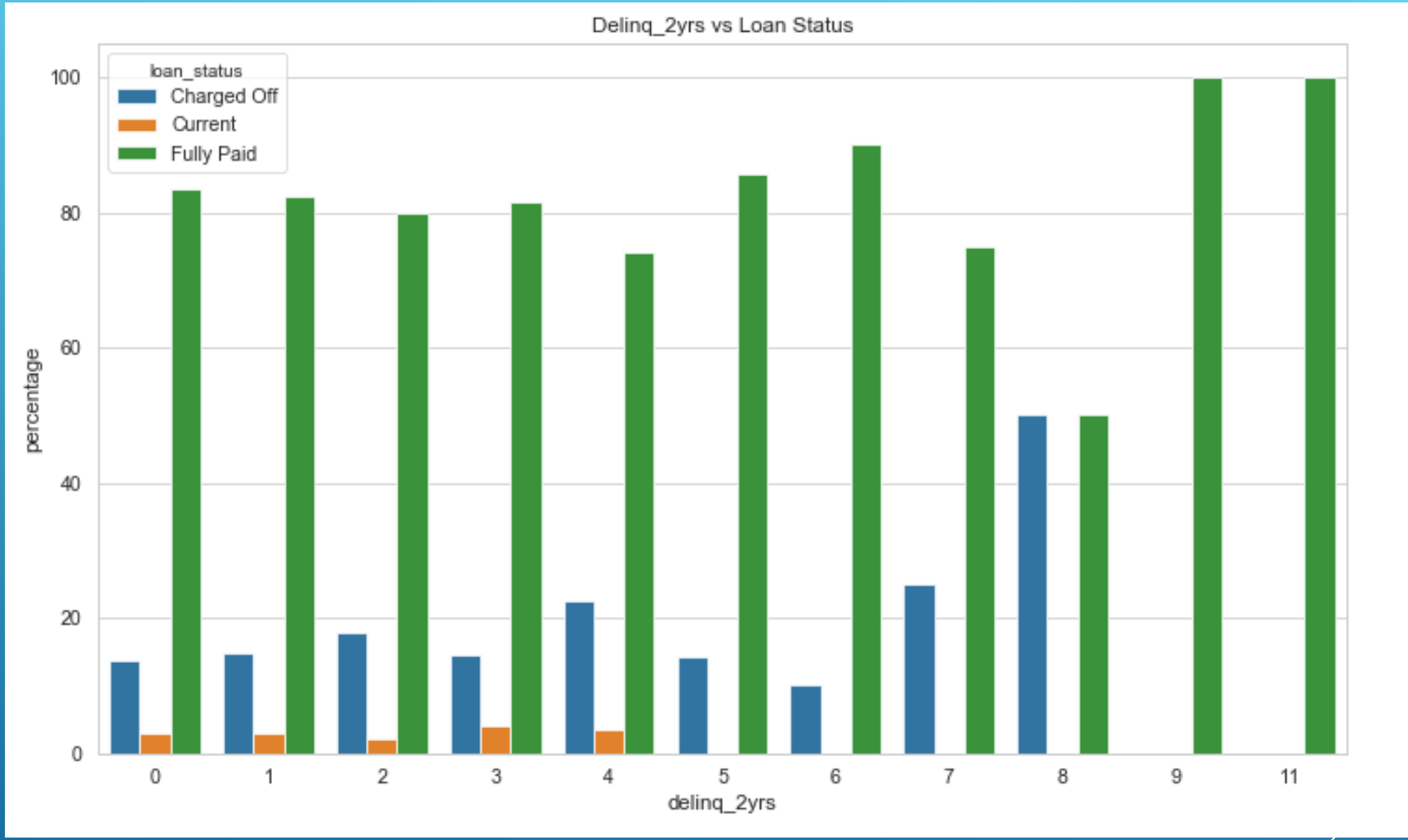
From above graph we can say that lower income group with 60 months term defaulted from paying the loans whereas 36 months term group with medium income paid their loans fully

Bivariate Analysis



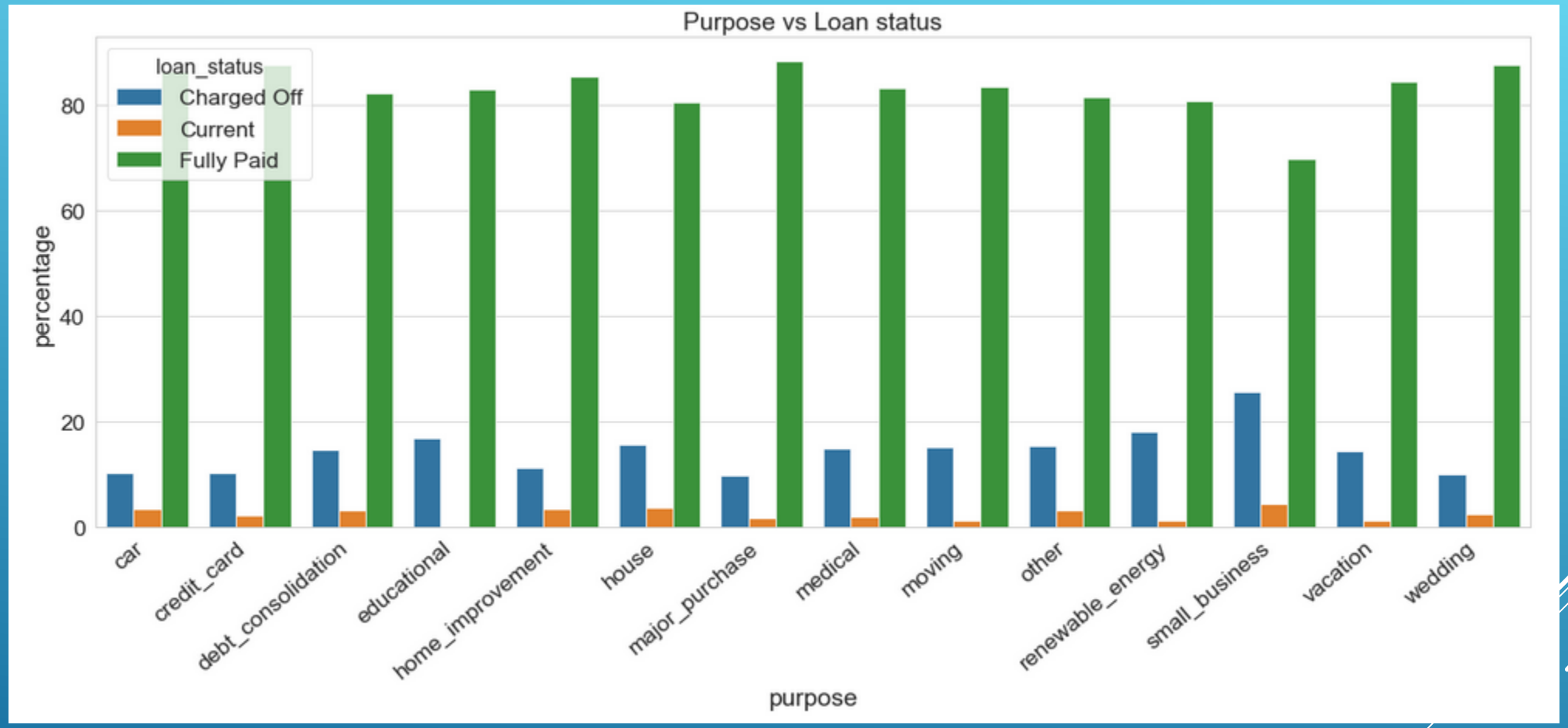
Borrowers with home_ownership as 'Other' are the most likely ones to default as seen from the above graph. We see that home_ownership => Other category showed a little high percentage of charged off loans

Bivariate Analysis



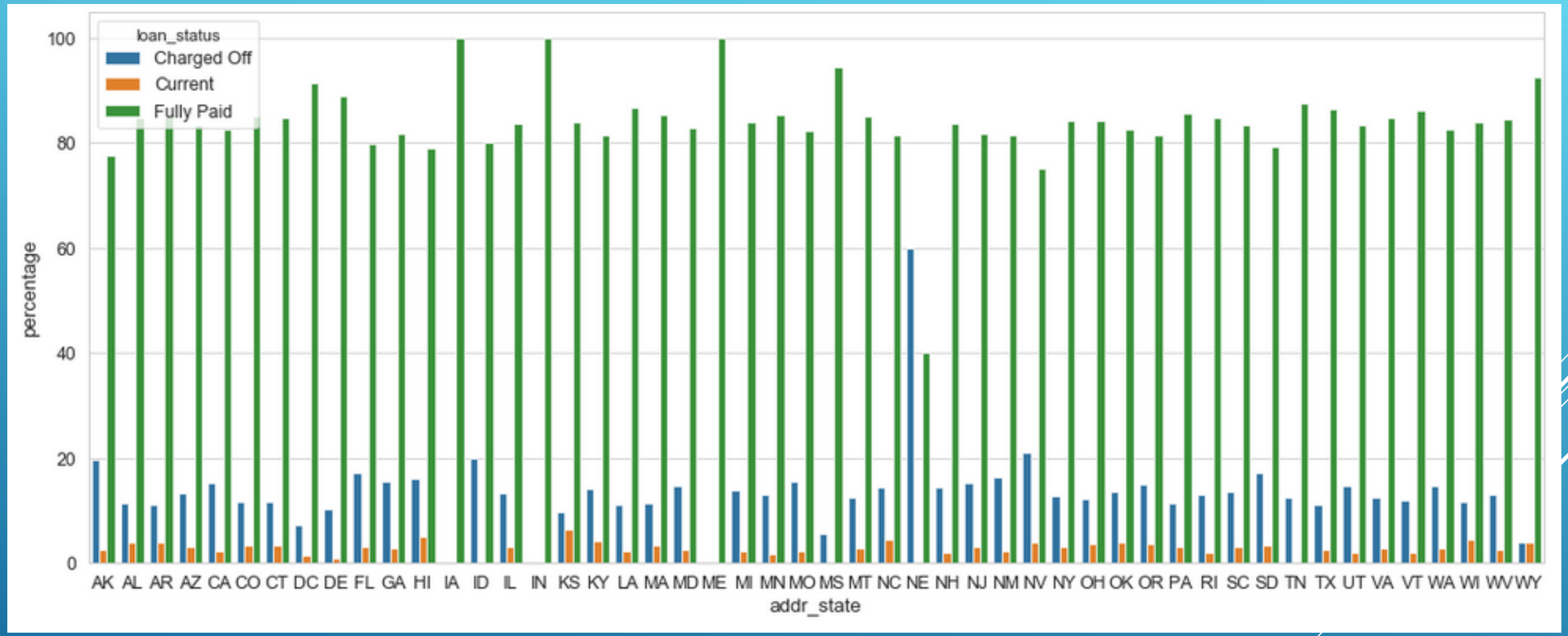
From above graph it is clear that as the incidences of delinquency in the borrower's credit file for the past 2 years increases there is no particular trend being followed. Thus we cannot conclude much from this attribute.

Bivariate Analysis



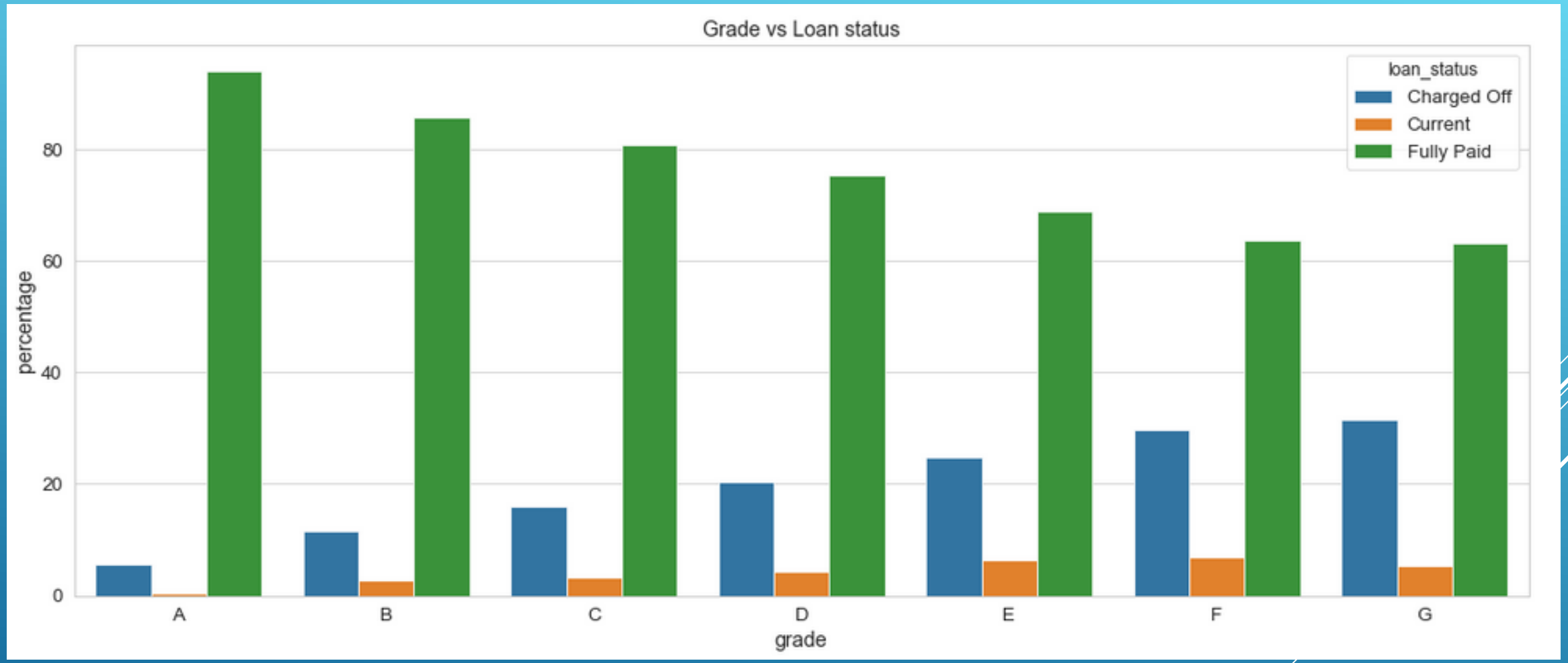
Percentage of charged off records is highest for the purpose of 'small business' and the least for 'major purchase'. This can be a deciding factor when sanctioning the loans.

Bivariate Analysis



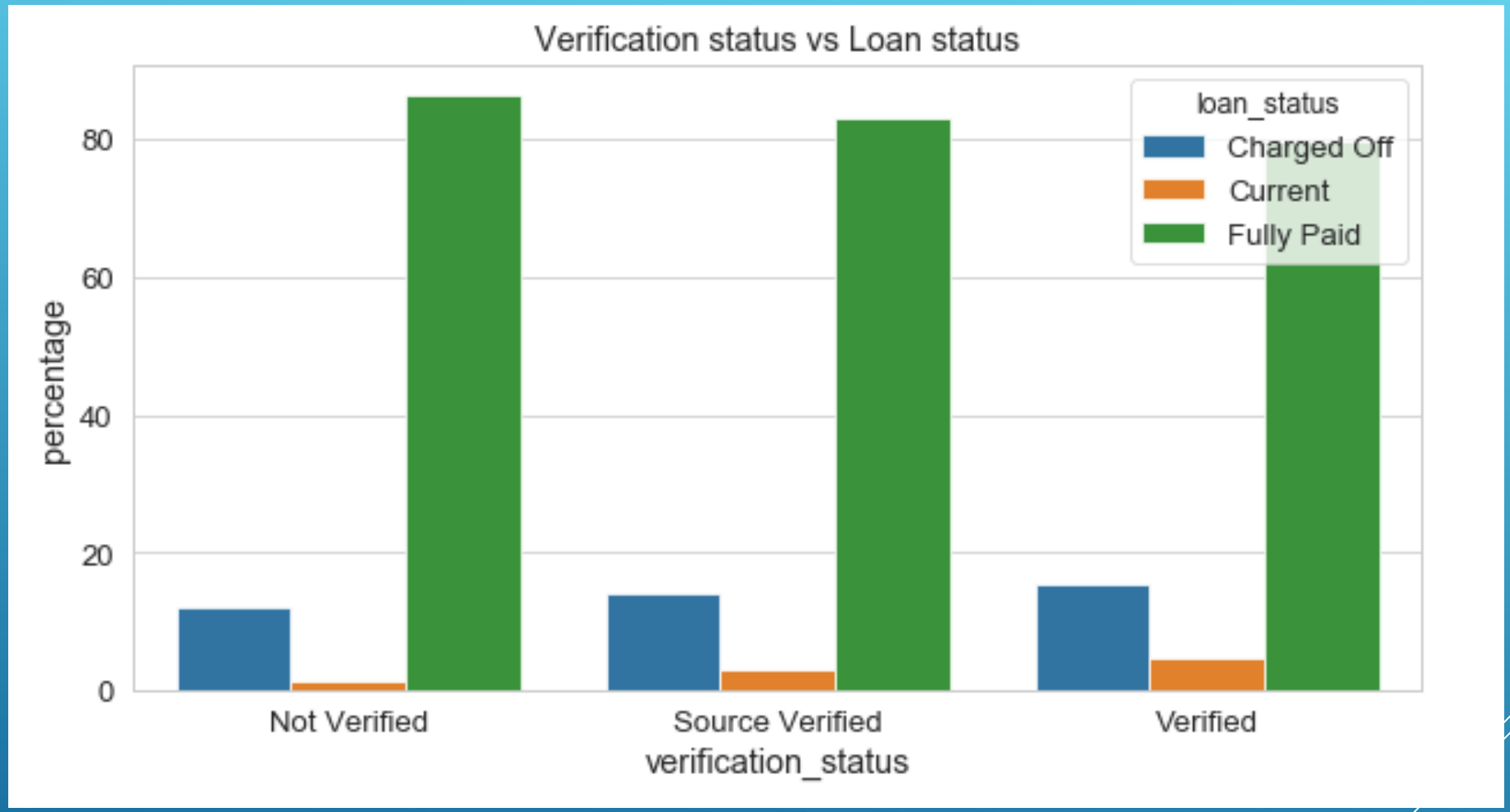
- From one of the above graphs (univariate for addr_state) it is clear that more applicants are from CA (California) state hence charged off records are more from that state.
- However when derived the percentage column based upon the loan_status from each state, NE (Nebraska) showed most charged off records. Hence it can be considered as risky state for loan applications.

Bivariate Analysis



As seen in the above plot, we see an incremental trend i.e. the percentage of charged off loans is increasing while going from A to G. Thus, we can rely on this attribute. The defaulting risk is highest for grade G and the lowest for A.

Bivariate Analysis



Verification status does not give much info about which borrower would default. As seen from the above graph, we have almost same percentages of defaulters among different Verification Statuses

SUGGESTIONS:

- It seems to be the case that the verification might not have been done correctly. Thus, the maximum defaulters under the verified category. So, the verification process must be strict enough to filter out the defaulters.
- The company should try to minimize the loan sanctions for loan Grade G applicants as it's the grade with highest number of defaulters.
- The company should minimize loan sanctions to the applicants from state NE i.e. Nebraska as it's the one with highest number of charged off loans.
- Loan applications for purpose 'Small Business' should be avoided.
- Loan applications from borrowers having home ownership status as 'others' should be strictly verified for other factors as well.
- Lower income groups that too for 60 months term should not be sanctioned loans easily.
- Borrowers with high interest rate with 60 months term should be avoided. The company should lower the interest rate with 60 months term to reduce the defaulters.


THANK YOU