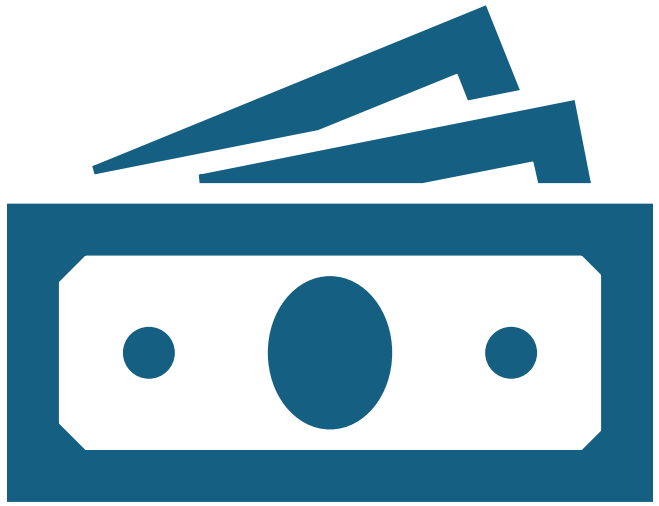


LENDING CLUB CASE STUDY

(RISK ANALYSIS)

Problem Statement

You work for a **consumer finance company** which specialises in lending various types of loans to urban customers.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's 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

The data given contains information about past loan applicants and whether they 'defaulted' or not.

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Business Objectives

Lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). 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’.

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 utilise this knowledge for its portfolio and risk assessment.

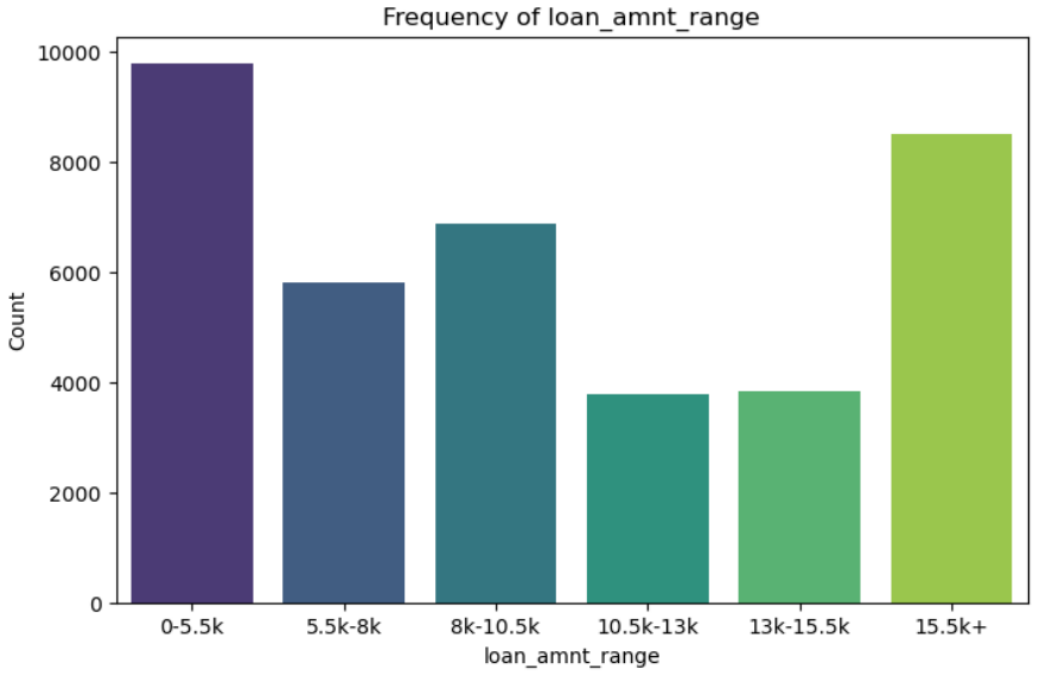
APPROACH FOR ANALYSIS

- Understanding the Problem statement and overviewing the data in an Excel file.
- Creating an ipynb notebook and loading the data as the pandas' data frame
- Data Cleaning:
    1. Removing columns with all null values(e.g. 'dti_joint','verification_status_joint', 'total_cur_bal', 'max_bal_bc', etc)
    2. Dropping the columns with unique values in the whole column(e.g. 'pymnt_plan', 'initial_list_status', etc)
    3. Dropping irrelevant/descriptive columns('member_id', 'id', 'url', etc)
    4. Removed unnecessary text from columns and converted those to required ones(i.e. object to int, object to float)
    5. Converted the columns with Dates as values to DateTime datatype.
    6. Imputed data where the values were missing.(e.g. 'emp_length')
    7. We have removed the rows with loan status as "Current" to get defaulter insights from "Fully Paid" and "Charged Off" applicants.
- Derived some columns from existing ones to get some clear insights.
- Added some columns to get bucketed values from the continuous range of values
- Performed Univariate Analysis to check the performance of each column wrt Frequency.
- Performed Bivariate Analysis to get insights into the interactions between two columns.

UNIVARIATE ANALYSIS

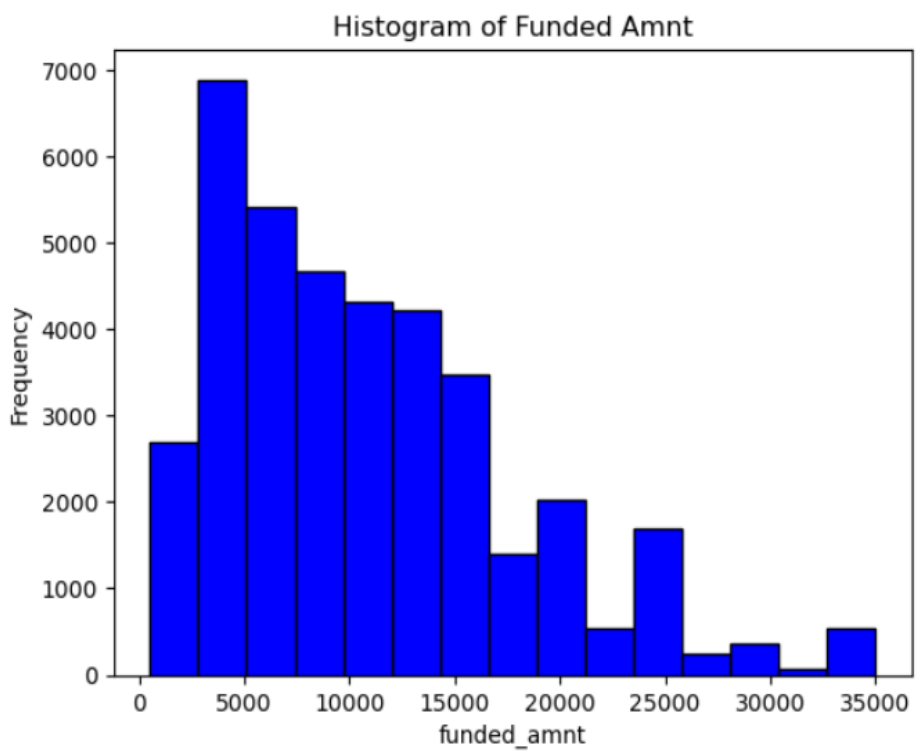
RANGE OF LOAN AMOUNT

Based on the Bar Plot we can say that most of the loan was taken for the ranges from 0 to 5000



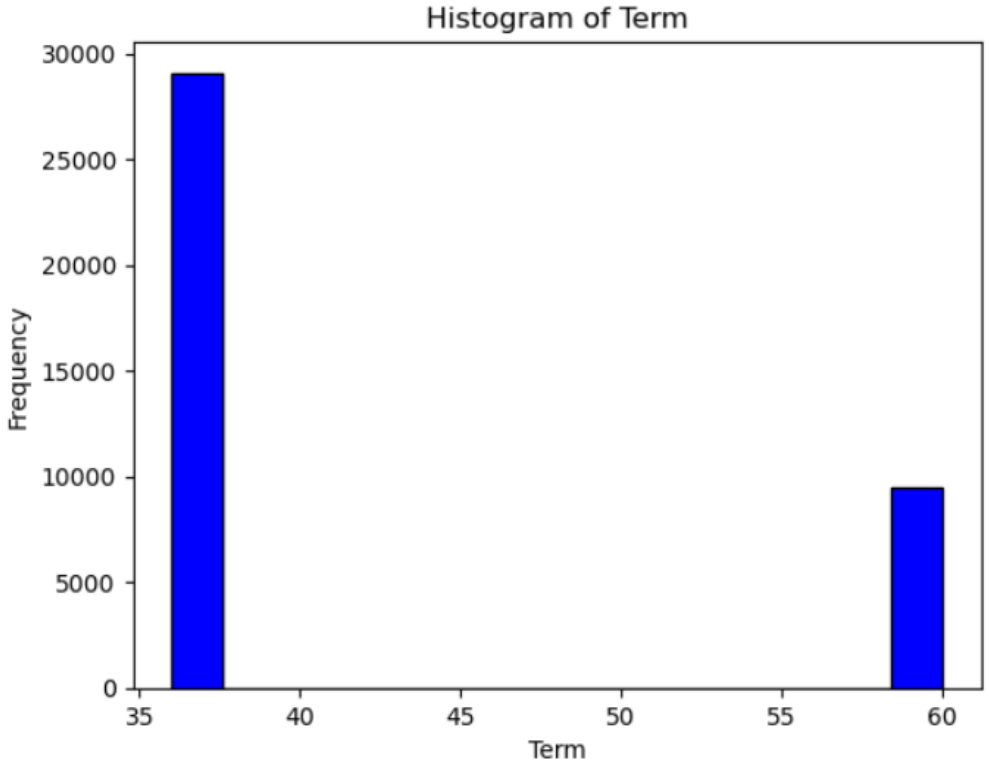
FUNDED AMOUNT

From the histogram we can see that most of the funded amount ranges from 0-5000



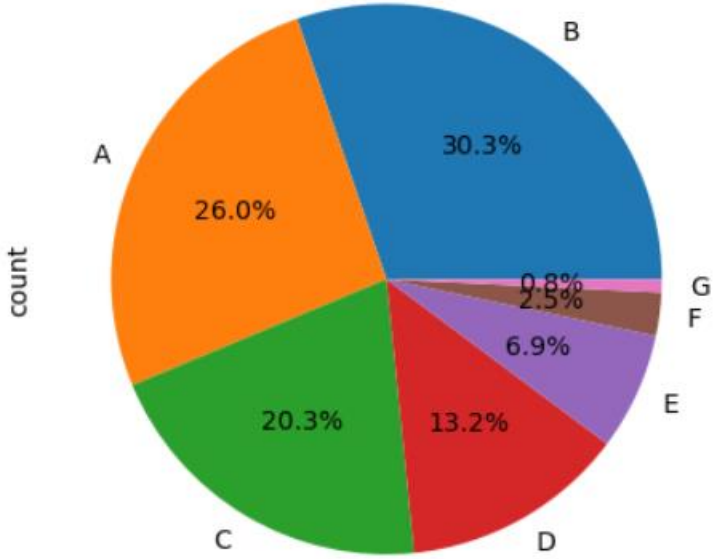
TERM

The histogram shows us that most people are taking a loan for a tenure of 36 months i.e. 2 years.



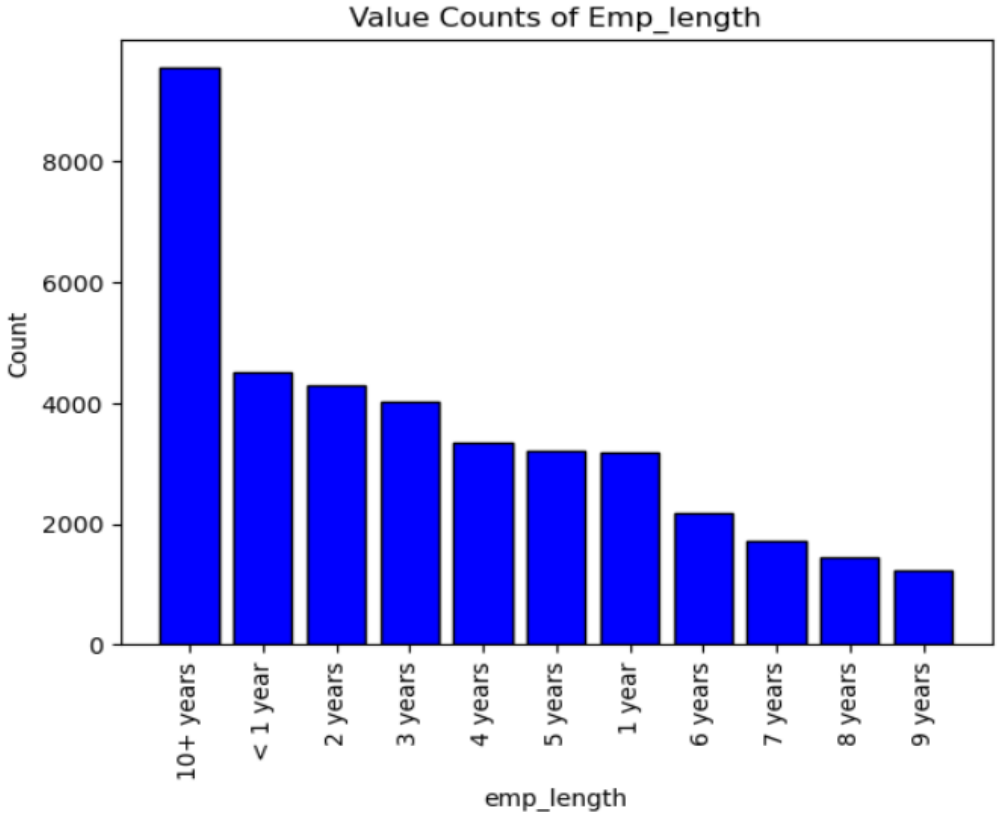
GRADE

We can see that a maximum i.e. 30.3% of borrowers are of B grade



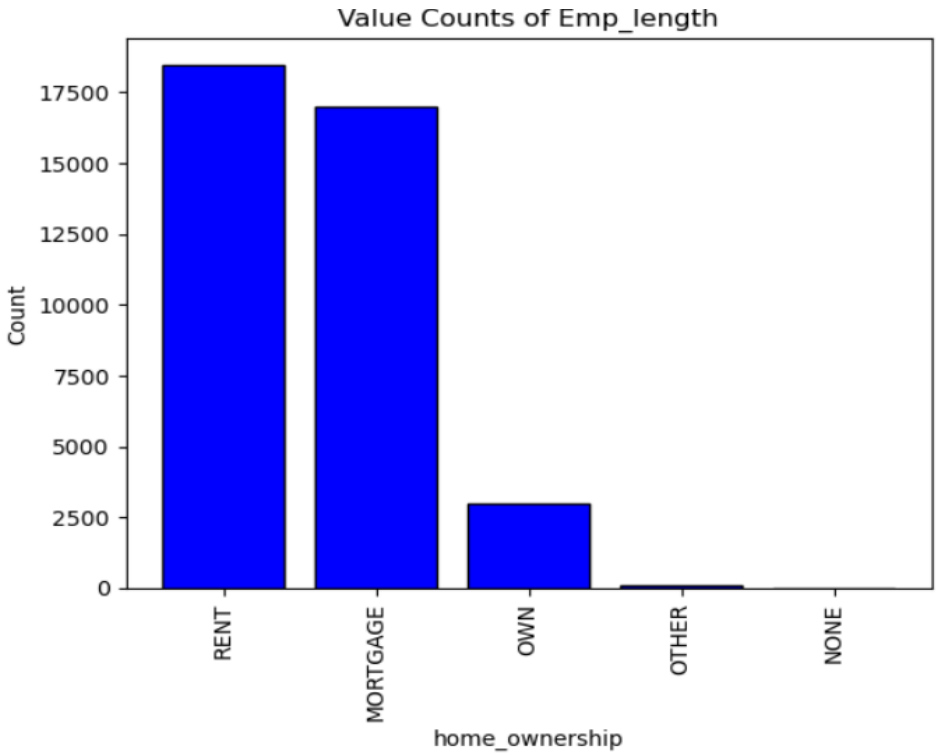
EMPLOYMENT LENGTH

Total years of employment has an impact on number of people taking loan such that more the number of employment period more is the chance to take the loan and the chance of taking a loan decreases with a year of experience. Loans are mostly taken by the employees with 10+ years of experience



HOMEOWNERSHIP

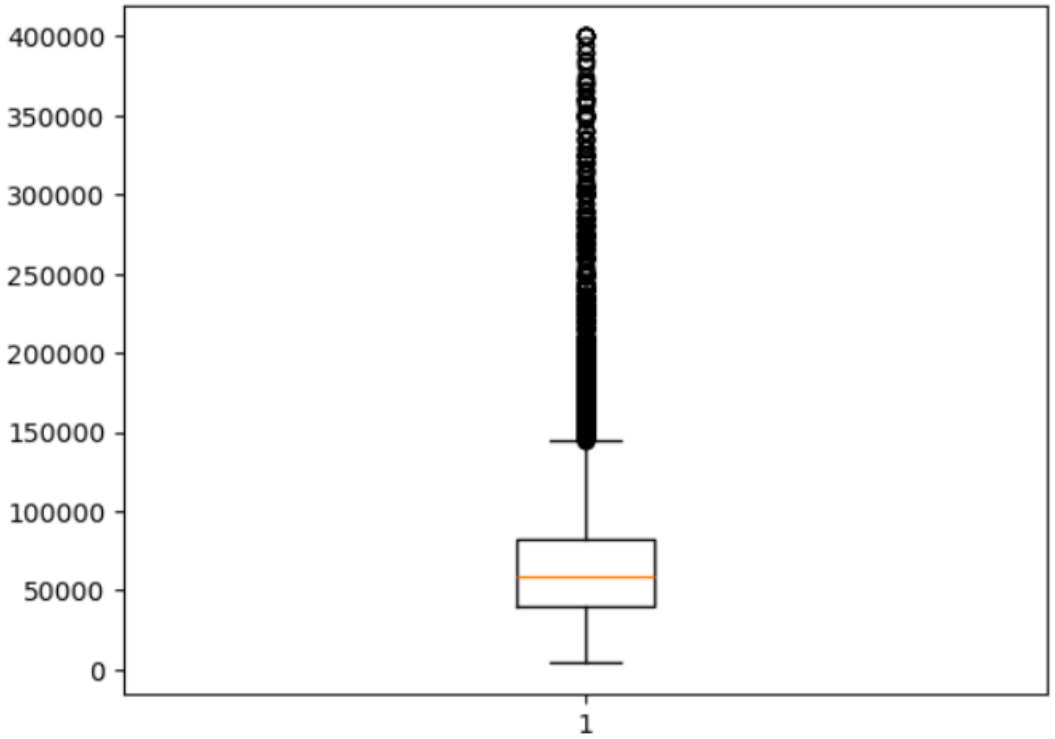
Majority of people taking loans prefer to live on rent



ANNUAL INCOME

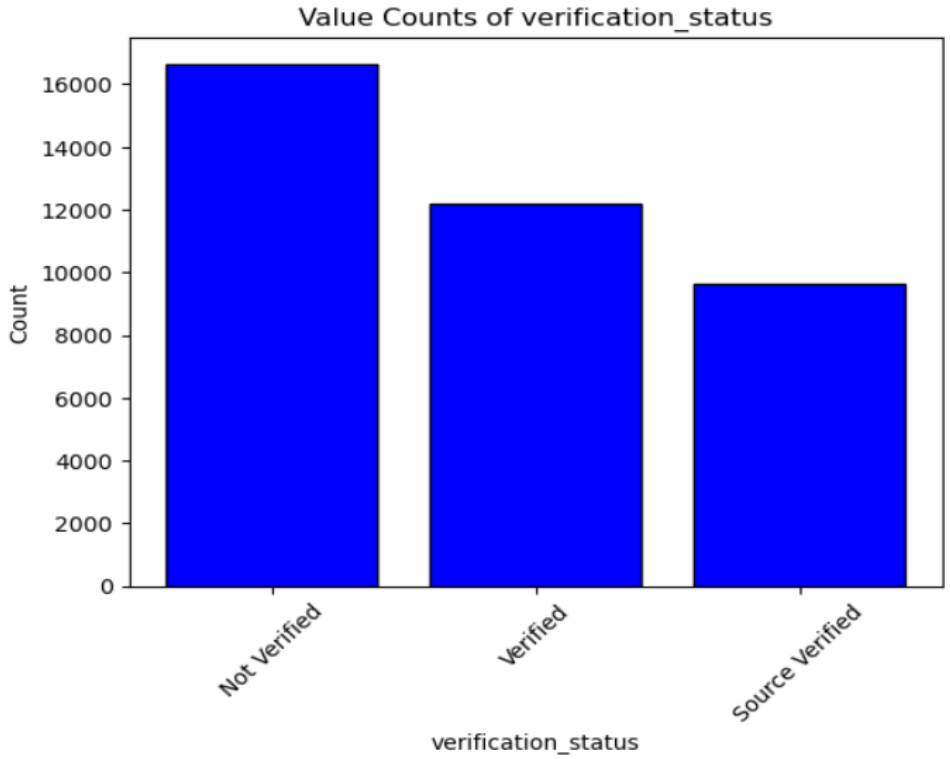
While doing univariate analysis for annual income we found **outliers** which we cleaned and the resultant box plot from the cleaned data is mentioned below.

From this we can deduce that most of the people have salary in range 40k to 80k



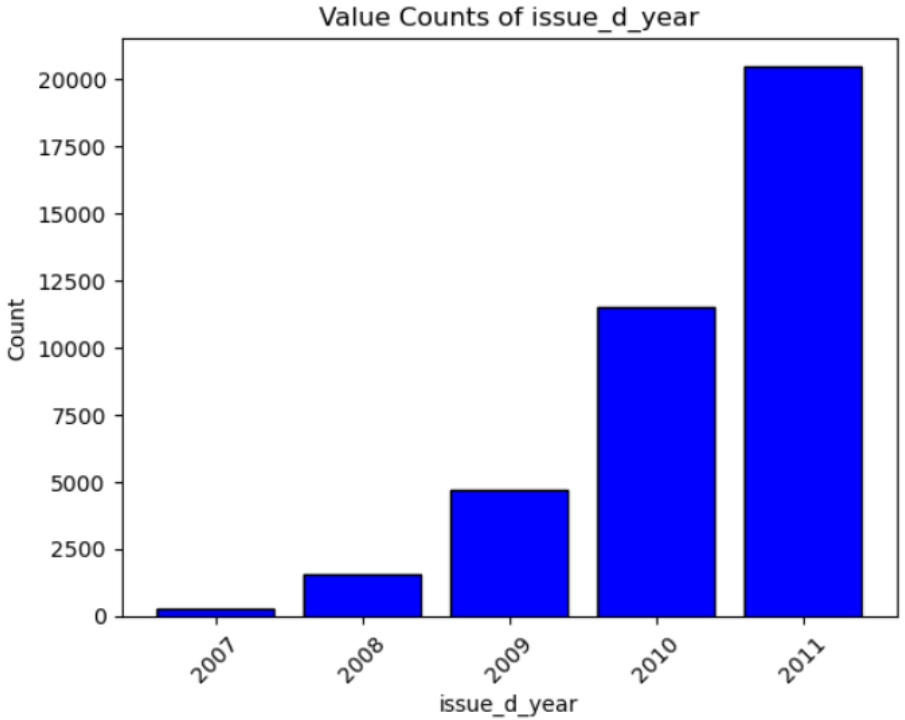
VERIFICATION STATUS

Very few applicants have the income source verified.



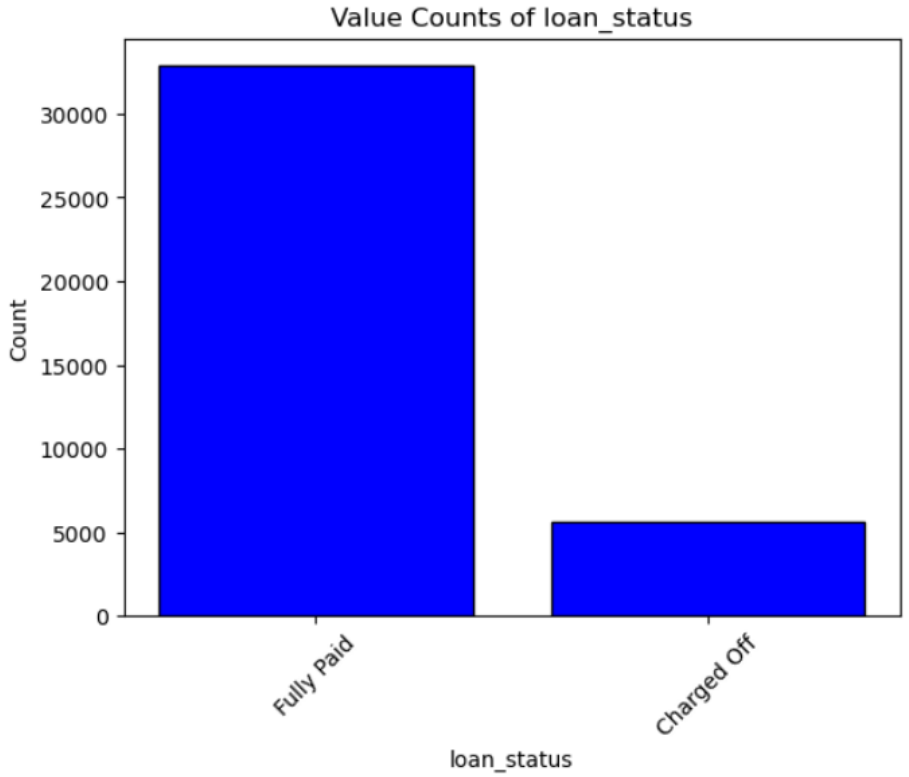
LOAN ISSUED YEAR

Number of loans taken peaked in the year 2011.



STATUS OF THE LOAN

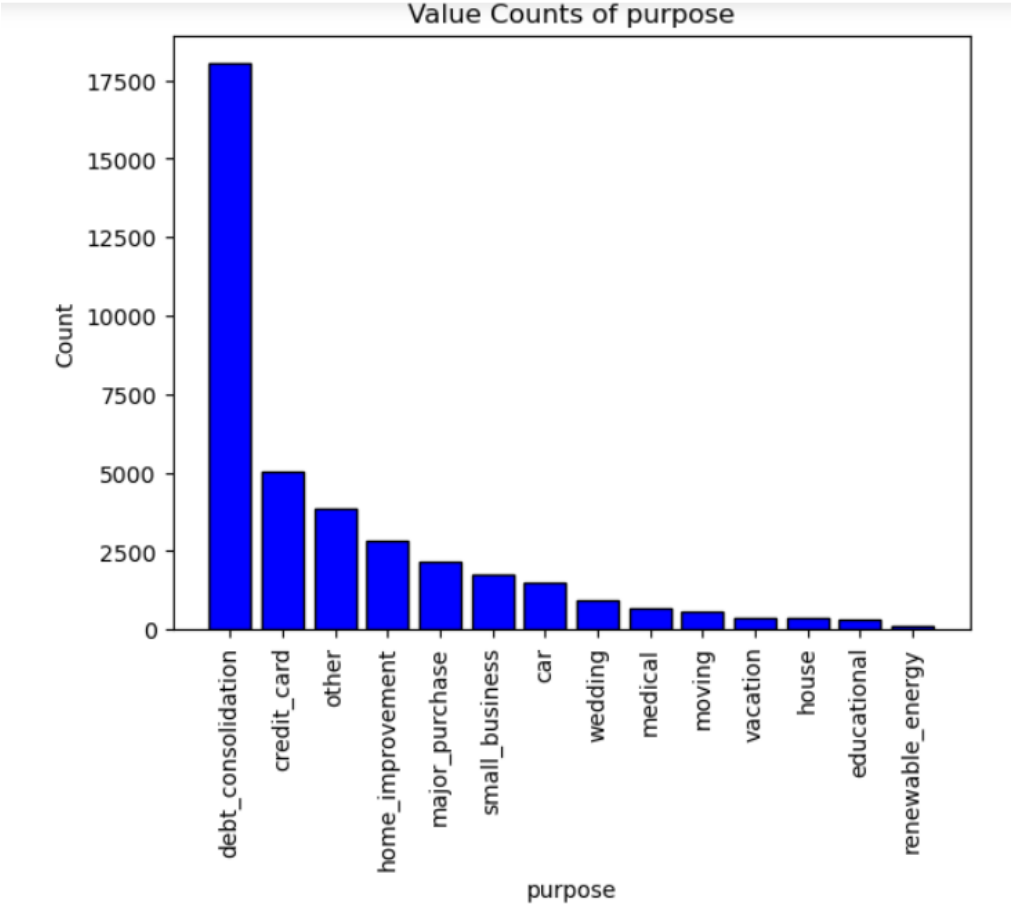
It was observed that majority of the applicants who have taken the loan were able to successfully repay it. There were around 5k applicants who were Charged Off.



PURPOSE OF THE LOAN

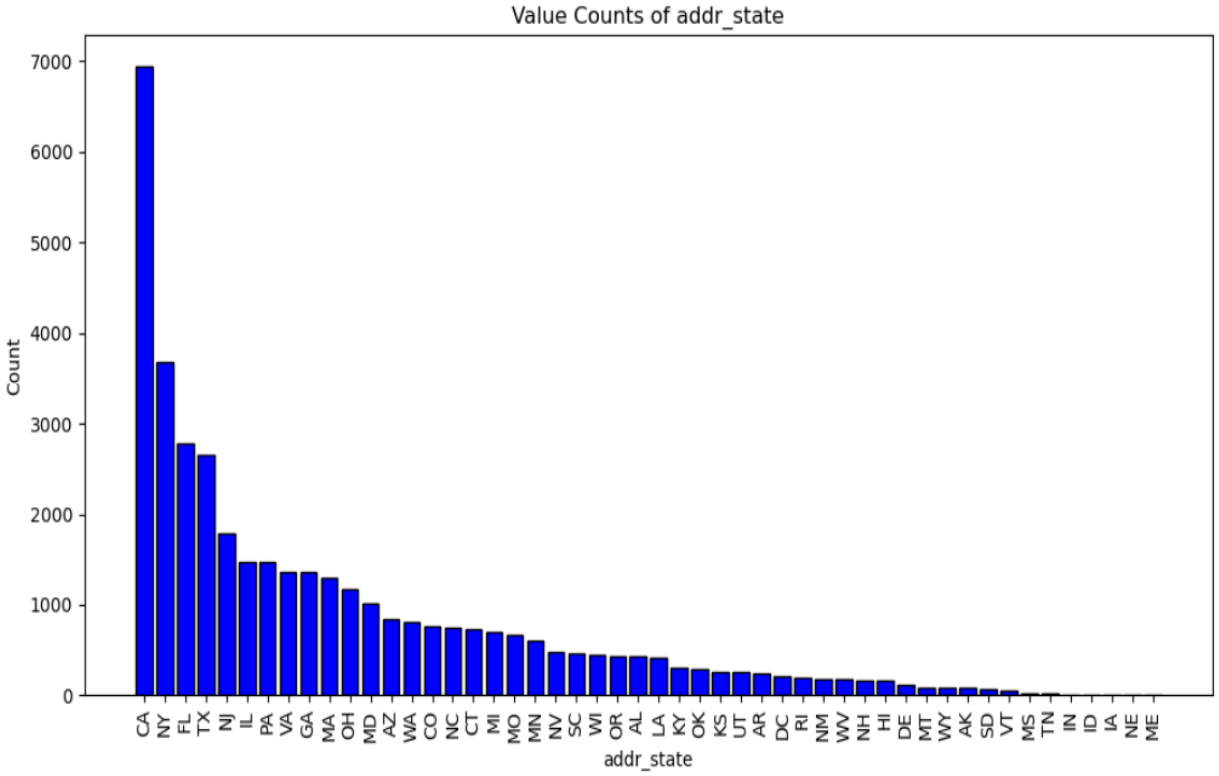
debt_consolidation was the popular reason of the loan amongst the applicants.

They must have taken the loan to take another loan



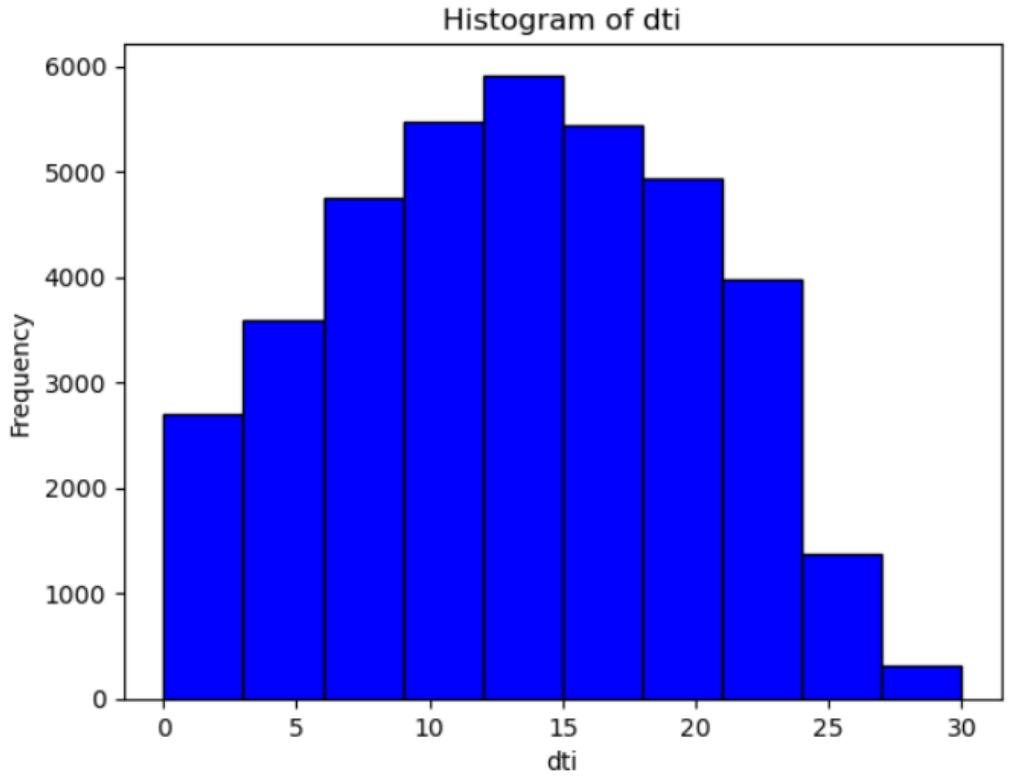
APPLICANT ADDRESS STATE

Maximum number of applicants who have taken the loan are from address state CA



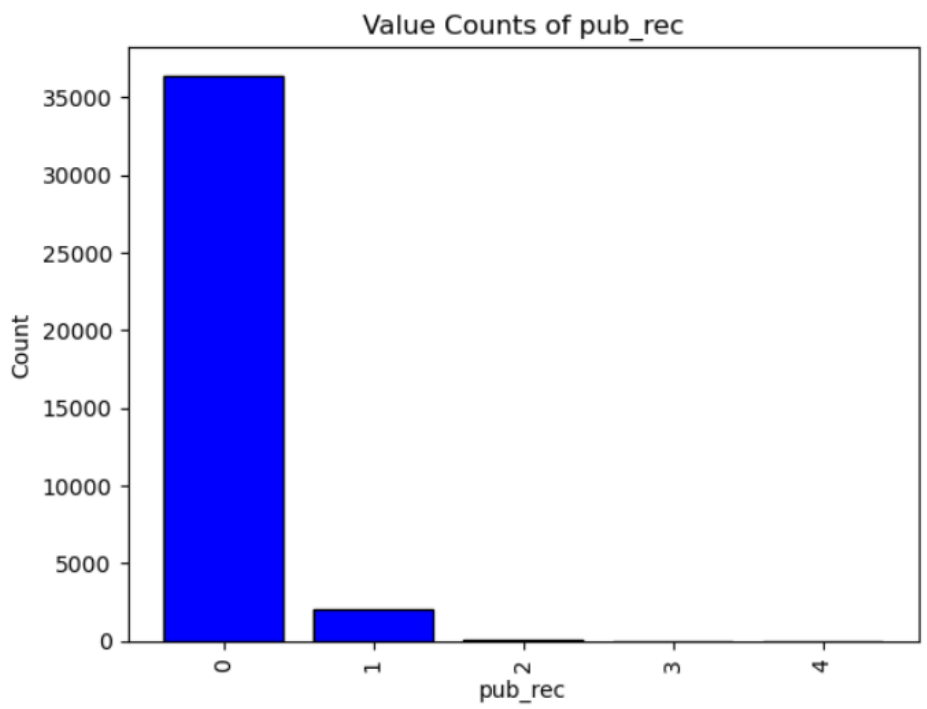
DEBT TO INCOME RATIO

Applicants with debt to income ratio between 12% to 15% are most likely to take the loan



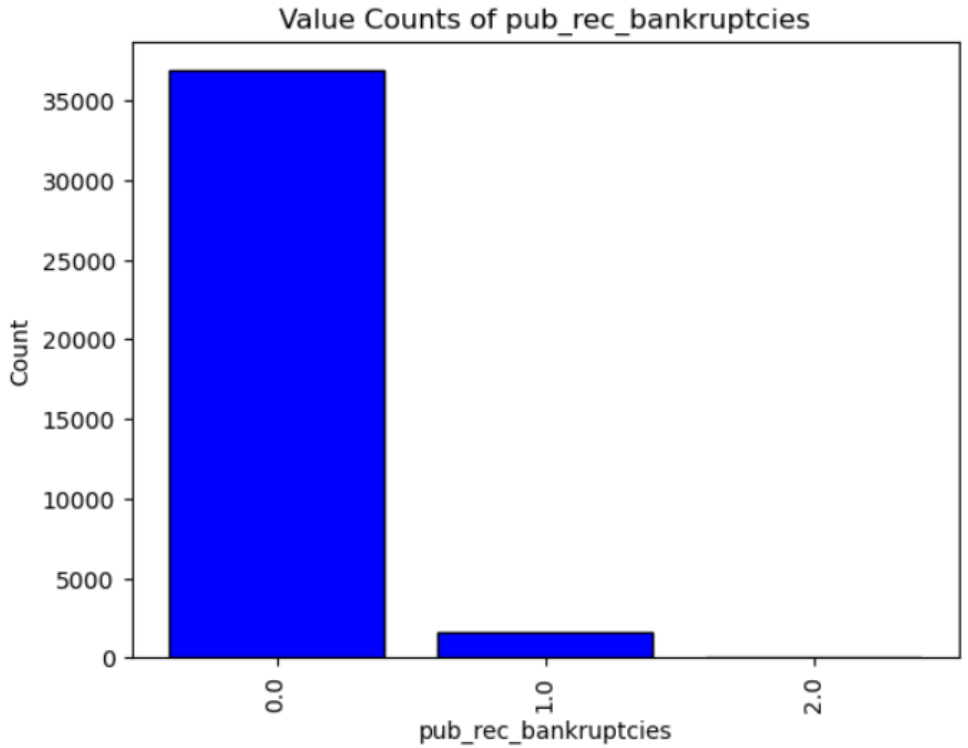
PUBLIC RECORDS

Most of the applicants have no public records against them



PUBLIC RECORDS BANKRUPTCIES

Most of the applicants also do not have bankruptcy records against them

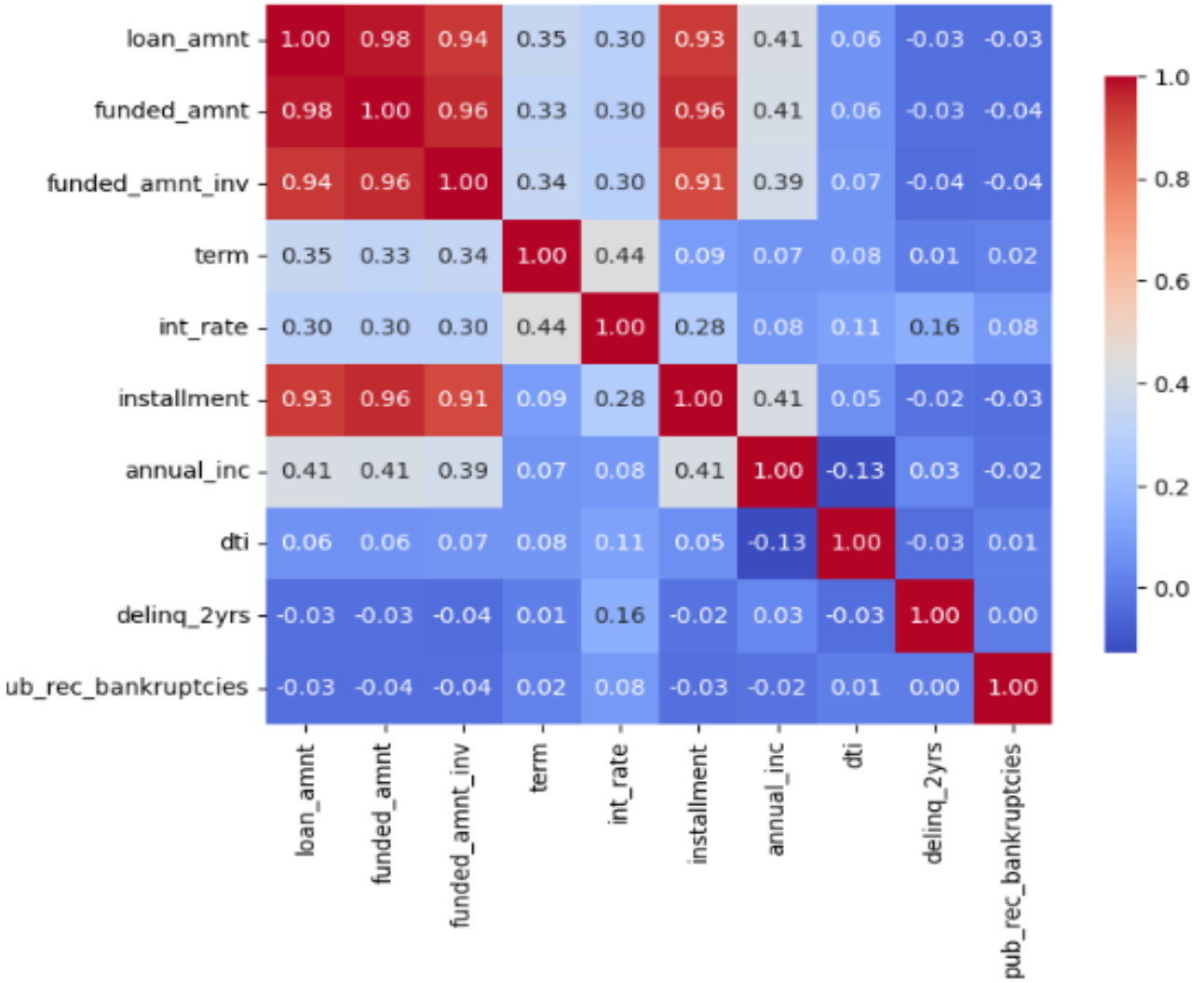


BIVARIATE ANALYSIS

Plotted a Correlation matrix in order to find the columns that are highly correlated.

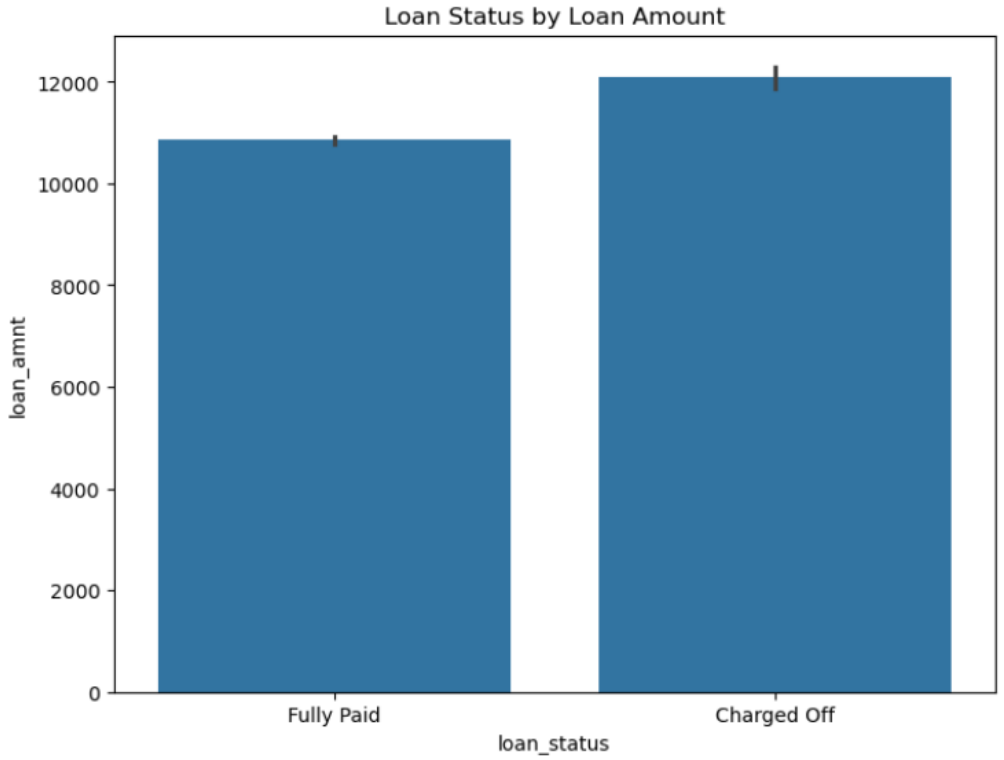
We found the following fields highly correlated :

1. funded_amnt & funded_amnt_inv
2. funded_amnt & loan_amnt
3. funded_amnt_inv & loan_amnt
4. Installment & funded_amnt
5. Installment & funded_amnt _inv
6. Installment & loan_amnt



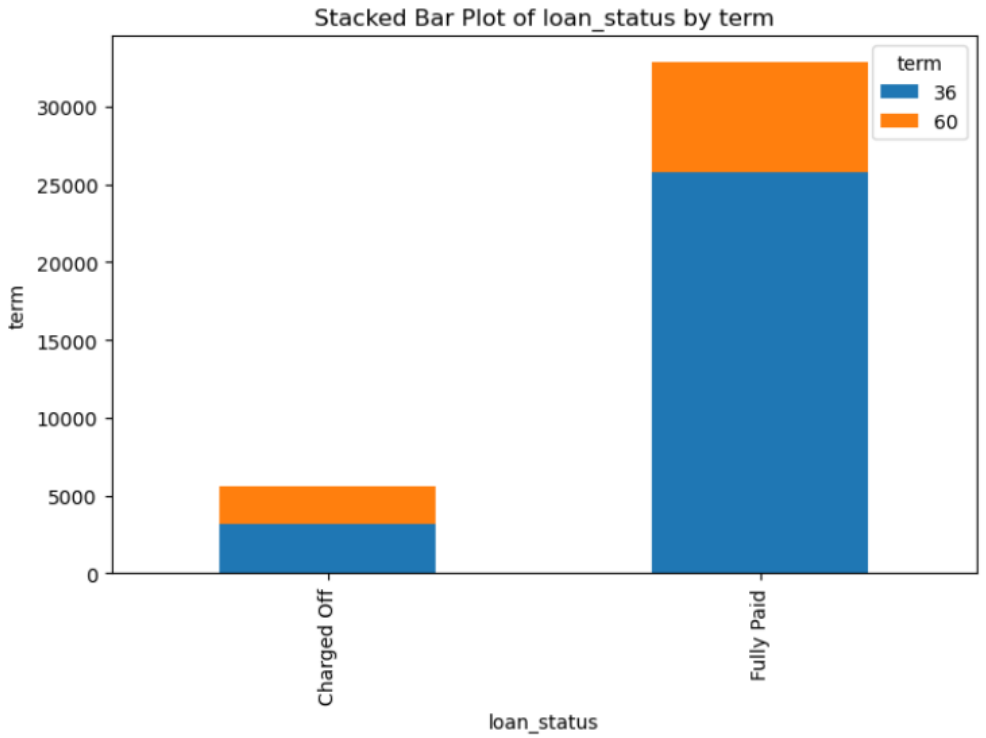
LOAN STATUS / LOAN AMOUNT

The higher the loan higher are the chances of being charged off

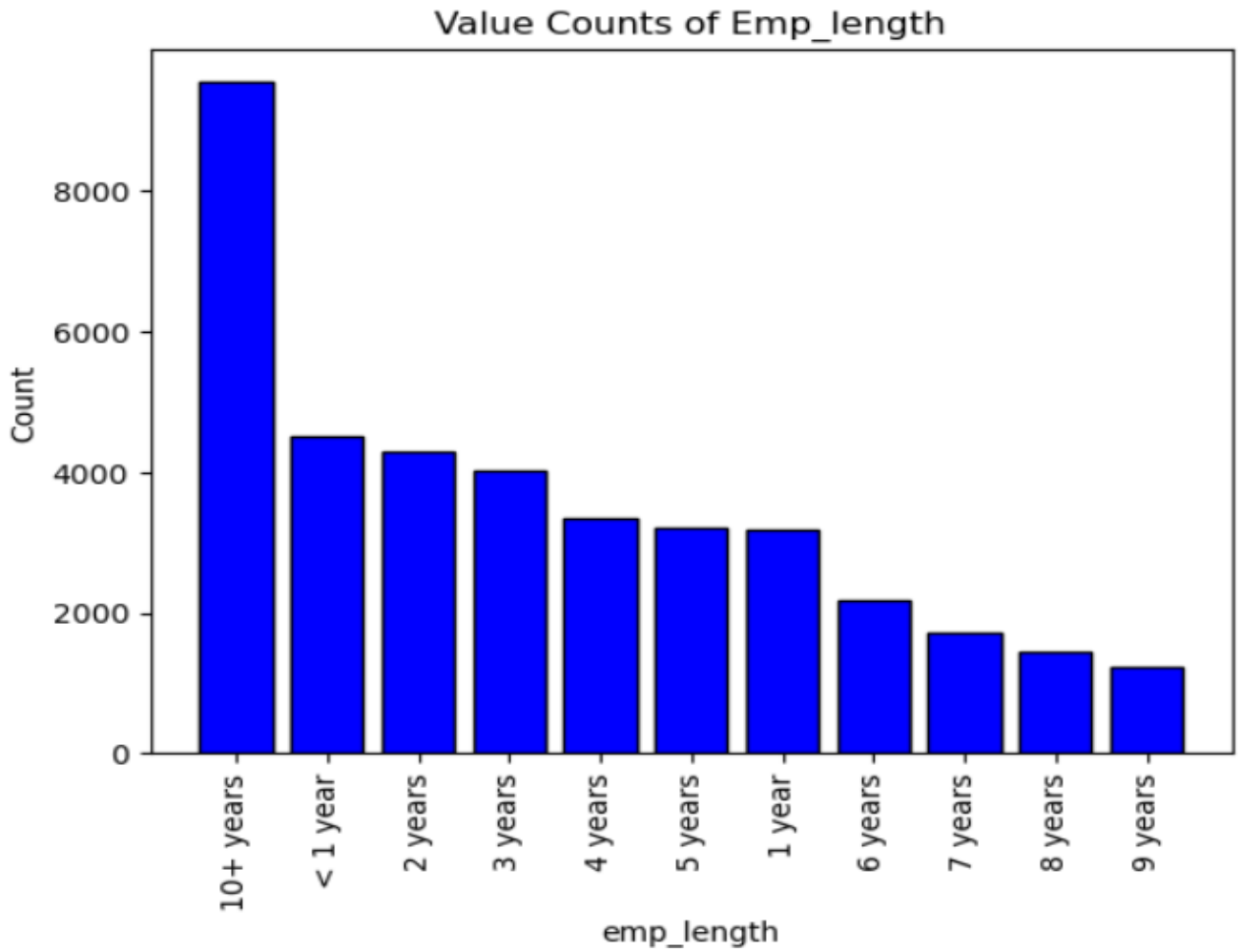


LOAN STATUS / TERM

The majority of borrowers have fully repaid their loans, with a preference for 2-year terms.

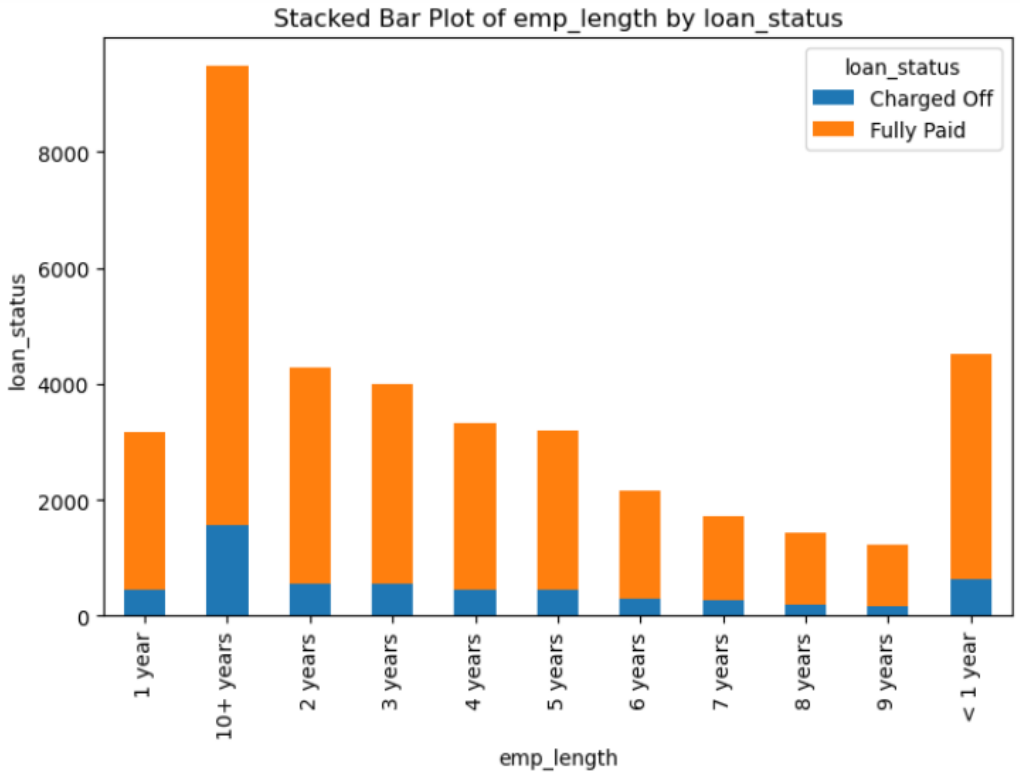


CHARGED OFF RATIO



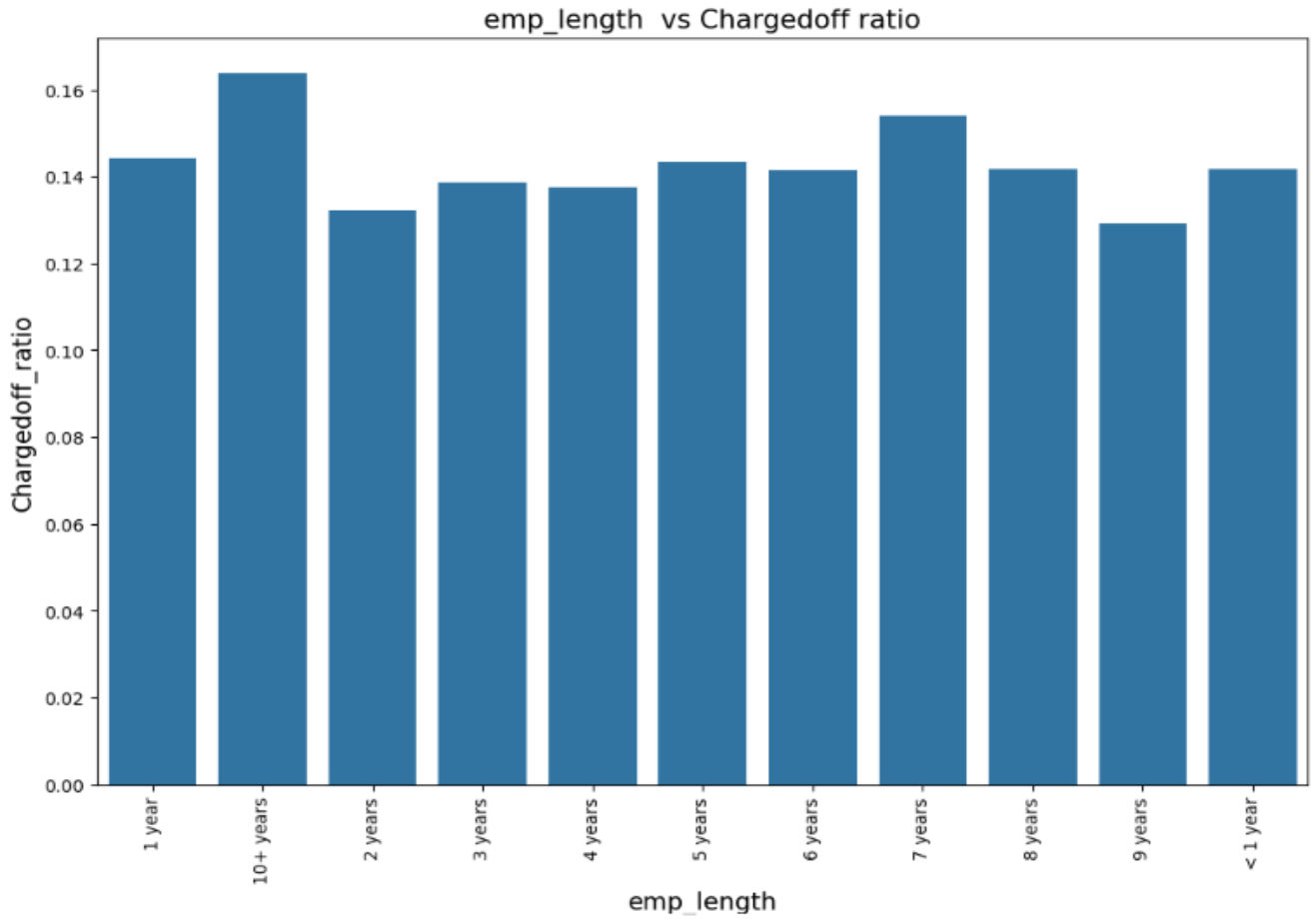
- While performing the analysis it was observed that some columns were having considerably high frequency for a particular value.
- Eg.frequency for "10+ years" in emp_length column was way higher than the other years.

CHARGED OFF RATIO



- Thus, when we performed bivariate analysis involving these columns, we could not find any distinguished data.
- Eg. If we plot emp_length against loan_status, since the frequency for "10+ years" emp_length is itself higher, naturally the frequency for applicants who has been "Charged Off" as well as “Fully Paid” will be higher for "10+ years" emp_length.

CHARGED OFF RATIO



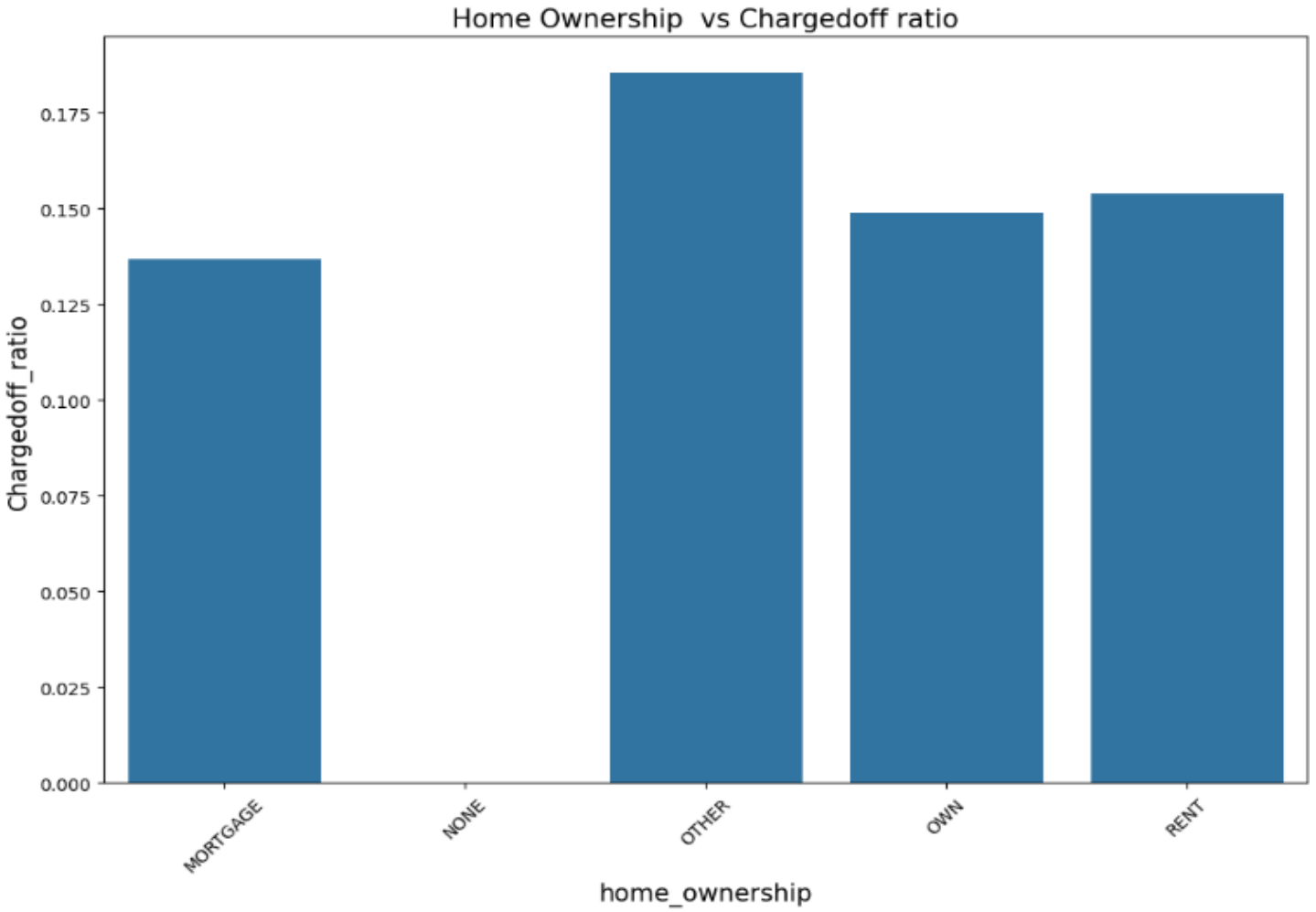
- In order to avoid such scenarios, we have used "Charged Off Ratio" instead of "Charged Off" data
- This "Charged Off Ratio" corresponds to : ("Charged Off" Applicants for a particular category) / ("Total" Applicants for a particular category)
- Charged Off Ratio focuses on applicants who failed to repay the loan on time and thus this data will give us clear insights on defaulters.
- In this representation we can see that applicants with **10+ years** of experience are more likely to get charged off

CHARGED OFF RATIO

Thus we have used the “Charged Off Ratio” for bivariate analysis of columns whose frequency for a particular value was seen to be considerably higher. These columns were

FIELD	HIGHEST FREQUENCY
Employment Length	10+ years
Home Ownership	Rent
Grade	B
Verification Status	Not Verified
Loan Issued Year	2011
Address	CA
Purpose	debt consolidation

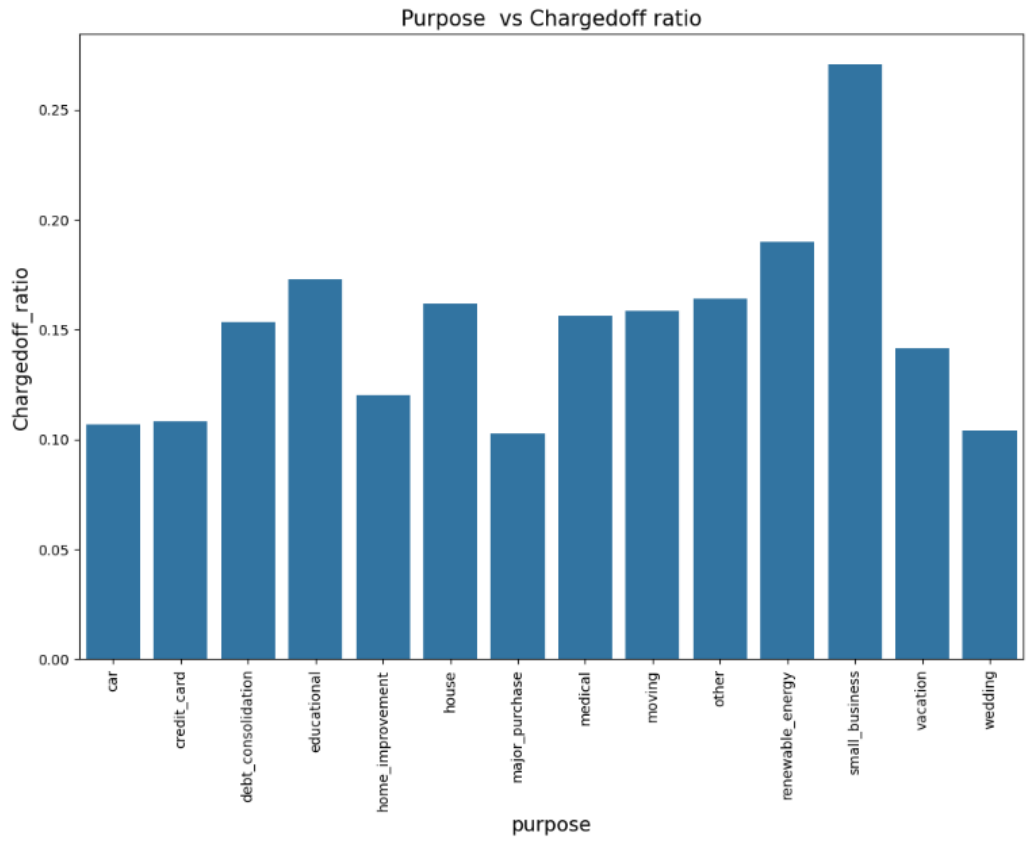
HOME OWNERSHIP / CHARGED OFF DATA



Since this representation focuses on Charged off applicants, we can deduce that those applicants categorized as "**Other**" in homeownership are at a higher risk of charge-offs.

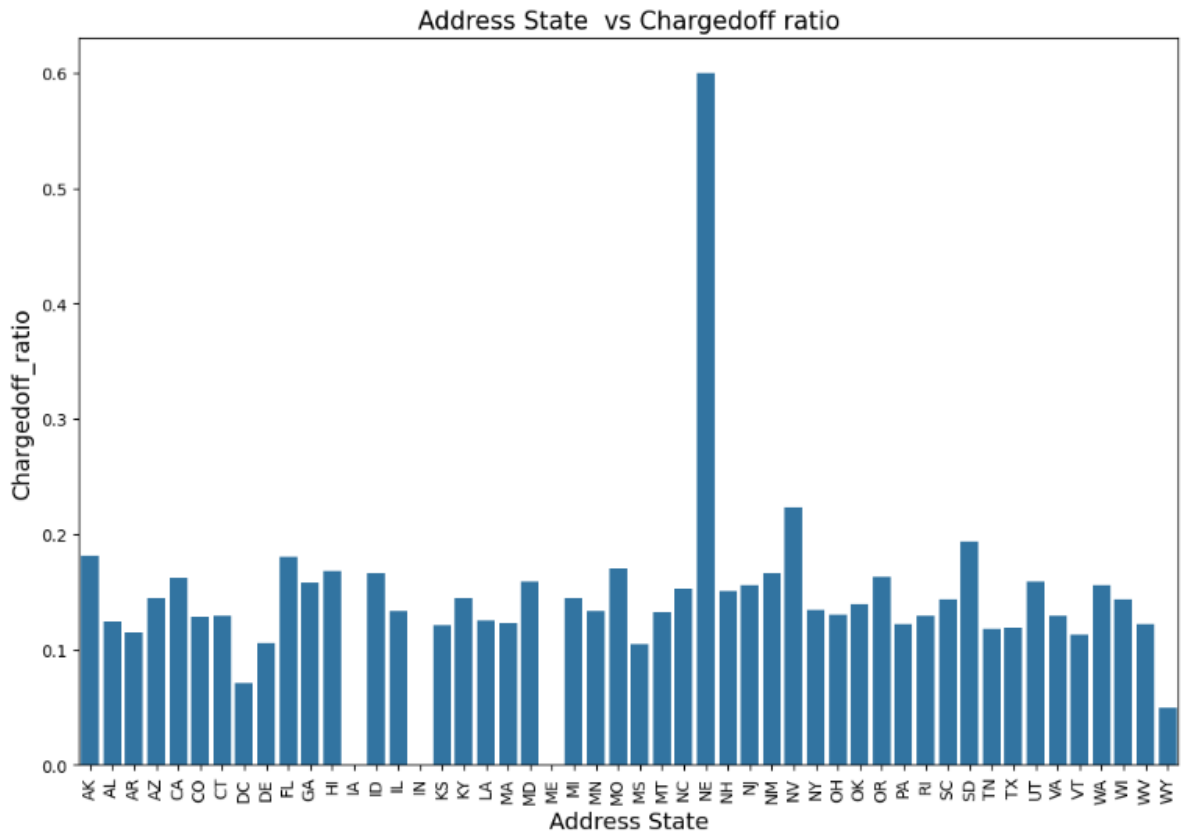
PURPOSE / CHARGED OFF DATA

Applicants seeking loans for **small_business** are **most** likely to be charged off



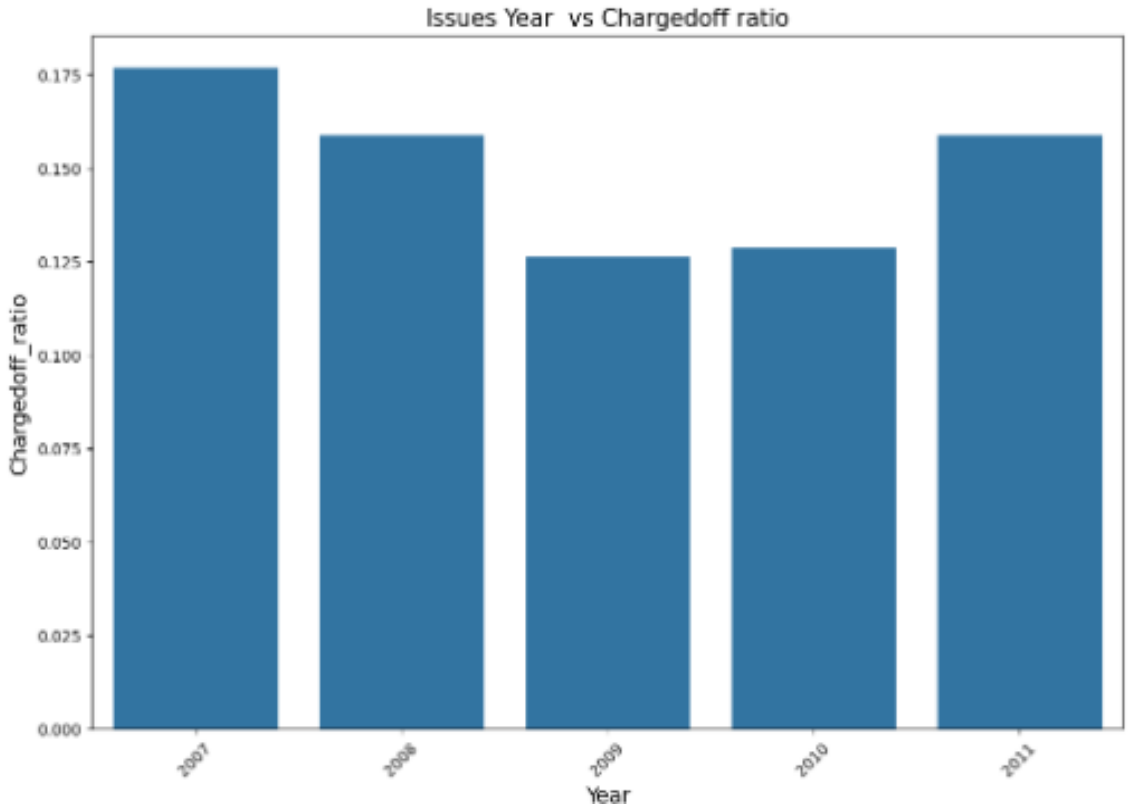
ADDRESS / CHARGED OFF DATA

NE State records the **highest** number of charged-off applicants whereas states **IL, IN, and ME** have seen **no charged-off** applicants since 2007 to 2011.



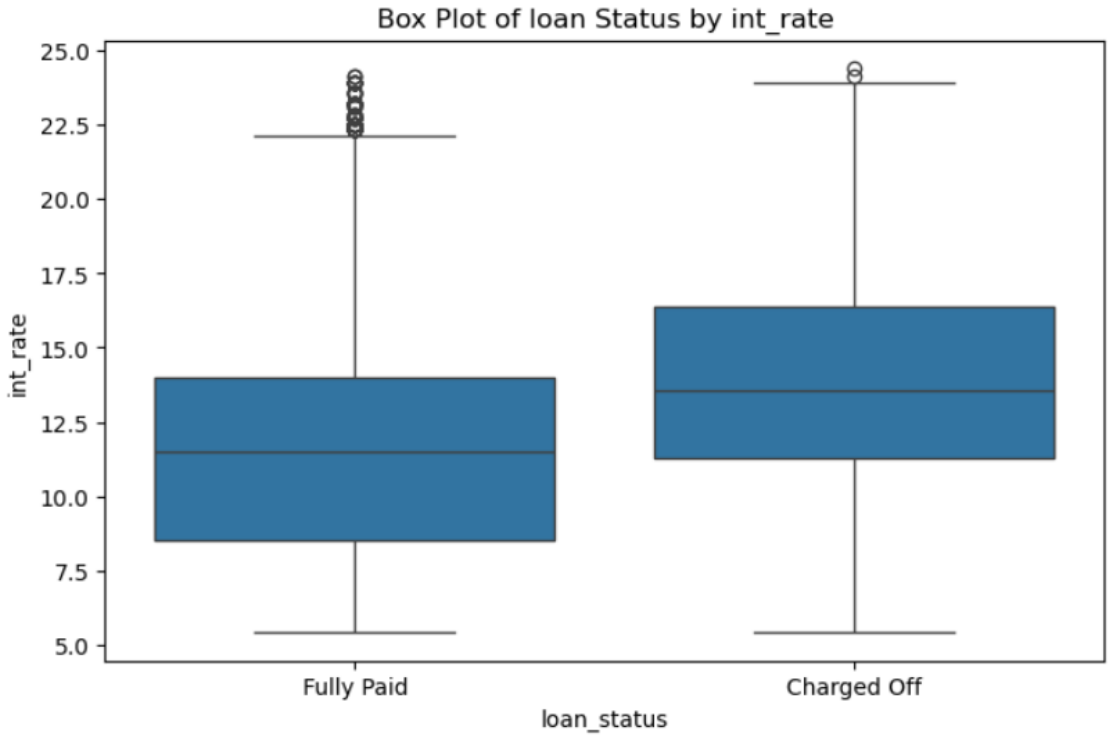
LOAN ISSUED YEAR / CHARGED OFF RATIO

The **highest** proportion of Charged Off Applicants **peaked** in the year **2007**



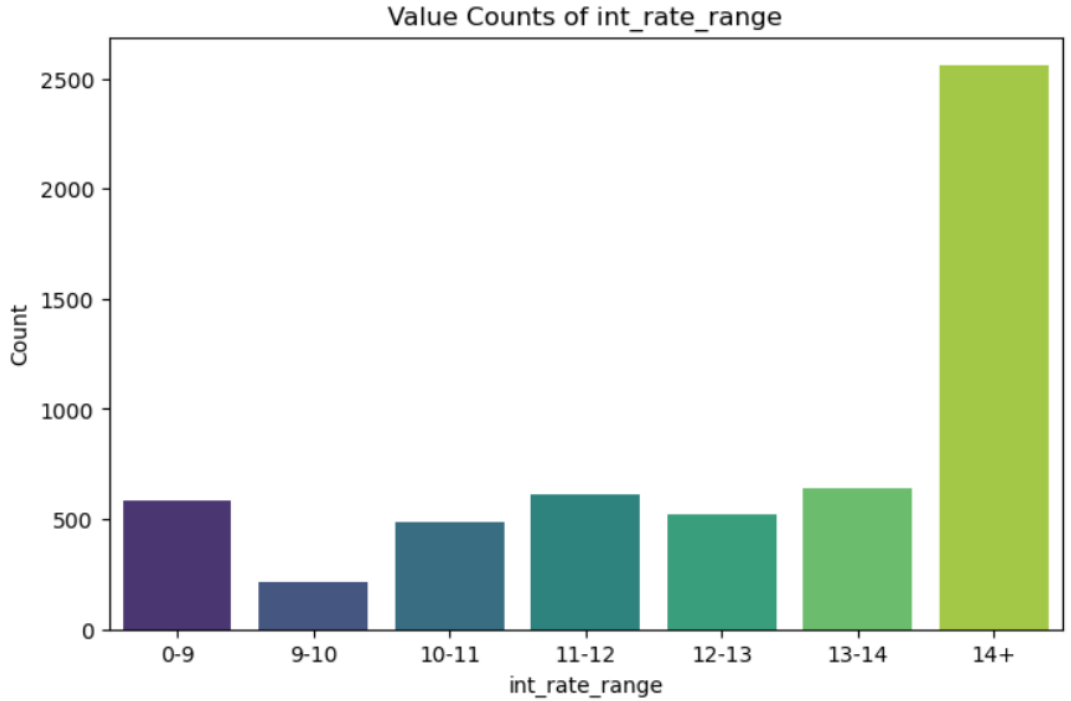
LOAN STATUS / INTEREST RATE

People who were able to pay the loan have **lesser interest** rates than the ones who are charged off



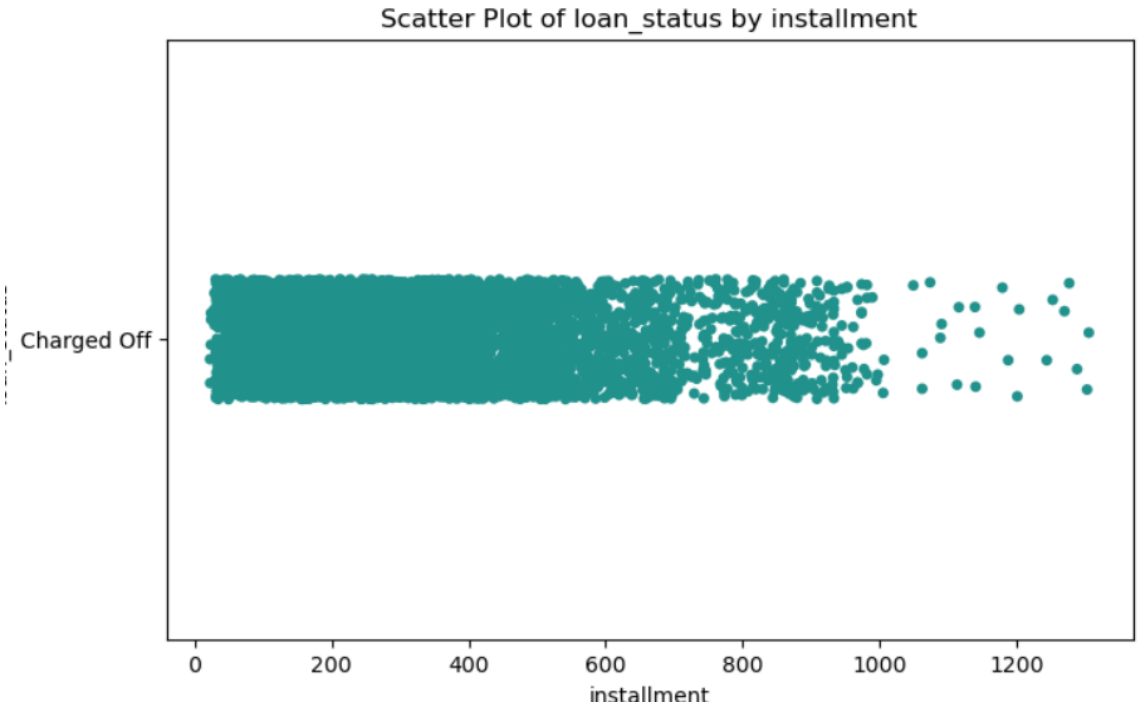
INTEREST RATE RANGE/ CHARGED OFF DATA

Applicants having interest rates of **14% and higher** are most likely to be charged off



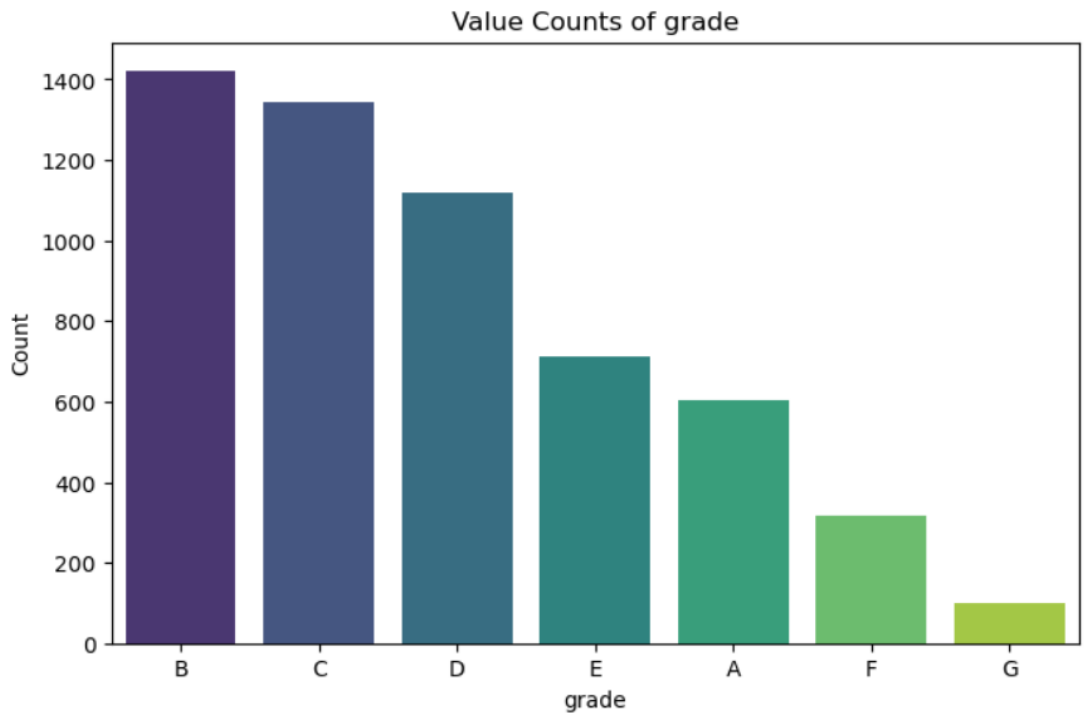
INSTALLMENTS / CHARGED OFF DATA

Installment does not actually impact if they are charged off or not as most of the person have less installment



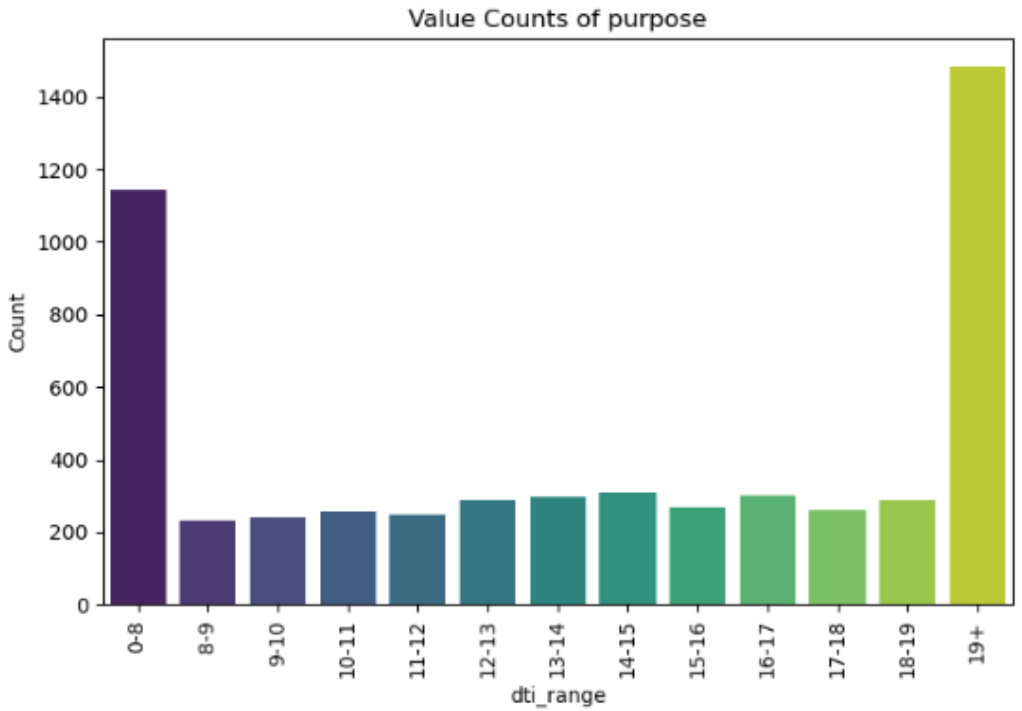
GRADE / CHARGED OFF DATA

Applicants with **grade B** are at higher risk of being charged off, next in line being grade C and grade D.



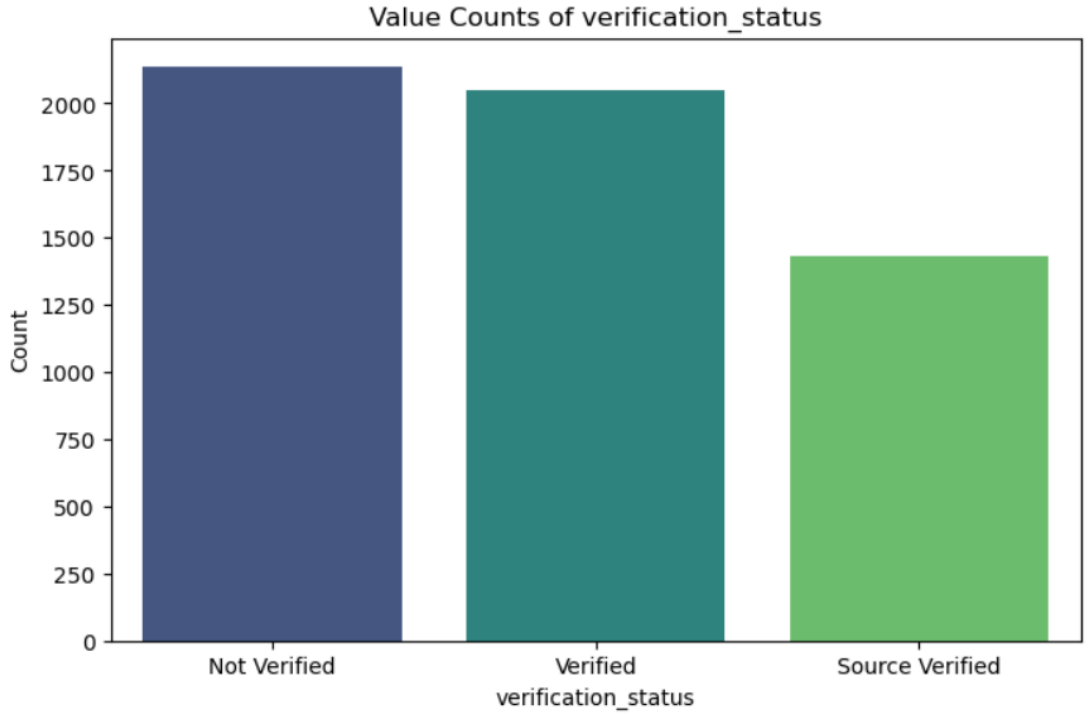
DEBT TO INCOME RATIO / CHARGED OFF DATA

People with high DTI are more prone to charge-offs'



VERIFICATION STATUS / CHARGED OFF DATA

Applicants with income **source verified** are **less** likely to get charged off



INSIGHTS BASED ON UNIVARIATE ANALYSIS

- Most loans are requested for amounts ranging between 0 to 5000.
- Most loans funded also fall within the range of 0 to 5000.
- Borrowers typically prefer a loan term of 2 years over 5 years.
- Majority of loan applicants are living on rent
- Most of the applicants have annual income in the range 40k to 80k
- Significantly large amount of loan applicants are not verified
- 2011 has proved to be the year in which most loans are taken
- Of the given data most of the applicants have successfully Fully Paid Off the loan
- Debt consolidation is the primary reason for borrowing, indicating many are using loans to pay off existing debts.
- CA was the state with the maximum number of loan applicants
- Applicants with debt to income ratio between 12% to 15% are mostly taking the loan

INSIGHTS BASED ON BIVARIATE ANALYSIS

- The higher the loan amount, the higher the chances of being charged off.
- The majority of borrowers have fully repaid their loans. These borrowers preferred the loan term of 2 years.
- People who have higher interest rates are at higher risk of being charged off .
- The applicants with home ownership as “OTHER” are at a higher risk of being charged off.
- Applicants with over 10 years of employment are more likely to get charged off.
- Applicants with income source verified are less likely to get charged off
- Applicants with grade B are at higher risk of being charged off, next in line being grade C and grade D.
- The proportion of Charged Off Applicants peaked in the year 2007
- Applicants seeking loans for small businesses are at higher risk of being charged off
- NE State records the highest proportion of charged-off applicants whereas states IL, IN, and ME have seen no charged-off applicants from 2007 to 2011.
- People with high DTI are more prone to charge-offs’.