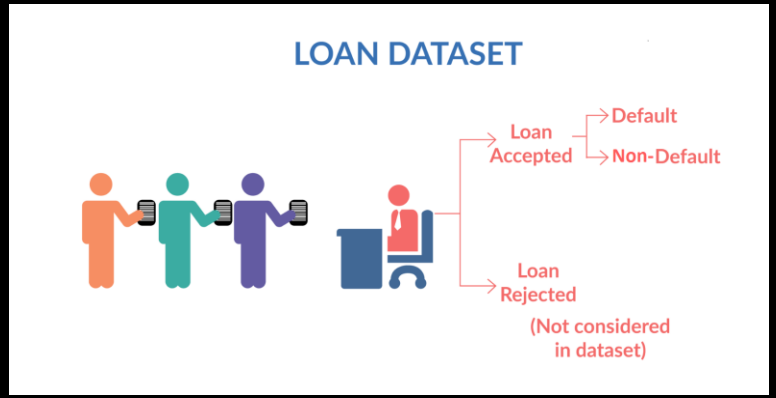
LENDING CLUB CASE STUDY

Contributor : Sayak Bhattacharjee

BUSINESS PURPOSE

The largest online lending club platform would like 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 STATEMENT

With the given data set that contains the complete loan data for all loans issued through the time 2007 t0 2011 we need to perform EDA to conclude the factors of default loan and thus control the credit risk.

ANALYSIS APPROACH

Domain knowledge and Data Dictionary Analysis

- Understanding of each column (out of 111 cols) of the data set
- Figuring out the driving variable which is loan_status

Data Clean up

- Duplicate rows
- Empty rows
- Completely Null rows and columns
- Mostly null rows and columns

Data Handling

- Convert columns to desired data type
- Create Derived column where required
- Identifying of non-critical columns

Univariate Analysis

- Distribution and variance of data
- Outliers analysis and removal

Bivariate Analysis

- Analyze the impact of different continuous and categorical variables on driver variable loan_status

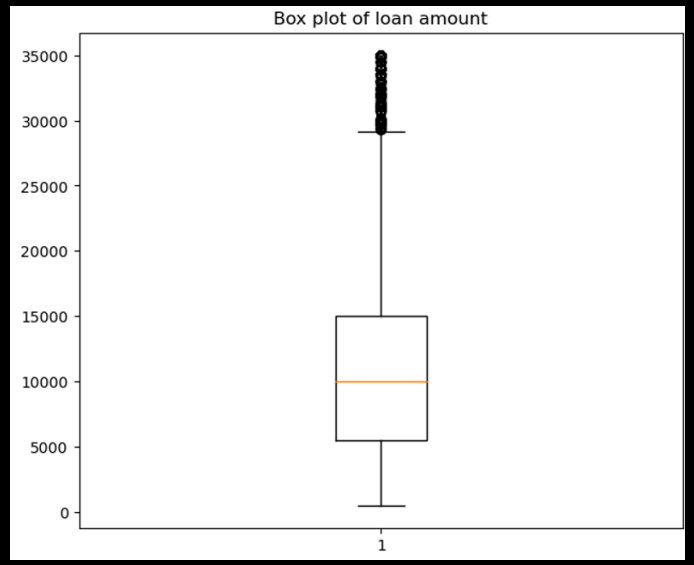
STRATEGY FOR BIVARIATE ANALYSIS

1. Categorical variable creation – We have created few of the categorical variables like annual_inc_group, int_rate_group …. Purpose was to create bins for the continuous variables to perform our analysis gracefully.
2. Charged Off Ratio over Absolute number – Instead of going with the exact number of Charged off loans we perform our analysis on the ratio between the charged off loans and total loans within a grouped data

UNIVARIATE ANALYSIS- 1

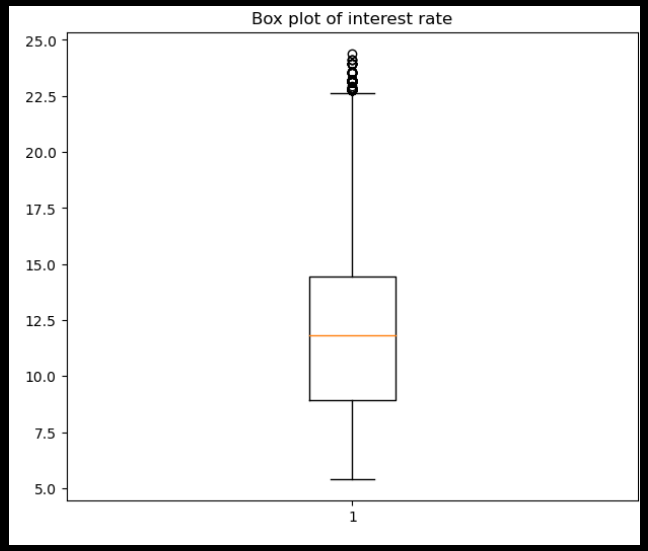
Loan Amount

no significant outliers for the continuous field `loan_amnt`. Specially from the box plot we find that the higher fence is at around 30000 whereas the max is around at 35000. This variation of data does not require outlier elimination



Interest Rate

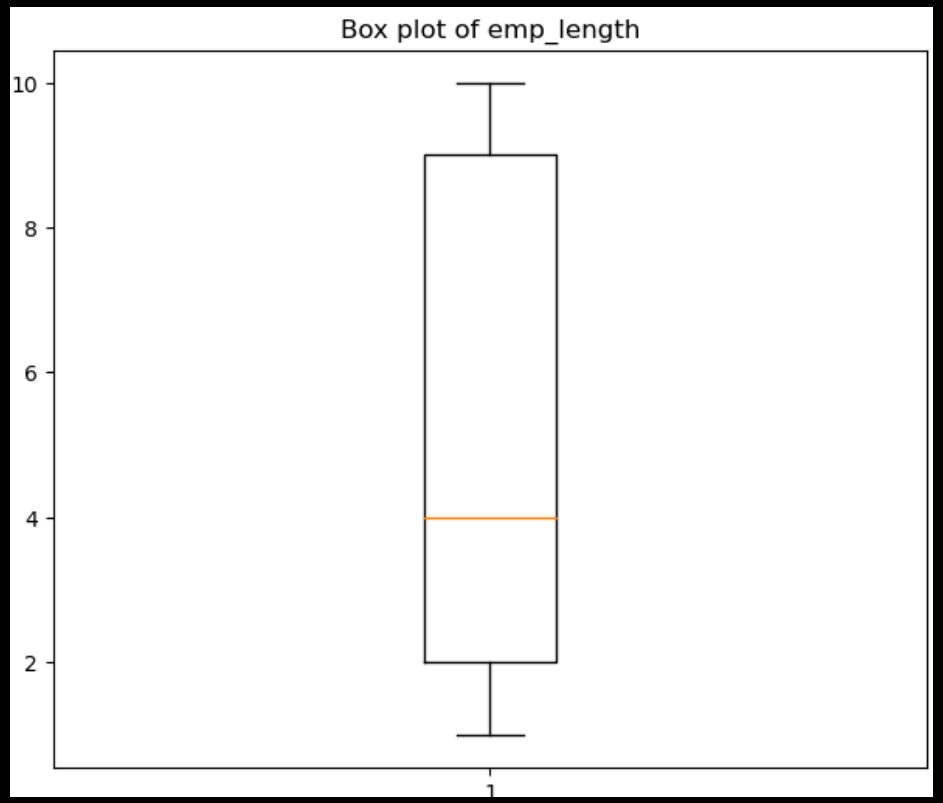
`int_rate` is that the upper fence value which is near to 22.5 and the max value which is near to 24. And thus, this variation does not require any outlier's treatment. We leave the data for this field as is for our analysis. Also, we see the major distribution of interest rate is in between 10%-14%



UNIVARIATE ANALYSIS- 2

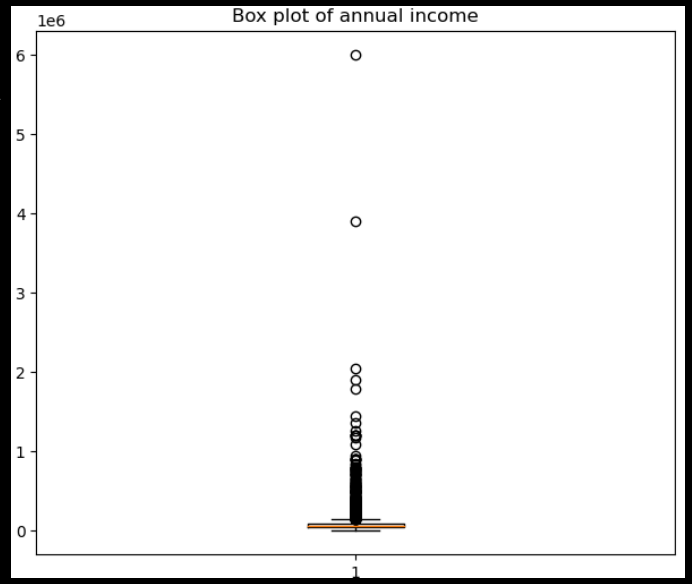
Years of Employment

No outliers (evident from the box plot) since all data within the upper fence. We leave the data for this field as is for our analysis. Also we see that most of the applicant's employment tenure is either less than 2 Years or 10 years and more



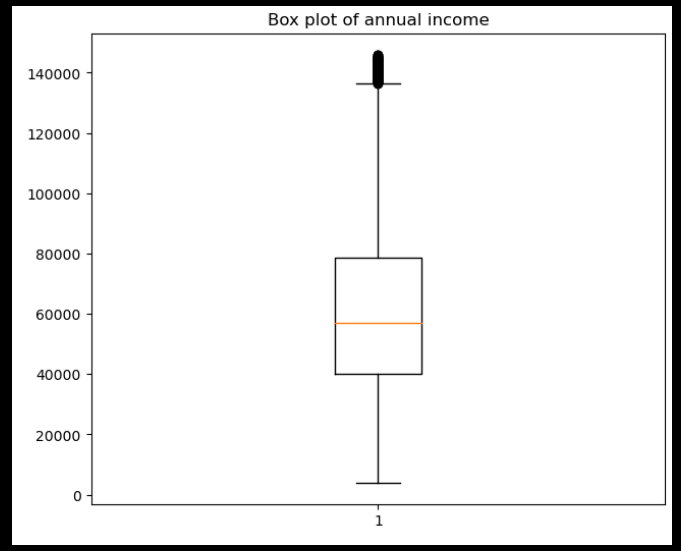
Annual Income

significant outliers for the field `annual_inc`



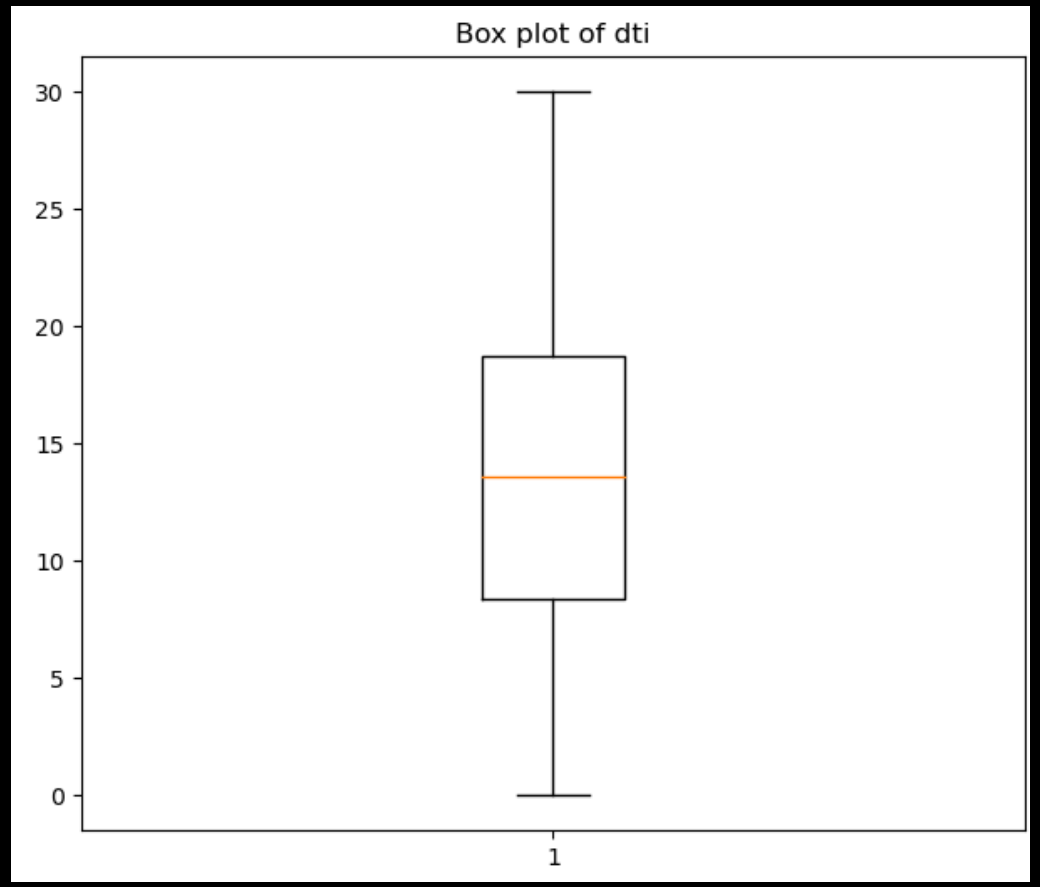
Outliers are removed based on IQR principle

Post outliers removal

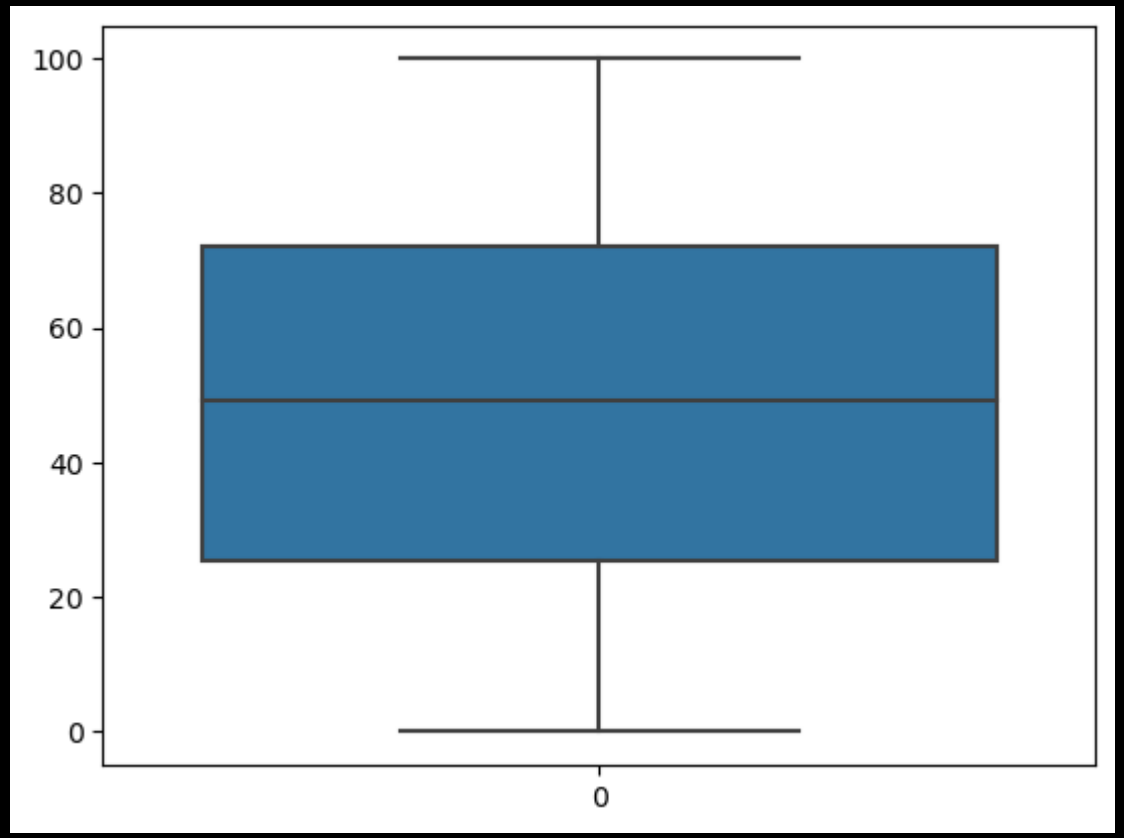


UNIVARIATE ANALYSIS -3

- DTI - Plot on `dti` looks pretty good, no otliers are found in the data set

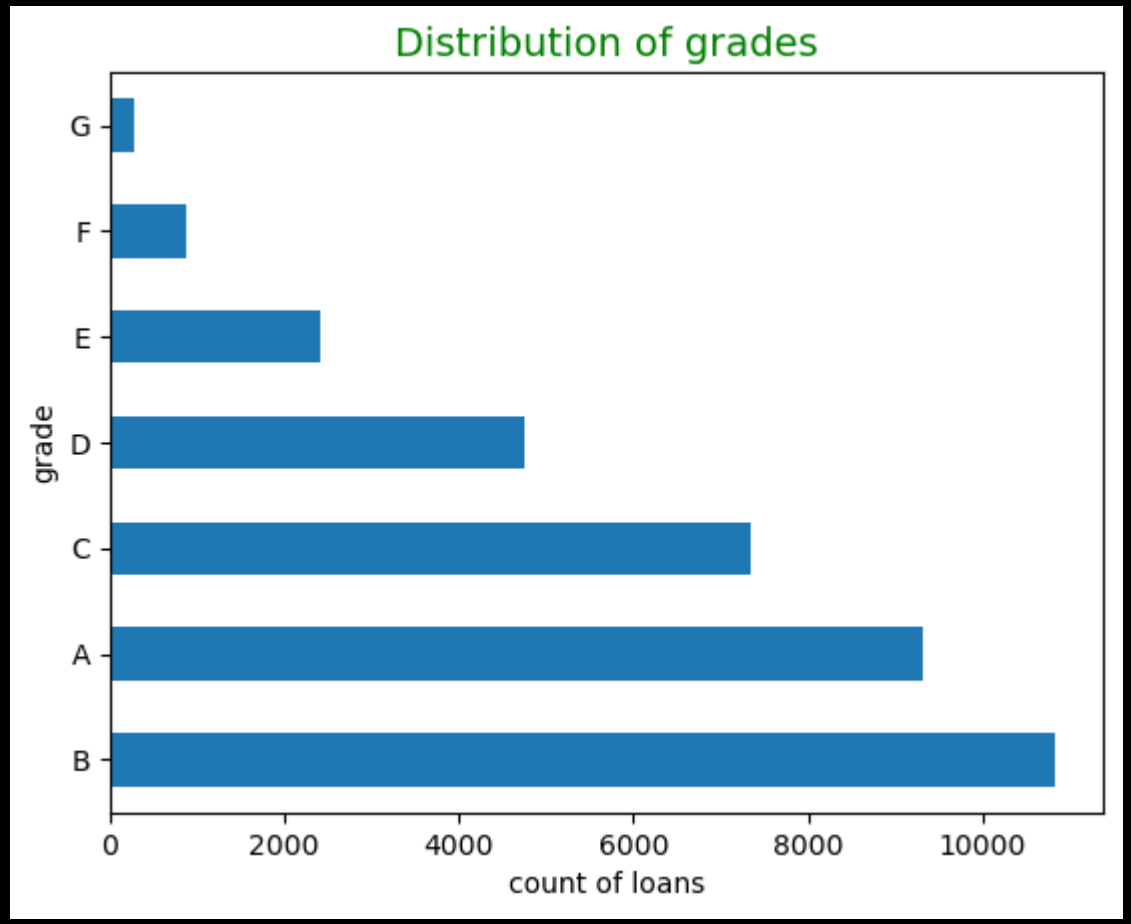


- Revol_Util - We see `revol_util` results a perfect boxplot and hence we keep thje data as is without any outlier treatment.

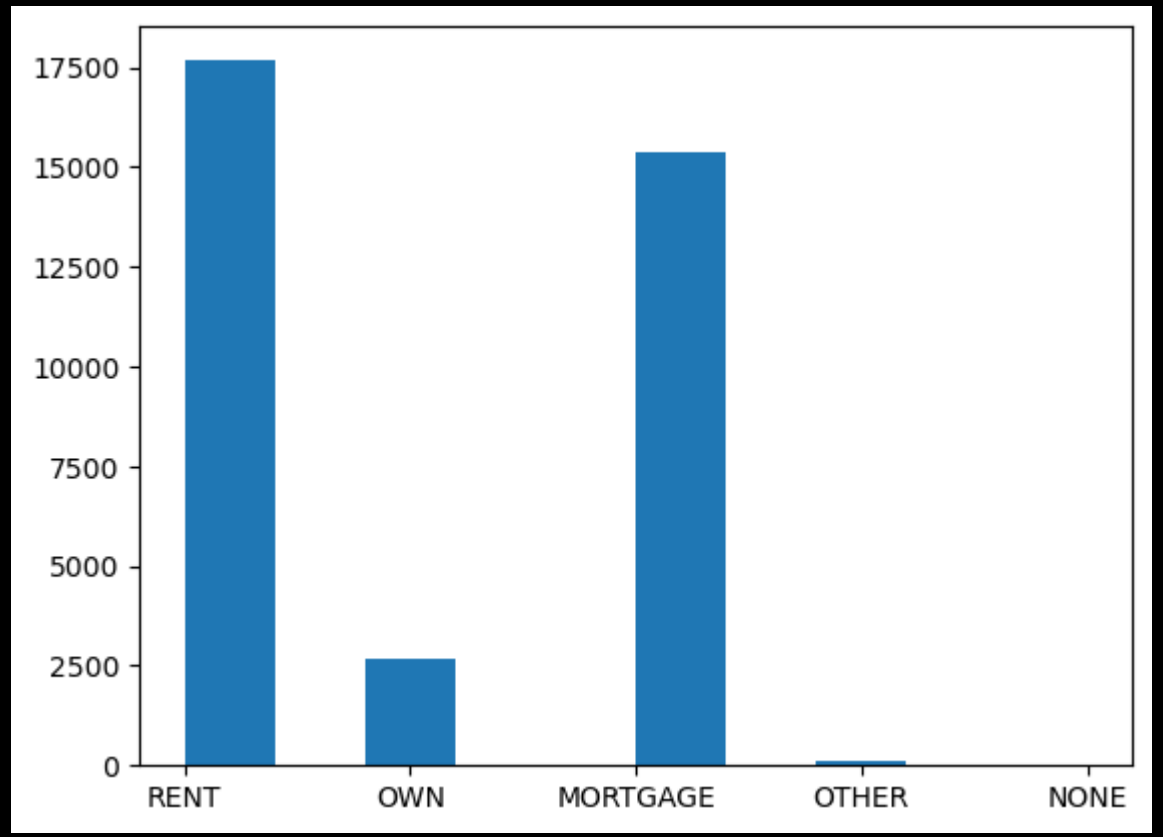


UNIVARIATE ANALYSIS- 6

Grade - Field `grade` A,B,C,D are quite dominant amongst all. Grade B counts more 10K , where as grade G is less than 500. We do not want to perform further univariate analysis on subgrade as that will not yield much value.

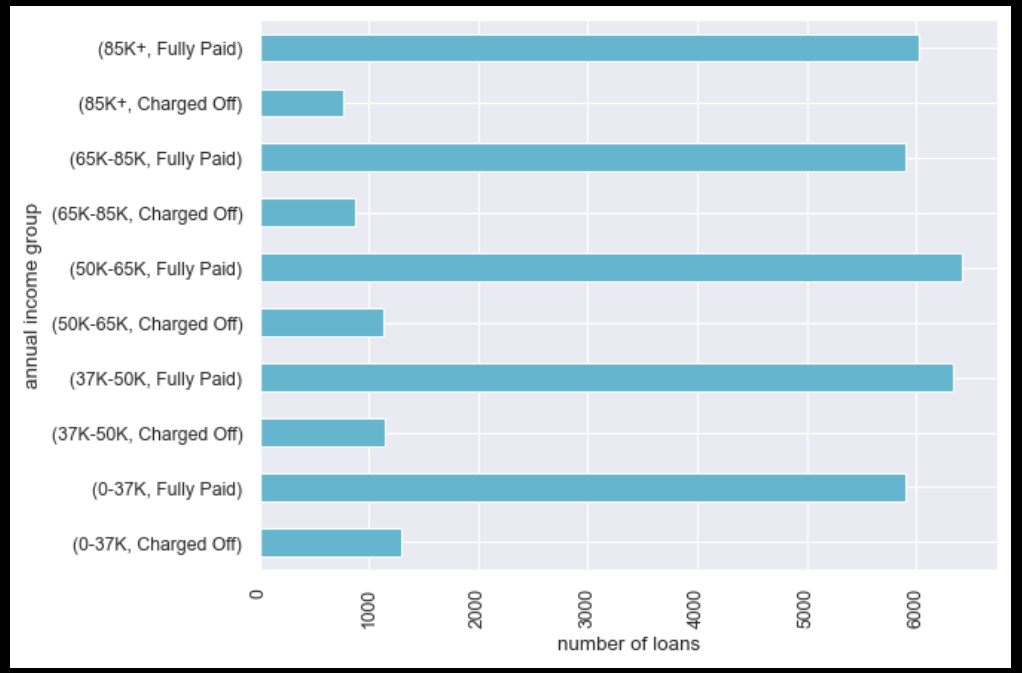


Home Ownership - Most of the loan applicants house type is either Rent or Mortgage. A very small portion of the applicants have their own houses.



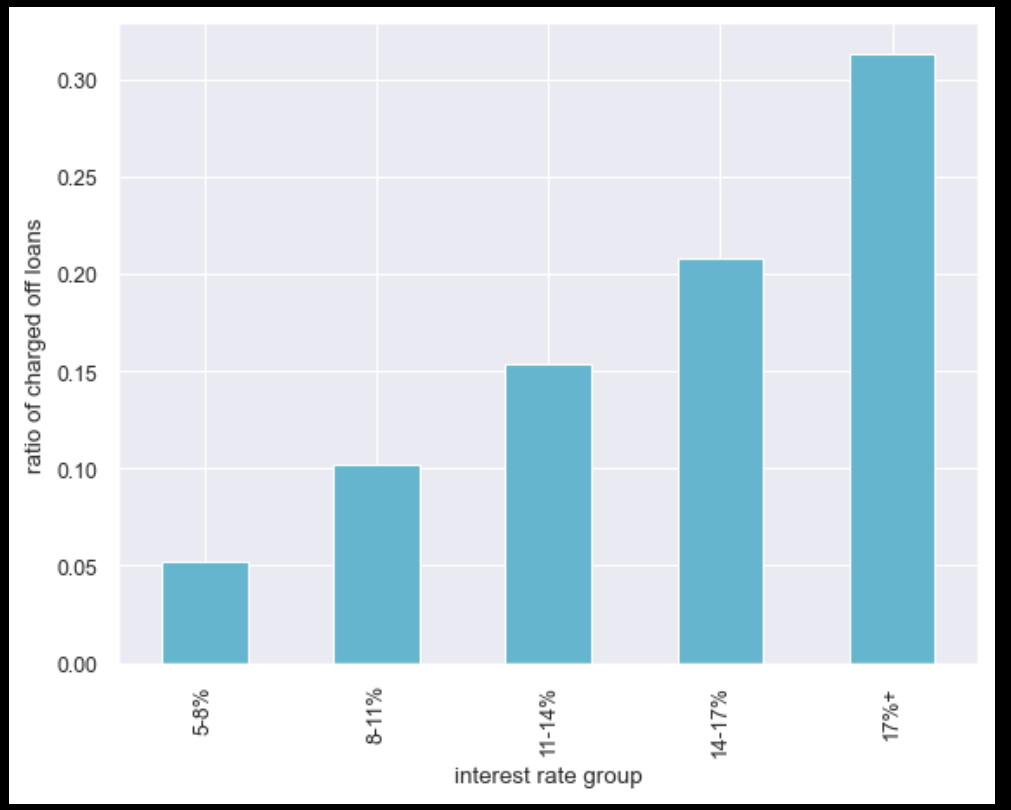
BIVARIATE ANALYSIS - 1

- Annual income and Charged Off Loan - We see with the increase of annual income the number of Charged Off loans are getting decreased.



- Interest Rate and Charged Off Loan Ratio

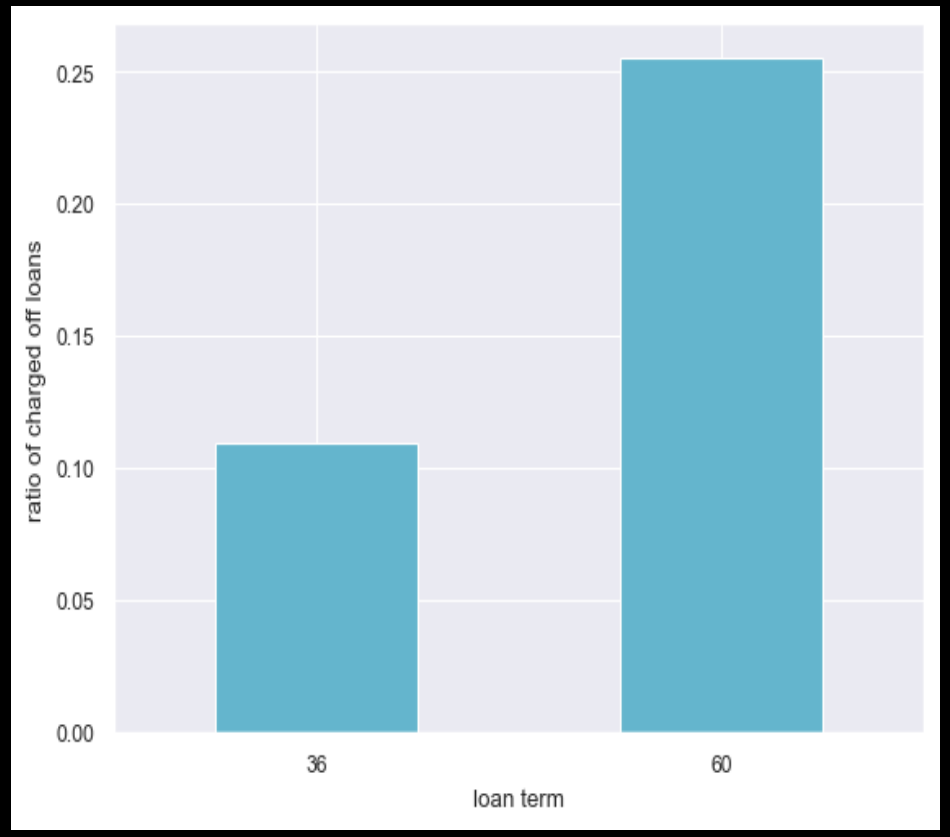
We get a clear indication that chances of loans getting charged off is very low if the interest rate is low(in rage 5-8%). It gets higher with the increase of interest rate and reaches the pick at 17% and more interest rate.



BIVARIATE ANALYSIS - 2

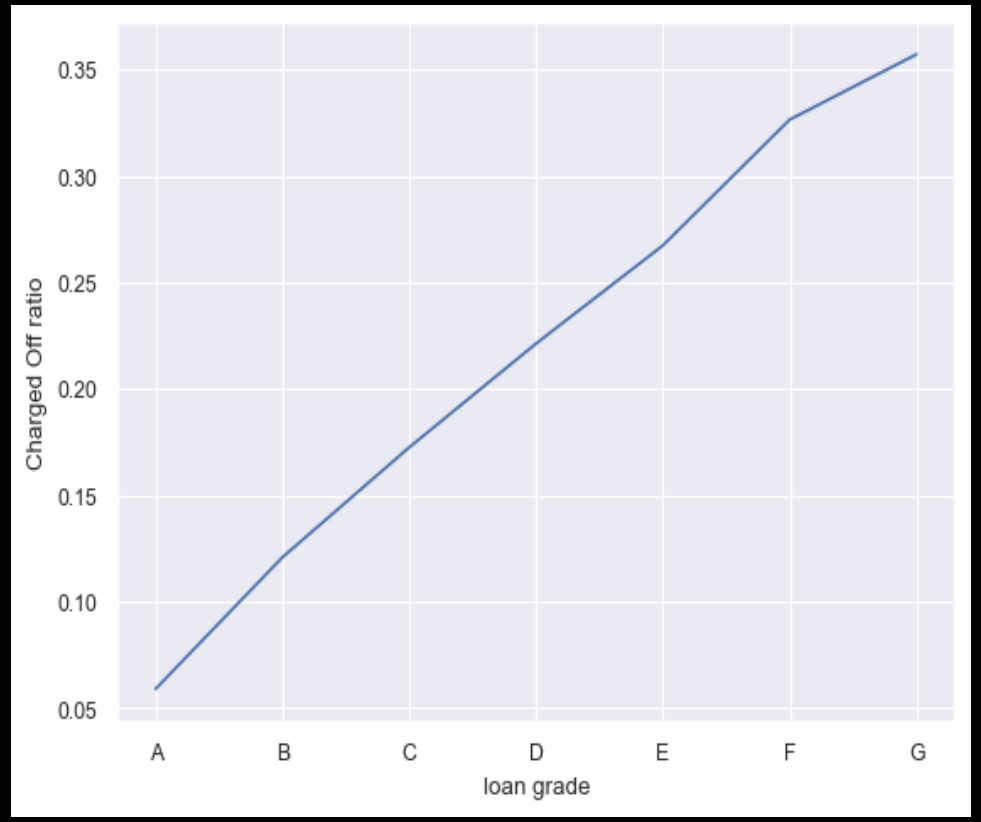
Loan term and charged off loan ratio-

The above plot reveals a very eminent trend that loan term 60 months are much more prone to get charged off than of the loan terms 36 months.



Grade and charged off loan ratio-

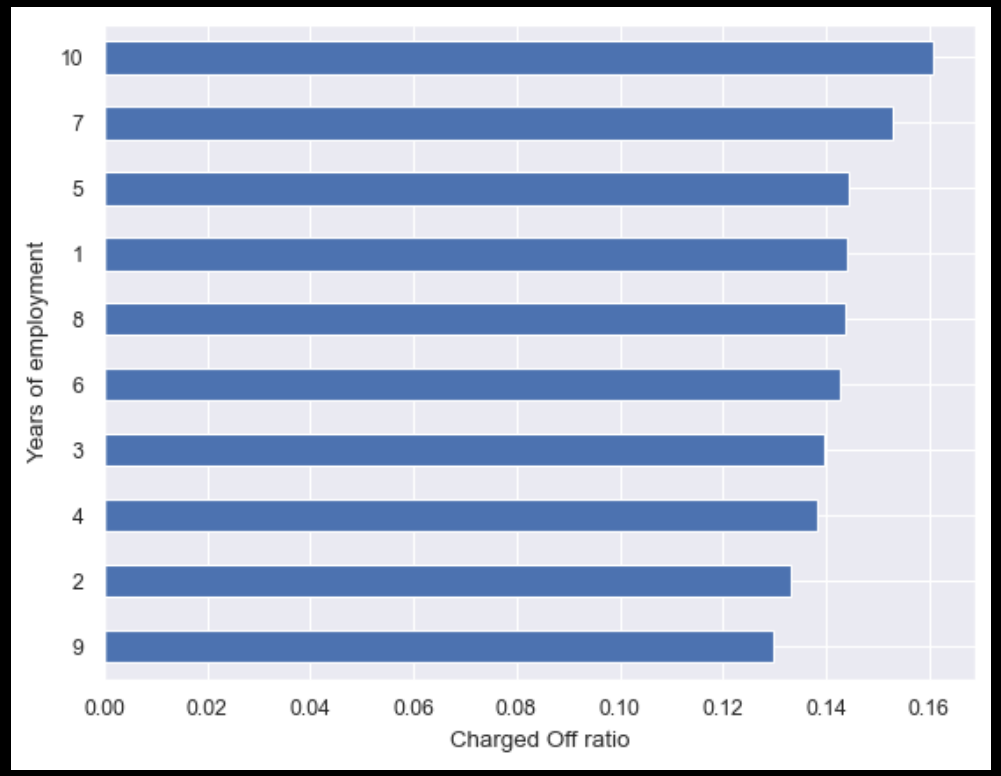
We find an almost linear line between loan grade and the charged off loans. As the loan grade increases the chances of a loan to be charged off increases. It also proves that higher grade loan bears much more credit risk for the lenders.



BIVARIATE ANALYSIS - 3

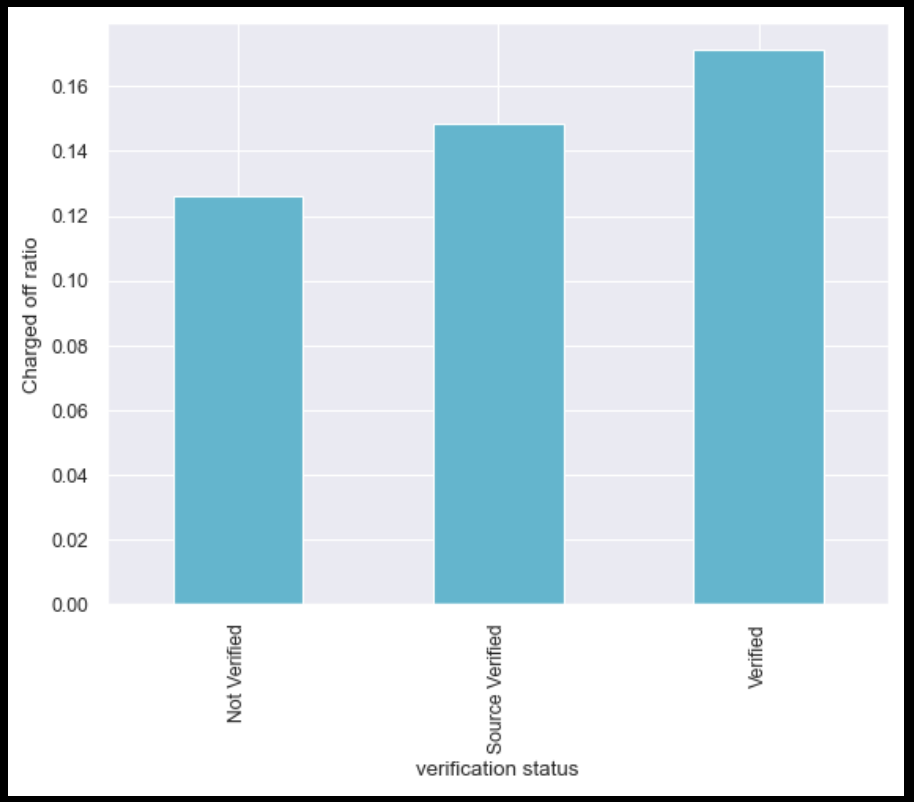
Employment length and Charged Off loan ratio

We don't see any clear trend between the possibility of a loan gets charged off with the years of employment of the borrower. We see most loans are charged off when the employment years is 10 years or above, whereas it is the least for 9 years and above. Also, the variation is too small to infer any conclusion between these two variables.



Verification status and charged off loan ratio-

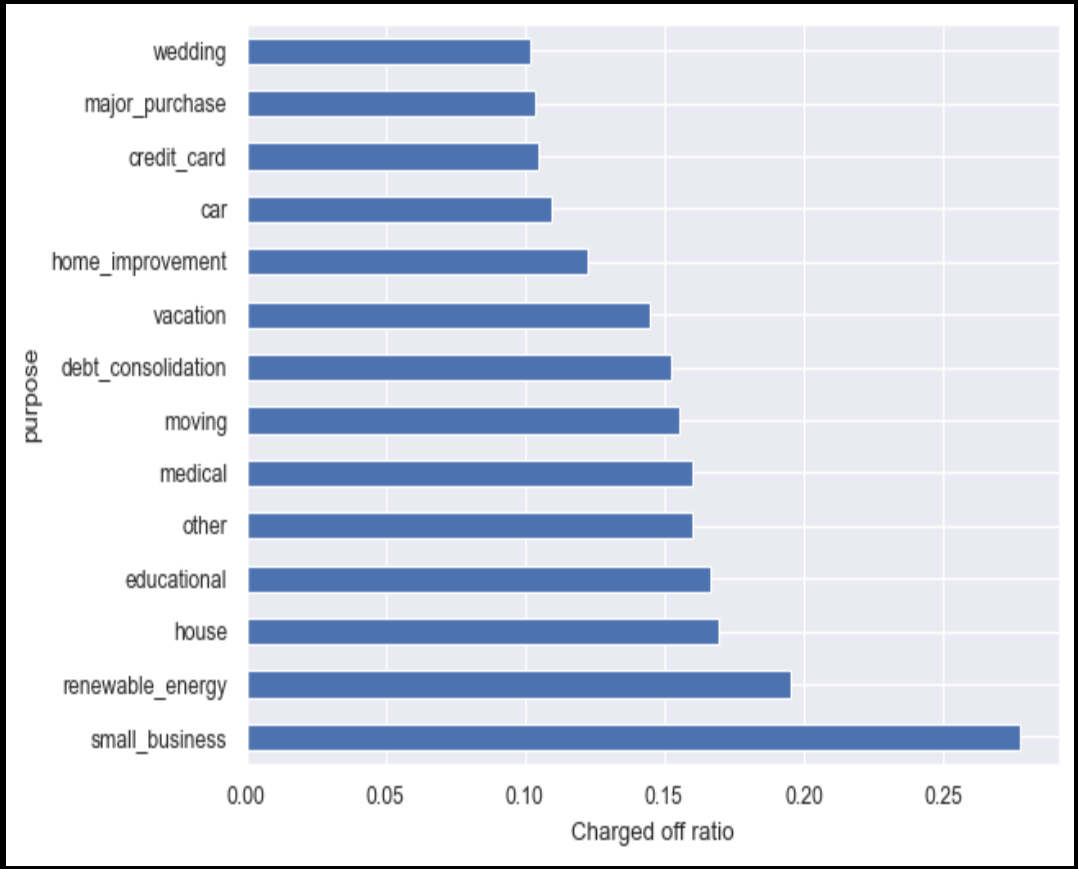
We see that verification status does not have much influence on Charged off loans.



BIVARIATE ANALYSIS - 4

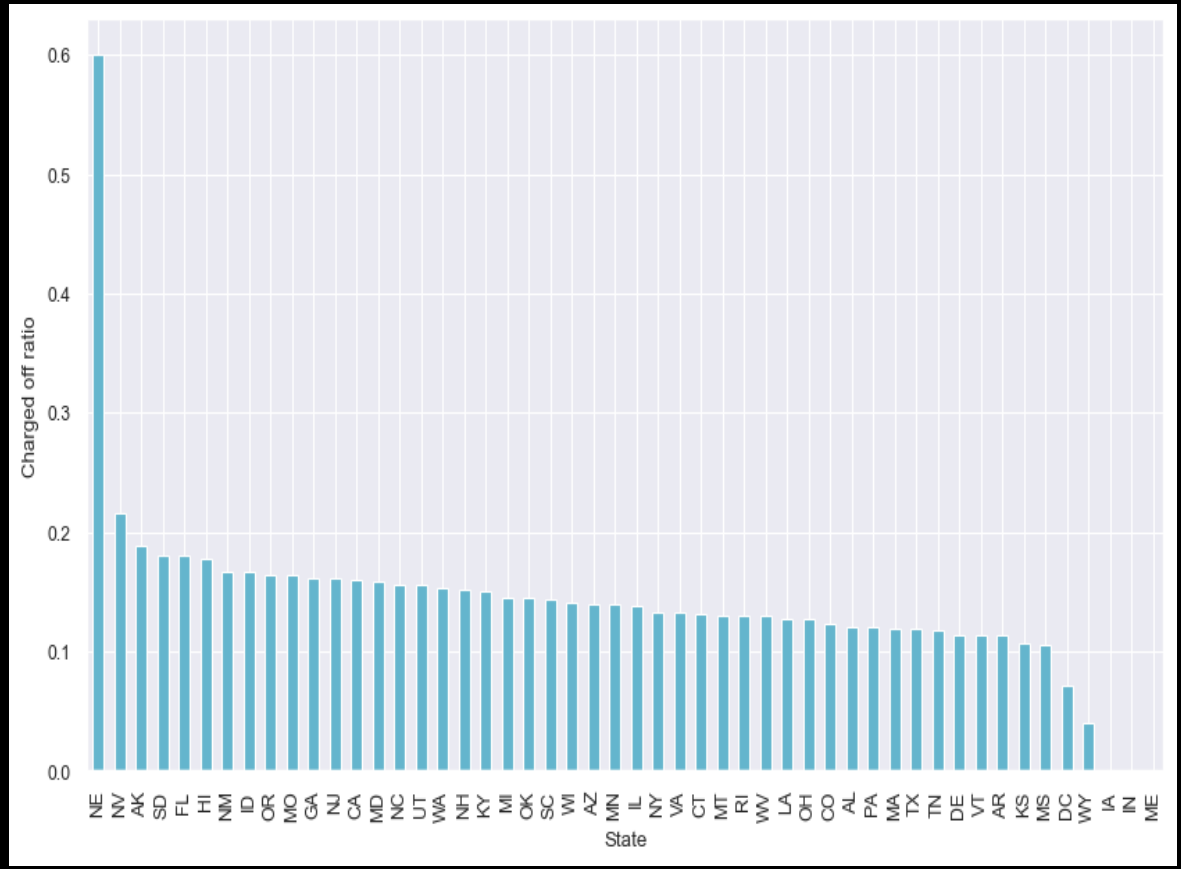
Purpose and Charged Off loan ratio

We see that for small business the chance is highest that the loan may get charged off. And it is the lowest for loan taken for marriage.



Address State and charged off loan ratio-

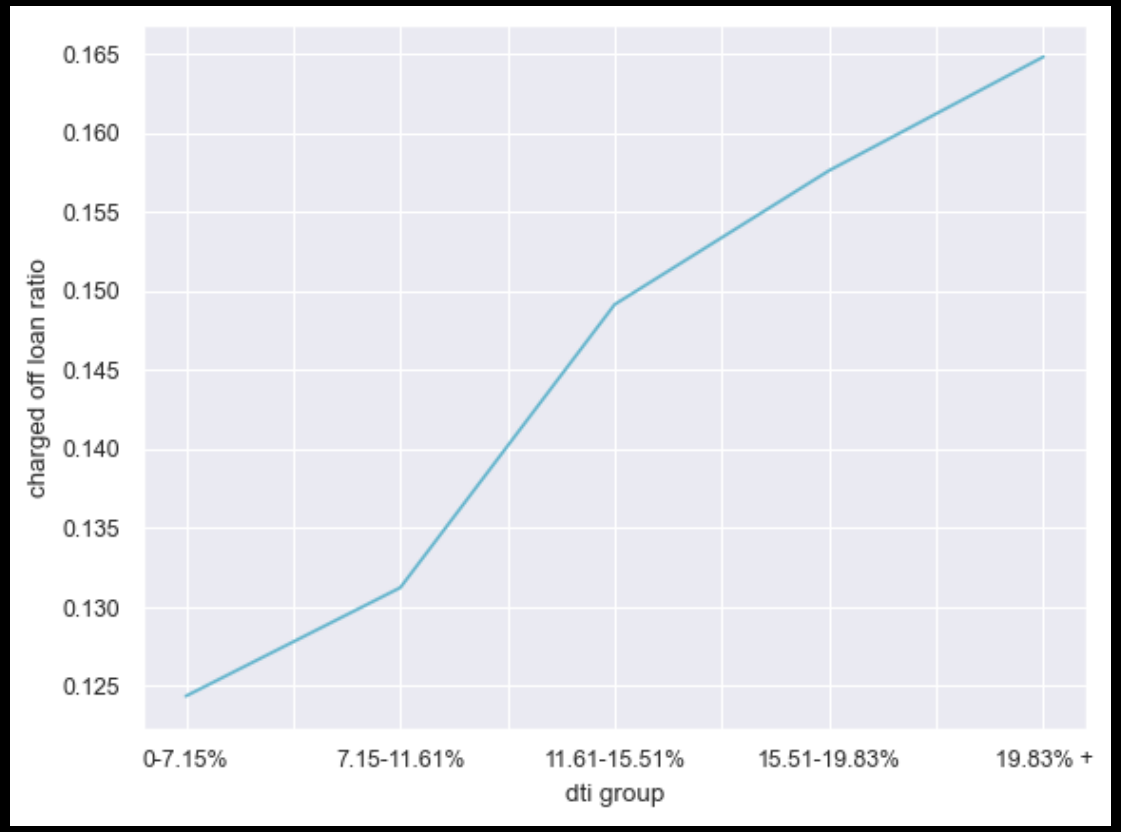
We see a significant high proportion of charged off loans in the state NY. Lender should be careful of this fact.



BIVARIATE ANALYSIS - 5

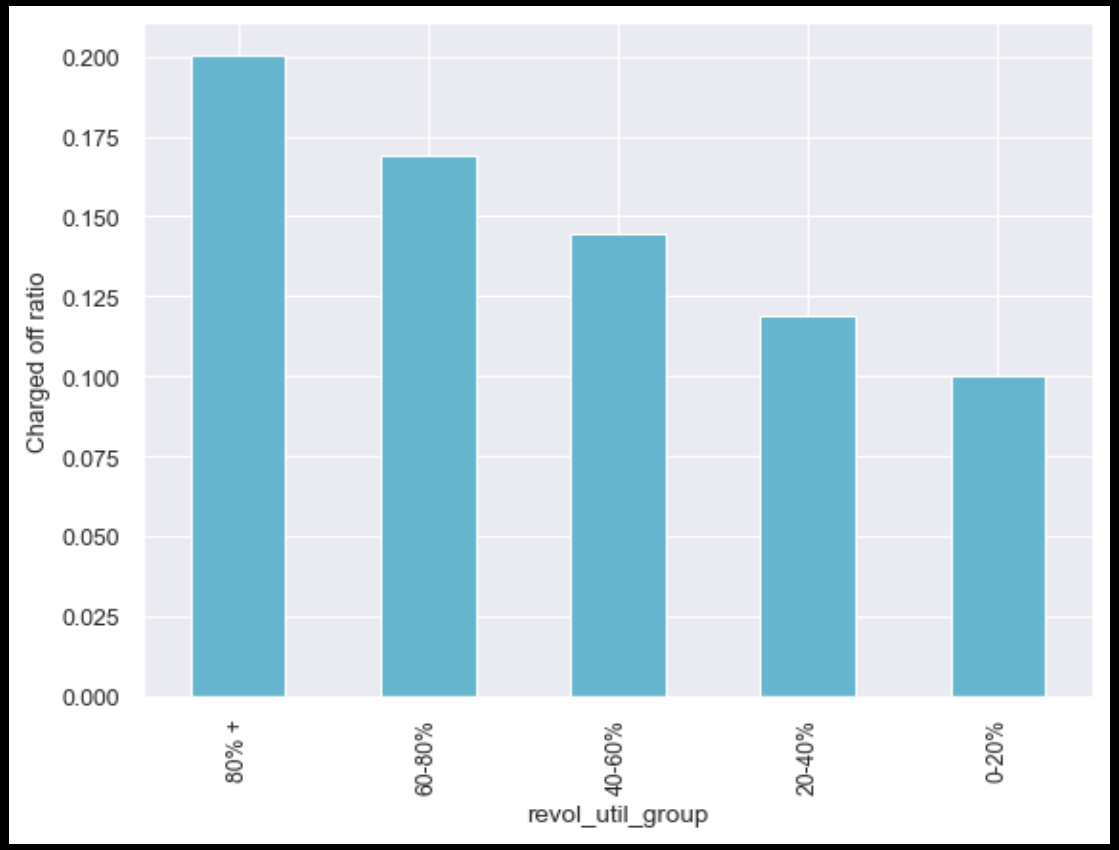
DTI and Charged Off loan ratio

We see a clear trend here that with the increase of dti possibility of loans getting charged off are increasing. The jump in this increase is highest between `7.15-11.61%` and `11.61-15.51%`.



Revolving utilization and charged off loan ratio-

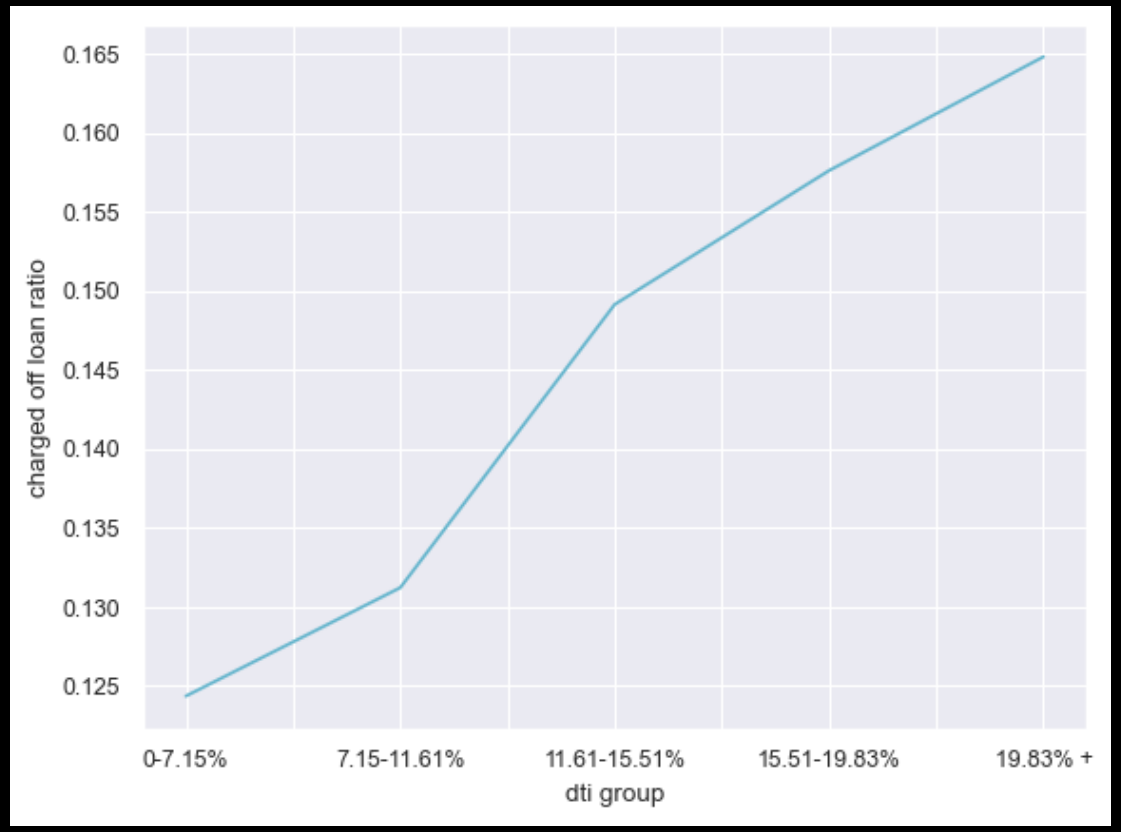
The observation is with the higher revolving utilization ratio possibilities of a loan gets charged off increases.



BIVARIATE ANALYSIS - 6

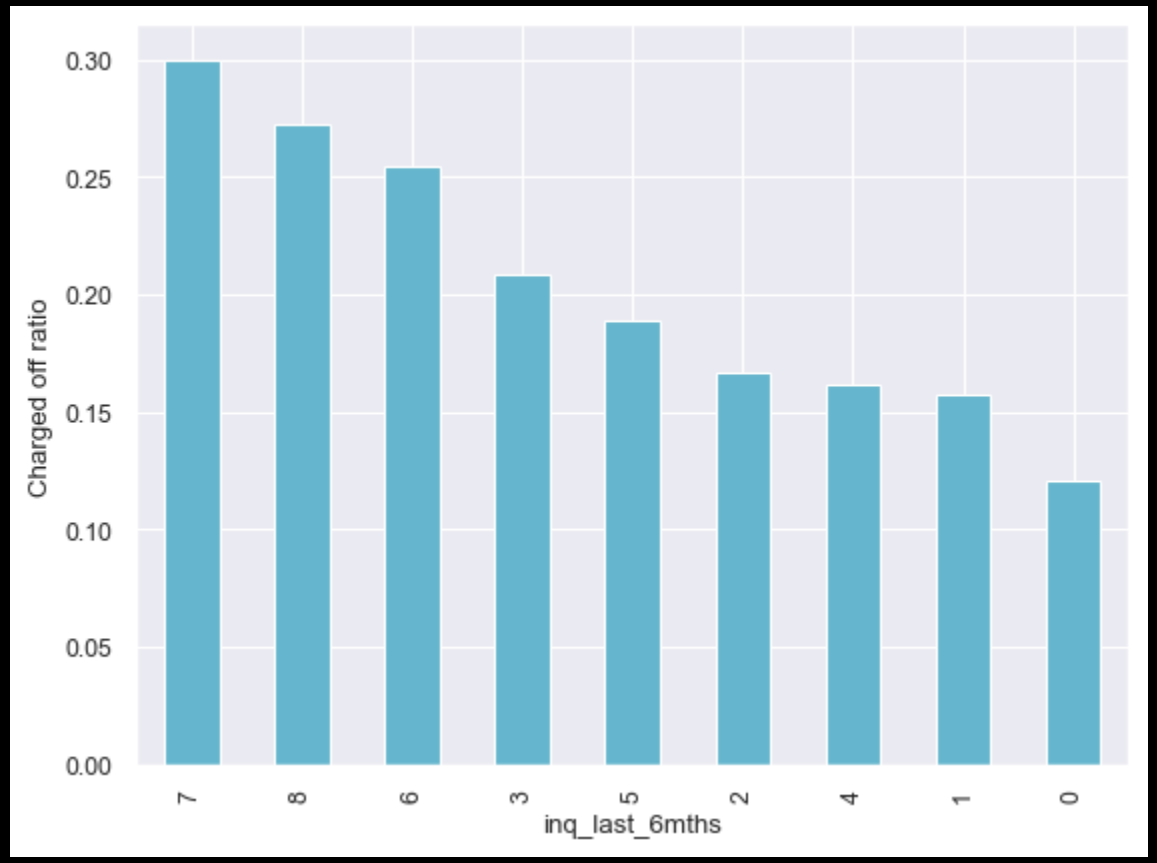
DTI and Charged Off loan ratio

We see a clear trend here that with the increase of dti possibility of loans getting charged off are increasing. The jump in this increase is highest between `7.15-11.61%` and `11.61-15.51%`.



inq_last_6mths and charged off loan ratio-

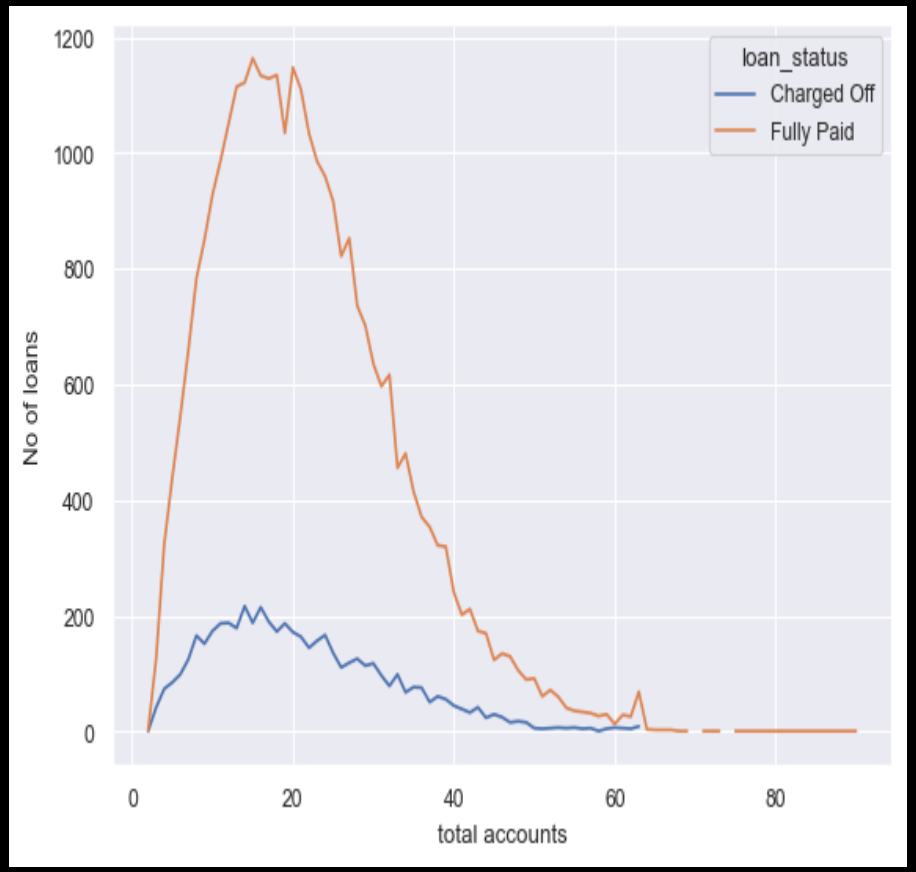
We see the trend that with higher number of credit inquiries charged off ratio is also getting increased.



BIVARIATE ANALYSIS - 7

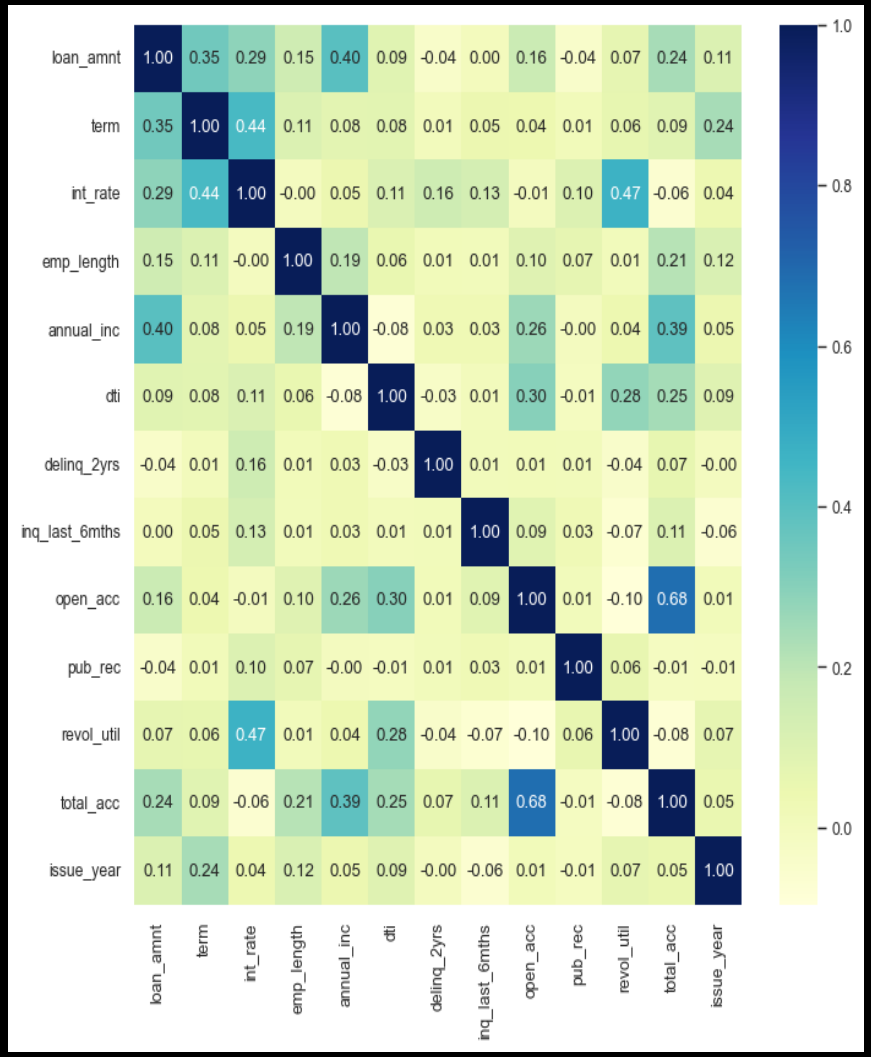
total_acc and Charged Off loan ratio

We see that when total accounts are increasing number of charged off loans are decreasing.



Heat Map Analysis

- Observation of positive co relation
  - `total_acc` and `open_acc`
  - `annual_inc` and `loan_amnt`
  - `revol_util` and `int_rate`
  - `term` and `loan_amnt`
  - `int_rate` and `term`
- Observation of negative co relation
  - `open_acc`and `revol_util`



THANK YOU