LENDING CLUB CASE STUDY

Group Members

Samyatirtha Bhattacharjee

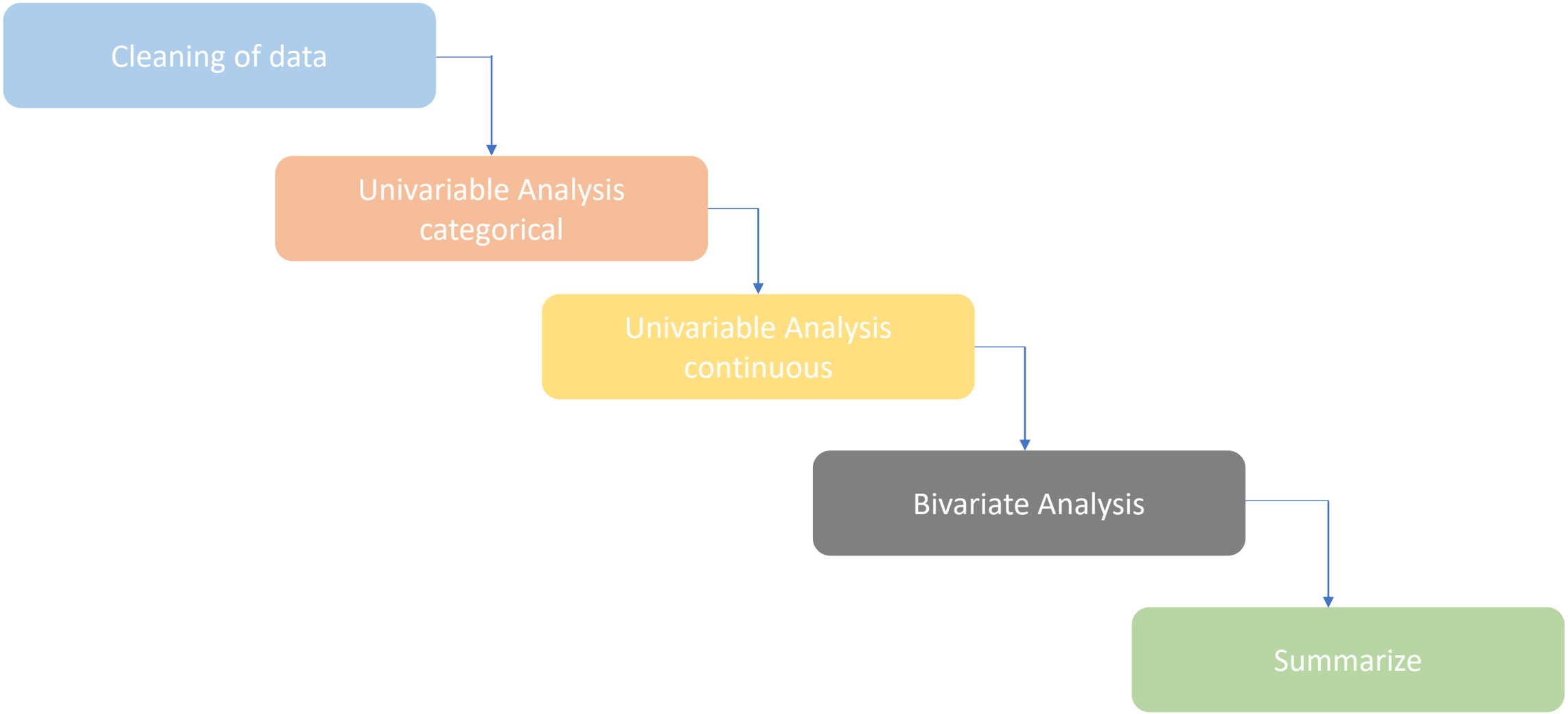
Shanu PN

Submission Date : 07/09/2022

Problem Statement

- Lending Club seeks to comprehend the driving forces underlying loan default, i.e., the driver variables that serve as reliable predictors of default.
- As a data scientist employed by Lending Club, you examine the dataset including data on previous loan applications using EDA to comprehend how consumer factors and loan attributes affect the likelihood of default.

Methodology



(38
38577

Data scrubbing

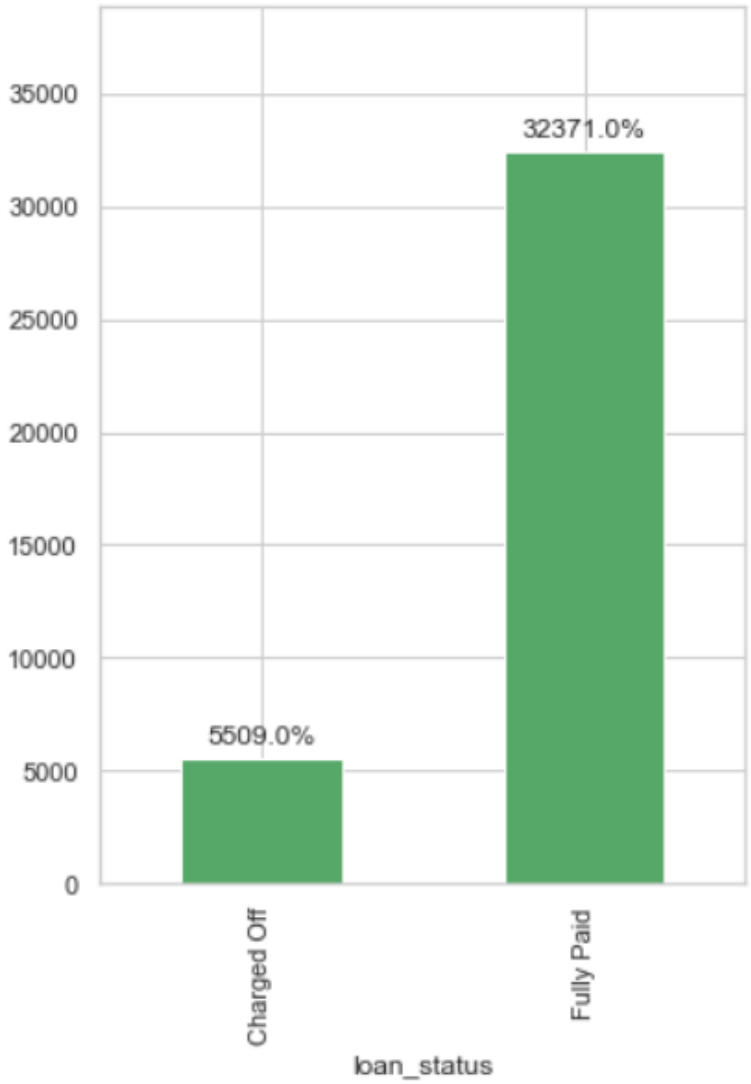
- Columns with more than 50% of the data missing were removed because the analysis based on the m might produce false conclusions Furthermore, the vast majority of the columns are with 100% missing data.
- Redundant columns with similar information’s and 0’s are dropped.
- Data uniformity is maintained across all the columns - Splitting and conversion
- Data of current loans are dropped as these might or might not be charged off and hence will not help for analyze.
- Derived new columns for better analysis, the final data set dimension for the analysis is (38577, 42).

Analysis

- Fully paid loans – 2371

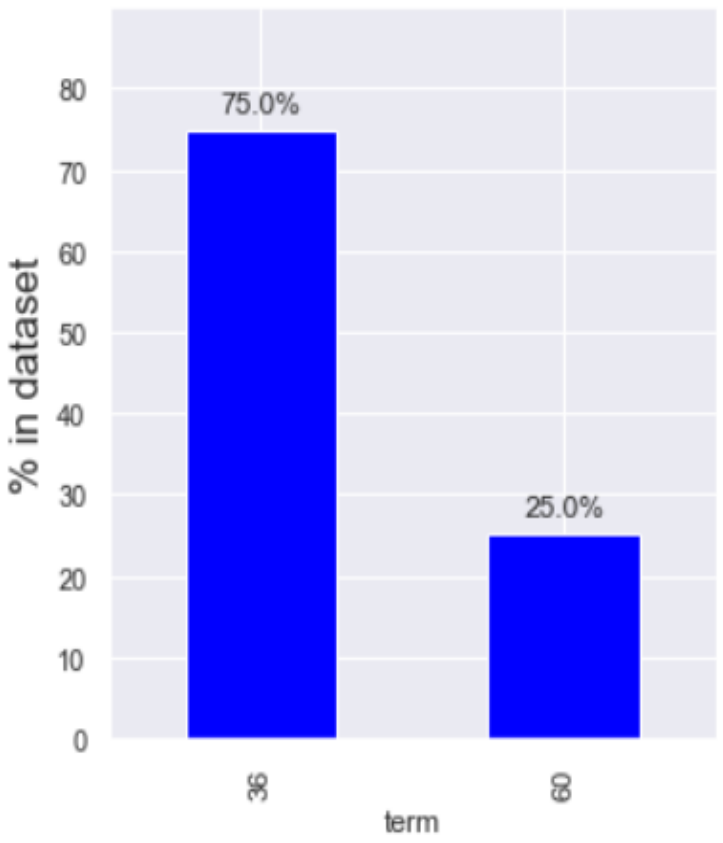
  charged off loans - 5509

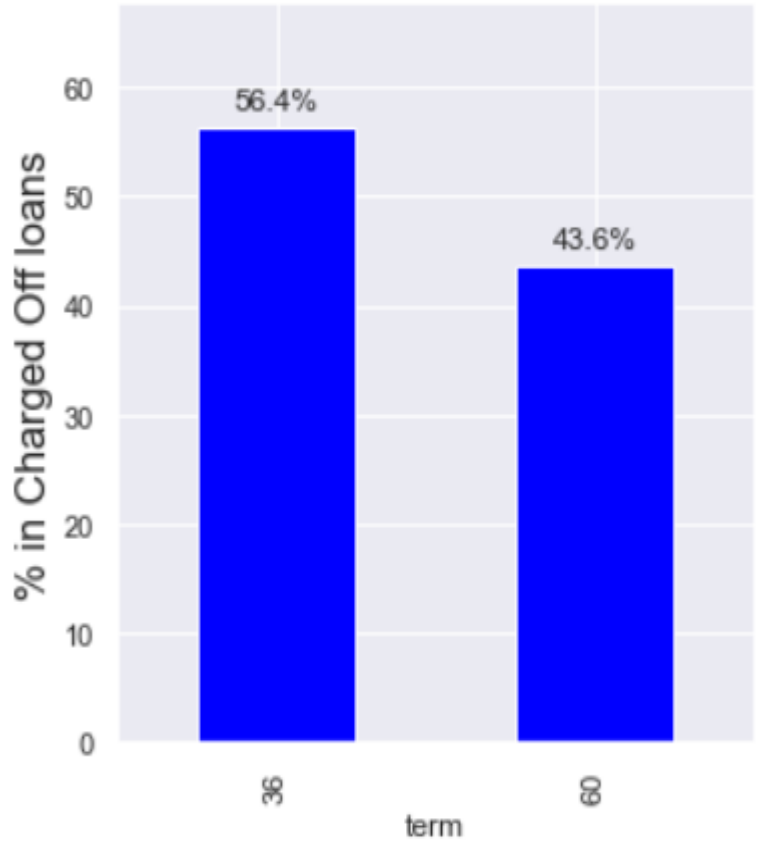
- In order to acquire deeper insights, cate gorical and continuous variables are als o analyzed for the charged off data set i n comparison to the entire data set.



Univariate Analysis - Term

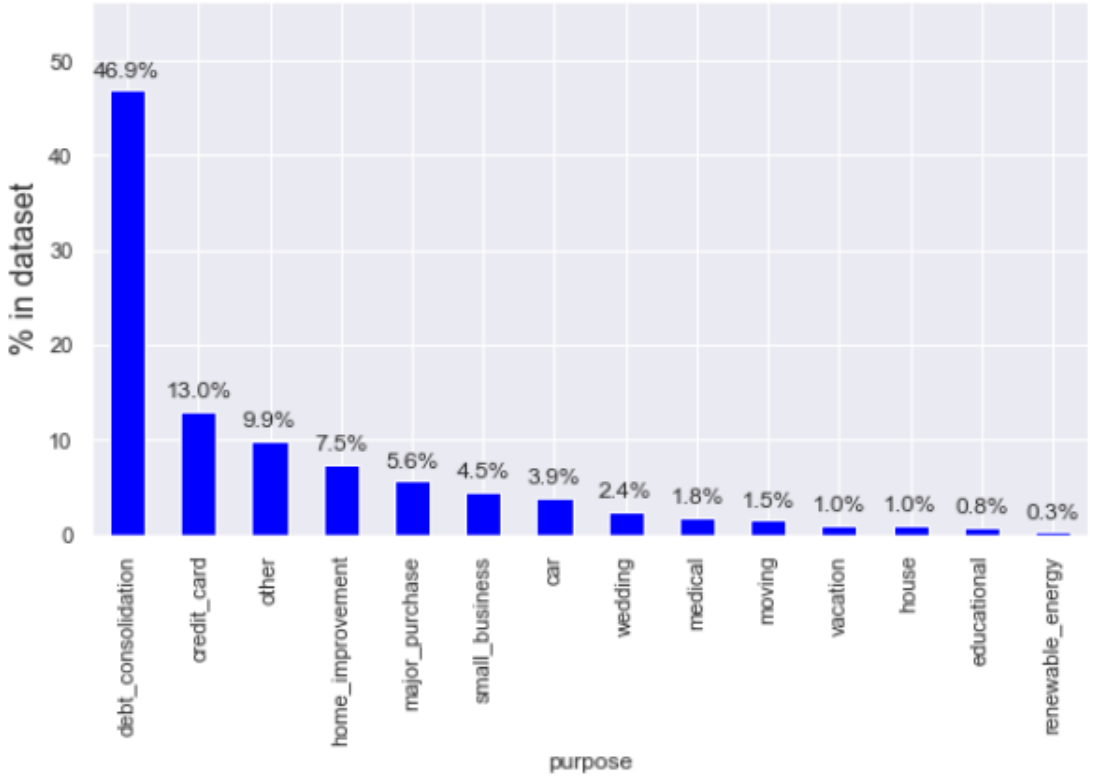
- The percentage of 60-term loans in the Charged Off dataset has significantly increased, indic ating that applicants for 60-term loans are more likely to fail.
- As a result, since a 60-month term is more likely to be c harged off than a 36-month term, the interest rate sho uld be higher for the 60-month term than it should be for the 36-month term.

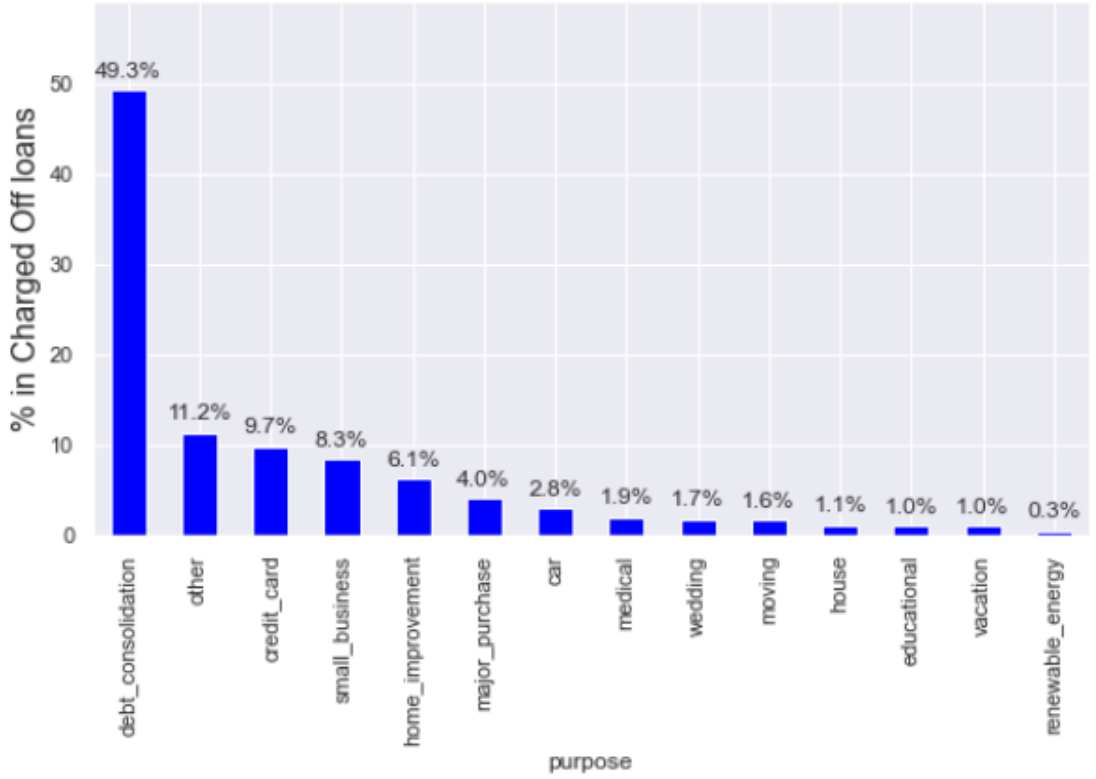




Univariate Analysis - purpose

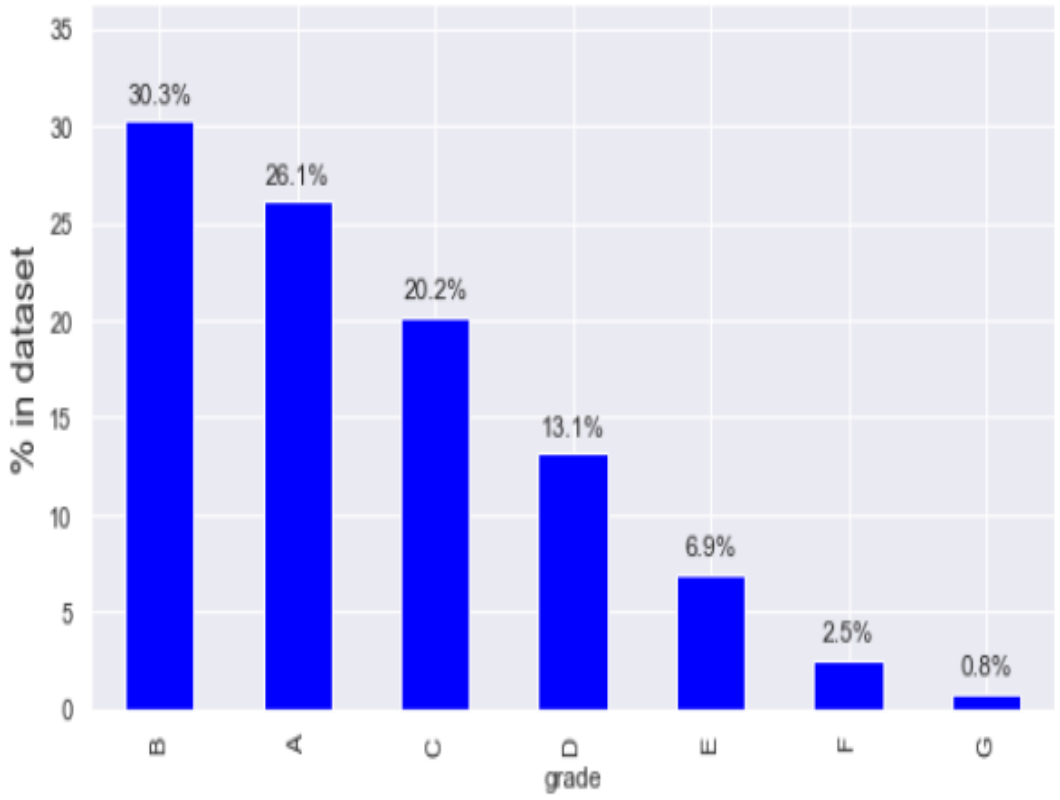
- From the below plot, loans for small business are more like to be defaulted. Hence small business loans interest rate can be more.

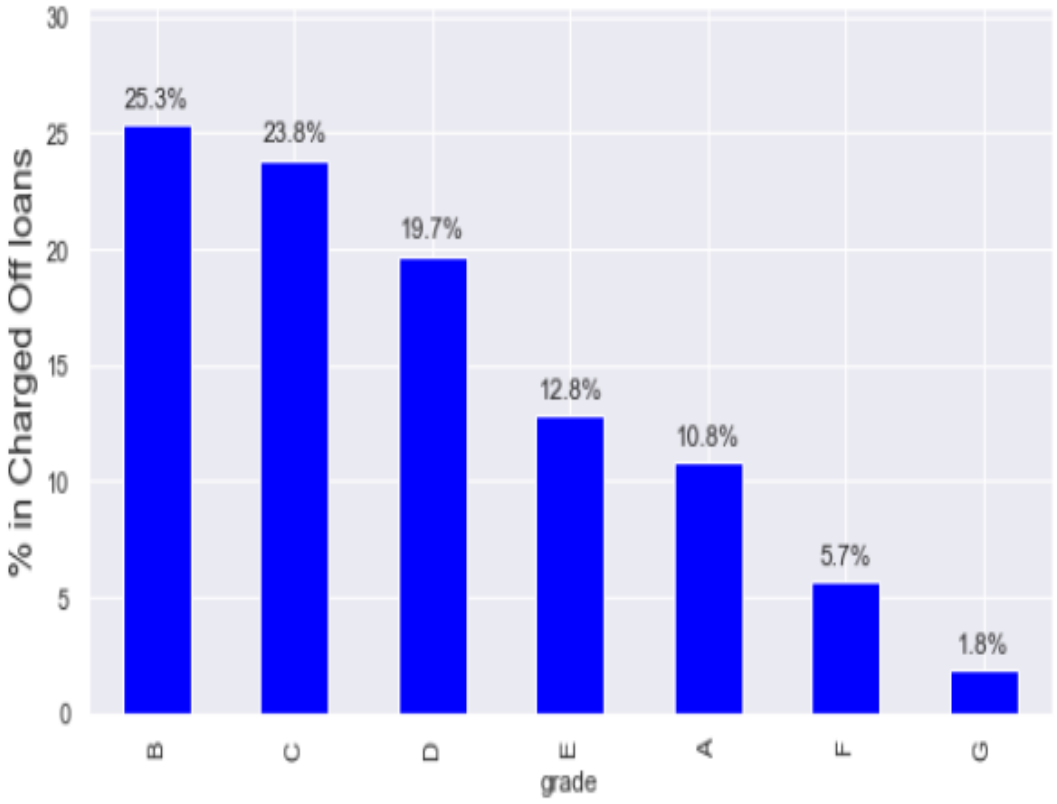




Univariate Analysis - Grade

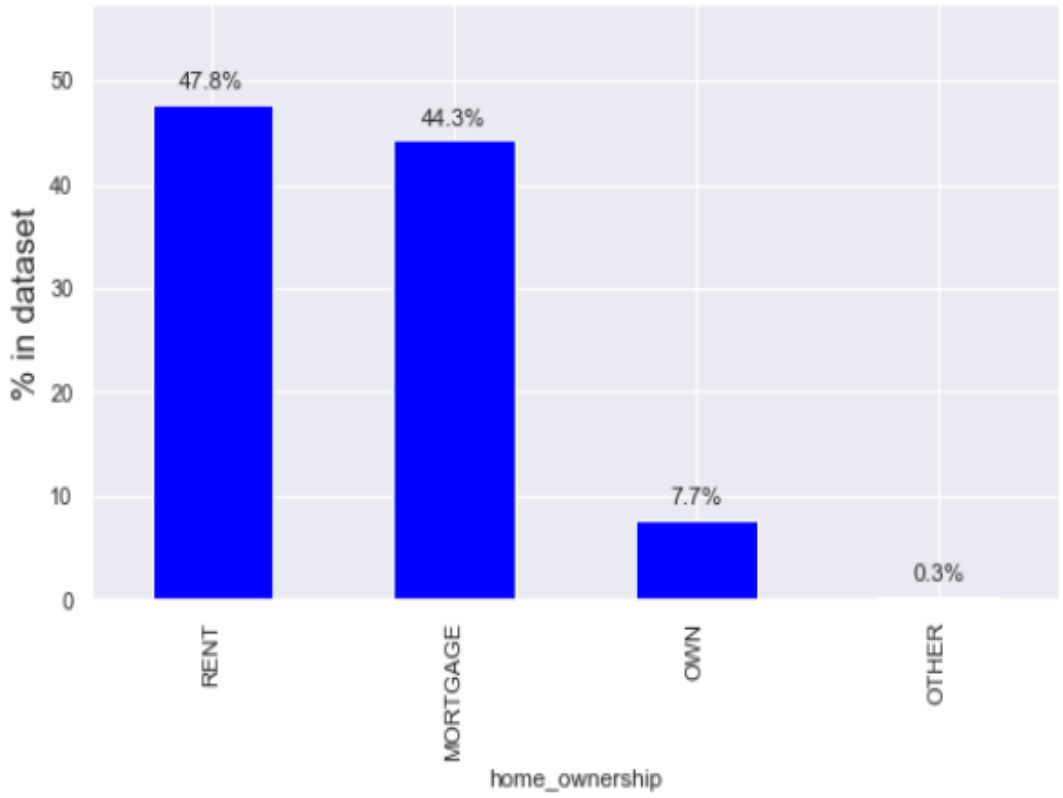
- Grade A is less likely to default whereas D,E,F are more likely to default while G grade is ignored due to very less volume.

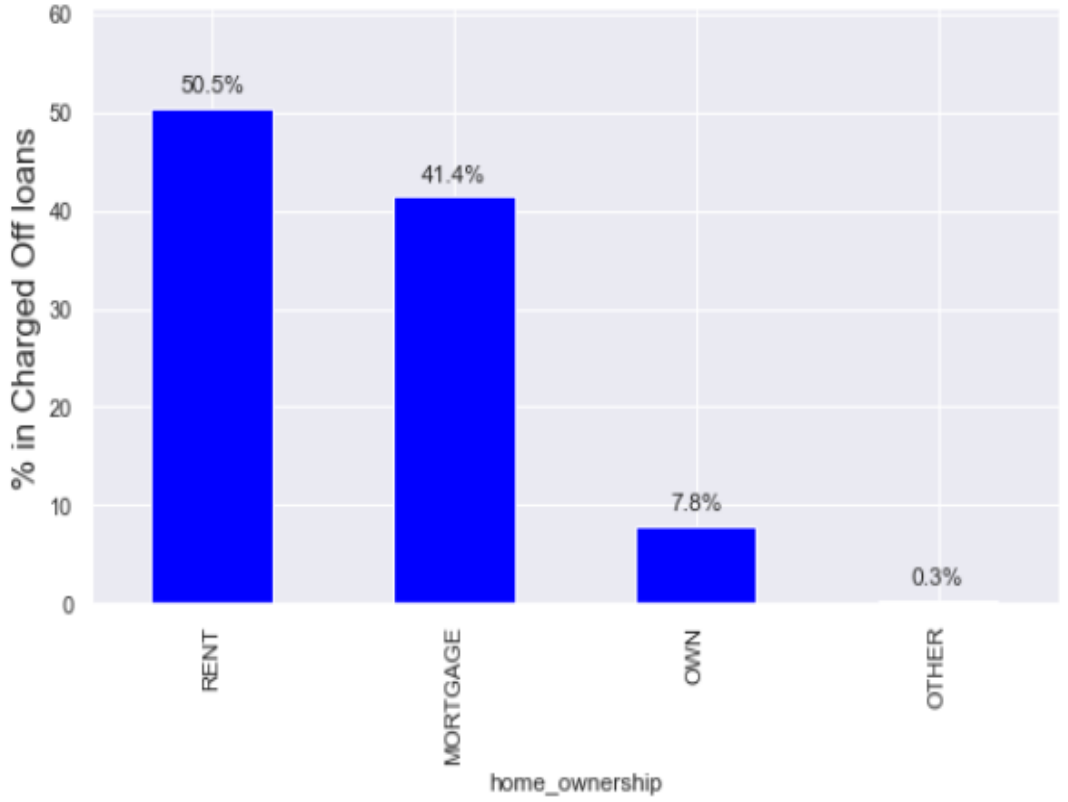




Univariate Analysis – home ownership

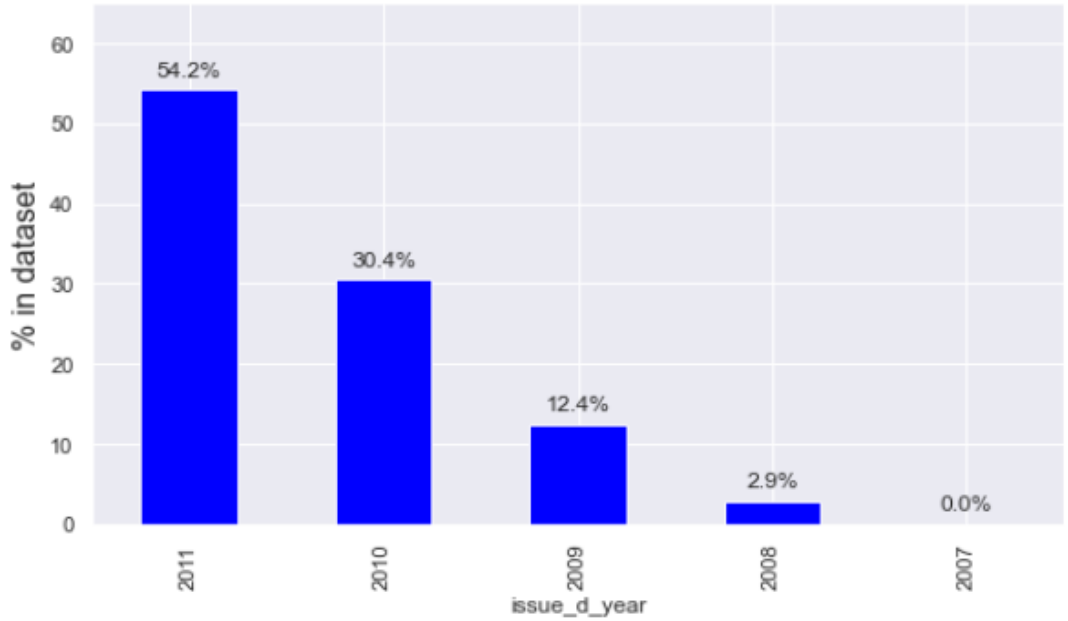
- There is no significant influence of home ownership on loan defaults

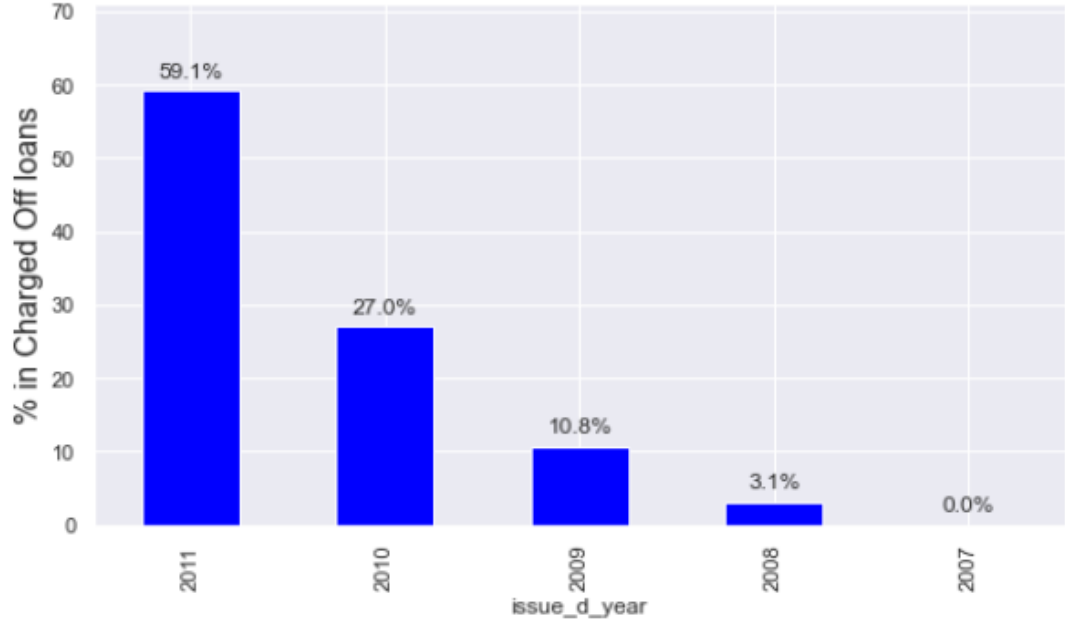




Univariate Analysis – Issue year

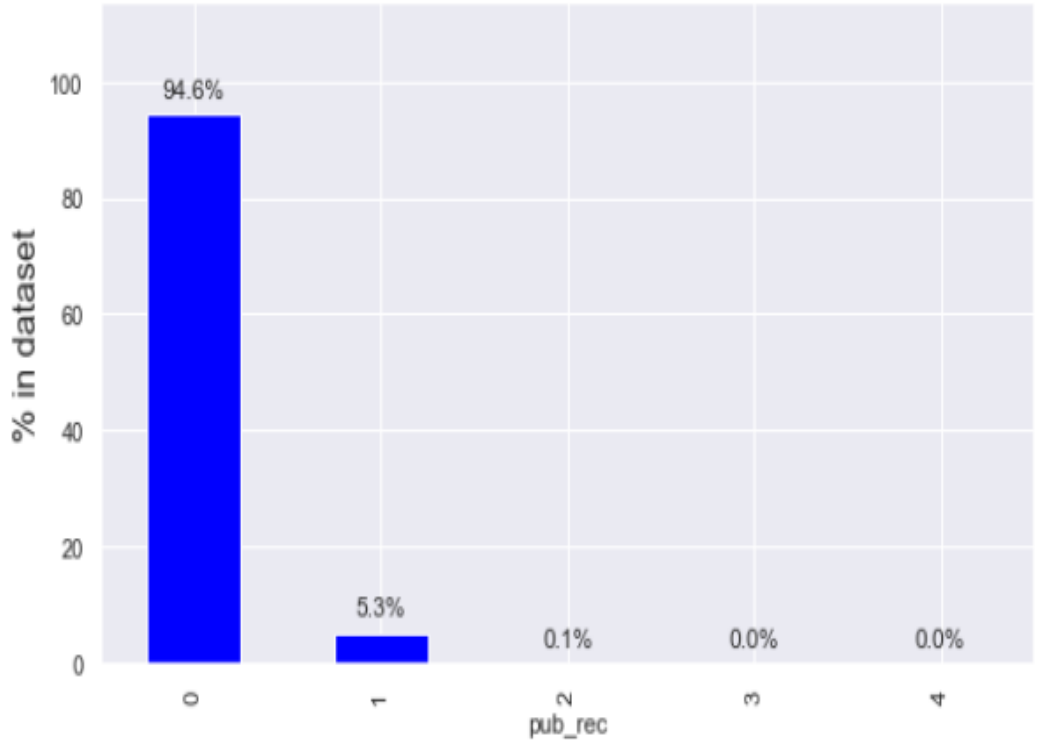
- There is no significant influence of loan issue year on loan default. Number of loan applications are increasing every year.
- Also there is significant increase in 2010 and 2011 years, might be corresponding to economic crisis.

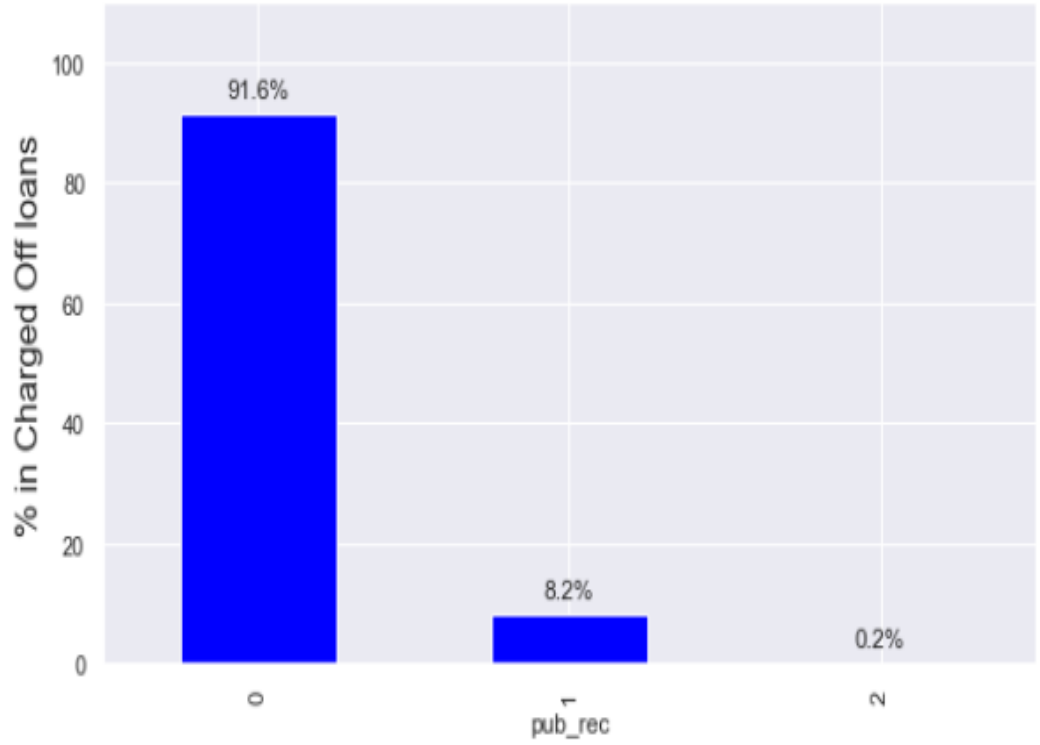




Univariate Analysis – pub_rec

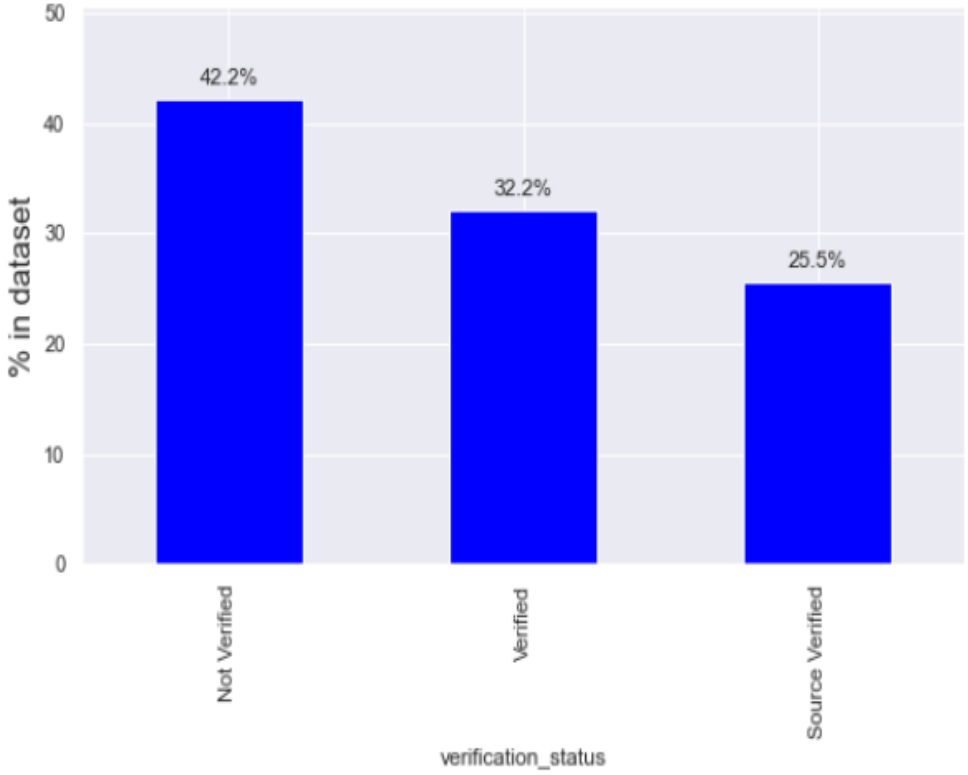
- We can observe that the loan applicants with more than one derogatory public record are likely to default
- Hence loan applications with prior derogatory public record should not be accepted

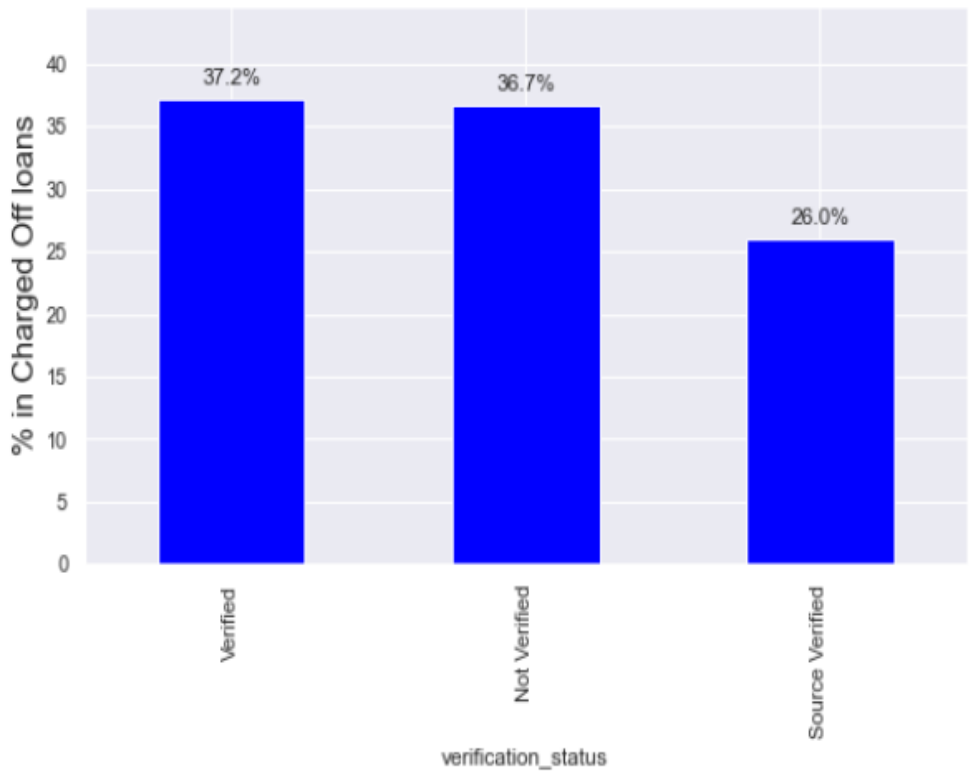




Univariate Analysis – verification status

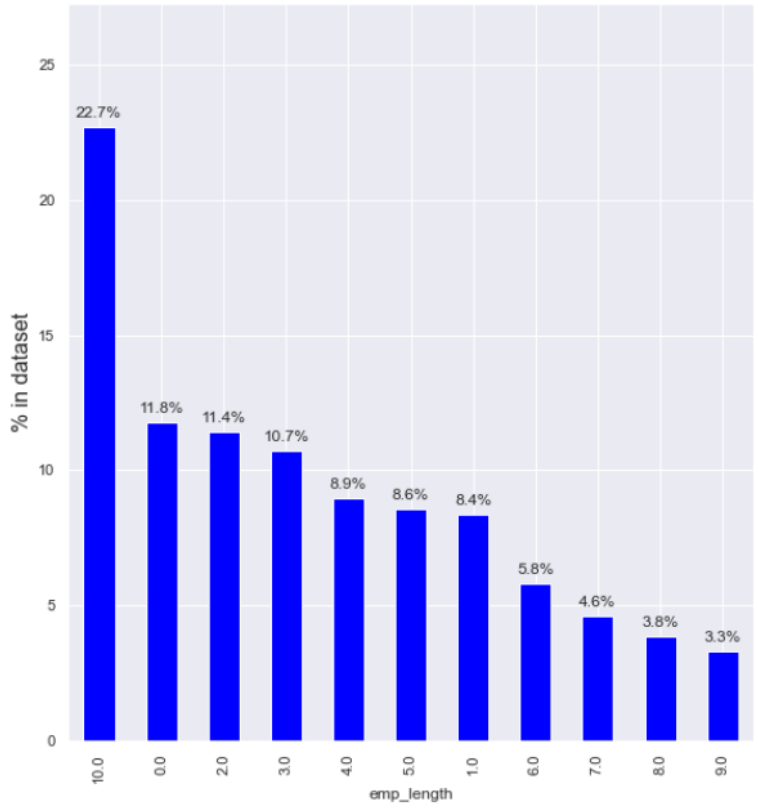
- Interestingly , loans which are verified are mostly defaulted compared to not verified.
- LC should revisit the verification process once.

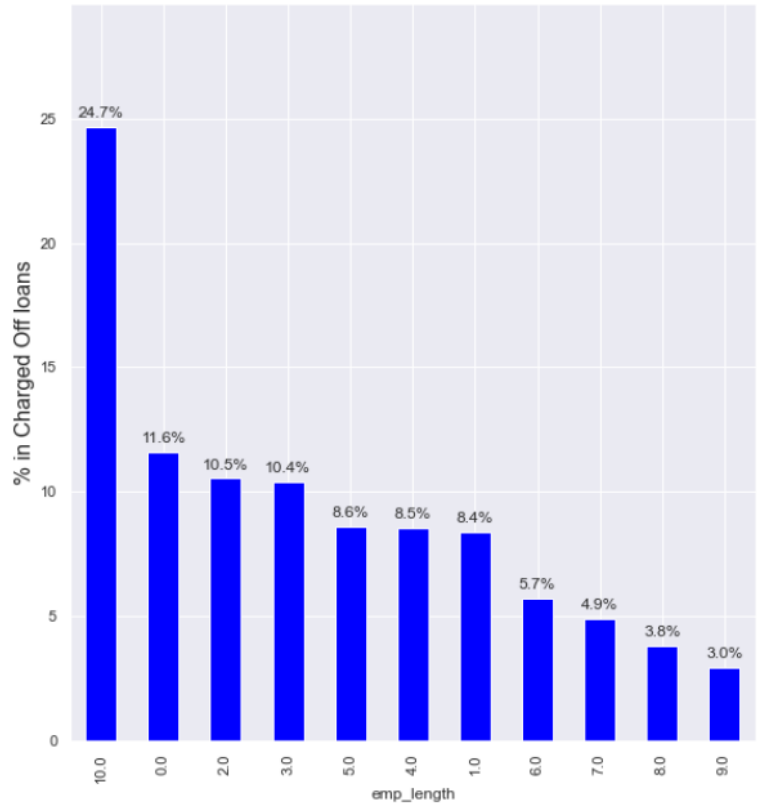




Univariate Analysis – employment length

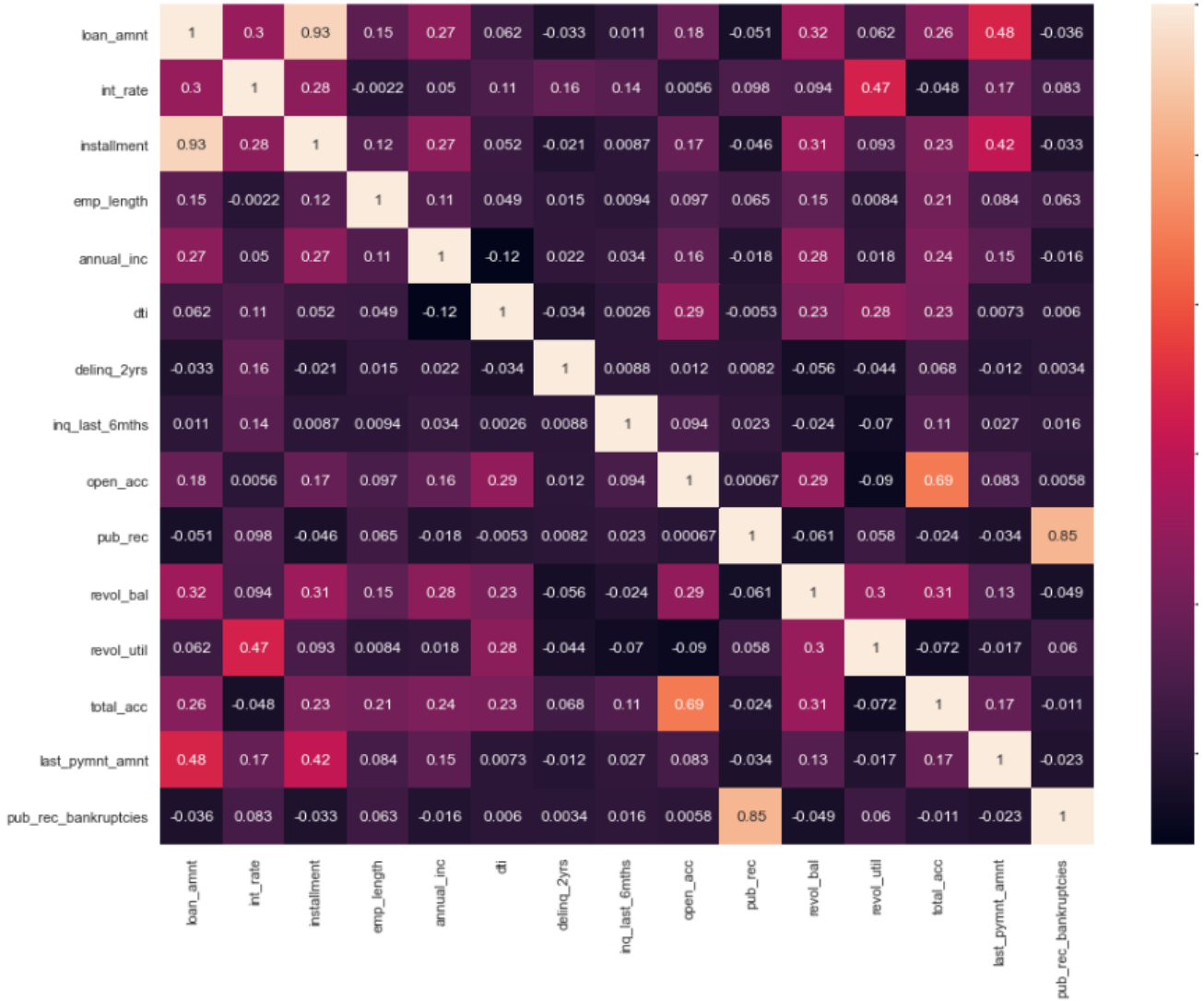
- Employee tenure has no bearing on loan default, while those with more than ten years' worth of experience take out more loans





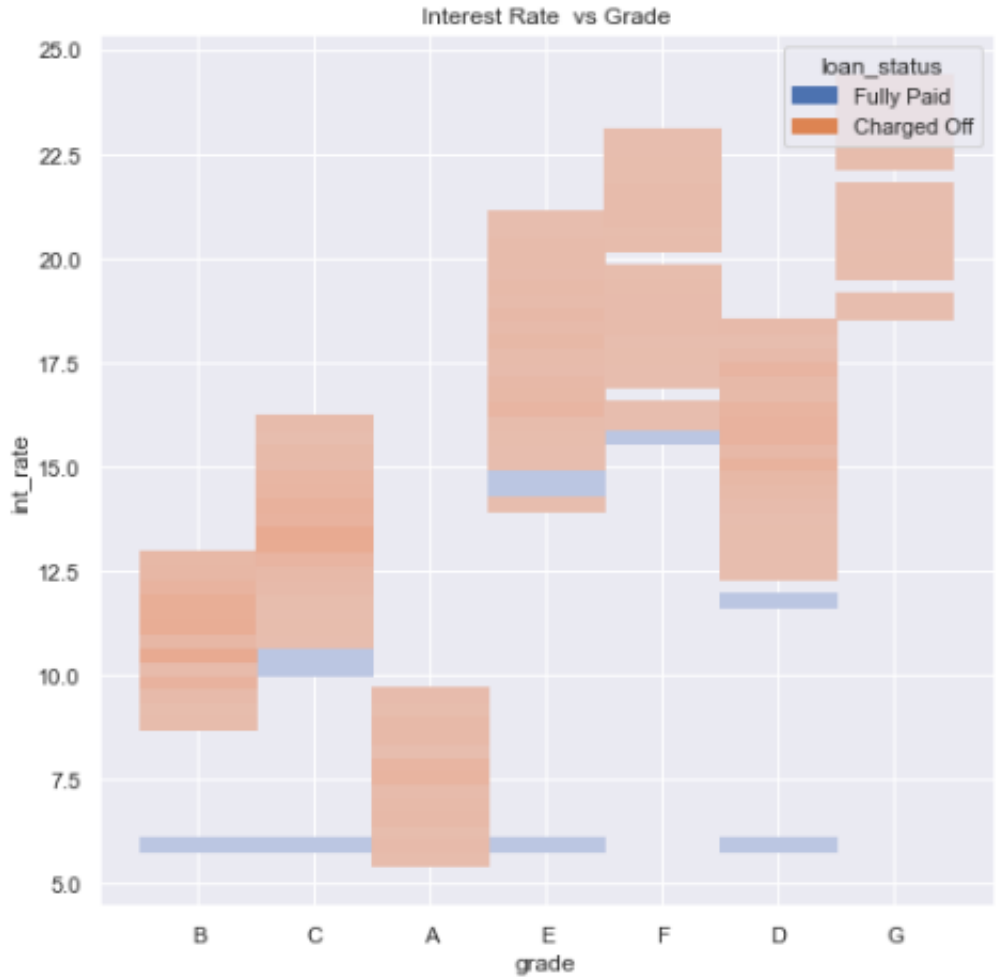
Bivariate Analysis – Heat Map

- loan_amnt is correlated to last_payment_amount with rfactor.48
- int_rate is correlated to revol_util with rfactor of .47. This is good, as company is charging higher interest from riskier loan.
- loan_amntrevol_bal are correlated with rfactor .32. This is not good as it suggests that higher loan amount is being approved to riskier borrowers.
- delinq_2yrs is totally uncorrelated with public record of bankruptcy. Therefore they represent distinct features with individual predictive value.
- loan_amnt is correlated with installmant with rfactor 0.93 which is expected



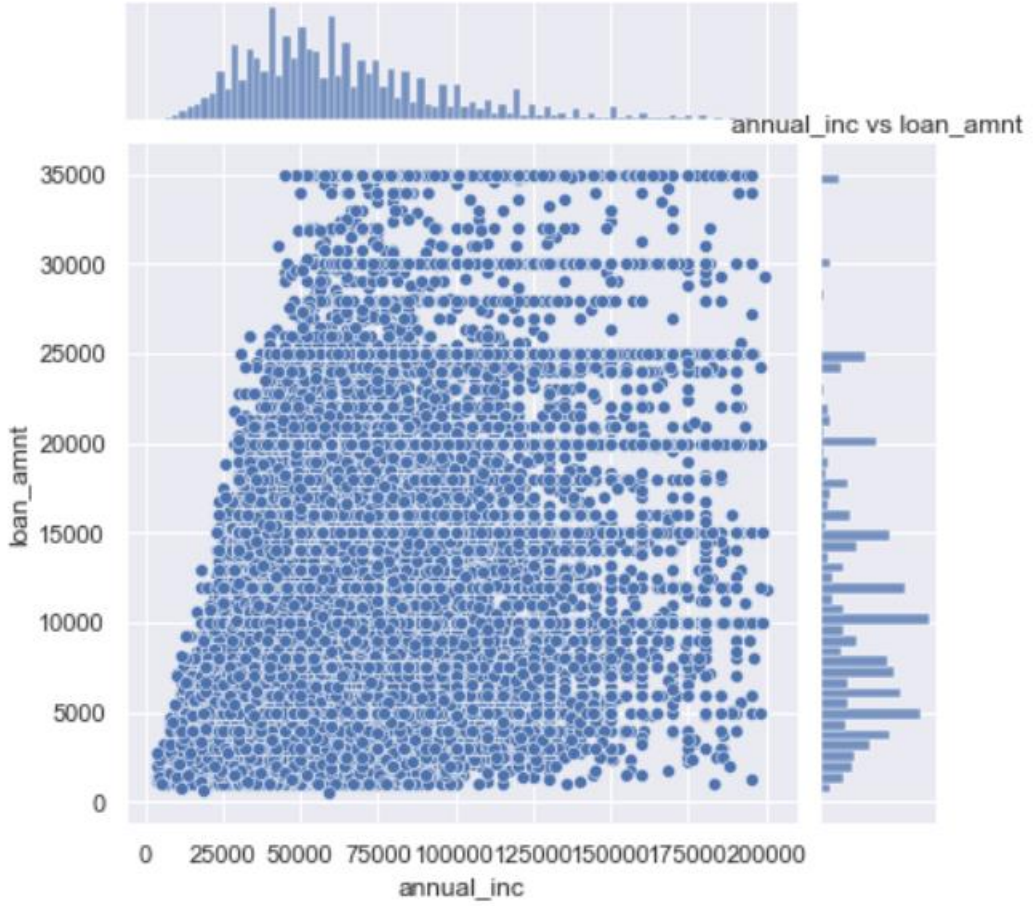
Bivariate Analysis – Grade- interest rate

- Grades B,C,A are having less interest rates and E,F,D,G are having more interest rates



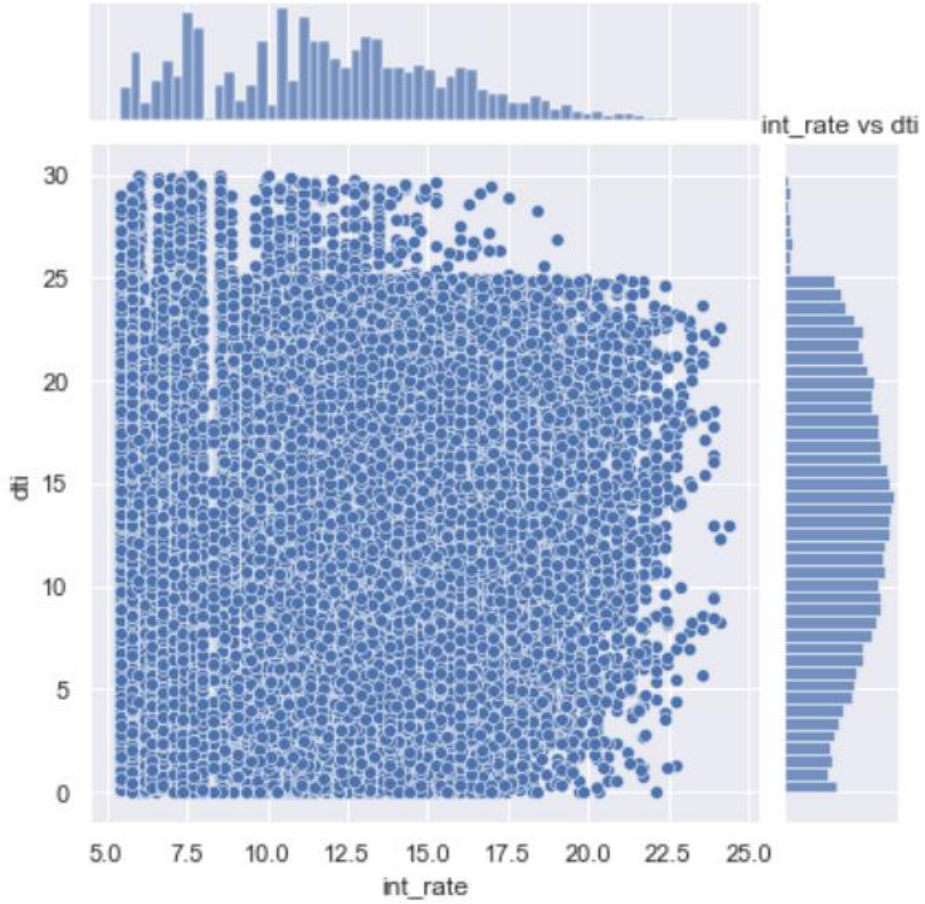
Bivariate Analysis – annual_inc vs loan_amnt

- There are people with average income lower than 50000 taking loans of 25000 or higher. These would be risky loans.



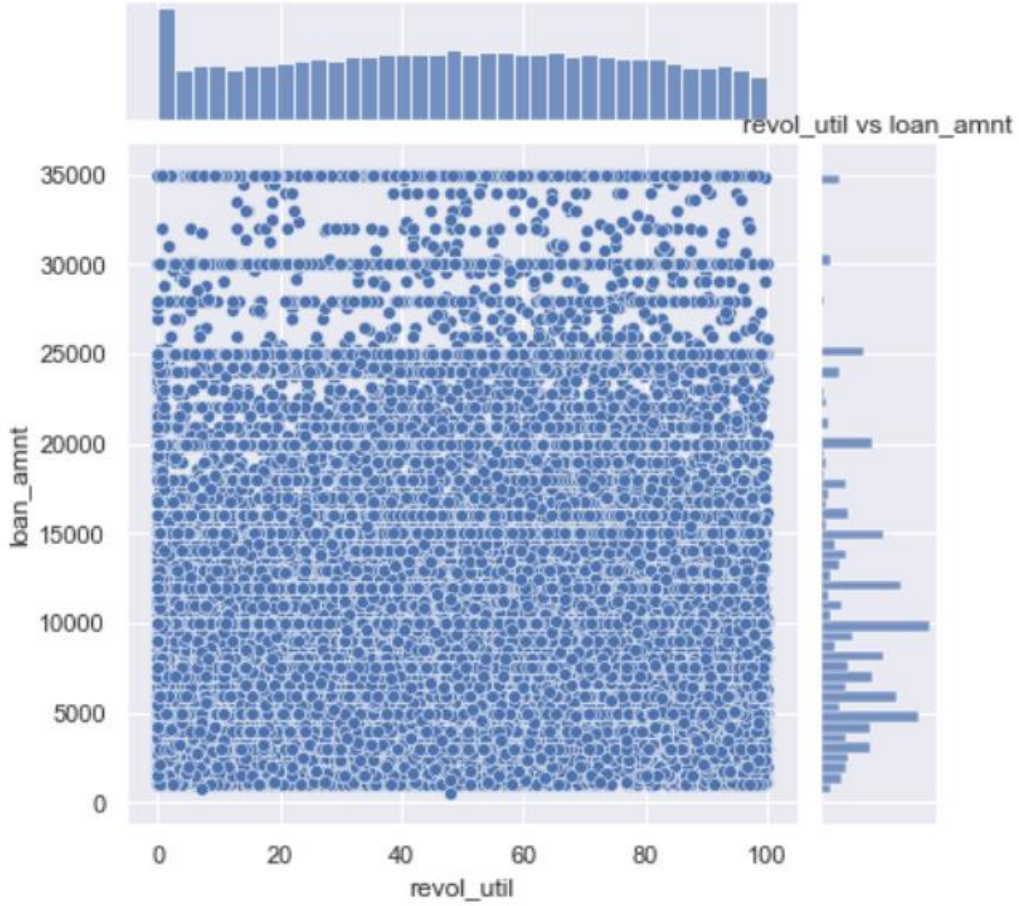
Bivariate Analysis – int_rate vs dti

- As projected, interest rates are rising till dti value hits the range of 15-20. Beyond this range, it actually starts to decline, which shouldn't be the case.



Bivariate Analysis –loan_amnt vs revol_util

- revol_util should not be approved when it is higher, however LC is clearly allowing huge loan amounts when it is between 65 and 75 percent.



SUMMARY

- Stop giving out large loans to those with modest annual incomes.
- Minimize the amount of approvals with a small business focus or raise the interest rate on these.
- Stop accepting loans for those having a history of defaulting on them, or at the very least, stop approving high-value loans.
- Stop accepting high-value loans when the revolving line utilization rate is above 75%.
- increasing the interest rates charged for loans with dti levels above 20