LENDING CLUB EDA CASE STUDY

SUBMISSION

Group Name:

1. Abhilash Mishra
2. Dibyanshu Pandey
3. Pramodini V Nayak

Abstract

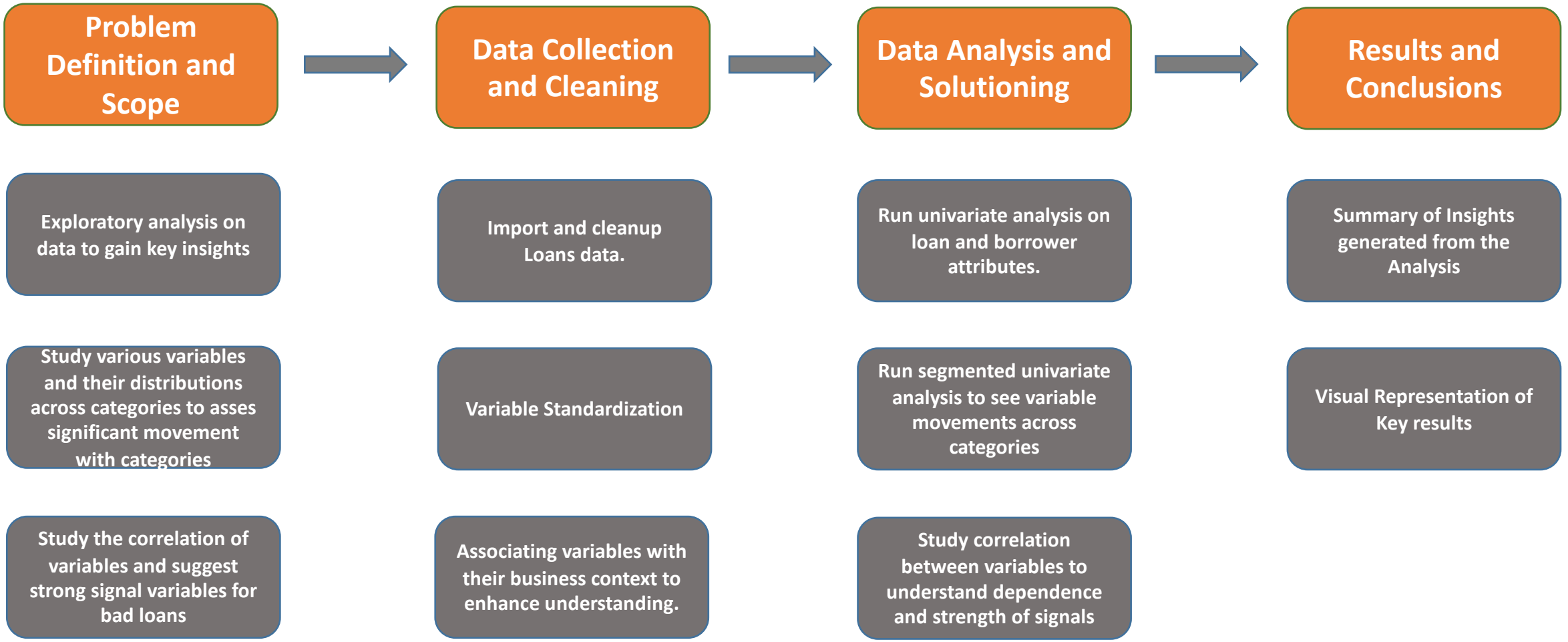
This study aims to explore a comprehensive loan dataset from Lending Club and derive insights from the same with the objective to narrow down on top drivers of bad loans.

Lending Club is the world's largest peer-to-peer lending platform. They offers lending service by connecting the borrowers with investors through an online marketplace. While this has became a popular source of passive income for investors, there is also risk of default, in which the borrowers fall behind on their repayments. In this study we use real market data from Lending Club to analyse over 39k loan transactions in the US along with numerous attributes and try to formulate a data backed approach to identify top drivers of bad loans. This could be used as features in models for risk assessment.

Our approach to solving this problem is quite simple and straight-forward. We pre process the data to get rid of redundant variables and features and then use univariate, segmented univariate and bi-variate analysis to identify variable distributions and patterns, We also study the dependency between set of two variables to find opportunities of correlating features that can predict risk.

Using this approach we finally narrow down on a set of features that can help us predict the probability of a loan going bad and hence minimising the risk for the lender.

Problem Solving Methodology



Analysis : High Level Call-outs

1.1 Loans Data at a Glance

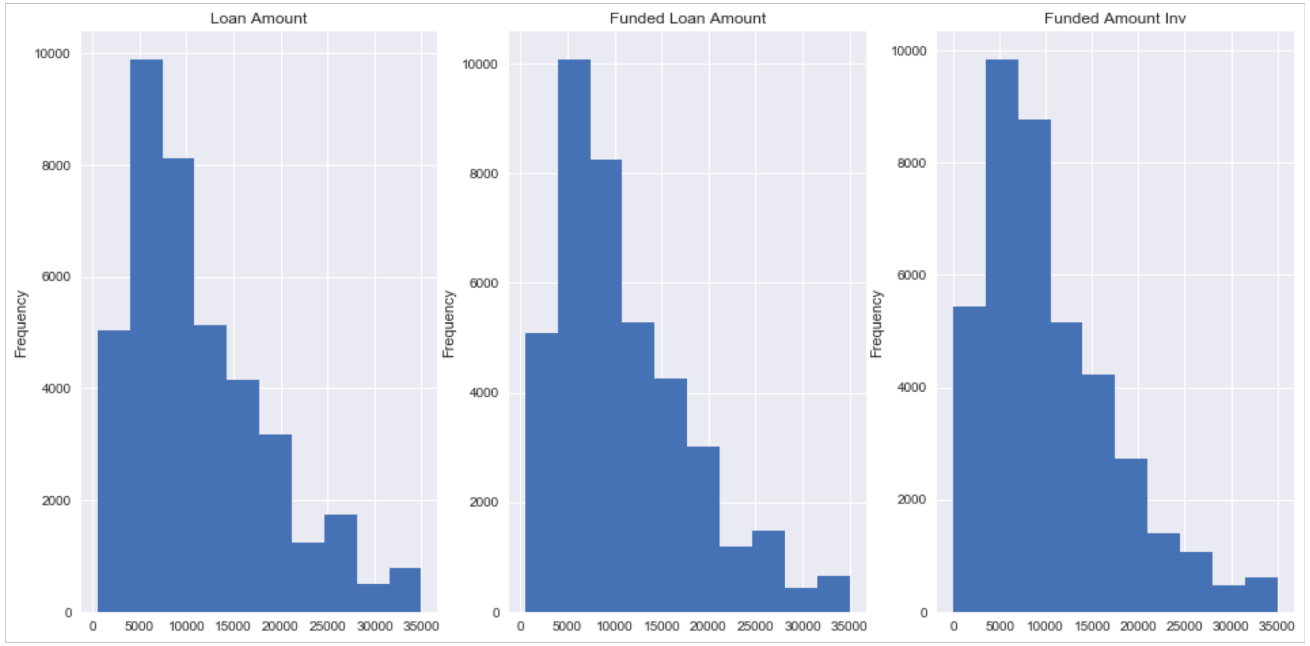
Total number of Accepted Loan Applications : 39717

Loan Status	Applicants	% of Total
Fully Paid	32950	82.96
Charged Off	5627	14.17
Current	1140	2.87

Loan Term

Term	Applicants	% of Total
36 Months	29096	73.26
60 Months	10621	26.74

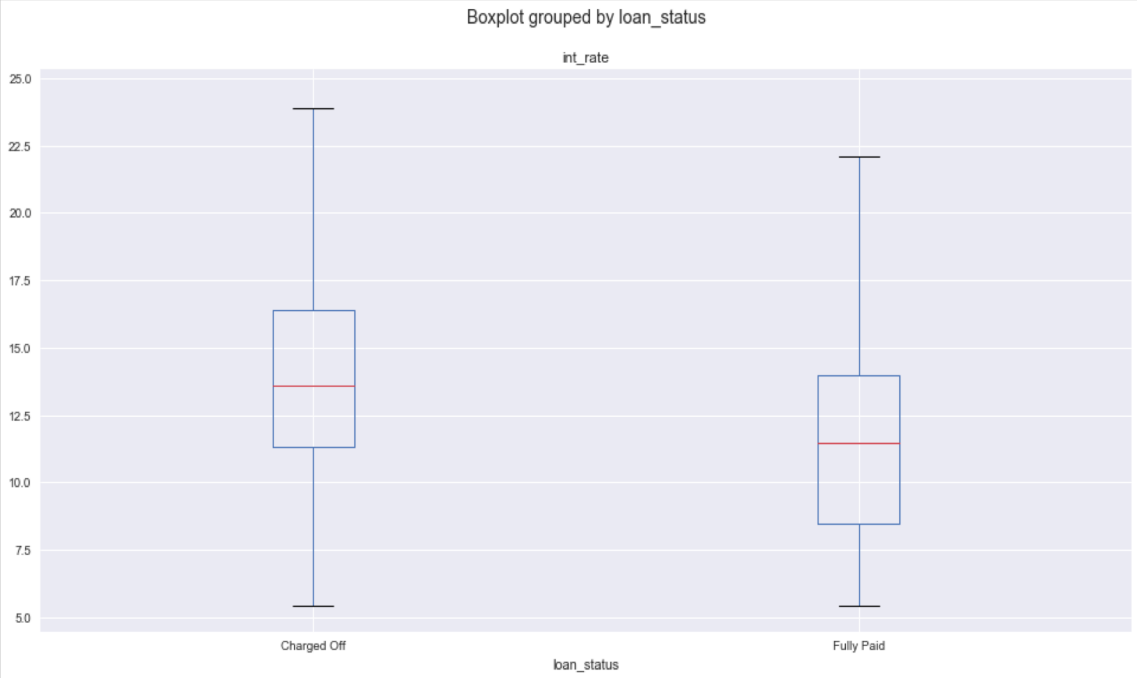
1.2 Distribution of Loan Amount



Most of the loan amount requested by borrowers is concentrated between 0 to 15000 USD. It also appears that the lenders fund the requested amount in almost all cases as the values are overlapping. Most loans fall in the 5k-10k USD bucket

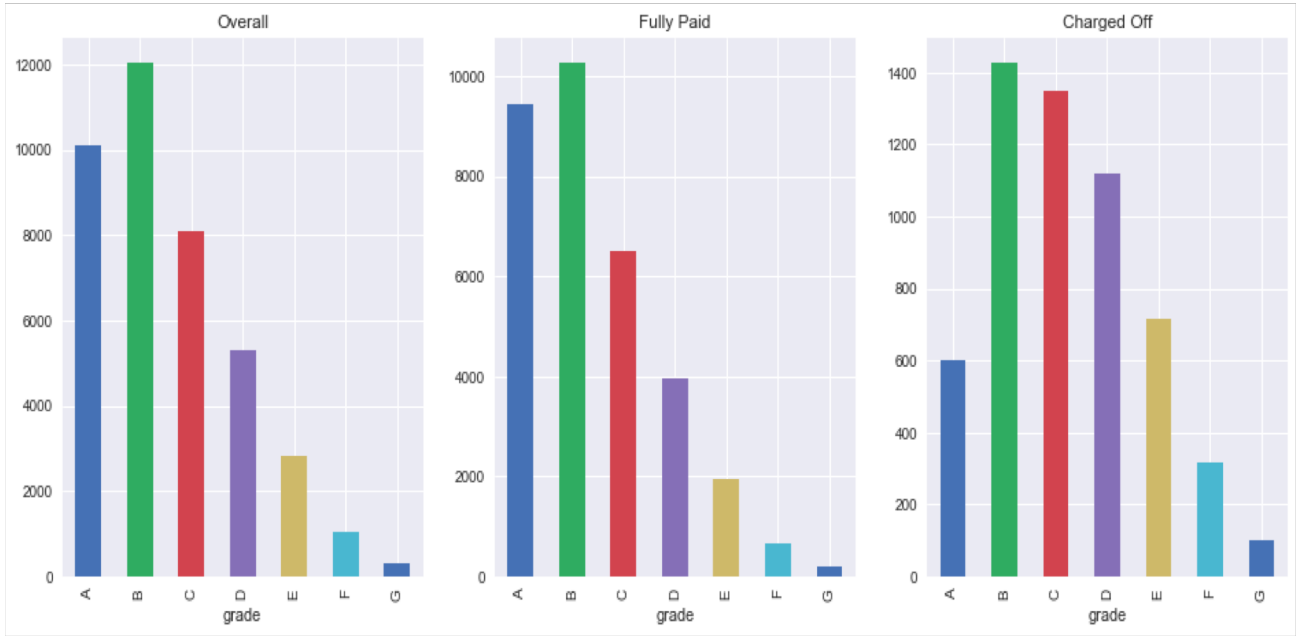
Analysis : High Level Call-outs

1.3 Distribution of Interest Rate across Loan Status



Very interesting results, we can see how the different the spread of interest rates is for Charged Off and Fully Paid loans. Defaulted or Charged Off loans have significantly higher interest rates when compared to fully paid loan. Looks like interest rate can be a good candidate to identify defaulters

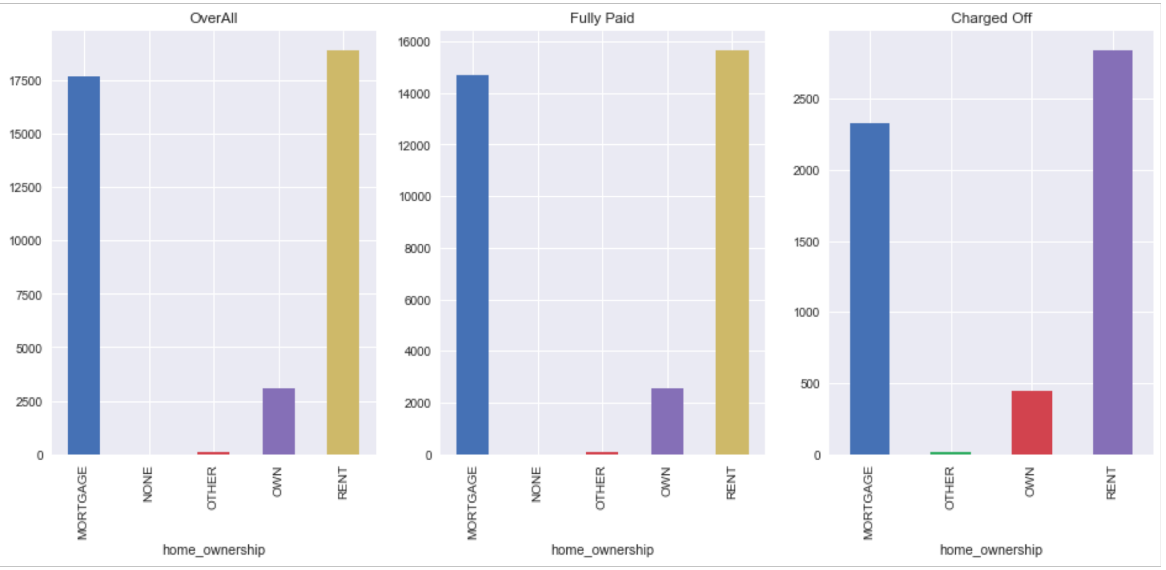
1.4 Borrower count across Loan Grades



Most(~75%) of the borrowers fall in Grade A,B and C. However, top defaulters categories in terms of absolute value are B, C and D

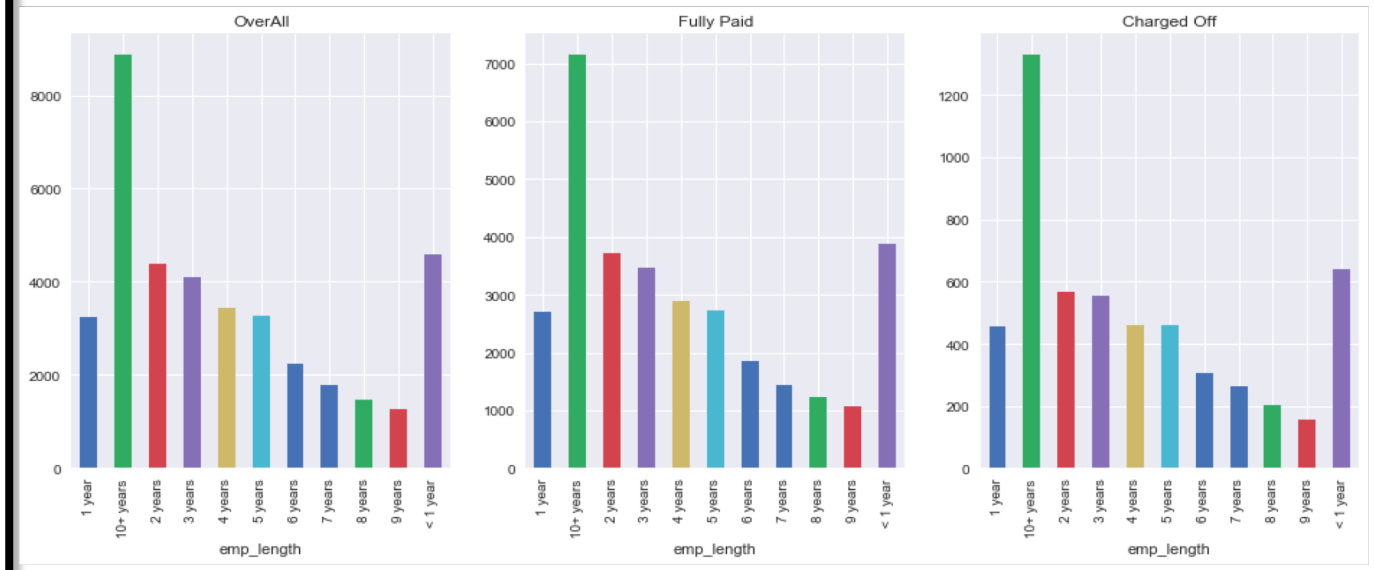
Analysis : High Level Call-outs

1.5 Borrower count across home ownership



Most of the borrowers have home ownership has either Rent or Mortgage. This can be said for both Fully Paid and Charged off loans in terms of absolute volume

1.6 Borrower count across Employment Length



Highest number of borrowers are people with more than 10 years of work experience, followed by people with <1 year of work experience. These two are the top categories for Non Defaults and Defaults as well, hence can't be said to be a strong signal for default without further analysis.

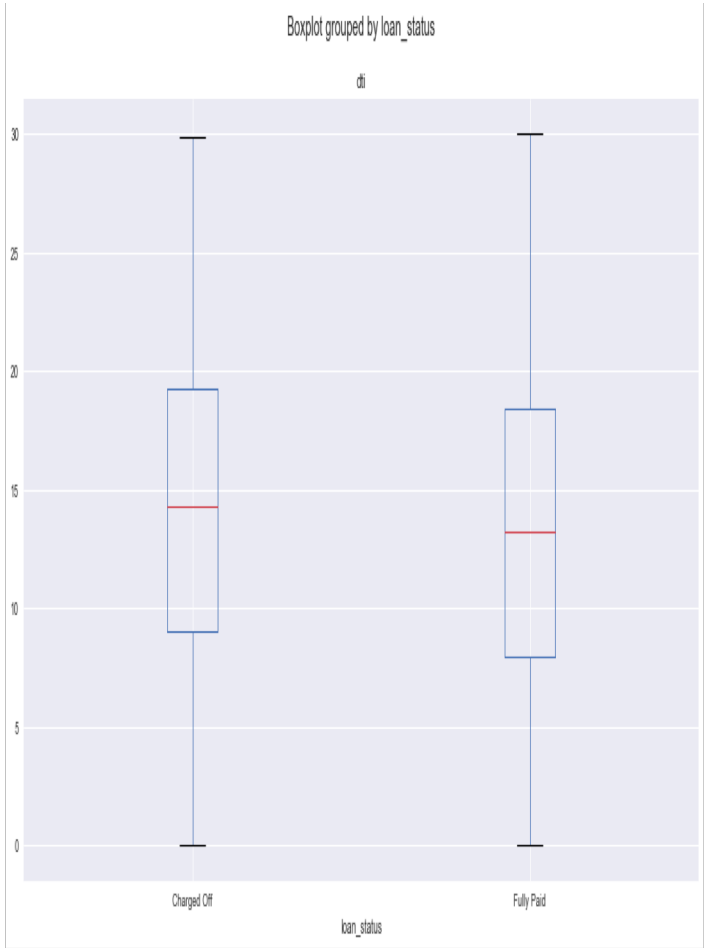
Analysis : Loan Default Signals

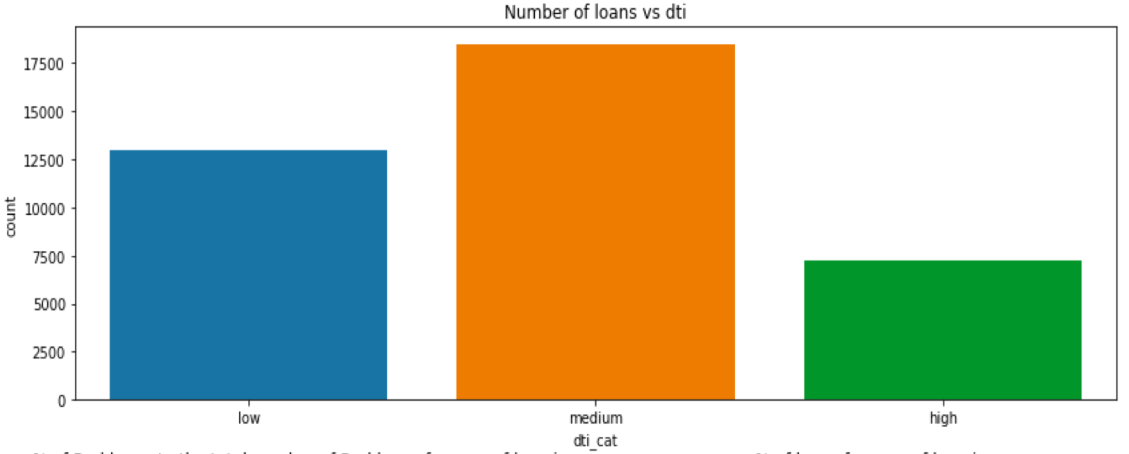
We carried out a thorough analysis of multiple variables to establish their signal strength to identify bad loans. In the following slides , we will be calling out only the relevant features that we found to be correlated to the probability of a loan going bad

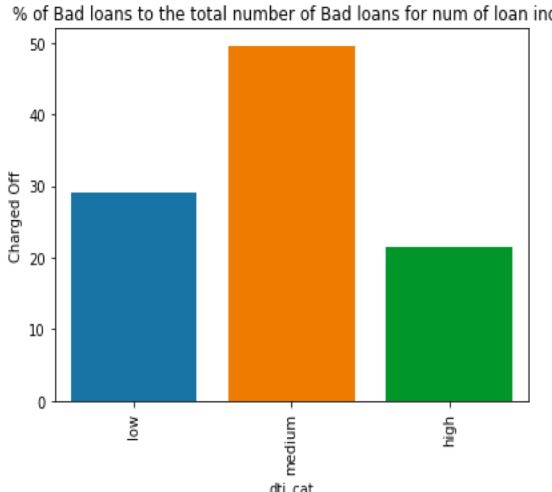
1. DTI

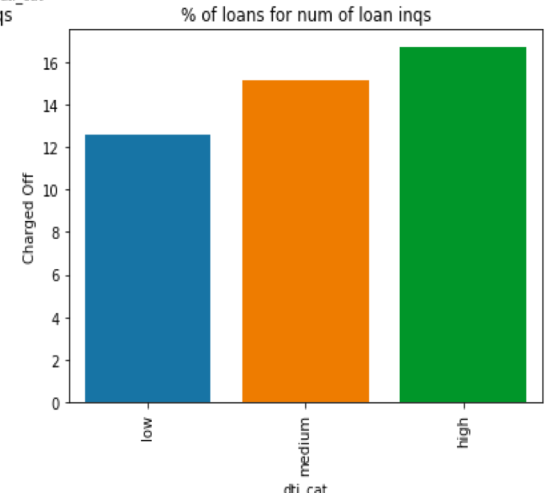
The boxplot shows the spread of DTI is higher for defaulted loans with the quartiles shifted above when compared to fully paid loans

There seems to be a trend in percentage of bad loans and dti. The higher the dti the higher is the percentage of bad loans. But the difference is not very high (about 1%)







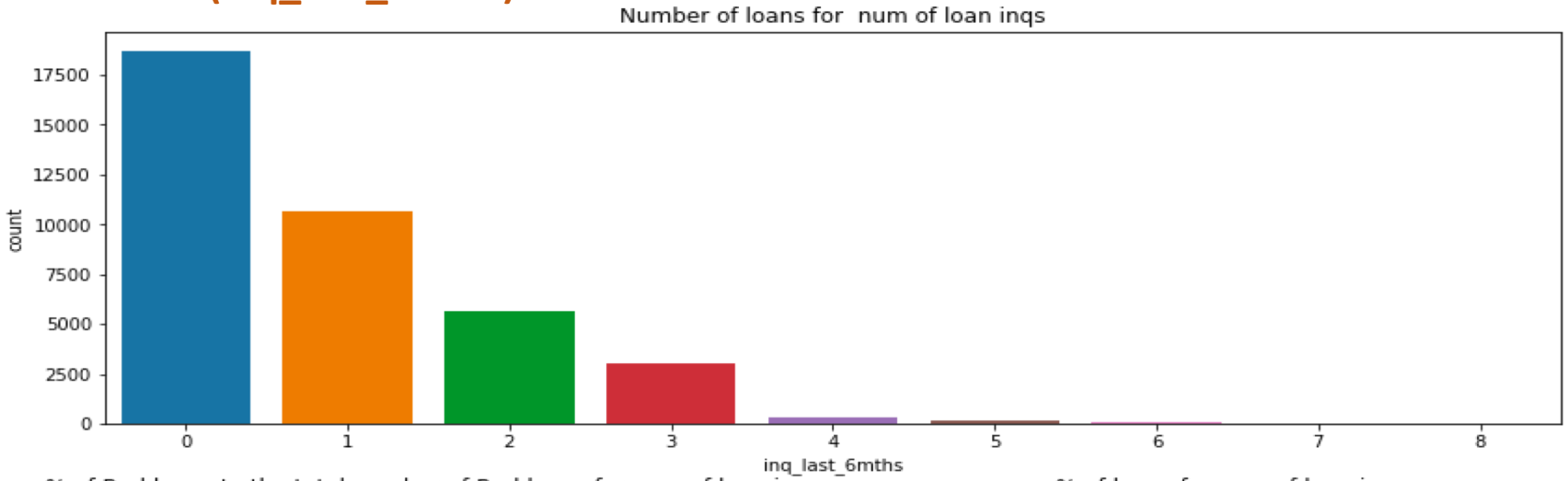


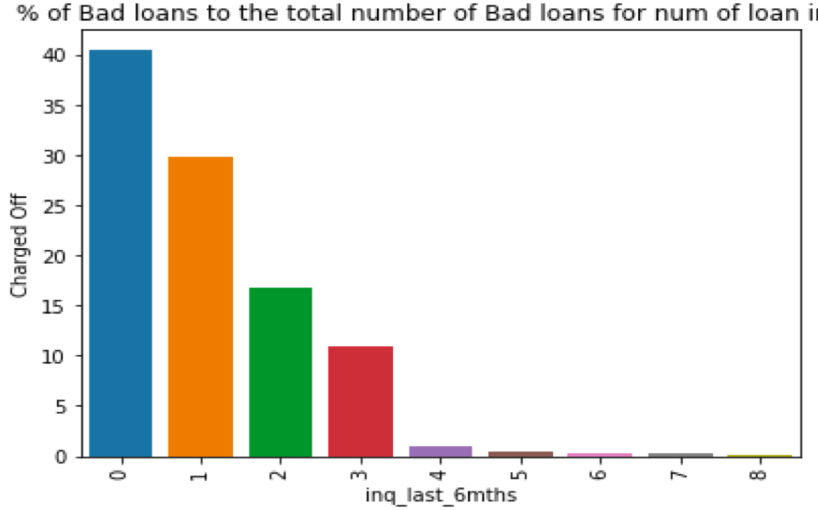
Analysis : Loan Default Signals

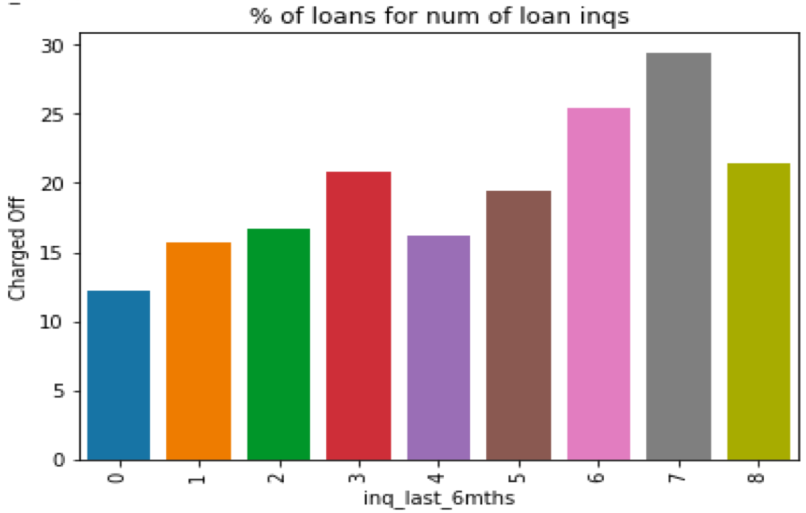
2. Number of loan inquiries in last 6 months (inq_last_6mths):

We can see that as number of loan enquires increases the percentage of bad loans tend to high (from plot 3

Default rate increases significantly as the number of inquires increases. If number of inquires are more than 2 then chances of loan being a bad loan are substantially high.





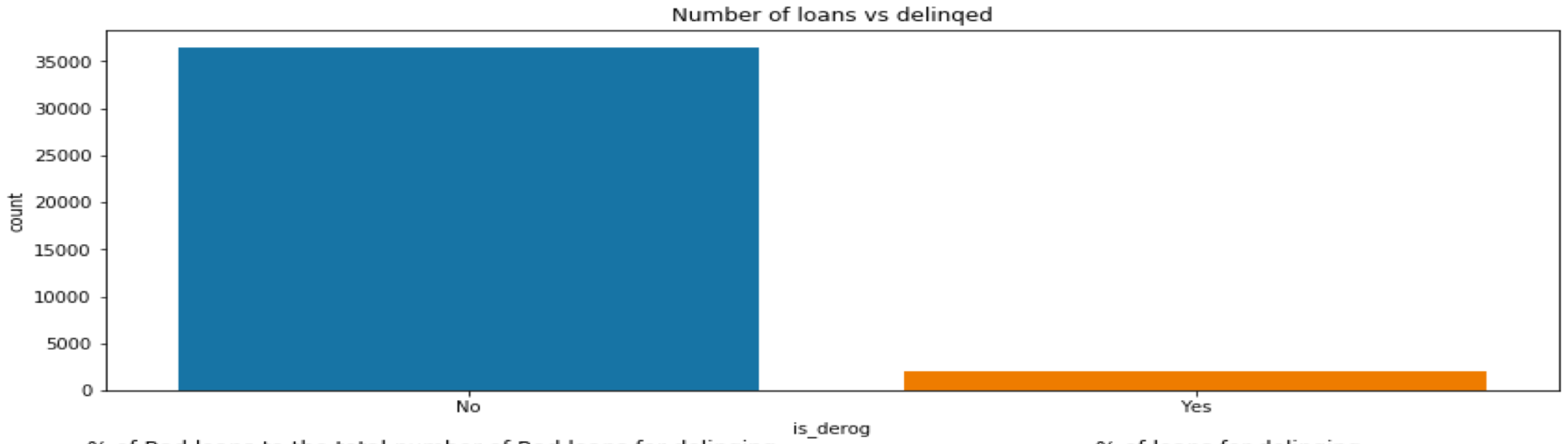


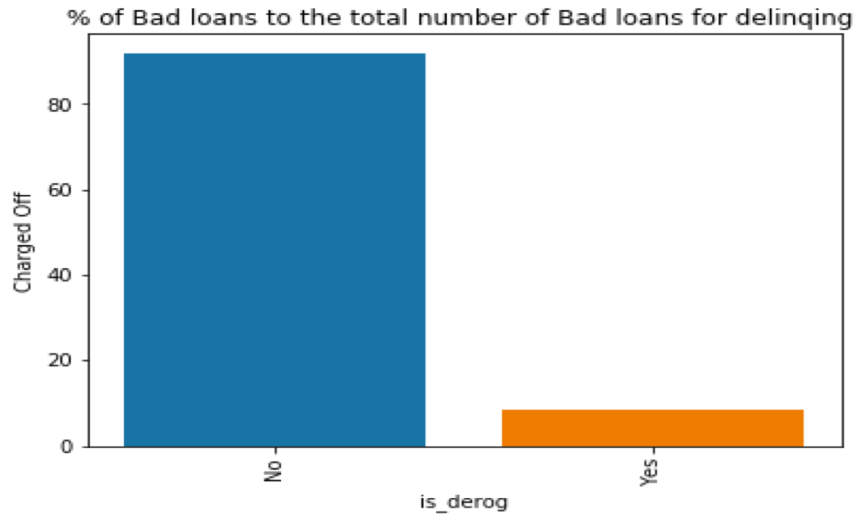
Analysis : Loan Default Signals

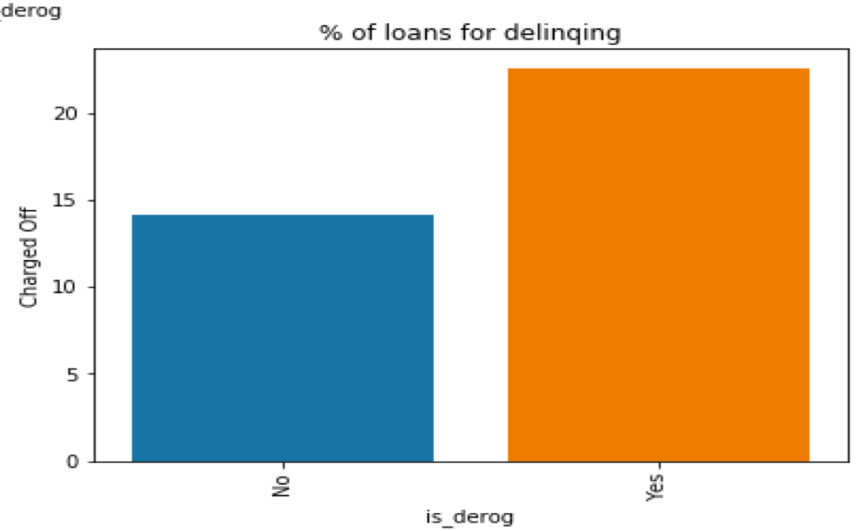
3. Number of derogatory records(pub_rec):

Most borrowers do not have a derogatory public record as can been seen in the top chart.

But if a borrower has a derogatory public record , there are high chances that he/she will default on their loan when compared to a borrower with no derogatory record.







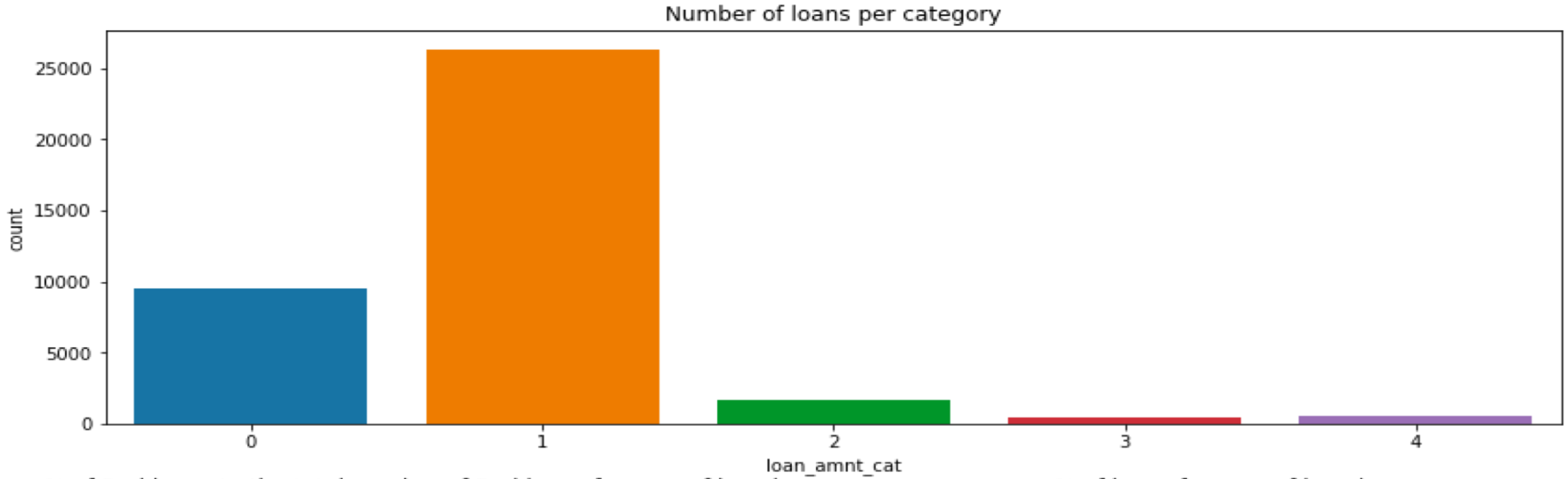
Analysis : Loan Default Signals

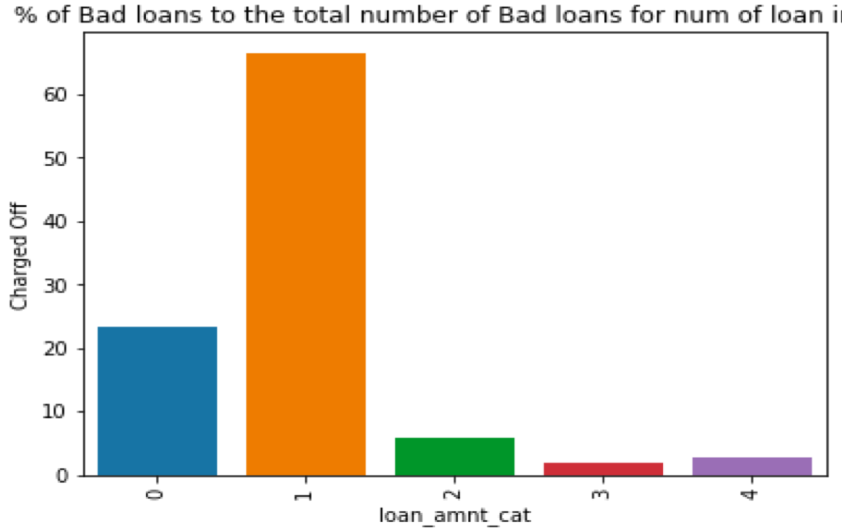
4. Loan Amount

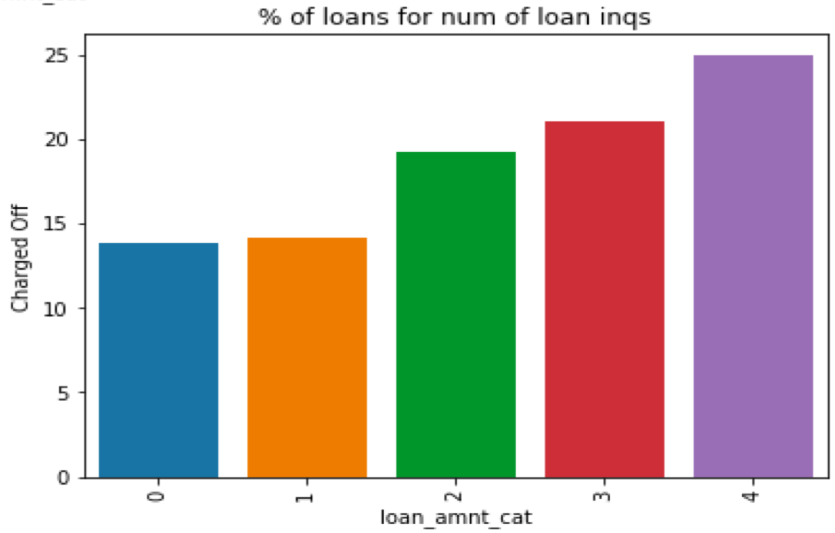
We segmented the loan amount into 5 different categories based on the amount

Most borrowers fall in category 1 : 5k-10k USD.

We can clearly see a trend that as loan amount increases the tendency of defaulting a loan also increases. This is partly also because of having higher interest rates as we move down the grades





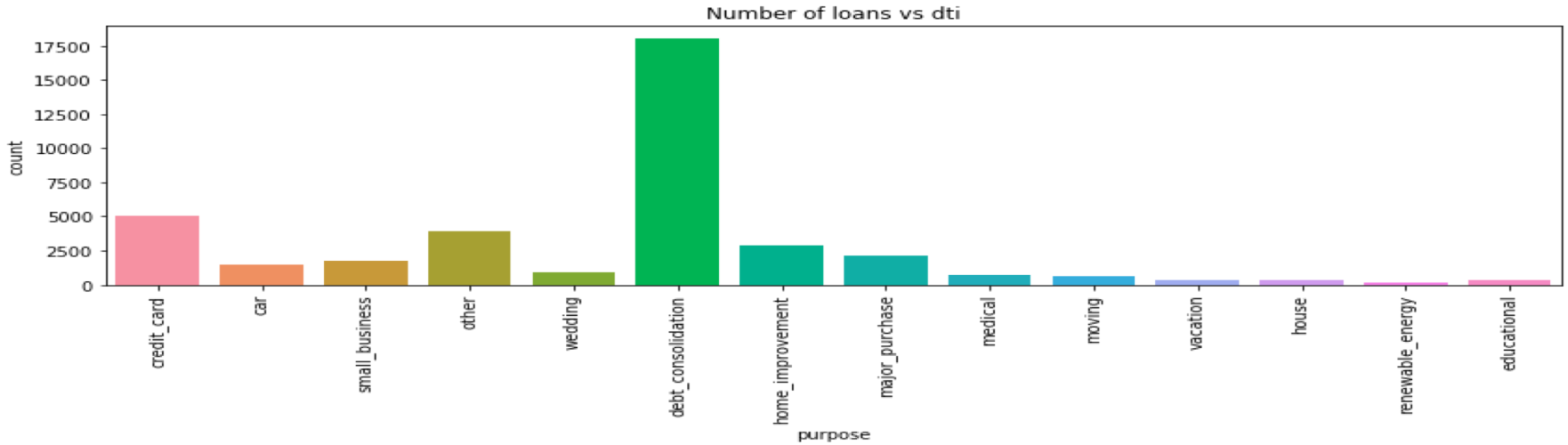


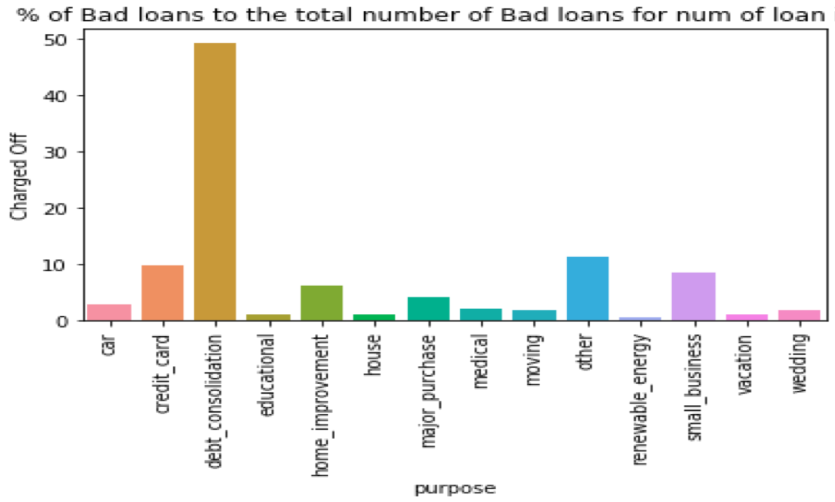
Analysis : Loan Default Signals

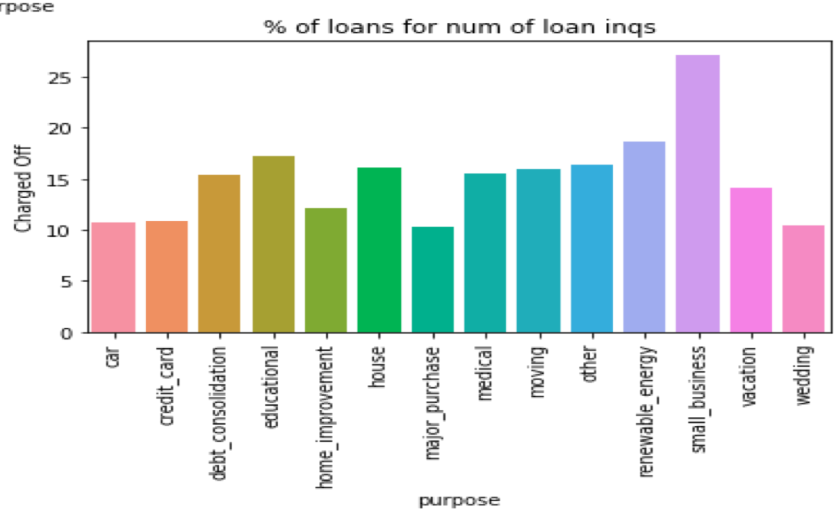
5. Purpose:

Debt Consolidation is the most common purpose of availing loans.

If the loan purpose is small business or Educational, the chances of them being defaulted on are significantly higher.(Plot 3)







Analysis : Loan Default Signals

6. Number of Open Credit Files:

We segmented the number of open credit files into three categories : Low(<5), Medium(>5 and <21), High(>21)

We can see that having open accounts less than 6 or more then 21 causes a higher percentage of loans to be defaulted