Lending Club Case Study

- Jaya Singh
and
Hiranmai

Agenda

- Problem Statement
- Tech Stacks Adopted
- Data Briefing
- Data Pre-processing
- Data Analysis Performed and relevant Inference
  - Histograms
  - Univariate & Segmented Univariate Analysis
  - Numerical variables analysis
  - Demographic analysis
  - Numerical or Categorical Variable Analysis wrt Target Variable
  - Analysis between 1 numerical and 1 categorical variable
  - Bivariate & Multivariate Analysis
  - Heatmap analysis
- Conclusion

Problem Statement

In the recent years, certain realities are highly visible in the Consumer Finance Company. The company is obliged to lend loans to customers, flying in loan application. Although, post thorough verification of the applicant has been done, there are situations where the company has to encounter the “defaulters” or “charged-off” candidates, who are unlikely to repay their loan. Thus, such failures to fulfill leads to financial loss for the company.

Hence, with the given dataset, aim is to identify patterns and/or factors which can help the company to either accept or deny lending loan to applicant/customer.

Tech Stacks Adopted

- python == 3.10
- numpy == 1.25.2
- pandas == 2.0.3
- matplotlib == 3.7.3
- seaborn == 0.12.2
- os
- warnings
- math

Data Briefing

- No. of rows: 39717
- No. of columns: 111
- data type of values: differs from column to column; either numeric or alphabetical or alphanumeric or NAN
- Multiple columns with NaN values
- Some columns with 92% missing values

Data Pre-processing

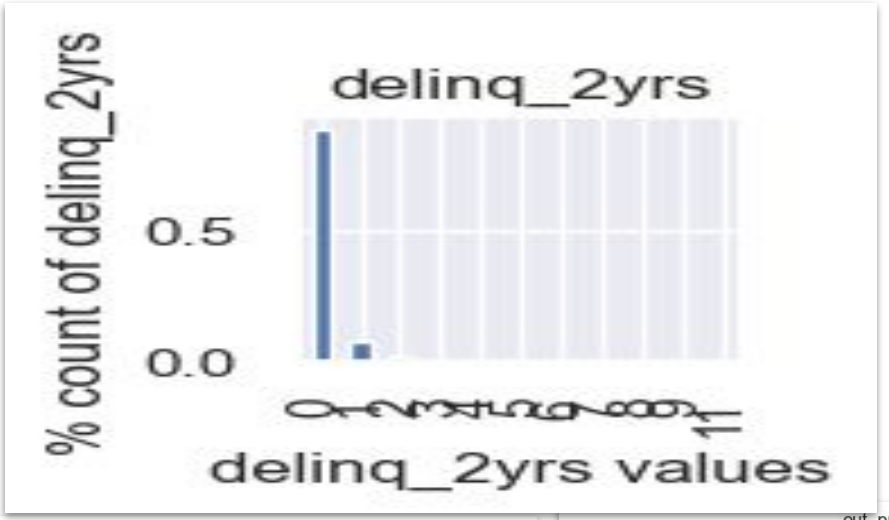
- Dropped columns with more than 25% missing values, hence reducing the number of columns to 53
- Dropped columns with lesser unique values, resulting in 42 columns
- Corrected parsing of the year
- Performed imputing for missing values in categorical columns
- converted values in categorical columns to proper data types
- Imputed for missing values in numerical column with median
- Removed outliers using IQR approach

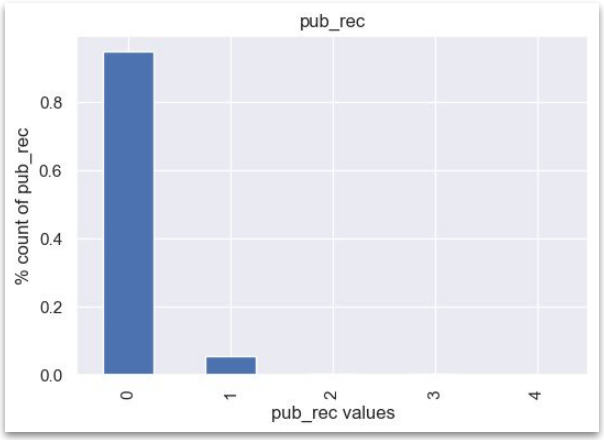
Data Pre-processing

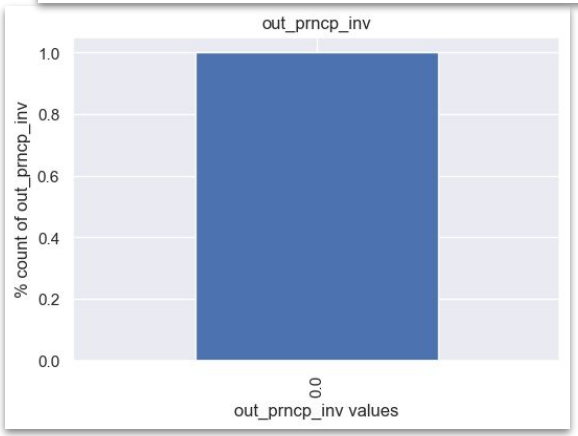
- Grouping columns by their usage type:
  - **Ignored and dropped columns** = ['id', 'member_id']
  - **Further dropped columns** = ['pymnt_plan', 'policy_code' , 'initial_list_status', 'collections_12_mths_ex_med', 'application_type', 'acc_now_delinq', 'chargeoff_within_12_mths', 'delinq_amnt', 'tax_liens', 'zip_code', 'url']
  - **Target variable** = ['loan_status']
  - **Categorical variables** = ['term', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'verification_status', 'purpose', 'addr_state', 'loan_month', 'loan_year']
  - **Tenure variables** = ['issue_d', 'earliest_cr_line', 'last_pymnt_d', 'last_credit_pull_d']
  - **Demographic variable** = ['addr_state']
  - **Numerical variables** = ['loan_amnt', 'funded_amnt', 'int_rate', 'installment', 'annual_inc', 'dti', 'delinq_2yrs', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc', 'out_prncp', 'out_prncp_inv', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt']

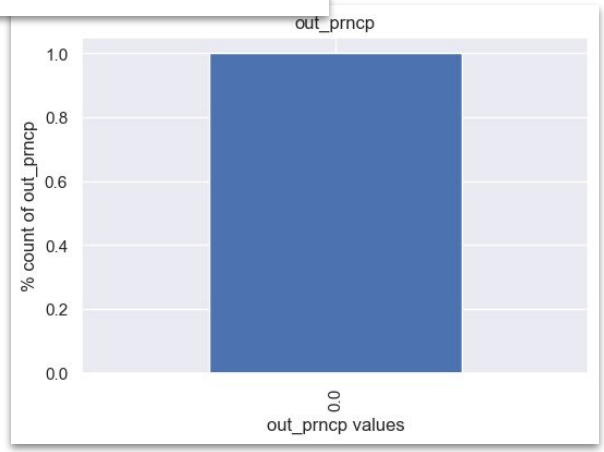
Data Analysis

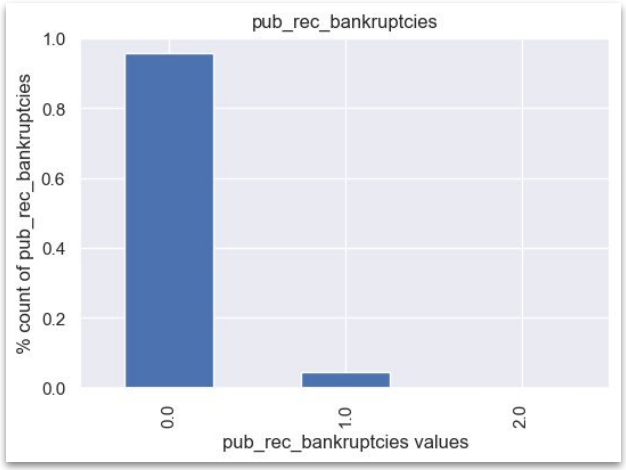
Histograms







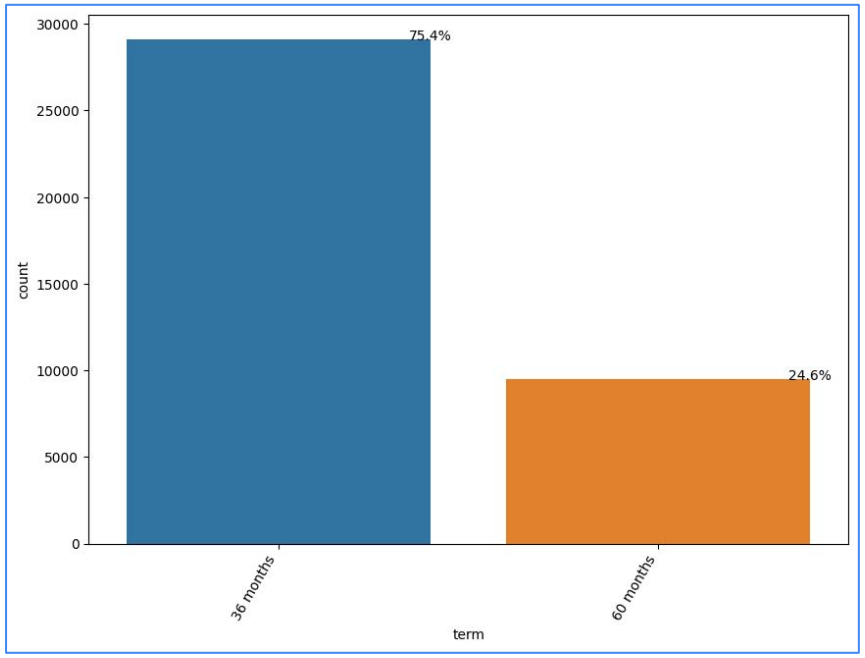


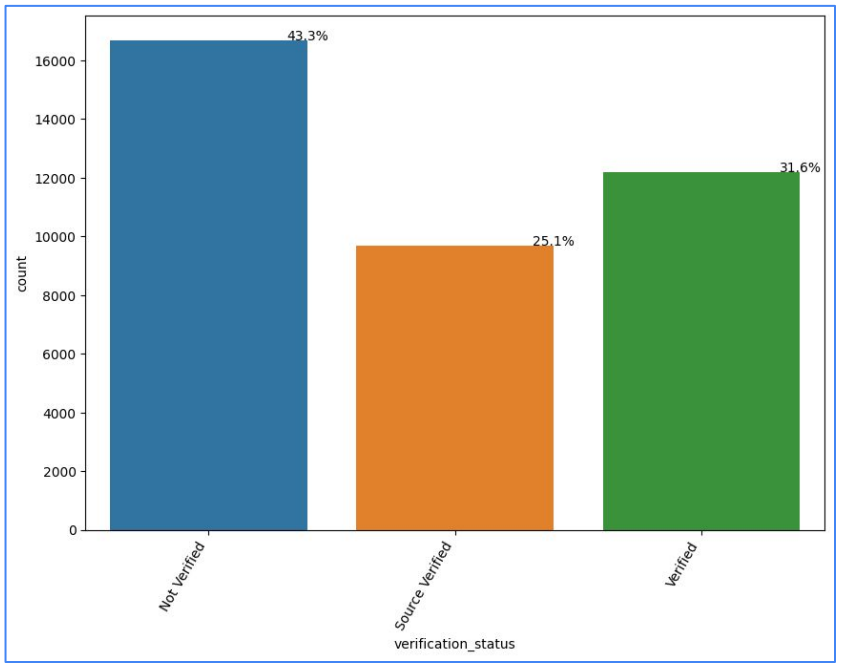


we see that columns delinq_2yrs, pub_rec, out_prncp, out_prncp_inv, pub_rec_bankruptcies, recoveries, collection_recovery_fee majority of the values are 0 and hence these columns are dropped from analysis as it wont result in meaningful analysis

Data Analysis

Univariate & Segmented Univariate



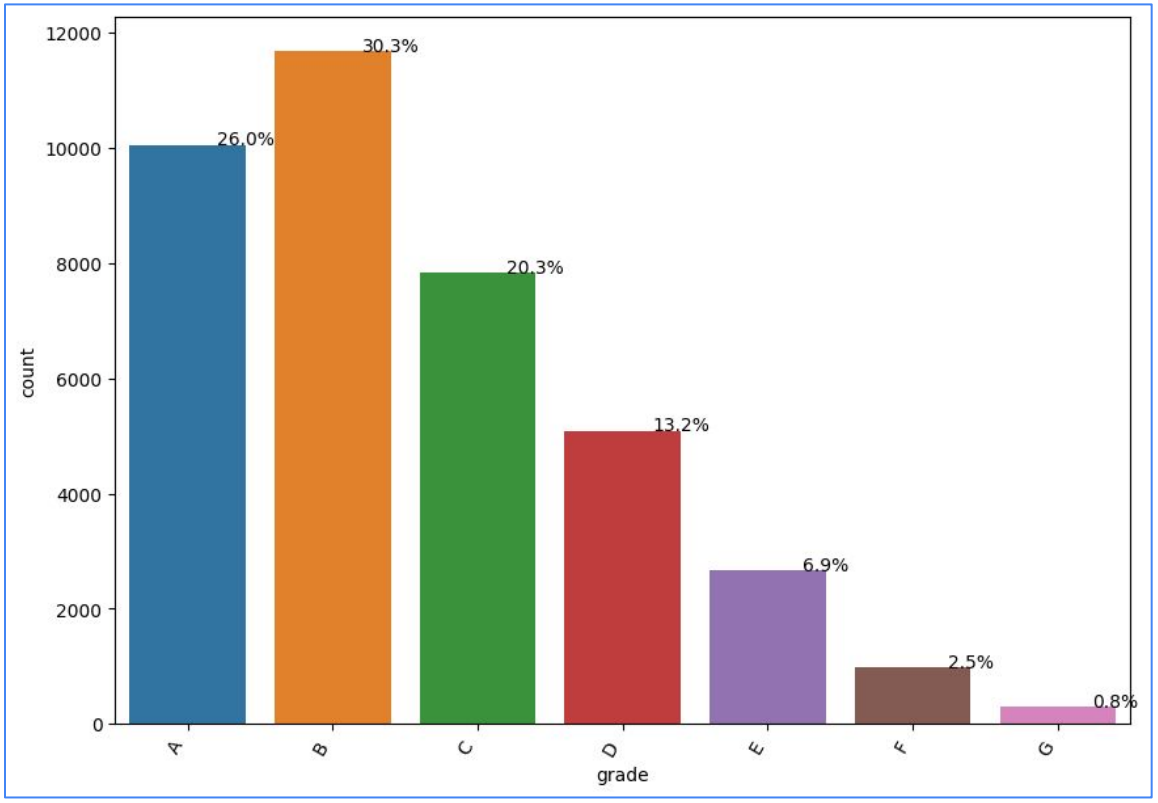


36 months has higher weightage compared to 60 months. More loans have been given for shorter duration

From the data pattern, it seems that majority of loans or defaulters are because these customers where not verified.
Customers whose income source was verified or were verified by LC they have lesser chances of defaulting.

Data Analysis

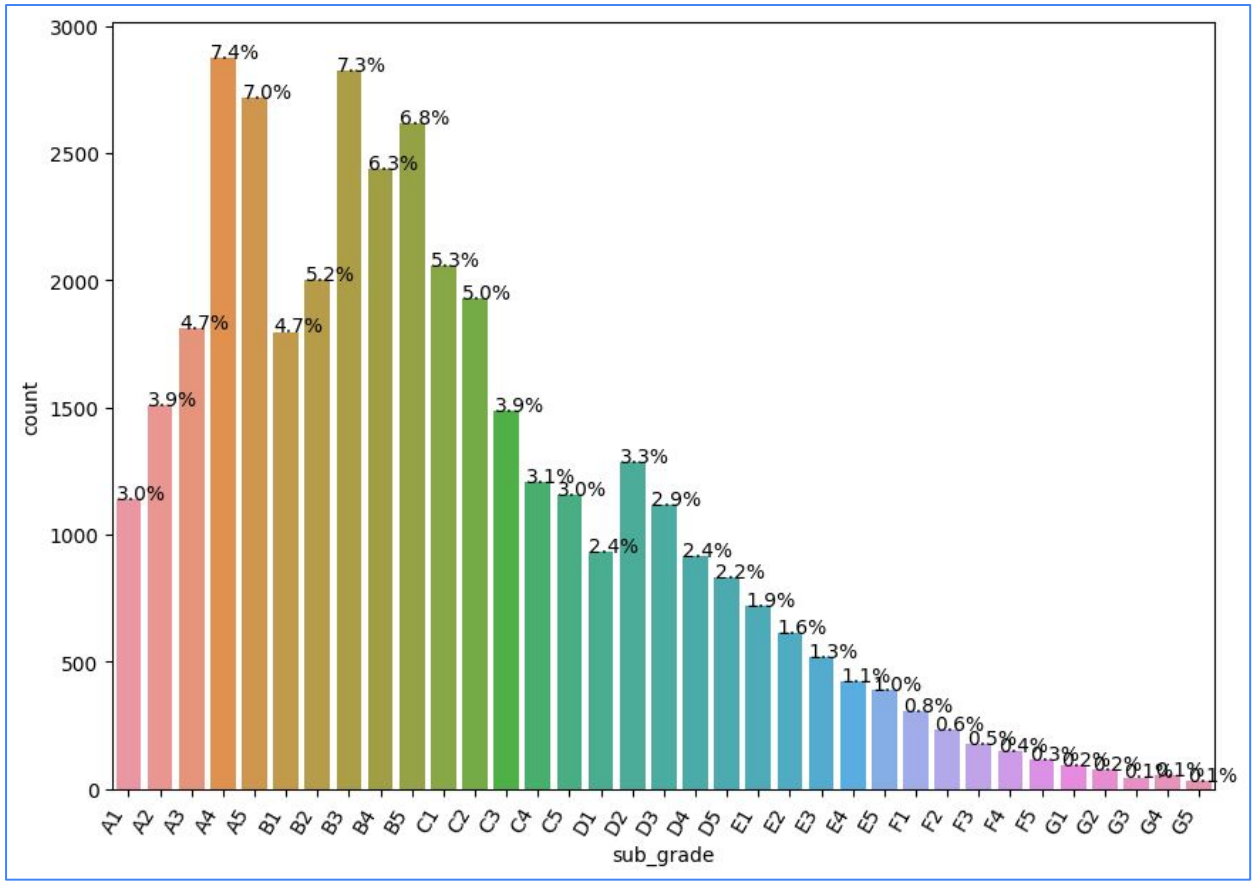
Univariate & Segmented Univariate



Grade Column: People belonging A and B grades have been favoured compared to other grades. This might be because A and B represents customers at a higher level(e.g VP, AVP etc.) compared to others.

Data Analysis

Univariate & Segmented Univariate

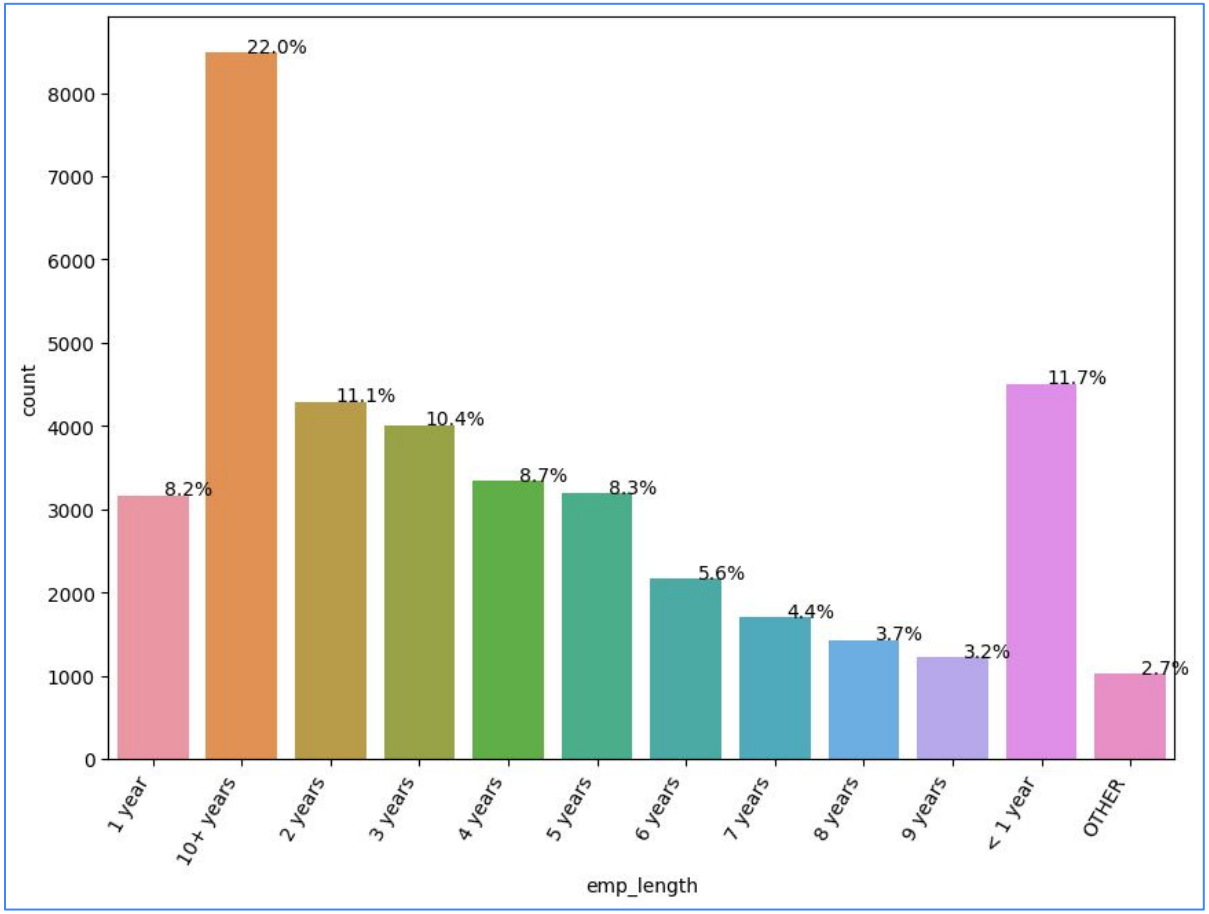


Each grades have then mutiple sub-grades. Among those, A4, B3, A5 have been favoured more compared to others. Followed by B5, B4.

From the data it seems like within subgrades, customer level follows ascending order. 5 is highest then 4 and so on.

Data Analysis

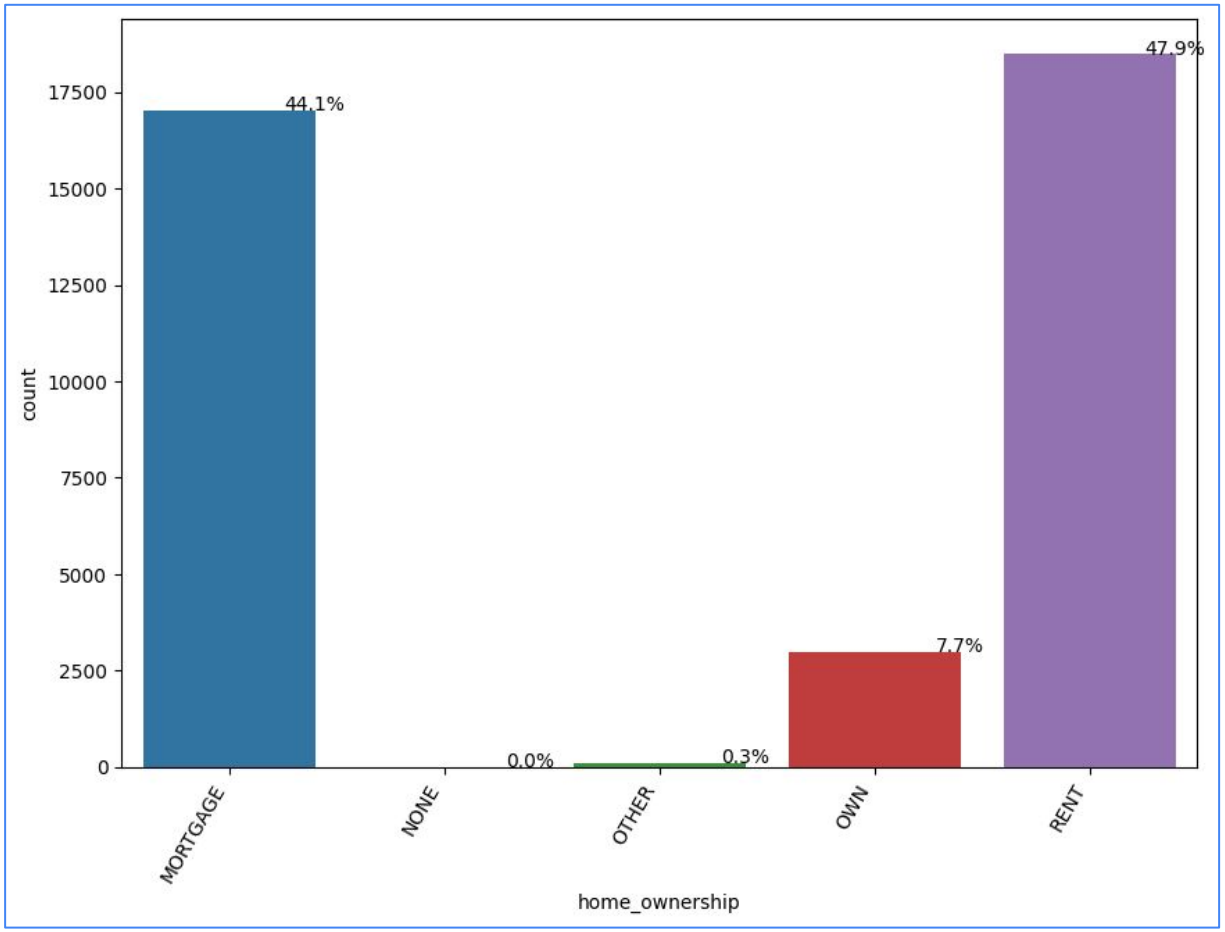
Univariate & Segmented Univariate



Customers who has more employment tenure are favoured. These customers are the best suited for lending Clubs as these customers will have consistent source of income resulting in repayment of loans.

Data Analysis

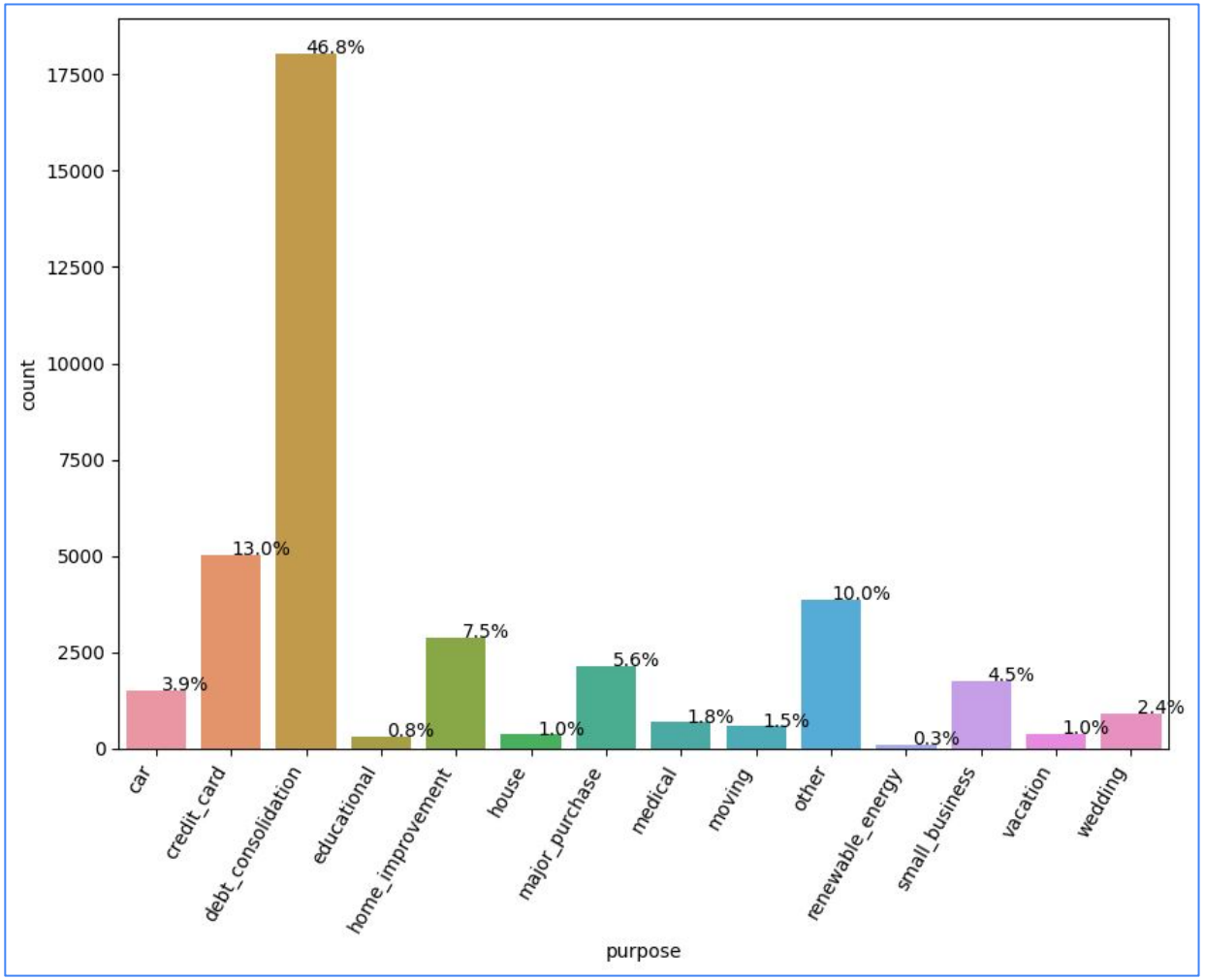
Univariate & Segmented Univariate



in this case, customers who are renting or paying mortgage have higher changes of taking loans. This could also be because they would want to or already purchased homes. Customers who owns the home mght not need to borrow from the banks/ lending clubs

Data Analysis

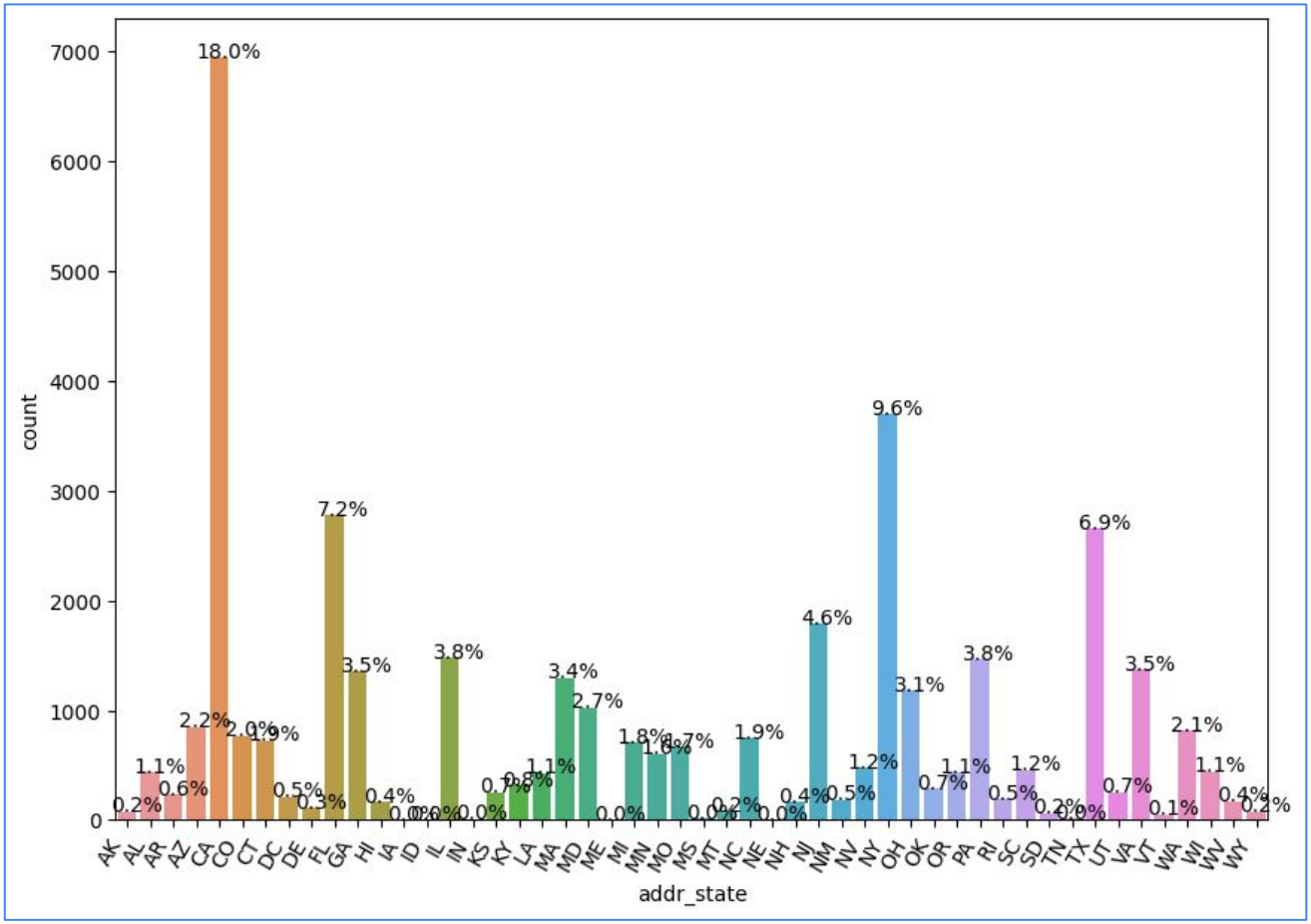
Univariate & Segmented Univariate



Majority of loans were taken by the customers belongs to the category of debt-consolidation followed by credit card.

Data Analysis

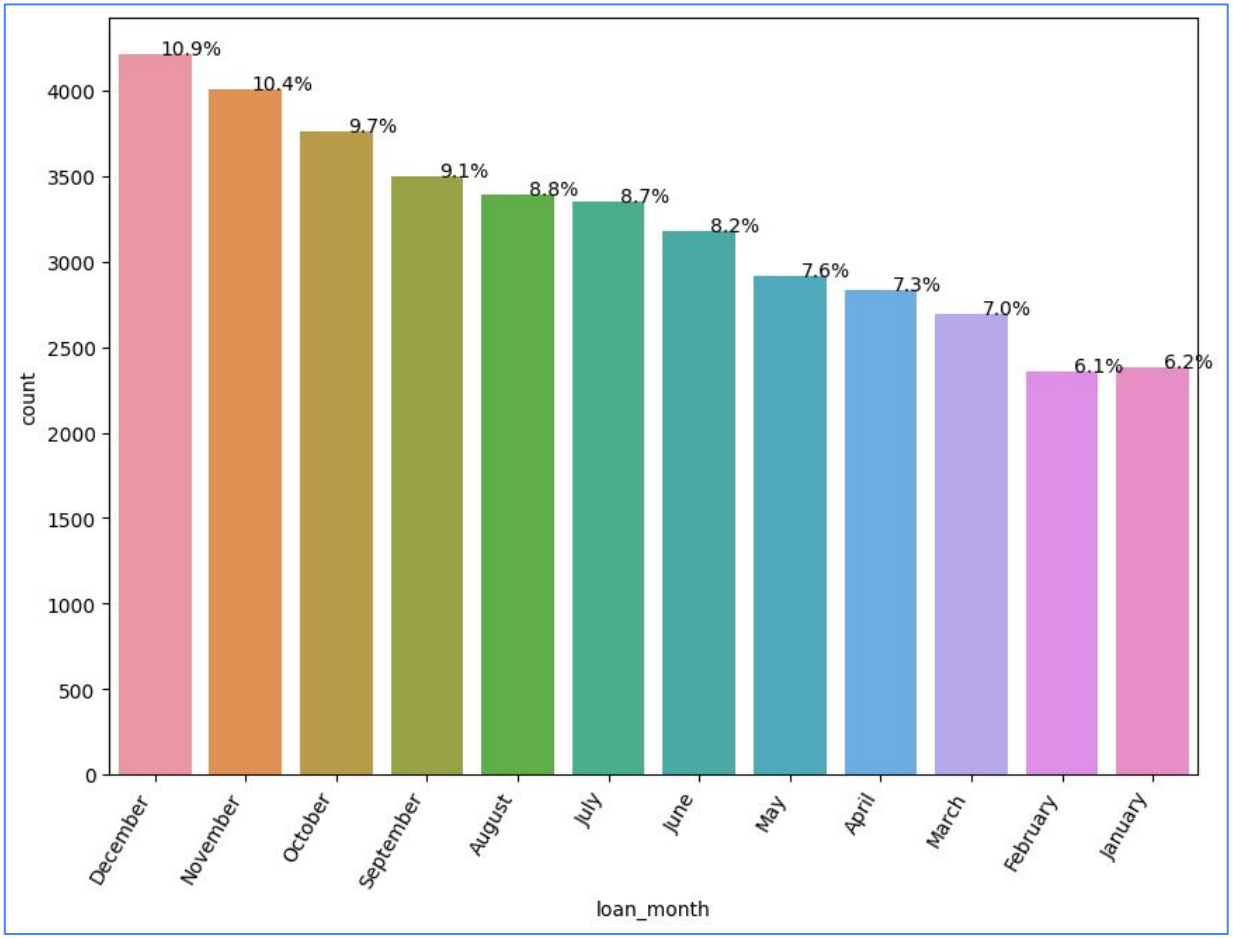
Univariate & Segmented Univariate



Majority of customers within US belongs to California Region.

Data Analysis

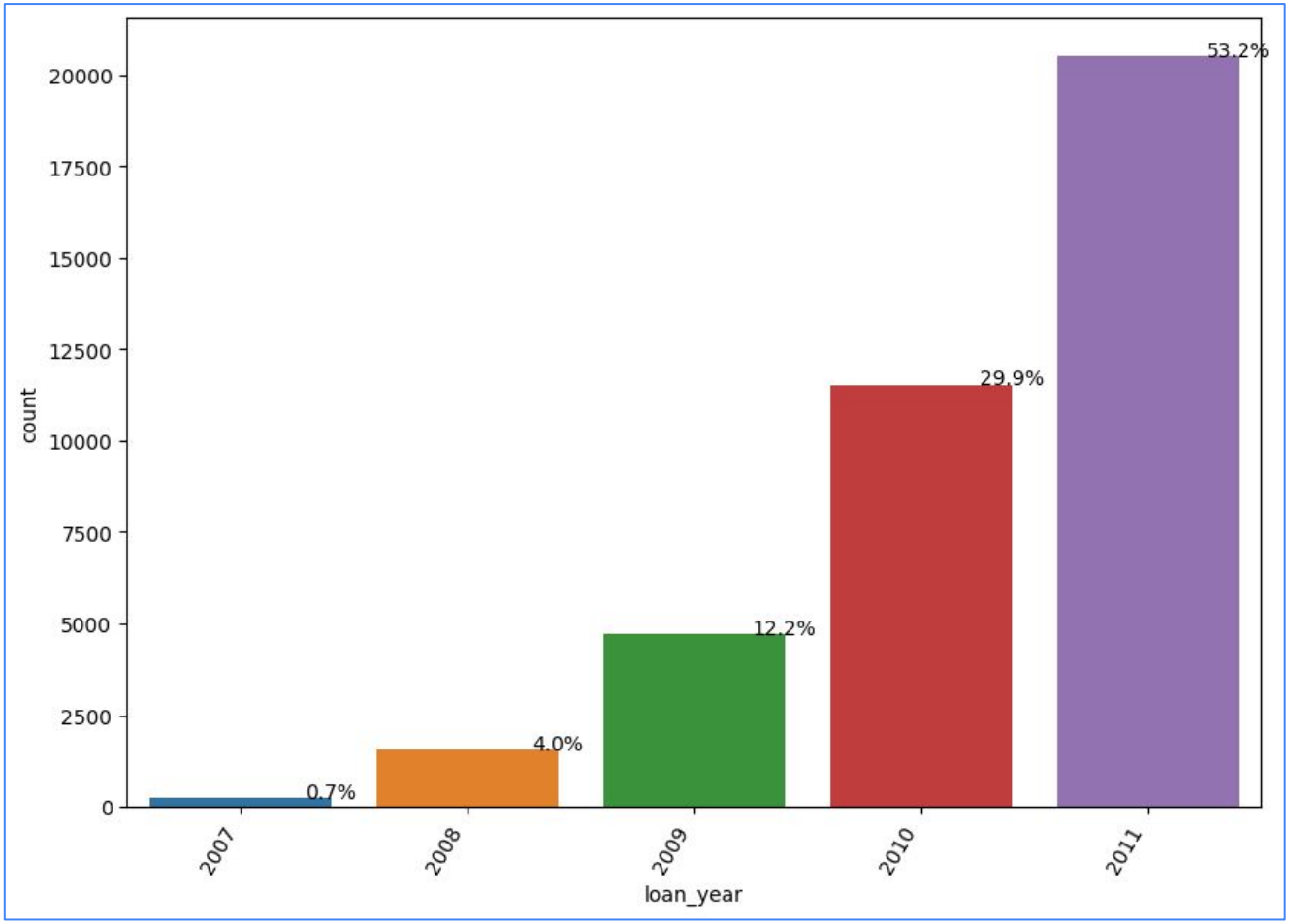
Univariate & Segmented Univariate



Majority of customers were issued loans during December, followed by Novemeber, October and so on. This might also be because customers would want to purchase house or have other plans for next year.

Data Analysis

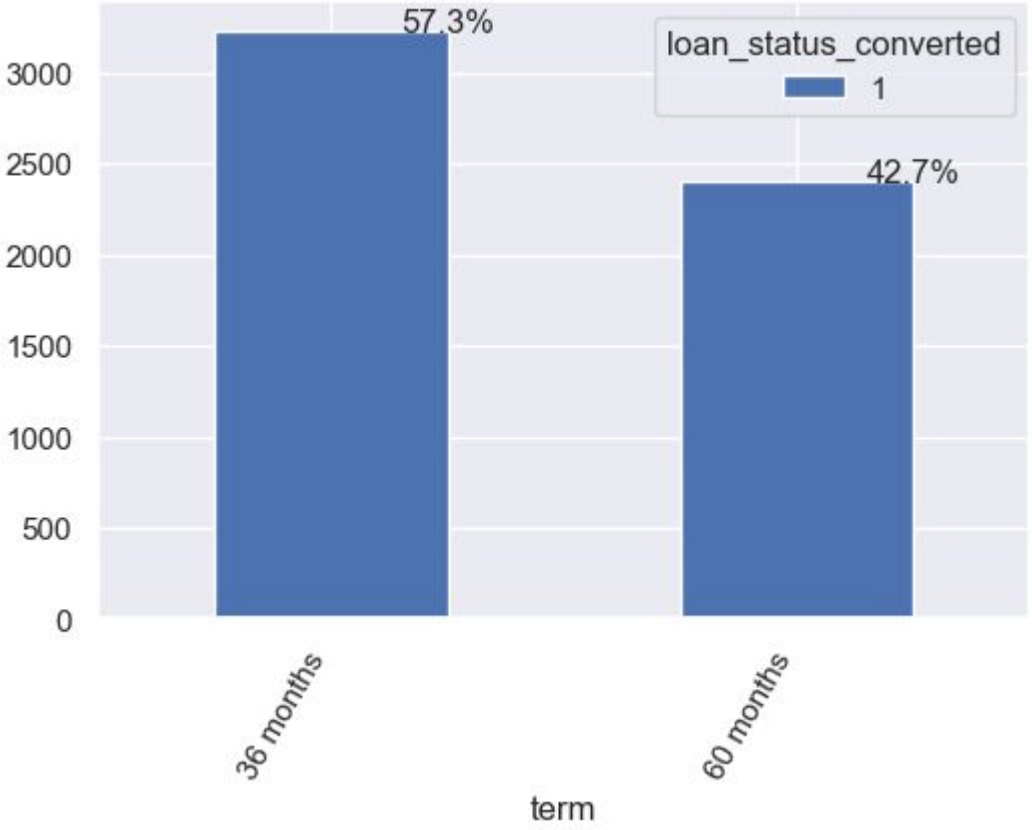
Univariate & Segmented Univariate



Over the period of time, we are seeing borrowing has considerably increased in 2011 compared to other years.

Data Analysis

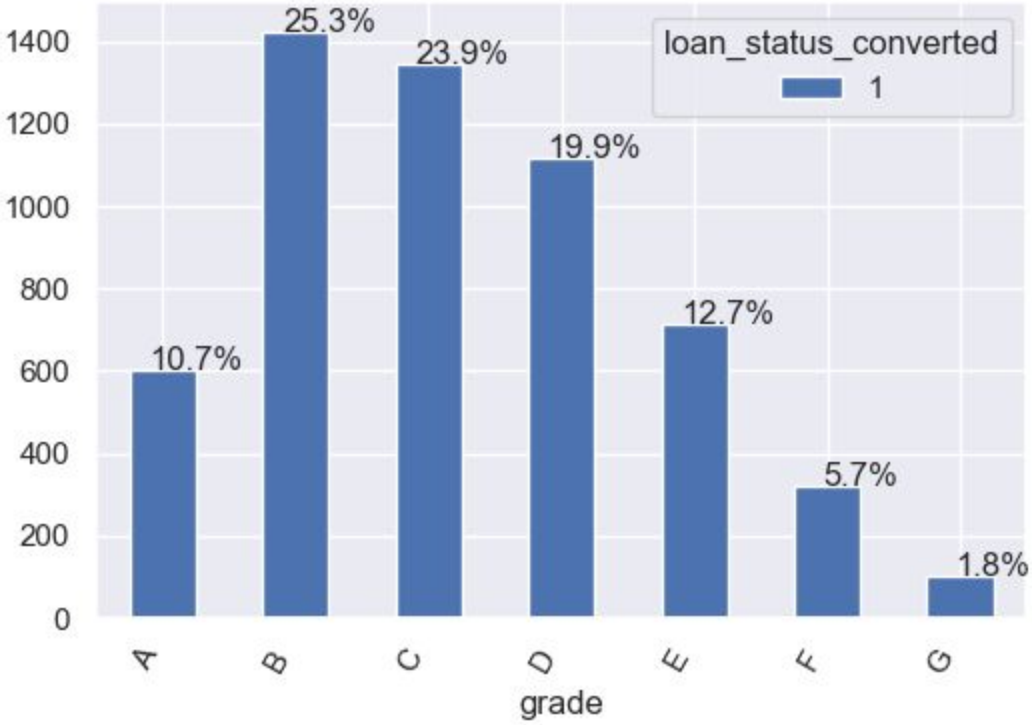
Target variable vs Numerical or Categorical variable



Since majority of borrowers has borrowed loan of lower duration they have high risk of defaulting as well.

Data Analysis

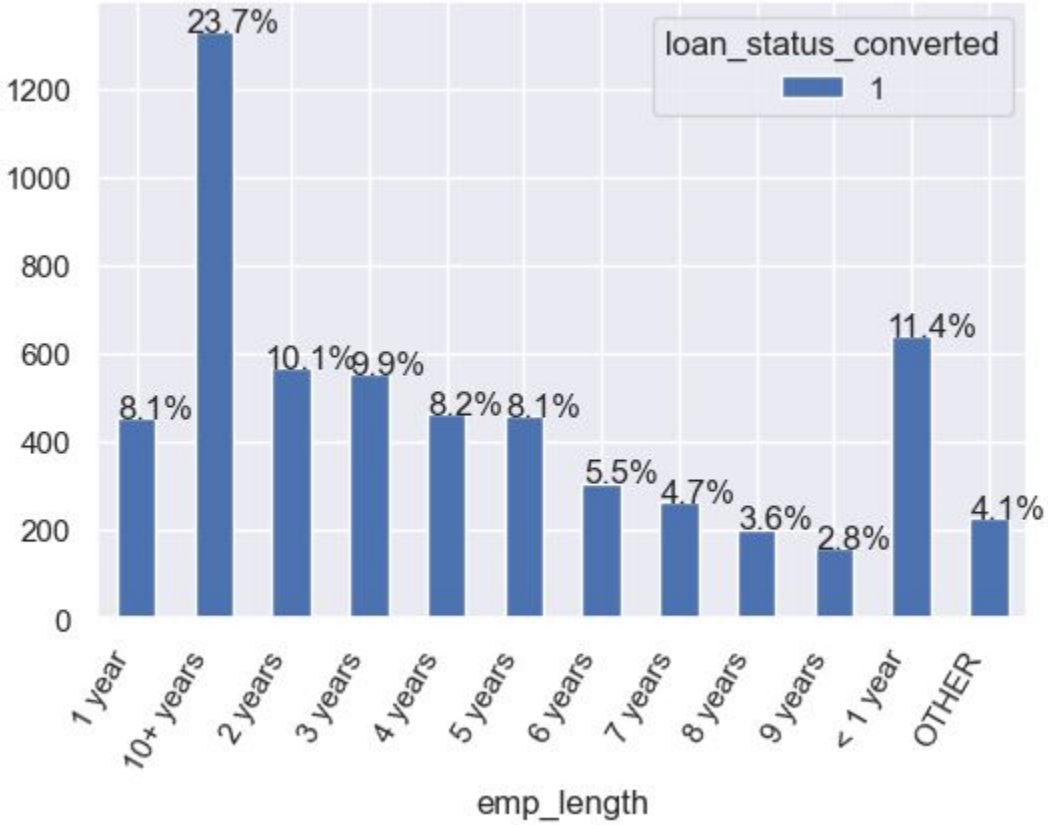
Target variable vs Numerical or Categorical variable



Majority of defaulters belong to B and C Category followed by D , E and then A.
Customers belonging to these categories can be classified as risky customers and other parameters should be checked

Data Analysis

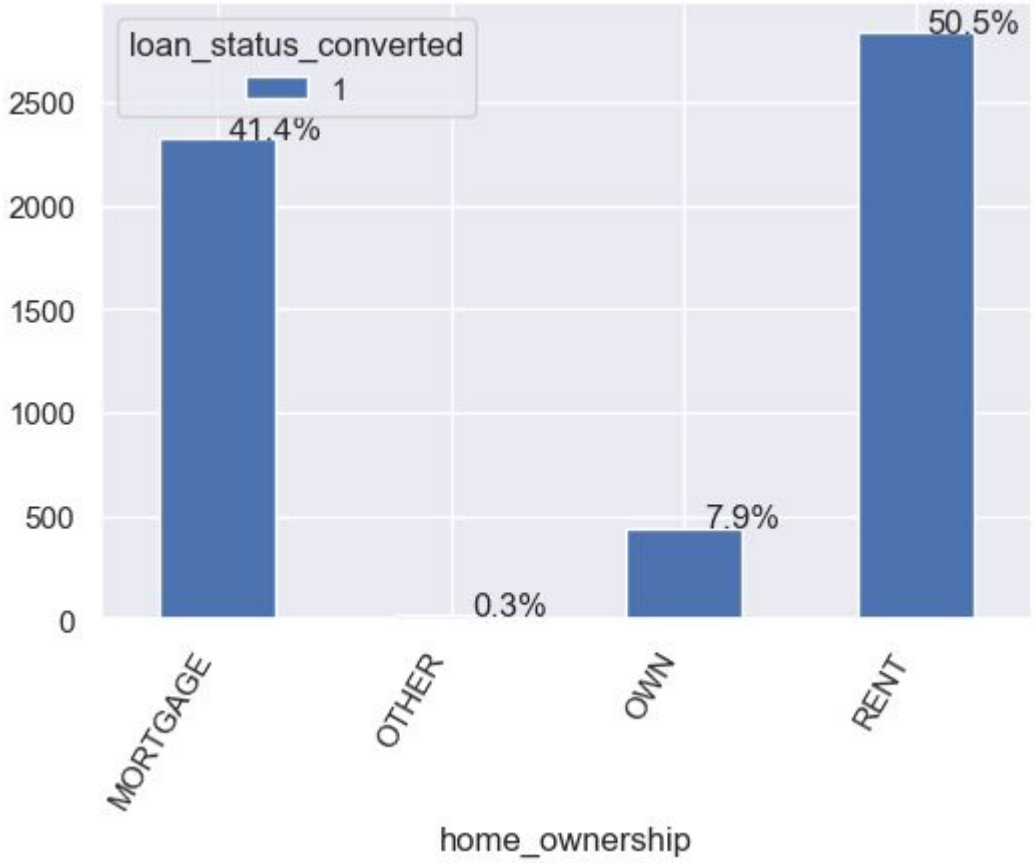
Target variable vs Numerical or Categorical variable



Majority of defaulters has an employment tenure of more than 10 yrs, followed by the ones who has lesser tenures (<1 yrs>). Borrowers with lesser employment tenure can be classified as risky borrowers.

Data Analysis

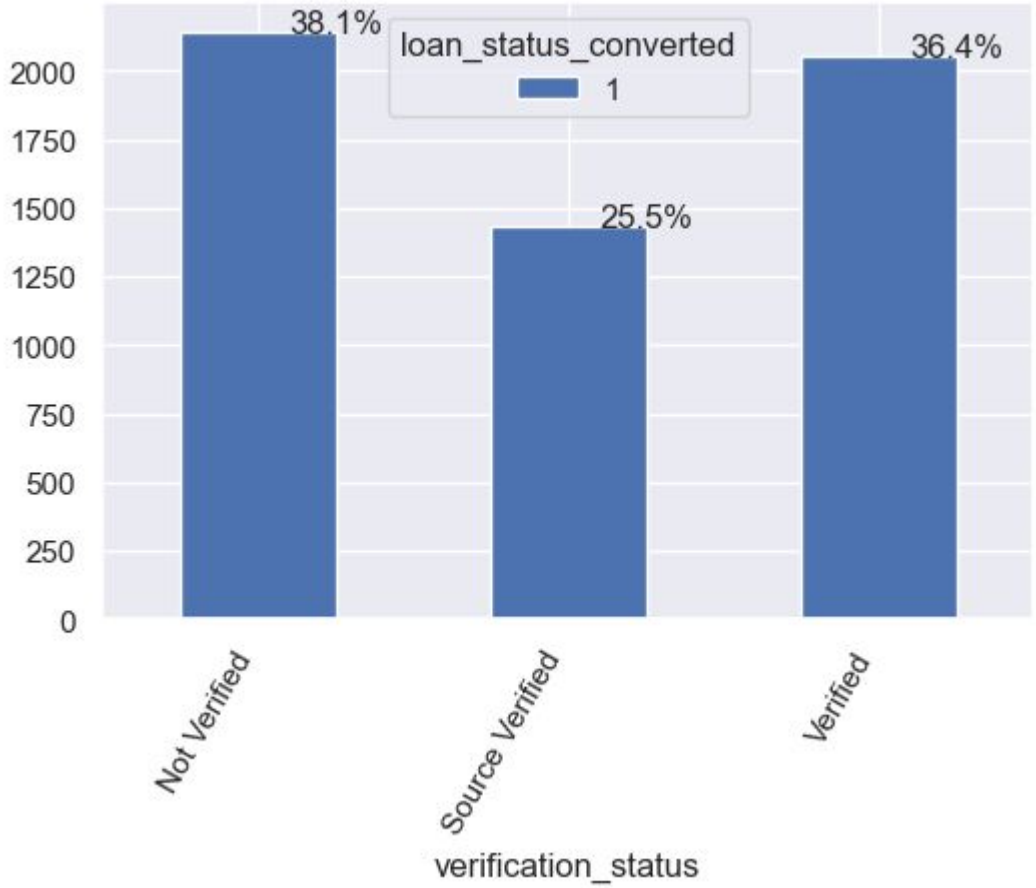
Target variable vs Numerical or Categorical variable



Majority of borrowers have either rented or paying mortgage and hence becomes the risk factor as well for lending.

Data Analysis

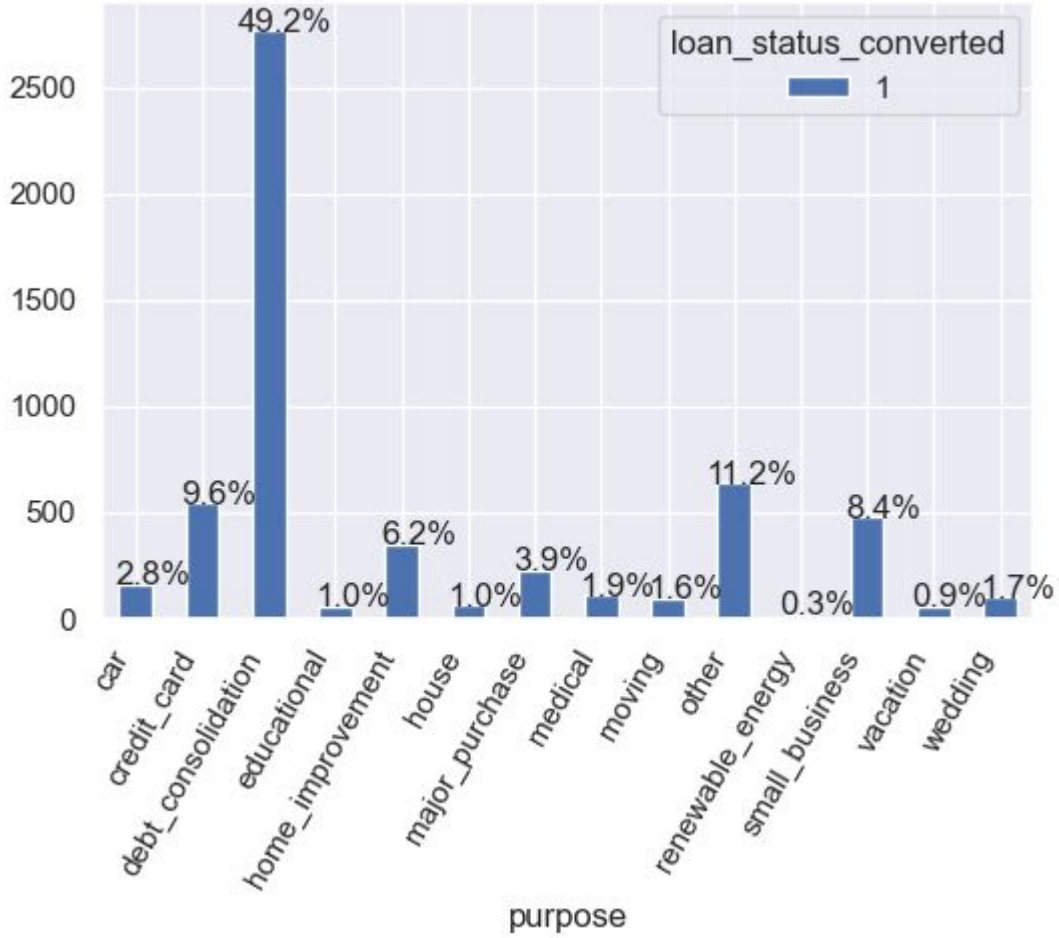
Target variable vs Numerical or Categorical variable



Borrowers which are not verified or LC verified becomes the high risk borrowers.
while the ones who are source verified are good borrowers

Data Analysis

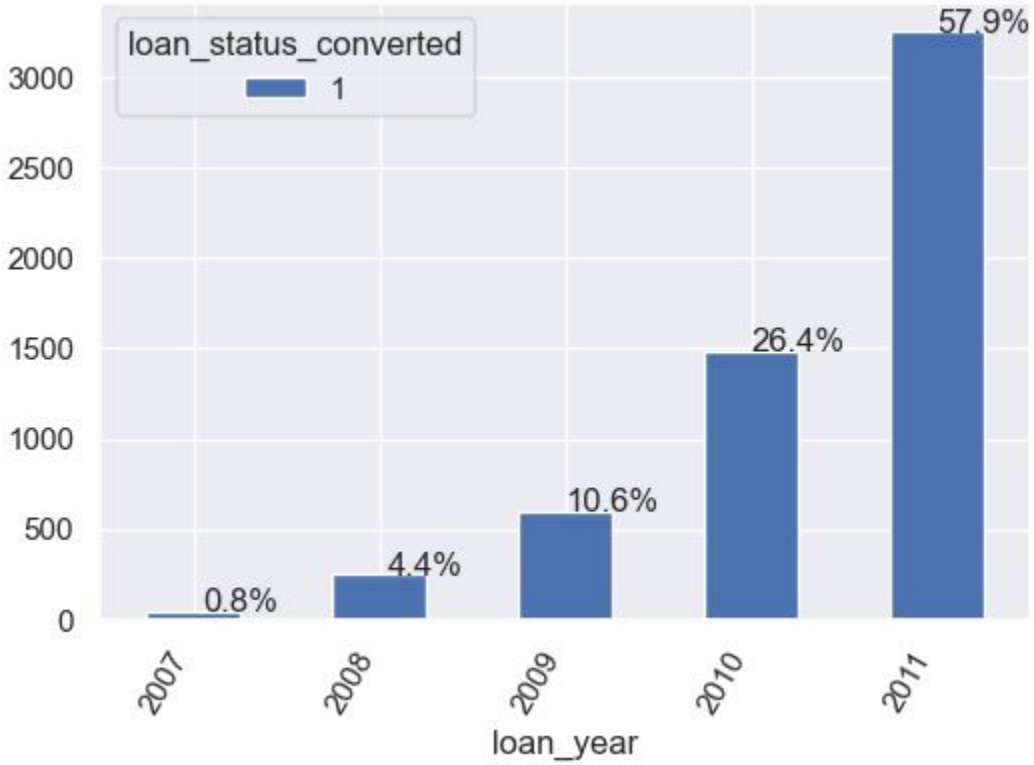
Target variable vs Numerical or Categorical variable



Since, majority of borrowers have borrowed for the purpose of debt_consolidation hence becomes the risk factor.

Data Analysis

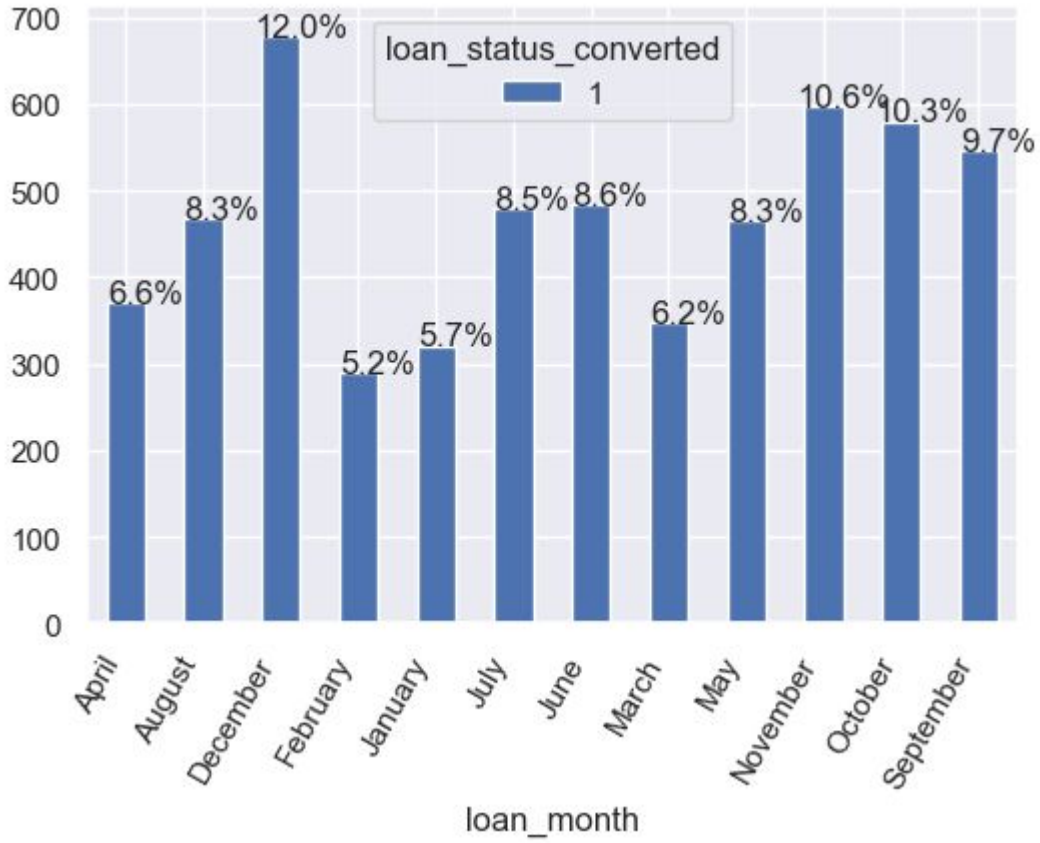
Target variable vs Numerical or Categorical variable



Borrowers who have been lended money in 2011 are risky borrowers and shows a high chance defaulting compared to other years

Data Analysis

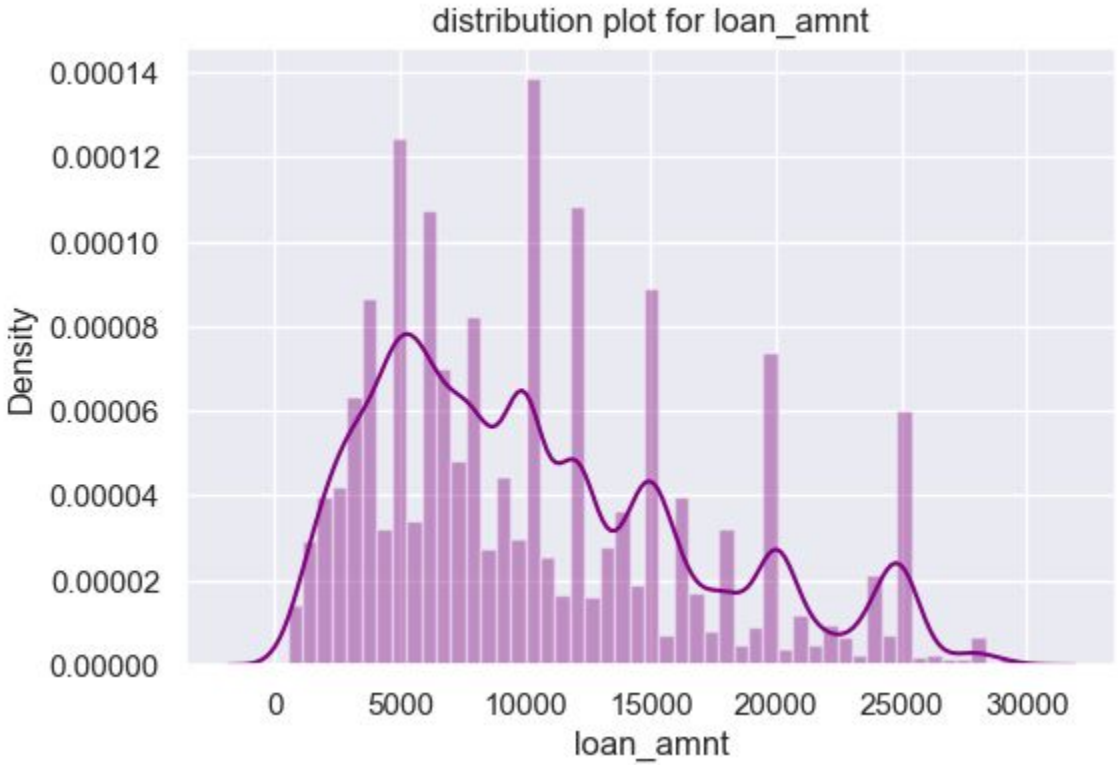
Target variable vs Numerical or Categorical variable

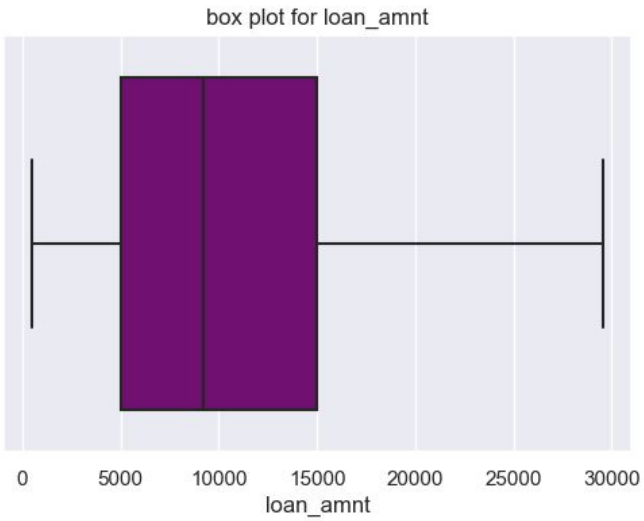


People who have borrowed in the month of December are risky borrowers followed by consecutive months in decreasing order

Data Analysis

Numerical variables

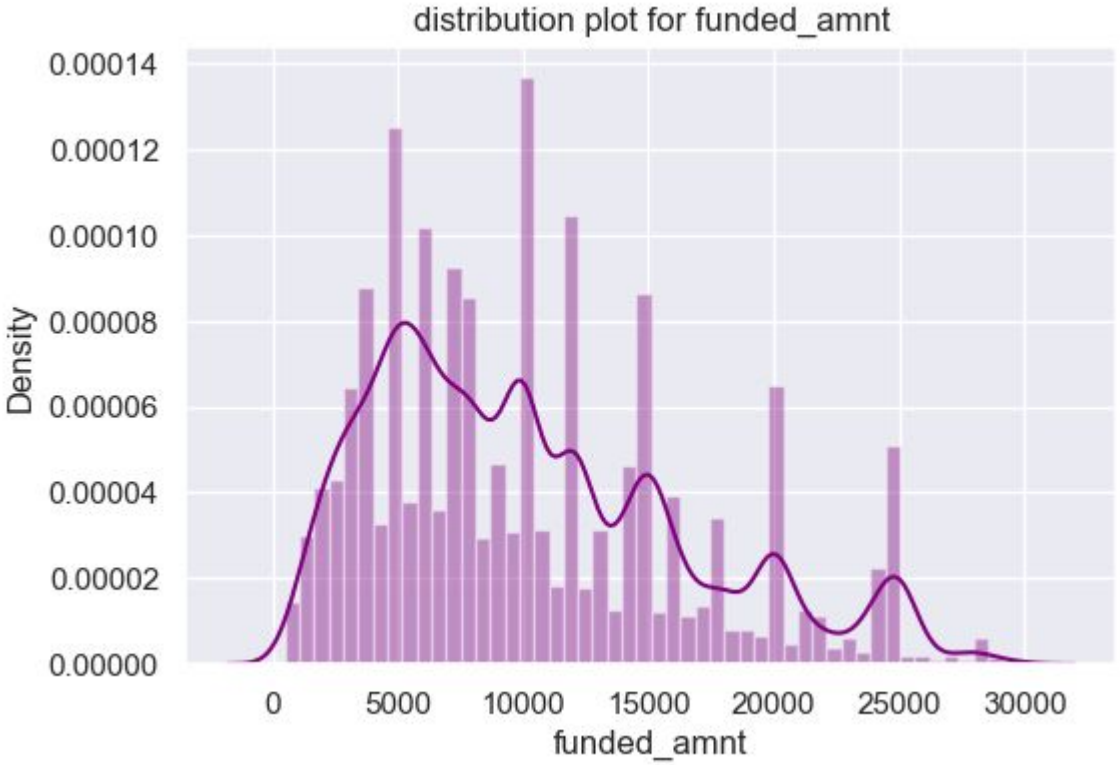


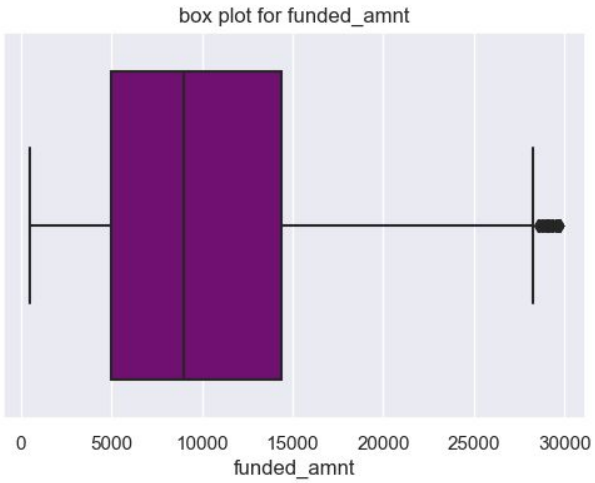


The distribution plot does not show a uniform distribution in data for loan amount.

Data Analysis

Numerical variables

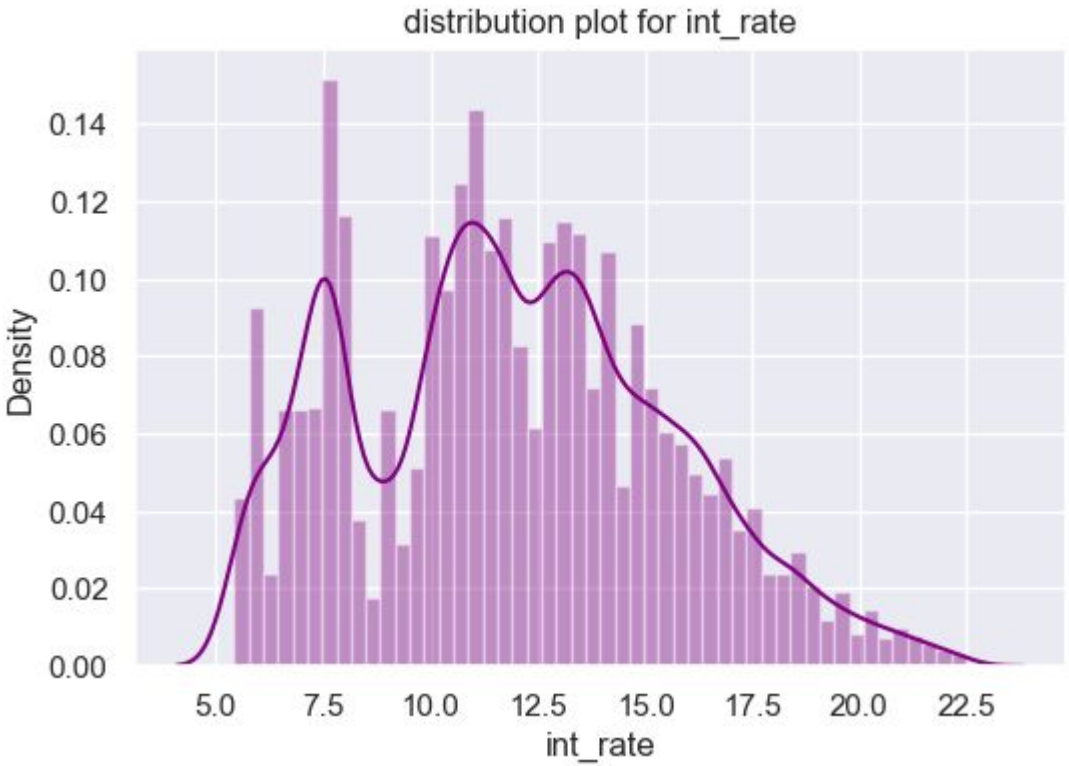


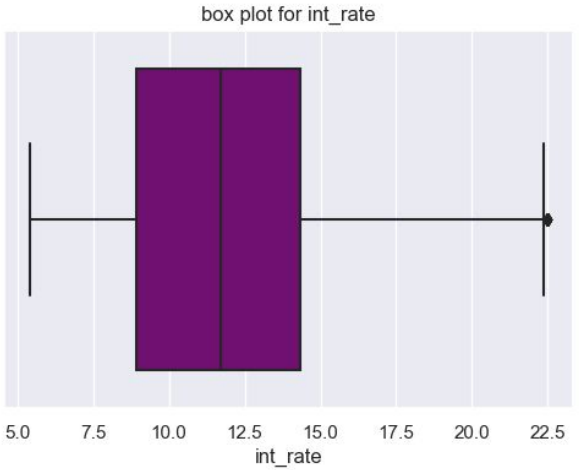


Distribution plot does not show a uniform distribution in data for funded amount. Loan funded lies between 5k to 14k. However , majority of borrowers are funded 10k as we can see from distribution plot and box plot as well.

Data Analysis

Numerical variables



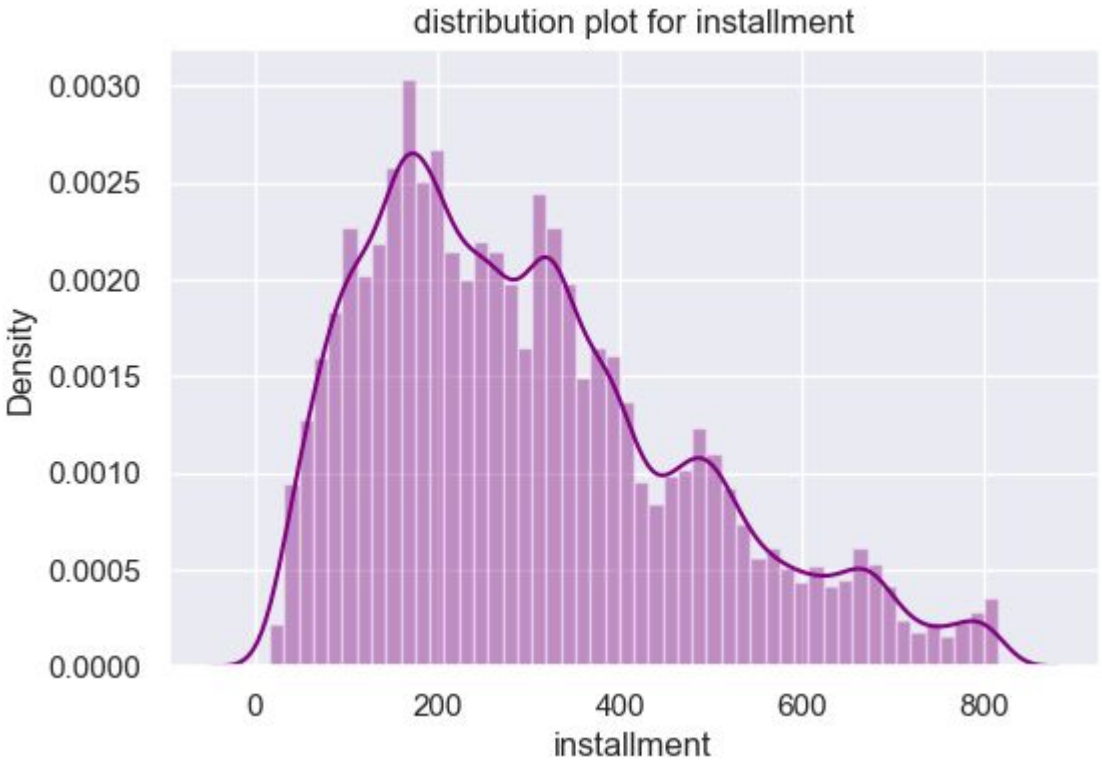


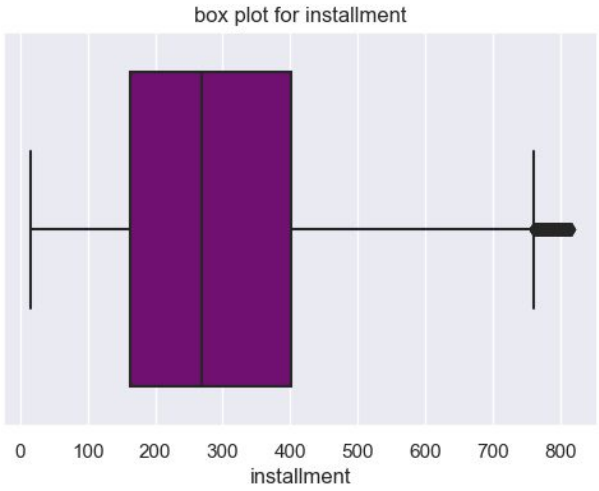
The distribution plot does not show a uniform distribution in data for interest rate.

However, majority of the customers are provided an interest rate of approx 11% which is quite high. This could be one of the factors for defaulting customers

Data Analysis

Numerical variables



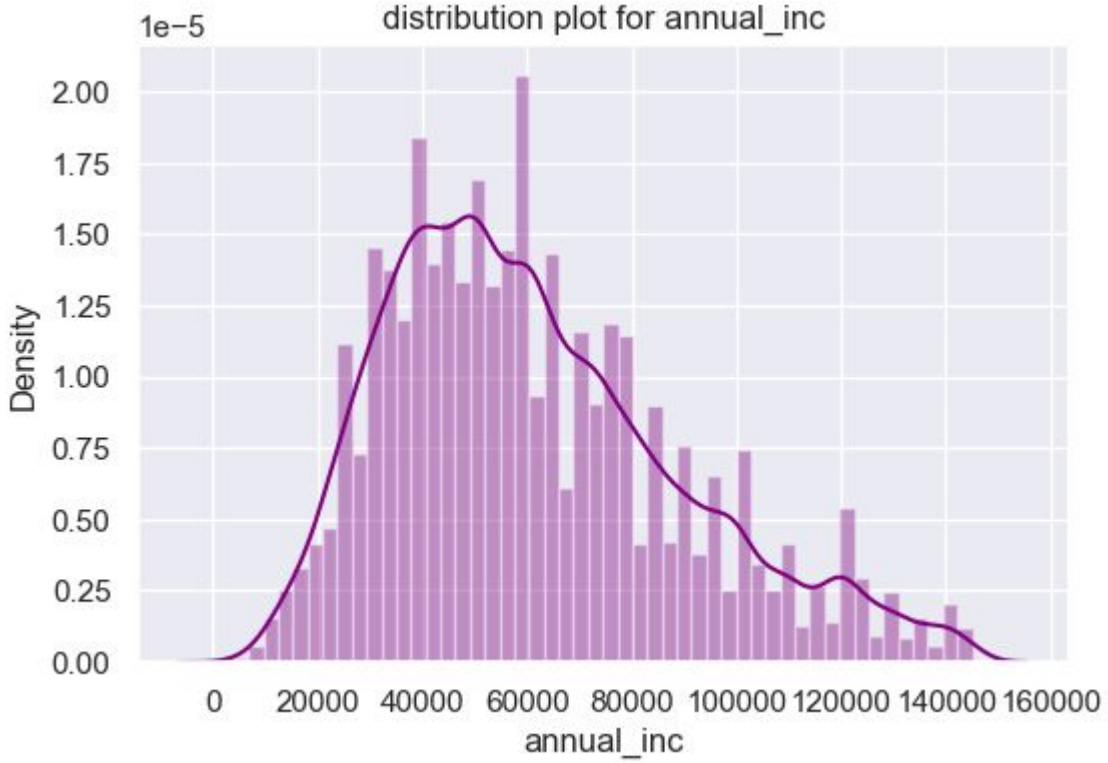


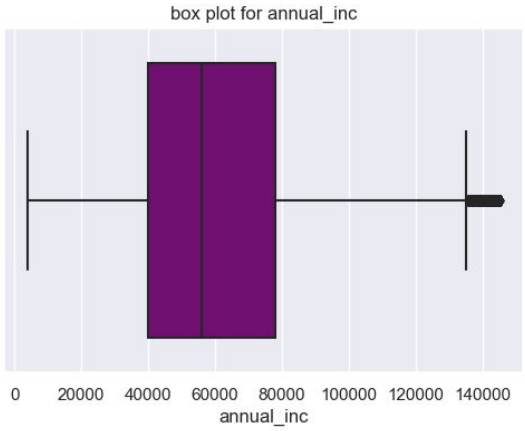
Majority of people pay installement between approx. 160 and 400k which is less.

The distribution of installement is also not uniform.

Data Analysis

Numerical variables

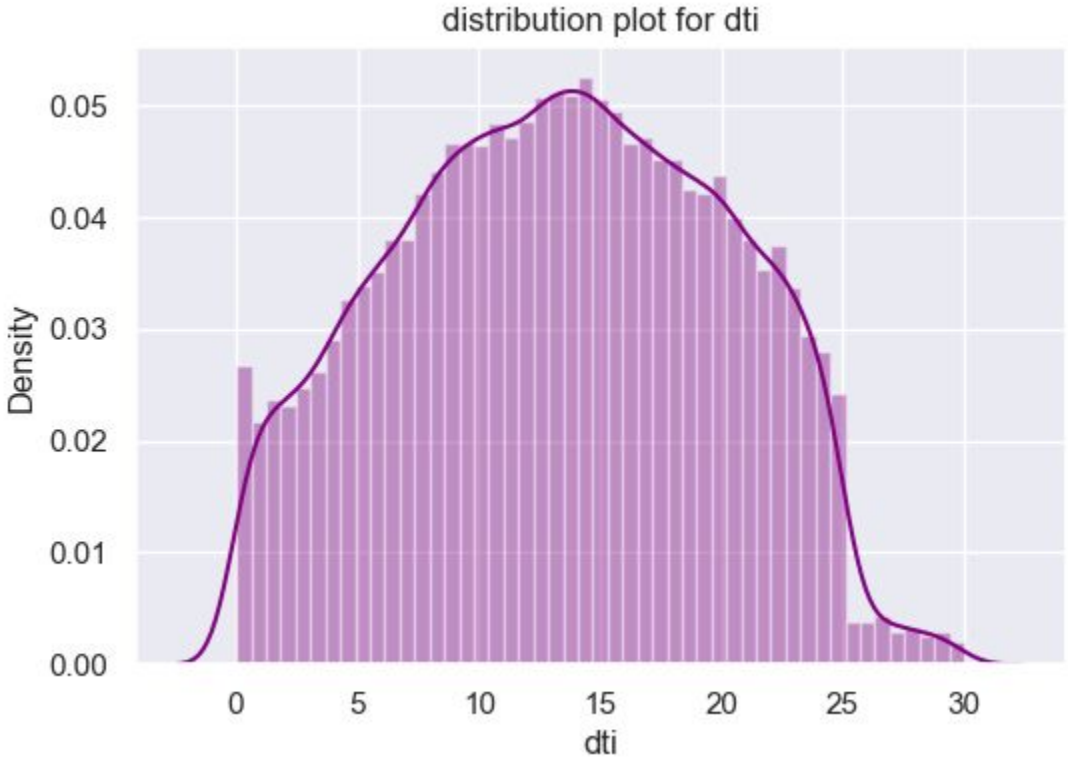


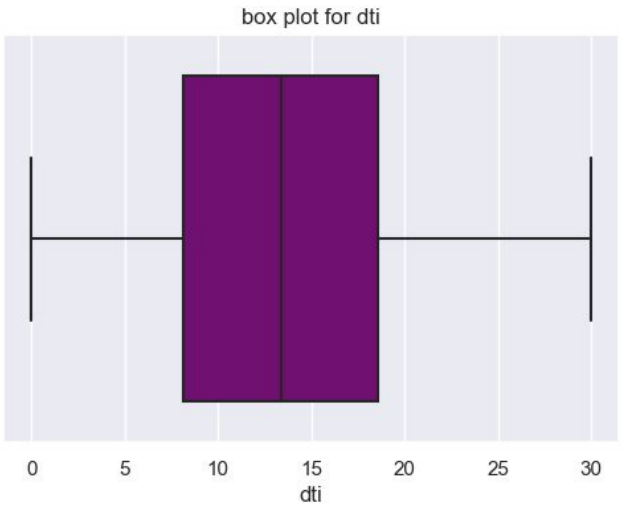


Borrower`s annual income starts from 40k to 78k and the favorables are the ones earling approx 56K

Data Analysis

Numerical variables



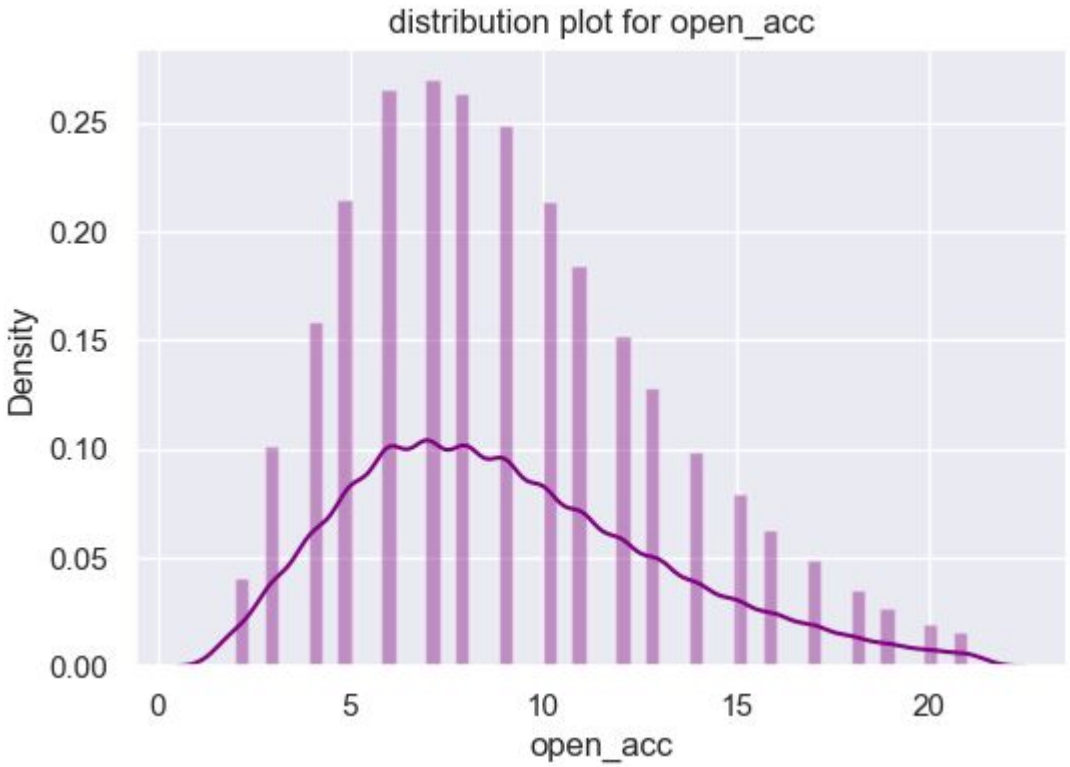


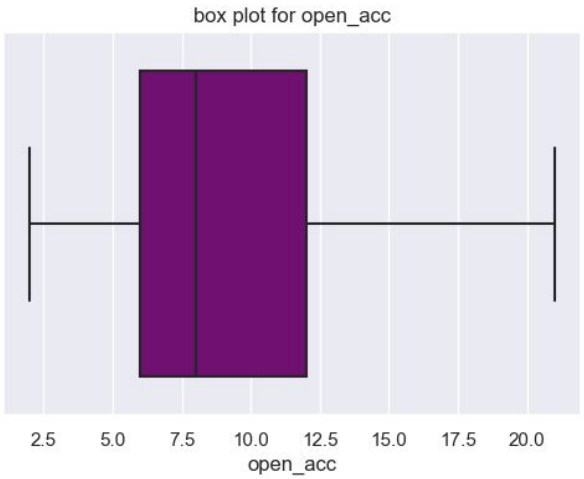
Borrower`s dti lies between approx. 8 to 18 and among them majority has dti of approx. 13

The distribution plot of dti is somewhat uniform.

Data Analysis

Numerical variables

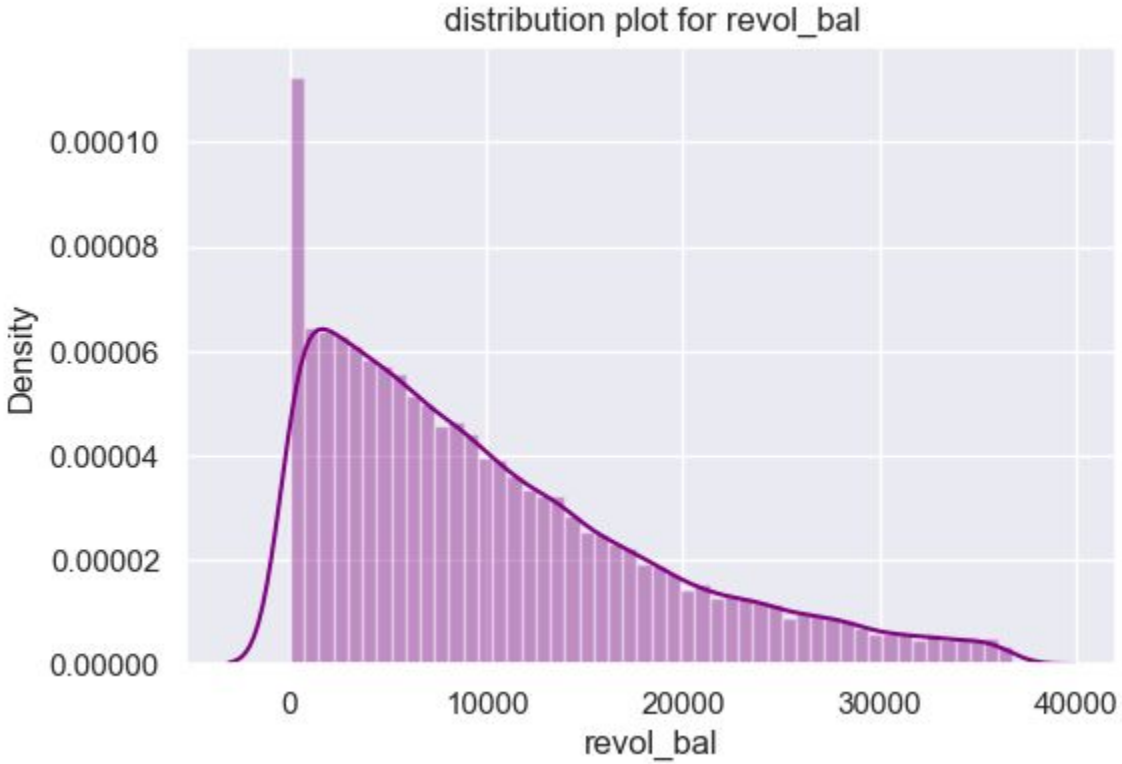


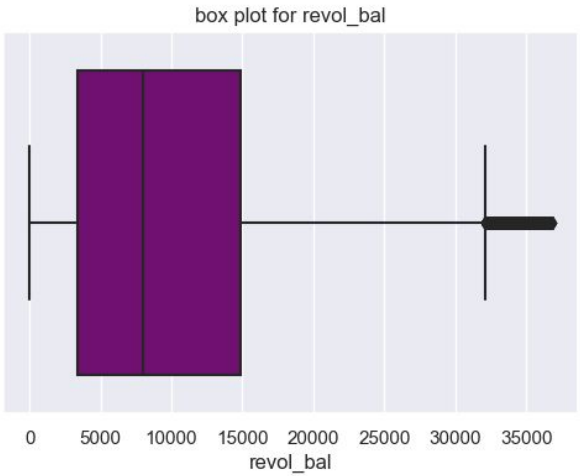


The distibution plot of open_acc is uniform with majority values ranging from 6 to 12.

Data Analysis

Numerical variables



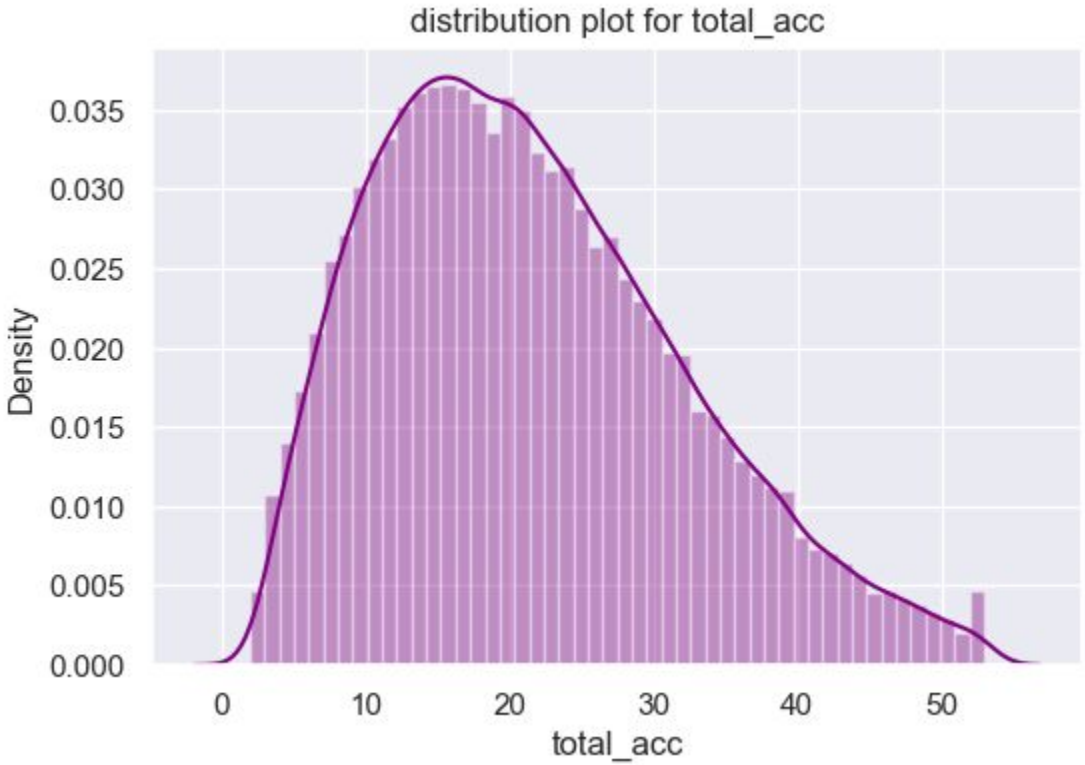


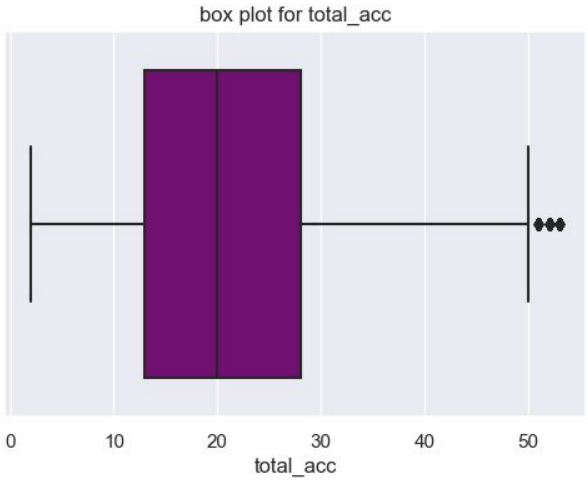
The distribution plot for revol_bal is skwed and majority values ranging from approx. 3k to 14k.

This is the credit borrower which is available to the borrower as he/she pays the balance.

Data Analysis

Numerical variables

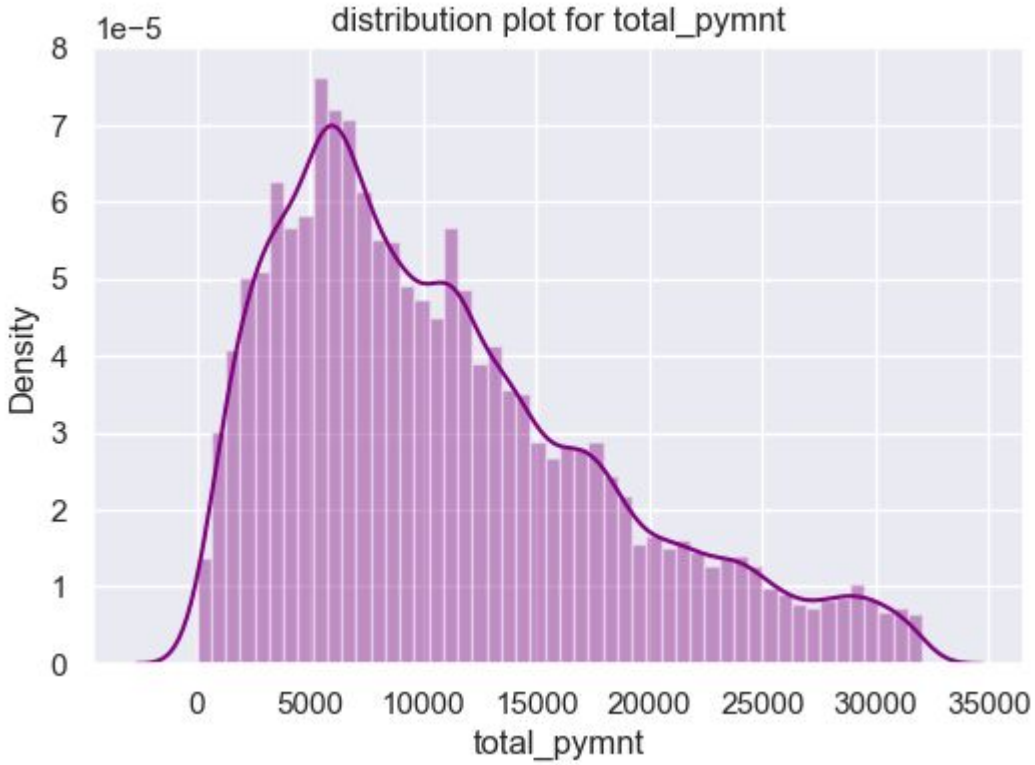


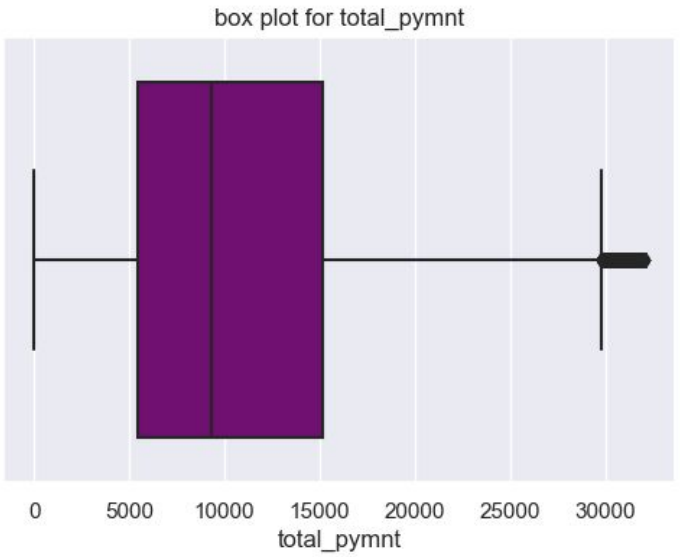


Values of otal_account is somewhat balanced with majority of them falling in the range of 13 to 53.
The median of total_accounts of borrower is 20.

Data Analysis

Numerical variables

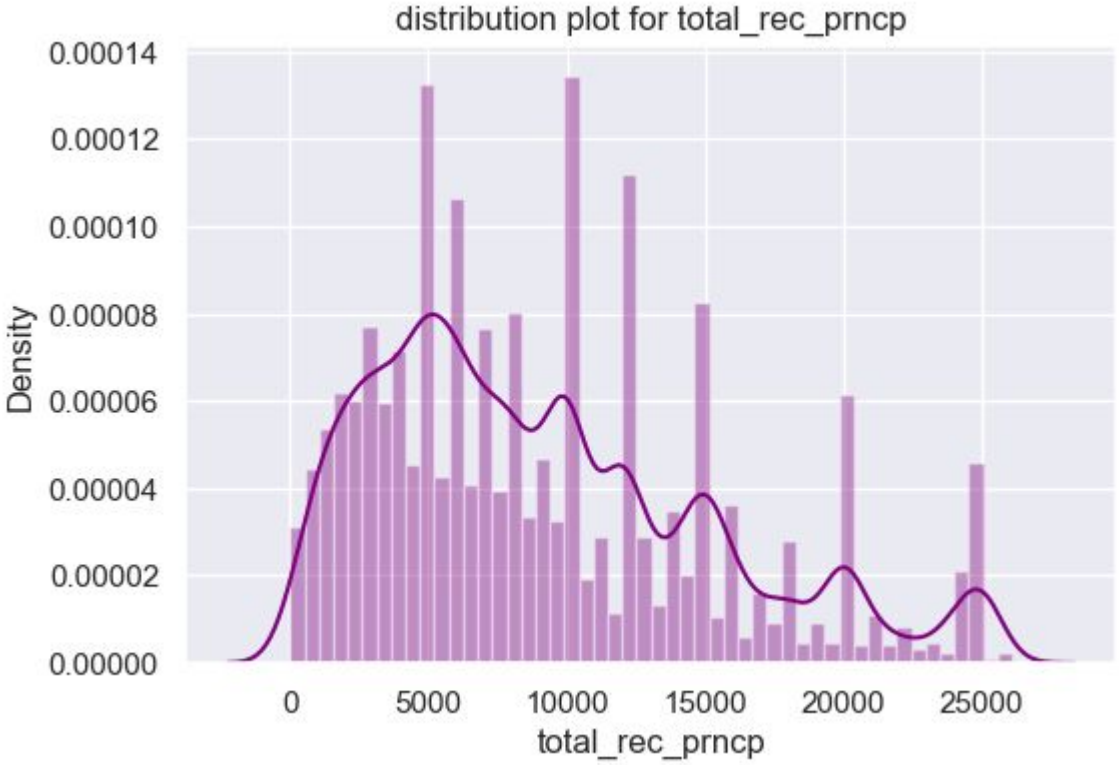


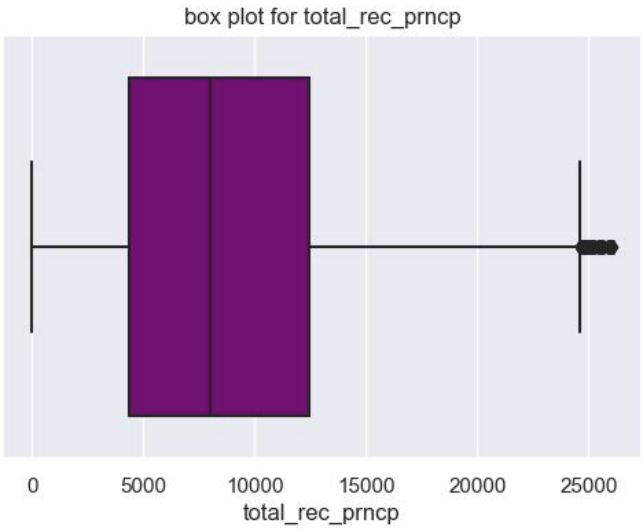


Total payment by borrower lies between approx 5k to 15k with skewed distribution. Majority of them pays approx. 9k.

Data Analysis

Numerical variables



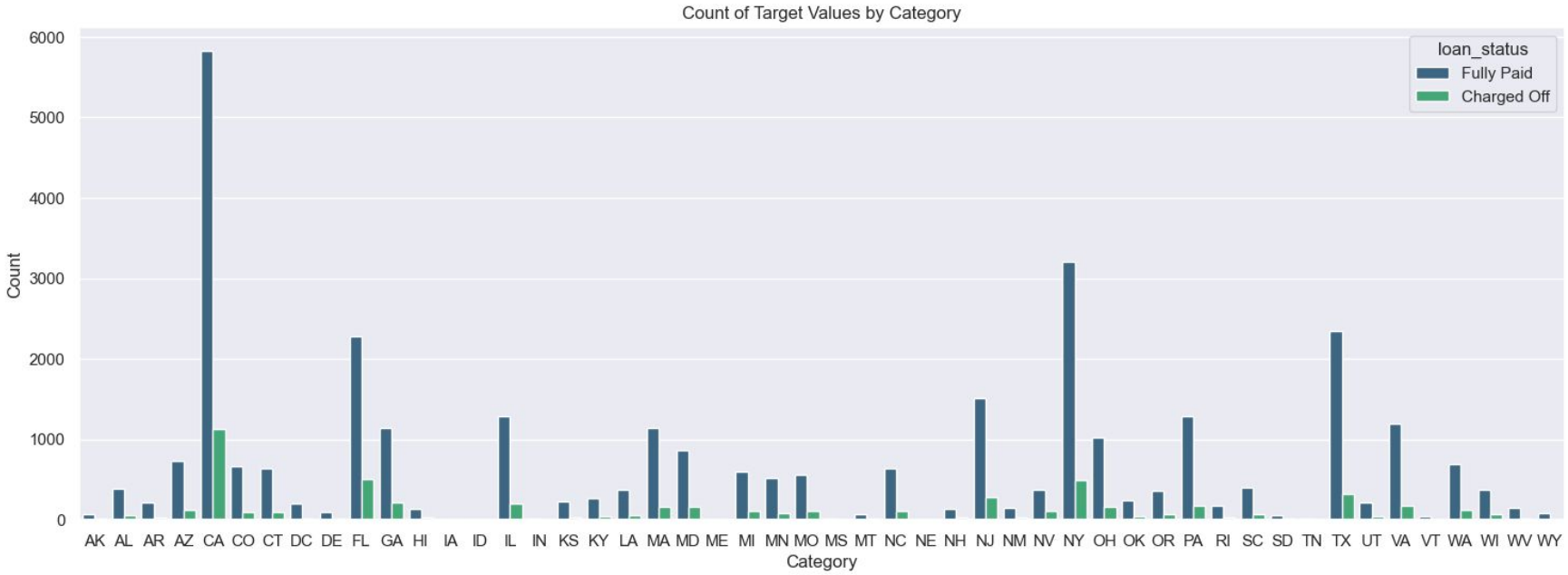


Total_rec_prncp is the total principal received till date. The payment values are not distributed uniformly.

The majority of payments are done between approx. 4k to 12k with median payments being 8k

Data Analysis

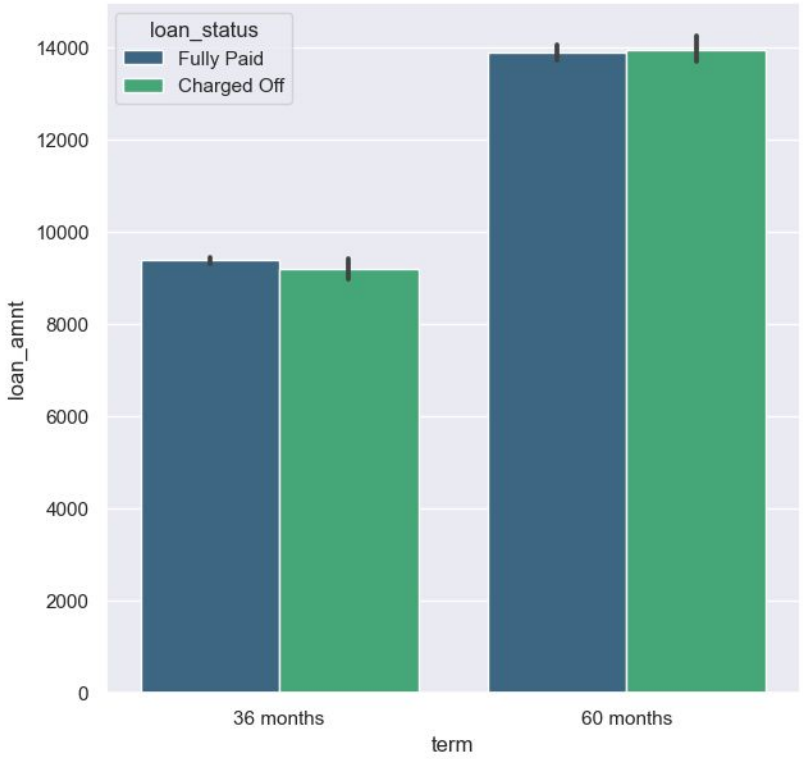
Demographic

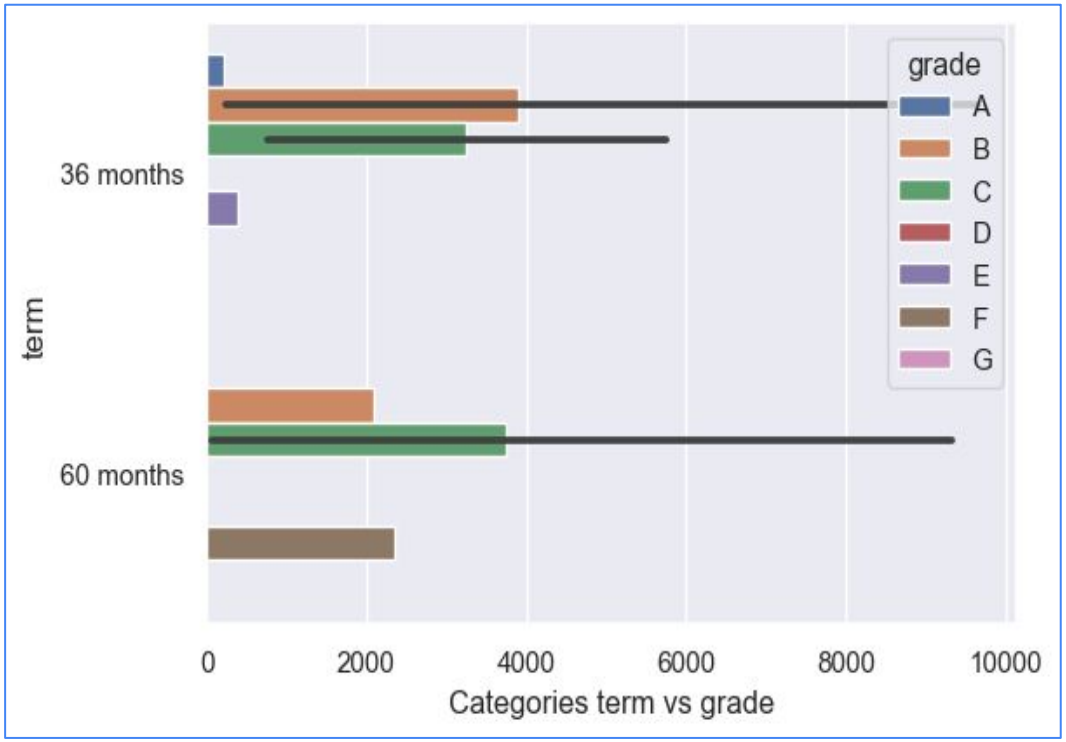


Based on the graph below, CA category has more number of borrowers.
Although, there are agreeable number of Fully paid candidates,
There are plenty of defaulters from the same category
There maybe various reasons to have borrowers from this category -> 1. High cost of living in the demographical place, 2. High need of expenditure

Data Analysis

B/w Numerical and Categorical



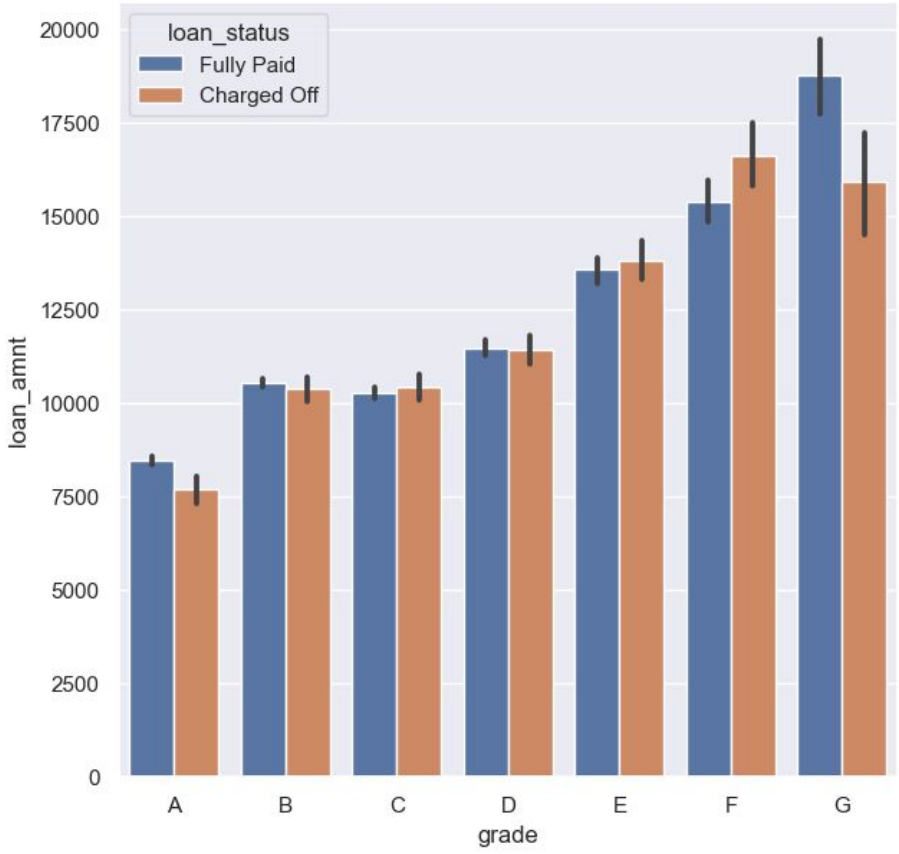


Considering loan _amount term and loan_status, higher loan_amount for longer duration, customers tend pay fully and default as well.

customers of category B and C have higher weightage in short and long duration

Data Analysis

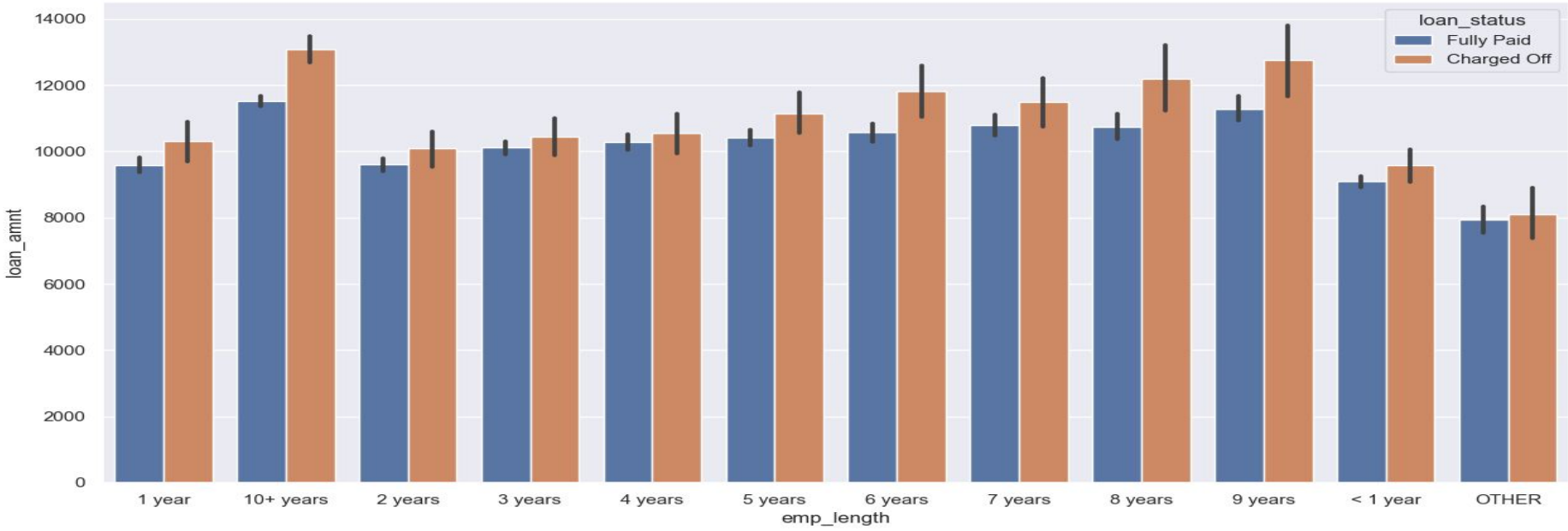
B/w Numerical and Categorical



so as the loan amount increases, customers belonging to lower grade (E & F) tend to default more and can be risky customers have same levels of lending loan to all customers

Data Analysis

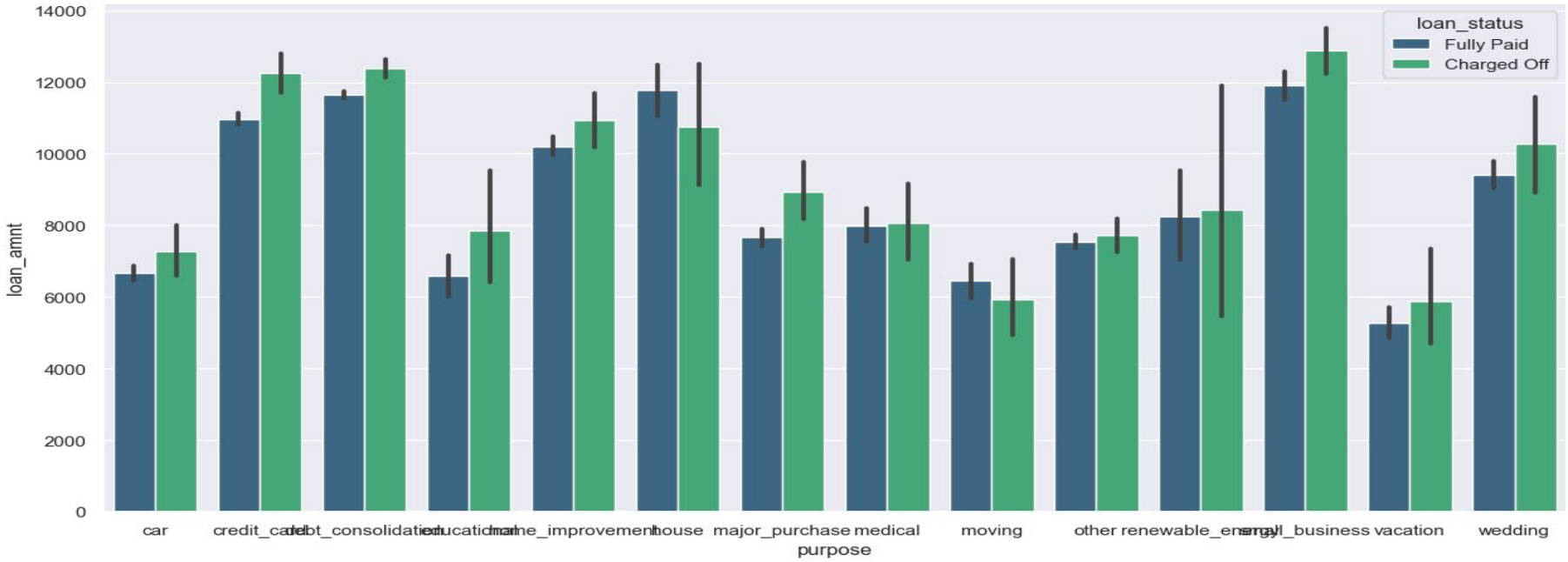
B/w Numerical and Categorical



Customers with 10+ yrs of employment tend to default more followed by customers with 9 and 8 yrs for employment tenure.

Data Analysis

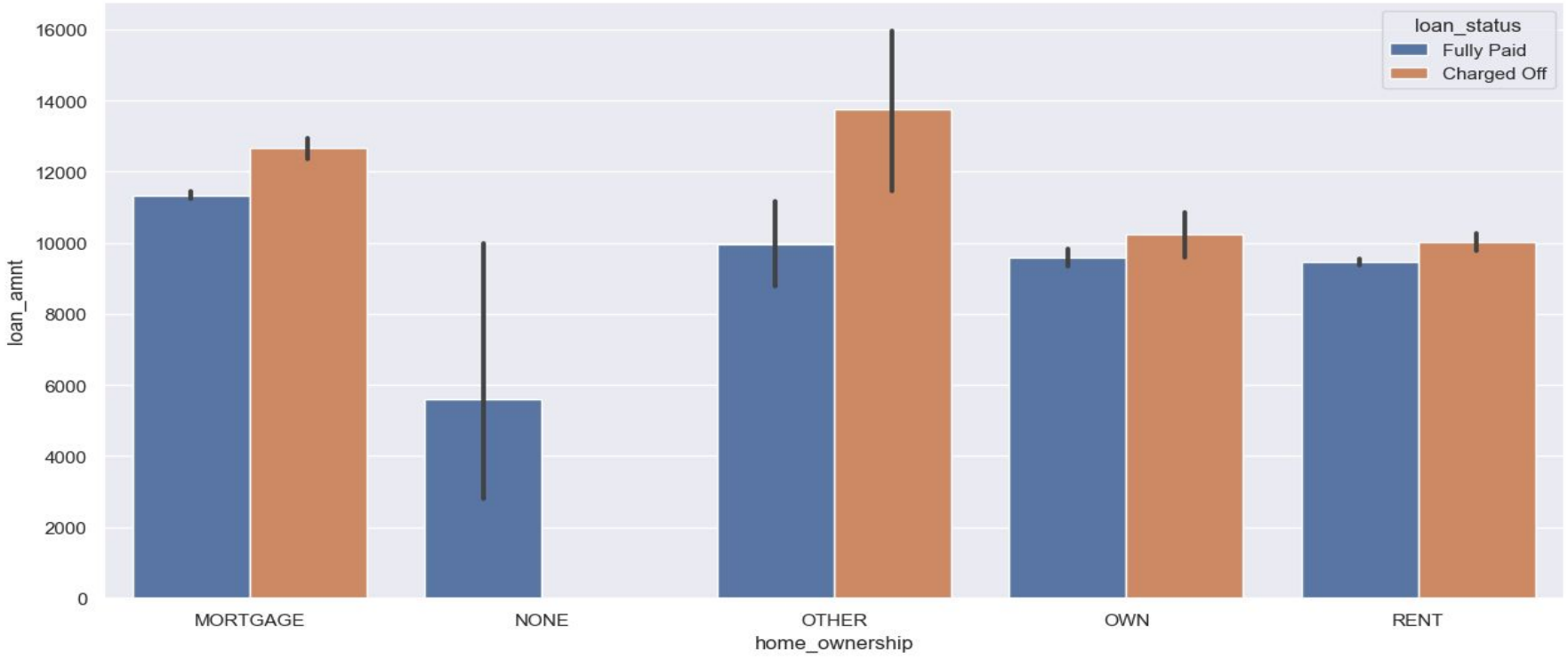
B/w Numerical and Categorical



higher Loan taken for small business are risky customers with chances of defaulting more compared to others

Data Analysis

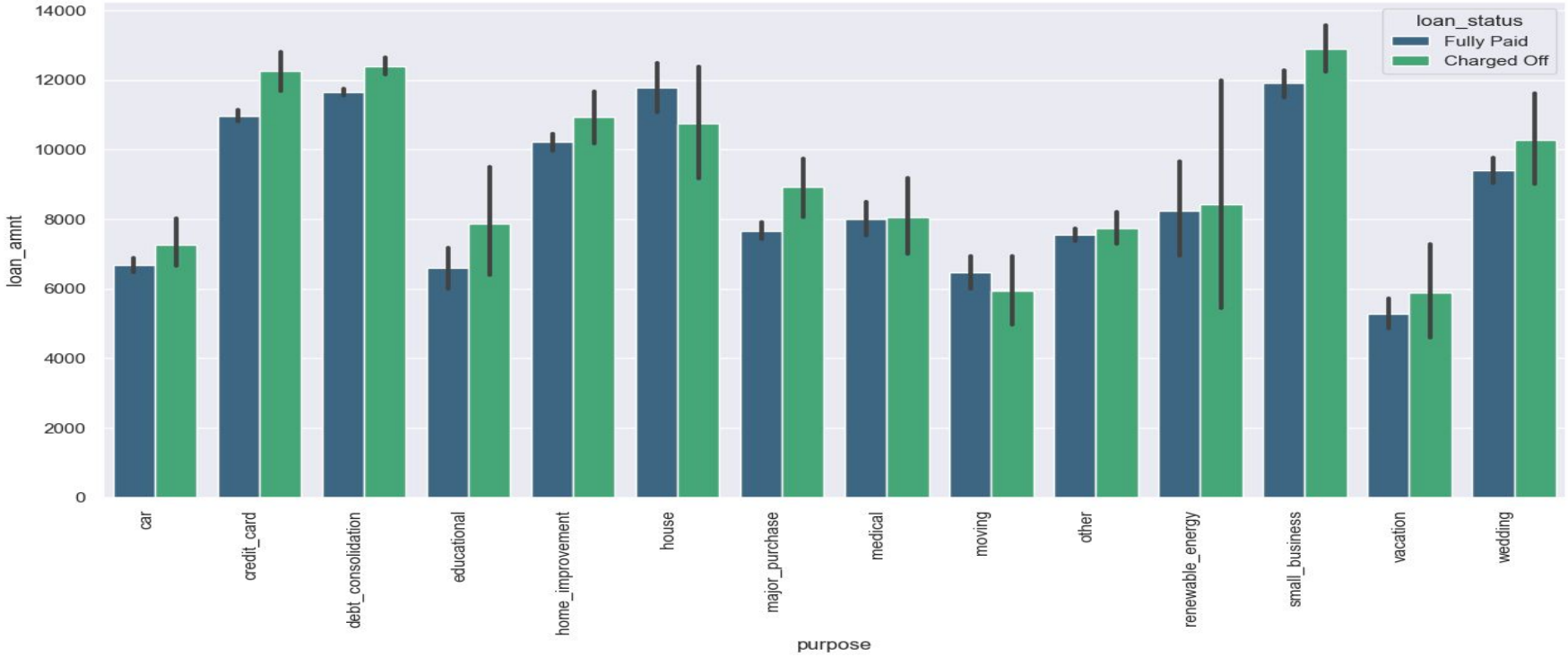
B/w Numerical and Categorical



Customers who are paying mortgage with higher loan_amount are risky customers and tend to default more

Data Analysis

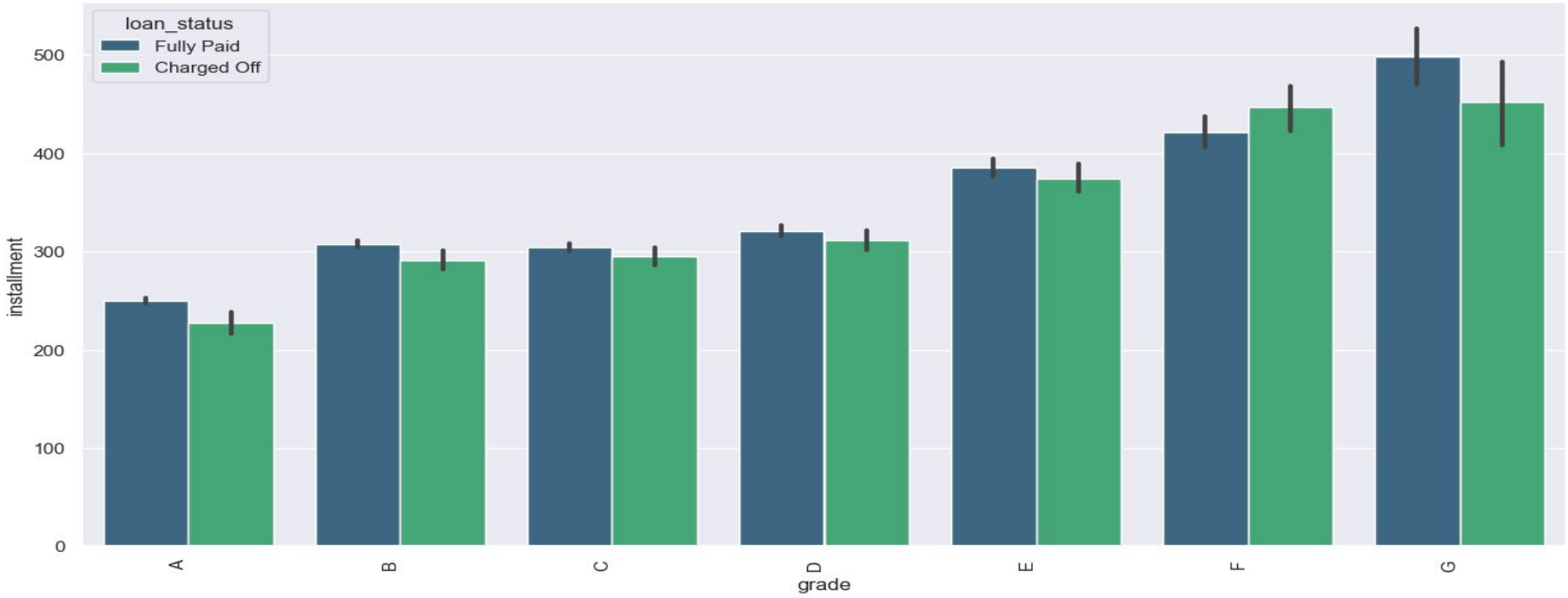
B/w Numerical and Categorical



higher Loan taken for small business are risky customers with chances of defaulting more compared to others

Data Analysis

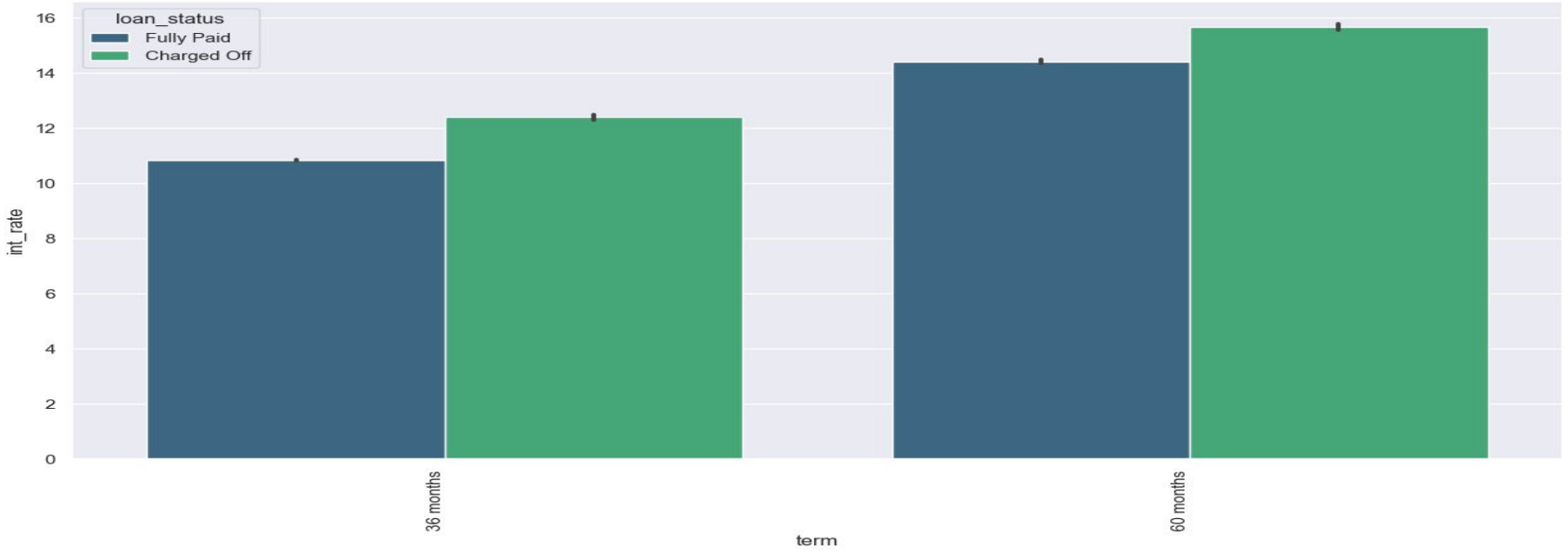
B/w Numerical and Categorical



customers Higher the installment amount for lower level of grade (e.g E) will tend to default more and are risky customers

Data Analysis

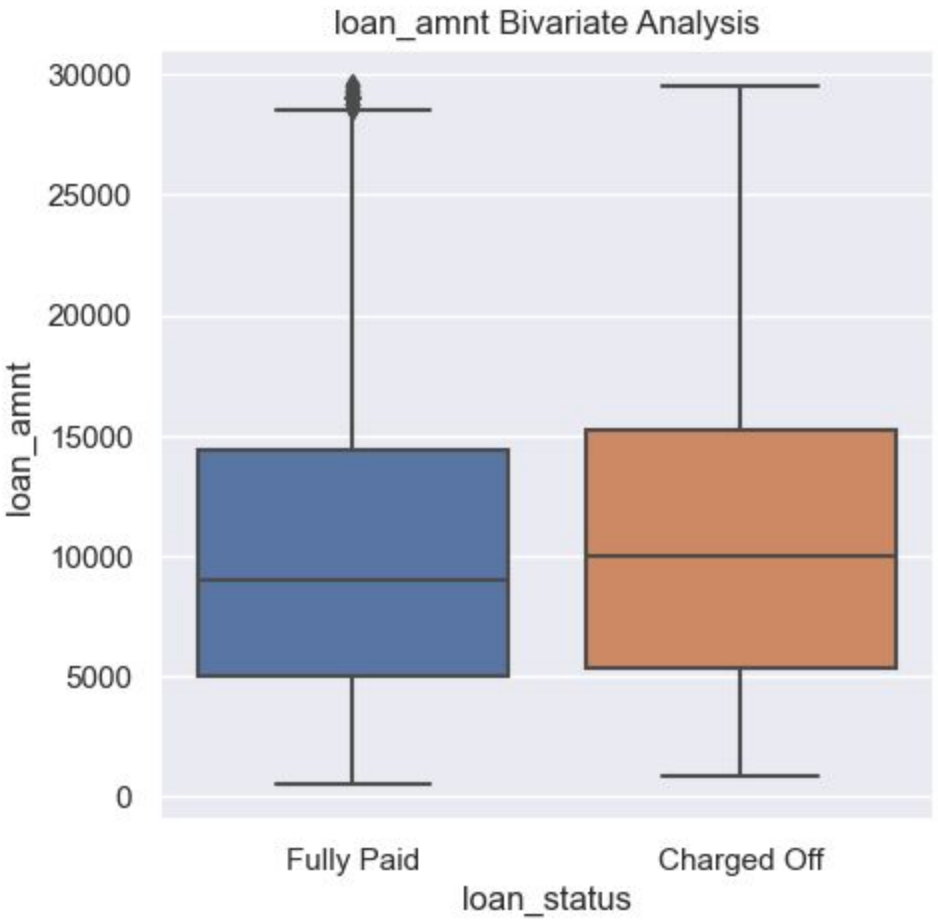
B/w Numerical and Categorical



Higher interest rate for longer duration of loan period(60 monhts) will have impact on customers and those customers will likely to default more.

Data Analysis

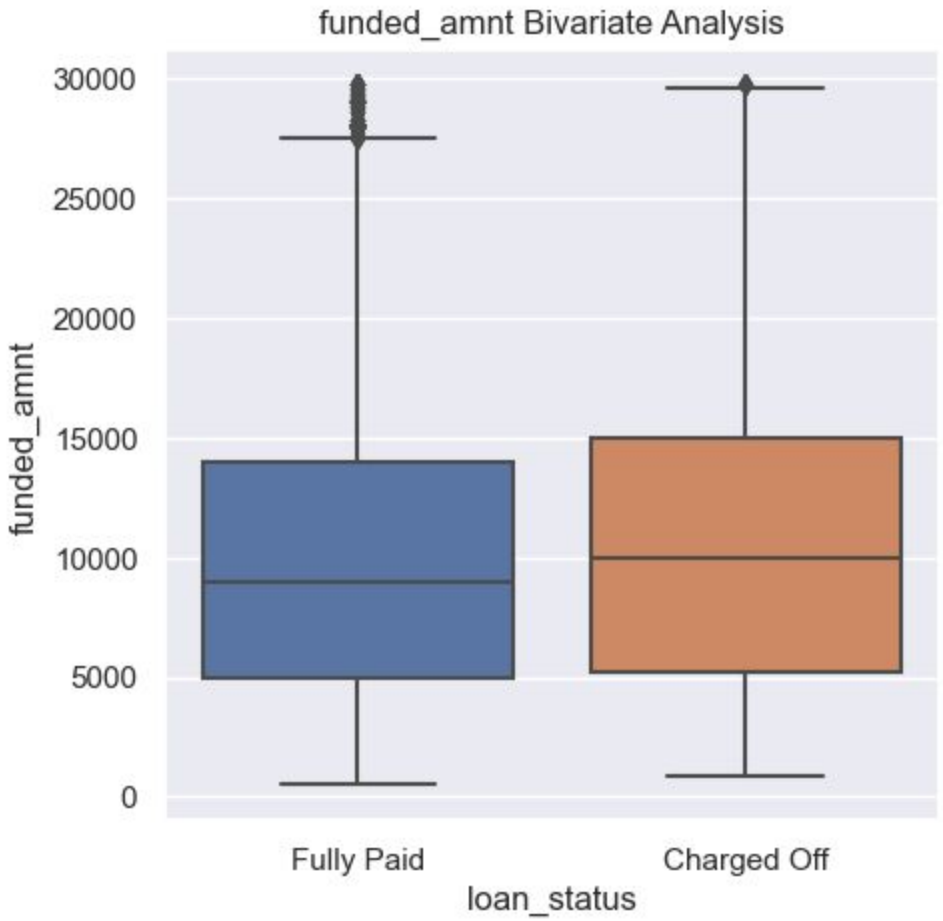
Bivariate and Multivariate



With loan amount, customers who are paying or defaulting doesnt have much difference, however still it can said that higher the loan amount higher are the chances of defaulting

Data Analysis

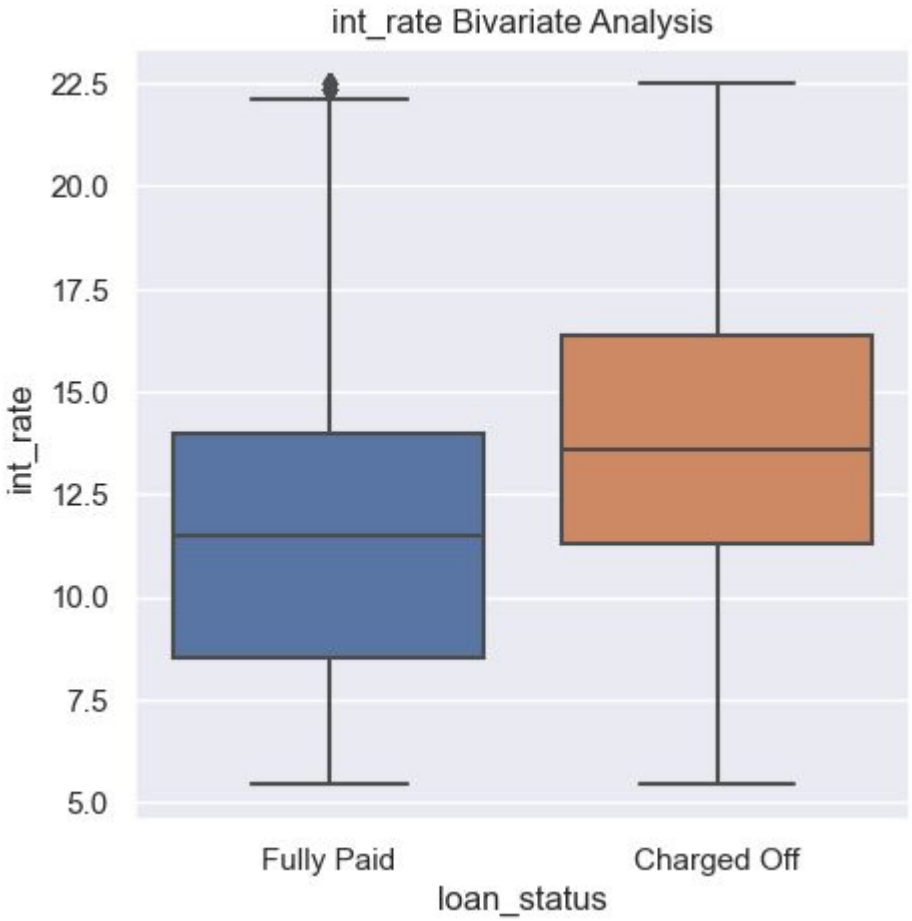
Bivariate and Multivariate



With funded amount, customers who are paying or defaulting doesnt have much difference, however still it can inferred that Higher the funding provided by Lending Clubs higher chances of defaulting.

Data Analysis

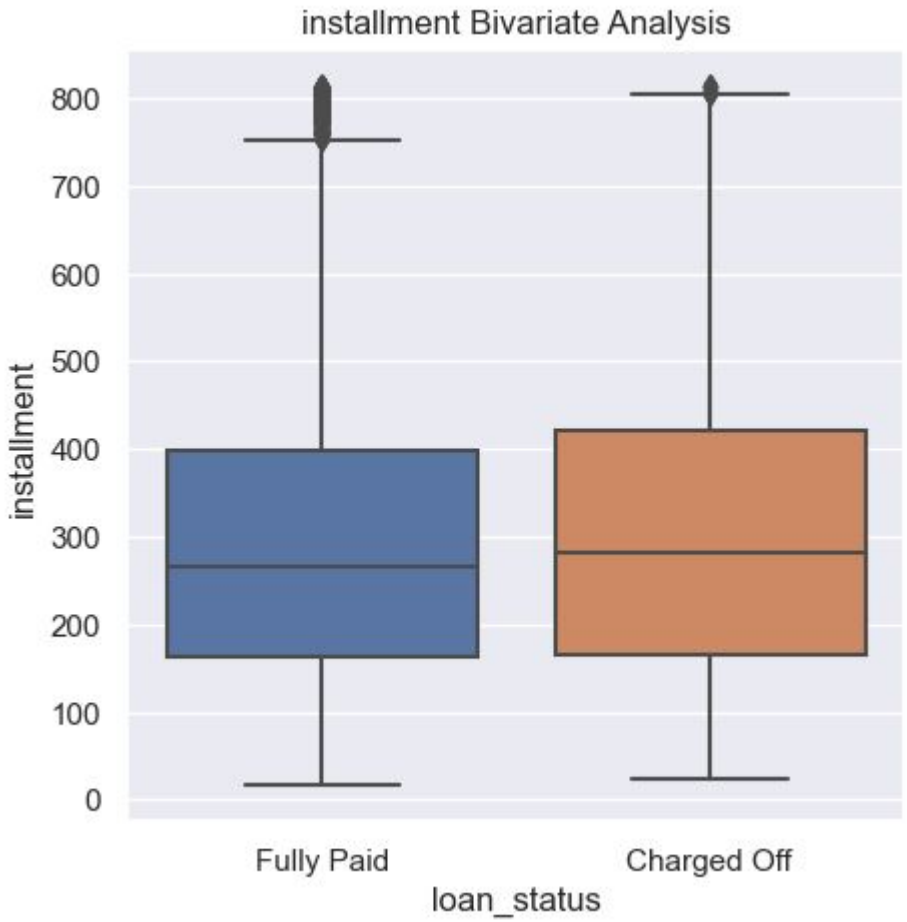
Bivariate and Multivariate



It is quite apparent higher interest rates which have higher risk of defaulting. So, lending clubs should keep interest rate low.

Data Analysis

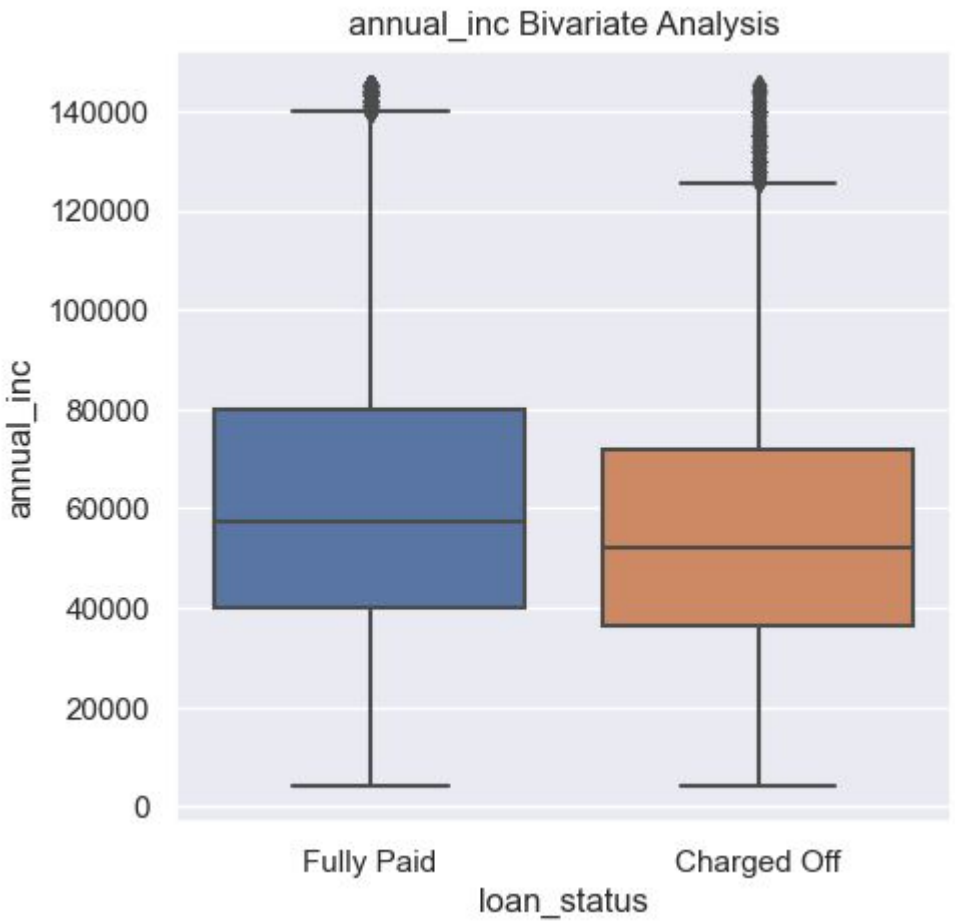
Bivariate and Multivariate



There isnt much difference in installments paid by fully paid customers or defaulters. So this column does not provide a clear distinction between defaulters and fully paid customers.

Data Analysis

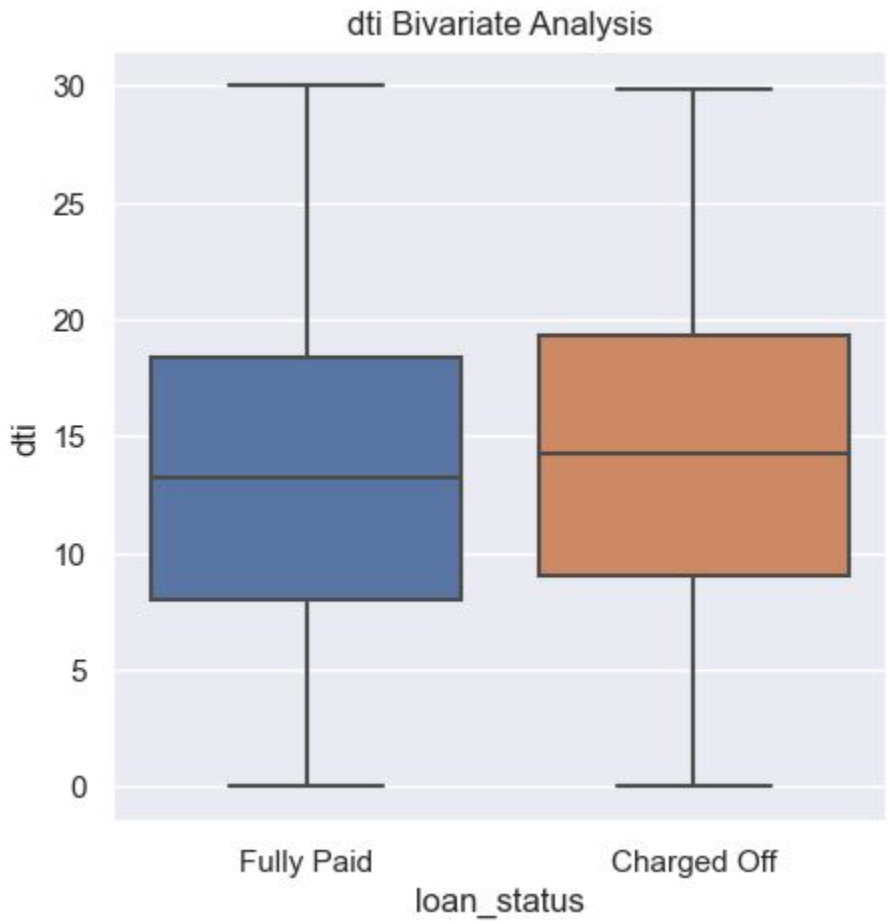
Bivariate and Multivariate



With more annual income, customers are fully paying the loan that they borrowed. People with lesser annual income can be risky customers

Data Analysis

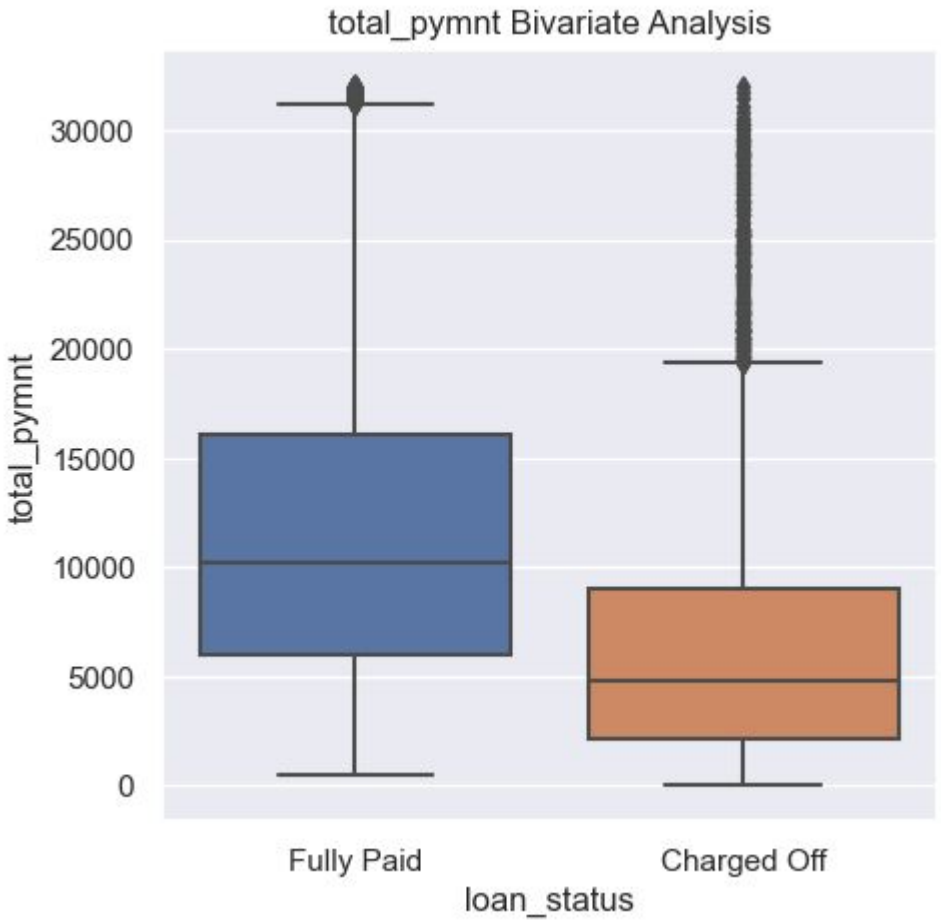
Bivariate and Multivariate



There is less distinguishing factor in dti with fully paying customers and defaulters

Data Analysis

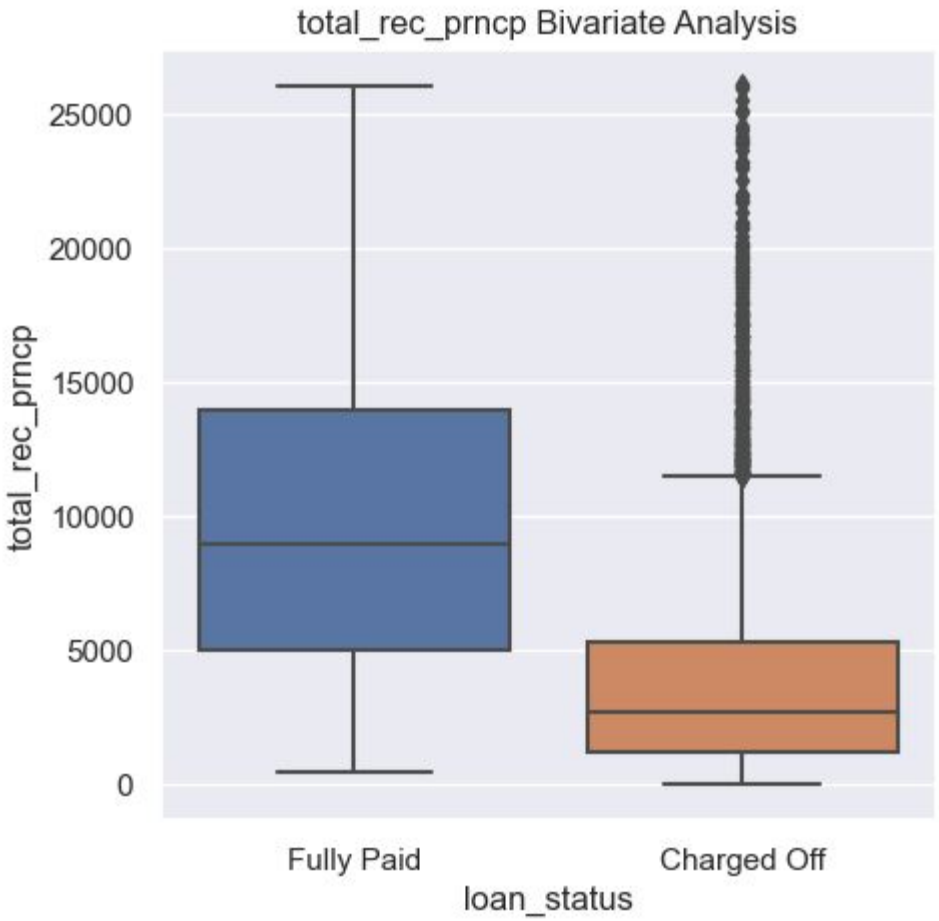
Bivariate and Multivariate



Customers/borrowers whose total_payments are more are good customers. The ones with lesser total_payments are risky customers and are more prone to default.

Data Analysis

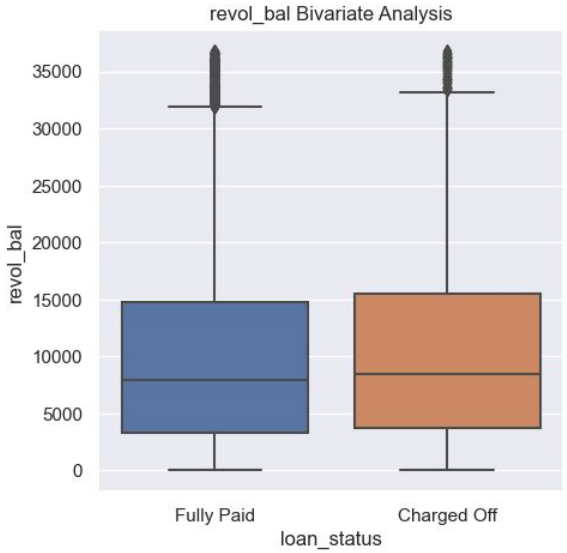
Bivariate and Multivariate

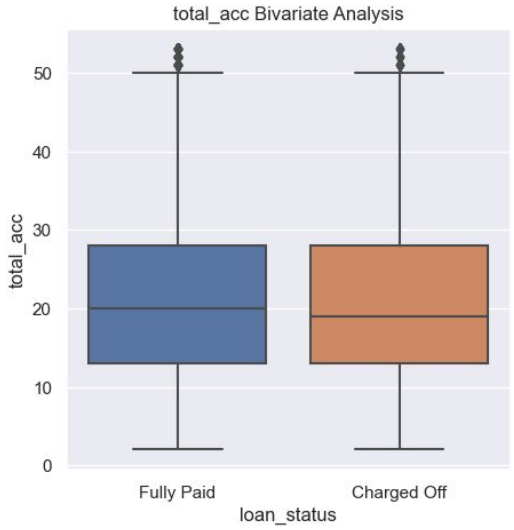


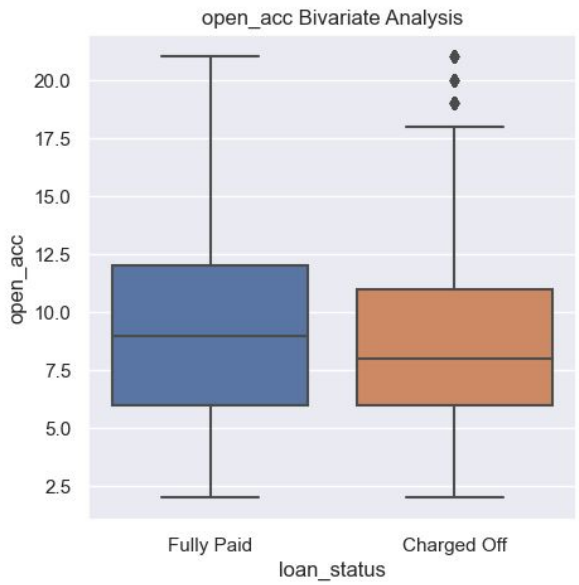
Higher the total principal amount higher changes of fully paying the loan and vice-versa. Customers with lesser value in total_rec_prncp have more chance of defaulting.

Data Analysis

Bivariate and Multivariate







open_account might not be good column for analysis as there is no difference in terms open accounts and fully paying or defaulting. Customers who more total_accounts are slightly better than those who do not have much accounts.

Data Analysis

Heatmap

Data Analysis

Heatmap

Loan amount and funded amount are highly correlated which indicates that loan_amount borrowed by the borrower and funded by the lender are almost similar or in proportion.
total_payment and total_rec_prncp are highly correlated which indicates that total_payment made will be higher for higher Principal received till date.
Total_payment made by the borrower also leads
loan_amnt and installments are highly correlated which means higher the loan amount higher the installment.
Loan amount, installment and total payment are highly correlated which indicates higher the loan taken, more number of installments would be made to repay the loan and accordingly total payment will also be high.

Loan amount and interest rate are not highly correlated which indicates that the interest rate would depend on how the bank or market is performing irrespective of the loan amount issued.

Total account and open account are not highly correlated to loan amount which indicates that the amount of loan taken or issued does not consider how many accounts customer has or opened.

Conclusion

- Post entire analysis, it is noticed that there are feature variables that can aide financial companies to identify defaulters
- Although these feature variables are marginally differentiate between fully charged and charge off categories, they do have peak point identification factors such as lending out minimal interest rates to applicants or based on grades of applicants, more term can be granted or based on demography and its spending power, applicants could have lesser stringent obligations to pay off their loans including customer with more number of accounts.
- Customers with higher loan amount and loan taken for small business ,belonging to grade E and F are risky customers.