LENDING CLUB CASE STUDY

ASHITA JAIN

ASIMANANDA MOHANTY

AGENDA

- Introduction
- About The Dataset
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Conclusion

INTRODUCTION

- The Lending Club Case Study is about a finance company, which mainly deals with lending various types of consumer loans.
- When a loan is given, not all loans are paid back fully. Some of them becomes bad loans.
- We have a task at hand to analyse the bad loans and find out the parameters/criteria that usually lead to bad loans.
- Once the parameters are identified, we must communicate it properly so that corrective actions can be taken i.e., the impacting parameters can be given more weightage before granting a loan.

ABOUT THE DATASET

- The dataset at hand consists of the granted loans only (not the one which are rejected) and it's a mix of loans that are fully paid back, those are running currently and the bad loans (called "Charged Off")
- The dataset has 39000+ records in it and comprises of 111 columns.
- The "loan_status" is the column which indicates the actual state of the loan, and we can treat this as our target variable (TV).
- Next, we will go ahead and start with analysing and cleaning the datset.

DATA CLEANING

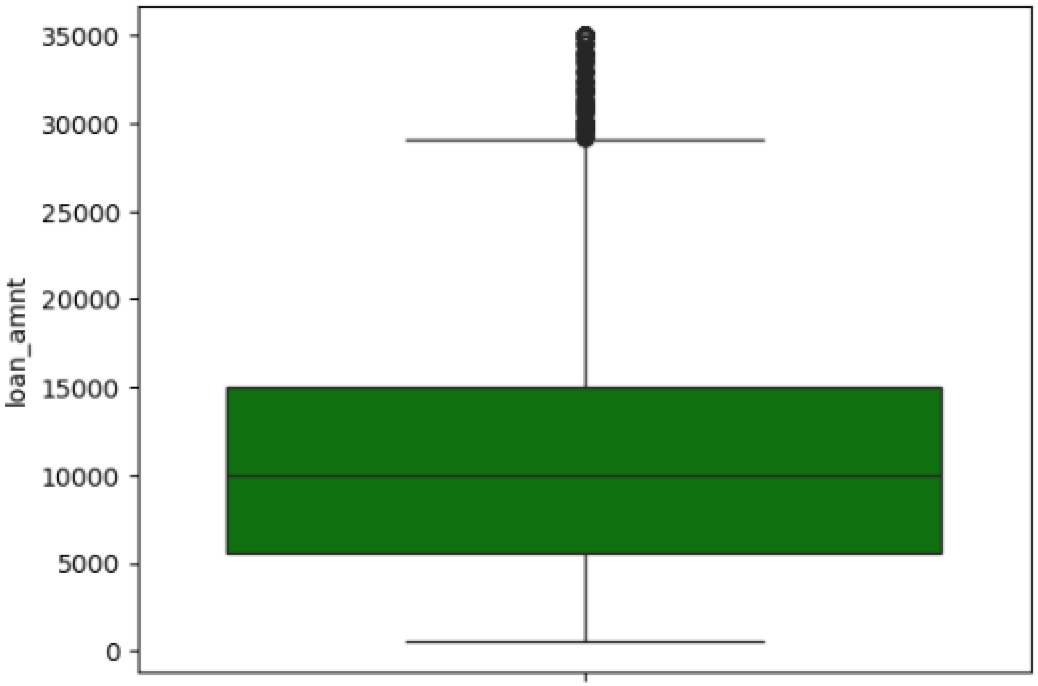
- We found 54 NULL columns, which had no value in analysis. So, these 54 columns were removed.
- Similarly, there were 14 columns which had some Null values in them, and we treated them to make sure we didn't have any Null values left.
  - Few columns were removed (%age of Null values were more or the columns themselves had no significance in our analysis).
  - For a few, the Null values were replaced with Statistical Mode.
  - The rows bearing Null values were deleted for some others.
  - For some columns, the only value it contained was 0 apart from Null values, We decided to delete those columns as well.
  - We found no rows having all 0 values after all the cleaning was done.

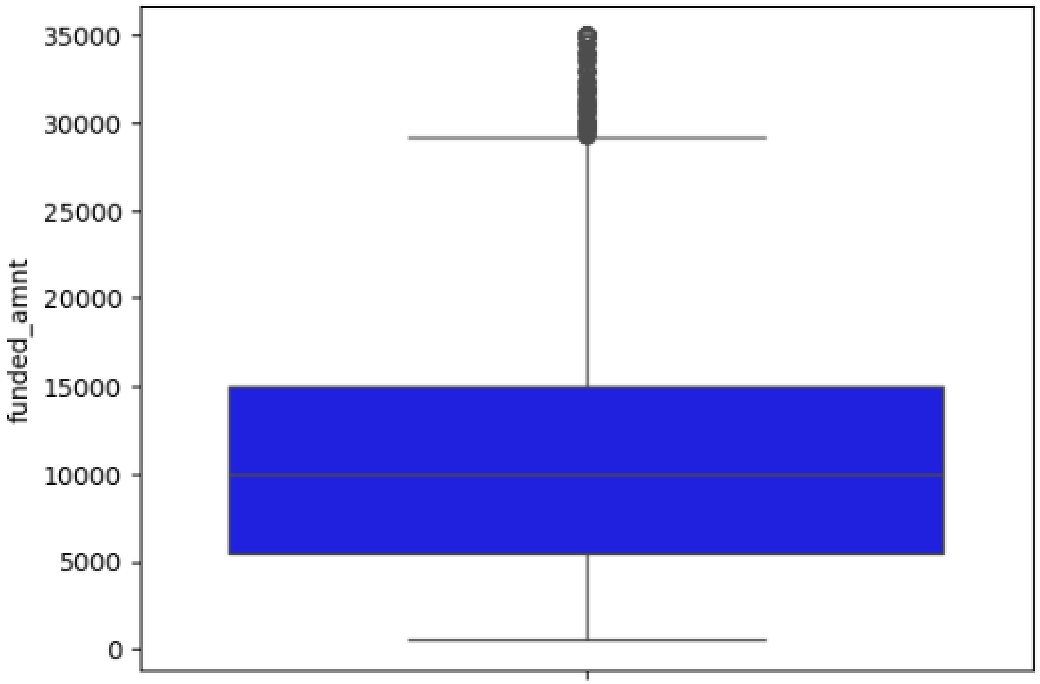
DATA CLEANING

- We then started analysing the individual data types and presence of any unwanted characters in the data.
- Did some changes like removing % symbol from features “int_rate” and “revol_util” etc.
- While checking the individual columns, we found columns with a singular value in it (e.g. “application_type”). This kind of columns are of no use due to lack of variation in data. Those columns were removed from the dataset as well.
- Some categorical variables had huge number of unique values (e.g. “emp_title”). These kind of columns are of no use either. We decided to remove the same too.
- We also dropped columns like “pymnt_plan”, “url” and “zip_code” due to no/low impact on the target variable.
- We maintained the Date columns (“last_pymnt_d”, “last_credit_pull_d” etc.) in a bid to find some relationship with the month/year of the loan. Also, we did split them into month and year for easier handling.

UNIVARIATE ANALYSIS

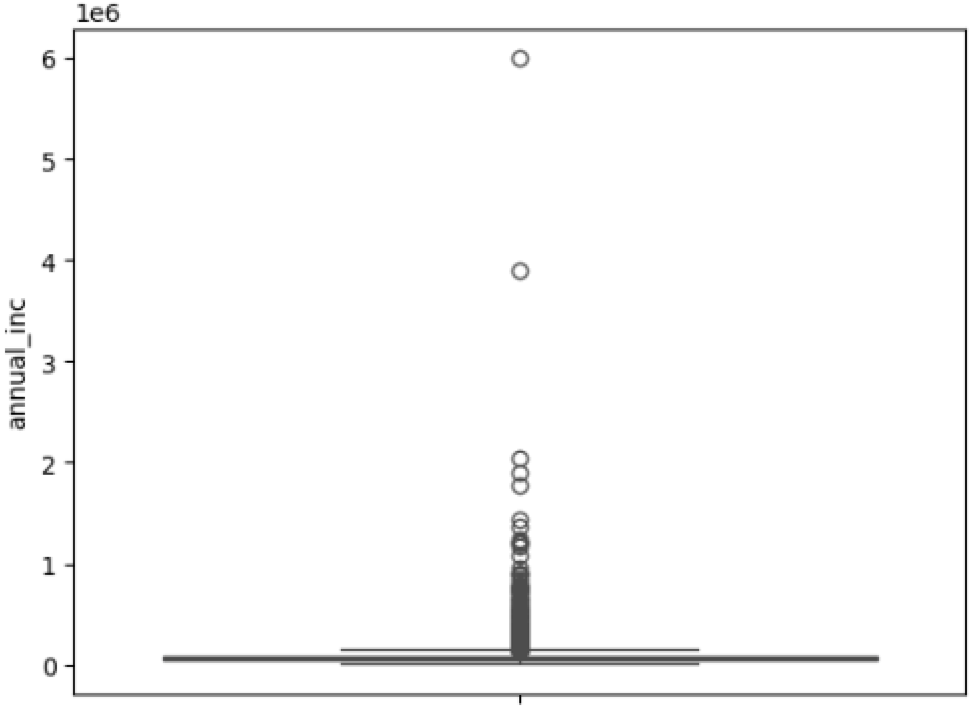
- For easier analysis, we separate numeric and categorical variables into 2 separate data frames (temporary).
- Used pandas “describe” method and plotted Box Plots for the numerical variables to find outliers in them.

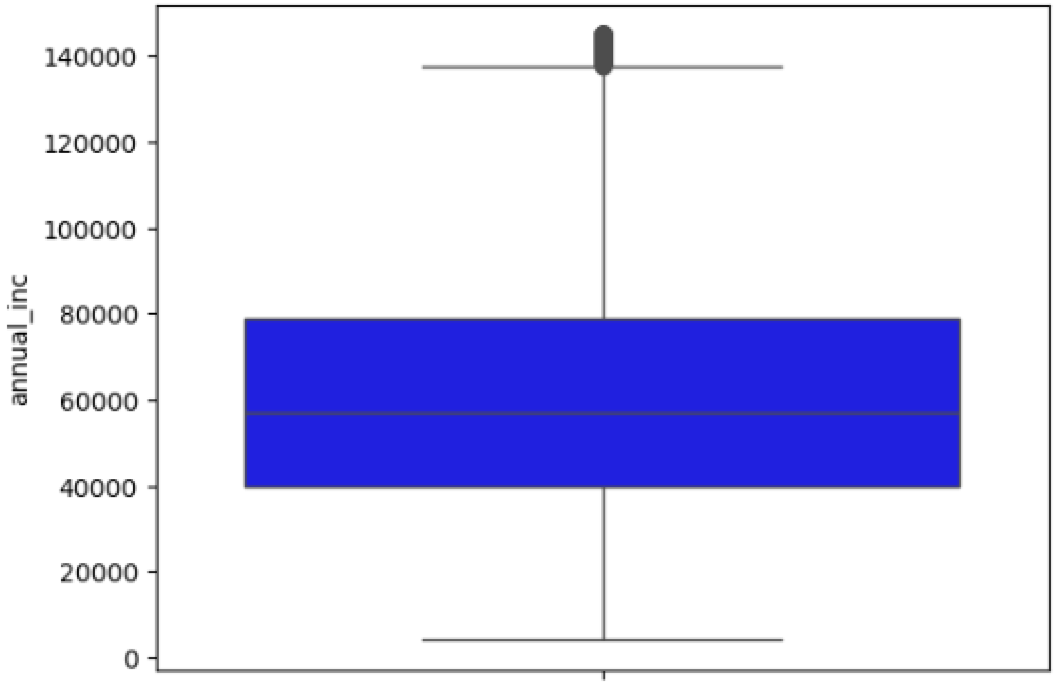




UNIVARIATE ANALYSIS

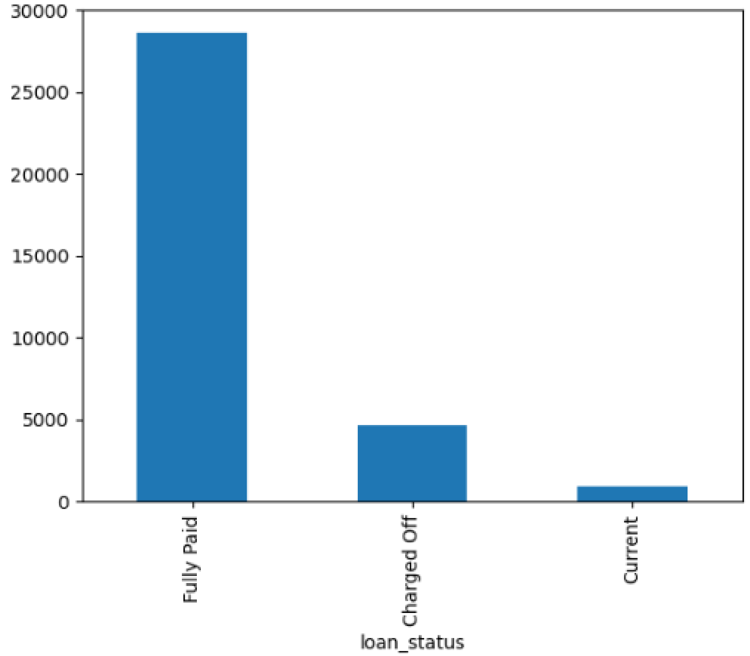
- For some, the outliers were in tolerance range. But for the few others, we found huge outliers.
- Applied IQR method to detect and clean the outliers.
- The following 2 diagrams show the Box Plots for the same variable “annual_inc” before and after outlier correction

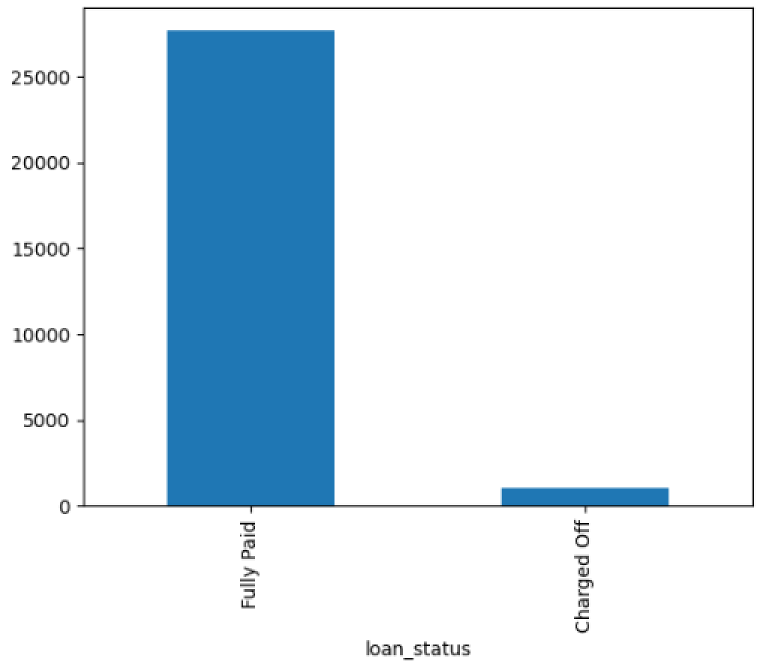




UNIVARIATE ANALYSIS

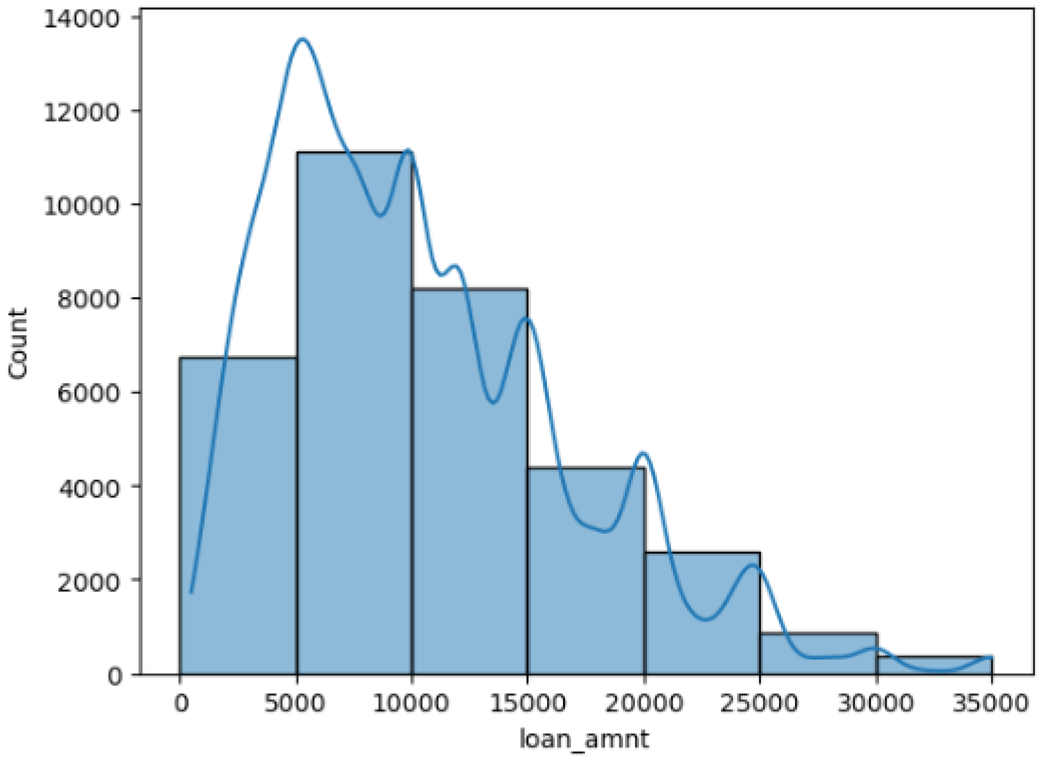
- The similar outlier treatment was done with many other variables too (e.g. “installment”, “delinq_2yrs” etc.)
- For some variables like “out_prncp” and “out_prncp_inv”, we found huge outliers too. But removing them caused more harm than good to the target variable “loan_status” (as can be seen from the plots below)
- To measure this, we started using 2 data frames in parallel for a while.
- At the end, it was seen that the target variable is severely impact after outlier removal for few variables like “out_prncp” and “out_prncp_inv”
- We decided to maintain them and later, the entire columns can be dropped if deemed suitable.

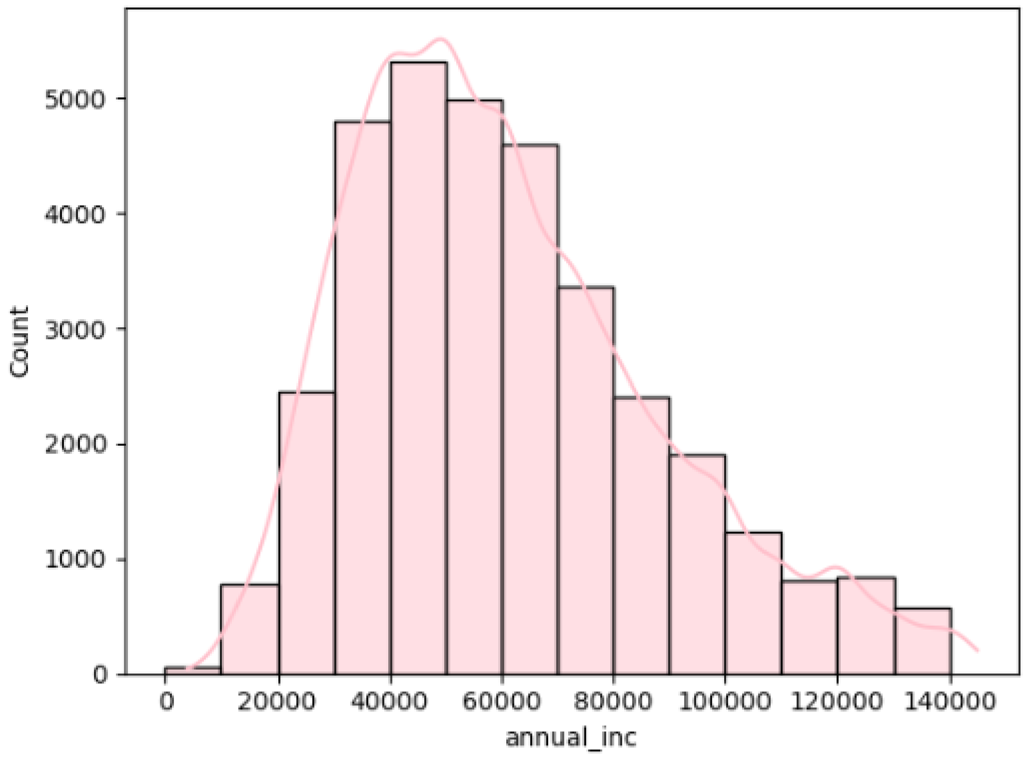




UNIVARIATE ANALYSIS

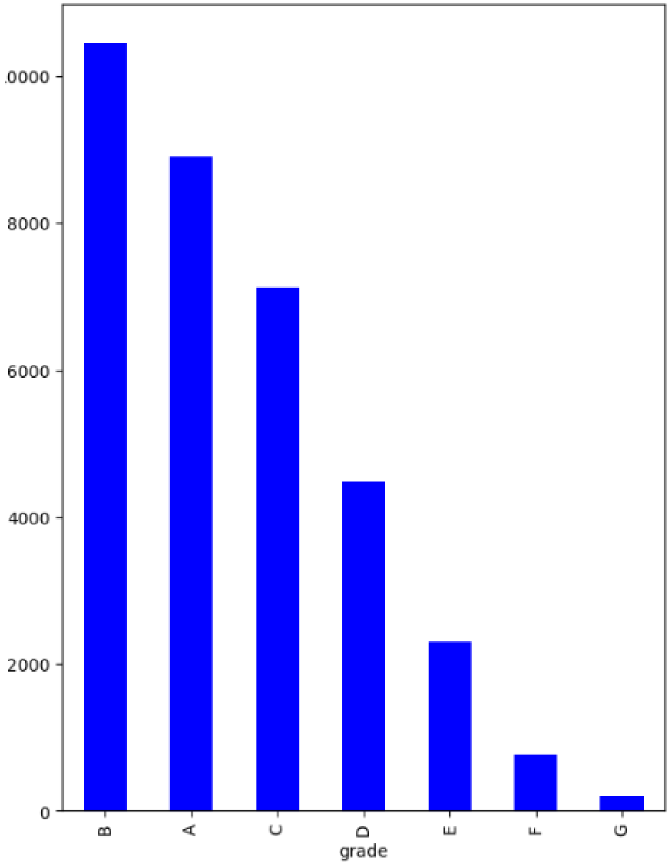
- Histograms were plotted for the numeric variables (some with custom bin size) to check the distribution of them.

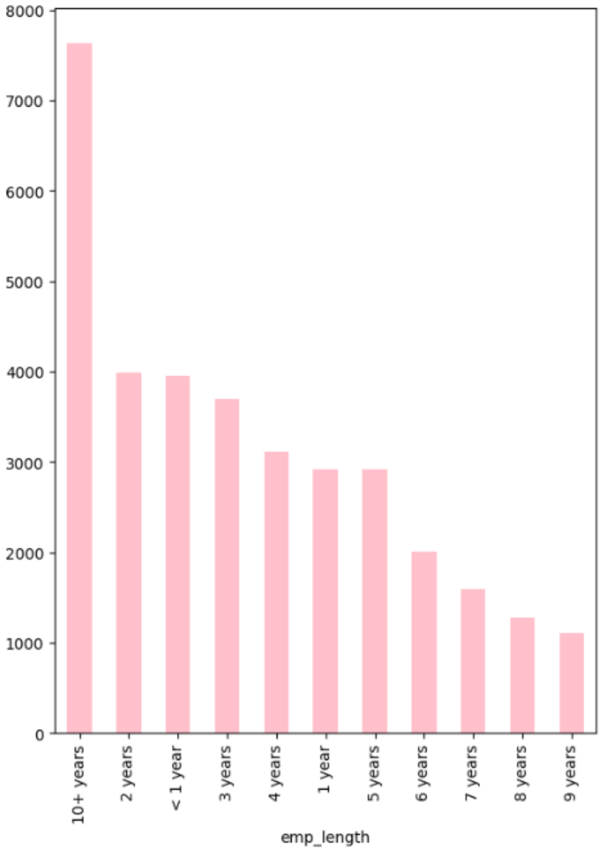




UNIVARIATE ANALYSIS

- Histograms were plotted for the categorical variables also to check the distribution of them.





UNIVARIATE ANALYSIS - OBSERVATIONS

Some Observations:

- "total_pymnt" and "total_pymnt_inv" exhibit similar behavior. Thus, there's not much difference in the regular account v/s accounts funded by investors
- Maximum people have experience more than 10 years
- Maximum accounts are from state CA
- The highest number of accounts are assigned Grade-B
- Max people who took loan are staying in Rented accommodation
- Maximum number of people have taken loans to clean another loan and to clearn credit card bills
- For max accounts, income source was not verified before giving loans
- A total of 13.6% loans are Charged Off (Bad Loans)

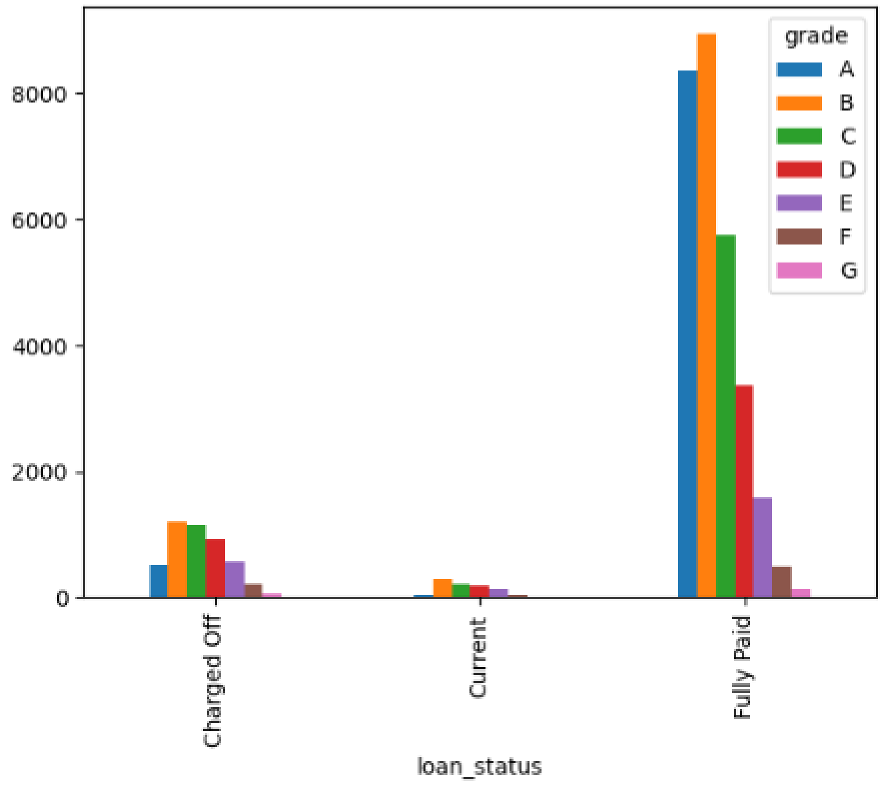
BIVARIATE ANALYSIS

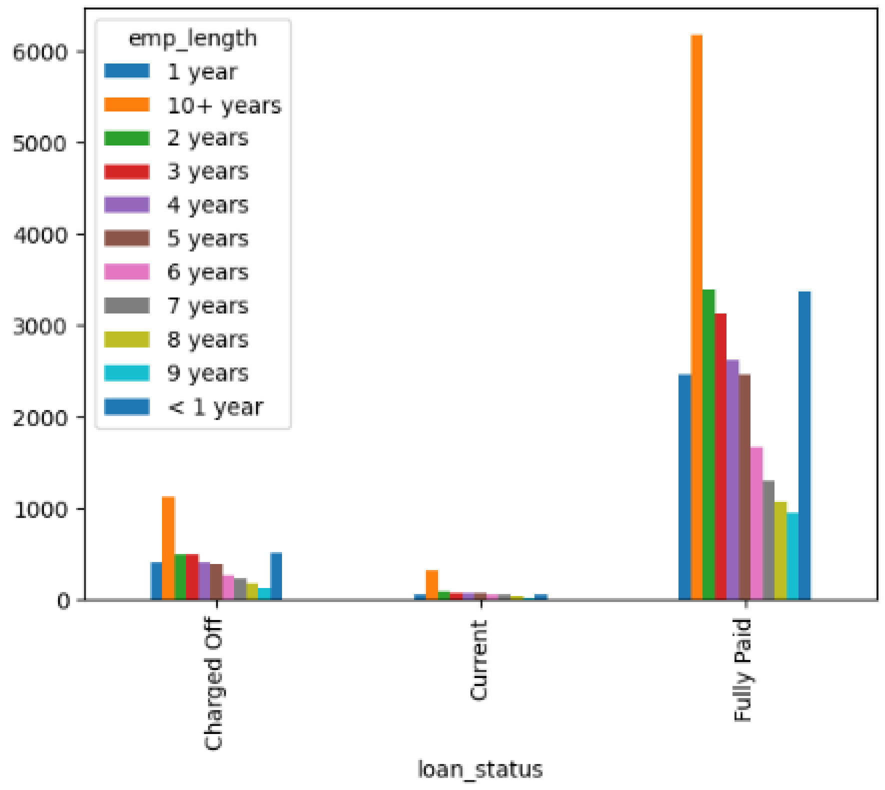
We plotted heatmap and following are some observations.

- Extreme negative correlation is absent in the dataset.
- Total Payment Received has a high Correlation with Loan and Funded Amount. More is the Loan/Funded Amount, more is the payment received.
- But Loan/Funded Amount has relatively lesser (as compared to Payment Received) correlation with Received Interest Amount.
- Outstanding Principal Amount for Total Amount Funded and Total Amount Funded by Investors are positively correlated
- Total Received Late Fee has negative correlation with Total Payment/Total Payment Inv/Total Received Principal, but has positive correlation with Total Interest Received

BIVARIATE ANALYSIS

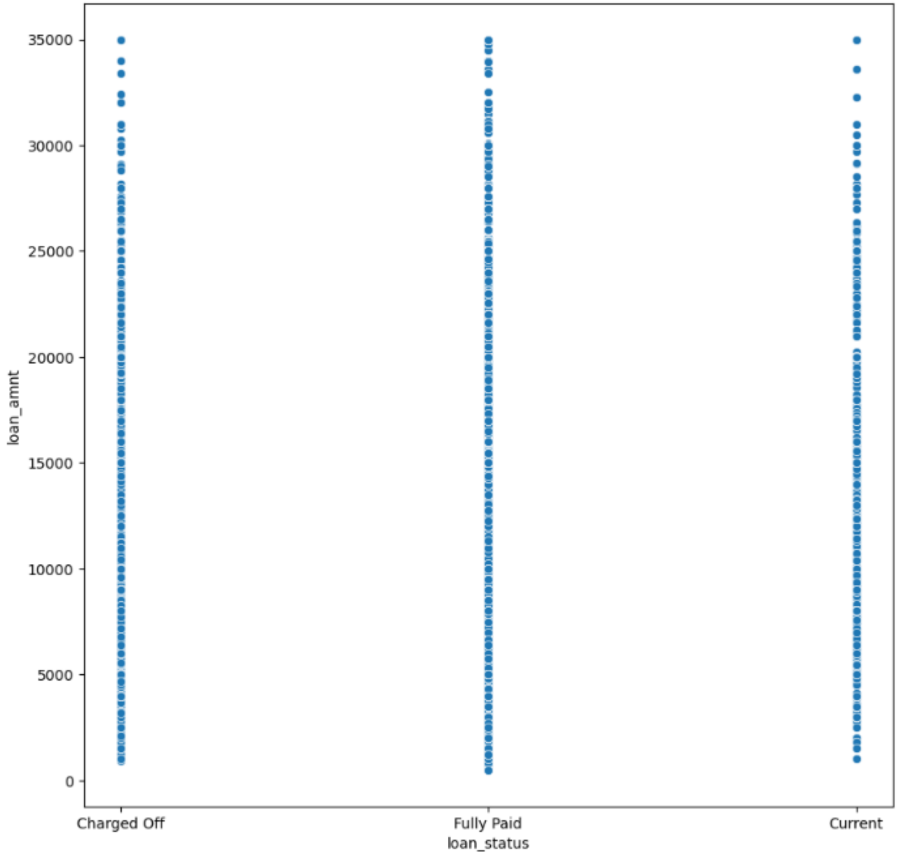
- Then some categorical variables were plotted against each other.

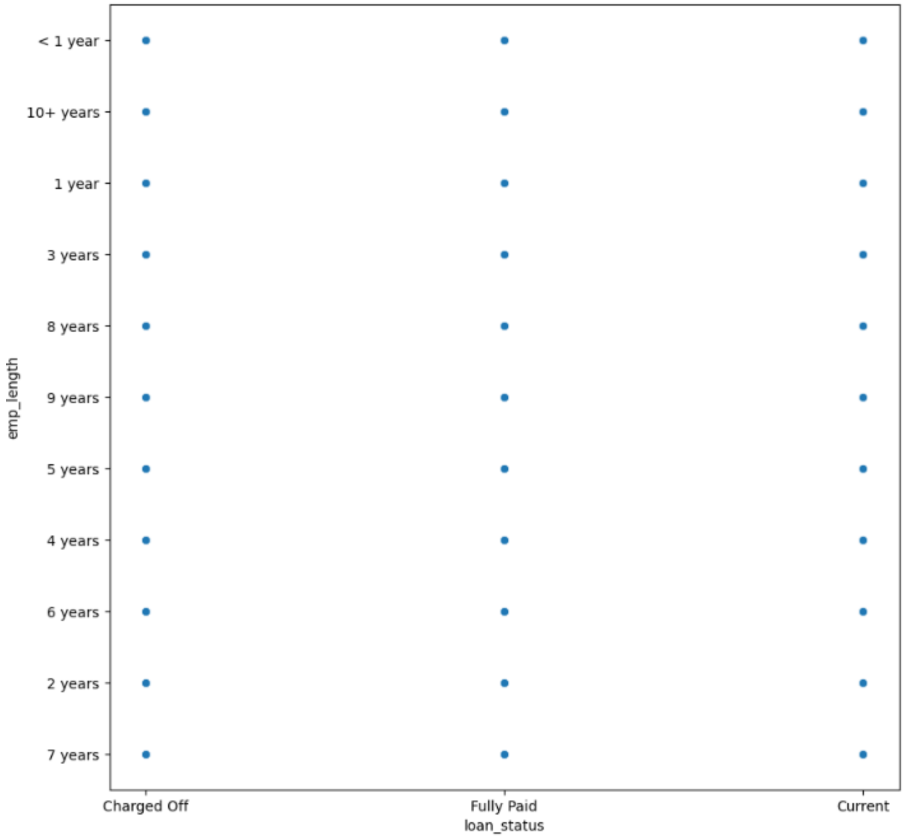




BIVARIATE ANALYSIS

- Numerical variables are plotted against the target variable, but not much insights were found.



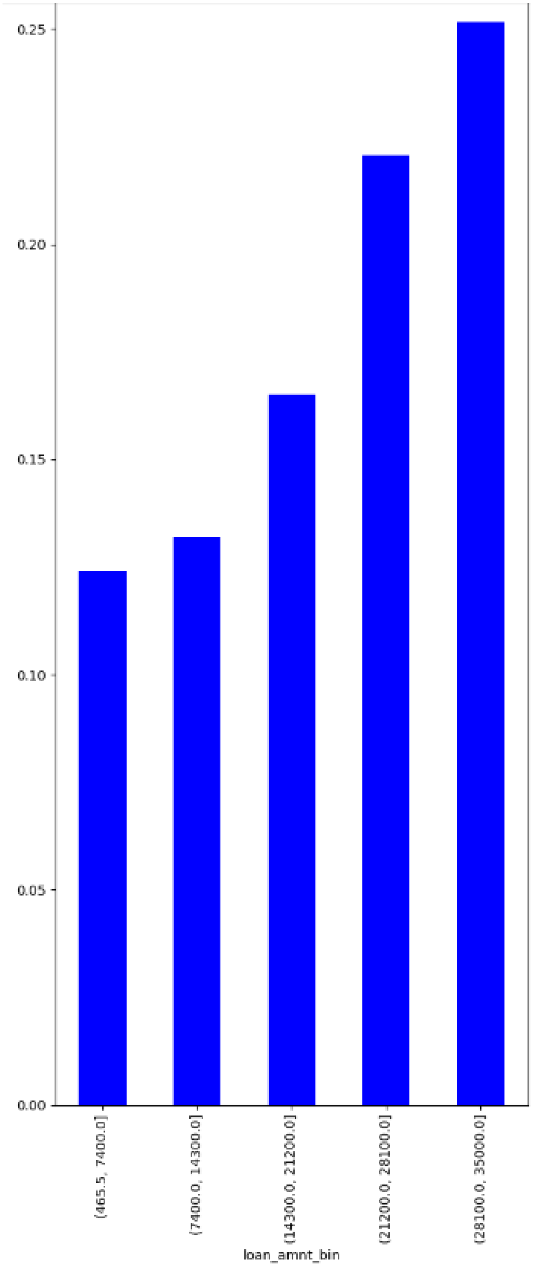


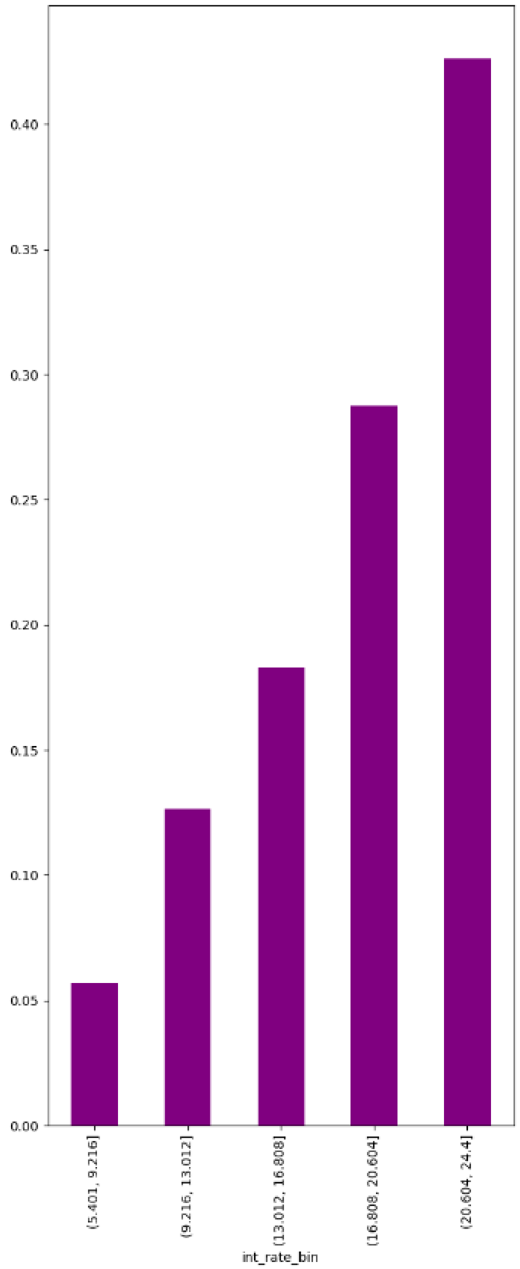
BIVARIATE ANALYSIS

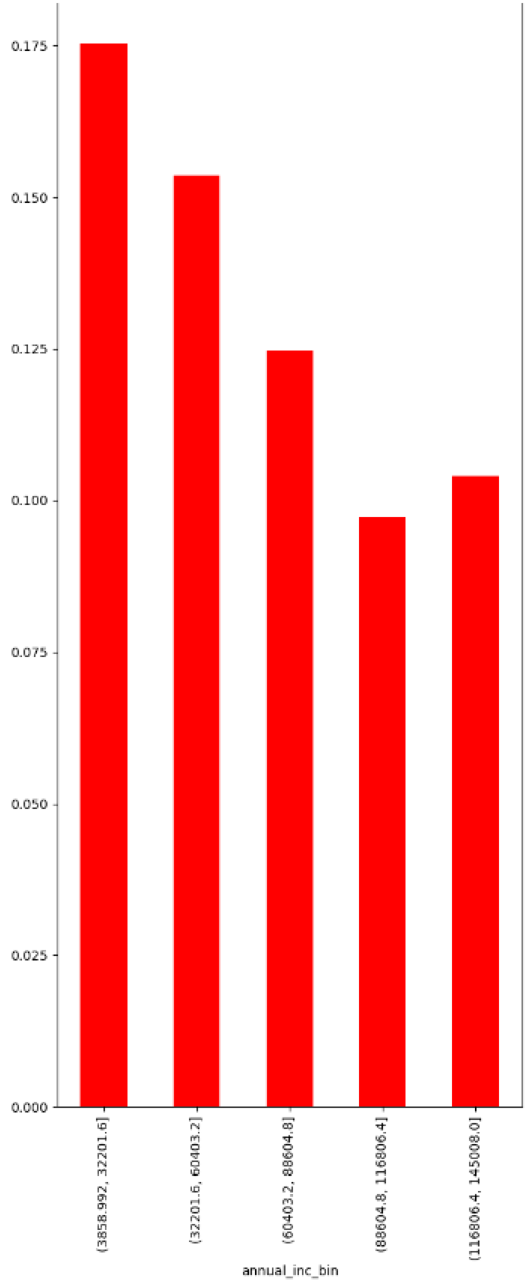
- Then the target variable "loan_status" was encoded into 1 and 0 after removal of "Current" loans (as they give no indication of whether a loan is good or bad). This new column was called "loan_status_co"
- The numerical variables were converted into bins using pd.cut() (as they were to be plotted against a categorical variable).
- Then the average of "loan_status_co" in each bin is plotted (for each numerical variable). This basically gave us the probability of presence of 1's in each bin.
- In other words, it provided us the %age of Charged Off loans against each bin.

BIVARIATE ANALYSIS

- Here are some plots.







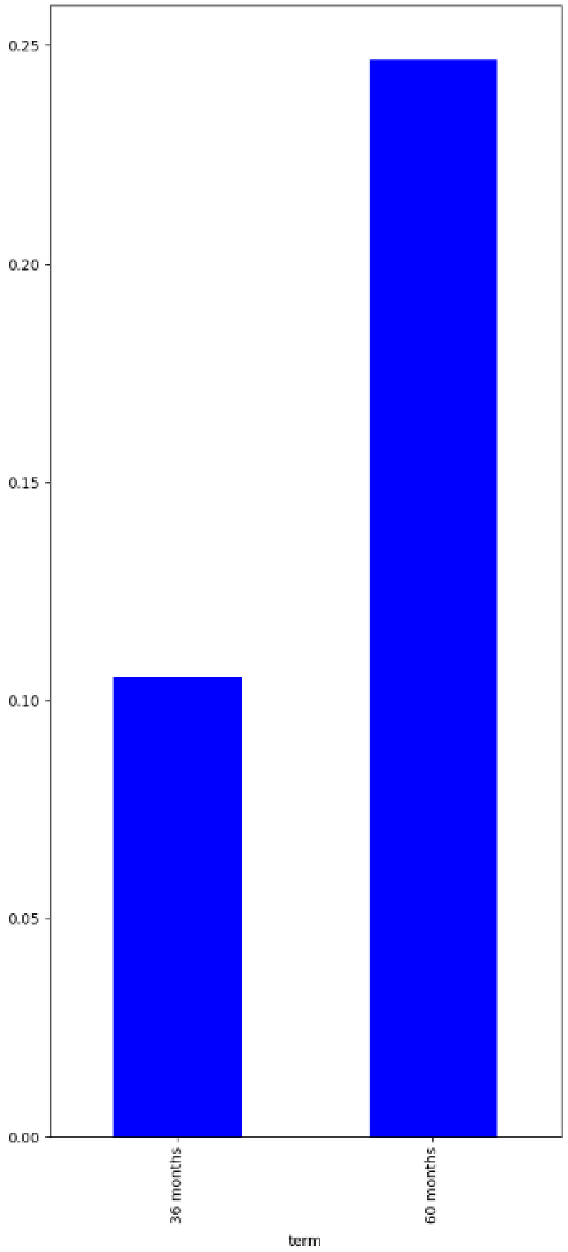
BIVARIATE ANALYSIS - OBSERVATIONS

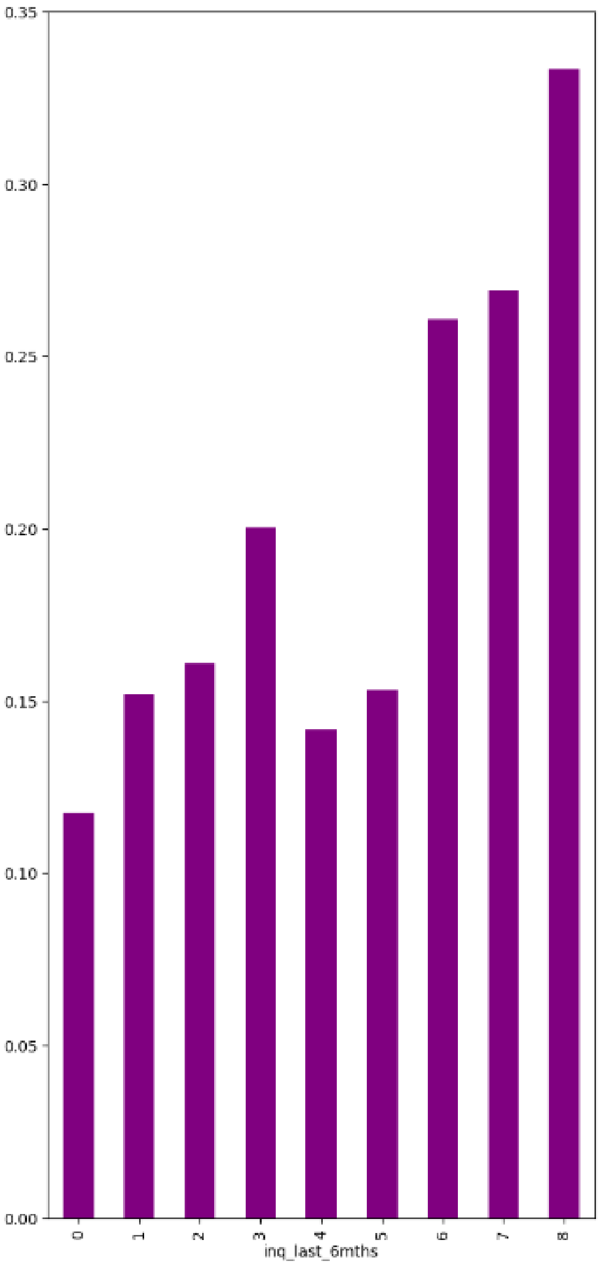
Observations:

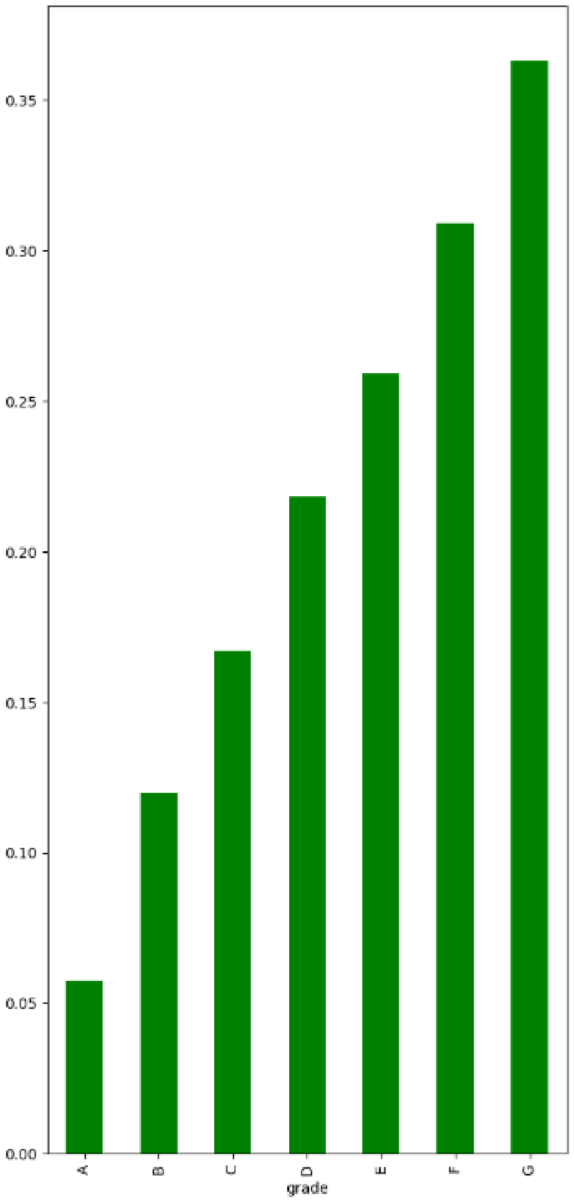
- Higher the loan_amount, funded_amount or funded_amount_inv, higher the chances of bad loans.
- Lower the Total Payment, higher the chances of Charged Off Loans
- Higher Interest Rate leads to Hinger Charged Off Loans
- Lower the Annual Income, higher is the chances of Charged Off Loans
- Higher the DTI, higher is the chance of Charged Off Loans
- Higher the Revolving Line Utilization Rate, higher is the chance of Charged Off Loans

BIVARIATE ANALYSIS

- Then the categorical variables were plotted against the same encoded target variable.







BIVARIATE ANALYSIS - OBSERVATIONS

<u>Observations:</u>

- Longer the loan tenure, higher the chances of Charged Off Loans¶
- Higher the Derogatory Public Record, Higher the chances of Charged Off Loans
- Lower the Grade/Sub-Grade assigned by LC, higher the chances of Charge Off
- Higher the number of public bankruptcies, higher the chance of Charge Off
- Small Business Loans have the higher chances of Charge Off
- Surprisingly, The Verified Loans have higher chances of Charge Off
- One more Surprising observation is related to House Ownership. Mortgaged House Owners have the lowest chances of Charge Off
- More the number of credit inquiries, more is the chance of Bad Loan

Note: Some of the results here might differ from the intermittent ones. This might be because of the removal of "Current" loans from the dataset.

THANK YOU !!!