Lending Club Case Study

Roshni Siddoju
Siva Rama Krishna R

Problem Statement

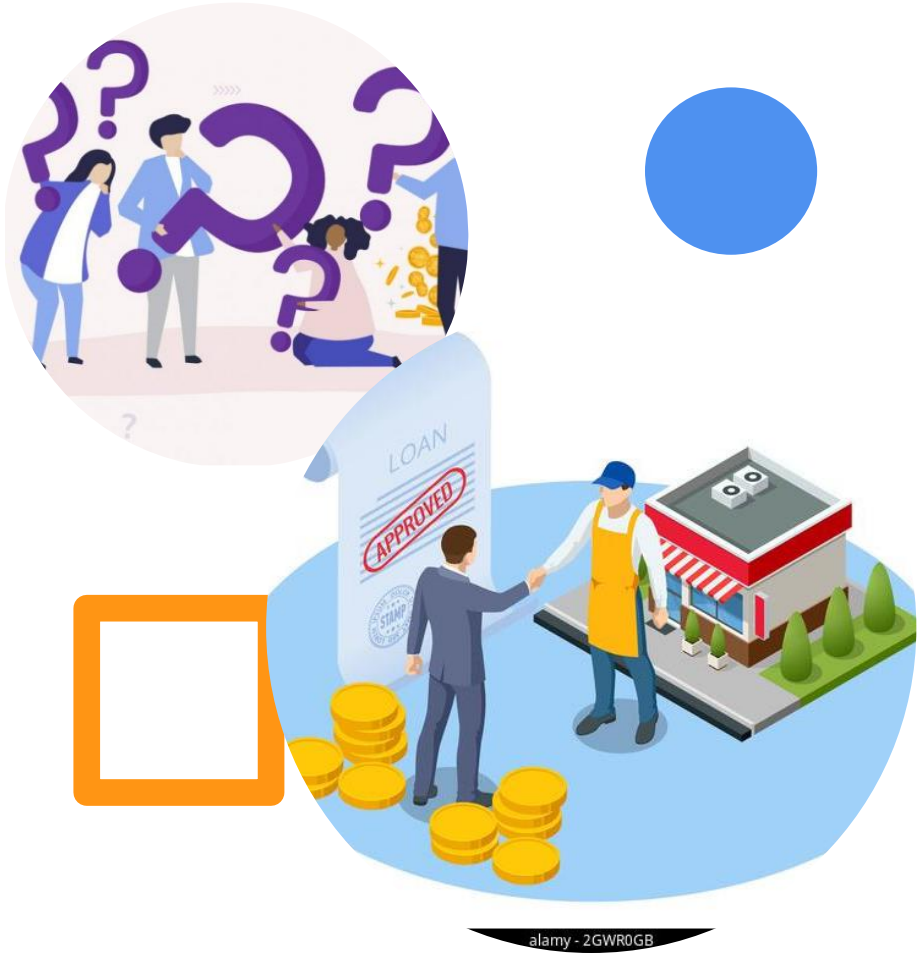
Minimize the risk of losing money while lending to customers

Business Understanding

Consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

•If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company

•If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company



- When a person applies for a loan, there are **two types of decisions** that could be taken by the company:
- **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:
    - **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    - **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
    - **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan
- **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Objective

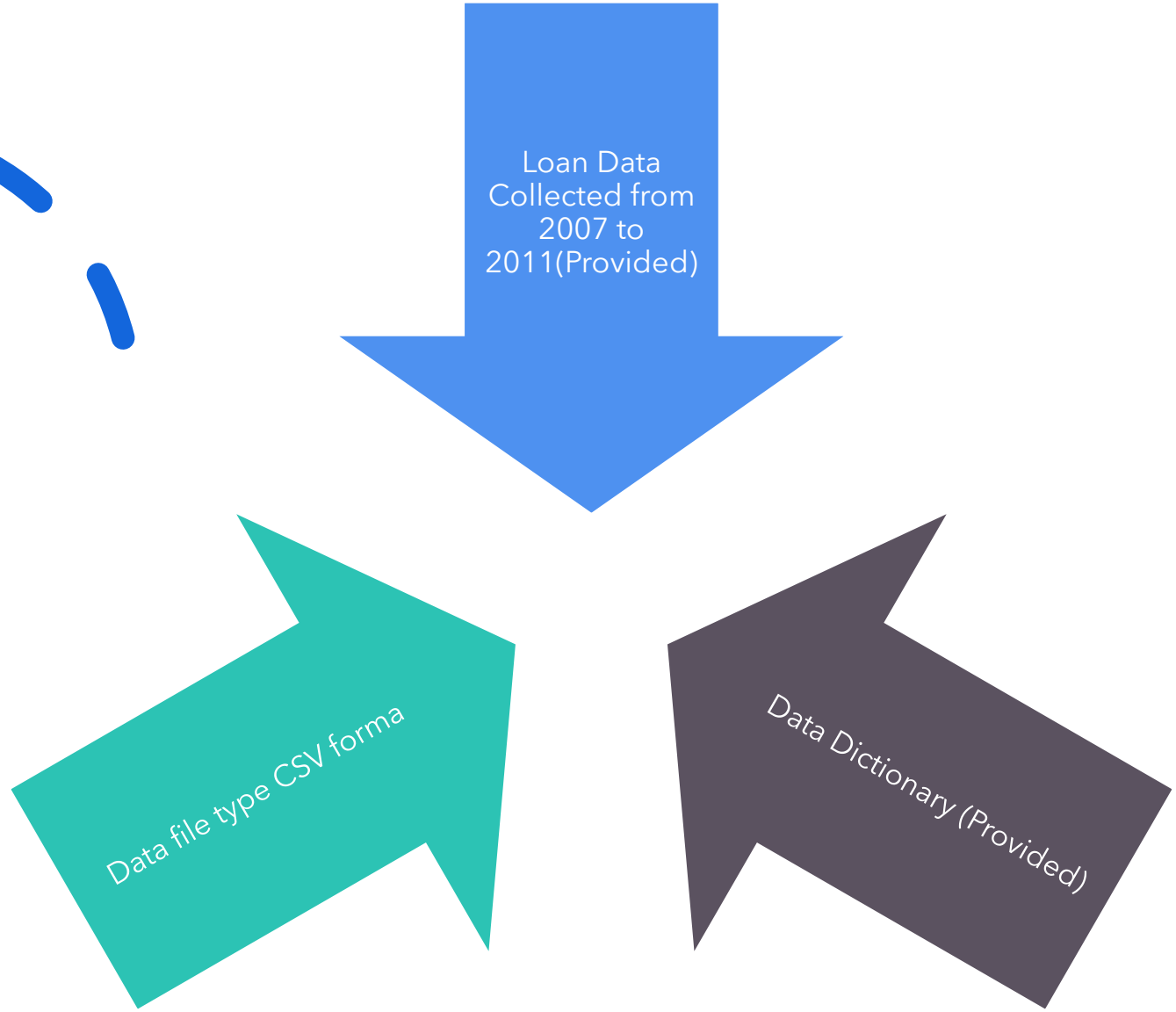
The Company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Approach

EDA Analysis

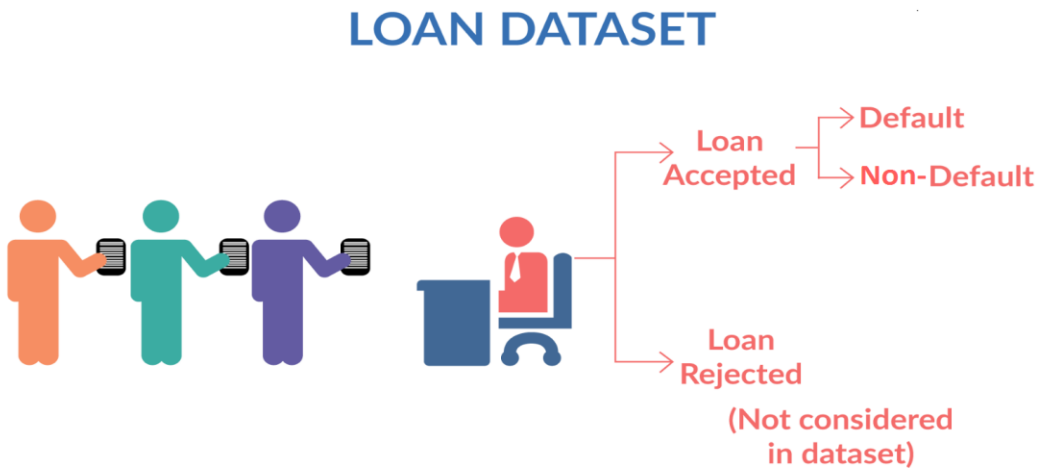
There are five major parts that are needed to be done for this case study:

Data Collection

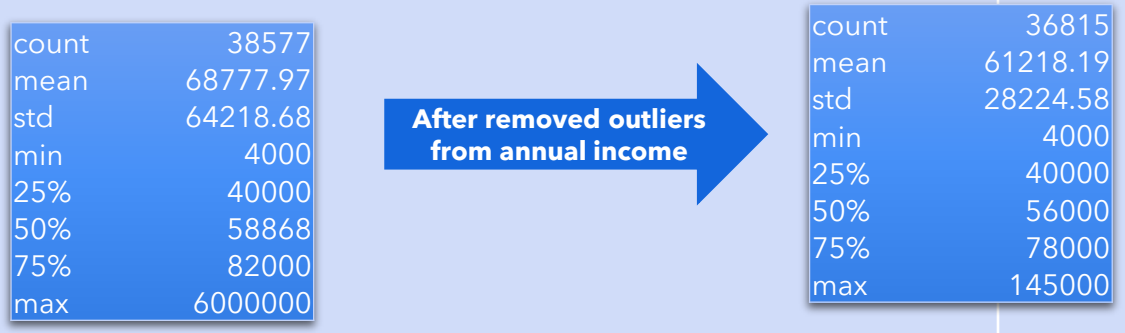


Data Understanding

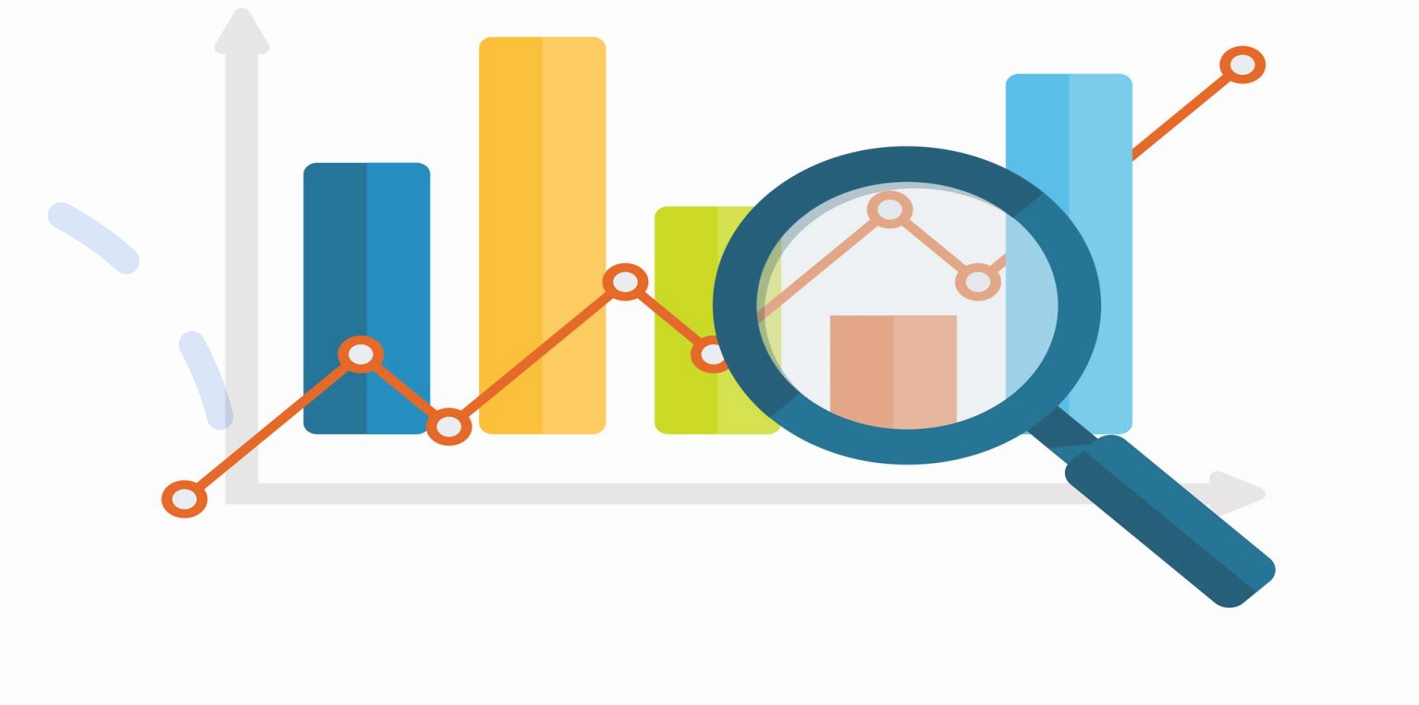
#	Observation	
1	Identified the dataset information	39717 X 111 [Rows X Columns]
2	54 columns in dataset are having 100% NA values	39717 X 57 [Rows X Columns]
3	35 columns does not provide any insights for the given problem statement which includes behavioral columns.	39717 X 24 [Rows X Columns]
4	Observed 1075 NA values in employee length	
5	Observed 697 NA values in pub_rec_bankruptcies	
6	Rest of the 22 columns are segmented into Continuous, Categorical & Id Columns	
	Columns in dataset	id, member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, pymnt_plan, purpose, zip_code, addr_state, dti, open_acc, revol_util, total_acc, pub_rec_bankruptcies
7	Loan Status having 'Current' values, this data is not useful to derive the defaulter , so these data needs to be removed	



Data Cleansing

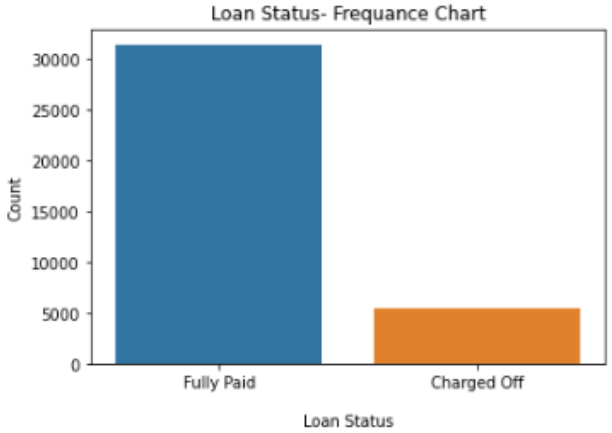
#	Action	Remark
1	Actual dataset shape	39717 X 111 [Rows X Columns]
2	Removed the columns from dataset which are having 100% NA values	39717 X 57 [Rows X Columns]
3	Removed the columns which does not provide any insights for the given problem statement.	39717 X 24 [Rows X Columns]
4	Handled NA values in emp_length with mode (Mode is '10+ years')	1075 values are replaced by '10+ years'
5	Handled NA values in pub_rec_bankruptcies with mode (Mode is 0)	697 values are replaced by '0'
6	Removed the records from dataset where loan status having 'Current'	38577 X 24 [Rows X Columns]
	Removed the outliers from dataset for annual_inc (one of the driving variable) using IQR method, mean that 25% to 75% of IQR range values are included in final data set for analysis 	36815 X 24 [Rows X Columns]
7	Handled NONE values in home_ownership	3 Values are replaced by 'other'

Data Analysis

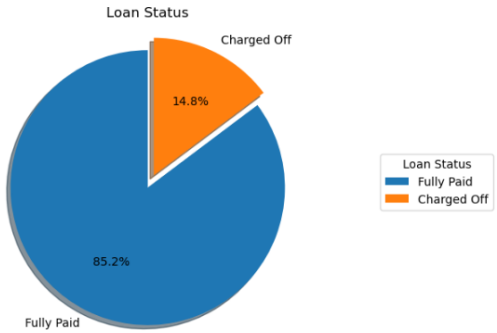


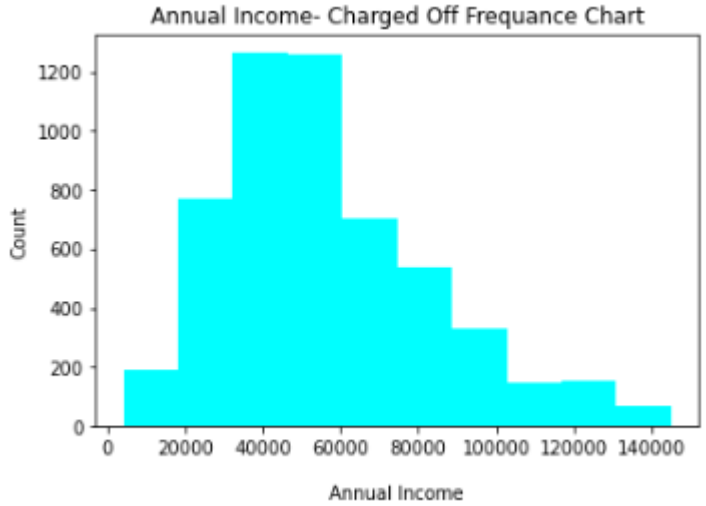
Univariate and segmented univariate analysis
Bivariate analysis

Univariate analysis & Observations for Charged off Data

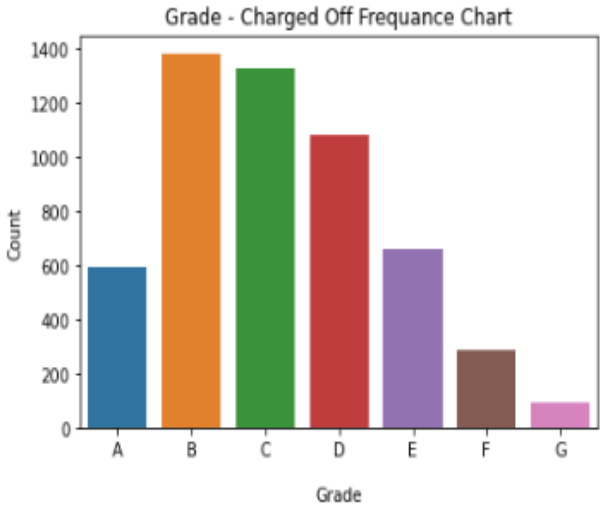


Loan Status frequency

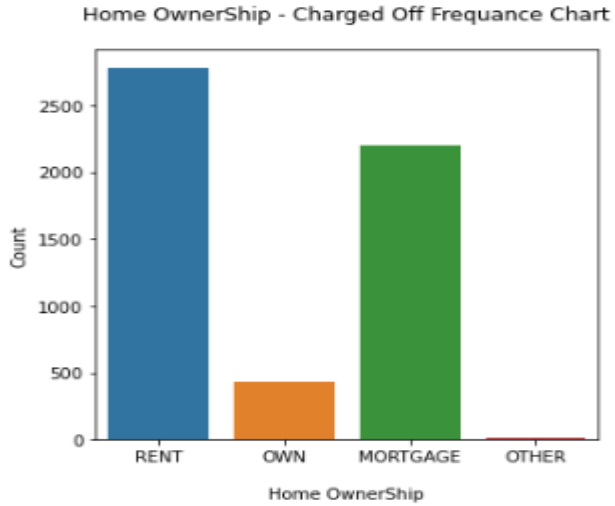




Observation 1 : Individuals who are having annual income has 40000 or 60000 having default chance

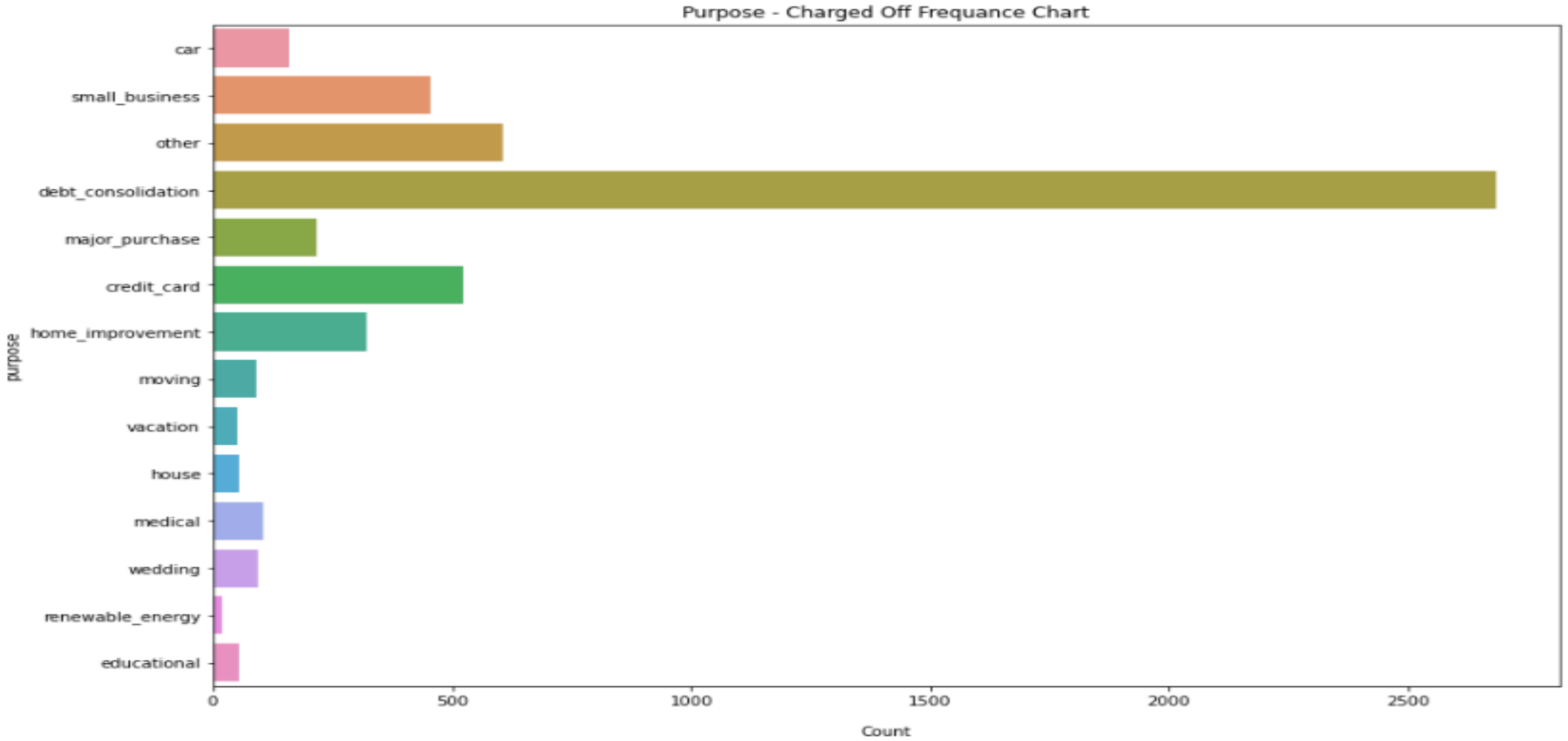


Observation 2 : Most of the Charged-Off Grade are B,C, D



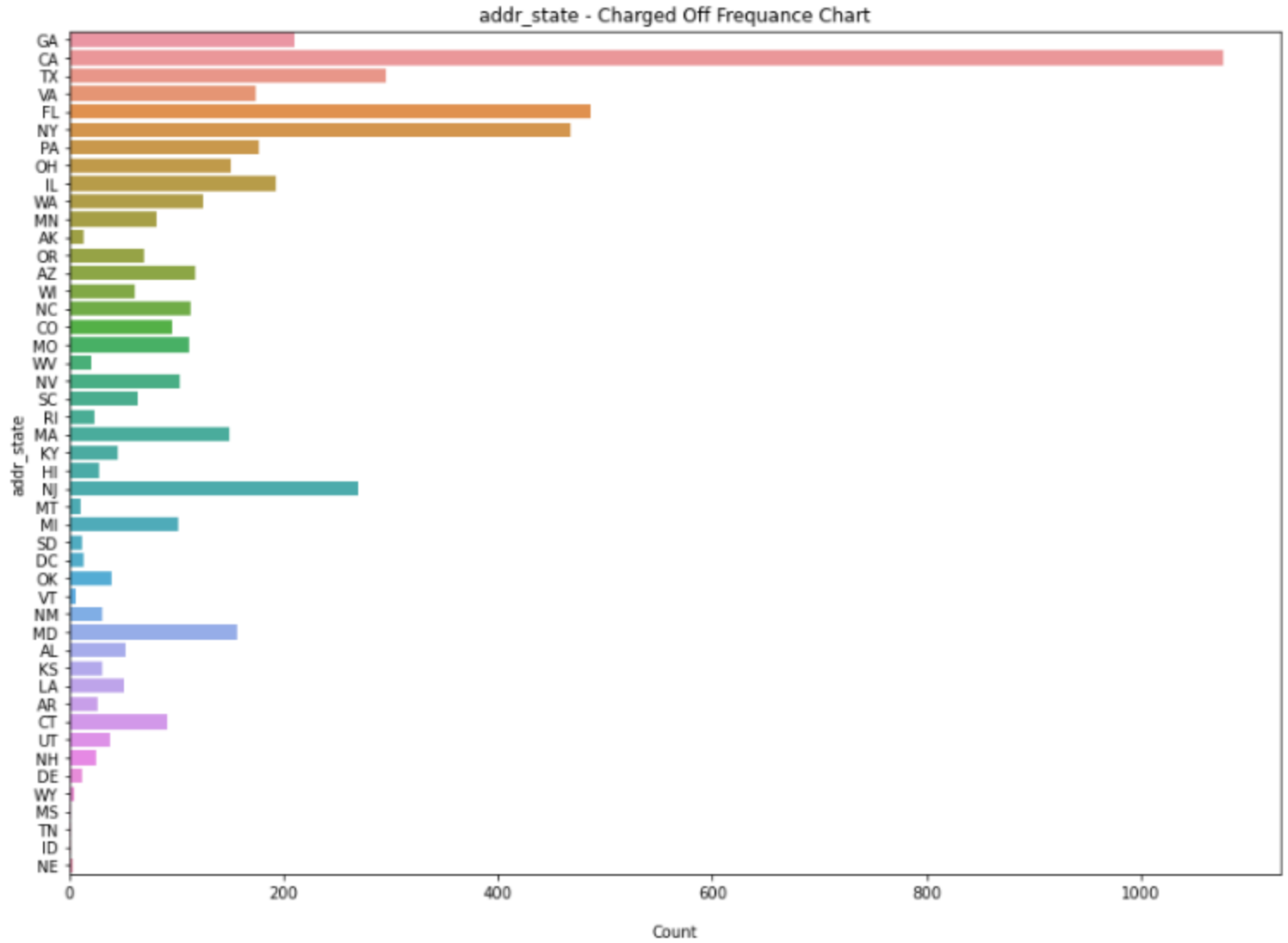
Observation 3 : If the home_ownership is RENT or MORTGAGE then scope of falling into charged off is more

Univariate analysis & Observations for Charged off Data



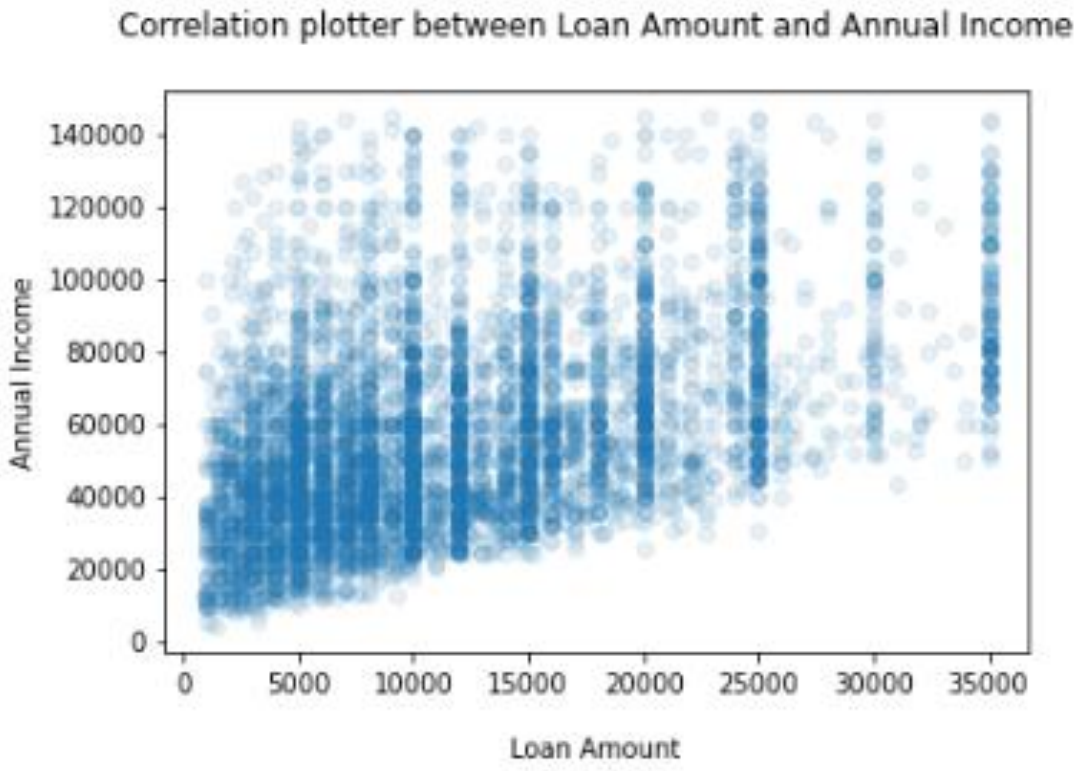
Observation 4 : From the purpose histogram it is clearly seen that debt_consolidation has more in number for falling into charged off

Univariate analysis & Observations for Charged off Data

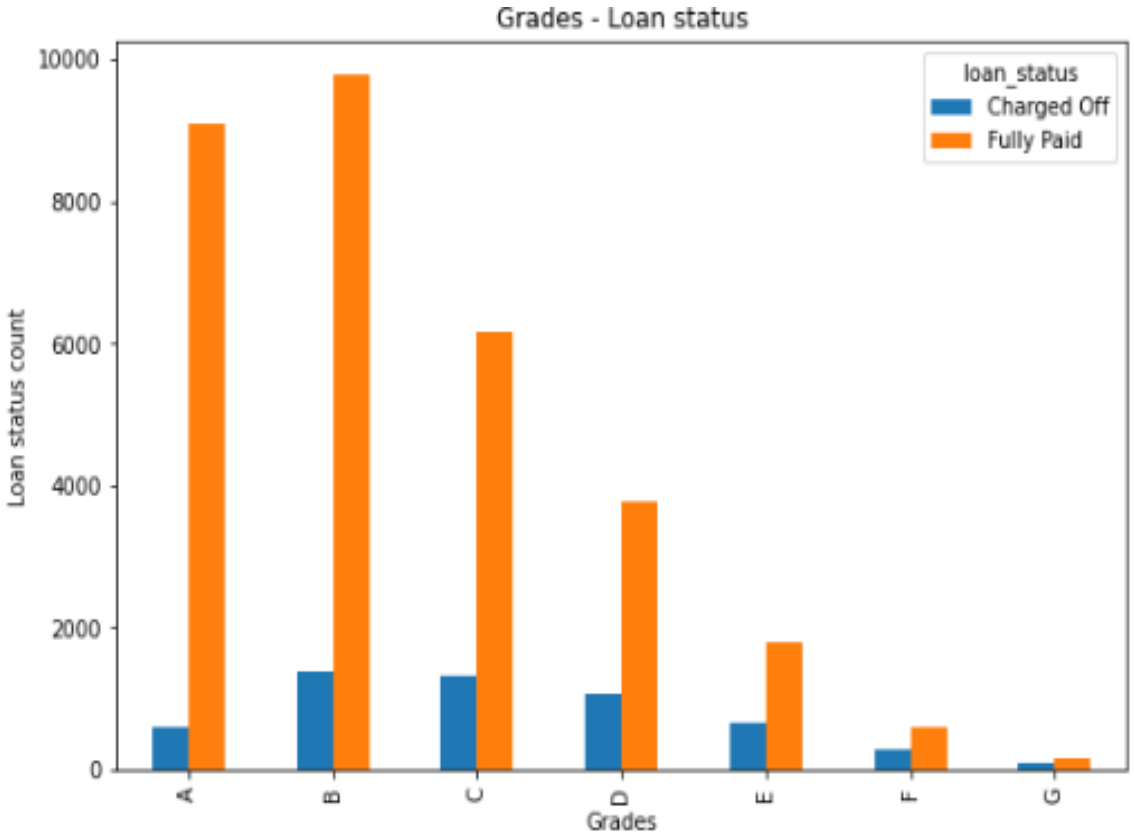


Observation 5 : From the address state CA tend to have more defaulters.

Bivariate analysis & Observations

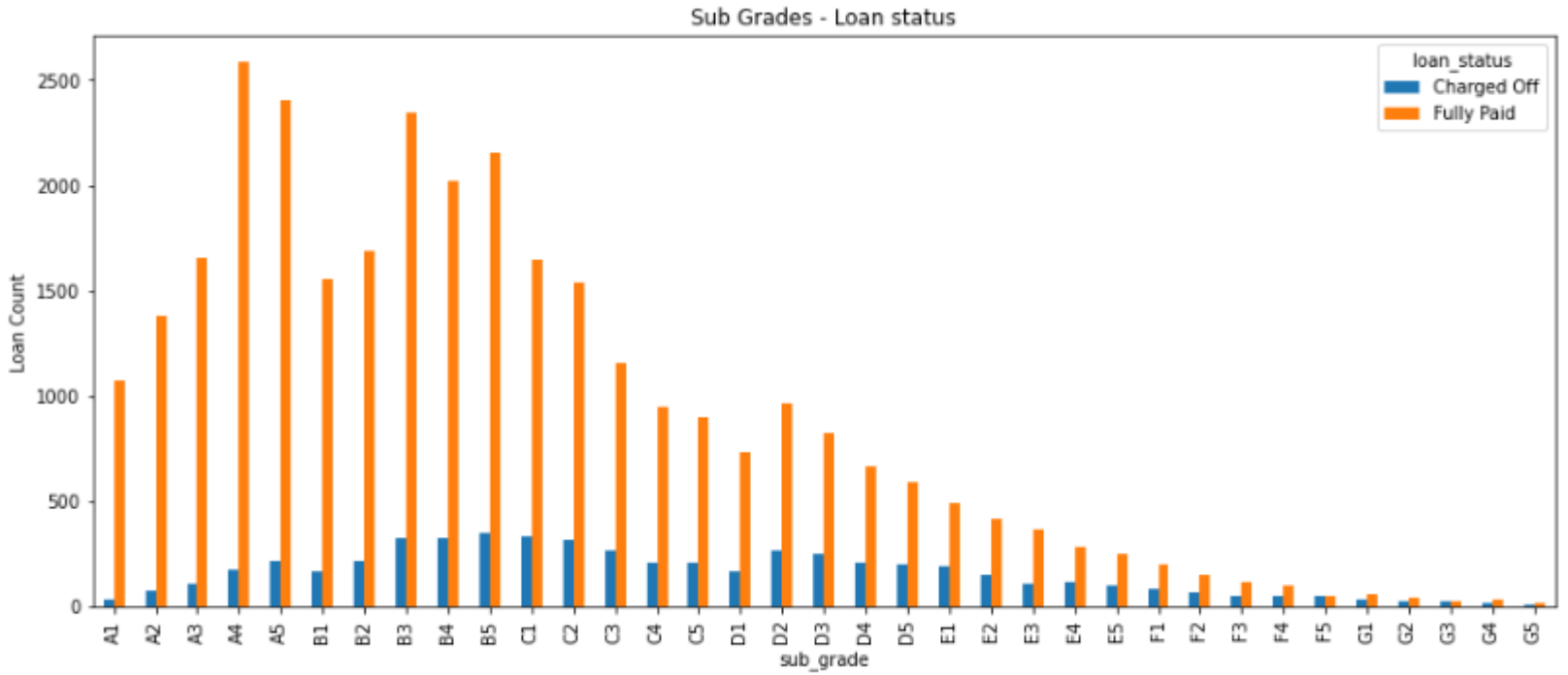


Observation 6 : From the above scatter plot we observe that density of the loan takers are in the range of 500-20000



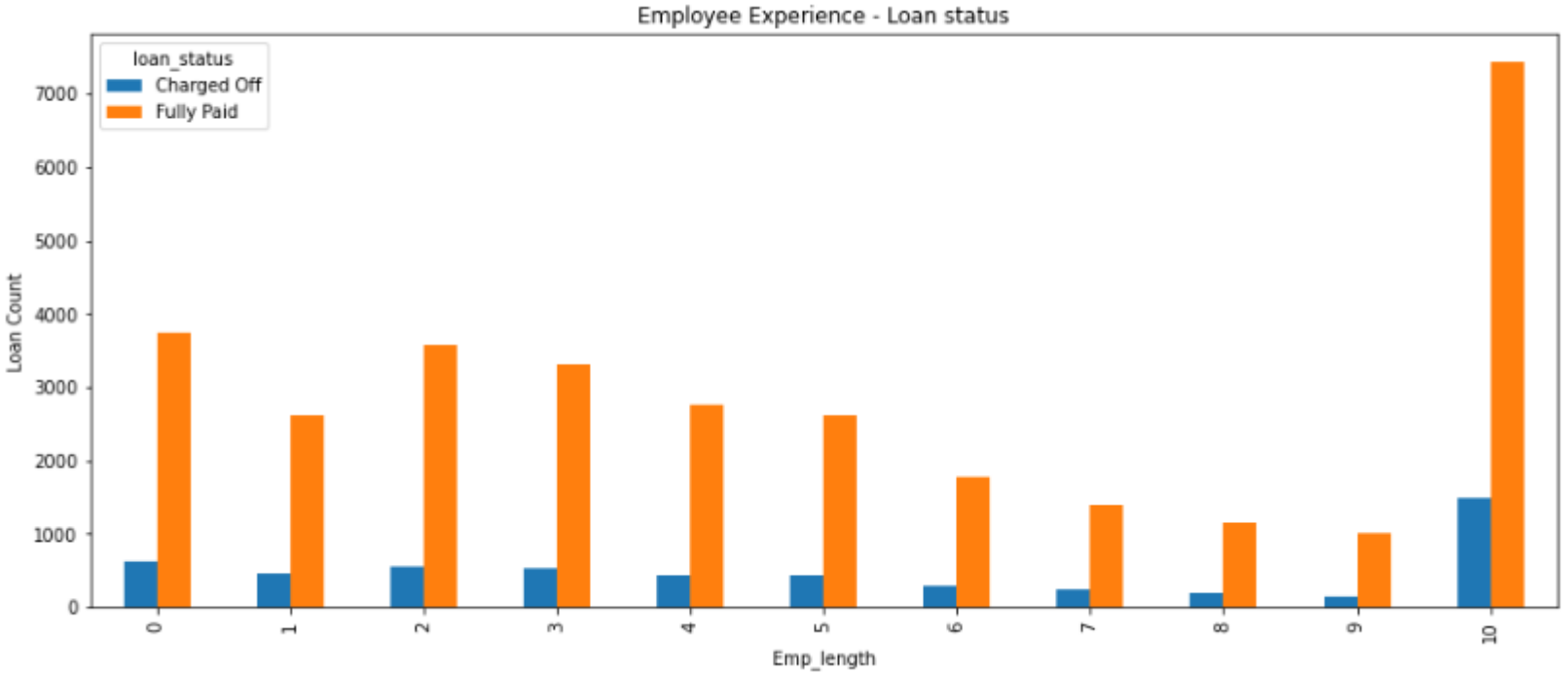
Observation 7 : Even from the bivariate analysis Most of the Charged-Off Grade are B,C, D. And we also observe that E, F, G have almost half of the defaulters when compared with Fully paid

Bivariate analysis & Observations



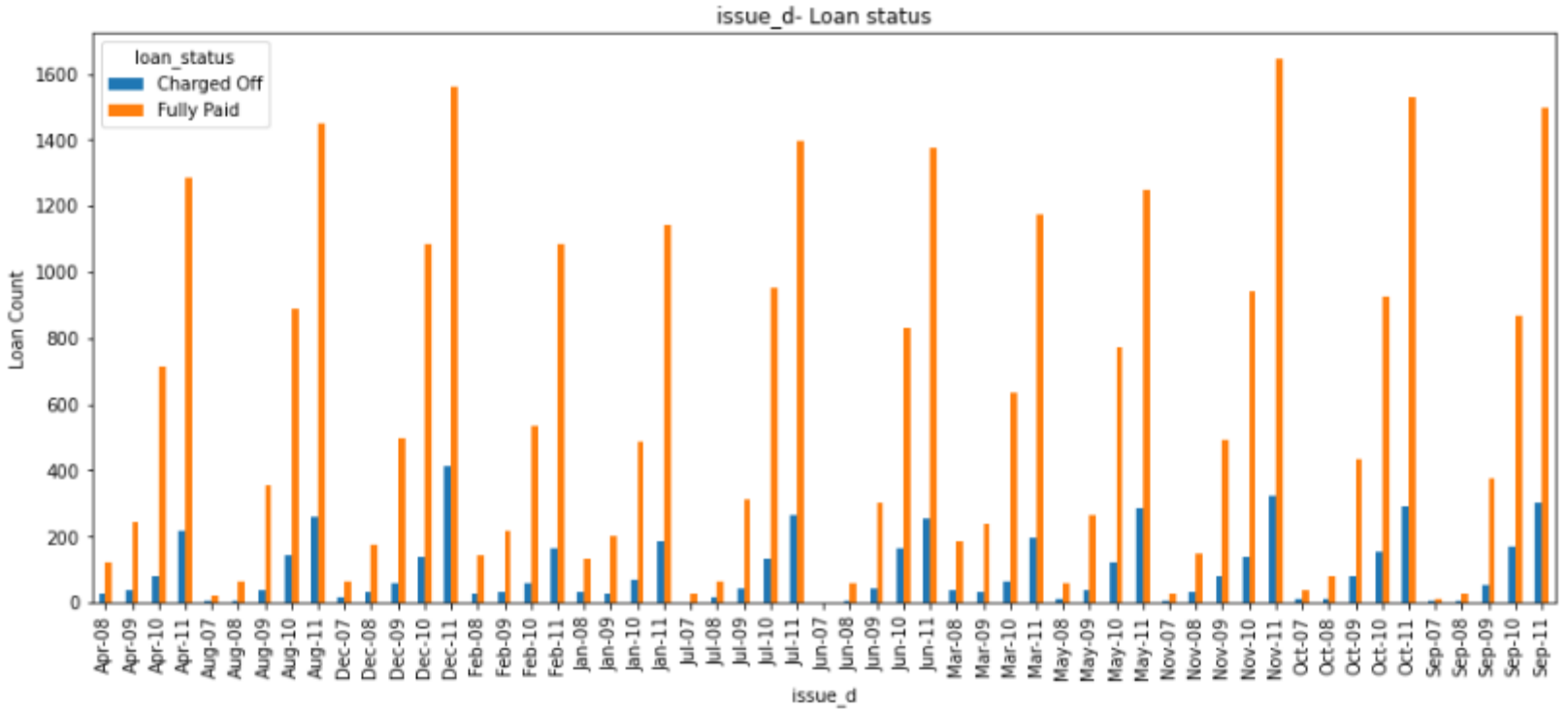
Observation 8 : Most of the Charged-Off Sub Grades are B,C, D. And we also observe that sub groups of E, F, G have almost half od the defaulters when compared with Fully paid

Bivariate analysis & Observations



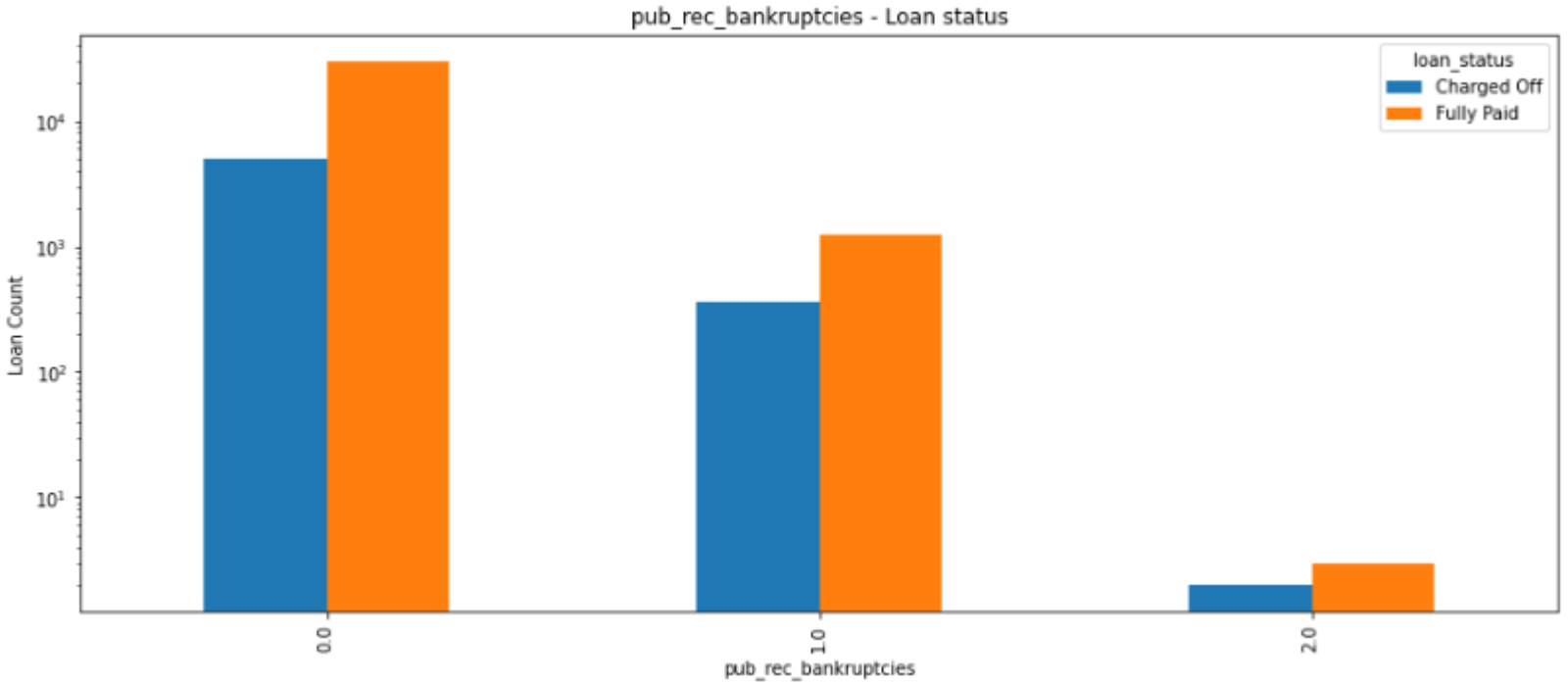
Observation 10 : From the above graph, we can say that If the emp_length is <1 year, 2 years, 3 years, +10 years then there is high change of falling into charged off

Bivariate analysis & Observations



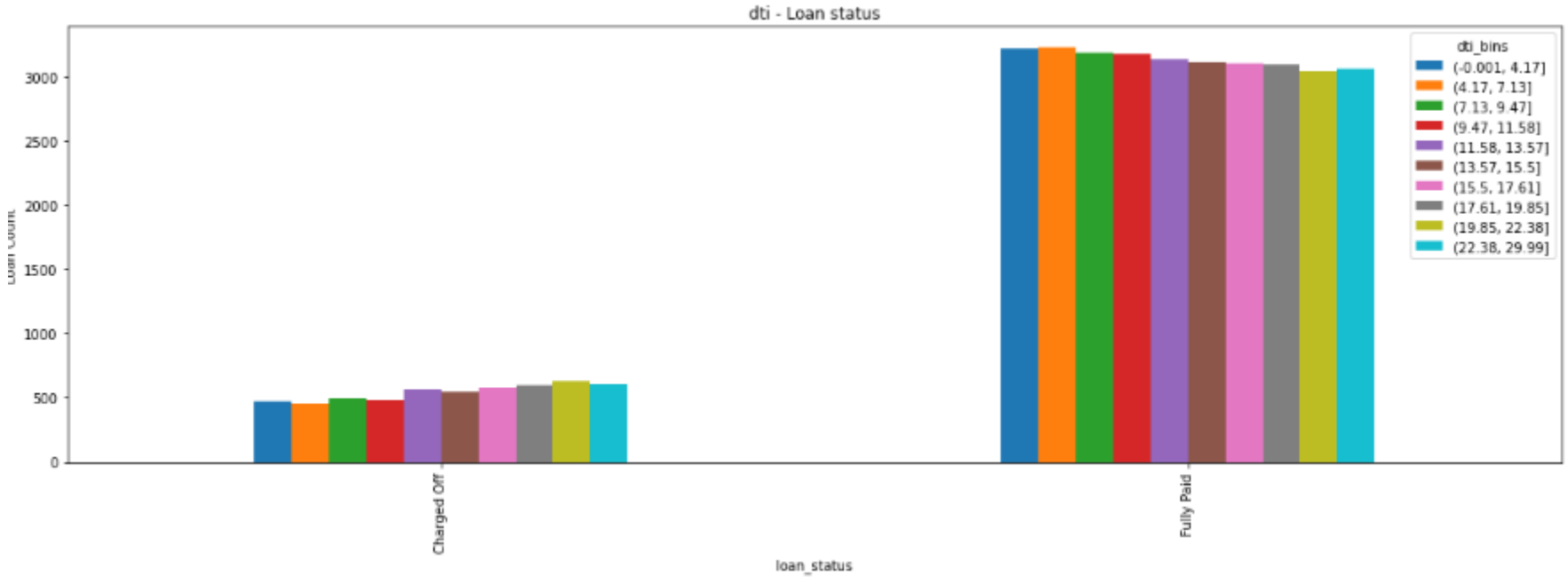
Observation 11 : From the above graph, we can clearly say that the defaulter rate is increasing with the successive months and years

Bivariate analysis & Observations



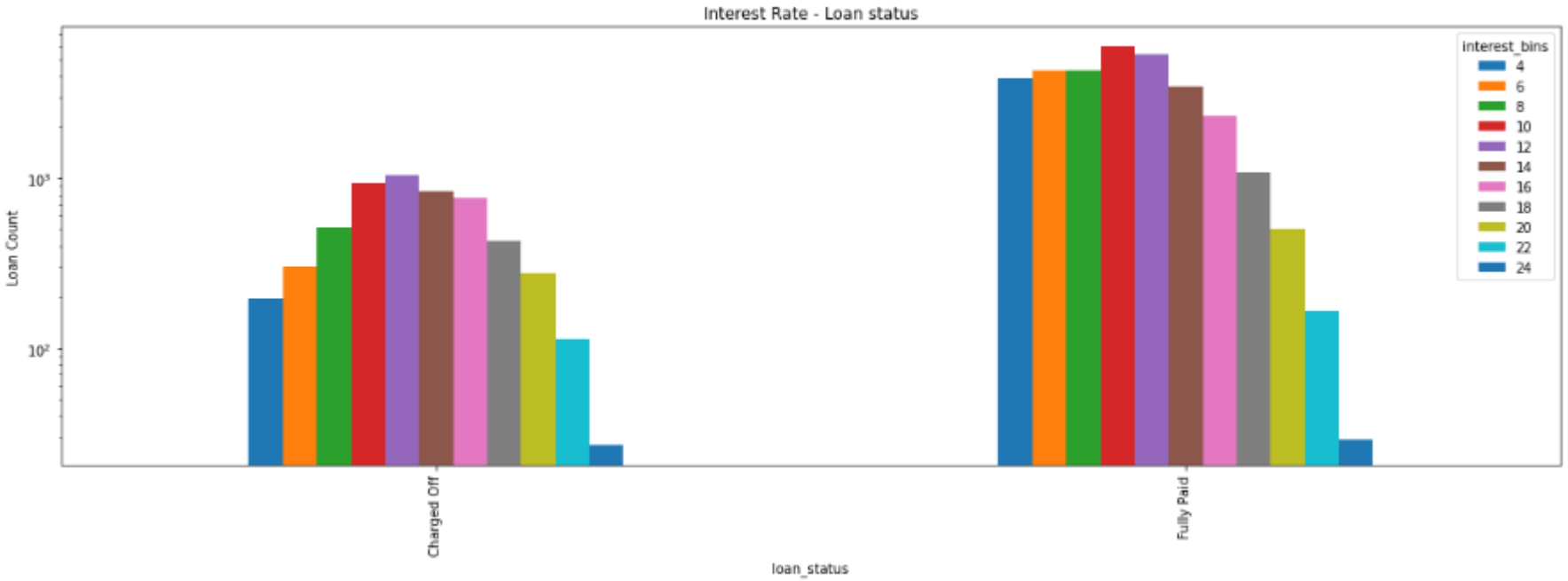
Observation 12 : From the above log-log scale for pub_rec_bankruptcies we see that;
If pub_rec_bankruptcies is 2 then almost half of the customers are tending to fall under Charged off categories

Bivariate analysis & Observations



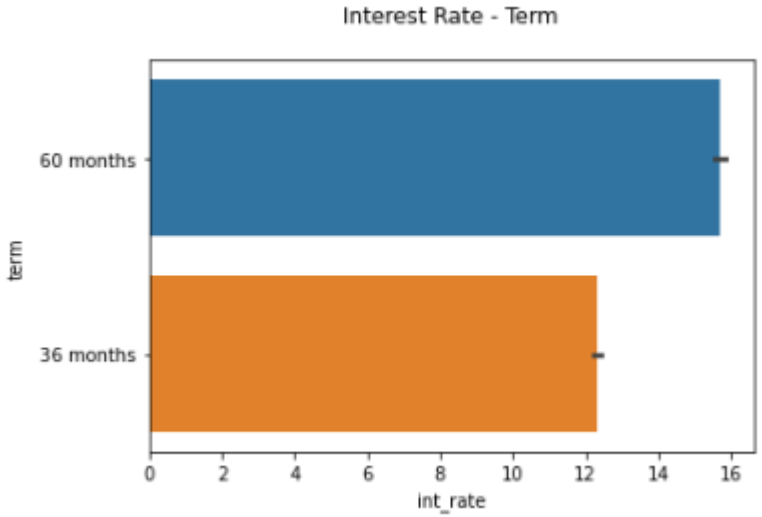
Observation 13 : From the above graph there is a slight increase of the defaulter when they have more dti. People having more dti tend to fall under charged off category

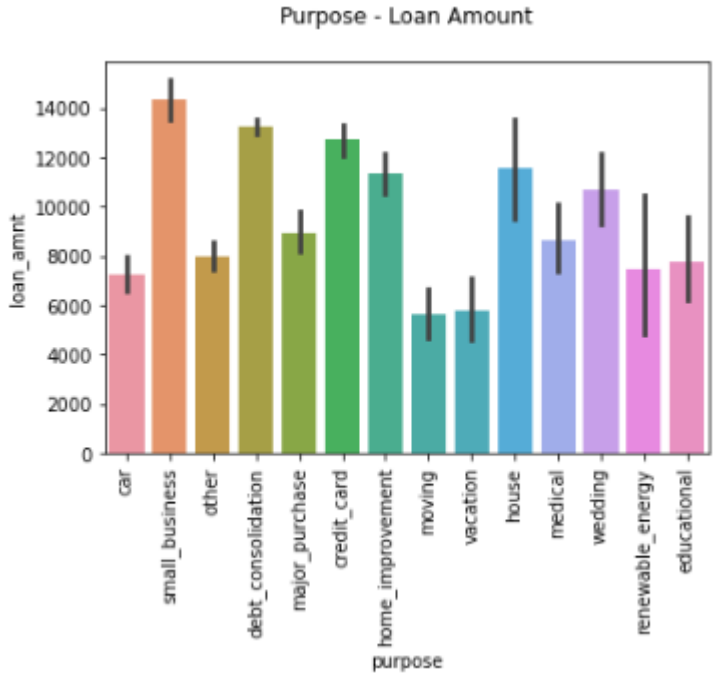
Bivariate analysis & Observations



Observation 14 : If the interest rate is increasing then the 'charged of's are more than the Fully paid individuals.

Bivariate analysis & Observations

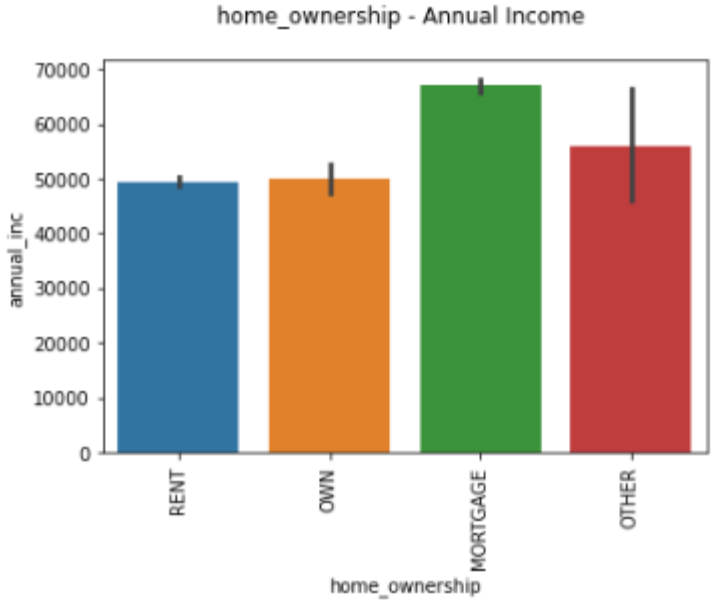




Observation 15 : In the above graph, we have considered data set which have only 'Charged-Off' records.
From that we observed that Term month with more interest rate is having more tendency to fall under defaulters

Observation 16: From the above graph, it is observed that if the loan amount is more than 10000 and if the purpose belongs to Small bussiness, debt_consolidation, credit card, home_improvemnt, house or wedding having chances to fall under defaulters record

Bivariate analysis & Observations



Observation 17 : From the above graph we observerd that if the annual income is more than 50000 then Home_ownership with Mortgage is having more chances to fall under default list

Recommendations

1. Good to avoid the applicant whose annual income is more with high mortgage value
2. Avoid B,C &D Grade applicants
3. Avoid A5 ,B3,B4,B5 & D2 Subgrade applicants
4. Avoid the applicants from CA state
5. Avoid purpose of the loan is either Small business, debt consolidation

Thank you