Lending club case study

H Sankritya Vedanabhatla

Sanyam Kumar

Problem Statement & Objective

Our business is a 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

We are concerned with the second type of loan. The data given in **loan.csv** contains information about past loan applicants and whether they 'defaulted' or not. The objective is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Solution approach (Python as the coding language)

1. We have checked the data and the data dictionary for meanings of each field present. This enabled us to drop many fields which are not relevant to the analysis in round 1 of arriving at the final data set for analysis.
2. In the second round, we cleaned the dataset by loading it into a dataframe. Cleaning involved various methods to impute/drop null values in the data, addressing non standard values (like % suffix), checking for duplicate records, addressing data types etc. It also involved eliminating certain records based on the business problem at hand, which do not help with the analysis in any way.
3. Subsequently, outlier analysis for continuous numeric variables was performed with a two-fold objective. Firstly, to check of any additional records (which might impact analysis due to high/low values) have to be eliminated . Secondly, it acts as a kind of univariate analysis for numeric variables. Based on this some records were eliminated.
4. In the next step, we renamed all column names to remove abbreviations and better readability. We also standardized the precision across numeric variables and arrived at the final dataframe for analysis.
5. We went ahead to plot frequency plots for all categorical variables to get some idea about the data. This is the Univariate analysis part for Categorical variables. We just got data-related observations from this part. No driver related conclusive findings
6. Next step Bivariate analysis using Box plots, count plots and stacked bar charts with Loan_status as a segment from which we draw good number of observations and identified the drivers of loan defaults. In this presentation, we will mostly be focusing the results from this analysis as all significant observations that we find as relevant drivers of loan default were from here only.

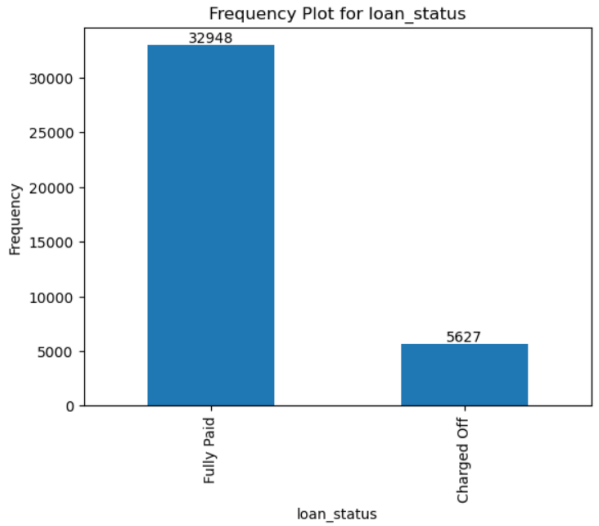
Metadata and summary analysis

- Below analysis is of the final data frame that we had arrived at after cleaning in Solution Approach Step 4

Total No of Records	38575
Total No of Columns	23

Some fields have been classified as categorical despite being numeric data type. This was because the value counts for such variables were less and it becomes easier to analyze

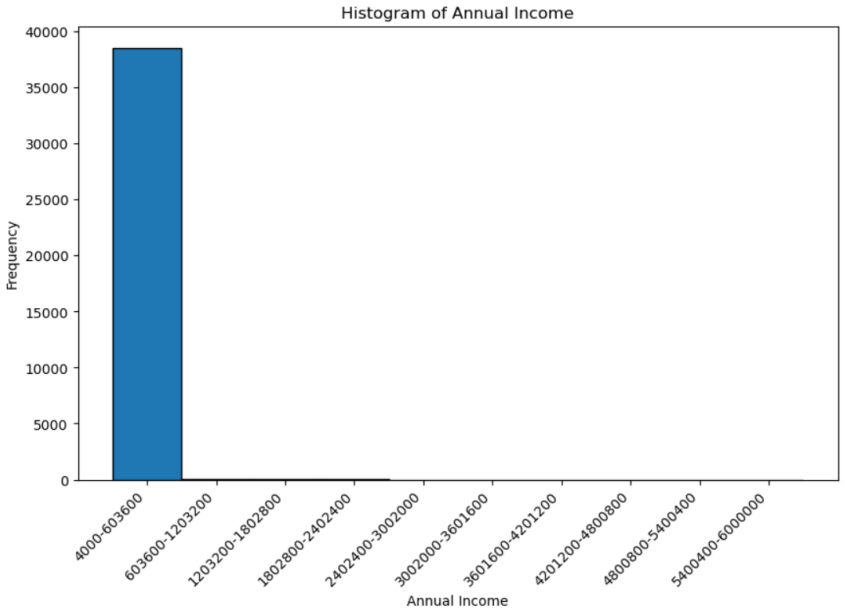
*Loan_status is the dependent variable. A simple frequency plot shows that 32948 loans were paid off while 5627 loans were charged off. See the Plot on the right for more details

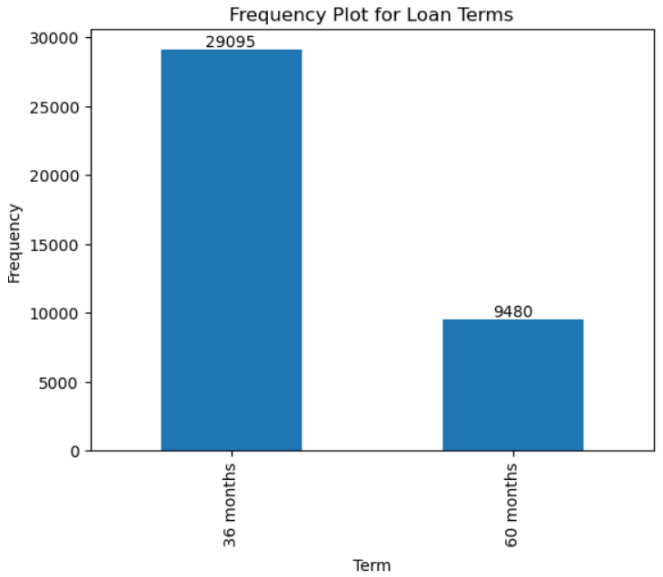


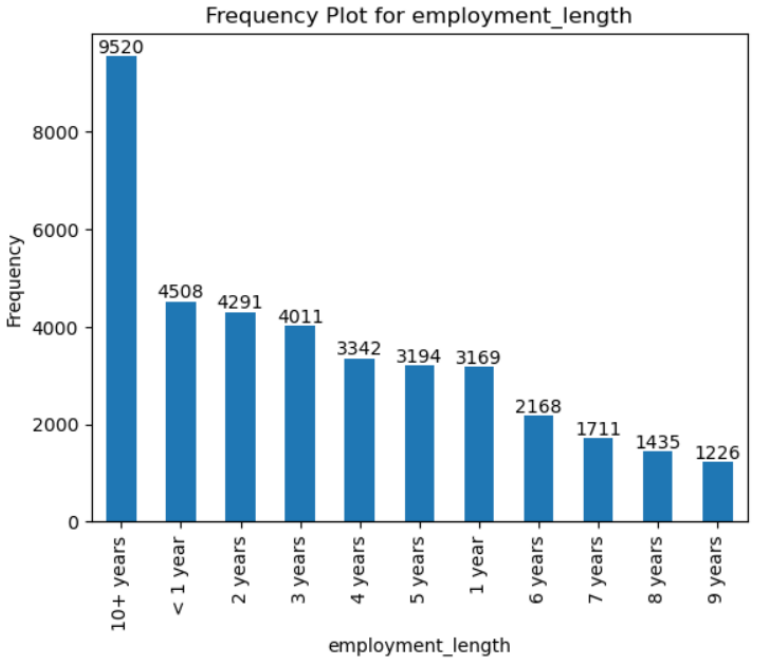
Field	Category	Variable Classification
member_id	Unique identifier	Categorical
loan_amount	Loan Attribute	Continuous
term	Loan Attribute	Categorical
interest_rate	Loan Attribute	Continuous
grade	Loan Attribute	Categorical
sub_grade	Loan Attribute	Categorical
employment_length	Customer Attribute	Categorical
home_ownership	Customer Attribute	Categorical
annual_income	Customer Attribute	Continuous
verification_status	Customer Attribute	Categorical
loan_status *	Loan Attribute	Categorical (Dependent variable)
purpose	Loan Attribute	Categorical
address_state	Customer Attribute	Categorical
debt_to_income_ratio	Customer Attribute	Continuous
delinquencies_2yrs	Customer Attribute	Categorical
inquiries_last_6_months	Customer Attribute	Categorical
open_accounts	Customer Attribute	Categorical
revolving_balance	Loan Attribute	Continuous
revolving_utilization	Customer Attribute	Continuous
total_accounts	Customer Attribute	Continuous
outstanding_principal	Field not used for analysis	NA
outstanding_principal_investor	Field not used for analysis	NA
public_record_bankruptcies	Customer Attribute	Categorical

Univariate Analysis for some variables where data skewness was observed

(not including plots for the variables for which data is evenly distributed)







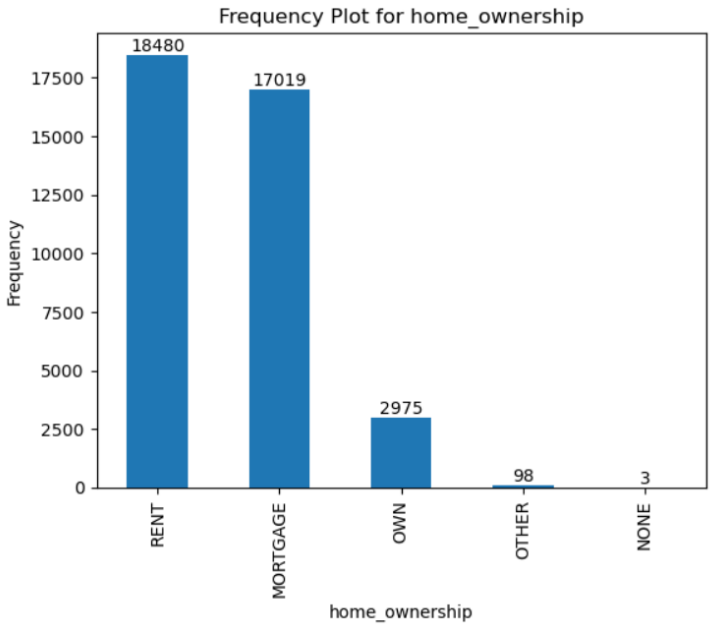
Most of the borrowers have annual incomes of < $ 603600

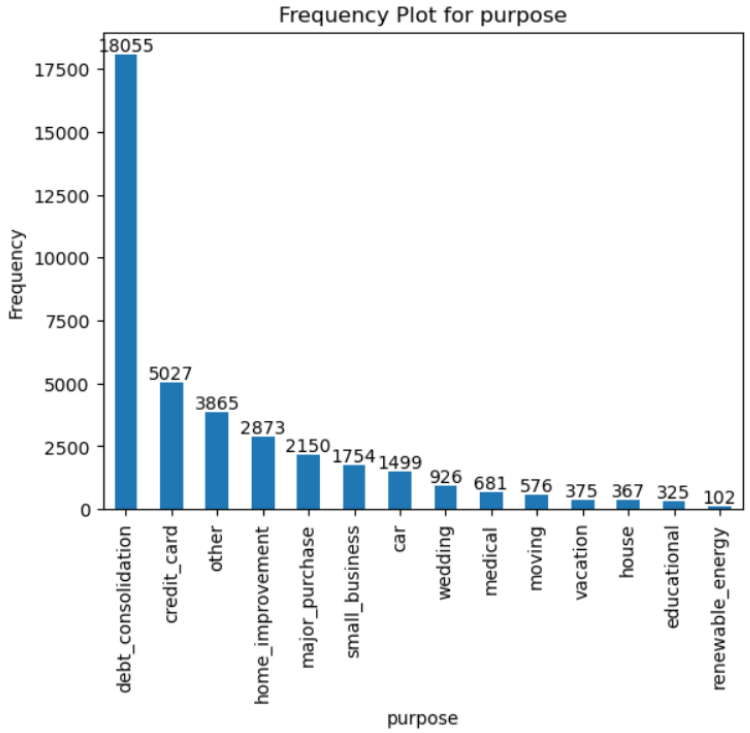
A majority of borrowers have loan term of 36 Months

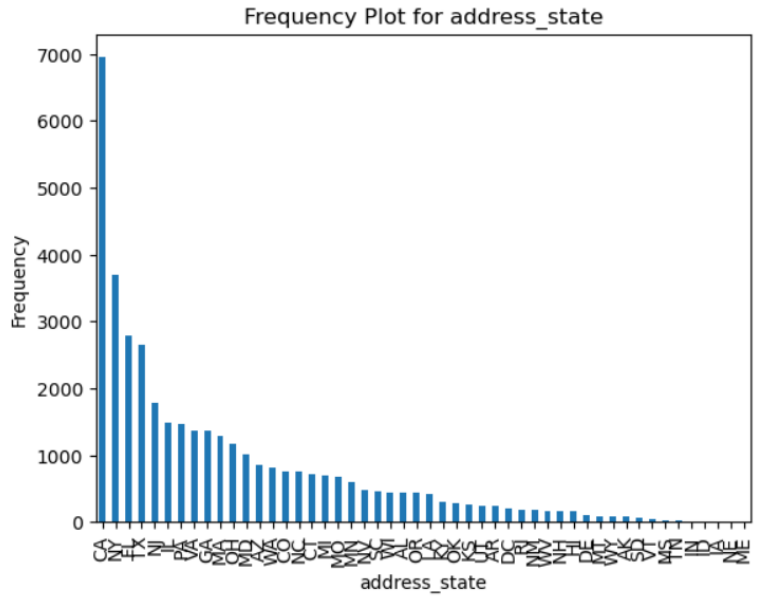
More borrowers have 10+ years of employment than any other category of employment length

Univariate Analysis for some variables where data skewness was observed

(not including plots for the variables for which data is evenly distributed)





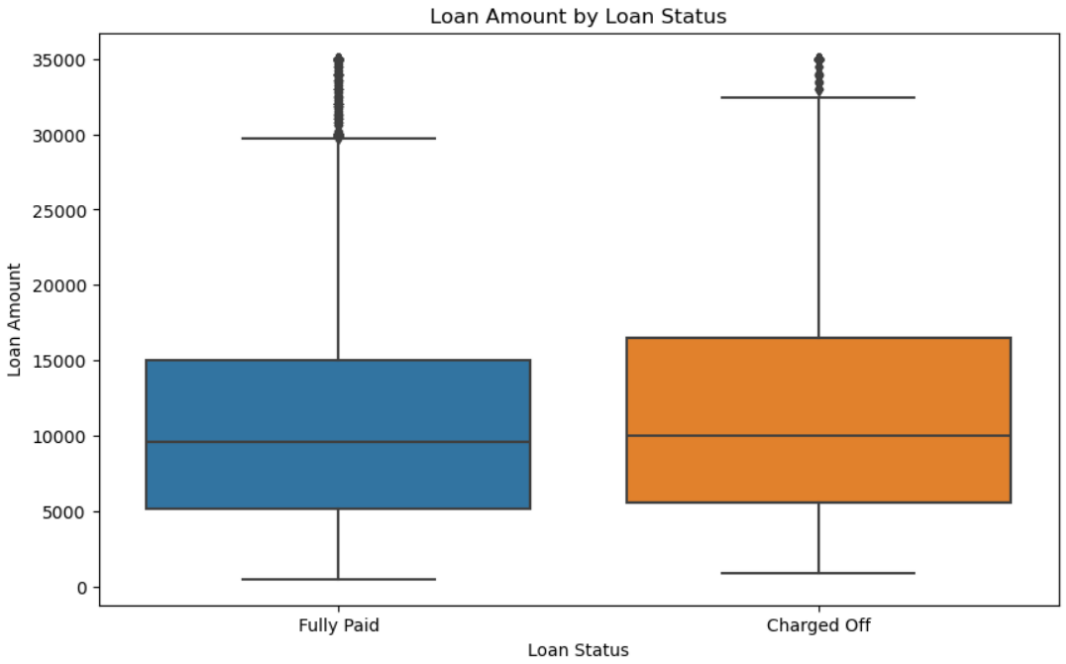


Most borrowers either live on rent or are on mortgage and a minority have their own house

Most borrowers borrowed money for consolidating their debt

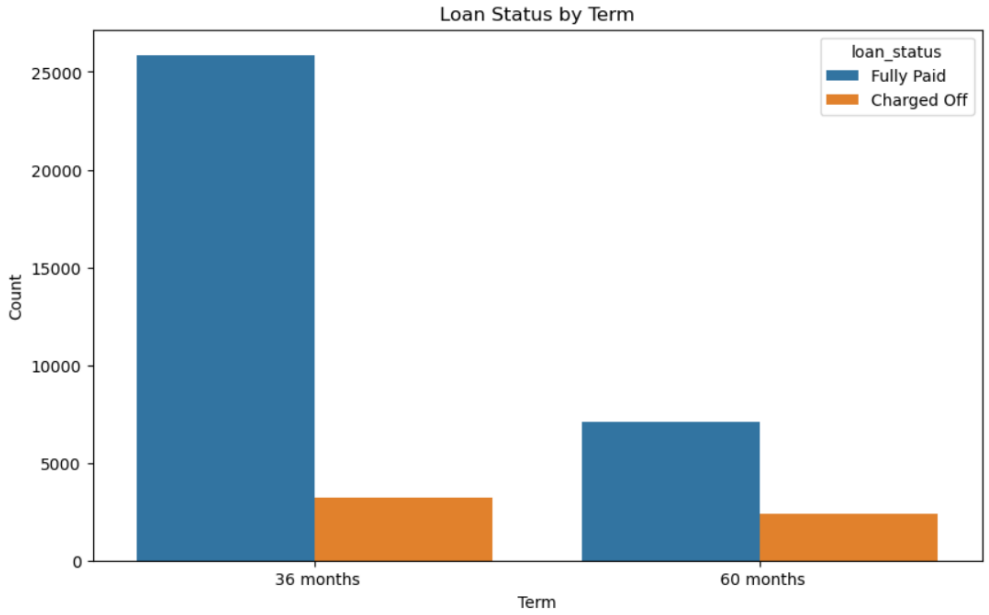
Most borrowers reside in California state (CA)

Bivariate analysis & Observations

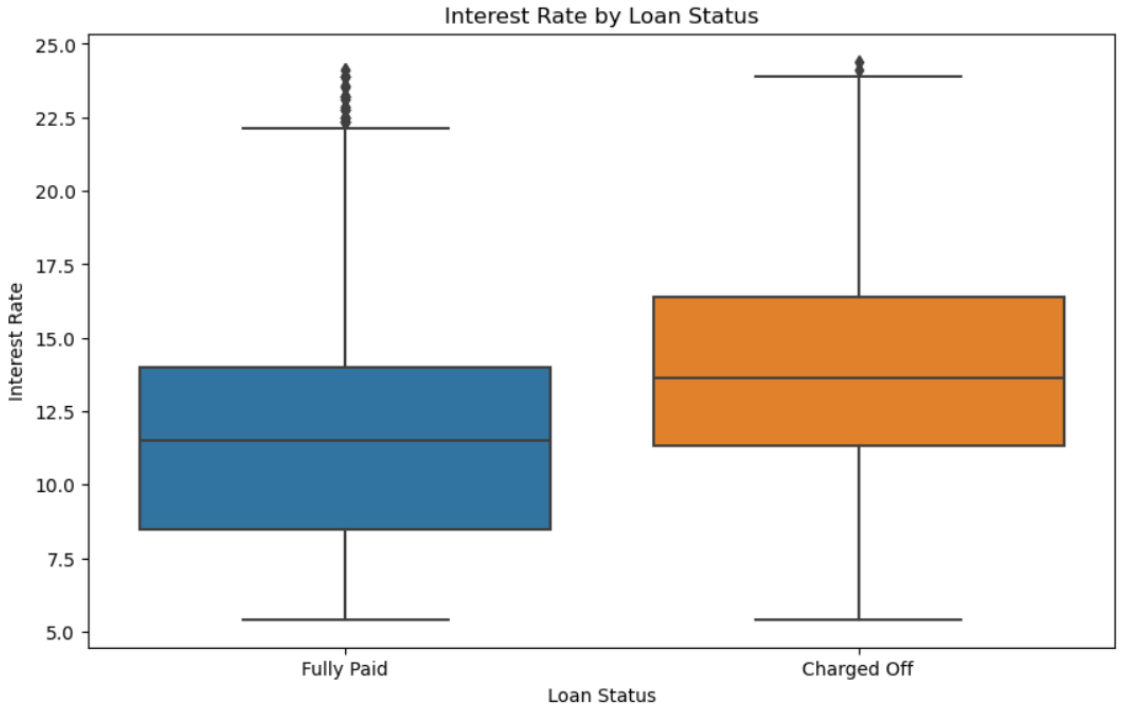


Observation 1 - Charged off loans are associated with slightly higher loan amounts than loan amounts associated with fully paid loans. Indicating that a higher loan amount is a risk factor

Observation 2 - Charged-off loans contribute a higher proportion of overall loans in 60 months term compared to 36 months term. Shorter loan terms have lesser chances of default

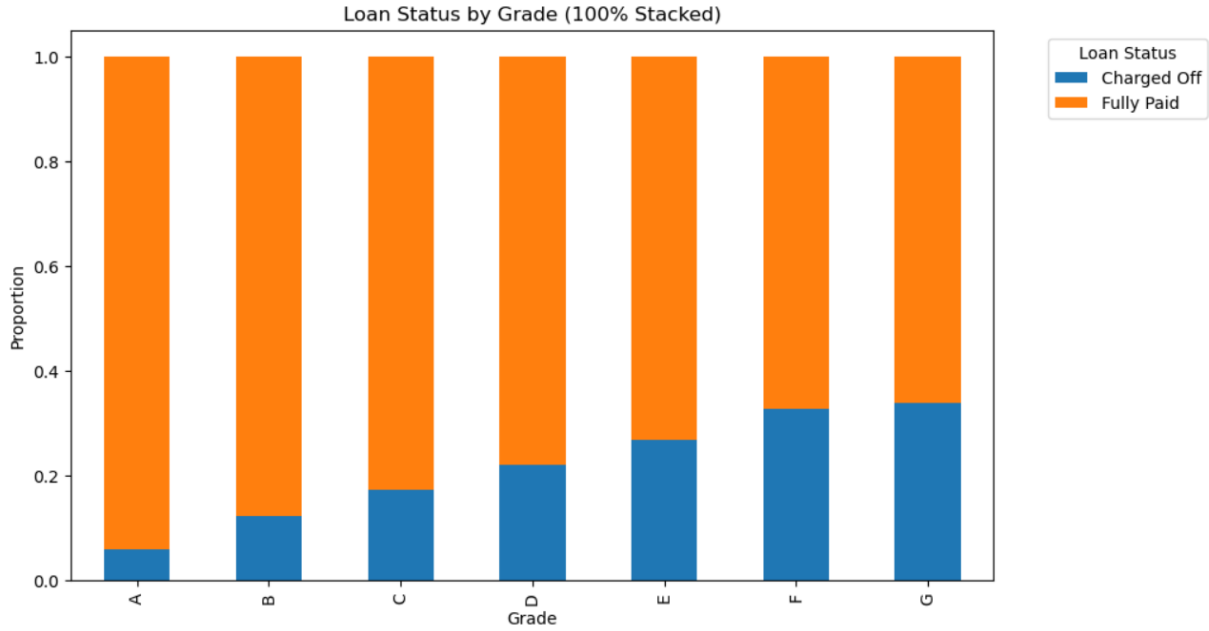


Bivariate analysis & Observations

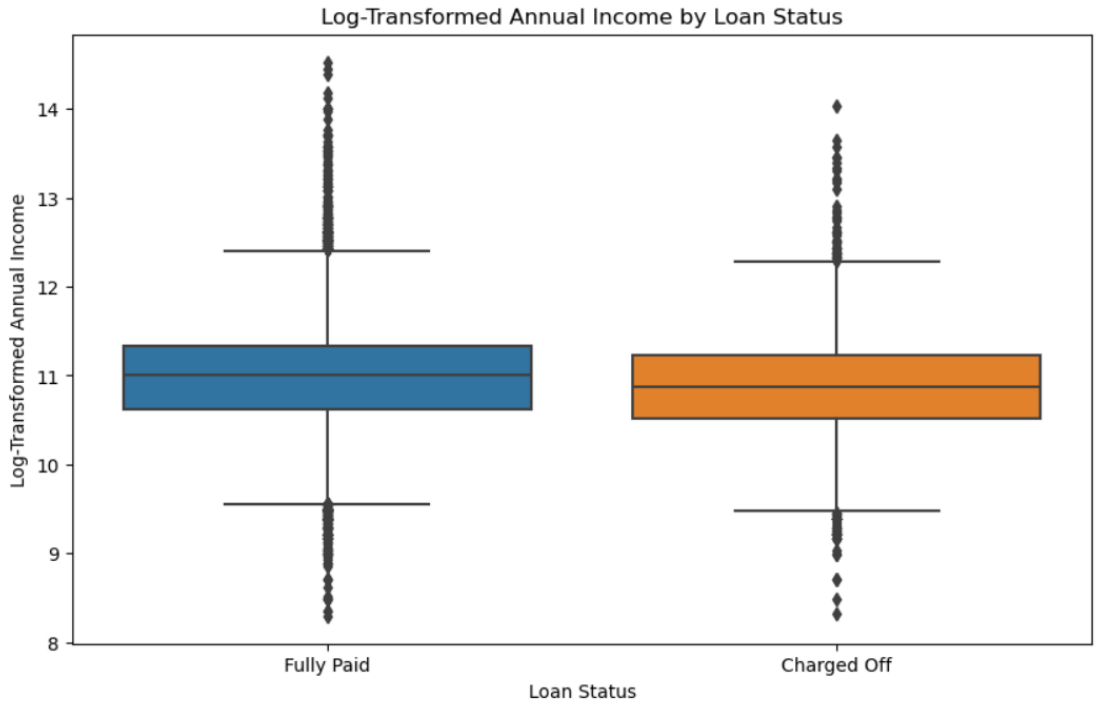


Observation 3 - Charged off loans is associated with significantly higher interest rate than fully paid ones, indicating that high interest rates are more at risk of default

Observation 4 - Lower grade loans have more proportion of defaults (Assumption as A being the highest grade and G being the lowest grade)

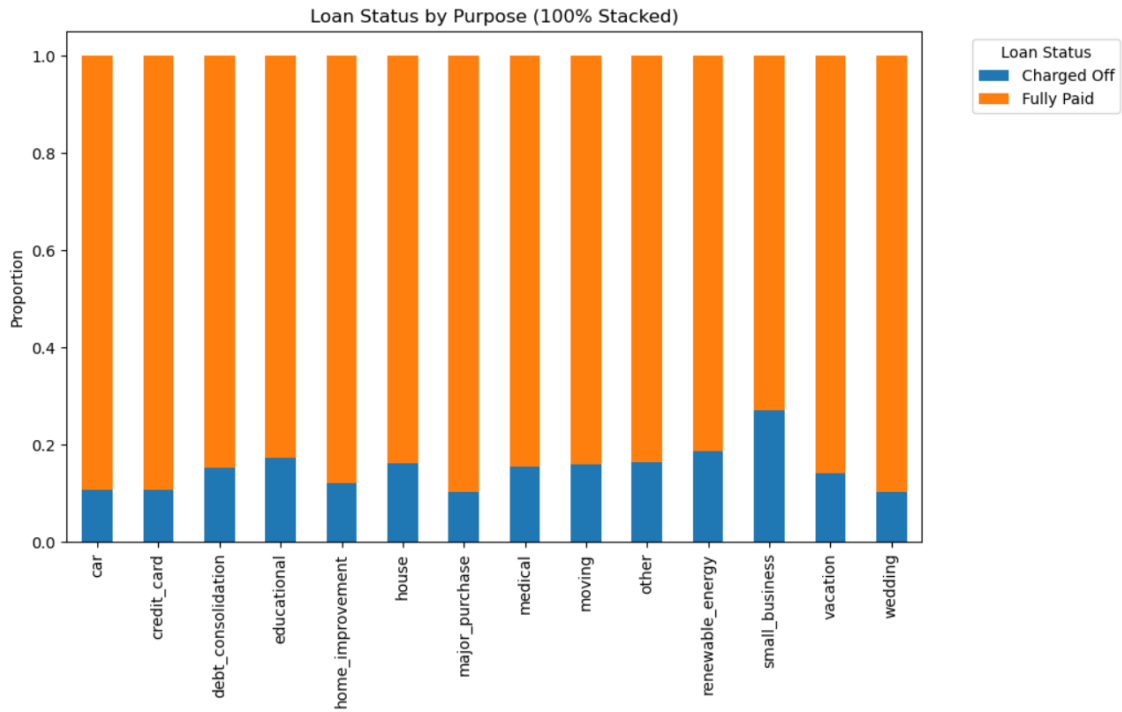


Bivariate analysis & Observations

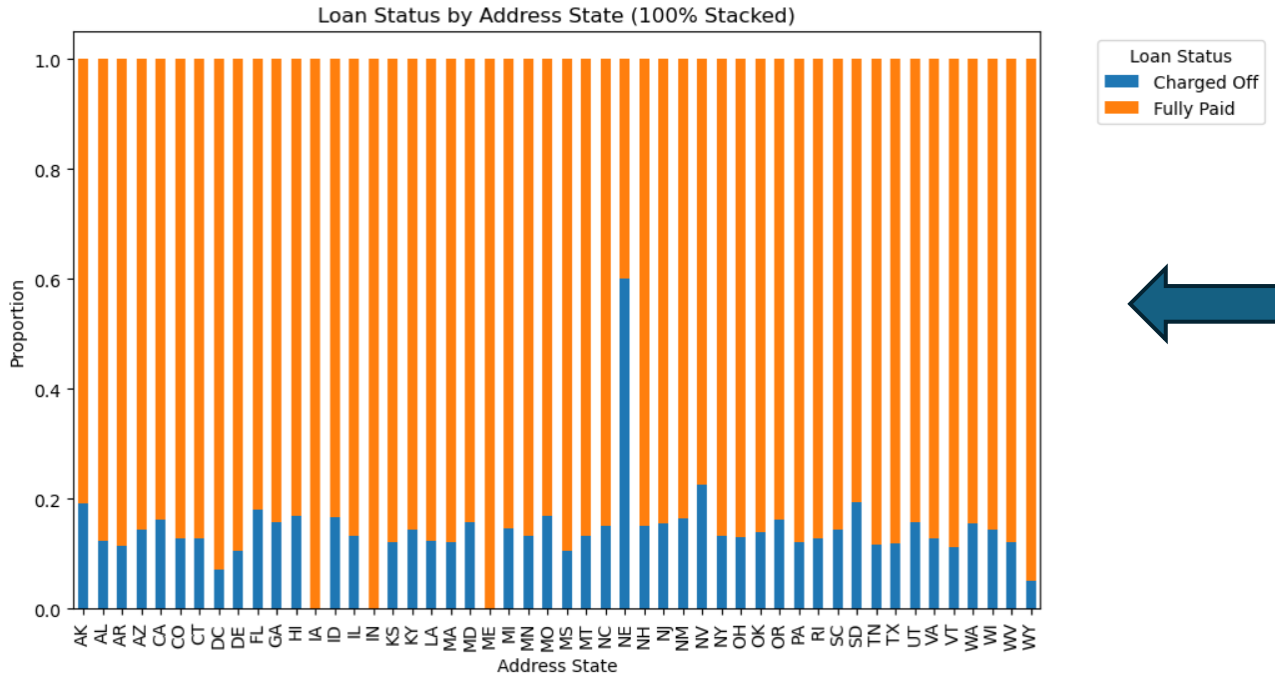


Observation 5 - Borrowers who defaulted on loans have on an average slightly lower annual incomes than fully paid loans

Observation 6 - Borrowers who have taken loans for small business have slightly higher chance of default, though not significant enough

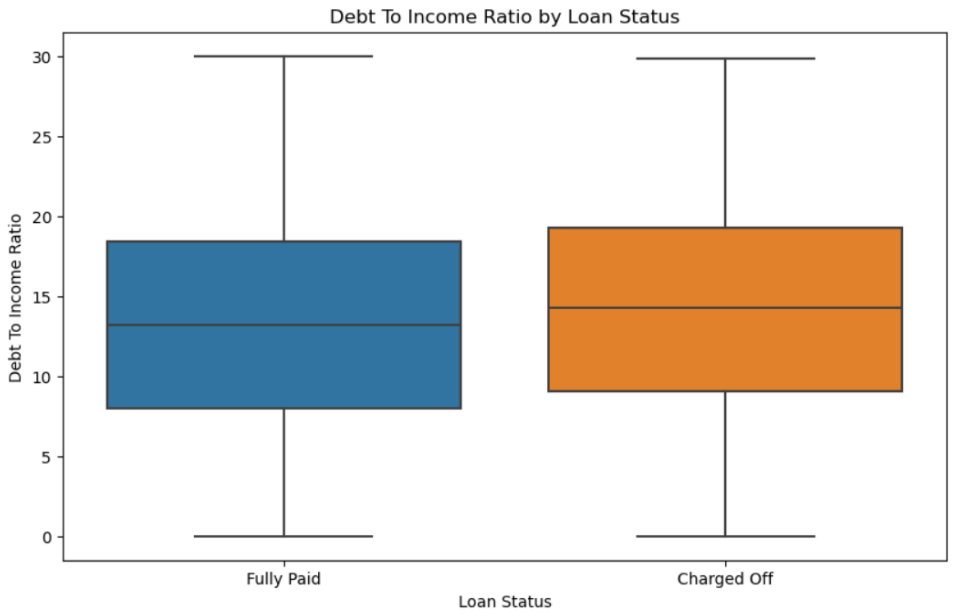


Bivariate analysis & Observations

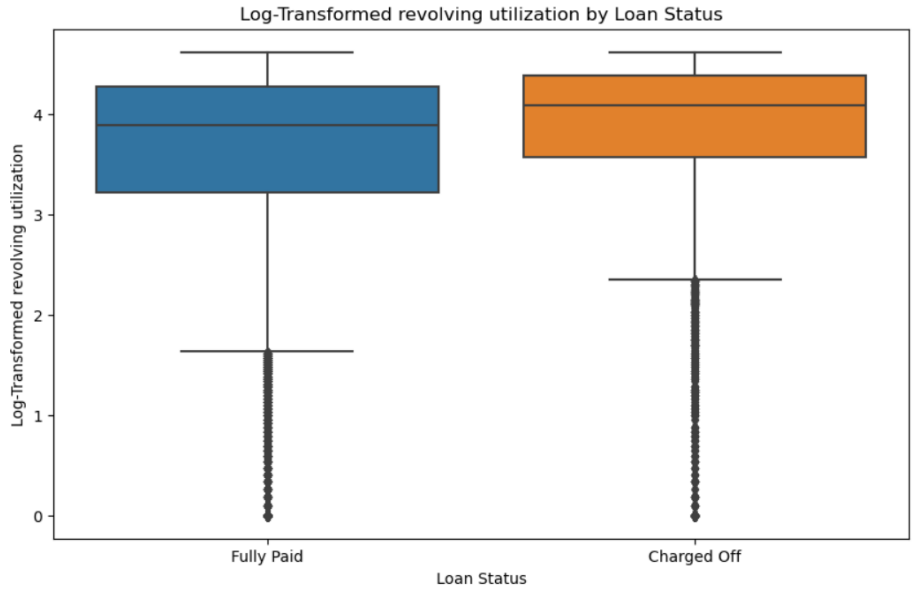


Observation 7 - Loans given to NE (Nevada) state residents have a significantly higher chance of default

Observation 8 - Members having higher debt-to income ratio have higher chance of defaulting as observed from the box plots mean and 75th percentiles

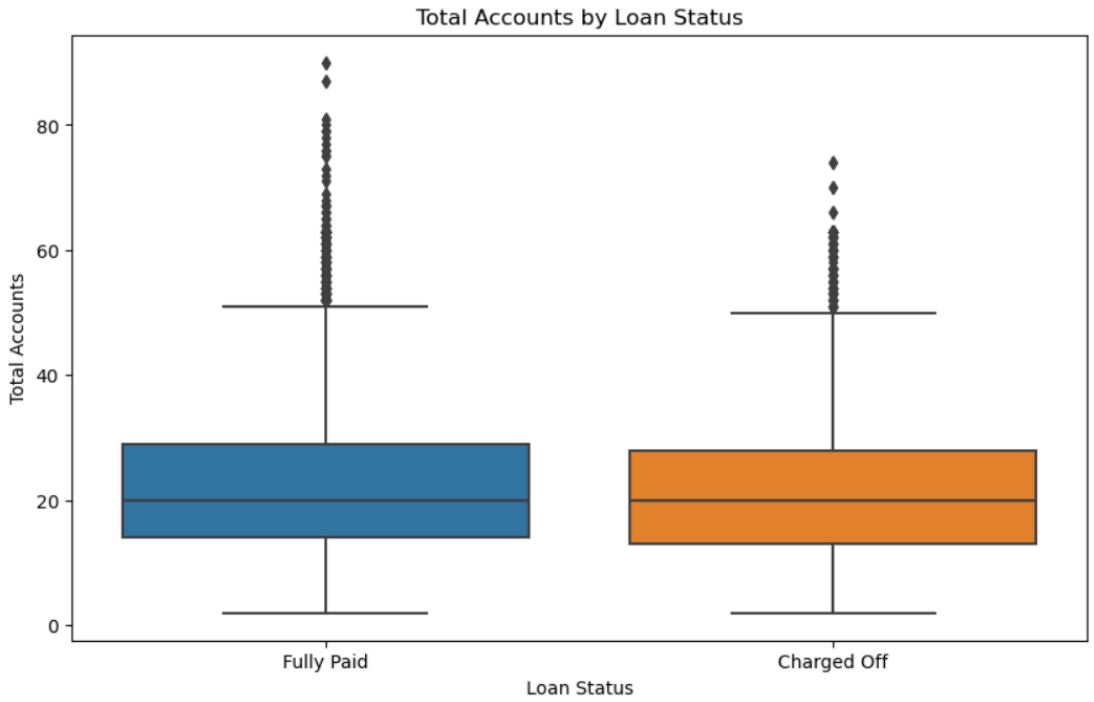


Bivariate analysis & Observations

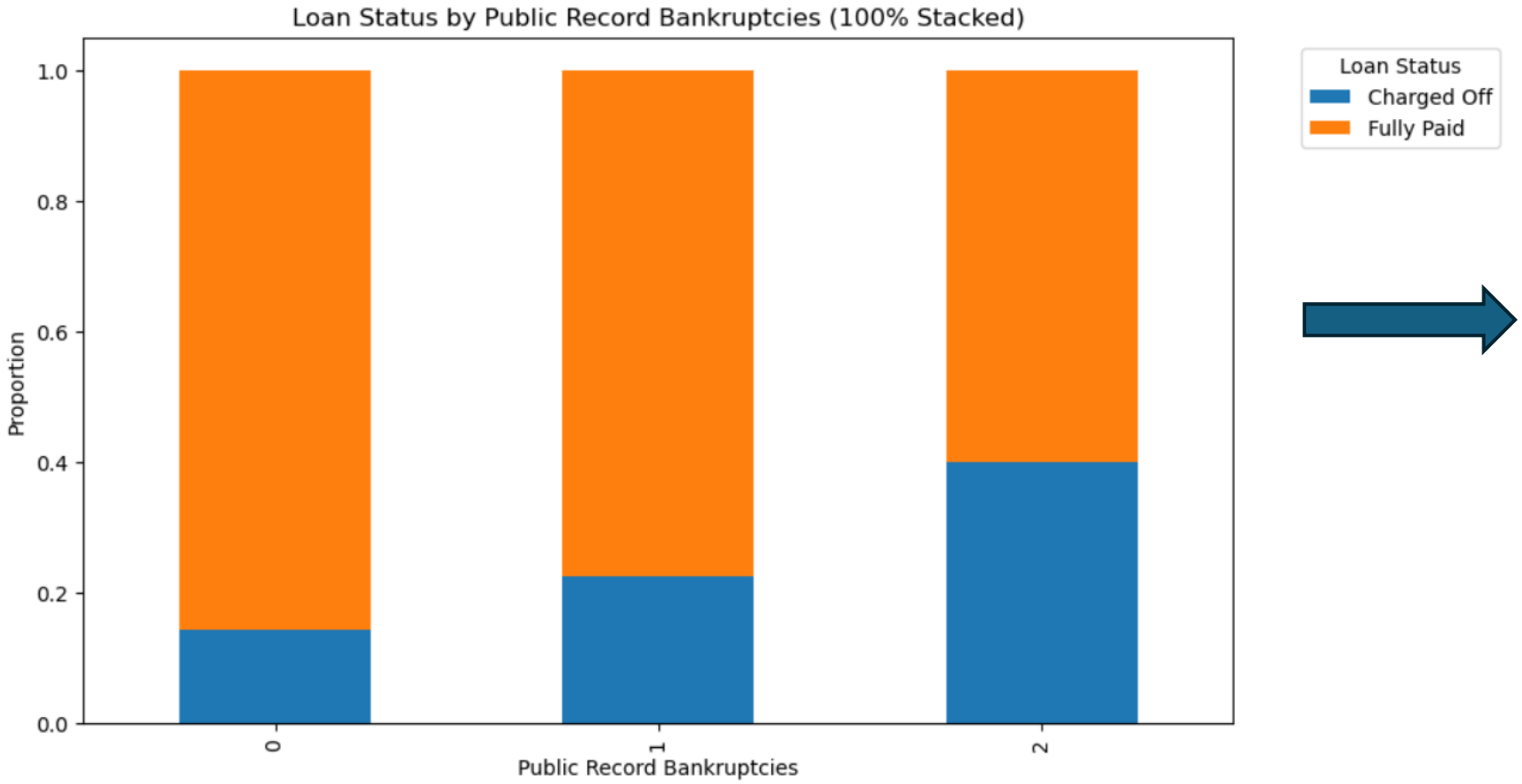


Observation 9 - Higher revolving utilization correlates with a higher number of loan defaults

Observation 10 - lower the total number of credit lines currently in the borrower's credit file, more chances of charged off loans. This finding is counter-intuitive to business understanding that more debt-ridden borrower tend to be more at a risk of default



Bivariate analysis & Observations



Observation 11 - the more a member has a public record of bankruptcies, more probability of the loan default

List of Drivers of Loan defaults

(summarizing all observations)

1. Higher Loan amount
2. Longer Term of loan
3. Higher Interest Rate
4. Lower Grade
5. Lower Annual Income
6. Loan taken for the Purpose of small business
7. Borrowers from a specific State
8. Higher Debt to Income ratio
9. Higher Revolving Utilization
10. Less number of Credit Lines
11. More Public records of Bankruptcies

Key assumptions leading to Business recommendations

There may be relationships between the independent variables. That is one variable might influence other variable which is not taken into consideration. All variables (other than loan_status) are being treated as being independent of each other. Example – Grading of loan may be dependent on Annual income and Debt to Income ratio but we are ignoring that inter-dependency

Business recommendation

If the company wants to reduce the risk by limiting the risky borrowers, it must thoroughly scrutinize all applications received from Nevada and must generally avoid lending to people already burdened with loan. That is, people with high revolving credit utilization, those having high debt to income ratio and people with a history of bankruptcies (People having more public record of bankruptcies).

Another way to look at it would be is that in general, it is risky to lend to borrowers with lower annual income and with high interest rates having lower grading. Preference should be given for shorter term loans as they have a good history of being paid off fully.

APPENDIX

Correlations of each independent variable with loan status

Variable	Correlation with Loan status (Fully paid = 1, Charged off = 0)
loan_amount	-0.059478
term	-0.173508
interest_rate	-0.21139
grade	-0.20187
sub_grade	-0.205317
employment_length	-0.023721
home_ownership	0.022204
annual_income	0.047355
verification_status	-0.048254
purpose	-0.010402
address_state	0.010881
debt_to_income_ratio	-0.04504
delinquencies_2yrs	-0.02009
inquiries_last_6_months	-0.071904
open_accounts	0.009152
revolving_balance	-0.005839
revolving_utilization	-0.099845
total_accounts	0.02261
public_record_bankruptcies	-0.046985
log_annual_income	0.069026
log_revolving_balance	-0.006846
log_revolving_utilization	-0.074589
public_record_bankruptcies	-0.046985
log_annual_income	0.069026
log_revolving_balance	-0.006846
log_revolving_utilization	-0.074589