LENDING CLUB CASE STUDY FOR LOAN APPROVAL RISK MANAGEMENT

BY: MAALOLAN K & RAJA SHEKHAR

Introduction:

We have to develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Overview:

We work for 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.

Problem Statement

Business Risk: Losing business from rejected repayers.
Financial Risk: Incurring losses from approved defaulters

AIM:

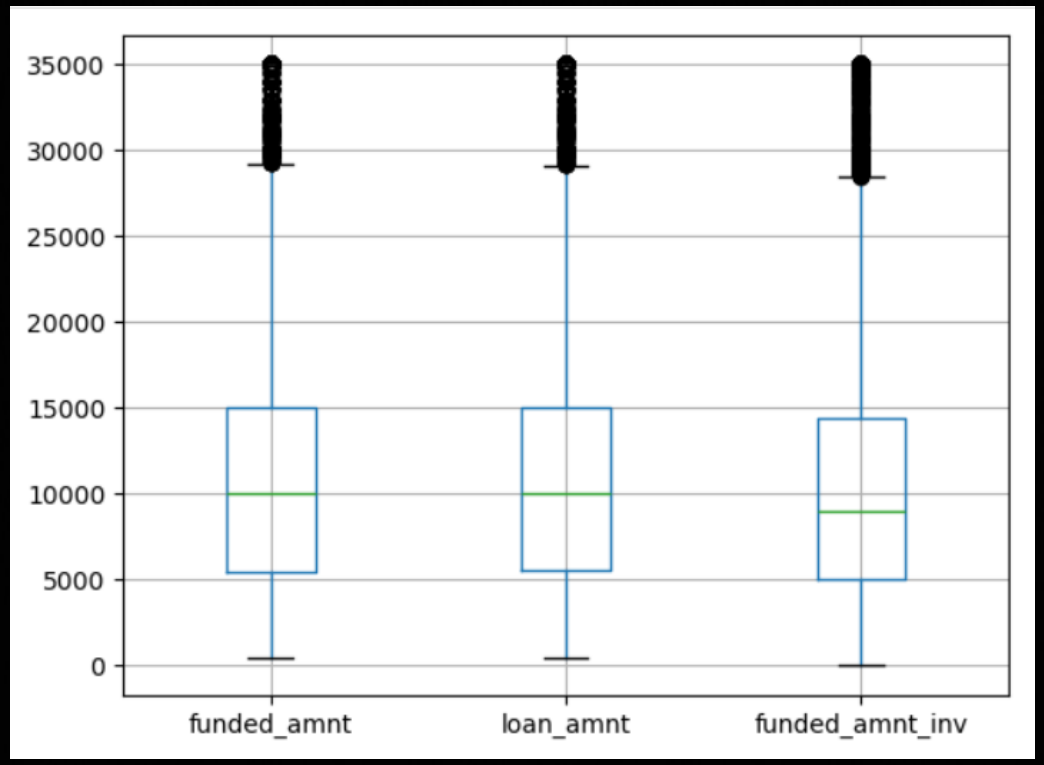
- To identify these risky loan applicants using EDA.
- Understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Borrowers who **default** cause the largest amount of loss to the lenders.

Data Understanding:

Data Sources: Applicant information, financial details, credit history, and historical loan data.
Key Features: Income, employment status, employment length existing debts, loan amount, and purpose.

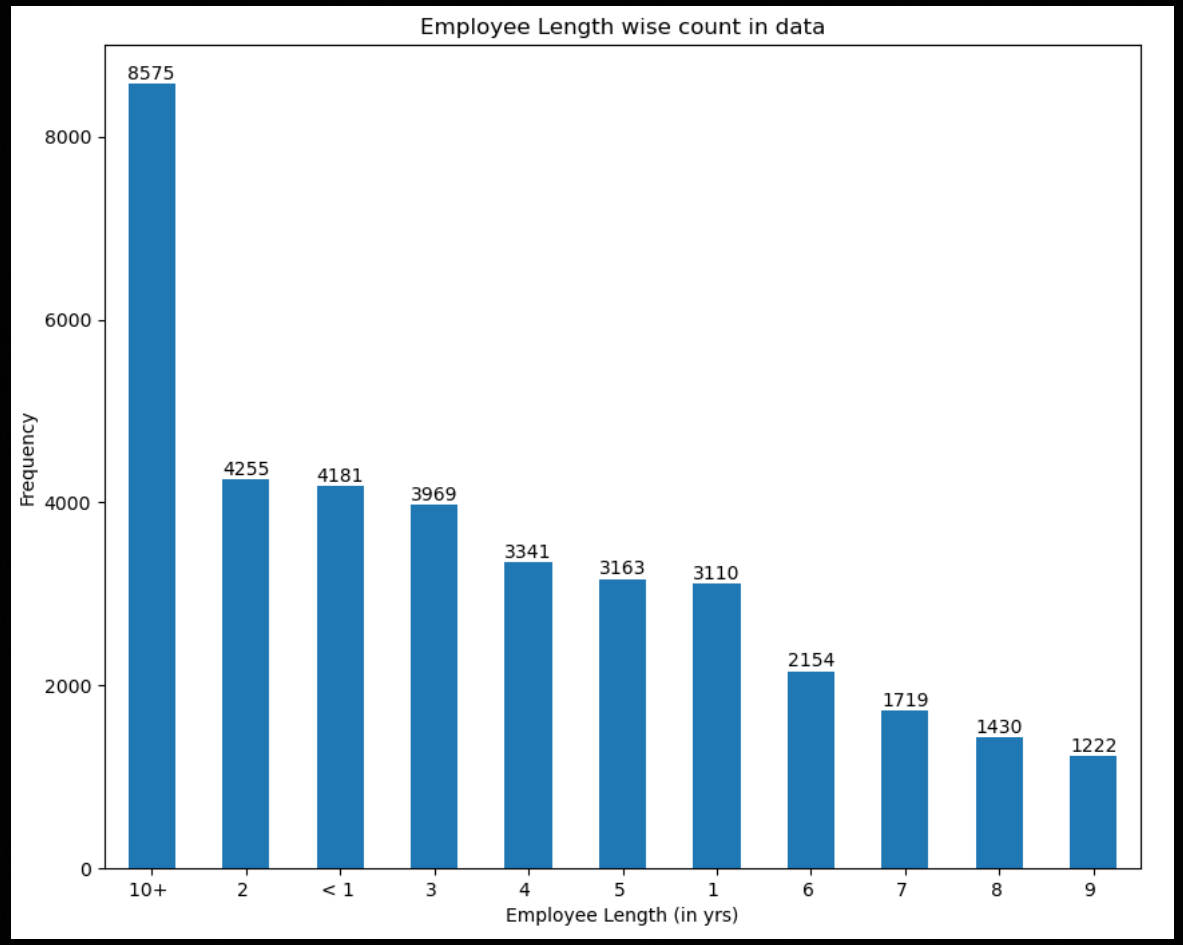
Funded_amount, loan_amount and funded_amount_inv column



It is observed that all these 3 columns have a very similar distribution so for later observations we can use either of these columns and not all 3 since they will add unnecessary bias towards analysis.
Also it is observed that amounts > 30000 seems to be outliers

Data Preprocessing:

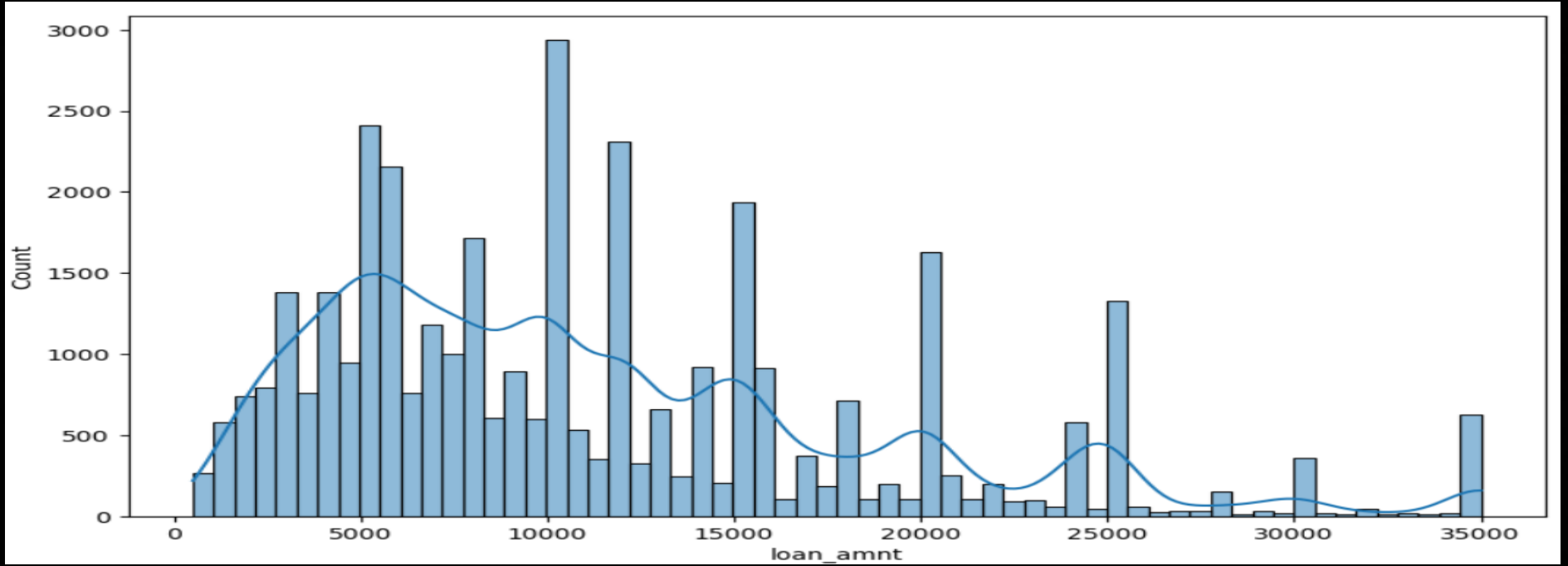
Cleaning: Handling missing values, correcting errors.
Feature Engineering: Debt-to-income ratio, loan-to-value ratio.
Normalization: Scaling numerical features for model uniformity.



We have maximum data of employees who were employed for more than 10 years

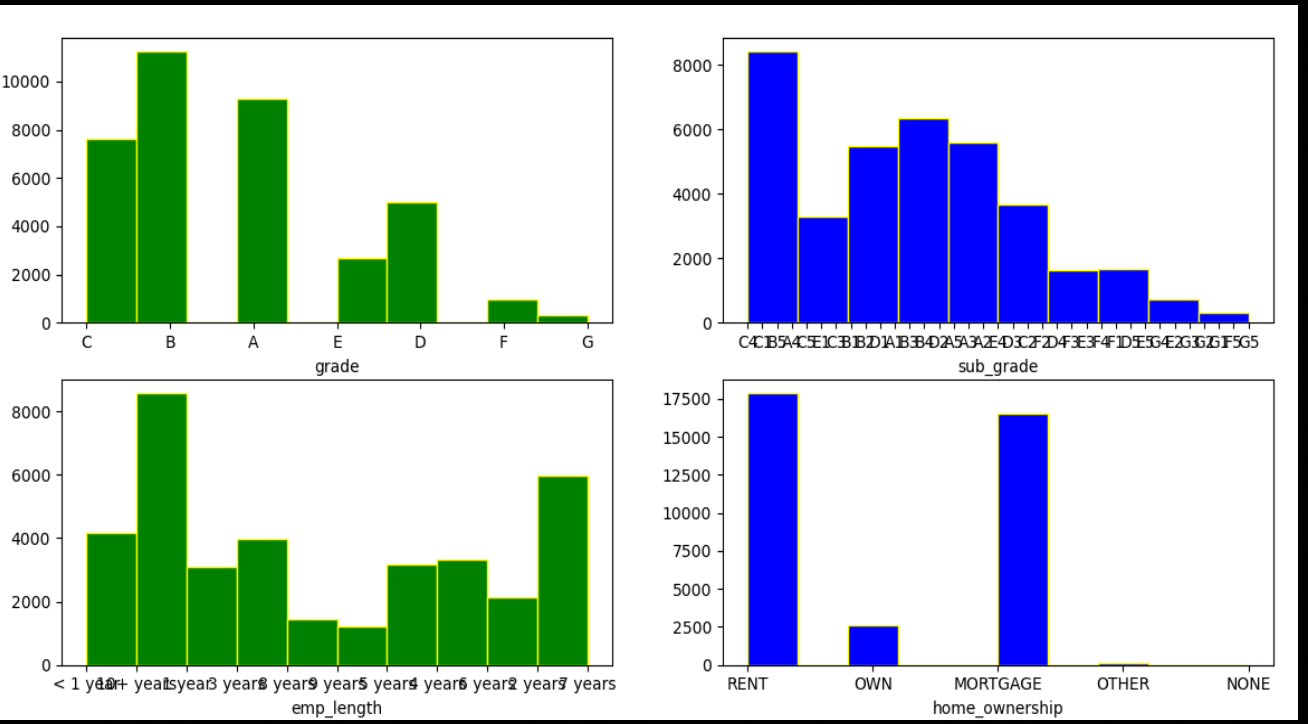
Exploratory Data Analysis (EDA)

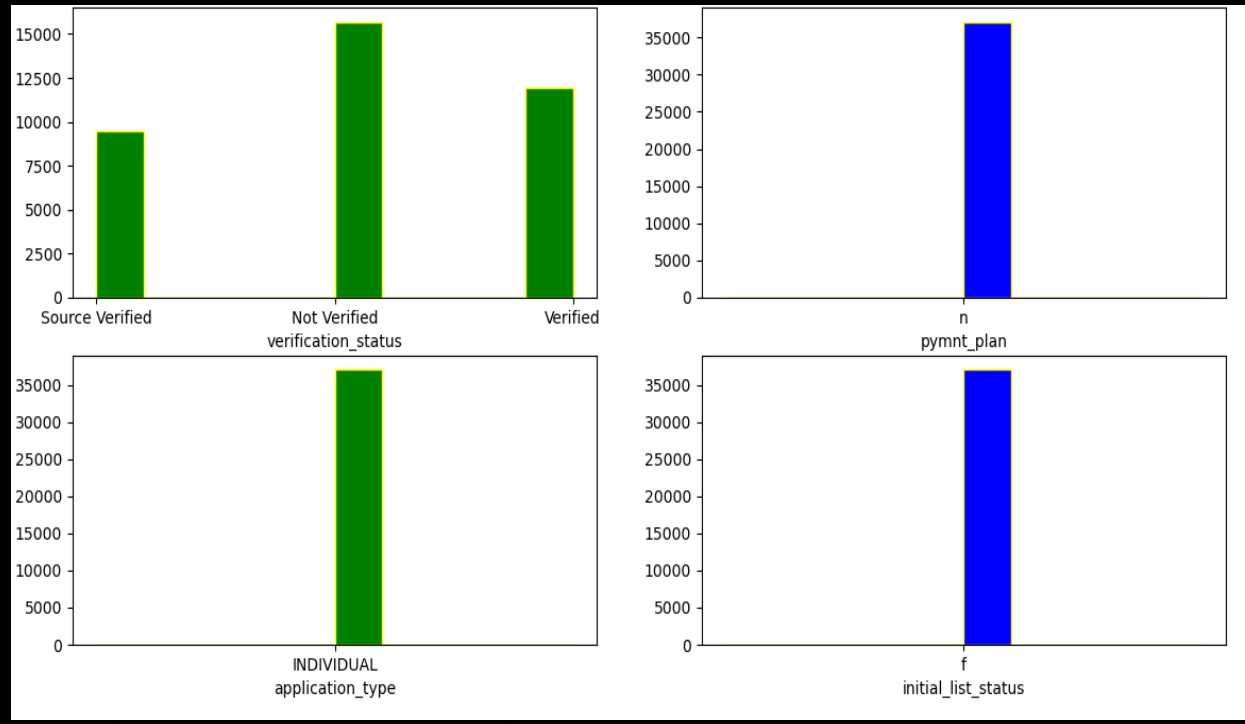
Distribution Analysis: Insights into applicant demographics and financial profiles.
Correlation Analysis: Key correlations between features and loan repayment likelihood.
Class Imbalance: Identification and handling of imbalance between repayers and defaulters



From the loan amount distribution it is observed that there are spikes near multiples of 5000s. This is because of rounding off from borrowers where if there is a necessity of loan amounts of for example 4676, they take a round amount of 5000.

Exploratory Data Analysis (EDA)





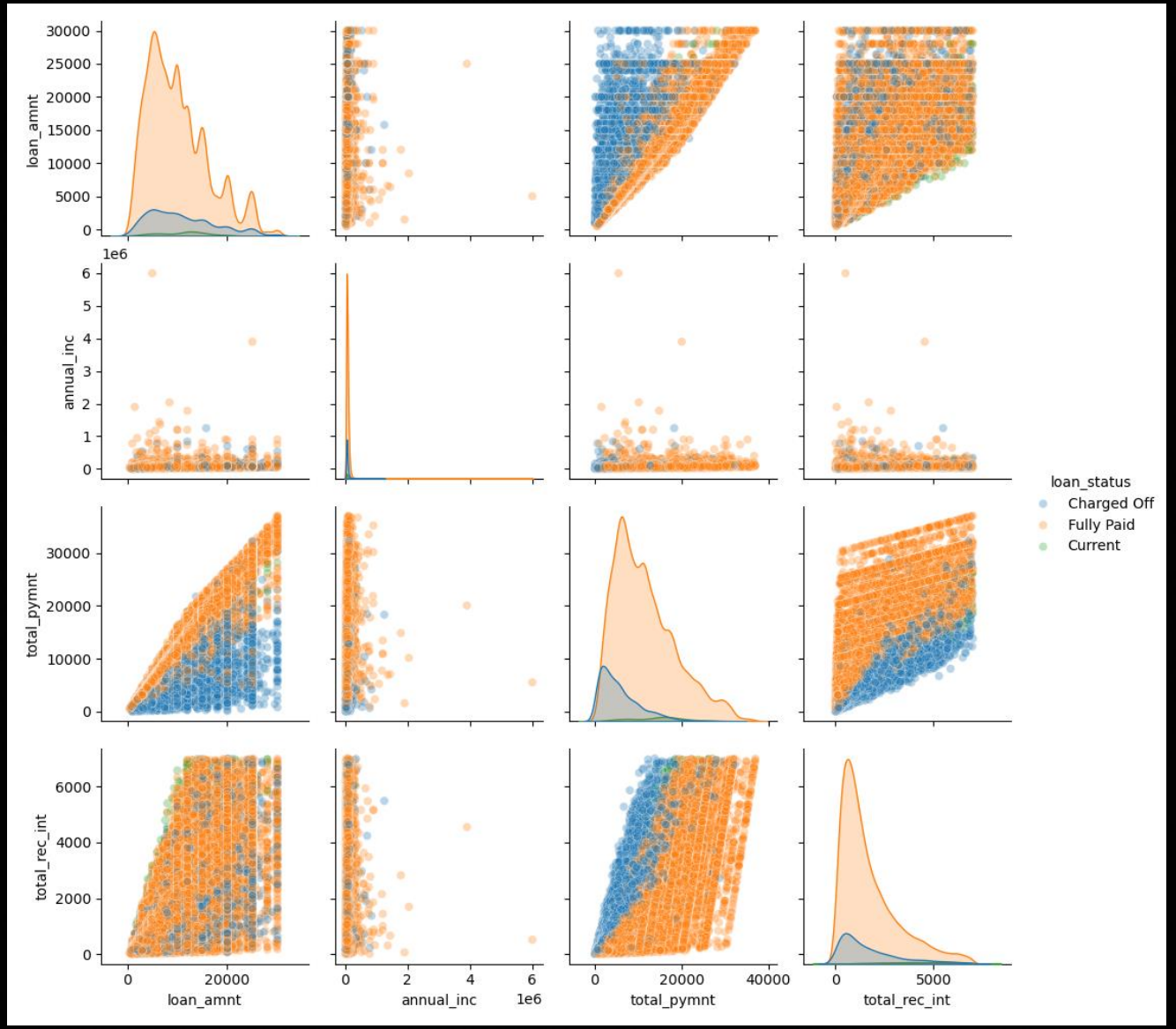
From the plots seems like 'grade','sub_grade','emp_length','home_ownership','verification_status' seem to be useful and columns like 'pymnt_plan','application_type','initial_list_status' are not much useful since they have only 1 value

Exploratory Data Analysis (EDA)

Scatter plot across few important numerical columns

Columns taken : loan_amnt, annual_inc, total_pymnt and total_rec_int

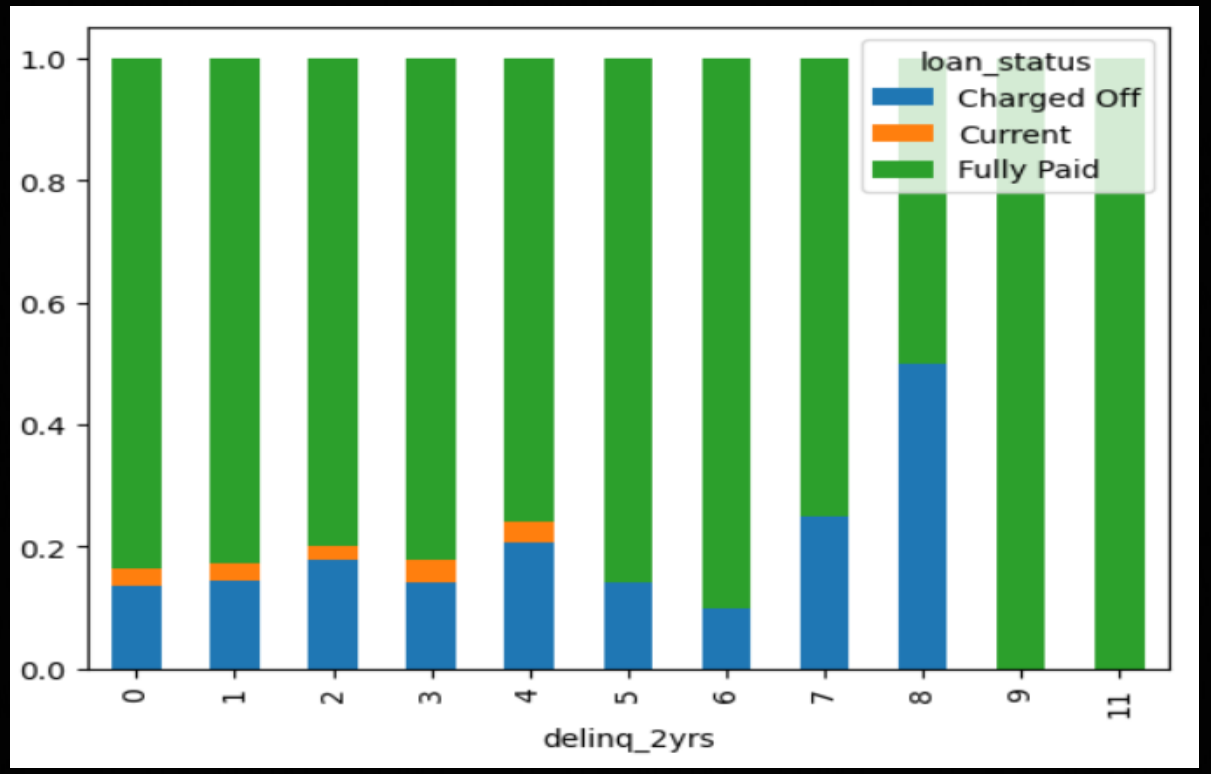
Total_pymnt seems to have a good correlation with the total_rec_int column. The total_rec_int is the total interest received and total payment is the total payment received, hence there is a correlation. We can observe that the charged off loans lie in the lower total payment range. Not much correlation is visible in the other columns



Case study

How does the value of delinquencies in the past 2 years affect the loan default?

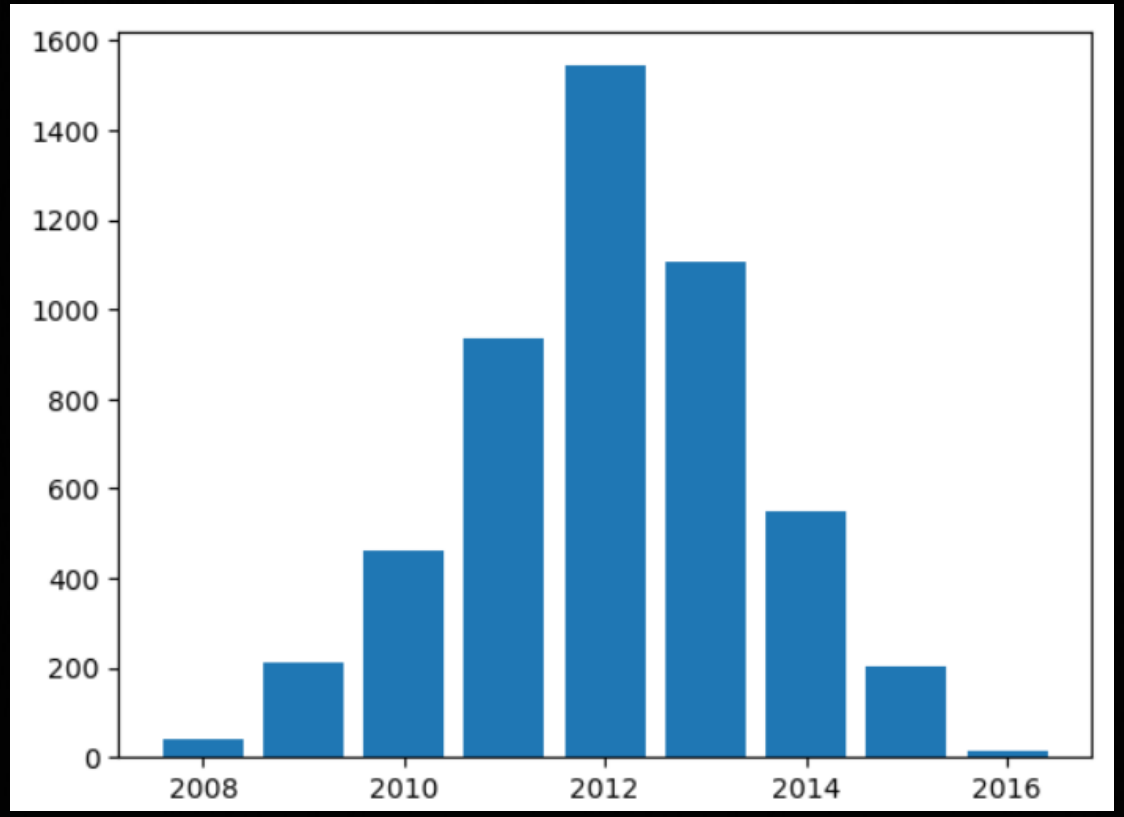
loan_status delinq_2yrs	Charged Off	Current	Fully Paid
0	4455.0	954.0	27671.0
1	449.0	89.0	2554.0
2	115.0	14.0	515.0
3	29.0	8.0	169.0
4	12.0	2.0	44.0
5	3.0	0.0	18.0
6	1.0	0.0	9.0
7	1.0	0.0	3.0
8	1.0	0.0	1.0
9	0.0	0.0	1.0
11	0.0	0.0	1.0



Since the number of data points we have for fully paid customers are very high, this graph might be a bit skewed, but the chances of customers not paying off the loan is clearly higher for those customers who have atleast 7 times 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.

Case study

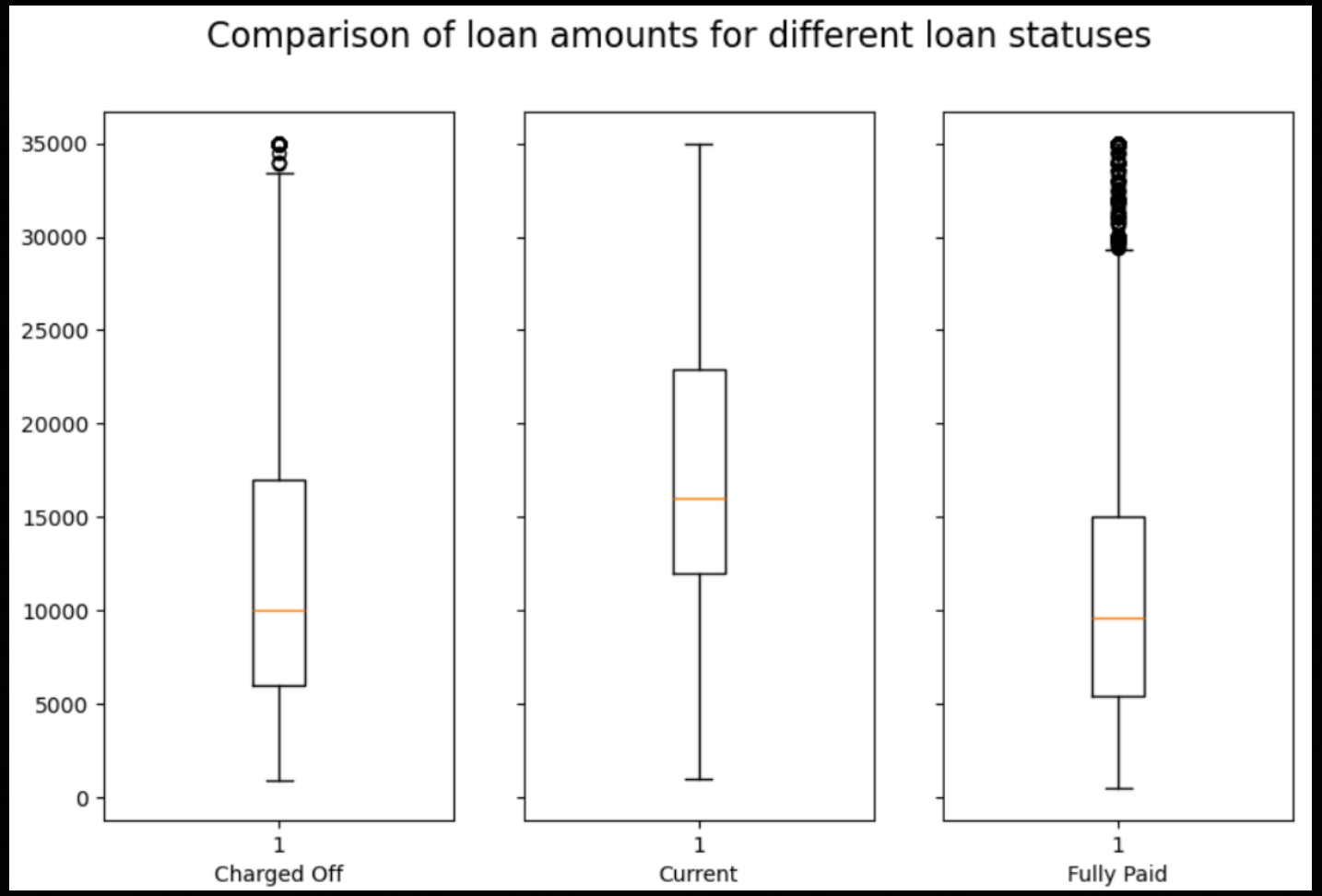
Which was the last payment year and month which lead to maximum defaulters?



last_pymt_year	last_pymt_month	
2012	10	161
	7	156
	6	142
	8	140
	2	134
...	...	...

The most number of defaulters had done their last payment in the month of October 2012. Most defaults have happened in 2012. Some event has occurred in 2012 which might have lead to many people not being able to continue their due installments.

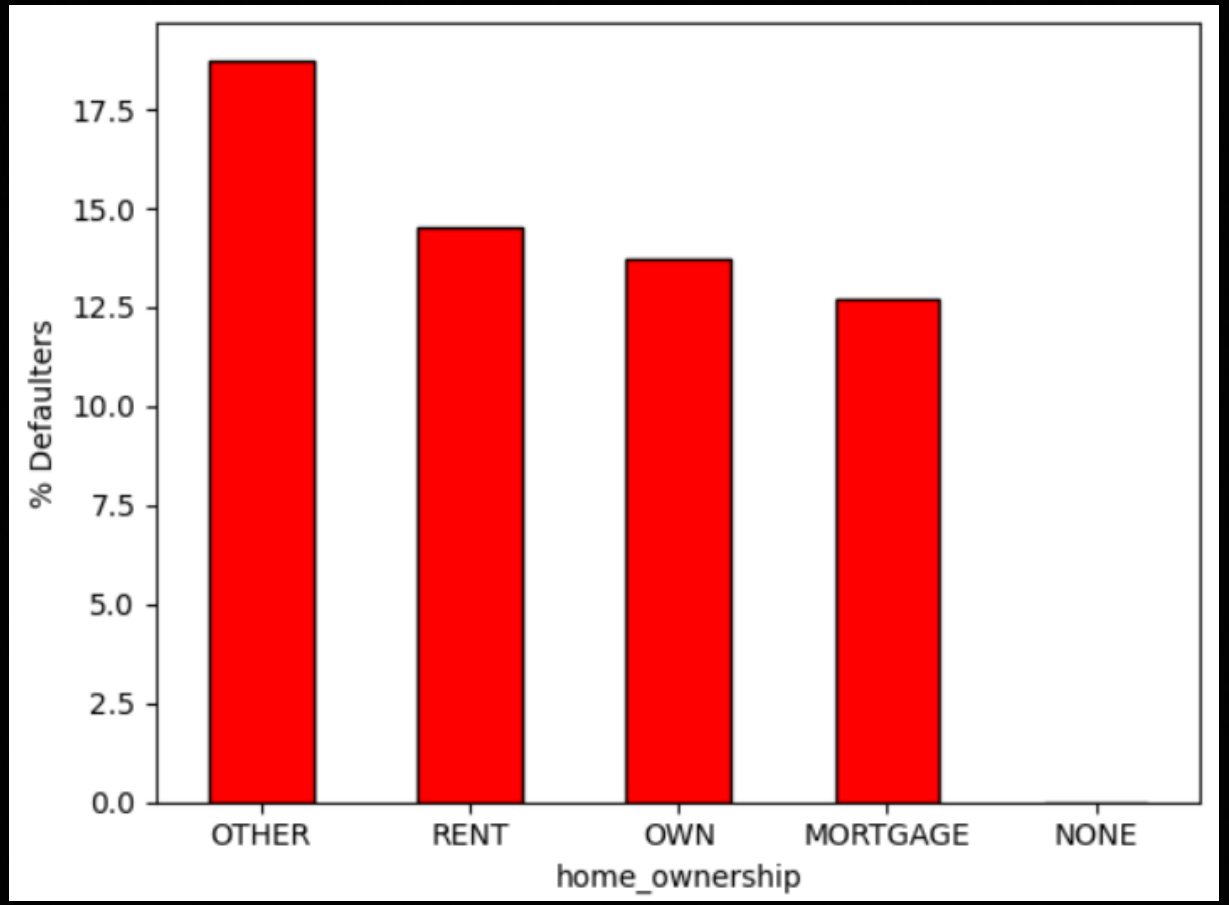
Case study



On an average higher amount of loans get charged off. The spread of charged off loans goes towards the higher loan amounts. The 75th percentile is above 15000 for charged off loans whereas it is lesser than 15000 for fully paid loans.
The average loan amount which gets charged off is 12277.23

Case study

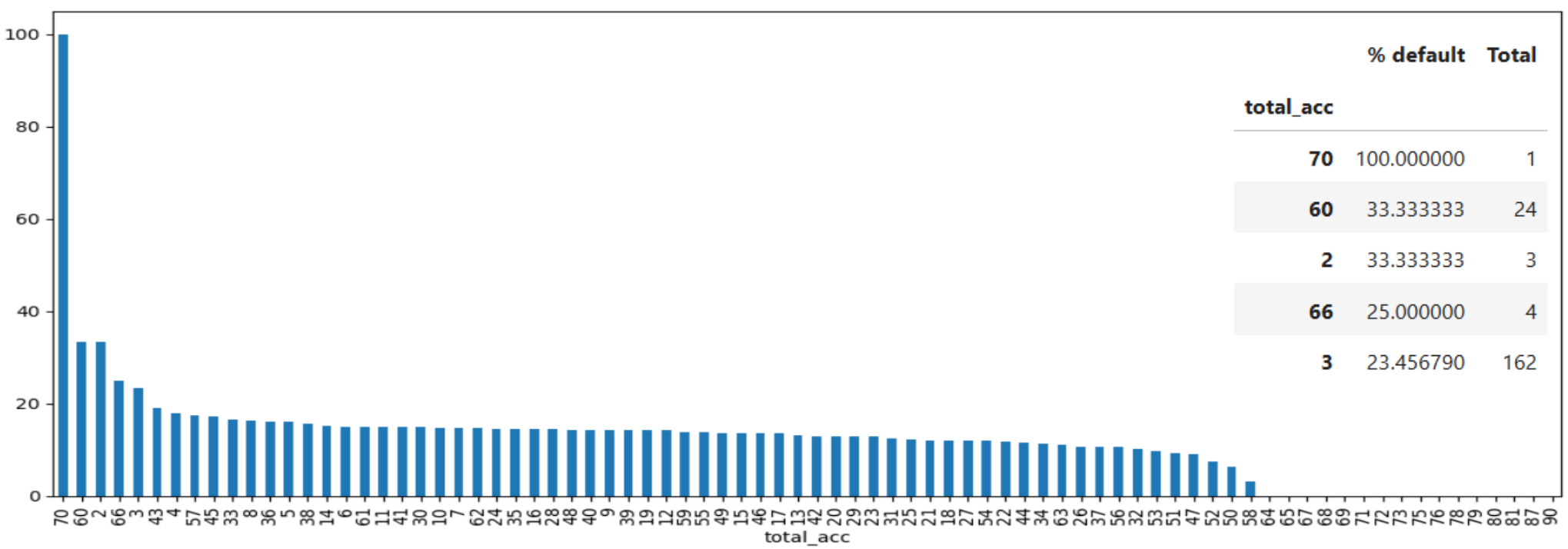
% of defaulters with respect to home ownership category.



The total % of borrowers who default is higher for borrowers who have homeownership type is 'OTHER'. But following that the next highest % of defaulters are from borrowers who take house on Rent. Thus these 2 categories are hot spots to look at when giving a loan.

Case study

% Defaulters compared to the total credit lines

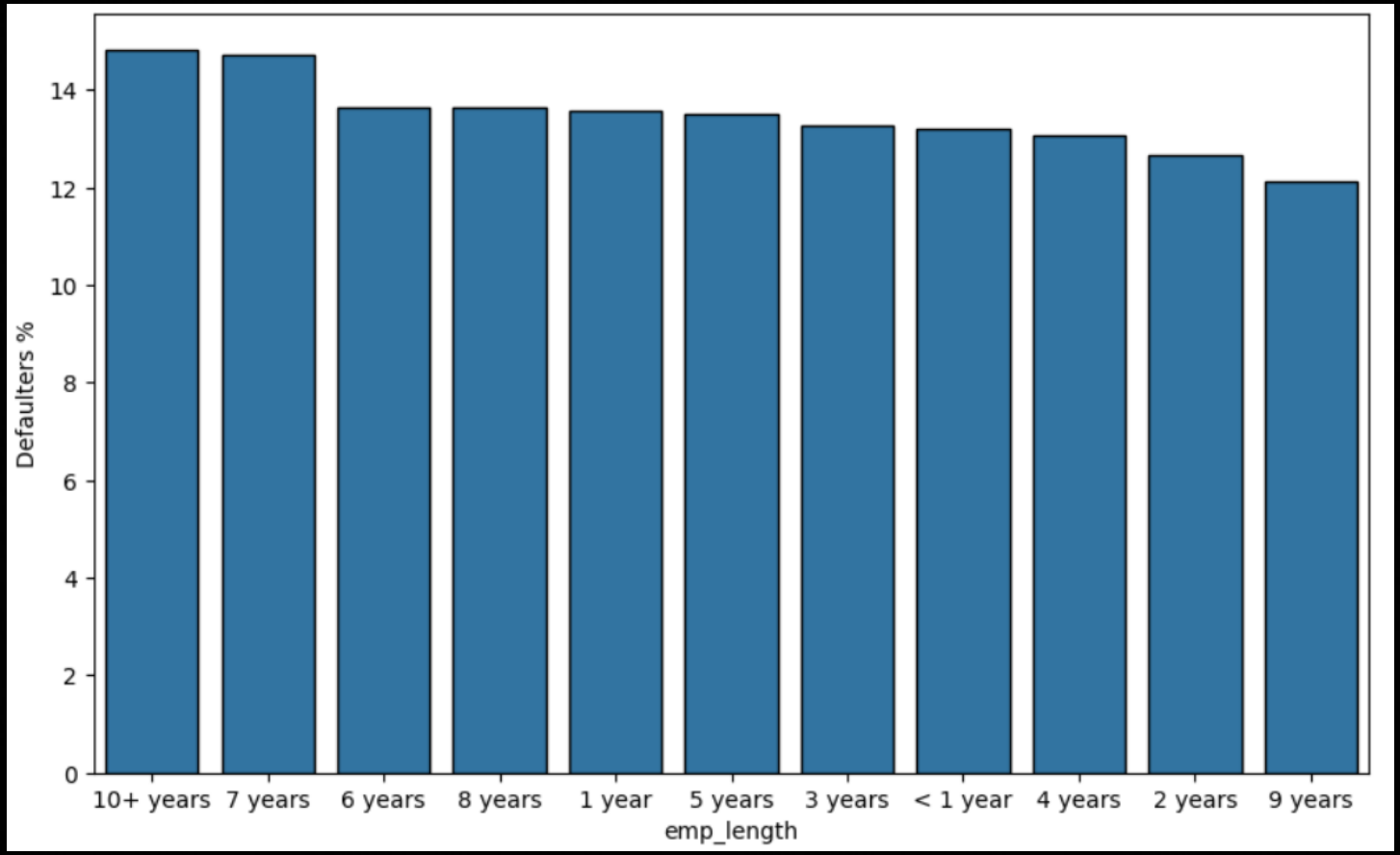


total_acc	% default	Total
70	100.000000	1
60	33.333333	24
2	33.333333	3
66	25.000000	4
3	23.456790	162

The only person who has 70 credit lines has defaulted. But we cannot take that as a definitive fact due to less number of data. But overall people who have either 60 credit lines or 3 credit lines seem to have defaulted with a good percentage. So these are few indicators to check whether loan should be given.

Case study

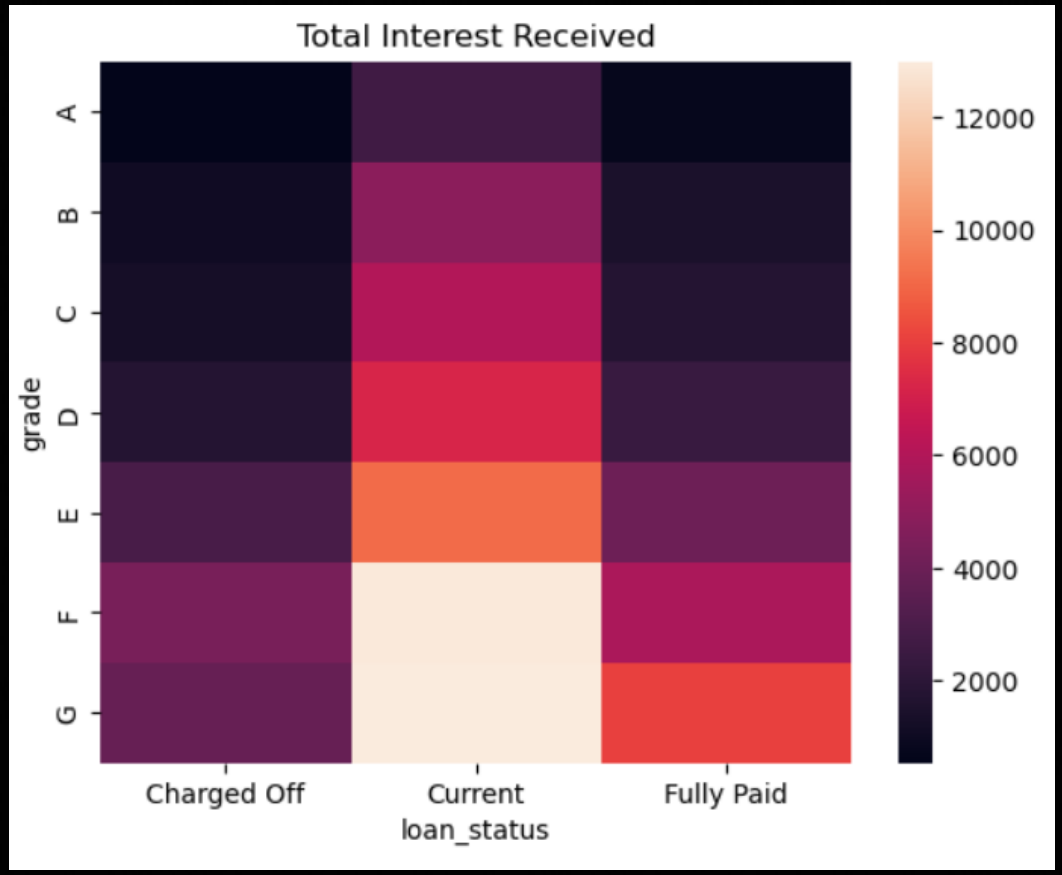
% Defaulters compared to the length of employment



There seems to be higher percentage of defaulters in employees who have worked for more than 10 years. It is quickly followed by an employment length of 7 years. It maybe the case that after 7 years of working people tend to apply loans for various categories like house loan, car loan or even loan for their businesses but after that are unable to pay back the same. But overall there is no stark difference for any category

Case study

Total interest received for different grades of loans and whether loans have got charged off



It is observed that there is a difference in the higher grades F and G. When the amount of interest received is high (6000 and above) those loans tend to get fully paid. If the interest received is around 4000 or lesser, those loans tend to get defaulted. Hence this is an indicator on when the loan might get defaulted.

THANK YOU