GRAMENER CASE STUDY

SUBMISSION

Group Name:

1. Amit Raheja
2. Sachita Chauhan
3. Prashant Sahu
4. Manmeet Singh

Objective

Lending Club is a consumer finance company which is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

When this company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss. The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders.

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.

Our objective as part of working with this organization is to analyze the loan data set provided and extract the major driving factors from this data set. Based on the analysis, possible solutions needs to be provided to the company.

Data Understanding and Manipulation

- There were 111 data columns (variables) in loan data set.
- Of these 111 variables, 54 variables had all observations as NA. These columns were removed.
- Two other columns had more than 90% observations as NA were also removed.
- The columns id and member_id uniquely identified each loan customer. Of these one of the column member_id was removed.
- Some columns had only one unique value across all observations. These columns were removed as well.
- Some more columns from where logically didn't look good for analysis, were also removed.
- After all above deletions, the resultant data set had 38 relevant data columns provided by data set.
- The data for columns int_rate and emp_length was cleaned so as to make it look ordered categorical data.
- The column **loan_status** which had three distinct values **"Charged Off"**, **"Current"** and **"Fully Paid"** was taken as base for the analysis and almost the complete analysis evolved in understanding the variations on this columns across different segments in other variables with more focus on **"Charged Off"** cases.
- The new derived bin columns were created from different measure columns in the provided data and these derived columns were also used for the analysis across bins.

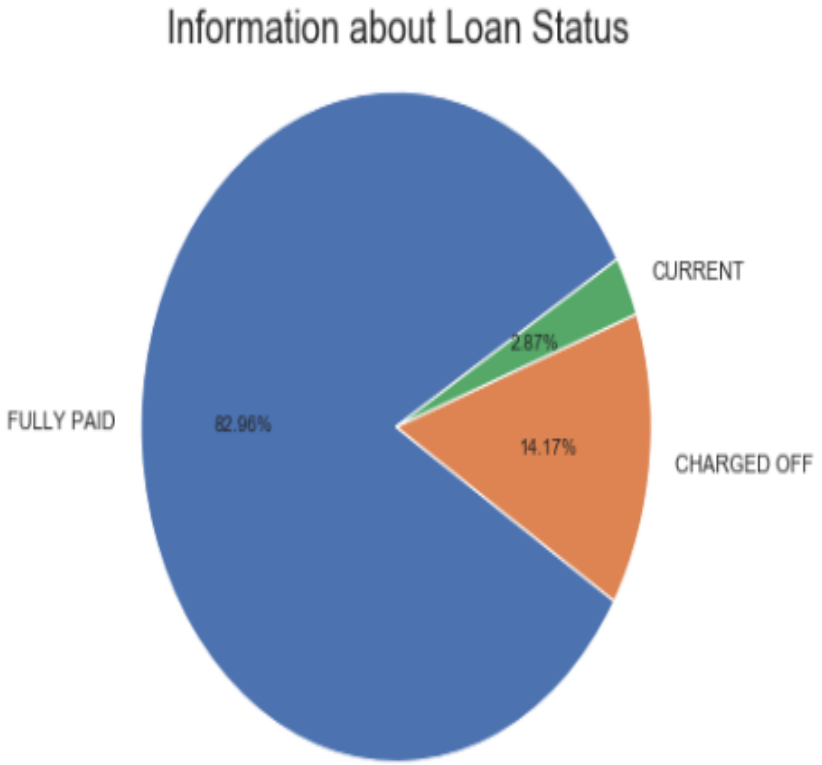
Data Analysis

Overall Distribution of Loan Status

The pie chart shows that the among 39717 borrower loan entries provided, **14.17%** of the borrowers defaulted the loans.

In other words, with respect to the probability we can say that the probability of borrower getting defaulted is 14.17%.

This is the overall understanding of loan data. This will be the basis of our analysis going further. The next steps will focus on understanding the variations of this percentages across different segments of other existing or derived variables. We will compare the increase and decrease in this variation across different segments to make conclusion on the data.



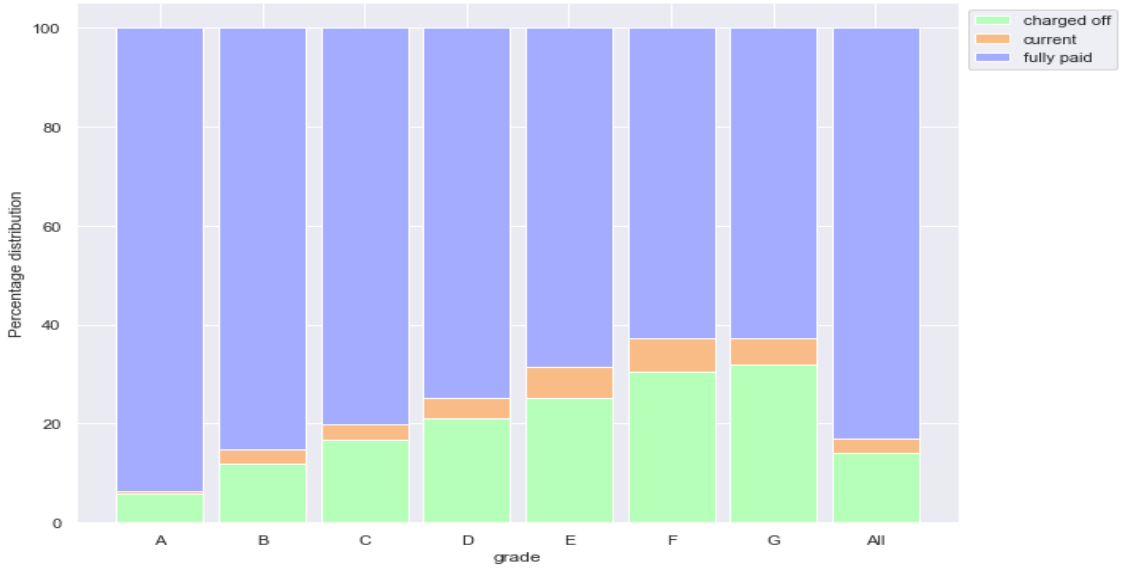
Segmented Univariate Analysis based on Grade

loan_status	CHARGED OFF	CURRENT	FULLY PAID	All
grade				
A	5.97	0.40	93.63	100.0
B	11.86	2.87	85.27	100.0
C	16.63	3.26	80.11	100.0
D	21.07	4.18	74.75	100.0
E	25.16	6.30	68.54	100.0
F	30.41	6.96	62.63	100.0
G	31.96	5.38	62.66	100.0
All	14.17	2.87	82.96	100.0

Assumption: The Grades assigned by Lending Club are ordered categorical from A to G.

Observation: The percentage of charged-off cases increases as grades change from A to G.

Suggestion: Lending Club should be aware about the scenario and more rigorous checks should be carried out while issuing / approving loans to high grade(like E, F, G) borrowers. The loan portfolio for such grades customers needs to be altered accordingly.

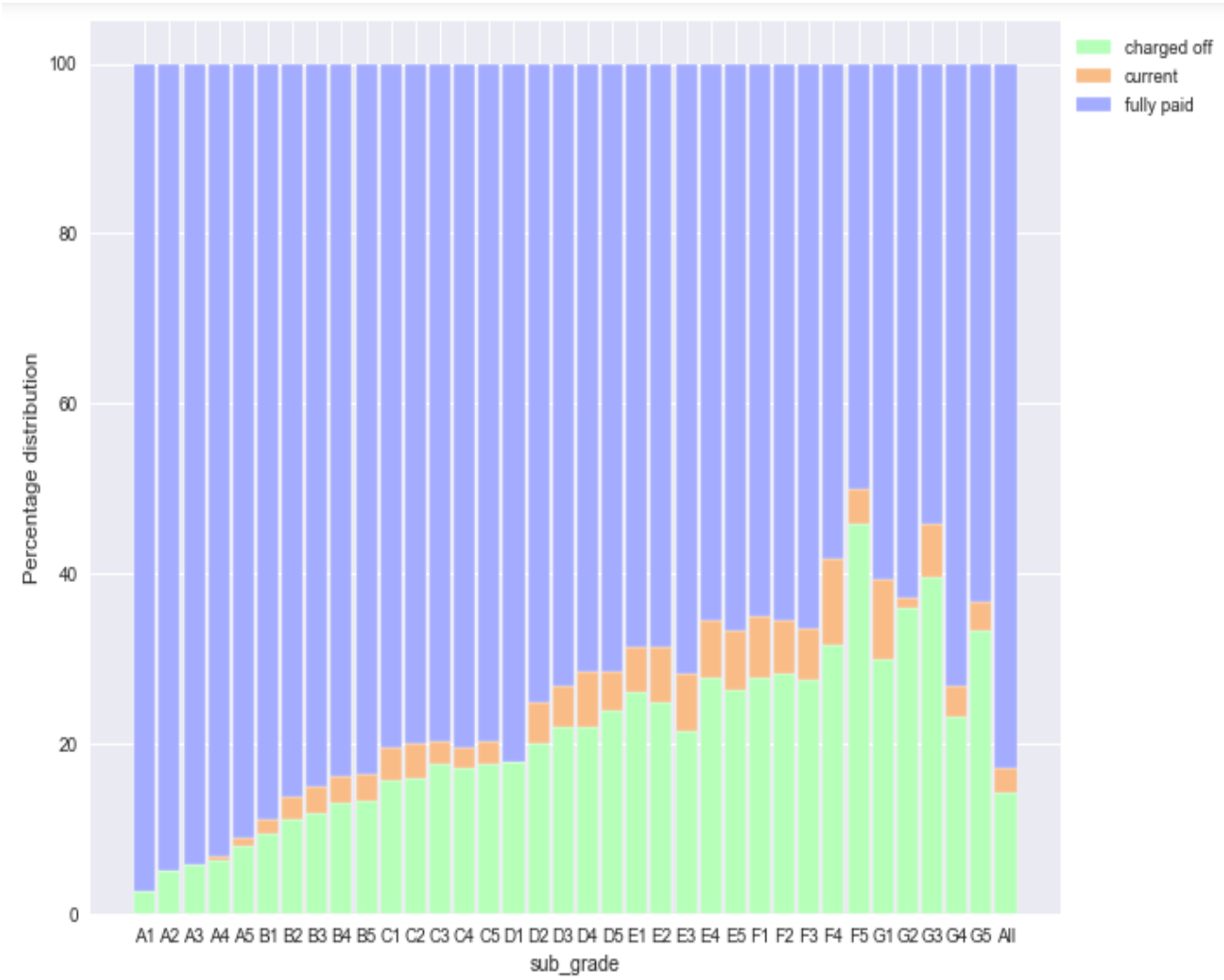


Segmented Univariate Analysis based on Sub Grade

Assumption: The Sub Grades assigned by Lending Club are ordered categorical corresponding to each grade (separate column). There are five sub-grades in each grade.

Observation: The percentage of charged-off cases increases as sub-grades change from one to five.

Suggestion: Lending Club should be aware about the scenario and more rigorous checks should be carried out while issuing / approving loans to high sub grades across each grade for borrowers. The loan portfolio should consider sub-grades into account and plan accordingly.

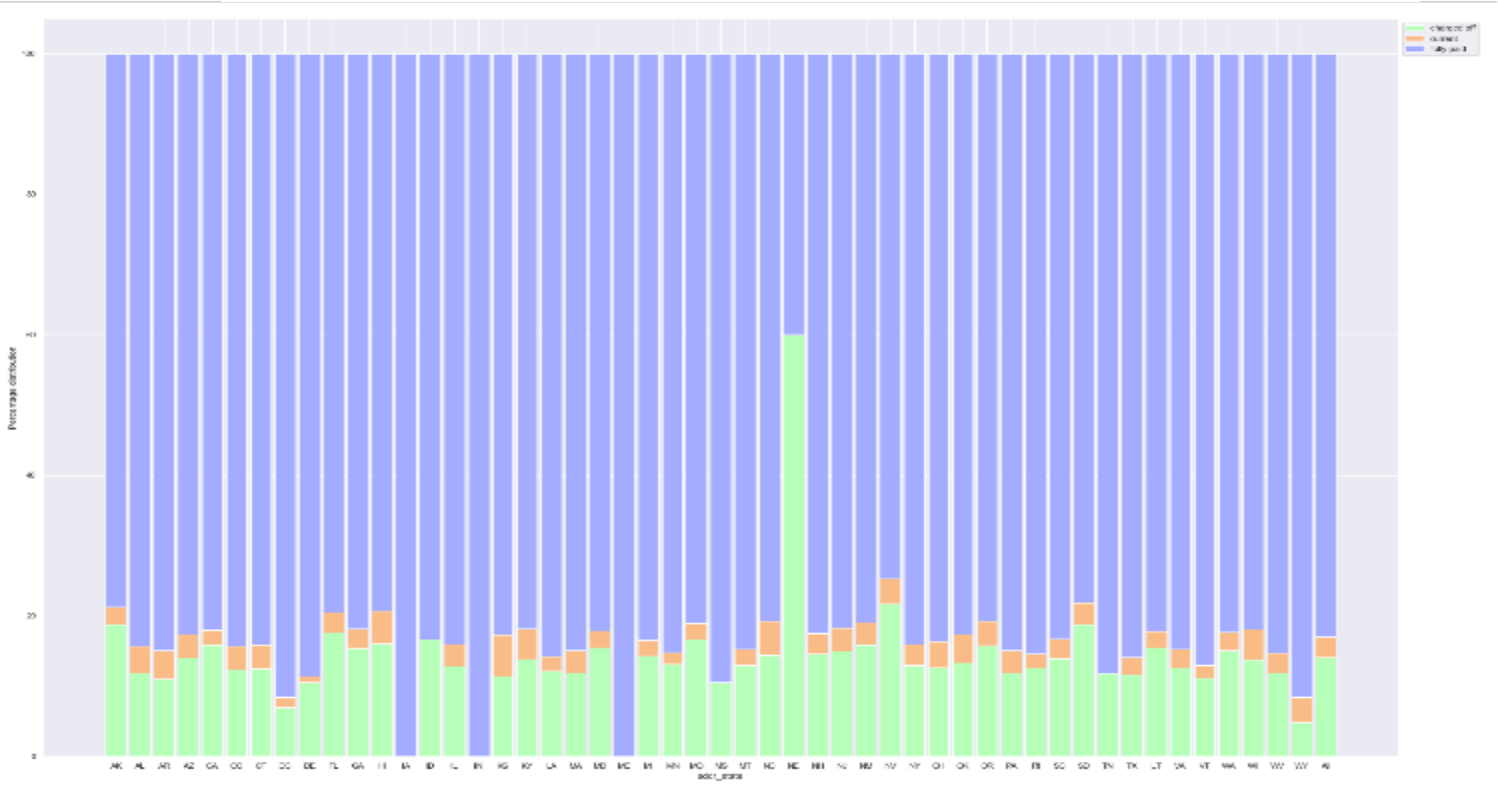


Segmented Univariate Analysis based on Address State

loan_status addr_state	CHARGED OFF	CURRENT	FULLY PAID	All
AK	18.750000	2.500000	78.750000	100.0
NE	60.000000	0.000000	40.000000	100.0
NV	21.730382	3.621730	74.647887	100.0
SD	18.750000	3.125000	78.125000	100.0

Observation: The states represented by AK, NE, NV and SD present relatively more percentage cases of defaulters over the other. Among these, NE state may be taken as an exception as total number of cases from this state were five of which three had defaulted.

Suggestion: The bank should consider the variable as more probability factor of default borrowers from such states. Separate portfolio based on these default risk states can also be designed.

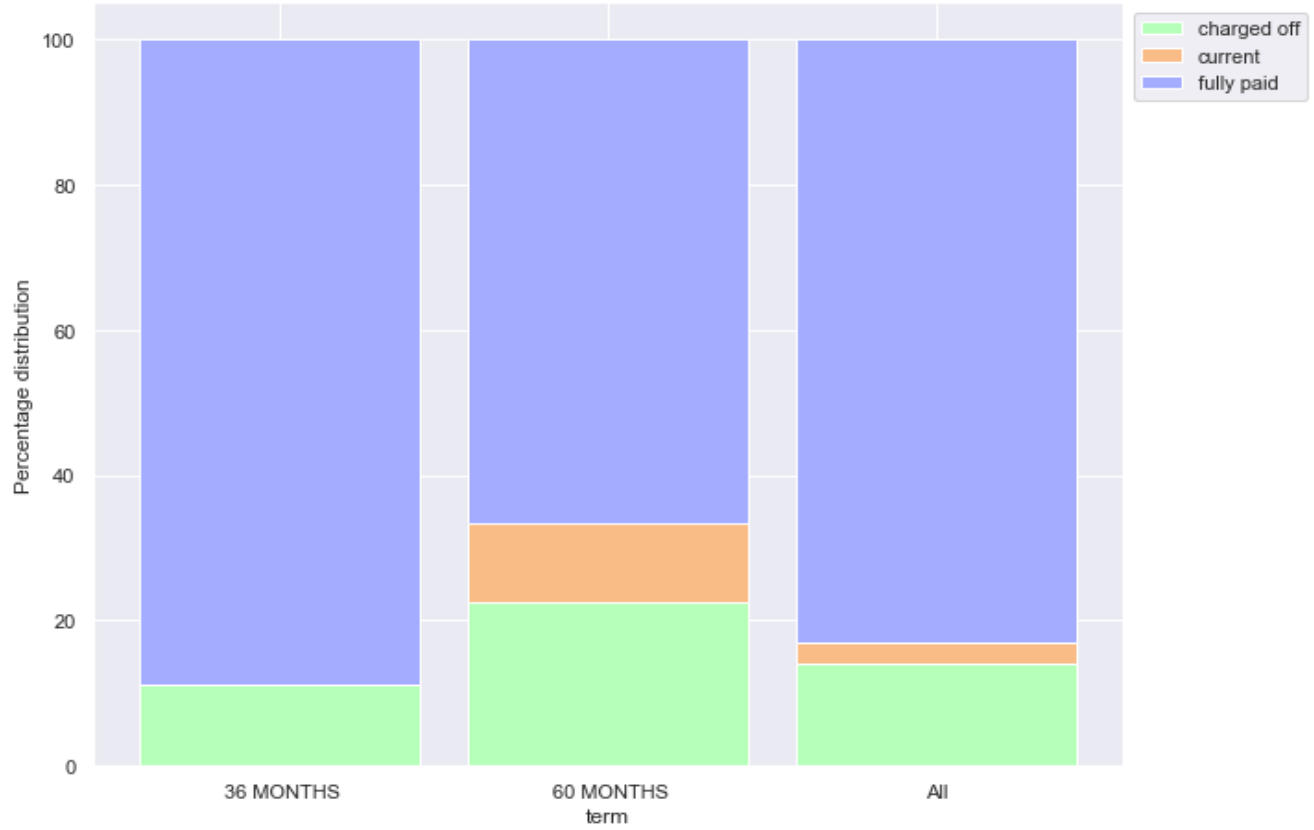


Segmented Univariate Analysis based on Loan Term

Observation: The percentage of defaulters within 60 months loan is relatively higher in comparison to 36 months home loan.

Suggestion: The bank needs to revamp the 60 months duration loan process. The revamp process should include enhanced checks should be applied while issuing loan.

loan_status	CHARGED OFF	CURRENT	FULLY PAID	All
term				
36 MONTHS	11.09	0.00	88.91	100.0
60 MONTHS	22.60	10.73	66.67	100.0
All	14.17	2.87	82.96	100.0

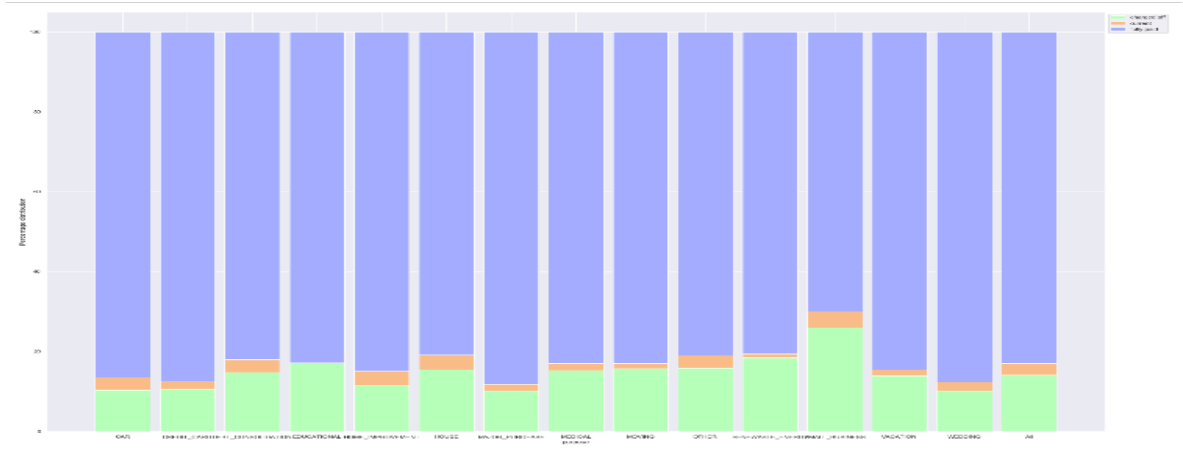


Segmented Univariate Analysis based on Purpose of Loan

Observation: The borrowers whose purpose of loan are Small Business, Renewable Energy, and Educational present relatively high percentage of charged-off cases.

Suggestion: The bank needs to be more vigilant while approving loans to such borrowers which present these reasons. More checks need to be applied for the borrower whose purpose of loan are one of these : Small Business , Renewable Sources and Educational.

loan_status purpose	CHARGED OFF	CURRENT	FULLY PAID	All
CAR	10.33	3.23	86.44	100.0
CREDIT_CARD	10.57	2.01	87.43	100.0
DEBT_CONSOLIDATION	14.84	3.14	82.01	100.0
EDUCATIONAL	17.23	0.00	82.77	100.0
HOME_IMPROVEMENT	11.66	3.39	84.95	100.0
HOUSE	15.49	3.67	80.84	100.0
MAJOR_PURCHASE	10.15	1.69	88.16	100.0
MEDICAL	15.30	1.73	82.97	100.0
MOVING	15.78	1.20	83.02	100.0
OTHER	15.85	3.21	80.94	100.0
RENEWABLE_ENERGY	18.45	0.97	80.58	100.0
SMALL_BUSINESS	25.98	4.05	69.97	100.0
VACATION	13.91	1.57	84.51	100.0
WEDDING	10.14	2.22	87.65	100.0
All	14.17	2.87	82.96	100.0



Segmented Univariate Analysis based on Interest Rates Quartile

Bin Creation: New column classifying Interest rates into low, medium and high was created. The value that spans across each bin is as below :

Low - 5.401 to 11.81

Medium - 11.81 to 18.2

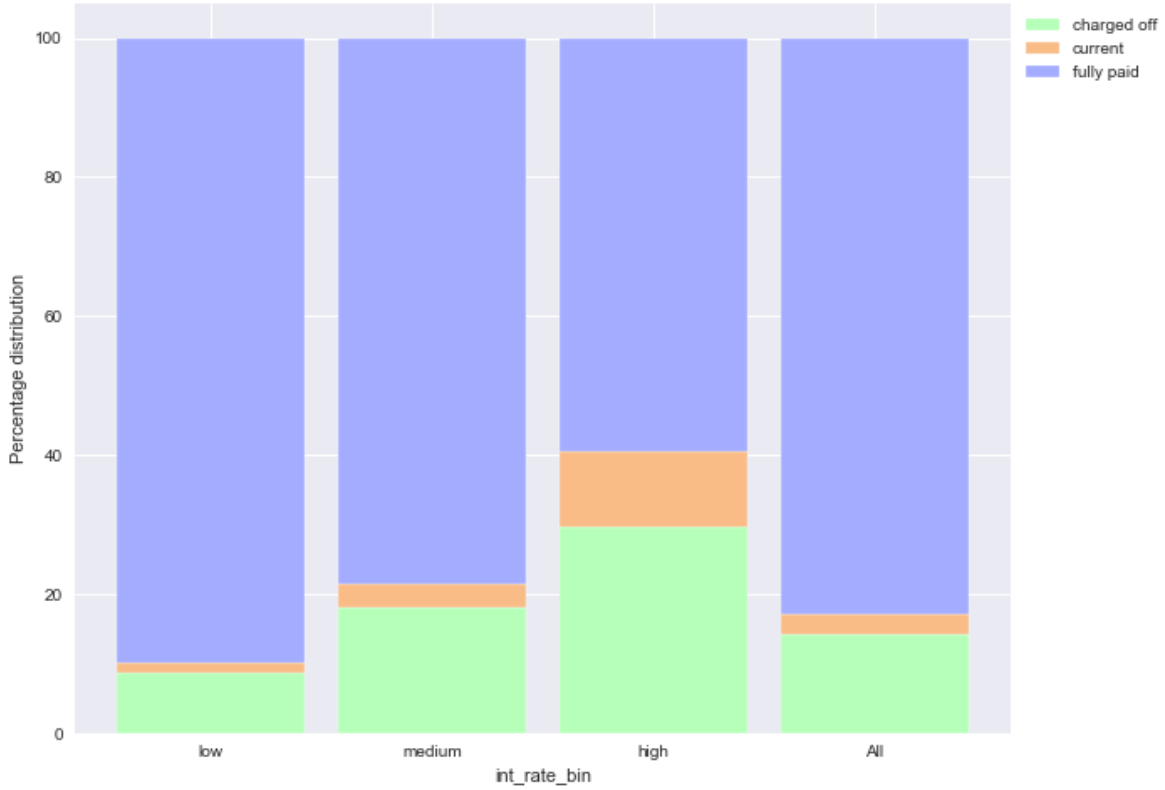
High - 18.2 to 24.59

The effect on loan status across different bins was analysed.

Observation: The percentage of charged-off cases for borrowers taking loan at high interest rates is relatively too high than percentage in other two bins.

Suggestion: Bank needs to consider this factor as probability of default while defining strategies for high values of rate of interest.

loan_status int_rate_bin	CHARGED OFF	CURRENT	FULLY PAID	All
low	8.74	1.32	89.94	100.0
medium	18.05	3.47	78.48	100.0
high	29.54	10.90	59.56	100.0
All	14.17	2.87	82.96	100.0



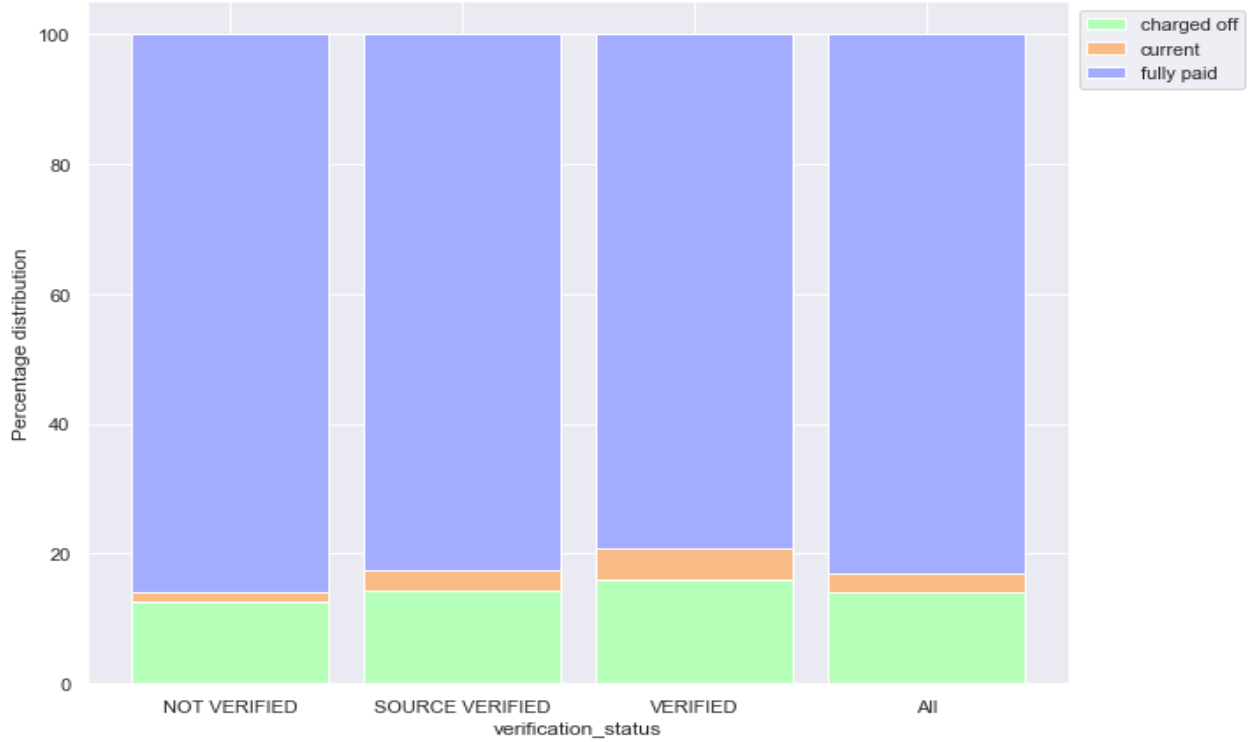
Segmented Univariate Analysis based on Verification Status

Assumption: Borrowers verified by the banks should have less probability of defaults over the borrowers who haven't be verified.

Observation: The percentage of charged-off cases across verified borrowers are relatively more than others. Although, the variation is small, however, it still questions on the quality of verification of the bank based on above assumption.

Suggestion: The bank needs to inspect the quality of its verification process.

loan_status	CHARGED OFF	CURRENT	FULLY PAID	All
verification_status				
NOT VERIFIED	12.66	1.34	86.00	100.0
SOURCE VERIFIED	14.36	3.10	82.54	100.0
VERIFIED	16.01	4.71	79.28	100.0
All	14.17	2.87	82.96	100.0



Segmented Univariate Analysis based on Loan Amount Quartile

Bin Creation: New column classifying Loan Amount into low, medium and high was created. The value that spans across each bin is as below :

Low – 465.5 to 12000

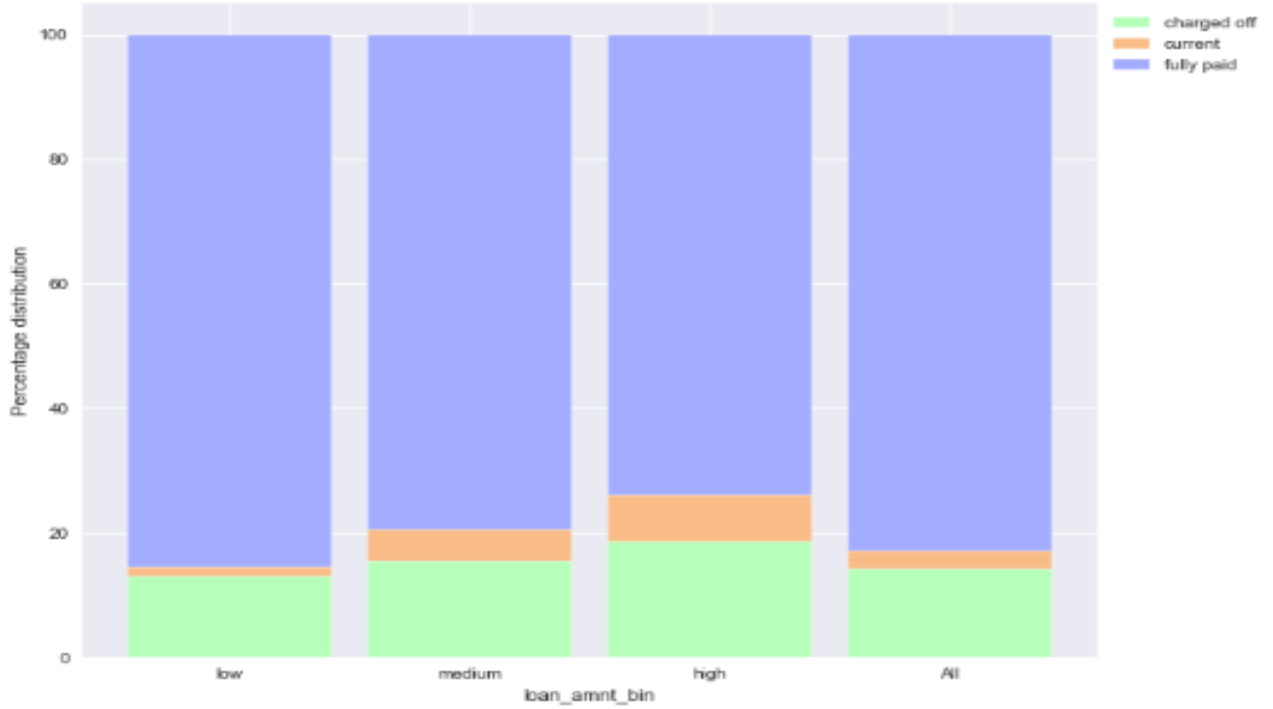
Medium - 12000 to 23500

High - 23500 to 35000

Observation: The percentage of charged-off cases for borrowers taking loan at high loan amount is relatively high than percentage in other two bins.

Suggestion: Bank needs to add this into the risk assessment factor while finalizing the loan amount.

loan_status	CHARGED OFF	CURRENT	FULLY PAID	All
loan_amnt_bin				
low	13.07	1.39	85.54	100.0
medium	15.41	5.03	79.56	100.0
high	18.58	7.56	73.86	100.0
All	14.17	2.87	82.96	100.0



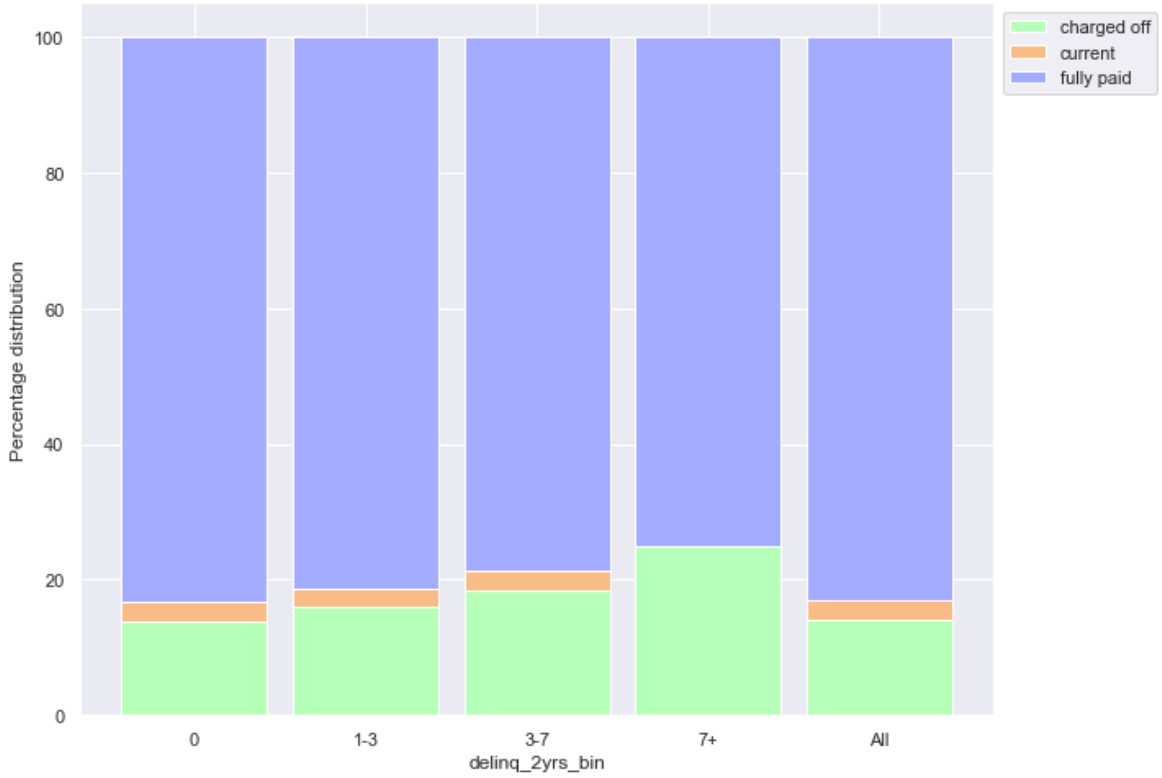
Segmented Univariate Analysis based on number of delinquency in past two years

Bin Creation: New column for classifying the number as 0, 1-3, 3-7 and 7+ was created. The effect on loan status across different bins was analysed.

Observation: The percentage of charged-off borrowers across 7+ delinquencies is higher.

Suggestion: Number of delinquencies over past two years are good indicators of defaulters especially if they are greater than seven. Banks need to make systems so that timely action can be taken.

loan_status	CHARGED OFF	CURRENT	FULLY PAID	All
delinq_2yrs_bin				
0	13.94	2.88	83.18	100.0
1-3	15.96	2.80	81.24	100.0
3-7	18.37	3.06	78.57	100.0
7+	25.00	0.00	75.00	100.0
All	14.17	2.87	82.96	100.0



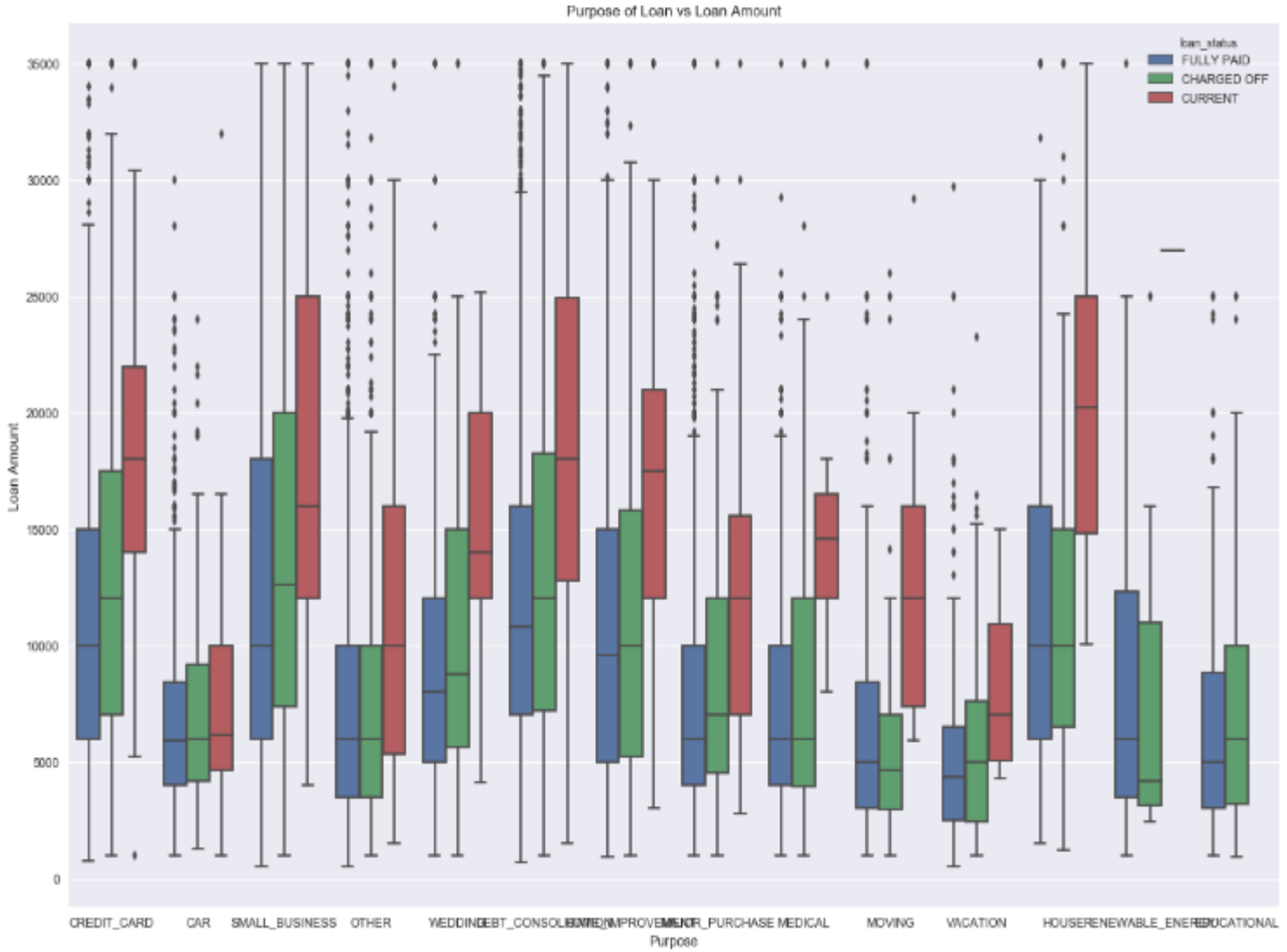
Bivariate Analysis based on Purpose and Loan Amount

Assumption: The amount of loan amount in case of default directly causes loss to bank. The bank is assumed to choose the loan amount carefully to as minimise the risk due to default.

We had already performed univariate showing that loan taken for small business are more probable of making default. Now with this bivariate, we need to check how much loss based on loan amount does it bring.

Observation: The first quartile, median, third quartile values for the charged off cases in small business is more than fully paid loans.

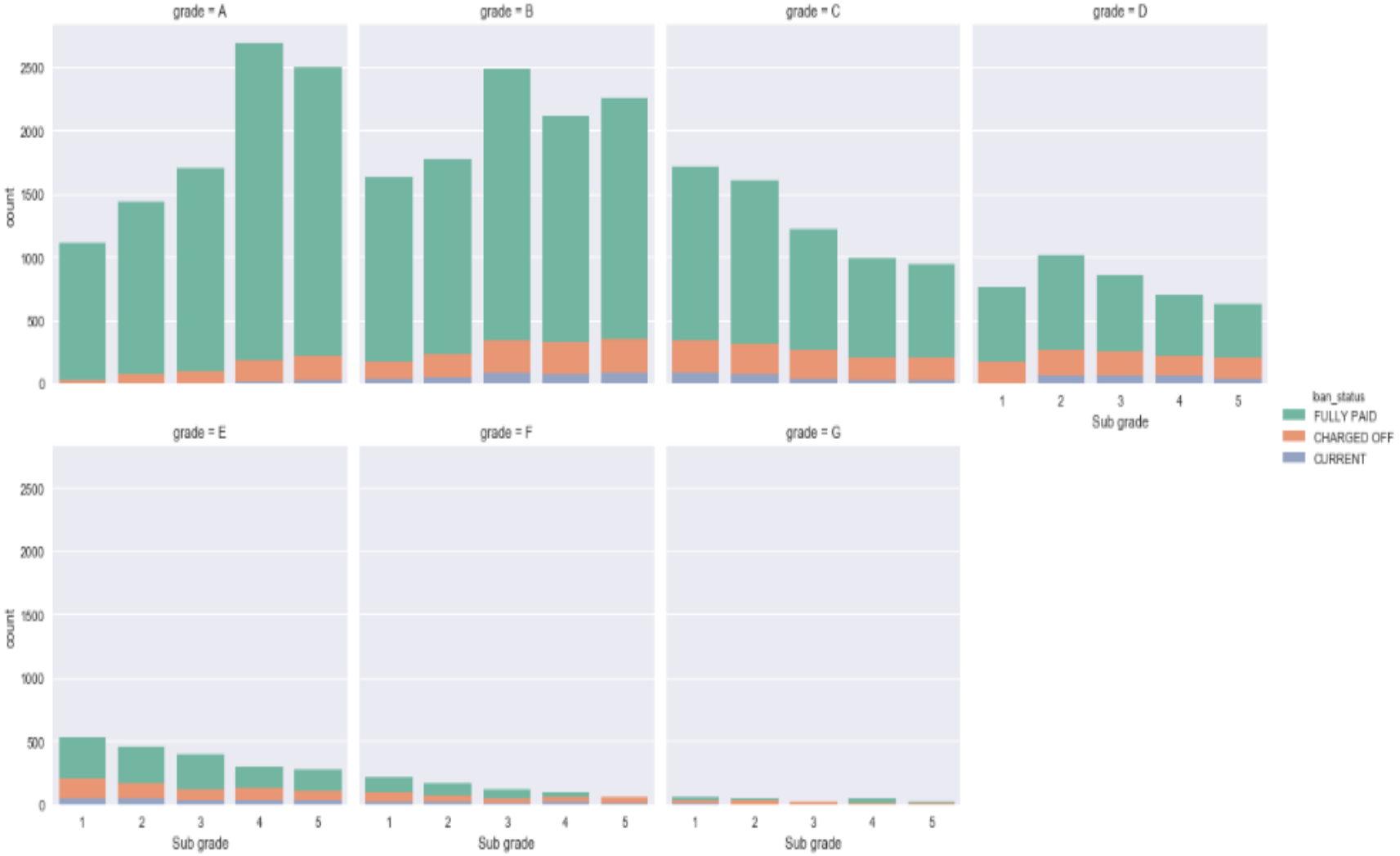
Suggestion: For loan purpose that present high risk of defaults, loan amount should be chosen carefully so as to minimise the risk impact that may arise due to default.



Bivariate Analysis based on Grades and Sub Grades

We had observed the clear variation in percentages across different grades and also sub-grades separately. Now we need to understand if together they present any combination that can impact.

Observation: There was no pattern found based on combination of grades and sub grades.



Conclusions

Based on the analysis conducted across multiple variables, following variables were found to be impacting the variations for loan status.

- Grades
- Sub Grades
- Address State
- Verification Status
- Term of Loan
- Purpose of Loan
- Interest Rate Quartile
- Loan Amount Quartile
- Delinquency in past two years
- Purpose of Loan and Loan amount (bivariate)

The suggestions have been provided corresponding to each findings and classified segments.

Thank You