Lending Club Case Group Study

Prepared by
Prashant Tariwal
Rajesh Balakrishnan

Objective

Understand how data is used to minimize the risk of losing money while lending to customers by using Exploratory Data Analysis.

Observation

Two types of risks are 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.

Two types of Data sets.

- **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:** If company not approved the loan because of not meeting the requirement's of the bank rules.

Summary

Univariate Analysis

- Positive Correlation with loan_amount, funded_amount, funded_amount_inv and installment.
- Interest rate is higher, the chance for 'charged off' is more.
- Loan amount is increasing, there is a high probability of loan getting default.
- *Higher the repayment term chances of charge off will be higher.*
- *Lower grades company default rate is much higher than others.*
- *Loan for small_business, renewable_energy and educational are the riskier ones.*
- *The presence of specific celebrations can be the reason for more charge off 7th & 11th Months.*

Bivariate Analysis

- *Higher installments for any income group have more number of defaults.*
- *As the interest rate and Loan term increases the chance of charged off is more.*
- *As the debt increases the chance to be default is getting increased*

Conclusion

- *Providing the loan to higher grade company employees.*
- *Less loan amount and less interest increase the less Charge off.*
- *Higher annual income lesser the charge off so provide the loan to higher income employees.*
- *Providing the cc loans its helps to when customer is doing Charge off so we can cut the loan from cc account so its very help full and customer can enjoy holidays and celebration without any guilty and loan pressure.*

Steps followed for the case study

Step 1: Data Understanding

Understanding the every column and its significant in the analysis.

Step 2: Data Cleaning

Removing the unwanted rows and columns after understanding the importance of that.

Step 2: Univariate Analysis

By using the every column analyzing the risk factors of the Charged off % and fully paid%.

Step 3: Segmented Univariate Analysis

By using single and multi column we are going to analyzing the risk factors of the Charged off % and fully paid% by clustering the data. Ex : Purpose of loan, Month, occupation, Company grade and state wise.

Step 4: Bivariate Analysis

By using multi column we are going to analyzing the risk factors of the Charged off % and fully paid% by clustering the data. Ex : Purpose of loan, Month, occupation, Company grade and state wise.

Step 5: Recommendations/Results

To reduce the risk factor and the importance of the business we are going to give recommendations and explain the result

Data Cleaning Steps

Delete columns: Delete unnecessary columns.
Remove outliers: Remove high and low values that would disproportionately affect the results of our analysis.
Missing values: Treat missing values with appropriate approach.
Duplicate data: Remove identical rows, remove rows where some columns are identical.
Filter rows: Filter by segment to get only the rows relevant to the analysis.

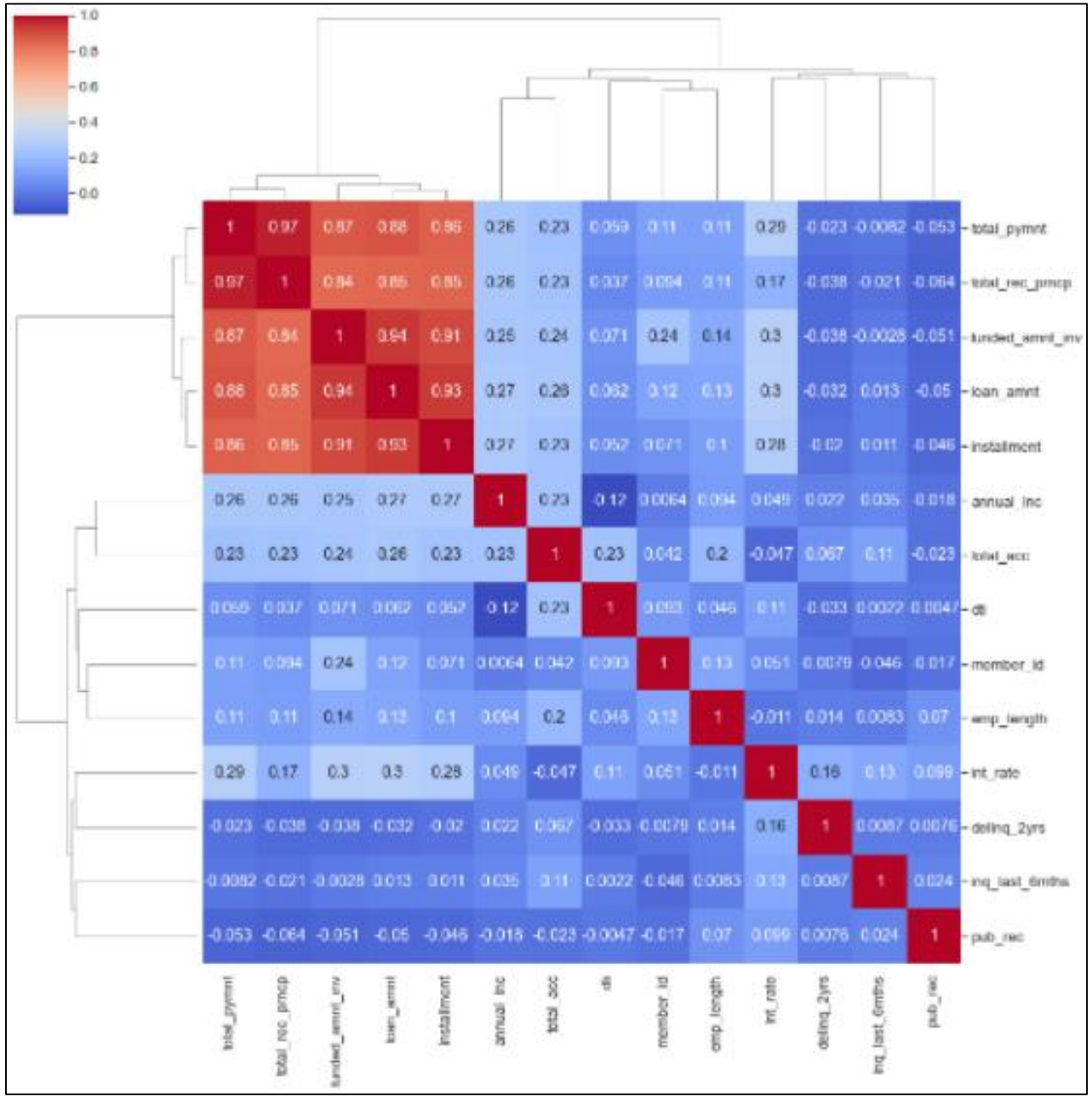
Analysis

- ❑ The whole project is to analyze and understand how consumer and loan attributes are influencing the tendency of Charge off.

- ❑ During analysis used the different plots.

- ○ Histograms and Bar charts to check out the distribution of all the variables
- ○ Box plots to detect the Outliers
- ○ Performed the Bivariate analysis to understand how different variables interact with each other and how its affecting the Charge off% and fully paid%.

Univariate Analysis



Let us visualize the data in a Heat Map

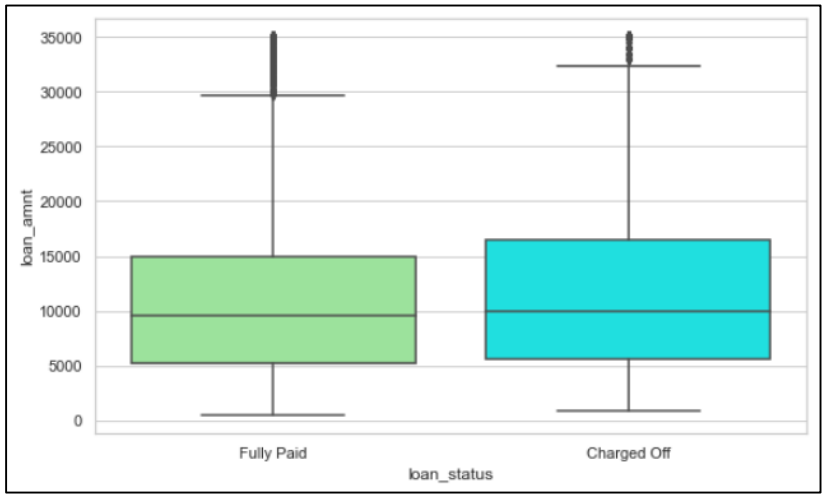
We have Positive Correlation Among Below Values.

loan_amount, funded_amount, funded_amount_inv and installment

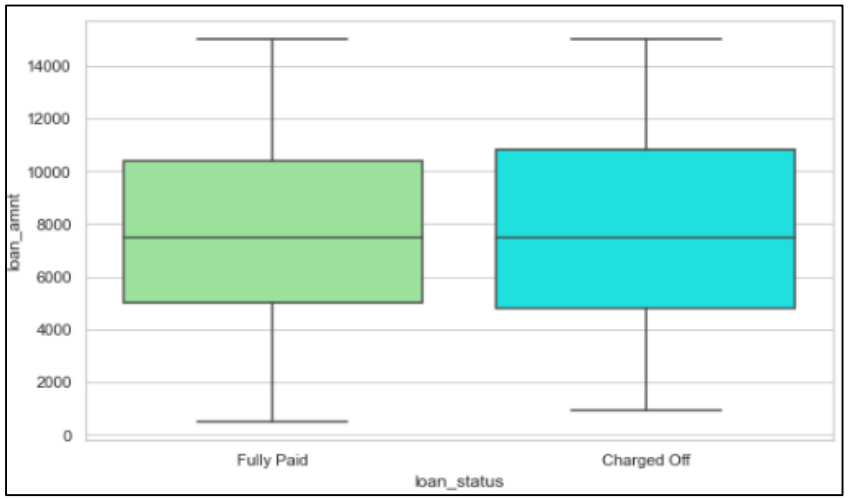
public records related fields pub_rec & pub_rec_bankrupcies and number of accounts related fields open_acc & total_acc are correlated

Univariate Analysis

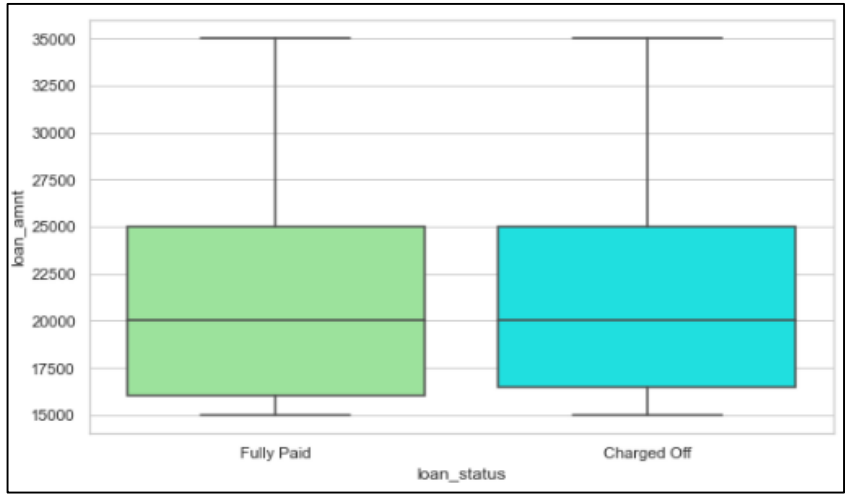
Loan amount took min to max for analysis



Loan amount took min to 75% for analysis



Loan amount took 75% to Max for analysis



count	38577.000000
mean	11047.025430
std	7348.441646
min	500.000000
25%	5300.000000
50%	9600.000000
75%	15000.000000
max	35000.000000

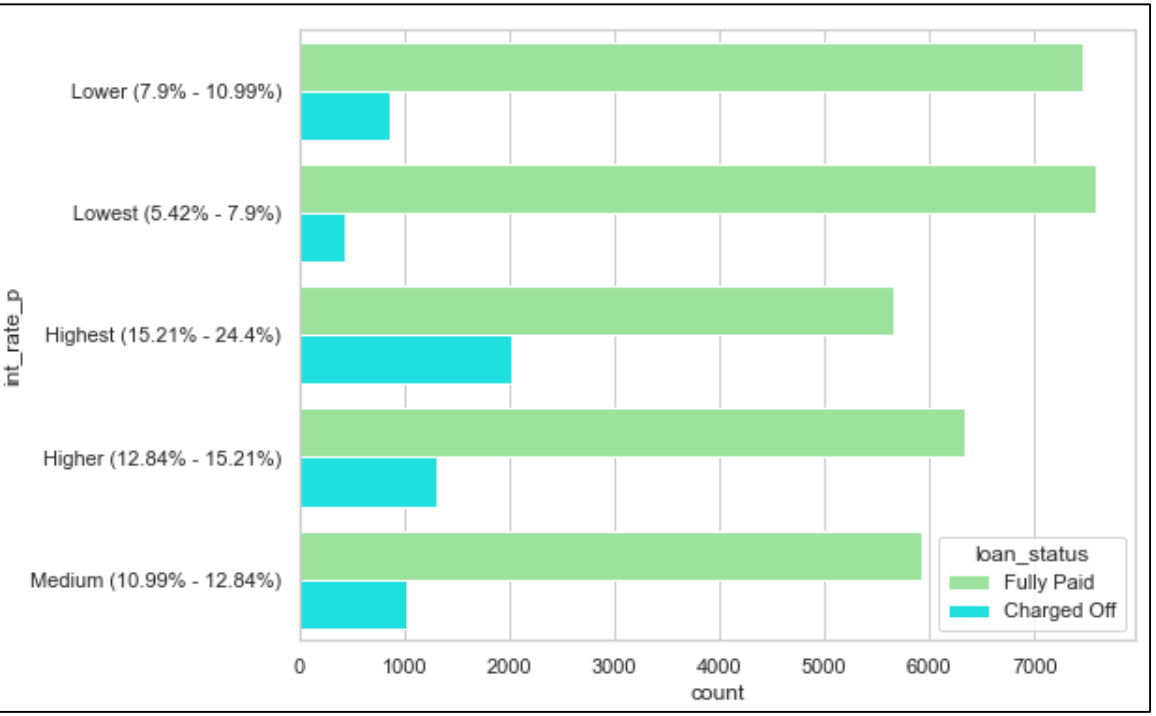
There is no clear correlation between loan_status, loan_amount, there is some outliers in loan amount

since it is giving picture , let us take date data above 75%

From above box plot we are not considering that "higher the amount" as it is insignificant

Univariate Segment Analysis

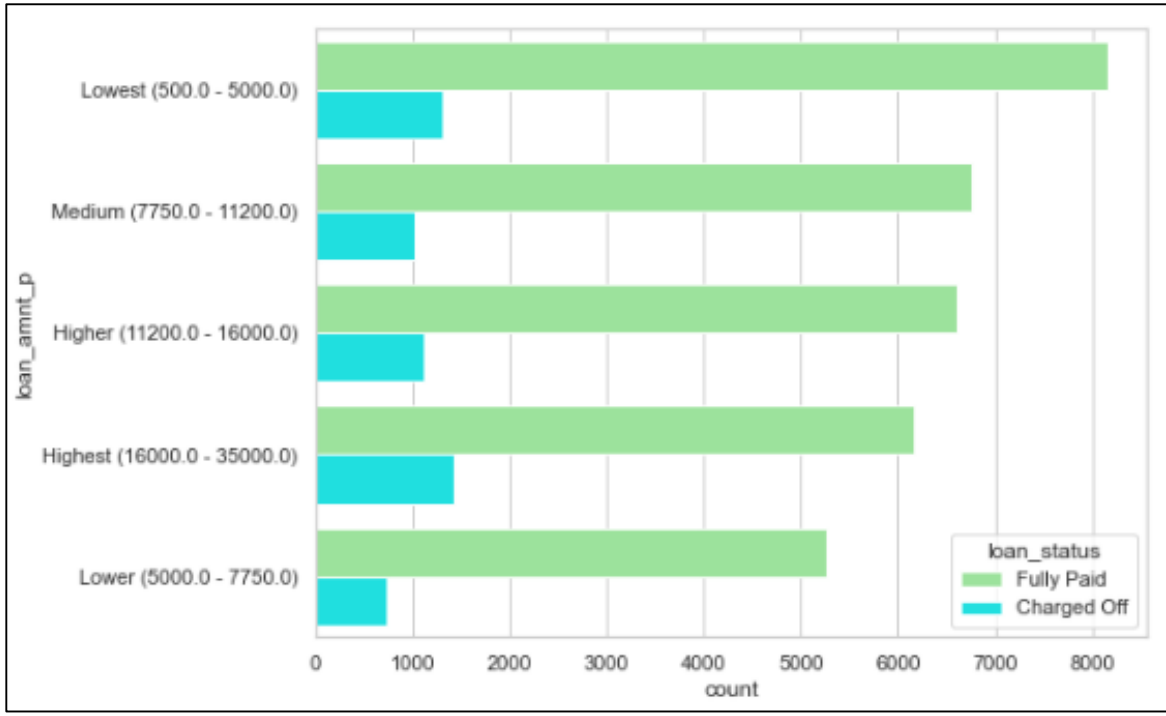
Interest rate and count is Correlation with Loan status



int_rate_p	Charged off %	Record count
Highest (15.21% - 24.4%)	0.262150	7675
Higher (12.84% - 15.21%)	0.171619	7639
Medium (10.99% - 12.84%)	0.146995	6939
Lower (7.9% - 10.99%)	0.102876	8311
Lowest (5.42% - 7.9%)	0.053538	8013

If the interest rate is higher, the chance for 'charged off' is more

Loan amount and count is Correlation with Loan status

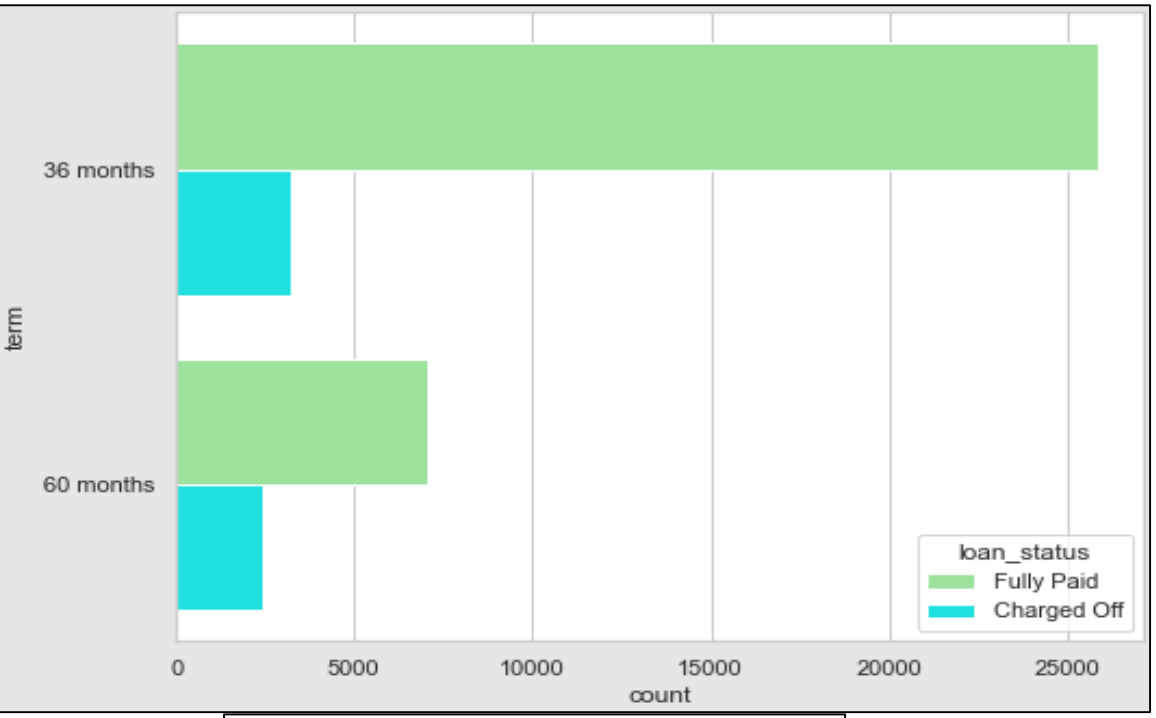


loan_amnt_p	Charged off %	Record count
Highest (16000.0 - 35000.0)	0.187624	7579
Higher (11200.0 - 16000.0)	0.145368	7739
Lowest (500.0 - 5000.0)	0.138725	9472
Medium (7750.0 - 11200.0)	0.131613	7788
Lower (5000.0 - 7750.0)	0.123521	5999

If the loan amount is increasing, there is a high probability of loan getting default.

Univariate Segment Analysis

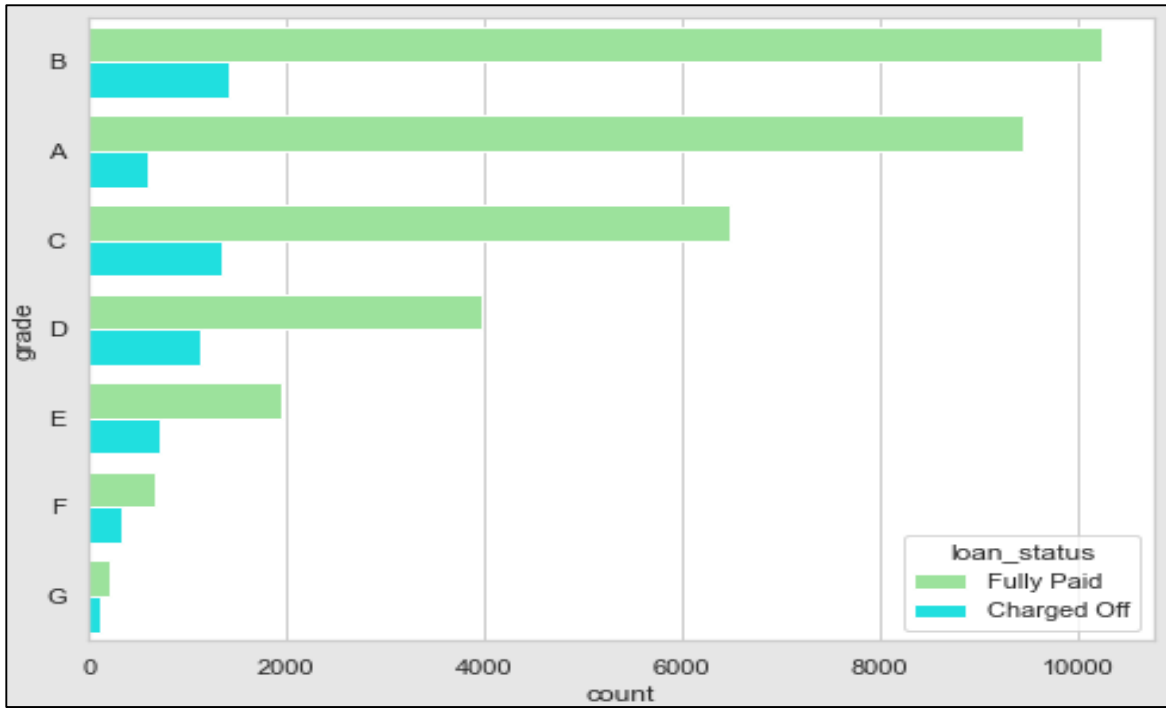
Loan term and count is Correlation with Loan status



term	Charged off %	Record count
60 months	0.253138	9481
36 months	0.110909	29096

For loans with 60 months term the default percent is 25%. and for 36 months term the default is only for 11% of the cases. If higher the repayment term chances of charge off will be higher

Company grade and count is Correlation with Loan status

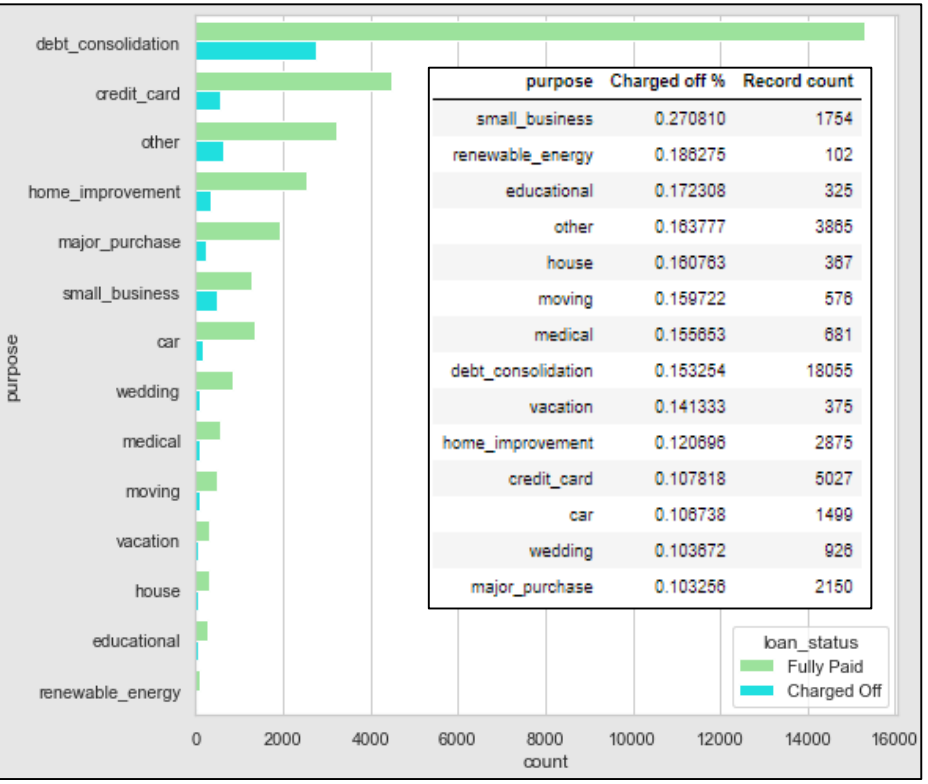


grade	Charged off %	Record count
G	0.337793	299
F	0.326844	976
E	0.268494	2663
D	0.219862	5085
C	0.171943	7834
B	0.122056	11675
A	0.059930	10045

We can clearly see that loan grades having highest default percentages. G, F, E and D form grades where default rate is much higher than others

Univariate Segment Analysis

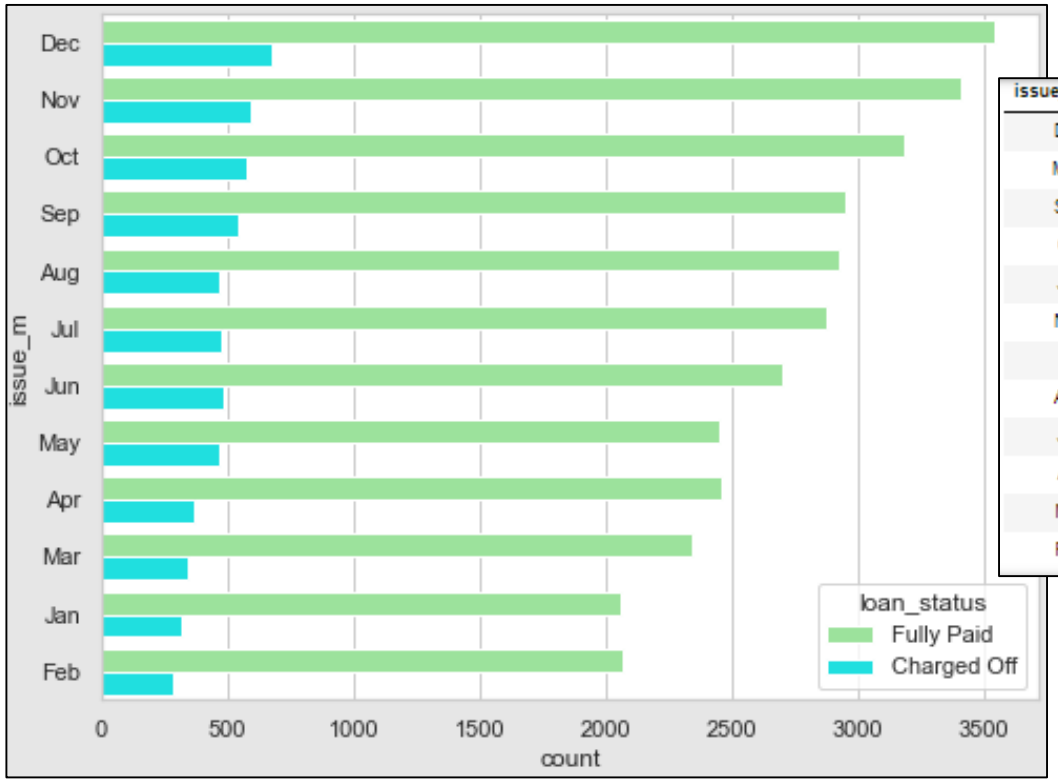
Purpose of loan and count is Correlation with Loan status



purpose	Charged off %	Record count
small_business	0.270810	1754
renewable_energy	0.186275	102
educational	0.172308	325
other	0.163777	3865
house	0.160763	367
moving	0.159722	576
medical	0.155653	681
debt_consolidation	0.153254	18055
vacation	0.141333	375
home_improvement	0.120696	2875
credit_card	0.107818	5027
car	0.106738	1499
wedding	0.103672	926
major_purchase	0.103256	2150

The plot shown loan for small_business, renewable_energy and educational are the riskier ones

Loan issue month and count is Correlation with Loan status

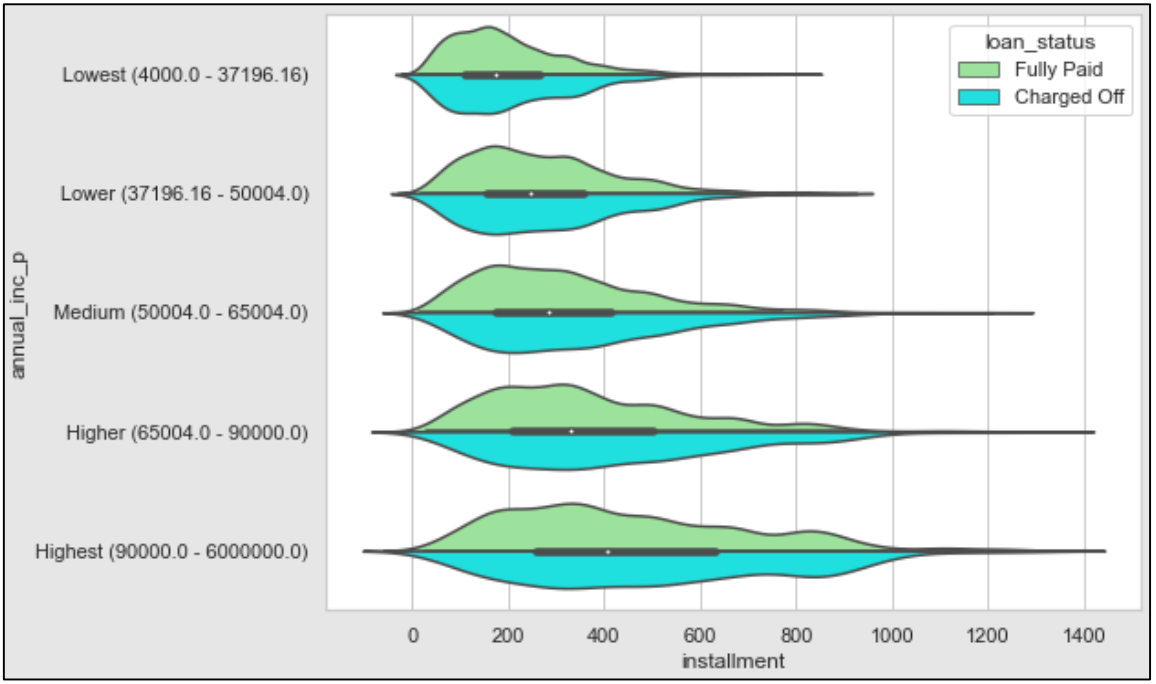


issue_m	Charged off %	Record count
Dec	0.160854	4215
May	0.159644	2919
Sep	0.156375	3498
Oct	0.154214	3761
Jun	0.151887	3180
Nov	0.149276	4006
Jul	0.142942	3351
Aug	0.138135	3388
Jan	0.134931	2379
Apr	0.130696	2831
Mar	0.128948	2691
Feb	0.122986	2358

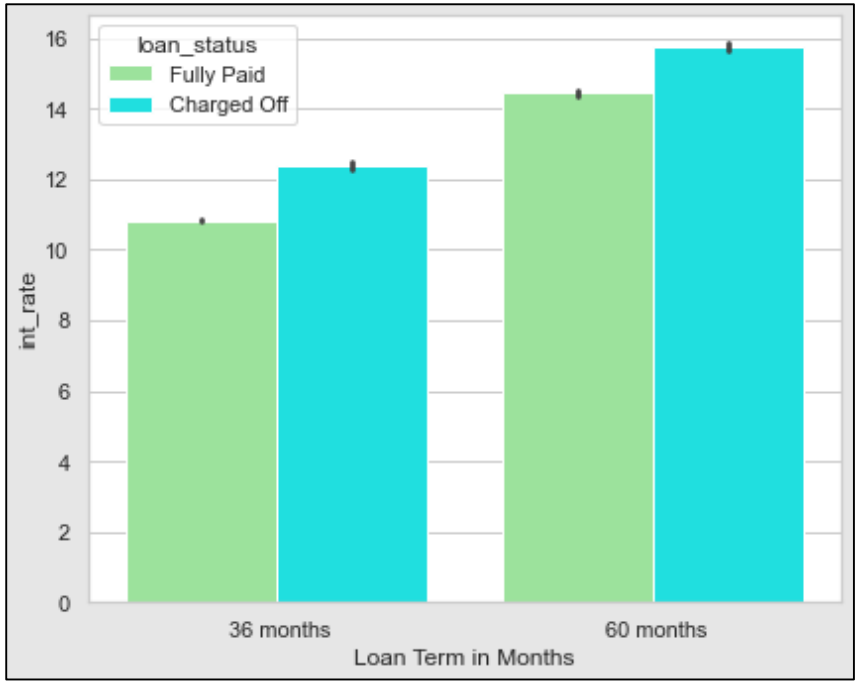
The presence of specific celebrations can be the reason for more charge of 7th & 11th Months

Bivariate Analysis

Annual income and installment Correlation with Loan status



Interest rate and Loan term is Correlation with Loan status

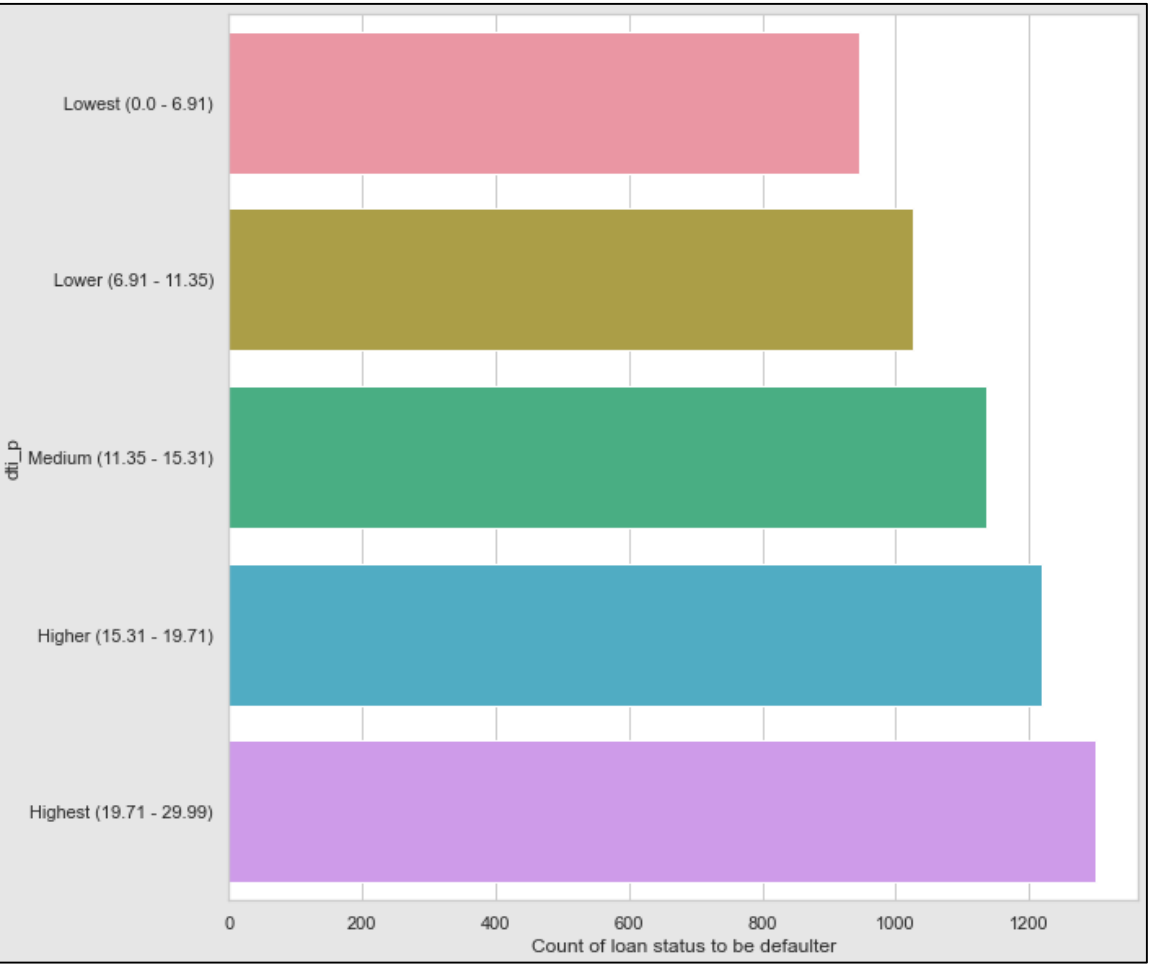


The Graphs is quite revealing the fact that shows that for higher installments for any income group have more number of defaults.

As the interest rate and Loan term increases the chance of charged off is more

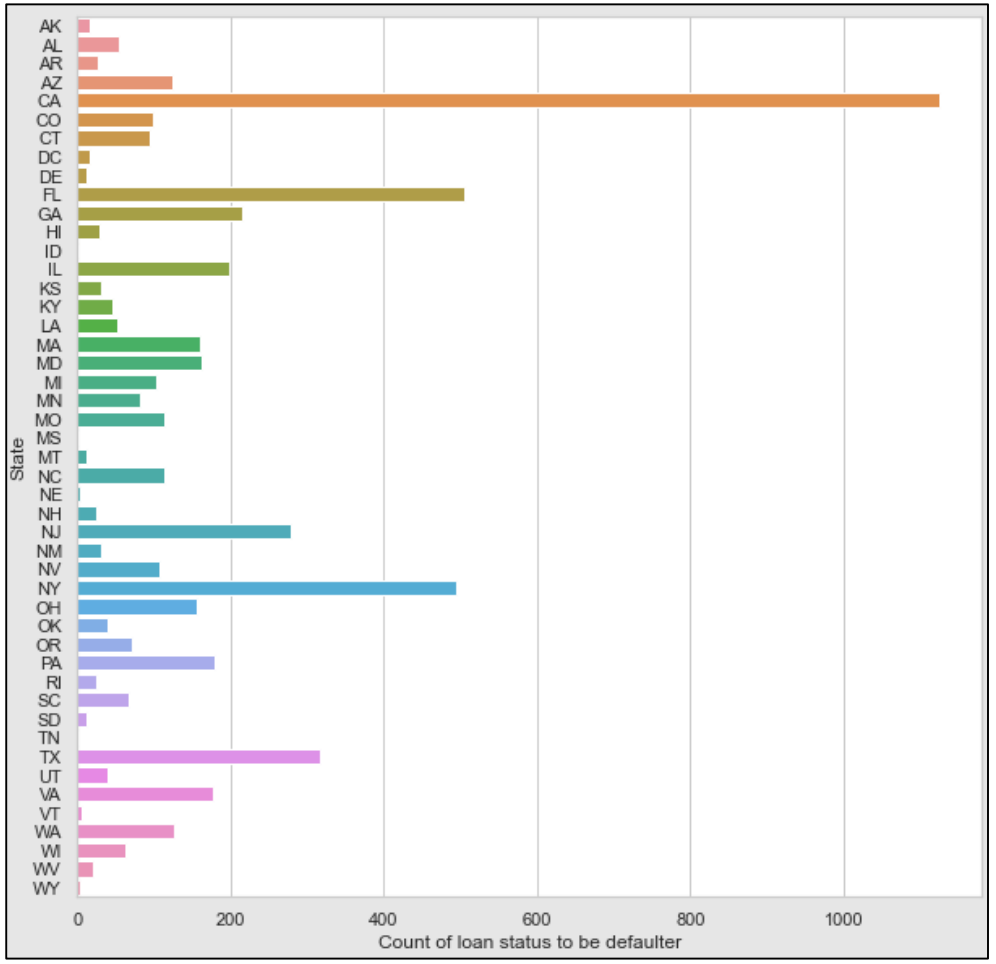
Bivariate Analysis

DTI and Loan status correlations



As the debt increases the chance to be default is getting increased

Interest rate and Loan term is Correlation with Loan status



There is a combined effect in case of the loan charged off VS Resident Area

Thank you