Lending Club Case Study

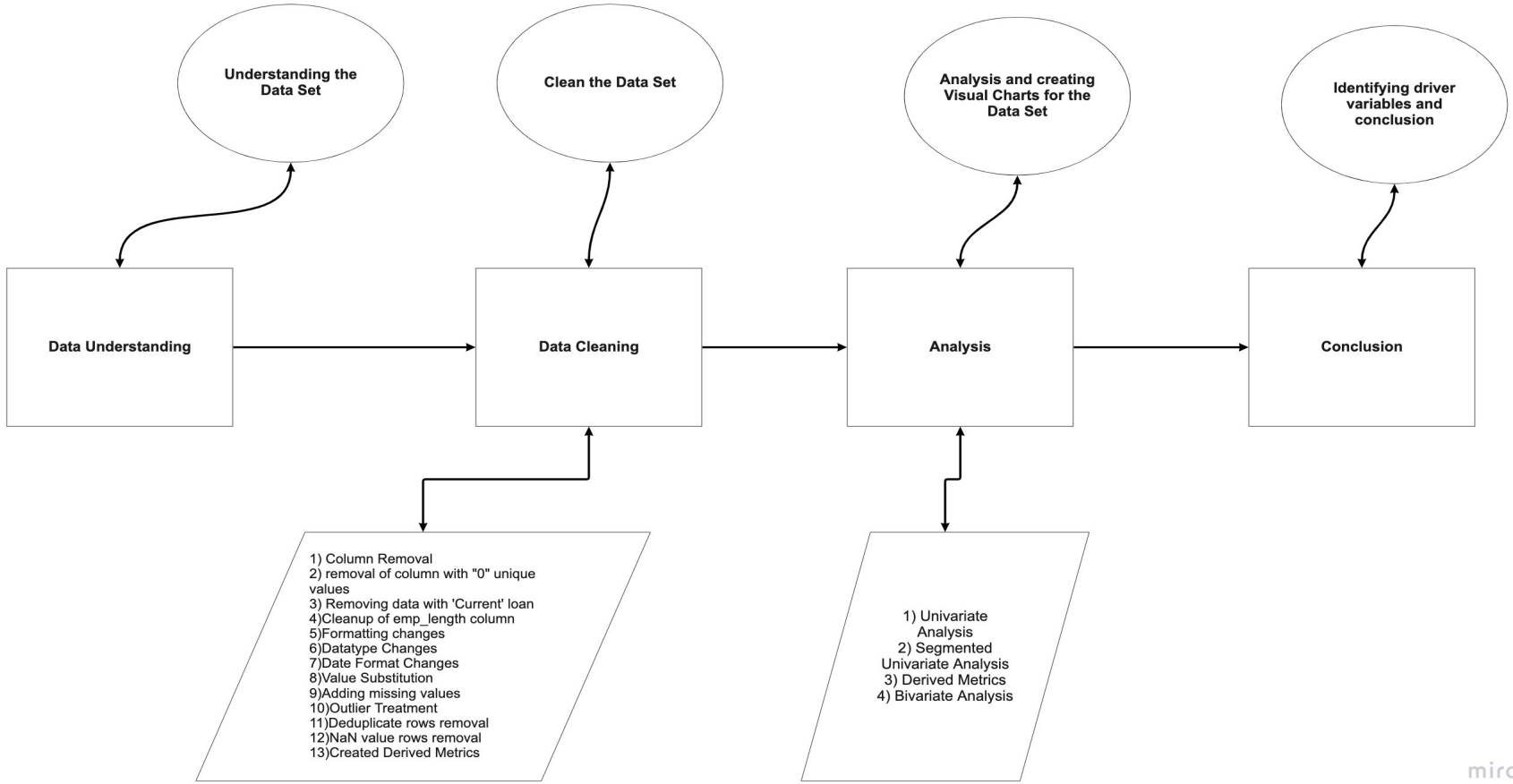
Submitted By: Rohit Sharma

Email ID: sh.rohit91@gmail.com

Objective

- To identify **driving factors** behind loan default, i.e the driver variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- Use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Analysis Approach



Data Understanding

Conclusion

- From the above data, we can see that our dataset has 39717 rows and 111 columns with first column as id and last column as total_il_high_credit_limit. We can see that many columns have 0 entries and NAN values which we will address in Data Cleaning. While doing the describe, the probability of outliers present for columns such as : annual_inc, loan_amount etc, we will analyze it using boxplot and remove the outliers if present.

Data Cleaning

Conclusion

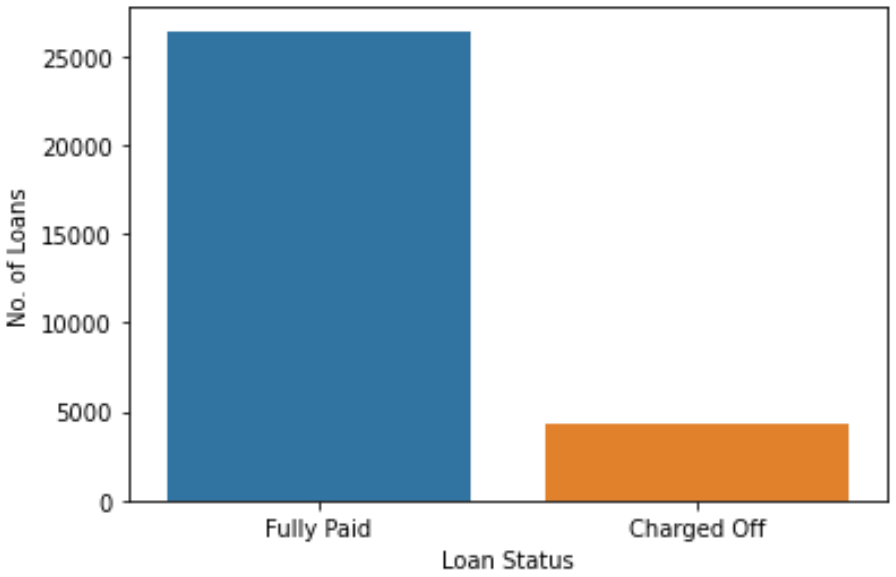
- We performed bunch of data cleaning techniques to clean our data, we used the following methods:
  - Identification and removal of Columns which are not useful
  - removal of column with "0" unique values
  - Removing data with 'Current' loan status
  - Cleanup of emp_length column
  - Formatting changes
  - Datatype Changes
  - Date Format Changes
  - Value Substitution
  - Adding missing values
  - Outlier Treatment using Quartile method
  - Deduplicate rows removal
  - NaN value rows removal
  - Created Derived Metrics

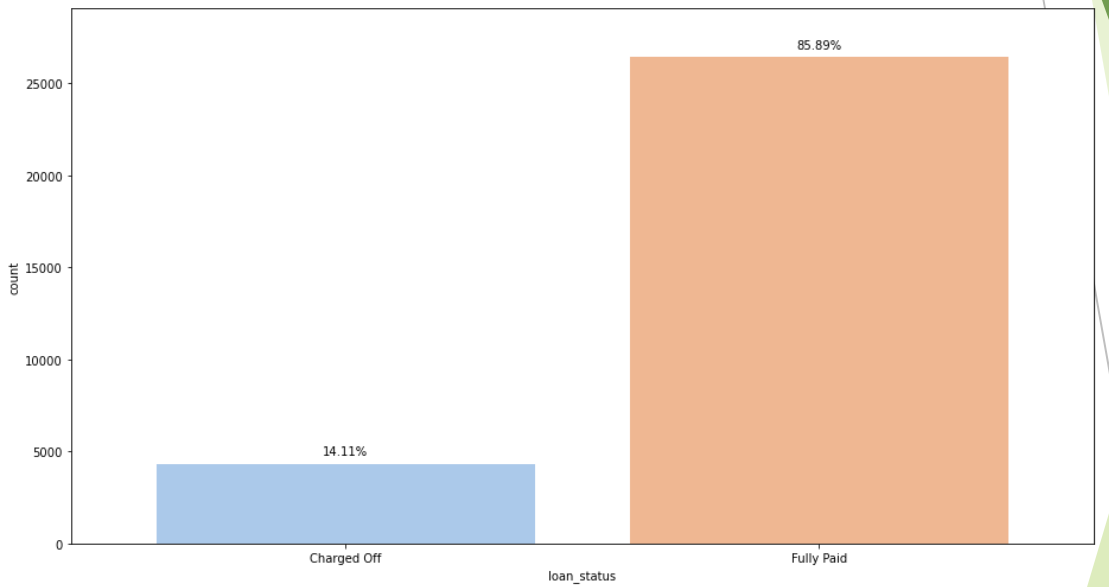
Now with the clean data, we can perform our analysis

Data Analysis

- In the following slides, we will showcase analysis done on multiple factors to determine loan default usecase and also how **consumer attributes** and **loan attributes** influence the tendency of default.
- We will use Univariate Analysis, Segmented Univariate Analysis, Derived Metrics, Bivariate Analysis for our analysis.

Loan Status Column¶

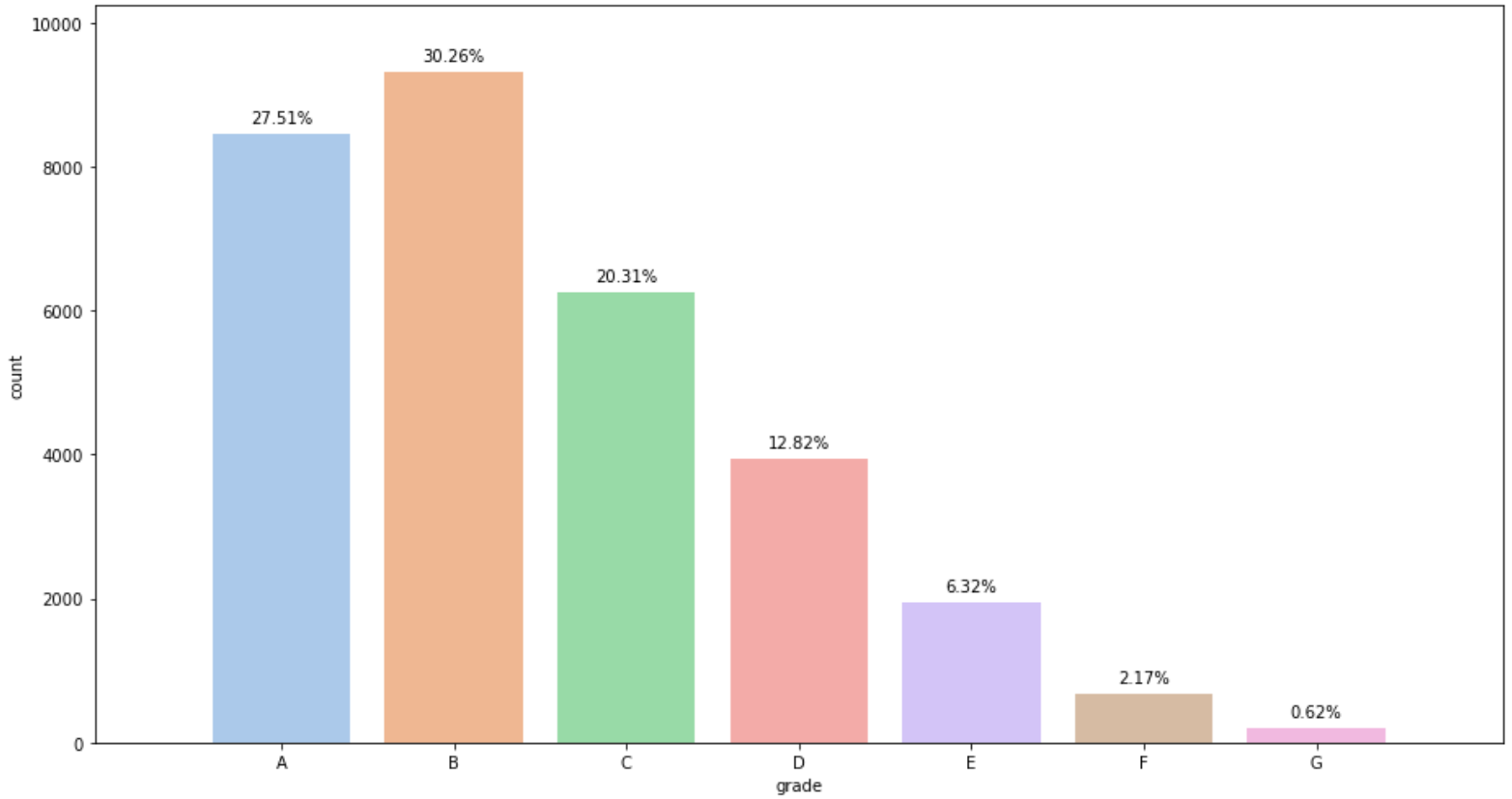




Analysis Result

- From the above analysis we can see more than 20000 loans have been fully paid which corresponds to 85.89% data and 14.11% loans are charged off

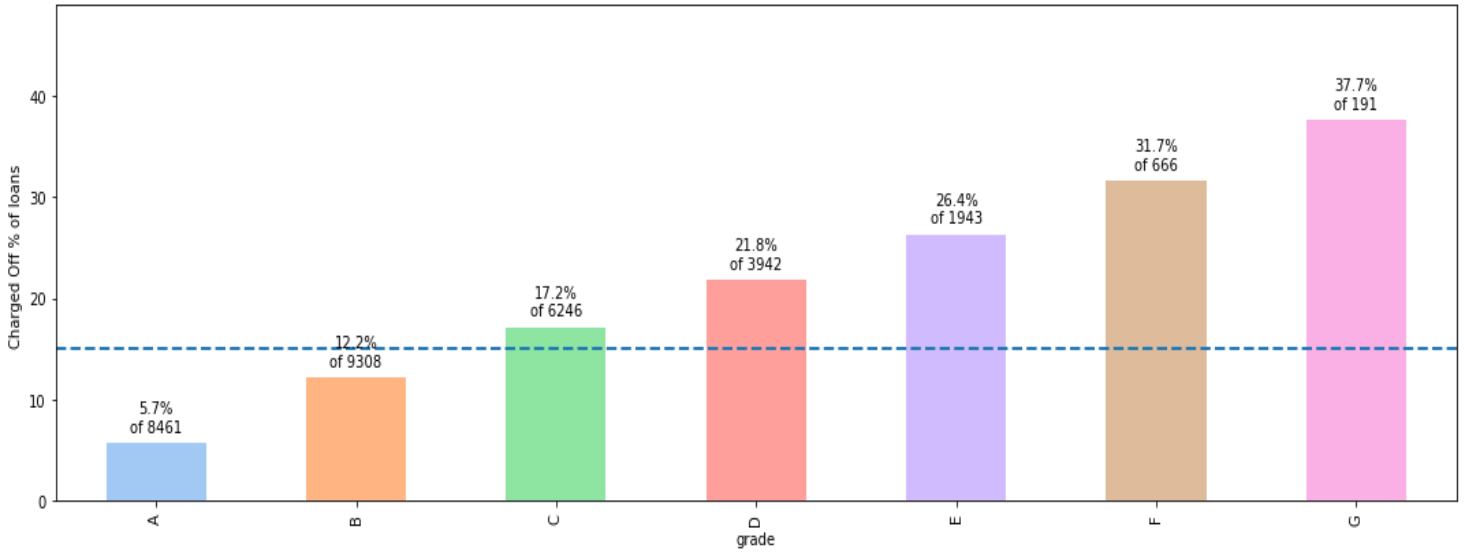
Grade Column

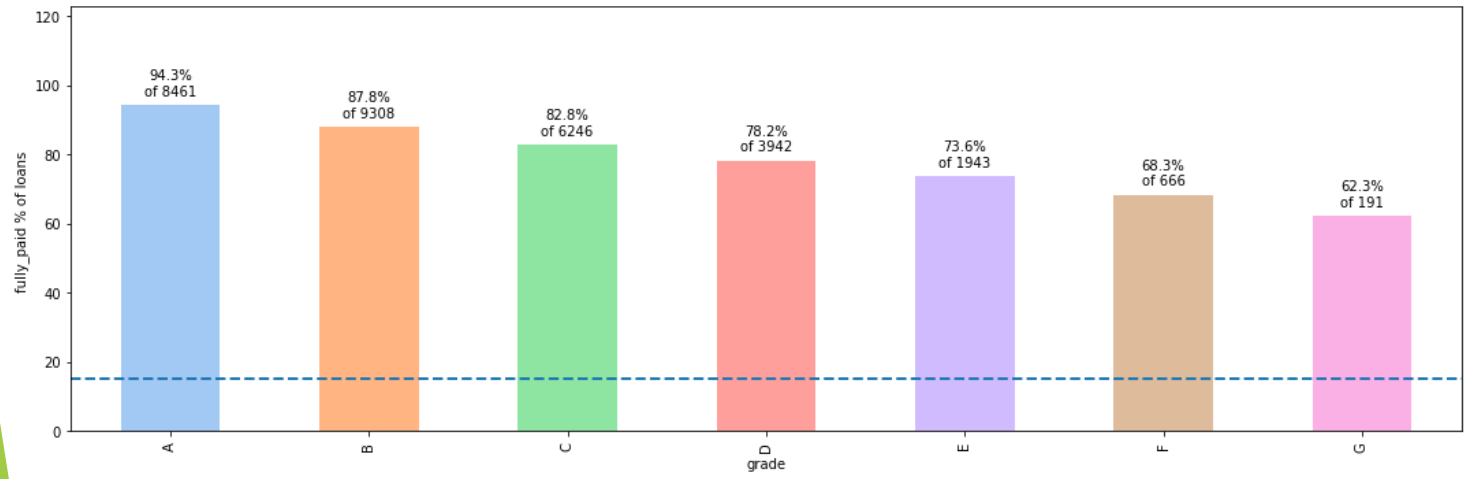


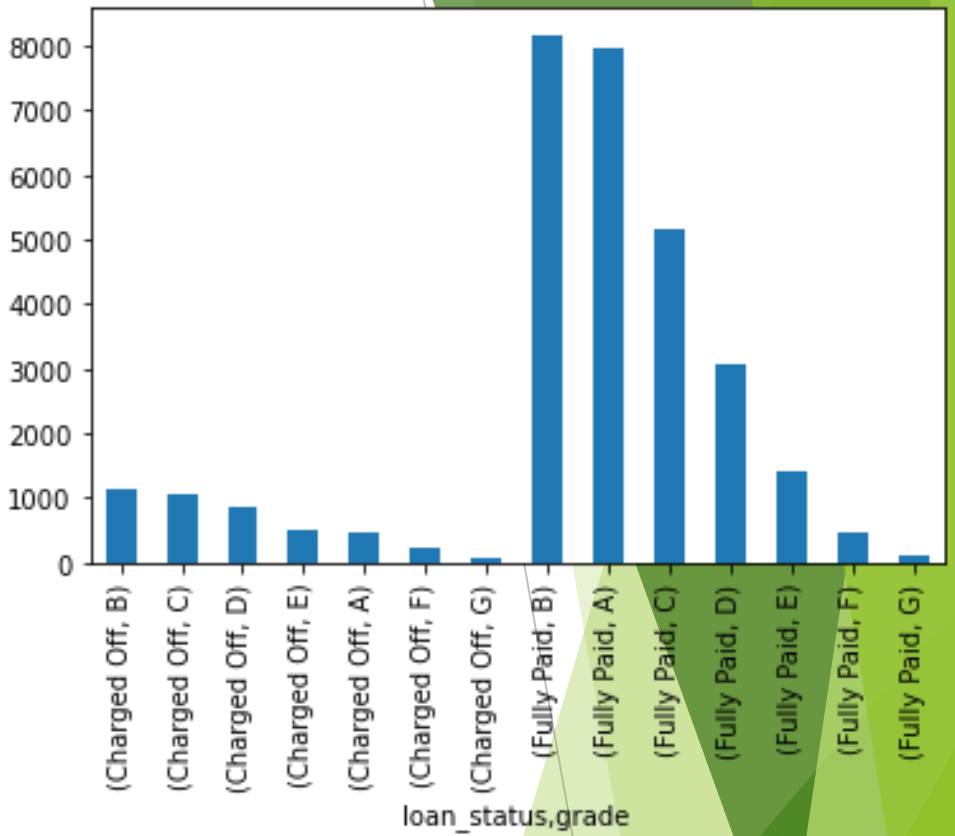
Analysis Result

- From the above, as we move to lower grade, the count decreases but it would be great to see 'grade' with the loan status

Loan_status vs Grade Column



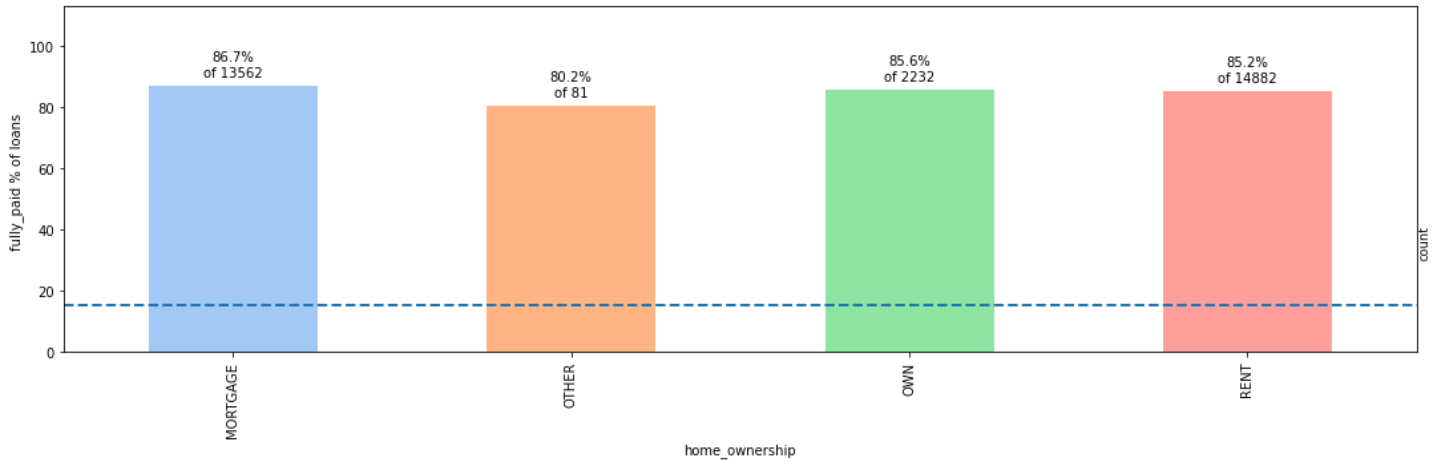


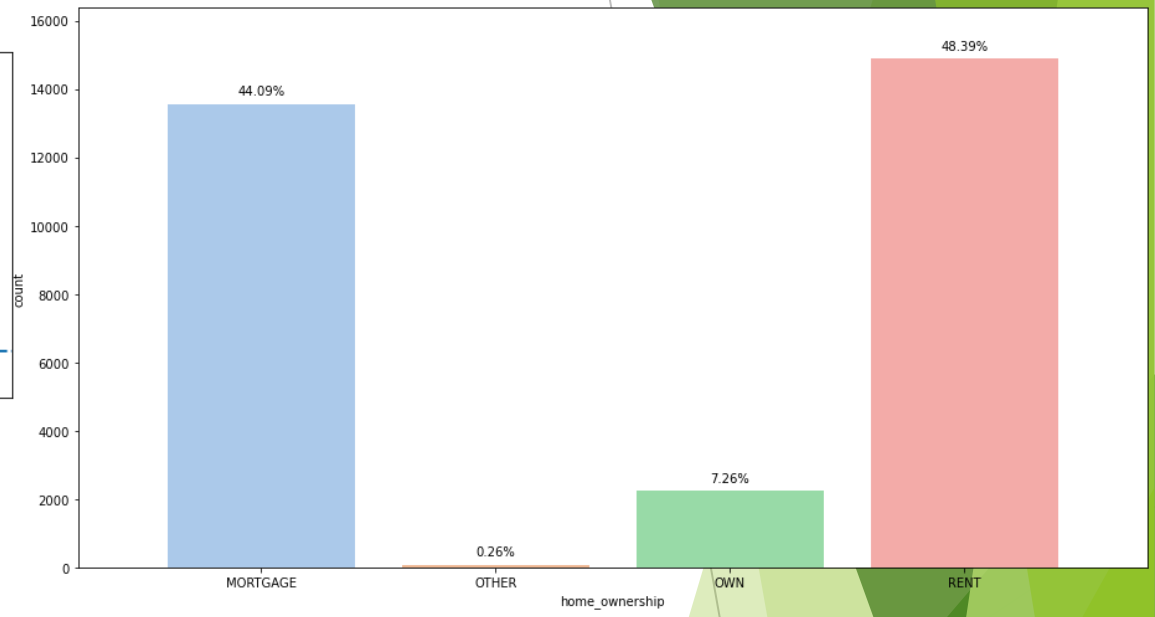


Analysis Result

Chances of defaulting increases with grade from A to G and subgrade from A1-A5, B1-B5, & so forth

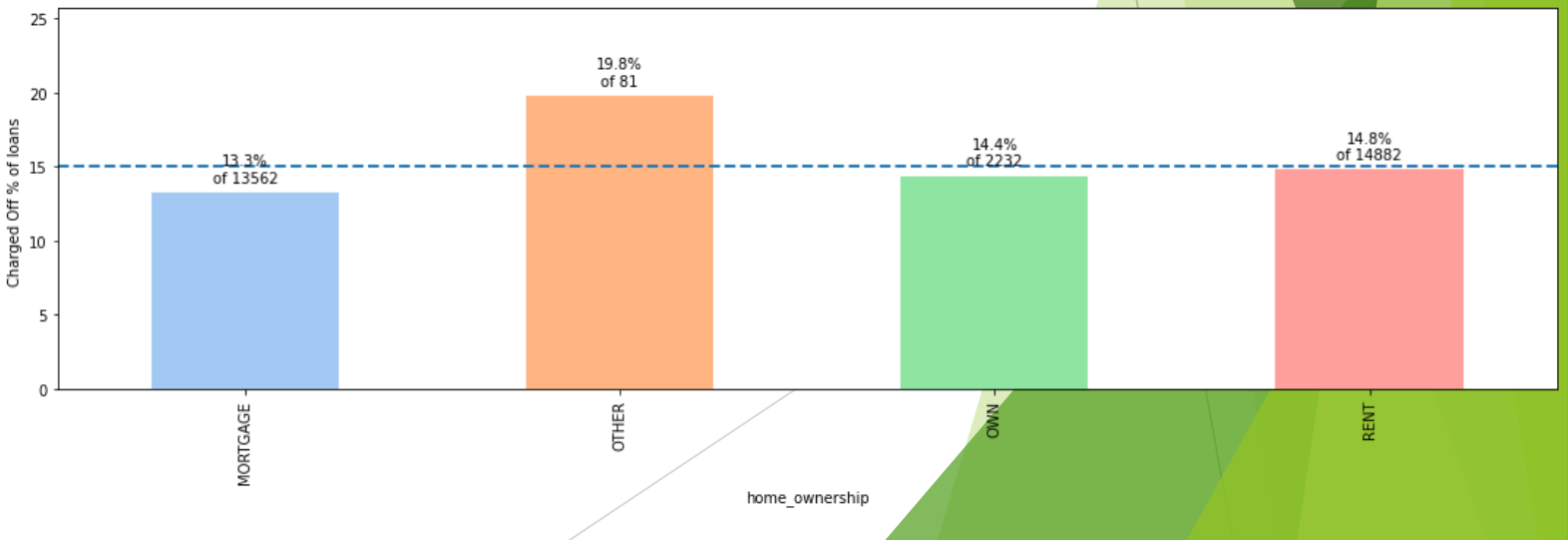
Home Ownership Column



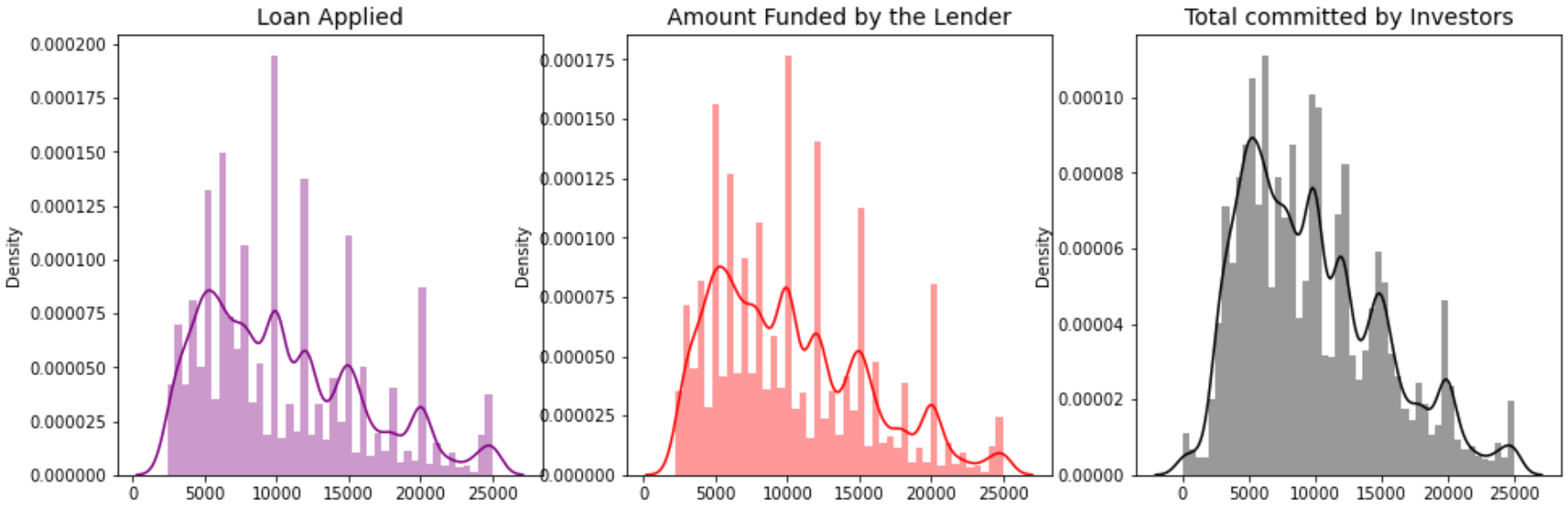


Analysis Result

- People with ownership type as Mortgage and Rent are the most who take loans.
- The default rate is high for people having home ownership type "other".
- The default rate is not significantly impacted by the home ownership status - Mortgage/Owned/Rent



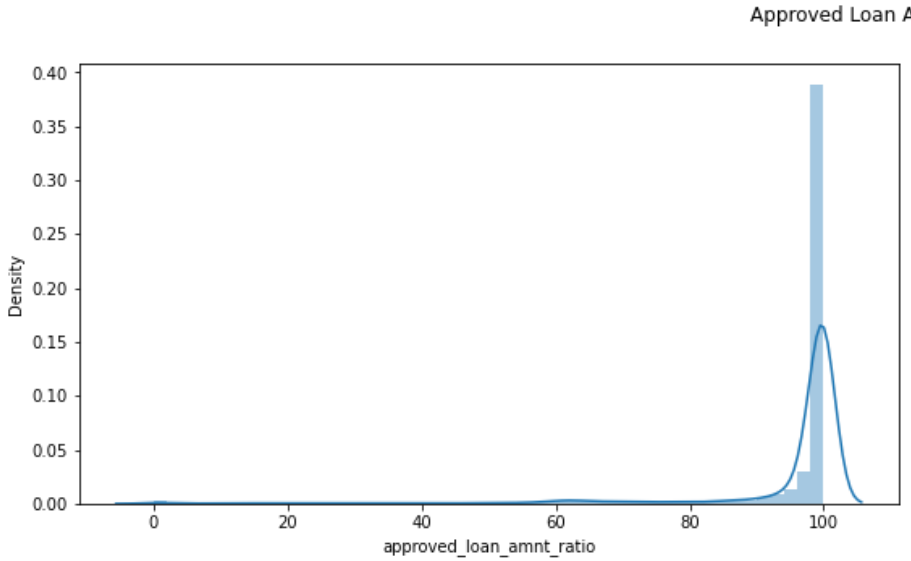
Loan Amount or Funded Amount? Which Column to use

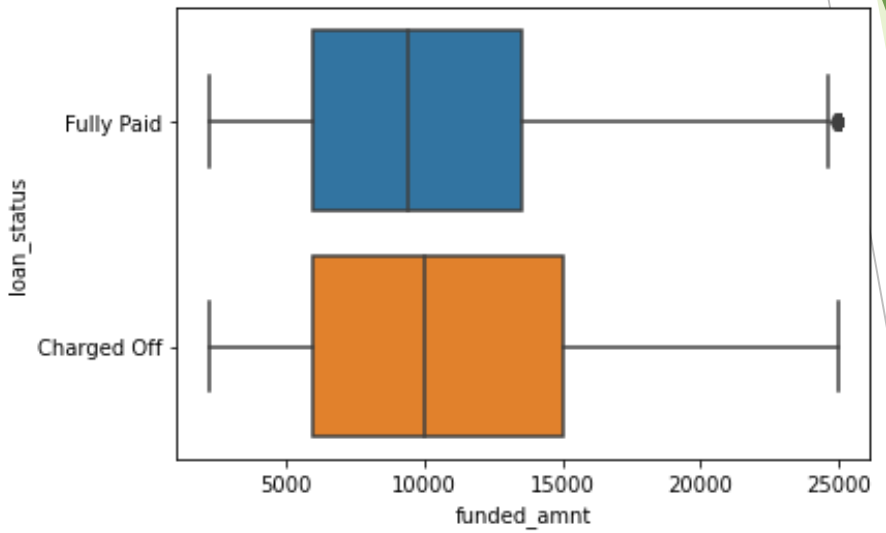


Analysis Result

We can see a spike in all the graphs between 5000 and 10000 for each figure, also there seems to be significant difference of rows between loan applied by the customer and the amount actually approved by the lender

funded_amnt Column¶

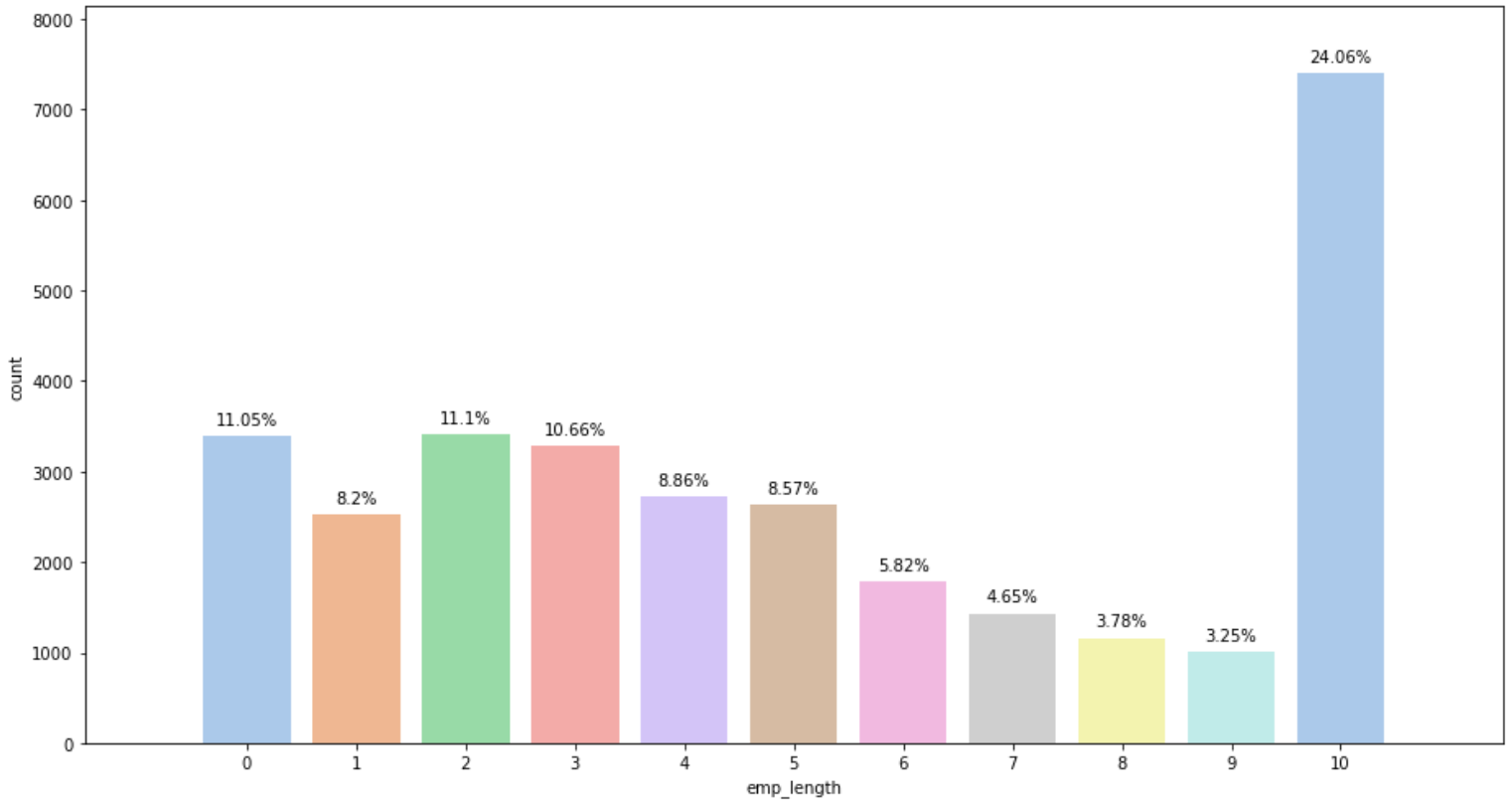




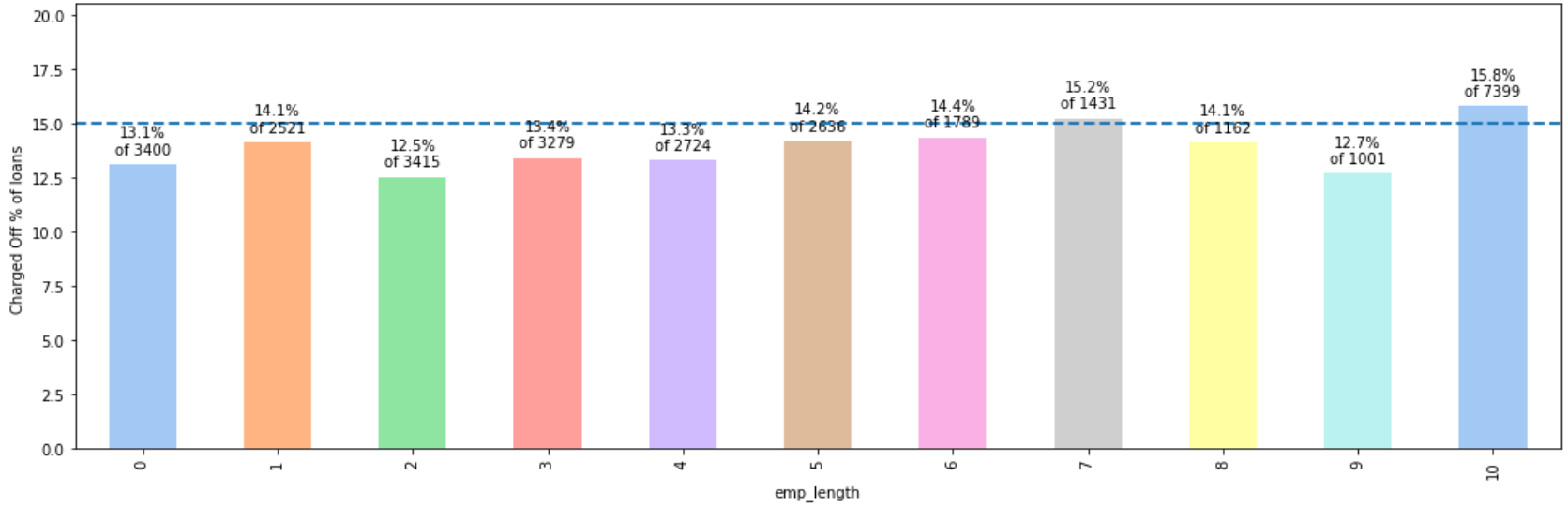
Analysis Result

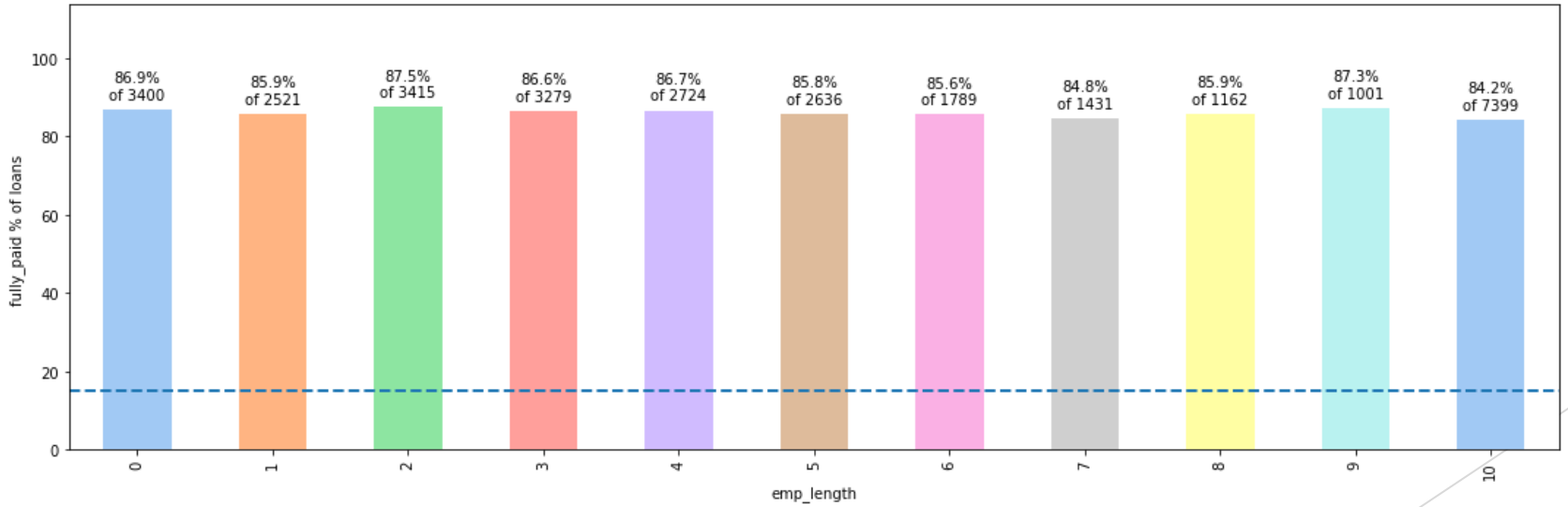
- We will be using funded_amnt for all the future investigation, as this is the actual amount approved by the investors
- Charged off loans have higher average of funded amount than the paid off loans.

emp_length Column



Contd…

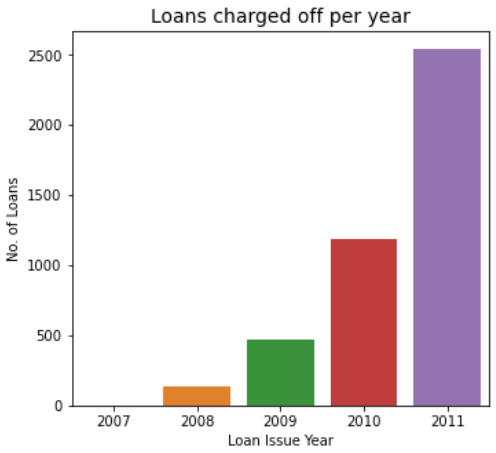


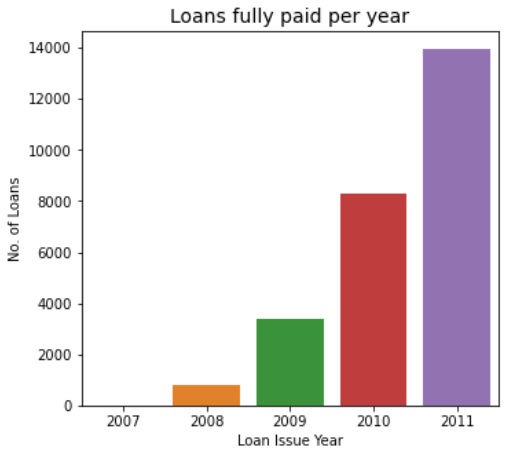


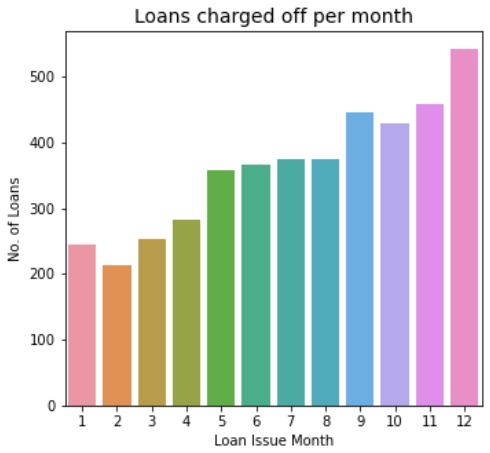
Analysis Result

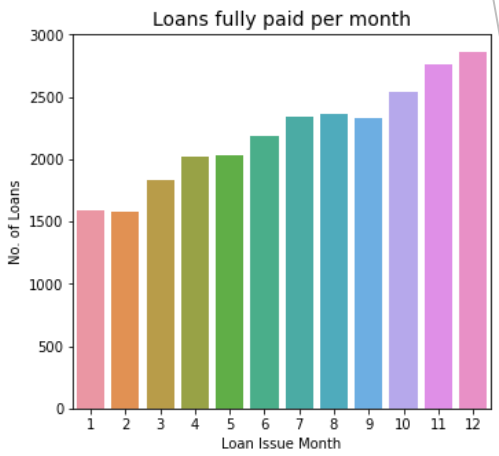
- Mostly 10 years and above employees apply for loan
- 10+ years employees are more likely to default on loan with 15.8%

issue_d_year, issue_d_month Columns





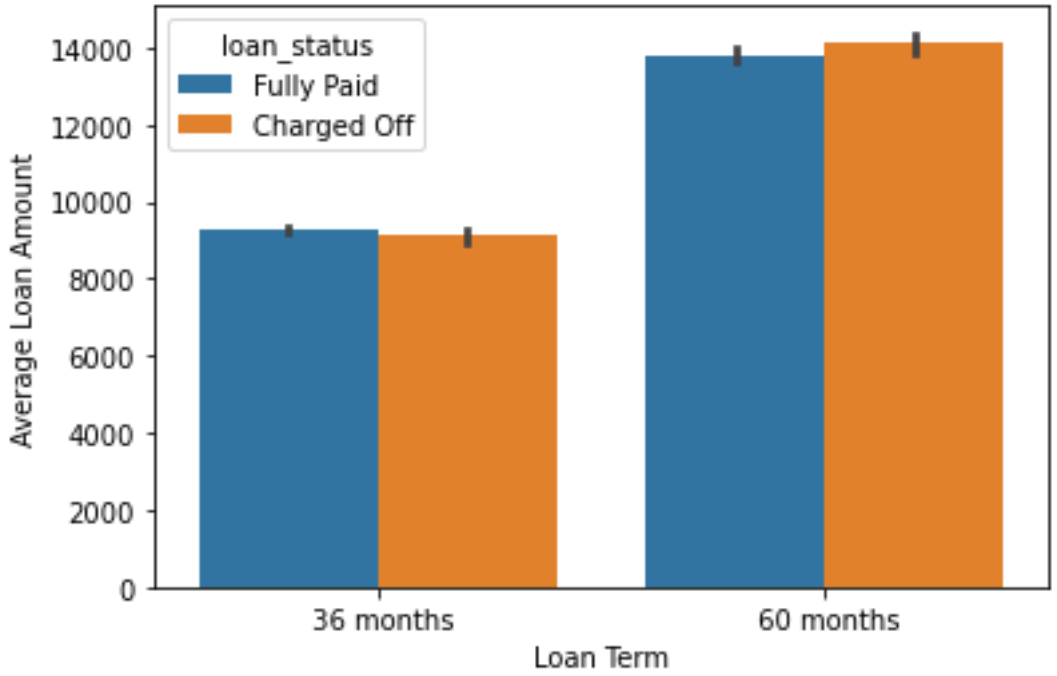




Analysis Result

- No. of loans, fully paid and charged off are increasing every year. They are at maximum in the year 2011. This is a very positive trend for Lending Club as the requirement of loans are increasing by each year.
- The month-wise trend shows that most loans are fully paid as well as charged off as the year comes to an end, maximum in the month of December clearly stating the importance of year-end.

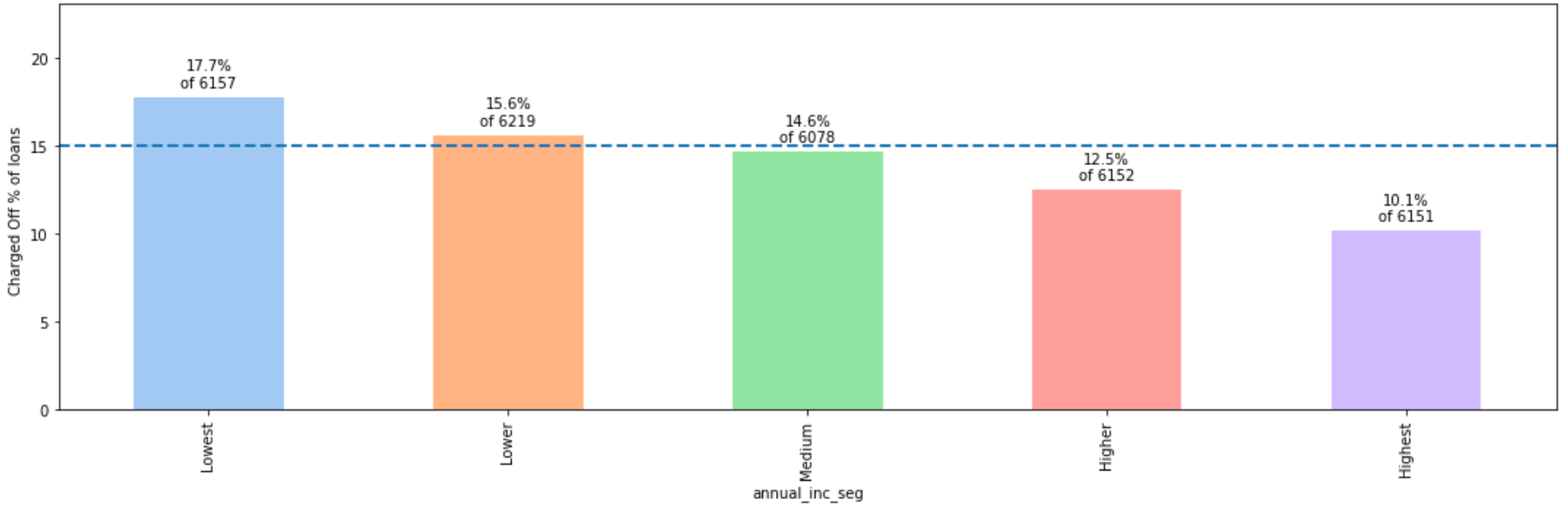
Term Column Analysis vs loan_amount vs loan_status

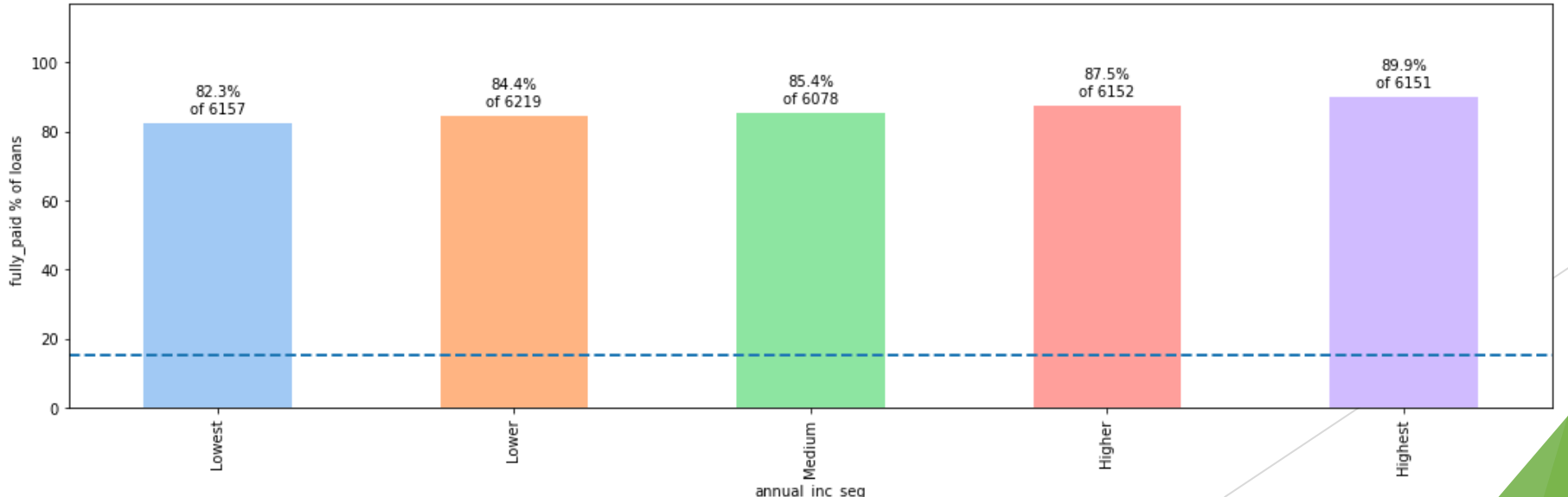


Analysis Result

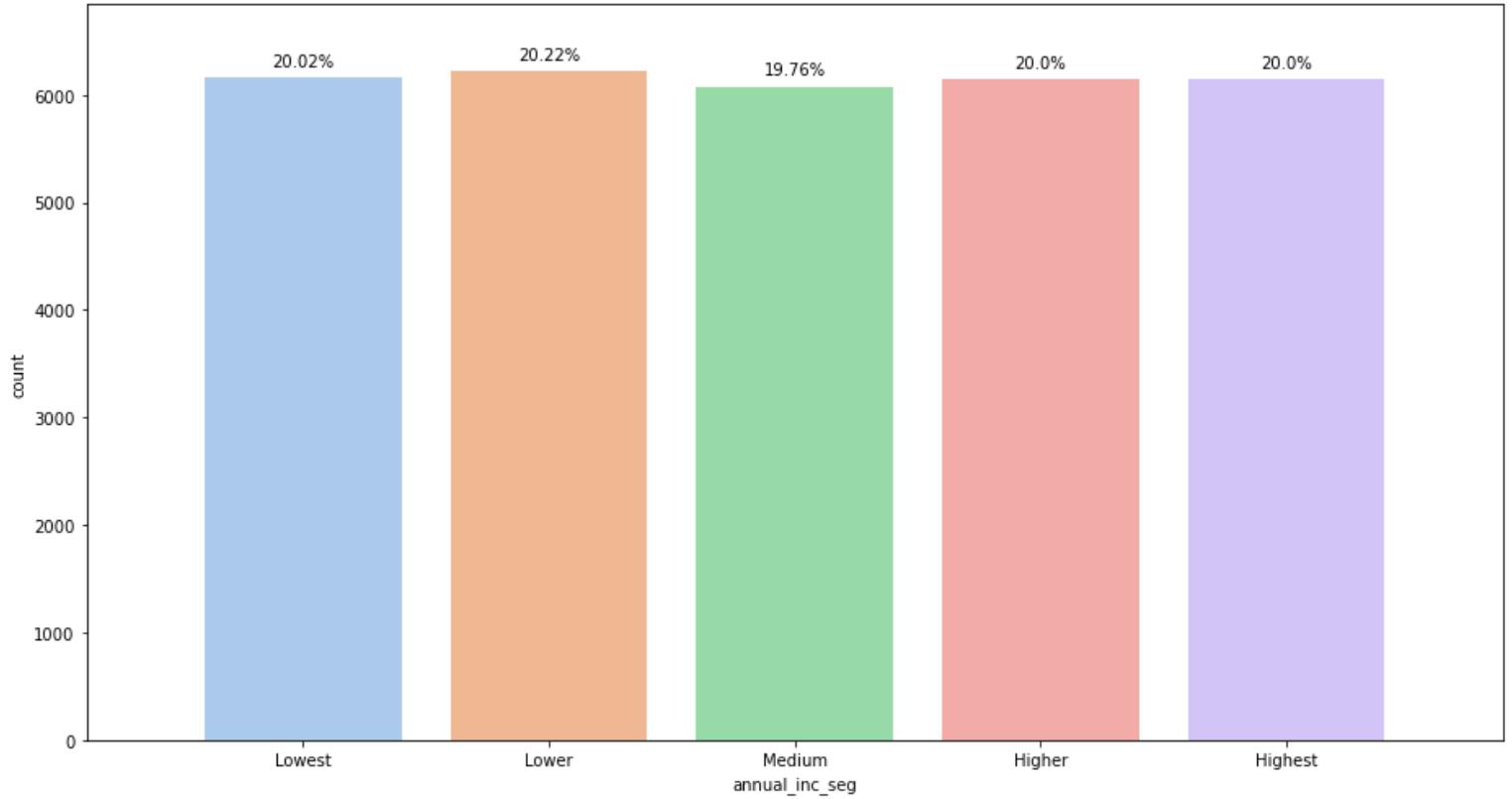
- The avg loan amt is almost same
- The avg loan amt is slightly greater for the 60 month term.
- Charge Off Percentage is higher for 60 months category

Annual Income





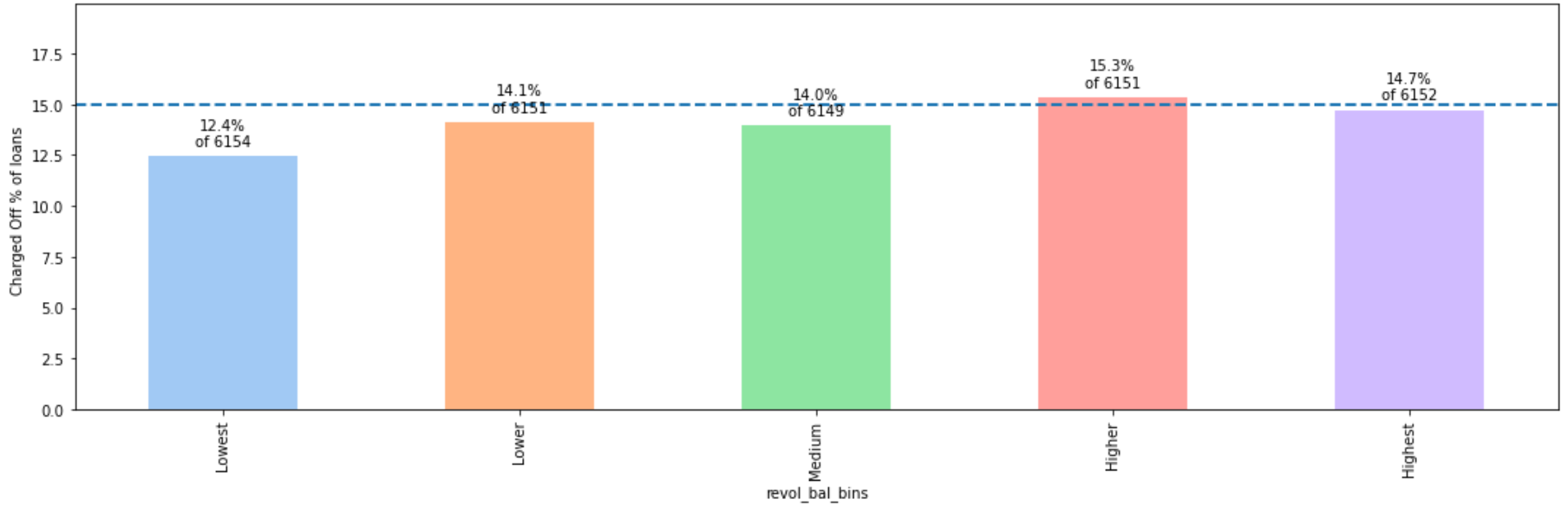
Contd…

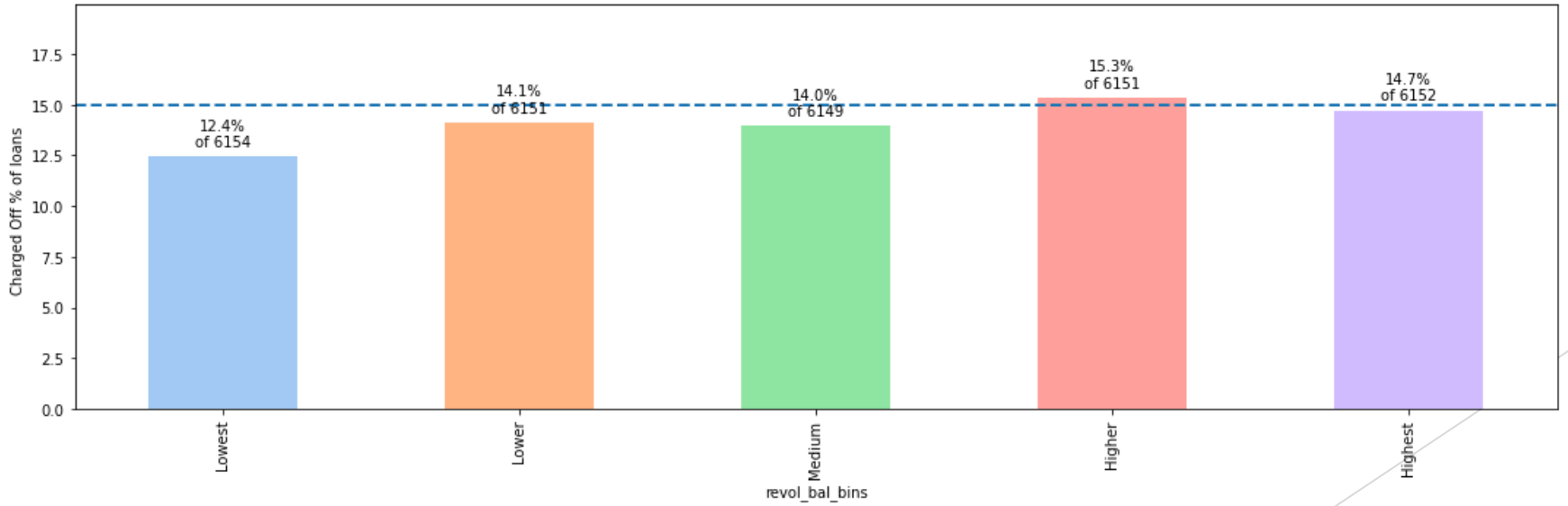


Analysis Result

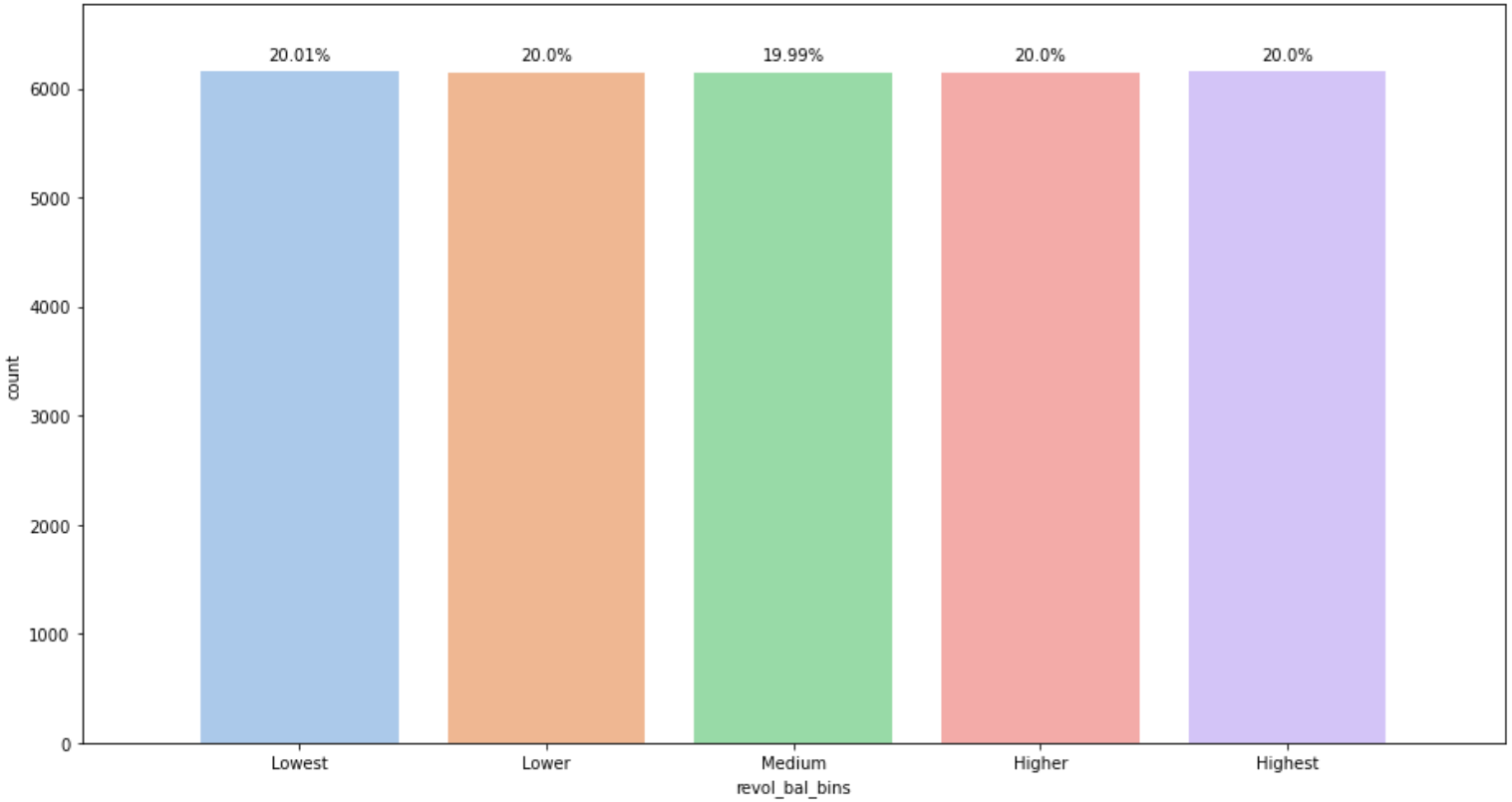
Approx 17.7% for lowest and 15.6% for lower income group are in the charged off category, implying they are more susceptible to charge off

revol_ban column analysis¶





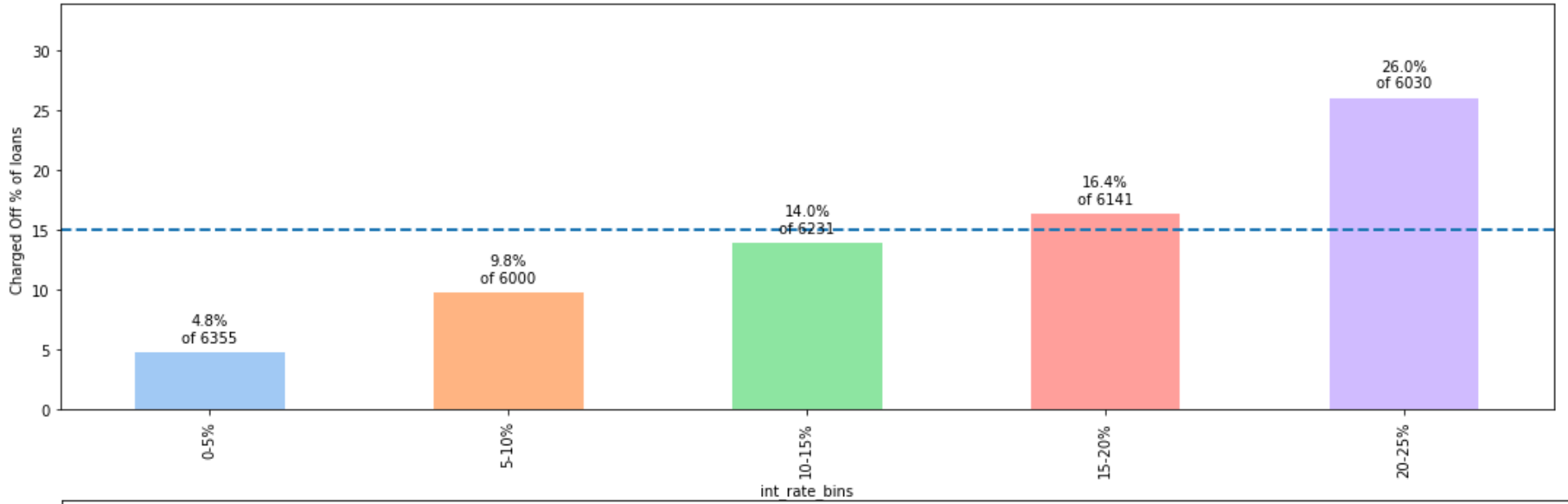
Contd…

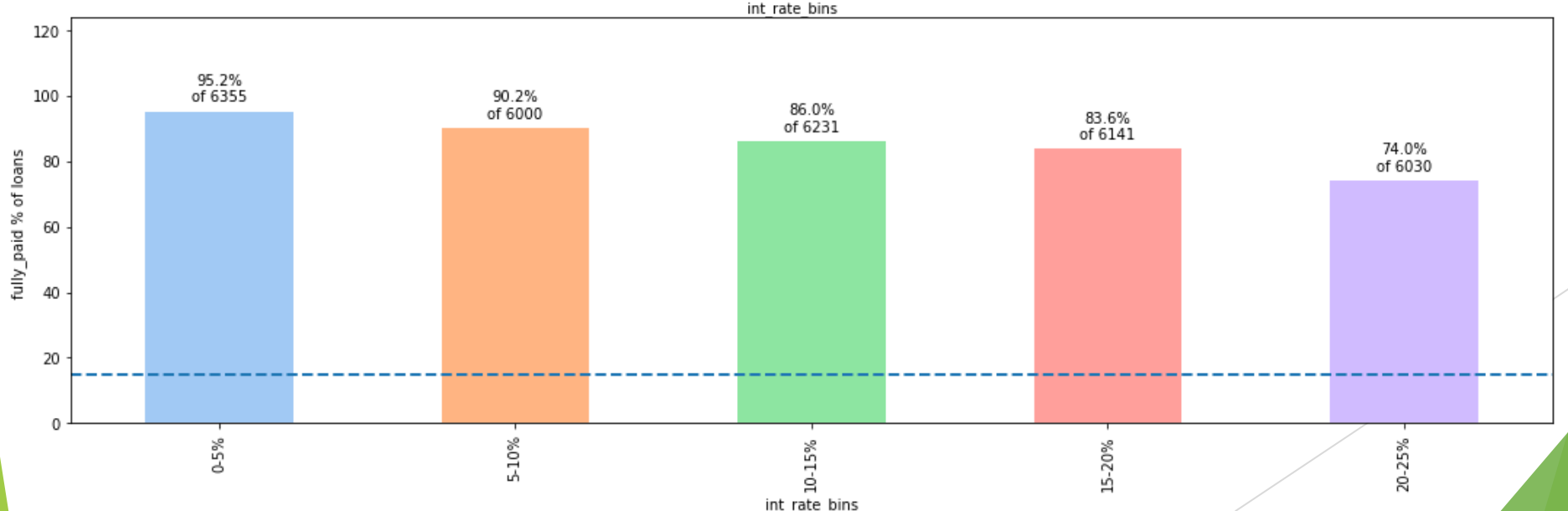


Analysis Result

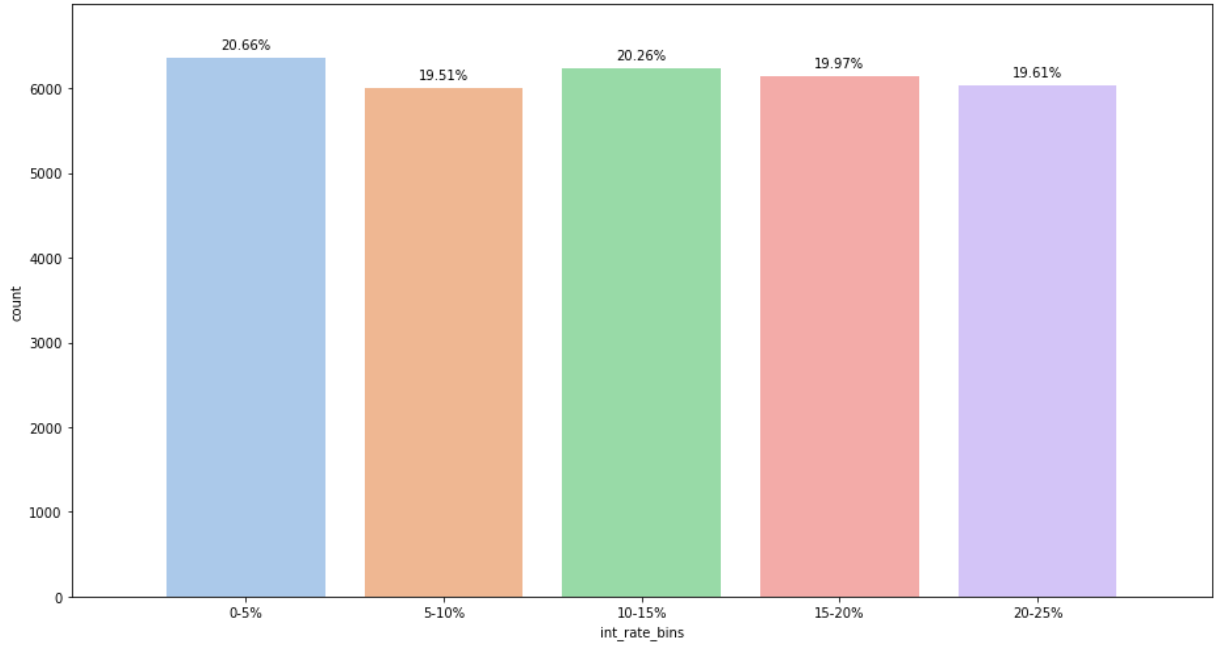
Higher the revolving balance, more is the probability of being charged off

int_rate columns





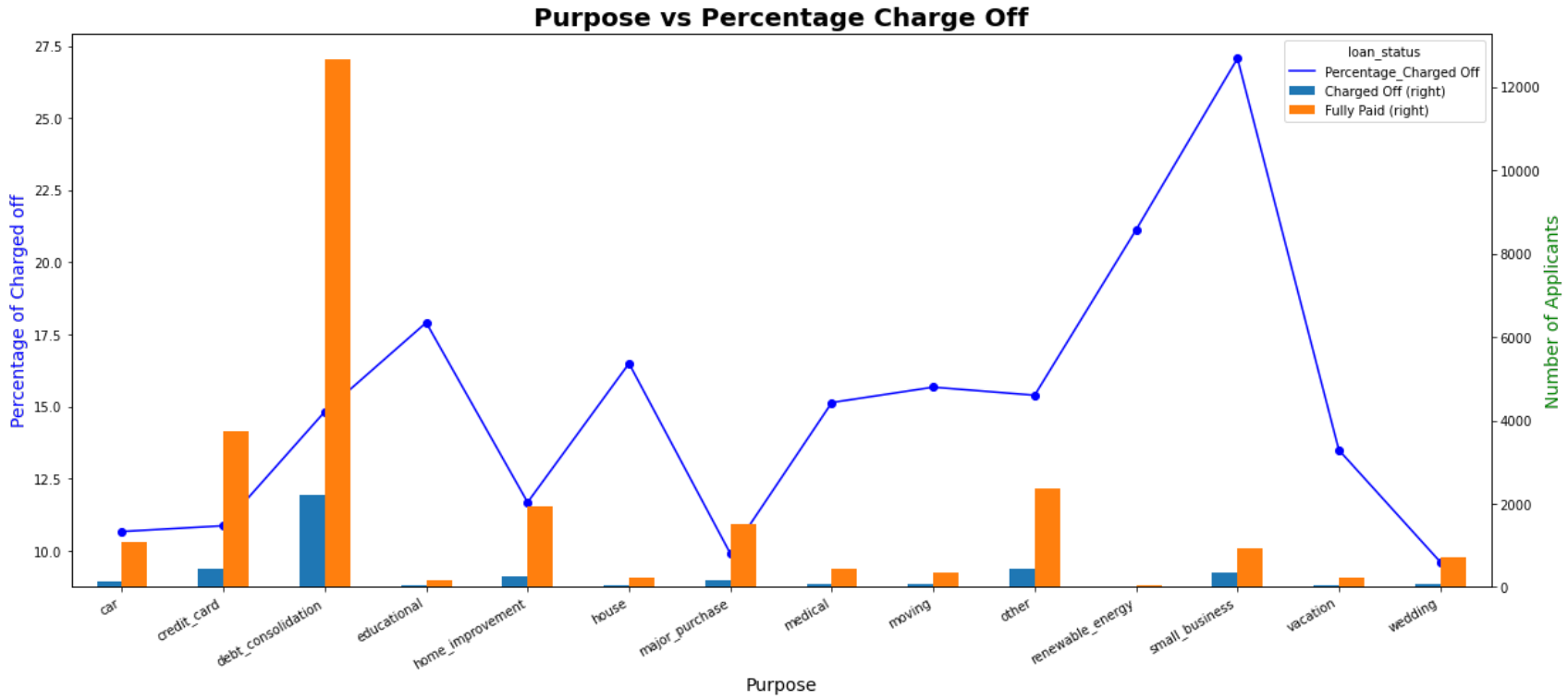
Contd…



Analysis Result

For a high percentage loans, they are more susceptible to default, loan with int rate as 20.25% had default rate of 26%

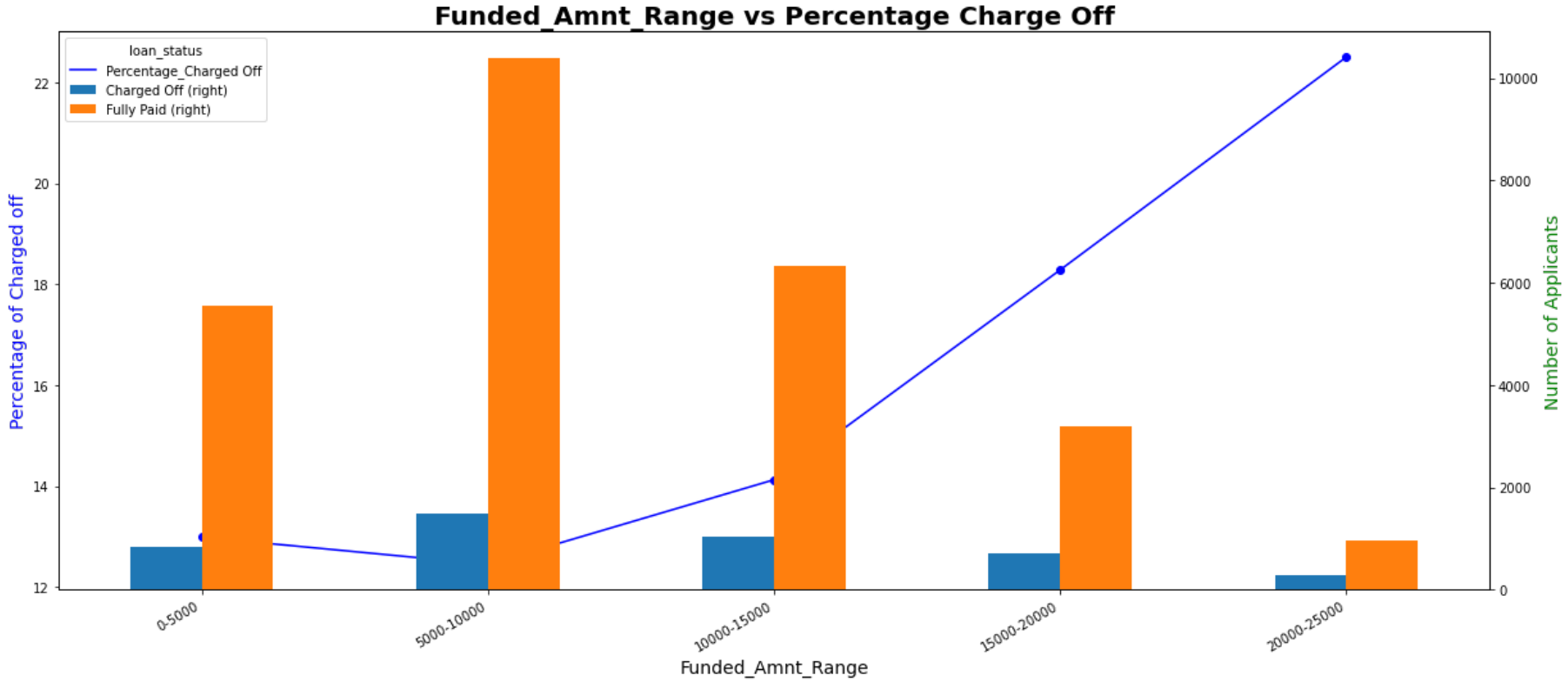
Purpose w.r.t % of loans charged off and no. of loan applicants



Analysis Result

- Max loans i.e. 14861 are accepted for the purpose of Debt Consolidation out of which 12658 are fully paid and 2203 are charged off i.e. 14.823%
- Maximum % of a loan being charged off is for the purpose of small business and the % is 27.067

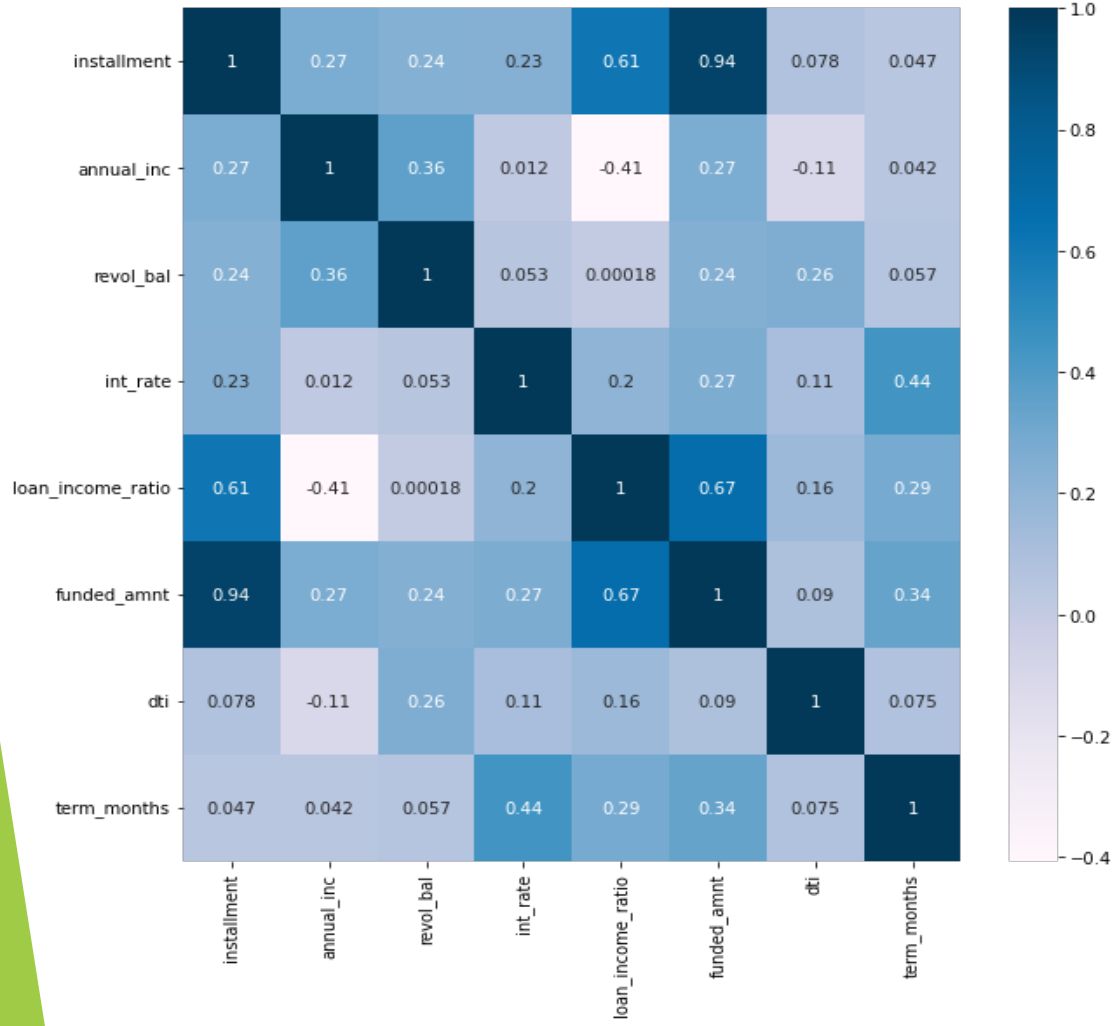
Analysis of funded_amnt vs Loan Applications



Analysis Result

- As the funded_amnt_range increases, The probablity of loan being charged off increases. It is maximum, 22.506% for the amount of 20000-25000
- The maximum no. of loans, 11864 are approved where the funded_amnt_range ranges between 5000-10000

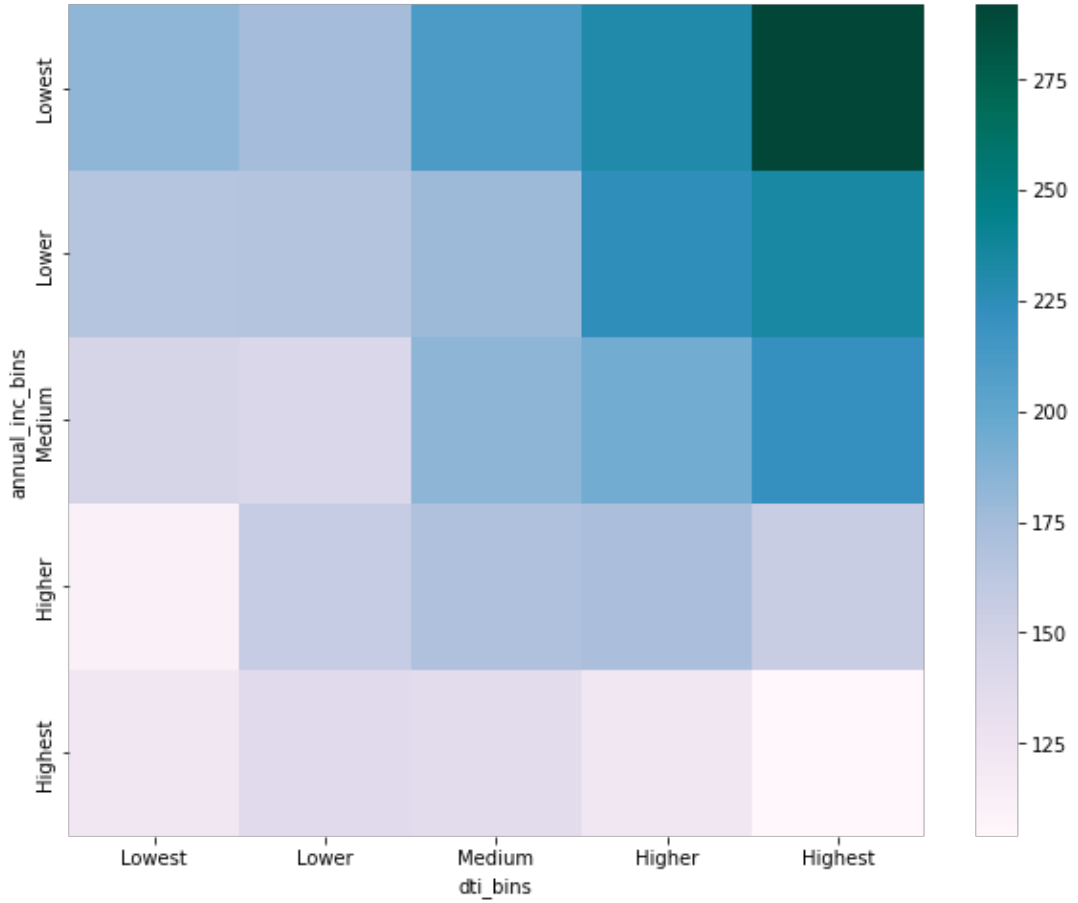
Correlation metrics - Bivariate on imp continous columns



Analysis Result

- Funded amount,installment have very high correlation with the loan income ratio.
- Term months and interest rate are positively correlated
- Annual income is negatively correlated to dti

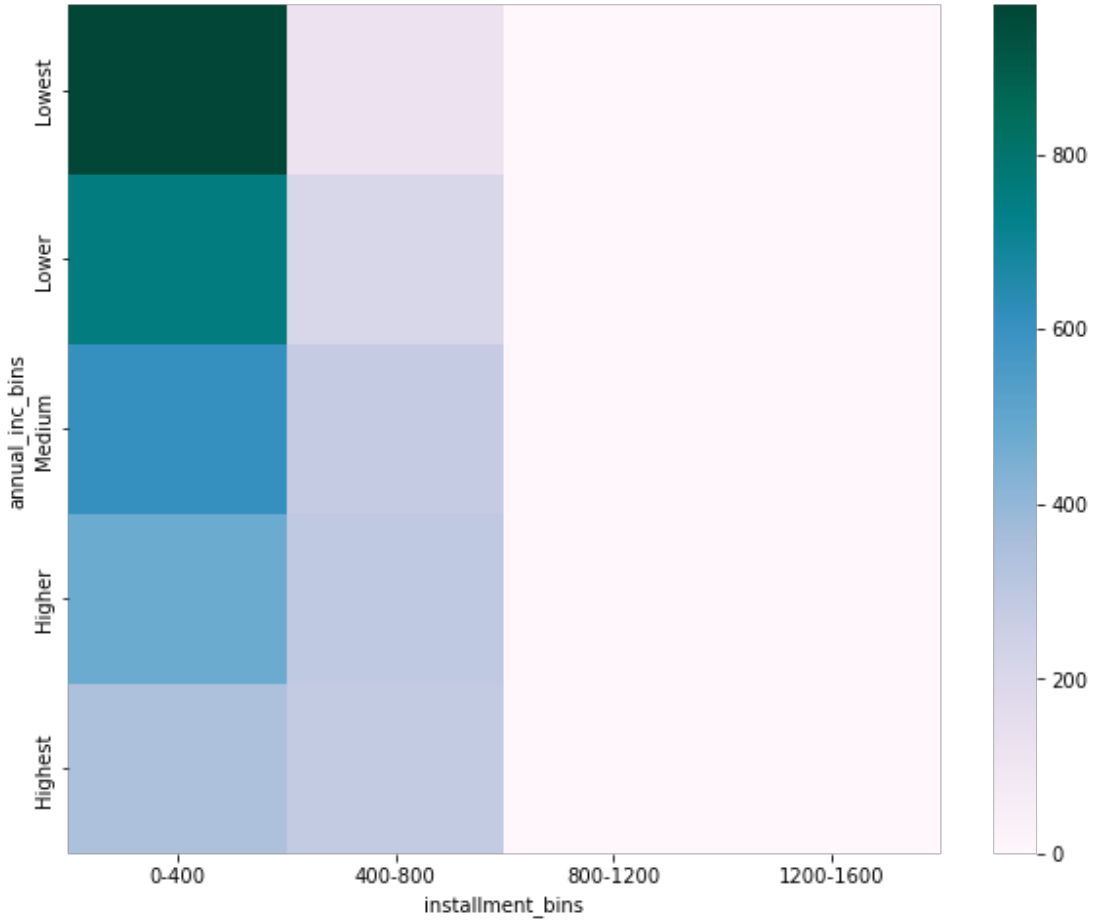
Analysis of dti and annual income



Analysis Result

Chances of defaulting is high for the borrower with lowest income and high dti(debt to income ratio).

Analysis of Installment and Annual Income



Analysis Result:
Chances of defaulting is high when the borrower has annual income in the lowest segment, even though the installment is lowest too.

Conclusion

- We used the following methodology:
  - **Univariate Analysis:** Home Ownership,Grade, Loan Status, emp_length, funded_amnt etc
  - **Segmented Univarate Analysis:** Annual Income, revol_bal,etc
  - **Derived Metrics:** issue_d_year, issue_d_month, etc
  - **Bivariate Analysis:** Correlation metrics on imp cols, heatmap for dti vs annual income, etc
  - Based on the above analysis, we made the following recommendations.

Recommendations

- Lending club should focus more on verifying loans for lower grade loans as the chances of defaulting increases as we move to lower grades of loans.
- Small Businesses Loans are mostly charged off, purpose seems to be one of the major driving factors for loan default.
- int_rate also seems to be one of the major driving factors behind loan defaults considering, with increase in int_rate, loans are defaulted more
- loans of tenure 60 months seems to default more and can be considered as a factor i.e longer duration loans
- Income Group is also one of the driving factors behind loan defaults with max number of loan defaults in lower income group, more guard rails can be put to increase the minimum criteria for loan disbursement
- Lender's whose employement length is greater than 10 years are mostly charged off, therefore employment_length can also be considered as a driving factor
- Home ownership is also one of the key driving factors in driving loan defaultment
- Lender's with low income and high dti are more susceptible to defaulting. Also higher installment leads to more defaults.