Lending Club Case Study

Name: Akshay Kachroo, Aloke Kumar Mukherjee

09-Aug-2023

Business Objective

Client :

Lending Club is a **consumer finance company** which specialises in lending various types of loans to urban customers

Lending loans to 'risky' applicants is the largest source of financial loss (called credit loss), like all Lending companies

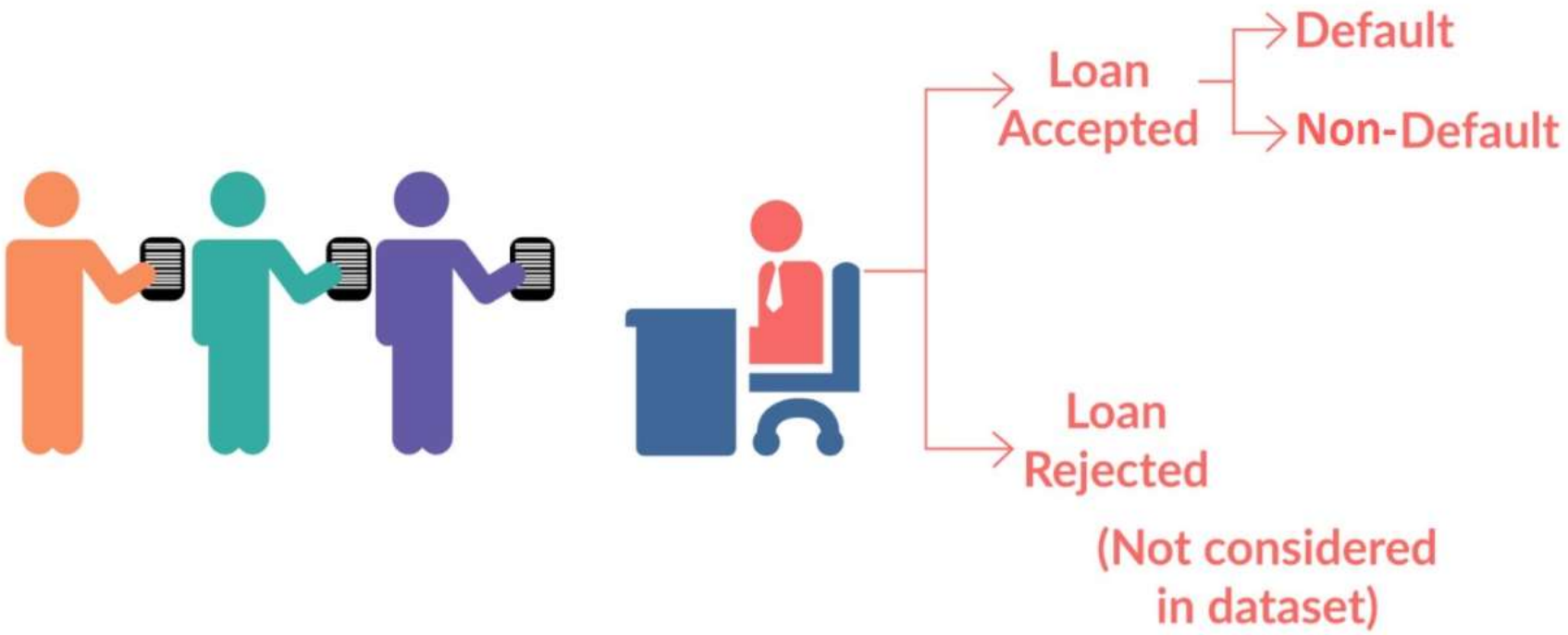
LC wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

Objective :

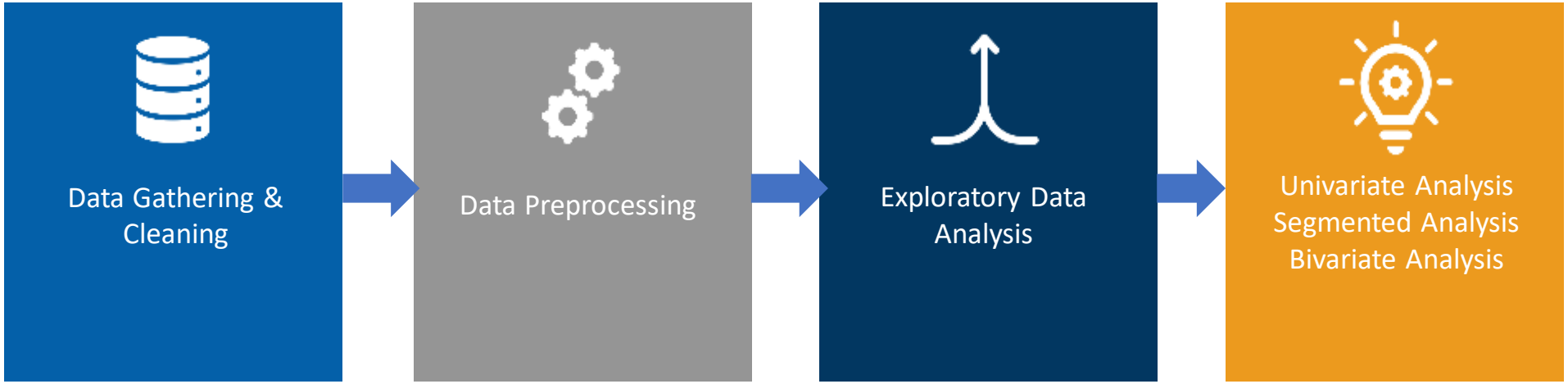
LC wants to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss

Use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Loan Dataset



Problem Solving Methodology



Data Cleaning

- Once we are done with loading the data and identified those columns which can be cleaned then we can proceed with some data cleaning steps which has been taught to us in DA.
- As while understanding the data, we found there are 54 columns which has all the values “NULL” or “NaN”, as a starting point of cleaning activity we dropped those columns as column with no values won’t be helpful for analysing the motive.
- There were some columns which had a lot of null values so we remove the columns having null more than 30%.
- After dropping all the NULL values column, we end up having 53 columns of data with no duplicates rows.
- After taking the look at the data as per the understanding giving by coach there are several rows which are specific to Customer Behaviour like deling_2_year, revol_bal, last_pymnt_d etc so we can drop those columns also since we are analysis a data for Diagnostic purpose and not performing any Prediction out of it so these variables came into picture if we are building Predictive model. As our main moto stays, with the existing data need to show who all are could be defaulter or not defaulters.
- After removing all the columns, we have 19 columns which are mainly focused on the borrowers characteristics like the loan amount, emp_length, emp_title, Grade, int_rate etc.

Data Preprocessing

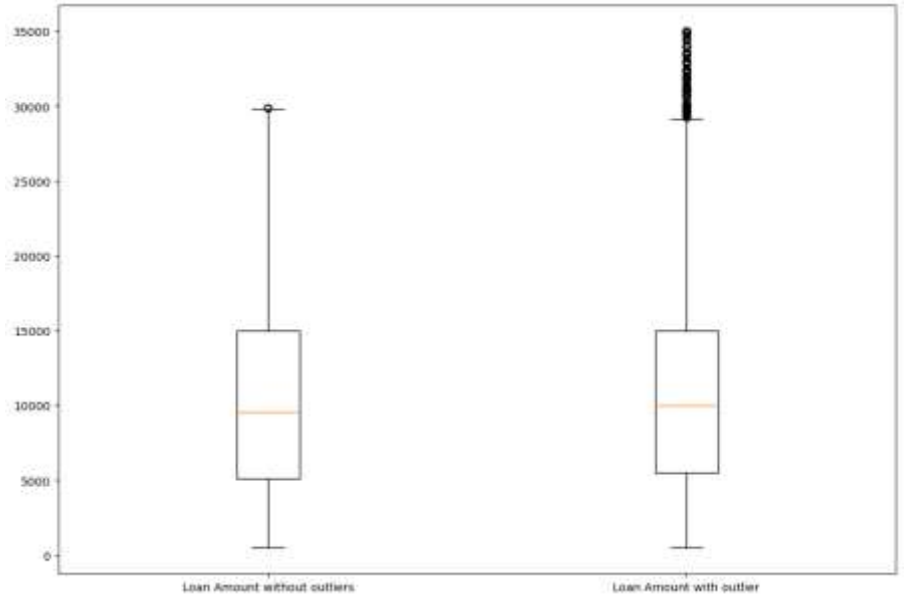
- After cleaning up all the unnecessary and invalid data we have the relevant data at our hand which is still raw and needs to be rectified for better analysis.
- We separated the numeric value from the string value using regex, and also changed the datatype of the same field from “object” to “int” so that the column will have numeric value all the way to end and to check the values statistically for columns like interest_rate etc.
- After handling such cases, went further and checked the issue_d column, which don’t have any NULL values, but since the values represented as Month-Year type, more like a date field, we can derive Month and Year from it. We derived the Month and Year from it, since I think it will helps to show the Interest Rate distribution month wise.
- After checking and deriving some metric we further proceeded with checking outlier, after going through int_rate, loan amount range variables we found annual income is one of the important field since its dealing with annual income of the borrower. While checking, found the there are outlier present in that field and to visualize the same, thought of displaying it using boxplot, since its very useful to display outliers.

Data Preprocessing

- No. of rows : 39717, No. of columns : 111 present in i/p data set
- No. of columns having no data : 54
- The data present in i/p dataset can be classified in three categories:
  - Customer demographic information (Annual income, No. of employment years etc.)
  - Loan related information (Loan amount, Interest rate, Loan status, Loan grade etc.)
  - Customer behaviour information (Purpose of loan, application type etc.)
- The target variable for the analysis is loan_status which has 3 values: “Charged off”, “Fully Paid”, “Current”
- After cleaning up all the unnecessary and invalid data we have the relevant data at our hand which is still raw and needs to be rectified for better analysis.
- We separated the numeric value from the string value using regex, and also changed the datatype of the same field from “object” to “int” so that the column will have numeric value all the way to end and to check the values statistically for columns like interest_rate etc.
- After handling such cases, went further and checked the issue_d column, which don’t have any NULL values, but since the values represented as Month-Year type, more like a date field, we can derive Month and Year from it. We derived the Month and Year from it, since I think it will helps to show the Interest Rate distribution month wise.
- After checking and deriving some metric we further proceeded with checking outlier, after going through int_rate, loan amount range variables we found annual income is one of the important field since its dealing with annual income of the borrower. While checking, found the there are outlier present in that field and to visualize the same, thought of displaying it using boxplot, since its very useful to display outliers.

Data Analysis

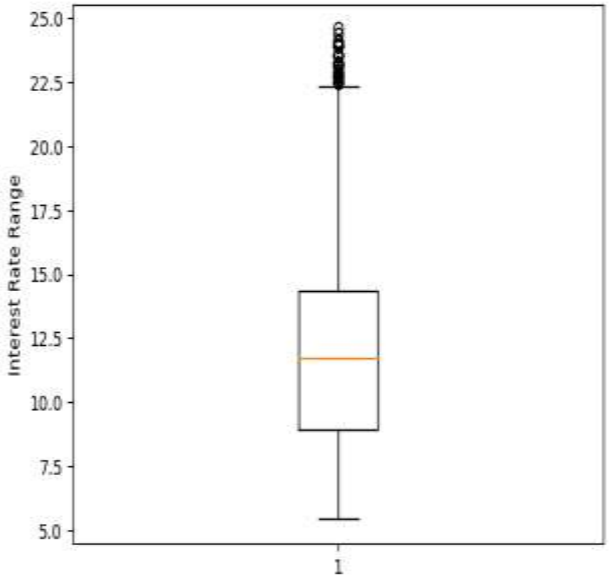
Loan Amount Range



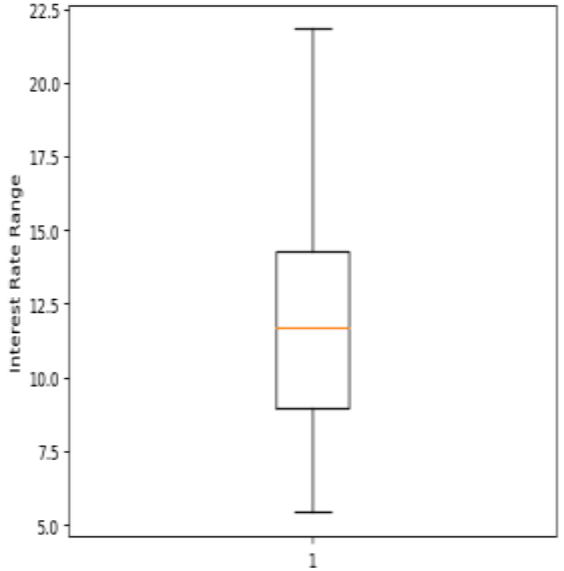
Before Outlier Removal

After Outlier Removal

Interest Rate Range



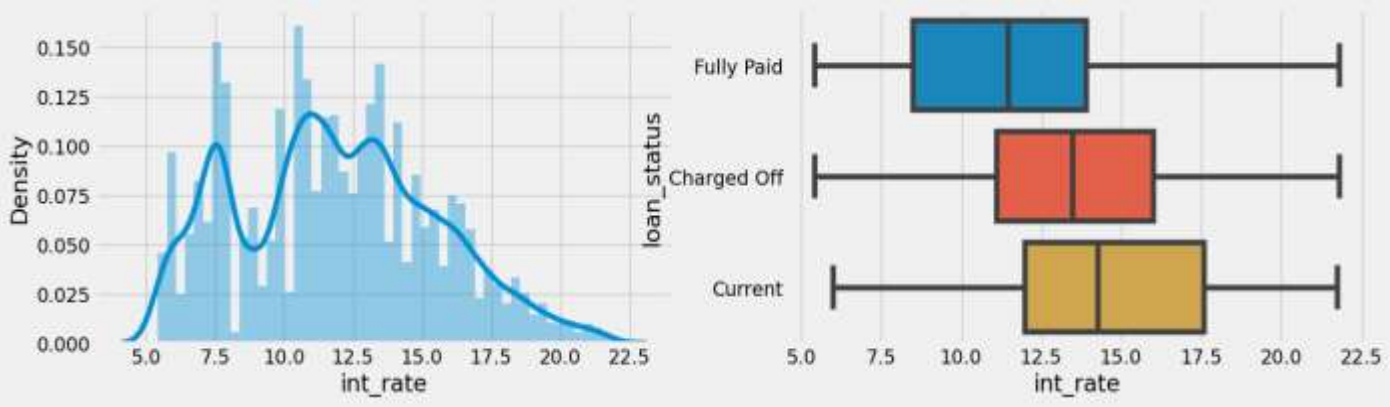
Before Outlier Removal



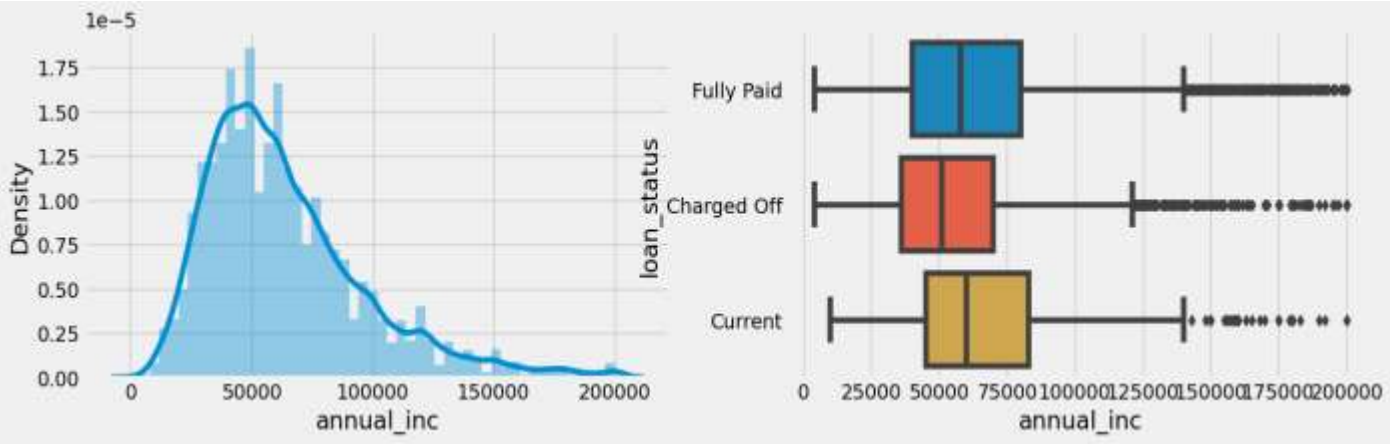
After Outlier Removal

Univariate Analysis

Distribution of Interest Rate across Loan Status

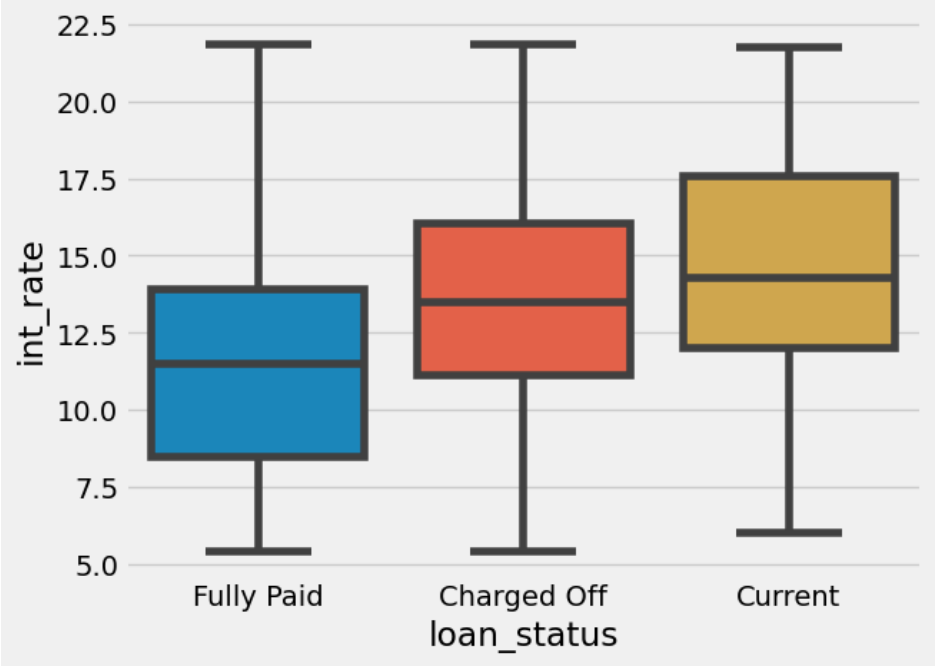


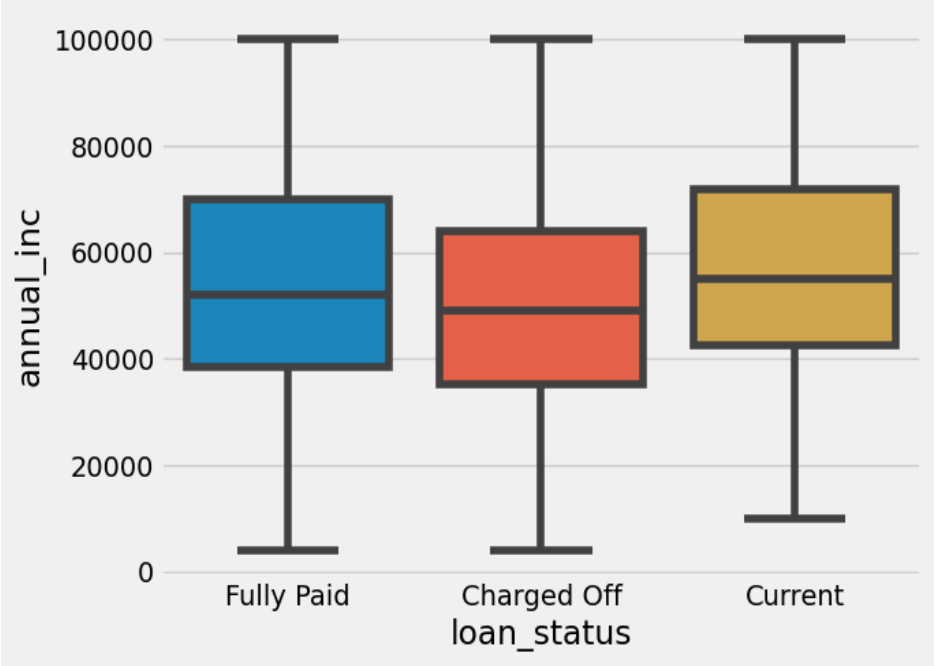
Distribution of Annual Income across Loan Status



- **Interest Rate**
  - Charged off loans are ~2% points higher than Fully paid loans.
  - LC is assessing risk correctly and charging riskier loans higher
- **Loan income**
  - As can be seen, between Charged off and Fully Paid category Fully Paid have got more annual income.
  - We can infer that majority of income is ~50000.

Univariate Analysis





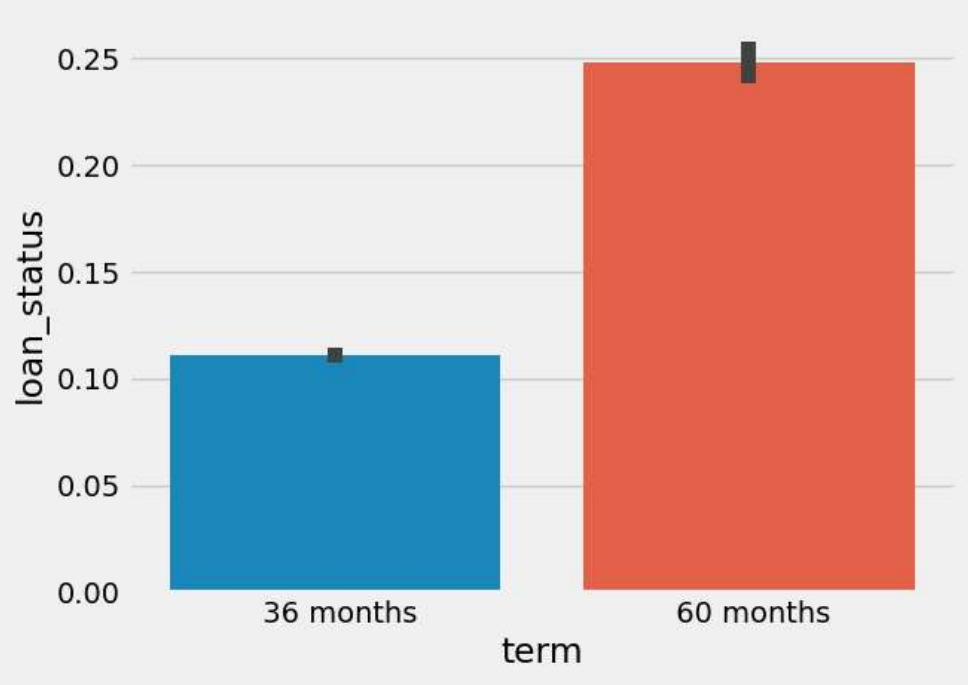
Interest Rates

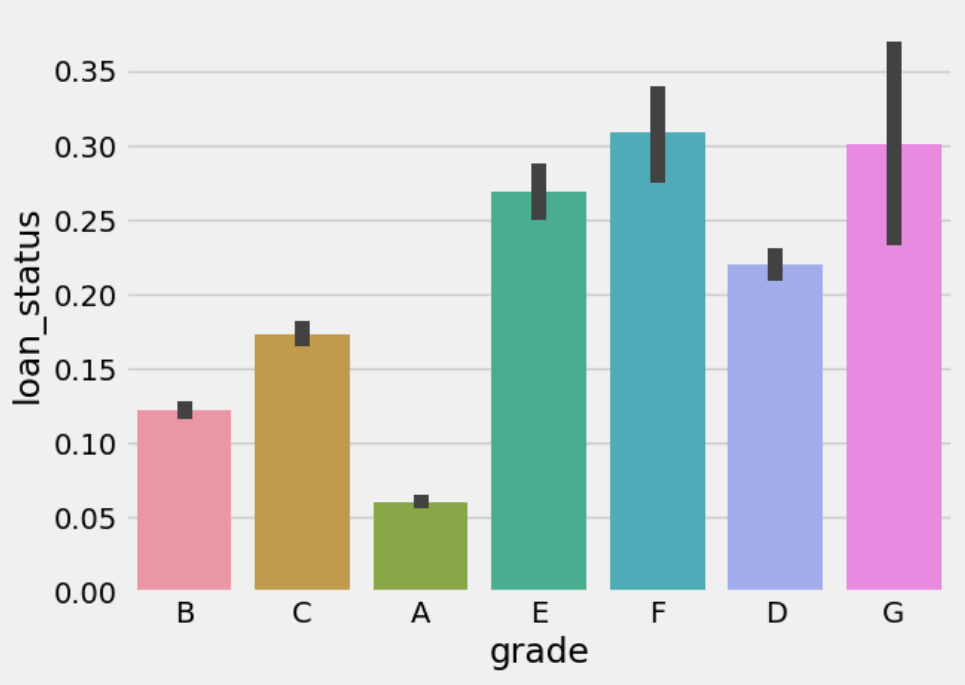
- Interest rate is categorized .. More chance of default if interest rate is very high

Annual Income

- The more chance of default if annual income is low. Current means the cases where loan repayment is happening , so it is not under analysis

Univariate Analysis





Loan Term

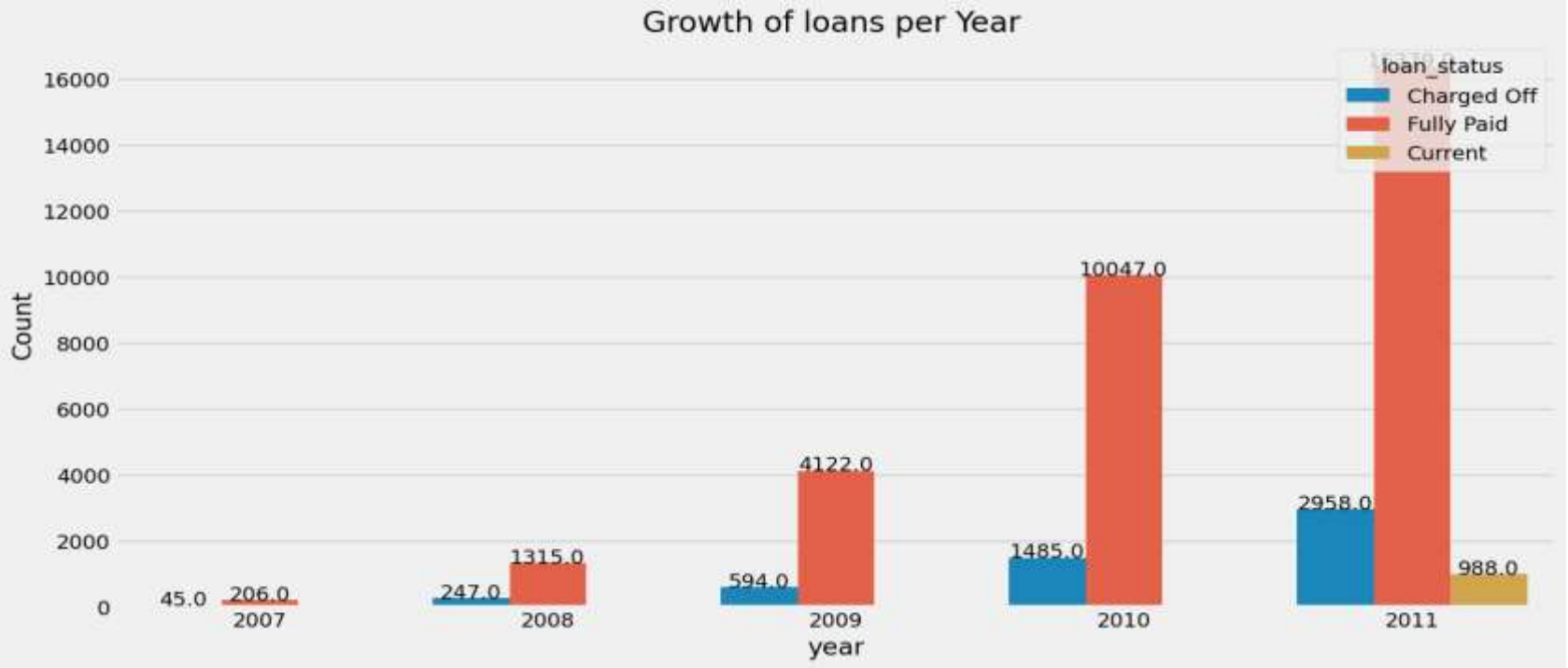
- The above plot is showing if loan term is more, the chance of being charged-off is more

Grade

- The above plot is showing that loan grade F has higher chance of being charged off

Univariate Analysis

Distribution of Loan Status across Years



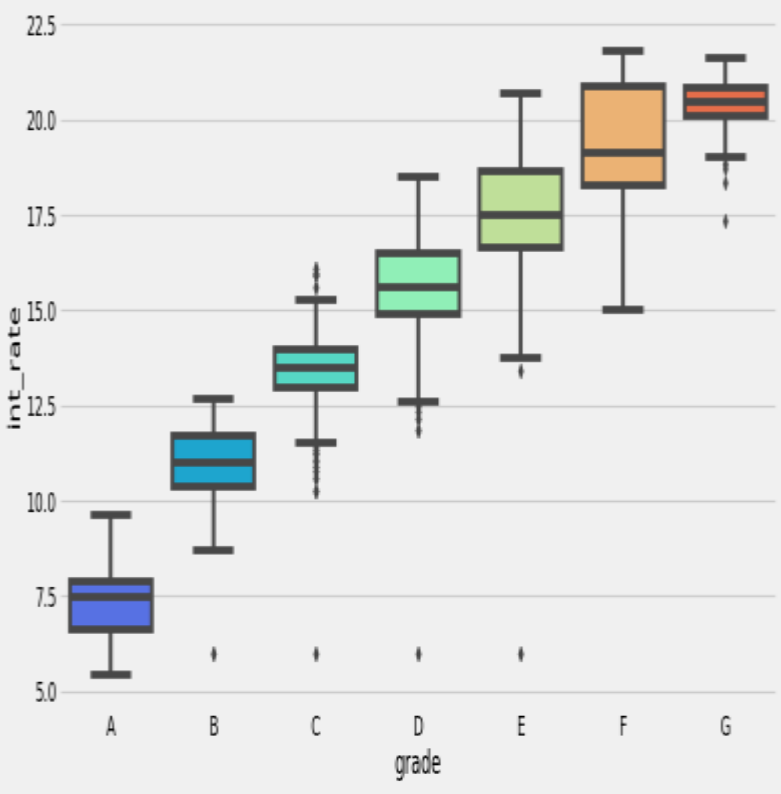
YOY GROWTH

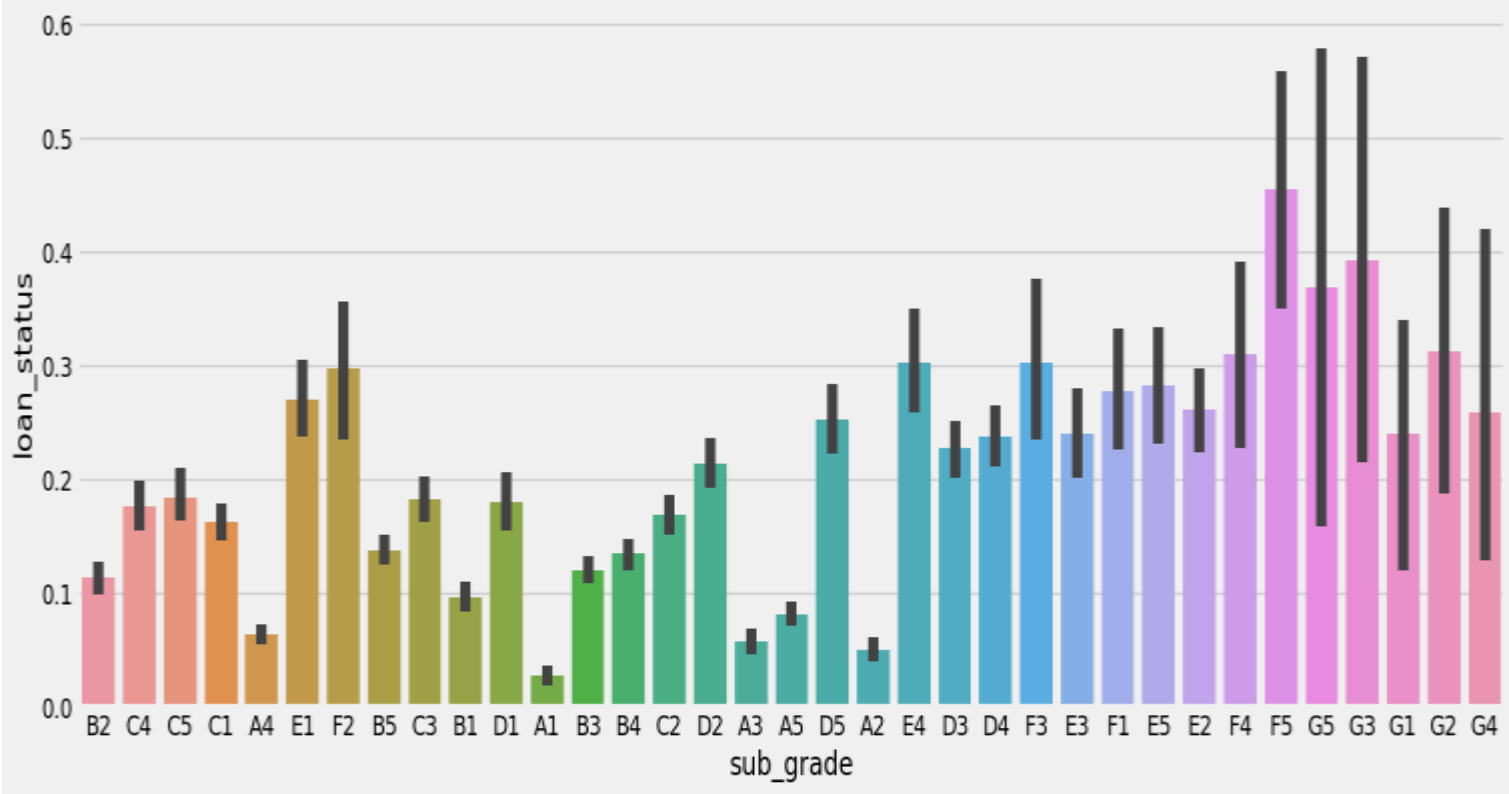
- The count of Fully paid and Charged off have kept on increasing every year with Fully paid increasing at strong rate.

Bi Variate Analysis

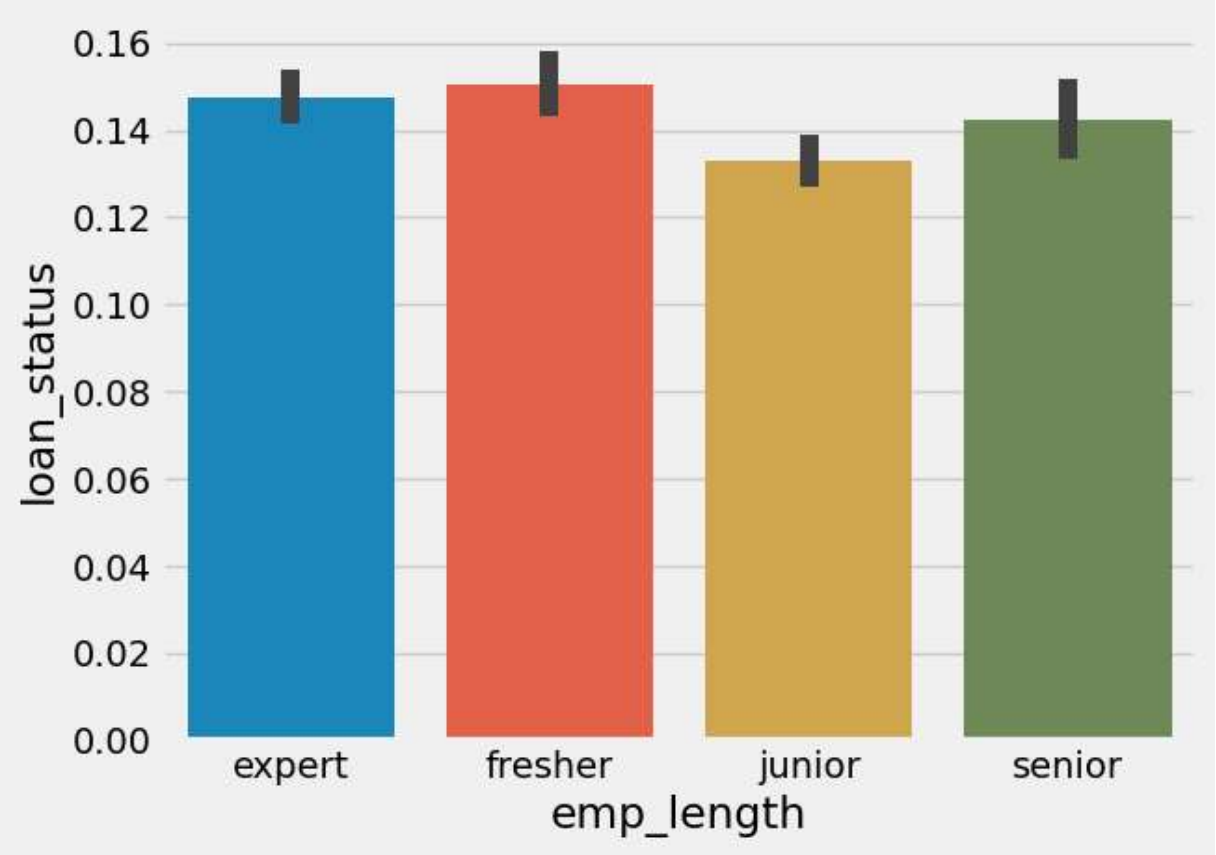
Grade and Sub grade of Default

- It can be seen the interest rates are high for the lower level grades E, F, D, G
- The more chance of default for Sub Grade F5, so LC should restrict that

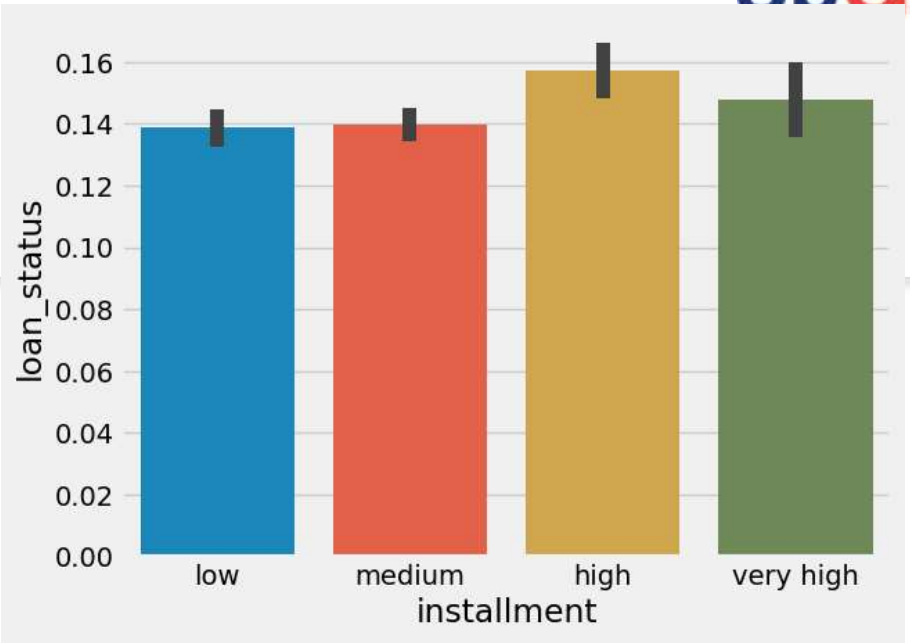




Bi Variate Analysis



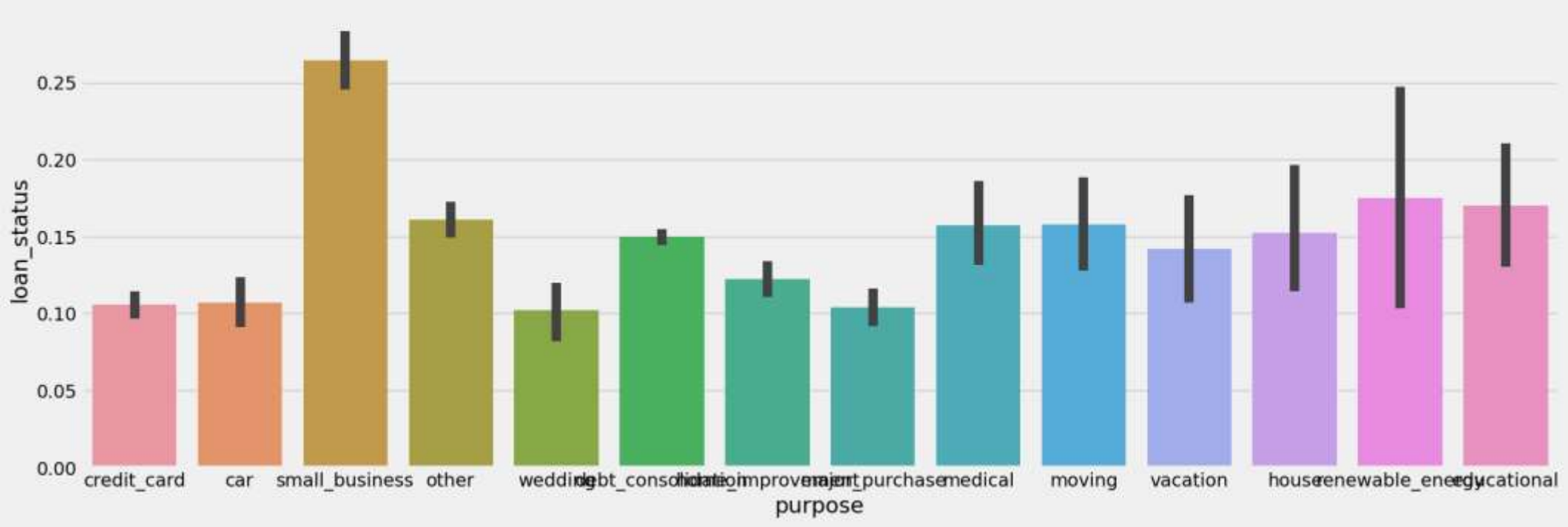
Installment Distribution



Employment Length and Installment of Default

- Installment is categorized .. More chance of default if no. of installment is high
- Employment length is categorized .. More chance of default for the case of fresher

Bi Variate Analysis

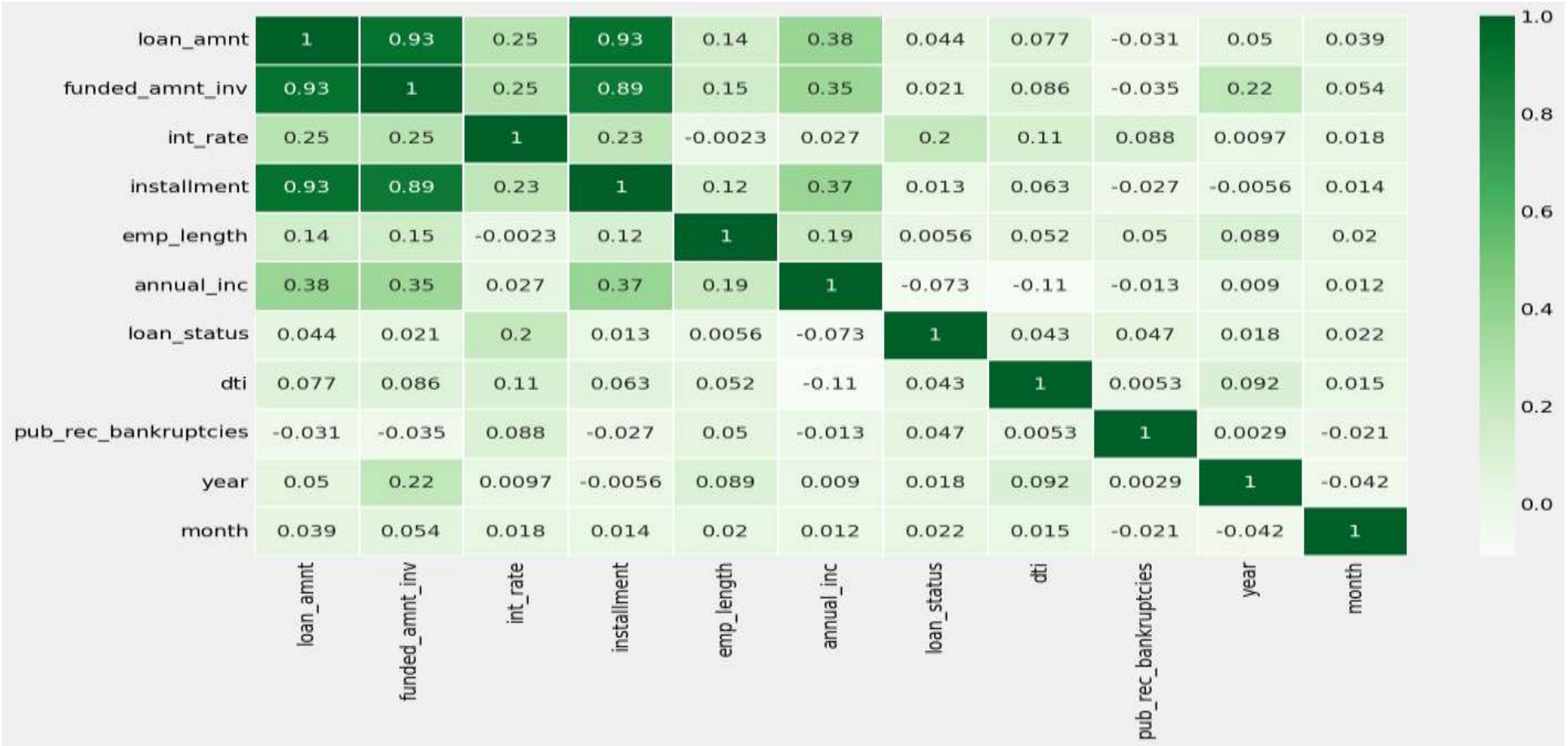


Loan Purpose in comparison to the charged off loan status

- The above plot is showing that loan approved for small business category has more chance of being charged off whereas loan approved for wedding category has least chance of being charged off

Bi Variate Analysis

Heatmap Co-relation Analysis for default category



Analysis Summary

Summary Points

- Interest rate - there is higher chance of Charged-off if the interest rate is high
- Anual_inc - there is higher chance of Charged-off if the annual income is low
- Grade - there is higher chance of Charged-off if loan grade is of F category
- Sub_grade - there is higher chance of Charged-off if loan grade is of F5 category
- Term - there is higher chance of Charged-off if no. of learn terms is 60
- Purpose - there is higher chance of Charged-off if loan is taken for small business
- Loan_amnt - there is higher chance of Charged-off if loan amount is very high
- Installment - there is higher chance of Charged-off if no. of installment is high
- Emp_length - there is higher chance of Charged-off for fresher