Loan Defaulters Prediction

By Shubham Kumar Gupta

Cohort ID: 4401

Problem Statement

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

- Like most other lending companies, lending to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or absconds with the money owed. In other words, borrowers who default cause the biggest losses to lenders. In this case, customers labelled as ‘charged-off’ are the ‘defaulters’.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- In other words, 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.

Approach for the analysis

1. Data Reading and Understanding
2. Data Cleaning
3. Fixing rows and columns
4. Data analysis

1. Data Reading and Understanding

- Read the dataset "loan.csv" into a dataframe variable named "loan_data".
- In Data Understanding, the information of all the column variables of the dataframe along with their numeric description such as mean, quantiles at 25%, 75% and 50%, min and max values have been displayed.
- *Observations*
  - There are total 111 columns and 397171 rows with different datatypes such as int, float and object.
  - According to the numeric description, many columns such as tot_hi_cred_lim ,total_bal_ex_mort, total_bc_limit ,total_il_high_credit_limit, etc. contained only null data indicating that they won't be helpful in data analysis.

2. Data Cleaning

2.1. Remove/Impute null values:

- Upon checking for the percentage of null values column wise, more than 90% of null values have been found in most of the columns which were then removed.
- Again upon checking the dataset for percentage of null values, two columns still left with more than 30% of null values so they were also removed.

2.2. Drop unnecessary columns:

- Even after deleting columns with null values there were still columns left which won't helpful in analysis such as:
  - Some of those columns were associated with personal data: "title", "url", "zip_code", "addr_state", etc.
  - Some columns were having only 1 input throughout the whole datset: "chargeoff_within_12_mths", "delinq_amnt","pymnt_plan", "tax_liens", etc.
- That's why they were removed.

2.3. Drop unnecessary rows:

- Upon checking the data row wise for more than 5 missing values, no such rows were found

3.Fixing rows and columns

- 3.1. Fix Data types
  - The data types of interest rate and term were being converted from object to numerical data type.
  - The month and year were being extracted from issue date into two columns seperately and removed the column of issue date.

4. Data Analysis

- Upon checking for the numerical and categorical variables, the target variable i.e. “loan status” contains the records of loans of previous occurred “charged off(defaulted)”, “fully paid” and “current”.
- The rows of current category loans were removed as they were neither paid nor defaulted.
- The Data analysis was being done in following categories along with their data visualizations:

  4.1. Univariate analysis

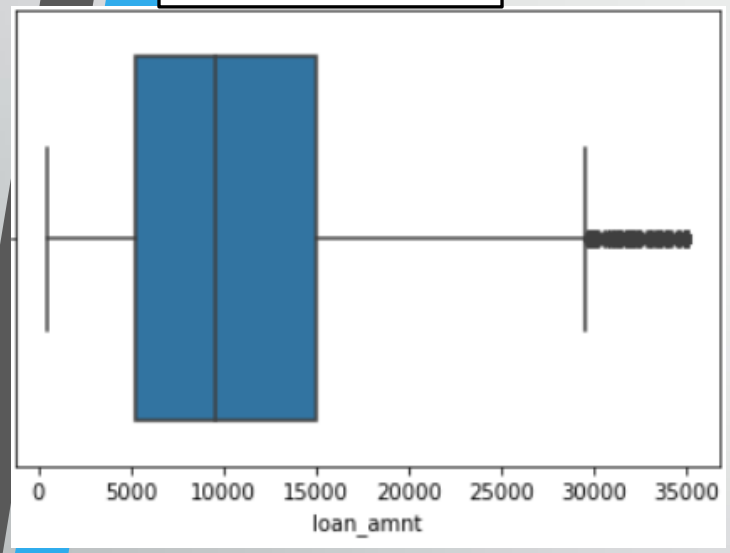
  4.2. Bivariate analysis

  4.3. Multivariate analysis

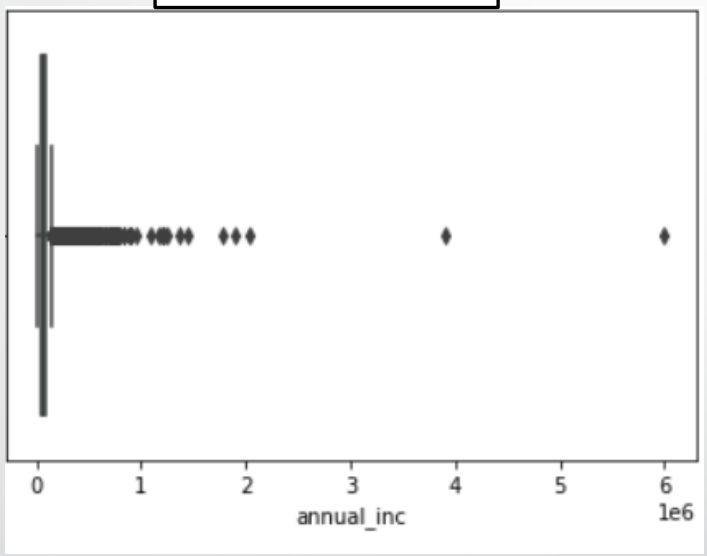
4.1. Univariate analysis

4.1.1. Numeric variables:

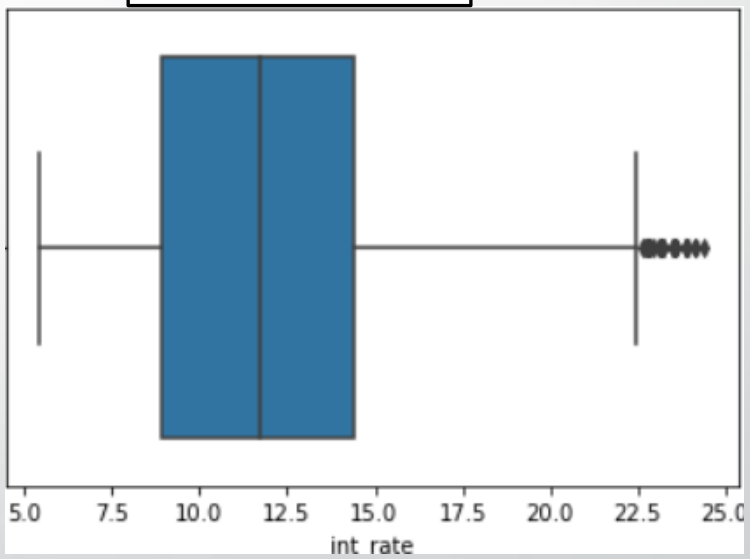
Loan amount



Annual income



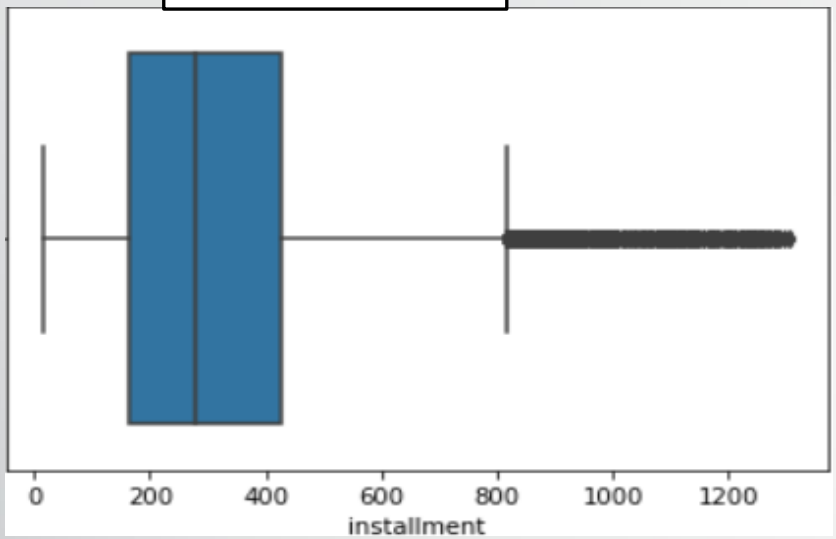
Interest rate



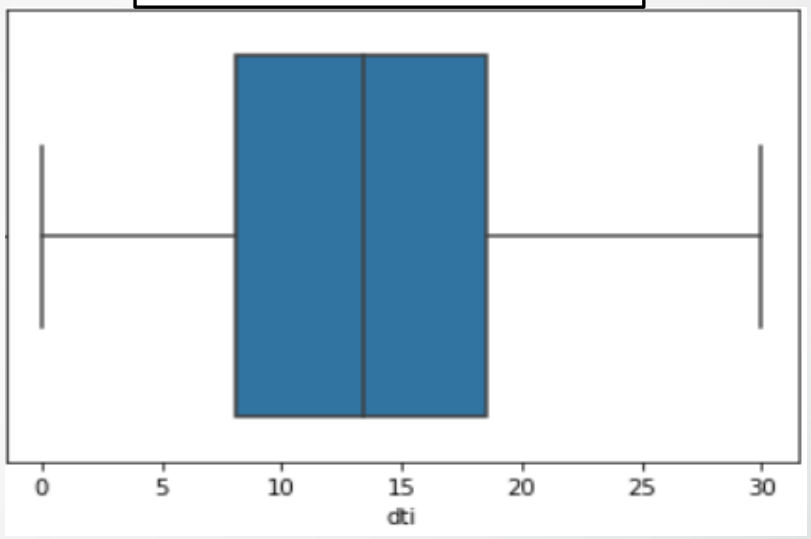
- There are outliers as shown in the above graphs but they are in continuity except annual income where certain income outliers are far away which indicate that there are certain borrowers with high incomes.

- The following graphs display the data of the Numeric variables:

Installment



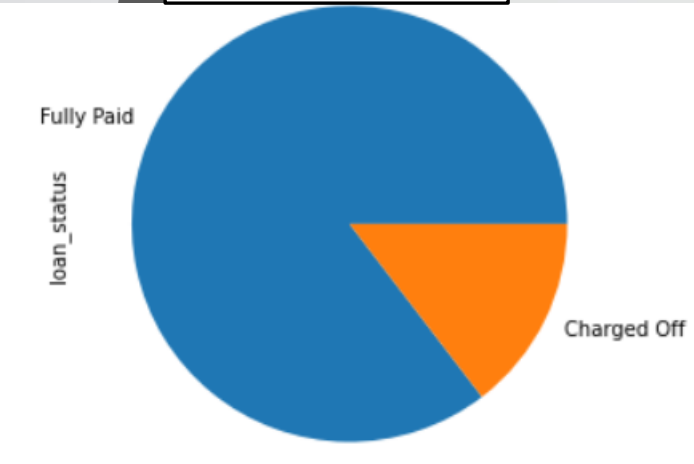
Debt to Income ratio



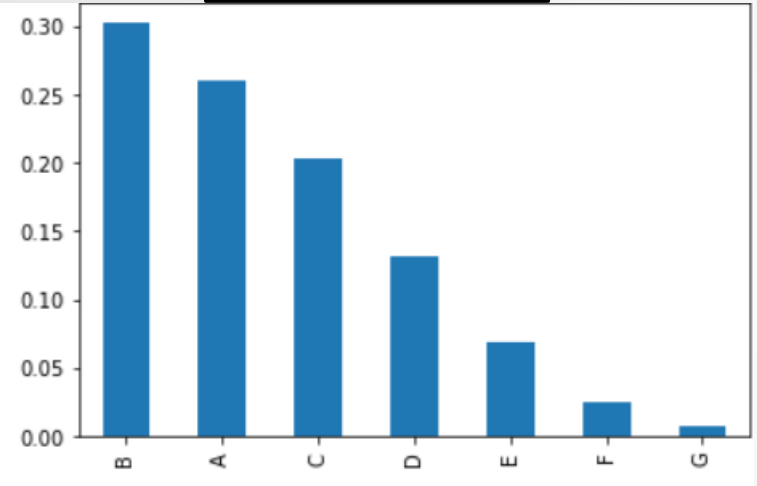
- The outliers in Installment shows continuity while the Debt to Income ratio does not contain any outliers.

4.1.2. Categorical variables:

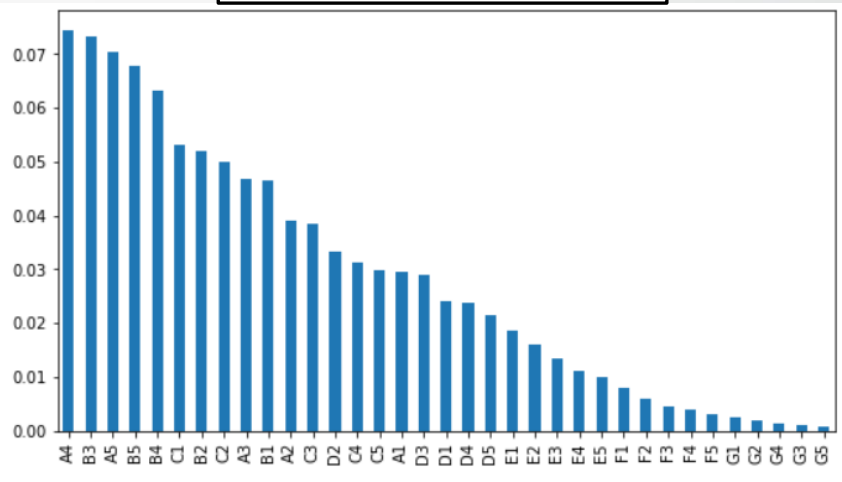
Loan status



Grade of loans

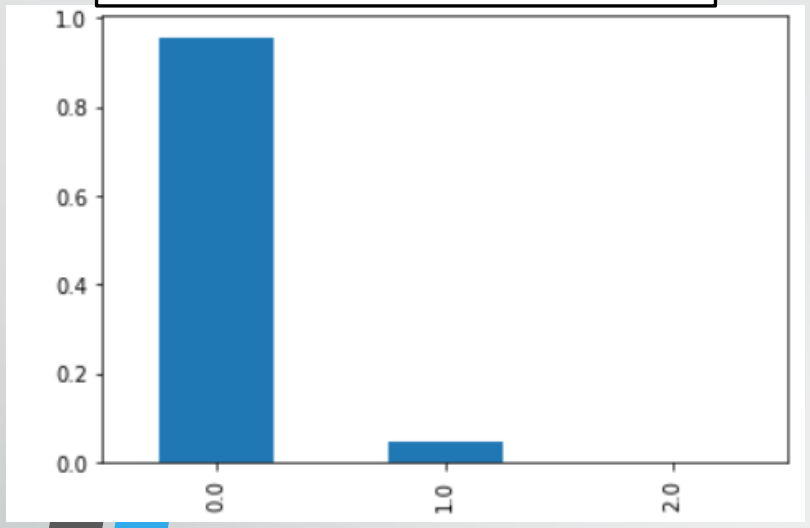


Sub Grade of loans

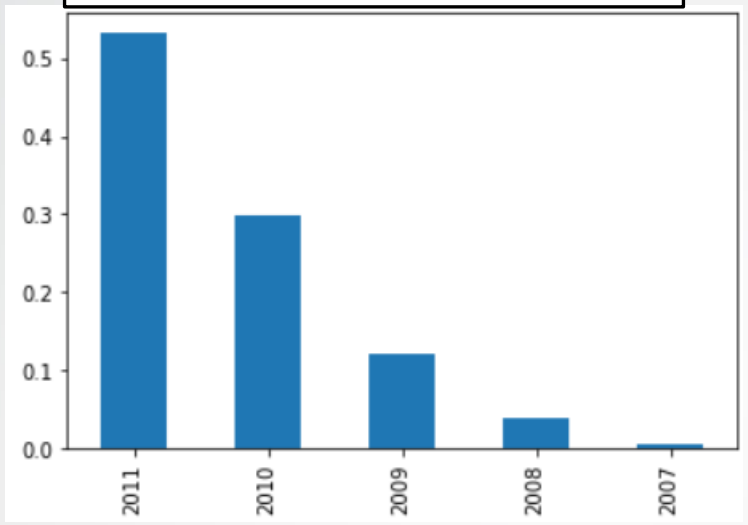


- The loan status is mainly the target variable for further analysis as on its basis of previous loan defaulters, it can be evaluated as who is most likely to default in future.
- The Grade and Sub Grade loans indicate that there are higher quantity of best grade loans i.e. ‘A’, ‘B’ than others.

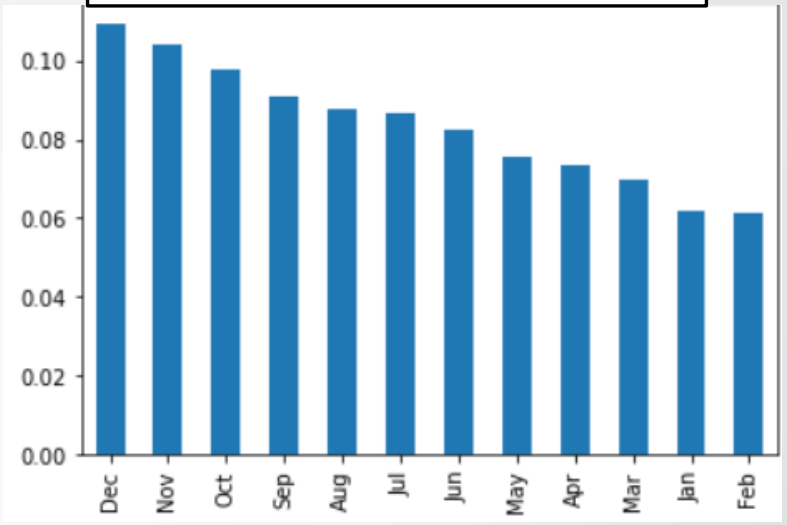
Public record bankruptcies



Issue year of loans



Issue month of loans

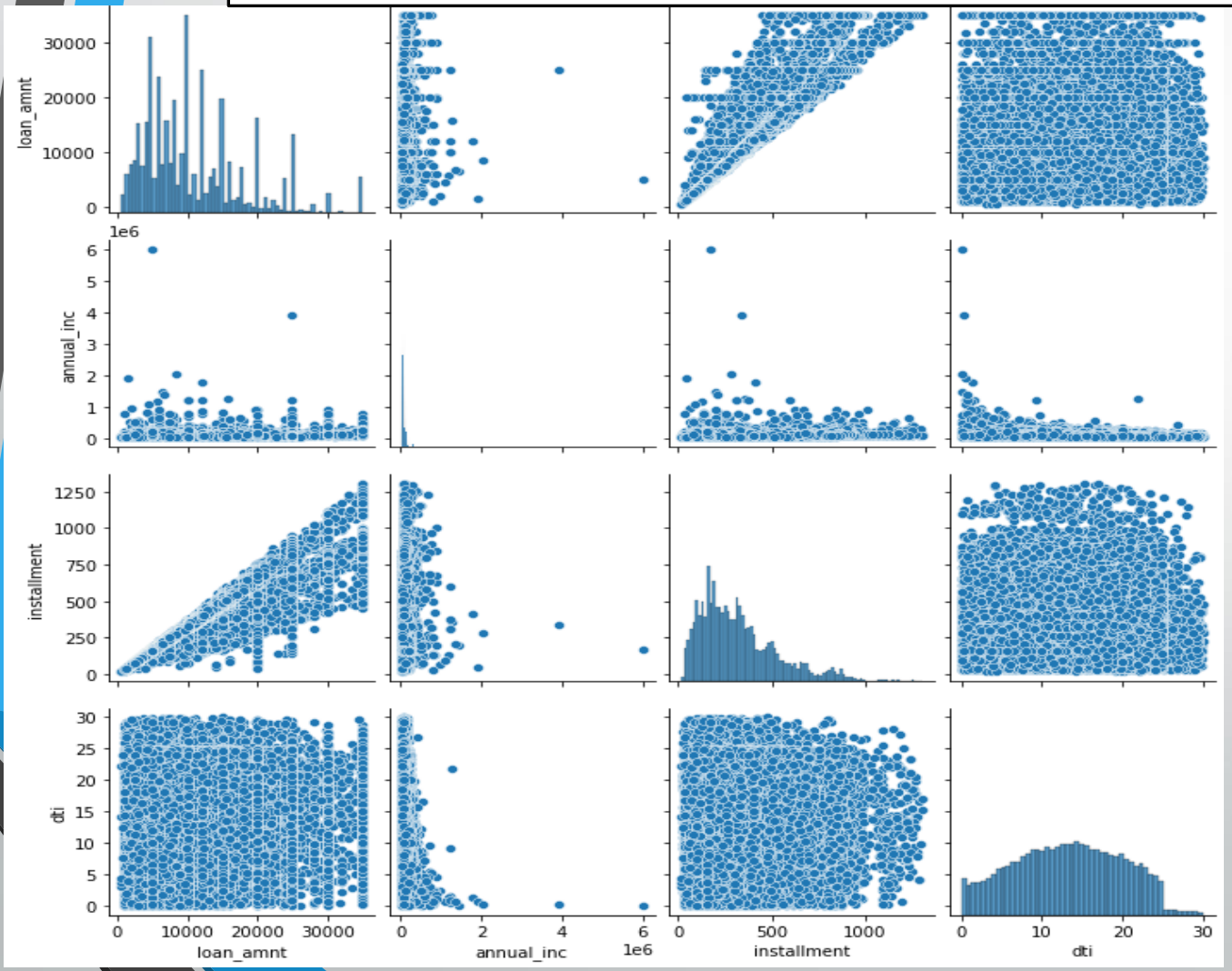


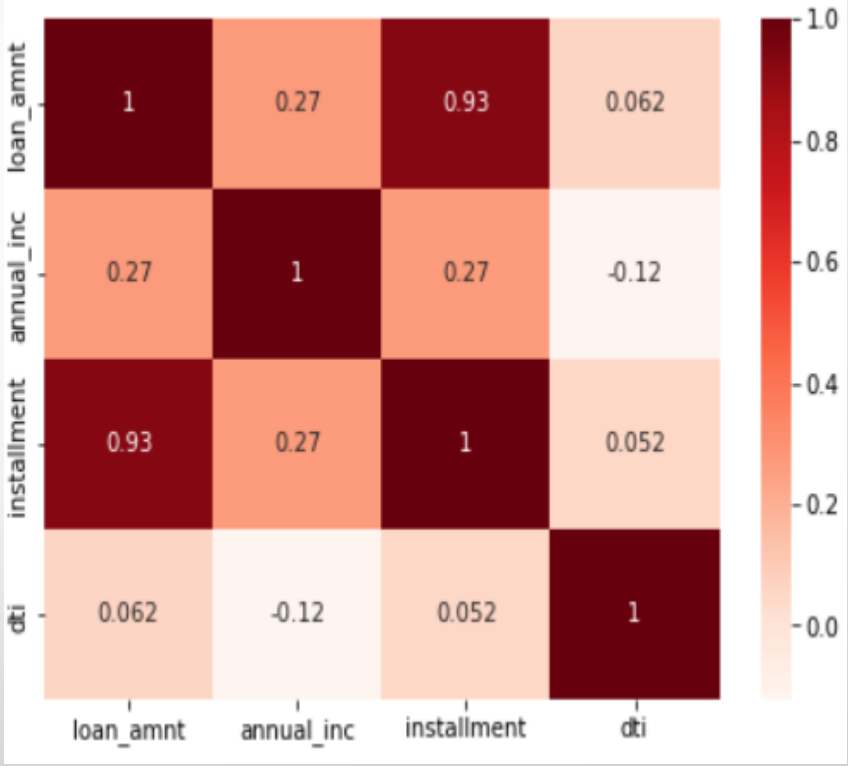
- Most borrowers have zero record of bankruptcies so maybe there is a higher chance of loan recovery.
- According to the issuing year and month of loans, most loans were being issued in 2011 and in Dec respectively.

4.2. Bivariate analysis

4.2.1. Numeric-numeric analysis:

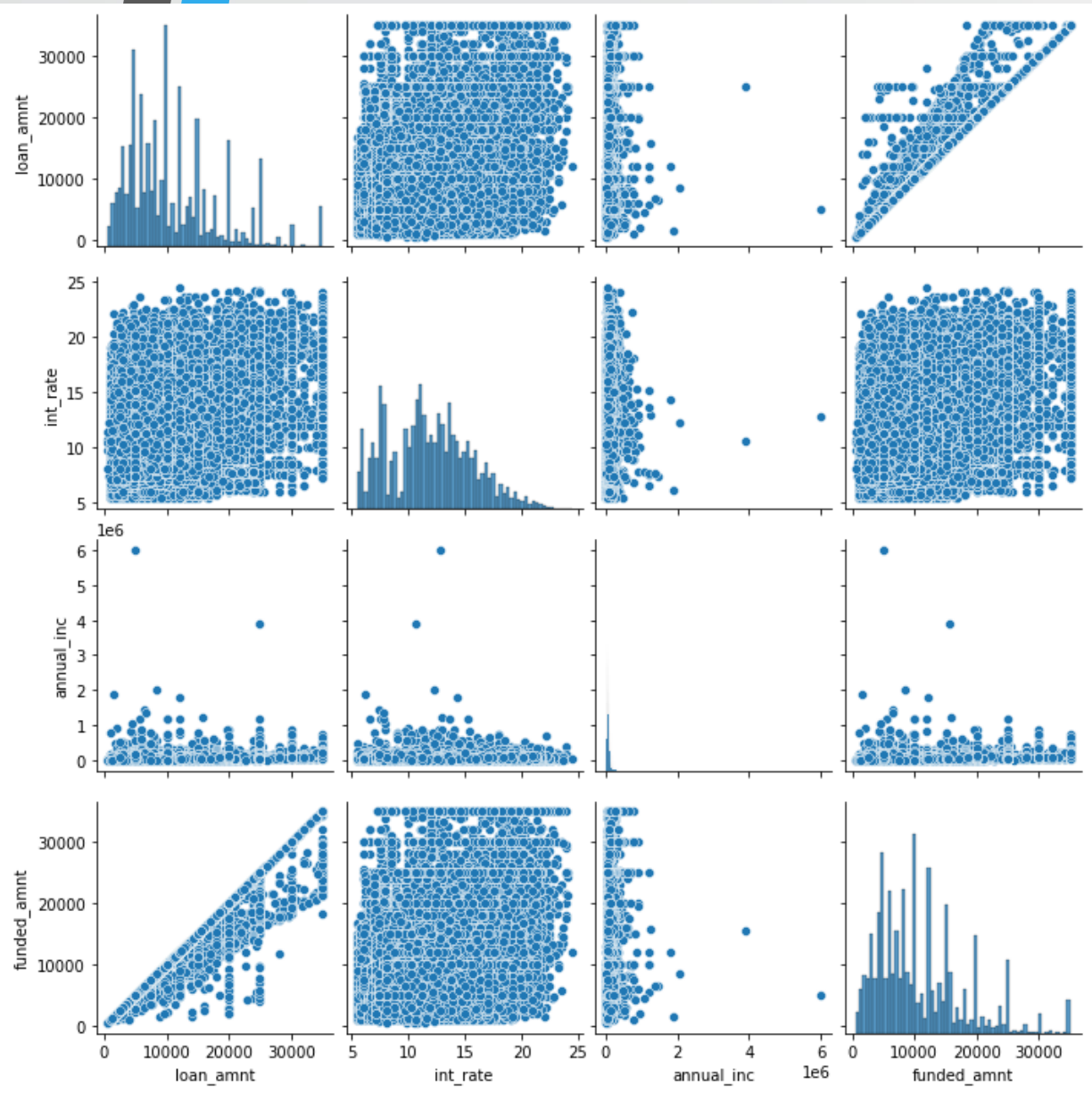
pair plot and heatmap of loan amount, annual income, installment and debt to income ratio





- Strong correlation between loan amount and installment.
- Weakest correlation between annual income and debt to income ratio

pair plot and heatmap of loan amount, interest rate, annual income and funded amount

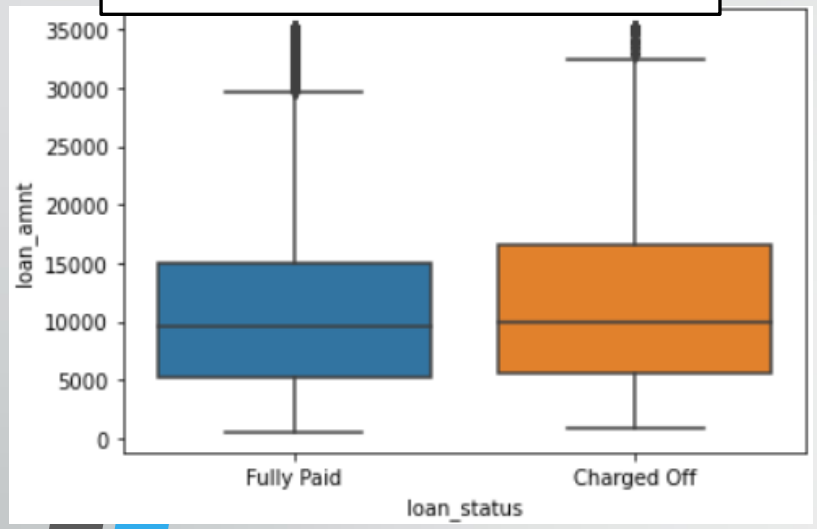


- Strong correlation between loan amount and funded amount.
- Weak correlation between annual income and interest rate.

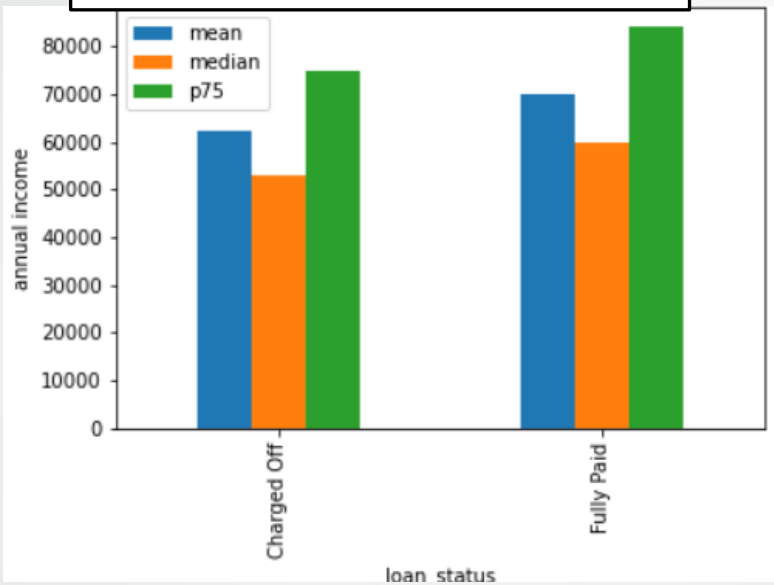
4.2.2. Numeric- Categorical analysis:

- All the numeric variables have been plotted on basis of loan status as it is the target variable.

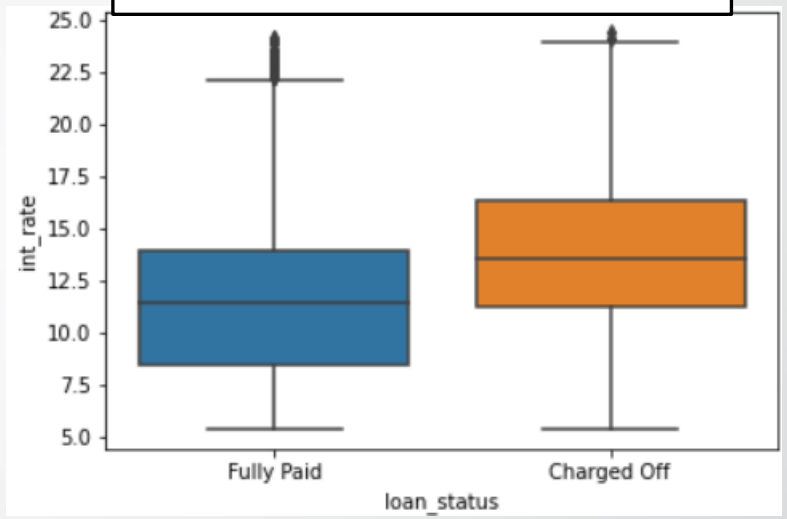
Loan amount v loan status



Annual income v loan status



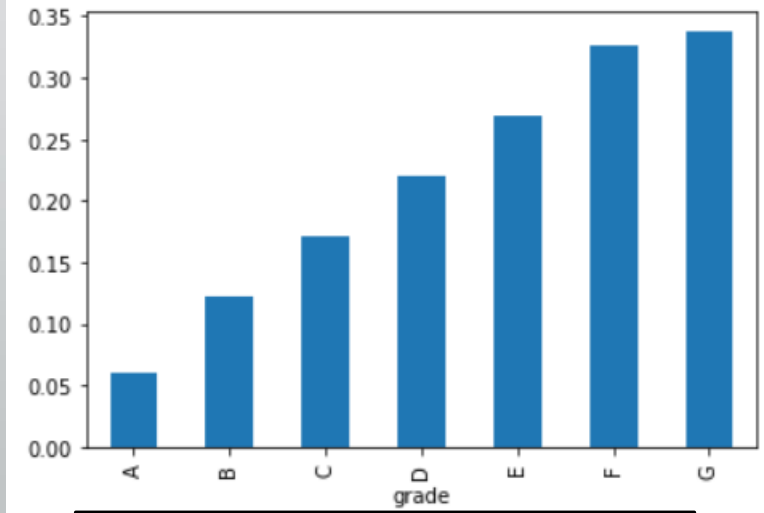
Interest rate v loan status



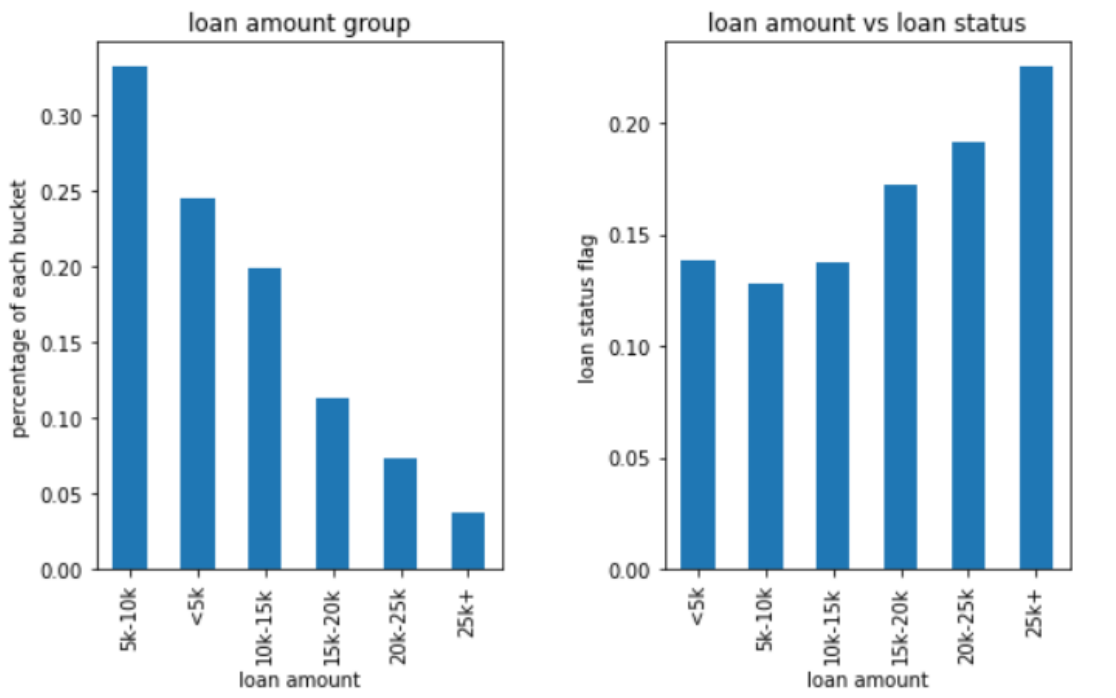
- Observations:
  - The borrowers with higher loan amount are more likely to get charged off than with the lower loan amount.
  - The borrowers with lower annual income are more likely to get charged off.
  - The borrowers with higher interest rate are more likely to get charged off.

4.2.3. Categorical-Categorical Analysis:

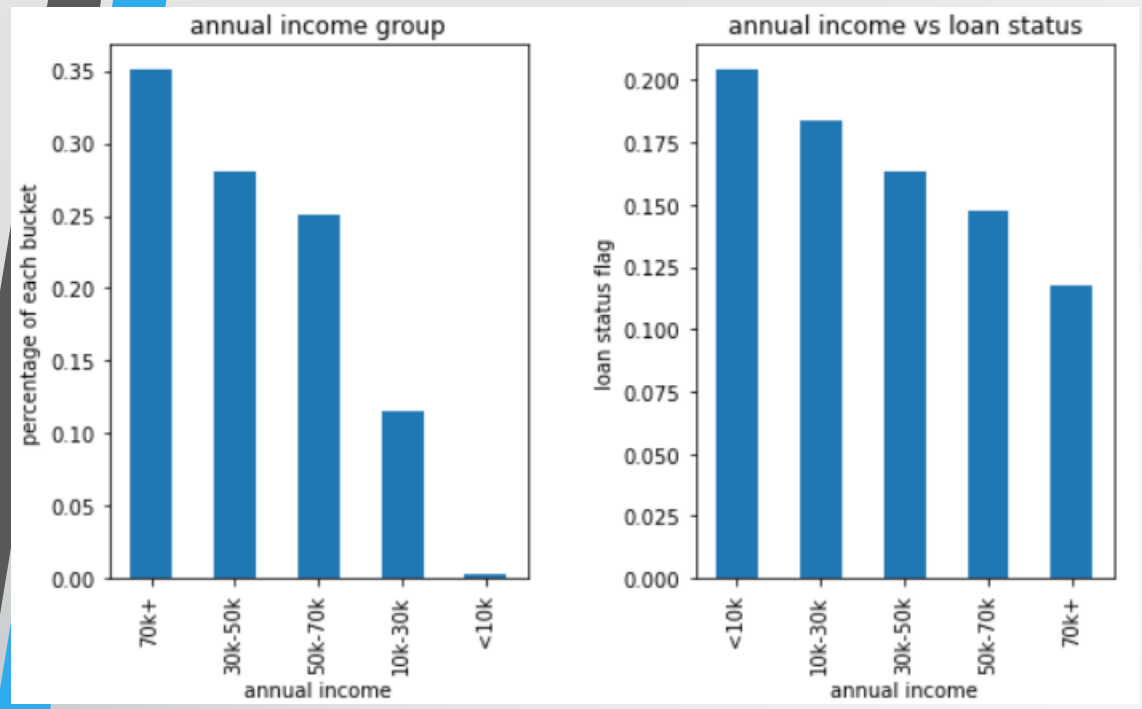
- A loan status flag variable has been created from loan status where charged off = 1; fully paid = 0 and on the basis of this flag variable all the analysis have been done to gain meaningful insights.

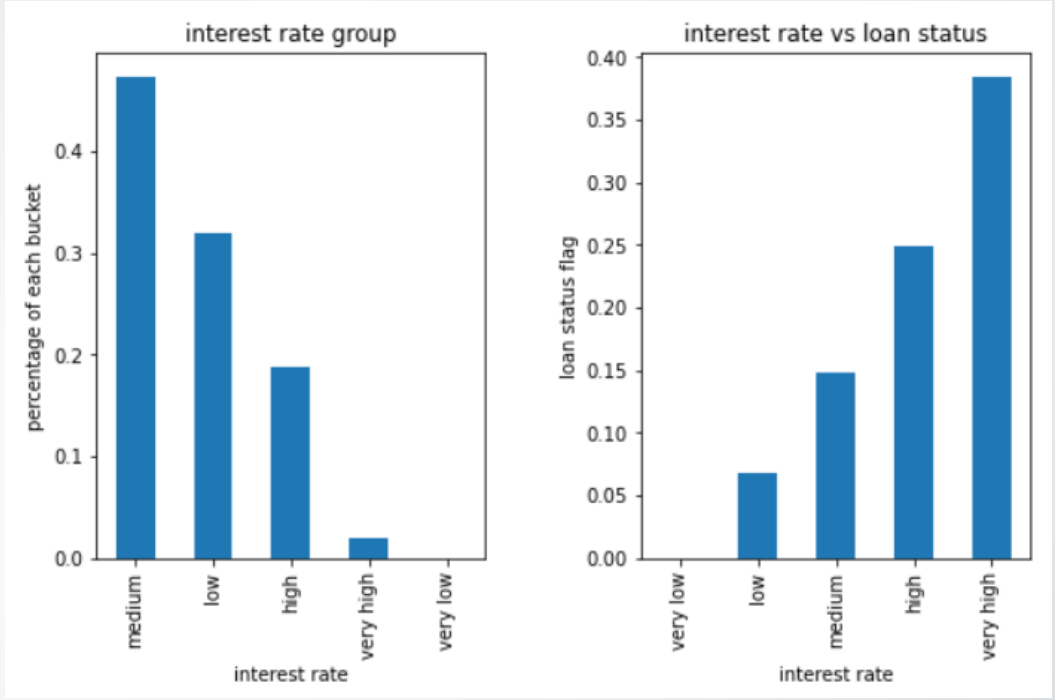


Grade v loan status flag

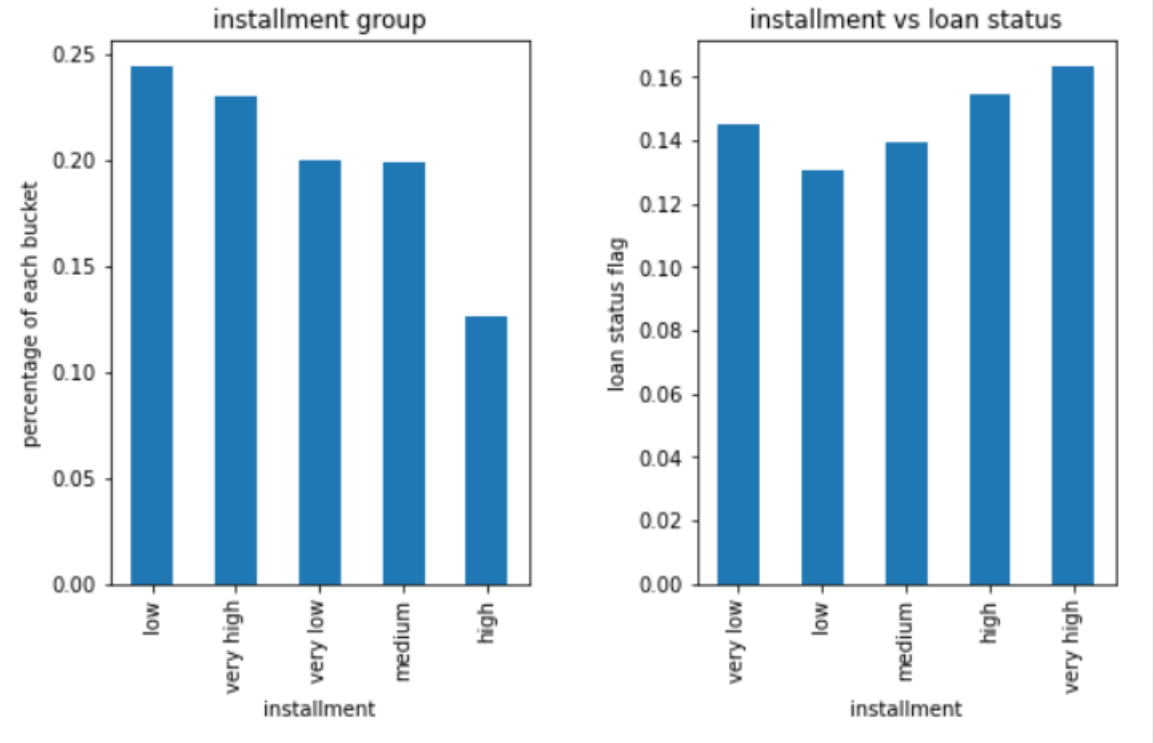


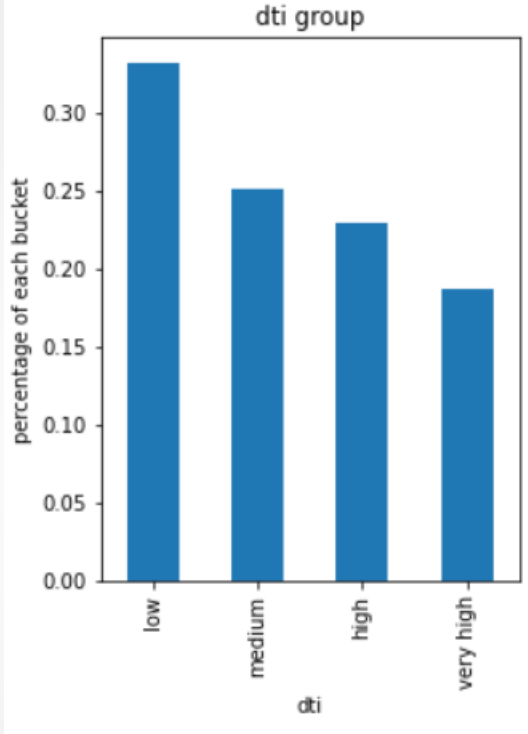
- The higher grade of loans are less likely to get charged off than the lower grade loans.
- The lower amount of loans will be less charged off than higher amount of loans i.e. there is an direct relation between loan amount and loan status.

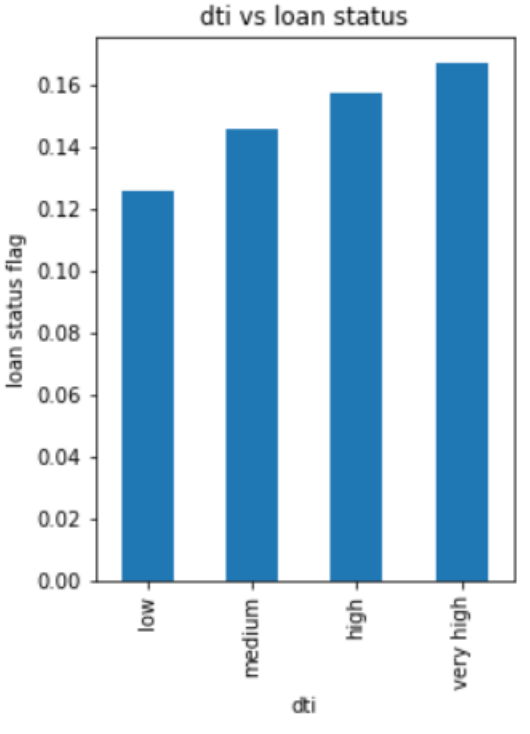




- The Higher the incomes, less the chances of getting charged off, indicating an inverse relation between the annual income and loan status.
- Direct relation between the loan interest rates and the loan status i.e. Lower the interest rates lesser the chances of charged off.





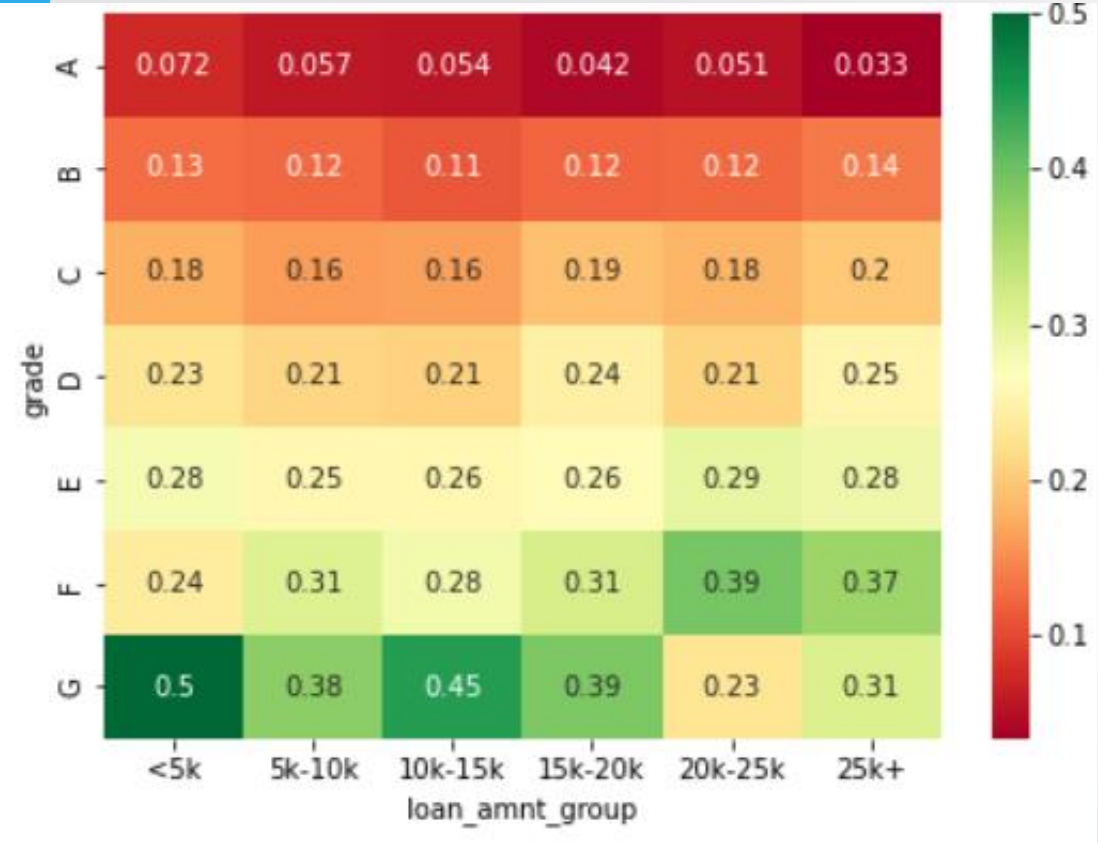


- Both installment and debt to income ratio display a direct relation with loan status.

4.3. Multivariate Analysis

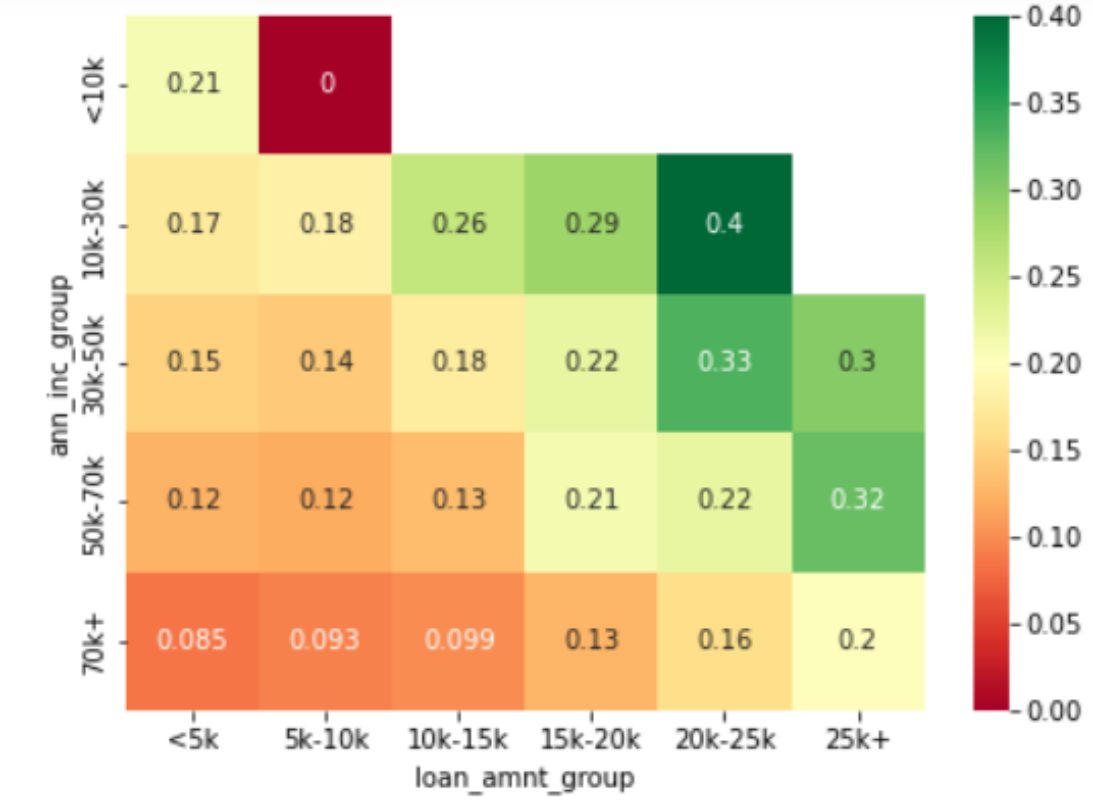
- Various heatmaps have been plotted with the help of pivot tables between different combination of variables. The most prominent results are as follows:

Grade v loan amount v loan status flag

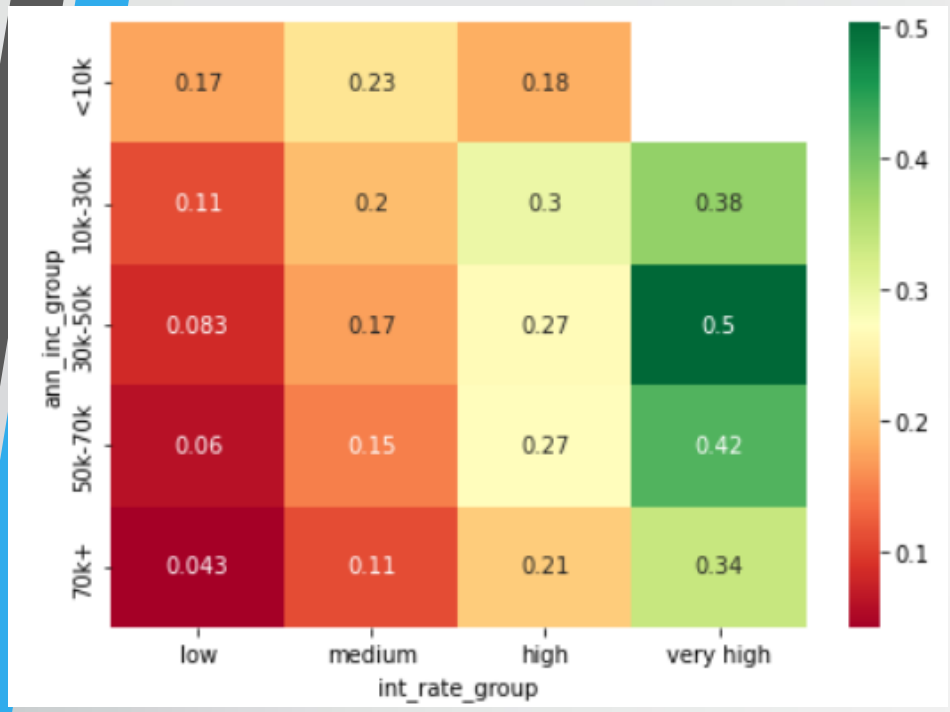


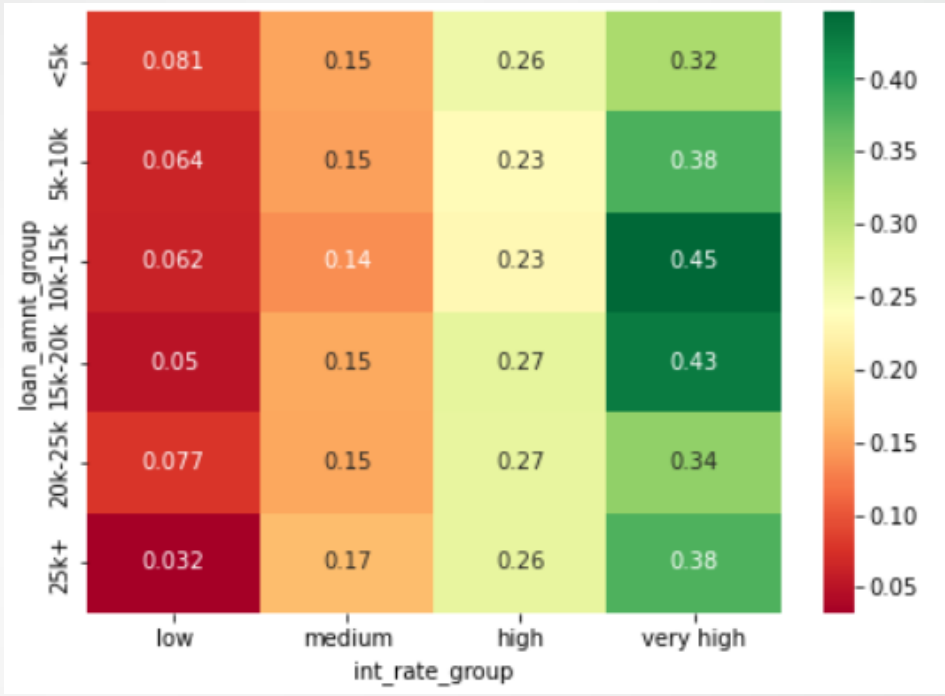
- This plot shows that higher grade along with lower amount of loans have minimum risk of getting defaulted.

Annual income v loan amount v loan status flag



- Both Higher income and lower income ones along with low amount of loans have minimum risk of getting defaulted.





- Both plots together indicate that lower income along with higher amount of loans with higher interest rates have a high risk of getting defaulted.

Conclusions

- The analysis shows that the ideal borrowers for loan recovery are those with:
  - Higher annual income
  - Low amount loans
  - Low interest rate
  - Low installment
  - Low record of bankruptcy
  - Low debt to income ratio