Lending Club Case Study

Exploratory Data Analysis

Meher Rupasri
Zurara Hasan

Problem Statement

Problem:

• You work for a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated 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

Business Understanding:

• When a person applies for a loan, there are two types of decisions that could be taken by the company:

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 installments, i.e. the tenure of the loan is not yet completed. These candidates are not labeled as 'defaulted'.
- **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

Loan rejected: The company had rejected the loan. Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company

Data Summary:

• Loan.csv file contains 39717 rows and 111 columns.
• There are two types of attributes Loan Attribute and Customer attributes.

Data Cleaning:

• There were 55 columns which is having all the rows values as null/blank and doesn’t participate in analyze has been removed.
• There were no duplicates rows found.
• There were 1140 rows present of loan status=‘current’ which has been deleted as loan status =‘current’ doesn’t participate in analysis.
• Some columns have the same value throughout. These columns don't provide valuable insights for our analysis, so we'll go ahead and remove them.(Ex: pymnt_plan,initial_list_status, policy_code)
• And we have 48 columns out of which some correspond to the post approval of loan, so we can delete them.
• Missing values: columns with missing values are "emp_length", "revol_util“ should be handled.

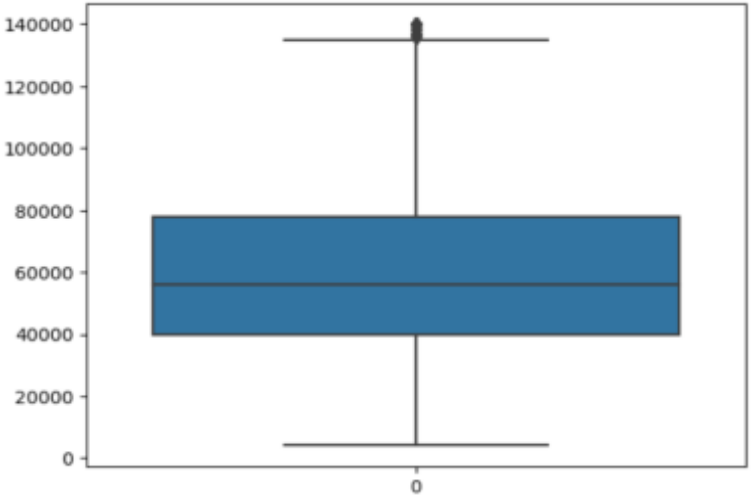
Data Standardizing:

• Column revol_util although described as an object column, it has continuous values.
• Convert int_rate from string to float.
• Changing format of emp_length column data

Univariate Analysis

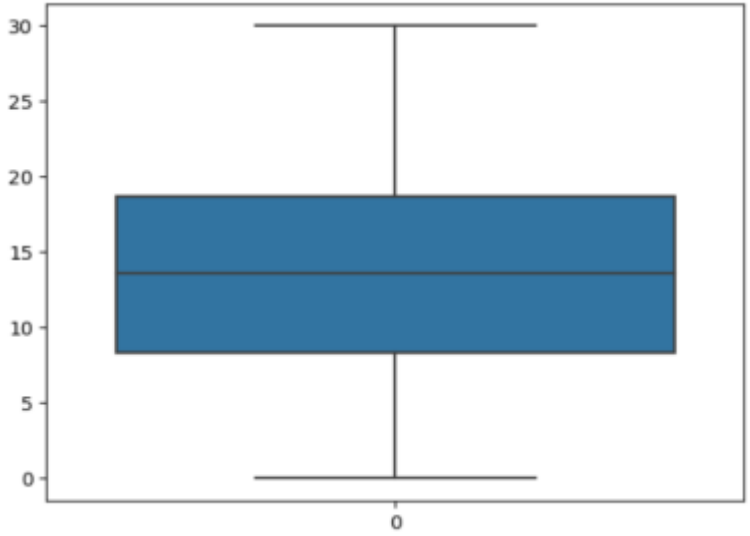
- **Fig1: annual_inc**

The Annual income of most if applicants lies between 40k-80k.
Average Annual Income is : 59883.0



- **Fig 2: dig**

The dig of most if applicants lies between 9 and 20
Average dig is : 14



Univariate Analysis

Fig 3: loan_amnt
Most of the loan amount applied was in the range of 5k-14k.

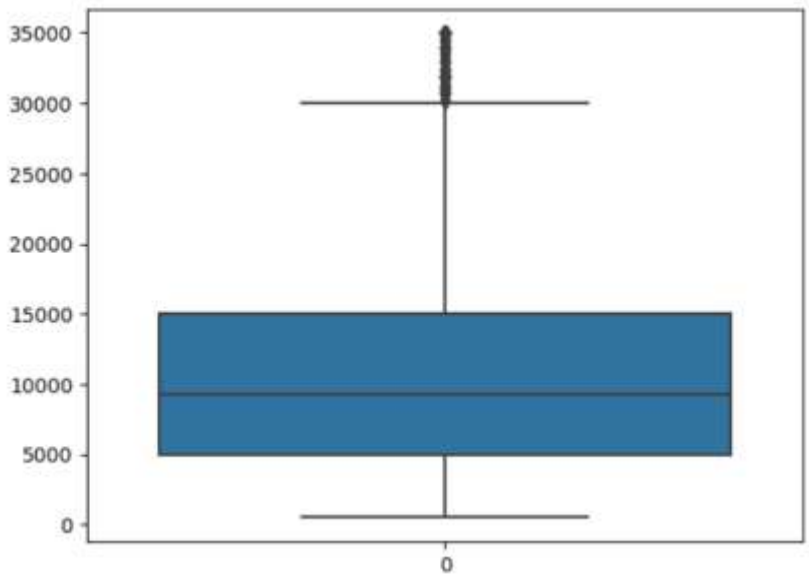
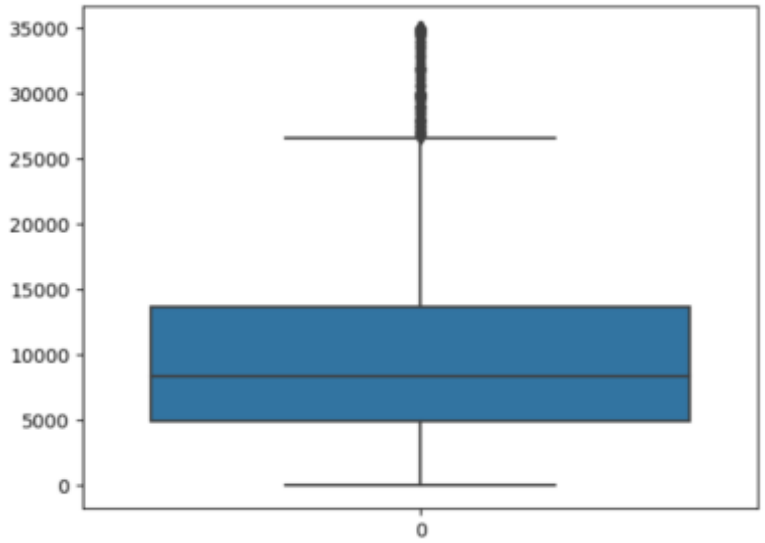


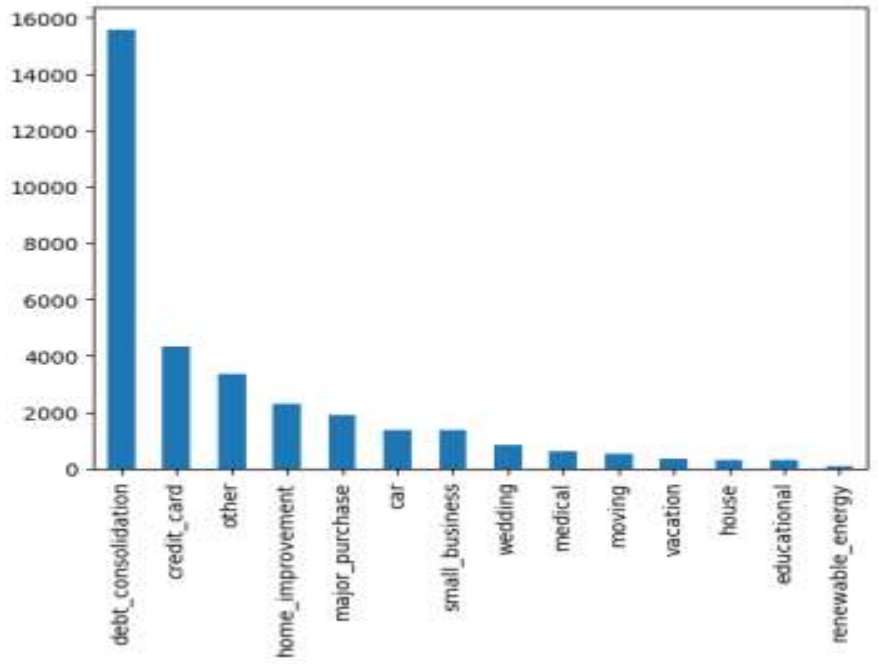
Fig 4: funded_amnt_inv
Most of the funded_amount_inv applied was in the range of 5k-14k

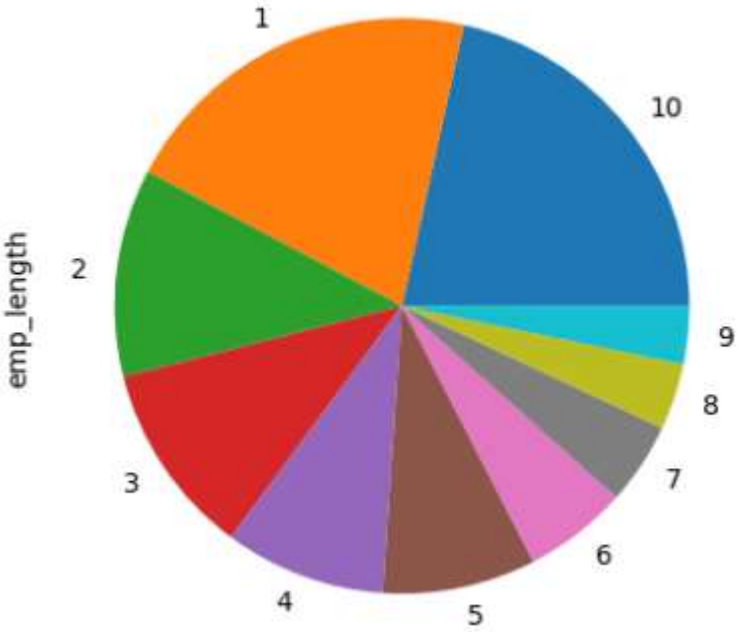


Unordered & Ordered Categorical Variable Analysis

Observations:

- Majority of loan applicants are either living on Rent or on Mortgage
- Most of loan applicants are for debt_consolidations
- Most of the Loan applicants are from CA(State).
- Most of the applications are having 10+ yrs of Exp.





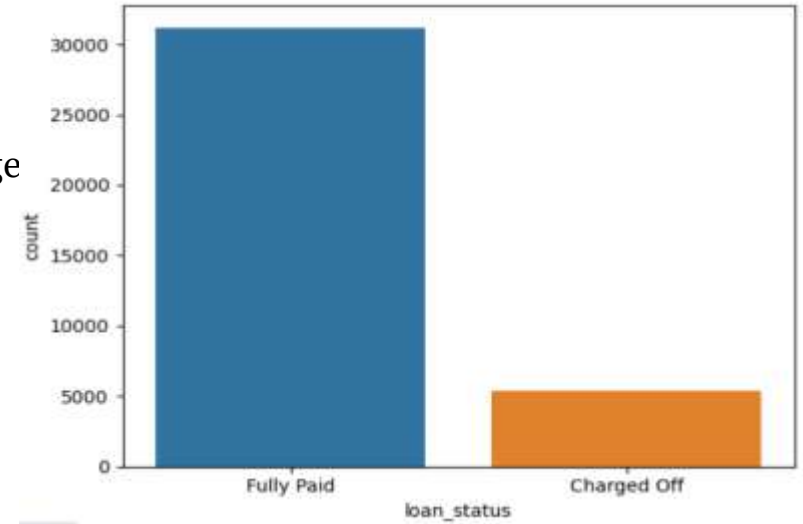
Bivariate Analysis

We are analyzing and visualizing only the defaulter data.

So subsetting the data while plotting only for 'Charge Off' loan_status for below plots

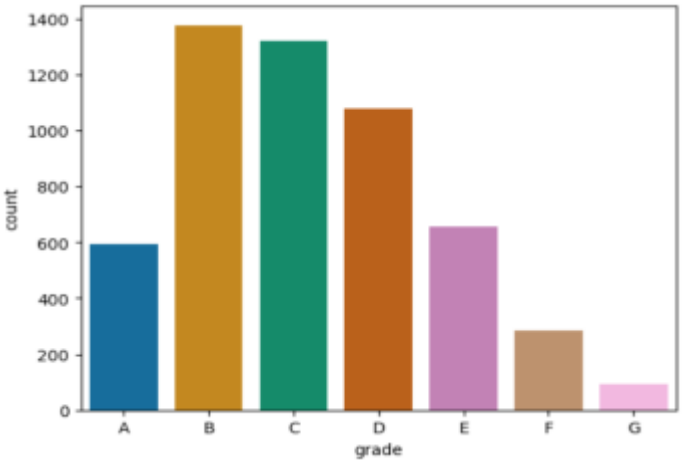
- **loan_status vs count**

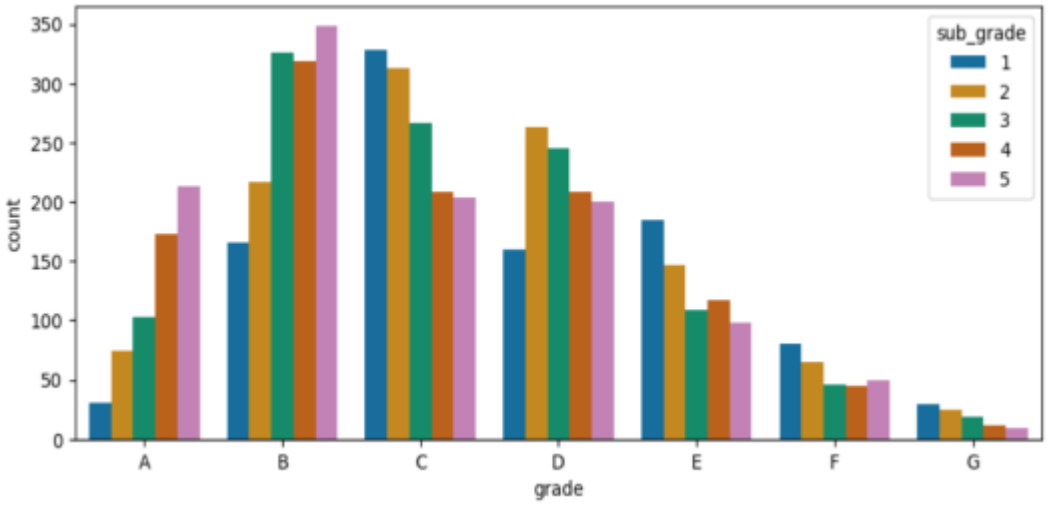
Applicants who are fully paid is having higher count than Charged off loan.



- **Grade vs count**

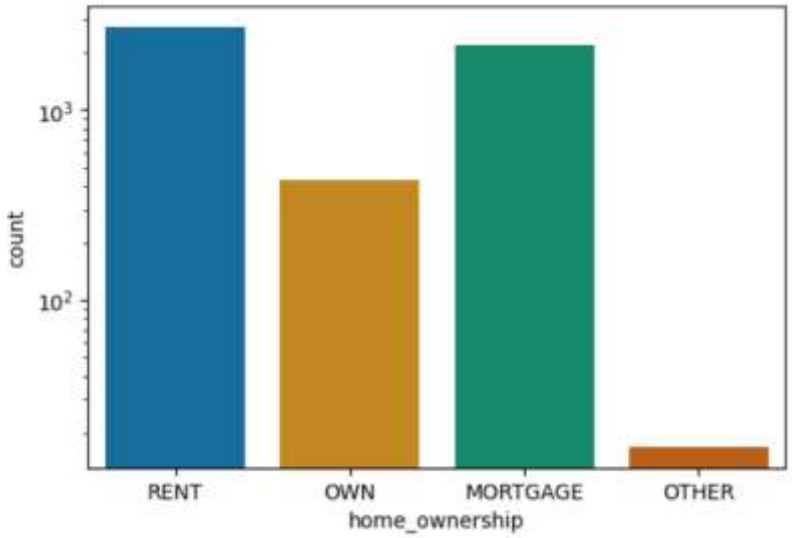
It Is identified that, customers with grade B,C and corresponding subgroups are having higher count.

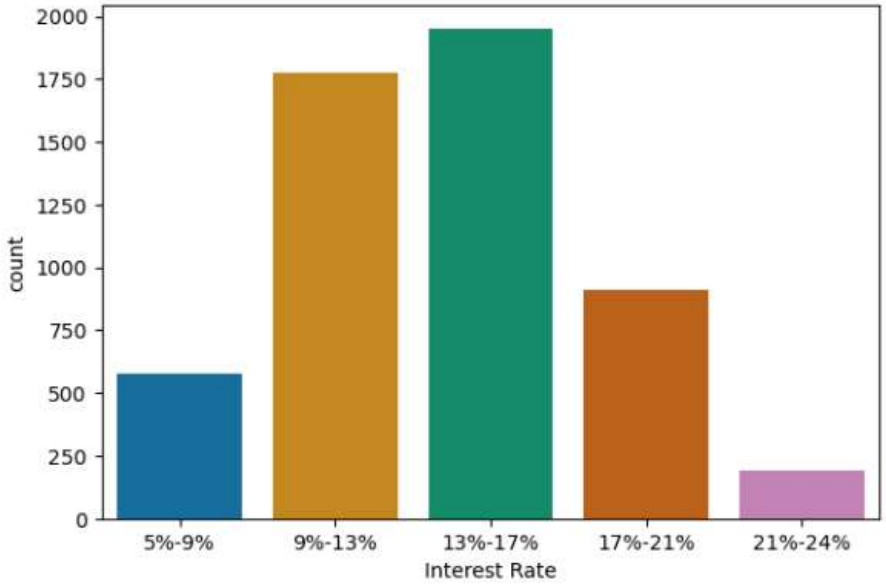




Observations:

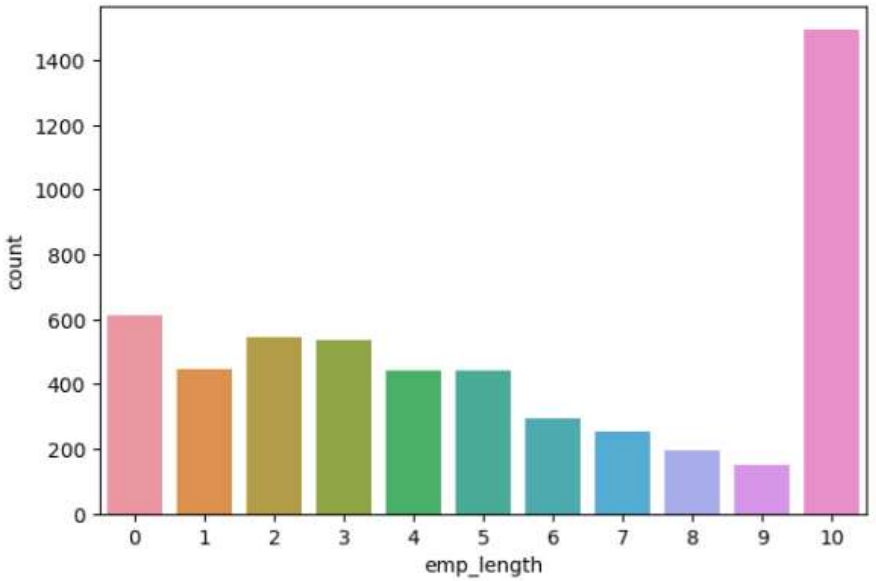
- It is observed that many of the customers are having Montage/Rented house
- And it is also observed that the highest interest rate is 13to 17 percentage

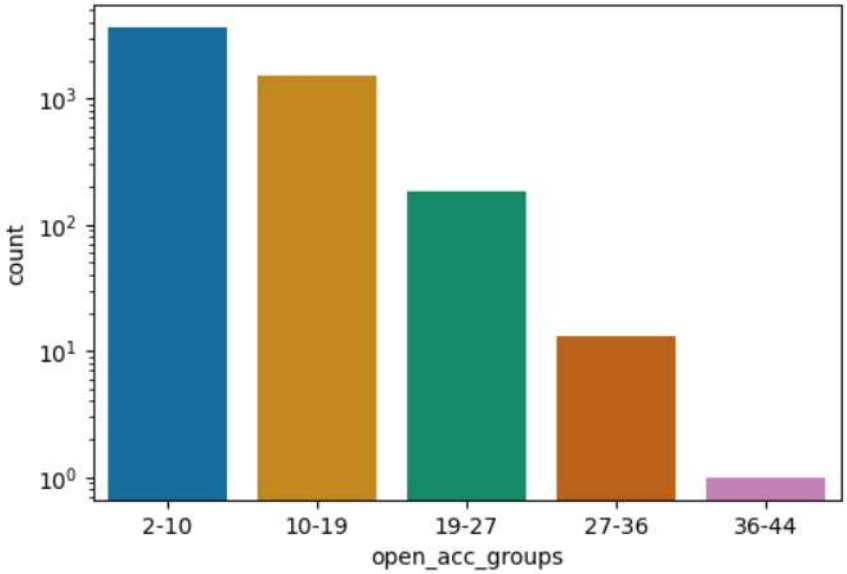




Observations:

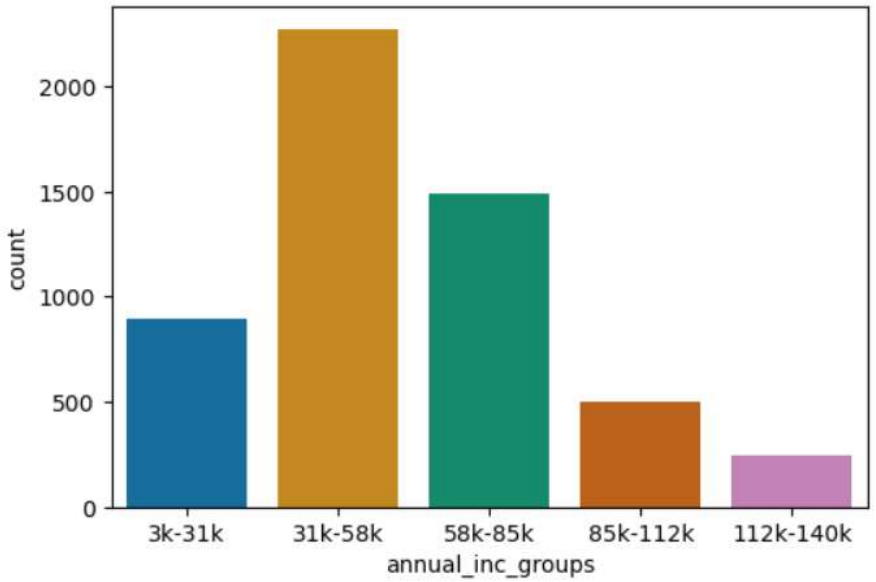
- We can see that the employee length is really high for 10
- It is observed that open account groups are gradually decreasing when count of group is increasing.

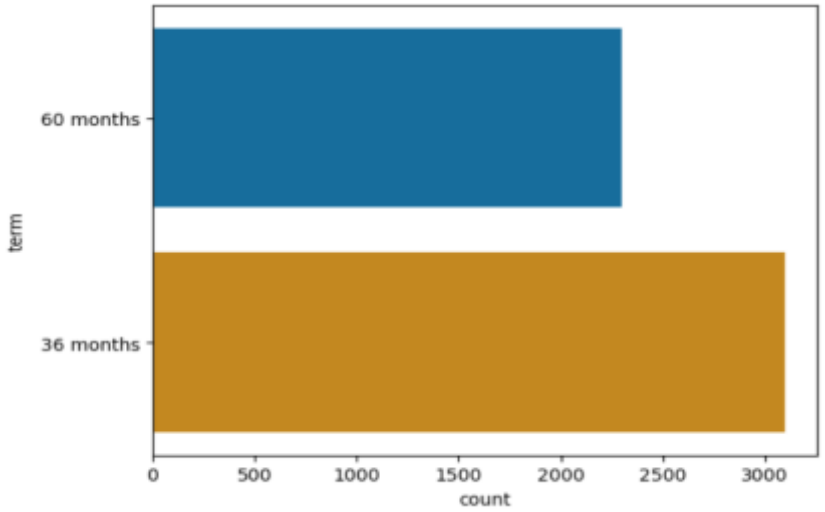




Observations:

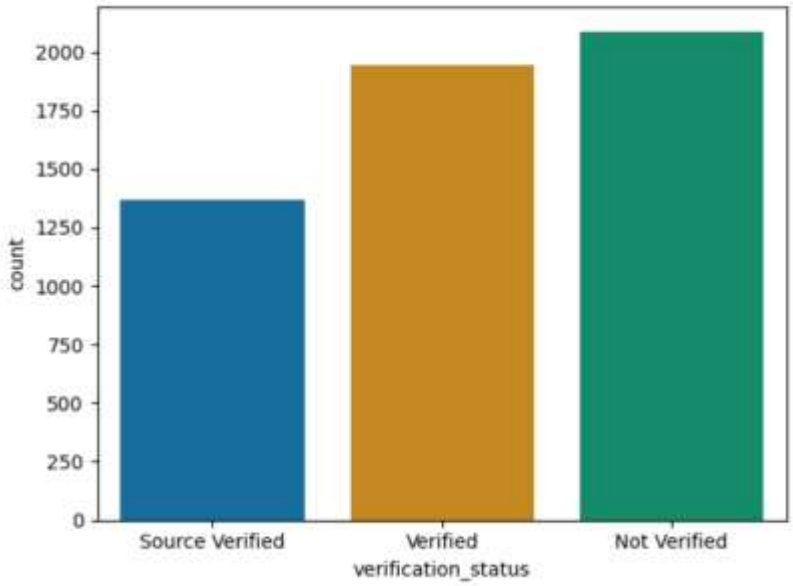
- We can say that annual inc groups are really higher for 31k to 58k
- It is also observed that many of the customers opted 36 months term

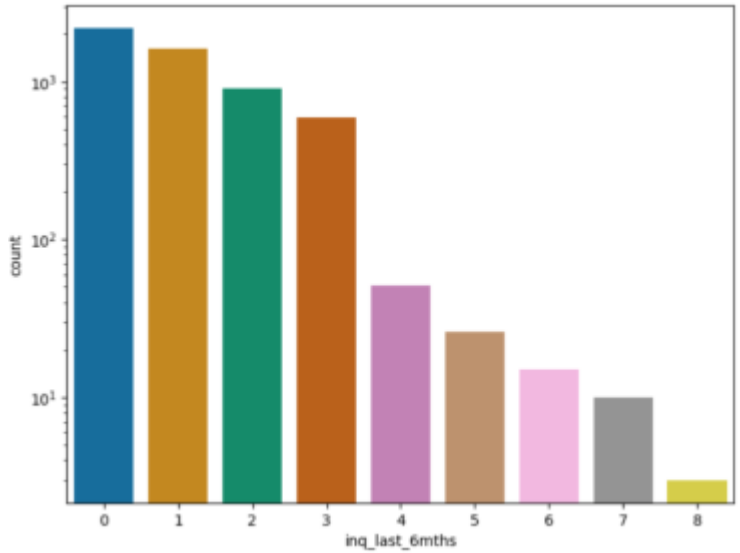




Observations:

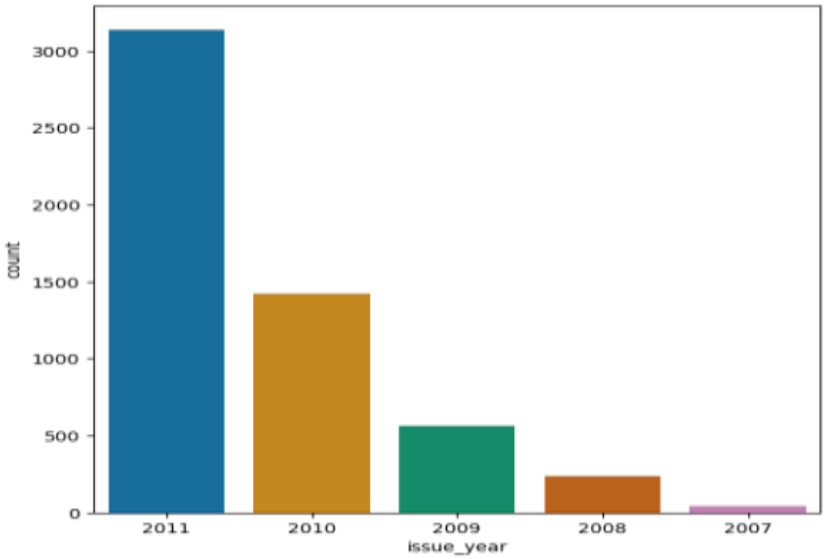
- It is observed that verification status of defaults are higher for not-verified
- It is observed that loan status of charged off defaults are higher for initial months

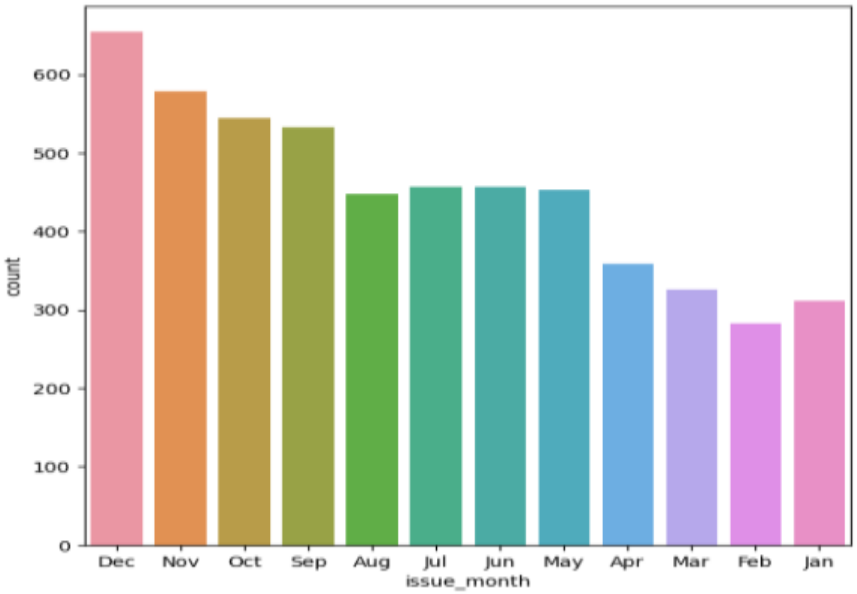




Observations:

- Year 2011 is highest loan defaults.
- 2007 is having lowest loan defaults
- Month December is highest for loan defaulters
- Month February is highest for loan defaulters

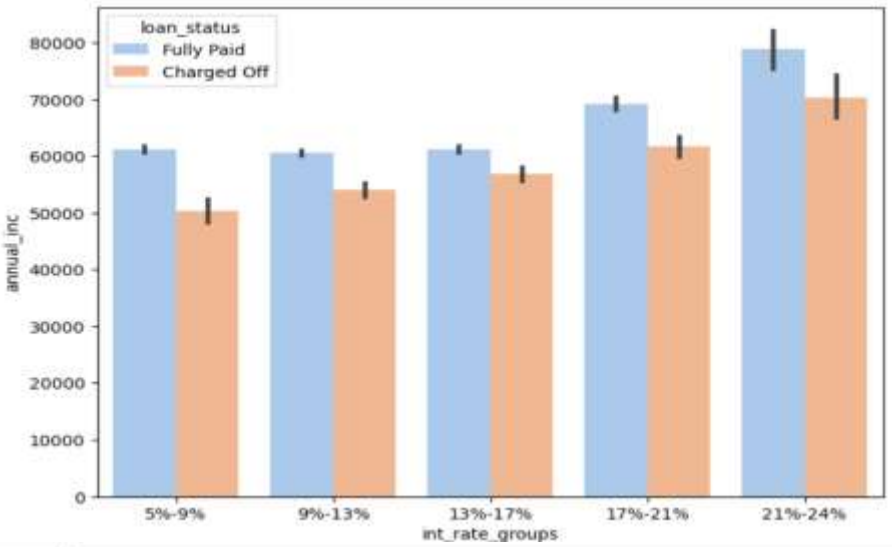


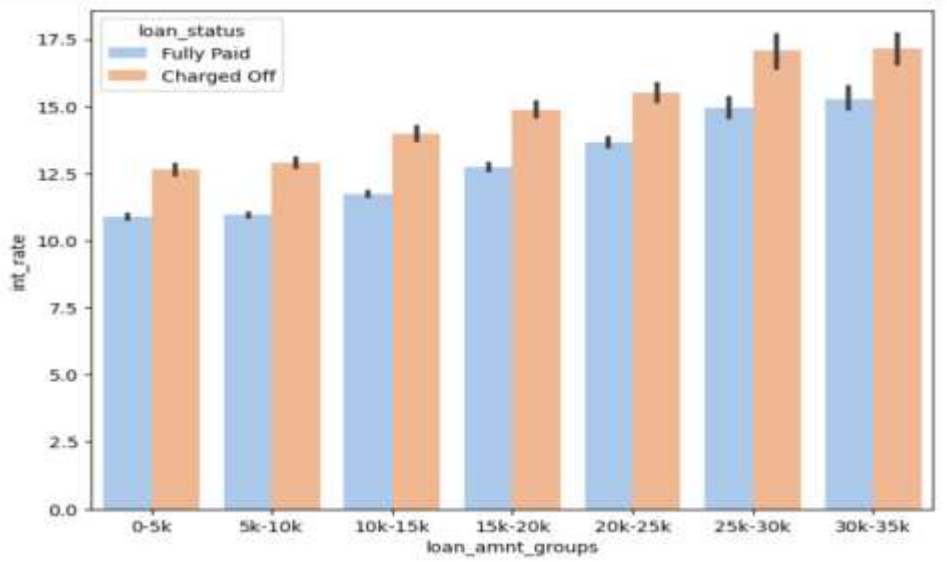


Group data analysis

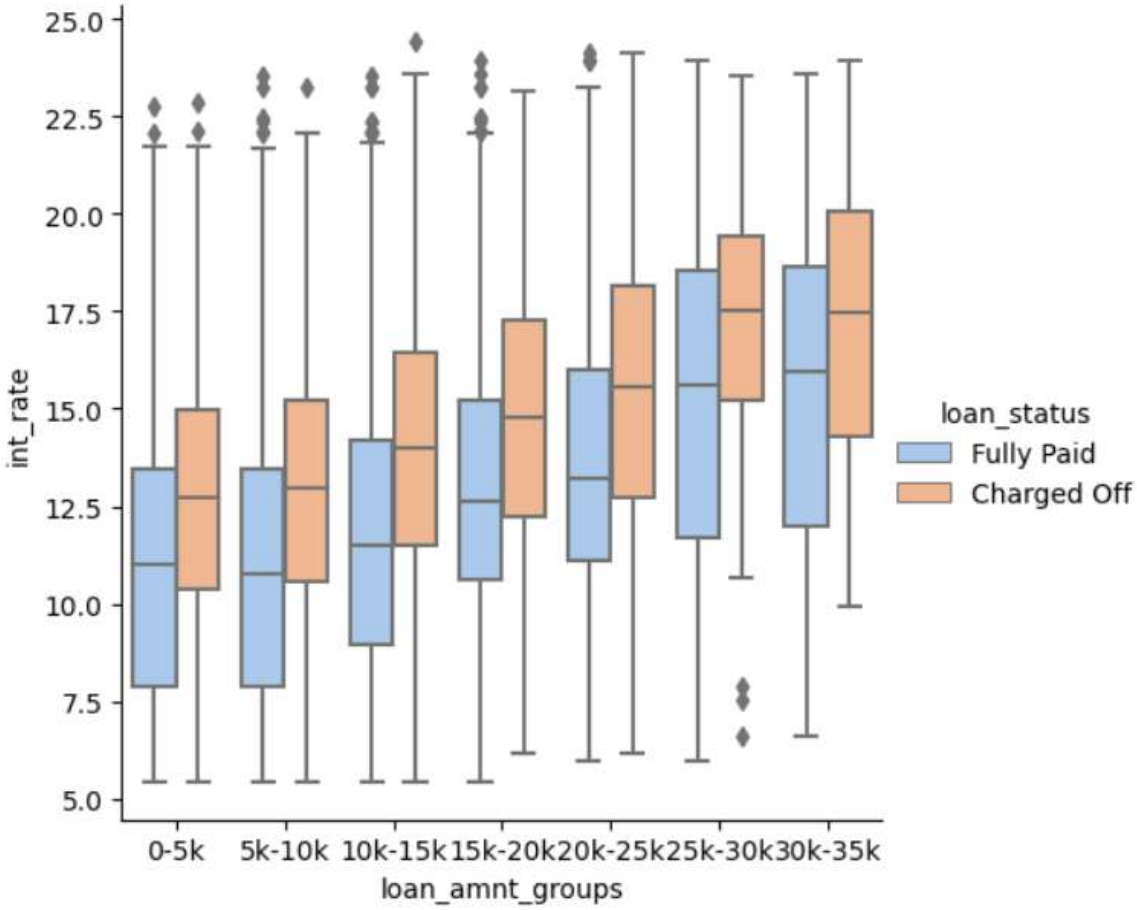
Analysing loan_amount with other columns for more insights.

- Across all the income groups, the loan_amount is higher for people who defaulted.
- It is observed that loan amount groups are higher when interest Rate is between 30 to 35k.





Looking at the verification status data, verified loan amount groups tend to have higher loan amount. Which might indicate that the firms are first verifying the loans with higher values.



Observations

- The above analysis with respect to the charged off loans. There is a more probability of defaulting when :
- Applicants seeking a loan for 'home improvement' with an income between 60k and 70k.
- Applicants owning a home with 'MORTGAGE' status and an income falling within the 60k-70k range.
- Applicants subjected to an interest rate between 21% and 24%, with an income ranging from 70k to 80k.
- Applicants securing a loan in the range of 30k to 35k, facing an interest rate between 15% and 17.5%.
- Applicants obtaining a loan for a small business with an amount exceeding 14k.
- Homeowners with 'MORTGAGE' status and a loan ranging from 14k to 16k.
- Instances where the loan grade is 'F,' and the loan amount falls between 15k and 20k.
- Applicants with a job tenure of 10 years seeking a loan between 12k and 14k.
- Loans that are verified and have an amount exceeding 16k.
- Cases involving grade 'G' loans with an interest rate more than 20%.