Lending Club Case Study

Group Facilitator Name: **Mallu Shivanna**

Other member of the Group: **Sandeepa D**

Problem Statement

- Lending club is a market place for personal loans that matches borrowers who are seeking loan with investors who are looking to lend money and make returns

Money flow diagram

- Once the request has come to the Lending club from Borrower for money, the Lending club should check following criterion

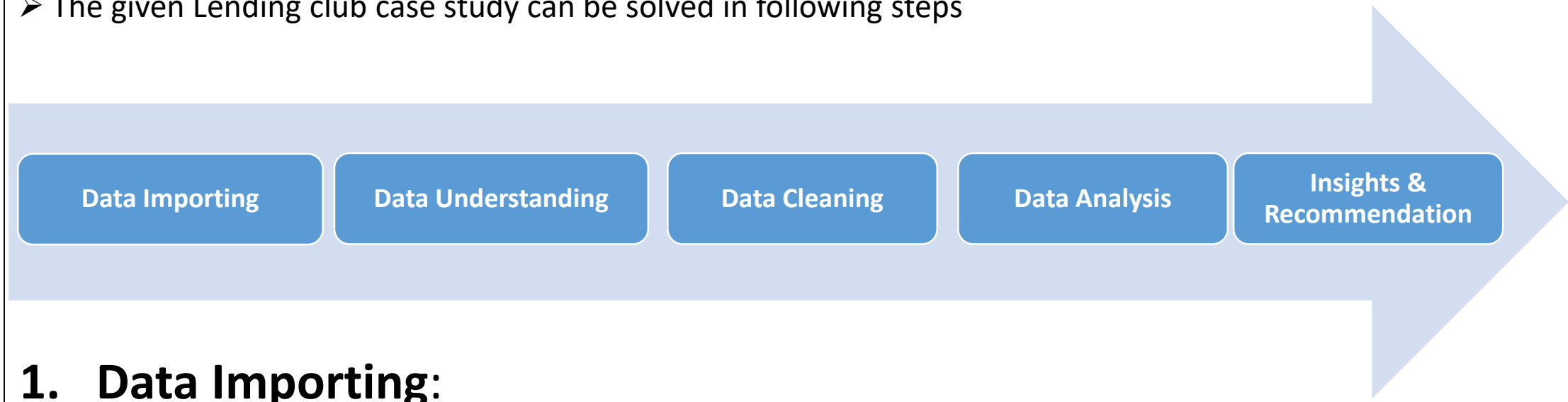
  a. **Borrower is likely to pay the money back with interest**
  b. **Borrower is likely not to pay the money back i.e. Default**

- After confirmation that he is likely to pay back the money, Lending club can issue him the Loan

- By this methods Credit loss to the Lending club can be avoided

Data Importing

- The given Lending club case study can be solved in following steps



1. Data Importing:

- We can use Numpy and Pandas Library to deal with the data set.
- To Plot the values we can use Seaborn and Matplot library
- Given CSV file is imported as Old Loan data.

Data Understanding and Cleaning

2. Data Understanding: By Manual inspection of data set, it is found that

- Some columns have all the Values as NA (Null values)
- Some Columns have all the Values as a Single Value
- Some Columns have missing values

Above types of data needs to be cleaned before going to further analysis so lets clean the data in the next step

3. Data Cleaning: It is done in 5 steps

a) **Column treatment**

b) **Row treatment**

c) **Missing value treatment**

d) **Standardization of Data**

e) **Outliers treatment**

Data Cleaning

a) **Column treatment:** Number of columns before column treatment=111

- Columns having null Values and Single values do not contribute to draw useful insights so they can be removed. (Number of columns after this step=48)

- Note: there are 4 different types of Attribute categories observed in the data given

- **Customer Attributes**: emp_title, emp_length, Home ownership, annual_inc,Verification_status,Purpose,title,addr_state, open_acc,pub_rec,total_acc

- **Loan Attributes:** loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, issue_d, loan_status, revol_util,

- **Bank Generated attributes** : id, member_id, url, desc, zip_code, dti, earliest_cr_line, inq_last_6mths

- **Post loan approval attributes** : delinq_2yrs, revol_bal, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv, total_rec_prncp, total_rec_int, total_rec_late_fee,recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, next_pymnt_d, chargeoff_within_12_mths, last_credit_pull_d,

Data Cleaning

- Since post loan approval attributes do not actually determine the criteria on whether to lend the loan or not, and hence these columns can also be removed. Also some bank generated attributes like id, member_id, url, desc, zip_code: customer attributes like emp_title,addr_state does not give much insight for default determination thus we can remove these columns as well
- Upon close observation of the columns funded_amnt and funded_amnt_inv it is found that what we need for analysis is ' total amount committed by investors', and hence 'funde_amnt' column can also be removed. The loan 'title' provided by the borrower is similar to 'purpose', we can keep only purpose column and delete the 'title‘
- **Finally after all possible column treatment we are left with 21 columns for our analysis**

Data Cleaning

b) **Row treatment:** Since our objective is to find who is likely to default , the data on **current** payments is not useful to draw any conclusion and hence we can remove the rows with current payment details

c) Missing Value treatment:

- Using panda functions It is found that only emp_length and revol_util has missing values
- Missing values can be assigned with the most repeated value in the column since the percentage of missing values found to be very less. (most repeated value can be found using **mode** function)

d) Missing Value treatment

- It is found that the columns 'int_rate' and 'revol_util' is represnting % values , however data set is a combination of int and string type, we can remove the % symbol.

Data Cleaning

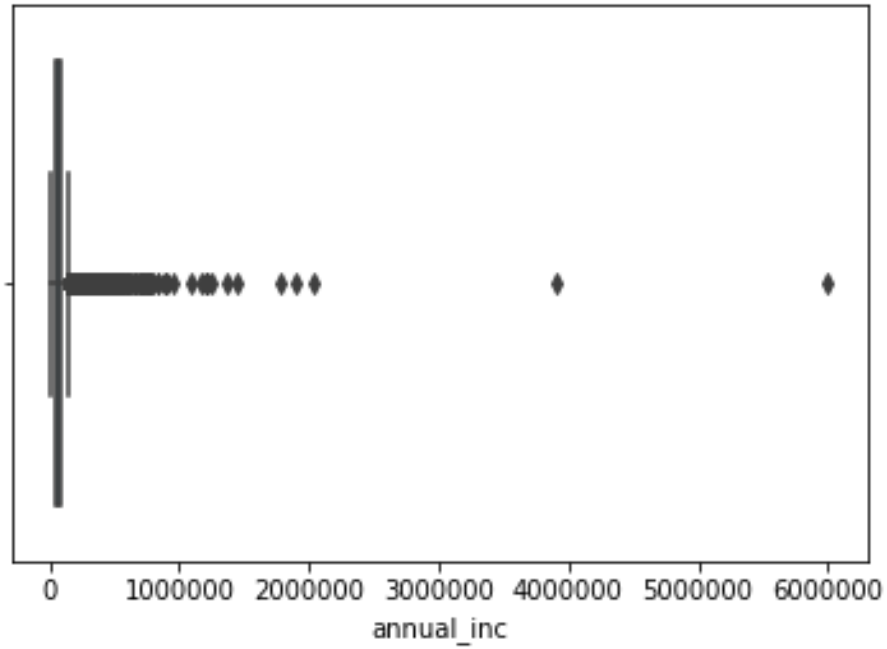
- The columns 'term', 'sub_grade' and 'emp_length' is a combination of text and int value, for 'sub_grade' and 'emp_length' we can separate the string values and only keep the int, however the column 'term' which only has 2 values 36 month and 60 month we don't need to remove string.

e) Outliers treatment: It is required to check outliers for the following columns

- loan_amnt, total_acc, revol_util, inq_last_6mths, annual_inc, funded_amnt_inv, open_acc, int_rate, installment, dti
- From quantile we found that values are continous for most of the columns , however for the columns 'anual income' after 95% the values are going out of range, we can visualize this using box plot

Data Cleaning

Annual_inc box plot



- From the box plot it is clear that outliers exist in annual income and it must be treated, and also after seeing quantile distribution variation is observed after 95% , we can remove the values after 95%.

➢ That ends the Data cleaning. It will be followed by Data analysis

Data Analysis

- **4. Data Analysis**: We have the following category of variables

➢ **Unordered Categorical Variables** : home_ownership, verification_status, loan_status, purpose,

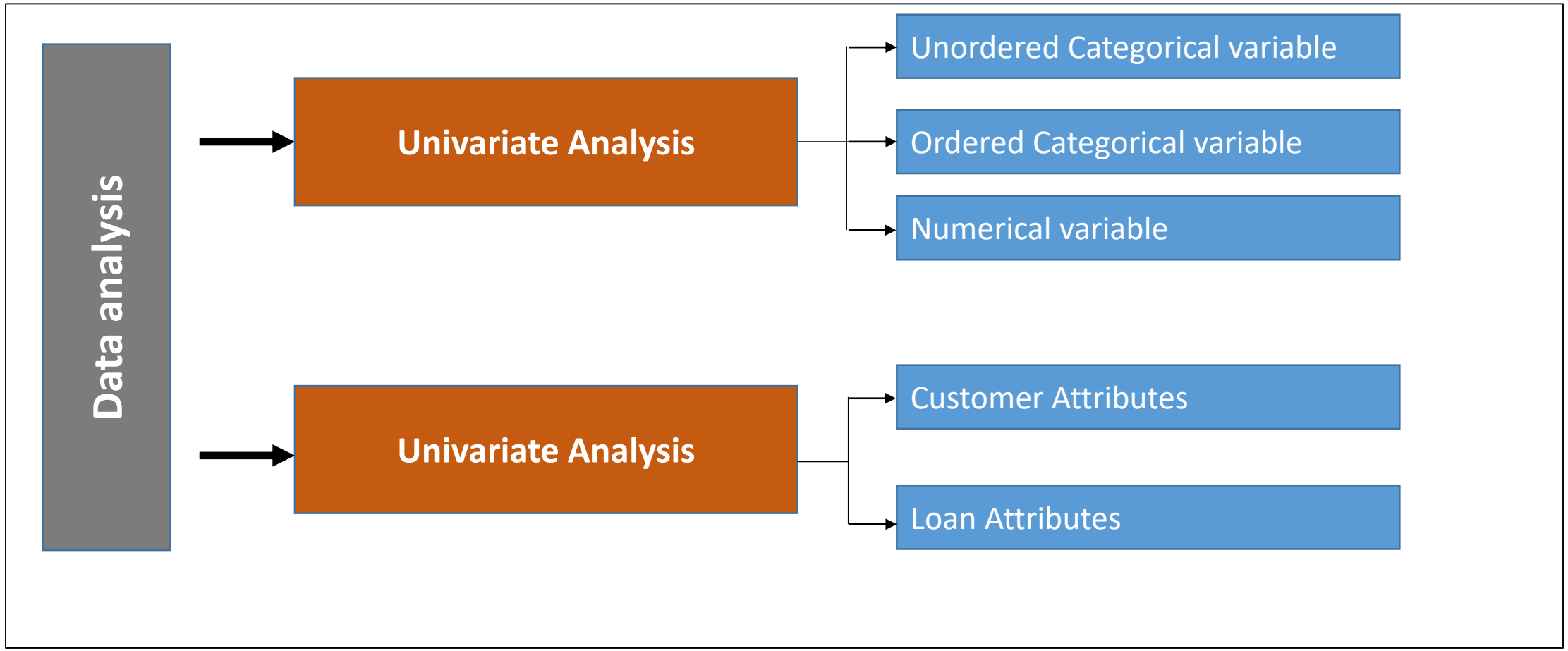
➢ **Ordered Categorical Variables** : term, grade, sub_grade, issue_d, earliest_cr_line,

➢ **Numerical Variables** : loan_amnt, funded_amnt_inv, int_rate, installment, emp_length, annual_inc, dti,inq_last_6mths, open_acc, pub_rec, revol_util, total_acc

➢ We perform Univariate and Bivariate analysis considering above variable

a) Univariate Analysis: Let us visualize the above category of variables and try to extract some meaningful patterns for charged off applicants

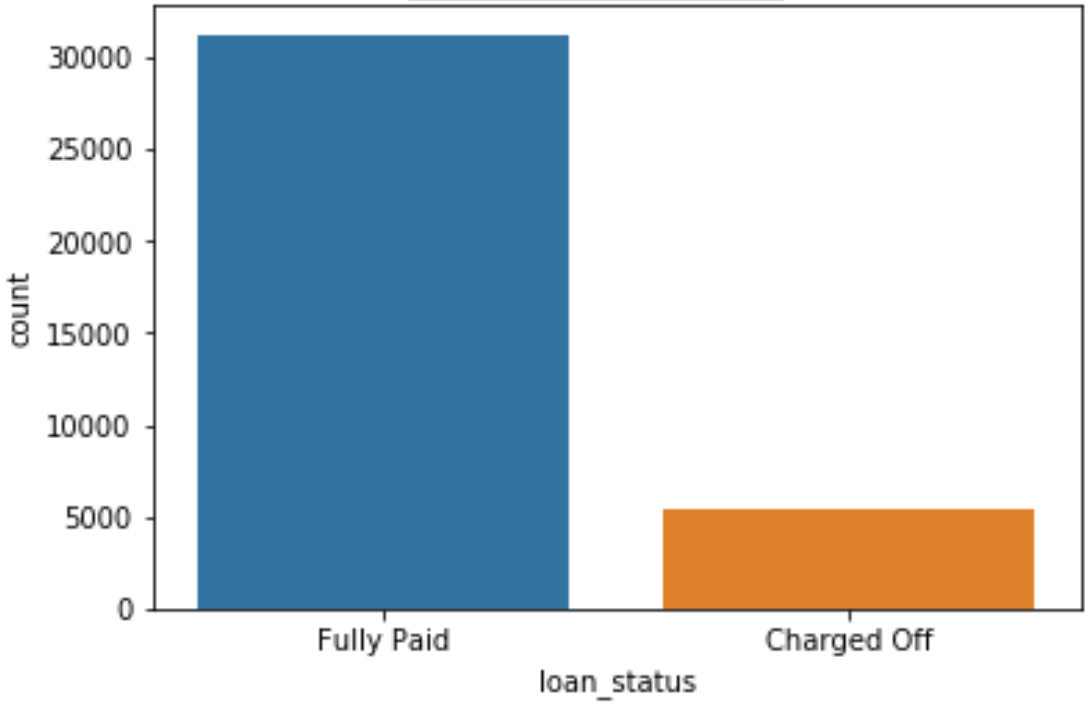
Data Analysis:



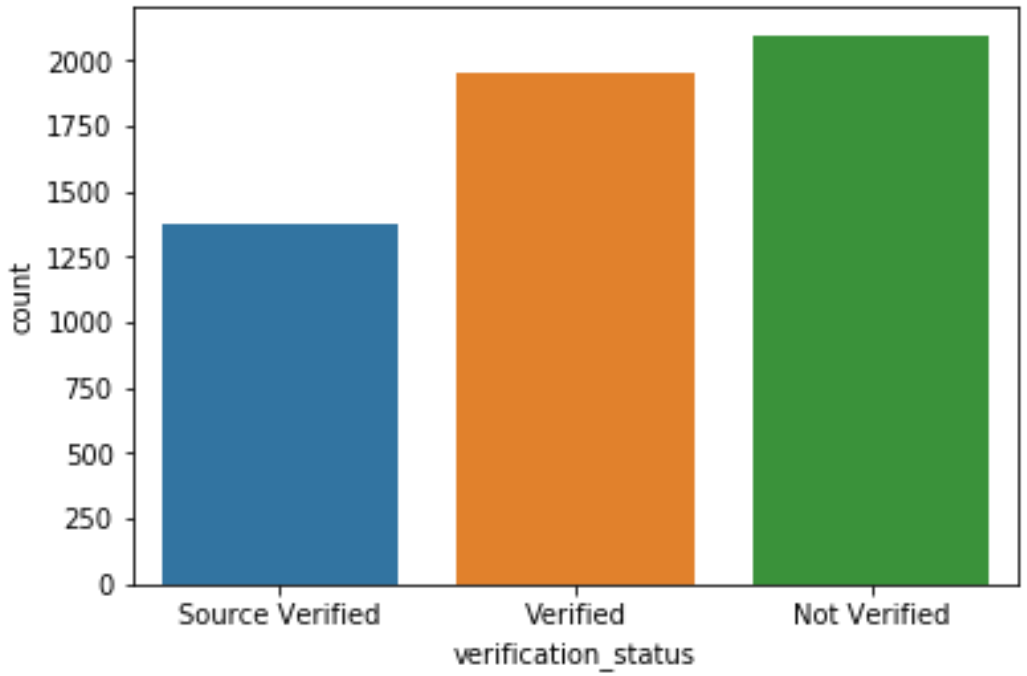
Data Analysis

1. Univariate Analysis with Unordered Categorical Variables:

Loan Status



Verification status

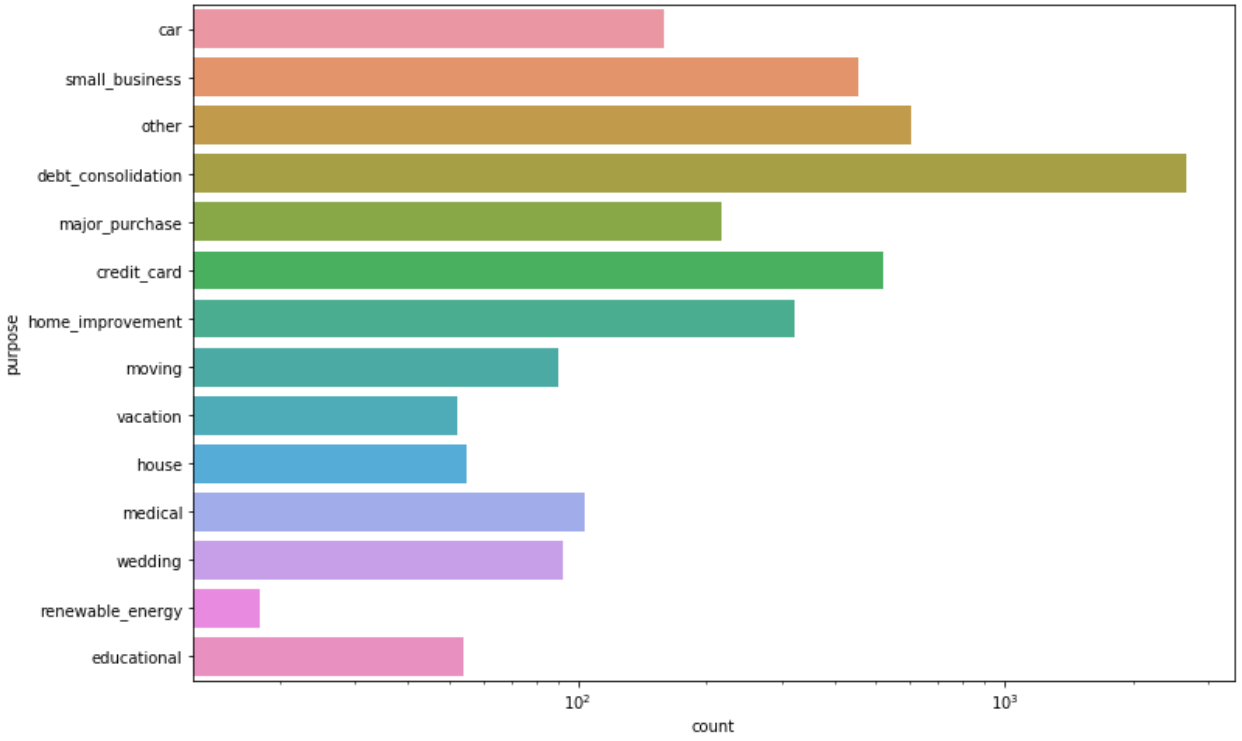


Data Analysis

1. Univariate Analysis with Unordered Categorical Variable

Home ownership

Purpose



Data Analysis

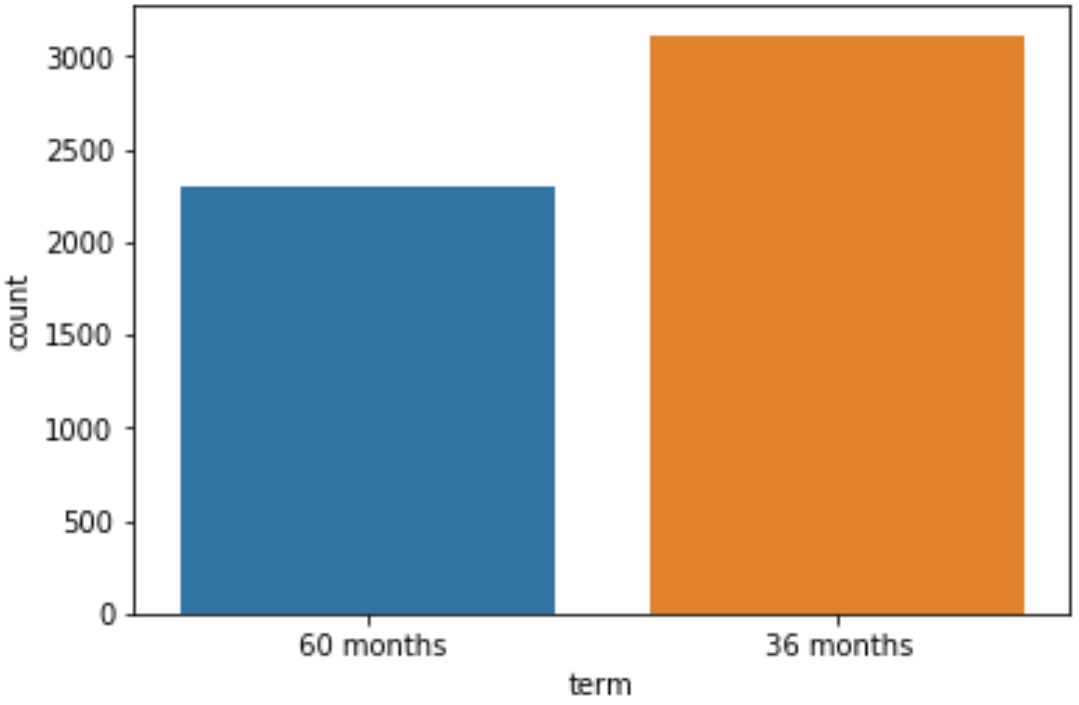
1. Univariate Analysis with Unordered Categorical Variable: Observations

- Most of the data sets are of fully paid applicants , only around 5000 applicants charged off, Analyzing the other variables for charged off applicants shows us that,
- Applicants who are not verified and whose source is not verified tend to default more, it is suggested to verify both income and income source before lending the loan
- Applicants who are living in the rented home and who mortgaged home tend to default more , it is suggested to avoid higher loans for those applicants.
- Applicants who are asking loan for debt consolidation and credit card payment tend to default.

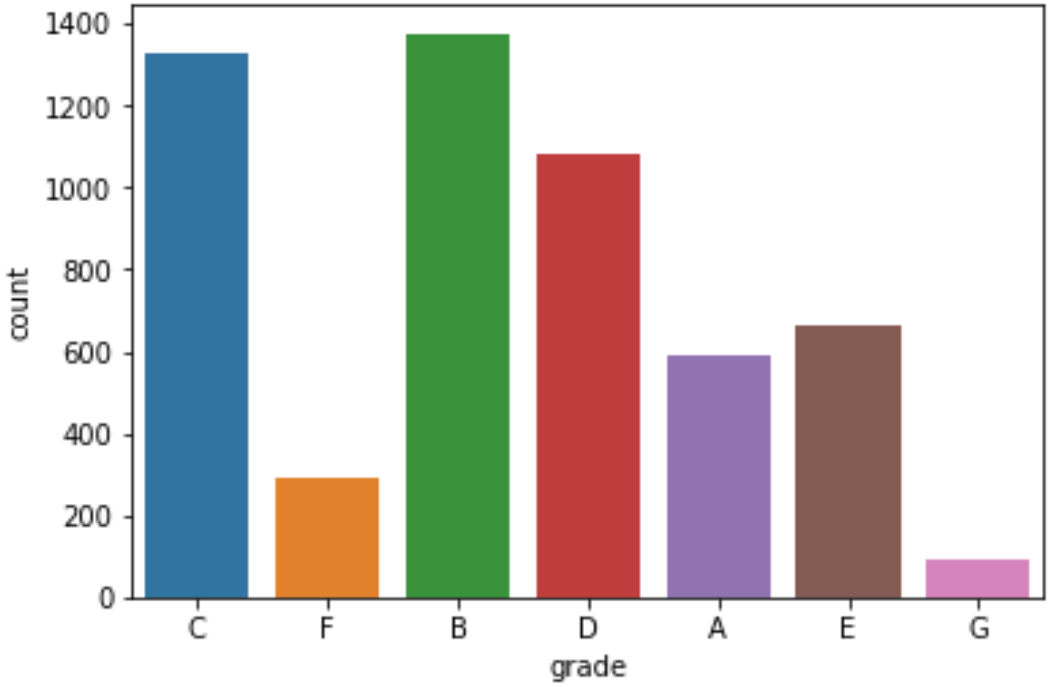
Data Analysis

2. Univariate Analysis with Ordered Categorical Variable:

Term



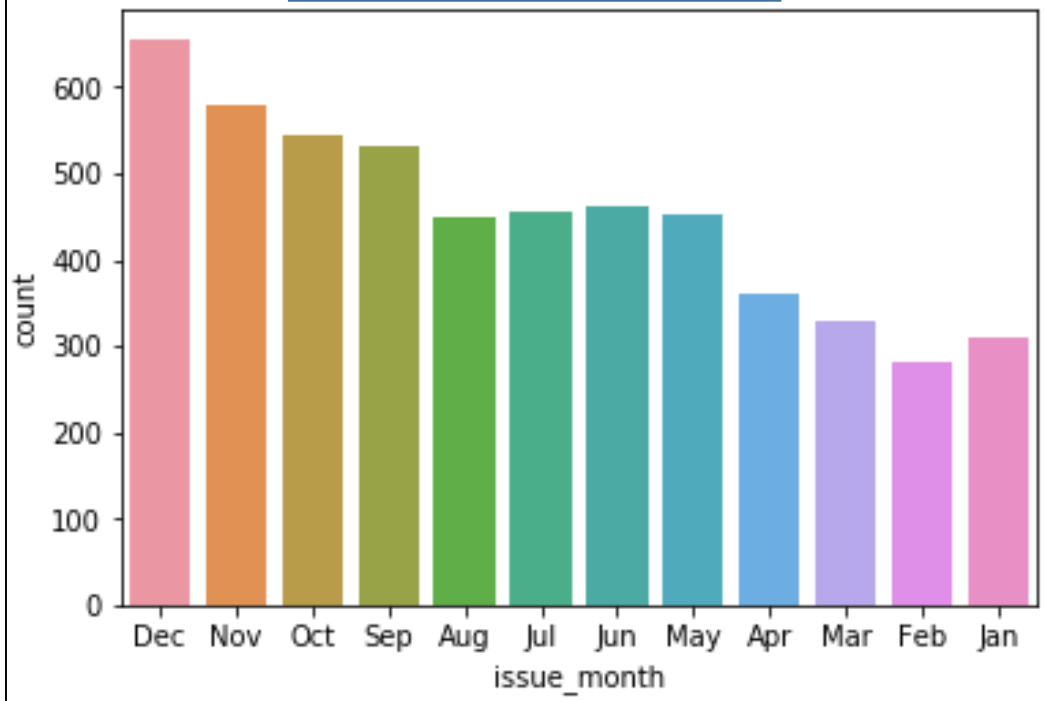
Grade



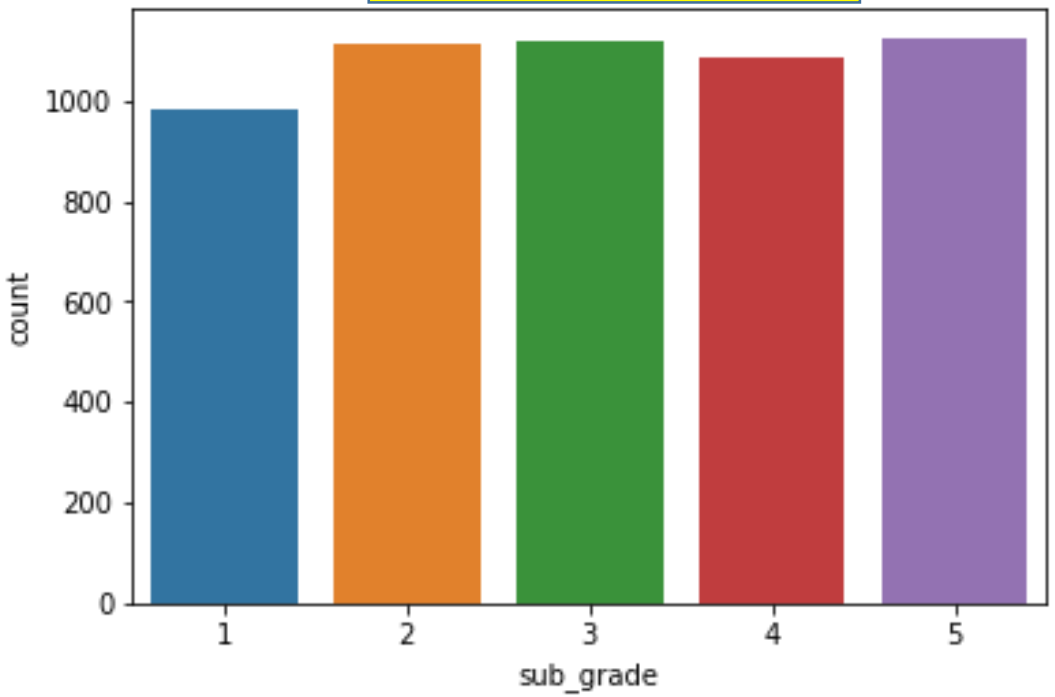
Data Analysis

2. Univariate Analysis with Ordered Categorical Variable

Issue Month

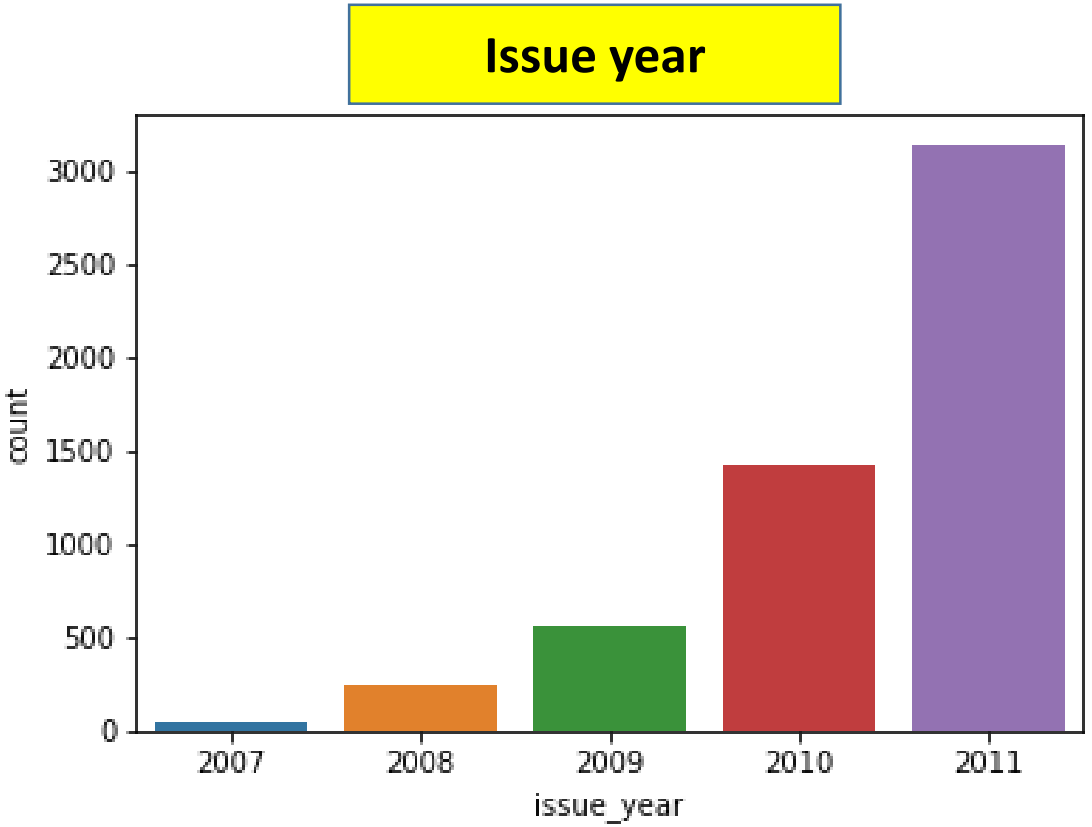


Sub grade



Data Analysis

2. Univariate Analysis with Ordered Categorical Variable: Observations



- Applicants with below attributed tend to default more
  - **Low term:** The EMI will be more so tend to fail to replay the loan
  - **B and C loan grade**
  - **Loan allotted in December:** May be due to some year end financial distress
  - **Issue year 2011:** May be due to economical crisis in the market

Data Analysis

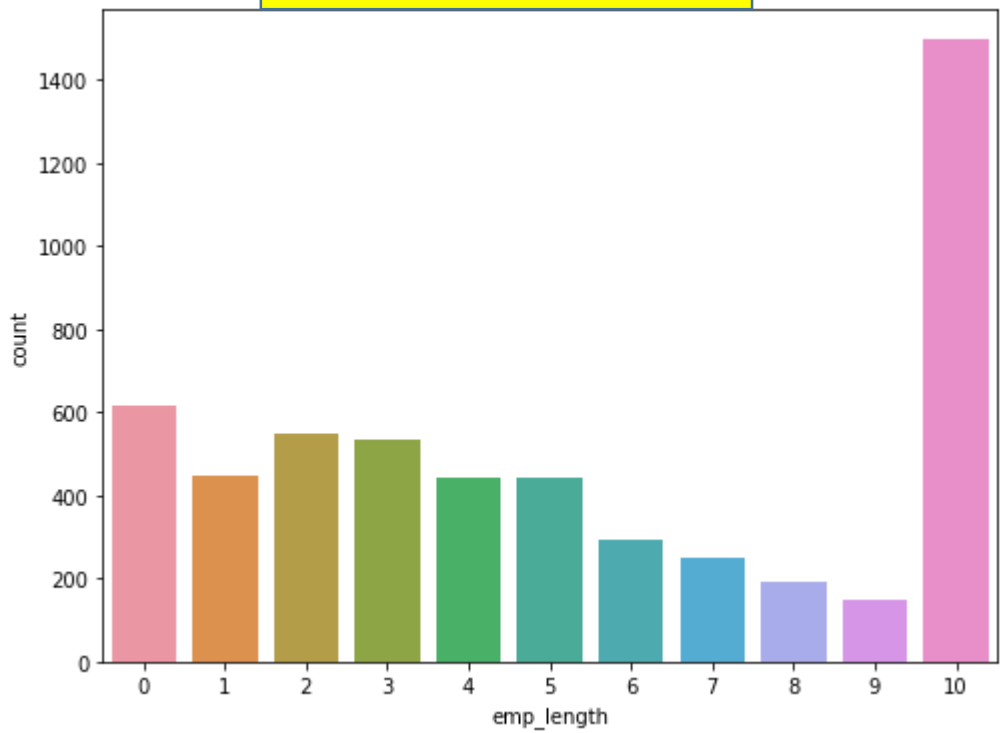
3. Univariate Analysis with Numerical Variable

- Numerical variables can be further split into 2 categories , in one such category we can directly use the numerical variable columns. In the second category we can use bins to segregate into separate groups
- **Category 1** : emp_length, pub_rec, inq_last_6mths,
- **Category 2** : loan_amnt_groups, funded_amnt_inv_group, int_rate_group, installment_groups, annual_inc_groups, dti_groups, open_acc_groups, revol_util_groups, total_acc_groups

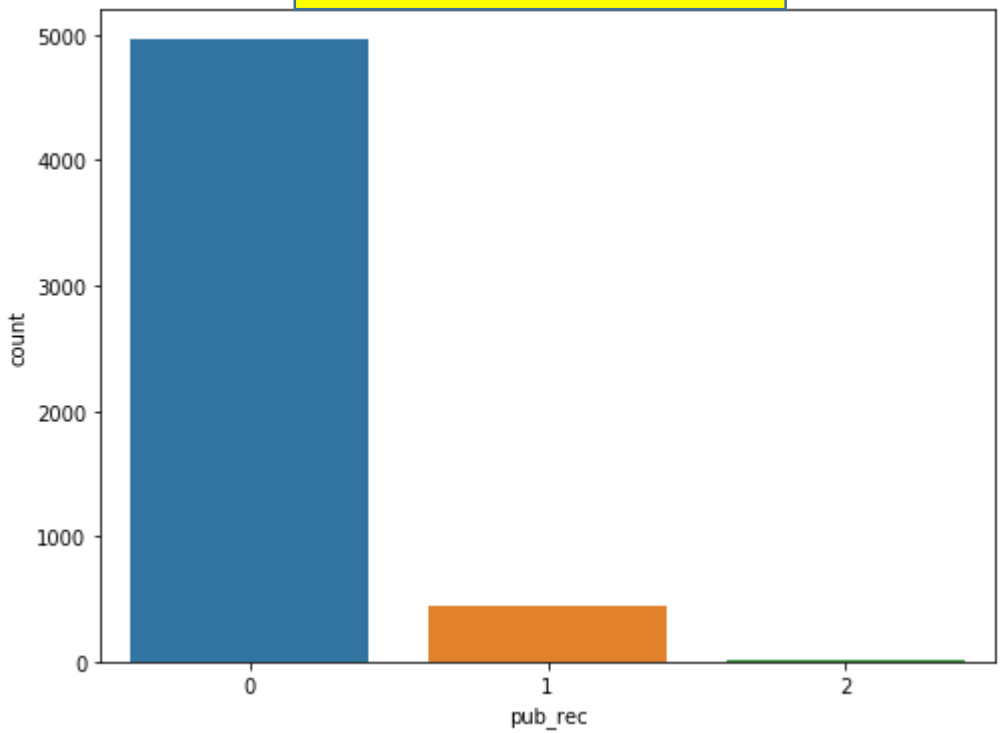
Data Analysis

3. Univariate Analysis with Numerical Variable

Employment length



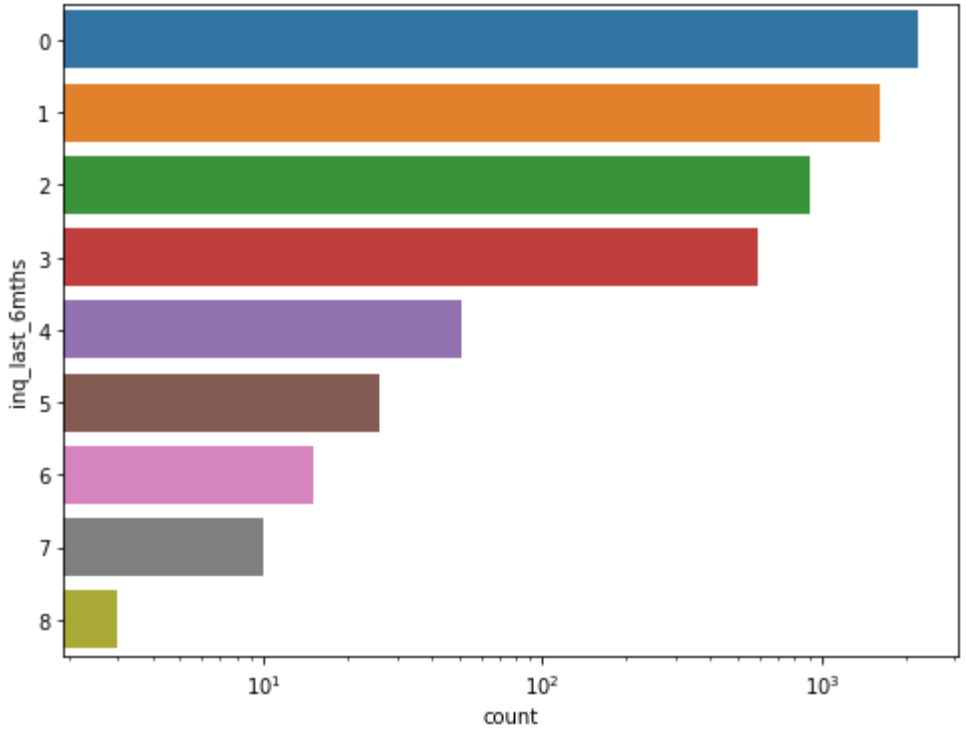
Public record



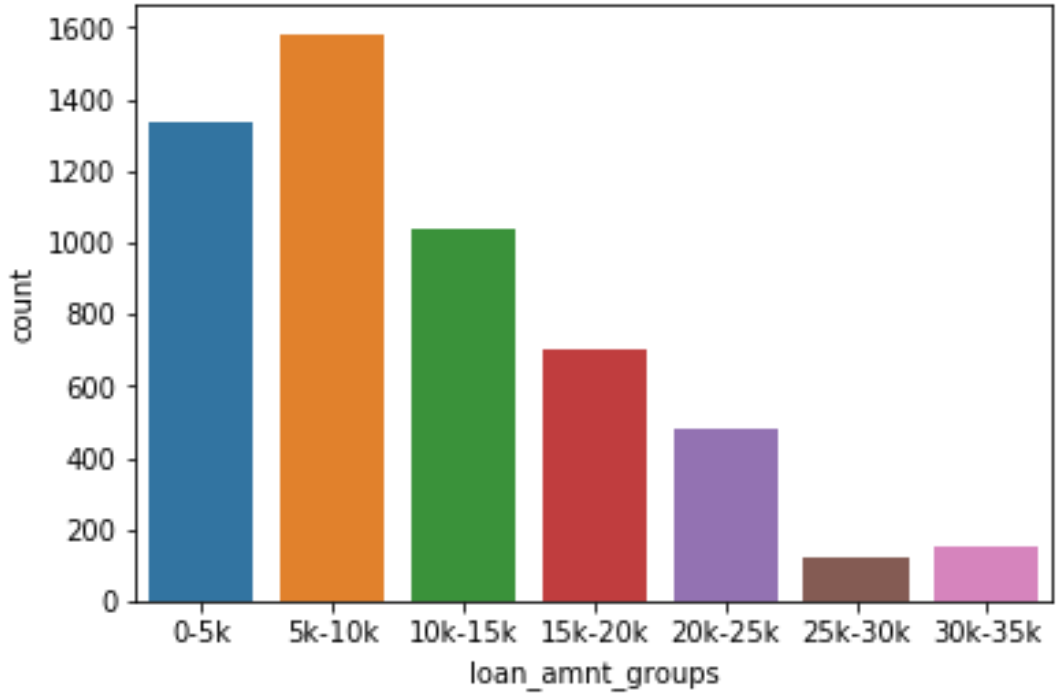
Data Analysis

3. Univariate Analysis with Numerical Variable

Inquiry in Last 6 months



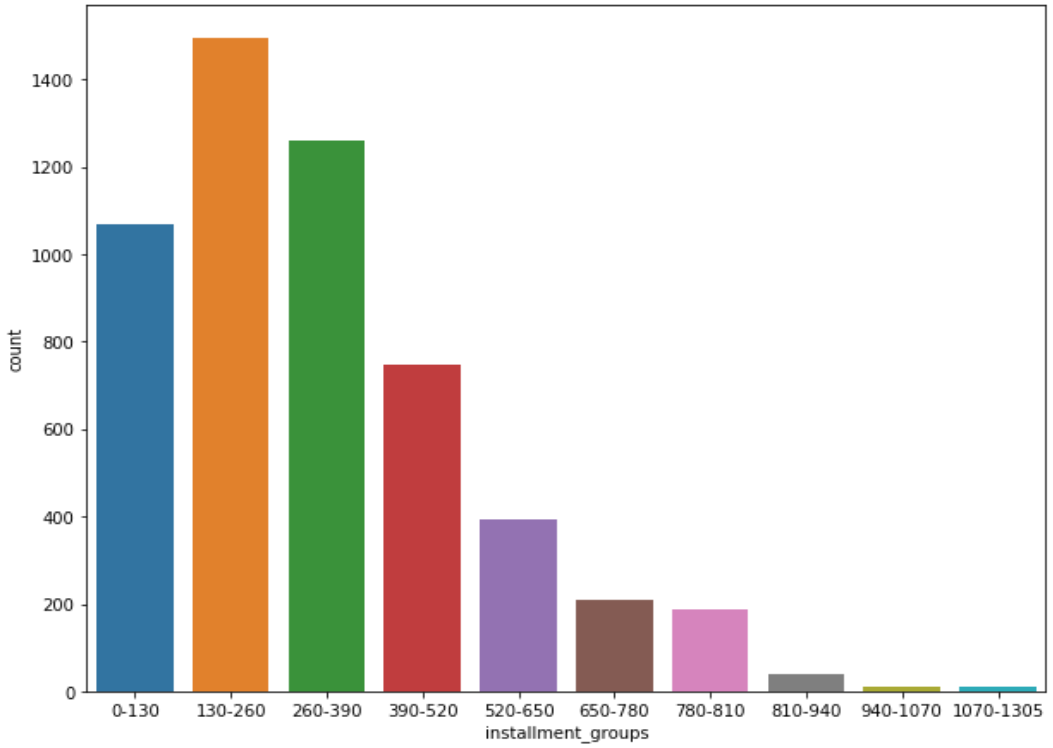
Loan amount



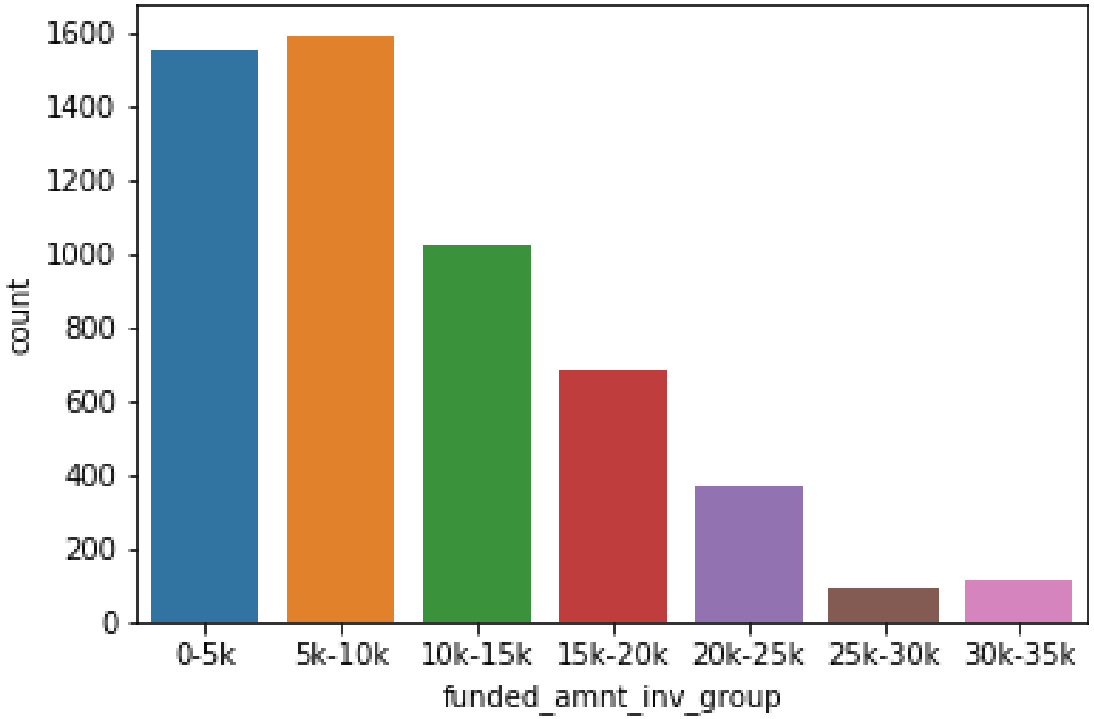
Data Analysis

3. Univariate Analysis with Numerical Variable

Installments



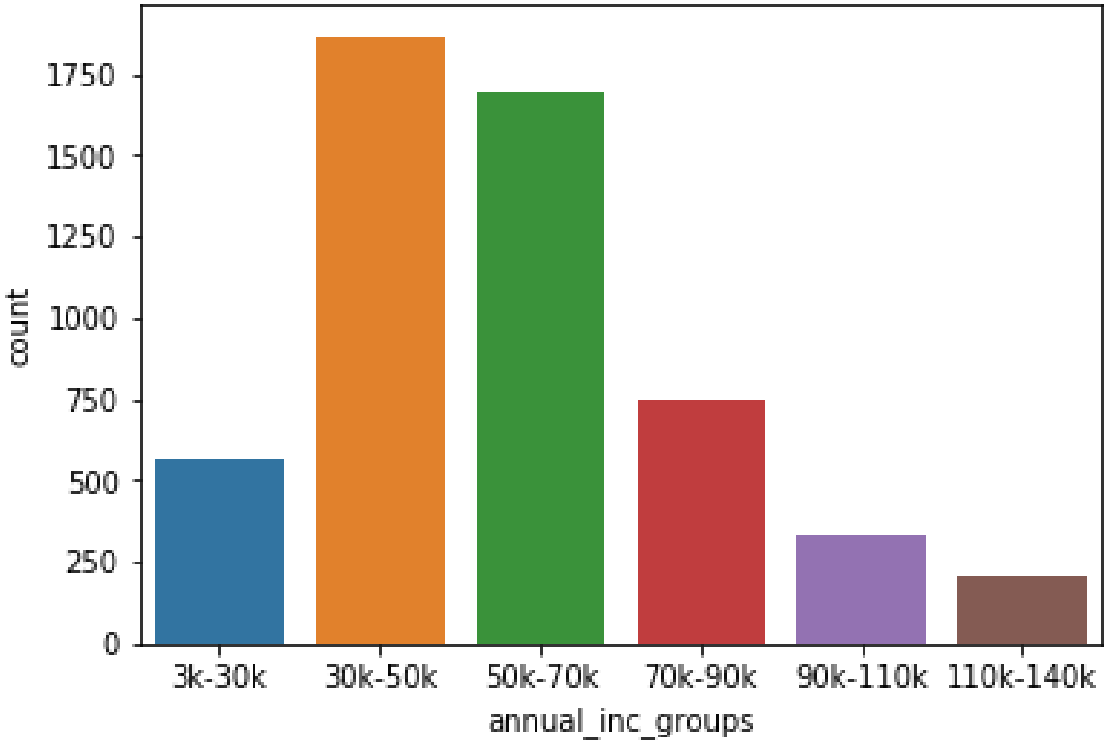
Funded amount investor



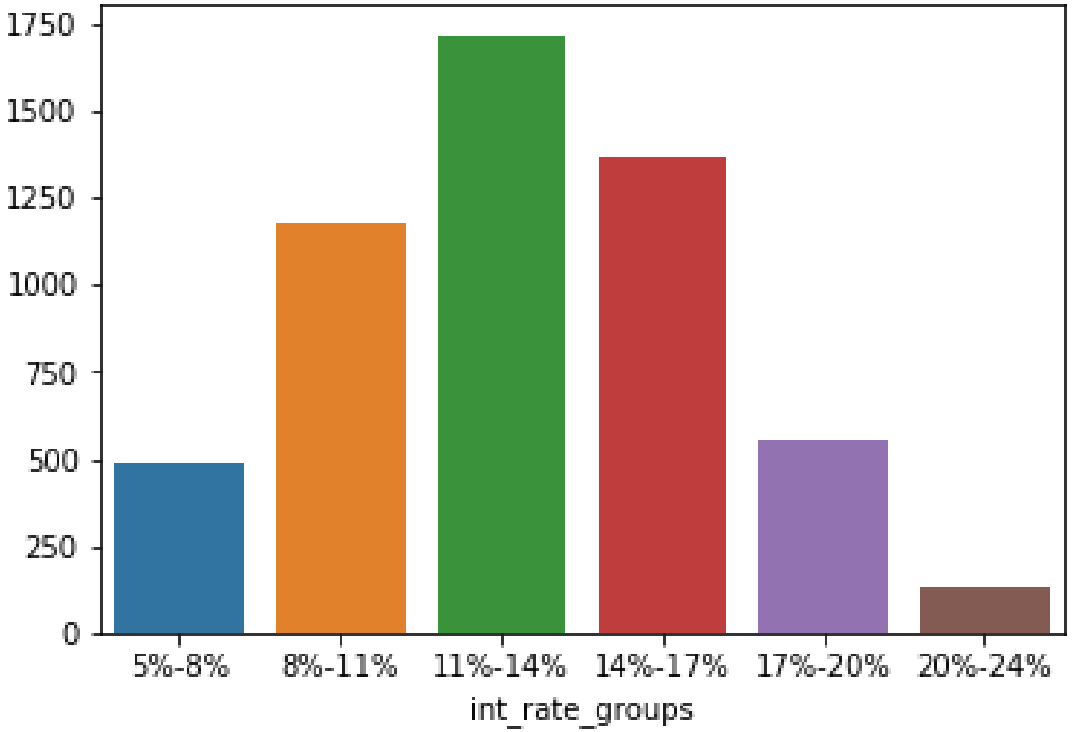
Data Analysis

3. Univariate Analysis with Numerical Variable

Annual income



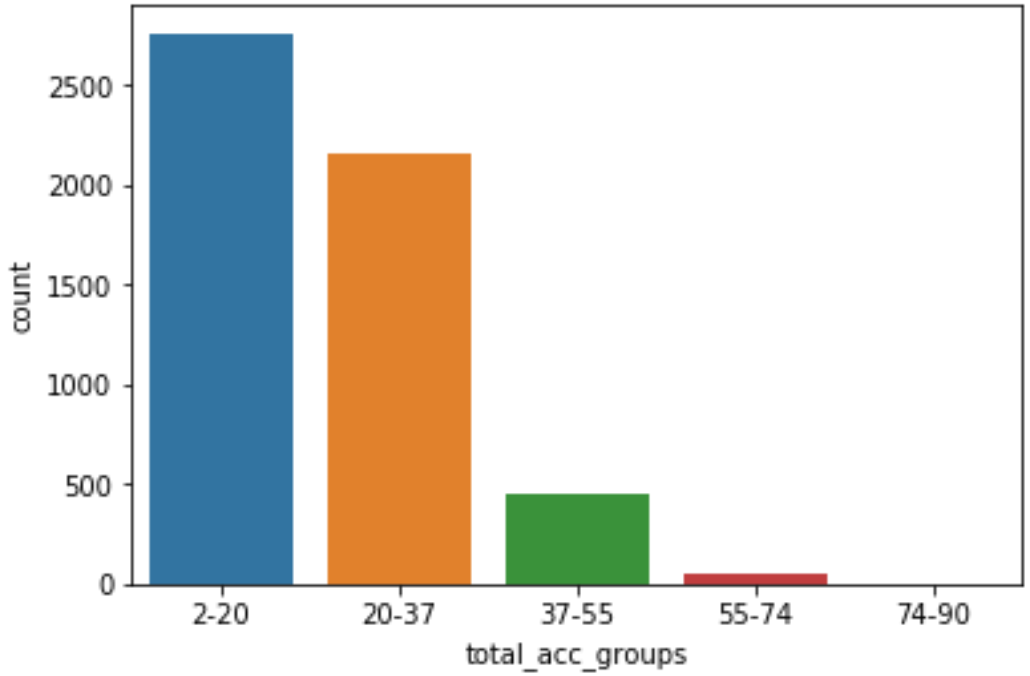
Interest rate



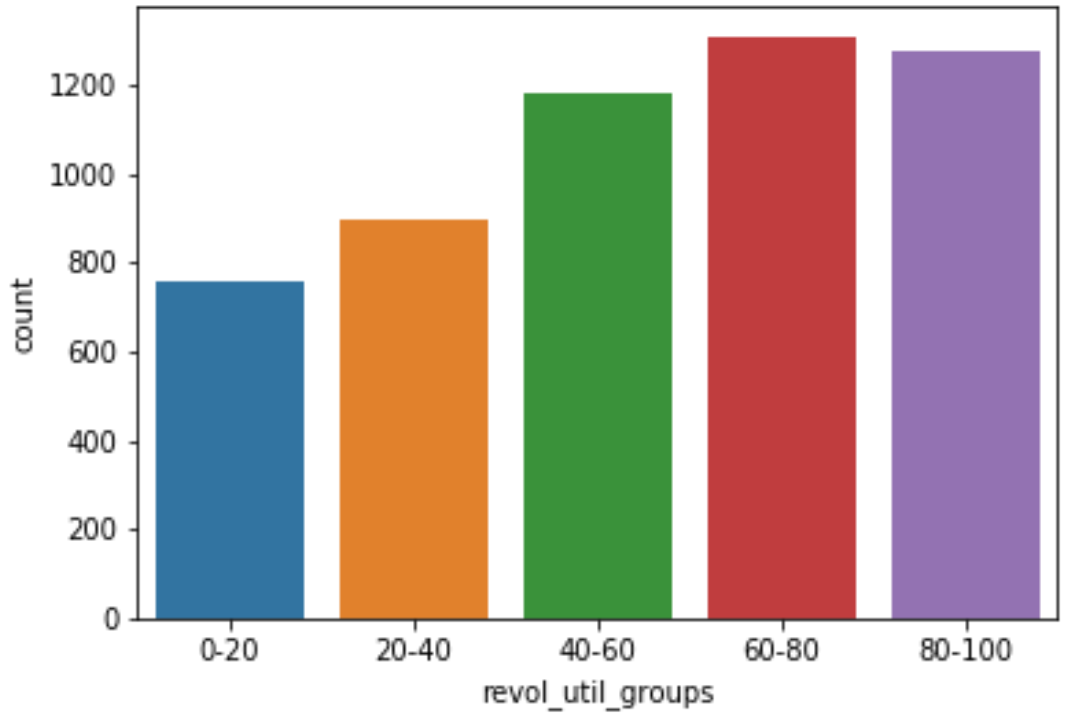
Data Analysis

3. Univariate Analysis with Numerical Variable

Total accounts

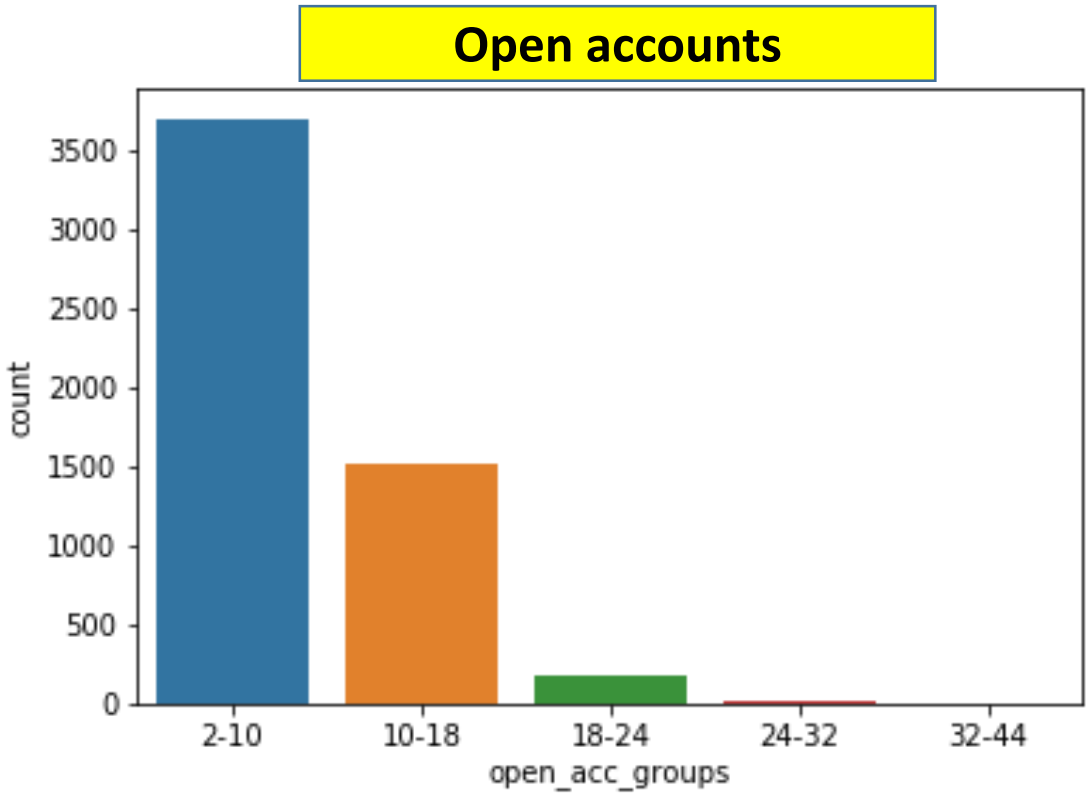


Rev Utilization



Data Analysis

3. Univariate Analysis with Numerical Variable: Observations



- Employment term more than 10 years tend to default more than other. May be due to family commitment and other financial burden.
- People with Zero public record are more defaulters.
- Applicants who enquired about loans more often and have good understanding of their loan payment paid properly compared to those who never enquired about the loan.
- Reason for the Default could be lack of awareness
- Applicants who are taking smaller loan amount up to 15k are the ones who are likely to default more.

Data Analysis

3. Univariate Analysis with Numerical Variable: Observations

- Funded amount also reflects the same behavior since small loans gets funded easily
- Since heavy interest rate is on heavy loan amount which is also given to high annual income people they tend to default.
- People who have middle income rate such as 30k-70k tend to default more
- People with Rev utilization rate are the more defaulter.
- People who have least total amount and open account are more defaulter.

Data Analysis

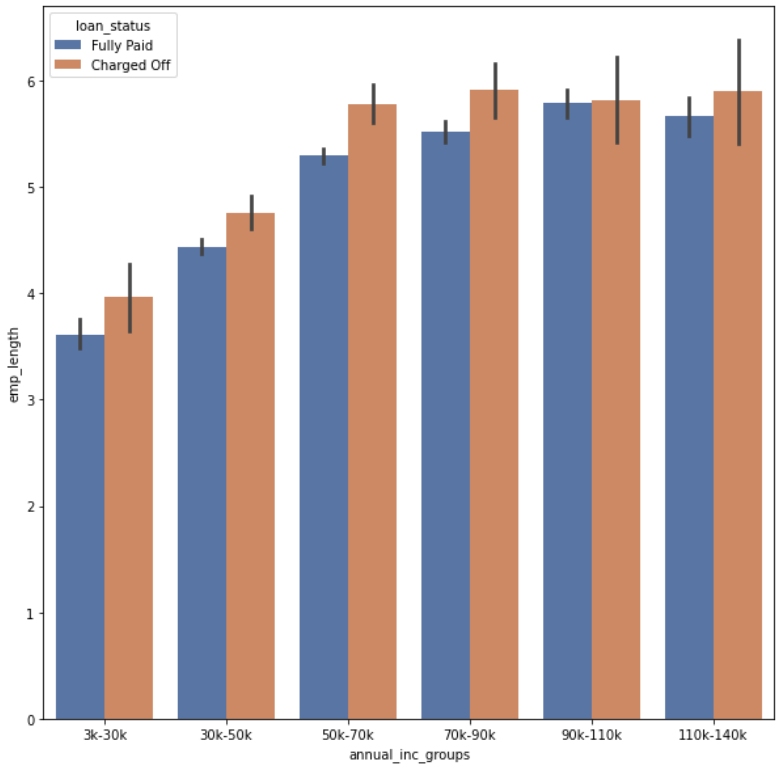
❑**Bivariate Analysis**

- Earlier we had categorized columns into different categories , 2 important categories are
- Customer Attributes : emp_length, home_ownership, annual_inc, verification_status, purpose, open_acc, pub_rec, total_acc
- Loan Attributes : loan_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, issue_d, loan_status, revol_util
- In the customer attributes the most important variable which decides whether a LC should lend money or not is the "annual income" since it decides the capacity of an applicant to repay the loan. And hence let us consider Annual income and pair with other variables to check who is likely to pay or default.
- In the Loan attributes the most important variable which decides whether a LC should lend money or not is the "Loan amount" since it is the riskiest parameter. And hence let us consider "Loan amount" and pair with other variables to check who is likely to pay or default.

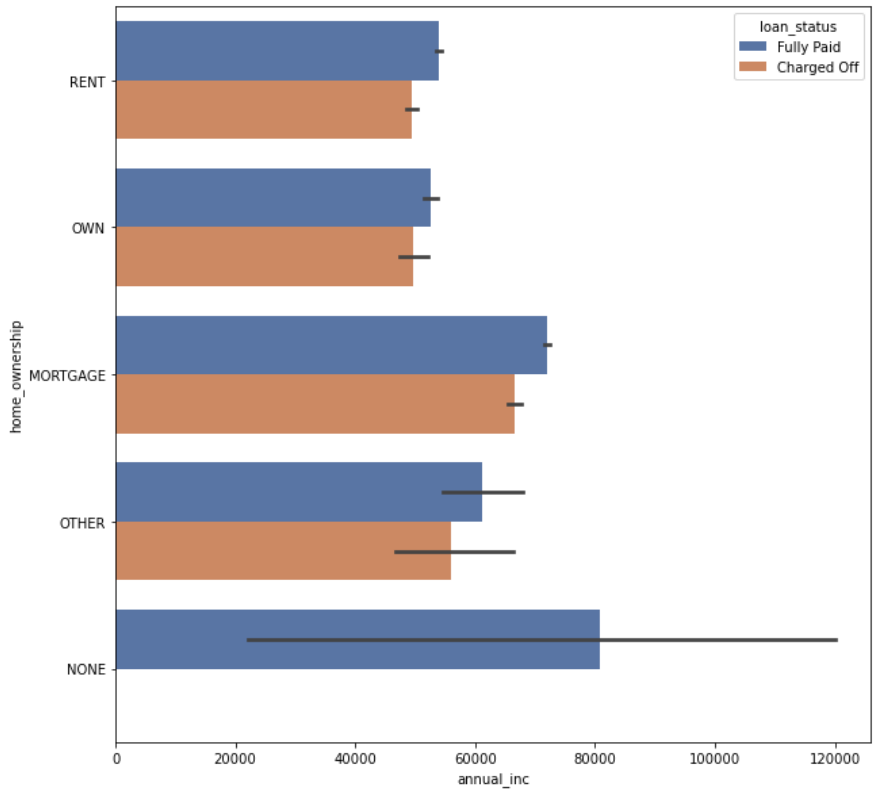
Data Analysis

❑**Bivariate Analysis:**

Annual income vs emp length



Annual income vs house ownership



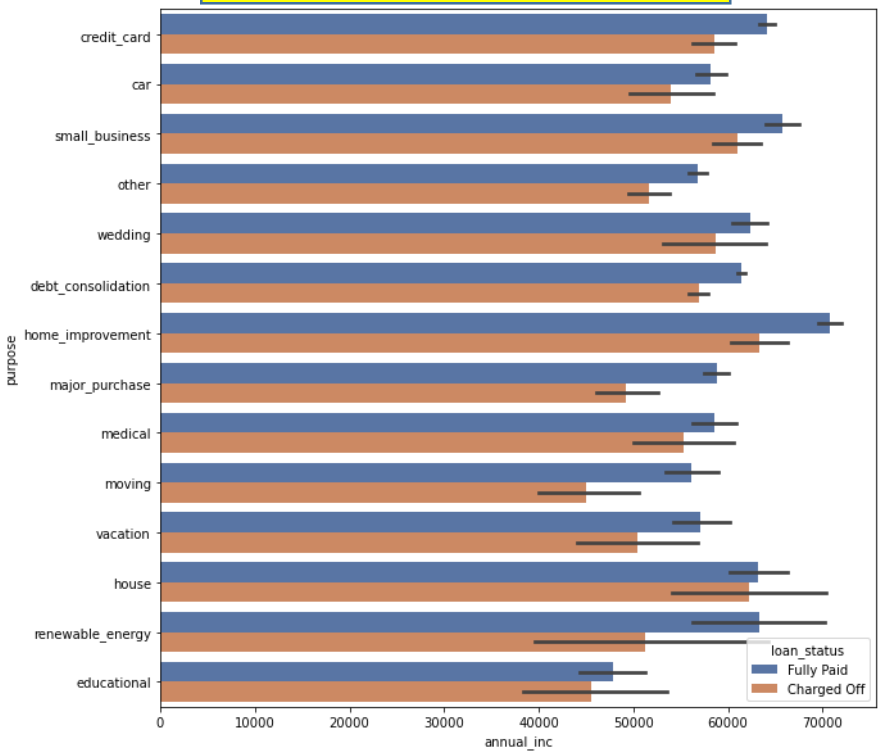
➢Applicants with more annual income are the one who are having more experience. Here people with more experience and with medium salary tend to default more.

➢Applicants who are having less annual income and living in rented house are mortgaged their house are more defaulters.

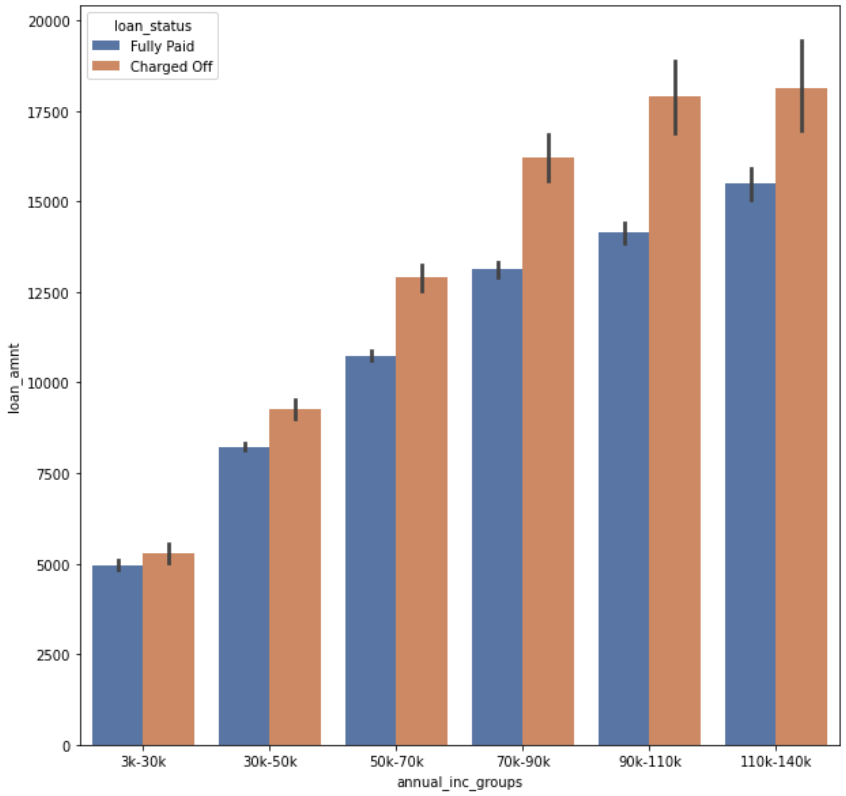
Data Analysis

❑**Bivariate Analysis:**

Annual income vs Purpose



Annual income vs Loan amount



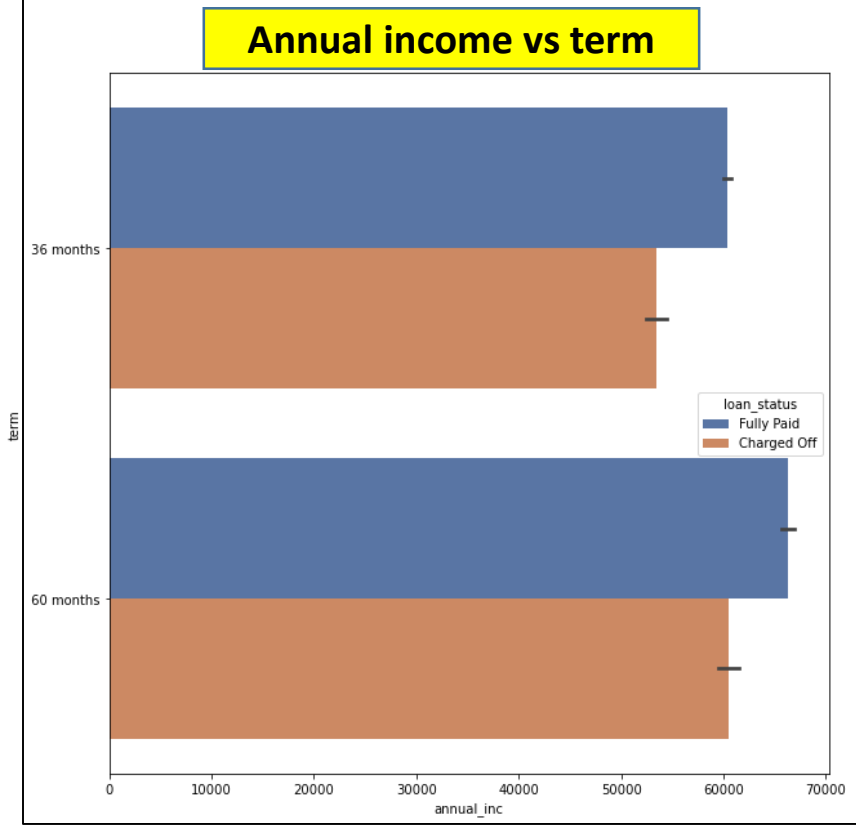
➢Applicants with high annual income rate like 70k are requesting loan for all purposes and they tend to pay.

➢Higher the annual income and the higher the loan amount requested and default rate is also more for such applicants.

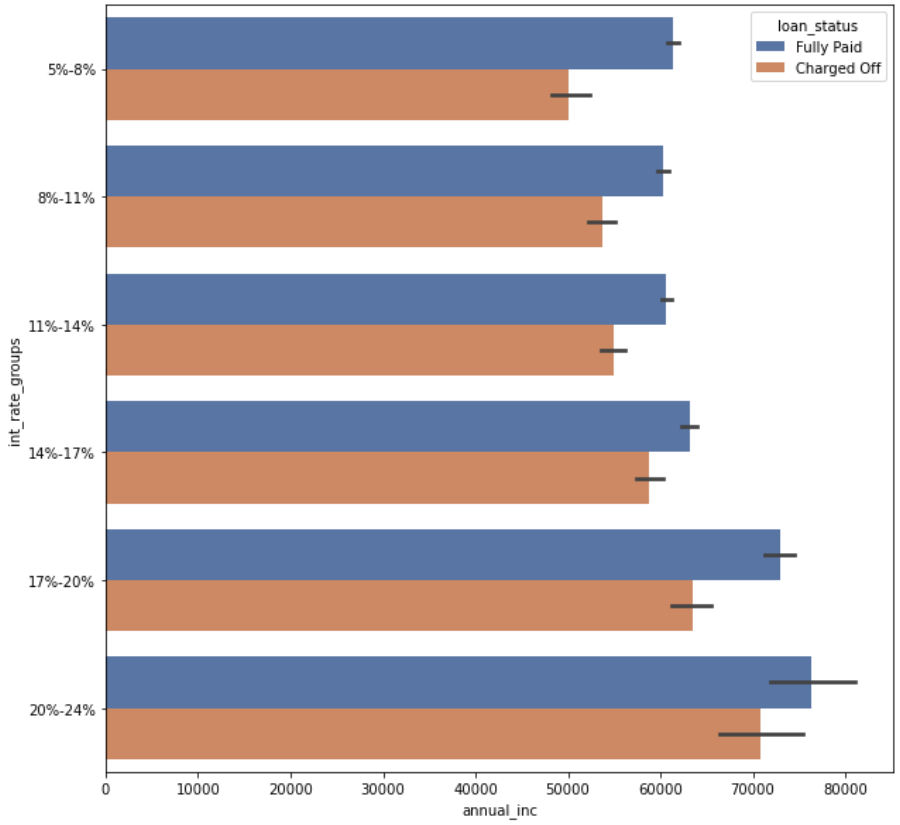
Data Analysis

❑**Bivariate Analysis:**

Annual income vs term



Annual income vs Interest rate



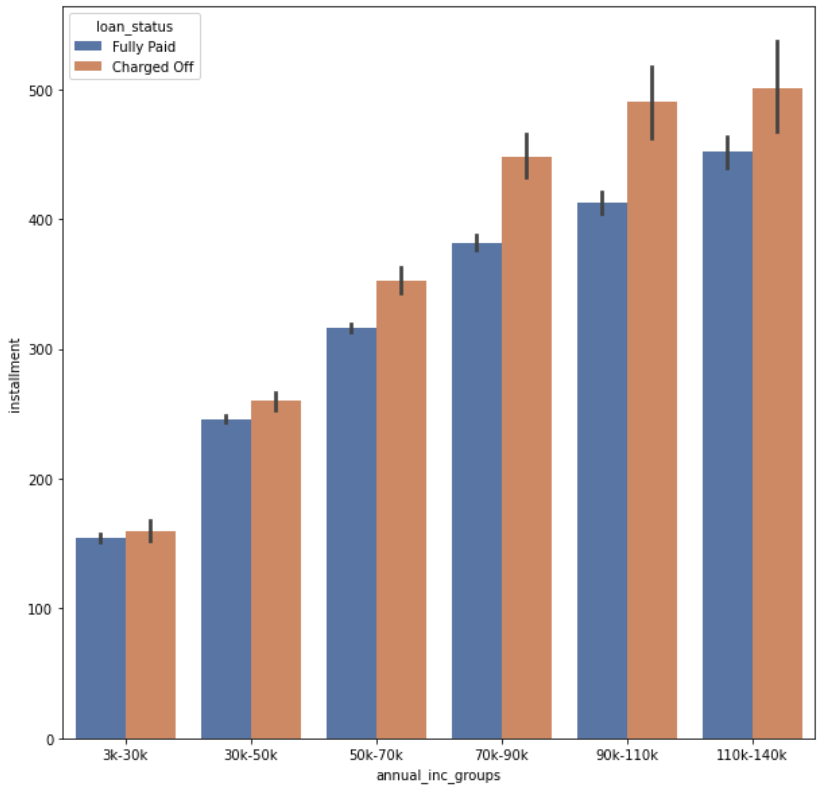
➢People with high annual income are taking more term and they tend to pay.

➢People with high annual income and low interest rate tend to pay more than high interest rate

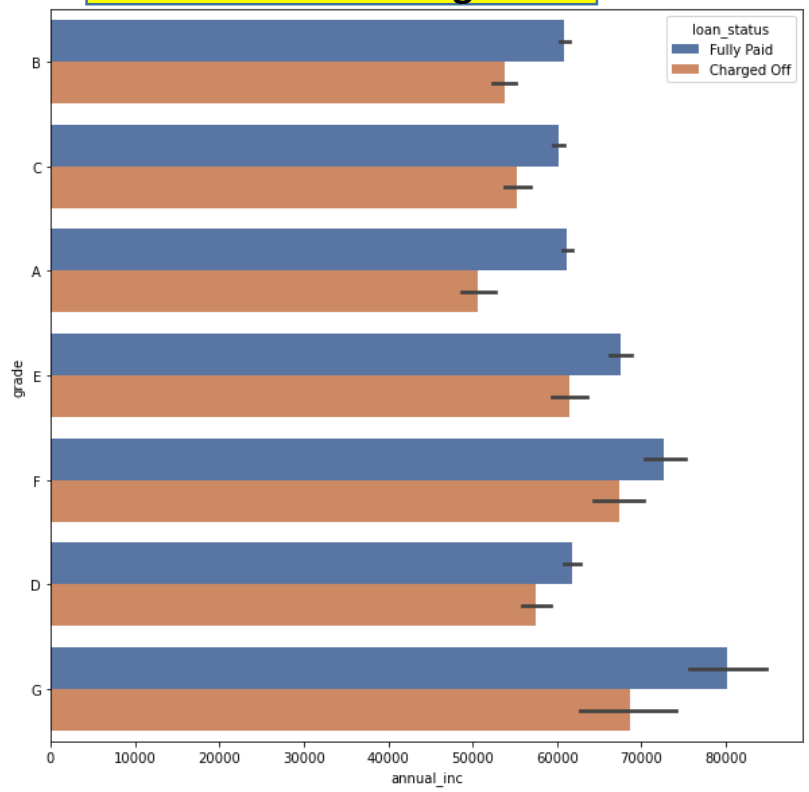
Data Analysis

❑**Bivariate Analysis:**

Annual income vs installment



Annual income vs grade



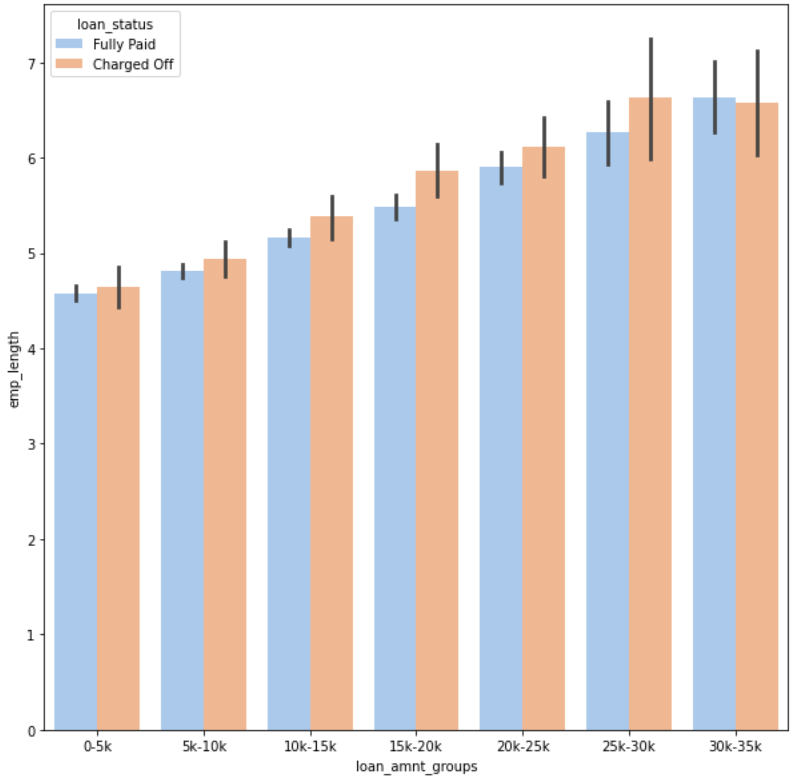
➢Irrespective of annual in come people with high installment tent to default more

➢People with high annual income are going for G grade loan and they tend to pay more.

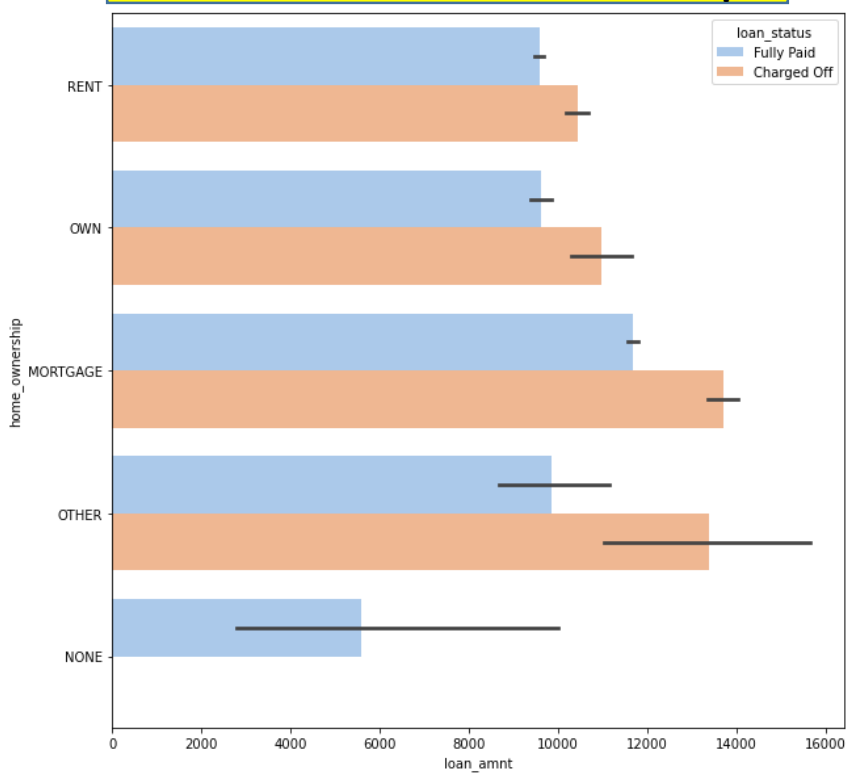
Data Analysis

❑Bivariate Analysis:

Loan amount vs employment length



Loan amount vs house ownership

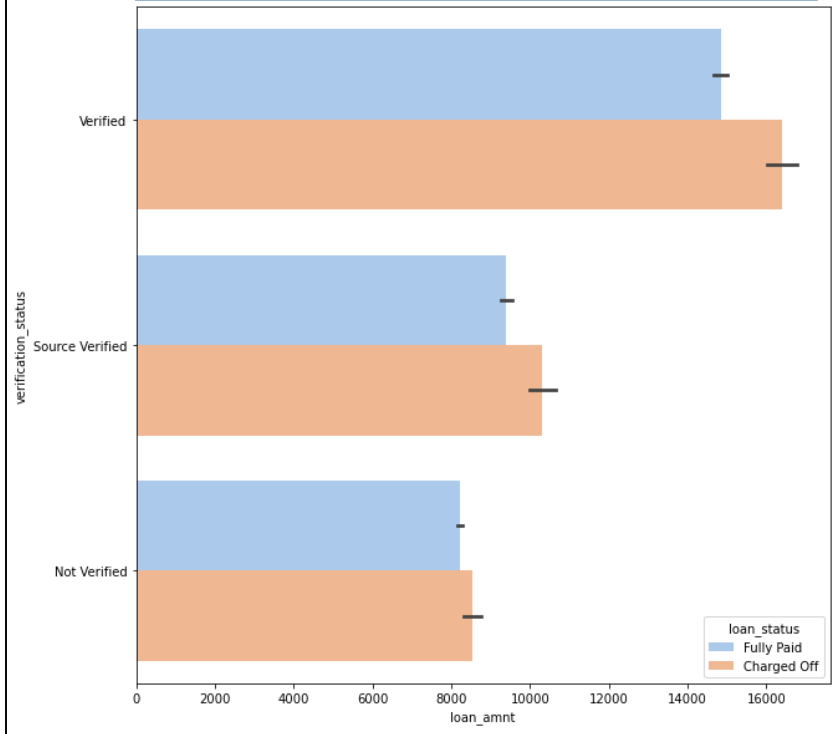


- As the employment length is more, applicants are requesting for more loan and there are more chances that they may go defaulter for such high loans
- Mortgage and other house people are also requested for high loans but there are high chances that they may go defaulter so it is suggested to avoid giving them high loans.

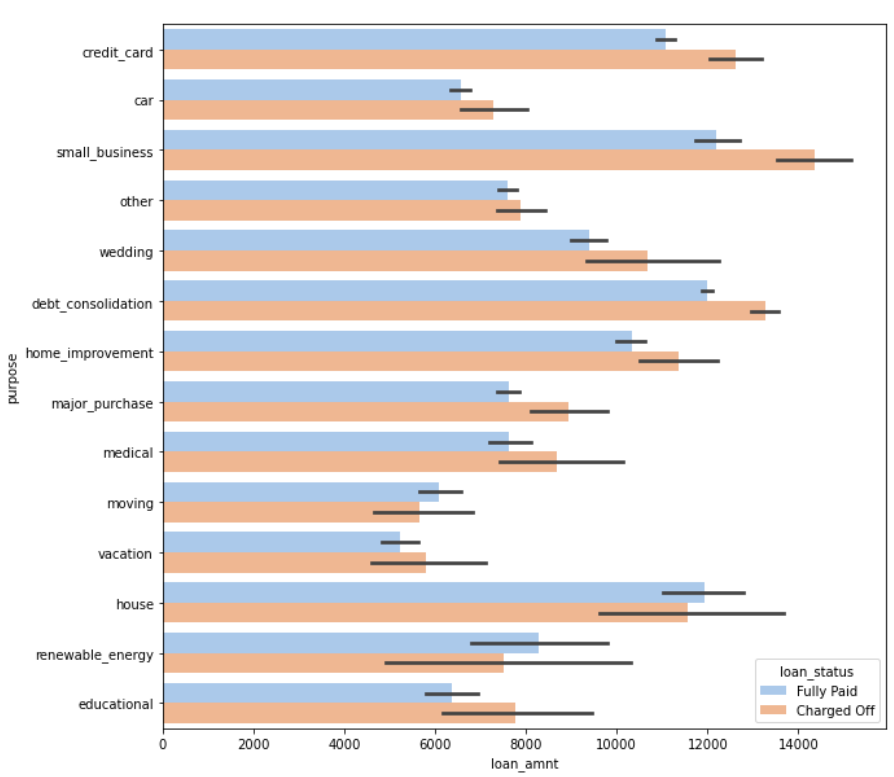
Data Analysis

❑Bivariate Analysis:

Loan amount vs verification status



Loan amount vs purpose

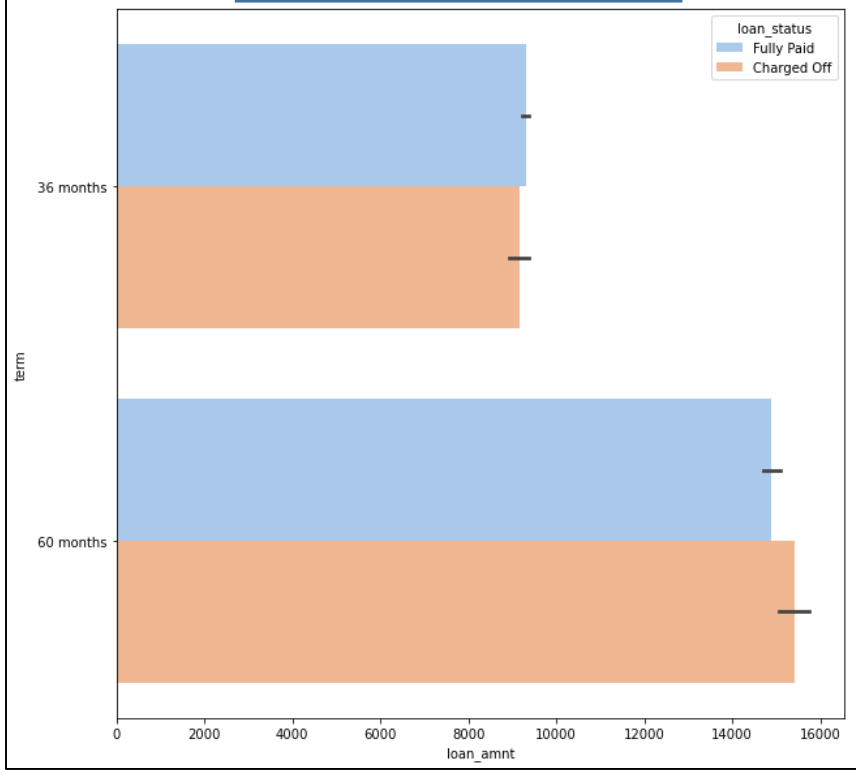


- Interesting pattern of verified applicants defaulting more is observed. It might be due to high loan amount allotted.
- People who are taking high amount of loan for small business, credit card and debt consolidation are more defaulters.it might be due to loss in business

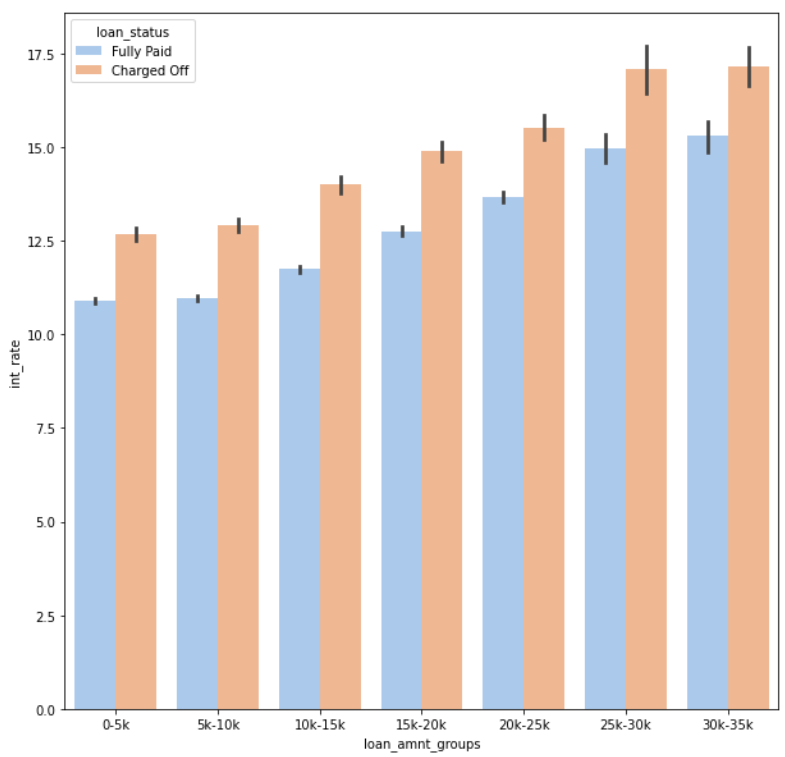
Data Analysis

❑**Bivariate Analysis:**

Loan amount vs term



Loan amount vs interest rate

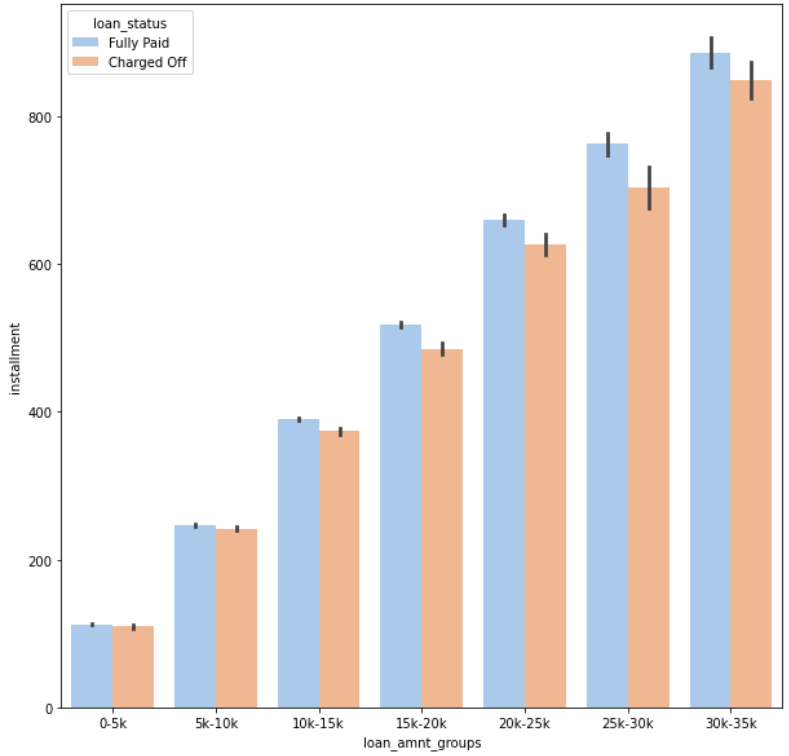


- Applicants having high loan amount are choosing high term defaulting more than those who are choosing low term.
- Irrespective loan amount people with high interest rate people tend to default.

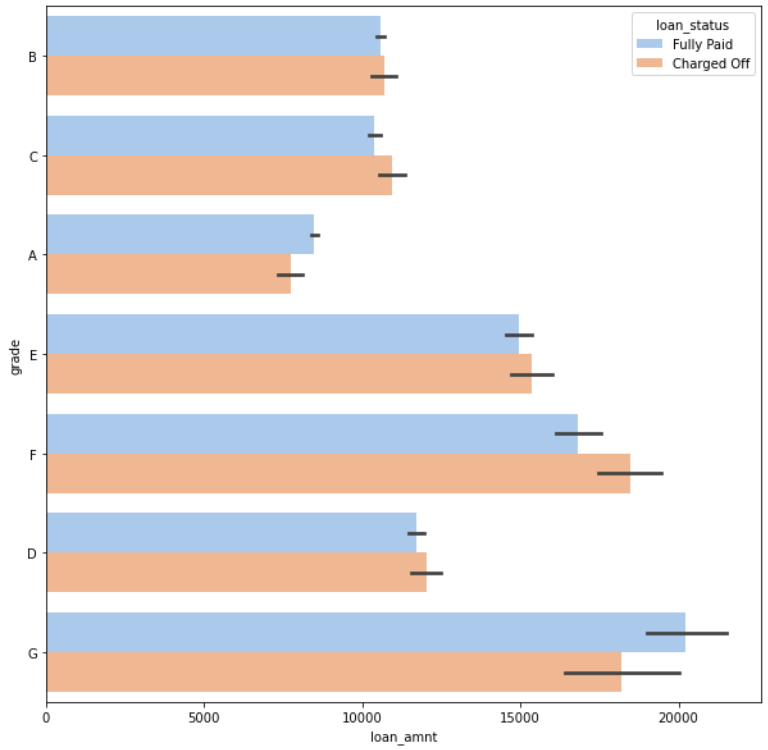
Data Analysis

❑**Bivariate Analysis:**

Loan amount vs installment



Loan amount vs grade

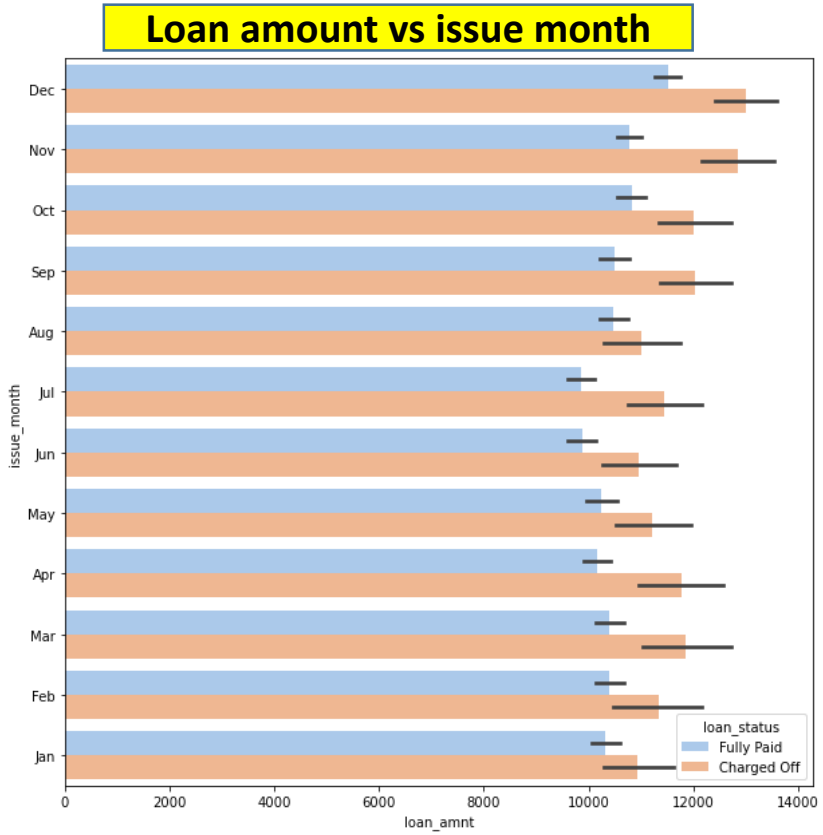


- Applicants with high loan amount are the ones who are having installment amount. For them its difficult to pay so they go defaulter.
- G and F grade are taking more amount . G grade loans are getting paid but F is defaulting more.

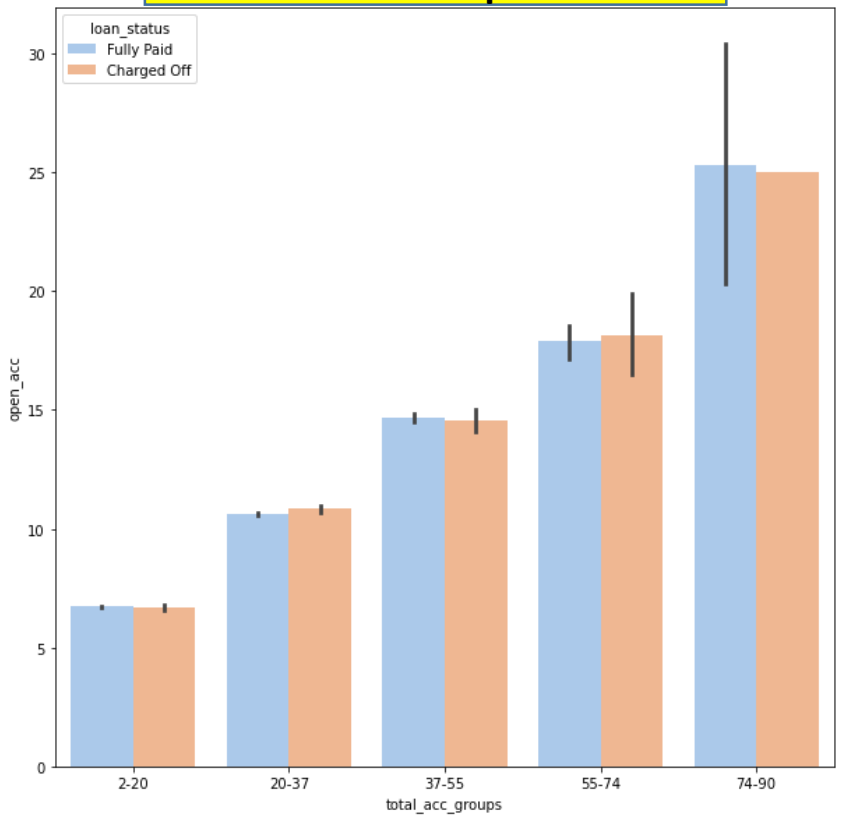
Data Analysis

❑**Bivariate Analysis:**

Loan amount vs issue month



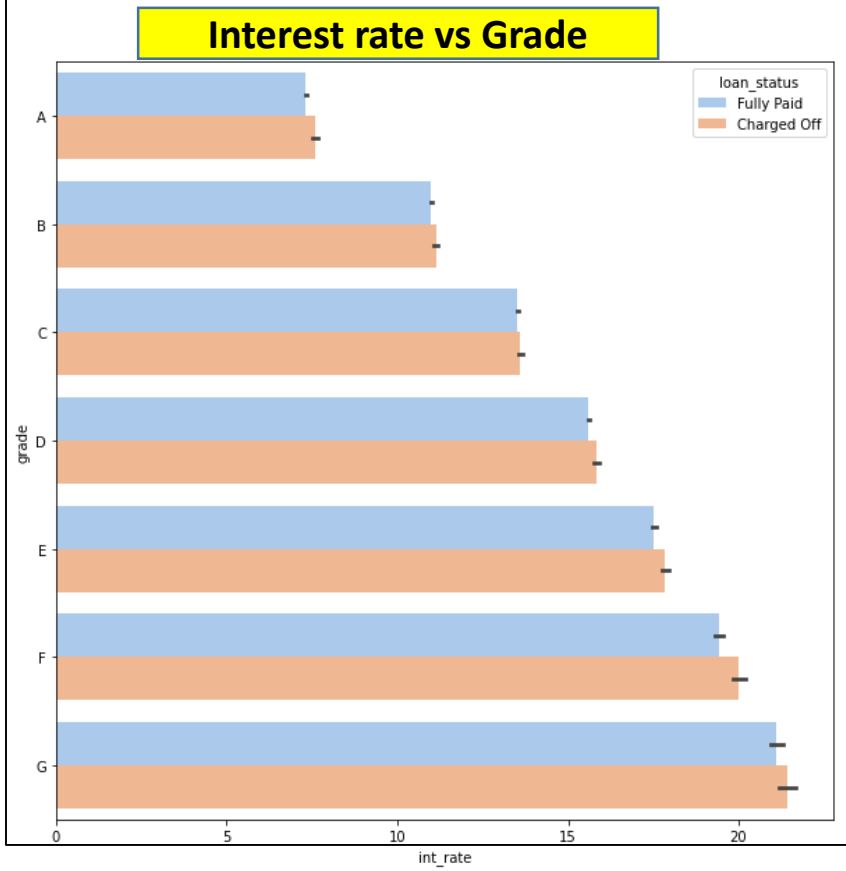
Total account vs open account



- December and November month are having highest issue in paying loan.
- As total accounts are increasing, open accounts are also increasing which is indicating that people who are making more loan and they are not paying earlier.

Data Analysis

❑**Bivariate Analysis:**



- G and F grades are the ones which are having high interest rate.

Insights and Recommendations

- It is highly suggested to verify income source for the applicants who are living in rented house, having medium level annual income (30k-70k) and requesting for high loan amount
- It is suggested to avoid loan allotments for the applicants for the purpose such as debt consolidation and credit card with more number of open credit lines.
- It is suggested to encourage to have more number of enquiry with LC officials so that they will be aware of loan repayment details.
- It is suggested to reduced the interest rate for medium range loan amounts.
- It is highly suggested to avoid high loan allotment to applicants having more than 10 years of experience and having mid level annual income (30k-70k)
- It is suggested to upgrade the term for people having high loan amount which reduced the installment amount and the chances of defaulting