Lending Club Case Study

Submitted by:

Snehal Yadav

Srikrishna Jana

Contents

Problem Statement

- **Lending Club**, a Consumer Finance marketplace specializing in offering a variety of loans to urban customers, faces a critical challenge in managing its loan approval process. When evaluating loan applications, the company must make sound decisions to minimize financial losses, primarily stemming from loans extended to applicants who are considered **"Risky".**
- These financial losses, referred to as **Credit Losses**, occur when borrowers fail to repay their loans or default. In simpler terms, borrowers labeled as **"Charged-Off"** are the ones responsible for the most significant losses to the company.
- The primary objective of this exercise is to assist Lending Club in mitigating credit losses. This challenge arises from two potential scenarios:

    1. **Identifying applicants likely to repay their loans is crucial, as they can generate profits for the company through interest payments. Rejecting such applicants would result in a loss of potential business.**
    2. **On the other hand, approving loans for applicants not likely to repay and at risk of default can lead to substantial financial losses for the company.**

- The objective is to pinpoint applicants at risk of defaulting on loans, enabling a reduction in credit losses. This case study aims to achieve this goal through Exploratory Data Analysis (EDA) using the provided dataset.
- In essence, 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.

Data Description

Lending Club provided us with customer's historical data. This dataset contained information pertaining to the borrower's past credit history and Lending Club loan information. The total dataset consisted of over 39717 records and 111 columns, which was sufficient for our team to conduct analysis. Variables present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement.

We required only the variables that had a direct or indirect response to a borrower's potential to default. To achieve this, we prepared the data by choosing select variables that would best fit this criteria.

Data Understanding

Dataset Attributes:

Primary Attribute

Loan Status: The Principal Attribute of Interest (loan_status). This column consists of three distinct values:

1. **Fully-Paid:** Signifies customers who have successfully repaid their loans.
2. **Charged-Off:** Indicates customers who have been labeled as "Charged-Off" or have defaulted on their loans.
3. **Current:** Represents customers whose loans are presently in progress and, thus, cannot provide conclusive evidence regarding future defaults.

For the purposes of this case study, rows with a "Current" status will be excluded from the analysis.

Decision Matrix:

Loan Acceptance Outcome - There are three potential scenarios:

Fully Paid - This category represents applicants who have successfully repaid both the principal and the interest rate of the loan.

Current - Applicants in this group are actively in the process of making loan installments; hence, the loan tenure has not yet concluded. These individuals are not categorized as 'defaulted.'

Charged-off - This classification pertains to applicants who have failed to make timely installments for an extended period, resulting in a 'default' on the loan.

Loan Rejection - In cases where the company has declined the loan application (usually due to the candidate not meeting their requirements), there is no transactional history available for these applicants. Consequently, this data is unavailable to the company and is not included in this dataset.

Data Understanding

Key Columns of Significance:

The provided columns serve as pivotal attributes, often referred to as predictors. These attributes, available during the loan application process, significantly contribute to predicting whether a loan will be approved or rejected. It's important to note that some of these columns may be excluded due to missing data in the dataset.

- **Customer Demographics:**
  - **Annual Income (annual_inc):** Reflects the customer's annual income. Typically, a higher income enhances the likelihood of loan approval.
  - **Home Ownership (home_ownership):** Indicates whether the customer owns a home or rents. Home ownership provides collateral, thereby increasing the probability of loan approval.
  - **Employment Length (emp_length):** Represents the customer's overall employment tenure. Longer tenures signify greater financial stability, leading to higher chances of loan approval.
  - **Debt to Income (dti):** Measures how much of a person's monthly income is already being used to pay off their debts. A lower DTI translates to a higher chance of loan approval.
  - **State (addr_state):** Denotes the customer's location and can be utilized for creating a generalized demographic analysis. It may reveal demographic trends related to delinquency or default rates.
- **Loan Characteristics:**
  - **Loan Amount (loan_amt):** Represents the amount of money requested by the borrower as a loan.
  - **Grade (grade):** Represents a rating assigned to the borrower based on their creditworthiness, indicating the level of risk associated with the loan.
  - **Term (term):** Duration of the loan, typically expressed in months.
  - **Loan Date (issue_d):** Date when the loan was issued or approved by the lender.
  - **Purpose of Loan (purpose):** Indicates the reason for which the borrower is seeking the loan, such as debt consolidation, home improvement, or other purposes.
  - **Verification Status (verification_status):** Represents whether the borrower's income and other information have been verified by the lender.
  - **Interest Rate (int_rate):** Represents the annual rate at which the borrower will be charged interest on the loan amount.
  - **Installment (installment):** Represents the regular monthly payment the borrower needs to make to repay the loan, including both principal and interest.
  - **Public Records (public_rec):** Refers to derogatory public records, which contribute to loan risk. A higher value in this column reduces the likelihood of loan approval.
  - **Public Records Bankruptcy (public_rec_bankruptcy):** Indicates the number of locally available bankruptcy records for the customer. A higher value in this column is associated with a lower success rate for loan approval.

Data Cleaning & Pre-processing

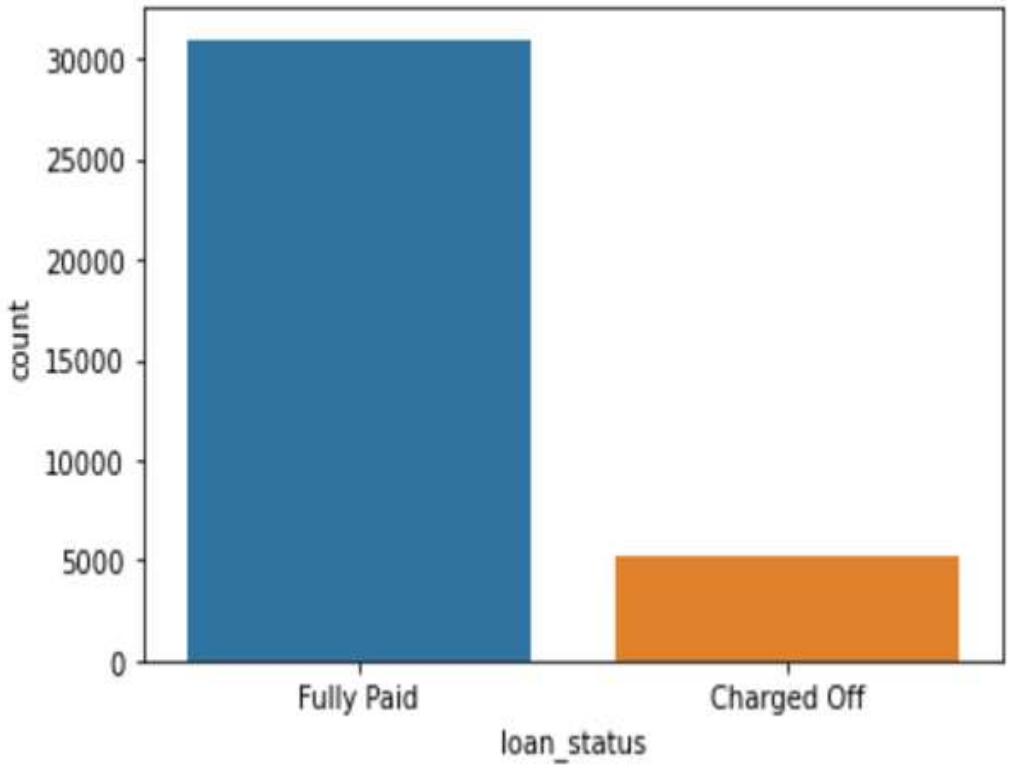
- Loading data from loan CSV
- Checking for null values in the dataset
- Checking for unique values
- Checking for duplicated rows in data
- Dropping Records & Columns
- Outlier Treatment
- Imputing values in Columns

Univariate Analysis

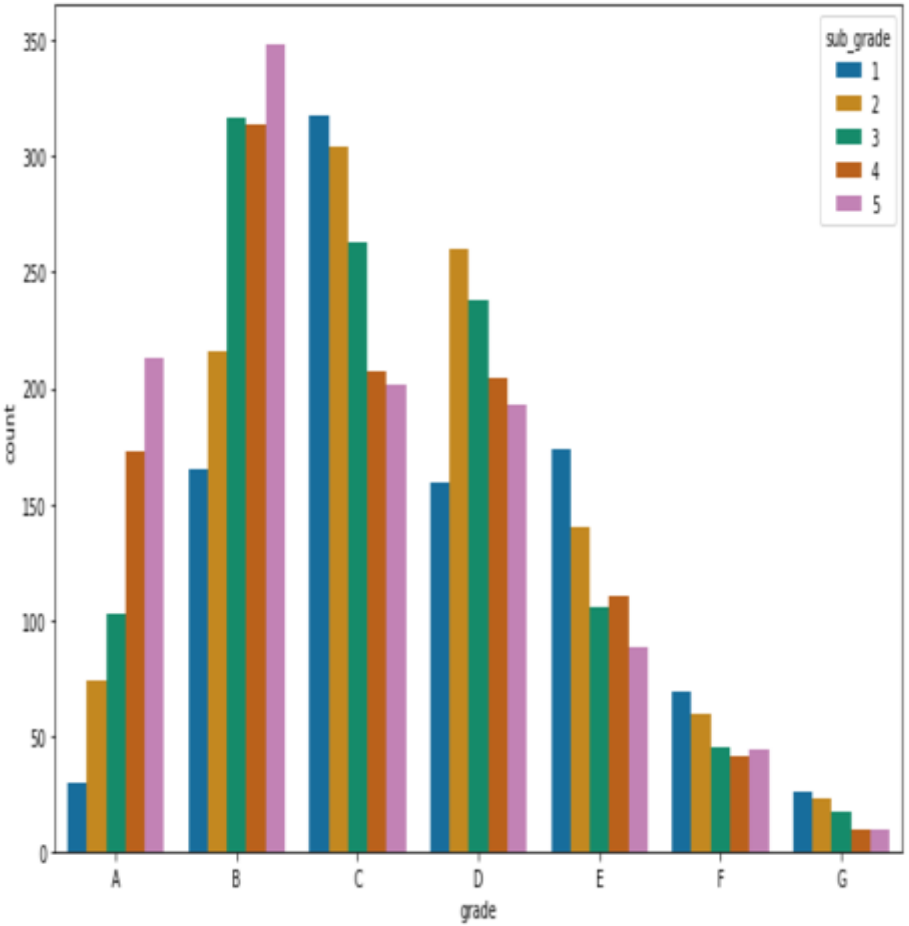
Univariate analysis is a statistical technique used to analyze and summarize the distribution, central tendency, and variability of a single variable. It is the simplest form of data analysis and focuses on understanding patterns, trends, and characteristics of one variable at a time without considering relationships with other variables.

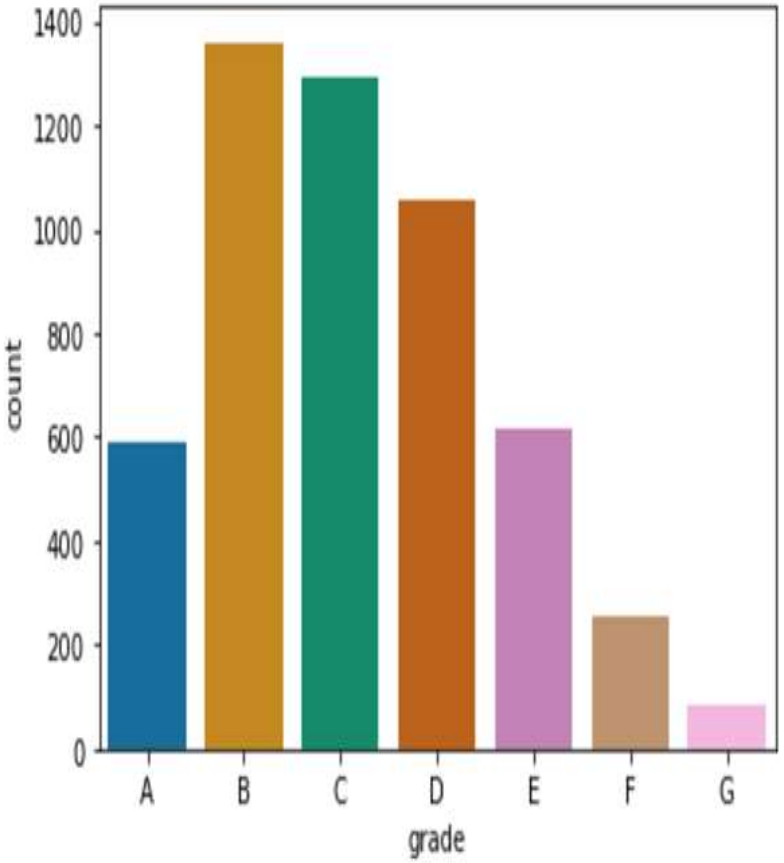
We are analyzing and visualizing only the defaulter data. So sub setting the data while plotting only for 'Charged Off' loan_status for below plots

Univariate Analysis



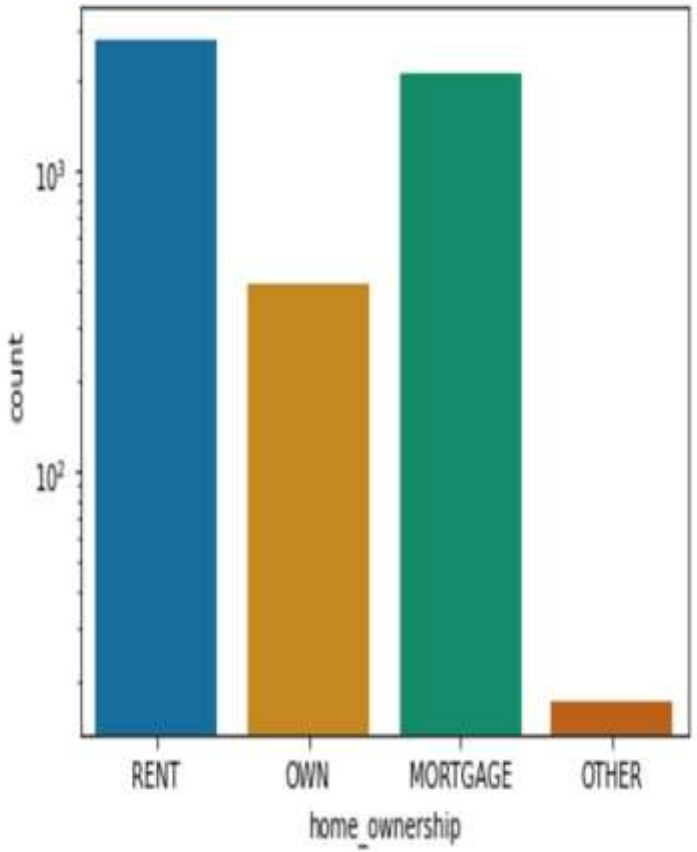
Univariate Analysis

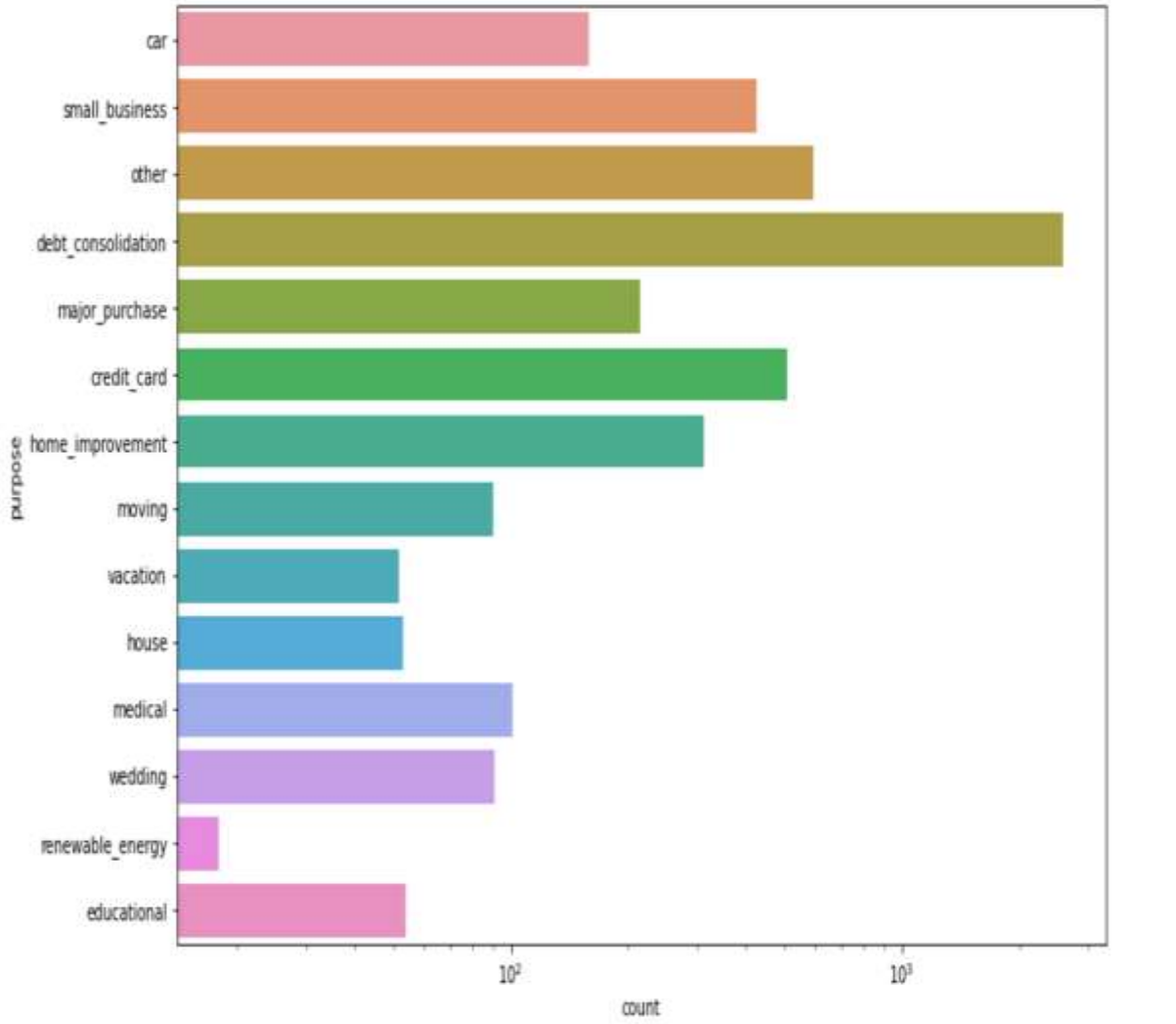




Grade and subgrade

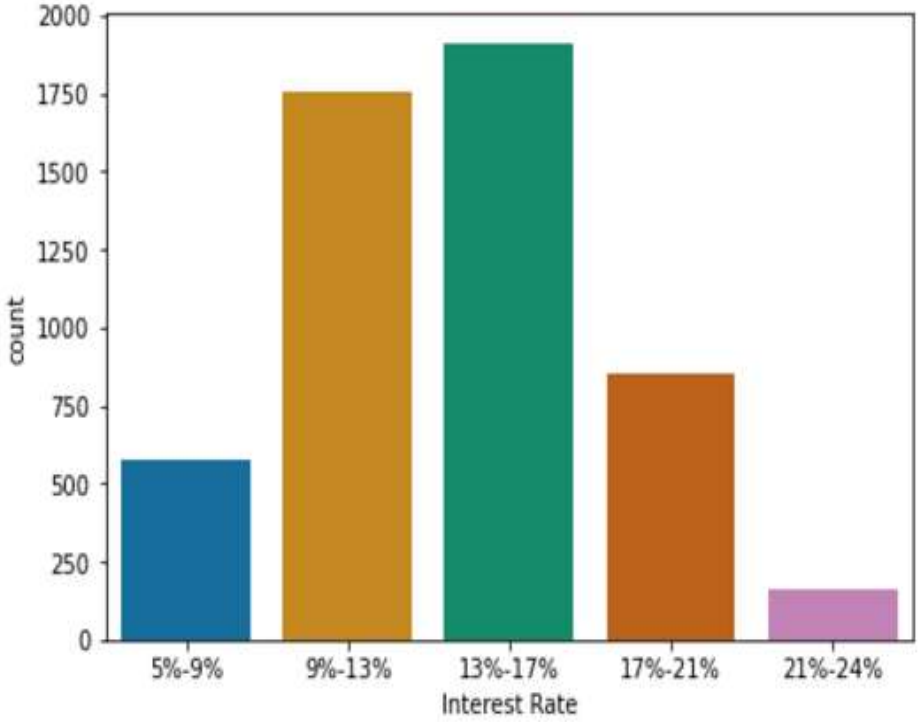
Univariate Analysis

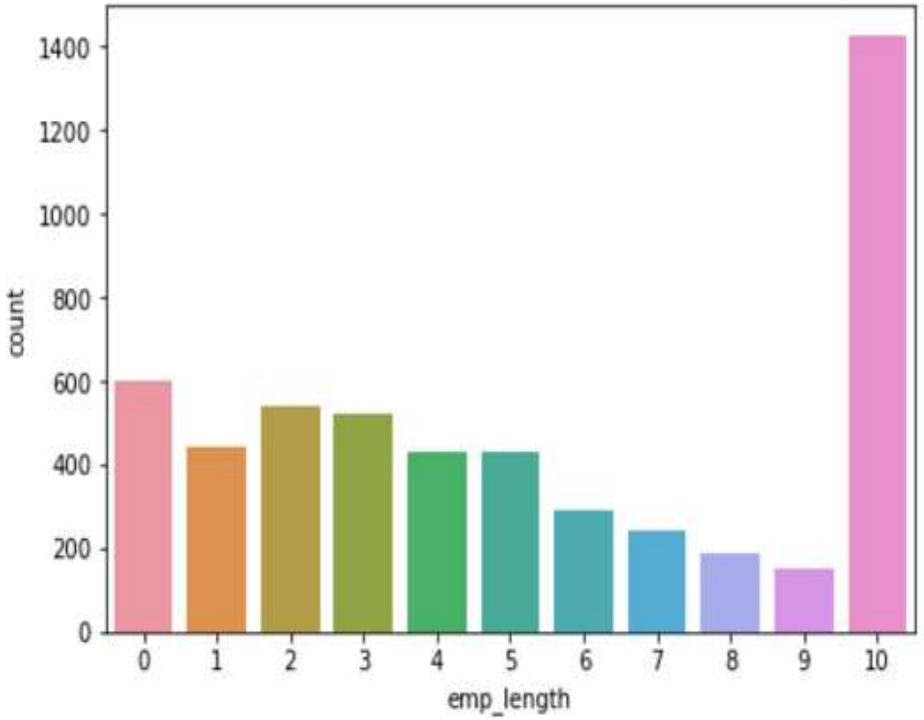




Home ownership and purpose

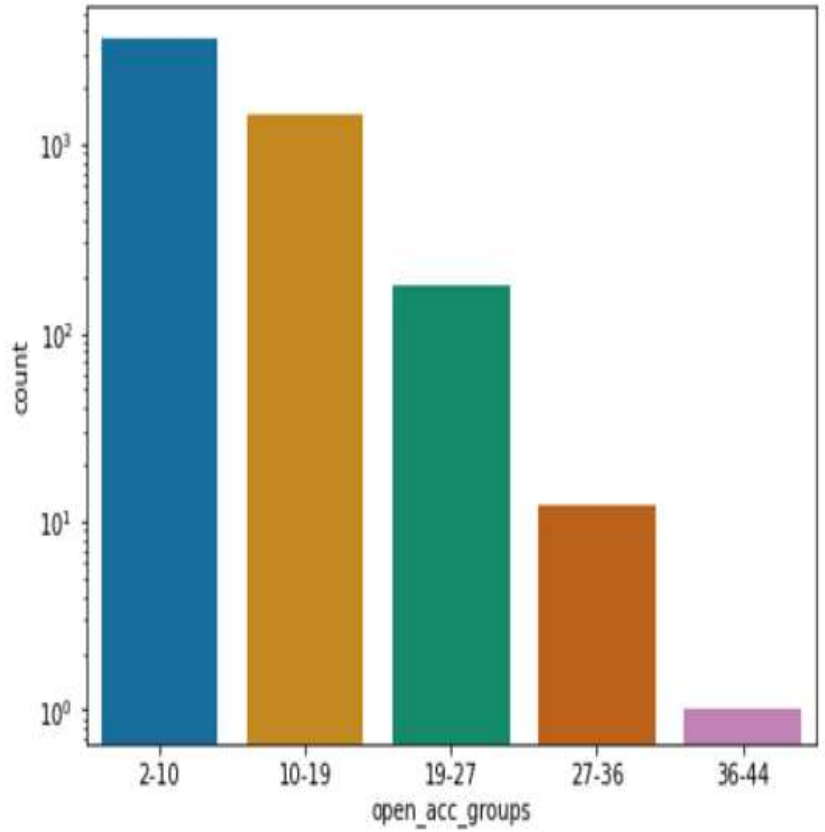
Univariate Analysis

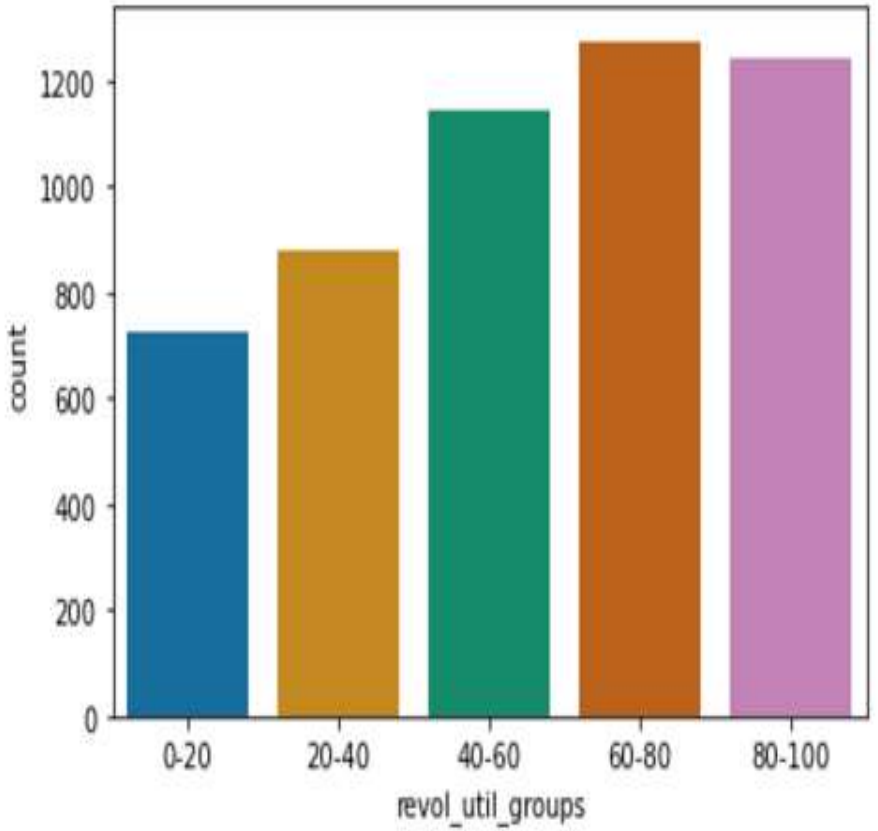




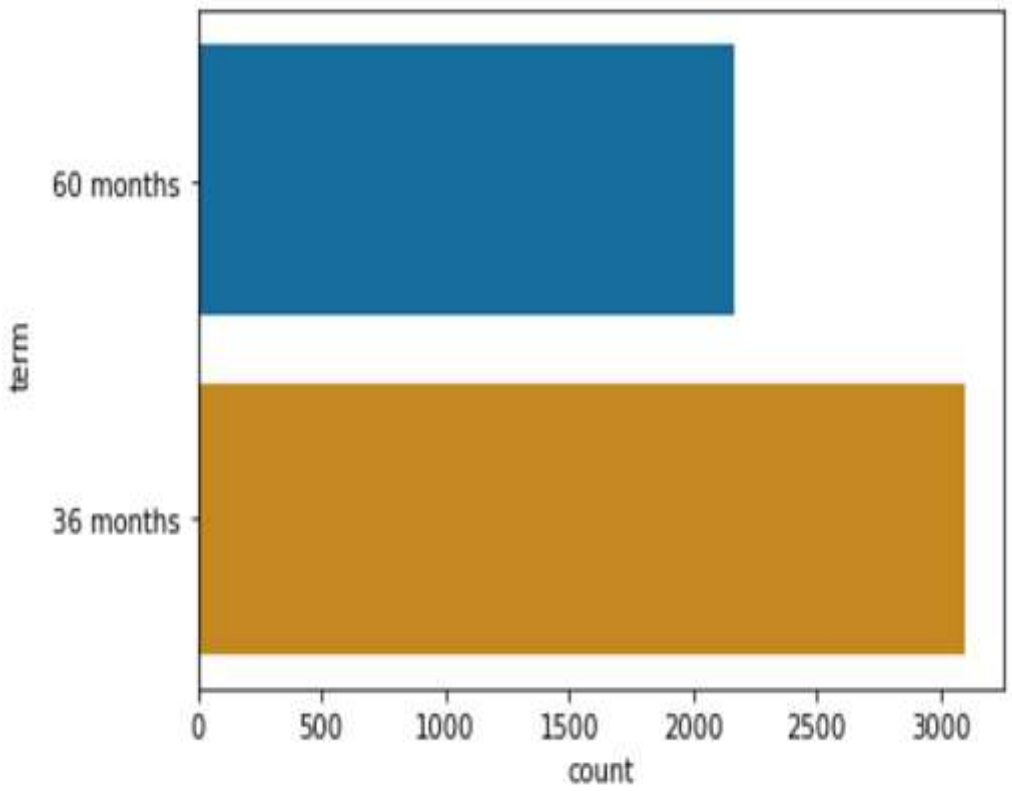
Interest rate and emp length

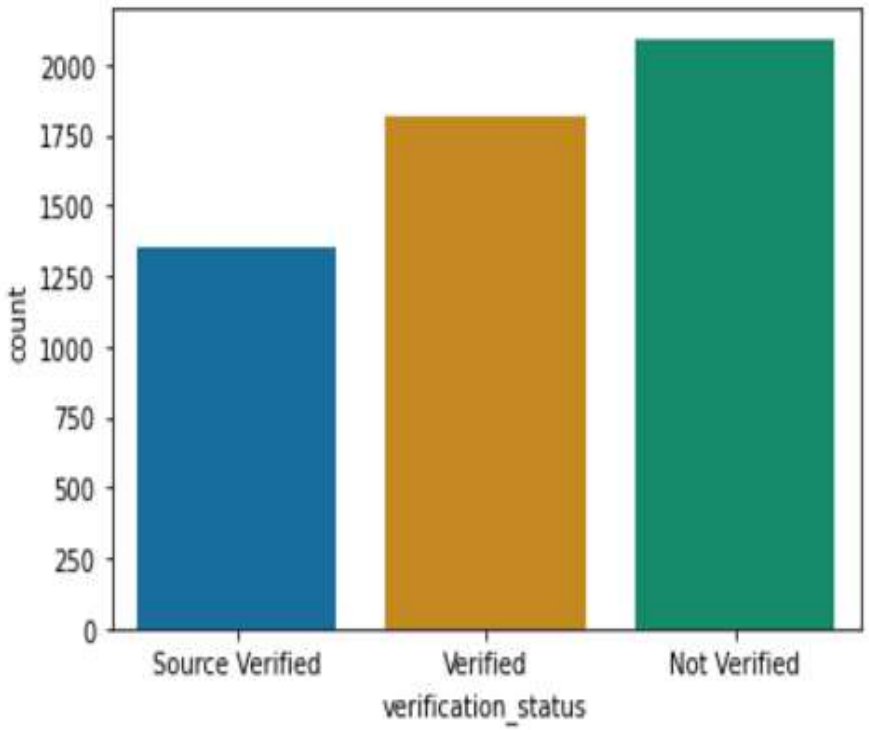
Univariate Analysis



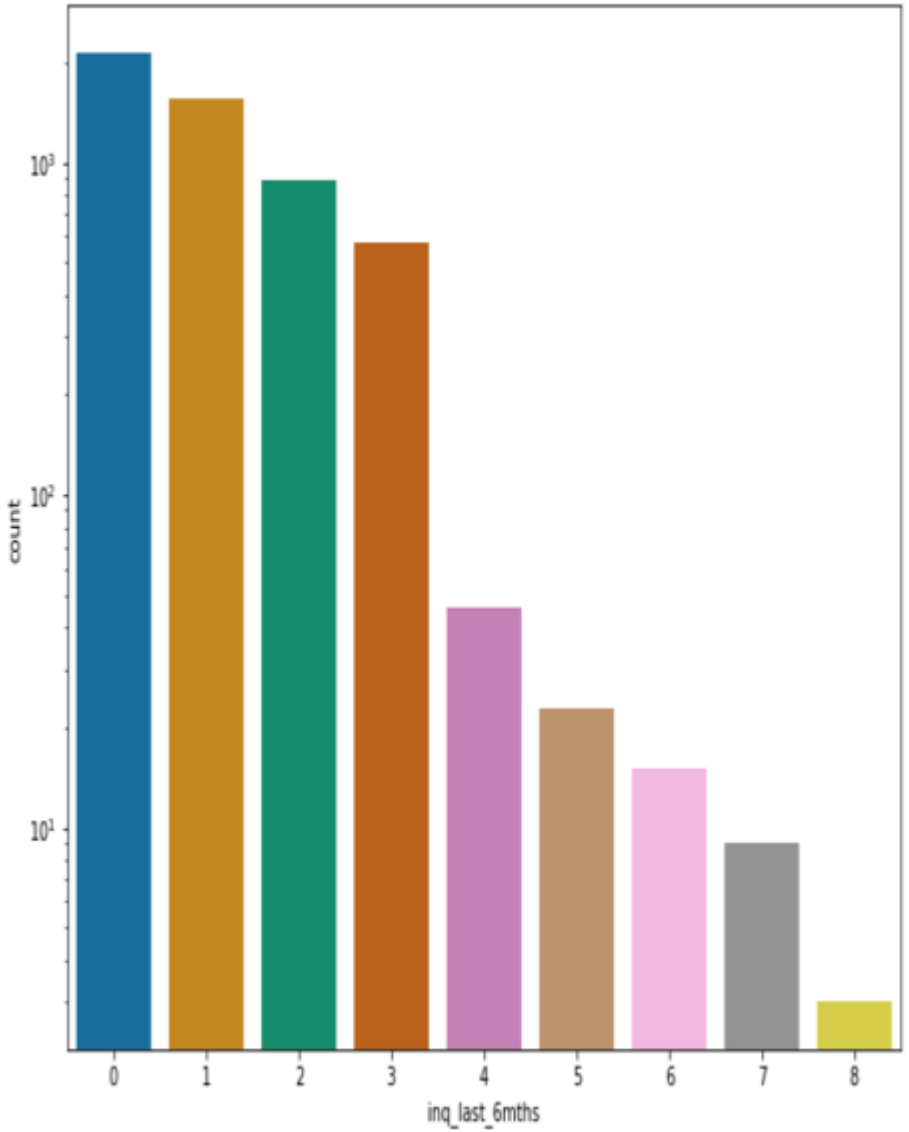


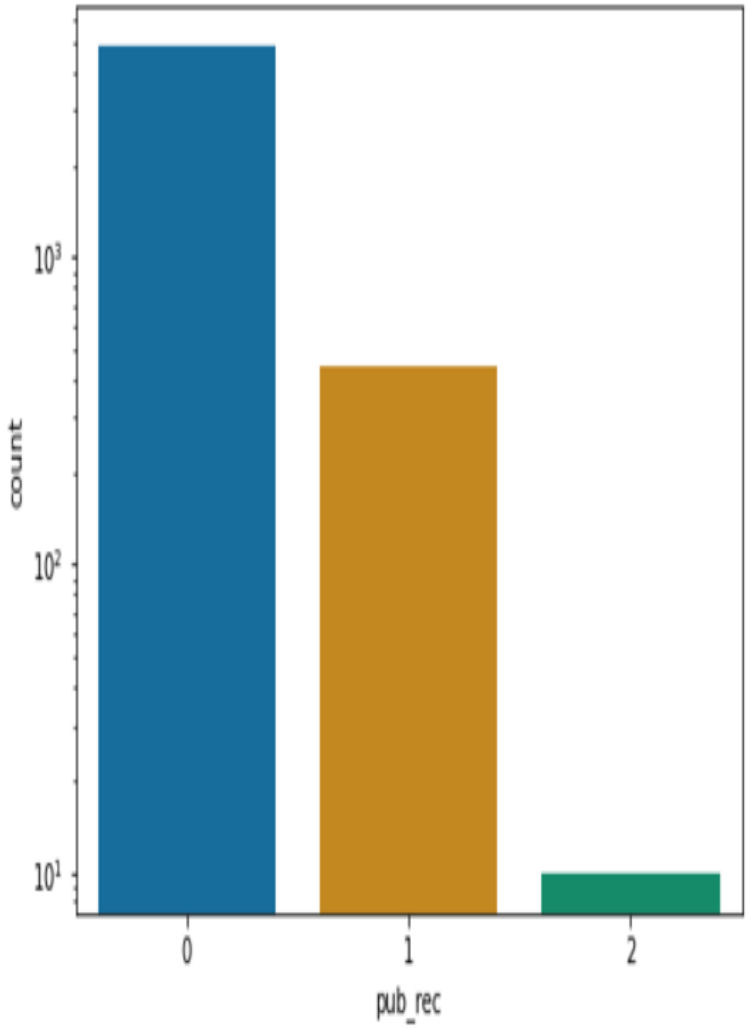
Univariate Analysis



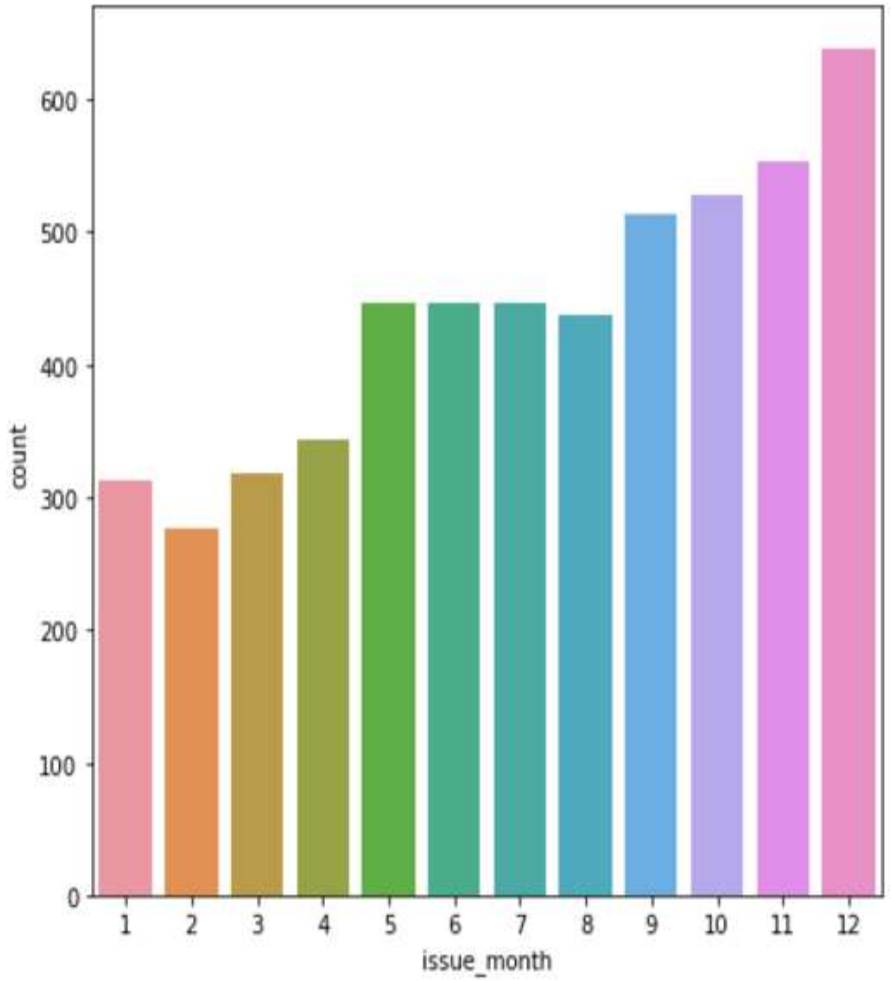


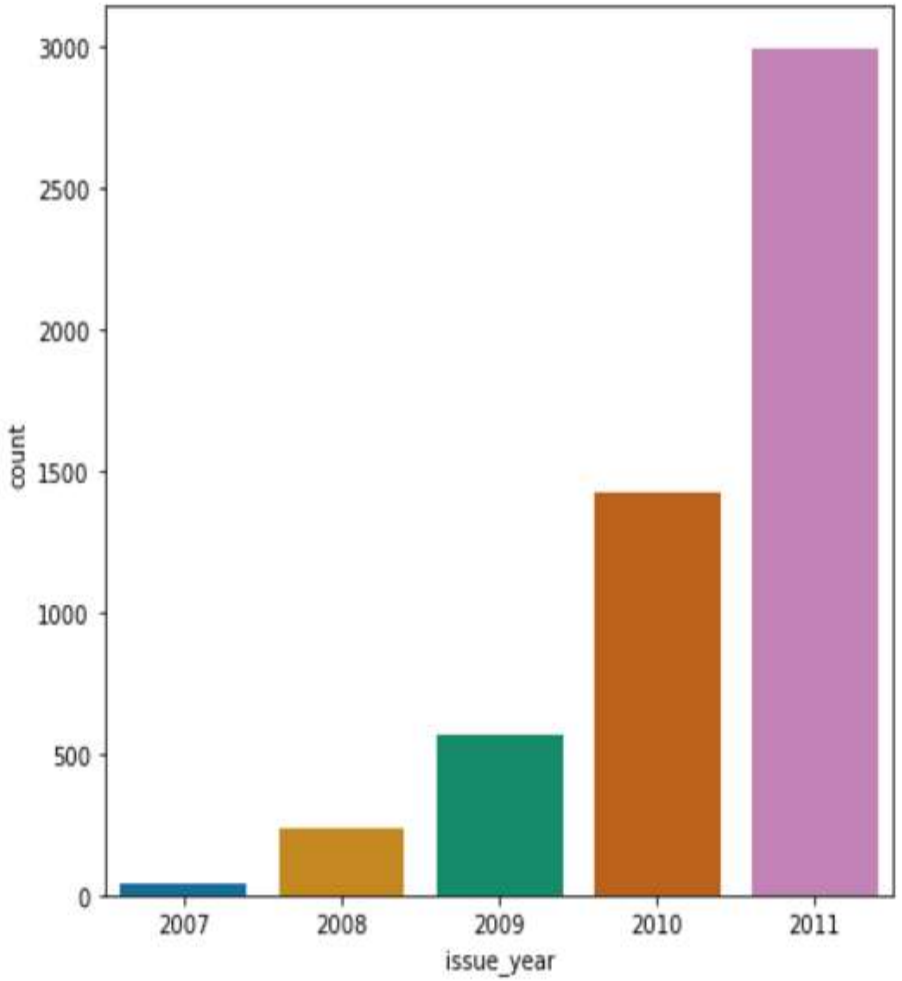
Univariate Analysis



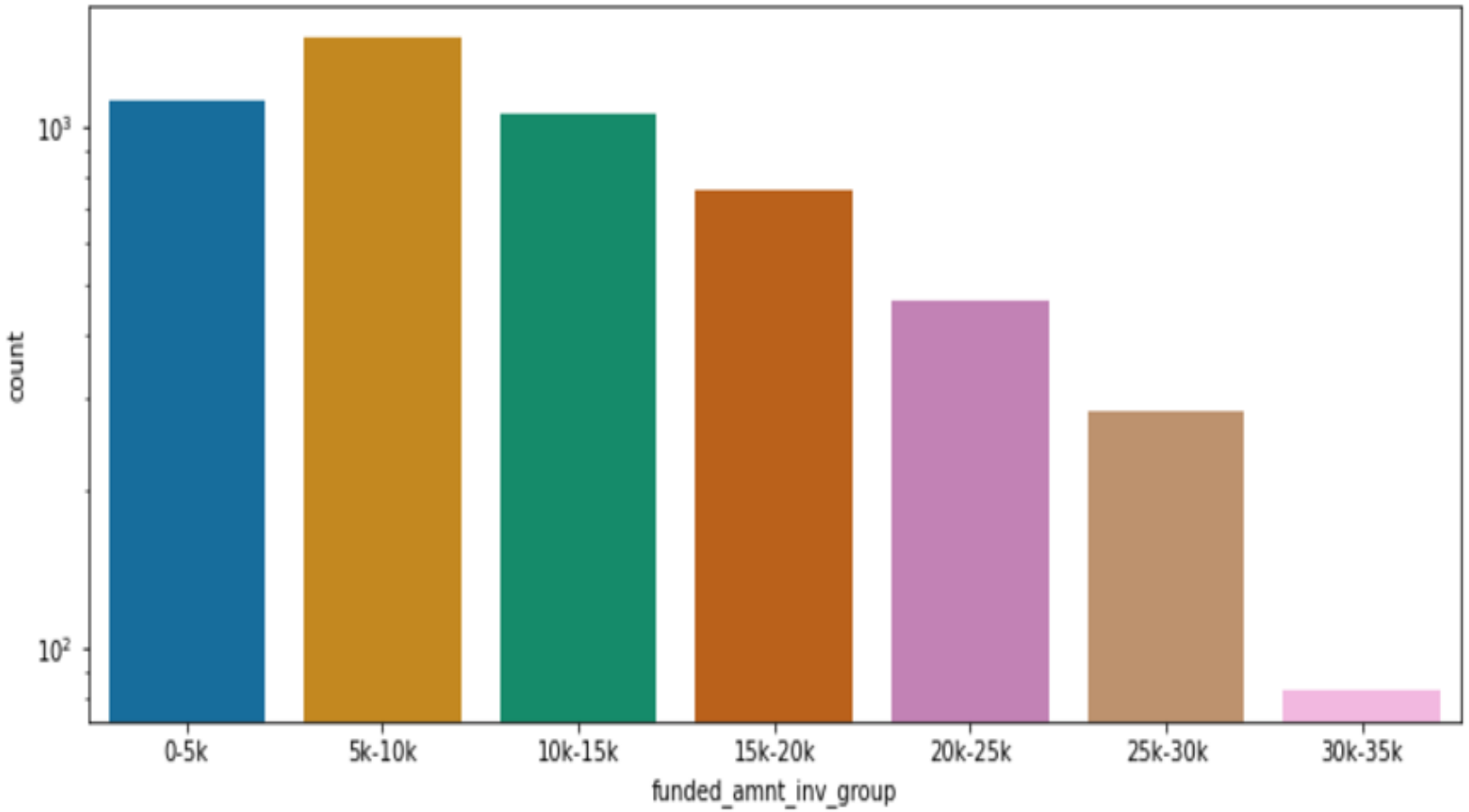


Univariate Analysis

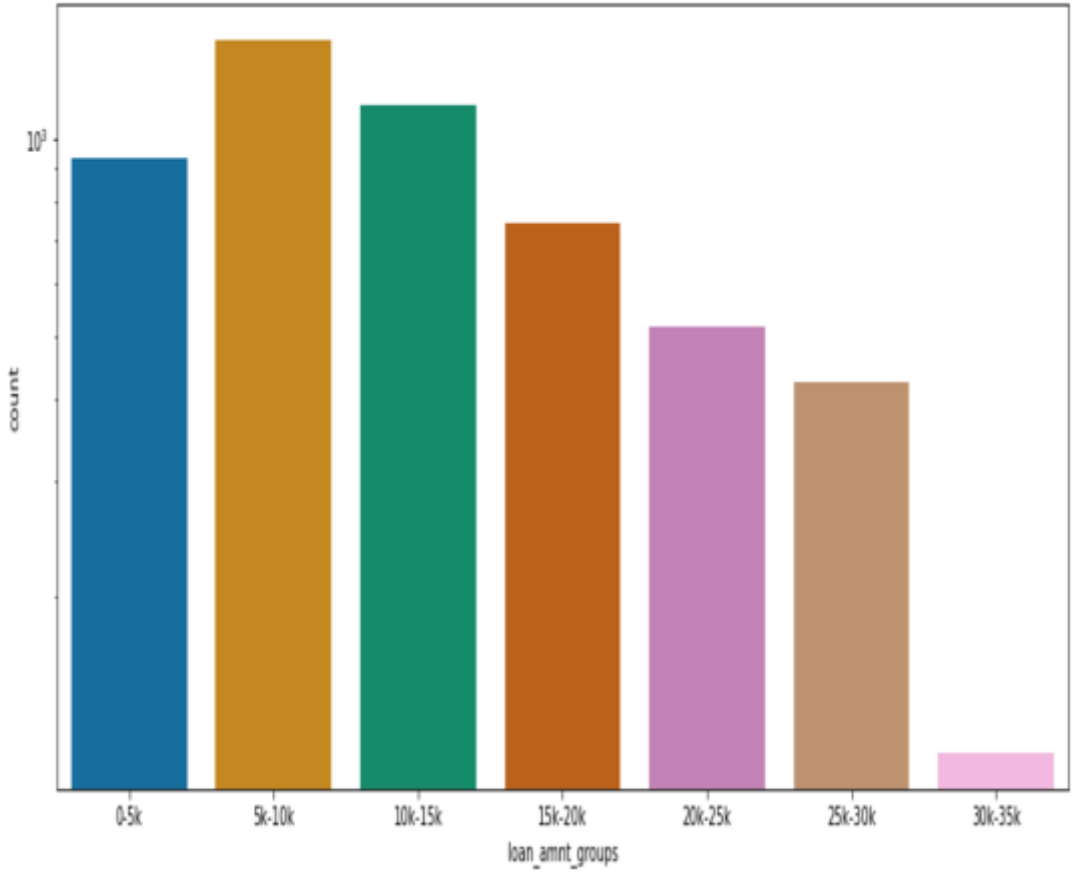


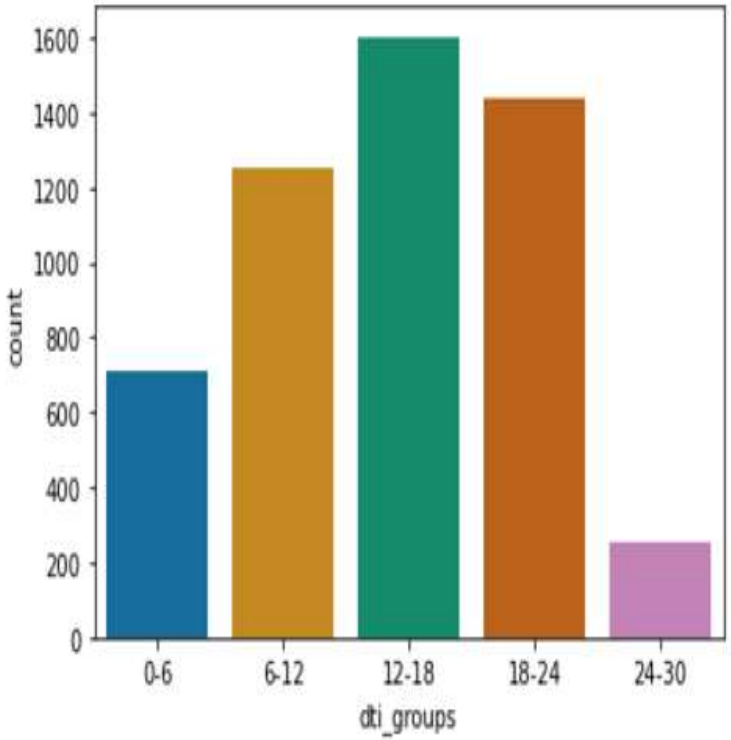


Univariate Analysis

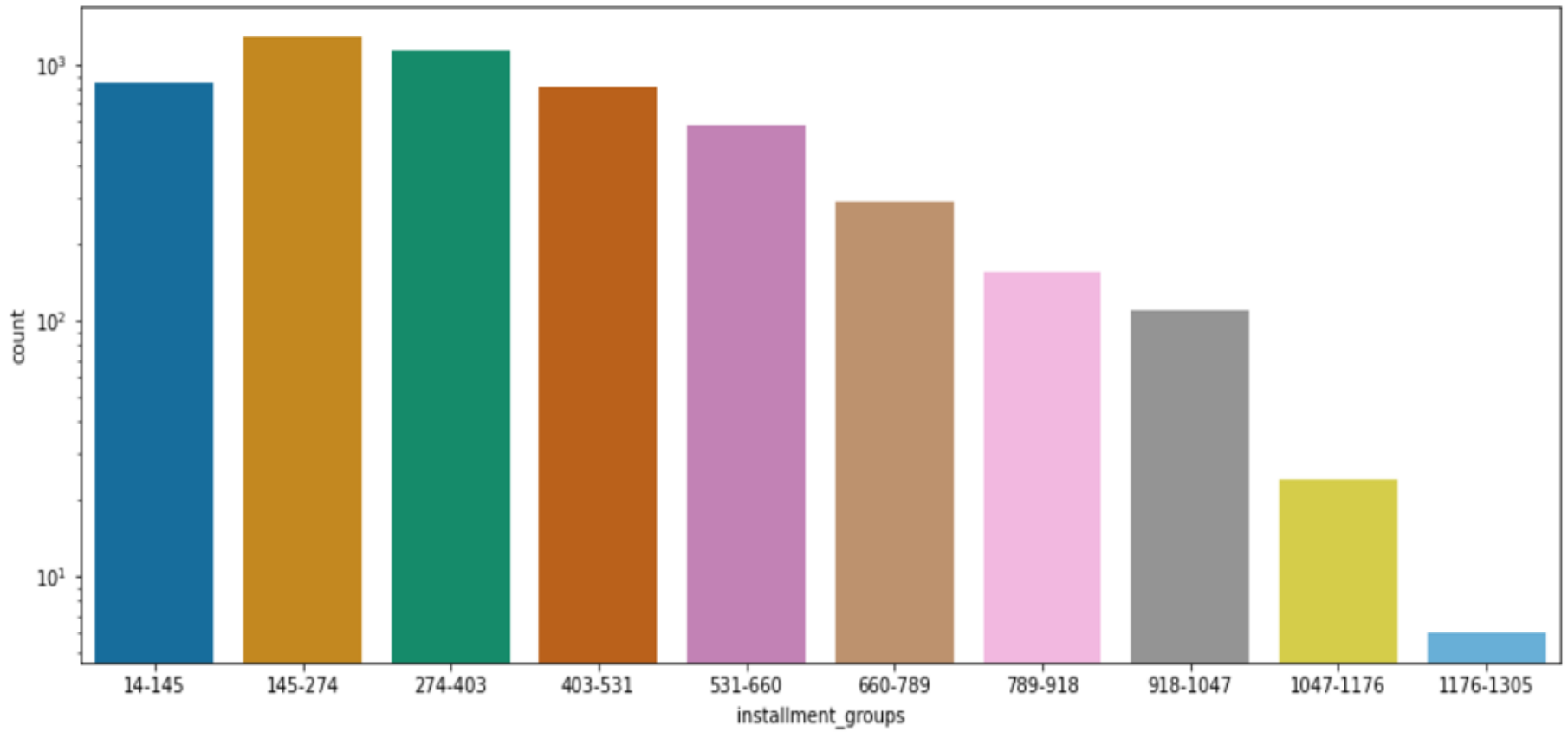


Univariate Analysis





Univariate Analysis



Univariate Analysis

Observations & Inferences:

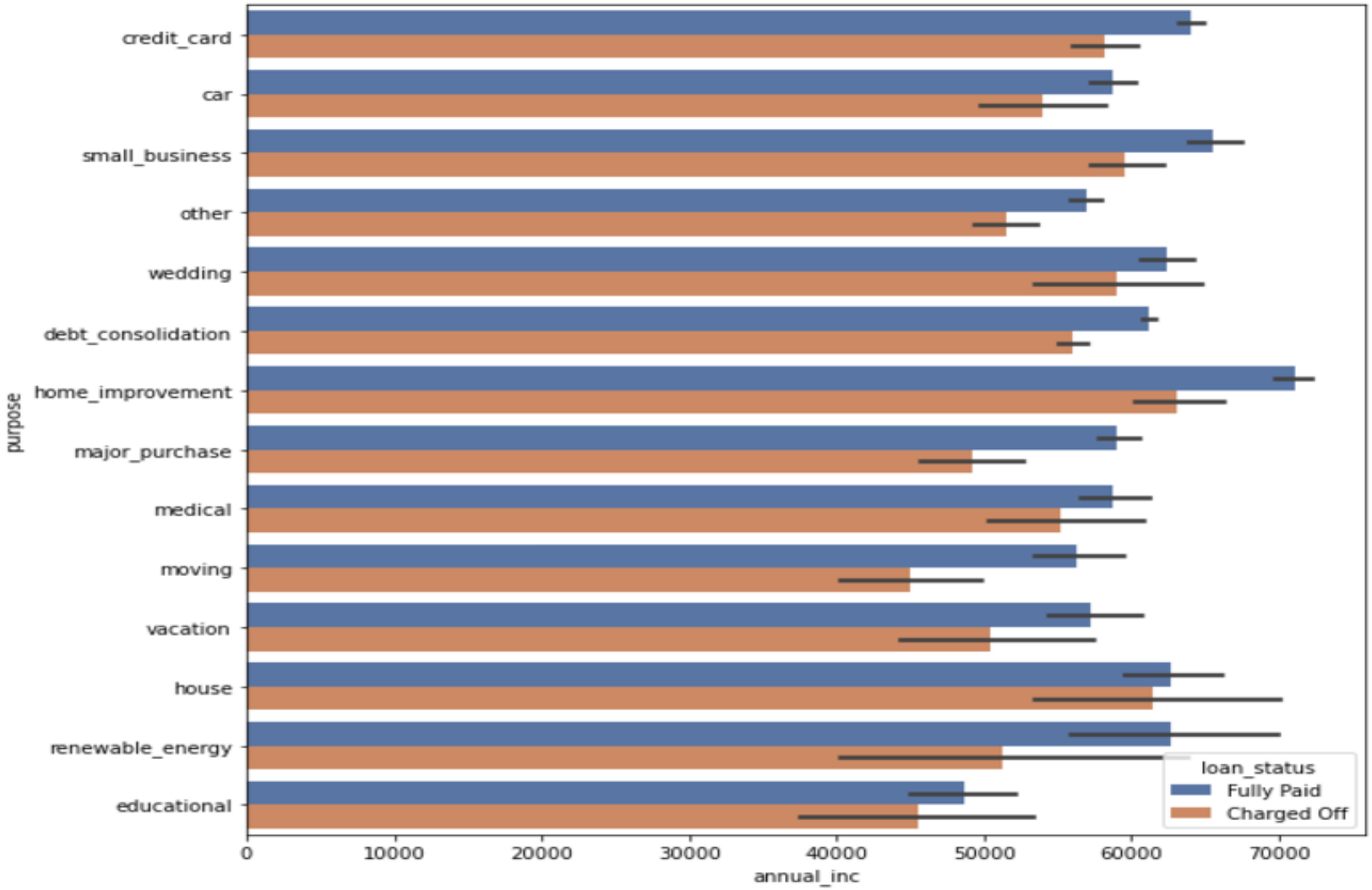
The above analysis with respect to the charged off loans for each variable suggests the following. There is a more probability of defaulting when :

- Applicants having house_ownership as 'RENT'
- Applicants who use the loan to clear other debts
- Applicants who receive interest at the rate of 13-17%
- Applicants who have an income of range 31201 - 58402
- Applicants who have 20-37 open_acc
- Applicants with employement length of 10
- When funded amount by investor is between 5000-10000
- Loan amount is between 5429 - 10357
- Dti is between 12-18
- When monthly installments are between 145-274
- Term of 36 months
- When the loan status is Not verified
- When the no of enquiries in last 6 months is 0
- When the number of derogatory public records is 0
- When the purpose is 'debt_consolidation'
- Grade is 'B'
- And a total grade of 'B5' level.

Bivariate Analysis

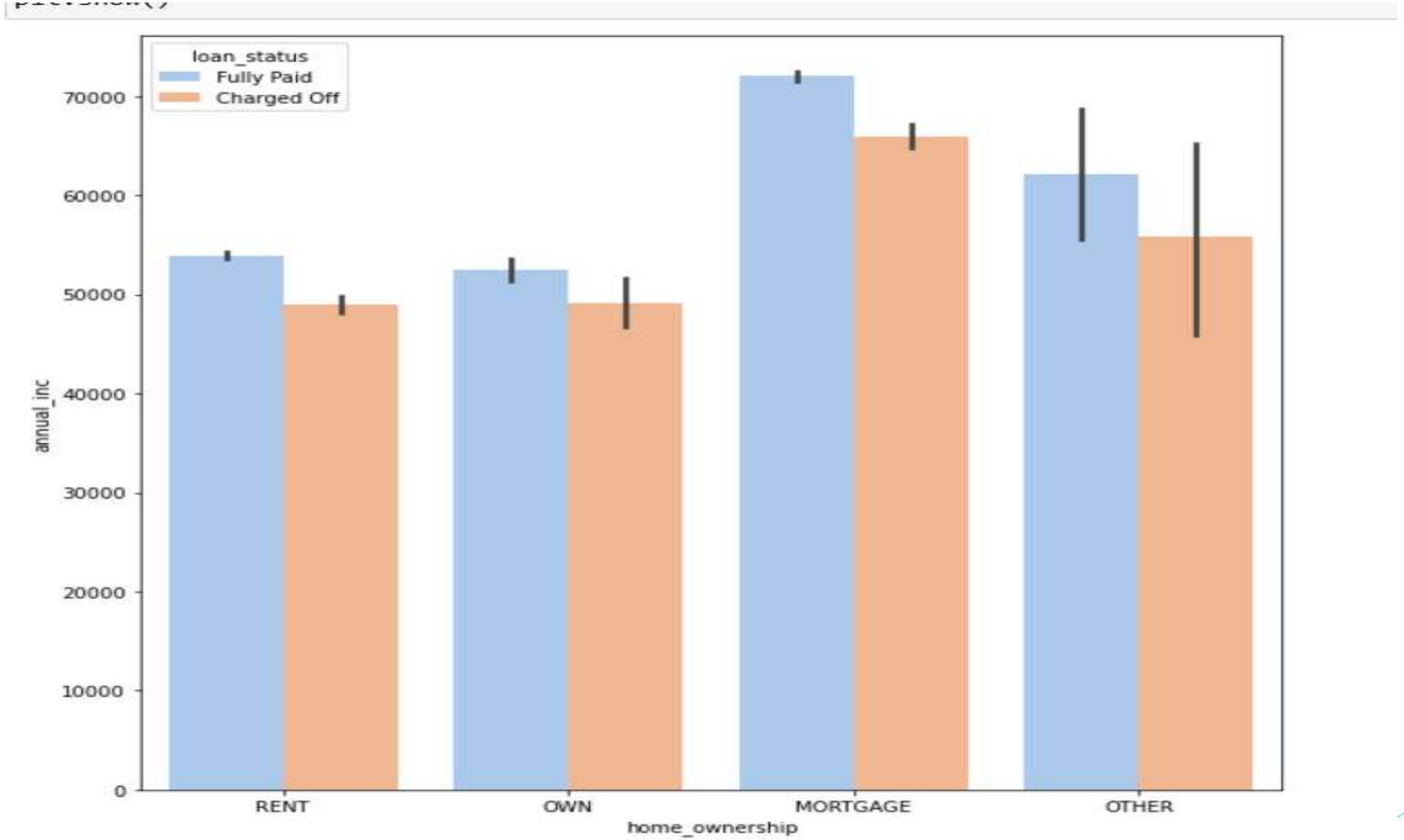
Bivariate analysis is a statistical method that involves the simultaneous analysis of two variables (factors). It aims to determine the empirical relationship between them. The analysis can be used to test hypotheses, identify patterns, or explore relationships between the variables.

Bivariate Analysis



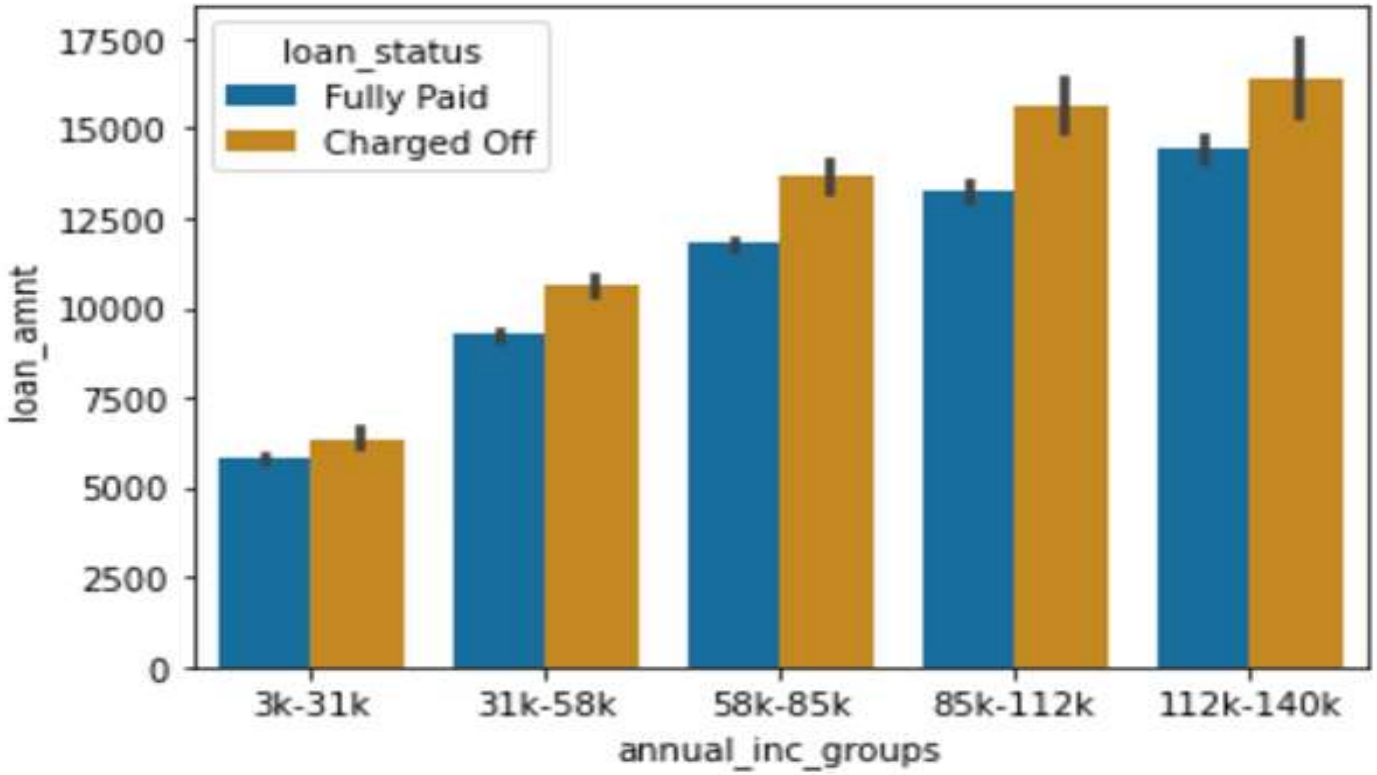
1.Annual income vs loan purpose

Bivariate Analysis



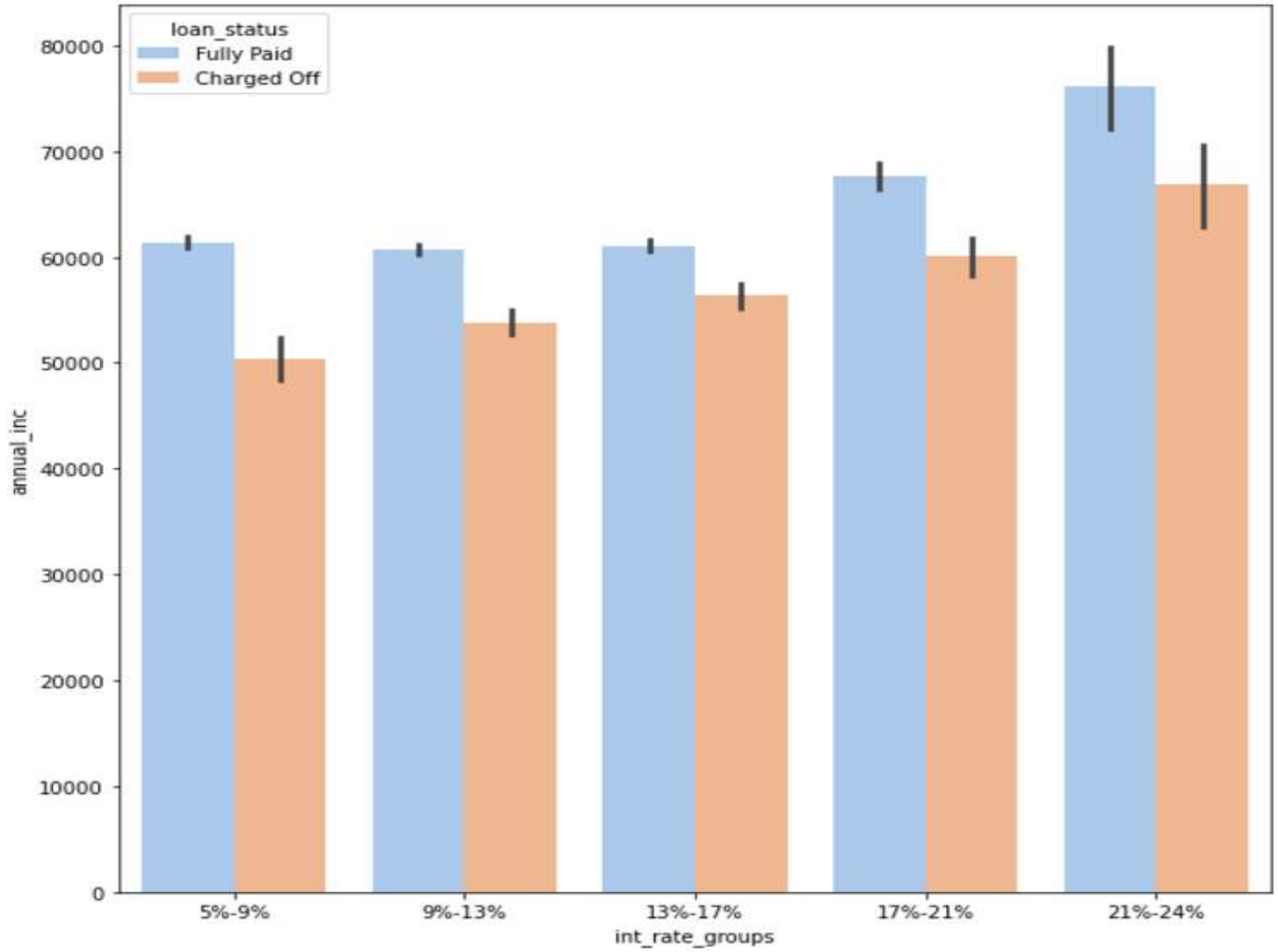
2.Annual income vs home ownership

Bivariate Analysis



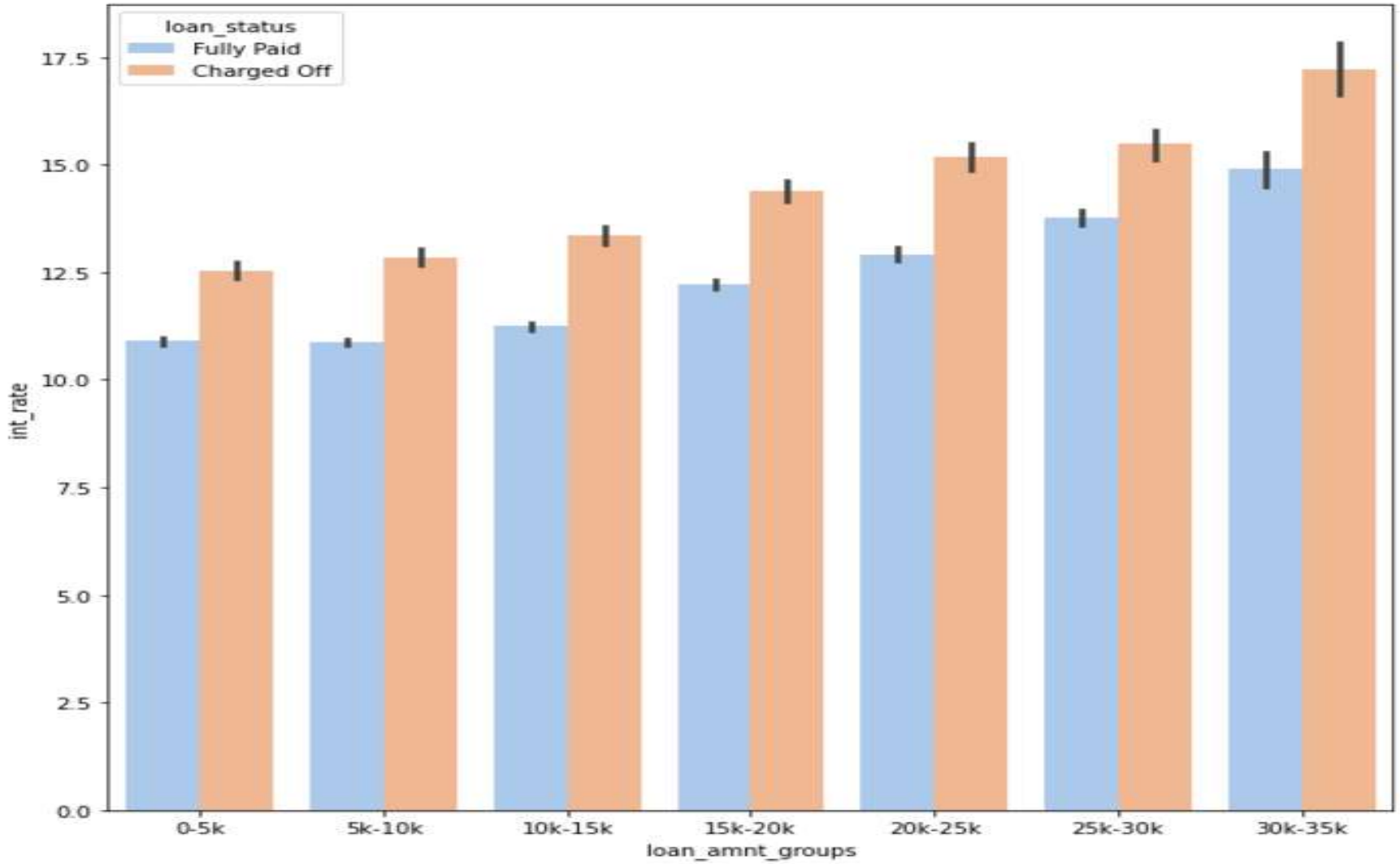
Annual Income vs Loan amount

Bivariate Analysis



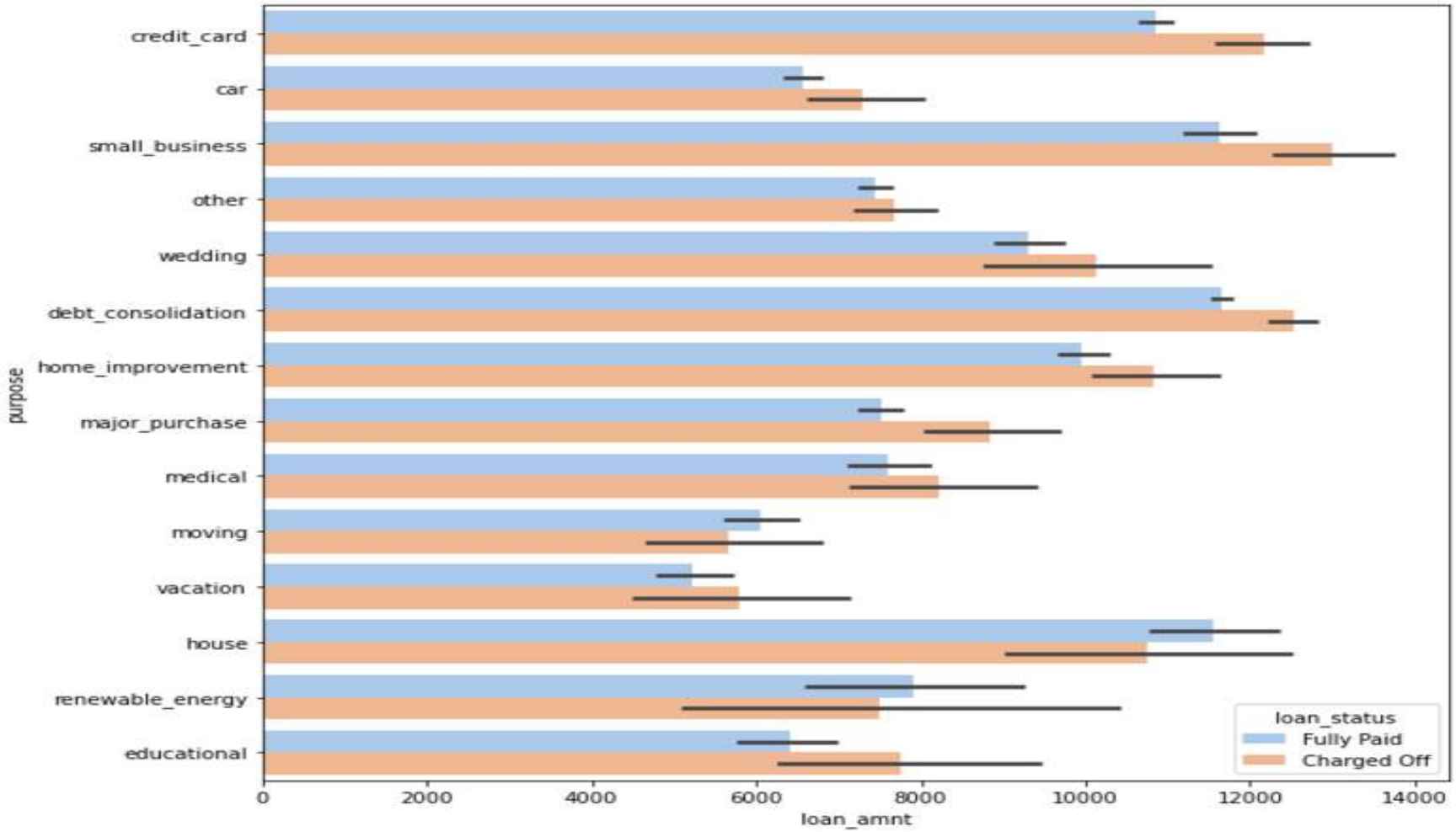
Annual income vs int_rate

Bivariate Analysis



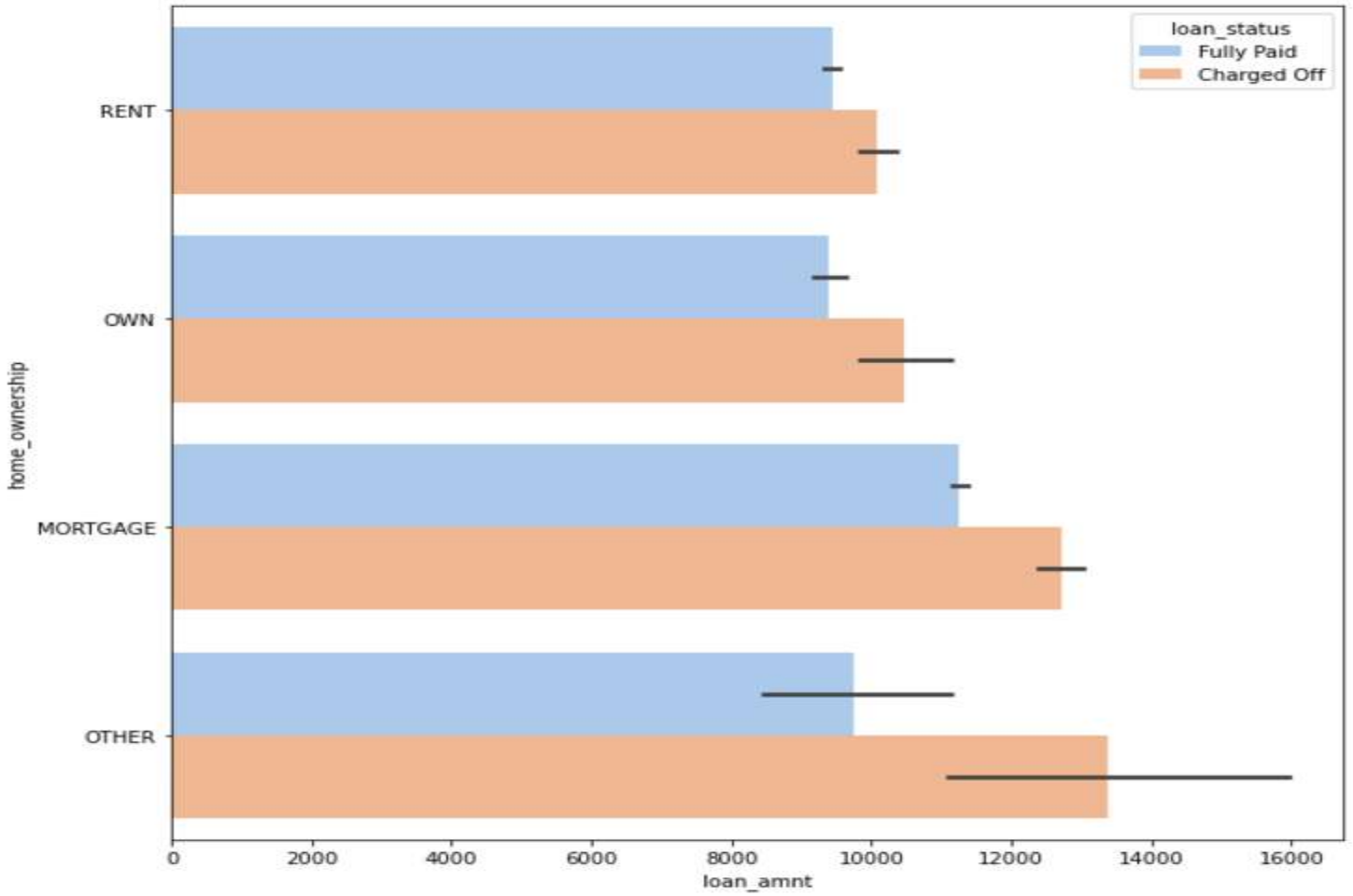
Loan Amount vs Interest Rate

Bivariate Analysis



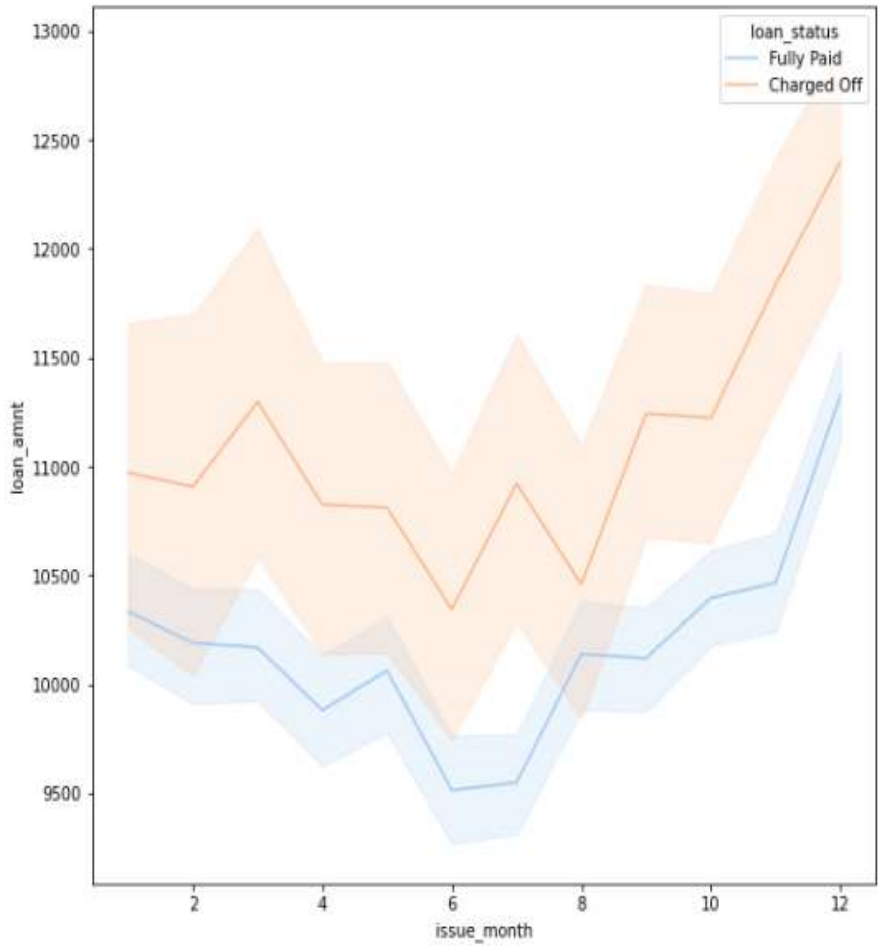
Loan vs Loan purpose

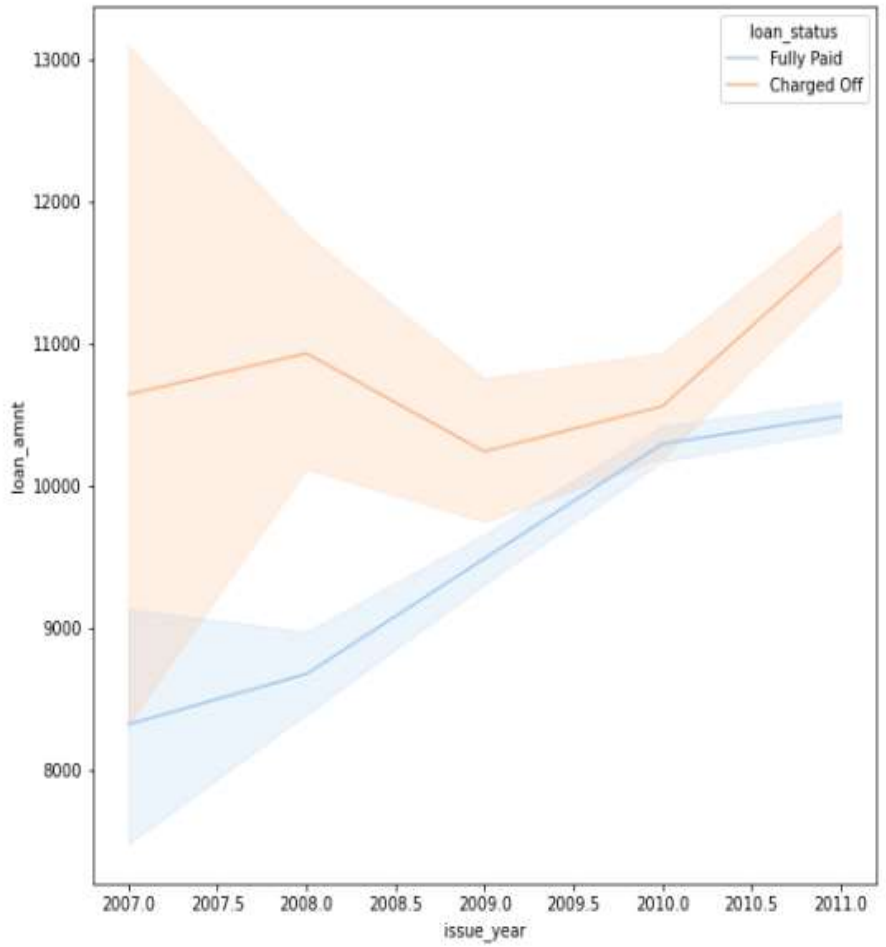
Bivariate Analysis



Loan vs House Ownership

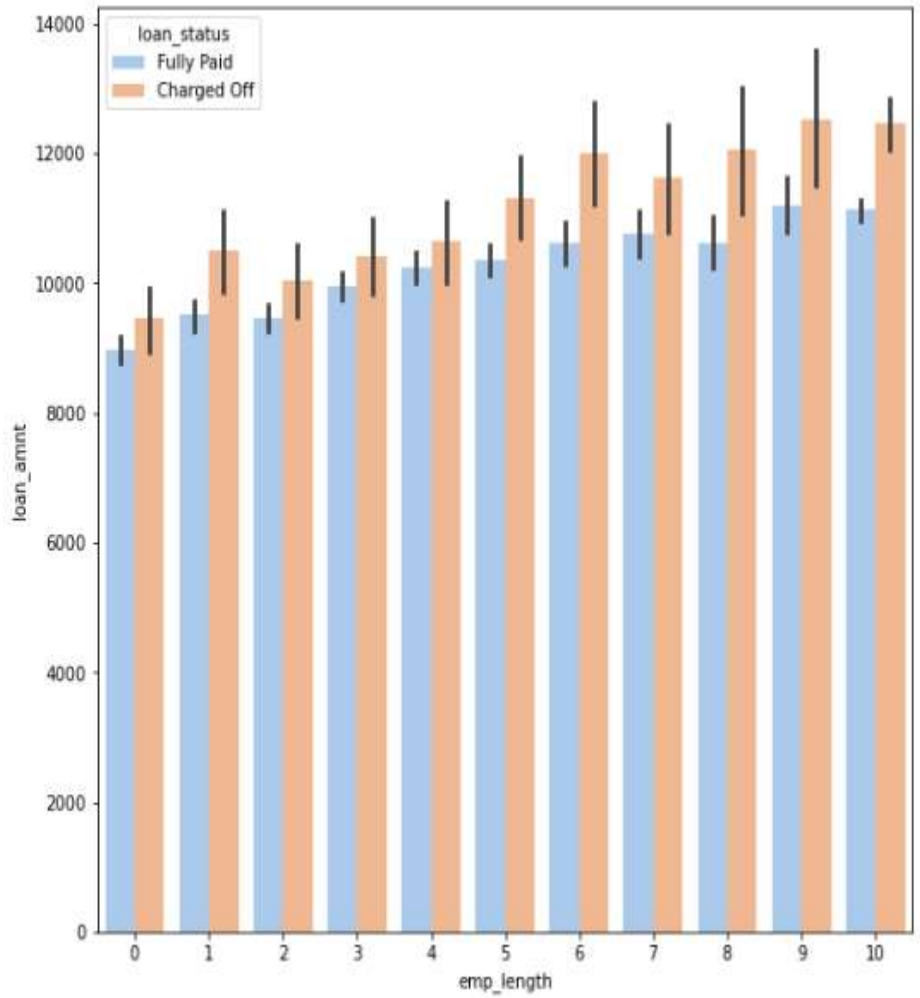
Bivariate Analysis

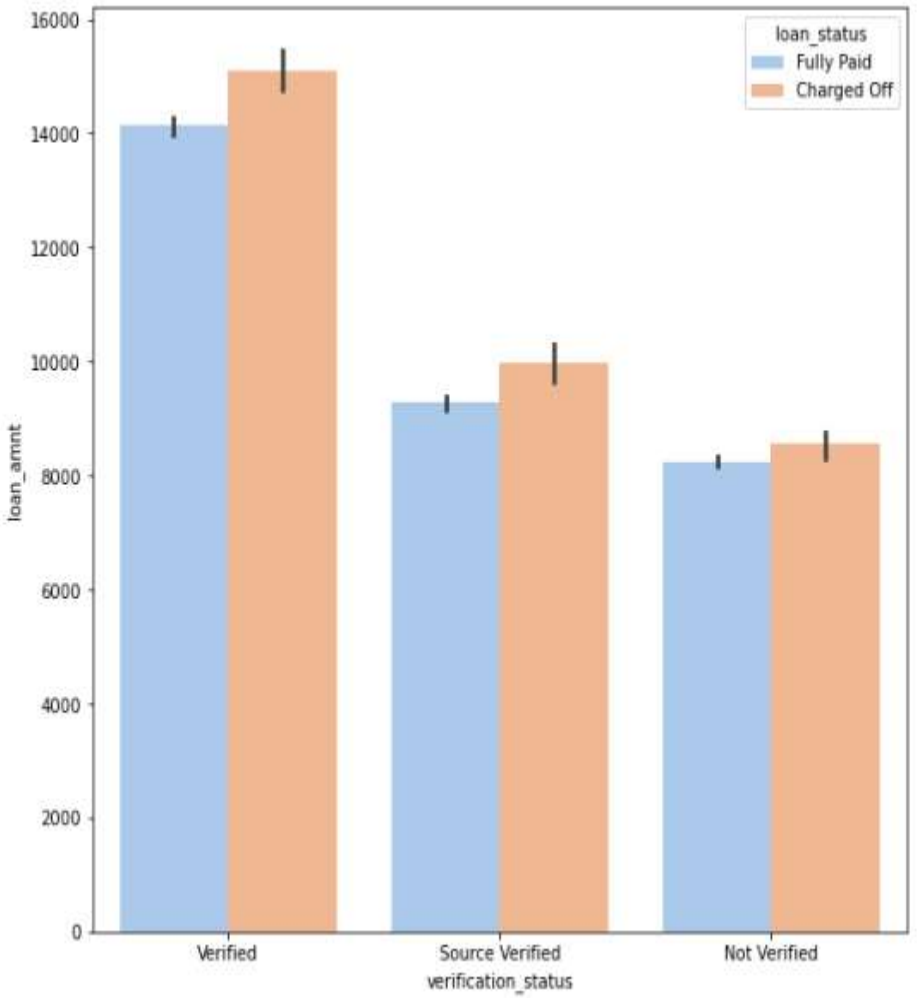




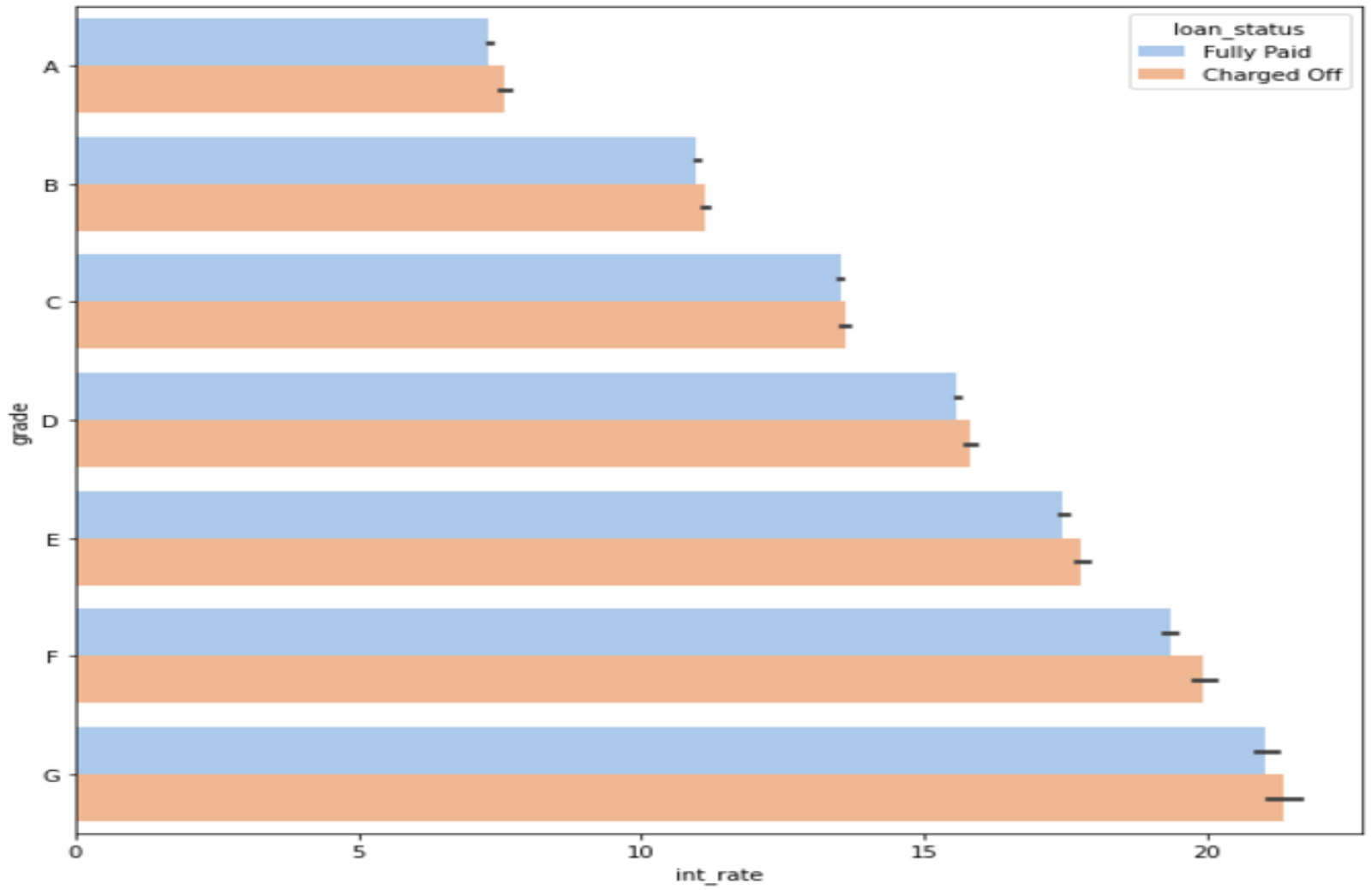
Loan amount vs month issued and year issued

Bivariate Analysis



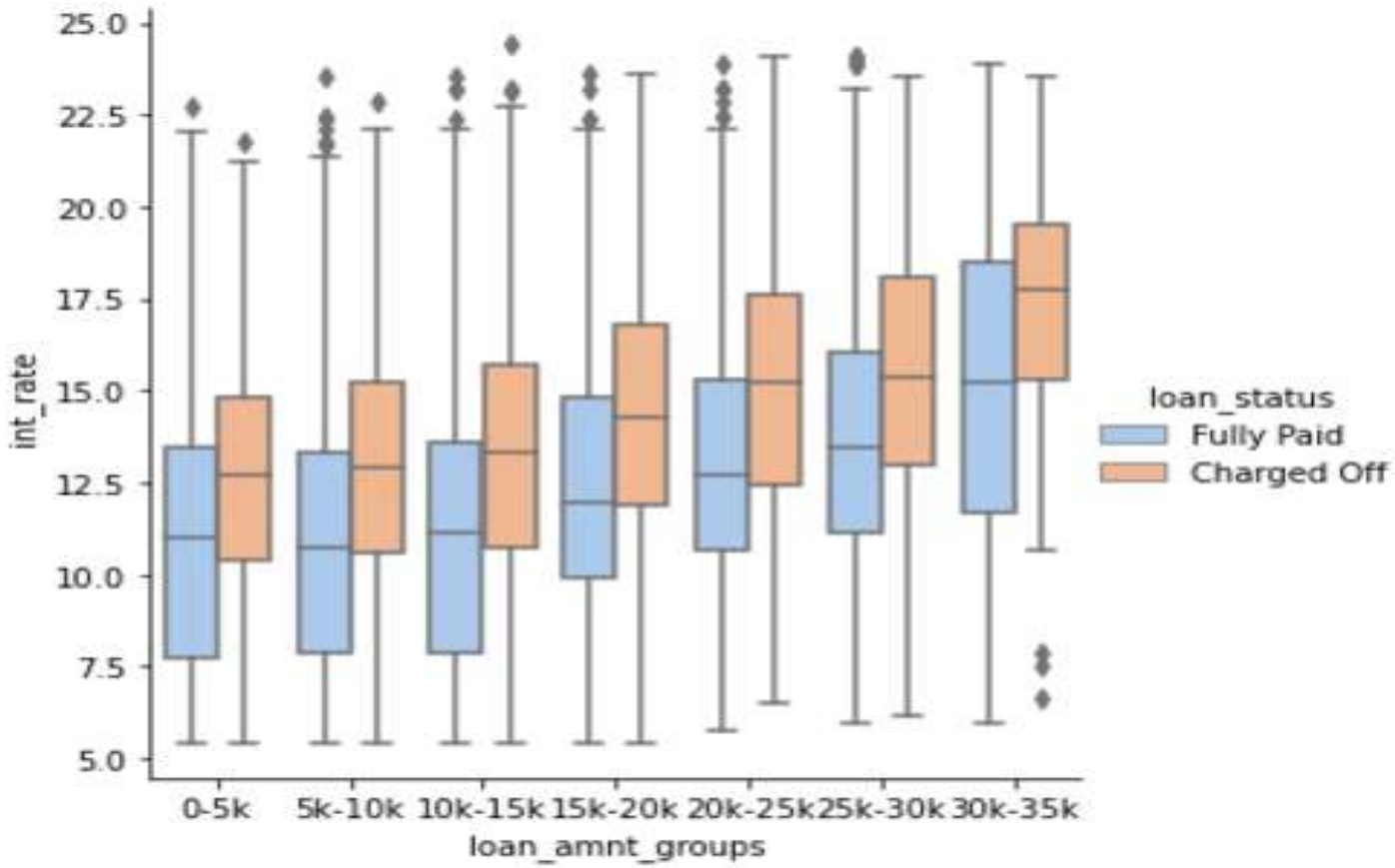


Bivariate Analysis

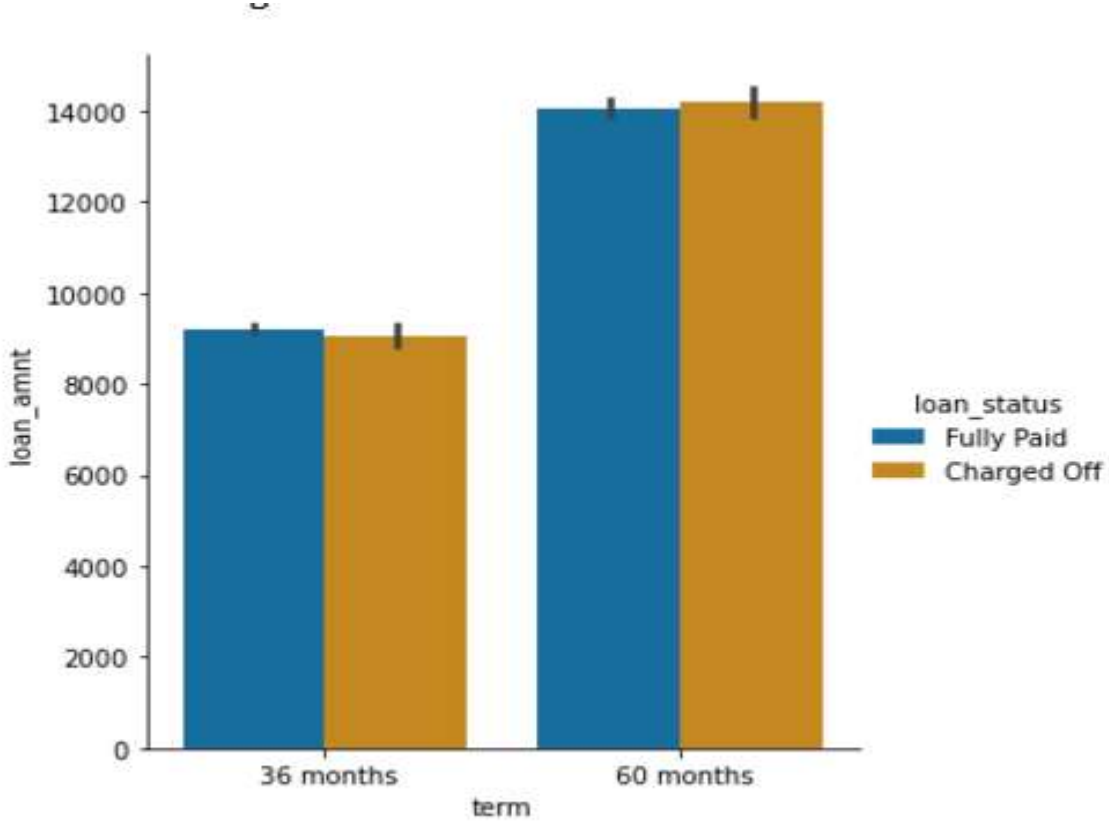


grade vs interest rate

Bivariate Analysis



Bivariate Analysis



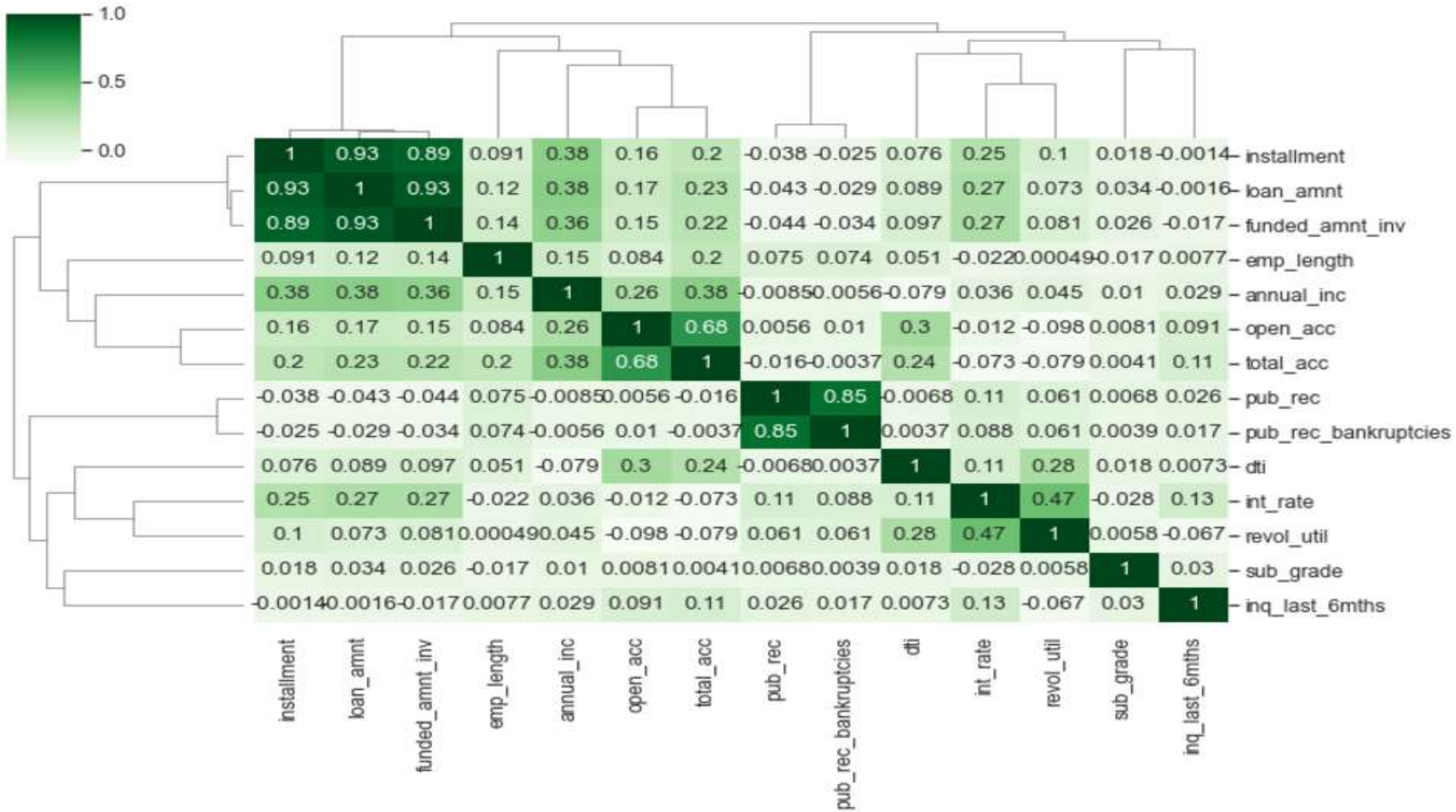
Bivariate Analysis

Observations

The above analysis with respect to the charged off loans. There is a more probability of defaulting when :

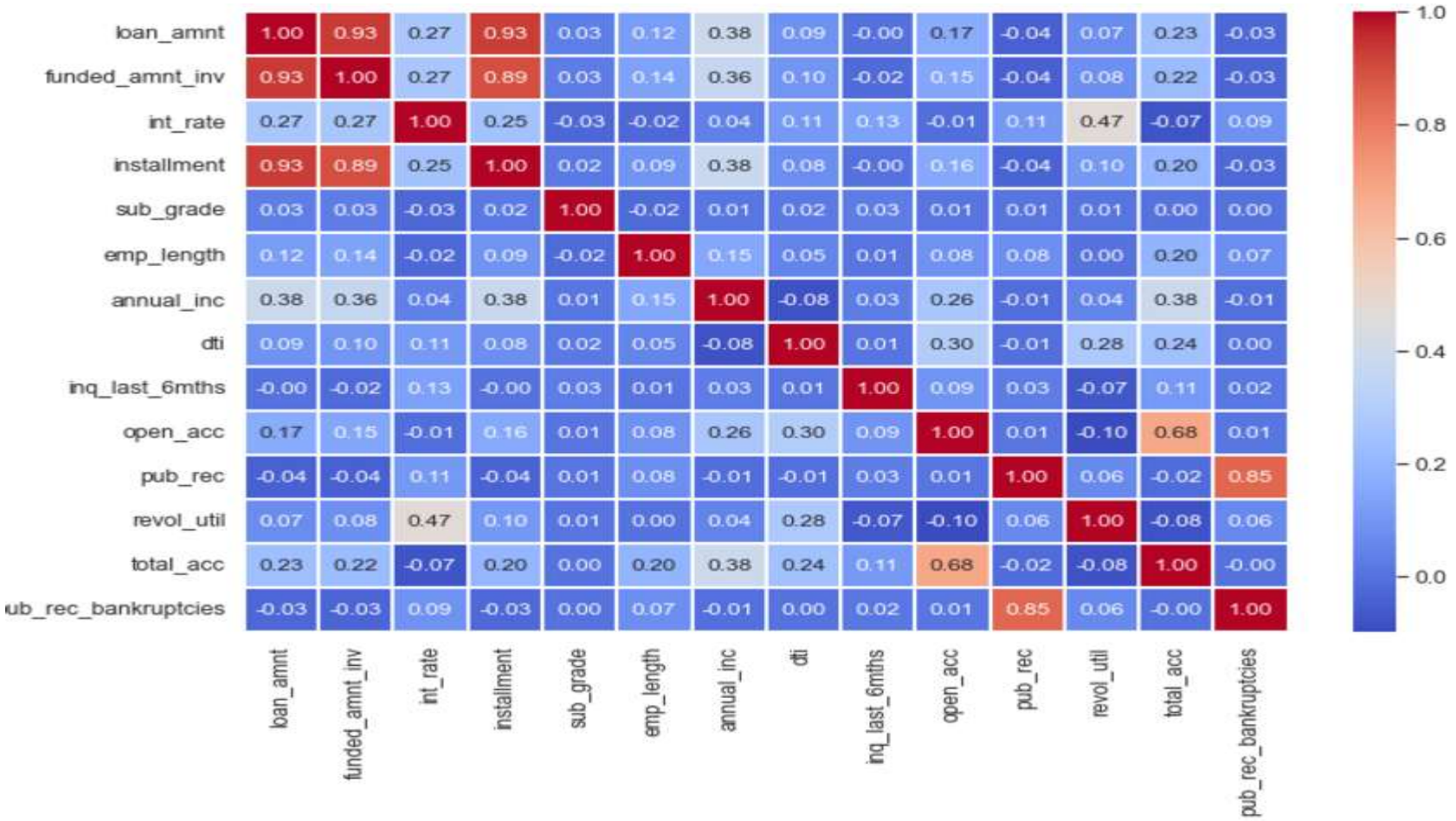
- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%

Correlation Analysis



Correlation Matrix among variables namely **installment, funded_amnt_inv, funded_amnt, loan_amnt, pub_rec_bankrupticies, annual_inc, emp_length, dti, int_rate**

Correlation Analysis



Correlation Matrix among variables namely **installment, funded_amnt_inv, funded_amnt, loan_amnt, pub_rec_bankrupticies, annual_inc, emp_length, dti, int_rate**

Correlation Analysis

Inferences from Correlation Metrics

Strong Correlation

- `installment` has a strong correlation with funded_amnt, loan_amnt, and funded_amnt_inv
- `annual_inc` has a strong correlation with loan_amount
- `open_acc` has a strong correlation with total_acc

Weak Correlation

- `dti` has weak correlation with most of the fields
- `emp_length` has weak correlation with most of the fields

Negative Correlation

- `pub_rec_bankrupticies` has a negative correlation with almost every field
- `annual_inc` has a negative correlation with dti

Thank You!