Lending Club Case Study

Submitted by:
Jarvis R

Contents

Problem Statement

Business Context: Understanding risk analytics in banking and financial services, particularly in the context of loan approval and minimizing financial loss.

Company Profile: Working for a consumer finance company specializing in lending various types of loans to urban customers, where loan approval decisions are crucial for minimizing business loss and financial risk.

Risk Factors: Two main risks associated with loan decisions: loss of business if a potential borrower who would repay the loan is denied, and financial loss if a borrower defaults on the loan after approval.

Objective: Using Exploratory Data Analysis (EDA) to identify patterns in past loan applicant data, aiming to predict default tendencies and minimize credit loss by identifying risky loan applicants.

Decision Outcomes: Loan applicants may be accepted or rejected, with accepted applicants possibly fully paying, being in the process of payment, or defaulting (charged-off).

Outcome Goals: The company seeks to understand the key factors driving loan default to enhance risk assessment and minimize credit loss, thus improving portfolio management and lending decisions.

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 us 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.

Data Understanding

Dataset Attributes:

Primary Attribute

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

- **Fully-Paid:** Signifies customers who have successfully repaid their loans.
- **Charged-Off:** Indicates customers who have been labeled as "Charged-Off" or have defaulted on their loans.
- **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.

Data Understanding

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.

Data Understanding

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 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 Understanding

Excluded Columns: In our analysis, we will not consider certain types of columns. It's important to note that this is a general categorization of the columns we will exclude from our approach, and it does not represent an exhaustive list.

- **Customer Behavior Columns**- Columns that describe customer behavior will not be factored into our analysis. The current analysis focuses on the loan application stage, while customer behavior variables pertain to post-approval actions. Consequently, these attributes will not influence the loan approval/rejection process.
- **Granular Data** - Columns providing highly detailed information that may not be necessary for our analysis will be omitted. For example, while the "grade" column may have relevance in creating business outcomes and visualizations, the "sub grade" column is excessively granular and will not be utilized in our analysis.
- Columns contain **NA** values only, and these columns will be removed.
- Columns (**emp_title, desc, title**) will be dropped as they contain descriptive text (nouns) and do not contribute to the analysis.
- Columns with **single value** that do not contribute to the analysis will be removed.
- Columns with more than **50%** of data being empty will be dropped.
- Columns (**id, member_id**) will be dropped as they are index variables with unique values and do not contribute to the analysis.
- The redundant column (**url**) will be dropped. Further analysis reveals that the URL is a static path with the loan ID appended as a query, making it redundant compared to the (**id**) column.

Data Cleaning and Pre-processing

- **Loading data from loan CSV:** While loading the dataset, some of the variables had mixed data types so they have to be converted accordingly as per analysis.
- **Checking for null values in the dataset:** There're many columns with null values. So they need to be dropped as they won't play a role in the analysis of the dataset. Roughly 48% of the columns were dropped.
- **Checking for unique values:** If the column has only a single unique value, it does not **make any sense to include it as part of our data analysis. We need to find out those** columns and drop them from the dataset.
- **Checking for single valued:** If the column has single valued, it does not **make any sense to include it as part of our data analysis. We need to find out those** columns and drop them from the dataset.
- **Checking for duplicated rows in data:** No duplicate rows were found.

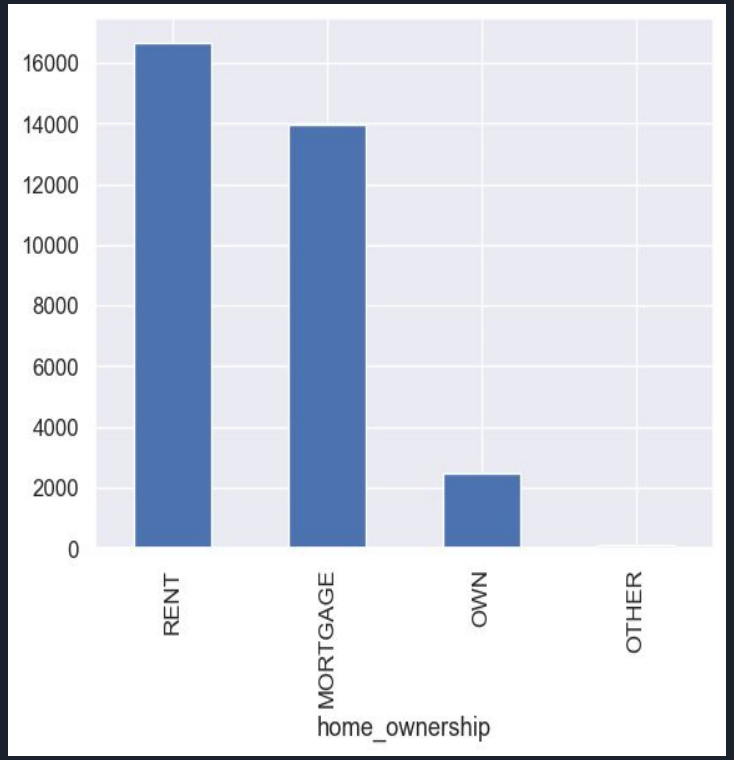
Data Cleaning and Pre-processing

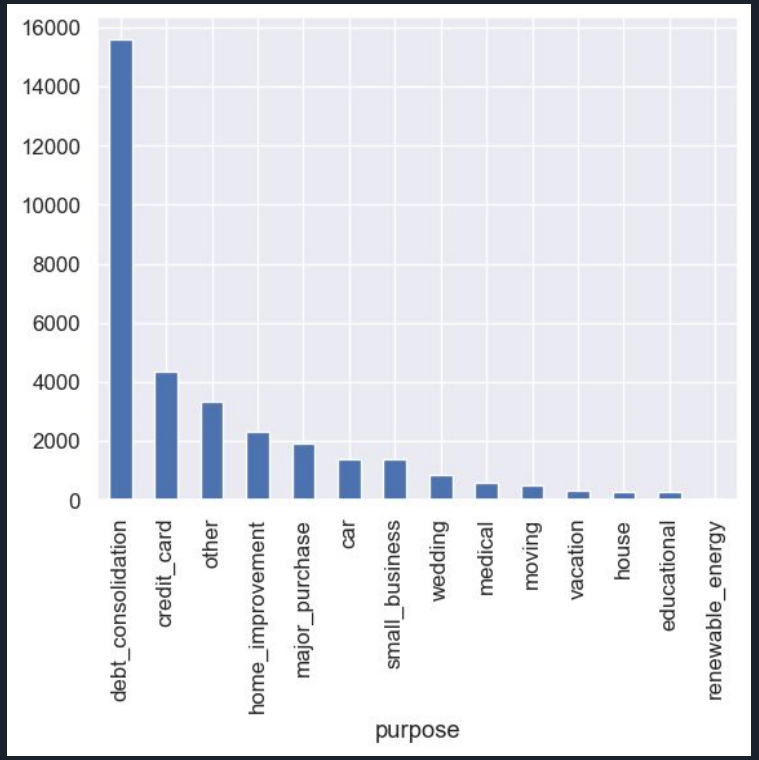
- **Dropped records where loan_status="Current" as the loan in progress cannot** provide us insights as to whether the borrower is likely to default or not.
- Dropping columns where missing data is **>=50%** as these columns will skew our data analysis and they need to be removed.
- Dropping extra columns containing text like **"delinq_2yrs", "earliest_cr_line", "last_pymnt_amnt", "inq_last_6mths", "open_acc", "pub_rec", "revol_bal", "revol_util", "total_acc", "out_prncp" , "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_late_fee", "recoveries", "collection_recovery_fee","application_type", "last_pymnt_d", "last_credit_pull_d", "total_rec_prncp", "total_rec_int", "emp_title", "zip_code"** as these will not contribute to loan pass or fail.
- **Common Functions:** Common functions were created for repeating common operations like plotting bar graphs, box plots, histograms, countplots, binning etc.
- **Data Conversion**: Converted columns such as **Term, interest rate(int_rate), debt to income (dti), loan amount (loan_amnt), funded amount (funded_amnt), grade, and employment length(emp_length)** to respective datatype to match the data. Also converted **loan date (issue_d)** to **DateTime (format: yyyy-mm-dd).**

Data Cleaning and Pre-processing

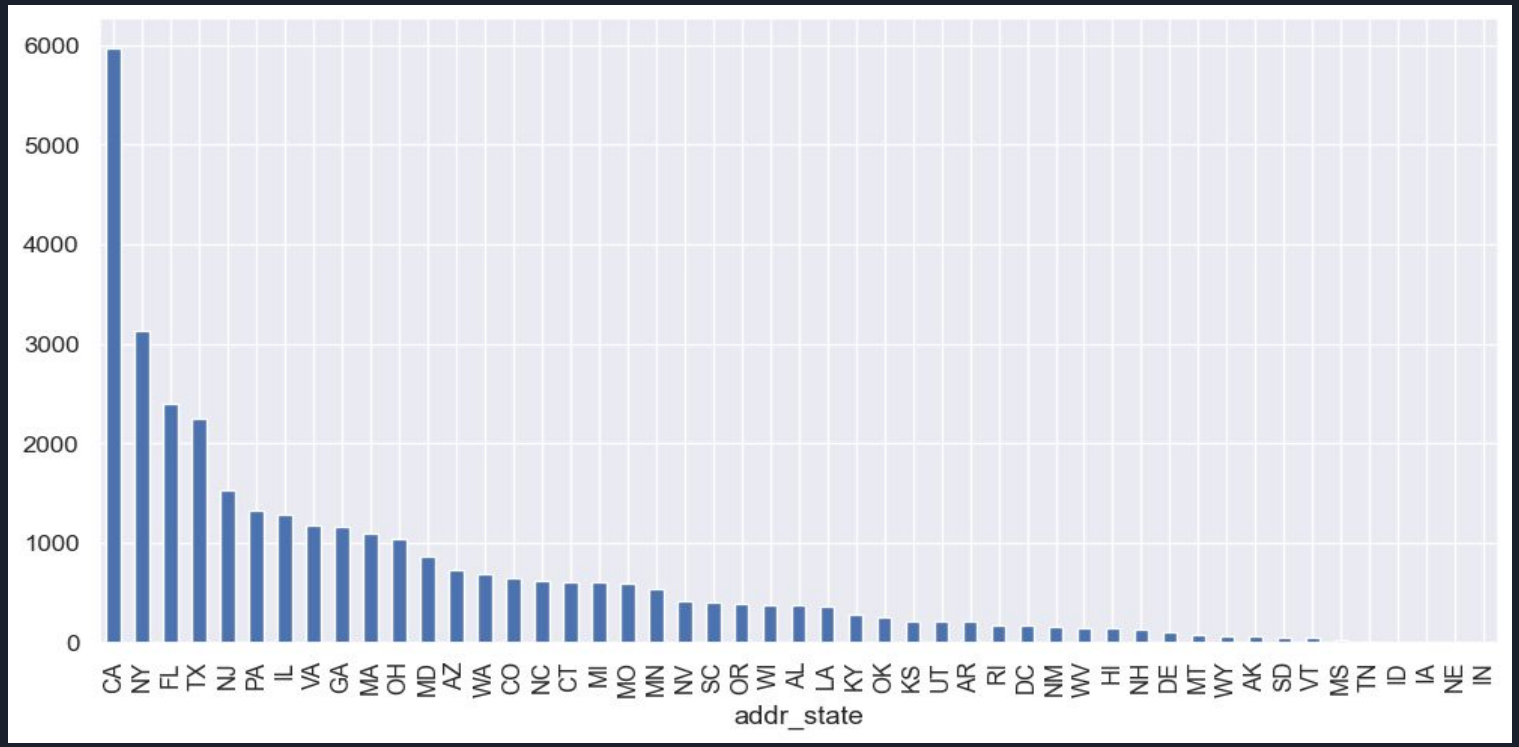
- **Imputing / Dropping the rows:** When checking the dataset for missing values, we can notice that the 'pub_rec_bankrupticies' column has 1.81% missing values, and the 'revol_util' column has 0.13%. Since the percentage of missing values in both columns is relatively low, we can to drop the rows containing missing values in the respective columns.
- **Outlier Treatment:** Calculated the **Interquartile Range (IQR)** and filtering out the outliers outside of lower and upper bound.

Univariate Analysis (Unordered Categorical)

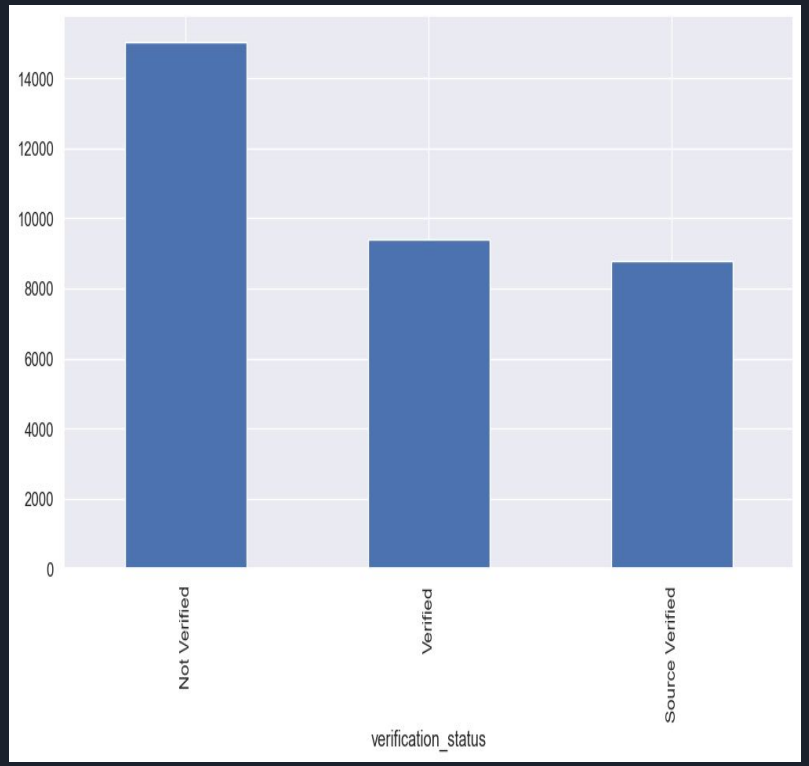


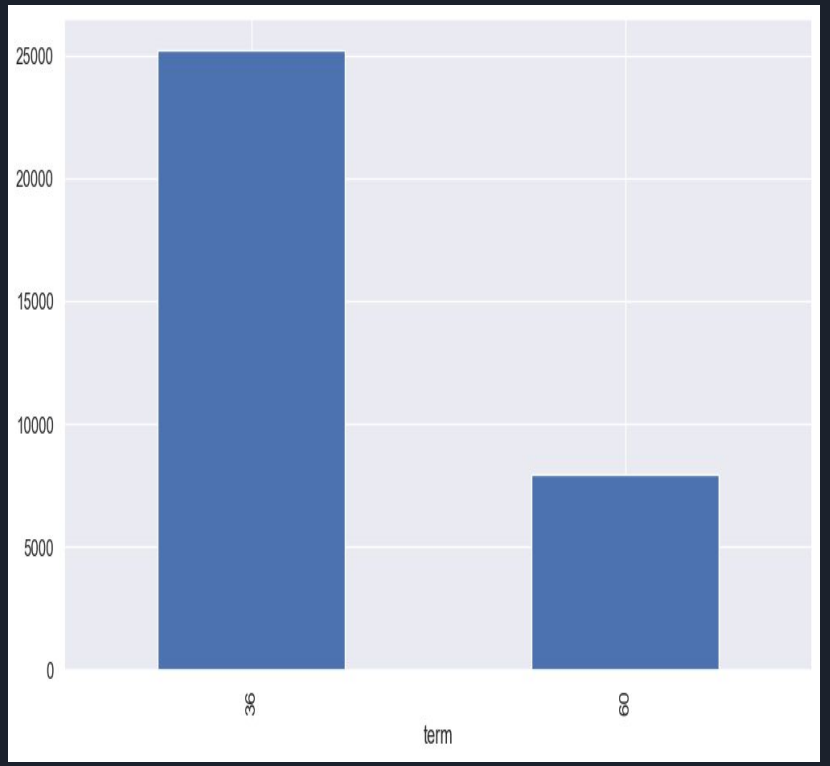


Univariate Analysis (Unordered Categorical)



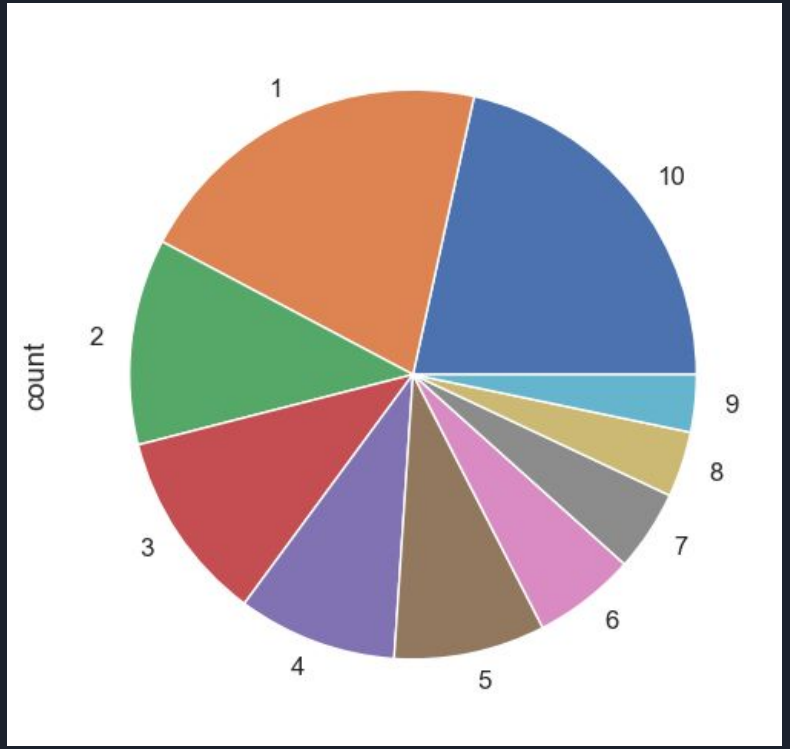
Univariate Analysis (Unordered Categorical)





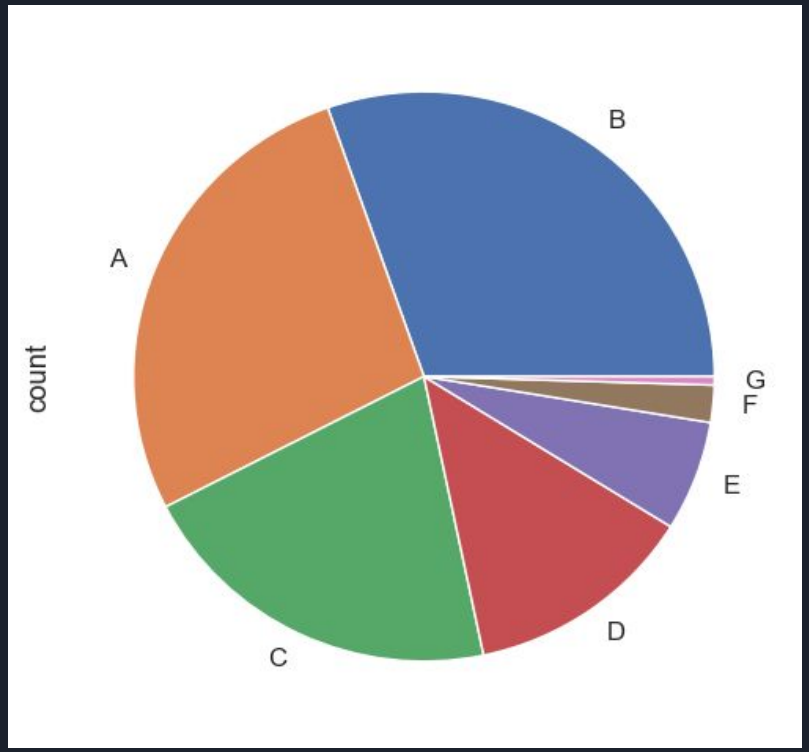
Univariate Analysis (Ordered Categorical)

Employment Length

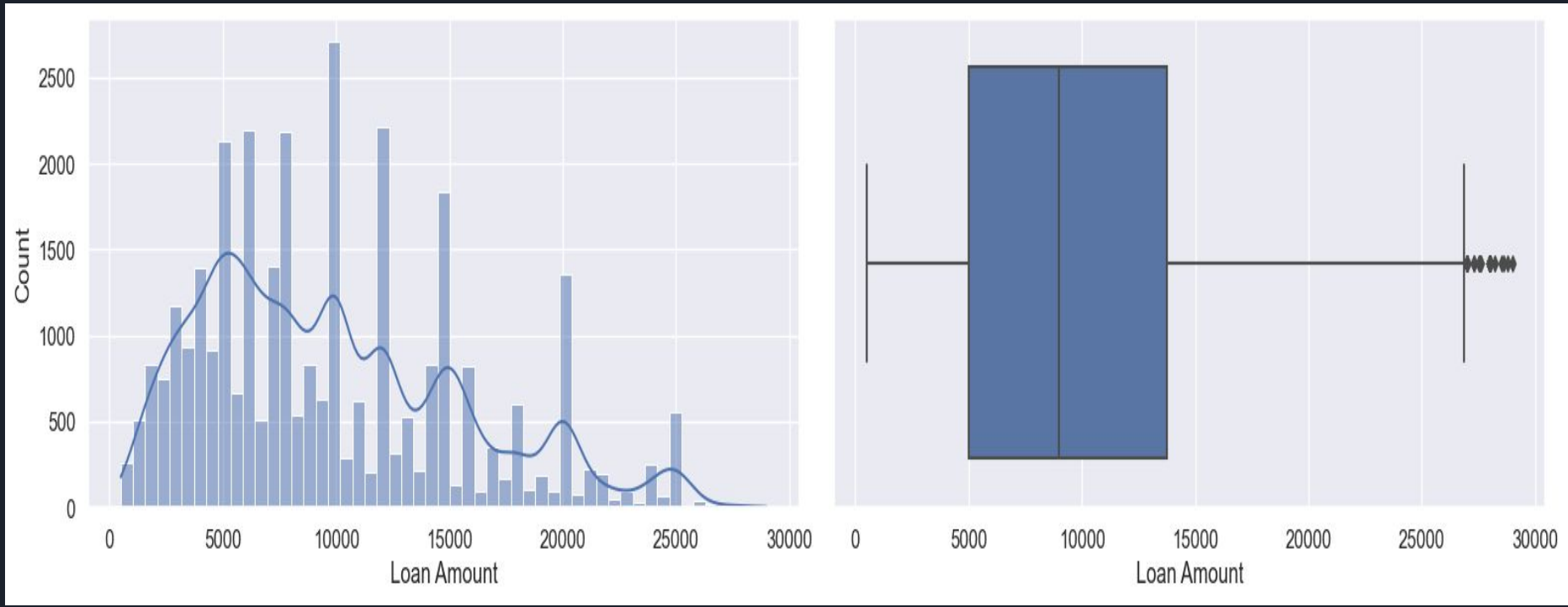


Univariate Analysis (Ordered Categorical)

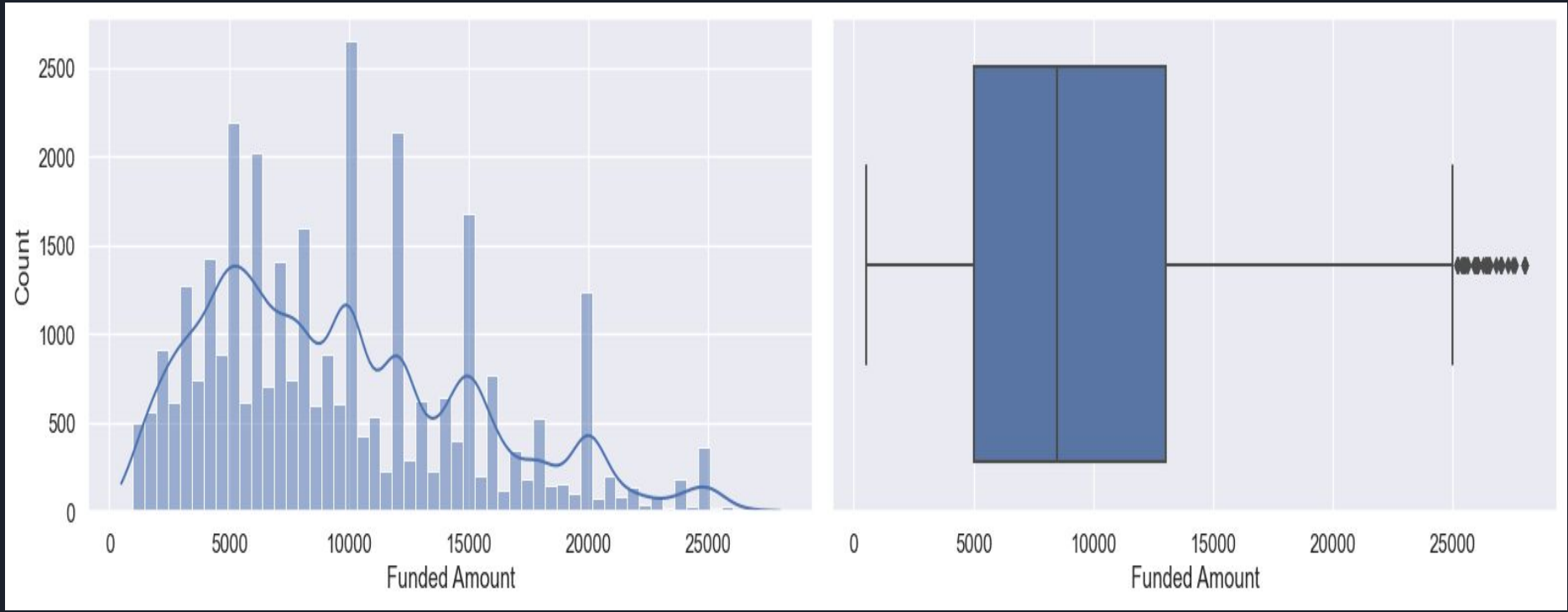
Grade



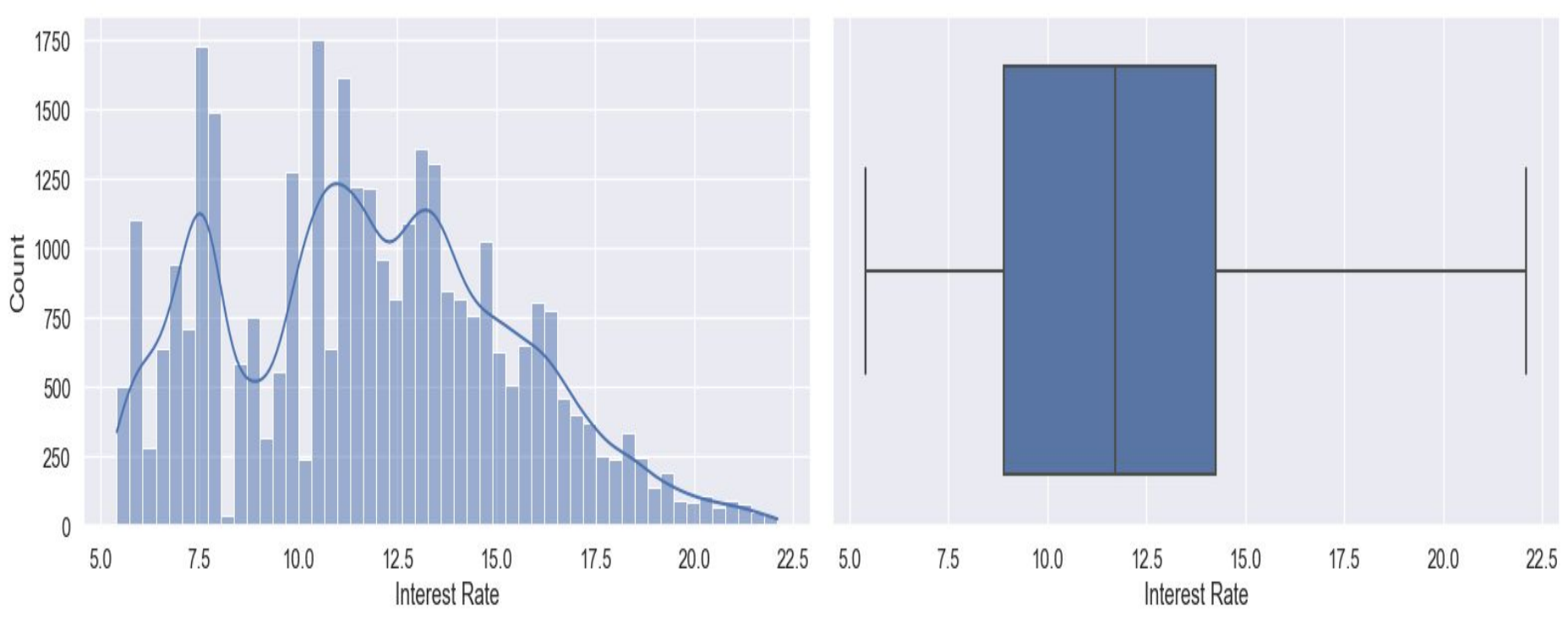
Univariate Analysis (Numerical)



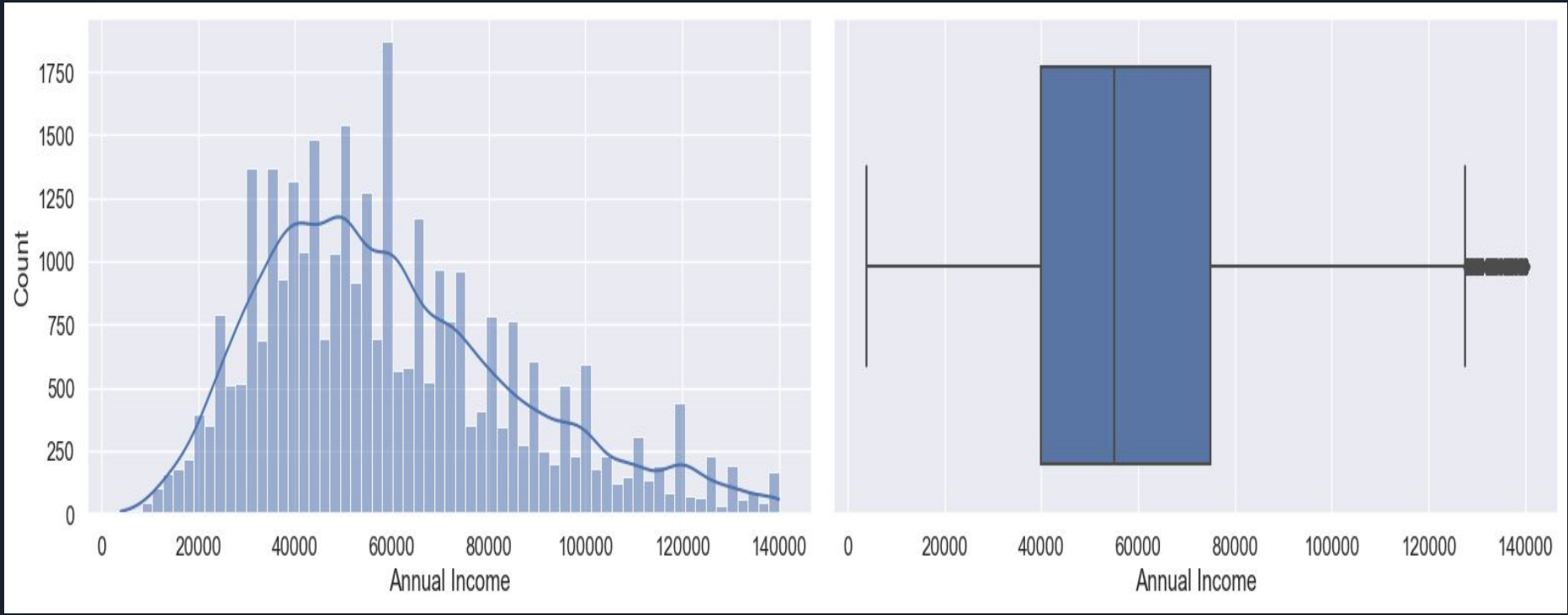
Univariate Analysis (Numerical)



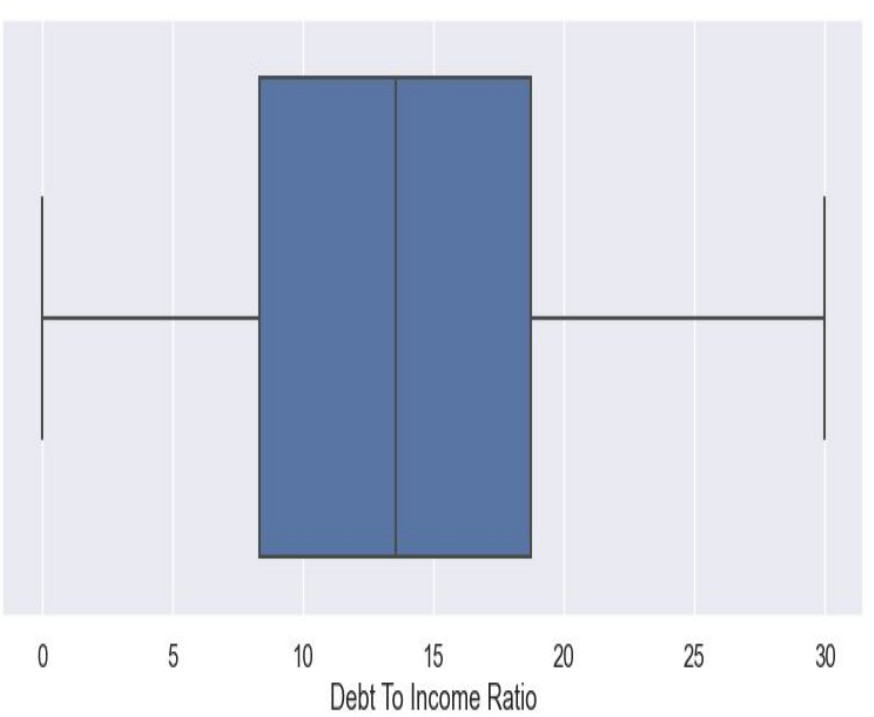
Univariate Analysis (Numerical)



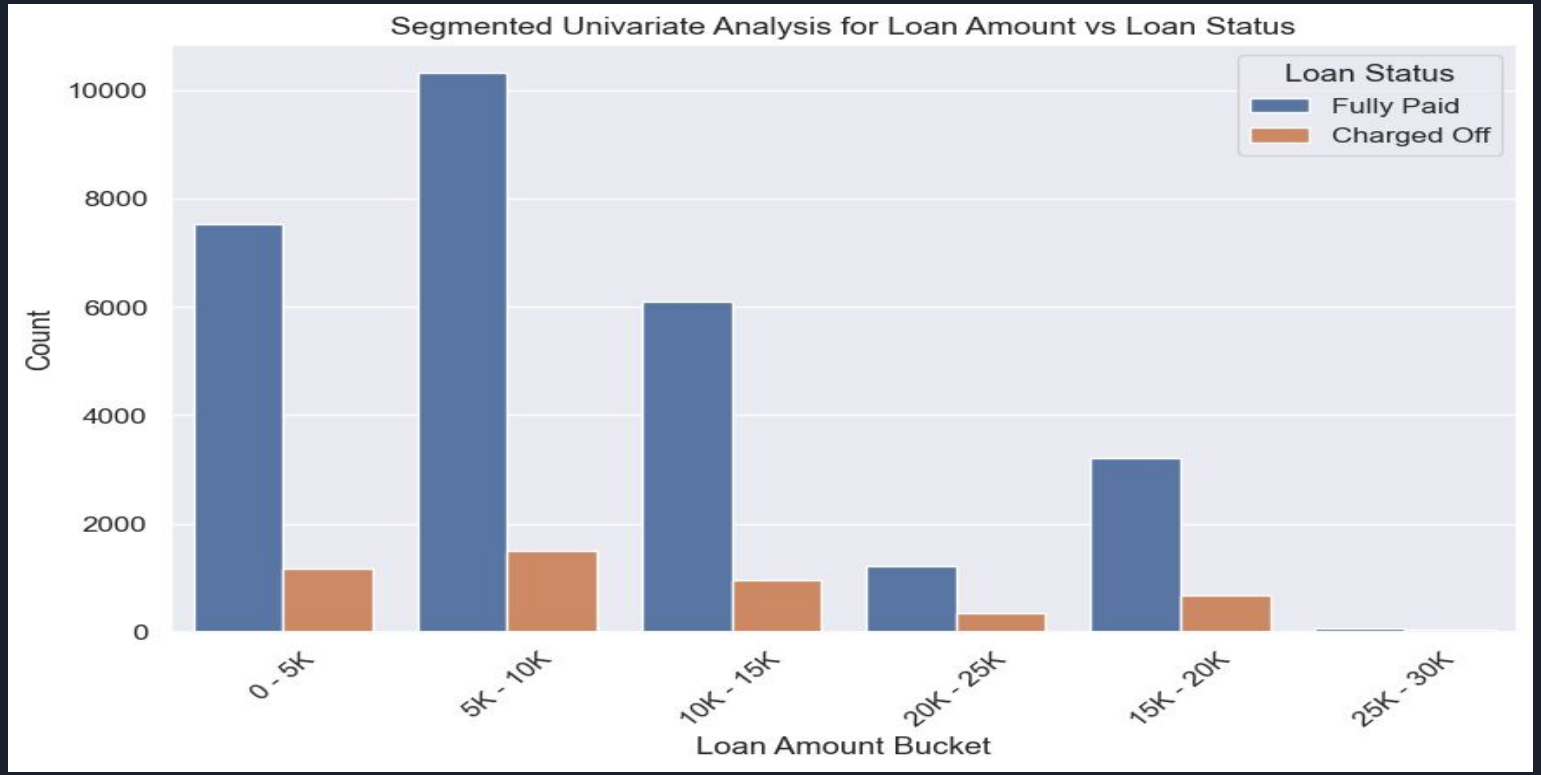
Univariate Analysis (Numerical)



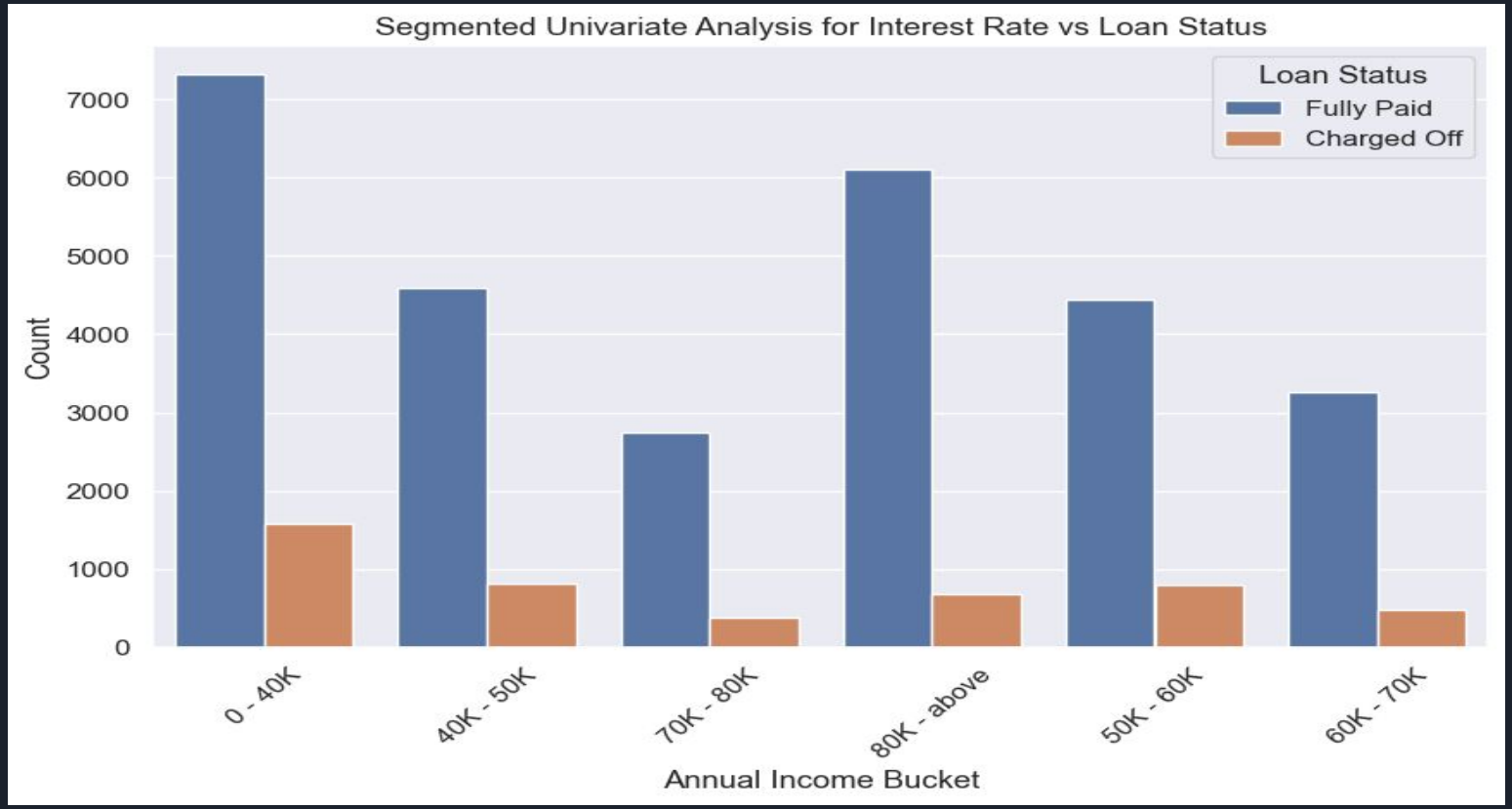
Univariate Analysis (Numerical)



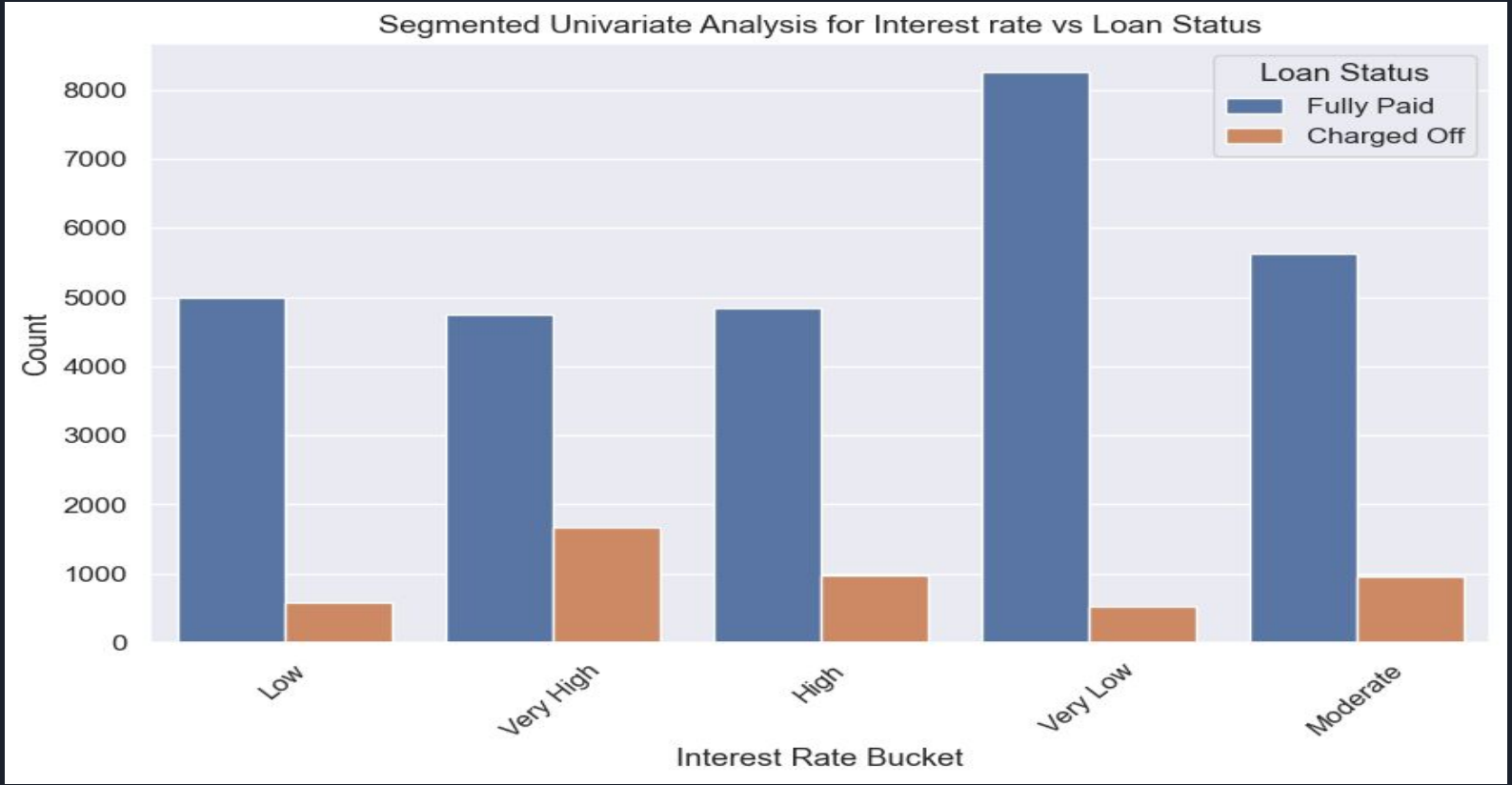
Segmented Univariate Analysis



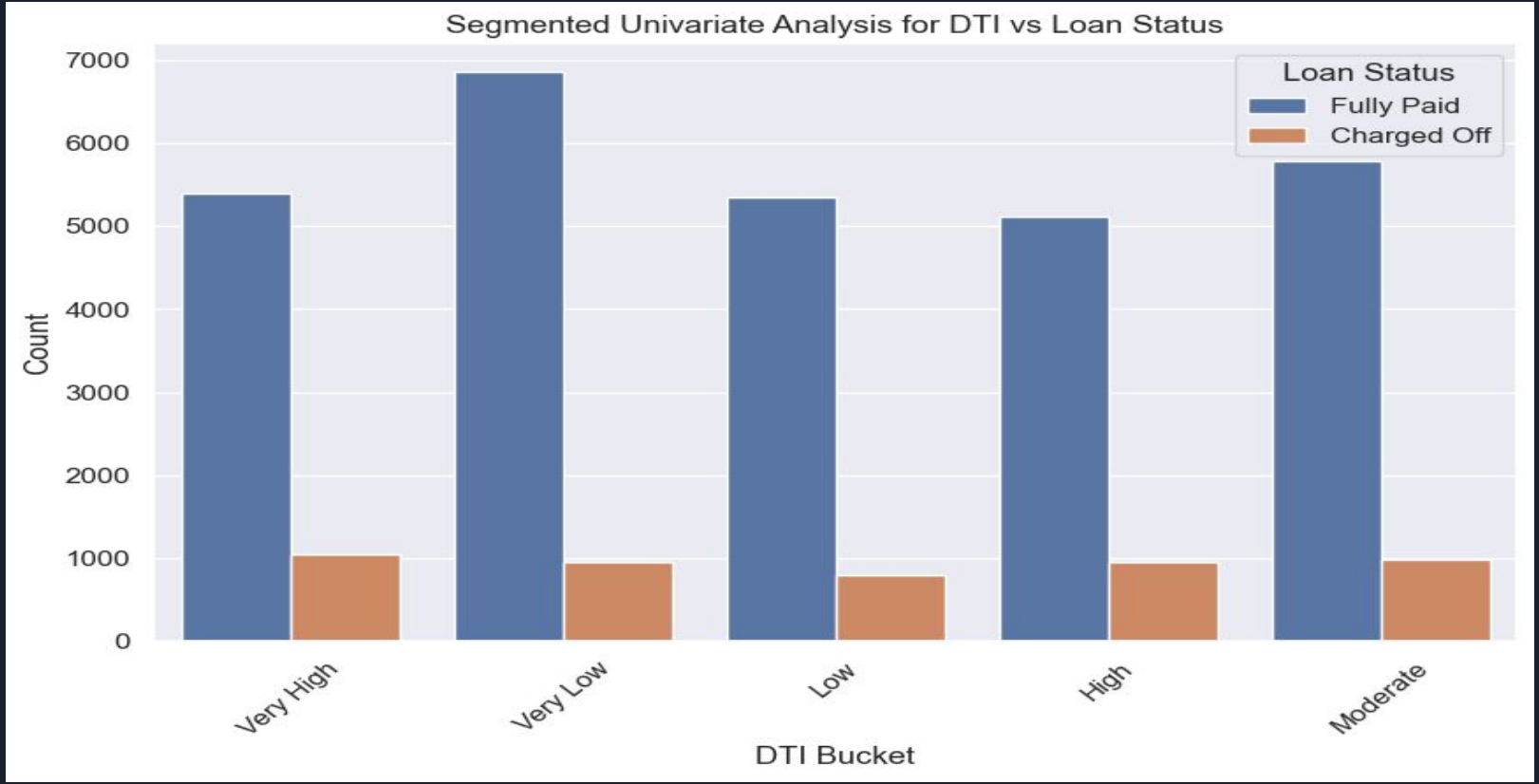
Segmented Univariate Analysis



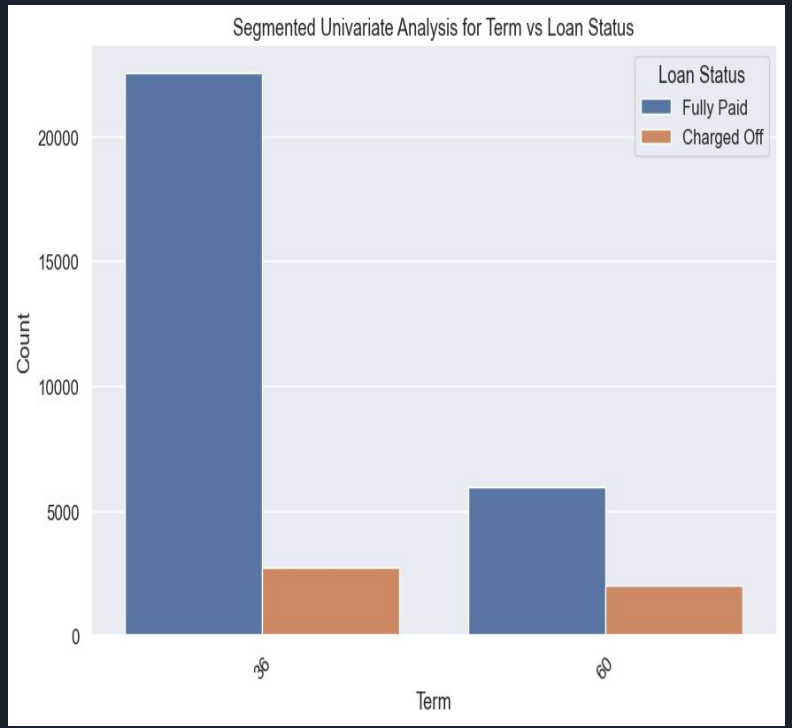
Segmented Univariate Analysis

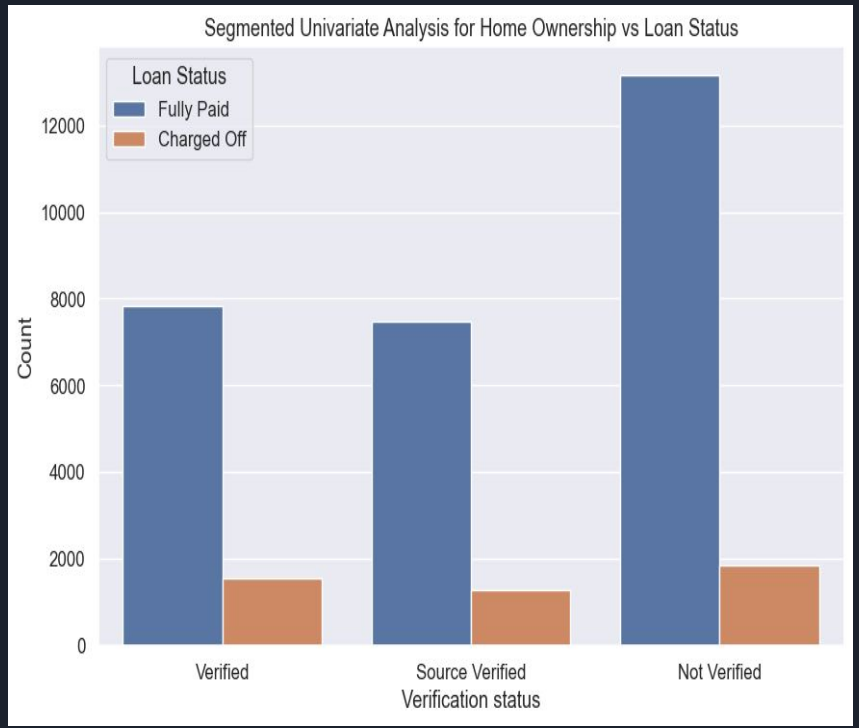


Segmented Univariate Analysis

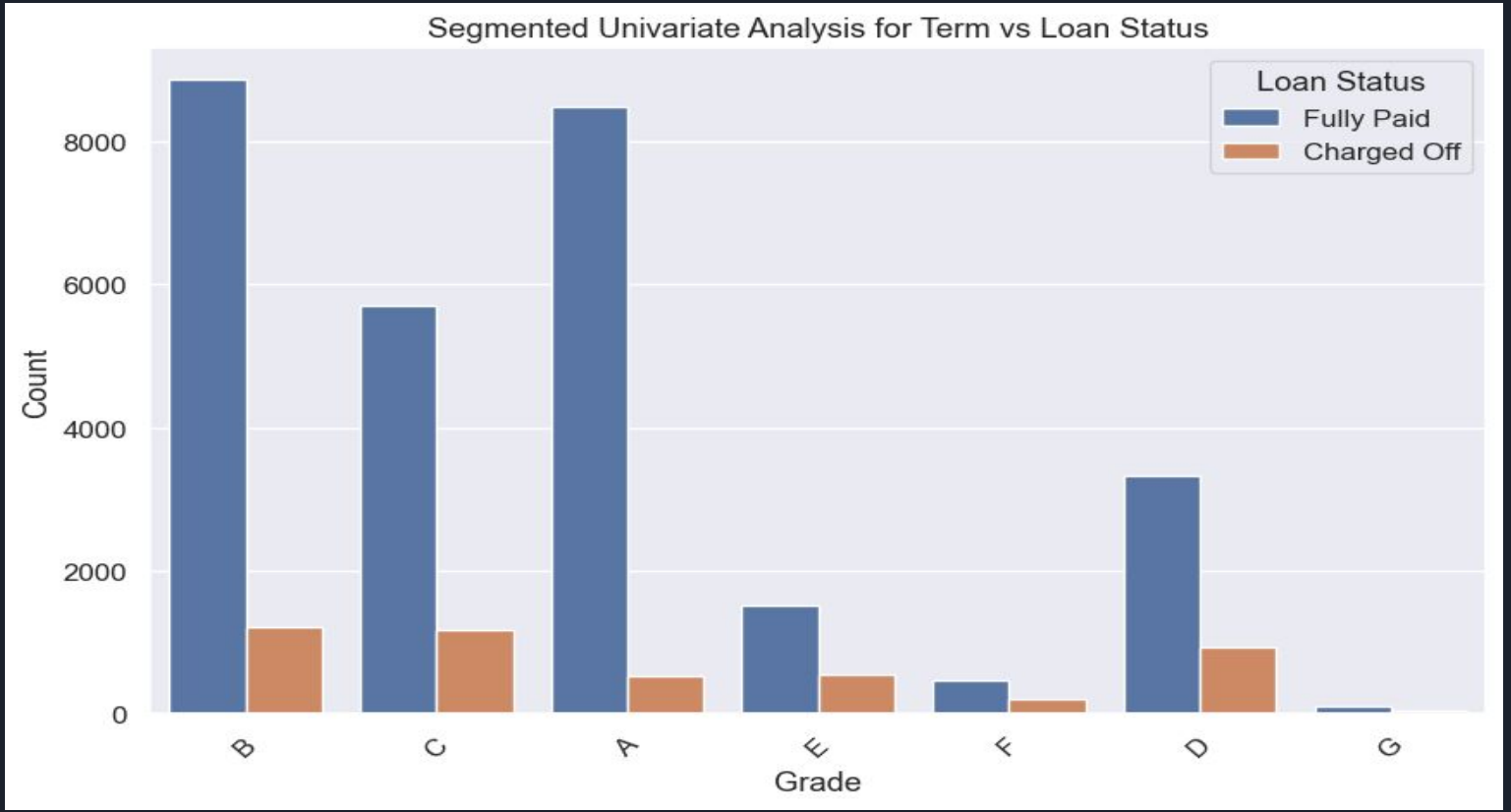


Segmented Univariate Analysis

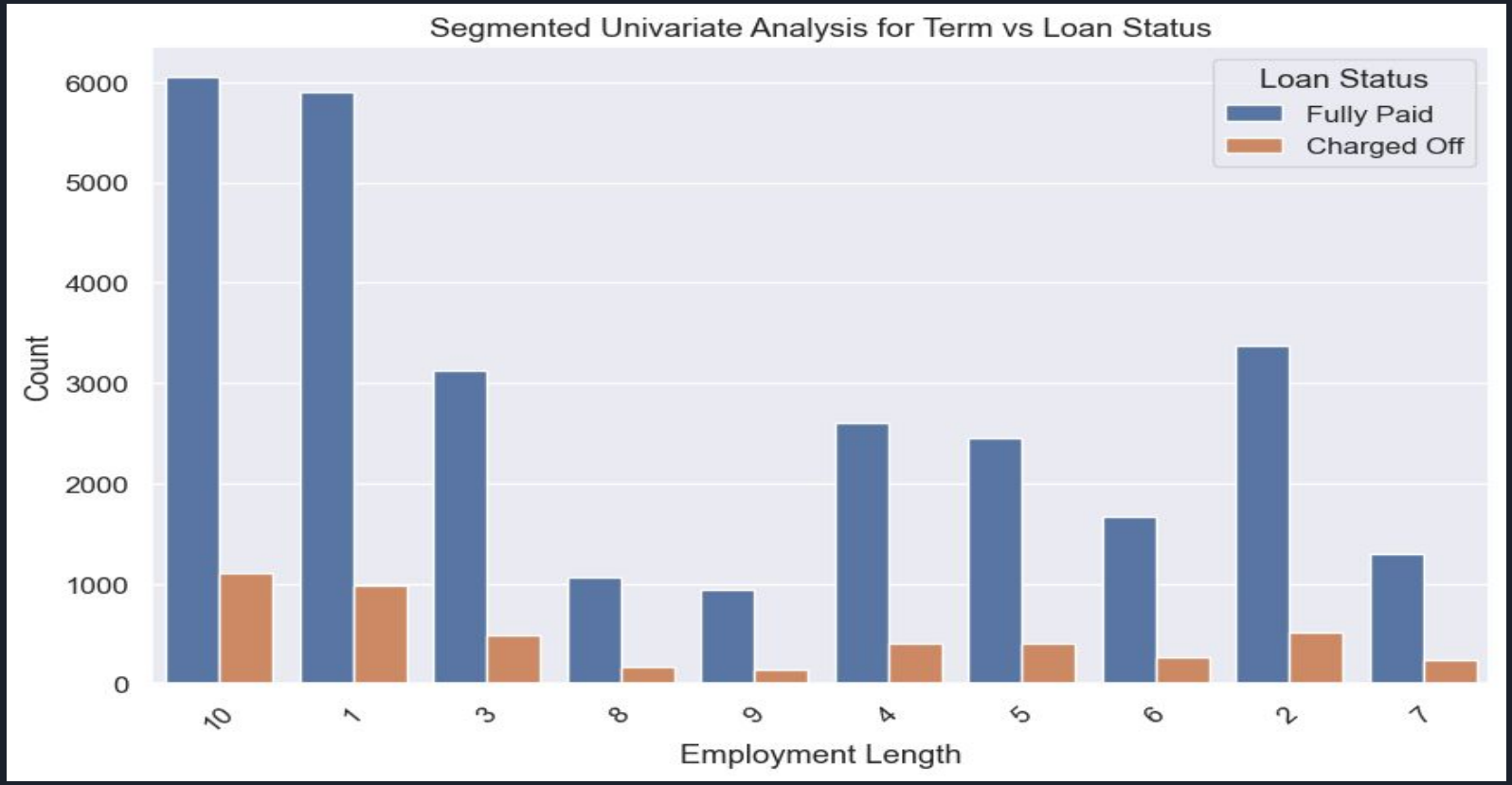




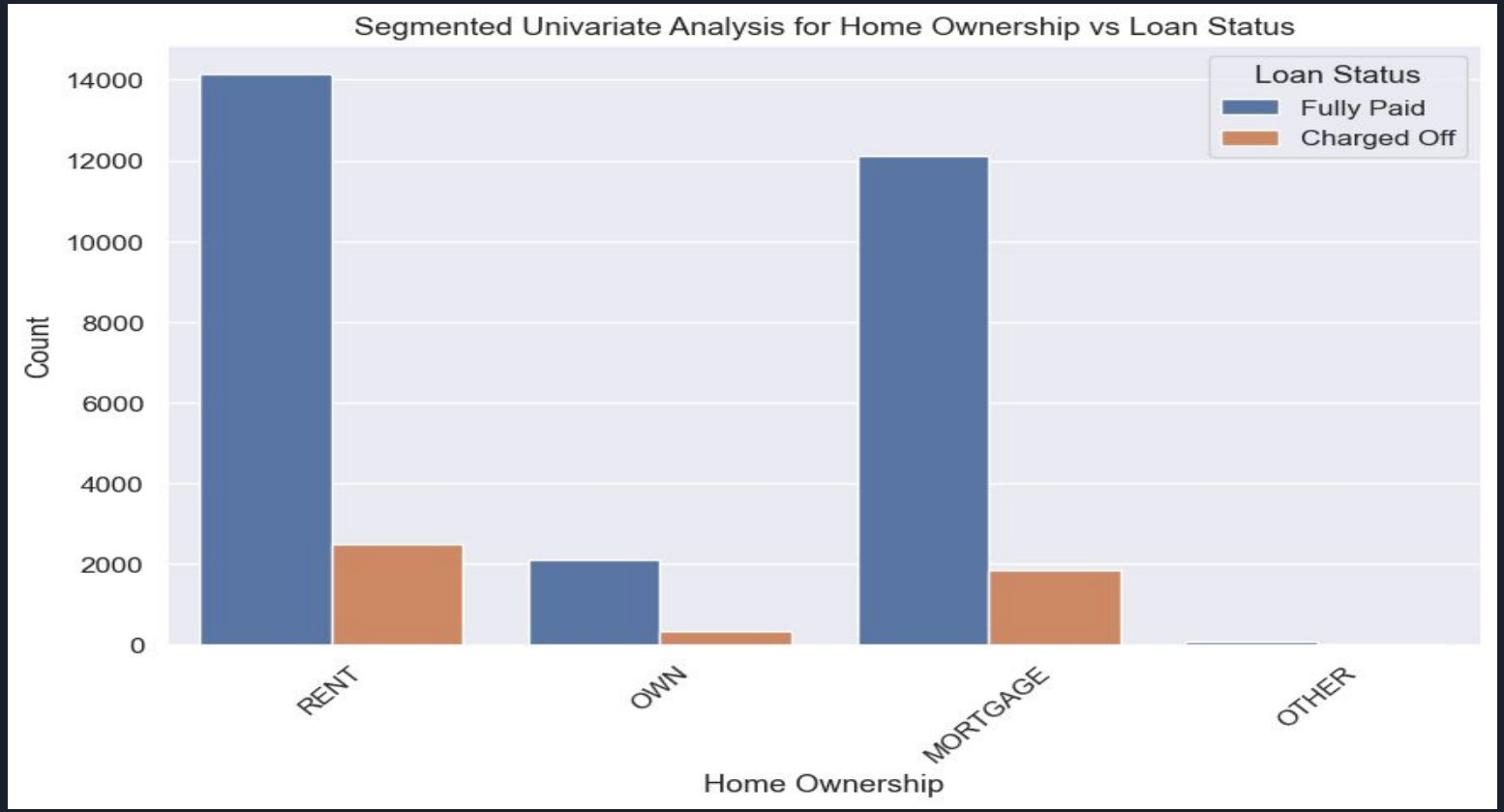
Segmented Univariate Analysis



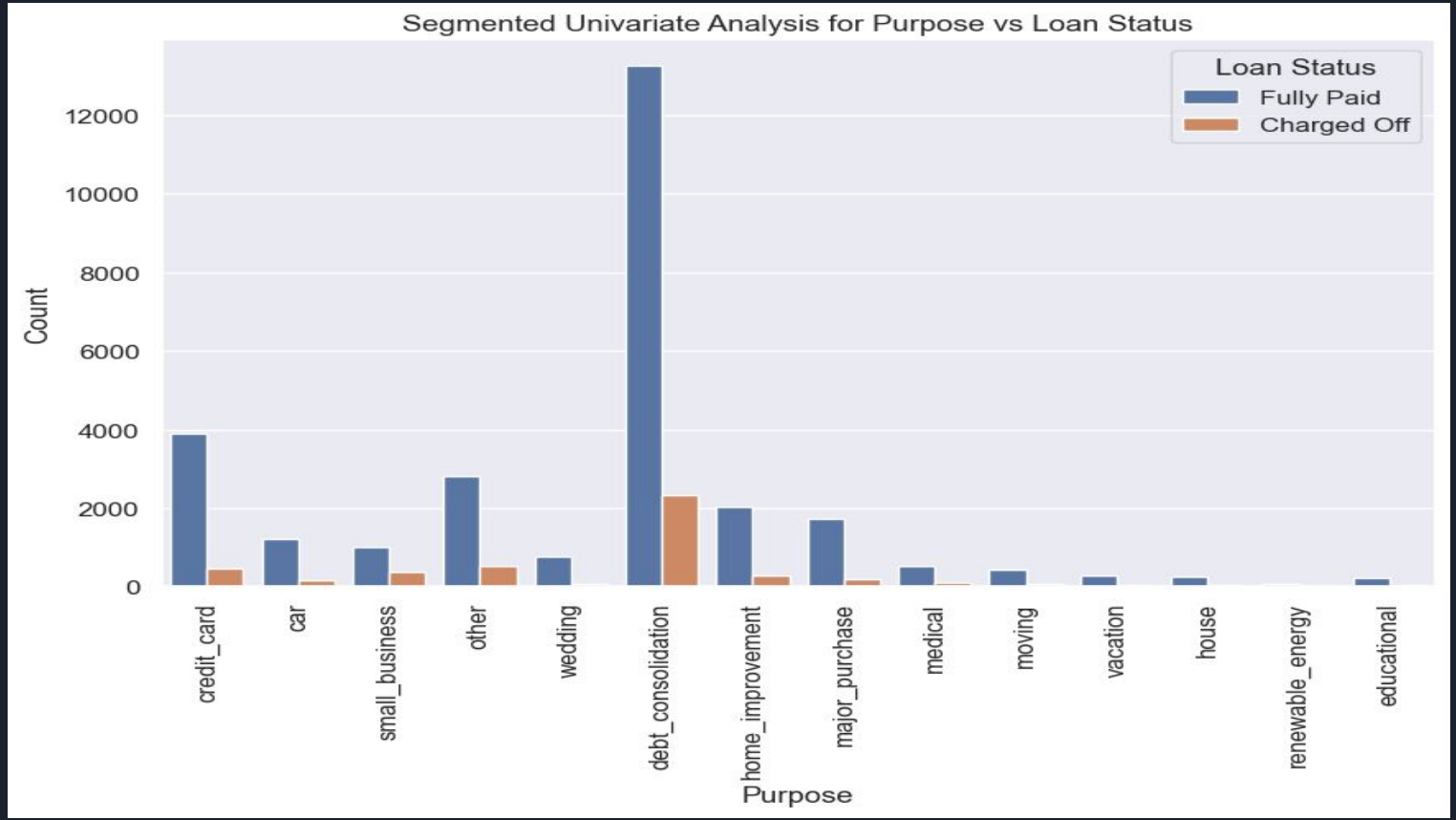
Segmented Univariate Analysis



Segmented Univariate Analysis



Segmented Univariate Analysis

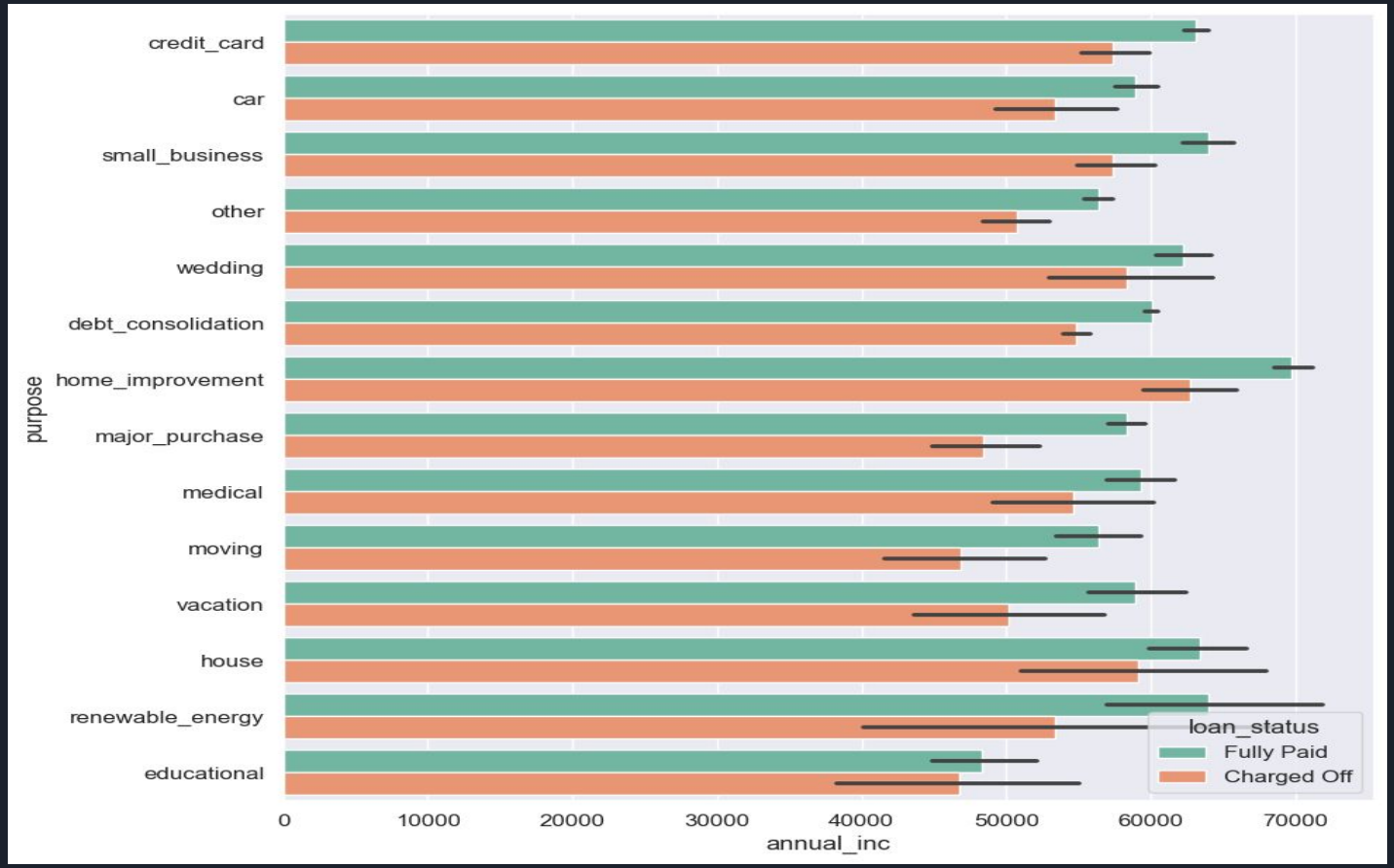


Univariate Analysis Observations

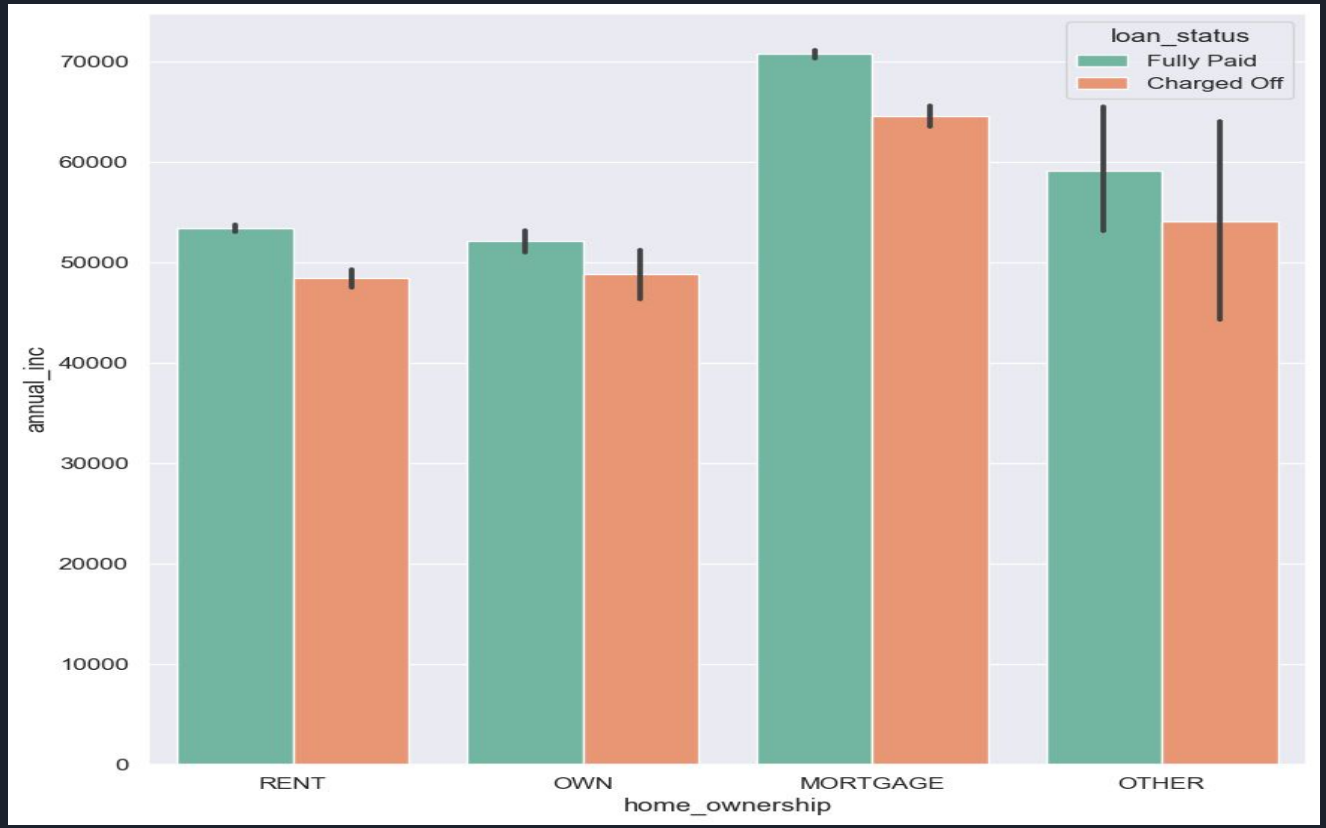
The above analysis with respect to the charged off loans for each variable suggests the following. There is a more probability of defaulting when:

- Applicants having house_ownership as 'RENT'
- When the purpose is 'debt_consolidation'
- When the loan status is Not verified
- Term of 36 months
- Applicants with employment length of 10
- Grade is 'B'
- Applicants who receive the loan amount ranging 5k - 10k
- Applicants who have an income of range 0 - 40k
- Applicants who receive interest at the rate above 13%
- Applicants who are from the state CA.

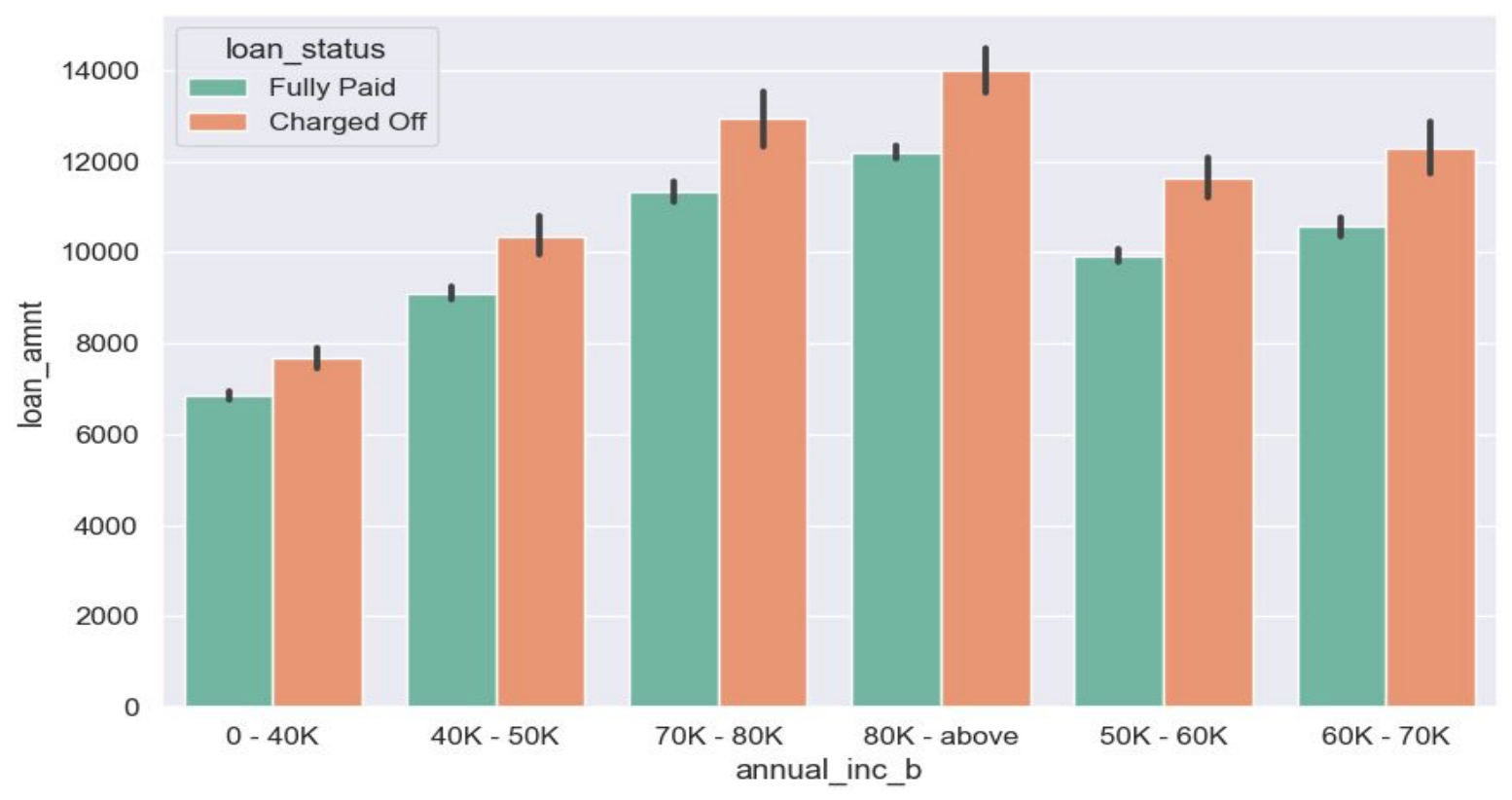
Bivariate Analysis



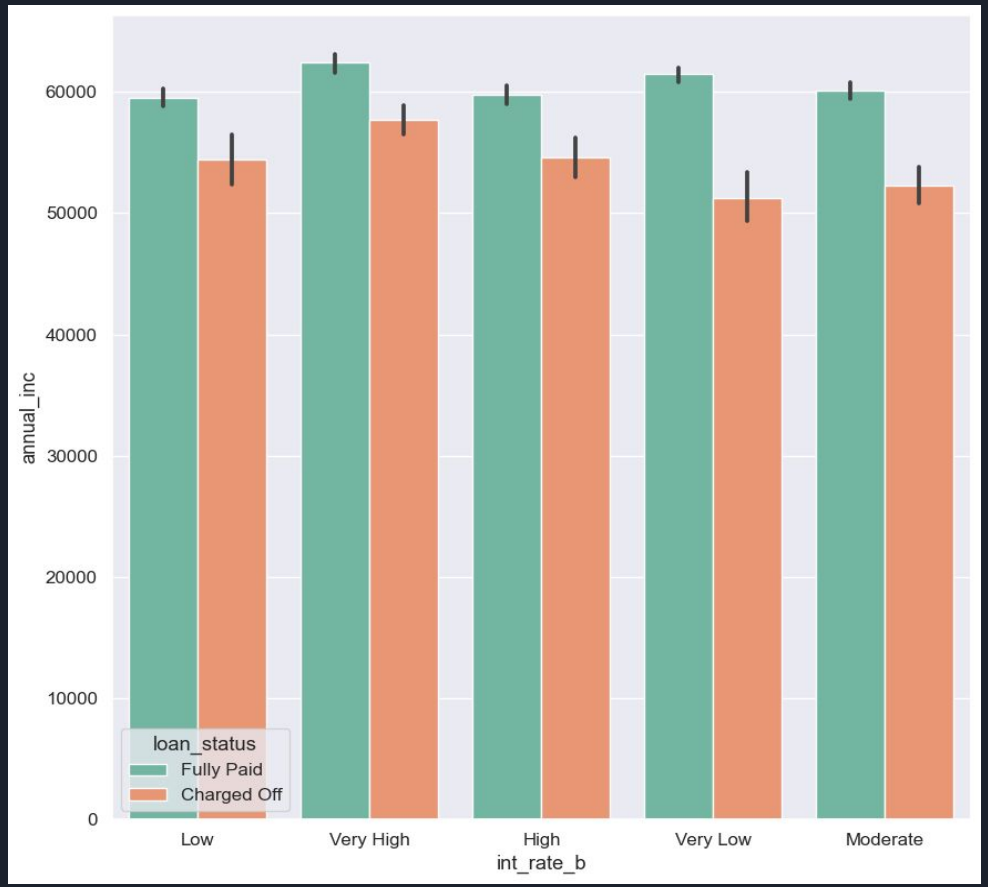
Bivariate Analysis



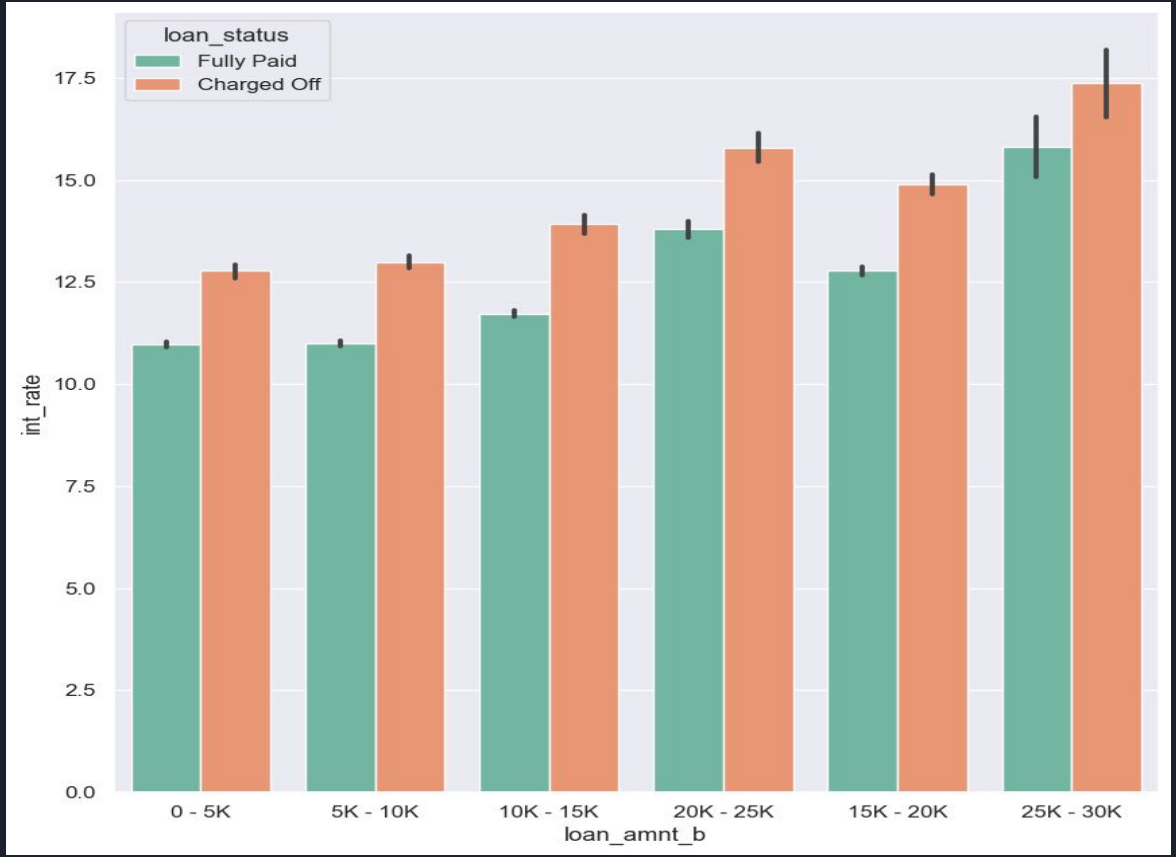
Bivariate Analysis



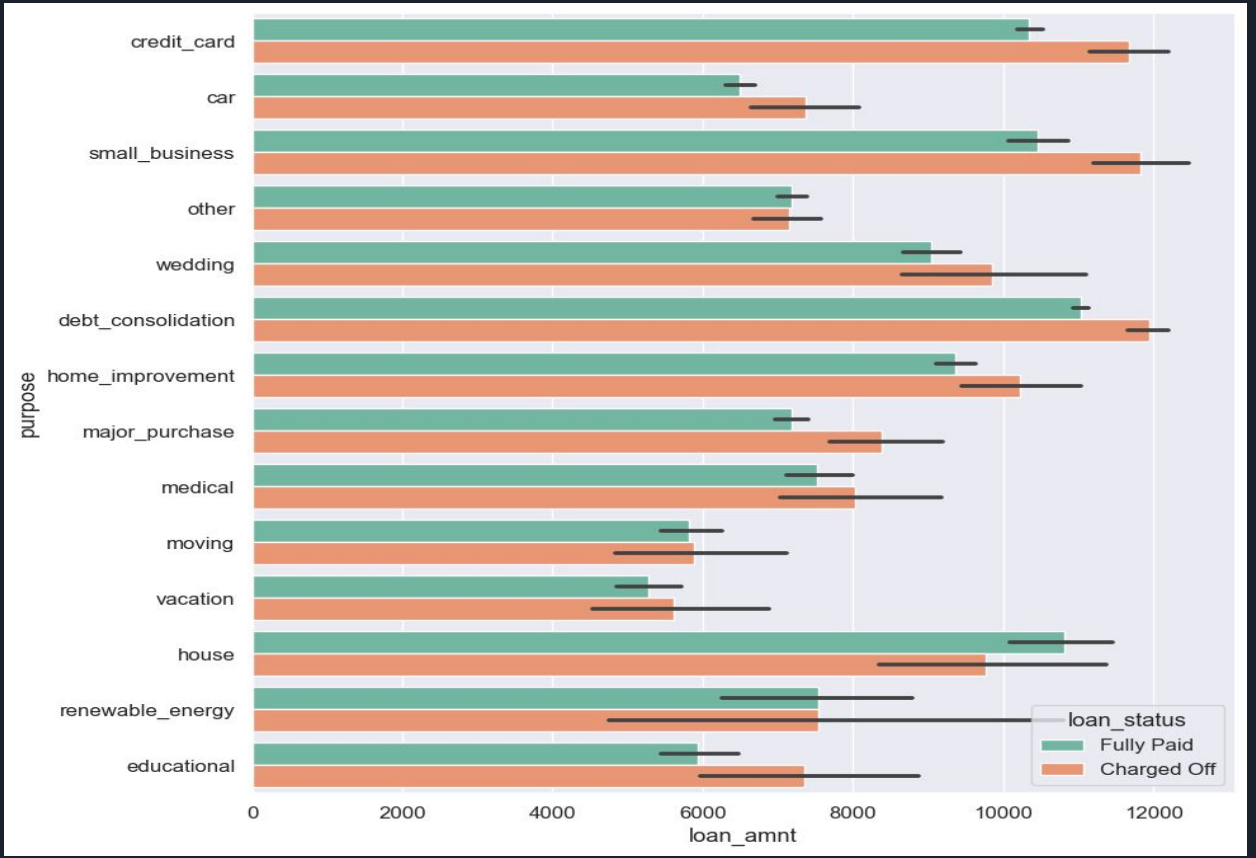
Bivariate Analysis



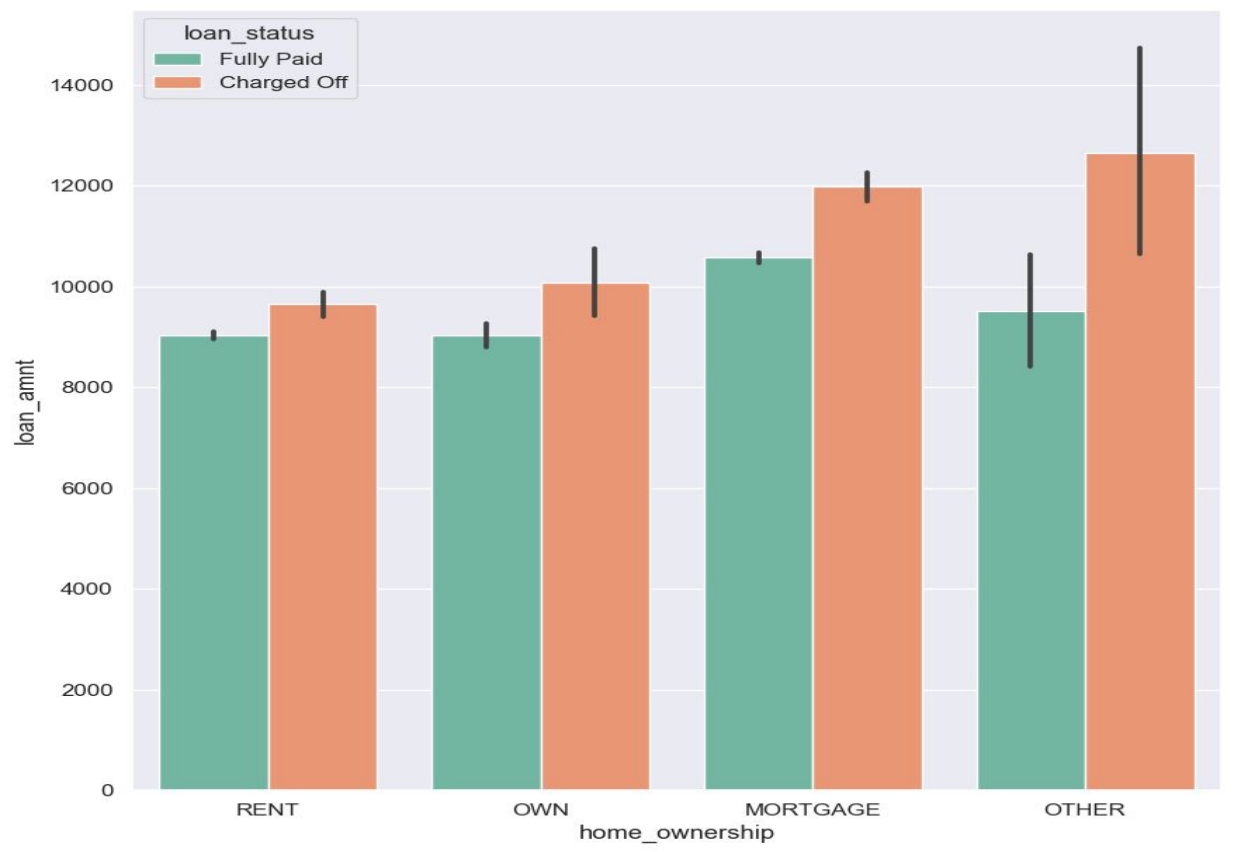
Bivariate Analysis



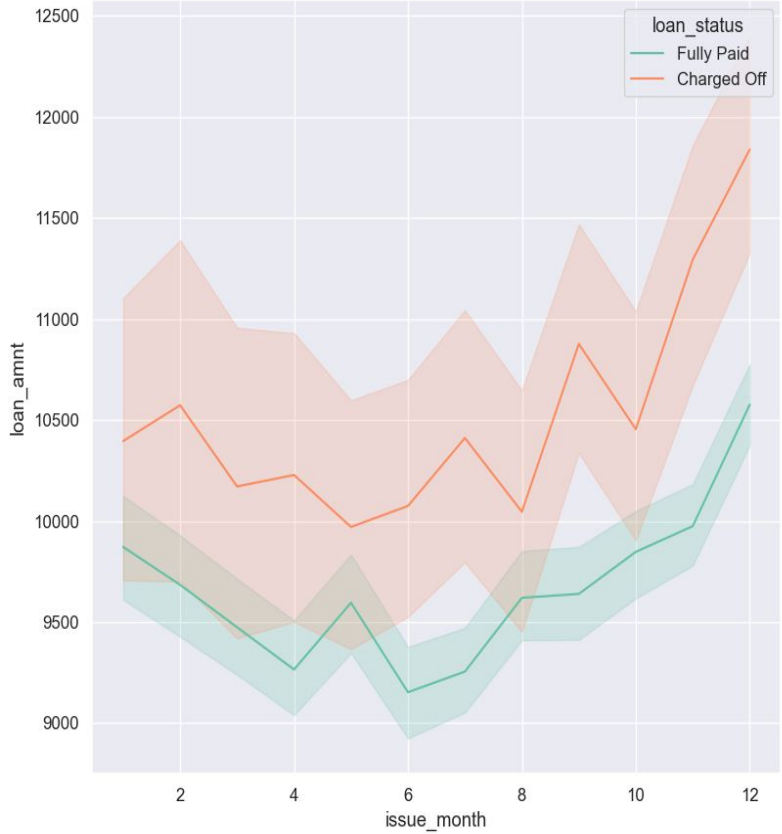
Bivariate Analysis

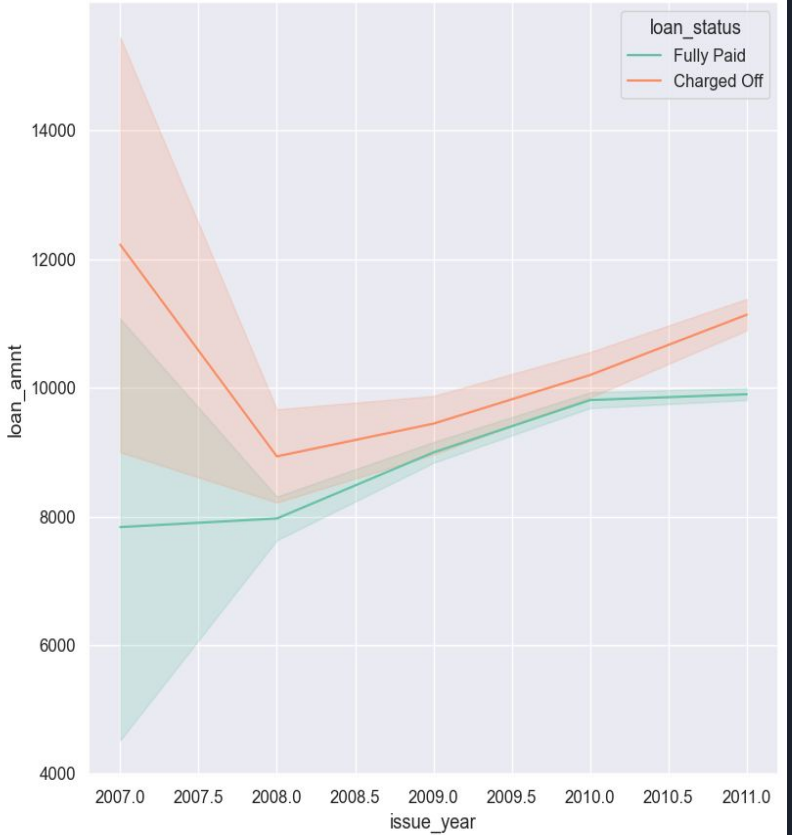


Bivariate Analysis

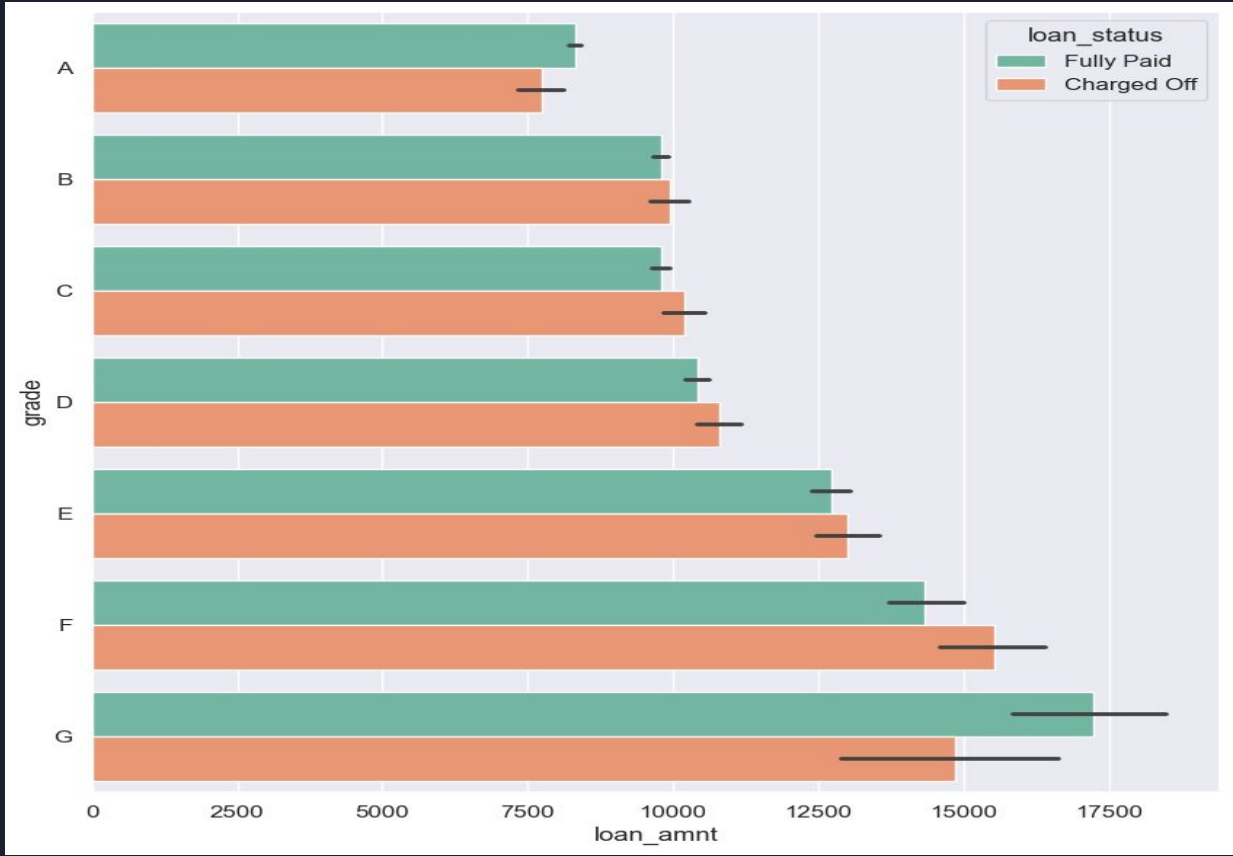


Bivariate Analysis

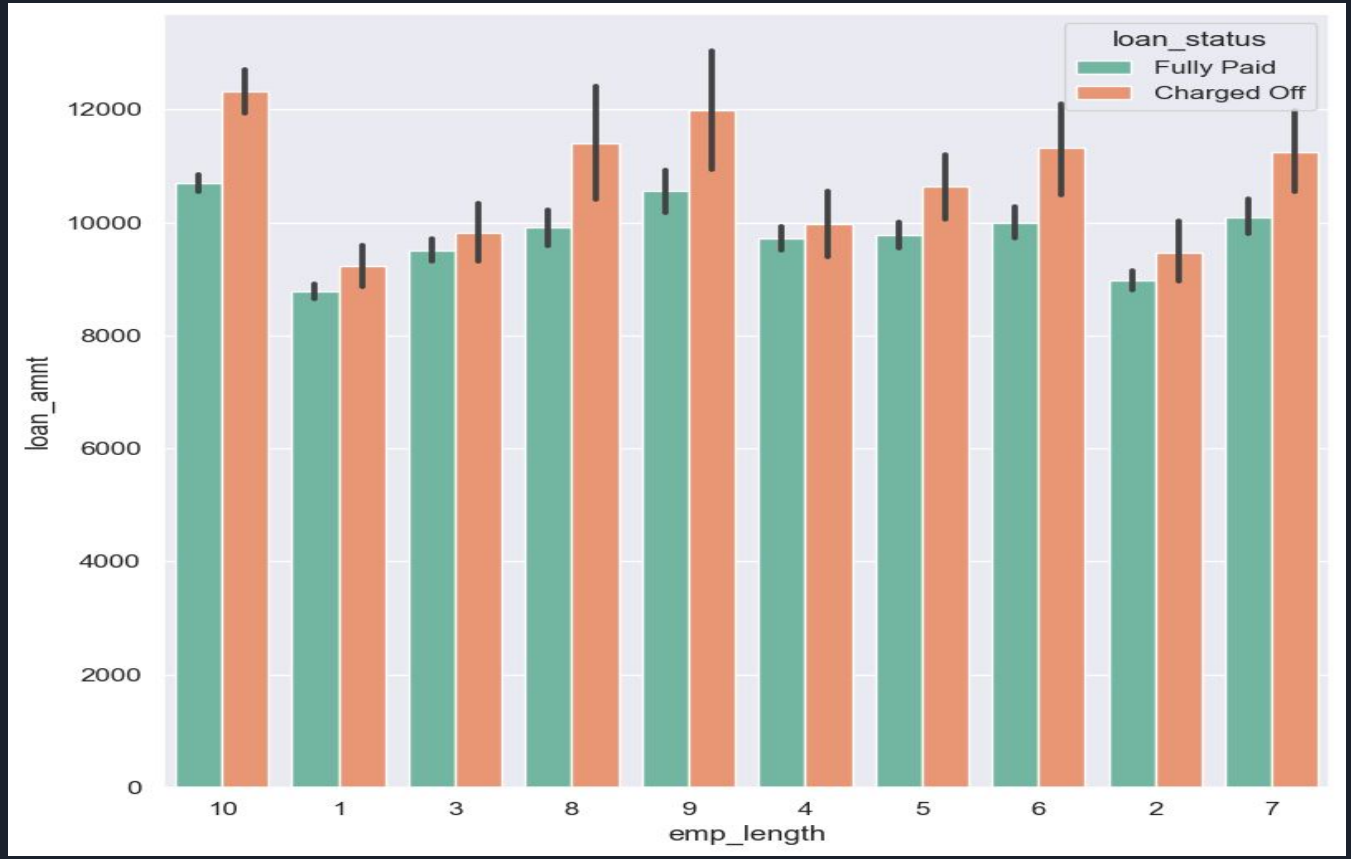




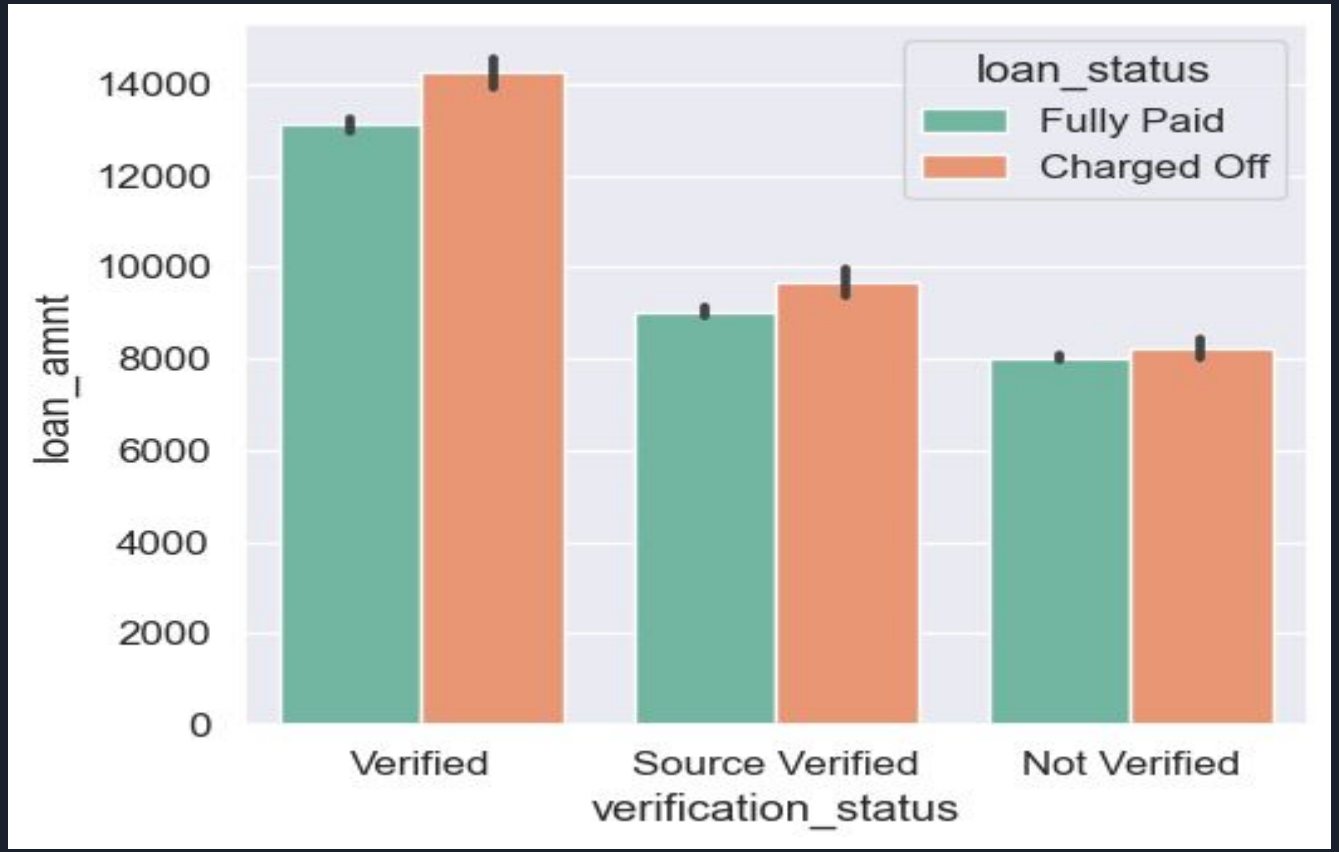
Bivariate Analysis



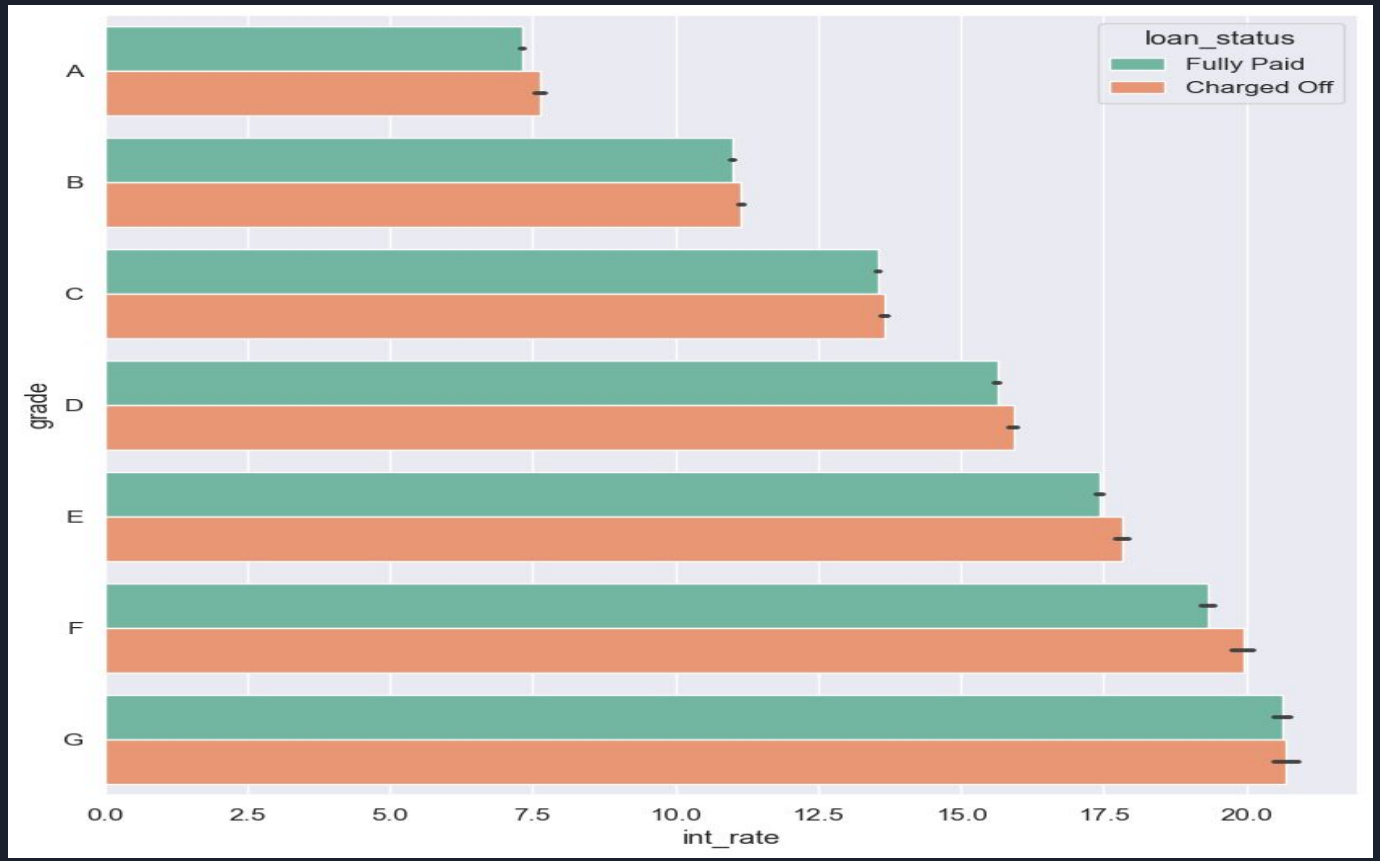
Bivariate Analysis



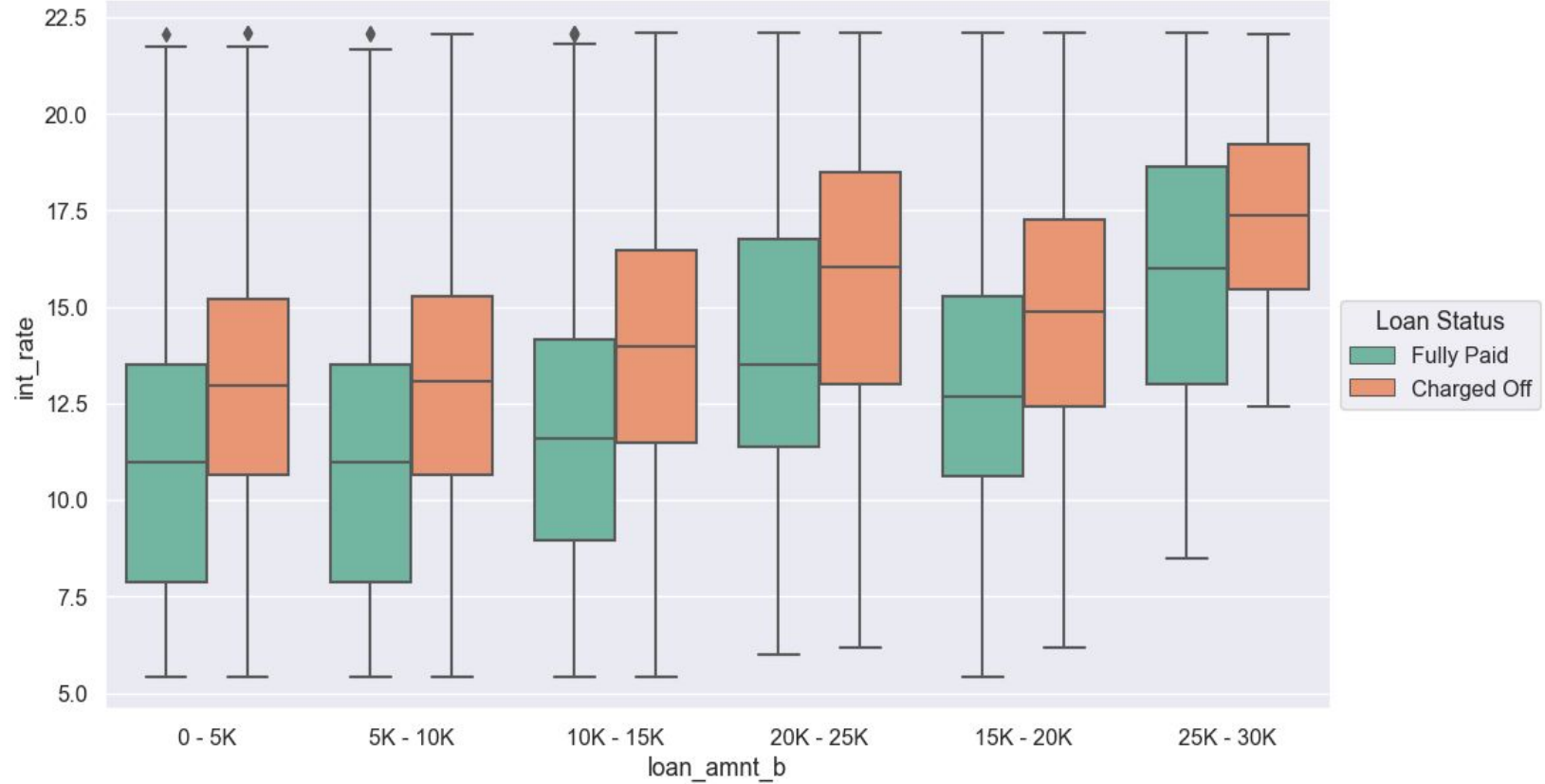
Bivariate Analysis



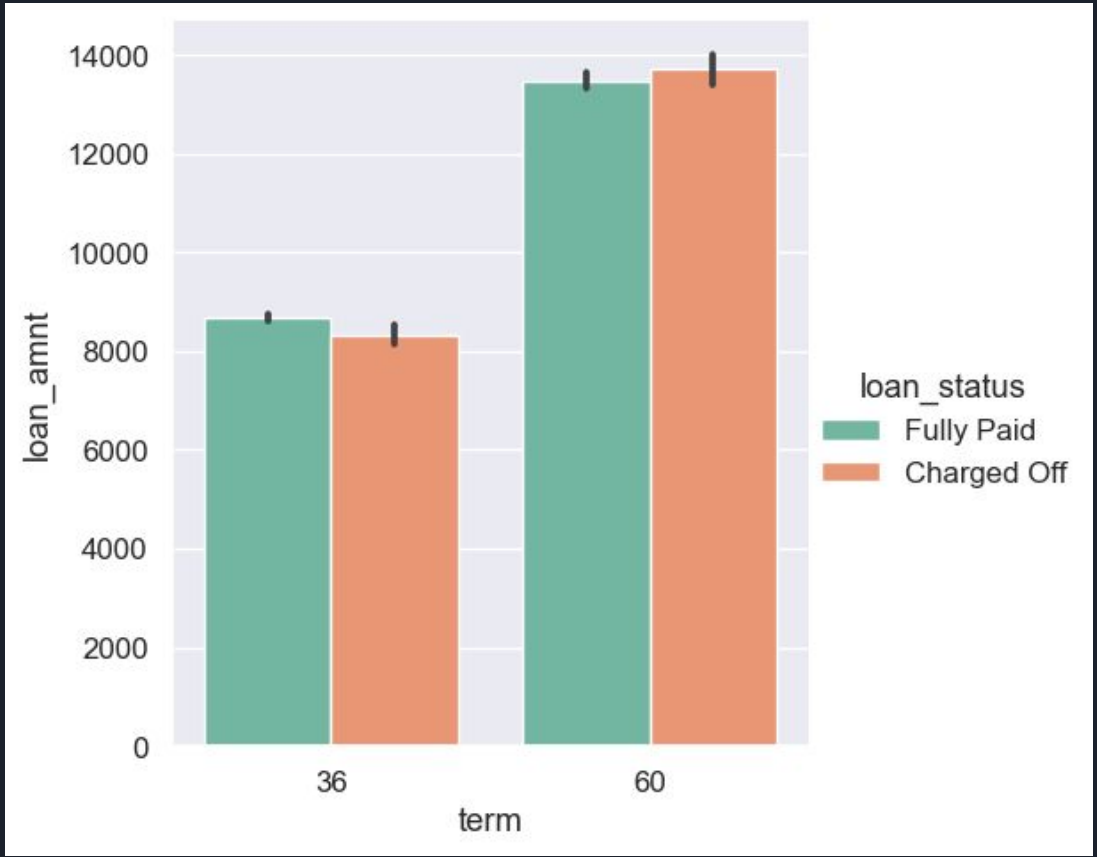
Bivariate Analysis



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 rate above 13% and have an income of 60k-70k
- Applicants who have taken a loan in the range 25k - 30k and are charged interest rate of 15-17.5%
- Applicants who have taken a loan for and dept consolidation small business and also the loan amount is greater than 10k
- Applicants whose home ownership is 'MORTGAGE and have loan of 10-12k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 10k-12k
- When the loan is verified and loan amount is above 14k
- For grade G and interest rate above 20%