Lending Club Case Study

Study Group Members:
Hema Lakshmi
Dushyant Bharadawaj

Introduction

- Lending Club is a leading online loan marketplace that provides personal and business loans, along with medical financing, through a fast and user-friendly interface offering lower interest rates to borrowers.

- The analysis aims to identify loan applicants who are likely to repay or default, as approving reliable borrowers maximizes profits, while approving high-risk applicants can lead to significant financial losses.

Problem solving methodology

- Understanding Problem Statement
- Data Understanding (using Data Dictionary Data)
- Data Cleaning & Pre-Processing: Identifying and Handling Missing Values and Outliers, addressing data Imbalance and performing sanity Checks
- Univariate Analysis: Target variable with one independent variable.
- Bivariate Analysis: Target variable with two independent variables.
- Multivariate Analysis: Target variable with more than two independent variables

Data Understanding

Based on the dataset attributes and decision matrix you've provided, here's a concise summary of how to approach the analysis and the data interpretation:

Dataset Attributes:

Loan Status (Primary Attribute of Interest):
Fully-Paid: Loans have been completely repaid.
Charged-Off: Loans have defaulted or are not repaid.
Current: Loans are still active and in progress.
Note: For analysis, exclude rows with "Current" status.

Decision Matrix

Fully Paid: Applicants who have successfully repaid both the principal and interest.
Charged-Off: Applicants who have defaulted on their loans due to non-payment over an extended period.
Current: Applicants who are still making payments, hence their loan status is ongoing and cannot be determined as defaulted or fully repaid yet.
Loan Rejection: Not included in this dataset; these are applicants who were declined for a loan and have no transactional history available.

Analysis Approach

Data Filtering: Remove rows with the "Current" loan status to focus only on loans that have been either fully repaid or charged off.
Outcome Evaluation: Focus on evaluating the factors influencing loans that are Fully Paid versus those that are Charged-Off.
Modeling and Insights: Use the filtered data to build predictive models or perform exploratory data analysis to understand the characteristics and patterns associated with fully paid versus charged-off loans.
Excluded Data: Loan Rejection data is not available in this dataset and thus cannot be used for any analysis. Ensure to exclude these cases from any assumptions or conclusions drawn.
If you need further details on how to analyze the data or specific methods for modeling and interpretation, let me know!

Data Cleaning & Pre-processing

1.Loading data from loan CSV

2.Checking for null values in the dataset

3.Checking for unique values

4.Checking for duplicated rows in data

5.Dropping Records & Columns

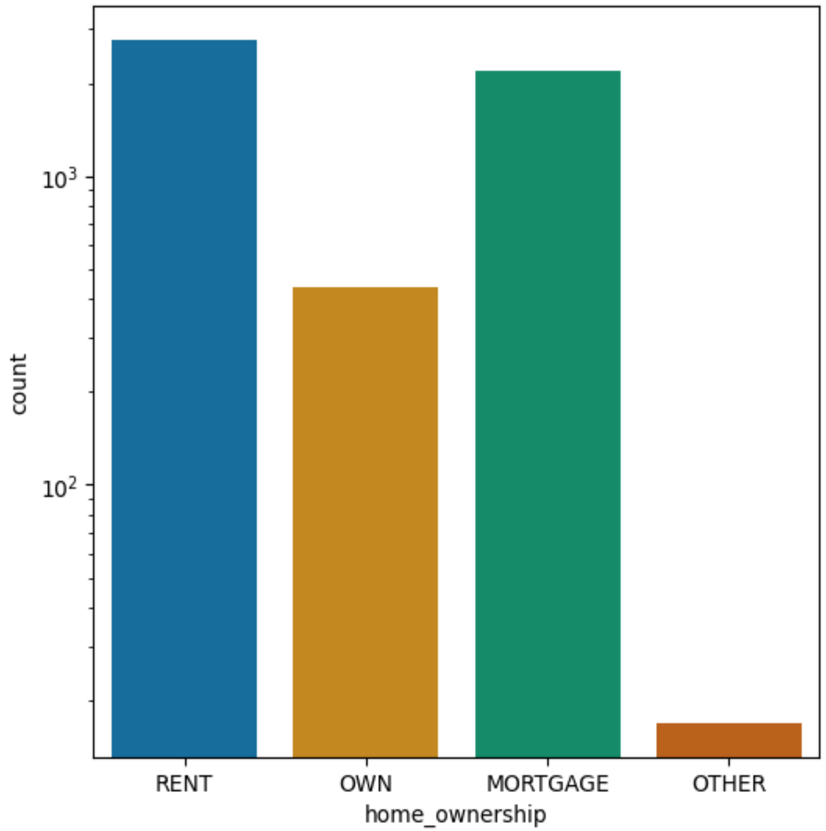
6.Common Functions

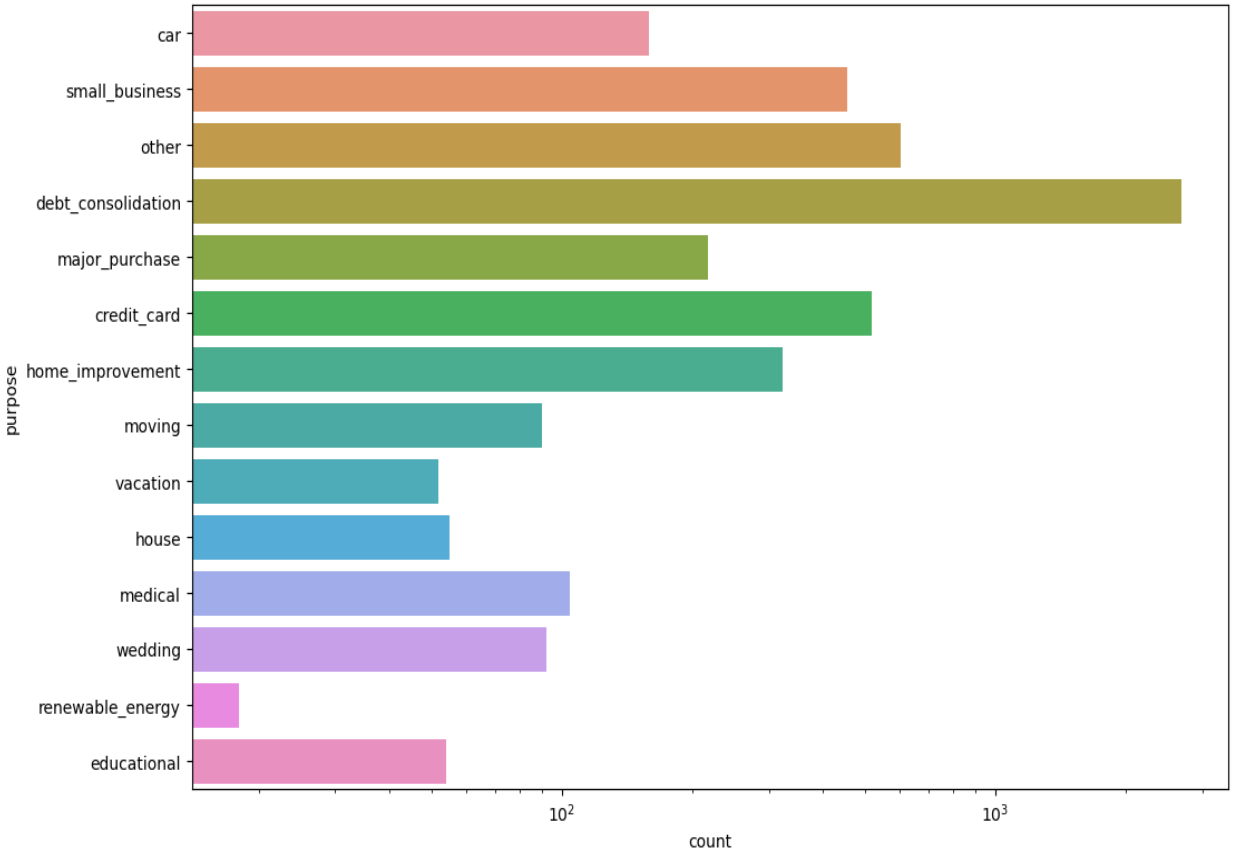
7.Data Conversion

8.Outlier Treatment

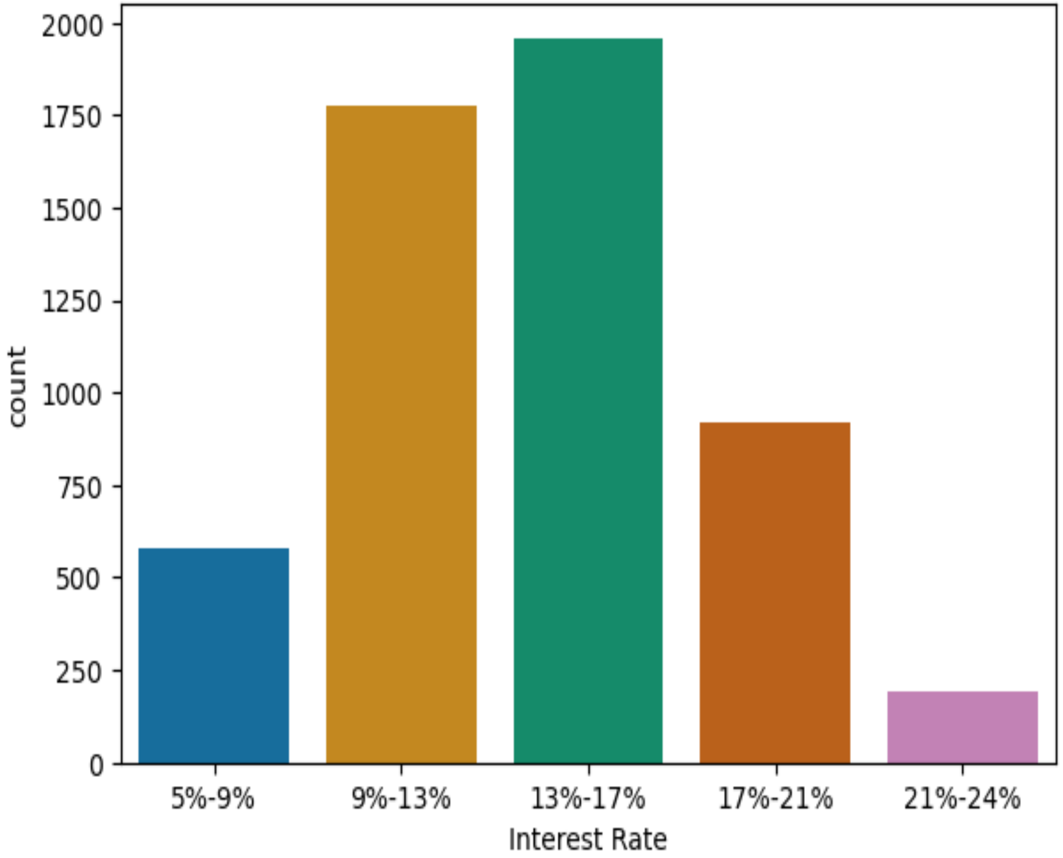
9.Imputing values in Columns

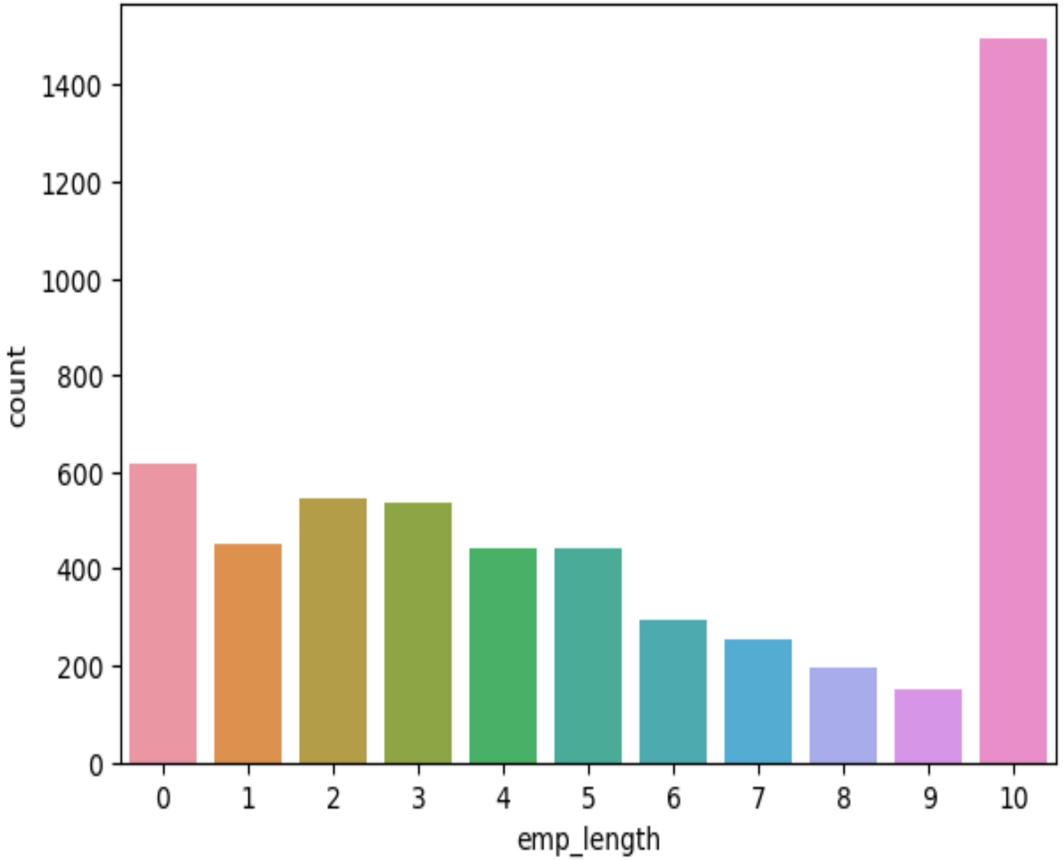
Univariate Analysis (Quantitative Variables)



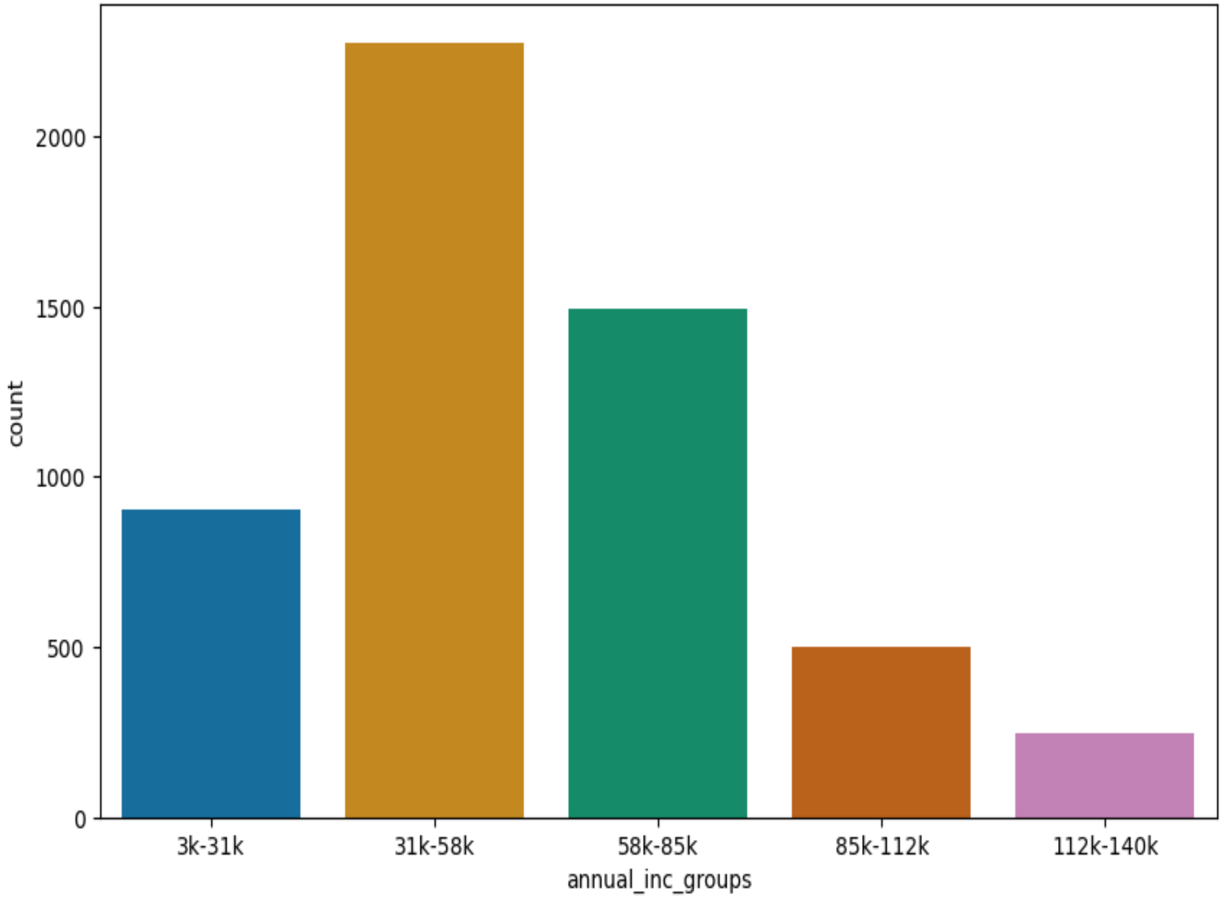


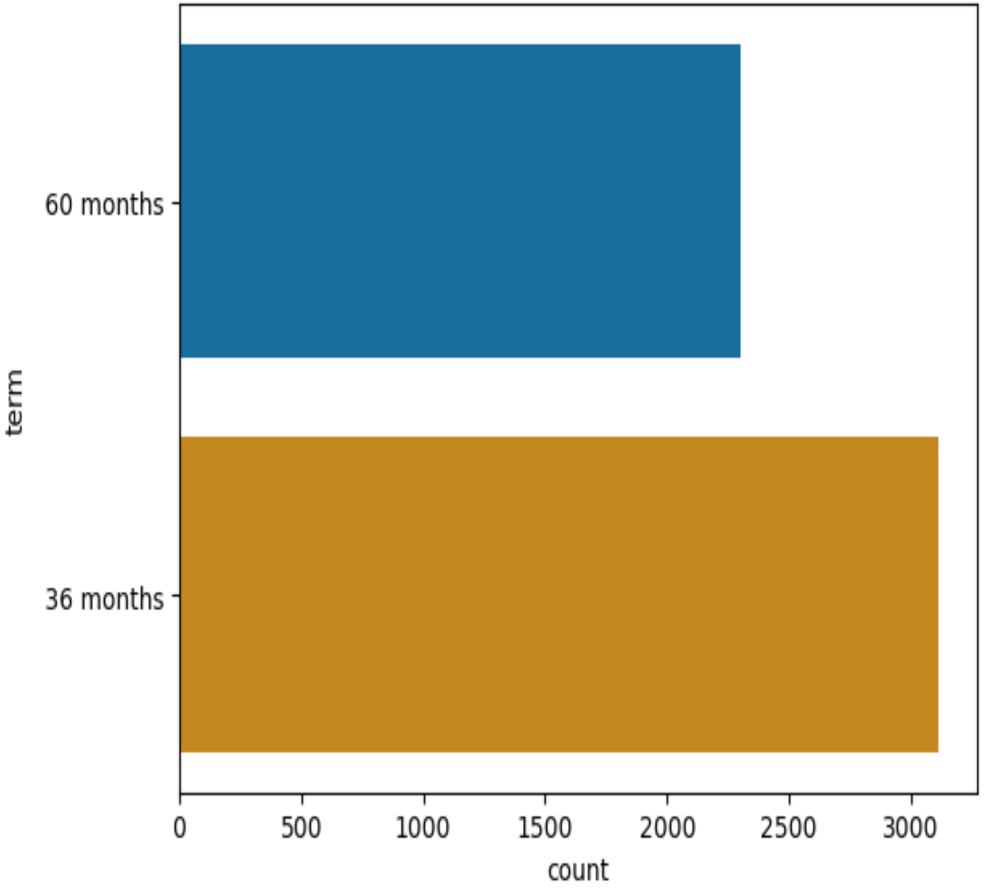
Univariate Analysis (Quantitative Variables)



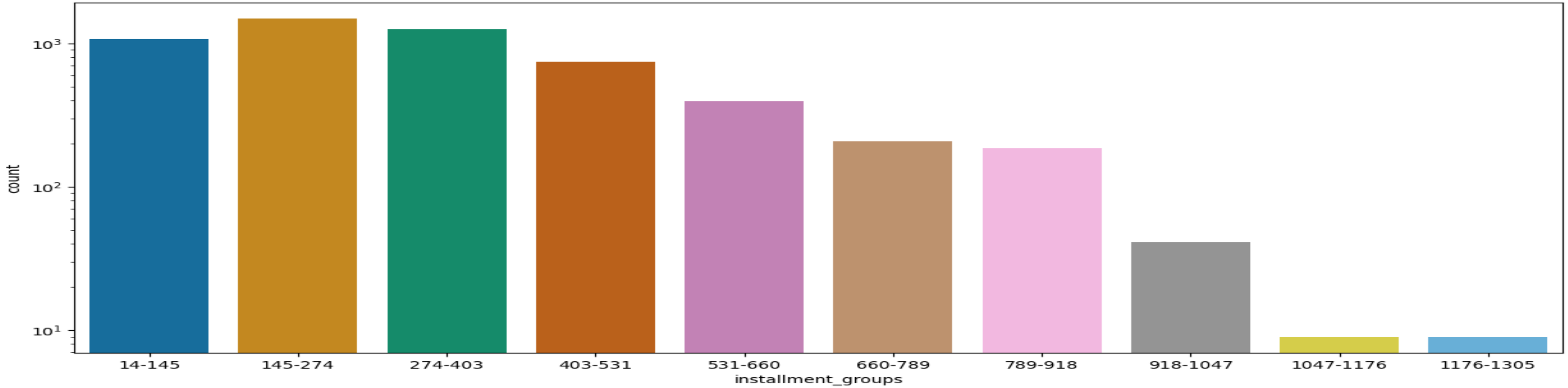


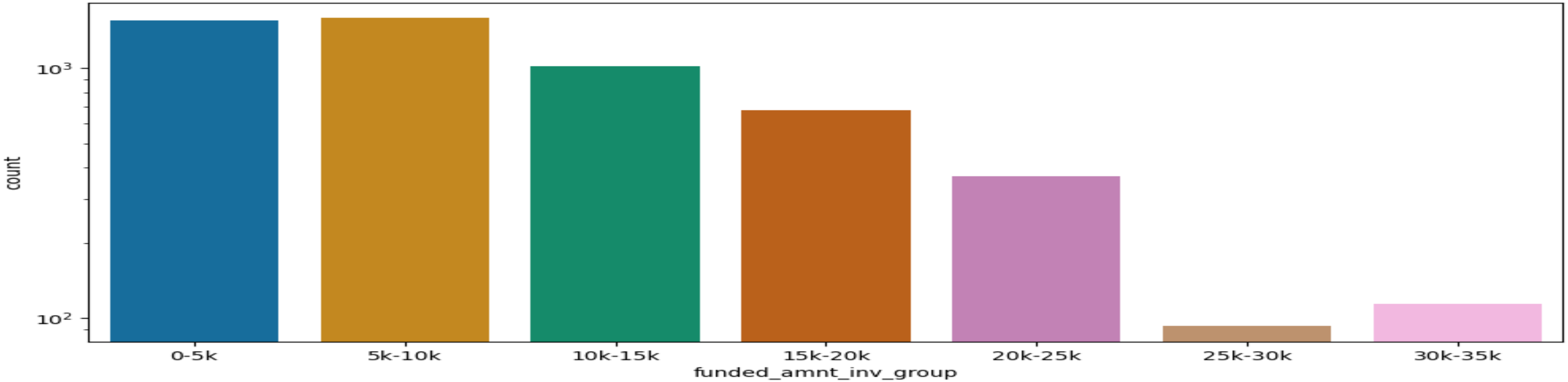
Univariate Analysis (Quantitative Variables)



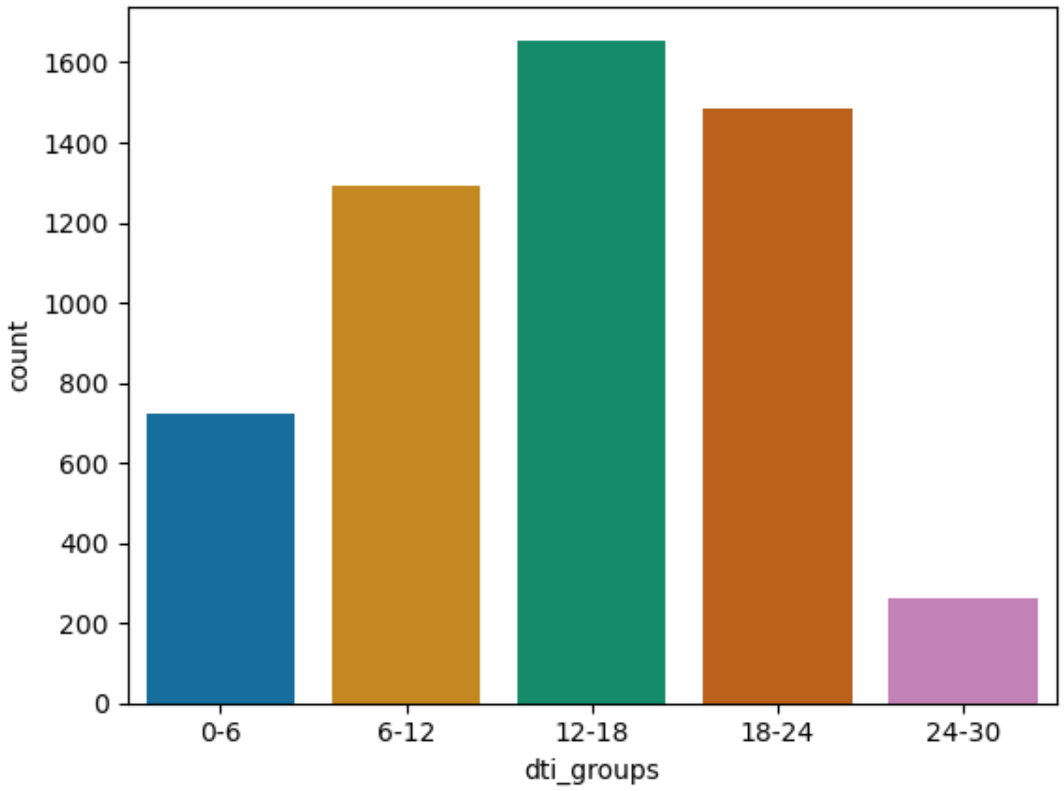


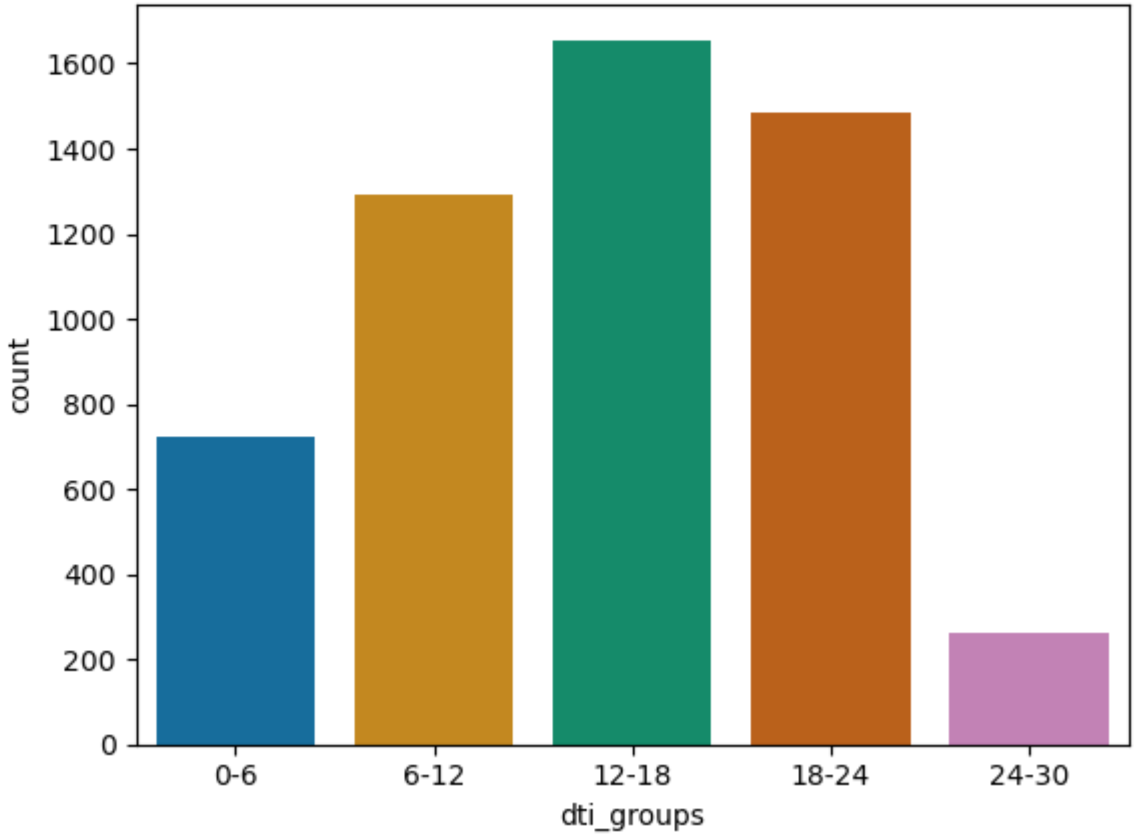
Univariate Analysis (Quantitative Variables)



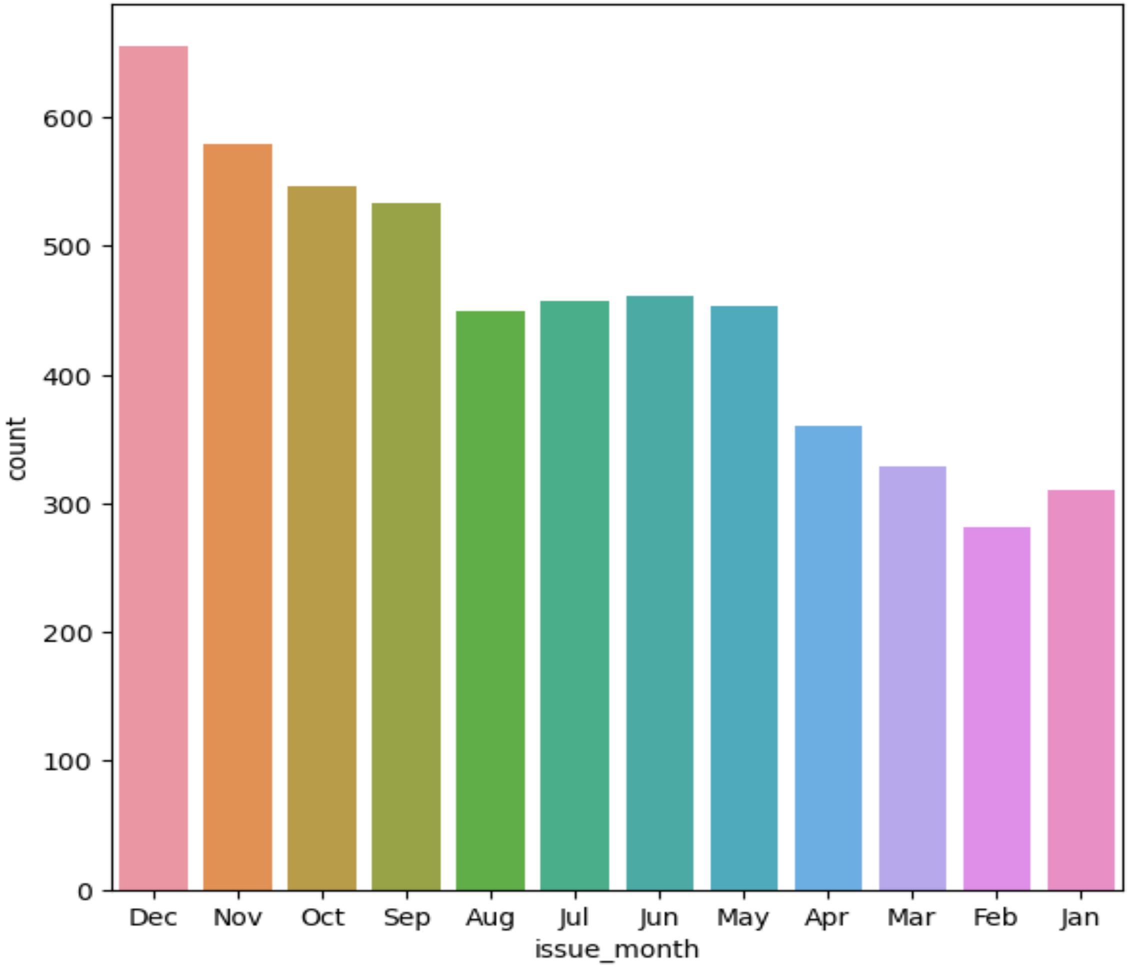


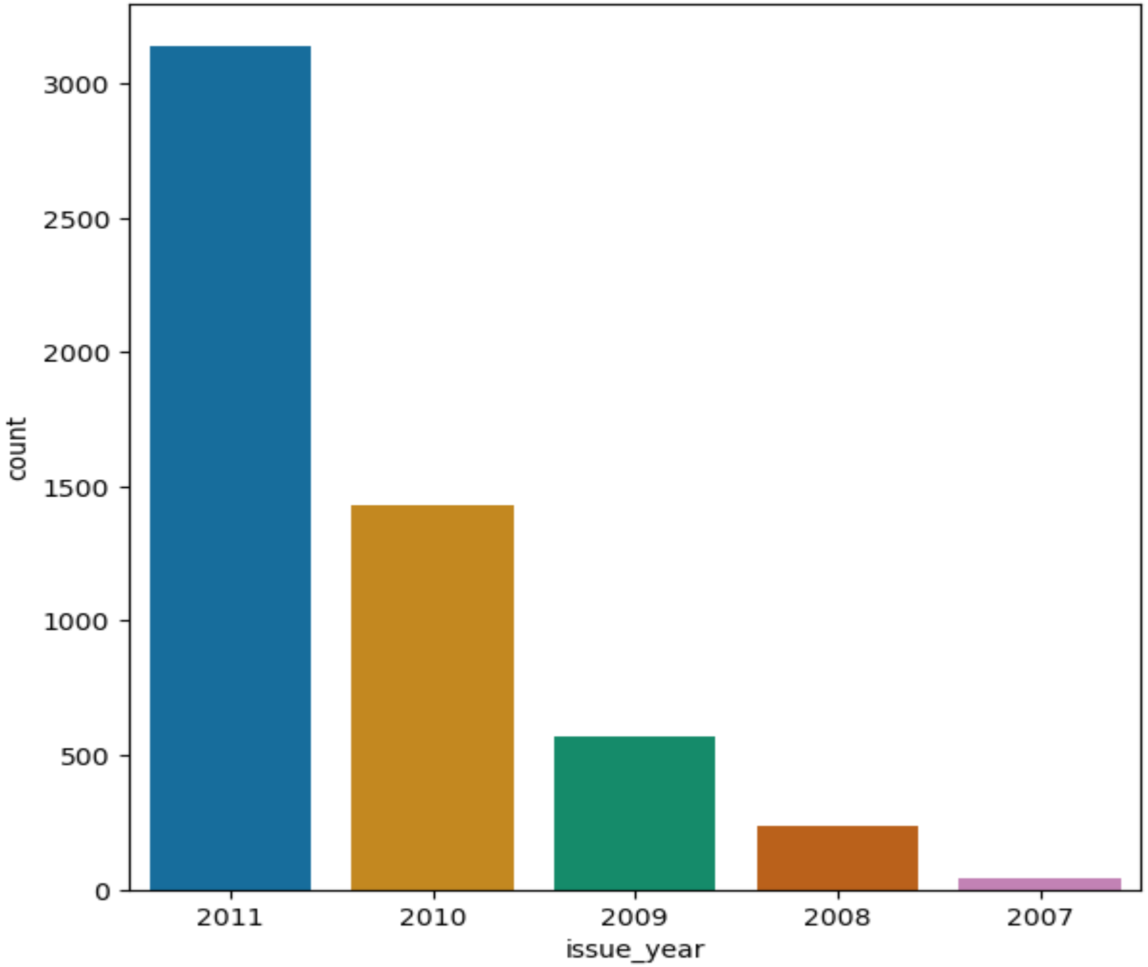
Univariate Analysis (Quantitative Variables)





Univariate Analysis (Quantitative Variables)





Univariate Analysis - Observations:

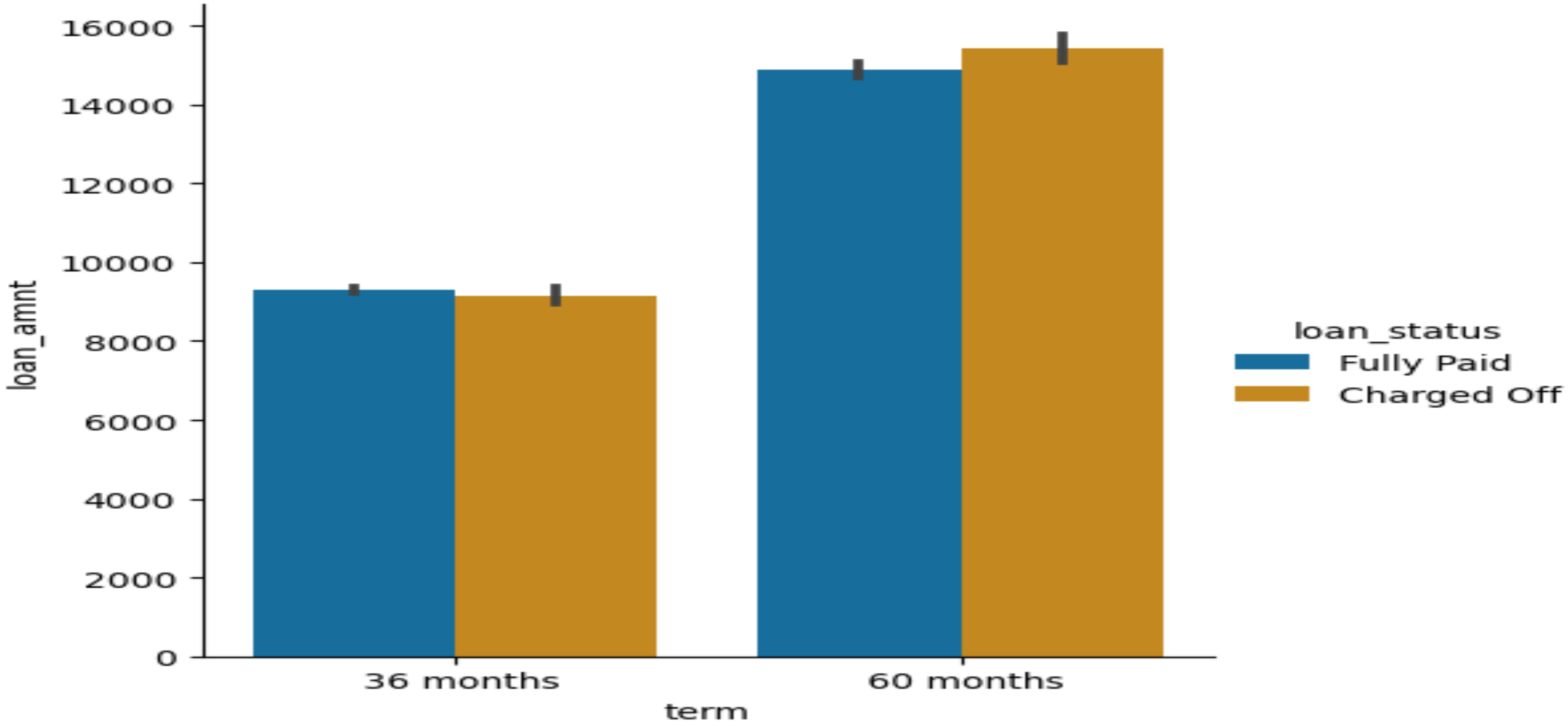
- 1,561 loan applicants who defaulted had annual salaries below $40,000. The lending company should be cautious when lending to individuals with lower annual incomes, ensuring thorough income verification and assessing repayment capacity more rigorously for this income group.
- Out of the 2,025 loan participants who defaulted, a significant number were in the 13%-17% interest rate bracket. To mitigate the risk of default, the lending company should consider offering loans at lower interest rates when feasible.
- 1,695 loan participants who defaulted had loan amounts of $15,000 or more. The lending company should carefully evaluate applicants requesting higher loan amounts, ensuring they have a strong credit history and the ability to repay larger loans.
- 1,608 loan participants who defaulted received funded amounts of $15,000 or more. The lending company should ensure that funded amounts match the borrower's financial capacity and conduct thorough credit assessments for larger loan requests.
- Among the loan participants who defaulted, 1,178 had very high debt-to-income ratios. The lending company should enforce strict debt-to-income ratio requirements to avoid lending to individuals with unsustainable levels of debt relative to their income.
- For those who defaulted, the majority had monthly installment amounts between $160 and $440. The lending company should closely monitor and assess applicants with similar installment amounts to reduce the risk of loan defaults.

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.

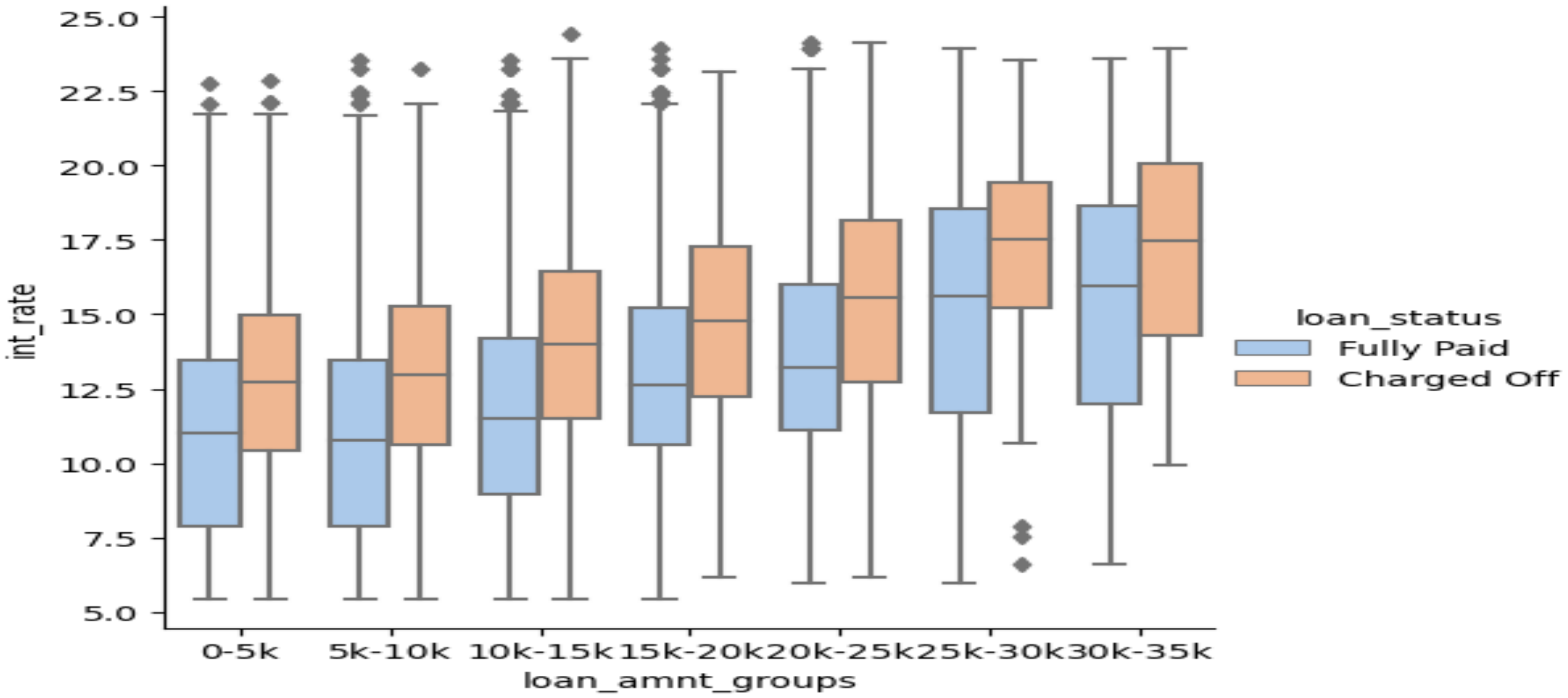
✓ It was carried out for both Categorical and Quantitative Variables

Bi Variate Analysis



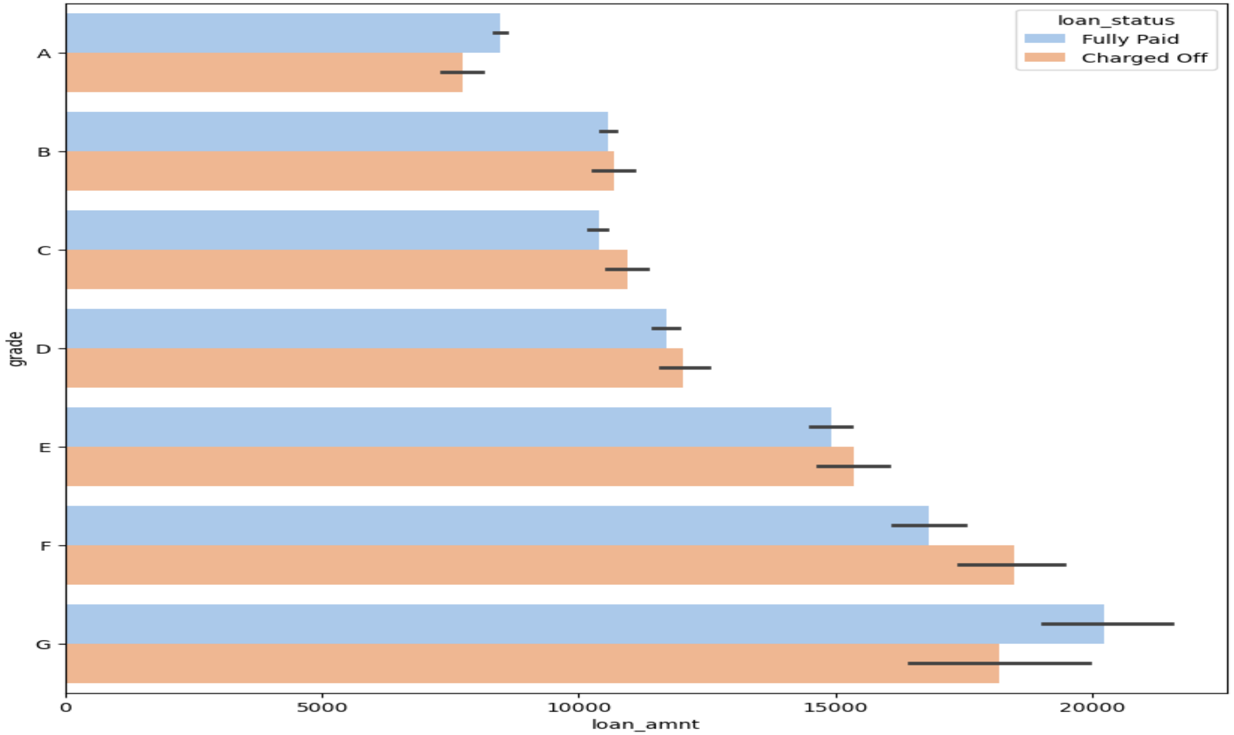
- Interestingly, the loan amount for Charged Off loans is slightly higher than that of Fully Paid loans for long term. This could indicate that higher loan amounts might be associated with a greater risk of default

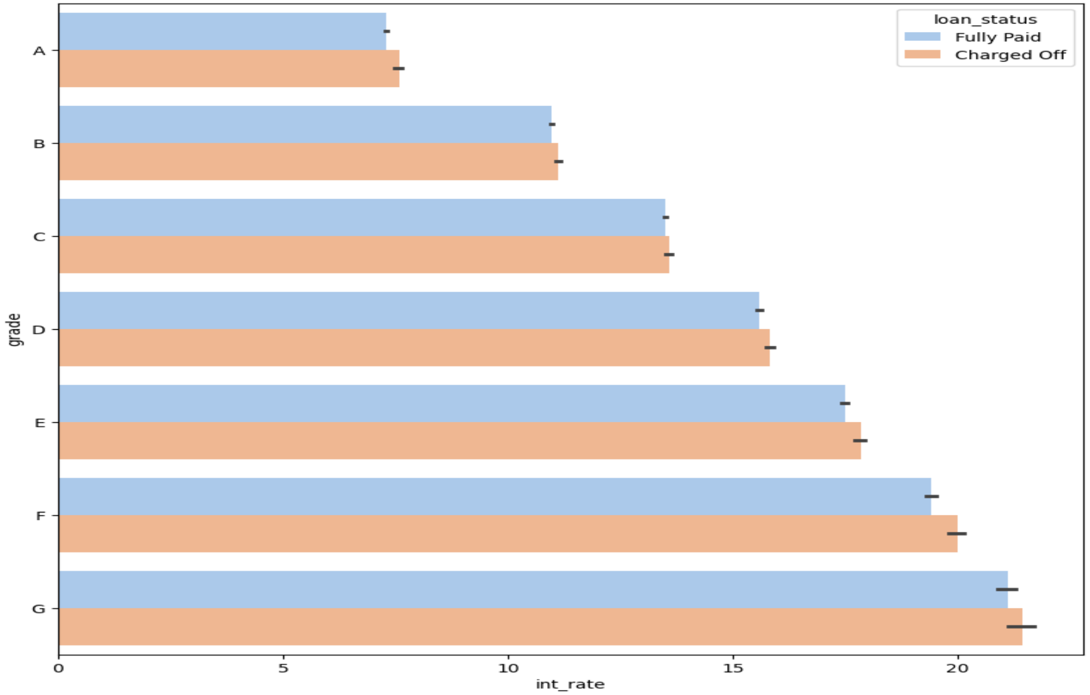
Bi Variate Analysis



- As the loan amount increases, the interest rate charged also tends to increase, regardless of whether the loan is eventually fully paid or charged off.

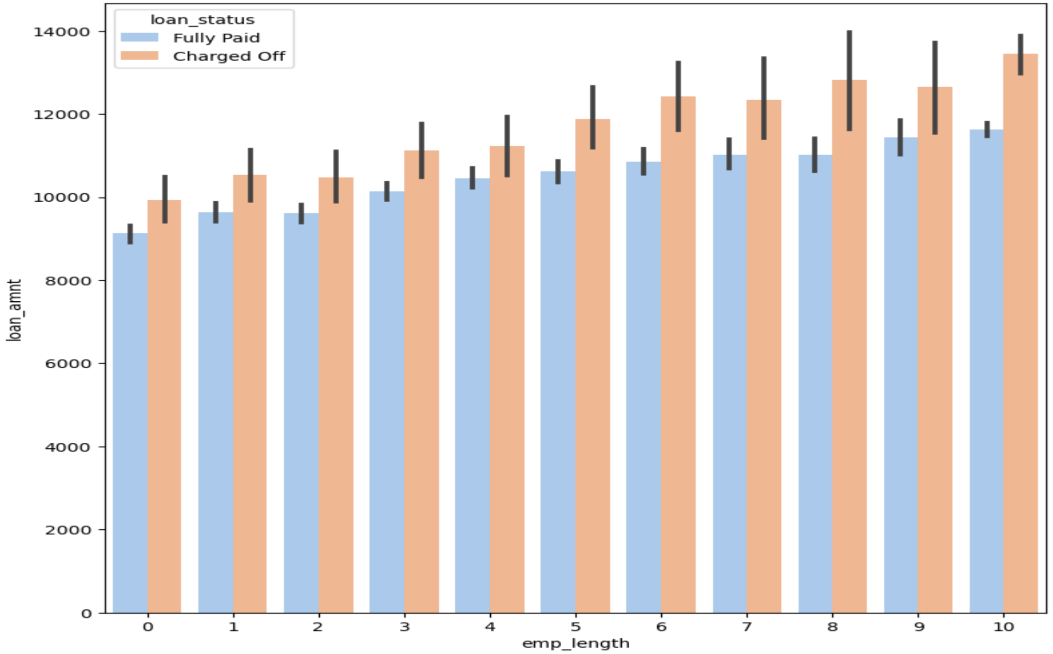
Bi Variate Analysis

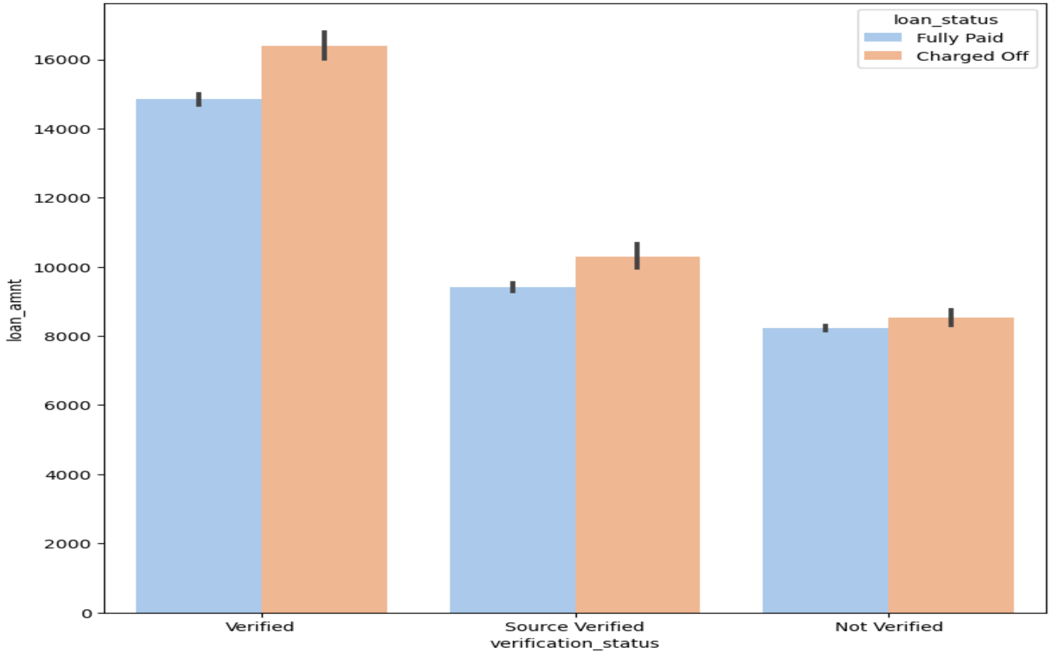




- Higher loan grades are associated with higher loan amounts.
- This indicates that lenders associate lower loan grades with higher risk, and therefore charge higher interest rates to compensate for that risk.

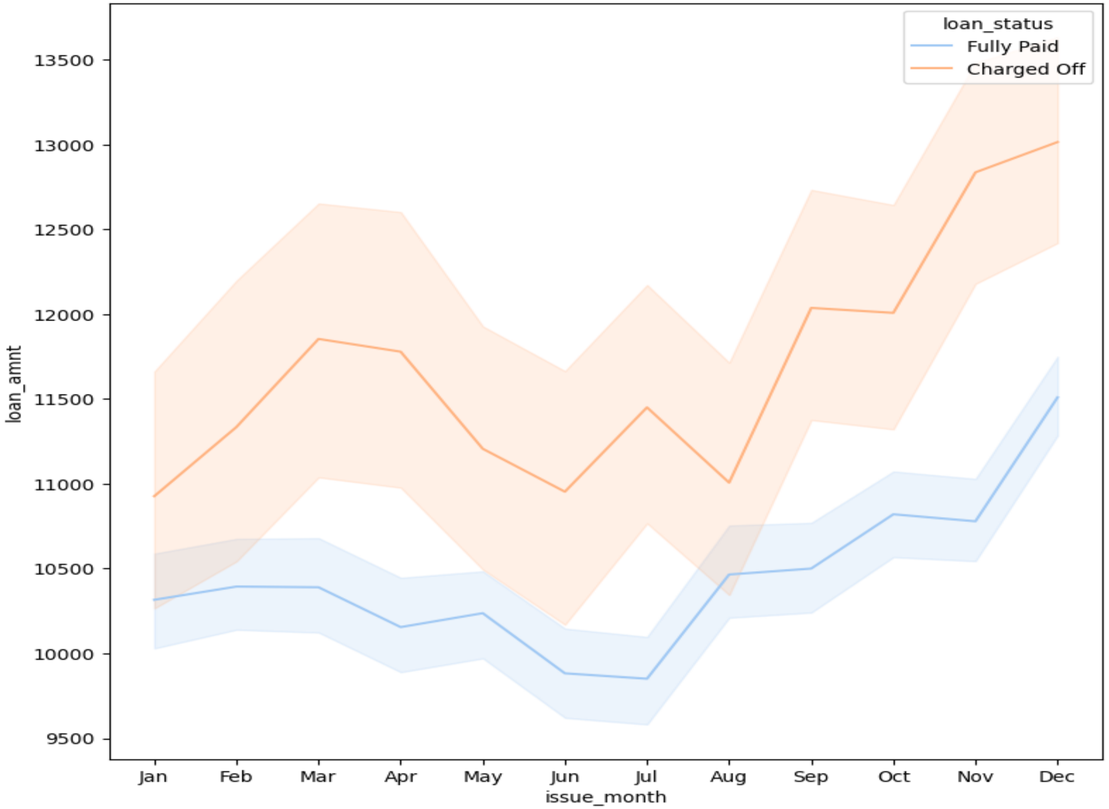
Bi Variate Analysis

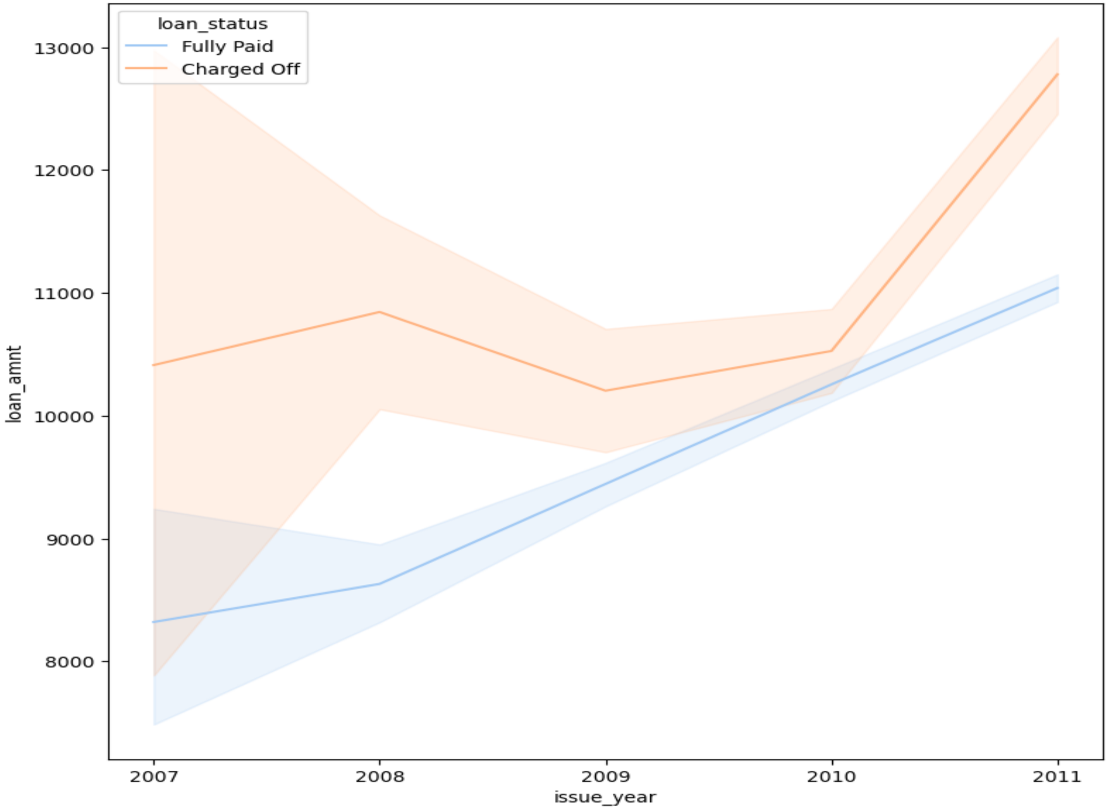




- These plots provide valuable insights into how loan amounts are influenced by employment length and verification status. While longer employment length is associated with slightly higher loan amounts, its predictive power for loan default seems limited.
- On the other hand, verification status appears to have a more noticeable impact, with verified borrowers receiving larger loans and potentially exhibiting a lower default risk compared to those who are not verified.

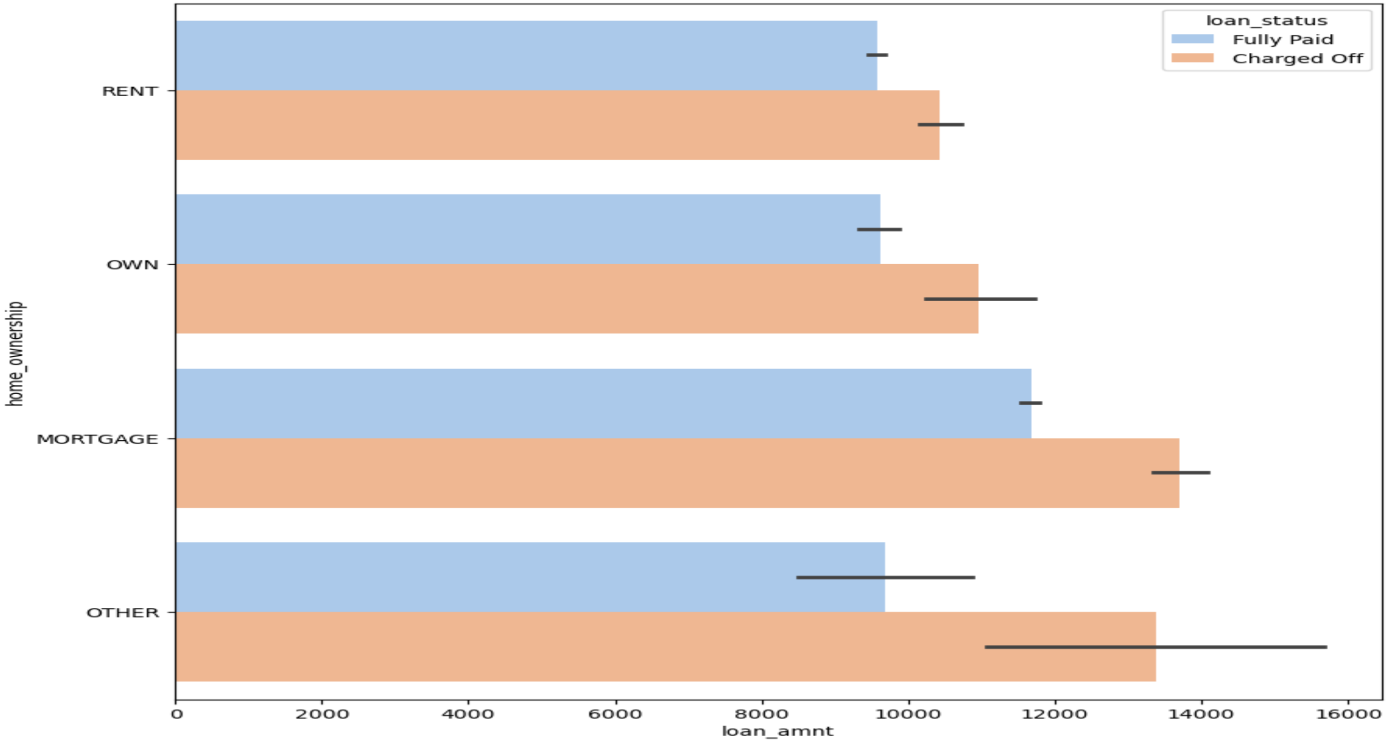
Bi Variate Analysis





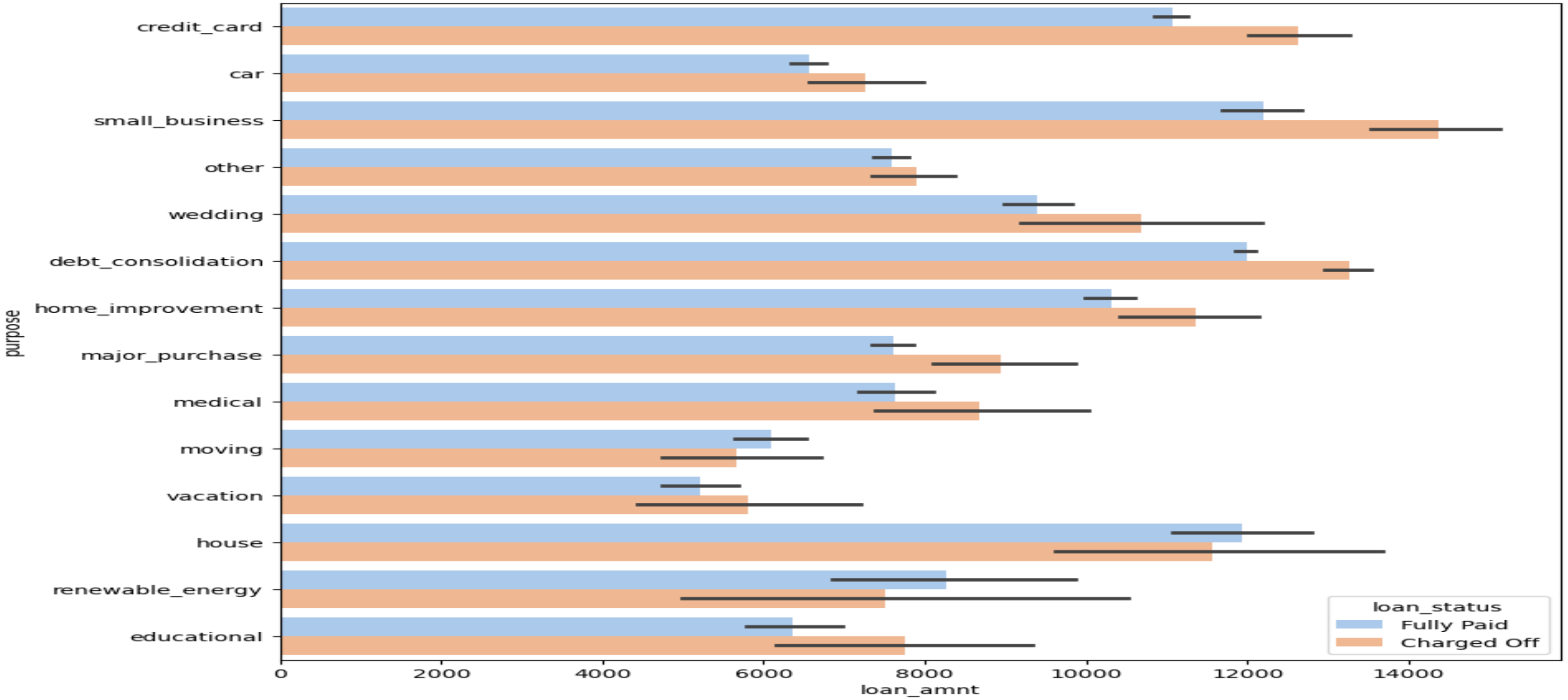
- Loan amounts tend to decrease from December to June and then slightly increase towards the year-end and decrease overall from 2007 to 2011.
- The average loan amount for charged-off loans is slightly higher, suggesting a potential correlation between higher loan amounts and increased default risk.

Bi Variate Analysis



Loans associated with **mortgages** tend to have higher amounts and show a notable inclination towards charge-offs, indicating a higher risk in this home ownership category.

Bi Variate Analysis



- **Debt consolidation and credit card loans** are the most common purposes for loans, with a significant proportion being fully paid off. However, they also show a notable amount of charge-offs, indicating a mixed risk profile.
- **Small business loans** have a relatively higher rate of charge-offs compared to other loan purposes, suggesting a higher risk associated with this loan type

Bivariate Analysis - Observations:

- The loan amount for Charged Off loans is slightly higher than that of Fully Paid loans for long term. This could indicate that higher loan amounts might be associated with a greater risk of default
- As the loan amount increases, the interest rate charged also tends to increase, regardless of whether the loan is eventually fully paid or charged off.
- Higher loan grades are associated with higher loan amounts.
- This indicates that lenders associate lower loan grades with higher risk, and therefore charge higher interest rates to compensate for that risk.
- These plots provide valuable insights into how loan amounts are influenced by employment length and verification status. While longer employment length is associated with slightly higher loan amounts, its predictive power for loan default seems limited.
- On the other hand, verification status appears to have a more noticeable impact, with verified borrowers receiving larger loans and potentially exhibiting a lower default risk compared to those who are not verified.

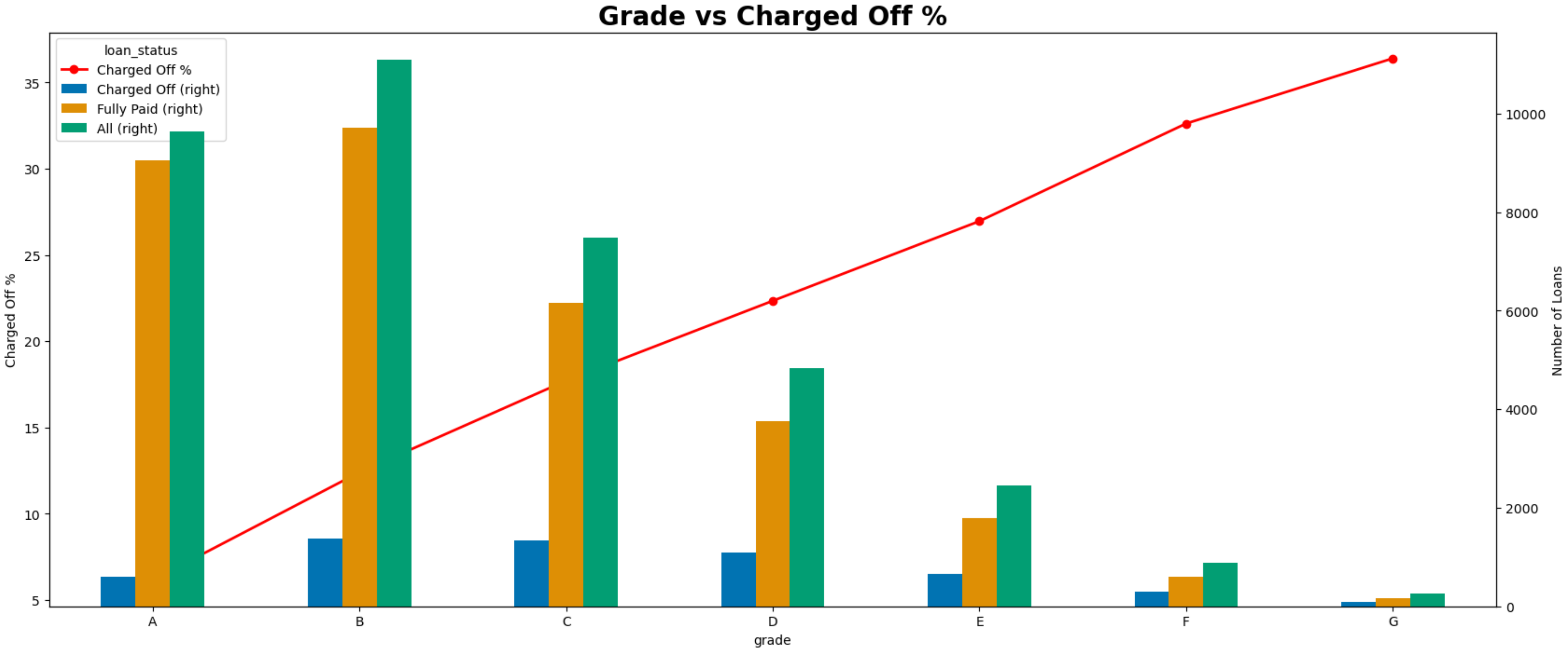
Bivariate Analysis – Observations(Contd):

- Loan amounts tend to decrease from December to June and then slightly increase towards the year-end and decrease overall from 2007 to 2011.
- The average loan amount for charged-off loans is slightly higher, suggesting a potential correlation between higher loan amounts and increased default risk.
- Loans associated with **mortgages** tend to have higher amounts and show a notable inclination towards charge-offs, indicating a higher risk in this home ownership category.
- **Debt consolidation and credit card loans** are the most common purposes for loans, with a significant proportion being fully paid off. However, they also show a notable amount of charge-offs, indicating a mixed risk profile.
- **Small business loans** have a relatively higher rate of charge-offs compared to other loan purposes, suggesting a higher risk associated with this loan type

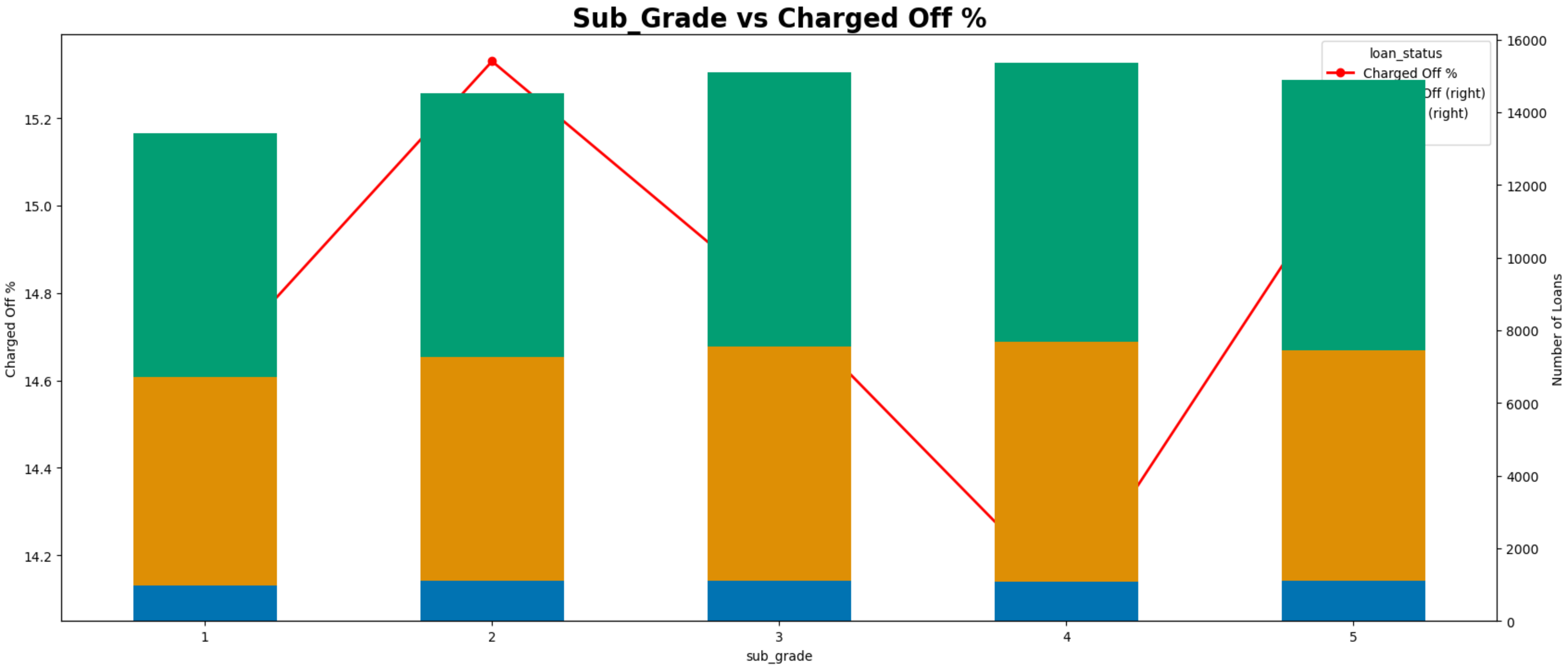
Multivariate Analysis

- Multivariate analysis can be used to predict a target variable based on multiple features, helping to understand the combined effect of different variables on the outcome.
- Multivariate analysis can be applied to both categorical and numerical variables:
  - **Numerical Variables**: Techniques like correlation matrices, pair plots, and multivariate regression are commonly used to analyse relationships and patterns among numerical variables.
  - **Categorical Variables**: Techniques such as Chi-square tests for independence, categorical regression models, and visualizations like stacked bar charts can be used to analyze relationships between categorical variables or between categorical and numerical variables

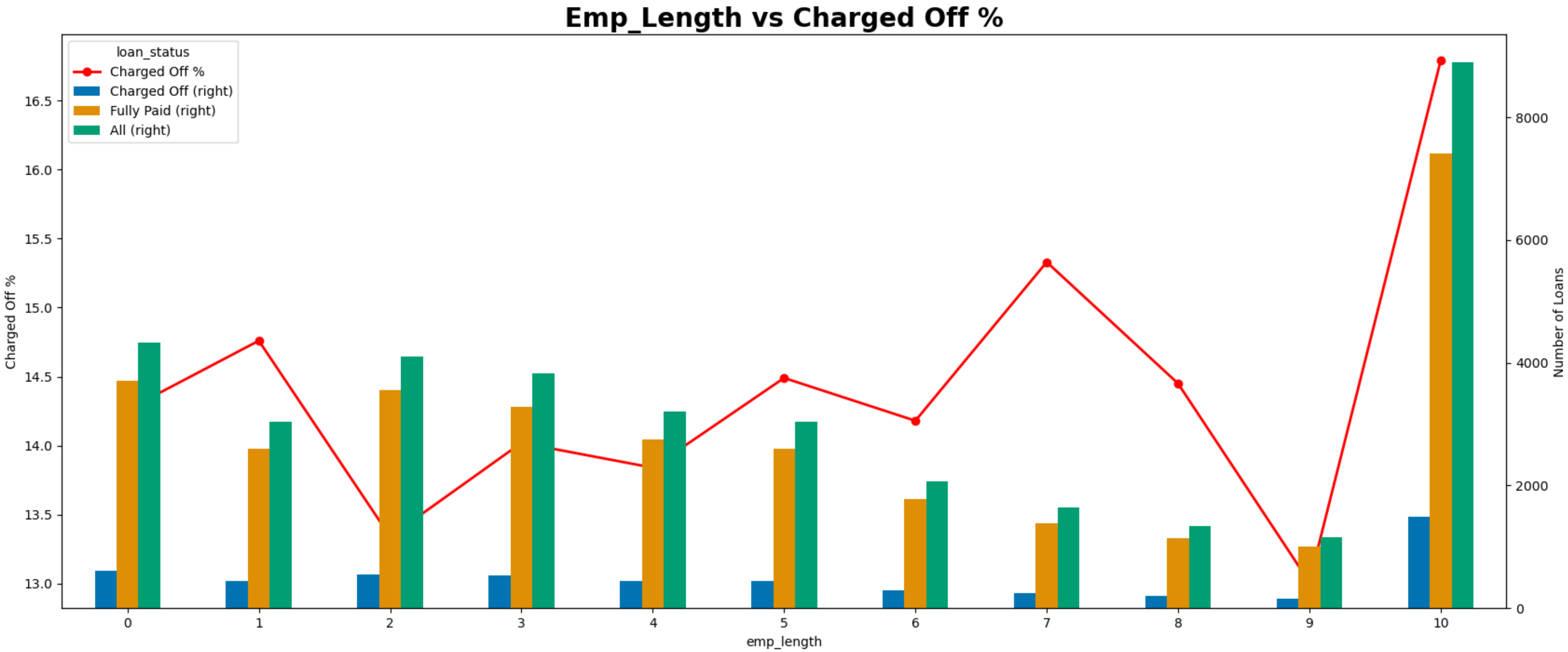
Multivariate Analysis



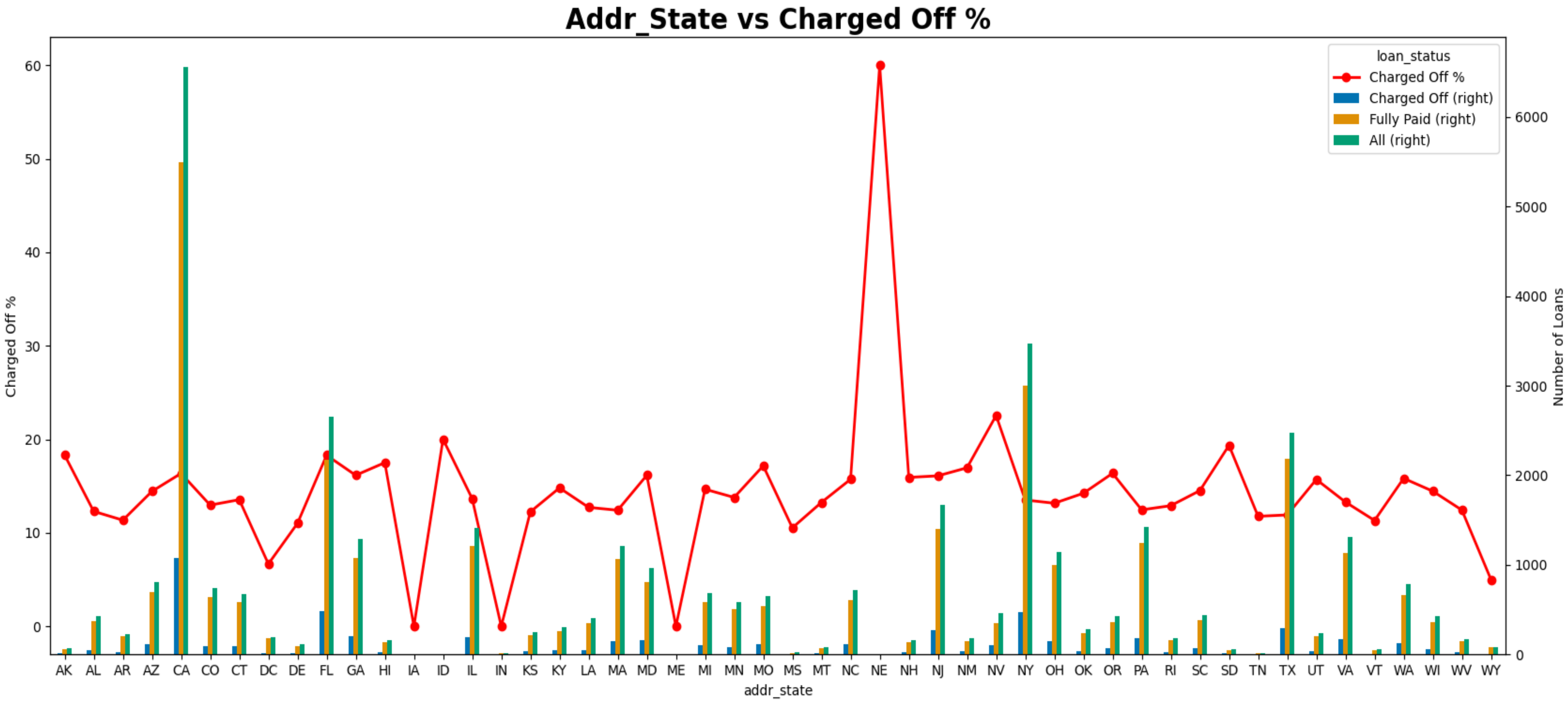
Multivariate Analysis



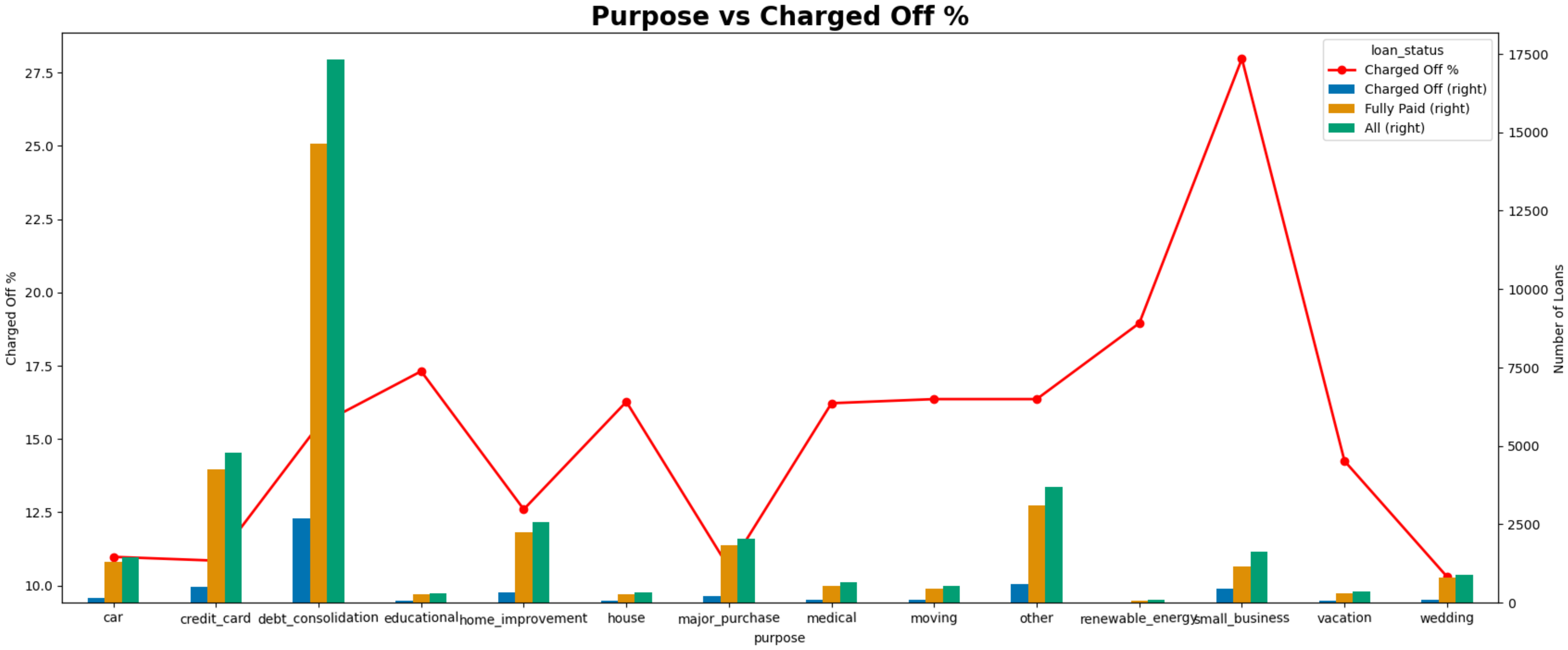
Multivariate Analysis



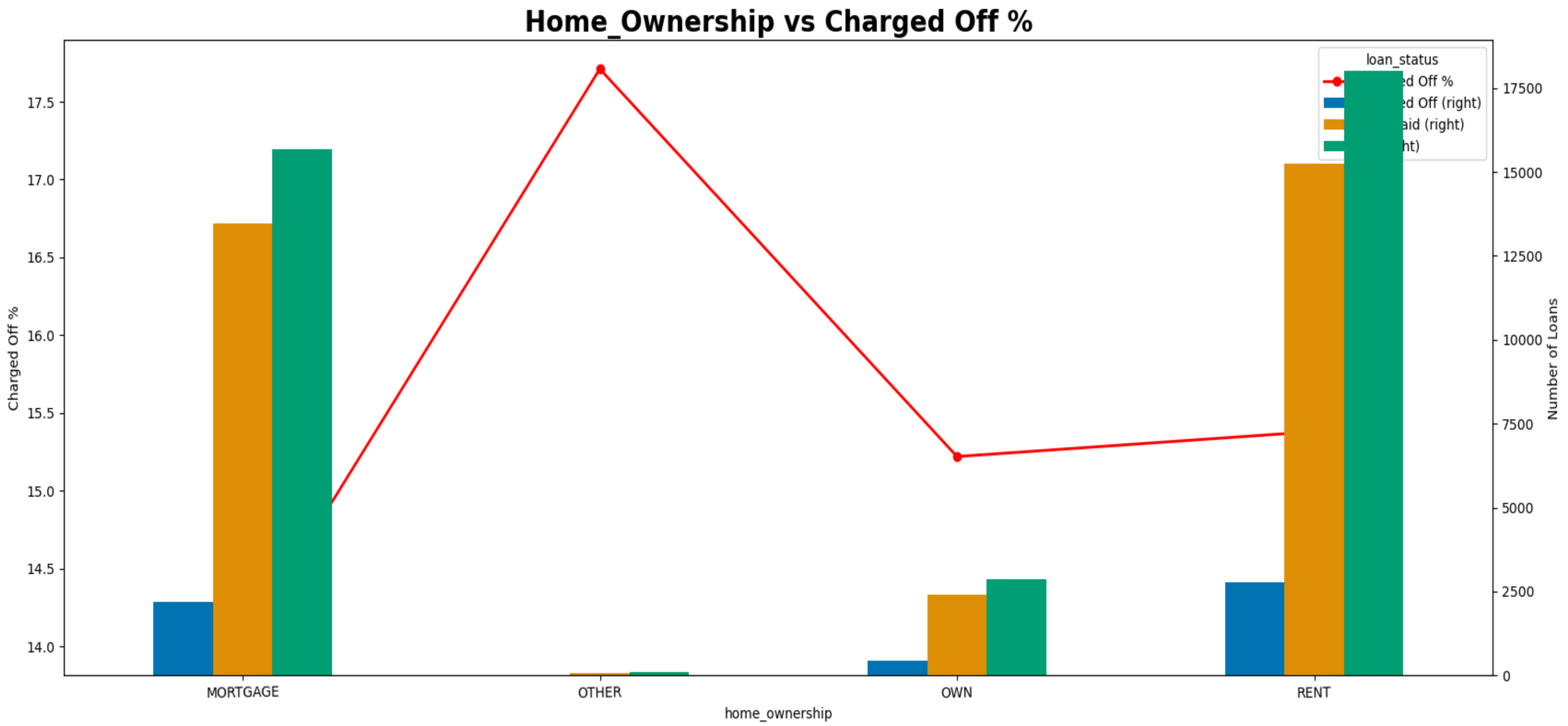
Multivariate Analysis



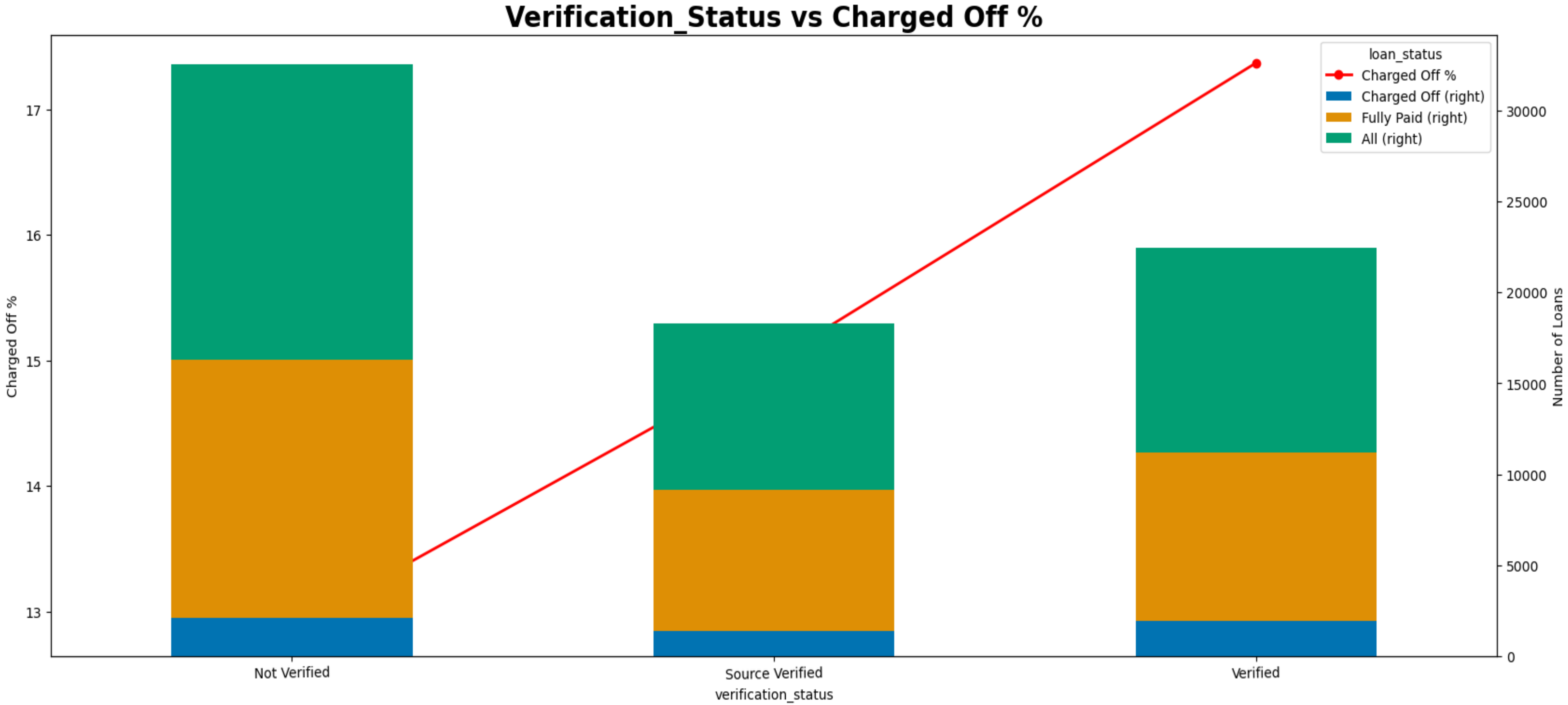
Multivariate Analysis



Multivariate Analysis



Multivariate Analysis



Multivariate Analysis

- The proportion of charged-off loans increases significantly as the loan grade decreases, indicating that lower loan grades are associated with a higher risk of default.
- As the sub-grade increases (from 1 to 5), the percentage of charged-off loans tends to increase, indicating higher default risk for higher sub-grades.
- The graph shows that borrowers with 0 or 10+ years of employment length have a lower percentage of charged-off loans compared to those with mid-range employment lengths (between 1 and 9 years).
- The state with the highest percentage of charged-off loans is **NE**, while the state with the lowest percentage is **HI**.**CA** has the highest number of loans, followed by **NY** and **TX**. While some states like **NE** and **NV** have high charged-off percentages, they have relatively low numbers of total loans.
- The purpose with the highest percentage of charged-off loans is 'small business', while ‘Renewable energy' has the lowest. 'Debt Consolidation' is the most common loan purpose, and it also has a noticeable percentage of charged-off loans.
- Renters have the highest percentage of charged-off loans compared to other homeownership categories. People with mortgages have the lowest percentage of charged-off loans. Most loans are given to renters, followed by those with mortgages.
- Loans with 'Verified' status have the lowest percentage of charged-off loans, followed by 'Source Verified', while 'Not Verified' loans have the highest charge-off rate. Most loans fall into the 'Verified’ category, indicating that most borrowers undergo income verification. Despite having the highest charge-off rate, 'Not Verified' loans represent the smallest portion of the total loans.