Lending Club Case Study

By Kakuli Saha & Kalyan Nath Somavarapu

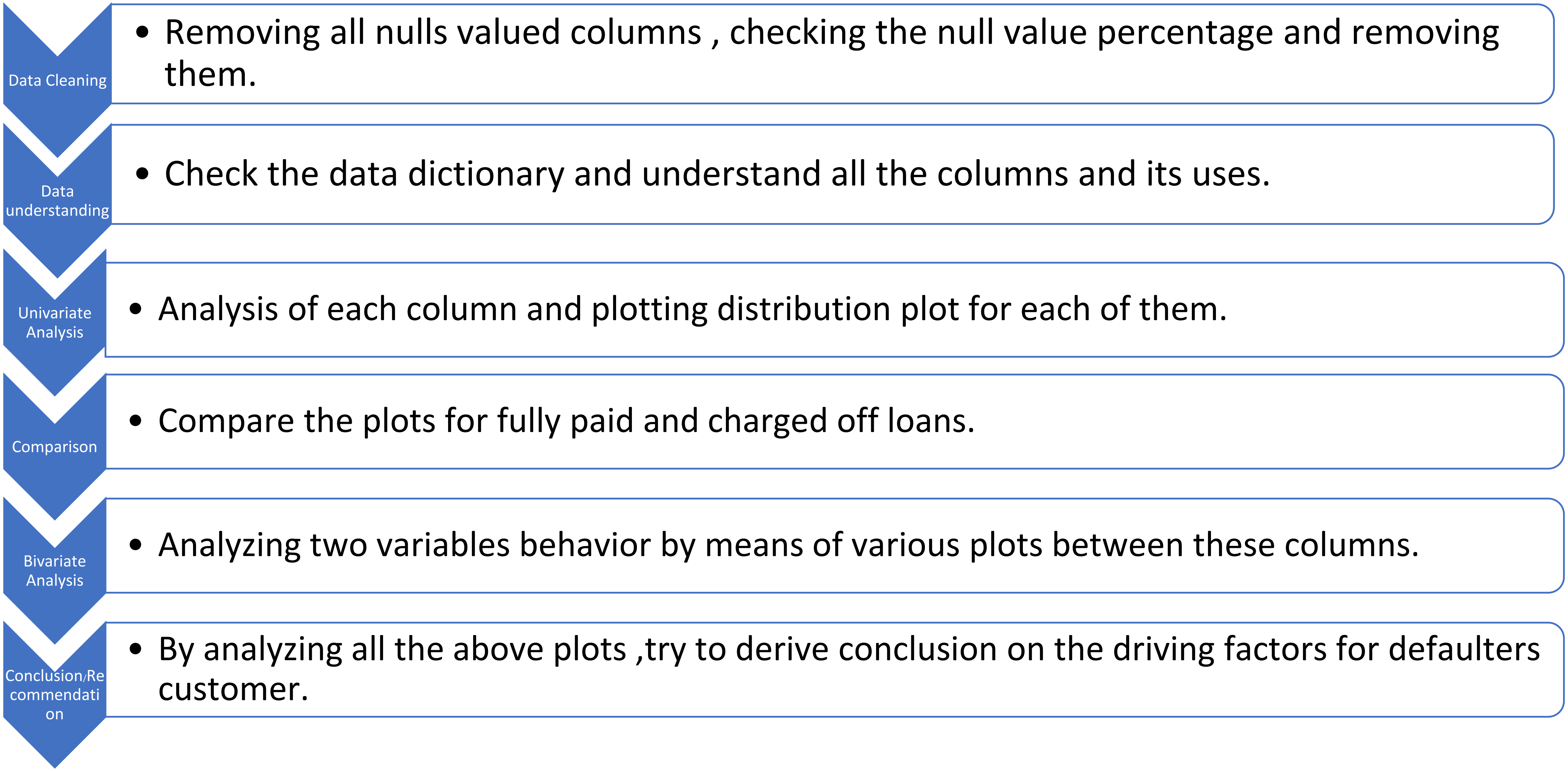
Problem Statement

1. Lending Club is an online platform that connects people who want loans with investors looking to lend money for a profit.
2. It offers different types of loans to city-based customers and needs to decide whether to approve each loan based on the applicant’s profile.
3. The biggest financial risk for the company is when borrowers don't repay their loans, leading to what’s called "credit loss." This happens when a borrower, known as a "defaulter," stops paying back their loan.
4. The main goal is to reduce these losses. The company must balance two risks:
    a. Rejecting good applicants who would repay and generate profit.
    b. Approving risky applicants who might default and cause a financial loss.

Objectives

- The goal is to identify risky loan applicants to reduce credit losses. This will be done through Exploratory Data Analysis (EDA) of the provided dataset.
- The company wants to understand the key factors that lead to loan defaults. By identifying these indicators, they can improve their risk assessment and manage their loan portfolio more effectively.

Analysis Approach



Data Cleaning

- Removing all nulls valued columns , checking the null value percentage and removing them.

Data understanding

- Check the data dictionary and understand all the columns and its uses.

Univariate Analysis

- Analysis of each column and plotting distribution plot for each of them.

Comparison

- Compare the plots for fully paid and charged off loans.

Bivariate Analysis

- Analyzing two variables behavior by means of various plots between these columns.

Conclusion/Recommendation

- By analyzing all the above plots ,try to derive conclusion on the driving factors for defaulters customer.

Data understanding

- The dataset provides information about past loan applicants and whether they defaulted. The goal is to spot patterns that indicate if someone is likely to default, which can guide decisions like denying loans, reducing loan amounts, or offering higher interest rates to risky applicants.
- The data focuses only on loans that were approved, so it doesn’t include any rejection criteria.
- The main objective is to find key indicators that contribute to defaults and use this analysis to form hypotheses. The loan process has three steps:
- A borrower requests a loan amount (loan_amnt).
- The approver decides to approve or reject based on past data (funded_amnt).
- The final loan amount is funded by the investor (funded_amnt_inv).

Data cleaning

- Rows where the **loan_status = CURRENT will be dropped** as CURRENT loans are in progress and will not contribute in the decision making of pass or fail of the loan. The rows are dropped before the column analysis as it also cleans up unnecessary column related to CURRENT early and columns with NA values can be cleaned in one go
- Find duplicate rows in the dataset and drop if there are.

Data cleaning

- There are multiple columns with NA values only. The columns will be dropped.
- This is evaluated after dropping rows with loan_status = Current
- There are multiple columns where the values are only zero, the columns will be dropped
- There are columns where the values are constant. They don't contribute to the analysis, columns will be dropped
- There are columns where the value is constant but the other values are NA. The column will be considered as constant.

Data cleaning

- **Drop customer behaviour columns which represent data post the approval of loan**
  - They contribute to the behaviour of the customer. Behaviour of the customer is recorded post approval of loan and not available at the time of loan approval. Thus these variables will not be considered in analysis and thus dropped
  - (delinq_2yrs, earliest_cr_line, 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_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, application_type)

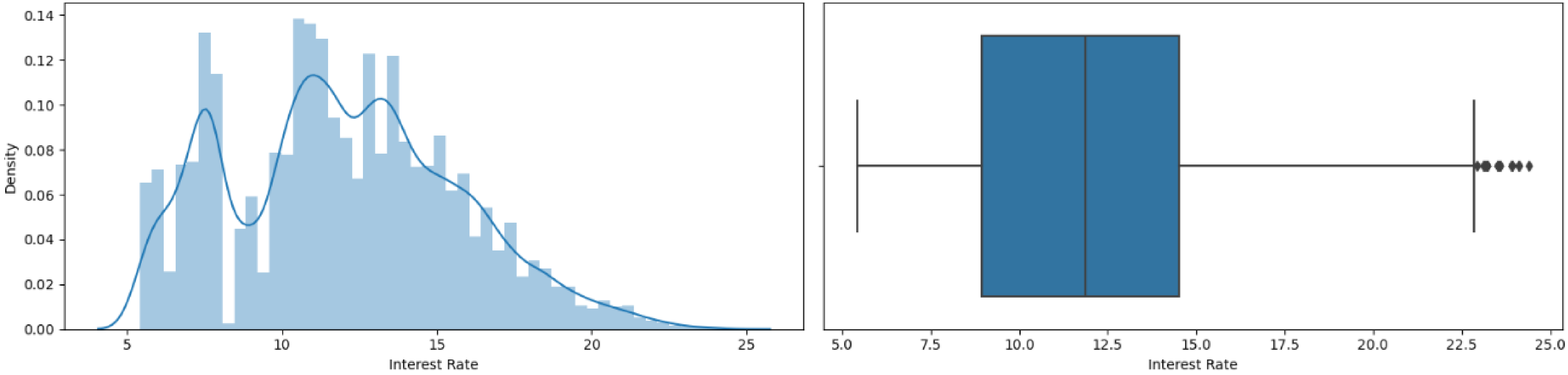
Data cleaning

- There are columns where more than 65% of data is empty (mths_since_last_delinq, mths_since_last_record) - columns will be dropped
- Drop columns (id, member_id) as they are index variables and have unique values and don't contribute to the analysis
- Drop columns (emp_title, desc, title) as they are descriptive and text (nouns) and don't contribute to analysis
- Drop redundant columns (url). On closer analysis url is a static path with the loan id appended as query. It's a redundant column to (id) column

Data cleaning

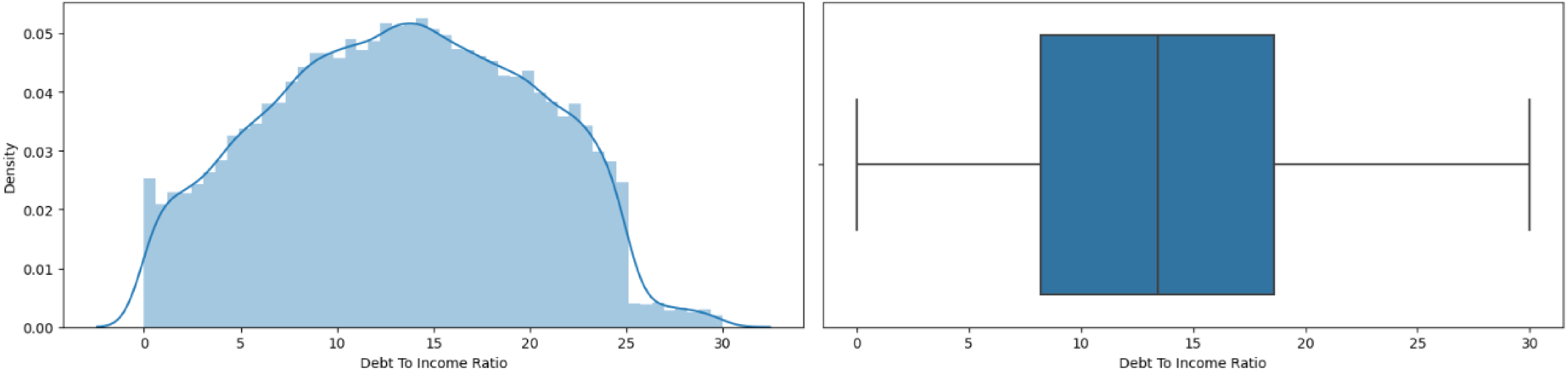
- `(loan_amnt, funded_amnt, funded_amnt_inv)` columns are Object and will be converted to float
- `(int_rate, installment, dti)` columns are Object and will be converted to float
- **strip "month"** text from `term` column and convert to integer
- Percentage columns `(int_rate)` are object. **Strip "%"** characters and convert column to float
- `issue_d` column **converted to datetime format**
- `loan_status` column converted to boolean **Charged Off = False and Fully Paid = True**. This converts the column into ordinal values
- `emp_length` converted to integer with following logic. Note < 1 year is converted to zero and 10+ converted to 10.

Univariate analysis



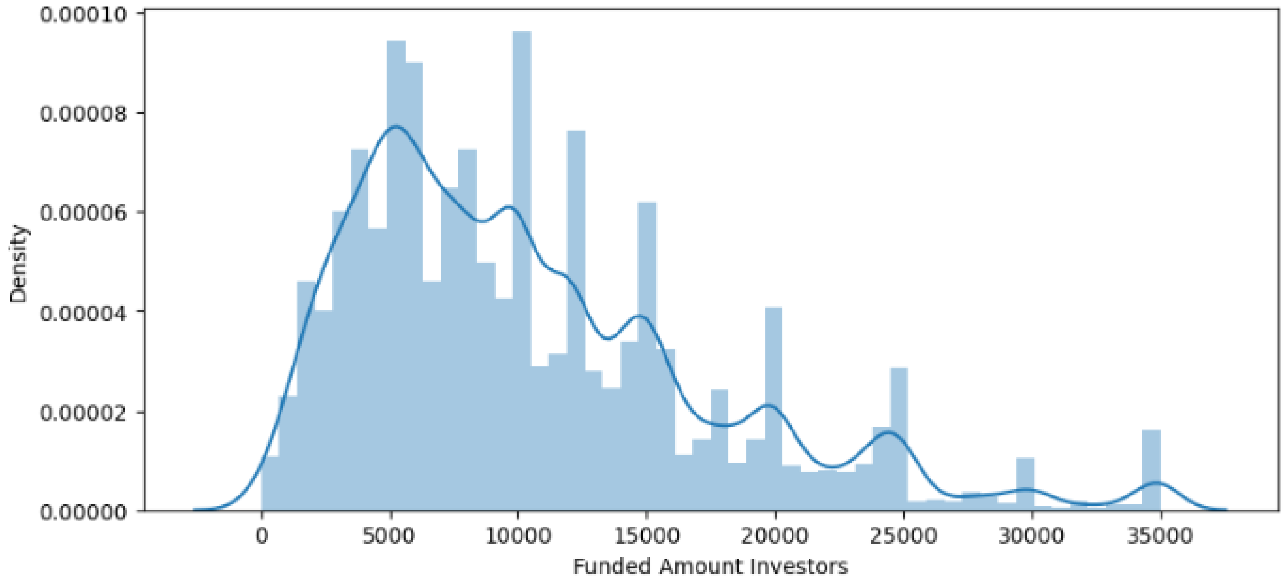
Majority of the interest rate is in the range of 5% to 16% going at the max to 22.5%.

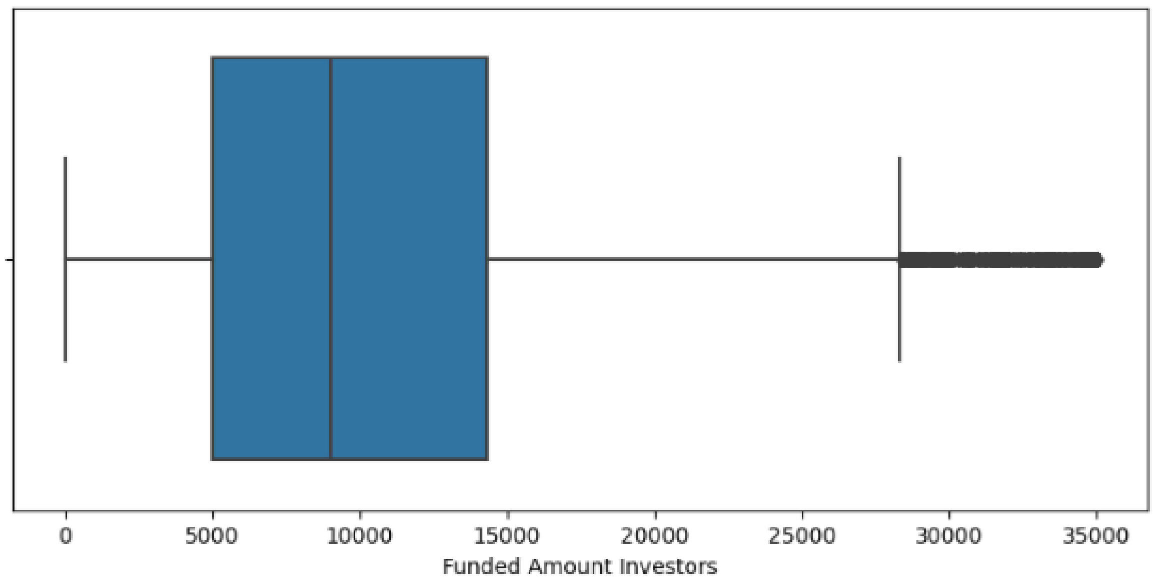
Univariate analysis



Majority of the debt to income is in the range of 10 to 20 going at the max to 30

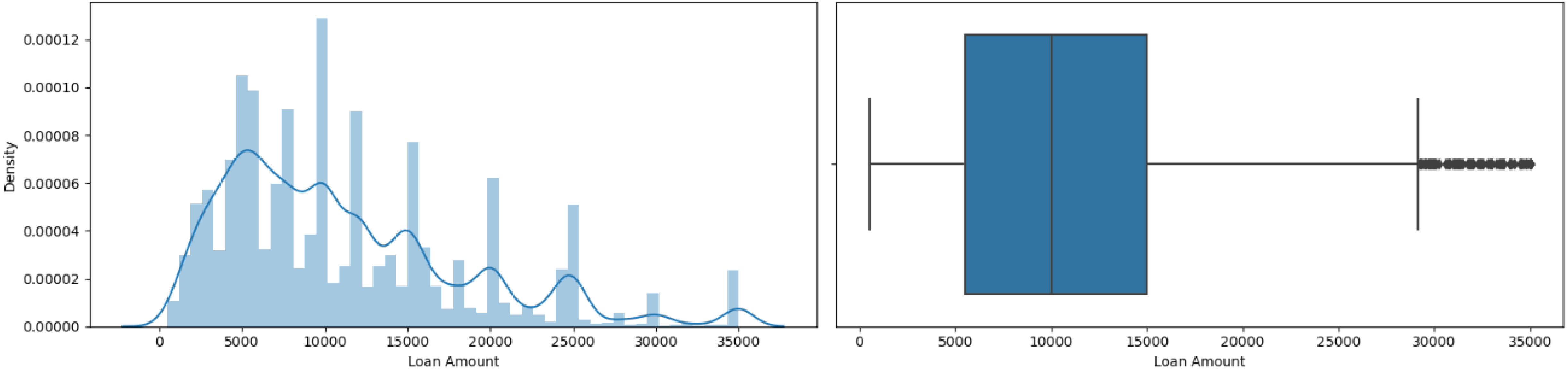
Univariate analysis





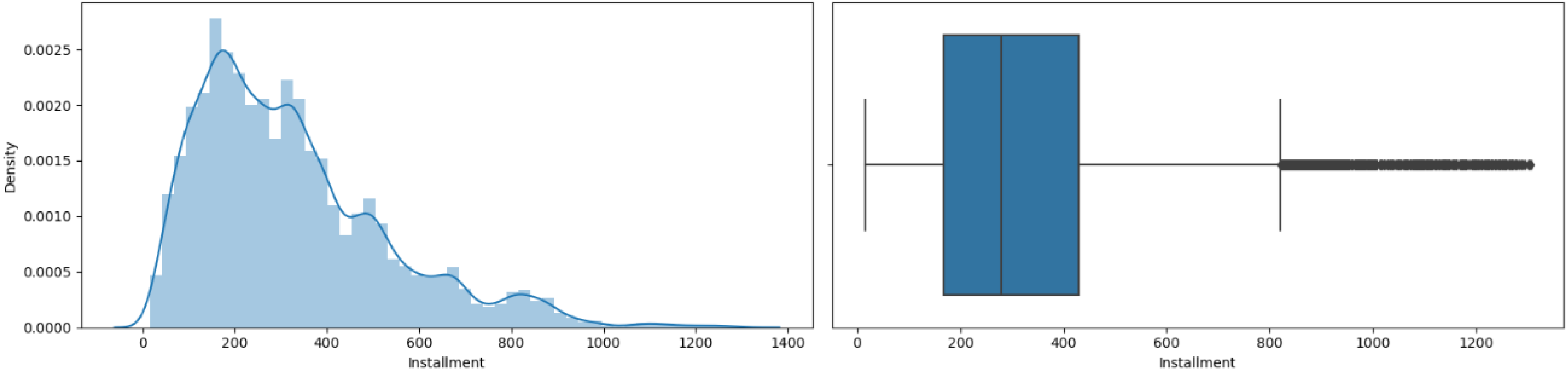
Majority of the funded_amnt_inv is in the range of 5K to 12K

Univariate analysis



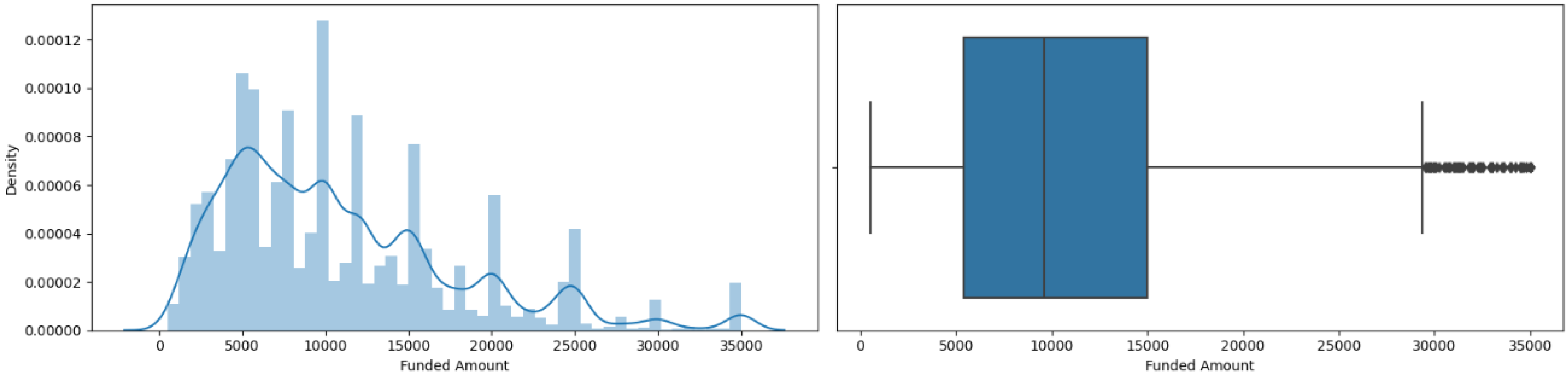
Majority of the loan_amount is in the range of 5K to 15K

Univariate analysis



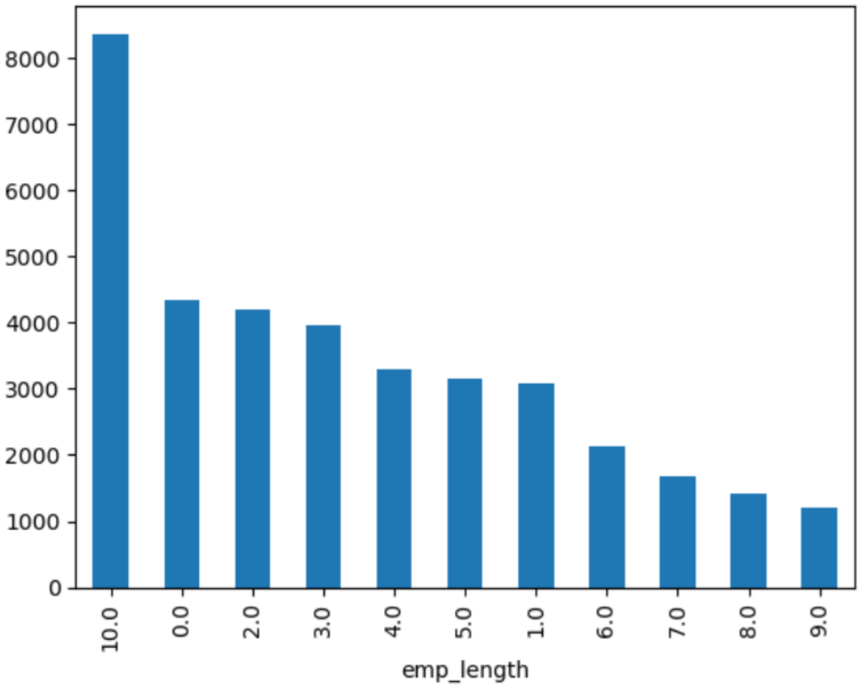
Majority of the installment is in the range of 20 to 400 going at the max to 700

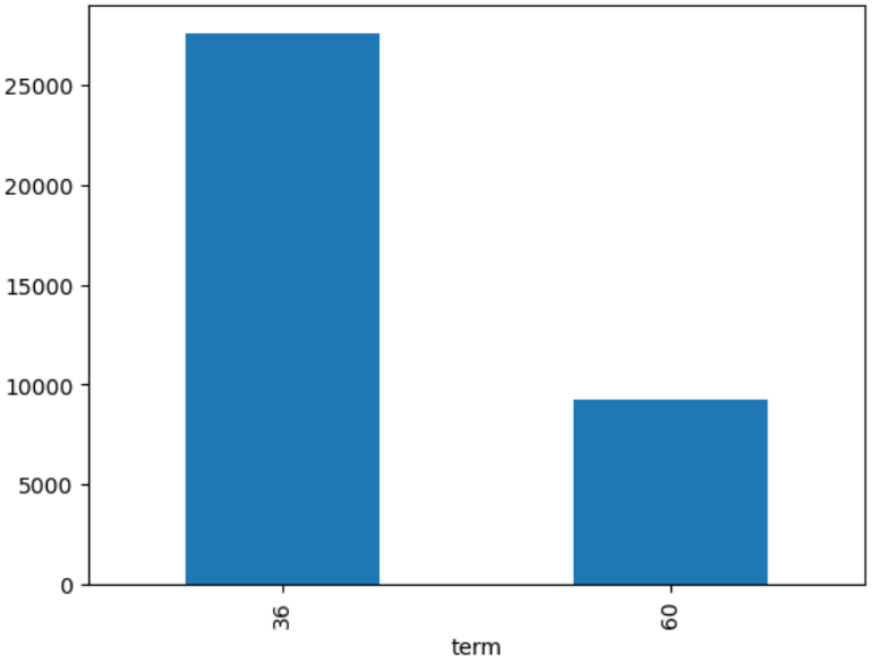
Univariate analysis



Majority of the funded_amnt is in the range of 5K to 15K

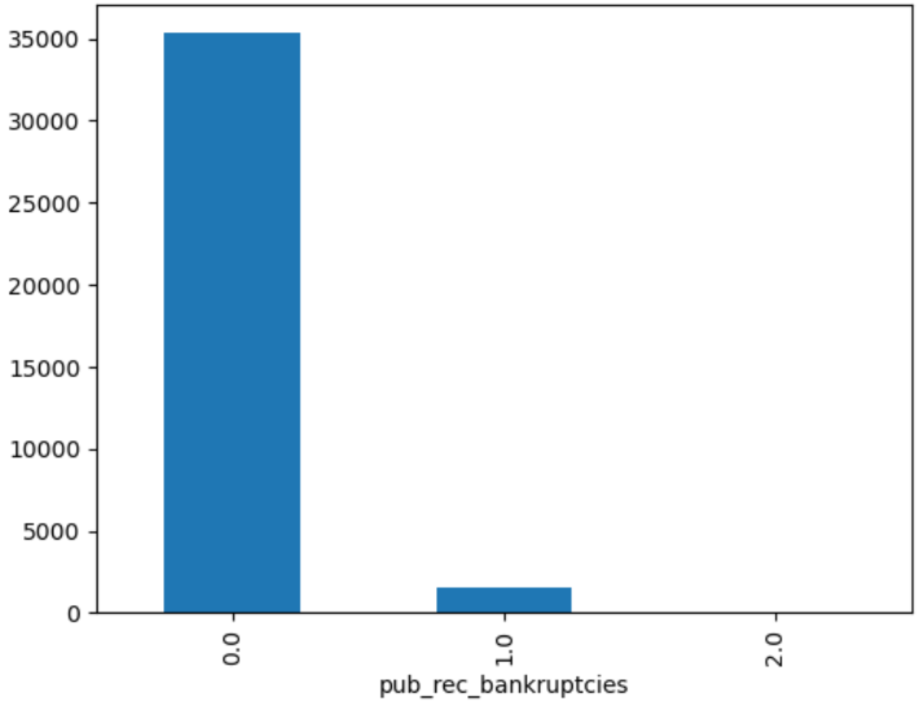
Univariate analysis

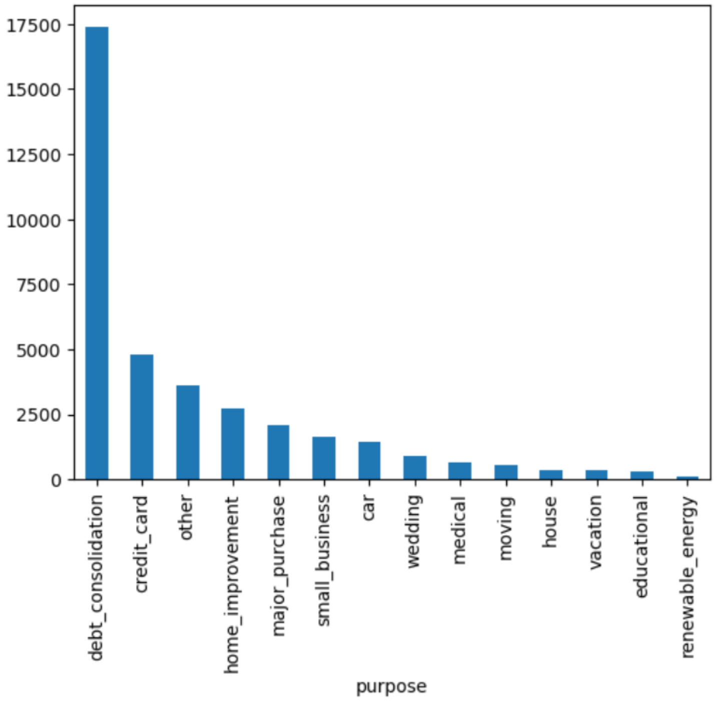




- Majority of the employment length of the customers are 10+ years and then in the range of 0-3 years
- Majority of the loan applications counts are in the term of 36 months.

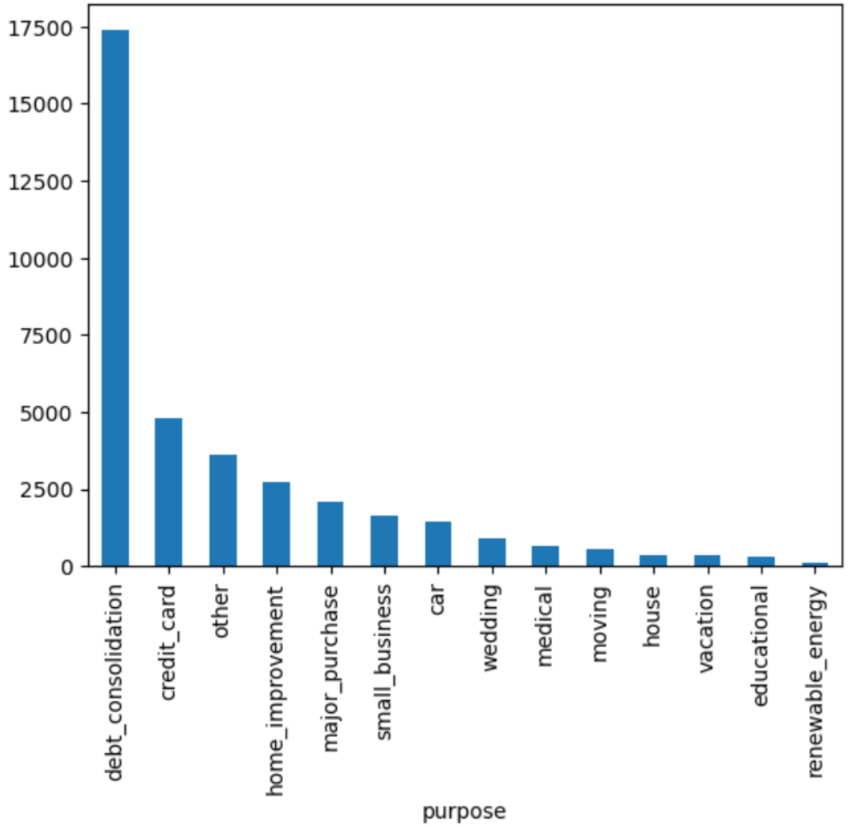
Univariate analysis

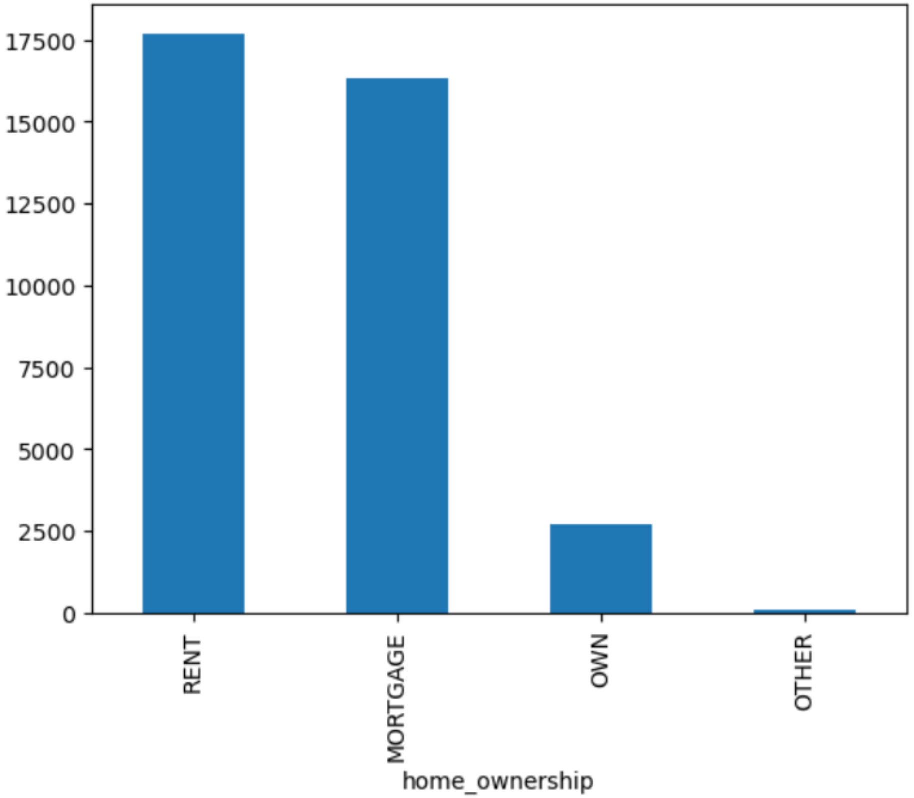




- Majority of the loan applicants are in the category of not having an public record of bankruptcies
- Majority of loan application are in the category of debt_consolidation

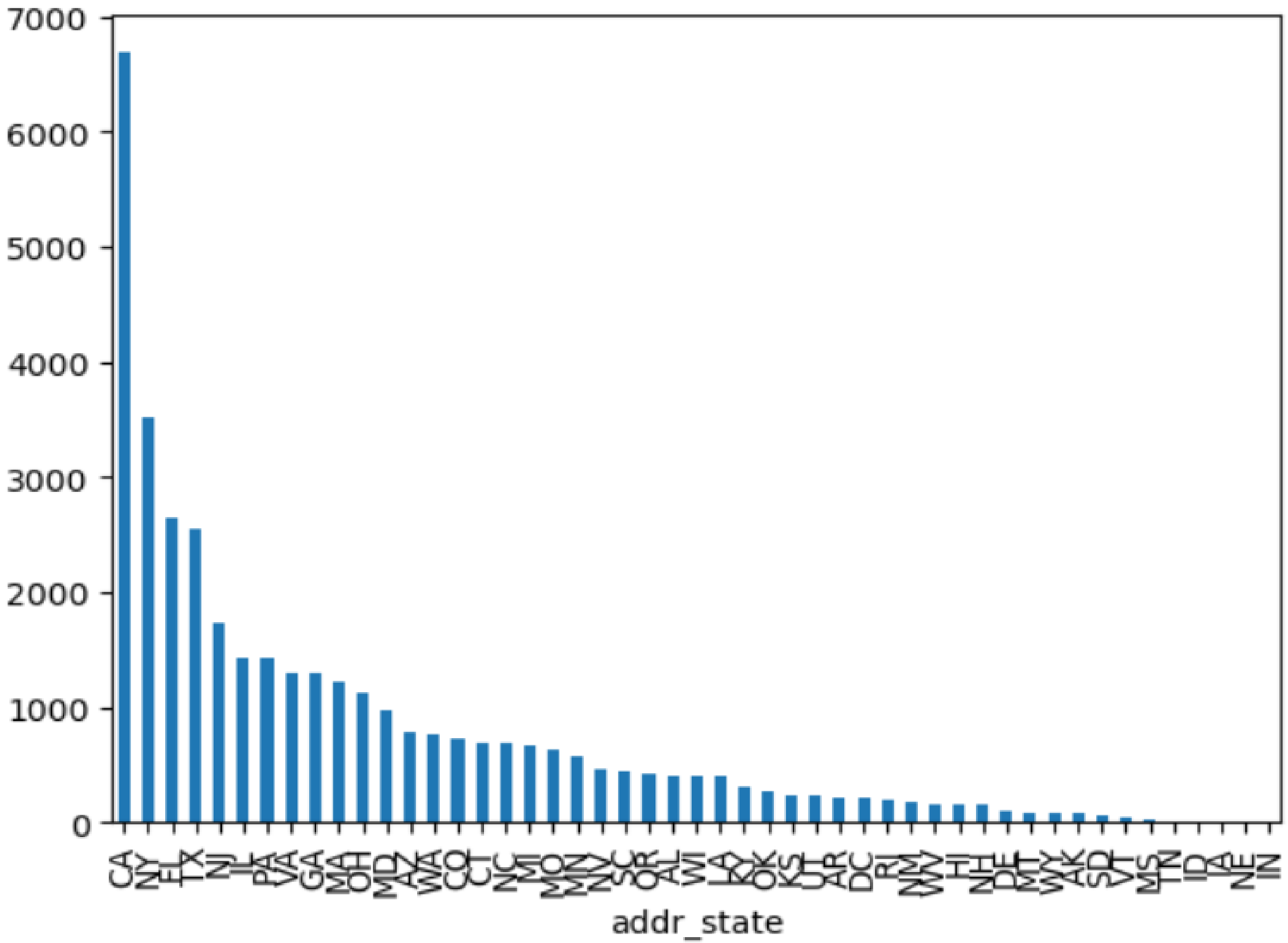
Univariate analysis





- Majority of loan application counts fall under the category of Grade B and A
- Majority of the home owner status are in status of RENT and MORTGAGE

Univariate analysis

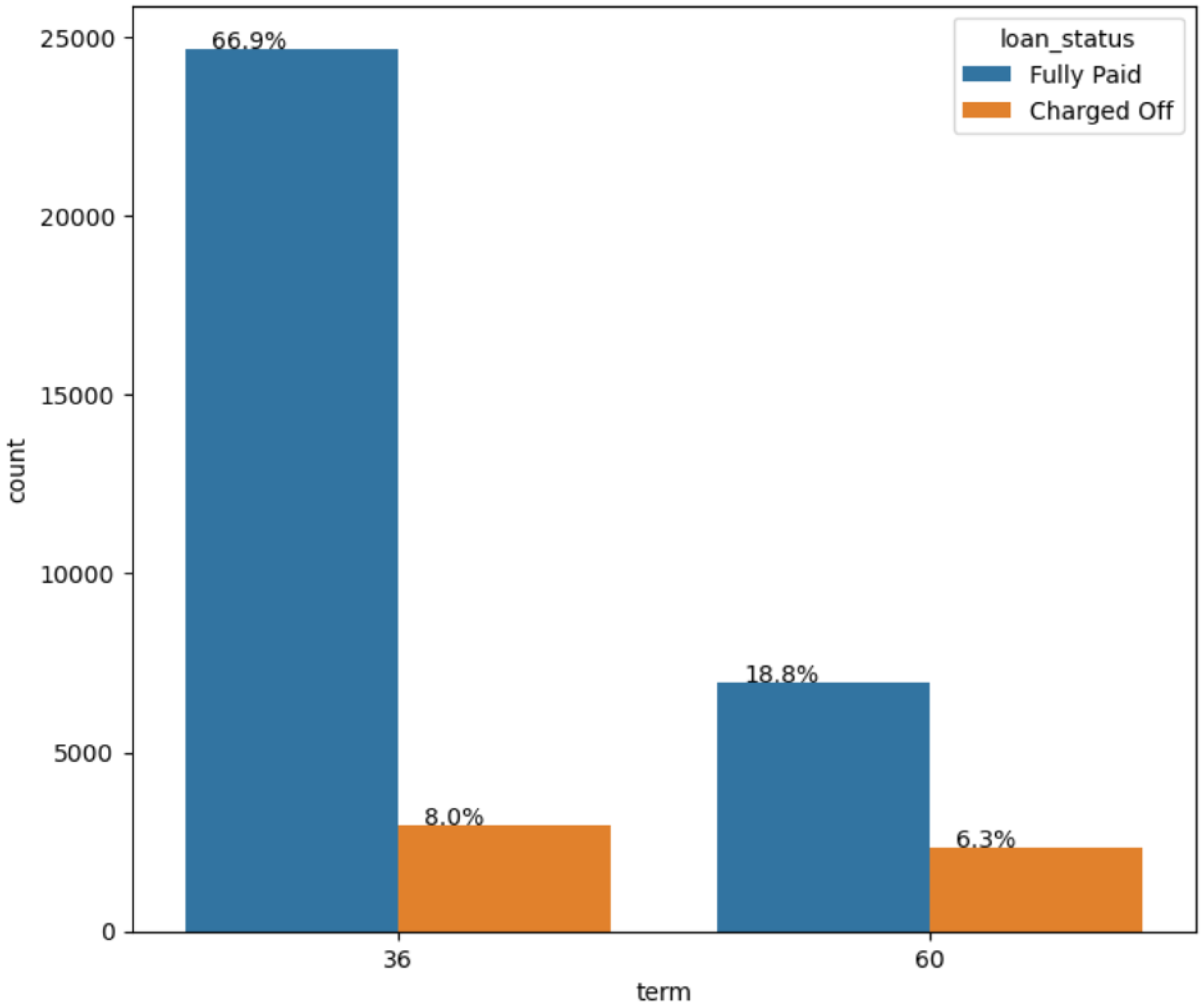


- CA state has the maximum amount of loan applications

Univariate Analysis Inferences

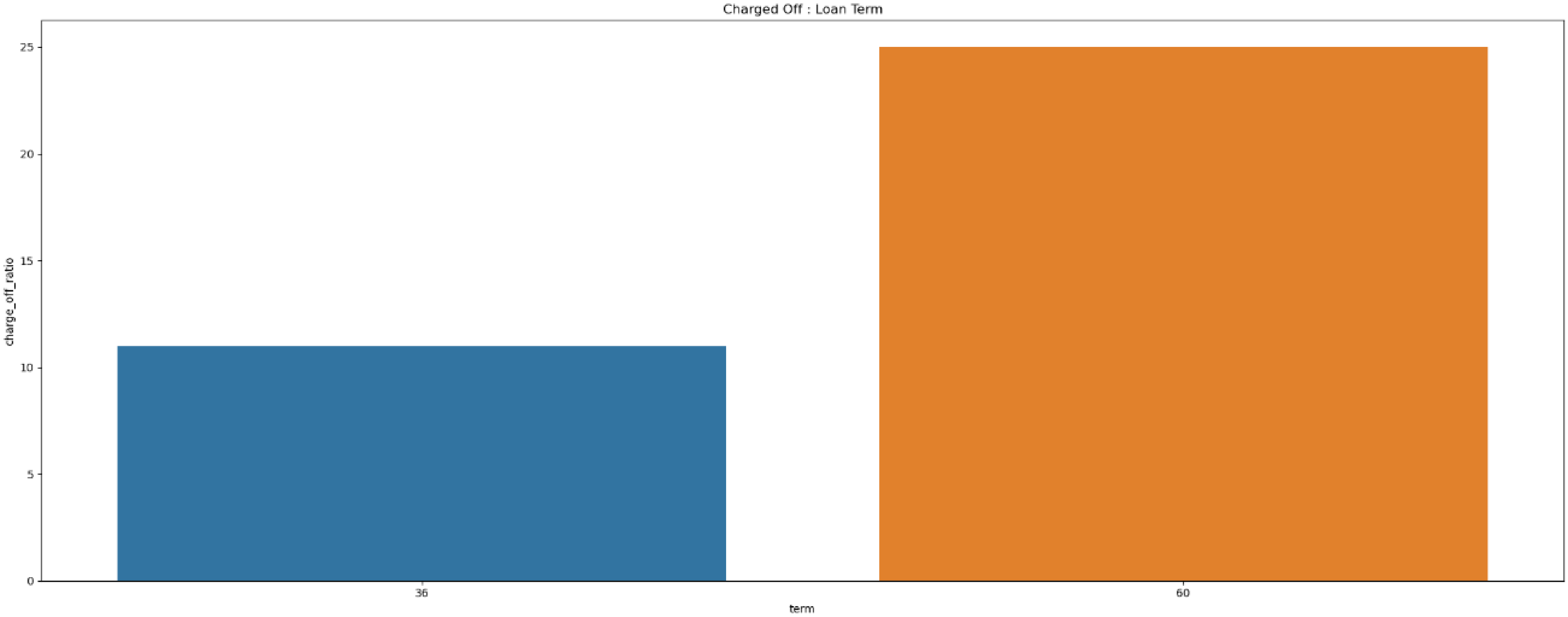
1. Most loan applicants earn between 0 and 40,000 annually.
2. The majority have a debt-to-income ratio between 0 and 20, with a few going up to 30.
3. Most people in the dataset are either renting their homes or have a mortgage.
4. The most common reason for loan applications is debt consolidation.
5. California has the highest number of loan applications.
6. Most applicants do not have a public record of bankruptcies.
7. The majority of applicants have either over 10 years or between 0-2 years of employment.
8. The most frequent loan amounts are between 5,000 and 10,000.
9. Most interest rates range from 5% to 16%, with a few going up to 22%.
10. The typical installment amount is around $20.
11. Most loans are applied for with a term of 36 months.
12. Most loans fall into Grade B.
13. Understanding customer demographics helps identify the most promising segments for targeting loan applications.
14. Further analysis is needed to explore why some categories have lower application rates compared to others.

Bivariate analysis



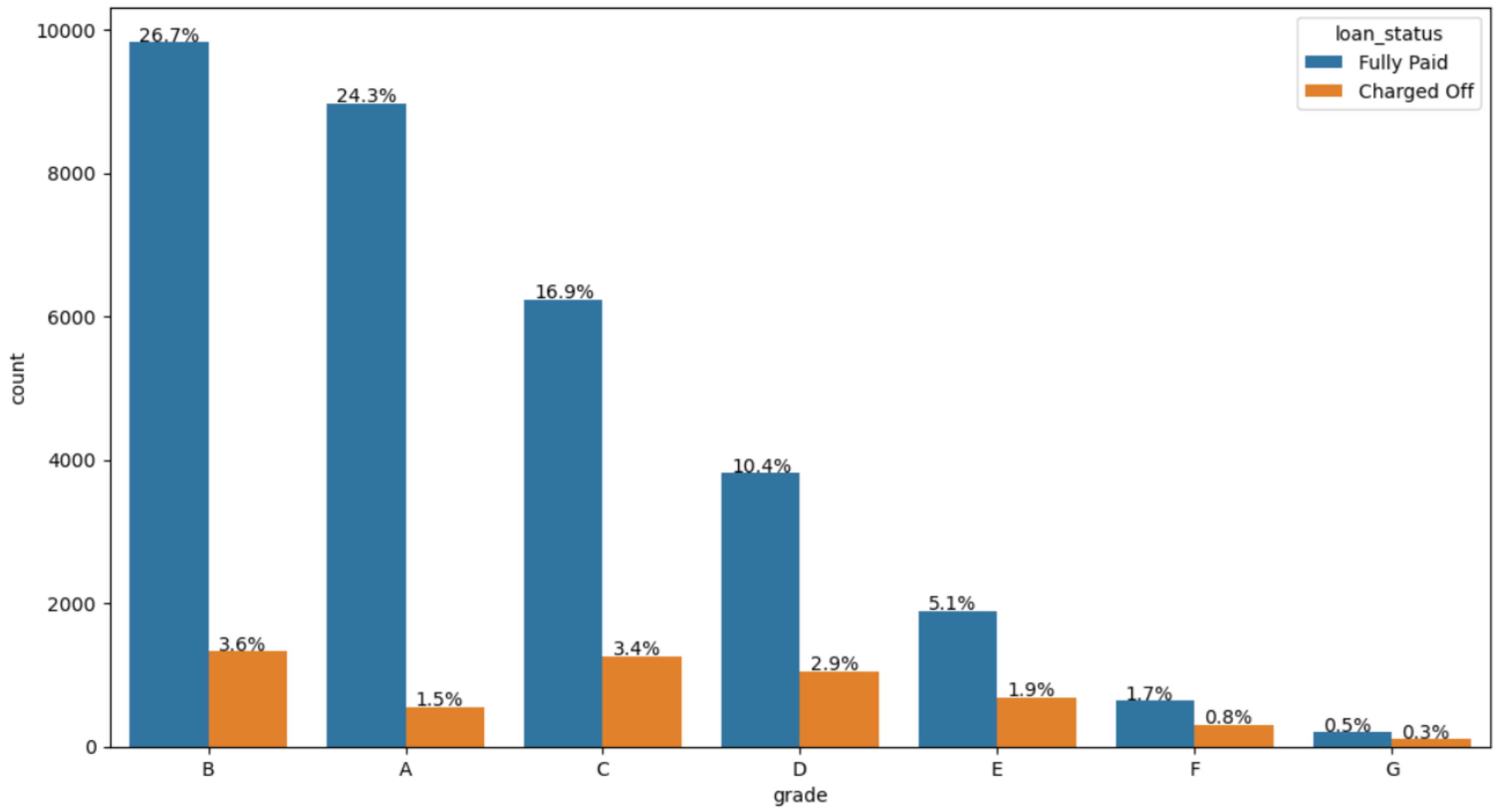
- The majority of loan applications are for a term of 36 months.
- The overall percentage of charge-offs is slightly higher for loans with a term of 36 months (8%) compared to those with a term of 60 months (6%).

Bivariate analysis



- when calculating the ratio of charge-offs within each term category, the ratio is significantly higher for loans with a term of 60 months (25%) compared to those with a term of 36 months (10%).
- Therefore, loans with a term of 60 months should be scrutinized more carefully.
- Applicants with a 60-month loan term are more likely to experience higher charge-offs.

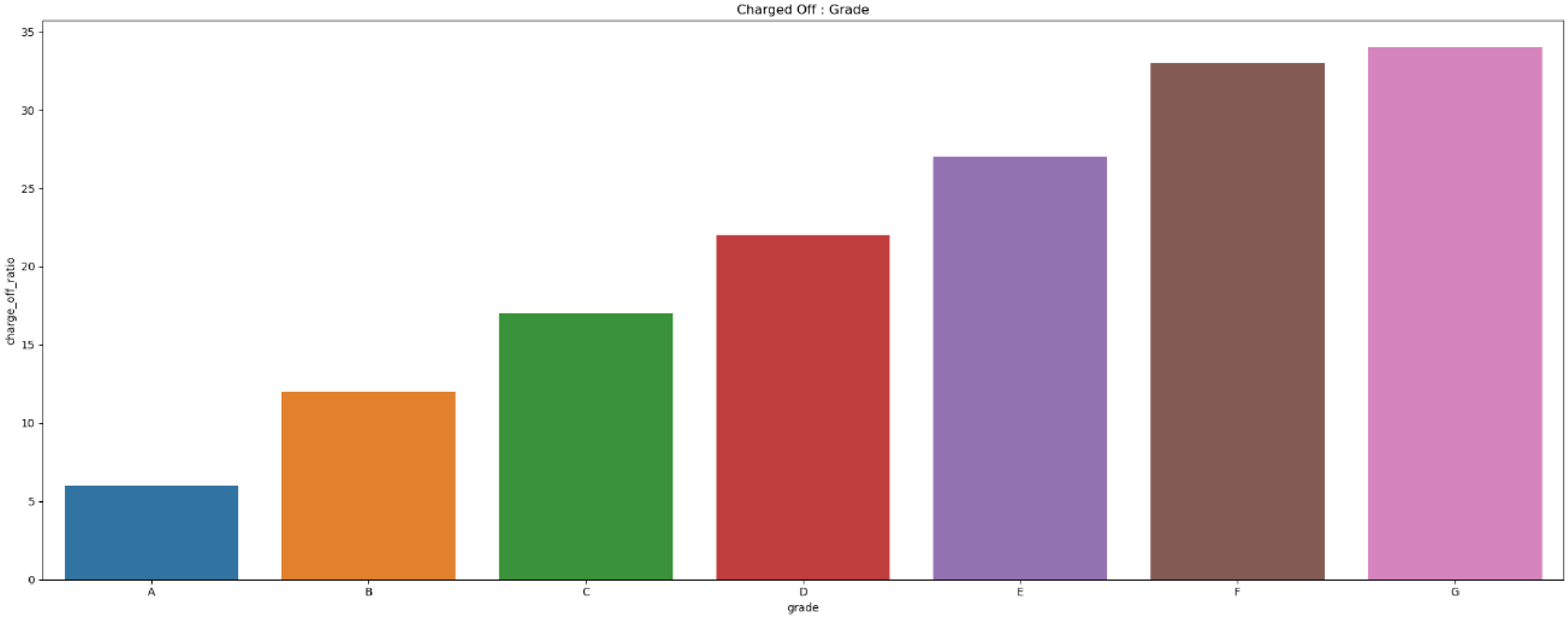
Bivariate analysis



- When analyzing the charge-off ratio within each grade, the highest percentage of charge-offs is observed in Grade G.
- The largest clusters of charge-offs are in Grades G and F, both exceeding 30%.

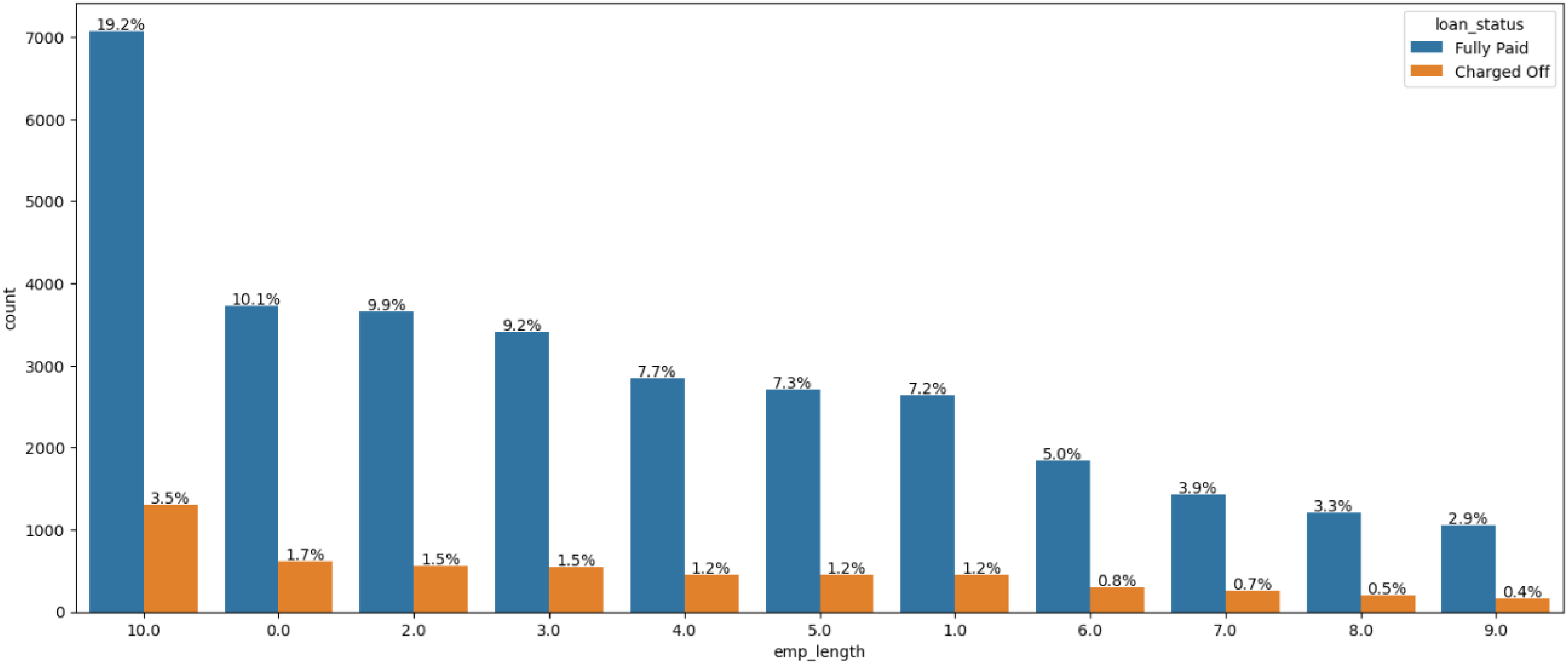
- The majority of loan volume is categorized as Grade B.
- The highest percentages of overall charge-offs are found in Grades B (3.7%) and C (3.6%).

Bivariate analysis



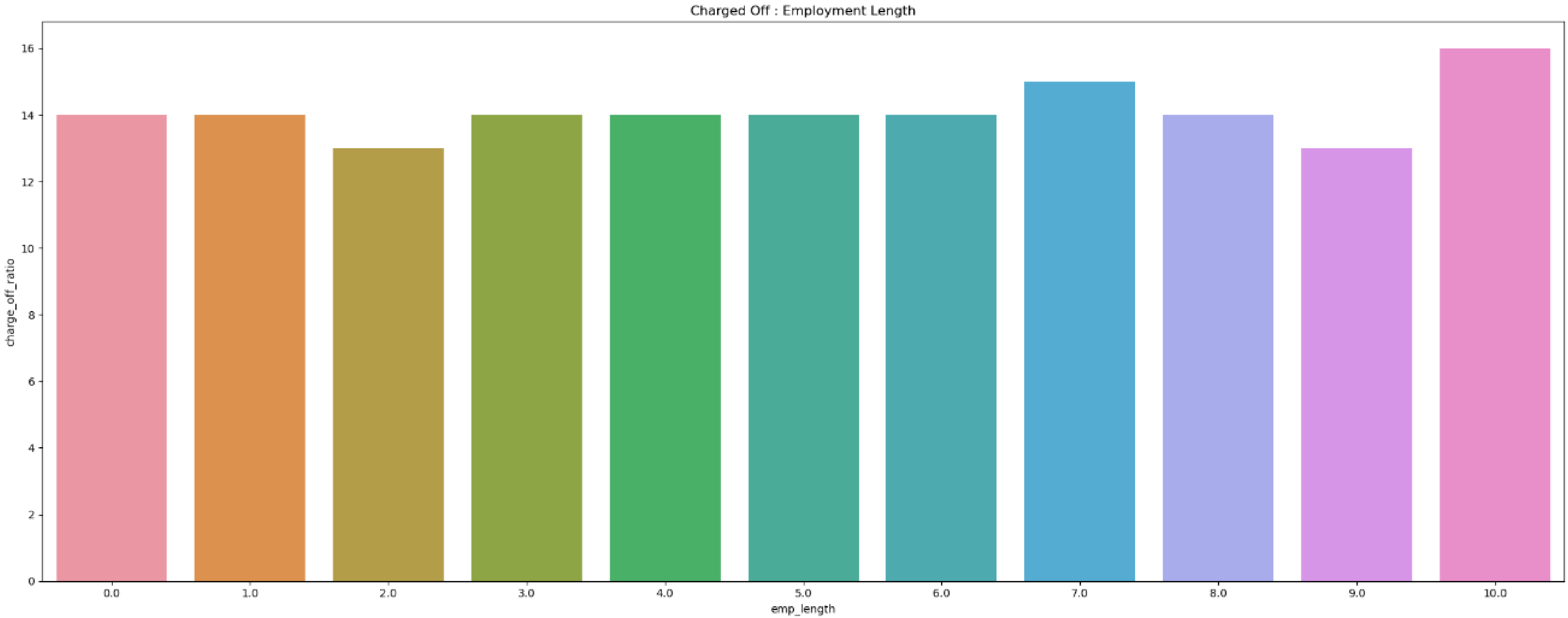
- the volume of loans in Grade G is very low (158), so it does not significantly impact the overall risk.
- The highest risk of charge-offs is associated with Grades B and C.
- Grades F and G have a very high likelihood of charge-offs, although their loan volumes are low.
- Grade A has a very low likelihood of charge-offs.
- The probability of charge-offs increases from Grade A to Grade G.

Bivariate analysis



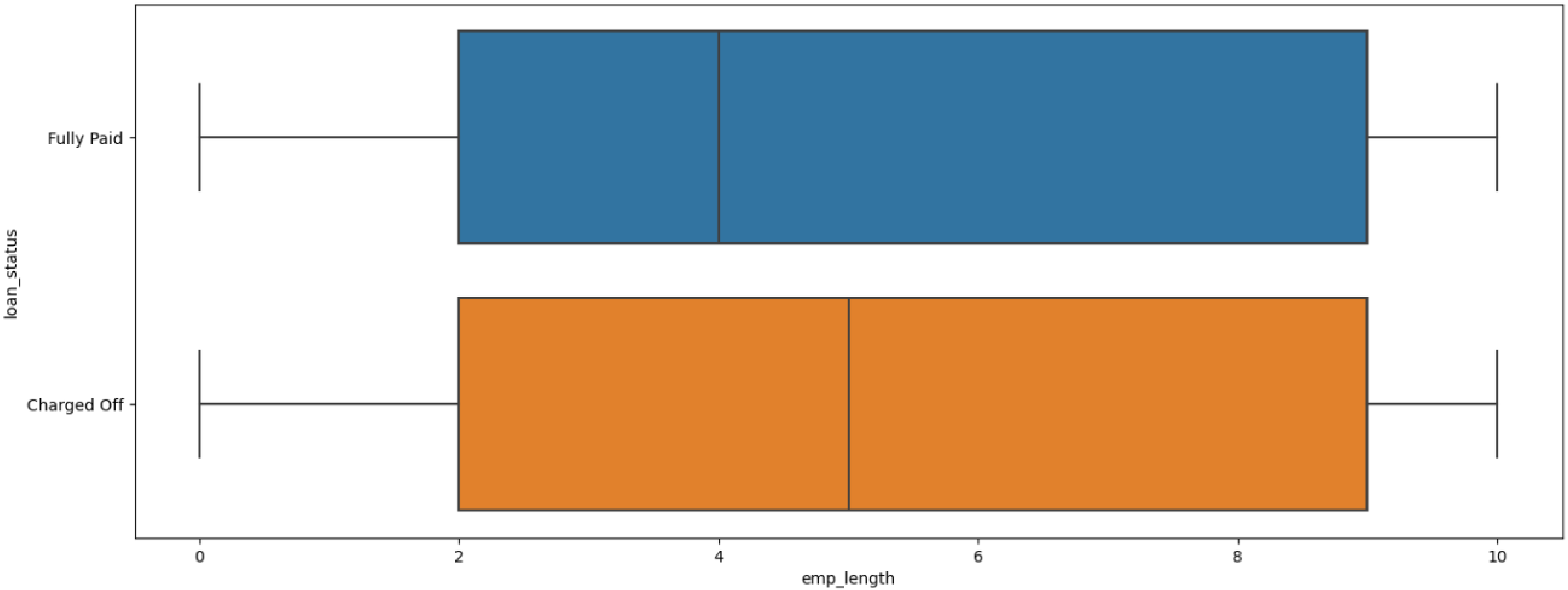
- The highest number of charge-offs is observed among employees with 10 years or more of tenure.
- The charge-off ratios within different employee length categories are similar and do not provide clear conclusions.

Bivariate analysis



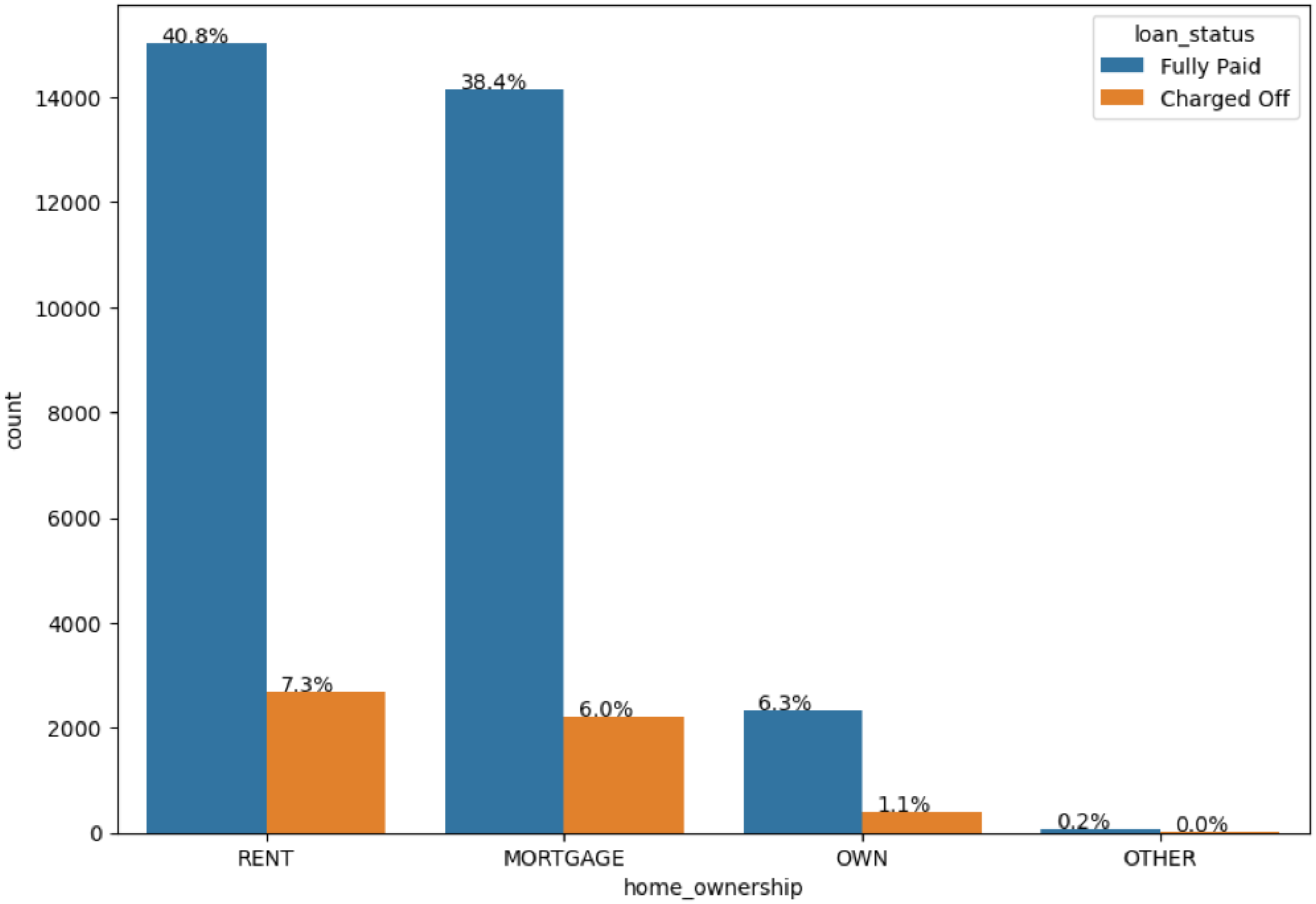
- The greatest number of charge-offs is seen among employees with 10 years or more of employment.

Bivariate analysis

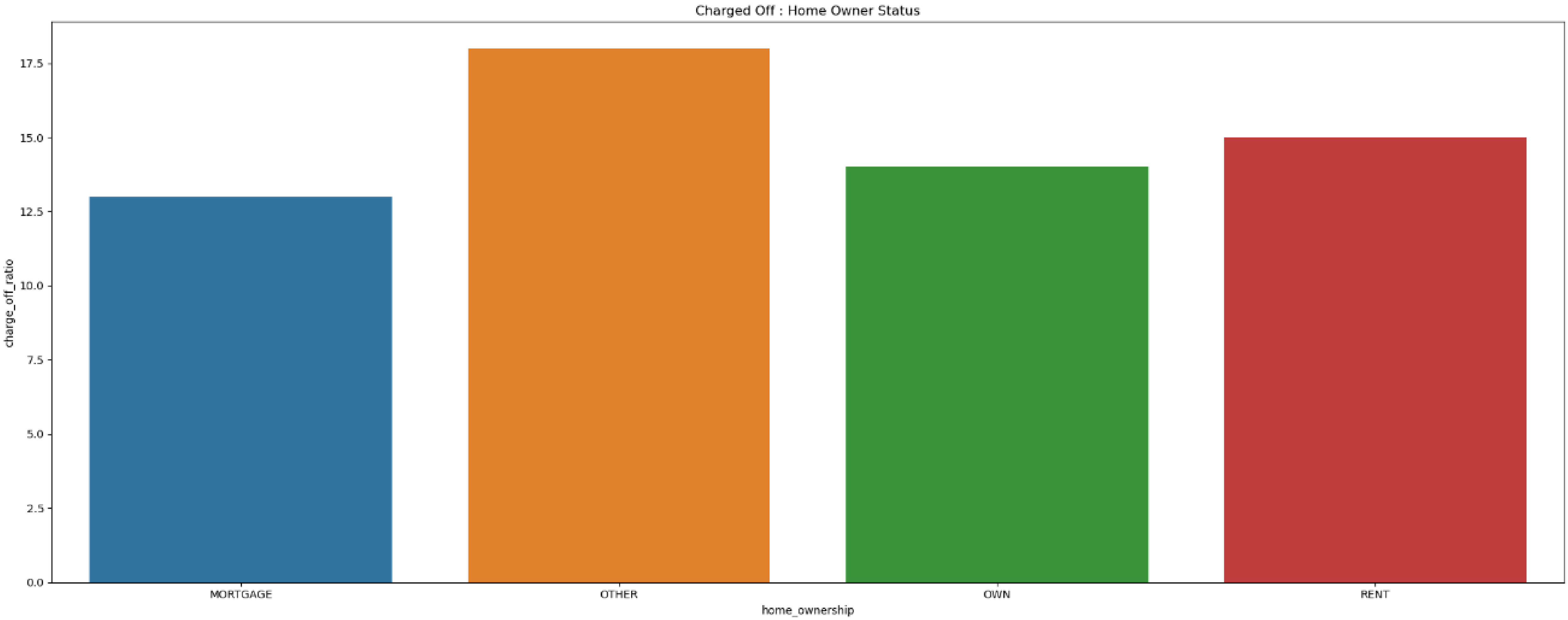


- There is a high probability of charge-offs for individuals with an income range of less than 1 year.
- The charge-off ratios within different income ranges are quite similar and inconclusive.

Bivariate analysis

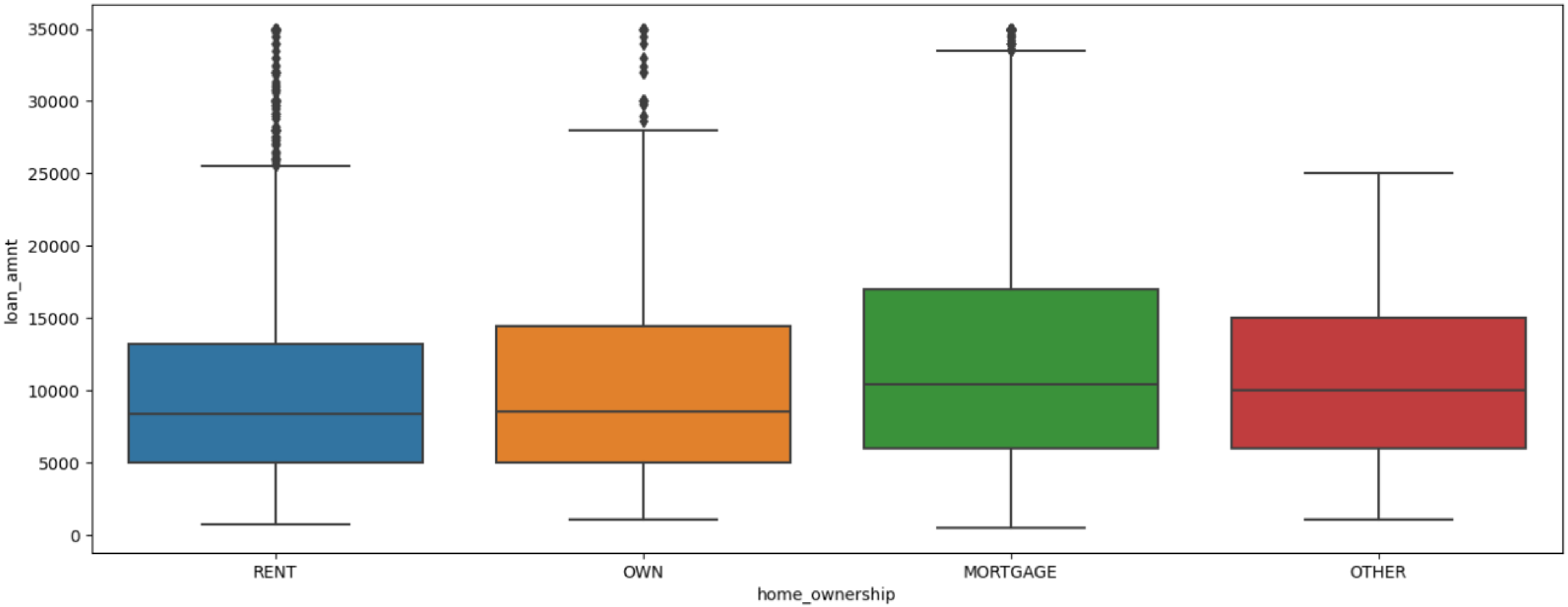


Bivariate analysis



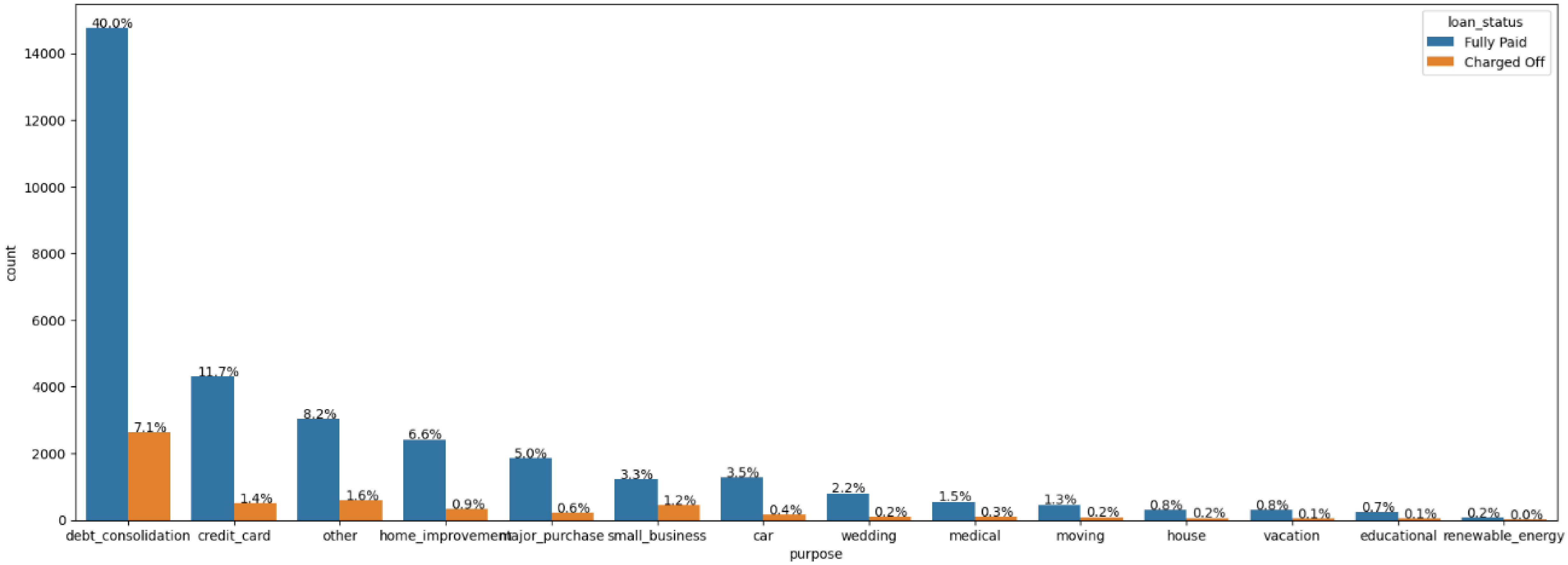
- The highest number of charge-offs is observed in the categories of RENT and MORTGAGE.
- Within each home ownership category, the charge-off ratio is highest for the "Other" category.

Bivariate analysis



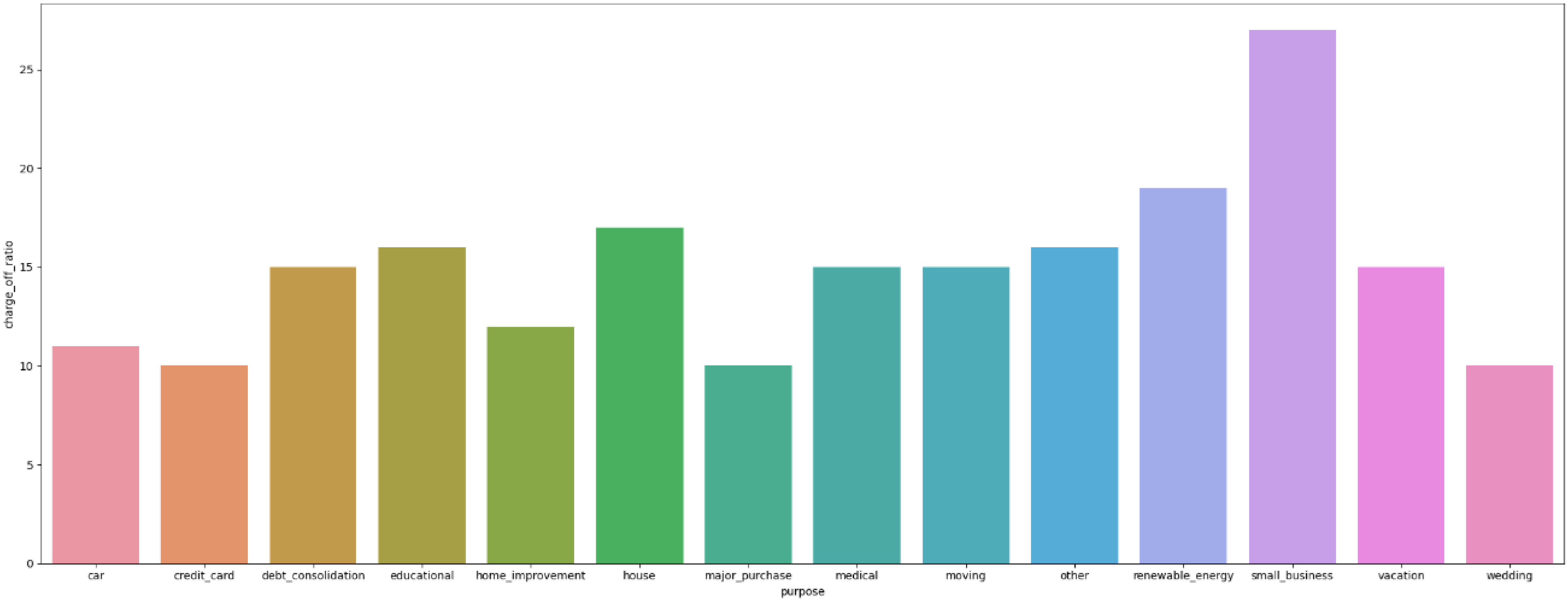
- Individuals with a home ownership status of MORTGAGE are at the highest risk for charge-offs.
- The MORTGAGE category also includes a wide range of loan amounts, which further increases the risk of charge-offs.

Bivariate analysis



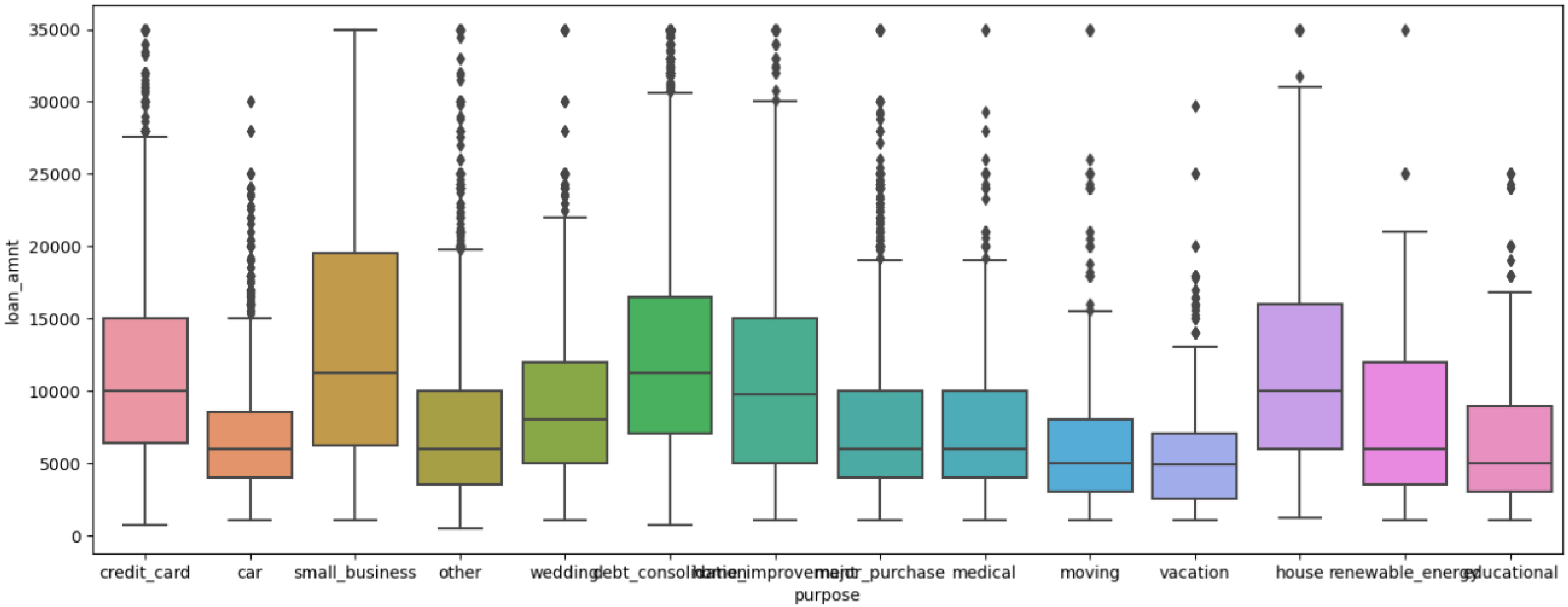
- The highest risk of charge-offs is associated with the category of debt_consolidation.
- The highest probability of charge-offs within a category is found in small_business, although the volume of such loans is extremely low.

Bivariate analysis



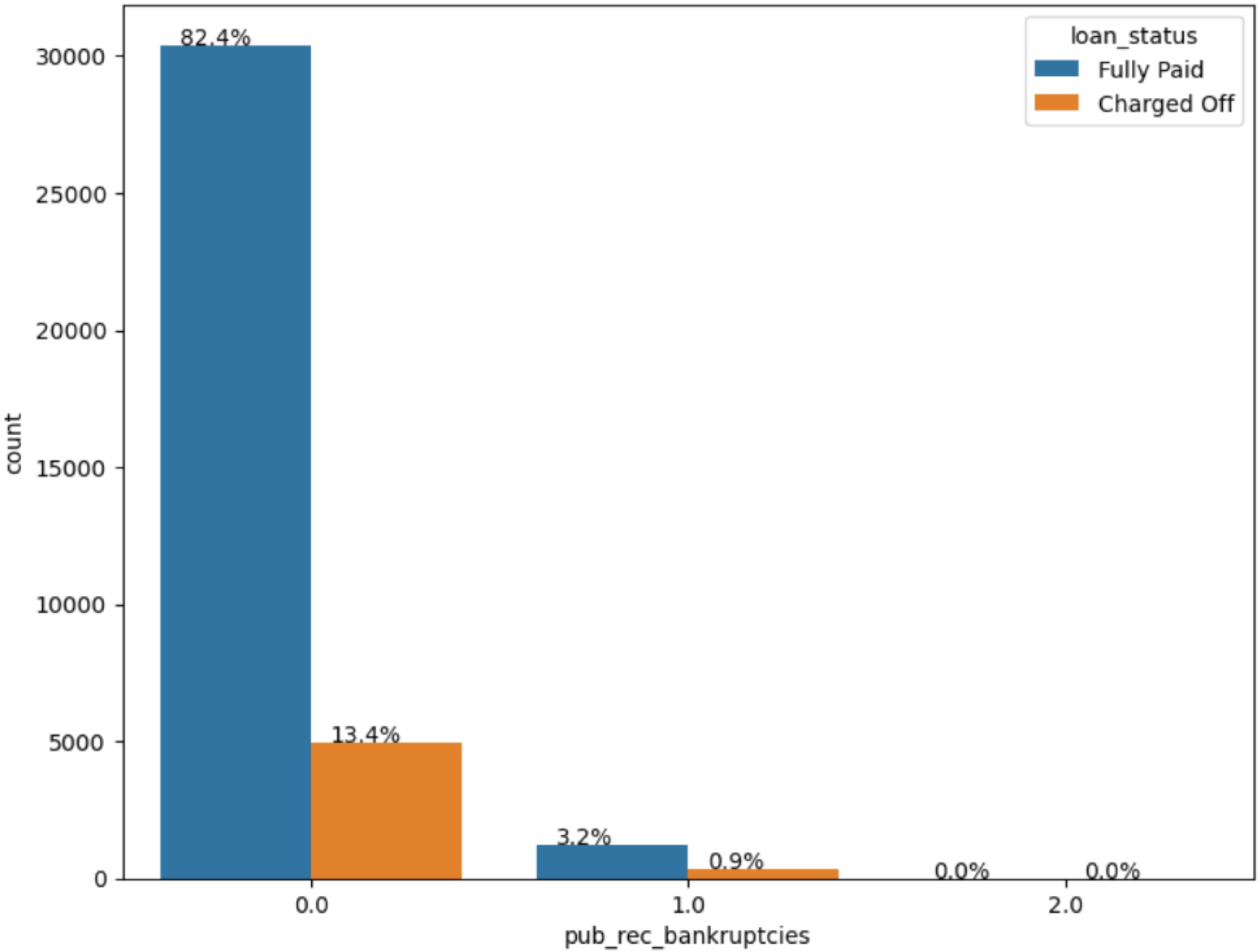
- The largest loan amounts are associated with the categories of small_business, debt_consolidation, and house.
- The greatest risk of charge-offs is linked to loans for debt_consolidation.

Bivariate analysis



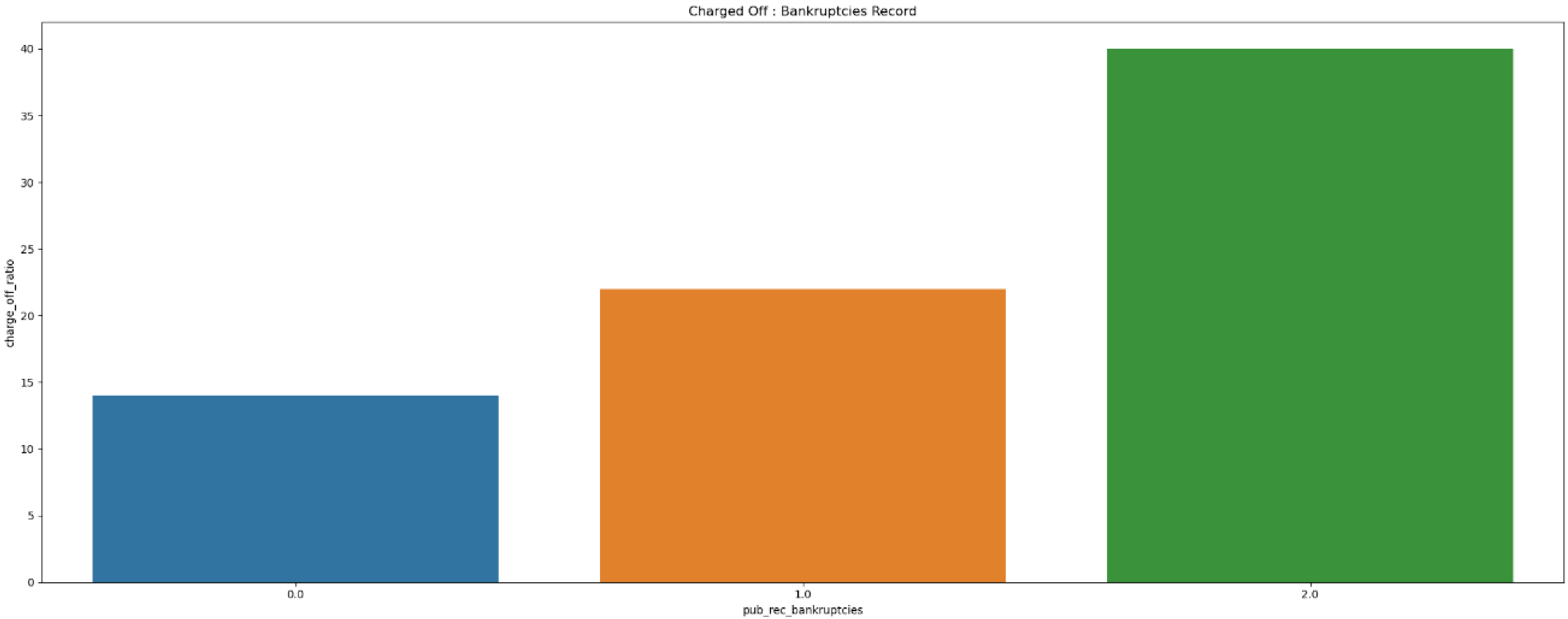
- Small Business applicants have a high likelihood of charge-offs, but their overall loan volume is low.
- Renewable_energy has the lowest risk of charge-offs relative to its loan volume.

Bivariate analysis



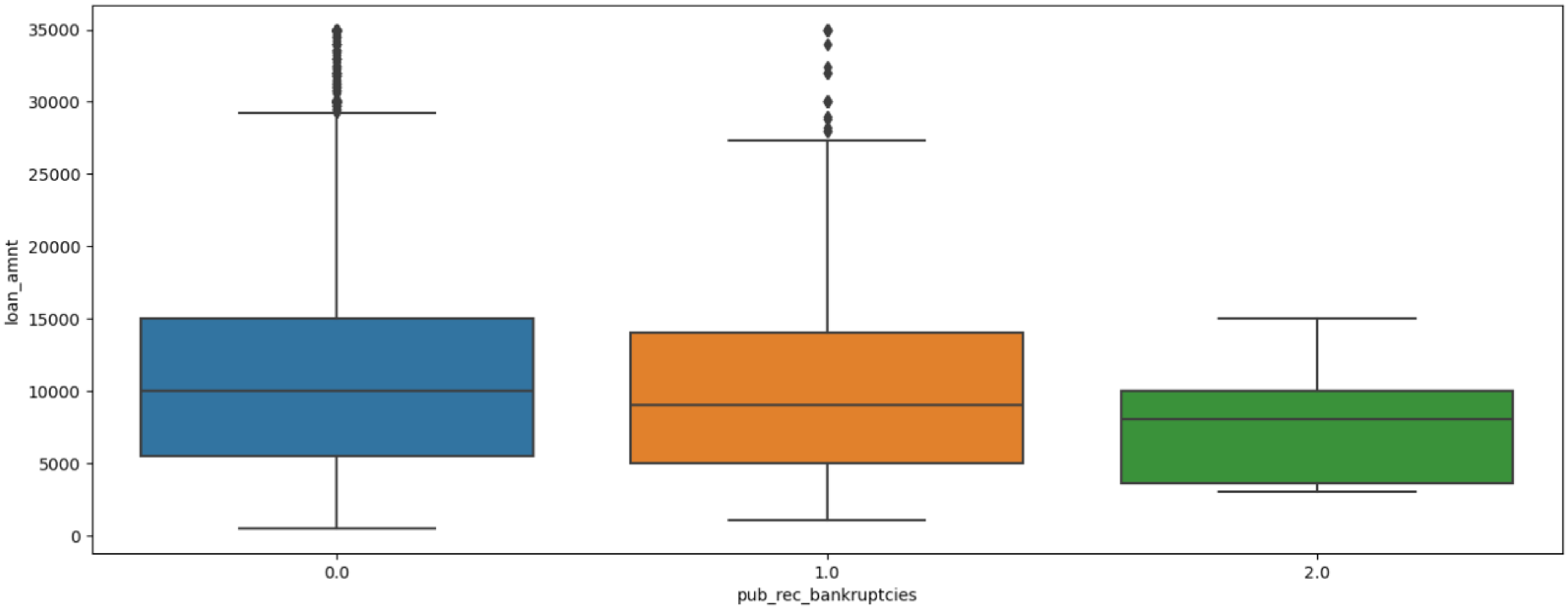
- Based purely on loan volumes, the highest number of charge-offs is observed in the category with no bankruptcy record (0).

Bivariate analysis



- When examining charge-off ratios within each category, customers with a bankruptcy record show a higher charge-off ratio.

Bivariate analysis



- Customers with a bankruptcy record are at a higher risk of charge-offs.
- Customers with a record of 2 bankruptcies have an even higher charge-off ratio.

Bivariate Analysis Inferences

- The overall percentage of charge-offs is slightly higher for loans with a term of 36 months (8%) compared to those with a term of 60 months (6%).However, when calculating the ratio of charge-offs within each term category, the ratio is significantly higher for loans with a term of 60 months (25%) compared to those with a term of 36 months (10%).
- The highest percentages of overall charge-offs are found in Grades B (3.7%) and C (3.6%). When analyzing the charge-off ratio within each grade, the highest percentage of charge-offs is observed in Grade G. However, the volume of loans in Grade G is very low (158), so it does not significantly impact the overall risk. The probability of charge-offs increases from Grade A to Grade G

Bivariate Analysis Inferences

- Small business loans have the highest chance of charge-offs but represent a small portion of total loans. Debt consolidation loans carry the greatest risk of charge-offs. Renewable energy loans have the lowest risk relative to their volume.
- California has the most loans and the highest number of charge-offs due to its volume. Nebraska has a high probability of charge-offs but low loan volume, while Nevada, California, and Florida also show high charge-off rates.
- Borrowers with a history of bankruptcy, especially those with two bankruptcies, are at a higher risk of defaulting on loans.
- Debt consolidation loans have the highest risk of charge-offs, while small business loans have a high likelihood but low volume. Renewable energy loans carry the lowest risk.

Bivariate Analysis Inferences

- Loans with a 60-month term have a higher charge-off ratio (25%) compared to 36-month loans (10%), so they require more scrutiny. Grades B and C carry the highest overall charge-off risk, while Grades F and G have the highest charge-off ratios but low volumes. The risk of charge-offs increases from Grade A to Grade G.
- Charge-offs are most frequent among employees with 10 or more years of tenure, although employee tenure and income levels do not clearly predict charge-offs. The greatest charge-offs are in RENT and MORTGAGE categories, with MORTGAGE loans being riskier due to their larger amounts.
- California leads in loan volume and charge-offs, while Nevada and Nebraska have high charge-off risks, though Nebraska's low volume lessens its overall impact. Borrowers with a bankruptcy history, especially those with two bankruptcies, are more likely to default.

Thank you…