# LendingClub Case Study

Analyzing Loan dataset and Default Patterns

Basavaraj G
Manasa Devadas

# Problem Statement

**Objective:** The goal of this case study is to analyze Lending Club's loan data to understand the factors influencing loan defaults and to develop strategies to improve loan performance and mitigate risk.

**Problem Statement:** Lending Club seeks to optimize its loan approval process and minimize the risk of loan defaults. Despite comprehensive credit assessments, the company experiences a significant rate of loan defaults, impacting profitability and investor confidence.

## Expected Outcomes:

1. Identification of key predictors of loan defaults.
2. Insights into borrower behaviors and loan characteristics that influence repayment.
3. Recommendations for improving loan approval processes and reducing default rates.
4. Enhanced strategies for investor risk mitigation and portfolio management.

By addressing these challenges and leveraging data-driven insights, Lending Club aims to enhance its operational efficiency, improve loan performance, and build stronger investor trust.

# Assumptions

- The loan_status  is the dependent variable which accurately categorizes loans into distinct states such as "Charged Off," "Fully Paid," and "Current", with a clear distinction between default and non-default outcomes.
- Currency is dollar in all the money related variables.
- The dataset is representative of the overall population of Lending Club borrowers, allowing for generalizable conclusions.
- The data collected over different time periods is consistent in terms of definitions, measurement units, and data collection methods.

# Approach

**Data Import**

- Import relevant datasets

**Data Cleaning**

- Remove duplicate records.
- Filter out current customer data, retaining only charged-off and fully paid loan information
- Identify and treat missing values appropriately

**Outlier Treatment**

- Detect and handle outliers using the Interquartile Range (IQR) method

**Feature Engineering**

- Create new variables from existing ones to enhance exploratory data analysis

**Data Type Verification**

- Ensure all variables have appropriate data types, converting where necessary

**Validation and Correction**

- Identify and rectify invalid data rows

# Approach Continued

**Variable Selection**

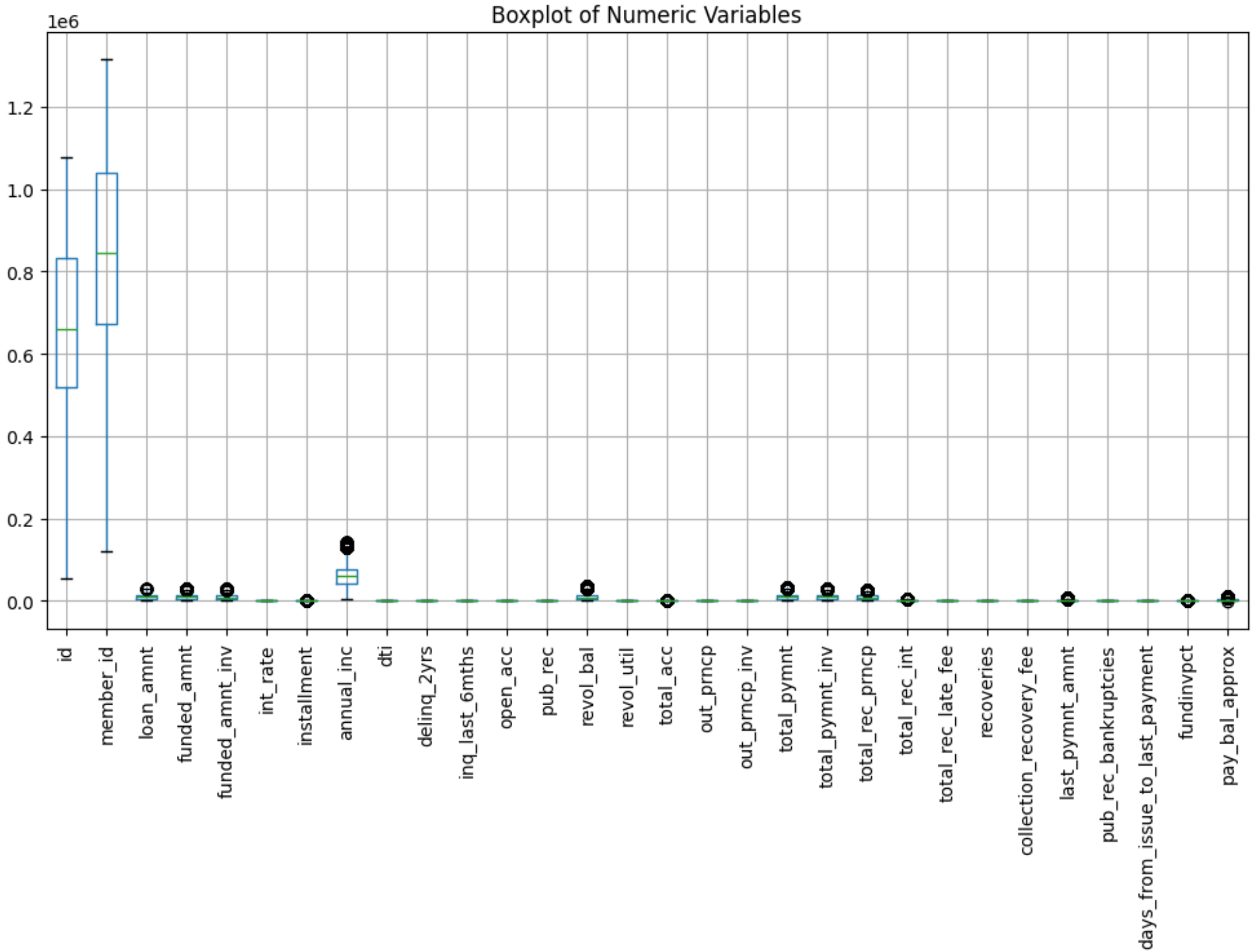
- Remove variables that do not contribute value to the analysis

**Exploratory Data Analysis (EDA)**

- Conduct univariate, segmented univariate, bivariate, and multivariate analysis

**Recommendations and Conclusions**

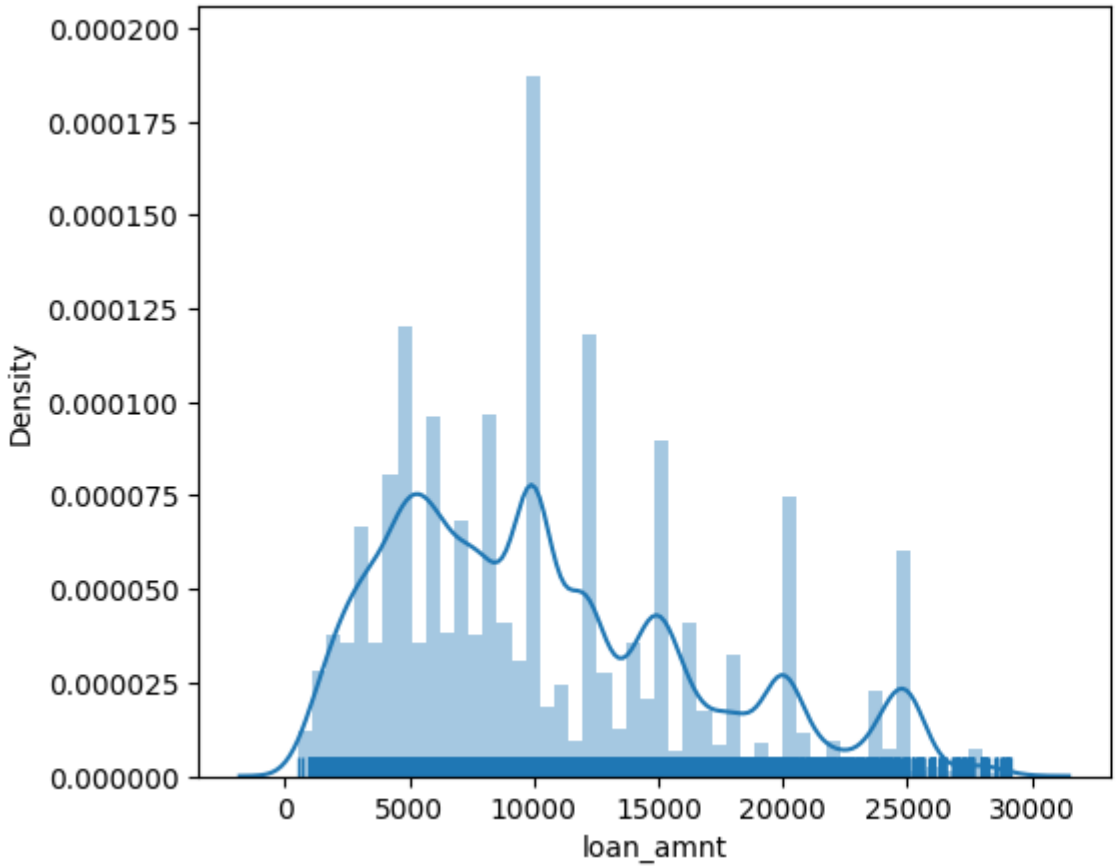
- Derive insights and formulate recommendations and conclusions based on the analysis

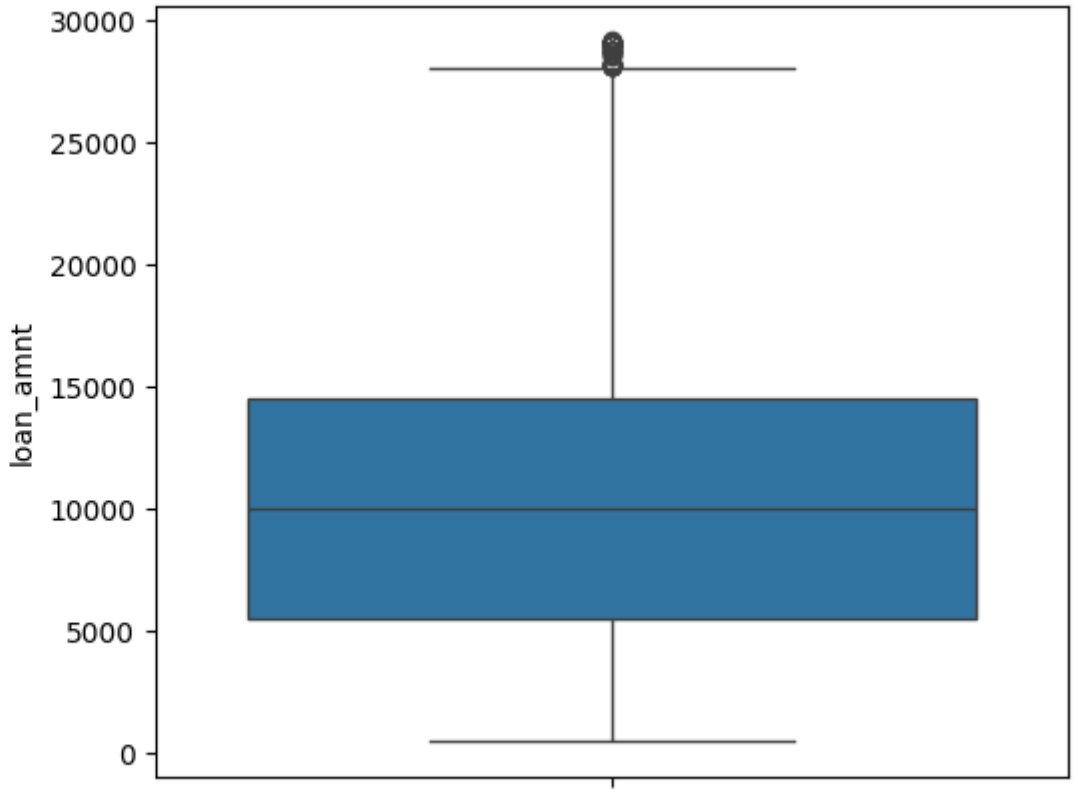



Outliers are identified in 14 numeric variables

# Graph Continuous Variables

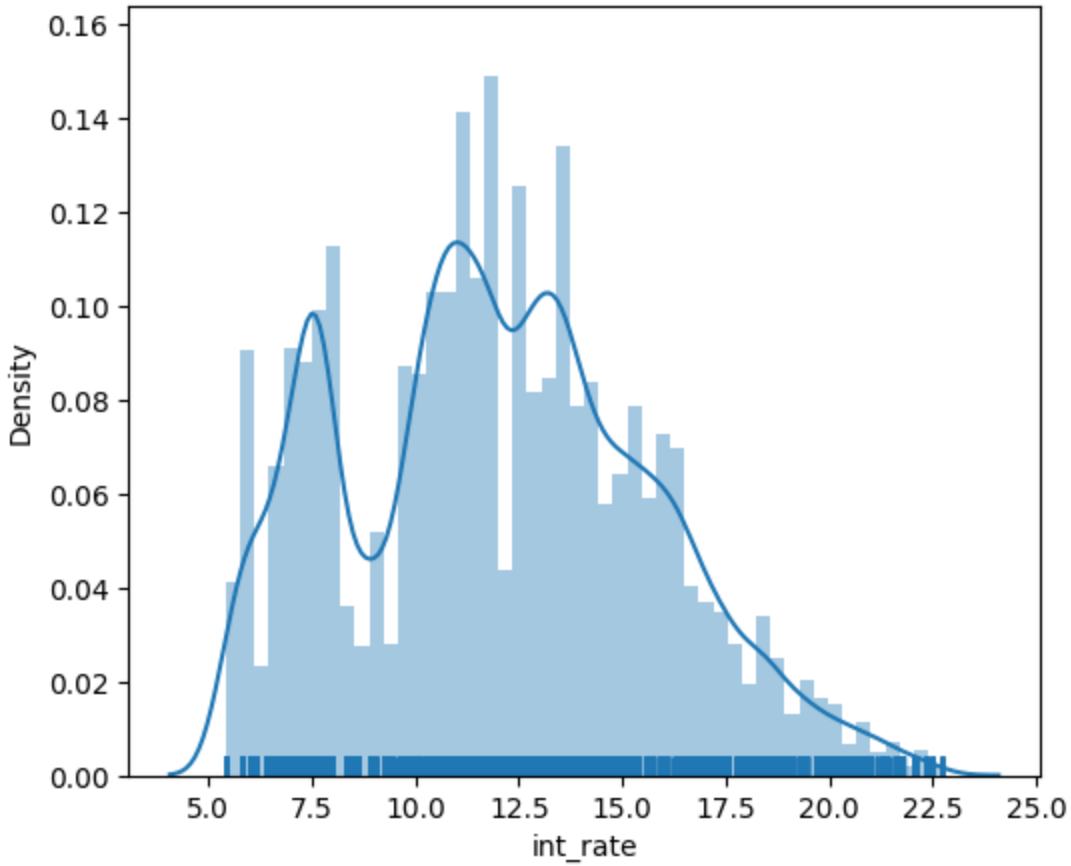
- Loan_amnt Variable

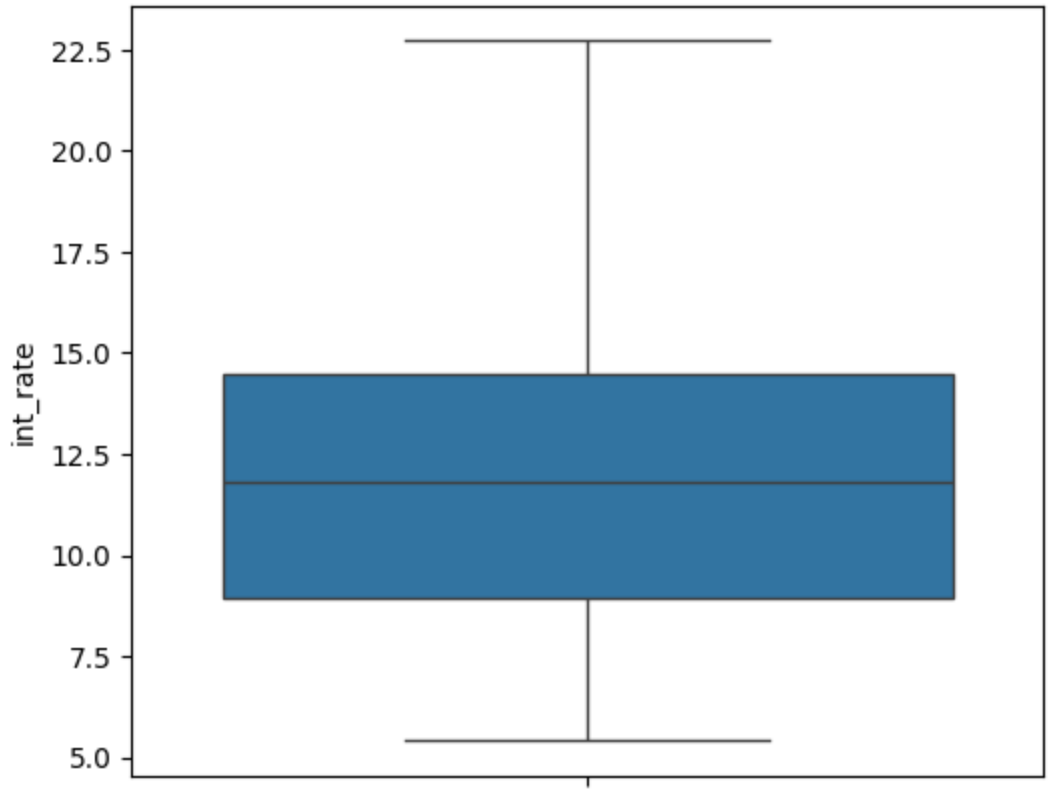



- Average Loan Amount Issued of all the loan application is 10475$

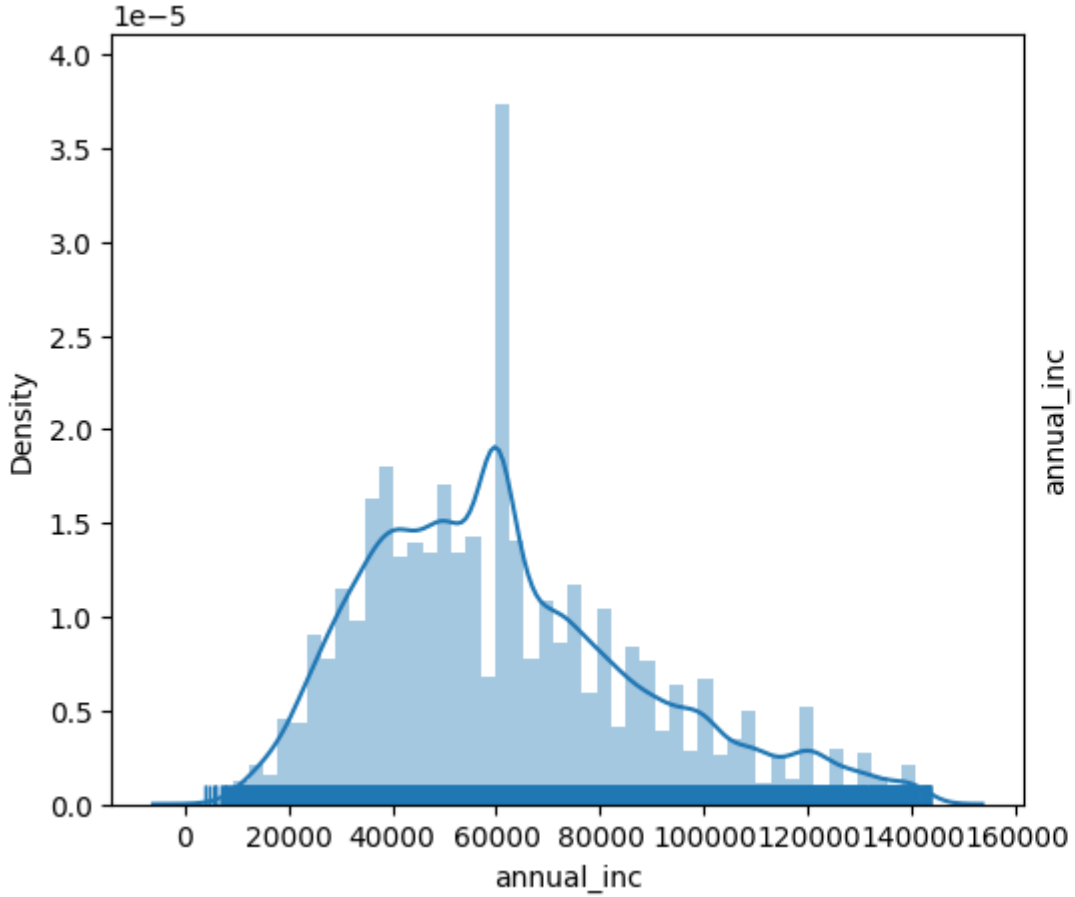
•Int_rate Variable

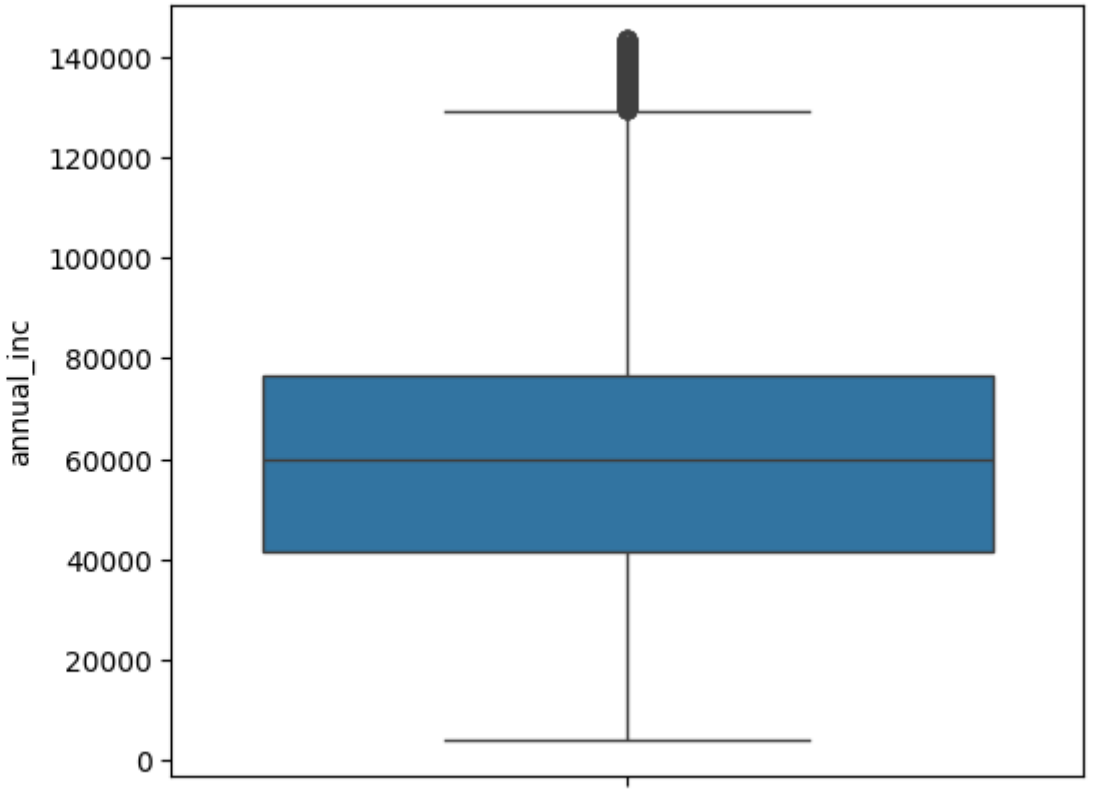



•Average Interest Rate of Loan Account is 11.96%

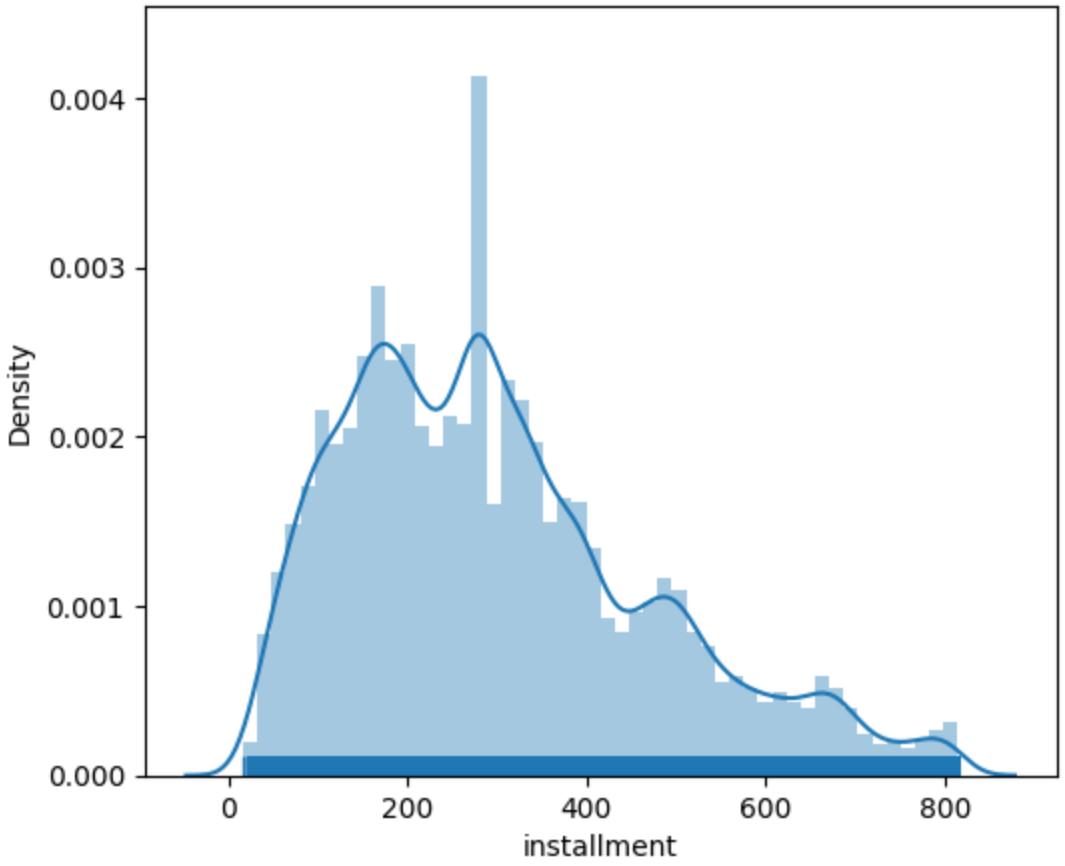
- annual_inc Variable

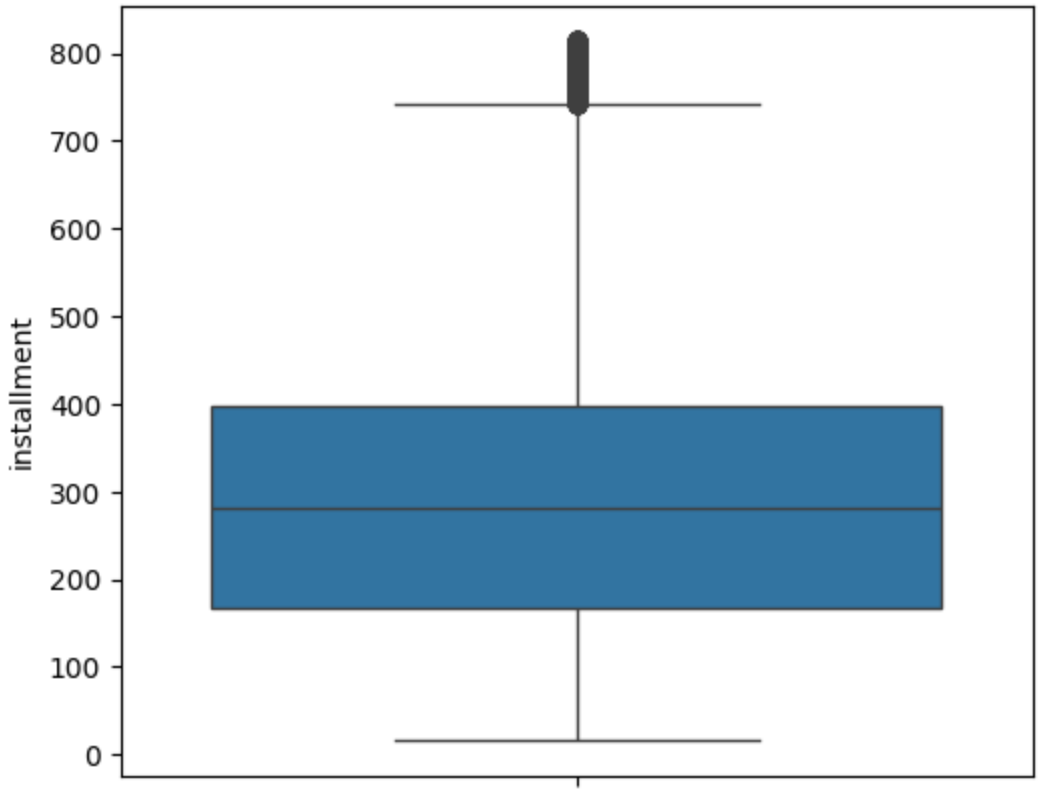



- Average Annual Income of loan customers is 61571$
- Majority of loan customers have salary between 40k-80k $

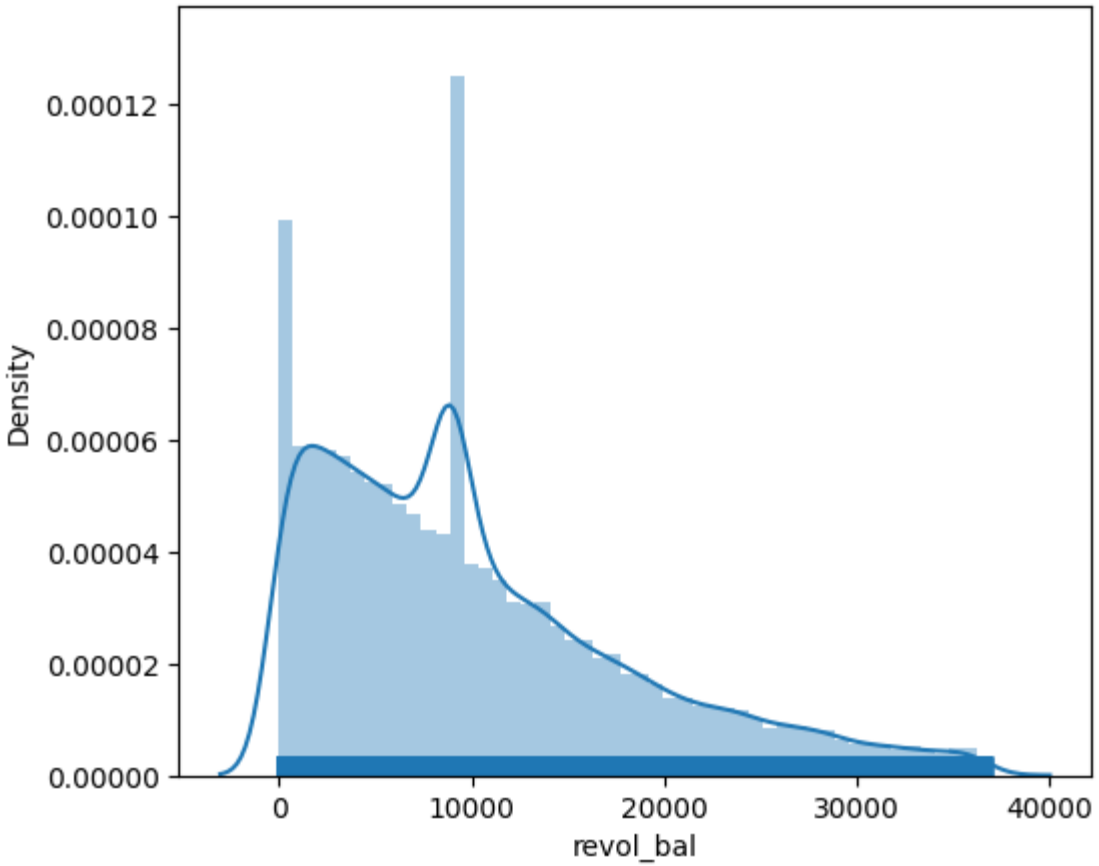
• installment variable

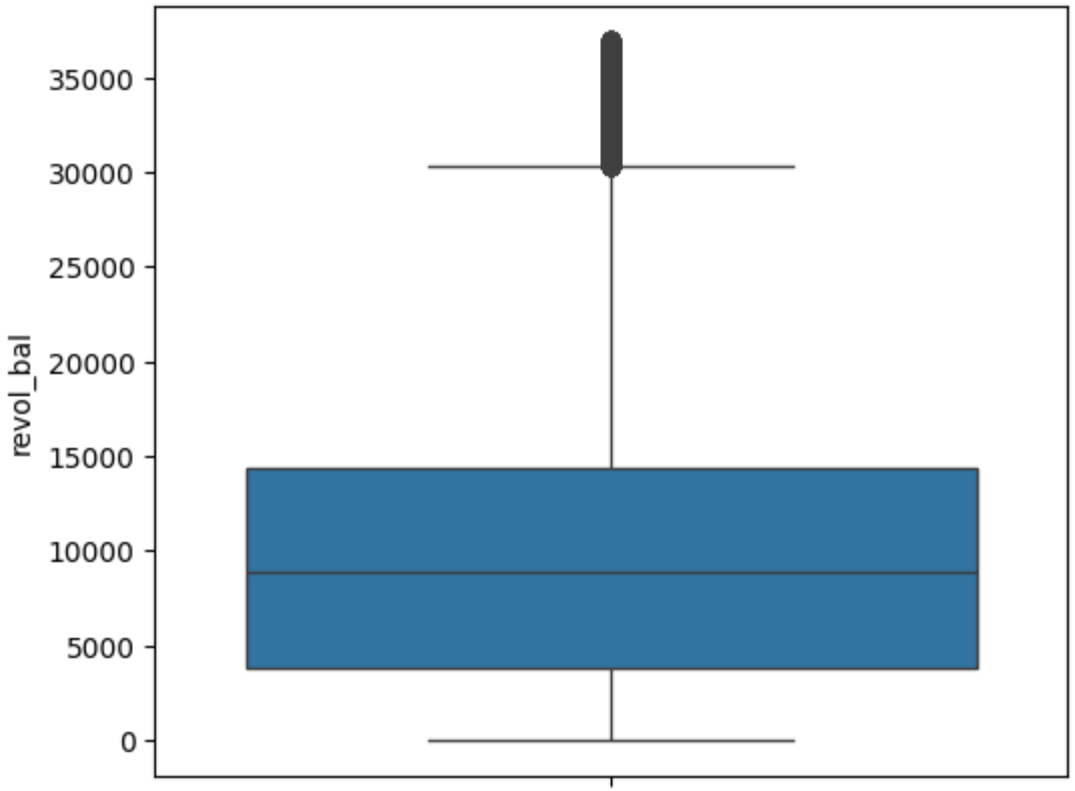



• Most of customers have installments between 100-300$

- Revol_Bal Variable



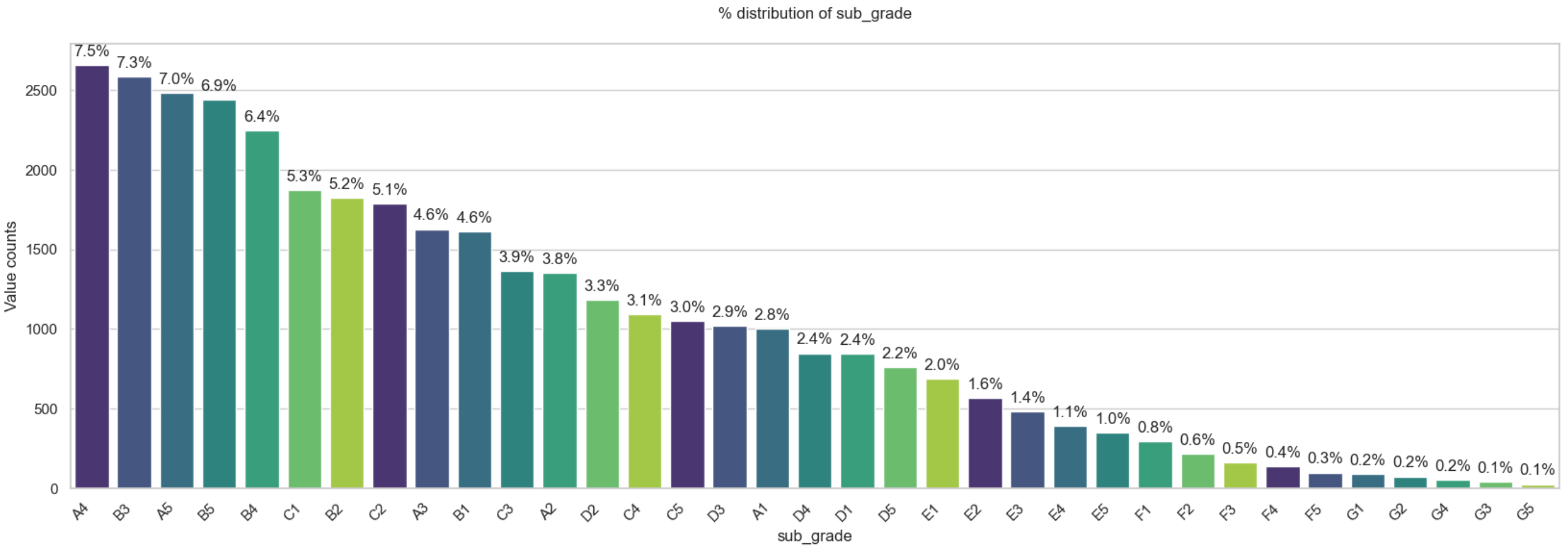




- Revol_Bal distribution is very much skewed towards left

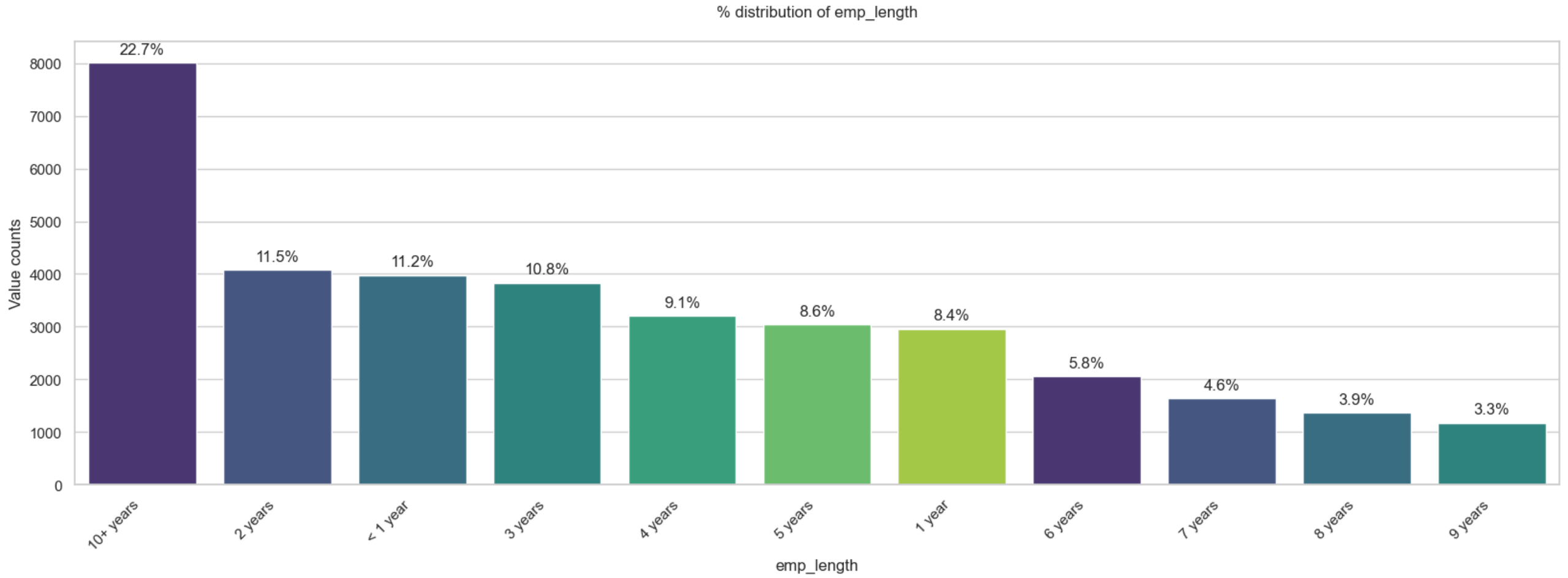
# Results

## Inferences From Descriptive Analysis of Continuous Variables**

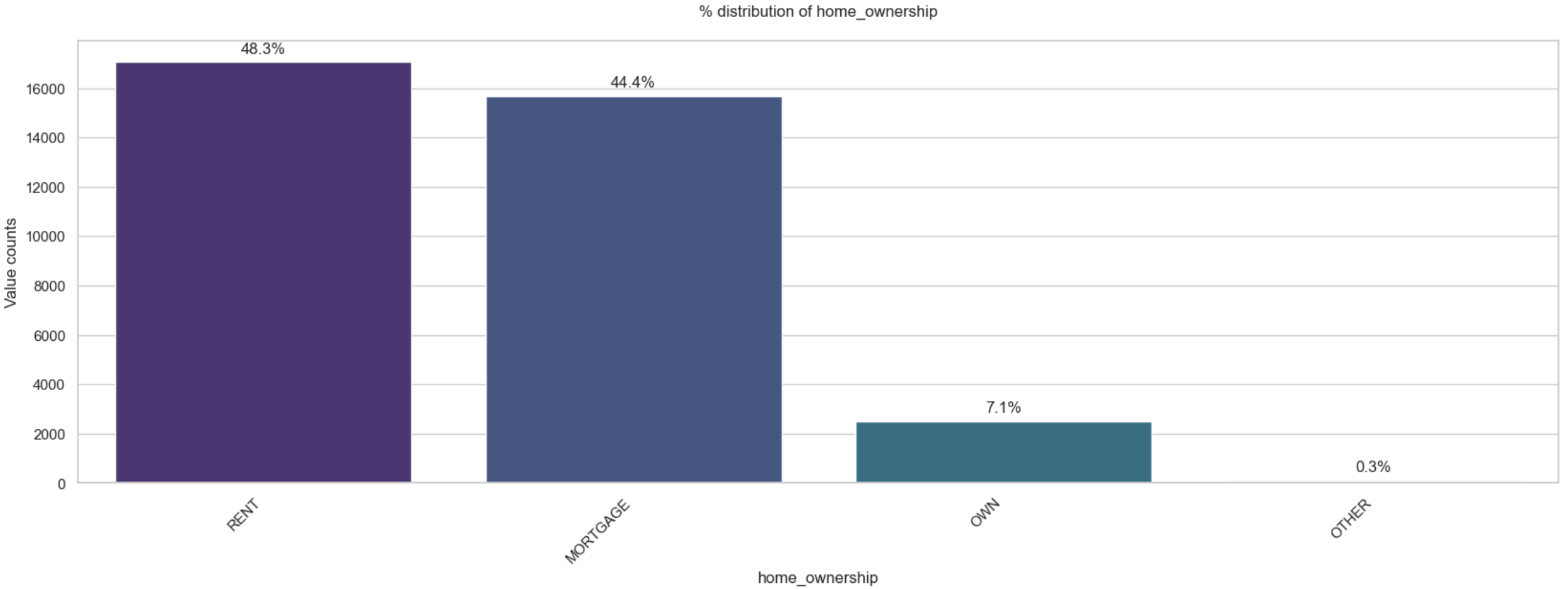
- Average Loan Amount Issued of all the loan application is 10475$
- Average Interest Rate of Loan Account is 11.96%
- Average Annual Income of loan customers is 61571$
- Majority of loan customers have salary between 40k-80k $
- Most of customers have installments between 100-300$
- Revol_Bal distribution is very much skewed towards left



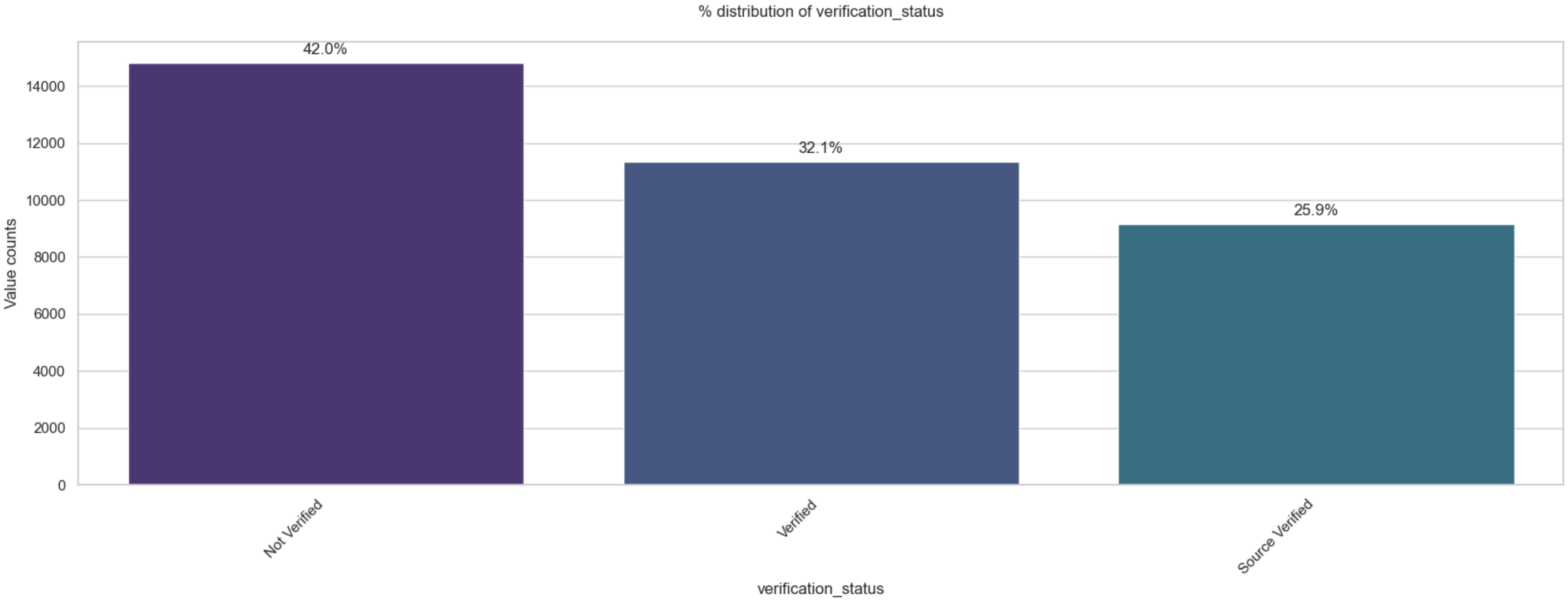

- Trend is very clear, Majority of the loans are approved for grade A, B & C



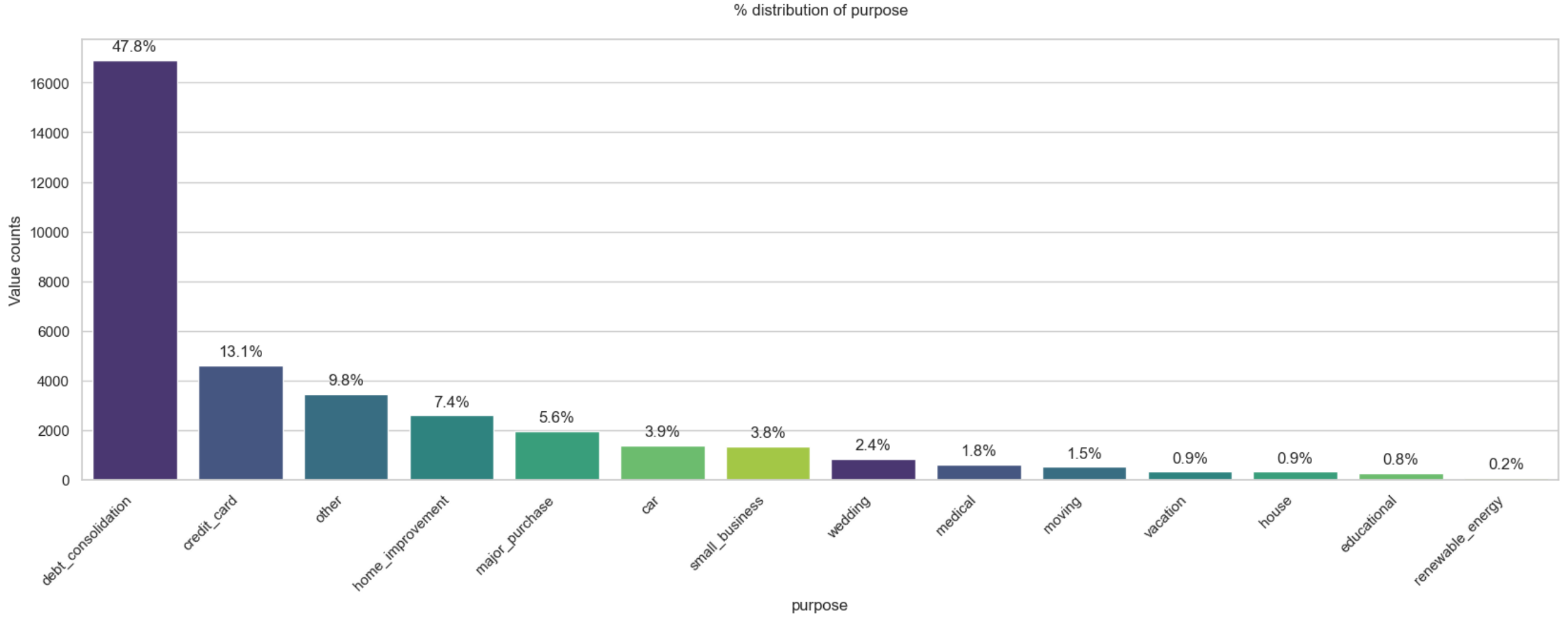

- 22.7% of loan is granted for the customers who has 10+ years of experience



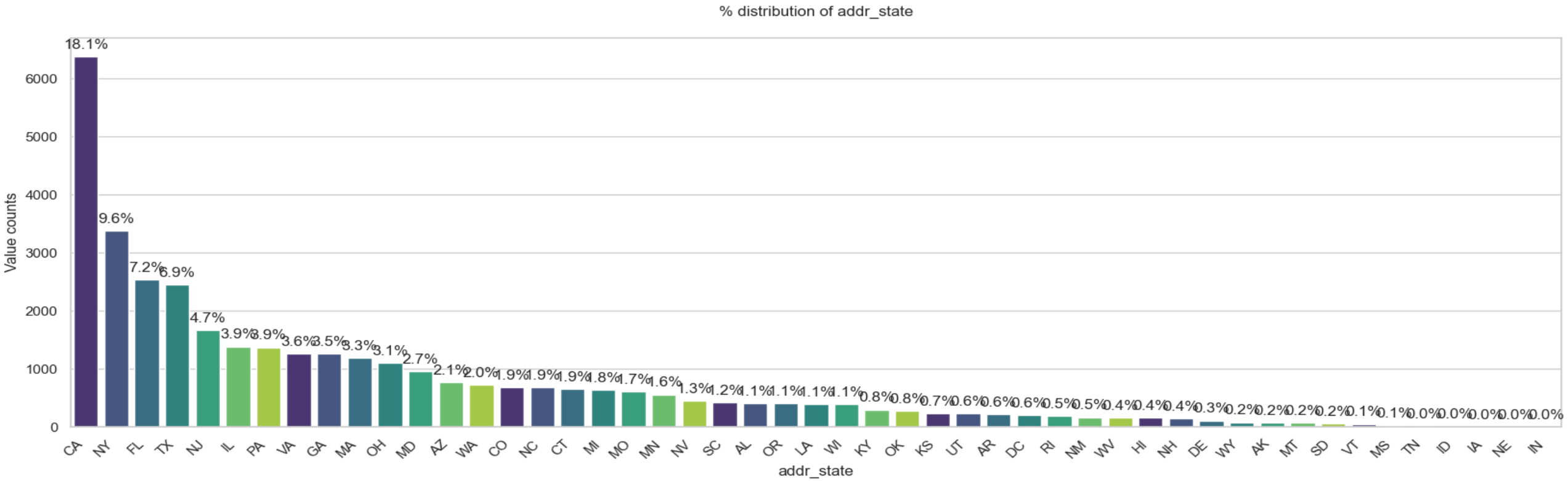

•Customers who are either staying in rented house or having mortgage loan contributes to 91% of loans.



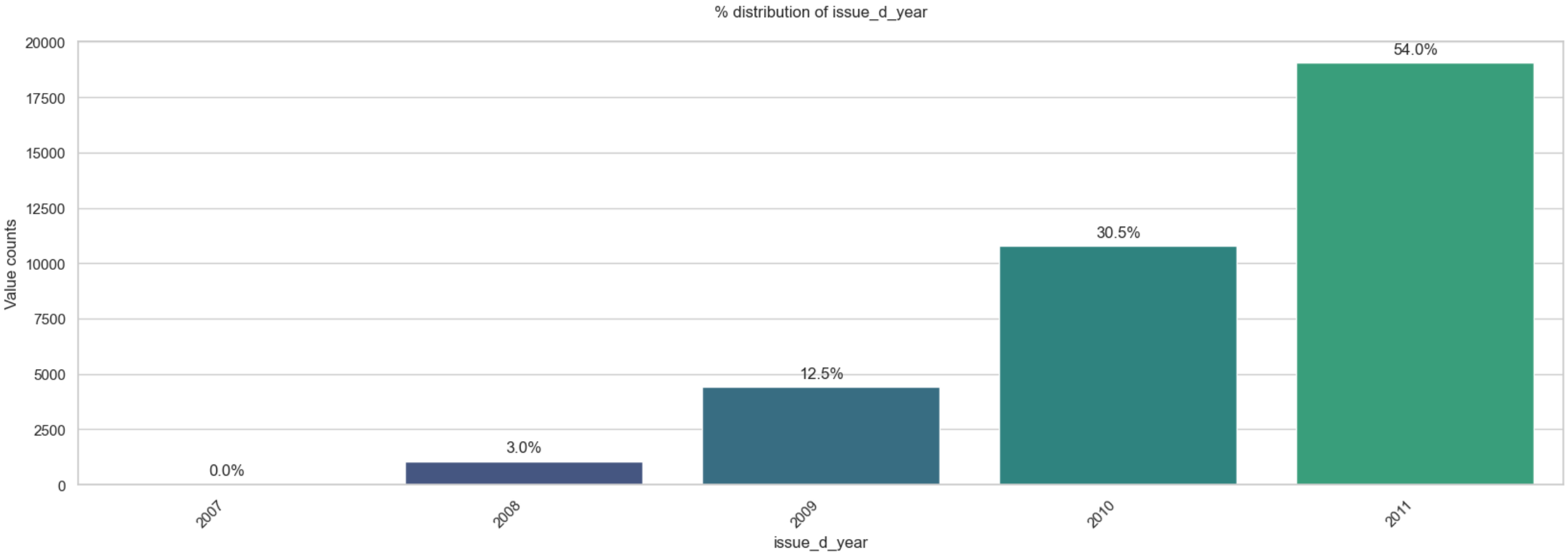

•Surprisingly 42% of loan granted to the customers whose income is not verified.



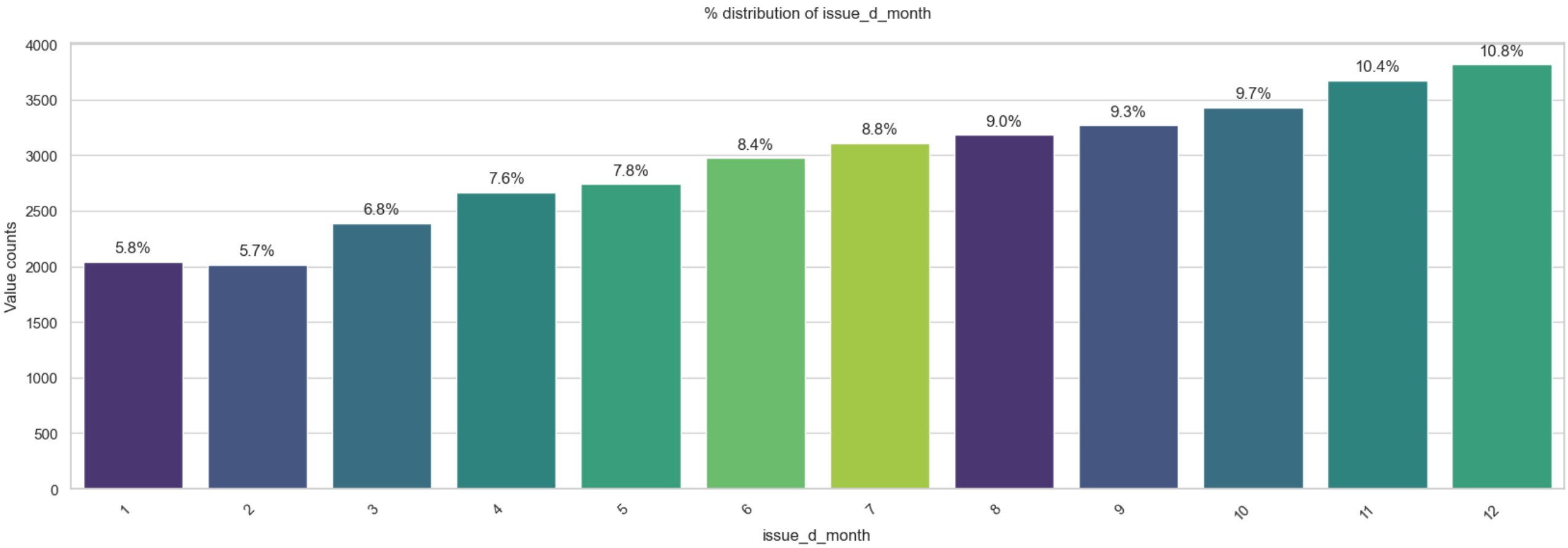

•47% of the loan taken by the customers to consolidate their debts




- 41% of the loan taken by the customers living in CALIFORNIA(CA),NEW YORK(NY),FLORIDA(FL) & (TEXAS)TX




•54 % of the loans are issued in the year 2011



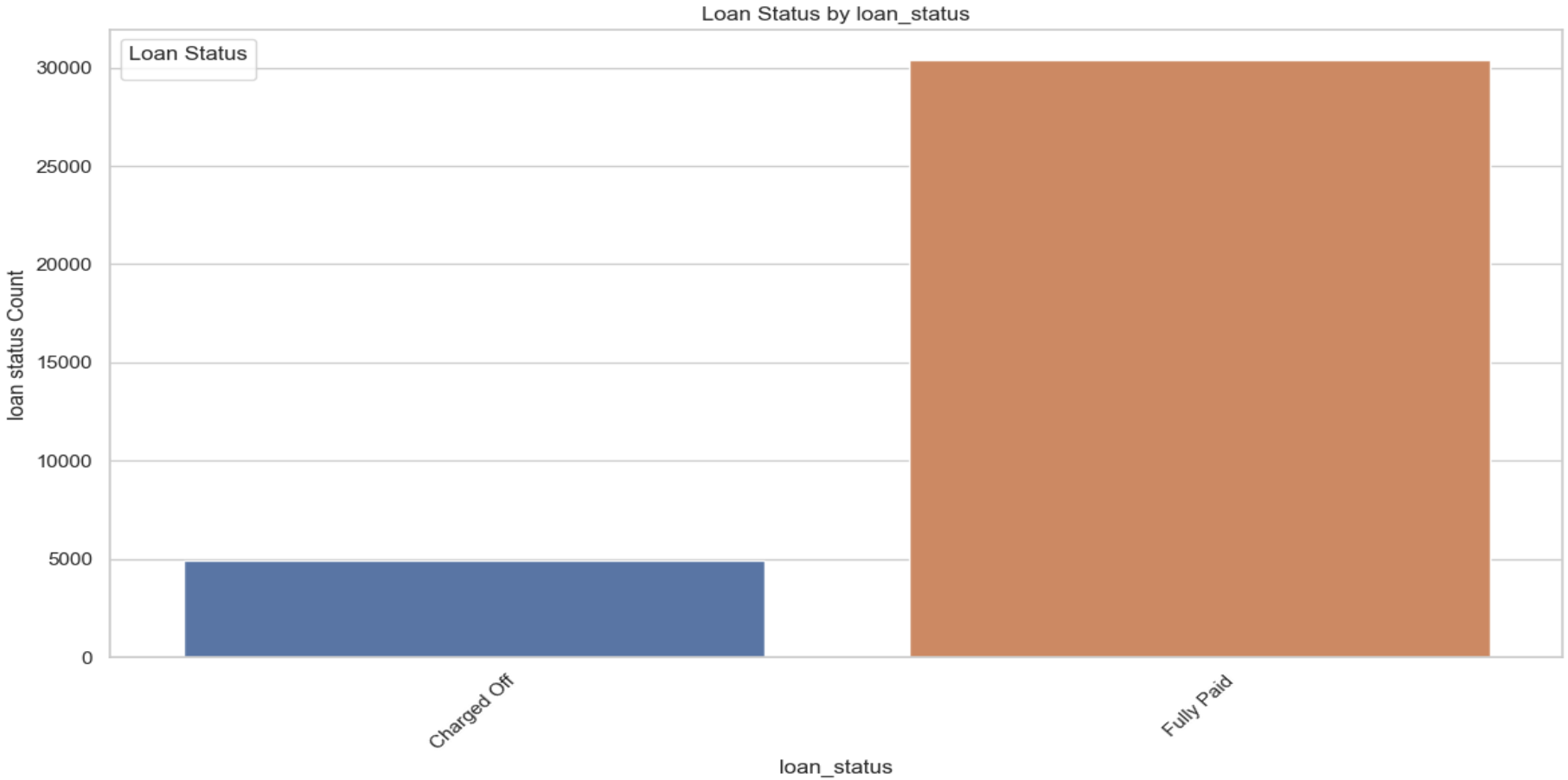

•Loans issued grows steadily over the year, most in December

# Results

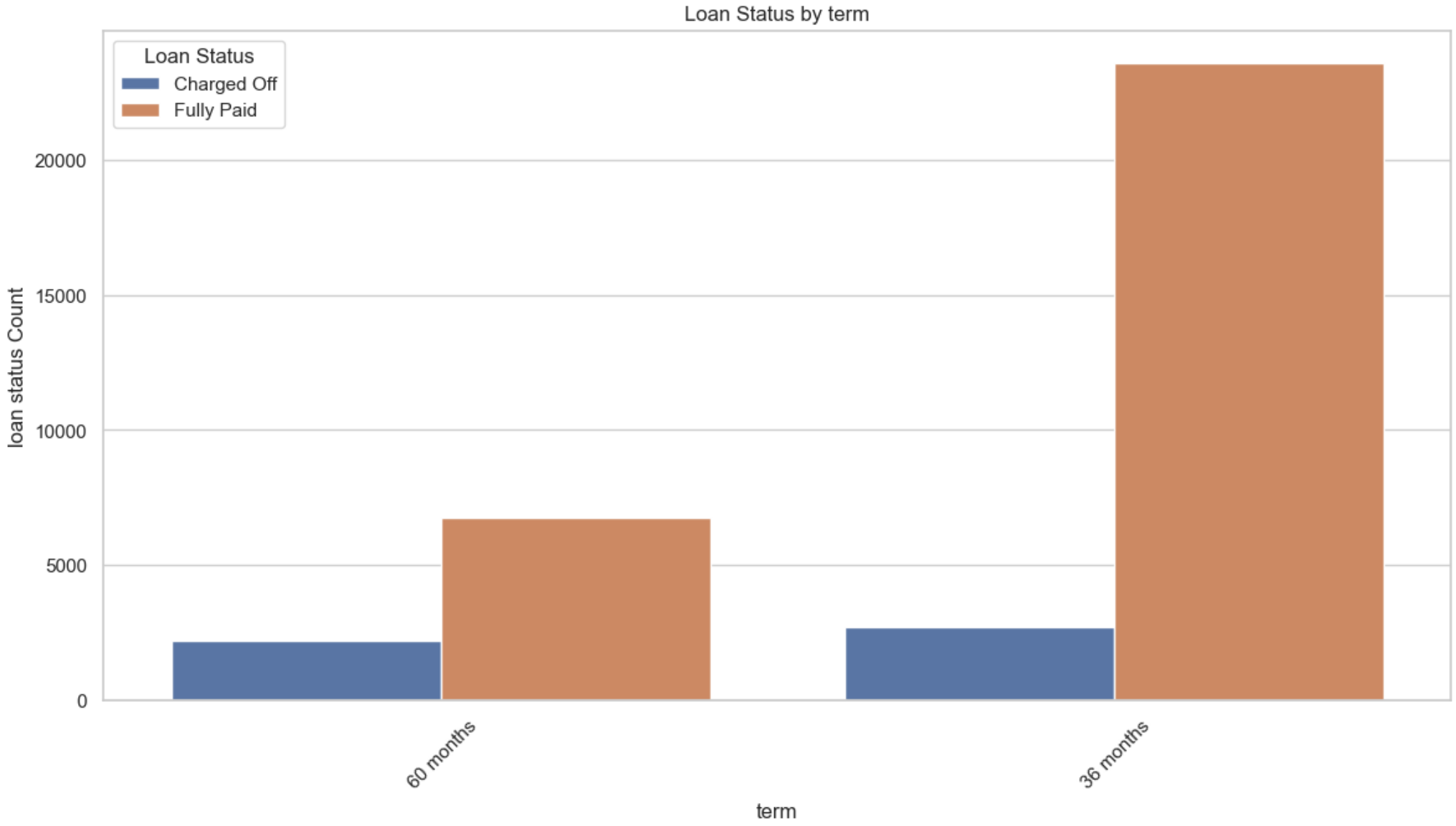
## Inferences From Univariate Analysis of Categorical Variables**

- 14% of loan accounts are charged off & rest all are fully paid by the customers
- Around 74.6% Loan Accounts Have Tenure 36months.
- Trend is very clear, Majority of the loans are approved for grade A, B & C
- 22.7% of loan is granted for the customers who has 10+ years of experience
- Customers who are either staying in rented house or having mortgage loan contributes to 91% of loans.
- Surprisingly 42% of loan granted to the customers whose income is not verified.
- 47% of the loan taken by the customers to consolidate their debts
- 41% of the loan taken by the customers living in CALIFORNIA(CA),NEW YORK(NY),FLORIDA(FL) & (TEXAS)TX
- 57 % of the loans are issued in the year 2011
- Loans issued grows steadily over the year, most in December

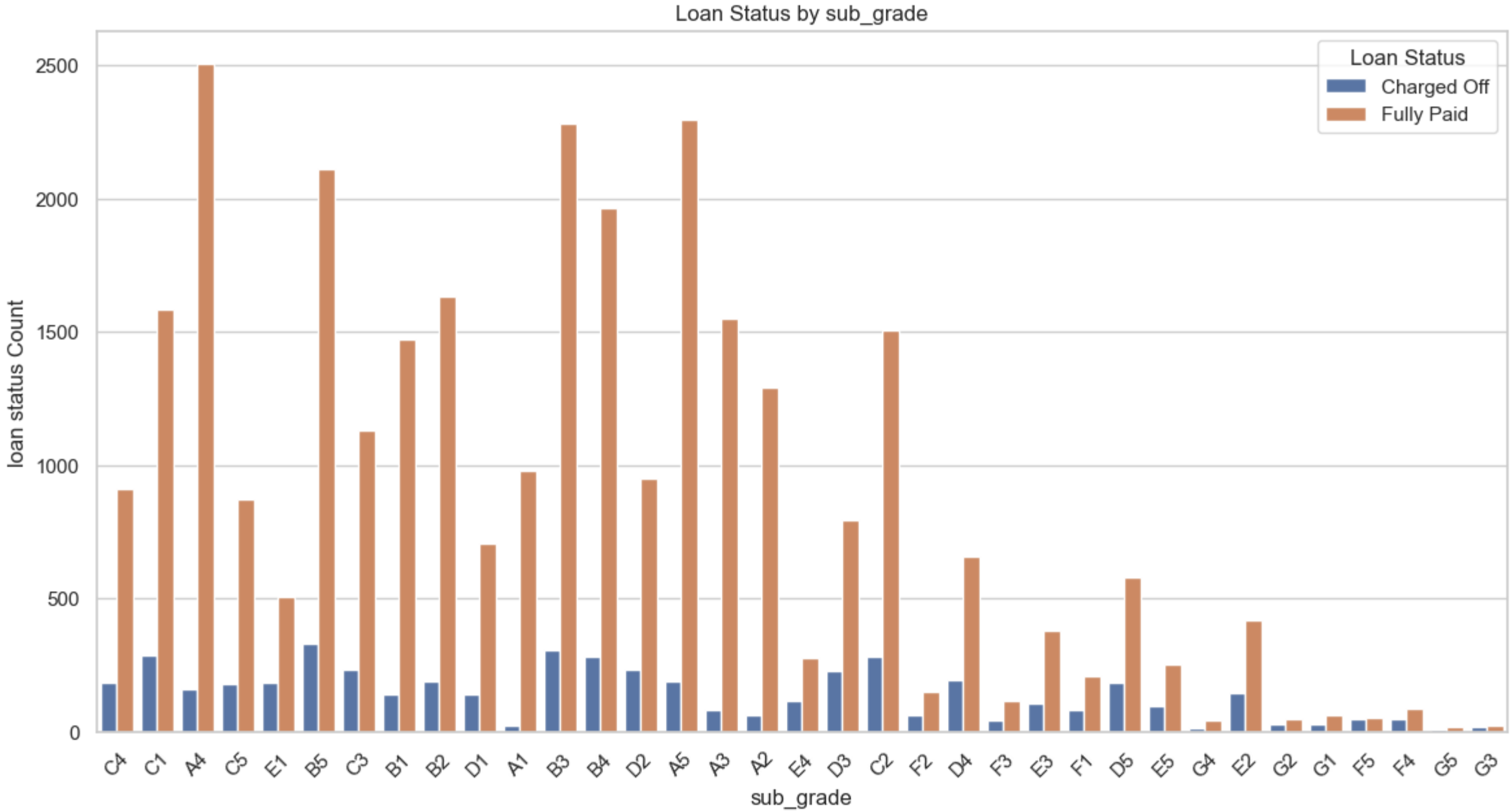
# Graphs – Bivariate Categorical (Loan stat)



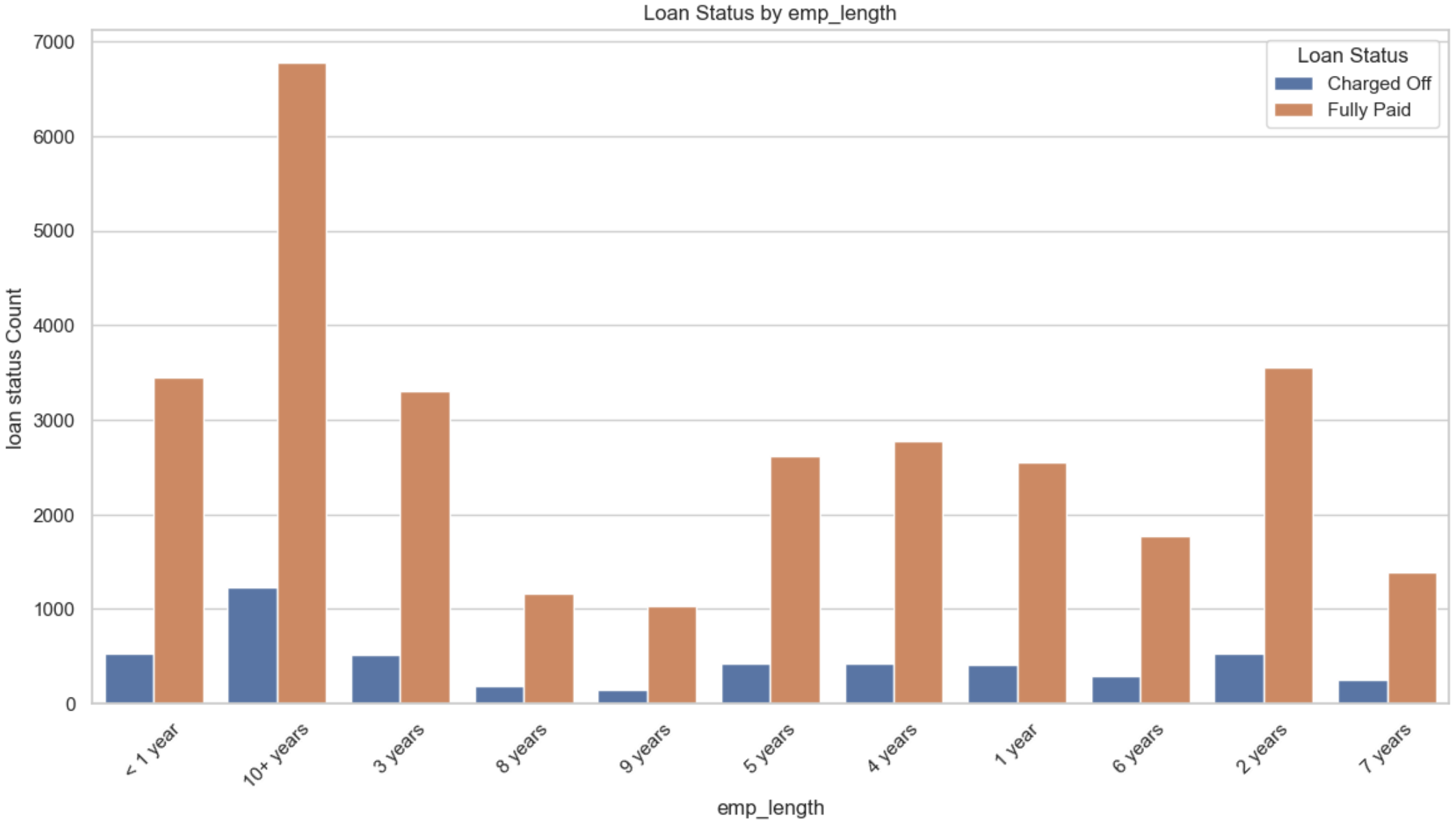

- 14% of loan accounts are charged off & rest all are fully paid by the customers
- We have more data of fully Paid, which means most customers tend to repay the loan.



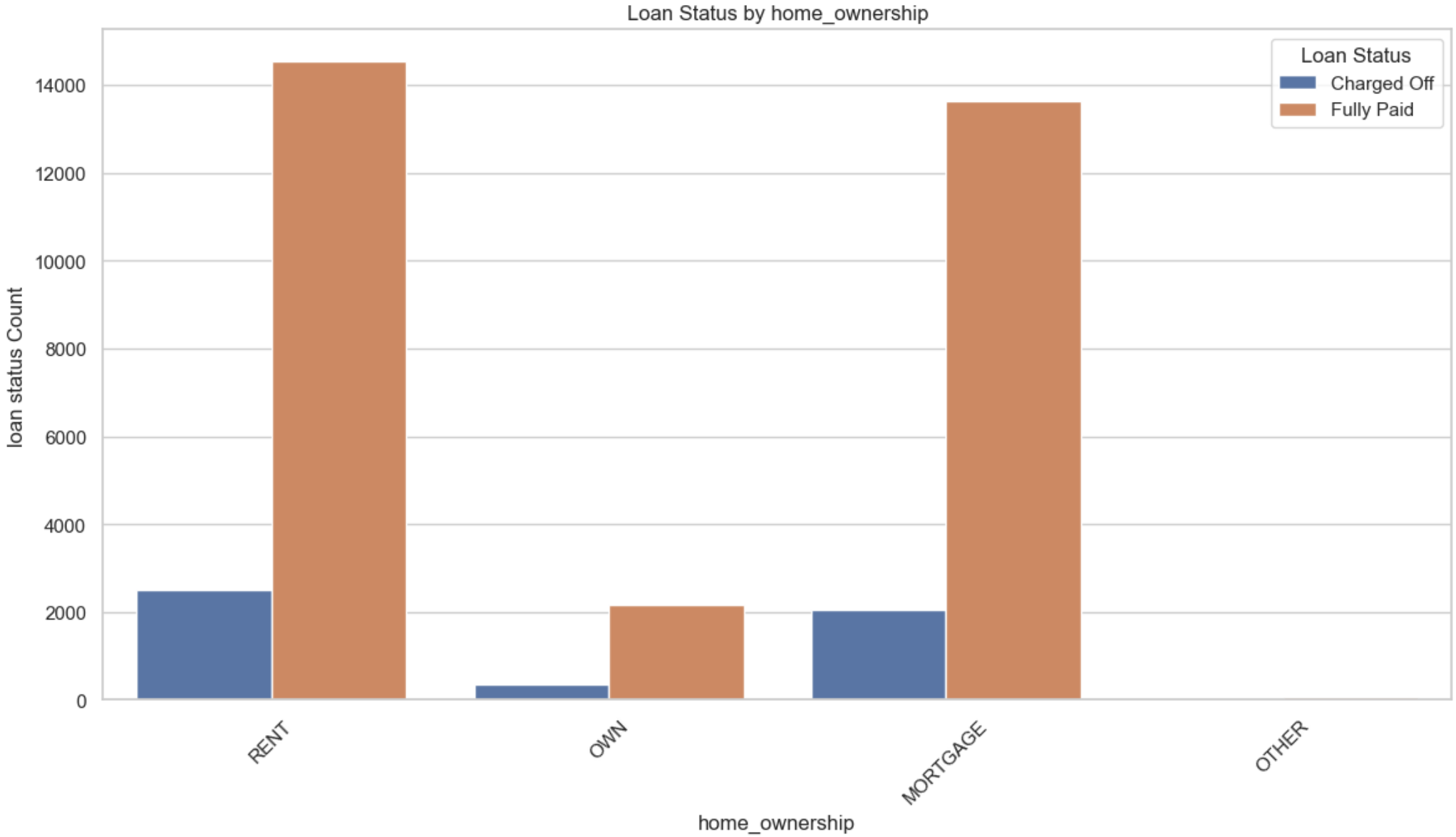

•Lesser the duration, higher chances of repayment.



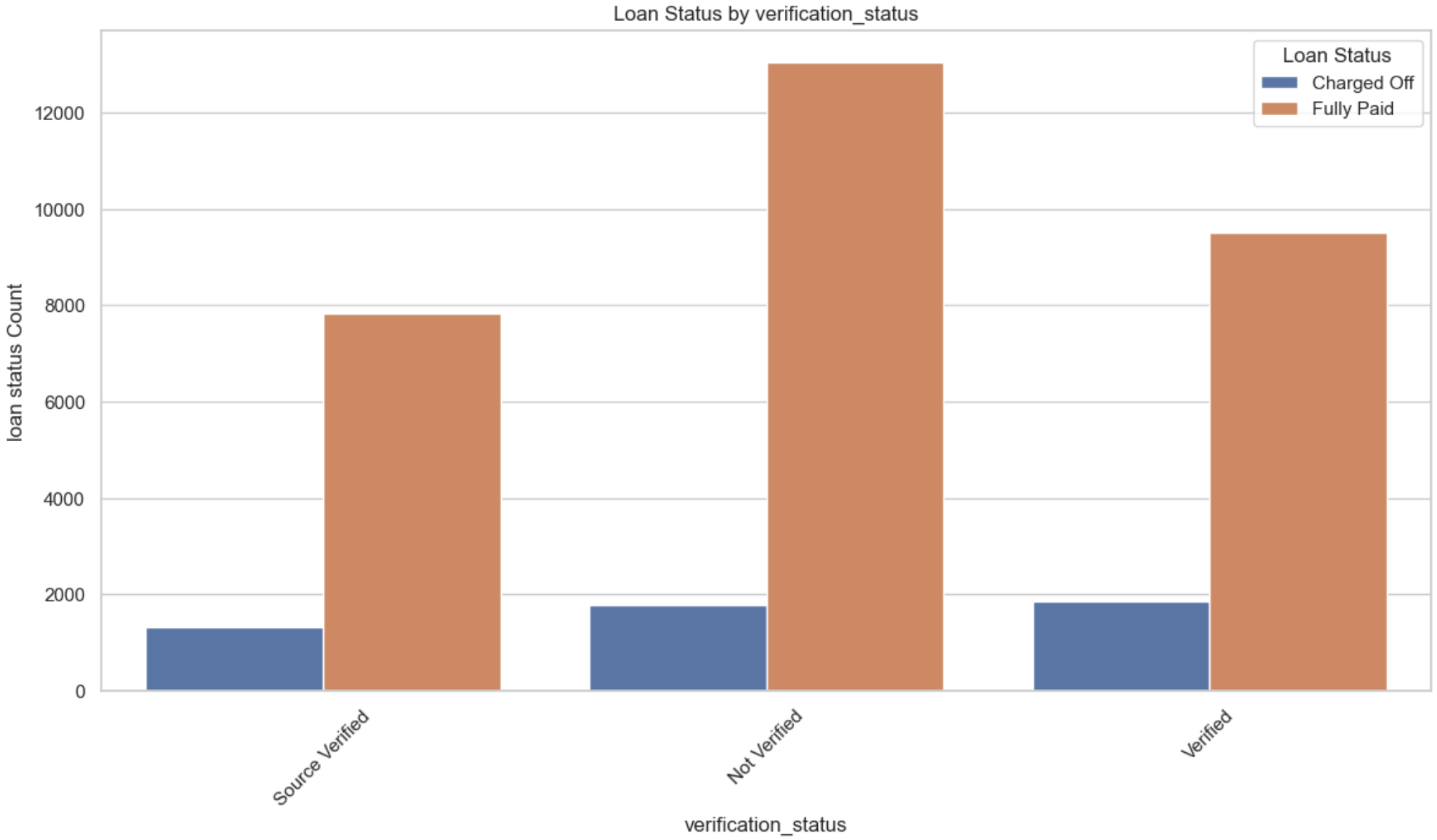

•A4 , A5 and A1 subcategories are highly likely to repay the loan. A1 category had less loan data, a potential growth area to concentrate.



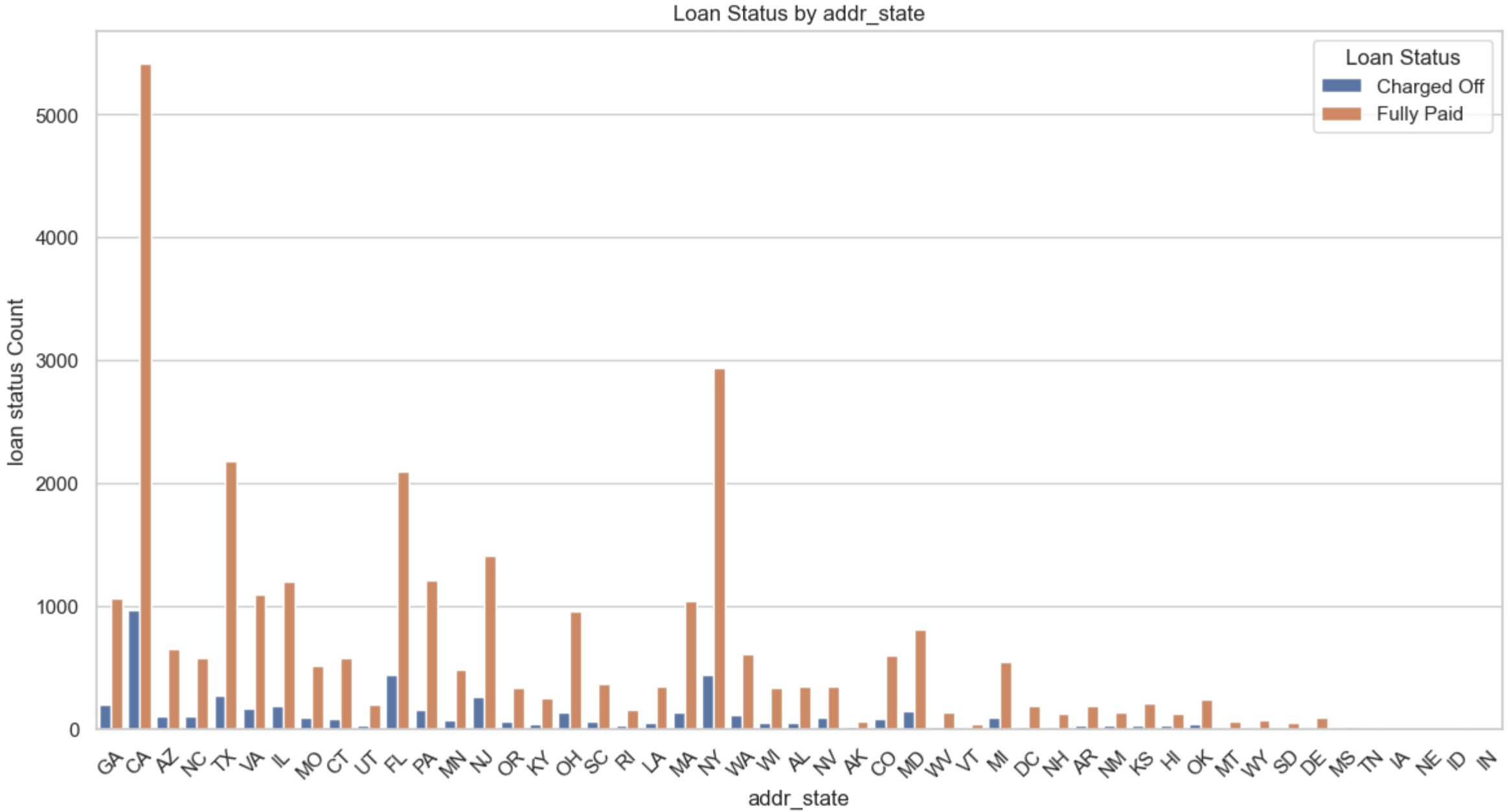

•Higher the employment length, higher chances of loan prepayment



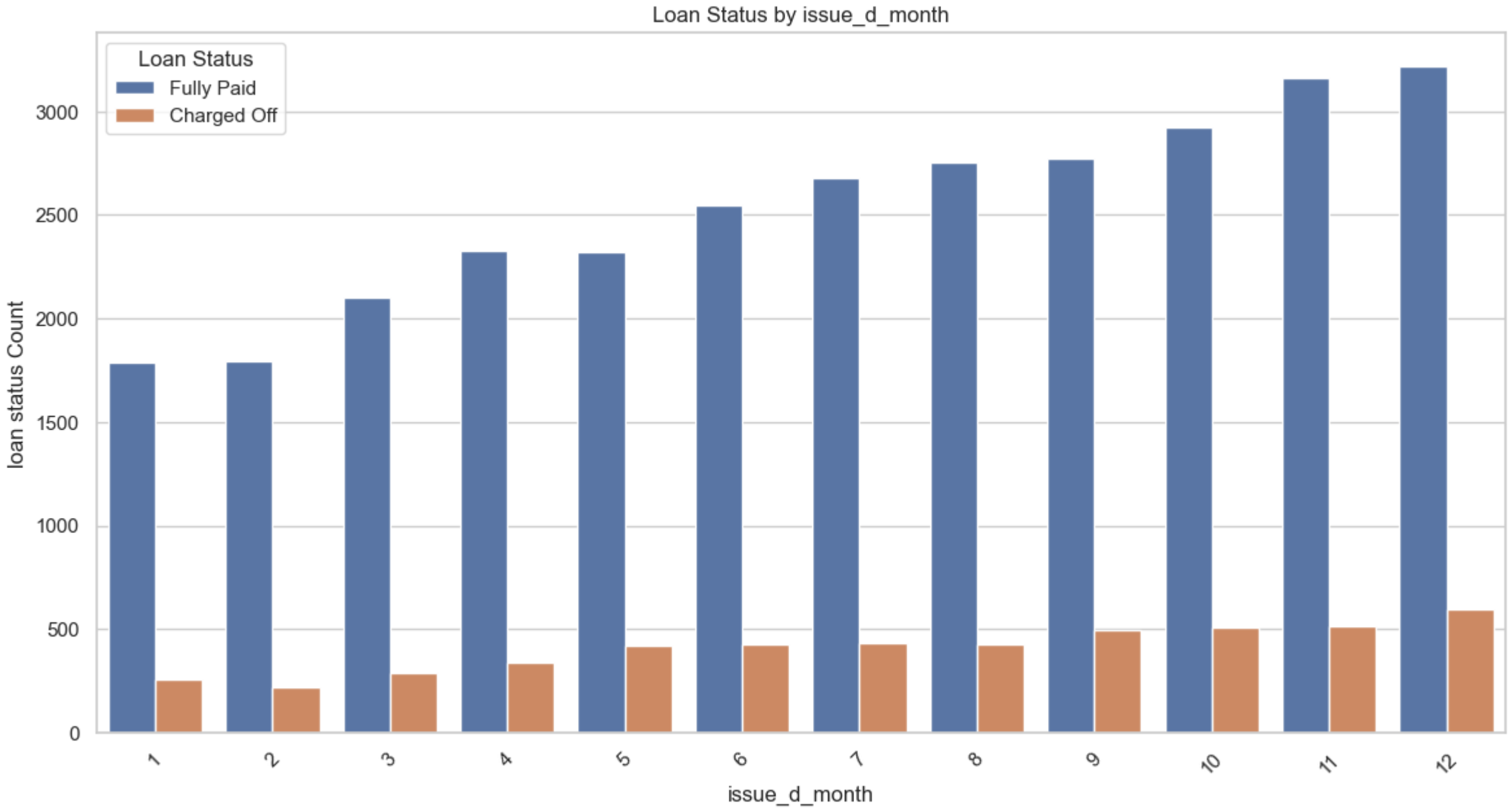

•Most people who take loans are on Rent, fully_paid % is also high.



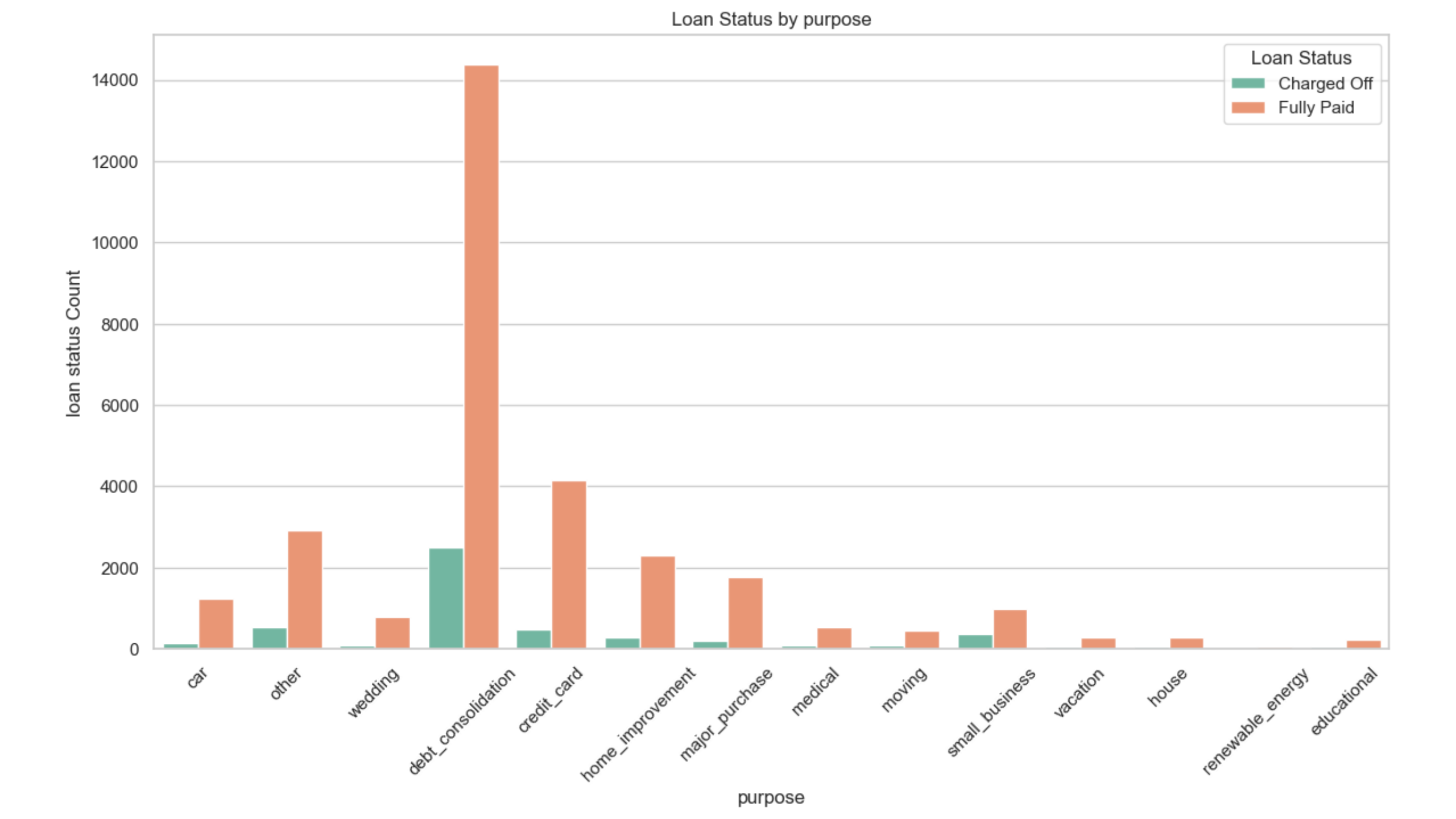

•Even though Source is not verified, the fully paid percentage is also high for the same




•Loans fully paid status percentage is highest in California




•Loans issued during Months of 3rd quarter of the Financial year and January have highest fully paid rate.

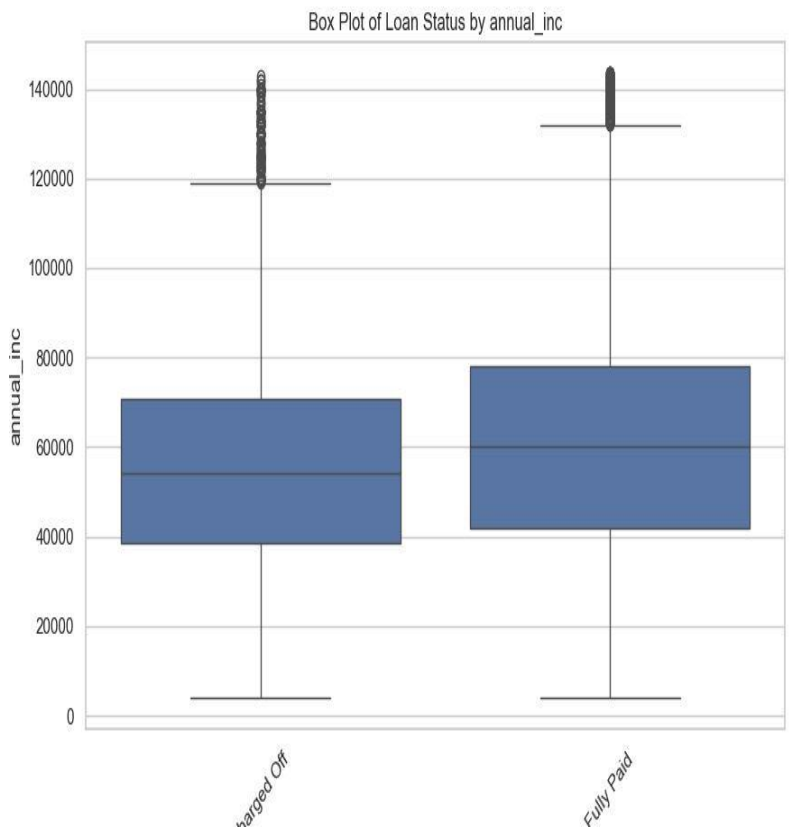

Loan Status by purpose
Loan Status
Charged Off
Fully Paid
loan status Count
14000
12000
10000
8000
6000
4000
2000
0
car
other
wedding
debt_consolidation
credit_card
home_improvement
major_purchase
medical
moving
small_business
vacation
house
renewable_energy
educational
purpose

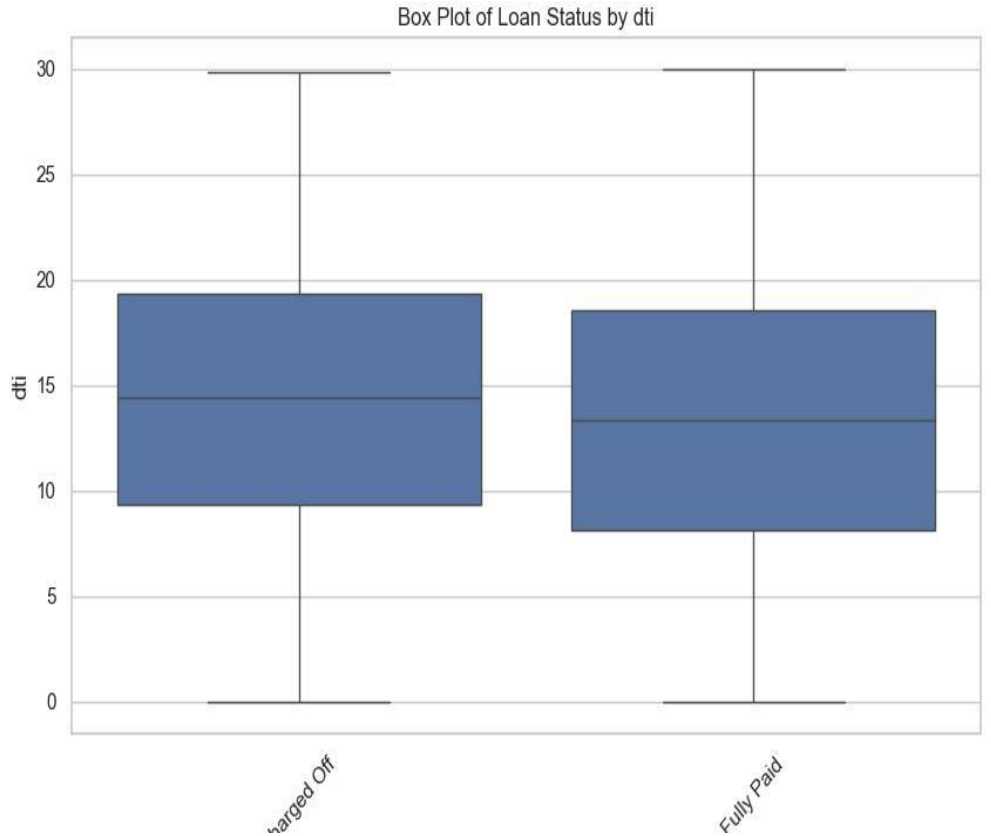
# Results

Inferences From Bivariate Analysis of Categorical Variables**

- We have more data of fully Paid, which means most customers tend to repay the loan.
- Lesser the duration, higher chances of repayment.
- A4 , A5 and A1 subcategories are highly likely to repay the loan. A1 category had less loan data, a potential growth area to concentrate.
- Higher the employment length, higher chances of loan prepayment
- Most people who take loans are on Rent, fully_paid % is also high.
- Eventhough Source is not verified, the fully paid percentage is also high for the same
- Loans fully paid status percentage is highest in California
- Loans issued during Months of last quarter of the year and January have highest fully paid rate.
- Most loans are for DEPT consolidation purpose

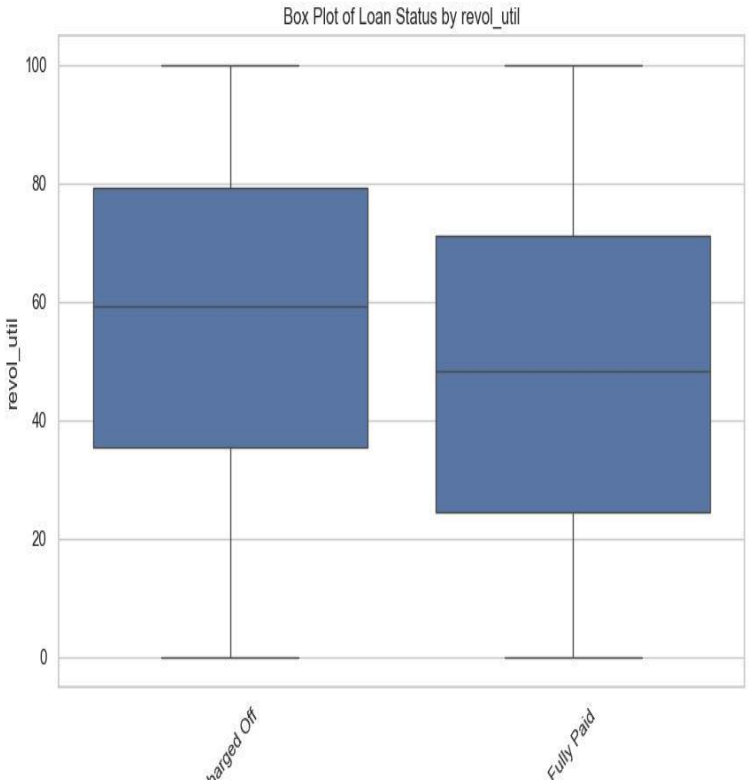
# Bivariate analysis of Continuous vars

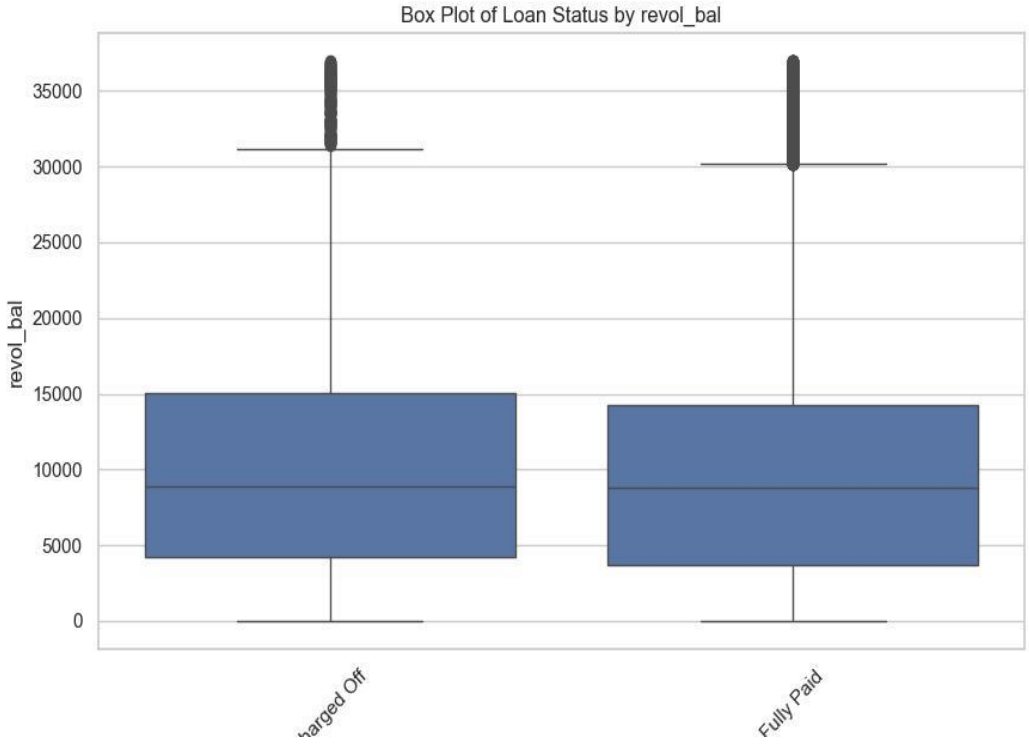


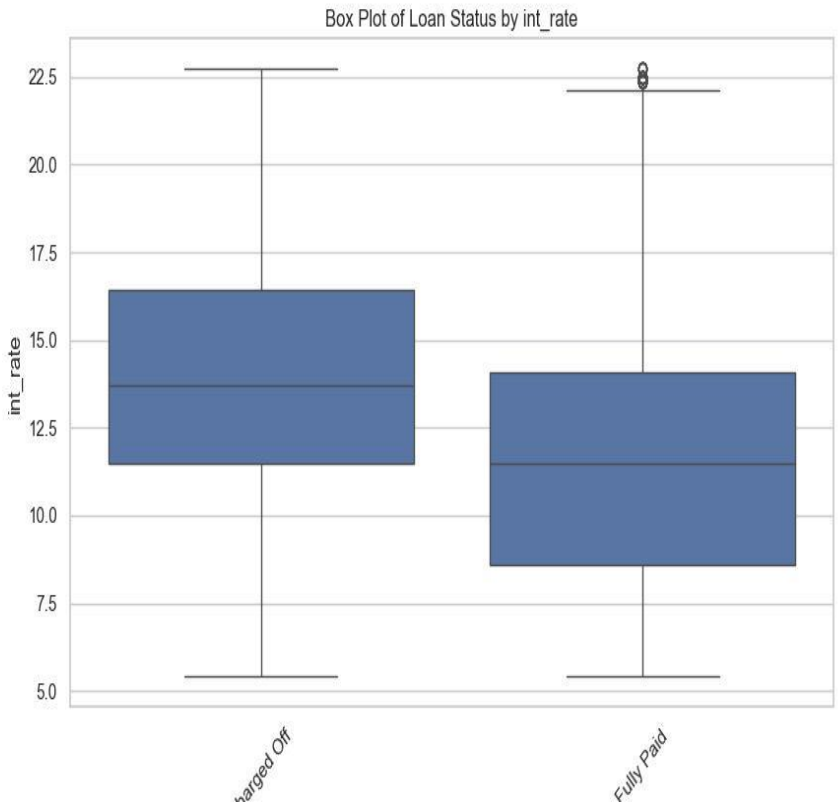

- Higher the annual income, higher the chances of full repayment.

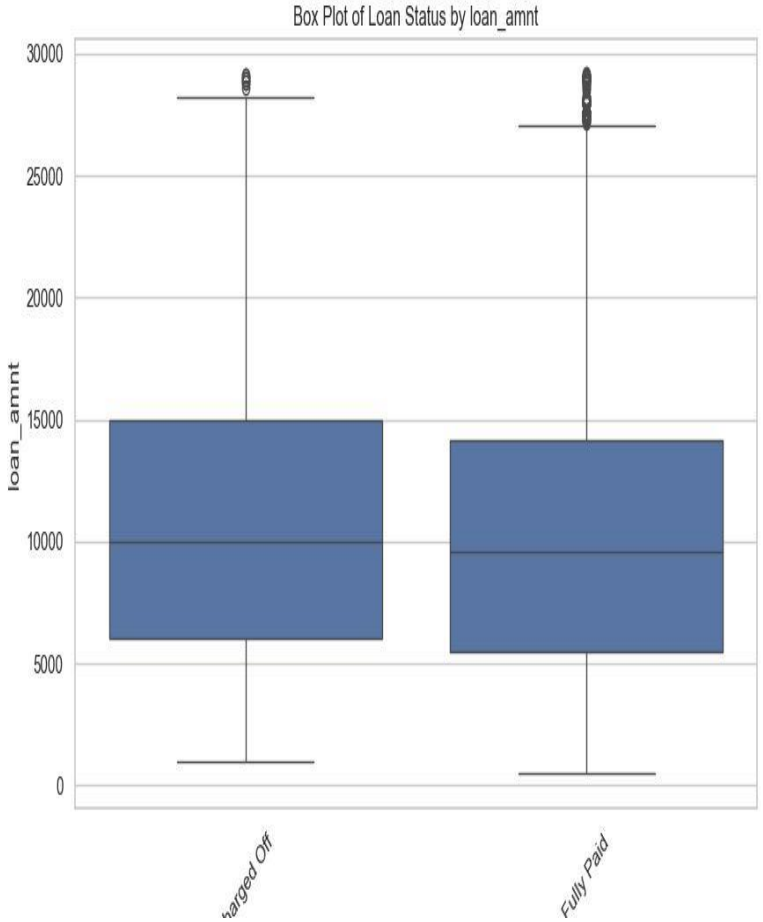
- Chances of defaulter if dti is high



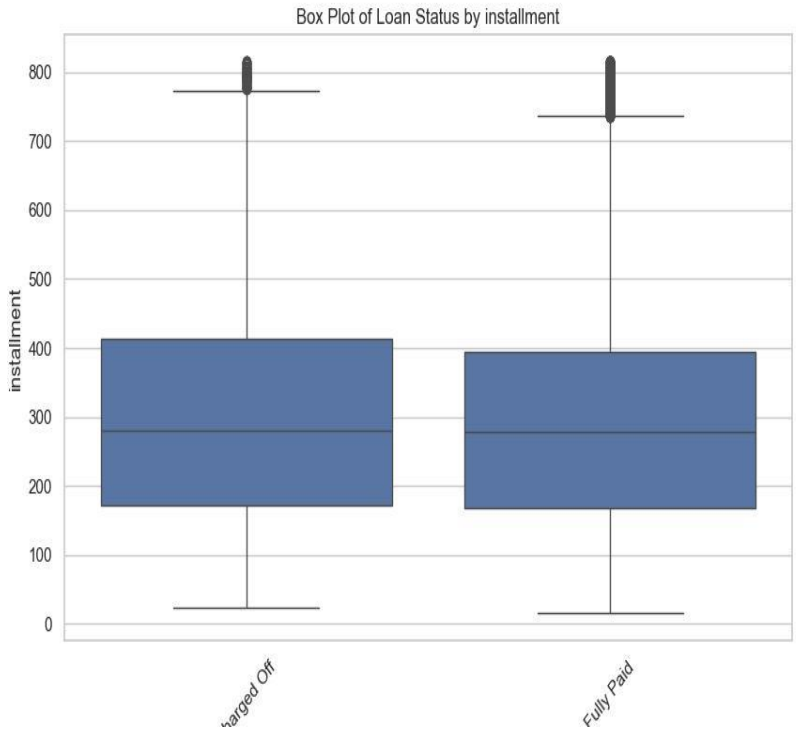

- Chances of defaulter if revol_util is high

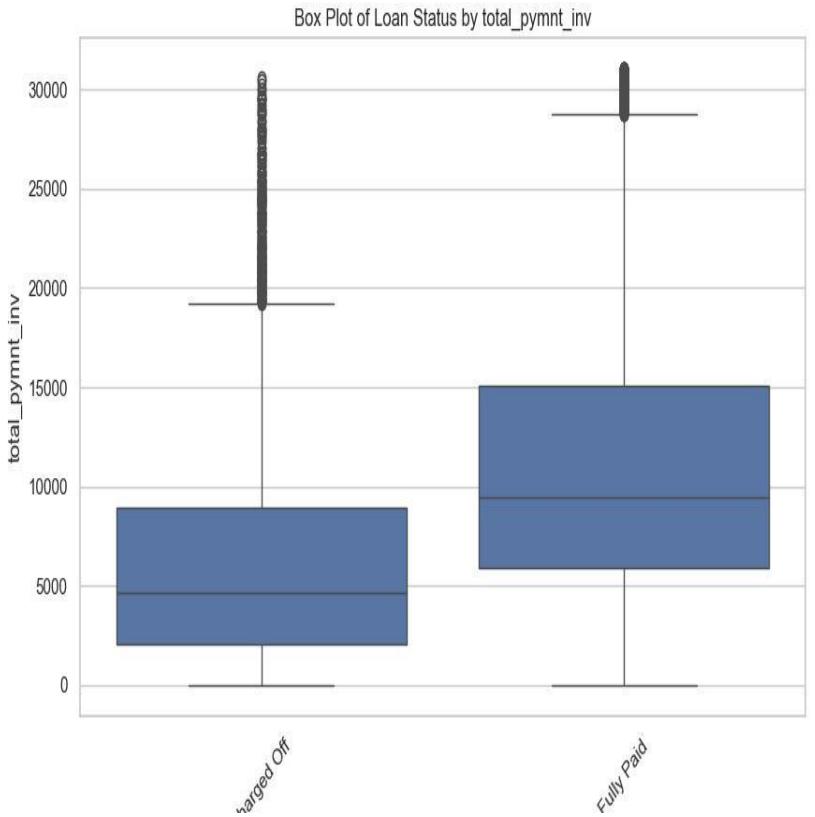



•High chances of defaulter if interest rate is high.

•Loan amount doesn’t make much difference on loanStatus.




•Instalment doesn’t make an impact on loan status

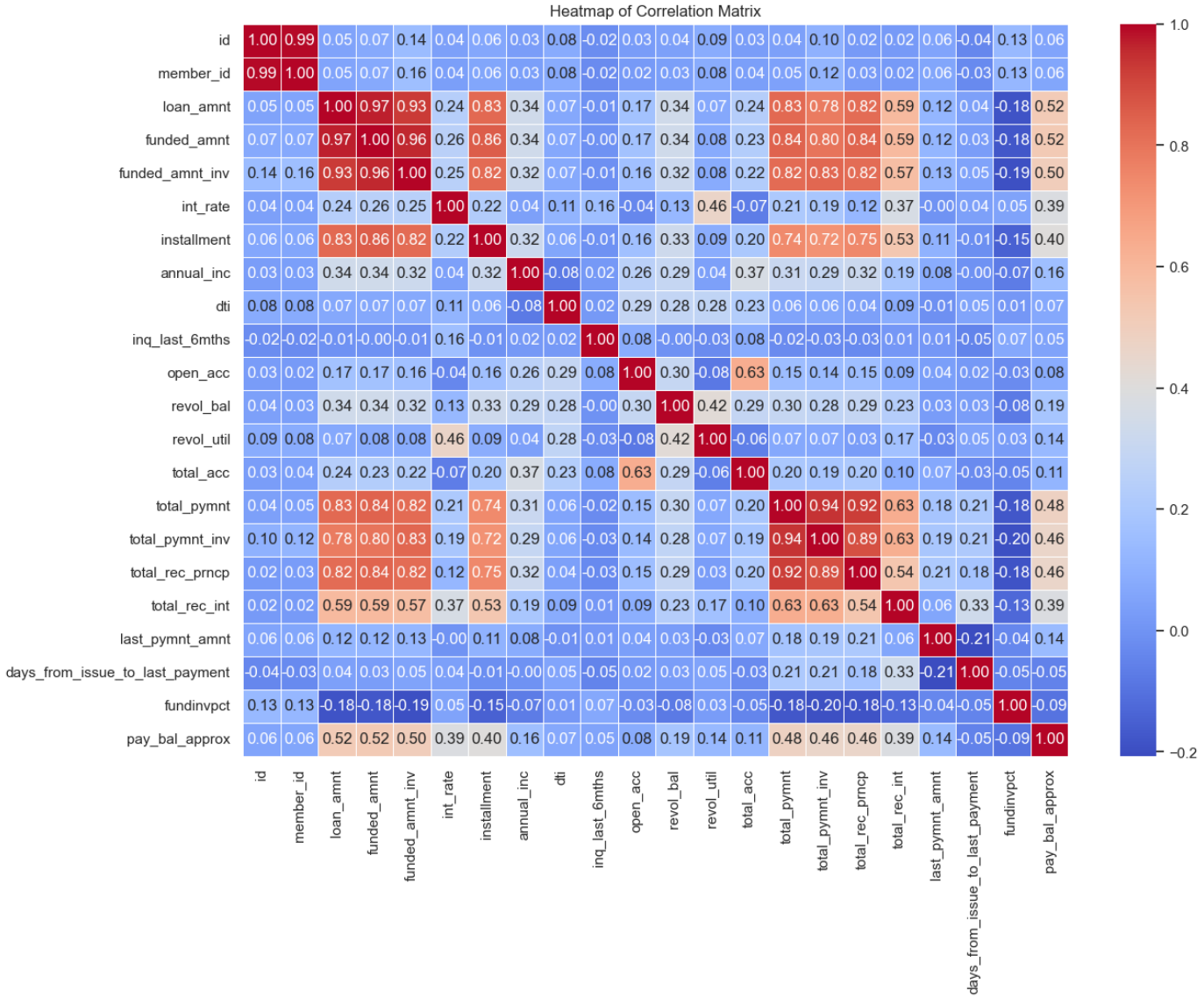
- If total payment by investors is high, the loan mostly will get fully paid.

# Results

## Inferences From Bivariate Analysis of Continuous Variables**

- Chances of defaulter if dti is high
- Chances of defaulter if revol_util is high
- If total_payment is high mostly the customer will close the loan.
- If total payment by investors is high, the loan mostly will get fully paid.
- High chances of defaulter if interest rate is high.
- Higher the annual income, higher the chances of full repayment.
- Loan amount doesn’t make much difference on loanStatus.
- Instalment doesn’t make an impact on loan status

# Multivariate HeatMap




## Inferences From Multivariate Analysis

- Loan Tenure has a good correlation with interest rate.
- Surprisingly, total_acc has negative correlation with earliest_cr_line_year.
- Revol_bal & revol_util doesn't have correlation with Loan Amount

# Recommendations

- **Market Expansion**: Explore opportunities to expand operations in regions exhibiting characteristics similar to California, where loan repayment rates are high.
- **Product Diversification**: Consider offering shorter-term loan options, especially in categories like **A1** where there's potential for growth.
- **Targeted Marketing**: Tailor marketing efforts towards people who are in rented houses, stable employment emphasizing the reliability of this segment in loan repayment.
- **Seasonal Strategy**: Capitalize on the trend of higher fully paid rates in the last quarter of the year and January by strategically timing loan issuance and marketing campaigns during these periods. Align product offerings and promotions with consumer behavior patterns to maximize engagement and repayment rates.
- **Risk Assessment**: Shorter loan durations, has higher chances of repayment. This suggests that offering shorter-term loans might mitigate default risks. Similarly for longer employment.

# Recommendations

**Verification Processes**: Despite unverified sources showing high repayment rates, it's crucial to maintain rigorous verification processes to mitigate fraud and default risks. Balance flexibility with risk management to ensure sustainable lending practices.

**Focus on Shorter-Term, lower interest rate Loans**: Since shorter loan durations and low interest rate correlate with higher repayment rates, consider offering more loan options with shorter terms.