Lending Club Case Study

Exploratory Data Analysis

Satyajeet Gupta

Introduction

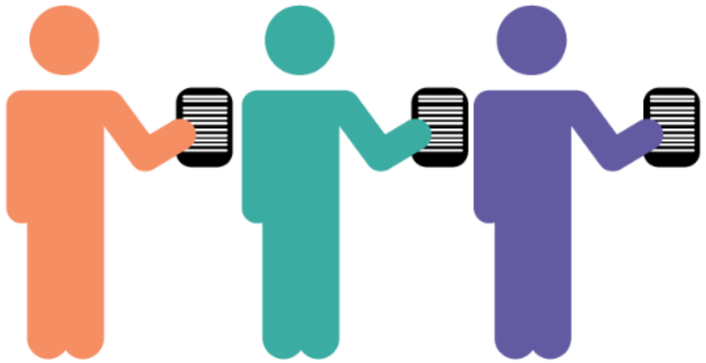
Business Understanding

- You work for a consumer finance company which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

LOAN DATASET



Loan Accepted → Default

Loan Accepted → Non-Default

Loan Rejected

(Not considered in dataset)

Types of Decisions

1. **Loan accepted:** If the company approves the loan, there are 3 possible scenarios described below:

   - **Fully paid:** Applicant has fully paid the loan (the principal and the interest rate)
   - **Current:** Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
   - **Charged-off:** Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

2. **Loan rejected:** The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objectives

- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed.

- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- In other words, 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 utilise this knowledge for its portfolio and risk assessment.

Steps of Analysis

- **Data Exploration:** Understand data characteristics, identify patterns, and uncover potential issues.
- **Data Preparation:** Cleanse, transform, and structure data for analysis.
- **Communication and Action:** Visualise findings, communicate results, and recommend actions.

Basic Information about Data

- **Data set:** *loan.csv*
- **Total Number of Columns:** 111
- **Total Number of Rows:** 39717
- **Total missing values:** 2263366
- **Total unique values:** 416800
- **Number of duplicates:** 0

Data Cleansing

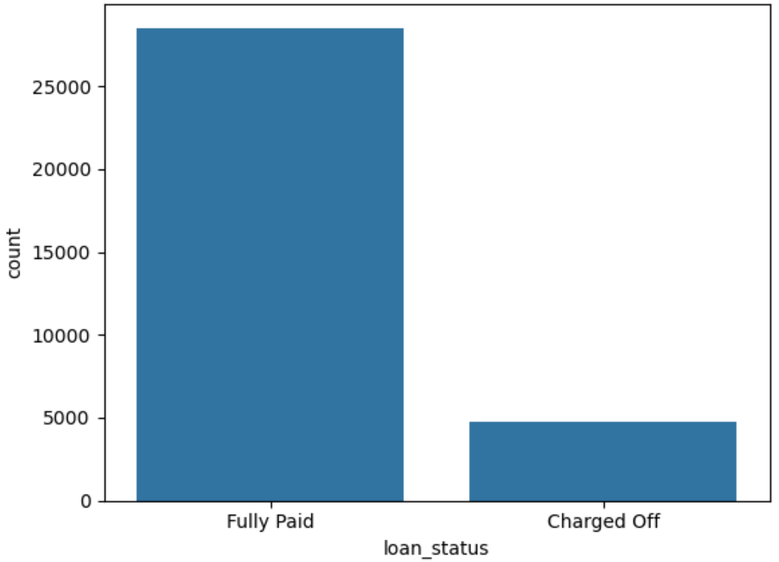
- Check for columns with null values and remove columns having more than 60% null values.
- Check for rows with null values, in our case we did not find any rows with all null values.
- Remove columns which does not contribute to analysis, e.g.
  - Some columns correspond to the post approval of loan, which do not contribute to our analysis.
  - Some columns are duplicate like id and member_id.
  - Some columns does not attribute to loan default analysis.
- Check for unique values in columns and remove columns having only 1 unique element.
- Remove the current loan data as it can not amount to analysis.

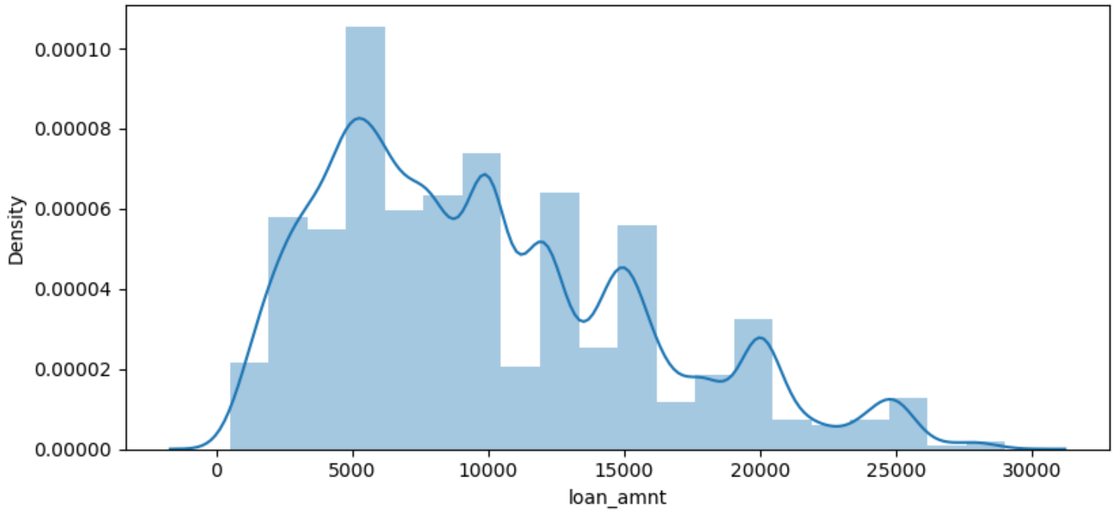
Data Transformation

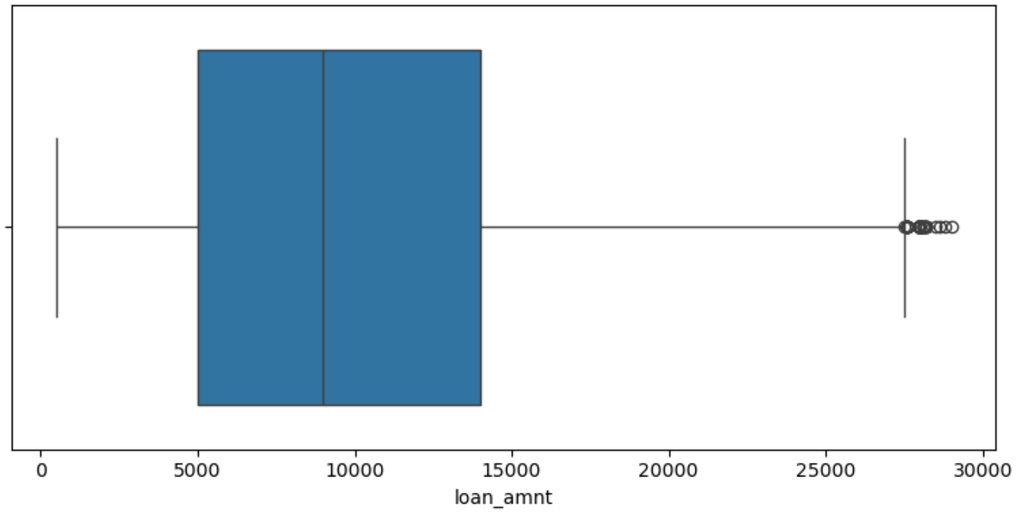
- Correcting the data type of columns: term, int_rate.
- Dropping the rows with null values for columns: emp_length, pub_rec_bankruptcies
- Reformatting the data in column: emp_length data.
- Adding new columns: issue_year, issue_month from column: issue_d
- Rounding off the columns: loan_amnt, funded_amnt, int_rate, dti to two decimals points.
- Checked and removed the outliers based on the IQR range keeping the threshold value to 1.5

Univariate Analysis

- Most applicants have applied for loan between 5k to 14 K
- Defaulters are less compared to fully paid loans

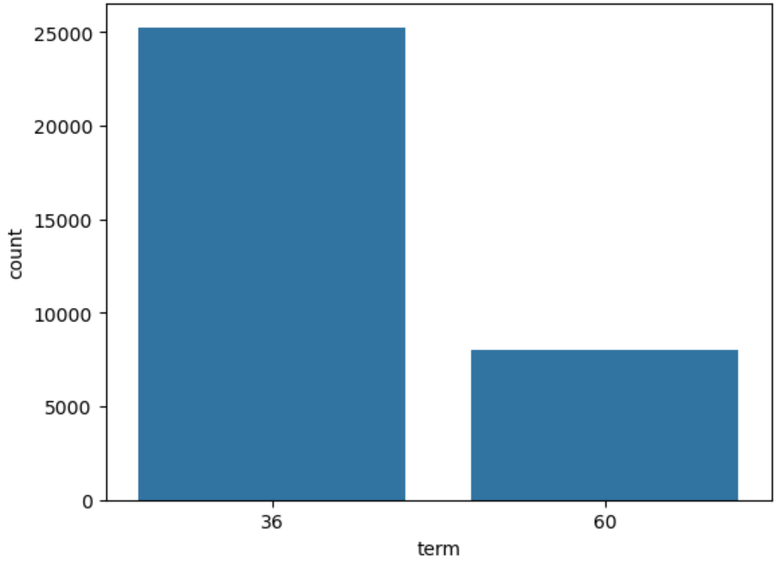


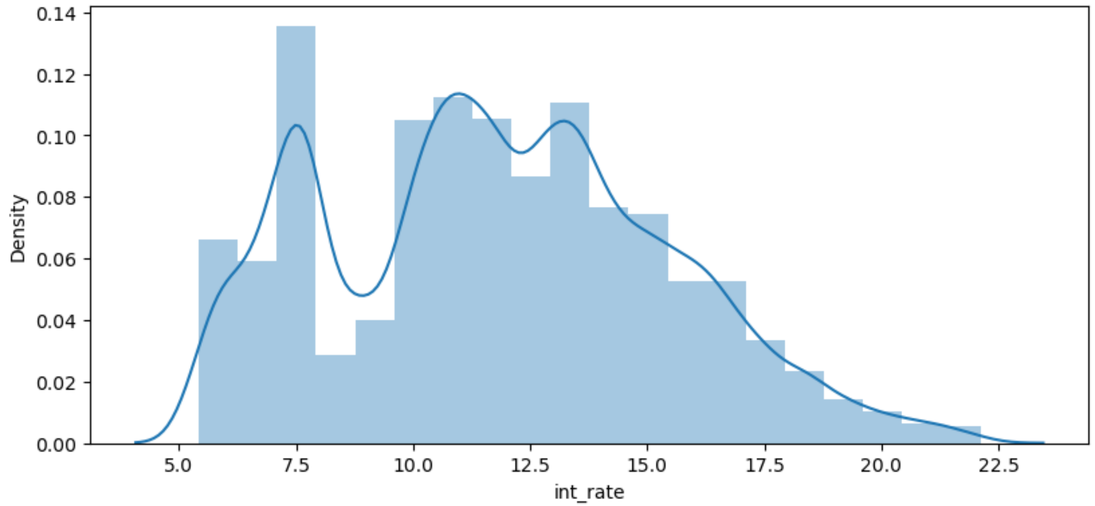


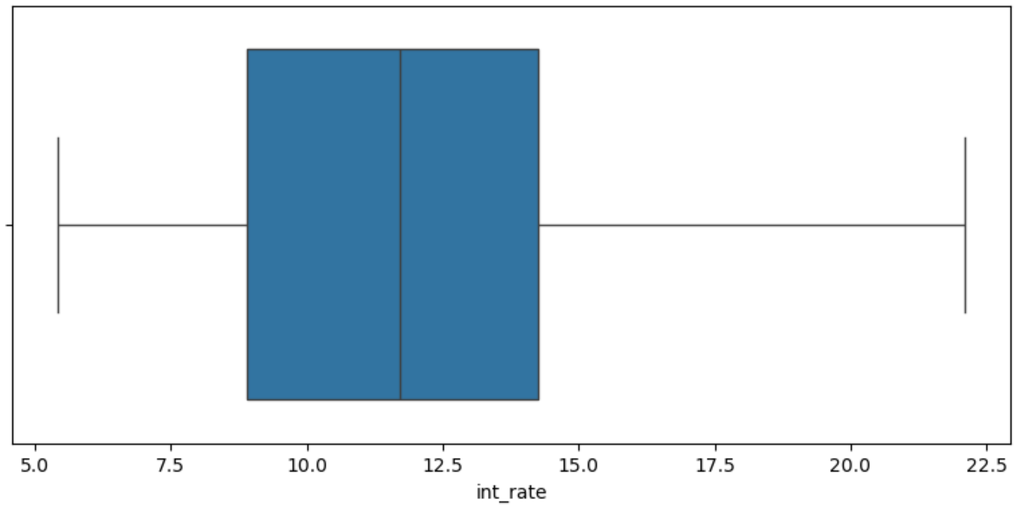


Univariate Analysis

- Most applicants have interest rate between 8.9 to 14.26 and avg Interest Rate is 11.79 %
- Most applicants have opted for 36 months tenure

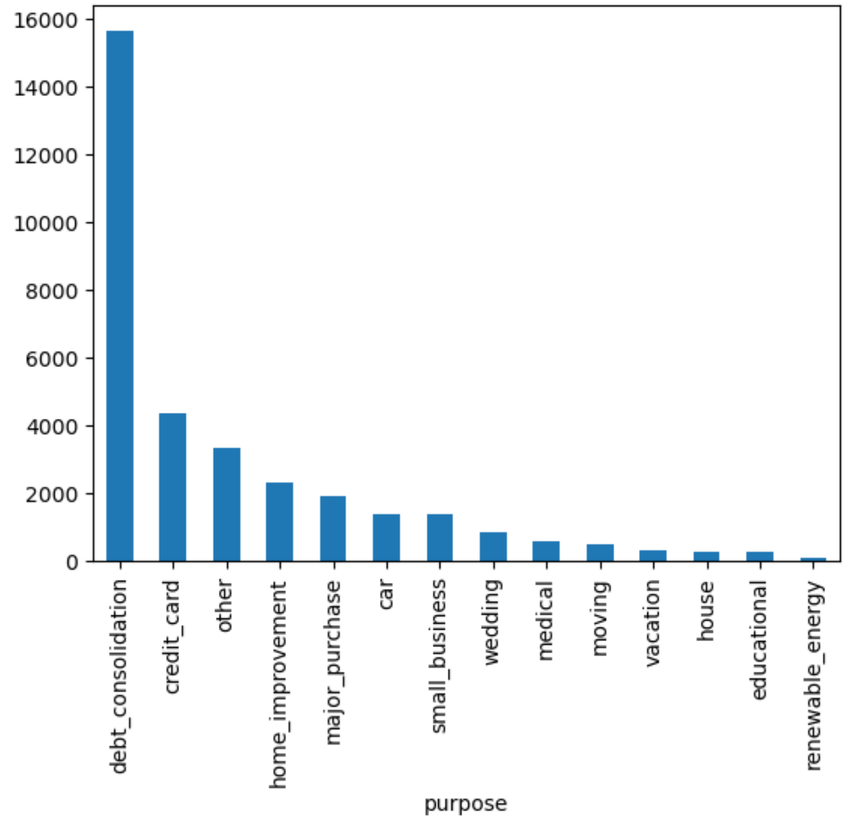


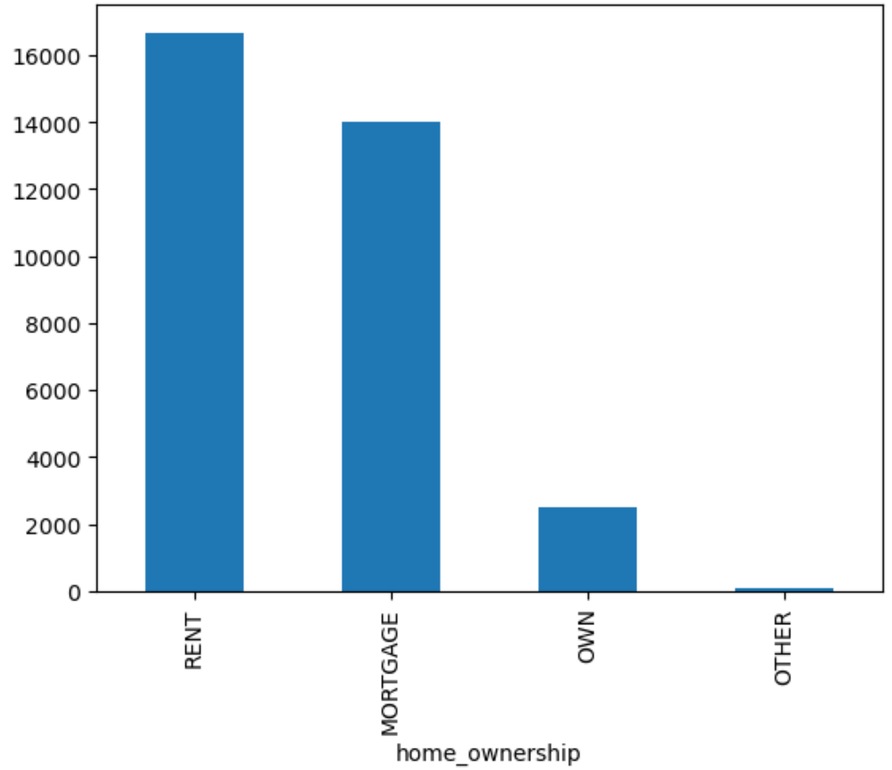




Univariate Analysis

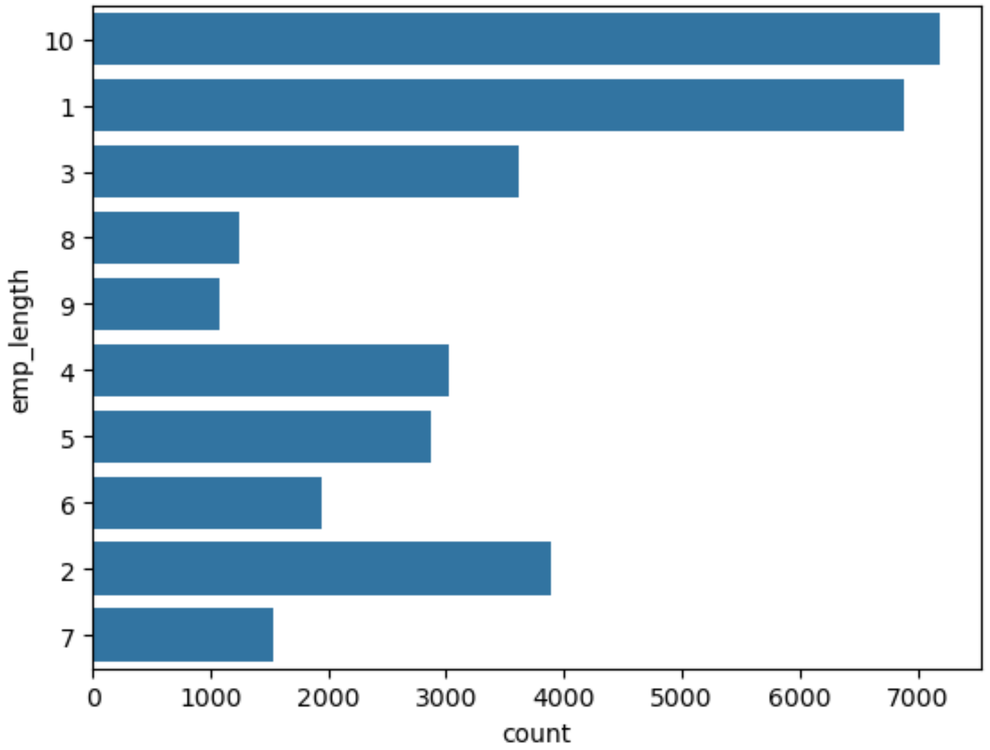
- Majority of applicants have taken loan for debt consolidation.
- Most loan applicants do not own their property

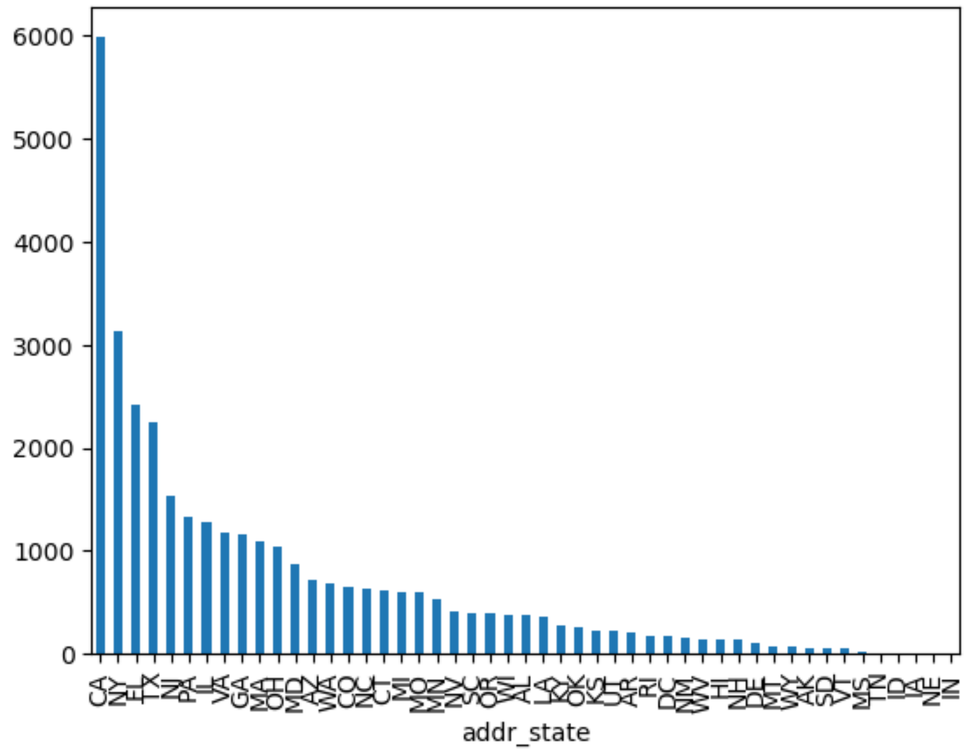




Univariate Analysis

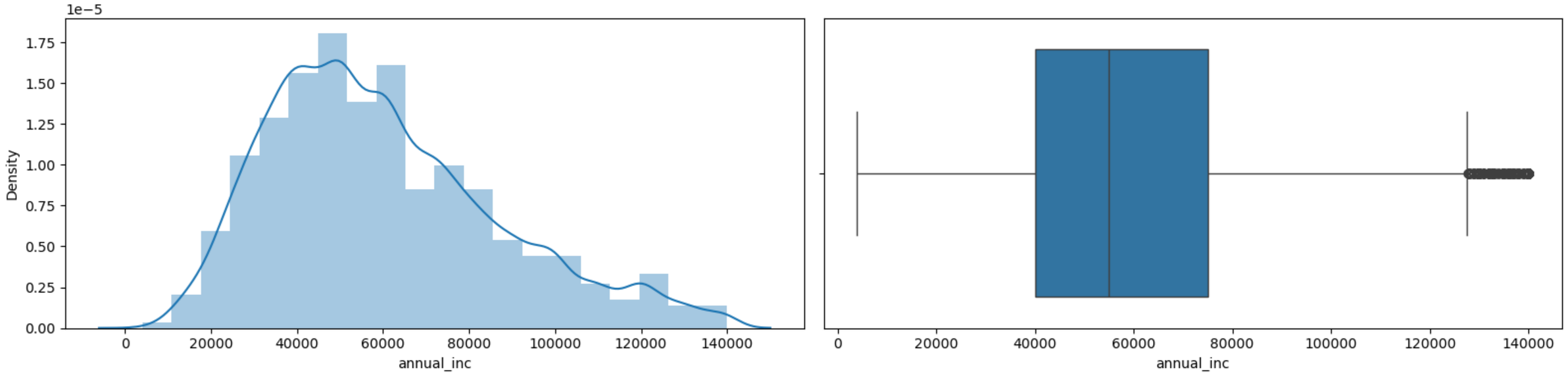
- People having employment length 10 or more are highest borrowers.
- Most loan applicants are from CA





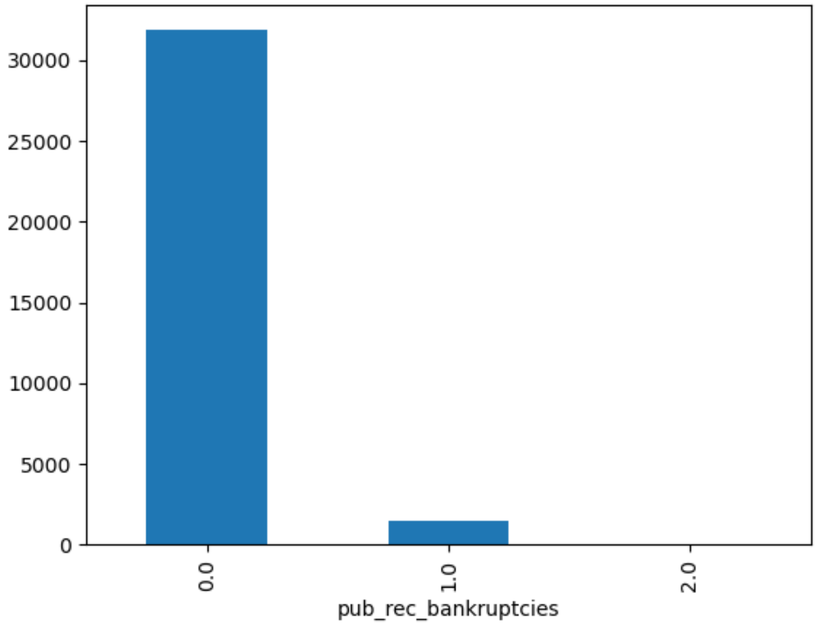
Univariate Analysis

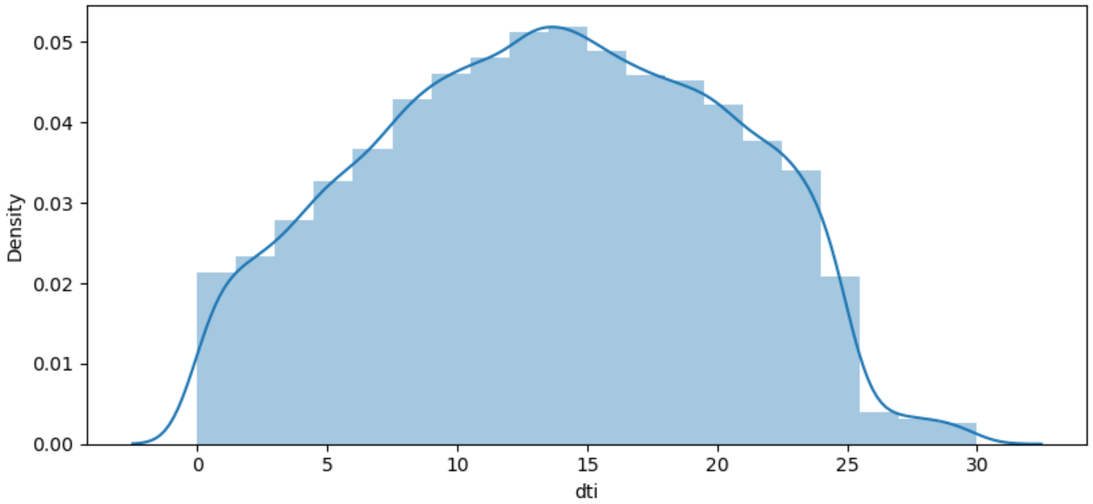
- Most loan applicants have annual income between 40k to 80 K.

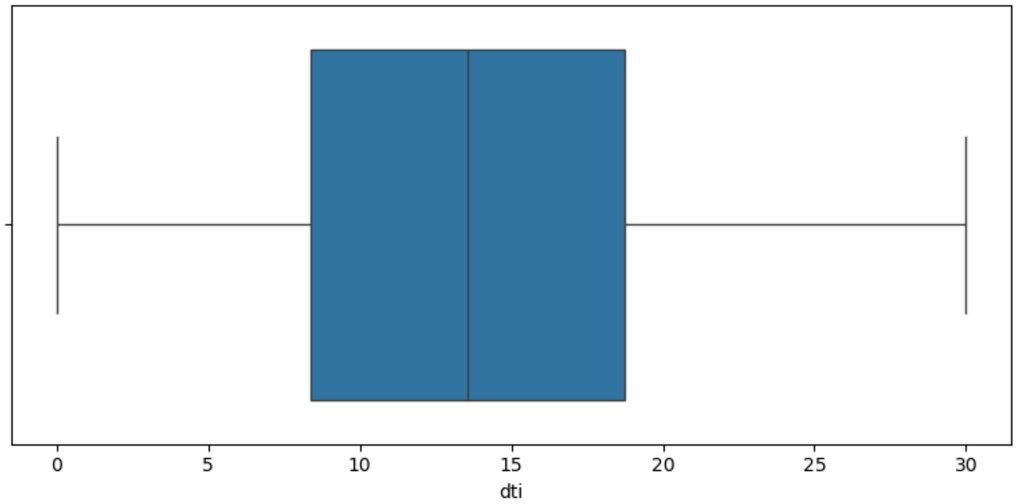


Univariate Analysis

- Most borrowers have no public record of bankruptcies.
- Most applicants have large debt to income ratio.

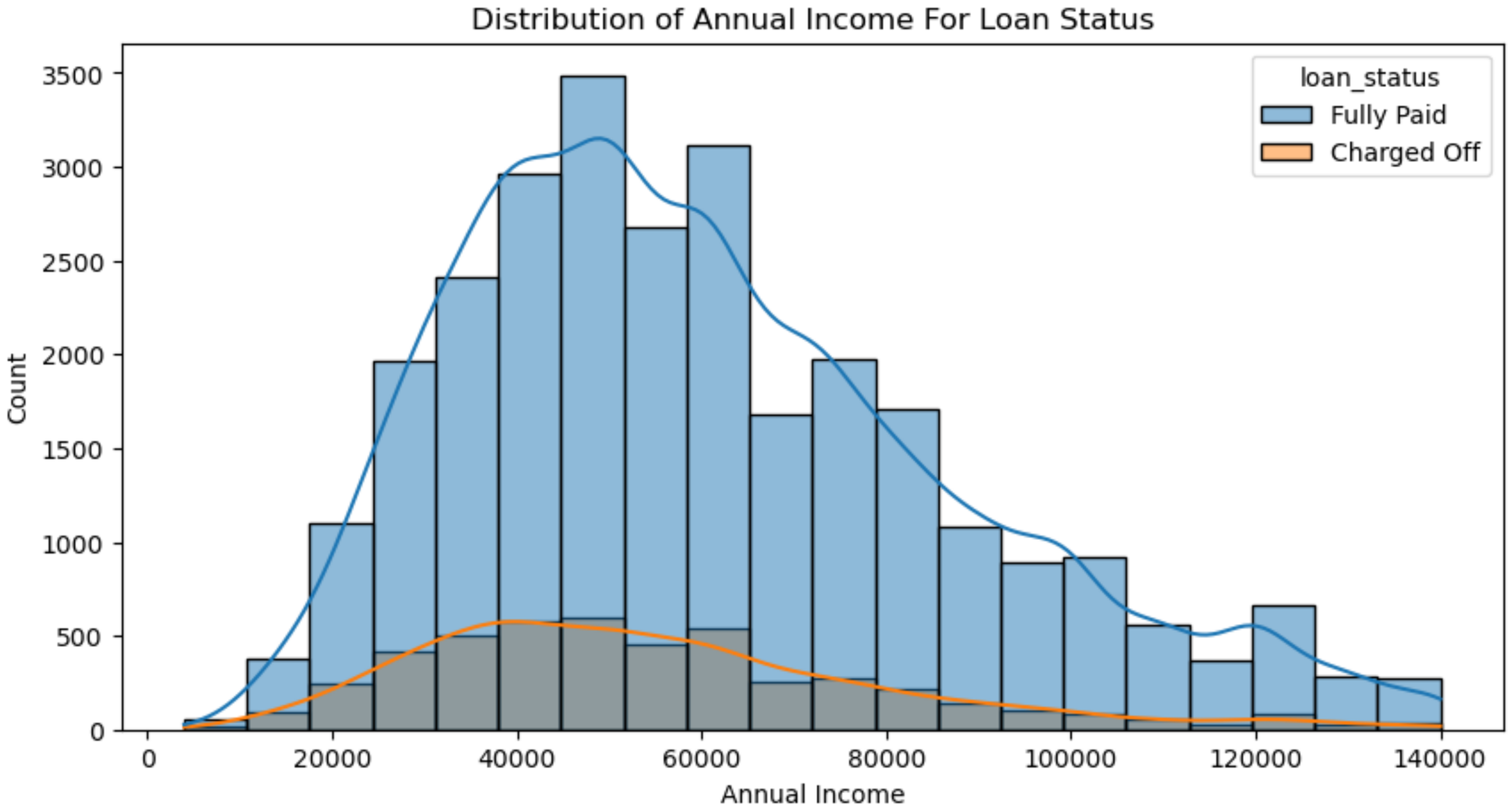






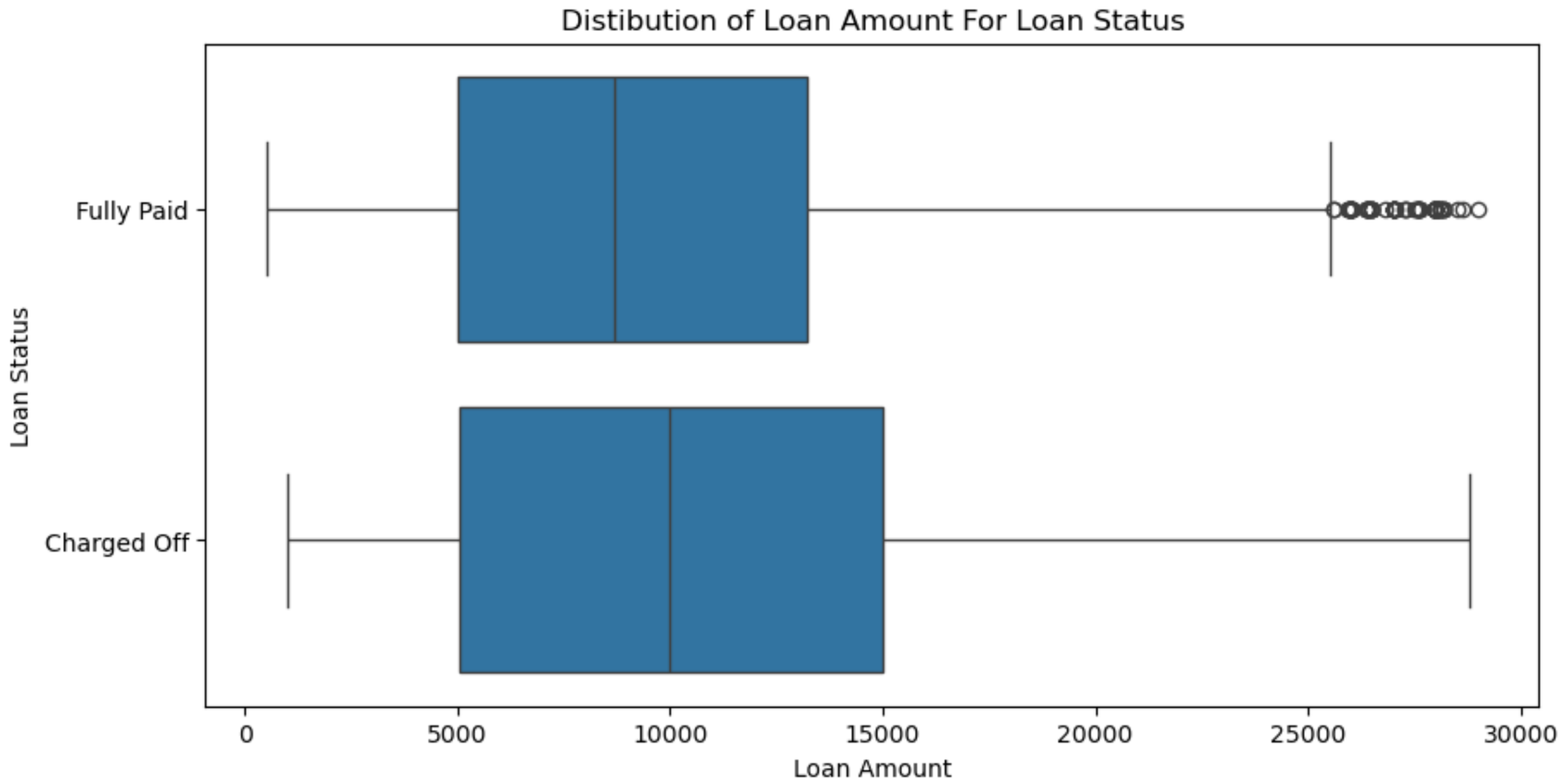
Bivariate Analysis

- Applicants having income in range 0-40 K are more likely to default.



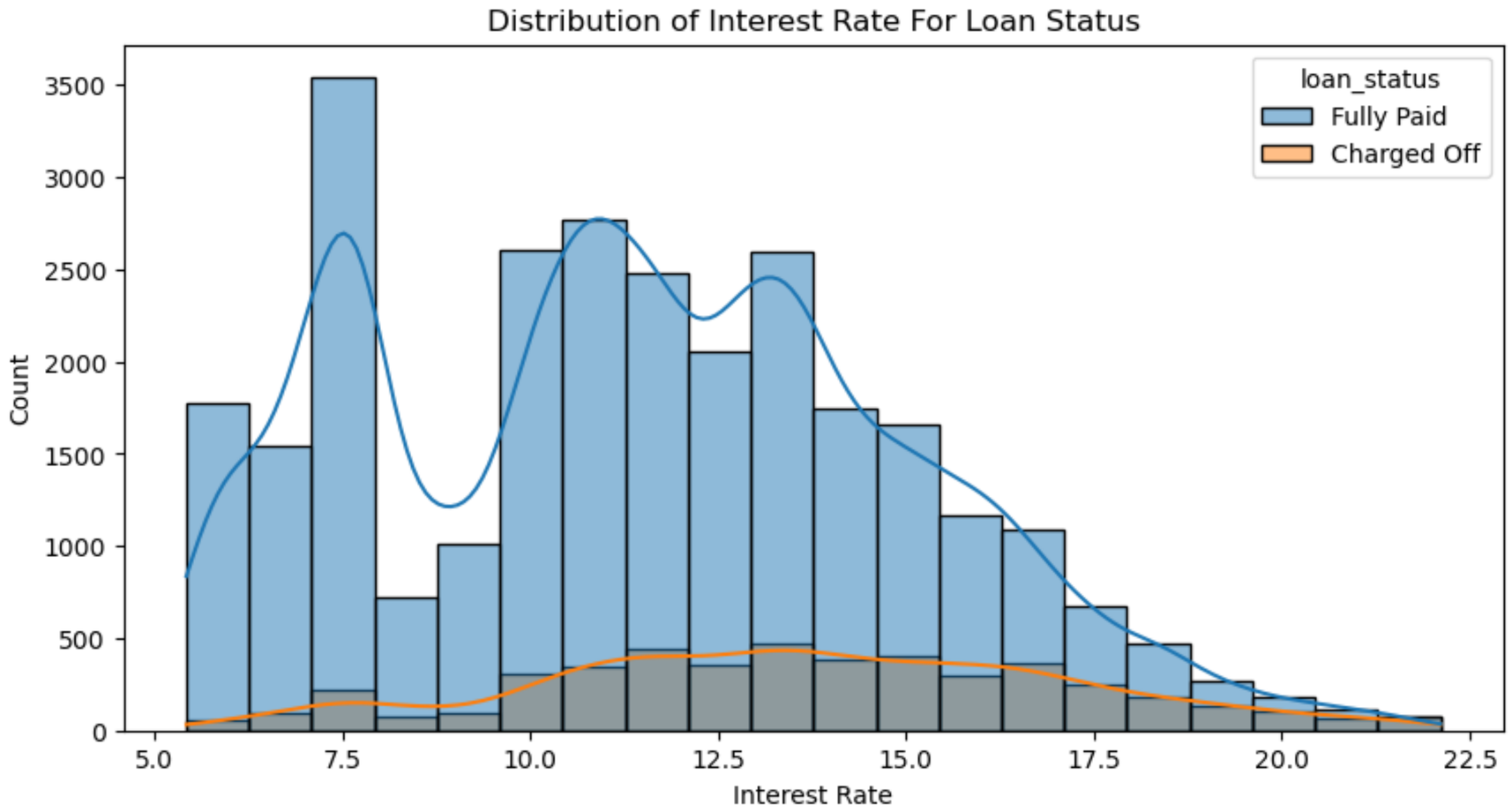
Bivariate Analysis

- Larger the loan amount, greater the chances of default.



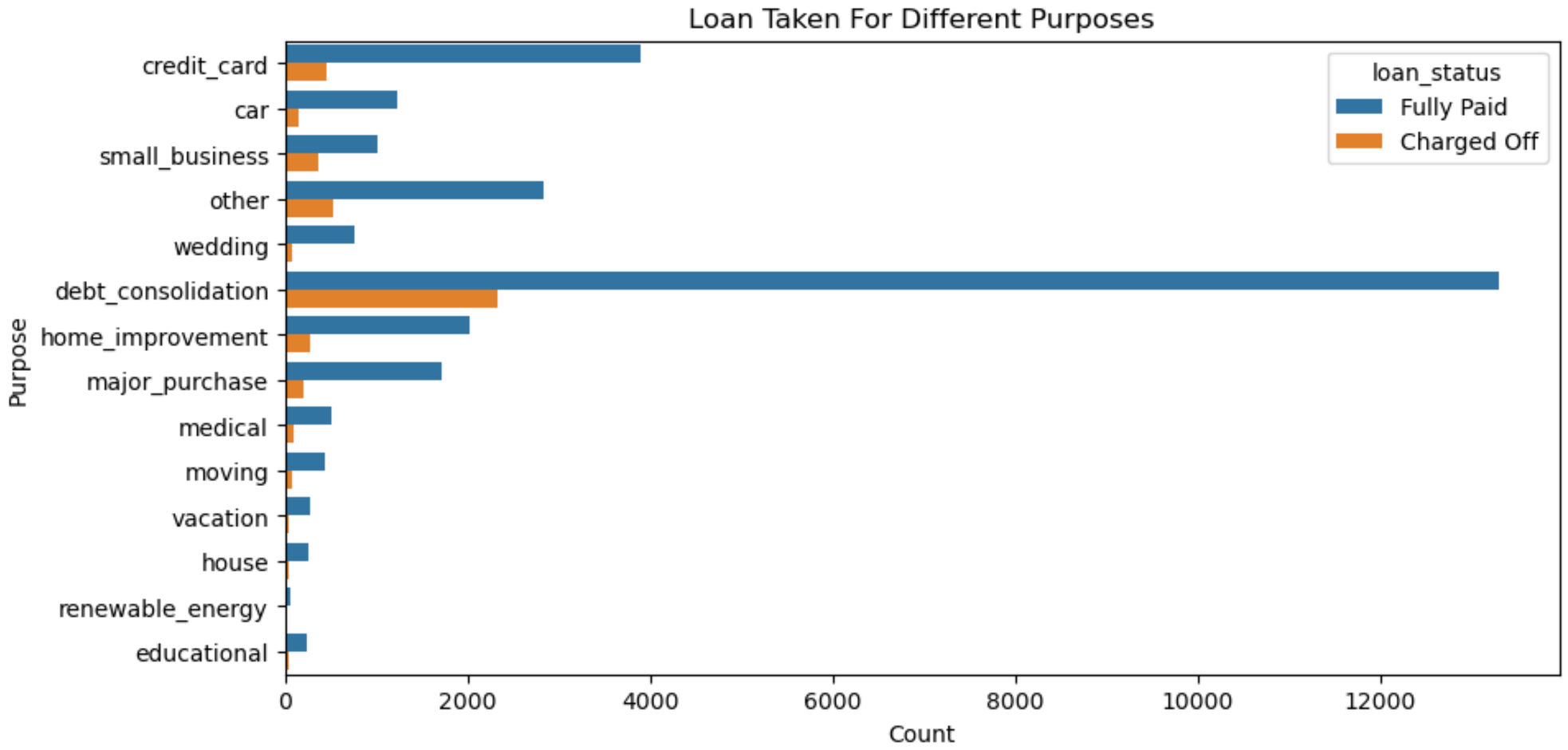
Bivariate Analysis

- Applicants with Interest Rate between 10% to 15% are more likely to default.



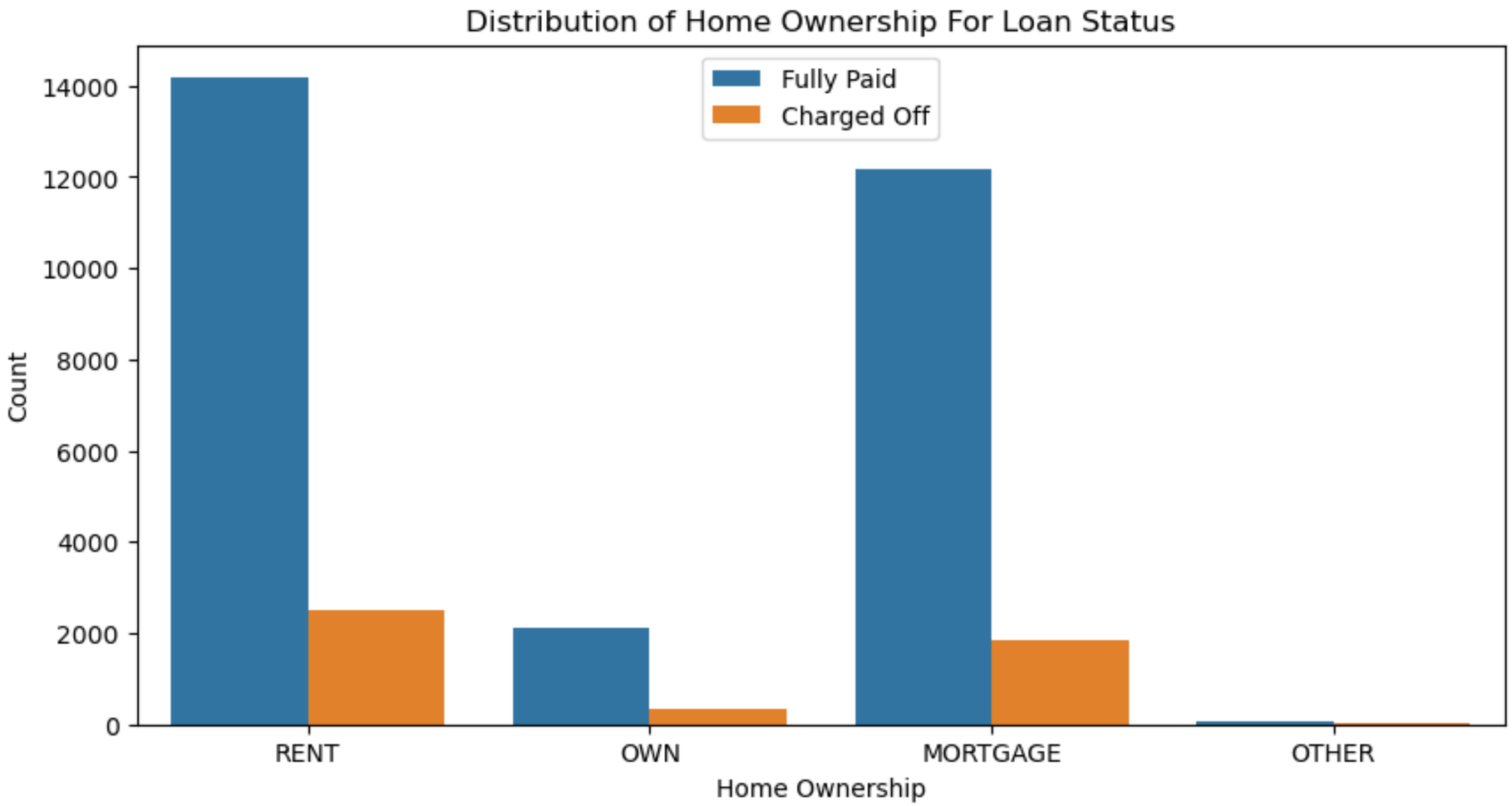
Bivariate Analysis

- People who take loan for debt consolidation are likely to charged of.



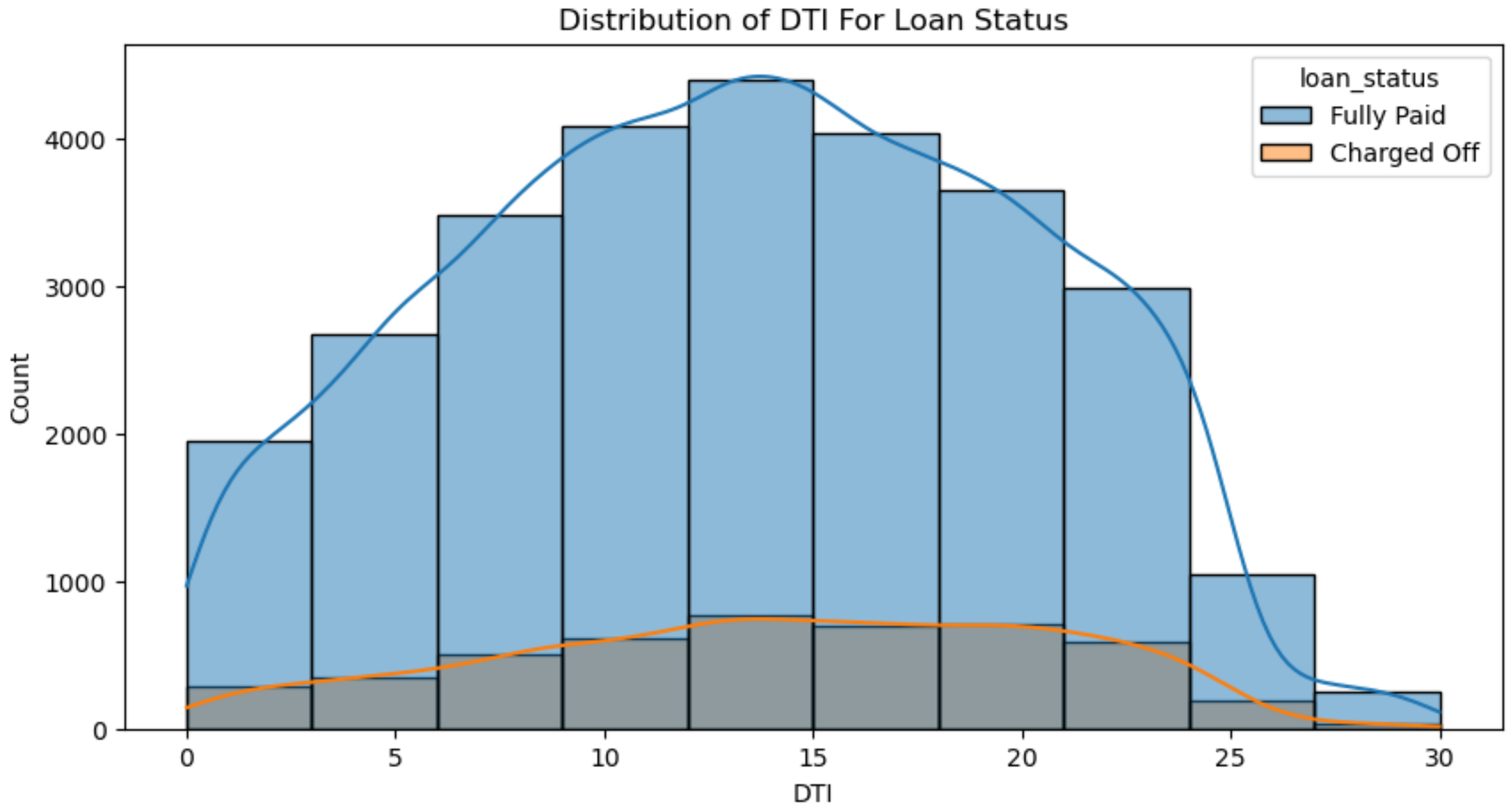
Bivariate Analysis

- People not having their own property are likely to charged off.



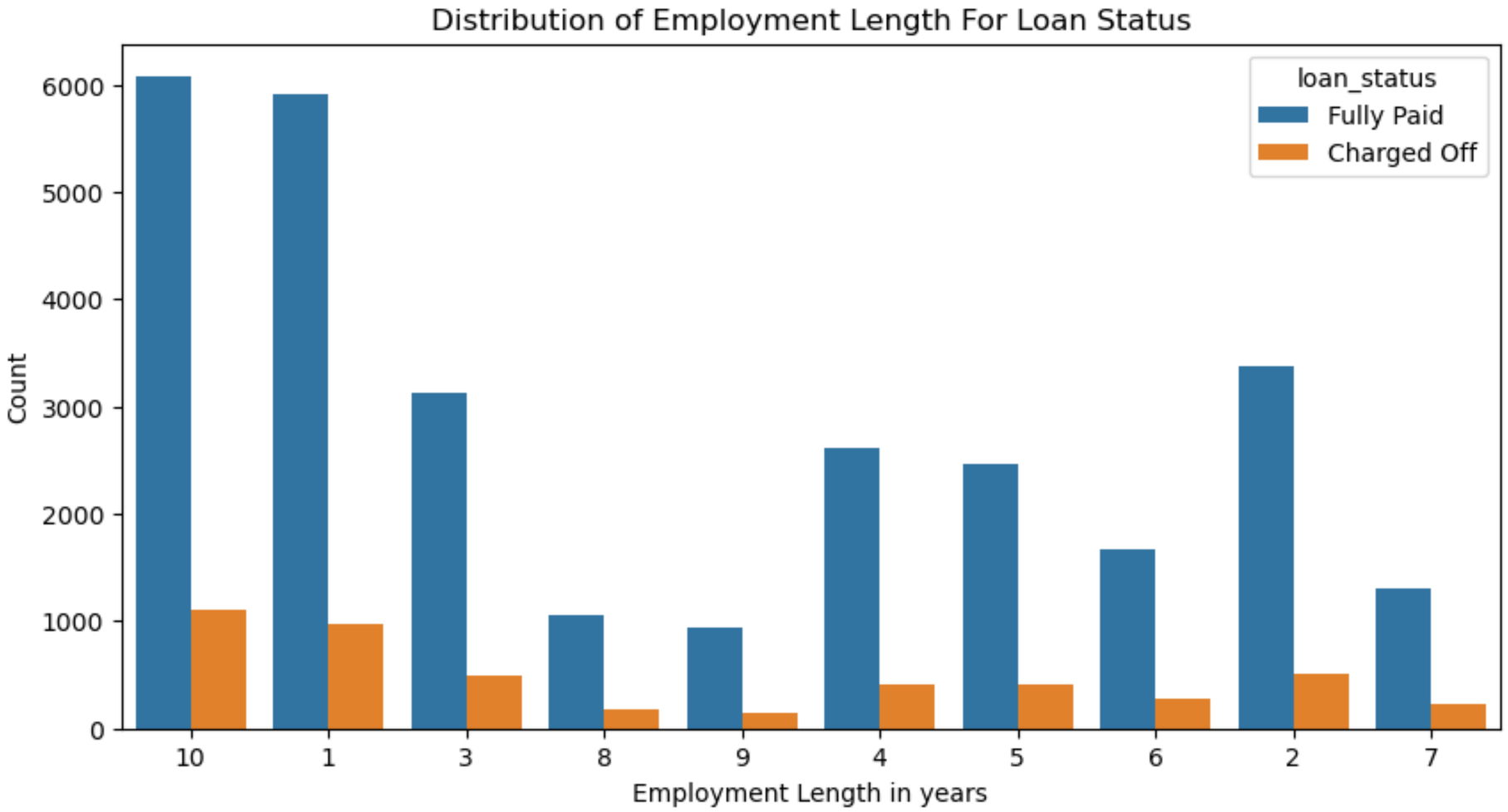
Bivariate Analysis

- Data shows high defaulters between 10 to 20 DTI range but High DTI has High charged off chances.



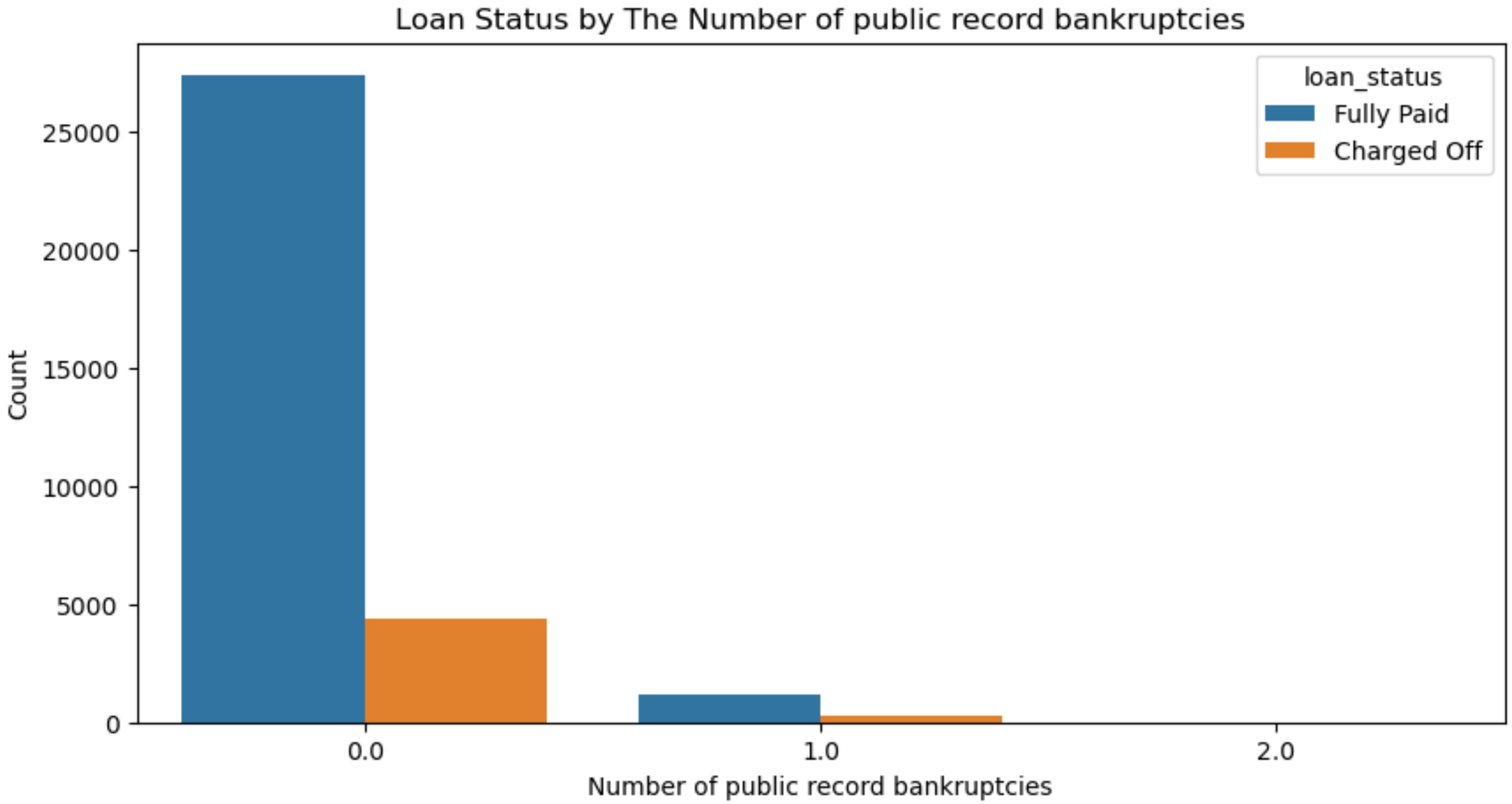
Bivariate Analysis

- Applicants with more than 10 years of experience have high default and as well paid cases.



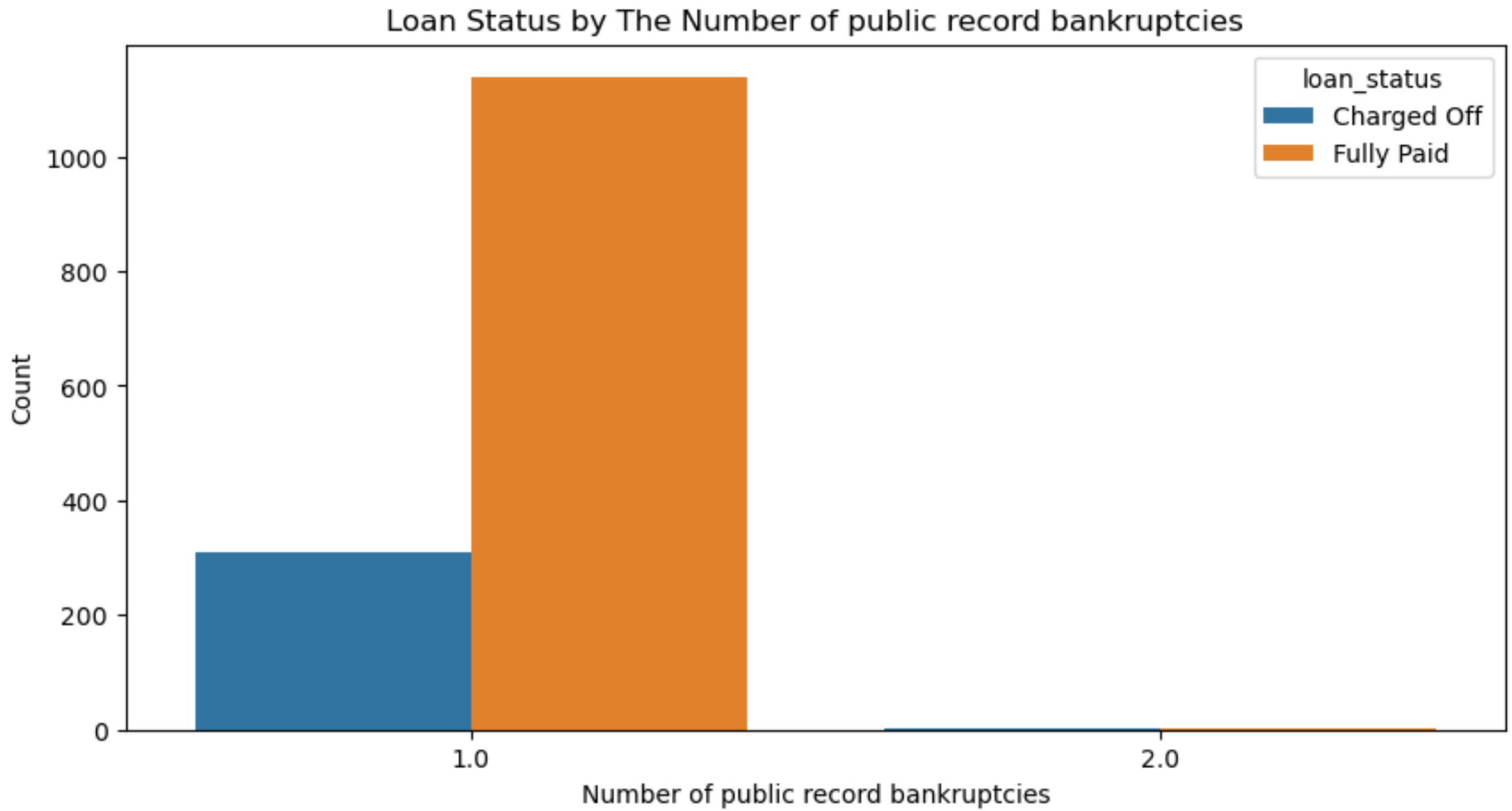
Bivariate Analysis

- Applicants with bankruptcy records are less likely to pay the loan back.



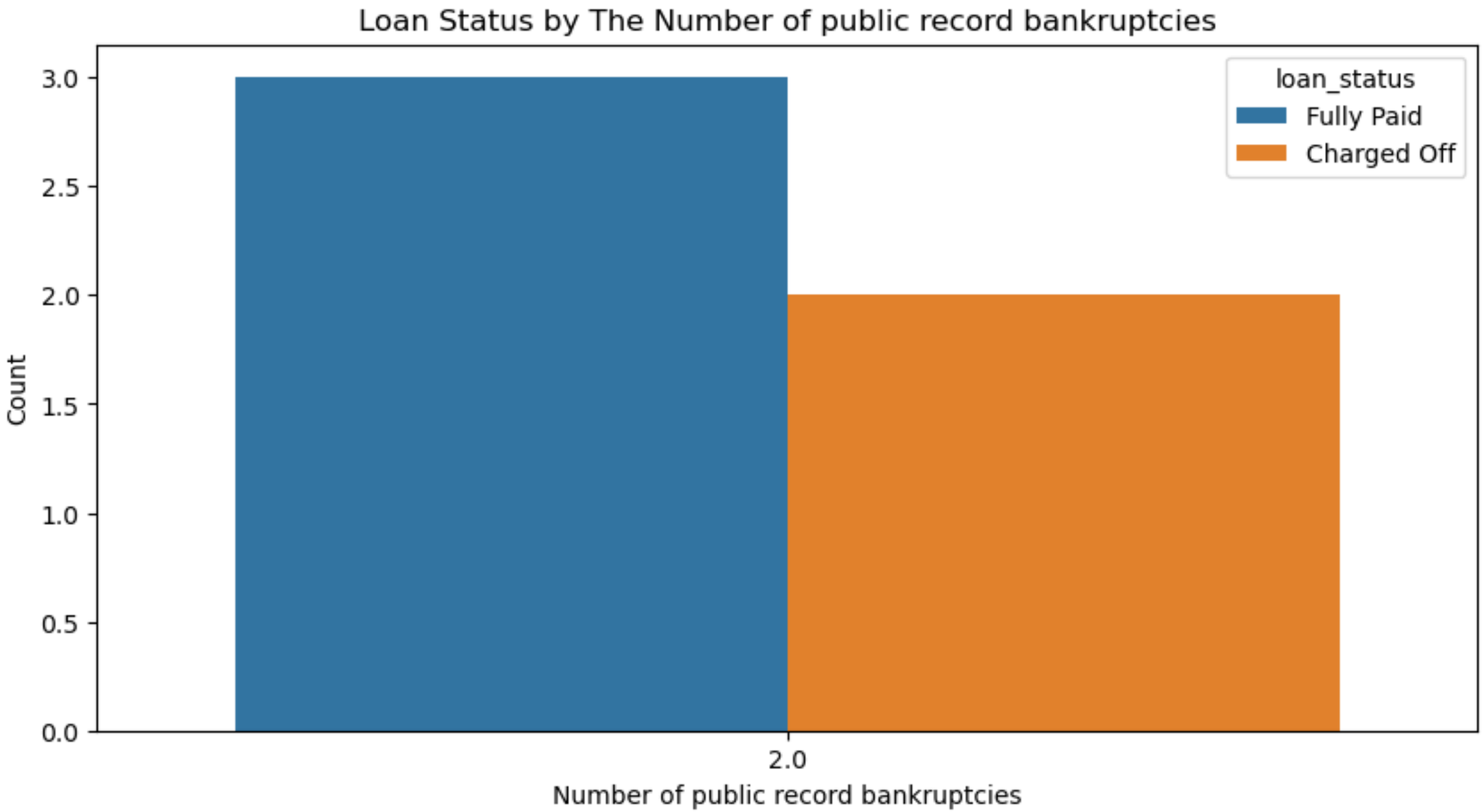
Bivariate Analysis

- Applicants with bankruptcies records are more likely to be charged of.



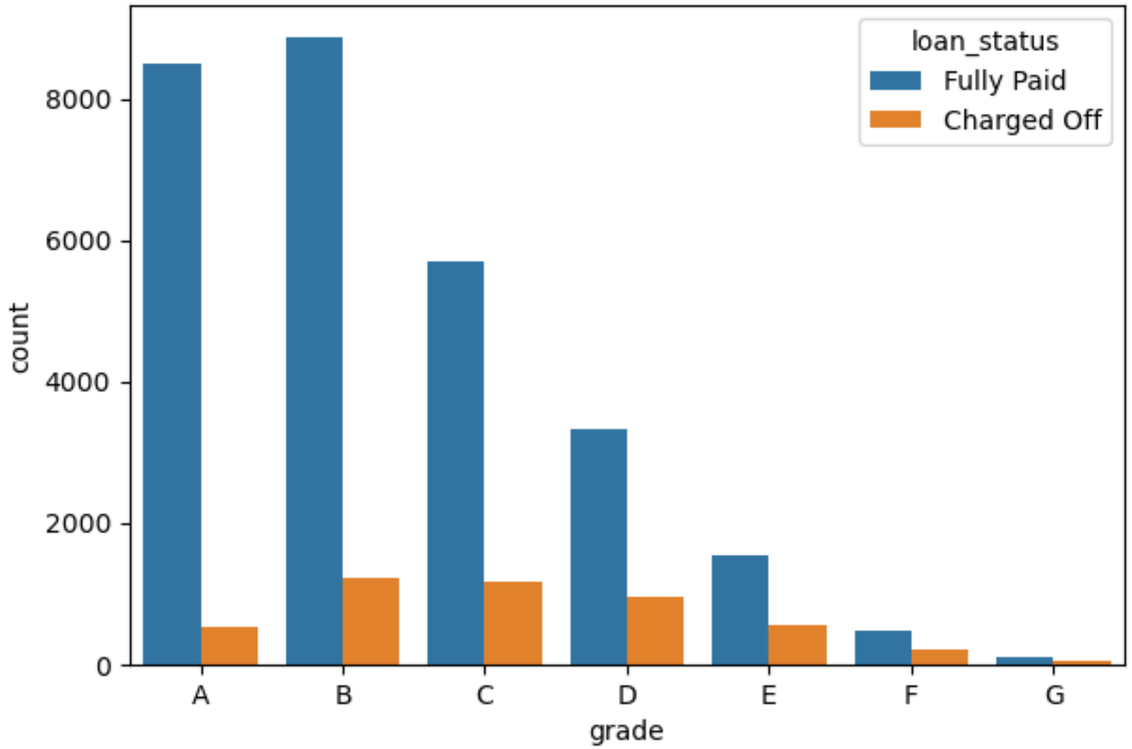
Bivariate Analysis

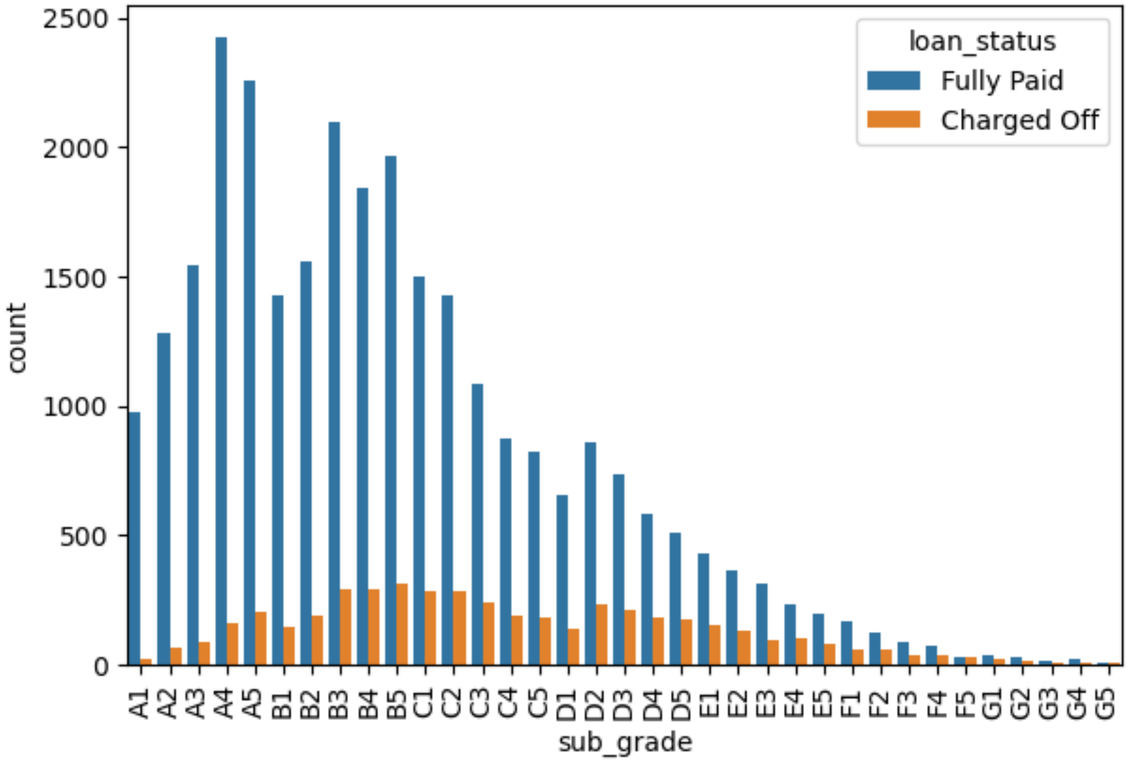
- Applicants with bankruptcies records are more likely to be charged of.



Bivariate Analysis

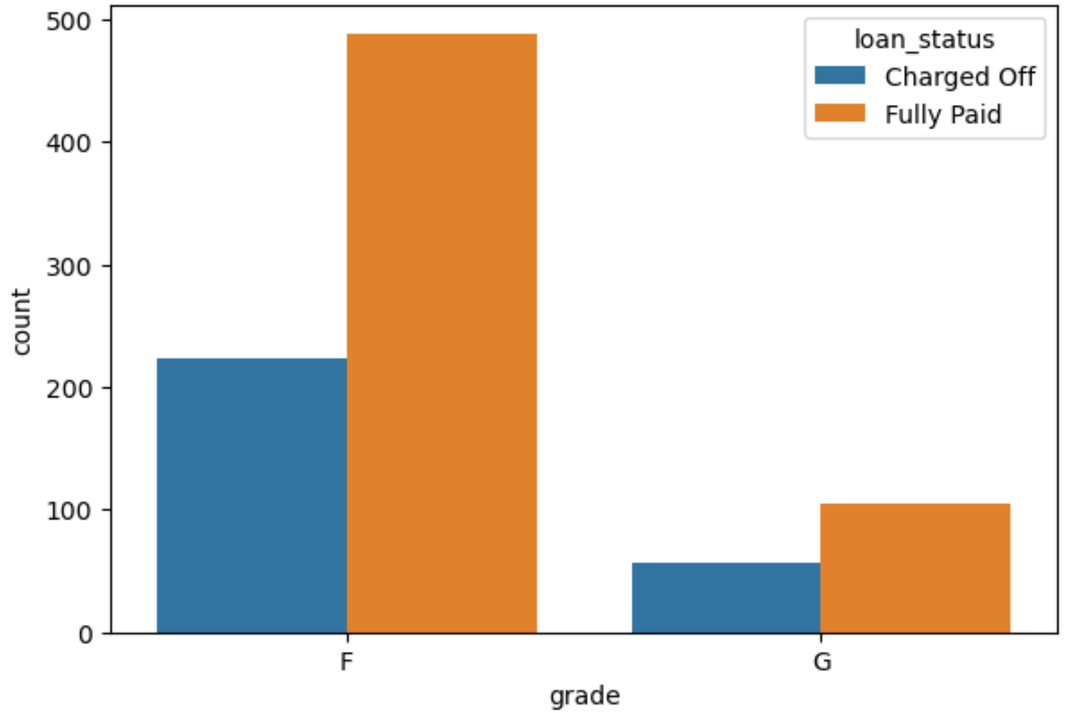
- Comparison of grade and sub-grade over loan status.

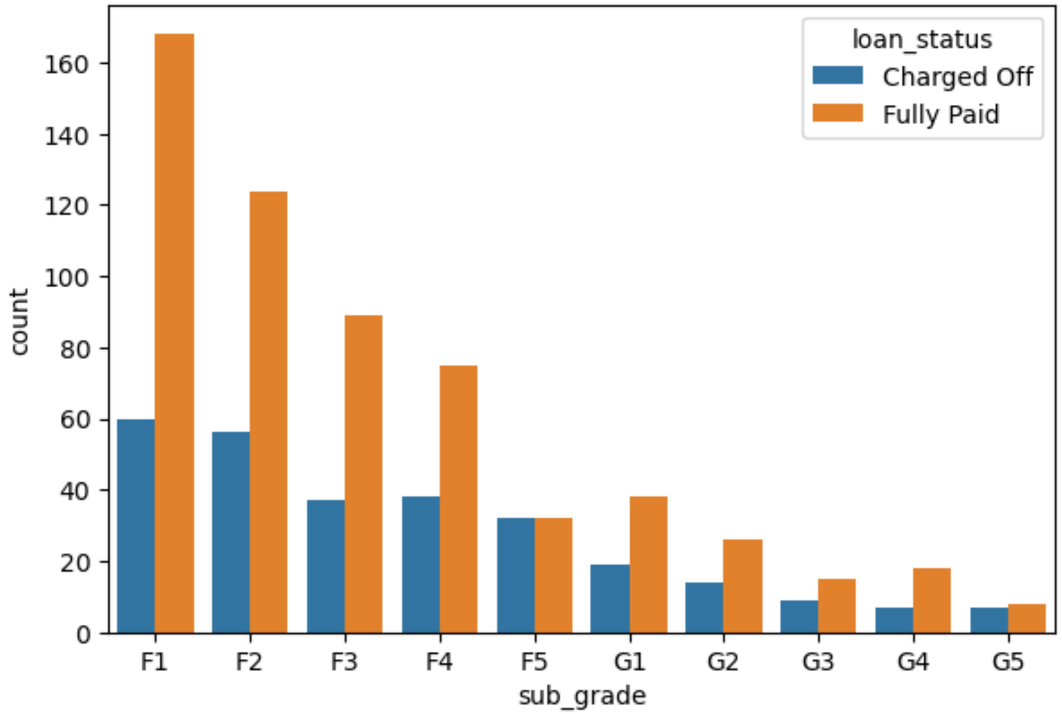




Bivariate Analysis

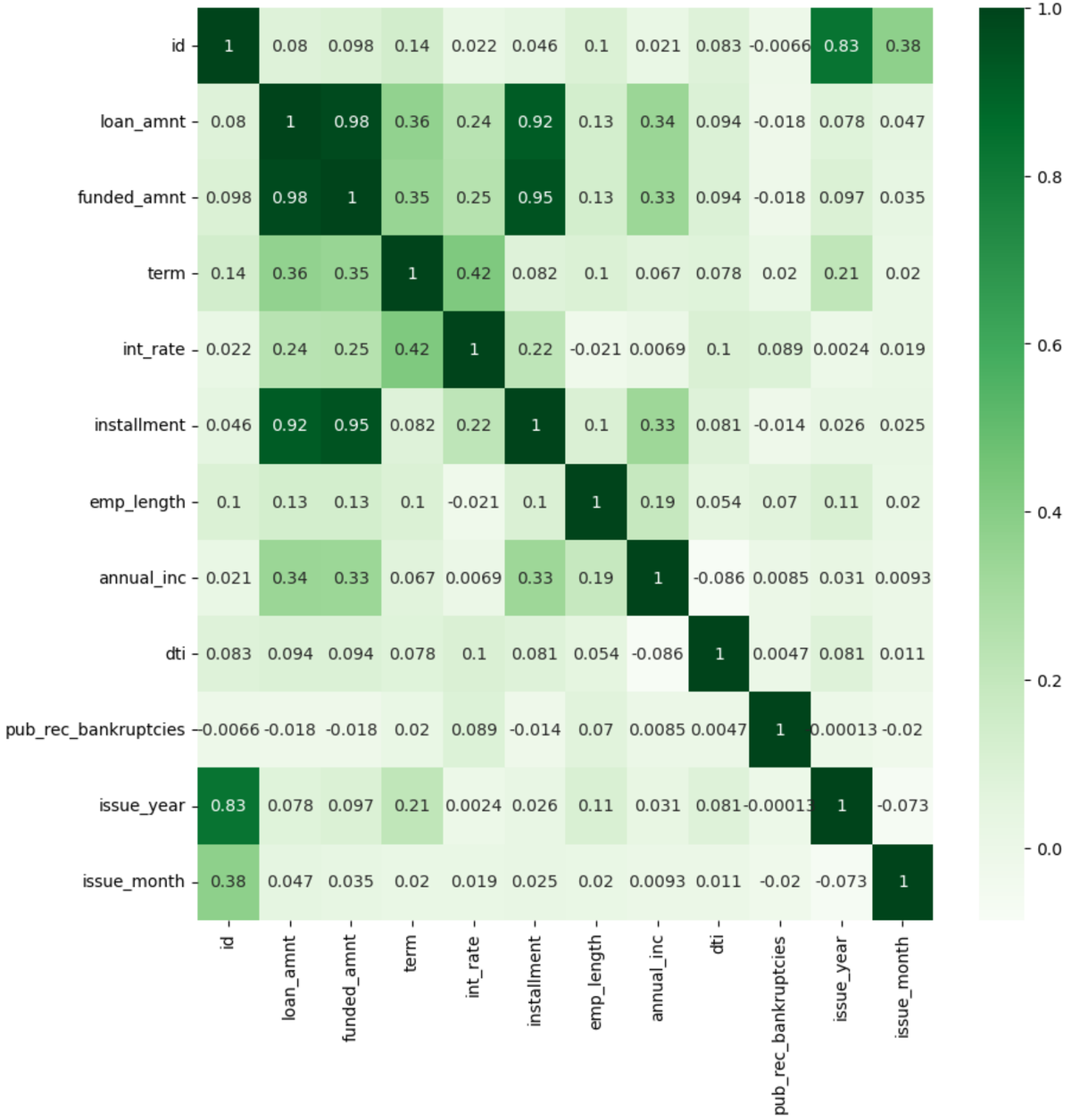
- Grade F and G loans are less likely to be paid back.
- Separating and visualising F and G grades, sub-grades.





Correlation Matrix

- Loan amount and annual income have positive correlation.
- Bankruptcies have negative correlation with loan amount.
- Funded amount has negative correlation with Bankruptcies.
- Loan amount and instalments are highly positively correlated.
- Employement length has postive correlation with annual income.



Driving Factors for Loan Defaults

- **Income Level:**
  - Applicants with annual incomes in the range of 40K are more likely to default compared to those earning between 80K.
- **Employment Stability:**
  - Those with 10 or more years of employment have a higher likelihood of borrowing, indicating stability, but paradoxically, they also show a significant risk of default.
- **Loan Amount:**
  - There is a clear trend where larger loan amounts correlate with a higher probability of default.
- **Interest Rates:**
  - Applicants with interest rates between 10% and 15% exhibit a greater tendency to default.
- **Purpose of Loan:**
  - Individuals taking loans for debt consolidation are more frequently associated with defaults.

Driving Factors for Loan Defaults

- **Property Ownership:**
  - Applicants without their own property are more likely to default, suggesting that home ownership may serve as a financial safety net.
- **Debt-to-Income Ratio (DTI):**
  - High default rates are observed within the 10 to 20 DTI range, although individuals with very high DTI ratios may have been filtered out beforehand or rejected prior to application.
- **Bankruptcy History:**
  - Individuals with bankruptcy records are significantly more likely to default on loans.
- **Loan Grades:**
  - Loans graded F and G have a higher propensity for defaults, indicating that lower-quality loans carry greater risk.

Conclusion

- Income, employment stability, loan amount, interest rate, property ownership, debt, and credit history are key determinants of loan default risk.
- Larger loans, higher interest rates, and borrowers with lower incomes or poor credit profiles pose greater challenges.
- Lenders can leverage these insights to refine underwriting models and risk management strategies.