Lending Club Case Study

Parvinder Kaur

Outline

Problem Statement

Data Cleaning

Datatype Conversions

Removing Outliers

Univariate Analysis

Problem Statement

- **To understand the driving factors behind loan default for a Consumer Finance Company for its portfolio and risk assessment.**
- **Business Understanding:**
  - 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 Outcome 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).

Data Cleaning

- **Columns:**
  - 54 Columns had 100% null values, removed those columns.
  - Dropped two descriptive columns ('desc' and 'title') since they do not impact the outcome.
  - Eliminated 9 columns with a single value, making them non-quantifiable.
  - Excluded 12 behavioral columns that were either unique or irrelevant for the outcome.
  - 'member_id' column has been removed; 'id' is now the primary key for analysis.
- **Rows:**
  - No duplicate rows found in the dataset.
  - Dropped 1140 rows with a loan status of 'Current,' as they are still in progress and not fully paid nor defaulted.
- After data cleaning step, the dataset now consists of 38577 rows and 34 columns.

Datatype Conversions, Formatting & Derived Columns

Conversion & Formatting

- Rounded all floating-point numbers in the dataset to two decimal points for consistency.
- Formatted the existing **'issue_d**' column to the date data type.
- Removed the percentage symbol from the '**int_rate**' and **'revol_util**' columns, converting them to the float data type for numerical analysis.

Derived Columns

- Extracted the **Year** and **Month** components from the existing **'issue_d**' date column to facilitate temporal analysis.
- Generated data **buckets** for essential numeric columns—specifically, **'annual_inc**,' **'loan_amnt**,' **'int_rate**,' and **'dti**.' These buckets will be utilized in Univariate analysis for a more granular examination of the distribution and patterns within these key numeric features.

Removing Outliers

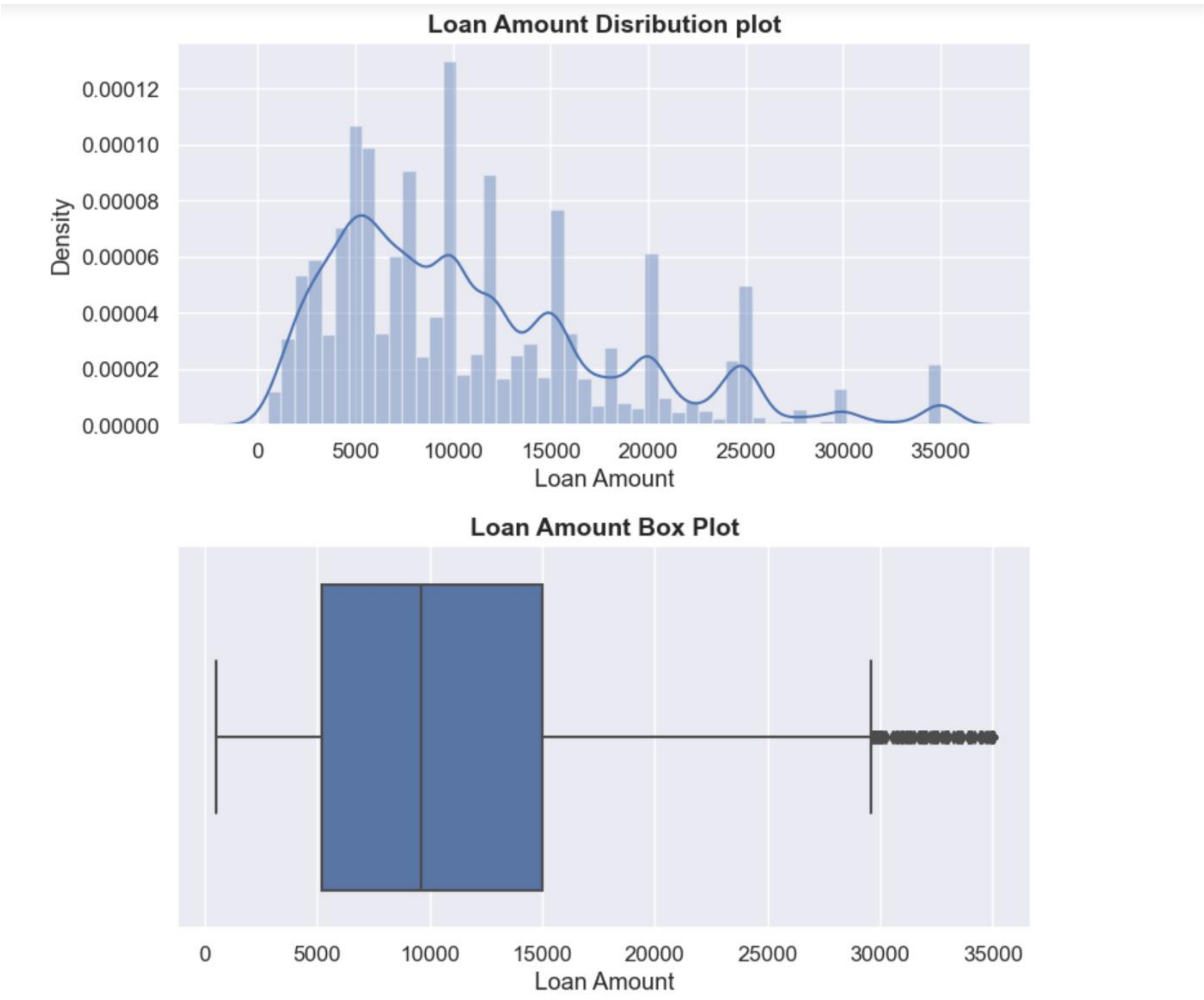
- Outliers have been identified in the annual income field, with 171 borrowers reporting annual incomes exceeding 300k. Given the relatively low number of such instances, their presence could potentially distort the overall outcome. Consequently, these data points have been excluded from the analysis.

  - annual_inc
  - loan_amnt
  - int_rate
  - dti

Univariate Analysis

- In the Lending Club case study, the examination of individual variables in isolation through univariate analysis is essential for predicting loan defaulters. This approach systematically reviews each variable independently, aiding in the identification of patterns, trends, and potential outliers that may signal default risk. Both boxplots and bar plots are used together in the univariate analysis process.

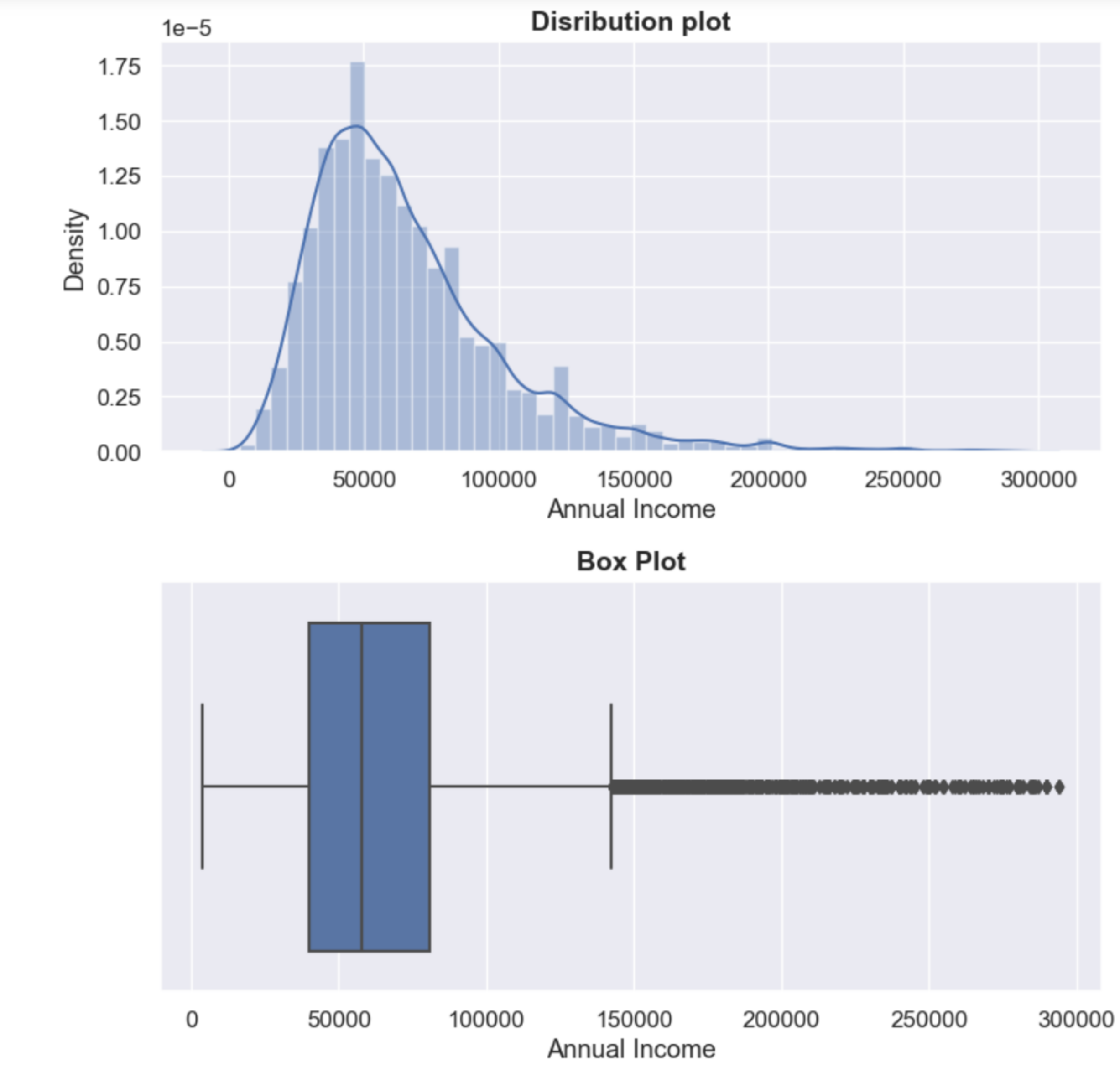
Loan Amount



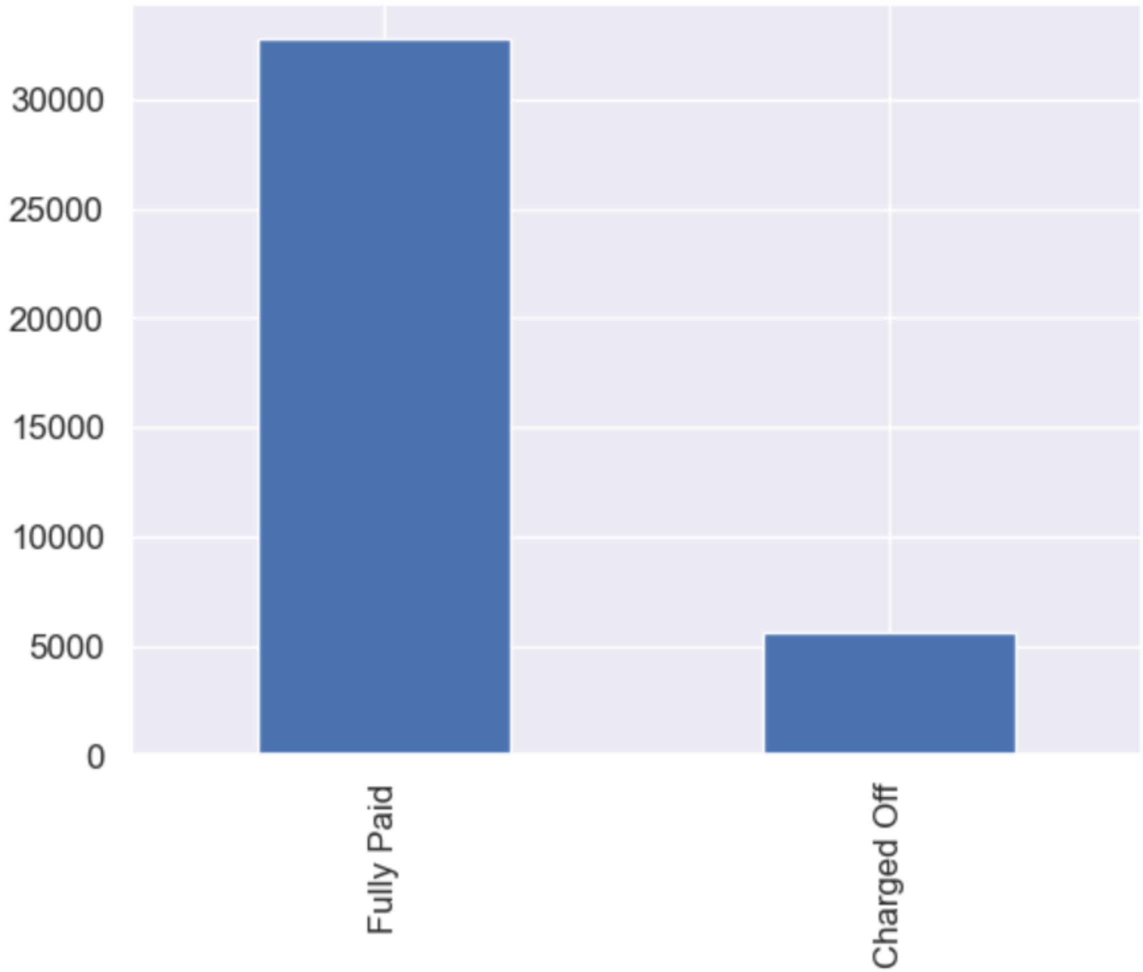
Loan Amount Disribution plot

Loan Amount Box Plot

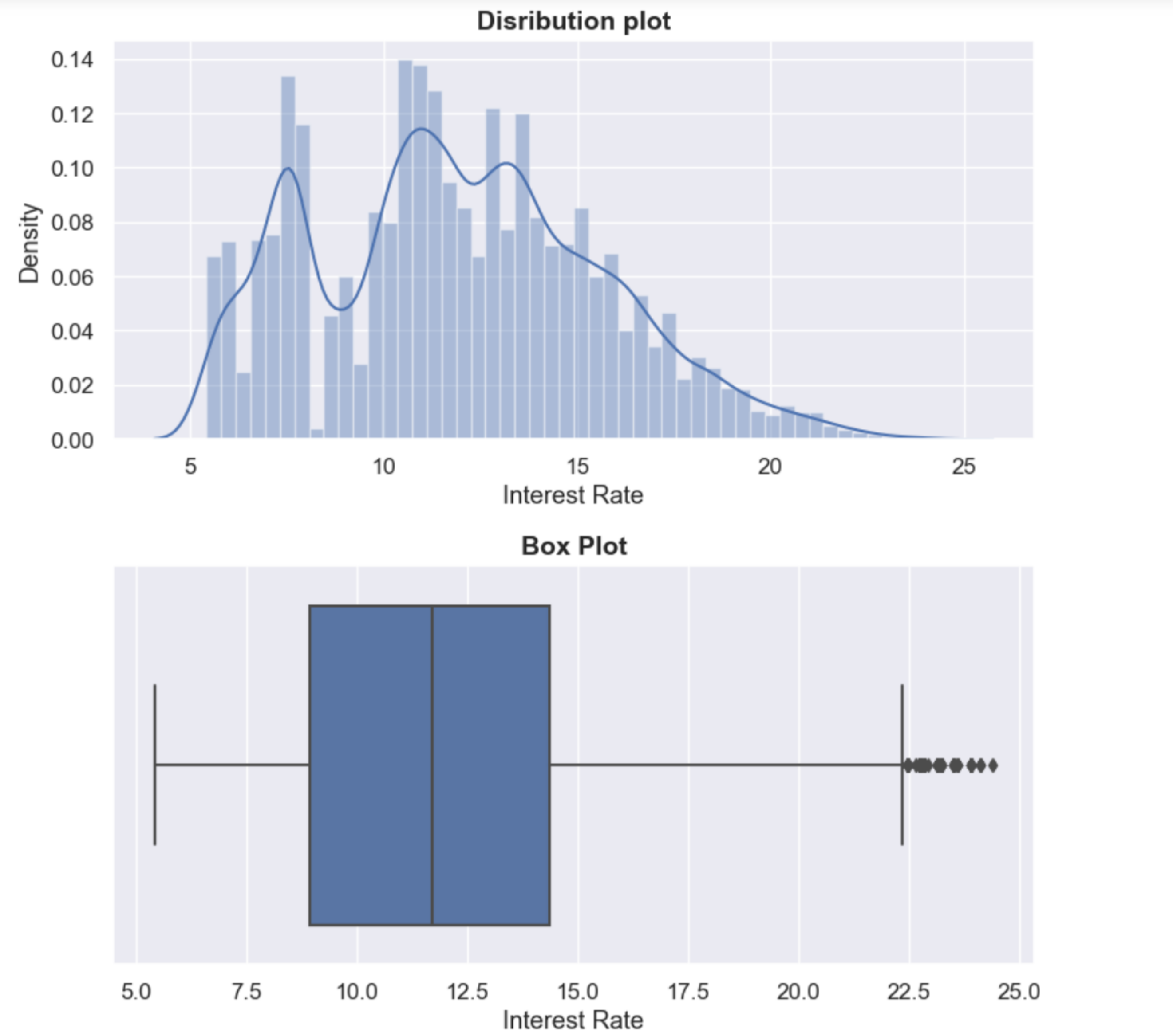
Annual Income



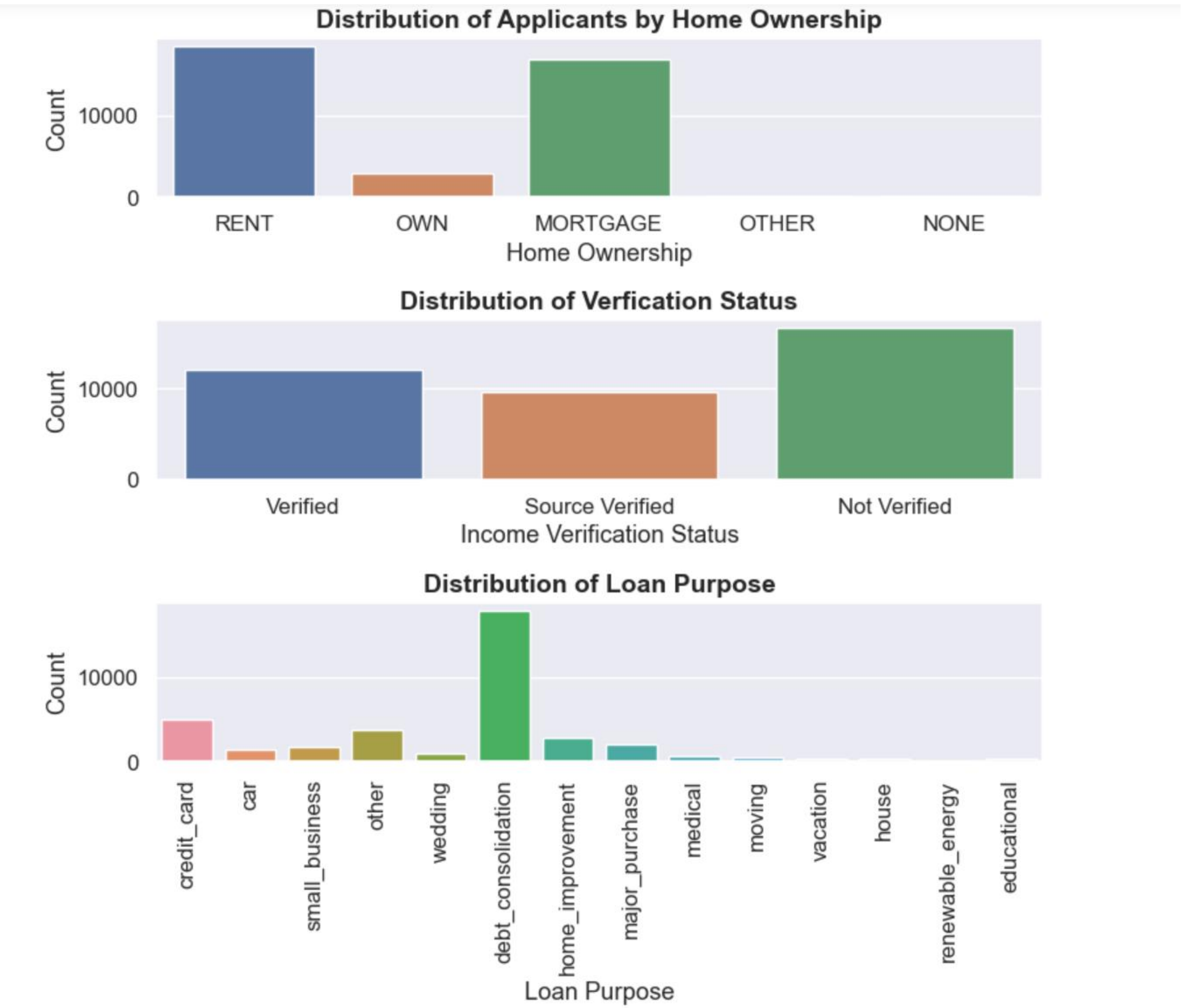
Loan Status



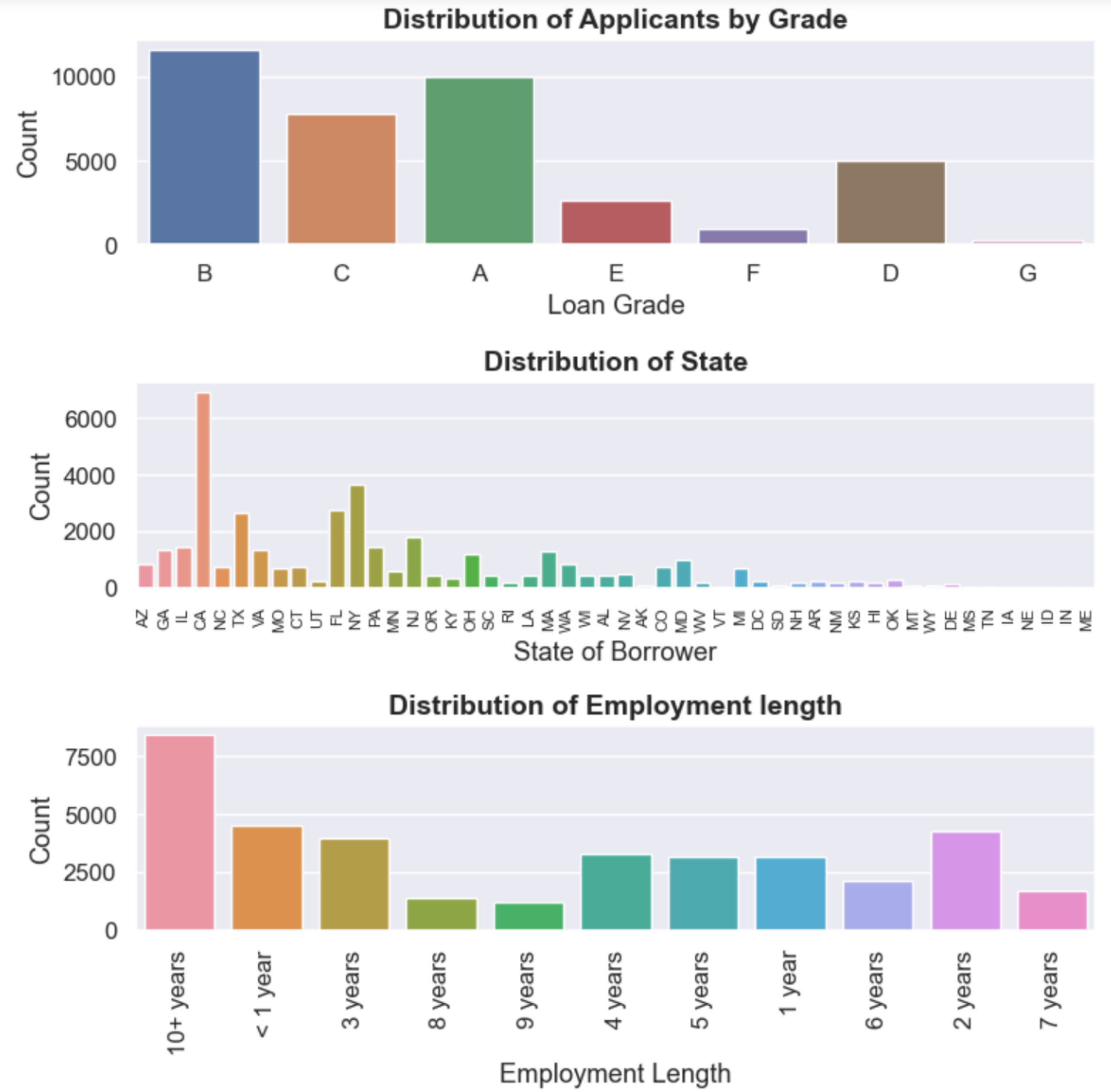
Interest Rate



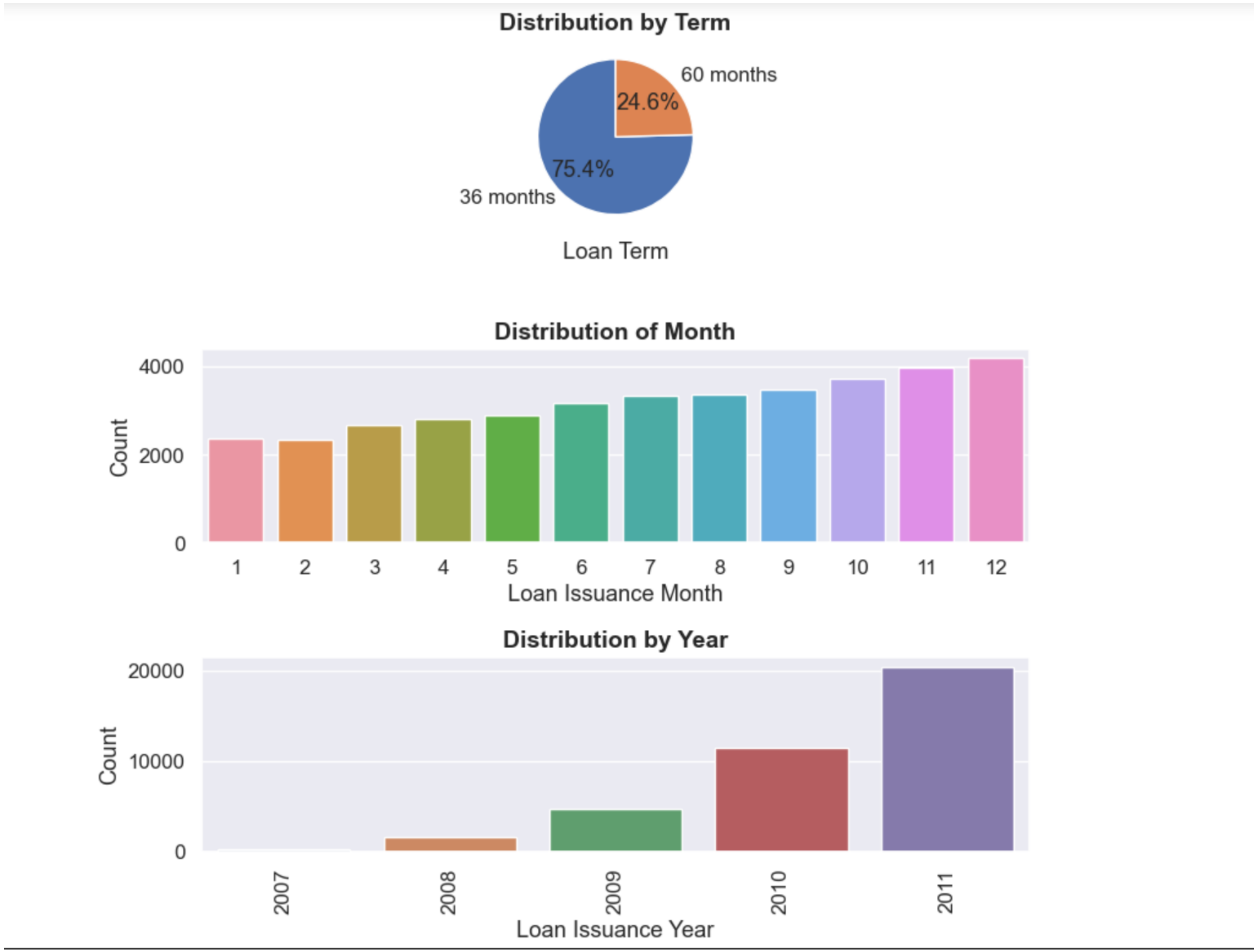
Categorical Variables



Categorical Variables



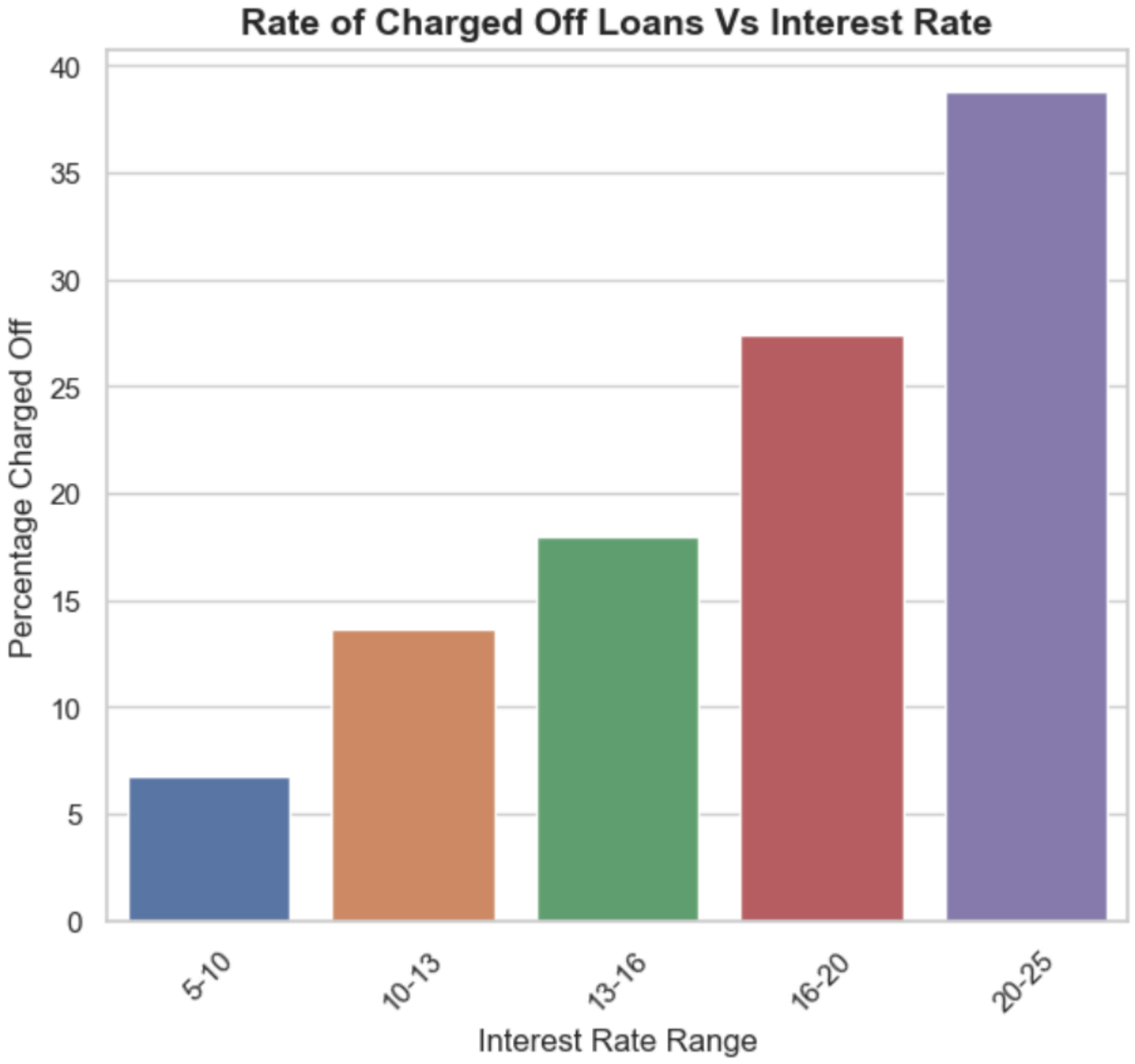
Categorical Variables



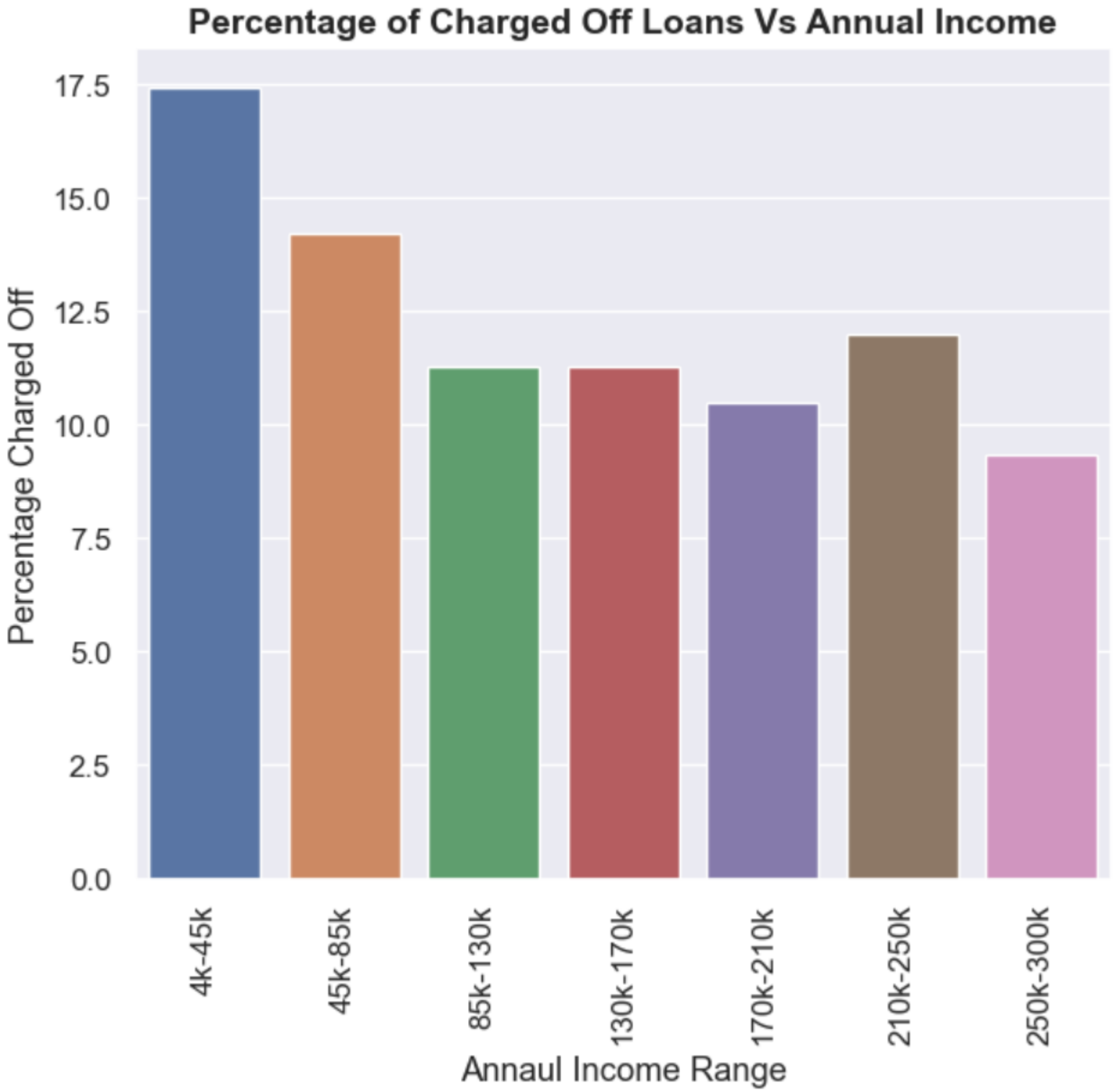
Bivariate Analysis

- In the Lending Club case study, bivariate analysis is employed to uncover relationships between crucial variables. This involves scrutinizing pairs like interest rate and loan status, income and loan amount, debt-to-income ratio and loan status, among others. Through this approach, it offers insights into potential correlations that may impact loan default.

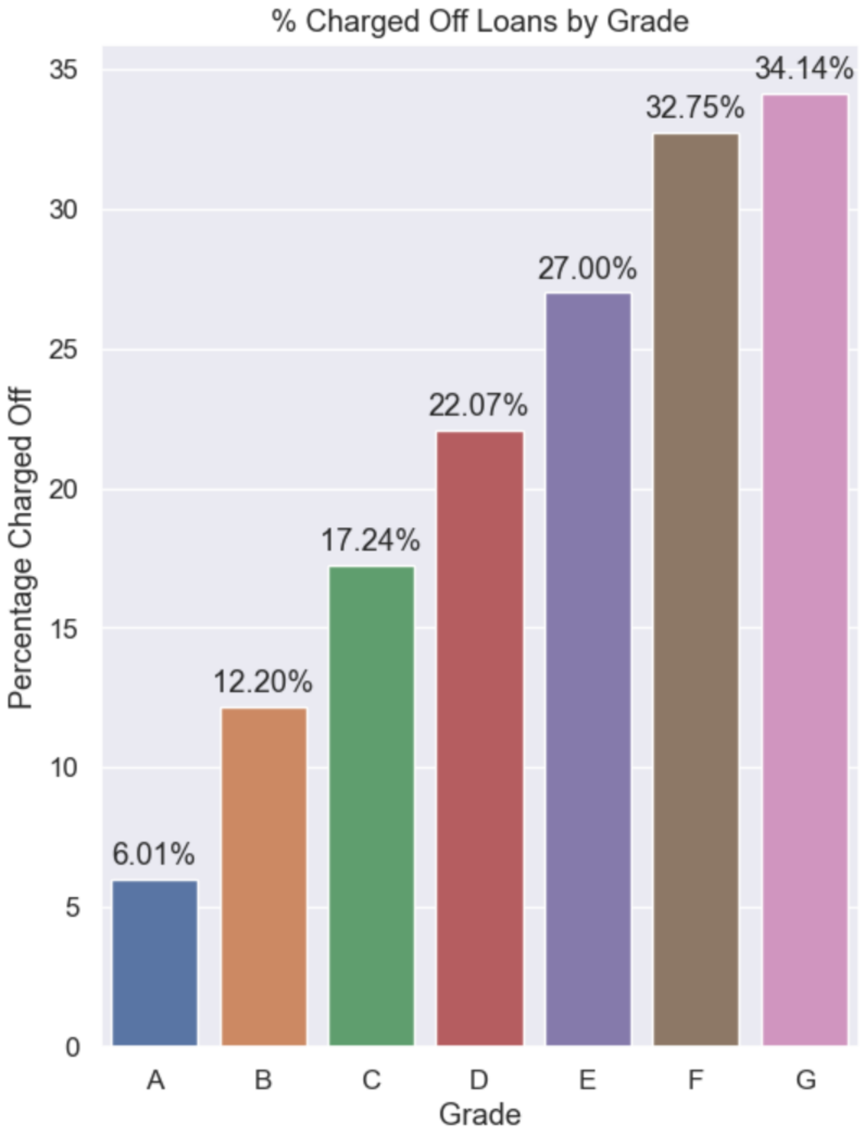
Interest Rate v/s Charged Off



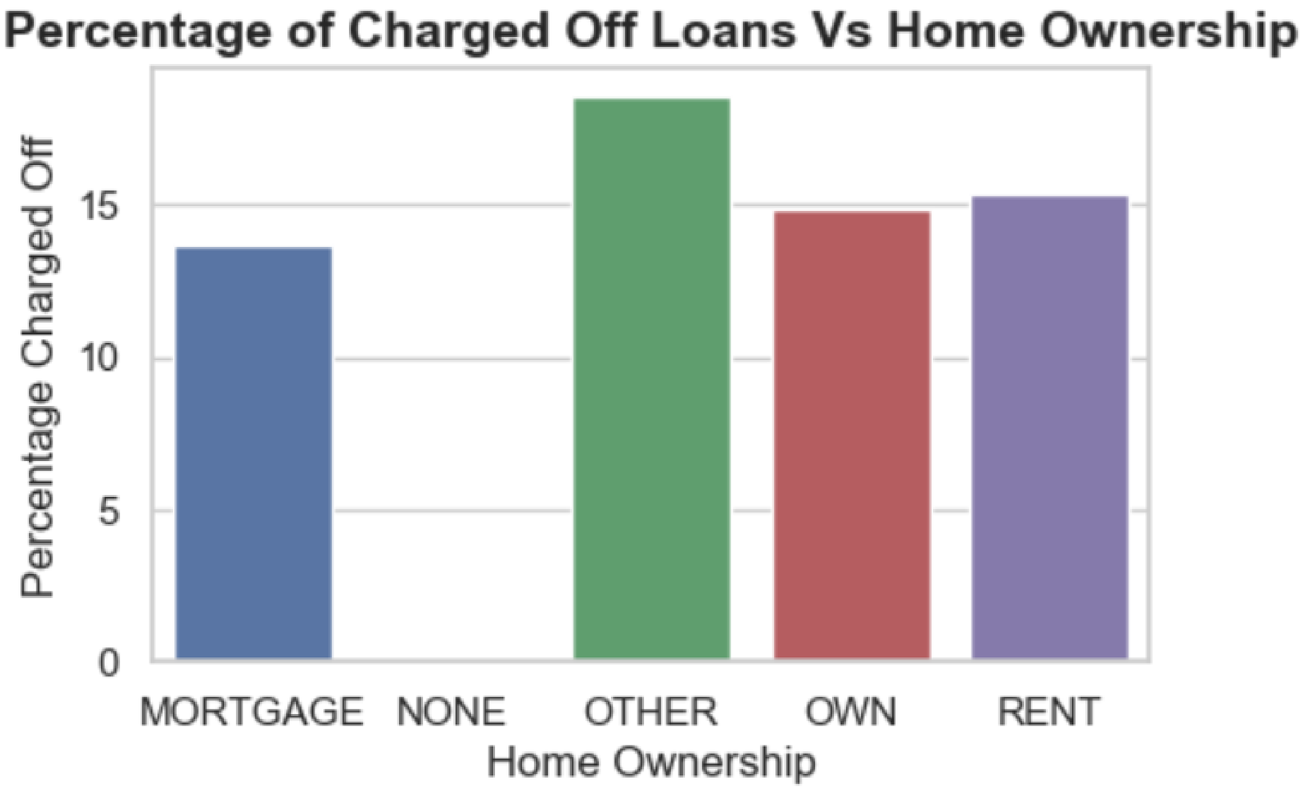
Annual Income v/s Charged Off



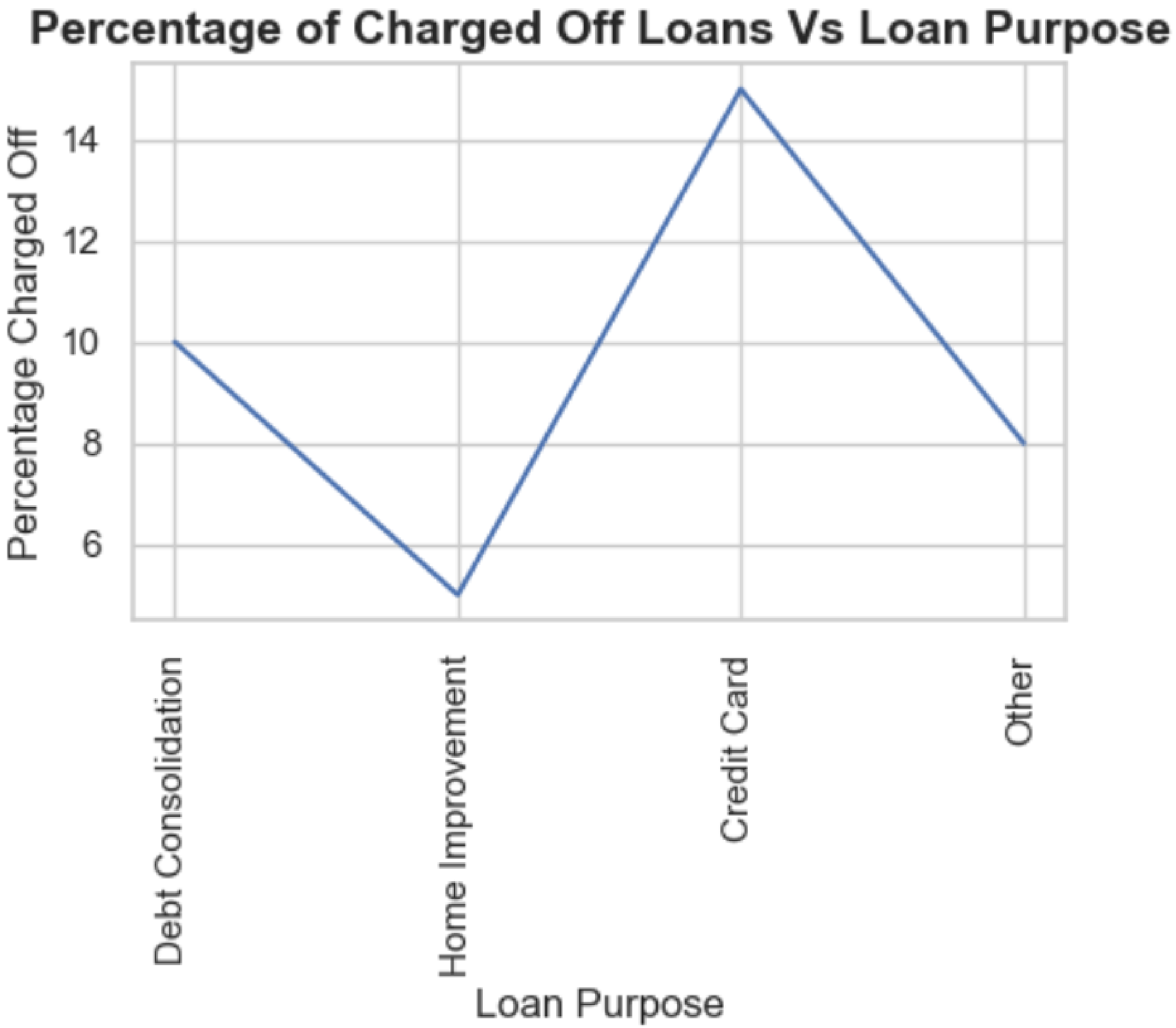
Grade v/s Charged Off



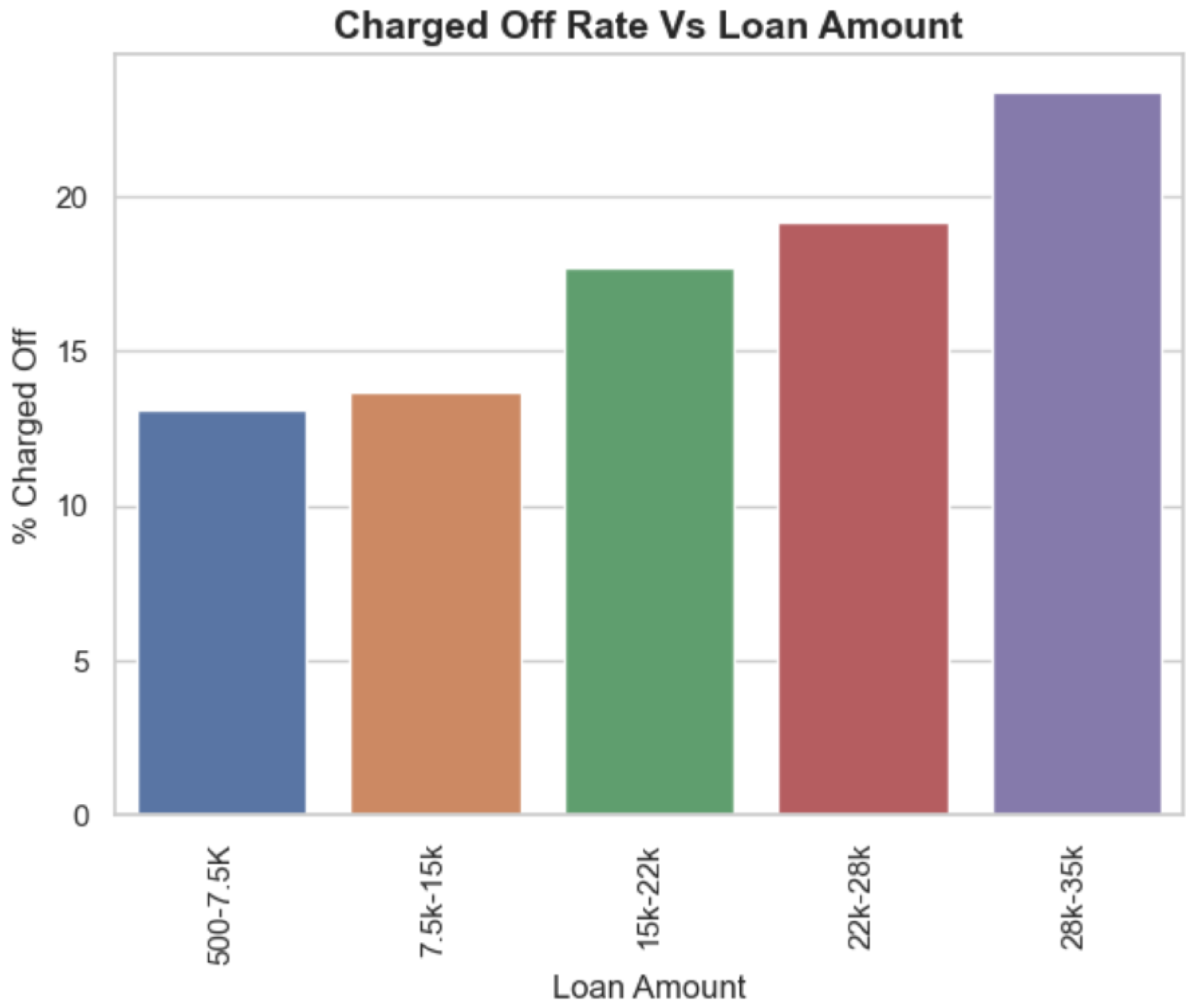
Home Ownership v/s Charged Off



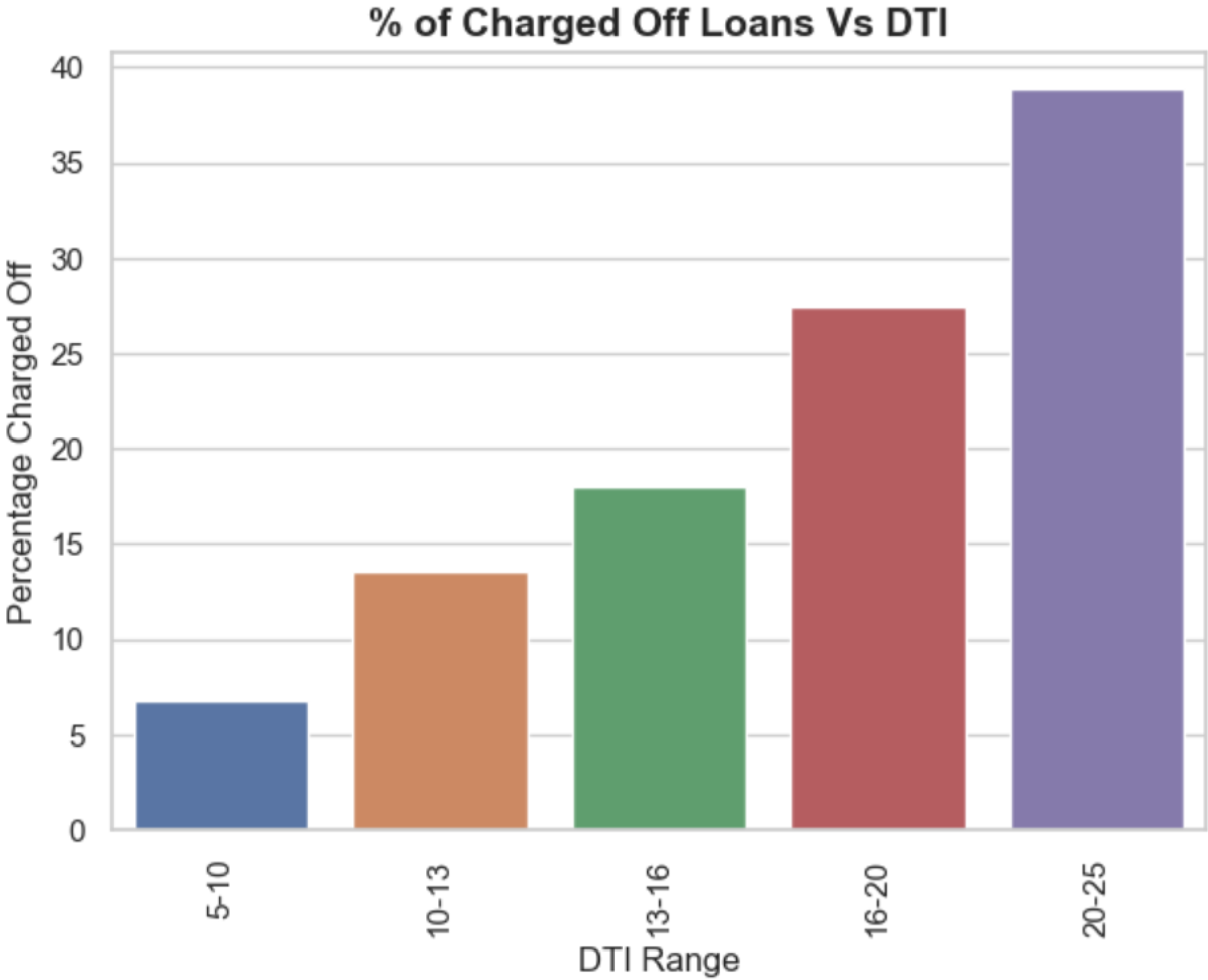
Loan Purpose v/s Charged Off



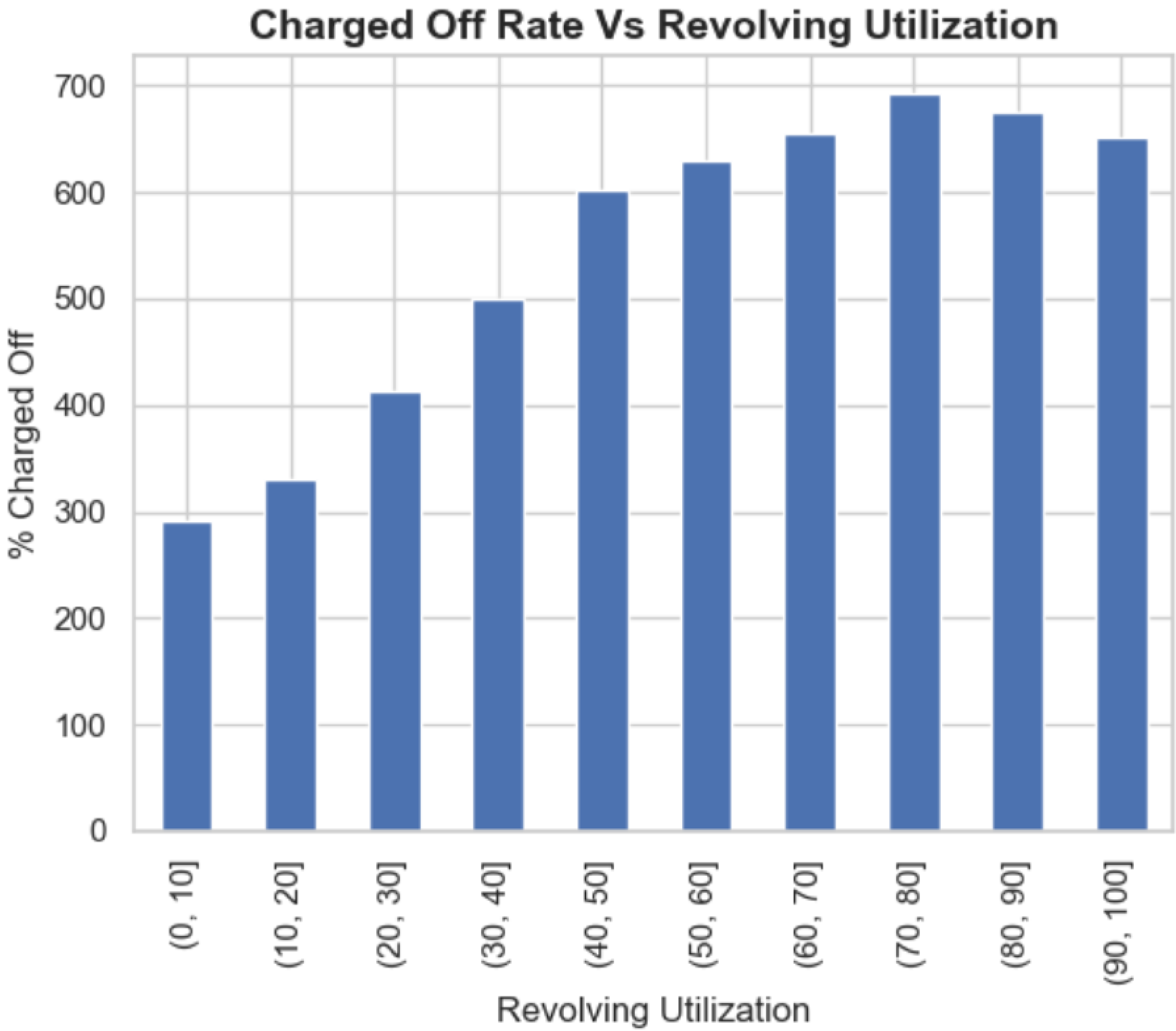
Total Loan Amount v/s Charged Off



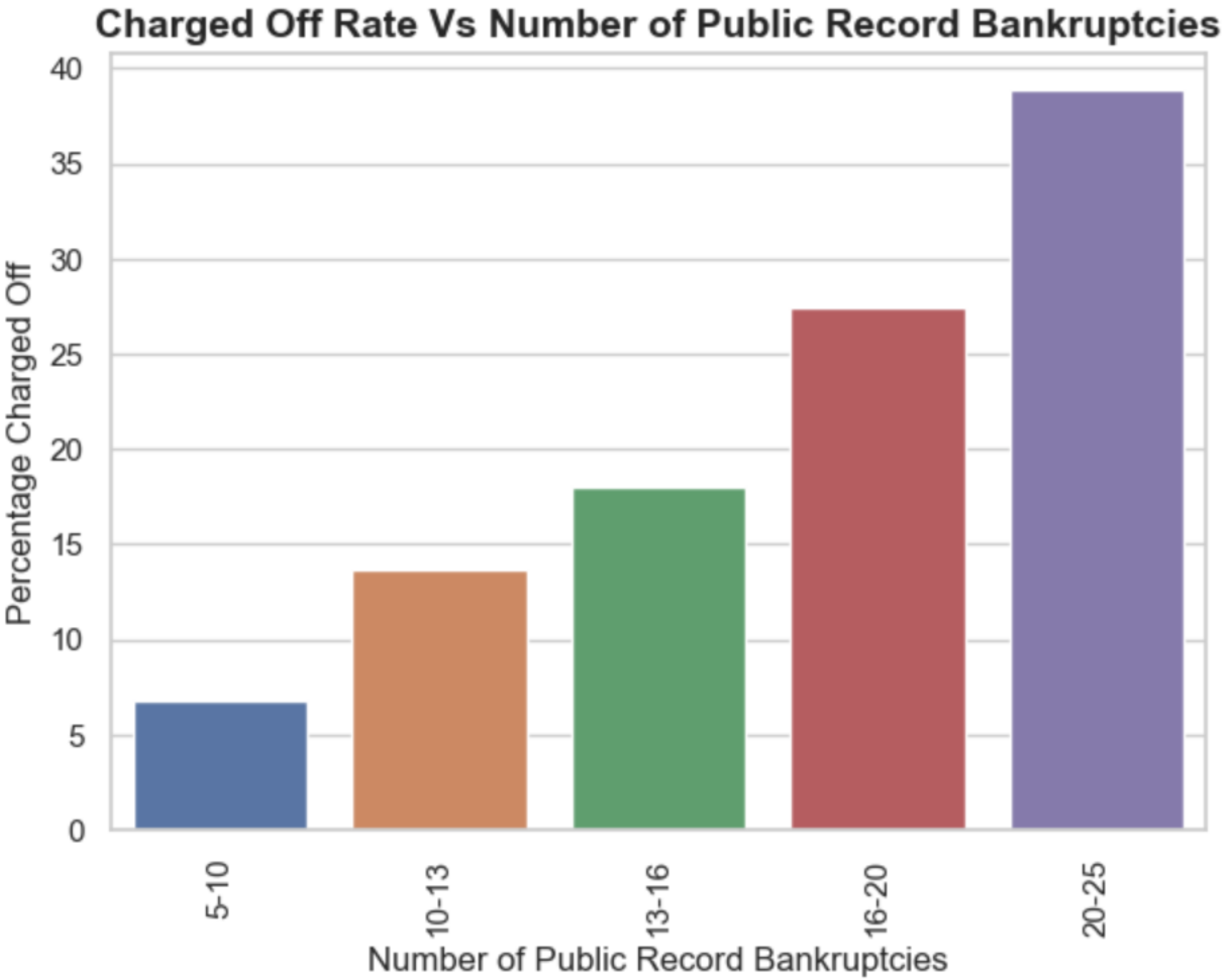
DTI v/s Charged Off



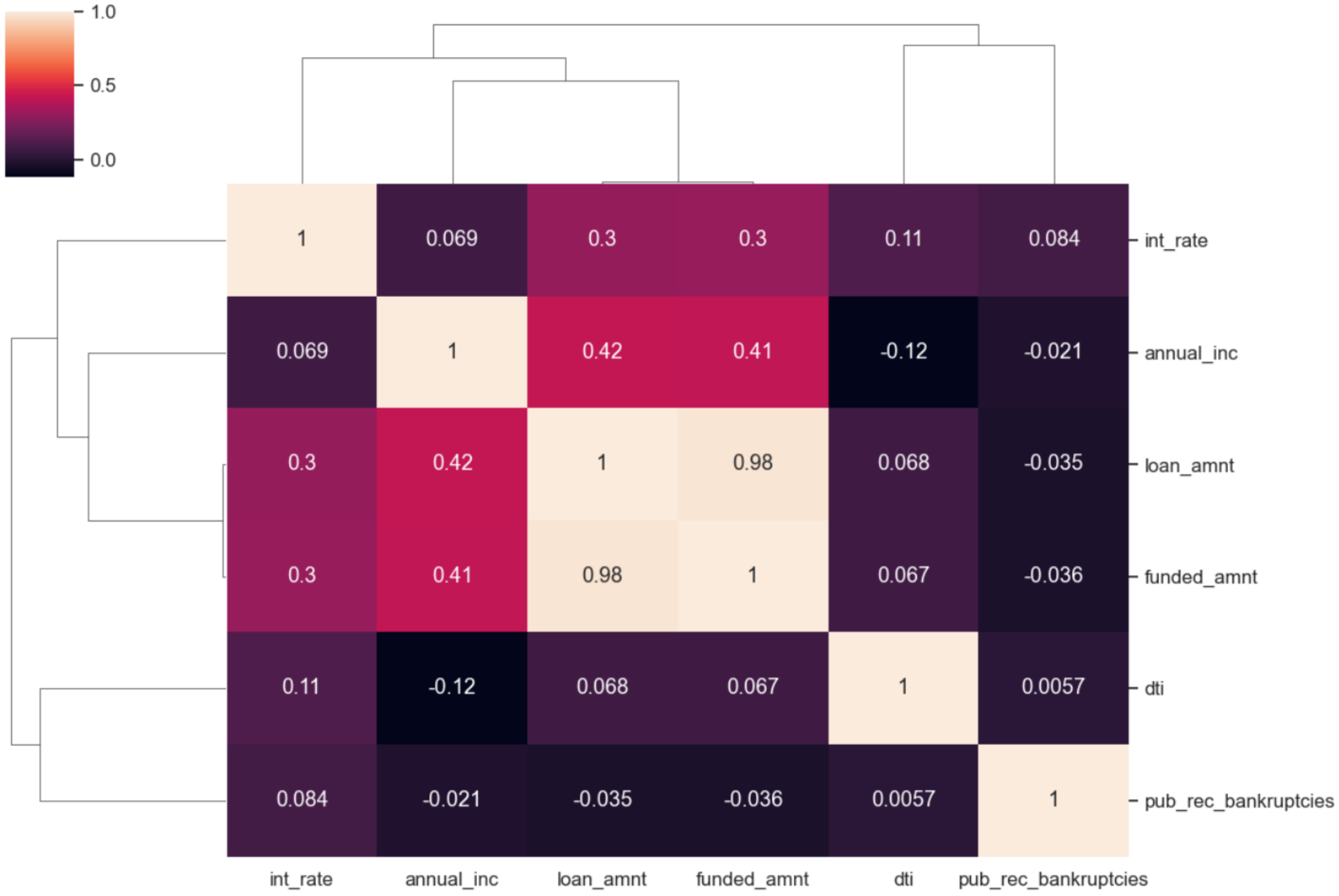
Revolving Utilization v/s Charged Off



Public Bankruptcies records v/s Charged Off



Correlation Plot



Conclusion

Taking out larger loans, especially in the range of $15,000 to $35,000, makes it more likely that the borrower won't be able to repay.

Borrowing more money increases the risk of not being able to pay it back.

People with yearly incomes between $4,000 and $45,000 are more likely to struggle with repayments, showing a link between lower income and repayment difficulties.

Loans for small businesses and renewable energy have a higher chance of not being repaid, maybe because these ventures come with more risks.

If you don't own property, like a house, you're more likely to have trouble repaying a loan.

Having a Debt-to-Income Ratio (DTI) between 24 and 30 significantly raises the chance of not being able to repay a loan, so it's important to check how much debt someone already has.

Credit grades F & G have a higher chance of not repaying, so we need to be careful when approving loans for people in these categories. Specifically, subgrades F5 and G3 are very likely to have trouble repaying.

If someone has a history of public bankruptcies, there's a higher chance they won't be able to repay a loan. That's why we need to carefully check their background before approving their loan.