Lending Club case Study

ML-C61

Arindam Mondal

March-xx-2024

Contents

- Background & Problem Statement
- Objective
- Data Understanding & cleaning
- Data Analysis
  – Univariate Analysis
  – Bivariate Analysis
- Conclusion

Background & Problem Statement

- A consumer finance company grants loans to urban customers, facing risks of both denying loans to good applicants and approving loans to defaulters.
- Given data on past applicants, the company aims to identify patterns predicting defaults for future decisions.
- These decisions include approving loans with various outcomes (fully paid, ongoing, defaulted) or rejecting them entirely.
- Rejected applicants' data isn't available since they have no transaction history with the company.
- This case study utilizes Exploratory Data Analysis (EDA) to understand how applicant and loan characteristics influence default likelihood.

Objective

- The main source of financial loss comes from borrowers who default on loans ("charged-off" customers).
- Identifying these "risky" applicants beforehand can significantly reduce financial losses.
- Understanding these "driver variables" will help the company assess risk and manage its loan portfolio.

Data understanding

- The CSV contains loan data with 111 columns and 39717 rows
- Missing Values
  - There are 56 columns with 90% missing values
- Data types of the columns
  - float64(74), int64(13), object(24)

Data Cleaning and Manipulation

- Dropped column with more than 90% missing value
- Converted Percentage Strings to Floats
- Parsed Date Strings to Datetime Objects
- Ensured Categorical Data is Appropriately Formatted

Data Cleaning and Manipulation

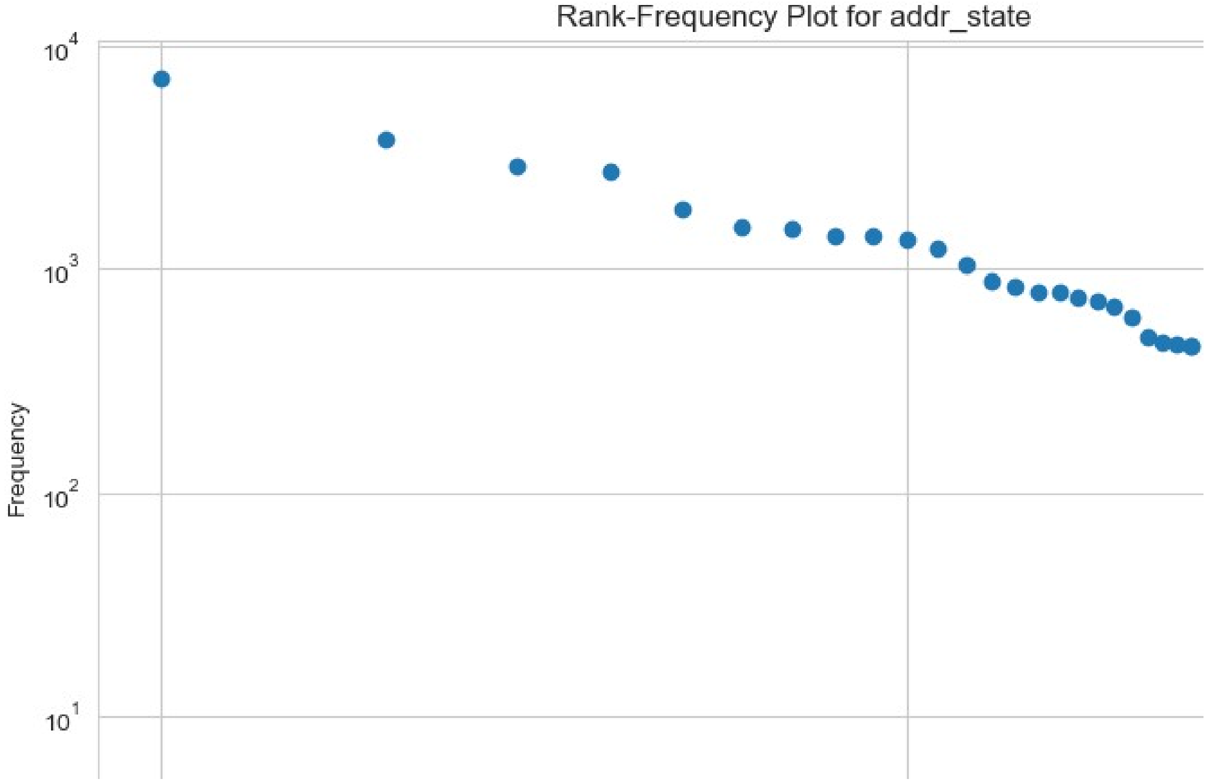
- Missing value treatment with following strategy
  - Impute numerical columns with the median
  - Impute categorical columns with the mode
- Outllier treatment
  - Outlier treatment for annual income

Data analysis

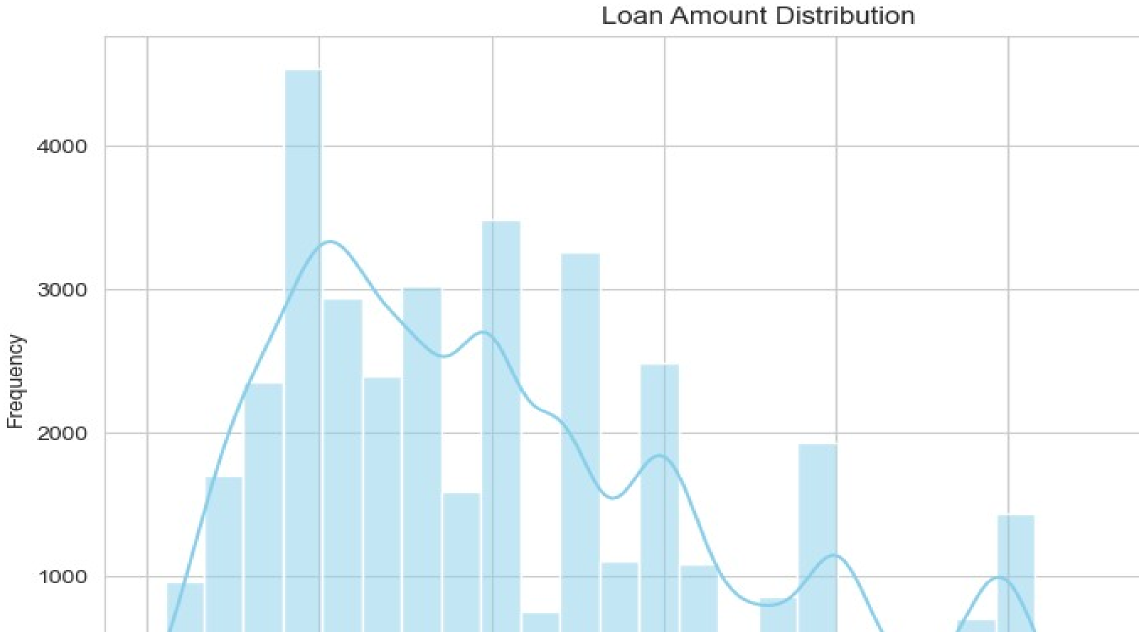
- Univariate Analysis
  - Unordered Categorical Variables
    - rank-frequency plots
      - addr_state shows power law distribution
  - Ordered Categorical Variables
    - Example: home_ownership
      - The 'RENT' and 'MORTGAGE' categories are the most common home ownership statuses among borrowers
  - Quantitative Variables
    - Example: Loan Amount Distribution, interest rates distribution

Univariate analysis: Unordered categorical variable

- rank-frequency plots: addr_state shows power law distribution

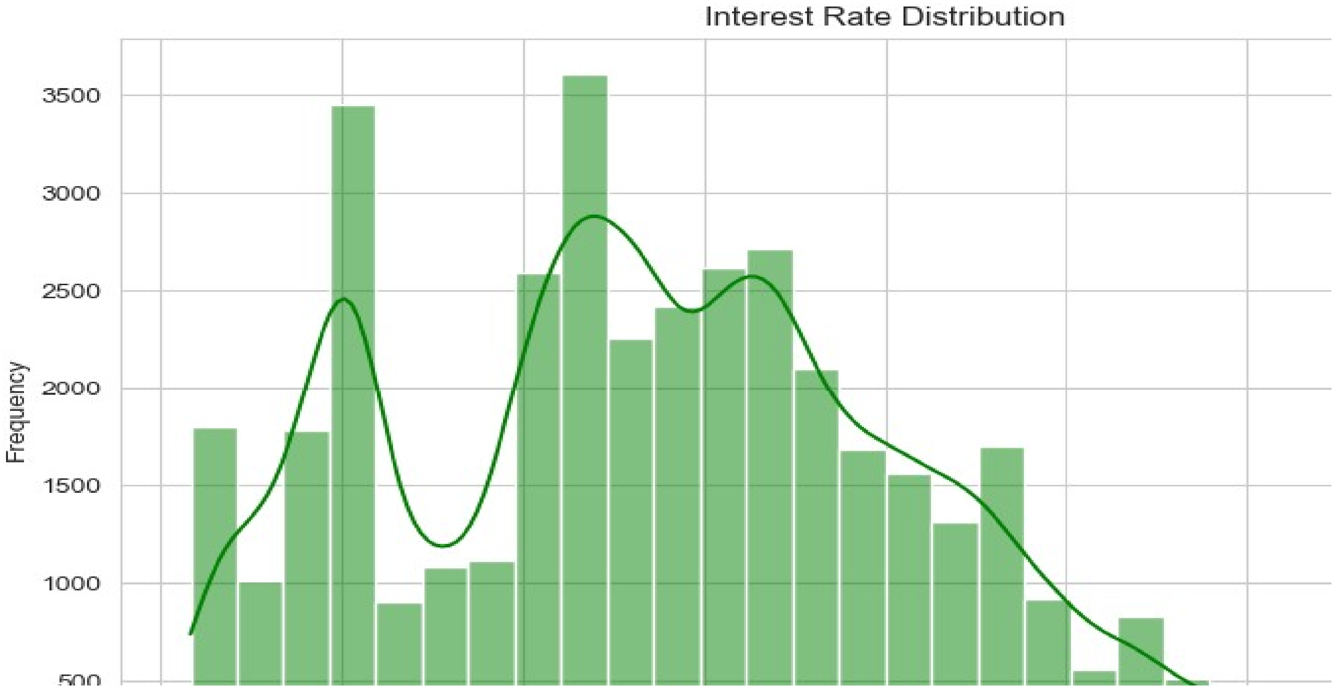


Univariate Analysis: Loan Amount Distribution



- The most frequently occurring loan amount, appears to be in the lower range, around $5,000 to $10,000.
- The frequency of loans decreases significantly for higher amounts, with relatively few loans above $25,000
- The loan amounts range from very small to $35,000

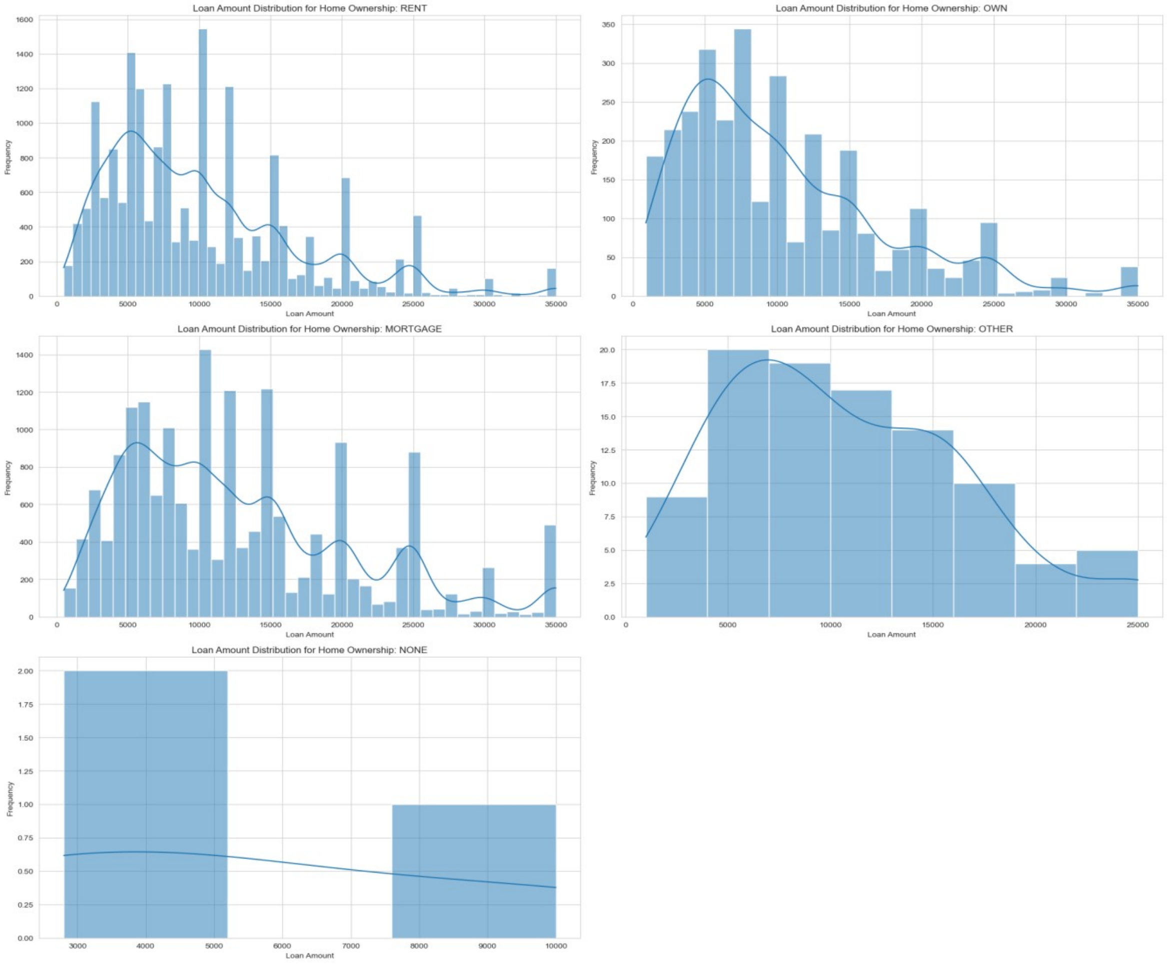
Univariate Analysis: interest rates distribution



- The most common interest rates, indicated by the peaks, seem to be around 7.5%, 10-12.5%, and around 15%.
- Interest rates range from around 5% to 25%, showing a wide variety of rates applied to the loans.
- The frequency of loans decreases as the interest rate.

Segmented Univariate Analysis

- Segmented univariate analysis on “home_ownership” provide following insights

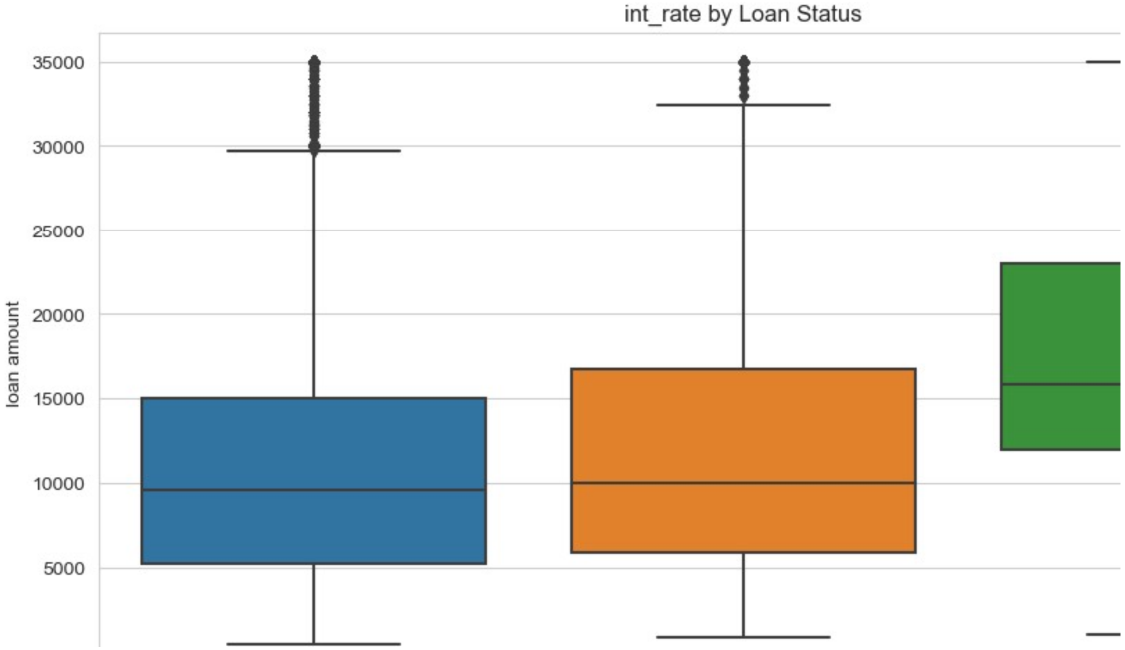


Segmented Univariate Analysis

- Segmented univariate analysis on “home_ownership” provide following insights
  - Borrowers with a mortgage have the highest average loan amount and funded amount
  - The “NONE” category has a very low average loan
  - The “OTHER” group has a moderate average loan amount and funded amount.
  - Those who own their homes have lower average loan and funded amounts compared to 'MORTGAGE
  - Renters have the lowest average loan and funded amounts, which might reflect lower credit limits or borrowing capacity.

Segmented Univariate Analysis

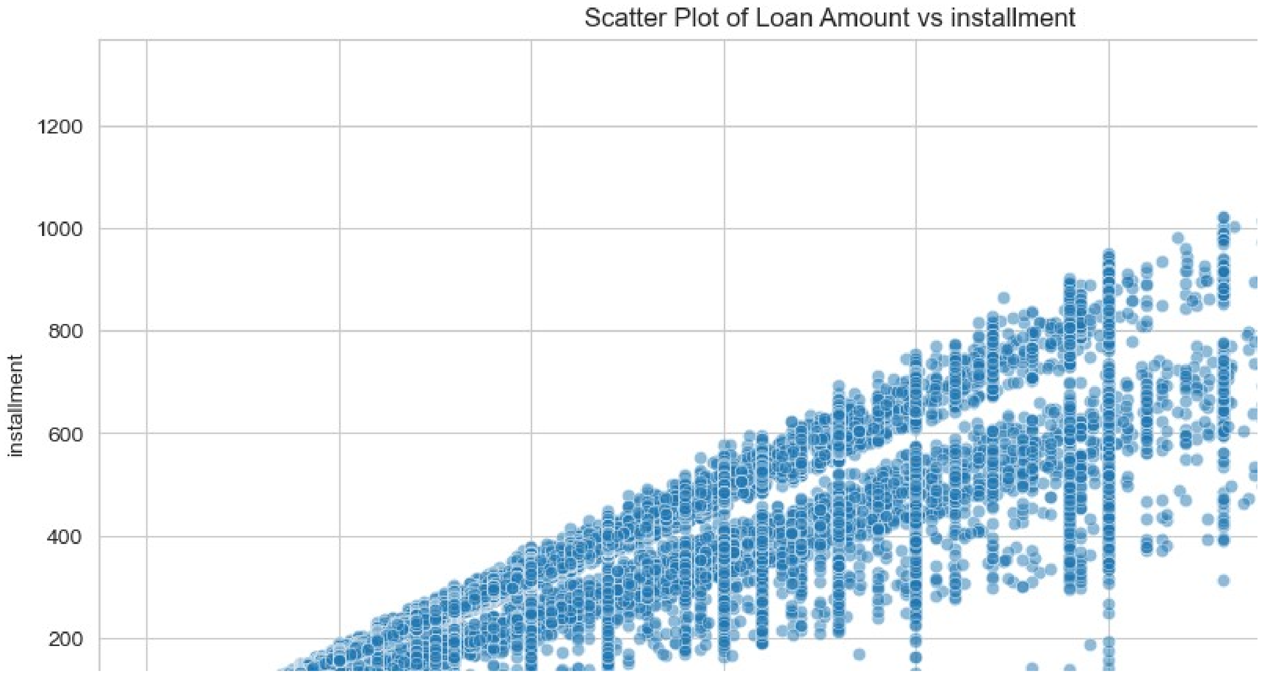
- Segmented univariate analysis on distribution of loan amounts across three different loan statuses: Fully Paid, Charged Off, and Current



•The 'Charged Off' loans have a slightly lower median loan amount compared to the 'Fully Paid' loans.

Bivariate Analysis

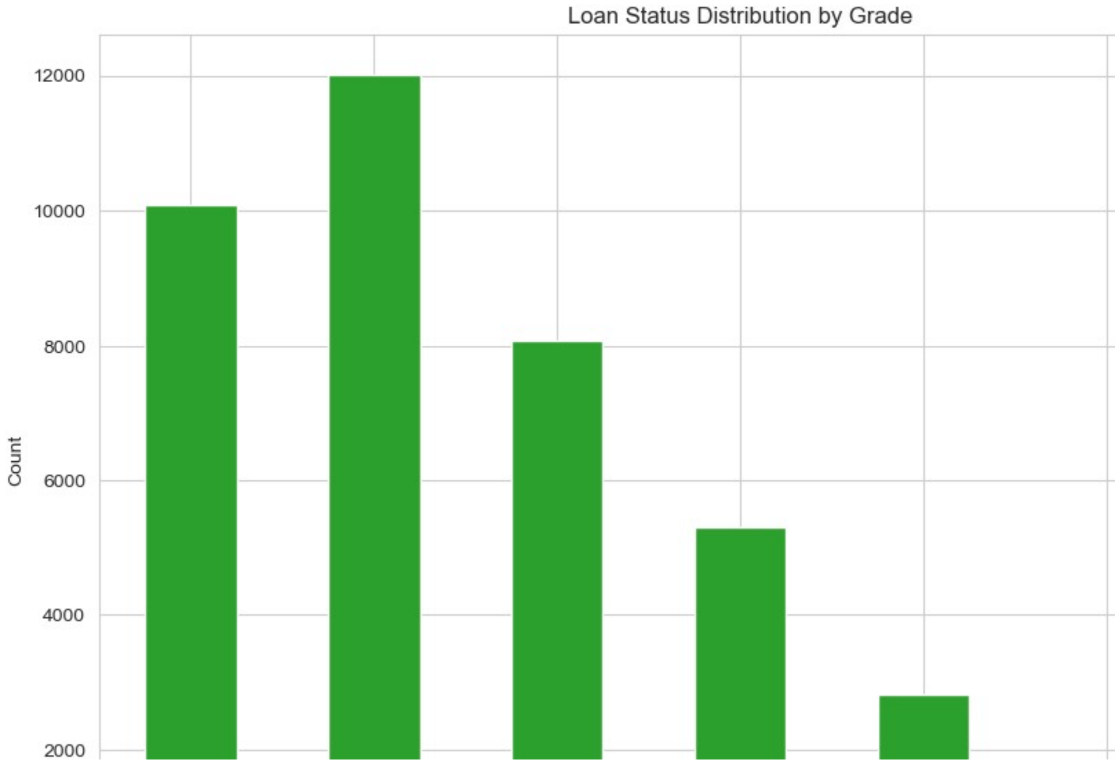
- Loan Amount vs installment



•The plot shows a positive linear relationship between loan amount and installment

Bivariate Analysis

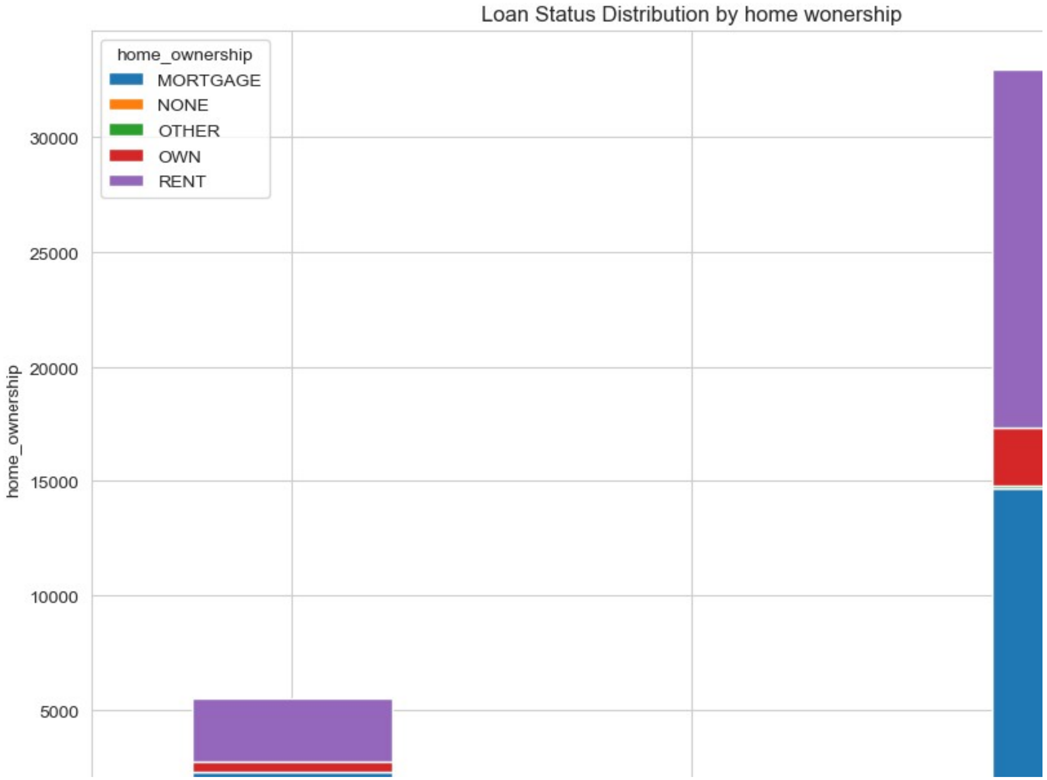
- Loan status vs. grade



•Grade A shows the highest number of loans that are fully paid and has a relatively small proportion of charged-off loans. This suggests that Grade A loans are lower risk.
•As the grades progress from A to G, there is a noticeable trend where the count of fully paid loans decreases

Bivariate Analysis

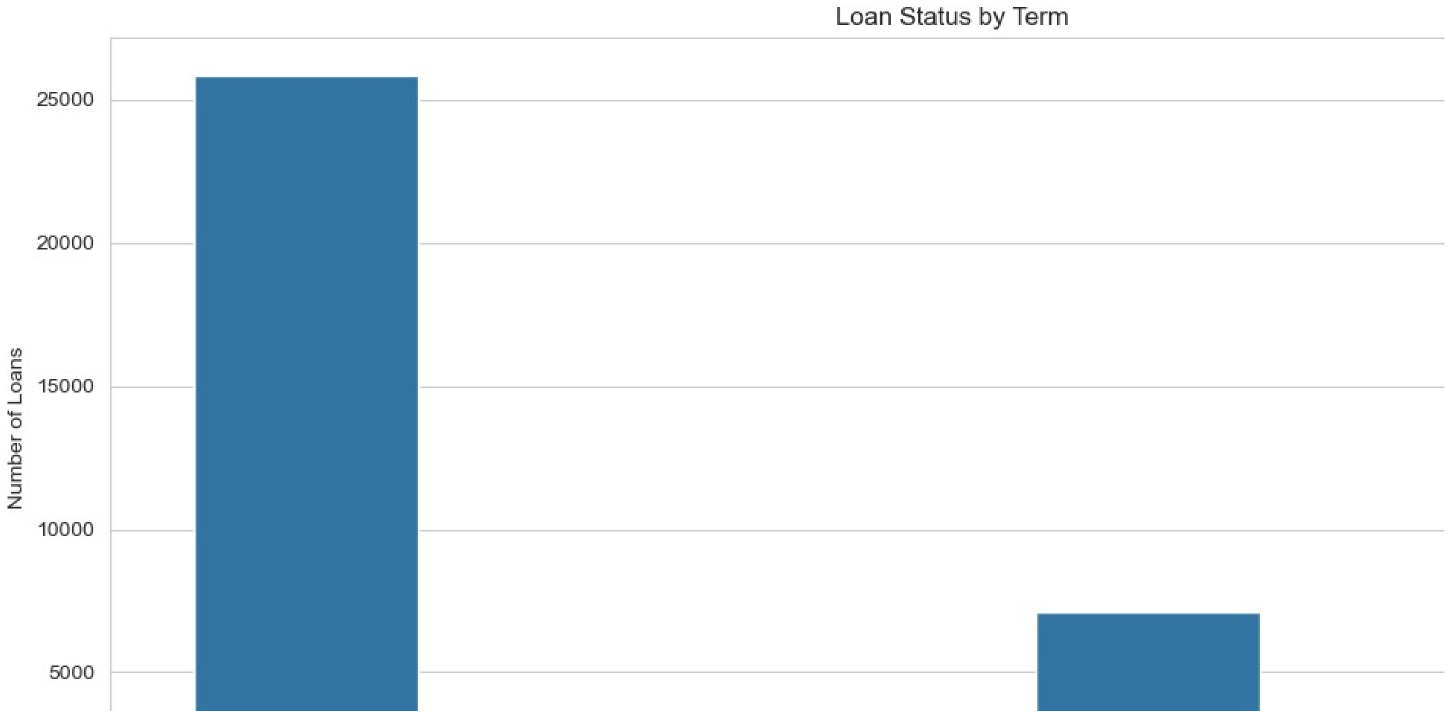
- **Loan Status Distribution by home wonership**



•In the charged off category, the loans are almost evenly distributed between 'Mortgage' and 'Rent', with 'Rent' having a slightly higher proportion.

Bivariate Analysis

- **Loan Status by Term**



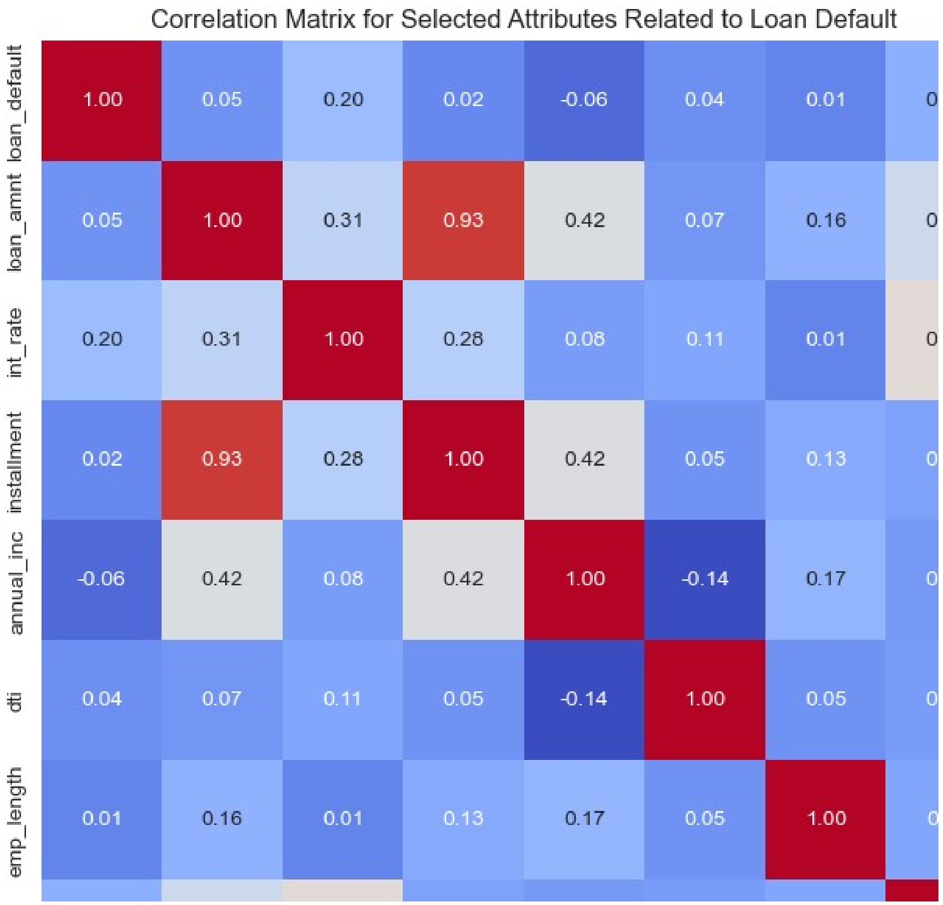
•The charge-off rate for the 36-month term appears to be lower than for the 60-month term
•This could suggest that shorter-term loans are less risky for lenders

Bivariate Analysis

- Correlation matrix

• There is a positive correlation between interest rate (int_rate) and loan default

•There is a strong positive correlation between loan amount and installment



Conclusion

- **DrivingFactors(or drivervariables):**
  - ✓ Grade: As the grades progress from A to G, there is a noticeable trend where the default count increases
  - ✓ Interest Rate: There is a positive correlation between interest rate and loan default
  - ✓ Term: The charge-off rate for the short term appears to be lower than for the long term