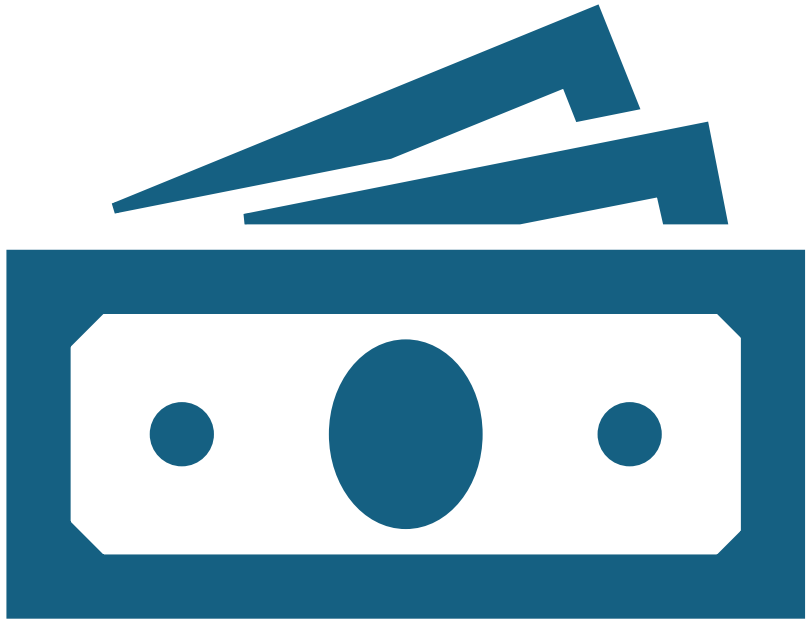
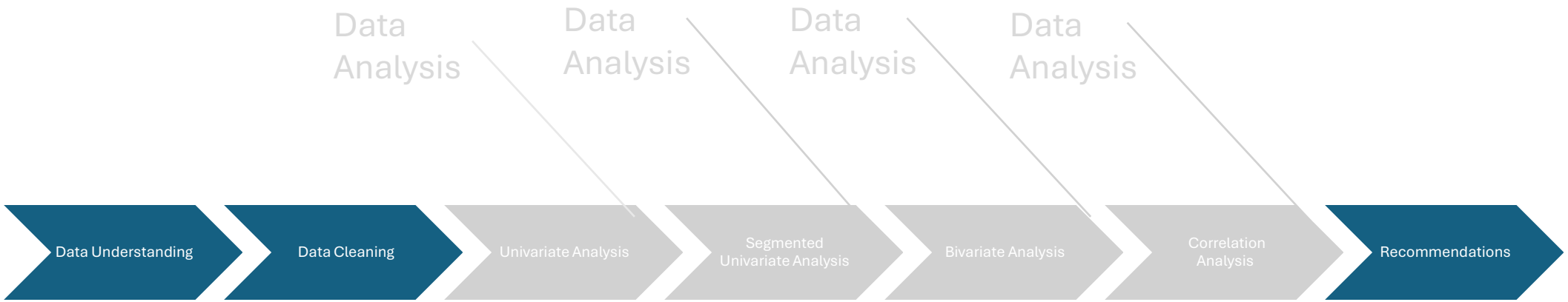


Upgrad Case Study

Lending Club

By
Varadhan Mariappan
Vijayamma Battala

EDA: Exploratory Data Analysis Process for Lending Club Case Study



Process of analyzing a dataset's structure, content, and context to ensure it aligns with the analytical objectives and to identify any potential patterns.

process of analyzing a dataset's structure, content, and context to ensure it aligns with the analytical objectives and to identify any potential quality issues or patterns.

Examination of a single variable to summarize its main characteristics, such as distribution, central tendency, and variability, often using visualizations and descriptive statistics.

Examining the distribution and characteristics of a single variable within specific subgroups or segments of the data to identify patterns or differences across those segments.

Examining the relationship between two variables to determine how they are correlated or how one may influence the other.

statistical technique used to assess the strength and direction of the relationship between two variables.

Suggestions or advice based on analysis or expertise, aimed at guiding decisions or actions to improve outcomes or solve problems.

Data Understanding

Problem Statement

- 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

Business Objectives

- Identification of such applicants using EDA is the aim of this case study.
- Identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.
- The company wants to understand the driving factors (or driver variables) behind loan default. The company can utilise this knowledge for its portfolio and risk assessment.

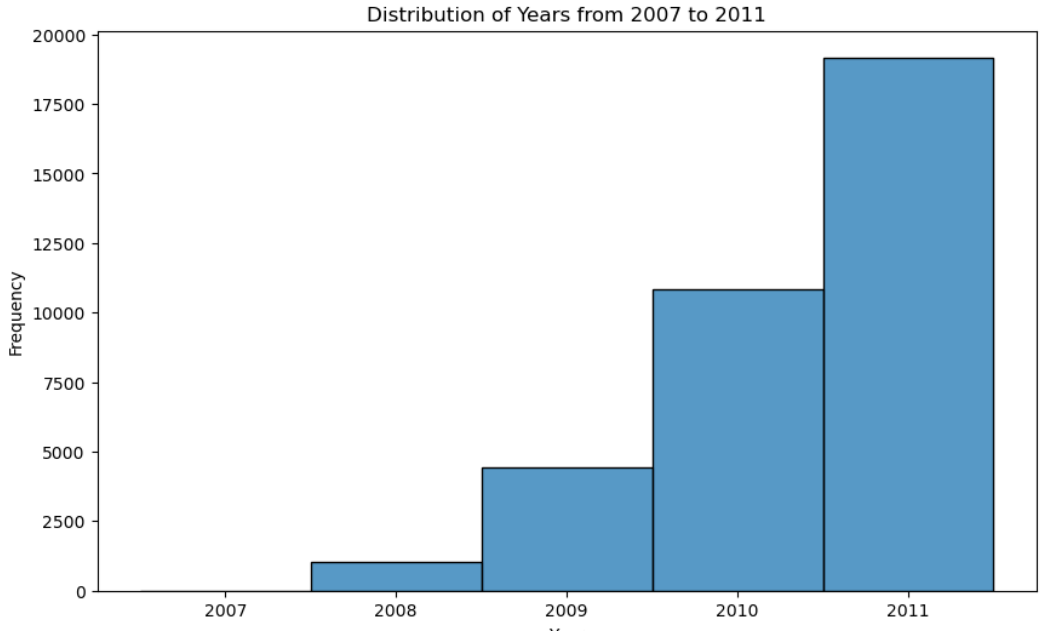
Understanding datasets

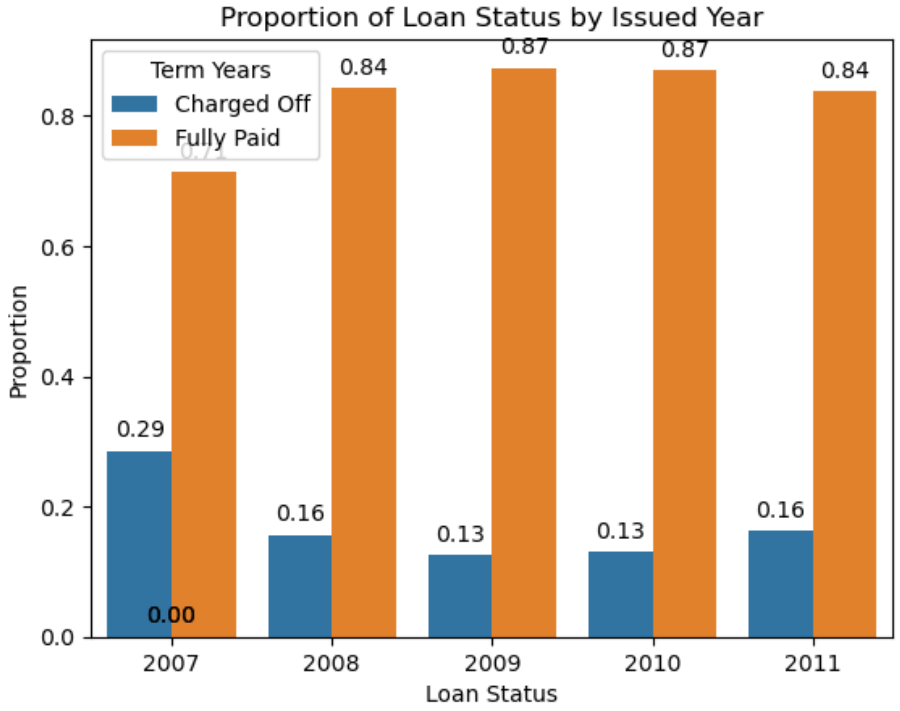
- Finance attributes: Identify and understand loan related attributes
- Borrower attributes: Identify and understand borrower related attributes

Data Cleaning

- **Initial Checks:**
  - Check for Header & Footer: Verify the presence of non-data elements in headers and footers.
  - Check for Empty Rows: Identify and handle any rows that contain no data.
  - Check for Null Values: Count missing data in each column and compare with total rows to find columns with all null values.
  - Remove Columns with High Missing Values: Drop columns with more than 60% missing values.
  - Remove Descriptive Columns: Exclude non-numeric or categorical columns that do not contribute to analysis or modeling.
  - Remove Unwanted Rows: Discard irrelevant, duplicate, or incorrect rows, such as those with all null values or outliers.
  - Remove Unwanted Columns: Ensure that only useful features remain for analysis.
- **Data Standardization:**
  - Verify Non-Numerical Values: Standardize categorical data by converting to consistent formats.
    - Convert Date Columns: Convert date columns to proper datetime format for accurate operations and analysis.
    - Remove Symbols from Columns: Remove '%' or other symbols to convert columns into numeric values.
    - Fix Decimal Points: Round numeric values to two decimal points for consistency.
    - Remove Strings from Numeric Variables: Ensure numeric variables are clean by removing non-numeric characters.
    - Replace Strings Indicating Nulls: Replace strings like "NaN," "NULL," or "NONE" with appropriate null values.
    - Merge Similar Values: Combine values that represent the same category for consistency.
    - Replace Null Values: Use relative values (mean, median, mode, or custom values) to fill nulls.
- **Derived Variables:**
  - Create Derived Values: Generate new features that provide additional insights.
    - Create Derived Value Categories: Create new variables from existing ones.
    - Add Date-Based Categories: Extract and add month, year, and quarter as separate columns.
    - Create Numerical-Based Categories: Group continuous values into discrete ranges.

Company Growth

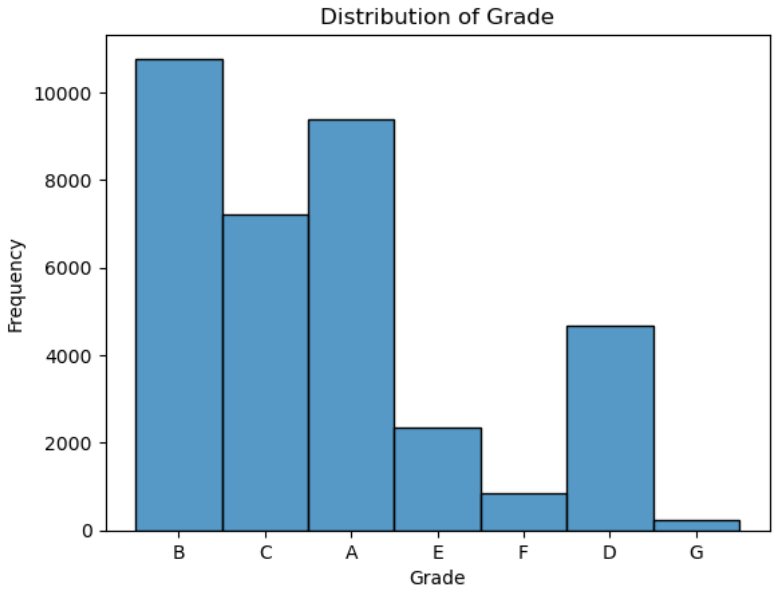


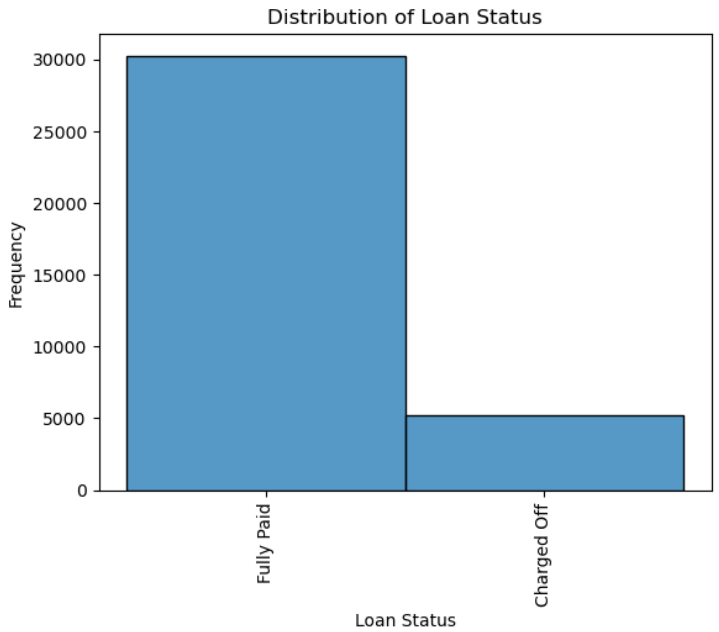


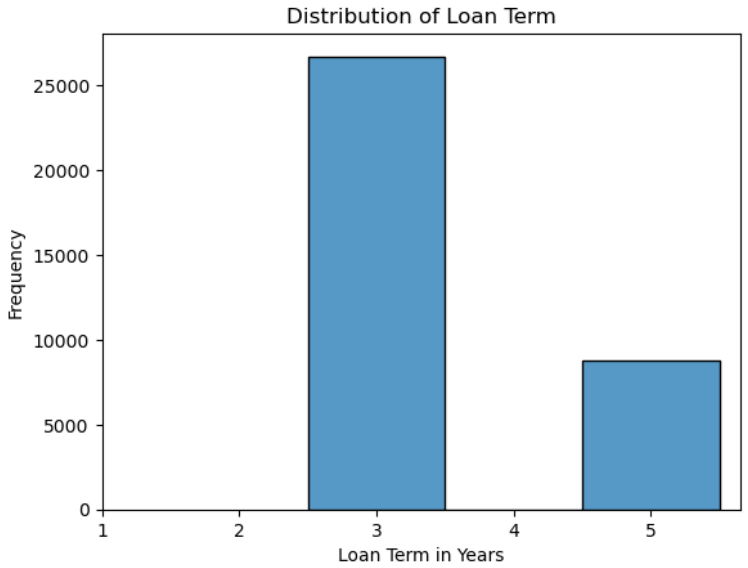
Observation

The company has been experiencing growth each year, with an increasing amount of loans being issued. Although the amount of defaulted loans decreased from 2007 onwards, there was a rise again in 2011. This issue needs to be addressed to mitigate financial losses.

Loan Amount Distribution



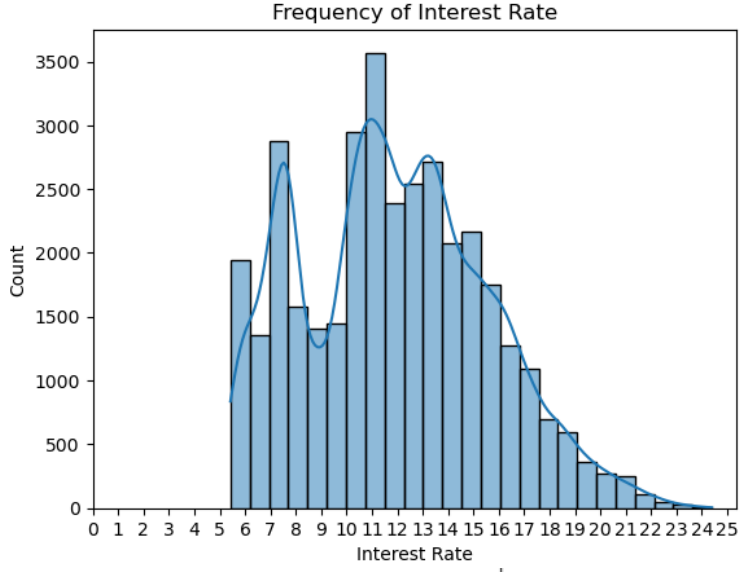


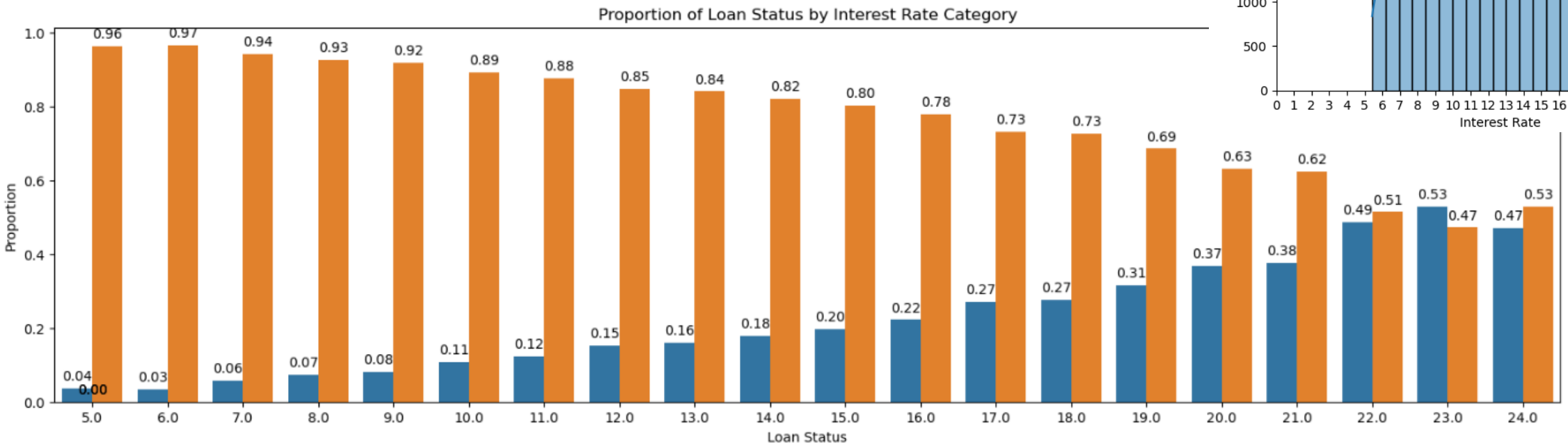


Observation

- The majority of loans are distributed to Grade B, followed by Grade A.
- Most loans have been fully paid.
- There is a significant proportion of loans (~15%) that have defaulted.
- Most of the loans issued have a term of 3 years.

Interest rate and default loans

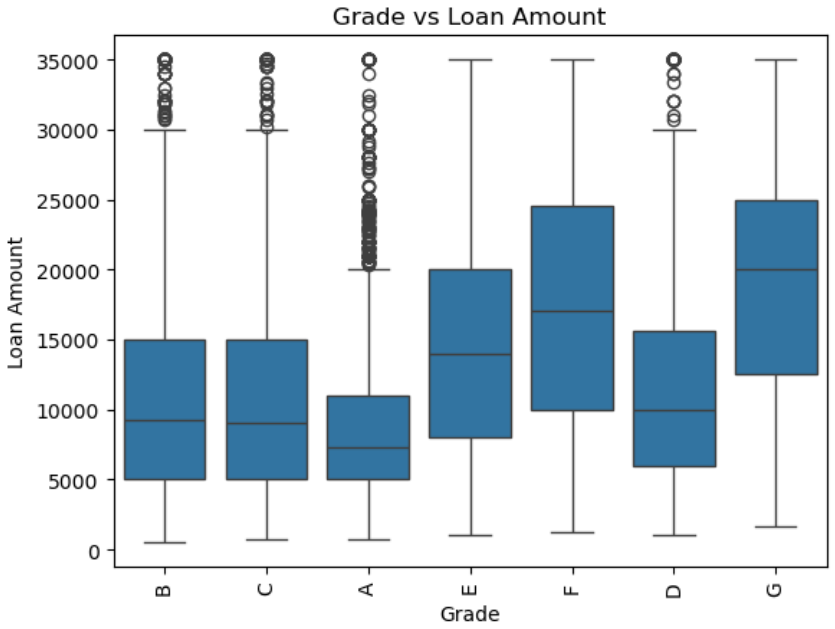


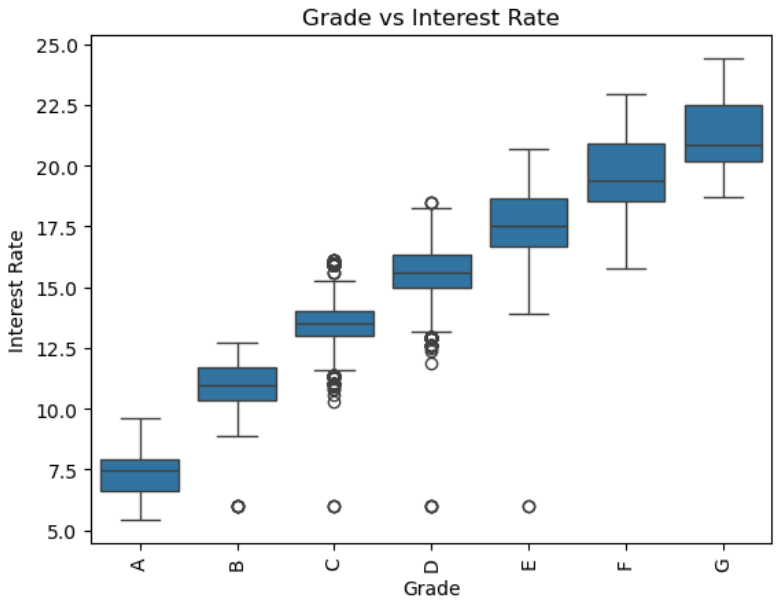


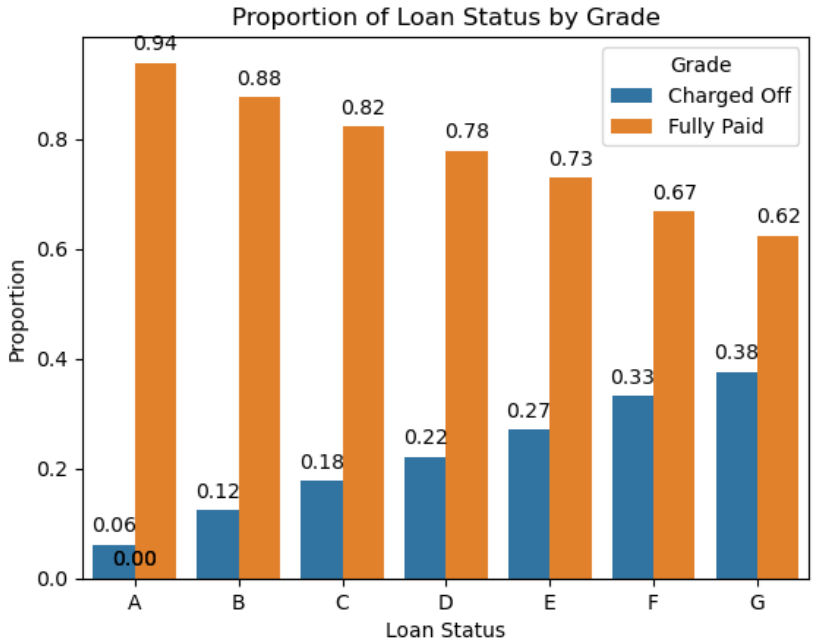
Observation

- Most applicants received interest rates of 11%, followed by 10% and 7%.
- There is an increase in default loans with higher interest rates.
- Interest rates between 22% and 25% have a higher incidence of default loans.

Loan amount vs Interest rate vs grade loans







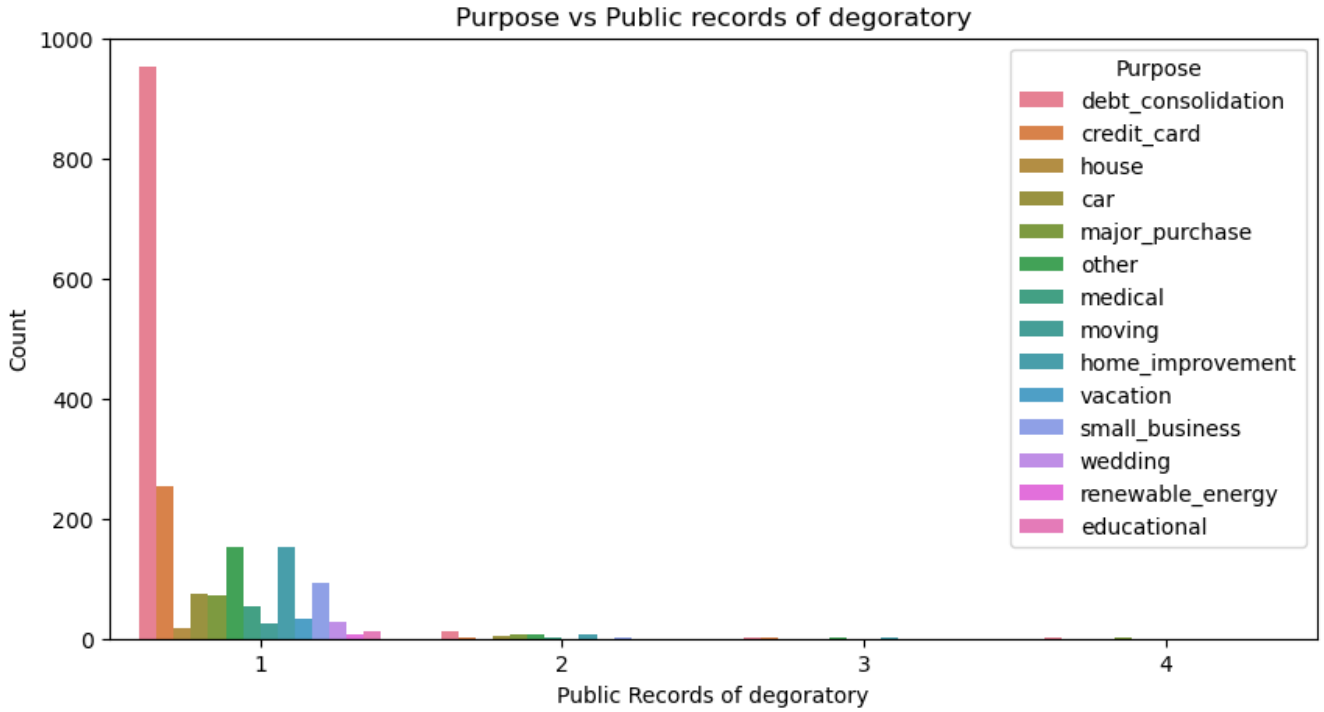
Observation

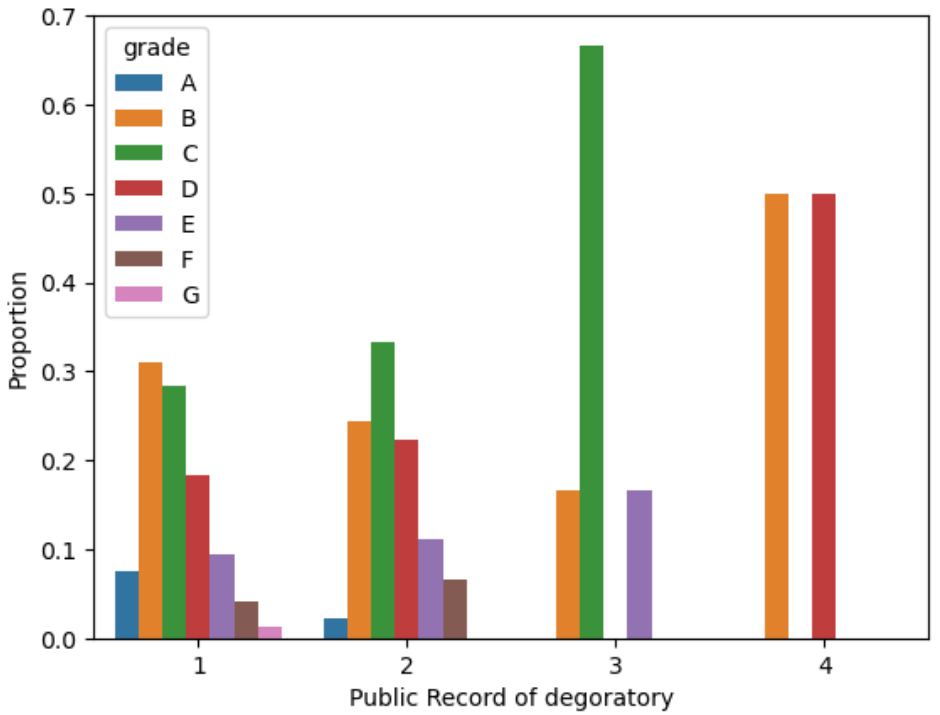
- **Grade vs Loan Amount:**
  - Grades A, B, C, and D have similar median loan amounts, around $10,000.
  - Grades F and G have higher median loan amounts, around $20,000.
  - Grade E has a median loan amount of $15,000.
  - Grade A has the most outliers, followed by Grades B and C.

Observation

- **Grade vs Interest Rate:**
  - Interest rates increase as the grade decreases.
- **Grade vs Loan Status:**
  - As the grade increases from A to G, the likelihood of loans defaulting also increases.

Public Record vs Purpose and Grade





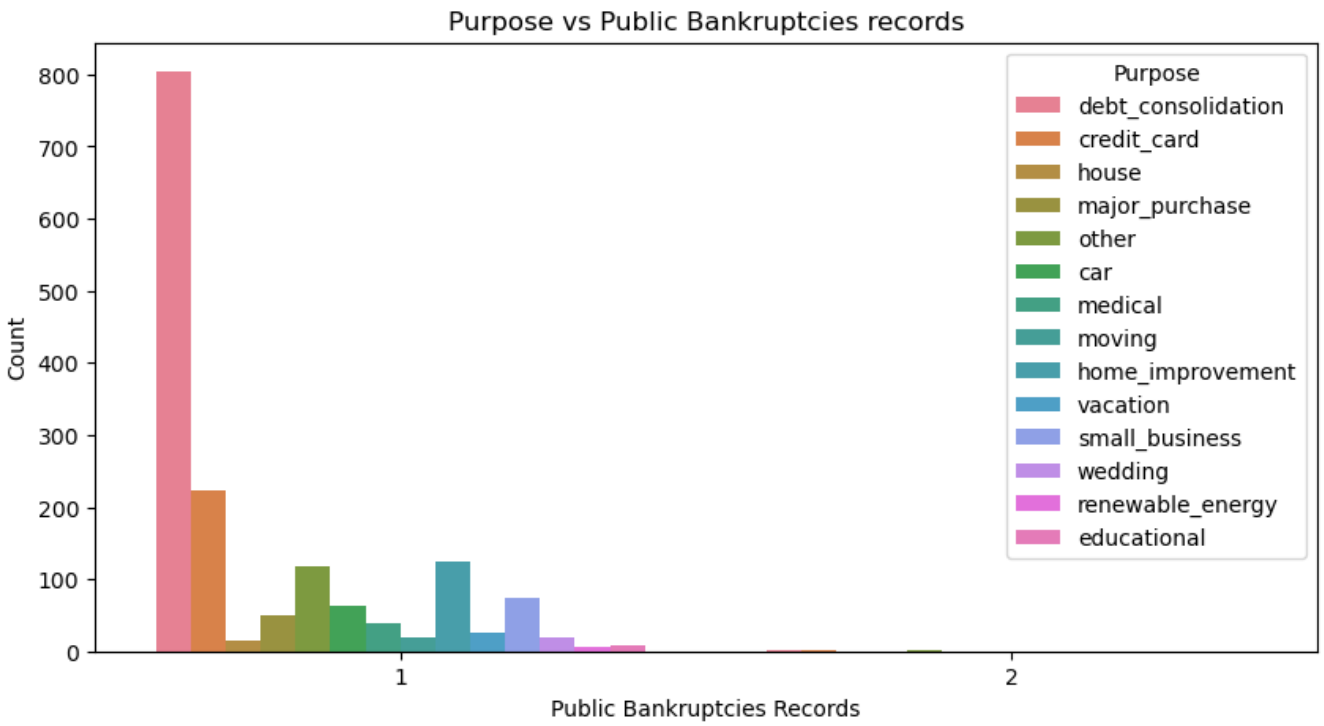
Observation

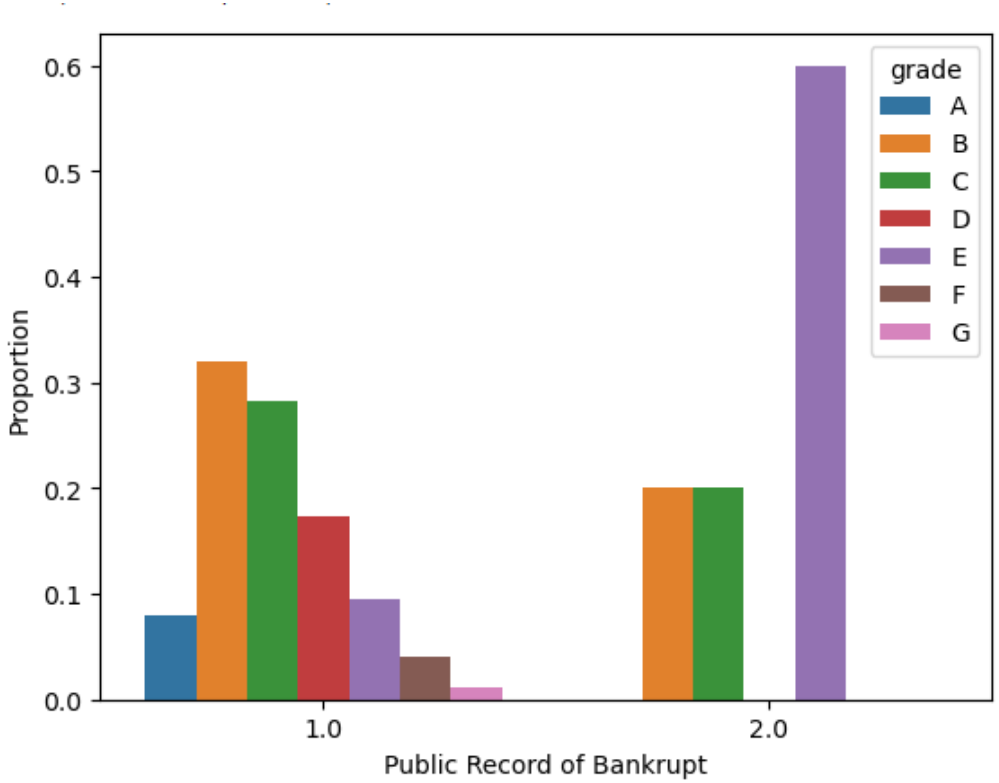
- **Purpose with Public Record of Derogatory**:
  - Debt consolidation loans have higher public records for derogatory items, followed by credit card loans.

Observation

- **Public Record of Derogatory Proportion with Grade**:
  - Grades B, C, D, and E have higher public records.
  - Grade A has low public records.
  - Grade G has very low public records, possibly due to fewer loans being approved.

Public Record Bankrupt vs Purpose and Grade





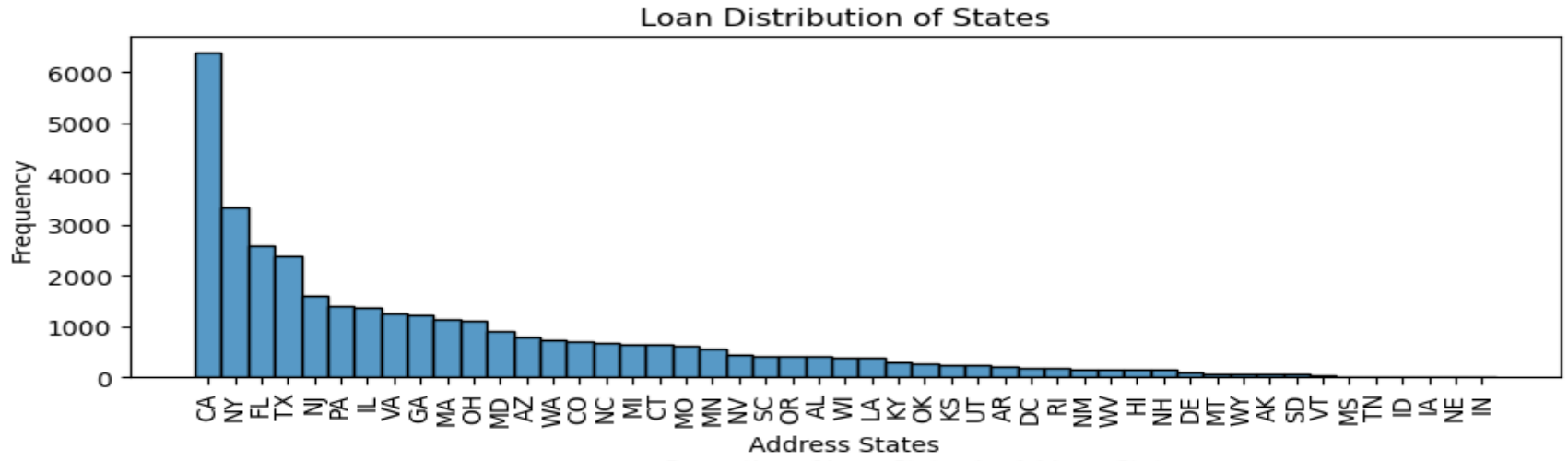
Observation

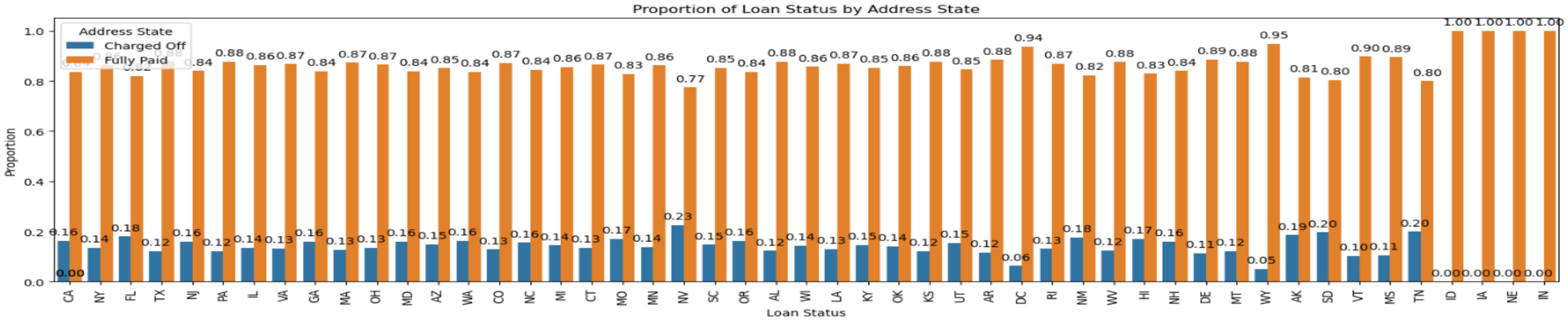
- **Purpose with Public Record of Bankrupt**:
  - Debt consolidation loans have higher public records for bankruptcies, followed by credit card loans.

Observation

- **Public Record of Bankrupt Proportion with Grade**:
  - Categories E, B, and C have high public records for bankruptcies, around 2.
  - Category G has fewer public records for bankruptcies, possibly due to fewer loans being approved

Distribution of Loan Amount and Default across states

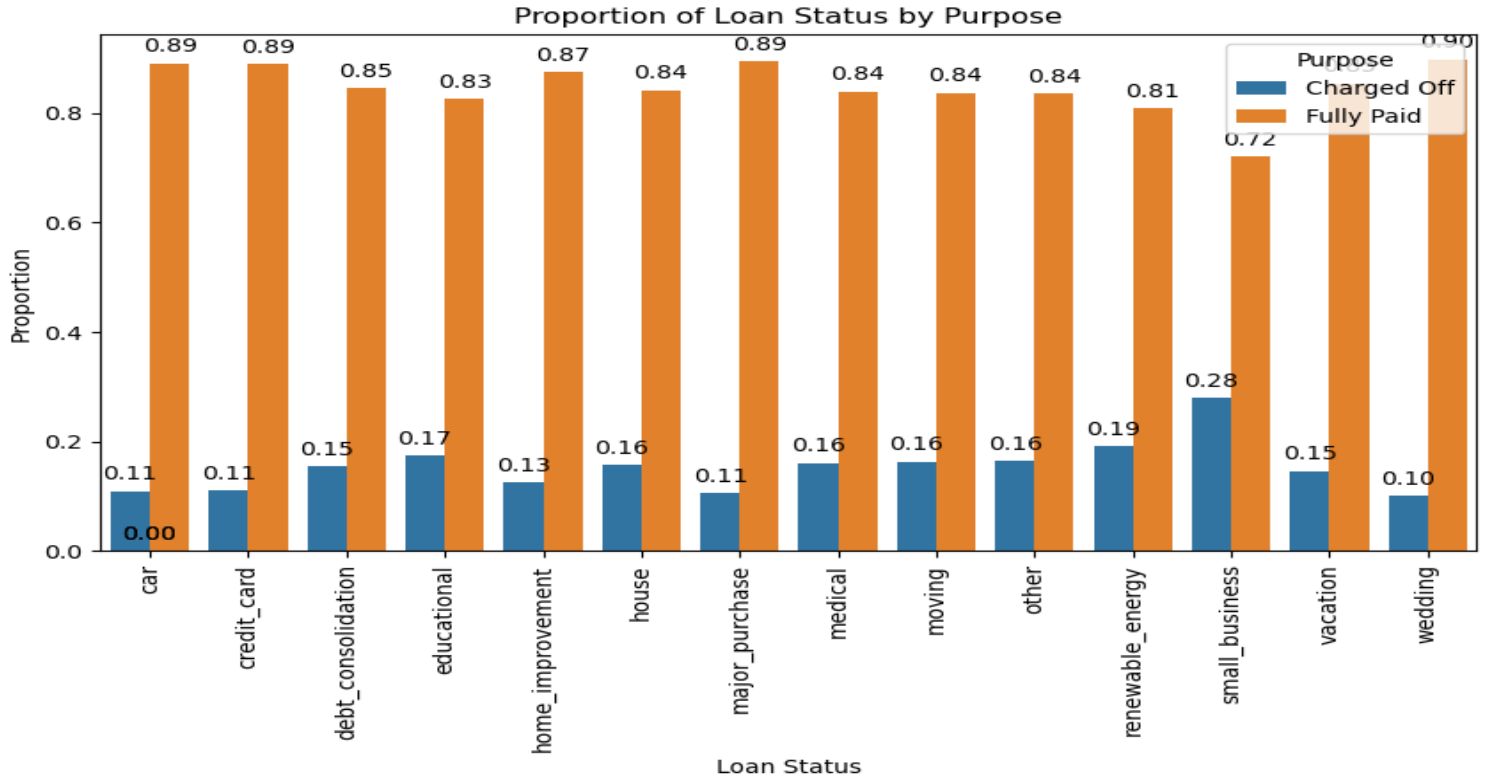




Observation

- State NE: More loans are charged-off and fewer are fully paid.
- States IA, IN, NE, ID: High proportion of loans are fully paid off with no defaults, but very few loans have been provided.
- States WY, DC: High proportion of loans are fully paid off.
- State CA: Proportion of charged-off loans is 0.16 and fully paid loans is 0.14, with a large number of loans provided to this state. CA has received more than 6000+ loan amount, followed by NY.

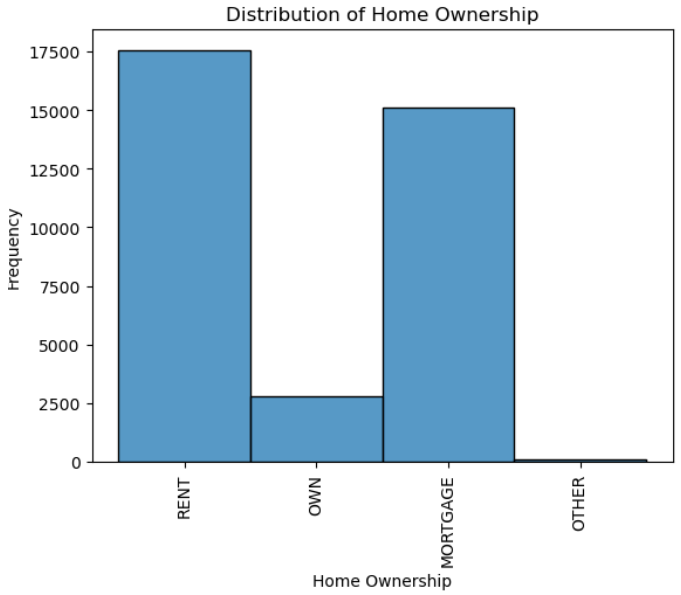
Purpose vs Loan Status

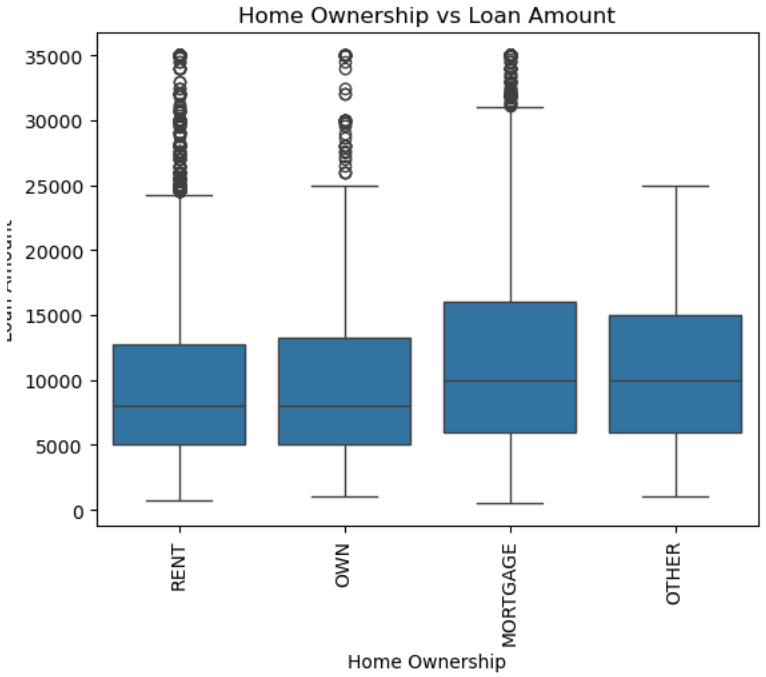


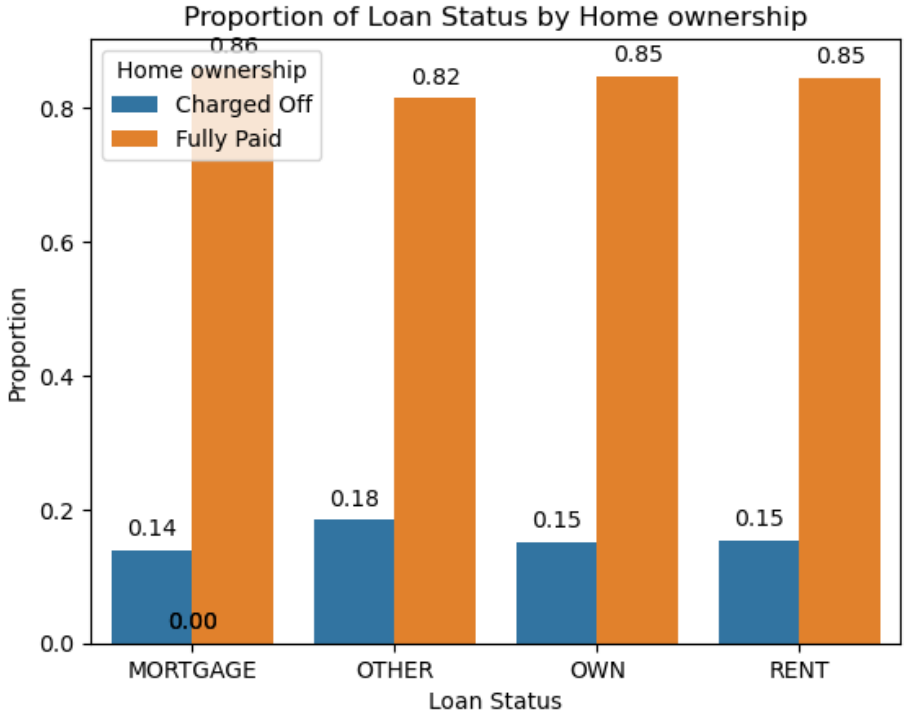
Observation

- Small business loans and Renewable energy have a very high default rate.
- Followed by Educational and debt consolidation loans have high default rates.
- Wedding loans have a very low default rate.
- Car, credit card, and major purchase loans have low default rates.

Purpose vs Loan Status







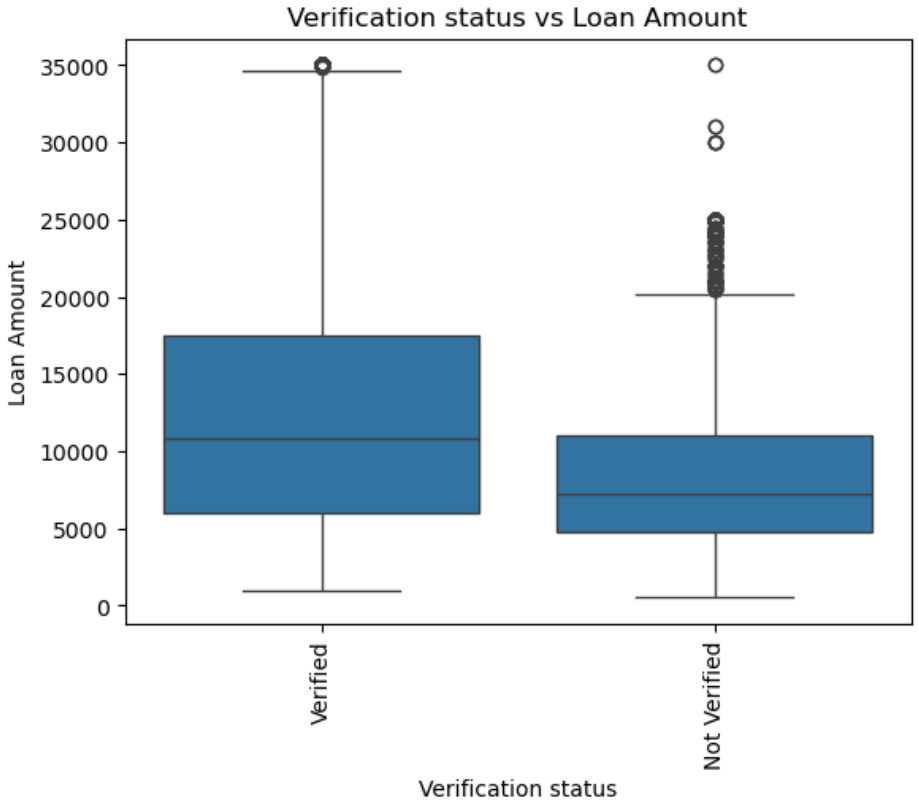
Observation

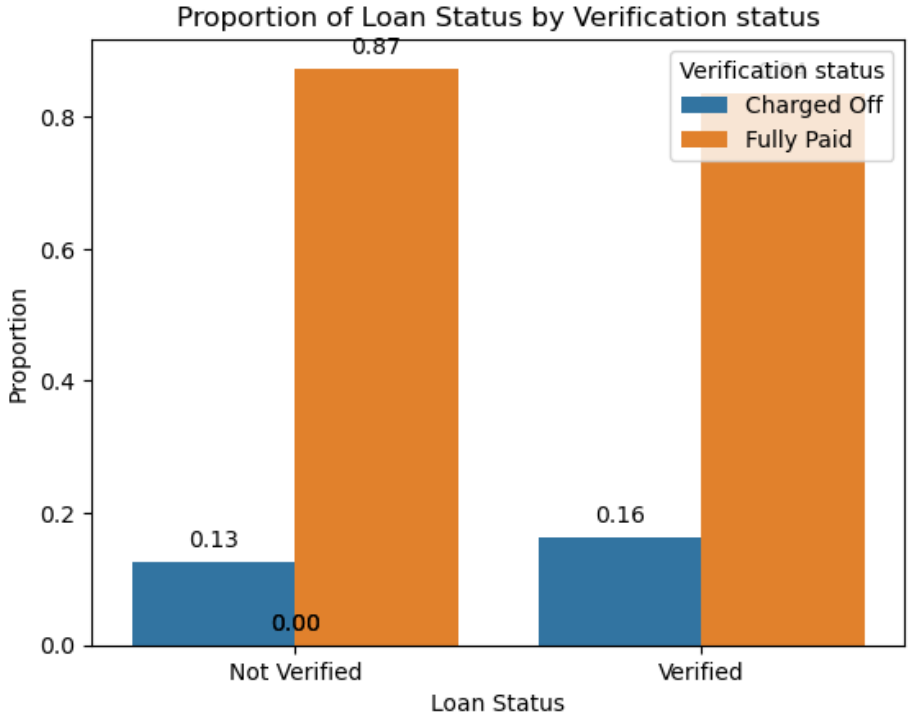
- **Distribution of Loan Amount with Home Ownership**:
  - Most loans are associated with rented homes, closely followed by mortgages.

Observation

- **Home Ownership vs Loan Amount**:
  - The median loan amount is almost the same across different home ownership segments.
  - Home ownership with mortgages received higher loan amounts.
  - There are more outliers for renters and mortgage home owners.
- **Proportion of Loan Status by Home Ownership**:
  - Loans with home ownership categorized as "OTHER" have higher default rates.
  - Other home ownership categories have similar rates of charged-off loans.

Verification Status Vs Loan Amount and Defaults



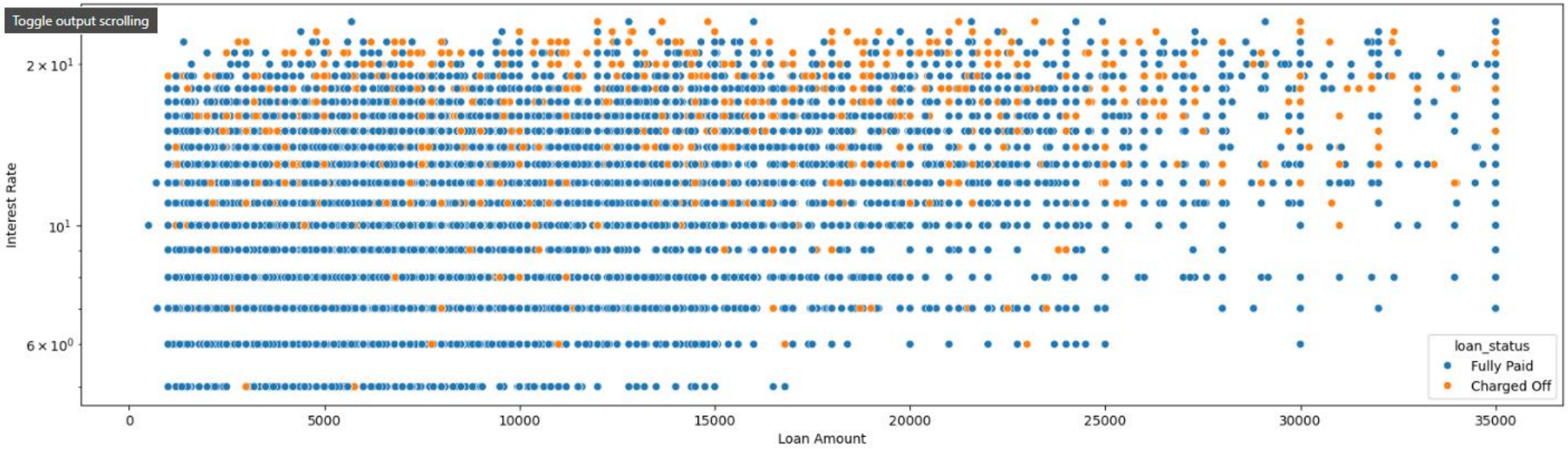


Observation

- **Verification Status vs Loan Amount:**
  - Verified loans have fewer outliers with a median amount of $11,000.
  - Not verified loans have more outliers with a median amount of $7,000.
- **Verification Status vs Charged-Off Loans:**
  - Verified loans show a marginally higher rate of default.

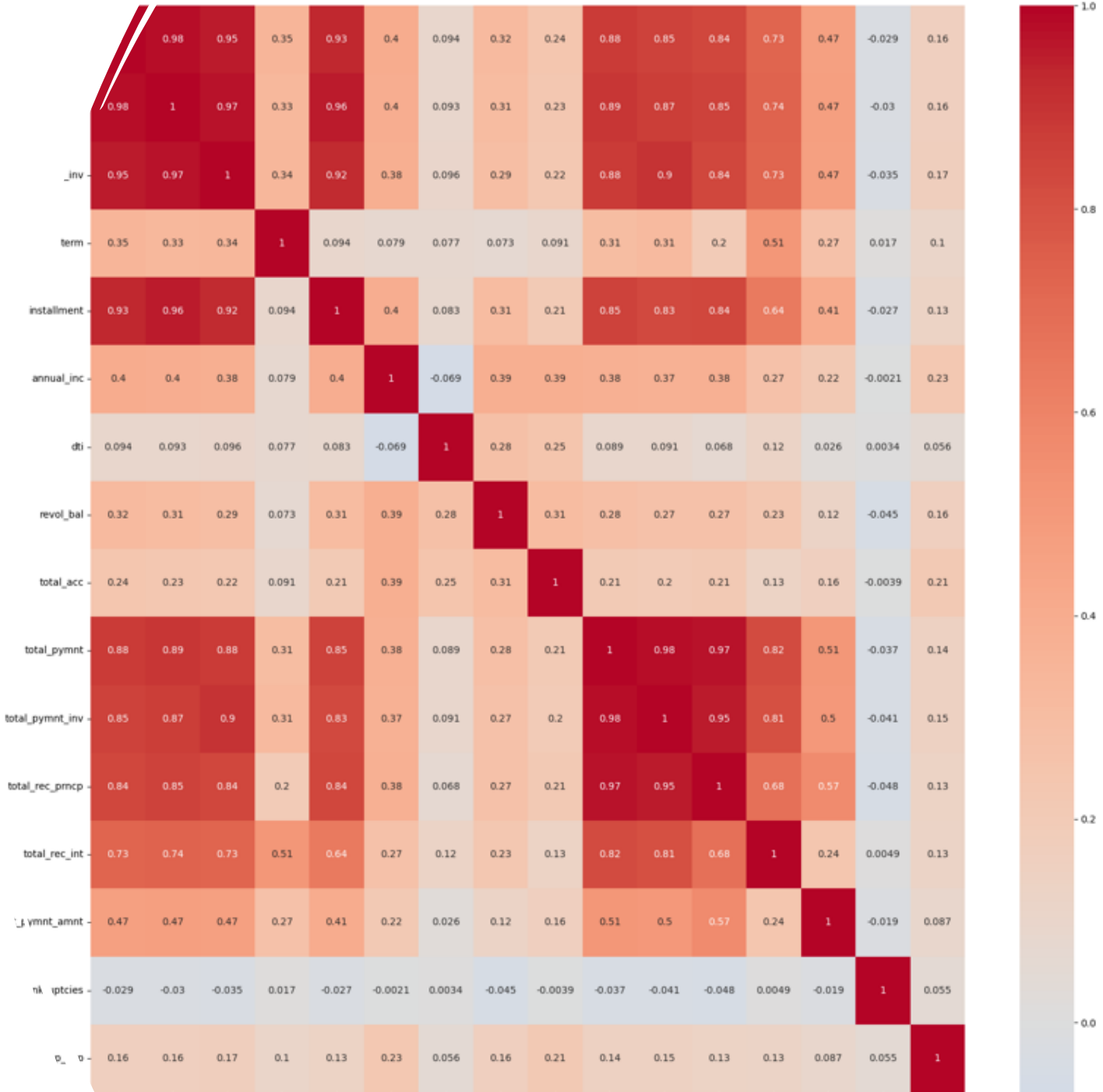
Loan Amount vs Interest Rate with Default Loan

- **Observation**
- Interest rates above 10% have a higher default rate.
- Interest rates above 20% have significantly more defaults compared to fully paid loans.
- There are fewer loans with lower interest rates for loan amounts greater than $15,000.



Correlation Matrix

- **Observation**
- **High Correlation**:
  - Loan amount, funded amount, and funded amount invested have very high correlation.
  - Installment has high correlation with loan amount, funded amount, and funded amount invested.
  - Total payment has high correlation with loan amount, funded amount, funded amount invested, installment, total payment invested, total received principal, and total received interest.
- **Moderate Correlation**:
  - Last payment amount has moderate correlation with loan amount, funded amount, funded amount invested, installment, total payment invested, total received principal, and total received interest.
- **Low Correlation**:
  - Revolving balance has low correlation with loan amount, funded amount, and funded amount invested.



Recommendations to Avoid Financial Loss on Bad Loans and Gain More Business

Attributes	Recommendation	Driving Factors
Grade vs Loan Amount	**Implement stricter underwriting criteria for higher grades (F and G)**	Higher loan amounts and riskier borrower profiles in lower grades (F, G) lead to more defaults.
Grade vs Interest Rate	**Adjust interest rates to better reflect the risk associated with each grade**	Higher interest rates correlate with higher default rates; borrowers in lower grades may struggle to repay high-interest loans.
Grade vs Loan Status	Increase monitoring and support for loans in grades D, E, F, and G	**Lower creditworthiness and higher risk profiles in higher grades lead to increased defaults.**
Public Records Proportion with Grade	Tighten approval criteria for borrowers with higher public records of derogatory marks in grades B, C, D, and E. **Be cautious in approving loans for grades E, B, and C; consider additional checks or higher interest rates**	Borrowers with derogatory public records are more likely to default. Higher historical bankruptcy rates in these grades indicate a greater risk of future defaults.
Purpose with Public Records	Pay closer attention to debt consolidation and credit card loan applications. **Implement stricter vetting for debt consolidation and credit card loans**	High debt levels and poor credit management practices in these categories. Borrowers seeking these types of loans may already be in financial distress.
Address State Analysis	Implement region-specific strategies based on loan performance in different states	**Economic conditions and borrower profiles vary by state, affecting default rates.**
Loan Purpose Analysis	**Be more cautious with small business and renewable energy loans; offer favorable terms for low-risk purposes**	Business risk and financial stability of borrowers in these categories.
Interest Rates Analysis	Avoid issuing high-interest loans (>10%) to borrowers with questionable creditworthiness	**High-interest rates increase the likelihood of default, especially on larger loan amounts.**
Verification Status	**Prioritize and enhance loan verification processes to reduce the risk of defaults and outliers.**	Enhancing the verification process can increase borrower trust and attract more reliable customers, thereby driving more business.

Possible Driving Factors for More Default Loans

1. **HIGH LOAN AMOUNTS**: BORROWERS MAY STRUGGLE TO REPAY LARGER LOANS, ESPECIALLY IF THEIR FINANCIAL SITUATION CHANGES.
2. **HIGH INTEREST RATES**: INCREASED REPAYMENT BURDEN CAN LEAD TO DEFAULTS, PARTICULARLY FOR HIGH-RISK BORROWERS.
3. **POOR CREDIT HISTORY**: BORROWERS WITH DEROGATORY MARKS OR BANKRUPTCIES ARE MORE LIKELY TO DEFAULT.
4. **ECONOMIC CONDITIONS**: REGIONAL ECONOMIC DOWNTURNS CAN IMPACT BORROWERS' ABILITY TO REPAY LOANS.
5. **LOAN PURPOSE**: CERTAIN PURPOSES LIKE SMALL BUSINESS AND DEBT CONSOLIDATION ARE INHERENTLY RISKIER.
6. **BORROWER PROFILE**: LOW CREDIT SCORES, HIGH DEBT-TO-INCOME RATIOS, AND UNSTABLE EMPLOYMENT CAN INCREASE DEFAULT RISK.