Lending Club Case Study

Ankit Kumar Surana

Introduction

As an employee of a consumer finance company specialising in lending various types of loans to urban clients, part of my responsibility involves facilitating loan approval decision-making. This entails evaluating application profiles and identifying potential risks associated with loan repayment. To accomplish this task, I need to analyse the data provided in "loan.csv", which contains historical information about past loan applicants along with their default status. The goal is to identify patterns that indicate the likelihood of an applicant defaulting, enabling us to take appropriate actions such as denying a loan, adjusting loan terms, or applying higher interest rates to risky applicants.

Through this analysis, my objective is to gain insights into the consumer and loan attributes that influence the likelihood of default, as well as to identify the key driving factors or variables behind loan defaults. By understanding these factors, the company can improve its portfolio management and risk assessment strategies.

Problem Statement

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

- **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

- **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)

DataSet

The dataset is provided in a CSV file named "loan.csv", and the corresponding data dictionary is available in “Data_Dictionary.xlsx".

The dataset comprises 39,717 loan records and encompasses about 111 columns, which include both consumer and loan data. Furthermore, the data dictionary contains approximately 115 entries and is structured with 2 columns: one for the column names and another for their respective descriptions.

Preliminary Wrangling & Assessment

- The dataset does not contain any duplicate rows.
- Approximately 54 columns have all NULL values.
- About 9 columns have the same values across all records.
- The dataset comprises approximately 7 categorical variables: term, grade, sub_grade, verification_status, loan_status, purpose, and home_ownership.

Data Cleaning

1. Remove all the columns that are being used post loan approval.
2. Identify and remove rows where the loan status == "Current".
3. Remove the columns with all NULL or NaN values.
4. Remove the textual or masked columns that is irrelevant to analysis.
5. Remove all columns that have identical values across all rows.
6. Cleanse the data within columns containing % symbols.
7. Strip the alphabets from the sub-grade column.
8. Standardise the values within the emp_length column.
9. Round-off the amount value/ interest rate to the nearest two decimal places.
10. Convert the data-type of date columns to appropriate date formats.
11. Convert the cleaned % data in columns to float data type.
12. Convert the data type of columns having categorical value to categorical data types.
13. Decompose the date columns into smaller units like month and year.
14. Derive categorical variables from the loan_amnt and int_rate.
15. Derive a column margin from (annual_inc-loan).
16. Rename the columns for clarity by using full terms instead of abbreviations.
17. Address the missing values in the dataset through imputation or deletion.
18. Handle the outliers in the data.

Following the cleaning process, the dataset now contains 38,577 records and 23 columns. This reflects a reduction of approximately 2.87% in the number of records and an 74% reduction in the number of columns with 6 new derived columns. Subsequently, we will employ this refined dataset for our Exploratory Data Analysis (EDA), encompassing univariate, segmented univariate, bivariate, and multivariate analyses.

Loan Attributes

- term → Categorical Data Type
- issue_d → DateTime Data Type
- grade → Categorical Data Type
- sub_grade → Categorical Data Type
- verification_status → Categorical Data Type
- loan_status → Categorical Data Type
- purpose → Categorical Data Type
- loan_amnt → Float Data Type
- funded_amnt → Float Data Type
- funded_amnt_inv → Float Data Type
- int_rate → Float Data Type
- installment → Float Data Type

Customer Attributes

- annual_inc → Float Data Type
- debt_to_income → Float Data Type
- pub_rec_bankruptcies → Float Data Type
- home_ownership → Categorical Data Type
- addr_state → String Data Type
- emp_length → Categorical Data Type

Derived Attributes

1. issue_d_year → Integer Data Type
2. issue_d_month → Categorical Data Type
3. loan_amnt_b → Categorical Data Type
4. debt_to_income_b → Categorical Data Type
5. int_rate_b → Categorical Data Type
6. margin → Float Data Type

Exploratory Data Analysis

Univariate Analysis

→ Mean, Median, Max, Min, Std, Variance, Count → Distribution (Histogram, CountPlot, BoxPlot)

Bivariate Analysis

→ Relationship Between 2 Variables (ScatterPlot, BoxPlot, BarPlot etc)

Multivariate Analysis

→ Relationship Between more variables (Heatmap etc.)

Numerical_Columns : loan_amnt, funded_amnt, funded_amnt_inv, int_rate, installment, annual_inc, debt_to_income, pub_rec_bankruptcies, issue_d_year, margin
Cateogrical_Columns : term, grade, sub_grade, emp_length, home_ownership, verification_status, loan_status, issue_d_month, loan_amnt_b, int_rate_b, debt_to_income_b
Extra_Columns : issue_d, purpose, addr_state

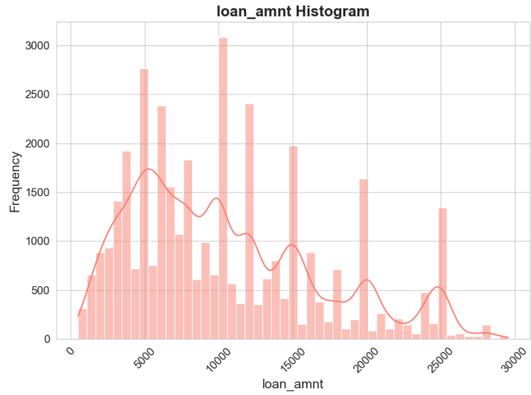
Univariate Analysis

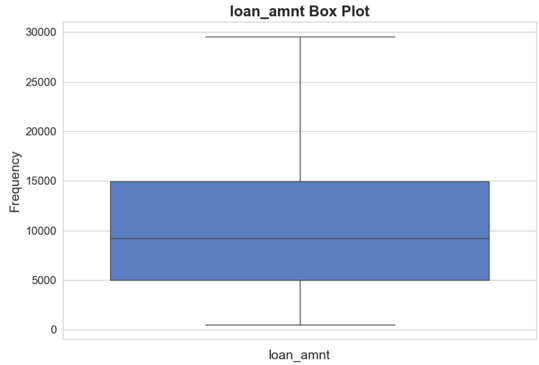
Exploratory Data Analysis

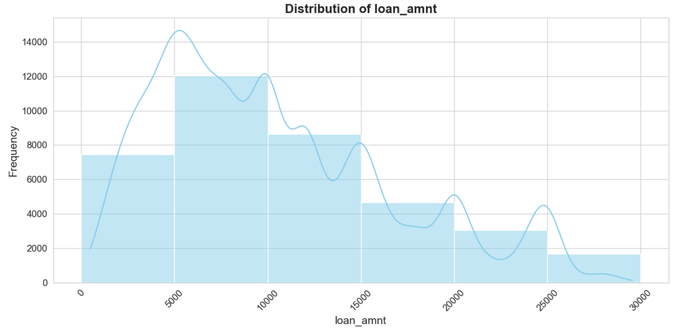
Univariate Analysis

- Statistical summary for loan_amnt:
  - count 37489.000000
  - mean 10408.101043
  - std 6398.162546
  - min 500.000000
  - 25% 5000.000000
  - 50% 9250.000000
  - 75% 14975.000000
  - max 29550.000000

- The mode of loan_amnt is: 10000.0







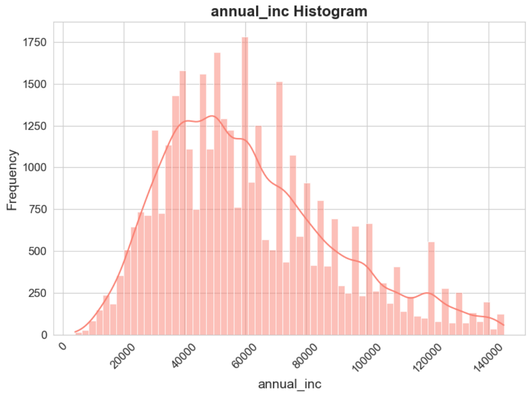
The distribution depicted above is right-skewed, indicating that the majority of loan application amounts fell between 5000-10000, followed by 0-5000 and then 10000-15000. However, the mean loan amount is 10678 and the mode is 10000.

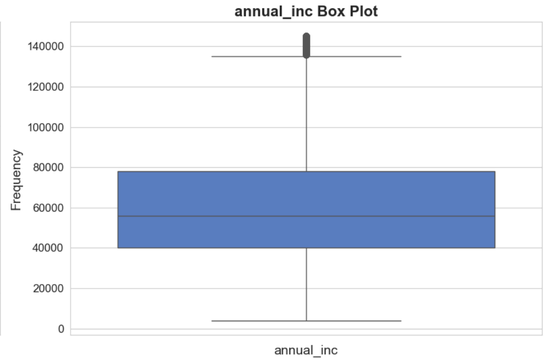
Exploratory Data Analysis

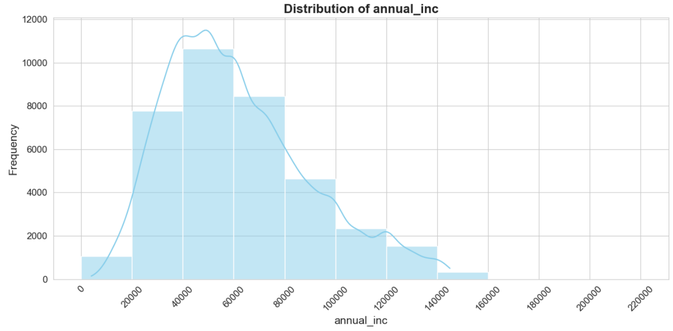
Univariate Analysis

- Statistical summary for annual_inc:
  - count 36815.000000
  - mean 61218.193490
  - std 28224.583784
  - min 4000.000000
  - 25% 40000.000000
  - 50% 56000.000000
  - 75% 78000.000000
  - max 145000.000000

- The mode of annual_inc is: 60000.0







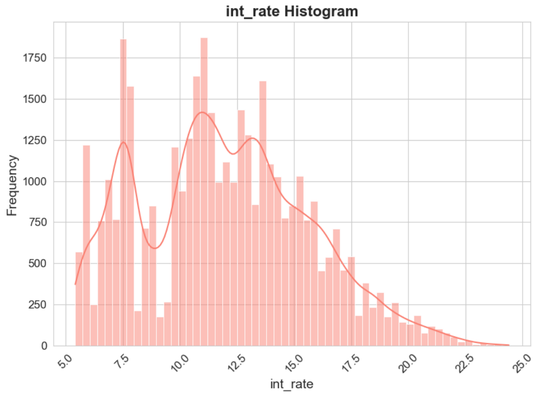
The above distribution is right skewed and most of the loan application where from customers whose annual income lies between 40000-60000. The mean of annual income of the customers is 61218 and the mode is 60000.

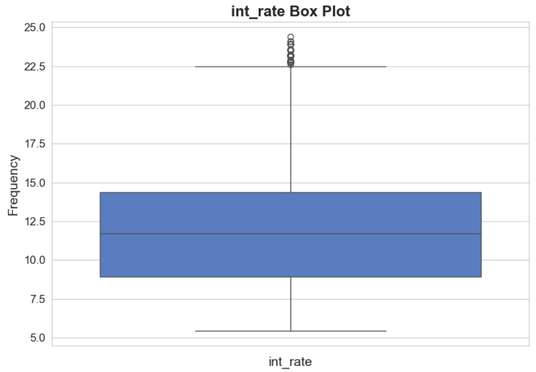
Exploratory Data Analysis

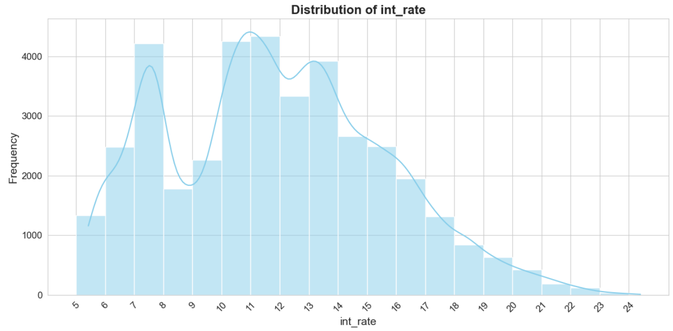
Univariate Analysis

- Statistical summary for int_rate:
  - count 38577.000000
  - mean 11.932219
  - std 3.691327
  - min 5.420000
  - 25% 8.940000
  - 50% 11.710000
  - 75% 14.380000
  - max 24.400000

- The mode of int_rate is: 10.99







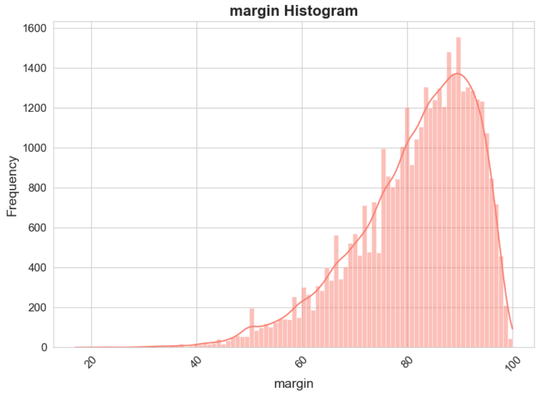
The above distribution is non-symmetric bimodal and most of the loans had an interest rate between 11%-12%, followed by 10%-11% and then 7%-8%. Also there is a sudden drop in loans where the interest rate is 7%-9%.

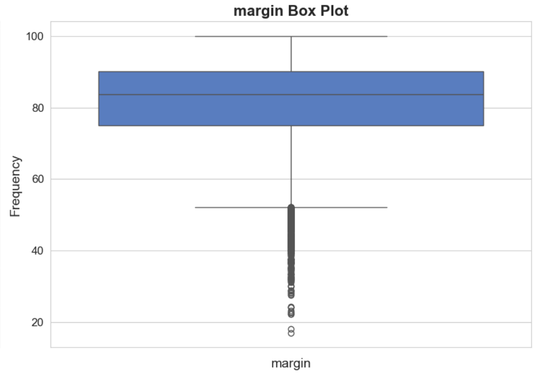
Exploratory Data Analysis

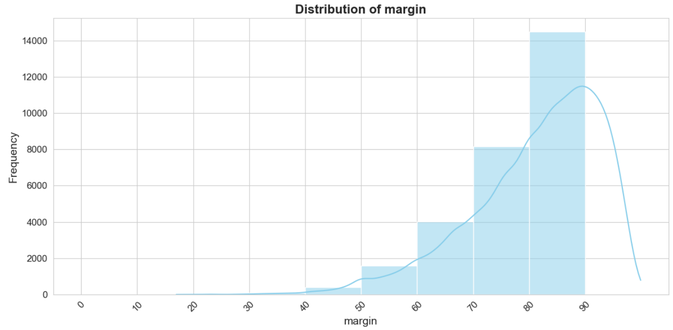
Univariate Analysis

- Statistical summary for margin:
  - count 38577.000000
  - mean 81.416662
  - std 11.524274
  - min 17.000000
  - 25% 75.000000
  - 50% 83.700000
  - 75% 90.200000
  - max 99.920000

- The mode of margin is: 80.0







Large chunk of margin is between 80%-90%, indicating that loan_amnt taken by customers is higher than their annual income.

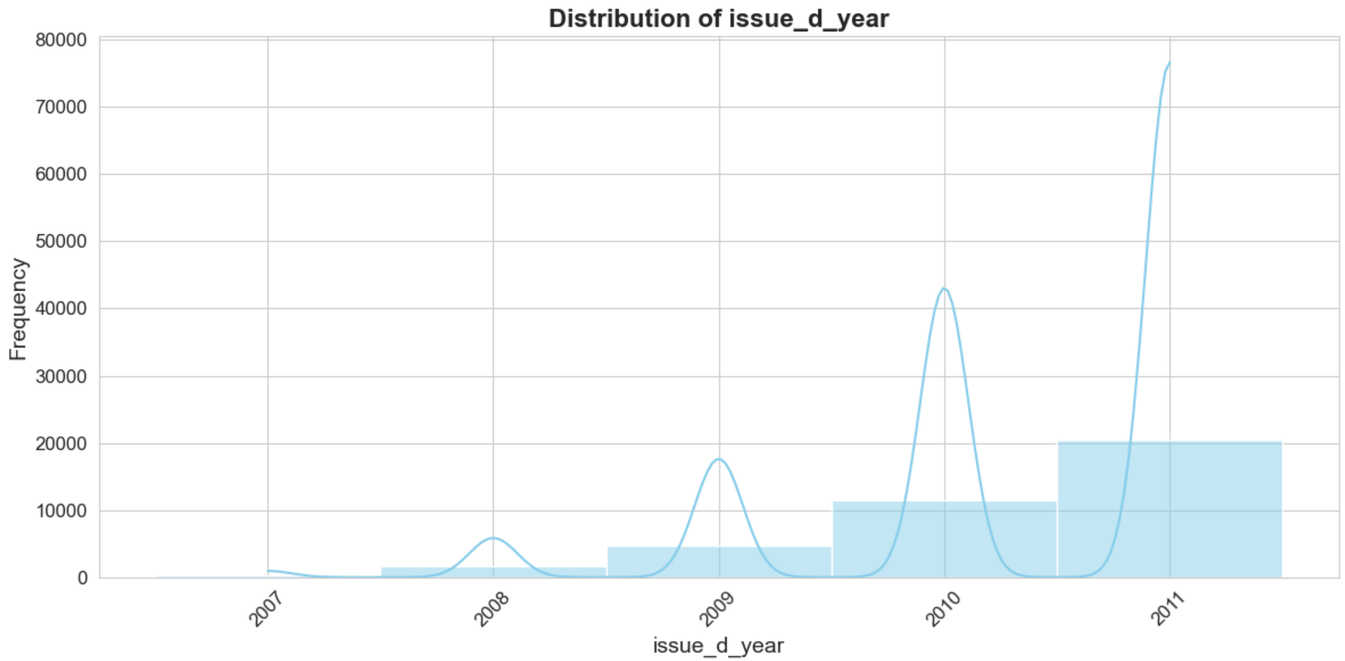
Exploratory Data Analysis

Univariate Analysis

- Statistical summary for issue_d_year:
  - count 38577.000000
  - mean 2010.309070
  - std 0.882658
  - min 2007.000000
  - 25% 2010.000000
  - 50% 2011.000000
  - 75% 2011.000000
  - max 2011.000000

- The mode of issue_d_year is: 2011

- Count summary for issue_d_year:
  - 2011 20516
  - 2010 11532
  - 2009 4716
  - 2008 1562
  - 2007 251



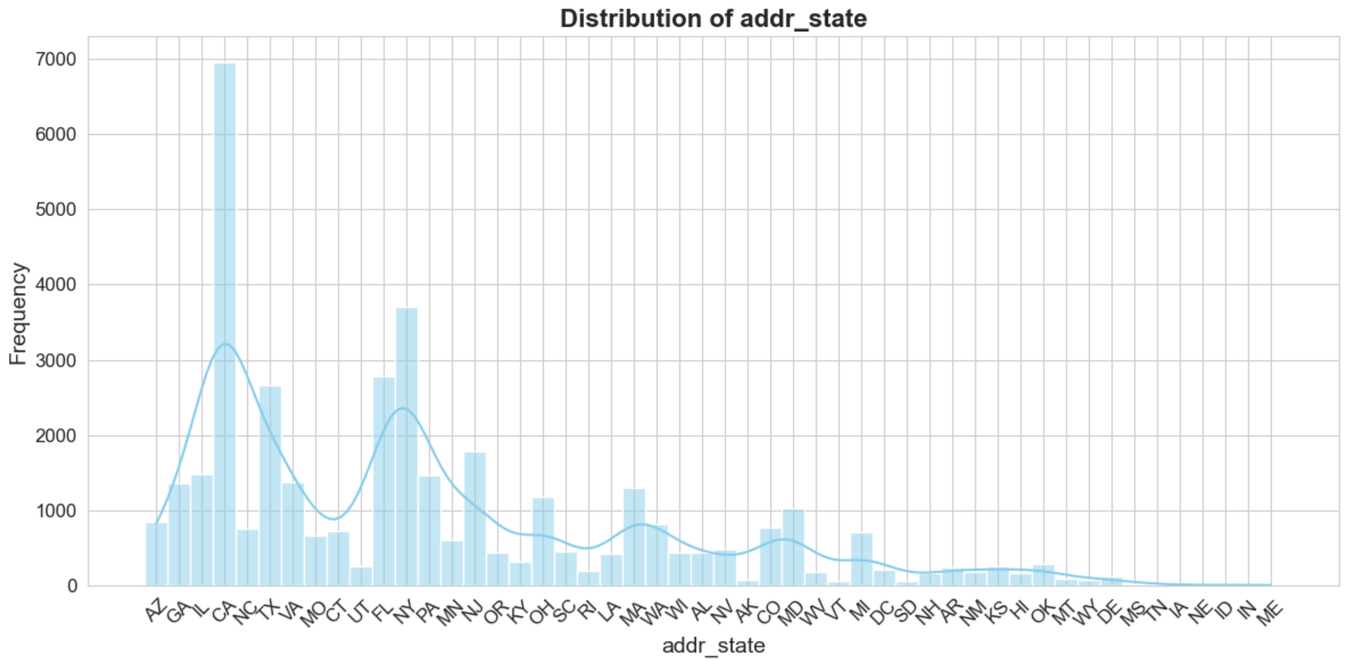
The distribution shown above exhibits left skewness, indicating that the majority of loans were issued in the year 2011, while the least number of loans were issued in 2007.

Exploratory Data Analysis

Univariate Analysis

- Statistical summary for addr_state:
  - count 38577
  - unique 50
  - top CA
  - freq 6949

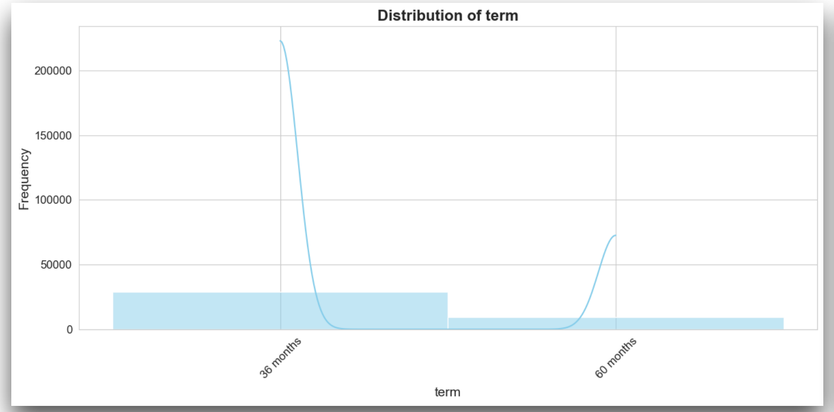
- The mode of addr_state is: CA

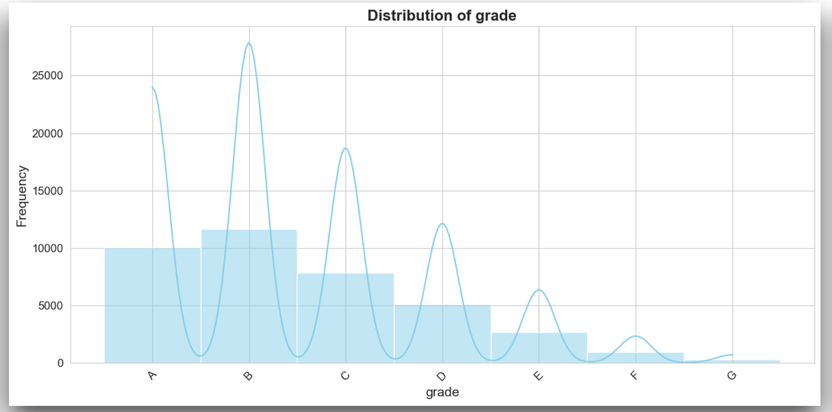


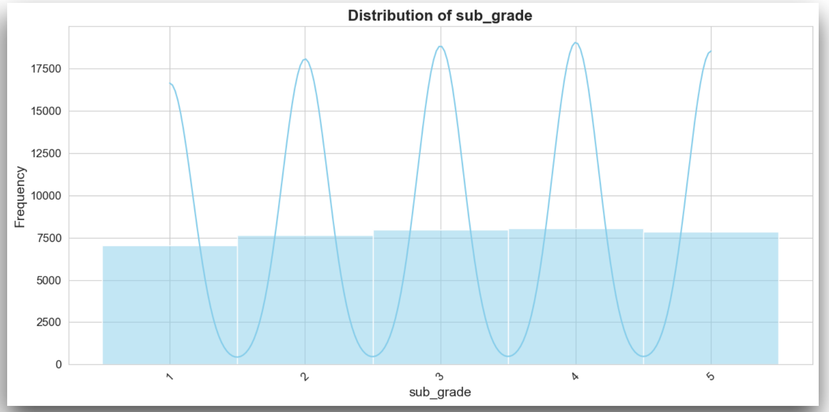
The top 5 state's from the dataset are CA, NY, FL, TX and NJ. However CA tops the list with most loan being taken there (around 6949 Loans).

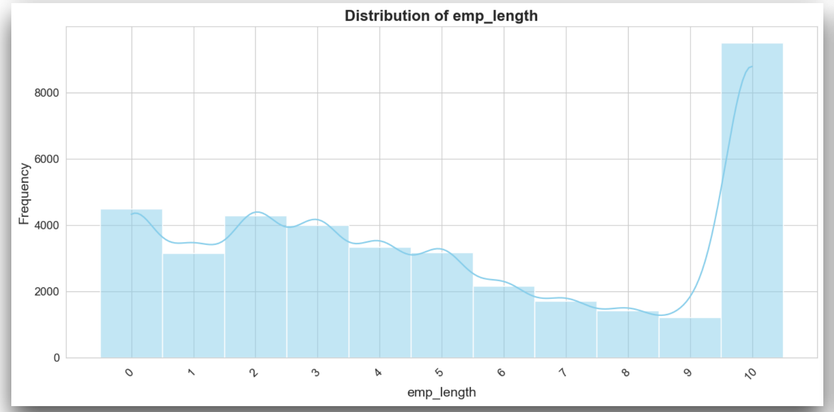
Exploratory Data Analysis

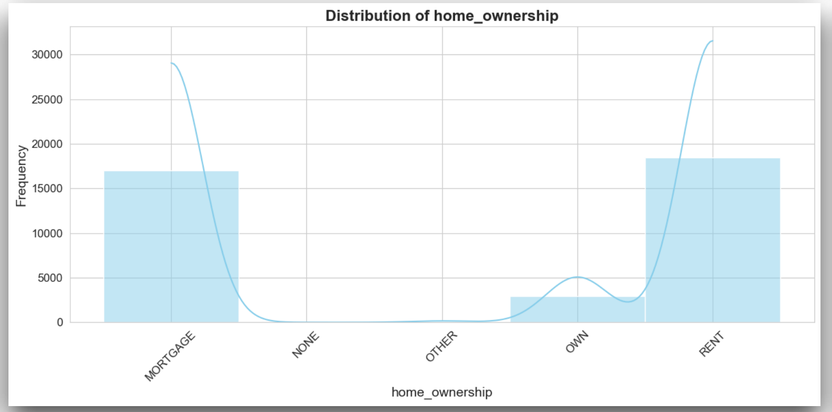
Univariate Analysis

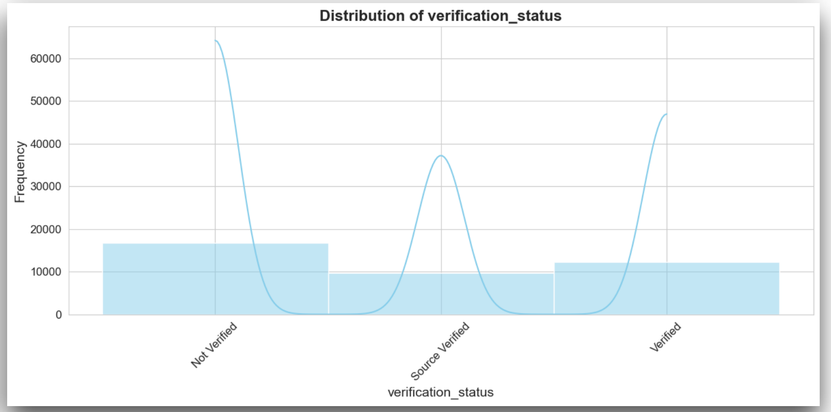






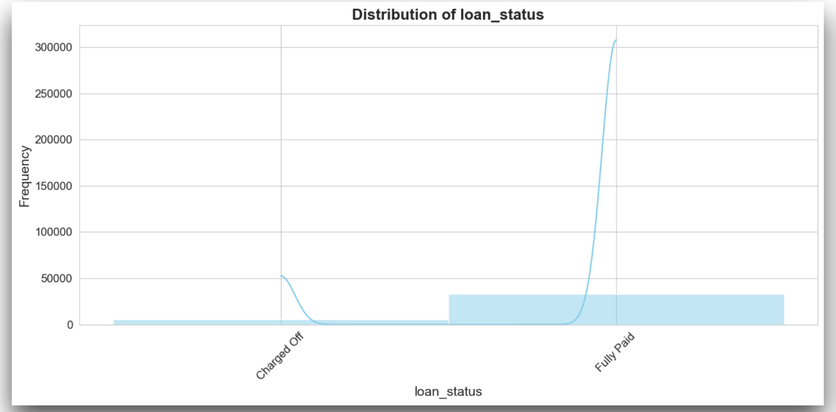


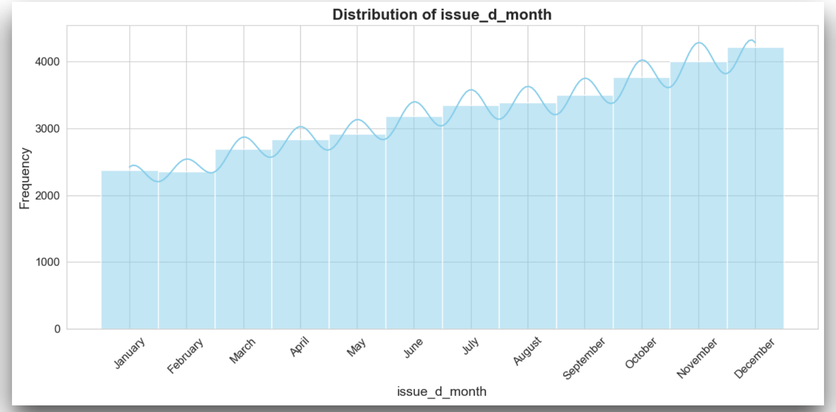


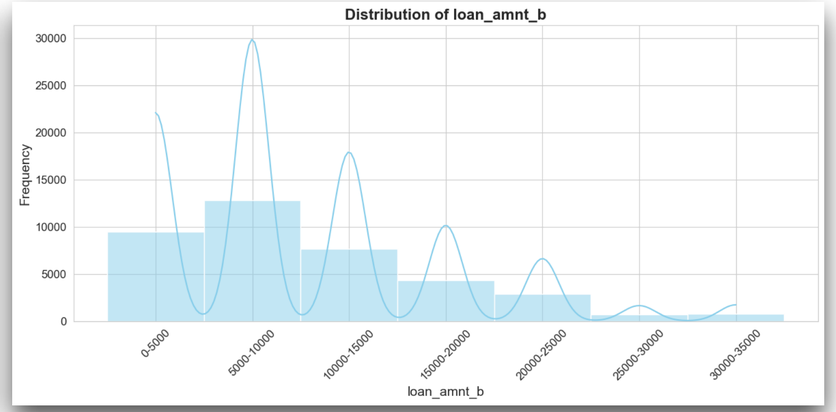


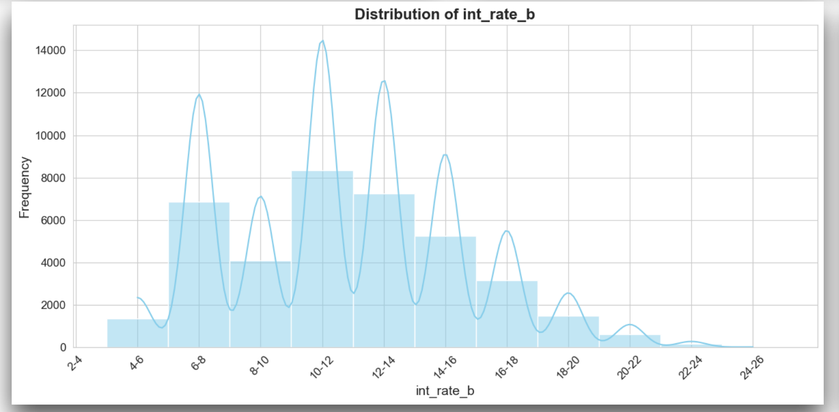
Exploratory Data Analysis

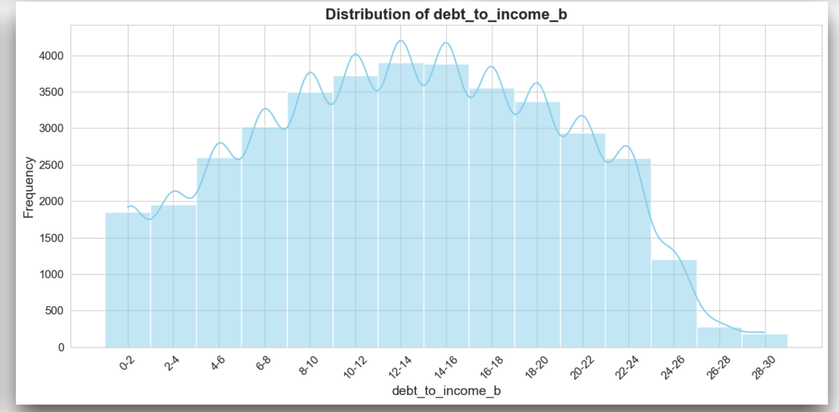
Univariate Analysis











Exploratory Data Analysis

Univariate Analysis

1. **debt_to_income_b**: It is a normally distributed graph with peak between 12 and 14 which indicates that the most of customers had a debt-to-income ratio within this range.
2. **loan_amount_b**: The graph indicates a right-skewed distribution with the peak indicating that most of loans are for amounts between 5,000 and 10,000.
3. **int_rate_b**: The distribution indicates that most of loans are between 10%-12%.
4. **issue_b_month**: The distribution is left-skewed, which indicates that most loan applications were issued in December,and the lowest in February.
5. **loan_status**: There are around 32,950 customers who have their loan status as fully paid while around 5,627 customers have loan status as charged-off.
6. **verification_status**: Around 16,694 customers who have been issued loans arent verified.
7. **home_ownership**: The majority of customers (around 18,480) have a home ownership status of "RENT," followed by 17,021 with "MORTGAGE," and only 2,975 who own their home.
8. **emp_length**: Customers with over 10 years of employment have applied for the most loans and that is around 9,131 applications.
9. **sub_grade**: Loans issued under subgrade 4 are the most common, while subgrade 1 has the least.
10. **grade**: The highest number of loans are issued under grade B (around 11,675), while the lowest is for grade G.
11. **term**: The majority of loans have a term of 36 months.

Segmented Univariate Analysis

Exploratory Data Analysis

Segmented Univariate Analysis

Segmenting the loan status into 'fully_paid' and 'charged_off' and analysing the variables that impact the loan status.

Loan Status → Fully Paid

- Created a new data frame "df_fully_paid" for analysis
- Number of Rows in df_fully_paid after Cleaning : 32950
- Number of columns in df_fully_paid after Cleaning: 24

Loan Status → Charged Off

- Created a new data frame "df_charged_off" for analysis
- Number of Rows in df_charged_off after Cleaning : 5627
- Number of columns in df_charged_off after Cleaning: 24

Exploratory Data Analysis

Segmented Univariate Analysis

- **Interest Rate (int_rate)**: Loans which have interest rates between 10%-12% have a higher chance of loan status being fully paid off. However, loans having interest rates between 12%-14% show a higher likelihood of charge-off.
- **Annual Income (annual_inc)**: Loans Status fully paid when the customer's annual income is between 45K-60K. In contrast, the customers having an annual income between 30K-45K tend to have more charged-off loans.
- **Grade (grade)**: Borrowers with 10 or more years of employment have the highest count of loans and also is the highest in fully paid and charged-off segments. This indicates that they take more loans and are also prone to default more.
- **Employment Length (emp_length)**: Borrowers having 10 or more years of employment have the highest loan counts in fully paid and default segments. This indicates they take more loans and also default more.
- **Verification Status (verification_status)**: Loans to borrowers who are not verified show higher peaks in being paid off compared to those who are verified. This indicates that unverified borrowers tend to repay their loans more frequently.
- **Home Ownership (home_ownership)**: Renters have a higher default rate, which might be explained by their higher expenses towards rent.

Bivariate Analysis

Exploratory Data Analysis

Bivariate Analysis - Numerical vs Numerical

Loan Amount vs. Interest Rate:
- Slight positive correlation: Higher loan amounts correlate with slightly lower interest rates.

Loan Amount vs. Installment:
- Positive correlation: Larger loans entail higher installment payments.

Loan Amount vs. Annual Income:
- Weak positive correlation: As annual income increases, loan amount tends to rise, albeit weakly.

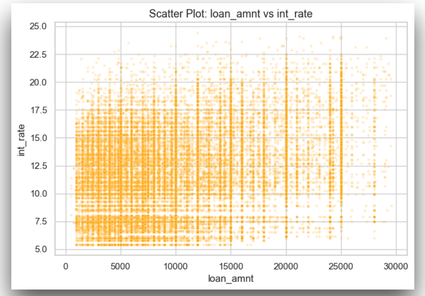
Loan Amount vs. Public Record Bankruptcies:
- Negative correlation: Public record bankruptcies minimally affect the loan amount qualification.

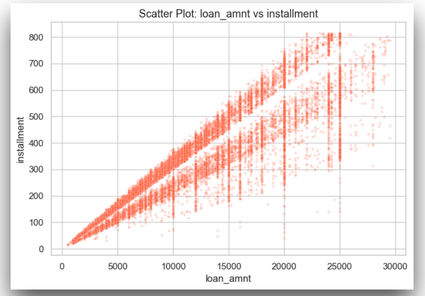
Annual Income vs. Interest Rate:
- Weak negative correlation: Drawing conclusions from this scatter plot alone is challenging due to the weak correlation.

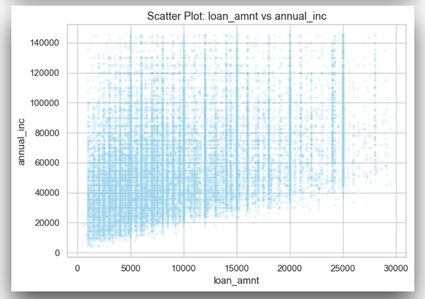
Annual Income vs. Debt-to-Income Ratio:
- Negative correlation: As annual income increases, the debt-to-income ratio tends to decrease.

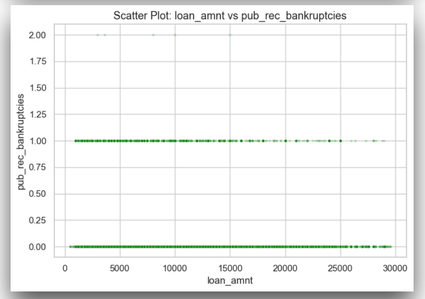
Margin vs. Debt-to-Income Ratio:
- As the margin increases, the debt-to-income ratios vary more widely.

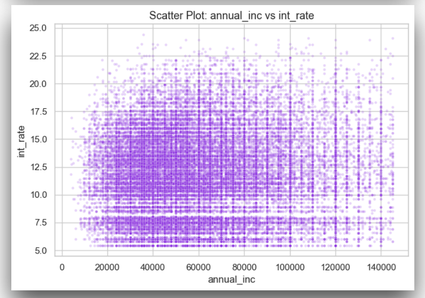
Margin vs. Int_Rate Ratio:
- The margin doesn't seem to affect the interest rate.

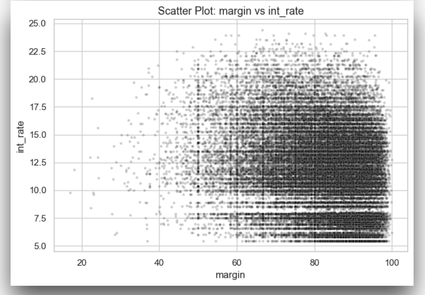












Exploratory Data Analysis

Bivariate Analysis - Categorical vs Numerical

Term vs Loan Amount:

- Loans with a 60-month term have higher average amounts as compared to those with a 36-month term.

Grade vs Loan Amount:

- Median of the loan amounts increases with higher grades (A to G except C), that indicates a strong positive correlation.
- Grades E, F, and G exhibit wider variety in loan amounts when compared to other grades.

Employment Length vs. Loan Amount:

- There is no clear correlation between employment length and the loan amount. Variety in loan amounts differs across employment lengths.

Loan Status vs. Loan Amount:

- Median of the loan amounts is almost same for charged-off and fully paid loans, but charged-off loans exhibit greater variety.

Loan Status vs. Interest Rate:

- Charged-off loans have higher median interest rates than the fully paid loans, thereby suggesting a potential correlation between interest rate and loan default.

Grade vs. Interest Rate:

- Median of the interest rate increases with lower loan grades (A to G), indicating that riskier loans carry higher rates.
- Some grades (e.g., B, C, D, and E) show wider variety in interest rates.

Verification Status vs. Loan Amount:

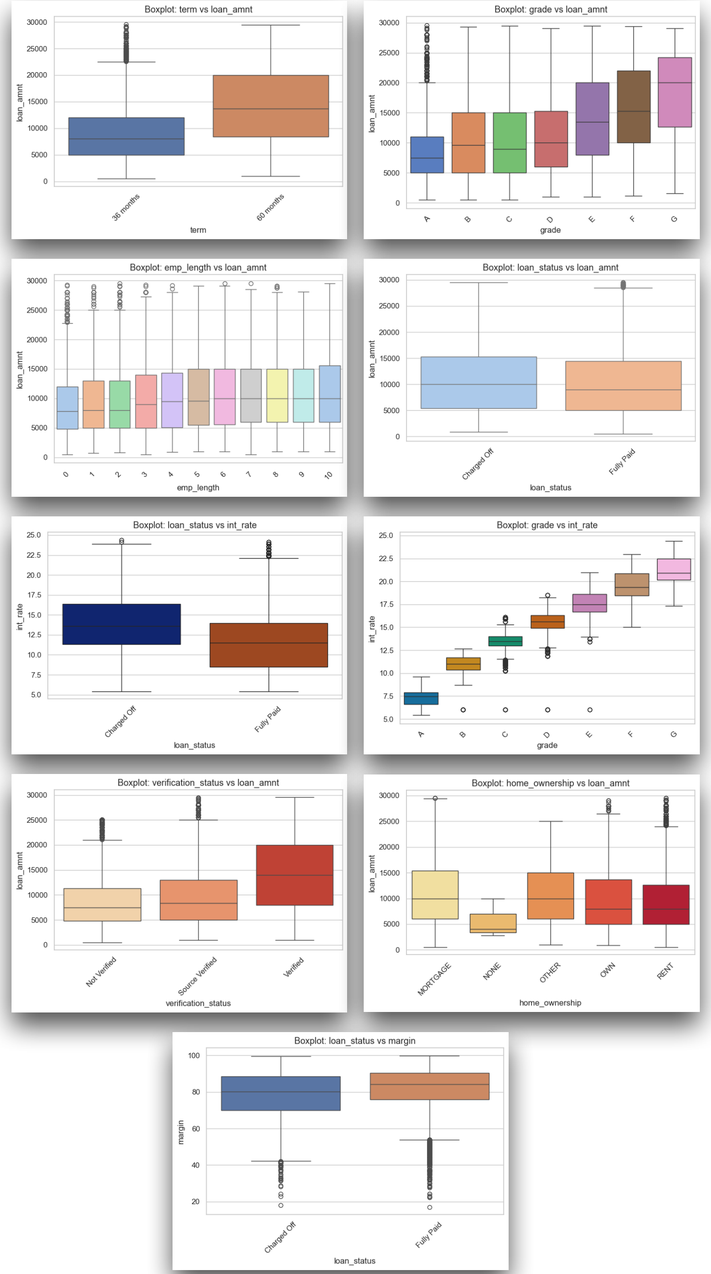
- There is no much difference in median income of verified and unverified customers. Unverified applicants show wider variety in income.

Home Ownership vs. Loan Amount:

- Median of the loan amounts is high for customers with mortgages as compared to renters and others, suggesting a correlation between homeownership and the loan amount.

Margin vs. Loan Status:

- People who fully repaid their loans had a bigger gap between their income and loan amount than those who didn't repay.



Exploratory Data Analysis

Bivariate Analysis - Categorical vs Categorical

- **Comparison between Loan Status and Term:**
  - Risk Levels (Charged Off Rates):
    - Less Risky: 36 months (11.11%)
    - More Risky: 60 months (25.05%)
  - Distribution of Loan Terms:
    - More Common: 36-month loans
    - Less Common: 60-month loans

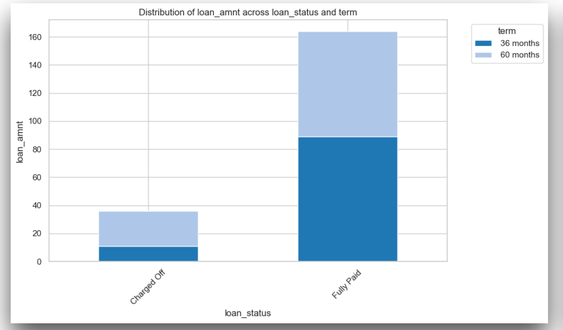
- **Comparison between Loan Status and Loan Amount:**
  - Risk Levels (Charged Off Rates):
    - Less Risky: 5000-10000 (12.83%), 0-5000 (13.87%), 10000-15000 (13.73%)
    - More Risky: 15000-20000 (17.27%), 20000-25000 (19.11%), 25000-30000 (20.46%)
  - Distribution of Loan Amounts:
    - More Common: Loans in the 5000-10000 and 0-5000 ranges
    - Less Common: Loans in the 25000-30000 and 20000-25000 ranges, indicating fewer larger loans.

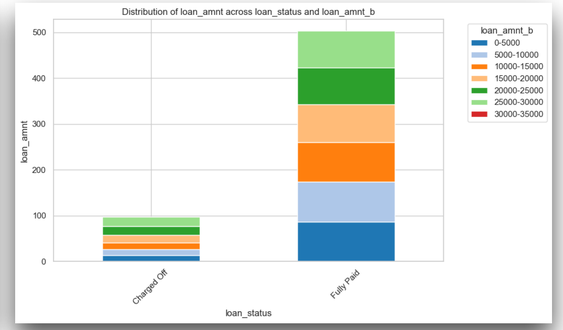
- **Comparison between Loan Status and Interest Rate:**
  - Risk Levels (Charged Off Rates):
    - Less Risky: 4-6% (4.09%), 6-8% (5.6%), 8-10% (9.61%)
    - More Risky: 16-18% (26.14%), 18-20% (30.34%), 20-22% (36.92%), 22-24% (46.08%)
  - Distribution of Interest Rates:
    - More Common: Loans with 10-12% and 6-8% interest rates
    - Less Common: Loans with 22-24% and 20-22% interest rates, indicating fewer high-risk, high-interest loans.

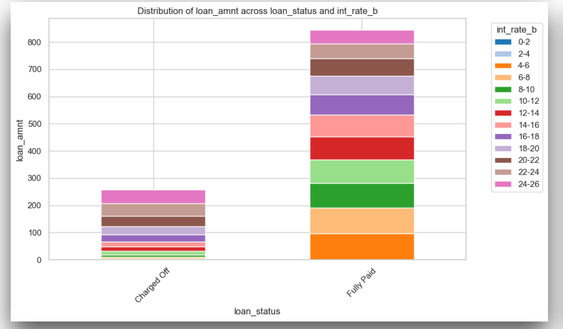
- **Comparison between Loan Status and Grade:**
  - Risk Levels (Charged Off Rates):
    - Less Risky: Grade A (6.01%), B (12.15%), and C (17.13%)
    - More Risky: Grade D (21.87%), E (26.89%), F (31.8%), and G (34.26%)
  - Distribution of Loan Grades:
    - More Common: Grades B and A
    - Less Common: Grades F and G, indicating fewer high-risk loans.

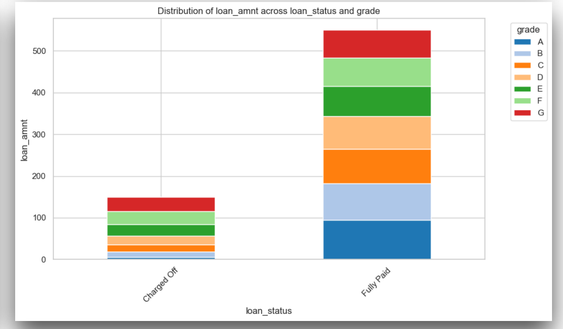
- **Comparison between Loan Status and Home Ownership:**
  - Risk Levels (Charged Off Rates):
    - More Risky: OTHER (18.37%), RENT (15.18%), OWN (14.67%)
    - Less Risky: MORTGAGE (13.31%), NONE (0.0%)
  - Distribution of Home Ownership:
    - High Common: RENT and MORTGAGE are most common.

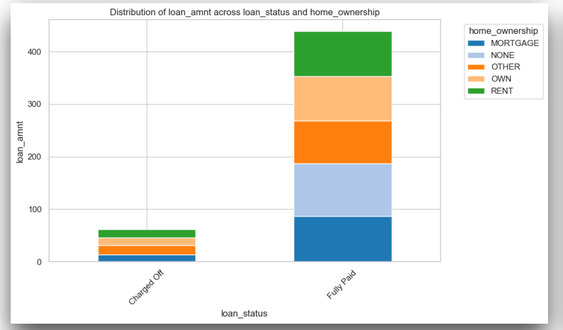
- **Comparison between Loan Status and Purpose:**
  - Risk:
    - High Risk: Small business, renewable energy, educational loans.
    - Low Risk: Wedding, car, credit card, major purchase loans.
  - Demand:
    - High Demand: Debt consolidation, credit card loans.
    - Low Demand: Renewable energy, educational loans.

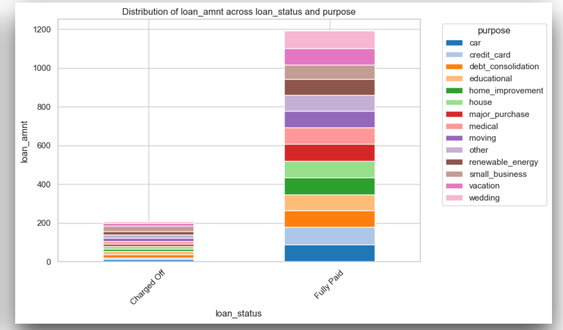












Multivariate Analysis

Multivariate Analysis

Loan Amounts:

- Strong positive correlations with funded amount and funded amount invested, moderate correlation with installment, and weak correlation with borrower characteristics and interest rates.

Funded Amount:

- Strong positive correlations with loan amount and funded amount invested, moderate correlation with installment, and weak correlation with borrower characteristics and interest rates.

Funded Amount Invested:

- Strong positive correlations with loan amount and funded amount, moderate correlation with installment, and weak correlation with borrower characteristics and interest rates.

Interest Rate:

- Weak correlations with loan attributes, indicating some association but not particularly strong.

Installment:

- Strong positive correlations with loan amount, funded amount, and funded amount invested, weak correlation with borrower characteristics and interest rates.

Annual Income:

- Weak correlations with loan attributes, indicating a slight association with loan amounts but not particularly strong.

Debt-to-Income Ratio:

- Weak correlations with loan attributes, suggesting some association but not particularly strong.

Public Record Bankruptcies:

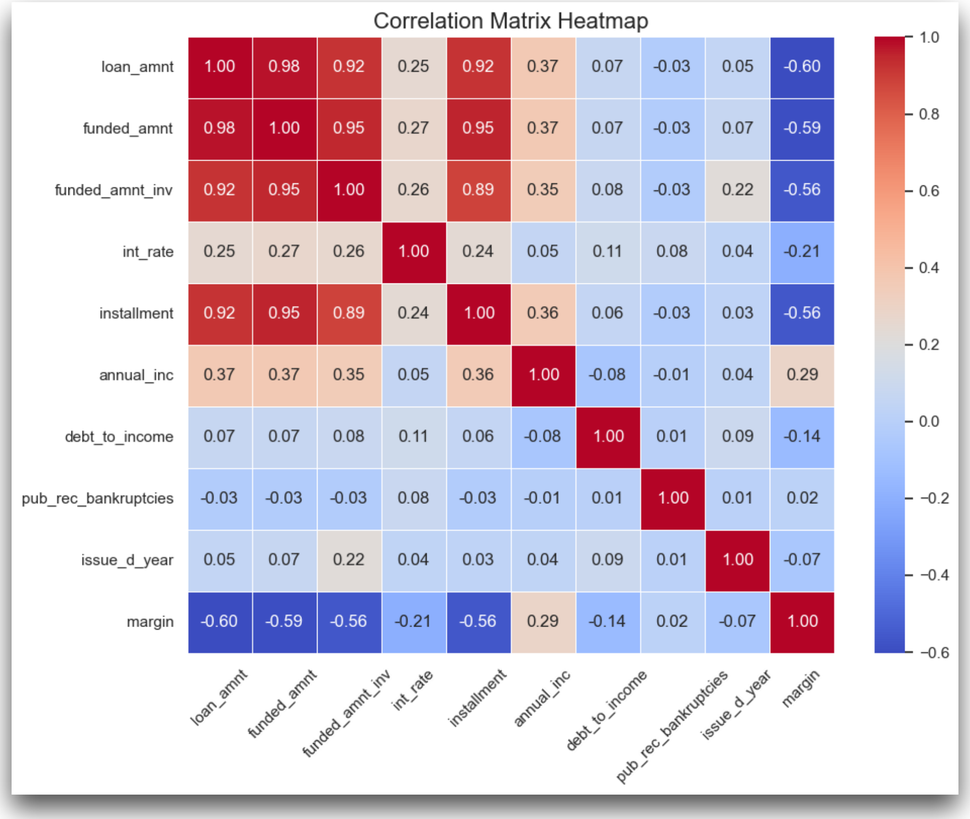
- Almost negligible correlation with loan attributes, indicating minimal association with borrower characteristics and loan terms.

Issue Date Year:

- Negligible correlation with loan attributes except for a somewhat stronger correlation with funded amount invested.

Margin:

- There's no correlation with loan attributes except for a somewhat positive correlation with annual_inc invested.



Recommendations

Based on the above analysis, below are the factors that influence loan defaults:

- **Loan Term**: Loans with longer terms (60 months) have higher default rates compared to shorter-term loans (36 months), indicating higher risk or borrower instability.

- **Loan Amount**: For loan amounts exceeding 15,000, there is a trend of higher default rates. Borrowers may find it difficult to make repayments for larger loans, leading to increased charge-offs.

- **Interest Rate**: Loans with higher interest rates, such as 16%, are associated with higher charge-off rates. High-interest loans may attract riskier customers.

- **Loan Grade**: Lower-grade loans (D, E, F, G) have higher charge-off rates compared to higher-grade loans (A, B, C).

- **Home Ownership**: Non-traditional home ownership, such as renting or other categories, is linked with higher charge-off rates compared to mortgage holders.

- **Loan Purpose**: Loans for small businesses, renewable energy, and education exhibit higher charge-off rates, while loans for weddings, cars, credit cards, and major purchases have lower charge-off rates.

These factors provide insights into customer behavior that can help lenders understand and mitigate loan charge-off risks.