Lending Club Case Study

Exploratory Data Analysis

Submitted by
Vineeth B

Problem Statement

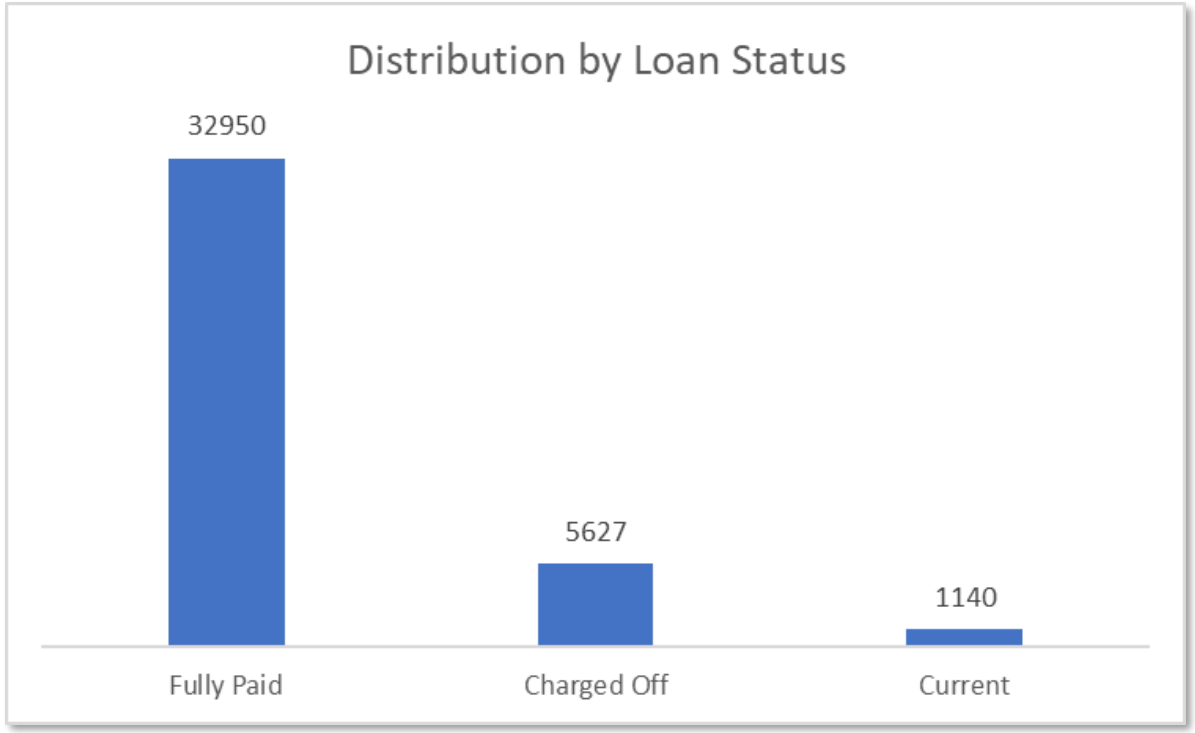
To understand the driving factors behind loan default, i.e. the variables which are strong indicators of default for a consumer financial services company to improve risk assessment.

Objectives:

- Identify patterns in default rates.
- Understand borrower demographics and financial characteristics.
- Provide insights for improving loan risk assessment.

Dataset Overview

- Dataset Size – 39717 rows (loans) and 111 columns (features)
- Dependent Variable – "loan_status"
- Default Rate - 14.59% (ignoring Loan Status = Current)



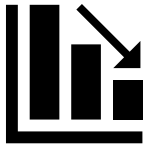
EDA Approach

1. Data Pre-Processing

- Remove columns that do not add value to analysis
- Missing Value Analysis and Treatment
- Data Type Conversion
- Derived Columns generation for better insights

2. Univariate Analysis

- Outlier Analysis of numerical variables through box plot Visualize Distribution of loans by binning numerical variables
- Visualize Distribution of loans by categorical variables



3. Bivariate Analysis

- Relation between categorical and numerical variables through box plot
- Correlation Matrix to understand correlation among numerical variables and with default rate
- Comparison of Default rate with all variables

Data Pre-Processing

1. Data Reduction Approach

- Loan status = current
- Columns that were >95% null
- Columns that had only one unique value
- Columns that are at application level like id, url, desc, title etc.,
- Columns that will not be available at application time – like recoveries

2. Missing Value Treatment & Data Cleaning

- Replace missing values with Mode for columns like emp_length, revol_util etc.,
- Convert date column like issue date into datetime format
- Convert emp length into numeric
- Remove % from rate columns like interest rate

3. Derived Columns

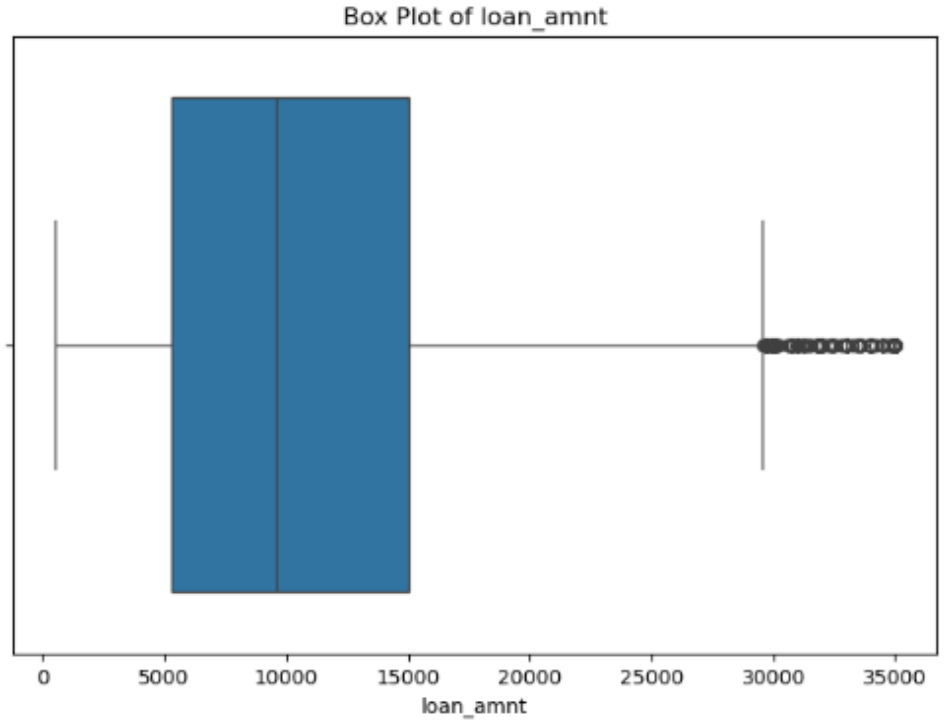
- Loan Status in Binary
- Issue Month and Year
- Never Delinquent Flag
- Credit History in Years
- Installment to Monthly Income Ratio
- Installment to Loan Amount Ratio

Data Understanding - Univariate & Bivariate Analysis

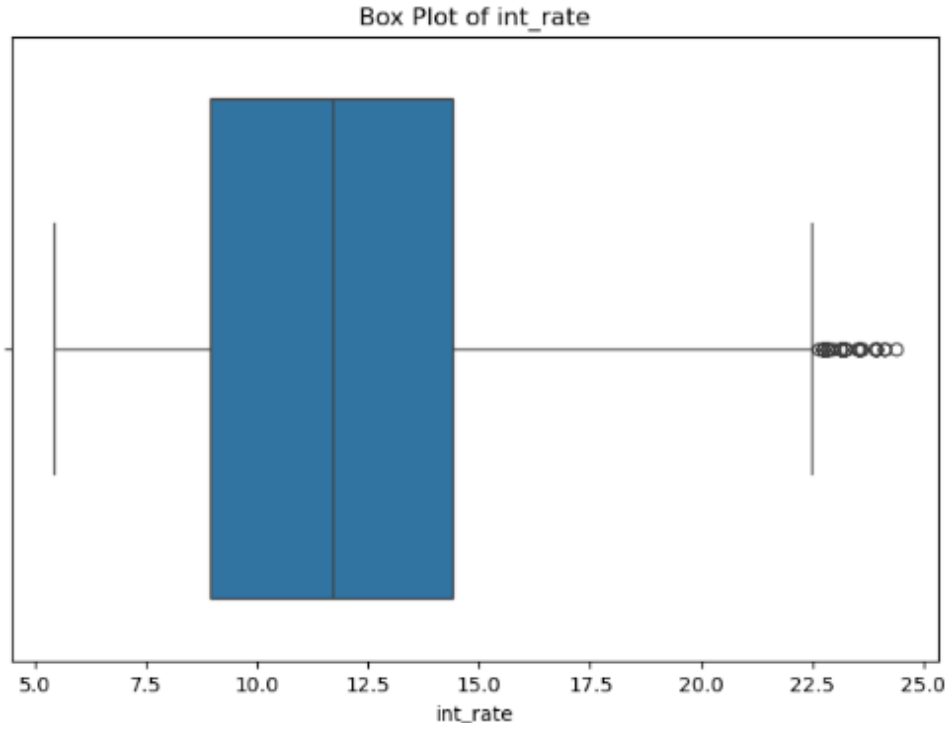
Univariate Analysis

Distribution of Major Numerical Columns

Median Loan Amount is ~$10K



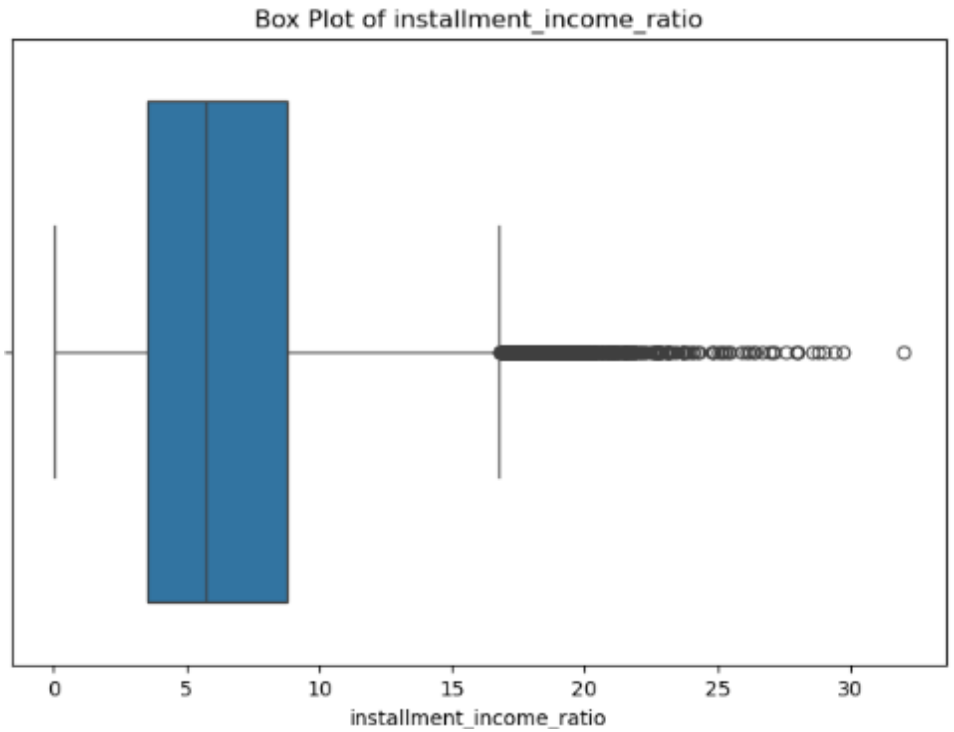
Median Interest Rate is ~12%



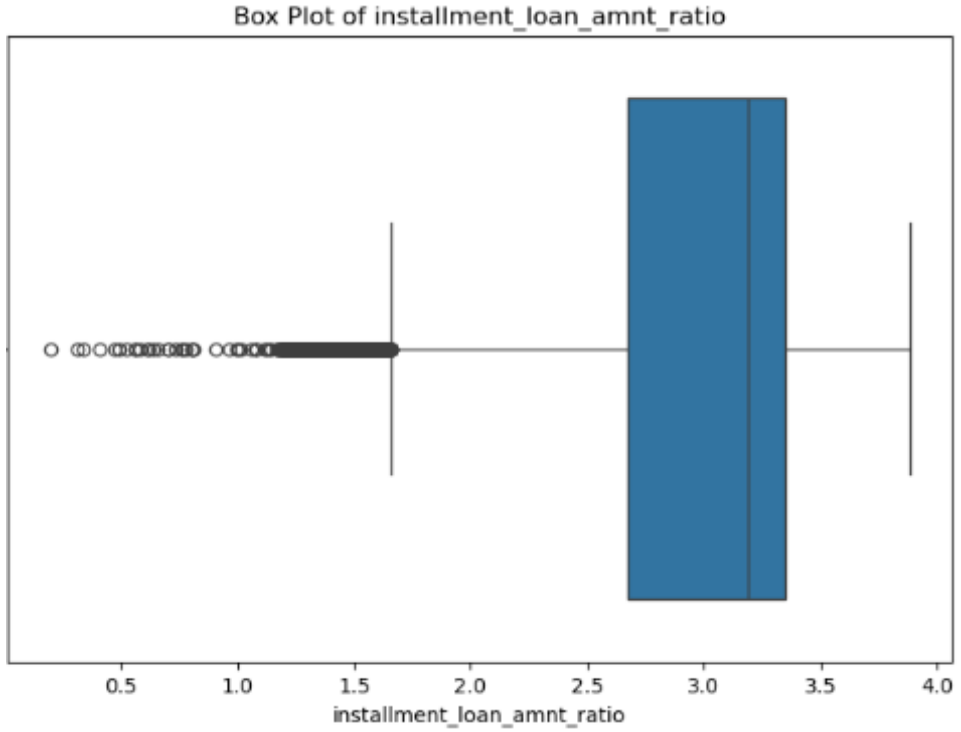
Univariate Analysis

Distribution of Major Numerical Columns

Installment amount is Less than 15% of Monthly Income for most of the applicants



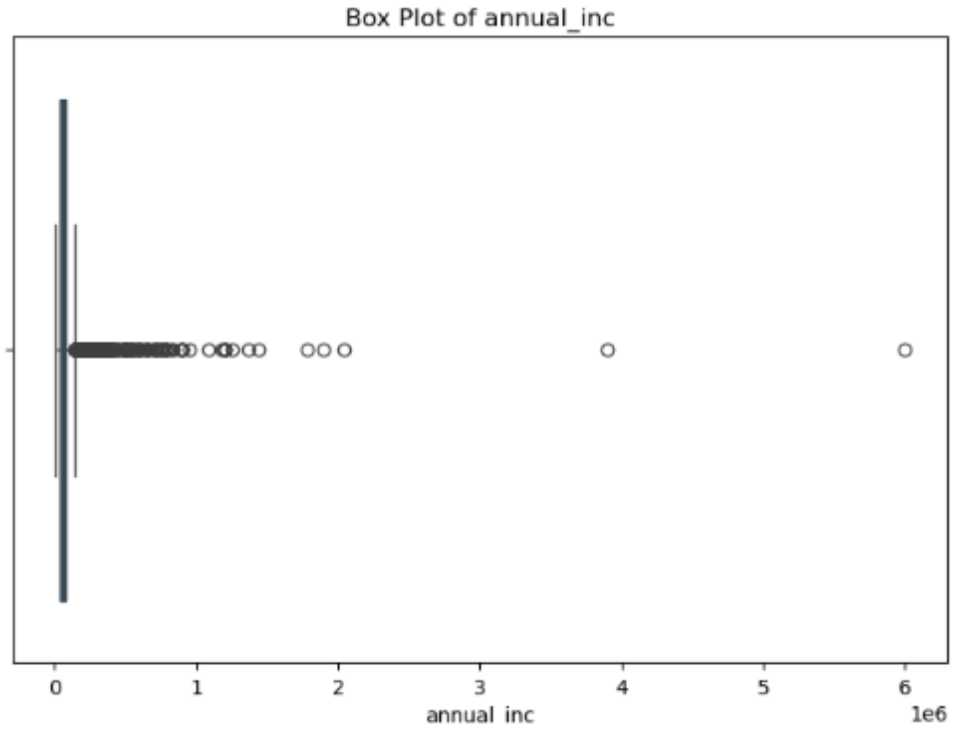
~60% applicants had installment value between 2.5% and 3.5% of loan amount



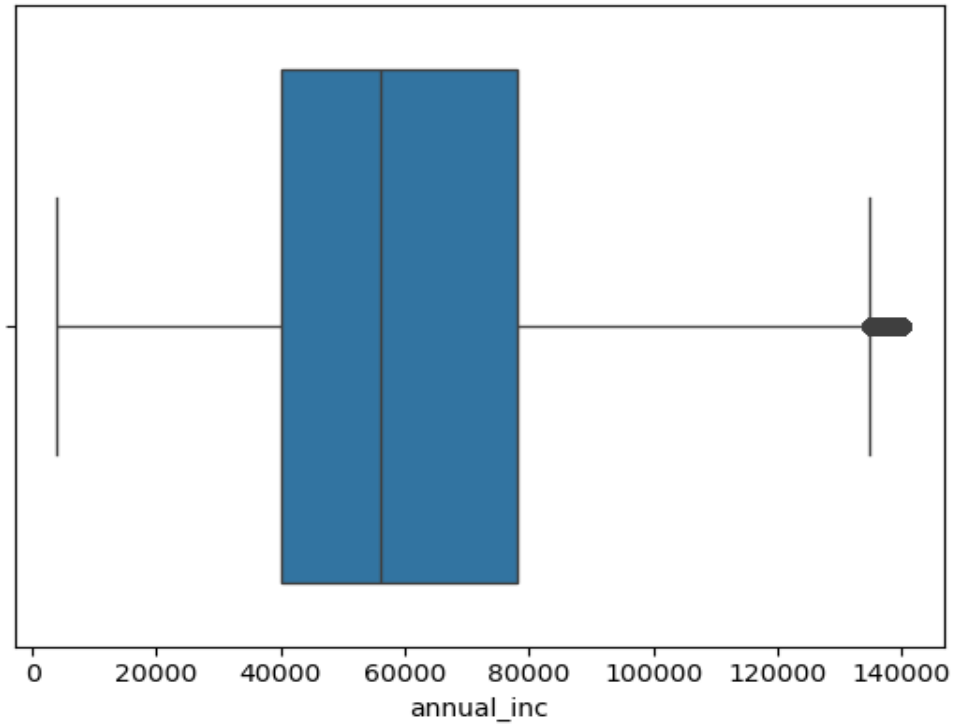
Univariate Analysis

Fixing Outlier for Income

Distribution on Annual Income - Before Outlier Treatment

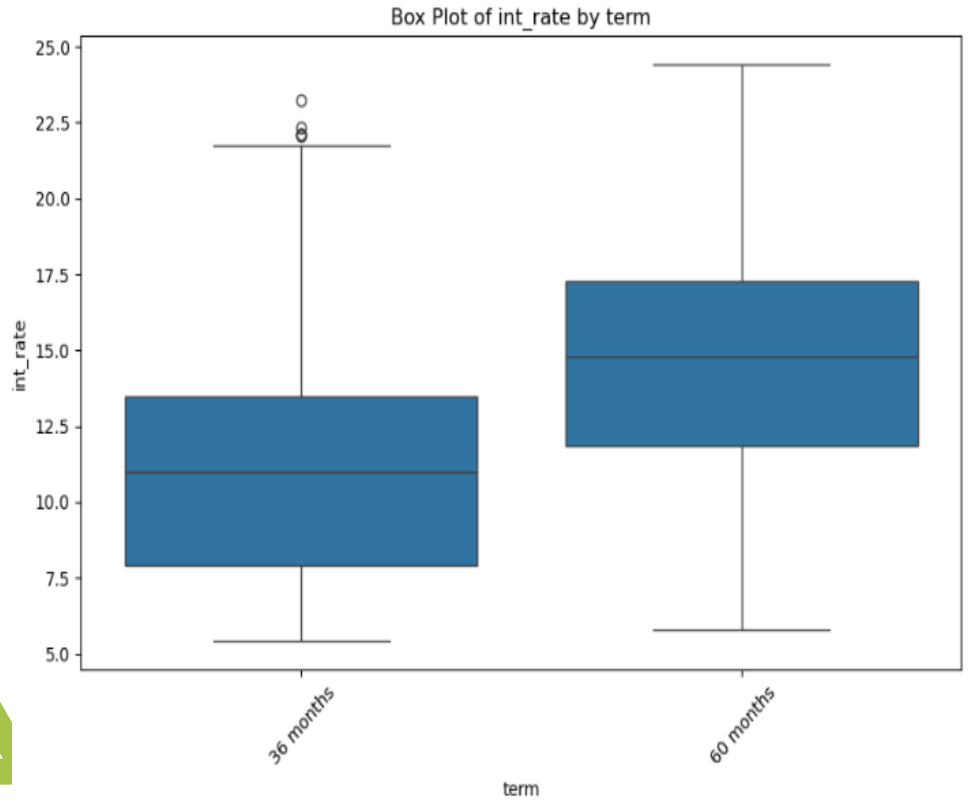


Distribution of Annual Income – After Outlier Treatment (with 95th percentile)

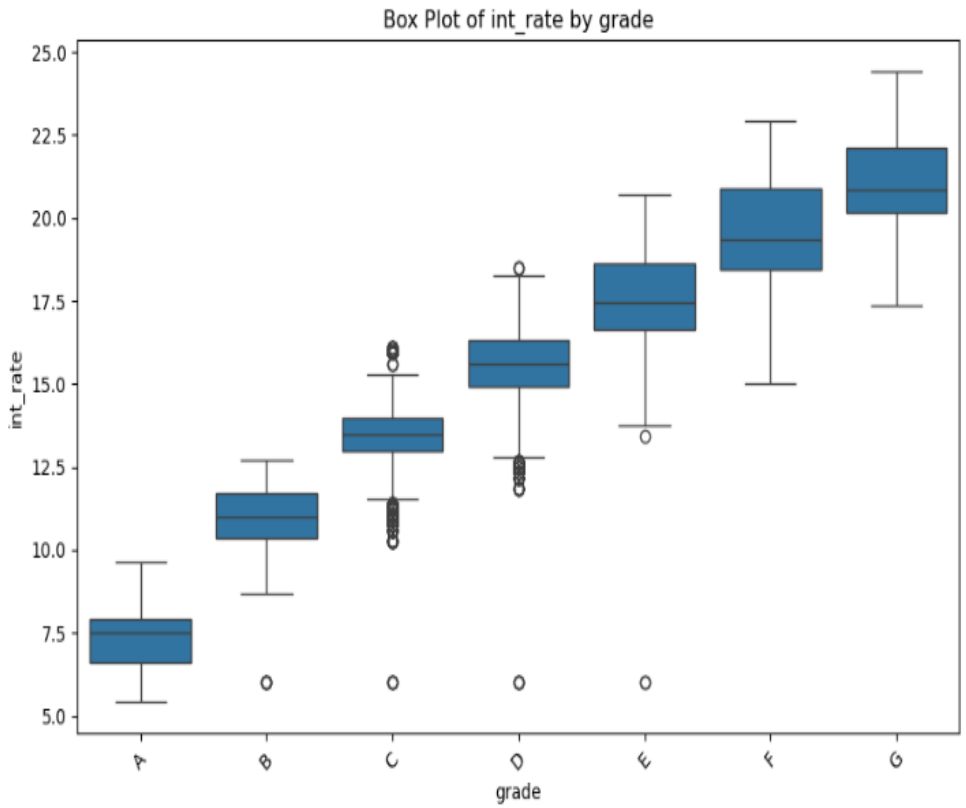


Bivariate Analysis

60 Month Term loans have significantly higher interest rate than 36 Month term

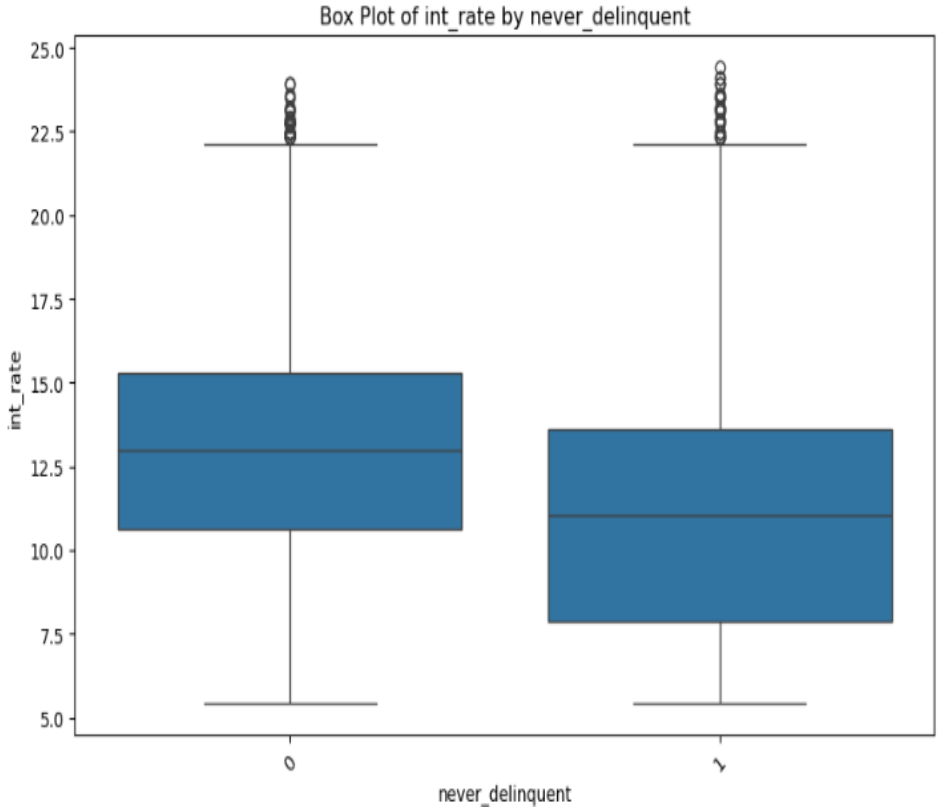


Interest Rate increased significantly for higher risky grade customers

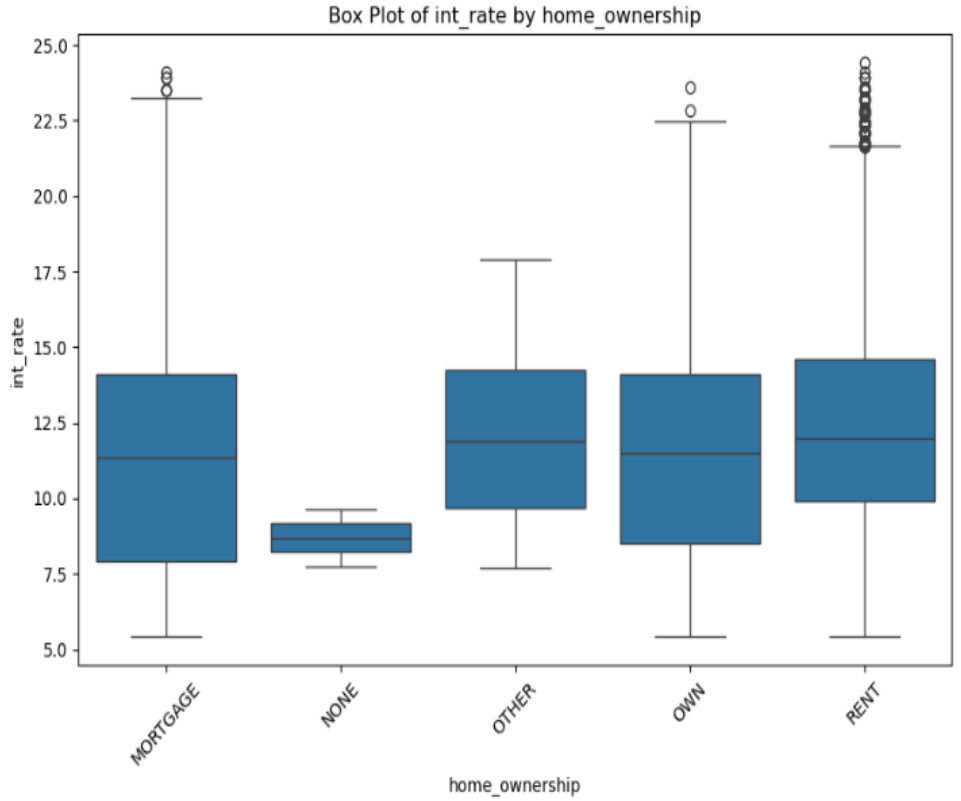


Bivariate Analysis

Those who were never delinquent had a lesser interest rate

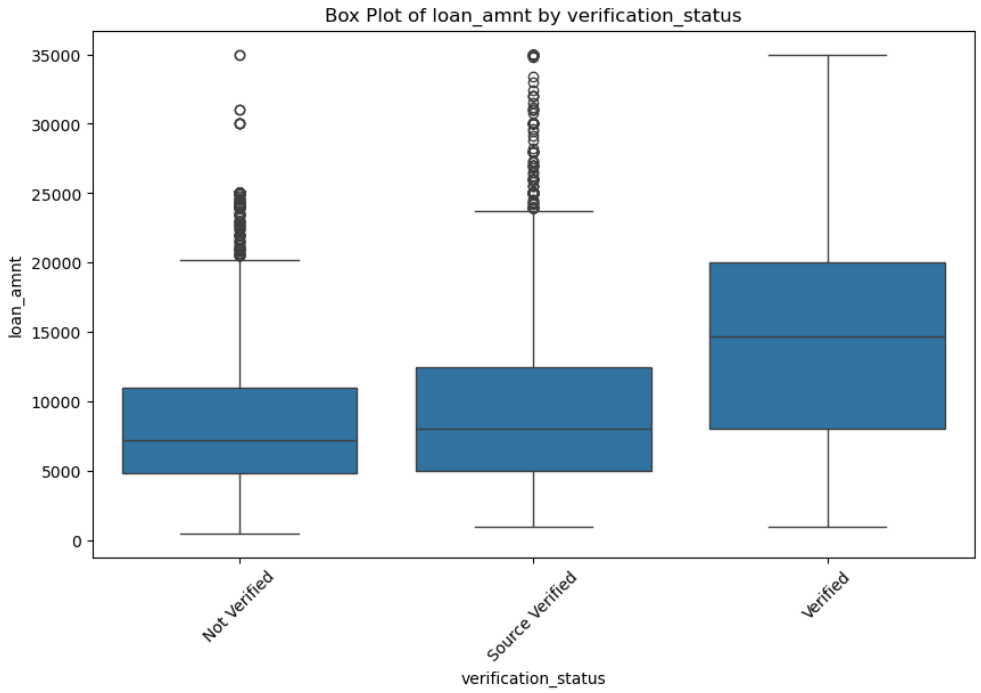


Looks like Home Ownership is not key factor behind interest rate decision

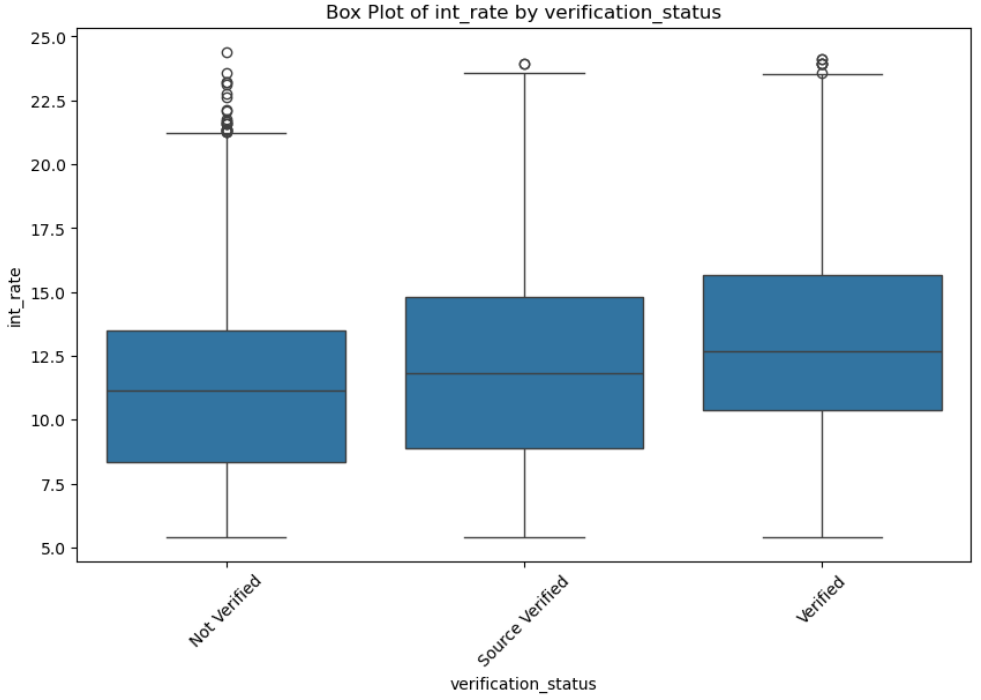


Bivariate Analysis

Looks like Verification process happens primarily for higher loan amounts

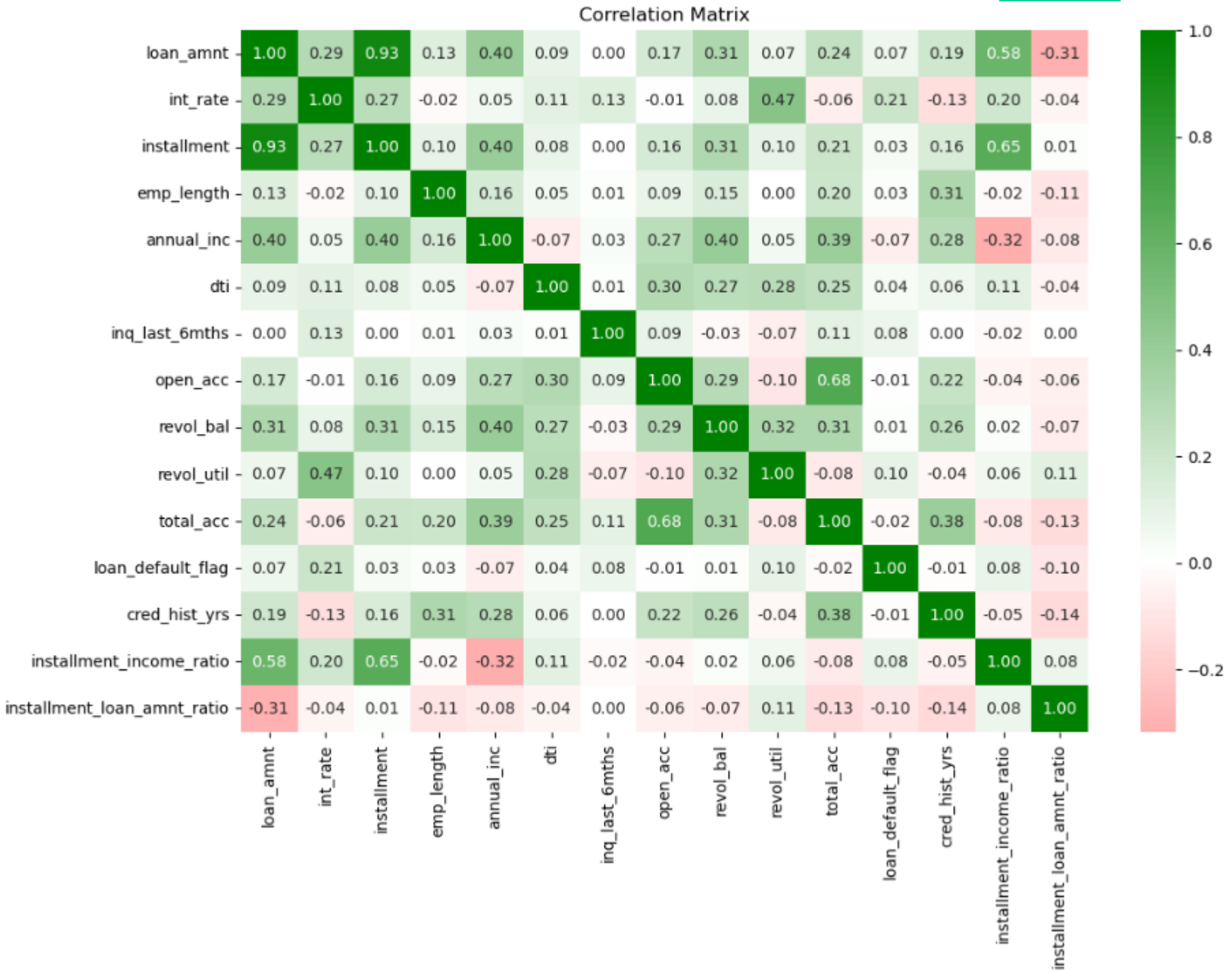


And hence, Verified users have higher interest rate than other groups



Correlation Matrix

- No strong correlation observed between major variables - Considering Industry Standard of 0.7 correlation factor
- Highest correlation for loan default is with interest rate = 0.21
- Income has some decent correlation with columns like interest rate, loan amount, revol bal (which is expected)



Observations from Univariate and Bivariate Analysis

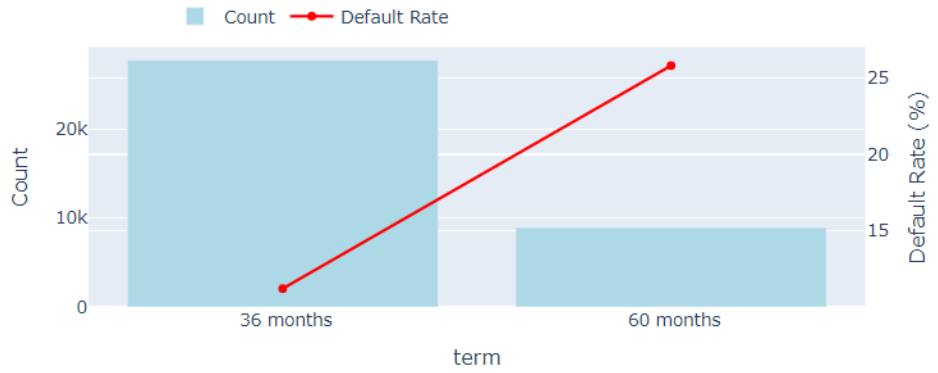
- Average Loan Amount is 10682, and max is 35000. IQR of loan amount - 5K to 15K.
- Average Interest Rate is 11.88% and max is 24.4%. IQR of interest rate - 9% to 14%.
- Average Installment is 322 and max is 1305. IQR of installment - 165 to 425.
- Average Annual Income is 68K. IQR of annual income - 40k to 82k
- Average Revolving credit utilization is 48%. IQR of credit utilization - 25% to 72%.
- 50% of them have at least 12 years of credit history. IQR of credit history - 9 years to 16 years.

- 73% loans were 36-month term loans.
- 30% loans were for grade B customers followed by 25% from grade A. Only 10% loans were grade E,F,G.
- 22% loans were given to employees with 10+ years history, and 20% were for those with less than 2 years.
- 48% loans were given to Rental home customers, followed by 44% for Mortgage homeowners.
- California and New York were the top 2 states with 18% and 10% loans issued.
- Number of loans issued are lesser during January and the count increases through the course of year and reaches maximum in December.

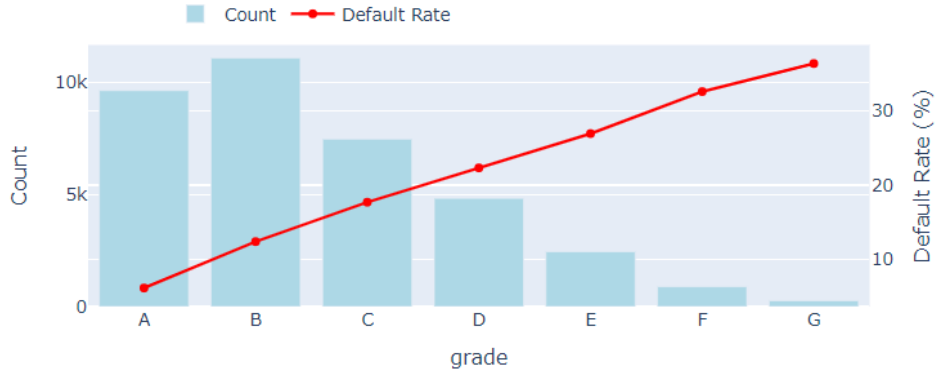
- 60 Month term had higher loan amount, higher interest rate than 36 Month term.
- Loan amount and Interest Rate increases for higher grade customers.
- Home ownership did not have significant changes for loan amount or interest rate.
- Never Delinquent customers had lower interest rate than those with delinquent history.

Key Factors driving Loan Default Rate

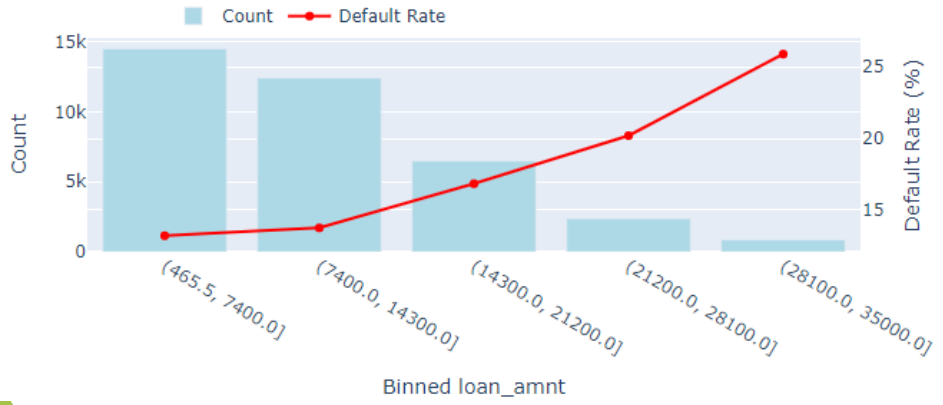
60 Month Term Loan have higher chance of default than 36 Month Term Loan



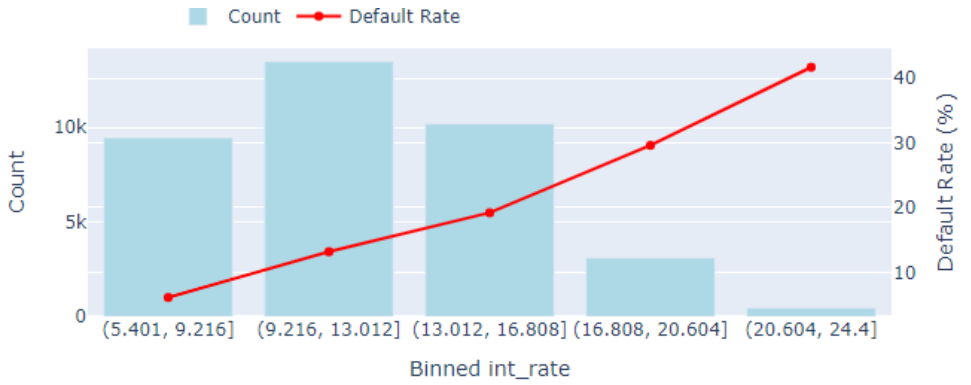
Higher the Grade, higher the chance of default



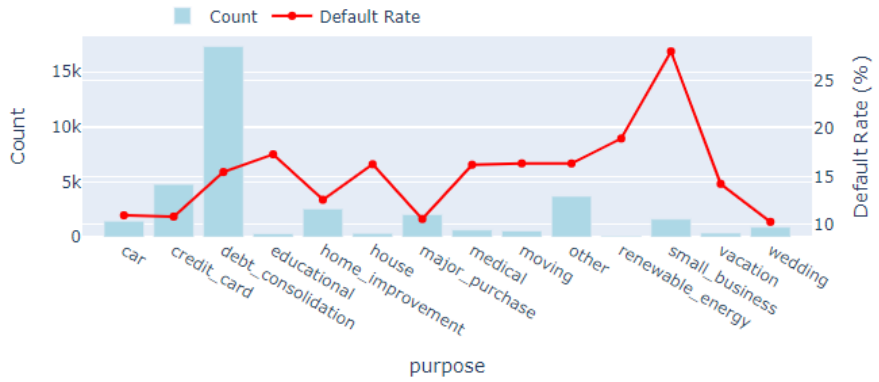
Higher the loan amount, higher the chance of default



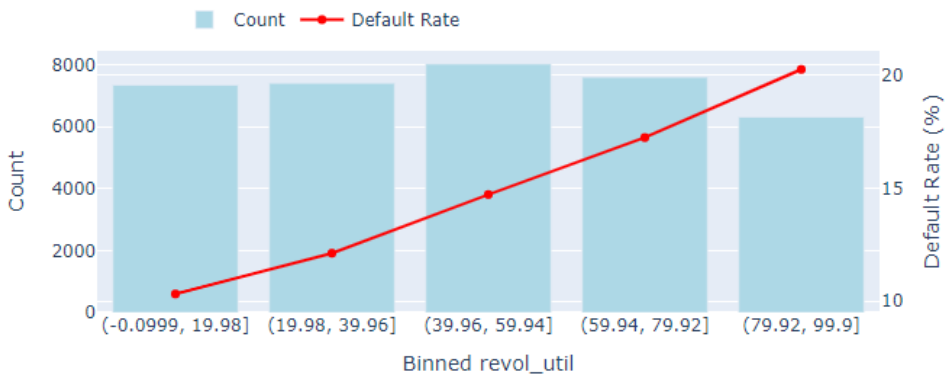
Higher the Interest Rate, higher the chance of default



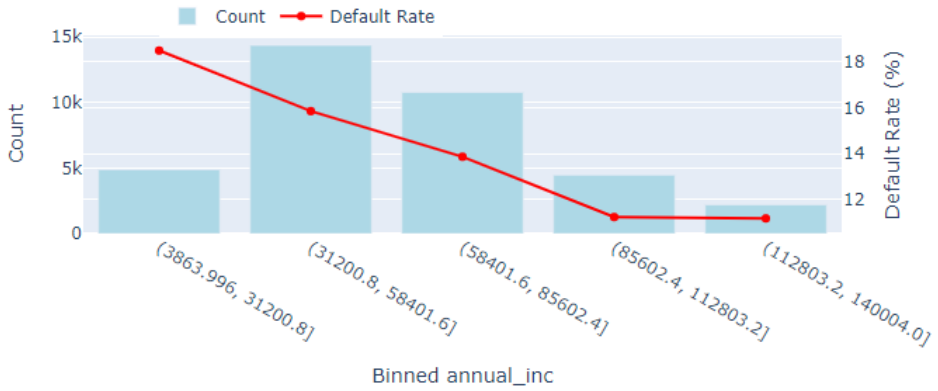
Small Business loans have significantly higher chance of default than others



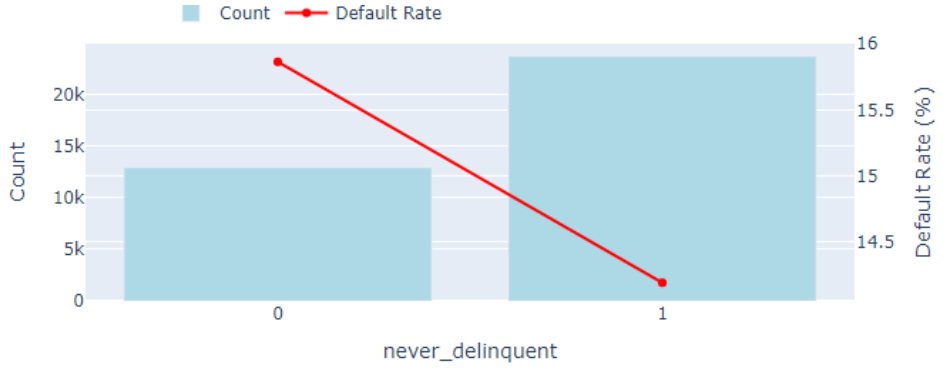
Higher the Credit Utilization, higher the chance of default



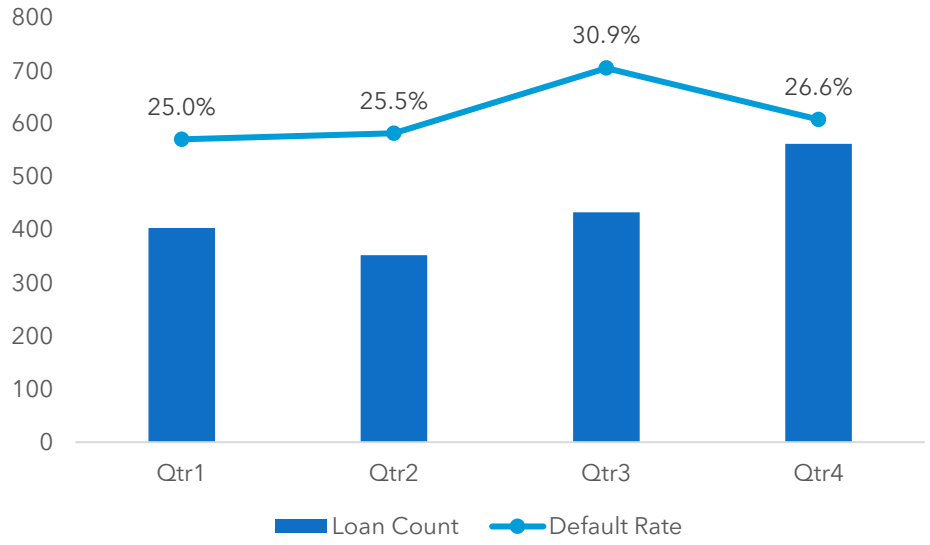
Lower the Income, higher the chance of default



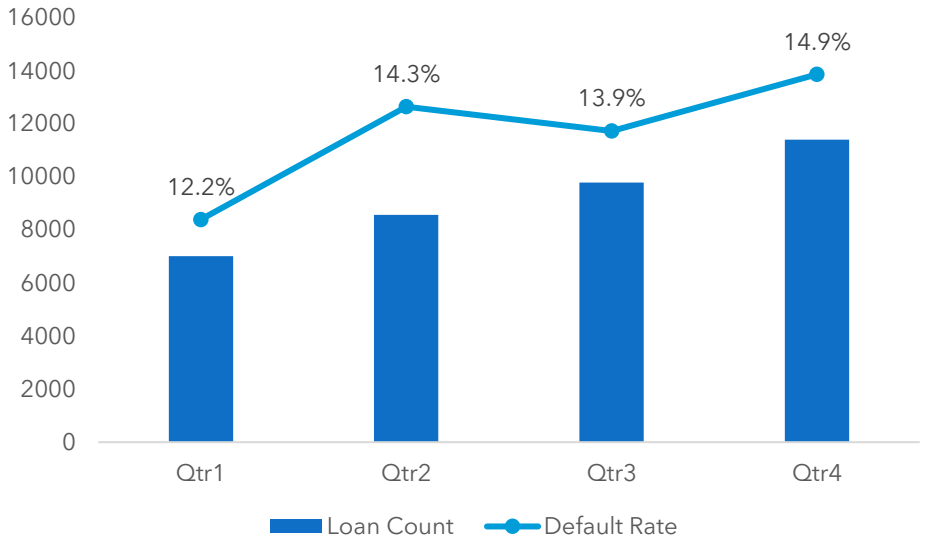
Applicants with delinquency history have higher chances of default



Historically Small Business loans issued in Q3 had higher default Rate



While Q3 had lower default Rate for all other purpose loans (non small business)



Hypothesis

Seasonal stress due to pre-holiday preparation and reduced cash flow may increase financial strain on small businesses during Q3, leading to higher default rates.

Recommendation

Enhance credit risk models by incorporating seasonality factors and industry-specific economic indicators that can predict potential cash flow challenges for small business in Q3.

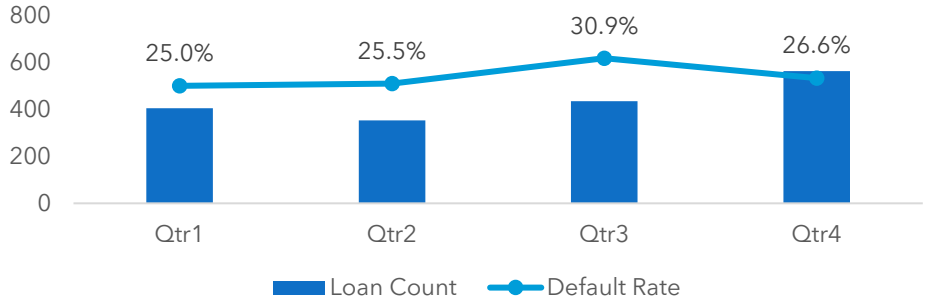
Observations on Factors impacting Default Rate

- Higher the loan amount, higher the chance of default
- Higher the interest rate, higher the chance of default
- Higher the income, lower the chance of default
- Higher the inquiries, higher the chance of default
- Higher revolving balance utilization, higher chance of default
- Higher the installment to loan amount ratio , lower the chance of default
- Higher the loan to income ratio, higher the chance of default
- Small Business have higher loan default rate.

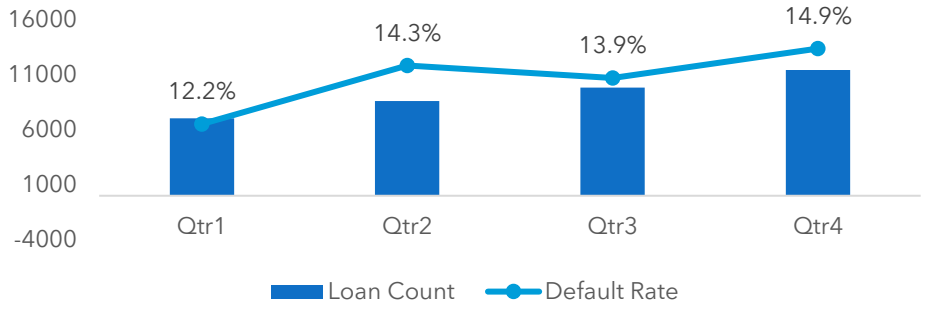
Recommendations to reduce Loan Default Rate

Focus on Small Business Loans

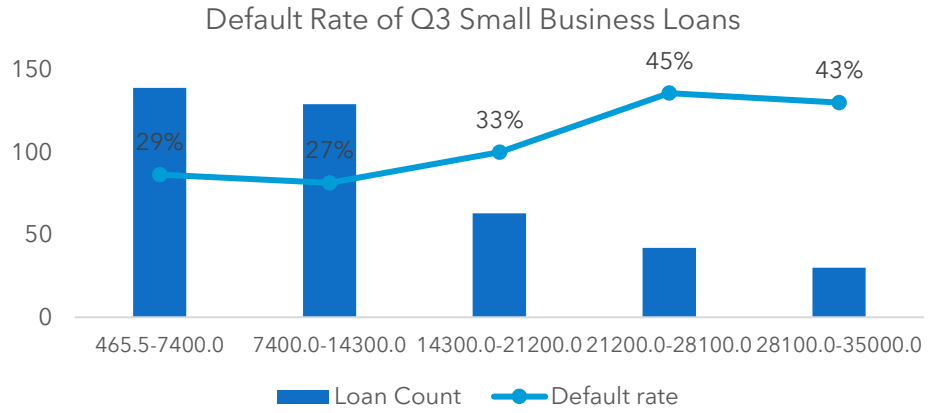
Historically Small Business loans issued in Q3 had higher default Rate

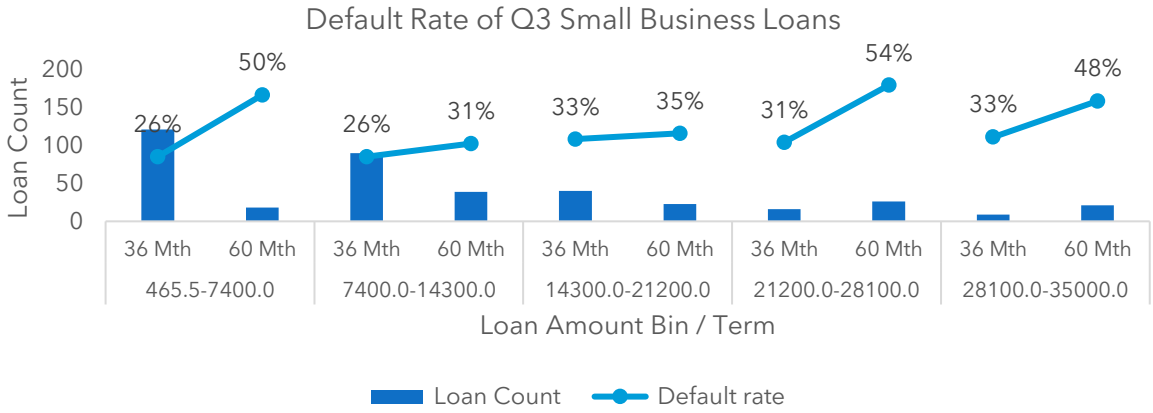


While Q3 had lower default Rate for all other purpose loans (non small business)



Among Q3 Small Business Loans, Lower loan amounts had lower default rate. And within each loan amount bin , 36 Month Term loans had lower default rate.





Recommendation

For Small business loans being issued in Q3, either approve a lower loan amount, or provide 36 Month term loan only

Generic Recommendations

- **Stricter Criteria for Larger Loans**: Limit high loan amounts to reduce default rates.
- **Promote 36-Month Loans**: Focus on shorter-term loans that are more manageable for borrowers.
- **Target Grade A and B Customers**: Tailor products to the needs of high-grade borrowers.
- **Segmentation Analysis**: Use customer segments based on their credit history, income, and loan amount to better predict default risk and tailor products accordingly.
- **Historical Credit Review**: Assess customers with delinquent histories better.
- **Small Business Loans**: Assess Small Business loans better especially those issued in Q3.