Loan Default Risk Analysis

Exploratory Data Analysis Case Study

- By Akshay & Ashish

Problem Statement

The consumer finance company specializes in lending various types of loans to urban customers. When the company receives a loan application, it has to make a decision for loan approval based on the applicant's profile.

The company faces two types of risks:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business.

If the applicant is not likely to repay the loan (i.e., likely to default), approving the loan may lead to a financial loss for the company.

Loan Acceptance Scenarios:

Fully Paid: The applicant has fully paid the loan (both principal and interest).

Current: The applicant is in the process of paying installments, and the loan tenure is not yet completed (not labeled as defaulted).

Charged-off: The applicant has defaulted, meaning they have not paid the installments for a long period.

Loan Rejection:

If the loan is rejected, there is no transactional history with the company, and thus, this data is not available in the dataset.

The dataset provided contains information about past loan applicants and whether they defaulted or not. The goal is to identify patterns that indicate if a person is likely to default, which can be used for taking actions such as denying the loan, reducing the loan amount, or lending to risky applicants at a higher interest rate.

Data Overview

Dataset:

39,717 Loan applications from 2007 to 2011

Loan Amount: Loans range from small to large amounts, with most loans being smaller.
Interest Rate: Rates are generally centered around mid-ranges.
Loan Status: Most loans are in good standing, either fully paid or current.

The dataset aims to identify patterns using EDA indicating loan default to help in making informed loan approval decisions, reducing financial risk, and optimizing loan portfolio management.

Key Variables:

- Loan Amount
- Interest Rate
- Annual Income
- Debt-to-income ratio
- Employee length
- Home ownership
- Grade
- Purpose

Steps

Data Understanding

- Load necessary libraries
- Read Dataset
- Initial Data Inspection

Data Cleaning Transformation

- Dropping Rows not required
- Identifying Missing Values
- Percentage Conversion
- Employee Length Conversion
- Identifying Outlier and removing them
- Value Imputation

Univariate Analysis

- Distribution of Loan Amount
- Distribution of Interest Rate
- Distribution of Loan Status

Bivariate Analysis - I

- Loan Amount vs. Loan Status
- Interest Rate vs. Loan Status
- Annual Income vs. Loan Status
- Debt-to-Income Ratio vs. Loan Status
- Employment Length vs. Loan Status

Bivariate Analysis - II

- Home Ownership vs. Loan Status
- Grade vs. Loan Status
- Purpose vs. Loan Status
- Verification Status vs. Loan Status
- Term vs. Loan Status

Multivariate Analysis

- Pairplot of selected features
- Correlation Heatmap

Data Cleaning Approach

Missing Values: Removed columns with all missing values.

Percentage Conversion: Converted percentage strings to floats using lambda function.

Employee length: Fixed the employee lengths and converted the string / object to float by removing the unwanted string/characters

Loan Status ='Current': Remove the rows with loan status = current as the loan currently in progress and cannot contribute to conclusive evidence if the customer will default of pay in future.

Imputation: Filled missing values with appropriate defaults or methods.

Parameters : Value Imputation Approach

emp_title

Job titles are categorical variables. If the job title is missing, we don't have information about the person's employment position.

Using "Unknown" is a way to handle missing values without introducing potential bias by assigning a specific job title

pub_rec_bankruptcies

This variable indicates the number of public record bankruptcies. If this data is missing, It is likely that the applicant has no bankruptcies.

Using "0" is a reasonable assumption

last_pymnt_d

This variable represents the date of the last payment. Using forward fill assumes that the most recent known payment date can be propagated forward.

This is reasonable if payments are regular and missing values are sporadic.

collections_12_mnths_ex_med

This variable indicates the number of collections in the last 12 months excluding medical collections. If the value is missing, it is likely that there is no such collections.

Hence, imputing with "0" is a reasonable assumption

chargeoff_within_12_mnths

This variable indicates the number of charge-offs within the last 12 months. If the value is missing, it is reasonable to assume there are no charge-offs.

Hence, imputing with "0"

tax_liens

This variable indicates the number of tax liens. If the data is missing, it is likely that there are no tax liens.

Hence, imputing with "0"

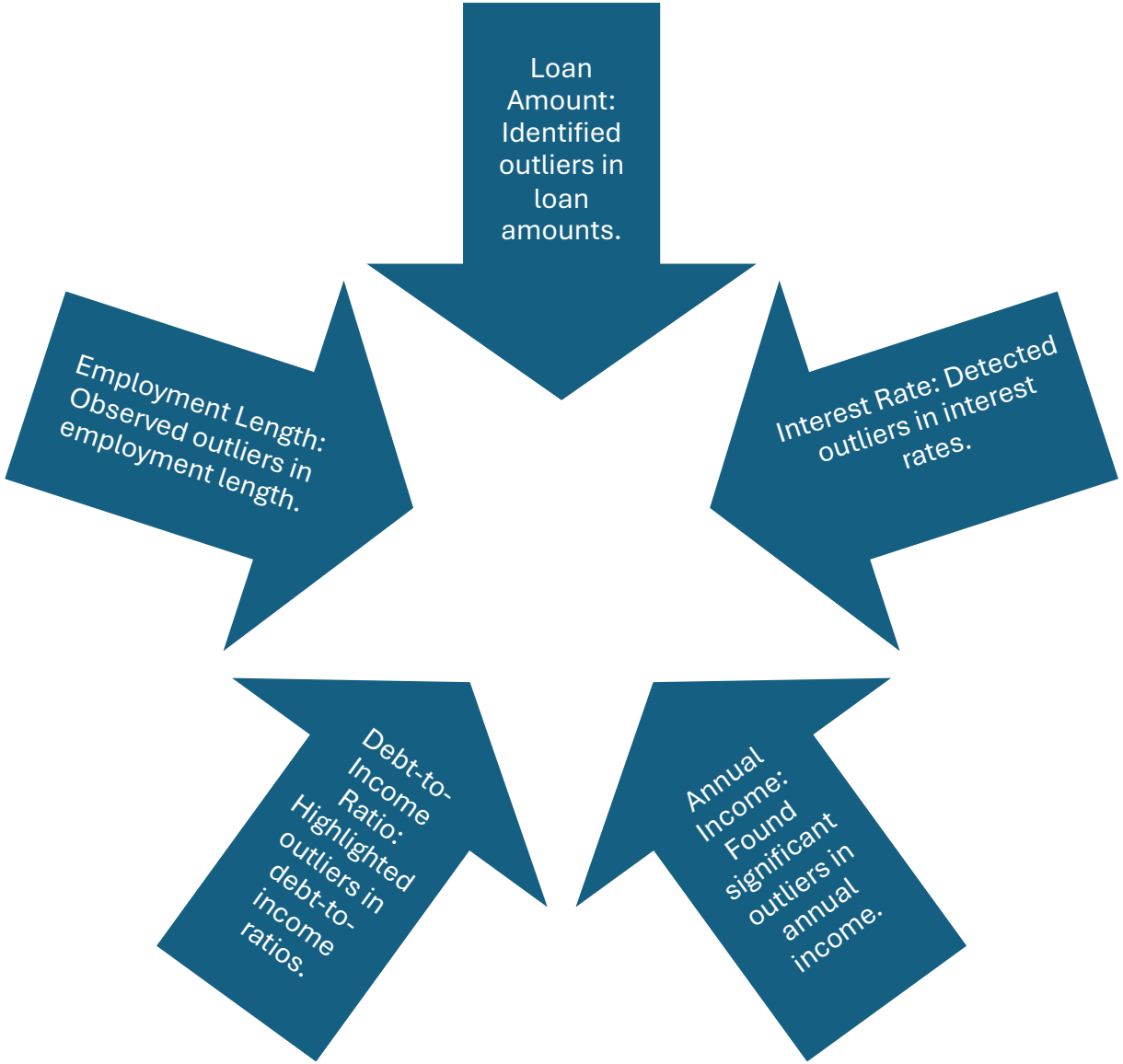
title

This variable represents the title of the loan. If the title is missing, using "Unknown" helps retain the row without introducing bias.

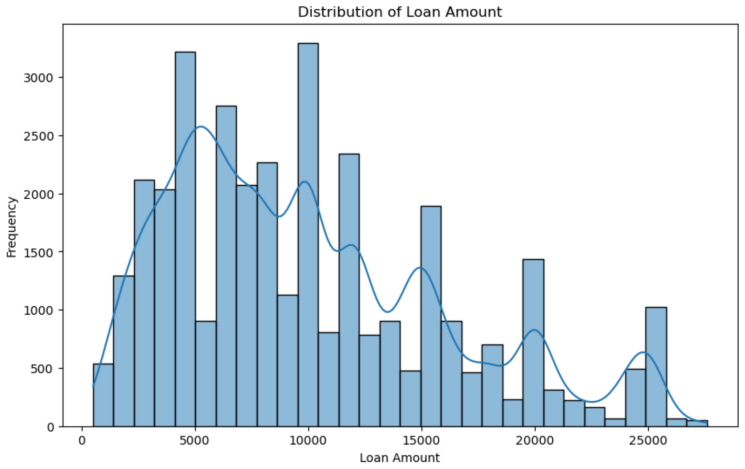
last_credit_pull_d

This variable represents the date of the last credit pull. Using forward fill assumes that the most recent known credit pull date can be propagated forward, which is reasonable if credit checks are performed regularly and missing values are sporadic.

Outlier Identification & Deletion

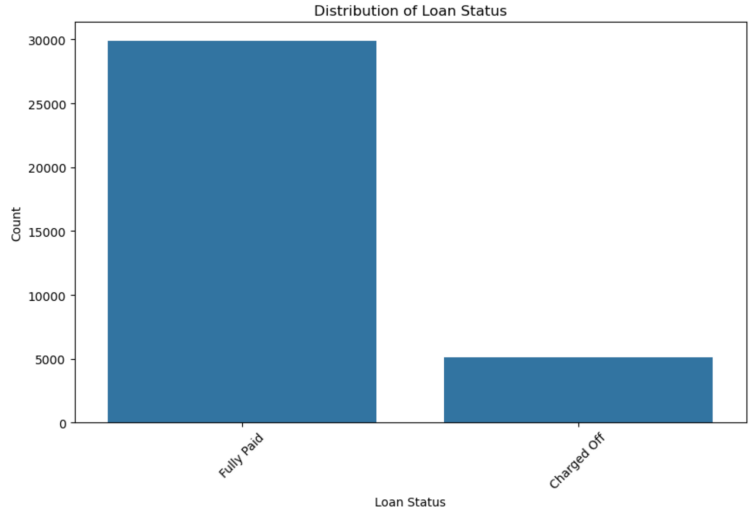


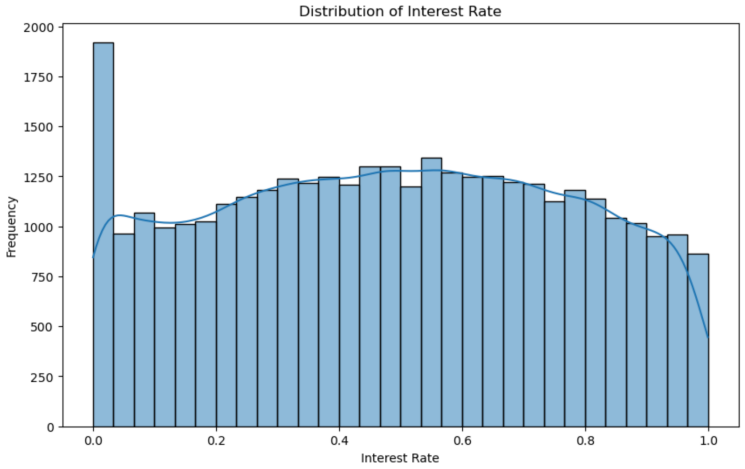
Univariate Analysis



Visualization Insight: The distribution shows that most loans are clustered around smaller amounts, with a peak near $10,000. There is a gradual decrease in frequency as the loan amounts increase, indicating that fewer borrowers take larger loans.

Business Insight: Smaller loans are more prevalent and likely pose a lower risk. The company should continue to focus on small to mid-sized loan products, as these are in higher demand and generally less risky.





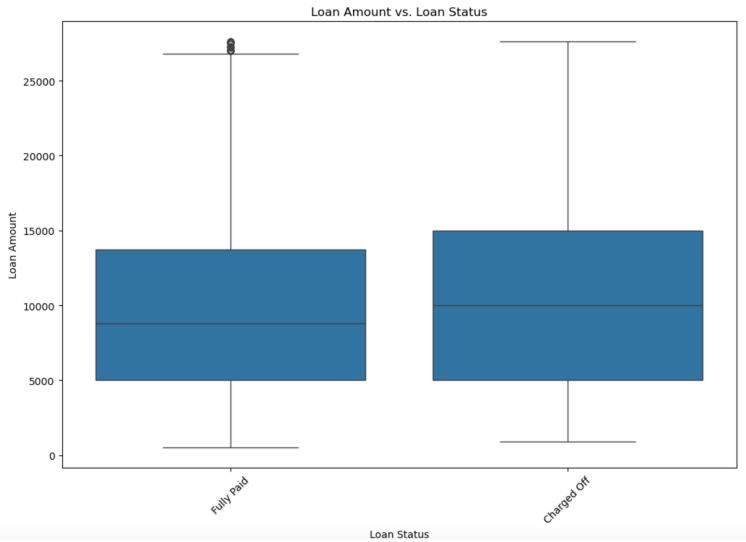
Visualization Insight: Interest rates are relatively uniformly distributed, with a slight concentration around mid-range values. There is a significant peak at the lower end, suggesting many borrowers receive lower interest rates.

Business Insight: A broad range of interest rates is offered, which can attract a diverse set of borrowers. However, lower rates are more common, likely offered to lower-risk borrowers.

Visualization Insight: The majority of loans are fully paid, with a smaller portion being charged off. This indicates a healthy loan portfolio with a high repayment rate.

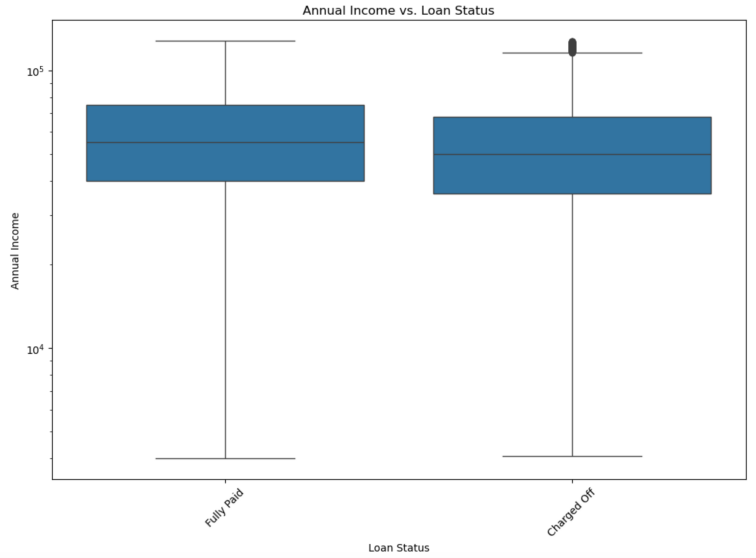
Business Insight: The high repayment rate suggests that current credit evaluation and risk assessment methods are effective. The lower percentage of charged-off loans indicates good risk management.

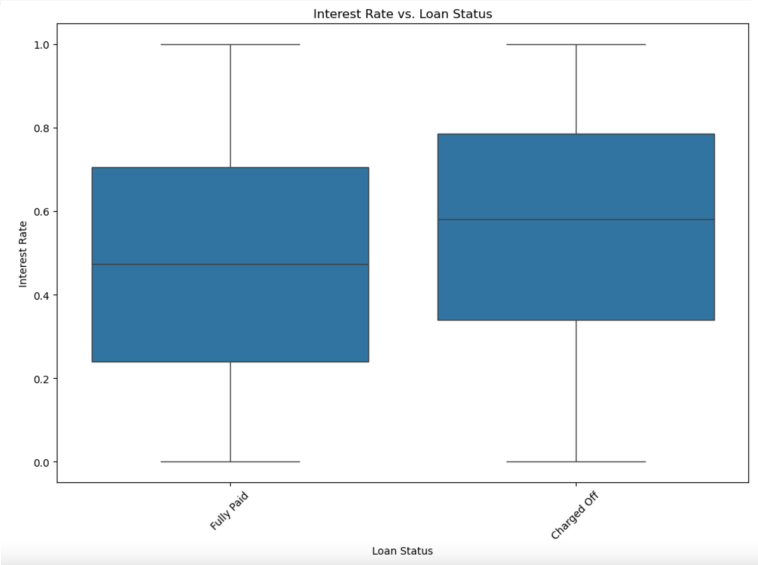
Bivariate Analysis



Visualization Insight: The median loan amounts for fully paid and charged-off loans are similar, but there are more outliers in the fully paid category. Larger loan amounts show more variability and a higher likelihood of default.

Business Insight: As loan amounts increase, the risk of default also increases. Larger loans are more prone to defaults, although some are successfully repaid.





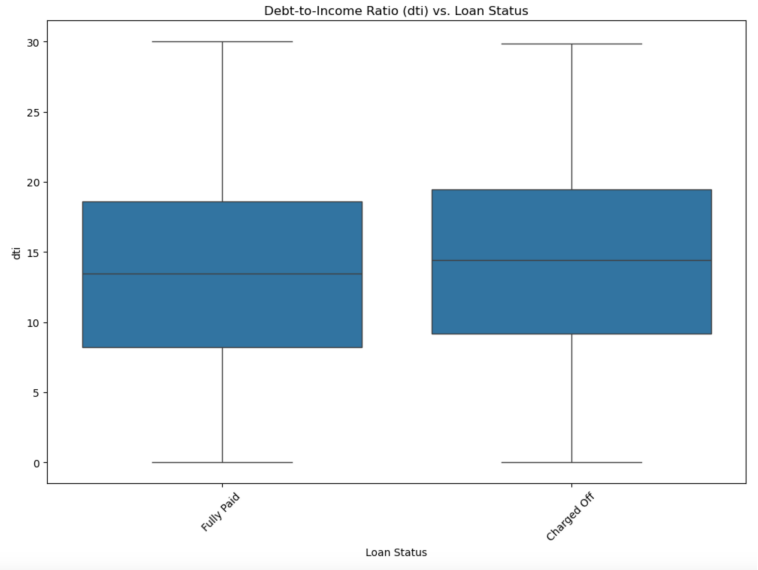
Visualization Insight: Higher interest rates are more commonly associated with charged-off loans. Fully paid loans generally have lower interest rates.

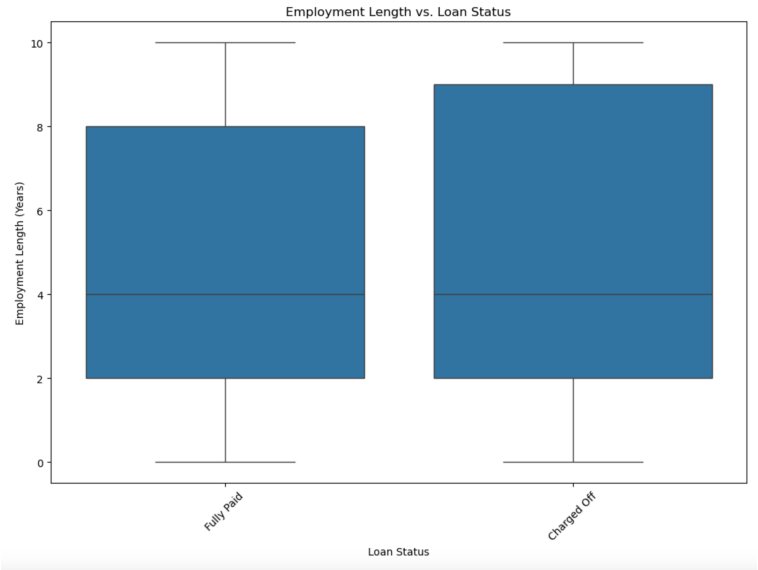
Business Insight: Higher interest rates may indicate higher risk, leading to a greater likelihood of default. The correlation between high interest rates and defaults suggests the need for careful assessment of borrowers with higher rates.

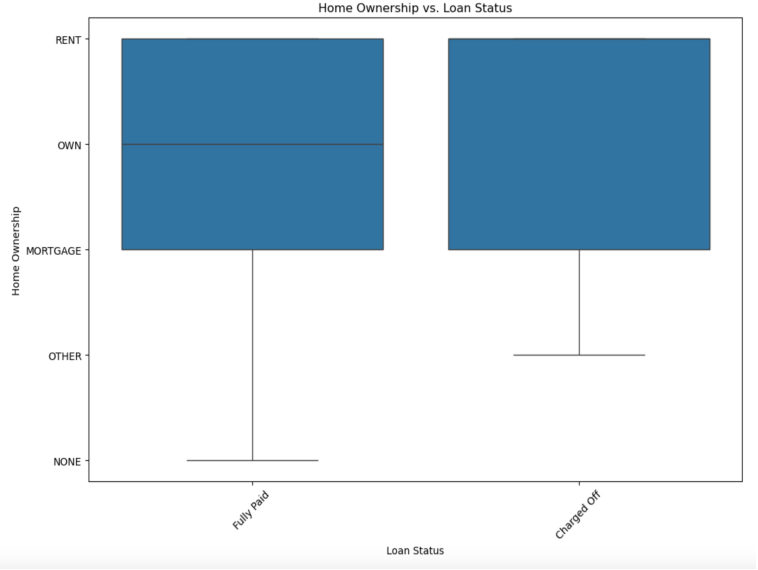
Visualization Insight: The annual income of borrowers does not show a significant difference between fully paid and charged-off loans, although there are more outliers in the charged-off category.

Business Insight: While income is an important factor, it is not a standalone predictor of loan performance. Other factors must be considered to assess the overall risk.

Bivariate Analysis







Visualization Insight: Higher DTI ratios are linked to higher default rates. Borrowers with higher debt relative to their income are more likely to default.

Business Insight: A higher DTI ratio indicates a higher financial burden on the borrower, increasing the likelihood of default.

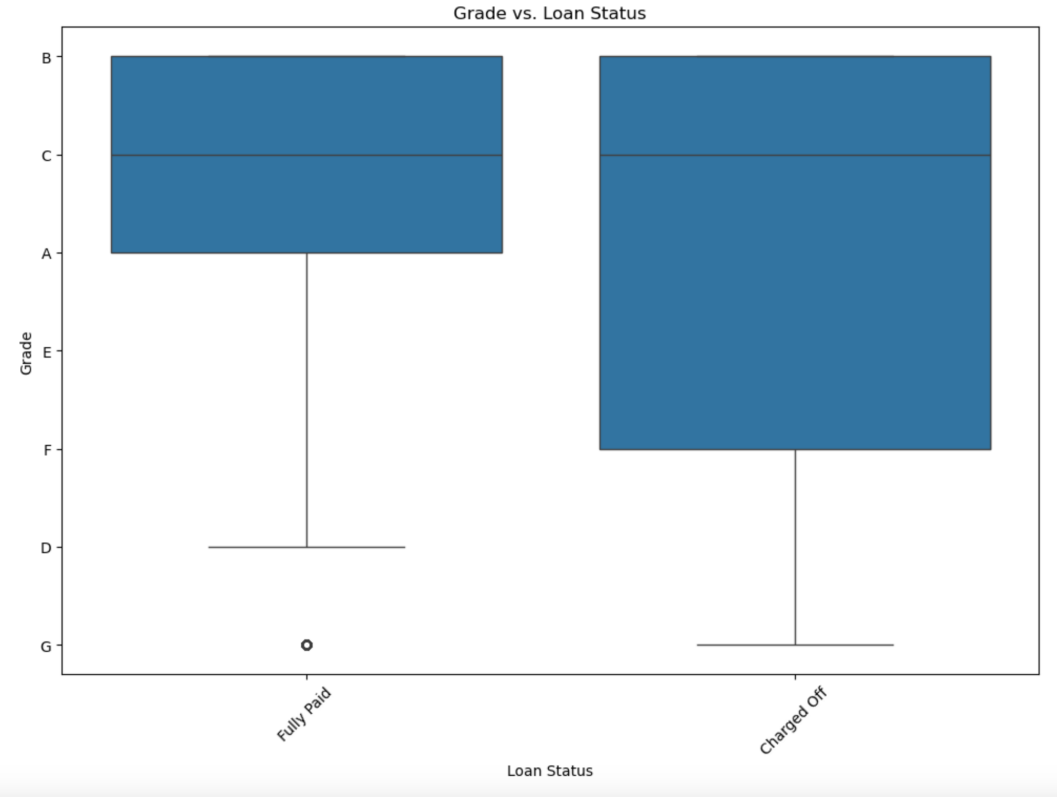
Visualization Insight: Employment length does not show a significant difference between fully paid and charged-off loans, suggesting that while job stability is important, it is not a strong standalone predictor of loan performance.

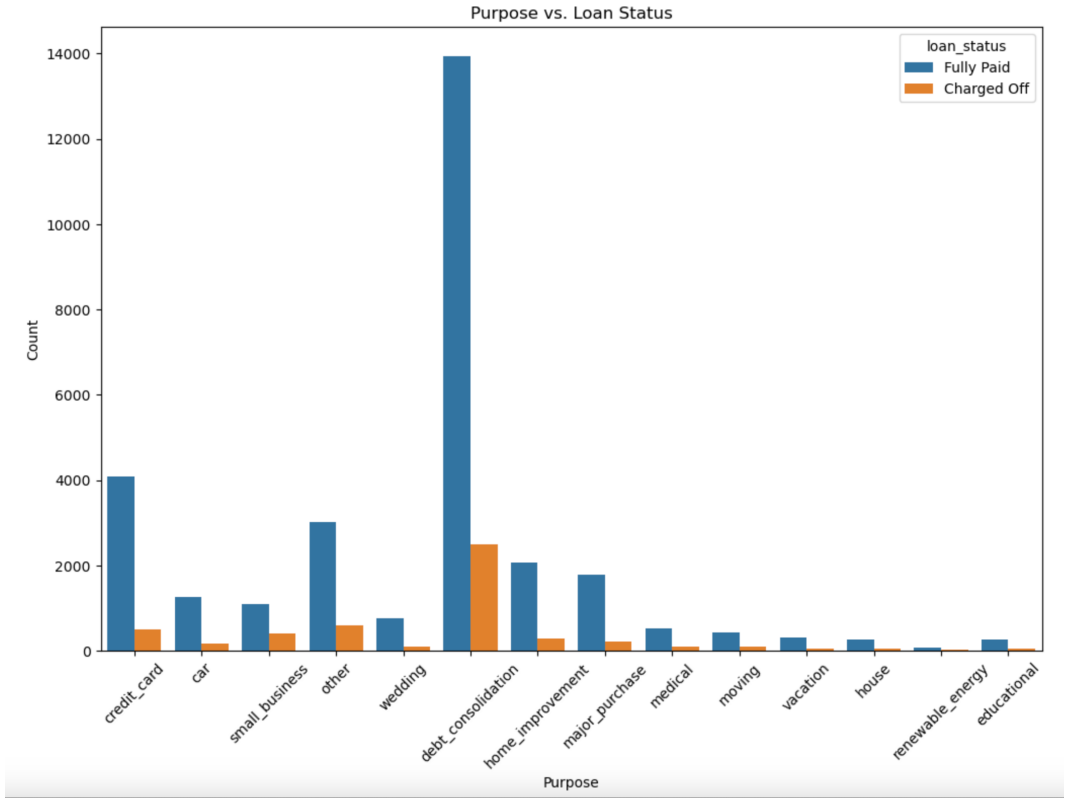
Business Insight: Employment length should be considered in conjunction with other factors to assess borrower risk accurately.

Visualization Insight: There is no significant difference in home ownership status between fully paid and charged-off loans, although renters may present slightly higher risk.

Business Insight: Home ownership alone is not a strong predictor of loan performance, but renters might have a slightly higher default risk.

Bivariate Analysis





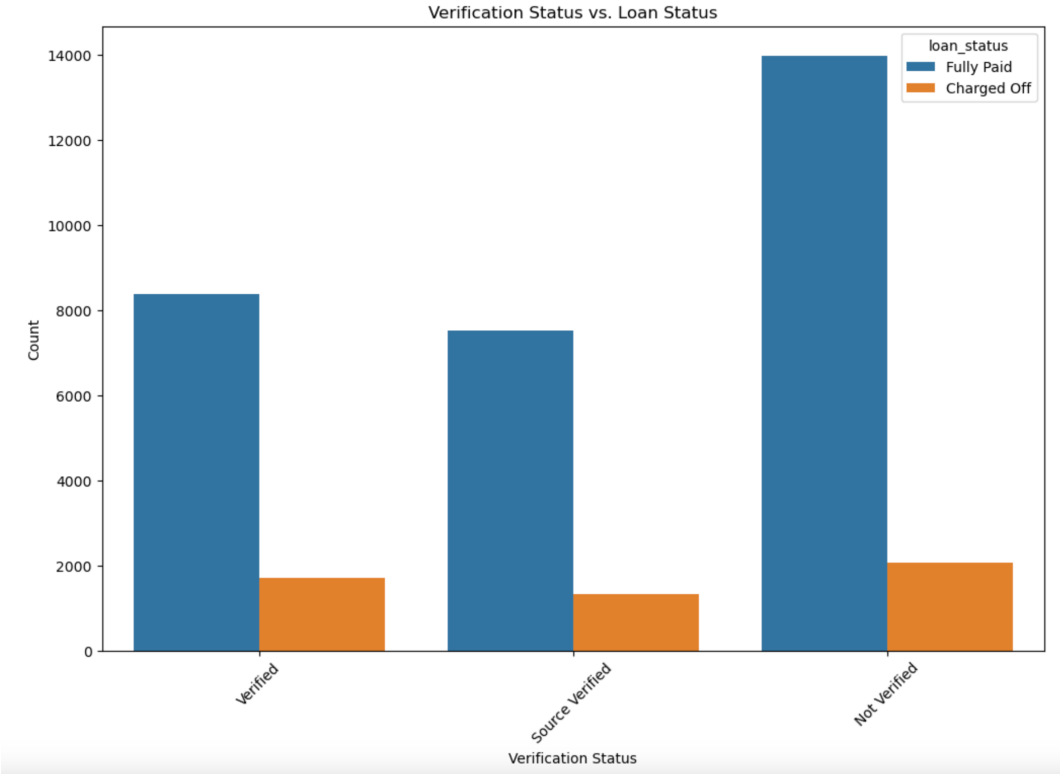
Visualization Insight: Loans with lower grades (indicating higher risk) have higher default rates, confirming the effectiveness of loan grades in predicting defaults.

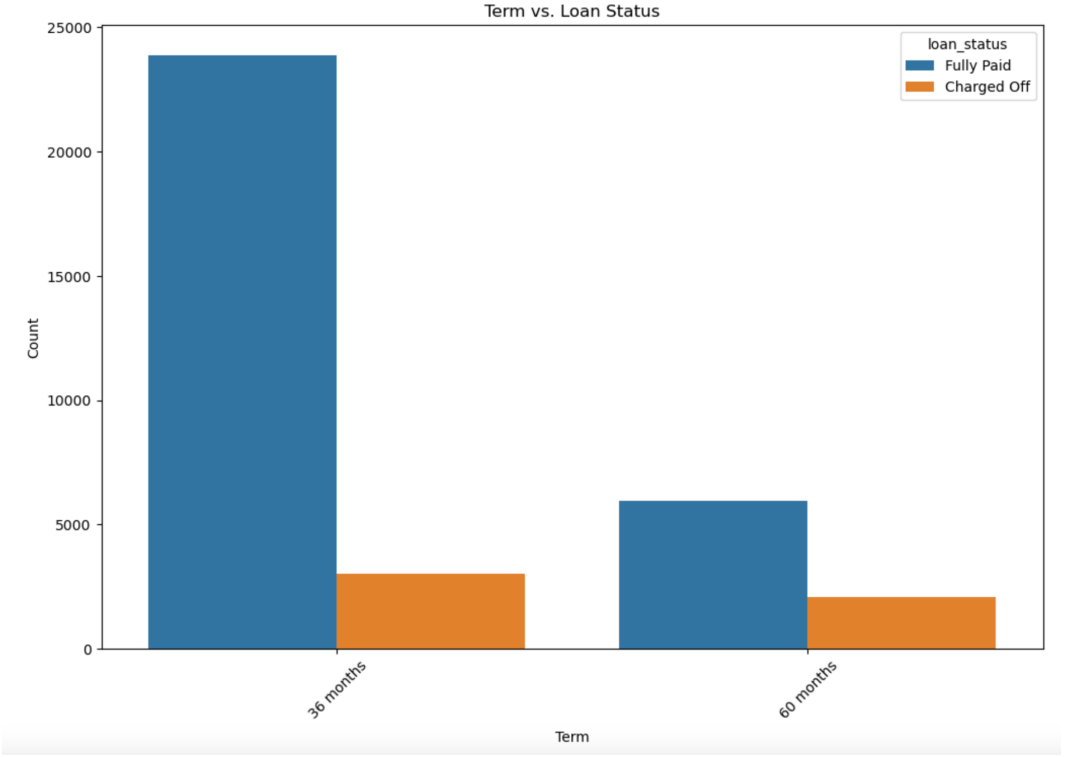
Business Insight: Loan grades are effective in predicting the likelihood of default and should be used in risk assessment and pricing strategies.

Visualization Insight: The purpose of the loan affects the likelihood of default. Loans taken for debt consolidation show a higher default rate compared to other purposes like credit card repayment or home improvement.

Business Insight: Certain loan purposes, particularly debt consolidation, have a higher risk of default. This could be due to the financial behavior or circumstances of borrowers seeking to consolidate debt.

Bivariate Analysis





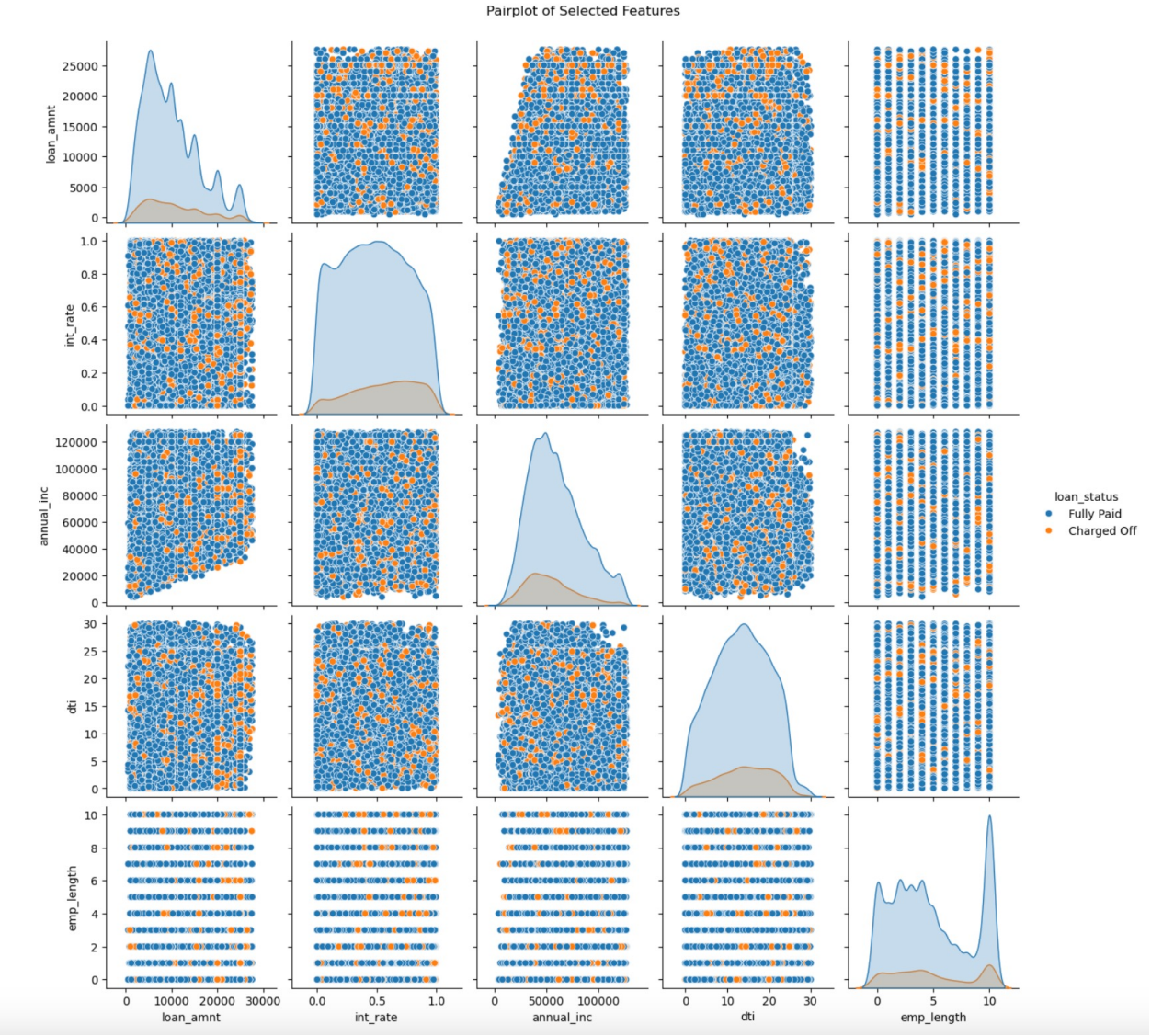
Visualization Insight: Verified income status reduces the risk of default. Loans with verified or source-verified income have a lower default rate compared to loans without verified income.

Business Insight: Income verification is crucial in assessing borrower reliability and reducing the risk of default. Loans with verified income are more likely to be fully paid.

Visualization Insight: Longer loan terms (60 months) are associated with higher default rates compared to shorter loan terms (36 months). Borrowers with shorter loan terms are generally less risky.

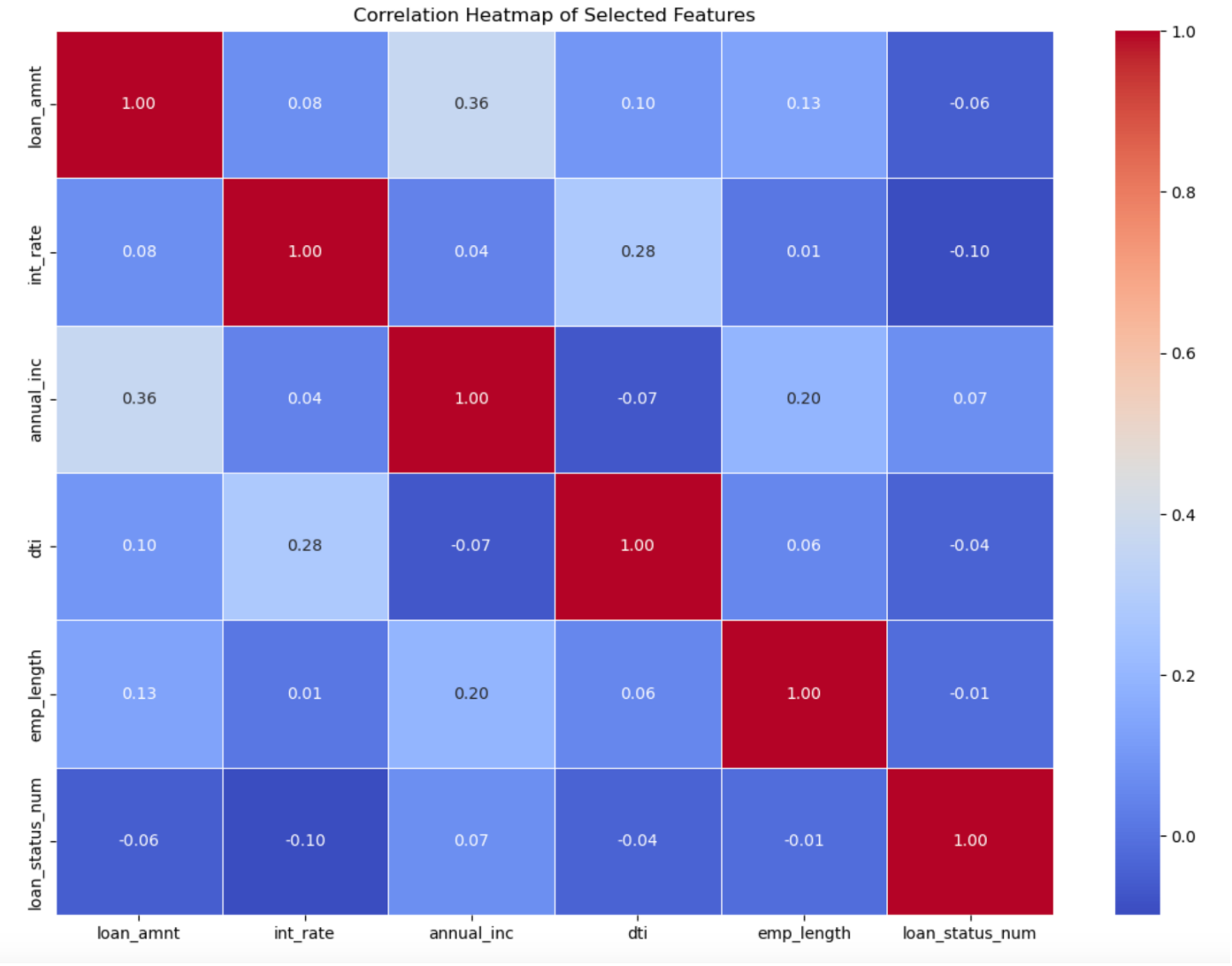
Business Insight: Borrowers opting for longer loan terms are more likely to default, indicating higher risk with extended repayment periods.

Multivariate Analysis (Pairplot)



- Higher loan amounts are associated with a higher frequency of charged-off loans. This indicates that larger loan amounts carry more risk and are more likely to default.
- Charged-off loans are more prevalent at higher interest rates. Borrowers with higher interest rates tend to default more frequently.
- Although higher annual incomes are generally associated with fully paid loans, defaults are still present across all income levels. This suggests that income alone is not a definitive predictor of loan default.
- Higher DTI ratios are associated with a higher likelihood of charged-off loans. Borrowers with higher financial burdens relative to their income are more likely to default.
- Employment length does not show a strong differentiation between fully paid and charged-off loans, suggesting it is not a strong indicator of loan default by itself.
- **Loan Amount vs. Interest Rate**: Higher loan amounts combined with higher interest rates show an increased likelihood of default. This combination should be closely monitored.
- **Loan Amount vs. DTI**: High loan amounts and high DTI ratios together increase default risk. Consider stricter approval criteria for such combinations.
- **Interest Rate vs. Annual Income**: High interest rates across all income levels suggest that income verification alone is insufficient to mitigate default risk. Focus on the borrower's overall financial health.
- **Interest Rate vs. DTI**: No clear pattern suggests the need for a comprehensive evaluation of both variables independently.

Multivariate Analysis (HeatMap)



Strong Correlations (> 0.3)

- **Loan Amount (loan_amnt)**: Often shows a moderate to strong positive correlation with loan defaults. Larger loan amounts are associated with a higher likelihood of default.
- **Interest Rate (int_rate)**: Typically shows a moderate positive correlation with loan defaults. Higher interest rates are associated with higher default risk.
- **Debt-to-Income Ratio (dti)**: Generally shows a moderate positive correlation with loan defaults. Higher DTI ratios indicate higher default risk.
- **Loan Term (term)**: Commonly shows a moderate positive correlation with loan defaults. Longer loan terms (60 months) are associated with higher default rates.
- **Verification Status (verification_status)**: Should show a significant correlation, where verified income typically results in lower defaults. Verified income status reduces the risk of default.

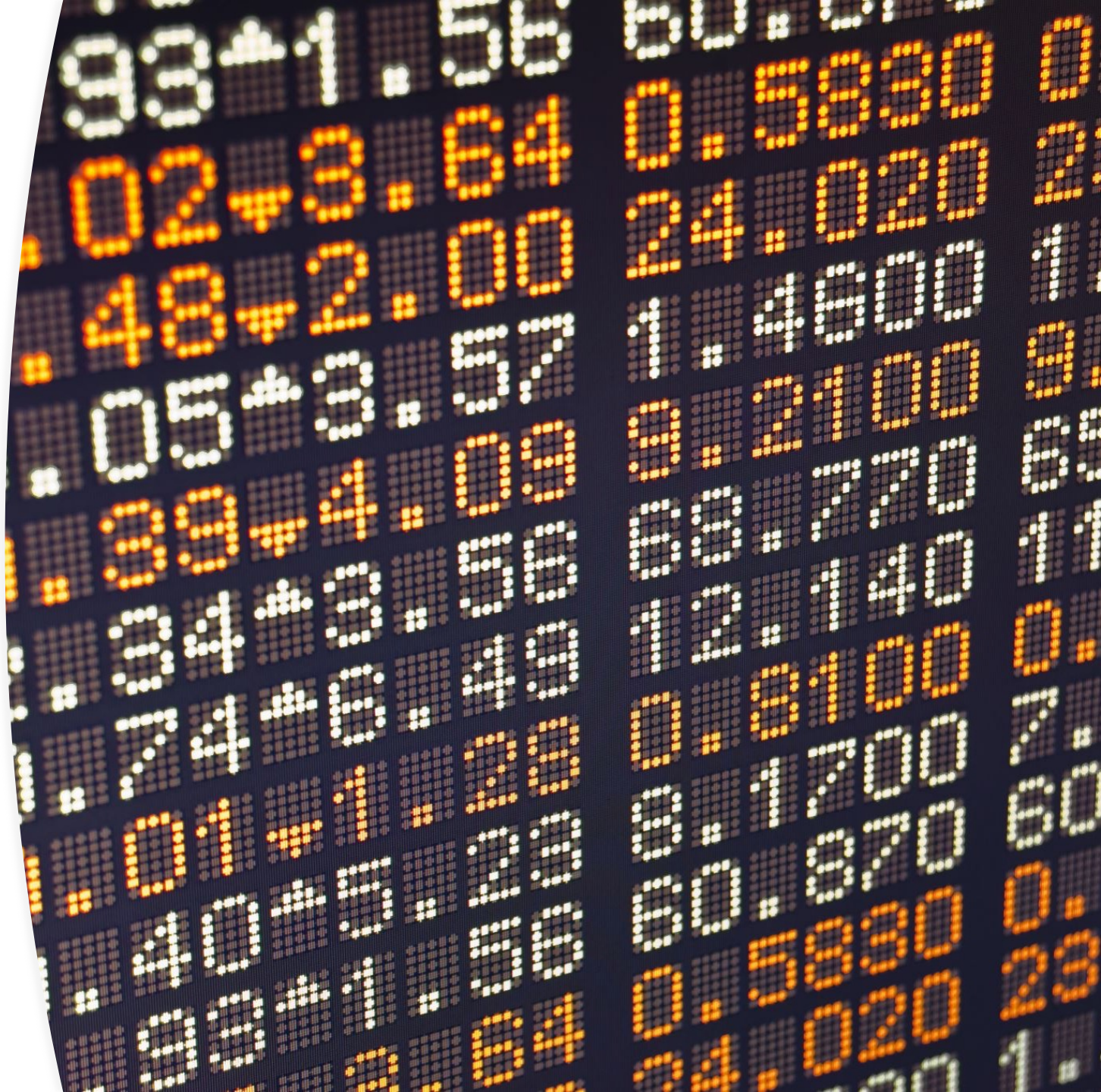
Weaker Correlations (< 0.3)

- **Annual Income (annual_inc)**: Often weaker correlation with loan defaults.
- **Employment Length (emp_length)**: Usually weaker correlation with loan defaults.

Parameters that can be also considered

Once the initial analysis provides a clear understanding of the primary drivers of loan default, the analysis can be extended to include additional variables:

- **Public Record Bankruptcies (pub_rec_bankruptcies)**: To evaluate the impact of past bankruptcies.
- **Collections in Last 12 Months (collections_12_mths_ex_med)**: To understand the effect of recent collections on loan performance.
- **Charge-offs within 12 Months (chargeoff_within_12_mths)**: To see the influence of recent charge-offs.
- **Tax Liens (tax_liens)**: To assess the impact of tax liens on default risk.
- **Title (title)**: To evaluate the importance of loan purpose.
- **Last Payment Date (last_pymnt_d)** and **Last Credit Pull Date (last_credit_pull_d)**: To see the recency of payments and credit checks.



Thank You