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"

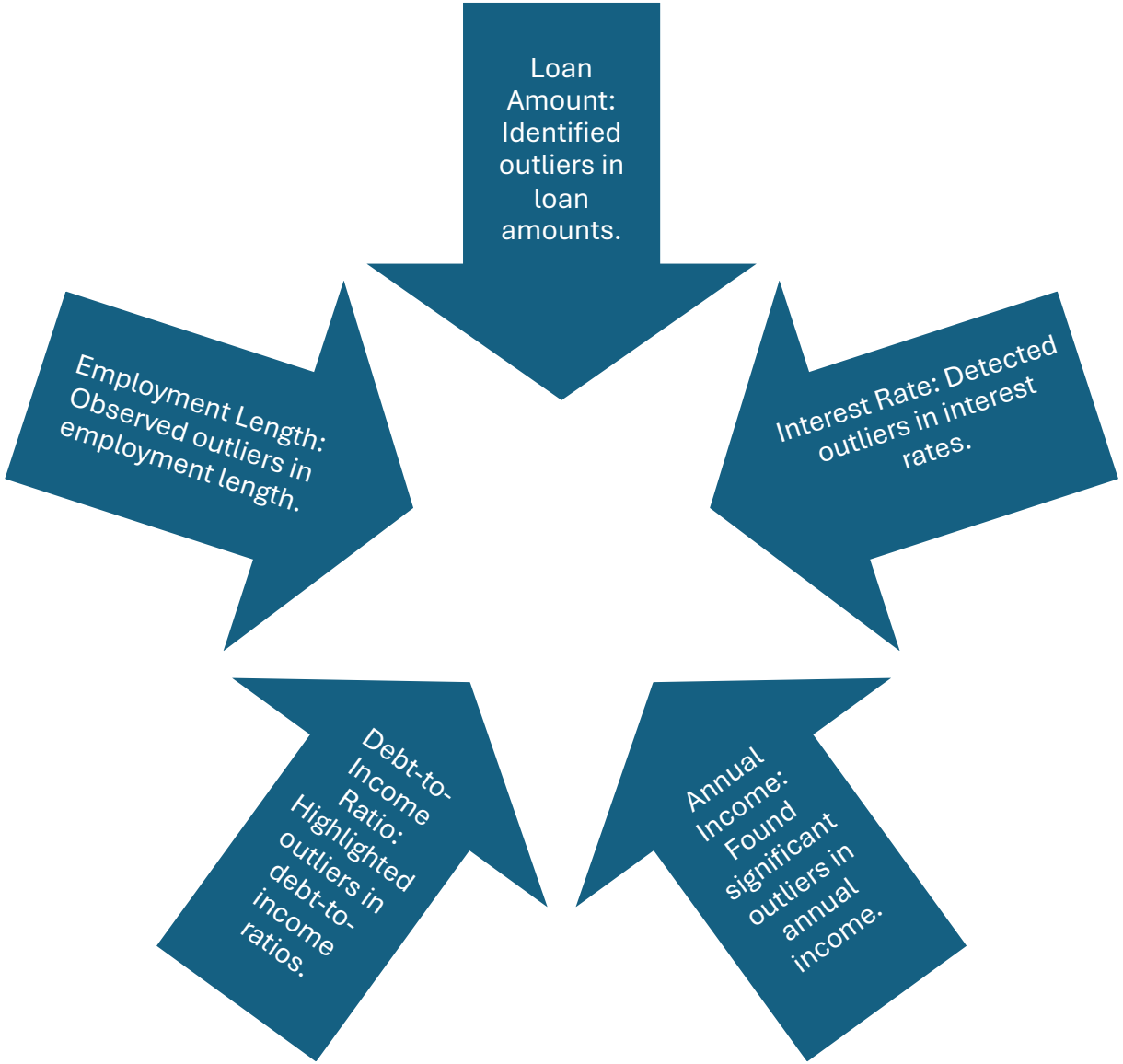
title

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

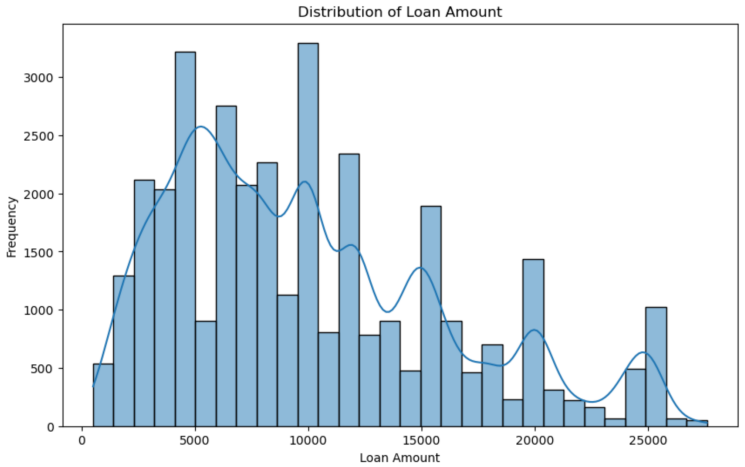
last_credit_pull_d

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

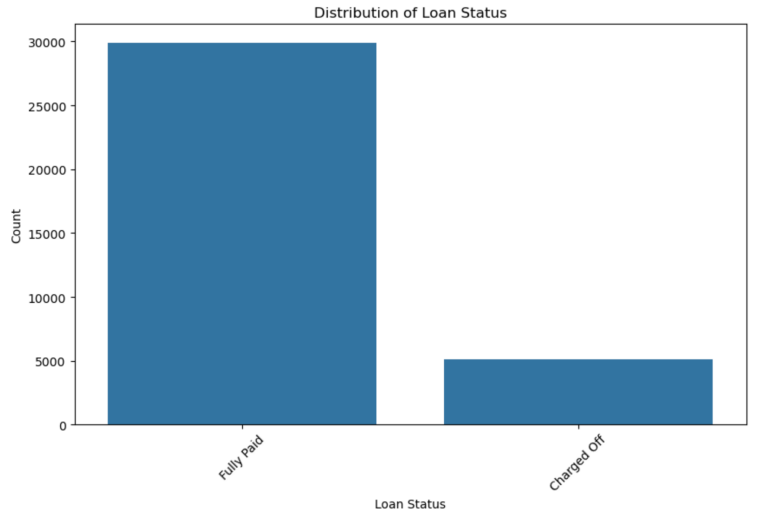


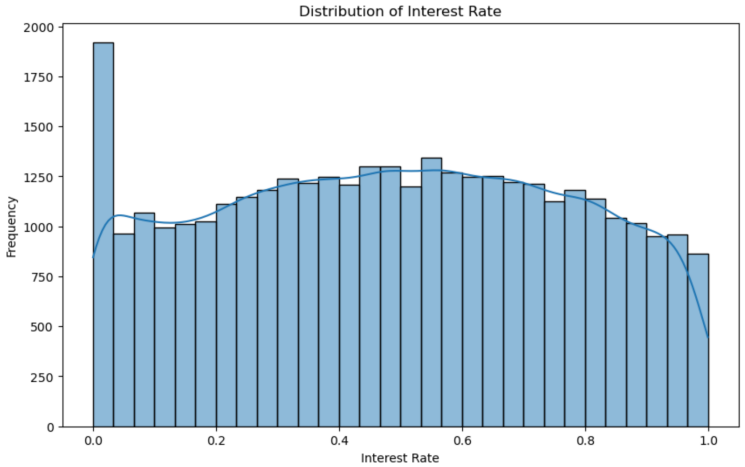
Univariate Analysis



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.

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.





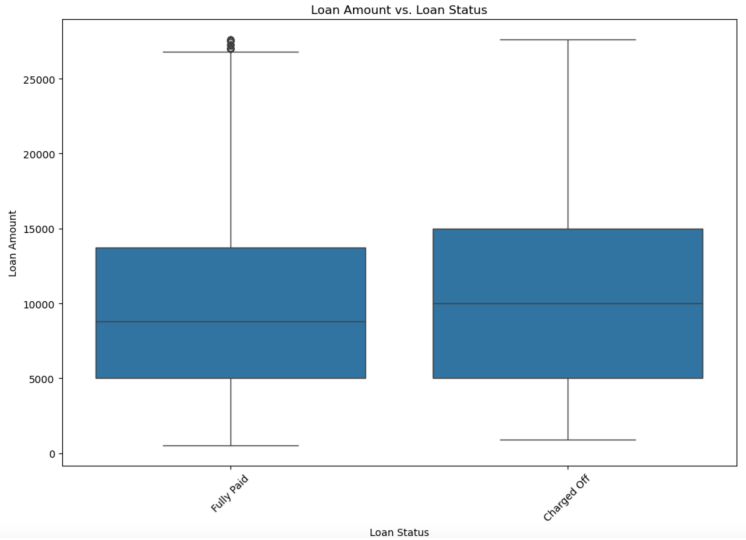
Interest rates are relatively uniformly distributed, with a slight concentration around mid-range values. There is a significant peak at the lower end, suggesting many borrowers receive lower interest rates.

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.

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.

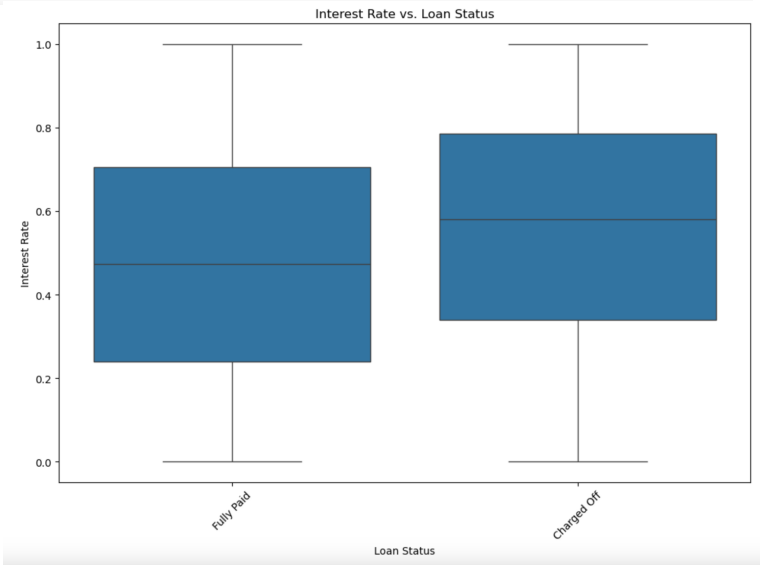
The high repayment rate suggests that current credit evaluation and risk assessment methods are effective. The lower percentage of charged-off loans indicates good risk management.

Bivariate Analysis



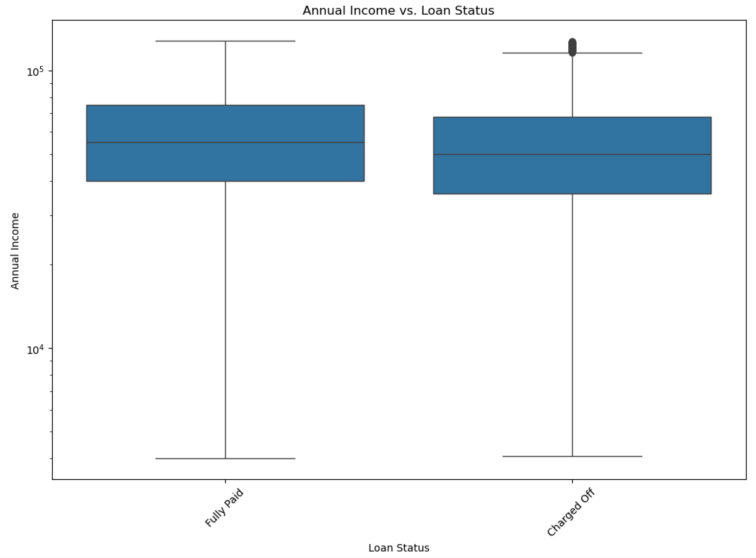
The median loan amount for "Charged Off" loans is slightly higher than that for "Fully Paid" loans. The interquartile range (IQR) and overall distribution for both categories are quite similar. "Fully Paid" loans have a few more high-value outliers compared to "Charged Off" loans.

Loan amounts for "Charged Off" and "Fully Paid" loans are similar, indicating that loan amount alone is not a decisive factor in predicting defaults. Borrowers with higher loan amounts do not necessarily default more often than those with lower loan amounts.



The median interest rate for "Charged Off" loans is higher than that for "Fully Paid" loans. The interquartile range (IQR) for "Charged Off" loans is also higher, indicating a wider spread of interest rates among defaults.

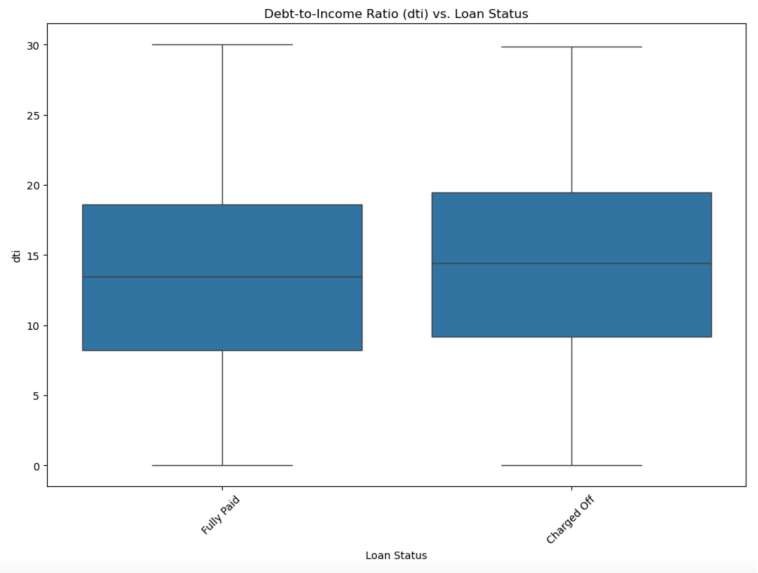
Higher interest rates are strongly associated with an increased likelihood of loan defaults. Borrowers facing higher interest rates tend to default more, suggesting that interest rate is a critical factor in loan performance.



The median annual income for "Fully Paid" loans is slightly higher than that for "Charged Off" loans. The distribution (IQR) for both categories is similar, with a few high-income outliers in the "Charged Off" category.

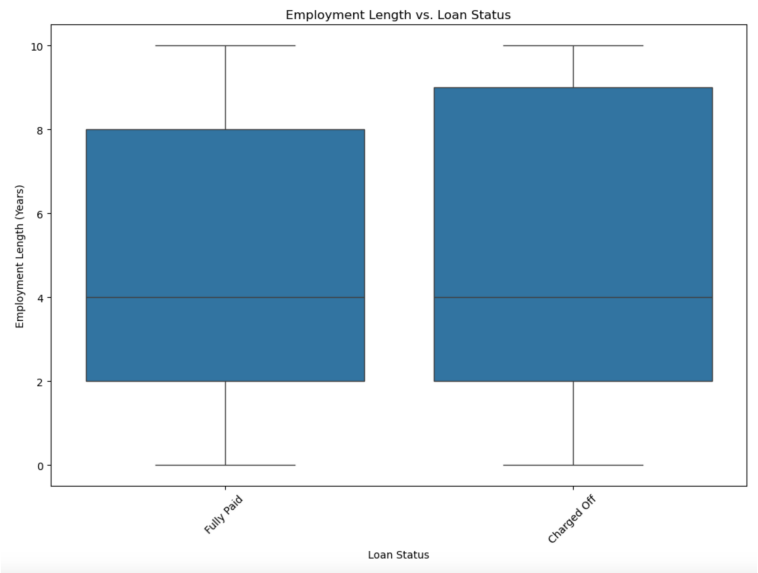
Lower annual incomes are associated with a higher likelihood of defaults. However, since there are high-income defaults as well, annual income is an important but not sole predictor of loan performance. Borrowers with lower incomes might struggle more with repayments.

Bivariate Analysis



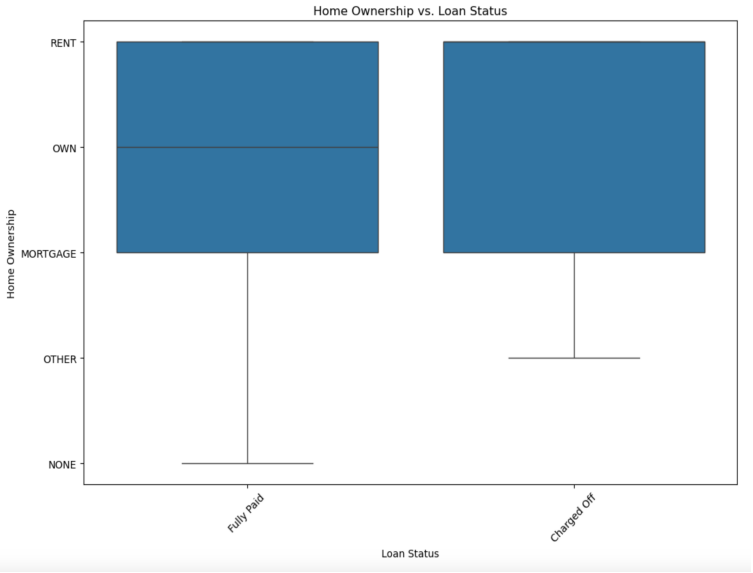
The median dti for "Charged Off" loans is slightly higher than for "Fully Paid" loans. Both "Fully Paid" and "Charged Off" loans have similar interquartile ranges (IQRs) and overall distributions.

While there is a slight increase in median dti for "Charged Off" loans, the similarity in distributions suggests that dti alone is not a strong predictor of loan defaults. However, a higher dti might still contribute to a marginally increased risk of default.



The median employment length for both "Fully Paid" and "Charged Off" loans is nearly identical. The interquartile ranges (IQRs) are similar for both categories, with "Charged Off" loans showing a slightly broader distribution.

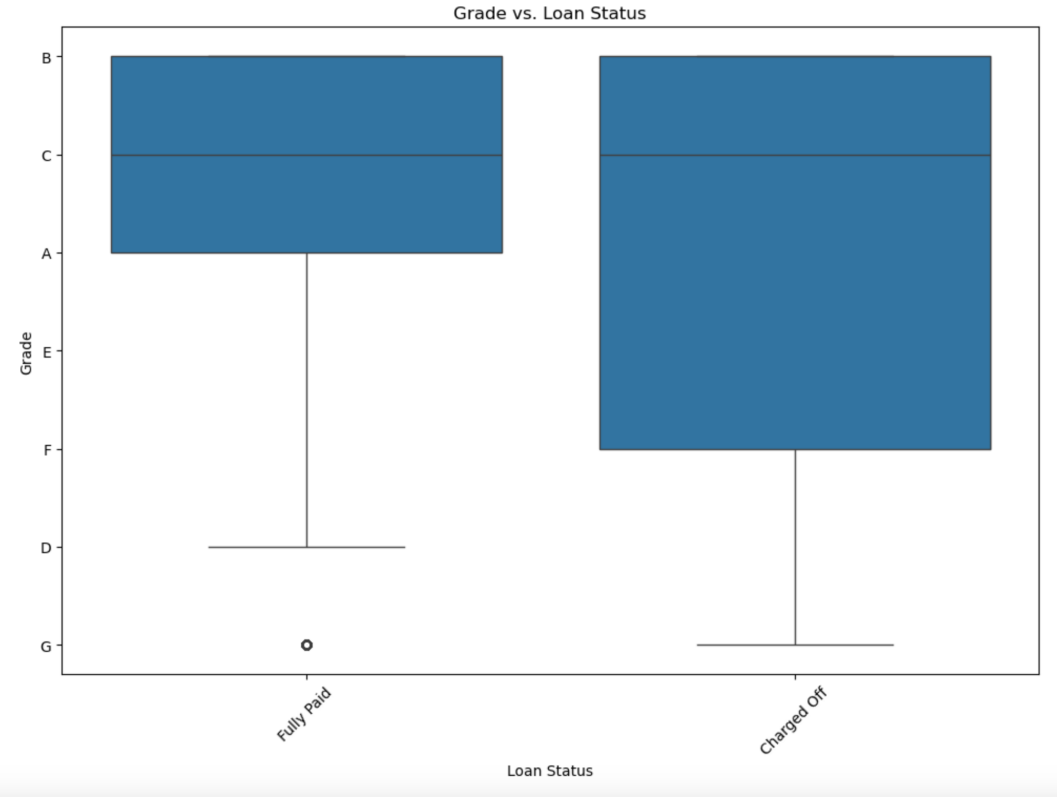
Employment length does not significantly differ between "Fully Paid" and "Charged Off" loans. This indicates that the length of employment is not a strong standalone predictor of loan defaults. Borrowers with various employment lengths have similar risks of defaulting

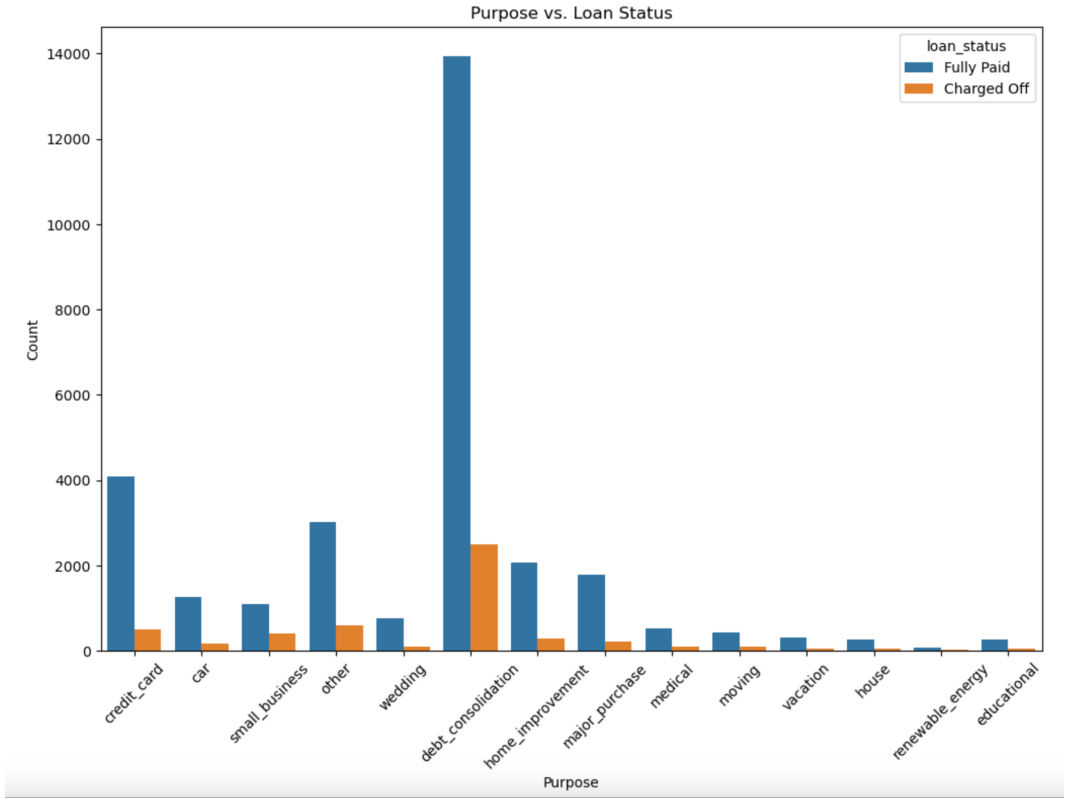


Both "Fully Paid" and "Charged Off" loans have similar distributions across home ownership categories (Rent, Own, Mortgage).The median values are slightly different but not significantly distinct.

Home ownership status (Rent, Own, Mortgage) does not significantly affect the likelihood of loan defaults. The similar distributions indicate that whether a borrower rents or owns their home does not strongly predict loan performance.

Bivariate Analysis





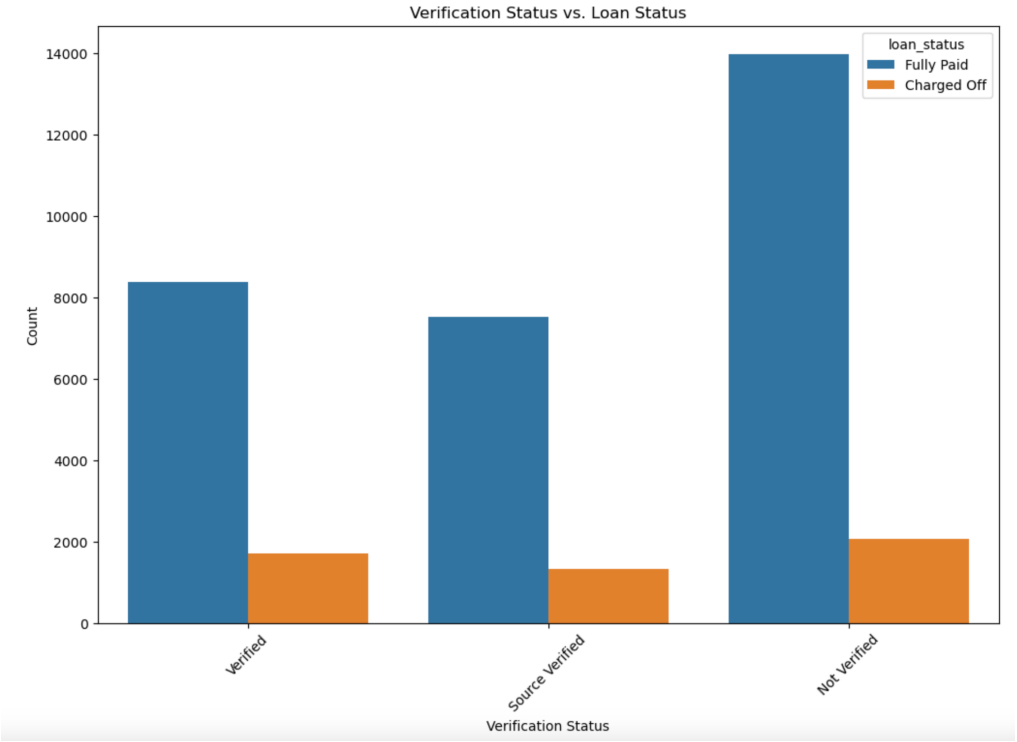
Both "Fully Paid" and "Charged Off" loans have a median grade around C. The IQR for "Fully Paid" loans extends from B to D, while for "Charged Off" loans, it extends from B to F.

Loan grade is a significant predictor of loan performance. Higher grades (B to C) are safer, while lower grades (D to F) have a higher risk of default. Lenders should be cautious with lower-grade loans.

Debt consolidation and small business loans show higher counts and proportions of "Charged Off" loans.

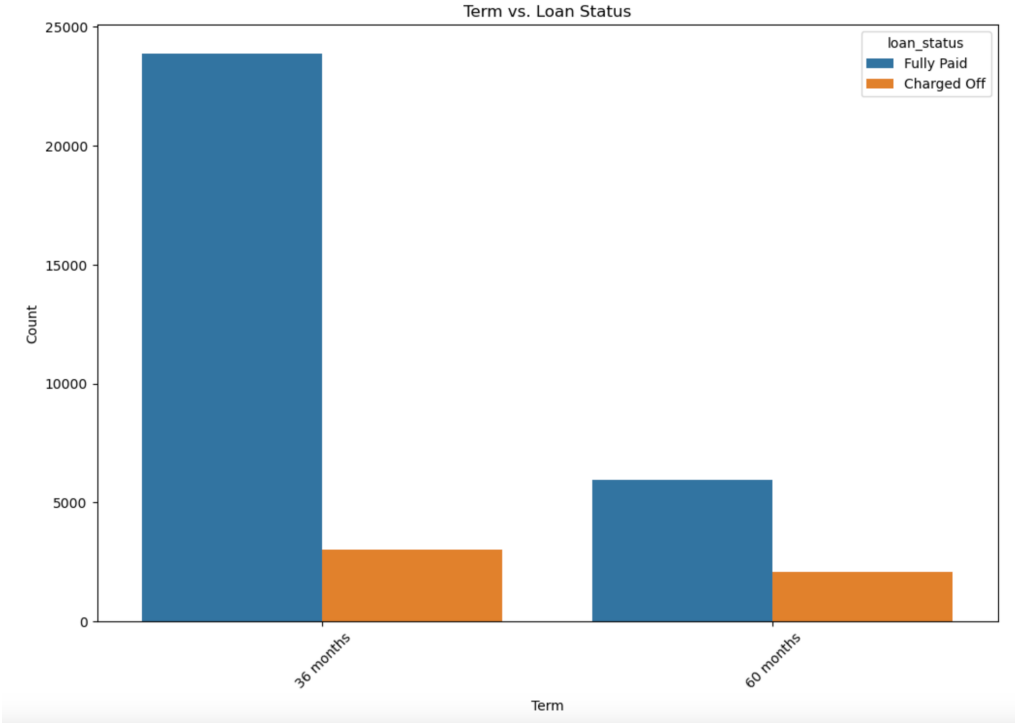
Loan purpose greatly influences default risk. Debt consolidation and small business loans are high-risk categories, requiring stricter evaluation and risk management. Other purposes like credit card and home improvement are relatively safer but still need careful consideration.

Bivariate Analysis



Across all verification statuses, "Fully Paid" loans (blue) are significantly higher than "Charged Off" loans (orange). "Not Verified" loans have a slightly higher count of defaults compared to "Verified" and "Source Verified" loans.

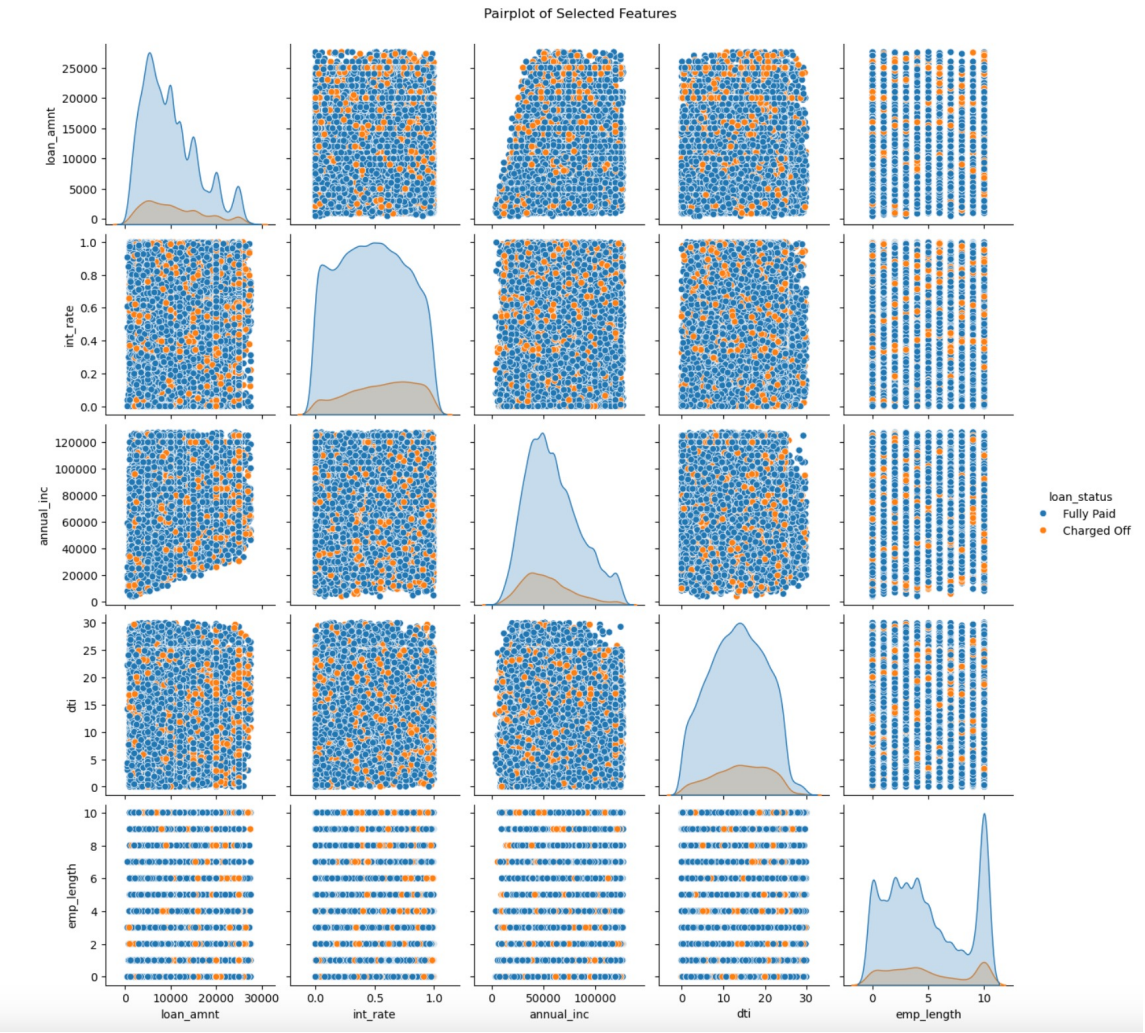
Income verification reduces the risk of loan defaults, but even "Not Verified" loans still have a substantial number of successful repayments. Verification status should still be a priority in the loan approval process to mitigate risk.



Loans with a 36-month term have a higher total number of defaults compared to loans with a 60-month term, even though the proportion of defaults is higher in 60-month loans. The majority of both "Fully Paid" and "Charged Off" loans fall under the 36-month term category.

While the proportion of defaults is higher for 60-month loans, the sheer volume of defaults in 36-month loans suggests that short-term loans are not immune to risk. Lenders need to balance loan term offerings, considering both volume and proportion of defaults, to manage risk effectively. Longer-term loans are riskier proportionally, but the higher volume of short-term loans requires careful monitoring as well.

Multivariate Analysis



Loan Amount (loan_amnt):
The scatter plots show that both "Fully Paid" and "Charged Off" loans are spread across all loan amounts. However, there is a higher density of "Fully Paid" loans at lower loan amounts. Higher loan amounts do not show a clear trend towards more defaults, but there is a slight increase in "Charged Off" loans at higher loan amounts.

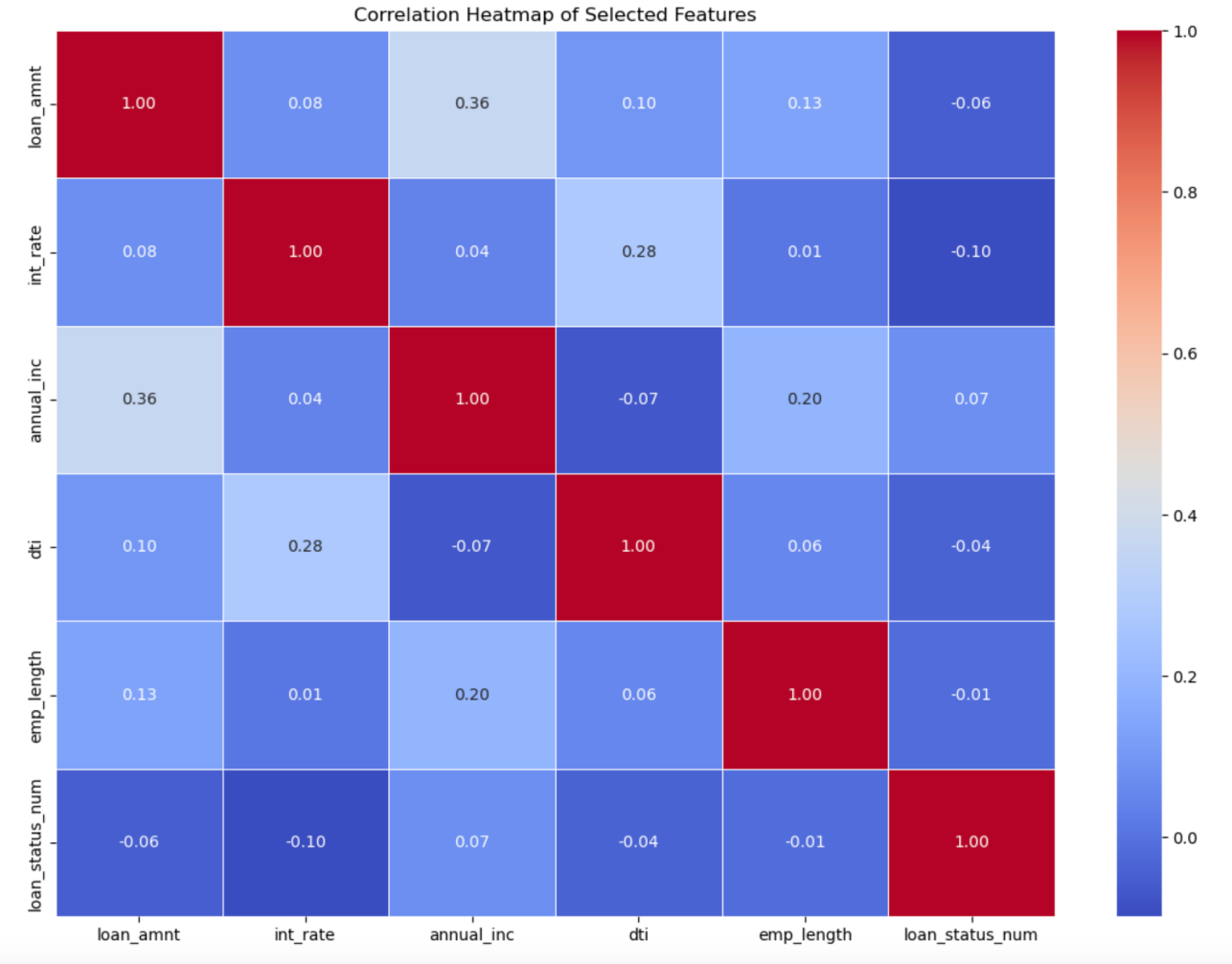
Interest Rate (int_rate):
There is a noticeable concentration of "Charged Off" loans at higher interest rates. The density plot shows a higher peak for "Charged Off" loans at higher interest rates compared to "Fully Paid" loans. Higher interest rates are associated with a higher likelihood of defaults.

Annual Income (annual_inc):
"Charged Off" loans are more frequent in the lower annual income ranges, though "Fully Paid" loans are also present across all income levels. The density plot shows a higher peak for "Fully Paid" loans at higher incomes.
Lower annual incomes are associated with a higher likelihood of defaults.

Debt-to-Income Ratio (dti):
There is a higher concentration of "Charged Off" loans at higher dti values. The density plot shows a higher peak for "Charged Off" loans at higher dti values compared to "Fully Paid" loans. Higher dti values are associated with a higher likelihood of defaults.

Employment Length (emp_length):
Both "Fully Paid" and "Charged Off" loans are distributed across all employment lengths, with a slight increase in "Charged Off" loans at shorter employment lengths. Shorter employment lengths might be associated with a higher likelihood of defaults, though this is not a strong predictor.

Multivariate Analysis



Loan Amount (loan_amnt):
Weak negative correlation with loan status (-0.06). Loan amount has a weak association with loan status, indicating it is not a strong predictor of defaults.

Interest Rate (int_rate):
Moderate negative correlation with loan status (-0.10). Higher interest rates are somewhat associated with higher defaults.

Annual Income (annual_inc):
Weak positive correlation with loan status (0.07). Higher incomes are slightly associated with lower defaults.

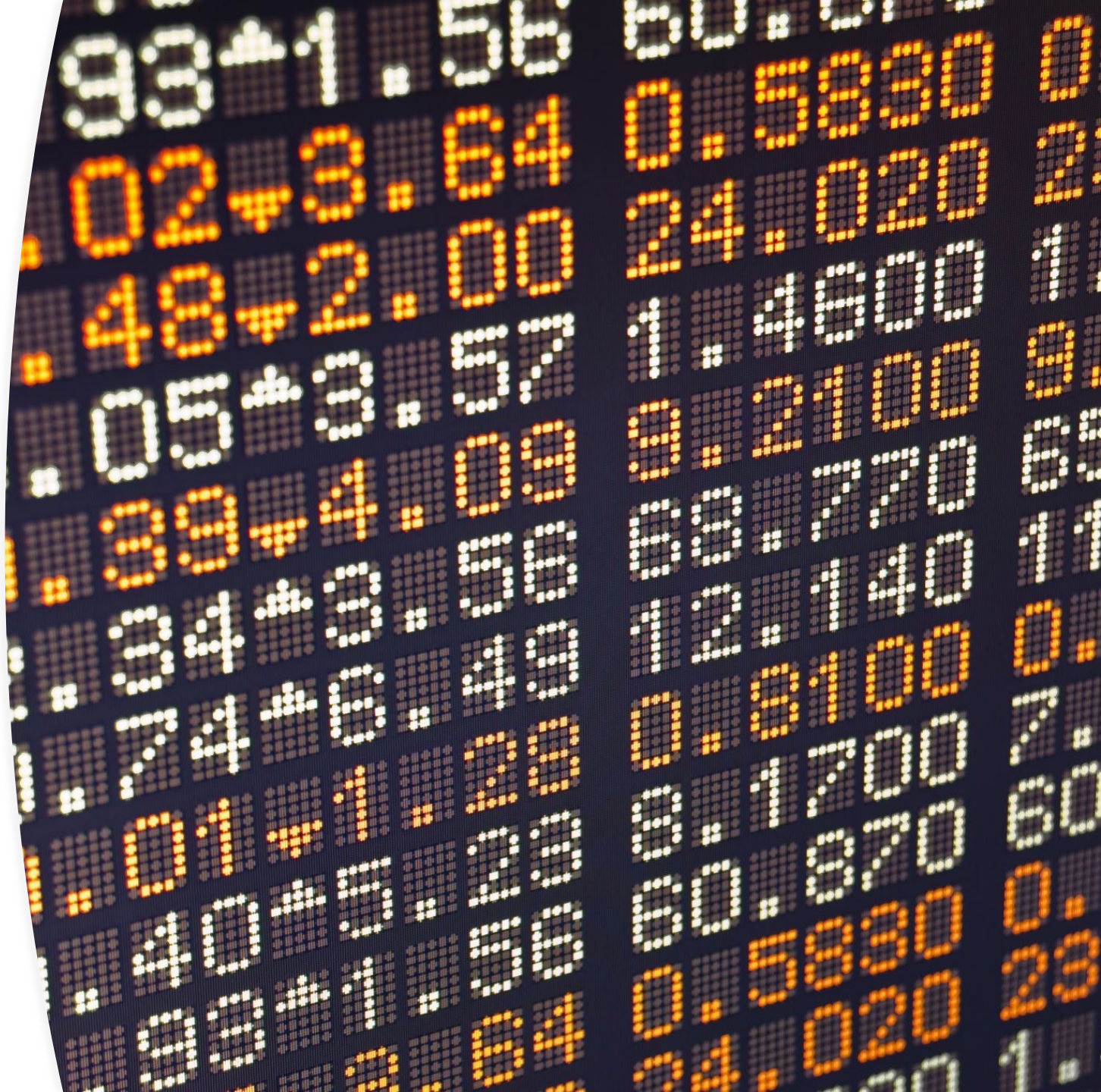
Debt-to-Income Ratio (dti):
Weak negative correlation with loan status (-0.04). Higher dti values have a slight association with higher defaults.

Employment Length (emp_length):
Very weak negative correlation with loan status (-0.01). Employment length has a very weak association with loan status.

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