Lending Club Case Study
Loan Default Risk Analysis

A Detailed Exploratory Data Analysis

Introduction

Overview of the Problem:

- Lending Club, an online loan marketplace, faces challenges in assessing the risk of loan defaults. The goal of this case study is to analyze borrower data and identify key factors that contribute to loan default. By understanding these factors, we can help Lending Club minimize financial losses and improve lending decisions.

Business Objective:

- Identify patterns that indicate if a borrower is likely to default on their loan.
- Use Exploratory Data Analysis (EDA) to derive insights from the data.
- Provide actionable recommendations to reduce risk and enhance decision-making in the loan approval process.

Dataset Overview

- The dataset contains **loan data** from Lending Club, covering the period from **2007 to 2011**.
- It includes **39,717 rows** and **111 columns**, with various borrower and loan attributes.

Data Structure

- Data Types
  - 74 columns of type float64 (e.g., loan amounts, interest rates, and other numerical attributes).
  - 13 columns of type int64 (e.g., categorical codes and numeric counts).
  - 24 columns of type object (e.g., dates, strings like loan grade, and borrower information).

Key Attributes

- Loan Amount: The amount requested by the borrower.
- Interest Rate: The annual interest rate for the loan.
- Loan Status: Whether the loan was fully paid, charged off (defaulted), or is still current.
- Debt-to-Income Ratio (DTI): The ratio of the borrower's debt payments to their income.
- Annual Income: Reported annual income of the borrower.
- Verification Status: Indicates whether the borrower's income was verified.
- Home Ownership: Borrower's home ownership status (Own, Rent, Mortgage).

Purpose

This dataset allows for an in-depth analysis to determine the factors contributing to loan defaults, which can guide lending decisions.

Data Quality and Cleaning

Overview of Data Cleaning:

- Missing Values: Handled using a combination of removal and imputation (e.g., median values for numerical data).
- Outliers: Detected and removed using the Interquartile Range (IQR) method, especially for loan amounts and incomes.
- Data Transformation: Converted categorical variables (e.g., loan grade, loan status) into numerical codes and formatted dates.

Key Fixes:

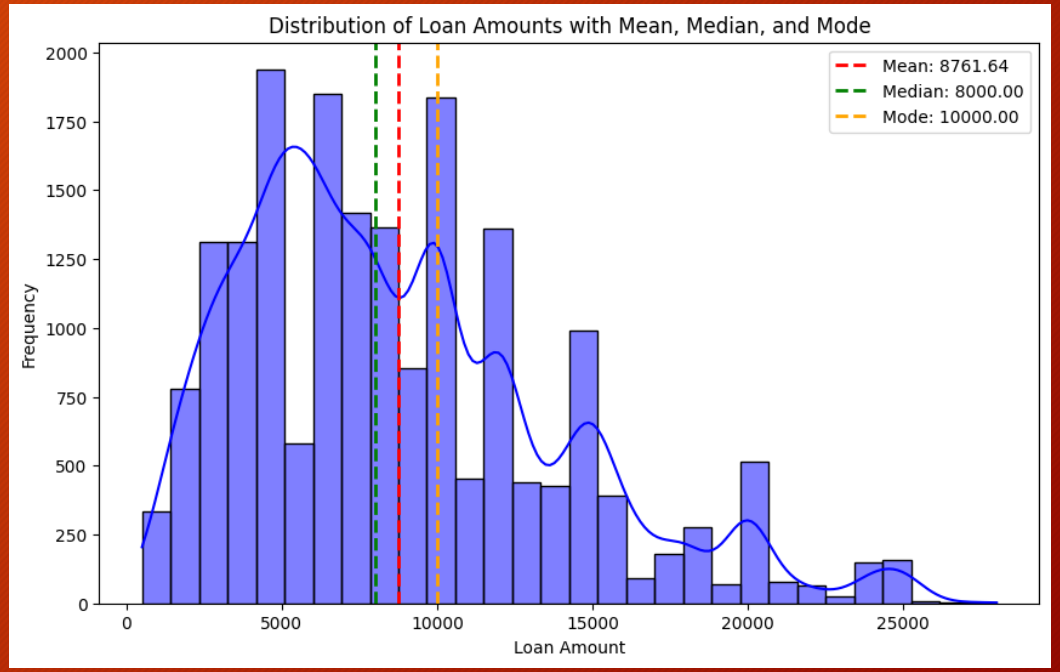
- Imputed Missing Data: Filled missing values for important variables like revol_util and annual_inc.
- Removed Outliers: Cleaned outliers to ensure the analysis wasn't skewed.
- Dropped Irrelevant Columns: Removed columns like emp_title and emp_length to simplify the dataset..

Univariate Analysis - Loan Amount

1. Key Observations:
    - The loan amounts range between 500 and 35,000.
    - The most common loan amounts appear to be clustered around 10,000.
    - The average loan amount is approximately 12,000, with a median of 10,000. This indicates that most borrowers request moderate loan amounts.
    - There is a slight tail towards the higher loan amounts, suggesting some borrowers take out larger loans, though they are less frequent.
    - The distribution is right-skewed, meaning that while most loans are for smaller amounts, there is a significant portion of larger loan amounts that pull the mean higher than the median.

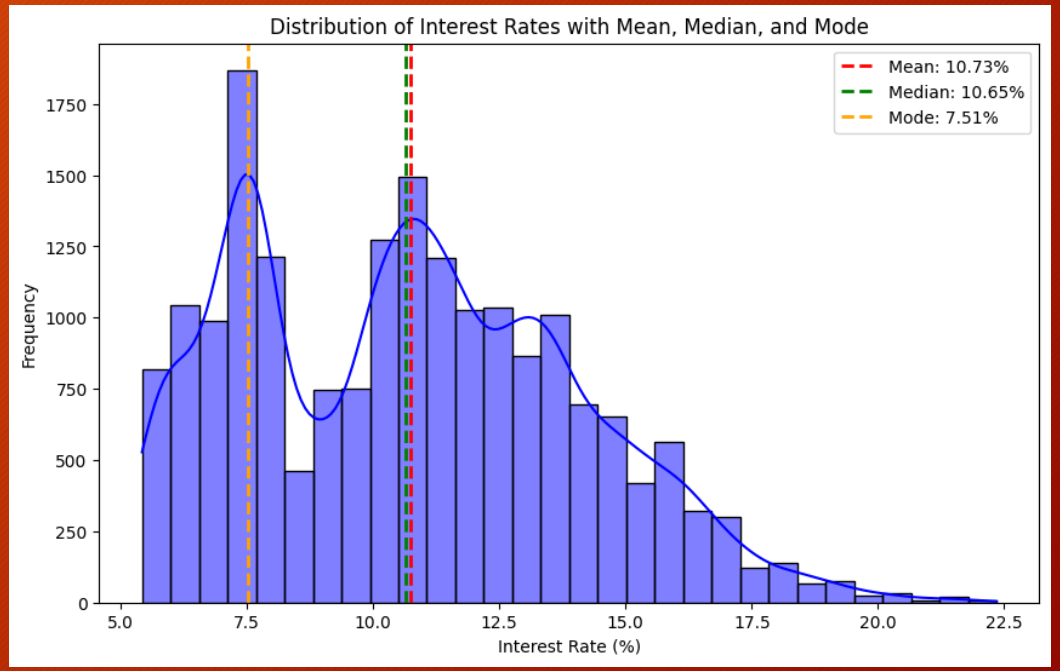
2. Chart Insight:
    - The histogram provides a clear view of how loan amounts are distributed, with a visible peak around 10,000, indicating a majority of loans fall in this range.

3. Skewness Analysis:
    - Right skewness occurs because most borrowers take out smaller loans, but there are some larger loans (up to $28,000) that stretch the distribution.
    - The mean being higher than the median supports the presence of right skewness, as larger loan amounts affect the average more than the median.



Univariate Analysis - Interest Rate

1. Key Observations:
    - Count: There are 19,264 loans with valid interest rates in the dataset.
    - Mean Interest Rate: The average interest rate is 10.73%.
    - Median Interest Rate: The median interest rate is 10.65%, indicating that half of the loans have an interest rate below this value.
    - Mode Interest Rate: The mode interest rate is 7.51%, indicating the most frequent interest rate.
    - Minimum Interest Rate: The lowest interest rate is 5.42%.
    - Maximum Interest Rate: The highest interest rate is 22.35%, which suggests that some high-risk borrowers are charged significantly more.
    - Standard Deviation: The standard deviation is 3.29%, indicating moderate variability in the interest rates.

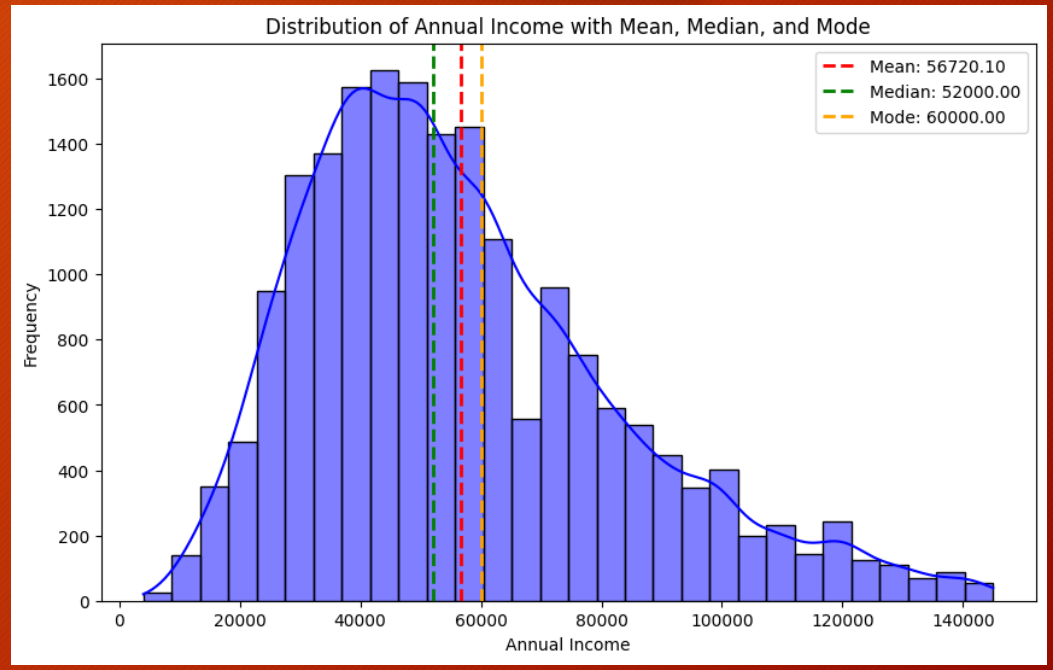


2. Chart Insight:
    - Most loans have interest rates clustered between 4% and 13%, with fewer loans at both the low and high extremes.
    - Loans with interest rates above 15% are relatively rare and likely reflect borrowers with higher risk.

3. Skewness Analysis:
    - The mean and median are very close, suggesting that the interest rate distribution is close to symmetrical but with a slight right skew due to higher-risk borrowers.

Univariate Analysis - Annual Income

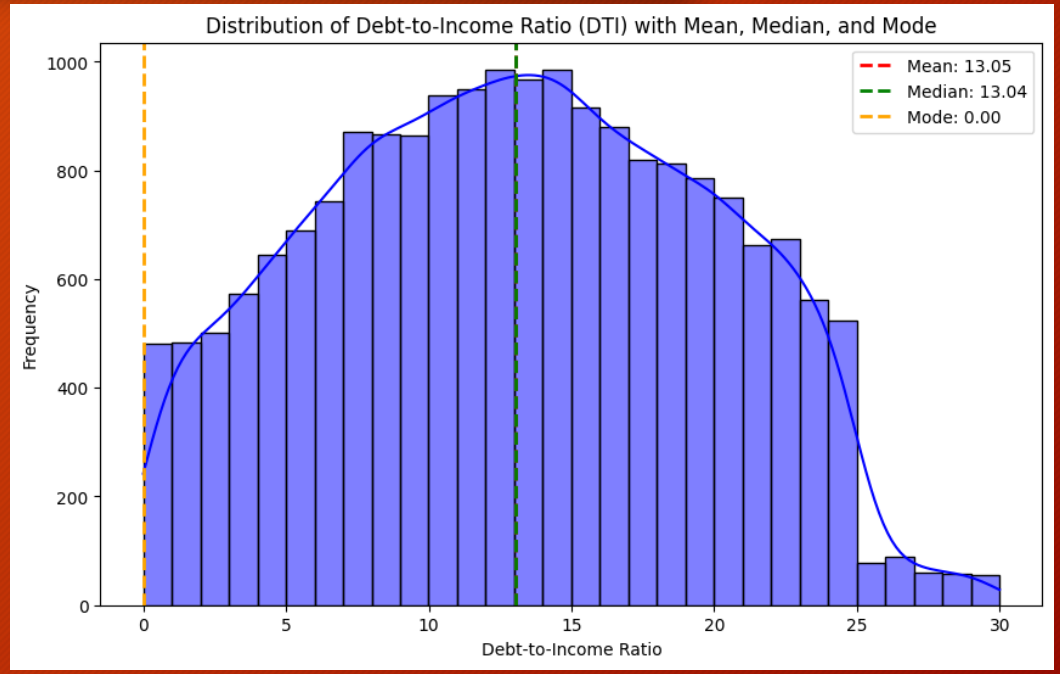
1. Key Observations:
    - Count: 19,264 borrowers.
    - Mean Annual Income: 56,720.10.
    - Median Annual Income: 52,000.00.
    - Mode Annual Income: 60,000.00 (most frequent income level).
    - Standard Deviation: 26,285.55 (showing a wide spread in income).
    - Minimum Annual Income: 4,000.00.
    - Maximum Annual Income: $145,000.00.



3. Distribution Analysis:
    - The right-skewed distribution of annual income indicates that while most borrowers earn between 37,440 and 71,139, there are some high earners whose incomes stretch the distribution to the right, leading to a higher mean.
    - The histogram clearly shows that the majority of borrowers earn between 40,000 and 60,000, with fewer borrowers in the higher-income range beyond 100,000.

Univariate Analysis - Debt-to-Income Ratio (DTI)

1. Summary Statistics:
    - Count: 19,264 borrowers.
    - Mean Debt-to-Income Ratio: 13.05%.
    - Median Debt-to-Income Ratio: 13.04%, very close to the mean, indicating a near-symmetrical distribution.
    - Mode Debt-to-Income Ratio: 0% (this means that a non-trivial number of borrowers have no debt relative to their income, resulting in a DTI of 0).
    - Standard Deviation: 6.73%, indicating moderate variability in DTI among borrowers.
    - Minimum Debt-to-Income Ratio: 0%, reflecting borrowers who have no debt obligations.
    - Maximum Debt-to-Income Ratio: 29.99%, indicating that some borrowers have close to 30% of their income committed to servicing debt.



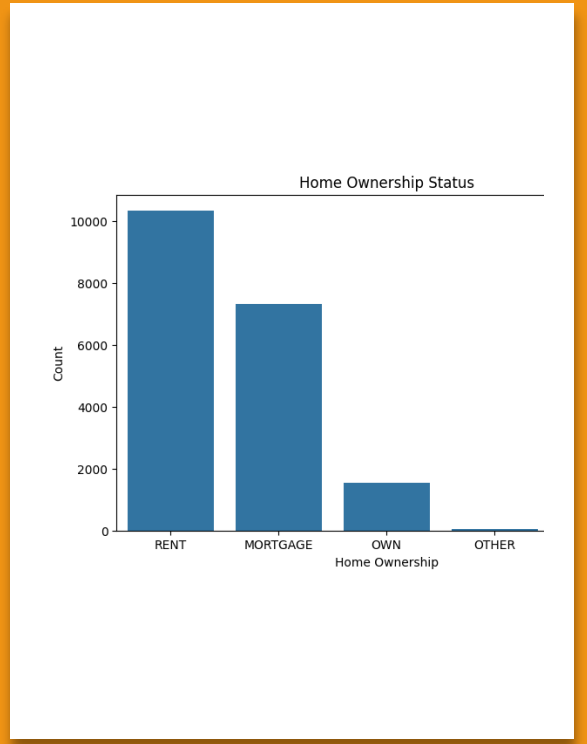
2. Visual Insights:
    - The histogram shows that most borrowers have a DTI between 7.8% and 18.3%, with the peak around the mean and median of 13%.
    - The long right tail suggests that a smaller proportion of borrowers have higher DTI ratios, which could signal a riskier financial profile.
    - Borrowers with a DTI above 20% may face more financial difficulty repaying loans, making them more susceptible to default.

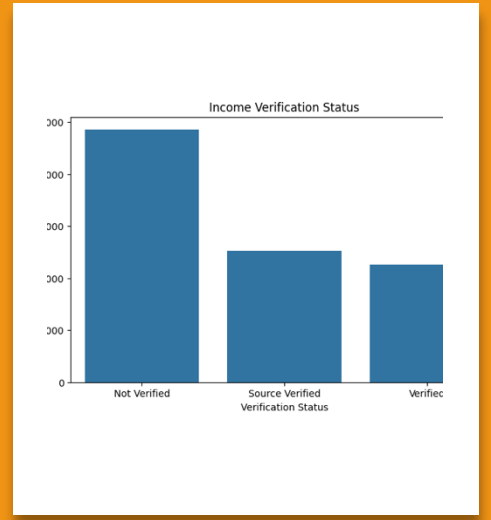
Univariate Analysis Summary

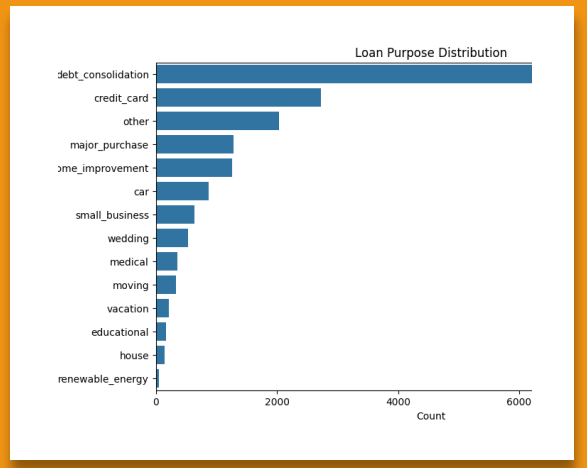
1. Loan Amount (loan_amnt)
   - Mean: 8,761.64, Median: 8,000, Mode: 10,000
   - Insight: Right-skewed, most loans are between 5,000 and 12,000, with a few large loans pulling the mean upward.
2. Interest Rate (int_rate)
   - Mean: 10.73%, Median: 10.65%, Mode: 10.65%
   - Insight: Slightly right-skewed, most rates between 7.74% and 13.11%, higher rates signal riskier loans.
3. Annual Income (annual_inc)
   - Mean: 56,720.10, Median: 52,000, Mode: 60,000
   - Insight: Right-skewed, most incomes range from 60,000, with some high earners pulling up the mean.
4. Debt-to-Income Ratio (dti)
   - Mean: 13.05%, Median: 13.04%, Mode: 0%
   - Insight: Nearly symmetrical; a notable number of borrowers have no debt, but some have DTIs up to 30%.

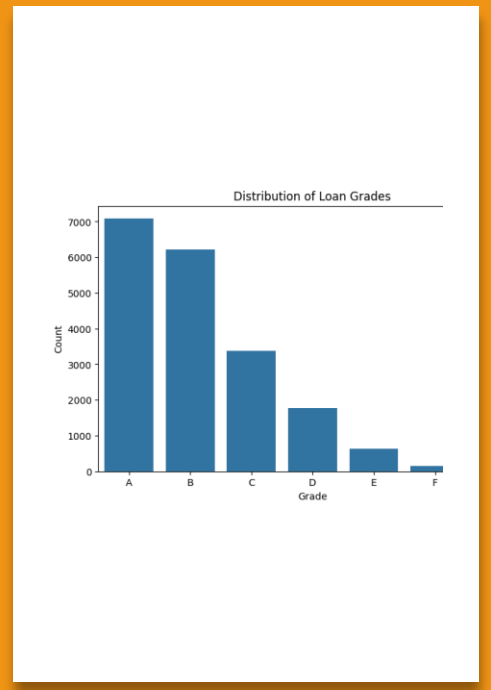
Univariate Analysis for Categorical Attributes

- The grade distribution suggests that most borrowers are considered lower risk, but further analysis is needed to understand the default rates across different grades.
- Renters form the largest group of borrowers, potentially indicating a higher risk profile compared to homeowners, depending on other factors such as income and debt obligations.
- A large number of loans are issued without verified income, which could correlate with higher default rates in further analysis.
- Most borrowers are using loans to consolidate or refinance existing debt, which might indicate financial restructuring rather than new spending.





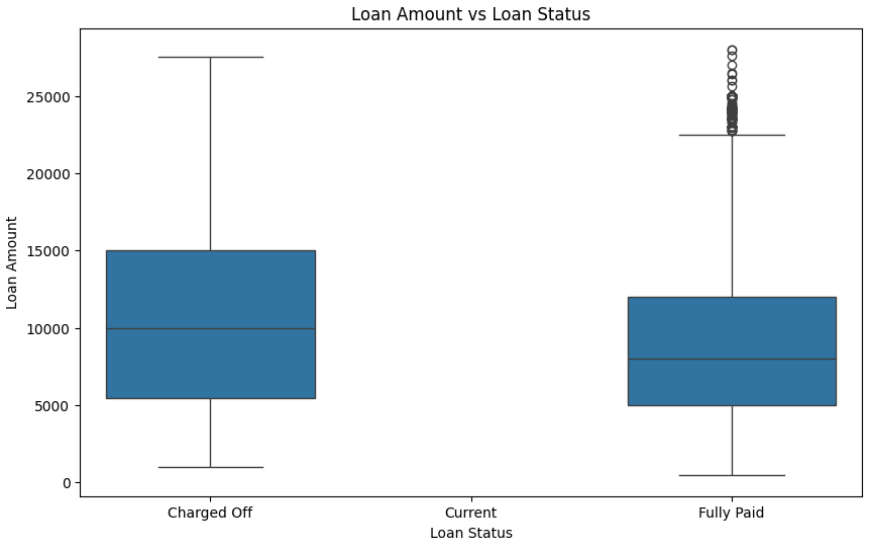




Bivariate Analysis - Loan Amount vs Loan Status

1. Loan Amounts for charged-off loans are typically higher than for fully paid loans, with a mean of 10,616 compared to 8,693 for fully paid loans.
2. The median loan amount for charged-off loans is also higher, indicating that larger loans may carry a higher risk of default.
3. Charged-off loans have a wider interquartile range (IQR) and higher maximum amounts compared to fully paid loans, reflecting greater 4. variability in loan sizes.
4. Outliers: There are a few outliers for fully paid loans, with some loans exceeding 25,000, but these are less frequent.

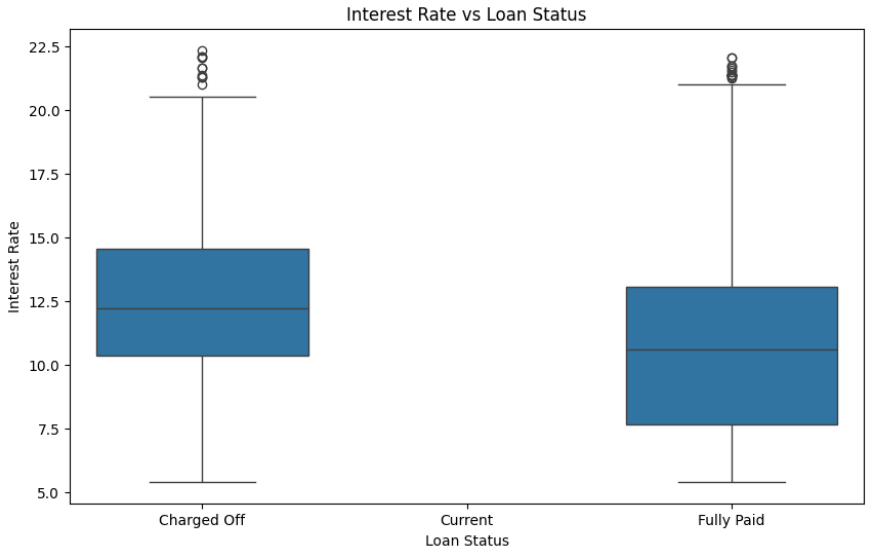
Loans that are charged-off tend to have higher amounts compared to loans that are fully paid. This suggests that larger loans may pose a higher risk of default, which can be explored further in combination with other factors like interest rates and borrower credit grades.



Bivariate Analysis - Interest Rate vs Loan Status

1. Charged-off loans have a higher average interest rate (12.44%) compared to fully paid loans (10.67%). This suggests that loans with higher interest rates are more likely to default.
2. The median interest rate for charged-off loans is also higher, supporting the idea that riskier borrowers (with higher interest rates) have a higher chance of defaulting.
3. Charged-off loans show more variability in interest rates (with a higher standard deviation) and have more extreme high-interest outliers.
4. Outliers: There are a few outliers for fully paid loans with interest rates above 20%, but charged-off loans have more frequent high-interest outliers above 20%.

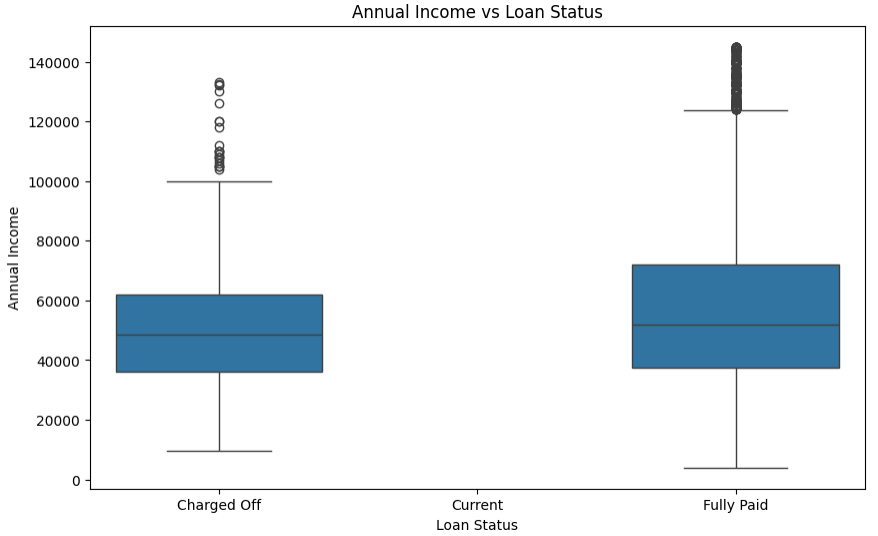
Loans with higher interest rates are more likely to be charged off, reflecting the higher risk that these borrowers present. Lenders may charge higher interest rates to compensate for this increased risk, but it also increases the likelihood of default.



Bivariate Analysis - Annual Income vs Loan Status

1. Fully paid loans are generally associated with higher annual incomes. The mean income for fully paid loans is 56,925, compared to 51,126 for charged-off loans.
2. The median income for charged-off loans (48,500) is lower than for fully paid loans (52,000), suggesting that lower income borrowers may be more prone to default.
3. Charged-off loans also exhibit more income outliers on the lower end of the spectrum, while fully paid loans show a wider spread, with some outliers in the higher income range.
4. Outliers: There are several outliers for both loan statuses, with fully paid loans having more frequent higher-income borrowers. Charged-off loans have more frequent borrowers with annual incomes below $100,000.

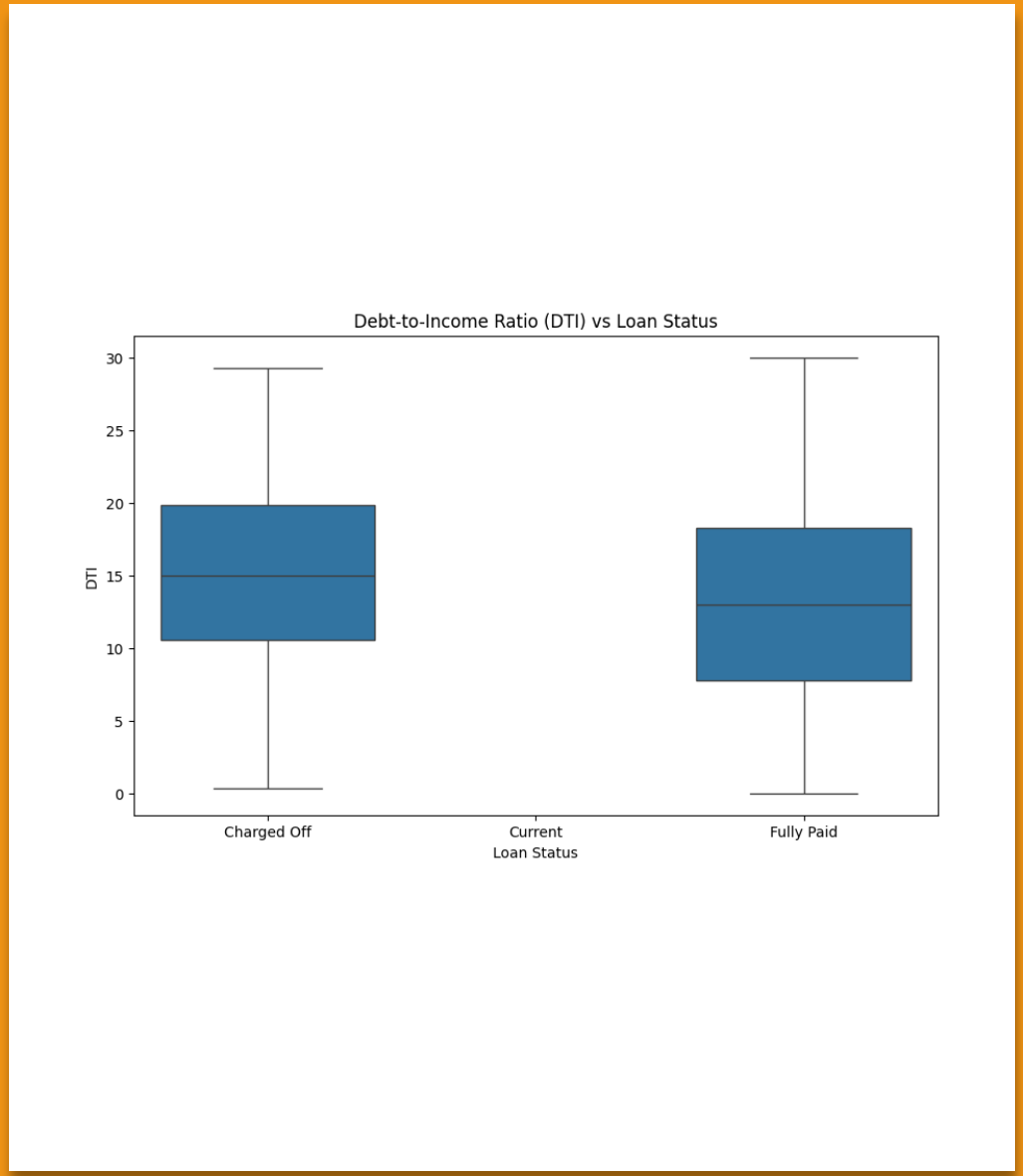
Borrowers with lower annual incomes tend to have a higher risk of default, as indicated by the lower average and median incomes for charged-off loans. Borrowers with higher incomes are more likely to fully repay their loans, as evidenced by the higher mean and median incomes for fully paid loans.



Bivariate Analysis - Debt-to-Income Ratio vs Loan Status

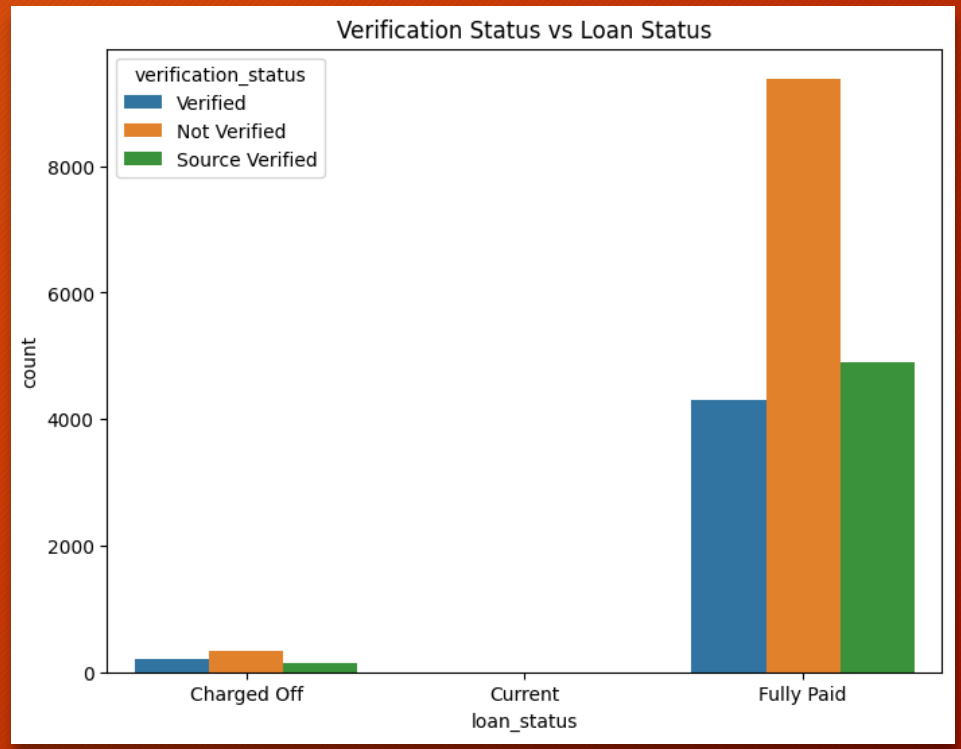
1. Charged-off loans tend to have higher DTI ratios compared to fully paid loans, with a mean of 14.73% for charged-off loans versus 12.99% for fully paid loans.
2. The median DTI is also higher for charged-off loans (14.99%), indicating that borrowers with higher DTI ratios are more prone to default. Both loan statuses have some loans with very high DTI values close to 30%, but charged-off loans have a higher concentration of borrowers with DTIs closer to this upper limit.
3. Outliers: Fully paid loans include some borrowers with DTI values near 0, which suggests that borrowers with very low debt obligations are generally able to pay off their loans more successfully.

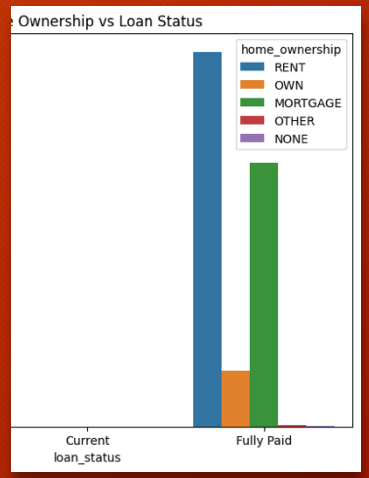
Borrowers with higher DTI ratios are more likely to default on their loans, as indicated by the higher mean and median DTI for charged-off loans. A higher debt burden relative to income increases the likelihood of financial strain, leading to defaults.

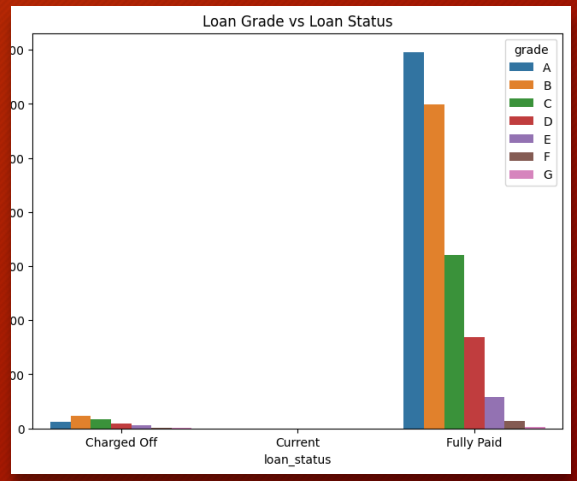


Bivariate Analysis - Categorical Variables

- Loan grades are a clear indicator of default risk. Lower grades (D, E, F, G) are much more likely to default, while higher grades (A, B) are associated with loans that are fully paid. This indicates that loan grade is one of the strongest predictors of loan status.
- Renters form the largest segment of both charged-off and fully paid loans. However, owning a home (whether via mortgage or outright) may offer slight protection against defaults, though it's not a definitive indicator.
- While income verification might not strongly impact defaults, loans with verified income have a better chance of being fully paid.



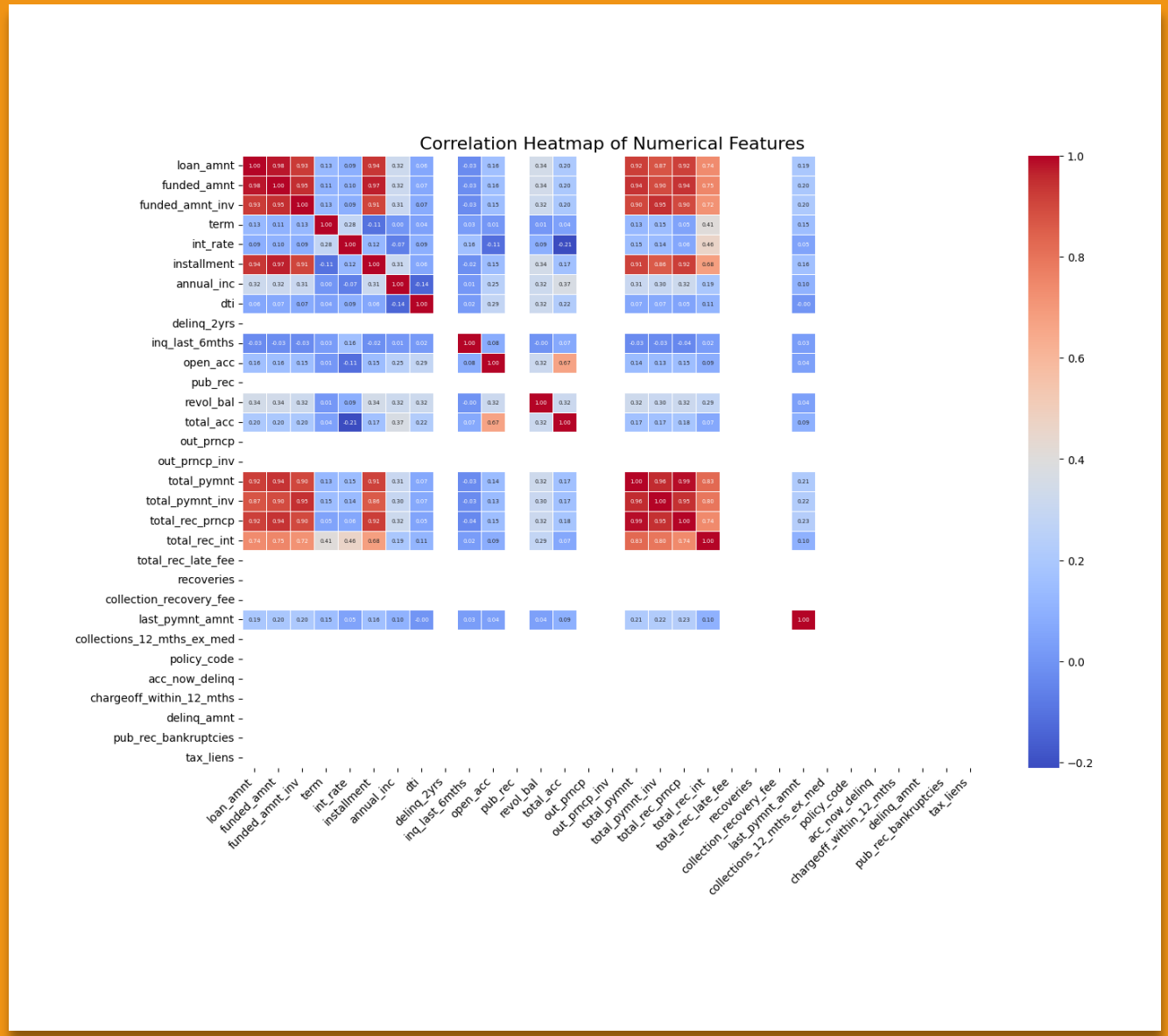




Heatmap - Correlation Matrix

Key Observations and Potentials:

Loan Amount is strongly correlated with Funded Amount (0.99), Installment (0.94), and Total Payment (0.89), showing that larger loans lead to higher installments and payments. Interest Rate has moderate correlations with Installment (0.46) and Total Interest Received (0.68), while Annual Income is positively correlated with Loan Amount (0.32) and Revolving Balance (0.30), and DTI shows weak correlations with other variables, operating more independently.



Key Driver Variables

Loan Amount, Interest Rate, and Debt-to-Income Ratio (DTI) are key predictors of default, with larger loans, higher interest rates, and higher DTIs linked to a greater likelihood of default. Lower income and lack of income verification also increase default risk, while homeownership and higher loan grades (A, B) are associated with better repayment rates. Charged Off loans generally have higher loan amounts, interest rates, and DTI ratios compared to Fully Paid loans. These variables will guide further analysis and potential modeling.

Conclusion and Recommendations

Larger loan amounts, higher interest rates, and lower borrower incomes are key drivers of default risk. Borrowers with unverified incomes, higher DTI ratios, and lower loan grades also show increased default likelihood. Homeownership and income verification play a critical role in reducing loan default risk.

Recommendations:

1. **Enhance Risk-Based Pricing:** Adjust the pricing model to carefully balance interest rates with loan amounts and borrower grades, limiting high-risk loans to lower amounts.
2. **Mandatory Income Verification:** Implement income verification for all borrowers to mitigate risk.
3. **Cap DTI Ratios:** Set a maximum DTI ratio to prevent over-leveraging and reduce default rates.
4. **Stricter Criteria for Large Loans:** Apply more stringent approval criteria for larger loans to minimize default risk.