Lending Club Case Study

Jayant Pratap Singh
Saurabh Rai

Contents

Introduction

Problem Statement

Lending Club, a peer-to-peer lending platform, is confronted with the challenge of optimizing its loan approval process to balance profit and risk effectively. The company needs to accurately identify which loan applicants are likely to default, as these defaults can lead to significant financial losses. Simultaneously, Lending Club must avoid rejecting creditworthy applicants who can contribute to the company's revenue through timely loan repayments.

Goal

The goal of this analysis is to assist Lending Club in refining its loan approval process by identifying key indicators of loan default. By understanding these risk factors, the company can make more informed decisions, reducing the likelihood of approving loans that may result in defaults and financial losses.

Objective

The primary objective of this case study is to conduct an Exploratory Data Analysis (EDA) on the provided dataset to uncover the variables that are most strongly associated with loan defaults. Through this analysis, Lending Club will gain insights into borrower characteristics and behaviors that predict defaults, enabling the company to enhance its risk management and loan approval strategies. The findings will help in minimizing credit losses while ensuring that profitable lending opportunities are not overlooked.

Data Understanding

The dataset contains loan information from 2007 to 2011. The data includes borrower details, loan attributes, and the status of the loan (fully paid, current, or charged-off). The aim is to understand which attributes are strong indicators of default.

The dataset originally contained over 39,717 records and 111 columns, offering a wealth of information for our analysis. These variables provided valuable insights into factors that might influence a borrower's ability to meet their loan obligations. To focus our analysis, we selected only the variables that directly or indirectly impact the likelihood of a borrower defaulting on their loan. This data preparation step involved carefully choosing the most relevant variables to ensure a more accurate and meaningful analysis.

Data Cleaning

Handling Missing Values

- **Identification of Missing Values**: The dataset was checked for missing values across all columns.
- **Dropping Columns with High Missing Data**: Columns with a significant amount of missing data that were not critical to the analysis were dropped to maintain data integrity.
- **Filling Missing Values**: For certain columns where missing values were minimal or the data was critical, values were imputed using appropriate methods (e.g., filling with the median, mode, or forward-fill techniques).
- **Dropping Rows with Missing Values:** Rows with missing values in critical columns such as `desc`, `emp_title`, and others were dropped. These columns were identified as essential for analysis, and missing values in these columns could lead to biased or incomplete results.

Data Type Conversion

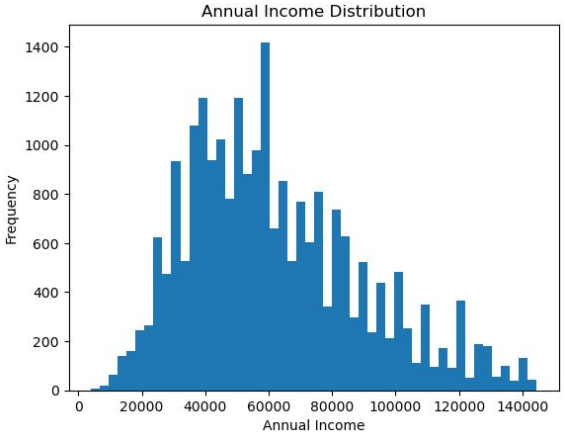
- **Columns Conversion**: Columns like `int_rate` and `revol_util`, which originally contained percentage values as objects, were converted to numeric types after removing the '%' symbol and dividing by 100.

Data Cleaning

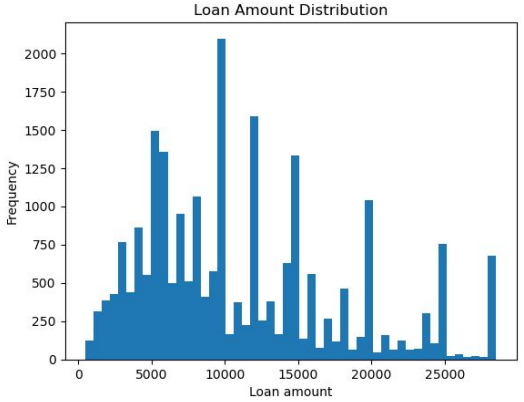
Outlier Detection and Handling

- **Identification of Outliers**: Continuous variables such as `annual_inc`, `loan_amnt`, and `revol_bal` were analyzed for outliers.
- **Removed Outliers**: Extreme outliers in certain variables were removed from the dataset. For instance, very high incomes were removed because they have high chance of paying back.
- **Capping Extreme Values**: Extreme outliers in certain variables were capped to reduce their influence on the analysis. For instance, loan amounts that were unrealistic or highly improbable were adjusted.

Univariate Analysis

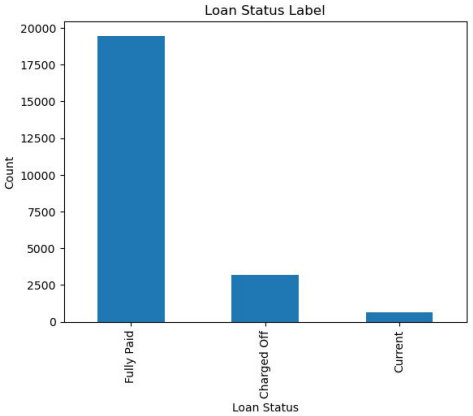


The histogram for annual income reveals that the majority of individuals earn between $40,000 and $60,000 per year.

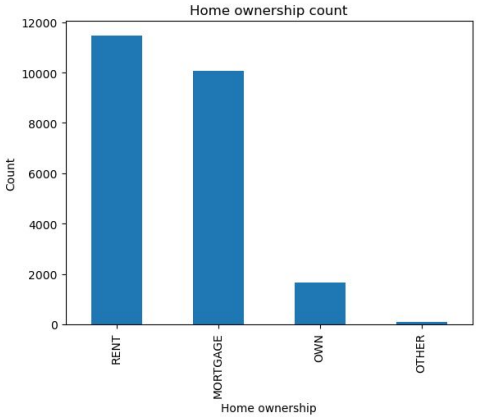


The data indicates that the company predominantly disburses loan amounts ranging from $5,000 to $15,000, with approximately $10,000 being the most common loan amount.

Univariate Analysis

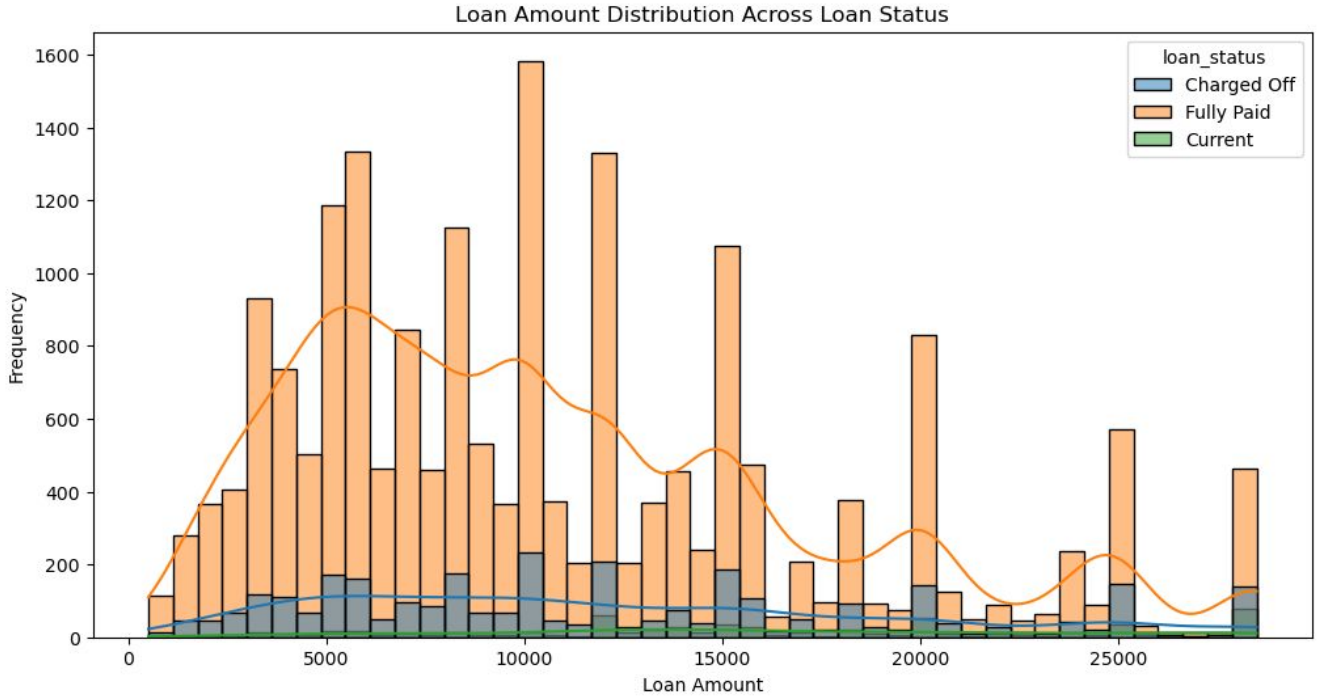


The majority of the loans provided by the company have been fully repaid

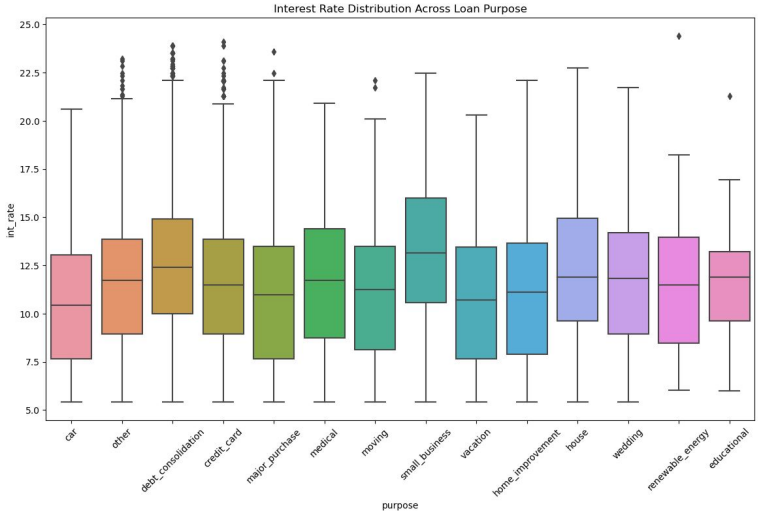


Most borrowers either live in rented housing or are residing in properties financed through a mortgage provided during registration.

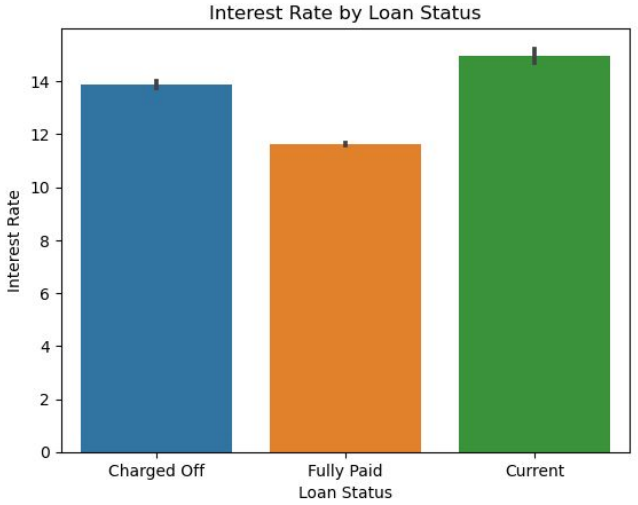
Univariate Analysis



Univariate Analysis

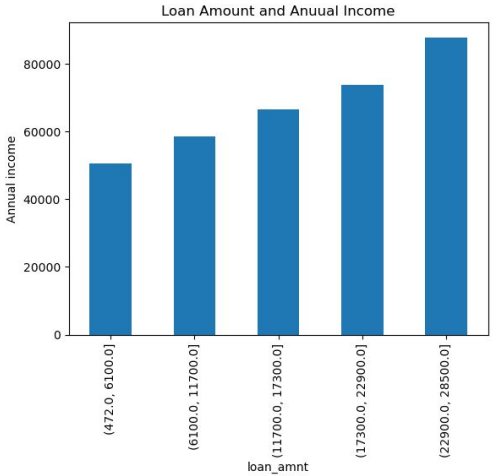


The median interest rate for vacation and car loans is lower compared to small business loans, which have a higher interest rate due to their higher risk.

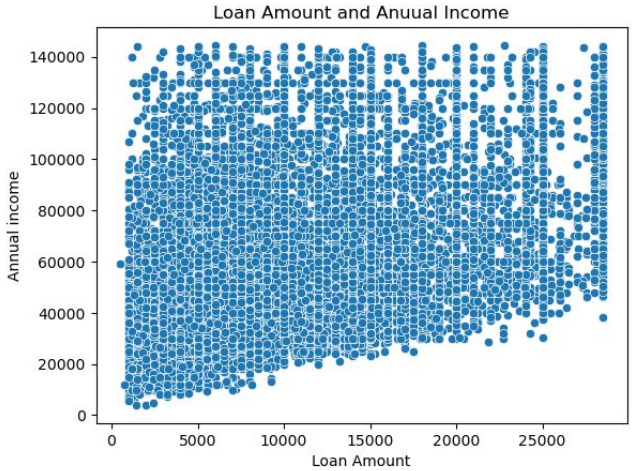


The graph indicate that loans offered to lower-risk borrowers, who fully repaid their loans, have lower interest rates. Conversely, loans to higher-risk borrowers, who defaulted, have higher interest rates.

Bivariate Analysis

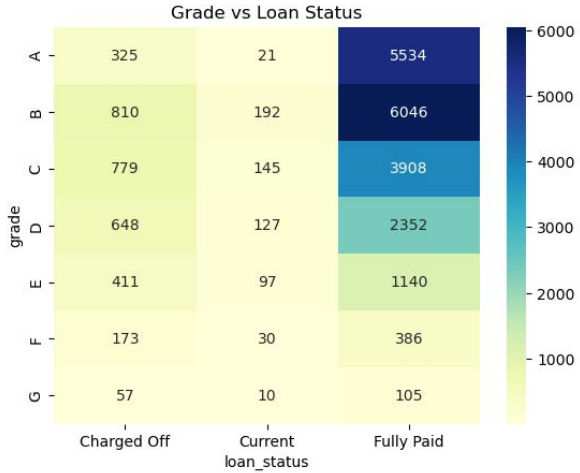


Grouping loan amounts into bins and analyzing them with a scatterplot shows that higher salaries are associated with larger loan amounts.



The scatterplot illustrates that the disbursed loan amounts are consistent across various salary ranges, with lower-salary individuals having minimal chances of securing a loan.

Bivariate Analysis

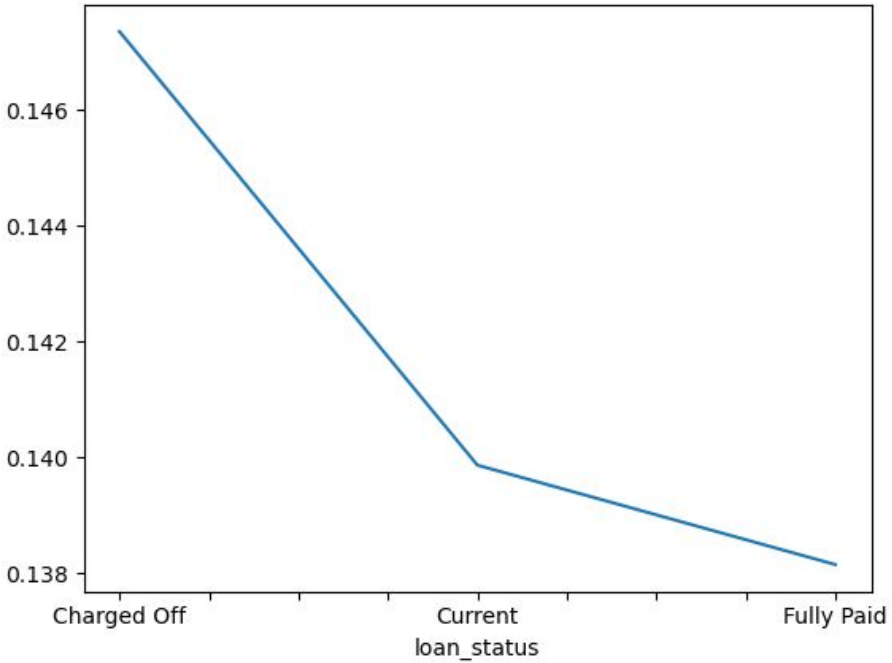


Borrowers with higher grades (A and B) not only secure more loans but also demonstrate a significantly higher number of fully repaid loans. This indicates that individuals with higher grades are more likely to fully repay their loans.

In the charged-off category, most loans are associated with higher grades, suggesting a higher likelihood of repayment.

A significant portion of the charged-off loans is linked to borrowers with home ownership, either through renting or mortgages. This poses a challenge for the organization in terms of loan recovery.

Bivariate Analysis



The charged-off loans have a higher number of "30+ days past-due" instances compared to other loans. This suggests that lending to borrowers who are fully paid and current is recommended for the organization.

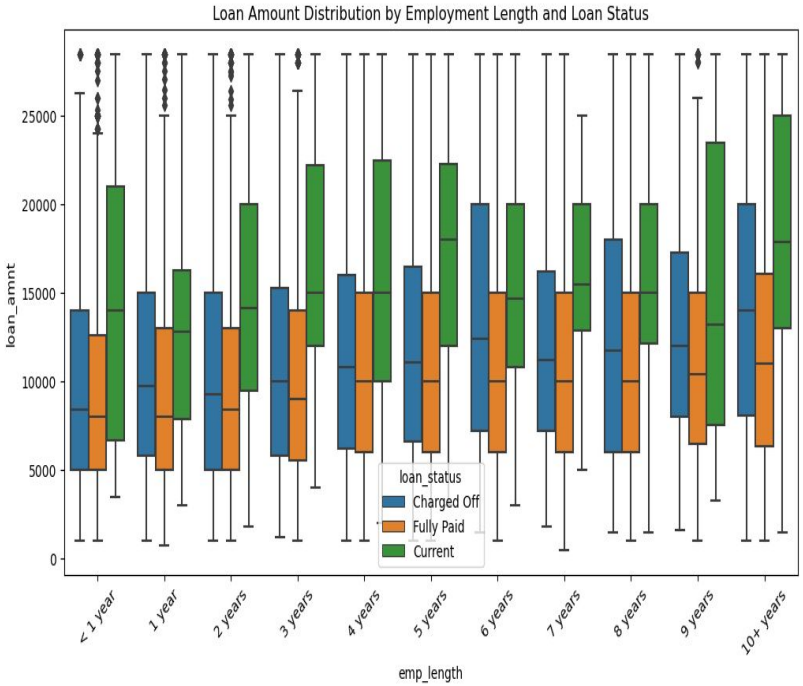
Bivariate Analysis

The graph you provided displays the distribution of loan amounts (loan_amnt) across different employment lengths (emp_length) and is grouped by loan status (loan_status) categories: Charged Off, Fully Paid, and Current. Here's an analysis based on the graph:

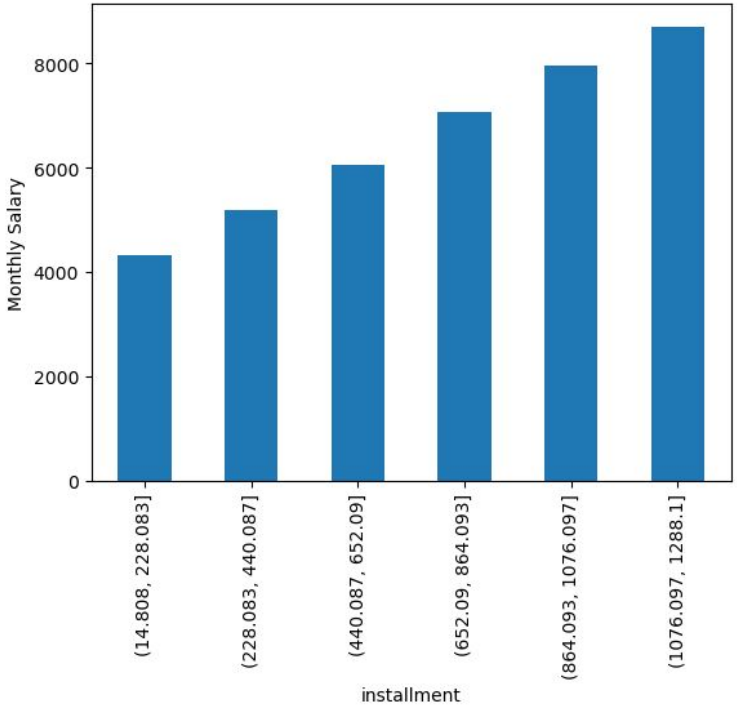
1. **Higher Loan Amounts for Longer Employment**: Borrowers

   with longer employment lengths, particularly those with 10+ years of employment, tend to receive higher loan amounts. The median loan amount for these groups is consistently higher compared to those with shorter employment lengths.

2. **Loan Default Risk**: The distribution shows that borrowers with shorter employment lengths (less than 1 year) have a higher proportion of Charged Off loans compared to those with longer employment lengths. This suggests that shorter employment lengths may be associated with higher default risk.



Derived Metrics



First, we will add a "monthly salary" field to the dataset to facilitate the calculation of the MI (monthly installment) ratio.

Individuals with higher monthly salaries tend to have larger installments. This new "monthly salary" column will assist lenders in evaluating and providing new loans to borrowers. Additionally, borrowers living in rental homes have a higher average salary compared to their installments, which could be a positive indicator for lenders when considering loan approvals.

Thank you