Lending Club: An EDA Case Study

Understanding Risk Analytics in Banking and Financial Services

GROUP MEMBERS

Md Sariq Sahazada

Vijay Mishra

Date: 8th May, 2024

Problem Statement:

- **Goal:** Analyze loan data from a Lending Club to identify patterns associated with loan defaults.
- **Objective:** Uncover key consumer and loan characteristics that influence a borrower's likelihood of repaying a loan.
- **Benefit:** This knowledge will refine the company's risk assessment strategies, leading to reduced credit losses.
- **Approach:** Utilize Exploratory Data Analysis (EDA) techniques to uncover hidden patterns within the loan data.
- **Focus:** Analyze the impact of various variable types (categorical, numerical) on loan default risk.
- **Outcome:** Build a robust system for predicting loan defaults, enabling better loan approval decisions.

Data Understanding

- **Data Shape:** The dataset has a dimension of 111 columns and 39,717 rows.
- **Missing Values:** We identified 55 columns with 100% missing values and 3 columns with over 30% missing values. These columns were removed for efficiency.
- **Loan Status Focus:** To focus on the core analysis of loan defaults, the data was filtered to include only loans categorized as "Fully Paid" or "Charged Off." This eliminates "Current" loans as they haven't reached a definitive outcome yet.

Data Segmentation

In our dataset, we segment columns based on their data type or characteristic. This segmentation helps us organize and analyze the data effectively. Here's how we categorize our columns:

- **Numerical Columns:** These columns consist of quantitative data such as loan amount, interest rate, income, and debt-to-income ratio.
- **Categorical Columns:** These columns represent qualitative data with discrete categories, including loan term, grade, employment length, and loan status.
- **Date Columns:** These columns contain temporal data, including issue date, last payment date, and earliest credit line.
- **Additional Columns:** Some columns, such as unique identifiers and irrelevant features, are excluded from analysis to focus on relevant insights.

Data Preprocessing

- **Preprocessing revol_util (Credit Card Utilization Ratio):** The revol_util column contains percentages, indicated by the "%" symbol. This symbol has been removed for proper numerical analysis.
- **Preprocessing emp_length (Employment Length):** The emp_length column contain values with suffixes like "years" or "year". These suffixes has been removed for consistent interpretation.

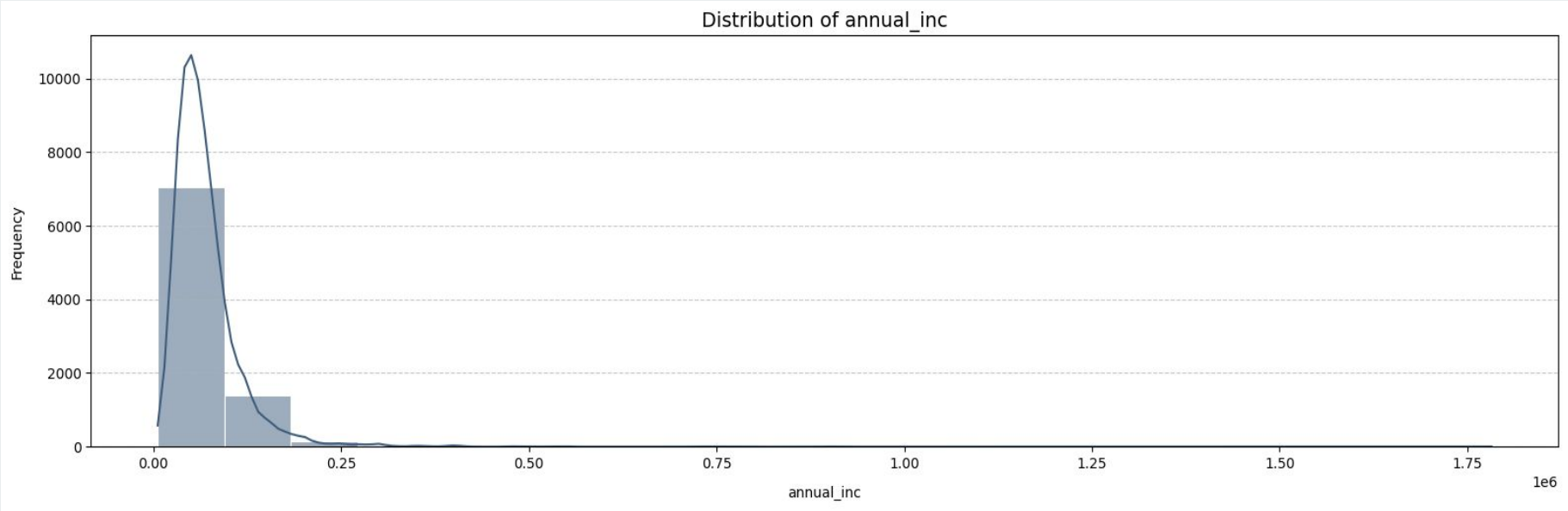
Note: While these examples highlight two specific columns, it's important to note that similar preprocessing steps have been applied to other columns as well. However, to maintain brevity and focus, only these two columns are discussed in detail in this presentation.

Handling Missing and Duplicate Values

Ensuring the integrity and completeness of our dataset is paramount. One crucial step involves addressing missing values, which could potentially skew our analysis or modeling efforts.

- For numerical attributes such as 'revol_util', we adopted a robust approach by replacing missing values with the median of the respective column. Why median? The median is less sensitive to outliers compared to the mean, making it a more reliable measure of central tendency, especially in datasets with skewed distributions.
- Categorical attributes like 'pub_rec_bankruptcies' and 'emp_length' required a different approach due to their discrete nature.
- In such cases, we opted to fill missing values with the mode of the respective column. The mode represents the most frequent value in the dataset and is a suitable choice for preserving the categorical distribution.
- We diligently checked for duplicate entries, ensuring each record was unique. The absence of duplicates enhances the reliability of our analysis and prevents skewed insights.

Univariate Analysis : Annual Income Distribution

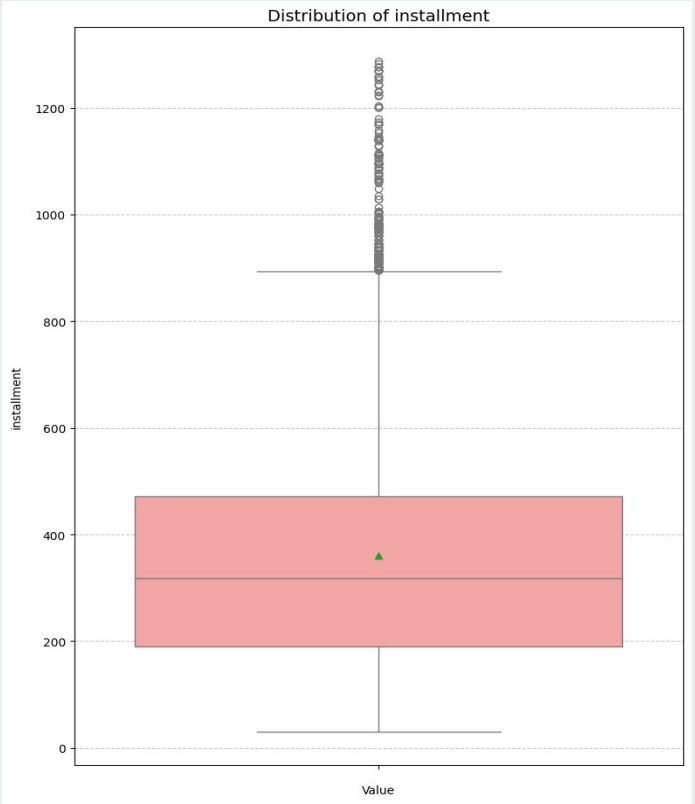


Univariate Analysis : Annual Income Distribution

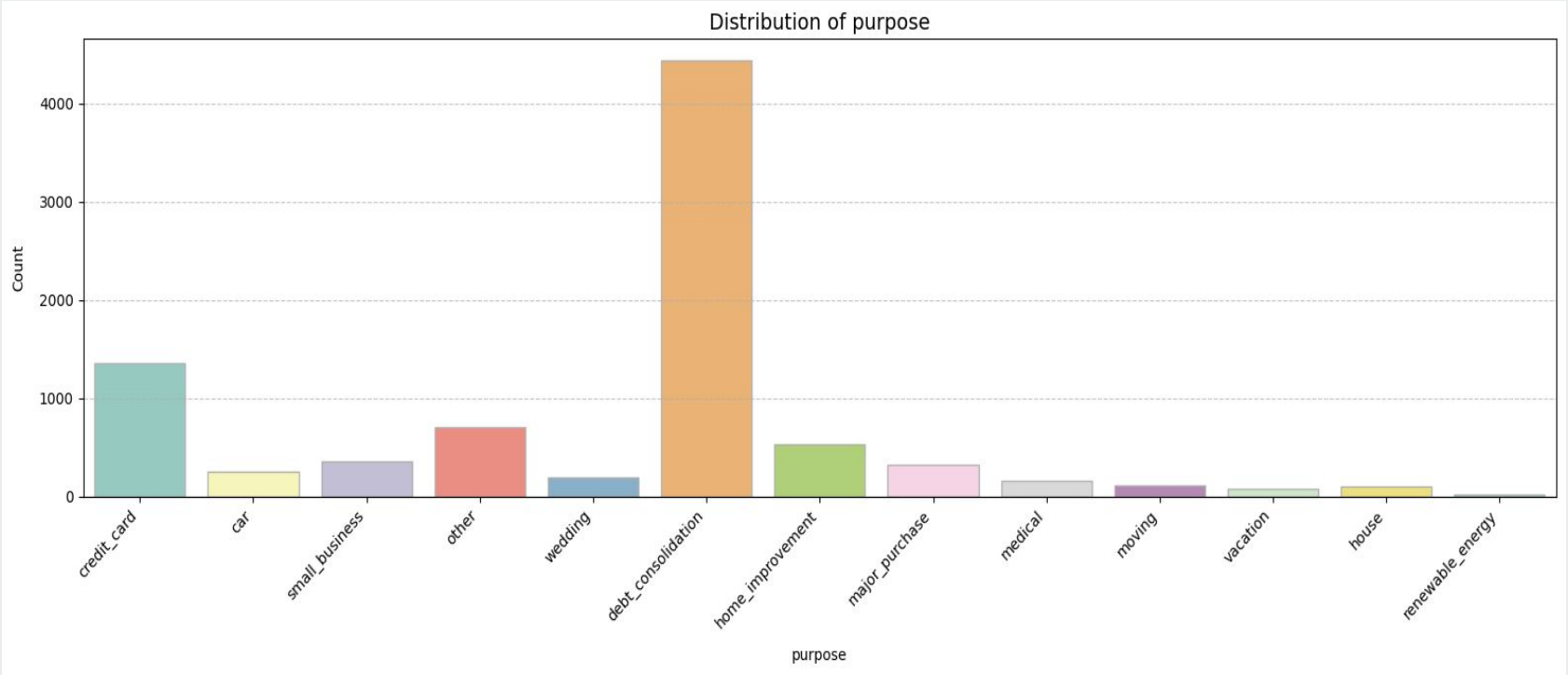
- **Observation:** Majority of annual incomes fall within the range of $40,000 to $60,000.
- **Analysis:**
- Utilizing a histogram plot, we observe the distribution of annual incomes.
- The peak of the distribution occurs in the $40,000 to $60,000 range, indicating that a significant portion of applicants fall within this income bracket.
- **Implications:**
- Understanding the distribution of annual incomes helps in identifying the income levels of loan applicants.
- This insight can be crucial in assessing the financial stability and repayment capacity of borrowers.

Univariate Analysis : Loan Installment Distribution

- Insight: Majority of loan installments were taken for a duration ranging from 200 to 400 months.
- The boxplot analysis of loan installments reveals that the most common duration for loan repayment falls within the range of 200 to 400 months.
- Understanding the distribution of loan installment durations can provide valuable insights into customer borrowing behavior and help in optimizing loan products and repayment strategies.



Univariate Analysis : Loan Purpose Distribution



Univariate Analysis : Loan Purpose Distribution

Insight: Majority of loans were taken for the purpose of debt consolidation.

Analysis:

The analysis of loan purposes reveals that the highest number of loans were used for debt consolidation, indicating a common financial need among borrowers.

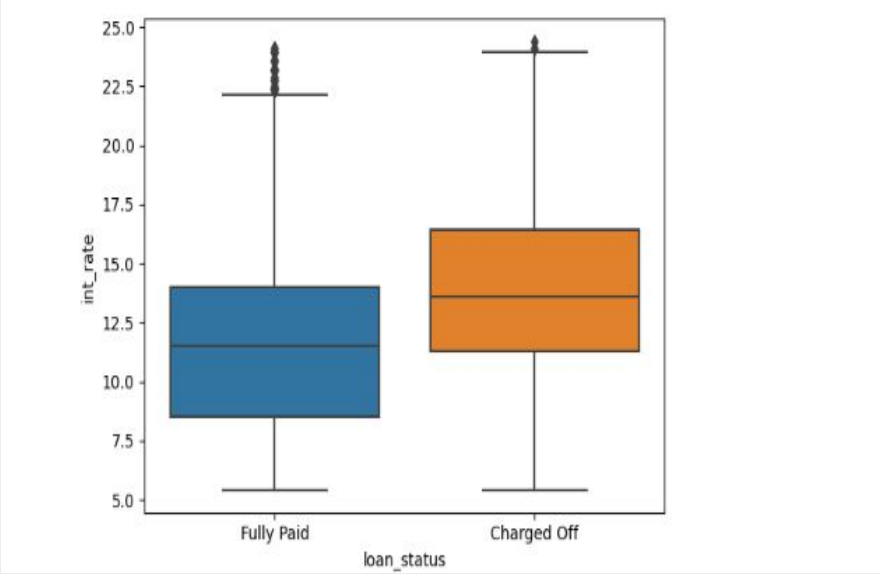
Bivariate Analysis : Defaulters and Interest Rates

Insight:

Defaulters tend to take loans at higher interest rates compared to non-defaulters.

Analysis:

The bivariate analysis of loan defaulters and interest rates indicates a positive correlation between default status and higher interest rates.

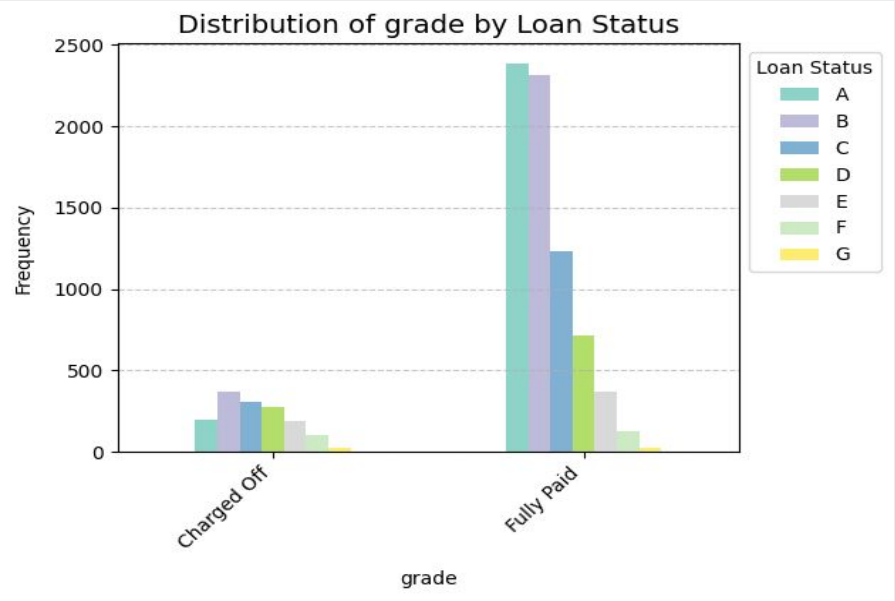


Bivariate Analysis : Loan Repayment by Member Grade

- **Insight:** Majority of borrowers who fully paid their loans belong to Grade A and Grade Majority of defaulters belong to Grade C and Grade D.

- **Analysis:**
- The analysis of loan repayment by member grade reveals distinct patterns based on borrower grades.
- Borrowers with higher grades (A and B) show a higher tendency to fully pay their loans.
- Conversely, borrowers with lower grades (C and D) are more likely to default on their loans.

- **Implication:**
- This observation suggests that borrower grade is a significant factor influencing loan repayment behavior.
- Lenders could consider adjusting their lending policies or risk assessment criteria based on borrower grades to minimize default risk.

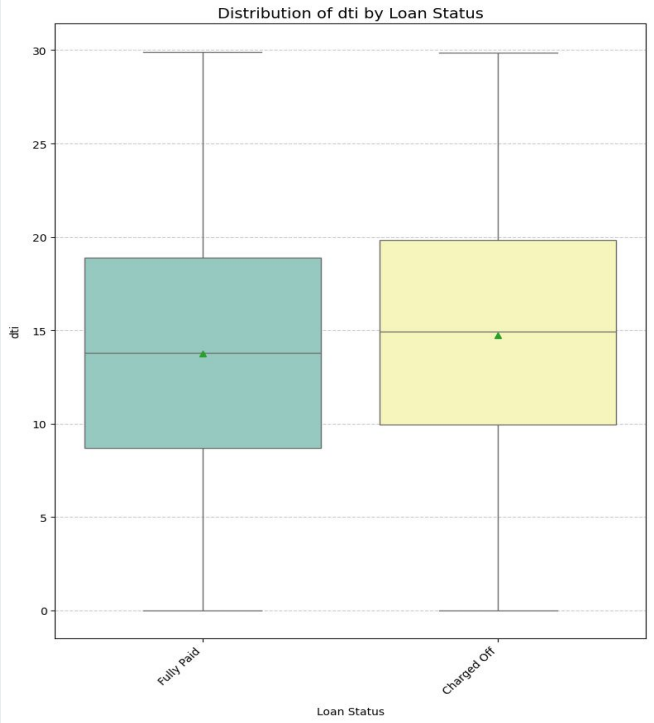


Bivariate Analysis : Debt-to-Income Ratio (DTI) and Loan Repayment

- **Insight:**
- Borrowers who were charged off have a higher Debt-to-Income Ratio (DTI) compared to those who paid the loan fully.

- **Analysis:**
- The analysis of DTI in relation to loan repayment status indicates a clear difference between borrowers who paid their loans fully and those who were charged off.
- Higher DTI may indicate financial strain, making it more challenging for borrowers to repay their loans.

- **Implication:**
- DTI can serve as a predictive factor for loan default risk assessment.
- Lenders may consider incorporating DTI into their risk assessment models to identify borrowers at higher risk of default and adjust lending policies accordingly.

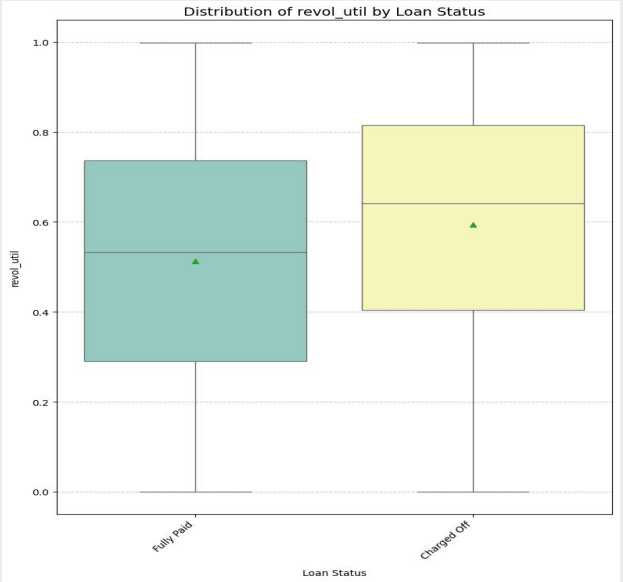


Bivariate Analysis : Revolving Line Utilization Rate and Loan Repayment

- **Insight:**
- A high ratio of revolving line utilization rate (revol_util) suggests financial strain or over-reliance on credit.
- People who were charged off had a higher average revol_util % compared to those who paid the loan fully.

- **Analysis:**
- The analysis of revol_util in relation to loan repayment status reveals a notable difference between borrowers who paid their loans fully and those who were charged off.
- Higher revol_util may indicate a higher level of debt relative to available credit, which could contribute to financial difficulties and loan default.

- **Implication:**
- Revolving line utilization rate can serve as a valuable indicator of financial health and credit risk.
- Lenders may incorporate revol_util into their risk assessment models to better identify borrowers at higher risk of default and tailor lending decisions accordingly.



Conclusion:

- **Interest Rates:** Borrowers who defaulted tended to have higher interest rates. Thus, it's advisable for the company to approve loans at lower interest rates to mitigate default risk.
- **Income Disparity:** The average annual income of defaulters is significantly lower than that of borrowers who made full payments. This suggests that income level is a critical factor in loan repayment capability.
- **Debt-to-Income Ratio (DTI):** Defaulters generally had higher DTI ratios compared to borrowers who paid the loan fully. Hence, the bank should exercise caution when lending to individuals with higher DTI ratios to minimize default risk.
- **Grade Distribution:** Majority of borrowers who fully paid their loans belong to Grade A, B, and C. For borrowers in lower grades, additional factors should be considered before approving loans.
- **Loan Purpose:** Debt consolidation was the most common purpose among defaulters. The bank should carefully assess the purpose of the loan before extending credit to borrowers.

Key Takeaways:

- **Risk Mitigation:** Lowering interest rates, scrutinizing income levels, and considering loan purposes can reduce default risk.
- **Targeted Approach:** Tailoring lending decisions based on borrower grades and DTI ratios can enhance risk management strategies.
- **Customer-Centric Approach:** By understanding borrower characteristics and behaviors, the bank can better align its lending practices with customer needs and financial capabilities.

Thank You