LENDING CLUB
CASE STUDY

A Data-Driven Approach for Risk Analysis

Submited By:
Shubham Sharma

AGENDA

Overview and EDA of Lending Club loan data to extract insights on borrower behavior, loan characteristics, and enhance risk assessment for better lending decisions.

- Introduction to Lending Club and the business problem.
- Overview of the dataset and key variables.
- Conducting Exploratory Data Analysis (EDA).
- Insights and implications for Lending Club.
- Conclusion
- Acknowledgements.

LENDING CLUB OVERVIEW

Lending Club is a financial technology company that operates an online marketplace lending platform. It connects borrowers seeking personal loans, small business loans, and other credit solutions with individual and institutional investors.

Key aspects of Lending Club's business model:

- Borrowers: It targets prime and near-prime urban borrowers, often for debt consolidation and personal financing.
- Investors: Both individual and institutional investors can fund loans and earn returns based on borrower interest.
- Loan Origination: Lending Club uses data analytics to assess creditworthiness and price risk effectively.
- Revenue: It generates fees from borrowers and investors, without direct lending risk.
- Regulation: Lending Club operates under SEC and state regulatory oversight.

Lending Club's peer-to-peer model aims to provide better rates for borrowers and returns for investors, leveraging technology. Effective risk management is critical to minimize loan defaults and credit losses.

BUSINESS PROBLEM

Based on the introduction provided, the key business problem that Lending Club is facing is managing its loan approval process to minimize credit losses.

The main aspects of the business problem are:

1. Identifying Creditworthy Borrowers
2. Mitigating Credit Losses
3. Balancing Loan Approval and Risk:

The primary objective is to use Exploratory Data Analysis (EDA) on the provided dataset to identify the key variables (or "driver variables") that are strong indicators of loan default. This knowledge can then be leveraged by Lending Club to improve its loan approval process and more effectively manage credit risk in its portfolio.

By addressing this business problem, Lending Club aims to enhance its profitability and sustainability by minimizing credit losses while still serving a broad range of creditworthy borrowers.

DATASET INTRODUCTION

The dataset provided appears to be a comprehensive loan dataset from Lending Club, a prominent peer-to-peer lending platform. The dataset contains detailed information about loan applications, borrower characteristics, and loan performance metrics.

Some key aspects of the dataset:

1. Borrower Demographics: The dataset includes various borrower attributes such as annual income, employment length, home ownership status, credit score range, and more. These variables provide insights into the borrower profile.

2. Loan Characteristics: The dataset contains information about the loan itself, including the loan amount, interest rate, loan term (36 or 60 months), loan status, and purpose of the loan (e.g., debt consolidation, home improvement).

3. Credit and Financial History: There are numerous variables related to the borrower's credit history, including the number of delinquencies, public records, credit inquiries, open accounts, credit utilization, and more. These variables can be used to assess the borrower's creditworthiness.

4. Repayment Performance: The dataset includes metrics related to loan repayment, such as the number of payments received, amount of principal and interest paid, and whether the loan has been charged off or is in collections.

DATASET KEY VARIABLES

- loan_amnt: The listed amount of the loan
- int_rate: The interest rate on the loan
- term: The number of payments on the loan (36 or 60 months)
- annual_inc: The borrower's self-reported annual income
- dti: The borrower's debt-to-income ratio
- loan_status: The current status of the loan (e.g., Fully Paid, Charged Off)
- total_pymnt: The total payments received on the loan
- total_rec_prncp: The total principal received on the loan
- grade: LC assigned loan grade
- sub_grade: LC assigned loan subgrade
- emp_length: Employment length in years
- home_ownership: The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER

all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_credit_pull_d	The most recent month LC pulled credit for this loan
last_fico_range_high	The upper boundary range the borrower's last FICO pulled belongs to.
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
max_bal_bc	Maximum current balance owed on all revolving accounts
member_id	A unique LC assigned Id for the borrower member.
mo_sin_old_il_acct	Months since oldest bank installment account opened

THIS DATASET PROVIDES A RICH SOURCE OF INFORMATION TO ANALYZE BORROWER CHARACTERISTICS, LOAN DYNAMICS, AND LOAN PERFORMANCE, WHICH CAN BE VALUABLE FOR LENDING CLUB'S CREDIT RISK MANAGEMENT AND PORTFOLIO OPTIMIZATION EFFORTS.

DATA CLEANING AND PRE-PROCESSING

1. Loading Data:
   - Loaded the loan dataset from the CSV file
   - Handled mixed data types in some variables
2. Handling Missing Values:
   - Identified columns with high percentage of missing values (>65%)
   - Dropped columns with excessive missing data
3. Unique Value Analysis:
   - Identified columns with only a single unique value
   - Dropped these columns as they do not provide meaningful insights
4. Duplicate Rows:
   - Checked for and found no duplicate rows in the dataset
5. Dropping Records:
   - Removed rows where the loan status was "Current" as they do not provide insights on default
6. Data Conversion:
   - Converted relevant columns to appropriate data types (e.g., float, datetime)

OUTLIER TREATMENT AND MISSING VALUE IMPUTATION

- **Removing the outliers**

Removing outliers is a critical step in data preprocessing aimed at improving the quality of the dataset and the robustness of subsequent analyses. Outliers are data points that deviate significantly from the rest of the data, potentially skewing results and leading to misleading conclusions.

DATA VISUALIZATION AND ANALYSIS

The analysis effectively addresses the right problem, aligning with the business needs, and is structured clearly for easy understanding. Univariate and segmented univariate analyses are conducted accurately, with realistic assumptions applied as needed. This approach successfully identifies at least five key driver variables that strongly indicate loan defaults.

1

Univariate Analysis:
The primary goal is to understand the distribution, central tendency, and spread of the data for that variable

2

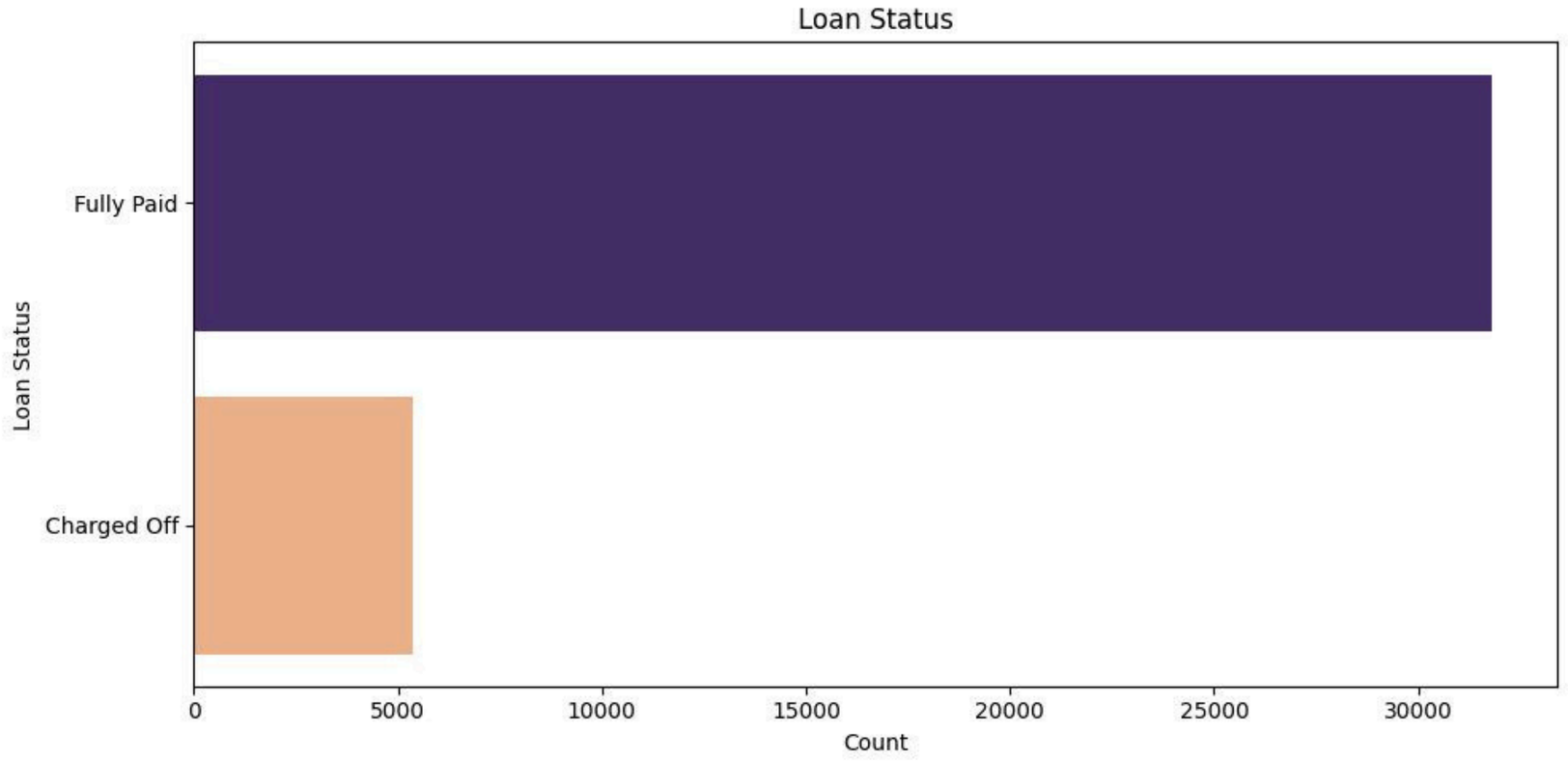
Segmented Univariate Analysis:
Segmented Univariate Analysis involves examining loan characteristics within specific segments, such as loan grades, employment length, or income levels.

3

Bivariate Analysis:
This involves examining the relationship between two variables to uncover any patterns or associations

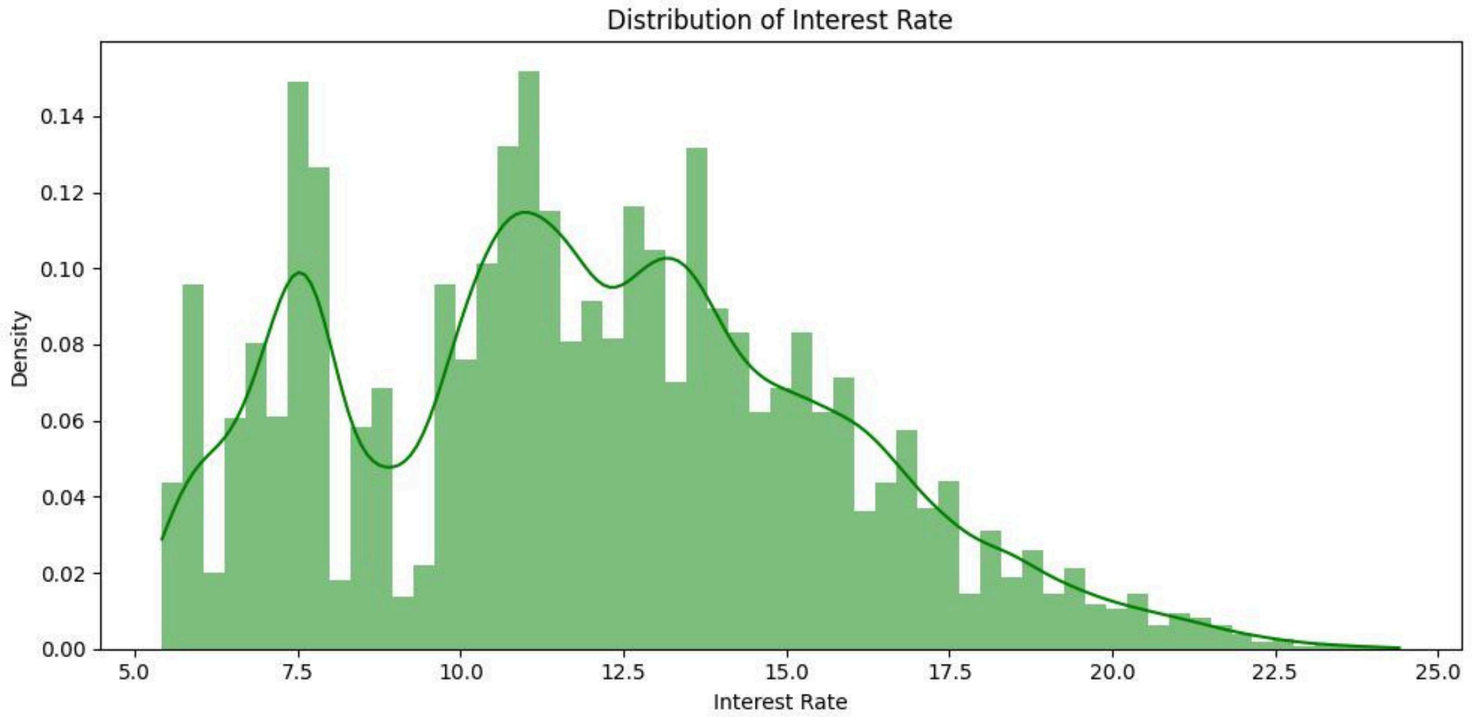
UNIVARIATE ANALYSIS

Observation: The majority of loans are small, with amounts varying from 500to35,000 and a median of $10,000. Although larger loans are less common, they have a higher likelihood of default. Despite this, the overall number of defaulted loans remains significantly lower compared to the number of fully paid loans.



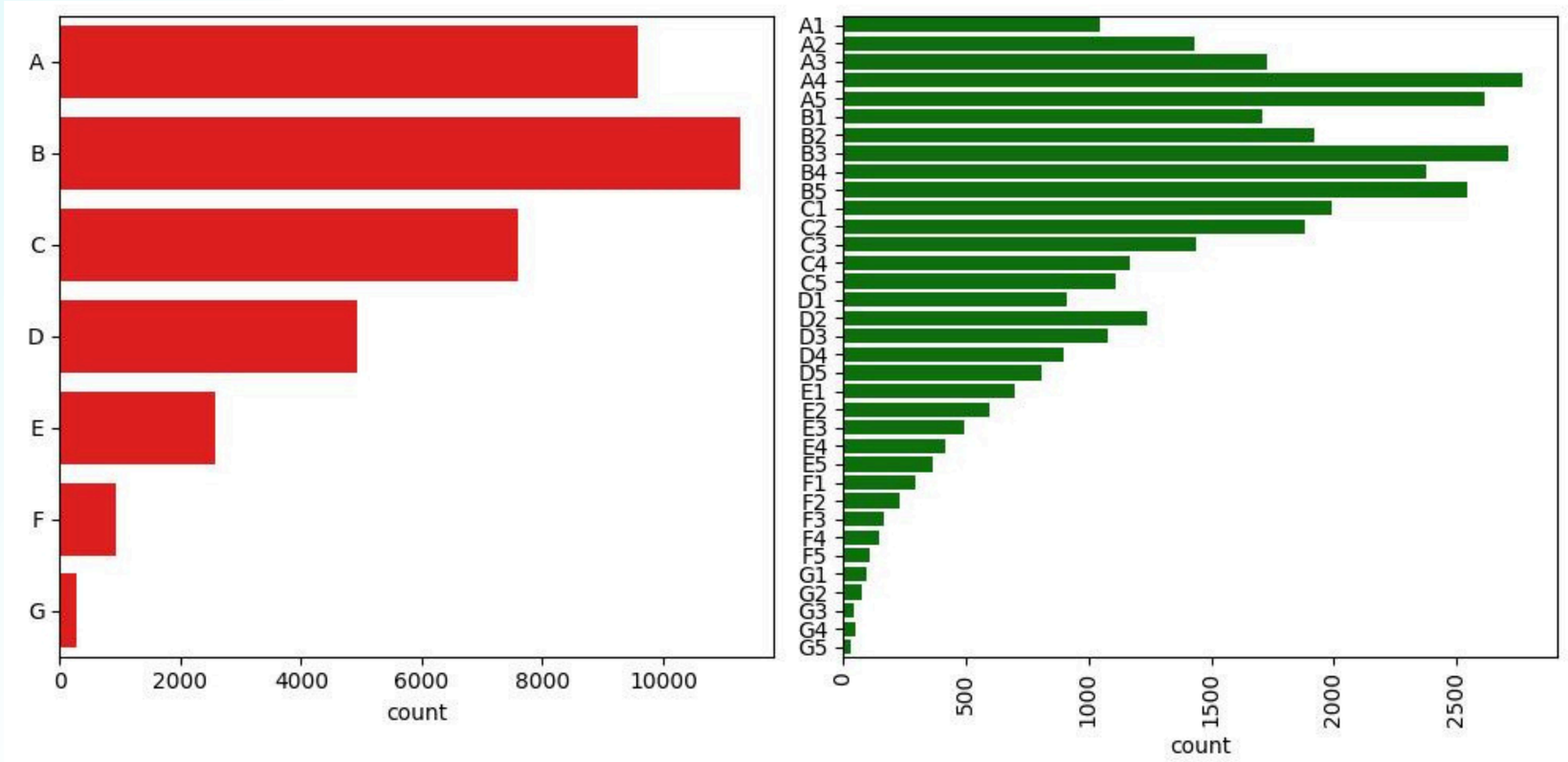
UNIVARIATE ANALYSIS

Observation: The interest rate is more crowded around 5-10 and 10-15 with a drop near 10.



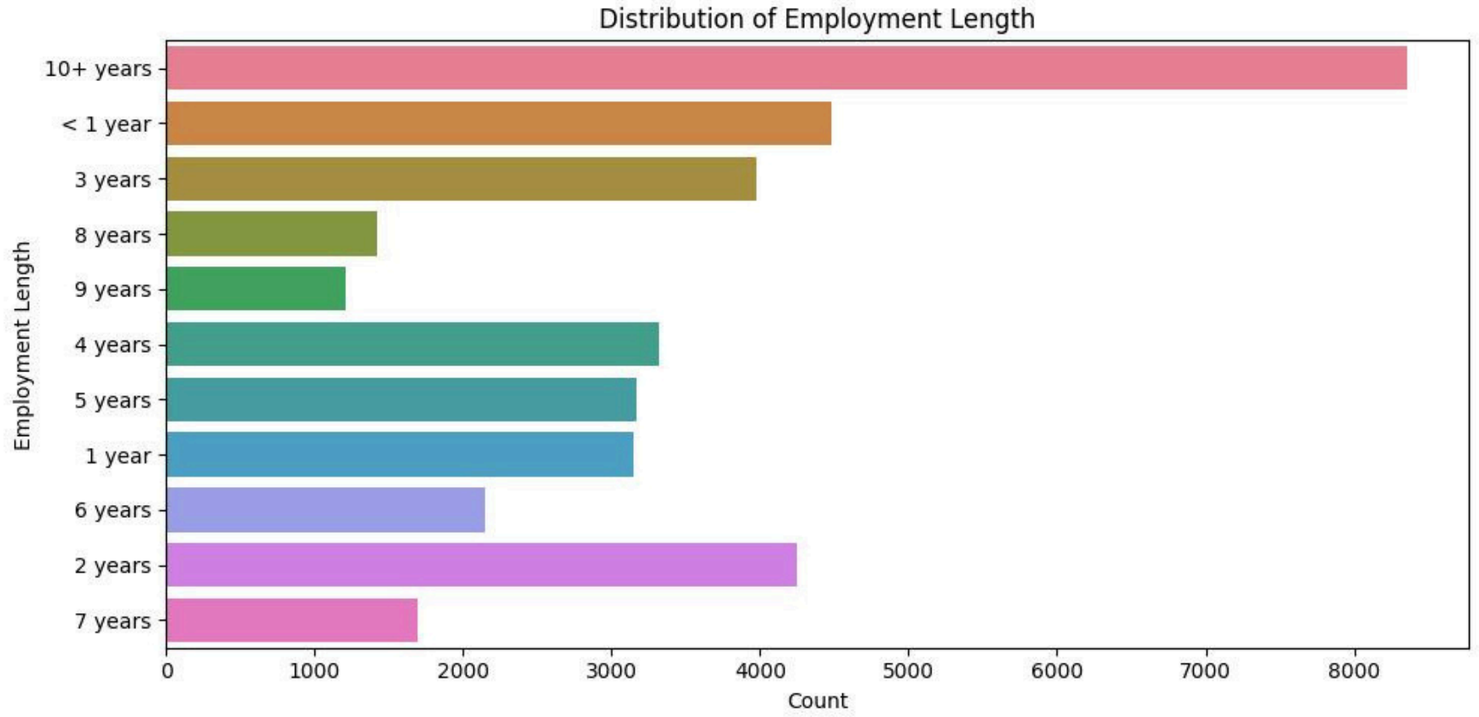
UNIVARIATE ANALYSIS

Observation: A significant number of loans fall under grades 'A' and 'B', particularly in the lower subgrades, indicating that most loans are of high quality. This pattern is consistent with the overall grade distribution.



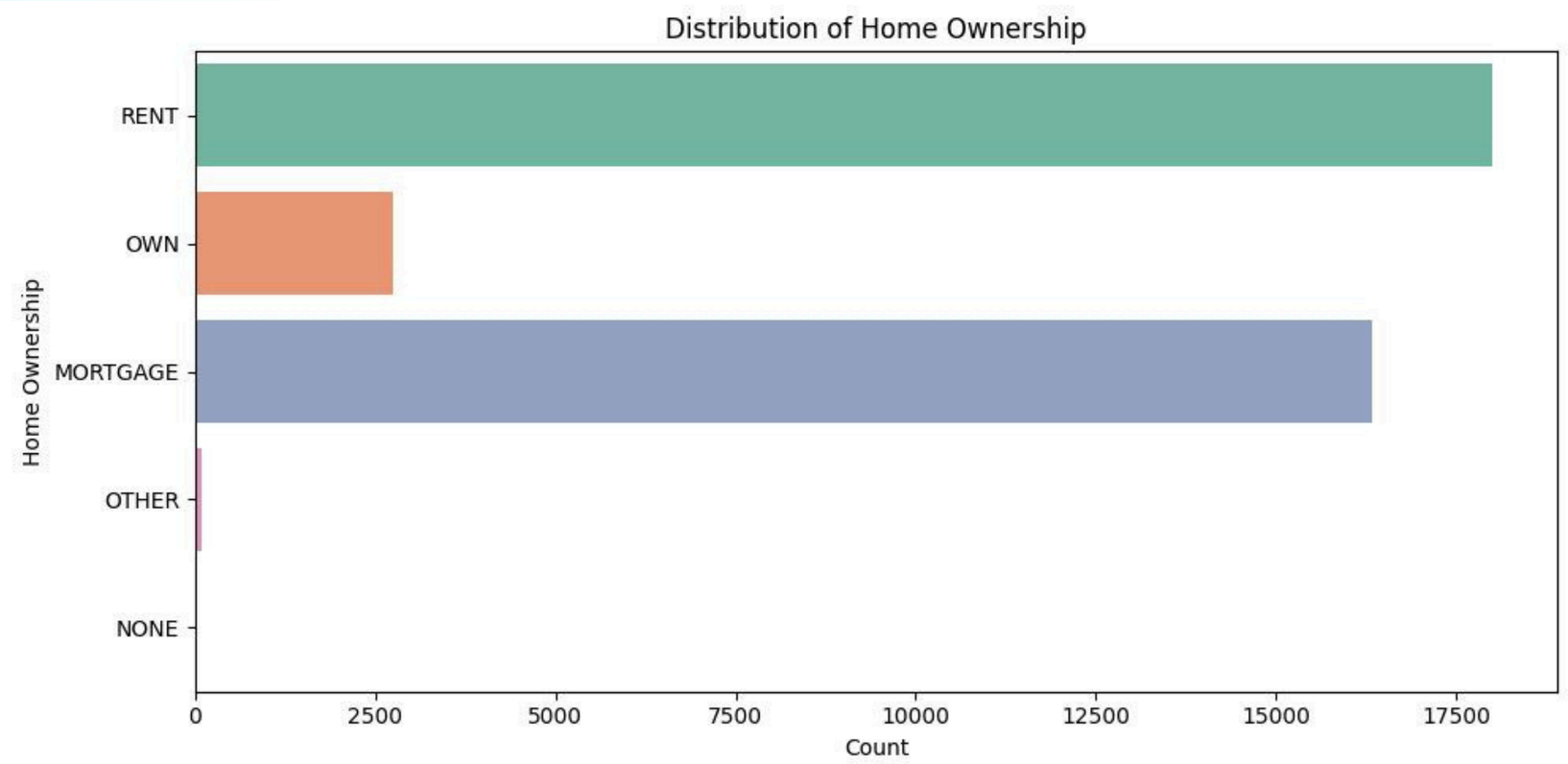
UNIVARIATE ANALYSIS

Observation: Majority of borrowsers have working experience greater than 10 years.



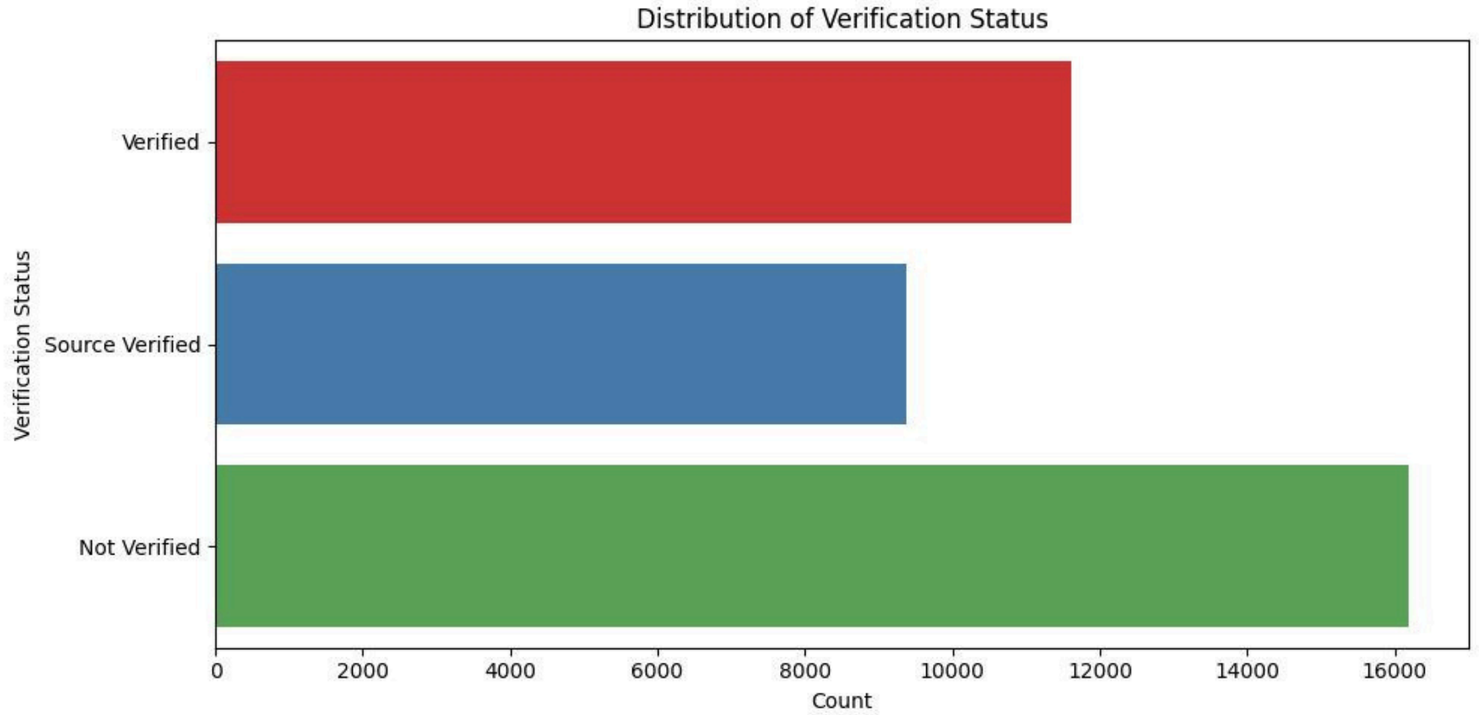
UNIVARIATE ANALYSIS

Observation: Majority of borrowsers don't posses property and are on mortage or rent.



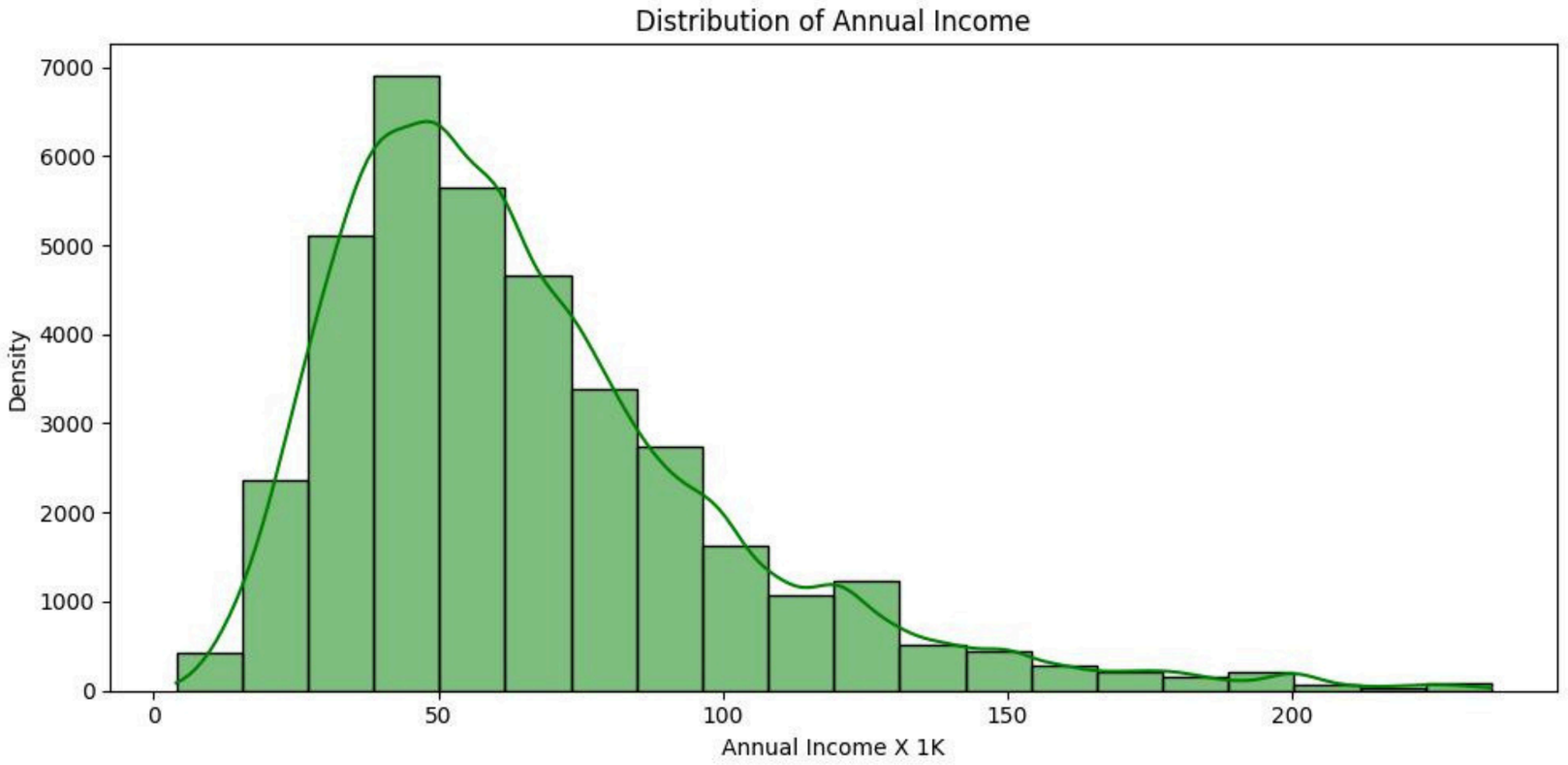
UNIVARIATE ANALYSIS

Observation: About 57% of the borrowers are verified by the company or have source verified.



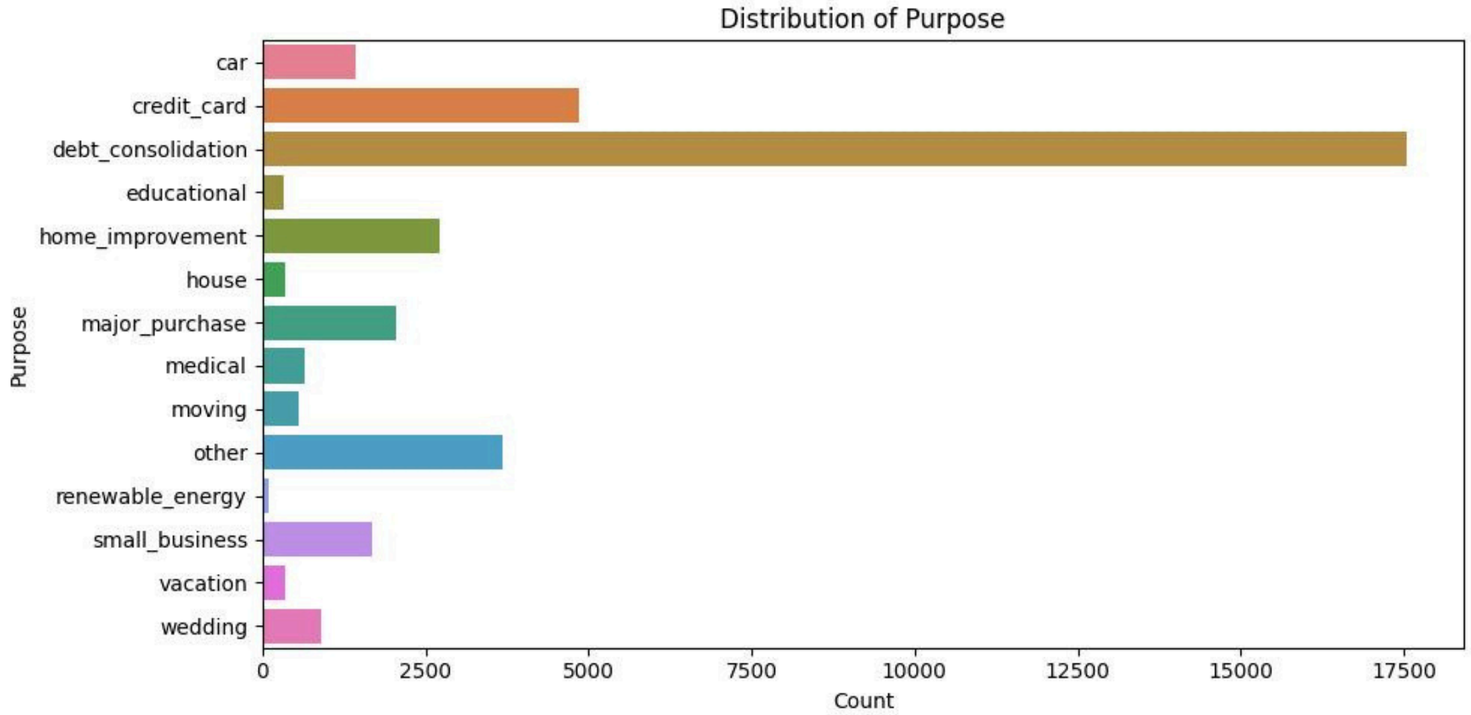
UNIVARIATE ANALYSIS

Observation: Annual Income shows left skewed normal distribution thus we can say that the majority of burrowers have income in range 25K-100K



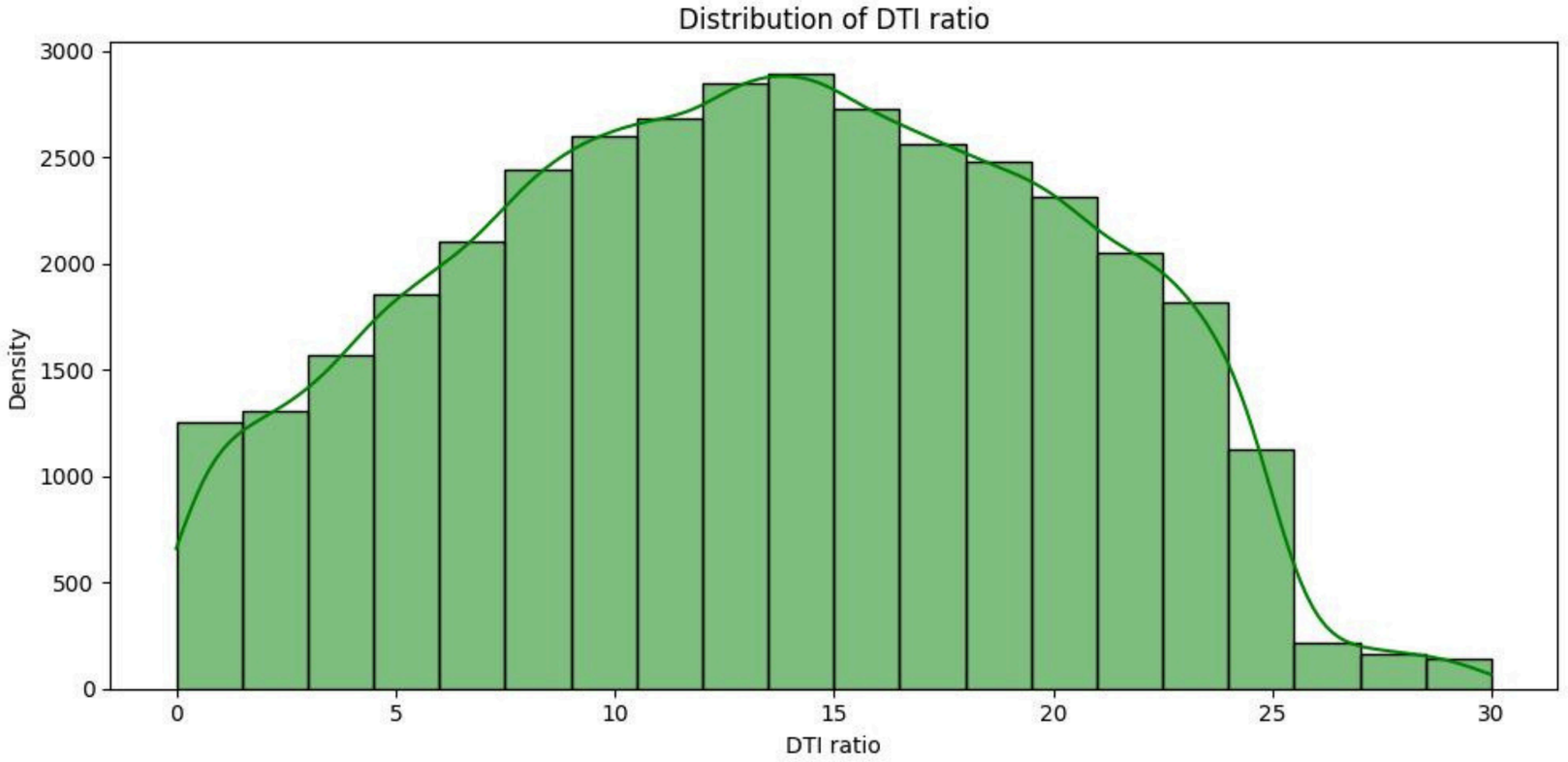
UNIVARIATE ANALYSIS

Observation: A large percentage of loans are taken for debt consolidation followed by credit card.



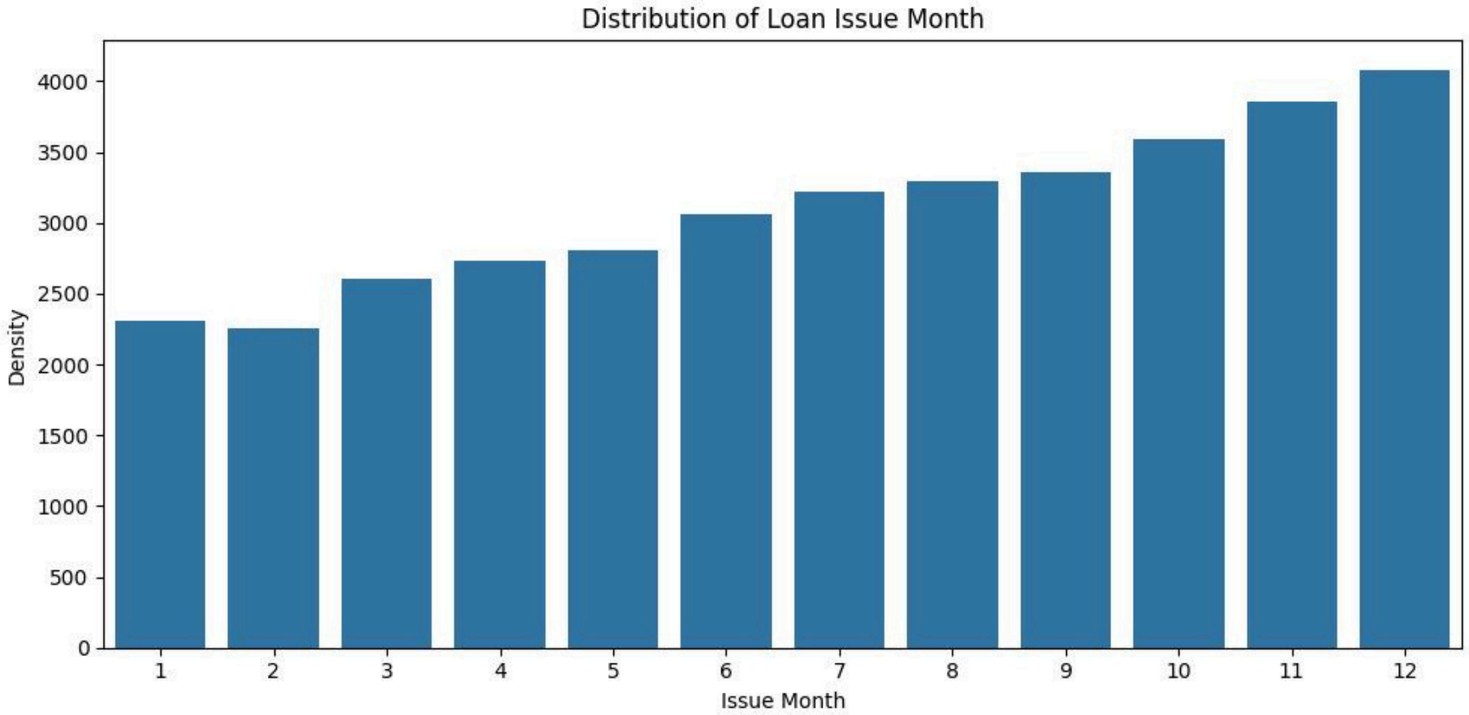
UNIVARIATE ANALYSIS

Observation: Majority of the borrowers have very large debt compared to the income registerd, concentrated in the 10-15 DTI ratio.



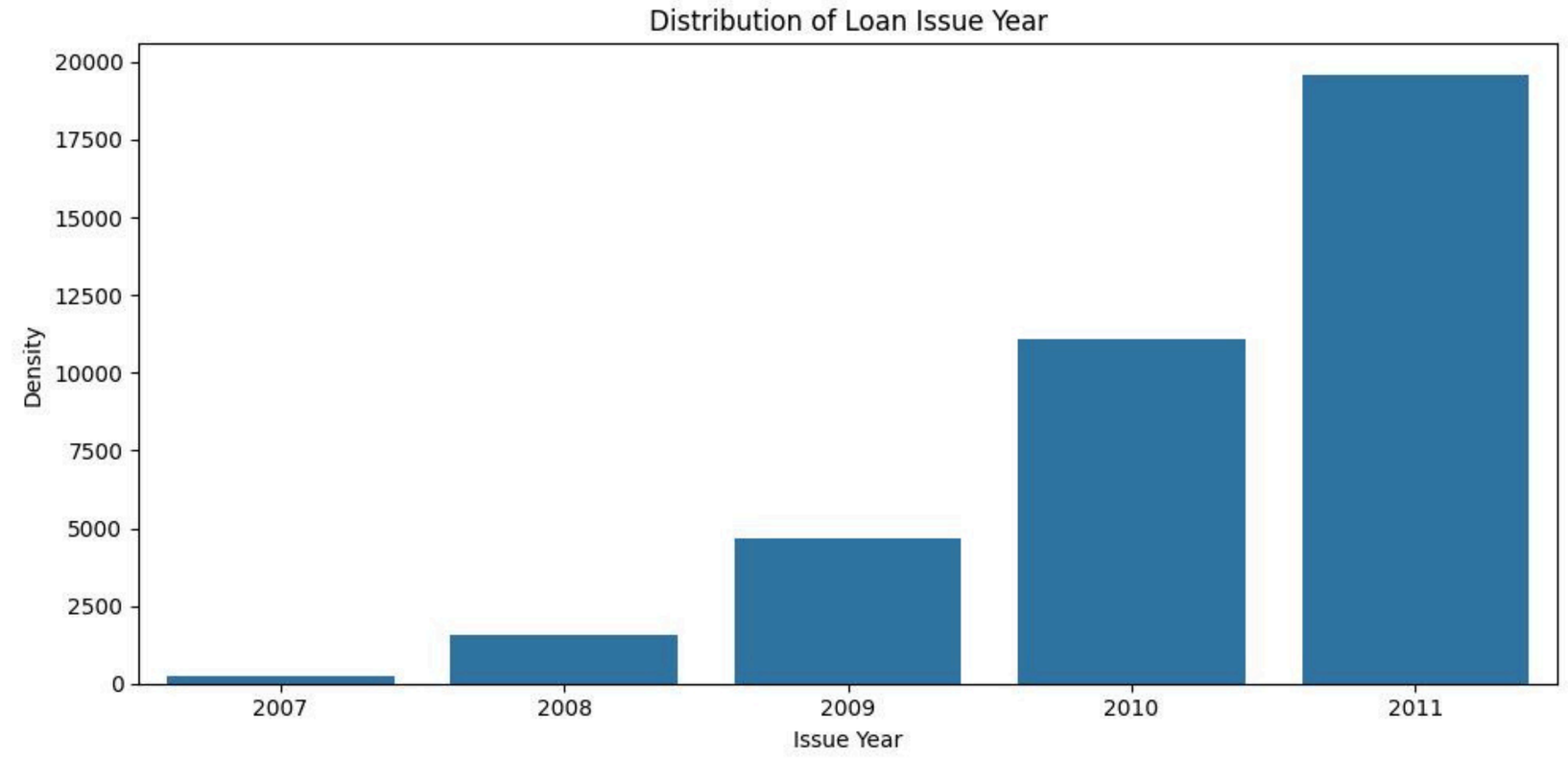
UNIVARIATE ANALYSIS

Obsservation: Majority of the loans are given in last quarter of the year.

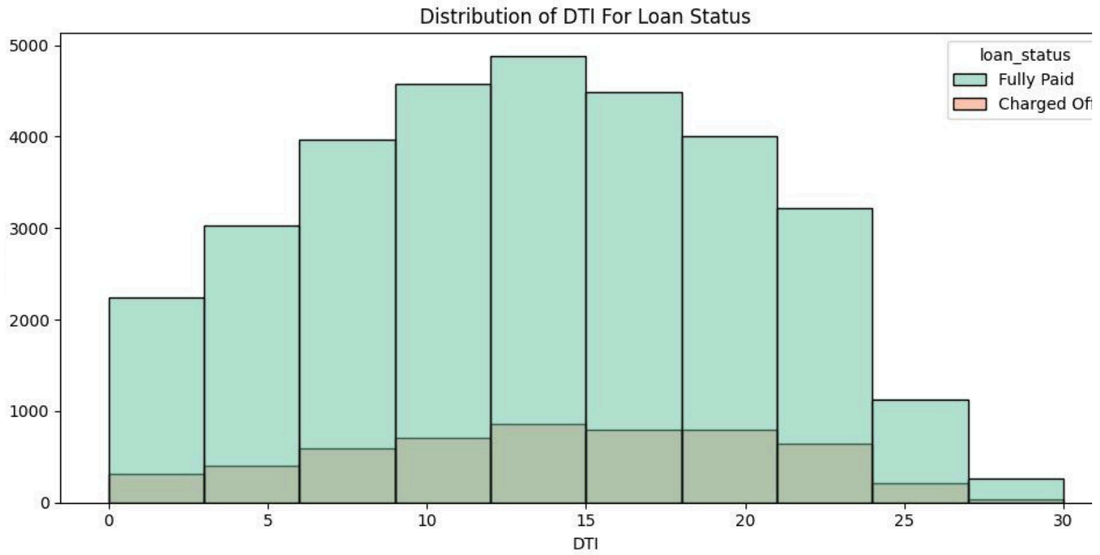


UNIVARIATE ANALYSIS

Observation: The number of loans approved increases with the time at expontential rate, thus we can say that the loan approval rate is increasing with the time.

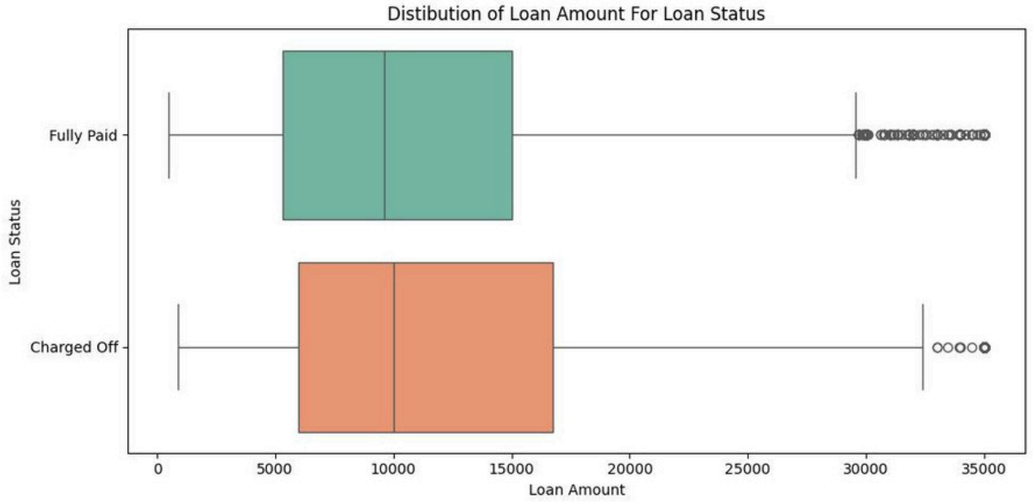


SEGMENTED UNIVARIATE ANALYSIS



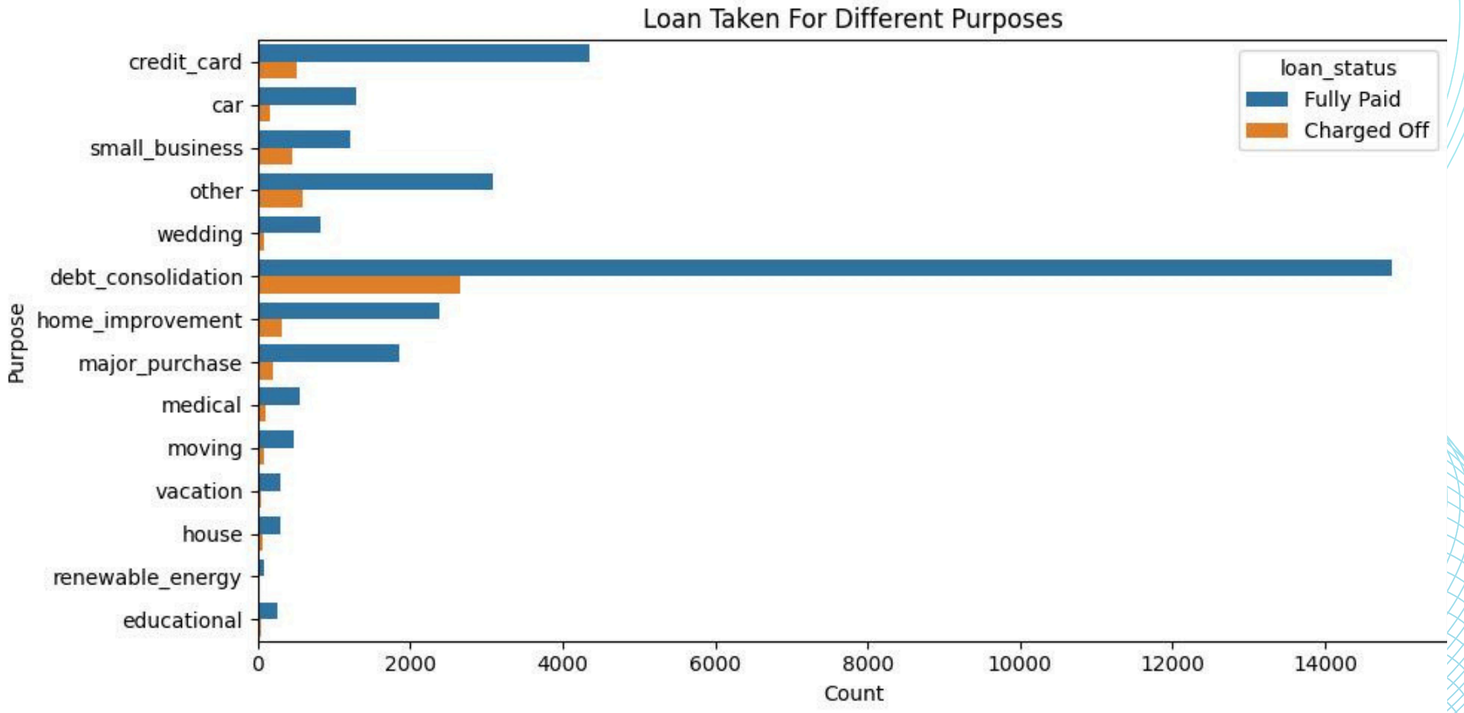
Observation:

- The Loan Status varies with DTI ratio, we can see that the loans in DTI ratio 10-15 have higher number of defaulted loan but higher dti has higher chance of defaulting.
- The mean and 25% percentile are same for both but we see larger 75% percentile in the defaulted loan which indicate large amount of loan has higher chance of defaulting.



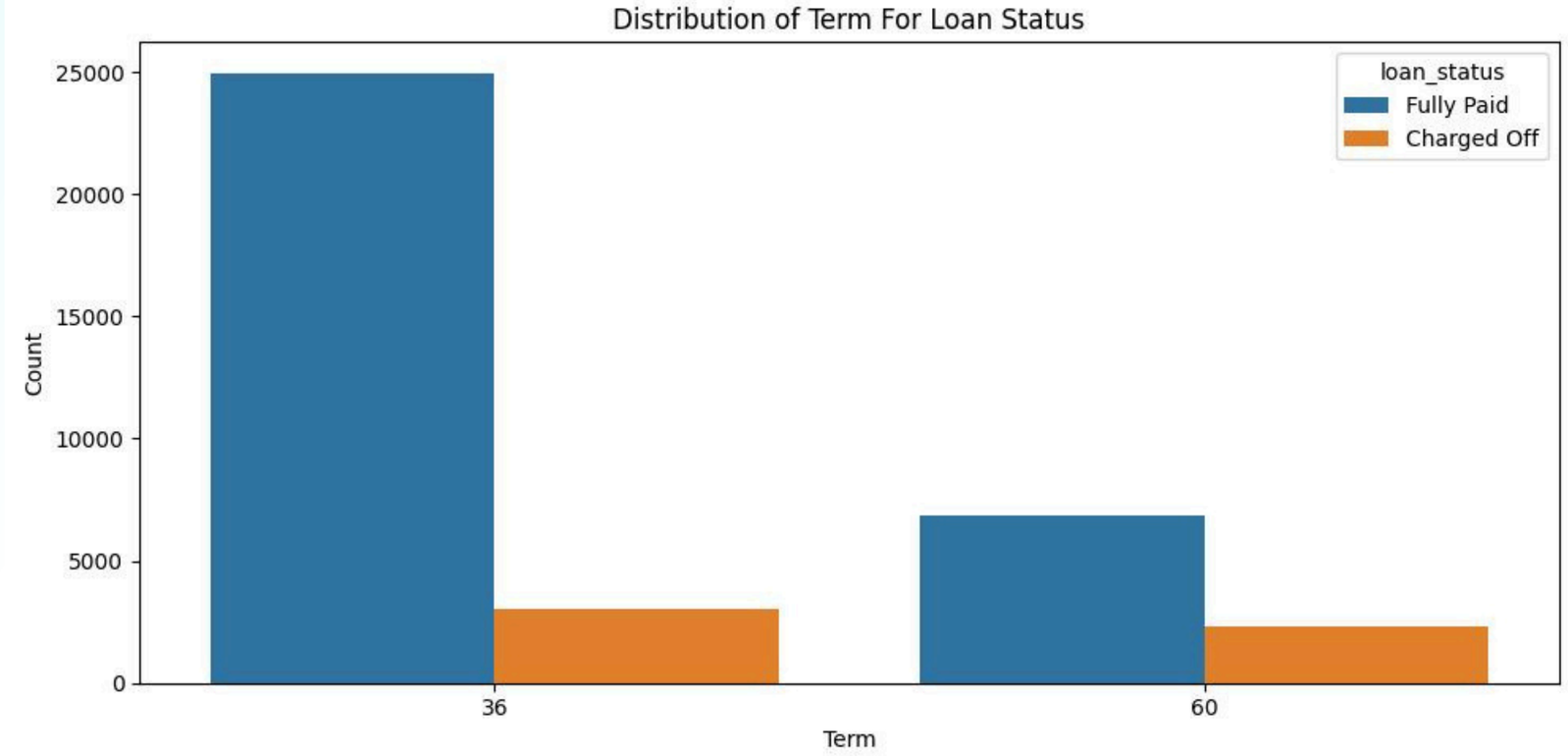
SEGMENTED UNIVARIATE ANALYSIS

Observation: Debt Consolidation is the most popular loan purpose and has highest number of fully paid loan and defaulted loan.

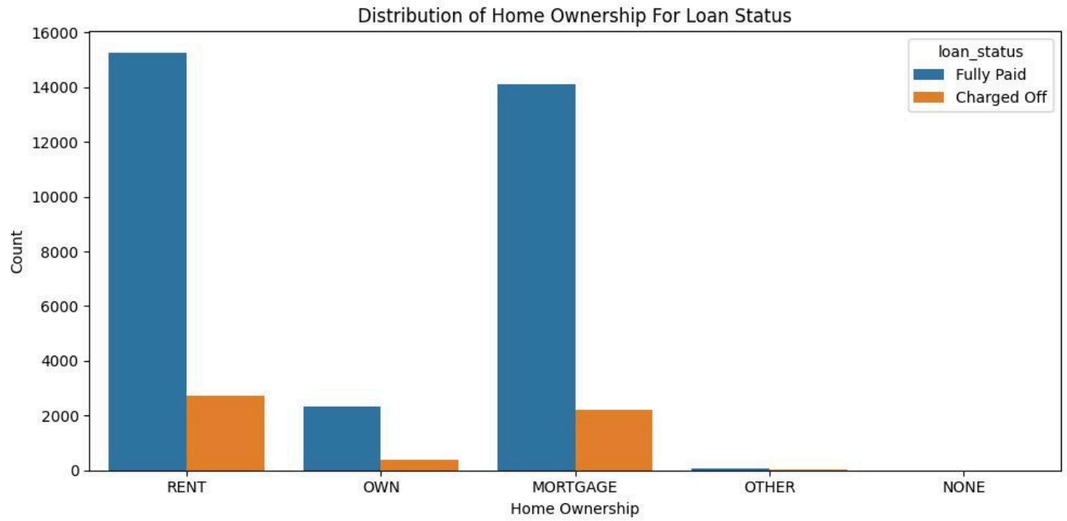


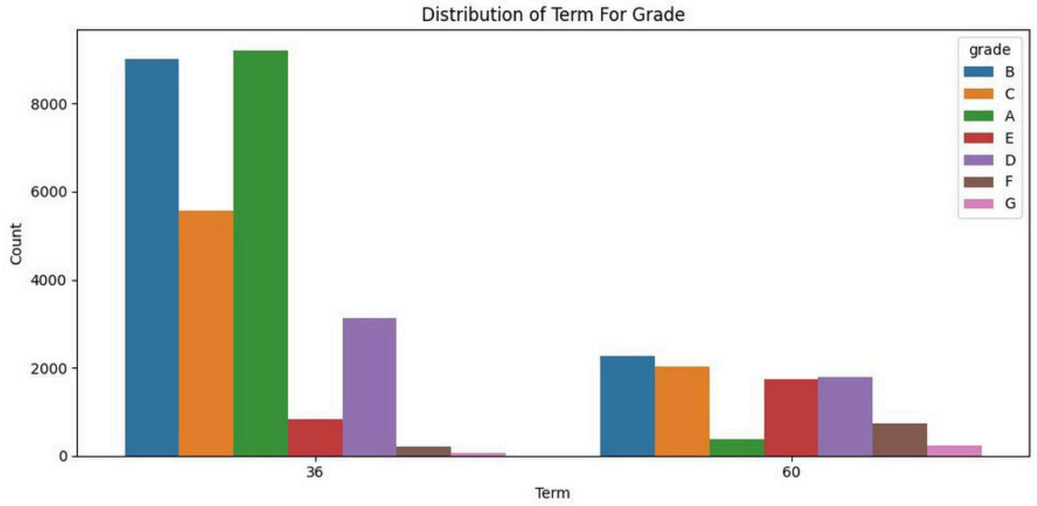
SEGMENTED UNIVARIATE ANALYSIS

Observation: The 60 month term has higher chance of defaulting than 36 month term whereas the 36 month term has higher chance of fully paid loan.



SEGMENTED UNIVARIATE ANALYSIS



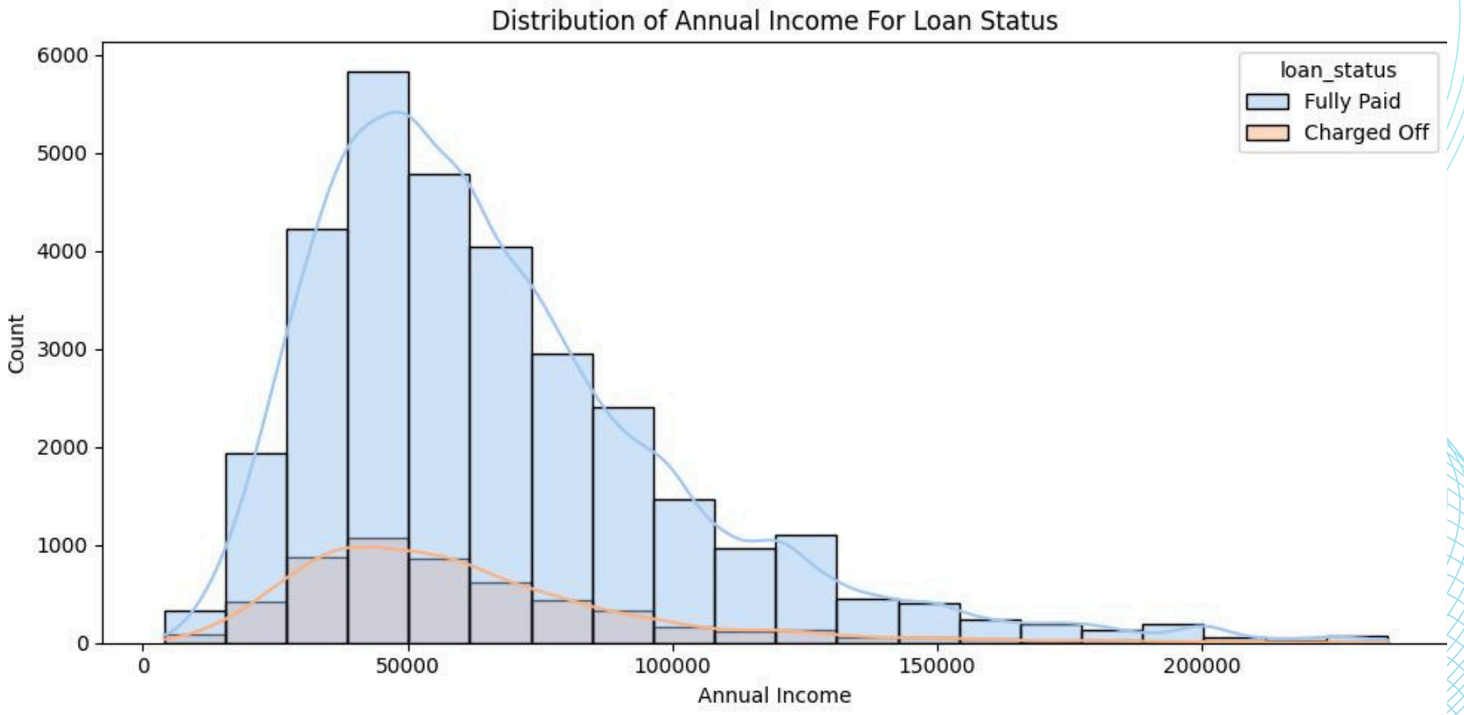


Observation:

1. The Defaulted loan are lower for the burrowers which own their property compared to on mortgage or rent.
2. The loans in 36 month term majorily consist of grade A and B loans whereas the loans in 60 month term mostly consist of grade B, C and D loans.

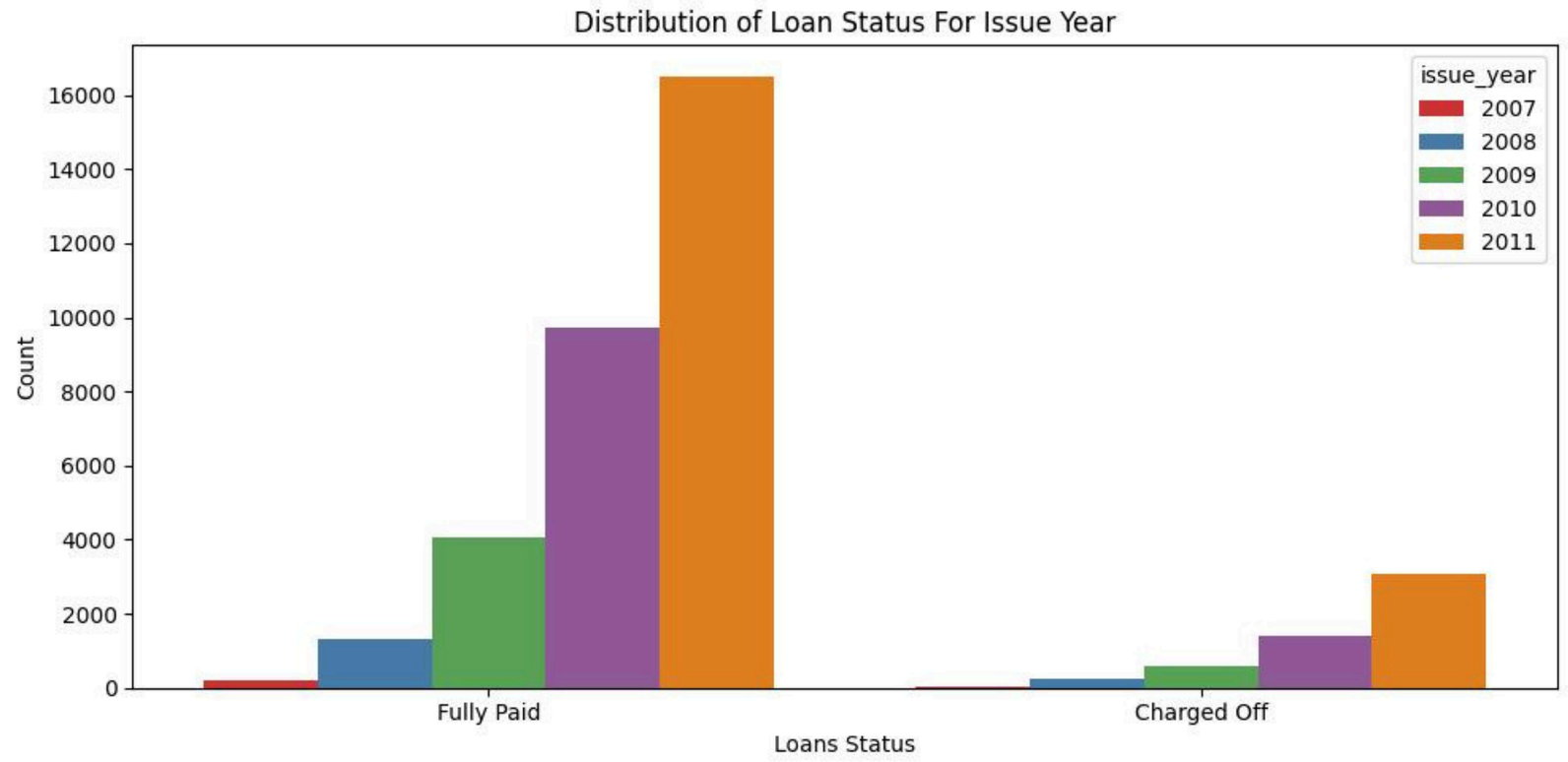
SEGMENTED UNIVARIATE ANALYSIS

Observation: Borrowers with an annual income below $50,000 are more likely to default on their loans, while those with higher annual incomes are less likely to default.

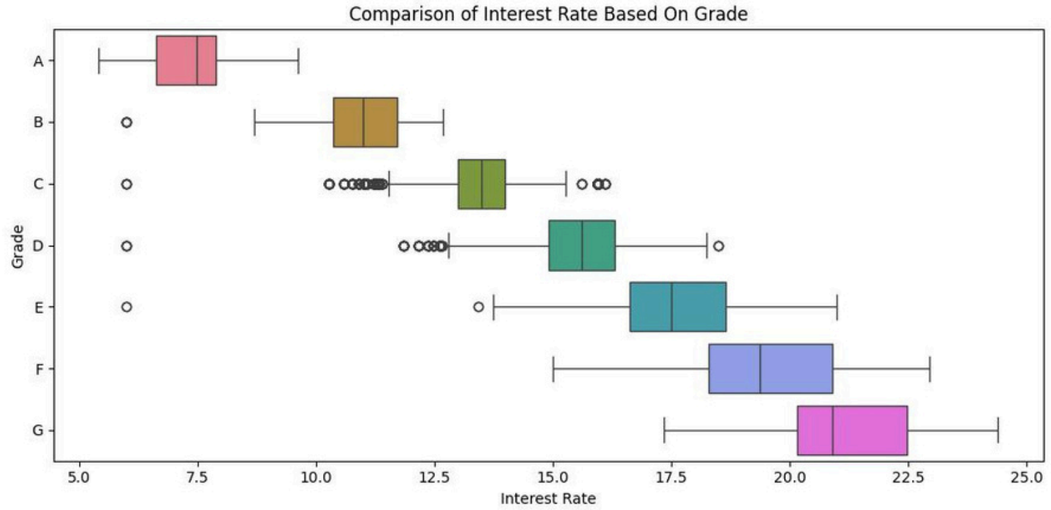


SEGMENTED UNIVARIATE ANALYSIS

Observation: The number of fully paid loans has been increasing exponentially over time, while the number of defaulted loans has not grown at the same rate.

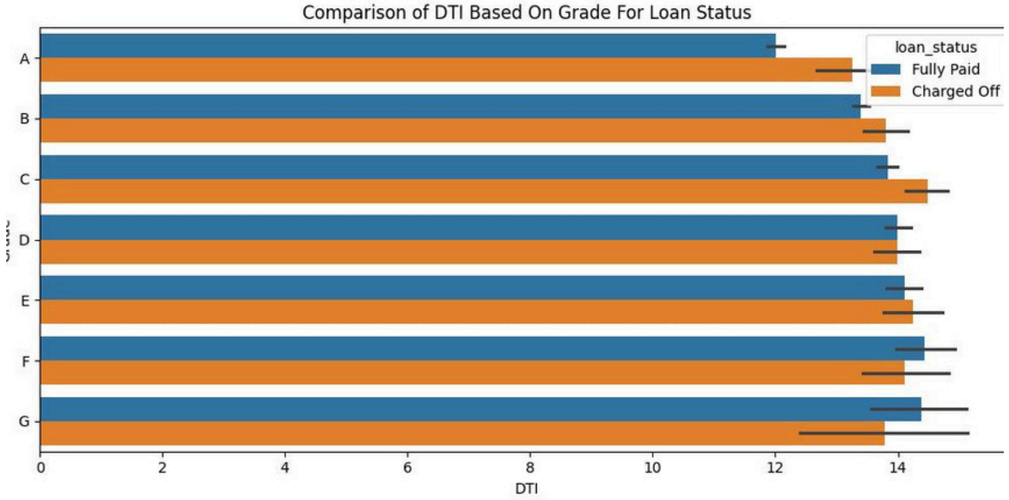


BIVARIATE ANALYSIS



Observation:

1. The loan grade reflects the level of risk, so as the risk increases, the interest rate also goes up.
2. Grade A, which represents the lowest risk, also has the lowest Debt-to-Income (DTI) ratio. This suggests that higher-grade loans are associated with a lower rate of default.



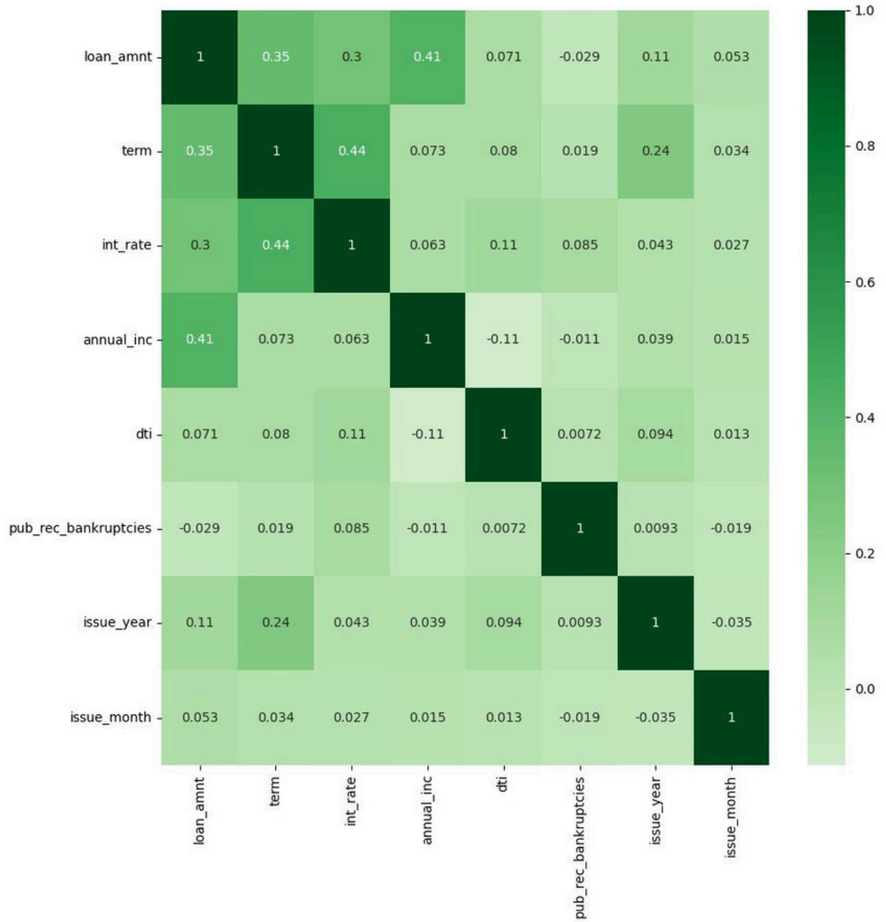
BIVARIATE ANALYSIS

Interest Rate:

The interest rate shows a moderate correlation with the loan term, with a correlation coefficient (r) of 0.44.

Loan Amount:

There's a moderate correlation between loan amount and annual income, with a correlation coefficient (r) of 0.41.



INSIGHTS AND IMPLICATIONS FOR LENDING CLUB.

- Loan Status: The majority of loans are successfully repaid, with defaulted loans being significantly lower in number.
- Loan Term: 36-month loans are more popular and less risky compared to 60-month loans, which have a higher default rate.
- Interest Rate Distribution: Interest rates cluster around 5-15%, with a notable decrease in defaults at rates above 17.5%.
- Loan Grades: Higher-grade loans (A and B) are predominant, reflecting lower-risk lending practices.
- Borrower Experience: Borrowers with over 10 years of experience have both higher repayment and default rates.
- Home Ownership: Borrowers who own homes have lower default rates compared to those renting or on a mortgage.
- Debt-to-Income Ratio (DTI): Borrowers with higher DTI ratios (especially above 15) show a greater likelihood of default.

CONCLUSION

- **Summary:**
  - Key insights reveal patterns in loan defaults and repayments, helping us better understand borrower behavior.
- **Recommendations:**
  - Focus on Debt-to-Income Ratio (DTI) and Loan Grades to predict defaults.
  - Ensure strong Verification and assess Annual Income carefully.
- **Additional Considerations:**
  - Caution with borrowers in lower grades or high DTI, especially those with over 10 years of experience.
- **Implications:**
  - Target these factors to reduce defaults and strengthen Lending Club’s loan portfolio.

ACKNOWLEDGEMENTS.

- This project was inspired by the need to better understand borrower behavior and loan performance at Lending Club.
- Data and analysis techniques were based on common practices in data science.
- Special thanks to the open-source community for providing the tools and resources used in this project.