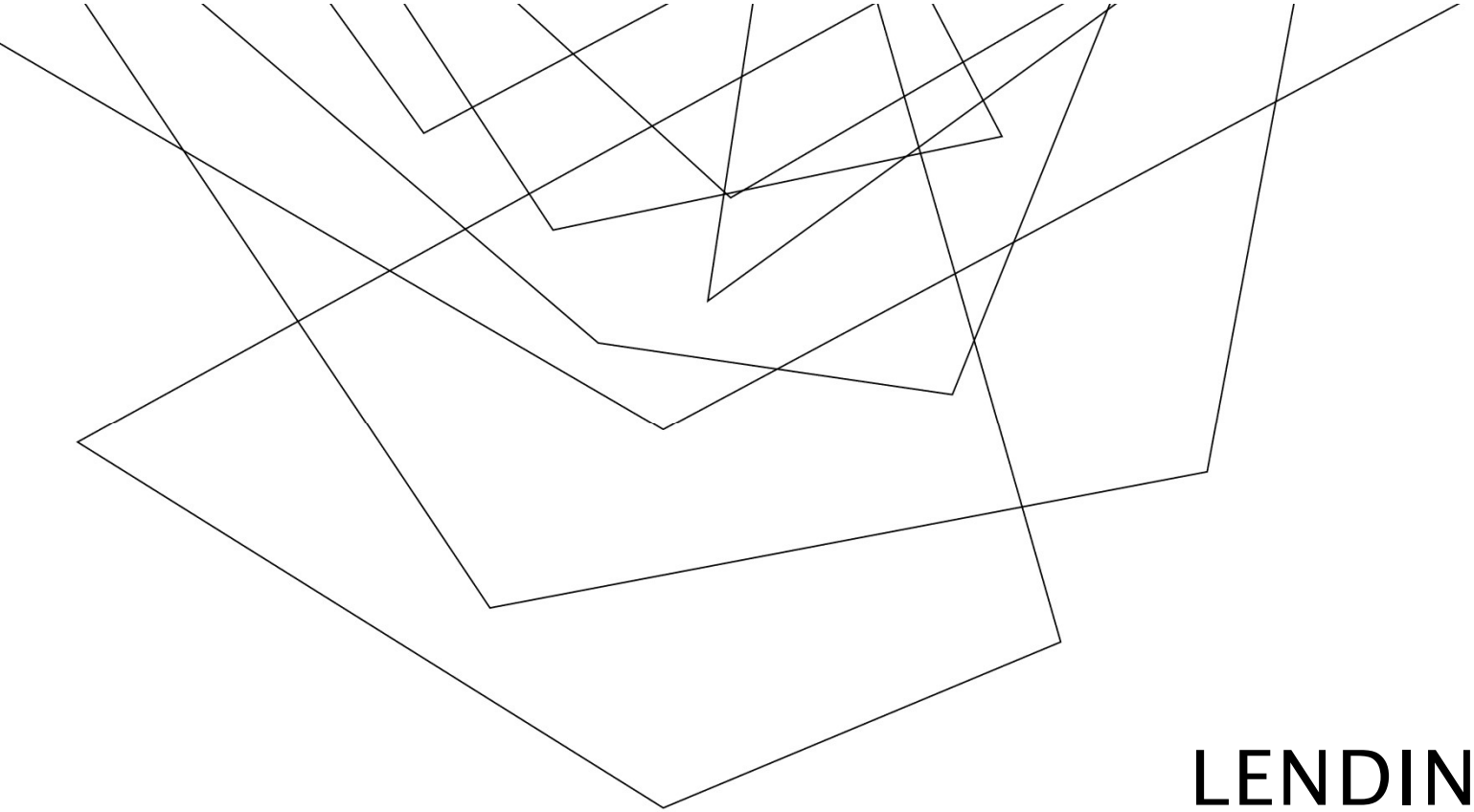


LENDING CLUB CASE STUDY

Group Members: Banashri Gogoi
Balaji R

AGENDA

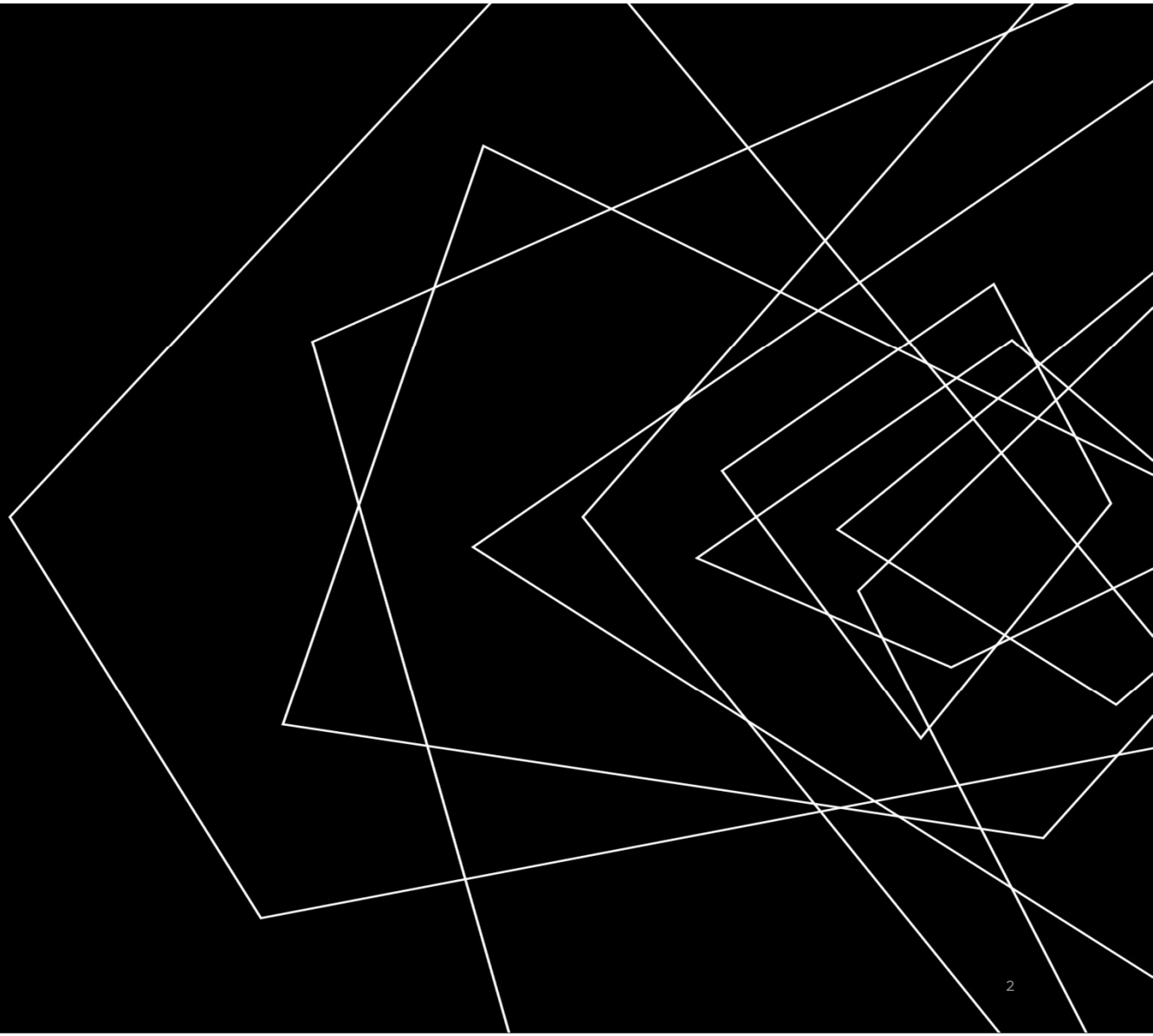
Topic Objective

Topic Business Understanding

Topic Univariate Analysis

Topic Bivariate Analysis

Topic Recommendations



OBJECTIVES

The Objective of this case study is to implement EDA technique on a real-world problem and understand the insights and present in a business first manner via presentation.

Benefits of the case study:
Gives an idea about how EDA is used in real life business problems.
It also develops a basic understanding of risk analytics in banking and financial services.
How the data is used to minimize loss of money while lending it. to clients.
It improves our understating of visualization and what charts to use for real life data.

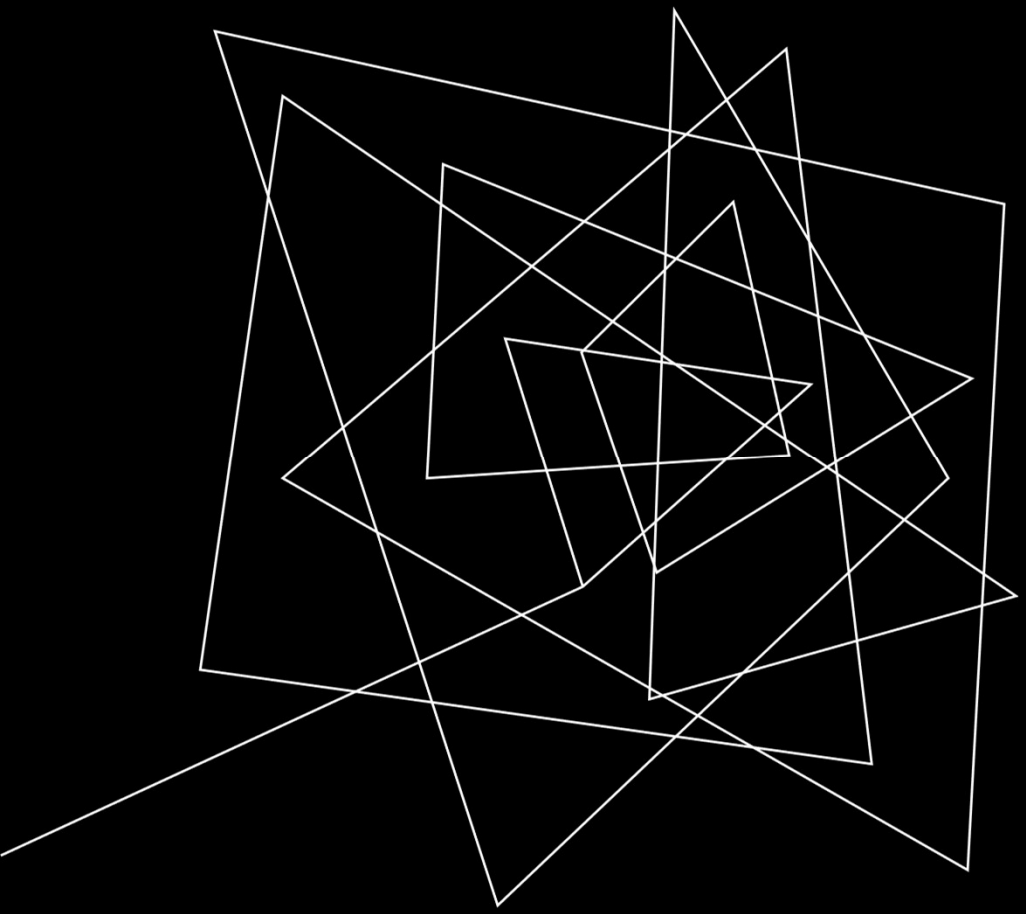
BUSINESS UNDERSTANDING

The business objective is to take a decision whenever they receive a loan application whether to reject or approve based on certain variables.

The data given below contains information about past loan applicants and whether they ‘defaulted’ or not. Data has details regarding approved loan not the rejected ones. It has 3 status of loan which is Fully Paid, Current and Charged-Off.

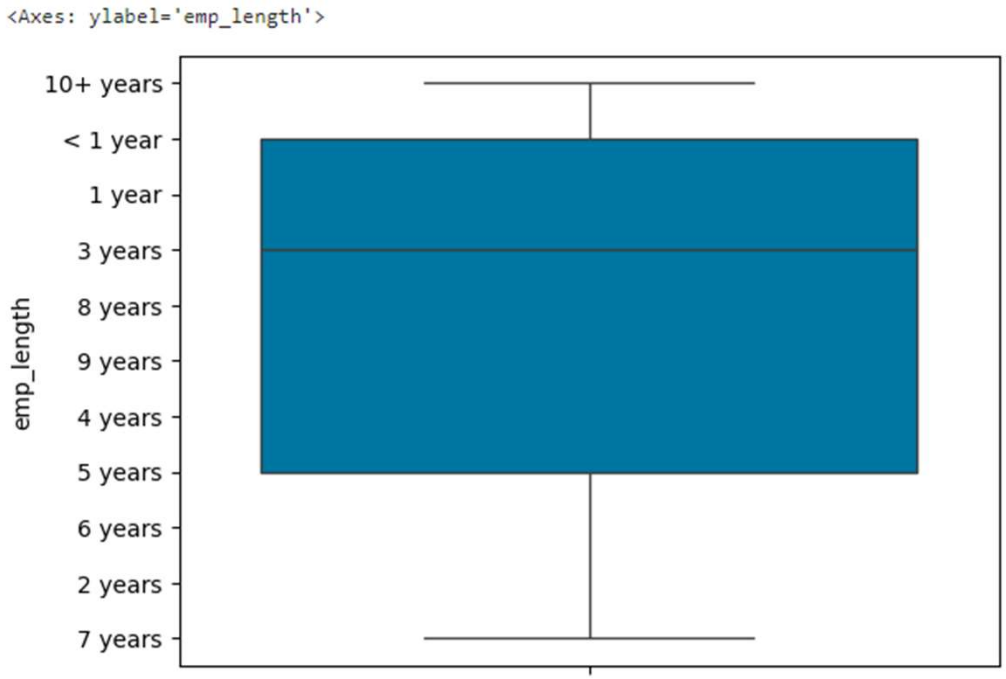
Data Cleanup and Preparation Process:

- Importing the Data
- Removing large null value columns
- Removing Duplicate Data
- Removing irrelevant columns
- Removing/Fixing null values
- Correcting data types and deriving new columns
- Filter Data for requirement.
- Removing outliers



UNIVARIATE ANALYSIS

ANALYSIS OF EMPLOYMENT LENGTH BY BOXPLOT



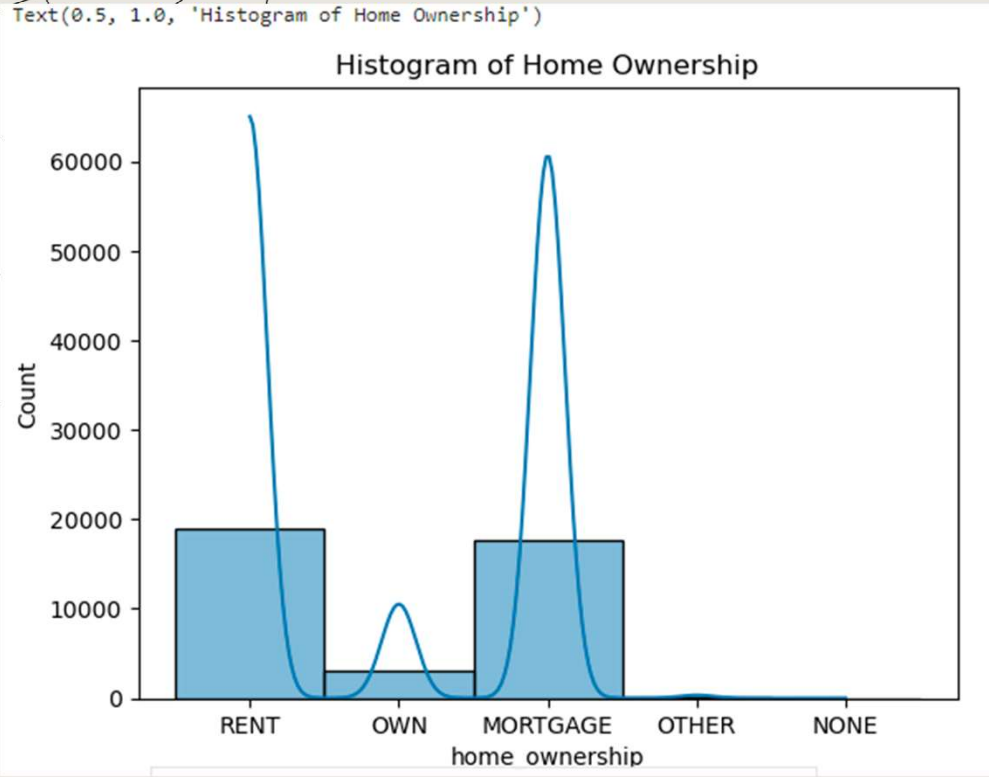
The plot suggests that most employees have been with their current employer for between 3 and 4 years. However, there is a significant portion of employees with longer tenures, and there are some outliers with much longer employment lengths.

The right-skewness indicates that there are fewer employees with shorter tenures compared to those with longer tenures.

The median employment length is between 3 and 4 years. This is indicated by the horizontal line inside the box.

The IQR, which is the range between the first quartile (25th percentile) and the third quartile (75th percentile), can be estimated from the box. It appears to be around 4 years. This means that 50% of the data points fall within a range of 4 years.

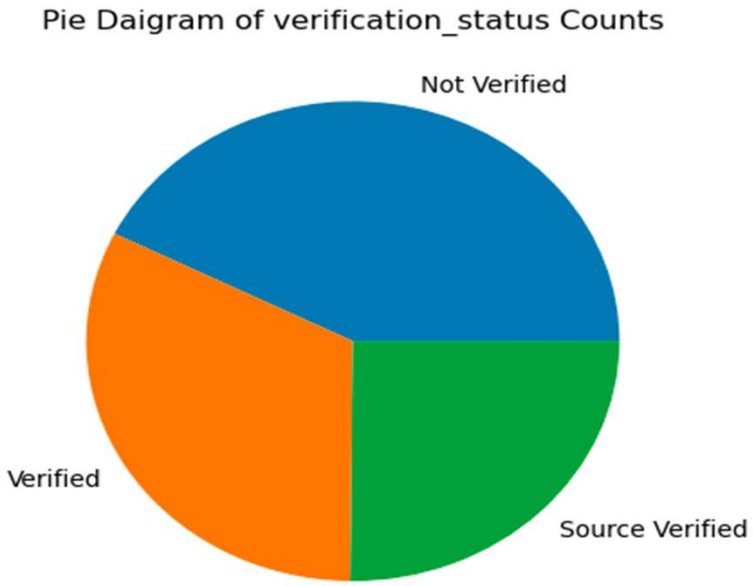
ANALYSIS OF HOMEOWNERS BY HISTOGRAM



The histogram provides a clear visual representation of the distribution of home ownership categories in the data. The right skew and the dominance of the "MORTGAGE" and "RENT" categories are key observations to consider for further analysis.

Most individuals in the dataset are either homeowners (MORTGAGE) or renters (RENT).
A smaller portion of individuals own their homes outright (OWN), while even fewer have "OTHER" or "NONE" as their home ownership status.
The right skew suggests that there are fewer individuals who fall into the "OTHER" and "NONE" categories compared to the "MORTGAGE" and "RENT" categories.

PIE DIAGRAM ANALYSIS OF VERIFICATION STATUS COUNTS



Overall Interpretation:

Based on the pie chart, most data points in the dataset have not been verified. A significant portion has been verified, while a smaller number have been verified at the source level. This information can be valuable for understanding the quality and reliability of the data.

Data Categories:

Not Verified: The largest segment of the pie chart, indicating that most of the data points have not been verified.

Verified: A significant portion of the data has been verified.

Source Verified: The smallest segment, suggesting that fewer data points have been verified at the source level.

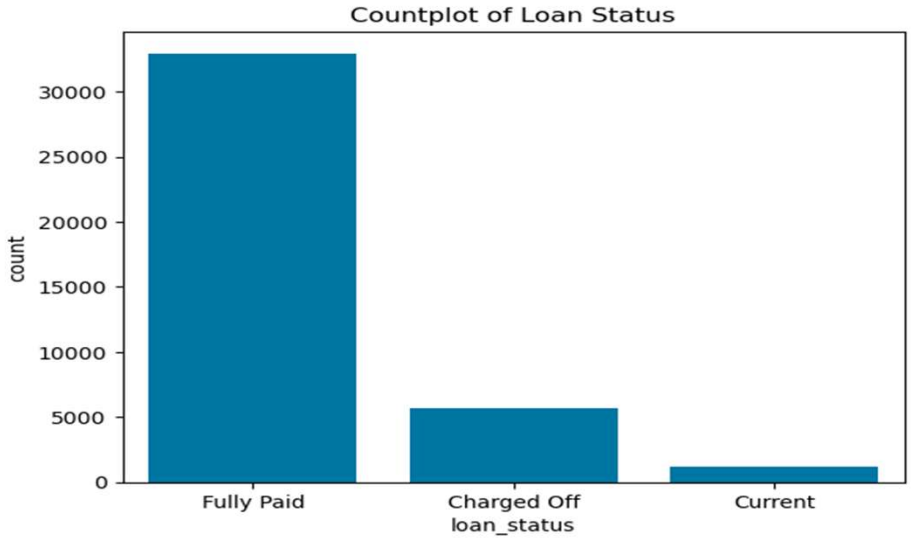
Visual Representation:

Colors: The chart uses distinct colors for each category, making it easy to differentiate between them.

Labels: The labels clearly identify each category, providing context for the data.

Proportions: The size of each segment visually represents the proportion of data points in that category.

COUNTER PLOT ANALYSIS OF LOAN STATUS



Overall Insights:

Based on the countplot, a majority of the loans in the dataset have been fully paid. A smaller number have been charged off, while an even smaller portion are currently active (not yet paid off or charged off). This information can provide valuable insights into the performance of the loan portfolio.

Bars: The plot displays vertical bars representing the frequency (count) of each loan status category.

X-axis: The x-axis labels the loan status categories: "Fully Paid," "Charged Off," and "Current."

Y-axis: The y-axis represents the count of loans in each category.

Colors: The bars are filled with a single color, likely for consistency.

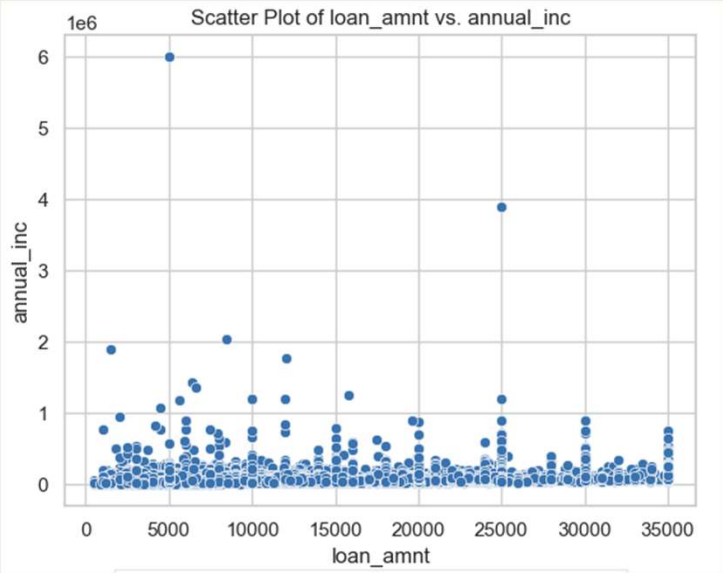
Data Interpretation:

Dominant Category: "Fully Paid" is the most frequent loan status, with a significantly higher count compared to the other two categories.

Secondary Categories: "Charged Off" and "Current" have considerably lower counts, suggesting that a smaller proportion of loans fall into these categories.

BIVARIATE ANALYSIS

SCATTER PLOT ANALYSIS OF LOAN AMOUNT VS. ANNUAL INCOME



Income-Based Lending: Lenders often consider a borrower's income when determining the maximum loan amount they can approve. This could explain the positive correlation between the two variables.

Debt-to-Income Ratio: Lenders may also consider a borrower's debt-to-income ratio, which is the ratio of monthly debt payments to monthly income. This could explain why some individuals with higher incomes may still have lower loan amounts, as they may have other financial obligations.

Individual Factors: Other factors, such as credit score, employment stability, and purpose of the loan, can also influence the loan amount, leading to some variability in the relationship.

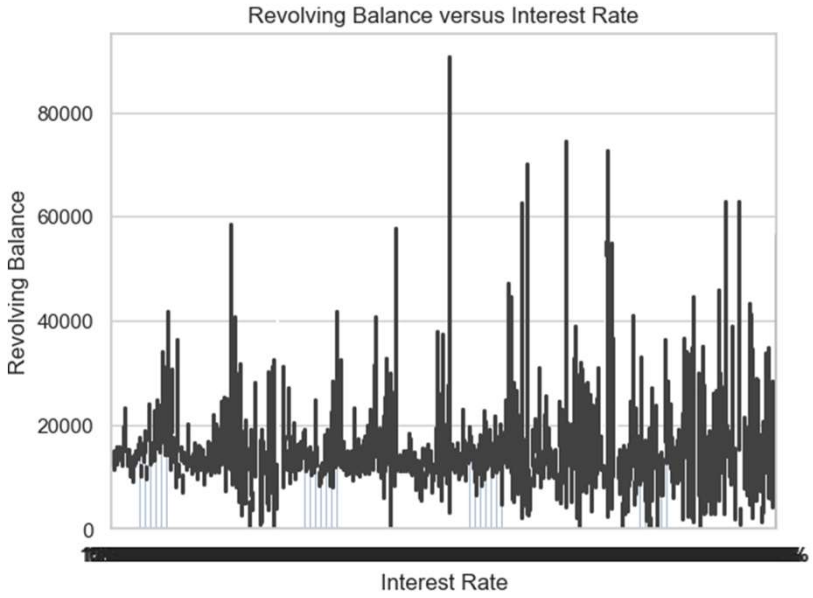
Positive Correlation: There seems to be a general positive correlation between loan amount (loan_amnt) and annual income (annual_inc). This means that as annual income increases, the loan amount tends to increase as well.

Clustering: The data points cluster heavily in the lower regions of the plot, indicating that most borrowers have lower loan amounts and annual incomes.

Outliers: A few outliers are visible, particularly on the higher end of both loan amount and annual income. These are the points that deviate significantly from the general trend.

Scatter: There is a moderate degree of scatter around the trend line, suggesting that while there is a general relationship, there are also individual variations in how loan amounts relate to annual income.

BAR PLOT ANALYSIS OF REVOLVING BALANCE VERSUS INTEREST RATE

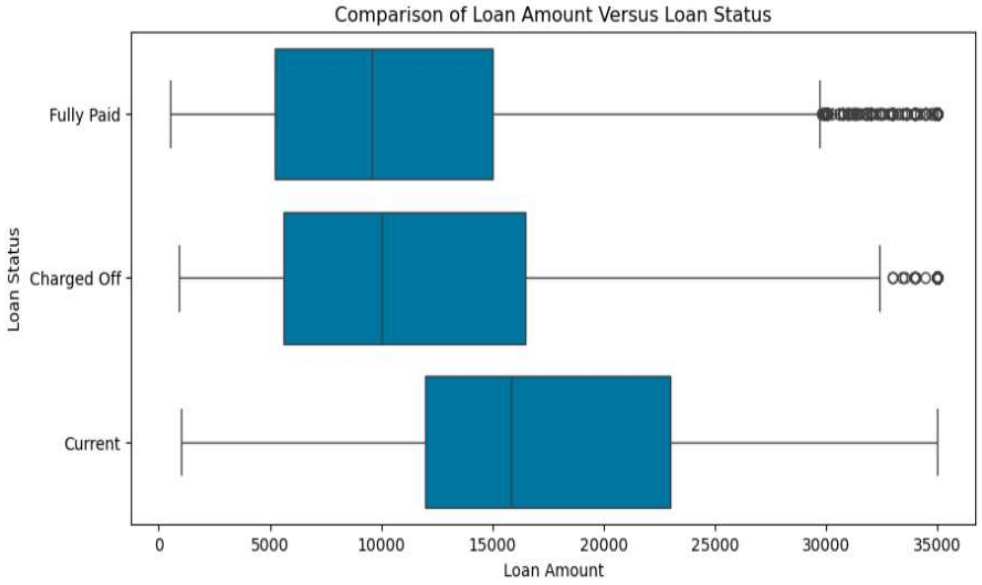


Grouping: To better understand the relationship, consider grouping the interest rates into ranges (e.g., low, medium, high) and analyzing the average or median revolving balance within each group.
Correlation: Calculate the correlation coefficient between revolving balance and interest rate to quantify the strength and direction of the relationship. However, keep in mind that correlation doesn't imply causation.

The bar plot provides a visual representation of the distribution of revolving balances across different interest rates. While there doesn't seem to be a strong linear relationship, further analysis is needed to understand the underlying factors influencing revolving balance and its connection to interest rate.

Distribution of Revolving Balance: The bars vary in height, indicating that revolving balances are not evenly distributed across interest rates. There are some interest rates with a higher concentration of borrowers with higher revolving balances, while others have a more spread-out distribution.
Outliers: There appear to be a few outliers with extremely high revolving balances, as evidenced by the very tall bars. These outliers might skew the overall distribution.
Relationship with Interest Rate: There doesn't seem to be a clear, linear relationship between revolving balance and interest rate. The bars are not consistently higher or lower for specific interest rate ranges. This suggests that other factors besides interest rate might influence revolving balance.

BOXPLOT ANALYSIS OF COMPARISON OF LOAN AMOUNT VERSUS LOAN STATUS

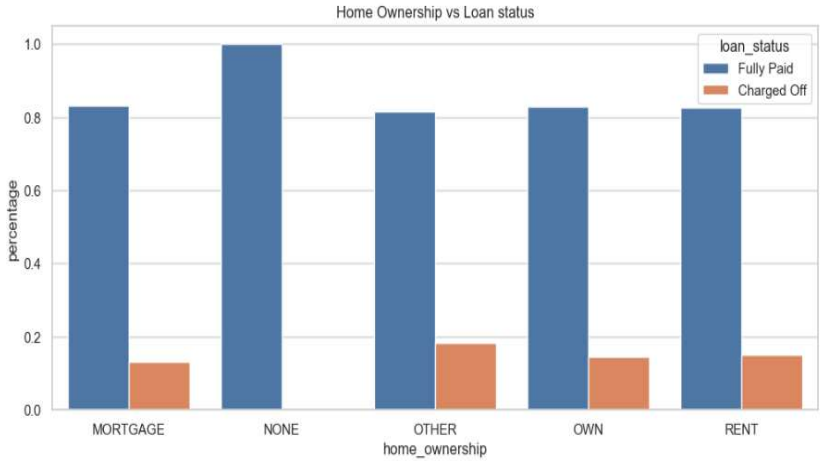


Loan Approval Criteria: Lenders may have different criteria for approving loans based on the loan status. For example, they might be more likely to approve larger loan amounts for applicants with a history of fully paying off loans.

Loan Amount Distribution: The box plots show the distribution of loan amounts for each loan status category: Fully Paid, Charged Off, and Current.
Median Loan Amounts: The median loan amount is highest for Fully Paid loans, followed by Charged Off loans, and then Current loans.
Interquartile Ranges (IQRs): The IQRs, represented by the boxes, show the variability within each loan status category. Fully Paid loans have a slightly larger IQR compared to Charged Off and Current loans.
Outliers: There are outliers present in all three categories, but they are more prominent in the Fully Paid and Charged Off categories. These outliers represent loans with significantly higher or lower amounts compared to most loans in their respective categories.

Repayment Capacity: Borrowers with larger loan amounts may have higher repayment capacities, which could increase their chances of fully paying off the loan.
Risk Assessment: Lenders may assess the risk associated with each loan status category differently. They might be more cautious about approving large loan amounts to borrowers with a history of charging off loans.

BARPLOT ANALYSIS OF HOME OWNERSHIP VS LOAN STATUS



Possible Explanations:

Financial Stability: Homeowners may have more financial stability due to equity in their homes, which could make them less likely to default on loans.

Creditworthiness: Lenders may consider homeownership as an indicator of financial responsibility and creditworthiness, leading to more favorable loan terms for homeowners.

Risk Assessment: Lenders might assess the risk associated with each home ownership category differently. They may perceive homeowners as having lower risk compared to those with other statuses.

Loan Status by Home Ownership: The bar plot compares the percentage of Fully Paid and Charged Off loans across different home ownership categories: MORTGAGE, NONE, OTHER, OWN, and RENT.

Fully Paid Loans: The highest percentage of Fully Paid loans is among homeowners (OWN), followed by those with a MORTGAGE. The lowest percentage of Fully Paid loans is among those with NONE or OTHER home ownership status.

Charged Off Loans: The highest percentage of Charged Off loans is among those with NONE or OTHER home ownership status, followed by renters (RENT). The lowest percentage of Charged Off loans is among homeowners (OWN).

Relationship: There seems to be a negative relationship between home ownership and the likelihood of a loan being Charged Off. Homeowners are more likely to fully pay off their loans compared to those with other home ownership statuses.

CONCLUSION

1. Burrowers not from large urban cities like California, new york, texas, floridaetc.

2. Burrowers having annual income in the range 50000-100000.

3. Burrowers with very high Debt to Income value.

4. Burrowers with working experience 10+ years.

5. Borrowers with mortgage home ownership are taking higher loans and defaulting the approved loans. Lending club should stop giving loans to this category when loan amount requested is more than 12000

6. Lending club should reduce the high interest loans for 60 months tenure, they are prone to loan default.