LENDING CLUB CASE STUDY

EDA - LOAN DATA SET

Nikhil Gupta

Nishant Singh

INDEX

BACKGROUND

There is a consumer finance company which specializes in lending various types of loans to urban customers. This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile.

PROBLEM STATEMENT

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the bank's decision:

- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as:

- ❑ Denying the loan
- ❑ Reducing the amount of loan
- ❑ Lending (to risky applicants) at a higher interest rate

BUSINESS OBJECTIVES

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
   - ❑ Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
   - ❑ Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
   - ❑ Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.).

- Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss)
- If one can identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicant's using EDA is the aim of this case study.
- **In other words, the objective of the company is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.**

DATA UNDERSTANDING & SUMMARY

INSPECTING ROWS AND COLUMNS

Initial loading of data set shows 39717 rows and 111 columns

DATATYPES INFORMATION

float64(74), int64(13), object(24)

NULL & SINGLE VALUES

Large no. of columns identified having 100% null & single values

DATA CLEANING & MANIPULATION

CONVERT DATA TYPES

- ‘int_rate’ & ‘revol_util’ has been converted from string to int. Additional ‘%’ has been trimmed.
- ‘issue_d’ & ‘earliest_cr_line’ has been converted to Date and time datatype.

UNNECESSARY COLUMNS & MISSING VALUES

- 57 columns having more than 40% missing values are dropped.
- Columns having single (same) values in all rows are removed.
- Irrelevant columns having no relevance to EDA removed (desc, zip_code, member_id, url, sub_grade, title.
- Using domain knowledge, variables which are generated post-approval of loan or are not directly indicative of the borrower's pre-approval financial status and creditworthiness are removed. 15 variables dropped.
- Columns having null rows less than 2% are dropped.
- Final Rows & Columns after data cleaning: 37164 ROWS, 24 COLUMNS

HEADERS & FOOTERS

- There were no header, footers, summary or Total rows found.

DERIVED COLUMNS

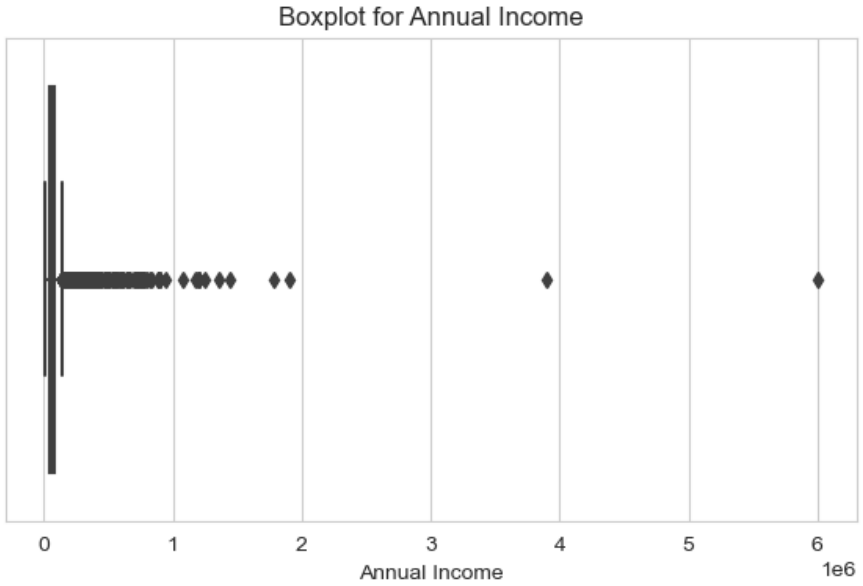
- Derived column ‘**loan_to_income_ratio**‘ is the new column created by dividing the loan amount by the annual income. This column represents the ratio of the loan amount to the borrower's annual income.
- Derived column **‘int_rate_category‘** is the new column created by categorizing the int_rate values into the defined bins and assigning the corresponding labels – Low, Medium etc.
- Derived column **‘dti_category‘** is the new column created by categorizing the dti values into the defined bins and assigning the corresponding labels – Low, Medium etc.

HANDLING DATA QUALITY ISSUES

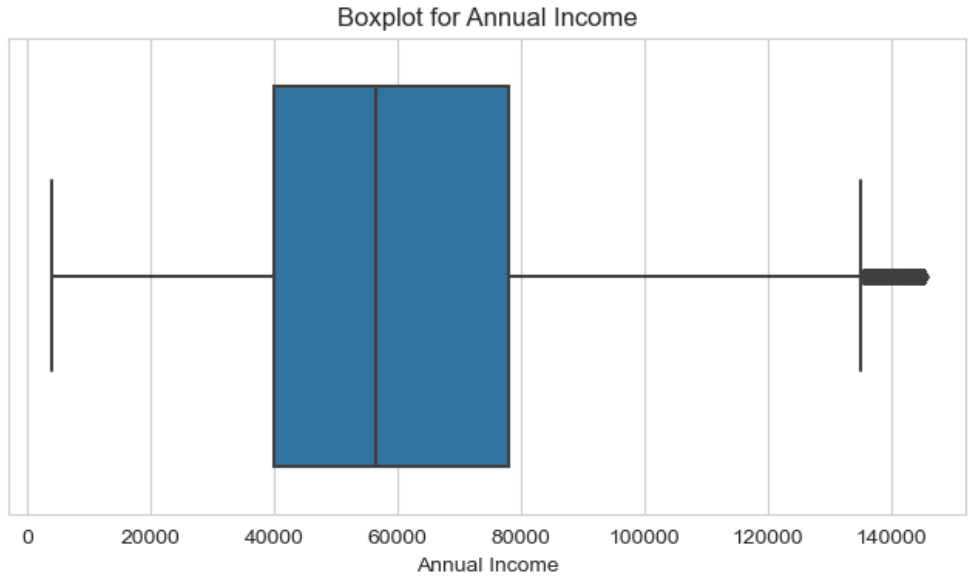
- Imputation: Column “emp_length” has 2.75 % null values, hence imputed with **mode** value.
- Standardize: “emp_length” has values in format of “10+ years”, “9 years” and so on, hence converted to numeric values 10, 9 and so on for better numerical analysis.

HANDLING OUTLIERS

Outliers were noticed on Annual Income which may deviate the analysis of defaulters due to extreme values. With the help of IQR range (IQR = Q3 - Q1) outliers were removed



BEFORE



AFTER

ANALYSIS APPROACH

1. The most important target variable identified in the dataset is Loan Status, which will help in identifying the defaulters.

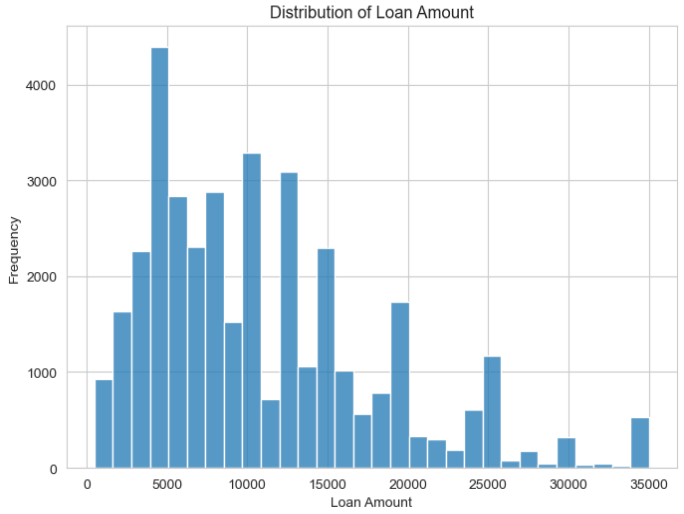
2. Loan Status has 3 outputs:
   - **Fully Paid**
   - **Current**
   - **Charged Off**

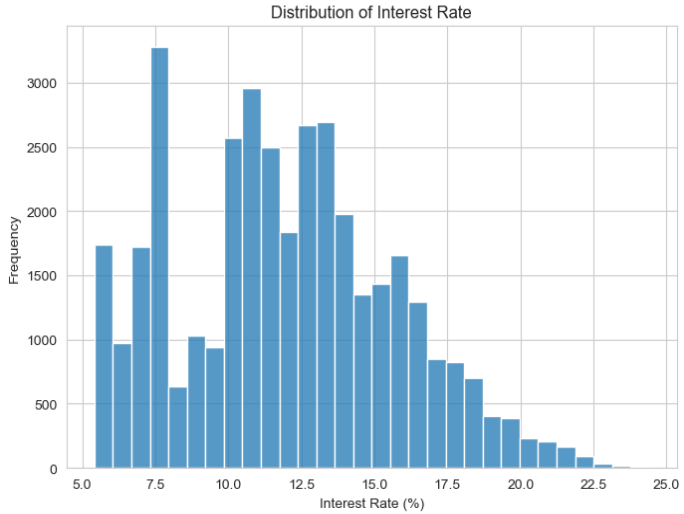
3. We ran the analysis through multiple variable columns to find its relationship with Charged off & Full paid loans

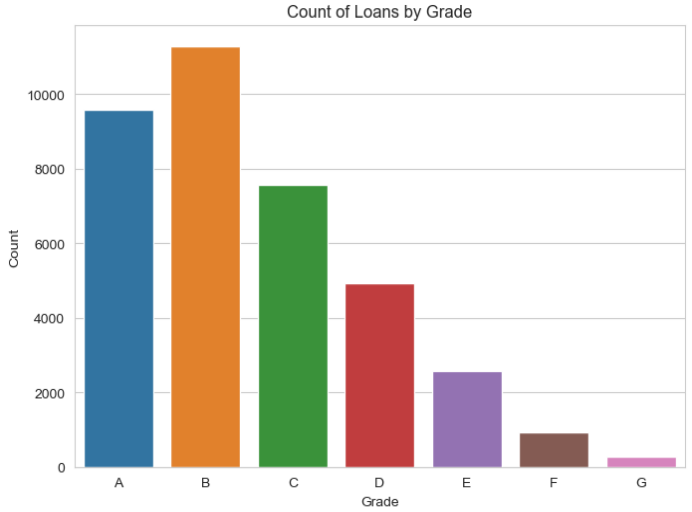
4. We have created multiple test case scenarios to identify loan default with respect to target variable “LOAN STATUS” and in some cases without target variable as well where between various numerical & categorical variables are analyzed to identify various domain trends.

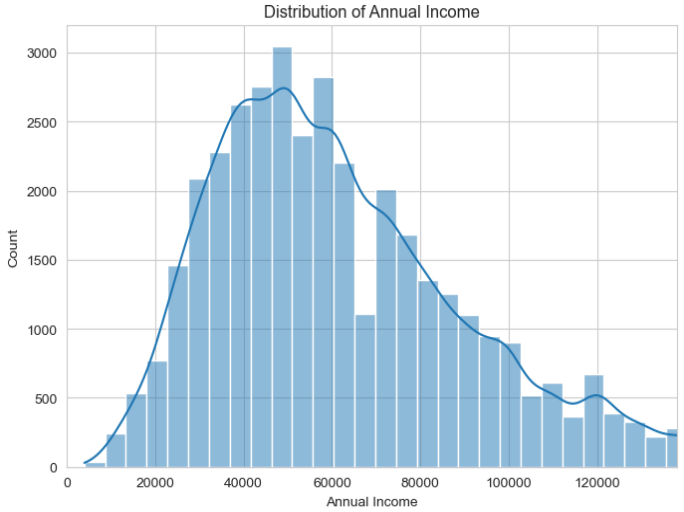
DATA ANALYSIS - UNIVARIATE ANALYSIS

LOAN AMOUNT, INTEREST RATE, GRADE, ANNUAL INCOME



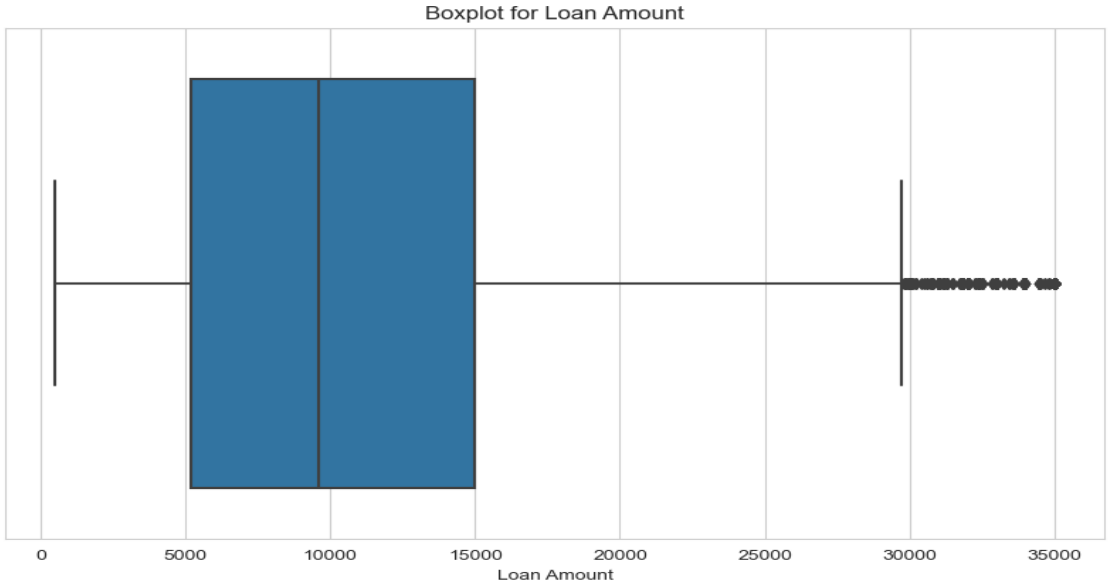




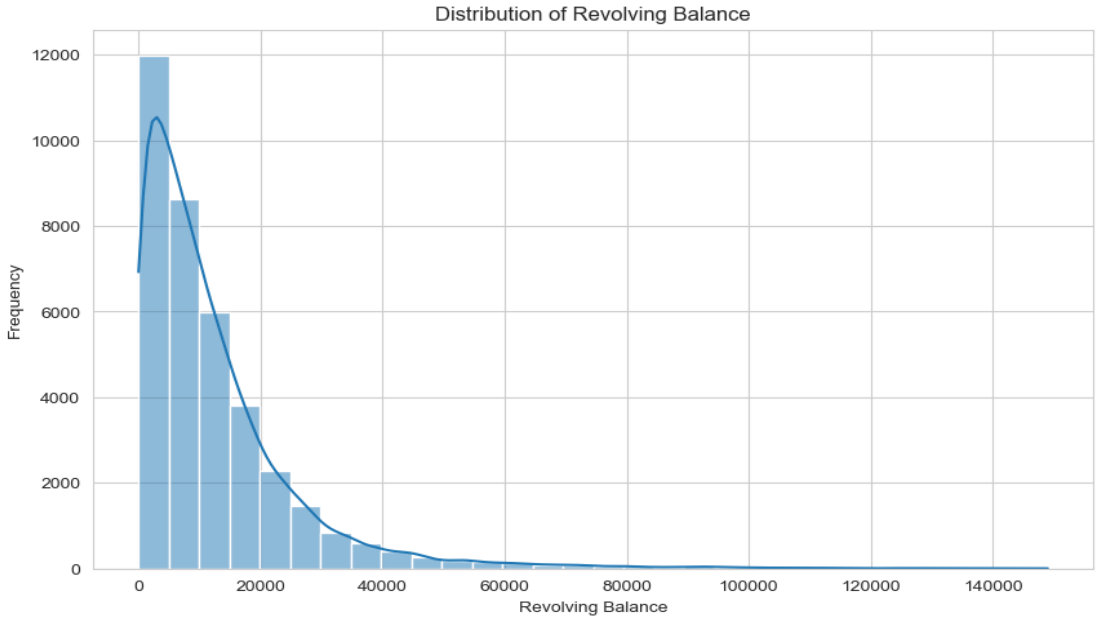


1. **Loan Amount Distribution:** Most loans are concentrated in lower amounts, suggesting that smaller loans are more common among borrowers.
2. **Interest Rate Distribution:** The interest rates are mostly clustered between 10% and 15%. This range might be indicative of the typical risk associated with the majority of loans.
3. **Loan Grades:** Loans are categorized across different grades (A to G). The count of loans decreases as the grade worsens, which likely reflects both the risk assessment of borrowers and the demand for loans across these risk categories.
4. **Annual Income:**
    - The distribution of annual income is heavily right-skewed with a long tail.
    - Most borrowers have annual incomes concentrated below a certain threshold, as indicated by the x-axis being limited to the 99th percentile to avoid extreme outliers.
    - The presence of a peak around the lower income ranges suggests that many borrowers fall into this category.

LOAN AMOUNT & REVOLVING BALANCE

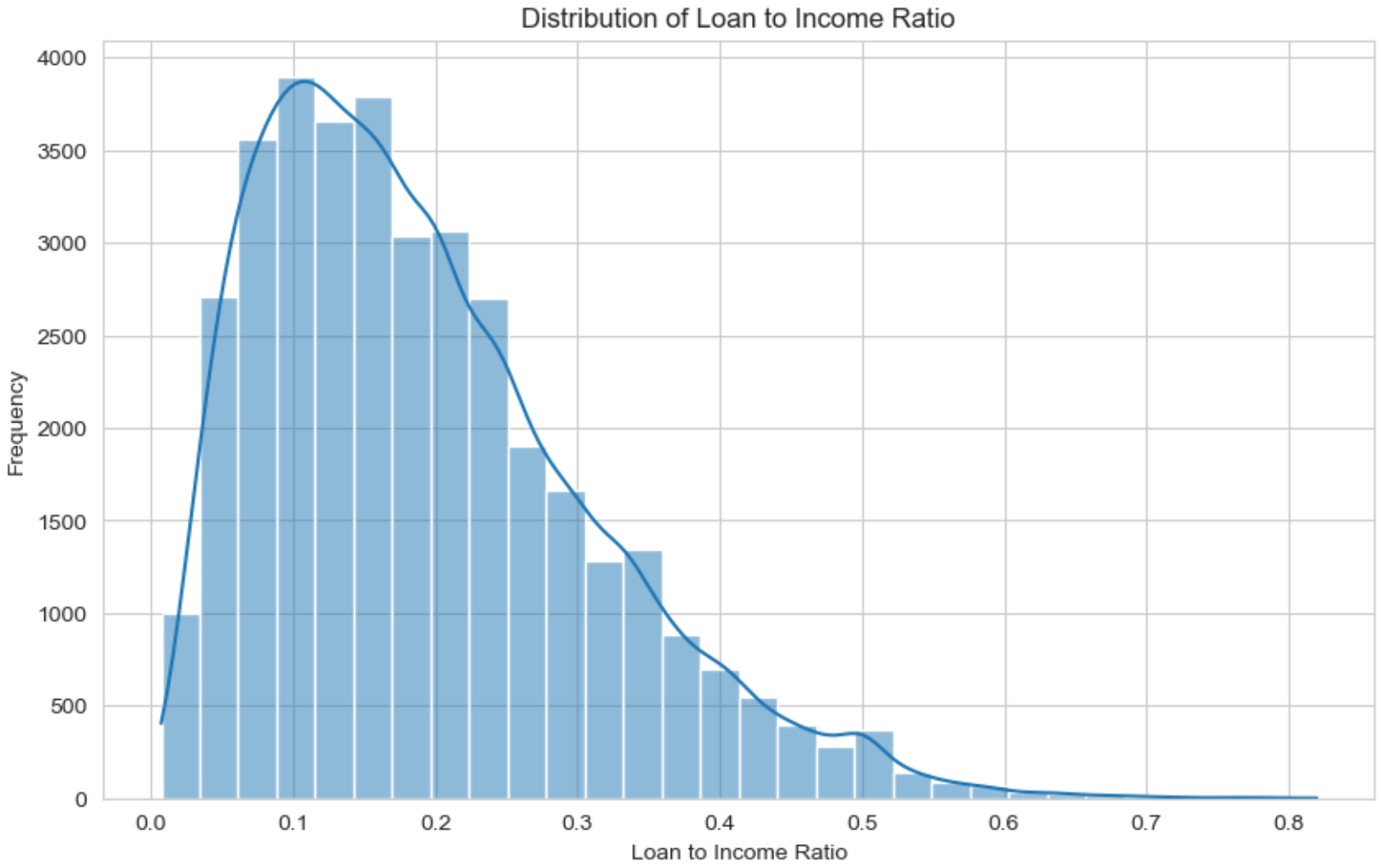


- The interquartile range (IQR), represented by the box, indicates that the middle 50% of loan amounts lie between approximately $5,000 and $20,000.
- The presence of several data points beyond the upper whisker suggests that there are numerous outliers in the dataset. These outliers are loans that are significantly higher than the majority of the dataset.
- The lower whisker starts around $0, indicating that there are no significant outliers on the lower end of loan amounts.



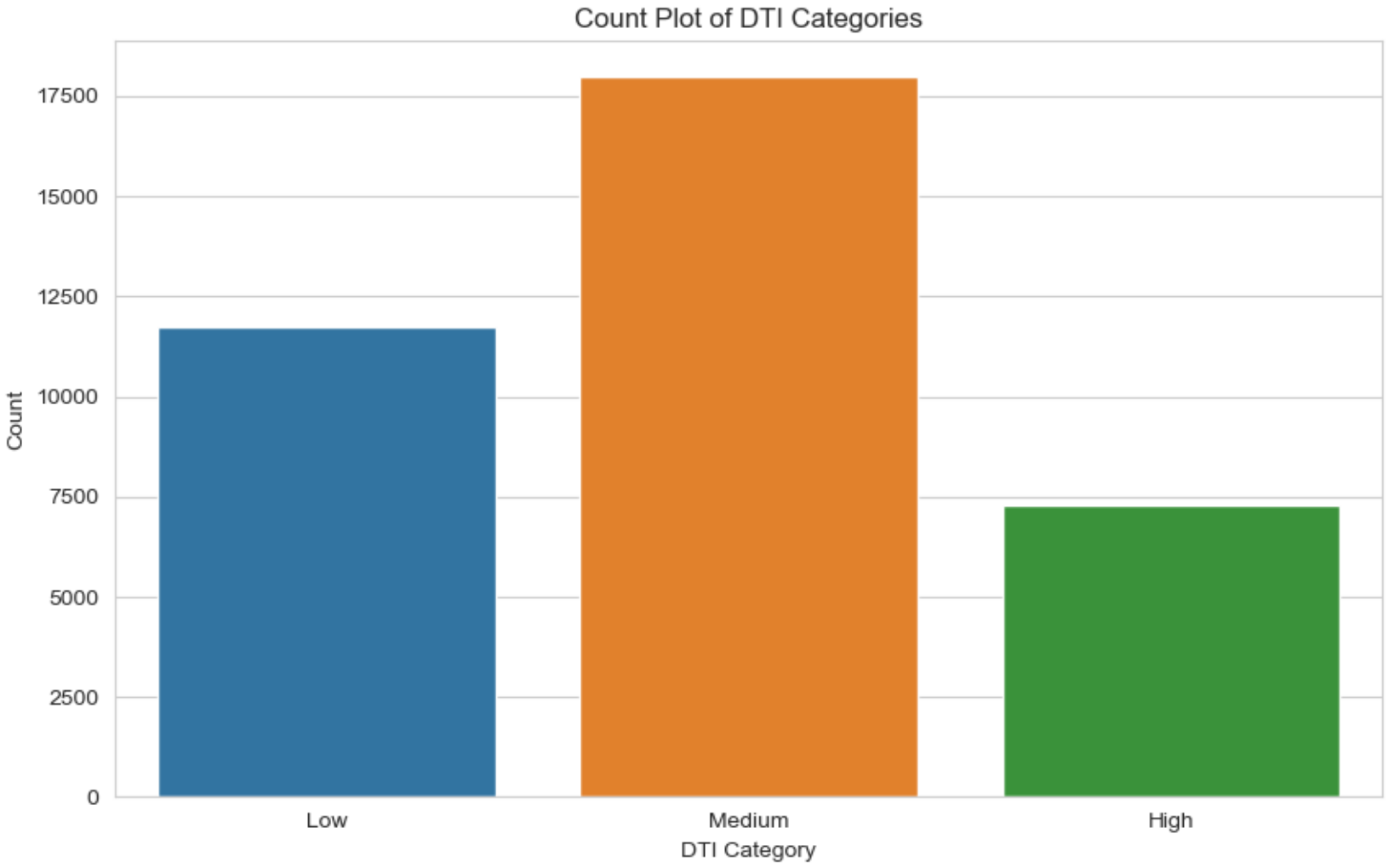
- Most borrowers have revolving balances that are relatively low, with a high frequency observed in the lower ranges of the balance.
- There is a long tail in the distribution, extending up to a maximum revolving balance of around 140,000,
- The concentration of borrowers with revolving balances in the lower range (0-5000) suggests that many borrowers maintain relatively low revolving balances compared to borrowers with higher balances. This could be seen as a positive indicator of responsible credit management.
- A high revolving balance can indicate that the borrower is heavily reliant on revolving credit accounts such as credit cards, which may affect their ability to repay loans.

LOAN TO INCOME RATIO



- The graph shows a peak in the loan to income ratio around 0.1 to 0.2. This suggests that many borrowers have loan amounts that are roughly 10% to 20% of their annual income.
- There is a long tail extending towards higher ratios, up to around 0.8. This tail suggests that although less common, there are still some borrowers with higher loan to income ratios.
- Borrowers with higher loan to income ratios might be at greater risk of financial stress, as a larger portion of their income is committed to loan repayments.

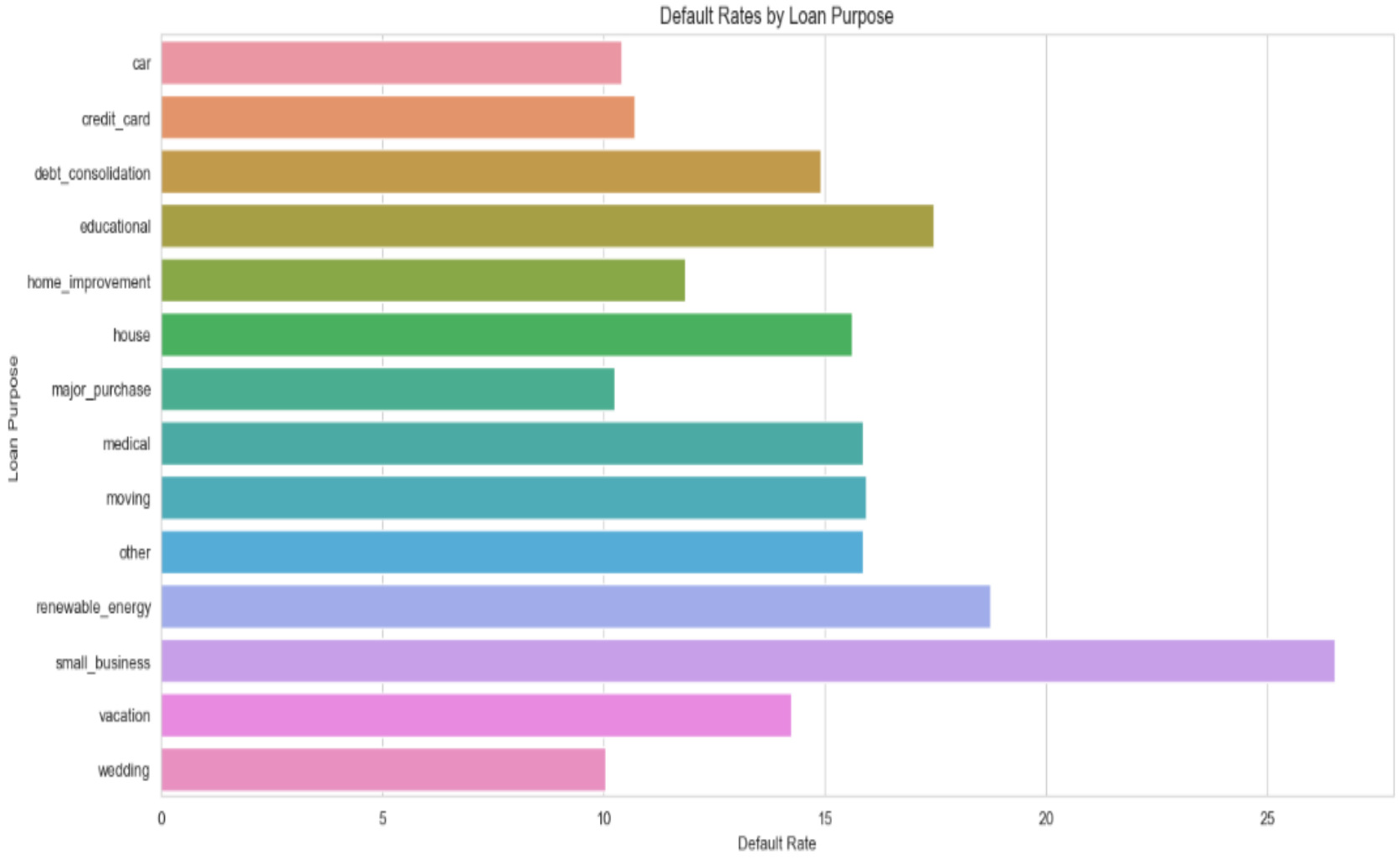
DTI CATEGORIES



- The "Low" DTI category has a considerable number of borrowers, with the count over 10,000, suggesting a good portion of borrowers maintain low debt levels compared to their income.
- The "High" DTI category has the least number of borrowers, with the count slightly over 7,500. This indicates that fewer borrowers are in high debt situations relative to their income.
- The majority of the loans fall into the "Medium" DTI (Debt-to-Income) category, with the count exceeding 17,500. This indicates that a significant portion of borrowers have a moderate level of debt relative to their income.
- **Borrowers in the "High" DTI category should be monitored closely, as they might be more vulnerable to financial stress and potential default**

DATA ANALYSIS - SEGMENTED UNIVARIATE ANALYSIS

DEFAULT RATES BY LOAN PURPOSE



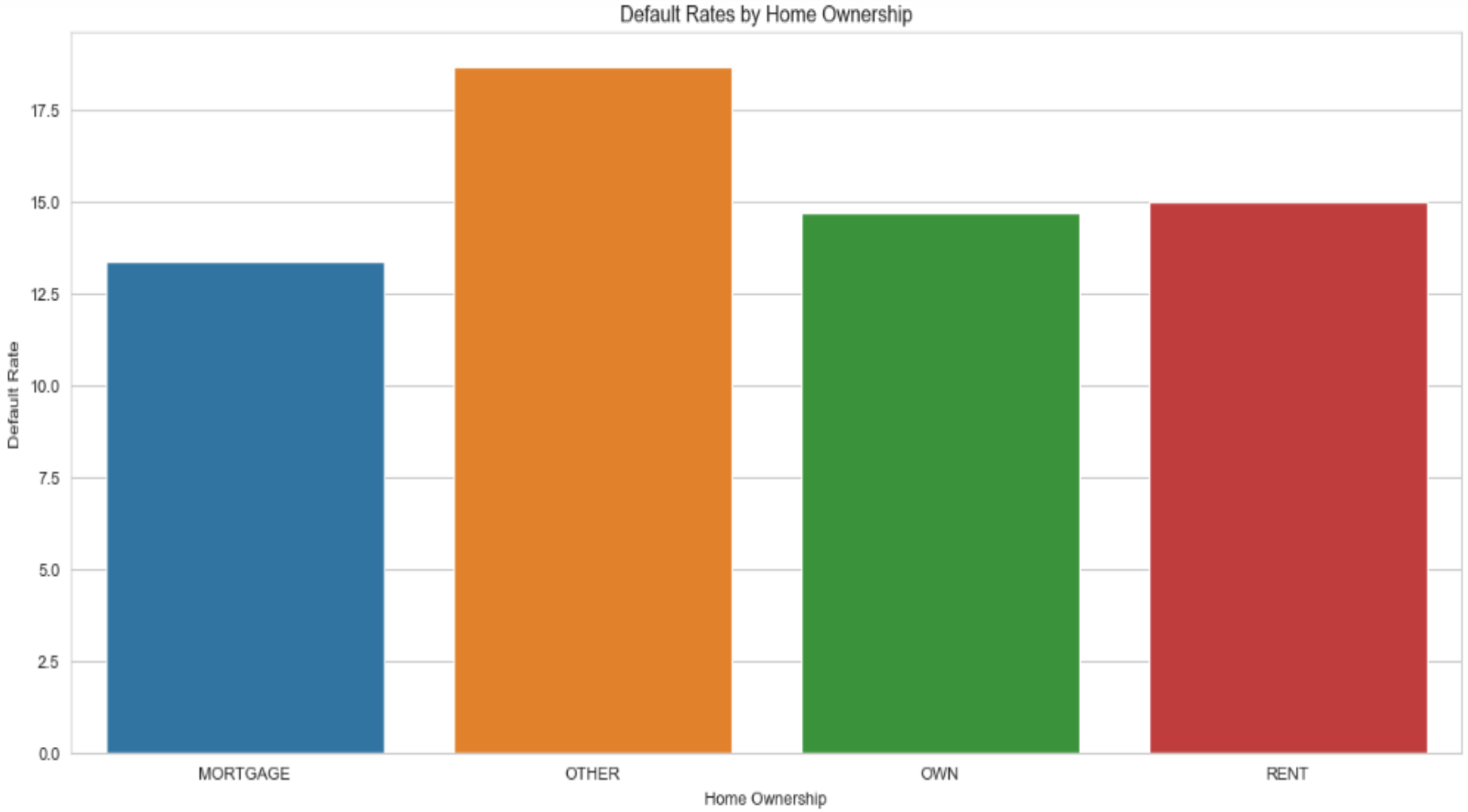
HIGH DEFAULTERS

- Loan purposes such as "small_business" and "renewable_energy" show higher default rates approximately 25.98% & 18.45% respectively compared to other categories. This indicates that borrowers taking loans for these purposes are more likely to default.

LOW DEFAULTERS

- Loan purposes such as Credit Card, Home Improvement, and Car Loans exhibit lower default rates, ranging from approximately 10.33% to 11.66%, suggesting these purposes might be more manageable for borrowers.

DEFAULT RATES BY HOME OWNERSHIP



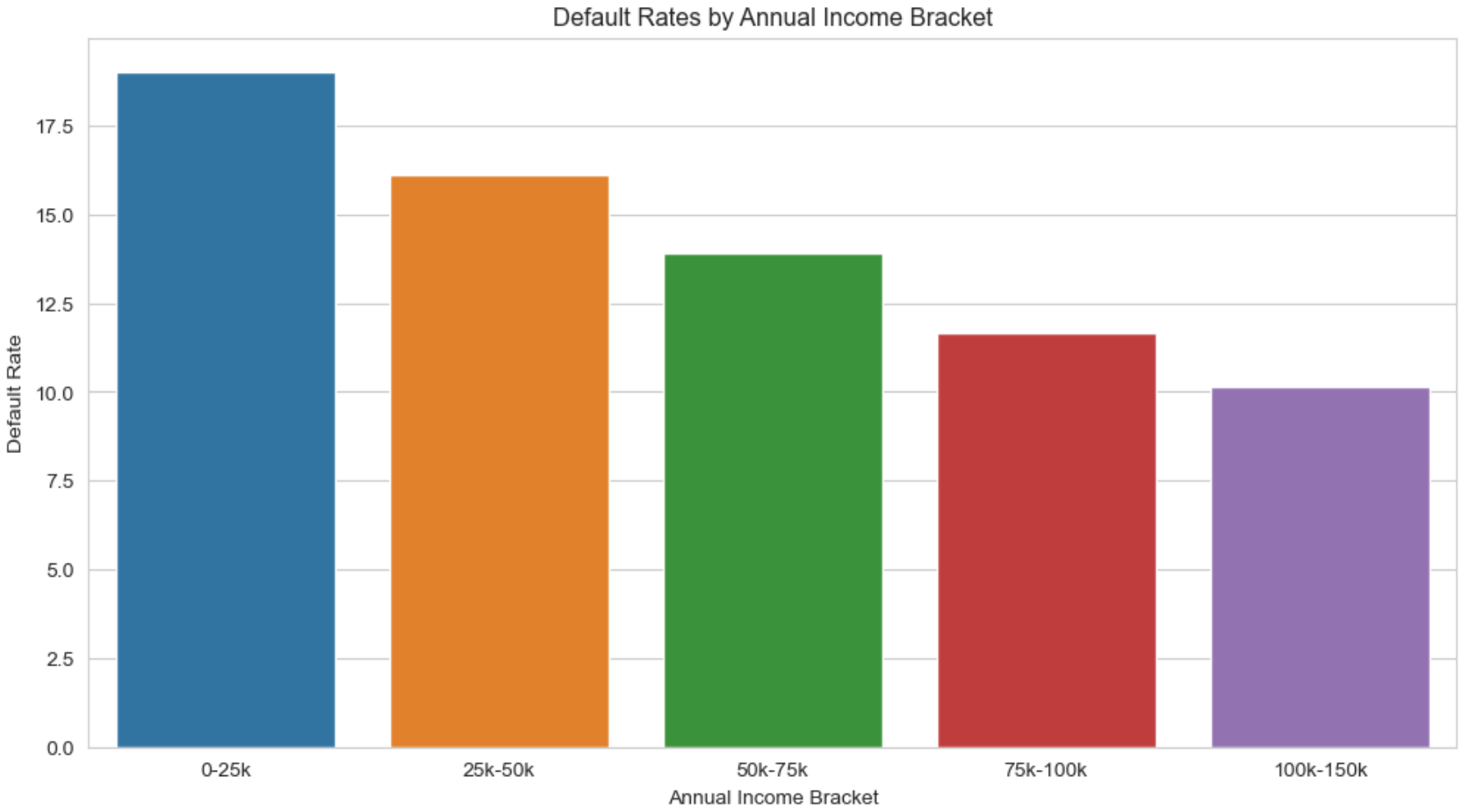
HIGH DEFAULTERS

- Home Ownership such as 'OTHER' shows the highest default rate at approximately 18.37%, suggesting higher risk for non-standard home ownership statuses.
- People on RENT have a default rate of around 15.02%, indicating that renters might face more financial challenges in repaying loans compared to homeowners

LOW DEFAULTERS

- Borrowers with mortgages have the lowest default rate at approximately 13.18%, indicating relatively lower risk compared to other categories

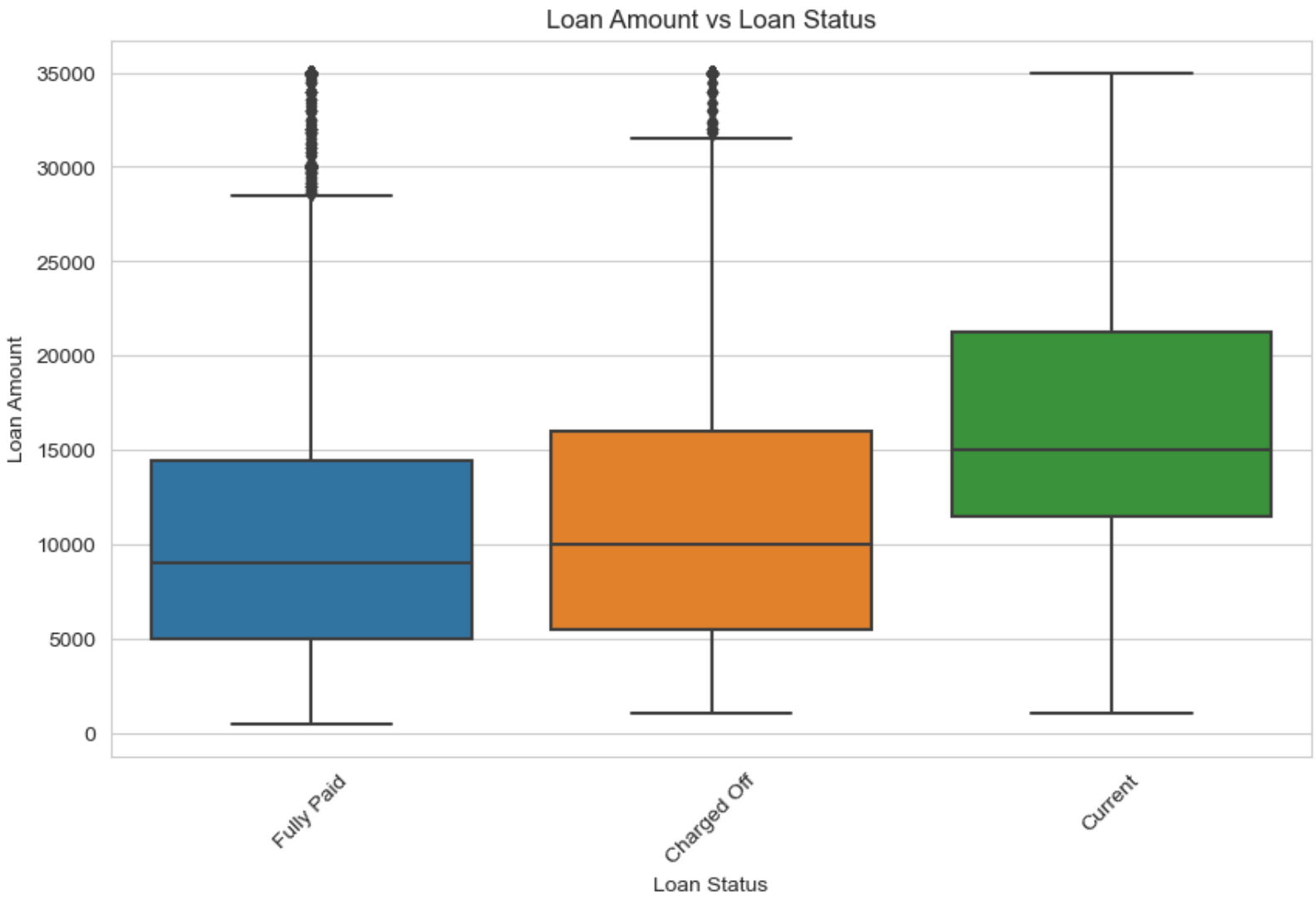
DEFAULT RATES BY ANNUAL INCOME BUCKET



- People with lower income bracket tend to have higher rate of Default.
- There is a noticeable trend of decreasing default rates as the income brackets increase
- These observations suggest a correlation between higher income levels and increased financial stability, leading to lower default rates. Borrowers in higher income brackets may have more disposable income and better capacity to repay loans.

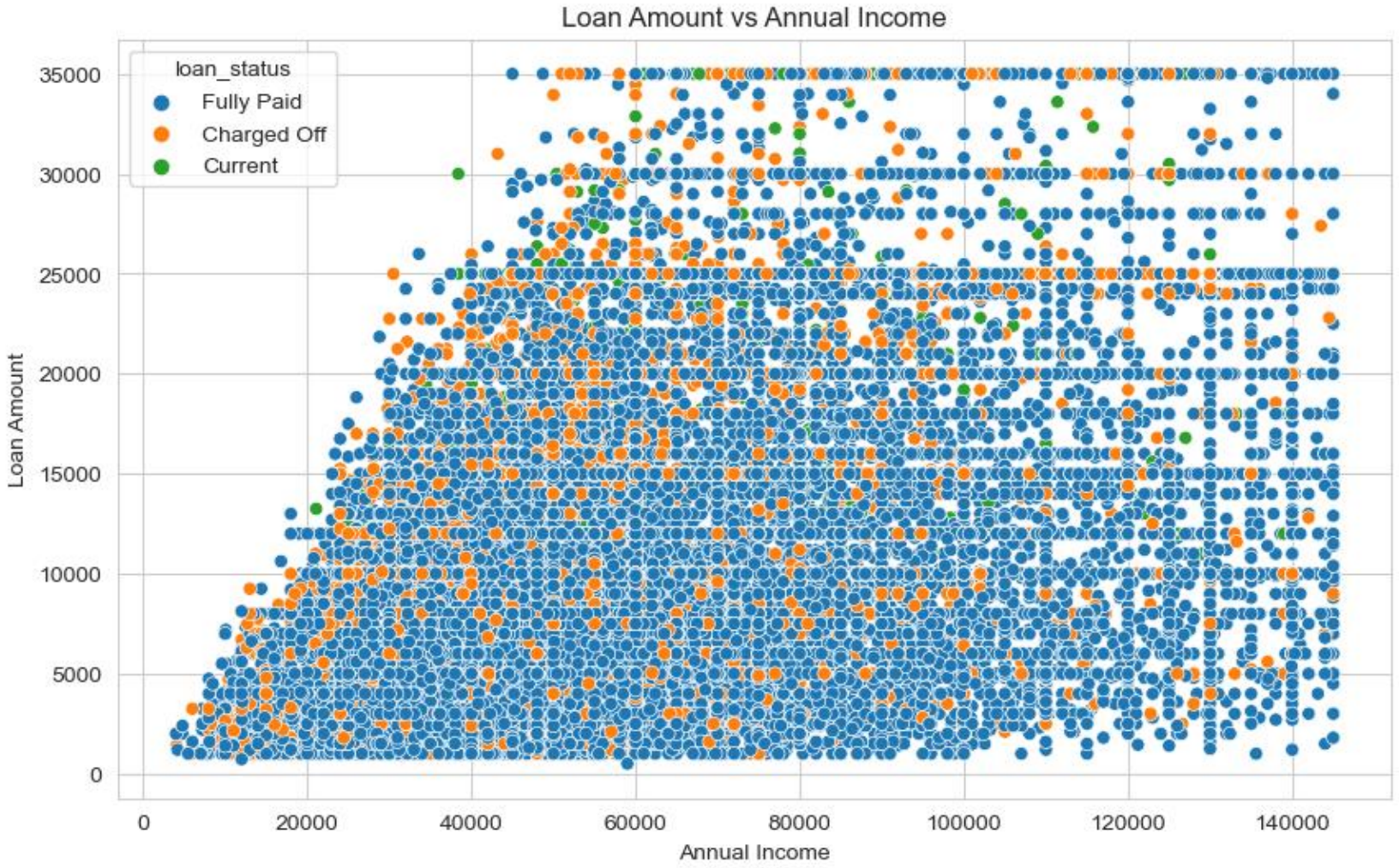
DATA ANALYSIS - BIVARIATE ANALYSIS

LOAN STATUS VS LOAN AMOUNT



1. Fully Paid Loans: The median loan amount for fully paid loans is lower compared to charged-off loans.
2. Charged-Off Loans: Charged-off loans tend to have higher loan amounts, indicating that higher loan amounts may be associated with a higher risk of default.
3. Current Loans: The whisker has a longer stretch with some loans extending up to $35,000, indicating that larger loans are still in repayment.

ANNUAL INCOME VS LOAN AMOUNT

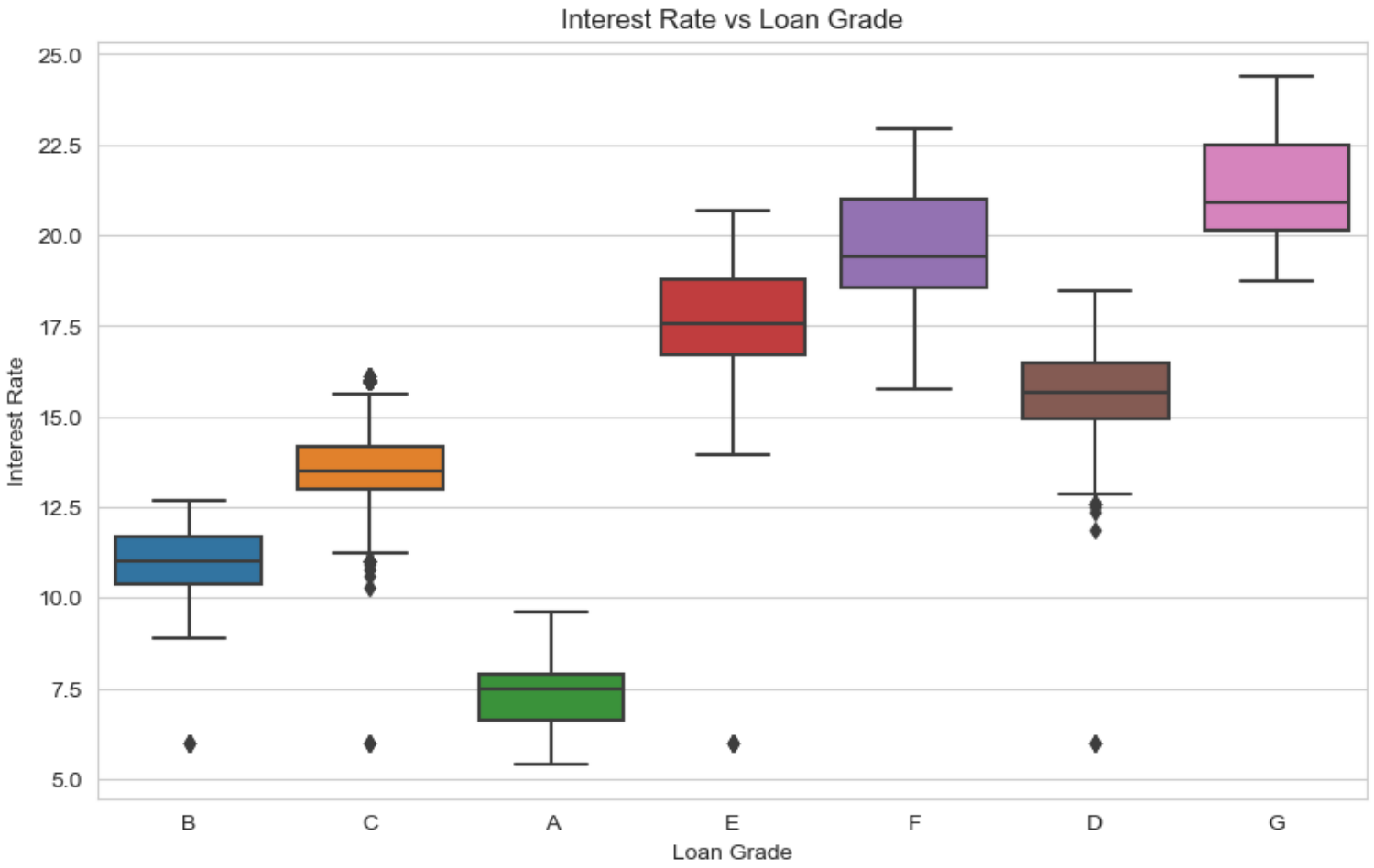


1. **Charged Off loans (orange points**) are spread across different income levels, but there is a noticeable concentration in lower income brackets and higher loan amounts.

2. There is a higher density of Charged Off loans in the lower and mid-income brackets ($20,000 to $60,000) and for loan amounts ranging from $5,000 to $20,000 suggesting that borrowers in these income brackets and loan amounts are at a higher risk of default.

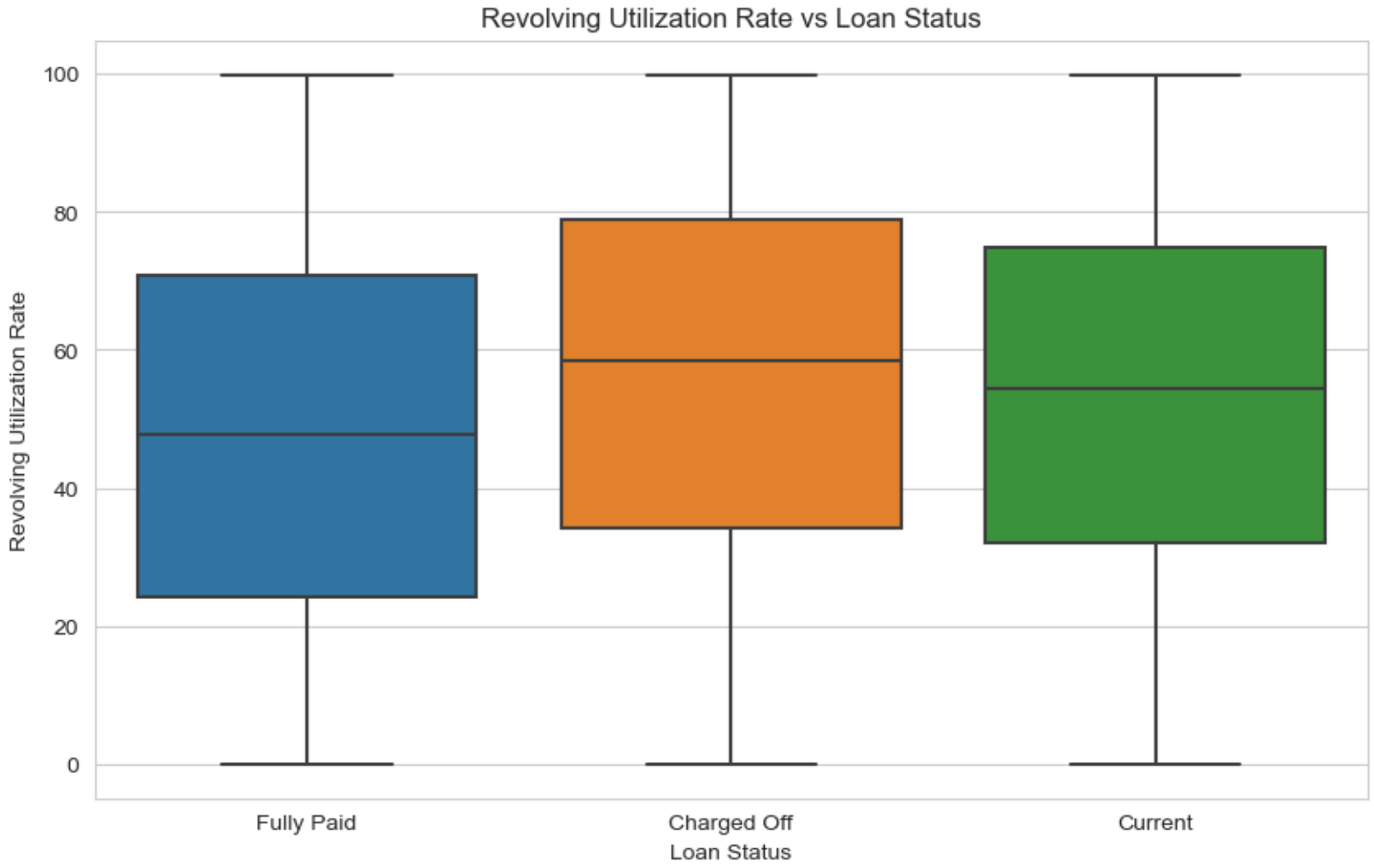
3. Borrowers with high income range >130000 tend to have more fully paid loans indicating better financial status and creditworthiness.

LOAN GRADE VS INTEREST RATE



- Lower grades (e.g., D, E, F) are associated with higher interest rates.
- The Loan applicants with loan Grade G are highest interest rated suggesting Loan Defaults.
- higher grades (A, B) have lower interest rates.
- The Loan applicants with loan Grade A are lowest interest rated suggesting lesser Loan Defaults.

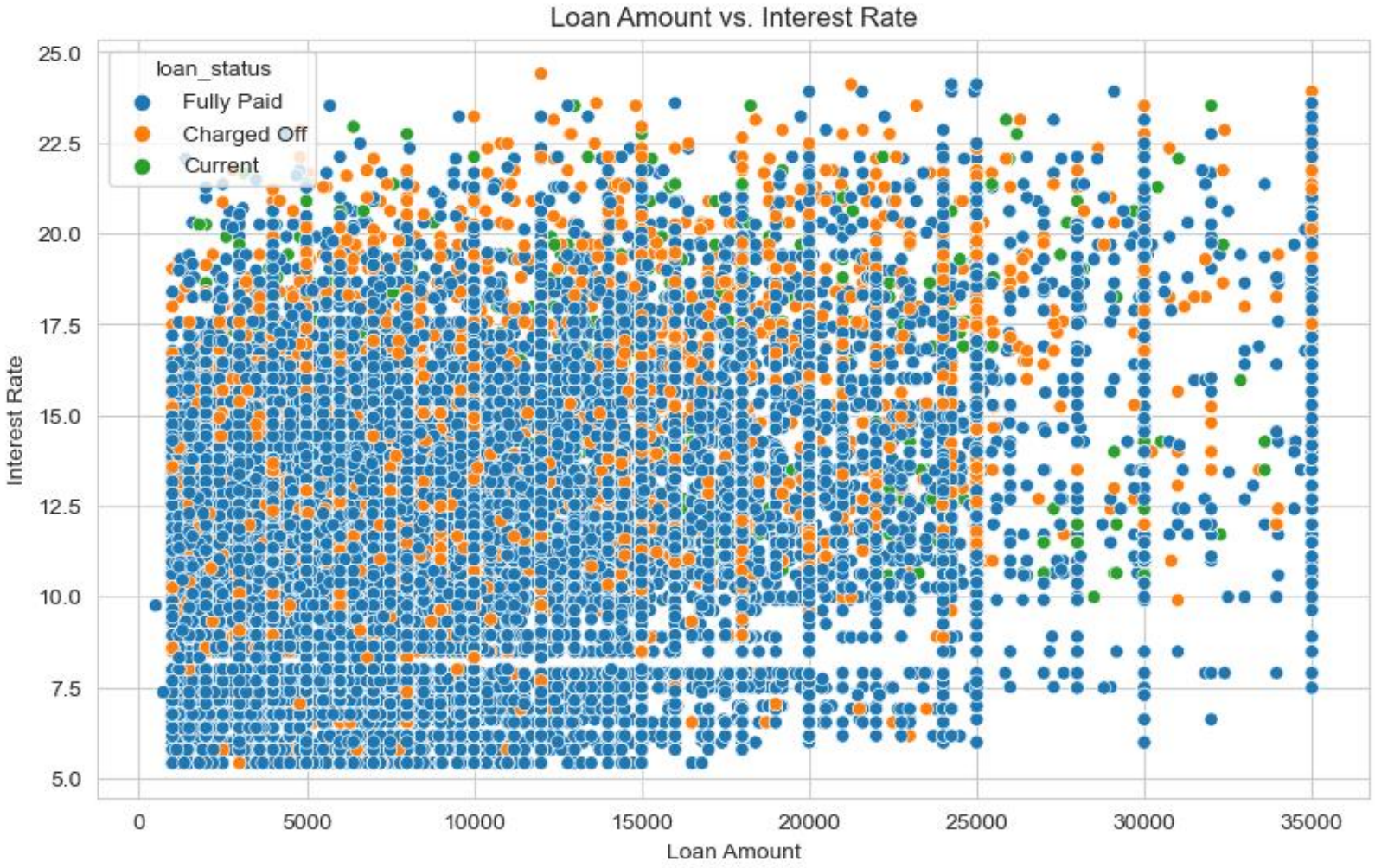
LOAN STATUS VS REVOLVING UTILIZATION RATE



Charged-off loans have higher median revolving utilization rates compared to fully paid and current loans, indicating borrower is heavily reliant on revolving credit accounts such as credit cards, which may affect their ability to repay loans.

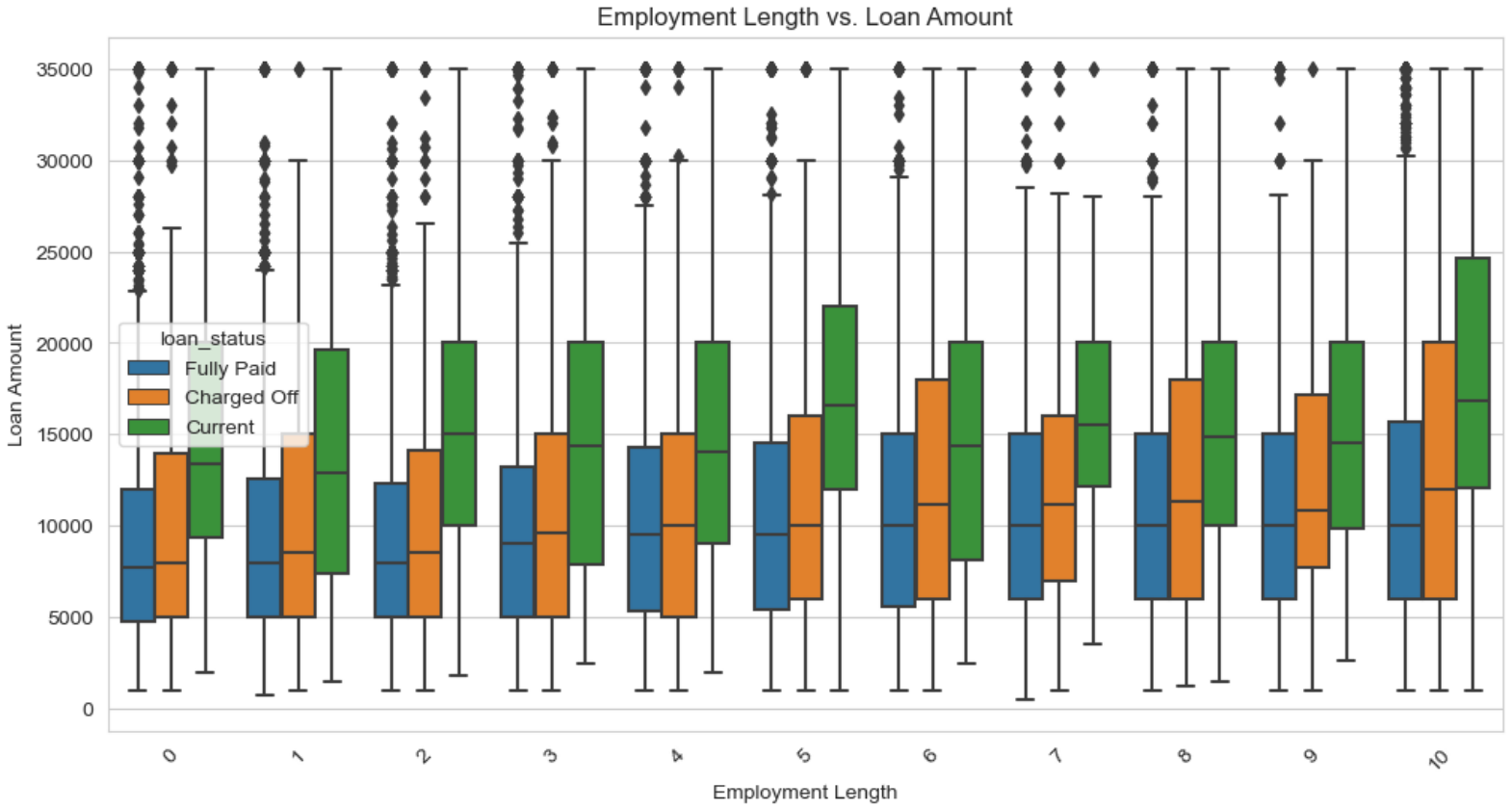
Fully Paid Loans tend to have lower revolving utilization rates, indicating better credit utilization management.

LOAN AMOUNT VS INTEREST RATE



- Loans that are "Charged Off" tend to have higher interest rates more frequently than "Fully Paid" loans, suggesting that higher interest rates might be associated with a higher risk of default.
- **This scatter plot illustrates that while higher interest rates might correlate with higher default risk, loan amounts do not have a strong influence on interest rates. This information is useful for understanding the risk profile and pricing strategies for loans.**

EMPLOYMENT LENGTH VS LOAN AMOUNT



- The median loan amount for **charged off loans** is higher than for fully paid loans leading to a higher likelihood of default.
- Borrowers with fully paid loans tend to have more consistent and typically lower loan amounts
- Current loans have higher medians than both fully paid and charged off loans indicating that borrowers are presently managing larger loans
- **Borrowers with "10+ years" of employment have taller box lengths. this suggests that stable, long-term employment may be associated with higher borrowing capacity**

DATA ANALYSIS - CORRELATION

CORRELATION HEATMAP

POSSITIVE CORRELATION

- Loan amount has positive correlation with annual income & revolving balance
- Interest rates has positive correlation with revolving utilization and loan amount
- DTI has positive correlation with open accounts

NEGATIVE CORRELATION

- DTI has a negative correlation with annual income
- Revolving utilization has a negative correlation with open accounts

DRIVER VARIABLES

The top parameter or driver variables for loan defaulters after thorough analysis through each univariate, segmented & multivariate analysis are below which company can utilize this knowledge for its portfolio and risk assessment.

- ❑ Loan Status
- ❑ Loan Grade
- ❑ Loan Purpose
- ❑ loan amount
- ❑ Home Ownership
- ❑ Annual Income
- ❑ Revolving balance
- ❑ Interest rate
- ❑ Revolving utilization rate
- ❑ DTI categories
- ❑ loan to income ratio

RESULTS/CONCLUSIONS IN BUSINESS TERMS

- Small Business loans have the highest default rate (~25.98%) followed by Renewable Energy and Educational loans showing significant default rates (~18.45% and ~17.23%, respectively).
- Borrowers categorized by home ownership under "Other" category have the highest default rate (~18.37%), followed by Renters (15.02%).
- Higher DTI ratios are associated with a higher likelihood of default.
- Borrowers with higher loan to income ratios are at a higher risk of default.
- Charged Off loans are more frequent in lower income brackets with higher loan amounts.
- Charged Off loans are more frequent at higher interest rates.
- Charged-off loans have higher median revolving utilization rates compared to fully paid and current loans.
- Higher DTI ratios are associated with a higher presence of Charged Off loans.
- The median loan amount for charged off loans is higher than for fully paid loans.
- Numerous outliers in loan amount indicate that some borrowers take out significantly larger loans regardless of employment length.
- States like CA, NY, and TX (Texas) show higher median loan amounts and wider interquartile ranges (IQR), indicating both higher borrowing and greater variability in loan amounts.
- There are more outliers in 60-month loans Term, suggesting that loan amounts for 60-month loans vary more significantly.

IDENTIFYING LIKELY DEFAULTERS

To identify likely defaulters, consider the following steps:

- ❑ High-Risk Loan Purposes
  - Prioritize monitoring and stricter approval processes for loans taken for Small Business, Renewable Energy, and Educational purposes.
- ❑ High DTI Ratios:
  - Identify borrowers with high DTI ratios and implement measures such as higher interest rates, additional collateral requirements, or denial of loan applications.
- ❑ High Loan to Income Ratios:
  - Set thresholds for loan to income ratios and decline applications that exceed these thresholds.
- ❑ Loan Amounts Relative to Income:
  - For low-income borrowers taking out high loan amounts, ensure rigorous credit checks and possibly higher interest rates or guarantees.
- ❑ Employment Stability:

  Give preference to borrowers with longer employment histories, but not at the expense of overlooking other high-risk indicators.
- ❑ Segmentation and Targeting:
  - Segment borrowers into risk categories based on the combination of these factors and design targeted strategies to mitigate default risks.

RECOMMENDATIONS

- Lenders could use this information to adjust their risk management strategies. For example, they might require stricter credit checks or additional guarantees for borrowers in the lower income brackets.
- Providing financial education and support to borrowers in lower income brackets to help them manage their finances better and reduce the risk of default.
- Monitor loans issued for high-risk purposes closely and consider more stringent approval criteria.
- Apply higher scrutiny to renters and those with other non-standard home ownership statuses.
- Factor in employment stability when assessing loan applications but consider it alongside other variables.
- Consider implementing stricter loan approval criteria for borrowers with high loan to income ratios to mitigate potential risks, as they might be more prone to default
- Regularly review and monitor the DTI ratios of borrowers to ensure early intervention if they move towards higher debt levels.
- The strong correlation between loan amount and loan to income ratio can be used for risk assessment. Higher ratios may indicate higher risk and warrant closer scrutiny.
- The moderate positive correlation between interest rate and loan to income ratio suggests that lenders might adjust interest rates based on the borrower's debt level relative to their income.
- Consider adjusting loan terms, such as interest rates and repayment periods, based on the borrower's income level and loan amount to manage risk better.
- Monitor borrowers with higher DTI ratios, open accounts, and revolving balances closely, as these factors are associated with higher risk profiles.
- Use the insights from the scatter plot to segment borrowers and design targeted strategies for each segment, such as offering financial counseling for high-risk segments.