Project Name

EDA - Landing Club case study

Table of Contents

General Information

- Provide general information about your project here.

  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.

- What is the background of your project?

  Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

  If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- What is the business problem that your project is trying to solve?

  Understanding the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Problem Statement:

- Conduct a comprehensive analysis of a dataset containing consumer attributes and loan attributes. Our goal is to gain insights into the factors influencing loan default rates and to develop strategies to mitigate risks associated with lending.
- What is the dataset that is being used?

  The dataset contains information about past loan applicants and whether they ‘defaulted’ or not. 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, etc.

Objective

- Importing necessary Modules

  Import the modules necessary for Data Manipulation and Visualization.
- Reading dataset

  Read the dataset containing loan applicant information.
- Exploring the Dataset

  Understand the Structure and various datatypes of the attributes within the dataset.
- Missing value analysis

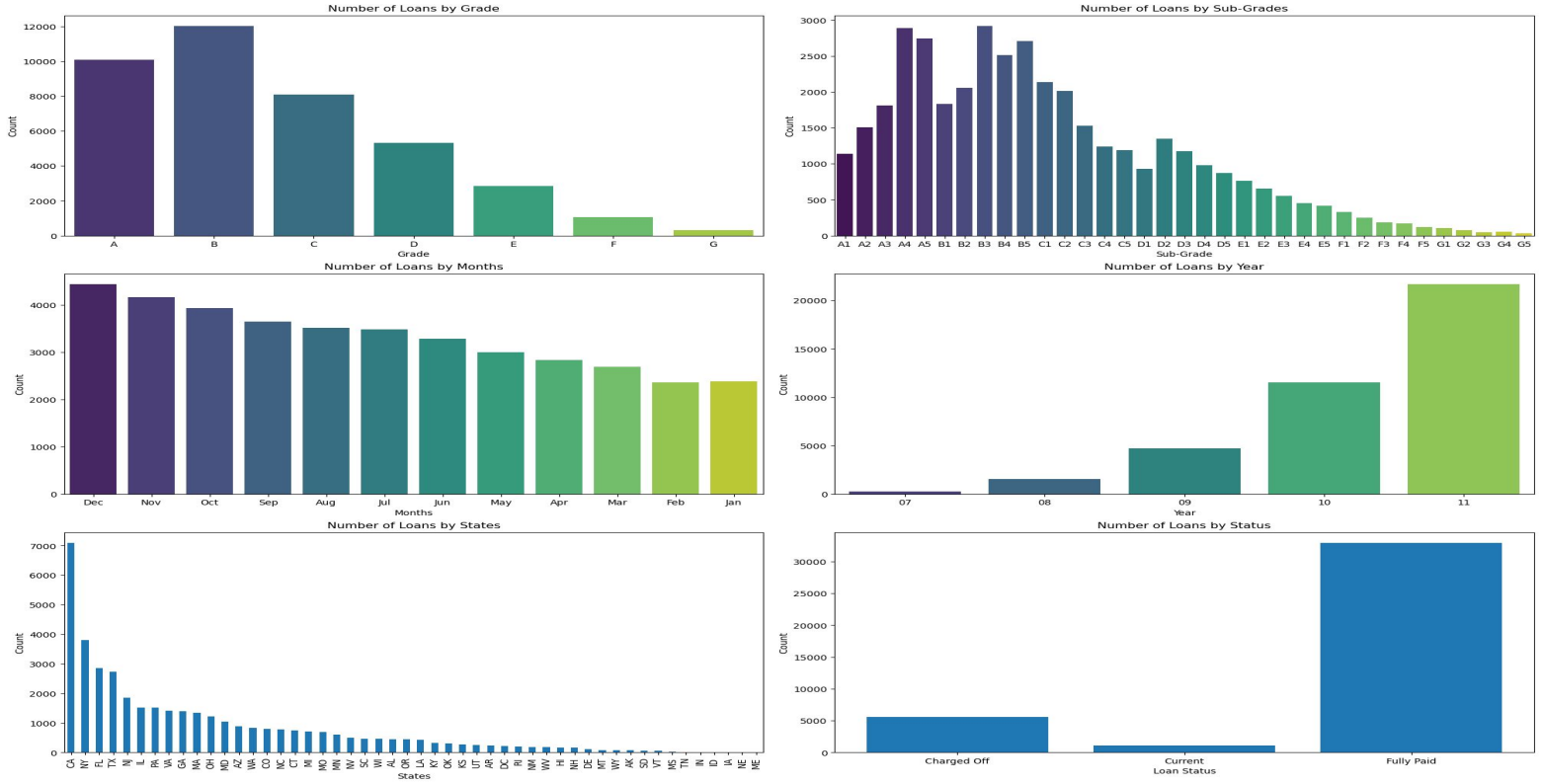
  Identify and analyze missing values in the dataset.
- Analysing categorical and numerical columns

  Analyze categorical and numerical columns to understand the statistical properties and relationships within the dataset.
- Univariate Analysis:

  Conduct univariate analysis to explore the distribution and characteristics of individual variables.
- Outliers:
  - Identify and analyze outliers within the dataset to understand their impact on the analysis.
- Bivariate analysis:
  - Conduct bivariate analysis to explore relationships between different variables and their impact on loan default rates.

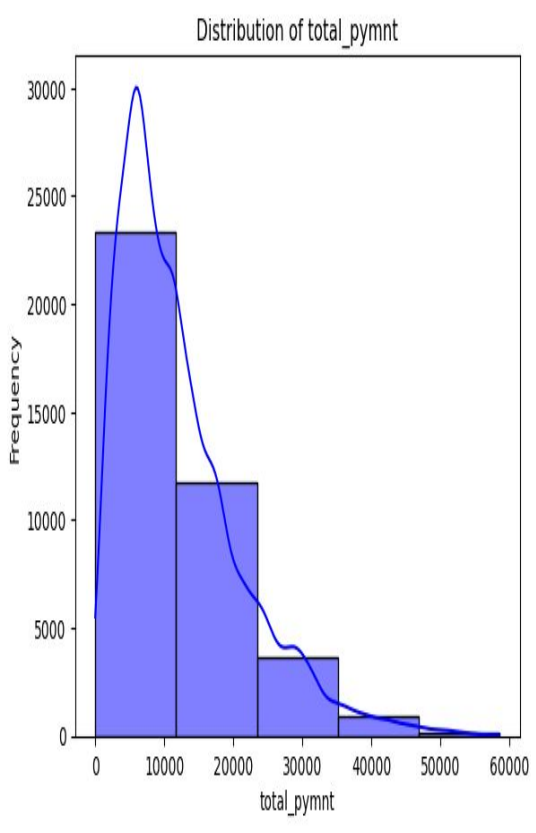
Conclusions

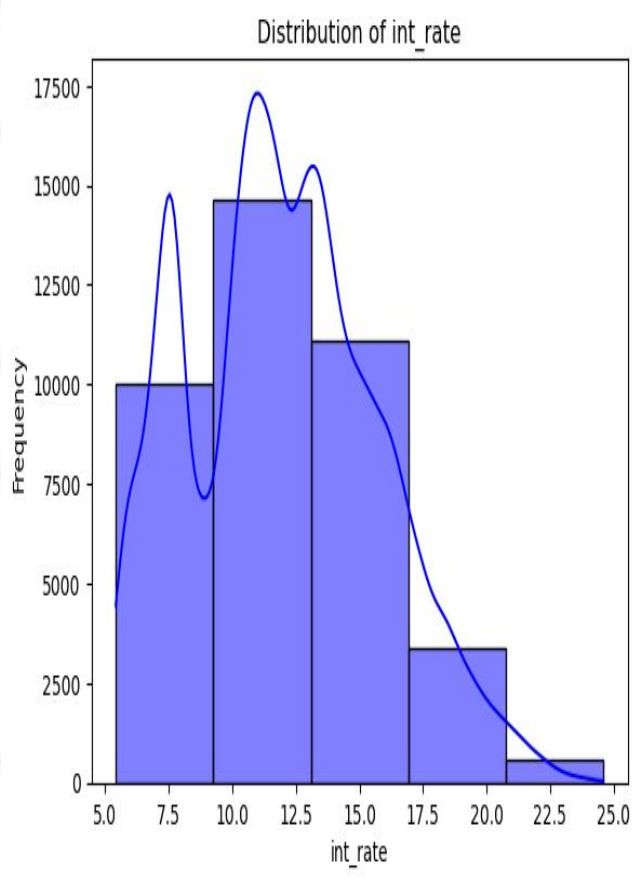
The Categorical Univariate Analysis

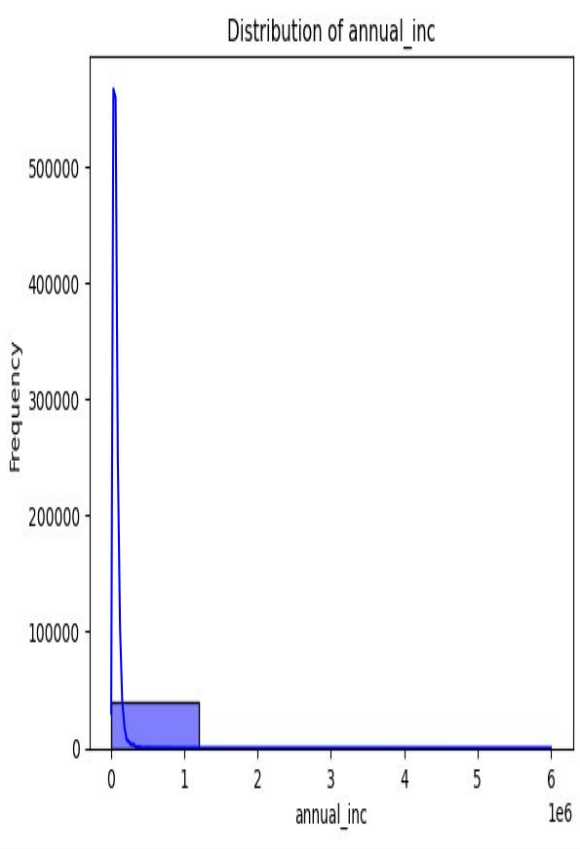


- Most of the loans comes under Grade and Subgrade A, B and C.
- Majority of loans had been issued in year 2011 and Jun-Dec months.
- California and Newyork state has the highest no of loans issued.

The Univariate Numerical Analysis 2

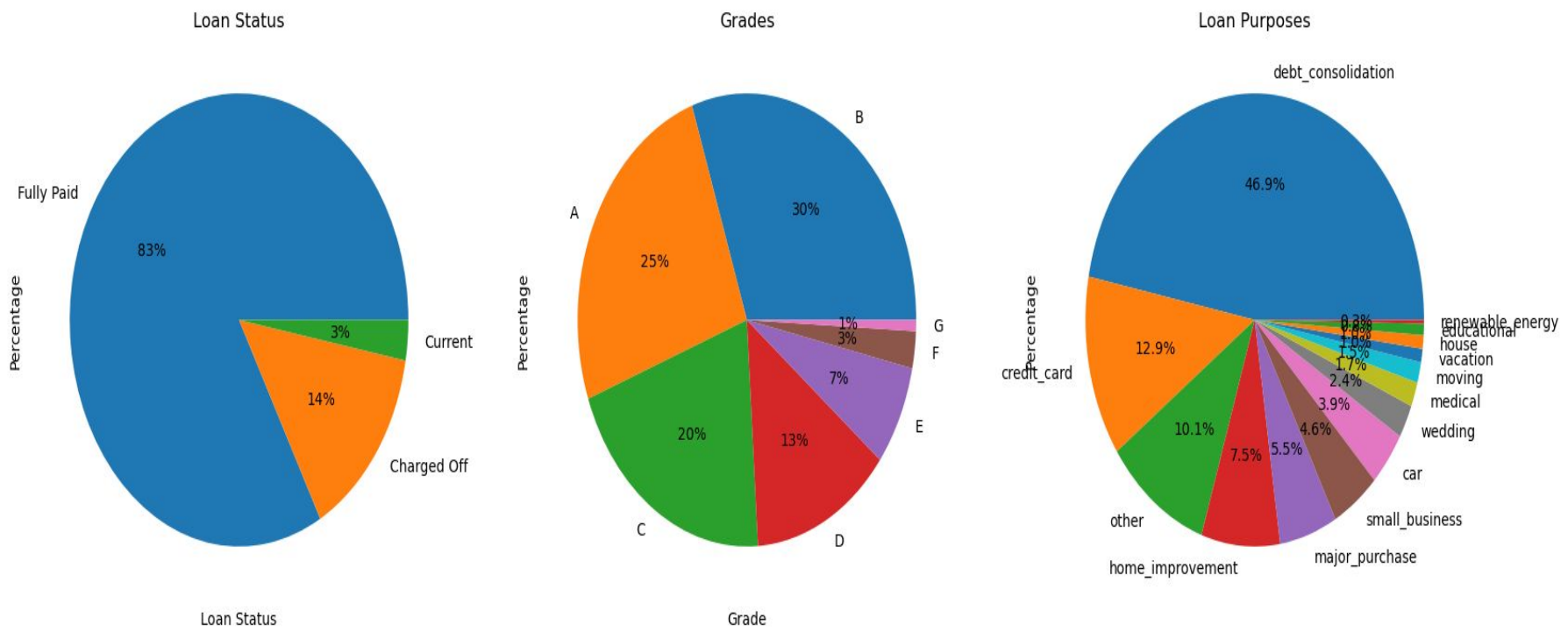






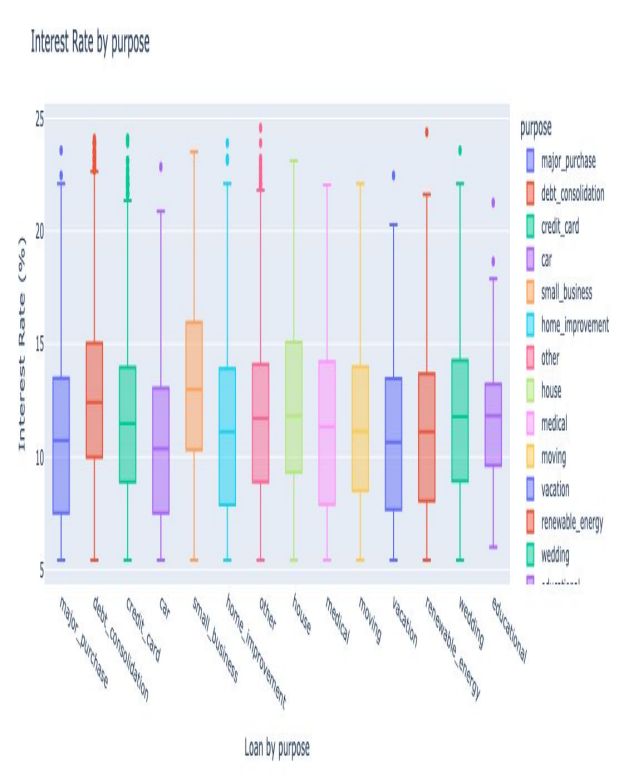
- **High Loan Amounts:** The right-skewed distribution of loan_amnt suggests that a significant proportion of loans are for larger amounts. Larger loans generally carry higher risk due to the potential for larger losses in case of default.
- **High Interest Rates:** The right-skewed distribution of int_rate indicates that a portion of loans have significantly higher interest rates. These loans are often associated with higher risk borrowers and may have a higher likelihood of default.
- **Income Distribution:** The right-skewed distribution of annual_inc suggests that a majority of borrowers have lower incomes. Borrowers with lower incomes may have limited capacity to repay loans, particularly if they have taken on larger amounts or have high interest rates.

Conclusion from Pie Plot (Loan Status, Grades and Purpose) Categorical Univariate Analysis 3



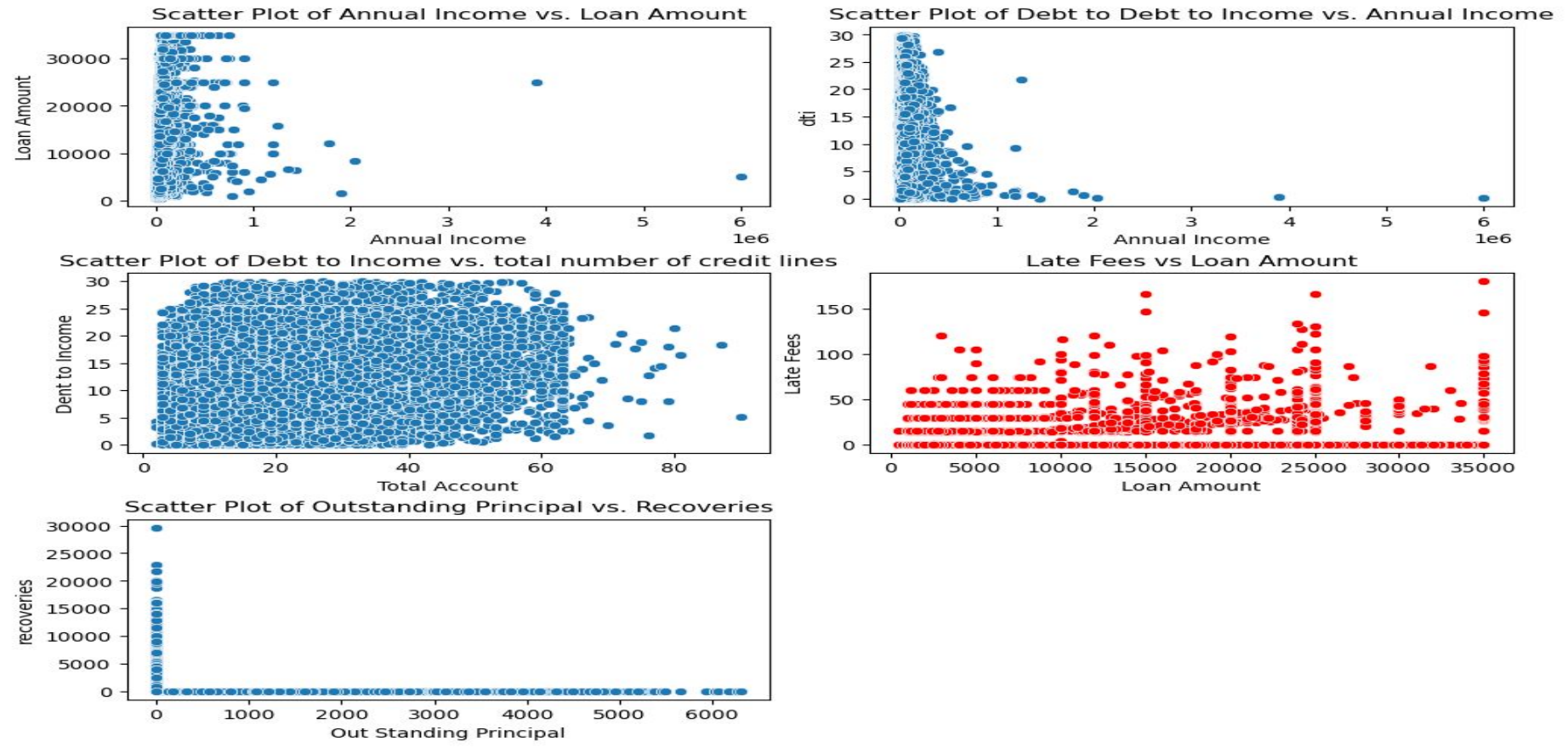
- This visualization can help us understand the proportion of loans that have been fully paid, charged off, or are currently in progress.
- We can see that the majority of loans have been fully paid.
- Most of the loans are for debt consolidation

Box Plot (Interest Rate by Grade, Sub Grade and Loan Status) Byvariate Categorical Analysis 4



- Current and default loan status have highr interest rates than fully paid.
- Analysis reveals significant variation in interest rates depending on the loan purpose. Notably, Debt Consolidation and Small Business loans tend to have higher interest rates compared to other loan categories.

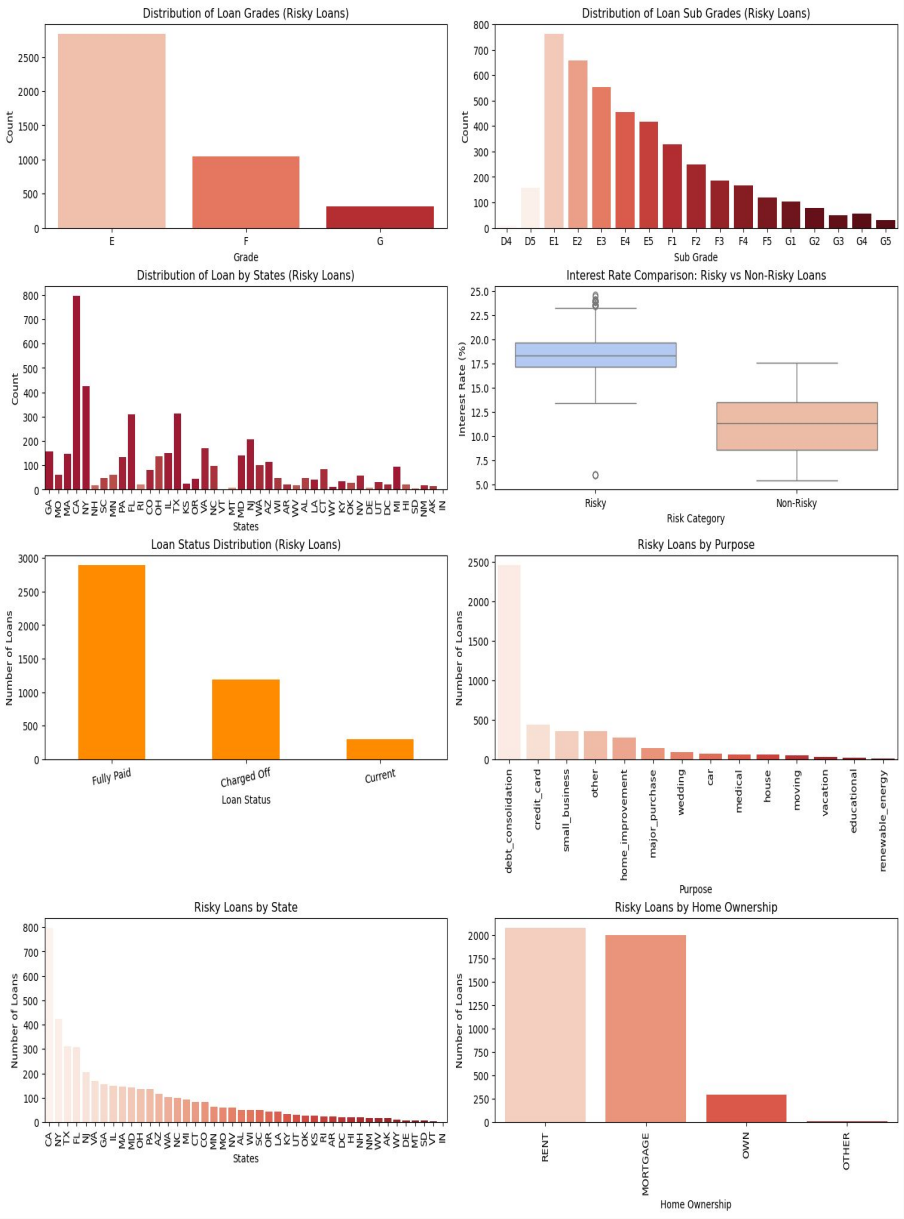
Scatter Plot Numerical Byvariate Analysis 5



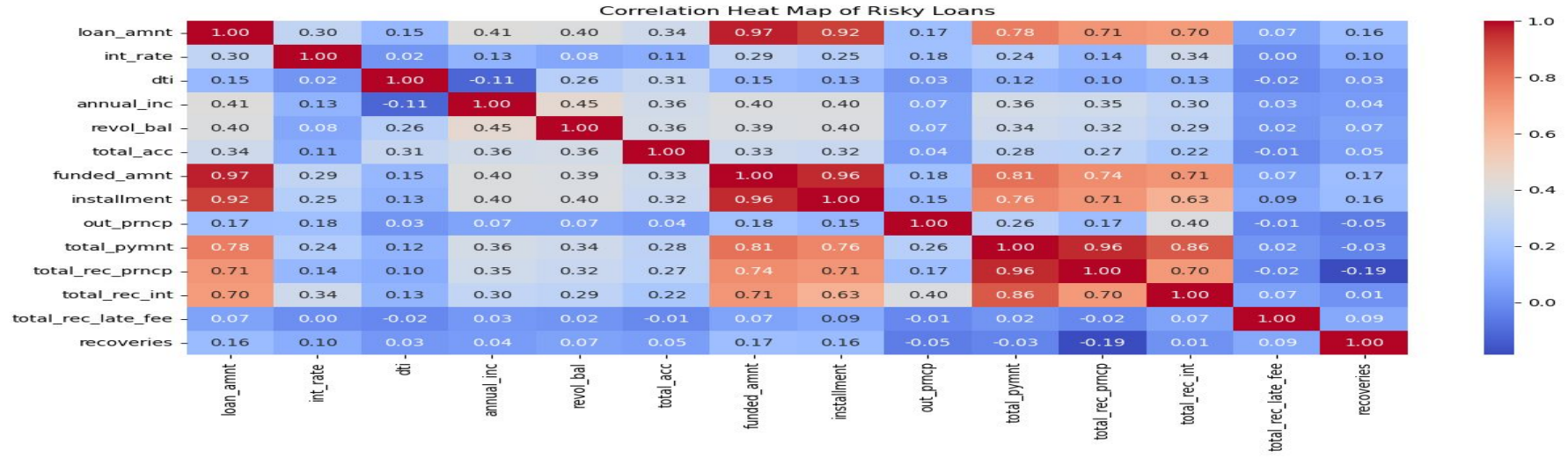
- The plot suggests no strong relationship between annual income and loan amount. Borrowers with similar incomes might request vastly different loan amounts.
- By analyzing these metrics together, you can assess the likelihood of a loan becoming risky or defaulting. High late fees, outstanding principal, and low recoveries, it signals a larger risk to lenders and investors.

Univariate and Byvariate Analysis 6 (Risky Loans)

- **Loan Grades & Sub-Grades:** The distribution shows a concentration of risky loans in the lower grades (E, F, G) and their corresponding sub-grades. This aligns with expectations as lower grades typically represent higher risk.
- **Loan Status:** A significant portion of risky loans are either "Charged Off" (defaulted) or still "Current" (ongoing). This highlights the ongoing risk associated with these loans.
- **Interest Rates:** Risky loans consistently have higher interest rates compared to non-risky loans, indicating the lender's assessment of increased risk.
- **Loan Purposes:** Debt consolidation and credit card loans appear to be the most common purposes for risky loans. This suggests that borrowers struggling with existing debt may be more likely to take on higher-risk loans.
- **State Distribution:** The distribution of risky loans across states is not uniform. Some states have a higher concentration of risky loans compared to others, potentially indicating regional differences in risk factors.



Correlation Heat Map (Numerical Features) Bivariate Analysis



- Negative correlation between DTI and annual income (-0.11):
- As income increases, the DTI ratio tends to decrease. A high DTI is a strong indicator of higher risk because it implies the borrower is already carrying a significant debt load relative to their income, which increases the likelihood of loan default.
- The weak positive correlation with loan amount (0.15) and revolving balance (0.26) also shows that higher loan amounts and higher balances tend to go hand-in-hand with a higher DTI, which is a risk factor for default.
- Correlation between loan_amnt and other variables:

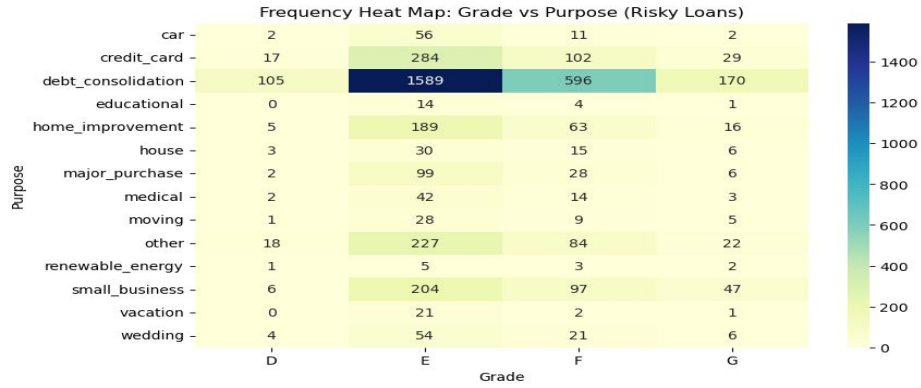
  The loan amount (loan_amnt) has a moderate positive correlation with the interest rate (int_rate) at 0.30 and a slightly higher positive correlation with annual income (annual_inc) at 0.41. It also has moderate positive correlations with revolving balance (revol_bal) and total accounts (total_acc).
- Interest rate correlations:

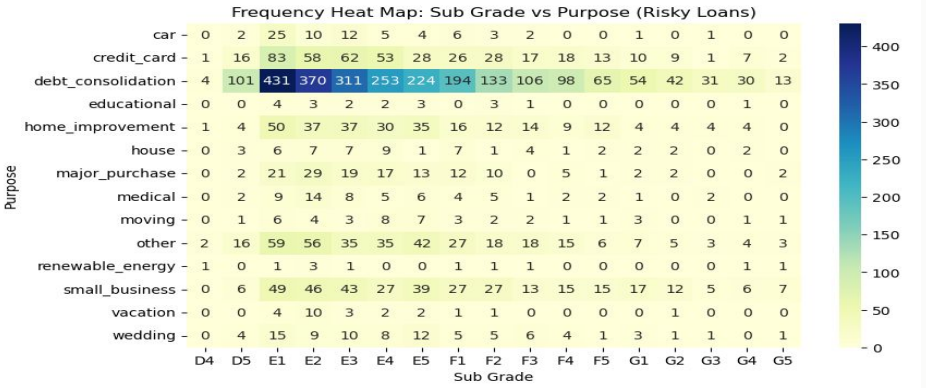
  The interest rate (int_rate) has a low positive correlation with loan amount (loan_amnt), but it has very low correlations with other variables like annual income (annual_inc) and revolving balance (revol_bal).

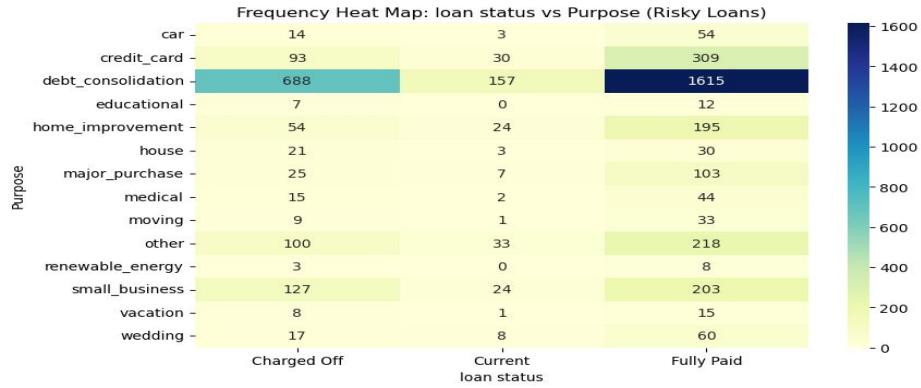
  The heatmap reveals that loan size, DTI, and revolving balance are strong indicators of risk. Larger loans, higher DTI ratios, and high revolving balances tend to increase the likelihood of default, while high income and low DTI ratios can mitigate this risk. Lenders should focus on these variables when assessing borrower risk to better predict the likelihood of loan repayment.

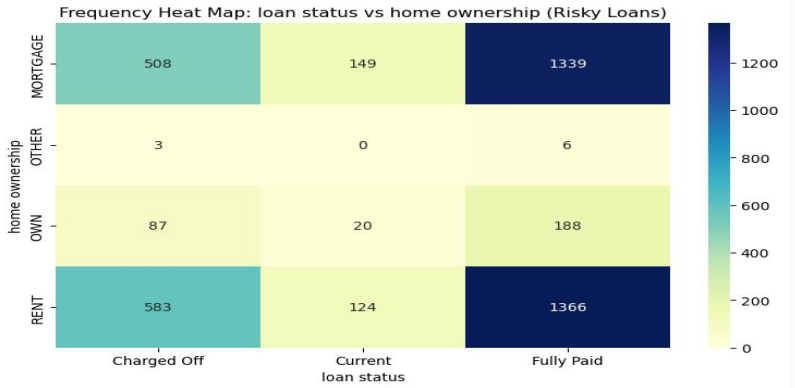
.

Correlation Heat Map (Categorical Features) Analysis



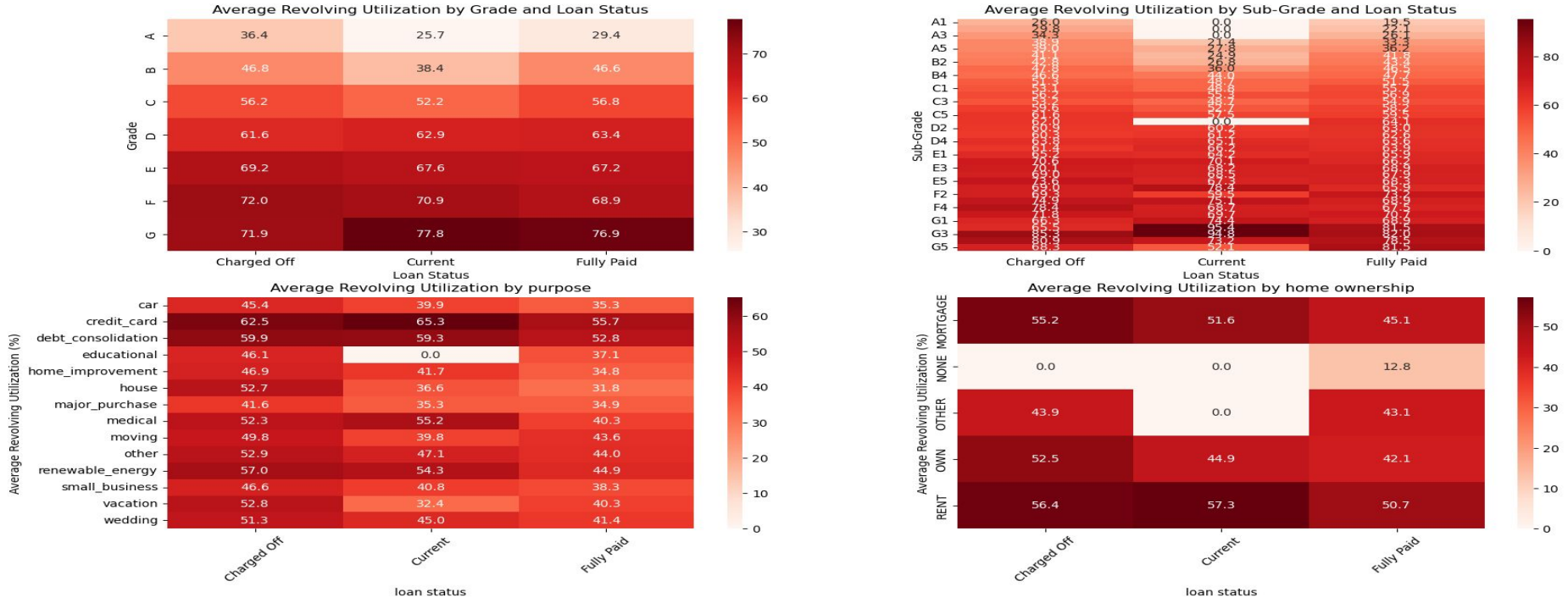






- Debt consolidation is strongly associated with higher-risk borrowers (grades D, E, F, G, sub-grades D4, D5). Borrowers in these groups are likely facing financial struggles, making them higher-risk for loan defaults.
- Debt consolidation is the primary purpose driving risky loans, often concentrated among renters and mortgaged homeowners.
- Renters exhibit higher risk levels compared to homeowners, which aligns with their larger proportion of "Charged Off" loans.

The Revolving Utilization Heatmap Analysis 7

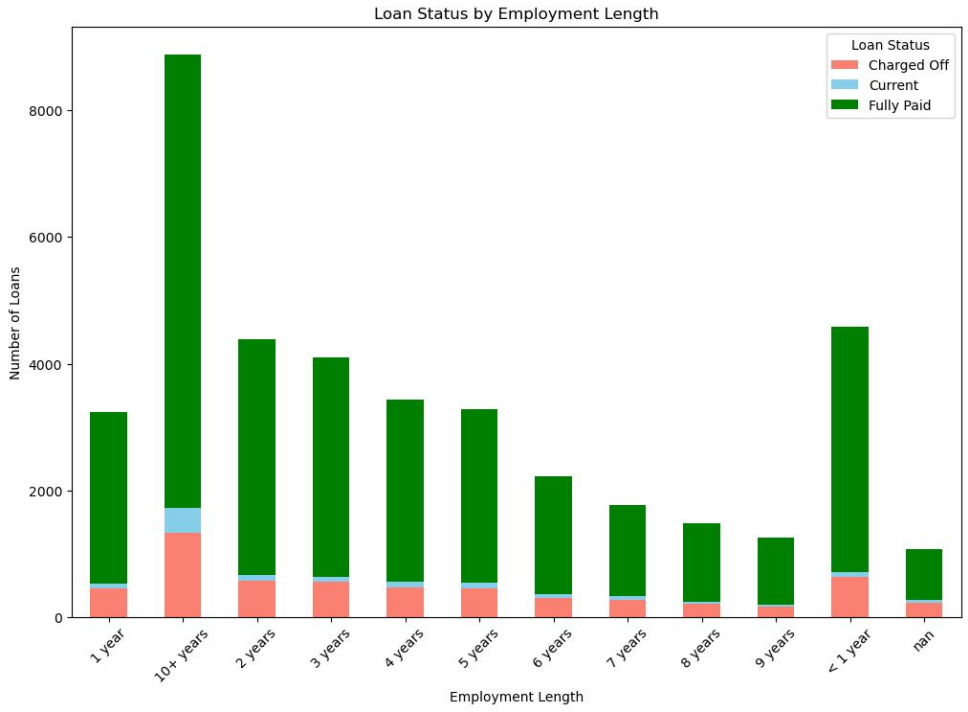


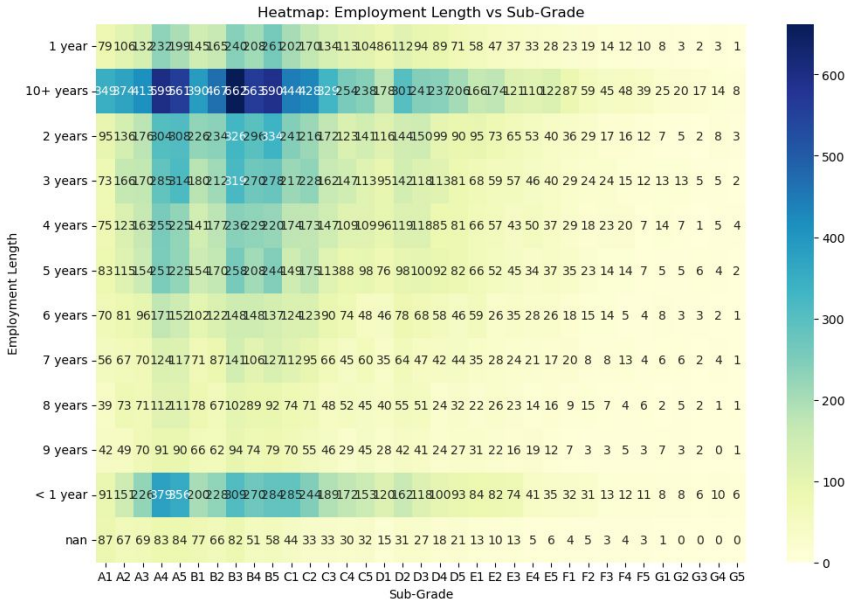
- Revolving Utilization and Loan Status: The heatmaps clearly show that revolving utilization tends to be higher for "Charged Off" loans compared to "Current" and "Fully Paid" loans across various categories (grade, sub-grade, purpose, home ownership).
- Grade and Sub-Grade: Revolving utilization generally increases as the loan grade and sub-grade deteriorate (from A to G). This is consistent with the expectation that higher-risk borrowers tend to have higher credit utilization.
- Loan Purpose: The heatmaps reveal that some loan purposes, such as "credit_card" and "debt_consolidation," tend to have higher revolving utilization compared to others. This suggests that borrowers taking out these types of loans might have higher existing debt burdens.
- Home Ownership: The heatmap for home ownership shows some interesting patterns. For instance, borrowers with "RENT" status seem to have higher revolving utilization compared to those with "MORTGAGE" or "OWN".

Box Plot (Loan Status vs Revolving Utilization) Analysis

- Revolving Utilization and Loan Status: The box plot clearly shows that "Charged Off" loans tend to have significantly higher revolving utilization compared to "Fully Paid" and "Current" loans. This suggests that borrowers with higher credit card debt relative to their credit limit are more likely to default on their loans

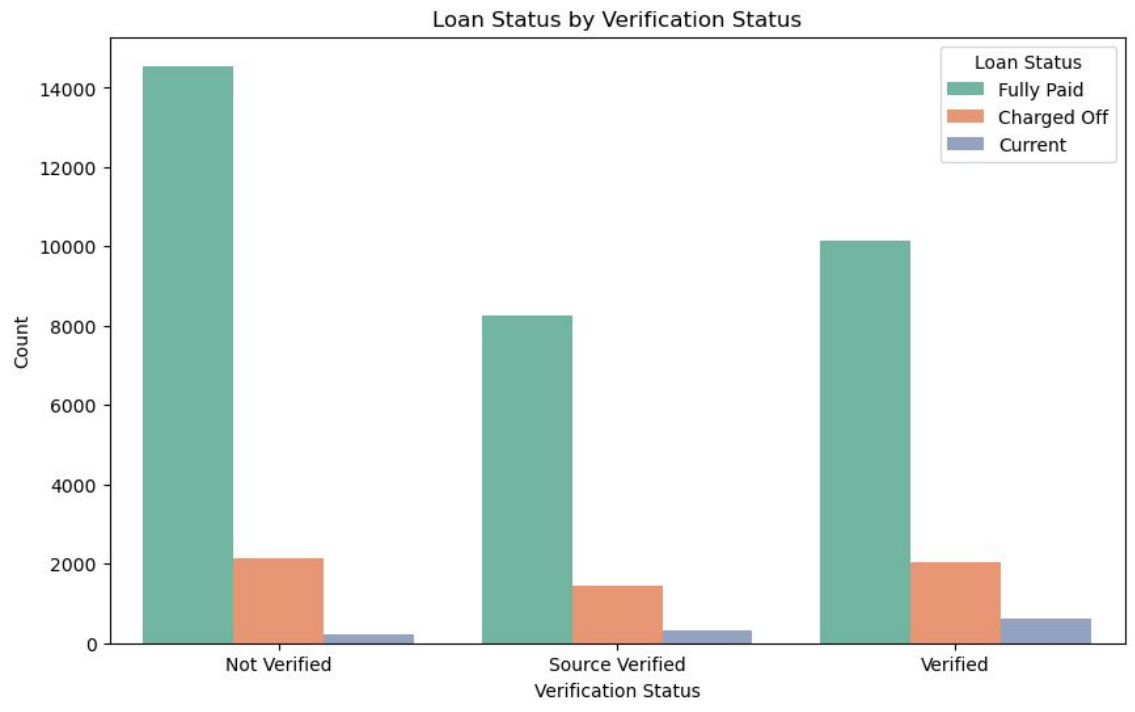
Bar Plot (Loan Status by Employment Length, Verification Status etc) analysis 8

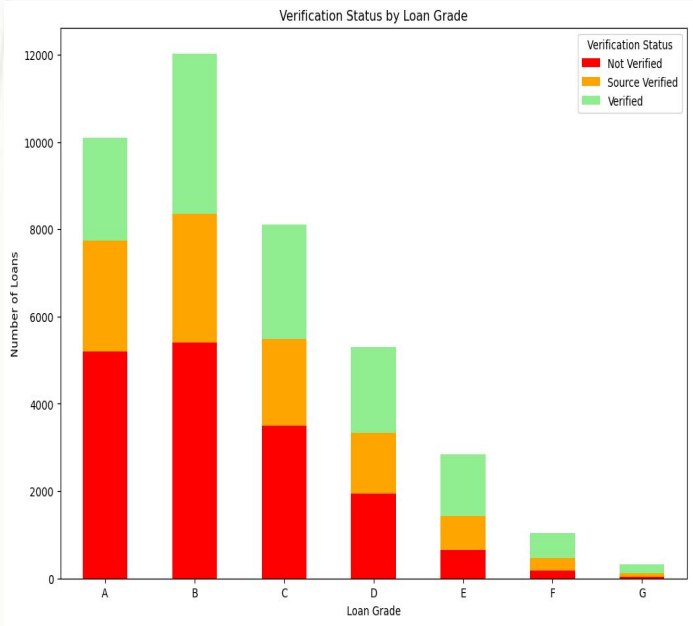




- Revolving Utilization and Loan Status: The heatmaps clearly show that revolving utilization tends to be higher for "Charged Off" loans compared to "Current" and "Fully Paid" loans across various categories (grade, sub-grade, purpose, home ownership).
- Grade and Sub-Grade: Revolving utilization generally increases as the loan grade and sub-grade deteriorate (from A to G). This is consistent with the expectation that higher-risk borrowers tend to have higher credit utilization.
- Loan Purpose: The heatmaps reveal that some loan purposes, such as "credit_card" and "debt_consolidation," tend to have higher revolving utilization compared to others. This suggests that borrowers taking out these types of loans might have higher existing debt burdens.
- Home Ownership: The heatmap for home ownership shows some interesting patterns. For instance, borrowers with "RENT" status seem to have higher revolving utilization compared to those with "MORTGAGE" or "OWN".

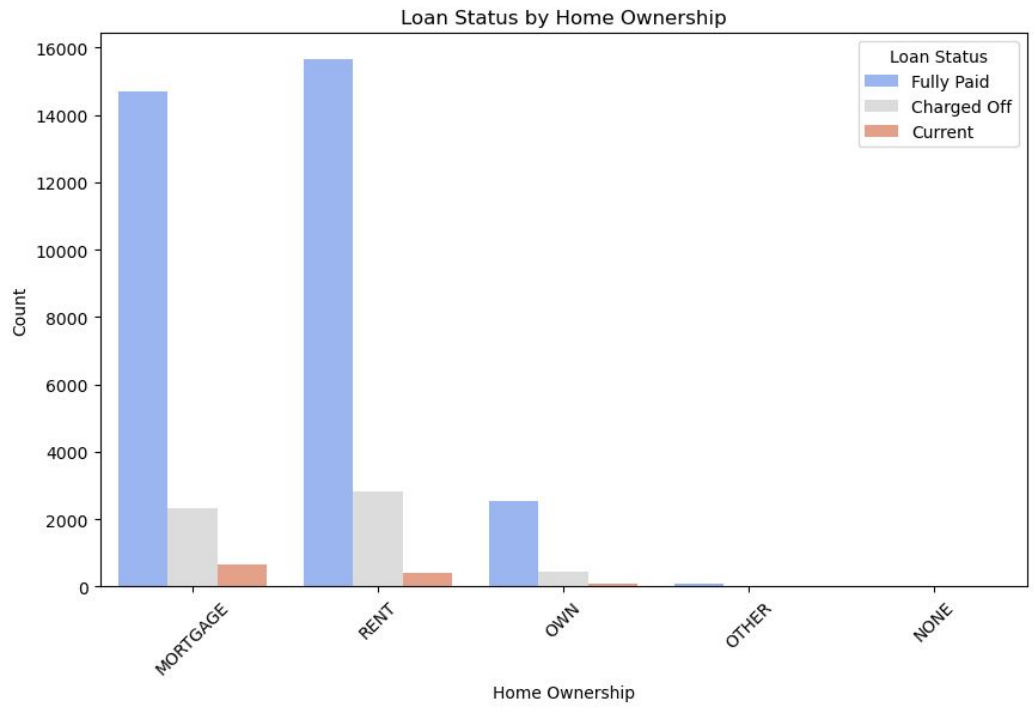
Bar Plot (Loan Status by Employment Length, Verification Status etc) analysis 8

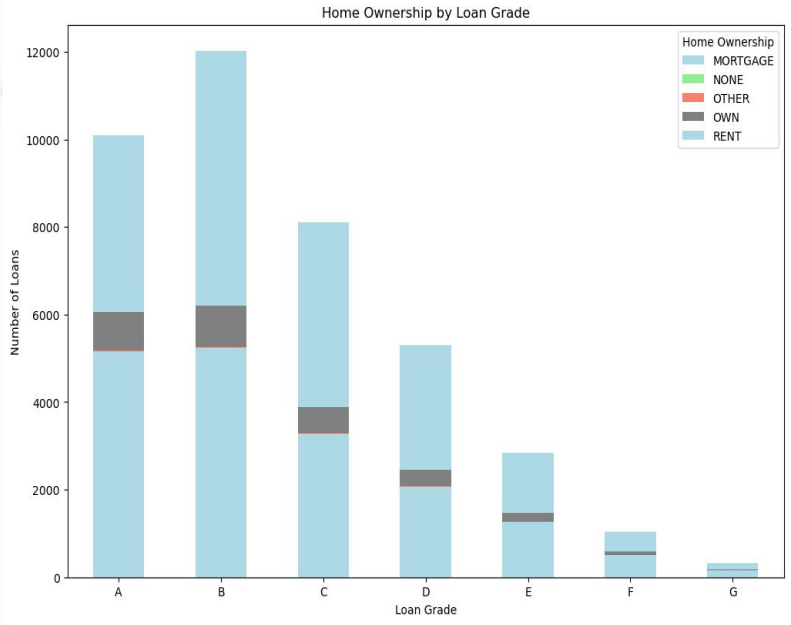




- Revolving Utilization and Loan Status: The heatmaps clearly show that revolving utilization tends to be higher for "Charged Off" loans compared to "Current" and "Fully Paid" loans across various categories (grade, sub-grade, purpose, home ownership).
- Grade and Sub-Grade: Revolving utilization generally increases as the loan grade and sub-grade deteriorate (from A to G). This is consistent with the expectation that higher-risk borrowers tend to have higher credit utilization.
- Loan Purpose: The heatmaps reveal that some loan purposes, such as "credit_card" and "debt_consolidation," tend to have higher revolving utilization compared to others. This suggests that borrowers taking out these types of loans might have higher existing debt burdens.
- Home Ownership: The heatmap for home ownership shows some interesting patterns. For instance, borrowers with "RENT" status seem to have higher revolving utilization compared to those with "MORTGAGE" or "OWN".

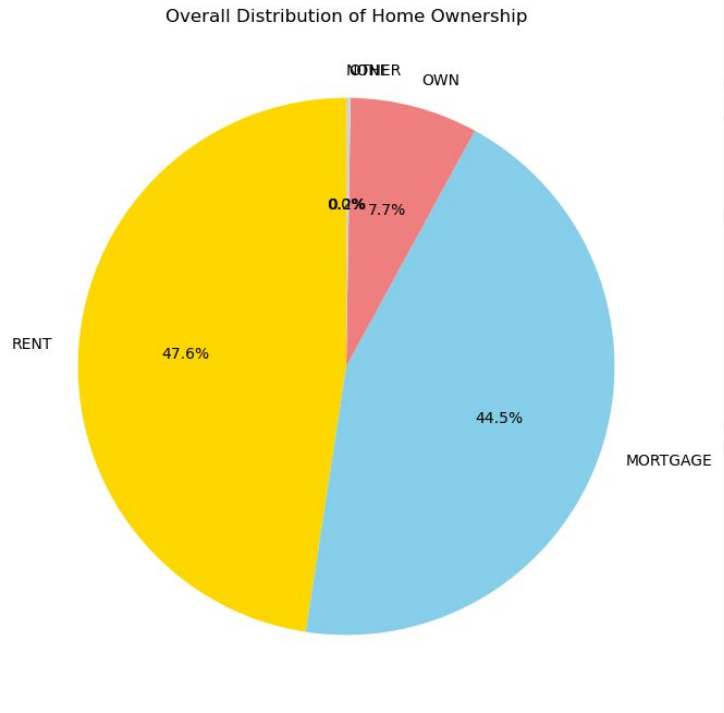
Conclusion from Bar Plot (Loan Status by Home Ownership) Analysis 9

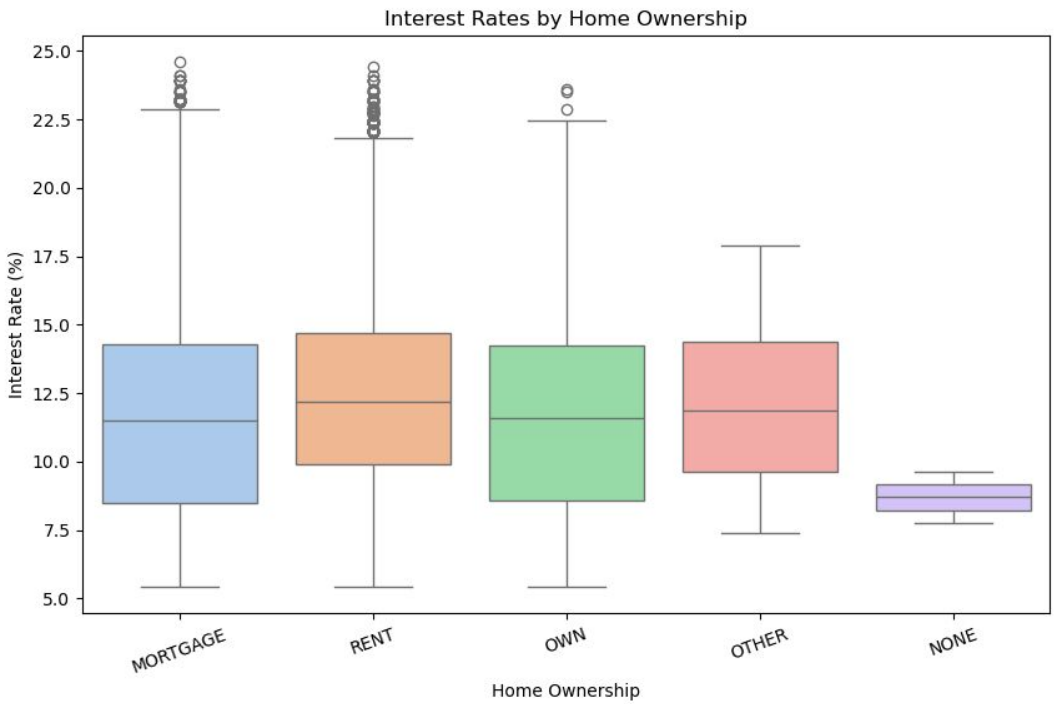




- Home Ownership and Interest Rates: The box plot suggests a general trend that individuals with "NONE" or "OTHER" home ownership status tend to have lower interest rates compared to those who mortgage, rent, or own their homes.
- Variability: There's significant variability in interest rates across all home ownership categories. This indicates that other factors besides home ownership likely influence interest rates.
- Outliers: The presence of outliers suggests that some individuals within each category have interest rates that deviate significantly from the typical range.
- Home Ownership: The chart suggests that individuals who rent or own their homes might have lower credit risk compared to those with "OTHER" or "NONE" home ownership status. This is because renting or owning a home often indicates financial stability and responsibility.
- Loan Status: The presence of "Charged Off" loans across all home ownership categories indicates that credit risk exists regardless of home ownership status.

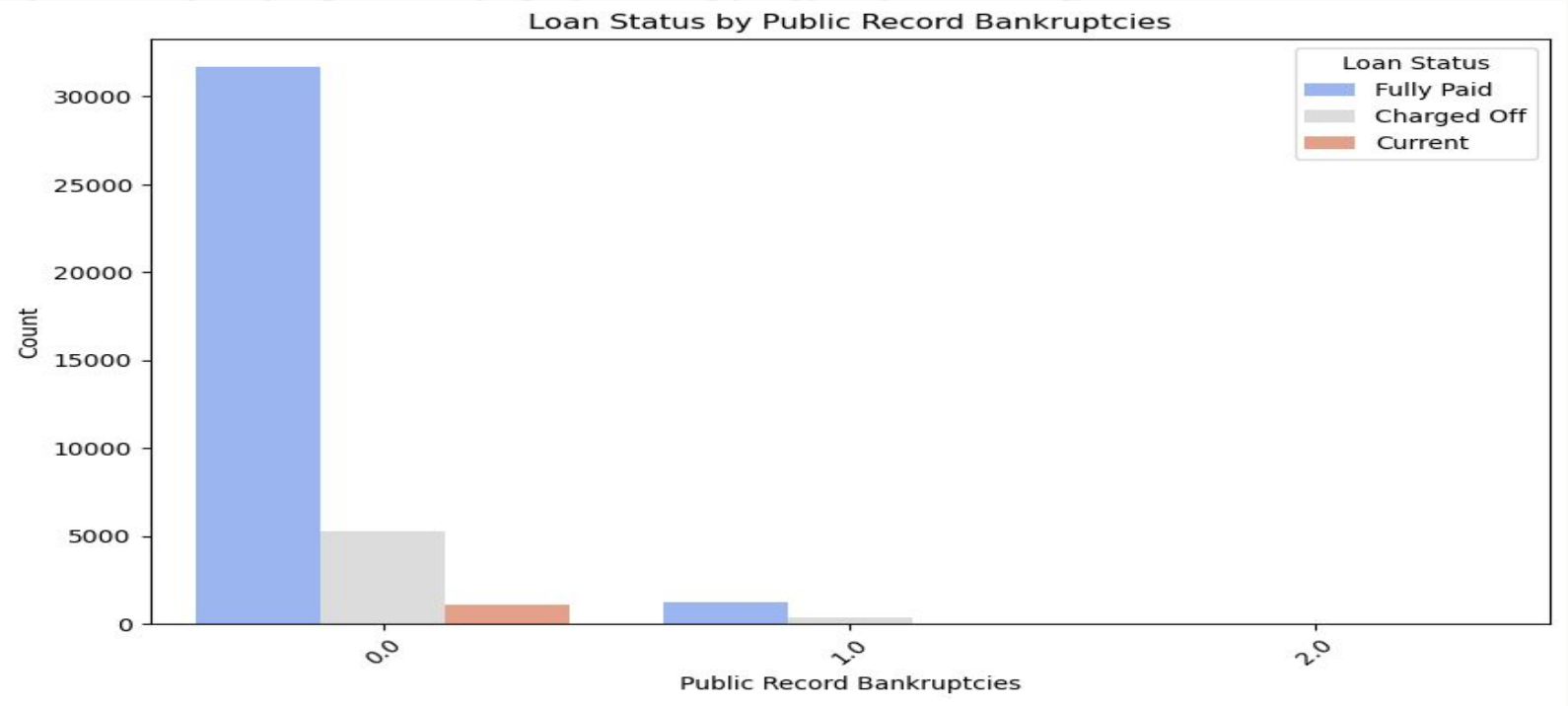
Conclusion from Bar Plot (Loan Status by Home Ownership) Analysis 9





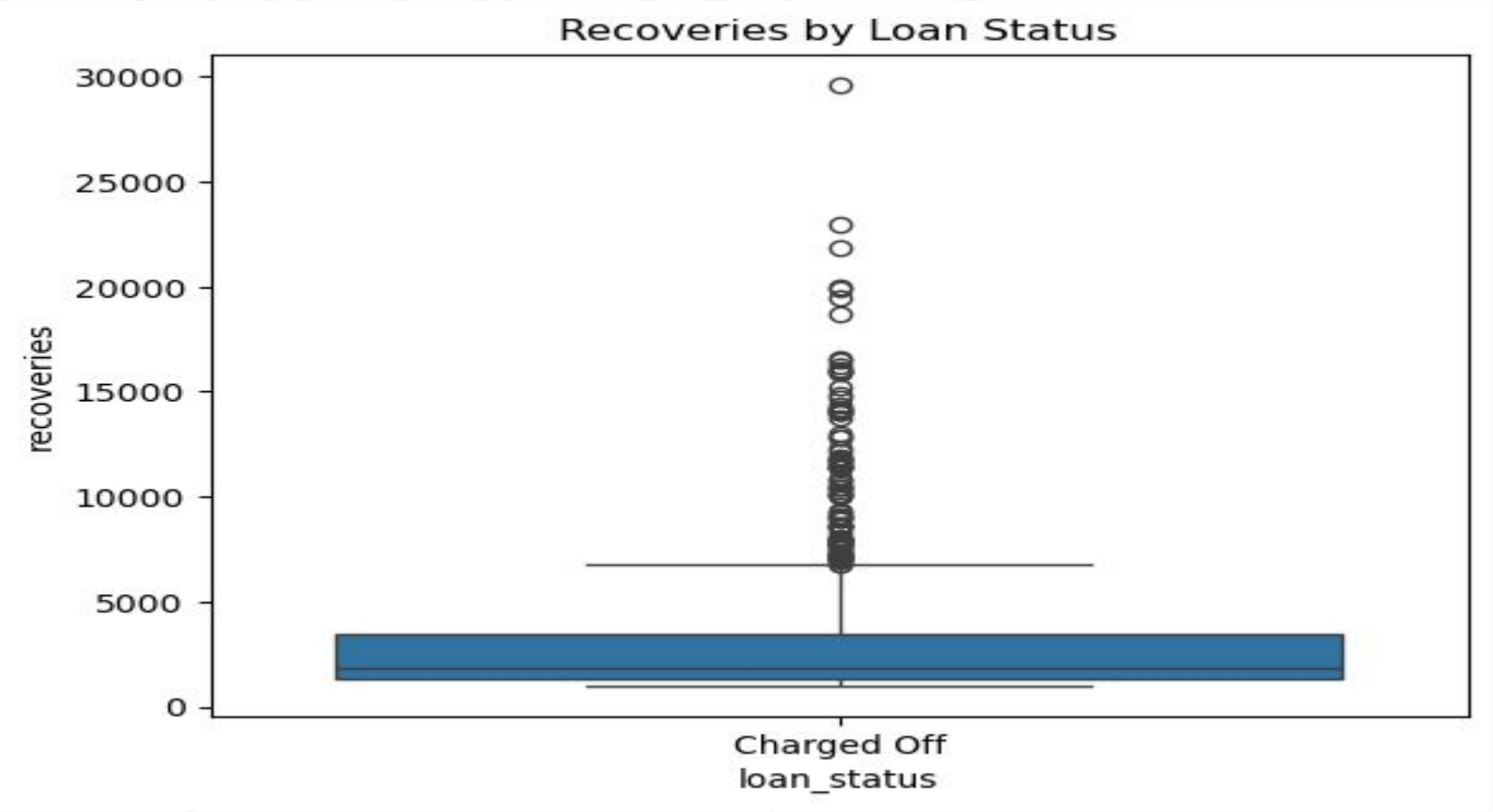
- Home Ownership and Interest Rates: The box plot suggests a general trend that individuals with "NONE" or "OTHER" home ownership status tend to have lower interest rates compared to those who mortgage, rent, or own their homes.
- Variability: There's significant variability in interest rates across all home ownership categories. This indicates that other factors besides home ownership likely influence interest rates.
- Outliers: The presence of outliers suggests that some individuals within each category have interest rates that deviate significantly from the typical range.
- Home Ownership: The chart suggests that individuals who rent or own their homes might have lower credit risk compared to those with "OTHER" or "NONE" home ownership status. This is because renting or owning a home often indicates financial stability and responsibility.
- Loan Status: The presence of "Charged Off" loans across all home ownership categories indicates that credit risk exists regardless of home ownership status.

Bar Chart (Bankruptcies and Loan Status) Analysis 10



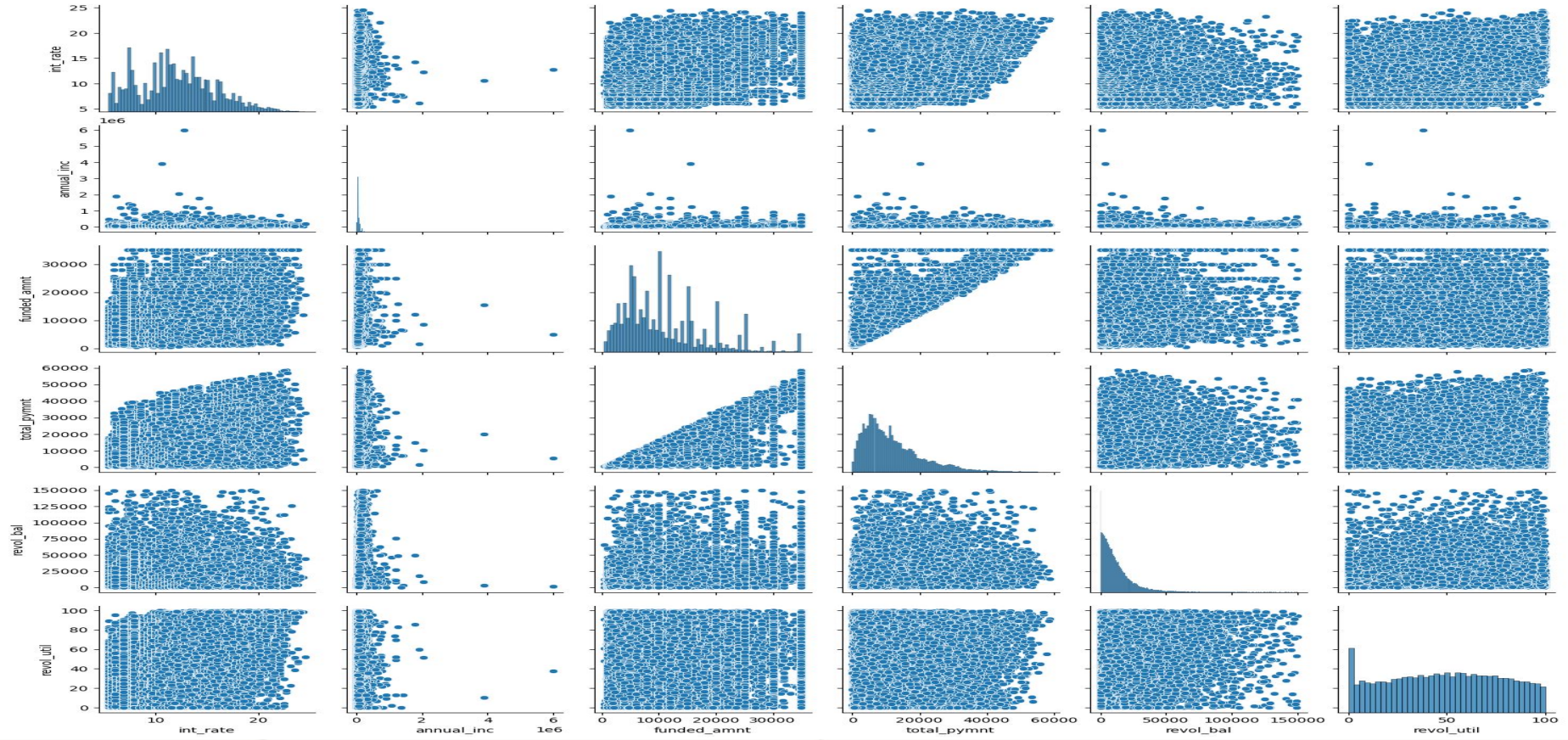
- Bankruptcies and Loan Status: The chart shows a clear trend: as the number of public record bankruptcies increases, the proportion of "Charged Off" loans also increases. This suggests that borrowers with a history of bankruptcies are more likely to default on their loans.
- Loan Distribution: The vast majority of loans are held by individuals with no public record bankruptcies. The number of loans decreases significantly as the number of bankruptcies increases.

Box plot Recovery Rate Analysis 11



- **Low Recovery Rates:** The box plot indicates that recoveries for "Charged Off" loans are generally low. This highlights the significant credit risk associated with these loans and the potential for substantial losses for the lender

Scatter Plot Analysis 12



- **Loan repayment (total_pymnt) strongly depends on the funded amount.**
- **Interest rates are relatively independent of borrowers' annual income, suggesting risk assessment uses other factors.**
- **High Revolving Balance & Utilization: A high revol_bal coupled with a high revol_util suggests that borrowers are heavily utilizing their available credit. This could indicate financial strain and a higher risk of default.**
- **Low Income: Borrowers with low annual_inc might have limited capacity to repay loans, particularly if they have taken on large funded_amnt.**
- **Loan Size: Large funded_amnt loans generally carry higher risk due to the potential for larger losses in case of default.**

Technologies Used

- Python - version 3.12.4
- Pandas - version 2.2.2
- Matplotlib - version 3.8.4
- Seaborn - version 0.13.2
- Plotly 5.22.0
- Jupyter Notebook
- Anaconda Navigator 2.6.3
- Visual Studio Code 1.96.0

Acknowledgements

- Give credit here.
- - This project was inspired by...
- - References if any...
- - This project was based on [this tutorial](https://www.upgrad.com/).

Contact

- Created by [Niranjan Singh and] - feel free to contact me!
- <!-- Optional -->
- <!-- ## License -->
- This project is open source and available under the [... License]().
- Created by [Niranjan Singh and Ameya Parab] - feel free to contact me!

Thank you