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

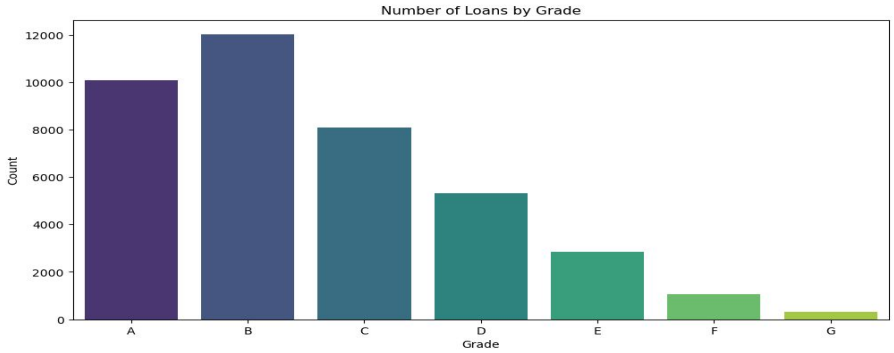
  Identify and analyze missing values in the dataset.
- Analysing categorical and numerical columns

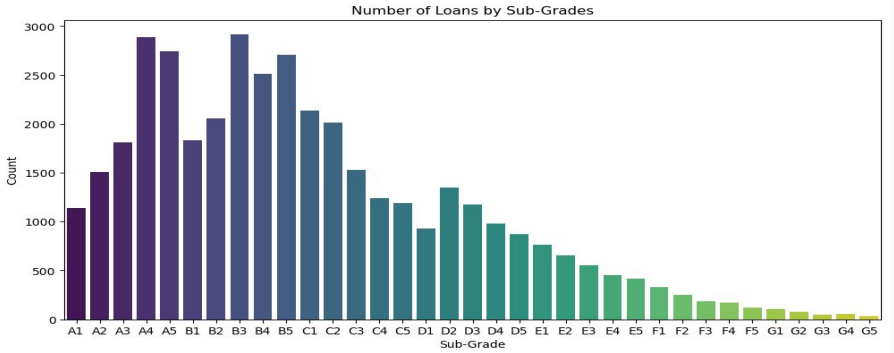
  Analyze categorical and numerical columns to understand the statistical properties and relationships within the dataset.
- Univariate Analysis:

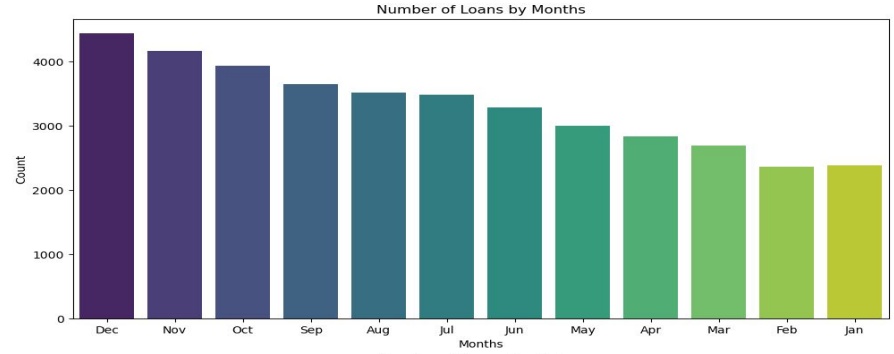
  Conduct univariate analysis to explore the distribution and characteristics of individual variables.
- Outliers:
  - Identify and analyze outliers within the dataset to understand their impact on the analysis.
- Bivariate analysis:
  - Conduct bivariate analysis to explore relationships between different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

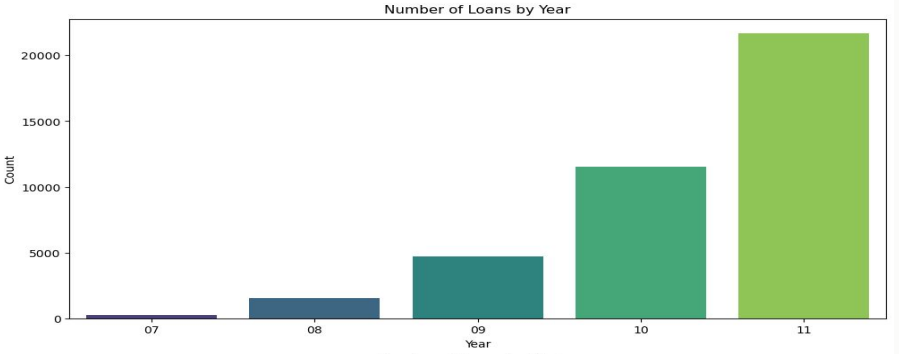
Conclusions

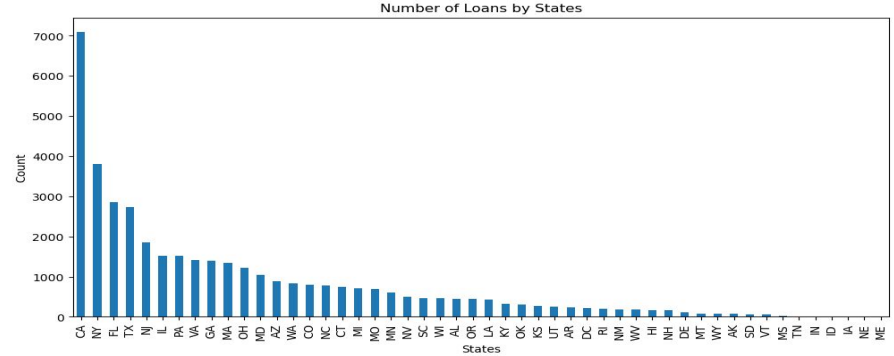
The Categorical Univariate Analysis

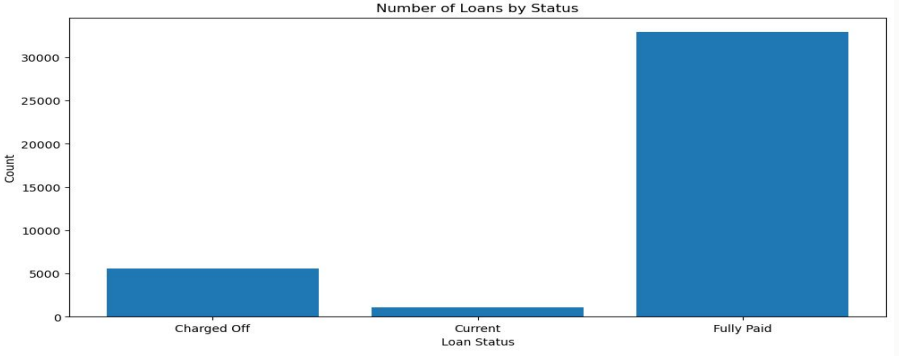




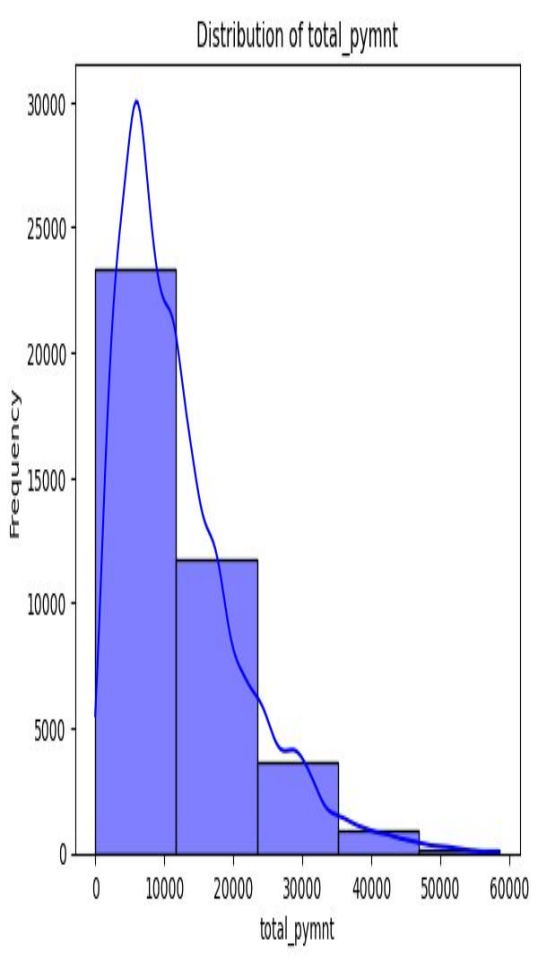


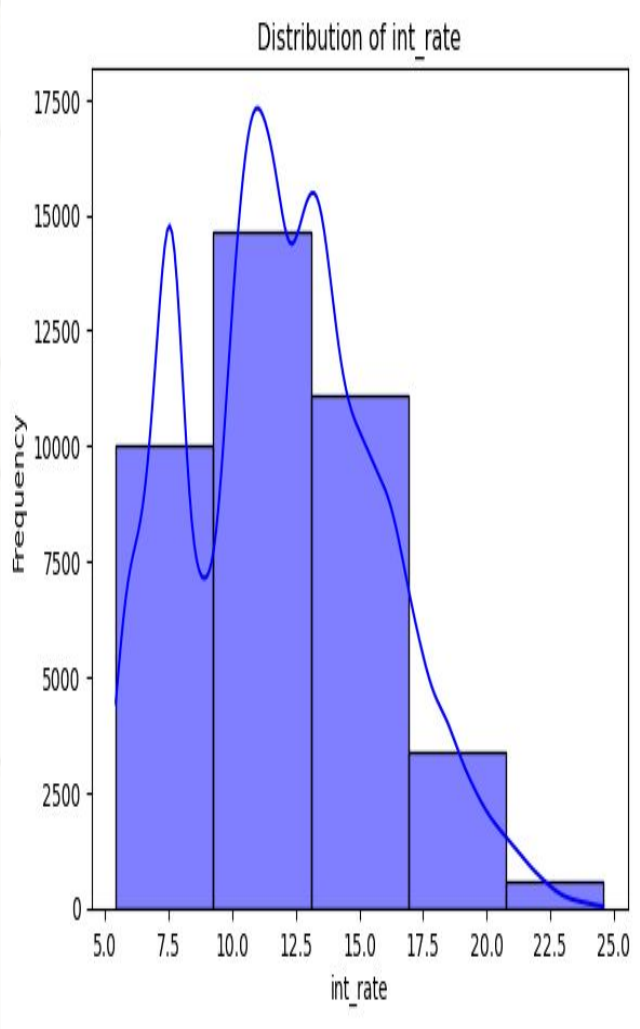


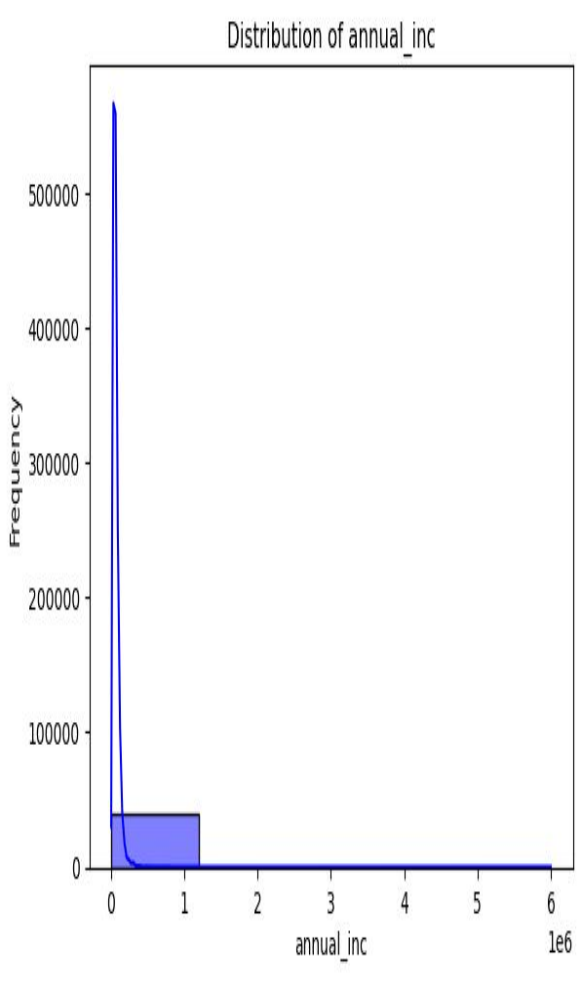




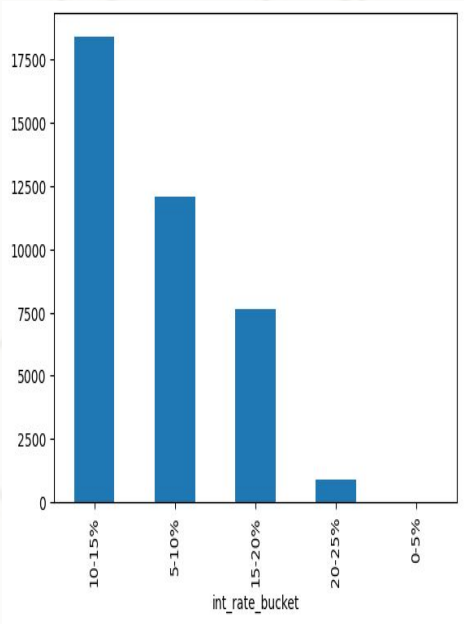
The Univariate Numerical Analysis

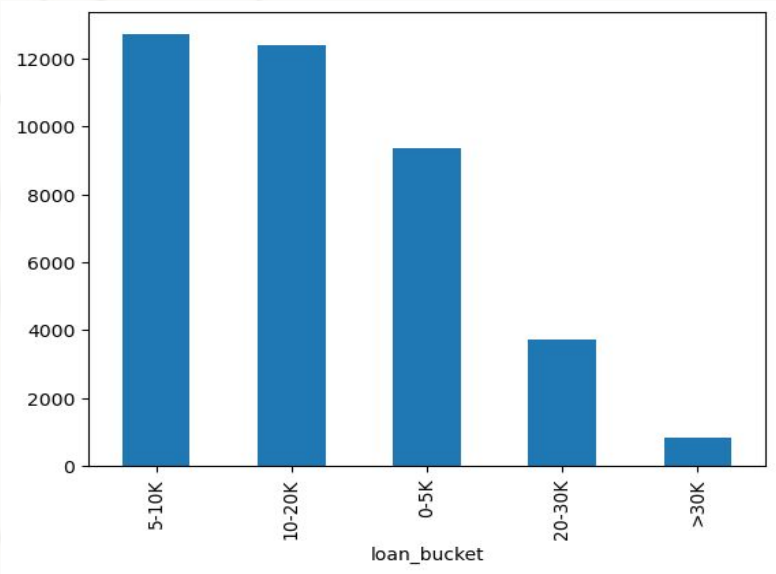




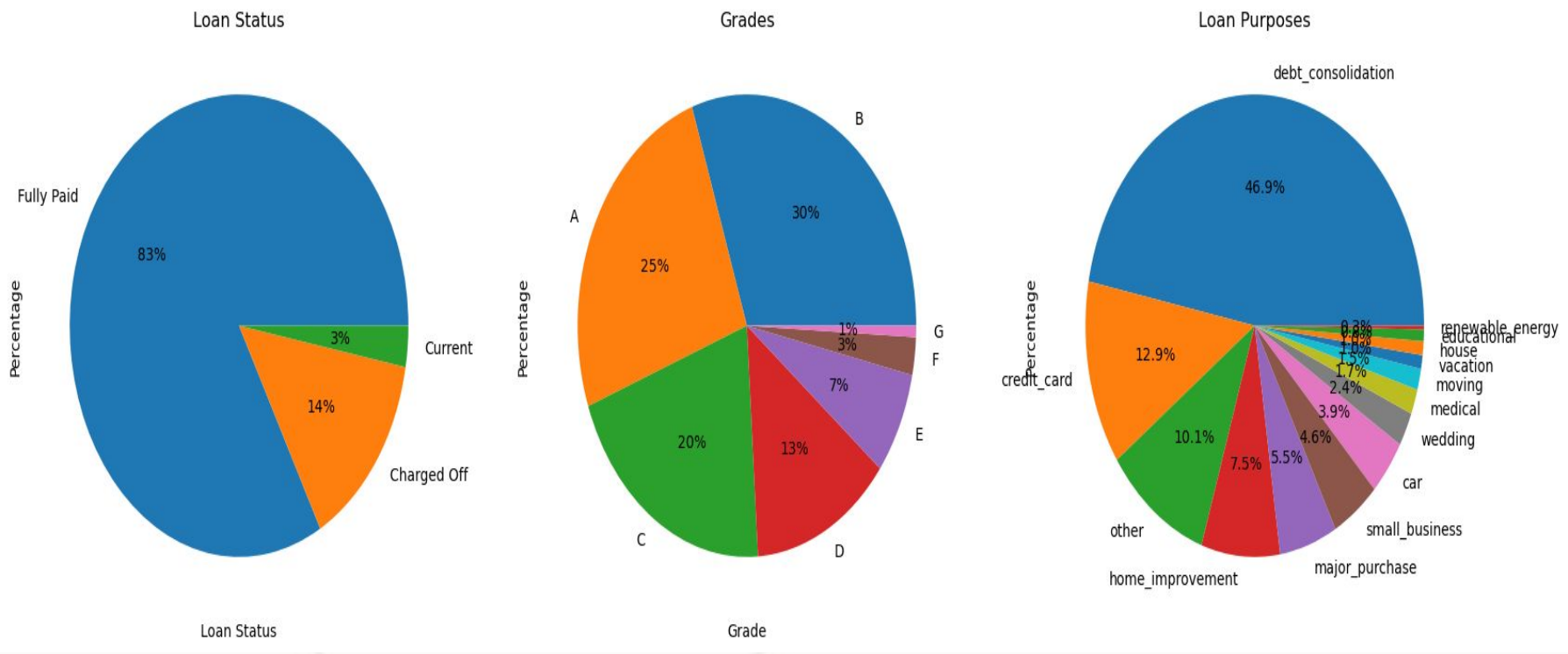


Segmented Univariate Analysis

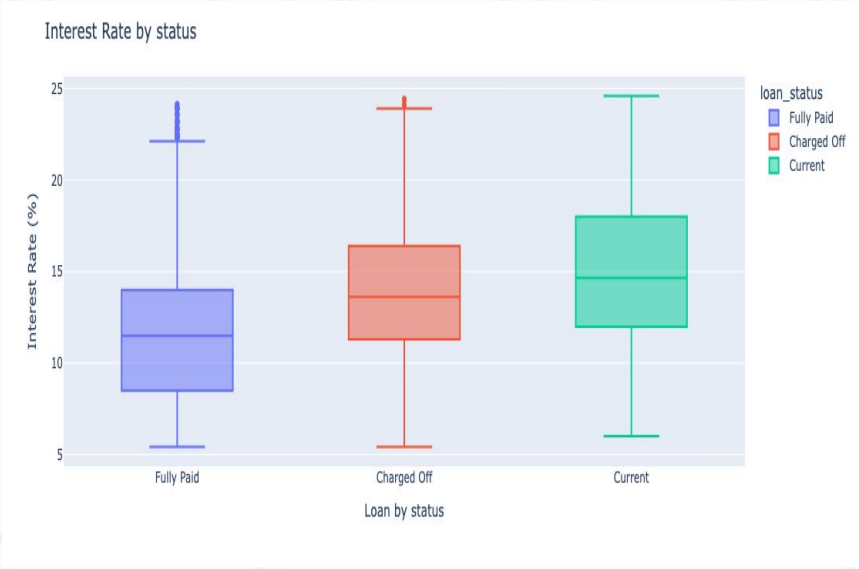


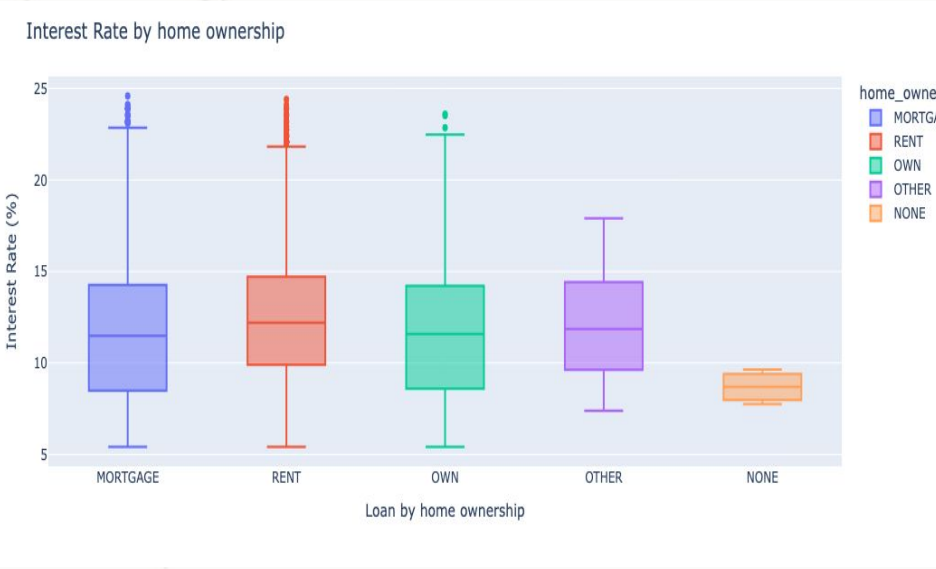


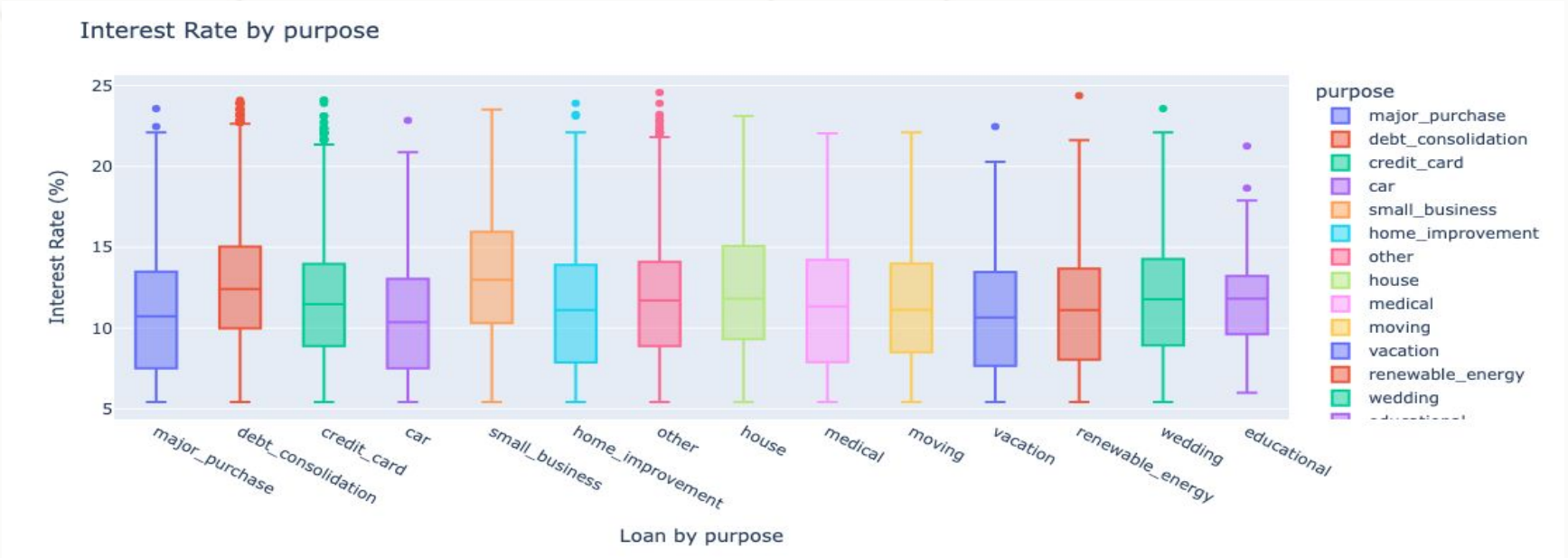
Pie Plot (Loan Status, Grades and Purpose) Categorical Univariate Analysis



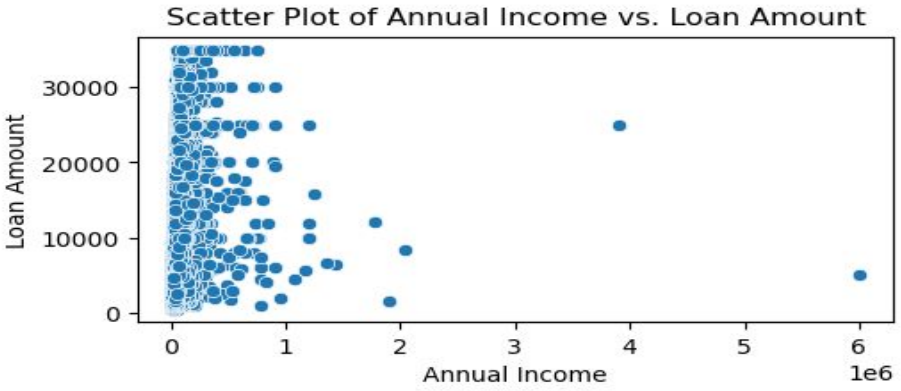
Box Plot (Interest Rate by Grade, Sub Grade and Loan Status) Byvariate Categorical Analysis

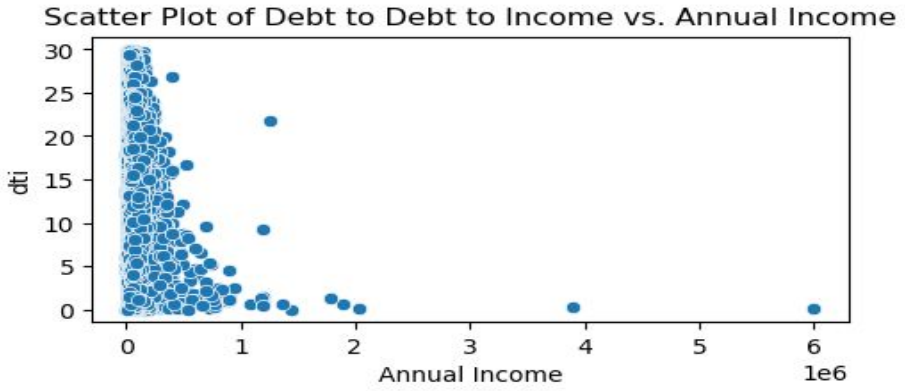


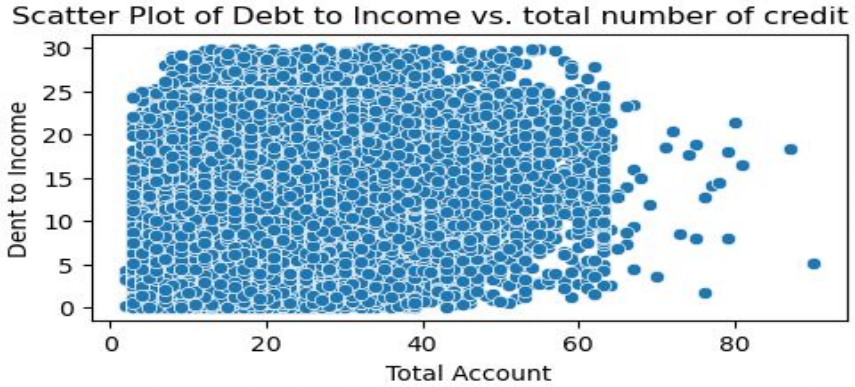


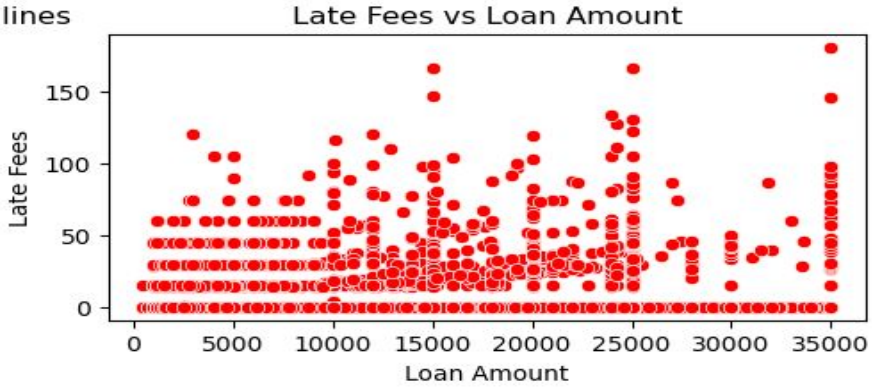


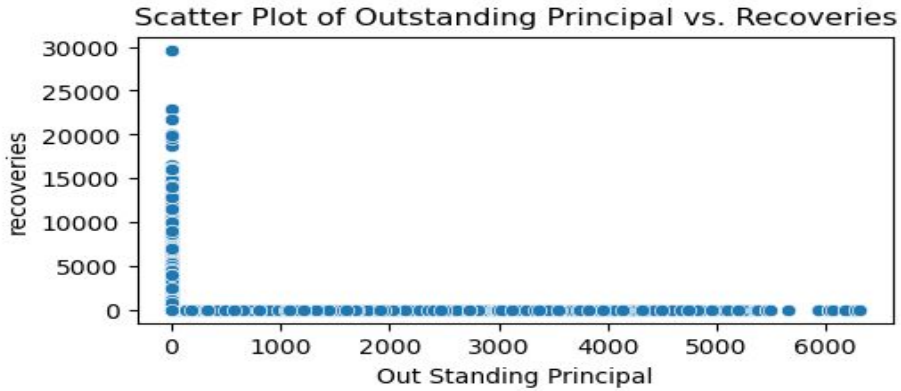
Scatter Plot Numerical Byvariate Analysis











Univariate and Byvariate Analysis (Risky Loans)

Grades B, C, and D

Grades B, C, and D have the highest "Charged Off" loans.Lower grades (B-1396,C-1313,D-1085) exhibit a higher proportion of "Charged Off" loans, suggesting a higher risk of default for borrowers with lower creditworthiness.

Subgrades C1, C2, B3, B4, and B5

Subgrades C1,C2, B3, B4, and B5 show a higher likelihood of defaults.Higher default tendencies in specific sub-grades indicate a need for more granular underwriting policies.

States

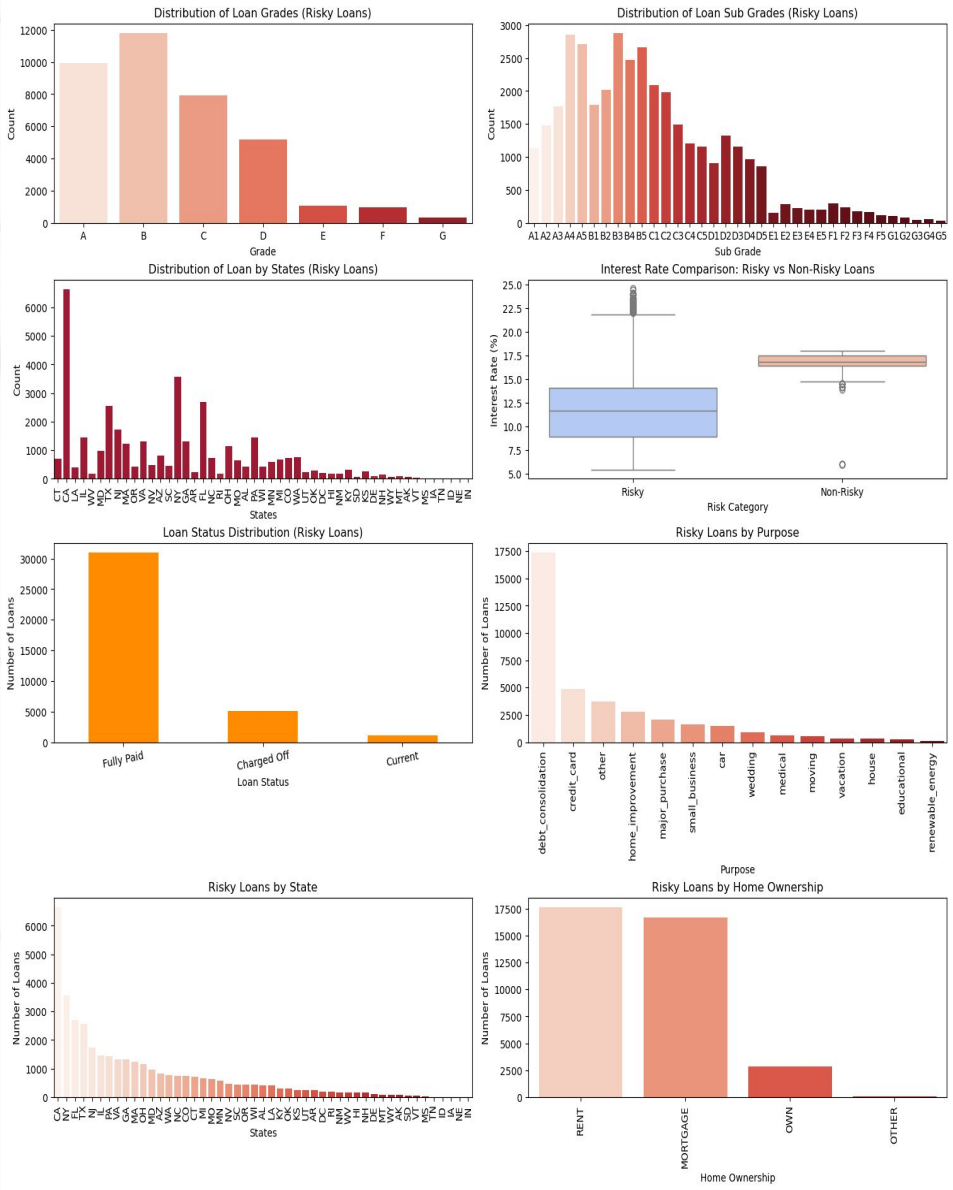
Applicants from states like California (CA), Florida (FL), and New York (NY) show higher default tendencies. Geographic concentration of risky loans could lead to localized portfolio risk exposure.

Interest Rate Comparison

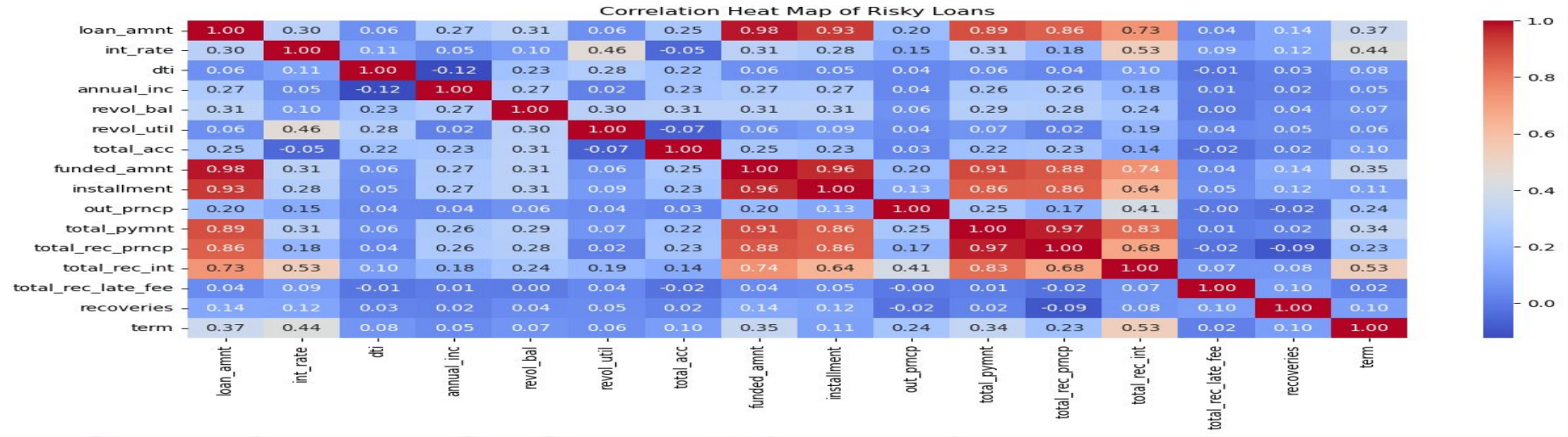
Risky loans are associated with higher interest rates compared to non-risky loans, but the spread between the two categories suggests that some risky loans may still be underpriced. Higher rates on risky loans might indicate an attempt to offset the increased probability of default, but affordability issues may arise.

Loan Status Distribution

Most risky loans have been fully paid, but a significant proportion are charged off or currently delinquent. Charged-off loans highlight direct financial losses.



Correlation Heat Map (Numerical Features) Bivariate Analysis



Key Observations:

Strong Positive Correlations:

Loan Amount (loan_amnt) and Funded Amount (funded_amnt): These two variables are highly correlated (0.98), which is expected as the funded amount is typically close to the loan amount requested.

Loan Amount (loan_amnt), Funded Amount (funded_amnt), and Installment (installment): These variables exhibit strong positive correlations, indicating that larger loan amounts generally result in higher monthly installments.

Loan Amount (loan_amnt), Funded Amount (funded_amnt), and Total Payment (total_pymnt): These variables also show strong positive correlations, suggesting that larger loans require larger total repayments.

Moderate Positive Correlations:

Interest Rate (int_rate) and Loan Amount (loan_amnt): A moderate positive correlation suggests that larger loans might be associated with higher interest rates

Interest Rate (int_rate) and Total Received Interest (total_rec_int): This correlation indicates that higher interest rates generally lead to higher total interest payments

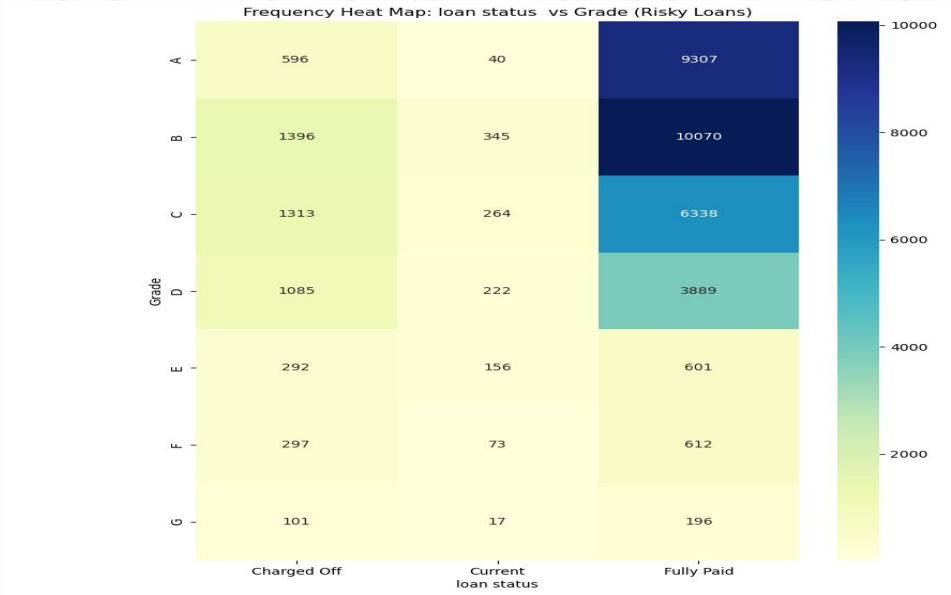
Revolving Balance (revol_bal) and Revolving Utilization (revol_util): A moderate positive correlation is expected, as higher revolving balances tend to increase credit utilization.

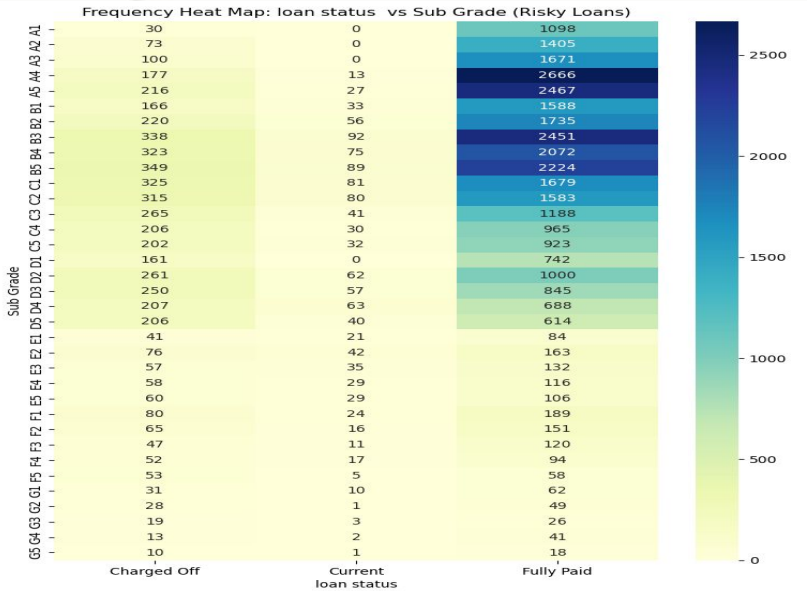
Negative Correlations:

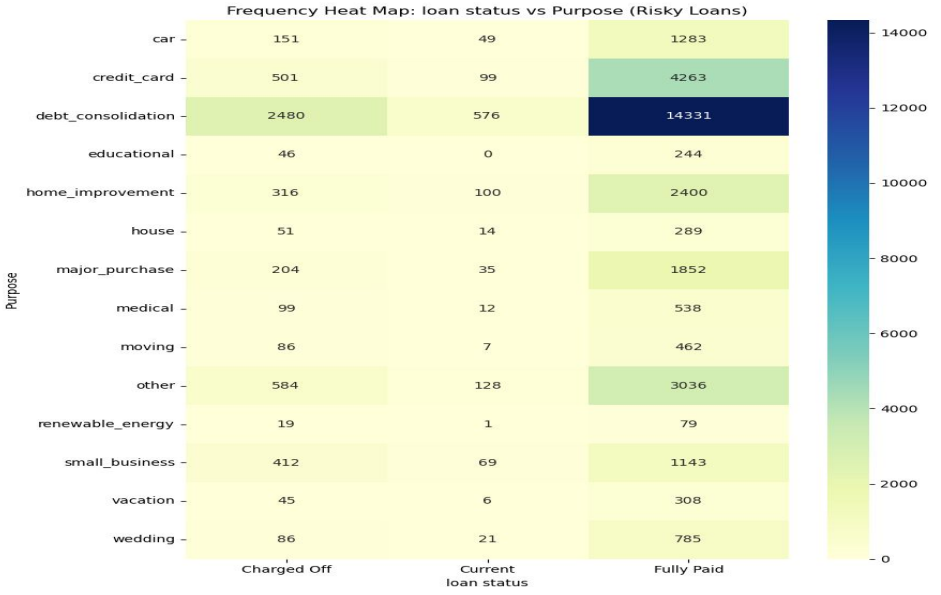
Debt-to-Income Ratio (dti) and Annual Income (annual_inc): A negative correlation suggests that individuals with higher incomes generally have lower debt-to-income ratios.

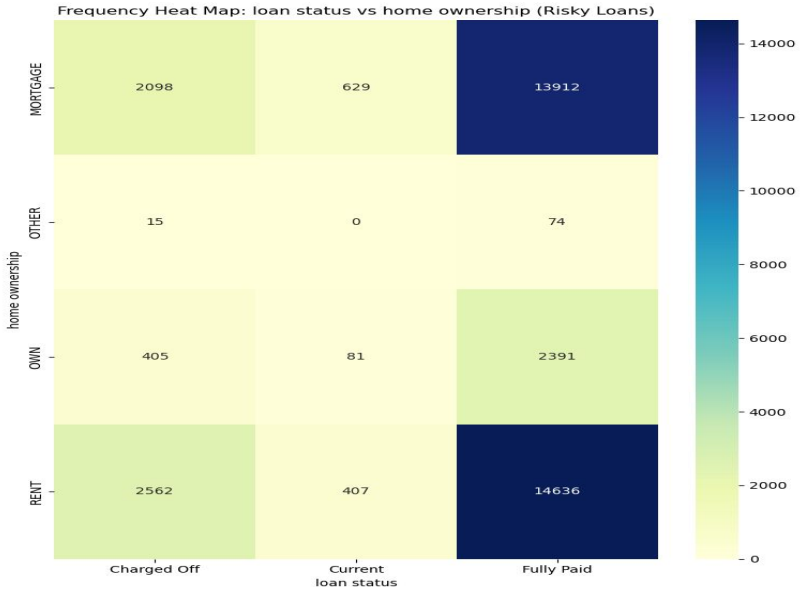
Revolving Utilization (revol_util) and Recovery Amounts (recoveries): A negative correlation indicates that loans with higher revolving utilization tend to have lower recovery rates in case of default.

Correlation Heat Map (Categorical Features) Analysis

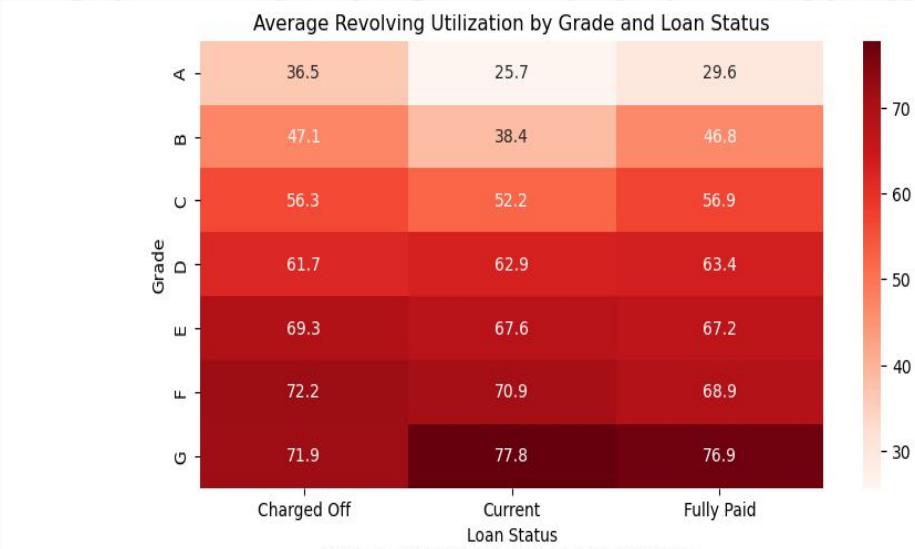


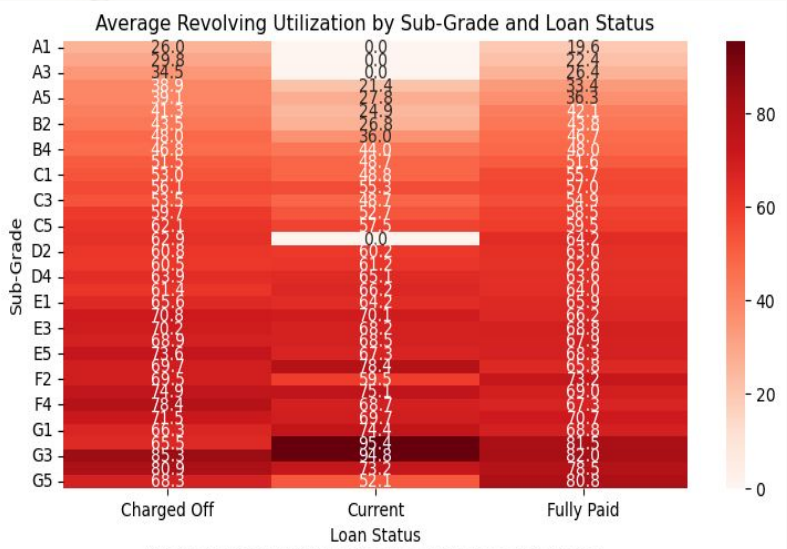


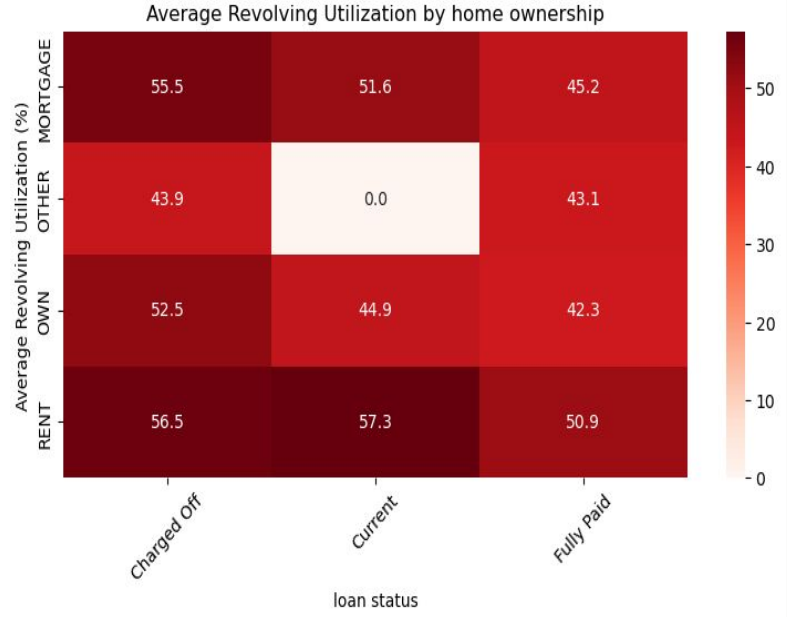




The Revolving Utilization Heatmap Analysis



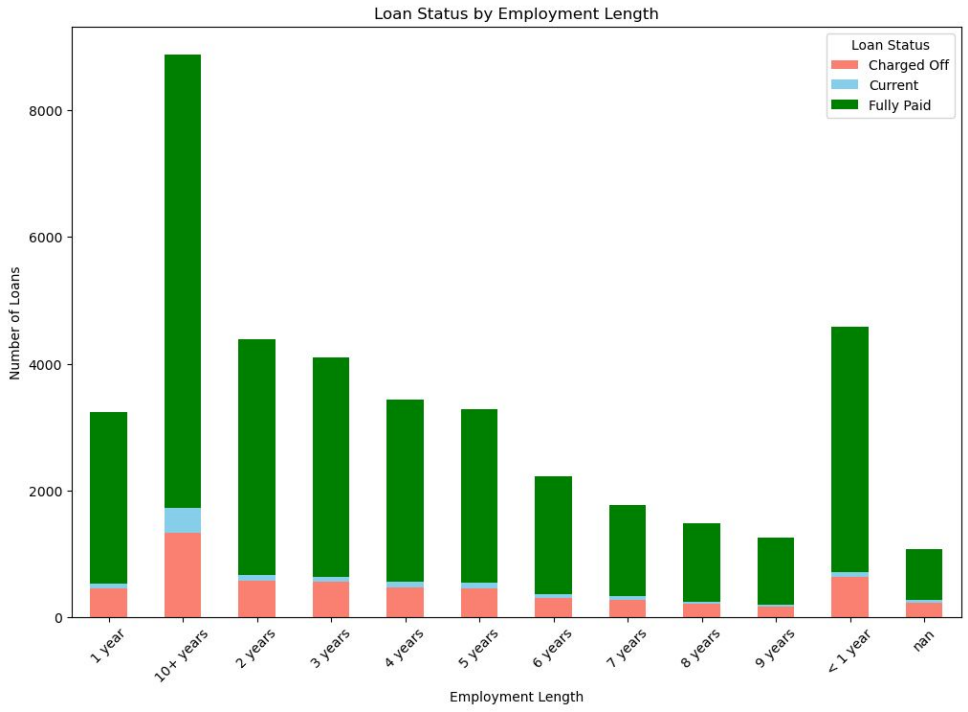


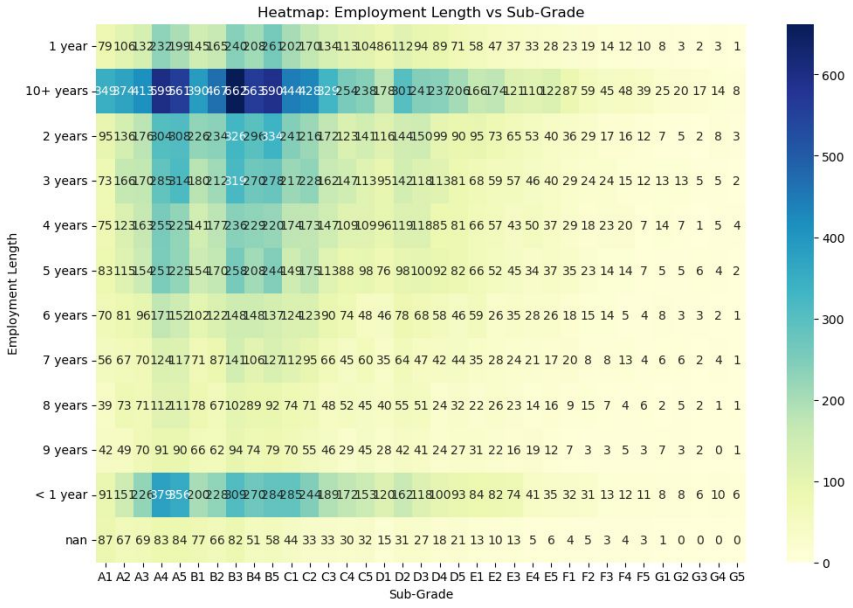


Box Plot (Loan Status vs Revolving Utilization) Analysis

- 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

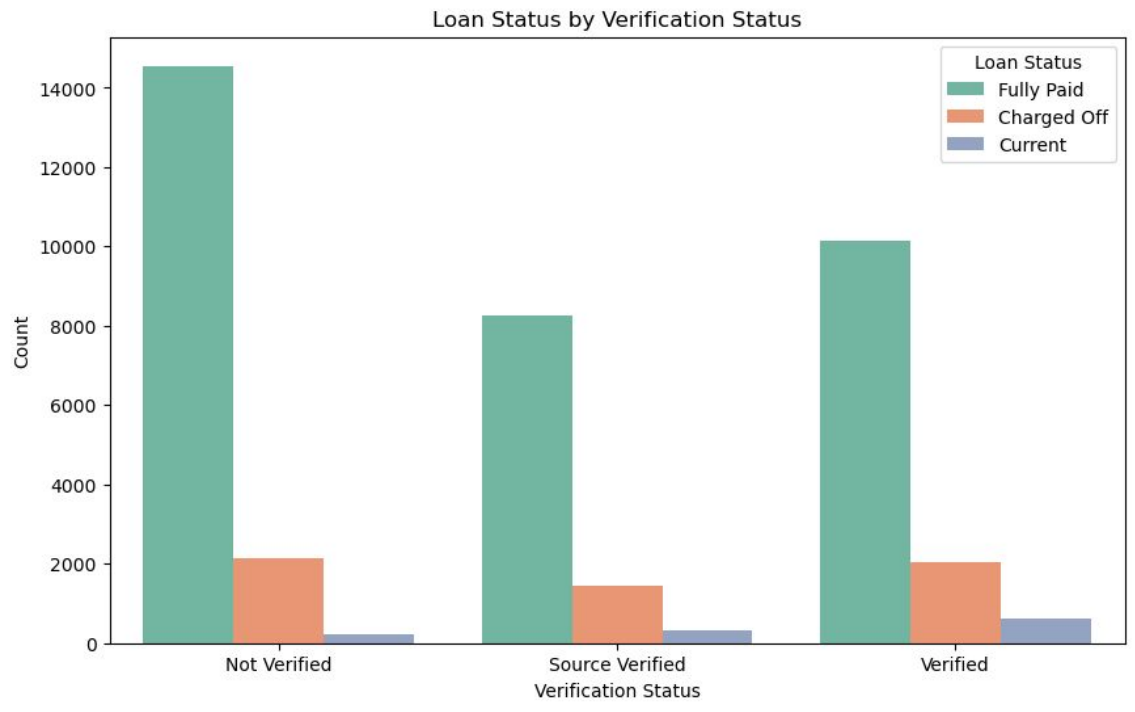
Bar Plot (Loan Status by Employment Length, Verification Status etc) analysis

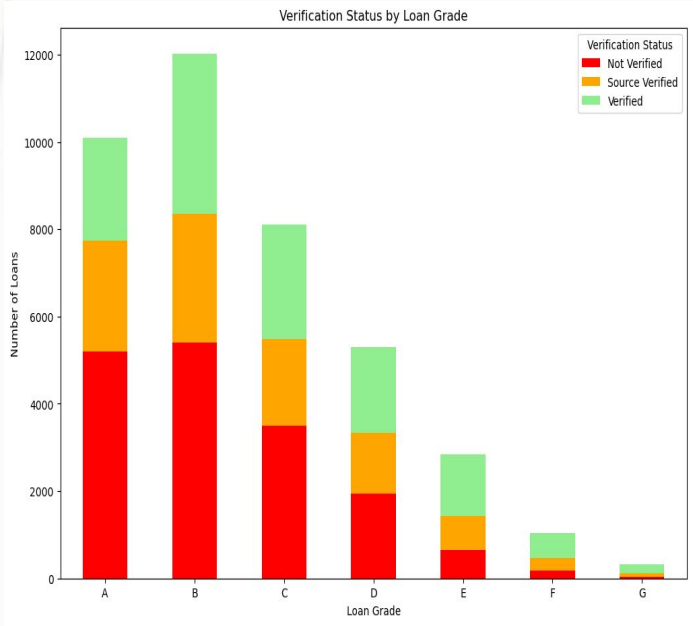




- Revolving Utilization and Loan Status: The heatmaps clearly show that revolving utilization tends to be higher for "Charged Off" loans compared to "Current" and "Fully Paid" loans across various categories (grade, sub-grade, purpose, home ownership).
- Grade and Sub-Grade: Revolving utilization generally 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".

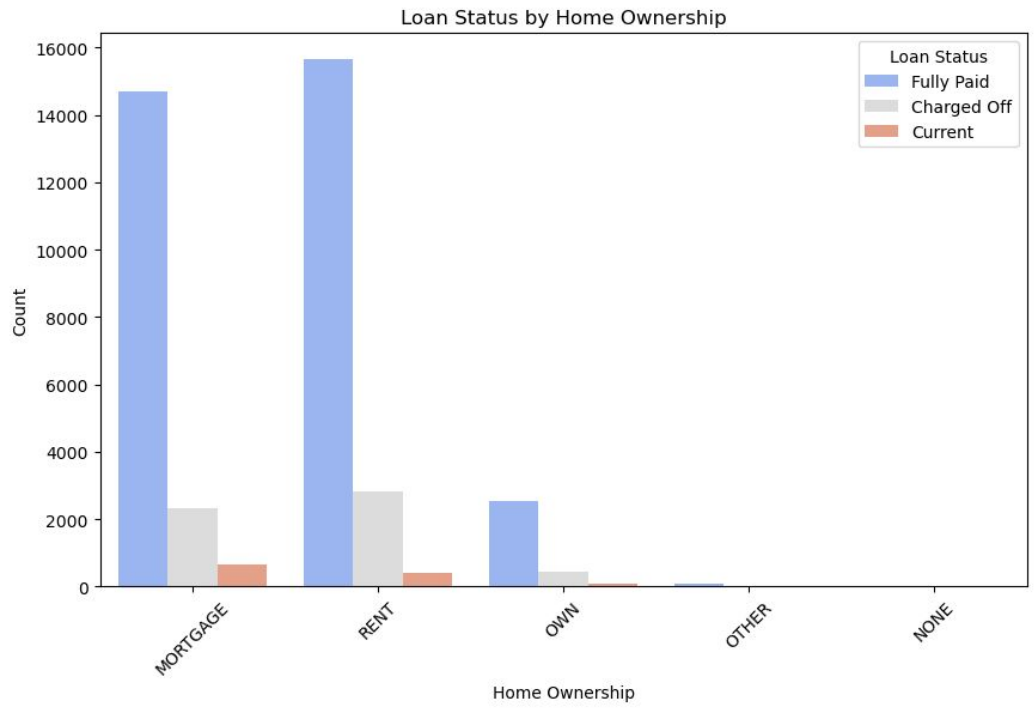
Bar Plot (Loan Status by Employment Length, Verification Status etc) analysi

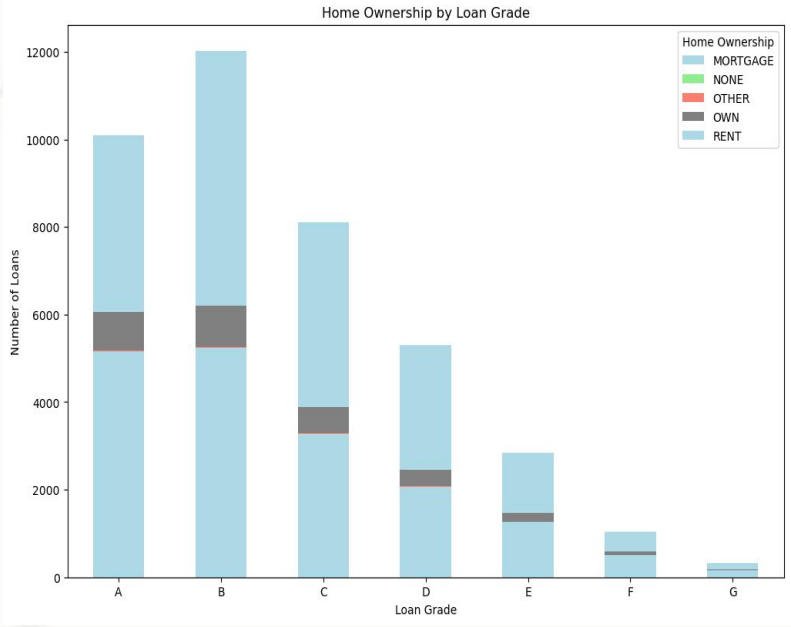




- Revolving Utilization and Loan Status: The heatmaps clearly show that revolving utilization tends to be higher for "Charged Off" loans compared to "Current" and "Fully Paid" loans across various categories (grade, sub-grade, purpose, home ownership).
- Grade and Sub-Grade: Revolving utilization generally 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".

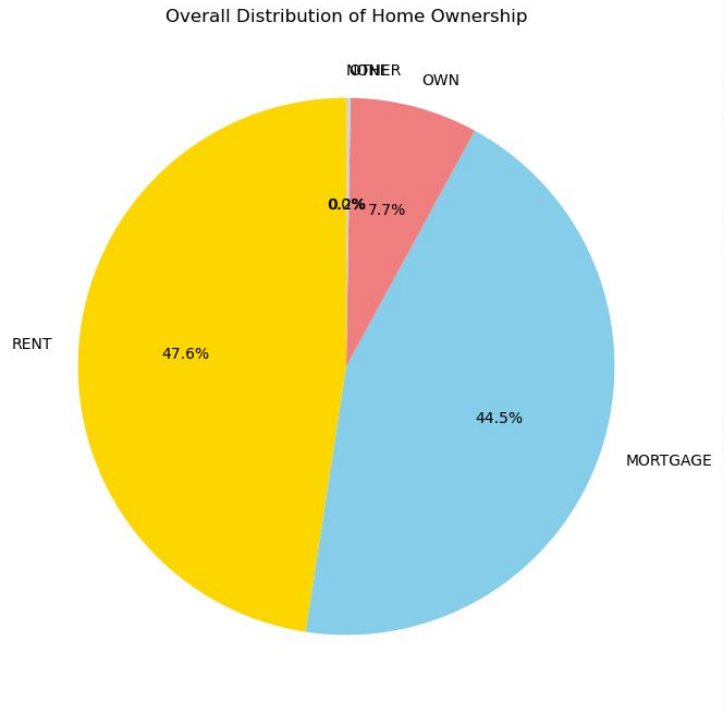
Bar Plot (Loan Status by Home Ownership) Analysis

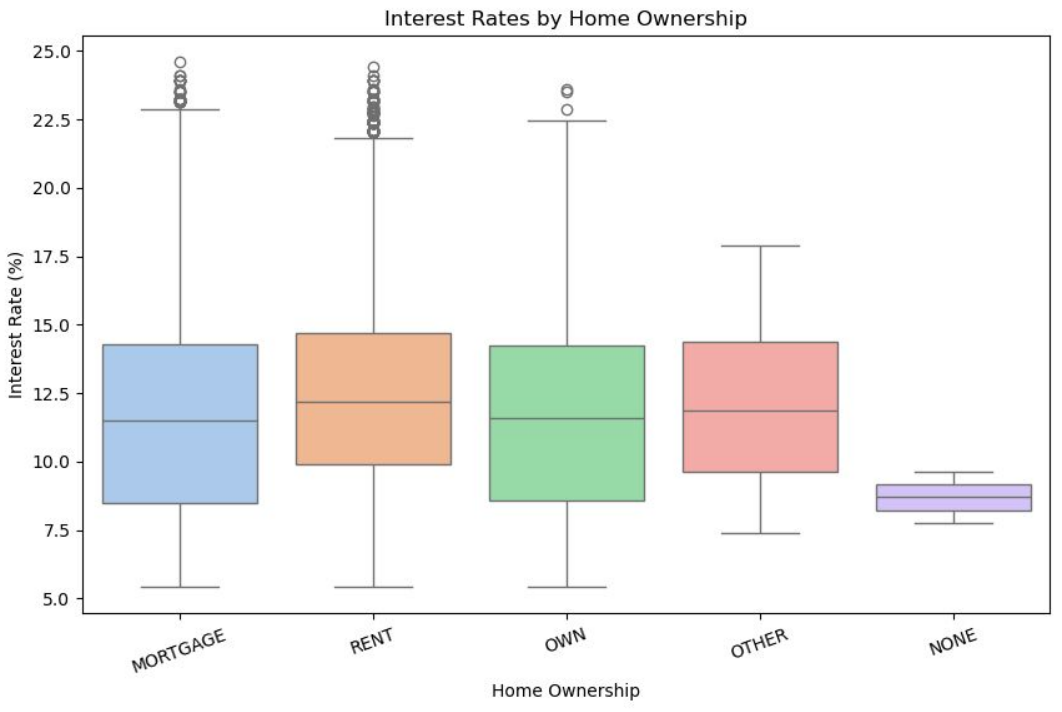




- Home Ownership and Interest Rates: The box plot suggests a general trend that individuals with "NONE" or "OTHER" home ownership status tend to have lower interest rates compared to those who mortgage, rent, or own their homes.
- Variability: There's significant variability in interest rates across 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.

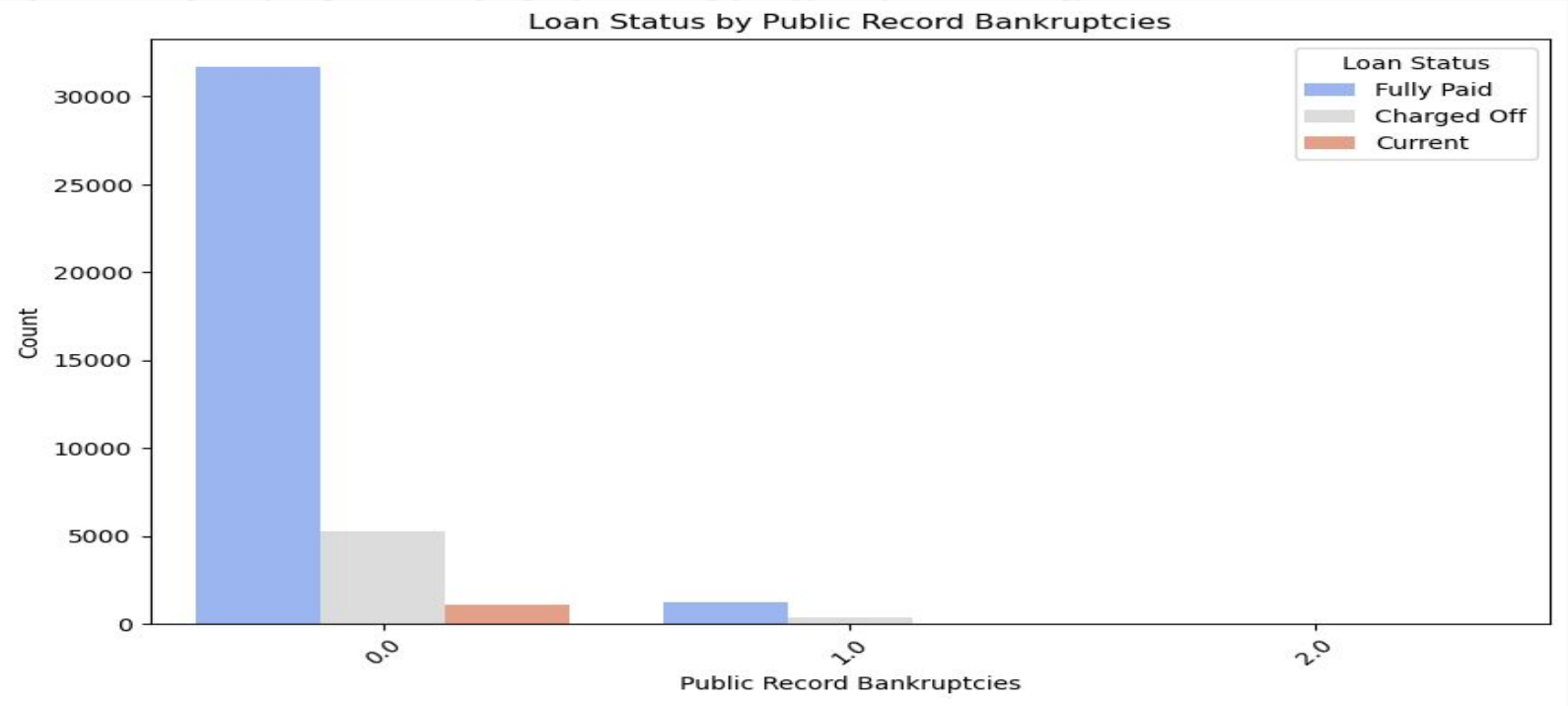
Bar Plot (Loan Status by Home Ownership) Analysis





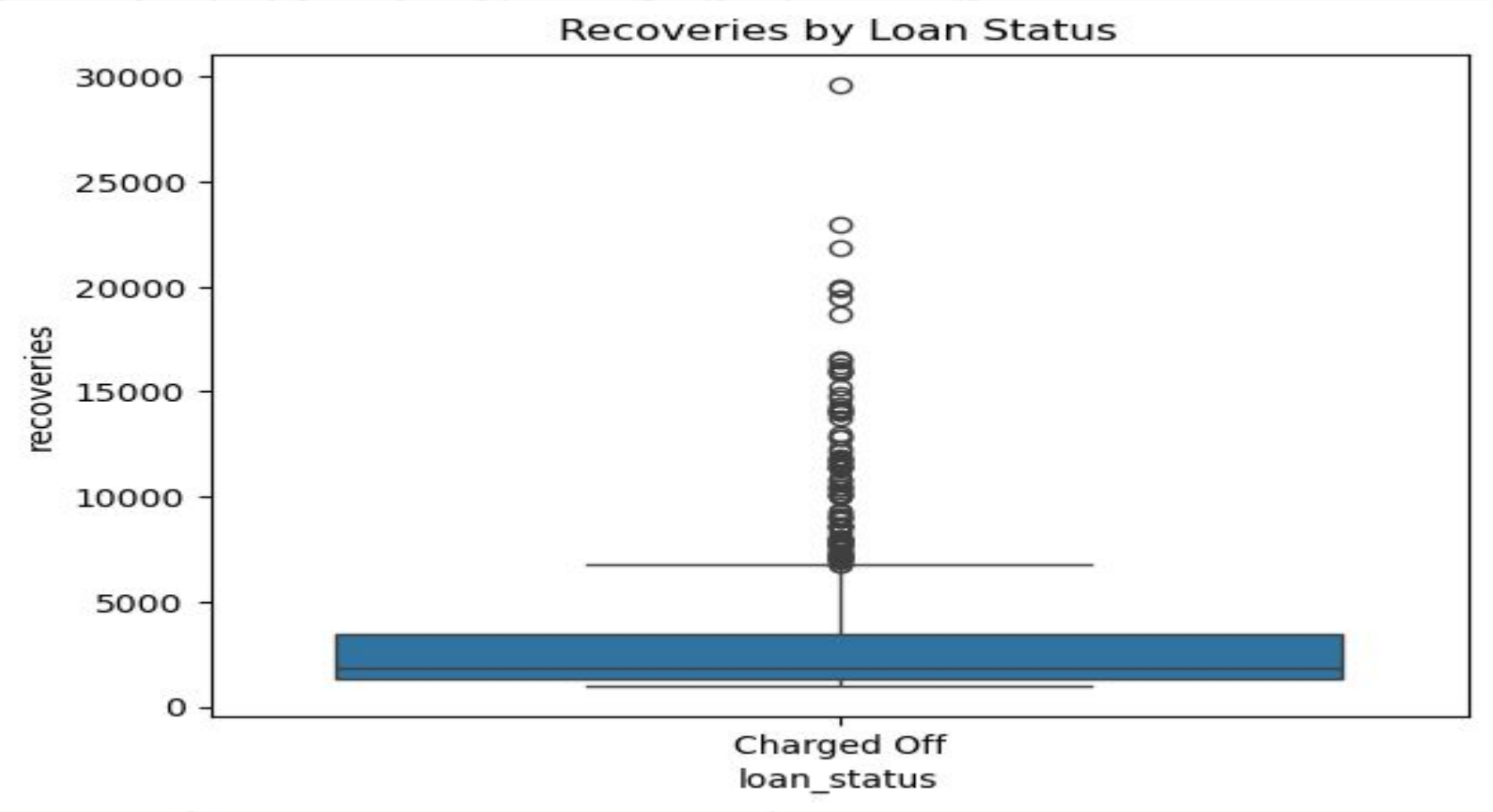
- Home Ownership and Interest Rates: The box plot suggests a general trend that individuals with "NONE" or "OTHER" home ownership status tend to have lower interest rates compared to those who mortgage, rent, or own their homes.
- Variability: There's significant variability in interest rates across 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



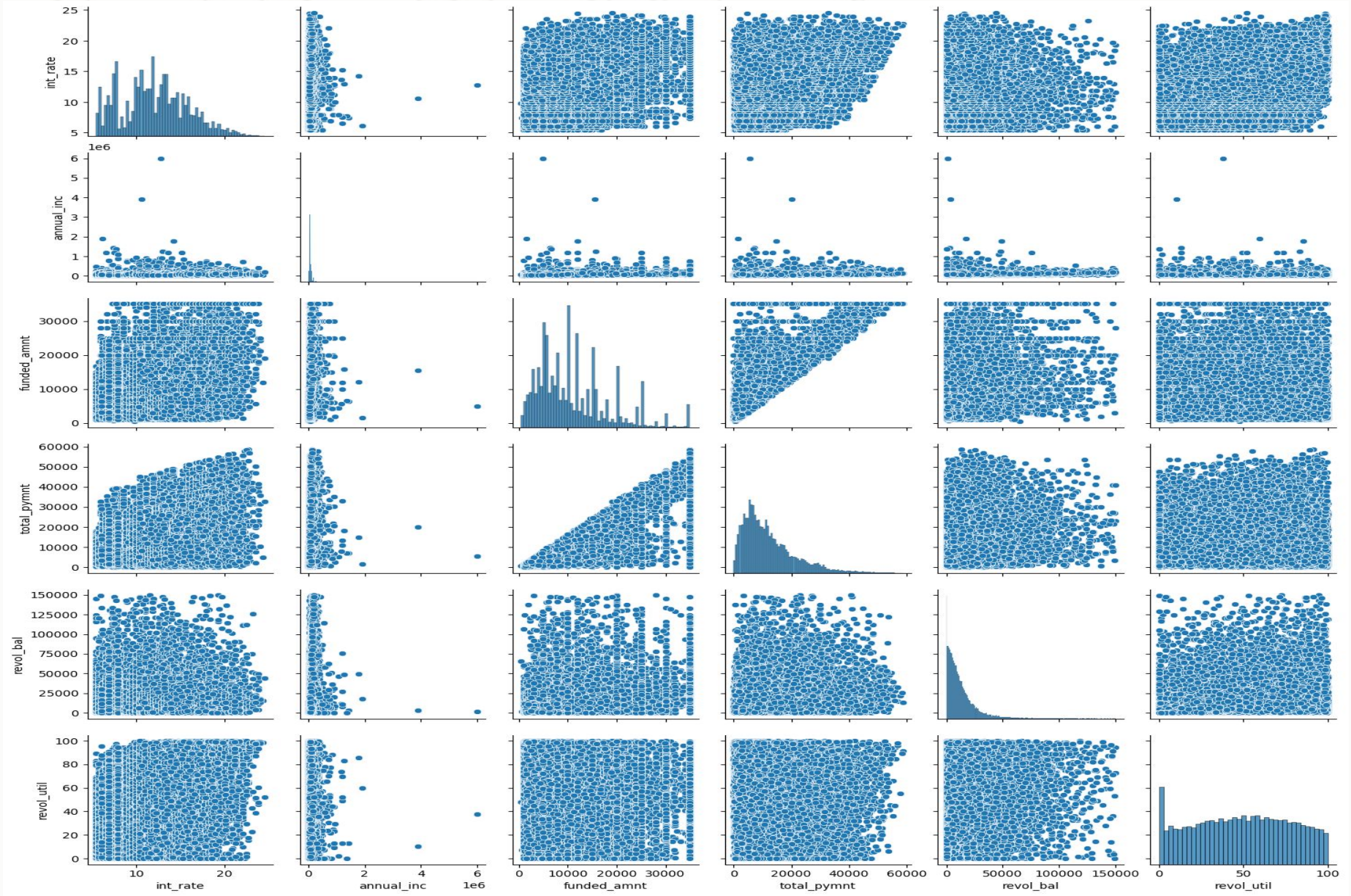
- Bankruptcies and Loan Status: The chart shows a clear trend: as the number of public record bankruptcies increases, the proportion of "Charged Off" loans also increases. This suggests that borrowers with a history of bankruptcies are more likely to default on their loans.
- Loan Distribution: The vast 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



- **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



Conclusions

Conclusion 1

Risk Assessment and Recommendations

Grades B, C, and D

Grades B, C, and D have the highest "Charged Off" loans.Lower grades (B-1396,C-1313,D-1085) exhibit a higher proportion of "Charged Off" loans, suggesting a higher risk of default for borrowers with lower creditworthiness.

Recommendation:

- Reassess the risk evaluation criteria for grades B and C to better distinguish higher-risk borrowers.
- Implement stricter risk assessment and underwriting criteria for applicants in these grades. Utilize additional credit metrics, such as repayment history, to screen high-risk applicants more effectively.

Subgrades C1, C2, B3, B4, and B5

Subgrades C1,C2, B3, B4, and B5 show a higher likelihood of defaults.Higher default tendencies in specific sub-grades indicate a need for more granular underwriting policies.

Recommendation: Focus on applicants in these subgrades by:

- Offering smaller loan amounts.
- Introducing risk-based pricing with higher interest rates to offset potential losses.
- Providing pre-loan financial counseling.

States

Applicants from states like California (CA), Florida (FL), and New York (NY) show higher default tendencies. Geographic concentration of risky loans could lead to localized portfolio risk exposure.

Recommendation:

- Introduce geographically adjusted lending strategies.
- Implement state-level risk mitigation strategies, such as stricter approval criteria for high-risk states.

Conclusion 2

Risk Assessment and Recommendations

Interest Rate Comparison

Risky loans are associated with higher interest rates compared to non-risky loans, but the spread between the two categories suggests that some risky loans may still be underpriced. Higher rates on risky loans might indicate an attempt to offset the increased probability of default, but affordability issues may arise.

Recommendation: Optimize pricing strategies for high-risk loans to balance profitability and affordability.

Loan Status Distribution

Most risky loans have been fully paid, but a significant proportion are charged off or currently delinquent. Charged-off loans highlight direct financial losses.

Recommendation: Focus on early warning signals to predict potential defaults and prevent charge-offs.

DTI Ratios and Interest Rates

Most loans have interest rates between 10% and 20%. Borrowers with rates above 15% might have higher default risks due to affordability issues.

Recommendation:

- Adjust interest rates based on borrower DTI ratios.
- Cap DTI ratios for loan approvals or offer smaller loan amounts for borrowers with high DTI.
- Ensure that interest rates align with the borrower’s repayment capacity.

Conclusion 3

Risk Assessment and Recommendations

Term Length

Applicants opting for 60-month loans are more likely to default.

Recommendation:

- Evaluate the risks of longer-term loans and consider limiting the maximum term.
- Offer 60-month loans at higher interest rates to account for elevated risk.

Applicant Experience

Applicants with 10+ years of experience are more prone to defaults, suggesting experience is not a standalone creditworthiness indicator.

Recommendation:

- Use a comprehensive credit scoring system that includes factors such as income, DTI ratio, and past loan behavior.

Market Trend

Loan applications have steadily increased between 2007 and 2011, indicating market growth.

Recommendation: Continue leveraging this growth by:

- Maintaining competitive rates and efficient loan processing.
- Enhancing risk management practices to sustain profitability.

Conclusion 4

Risk Assessment and Recommendations

Loan Demand

*Q4 and December see a peak in loan applications, likely due to holiday-related spending.

Recommendation: Prepare for seasonal demand surges by:

- Allocating additional resources to ensure timely processing.
- Offering promotional rates or tailored products during peak seasons.

Debt Consolidation Loans

Debt consolidation accounts for the highest loan volume and default rates.

Recommendation: - Apply stricter eligibility criteria for debt consolidation loans. - Provide financial counseling to help applicants manage existing debts effectively.

Housing Status

Renters and mortgaged homeowners are more likely to default. Renters may face more financial instability compared to homeowners, while mortgage holders might be over-leveraged.

Recommendation: Include housing stability as a key factor in the underwriting process. Consider:

- Adjusting interest rates or loan amounts based on housing status.
- Encouraging applicants to secure cosigners if housing status poses a risk.
- Applicants with higher housing costs (Rent or Mortgage) may require lower loan amounts or higher scrutiny.

Conclusion 5

Risk Assessment and Recommendations

Verification Process

Verified applicants are defaulting more frequently than unverified ones.

Recommendation: Review the current verification process to:

- Improve its accuracy in identifying risky borrowers.
- Incorporate additional metrics, such as alternative credit data.

Loan Amounts

Applicants with loan amounts of $15,000 or higher are more likely to default.

Recommendation:

- Set stricter caps on loan amounts for higher-risk borrowers.
- Require additional collateral or cosigners for larger loans.

Annual Income Levels

Borrowers earning less than $38,000 annually have a higher likelihood of defaulting.

Recommendation:

- Provide financial education and budgeting resources for lower-income borrowers.
- Restrict maximum loan amounts or shorten loan terms for low-income applicants.

Conclusion 5

Risk Assessment and Recommendations

Revolving Utilization (revol_util)

Revolving utilization is distributed across a wide range, with many borrowers exceeding 50%.High utilization indicates greater dependence on credit, correlating with higher default likelihood.

Recommendation: Reduce loan amounts or increase interest rates for high-utilization borrowers.

Months Since Last Delinquency (mths_since_last_delinq)

Many borrowers have had delinquencies within the past 60 months. Recent delinquencies are strong indicators of repayment challenges.

Recommendation: Require additional proof of creditworthiness from borrowers with recent delinquencies.

Risky Loans by Purpose

- Debt consolidation and credit card refinancing are the top purposes for risky loans.
- Home improvement and small business loans also show significant representation.
- These categories often reflect financial distress or speculative investments.

Recommendation: Tighten evaluation for loans aimed at debt consolidation and small businesses, where risks are inherently higher

Conclusion 6

Risk Assessment and Recommendations

- Stricter Criteria for High Utilization Borrowers: Applicants with utilization rates above 50% should be closely monitored or required to lower their utilization before loan approval.
- Financial Counseling and Education: Provide resources to help borrowers manage credit utilization and understand its impact on loan approvals and creditworthiness.
- Dynamic Interest Rates: Adjust interest rates based on revolving utilization levels, with higher rates for higher utilization borrowers to offset risk.

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