Lending Club Case Study

EXPLORATORY DATA ANALYSIS

SWAPNIL KELKAR
KRITIKA SINGH

Objective

The objective of this case study is to apply Exploratory Data Analysis (EDA) techniques to a real-world problem, derive insights, and present these findings in a business-centric manner through a presentation.

Benefits of the Case Study:

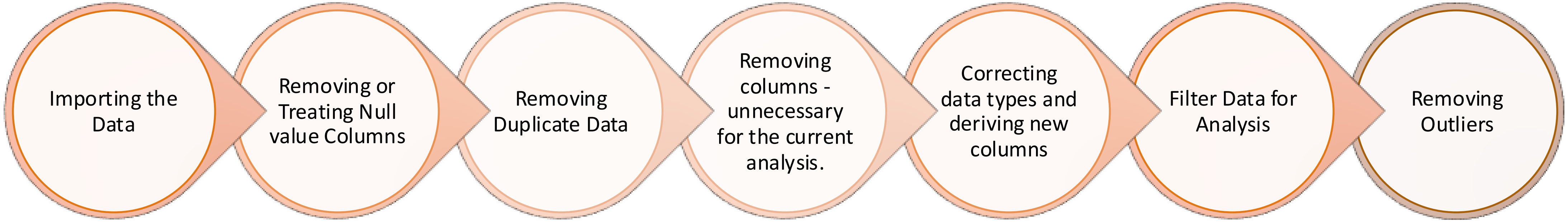
1. **Practical Application of EDA**: Provides a comprehensive understanding of how EDA is utilized in addressing real-life business challenges.
2. **Foundational Knowledge in Risk Analytics**: Enhances basic knowledge of risk analytics within the banking and financial services sectors.
3. **Data-Driven Decision Making**: Demonstrates how data is leveraged to minimize financial losses in lending processes.
4. **Improved Visualization Skills**: Enhances understanding of effective visualization techniques and the appropriate use of charts for real-world data.

Business Understanding

Business Objective: The business objective is to make informed decisions on whether to approve or reject loan applications based on specific variables.

Dataset Details: The dataset contains information about past loan applicants and their default status. It includes details of approved loans, not rejected ones. The dataset categorizes loan statuses into three types: Fully Paid, Current, and Charged-Off.

Data Clean-up and preparation process:

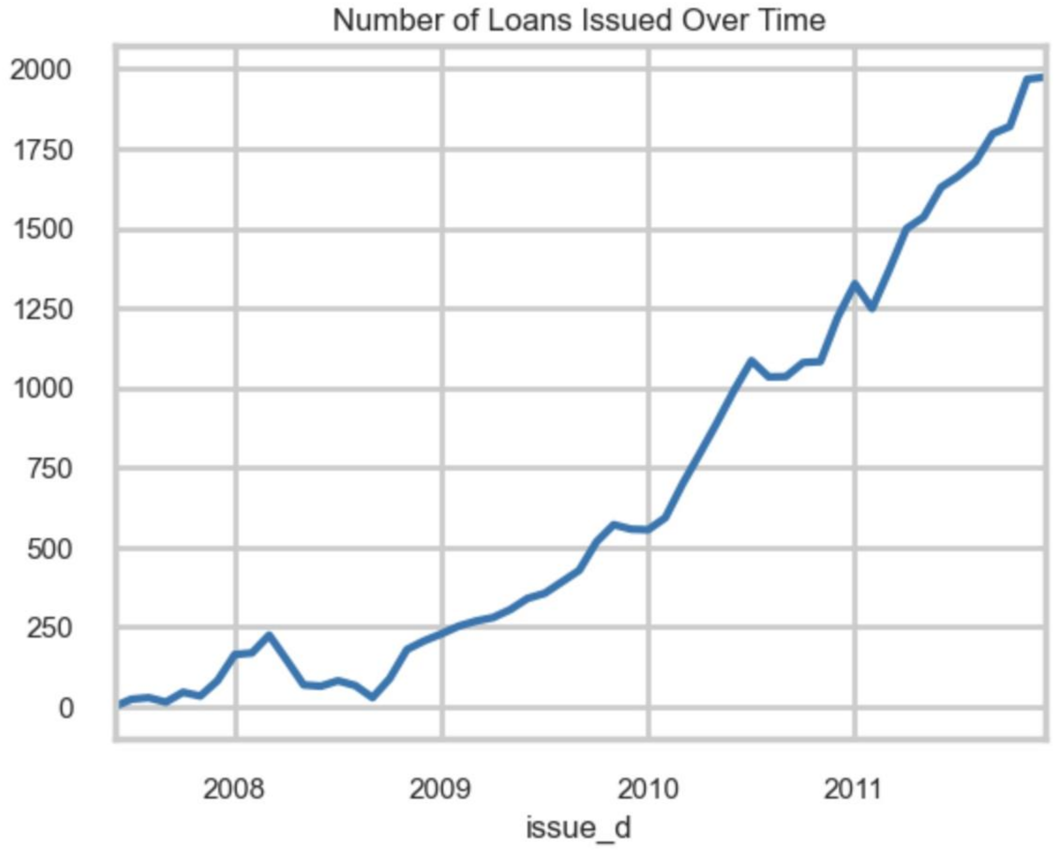


Visual Representations and Recommendations

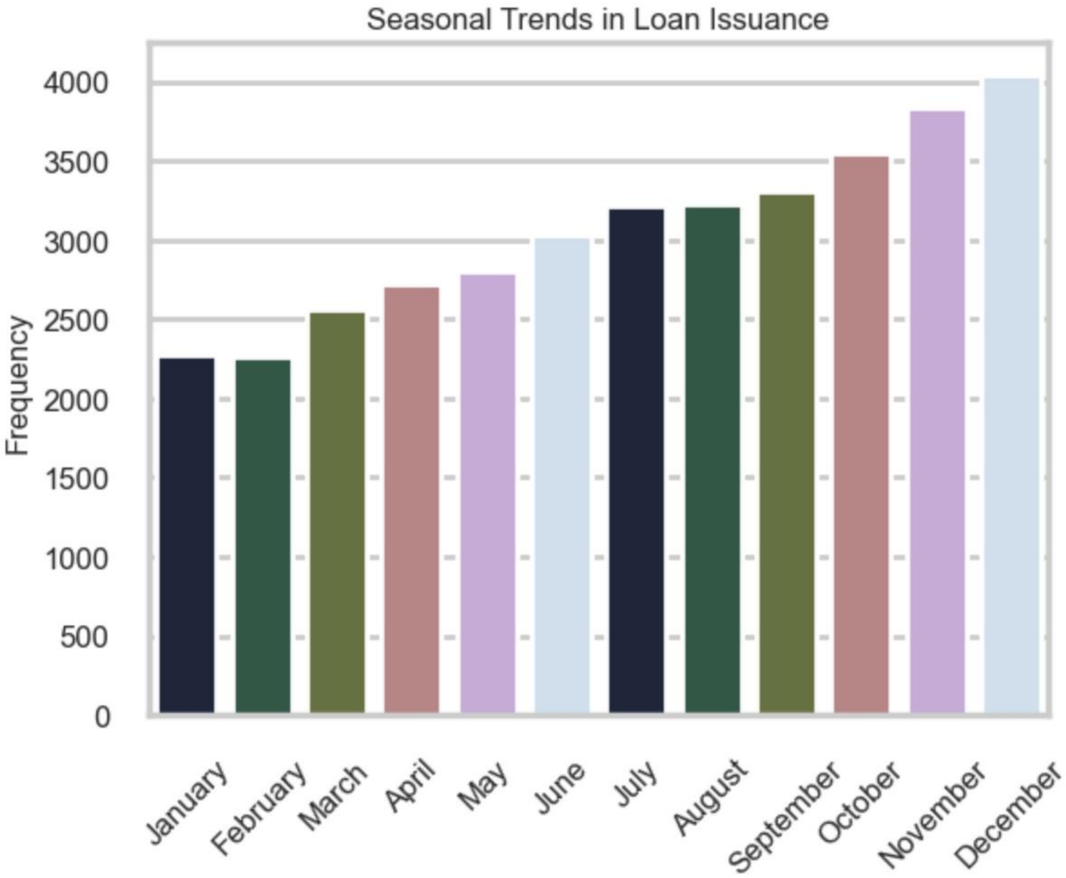
Univariate Analysis

The following key variables will be analysed as part of our Univariate Analyses:

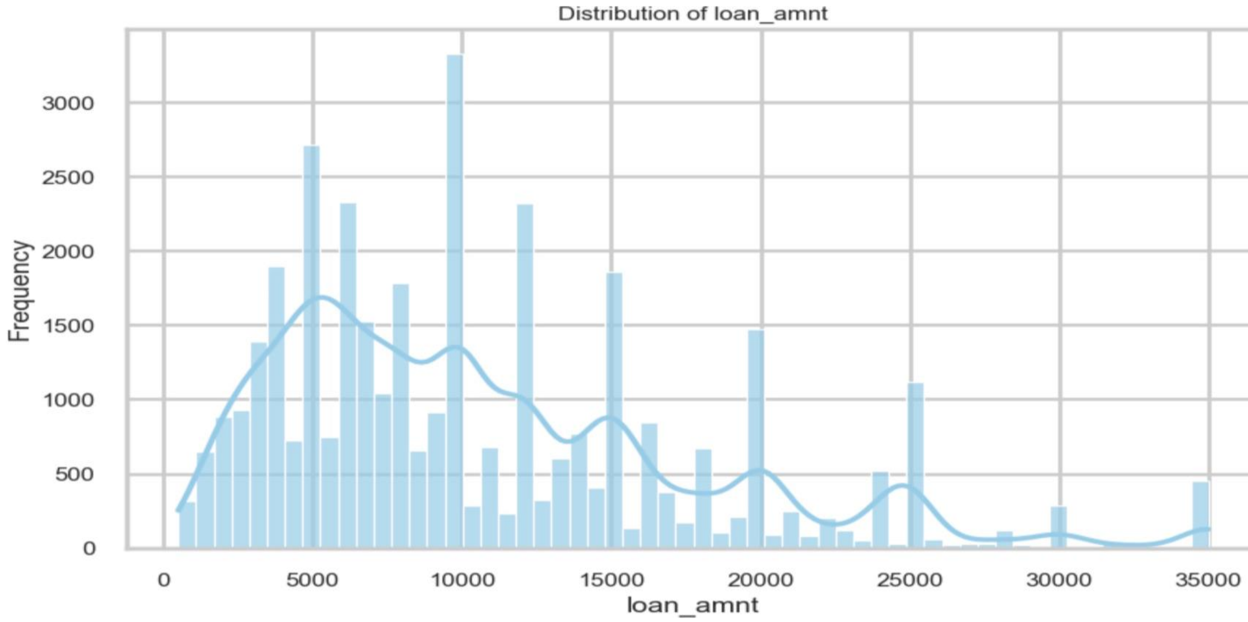
1. Loan Amount (loan_amnt)
2. Interest Rate (int_rate)
3. Loan Term (term)
4. Loan Grade (grade)
5. Employment Length (emp_length)
6. Home Ownership (home_ownership)
7. Annual Income (annual_inc)
8. Debt-to-Income Ratio (dti)
9. Loan Status (loan_status)
10. Issue Date (issue_d)
11. Issue Date Year (issue_d_year)
12. Issue Date Month (issue_d_month)

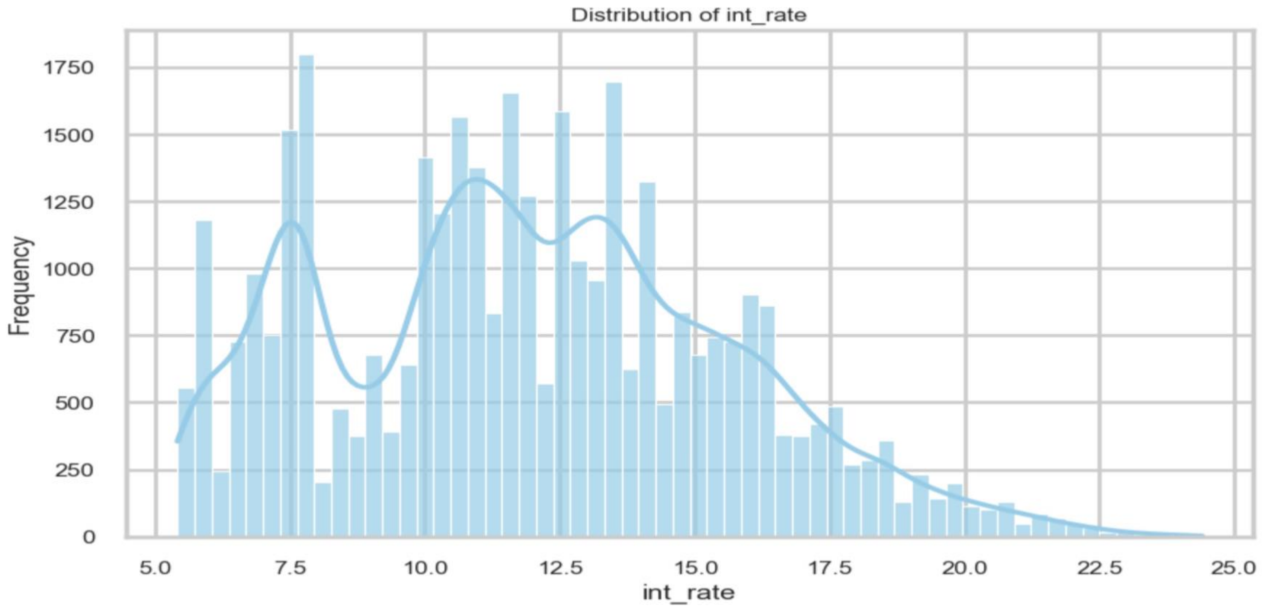


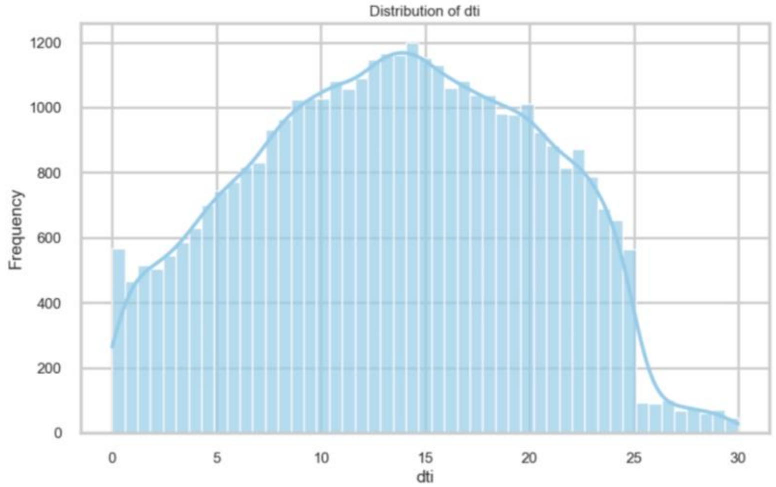
Looking at the figure, we can say that the number of loans issued over time increased almost exponentially from 2007 to 2011. This could suggest that the Lending Club business is expanding, and thus presents a greater need to be cautious about who loans are being issued to.



From this plot, we can infer that the number of loans steadily increases towards the end of the month. This could be as people are gearing up for holiday season or could also be because people usually like to make bigger purchases at the start of a new year.



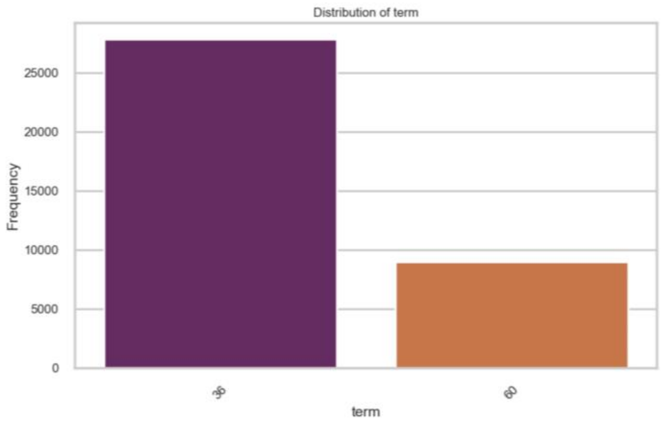


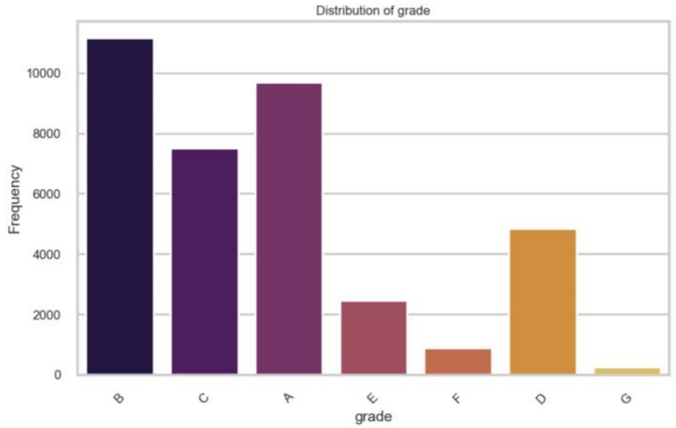


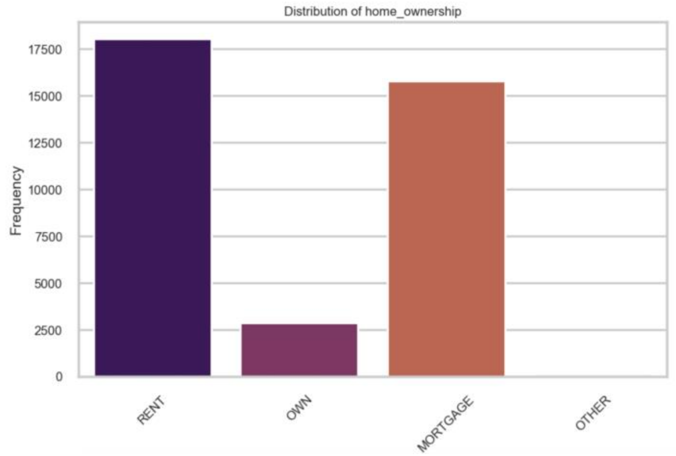
1. Loan Amount: Most loans are between 0-10,000$, however, there are also some loans that are higher in amounts and might indicate a higher risk if not managed properly. We also notice spikes in amounts that are multiples of 5000, as most people probably like to round of how much money they are borrowing in increments of 5000.

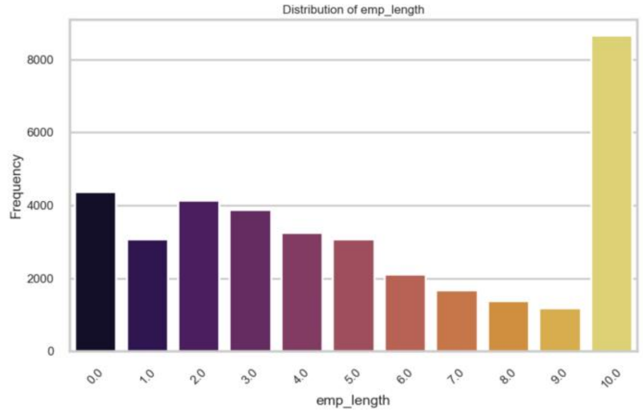
2. DTI: Ideally, a debt-to-income (DTI) ratio below 36% is deemed favorable and indicative of financial health. Most DTI ratios observed fall within a commendable interquartile range (IQR) of 8.3% to 18.7%. This suggests that the borrowers are generally opting for loans that are financially manageable and unlikely to pose repayment challenges.

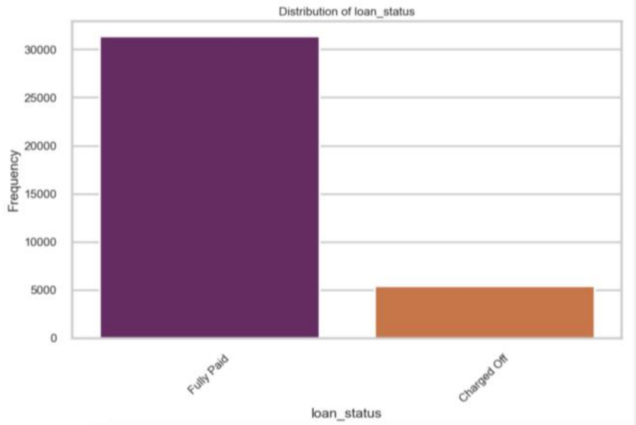
3. Interest Rates: The mean interest rate is 11.8%, whereas the highest interest rate is 24.4%. This is a very high interest rate and could indicate that some of the borrowers have bad credit scores and are taking risky loans. It could also be related to the Loan amount, as the graphs follow similar patters. Higher loan amounts result in higher interest rates.











1. **Grade**: Although most loans are between grades of A, B and C, there are a substantial number of grade D and below that could have higher default rates

2. **Loan Duration**: Although most loans are of 36 months, there a significant number of long term (60 month) loans, these could be riskier compared to short term ones.

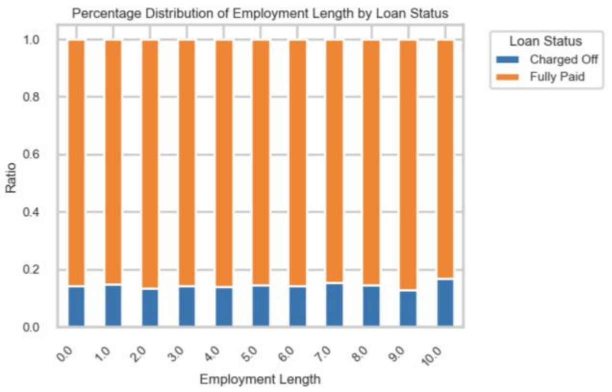
3. **Home Ownership**: Renters could potentially have higher default rates as compared to owners. Further, owners also have a higher collateral, and could get better loans with lower interest rates.

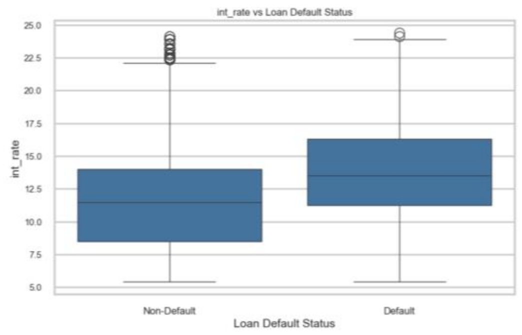
4. **Employment Length**: Borrowers who have been working for a lesser period could be associated with higher default rates as compared to those who have been working for longer. We see that people who have been working 10 or more years correspond to most of our borrowers.

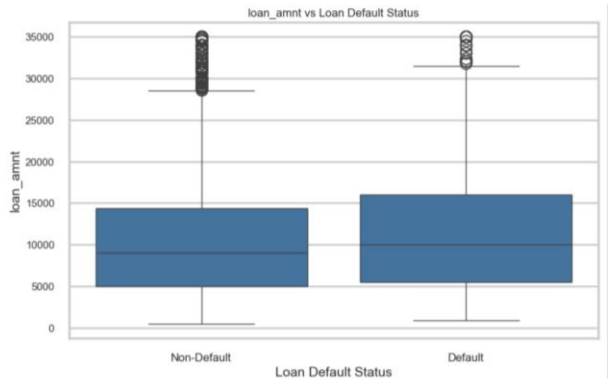
Bivariate Analysis

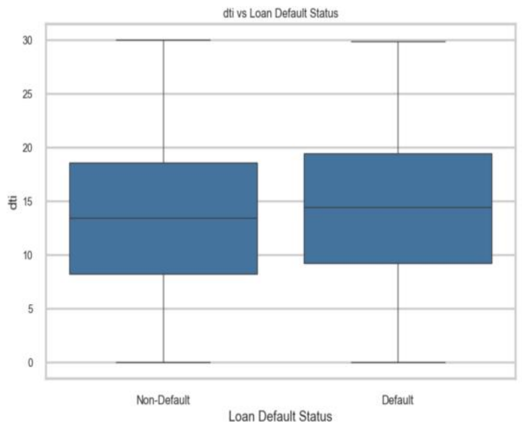
The following bivariate analysis will be considered:

1. Quantitative Variables
   A. Loan Amount vs Loan Status (loan_amnt vs loan_status)
   B. Interest Rate vs Loan Status (int_rate vs loan_status)
   C. Annual Income vs Loan Status (annual_inc vs loan_status)
   D. Debt-to-Income Ratio vs Loan Status (dti vs loan_status)
   E. Employment Length vs Loan Status (emp_length vs loan_status)
2. Categorical Variables
   A. Home Ownership vs Loan Status (home_ownership vs loan_status)
   B. Grade vs Loan Status (grade vs loan_status)
   C. Verification Status vs Loan Status (verification_status vs loan_status)
   D. Term Length vs Loan Status (term vs loan_status)
   E. Address State vs Loan Status (addr_state vs loan_status)
   F. Loan Purpose vs Loan Status (purpose vs loan_status)







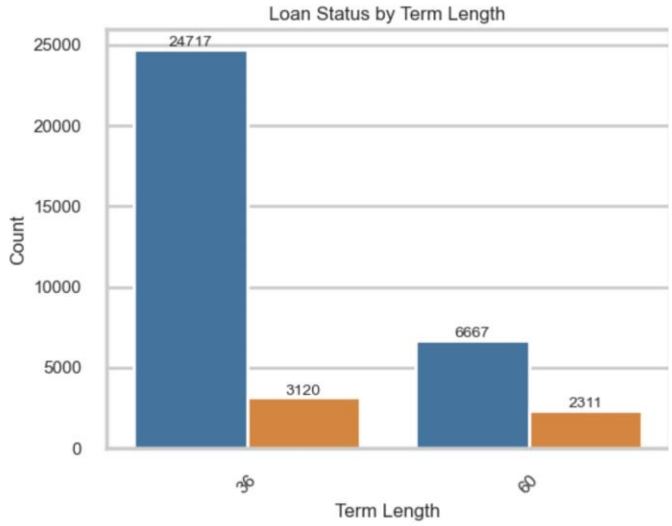


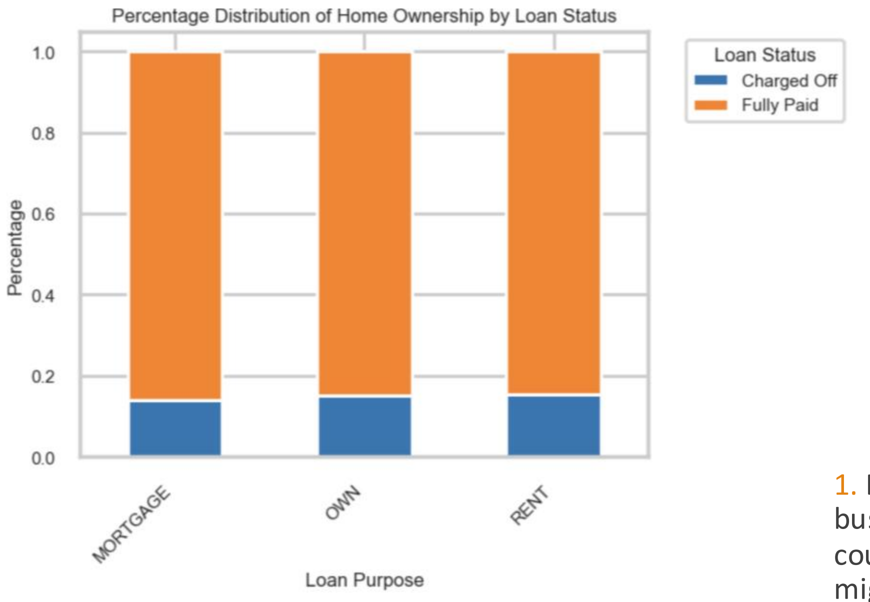
1. Higher loan amounts tend to default more than lower loan amounts. Further, higher interest rates also contribute to a higher default rate. This could mean that the people who take higher loan amounts are usually the ones who also receive a higher interest rate, and thus are more likely to default due to the higher loan burden.

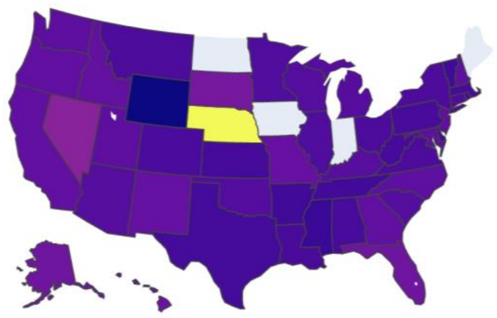
2. Similarly, a higher DTI usually results in a higher default rate as seen from the graph.

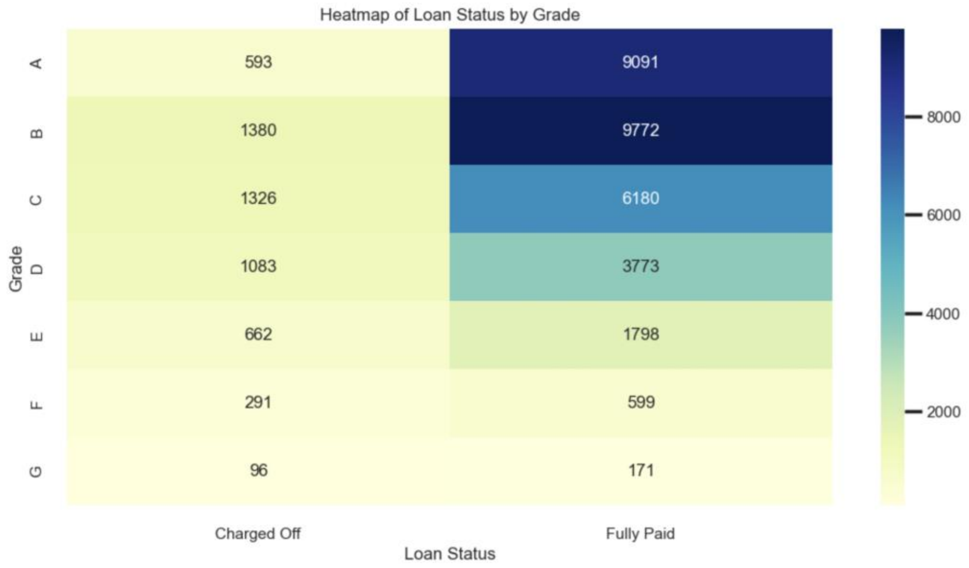
3. The median income of people who defaulted is about 10,000$ lesser than the median income of people who fully paid.

4. Although most of the borrowers who defaulted have been employed for 10 or more years, Employment Length is not much of a differentiator for knowing whether borrowers will default or not. When we plot ratios, they are pretty much equal for all lengths of employment.









1. Borrowers whose primary purpose for the loan is "small business" default significantly more than other categories. This could be since a business is naturally a risky venture, and it might be possible that several of these business have gone bust, resulting in borrowers not making loan payments on time.

2. From the state-map, we notice that Nebraska has a very high default rate of 0.6, followed by Nevada (0.22) and New Mexico (0.17).

3. We also notice that longer loan term of 60 months are a lot more like to be defaulted on as compared to shorter terms of 36 months.

4. In terms of home ownership, most of the borrowers are Renters or are paying Mortgages, and very few people own homes. Although the number of renters who default is high, the ratio is mostly the same and thus is not conclusive enough evidence.

5. Out of all the loan grades that were charged off, grade B, C and D have the highest default rates.

Thank You