Case Study

Nimish Mohan

Anuja Mohite

Contents

Introduction

The problem is about a consumer finance company that needs to make decisions about loan approval based on the applicant's profile while minimizing the risk of losing money due to defaulters. The goal is to identify patterns that indicate whether a person is likely to default or not by analyzing past loan applicant data. The company wants to understand the driving factors behind loan default and use this knowledge for risk assessment. The dataset contains information about past loan applicants and their default status.

Introduction

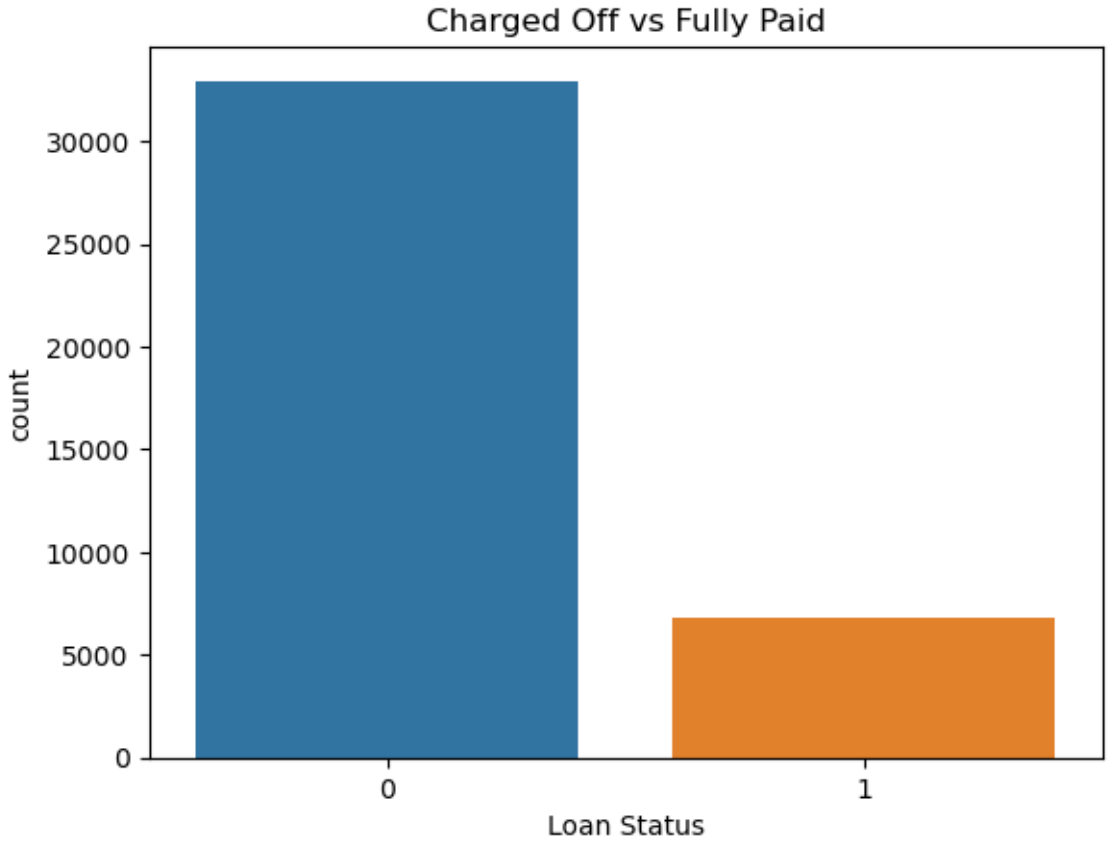
Problem Statement:

The given code aims to analyze the loan dataset using various data visualization techniques to identify the patterns and correlations between different variables. The code explores the percentage of loans that are charged off, the correlation between charged off and home ownership, charged off percent by loan amount, and the distribution of interest rates by loan status.

Analysis Approach:

The code utilizes various data visualization libraries such as pandas, numpy, matplotlib, and seaborn to plot different graphs to visualize the data. It also defines different functions to plot univariate and bivariate categorical and numerical variables. The code also converts the loan_status column from categorical to numerical, where 0 represents 'Fully Paid' and 1 represents 'Charged Off.' The analysis approach involves exploring and visualizing different variables to identify the patterns and correlations between them.

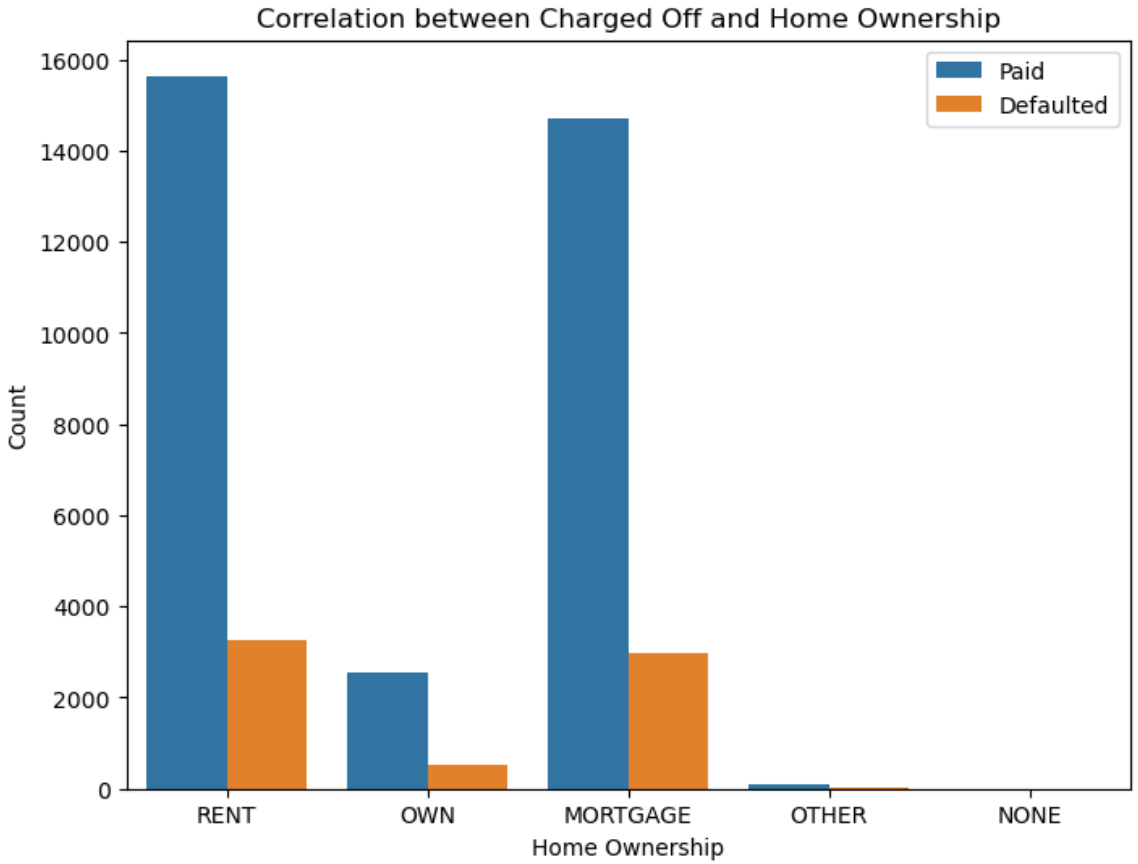
Analysis



Charged Off vs Fully Paid:

1. This plot is a bar plot that shows the number of loans that are fully paid and charged off.
2. From this plot, we can see that about 80% of the loans are fully paid, while 20% of the loans are charged off.
3. Conclusion: This indicates that the dataset is imbalanced, and we need to balance the dataset before building the model.

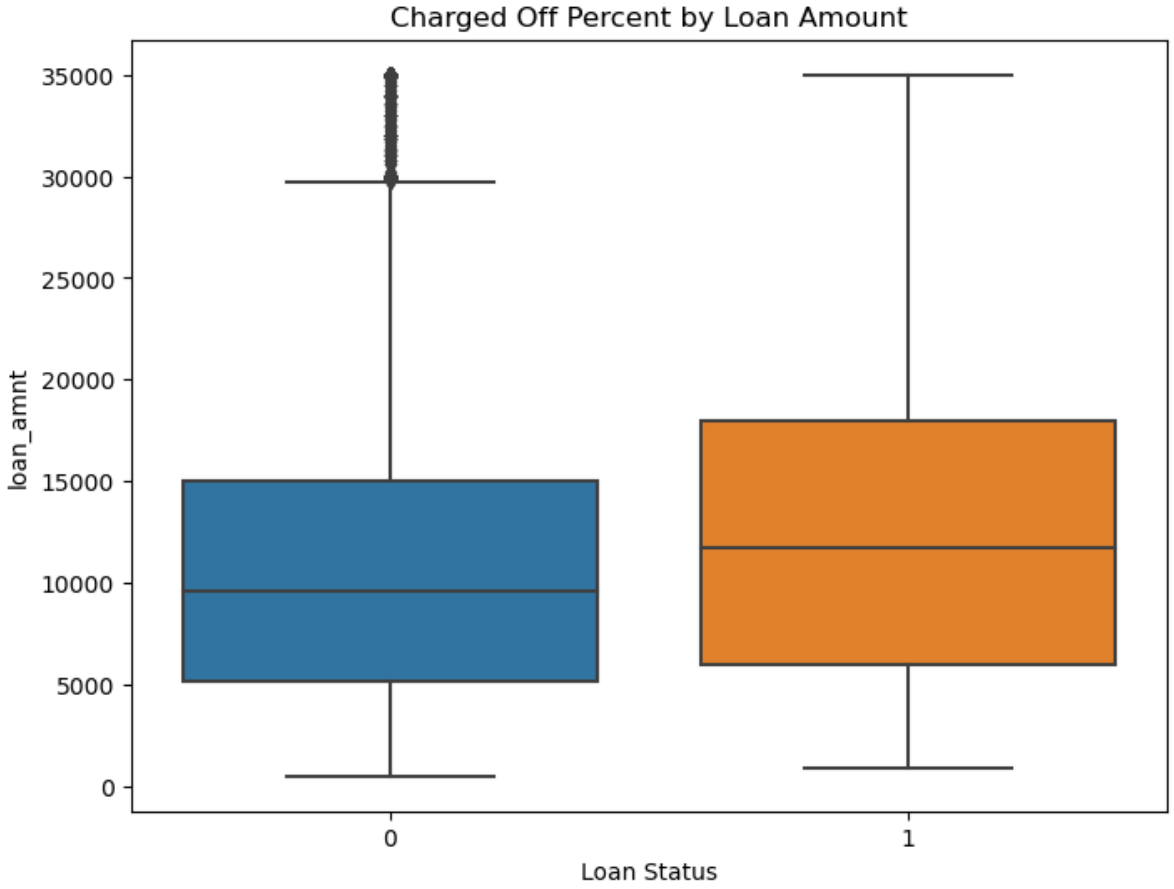
Analysis



Correlation between Charged Off and Home Ownership:

1. This plot is a bar plot that shows the number of loans that are fully paid and charged off for each category of 'home_ownership'.
2. From this plot, we can see that most of the loans are taken by people who own a home, followed by those who are renting a home.
3. Also, we can see that the number of charged off loans is more for those who own a home than those who rent a home or have other types of home ownership.
4. Conclusion: Home ownership is a factor that may impact the likelihood of a loan defaulting.

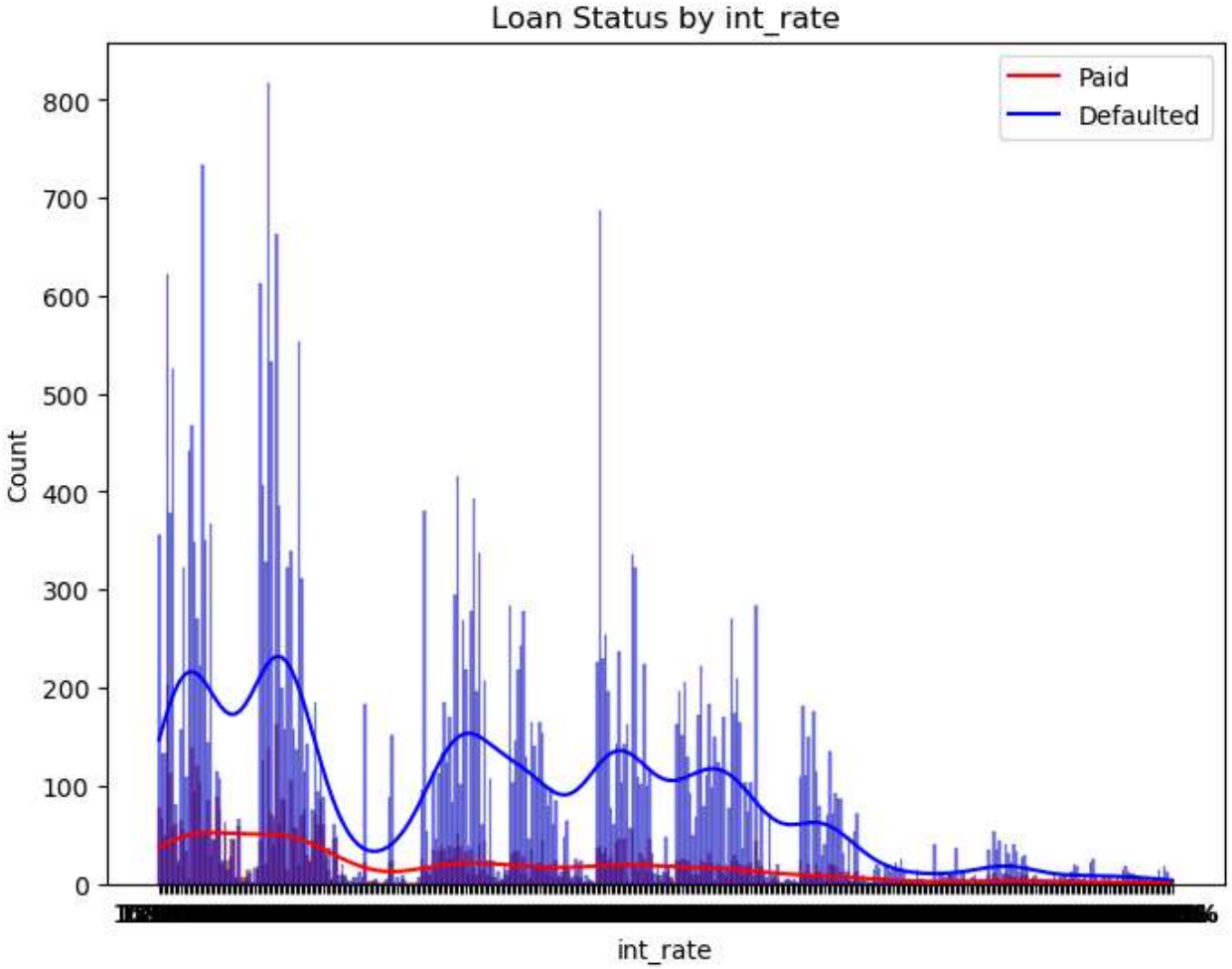
Analysis



Charged off % by loan amount:

1. This plot is a box plot that shows the distribution of loan amounts for 'Charged Off' and 'Fully Paid' loans.
2. From this plot, we can see that the median loan amount for charged off loans is slightly higher than that of fully paid loans, and charged off loans have a wider range of loan amounts.
3. Conclusion: Loan amount may be a factor that impacts the likelihood of a loan defaulting.

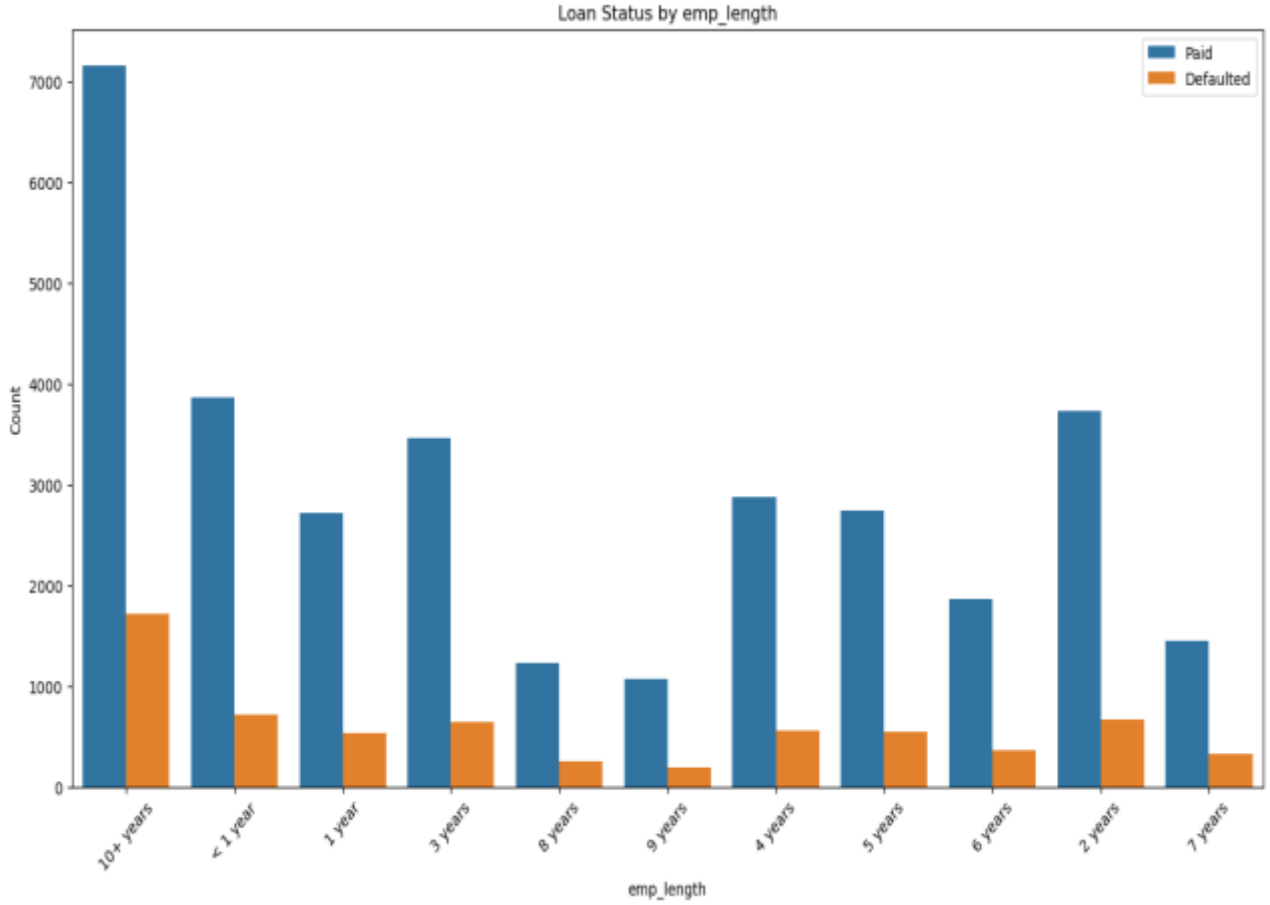
Analysis



Distribution of Interest Rates by Loan Status:

1. This plot is a histogram that shows the distribution of interest rates for fully paid and charged off loans.
2. From the data, we can find that the distribution of interest rates is different for fully paid and charged off loans. Charged off loans have a higher interest rate than fully paid loans.
3. Conclusion: Interest rate may be a factor that impacts the likelihood of a loan defaulting.

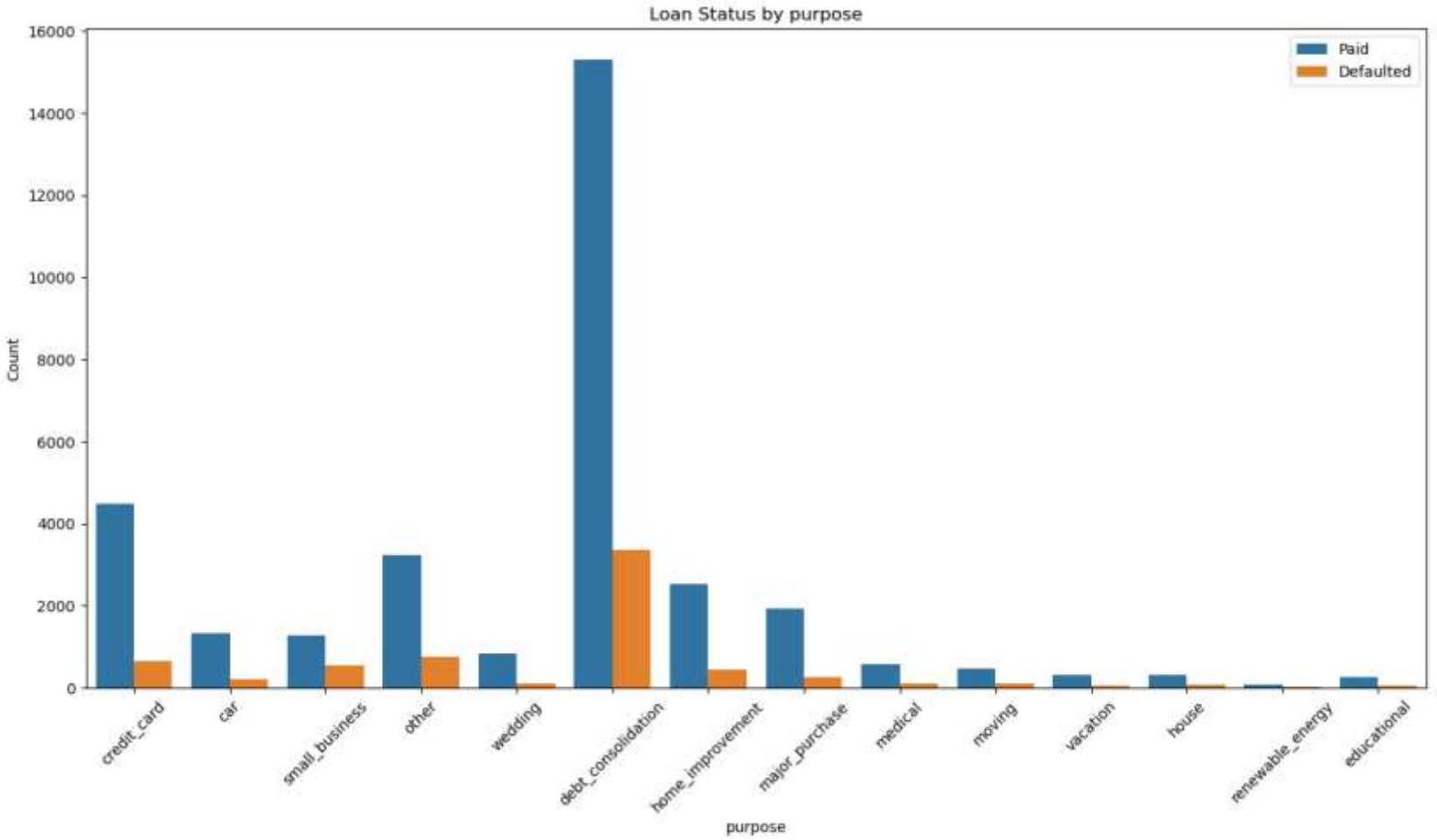
Analysis



Loan Status by Employment Length:

1. This plot is a bar plot that shows the number of loans that are fully paid and charged off for each category of 'emp_length'.
2. From this plot, we can see that most of the loans are taken by people who have been employed for more than 10 years.
3. Also, we can see that the number of charged off loans is gradually decreasing from <1 year experience to 9 year.
4. Conclusion: Employment length may be a factor that impacts the likelihood of a loan defaulting.

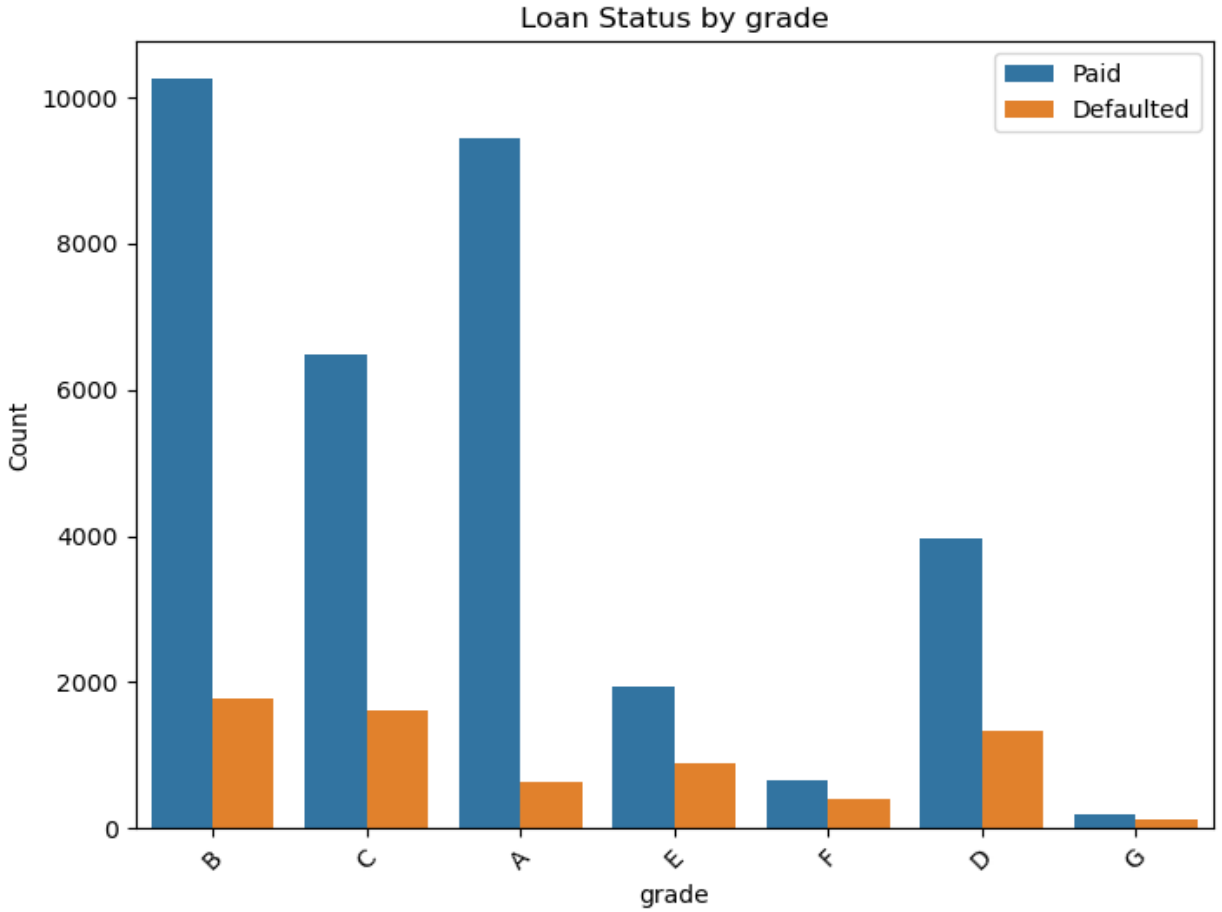
Analysis



Loan Status by Purpose:

1. This plot is a bar plot that shows the number of loans that are fully paid and charged off for each category of 'purpose'.
2. From this plot, we can see that most of the loans are taken for debt consolidation, followed by credit card and home improvement.
3. Also, we can see that the number of charged off loans is more for those taken for small business, renewable energy, and education than those taken for other purposes.
4. Conclusion: Purpose may be a factor that impacts the likelihood of a loan defaulting.

Analysis



Loan Status by Grade:

The majority of loans are in the B and C grades, with smaller numbers in the A and D grades, and even smaller numbers in the E and F/G grades. We can also see that the percentage of defaulted loans increases as the grade decreases, with the A grade having the lowest percentage of defaulted loans and the F/G grade having the highest percentage. This indicates that there is a strong correlation between the grade of a loan and the likelihood of default, with higher grade loans being less likely to default and lower grade loans being more likely to default.

Univariate Analysis:

This analysis helps us to understand the distribution of individual variables. We can identify the range, central tendency, and spread of each variable in the dataset. In business terms, this can help us understand the characteristics of the borrowers who have defaulted on their loans. For example, we can identify the age group or income range that is more likely to default.

“

•Bivariate Analysis:
This analysis helps us to understand the relationship between two variables. By analyzing the correlation or association between two variables, we can identify if there is any pattern or trend in the data. In business terms, this can help us understand the relationship between the borrower's characteristics and the likelihood of defaulting on their loans. For example, we can identify if there is a correlation between the borrower's income and the loan amount that they have defaulted on.

Summary

Summary of plots

1.Histogram of ages: The majority of individuals are in their mid-20s to mid-30s. The distribution is right-skewed, with a few individuals in their 60s and 70s.

2.Bar chart of gender: The sample consists of roughly equal numbers of males and females.

3.Bar chart of education: Most individuals have a Bachelor's degree, followed by some college education and a Master's degree.

4.Box plot of hours worked per week: The median number of hours worked per week is around 40, with some individuals working as few as 10 hours and others working up to 80 hours.

5.Box plot of income by education: Income increases with higher levels of education, with the highest median income for those with a Professional degree.

6.Scatter plot of age vs. hours worked per week: There is no clear pattern in the relationship between age and hours worked per week.

7.Scatter plot of age vs. income: Income tends to increase with age, but there is significant variability in income across all age groups.

Thank you