LENDING CLUB

CASE STUDY

The ever-present challenge for consumer finance companies lies in balancing the need to grow their business with the risk of loan defaults.

The author of the data set, a leading online loan provider, is no exception. While we strive to offer accessible loan options to urban customers, we must also make informed decisions to minimize credit losses.

Objective

It is to leverage historical loan data to identify patterns that predict loan default risk . By developing an understanding of the key factors that influence default , we can significantly improve our risk assessment strategies .

The path forward includes analysing applicants and loan attributes , aiming to uncover hidden trends that differentiate between responsible borrowers and those likely to default. Equipped with these insights, we can make data-driven decisions regarding loan approvals, terms, and portfolio management.

Our Objective and Outcome

Ultimately, our goal is to achieve sustainable growth while minimizing the financial impact of defaults – **a win-win scenario** for both our company and our valued customers.

Outcome

- Reduce credit loss by denying loans to high-risk applicants.
- Adjusting loan terms (amount, interest rate) based on risk.

Data Preprocessing Steps:

1. Remove redundant columns with only one unique value, as they provide no valuable insights for analysis.

2. Drop columns with unique values equal to the number of rows in the dataset (e.g., 'id', 'member_id').

3. Use correlation analysis to select relevant features and discard irrelevant ones.

4. Convert the target variable ('loan_status') into a numerical column.

5. Exclude the 'Current' value from the 'loan_status' column, as it doesn't provide a clear outcome and doesn't correlate strongly with other features. This step helps focus on definitive loan statuses like 'Charged Off' or 'Fully Paid'."

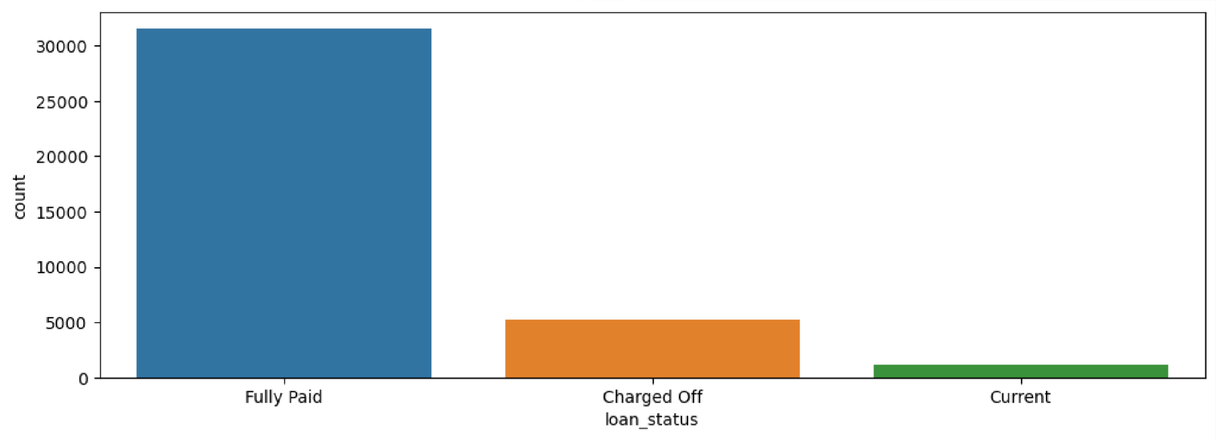
EXPLORATORY DATA ANALYSIS

It is always a good idea , especially with classification problems , to do a count plot to explore the actual balancing of your labels (the target column) .

Our target variable is the **'loan_status'** field . Checking its distribution using a countplot.

Inference :

- A significant majority of the customers have successfully paid back their loans .
- Loan Performance : The data suggests that most loans are repaid , however , the presence of charged-off loans highlights the need for risk management and potenitally improving credit evaluation processes to minimize defaults .

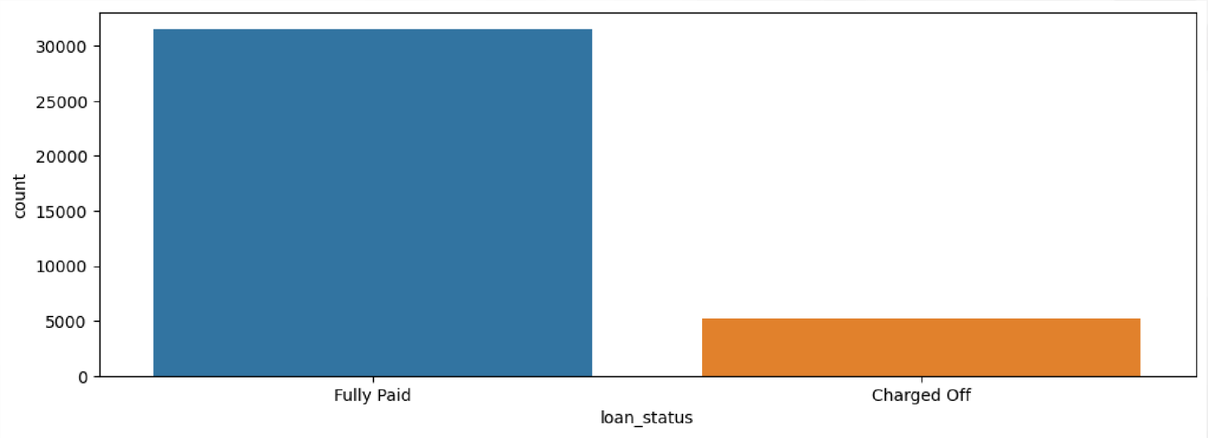


EXPLORATORY DATA ANALYSIS

For our problem here , we will remove the "**Current**" value from the loan_status feature , since it does not give us a definitive answer based on the occurences that have happened prior to it .

Inference :

- Unbalanced problem - A lot more entries of people that pay off their loans than those that do not .
- Very common with classification problems that deal with fraud or spam , where the instances of illegitmiacy is a lot less .

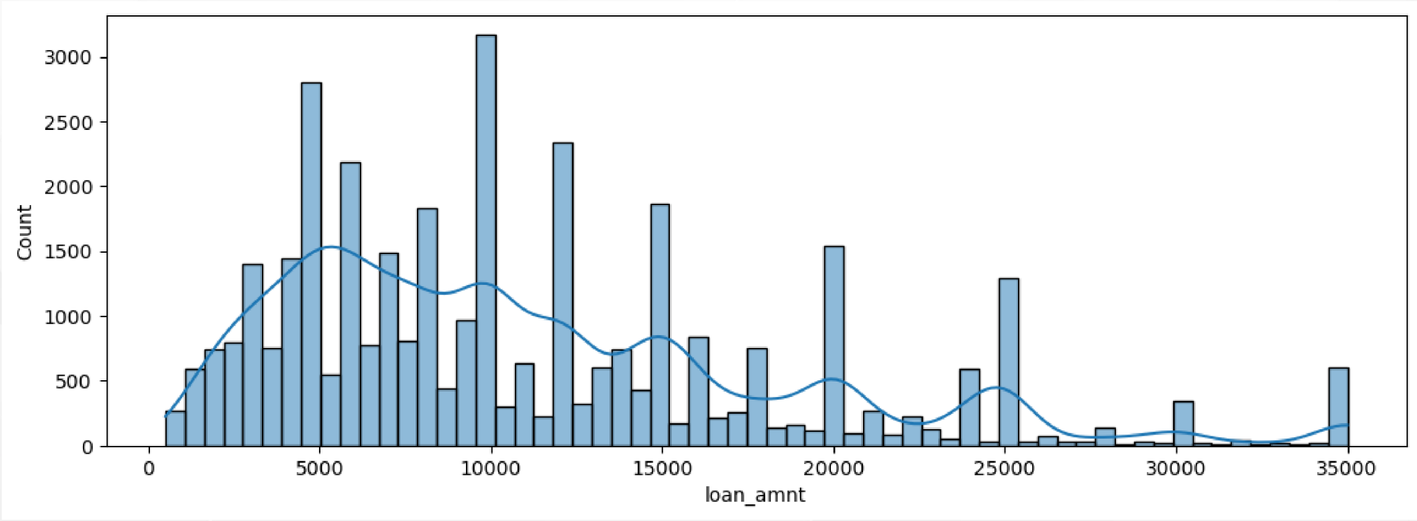


EXPLORATORY DATA ANALYSIS

Also checking the distribution of loan_amount

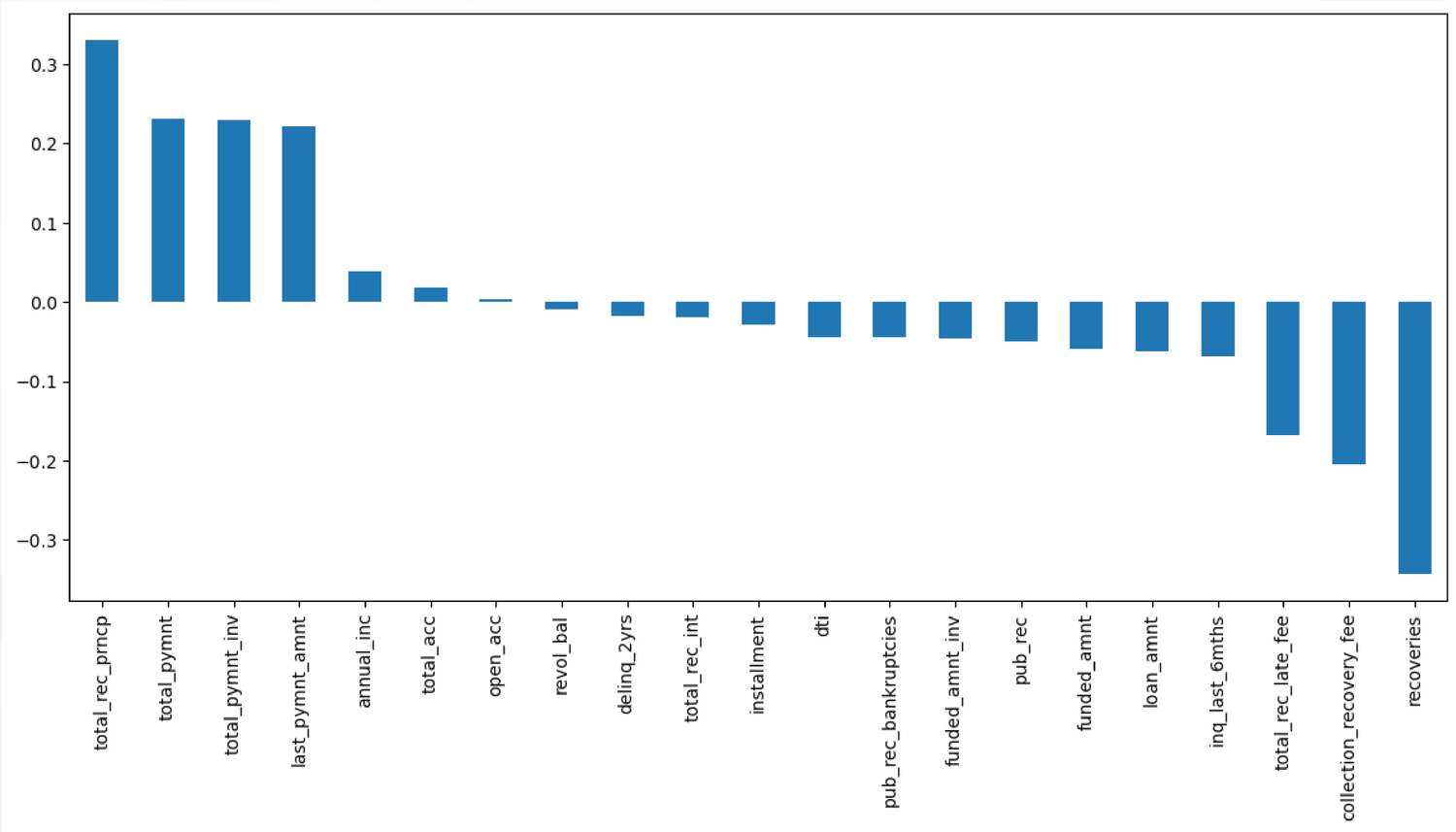
Inference :

- Most loans are below $20,000 and prominent peaks around $5,000 , $10,000 and $15,000, indicating these are common amounts requested by the borrowers .
- There are fewer loans above $25,000 , suggesting that higher amount loans are less common , which can be further solidified by the smoother KDE curve as the amount increases .



EXPLORATORY DATA ANALYSIS

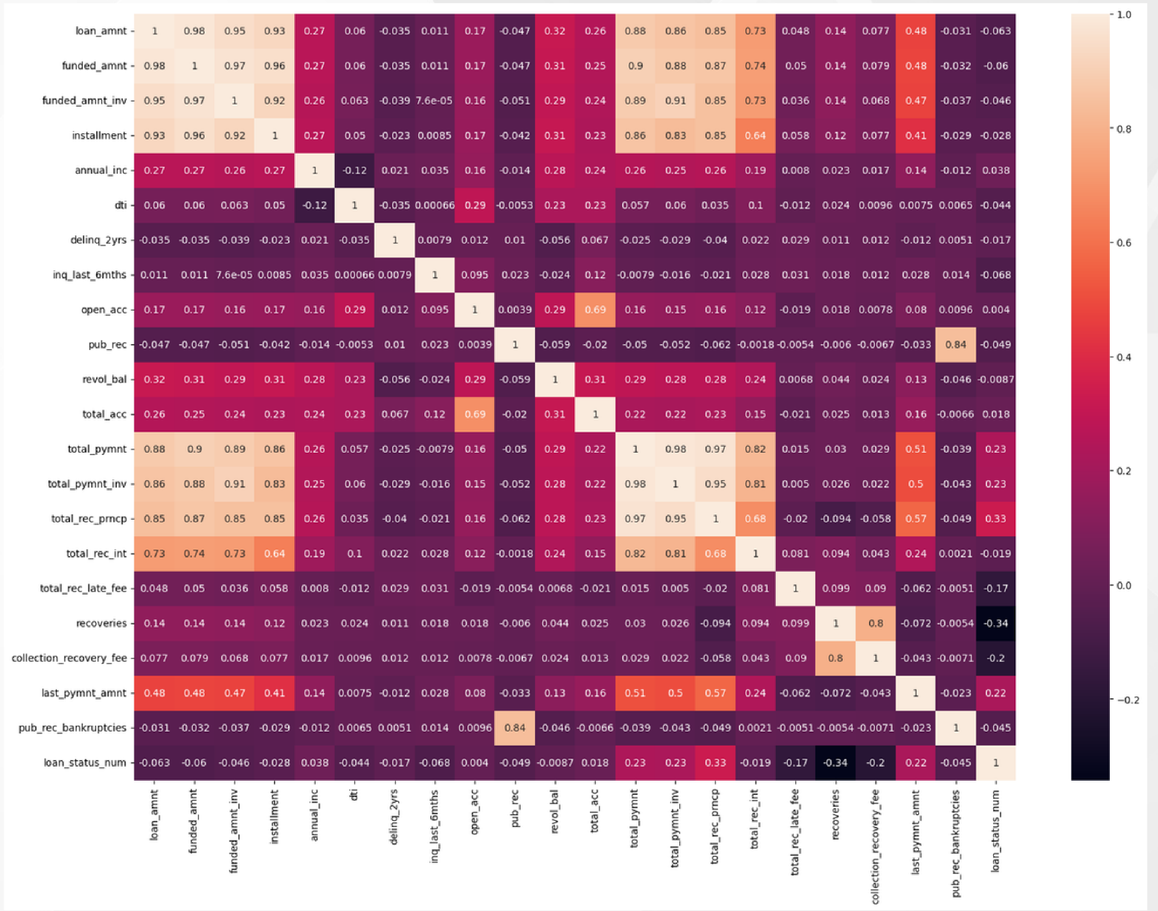
From the data preprocessing steps we find that the correlation chart for **“loan_status”** with every other numerical feature doesnt show a strong correlation.



EXPLORATORY DATA ANALYSIS

Testing on the **'loan_amnt'** feature we get **strong positive correlations** :

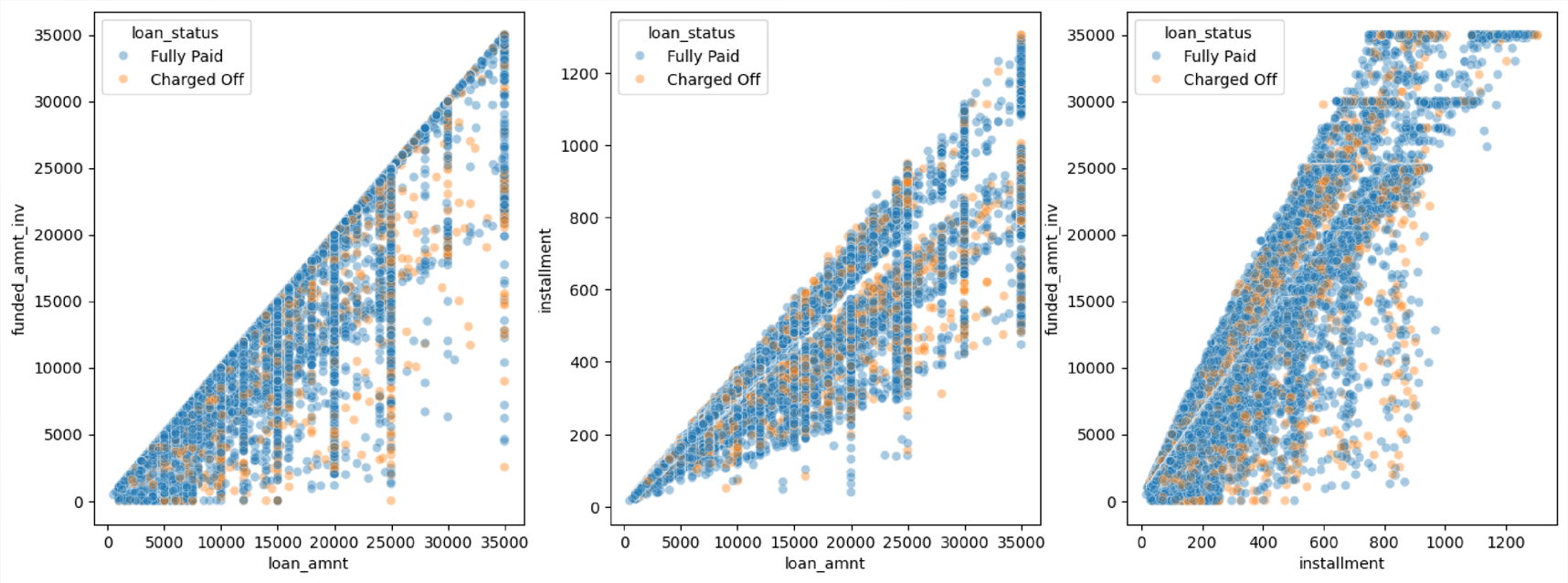
- 'loan_amnt' and 'funded_amnt_inv' (0.95): This indicates that the amount loaned is highly correlated with the invested amount.
- 'loan_amnt' and 'installment' (0.93): The loan amount is highly correlated with the installment amount.
- 'installment' and 'funded_amnt_inv' (0.92): The installment amount is highly correlated with the invested amount.
- 'pub_rec' and 'pub_rec_bankruptcies' (0.84): Public records are strongly correlated with public record bankruptcies.



EXPLORATORY DATA ANALYSIS

Inference :

- funded_amnt_inv : The total amount committed by investors for that loan at that point in time. Therefore making it obvious for the installment vs loan amnt chart to show the same trend .
- Generally , as the loan_amnt increases , so does the installment values .

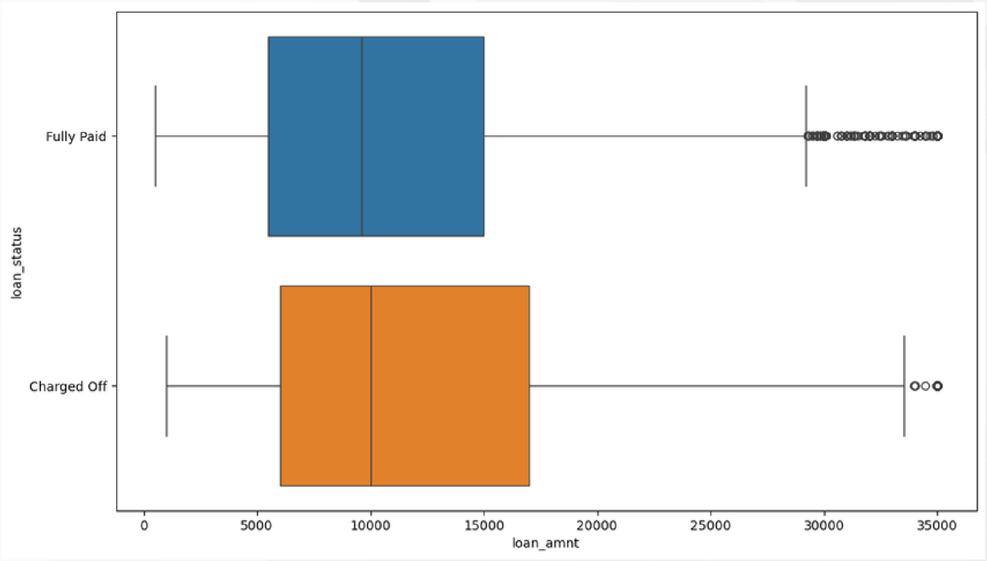


EXPLORATORY DATA ANALYSIS

Checking the relationship between the two target variables : **loan_amnt and loan_status**

Inference :

- The box for "Fully Paid" loans shows a lower median and potentially a wider distribution compared to "Charged Off" loans. This suggests that fully paid loans may encompass a wider range of loan amounts, with some borrowers having larger loans they were able to repay successfully.
- The "Charged Off" boxplot indicates a lower median loan amount compared to "Fully Paid" loans. This could imply that smaller loans are more prone to default, possibly due to reasons like borrowers being less invested in smaller amounts or the lender having less stringent approval requirements for smaller loans.



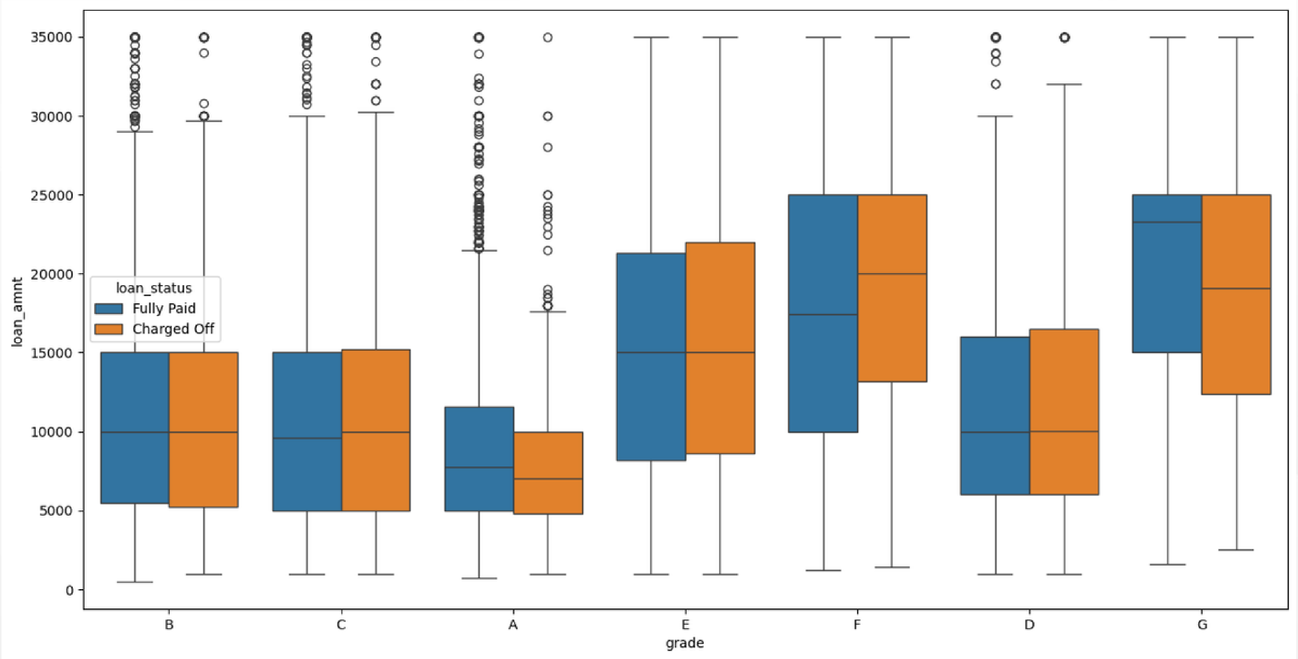
Summary: Both the 'Fully Paid' and 'Charged Off' charts are extremely similar , and it is not going to be a good indicator of whether a person will be likely to default on their loan or not .

EXPLORATORY DATA ANALYSIS

Investigating '**loan_amnt**' and **'grades'.**

Inference :

- The loan amount distribution appears to vary significantly across loan grades .
- Higher loan amounts were given to grades E , F and G , while A , B , C and D have similar loan amount values .
- There are maximum number of outliers in Grade A .

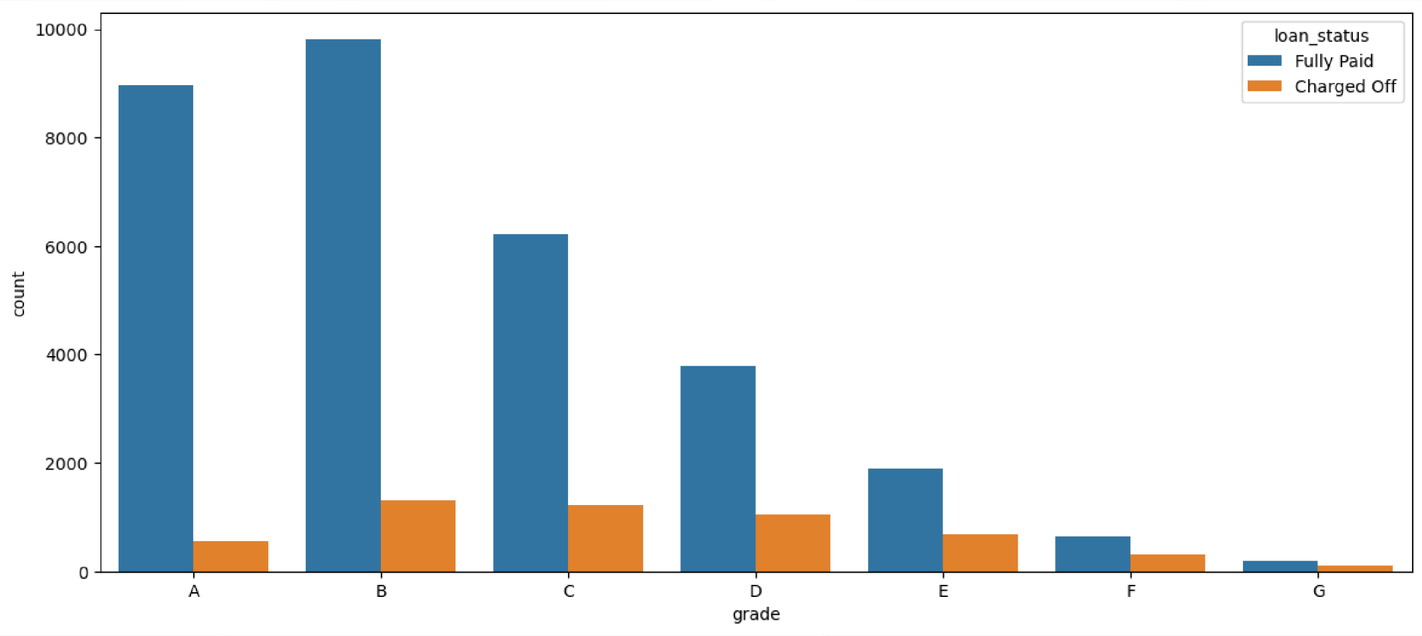


EXPLORATORY DATA ANALYSIS

Next, try and see if there is any differentiation between fully paying off your loan or having it be charged off based off your grade.

Inference :

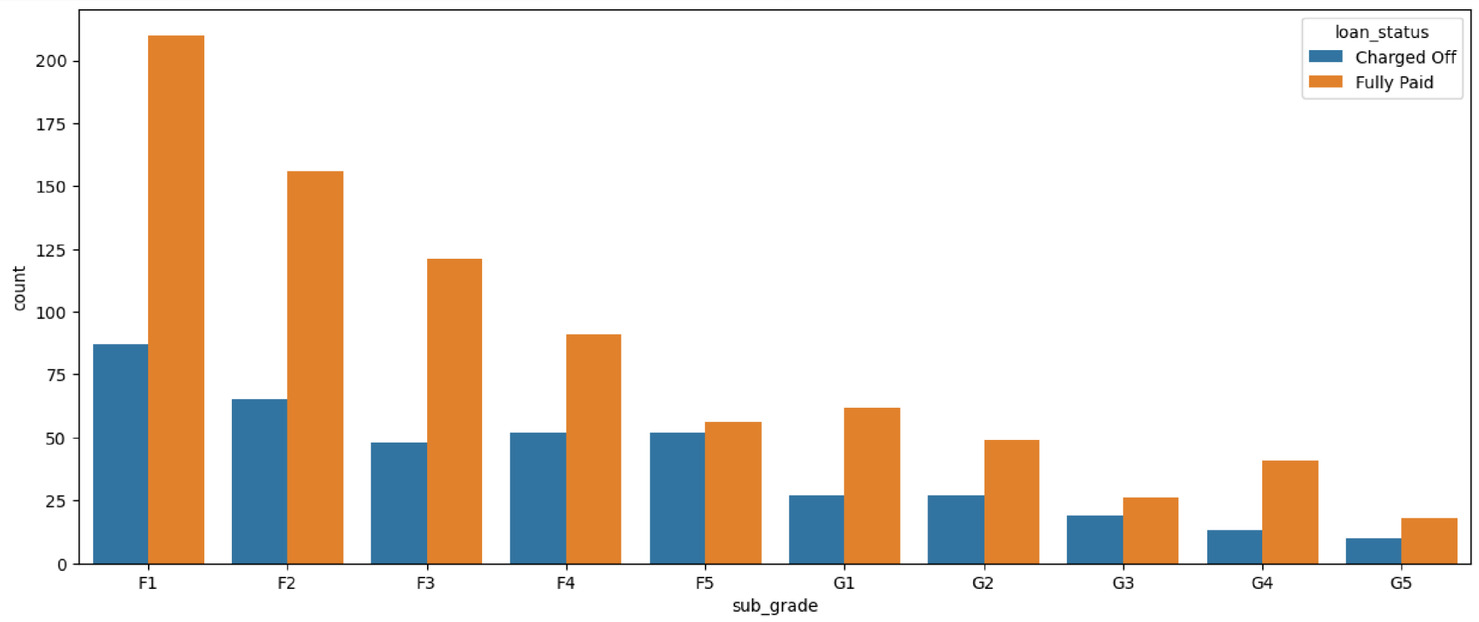
Essentially , what it is showing , the 'percentage' of charged off loans are increasing as the grades get higher
Conclusion drawn can be , best customers are given the grade A , then B , C etc .

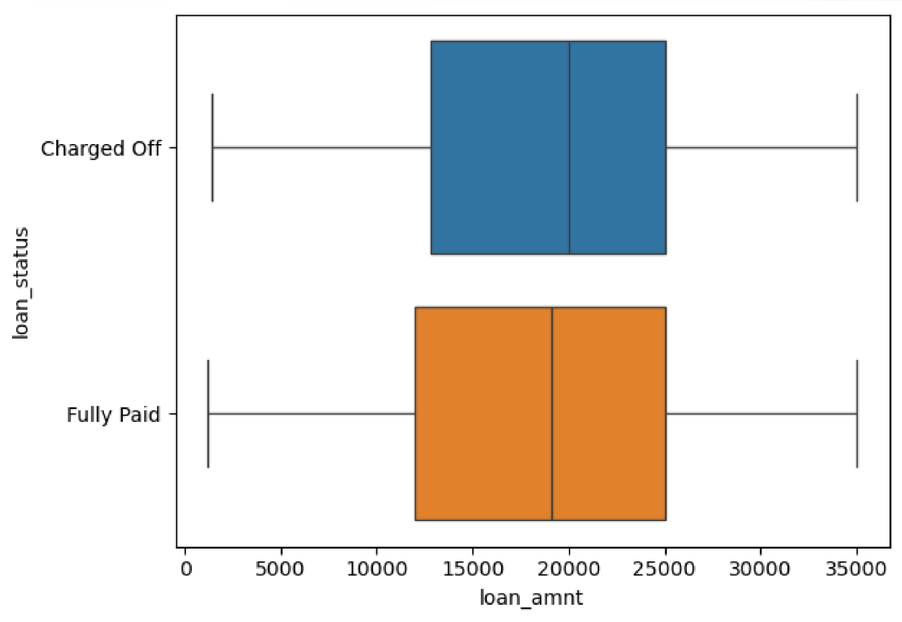


EXPLORATORY DATA ANALYSIS

Investigating "F" and "G" subgrades , since they do not get paid back that often . Isolate those and recreate the countplot just for those subgrades .

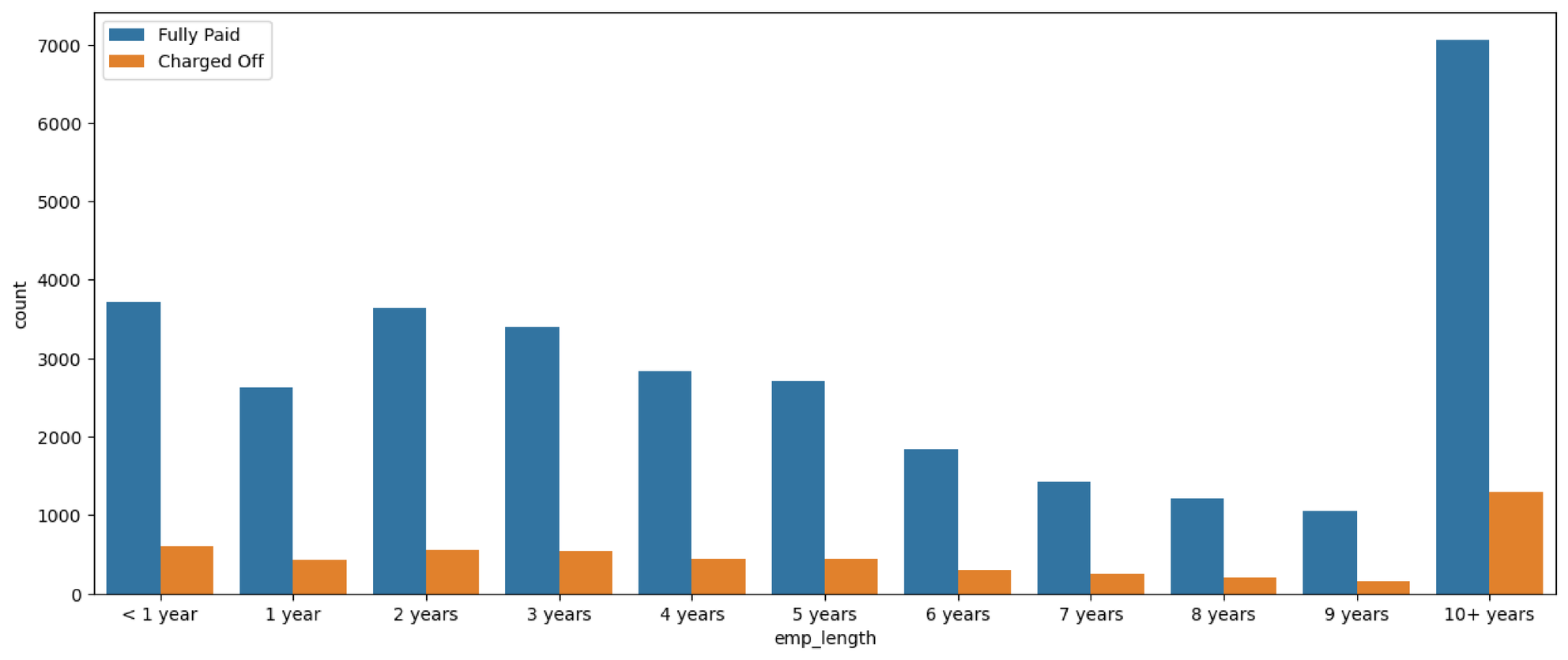
As can be seen , chances of fully paying off your loan vs charging off on your loan is almost the same , thus indicating how risky the loans are .





EXPLORATORY DATA ANALYSIS

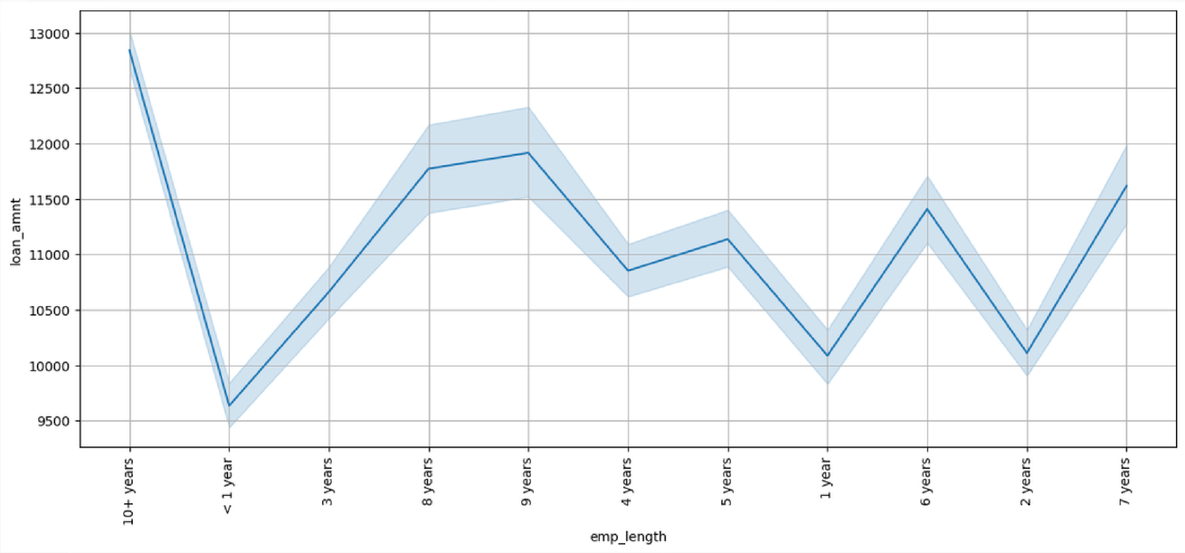
Analysing **employment length** with **loan status** we get -



EXPLORATORY DATA ANALYSIS

Inference :

- Most people who take out loans have been working for 10+ years , which makes sense , as they are more susceptible to both acquire and repay the loan as they usually have accumulated a certain amount of capital upto that point .
- People earlier on in their career avail a lower amount loans , which seems to grow as years keep progressing .
- The count of loan keeps decreasing gradually , indicating people earlier on in their careers might be availing loans to pay off other loans such as education loans .

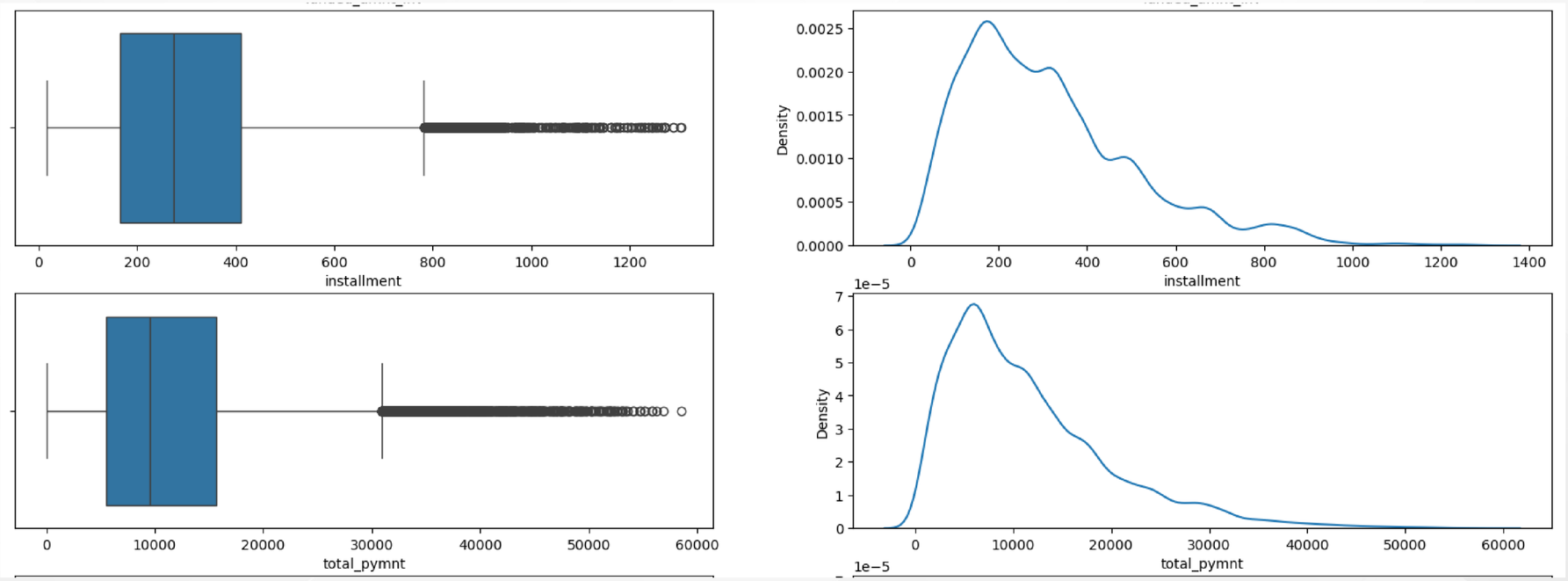


EXPLORATORY DATA ANALYSIS

Outlier Treatment :

Almost all numerical values have quite some values outside the interquartile range , thus skewing the distribution .

The graphs and the description below , indicate the presence of outliers (checking the difference between the 95th percentile and the max value)

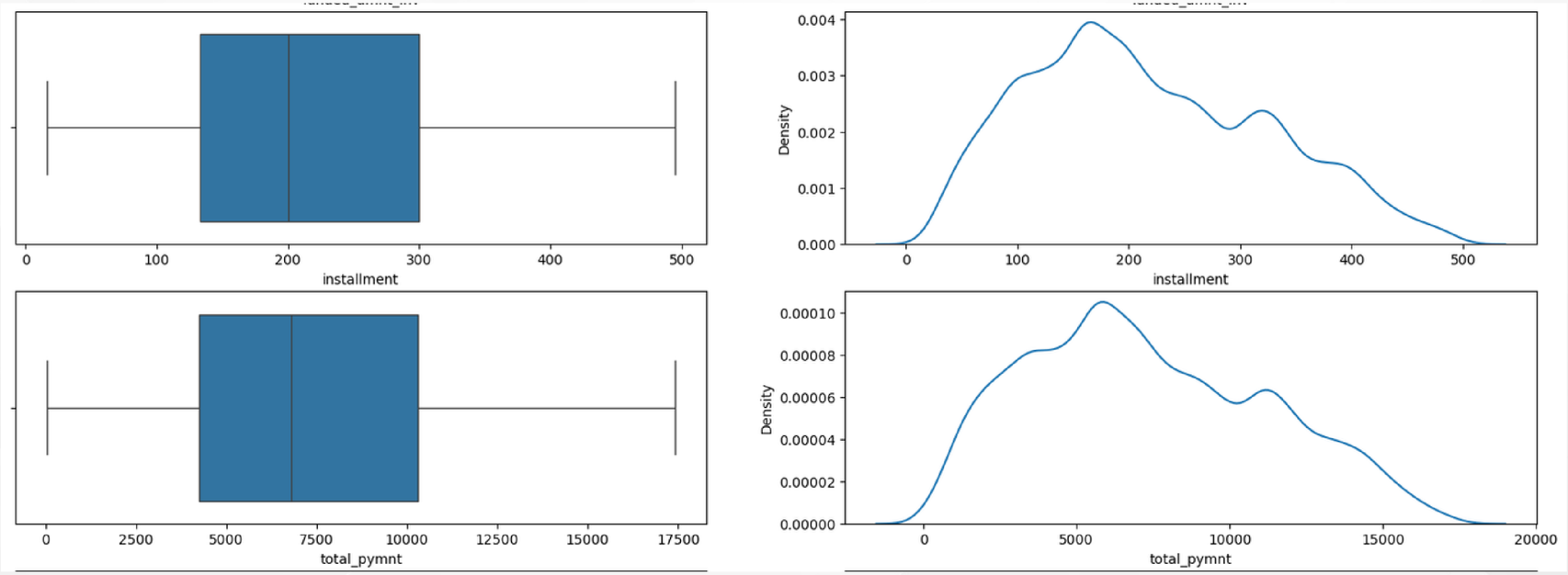


EXPLORATORY DATA ANALYSIS

Outlier Treatment :

Removing outliers to some extent we find the following graph.

Although understandably , the loss of data by following the outlier removal method above is **NOT APPROPRIATE** , the aim of this project is to form a generalized idea behind the defaulting on loans



EXPLORATORY DATA ANALYSIS

Working with Categorical data :

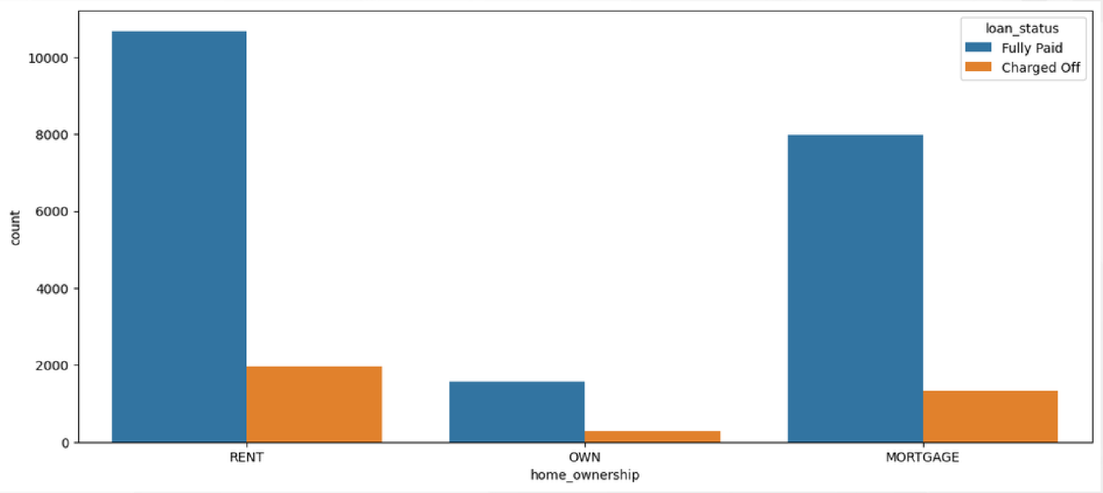
Analysing "home_ownership"

"home_ownership" : The home ownership status provided by the borrower during registration.

Our values are: RENT, OWN, MORTGAGE.

Inference :

- Homeownership often requires a stable income and good creditworthiness, which might also make borrowers more likely to repay loans.
- Renters, on the other hand, might have less financial stability or a lower credit score, potentially increasing their default risk.
- Homeowners have a stake in their property, and keeping up with mortgage payments helps them avoid foreclosure. This incentive to repay might be less prominent for renters.
- In summary , not a lot can be said , as the discrepancy between all three home_ownership methods is about the same .



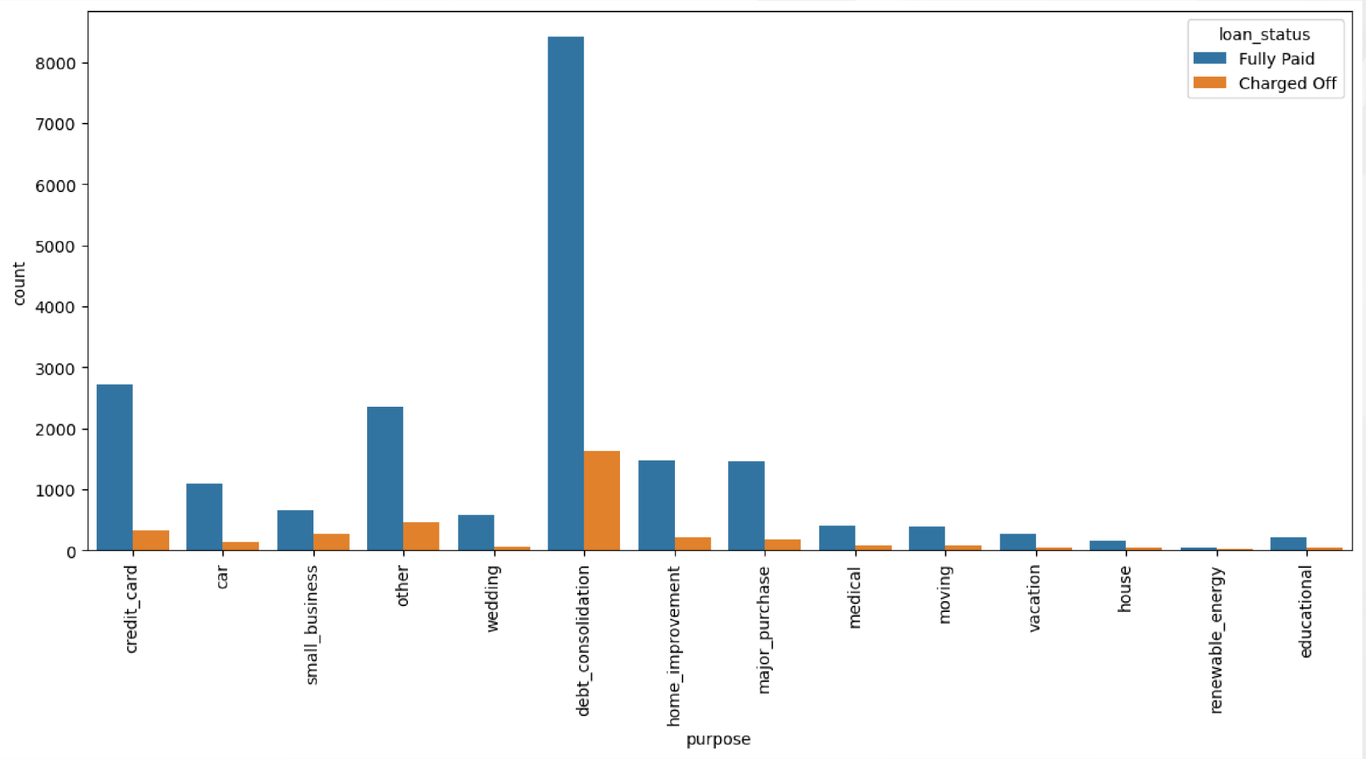
EXPLORATORY DATA ANALYSIS

Analysing "purpose":

"purpose" description : A category provided by the borrower for the loan request.

Inference :

- Although no distinct seperation can be drawn , 'debt_consolidation' seems like the leading factor behind taking out a loan i.e. paying off existing loans by availing another loan .
- We can further draw insights on defaulters by looking at the number of loans they have availed prior to availing this loan .
- Another purpose is taking out a loan for 'credit_card' payments .



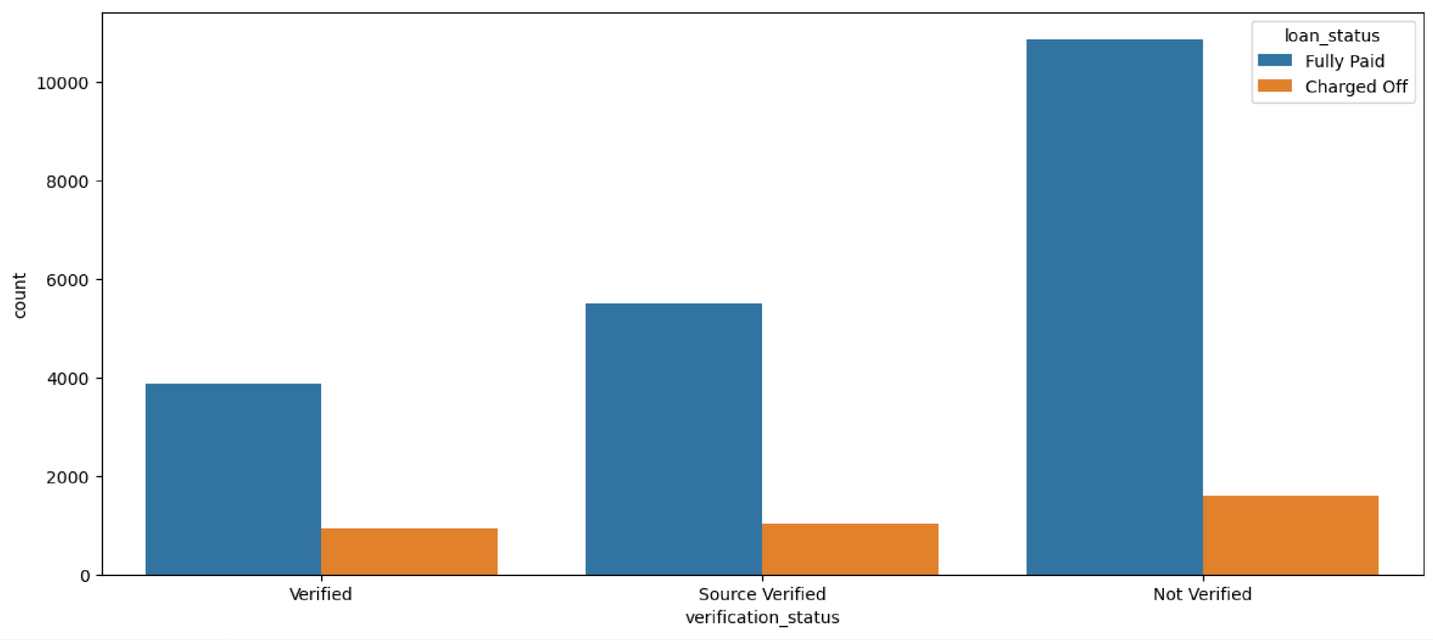
EXPLORATORY DATA ANALYSIS

Analysing 'verification_status':

Description : Indicates if income was verified by LC, not verified, or if the income source was verified

Inference :

- No distince seperation between defaulters w.r.t the applicant being vetted by the Bank or not .



EXPLORATORY DATA ANALYSIS

Analysing 'term':

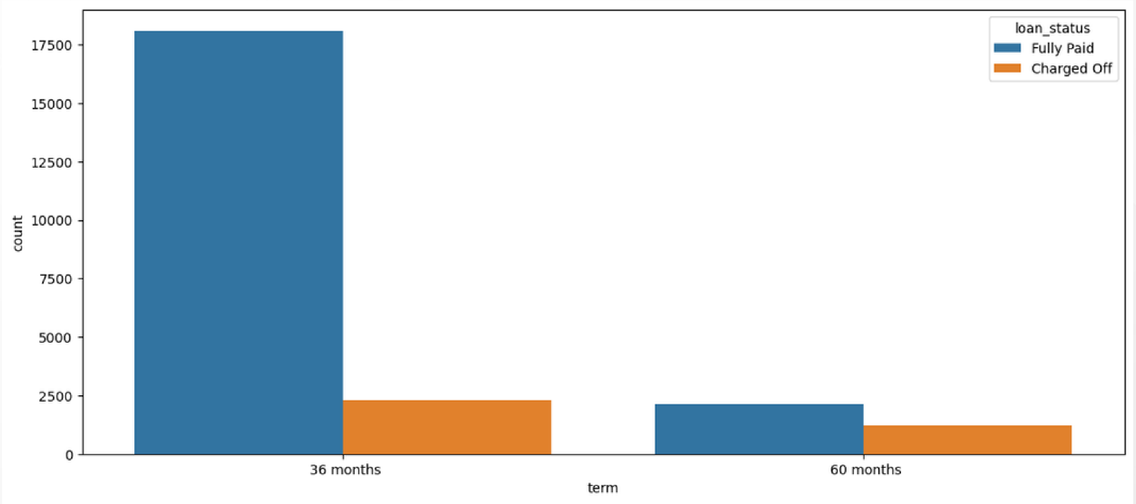
Description : The number of payments on the loan. Values are in months and can be either 36 or 60.

Inference:

People tend to default more on the '60 month' payment plan .

Possible reasons :

- Financial Instability : Over a longer period , individuals are more likely to encounter financial hardships such as job loss , medican emergencies etc . which can impact their ability to make constant payments .
- Interest Accumulation : Longer-term loans accrue more interest .
- Borrower overconifdence and Economic Changes



EXPLORATORY DATA ANALYSIS

Processing the 'issue_d' feature:

For drawing analysis based on month

Description : "The month which the loan was funded"

Inference:

- Maximum loan transactions happened in the month of Novemeber and December , with the lowest being in January .
- Maximum defaulters where the ones who availed the loan in December and Novemeber , with the least defaulters being in the month of Feb , Jan .

Reasons being :

1. Holiday spending
2. Year-end financial strain
3. Seasonal Unemployment
4. Bonus and Tax Refunds

Closing Statement for the Lending Club Case Study

In conclusion, our comprehensive analysis of the given dataset has provided valuable insights into the patterns and factors influencing loan performance.

Through various visualizations and statistical analyses, we have uncovered significant trends and correlations that can inform better decision-making for both lenders and borrowers.