Lending Club Case Study

Group Members: Pranjal Verma

Objective:

- This project aims to perform an Exploratory Data Analysis (EDA) on a given dataset of past loan applicants which have paid or defaulted on their loan.
- The EDA process involves cleaning and analysing the dataset,
- cleaning involves preparing the data for analysis. such as its distribution, missing values, outliers, and relationships between variables.
- The goal is to gain insights and understanding of the data using it to taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

Understanding the dataset

- The original dataset contained about 111 columns and over 40k rows
- After perusing through the data I found that alot of the columns had 0 values
- Some of rows were null and alot of columns were having null values
- After dropping the invalid data that would not contribute to the analysis. We got left with 37k rows with 32 columns.

Understanding the dataset

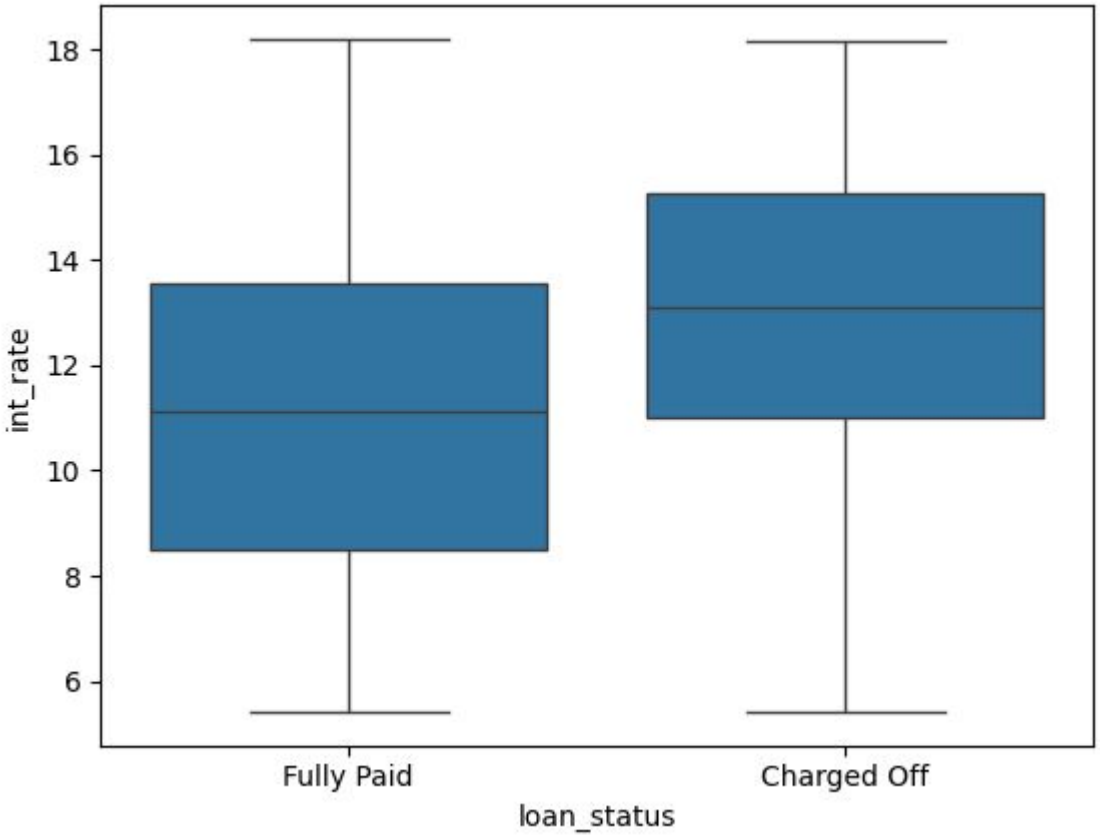
- Columns like id, member_id, url are from transactional database and does not add any doesn't any value or contribute to risk assessment
- Columns emp_title, issue_d
- There seems to be a redundancy between, funded_amnt and funded_amnt_inv. Since we are only interested in the funds the were provided by us, we can drop the former.
- Columns like initial_list_status, collection_recovery_fee, next_pymnt_d carry less significance

Important for risk analysis:

- loan_amnt, int_rate, installment, grade, sub_grade: These are directly tied to the loan and risk assessment.
- emp_title, emp_length, annual_inc: Important for evaluating the borrower's financial stability.
- home_ownership, verification_status: Useful for understanding the borrower’s financial background.
- dti (Debt-to-Income Ratio), delinq_2yrs, inq_last_6mths, mths_since_last_delinq, mths_since_last_record: Important metrics for understanding the borrower's creditworthiness and history of delinquencies.
- open_acc, total_acc, pub_rec, pub_rec_bankruptcies, revol_bal, revol_util: All relate to credit history and available credit, crucial for risk.
- recoveries, total_pymnt, out_prncp: Payment-related columns that reflect on the borrower’s payment behavior.
- earliest_cr_line, last_credit_pull_d: Useful for assessing the length of credit history.

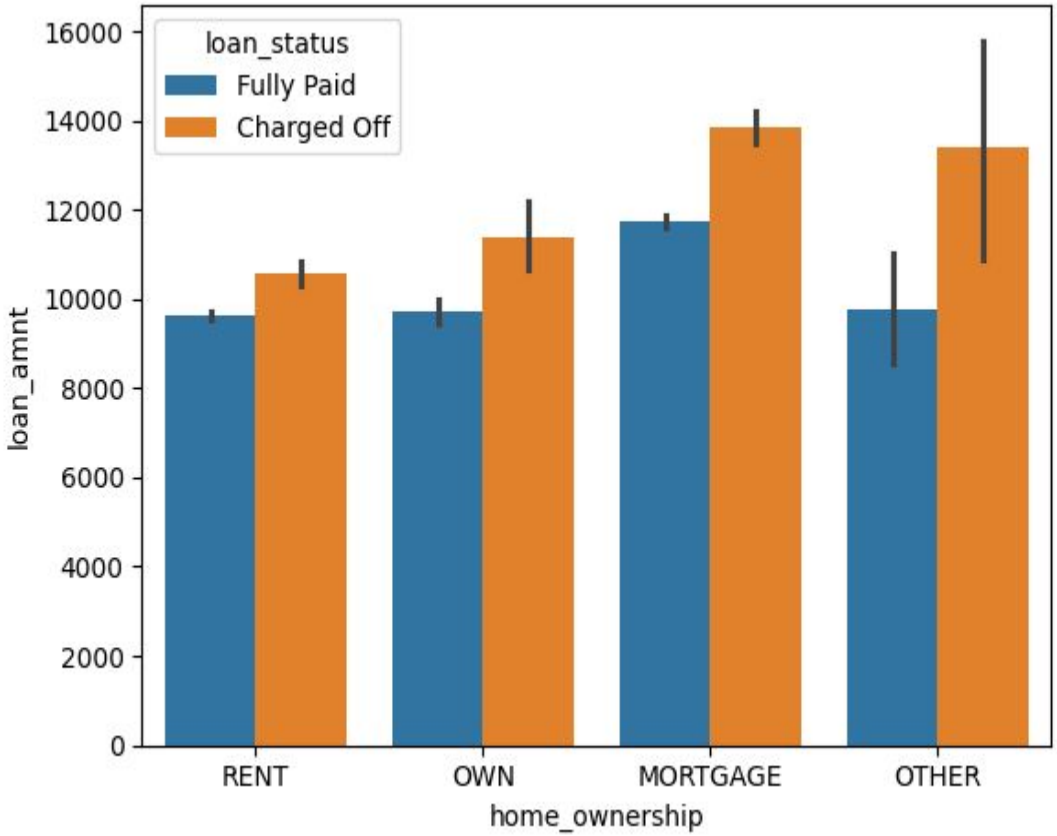
Univariate Analysis

1. Interest Rate: Interests rate have a median value of 11.86 with max going up to 24. When grouped by the loan status, it appears the people who defaulted on their loans had median of 13.610000 while who were to able to pay it off had an interest rate of 11.4.



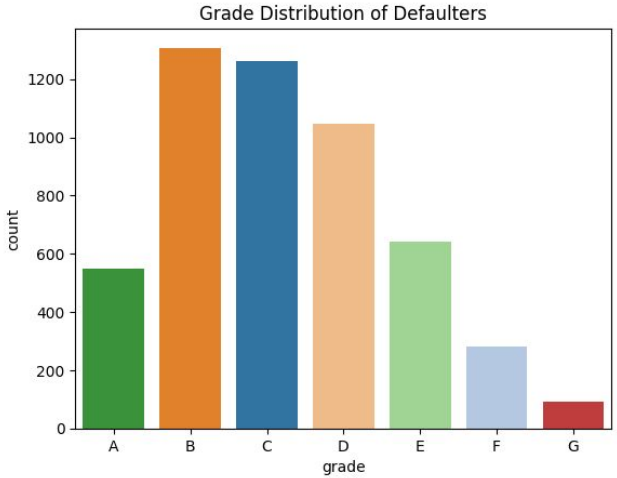
2. Home Ownership:

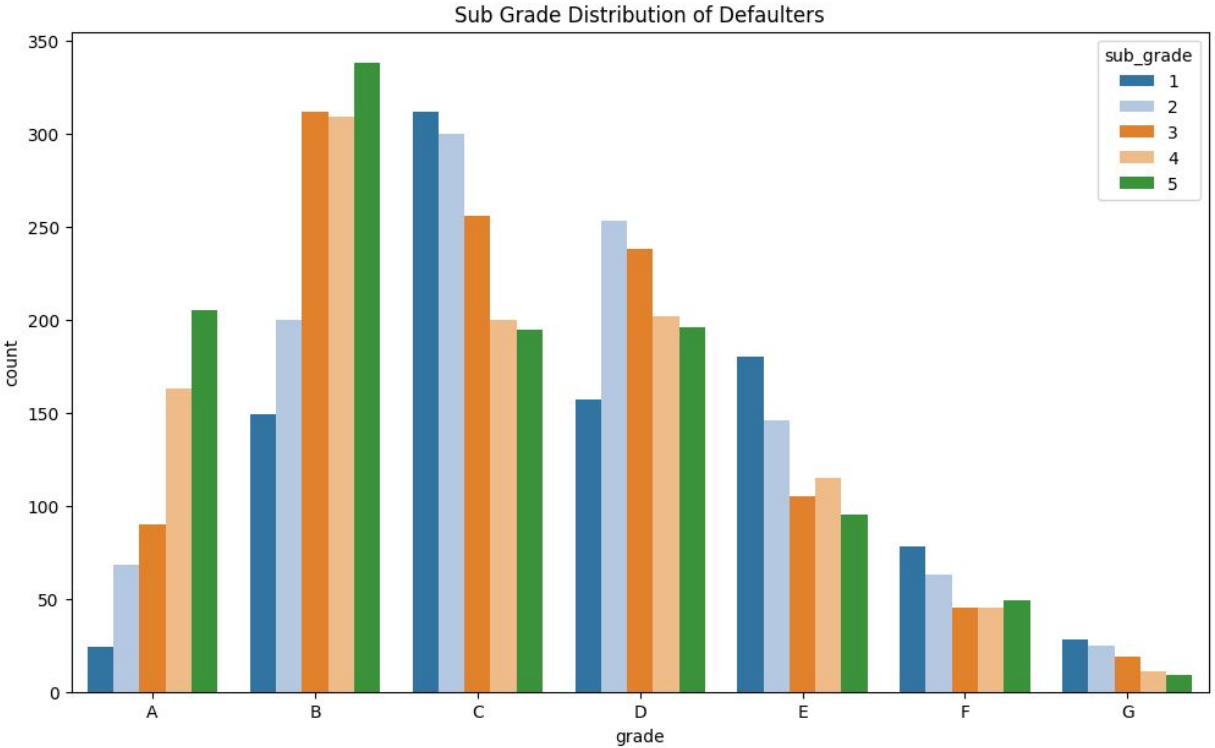
From the graph,Borrowers who have have EMIs on their home are most likely to default on their loans.



3. Grade and Subgrade

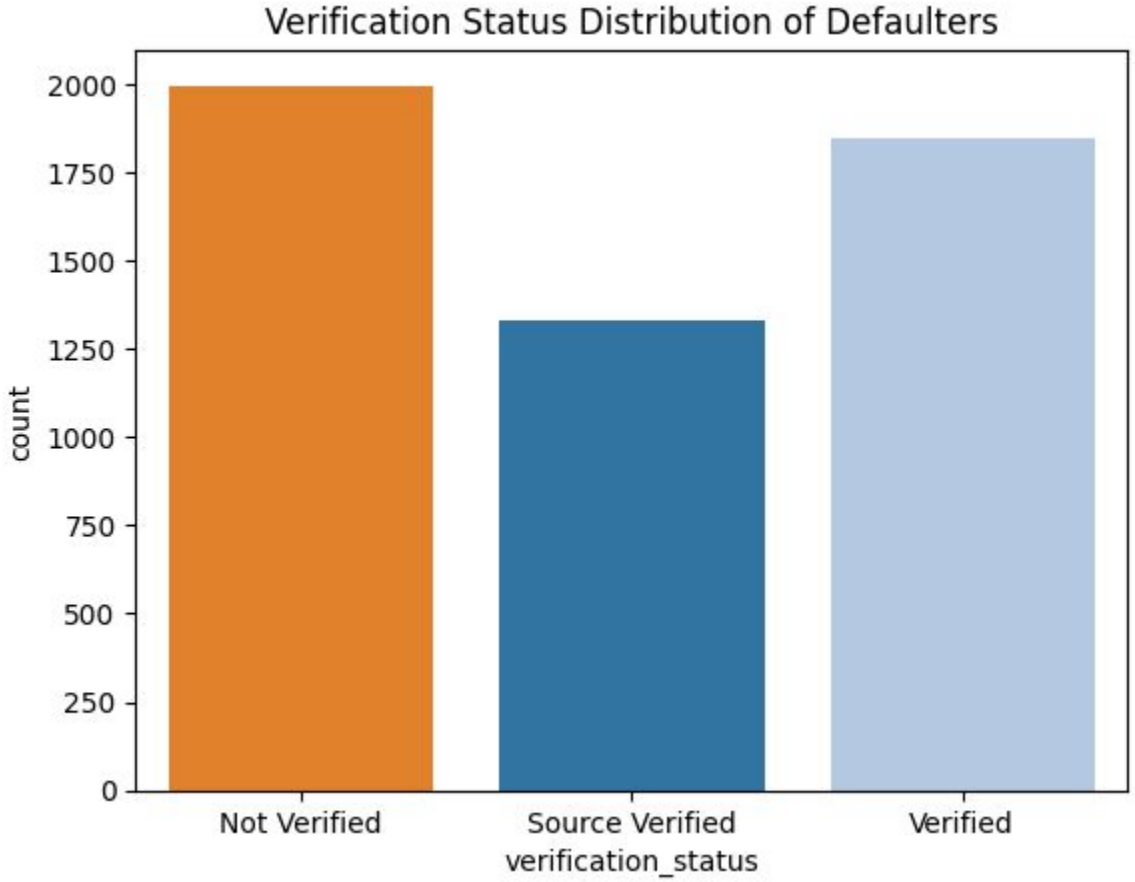
Borrowers in grades B, C, and D, particularly in sub-grades 4 and 5, appear to be at higher risk of defaulting.





4.Verification Status

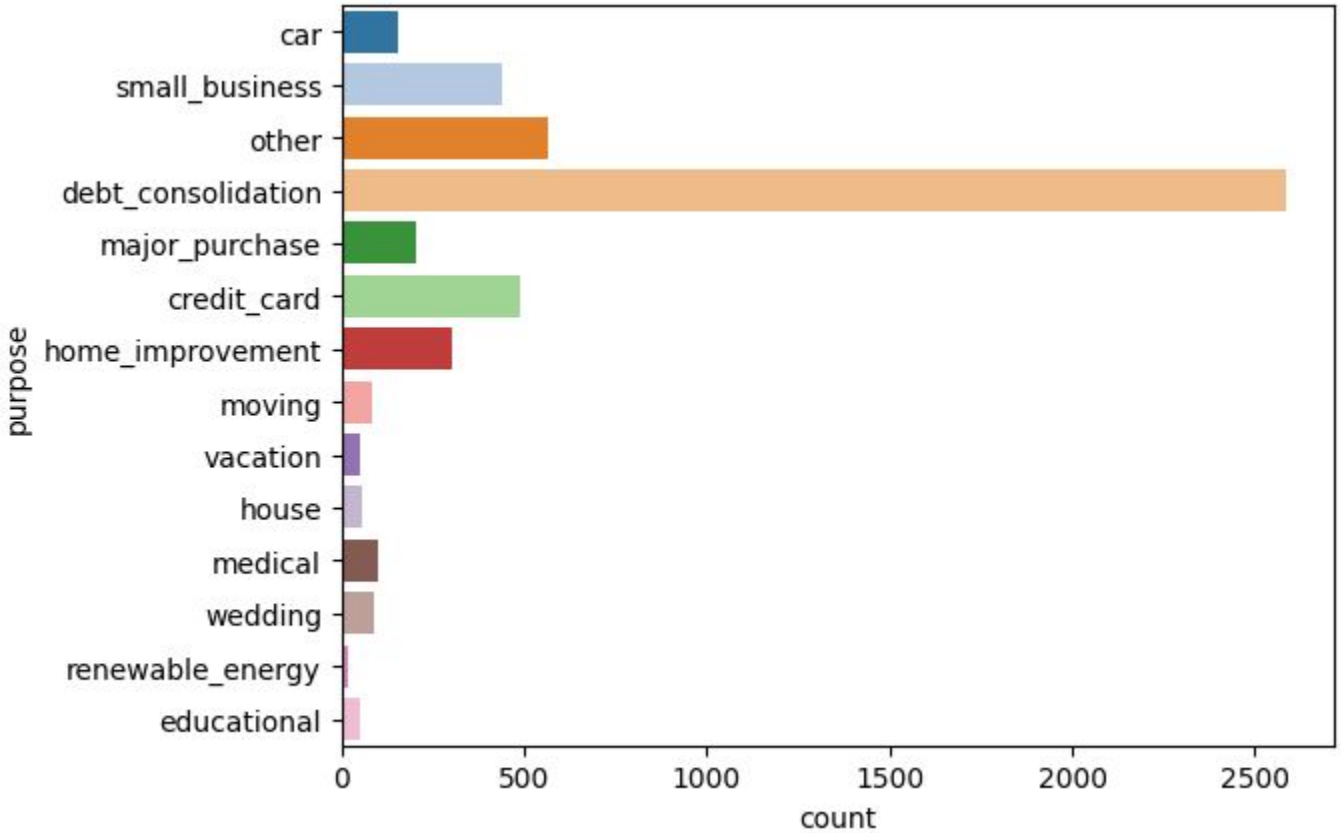
Unverified loans are most likely to default however even verified loans are closely behind.



5. Loan Purpose

The most common loan that defaults is taken for debt consolidation, with significantly more defaulters compared to any other loan purpose.

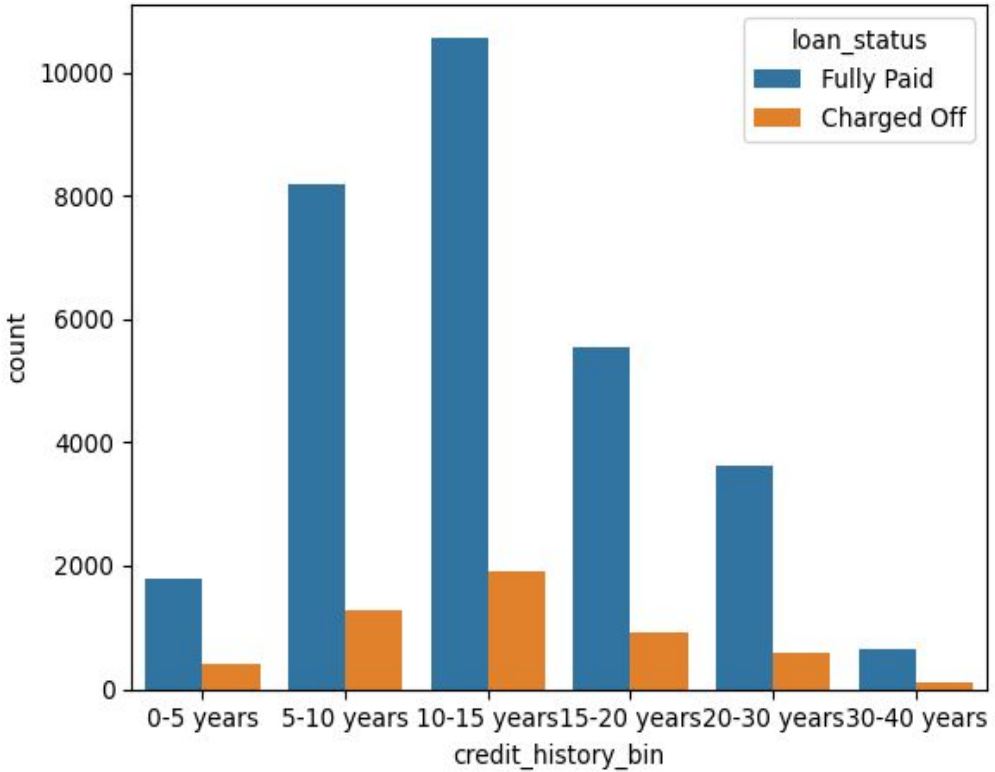
This suggests that borrowers consolidating debt may already be financially stressed



6. Credit History:

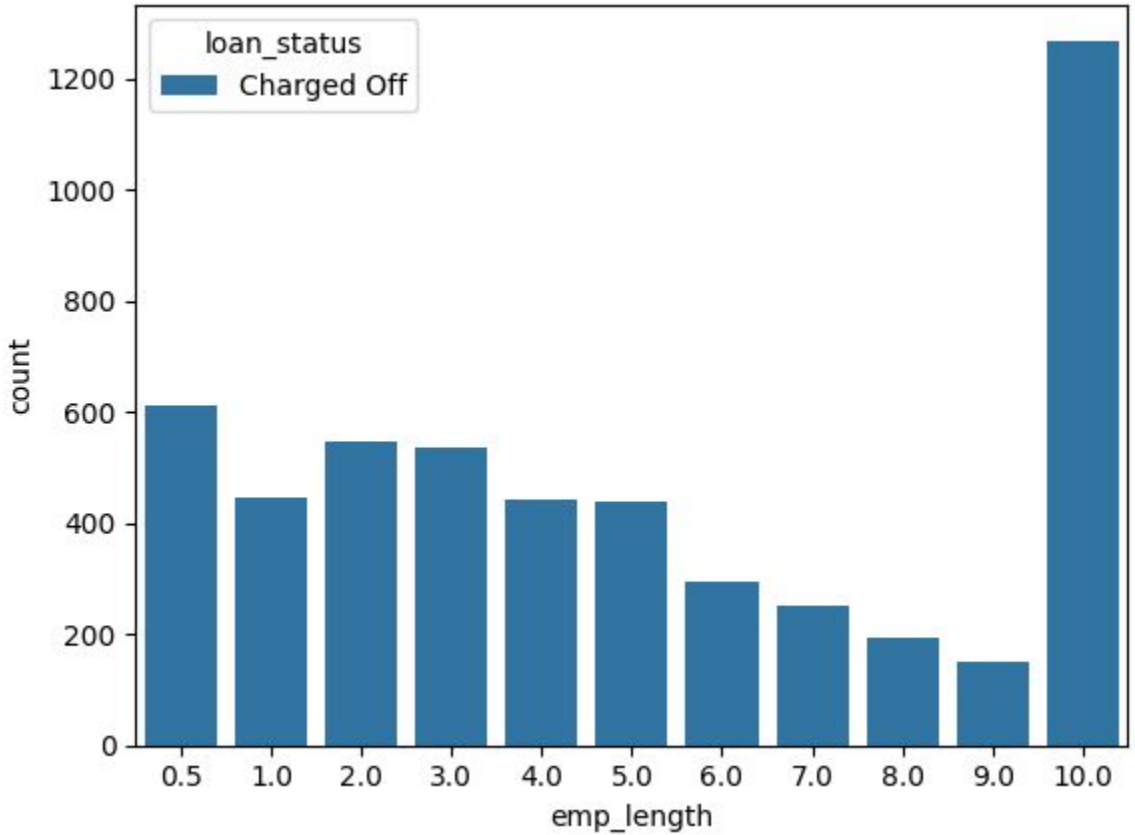
The longer the credit history, the more experience the borrower likely has managing debt. This is a Derived Metric and can be calculated by subtracting earliest_cr_line from loan issue date

Most borrowers fall into the 10-15 and 5-10 years of credit history range. Borrowers with less experience handling credit i.e 0-15 years are more likely to default than seasoned borrowoeres i.e 30-40 years of experience.

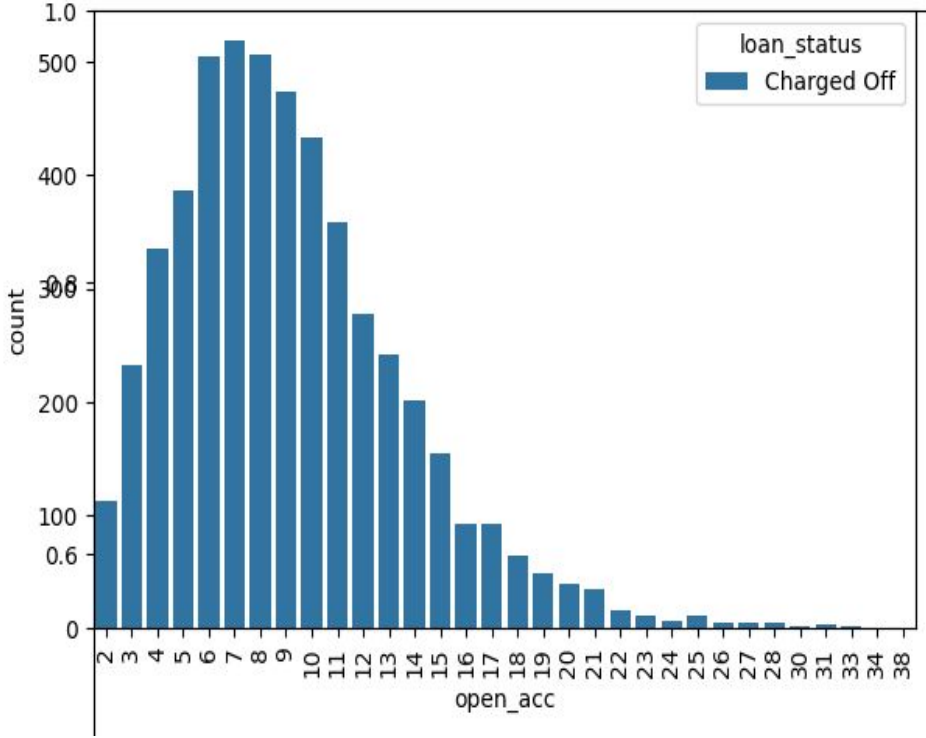


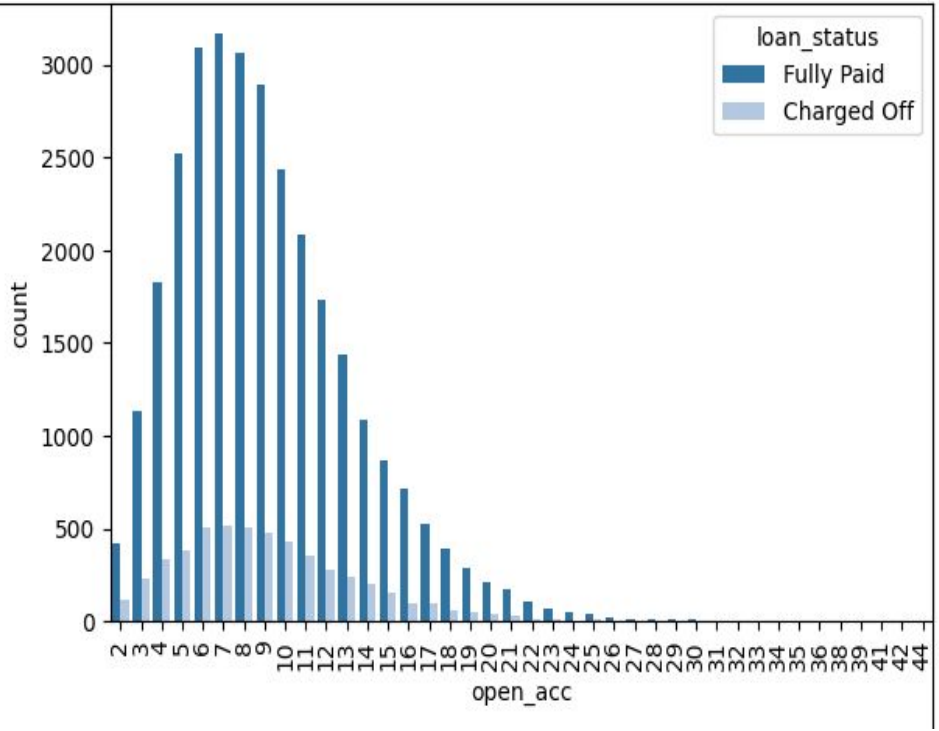
7. Employment Length

This looks like an anomaly, borrowers with <1 YOE and 10+ YOE are most likely to default on their loans. Perhaps pairing this column with other columns such as DTI will give us move information



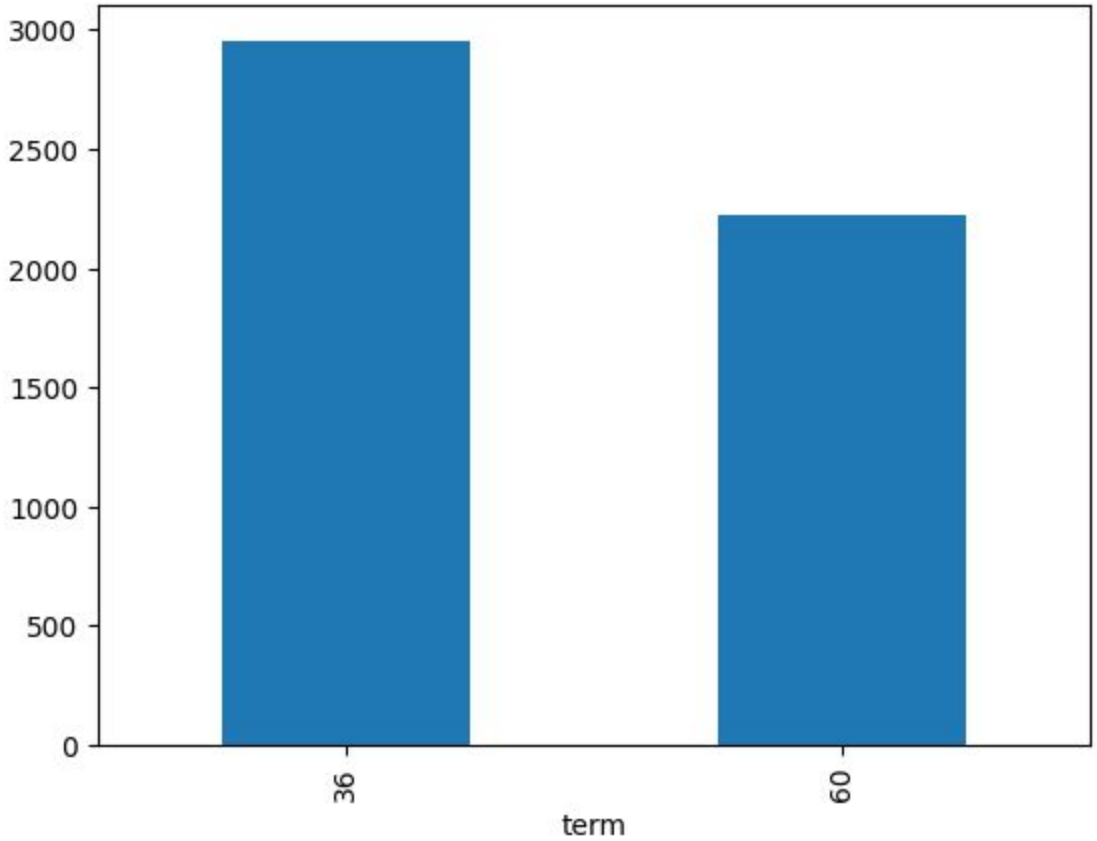
8,9 Open lines of credit and public derogatory accounts





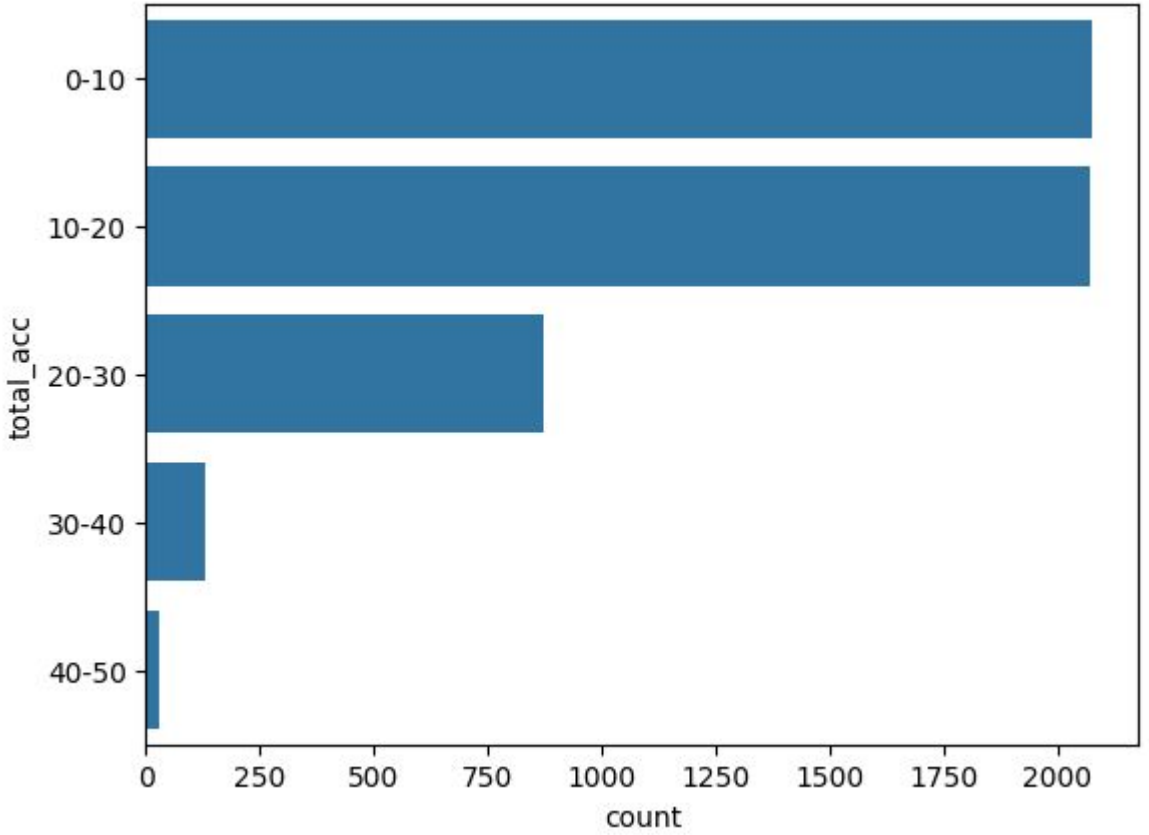
10. Loan Term

Defaulters have had a shorter loan term



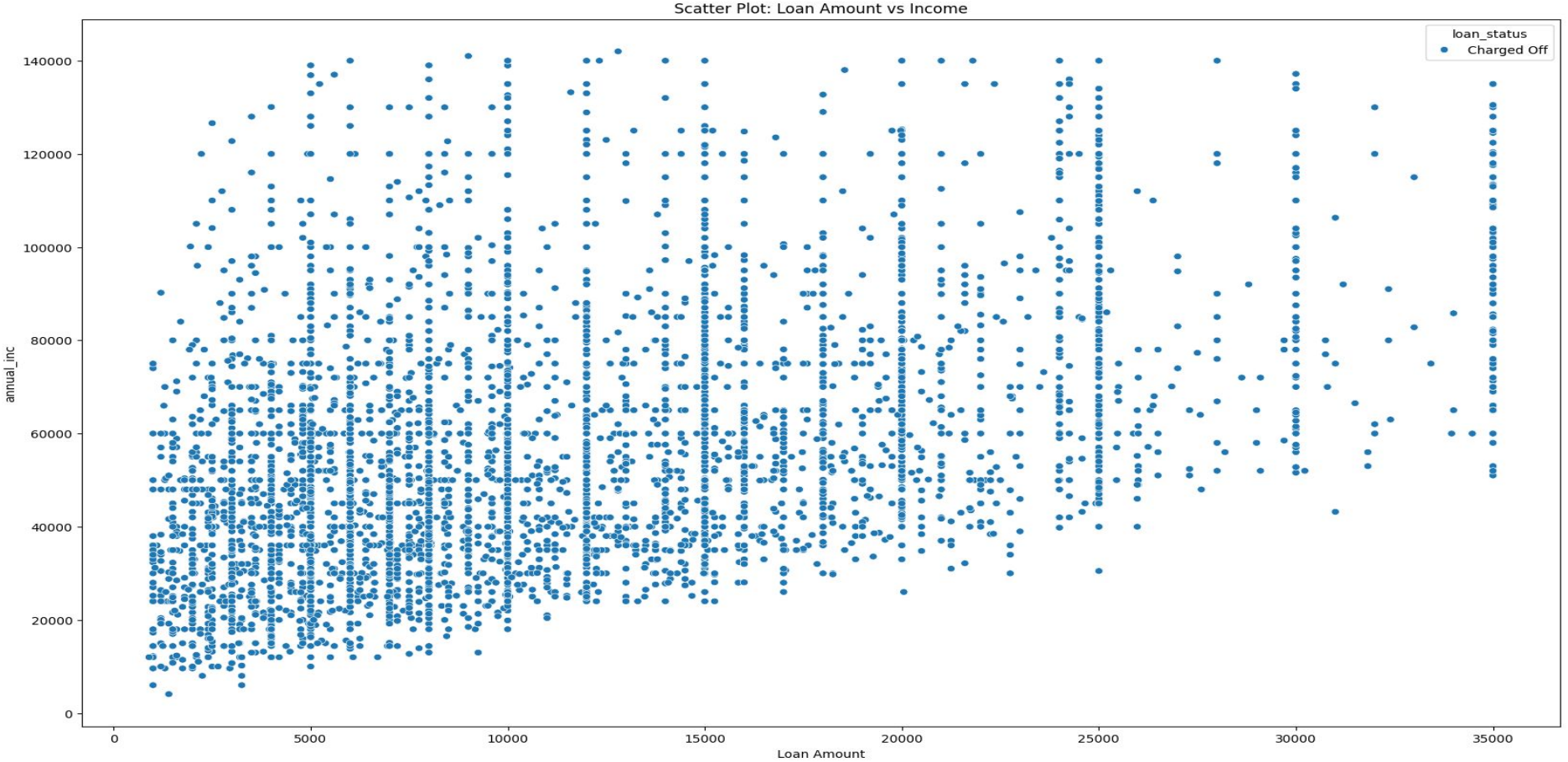
11. Total Accounts

Most borrowers in this dataset fall within 0-20 credit accounts, while higher numbers of accounts become incresingly rate. This could suggest that most loan applicants are not extremely active in terms of credit utilization



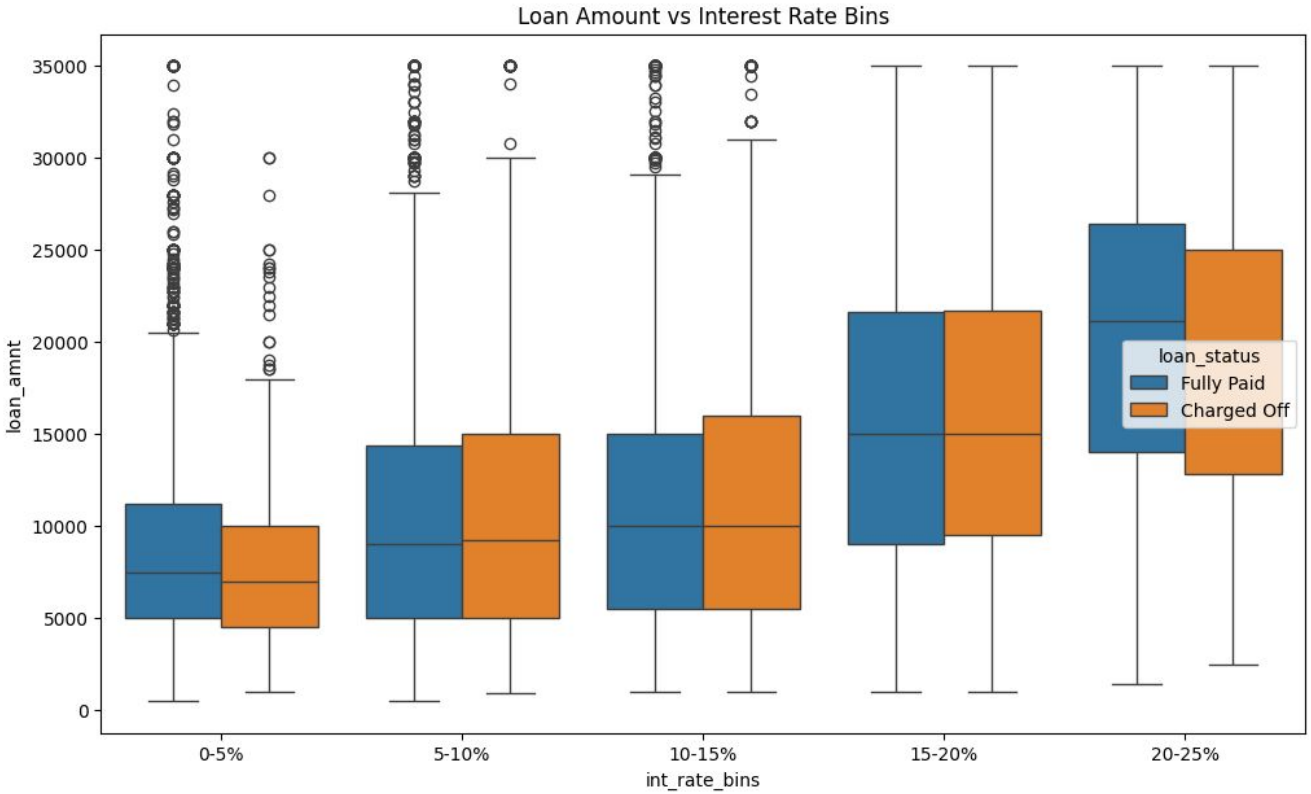
Bivariate Analysis

1.Loan Amount and income



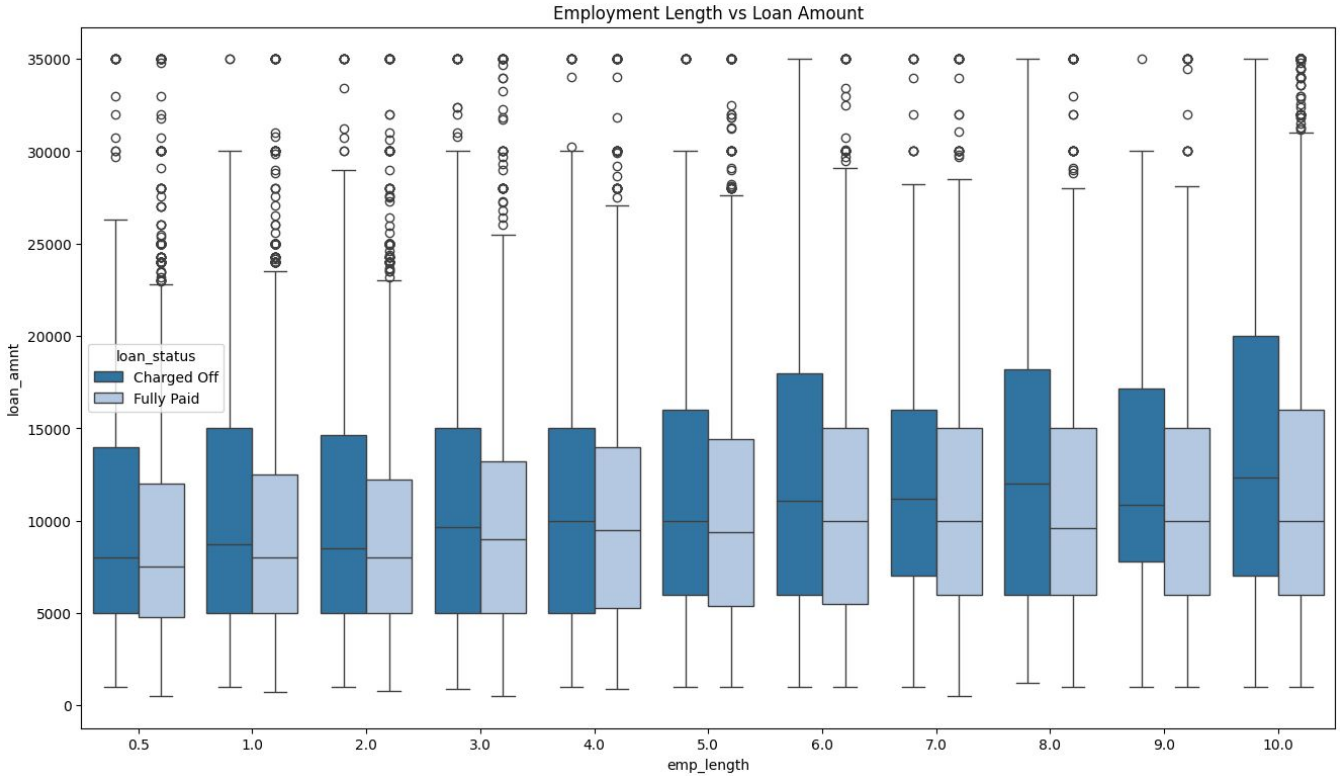
2.Interest Rate and Loan Amount

From the scatter plot we can see that alot of interest rate lies in the 15% and 10,000 income. From the heatmap,There isn't a strong co relation between loan amount and interest rate. No Strong Correlation



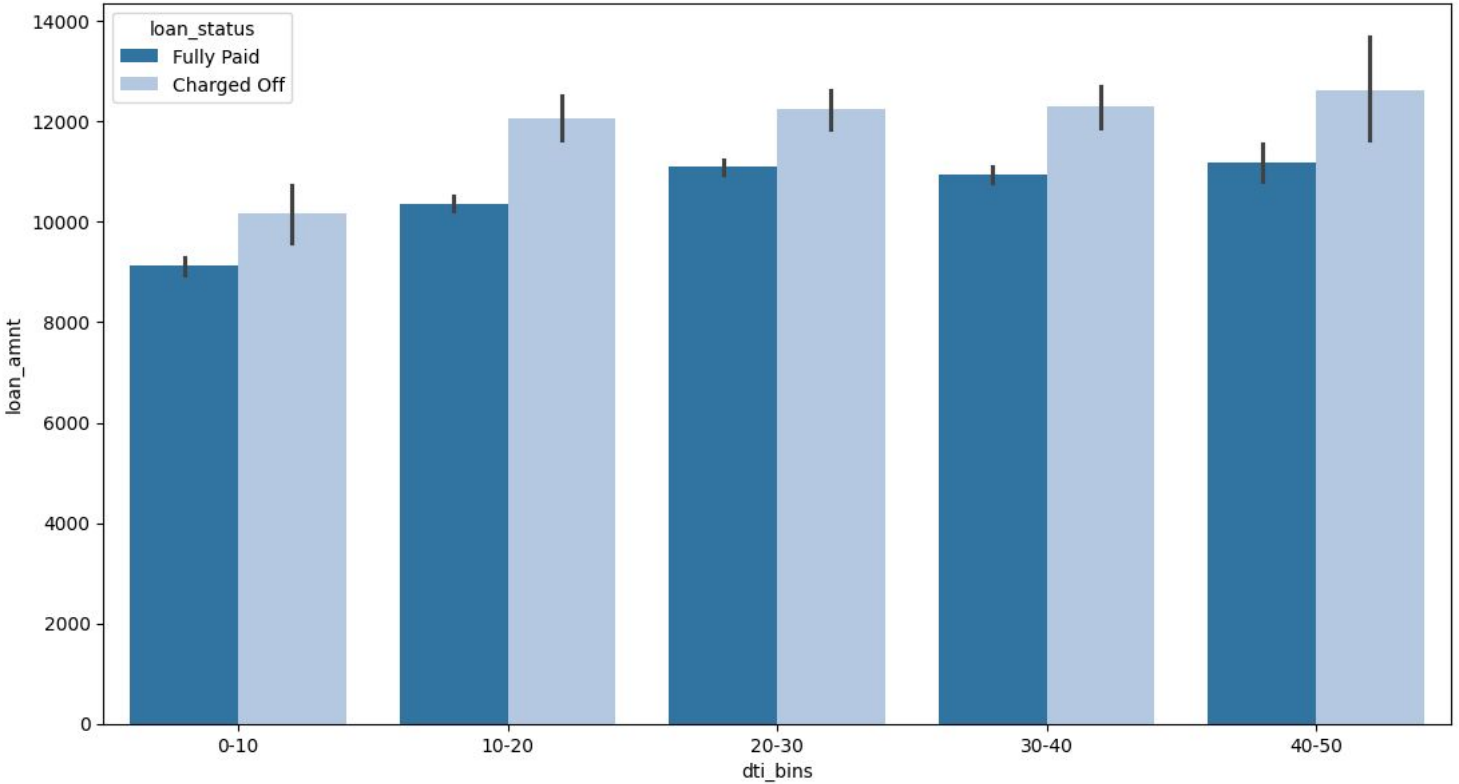
3. emp_length and loan amount

The consistently highest loans have been granted to people with 10+ YOE. Employment length doesn't seem to have a strong impact on whether a loan is Charged Off or Fully Paid. Borrowers with longer employment histories (10+ years) still default, indicating that job tenure alone is not a strong predictor of loan performance. This is a reflection on personal finances and financial literacy of peopls



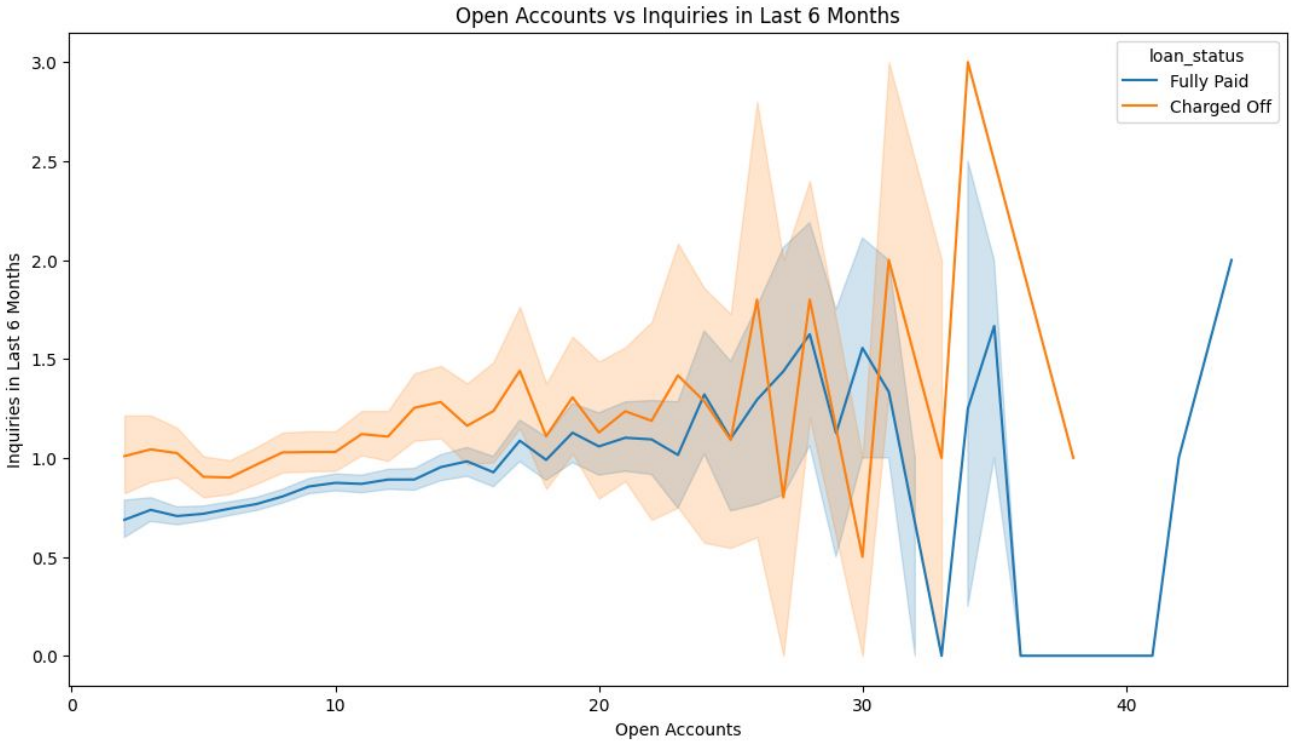
4. DTI vs Loan Amount

From this plot we can see that interest rate increases as grade decreases. The median interest rate of defaulted loans is slightly greater in high grade loans and it is significantly greater in lower grade loans. A low grade high interest loan is very likely to default.



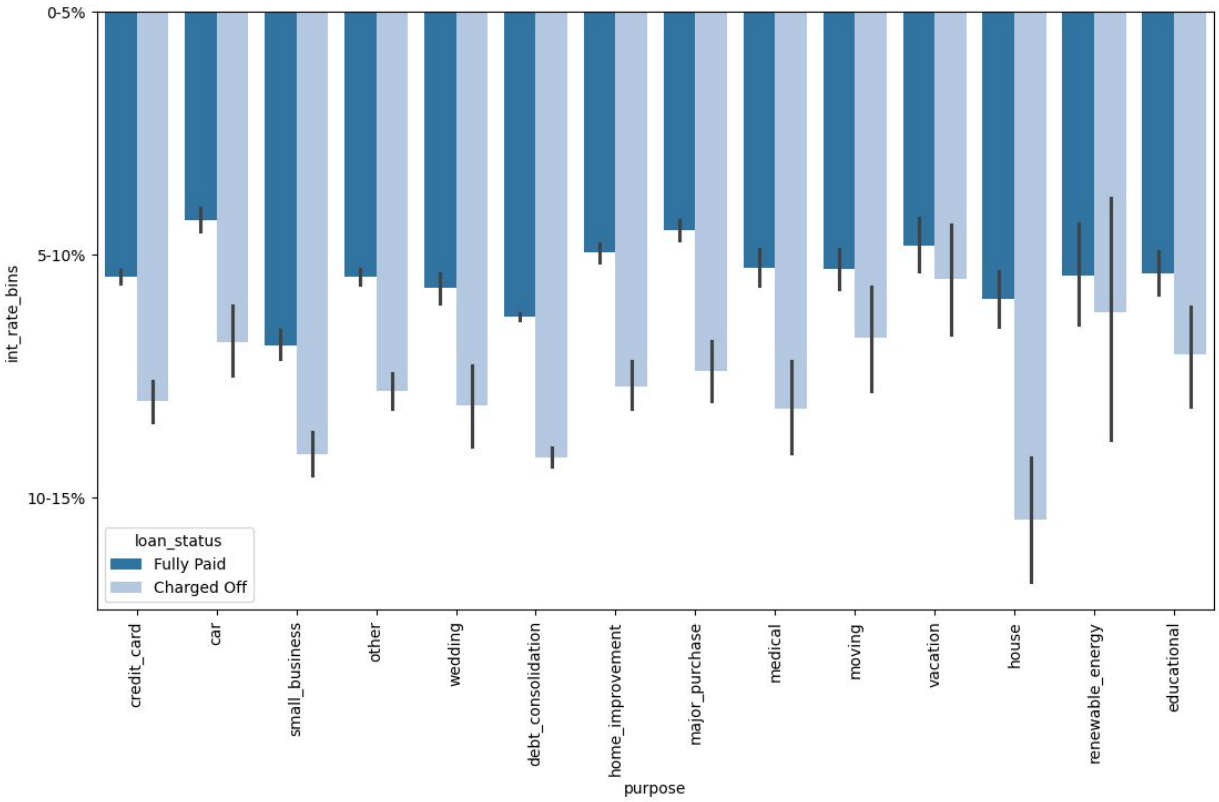
6. Open accounts vs Inquiries in last six months

This is an interesting insight. Borrowers who defaulted have consistently made more inquiries in last 6 months at the time of taking loan, the gap increases as the number of open accounts increases.



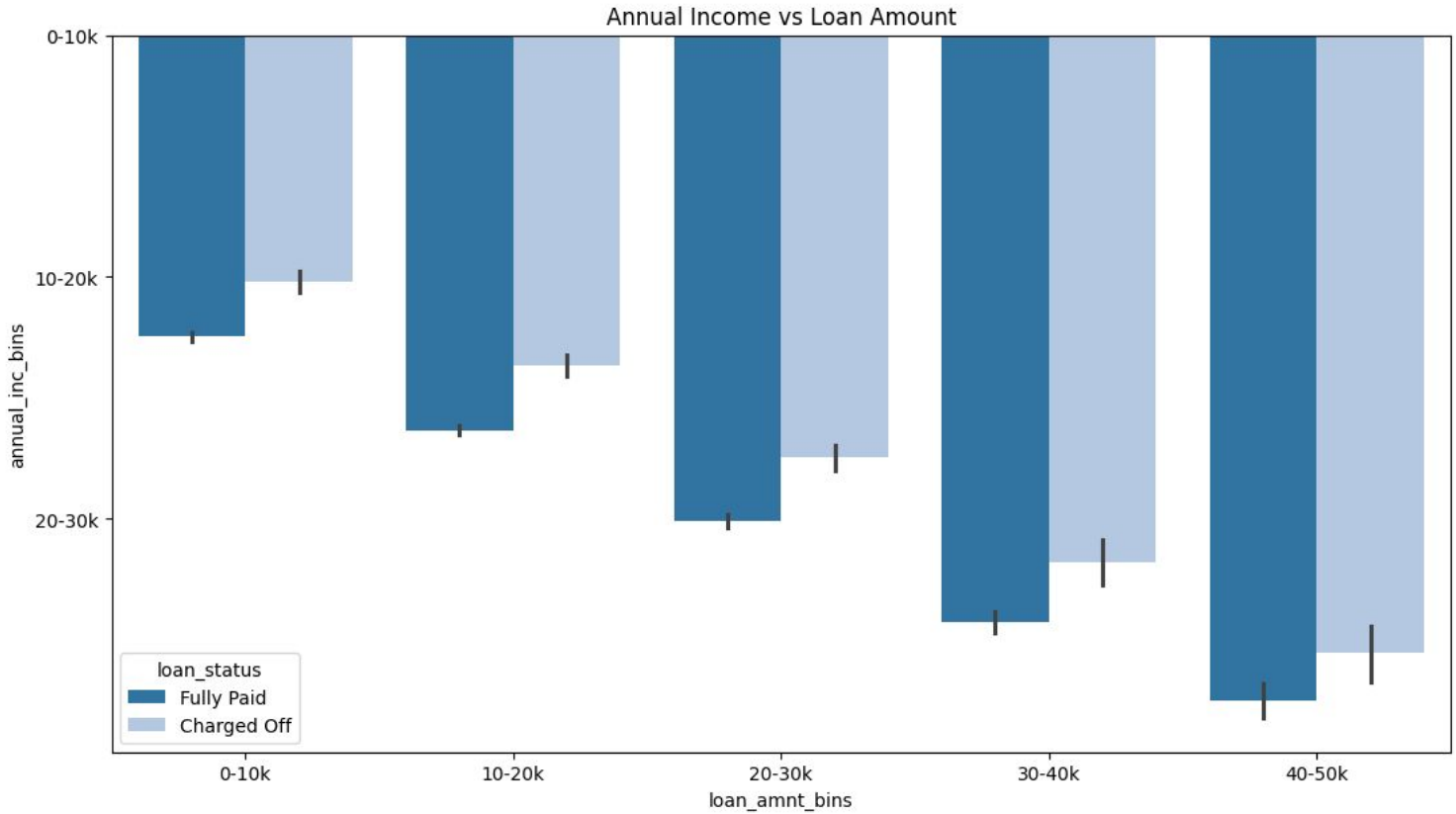
7. Interest rate and loan purpose

A home loan with high interest is most likely to default, right next to debt consolidation.



8.Loan amount vs income

Higher Loan Amounts Lead to Higher Charge-Off Rate



Conclusions

1. Higher loans lead to higher charge off rate

2. If loan amount is more than double then loan is more likely to default

3. A home loan with high interest rate (15%) is very likely to default, followed by debt consolidation

4. Borrowers who defaulted have consistently made more inquiries in last 6 months at the time of taking loan, the gap increases as the number of open accounts increases.

5. Likelihood to default increases with interest rate and lowering loan grade. A loan with G grade and high interest is very likely to default

6. A high DTI is directly related to defaulting. People who have paid of loans have had low DTI compared to thsoe who defaulted.

7. Employment lenght does not mean loan security. People who 10+ YOE have still defaulted

8. Likelihood of defaulting increases with loan amount and interest. High loans with high interest rates are most likely to default. The pattern is linear