Lending Club Case Study

Group Facilitator: Anupam Maiti

Team Member: Arpita Datta

Problem Statement

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The shared data 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.

In this case study, EDA is used to understand how consumer attributes and loan attributes influence the tendency of default.

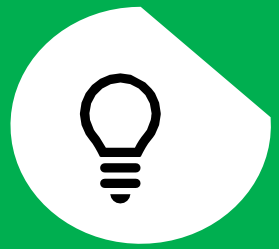
Abstract

The analysis is broadly divided into three steps

- Exploratory analysis to understand the data set, clean the data set, and identify key column for the analysis
- Perform univariate, bivariate, and multivariate analysis for target categorical and numerical variables
- Identifying the driver variables which can help in predicting high risk customers for future loan application.

Exploring Data Sets

The initial phase of the analysis process involved the preparation of the dataset to facilitate predictive analysis.

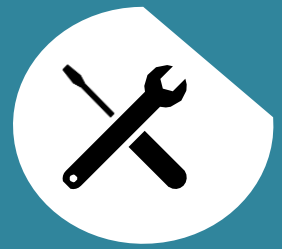


Understand the

- ***data dimensions***
- ***Column types***
- ***Identify the columns with missing value and outliers***

This step was focus on

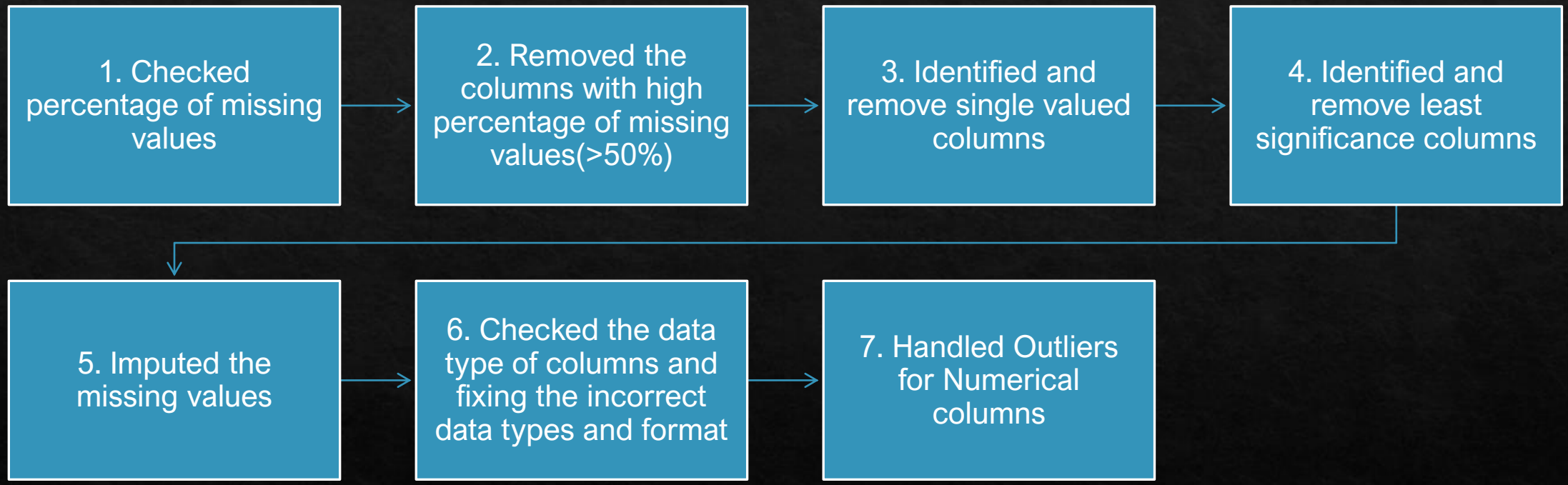
- ***Dropping columns with missing value***
- ***Develop approach to treat outliers***
- ***Fixing the column with incorrect data types***



In this step, objective was

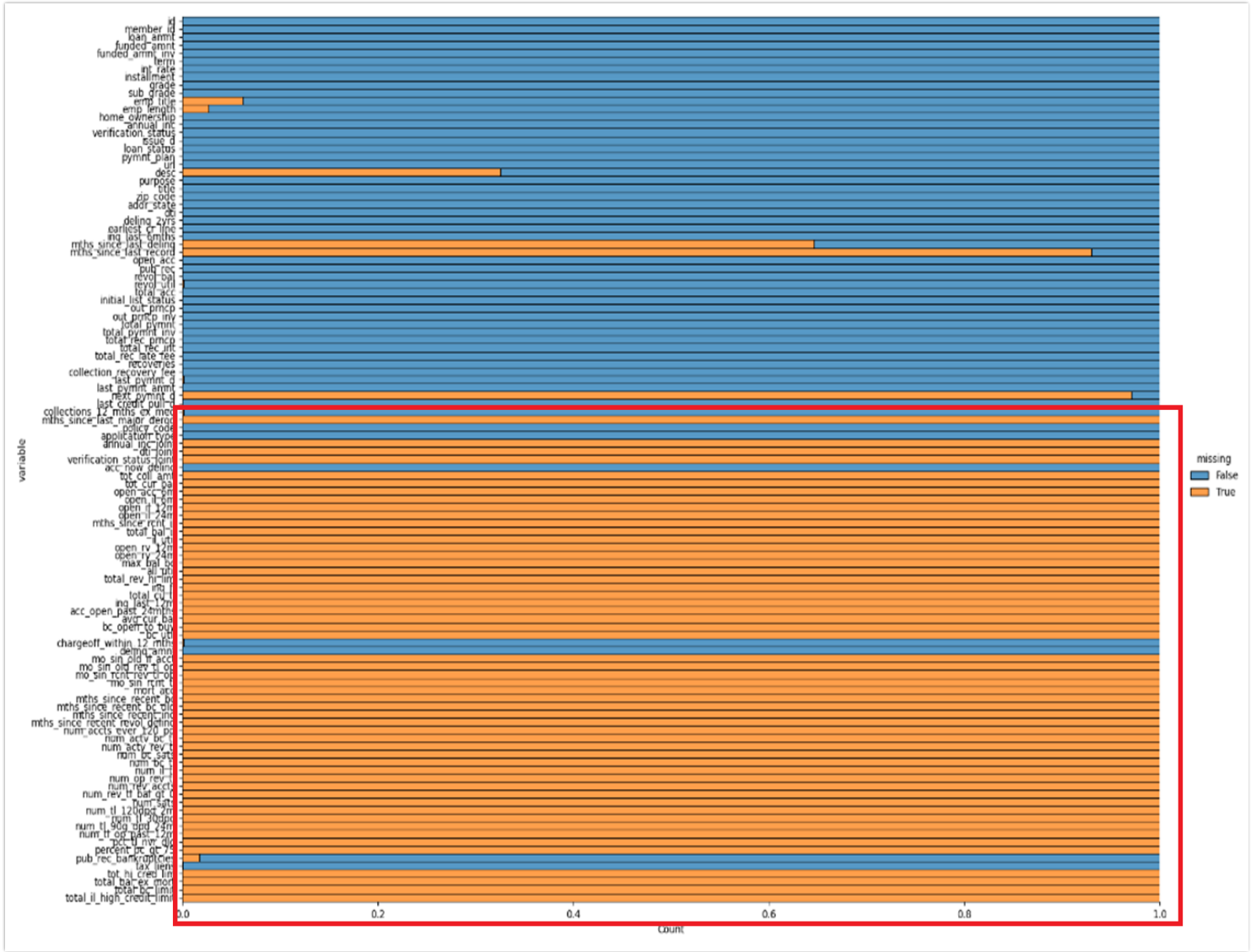
- ***Creating required columns(binning)***
- ***Dropping irrelevant columns***
- ***Creating required data frame***

Data Understanding, Cleaning and Manipulating



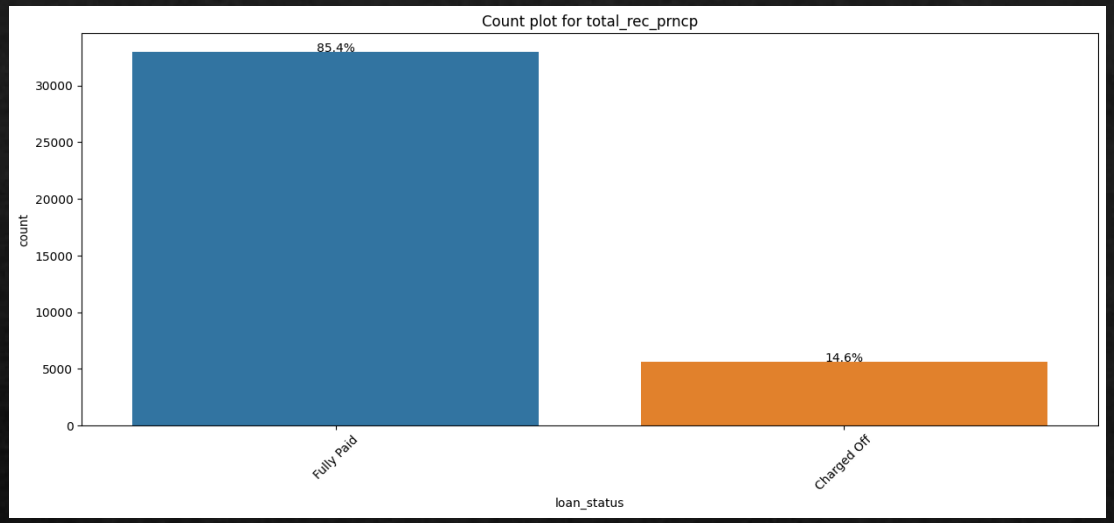
In the graph, we observed that some of the columns has all NULL value, and these columns won't help us in data analysis so discarding them.

Generally, if a column has more than 50% of missing value, we can drop those columns

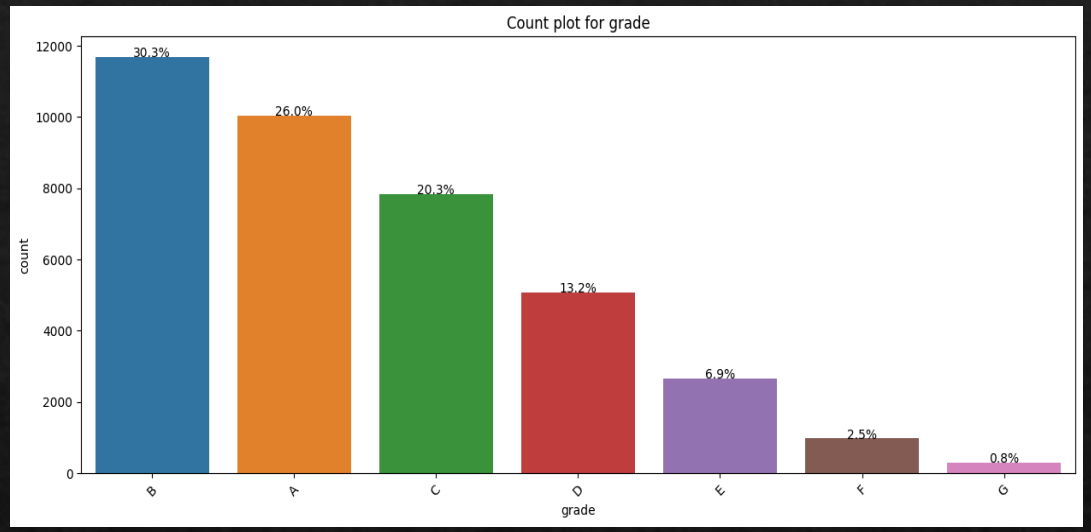


Exploring Data Sets - Univariate Analysis

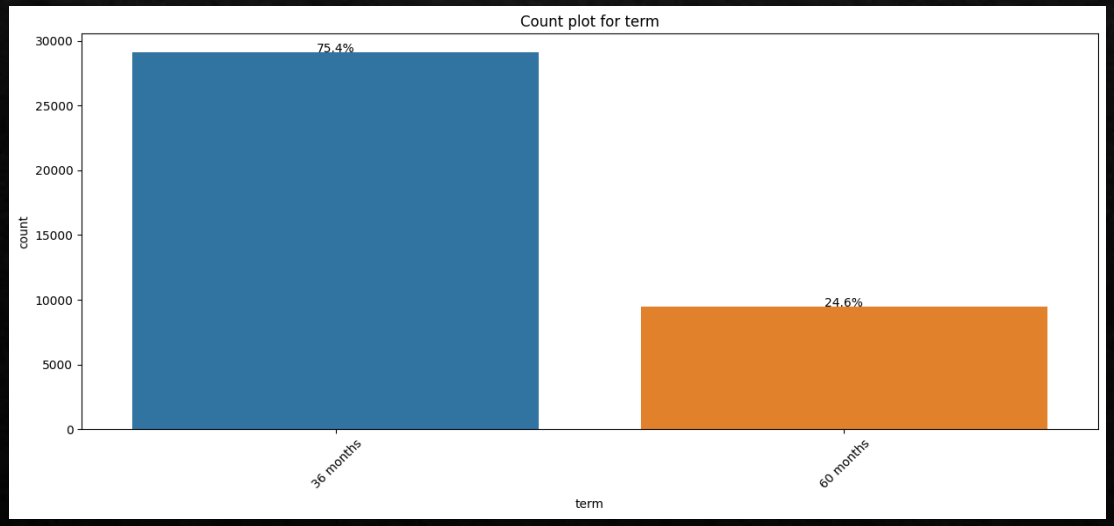
Loan Status: Only 15% of the borrowers are 'Defaulters'



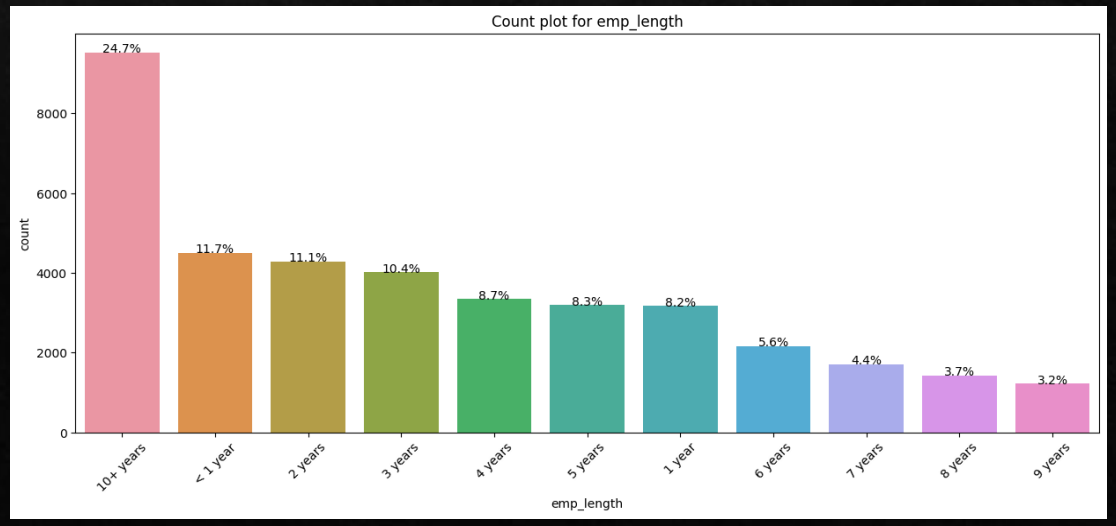
Grade: Most common loan grades are B(30%), followed by A(26%) and C(20%)



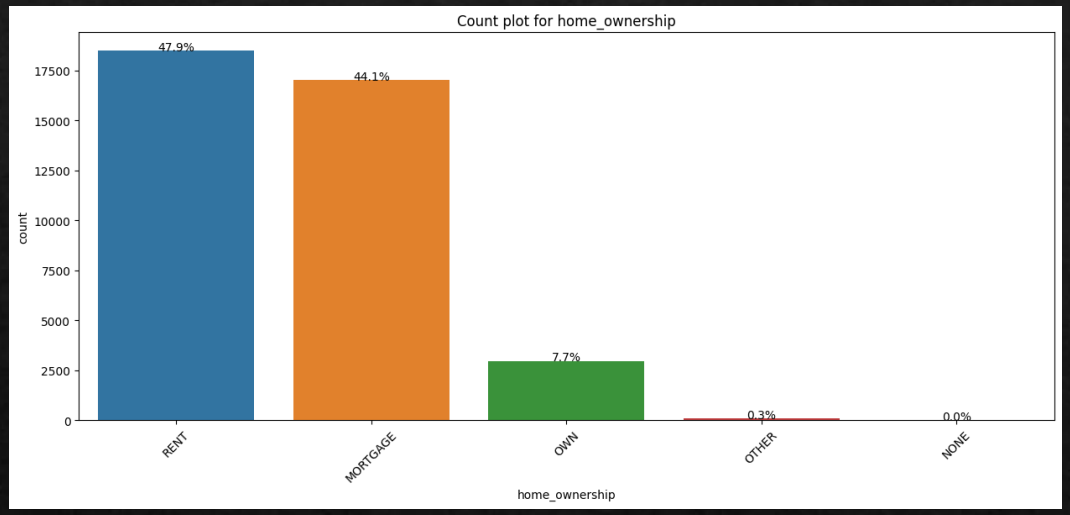
Term: Most of the loans (73%) have 36 months tenure whereas only 26% loans have 60 months tenure

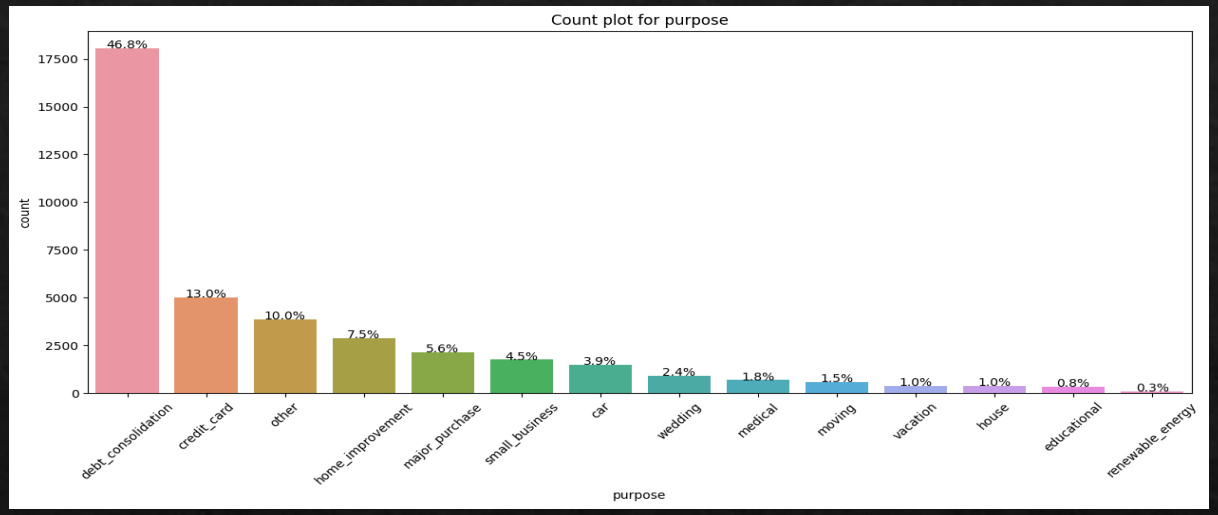


Employment Length: Most applicants have more than 10 years of employment, followed by less than 1 year

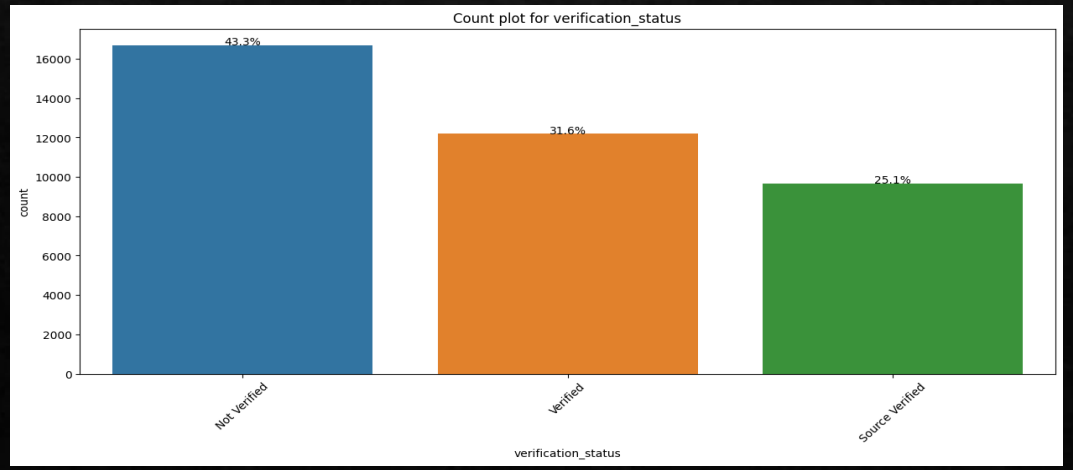


Home Ownership, purpose: Most of the borrowers don't have their own house. 90% borrowers either pay "Rent" or "Mortgage" their house. 47% borrowers' purpose is 'debt_consolidation'. But only 1% borrowers"purpose' of the loan is 'house'.

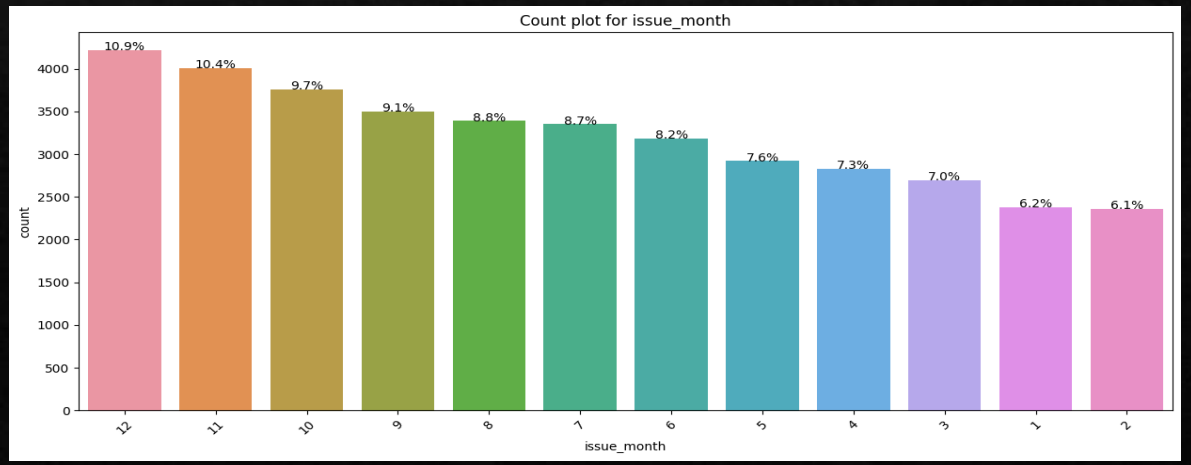




Verification Status: 44% of the borrowers' income is 'not verified' which is risky.



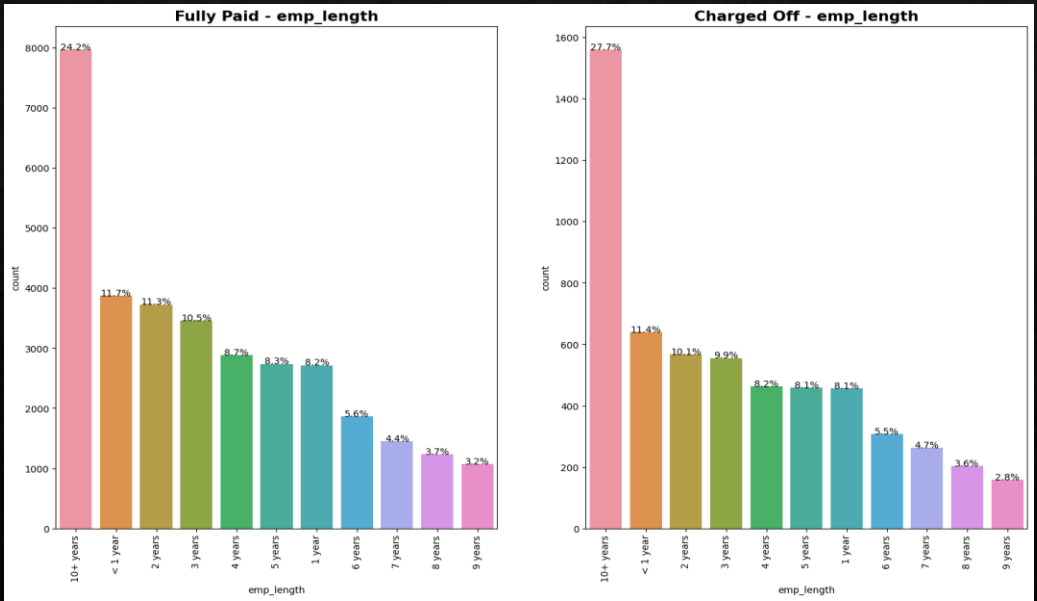
Issue Month: Most of the loan issued towards end of the year



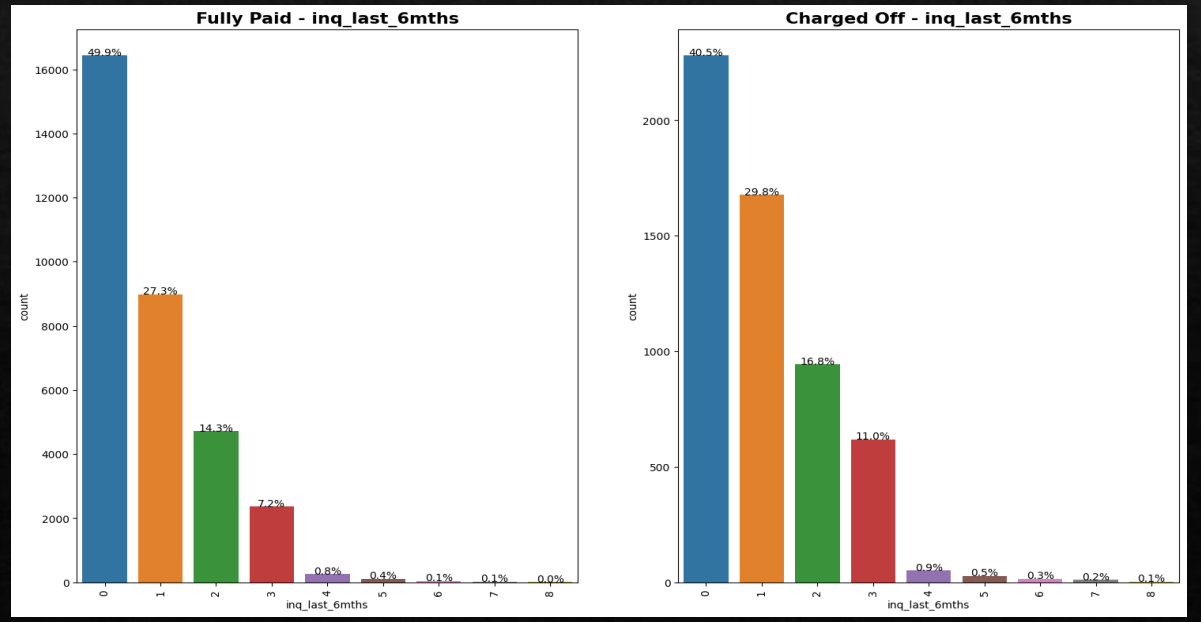
Exploring Data Sets - Segmented Univariate Analysis

We have conducted segmented univariate analysis for multiple categorical columns to explore how the distribution of loan statuses ("Fully Paid" and "Charged Off") varies within each category of the chosen variable. Here are key observations:

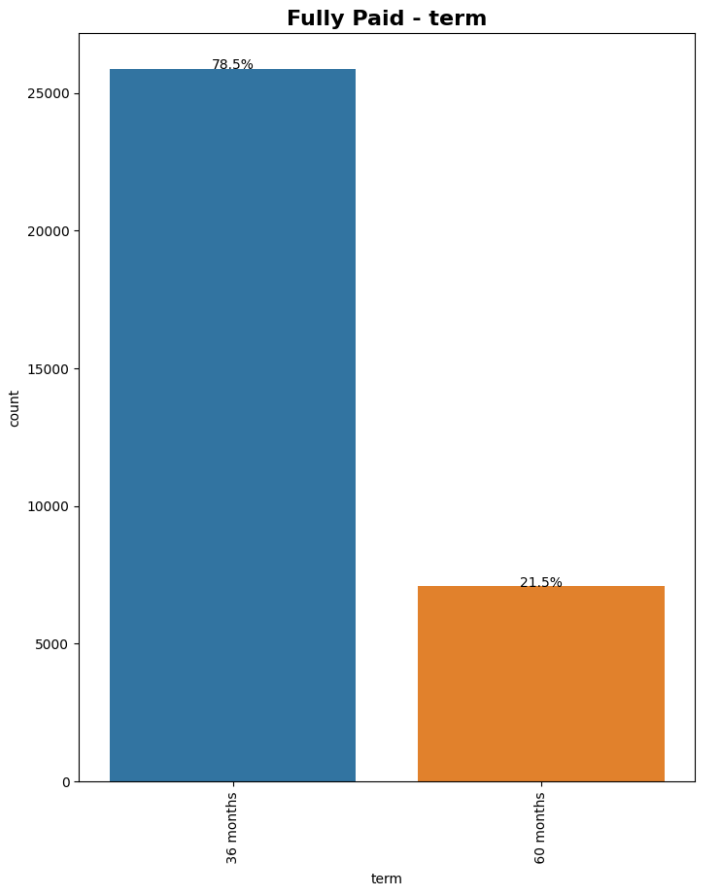
There is a 4% chance of loan getting default when employment length of the borrower is 10+ years.

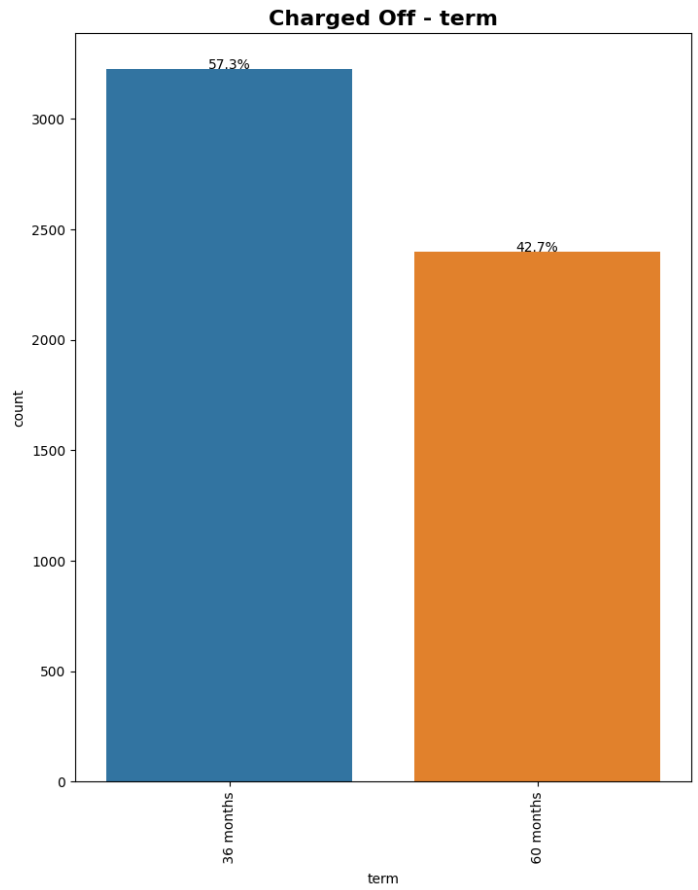


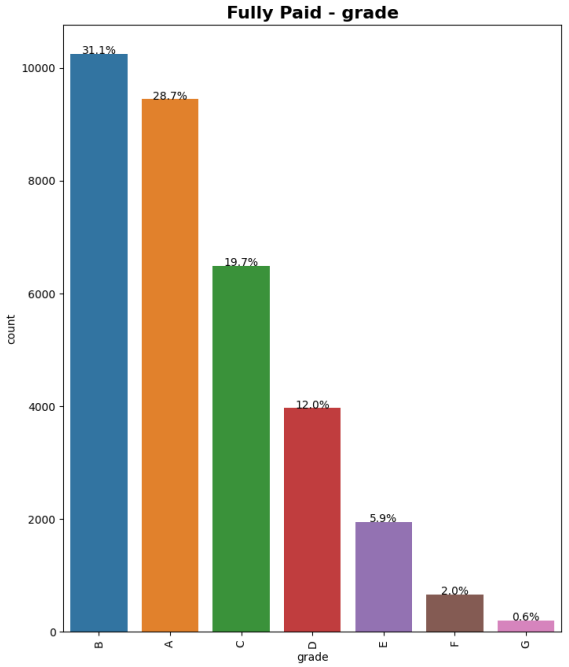
Higher number of inquiries is indication of high chances of defaulted loan.

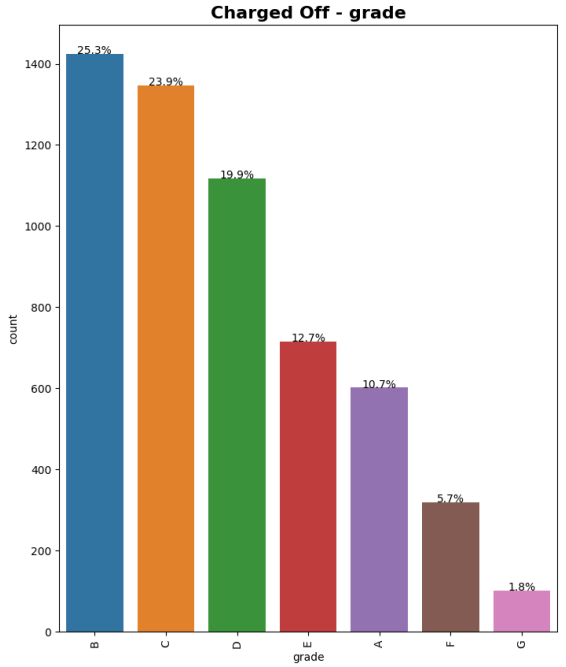


Loans with 60 months tenure have double chances (21%) of getting default than loans with 36 months tenure





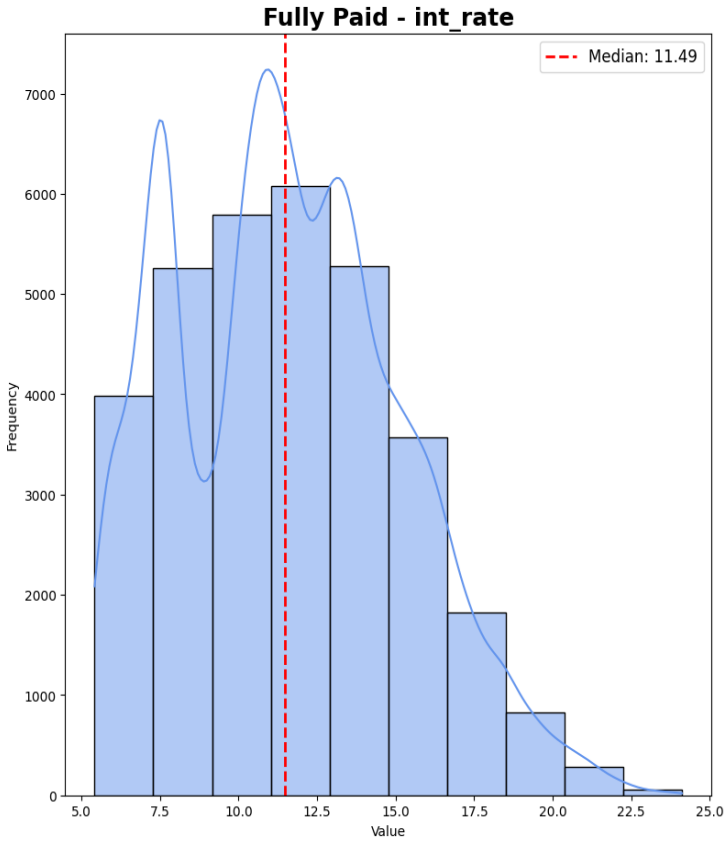


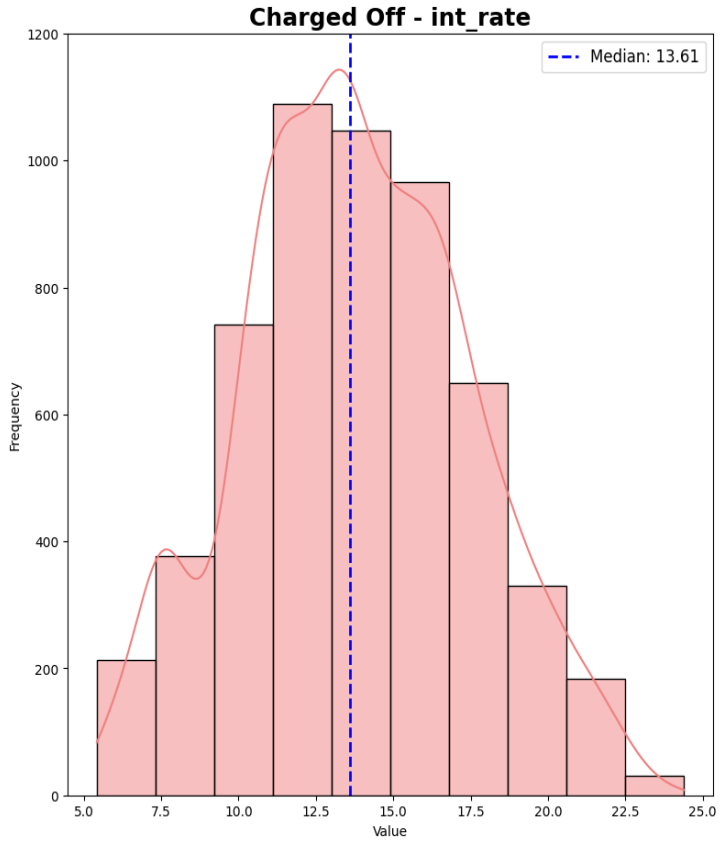


Grade C has 5% more and grade D has 8% more defaulter in comparison to non-defaulters.

But Grade G has 3 times more defaulter in comparison to non-defaulters

Defaulter are more when Loan has higher interest rate. On an average default loan interest rate is 13.50% (2% higher) whereas for non default loan interest rate is 11.50%.





The debt-to-income ratio is more for defaulter.

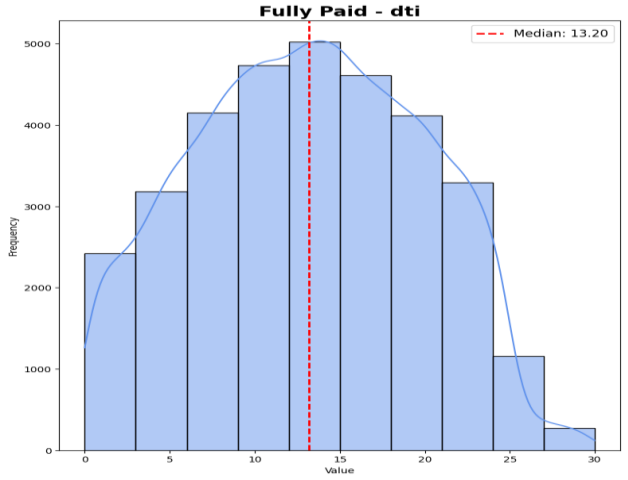
- ✓ Fully Paid - 13.43 (median)
- ✓ Charged Off - 14.40 (median)

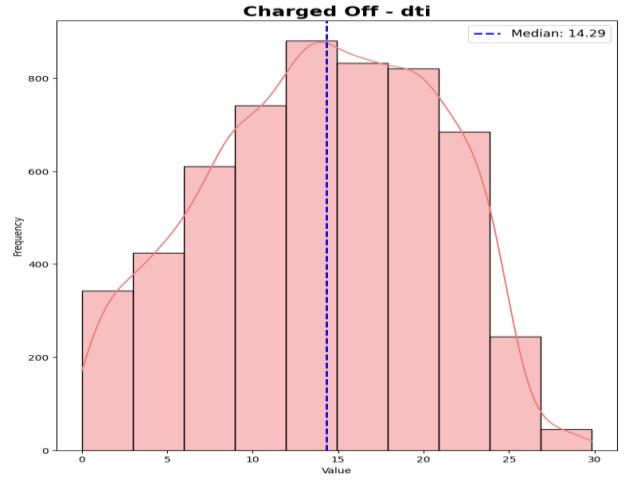
Higher revolving line utilization rates are linked to higher default rates.

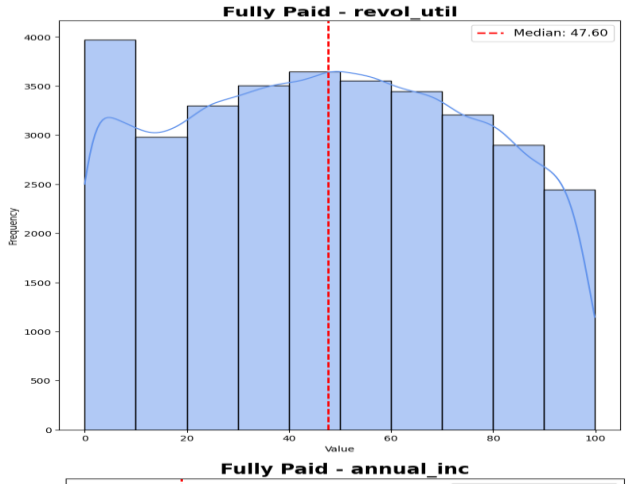
- ✓ Fully Paid - 47 (median)
- ✓ Charged Off - 58 (median)

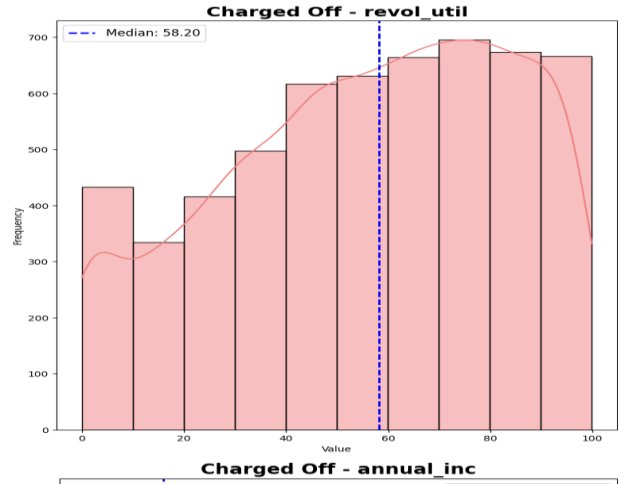
The defaulter annual income is less compared to non-defaulter borrowers.

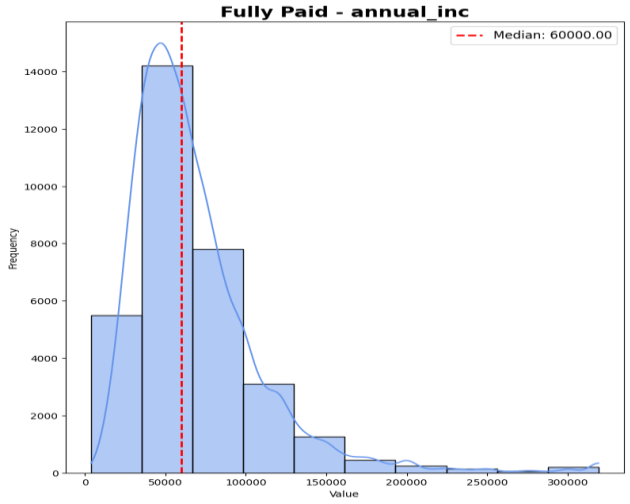
- ✓ Fully Paid - 57,000 (median)
- ✓ Charged Off - 52,000 (median)

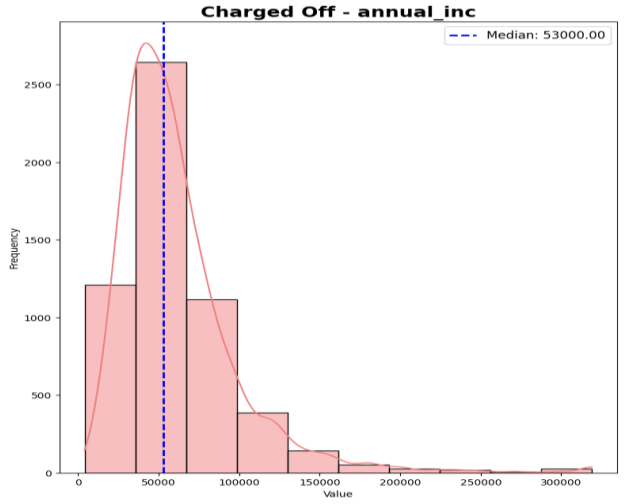










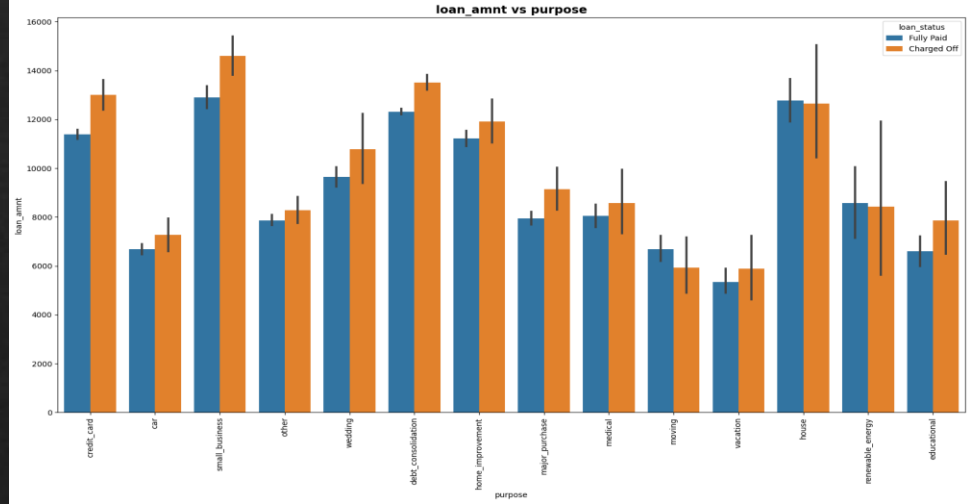


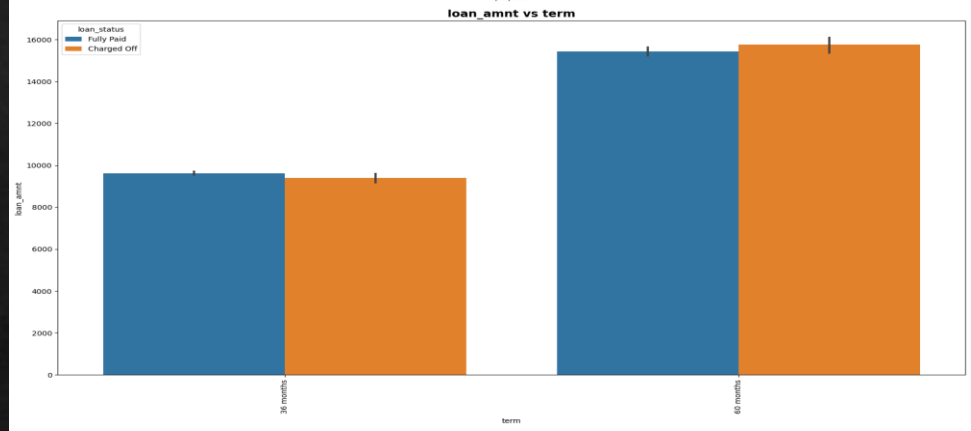
Exploring Data Sets - Bivariate Analysis

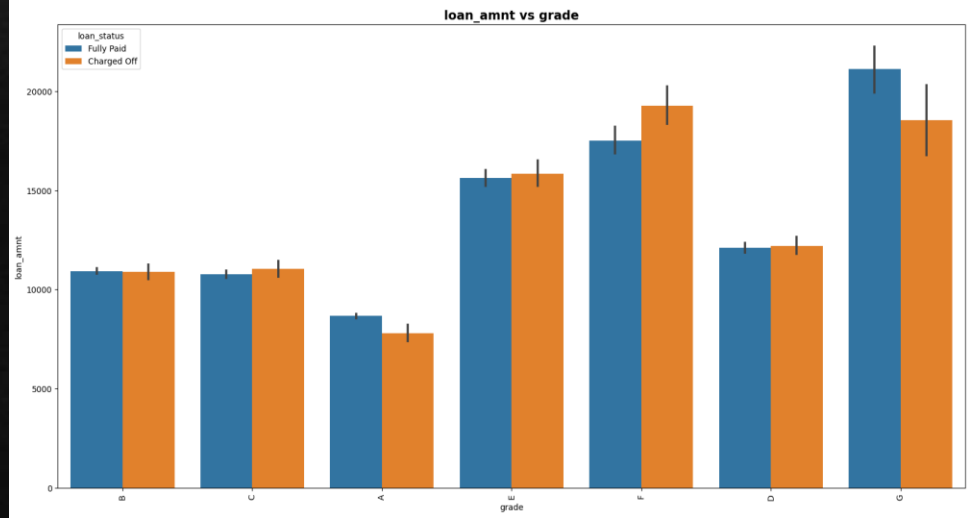
Loan Amount vs Purpose:
Loans taken more for small business, debt consolidation and credit card purposes are likely to result in defaults.

Loan Amount vs Term:
The loan amount is higher for 60-month tenure loans compared to 36-month tenure loans.

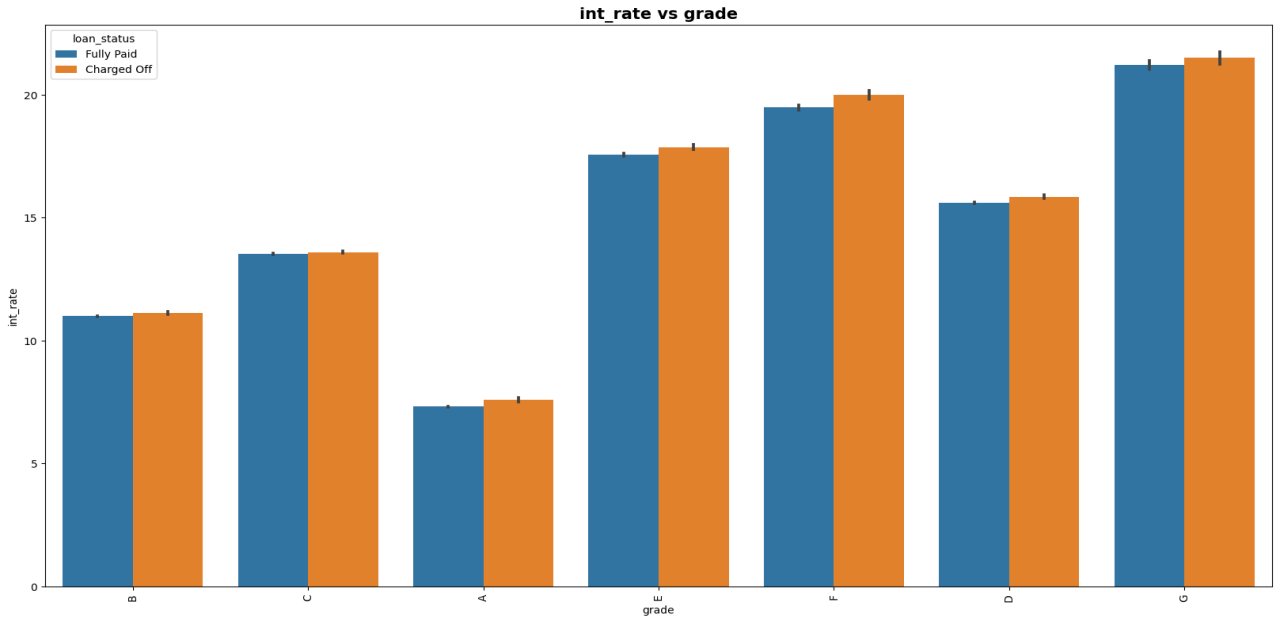
Loan Amount vs Grade:
Loans with grades G, F, and E tend to have higher loan amounts.

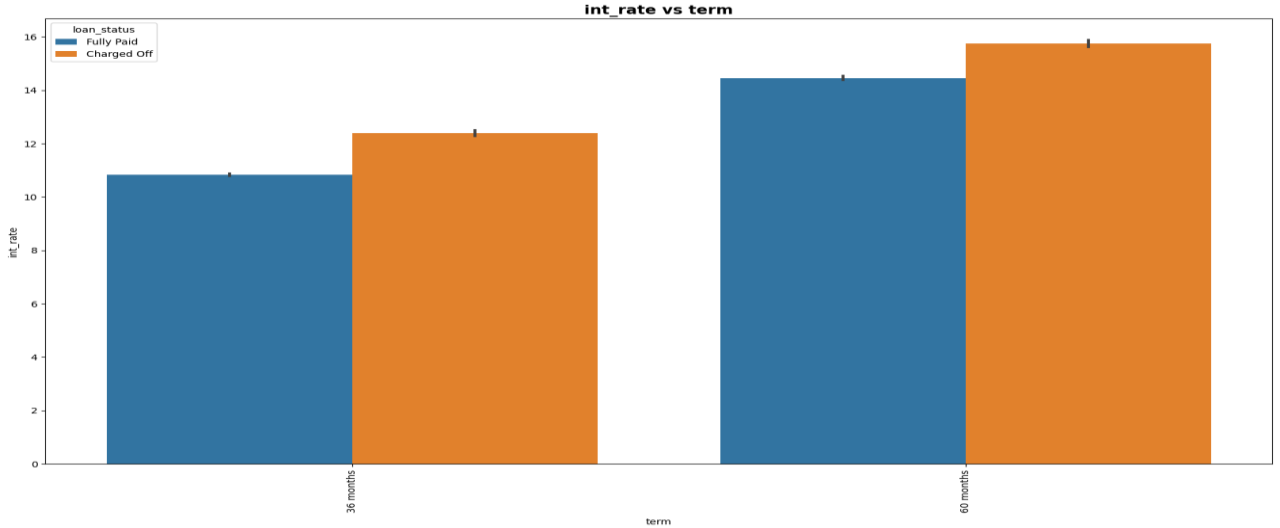






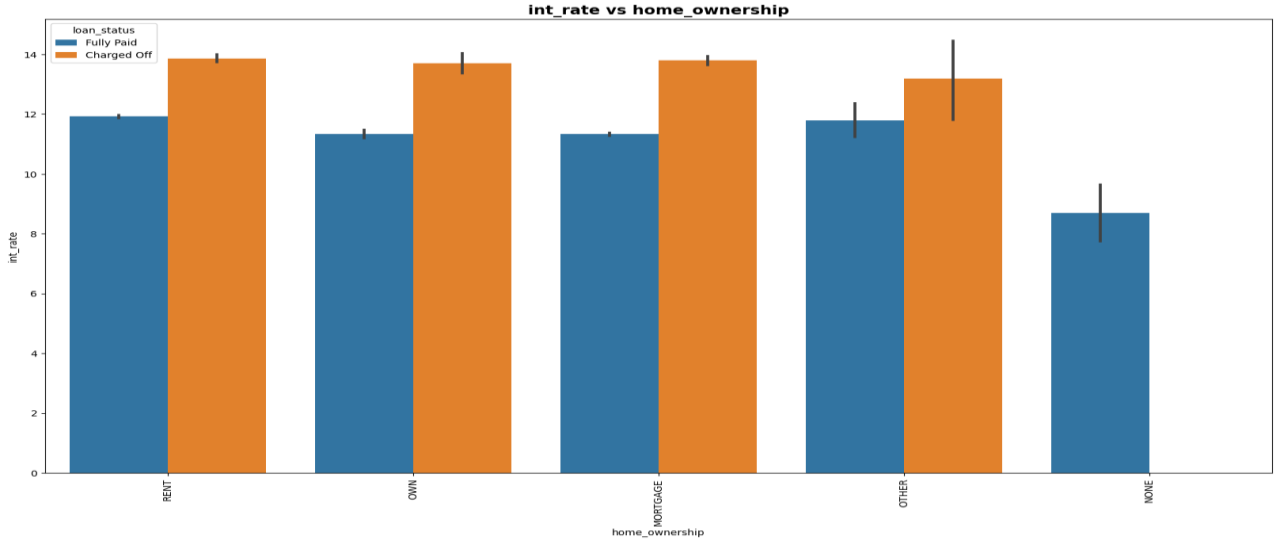
Interest Rate vs Term:
Interest rates are higher for 60-month tenure loans compared to 36-month tenure loans.

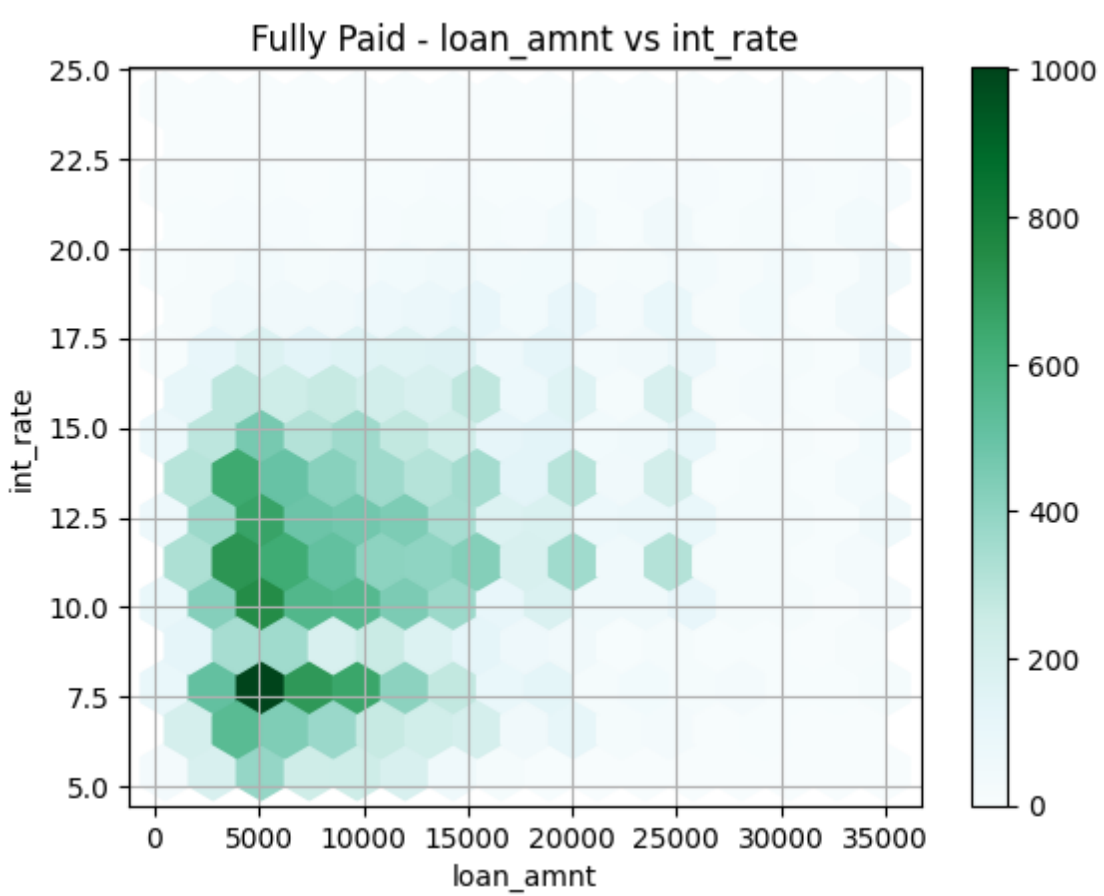


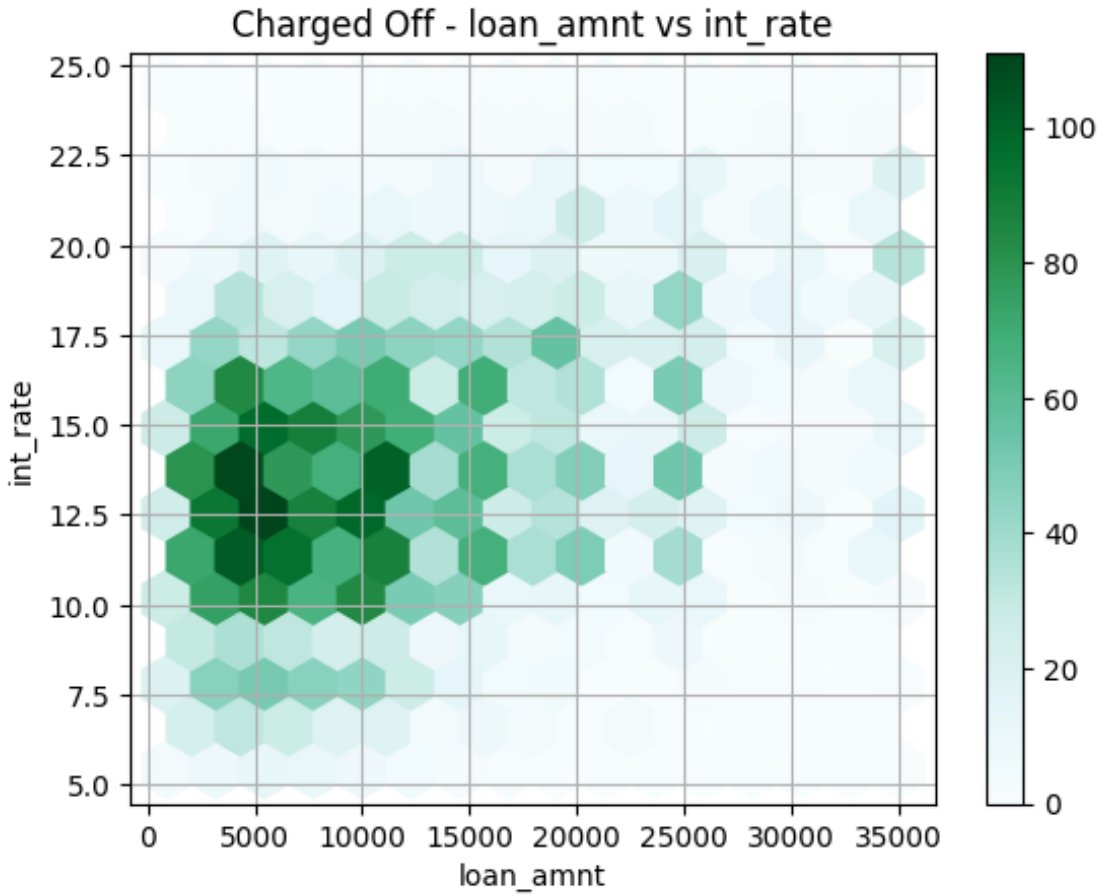


Interest Rate vs Grade:
Interest rates gradually increase for loans with grades A, B, C, D, E, F, and G. G grade loans have the highest interest rates, while A grade loans have the lowest. The interest rates for grade 'G' loans typically range from 20% to 22%.

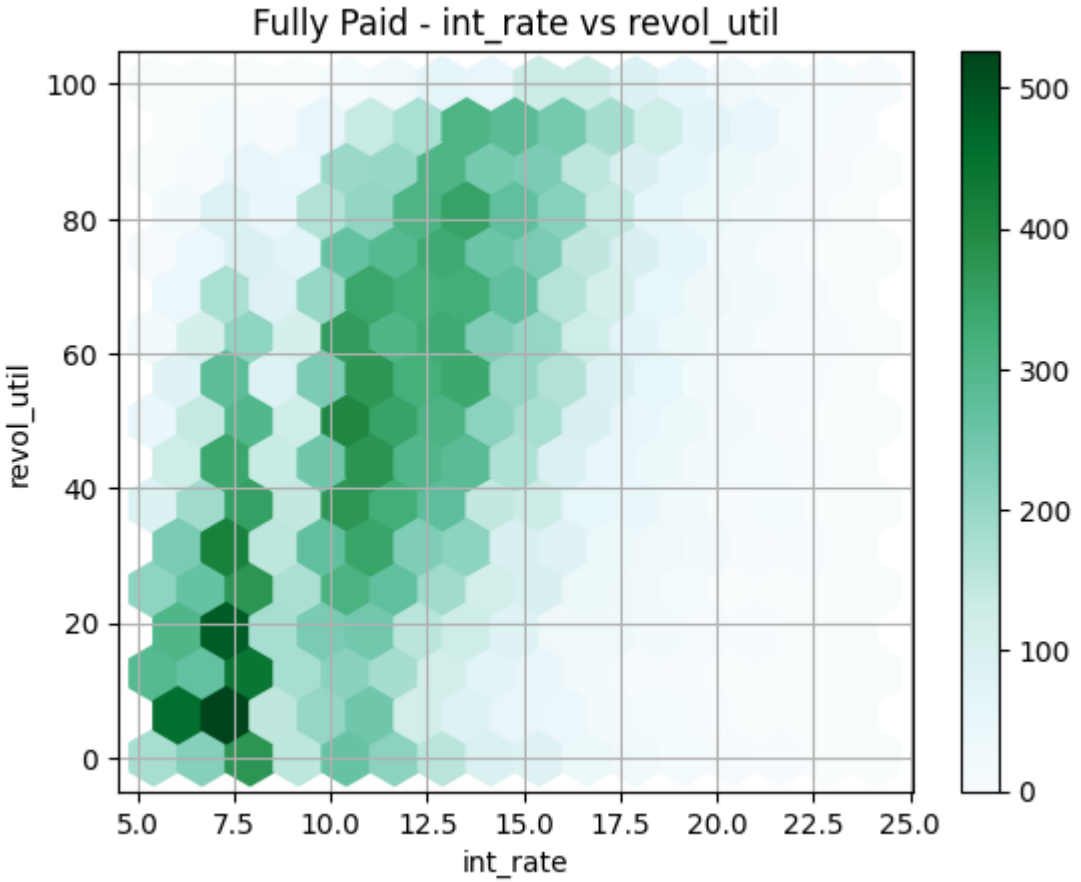
Interest Rate vs Home Relationship:
Regardless of the home relationship, higher interest rates are associated with a higher likelihood of default.

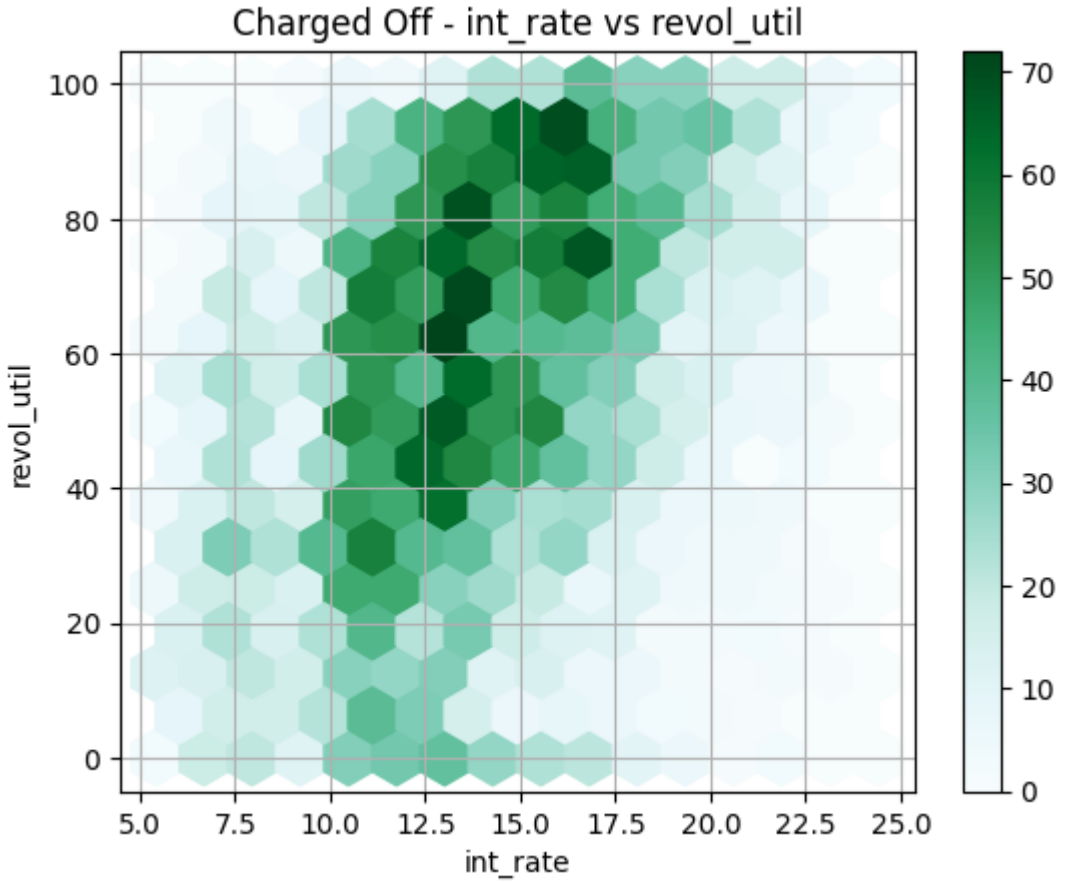






Loan Amount vs Interest Rate:
An increase in the loan amount is associated with an increase in the interest rate, leading to a higher percentage of defaulters.



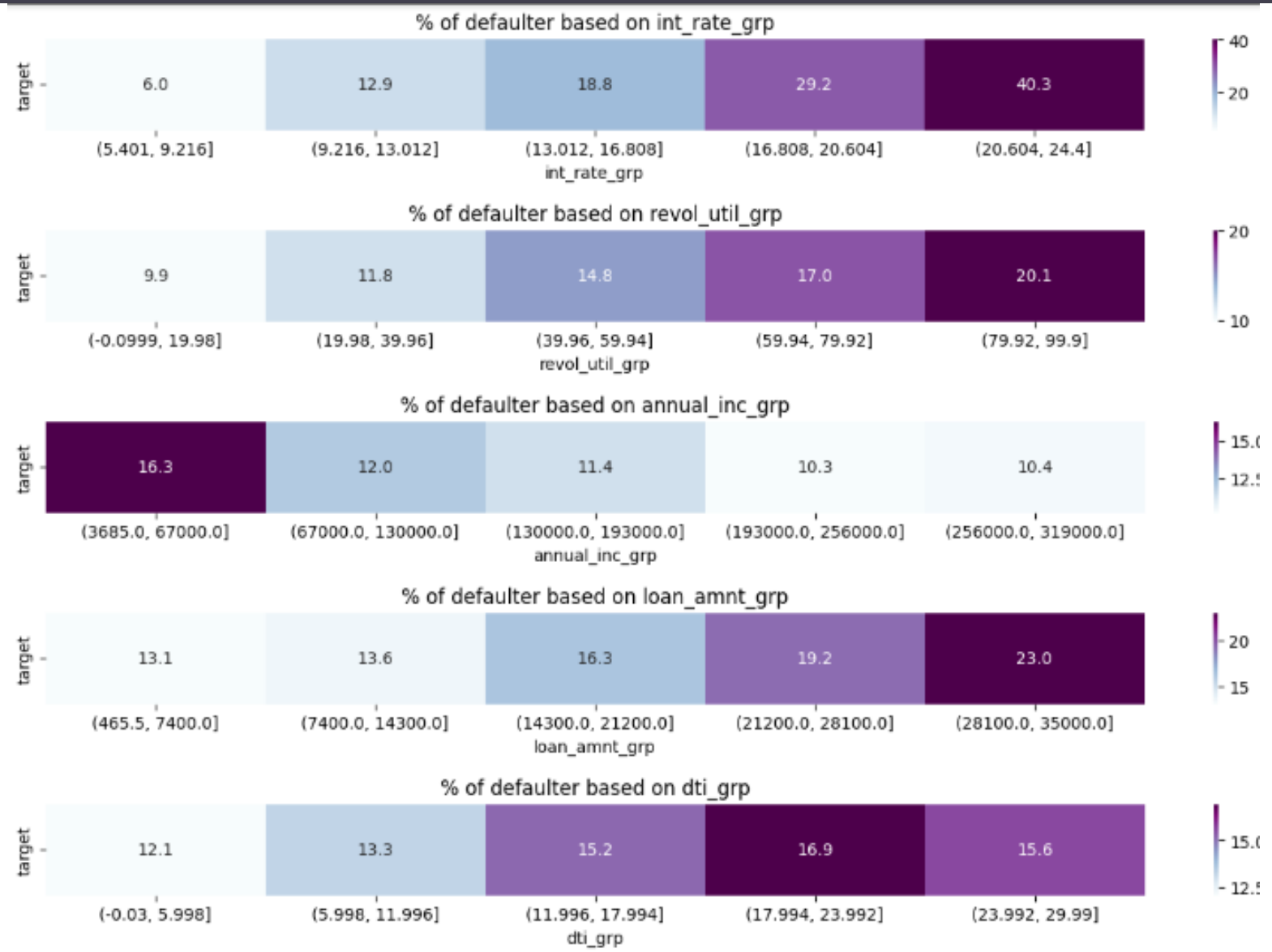


Interest Rate vs Revolving Line Utilization:
Customers with a higher revolving line utilization rate (60-80%) are more likely to become defaulters compared to those with a lower rate (10%-30%) and experience higher interest rates because of their elevated revolving line utilization rate.

Segmented Numerical vs Charged Off/Target

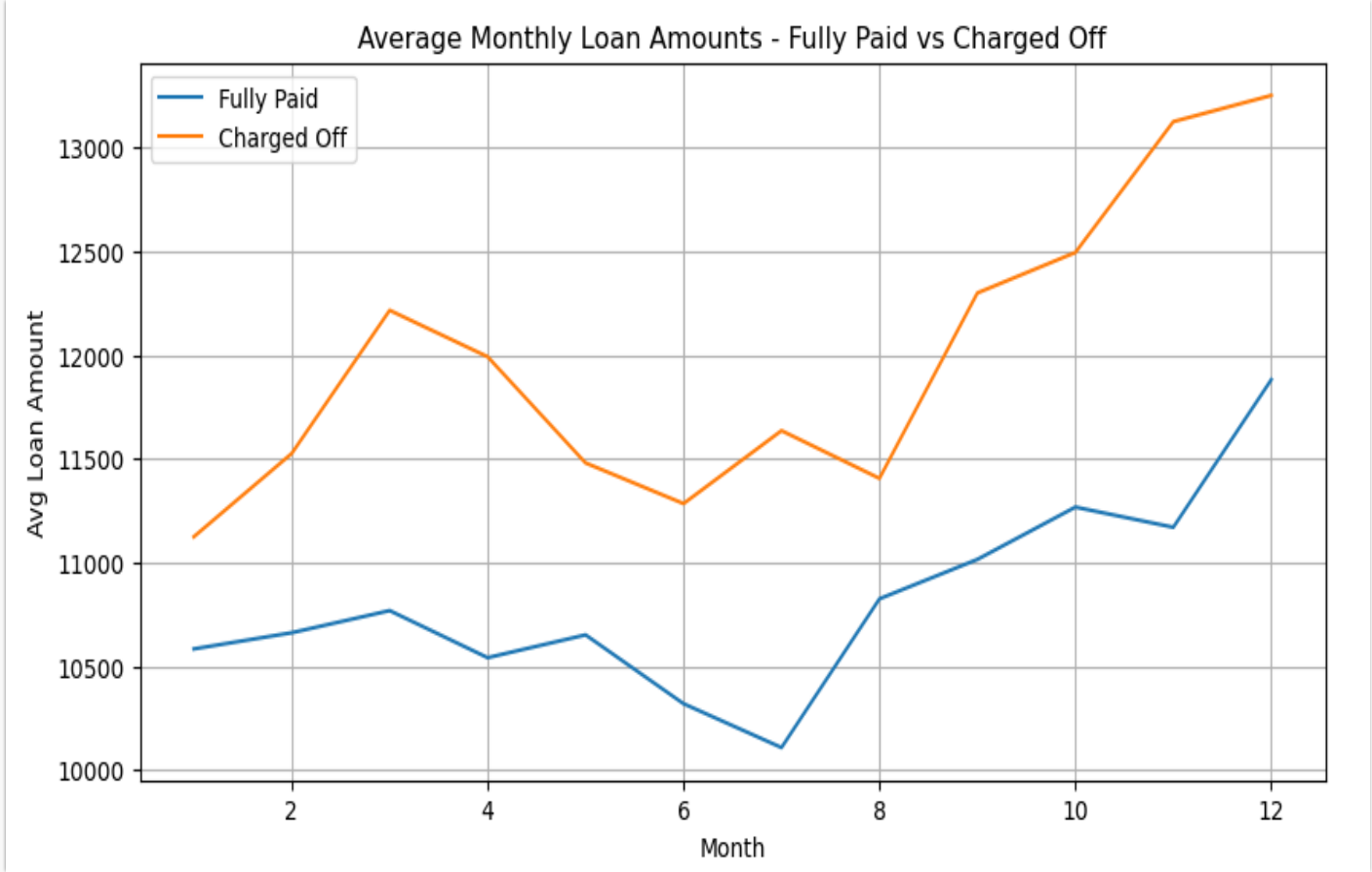
Most of the defaulter belongs to below group

1. Interest rate > 20 %
2. Revolving Line Utilization Rate > 60 %
3. Annual Income < 60K
4. Loan Amount > 25K
5. Debt-to-Income Ratio > 17



Loan Amount vs Loan Issue Month

- In the 4th Quarter of the year with the loan amount ranging from 12k-13k
- Loan amount below 11K in this throughout the year significantly lower the risk.

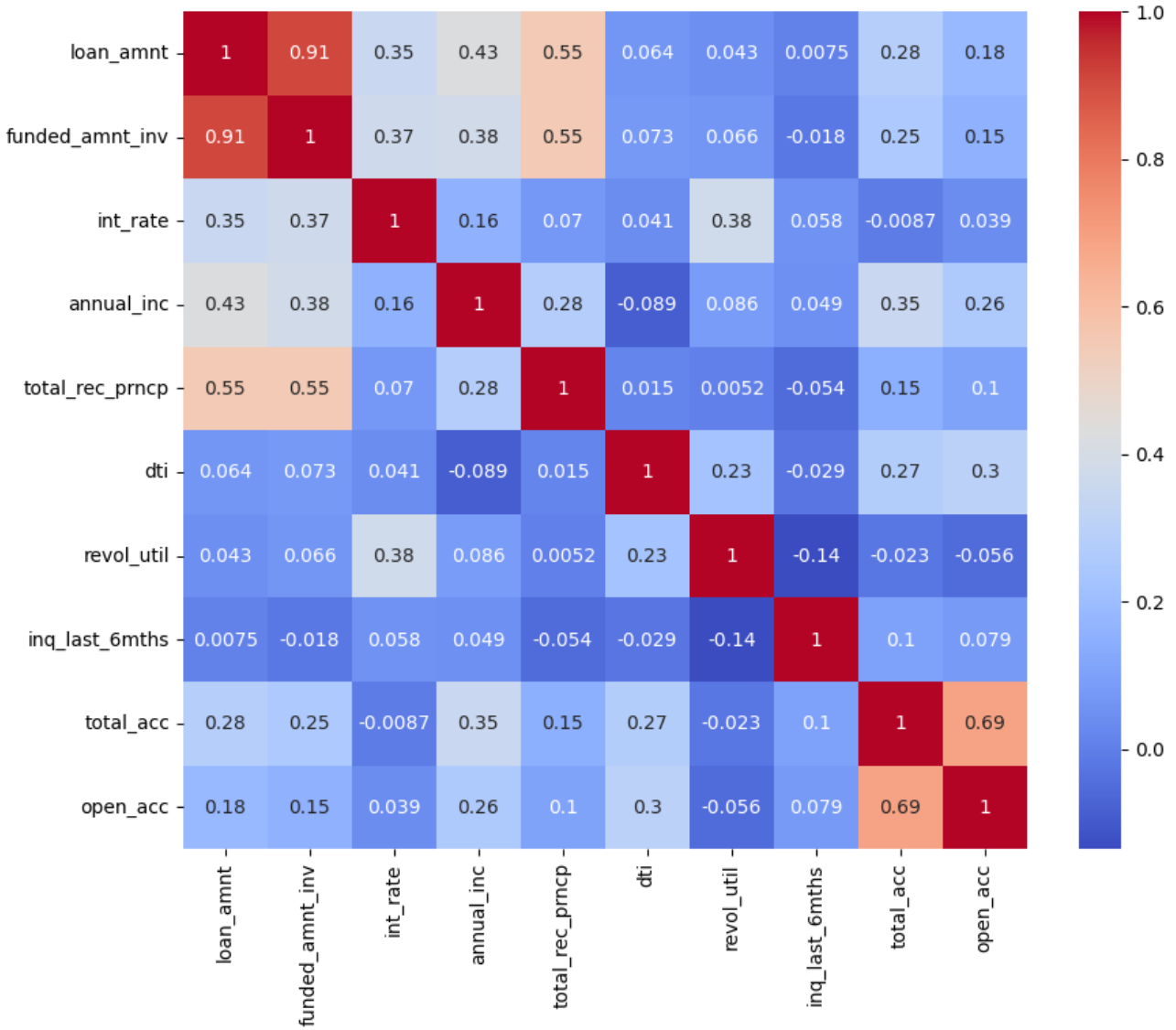


Multivariate Analysis

inq_last_6mths: Enquiry made in last 6 months has a weak or negative correlation with all other numerical columns which indicates one of the driving factor for loan defaulter.

loan_amnt, funded_amnt_inv : A strong correlation between the loan amount requested by the borrower and the amount of the loan that the investor commits to provide.

total_acc, open_acc : The number of open credit lines is highly correlated with the total count of credit lines in the borrower's credit history.



Summary and Recommendations

Loan Term

Loans with a 60-month tenure have a higher risk of default compared to 36-month loans.

Interest Rate

Loans with higher interest rates (>12 %) have a higher chance of defaulting.

Inquiries in the Last 6 Months

A higher number of inquiries made in the last 6 months is associated with a higher likelihood of loan defaults.

Grade

Higher the grade, higher is the interest rate, and high chances of getting default. Loans with grades G have a higher chance of defaulting as they are paying more interest rate (> 20 % interest rate).

Annual Income

Lower annual income is associated with a higher risk of loan defaults.

Debt-to-Income Ratio (DTI)

Higher DTI ratios are associated with a higher likelihood of loan defaults.

Revolving Line Utilization Rate

Higher revolving line utilization rates (> 50 %) are linked to higher default rates.

Employment Length

Borrowers with 10+ years of employment are at higher risk of default.

Conclusion

Larger loan amounts are associated with 60-month tenure loans, loans of higher grades, and loans designated for small businesses, all of which exhibit a higher likelihood of default.

Borrowers with over a decade of employment history and verified income have an elevated default rate.

Loans with superior grades tend to have larger loan amounts applied by borrowers, resulting in higher potential losses if the loan defaults.

Higher interest rates are linked to loans with 60-month tenures, as compared to the lower rates observed for 36-month loans.