Lending Club Case Study

IIITB / upGrad – Machine Learning and Artificial Intelligence – EPGP Program - March 2022

Course Module - Statistics Essentials

- Individual Submission by Vaibhav Jain csevaibhavjain@gmail.com

Problem Statement

Use EDA to understand how consumer attributes and loan attributes influence the tendency of default on a loan.

- Use given dataset which has the 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.

- If one is able to identify risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Business Understanding

A consumer finance company specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make critical decision for loan approval based on applicant's profile. Two types of risks are associated with the company's decision:

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

Lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

Data Analysis Approach

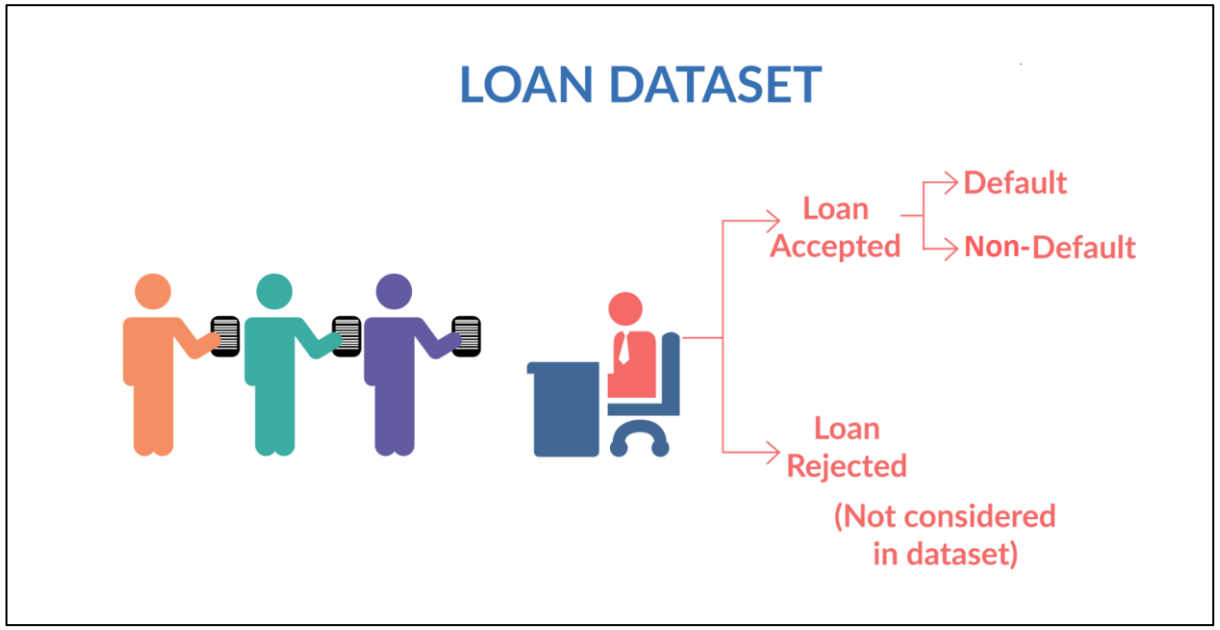
As per exploratory data analysis learning, my intention is to follow these steps in given order to analyze the given dataset

- Data sourcing
- Data cleaning
- Outliers detection
- Univariate analysis
- Bivariate analysis

Data Sourcing

Dataset loan.csv is already provided as part of the case study. This dataset contains the complete loan data for all the loans issued through the time period 2007 to 2011.

- A data dictionary describing the meaning of column names \ variables is also provided Data_Dictionary.xlsx along with the dataset.



When a person applies for a loan, there are two types of decisions that could be taken by the company:

1.Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

2.Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Data Cleaning

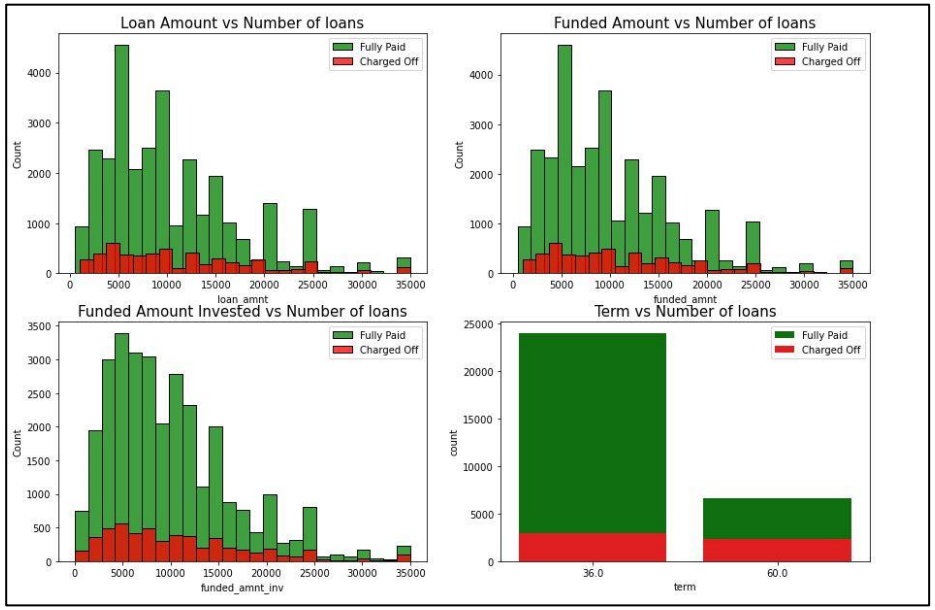
Data cleaning involves understanding and resolving quality issues with the data. This is the most time-consuming tasks of data analysis. I will be identifying formatting errors, missing values, repeated rows, spelling inconsistencies etc. With bad data it is very difficult to perform analysis. This could lead to errors and irrelevant results.

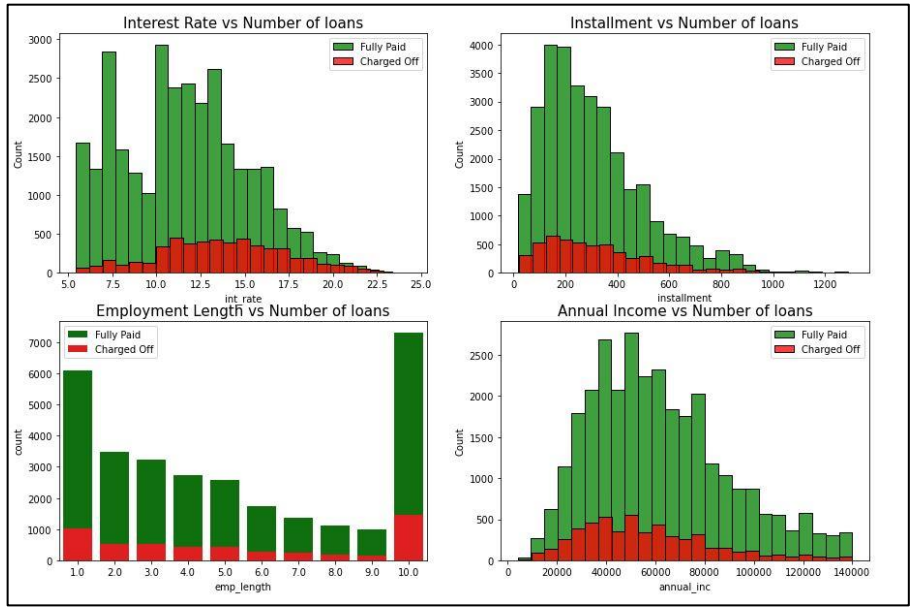
Overall steps to follow for data cleaning:

- Fix rows and columns
- Fix missing values
- Standardize values
- Fix invalid values
- Filter data

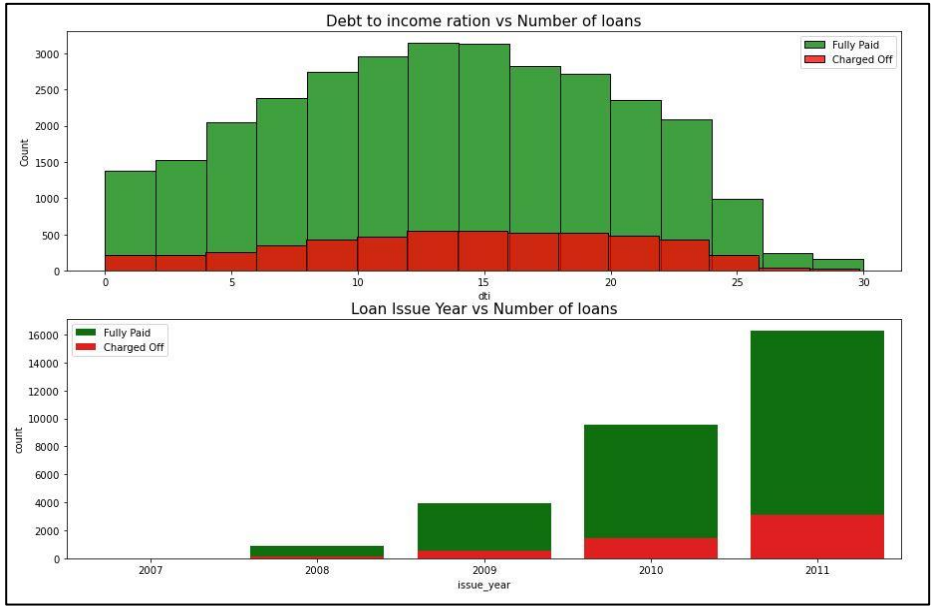
Provided dataset has data for about 40000 loans with loans having more than 110 data points captured for each loan. We need to drop irrelevant columns and keep the key indicators to identify defaulter characteristics.

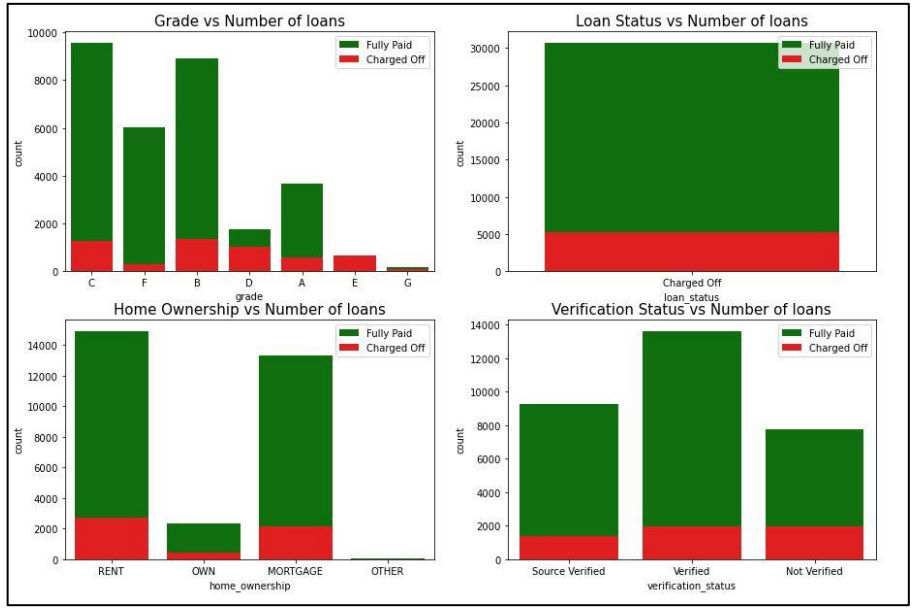
Analysis





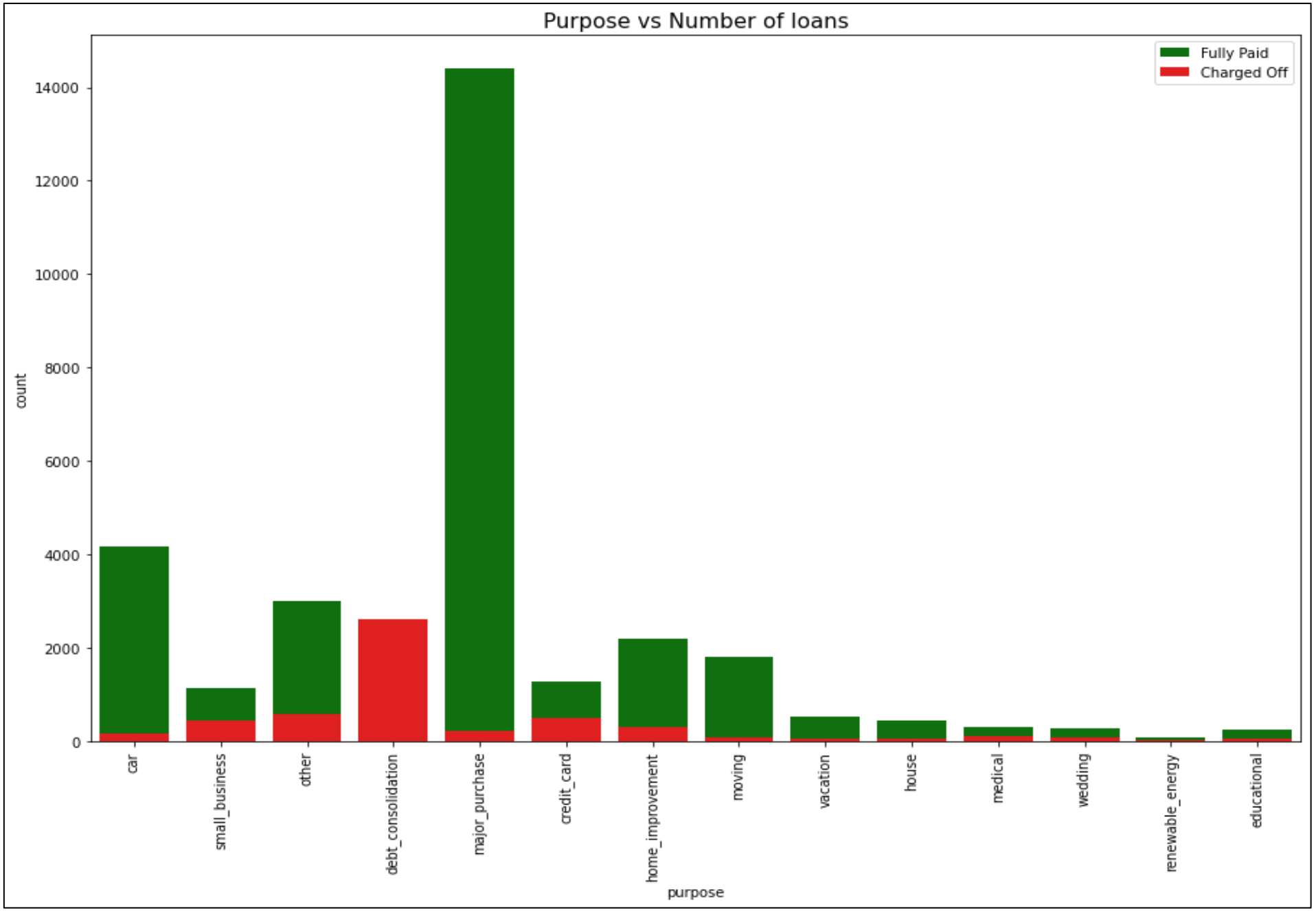
These charts show count of loans across various columns





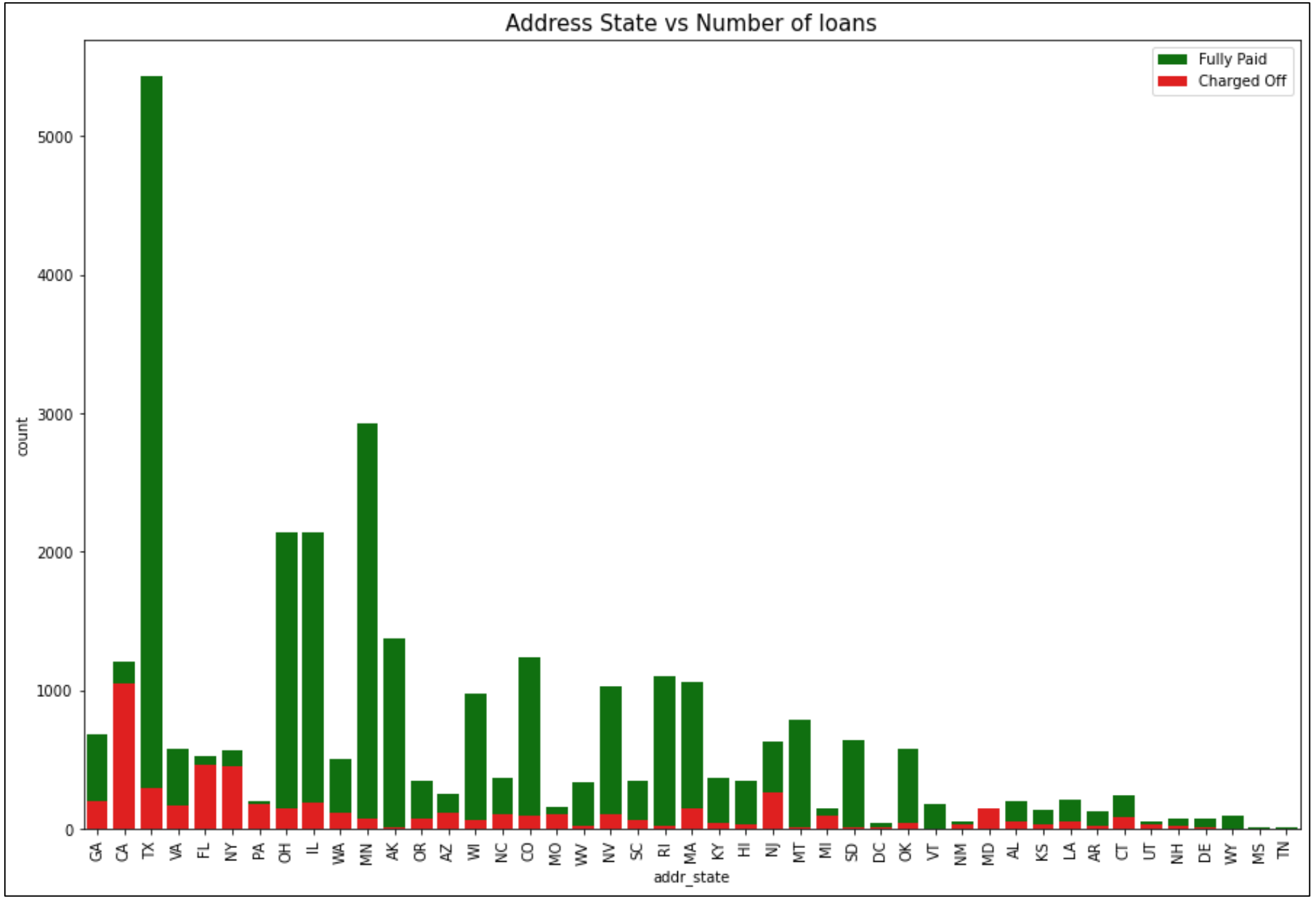
Observations

- Most requested loans are –
  - With loan amount around 5000
  - Term of 36 months
  - Interest Rate 10 % – 13 %
  - Instalment 150 – 350
  - Employment length less than 1 year or more than 10 years
  - Annual income range 35K – 60K
  - Most loans issued in 2011
  - Grades B and C are most common
  - Most loans are verified



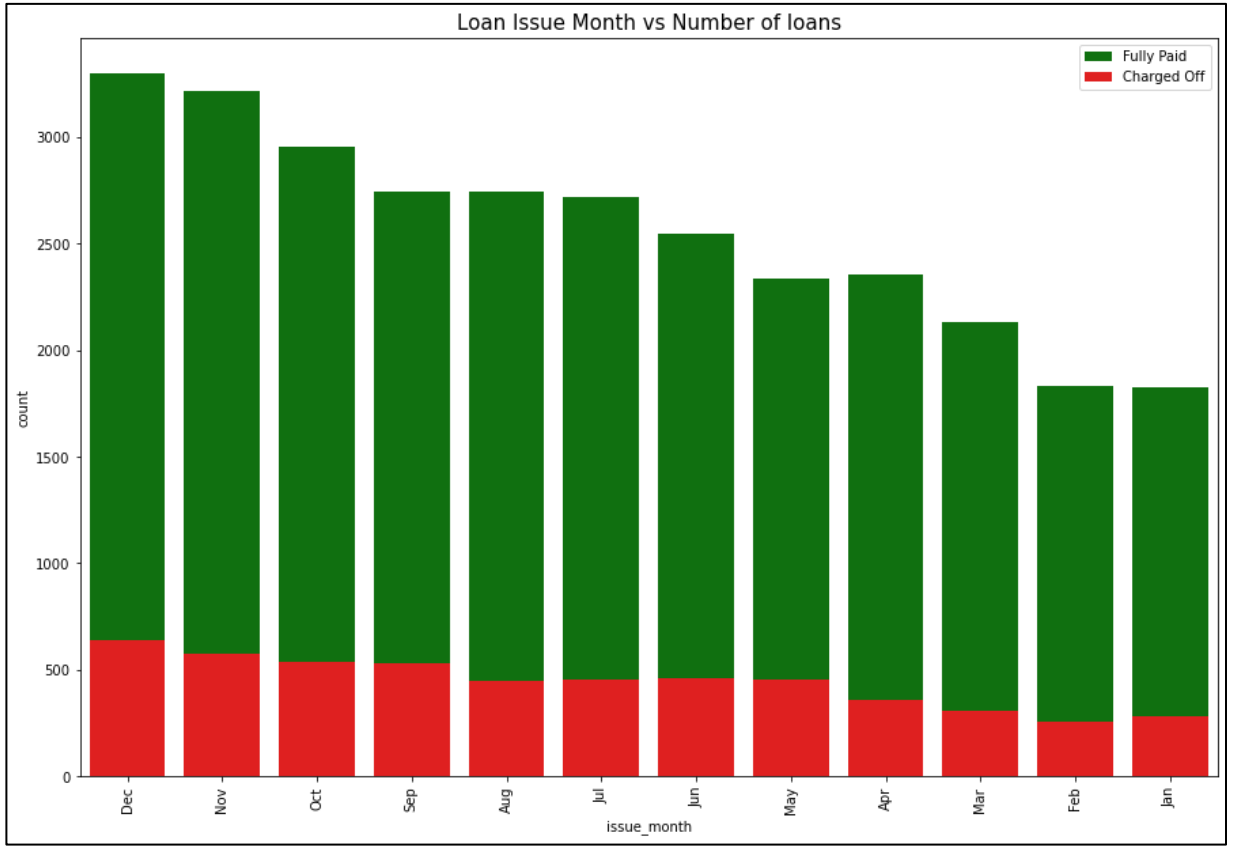
Observations

- Debt_consolidation has most defaulted loans



Observations

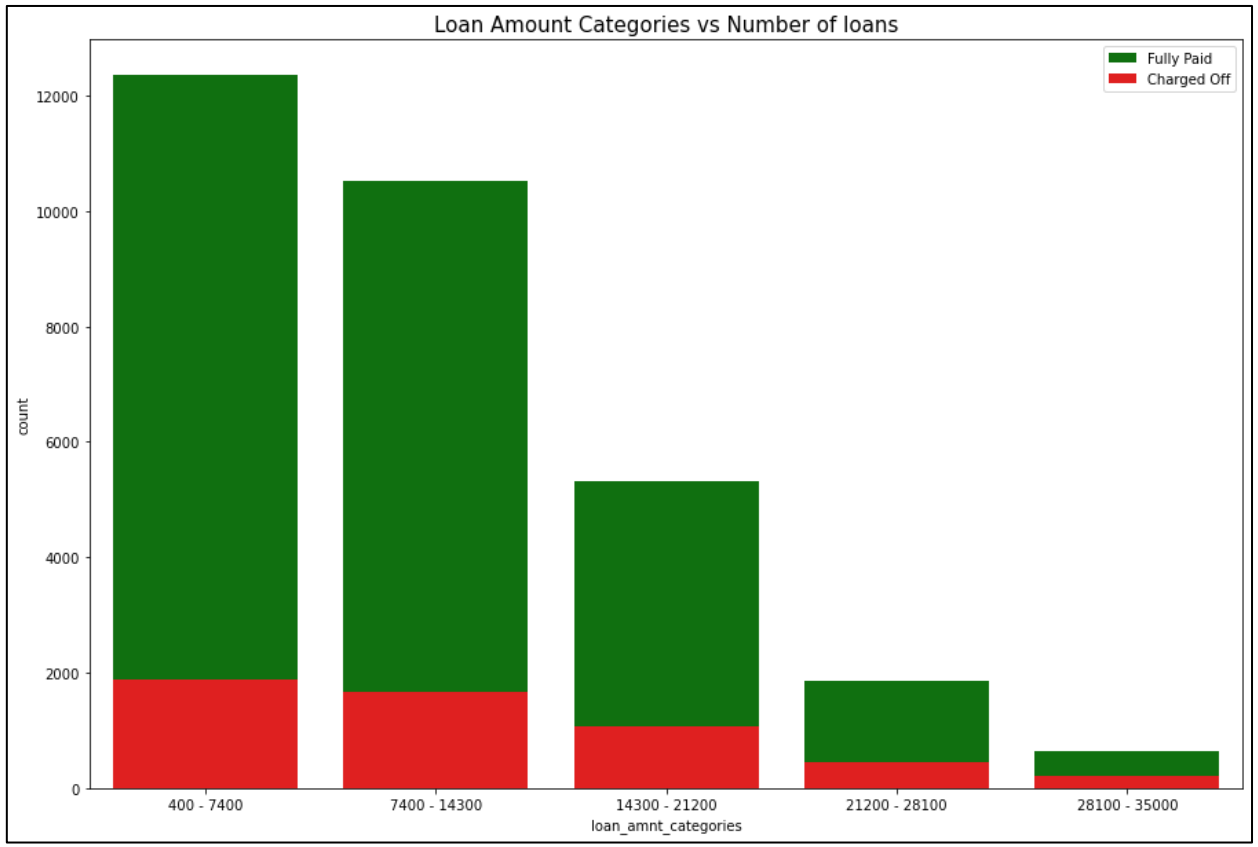
- Most loans are issued in CA, TX, MN state
- CA State has most defaulted loans, followed by FL and NY

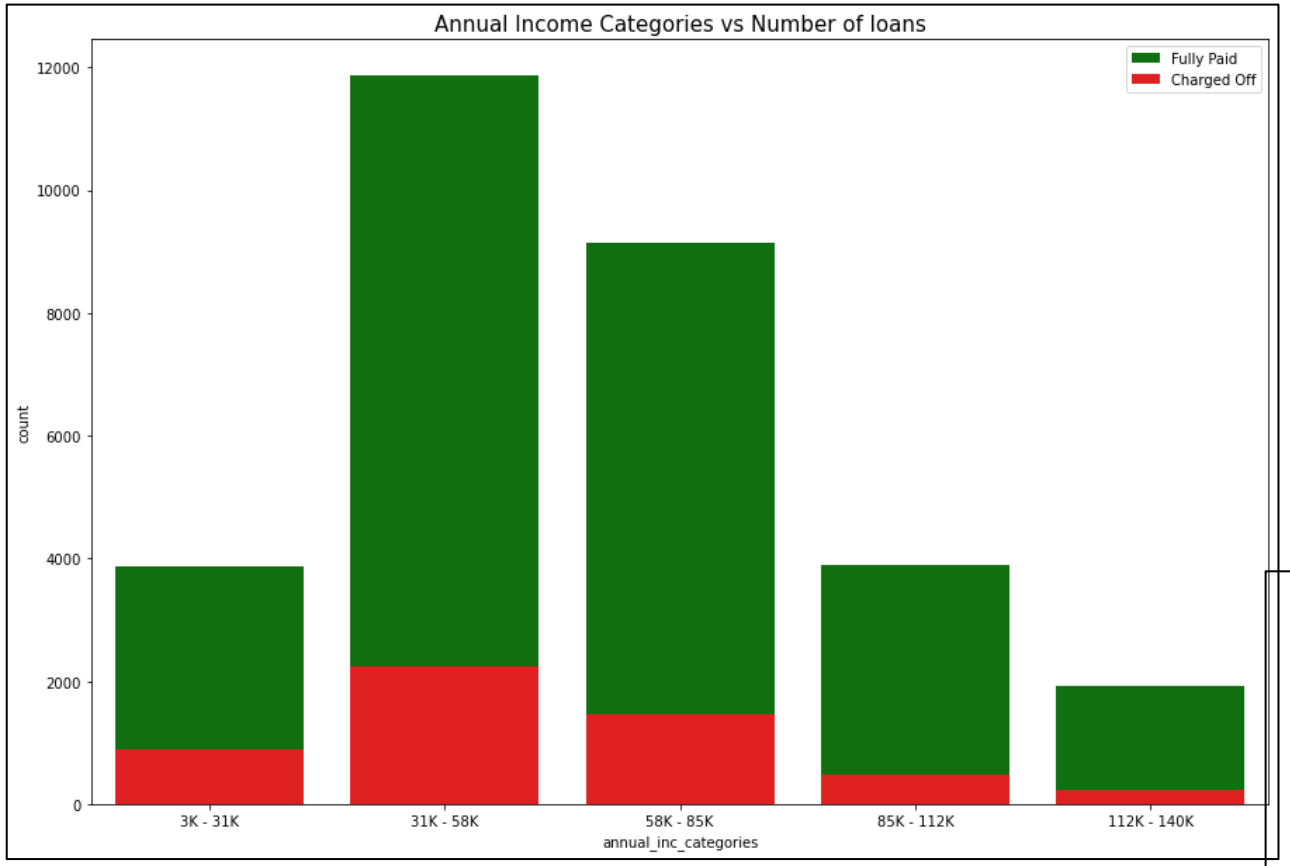


Observations

- Most loans are issued in December
- Dec and Nov month loans have high default rate

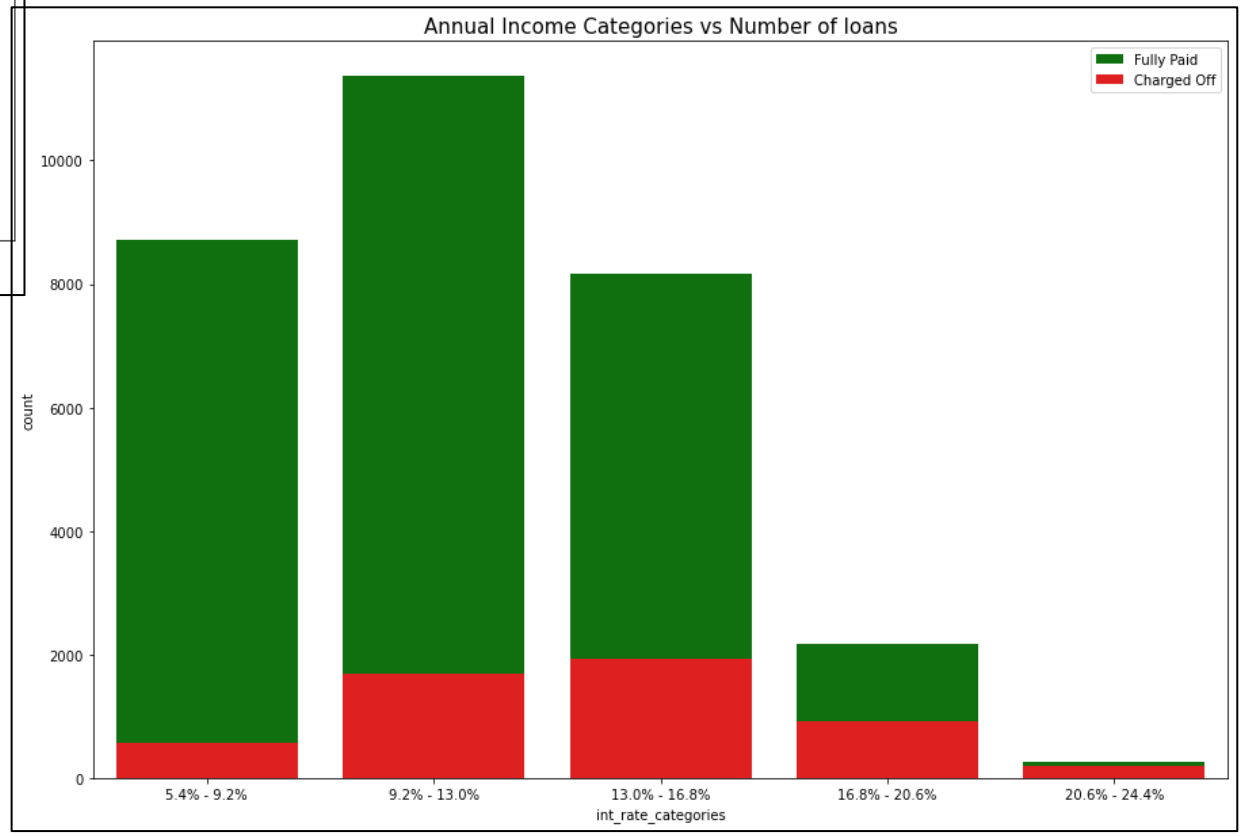
- Most loans are issued for amount range 400 - 7400





Observations

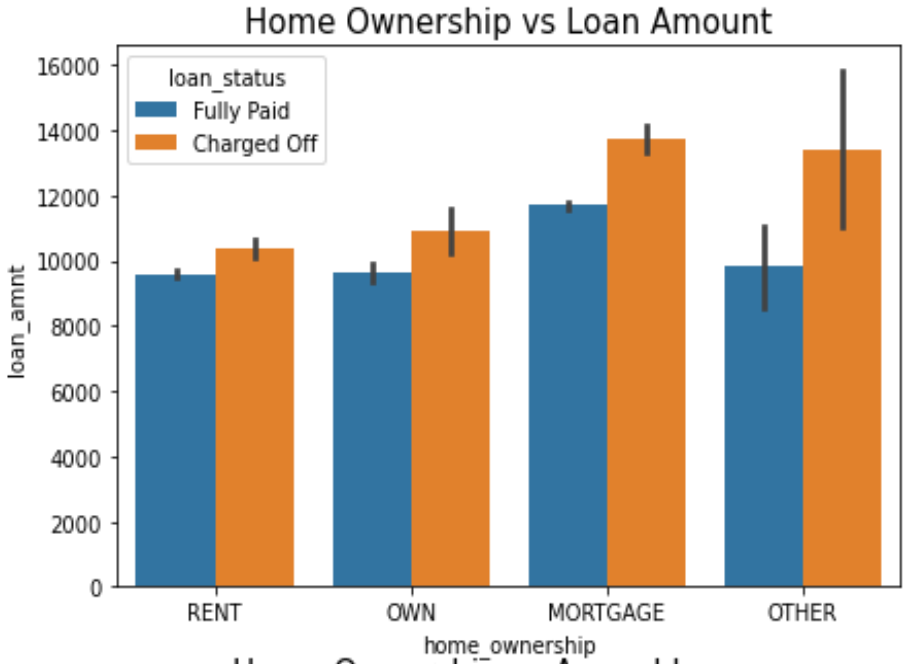
- Annual income range for most borrowers is 31K – 58K and same income range has most defaulters
- Almost all loans with High interest rate > 20% are defaulters

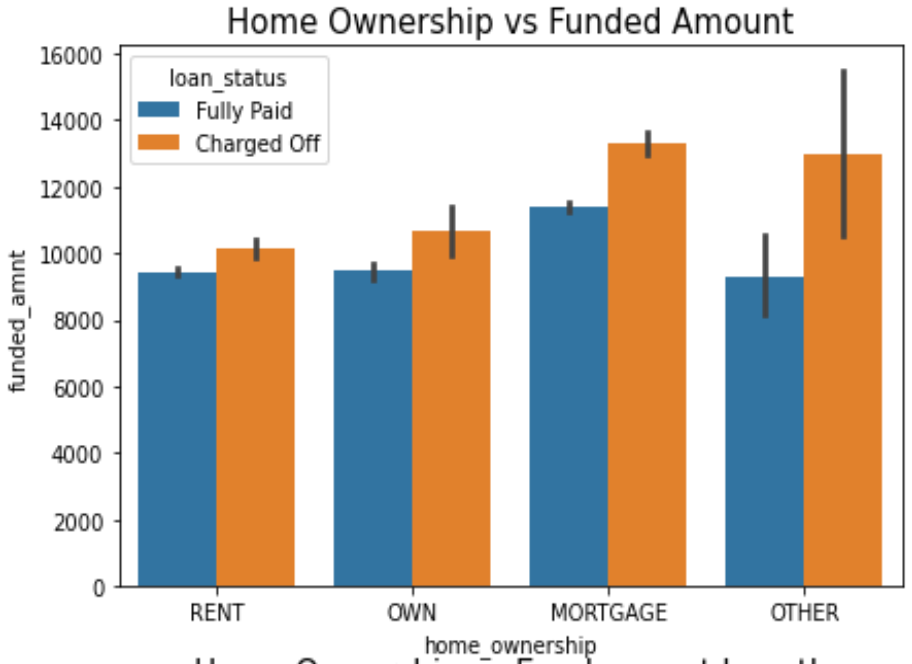


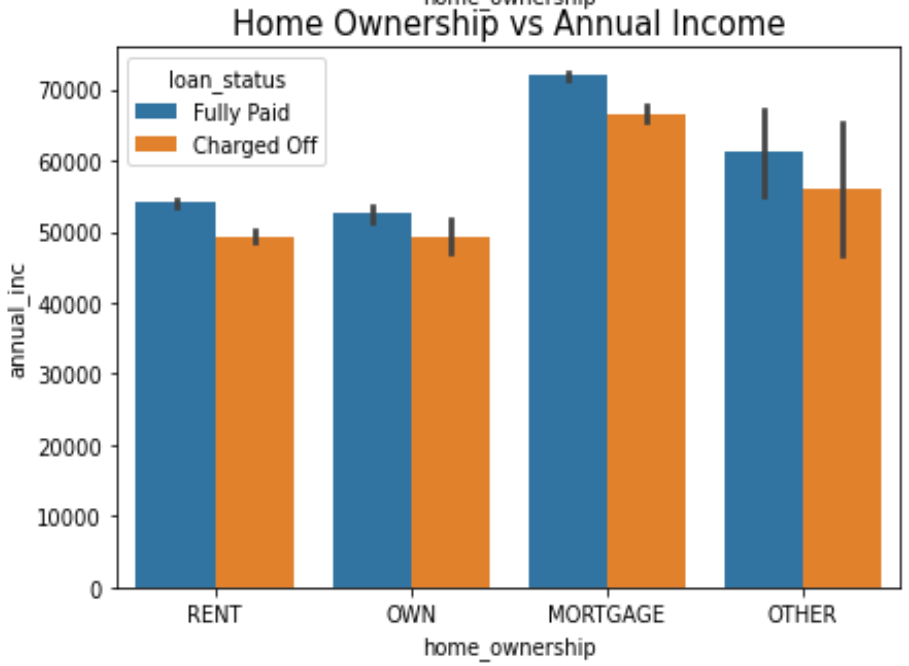
Observations

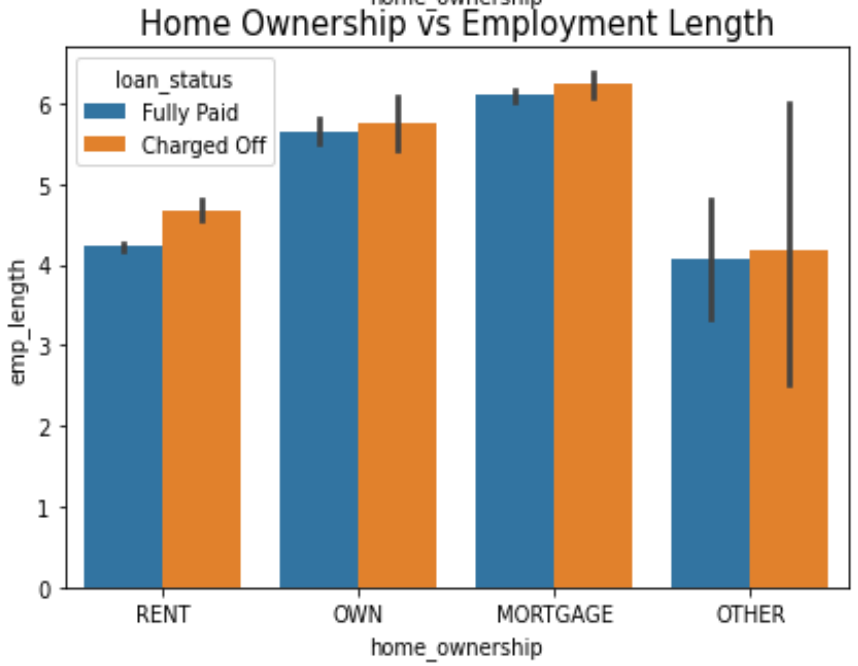
Based on number of loans in each of the above visualizations, we can infer that loan applications with below features have more chances to be defaulters :

- Loan amount up to 15000
- Having interest rate in range of 9% - 13%,
- Having interest rate > 20.6% there is very high probability that loan will be defaulted
- Having experience less than 1 year or more than 10 years
- Is a 36 months loan
- Annual Income within range of 31K - 58K
- DTI range between 10 to 25
- loan applied in 2011
- Grade B, C, or D
- Home ownership of Rent or Mortgage
- Verification status of 'Not Verified' or 'verified'
- Loan purpose of 'debt_condolidation'
- Are from CA, TX, FL, NY, NJ states
- Loan applied in the month of December, November



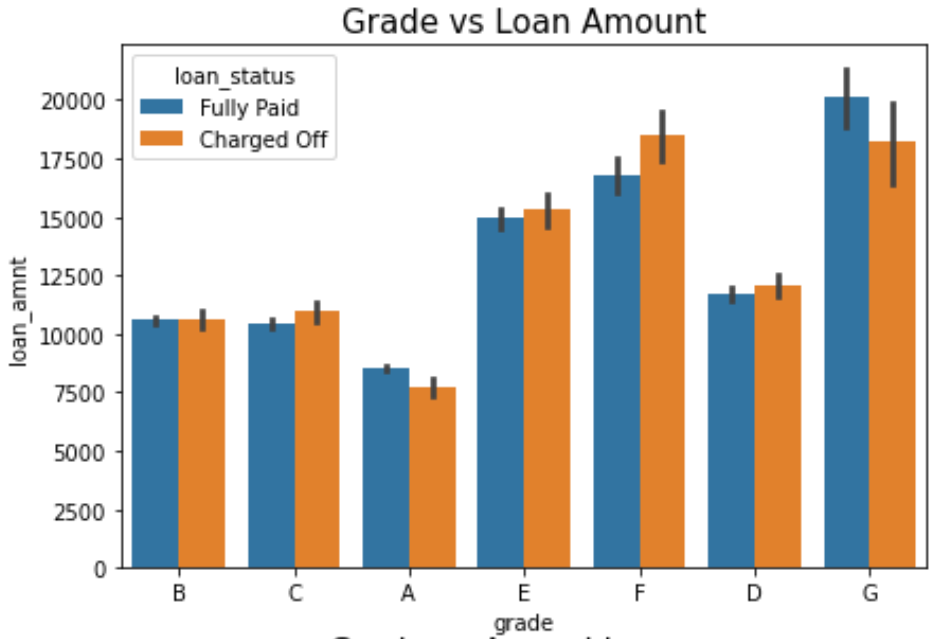


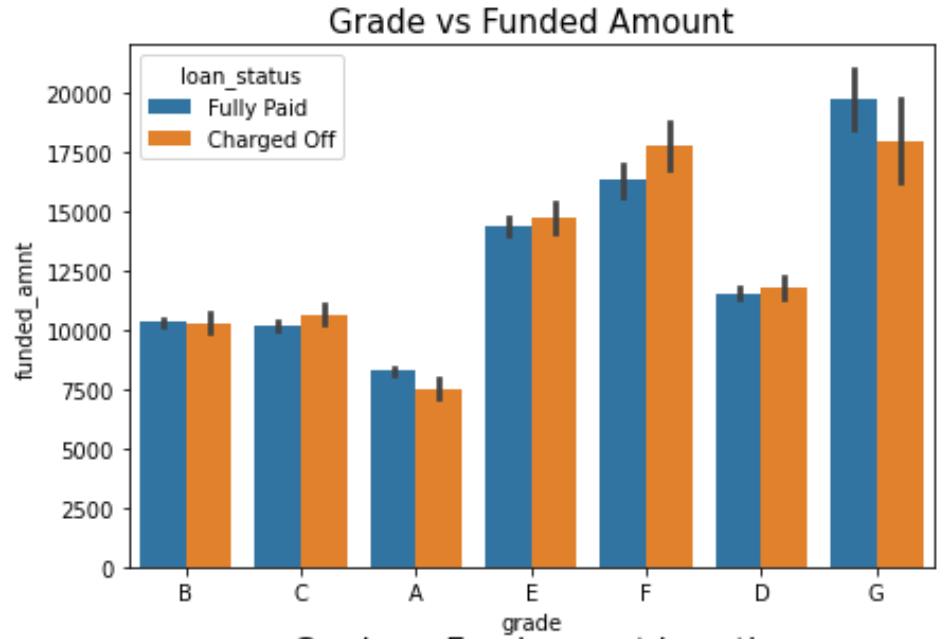


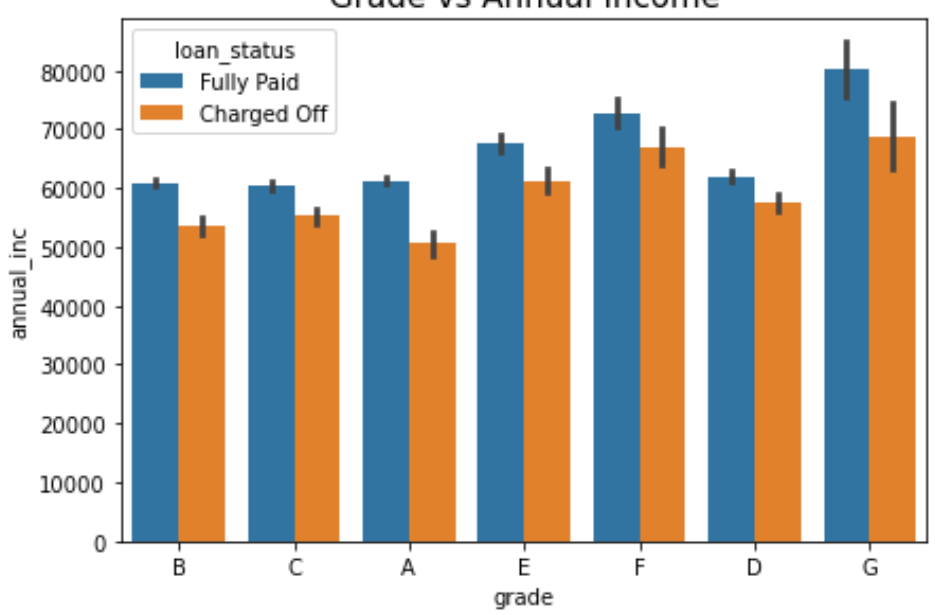


Observations

- Most defaulters either have Mortgage or Other as home ownership with loan amount > 10K

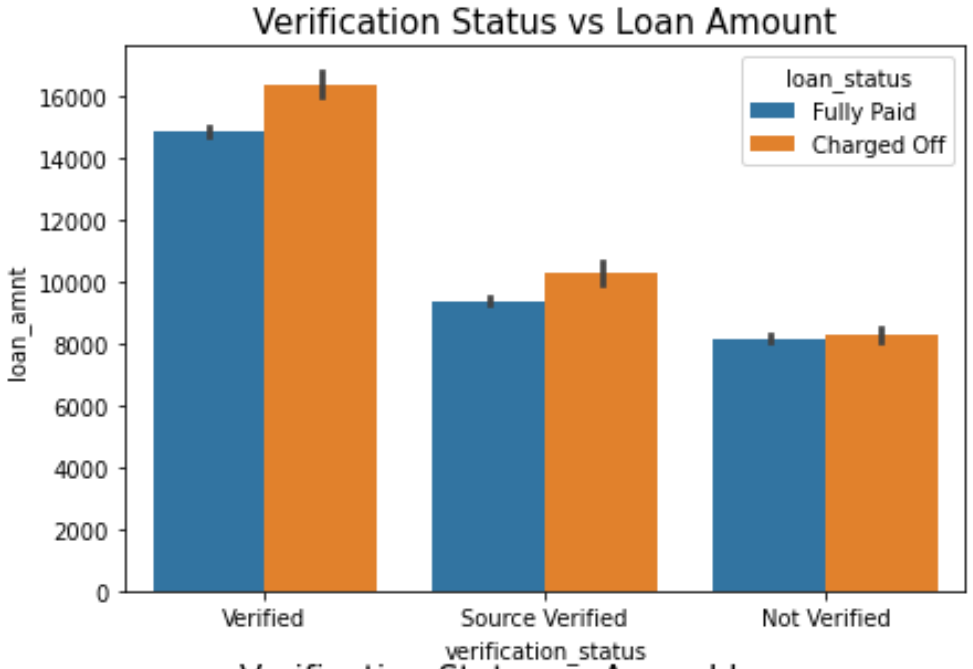


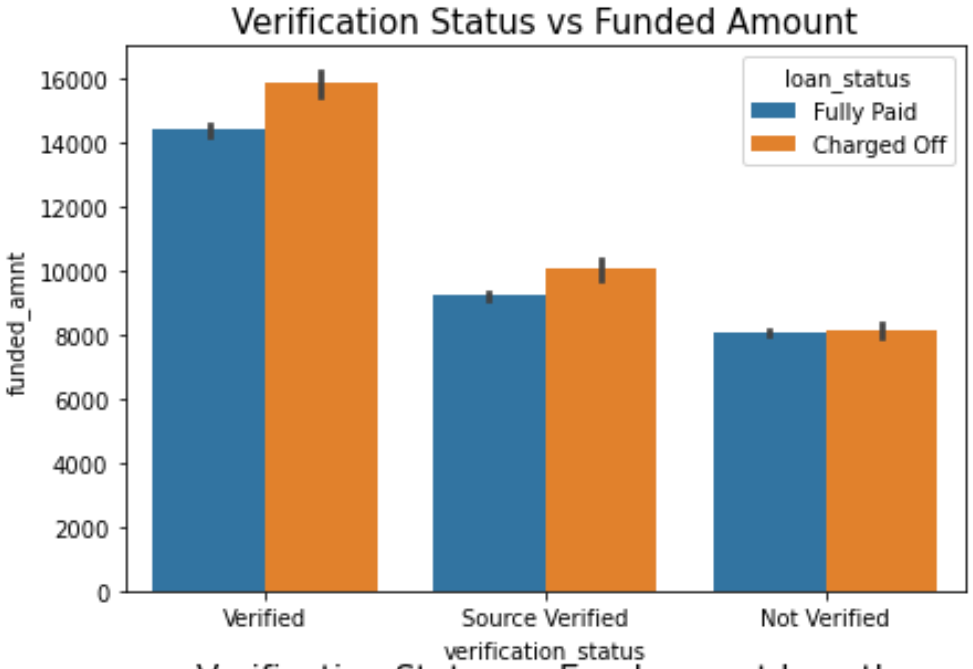


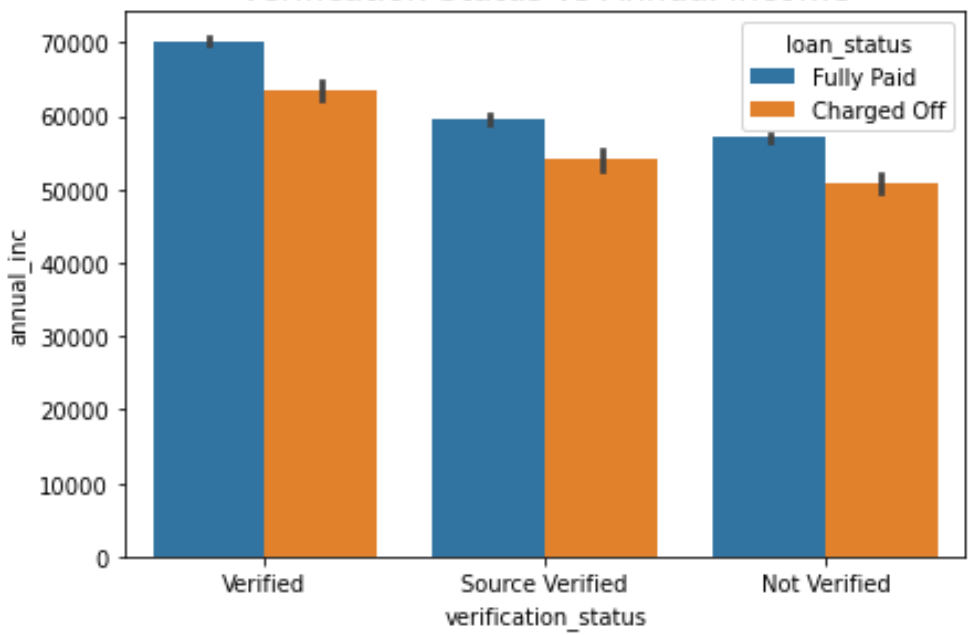


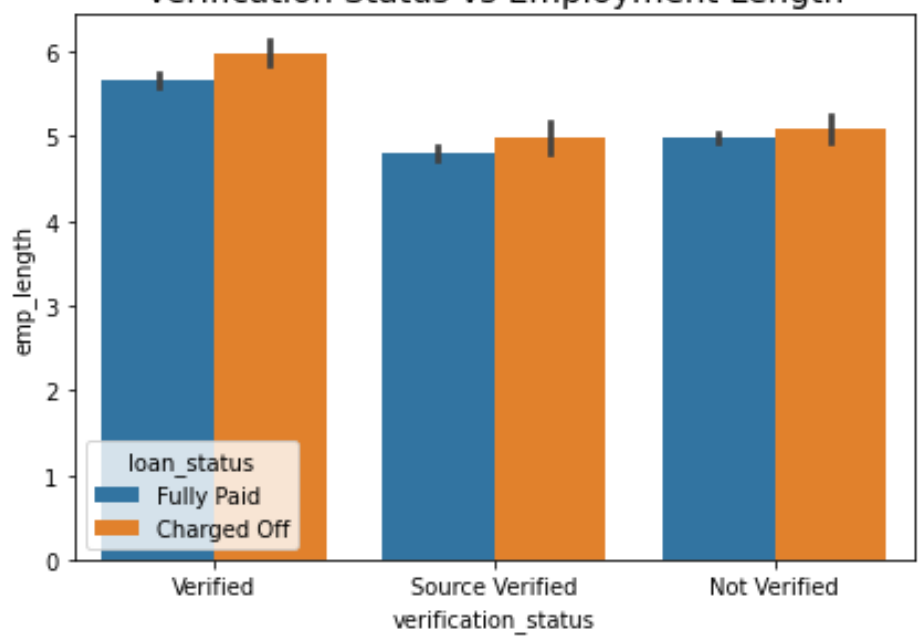
Observations

- Most defaulters are in Grade F or G



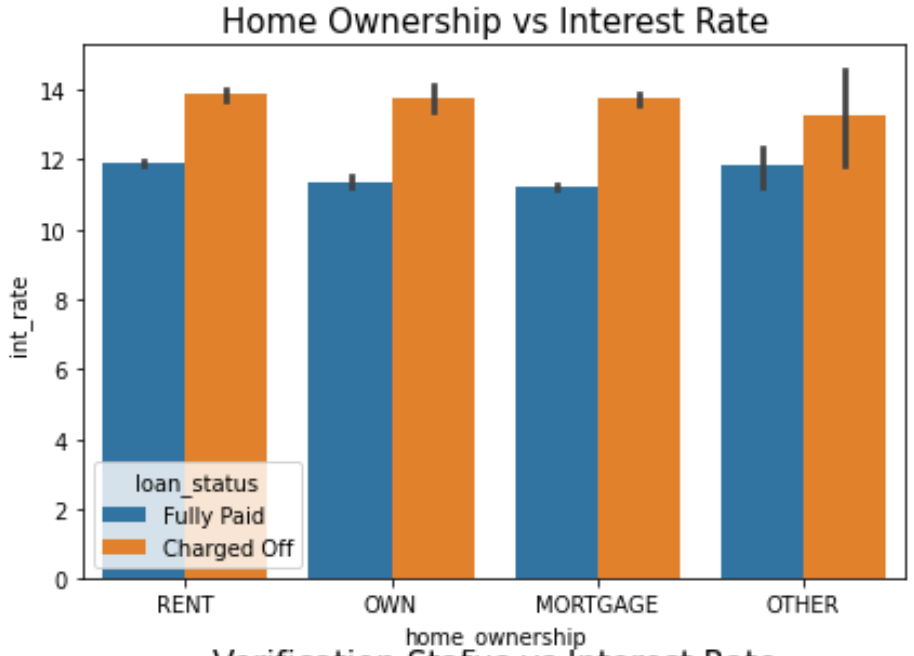


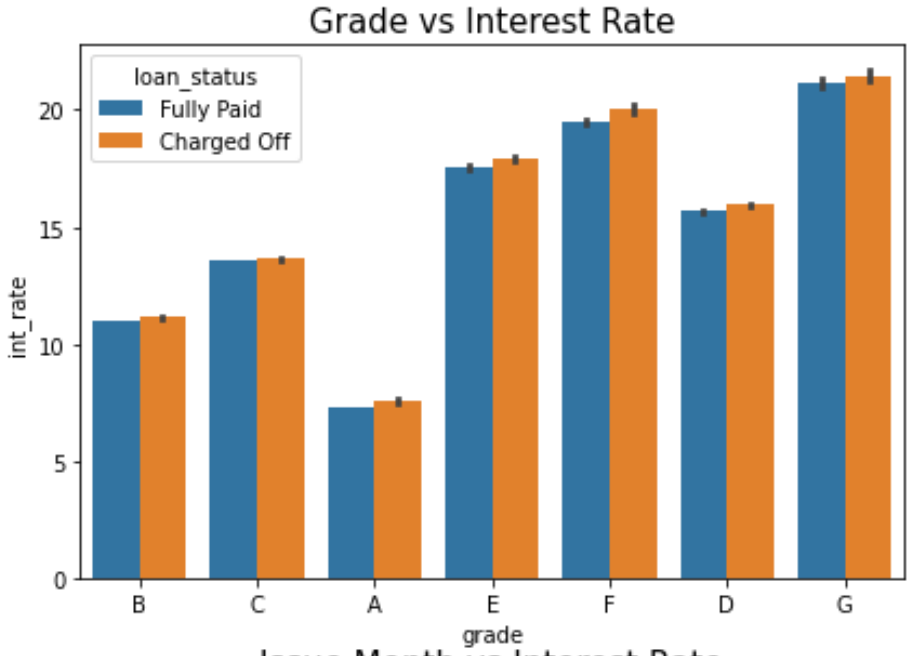


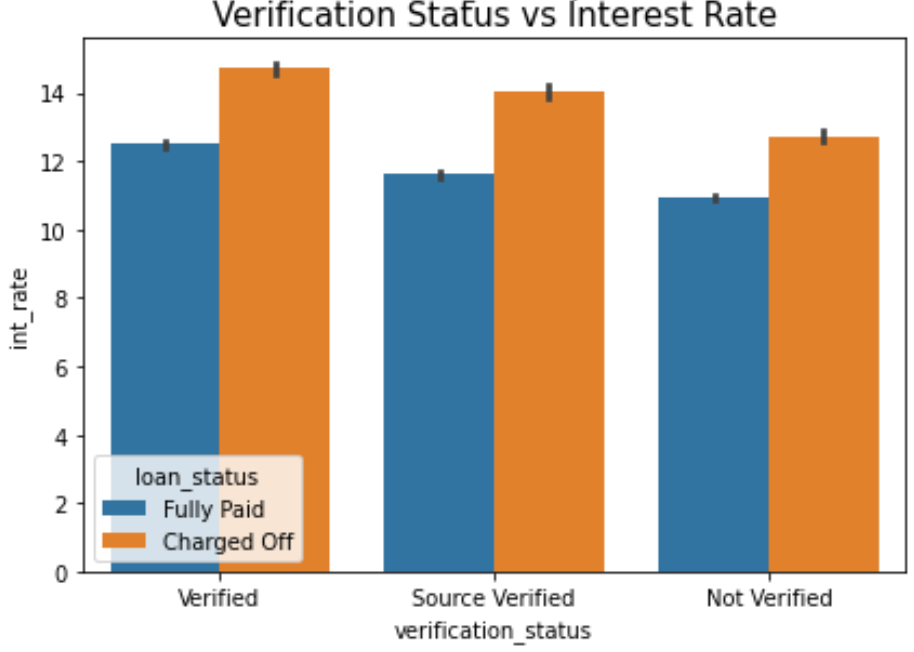


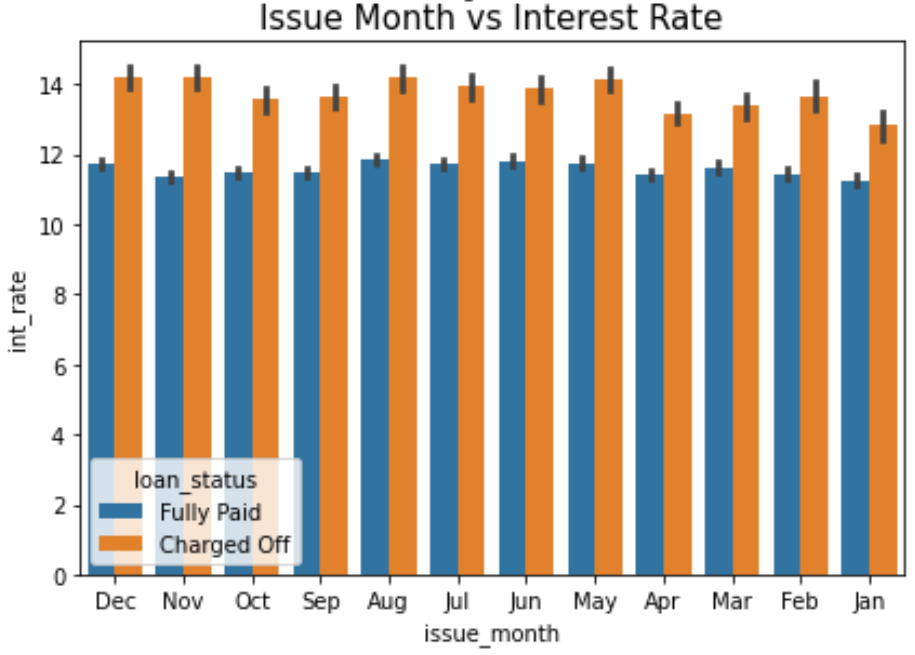
Observations

- Most default loans have Verified status









Observations

- High Interest rate leads to high default percentage

Observations

Based on above visualizations, we can infer that loan applications with below features have more chances to be defaulters:

- Loan amount up to 15000 and have Home ownership of Mortgage or Other
- Loan amount of 13000 - 18000 and Grade F or E
- Annual income around 50000 and Grade F or E
- Verification status of 'Verified' and Loan Amount around 15000
- Verification status of 'Verified' and employment length of about 5 years
- Interest rate > 12%

Conclusion

Based on the EDA analysis done, Loans with below characteristics are most likely to be defaulters:

- Loan purpose is debt-consolidation, is a 36 months loan, and loan being applied for in month of December
- Loan amount is less than 15000, borrower home ownership be one of rent or mortgage
- Loan is on high interest rate >15%, with borrower annual income in range of 31K - 58K
- Loan Grade is E or F, Borrower state is CA