Lending Club Case Study

NAGA MANASA SURIKUCHI

SRIRAM MANNAM

Table of Contents

Problem Details

A consumer finance company which specializes in lending various types of loans to urban customers, wants to minimize the risk of going into credit loss. The main focus is on the loss occurred when the borrower refuses to pay or runs away with the money owed.

The company wants to understand the **driving factors** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Provided:

The loan information about past loan applicants and whether they ‘defaulted’ or not is provided along with other details about the applicants.

Understanding & Objective

Understanding:

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

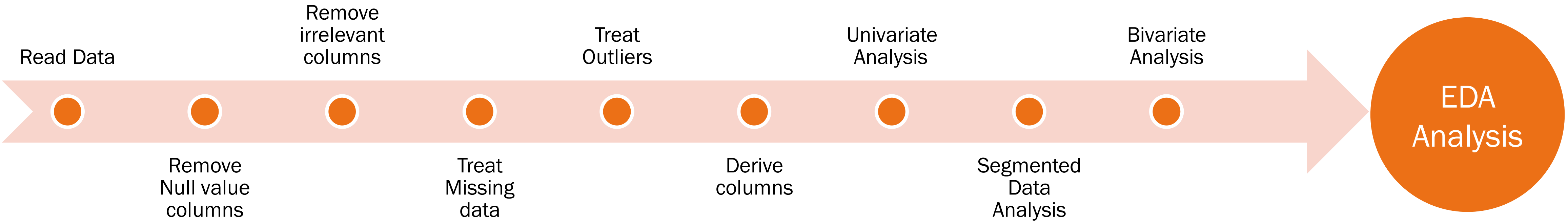
Objective:

Analyze and Understand how **consumer attributes** and **loan attributes** influence the tendency of default.

Approach & Methodology

In order to achieve meaningful insights about the data and the driving factors for loan defaulters, EDA (Exploratory Data Analysis) has to be performed on the given Dataset.

Below are the methods applied in the course of analysis.



Data Cleaning and Handling

- The dataset given has 111 columns and 39717 rows.
- There are 54 columns with null values. Removed all the columns as they don't contribute to our analysis.
- Columns with Identifier data like do not contribute to the analysis and hence removed.
- Some columns have similar and highly correlated data with other columns and such columns are removed.
  - **Identifier columns**
    - Id, member_id
  - **Columns that are highly correlated**
    - Loan_amt, ~~funded_amt, funded_amt_inv~~
    - Term, int_rate and ~~Installment~~
    - Grade, ~~sub_grade~~
    - Addr_stat, ~~Zip_code~~

Data Cleaning and Handling

- Columns that are having only one value are removed
  - Pymnt_plan → having only ‘n’
  - policy_code → 1
  - application_type → INDIVIDUAL
  - acc_now_delinq → 0
  - chargeoff_within_12_mths -> 0
  - delinq_amnt → 0
  - tax_liens → 0 or NA

Columns related to defaulted are removed
  - URL, desc

Date columns
  - last_pymnt_d, last_pymnt_amnt, next_pymnt_d, last_credit_pull_d

Data Cleaning and Handling

- Some columns with missing data, but required for our analysis are treated with filling the missing values with mode as it does not change the integrity of the dataset.
- Then the datatypes of some columns like int_rate, emp_length are changed from object to numeric as it should be.

Standardizing columns

- Term -> 36 months, 60 months → 36, 60
- Int_rate -> 10.0% -> 10
- Emp_length -> 3 years, 6 years, 10+ years → 3 ,6, 10
- Revol_util -> 45% -> 45

- loan_status column has three values namely “Fully Paid”, “Charged off” and “Current”. Our analysis is based on the comparison between “Fully paid” and “Charged off” data. Hence we can ignore the data in rows corresponding to “Current” data.
- Univariate, Segmented and Bivariate analysis is performed for further analysis of data.

Data Analysis

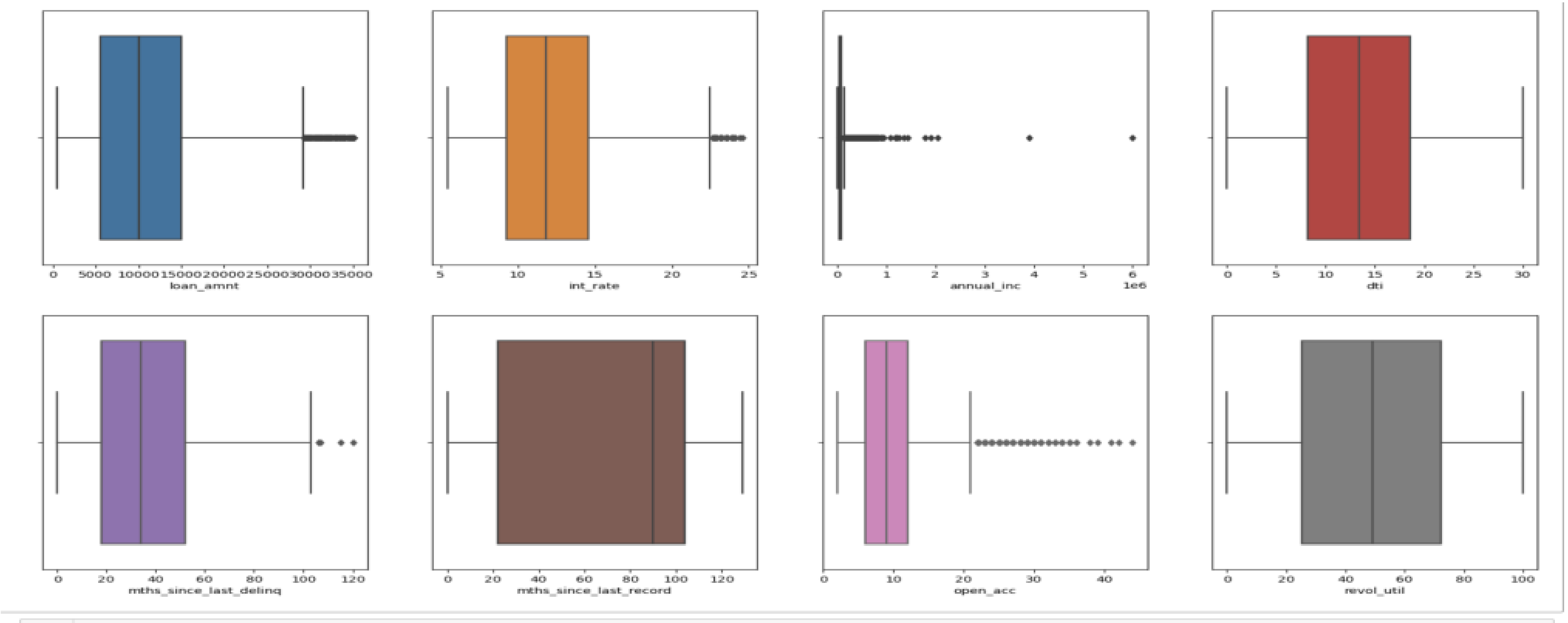
The remaining columns after Data cleaning are divided into numerical and categorical variables.

```
catagorical_columns = [
    "term",
    "grade",
    "emp_length",
    "home_ownership",
    "verification_status",
    "purpose",
    "addr_state",
    "delinq_2yrs",
    "inq_last_6mths",
    "pub_rec",
    "loan_status",
    "pub_rec_bankruptcies"
    ]
```

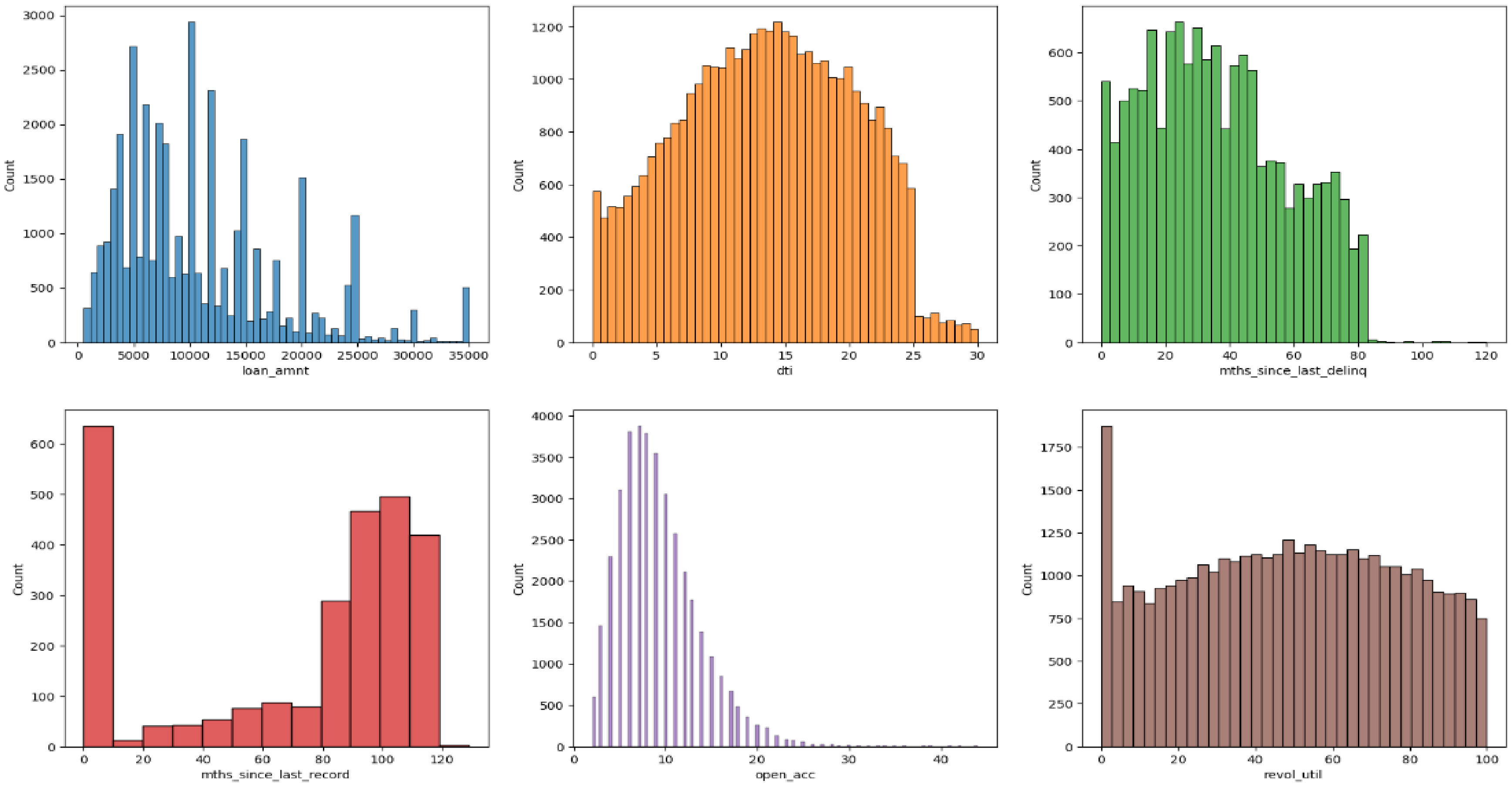
```
numerical_columns = [
    "loan_amnt",
    "int_rate",
    "annual_inc",
    "dti",
    "mths_since_last_delinq",
    "mths_since_last_record",
    "open_acc",
    "revol_util",
  ]
```

Outliers Treatment

Outliers in columns like annual_inc are removed as they mislead the analysis results.

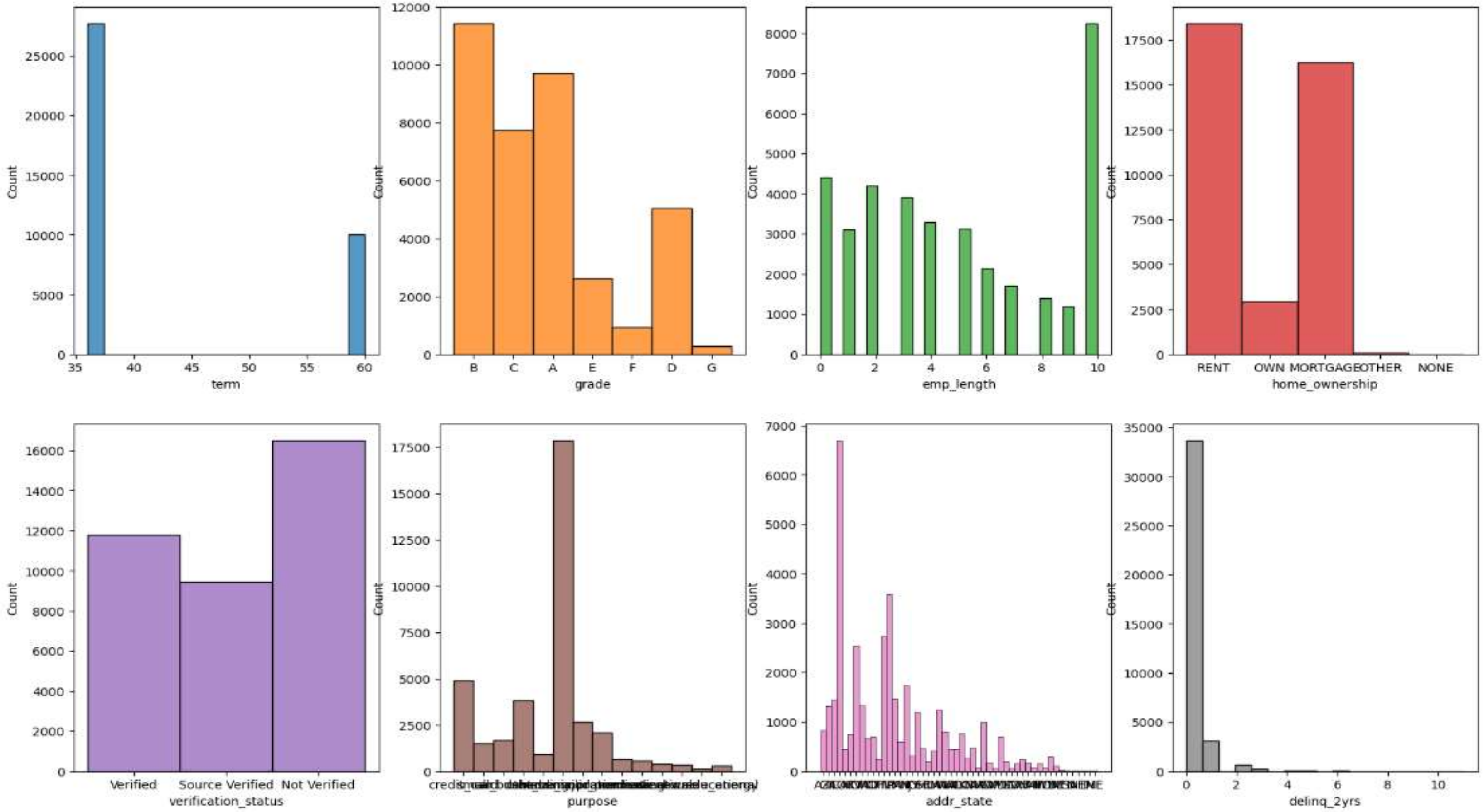


Univariate Analysis - Numerical Columns



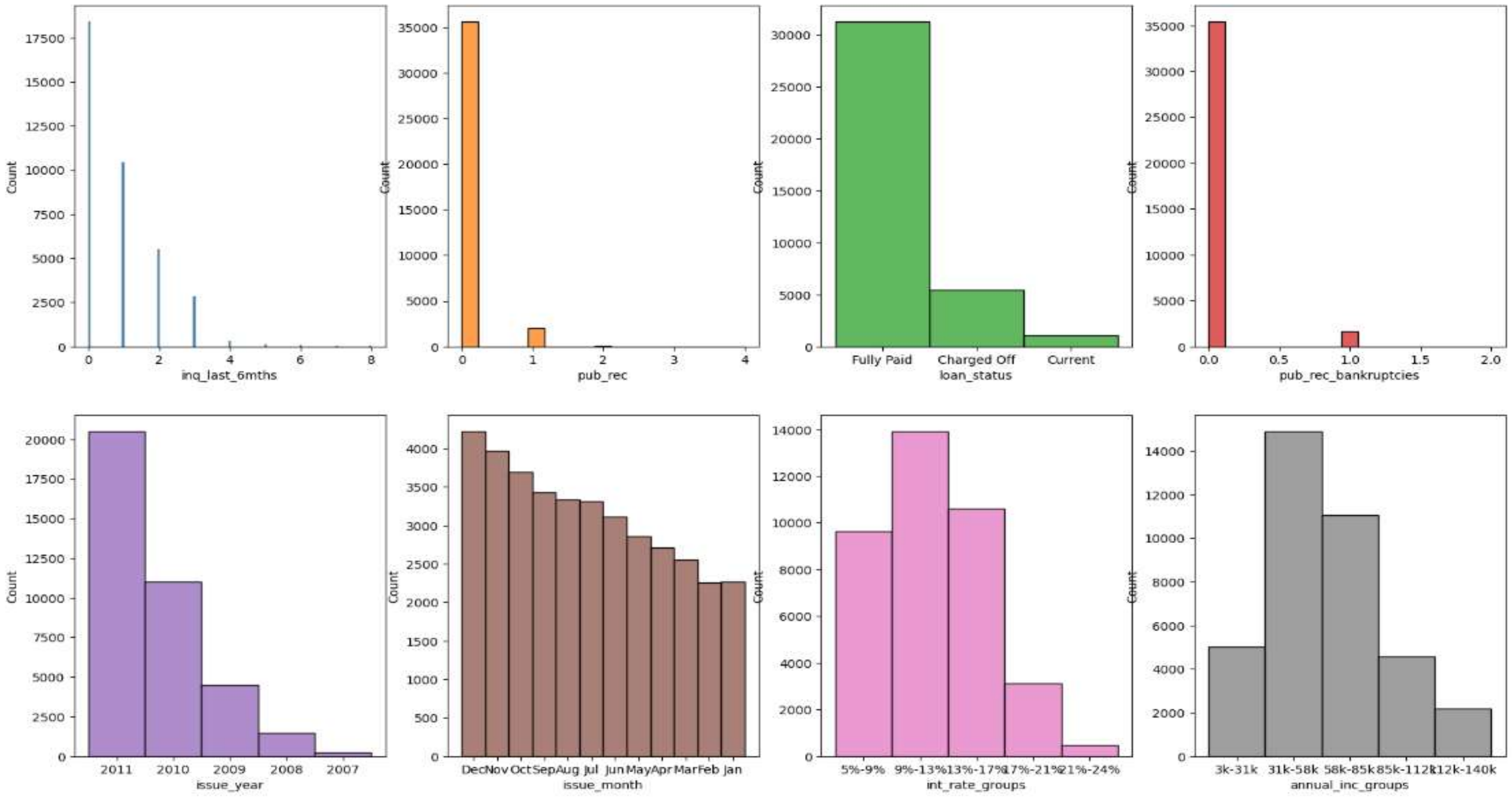
These graphs depict the univariate analysis of numerical columns for the entire dataset.

Univariate Analysis - Categorical Columns



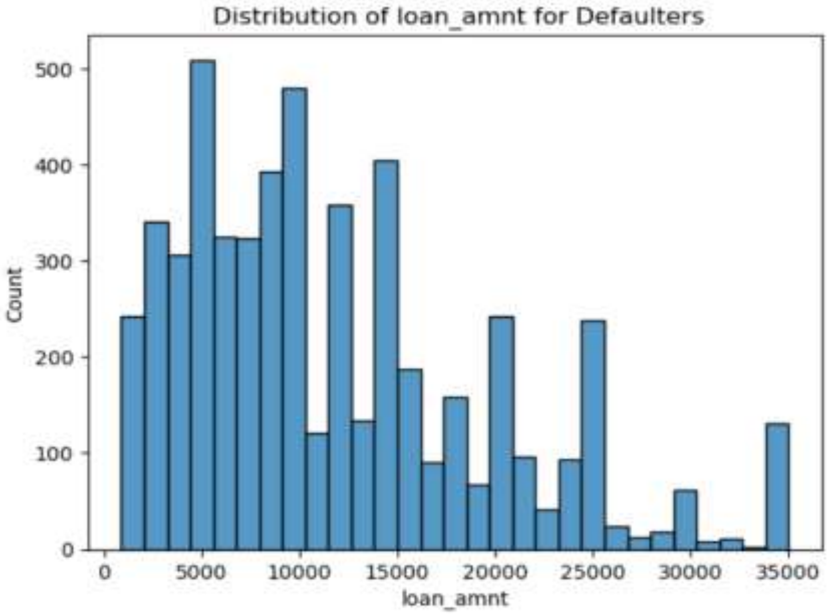
These graphs depict the univariate analysis of catagorical columns for the entire dataset.

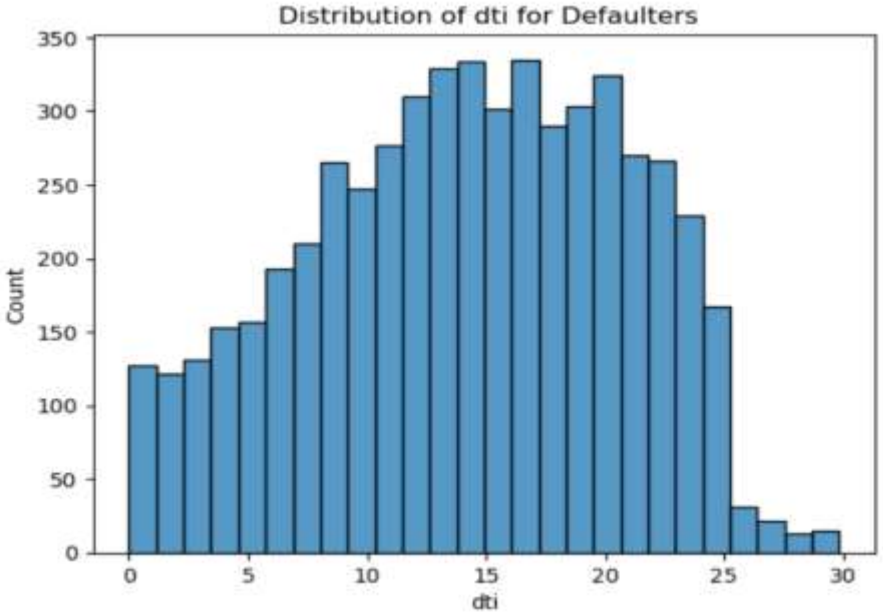
Univariate Analysis - Categorical Columns



These graphs depict the univariate analysis of categorical columns for the entire dataset.

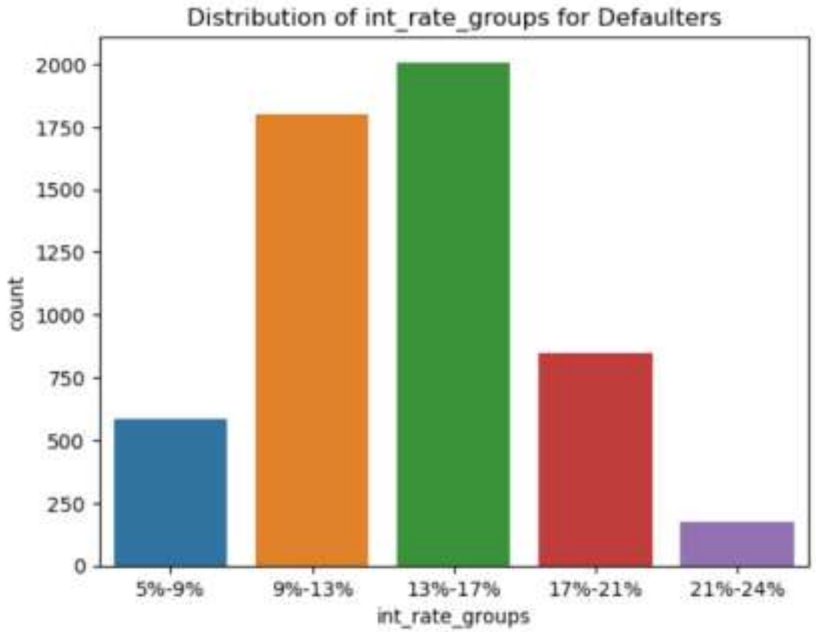
Univariate Data Analysis

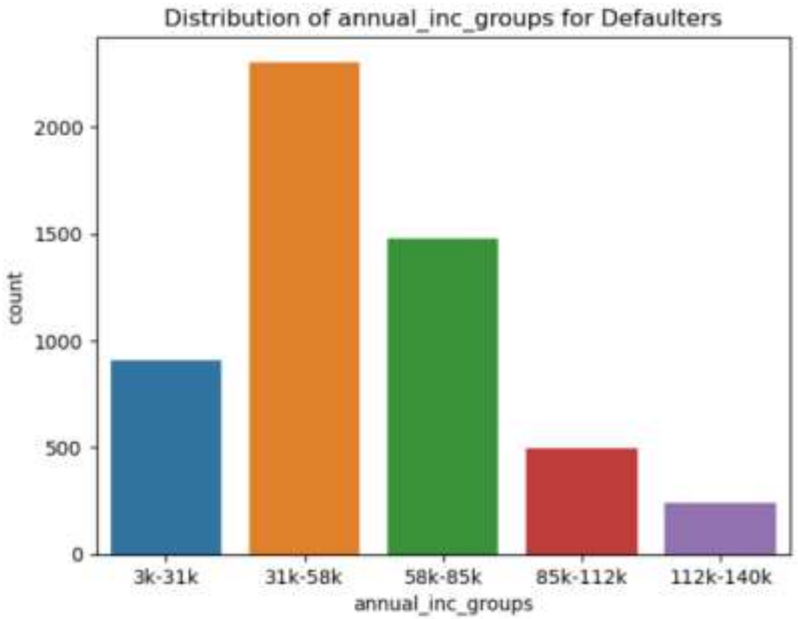




- As the loan amount increases, there is less count of defaulters
- The debt to income ratio for defaulters is at it peak between 10 to 25%

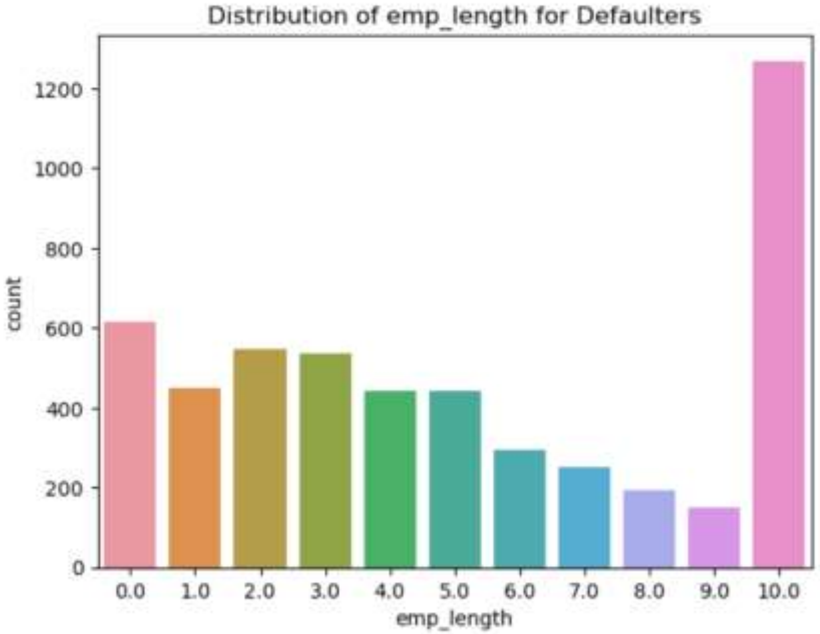
Univariate Data Analysis contd…

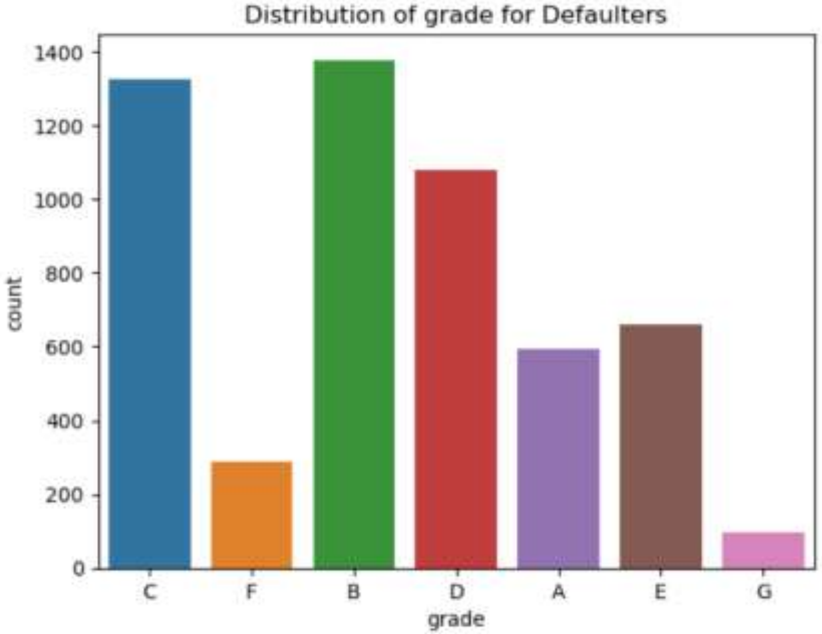




- Most of the defaulters are the people whose loan interest rate is between 13 to 17%
- The self reported annual income for most of the defaulters lies between 31k to 58k

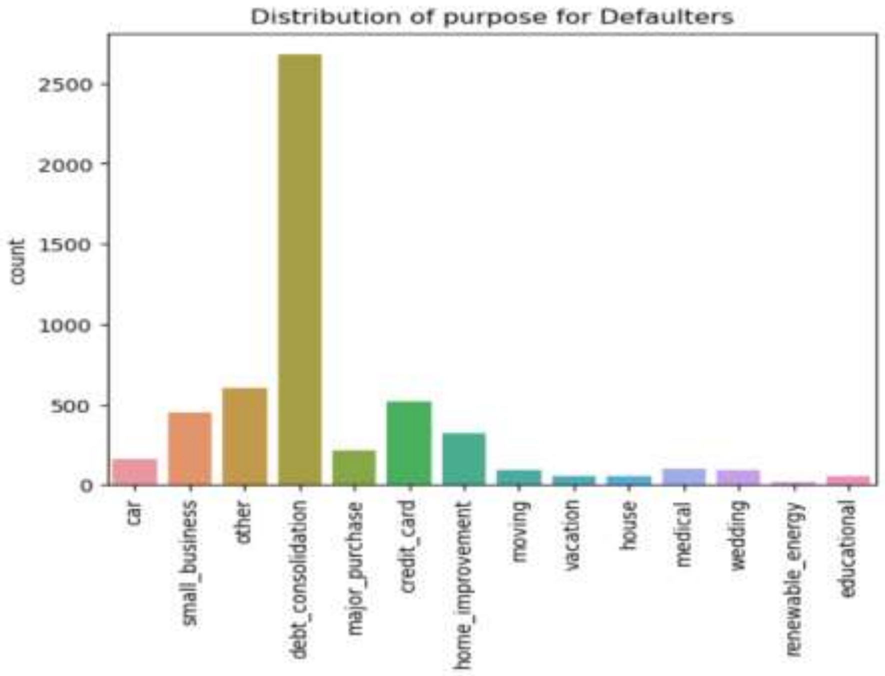
Univariate Data Analysis contd…

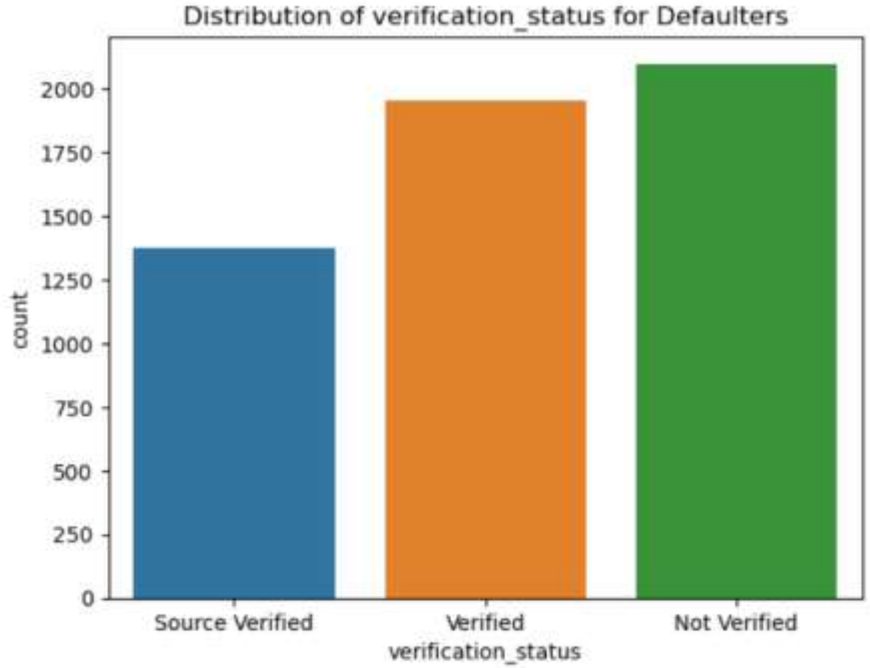




- More defaulters are having the employment experience more than 10 years
- Applicants with grade B and C are more defaulters compared to applicants of other grades.

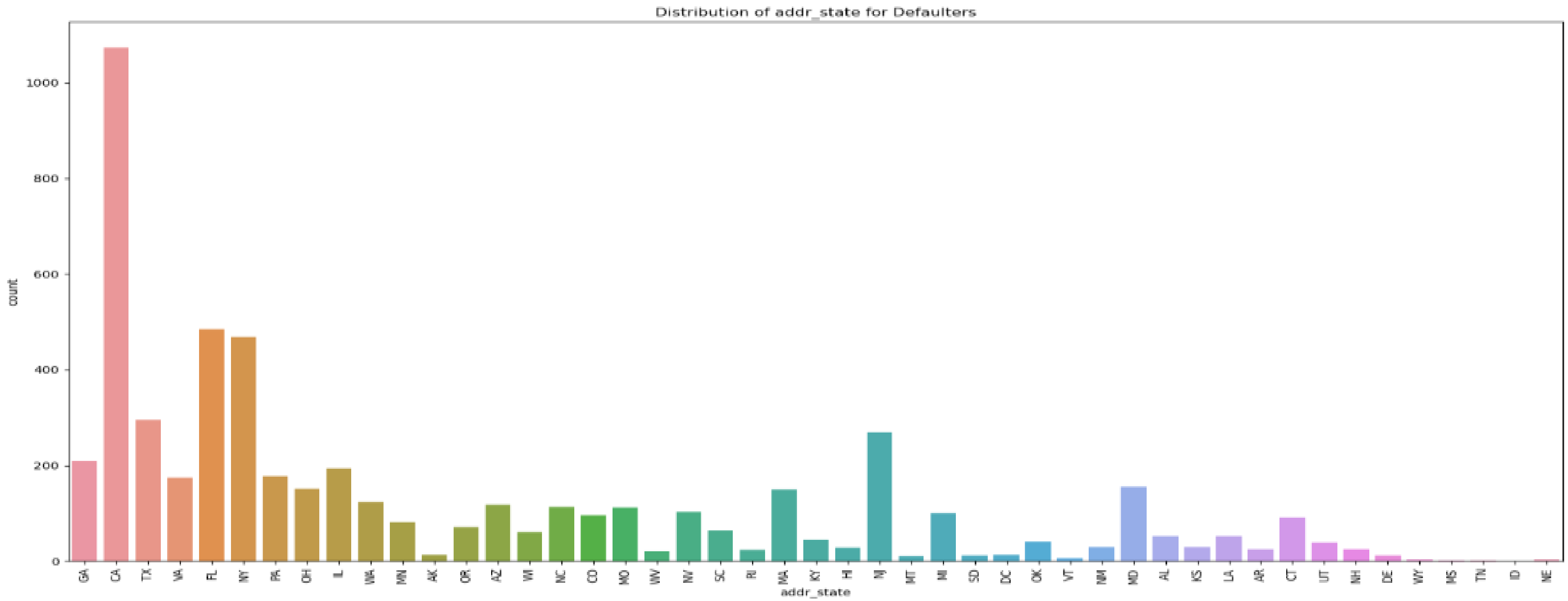
Univariate Data Analysis contd…





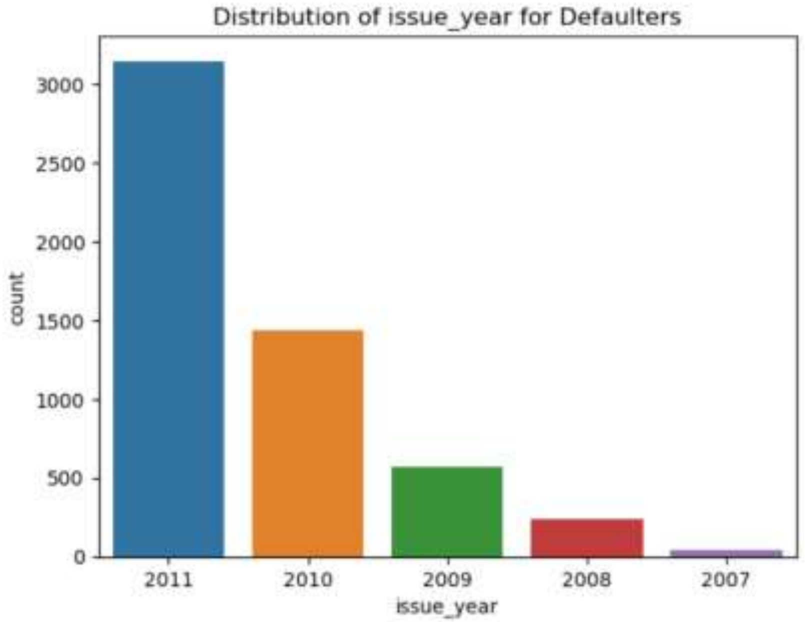
- Applicants who has taken loans for debt_consolidation purpose are more prone to be defaulters.
- Not Verified applicants are more prone to be defaulters.

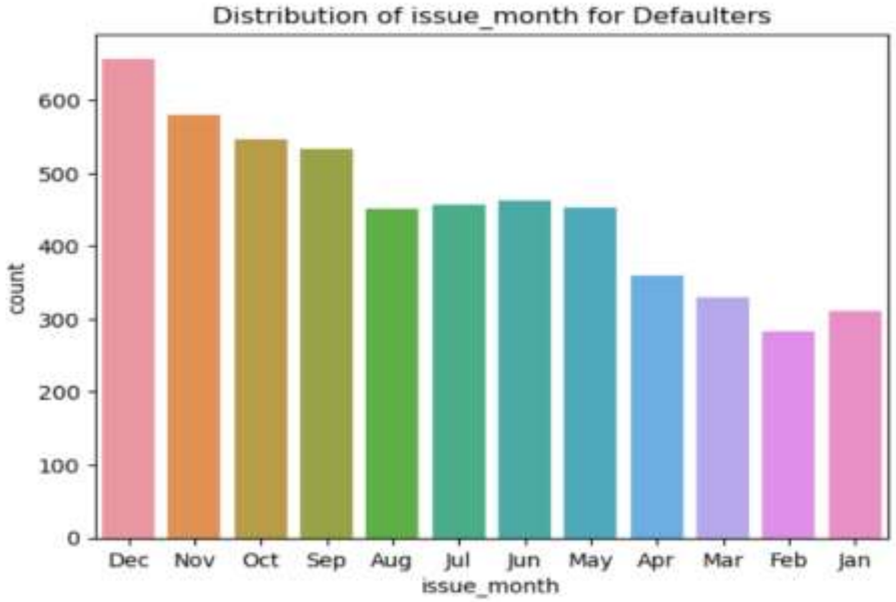
Univariate Data Analysis contd…



- Applicants from states CA, FL and NY are more defaulters compared to other states.

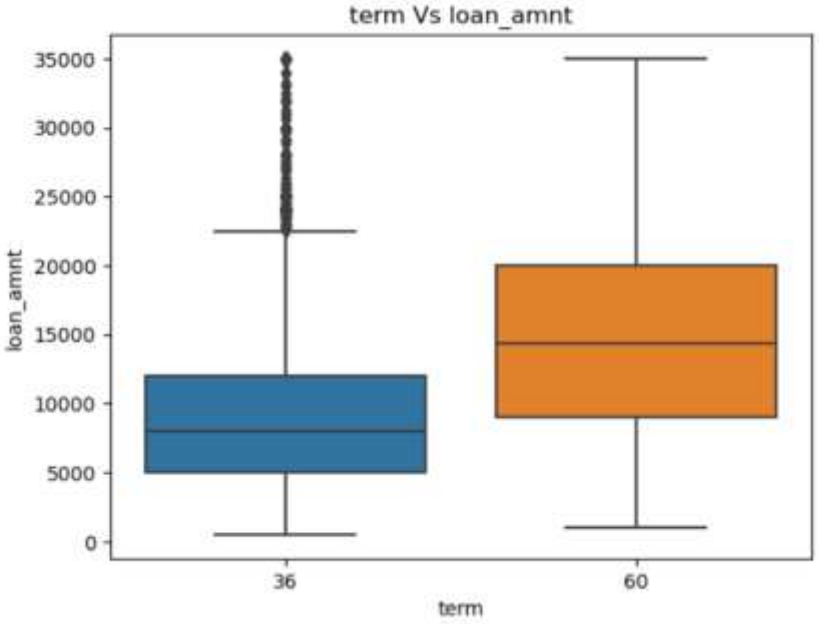
Univariate Data Analysis contd…

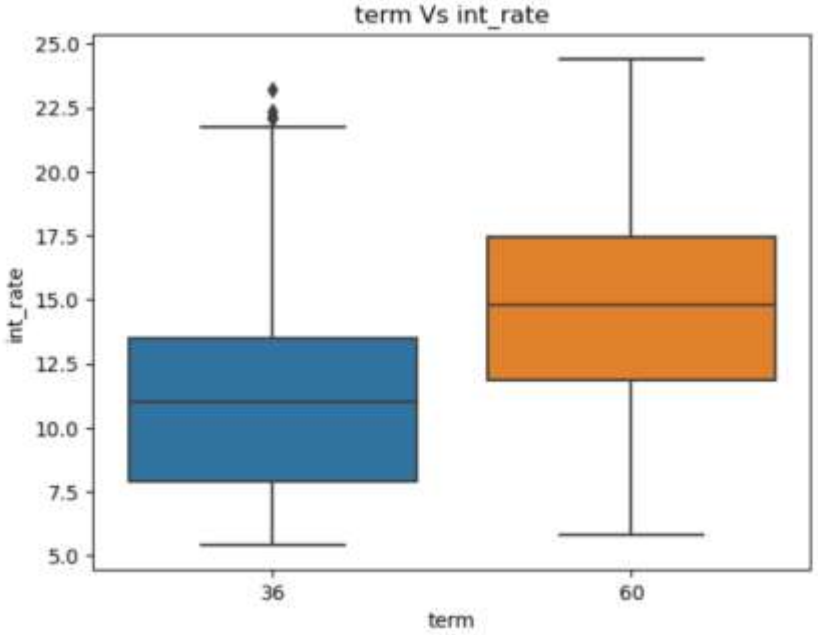




- 2011 has the highest defaulters.
- Applicants who has taken loan towards the end of the year are more prone to be defaulters.

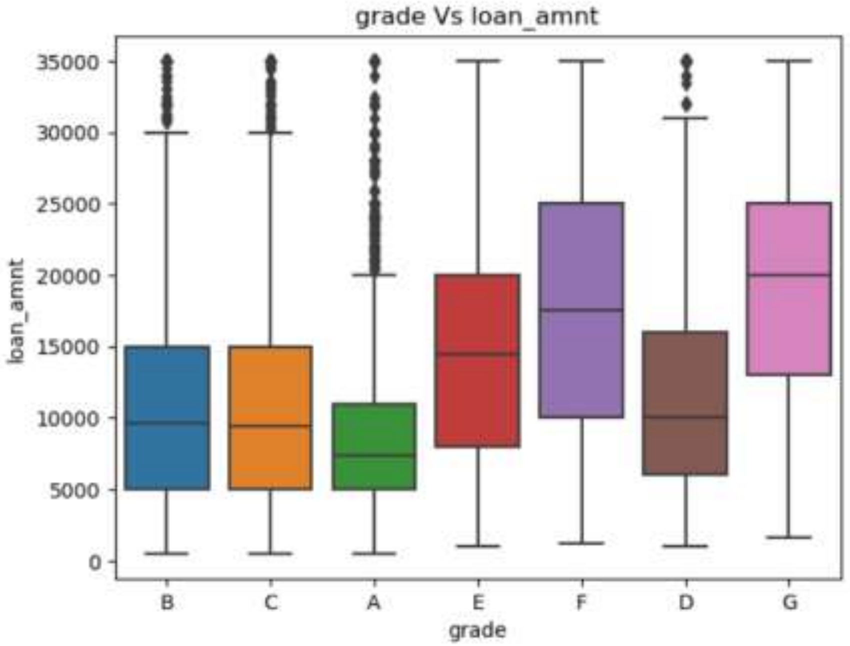
Bivariate Analysis

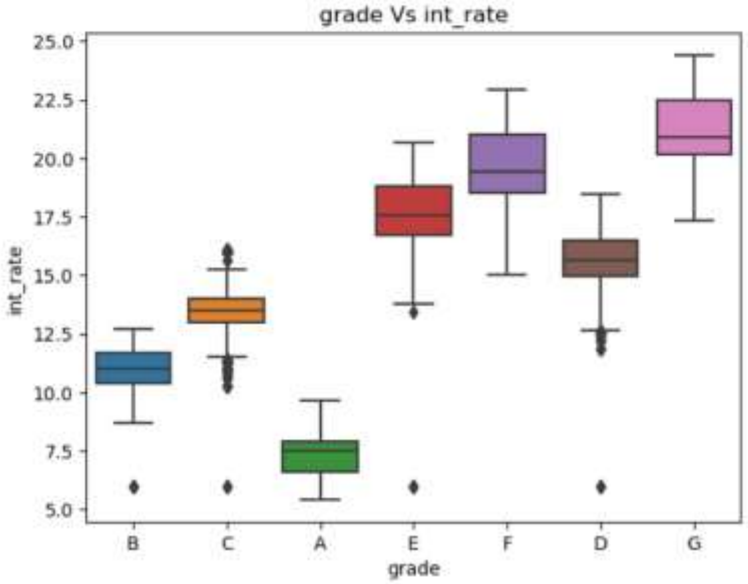




- The applicants who opt for higher term are provided more loan amount at the same time they are charged with higher interest rates.

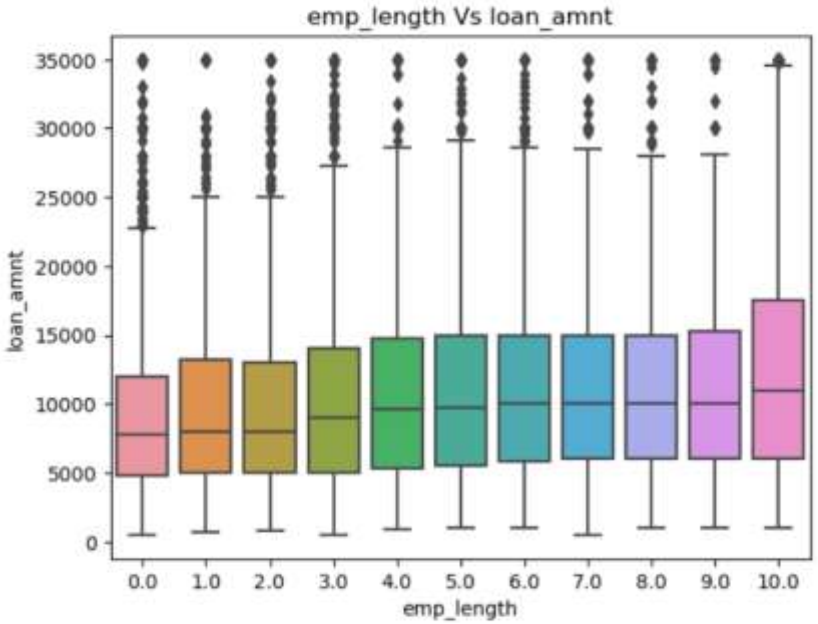
Bivariate Analysis contd…

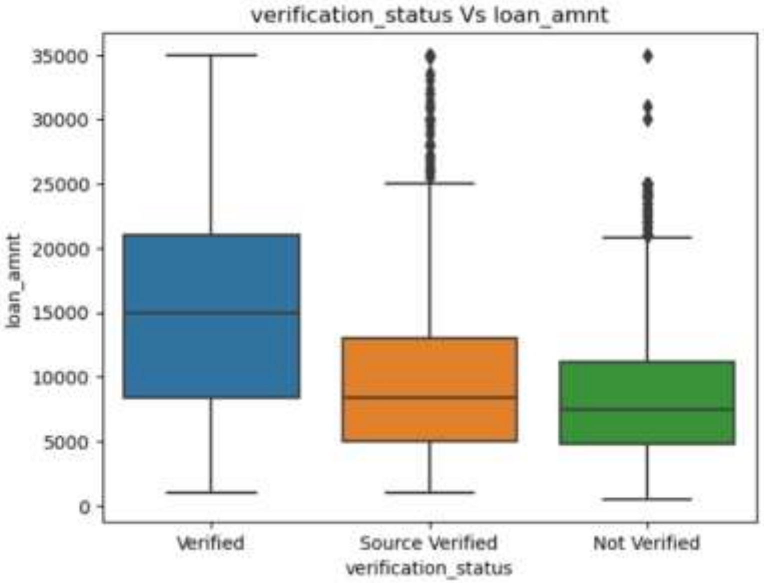




- The applicants who are of Grade ‘F’ and ‘G’ are provided higher loan amount with higher interest rates.

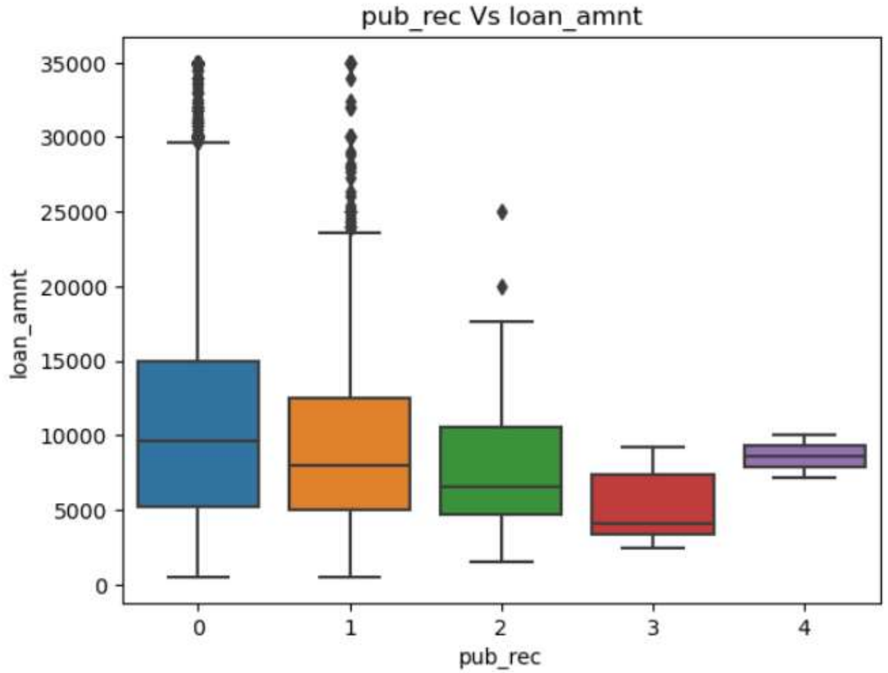
Bivariate Analysis contd…

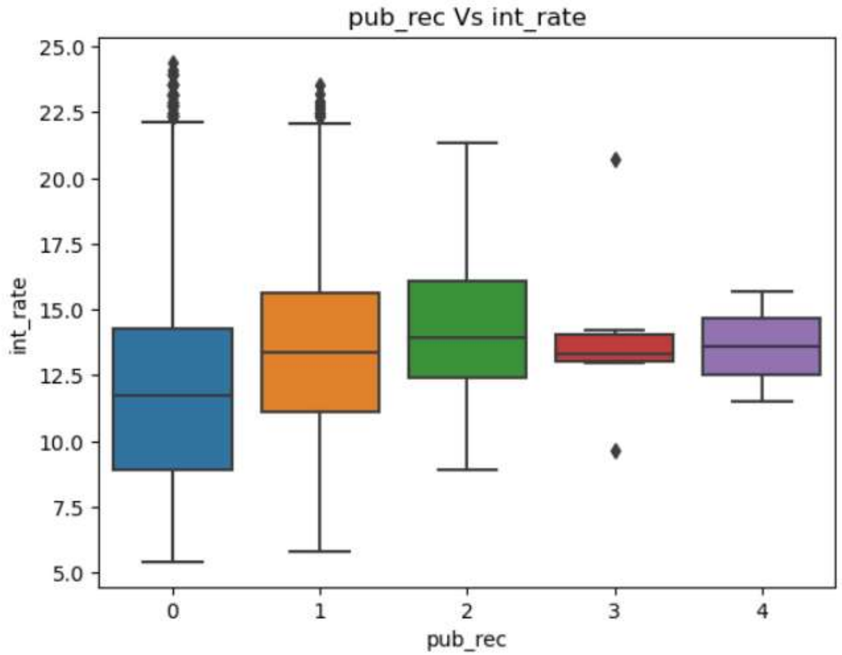




- The applicants who has 10 Plus years experience are offered higher loan amount.
- Higher loan amount is provided with verification.

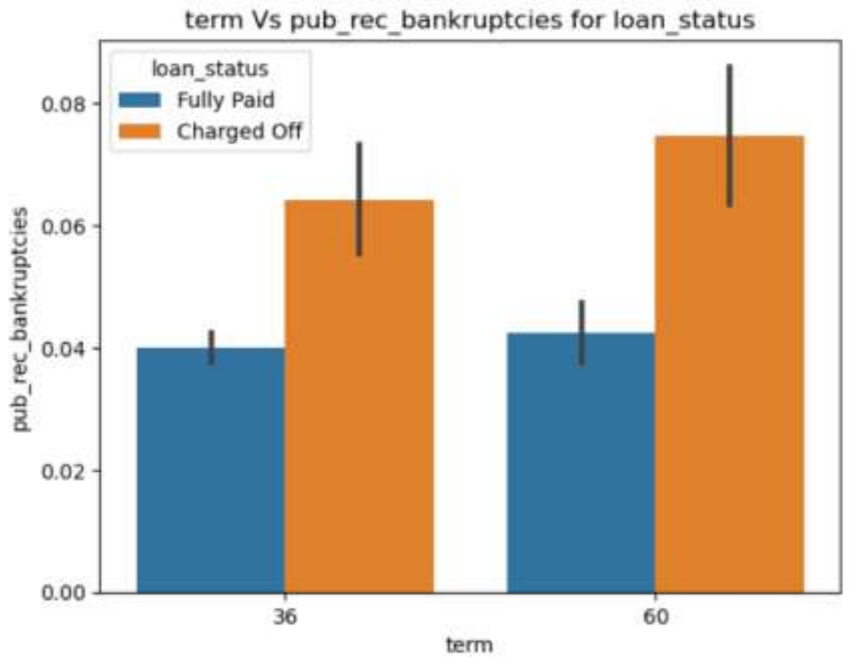
Bivariate Analysis contd…

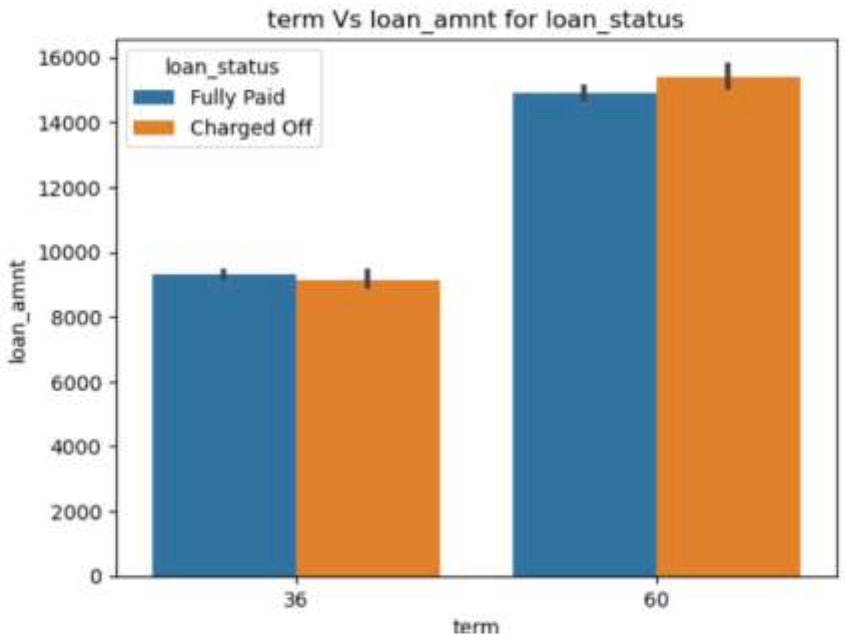




- The applicants who does not have any derogatory public records are provided more loan amount with less interest.

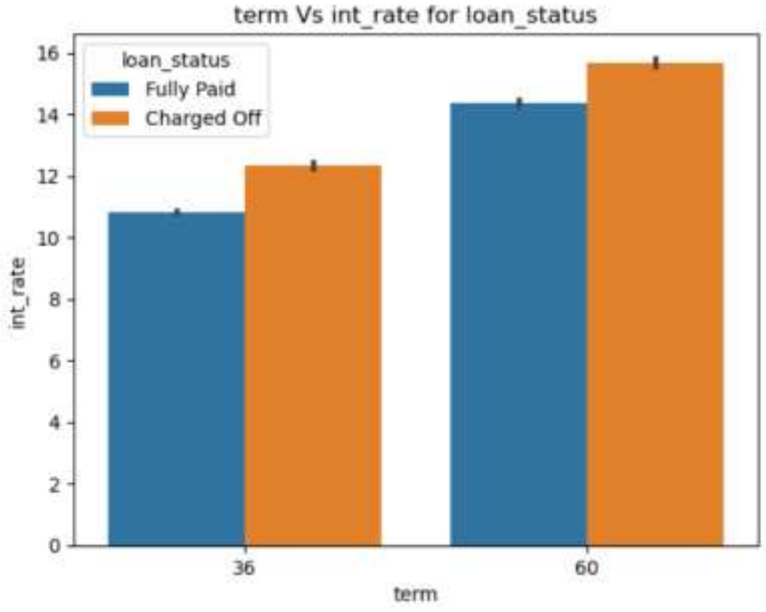
Loan Status Vs other columns

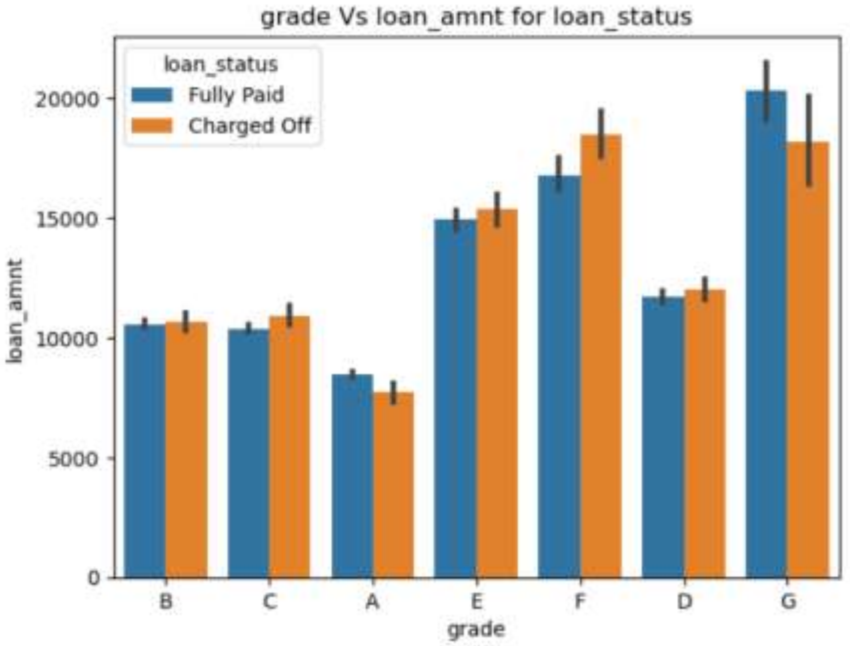




- The applicants who defaulted for any term has more public recorded bankruptcies compared to the fully paid applicants.
- Applicants who applied and defaulted have no significant difference in loan amount which means the applicants applying for long term has applied for more loan.

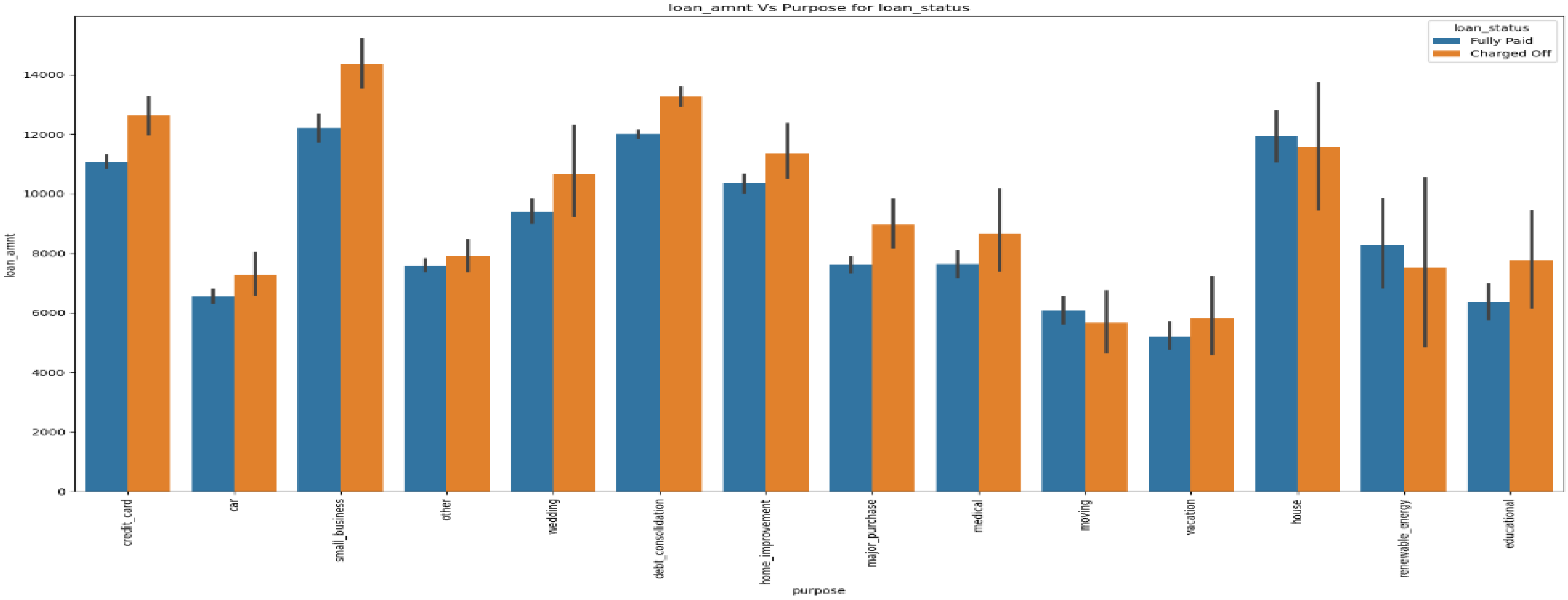
Loan Status Vs other columns





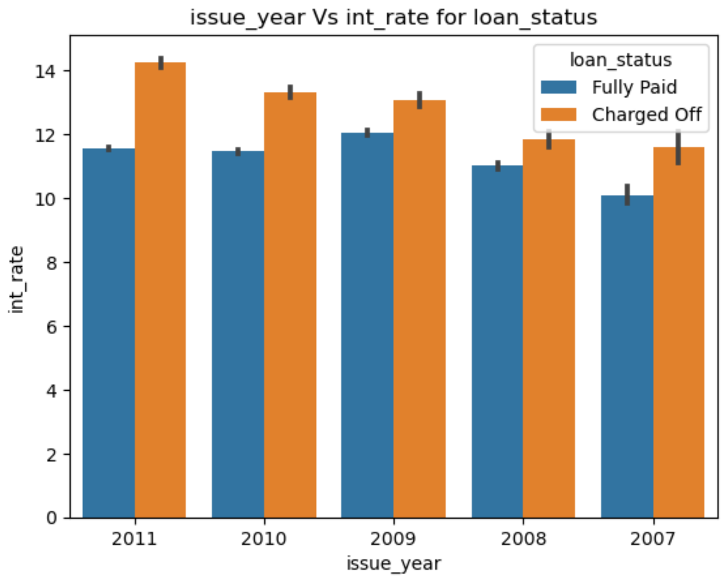
- The more the term of loan the more is the interest rate and the more are the defaulters.
- Applicants of F grade has taken more loan amount and are more defaulter compared to other grades.

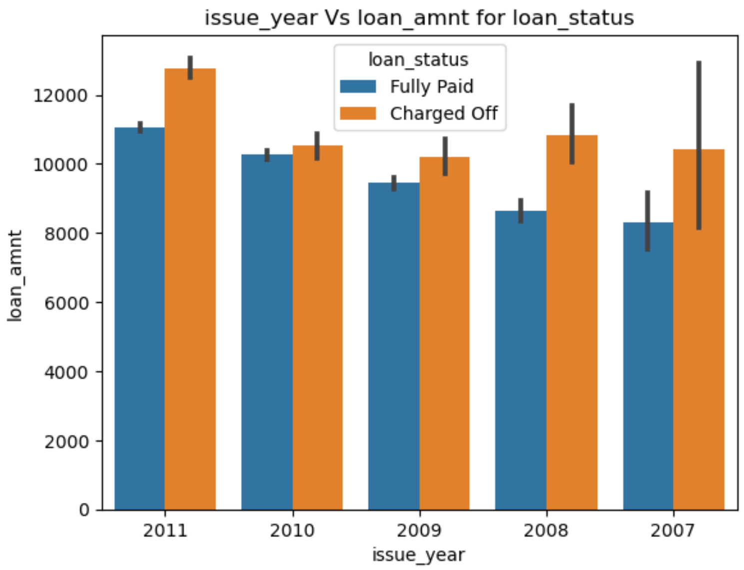
Loan Status Vs other columns



- The applicants to has taken loans for small business purpose defaulted the most compared to other applicants

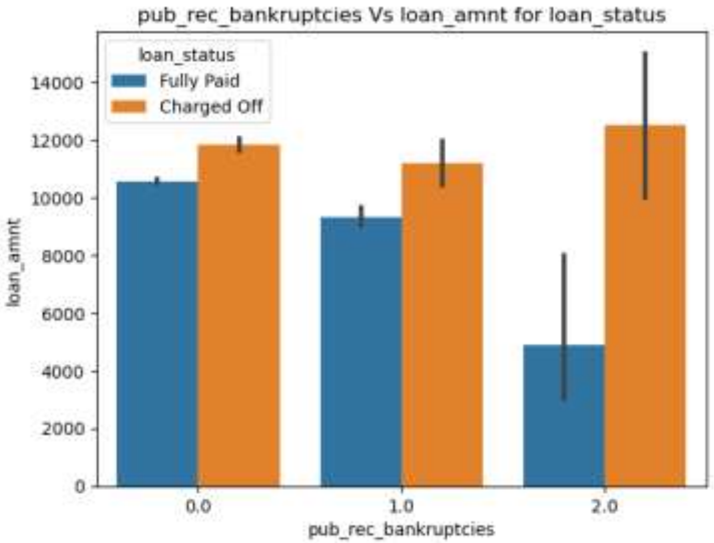
Loan Status Vs other columns

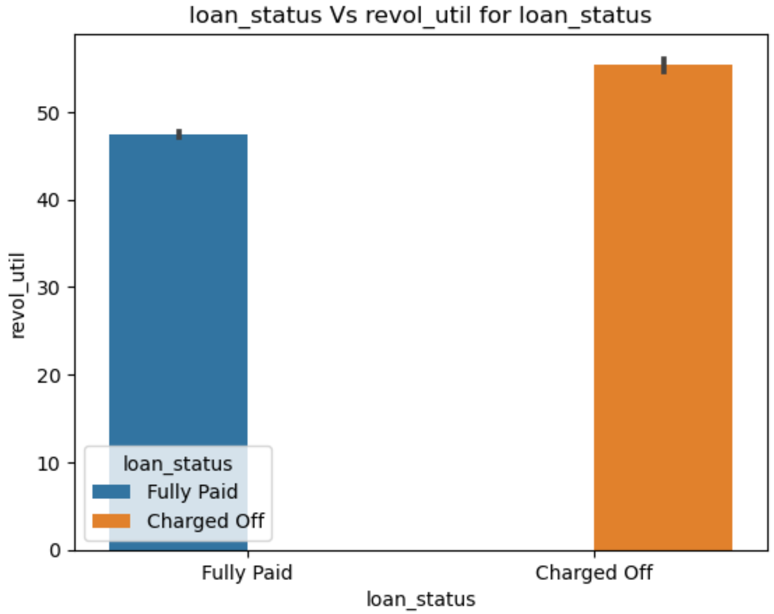




- Due to the US financial crisis in 2011, higher loan amounts with higher interest rates are disbursed. Hence more defaulters are observed in case the country is facing any abnormal issues.

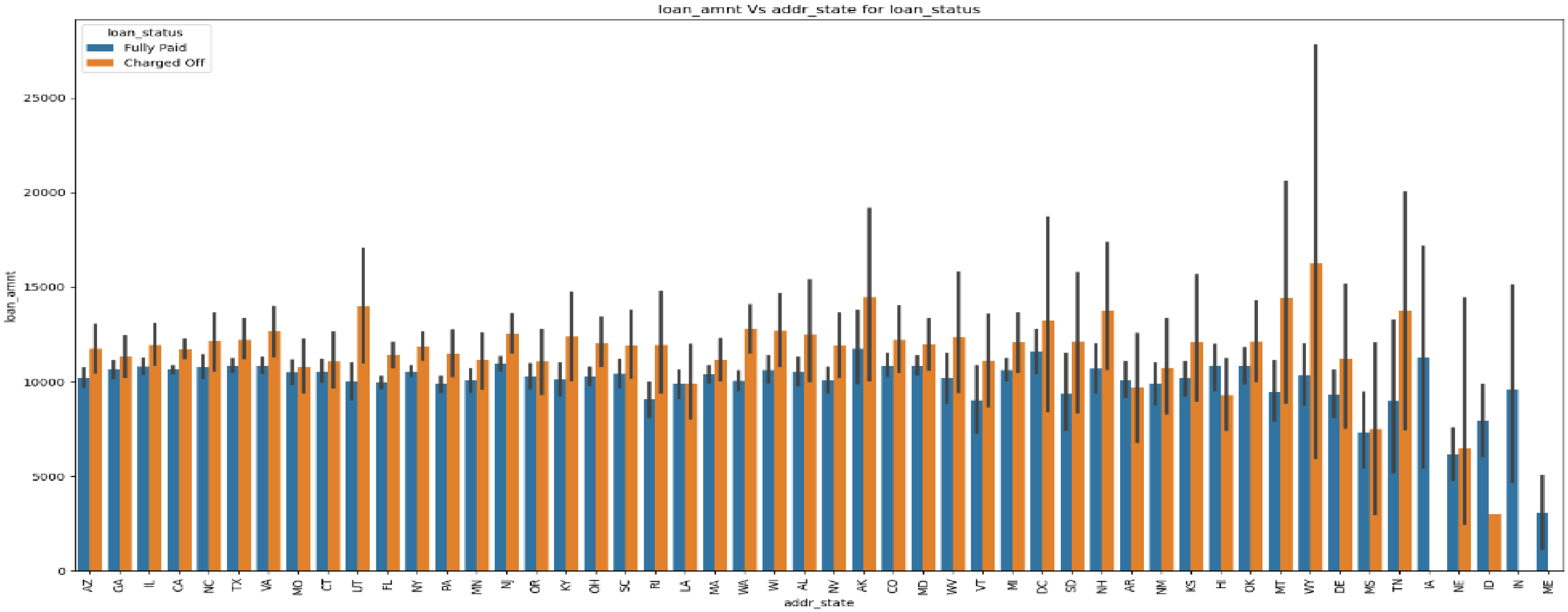
Loan Status Vs other columns





- Applicants who has more bankruptcies are more prone to be defaulters.
- Applicants who has more revolving line utilization rate are more prone to be defaulters.

Loan Status Vs other columns



Among the applicants of different states, WY state people have take more loan amount and has more defaulters

Conclusion

The key attributes leading to applicants becoming defaulters are :

1. Applicants of F grade has taken more loan amount and are more defaulter compared to other grades. (Ref slide 26)
2. Loans taken for purpose Small business are more prone to be defaulters. (Ref slide 27)
3. If there are any natural or financial Crices there could be more defaulters among the people taking loan in this duration (Ref slide 28)
4. The self reported annual income for most of the defaulters lies between 31k to 58k. (Ref slide 16)
5. Applicants who have higher public derogatory records have chances to default but they are provided loans with higher interest rates.

Key Attributes:

1. Loan Purpose
2. Employee Grade
3. Annual Income
4. Public Recorded bankruptcies