By:

- Nishita K Murthy
- Hemja Singhal

Problem Definition

The aim is to identify patterns which indicate if a person is likely to default (quits repaying the loan borrowed), 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.

To achieve this goal, we will be using EDA (Exploratory Data Analysis) which would involve understanding the dataset, collecting and cleaning of data, analyzing the data with multiple graphs/plots and finally providing a few observations and recommendations to the club.

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

- **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

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)

Cleaning and standardize data

CLEANING:

- Removed multiple missing value columns
- Eliminating columns with more than 90% null values
- Check for rows with missing values. Since not many rows have too many missing values, we have not eliminated any.

STANDARDIZE:

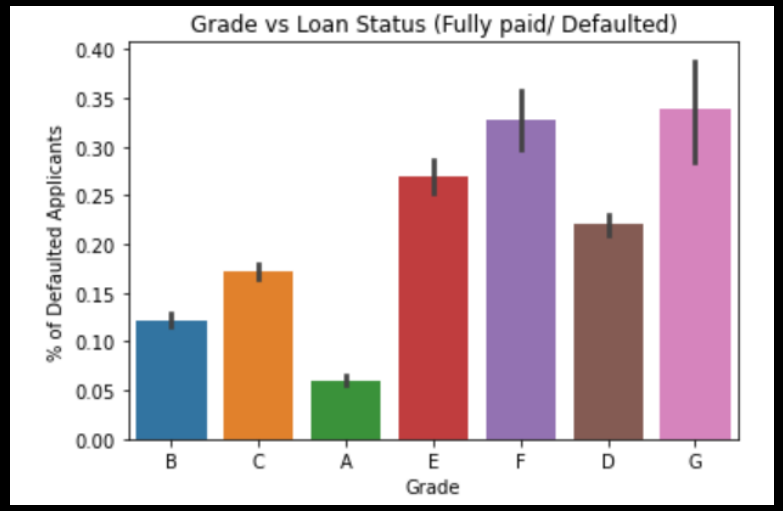
- Type casting the int_rate column from string to type float to get rid of the ‘%’ character.
- Filtering rows: eliminating the rows with ‘current’ as its loan_status value.

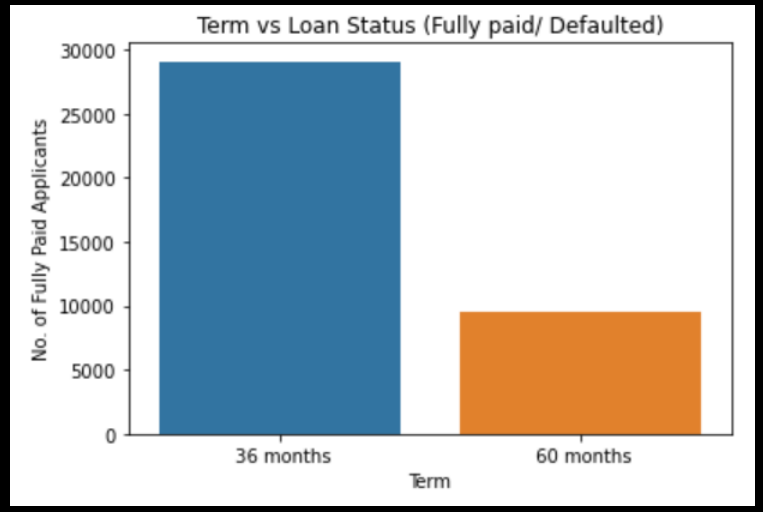
Univariate Analysis

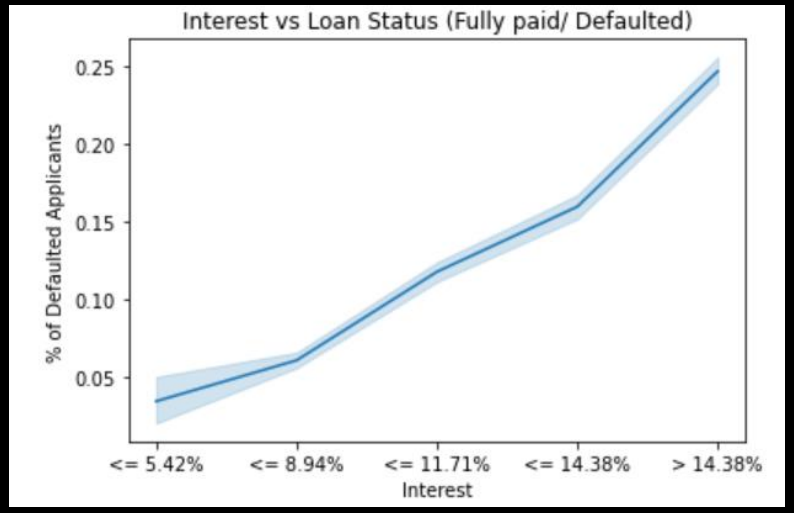
Below are the metrics used to perform the univariate analysis

- Grade vs loan_status
- Term vs loan_status
- Purpose vs % of defaulting applicants
- Interest rate vs % of defaulting applicants
- Verification status vs grade
- Home ownership vs % of defaulting applicants

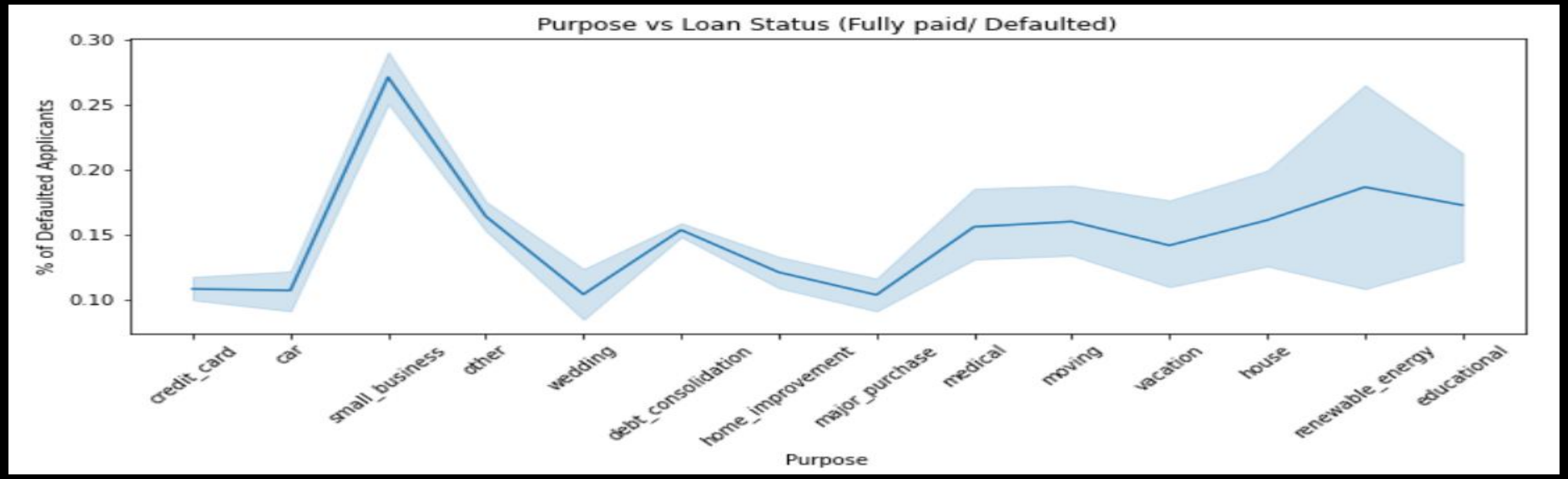
Univariate analysis corresponding graphs (cont..)

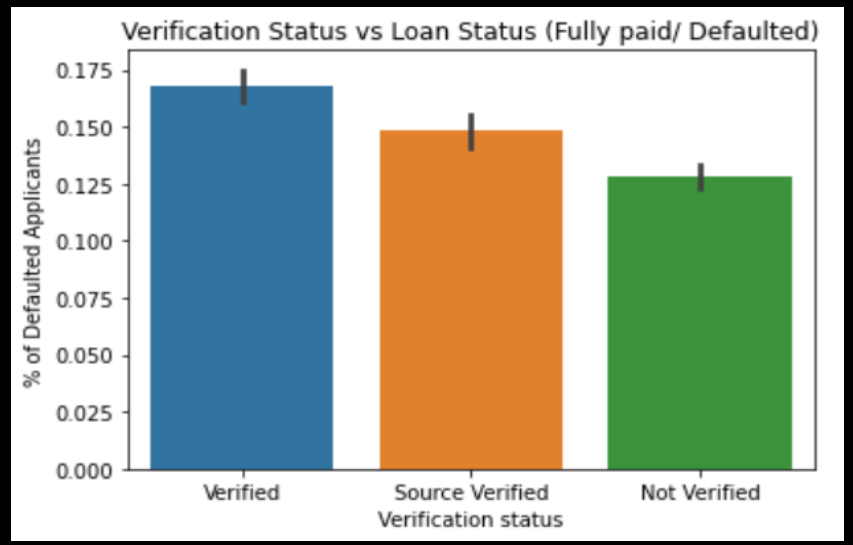


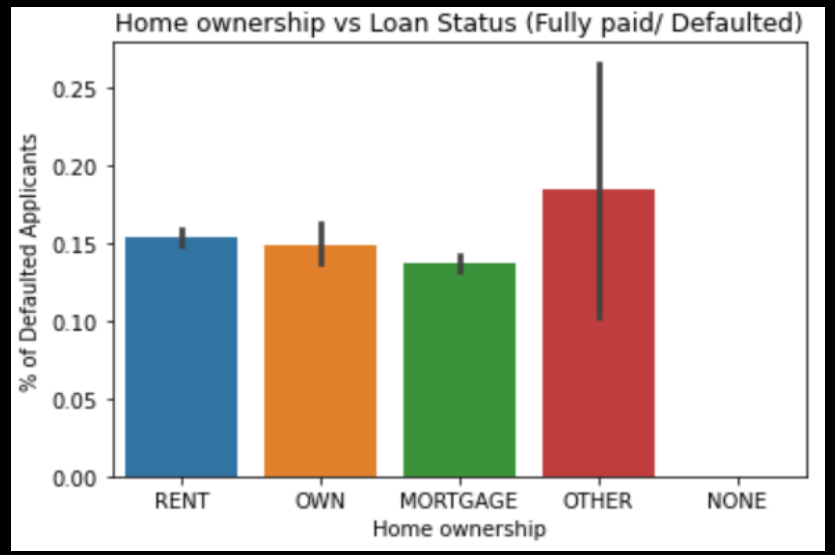




Univariate analysis corresponding graphs

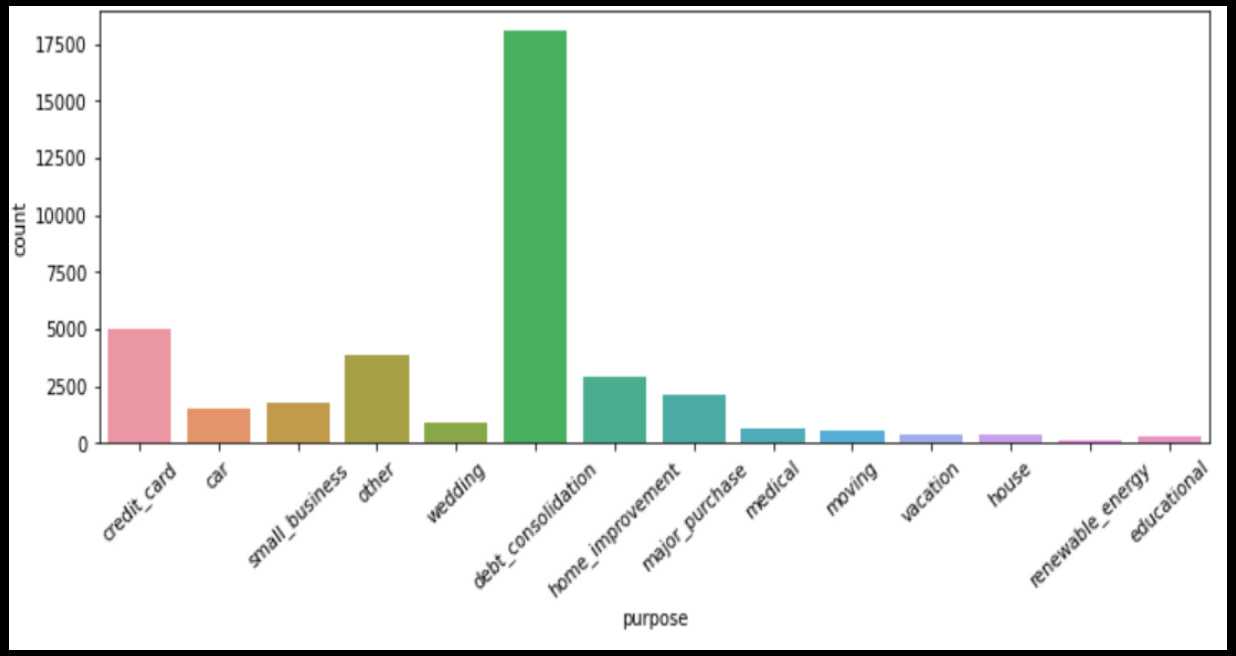


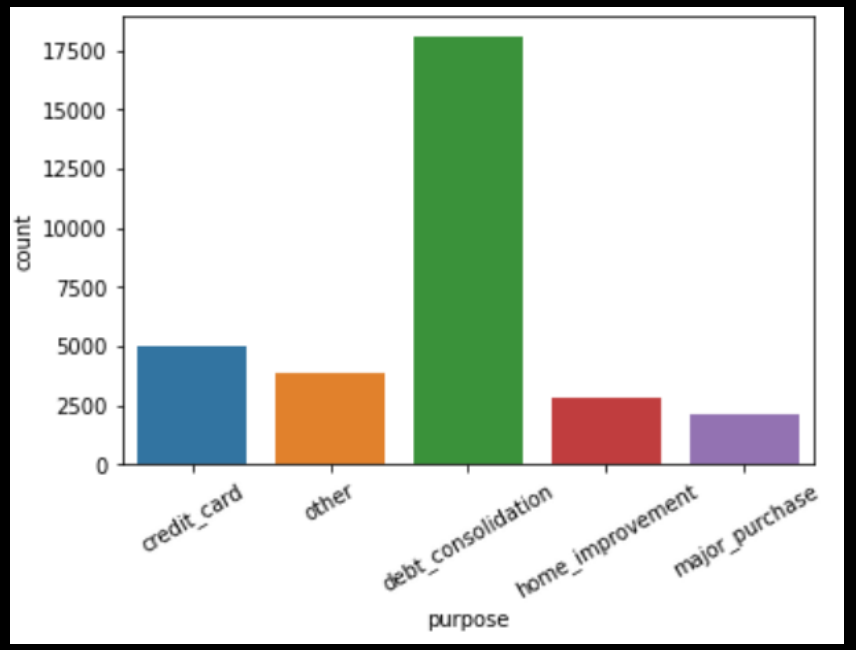




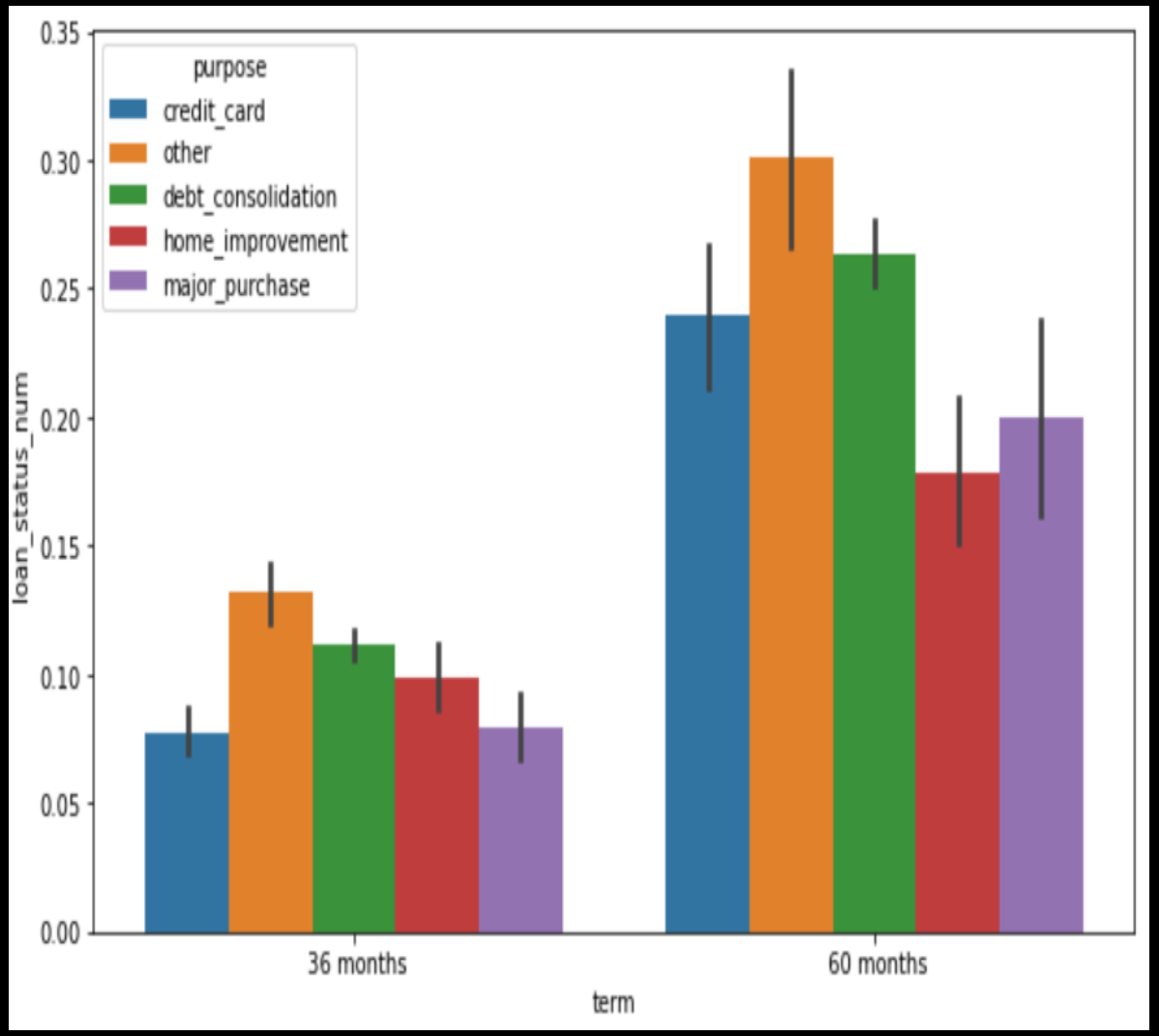
Segmented Univariate Analysis

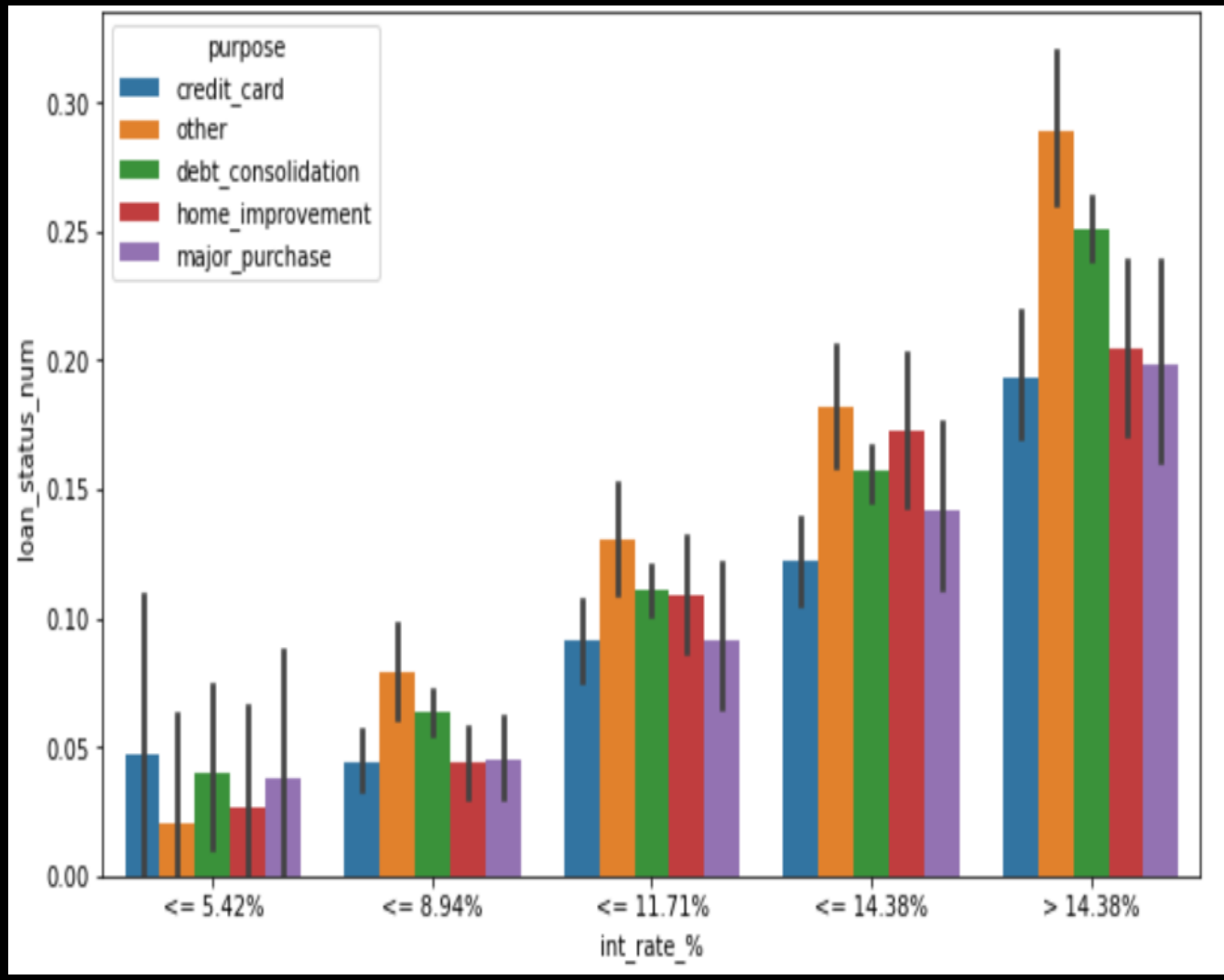
- Using the set of values under the purpose column against multiple variables was performed to understand various default rate reasons.
- Fig. 2 is the chosen segment of the purpose column (consisting of: debt_consolidation, credit_card, other, home_improvement and major_purchase)





Segmented Univariate Analysis Graphs



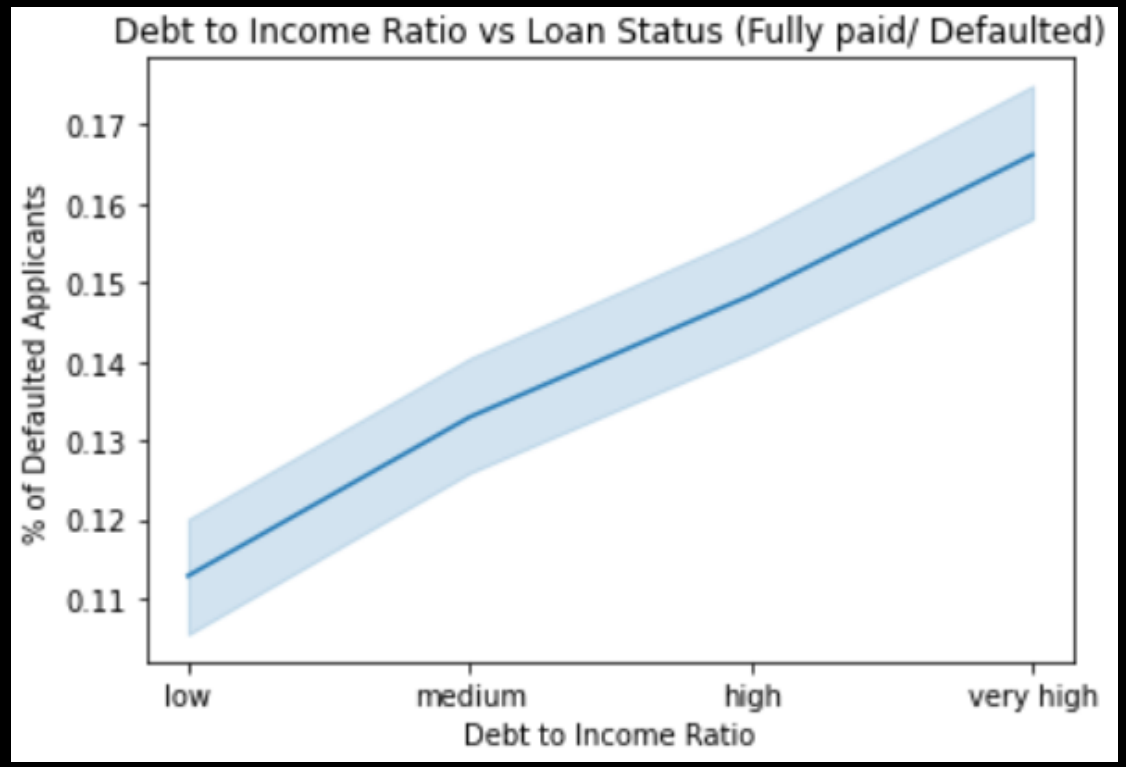


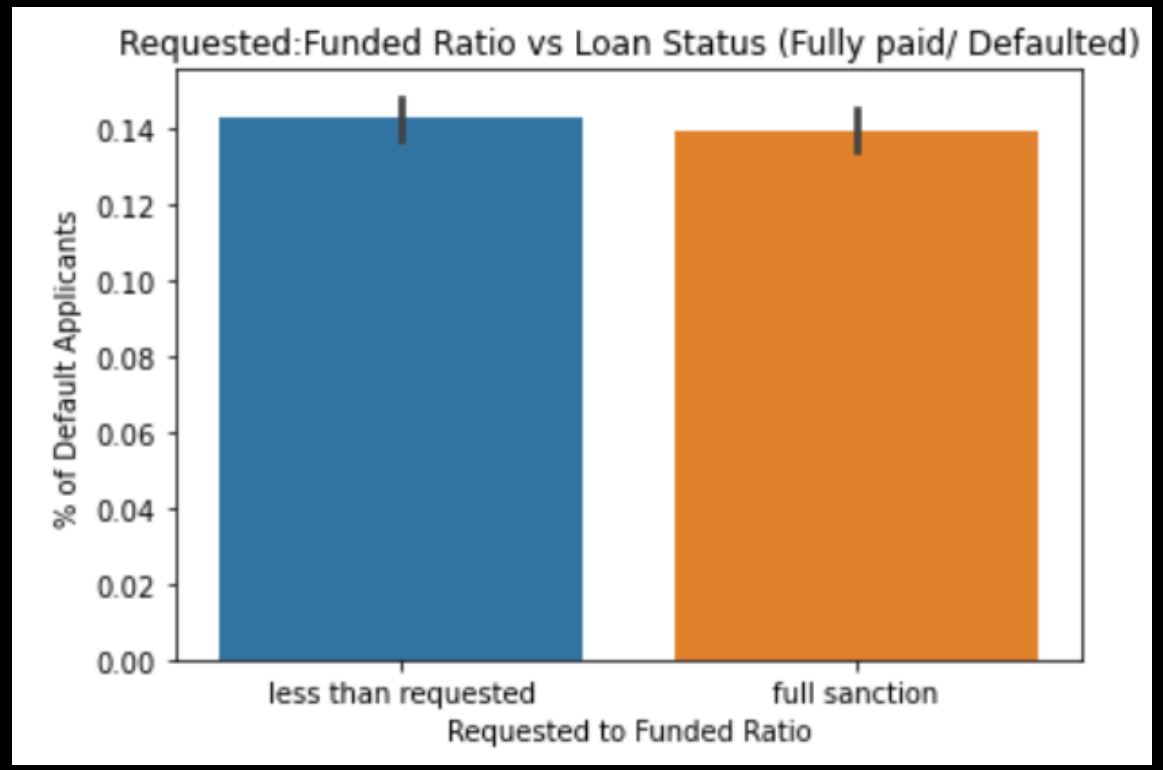
Bivariate Analysis

Below are the metrics used to perform bivariate analysis

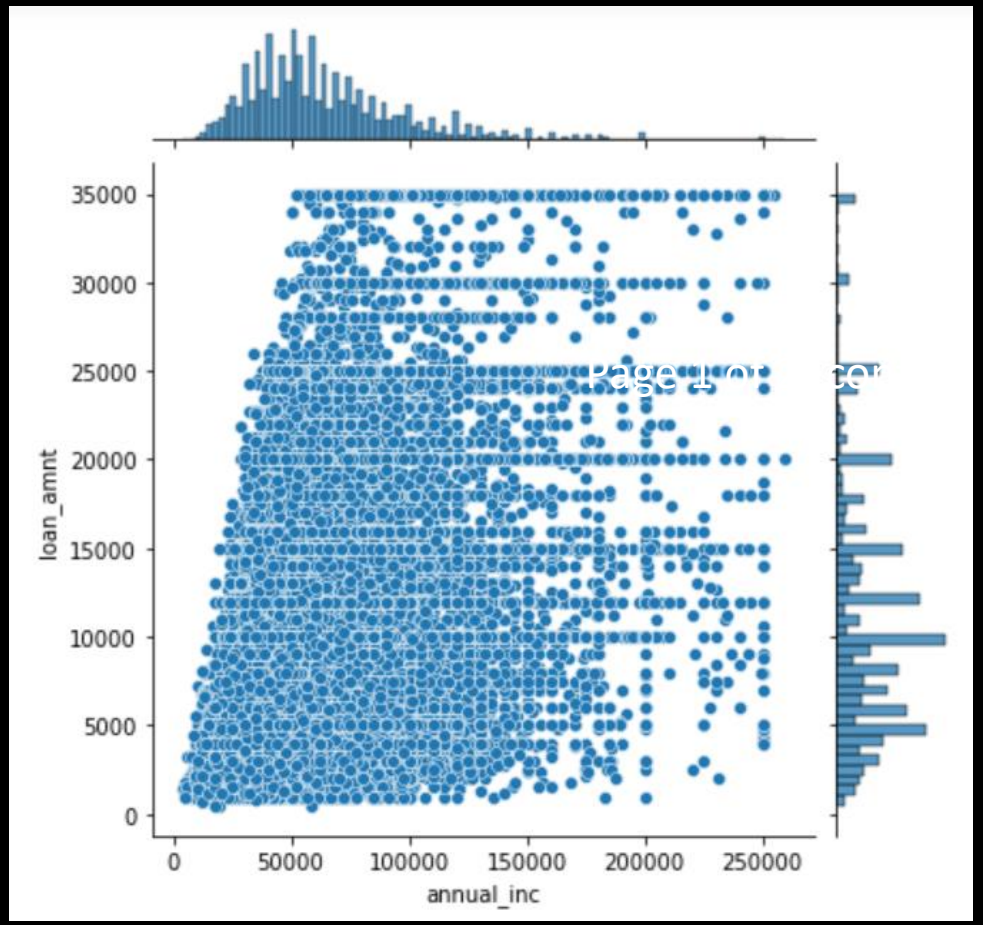
- Debt vs annual income ratio
- Requested amount vs funded amount
- Annual income vs loan amount
- Loan status vs verification status vs loan amount
- Loan amount vs state vs loan amount
- Interest rate vs delinquent in 2 years vs loan status

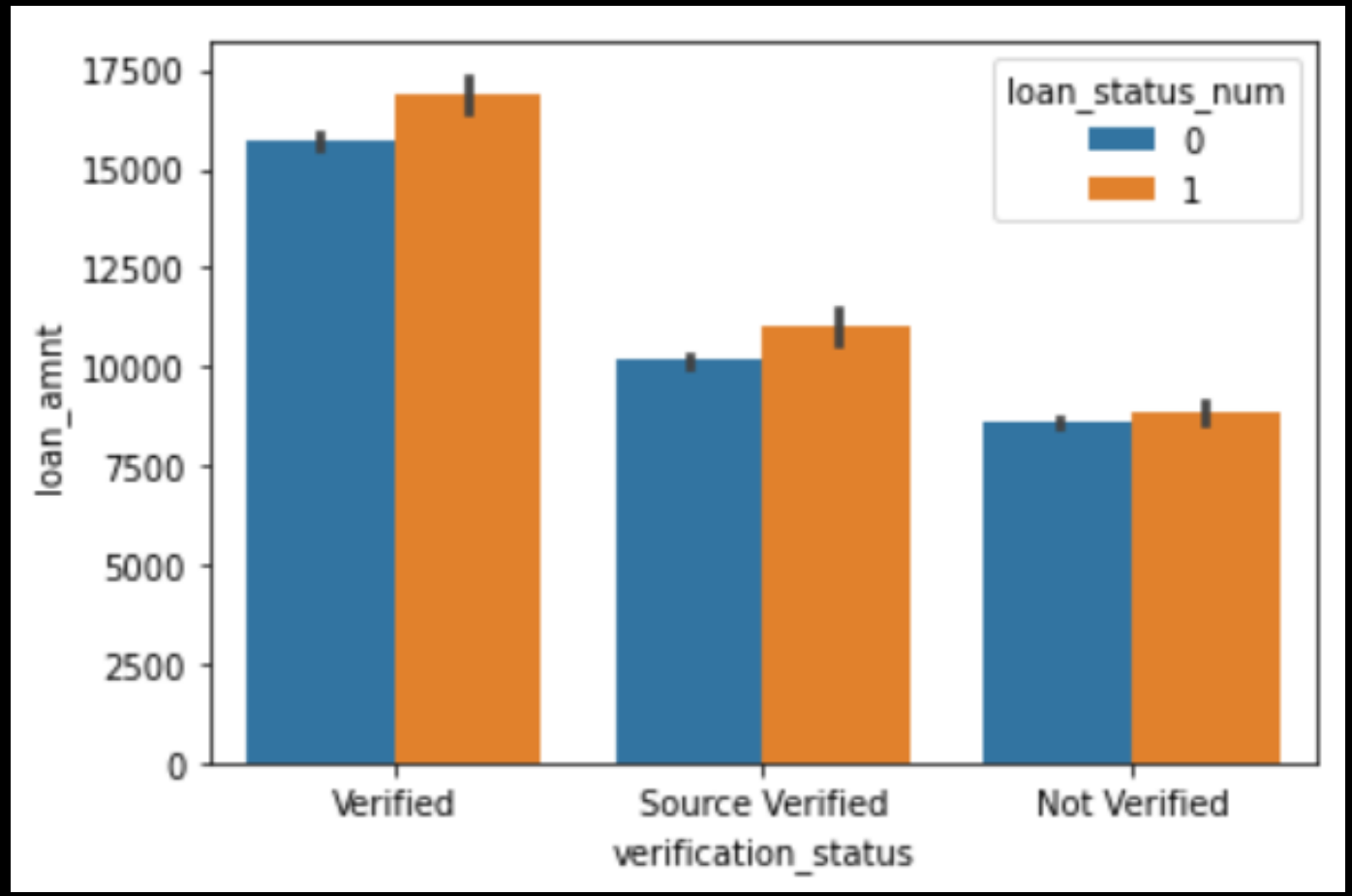
Bivariate Analysis Graphs (cont..)



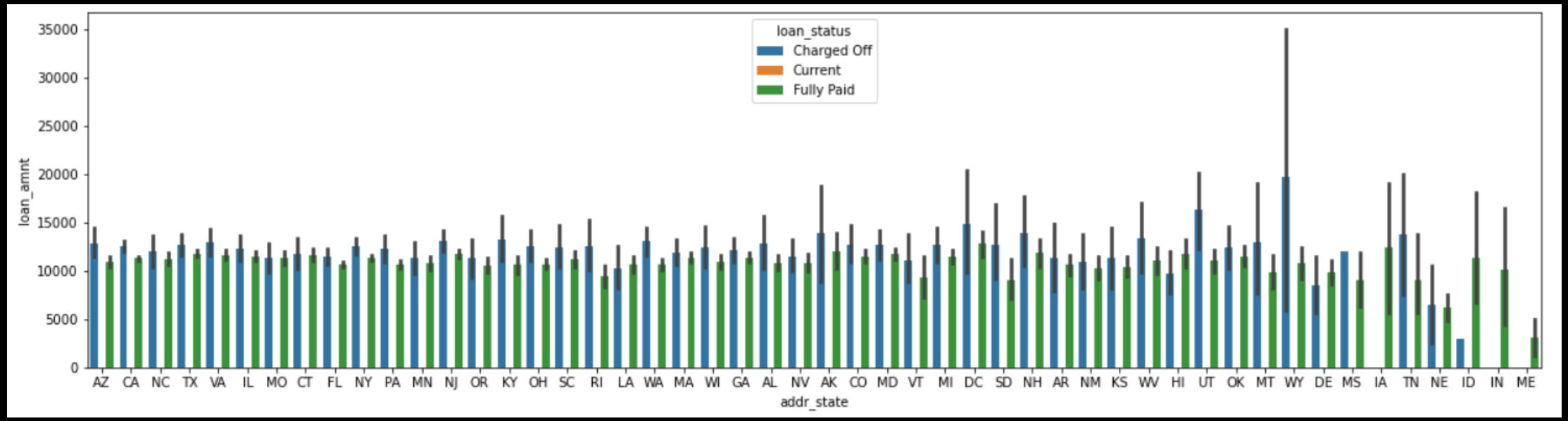


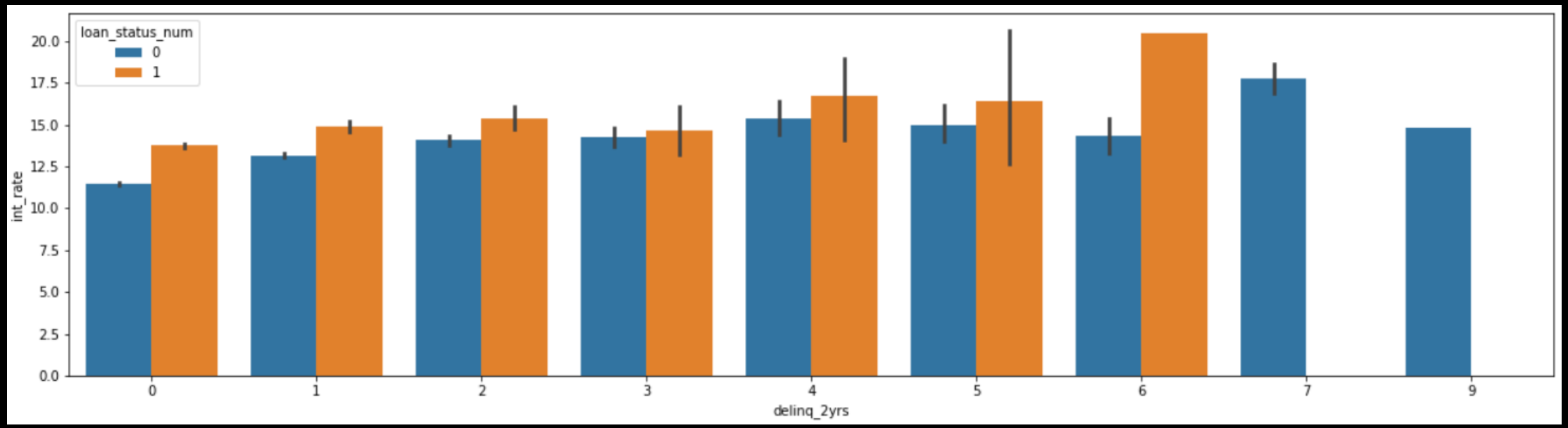
Bivariate Analysis Graphs (cont..)





Bivariate Analysis Graphs (cont..)





Recommendations

From the analysis, below are the recommendations for the club:

- Verification and background check for large loan applications is of prime importance as verified loans (large loans are more likely to default)
- Need to check other sources of income, pervious or ongoing debts with other parties, or assets for applicants with low annual income.
- Check why certain states have a high default rate when compared to others.

THANK YOU