Lending Club Case Study

SUBMISSION

Names:

APRAJITA JHA

KODUGANTI SAI SUDHA MADHAVI

Insights into Risk Analysis for "Lending Club", a consumer finance company.

Business Understanding

Lending Club is a **consumer finance company** which specialises in lending various types of loans to urban customers. 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

Goals of Exploratory Data Analysis:

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.

Loading the Data

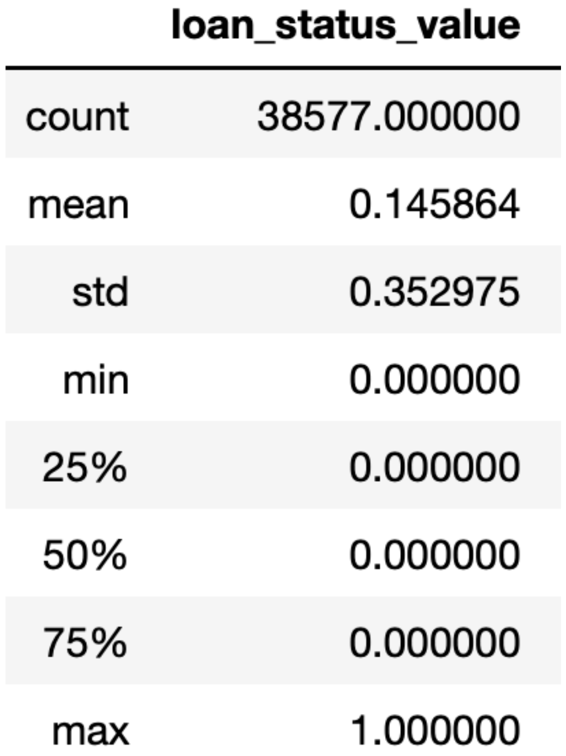
- All the data necessary for analysis has been provided as one dataset, with a supporting data dictionary to understand the meaning of variables
- The provided dataset for analysis is loaded into a dataframe to proceed with the analysis

Data Cleaning

- There are multiple columns i.e, more than 50% columns with no inputs. Hence dropping all such columns is to be the first step of cleaning
- Taking a look at missing values in each column, there were two columns that had significantly great number of missing values which are dropped
- Since we are to look at the loan status with either Fully_Paid or Charged_off, we can drop off the rows with Loan status Current
- Any columns with customer behaviour data and loan repayment data, which are not expected to be available at the time of application of loan, can be dropped off.
- Data Formatting of the remaning columns with interest rate, dates, which are represented as string in the dataframe is to be done.
  - The Target column of whether or not a loan was defaulted is formatted with values 0 and 1, with 0 implying "Fully paid" and 1 implying "Charged Off" as loan status.
  - All the analysis that is done regarding loan status is done on this column – which has boolean like values.
- Proceeding with columns which are then present, with missing values, imputation can be done.
- In the Analysis done, two columns were imputed
  - emp_length 2.6% missing values were imputed with median value of the available values
  - pub_rec_bankruptcies 1.8% missing values were imputed with the mode value of the available values

Exploratory Data Analysis – Univariate Analysis

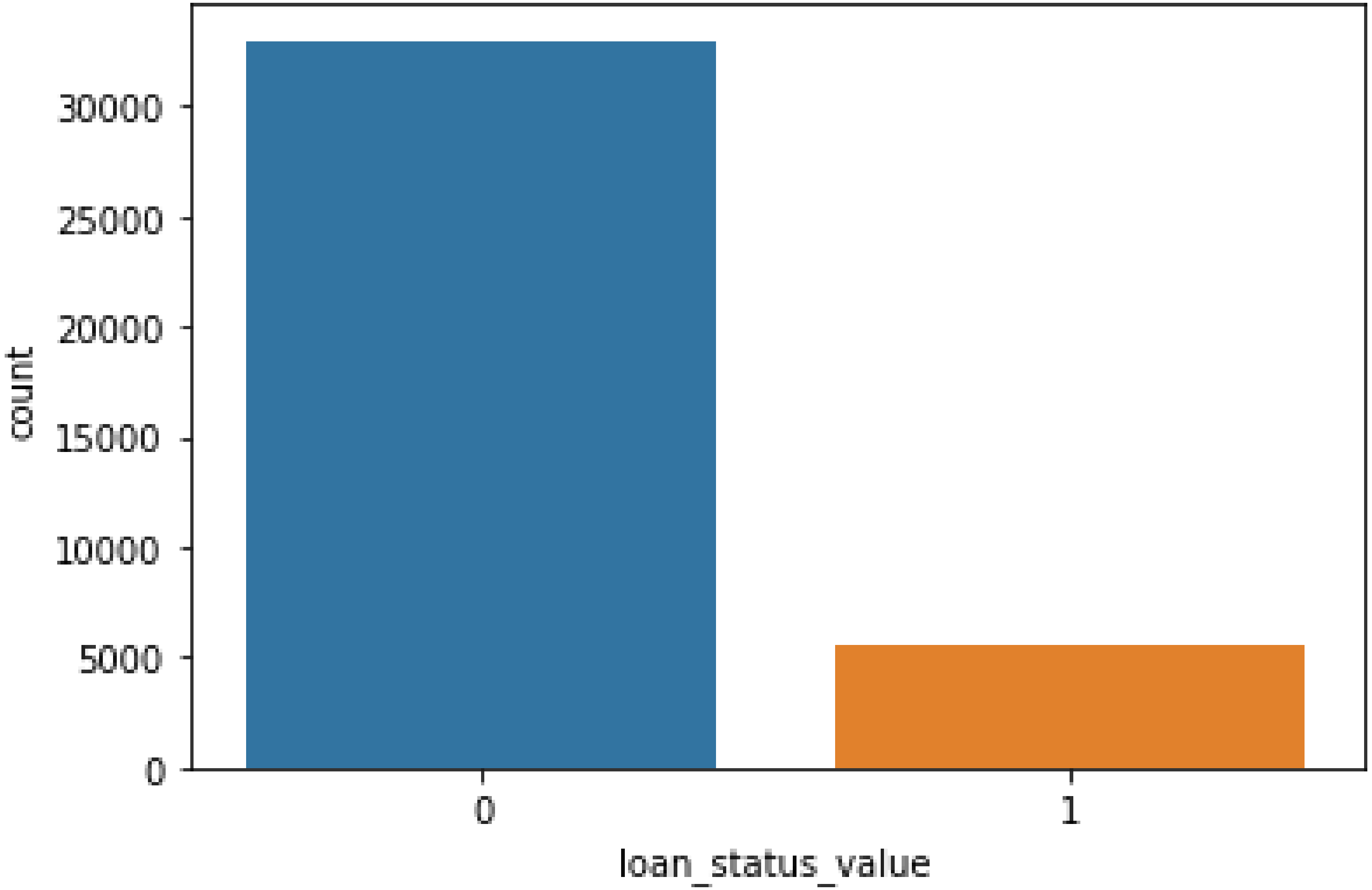
- The Analysis is done on loan_status column with values of 0 and 1.



	loan_status_value
count	38577.000000
mean	0.145864
std	0.352975
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

Exploratory Data Analysis – Univariate Analysis Results:

It is observed that 14.6% of the applicants who are approved for loan have historically Charged-off (or defaulted).

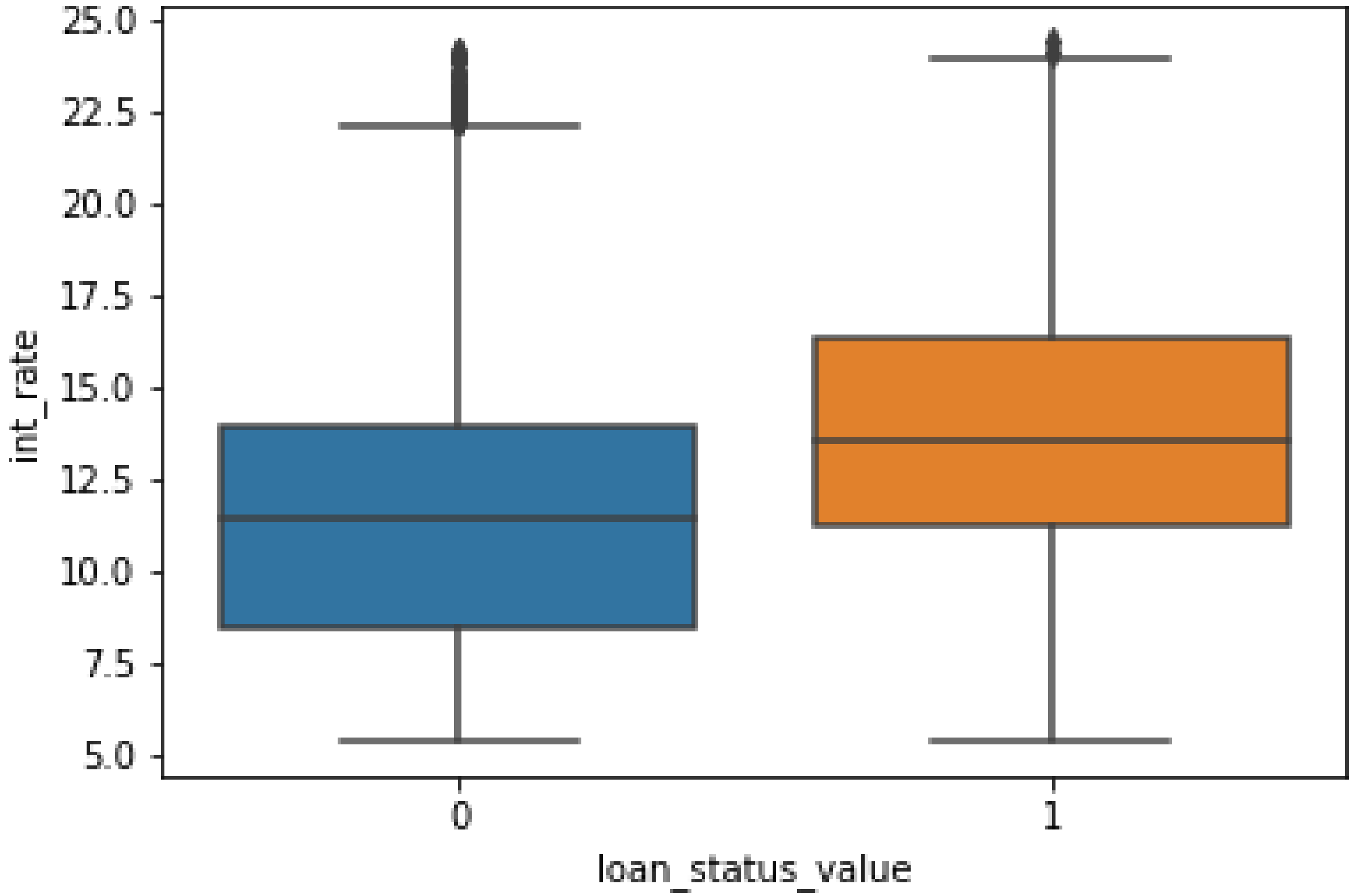


Exploratory Data Analysis
Segmented Univariate Analysis

- Goals of the analysis:
  - To uncover the strongest drivers behind loan defaults
  - We have proceeded with segmenting the loan default metrics with respect to every other available relevant variable, and listed out the drivers.
  - The metric based on which we analysed the segmented data is the average/mean rate of default.
    - Mean is better metric than median as the result set is to have only 2 values {0,1}
    - Since there can be no outlier, the mean is expected to be a fair estimator.
- Observed Strong Drivers of loan Default:
  - Subgrade and hence Grade
    - High grade conforms to higher default rate
    - High implies lexicographically greater. Example B is higher than A
  - int_rate
    - Positively corelated
  - pub_rec_bankruptcies
    - Positively corelated
  - inq_last_6mths
    - Positively corelated
  - loan_amount and hence funded_amnt_inv
    - Positively corelated
  - term
    - Higher default rate observed for 60 month term

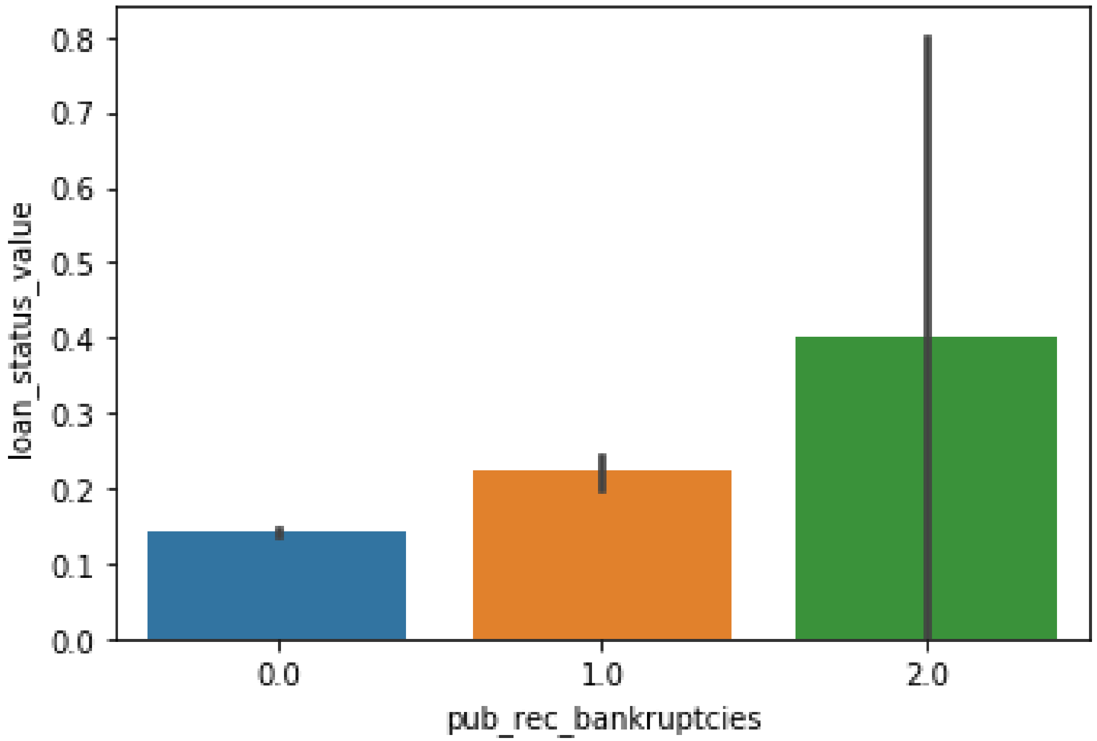
Variable - int_rate

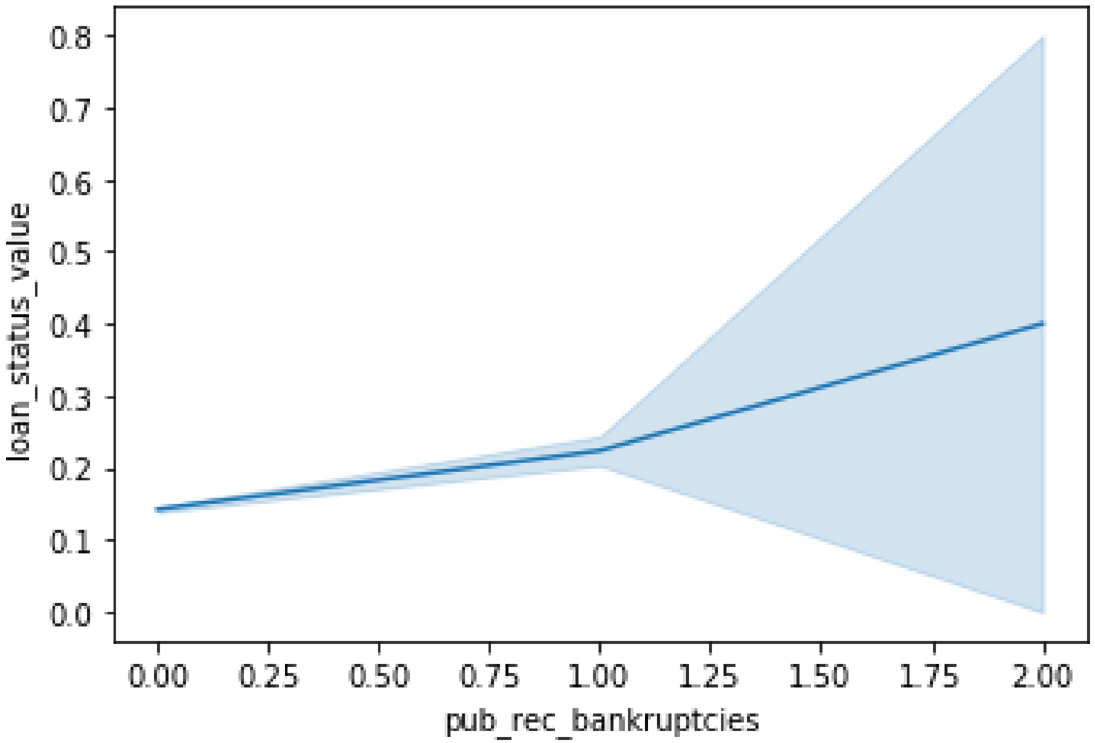
The higher interest rate is frequent in the cases of defaulting



Variable - pub_rec_bankruptcies

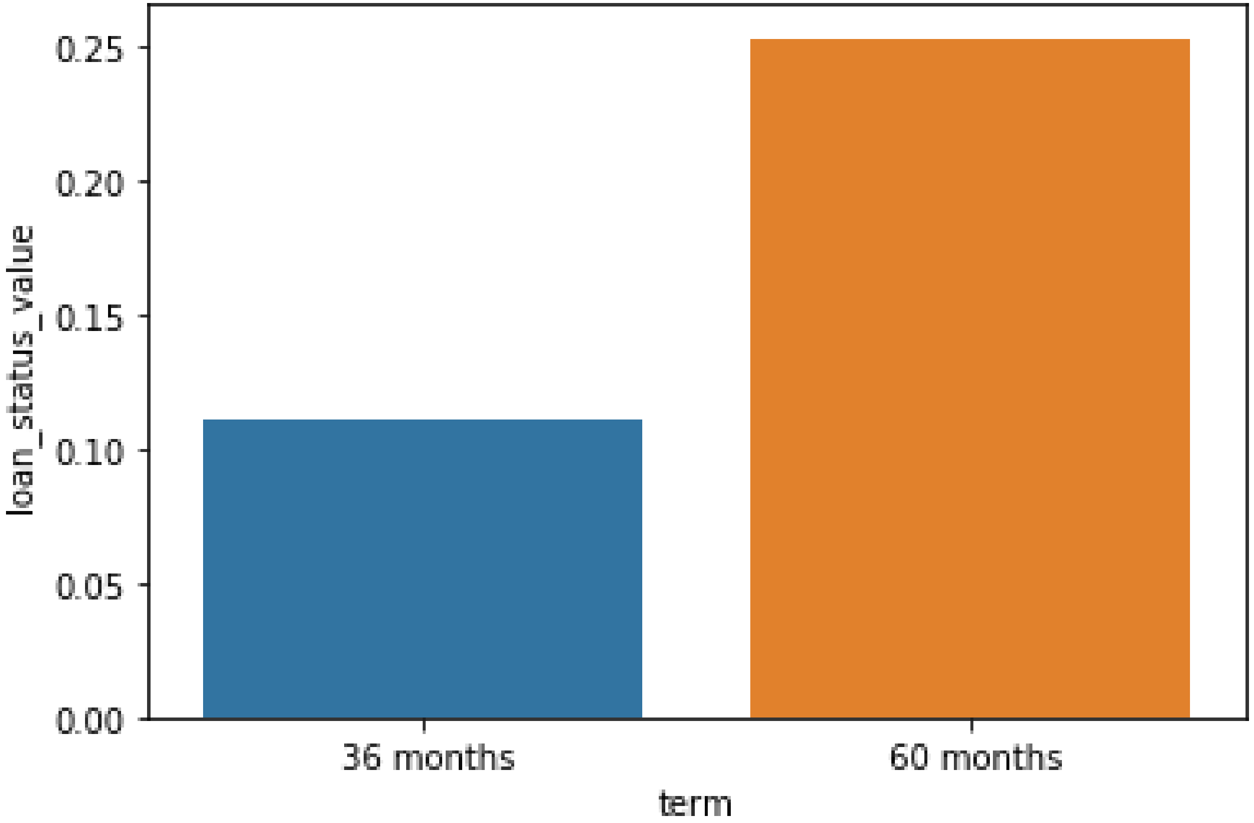
As the annual income increases, rate of default is observed to be decreasing, steadily until the annual income is <= 200000





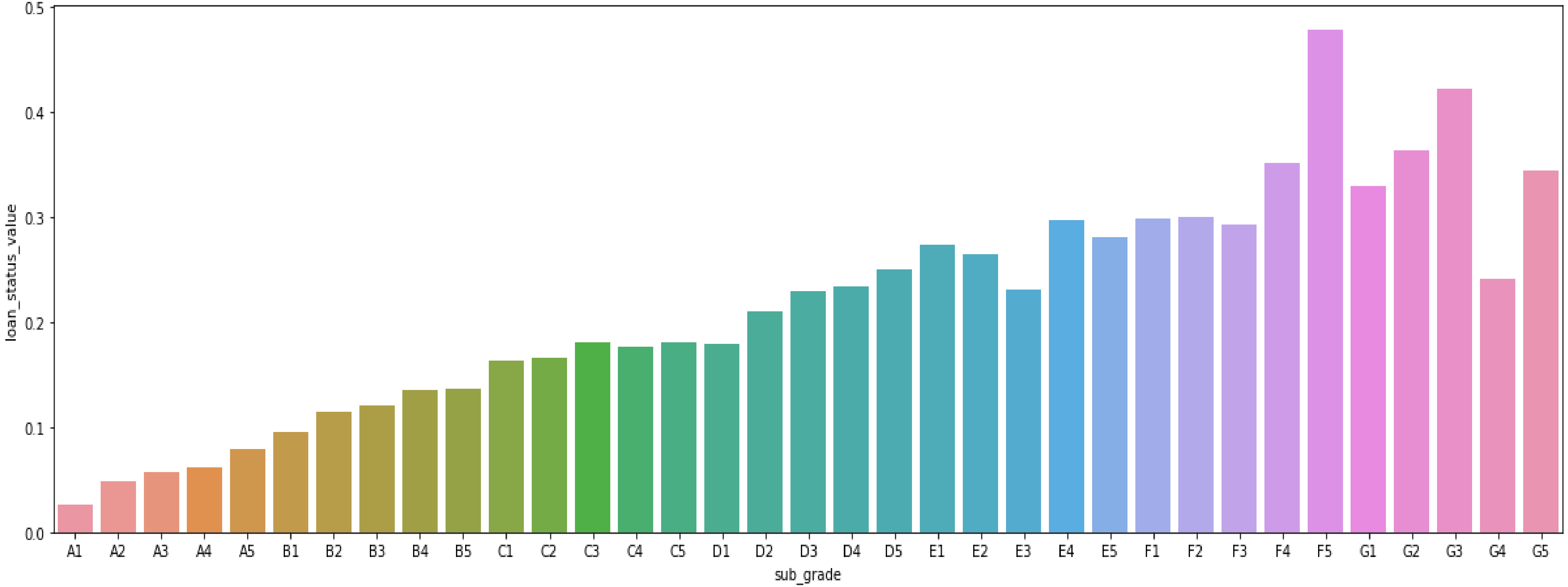
Variable – term

From the two available terms, the average rate of default is much higher for 60 month term, almost by 2 times



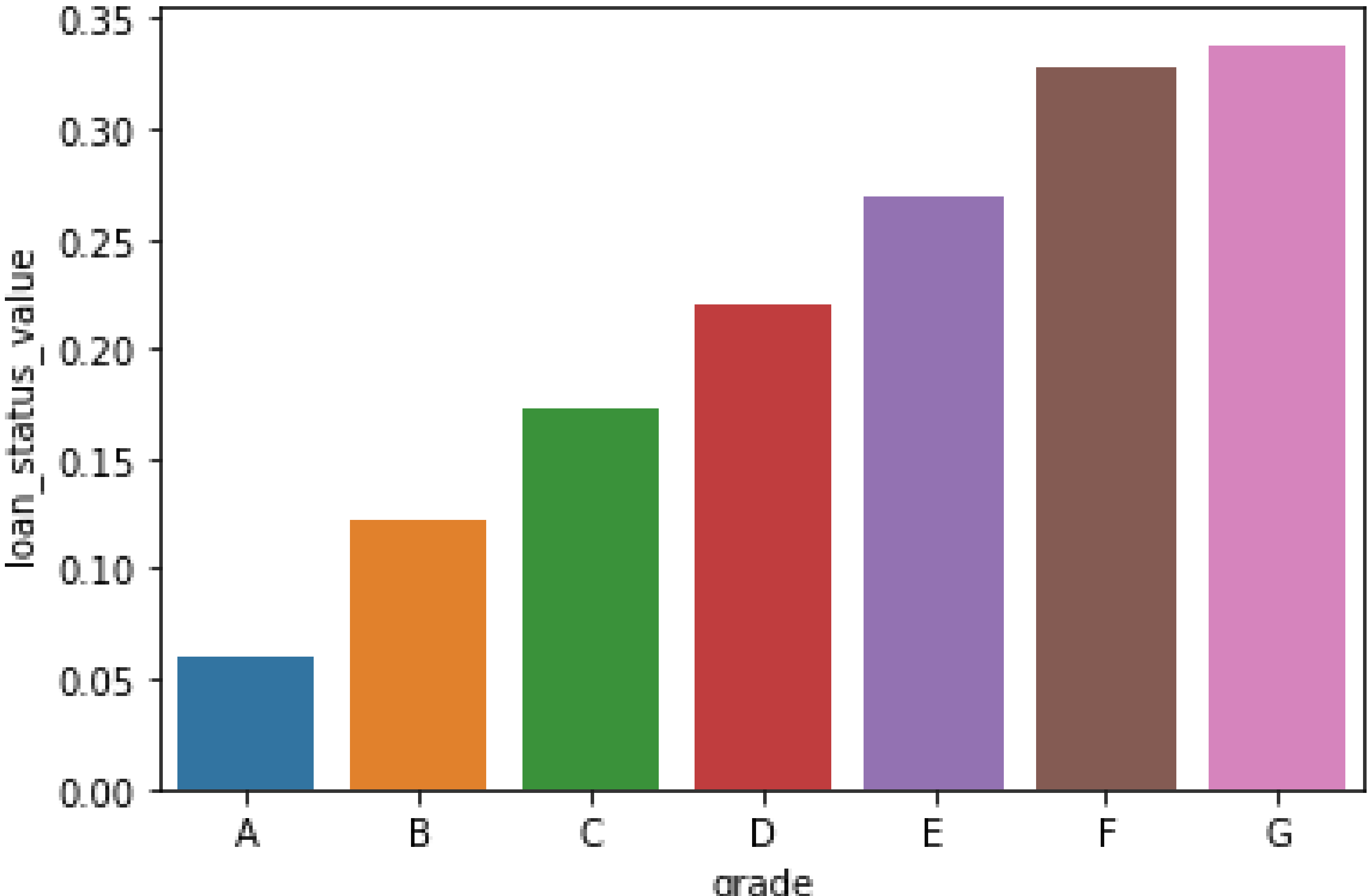
Variable - sub_grade

Similar to the grade we can see that as we move towards the right, the trend is that the rate of default increases



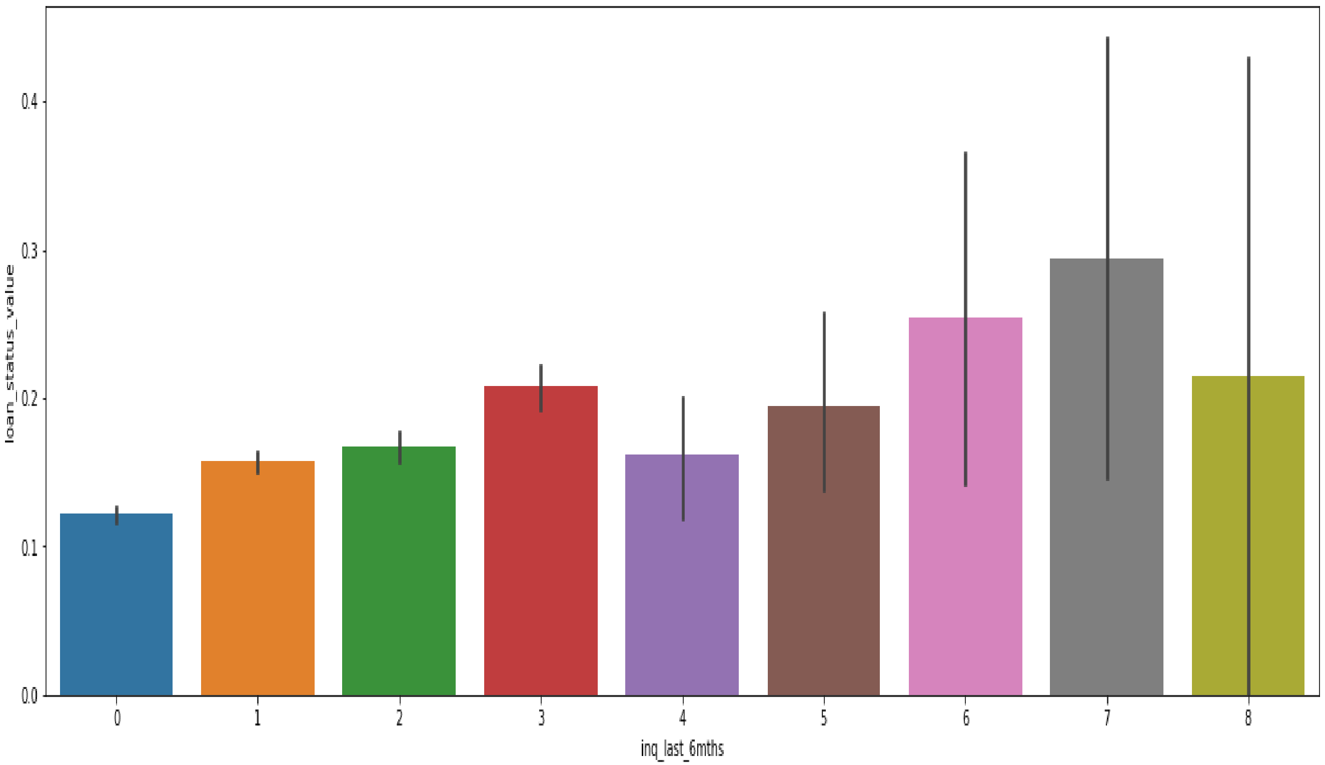
Variable - grade

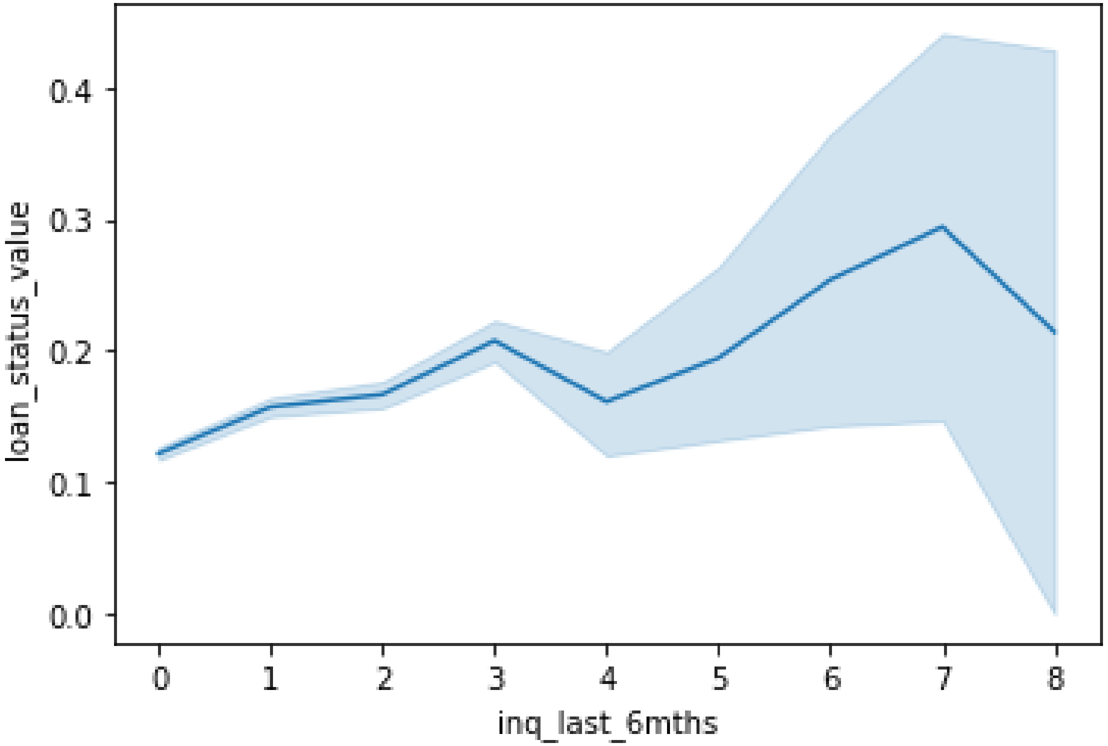
Grades from E to G are at significant risk of defaulting



Variable - inq_last_6mths

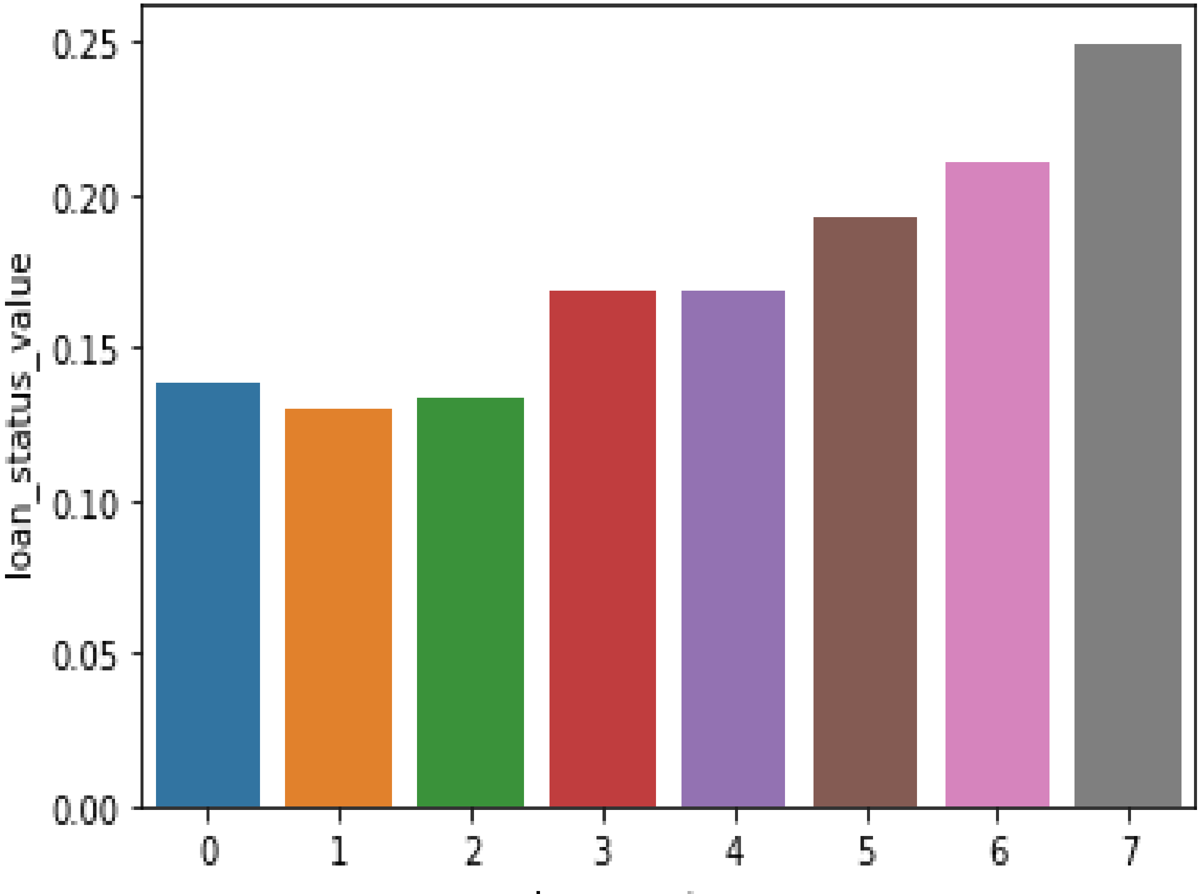
The lineplot shows there is positive corelation - as the default rate increased along with increase in inq_last_6mths

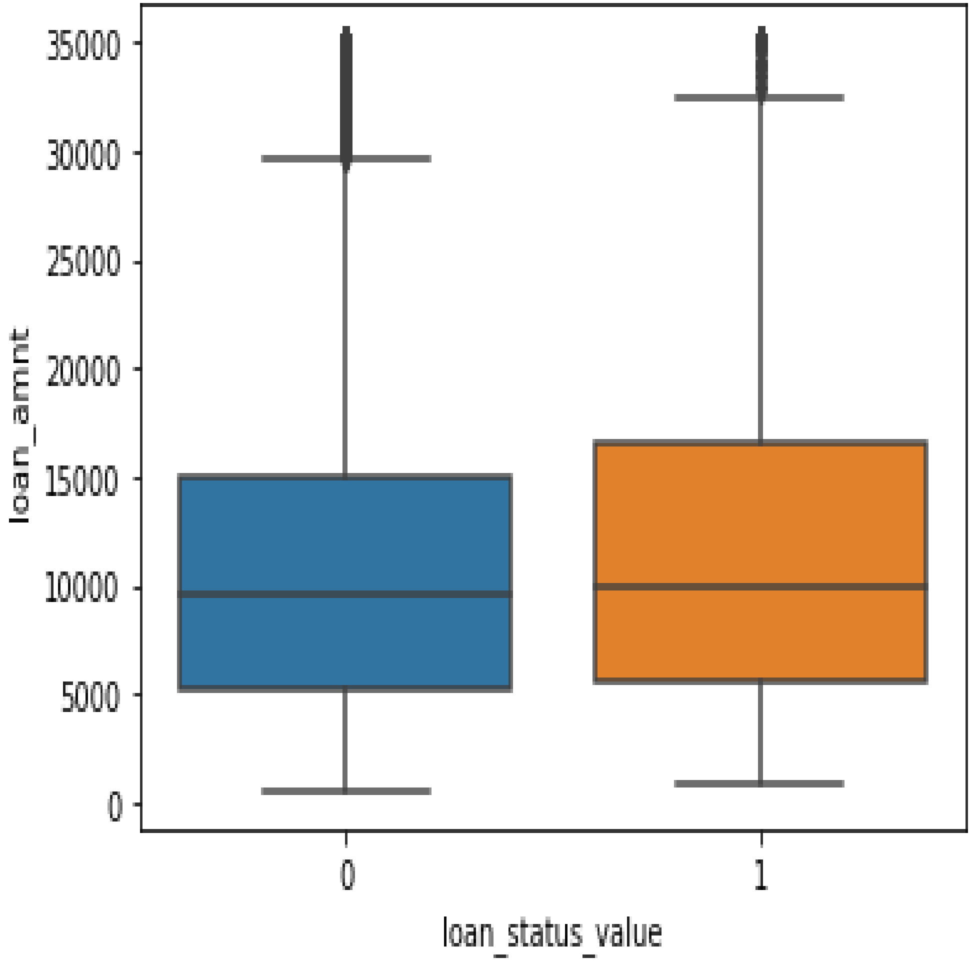




Variable – loan_amnt

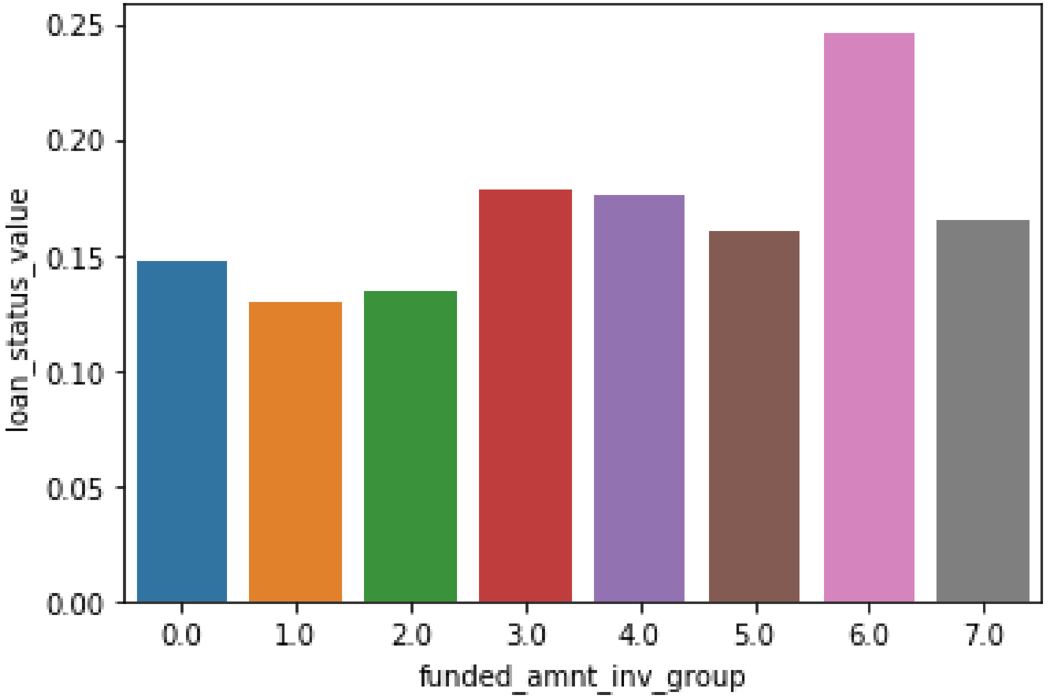
The average rate of loan default increases as the loan_amount increases.

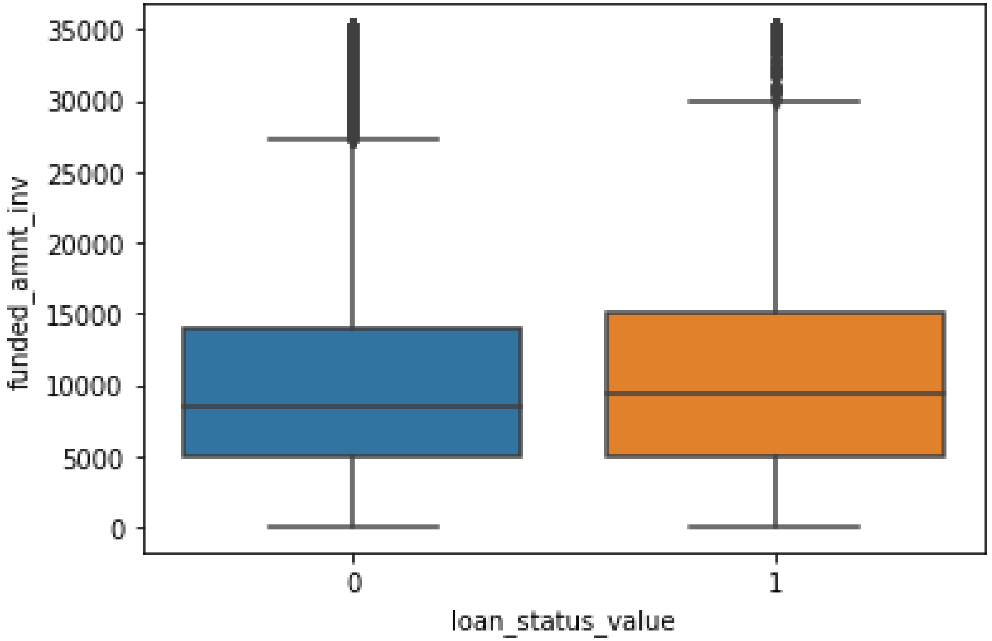




Variable - funded_amnt_inv

The results show that the average rate of loan default increases as the funded_amnt_inv increases





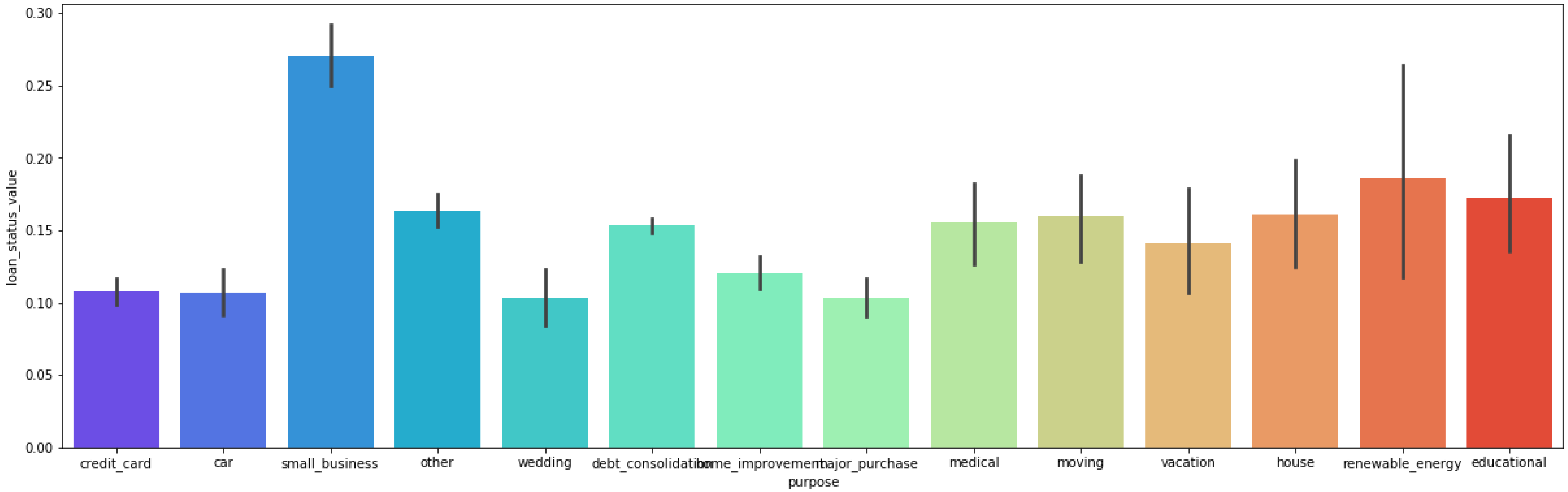
Exploratory Data Analysis
Segmented Univariate Analysis

Observed Weak to Mid-Range Drivers in that order:

- open_acc to total_acc ratio
- emp_length
- verification_status_range
- dti
- home_ownership
- Installment
- annual_inc
- purpose

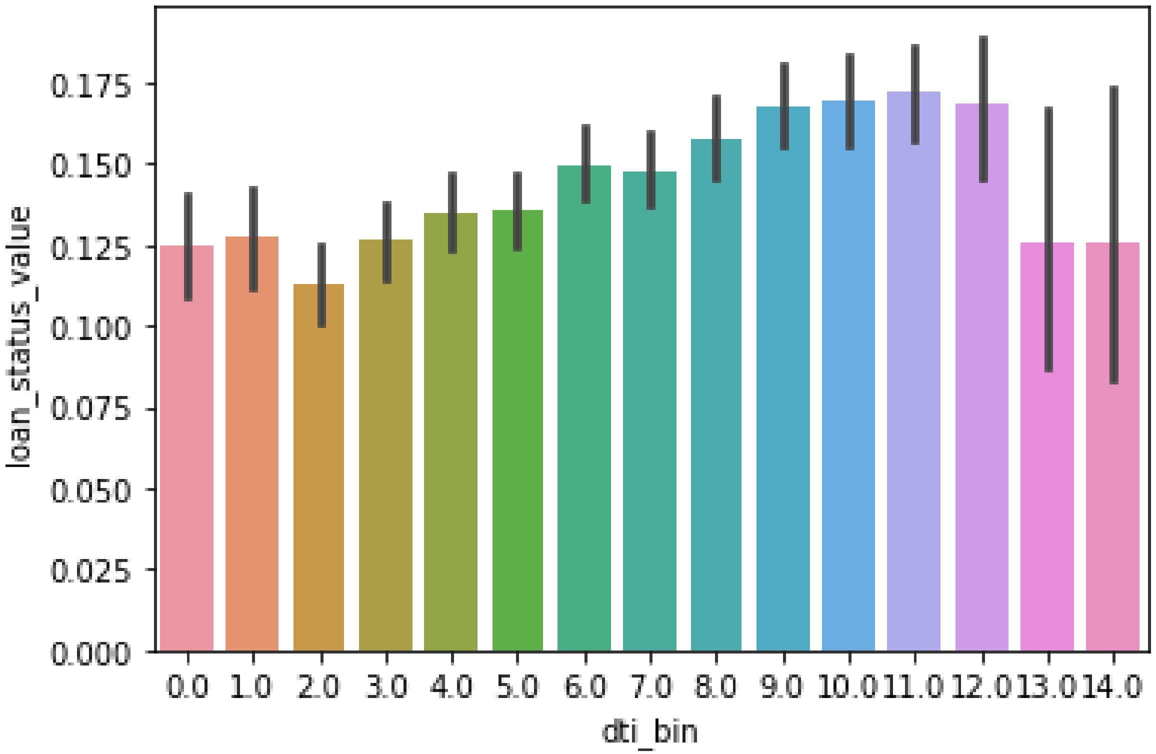
Variable –Purpose

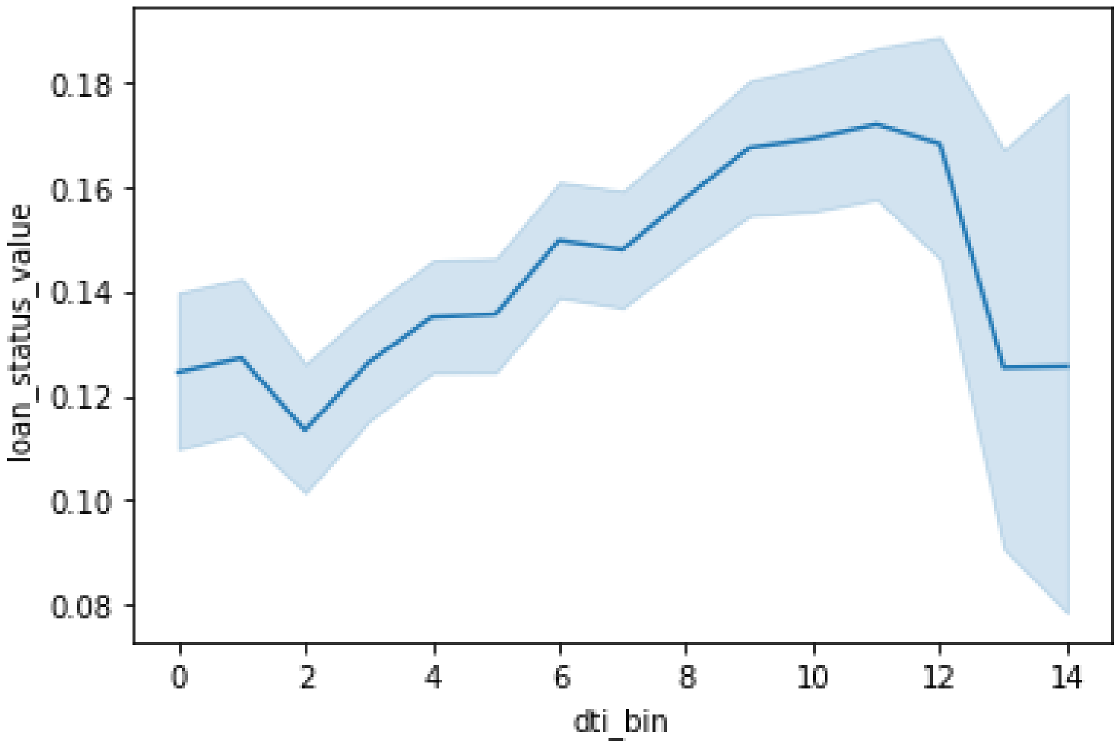
Small_business and renewable_energy tend to show high rate of defaulting the loan



Variable - dti

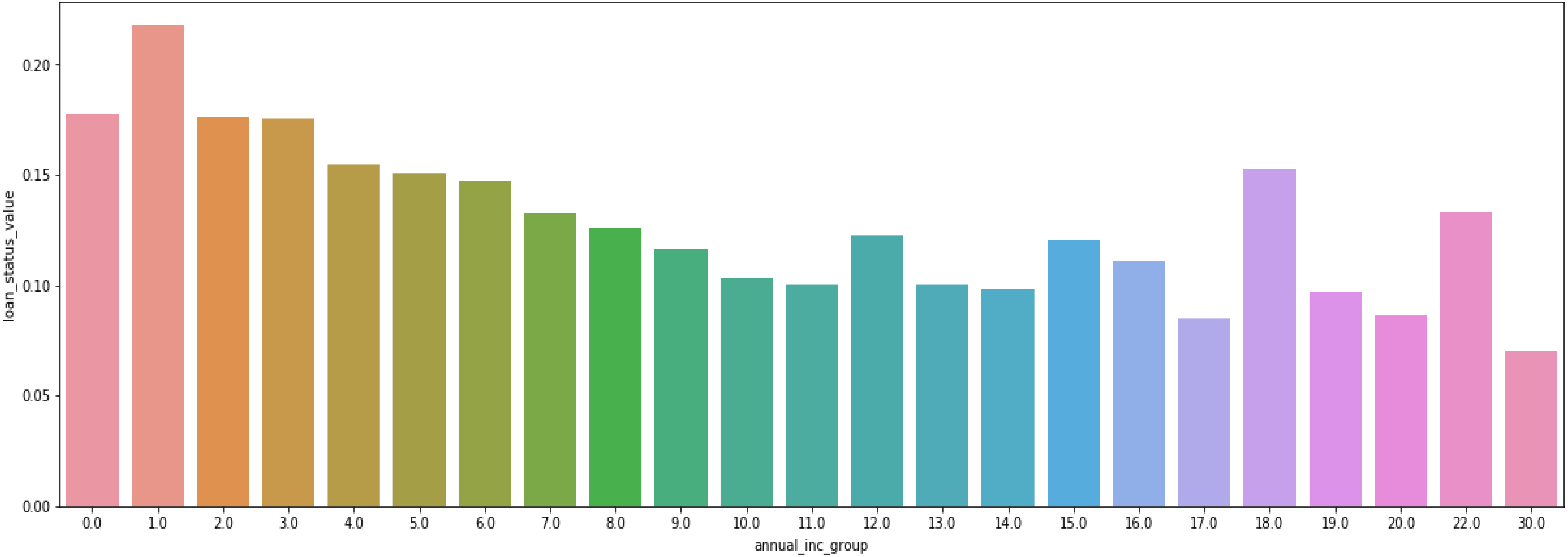
Dti is a weak positively corelated driver for default rate





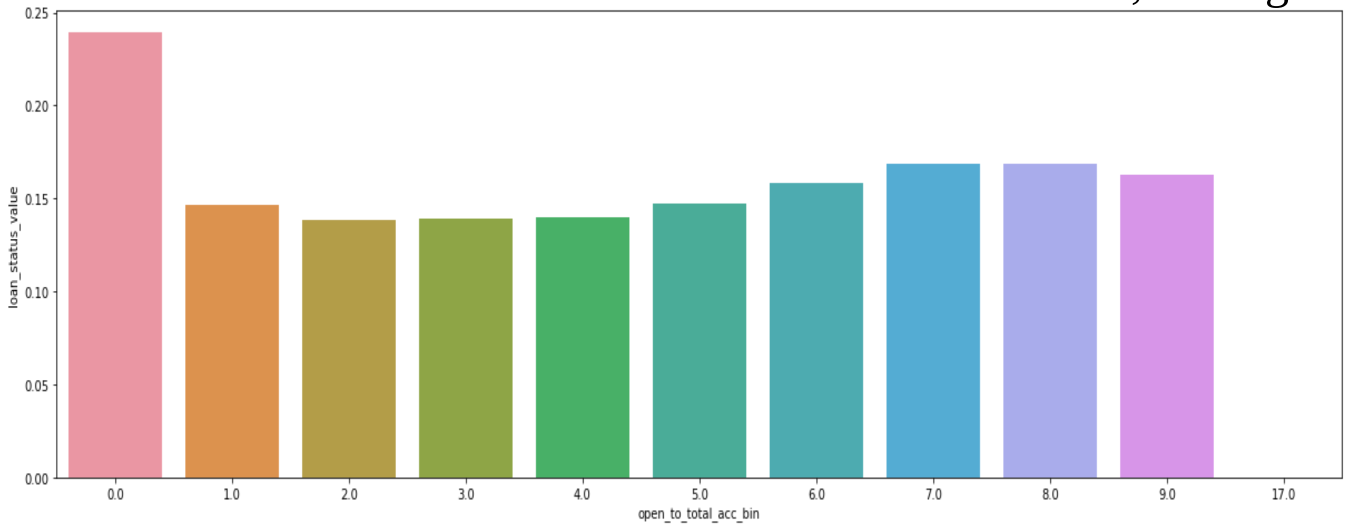
Variable - annual_inc

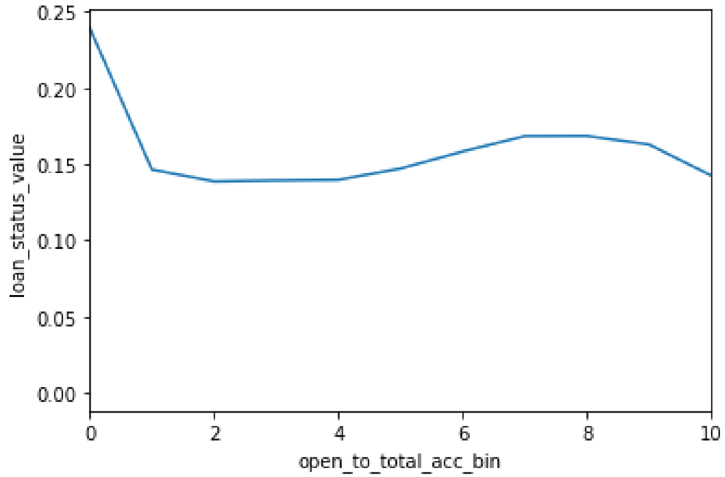
As the annual income increases, rate of default is observed to be decreasing, steadily until the annual income is <= 200000



Derived metrics

- Type metric: Binned variables
  - To be able to aggregate and draw conclusions of the default rate over different quantitative variables, we have decided to bin the same into intervals. By creating the intervals, we were able to see correlation of the ranges with the default rate.
  - Some Variables on which we did binning - installments, income, loan amount, dti, employment length
- Business metric: days_since_earliest_cr_line
  - Out of the domain knowledge that a person holding a credit line for a long duration implies financial discipline, we have decided to compute the number of days until the loan issue date(issue_d) that a credit line was open.
- Data driven metric: open_to_total_acc
  - Upon finding no significant trend with respect to either open accounts and total accounts, we have decided to go with considering the ratio of open_acc to total_acc.
  - We noticed that this metric is a weak driver for default rate, as long as the atleast one account is opened

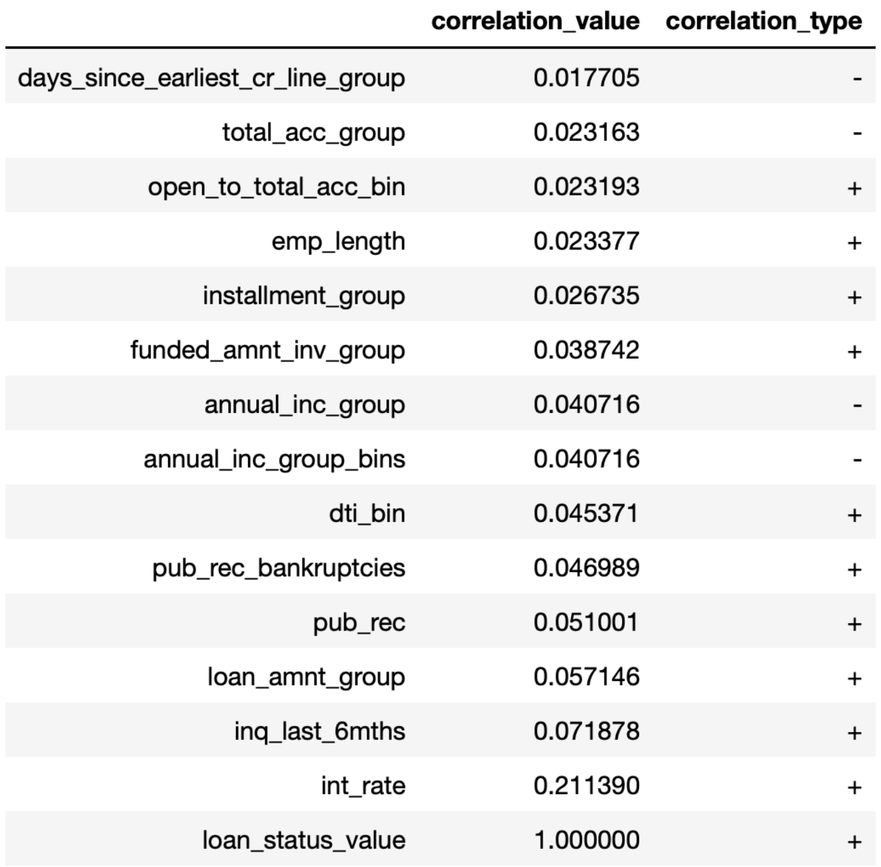




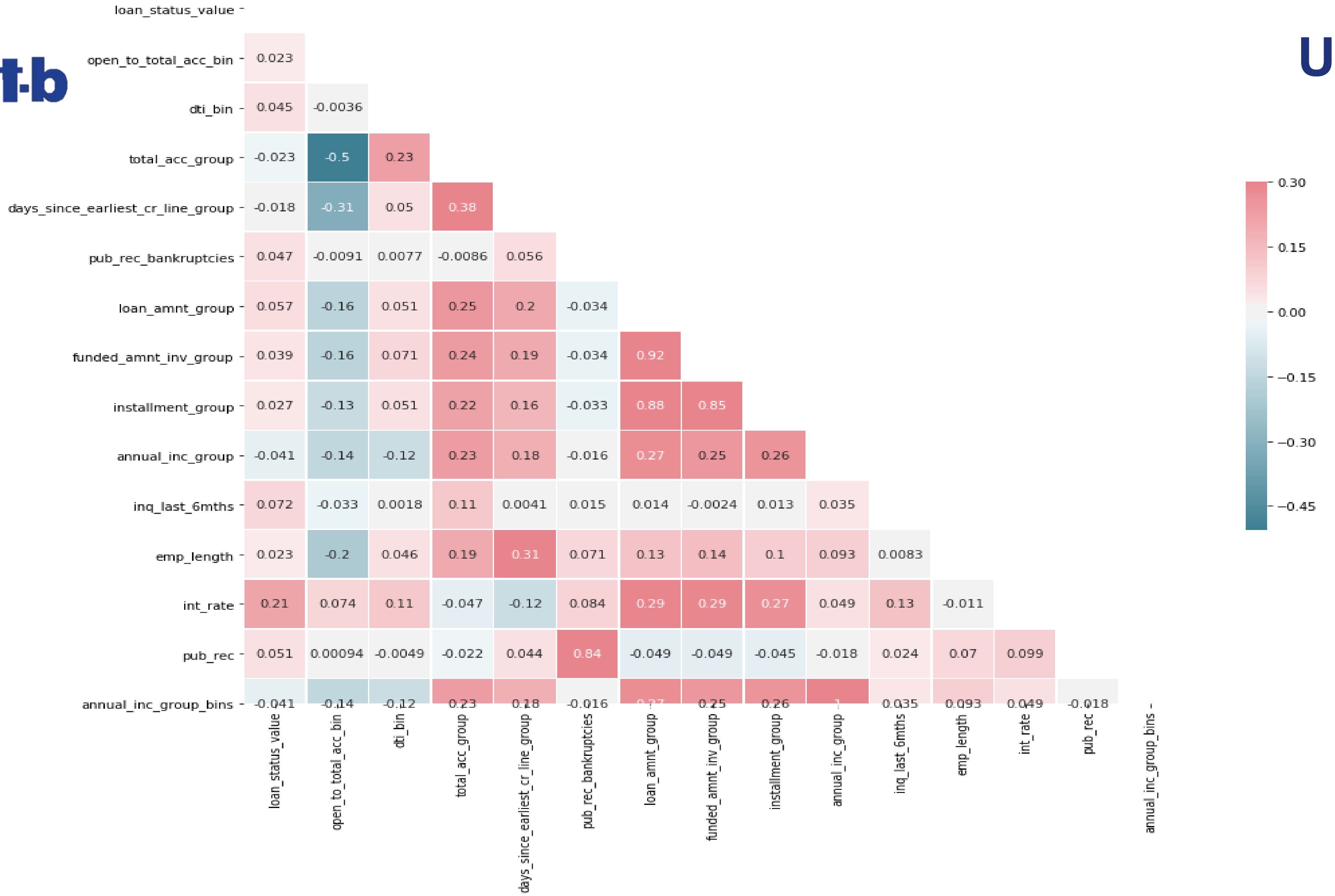
Observations of Segmented Univariate Analysis

- Term – 60 month term is riskier
- Purpose – Small business as purpose shows higher risk
- Higher value of the following shows higher risk:
  - Public record bankruptcies
  - Loan amount
  - Interest rate
  - Installments
- Higher income shows lower risk of default
- State NE tends to show higher risk of default

Correlation matrix – Bivariate analysis



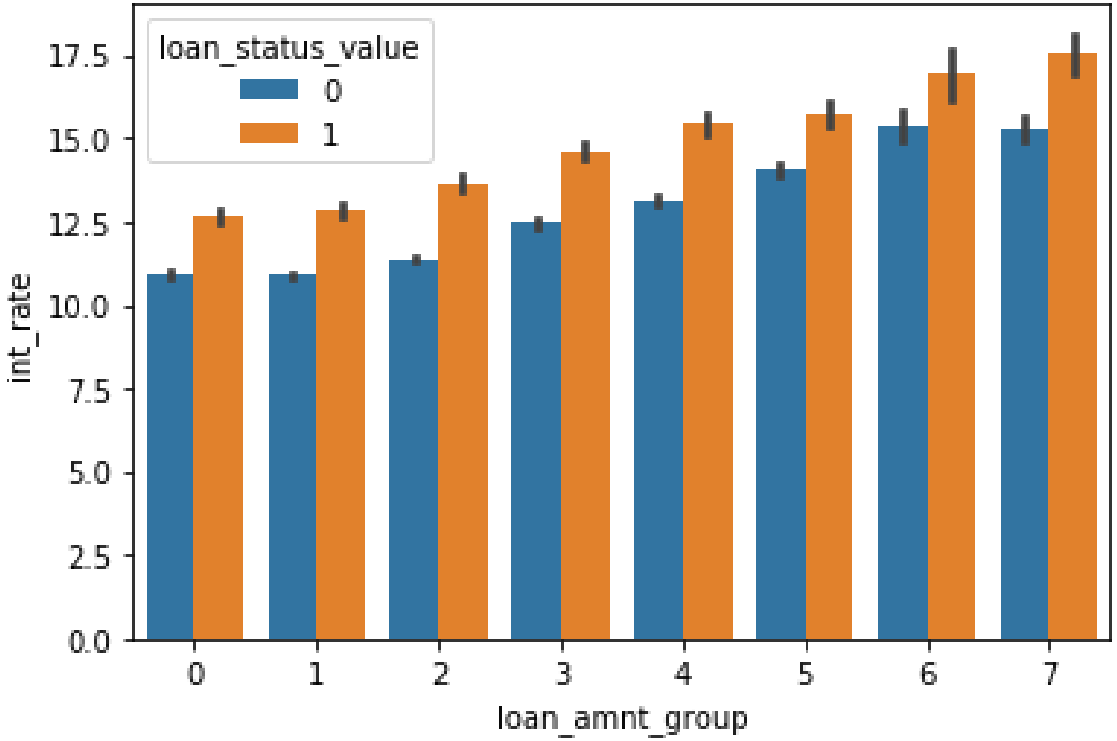
	correlation_value	correlation_type
days_since_earliest_cr_line_group	0.017705	-
total_acc_group	0.023163	-
open_to_total_acc_bin	0.023193	+
emp_length	0.023377	+
installment_group	0.026735	+
funded_amnt_inv_group	0.038742	+
annual_inc_group	0.040716	-
annual_inc_group_bins	0.040716	-
dti_bin	0.045371	+
pub_rec_bankruptcies	0.046989	+
pub_rec	0.051001	+
loan_amnt_group	0.057146	+
inq_last_6mths	0.071878	+
int_rate	0.211390	+
loan_status_value	1.000000	+



loan_status_value
open_to_total_acc_bin 0.023
dti_bin 0.045 -0.0036
total_acc_group -0.023 -0.5 0.23
days_since_earliest_cr_line_group -0.018 -0.31 0.05 0.38
pub_rec_bankruptcies 0.047 -0.0091 0.0077 -0.0086 0.056
loan_amnt_group 0.057 -0.16 0.051 0.25 0.2 -0.034
funded_amnt_inv_group 0.039 -0.16 0.071 0.24 0.19 -0.034 0.92
installment_group 0.027 -0.13 0.051 0.22 0.16 -0.033 0.88 0.85
annual_inc_group -0.041 -0.14 -0.12 0.23 0.18 -0.016 0.27 0.25 0.26
inq_last_6mths 0.072 -0.033 0.0018 0.11 0.0041 0.015 0.014 -0.0024 0.013 0.035
emp_length 0.023 -0.2 0.046 0.19 0.31 0.071 0.13 0.14 0.1 0.093 0.0083
int_rate 0.21 0.074 0.11 -0.047 -0.12 0.084 0.29 0.29 0.27 0.049 0.13 -0.011
pub_rec 0.051 0.00094 -0.0049 -0.022 0.044 0.84 -0.049 -0.049 -0.045 -0.018 0.024 0.07 0.099
annual_inc_group_bins -0.041 -0.14 -0.12 0.23 0.18 -0.016 0.27 0.25 0.26 1 0.035 0.093 0.049 -0.018
loan_status_value
open_to_total_acc_bin
dti_bin
total_acc_group
days_since_earliest_cr_line_group
pub_rec_bankruptcies
loan_amnt_group
funded_amnt_inv_group
installment_group
annual_inc_group
inq_last_6mths
emp_length
int_rate
pub_rec
annual_inc_group_bins
0.30
0.15
0.00
−0.15
−0.30
−0.45

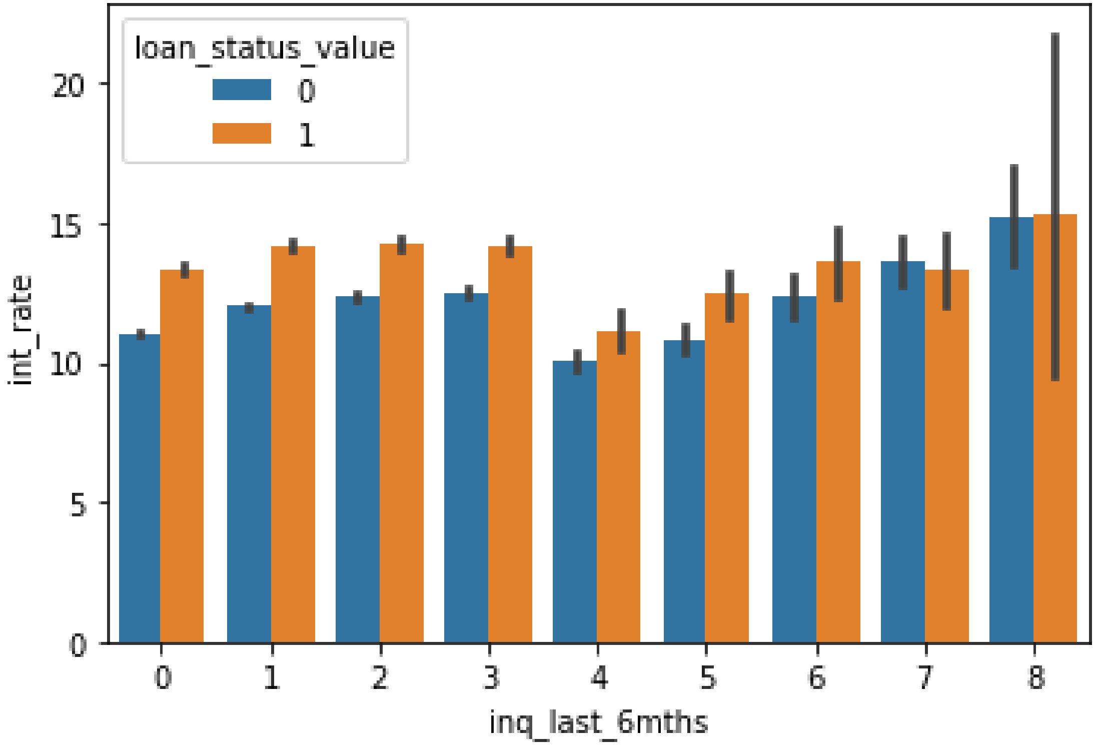
BIVARIATE AND MULTI VARIATE ANALYSIS
Quantitative variables

- Bivariate analysis between loan interest rate and loan amount:
  - At any loan amount group, the defaulted loans have shown higher interest rate.
  - Recommendation:
    - Try to figure out how reducing interest rate might impact default rate



BIVARIATE AND MULTI VARIATE ANALYSIS
Quantitative variables

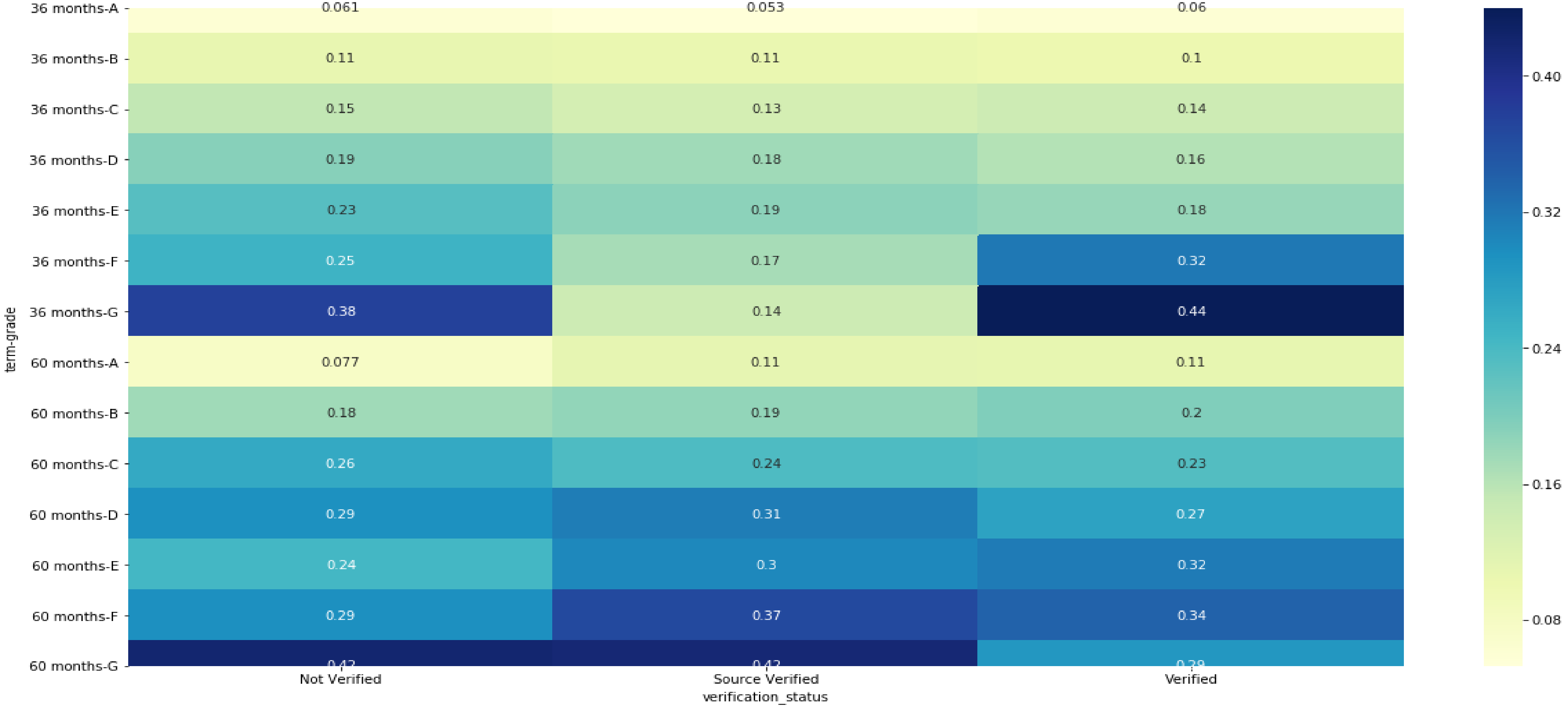
- Bivariate analysis between loan interest rate and number of inquiries:
  - Even the reduction of interest rate did not make impact on default rate when number of inquiries is high.
  - Recommendation:
    - If there had been more enquiries, reduction in interest rate might not make a difference.



BIVARIATE AND MULTI VARIATE ANALYSIS
Categorical variables

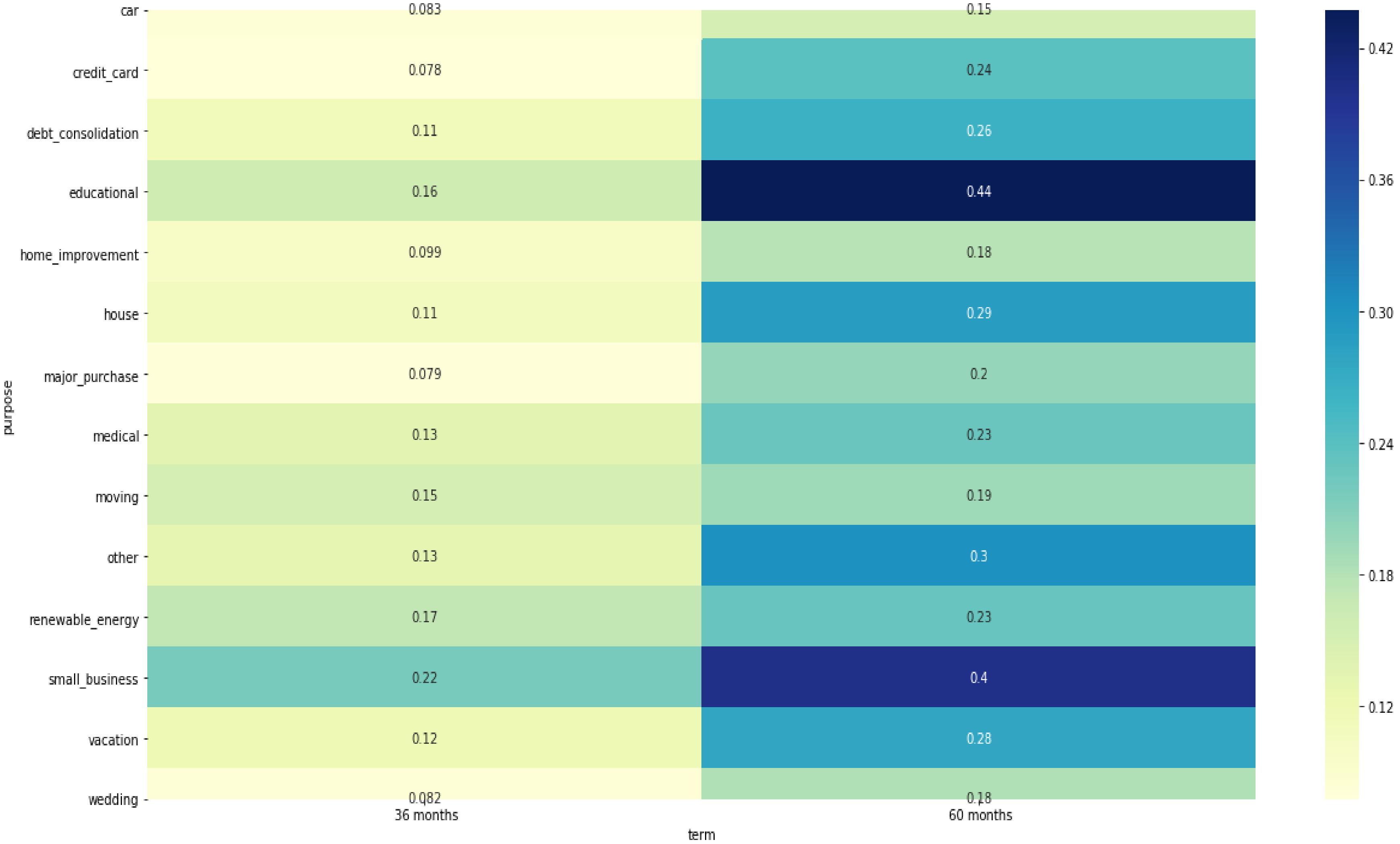
Multivariate analysis between term, grade and verification status:

- The default rate is high even the status is verified, when the Grade is high(F, G).
- While 60 month term is usually risky across all values of verification status



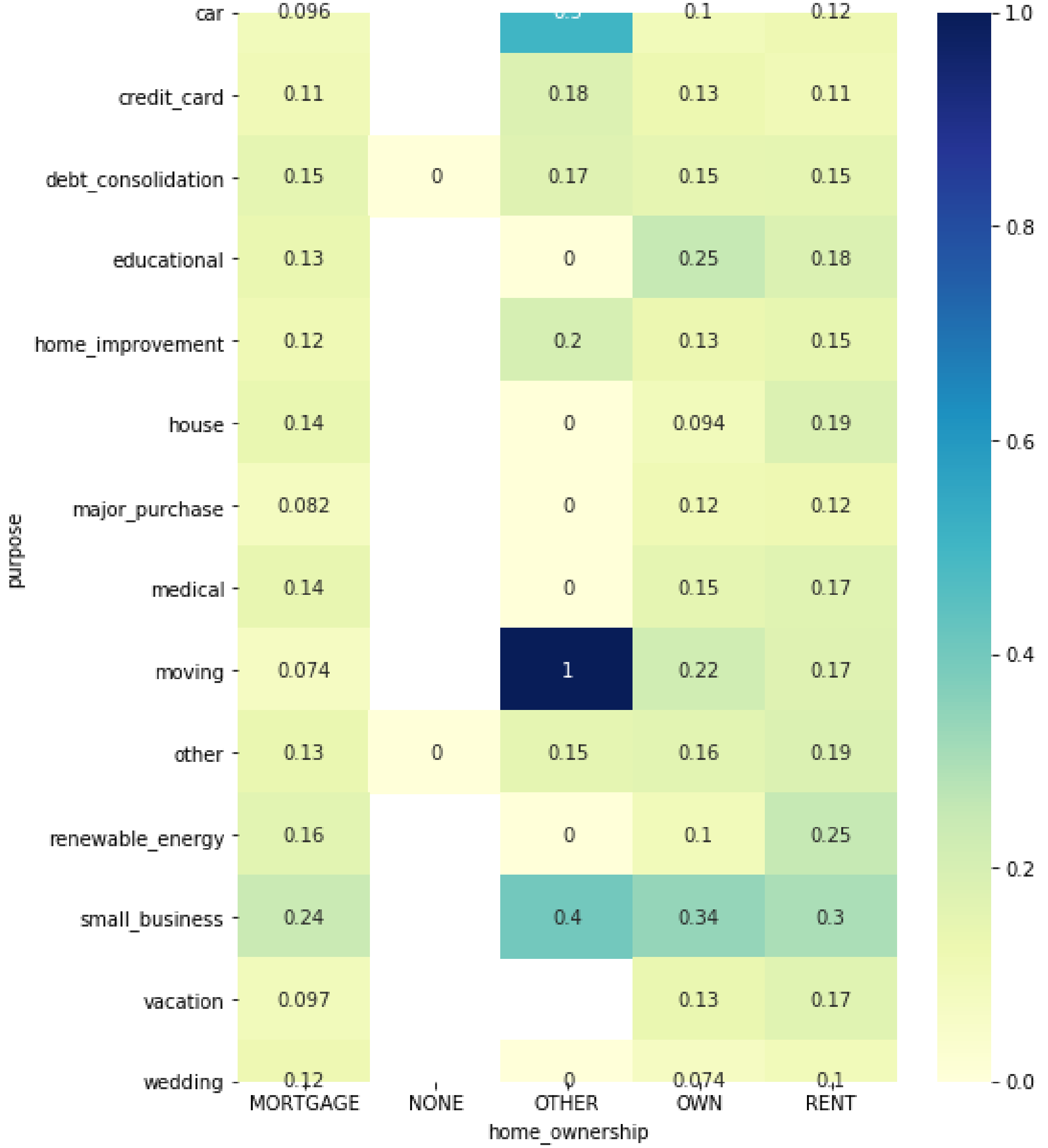
BIVARIATE AND MULTI VARIATE ANALYSIS
Categorical variables

- Bivariate analysis between term and purpose
- Recommendation:
  - Under 36 month term:
    - small_business is a risky investment
  - Under 60 month term:
    - education and small_business are risky investments



BIVARIATE AND MULTI VARIATE ANALYSIS
Categorical variables

- Bivariate analysis between home_ownership and purpose
- Recommendation:
  - It is risky to approve loan for purpose of moving when home_ownership is other



Thank You!