Lending club case study

Objective and problem statement:

- EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of loan default.
- If one is able to identify these 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, EDA to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default.

Analysis approach

- Excel was analyzed at high level along with metadata file
- Then time was spent on understanding columns in more depth by google search to understand their business significance
- Around 50+ columns had only NA, and hence made no sense for any analysis and could be removed
- No empty values were found, few columns has NA values on few rows but removing them was not necessary nor those could be enriched from external sources
- Many columns have categorial values already in them, removing outliers in them did not felt necessary considering very low number of outliers as well as considering them to be categorial in themself. For example: emp_length column which has values ranging from 1 year to 10+ years

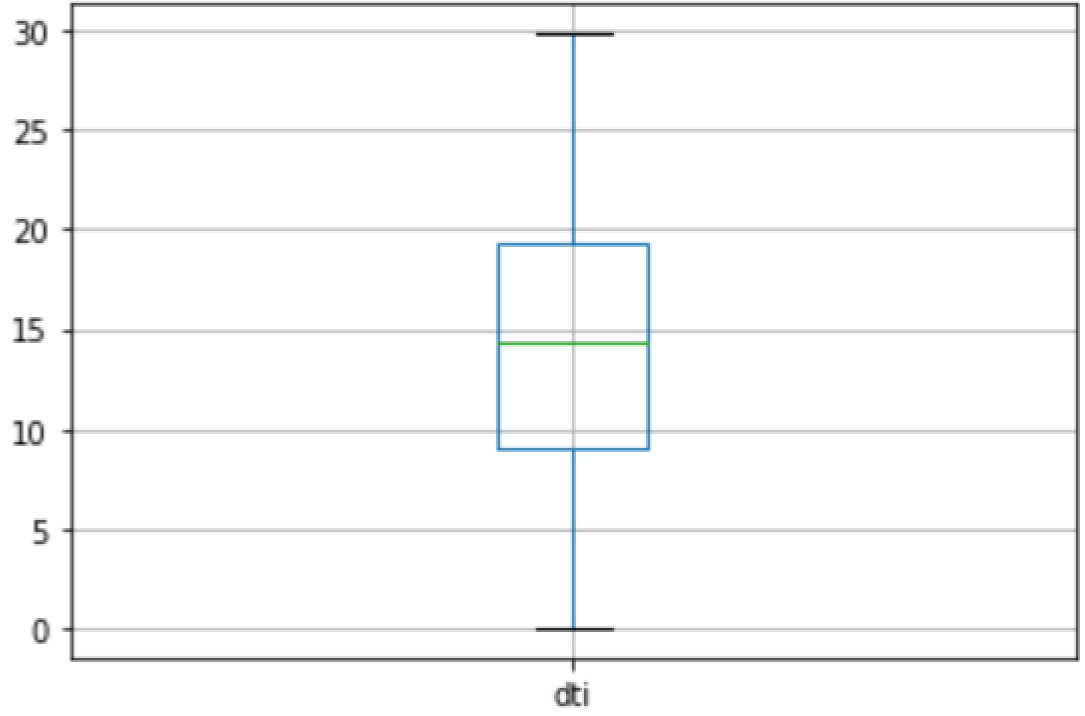
Univariate analysis

Below columns were studied, analyzed in detail as those seemed relevant for problem under discussion, remaining columns were not found significant hence were not studied in further detail other than high glance

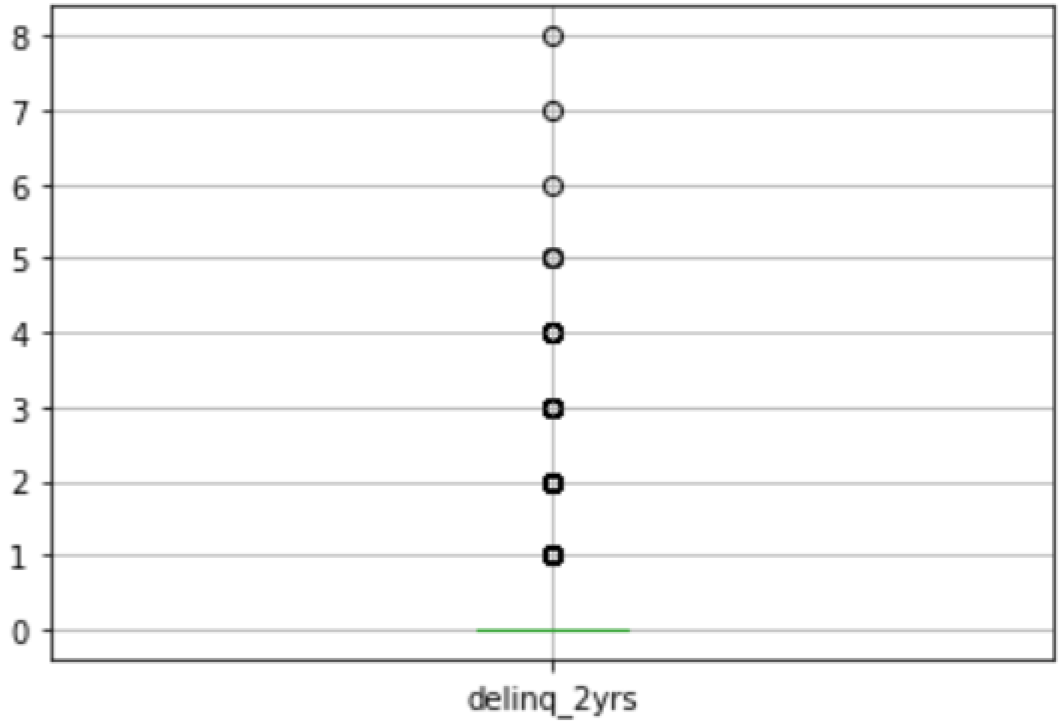
- All loan ids were found to be unique, so no duplicates
- Emp len values ranging from <1 year to 10+ years
- Loan term either 36 months or 60 months
- Interest rate ranging from 5% to 25%
- Loan grades ranging from A to G, with subgrades 1-5
- Home ownership RENT, MORTGAGE, OTHER, OWN
- verification_status - Source Verified, Verified, Not Verified
- Loan purpose had around 14 separate categories
- Different state and zip codes
- revol_util % ranging from 0 to 99
- delinq_2yrs, inq_last_6mths, mths_since_last_delinq, open_acc

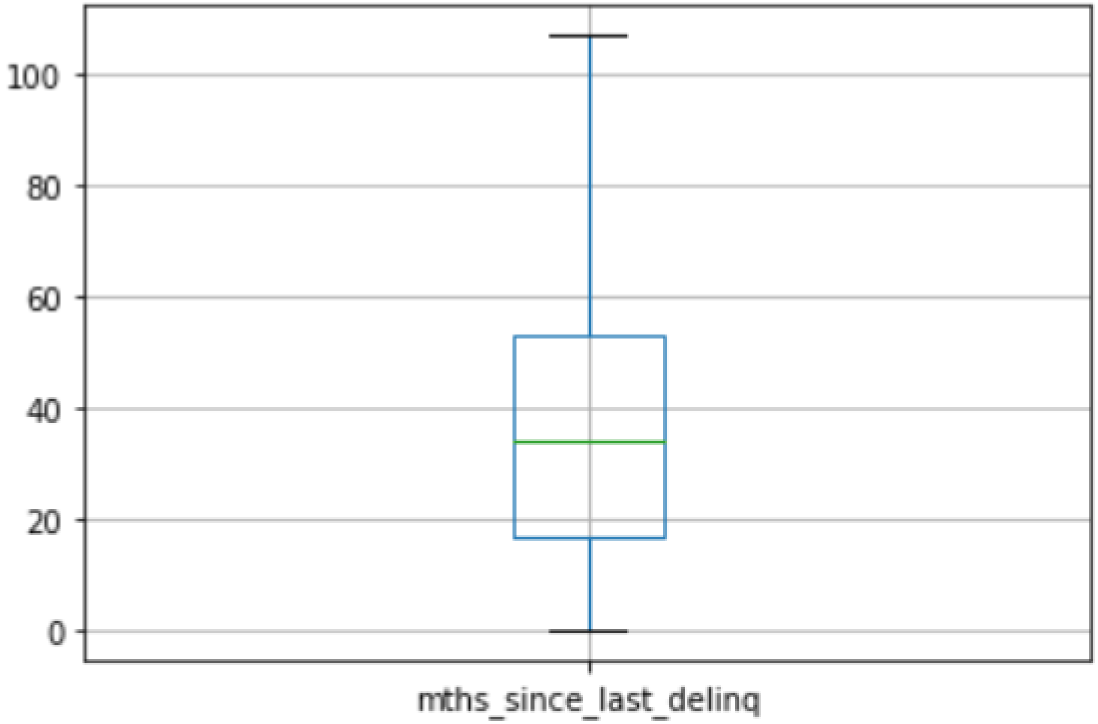
Outlier detection/analysis on selected columns:

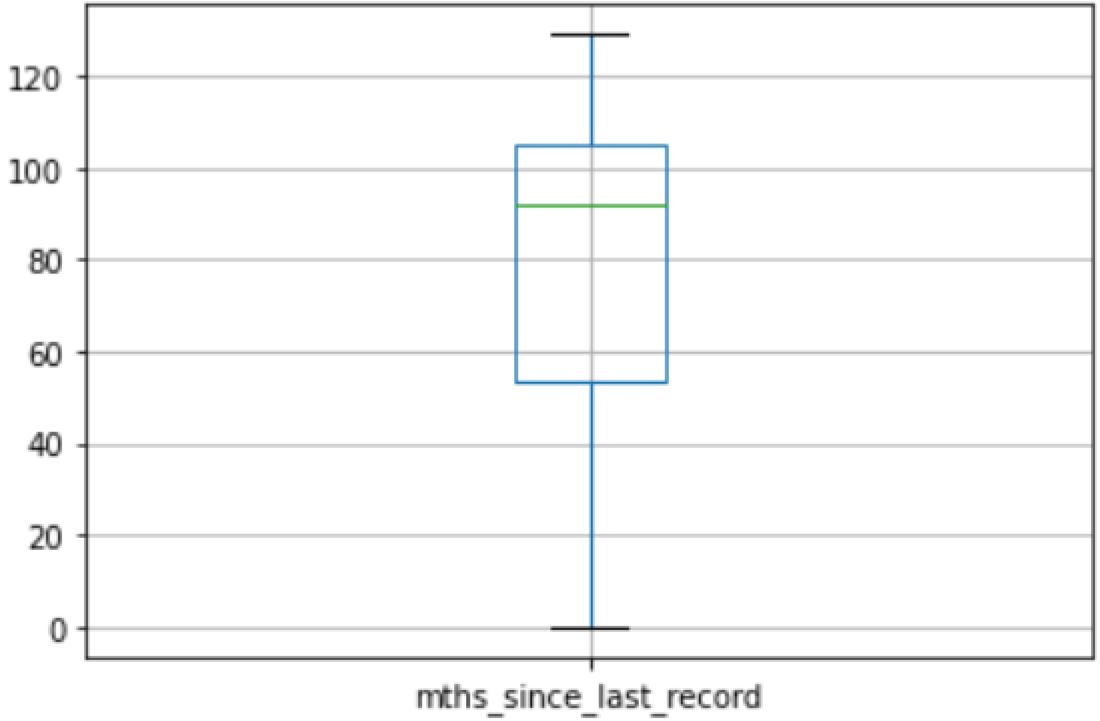
DTI



Delinq_2yrs



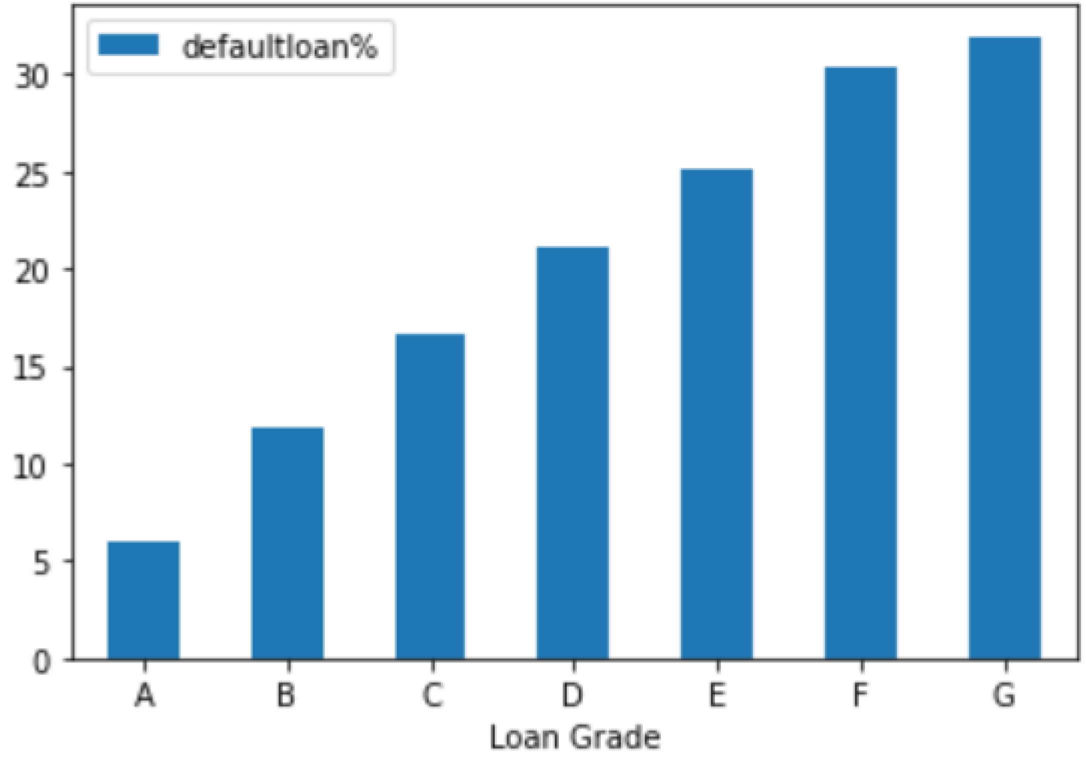

100
80
60
40
20
0
mths_since_last_delinq



120
100
80
60
40
20
0
mths_since_last_record

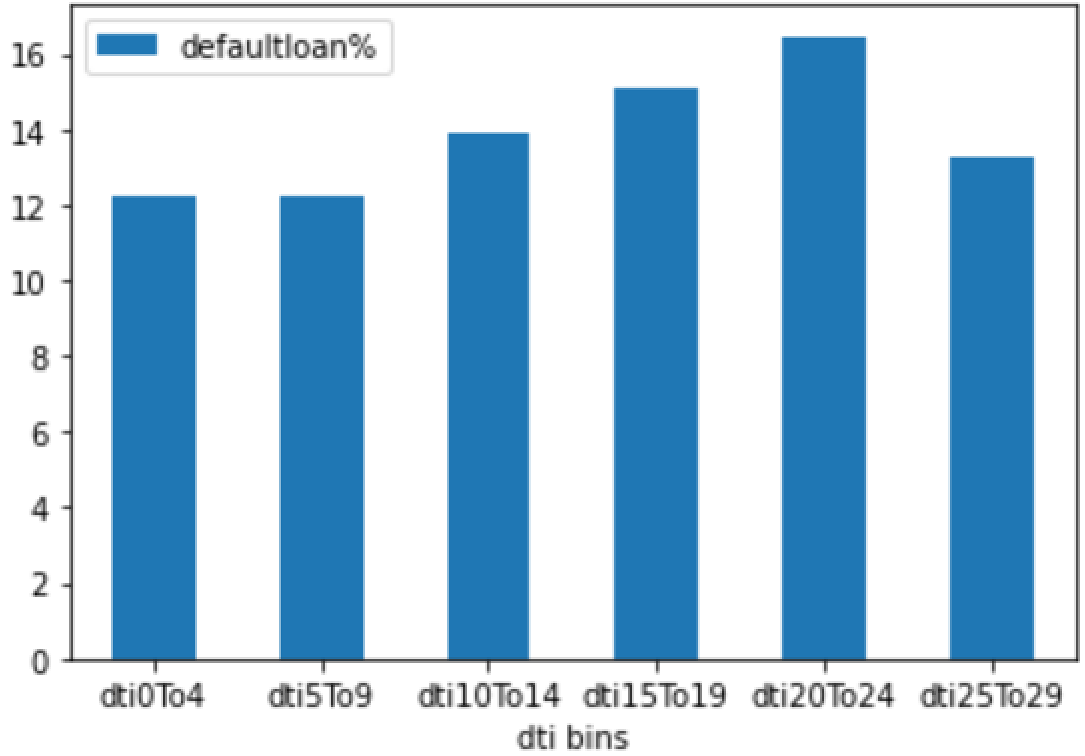
Bivariate Analysis – Loan grade <-> loan default %

Loan grade seems to impact loan default %, as high as 30% for G grade



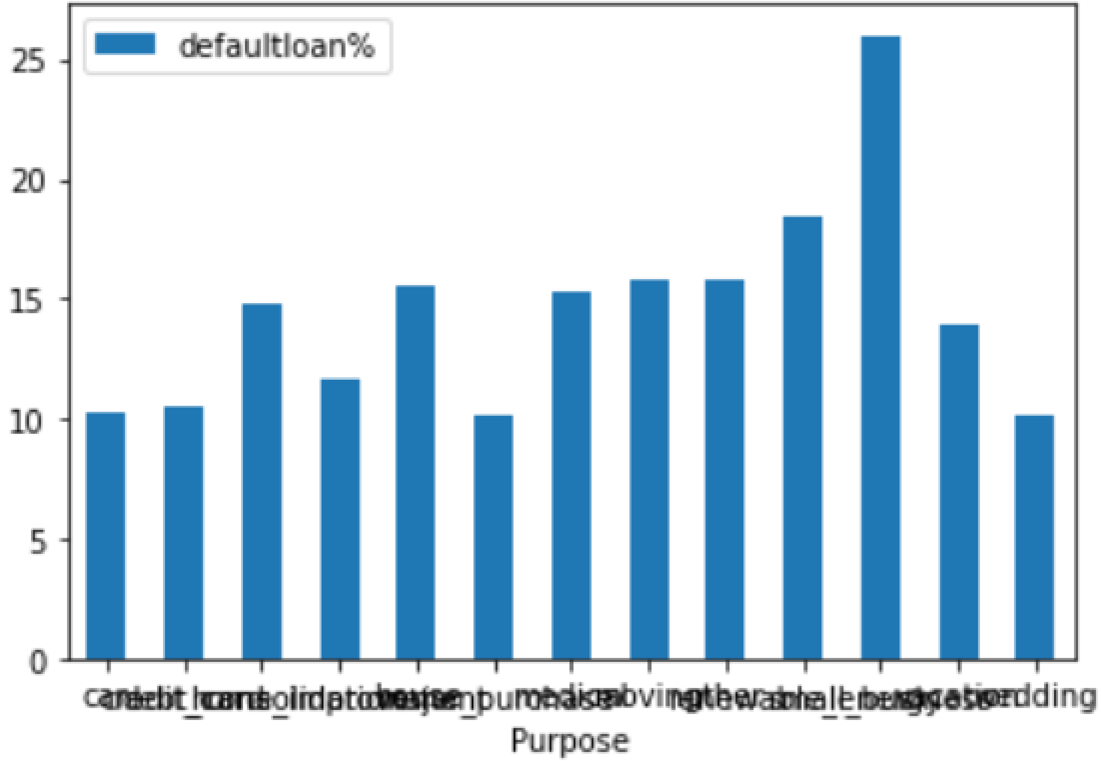
Bivariate Analysis – dti <-> loan default %

Dti grouped/binned – dti were grouped by internal of 5, and then for each interval/group total defaulter % was calculated against total rows in that group, higher dti seems to result in more defaulters except last group of 25-29



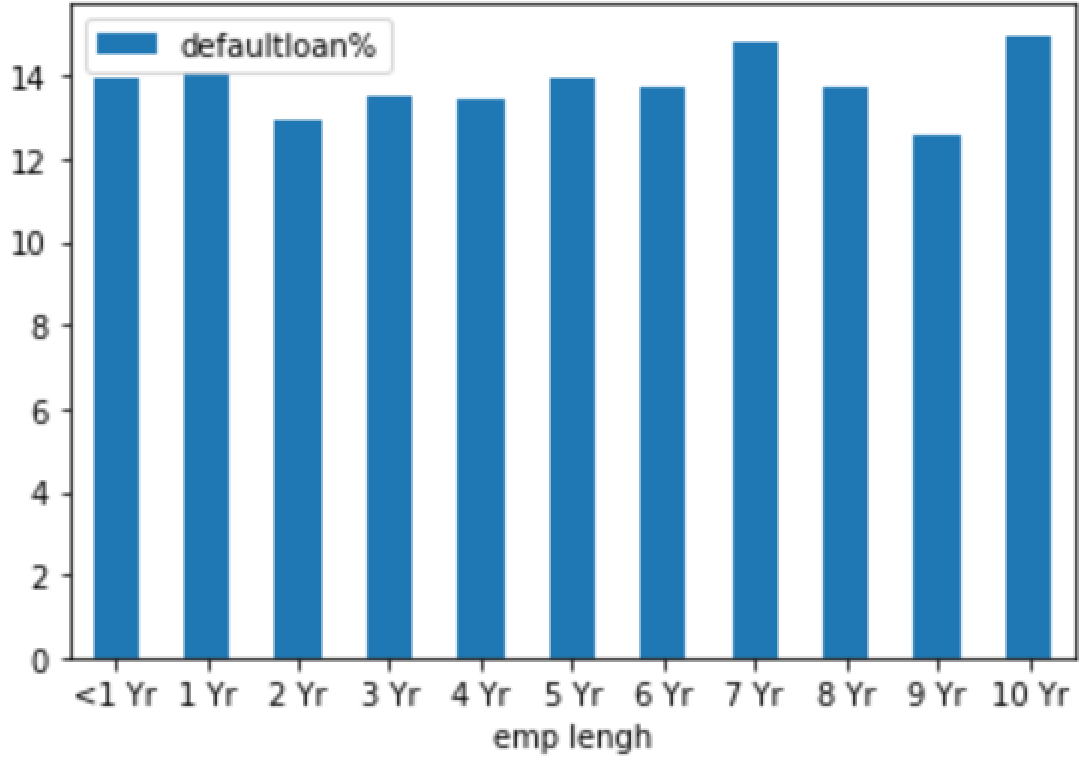
Bivariate Analysis – Home ownership <-> loan default %

- Purpose seem to impact loan default, small business loans resulting in default as much as 25%, y axis titles not visible, but could figure out the names from order in which they were displayed



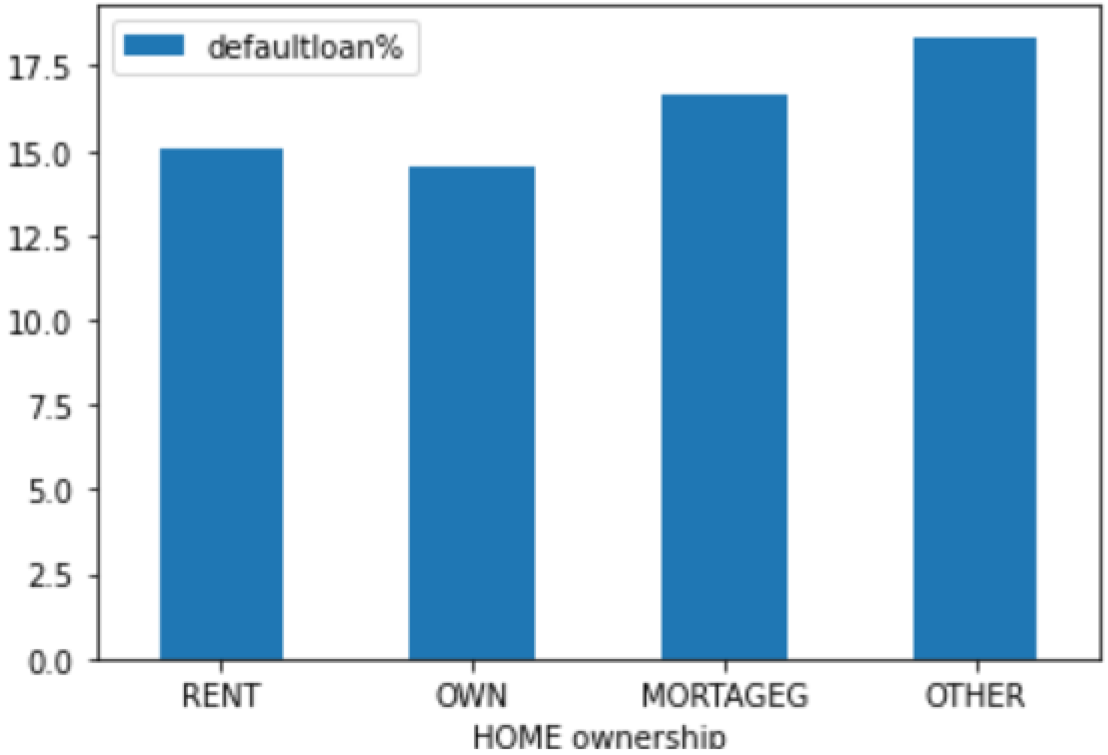
Bivariate Analysis – emp length <-> loan default%

- Emp length does not seem to impact loan default

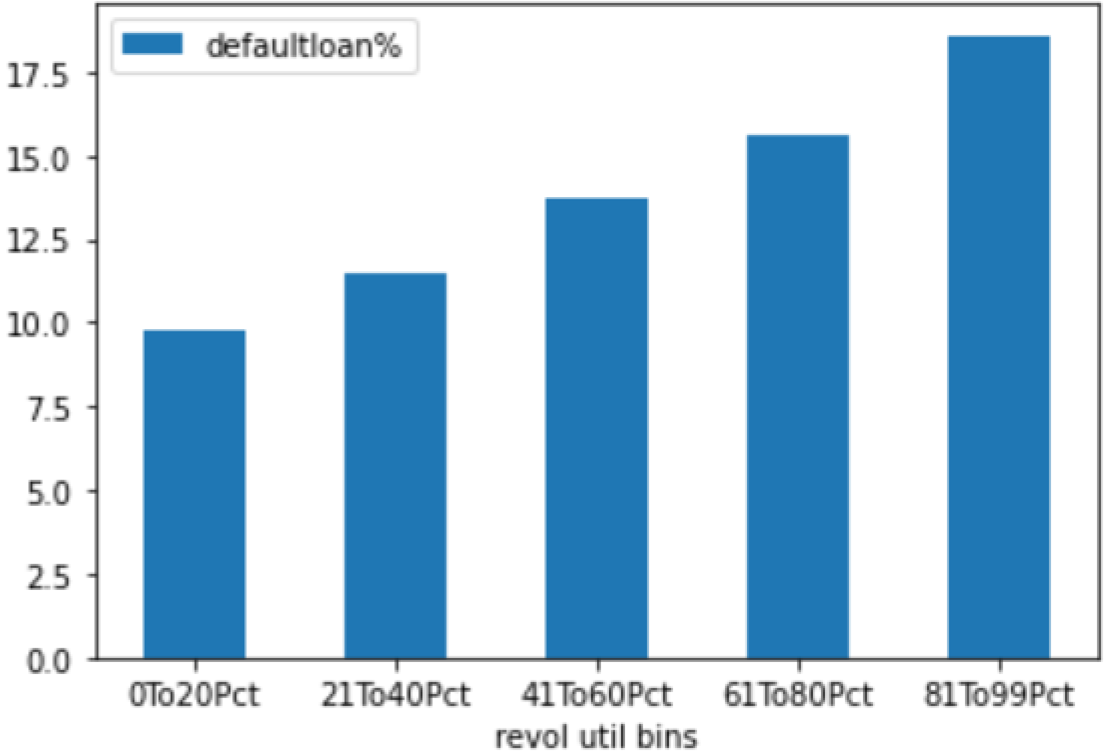


Bivariate Analysis – Home ownership <-> loan default %

- Home ownership does not seem to impact much except OTHER category



Bivariate Analysis – revol util binned <-> loan default %



Summary

- Higher loan grades have tendency for more loan default
- Loan purpose of small business, renewable_energy debt_consolidation, house, medical, moving, other seem to have more loan default % as compared to remaining, in the order of listing, small business seems to be having highest loan default %
- Higher Revol has more tendency for loan default
- increase in dti seems to corelated with higher changes of loan default