Lending Club Case Study

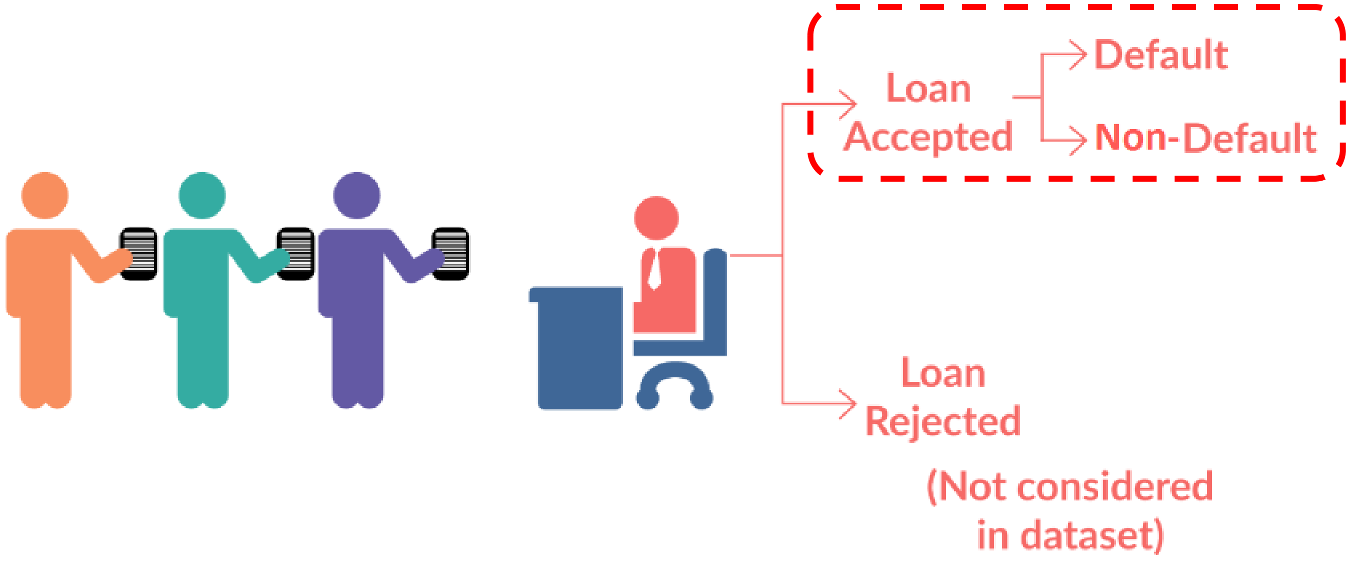
Prepared By: Ashutosh V.
Date: 03 Mar 2024

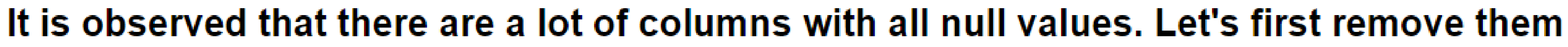
Problem Statement:

At consumer finance company, when the company receives a loan application, to make a decision for loan approval based on the applicant's profile.

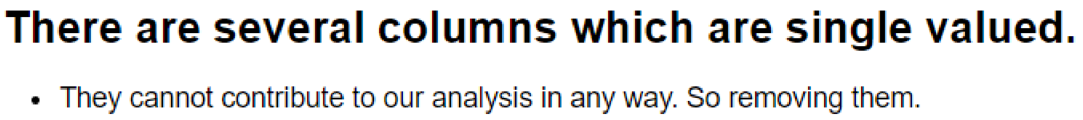
Loan application is accepted, now I have to find if the applicant will be default or not based on the loan data provided.

LOAN DATASET





It is observed that there are a lot of columns with all null values. Let's first remove them



There are several columns which are single valued.

- They cannot contribute to our analysis in any way. So removing them.

Now we have 48 columns out of which some correspond to the post approval of loan

•We are analyzing the user details and the driving factors of loan defaulting before approving loan.

•So we can safely remove the columns / variables corresponding to that scenario.

•Also there are some columns such as "id", "member_id", "url", "title", "emp_title", "zip_code", "last_credit_pull_d", "addr_state".

•The above features or columns doesnt contribute to the loan defaulting in any way due to irrelevant information. So removing them.

•"desc" has description (text data) which we cannot do anythhing about for now. So removing the column.

•"out_prncp_inv" , "total_pymnt_inv " are useful for investors but not contributing to the loan defaulting analysis. So removing them.

•"funded_amnt" is not needed because we only need info as to how much is funded in actual. As we have "funded_amnt_inv" , we can remove the earlier column.

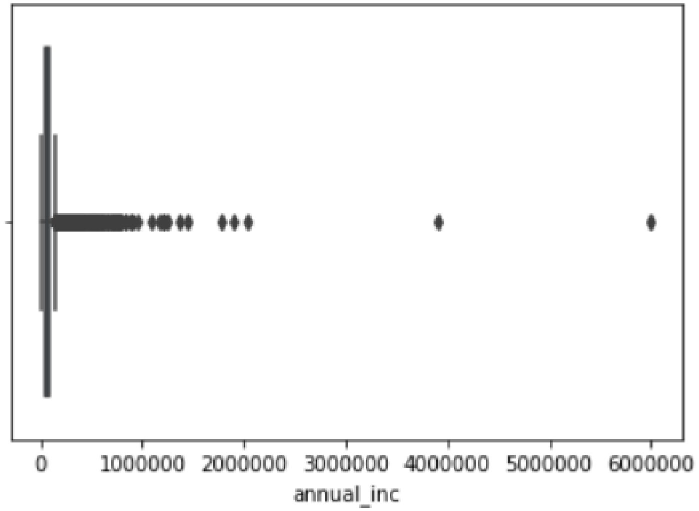
Handling Missing values

- columns with missing values are "emp_length", "revol_util".
- So before doing that, lets see what kind of data each column has.

Standardizing the data

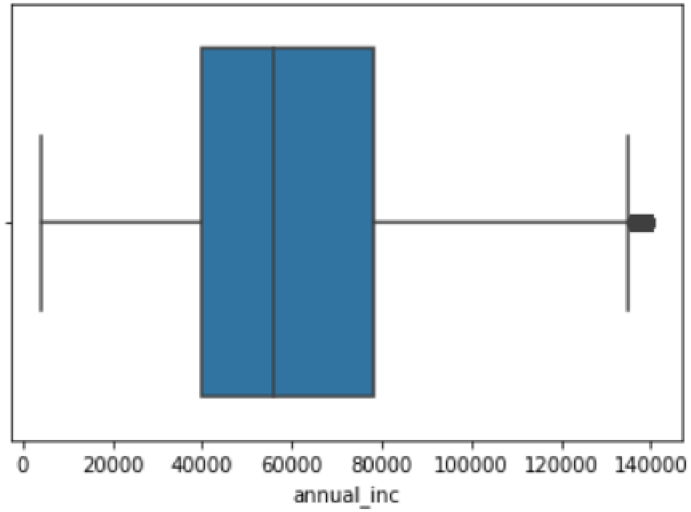
- "revol_util" column although described as an object column, it has continous values.
- So we need to standardize the data in this column
- "int_rate" is one such column.
- "emp_length" --> { (< 1 year) is assumed as 0 and 10+ years is assumed as 10 }
- Although the datatype of "term" is arguable to be an integer, there are only two values in the whole column and it might as well be declared a categorical variable.

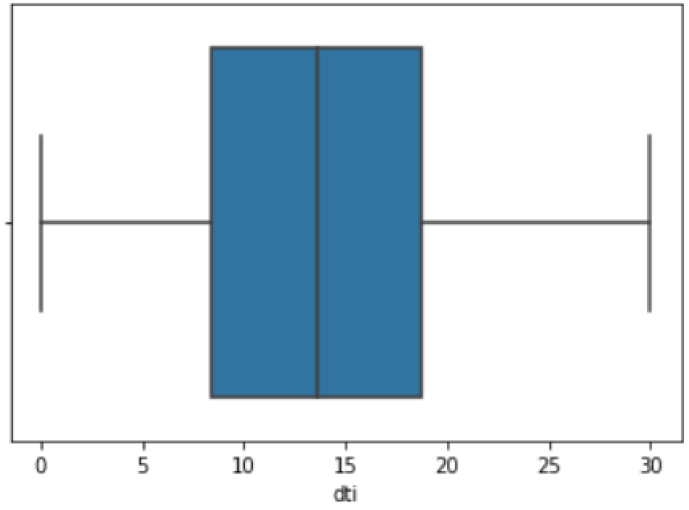
Outlier Treatment

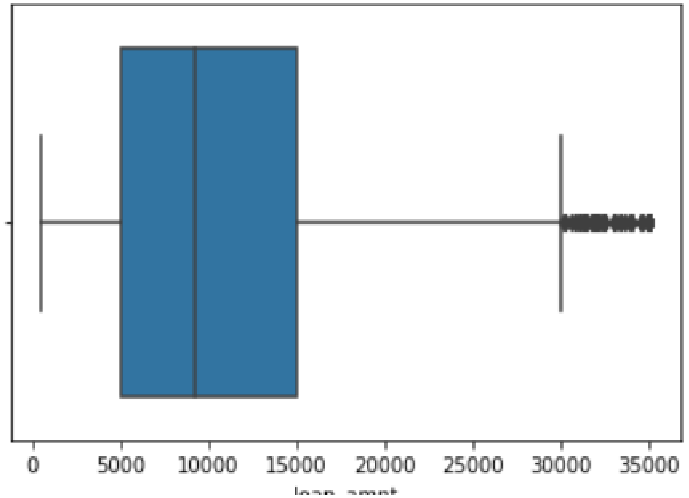


Clearly indincating the presence of outliers.

- So, Removing them.
- Let's see the quantile info and take an appropriate action.
- The values after 95 percentile seems to be disconected from the general distribution and also there is huge increase in the value for small quantile variation.
- So, considering threshold for removing outliers as 0.95

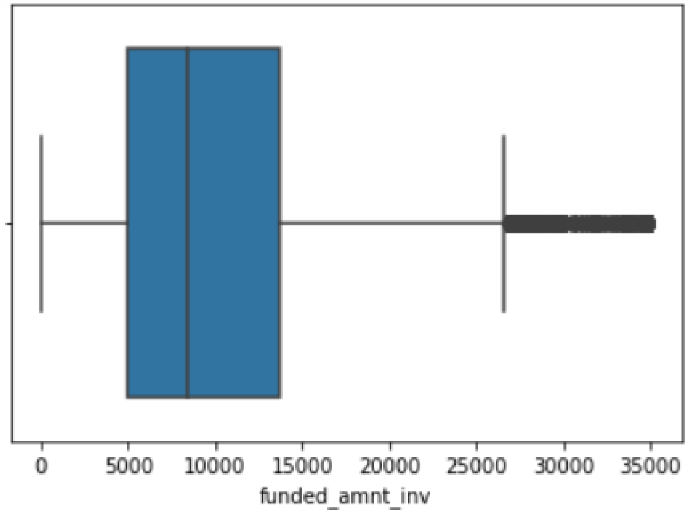






Now the "annual_inc" data looks good and proceeding next.

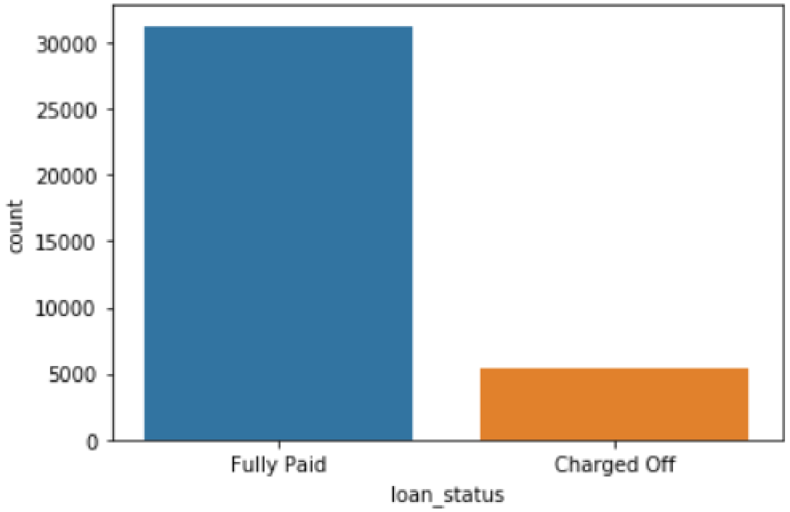
- Let's analyze other numerical variables which could possibly have outliers.
- dti
- loan_amnt
- funded_amnt_inv

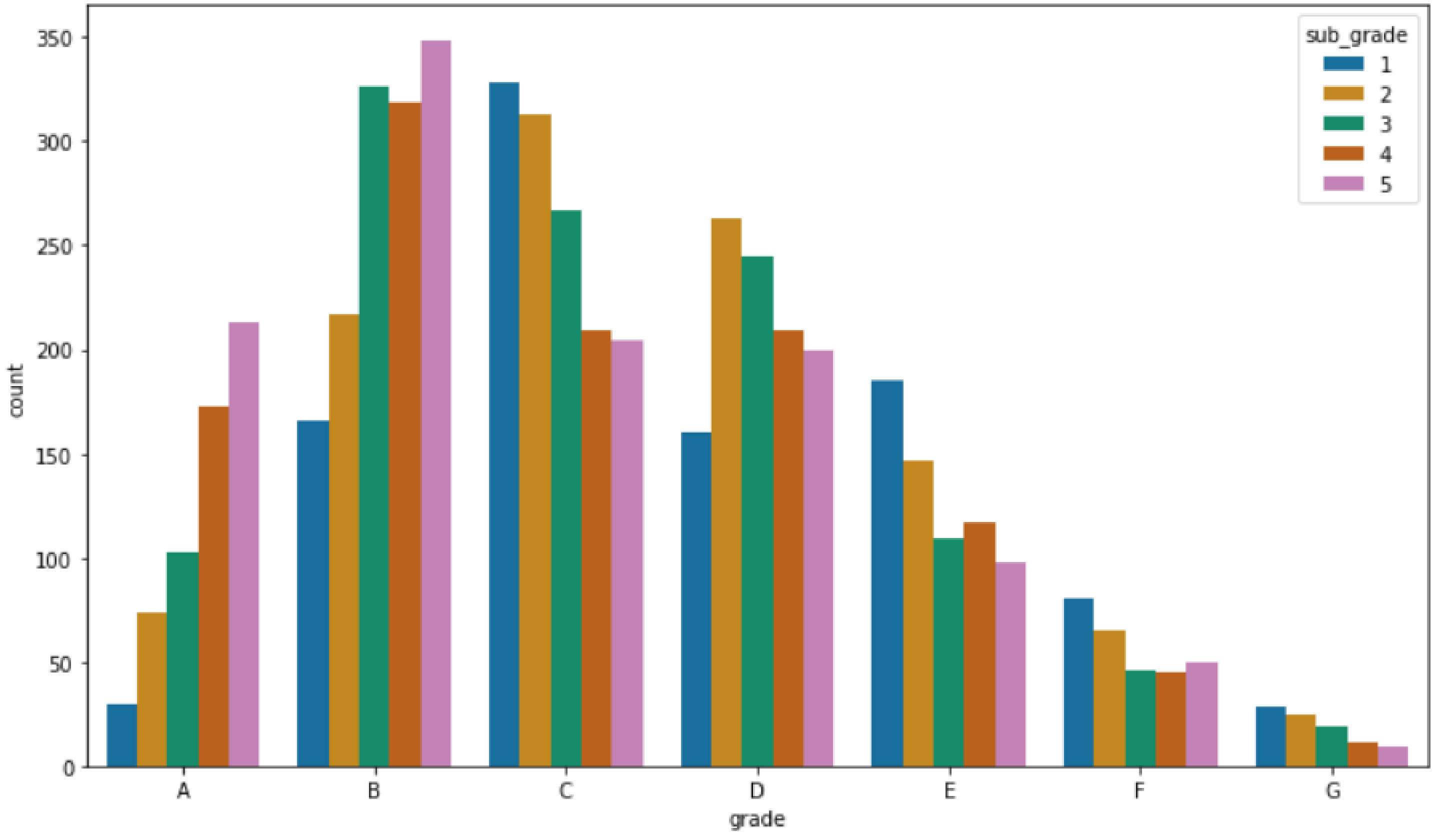


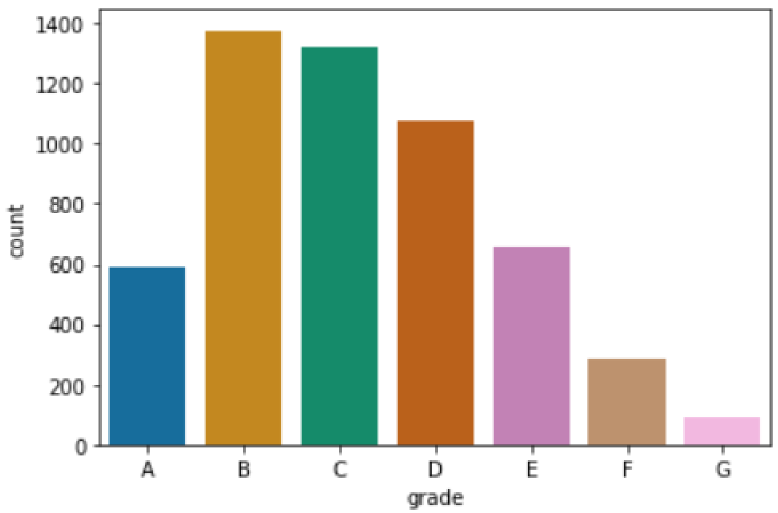
Visualizing Categorical Data

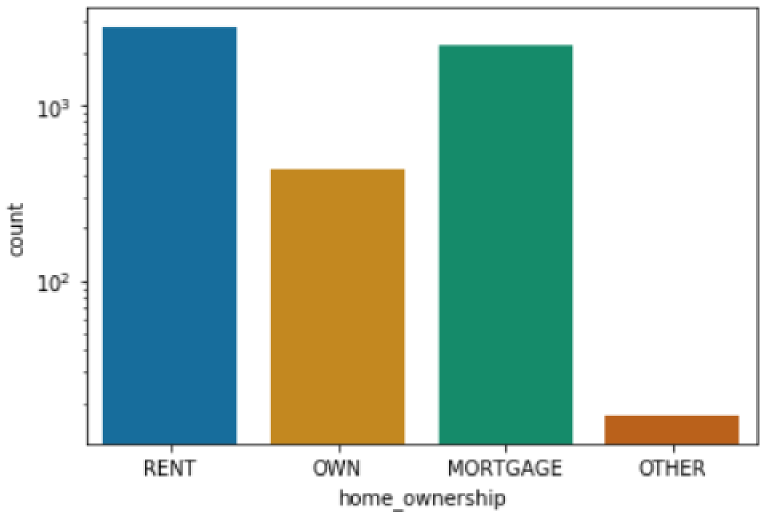
As we already have grade column, extracting only subgrade (int level value) from the sub_grade variable

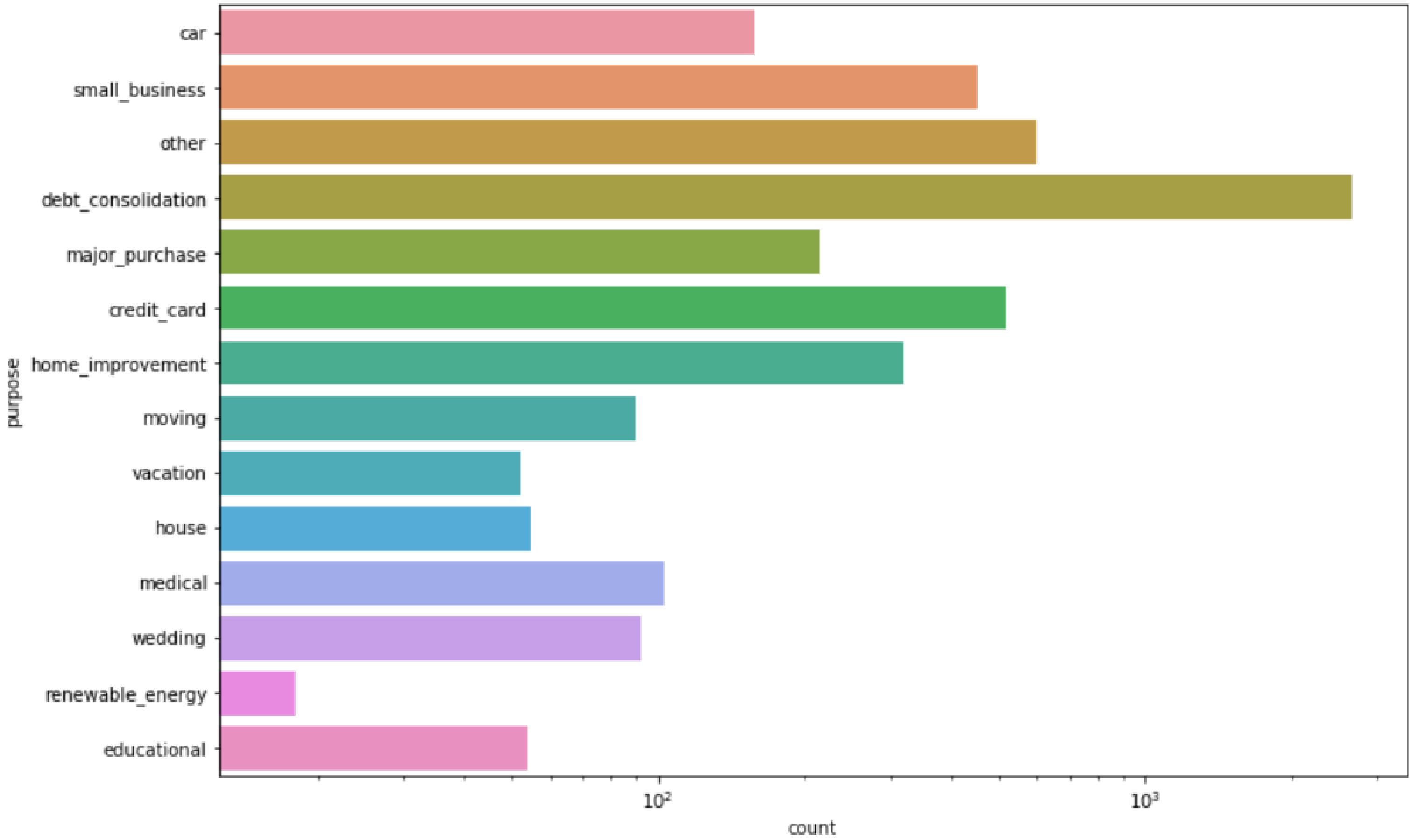
- We are analyzing and visualizing only the defaulter data. So subsetting the data while plotting only for 'Charged Off' loan_status for below plots








count
10^3
10^2
RENT
OWN
MORTGAGE
OTHER
home_ownership


purpose
car
small_business
other
debt_consolidation
major_purchase
credit_card
home_improvement
moving
vacation
house
medical
wedding
renewable_energy
educational
10^2
10^3
count

Applicants who applied and defaulted have no significant difference in loan_amounts.

- Which means that applicants applying for long term has applied for more loan.

Observations

The above analysis with respect to the charged off loans. There is a more probability of defaulting when :

- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %
- Applicants who have taken a loan for small business and the loan amount is greater than 14k
- Applicants whose home ownership is 'MORTGAGE and have loan of 14-16k
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%