CREDIT EDA CASE STUDY

NITIN HUDA

TEJALKAUR GULATI

Data cleaning and imputation

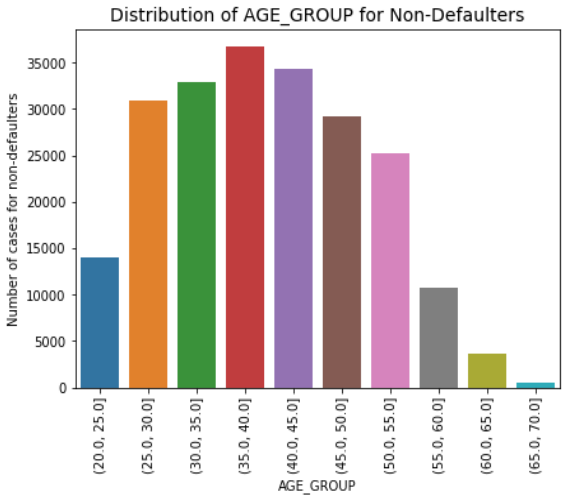
- Total no of columns with missing values more than 50%:41
  - Steps:- Deleted columns from dataset
- Total no of columns with missing values less than 13%
  - Steps:- Data imputation
- Data imputation methods used

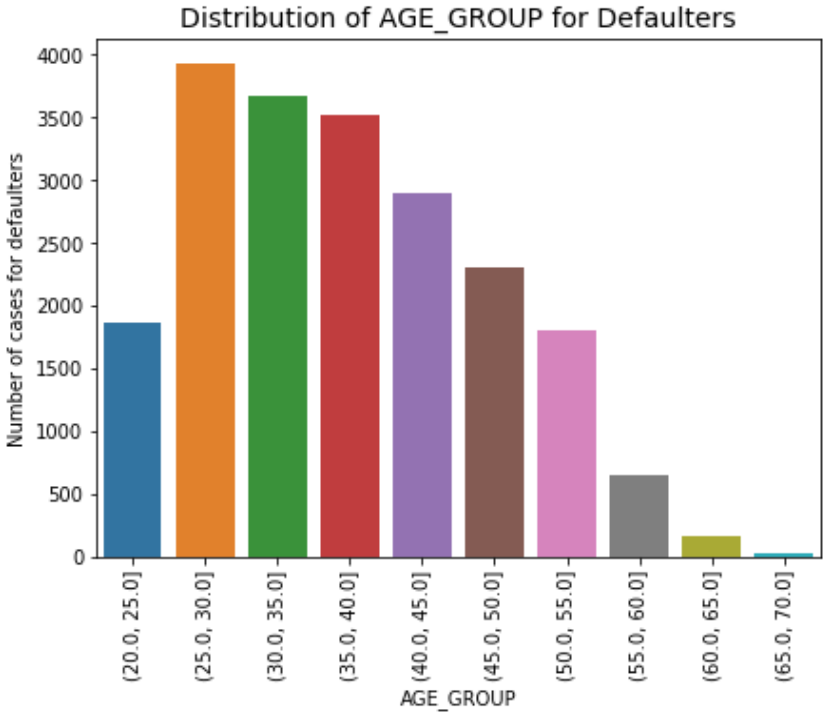
1. For categorical variables:- Used value with the highest frequency
2. For numerical variables:- Used mean/median value as per the data.

- Data cleaning and formatting
  - Changed data type from float to integer for discrete values
  - Changed negative values to absolute values
  - Changed default values to null

Numerical to categorical

- Age and employment days are to convert the data into categorical data for further analysis.





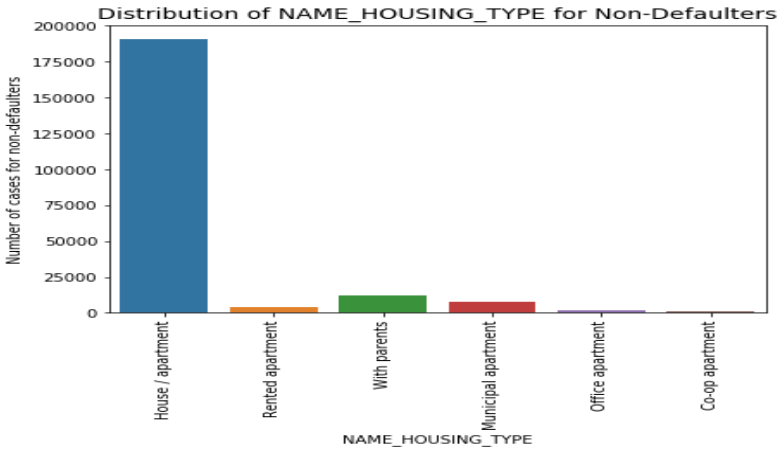
Data imbalance

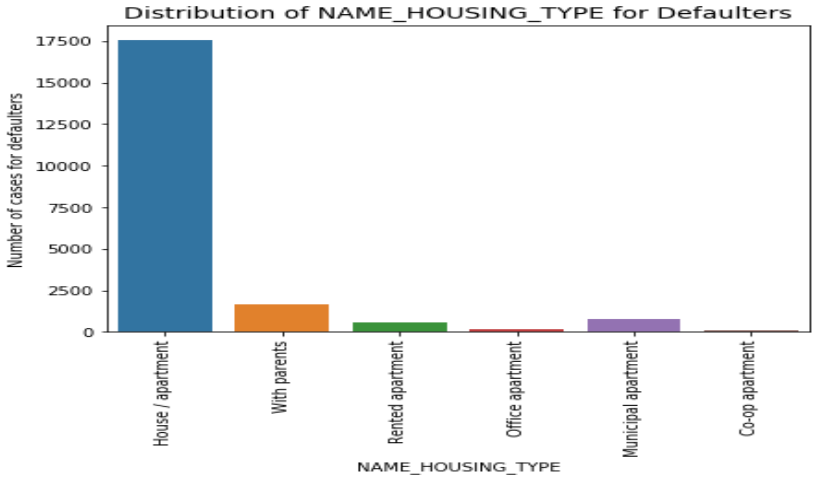
- The data is cleanup which is highly imbalanced with around 8.35% data for loan defaulters(target=1) and remaining for non-defaulters with around 91.65%.
- Data sampling technique can be used to do over sampling/under sampling which can reduce the bias introduced due to imbalance.

Segmented univariate analysis

•Few observations for segmented univariate analysis on Family status, housing type and education type.

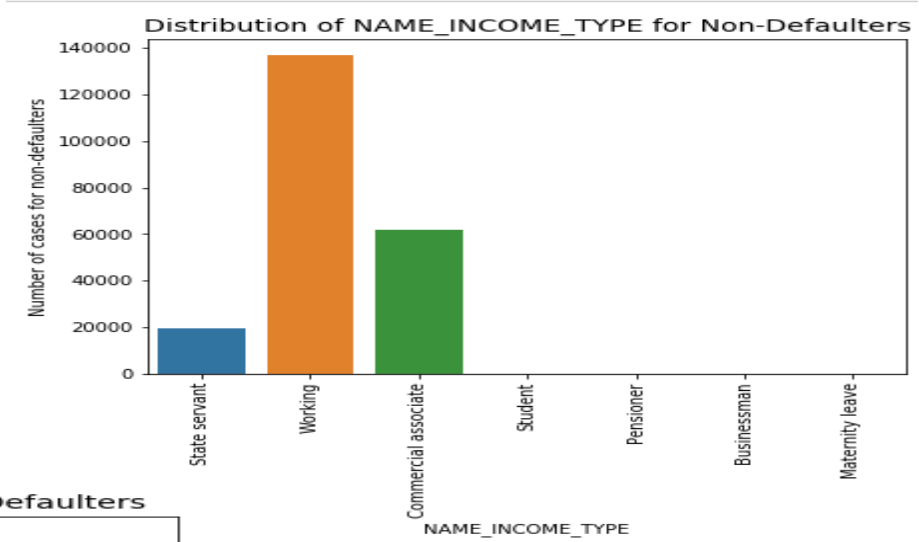
- It is observed customer living with parents have little more proportion of defaulting compared to non-defaulters.
- Likewise, rented apartment and municipal shows slowly higher proportions towards defaulting.

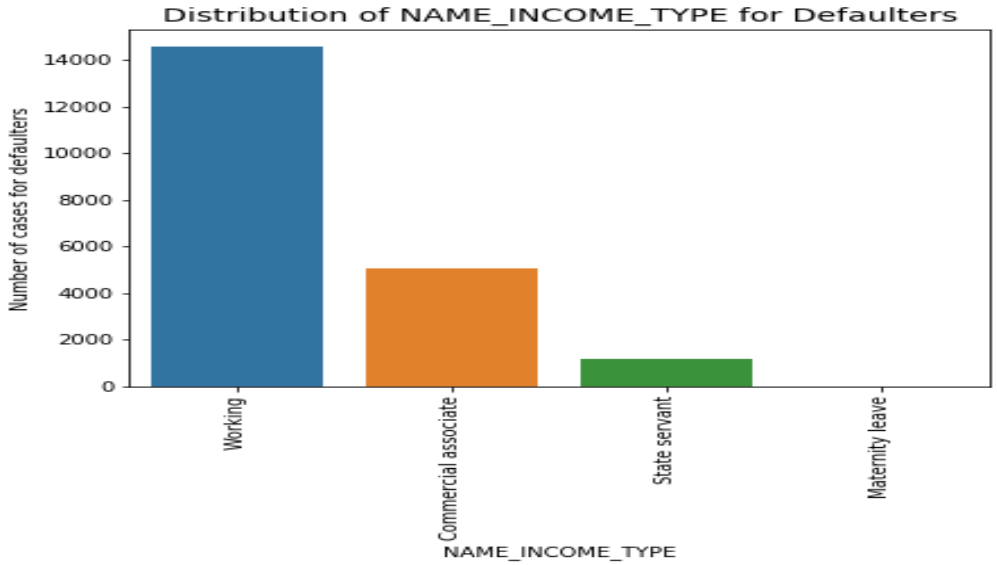




Segmented univariate analysis

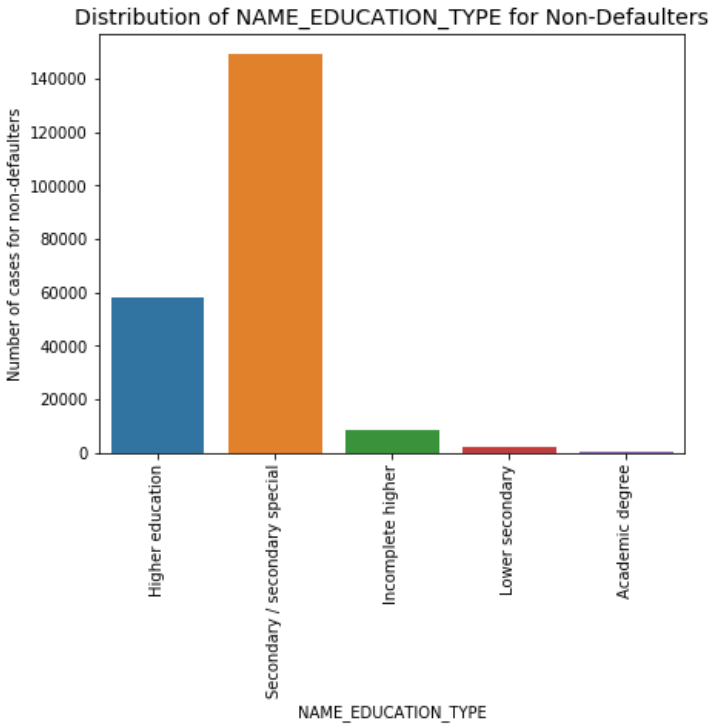
- Customers who are currently working have higher proportion of defaulters.
- Pensioners seems to be pay back loan, so their proportion is lesson defaulters .
- State servants are comparatively show less tendency towards defaulting.

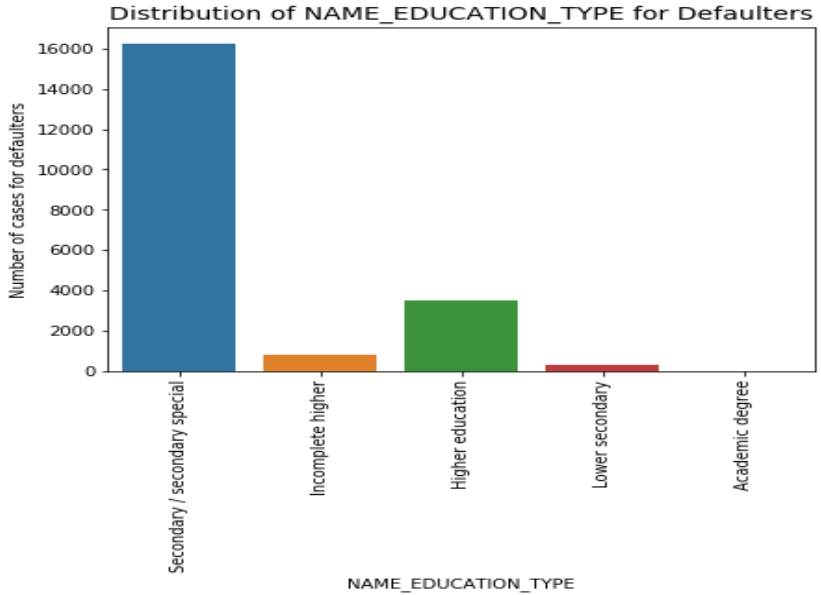




Segmented univariate analysis

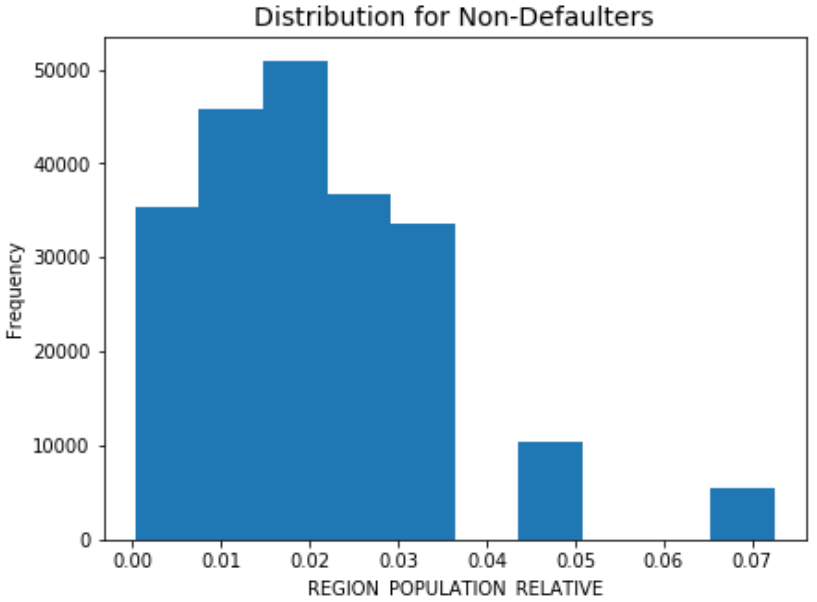
- Customers with Secondary education have high proportion of defaulting if compared to non-defaulters
- Customers with higher education tend to default less as their proportion is reduced.

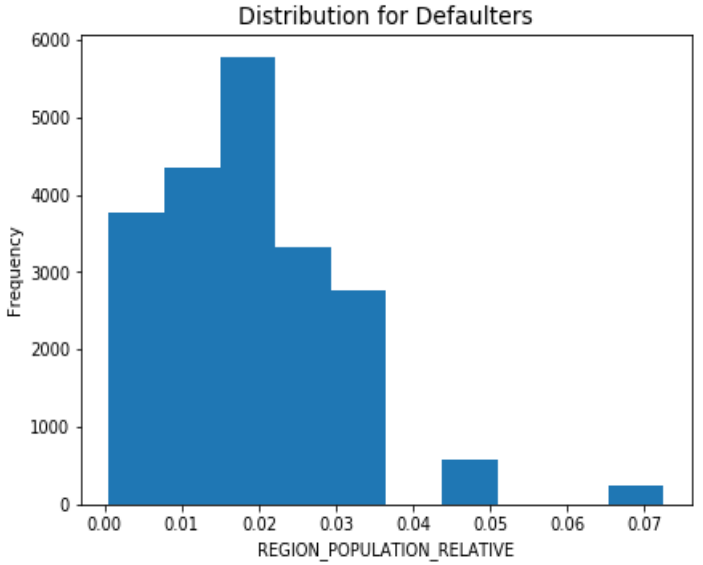




Univariate analysis

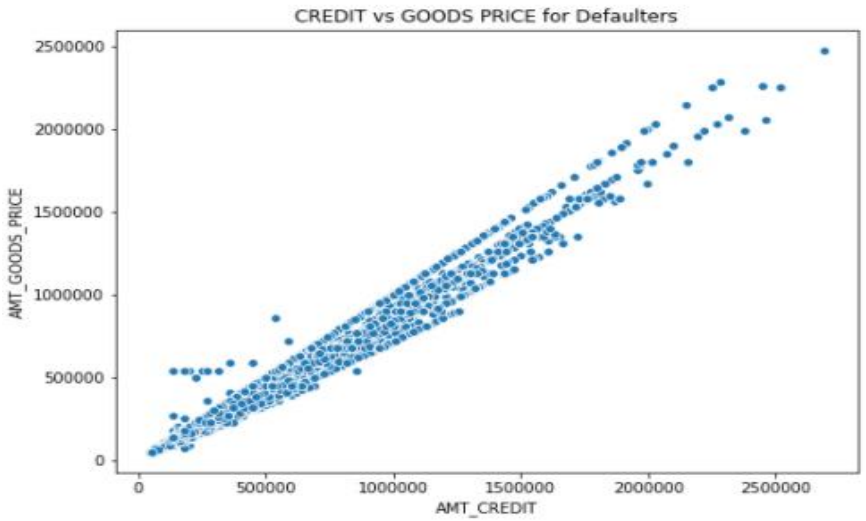
- Observations for the univariate analysis for the population were proportion of defaulters is more as compared to non-defaulters.





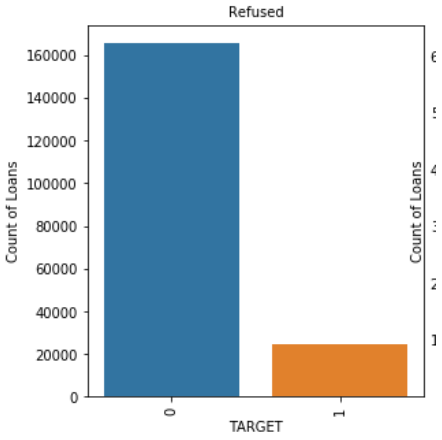
Bivariate Analysis

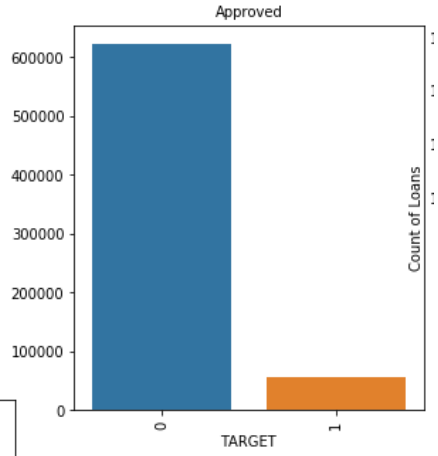
- AMT_GOODS_PRICE and AMT_CREDIT
- The correlation between property price and loan amount of non-defaulters is 0.9816 but for defaulters is 0.9776.
- Credit amount and goods price are highly correlated variables for both defaulters and non-defaulters. So as the home price increases the loan amount also increases.

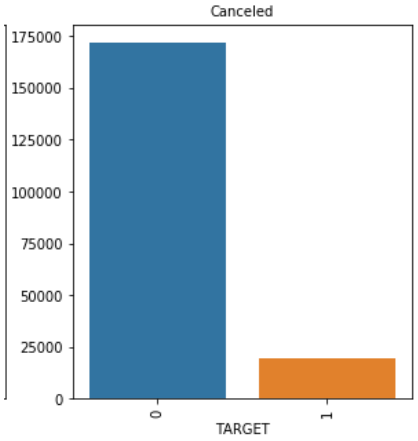


Previous application data analysis

- The previously refused % of applications for non-defaulters is 16.75%.
- The previously refused % applications for defaulters is 23.96
- Below is the distribution of contract status divided in target 0 and target 1.
- The distribution of target 1 is maximum for refused state and target 0 for approved stage.







Thank You