Problem Statement

For the banking data set case study, the problem statement is to identify the potential defaulters/applicants who are struggling to repay the loan.

So that bank can take decisions based on the Detailed EDA on the loan customer dataset.

Few of the major salient points are mentioned below:

1. Based on the thorough analysis, bank will decide whether to provide loans or to reduce the loan amount or to increase the interest rates and so on.

2. Bank can also decide based on the previous loans repayments attemepted by consumers and its respective installments.

3. It ensures applicants application should not get rejected and

4. At the same time, bank can make profits.

5. It will also help the bank to understand the potential defaulters and provides valid reasons to reject the loan application.

Assumptions:

As far as the EDA data analysis is concerned, there are lots of assumptions fetched or made from the entire case study which will eventually help the bank to take business decision:

- I have looked into the income type of the applicant which states the financial condition.
- I have analysed the applicant marital status and their earning capabilities which further drill downs to the risk of taking/givng loans.
- I have also analysed the companies in which the applicant is working and for how long.
- Previous loans repayment on timly installation also clear the air of the consumer financial well being.
- Age group is also thoroughly stufied to understand the minimal risk asociated with it.

Approaches and Methodologies

Approaches:

- ➢ 1. Started of with understanding and analysing the multiple variables/attributes of the datasets.
- ➢ 2. Identifying the loop holes and resolving them one by one.
- ➢ 2. Post that data clean up/ imputing activities which helps to get rid of data polution.
- ➢ 3. Keep the data precise and clean for better hypothesis and analysis.
- ➢ 4. Ensuring no null values and missing values on the continuous numerical columns and catergorical nominal columns.
- ➢ 5. Perform univariate analysis using different graphs and plots.
- ➢ 6. Followed by performing bivariate analysis to understand the behaviour of the data when shows up in the charts and graph format.

Methodologies

Methodologies includes :

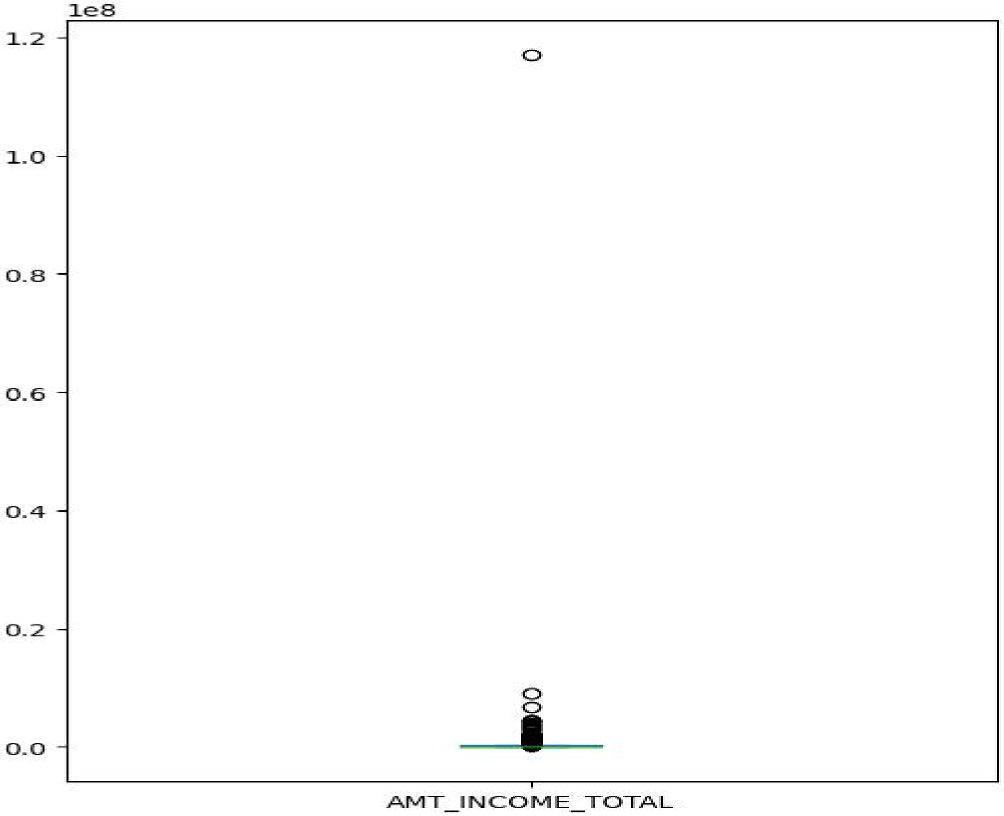
- First thing is getting rid of missing values and null values from the rows and columns.
- Dropping the columns which has more than 40 % of null/missing values.
- Deleting the rows from those attributes which has less than 40 % of missing values
- Or replacing them with the mean or mode value of that particular column data.(Numerical data)
- When there are no null values in any of the attributes then perform visualization method to undestand the correlation between two variables.
- For that I used, bar chart, pie chart, box plot, hist plot, scatter plot, count plot and heat map.
- I used numpy and pandas library to recite the code.
- Seaborn library is used to add aesthetic outlook to the entire case study.

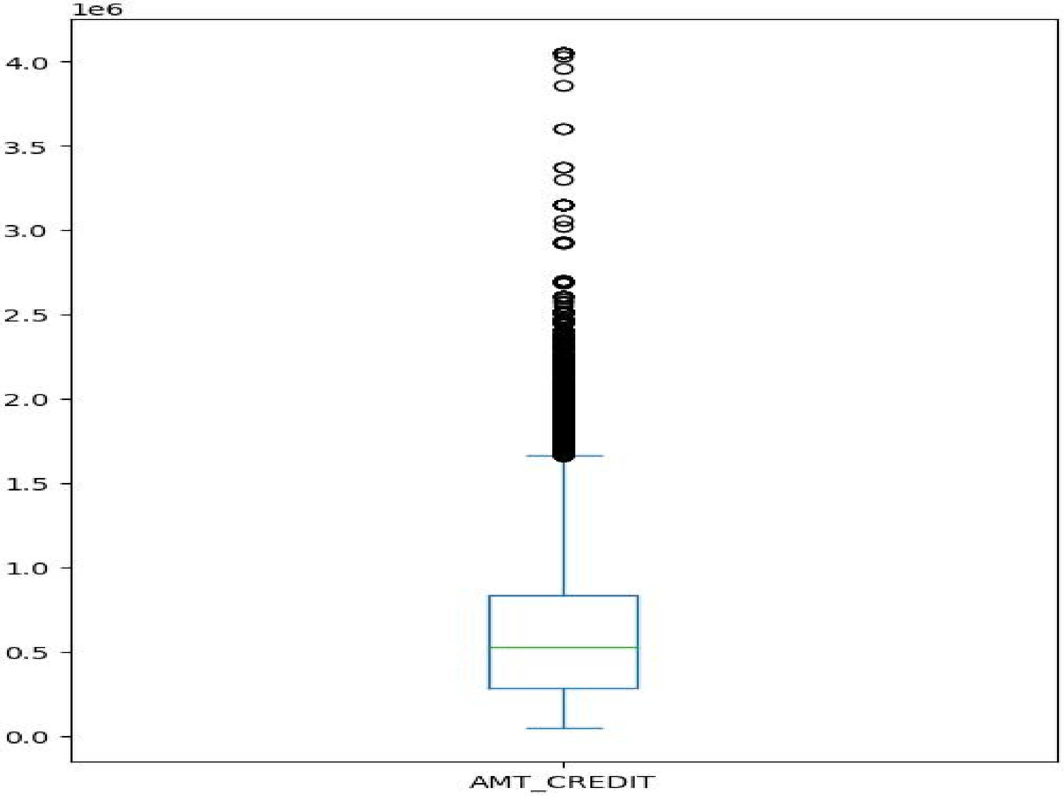
Graphs and insights

Multiple graphs are created to understand the relation ship between two or more variables.

In univariate analysis,behaviour one variable acorss the target atttribute is depicted in the below graphs. Few of those are mentioned below:

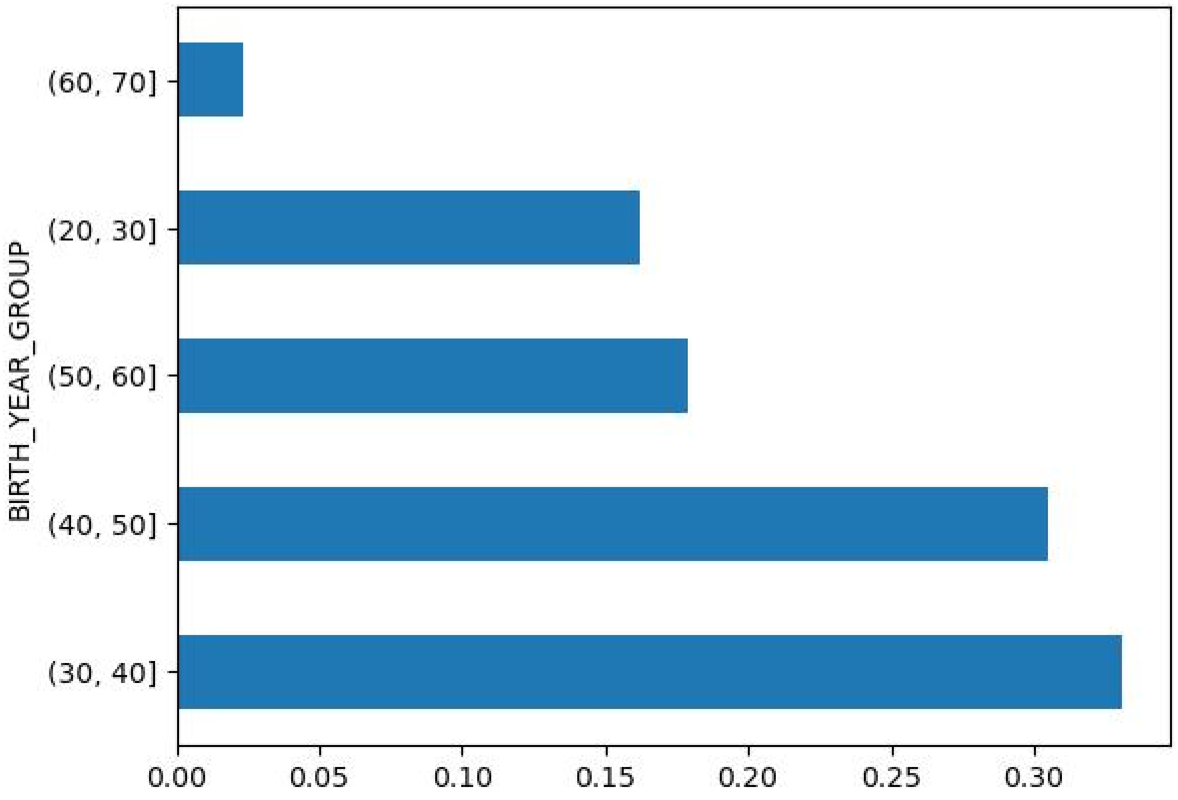
1. Outliers in first dataframe - application_dataset - (AMT_INCOME_TOTAL, AMT_CREDIT)

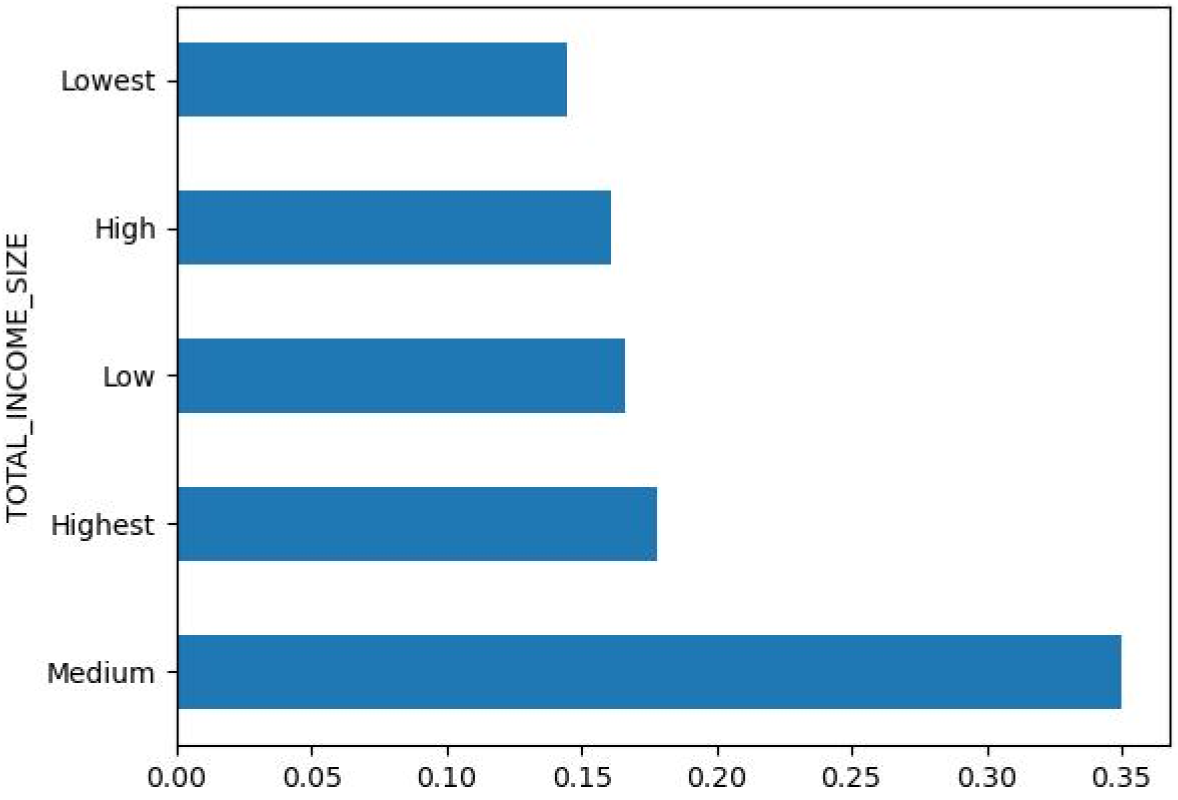




Graphs and Insights

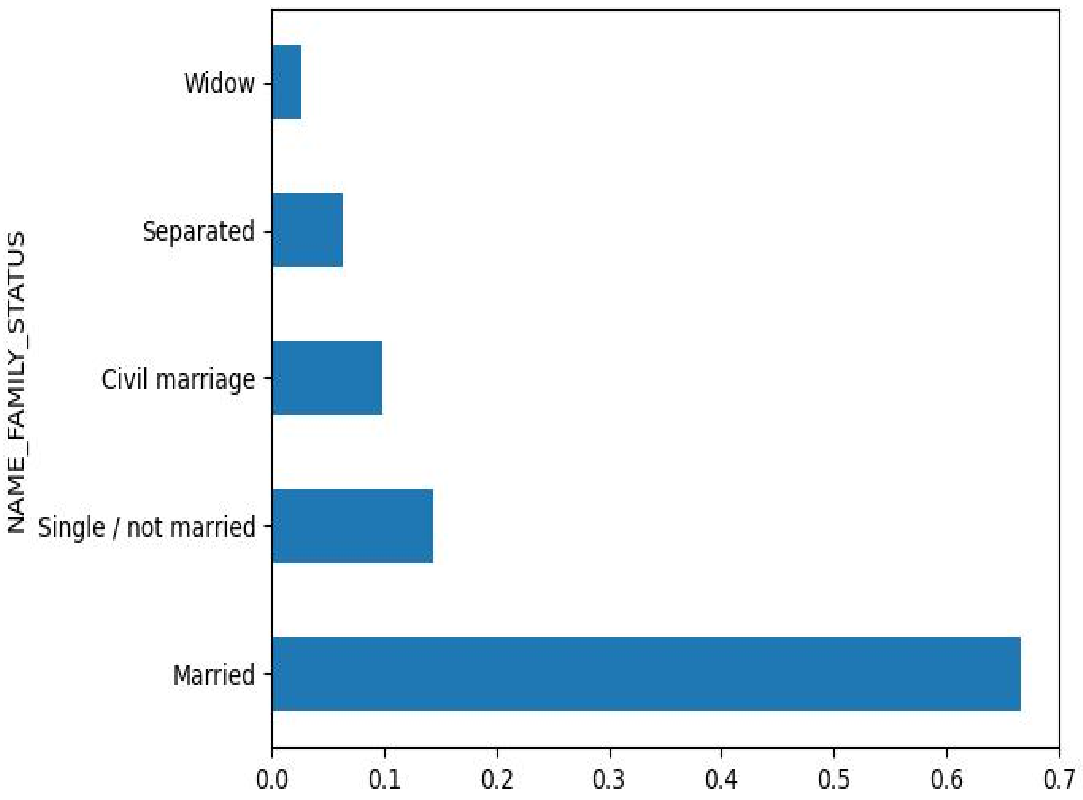
Creating buckets for age group and income group:

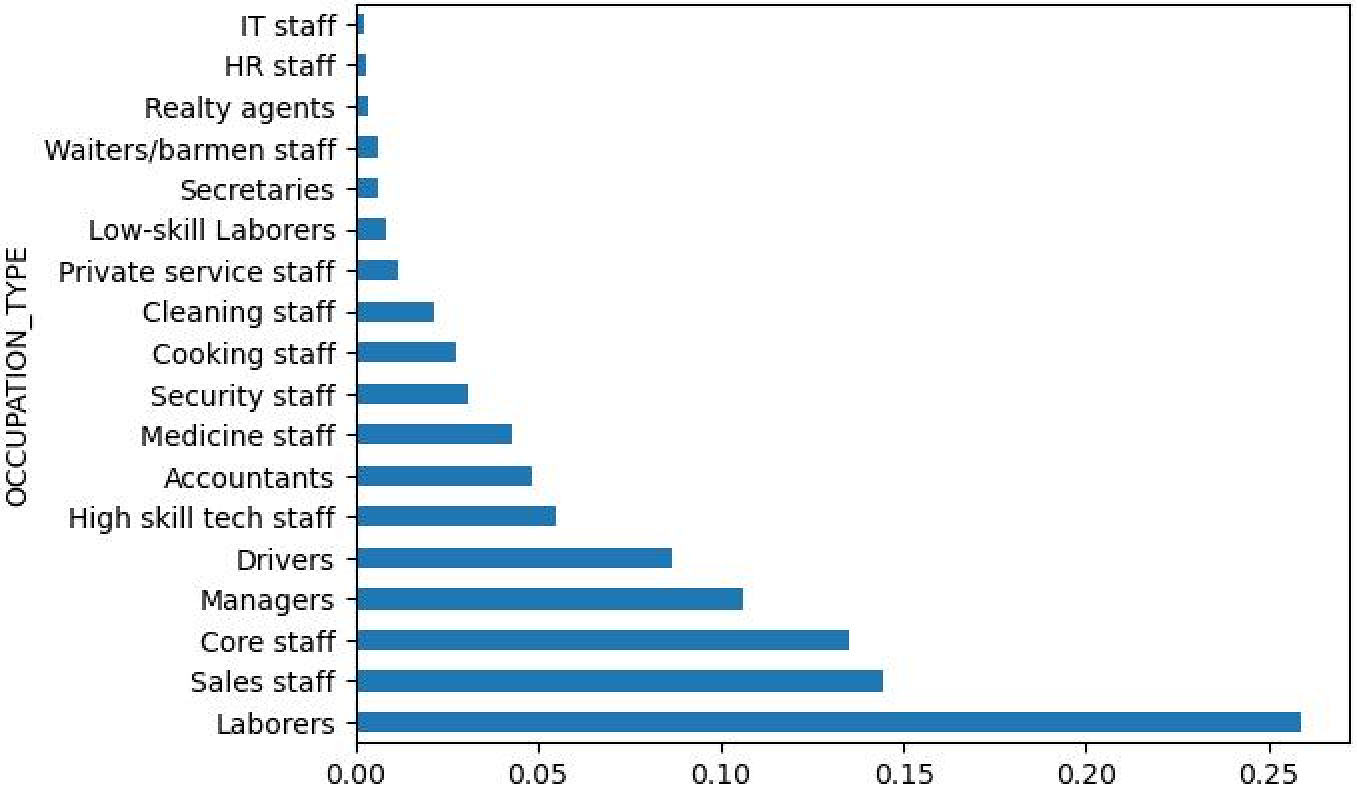




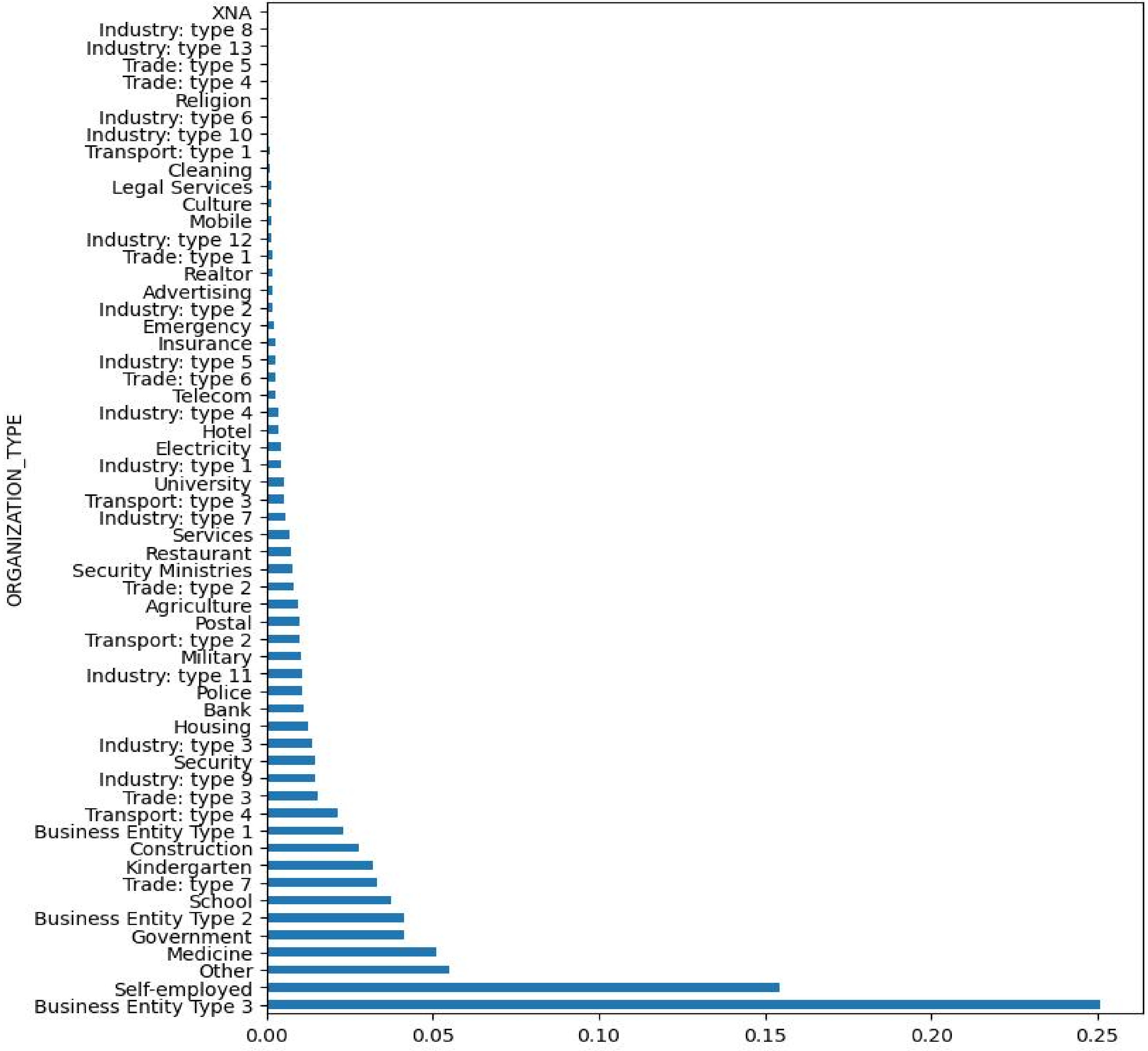
Categorical unordered univariate analysis

1. Type of loan taken by a person = home, personal, auto etc.
2. Organisation= Sales, marketing, HR etc.
3. Occupation type category
4. NAME_FAMILY_STATUS status
5. Housing type of target variable = Own/rented



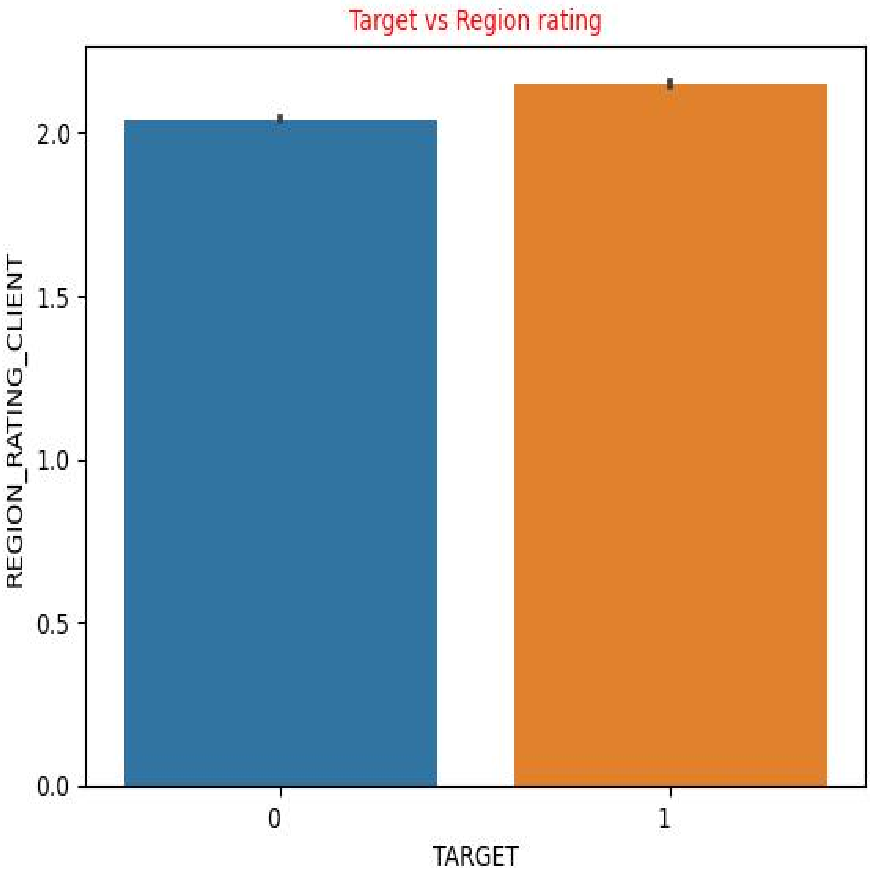


Organization:



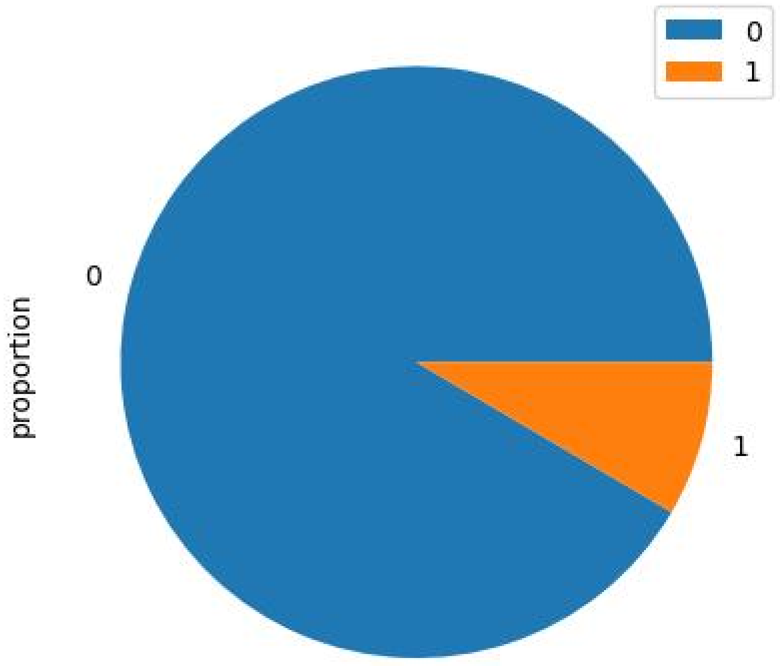
Region Specific Analysis for target variable

Plot the bar chart for the Region rating category vs target



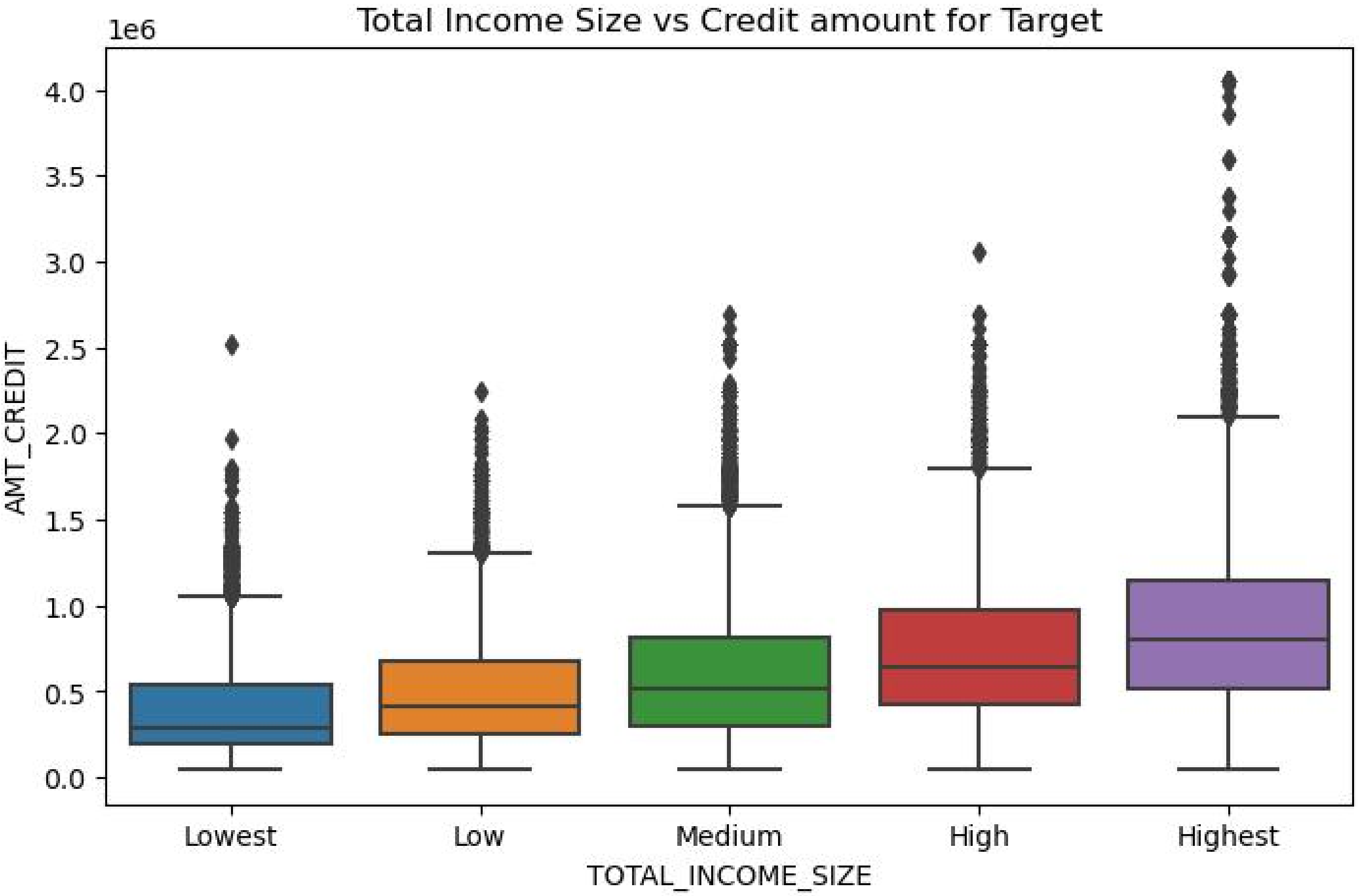
Target variable/defaulter

- plot the pie chart of target/defaulter categories



Bivariate Analysis: Categorical - Numerical Variable

Analysing Income Group vs Credit amount of the loan variable in df dataframe



Conclusion:

All the below variables were concluded while performing analysis of Application dataframe as leading to default.

Checked these against the Approved loans which have defaults, and it proves to be correct

-Medium income

-25-35 years olds , followed by 35-45 years age group

-Male

-Unemployed

-Own House - No

few other important points to be considered

-Previous applications with Refused, Cancelled, Unused loans also have default which is a matter of concern.