EDA Credit Assignment
by
Shubham Sharma

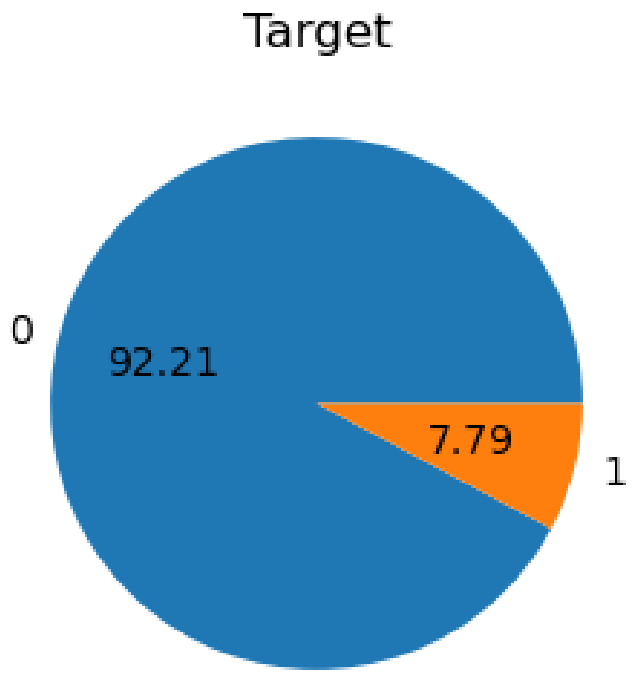
Problem Statement

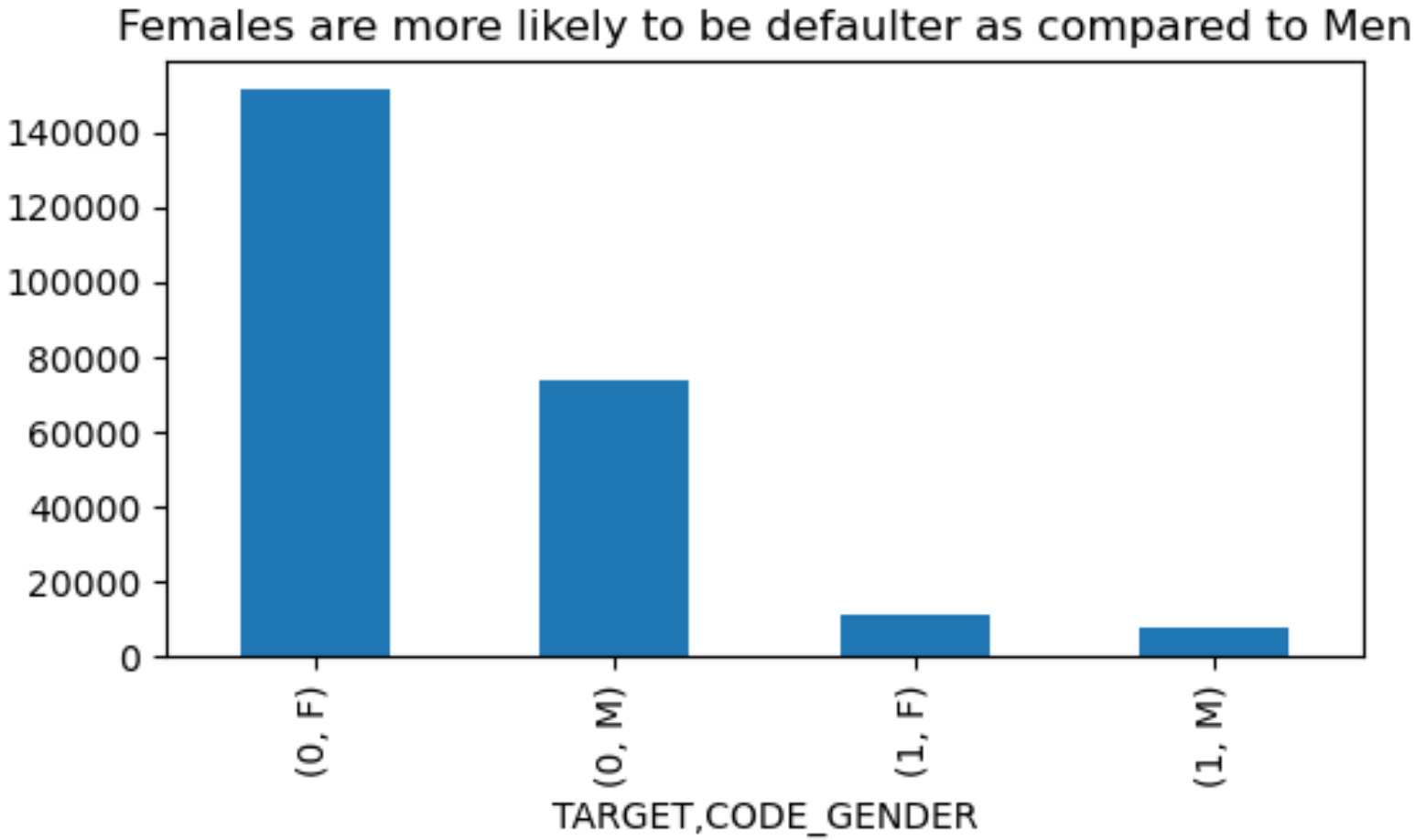
The loan-providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming defaulters. Suppose you work for a consumer finance company that specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. When the company receives a loan application, the company has to decide on 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. The data given below contains information about the loan application at the time of applying for the loan. It contains two types of scenarios: The client with payment difficulties: he/she had a late payment of more than X days on at least one of the first Y installments of the loan in our sample, All other cases: All other cases when the payment is paid on time.

Overall Approach

- Importing data
- Understanding dependent variables
- Checking the structure of data dealing with missing or null values.
- Removed columns with more than 30% null values in them.
- Checking for Outliers.
- Univariate analysis on Categorical and Numeric data. (Pie charts, bar charts, count plots, box plots, etc.)
- Bivariate or Multivariate analysis on Categorical and Numeric data (Scatter plot, pair plot, heatmap, etc.)

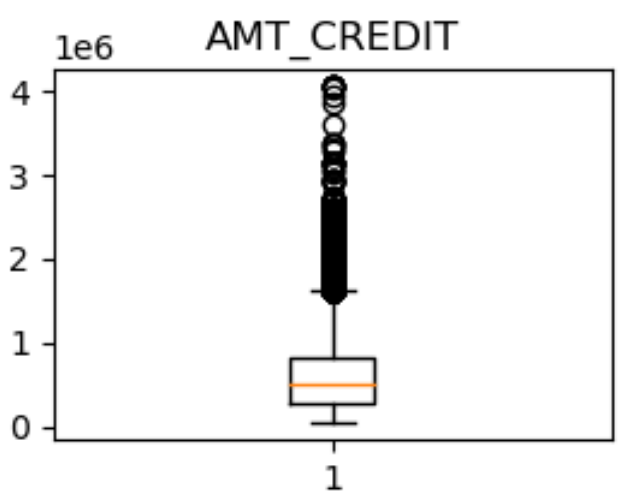
The distribution of the Target variable is quite imbalanced. As 92% are Genuine and 8% are defaulters.

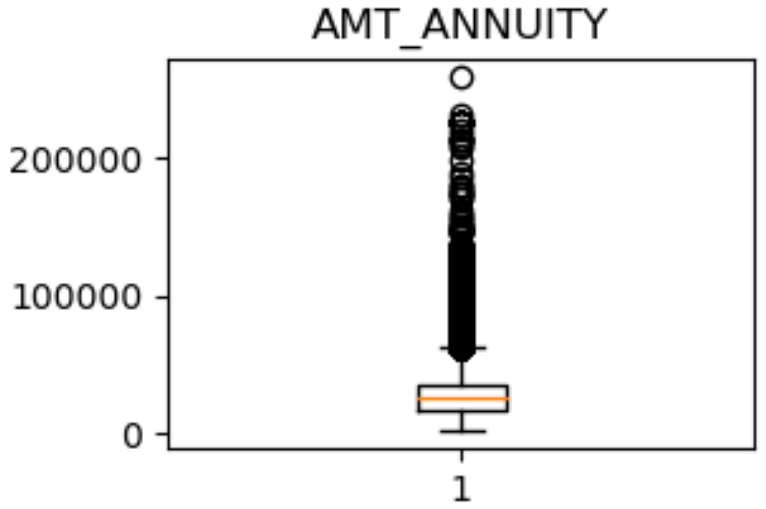


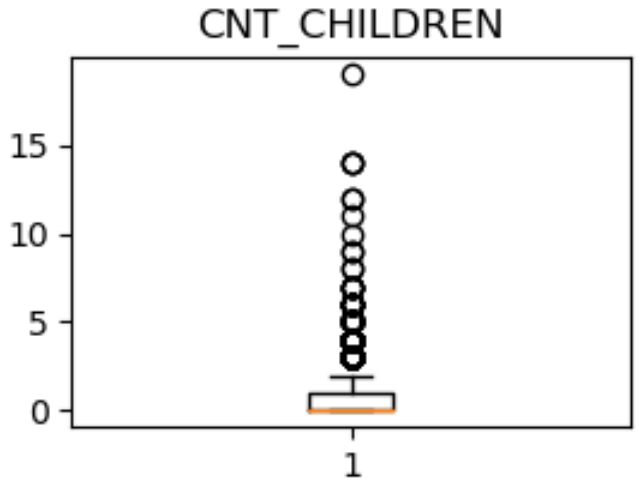


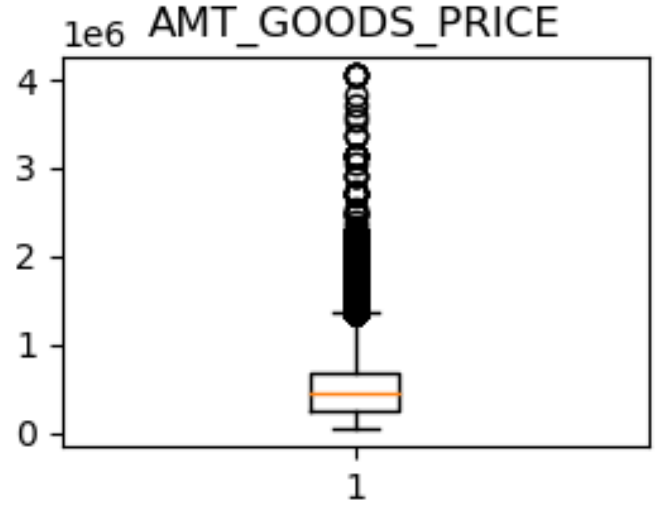
Univariate Analysis

- CNT_CHILDREN , AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE have outliers.



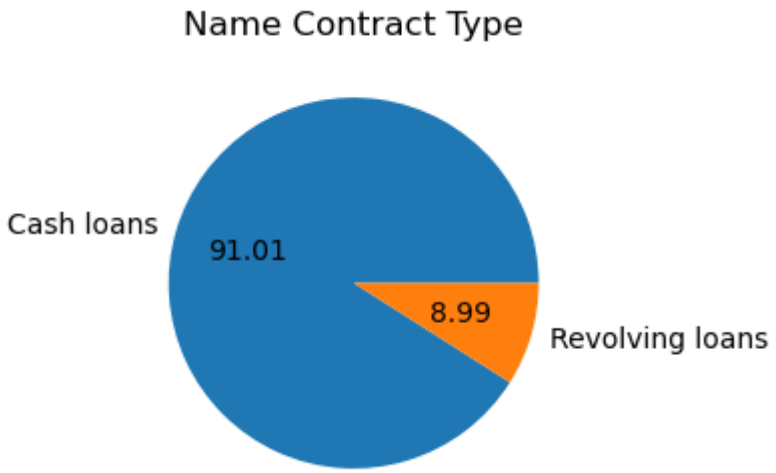


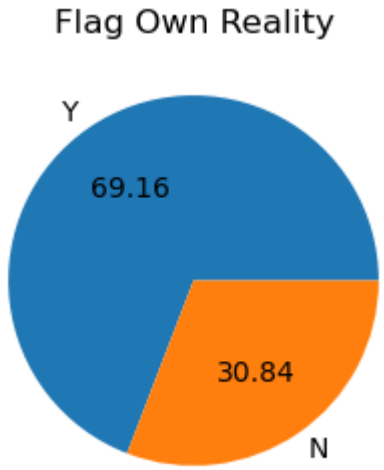


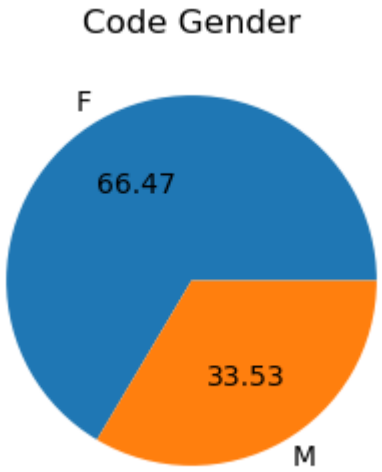


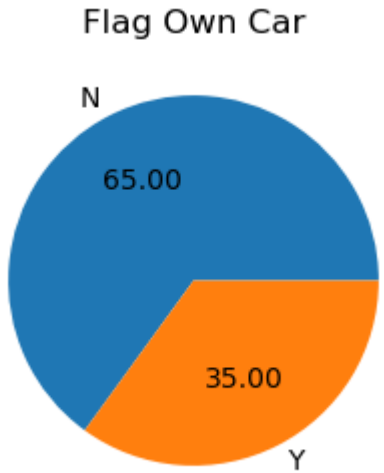
Univariate Analysis

- Name Contract Type
- Flag Own Reality
- Code Gender
- Flag Own Car



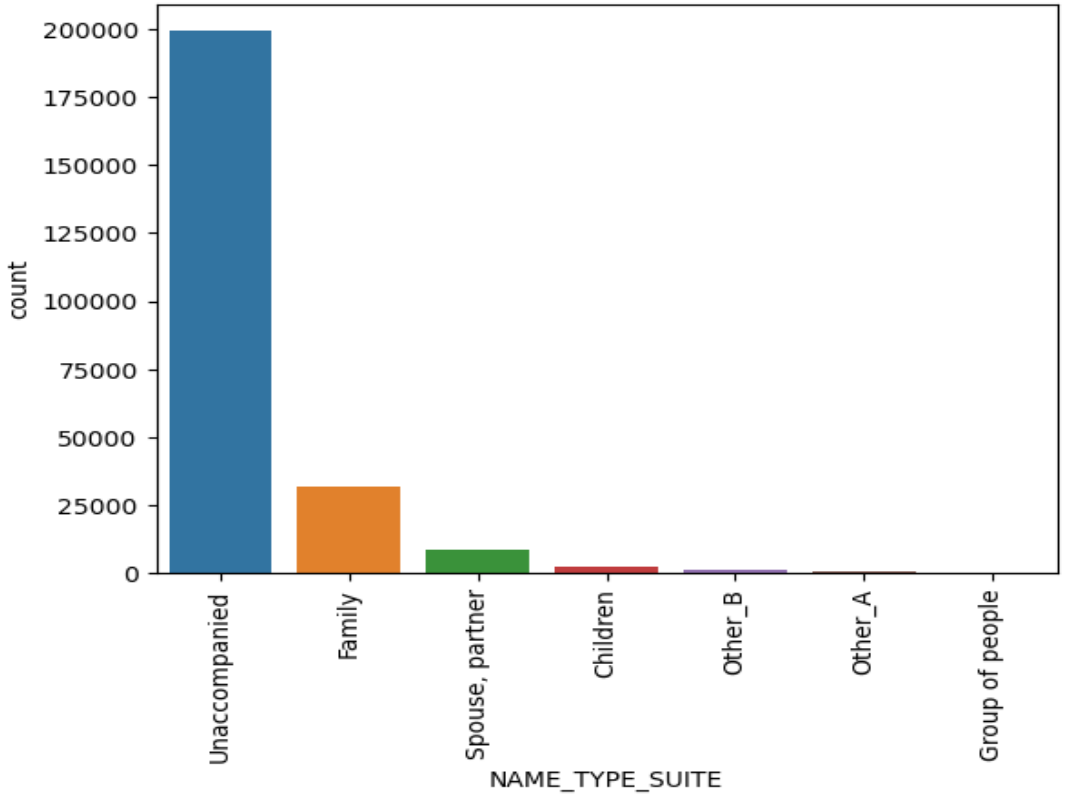


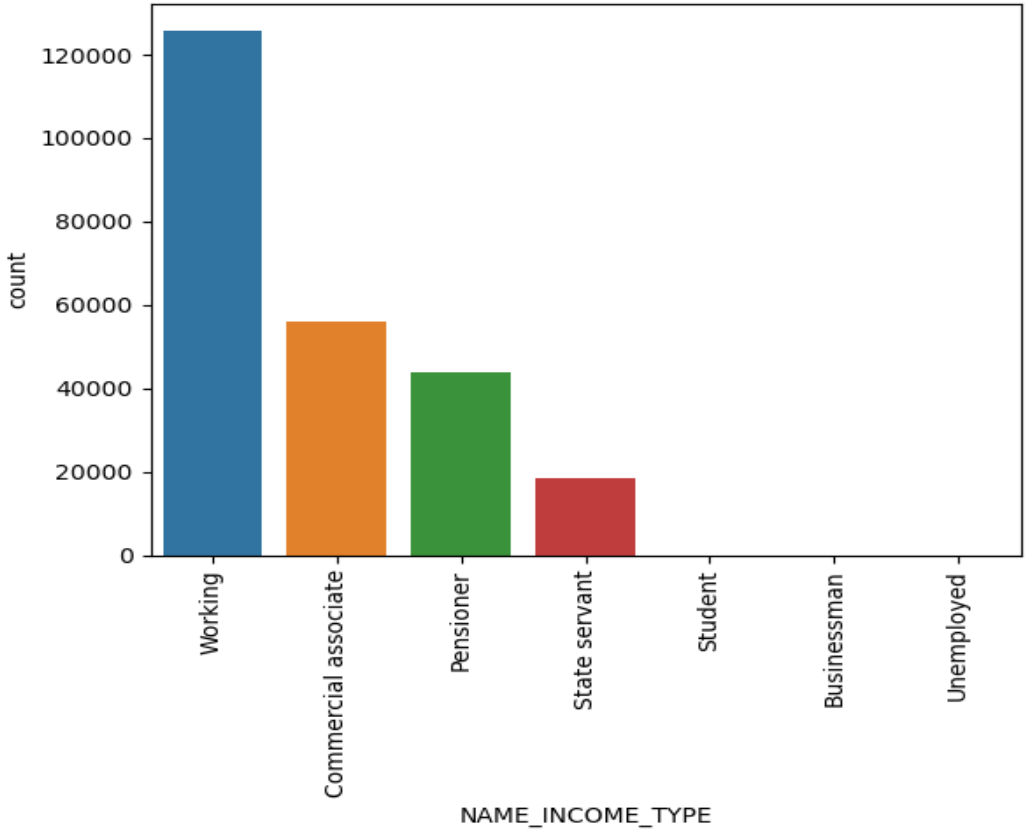




Univariate Analysis

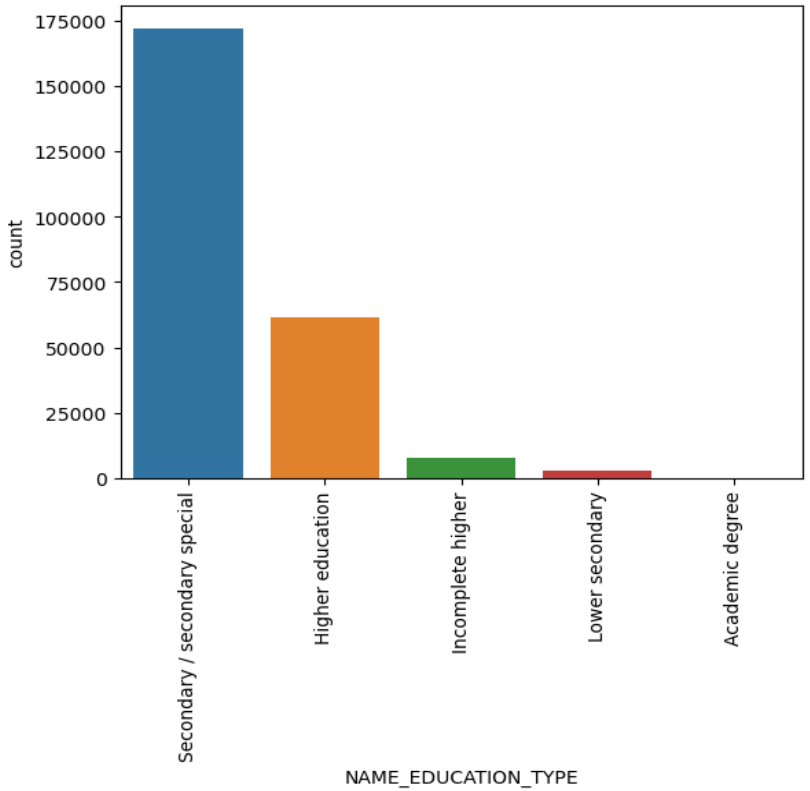
- Name Suite Type
- Name Income Type

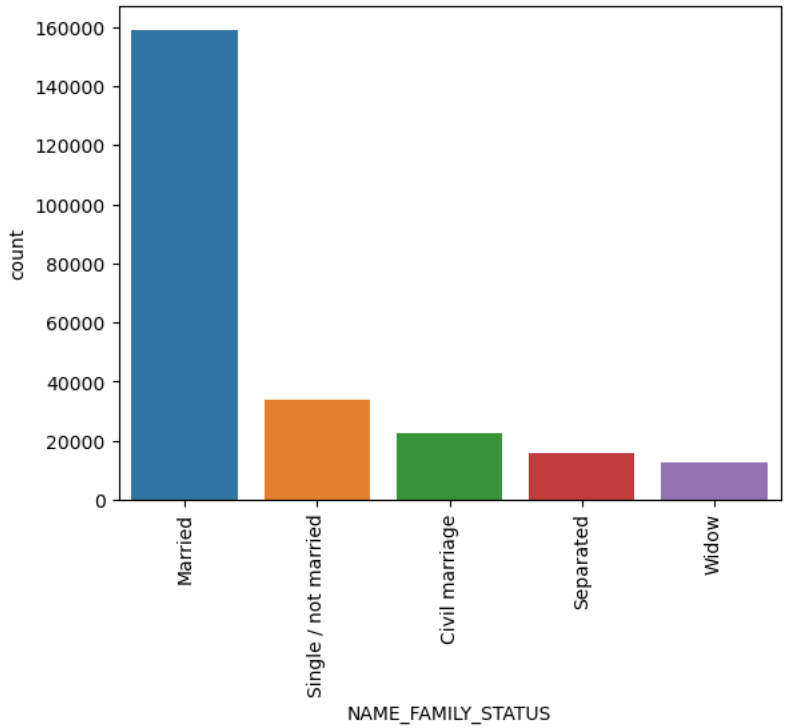


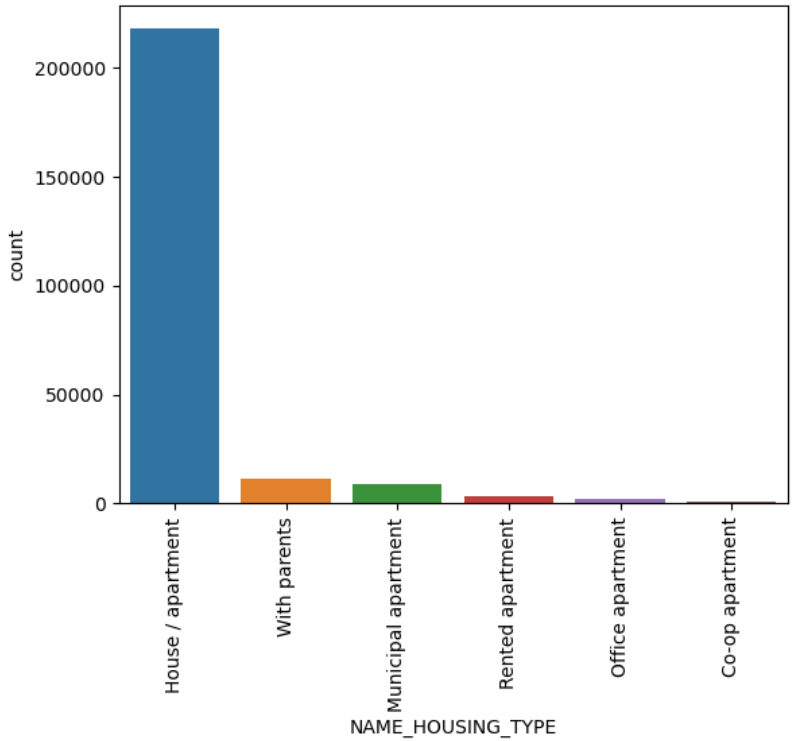


Univariate Analysis

- Name Education Type
- Name Family Status
- Name Housing Type

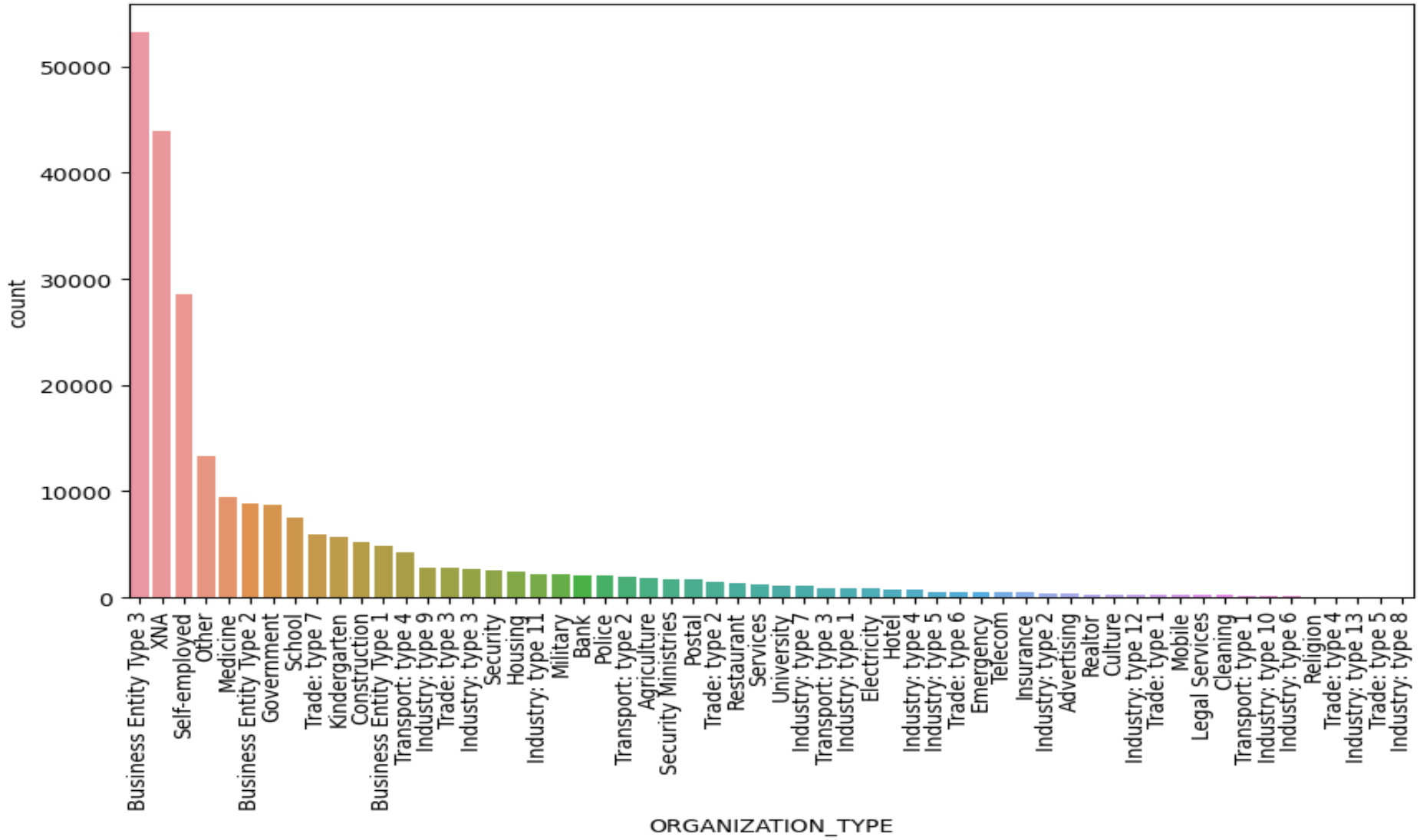






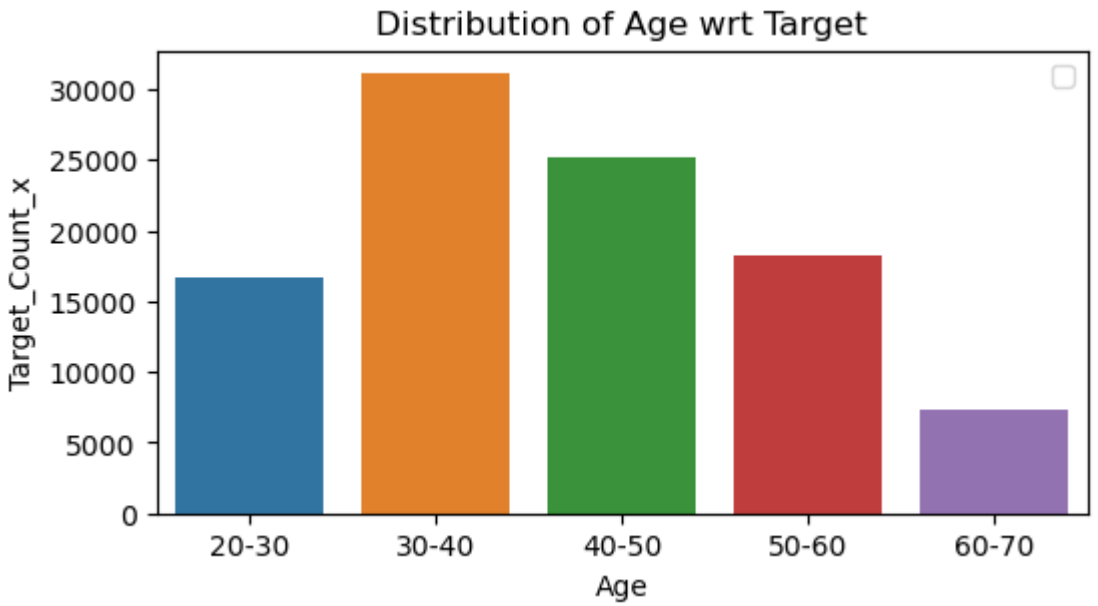
Univariate Analysis

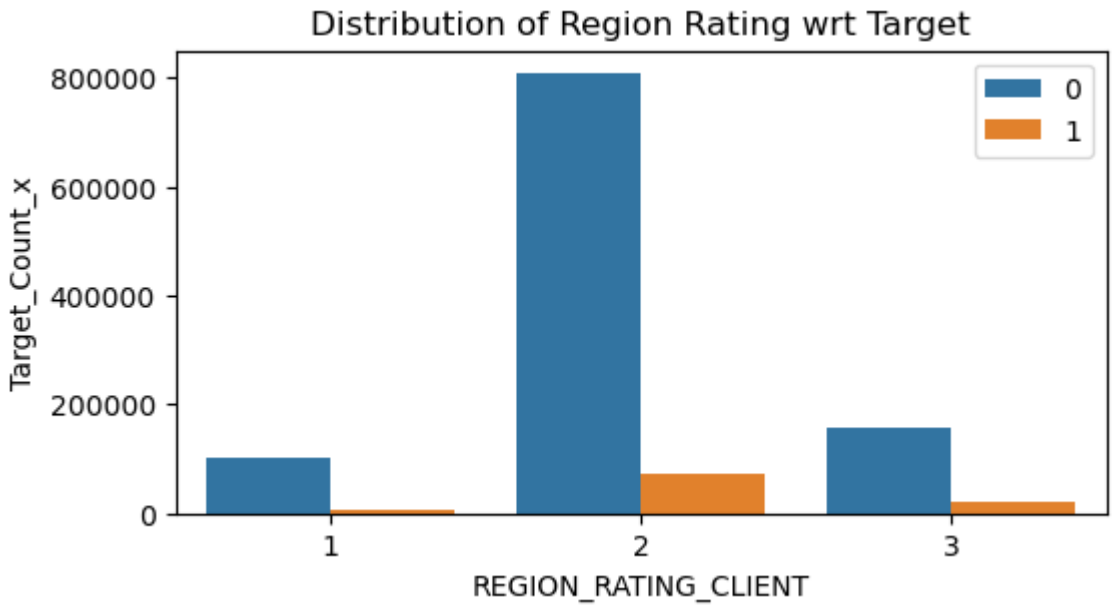
- Organization Type



Bivariate Analysis

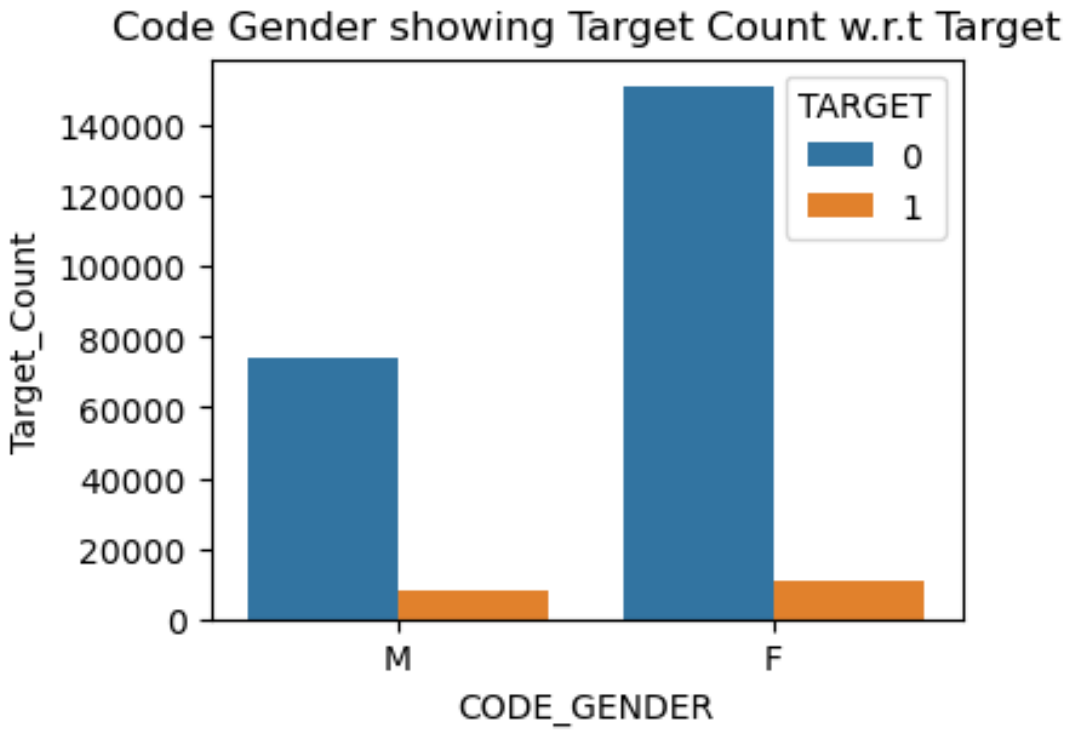
- Distribution of Age w.r.t Target
- Distribution of Region Rating w.r.t Target.

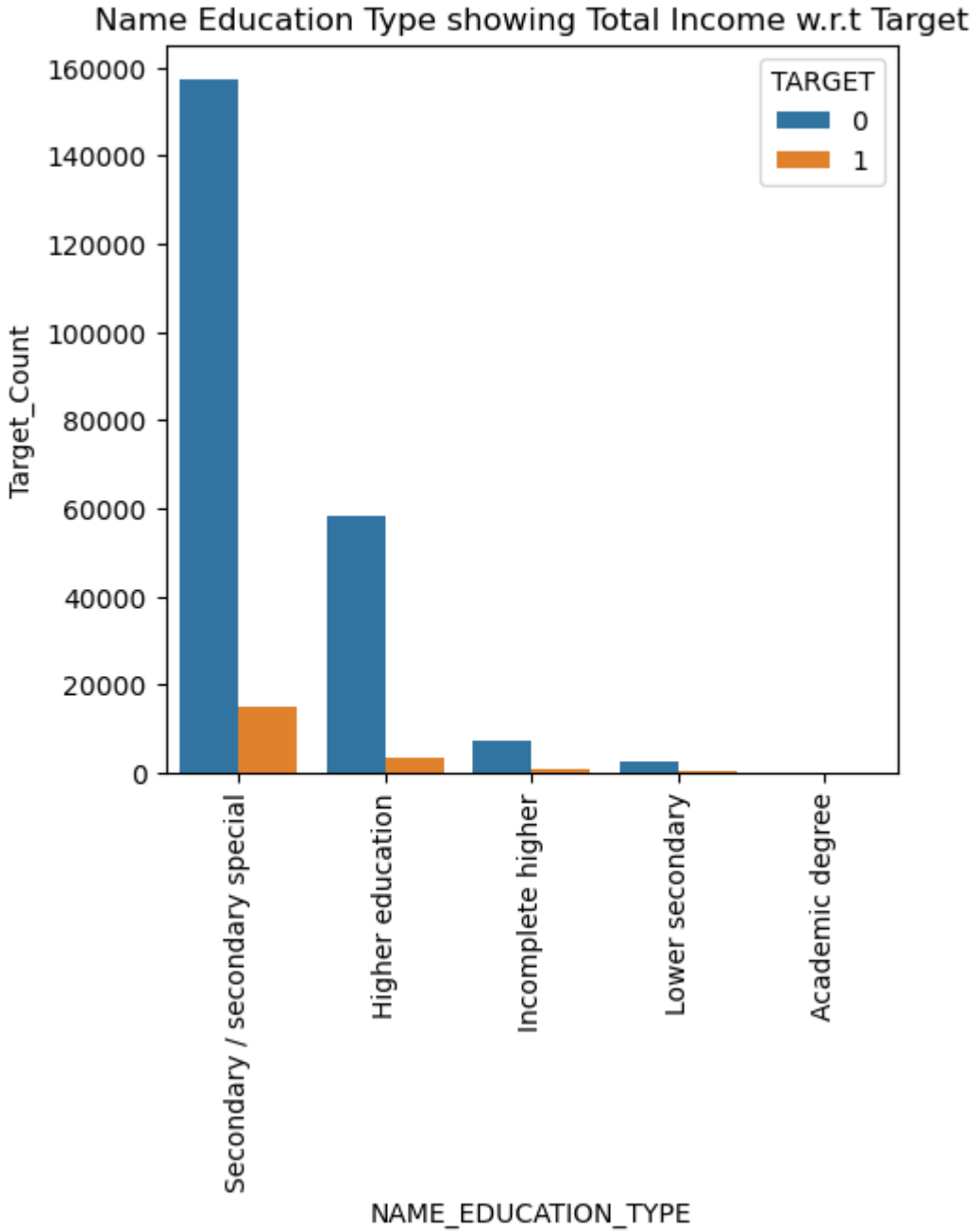




Bivariate Analysis

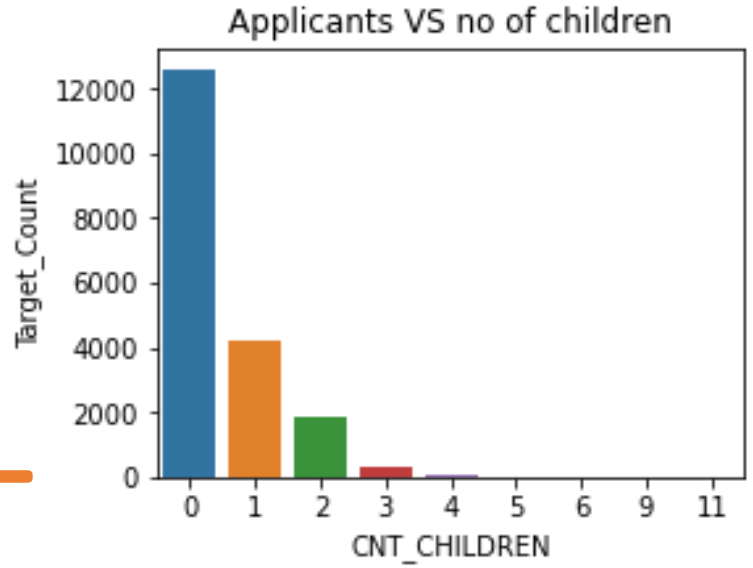
- Code Gender showing Target Count w.r.t Target.
- Name education type showing total Income w.r.t Target.

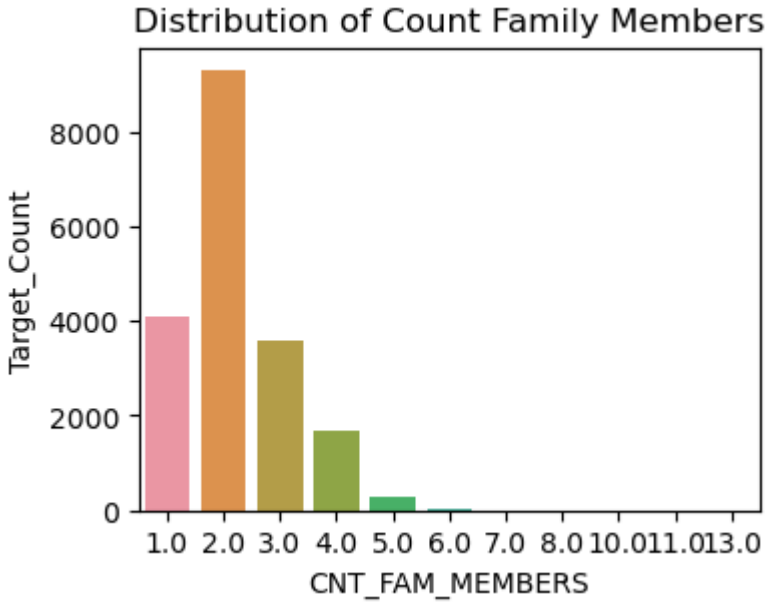


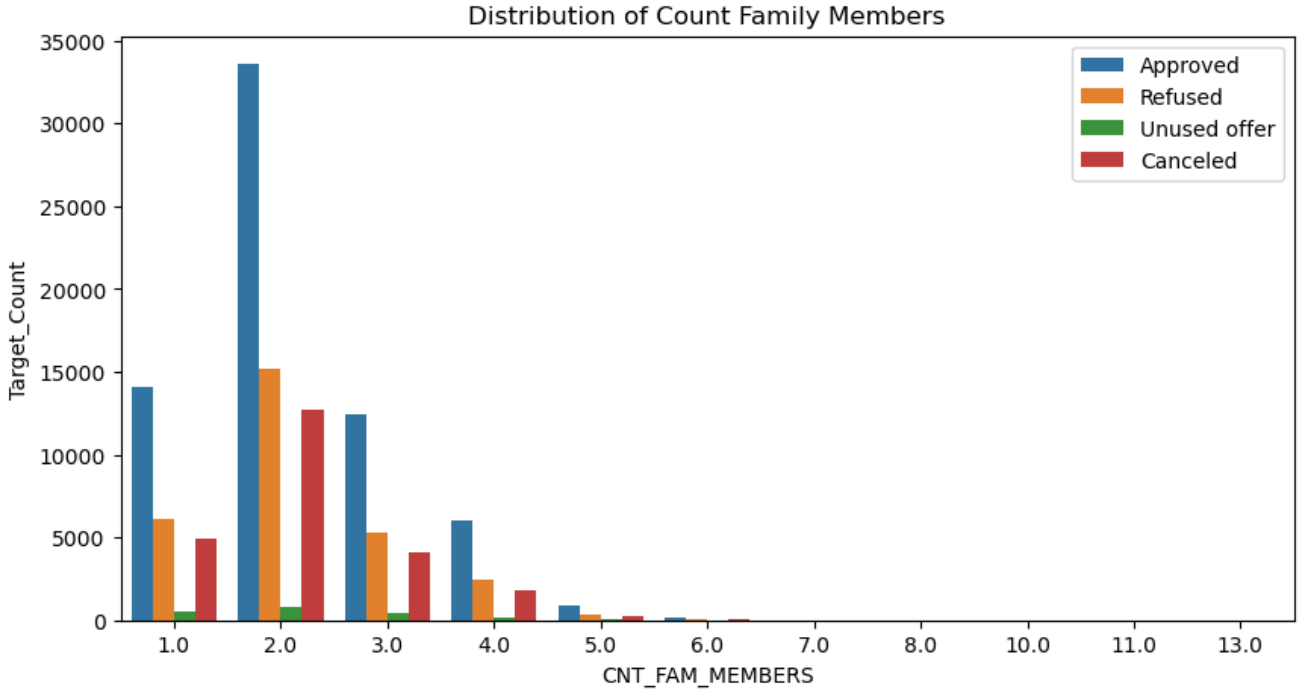


Bivariate Analysis

- Applicants vs CNT_CHILDREN
- Distribution of count of family members.
- Distribution of count family members w.r.t name contract status.

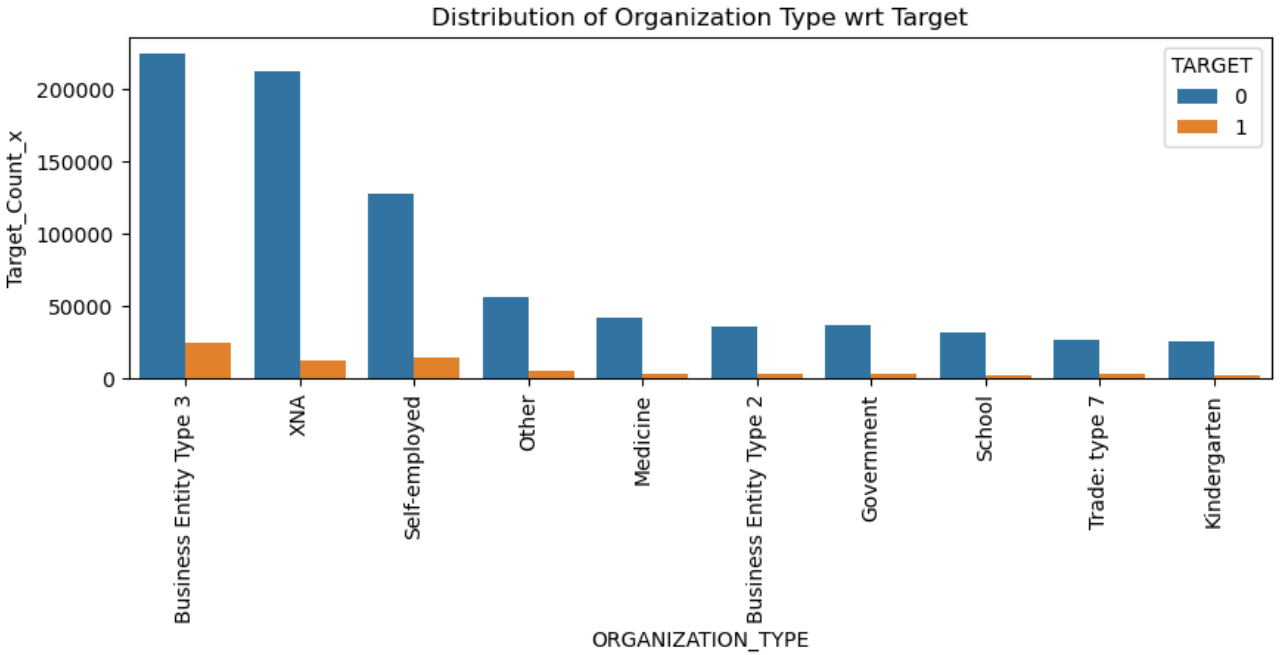


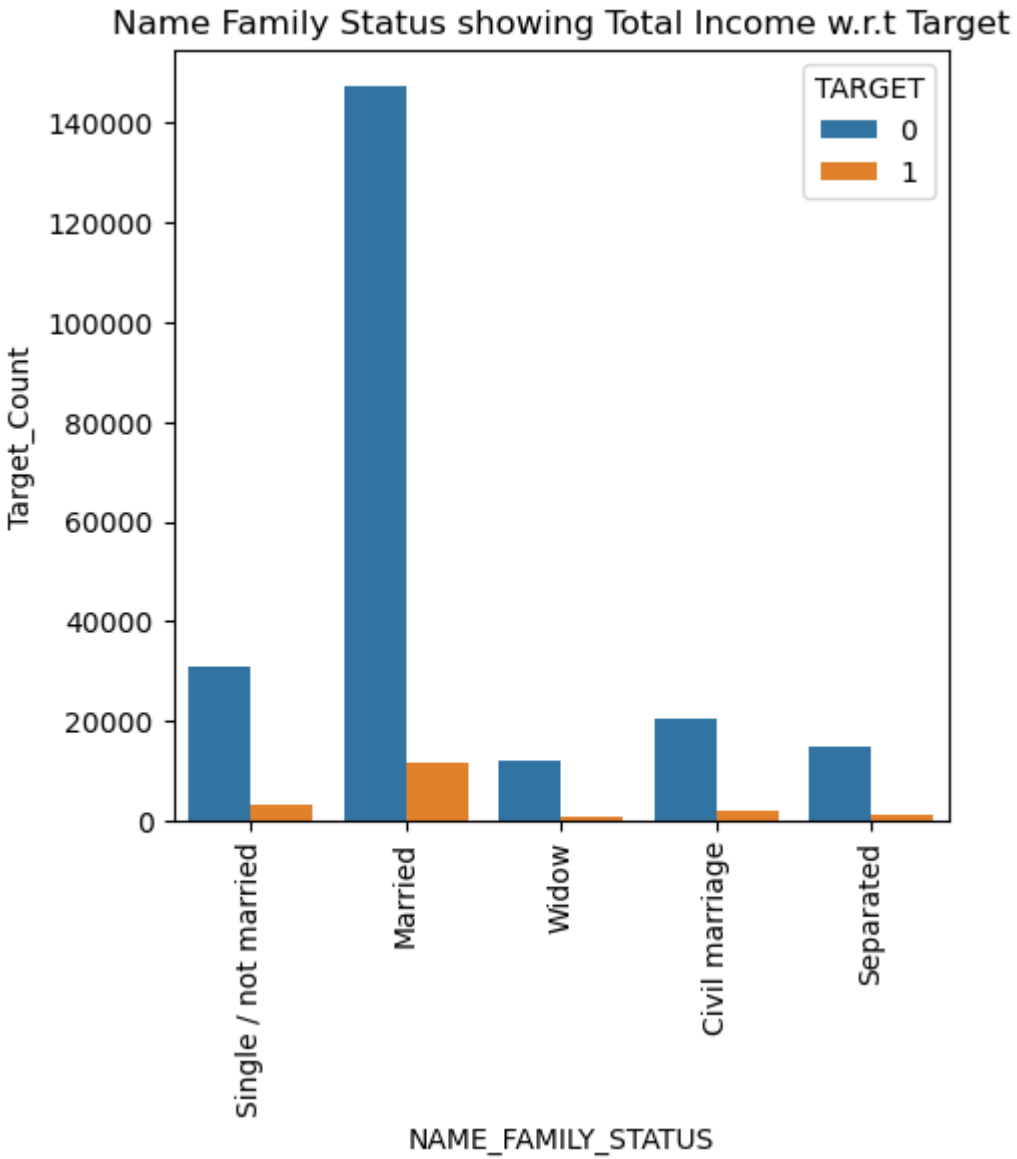




Bivariate Analysis

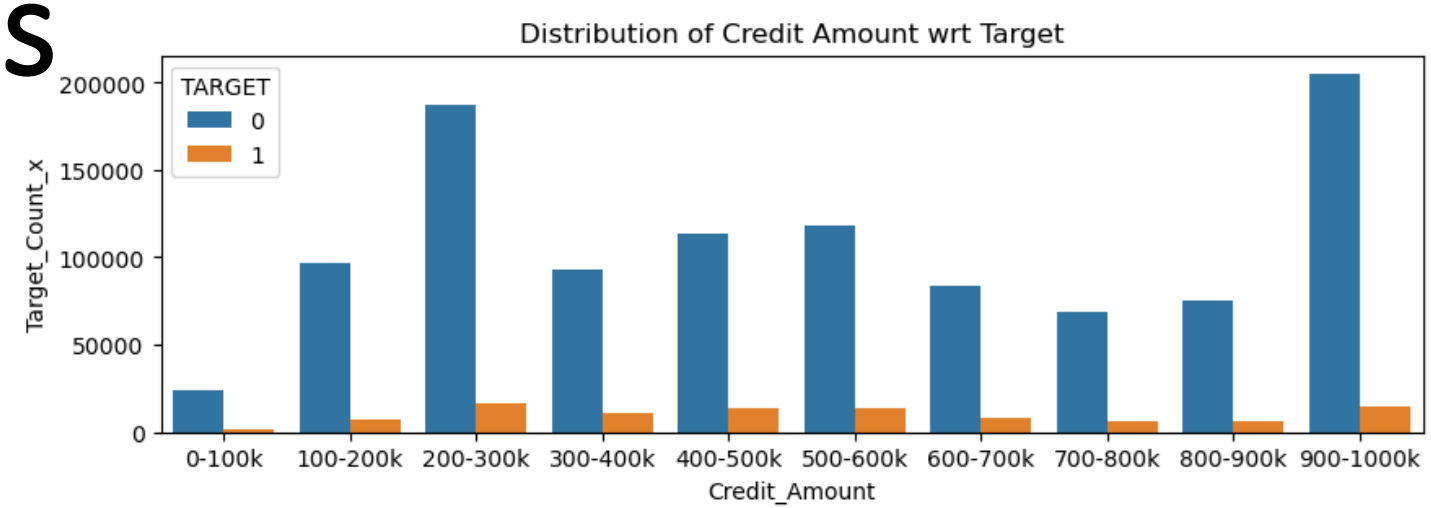
- Distribution of Organization Type w.r.t Target.
- Name family status showing total Income w.r.t Target.

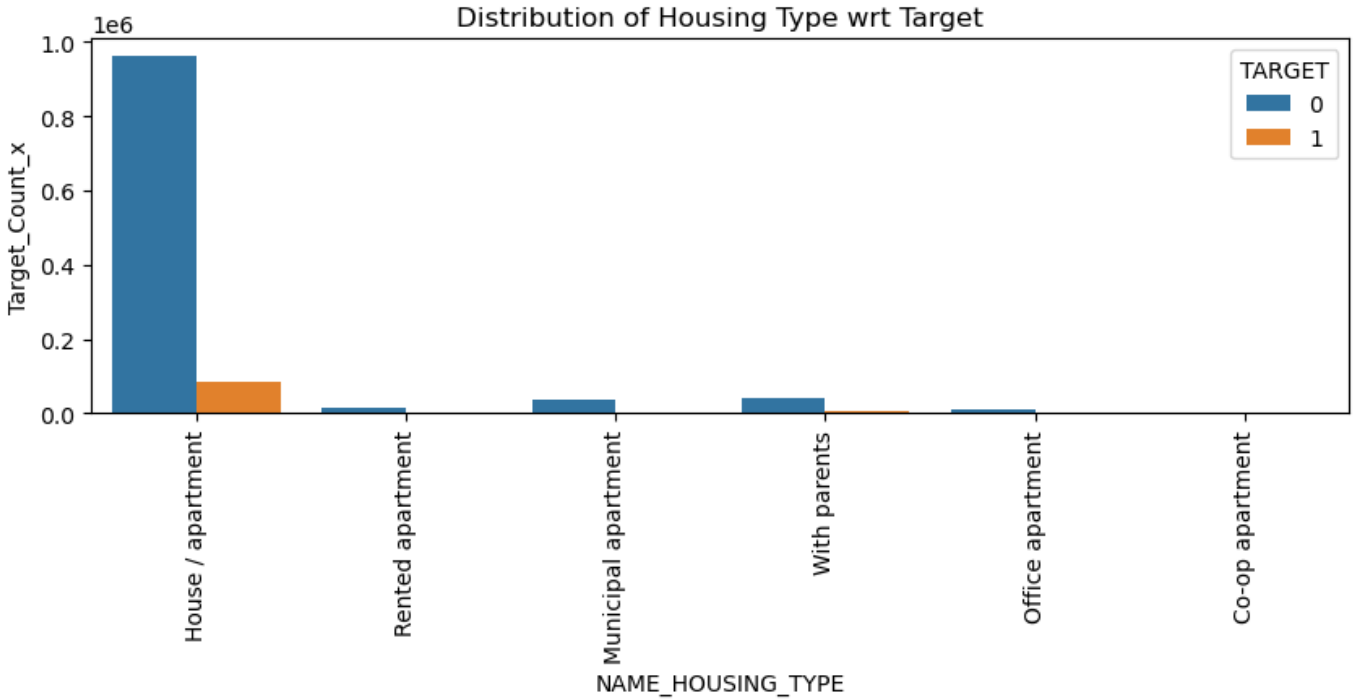




Bivariate Analysis

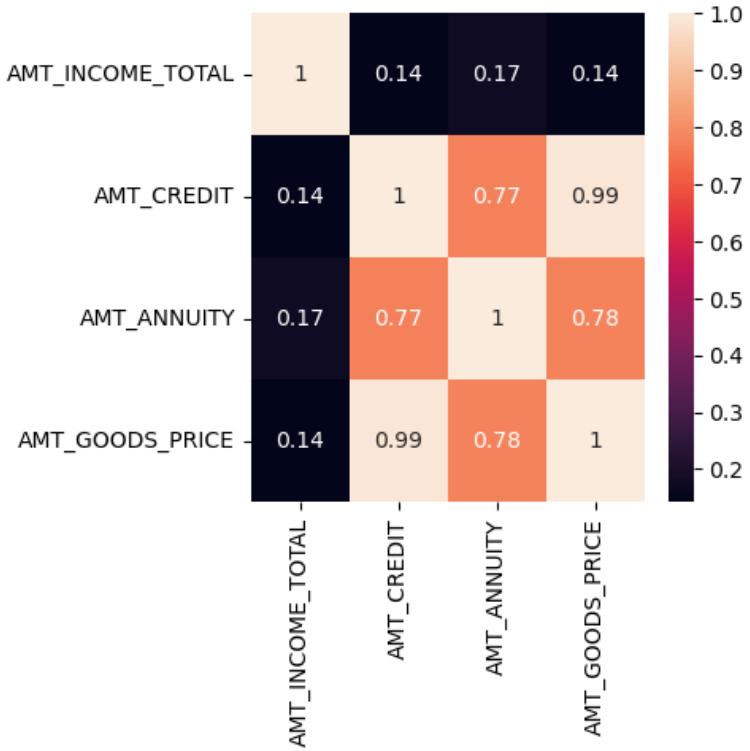
- Distribution of Credit Amount w.r.t Target.
- Distribution of Housing Type w.r.t Target.

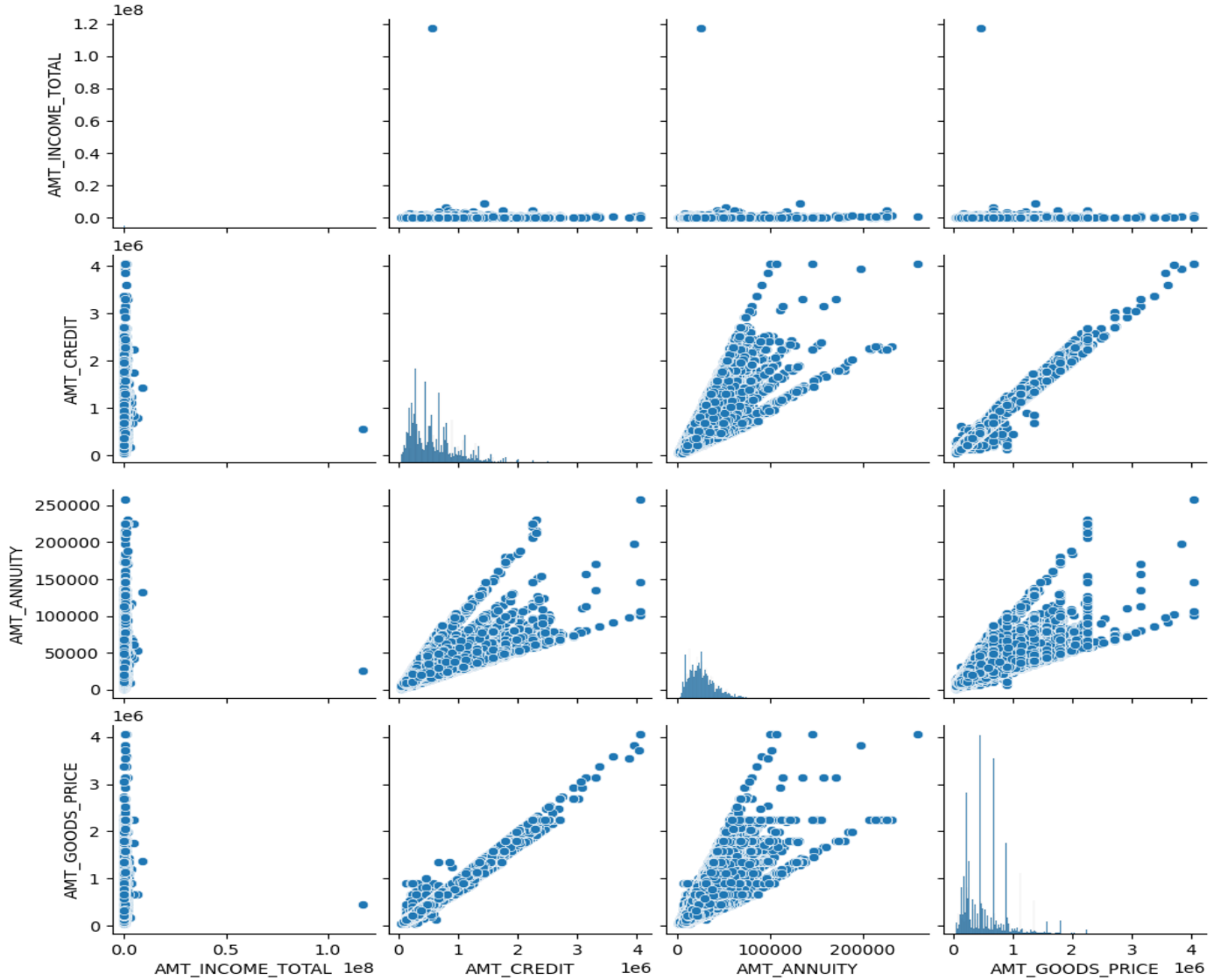




Multivariate Analysis

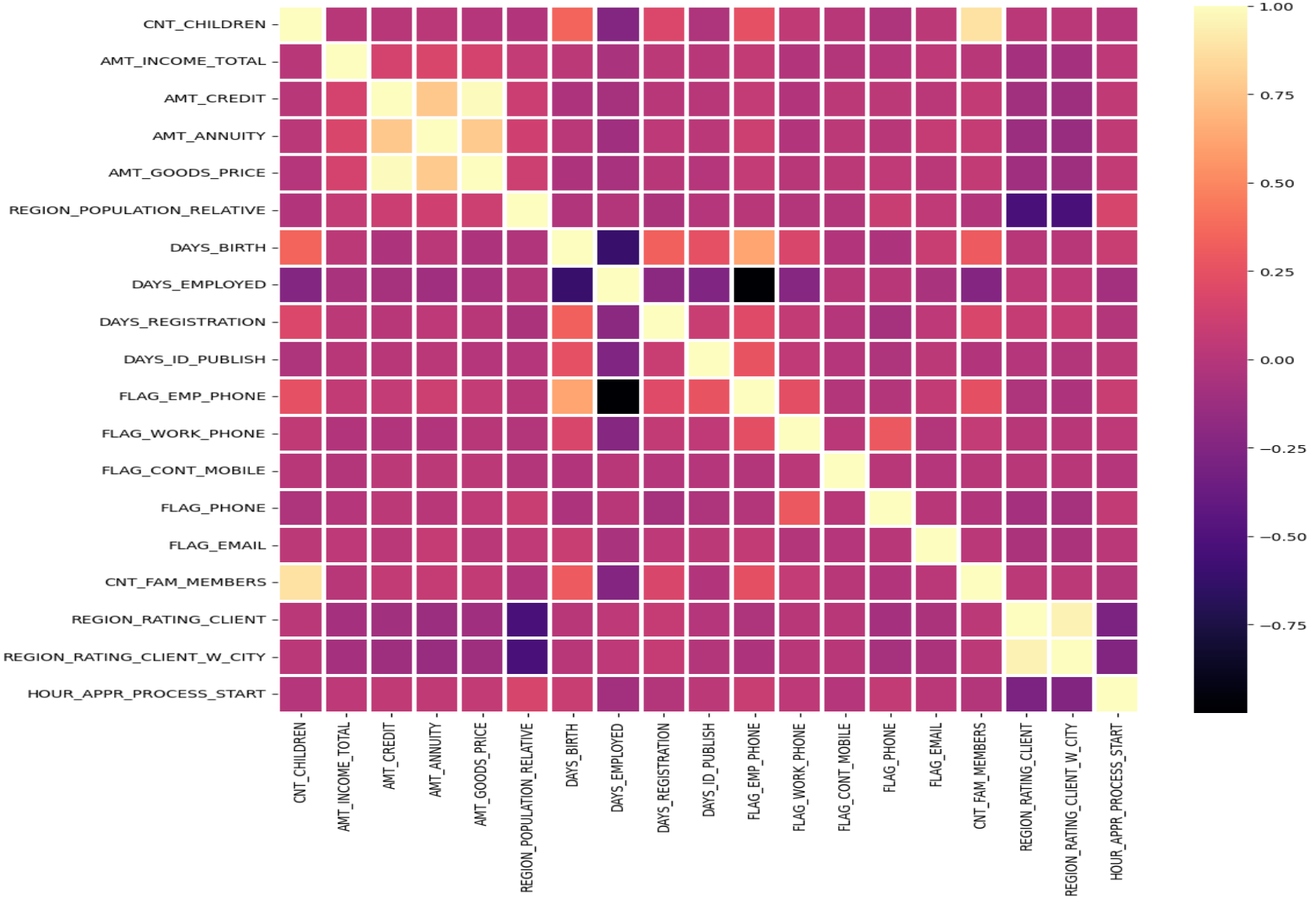
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE





Multivariate Analysis

- CNT_CHILDREN
- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_ANNUITY
- AMT_GOODS_PRICE
- REGION_POPULATION
- DAYS_BIRTH
- DAYS_EMPLOYED
- DAYS_REGISTRATION
- DAYS_ID_PUBLISH
- FLAG_EMP_PHONE
- FLAG_WORK_PHONE
- FLAG_CONT_MOBILE
- FLAG_PHONE
- FLAG_EMAIL
- CNT_FAM_MEMBERS
- REGION_RATING_CLIENT
- REGION_RATING_CLIENT_W_C ITY
- HOUR_APPR_PROCESS_START



Relevant Results

- CNT_CHILDREN , AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE have outliers.
- Removed columns with more than 30% null values and also removed all the null values after removing columns.
- Distribution of the Target variable is quite imbalanced. As 92% are Genuine and 8% are defaulters.

Relevant Results

- Applicants between the age group **30-40** are more likely to be defaulters.
- Applicants living in Region 3 are more likely to be defaulters.
- Females are more defaulters than Males analysed by the given data. As the population of females is more than men.
- Applicants having Academic degree are less likely to be defaulters.
- Applicants with 0 children have a high rate of being defaulters.
- Applicants having 2 family members are more likely to be defaulters.
- Defaulter Applicants having 2 children have the highest no. of approved and refused loans.

Relevant Results

- Maximum loans are taken from Business Entity 3 and Self employed are more likely to be defaulters as compared to others.
- Married people have taken maximum loans but singles and civil marriages are more likely to be defaulters.
- Most of the applicants lives in their own houses and very less amount of them are defaulters.
- AMT_CREDIT and AMT_GOODS_PRICE are highly and positively correlated.