EDA CASE STUDY

By,
Mrs. Mamatha E
Mr. Sayantan Dutta

Problem Statement

- Types of risks associated with the bank’s decision:
  1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company, i.e., Interest Loss.
  2. 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, i.e., Credit Loss.

Missing Data and its Treatment

- For higher percentage of missing values i.e., > 50%, dropping the columns was the approach used as they effect the analysis.
- Columns having lower percentage of null values (i.e., < 13%), impute missing data either with mean, median and mode values.
- Before imputing we need to categorize the column as continuous and categorical variables.

- Measures to take for handling missing values < 13%
  - **Continuous Variables:**
    1. AMT_ANNUITY: Impute with median value, i.e., 10791.0.
    2. AMT_GOODS_PRICE: Impute with median value, i.e., 270000.0.
    3. EXT_SOURCE_2 : Impute with mean value, i.e., 0.51.
    4. DAYS_LAST_PHONE_CHANGE: Impute with median value, i.e., -757.0
  - **Categorical Variables:** For the categorical variables listed below we can impute the missing values with most frequently occurring value.
    1. NAME_TYPE_SUITE
    2. CNT_FAM_MEMBERS
    3. OBS_30_CNT_SOCIAL_CIRCLE
    4. DEF_30_CNT_SOCIAL_CIRCLE
    5. OBS_30_CNT_SOCIAL_CIRCLE
    6. DEF_30_CNT_SOCIAL_CIRCLE

Identify Outliers in the Data Set

- To identify outliers we have used box plots and found that the following columns had outliers:
  - **AMT_ANNUITY:**
    - There were many data points at the upper fence of the boxplot. The chunk of outliers can be handled with the binning method and one data point which is after 250000 can be deleted as it is far from the actual range.
  - **AMT_CREDIT:**
    - Here also you can find that there are many data points at the upper fence of the boxplot which can be handled with the binning method.
  - **AMT_INCOME:**
    - One exceptional outlier was present which was far from the actual range and hence it has to be deleted as it might effect the analysis.

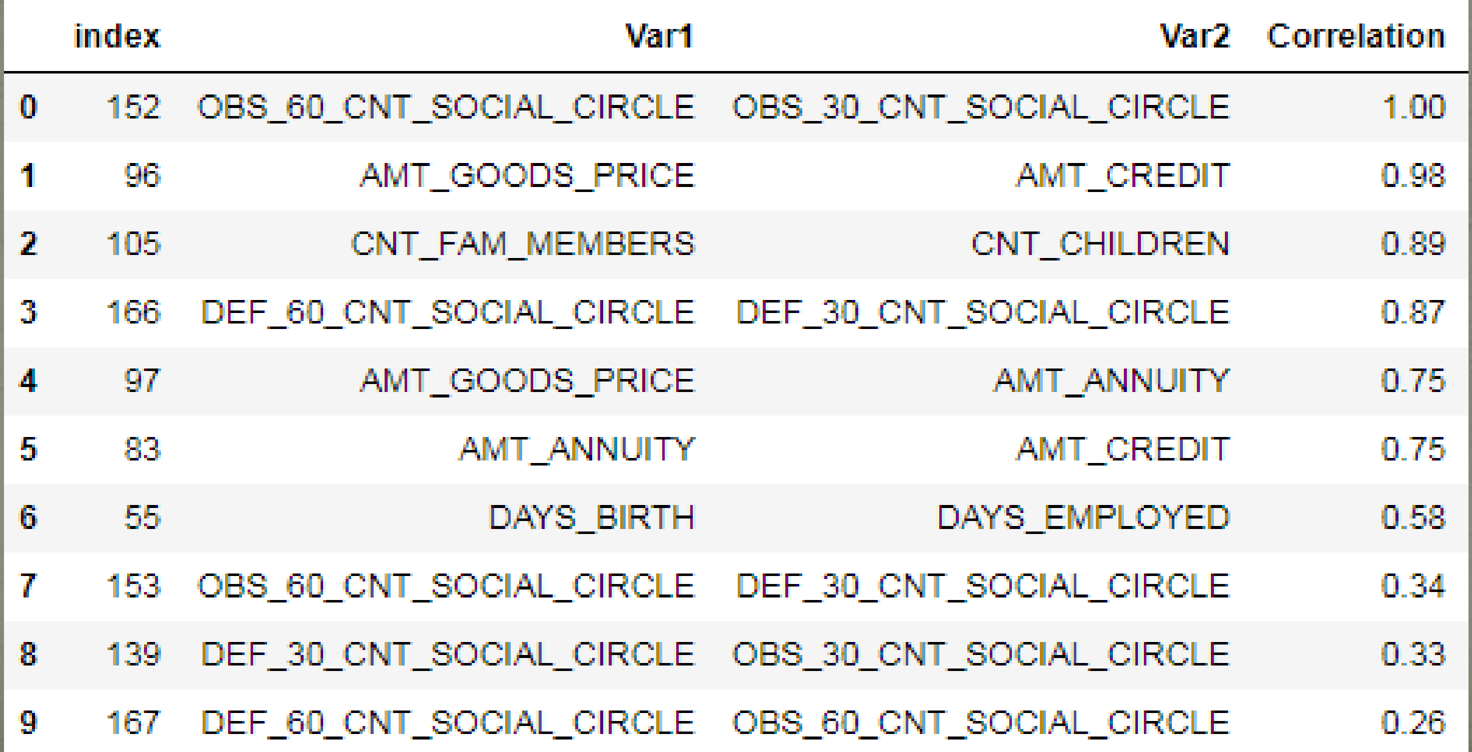
- **CNT_CHILDREN:**
  - Though there are some data points above the upper fence those may be the genuine values and cannot be treated as an outlier so we handle those values using capping method.
- **DAYS_EMPLOYED:**
  - The outlier was present after 350000. This has to be deleted as the days before the application that the person started his current employment cannot be so many days.

Data Imbalance and its Ratio

- The target variable in our dataset is TARGET column which tells whether the customer has defaulted(1) or not (0).
- There was imbalance ratio with respect to this column which is as follows:
  - TARGET = 0 → 91.93%
  - TARGET = 1 → 8.07%
- Which tells that there were 8.07% customers who were defaulted customers.

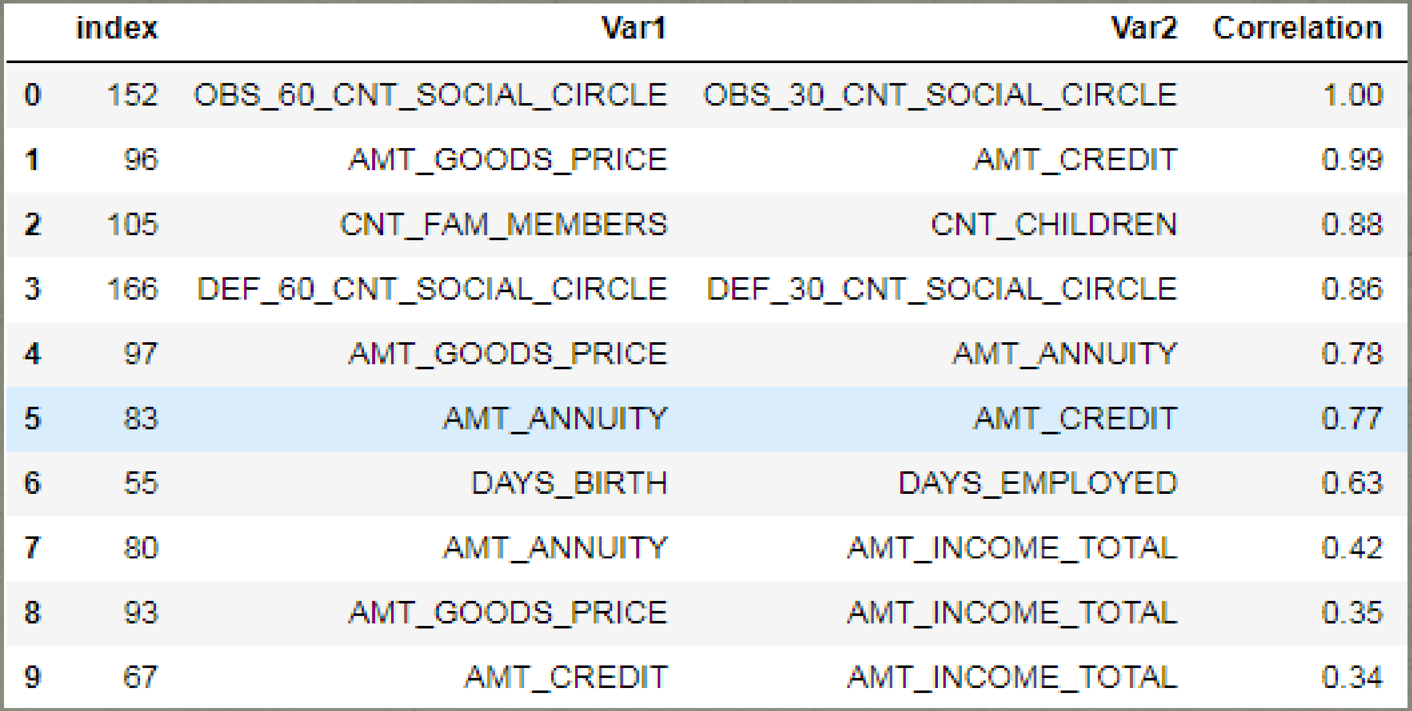
Analysis

- Correlation between numerical variables to find better relationship between them with respect to TARGET=1 & TARGET=0.
    - **Top 10 Correlation for TARGET=1**



	index	Var1	Var2	Correlation
0	152	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
1	96	AMT_GOODS_PRICE	AMT_CREDIT	0.98
2	105	CNT_FAM_MEMBERS	CNT_CHILDREN	0.89
3	166	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87
4	97	AMT_GOODS_PRICE	AMT_ANNUITY	0.75
5	83	AMT_ANNUITY	AMT_CREDIT	0.75
6	55	DAYS_BIRTH	DAYS_EMPLOYED	0.58
7	153	OBS_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.34
8	139	DEF_30_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.33
9	167	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.26

- **Top 10 Correlation for TARGET=0**



	index	Var1	Var2	Correlation
0	152	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00
1	96	AMT_GOODS_PRICE	AMT_CREDIT	0.99
2	105	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
3	166	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86
4	97	AMT_GOODS_PRICE	AMT_ANNUITY	0.78
5	83	AMT_ANNUITY	AMT_CREDIT	0.77
6	55	DAYS_BIRTH	DAYS_EMPLOYED	0.63
7	80	AMT_ANNUITY	AMT_INCOME_TOTAL	0.42
8	93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.35
9	67	AMT_CREDIT	AMT_INCOME_TOTAL	0.34

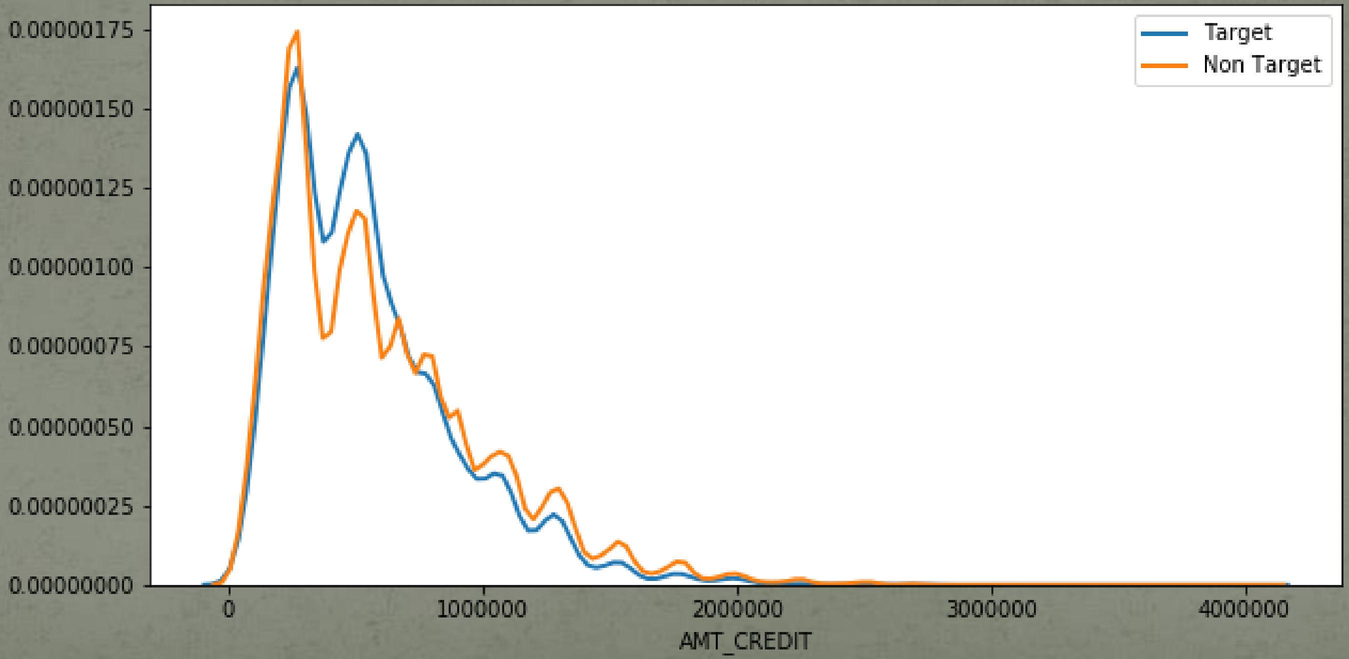
- The correlation is more or less same for the variables upto the 7th row in both dataframes, but the remaining rows has different parameters in both data frames.

- To carry out the further analysis with respect to data imbalance various types of analysis where carried out:

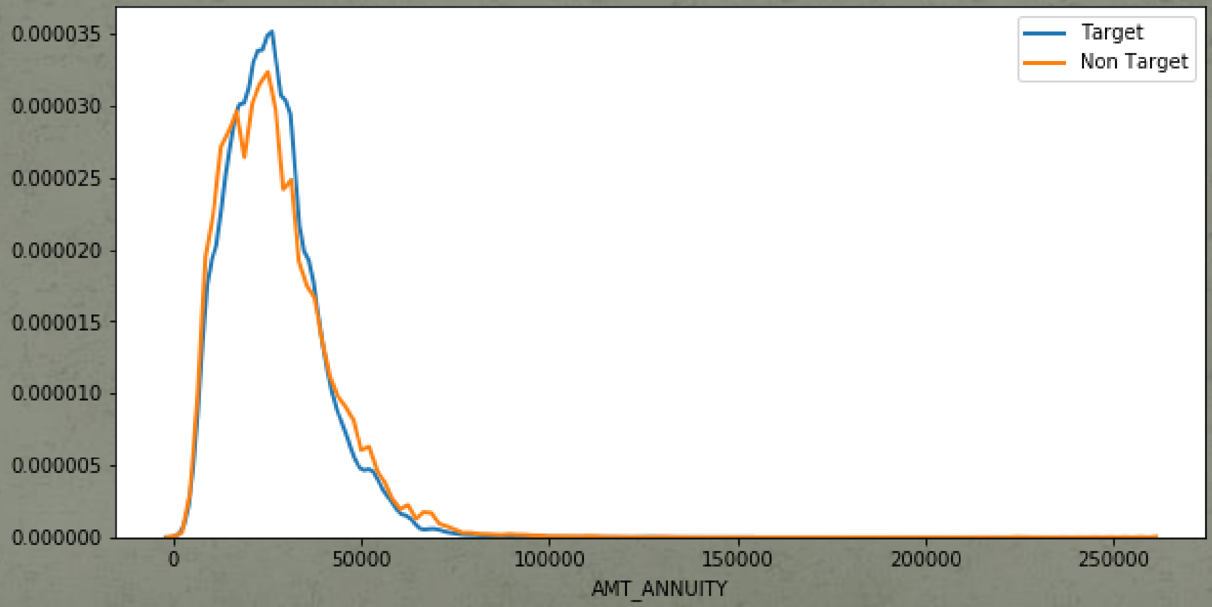
  1. **Univariate Analysis:**
     a. Continuous
     b. Categorical
  2. **Bivariate Analysis:**
     a. Continuous-Continuous
     b. Continuous-Categorical
     c. Categorical-Categorical

Univariate Analysis (Continuous)

- **AMT_CREDIT:**
  - The pattern is same but the 2nd spike is more for TARGET than NON TARGET indicating that the credit amount taken TARGET customers is more than NON TARGET.



- **AMT_ANNUITY:**
  - Though the pattern is same but the AMT_ANNUITY is slightly more for TARGET than NON TARGET customers indicated by spike.

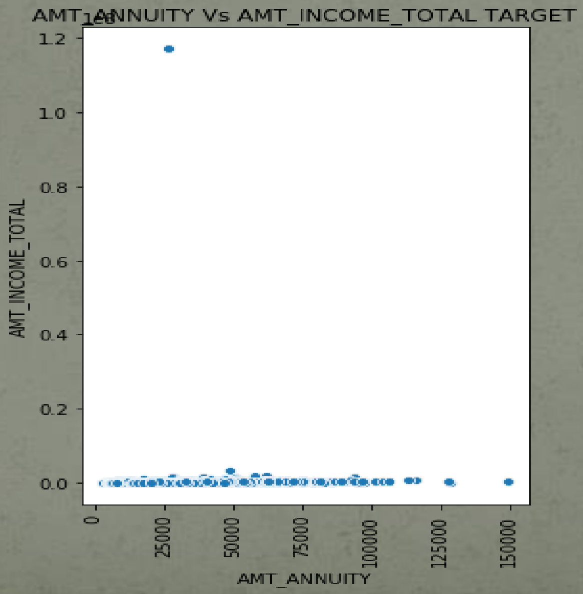


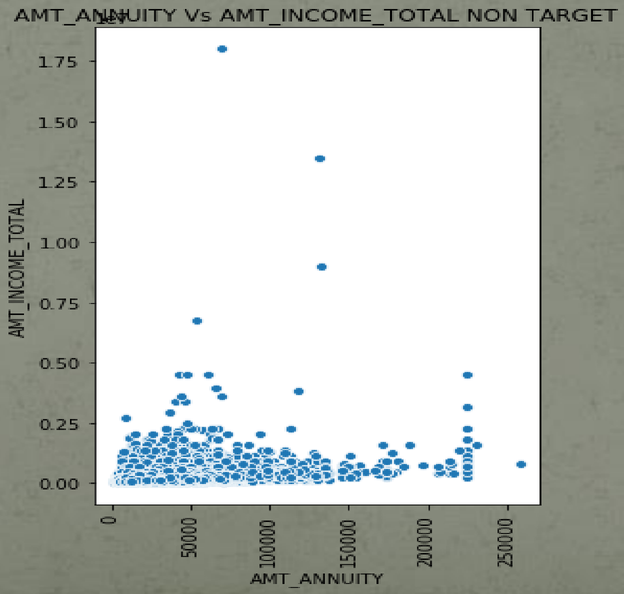
Univariate Analysis (Categorical)

- Following columns were considered for analysis:
  - NAME_CONTRACT_TYPE
  - NAME_INCOME_TYPE
  - NAME_EDUCATION_TYPE
  - NAME_FAMILY_STATUS
  - OCCUPATION_TYPE
- The count in each of the univariate analysis variables are more in TARGET=0 than TARGET=1 but, the pattern for all 5 variable analysis is more or less same.
- **Minor Observation:** % for laborers are more for TARGET=1 (i.e., 31%) when compared to TARGET=0 (i.e., 26%).

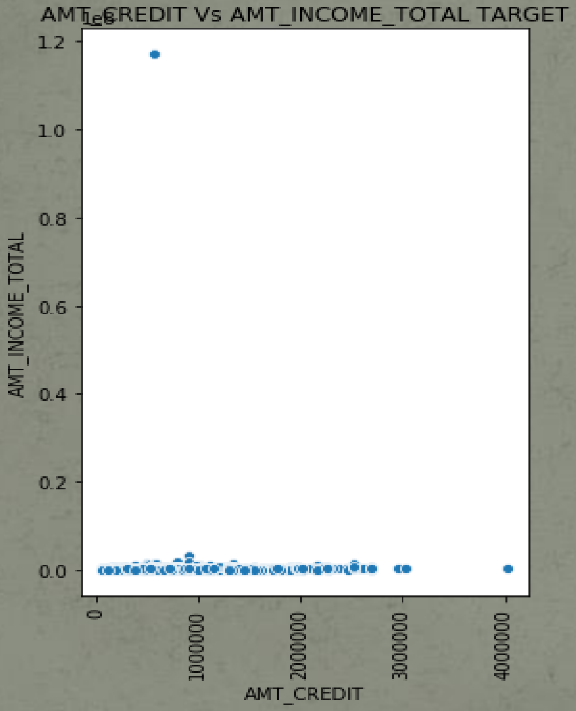
Bivariate Analysis (Continuous-Continuous)

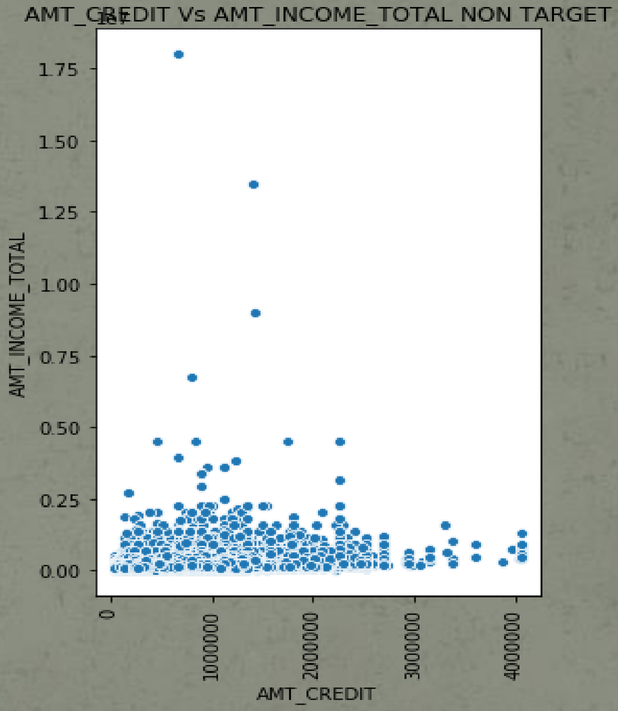
- **AMT_ANNUITY vs AMT_INCOME_TOTAL:**
  - For TARGET, as the customers income is less they are finding difficult in paying the loan amount on time.
  - Also, this might be the reason that this variable combination was not present in top ten of the correlation data-frame that we had found earlier.





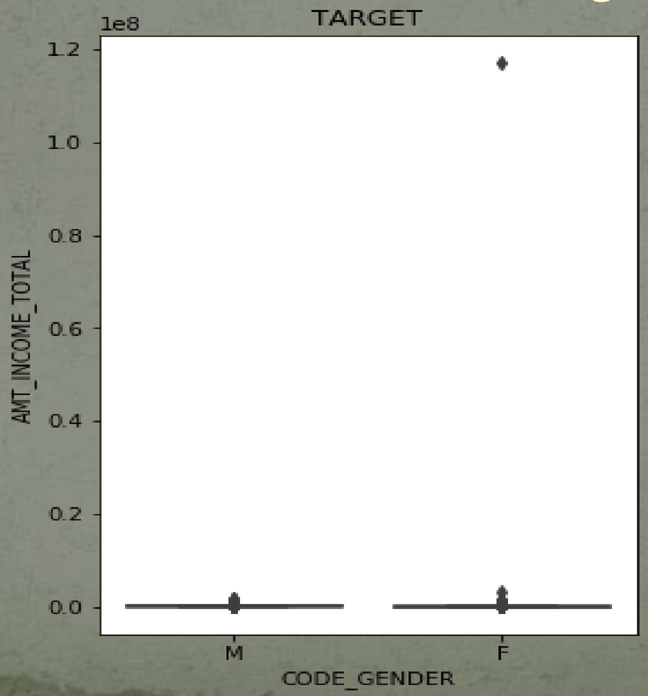
- **AMT_CREDIT vs AMT_INCOME_TOTAL:**
  - The plot looks similar to the plot of AMT_ANNUITY vs AMT_INCOME_TOTAL. Here also, the TARGET customers having low income have taken credit amount of upto 2 to 3 lakhs.

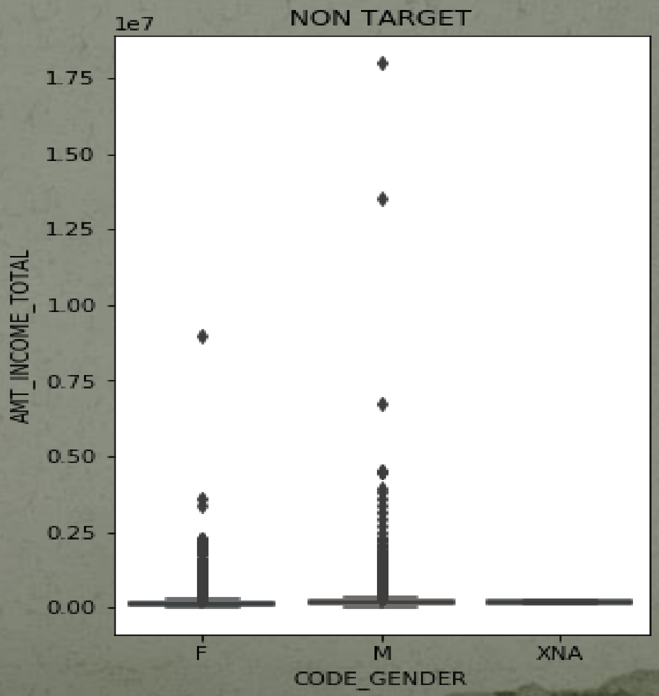




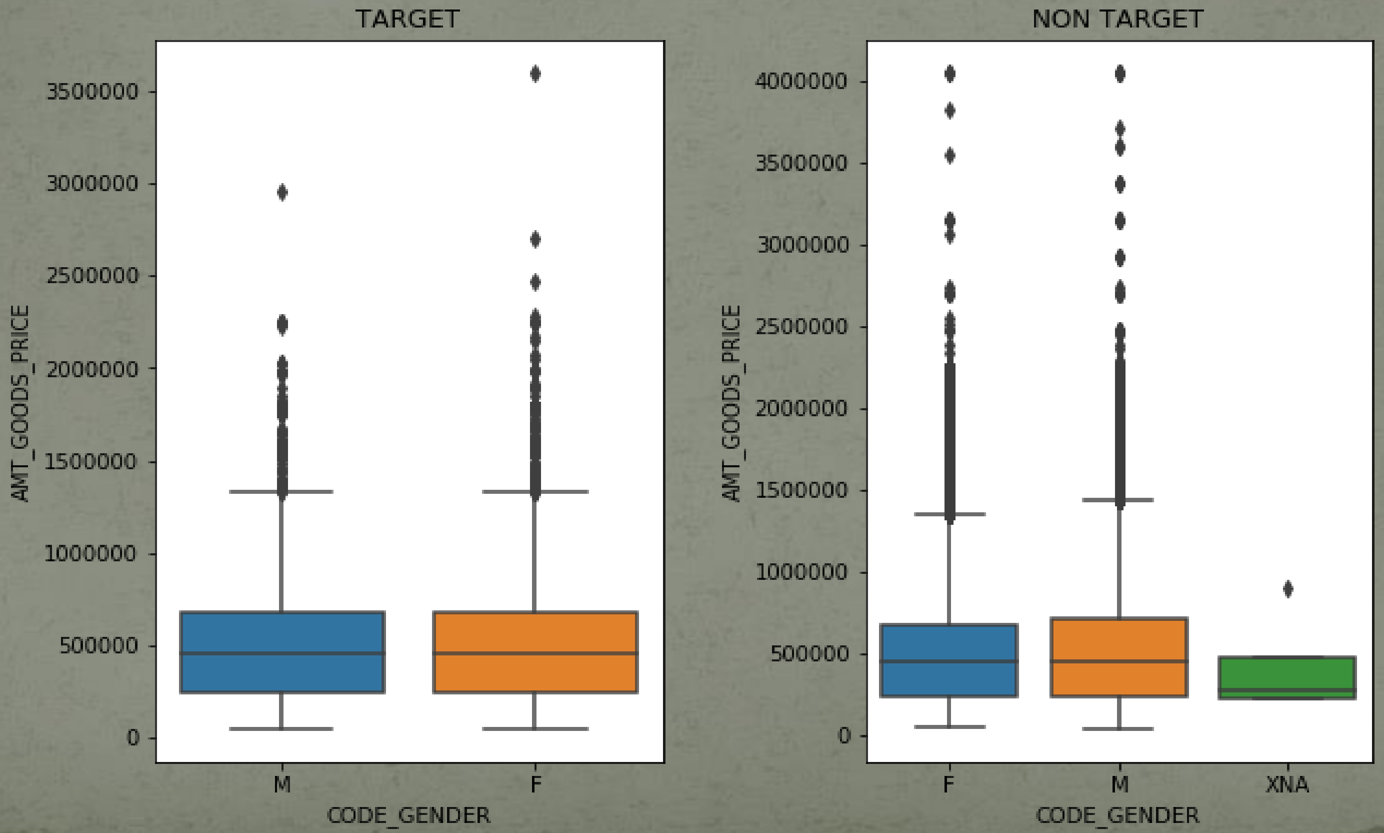
Bivariate Analysis (Continuous-Categorical)

- **CODE_GENDER vs AMT_INCOME_TOTAL:**
  - In TARGET, the outlier has to be removed.
  - In NON TARGET, the four outliers can be removed and the outliers present as a group can handled using binning or capping methods.
  - XNA to be treated as missing value.

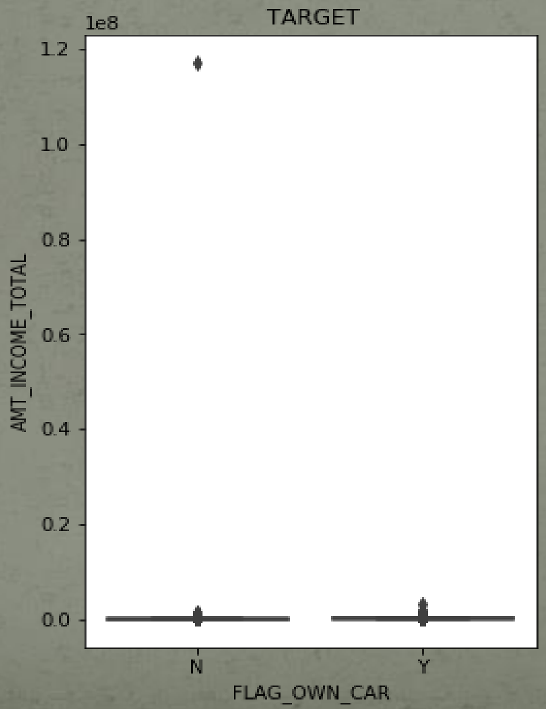


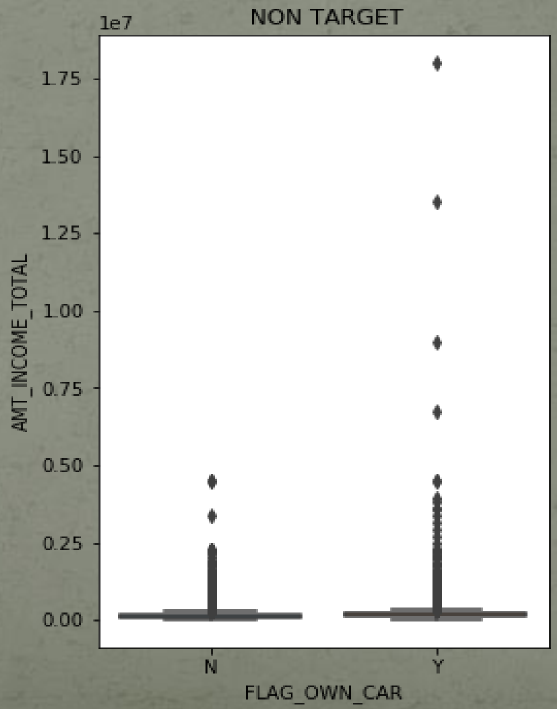


- **CODE_GENDER vs AMT_GOODS_PRICE:**
  - The mean and IQR range is same for both TARGET and NON TARGET.
  - Also, outliers to be treated using binning or capping methods and XNA to be considered as missing values.

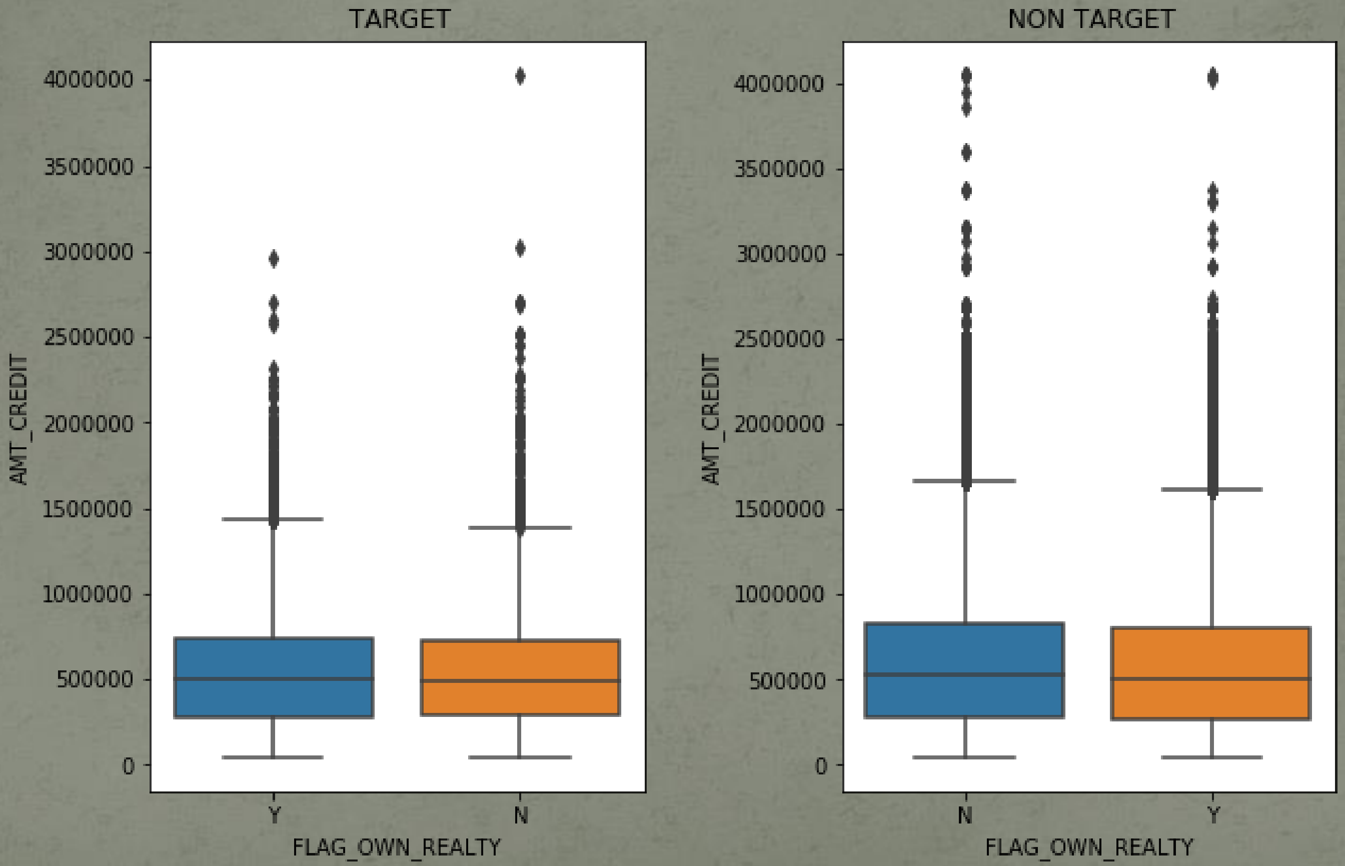


- **FLAG_OWN_CAR vs AMT_INCOME_TOTAL:**
  - In TARGET, for FLAG_OWN_CAR = N, there is a data point at an extreme end. A customer having very good salary but not owning a car? Either the AMT_INCOME_TOTAL or FLAG_OWN_CAR value might be wrong, in which case it can be treated as an outlier.
  - In NON TARGET, the outliers should be treated either by binning or capping methods.

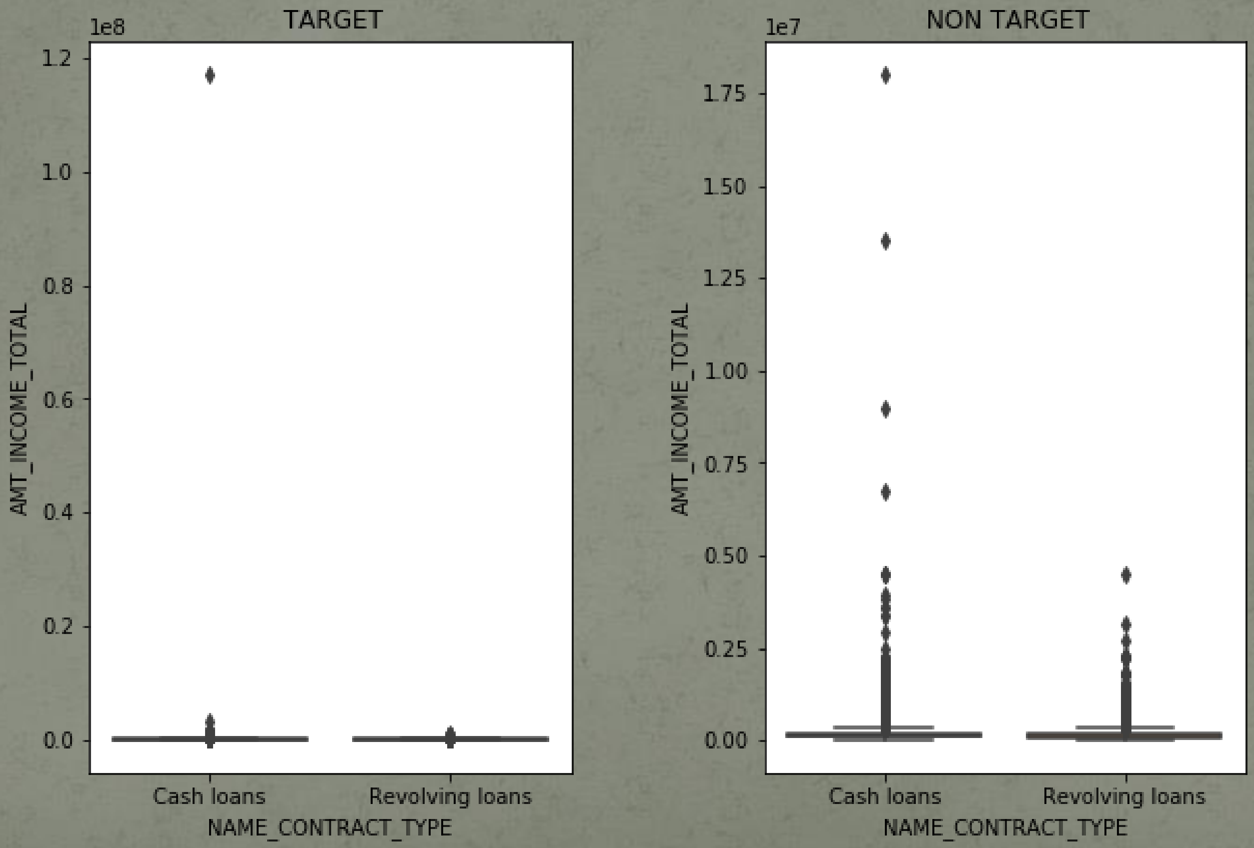




- **FLAG_OWN_REALTY vs AMT_CREDIT:**
  - The mean and IQR is more for TARGET than NON TARGET. Outliers to be handled using capping method.

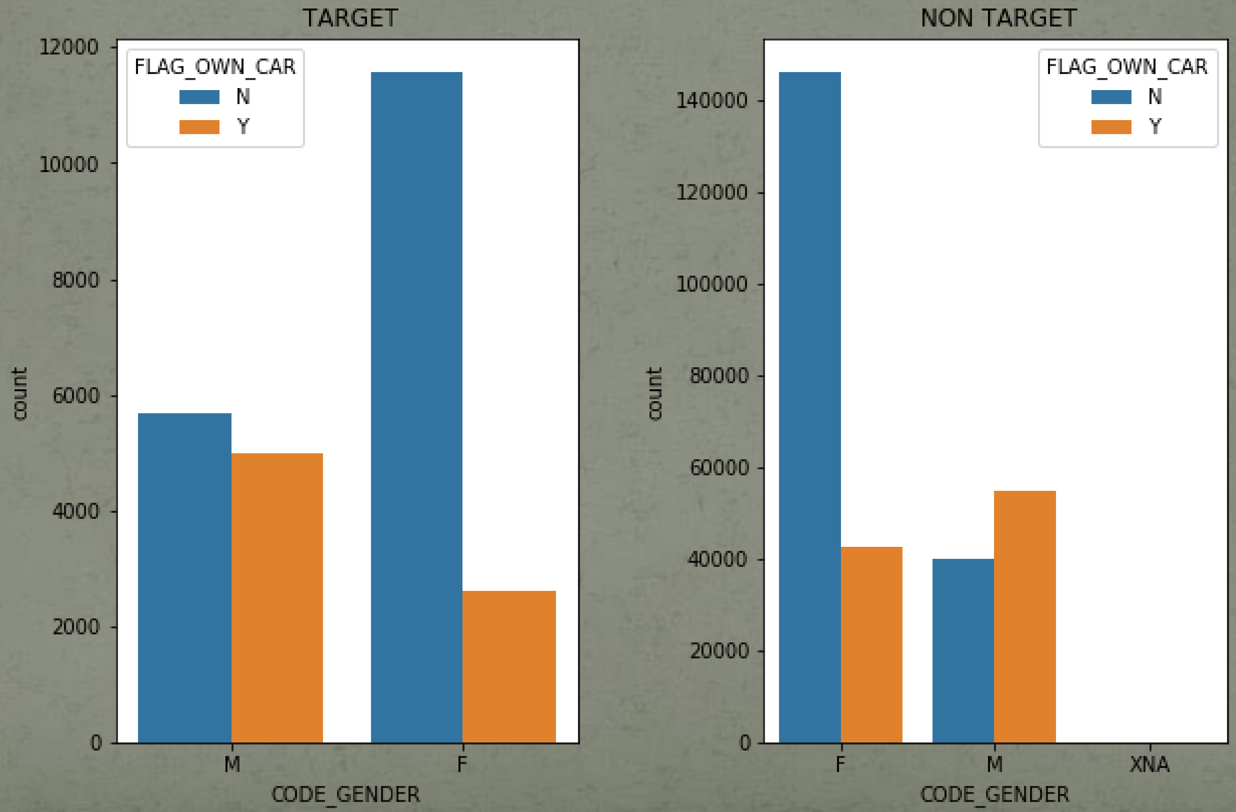


- **NAME_CONTRACT_TYPE vs AMT_INCOME_TOTAL:**
  - Except with the outliers which can be handled with capping and extreme one to be deleted, Cash loans are more when compared to Revolving loans in both plots.

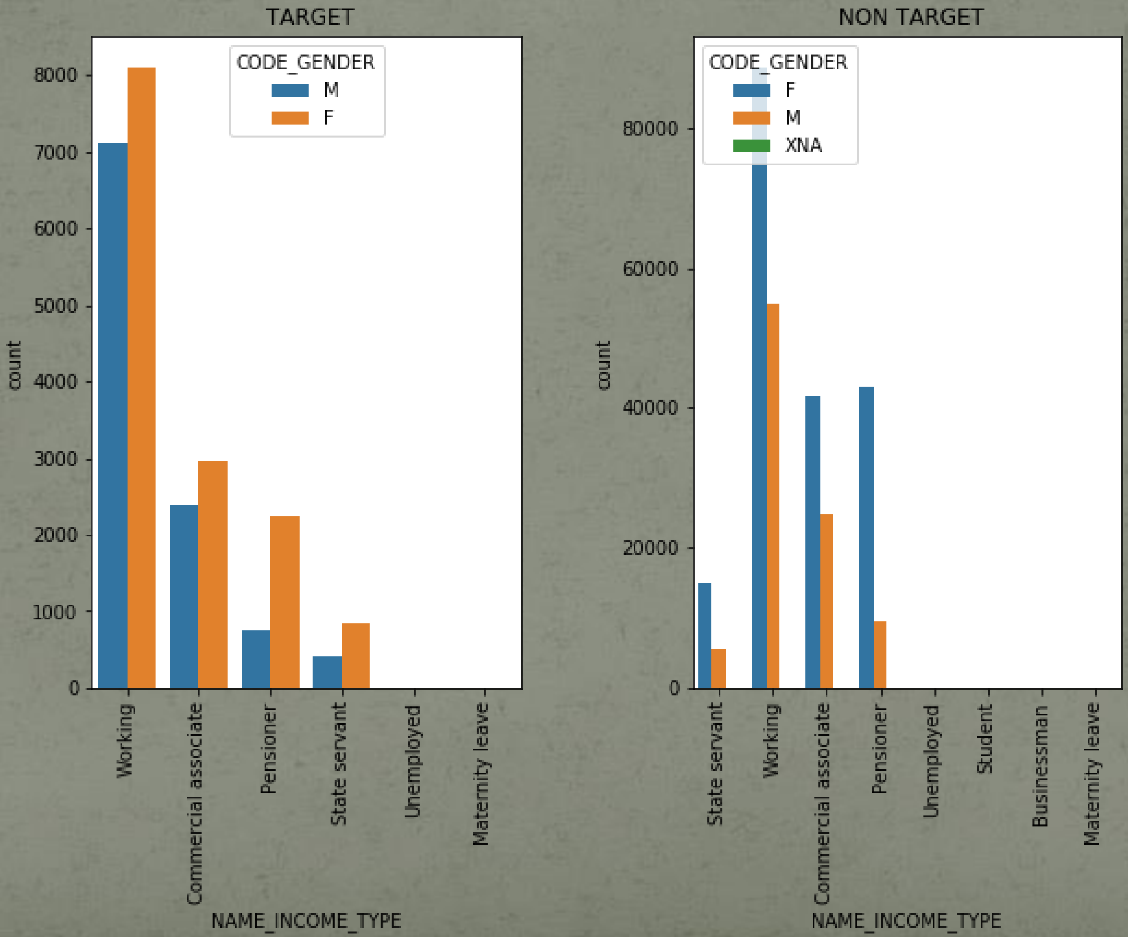


Bivariate Analysis (Categorical-Categorical)

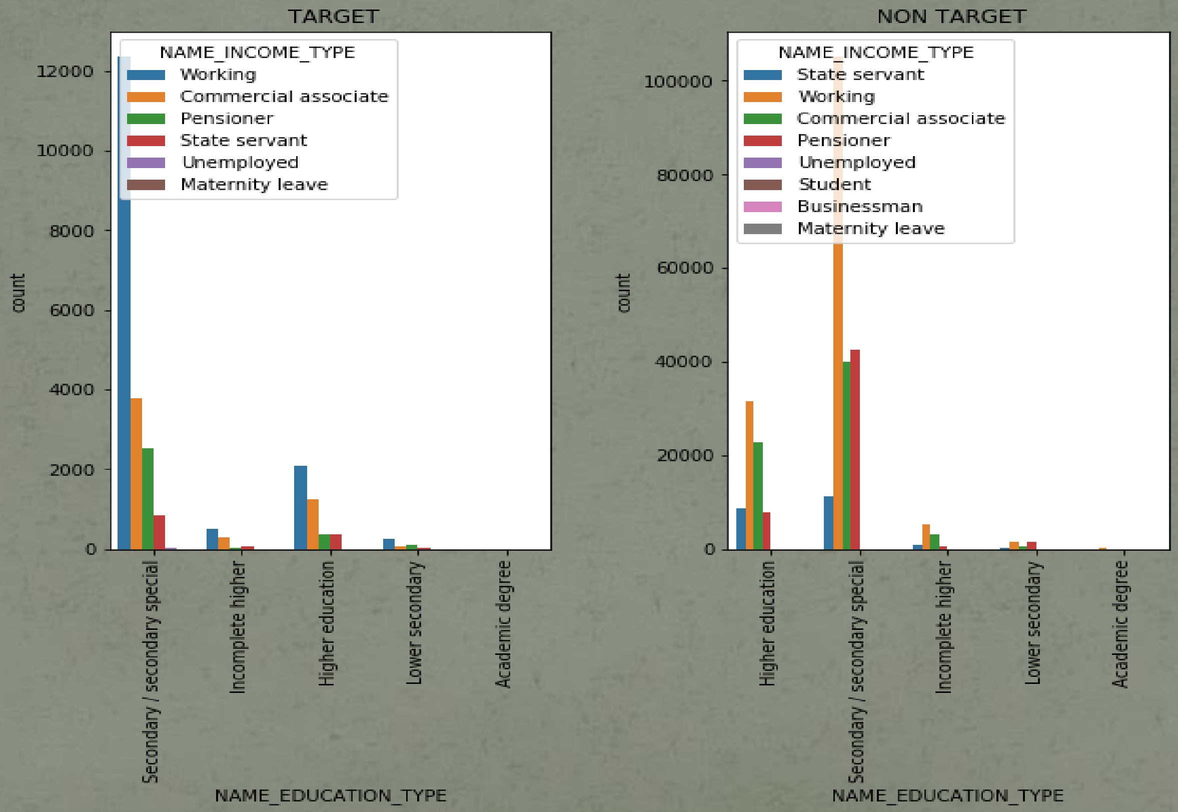
- CODE_GENDER vs FLAG_OWN_CAR:
  - Irrespective of TARGET or NON TARGET, there are more no. of male customers who own car. Also, it is more in TARGET.



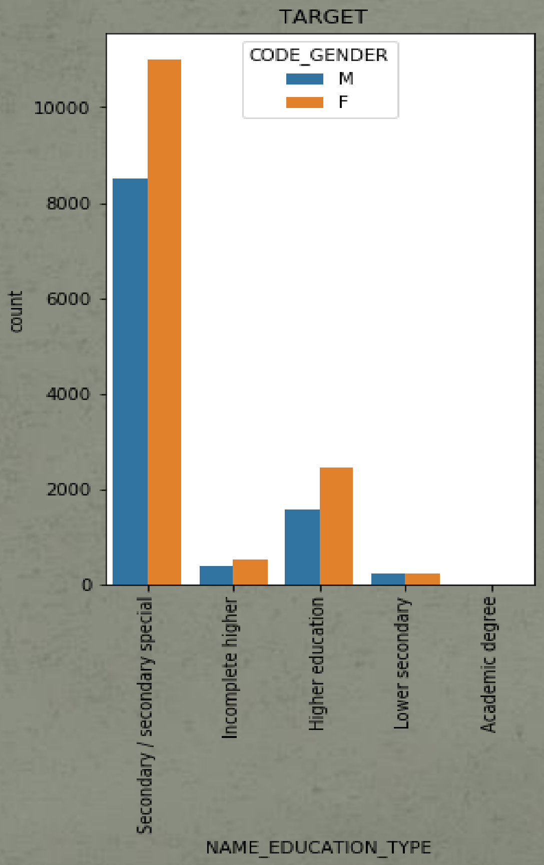
- **NAME_INCOME_TYPE vs CODE_GENDER:**
  - It can be seen that female customer numbers are more in being employed irrespective of TARGET variable.

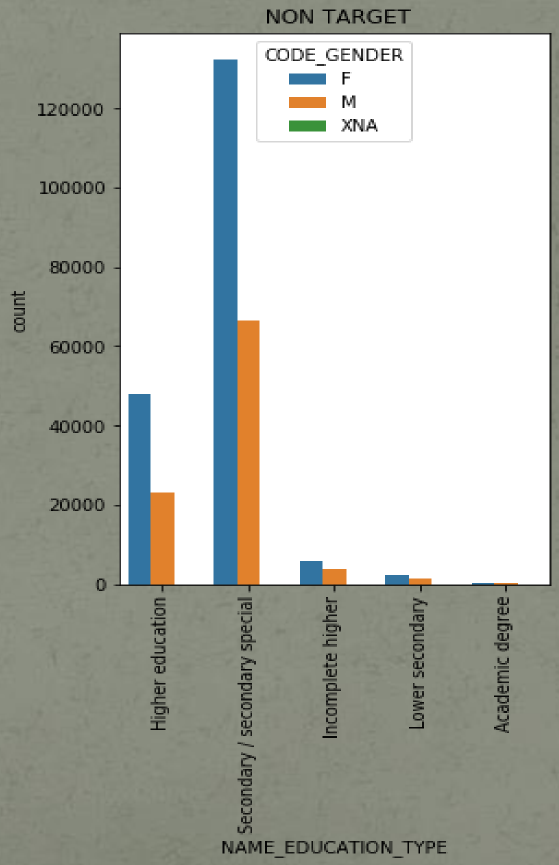


- **NAME_EDUCATION_TYPE vs NAME_INCOME_TYPE:**
  - The no. of customers who have 'Incomplete Higher' and 'Lower Secondary' education is more in TARGET than NON TARGET which in turn results in less no. of people getting employed or getting income.



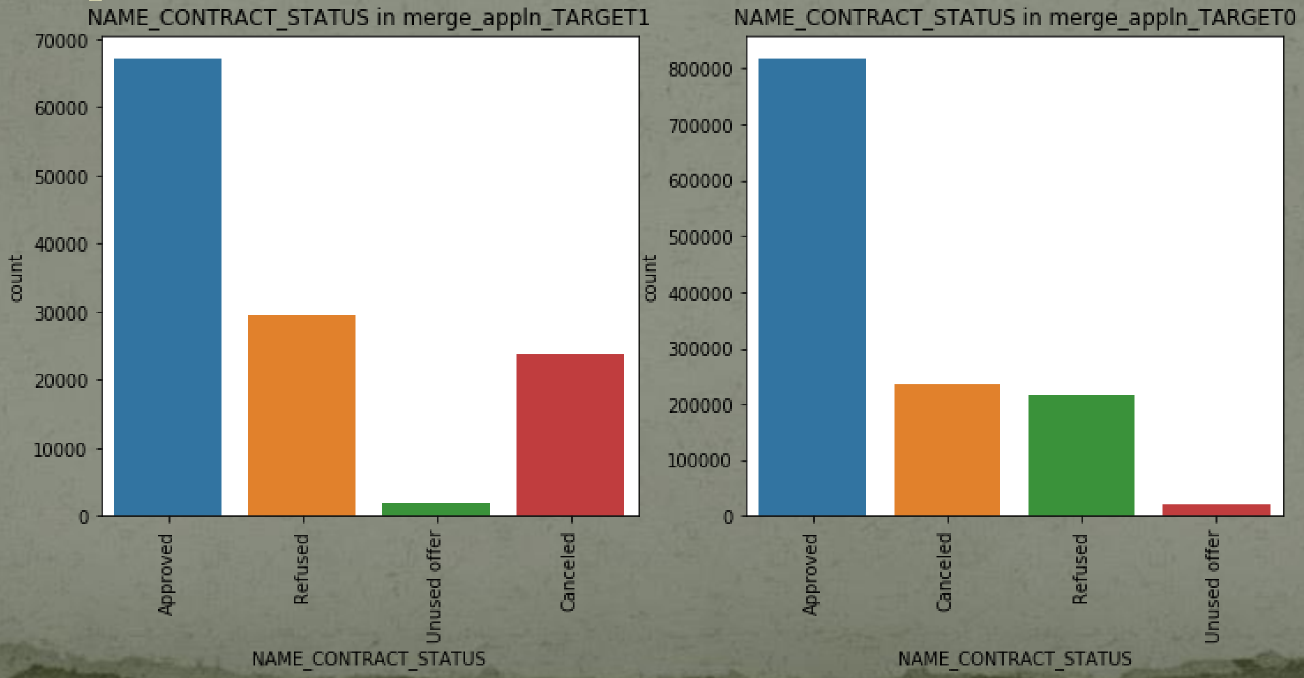
- **NAME_EDUCATION_TYPE vs CODE_GENDER:**
  - No much inferences from the plots. Only thing is the education level of Male customers number is less when compared to female customers.


TARGET

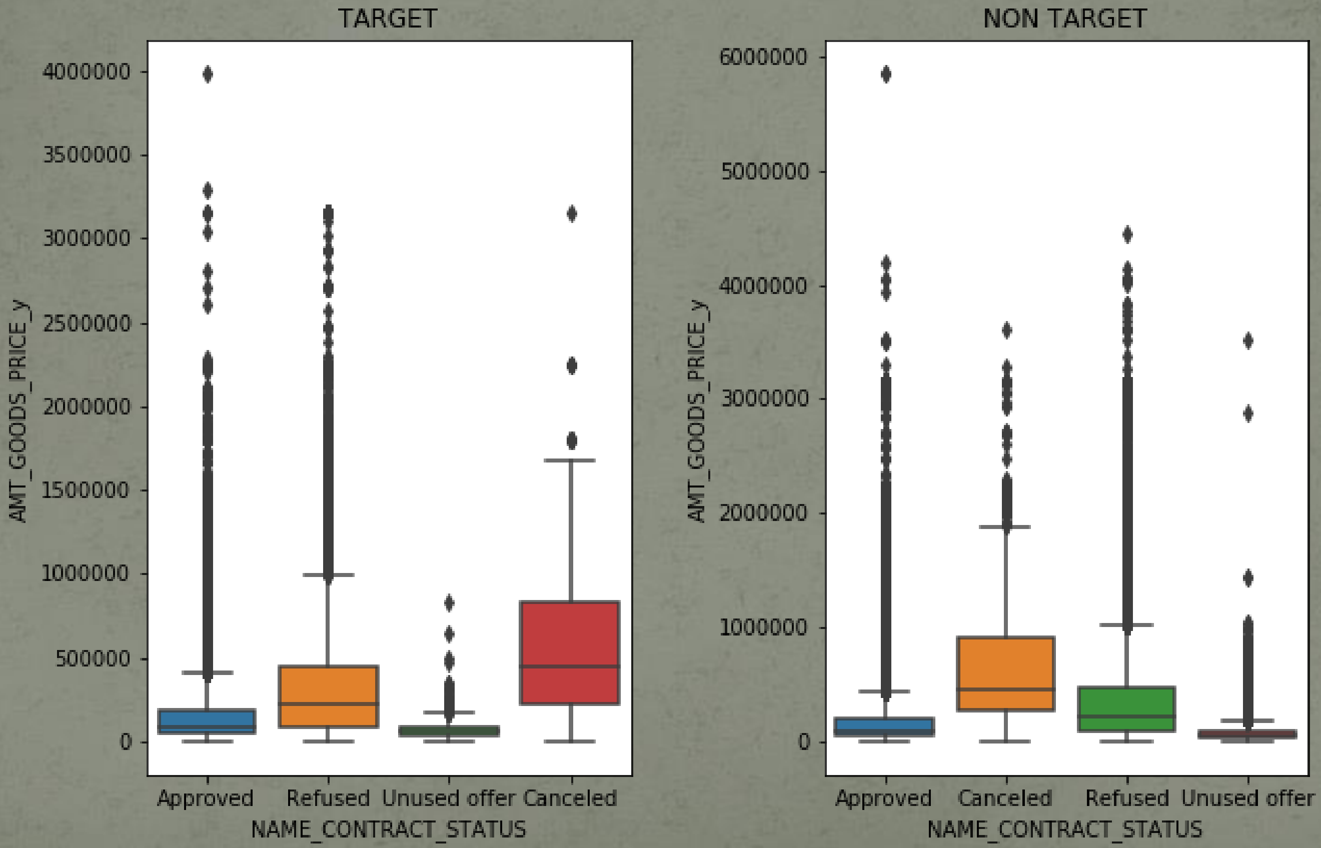

NON TARGET

Analysis on Previous Application Dataset:

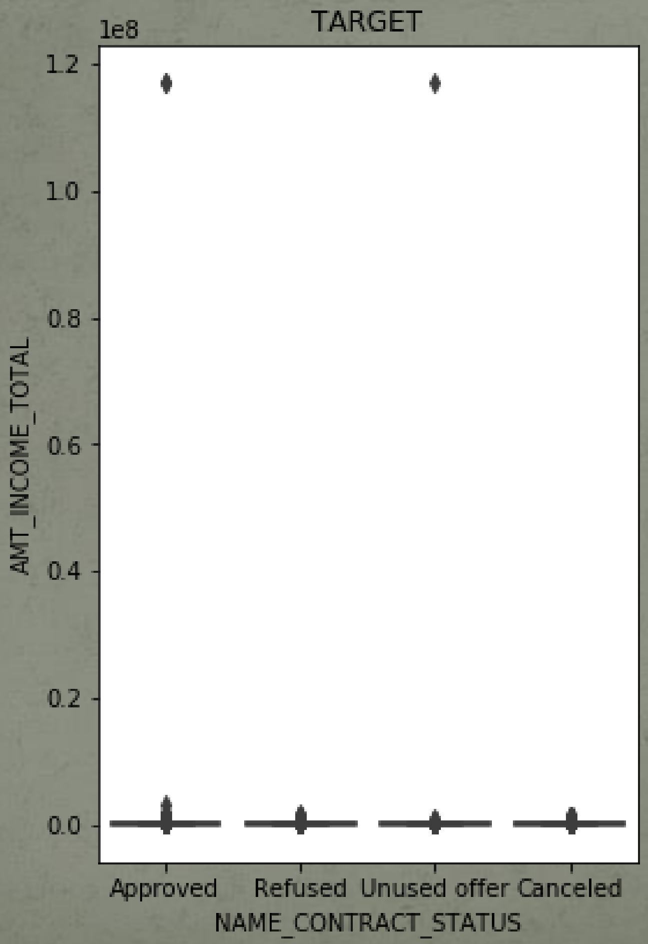
- **NAME_CONTRACT_STATUS:**
  - The percentage of refusal is more for TARGET as the requirements is not met by the client when compared to NON TARGET customers.
  - The percentage of approval is more in NON TARGET when compared to TARGET customers.

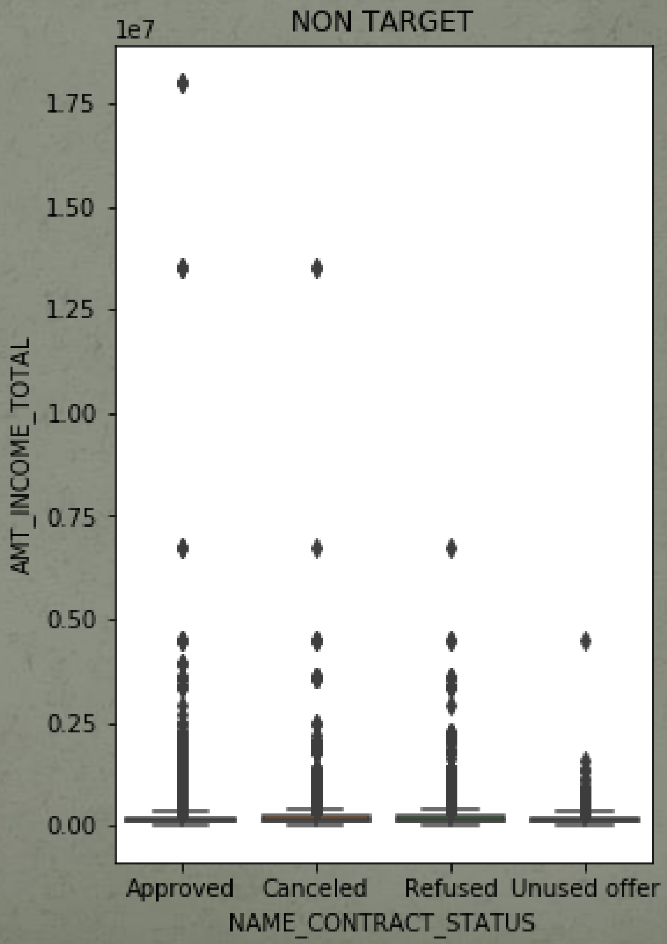


- **NAME_CONTRACT_STATUS vs AMT_GOODS_PRICE_y:**
  - The IQR of 'Canceled' application as well as 'Refused' application is more in TARGET plot when compared to NON TARGET customers.



- **NAME_CONTRACT_STATUS vs AMT_INCOME_TOTAL:**
  - The income for NON TARGET customers is more when compared to TARGET customers.





Final Words

- With all these analysis it can be inferred that the following columns might be the driving factor for loan default:
  - AMT_INCOME_TOTAL
  - AMT_ANNUITY
  - AMT_CREDIT
  - NAME_EDUCATION_TYPE
  - NAME_INCOME_TYPE
- These factors are related to each other.
  - AMT_INCOME_TOTAL is dependent on NAME_INCOME_TYPE which is in turn dependent on NAME_EDUCATION_TYPE.
  - AMT_CREDIT is related to AMT_INCOME_TOTAL.
  - AMT_ANNUITY is dependent on AMT_CREDIT.

- Considering the plots, NAME_CONTRACT_STATUS vs AMT_GOODS_PRICE_y and NAME_CONTRACT_STATUS vs AMT_INCOME_TOTAL, it can be infered that, to decrease the credit loss for NON TARGET customers, the no. of 'Refused' application needs to be minimised with certain measures that can be taken by the bank as the total income of NON TARGET customers are fairly good when compared to TARGET customers thereby increasing the 'Approved' application count.

- Also, with the same parameter i.e., AMT_INCOME_TOTAL, the interest loss can be minimised for TARGET customers.

THANK YOU