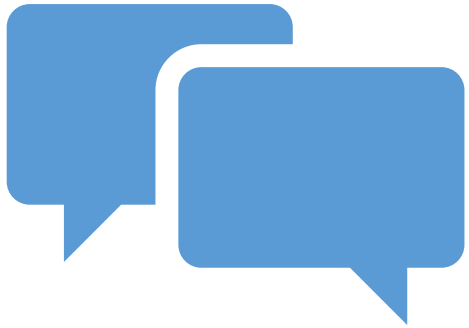



Credit EDA Case Study

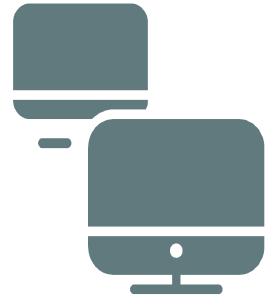
Problem Statement:

To analyse the patterns in dataset of loan providing companies and consumers, which will be used for reducing risk against defaulters and ensuring that the consumers capable of repaying the loan are not rejected.

Analysis Approach:

Following EDA steps starting from inspecting both the dataframe and doing data cleaning, variable transformation with missing value and outliers treatment then Non-Graphical Univariate Analysis to Graphical Univariate Analysis continuing with Bivariate Analysis and at last Correlation analysis.

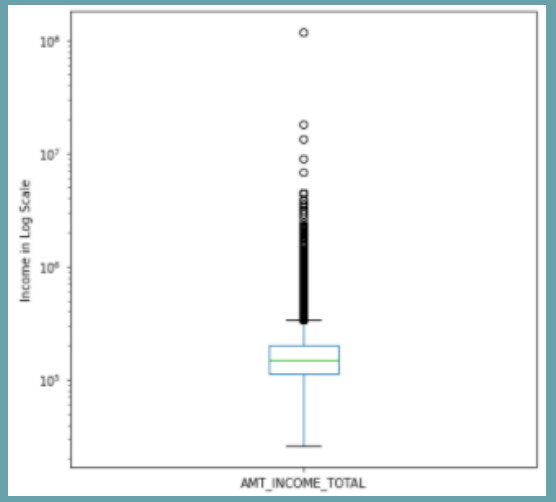
- INSPECTING THE DATAFRAME (APPLICATION)

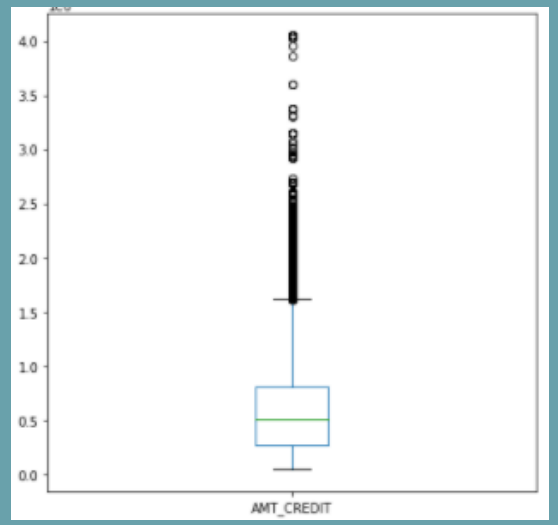


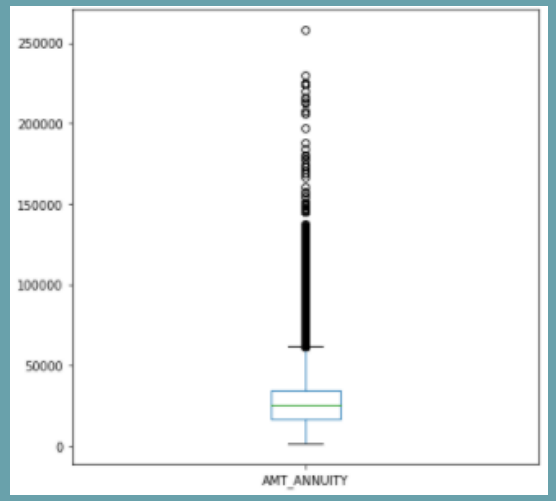
- DATA CLEANING
  - Removing columns and rows where NA values are >=30%
  - Filling up null values for columns

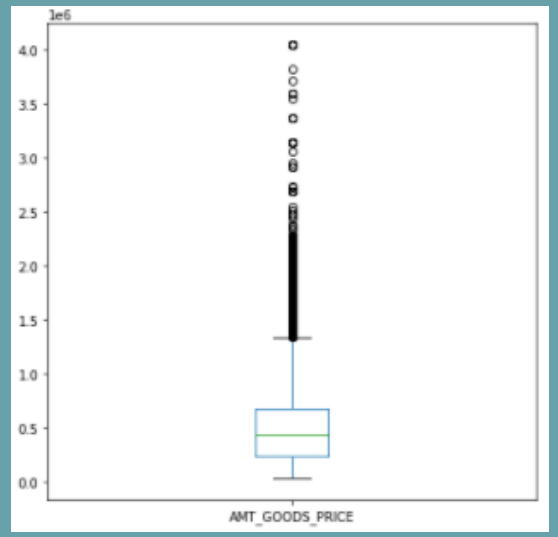
Columns with NA	Impute By	Inference
AMT_ANNUITY, AMT_GOODS_PRICE	MEDIAN	Difference between .95 quantile and max value.
CNT_FAM_MEMBERS, EXT_SOURCE_2	MEAN	Estimated average of the data
CODE_GENDER , NAME_FAMILY_STATUS, ORGANIZATION_TYPE , NAME_TYPE_SUITE	MODE	Since non numeric and categorically unordered variable.

  - Converting columns to absolute value
    DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE





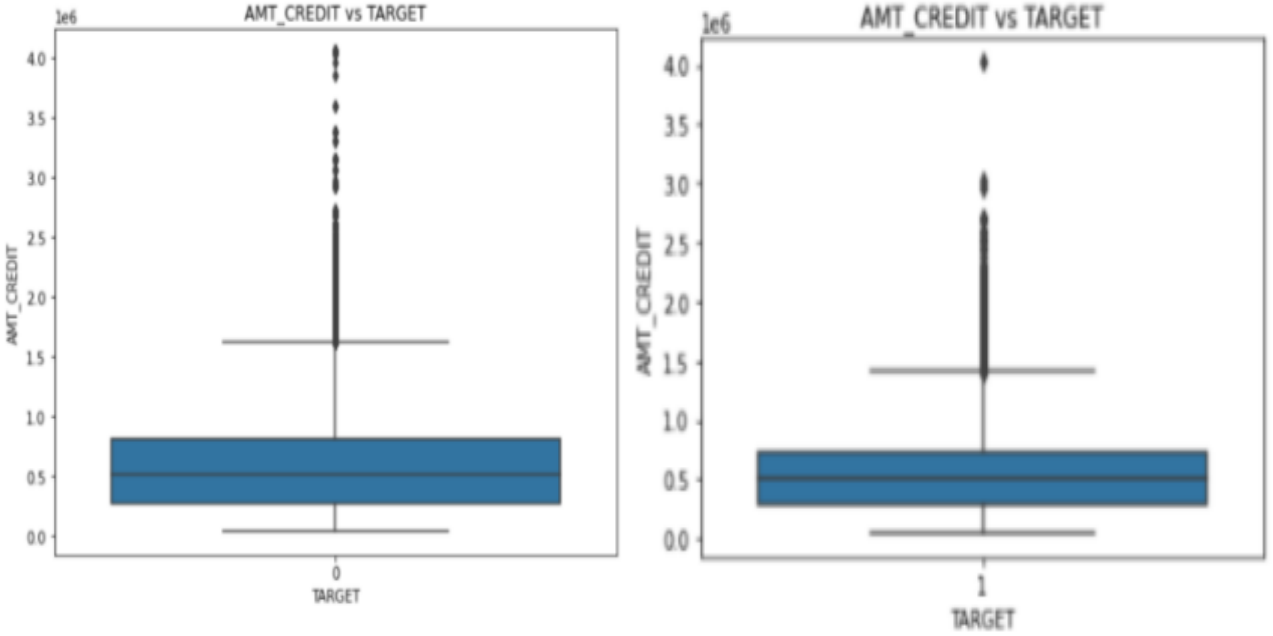




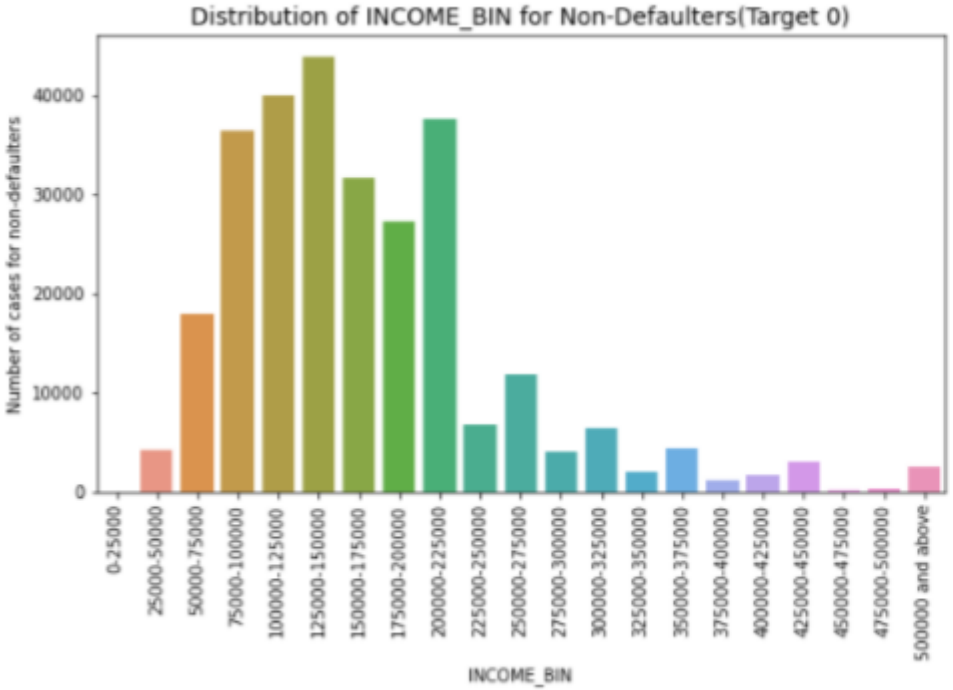
- Derived column AGE_YEARS and AGE_GROUP from DAYS_BIRTH column.
- Capping numeric variable outliers between .95 quantile and max value to .95 quantile.
- Binning for columns AMT_INCOME_TOTAL, AMT_CREDIT, AGE_YEARS.
- Data Imbalance Check: The data is skewed towards non defaulted population i.e. TARGET 0

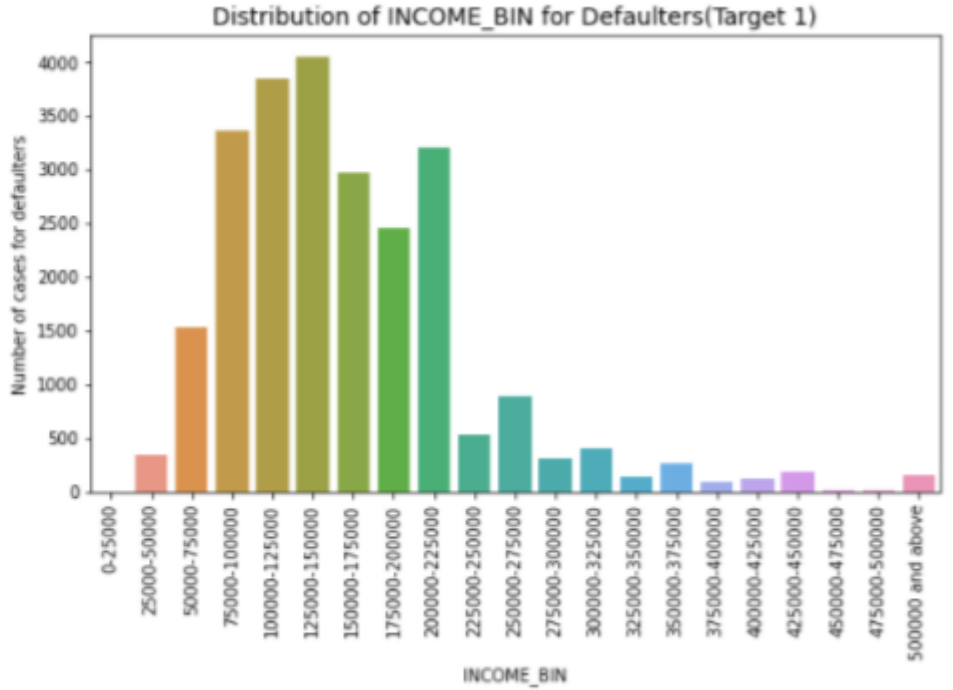
- ANALYSIS

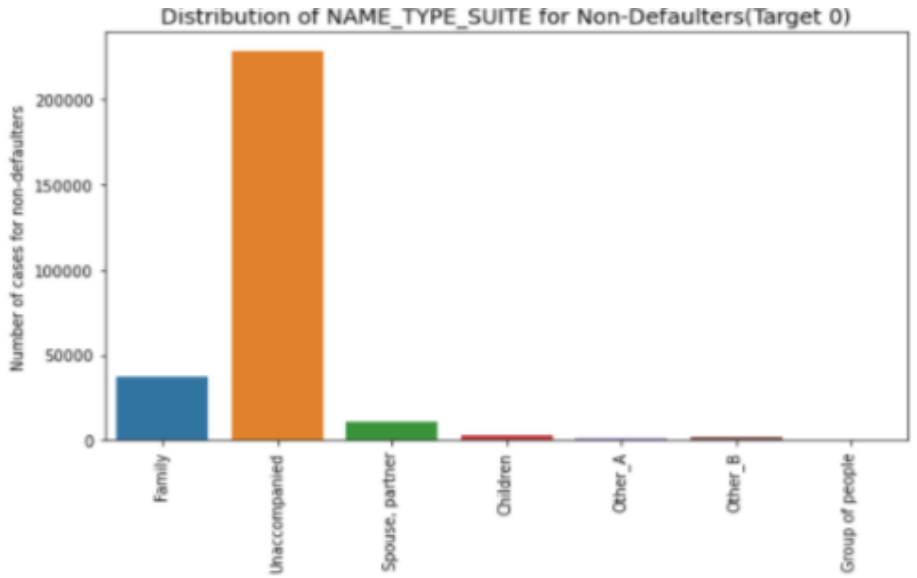
Median of AMT_CREDIT is similar for both the targets. People with higher amount of credit have more chances to fall in the category of ‘All other cases’



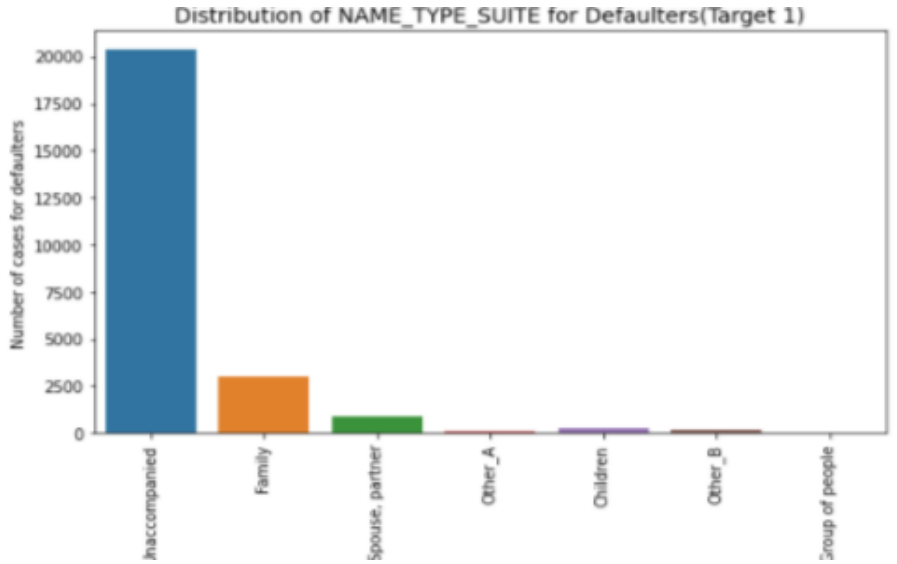
No effect of income on defaulters as the proportion is same 10:1 which is same as population proportion as calculated in the Data Imbalance.

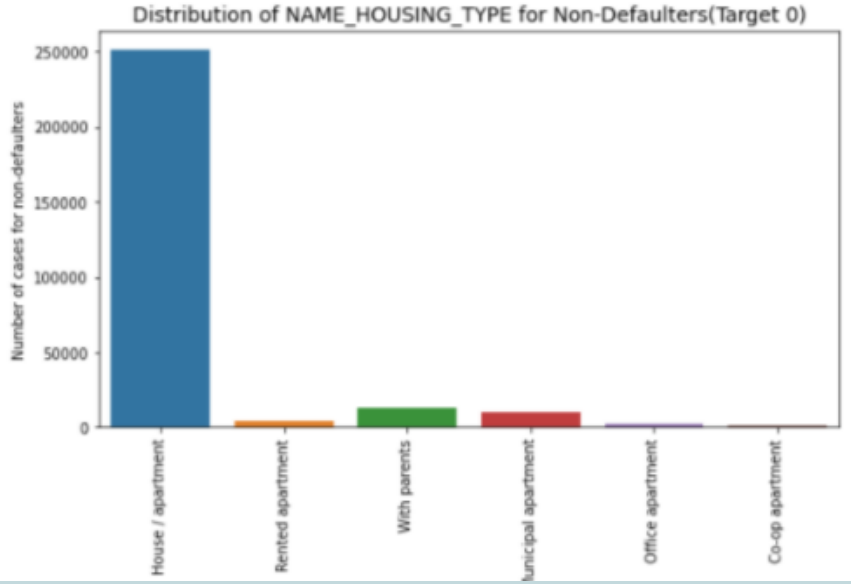




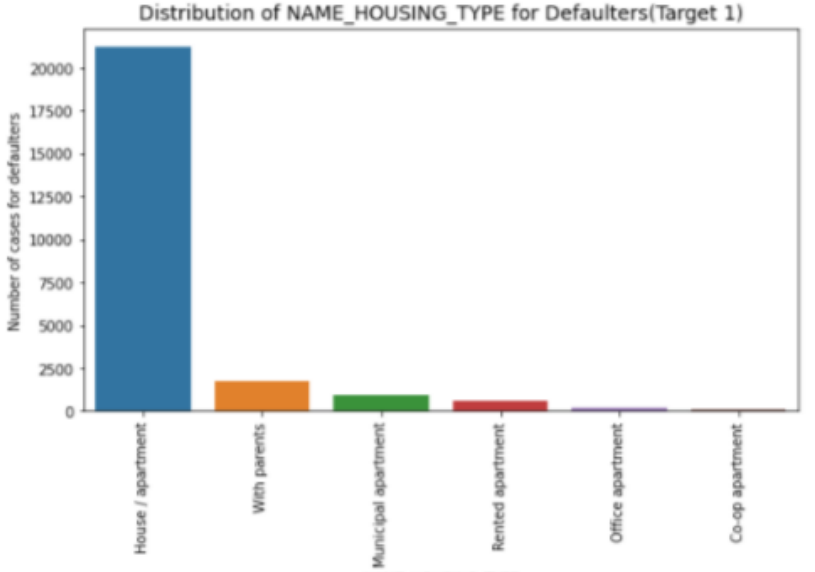


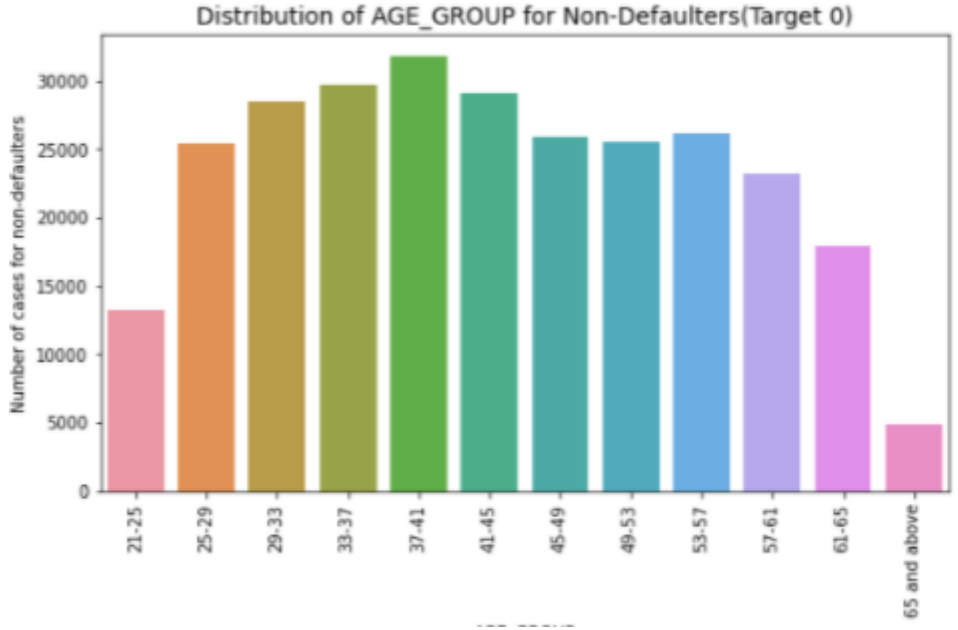
There is no impact of NAME_TYPE_SUITE on defaulters as the proportion is same 10:1 which is same as population proportion as calculated in the Data Imbalance.

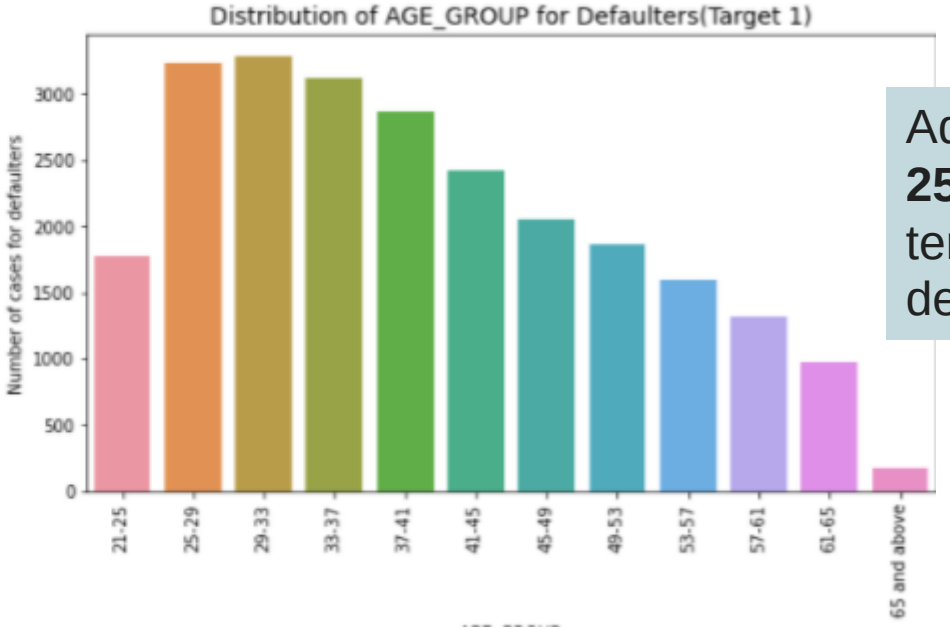




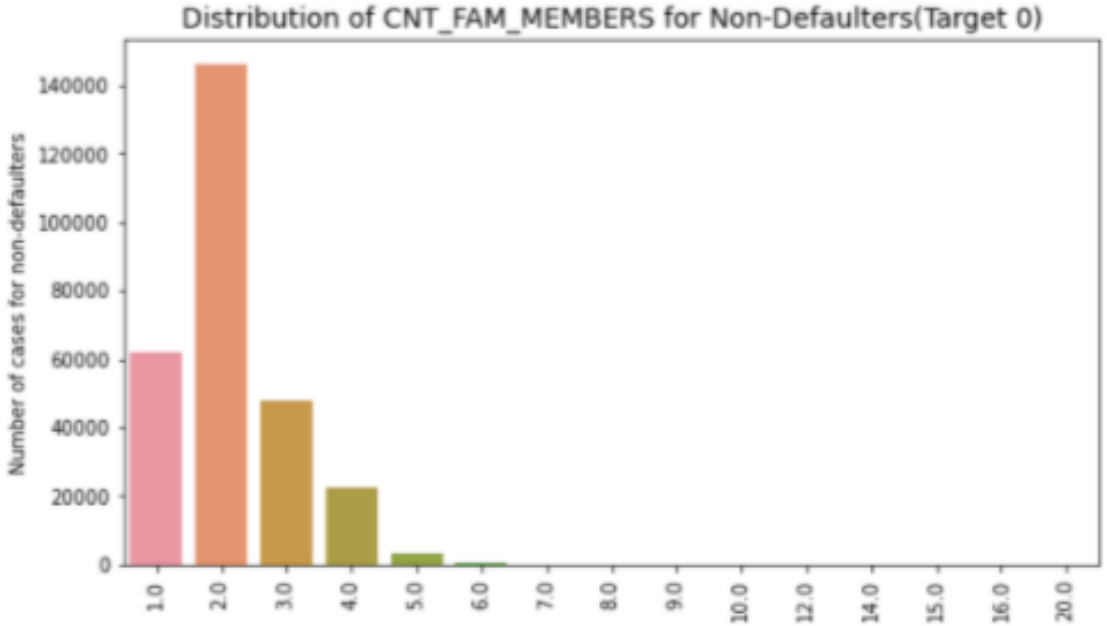
Living in Rented apartments and those living with parents have higher default rate as they have higher proportion in T1 (Defaulted population) in comparison to T0 (non)

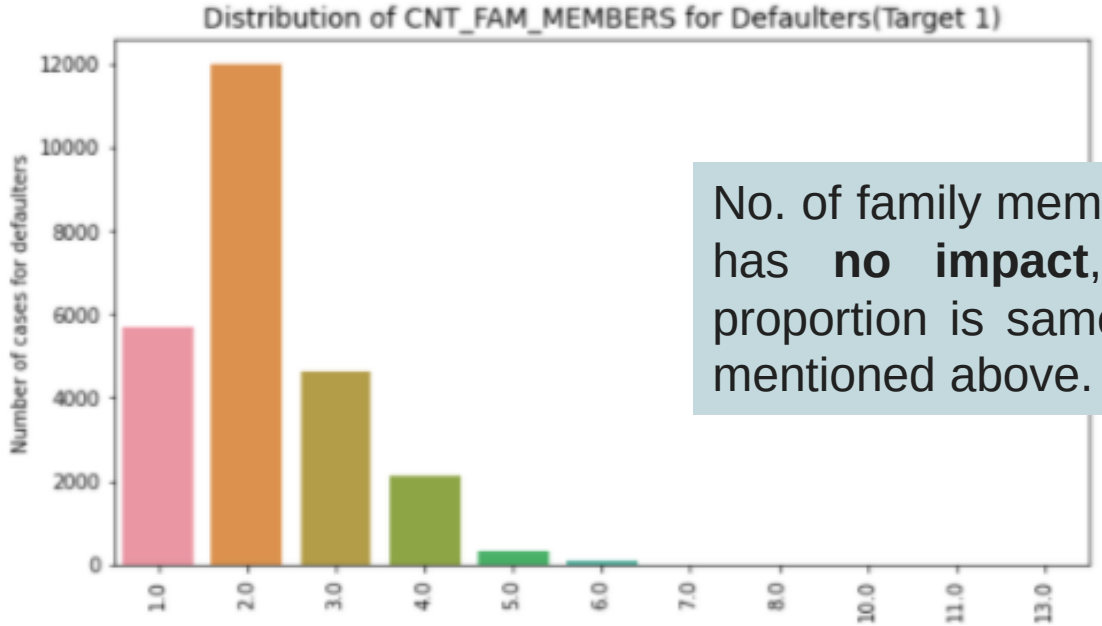






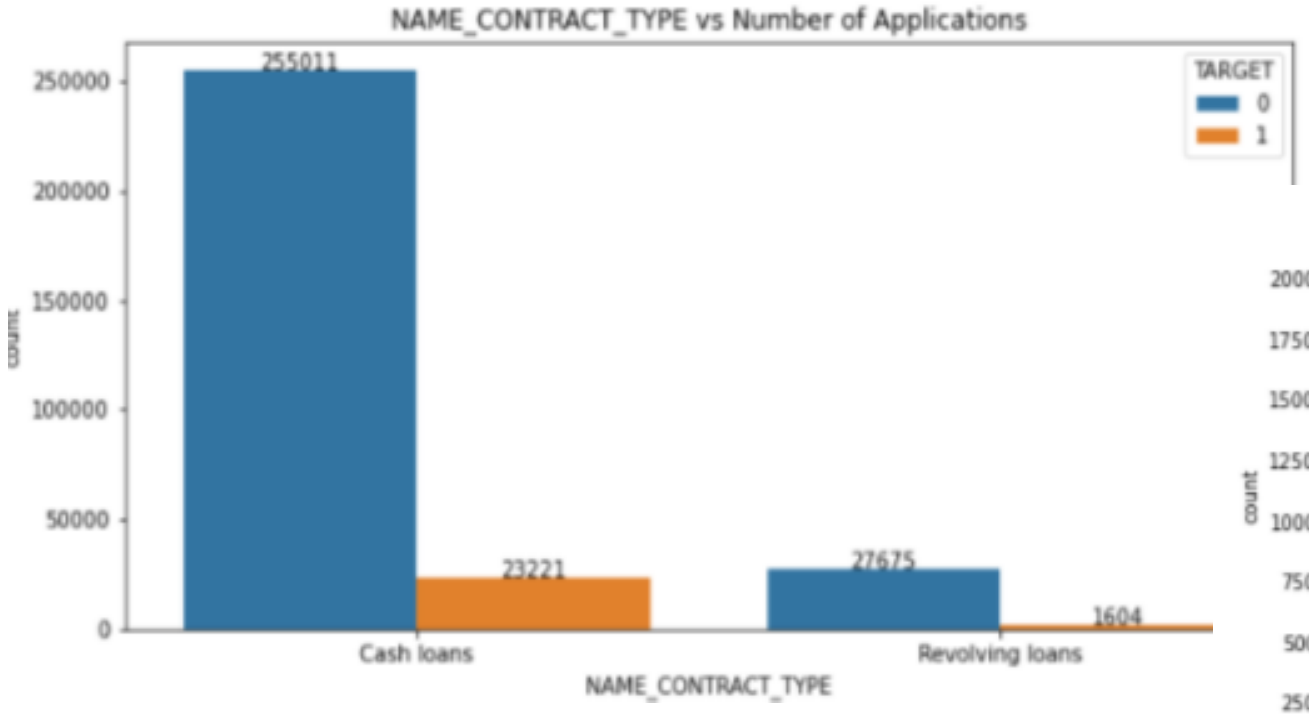
Adults age between **25-45** have higher tendency to be defaulters.

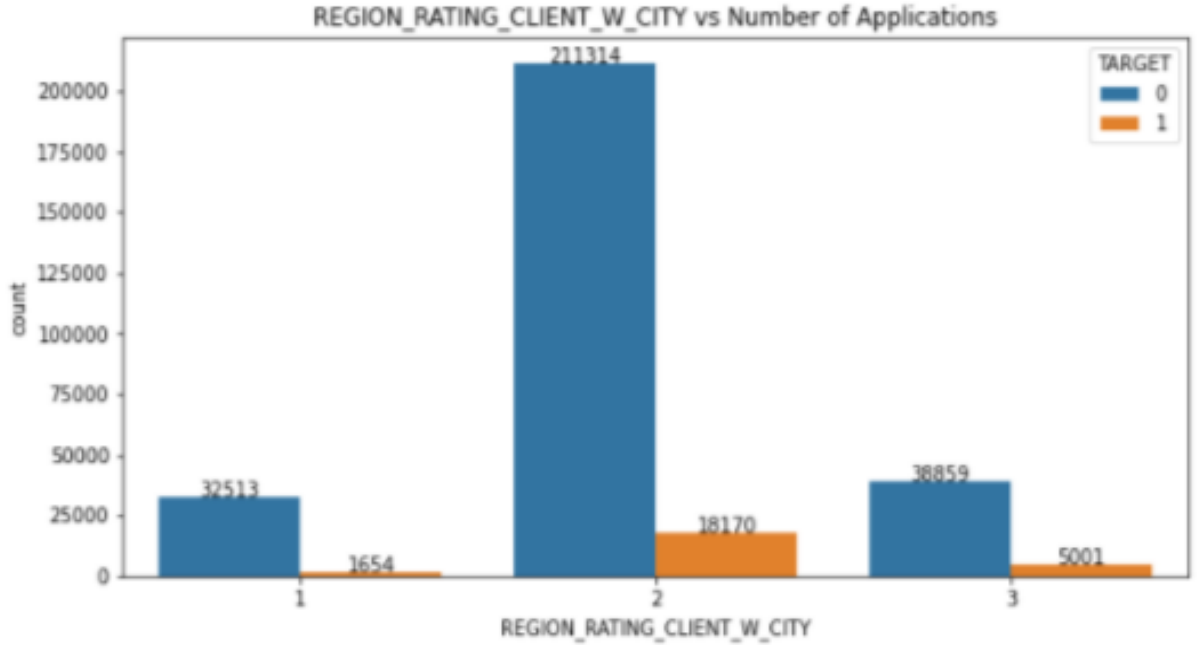




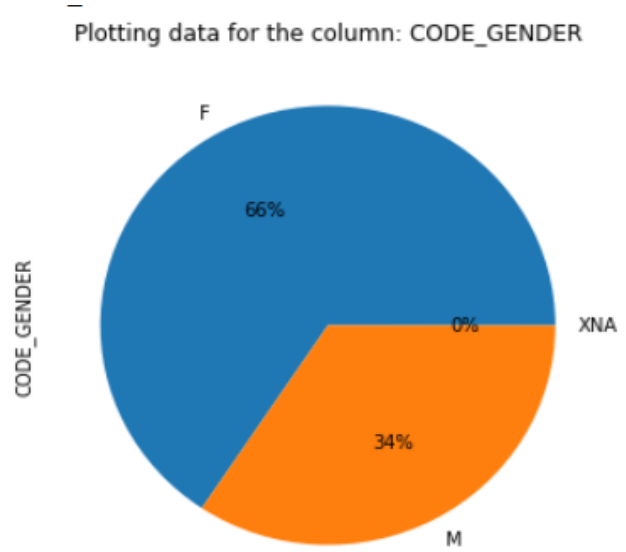
No. of family members has **no impact**, as proportion is same as mentioned above.

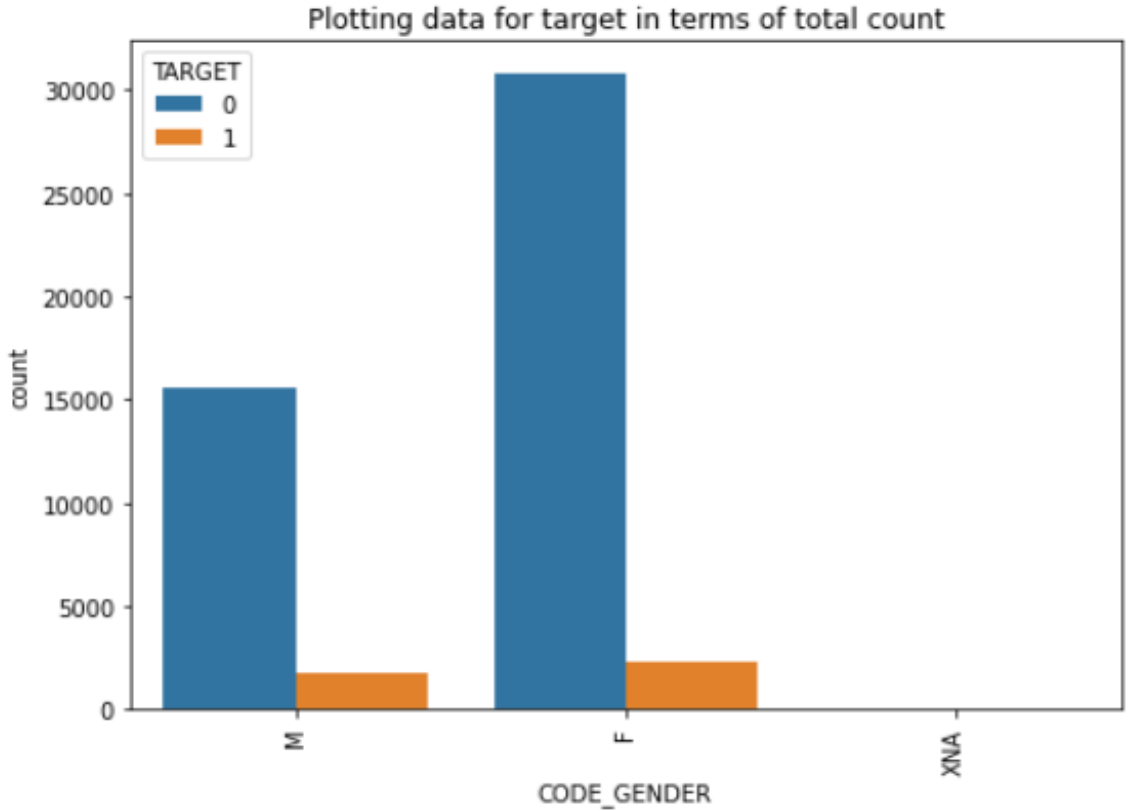
The revolving loans are lesser in Target: 1 for defaulters which means they have less chance to make a financial impact on the business.

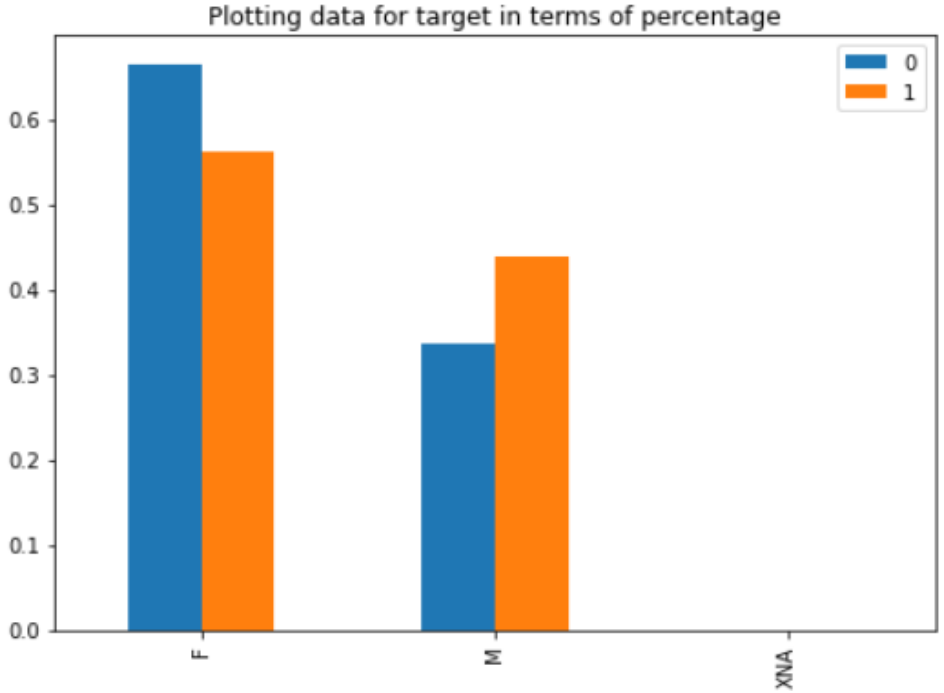




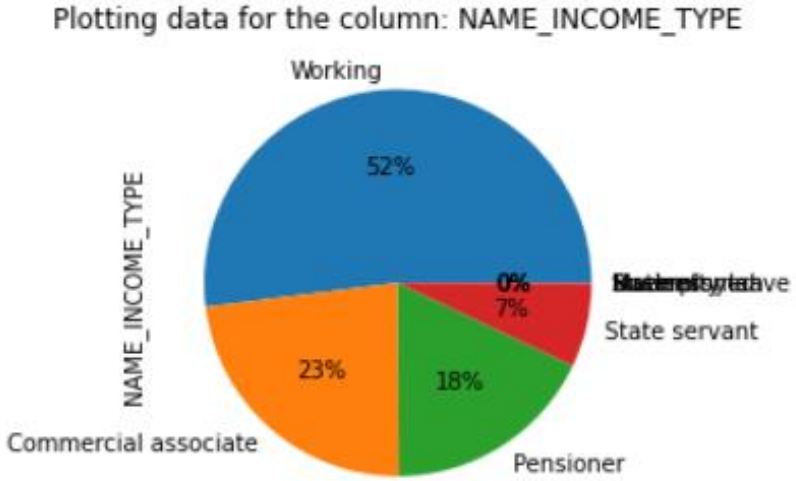
There are higher number of Payment Difficulties in category 2 cities compared to All other cases.

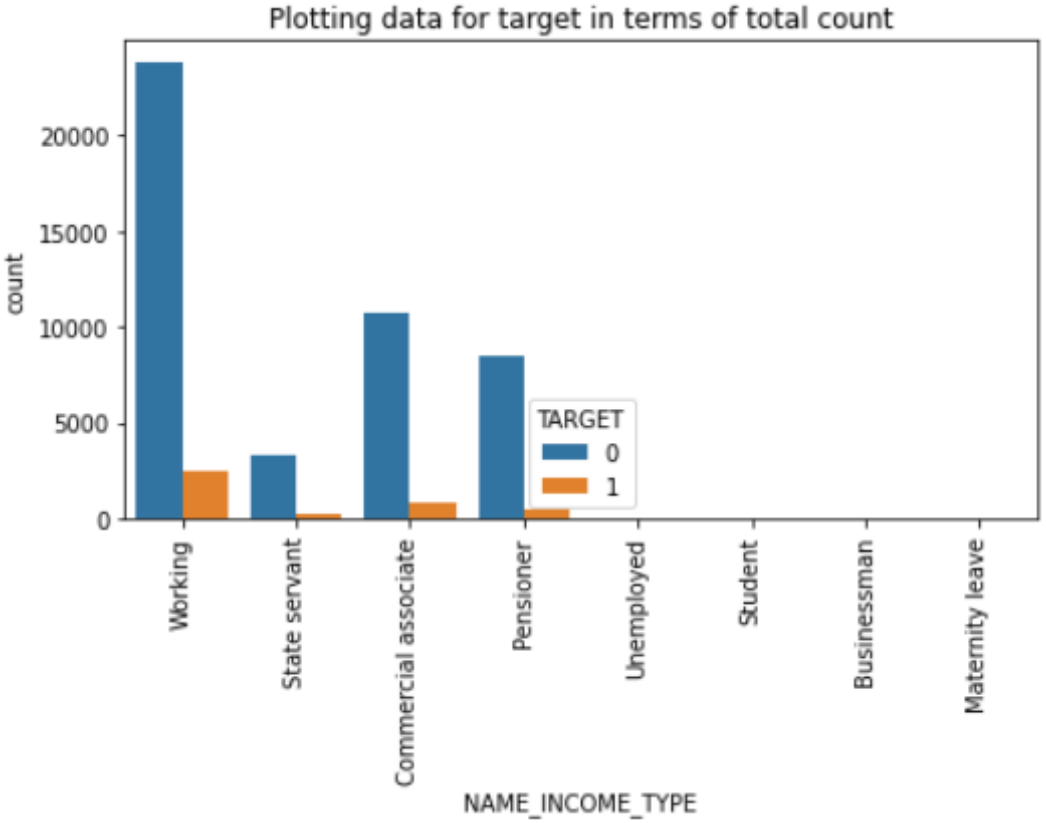


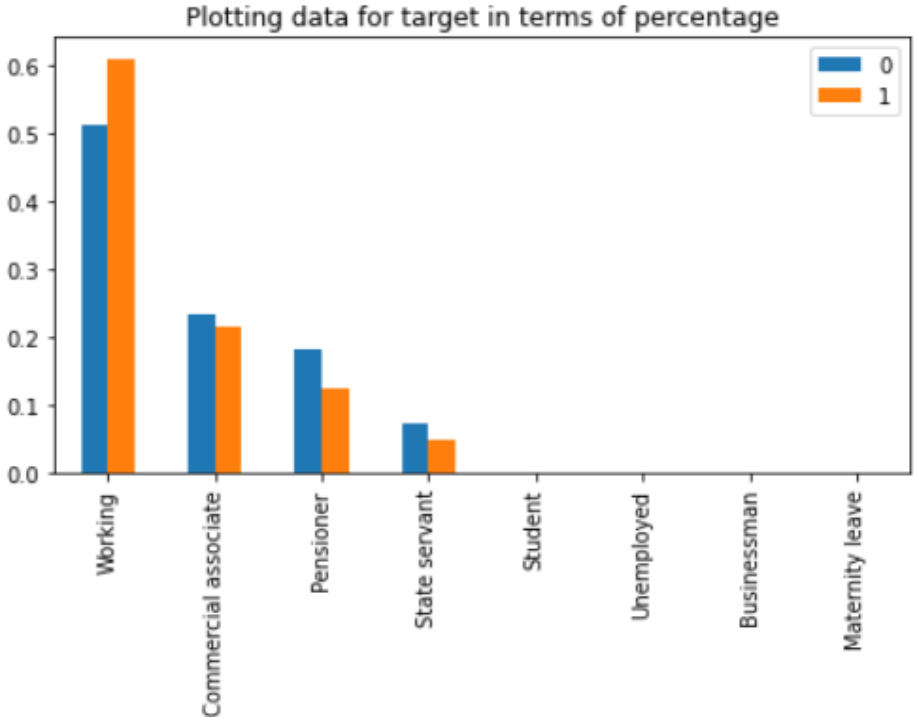




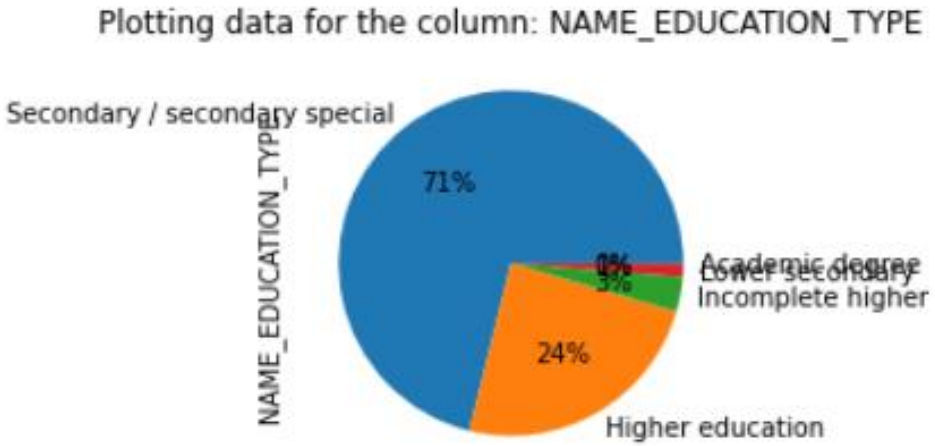
CODE_GENDER:
Less number of males take loan but the defaulters are higher in case of males.

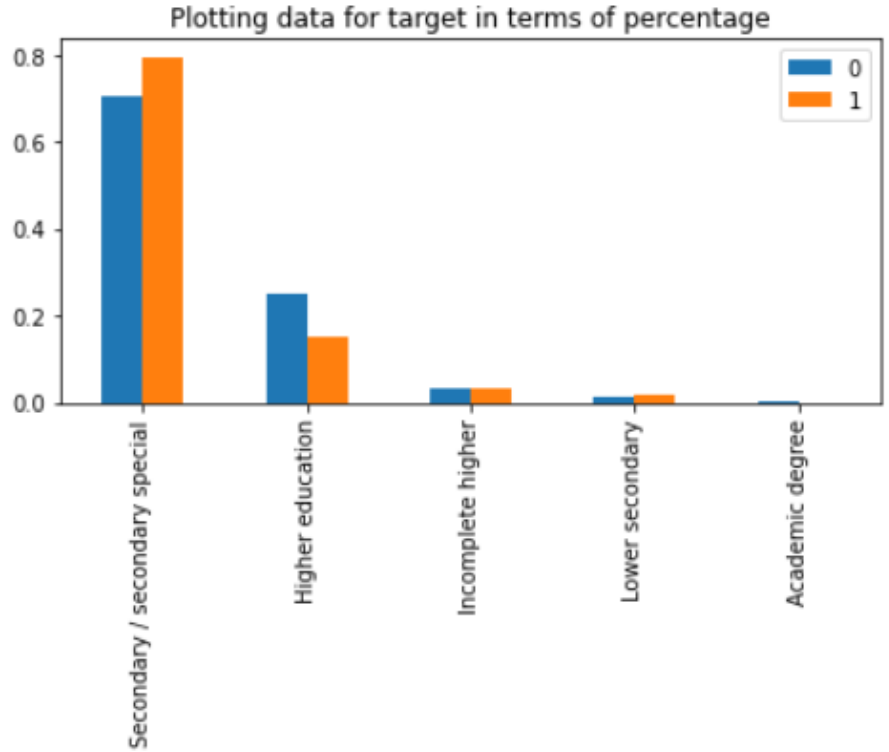


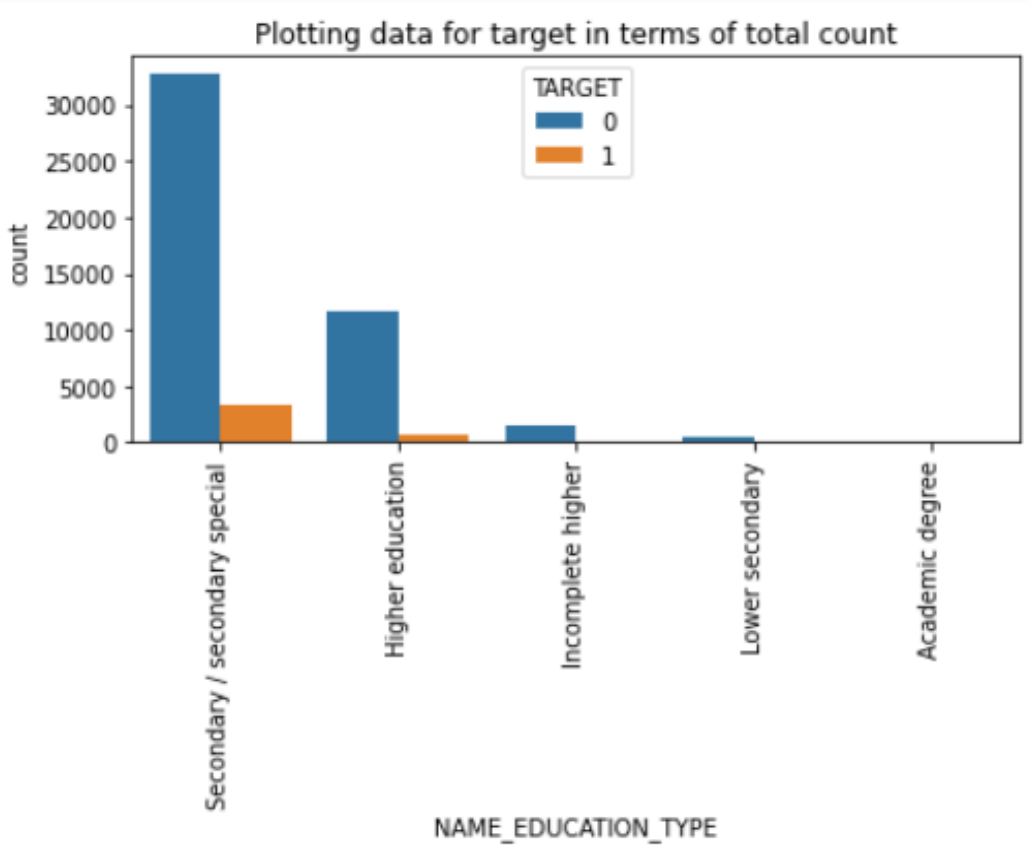




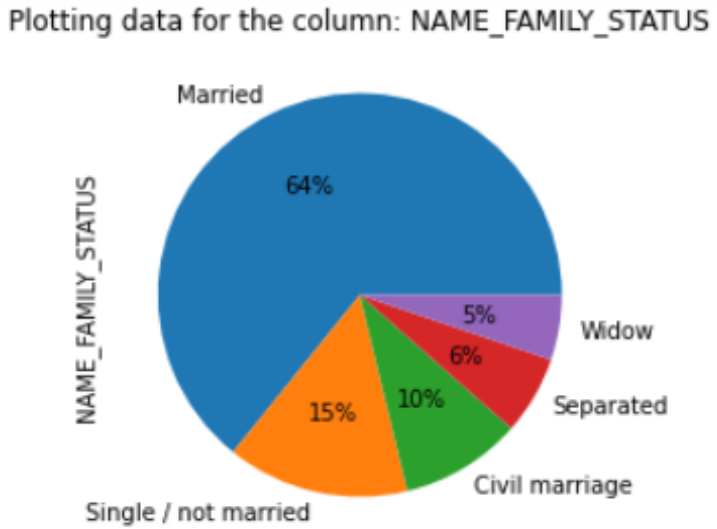
NAME_INCOME_TYPE:
Pensioner defaulter is lower than non-defaulter.

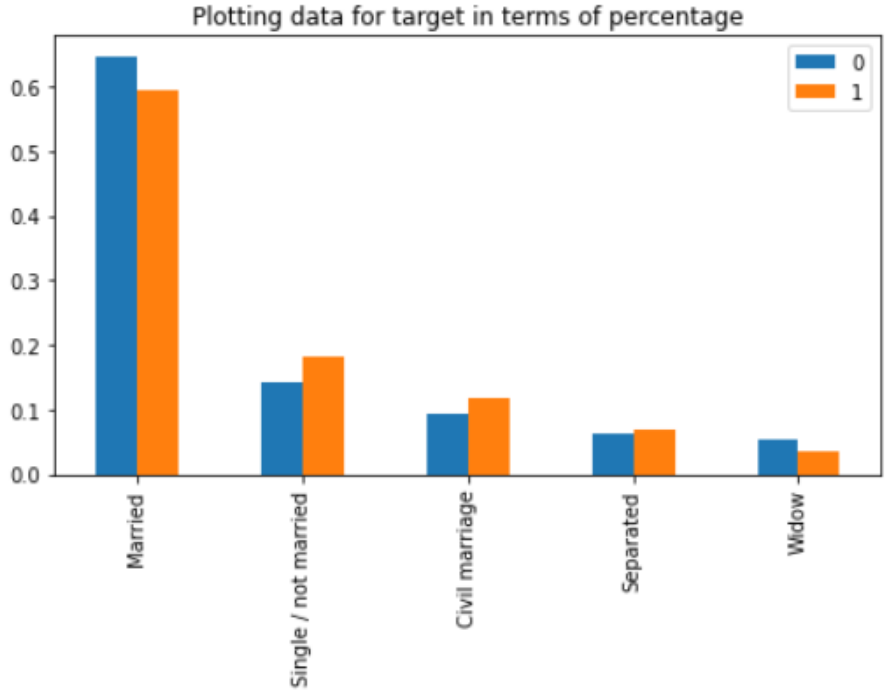


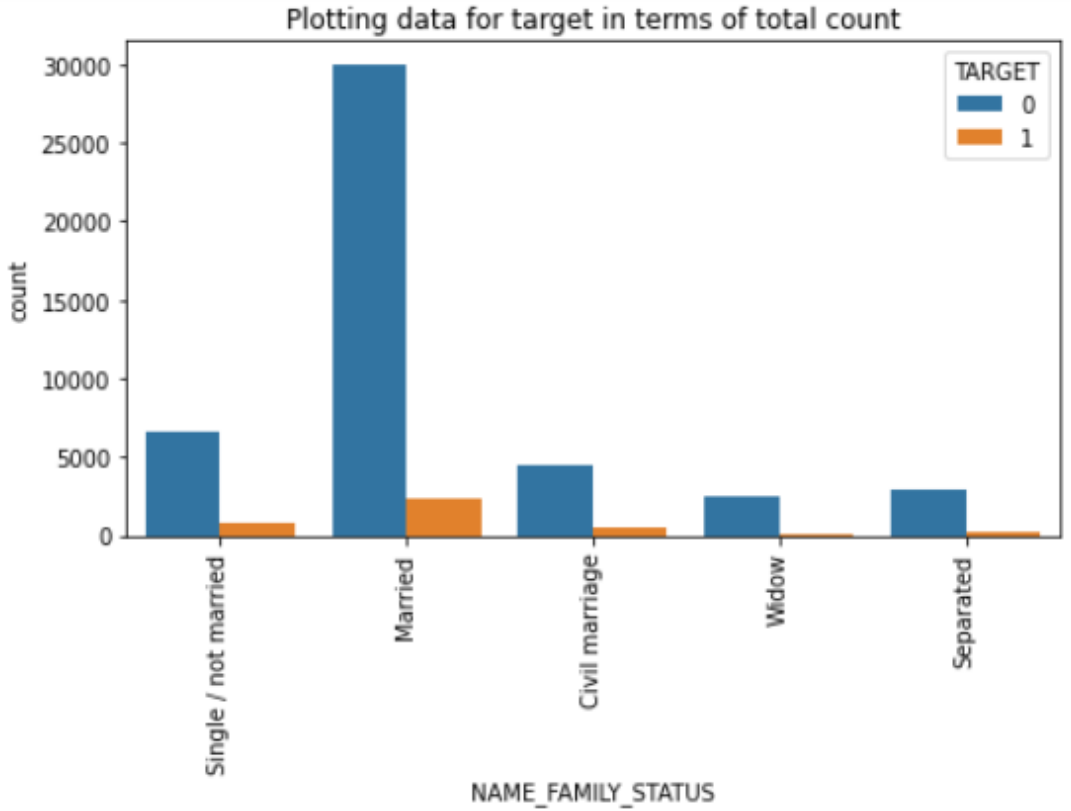




NAME_EDUCATION_TYPE:
Most client take loan for secondary education followed by higher education. But the default rate in secondary education is much high and for higher education is much low.





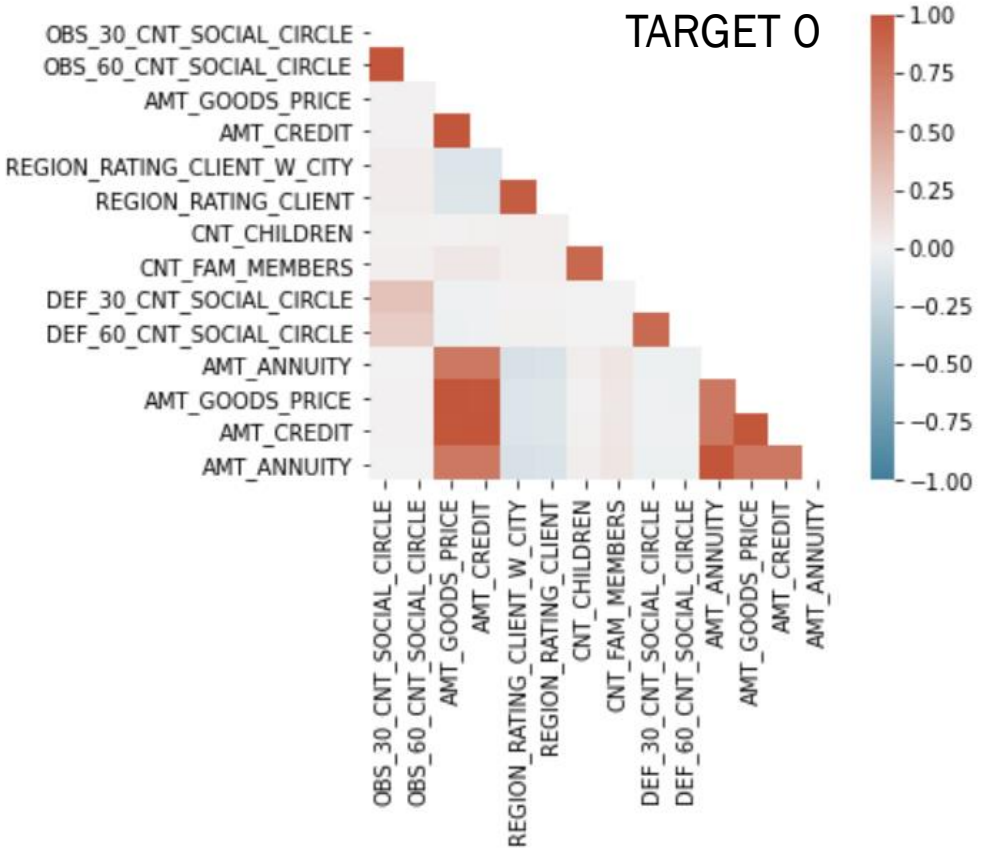


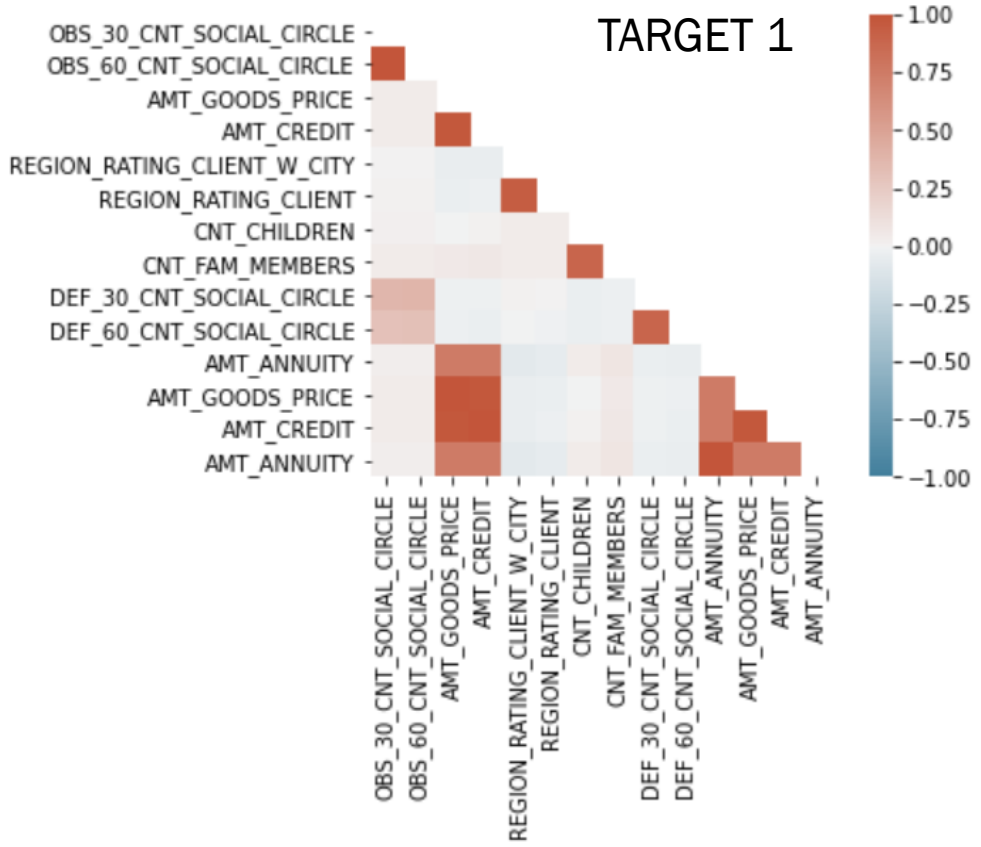
NAME_FAMILY_STATUS:
Most married people apply for loan, and mostly they are not defaulters. Single and civil marriage turns out to be more defaulter.

- CORRELATION

– The top few correlation variables both for target 0 and target 1 are:

OBS_30_CNT_SOCIAL_CIRCLE - OBS_60_CNT_SOCIAL_CIRCLE , AMT_GOODS_PRICE - AMT_CREDIT , REGION_RATING_CLIENT_W_CITY - REGION_RATING_CLIENT , CNT_CHILDREN - CNT_FAM_MEMBERS , DEF_30_CNT_SOCIAL_CIRCLE - DEF_60_CNT_SOCIAL_CIRCLE , AMT_ANNUITY - AMT_GOODS_PRICE , AMT_CREDIT - AMT_ANNUITY

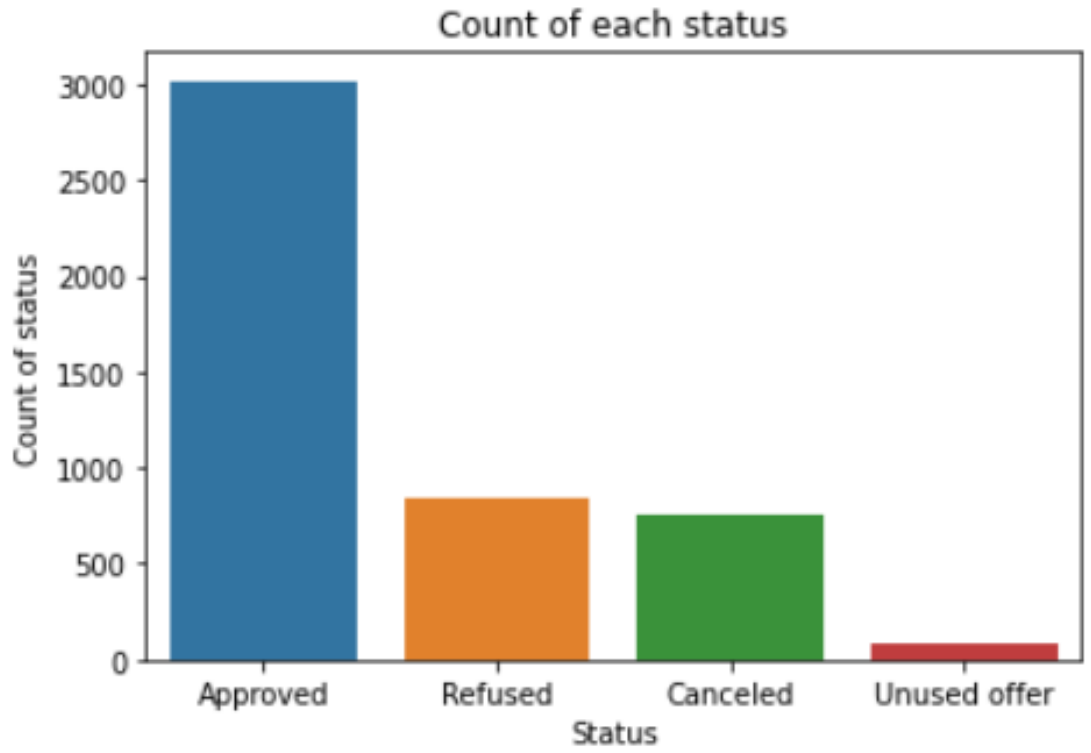




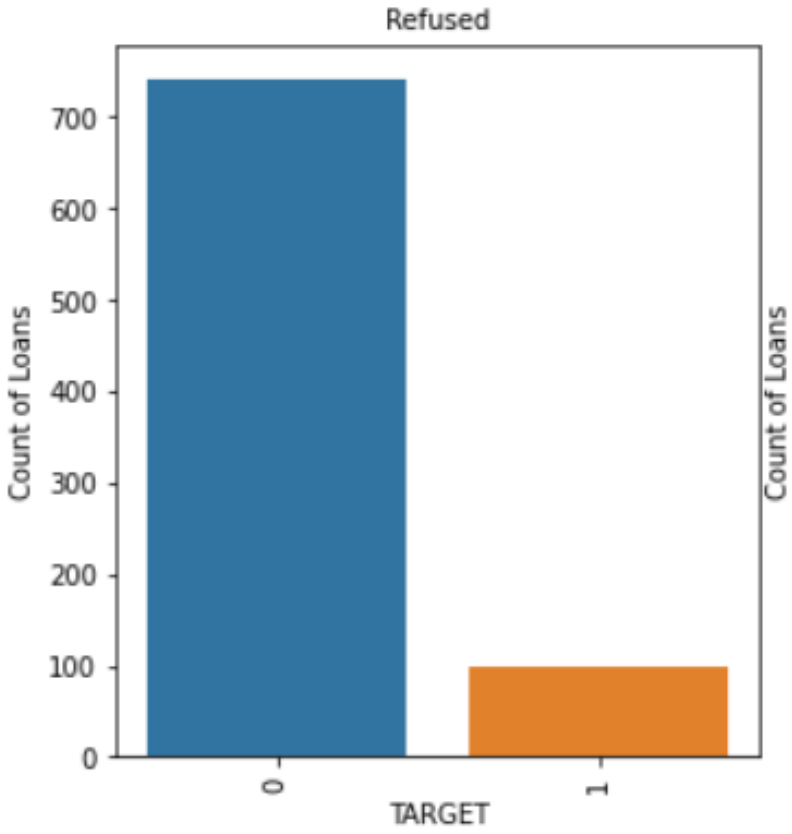
- INSPECTING THE DATAFRAME (Previous)

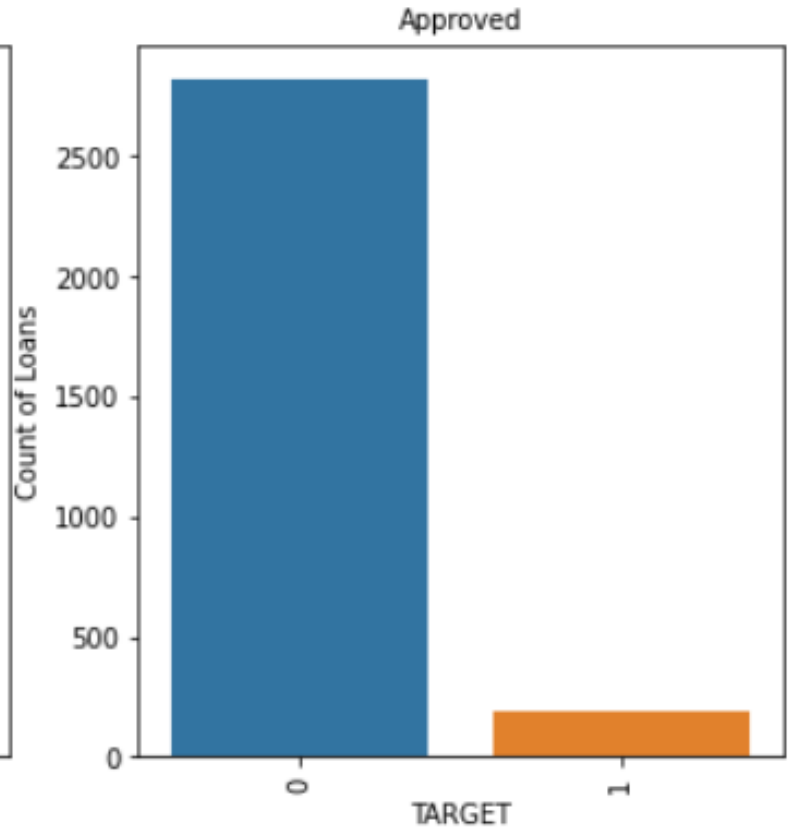
- DATA CLEANING

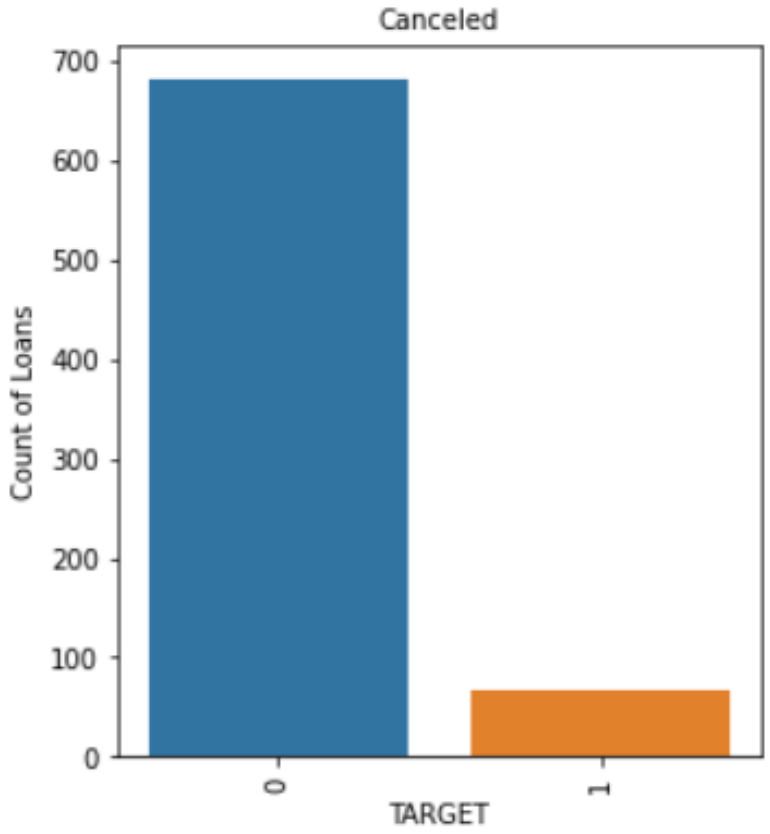
  - Removing columns and rows where NA values are >=20%
  - Merging table Application and Previous on “SK_ID_CURR” using left join
  - Splitting merged table on “NAME_CONTRACT_STATUS” each category



NUMBER OF LOANS PER CATEGORY CONTRACT STATUS OF TARGET 0 & 1





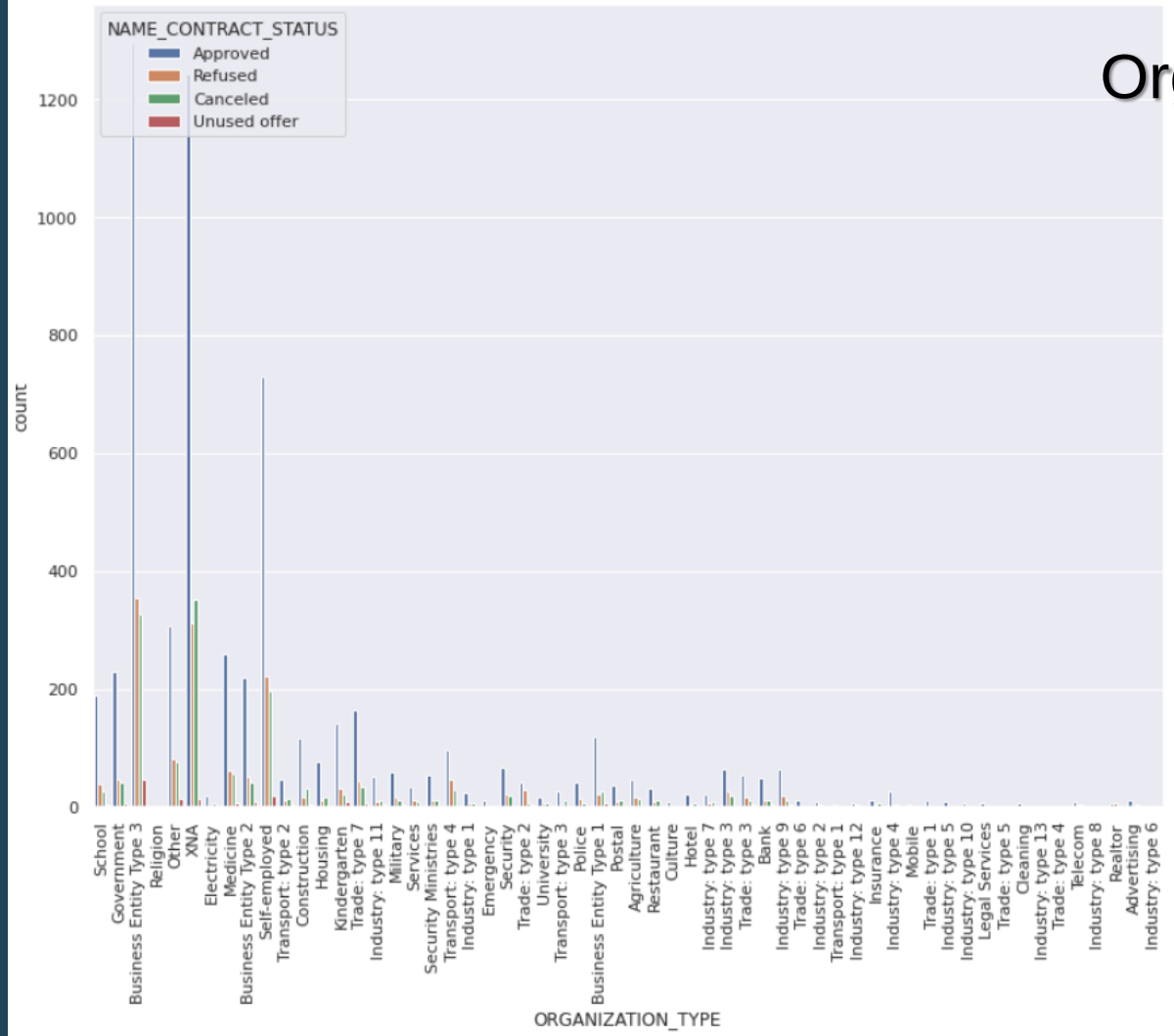


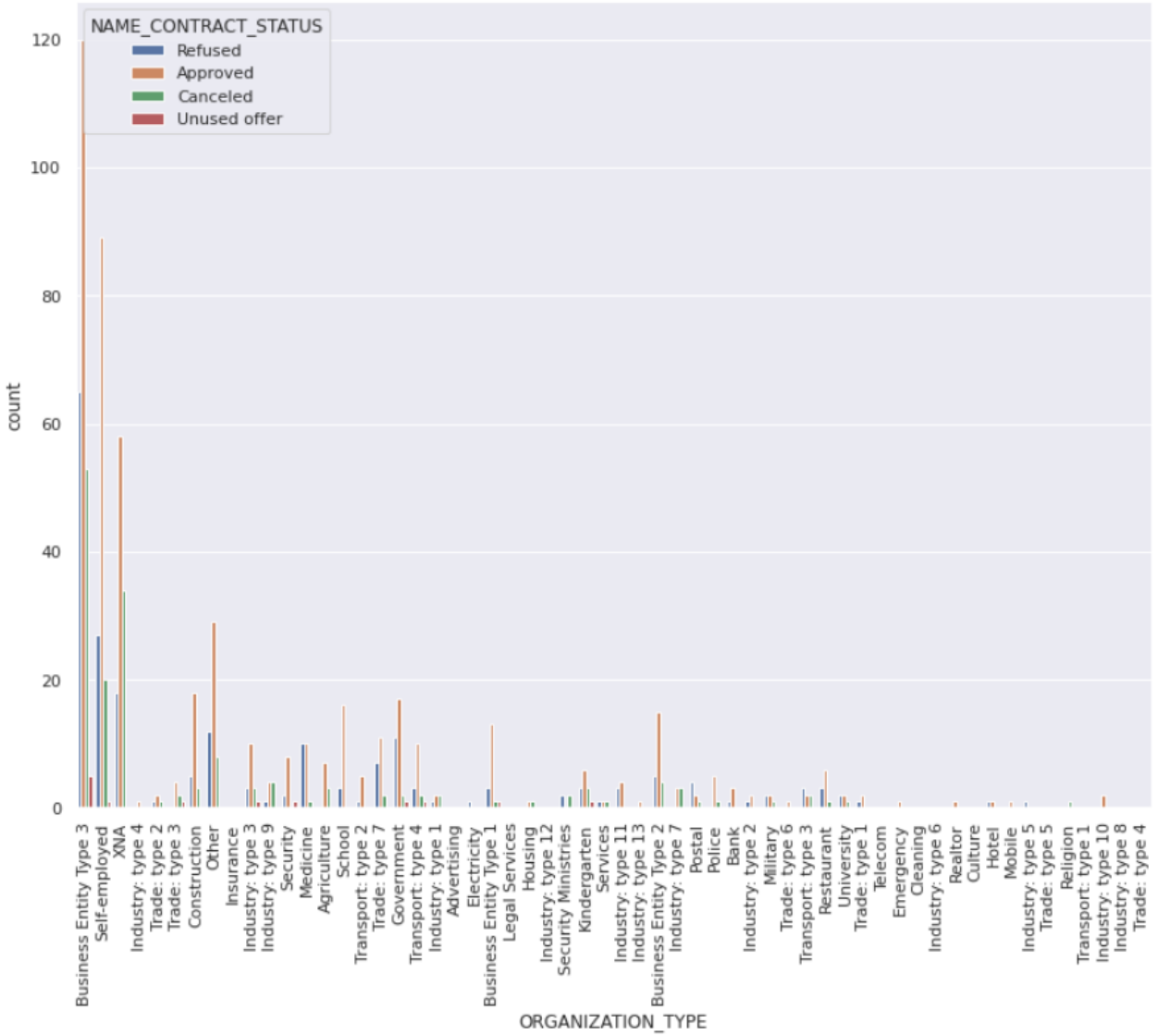
Approved Status (%)
T0 – 93.5698
T1 - 06.4302

Refused Status (%)
T0 – 88.2283
T1 – 11.7717

Cancelled Status (%)
T0 – 90.9333
T1 – 09.0667

Organization_Type VS Name_Contract_Status





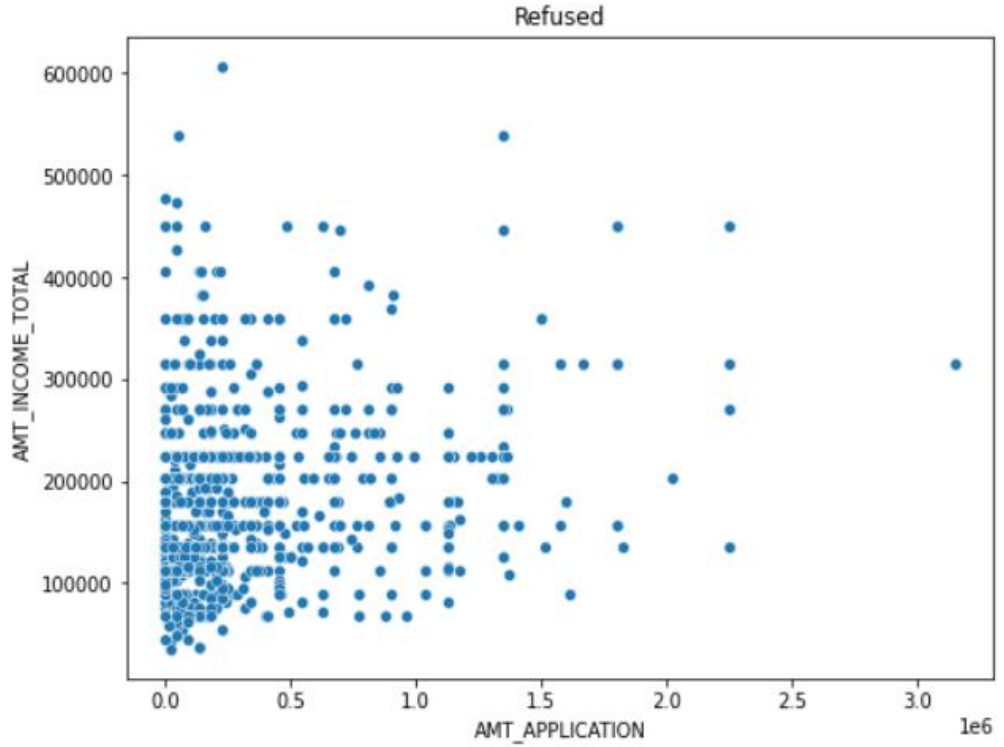
Business Entity Type 3 and Self-employed tends to be the maximum defaulter. The univariate analysis of this dataset was more fruitful than bivariate analysis.

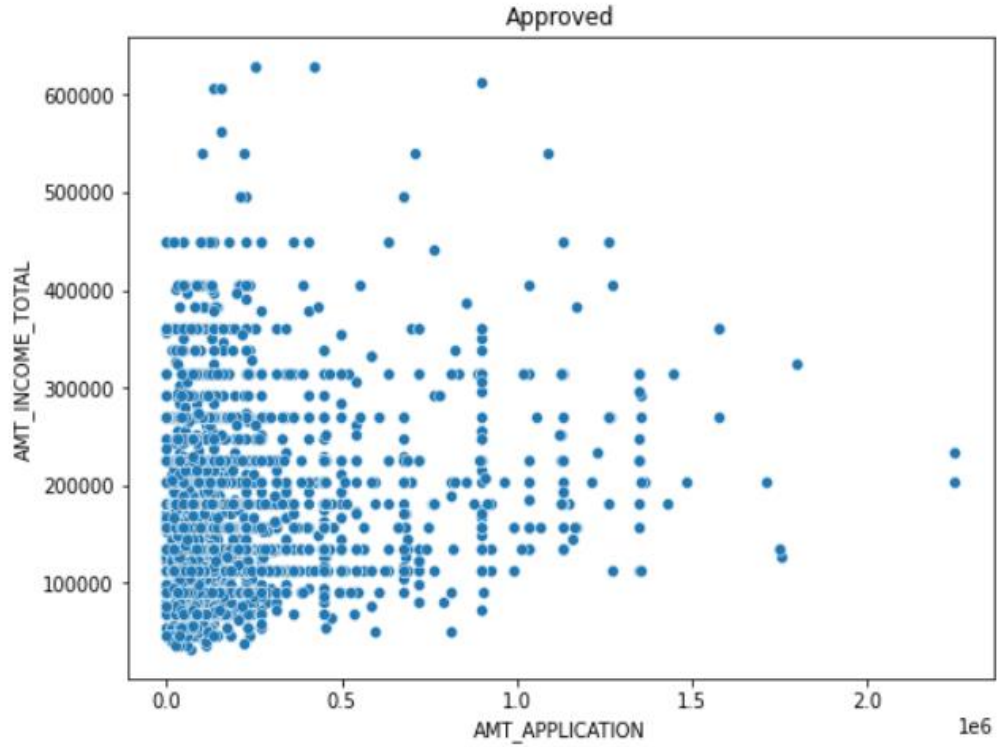
- AMT_INCOME_TOTAL quantile values for merged table:

QUANTILE	VALUE
0.100	81000.0
0.200	99000.0
0.300	112500.0
0.400	135000.0
0.500	148500.0
0.600	166500.0
0.700	189000.0
0.800	225000.0
0.900	270000.0
0.950	337500.0
0.990	472500.0
0.995	630000.0
0.998	765000.0
1.000	117000000.0

Thus we can remove more than 99.8 percentile data as the difference is very high.

Effect of AMT_APPLICATION & AMT_INCOME_TOTAL on LOAN STATUS





AMT_APPLICATION higher than 200k had a higher rejection rate. Also loan rejection rate was much lower if the income was higher than 550k

Thank You