PRESENTATION ON CREDIT EDA

By-Saumrit Saurav Parida

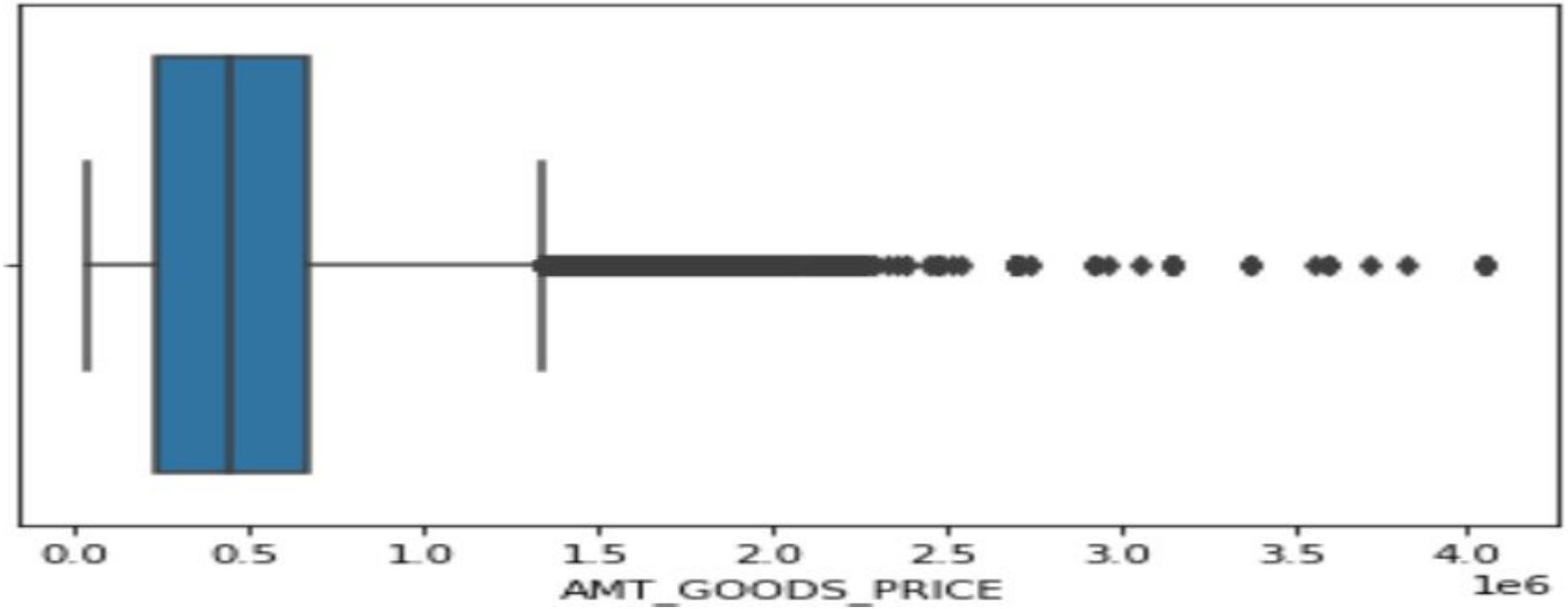
OBJECTIVES

- This analysis is to credit risk analysis to help company to make a decision on approving loan to the right applicant based on applicant's profile which means to look at the outcome of default and non-default applicants.

DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

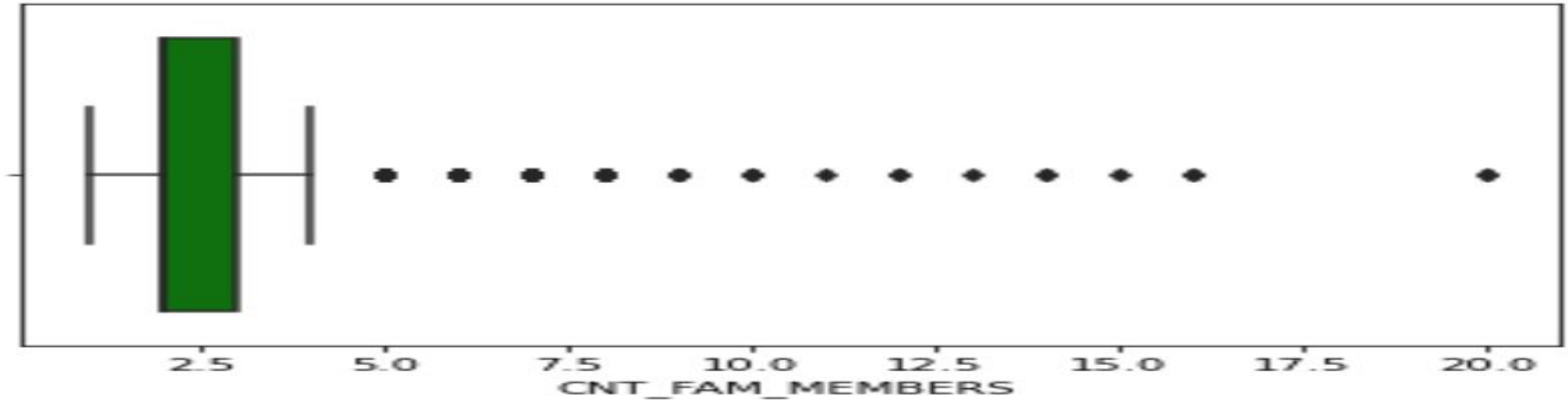
- AMT_ANUITY(Loan annuity) was having outlier null value or unexpelow diagram, those cted value that were exceeding to 250000 amount as shown in bevalues has been imputed with median value i.e., 24903.

➔ **AMT_GOODS_PRICE For consumer Loans it is the price of the goods for which the loan is given was having outlier null value that were exceeding to 450000 as shown diagram, those values has been imputed with median value.e450000**

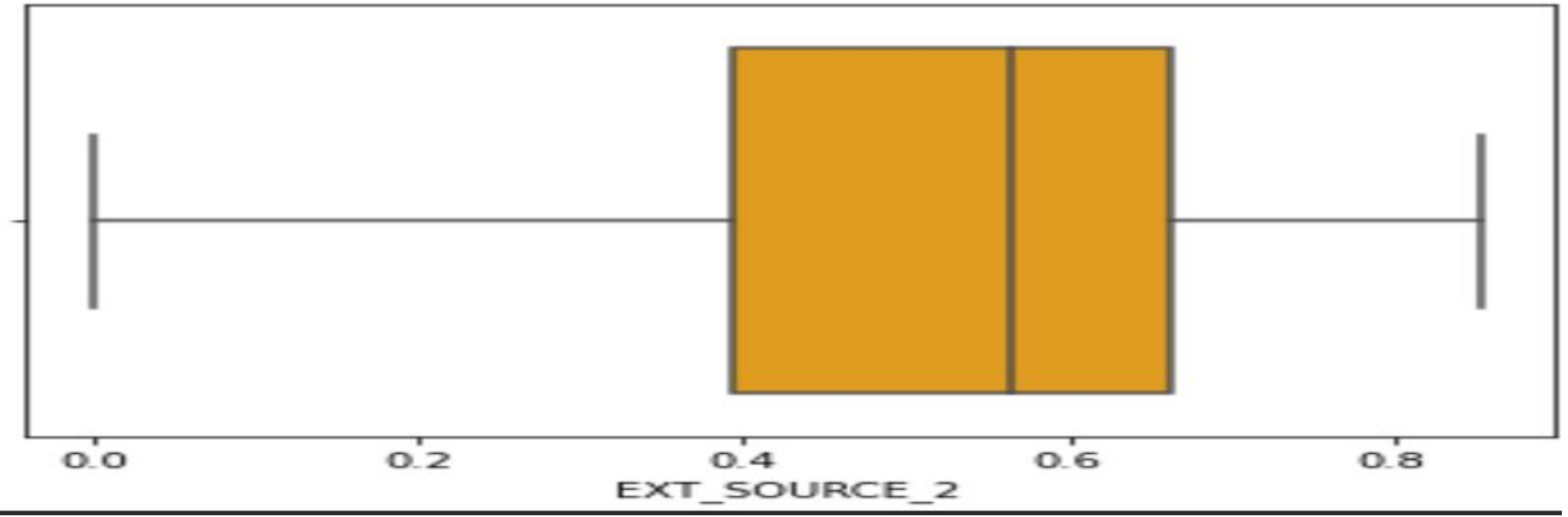


DATA ANALYSIS ON CONTINOUS VARIABLE OF APPLICATION DATA

➔ **CNT_FAM_MEMBERS Count of family members client have was having outlier null value that were exceeding to 20 as shown in below diagram, those values has been imputed with median value i.e..,2.**

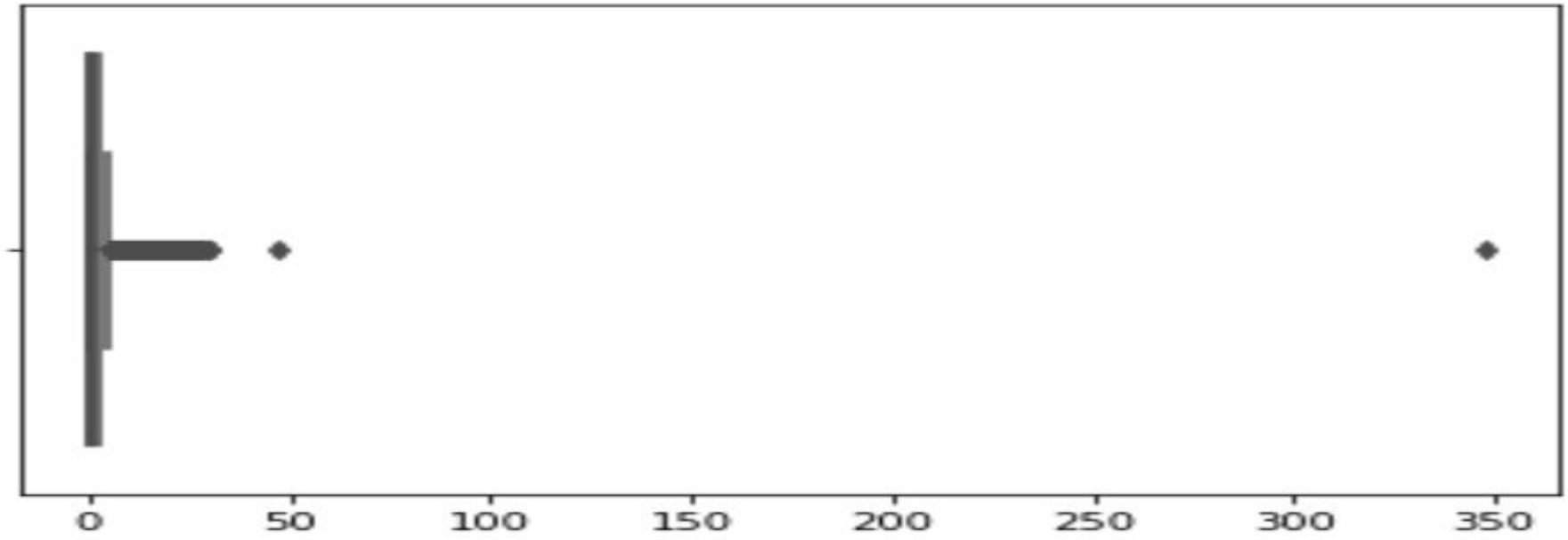


➔ **EXT_SOURCE_2 (Normalized score from external data source) was not having any outlier null value as shown in below diagram. Hence, those values has been imputed with mean value i.e., 1**

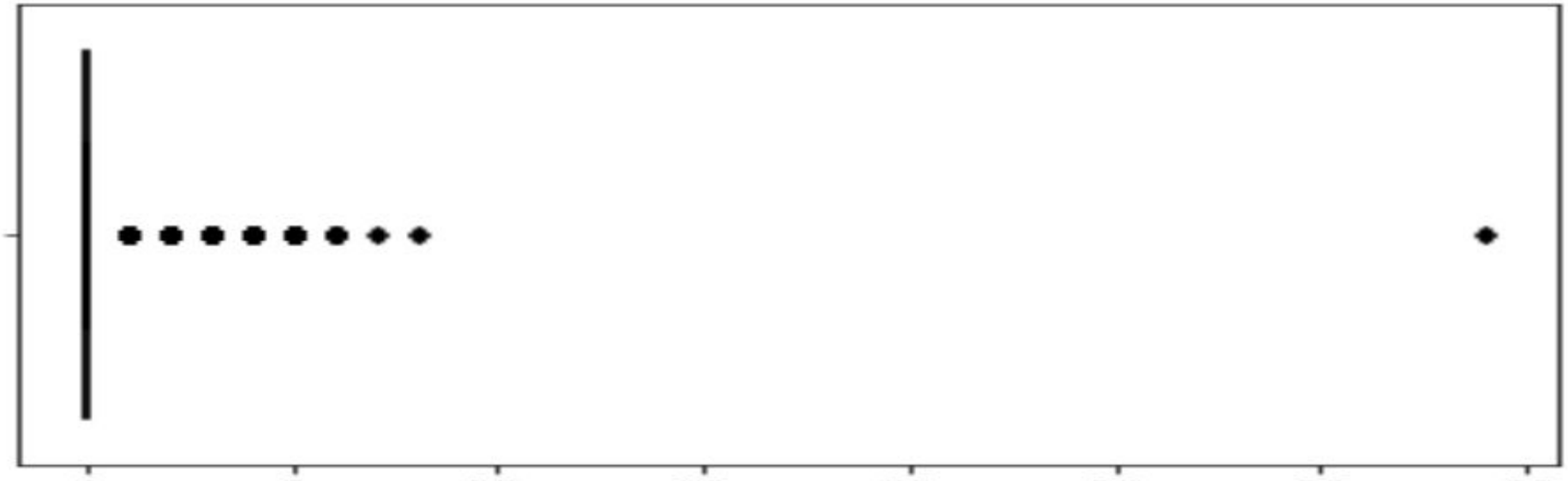


DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

➔ OBS_30_CNT_SOCIAL_CIRCLE Count of observation of client's social surroundings with observable 30 days past due default was having outlier null value that were 350 approx. as shown in below diagram. Those values has been imputed with median

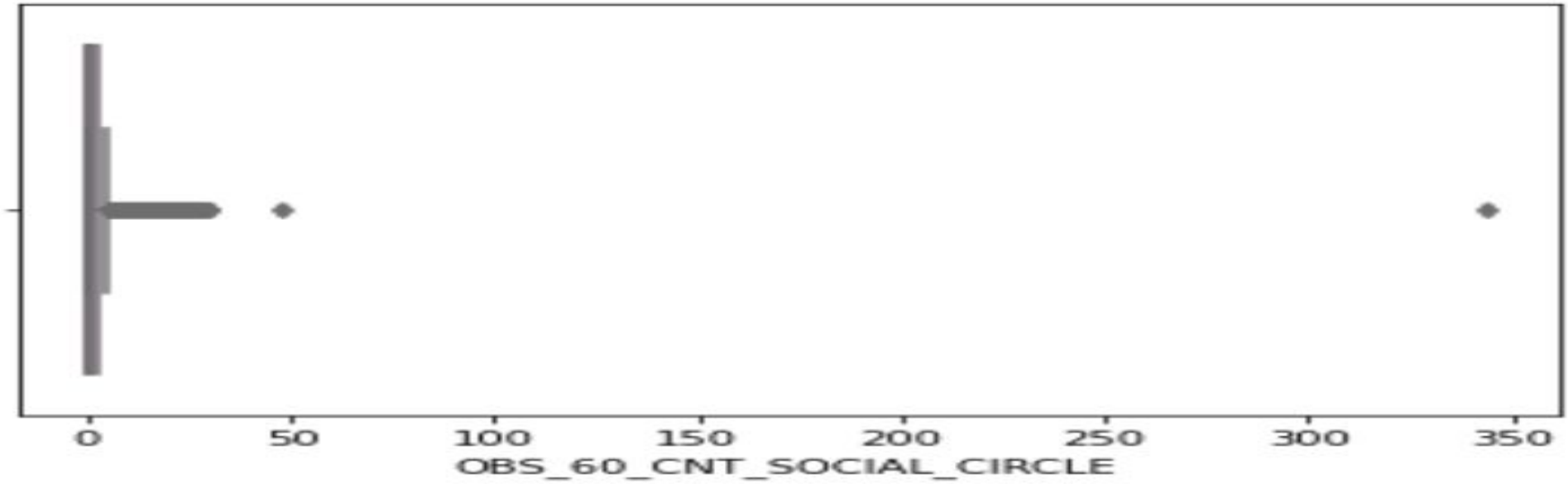


➔ DEF_30_CNT_SOCIAL_CIRCLE Count of observation of client's social surroundings defaulted on 30 DPD days past due was having outlier null value that were 34 approx, as shown in below diagram. Those values has been imputed with median value i.e., 0.

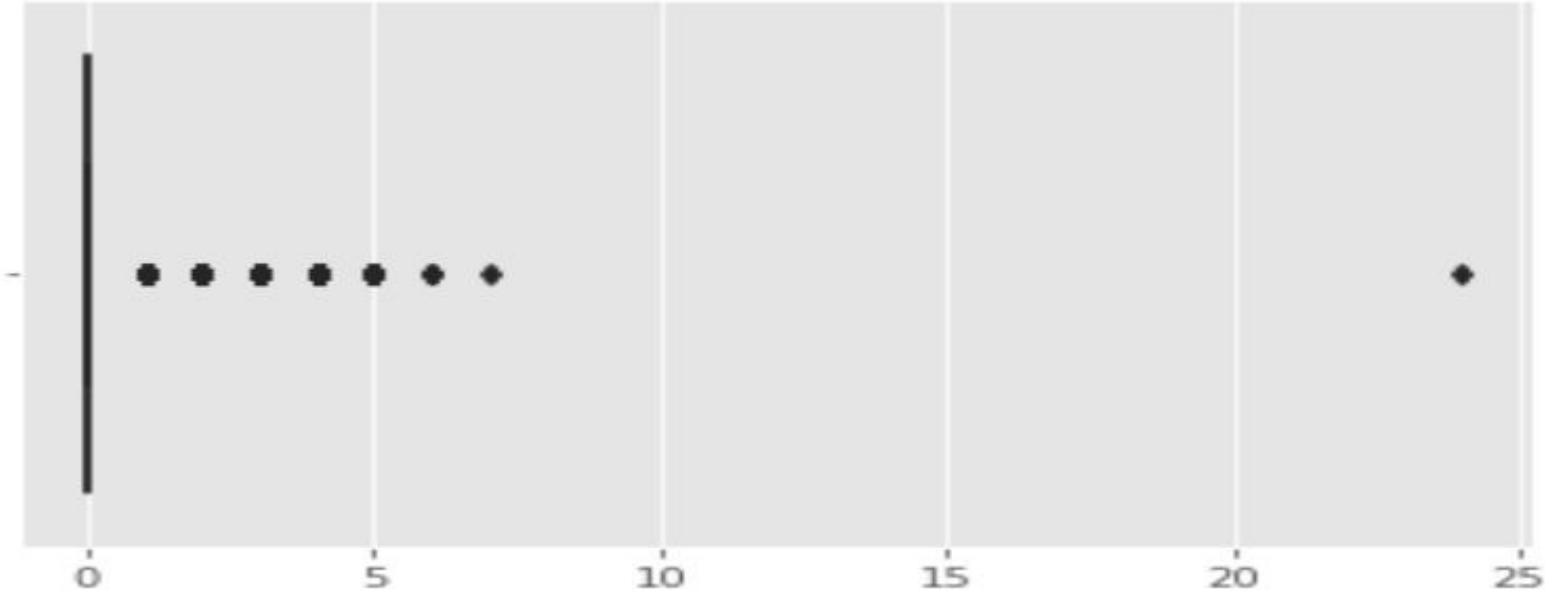


DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

➔ **OBS_60_CNT_SOCIAL_CIRCLE Count of observation of client's social surroundings with observable 60 days past due default was having outlier null value that were 350 approx, as shown in below diagram. Those values has been imputed with median**

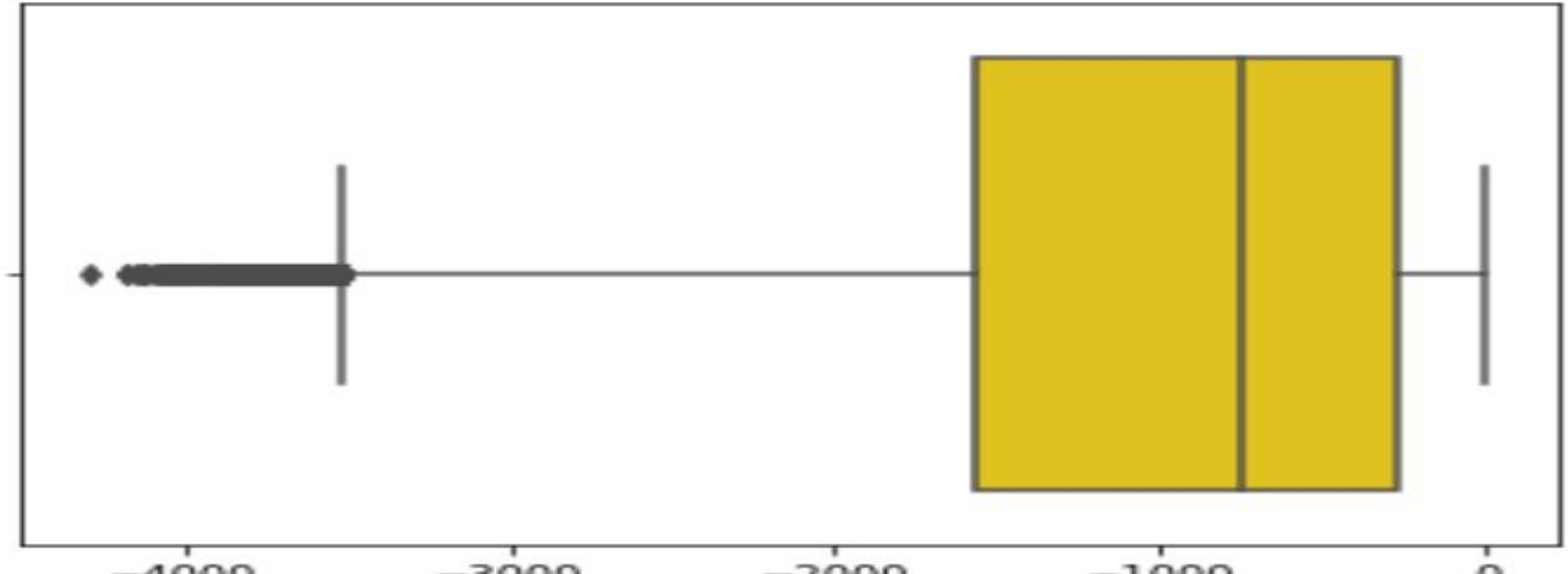


➔ DEF_60_CNT_SOCIAL_CIRCLE (Count of observation of client's social surroundings defaulted on 60 days past due) was having outlier null value that were 24 approx, as shown in below diagram. Those values has been imputed with median value i.e., 0



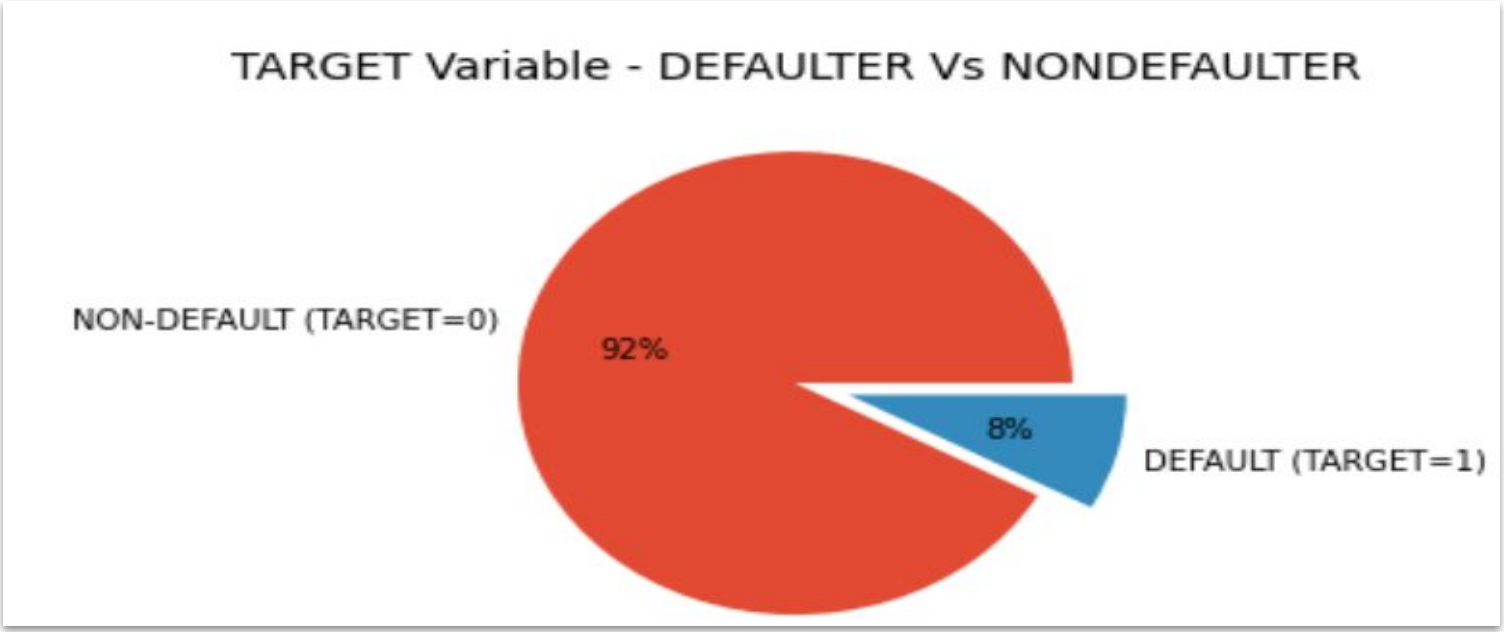
DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

➔ DAYS_LAST_PHONE_CHANGE Count of days before application did client change phone was having outlier null value that were 24 approx, as shown in below diagram. Those values has been imputed with median value i.e., -757



DATA ANALYSIS ON CATAGORICAL VARIABLES OF APPLICATION DATA

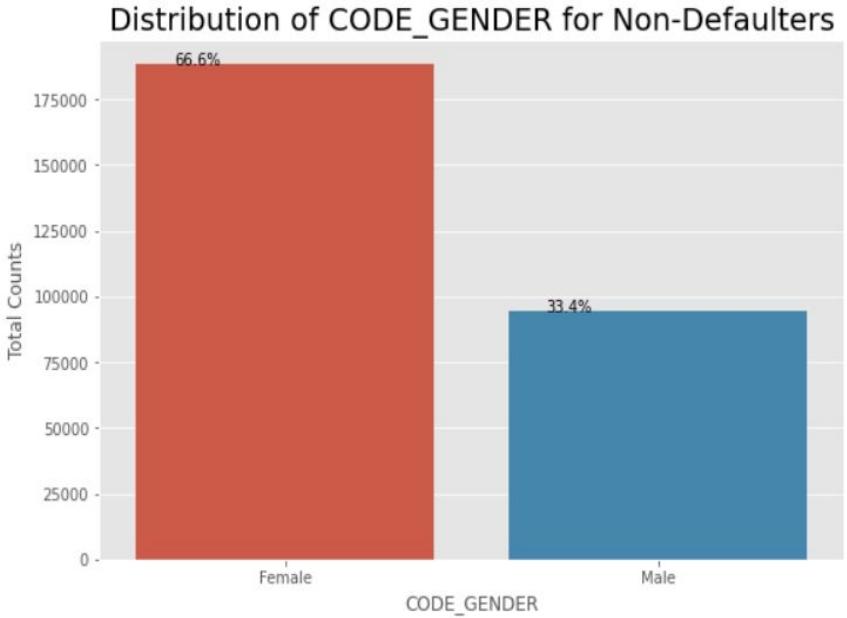
- ➔ NAME_TYPE_SUITE (Who was accompanying client when he was applying for the loan) can be imputed with value with mode of the column which is Unaccompanied

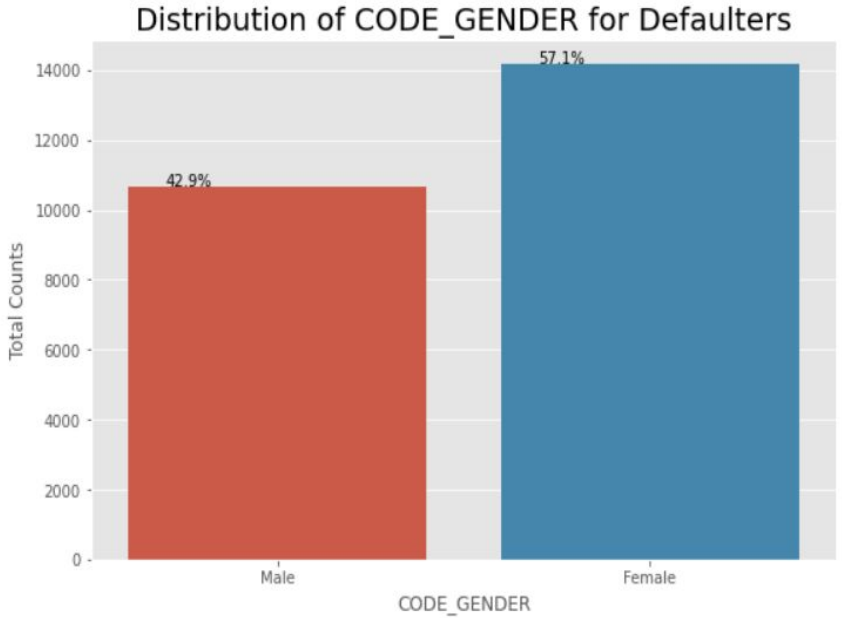


➔ Using this pie plot, we have a clear understanding that there are huge imbalance between defaulters and non-defaulters. This plot states that 92% of applicants are non-defaulters whereas 8% of applicants are defaulter

UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

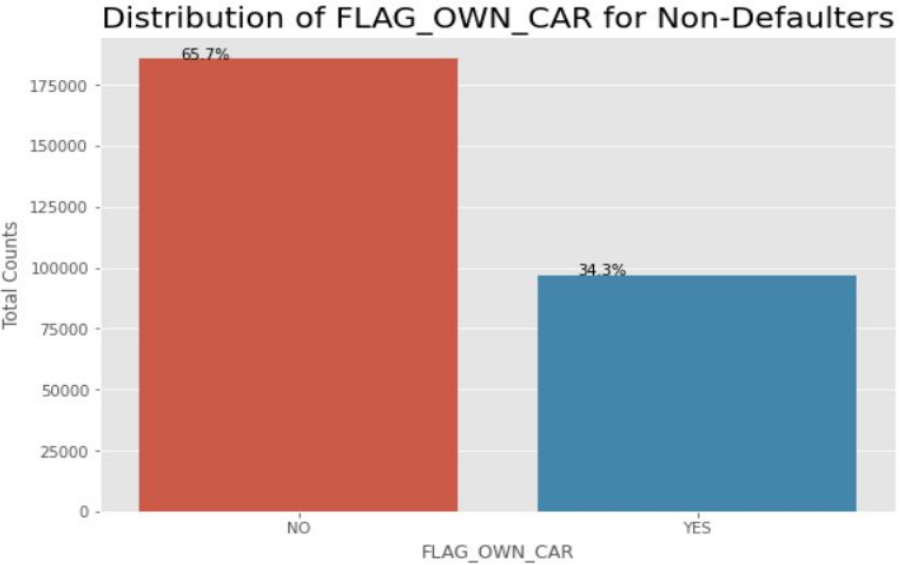
CODE _GENDER

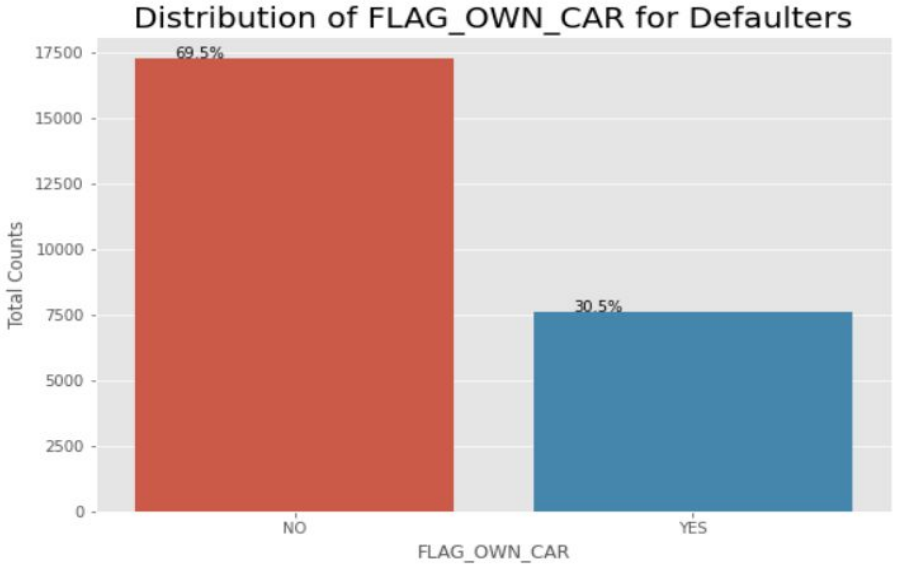




Through this plot diagram we can understand that the percentage of

FLAG_OWN_CAR

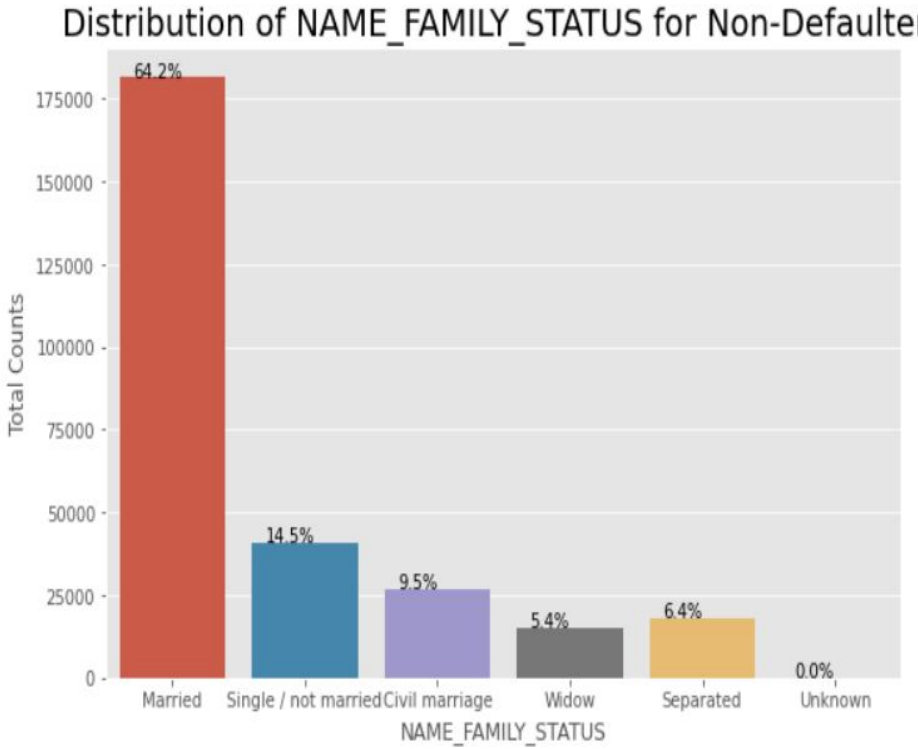


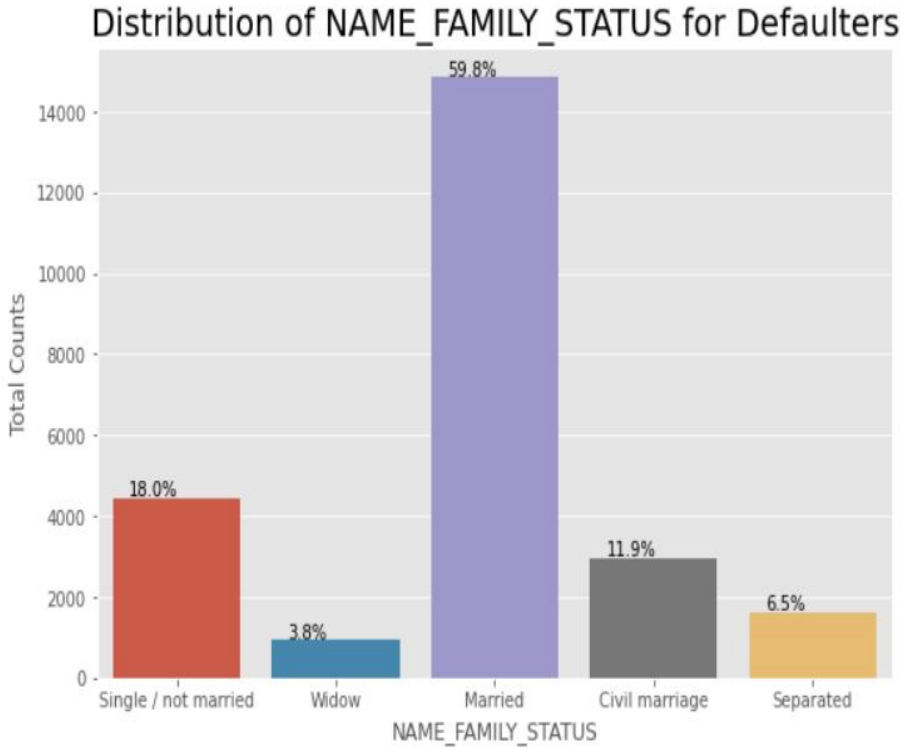


➔ We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters. While people who have car default more often, the reason could be there are simply more people without cars. Looking at the

UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

NAME_FAMILIY_STATUS

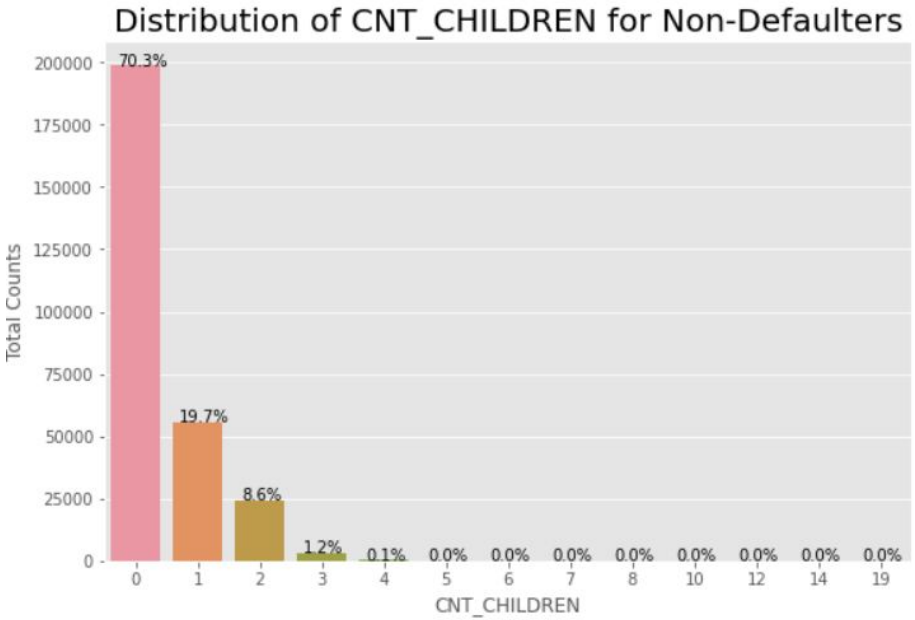


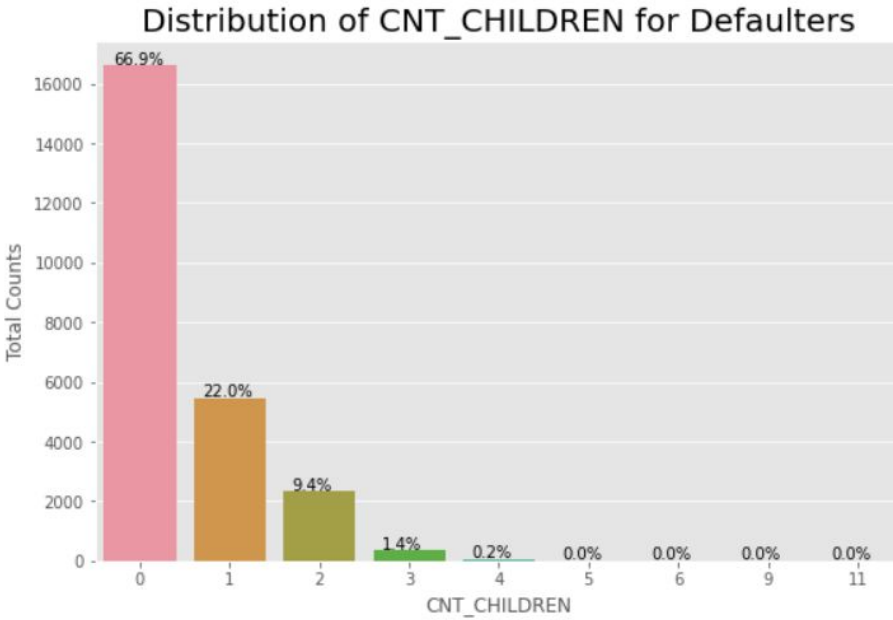


➔ It is clear from the graph that people who have House/Appartment,

UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

CNT_CHILDREN

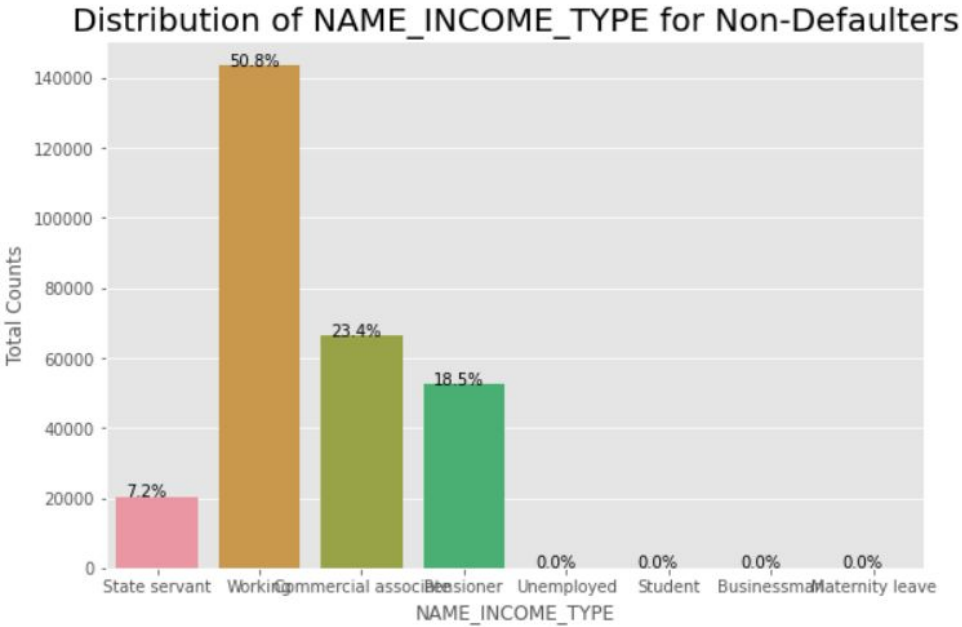


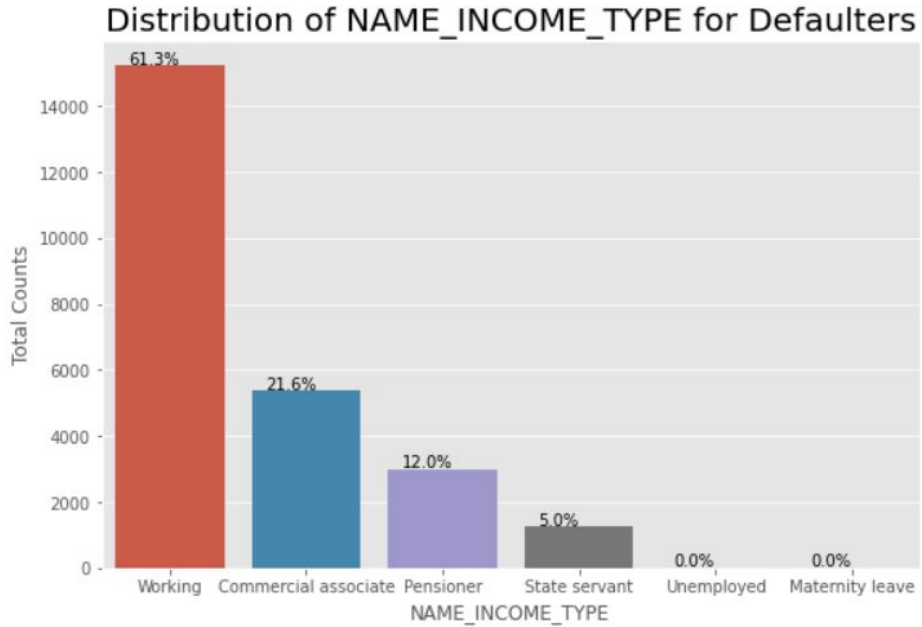


➔ Approx 70% of applicants who have children are counted and non-Defaulters and 66% applicant who are having children are defaulters

UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

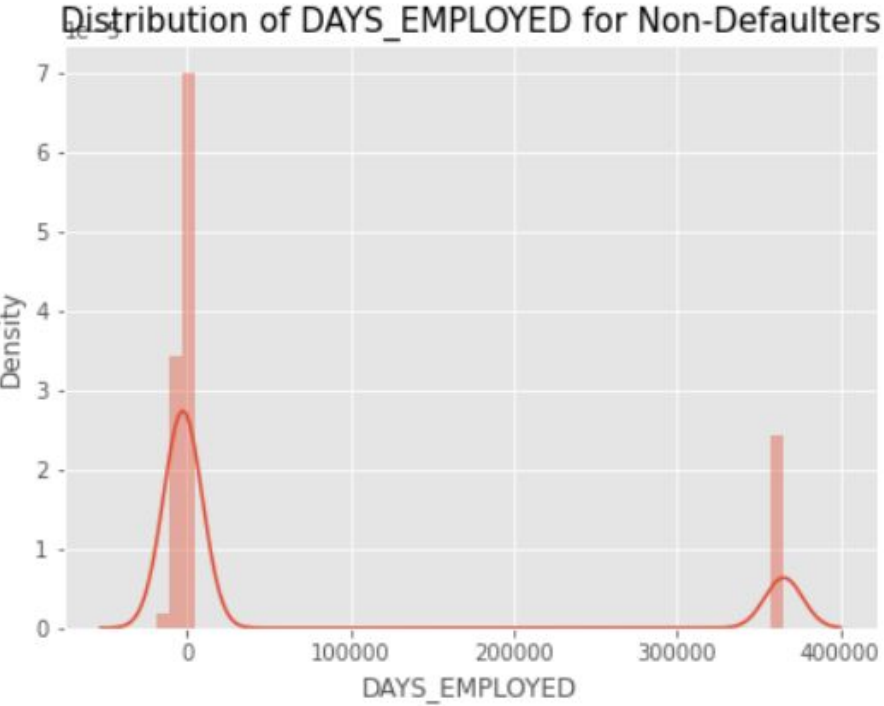
NAME_INCOME_TYPE

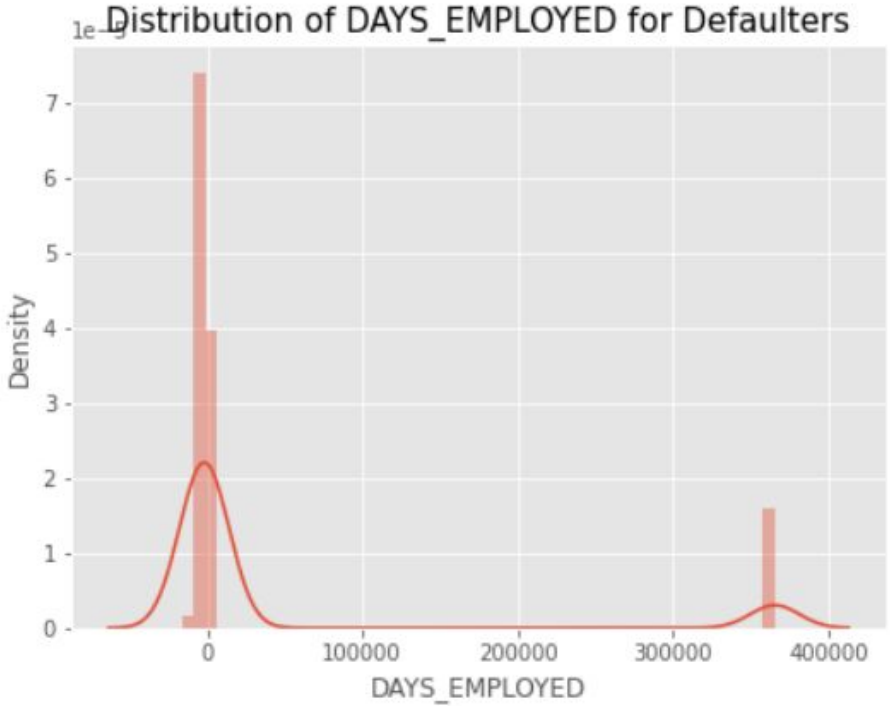




UNIVARIATE ANALYSIS ON CONTINUOUS VALUE

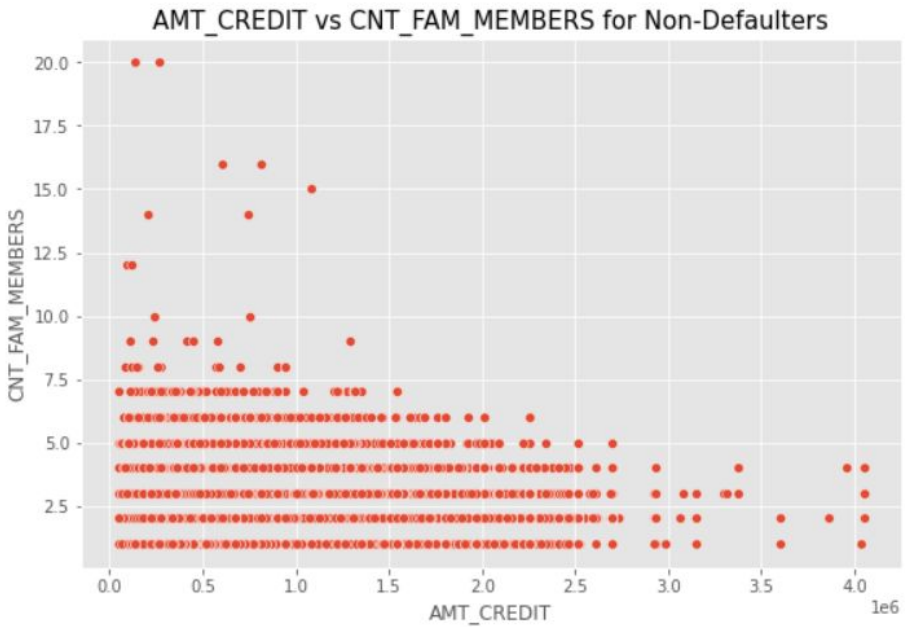
DAYS_EMPLOYED

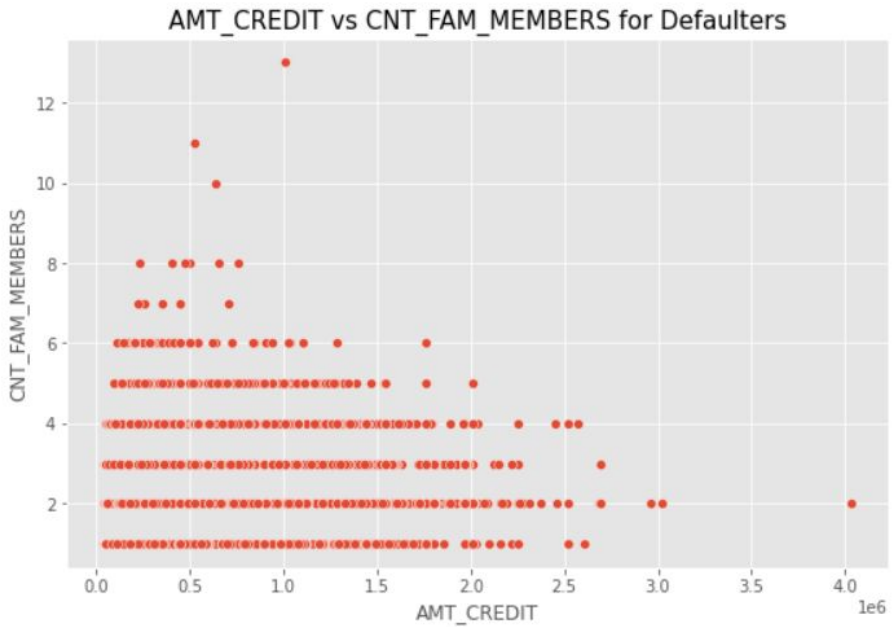




UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

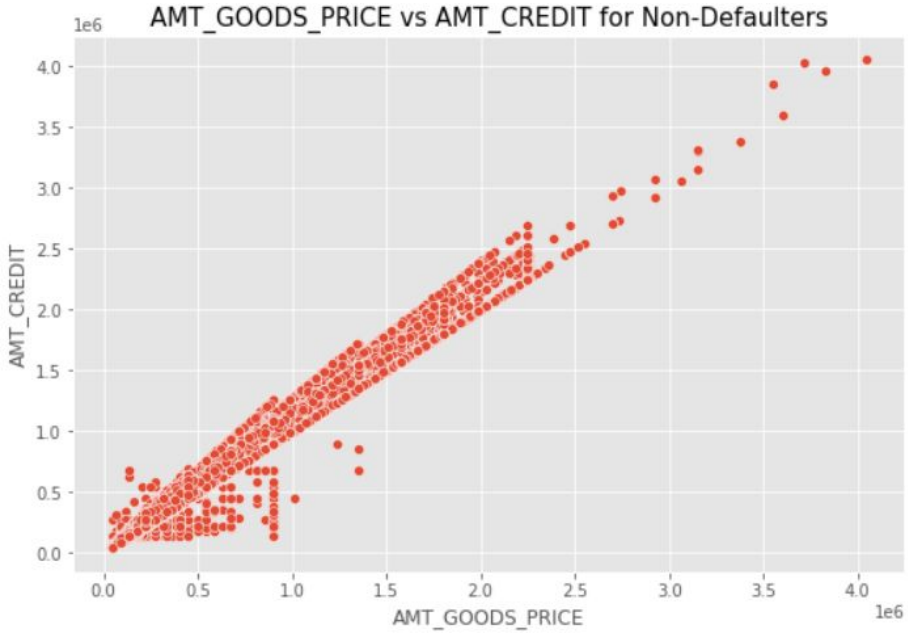
AMT_CREDIT and CNT_FAM_MEMBERS

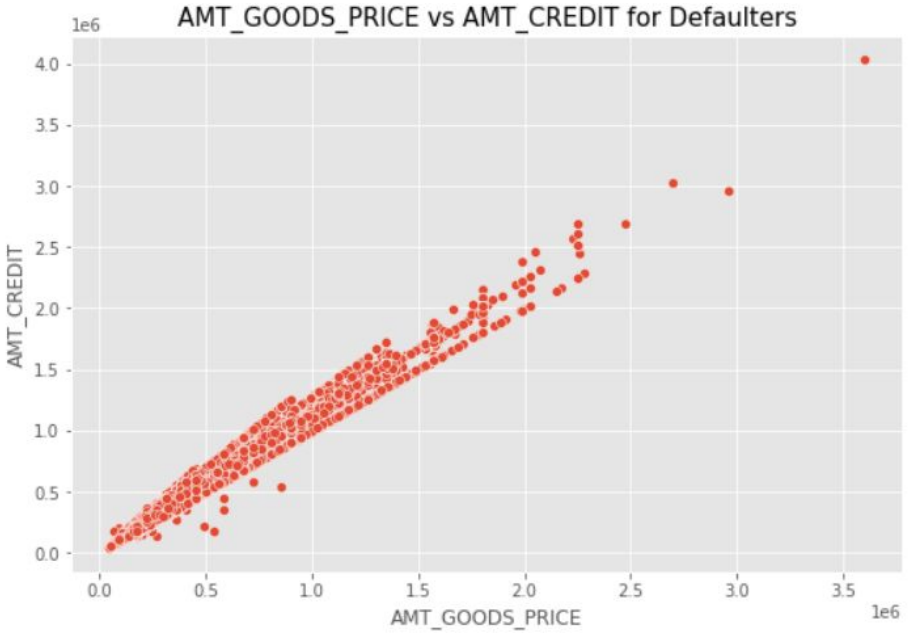




UNIVARIATE ANALYSIS ON CATEGORICAL VALUE

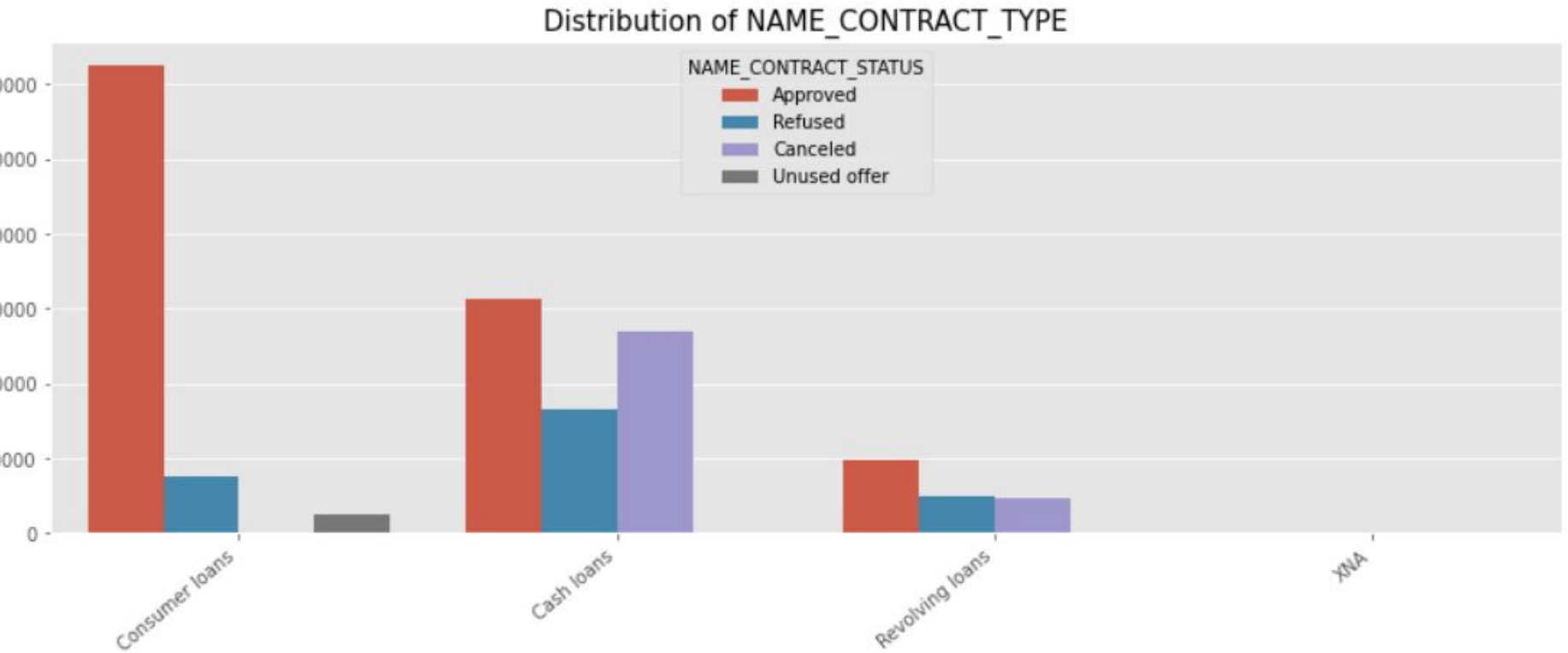
AMT_GOODS_PRICE and AMT_CREDIT



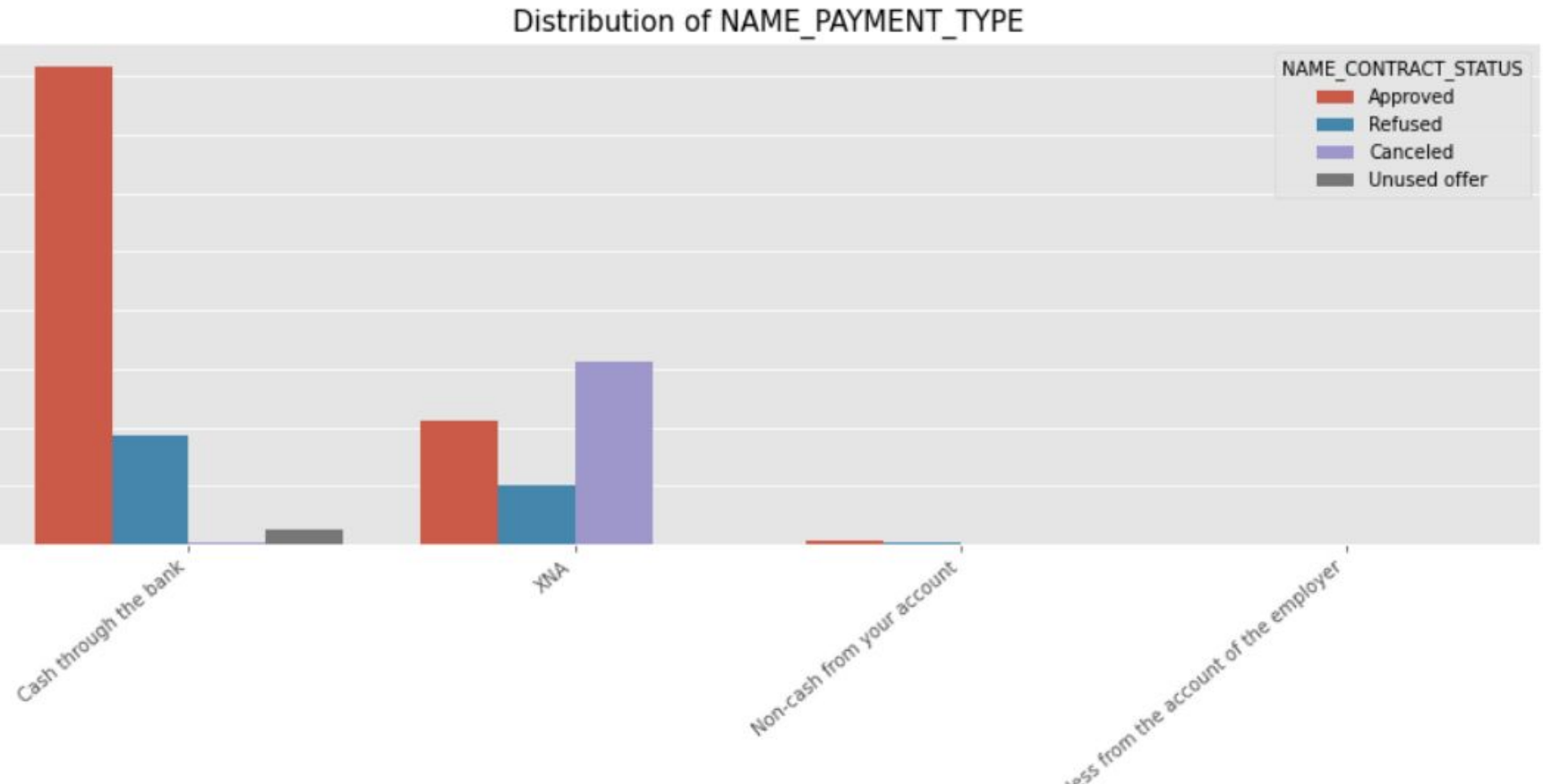


DATA ANALYSIS IN PREVIOUS APPLICATION DATA

NAME_CONTRACT_TYPE

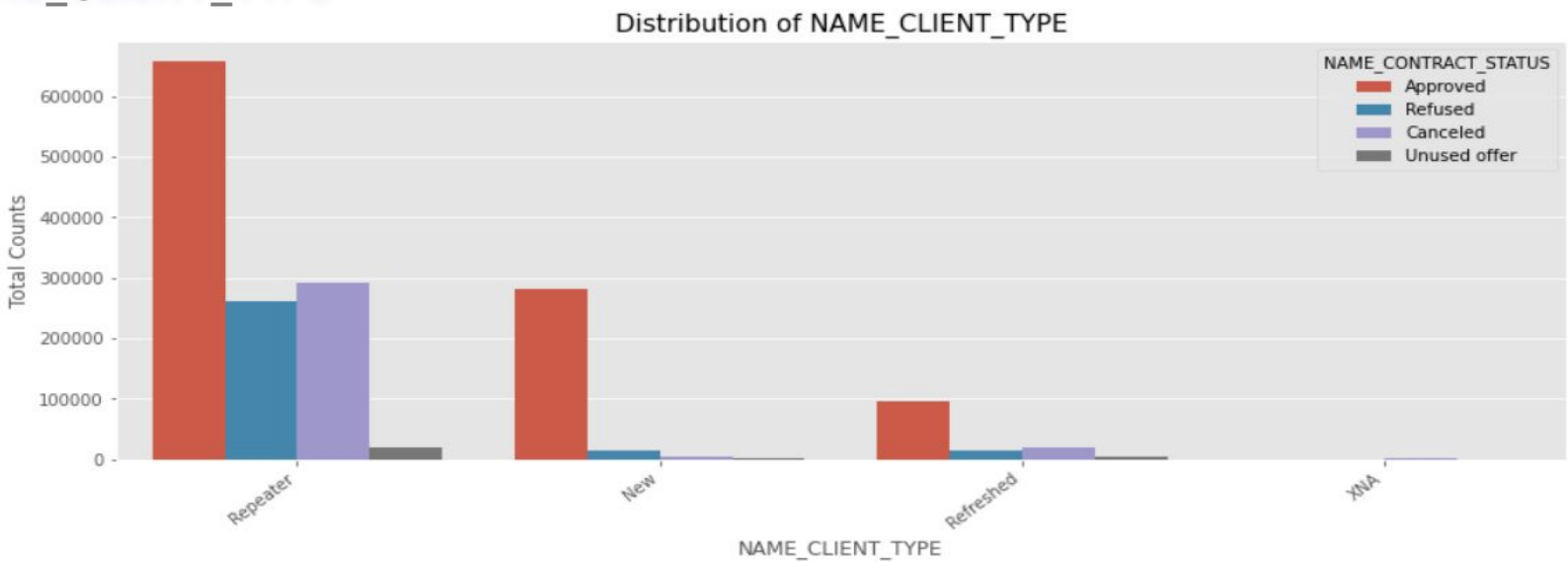


NAME_PAYMENT_TYPE



DATA ANALYSIS IN PREVIOUS APPLICATION DATA

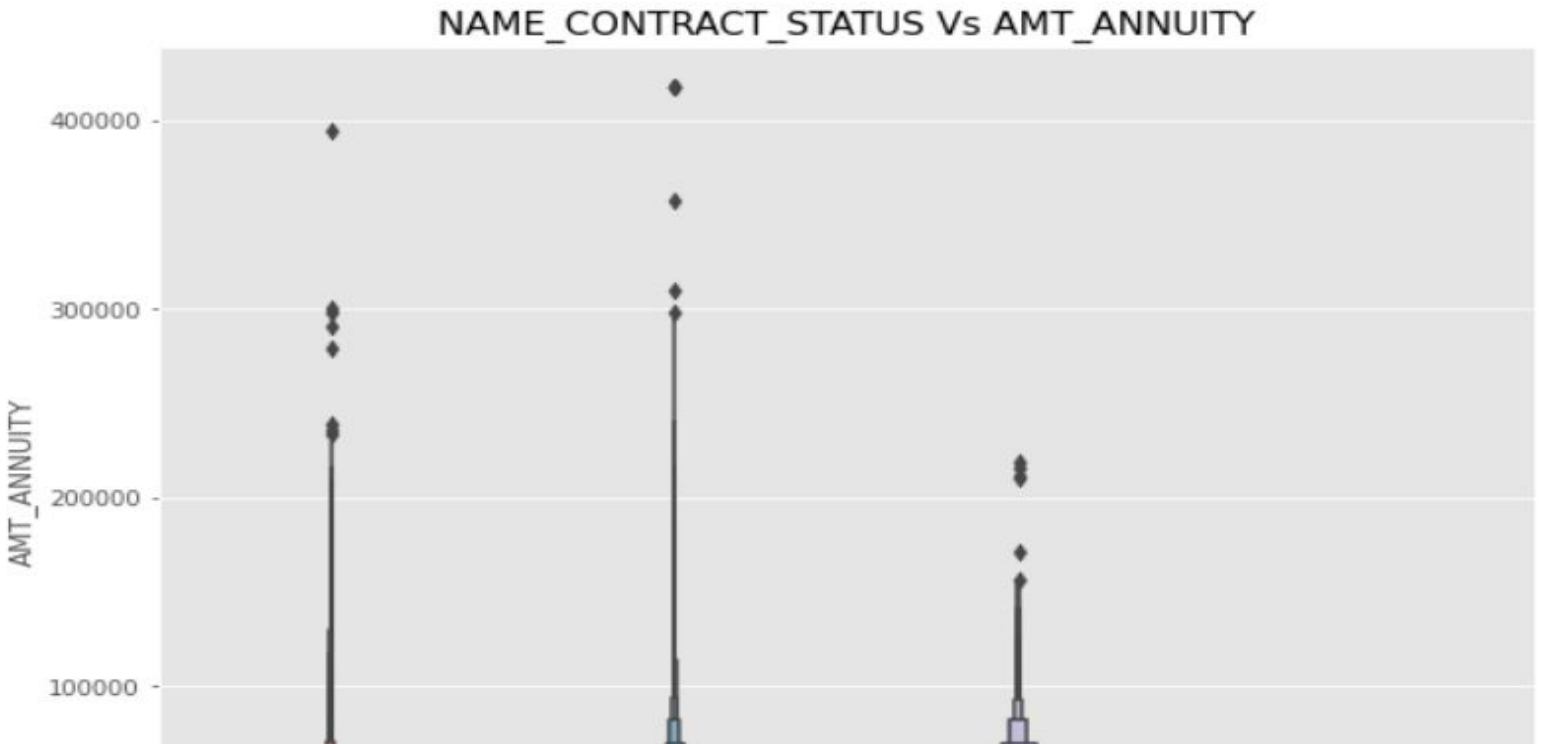
NAME_CLIENT_TYPE



➔ Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get

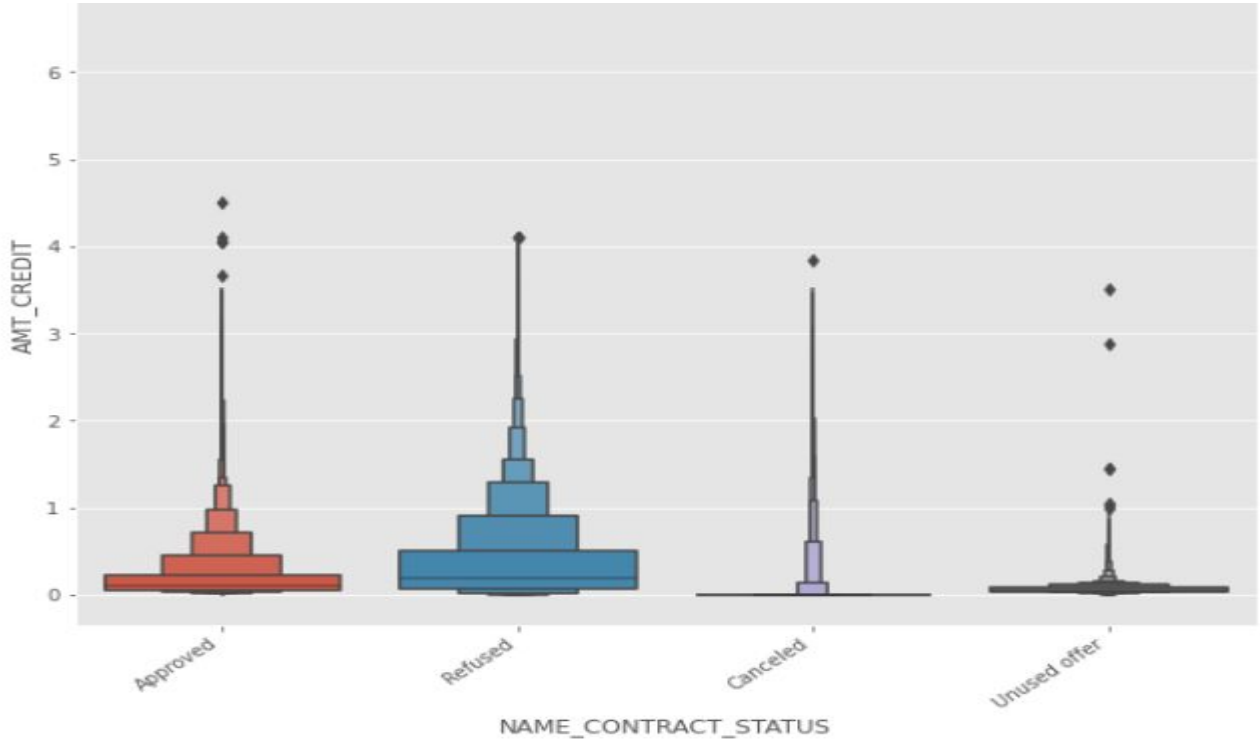
USING BOX PLOT TO DO BIVARIATE ANALYSIS ON CATAGORICAL AND NUMERIC COLUMNS

NAME_CONTRACT_STATUS and AMT_ANNUITY



USING BOX PLOT TO DO BIVARIATE ANALYSIS ON CATAGORICAL AND NUMERIC COLUMNS

NAME_CONTRACT_STATUS and AMT_CREDIT



MERGING APPLICATION DATA AND PREVIOUS APPLICATION DATA

FLAG_OWN_CAR and NAME_CONTRACT_STATUS

MERGING APPLICATION DATA AND PREVIOUS APPLICATION DATA

TARGET and NAME_CONTRACT_STATUS