PRESENTATION ON CREDIT EDA

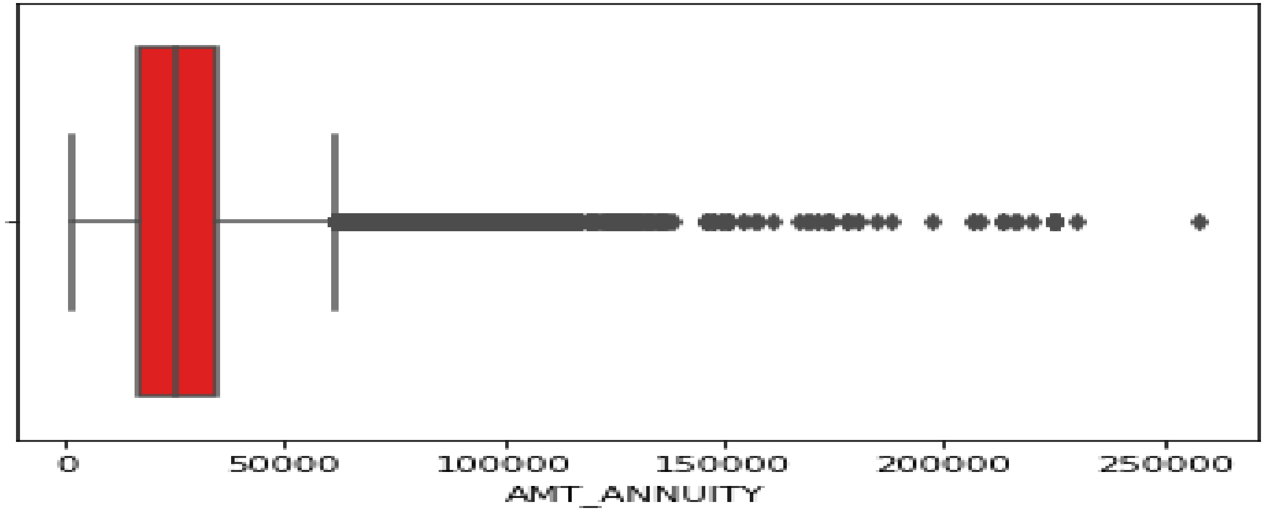
By Tushar Anand

OBJECTIVES:

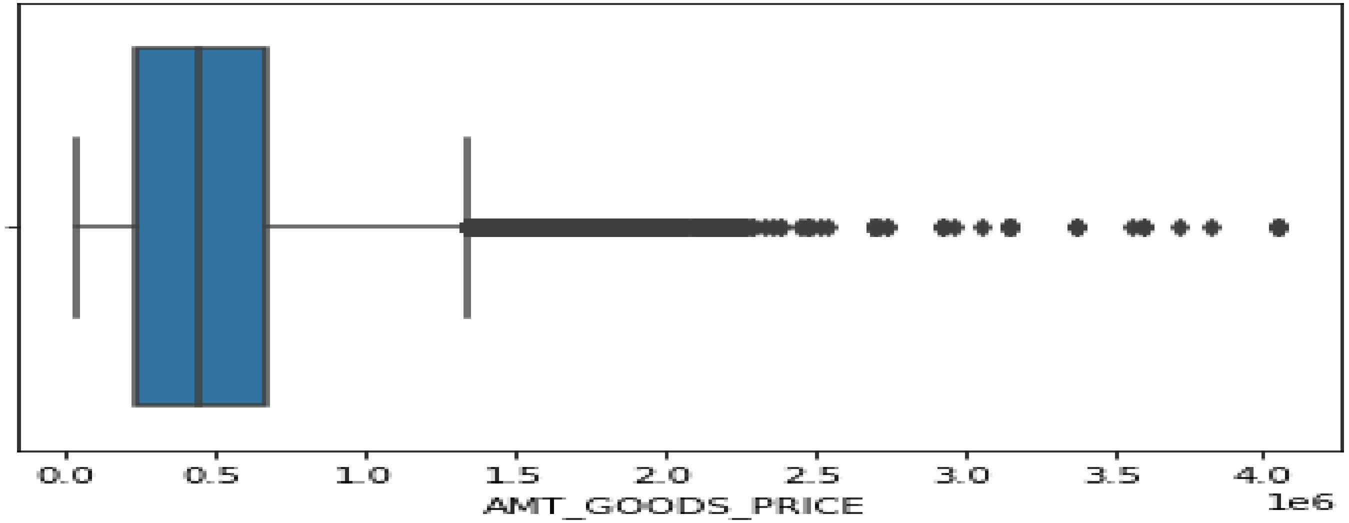
- This analysis refers to credit risk analysis to help company to make a decision on approving loan to the right applicant based on applicant's profile which means to look at the outcome of default and non-default applicants.

DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

✓ AMT_ANNUITY (Loan annuity) was having outlier null value or unexpected value that were exceeding to 250000 amount as shown in below diagram, those values has been imputed with median value i.e., 24903.

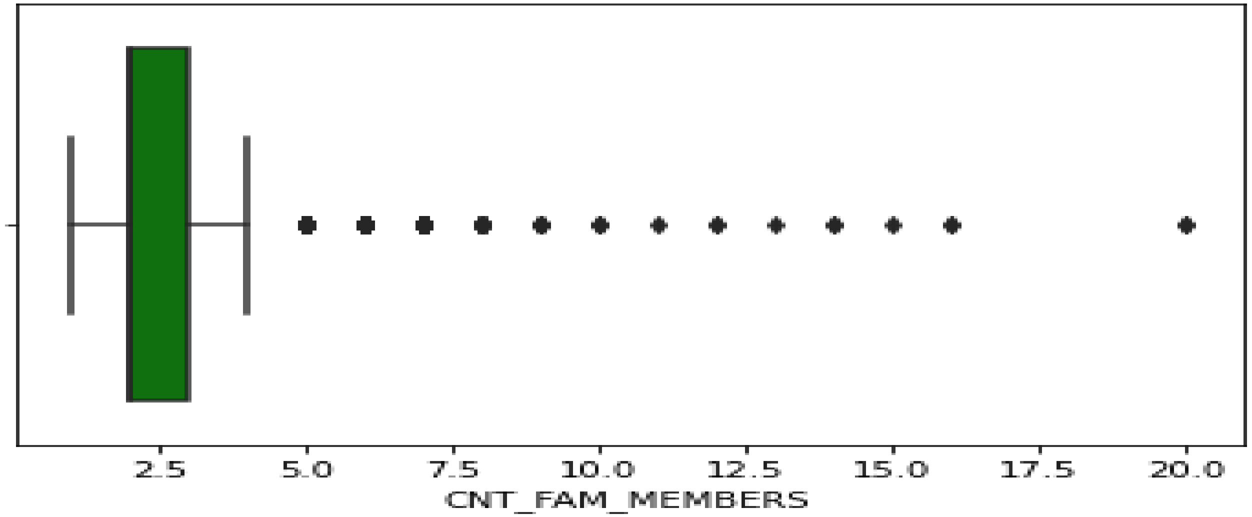


✓ AMT_GOODS_PRICE (For consumer loans it is the price of the goods for which the loan is given) was having outlier null value that were exceeding to 450000 as shown in below diagram, those values has been imputed with median value i.e., 450000.

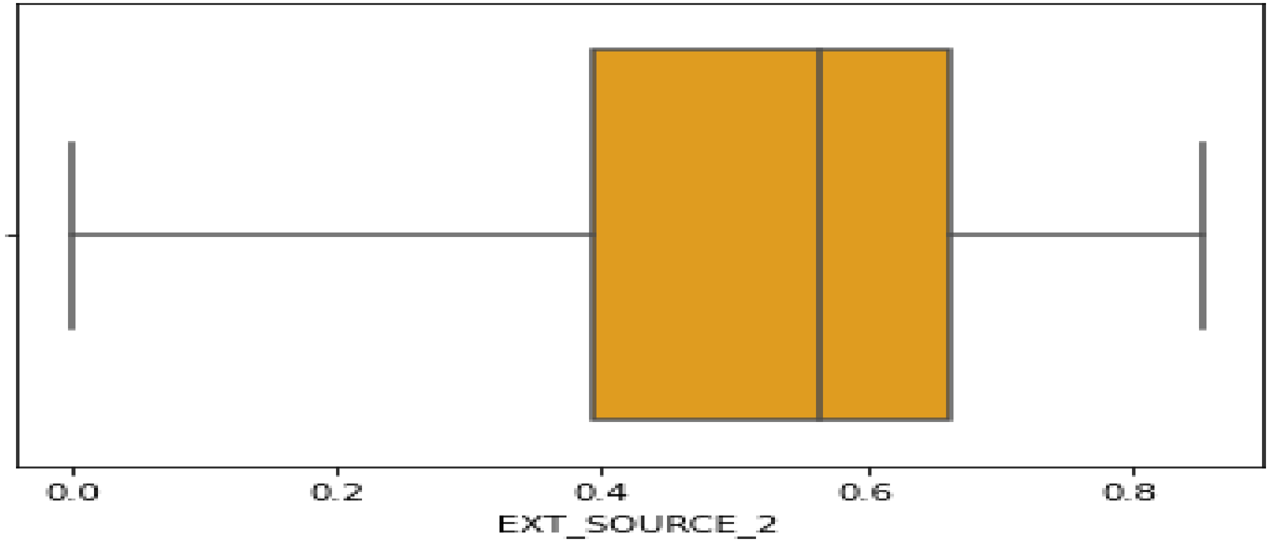


DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

- 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.

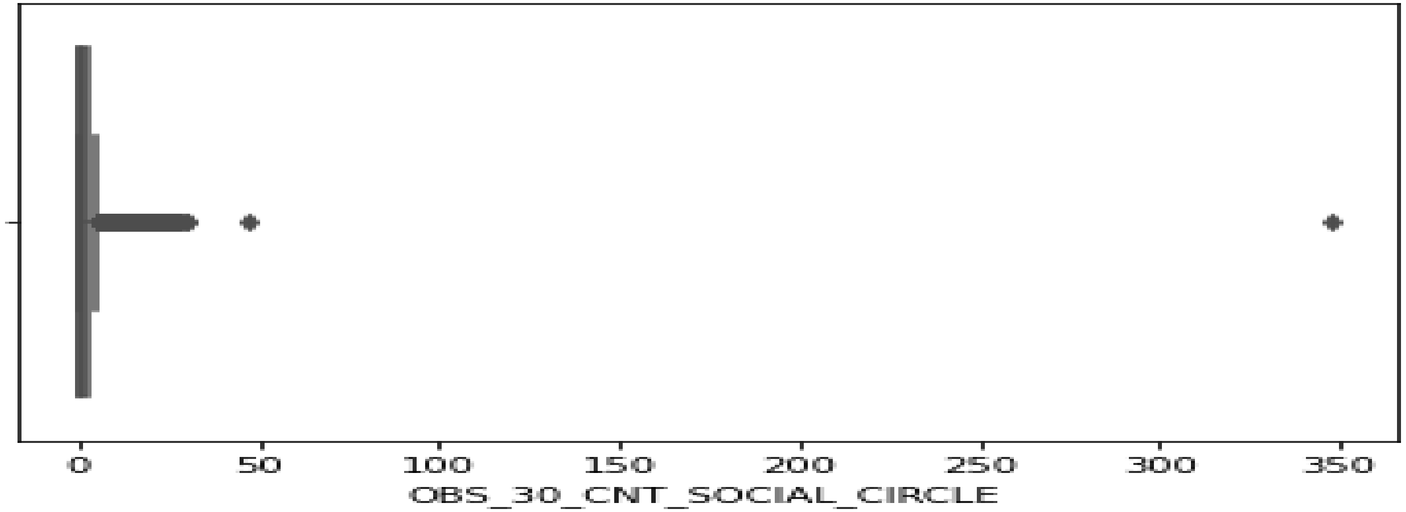


- EXT_SOURCE_2 (Normalized score from external data source) was not having any outlier null value as shown in below diagram. Hence, those values has been imputed with mean value i.e., 1.

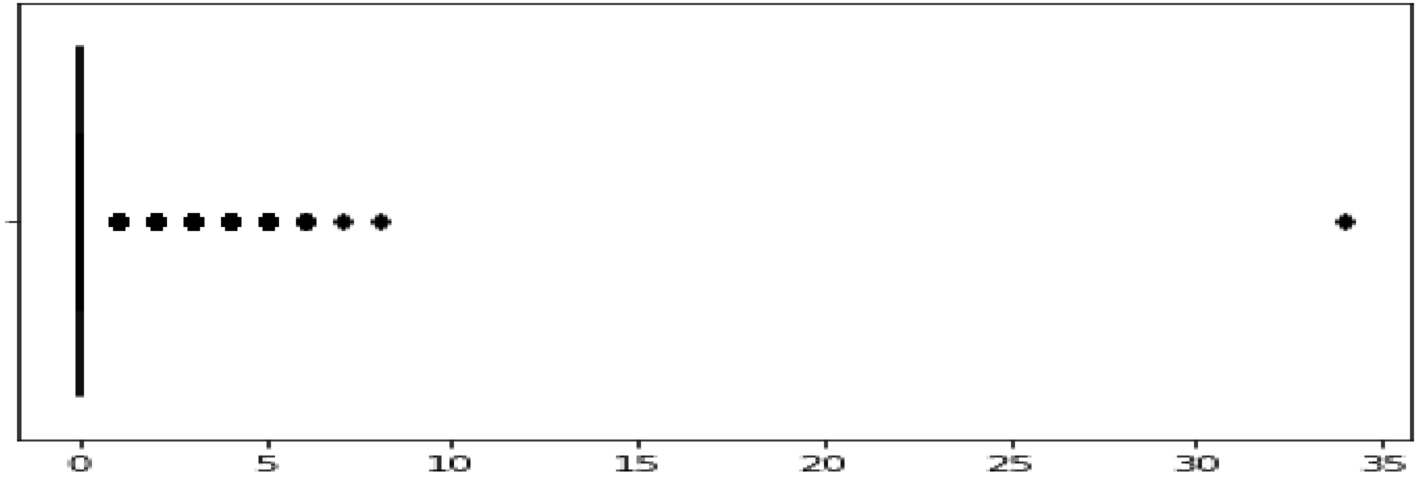


DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

- 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 value i.e., 0.

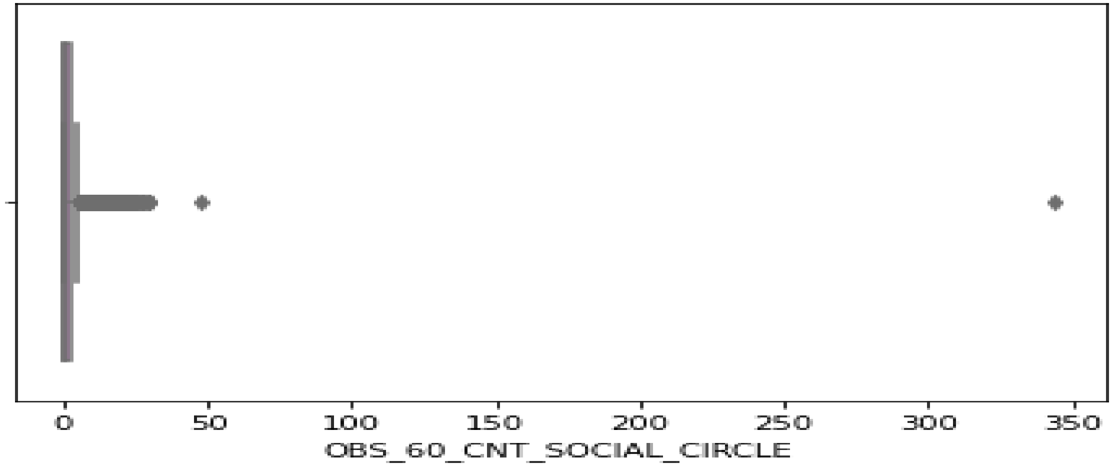


- DEF_30_CNT_SOCIAL_CIRCLE (Count of observation of client's social surroundings defaulted on 30 DPD days past due) was having outlier null value that were 34 approx, as shown in below diagram. Those values has been imputed with median value i.e., 0.

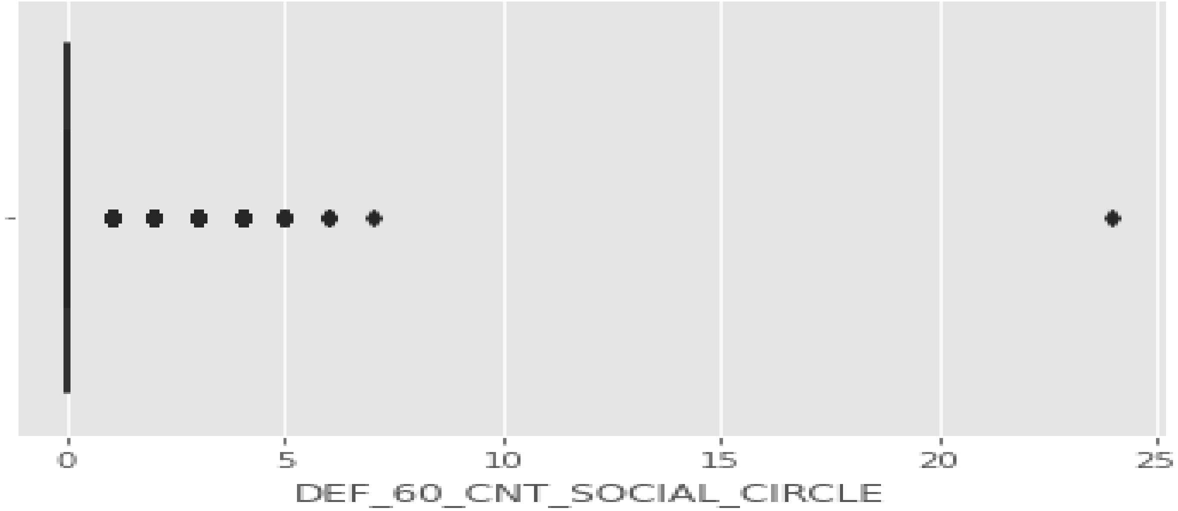


DATA ANALYSIS ON CONTINUOUS VARIABLE OF APPLICATION DATA

- 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 value i.e., 0.

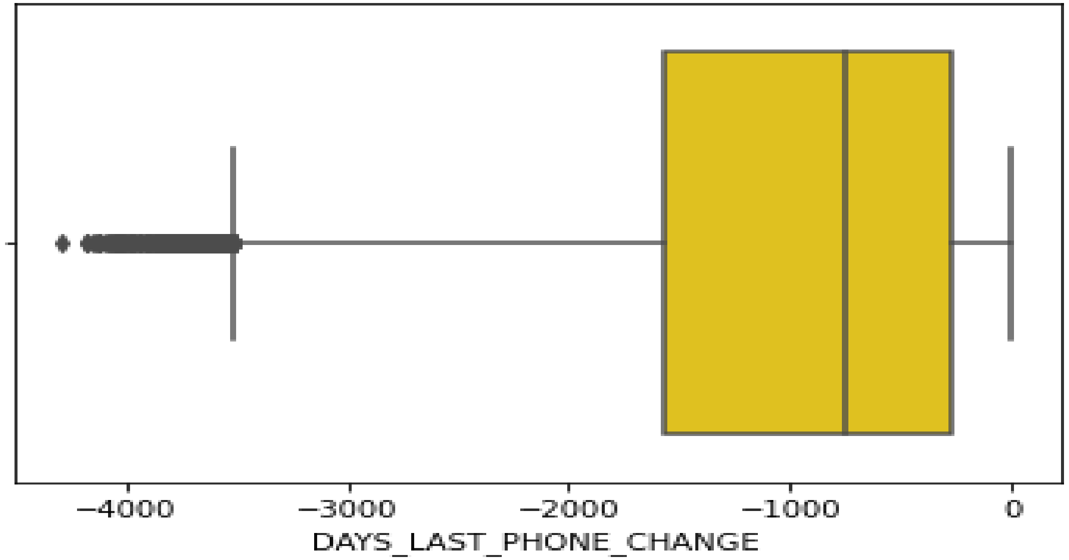


- DEF_60_CNT_SOCIAL_CIRCLE (Count of observation of client's social surroundings defaulted on 60 days past due) was having outlier null value that were 24 approx, as shown in below diagram. Those values has been imputed with median value i.e., 0.



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.



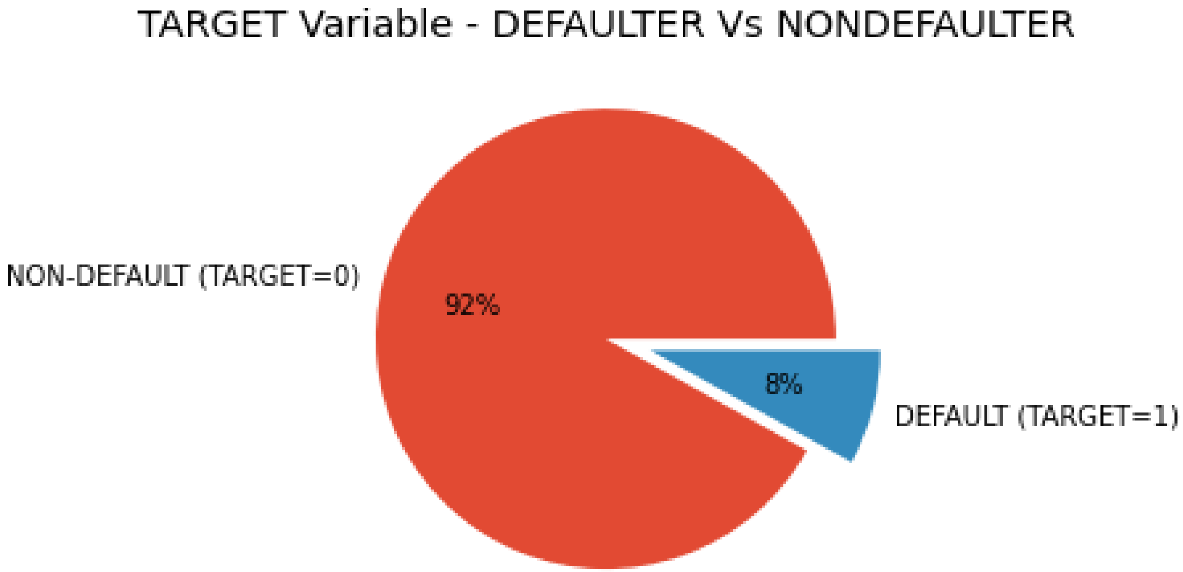
DATA ANALYSIS ON CATEGORICAL VARIABLE OF APPLICATION DATA

- 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.

CHECKING FOR DATATYPES AND CHANGING TO CORRECT TYPE

- Those columns which were having incorrect datatypes have been changed to the correct datatypes. Also, renamed values of some column with full form of it.
- For example – *In* CODE_GENDER column the values were Present as ‘M’, ‘F’ and ‘XNA’. So, XNA has been ignored and changed the value of ‘M’, ‘F’ as ‘Male’ and ‘Female’.
- In FLAG_OWN_CAR column the values were present as ‘Y’ and ‘N’. So, renamed the values as ‘Yes’ and ‘No’.
- In FLAG_OWN_REALTY column the values were present as ‘Y’ and ‘N’. So, renamed the values as ‘Yes’ and ‘No’.
- New column has been created that is Age with the age of every applicants.

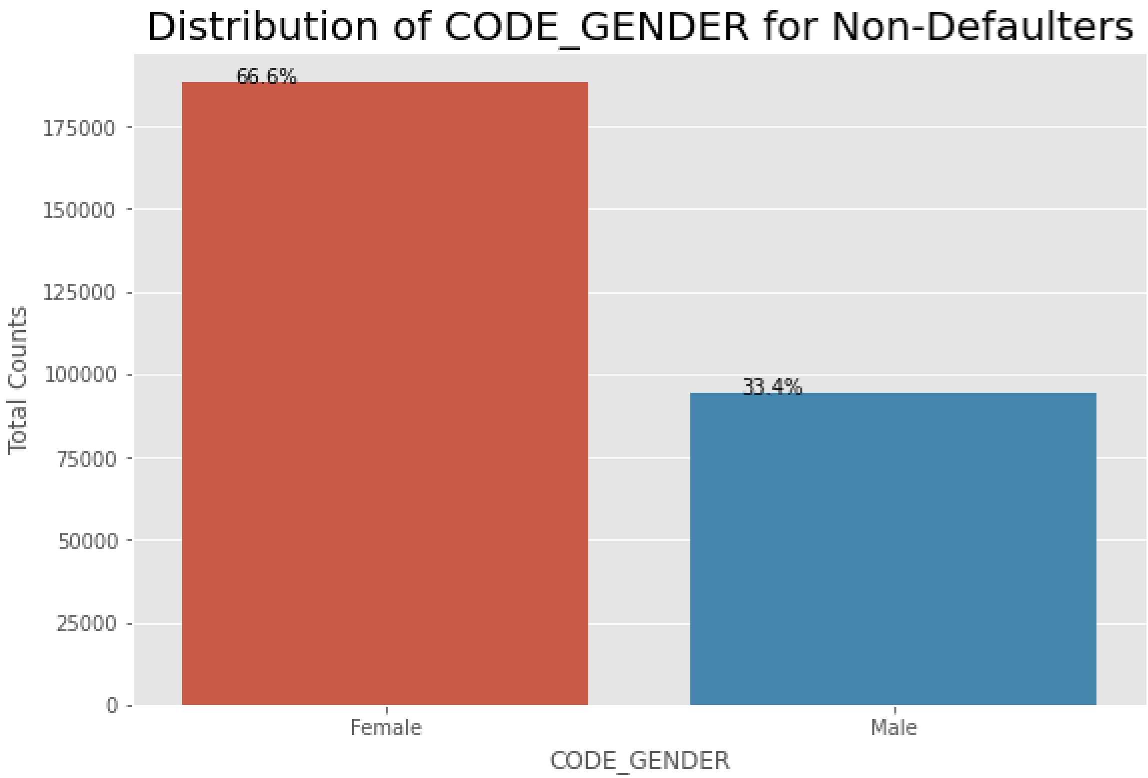
CHECKING IMBALANCE IN TARGET COLUMN

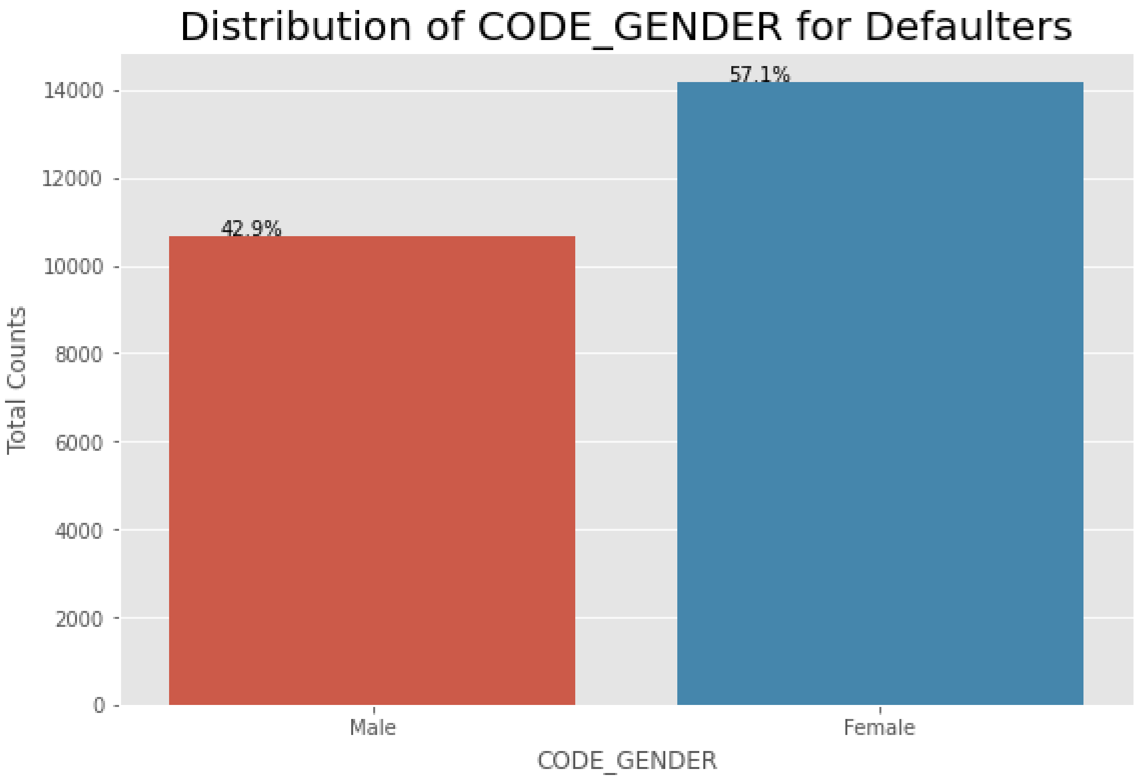


- Using this pie plot, we have a clear understanding that there are huge imbalance between defaulters and non-defaulters. This plot states that 92% of applicants are non-defaulters whereas 8% of applicants are defaulters

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- CODE_GENDER

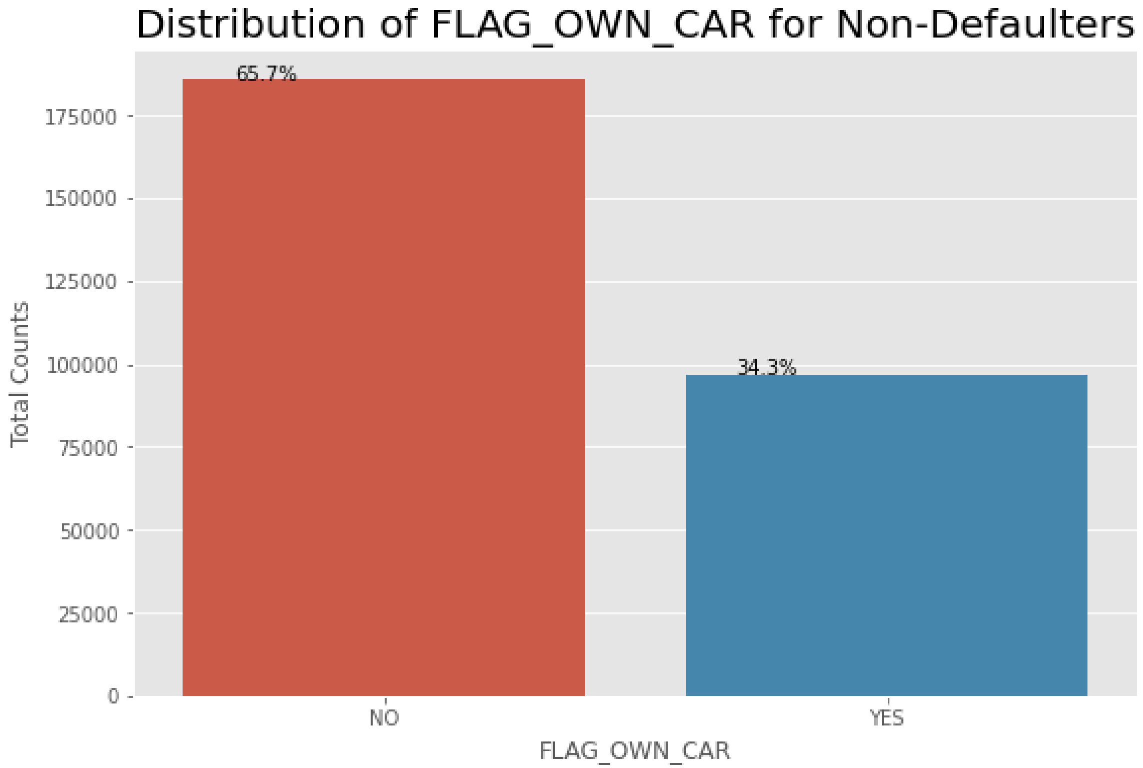

Distribution of CODE_GENDER for Non-Defaulters

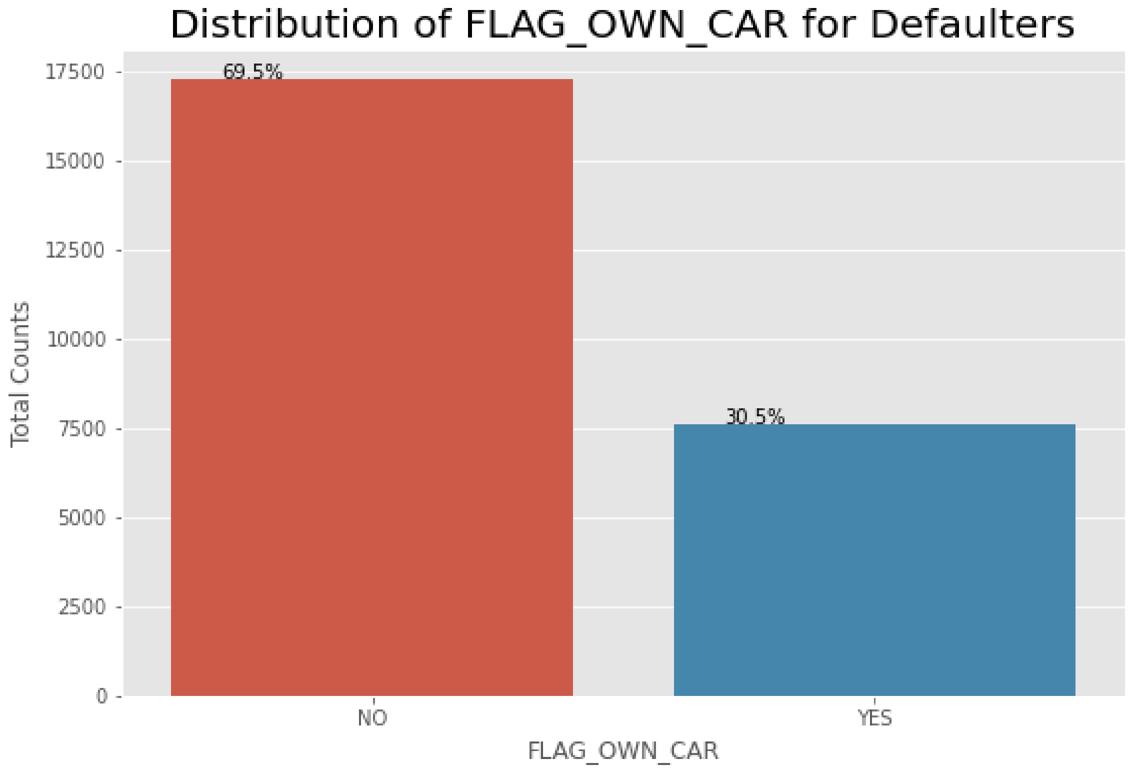

Distribution of CODE_GENDER for Defaulters

- Through this plot diagram we can understand that the percentage of female applicants are higher in comparison of male applicants to the non-defaulters i.e., approx 67% of female are non defaulters whereas 33% of male are non-defaulters, whereas percentage of females are higher than males in defaulters diagram also, this means count of female applicants are more than that of male applicants.

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- FLAG_OWN_CAR

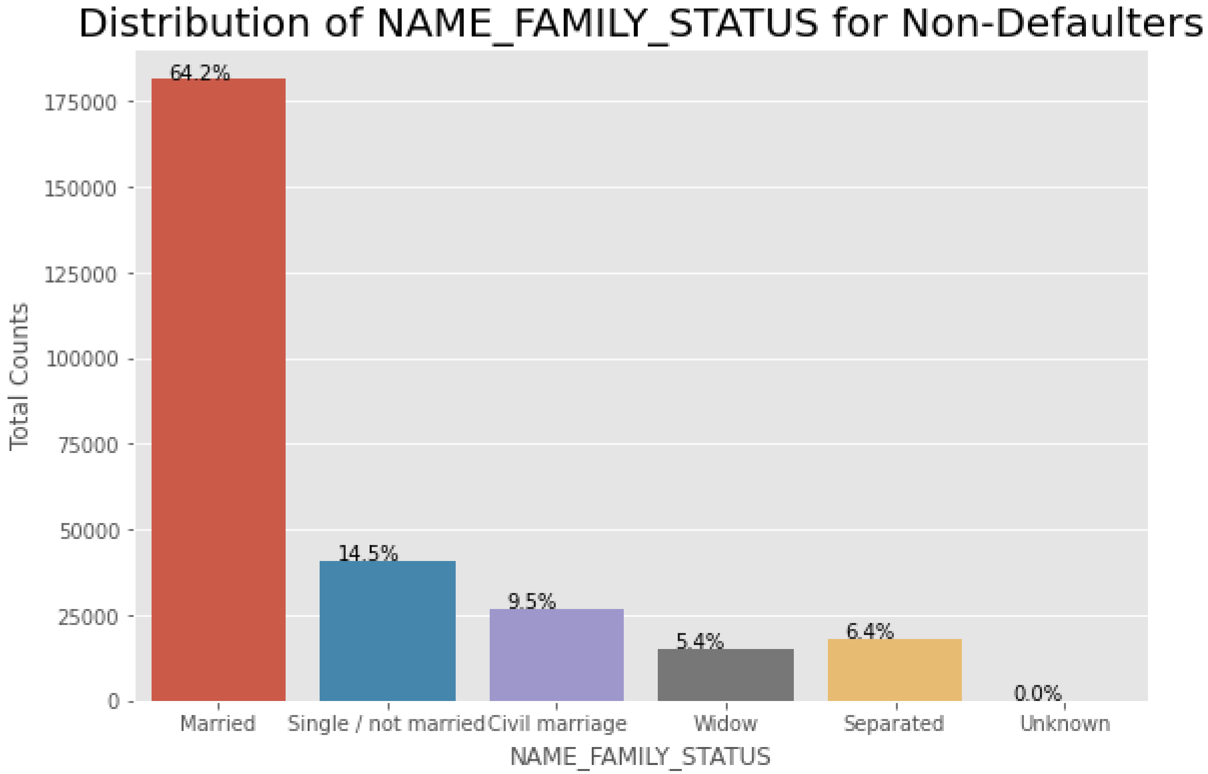

Distribution of FLAG_OWN_CAR for Non-Defaulters

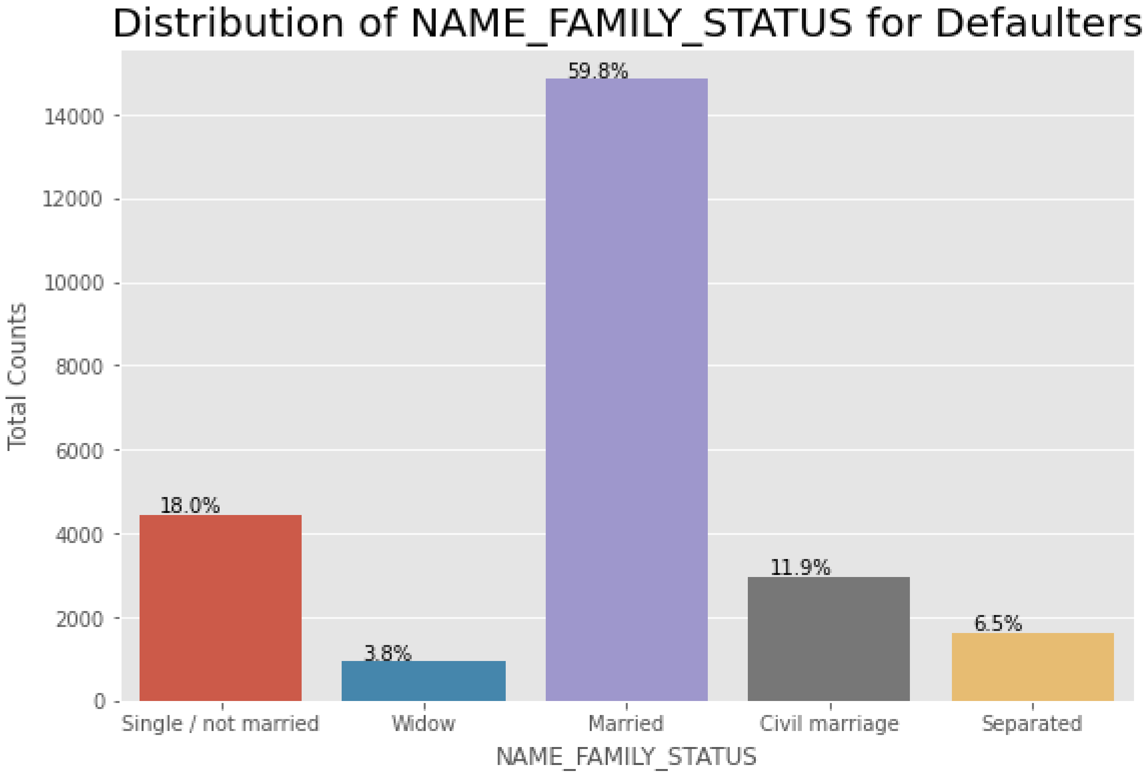

Distribution of FLAG_OWN_CAR for Defaulters

- 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 percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- NAME_FAMILY_STATUS

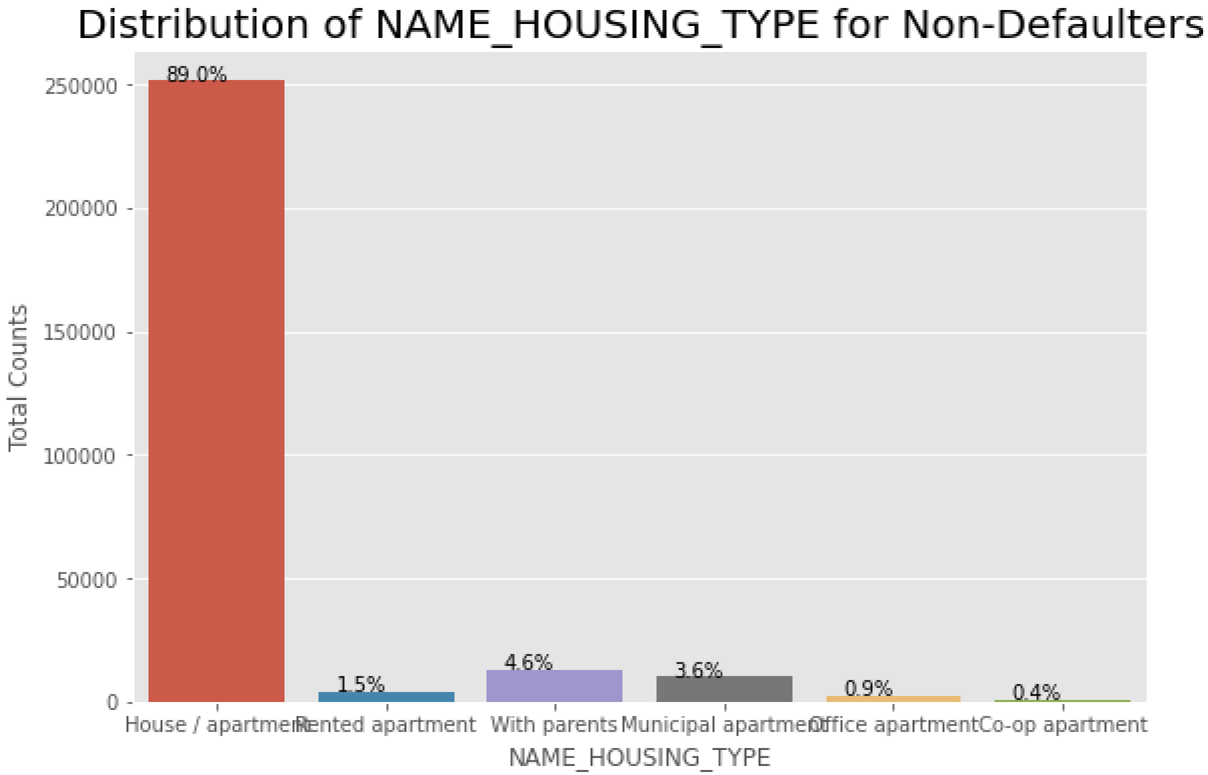


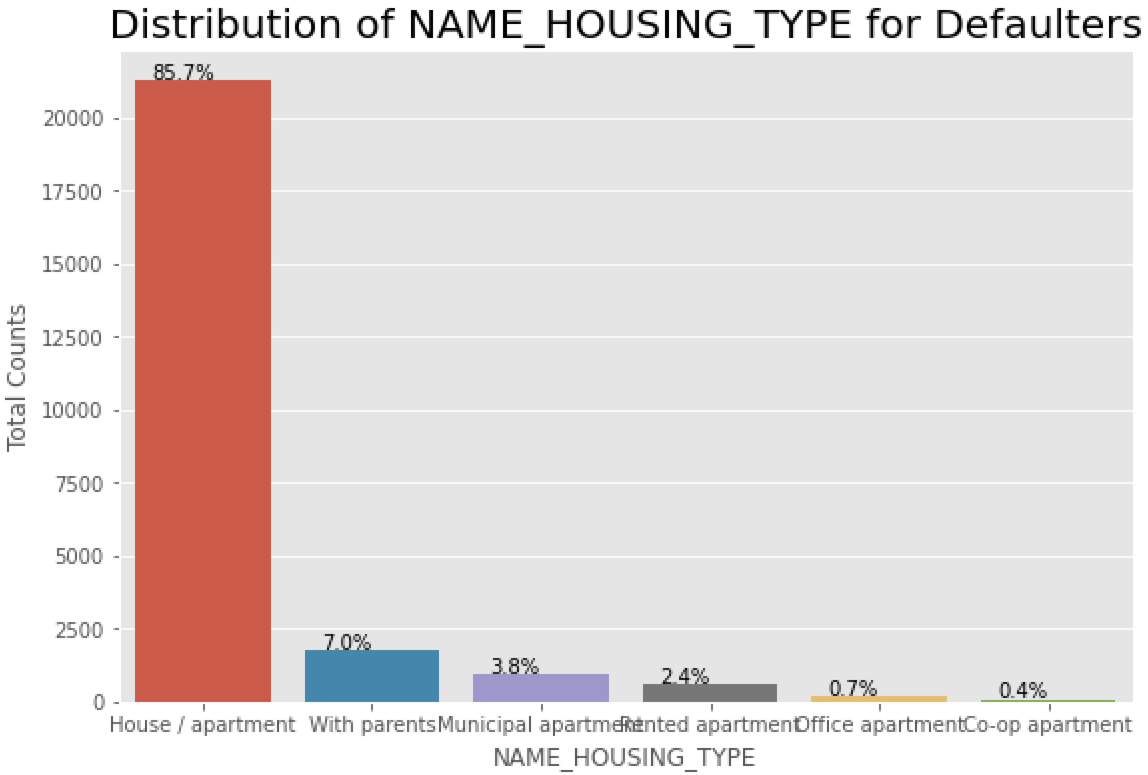


- Married people tend to apply for more loans comparatively.
- But from the graph we see that Single/non Married people contribute 14.5% to Non Defaulters and 18% to the defaulters. So there is more risk associated with them.

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- NAME_HOUSING_TYPE

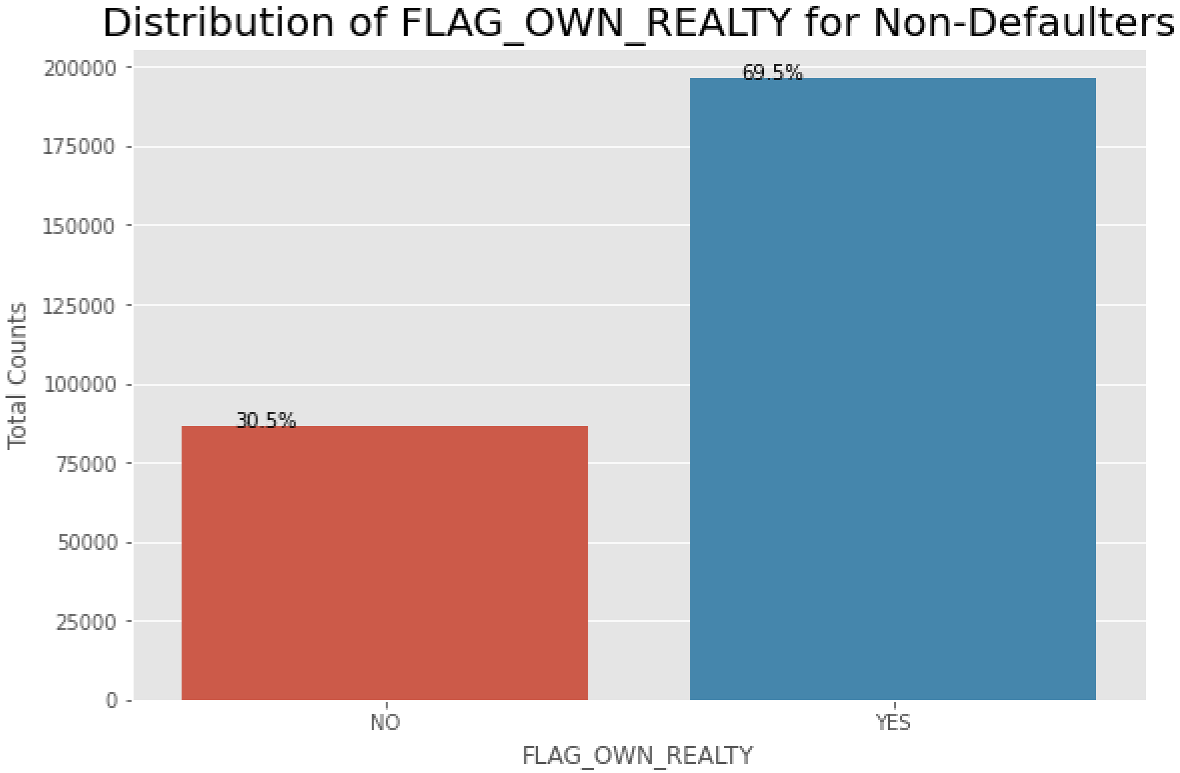

Distribution of NAME_HOUSING_TYPE for Non-Defaulters

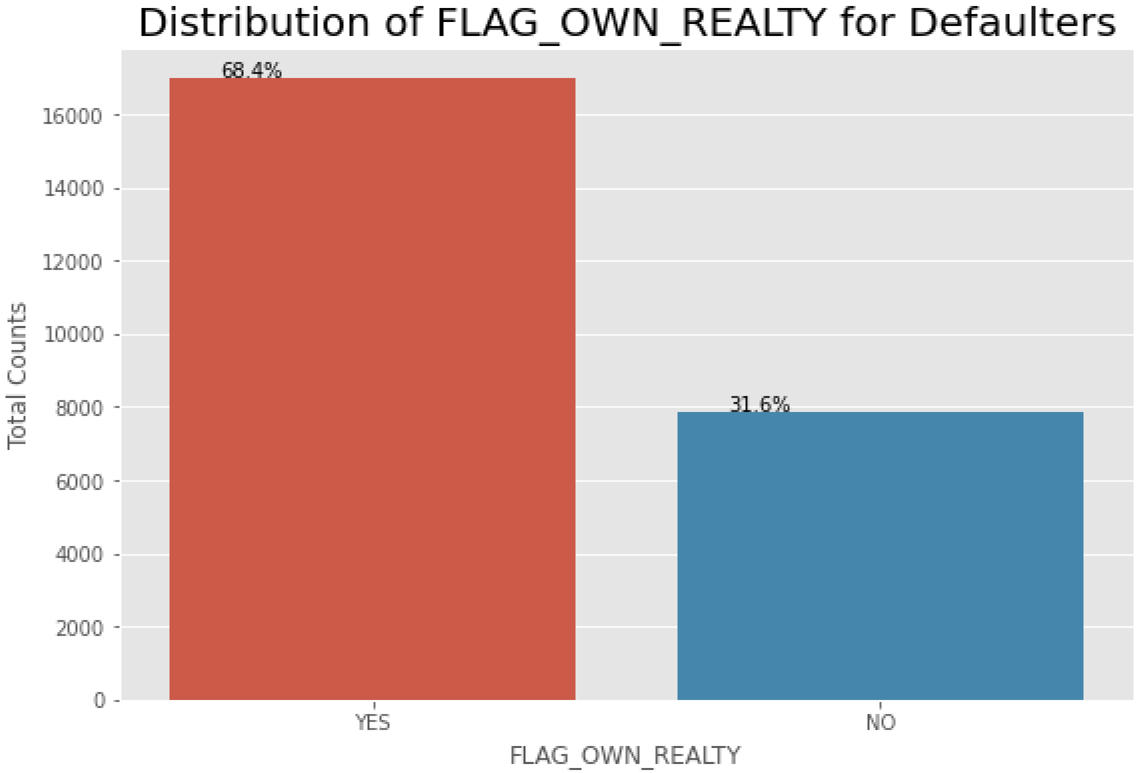

Distribution of NAME_HOUSING_TYPE for Defaulters

- It is clear from the graph that people who have House/Appartment, tend to apply for more loans.
- People living with parents tend to default more often when compared with others.The reason could be their living expenses are more due to their parents living with them

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- FLAG_OWN_REALTY

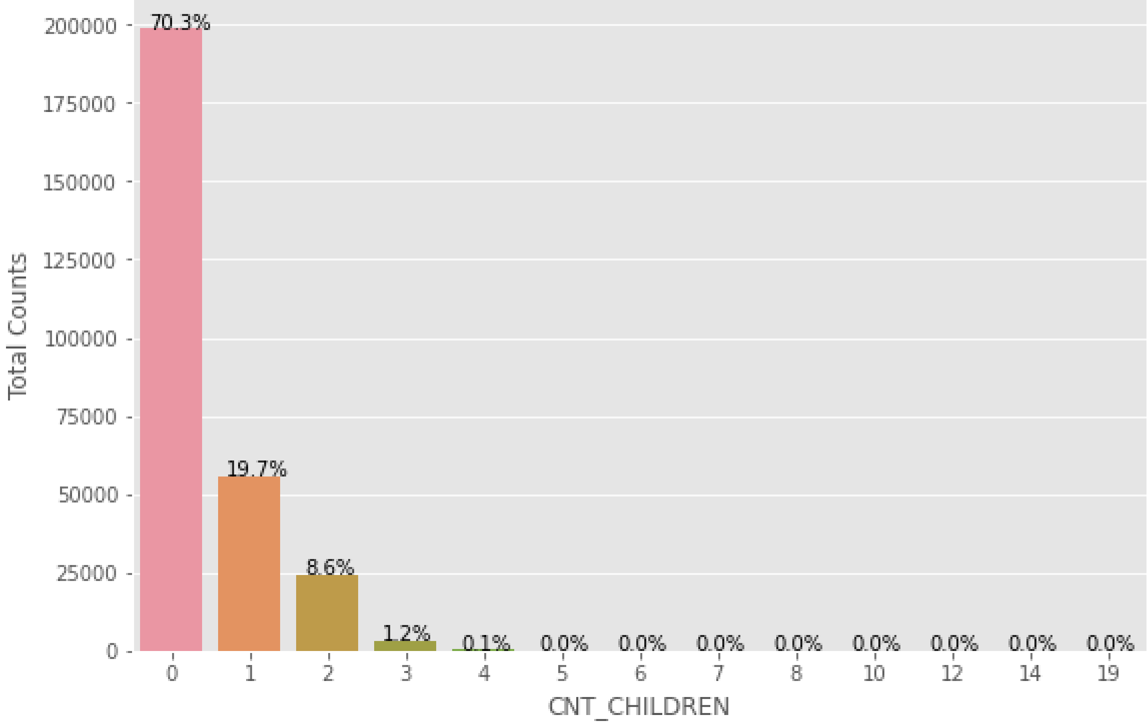




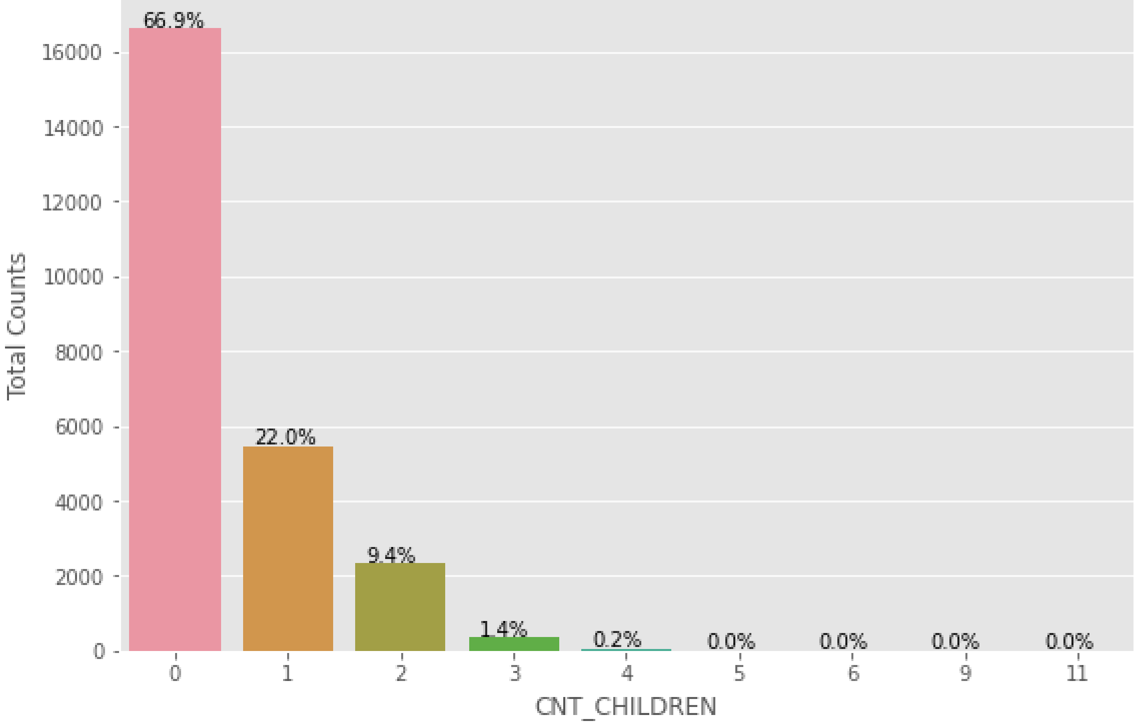
- Those applicants are higher in percent of non-defaulters who has their own houses/property.
- Same result as well for defaulters also that means the applicants who have their own property has taken the filled maximum application for the loan.

UNIVARIATE ANALYSIS ON CATEGORICAL VARIABLE

- CNT_CHILDREN



Distribution of CNT_CHILDREN for Non-Defaulters

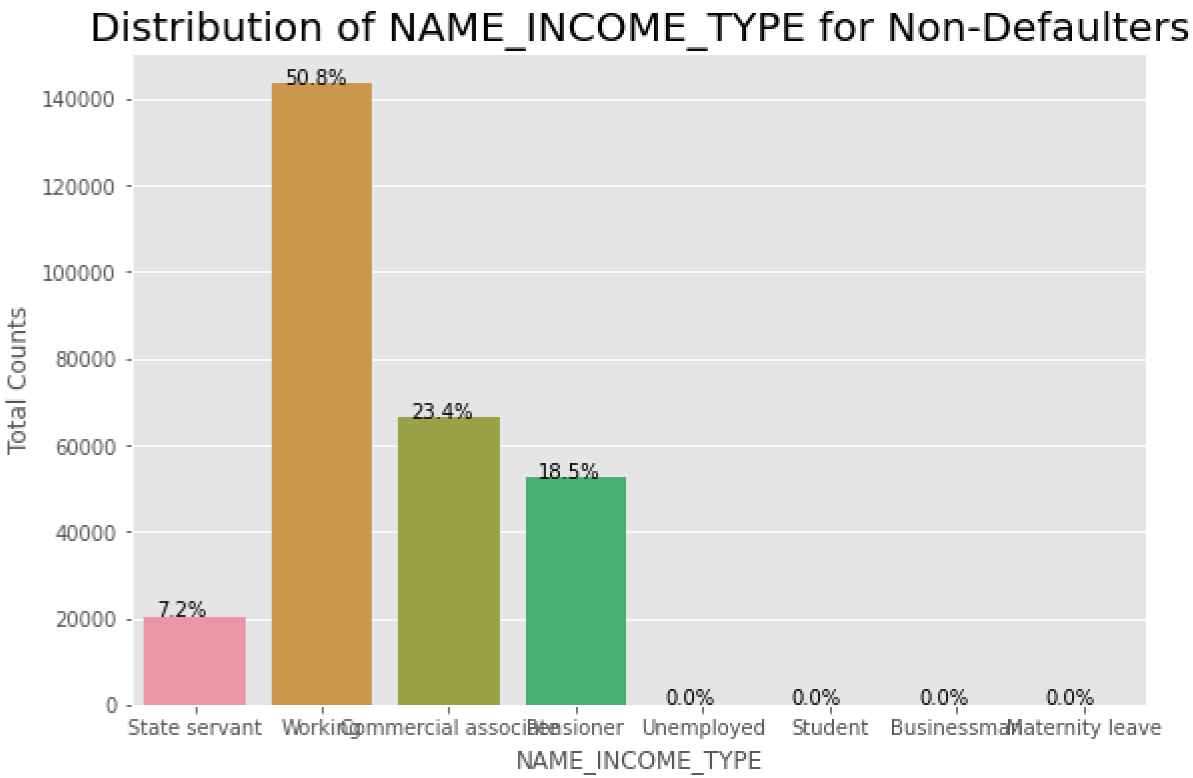


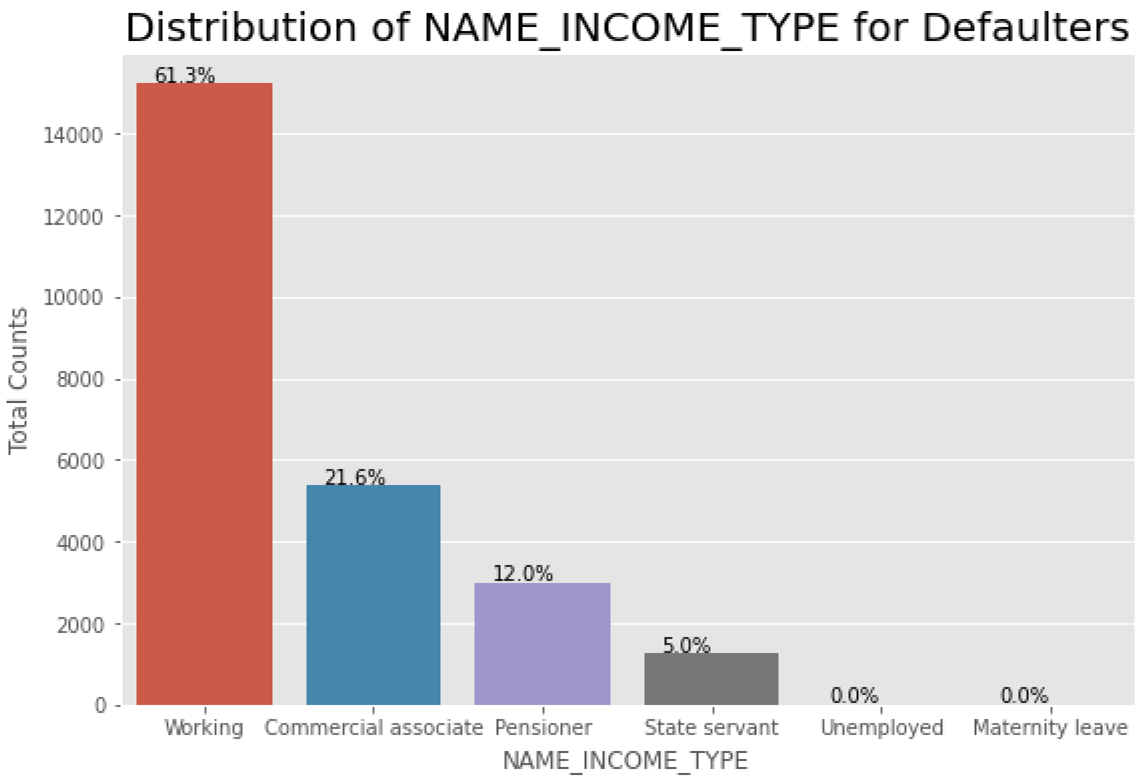
Distribution of CNT_CHILDREN for Defaulters

- 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 VARIABLE

- NAME_INCOME_TYPE

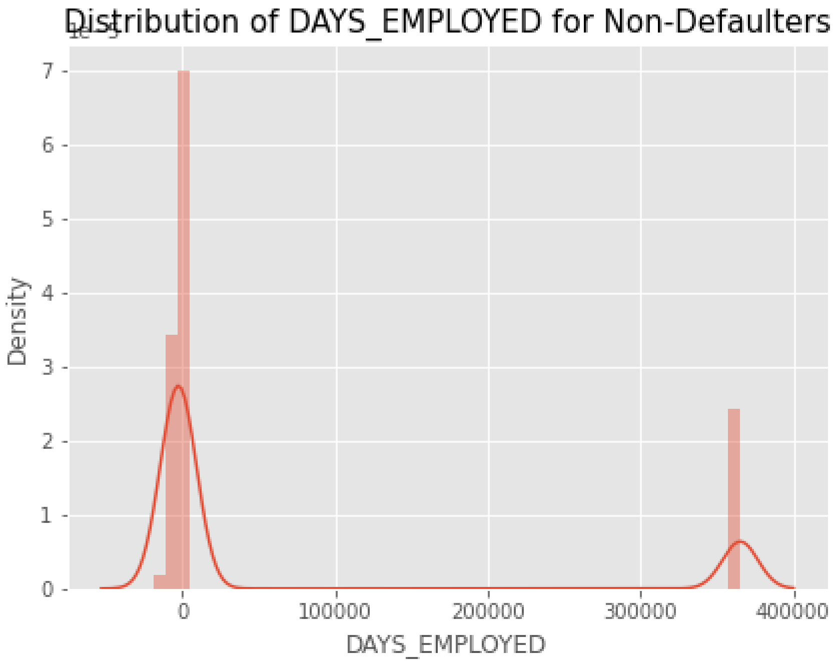


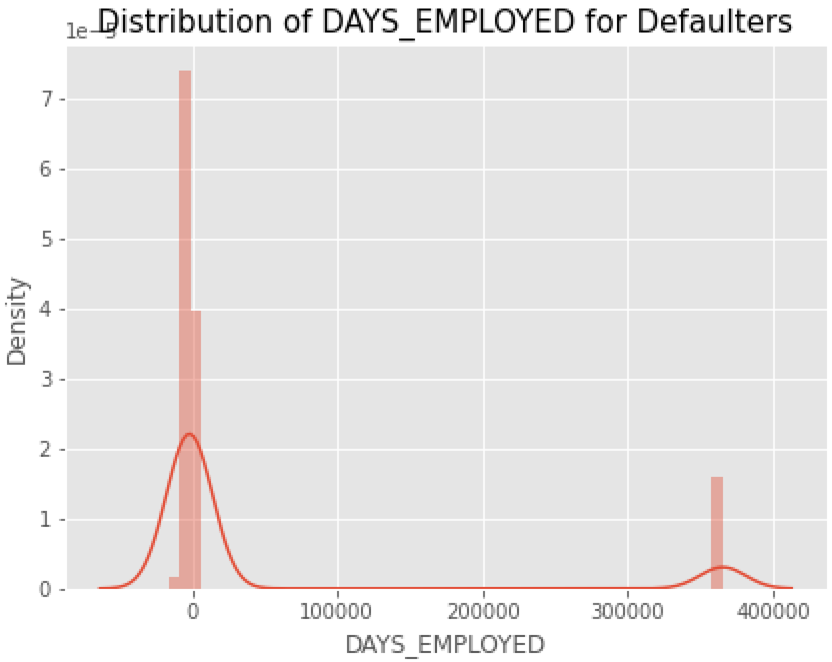


- Income distribution represents that 50% of applicants who are working are counted as non defaulters and 23% who are commercial, 18% who are pensioners, 7% who are state servants are counted as non defaulters. Whereas, 61% of applicants who are working are counted as defaulters, commercial 21% commercial, 12% pensioners, 5% state servants are counted as defaulters.

UNIVARIATE ANALYSIS ON CONTINUOUS VARIABLE

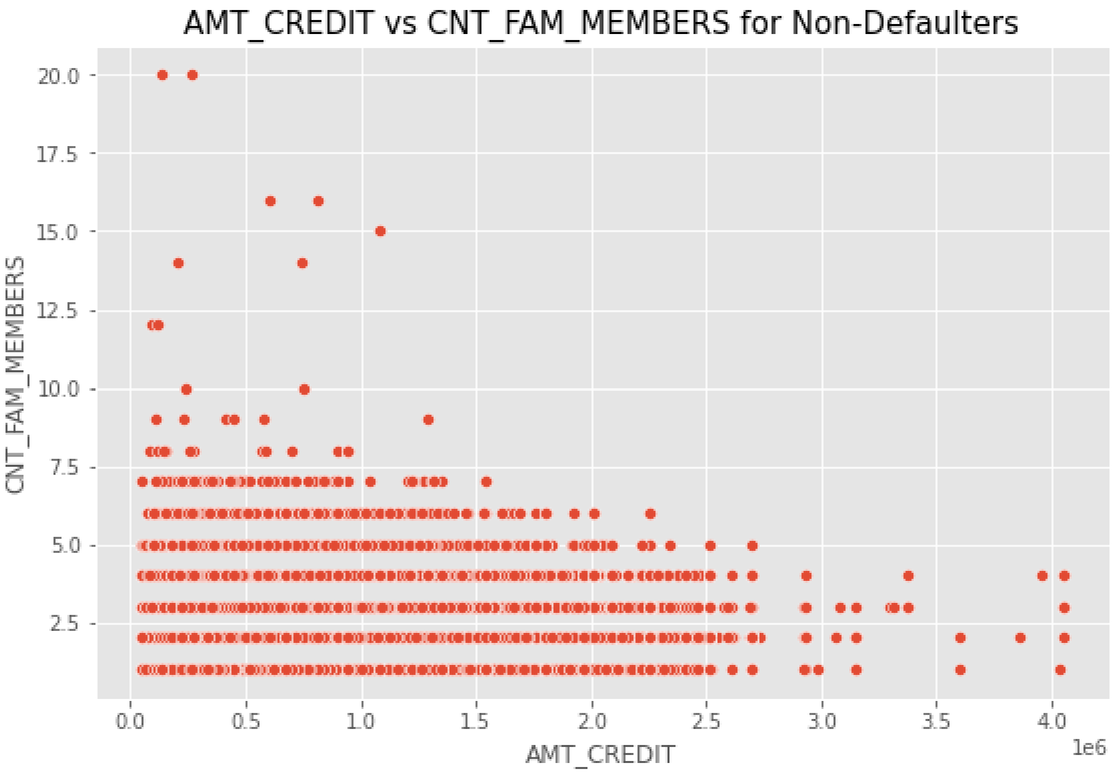
- DAYS_EMPLOYED
-

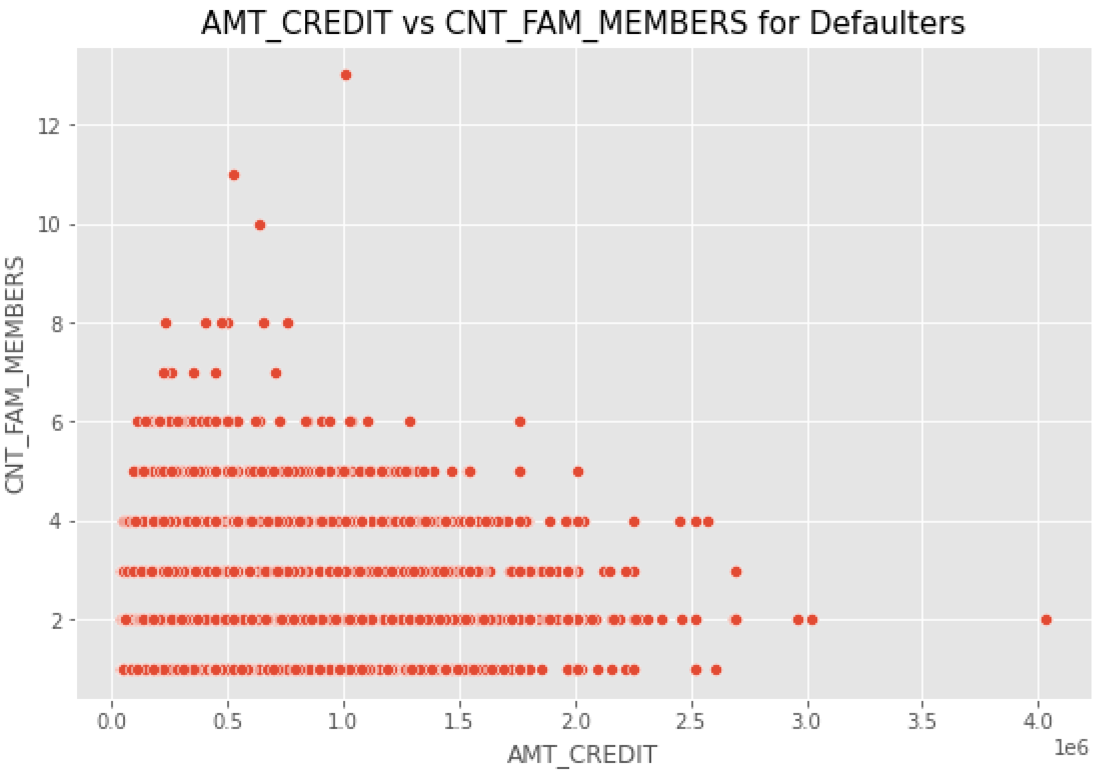




BIVARIATE ANALYSIS OF CONTINUOUS VARIABLES

- AMT_CREDIT and CNT_FAM_MEMBERS
-

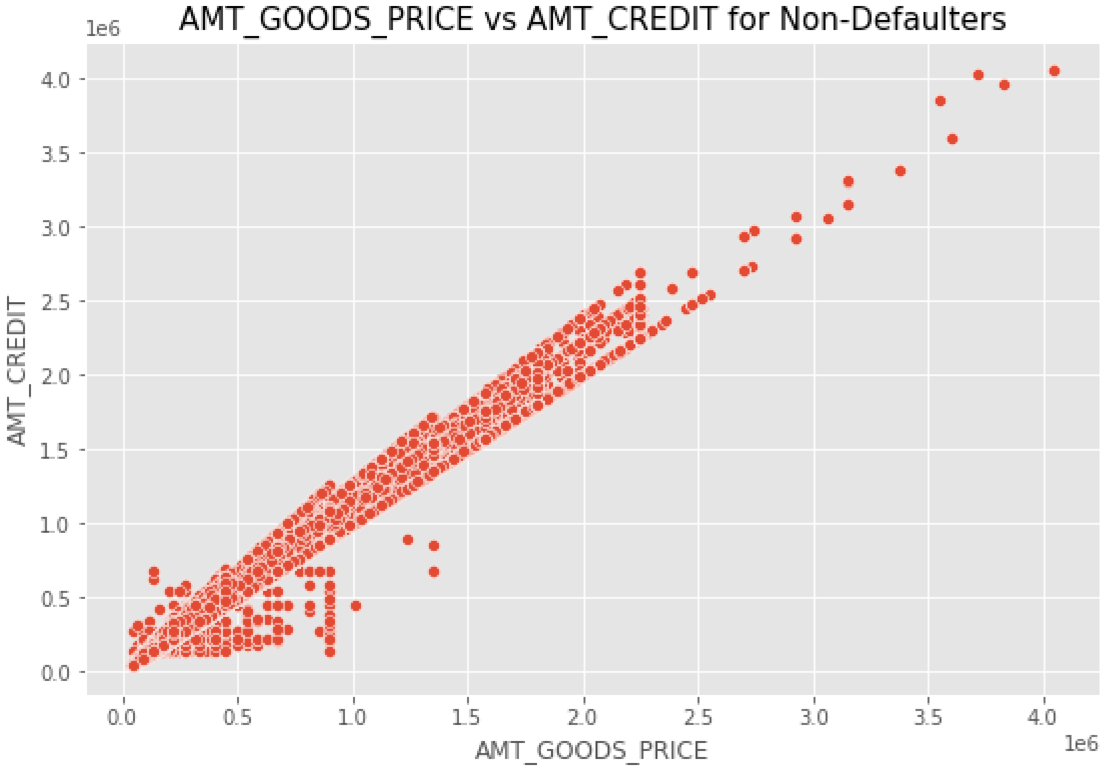




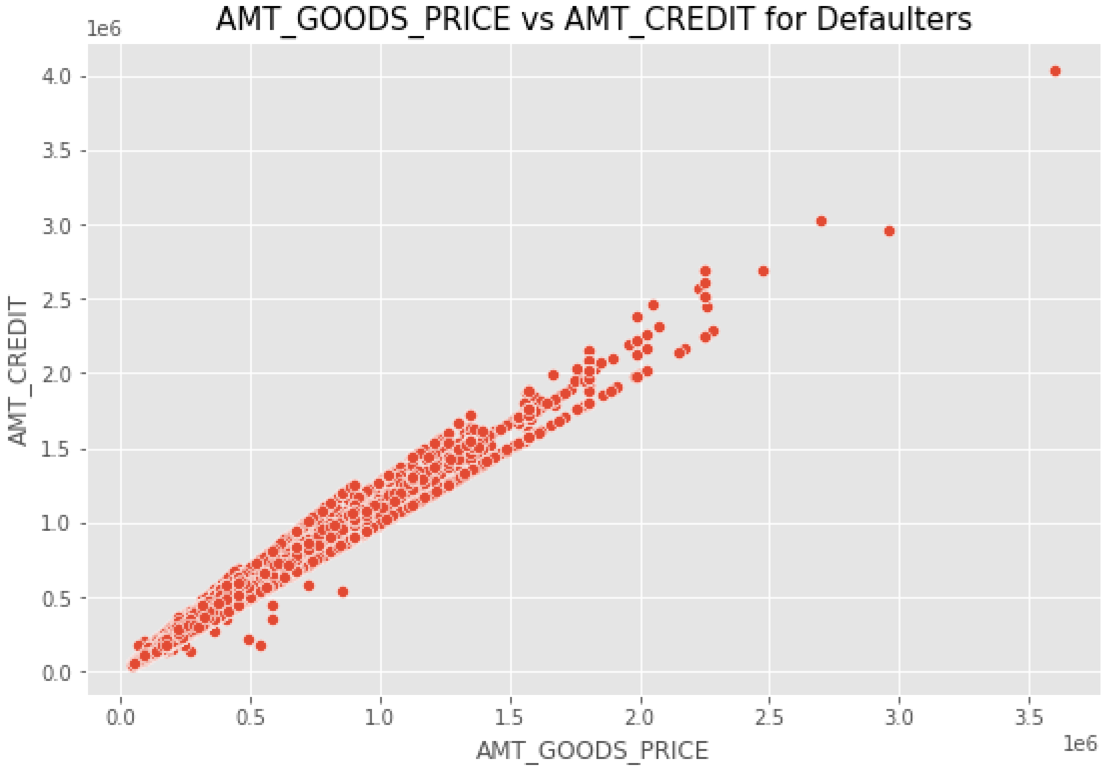
- We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often.

BIVARIATE ANALYSIS OF CONTINUOUS VARIABLES

- AMT_GOODS_PRICE and AMT_CREDIT



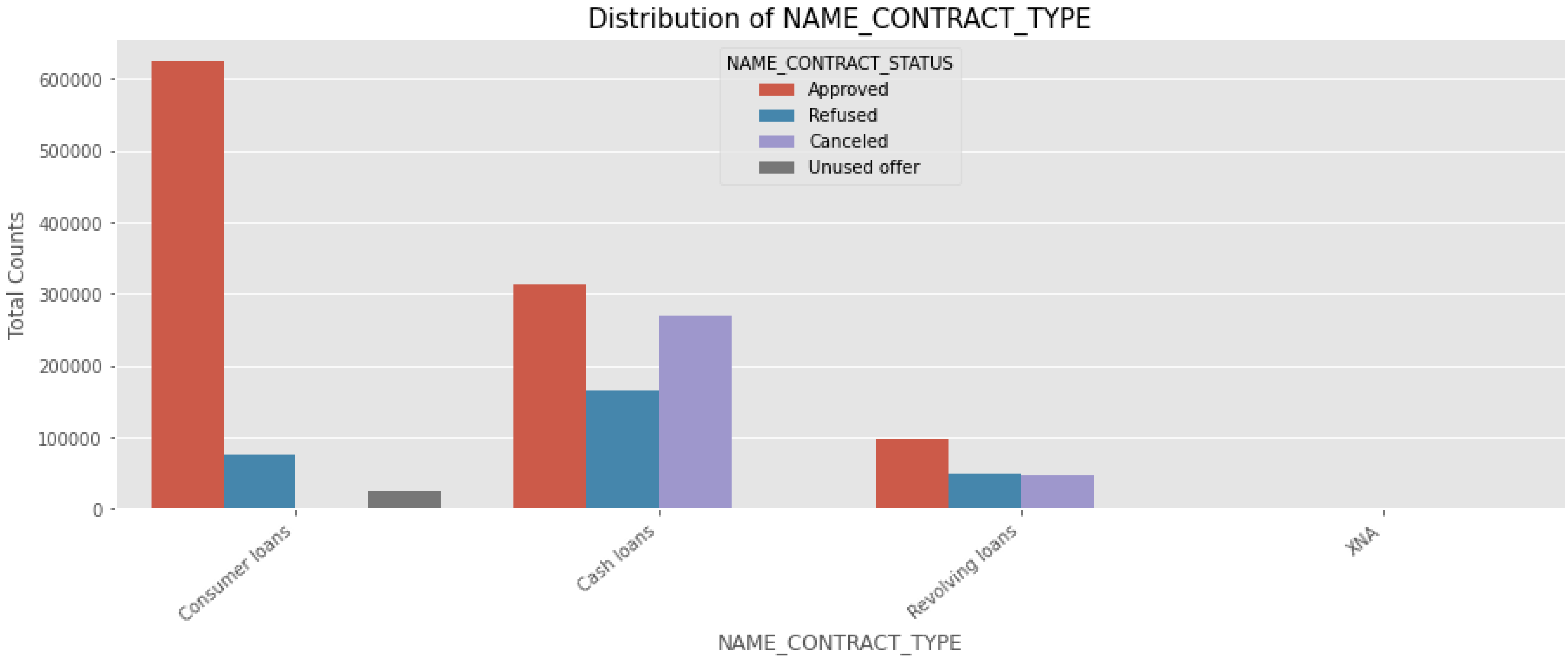
AMT_GOODS_PRICE vs AMT_CREDIT for Non-Defaulters



AMT_GOODS_PRICE vs AMT_CREDIT for Defaulters

DATA ANALYSIS FOR PREVIOUS APPLICATION DATA

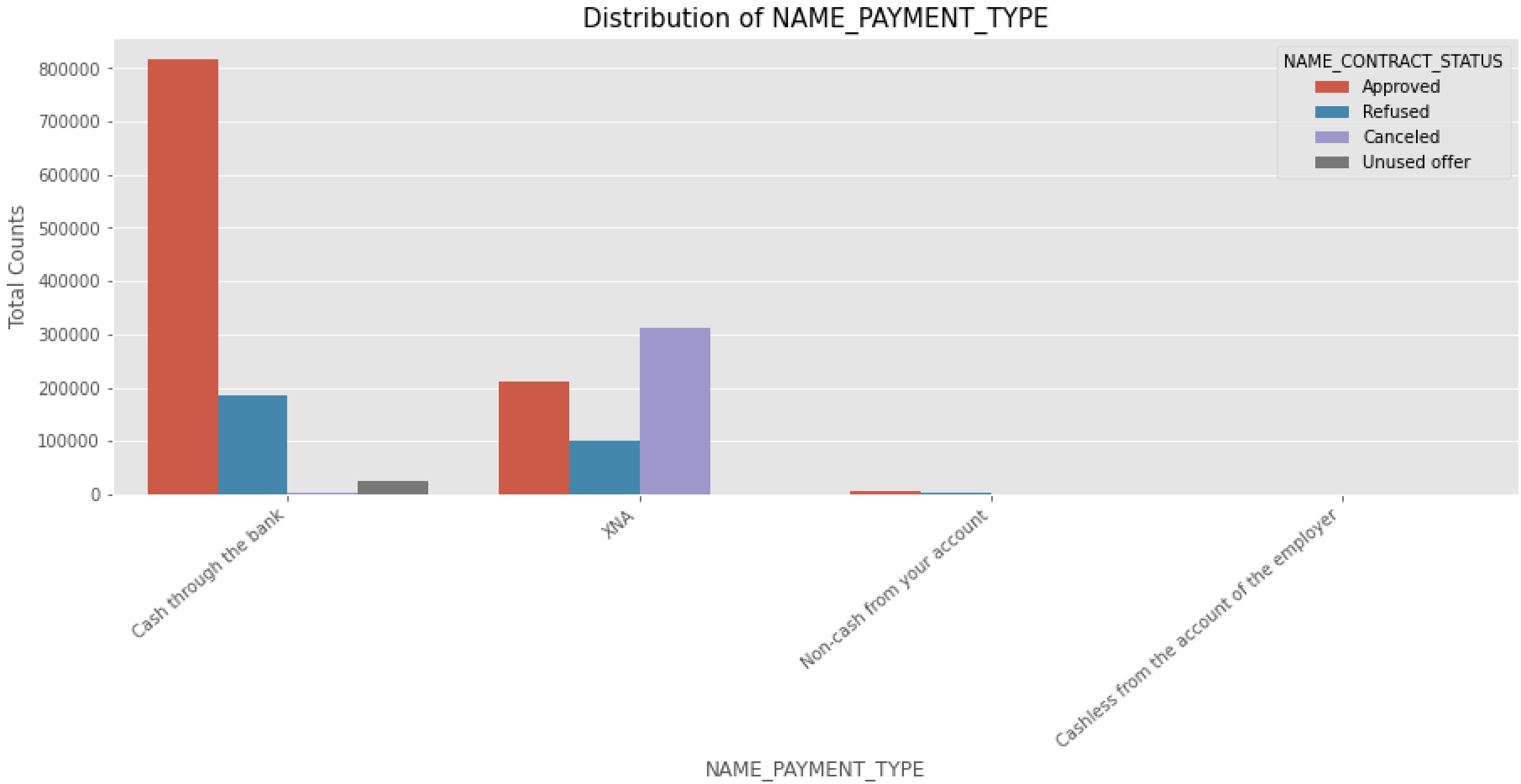
- NAME_CONTRACT_TYPE
-



- From the above chart, we can infer that, most of the applications are for 'Cash loan' and 'Consumer loan'. Although the cash loans are refused more often than others

DATA ANALYSIS FOR PREVIOUS APPLICATION DATA

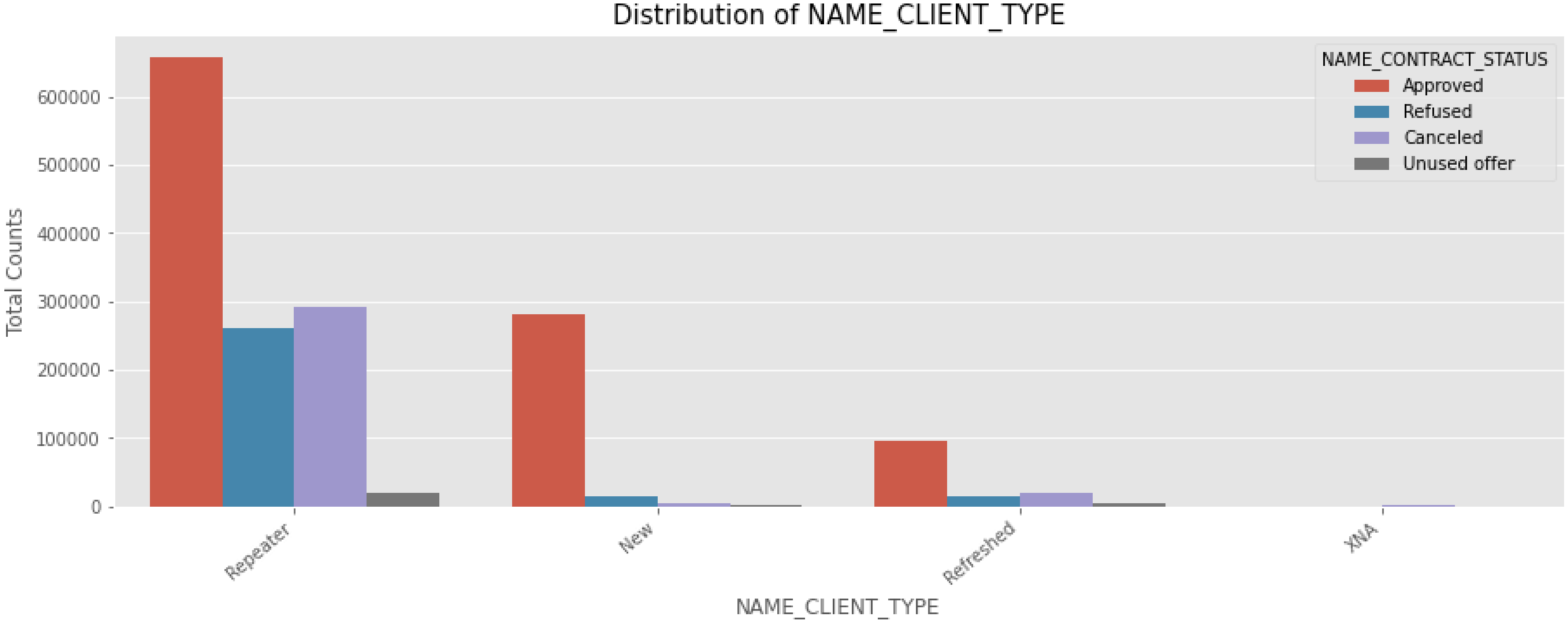
- NAME_PAYMENT_TYPE
-



- From the above chart, we can infer that most of the clients chose to repay the loan using the 'Cash through the bank' option We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers.

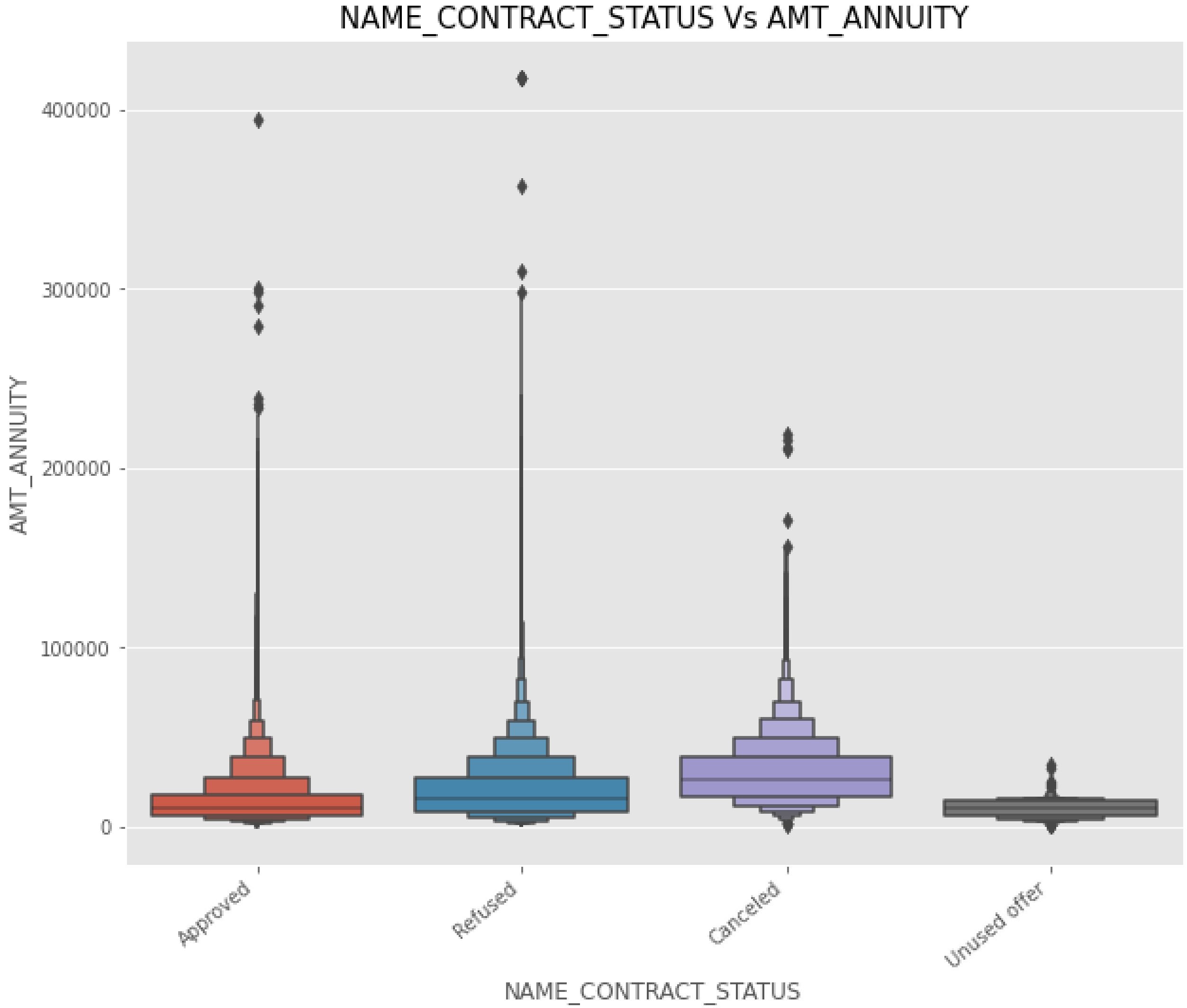
DATA ANALYSIS FOR PREVIOUS APPLICATION DATA

- 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 refused most often.

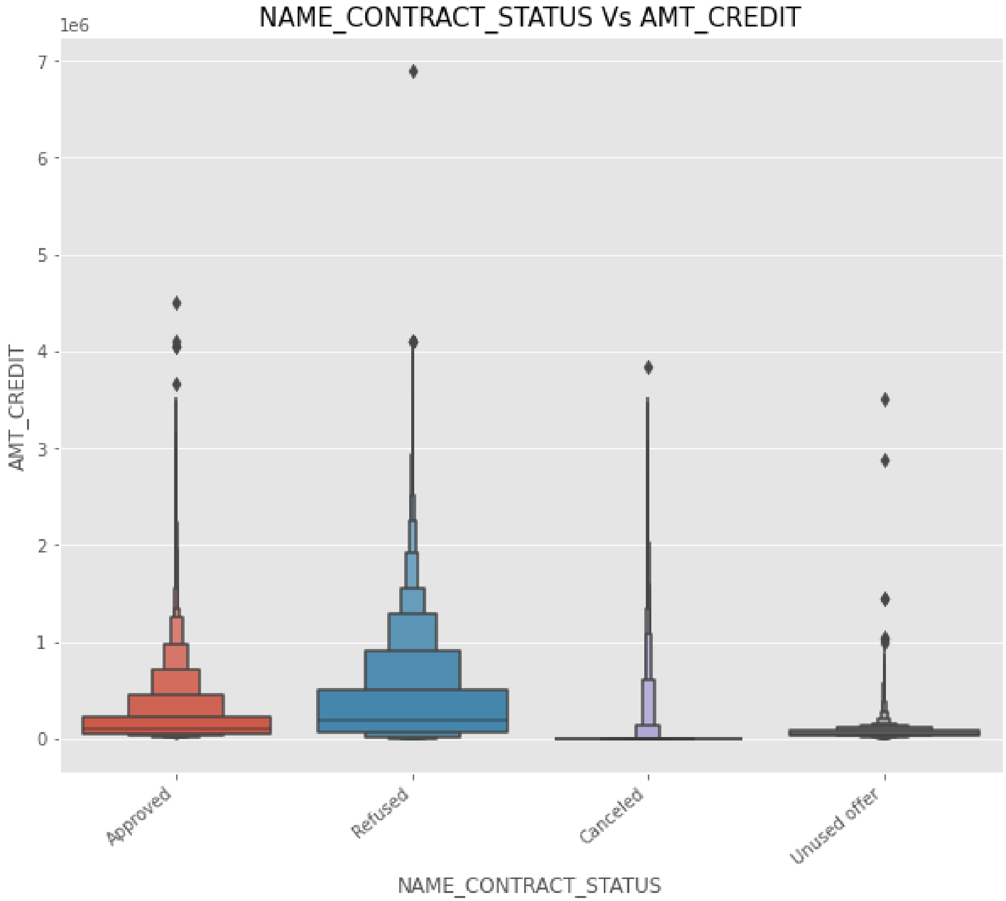
USING BOX PLOT TO DO BIVARIATE ANALYSIS ON CATEGORICAL AND NUMERIC COLUMNS

- NAME_CONTRACT_STATUS and AMT_ANNUITY
- 

- From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time.
We also see that applications with too high AMT ANNUITY also got refused more often than others.

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

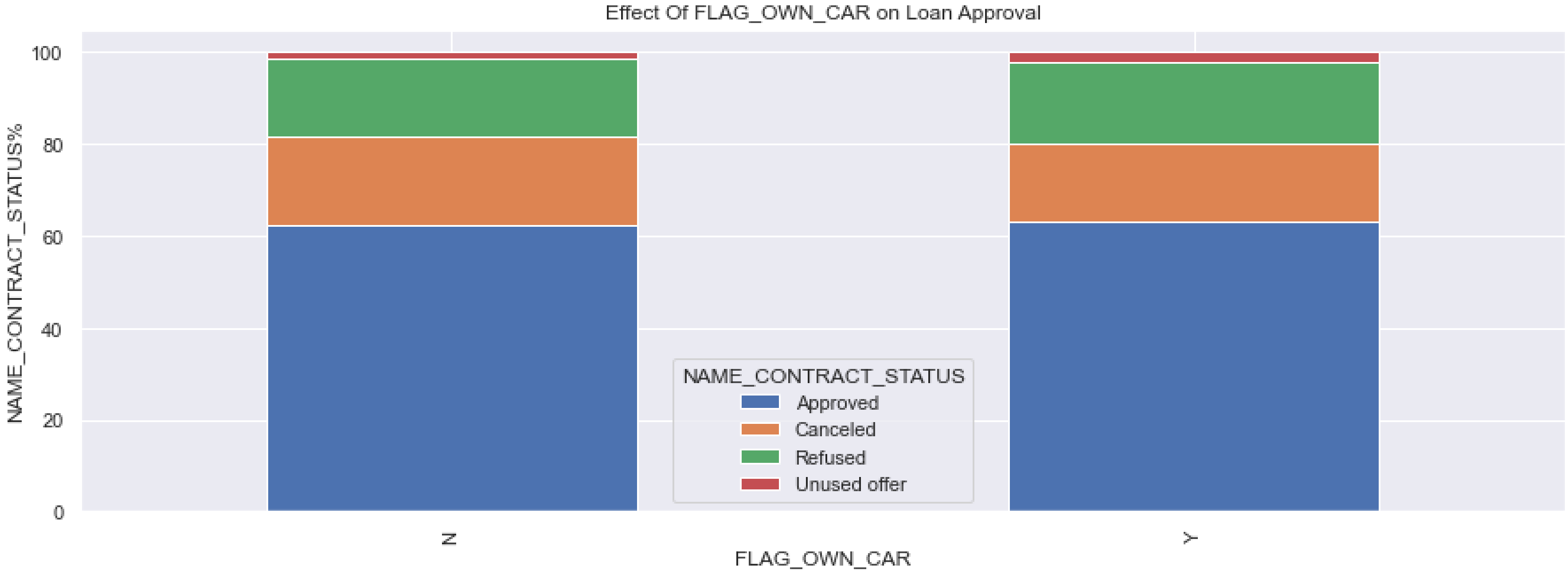
- NAME_CONTRACT_STATUS and AMT_CREDIT
-



- We can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

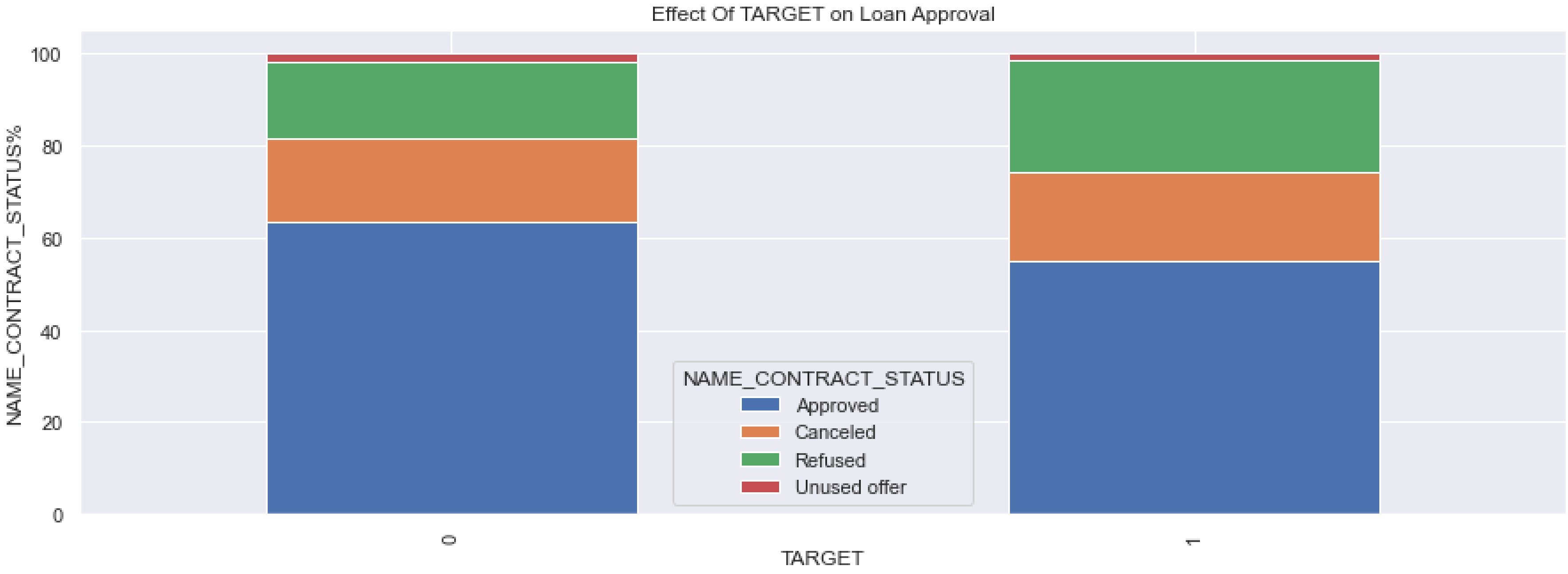
MERGING APPLICATION DATA AND PREVIOUS APPLICATION DATA

- FLAG_OWN_CAR and NAME_CONTRACT_STATUS

- 

- We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default. The bank can add more weightage to car ownership while approving a loan amount.

MERGING APPLICATION DATA AND PREVIOUS APPLICATION DATA

- TARGET and NAME_CONTRACT_STATUS
- 

- We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.