CREDIT EDA CASE STUDY

By Abin John & Asha Sasidharan

DATA EXPLORATION

There are 122 variables in the Current Application data dataset.

Problem Statement:

To find a solution to minimize the risk of losing money while lending to customers. So we need to particularly concentrate on the charged off case.

DATA CLEANING AND MANIPULATION

Dealing with Missing values :

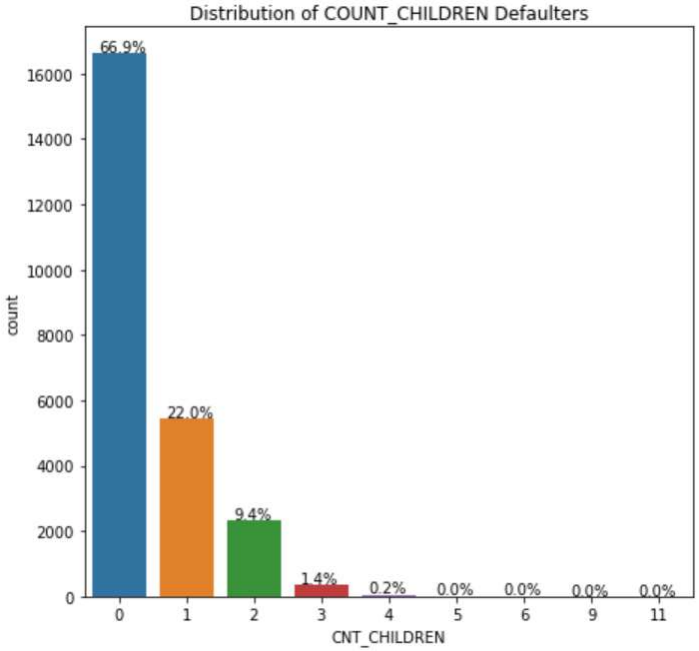
- Removed all the columns with more than 50% nulls values
- Selected columns with less or equal to than 10% null values
  10 Such columns were selected..
- Replaced null values on the fist 5 columns from above list using possible solutions.

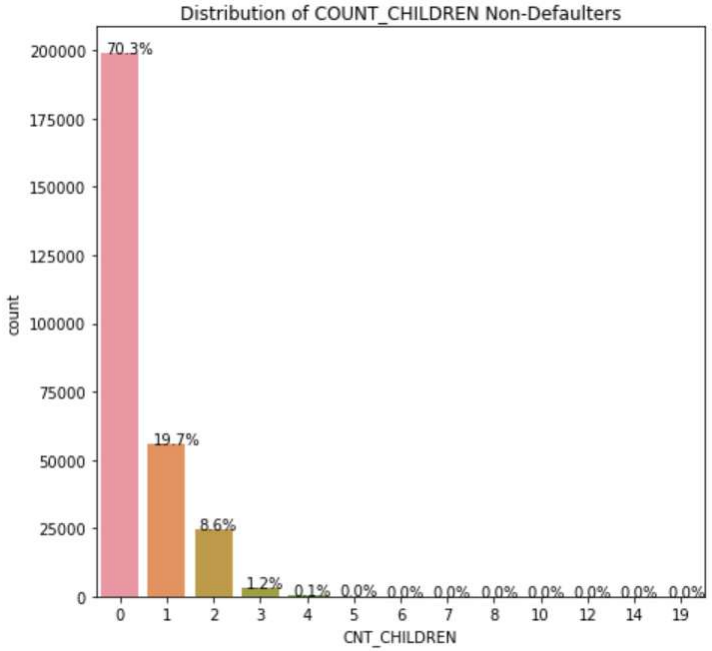
Manipulation on given data set for the ease of Analysis:

- A new data set of 30 columns derived from **CurrentApplication** for Analysis.
- Age is categorized into two types AGE_GROUP(NUMERICAL),AGE_SLOT(AGE CATEGORY)
- Categorize AMT_INCOME_TOTAL into 5 types based on certain conditions.
- Newly derived data set is divided into 2 data set based on Target variable

Below Analysis performed on the Newly derived data sets.

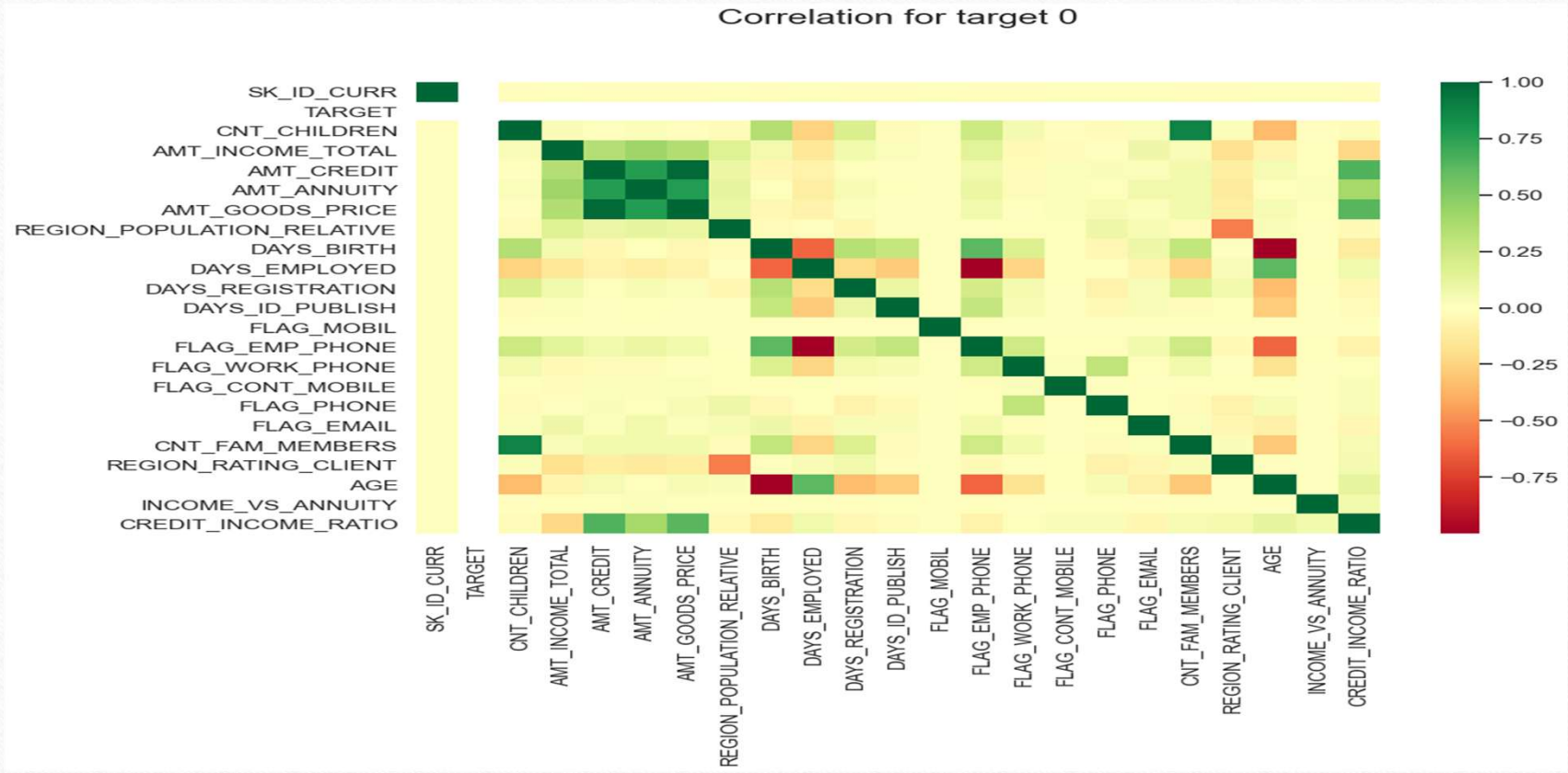
- **Univariate Categorical Analysis**
- **Univariate Continuous Analysis**
- **Correlation check in Analysis_df0 and Analysis_df1 dataset**
- **Bivariate Analysis of numerical variables**

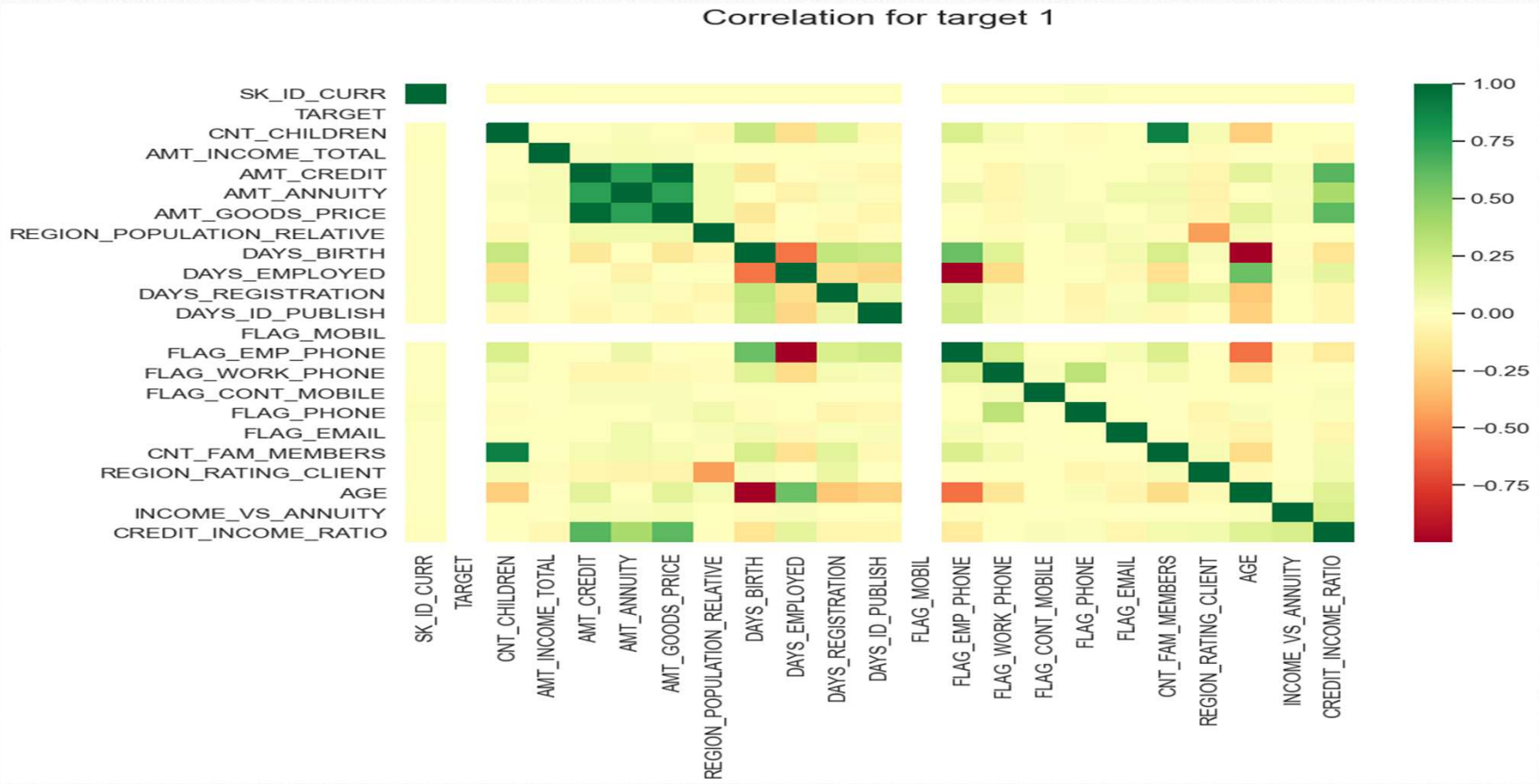




Inference:

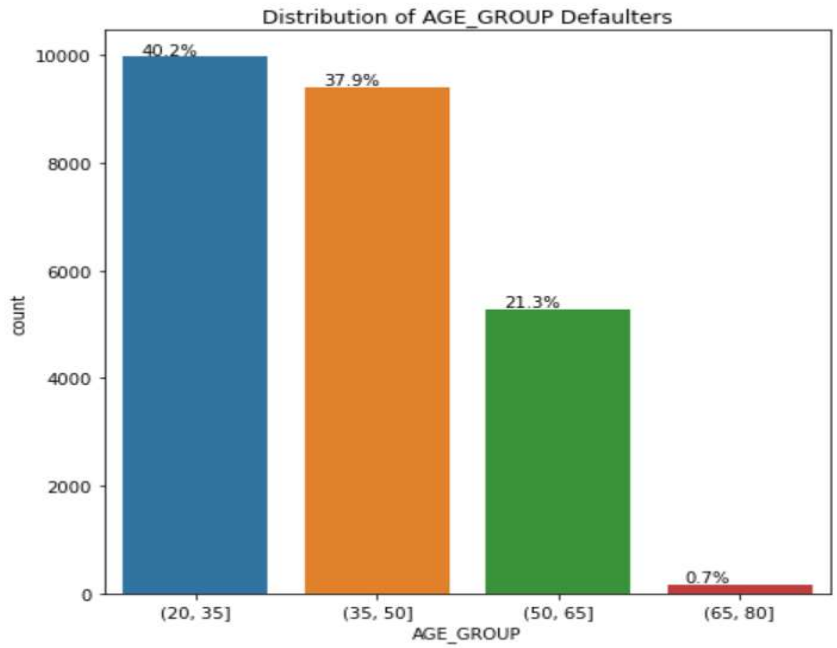
We can see that low child count maximizes that chances of both being a defaulter and also non defaulter. So we cannot conclude any specifics from this exploration.

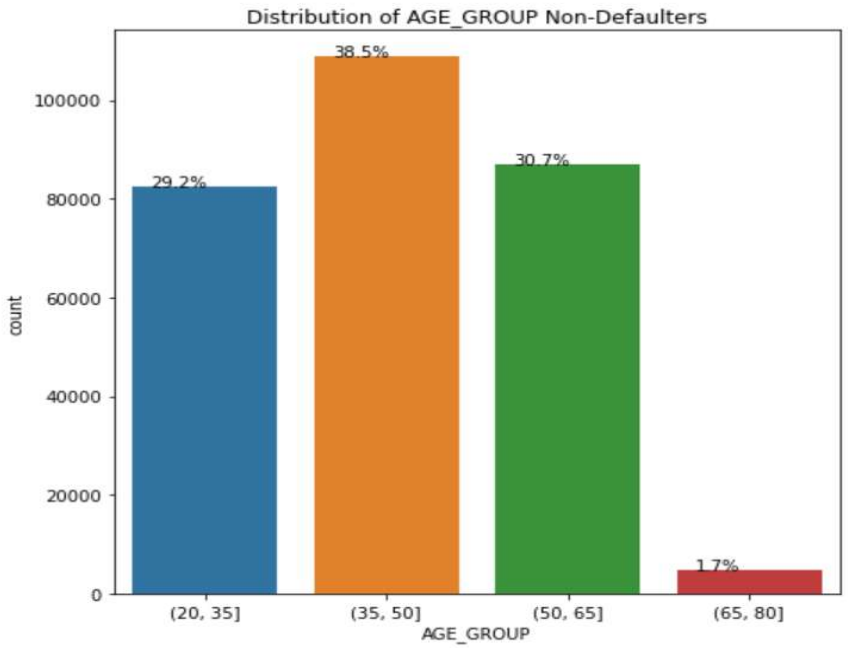

Correlation for target 0
SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
AGE
INCOME_VS_ANNUITY
CREDIT_INCOME_RATIO
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Correlation for target 1
SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
AGE
INCOME_VS_ANNUITY
CREDIT_INCOME_RATIO
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

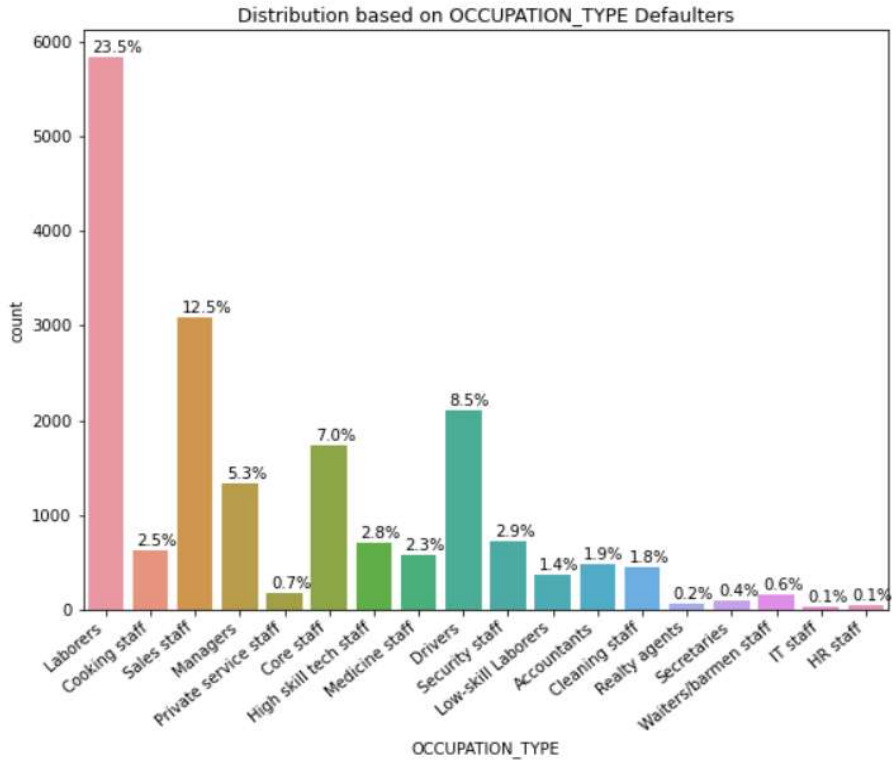
Below observations we can point out from Correlation heat map

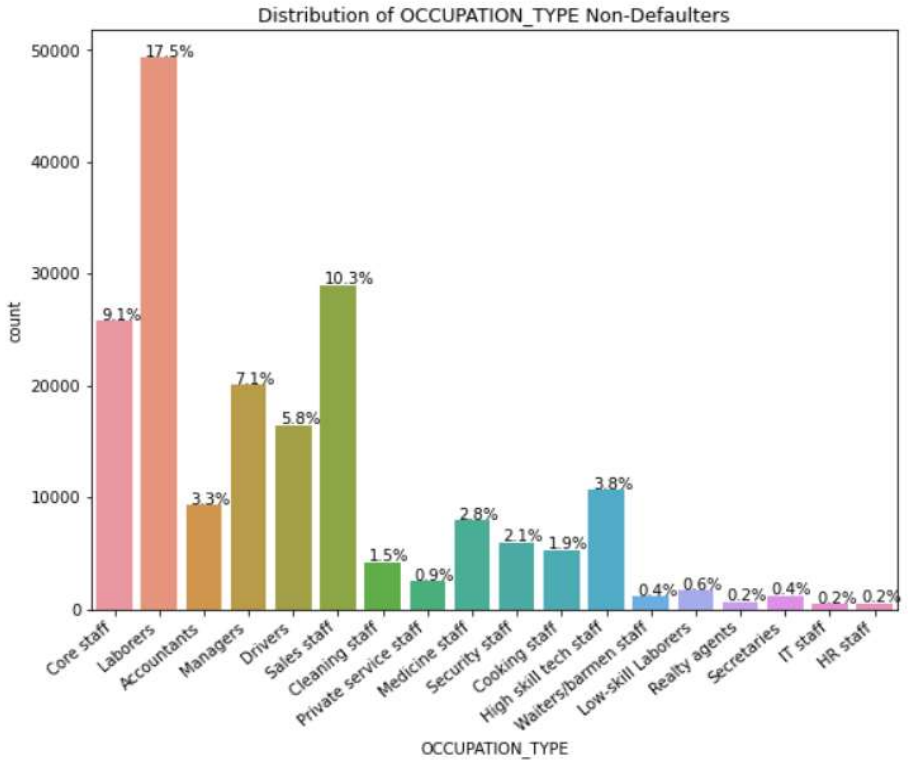
- 1-Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
- 2-Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.
- 3-Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.
- 4-Credit amount is higher to densely populated area.
- 5-The income is also higher in densely populated area.



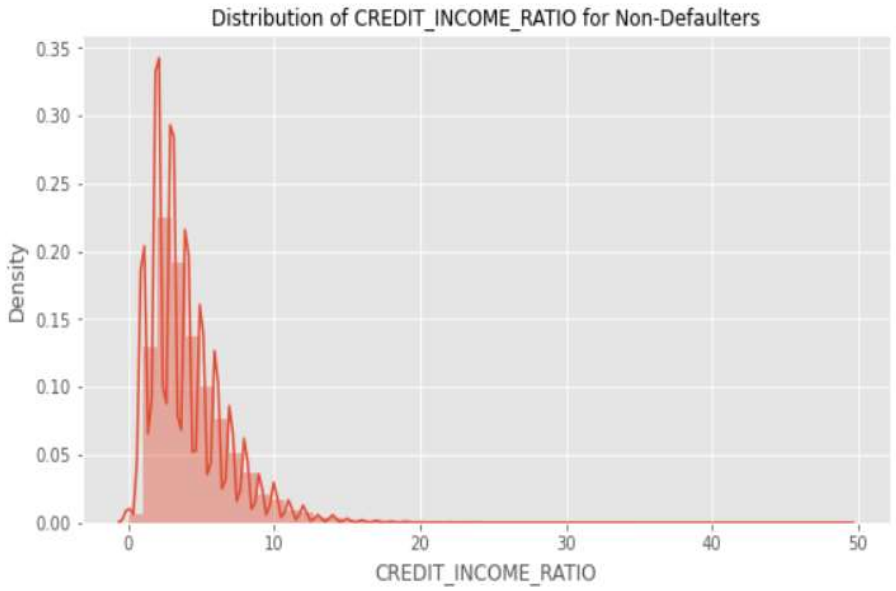


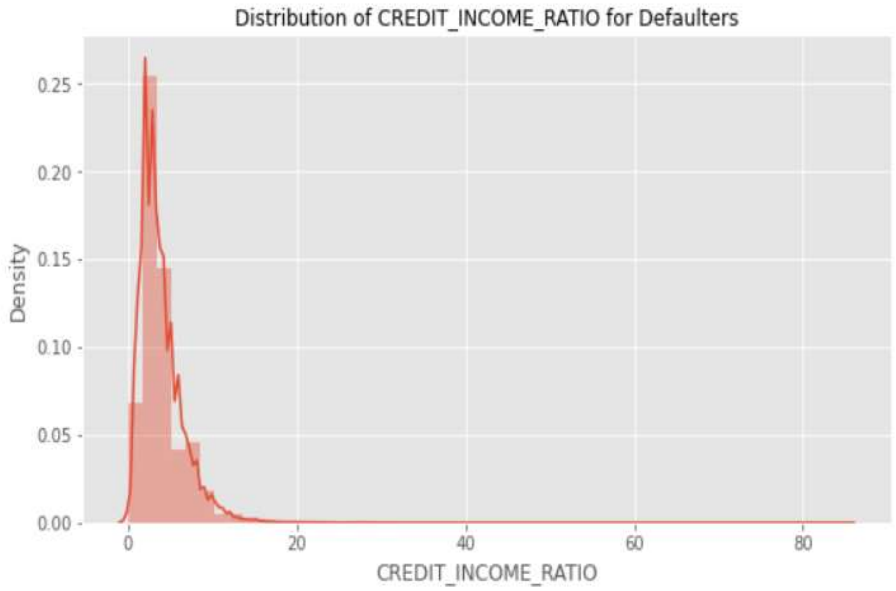
People aged between 20-35 have more number of defaulters.
People aged between 35-50 have more number of non-defaulters.





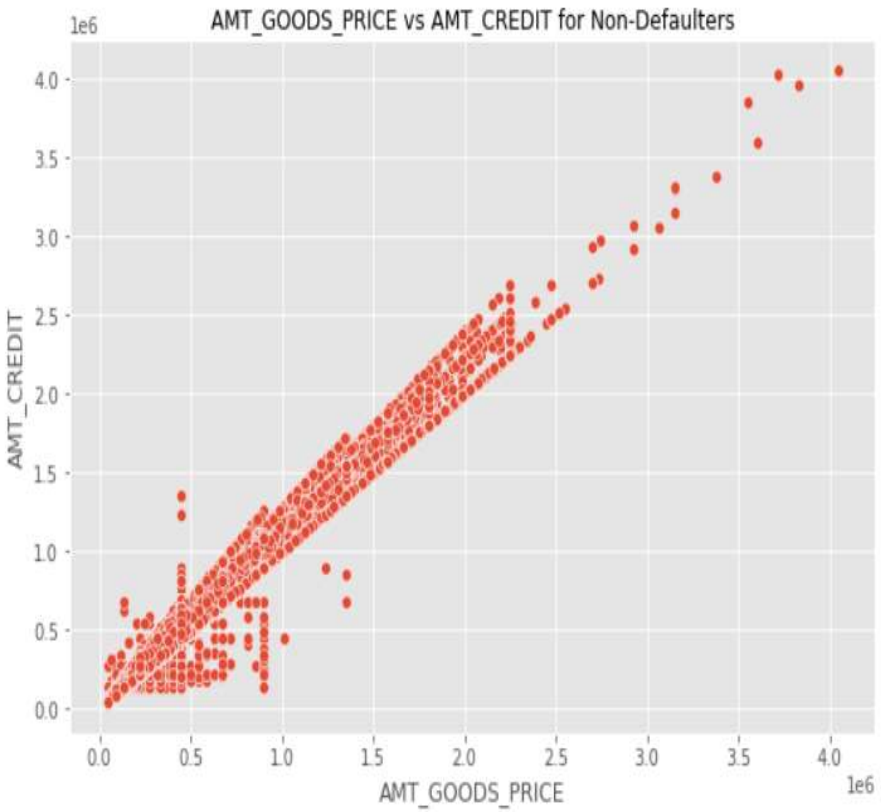
Maximum number of applications are from Laborers(Occupation_Type) in both defaulters and non-defaulters.

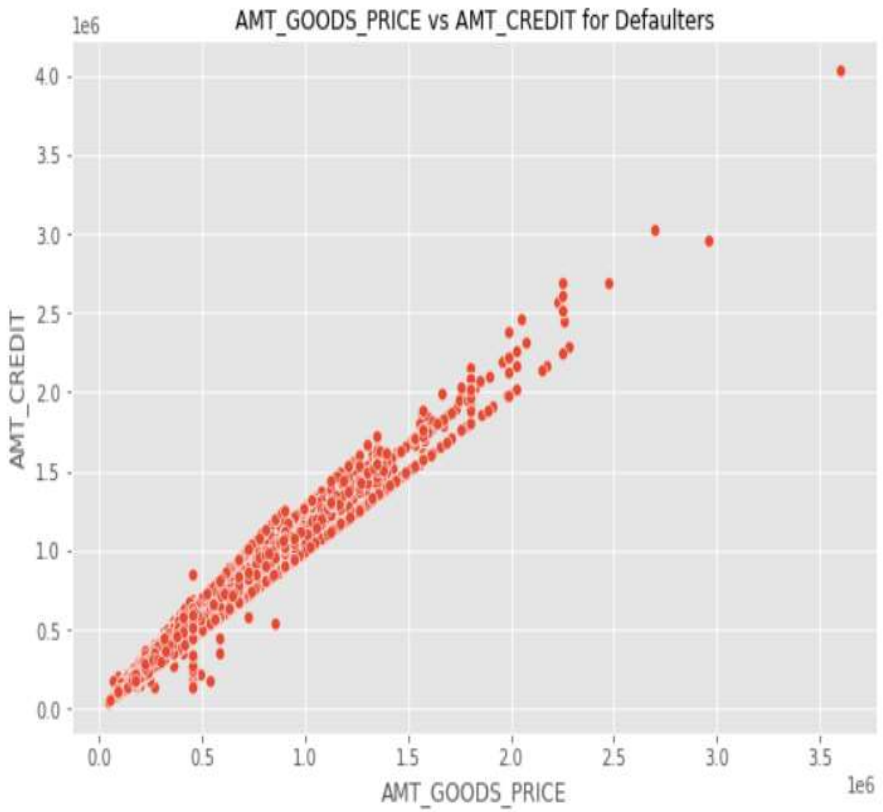




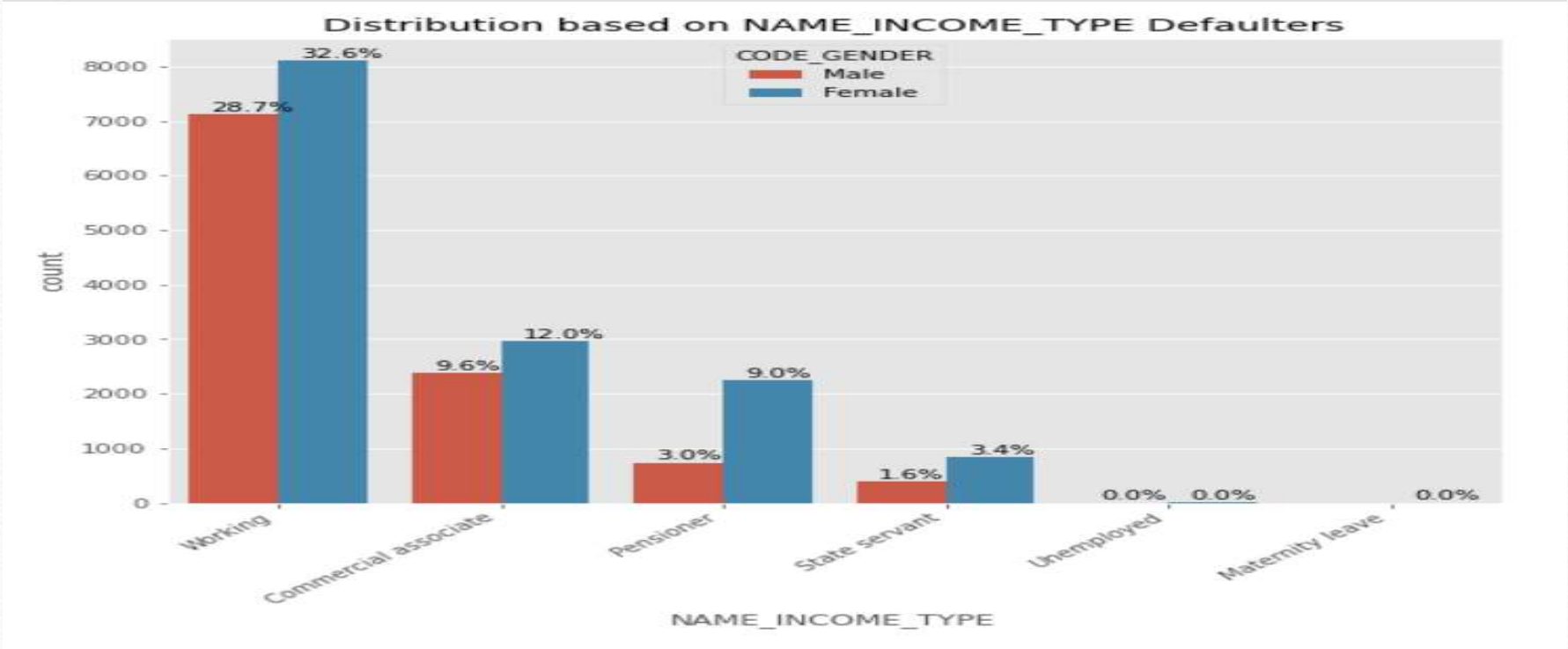
Inference:

we can see that the density of the credit income ratio is more than 50.





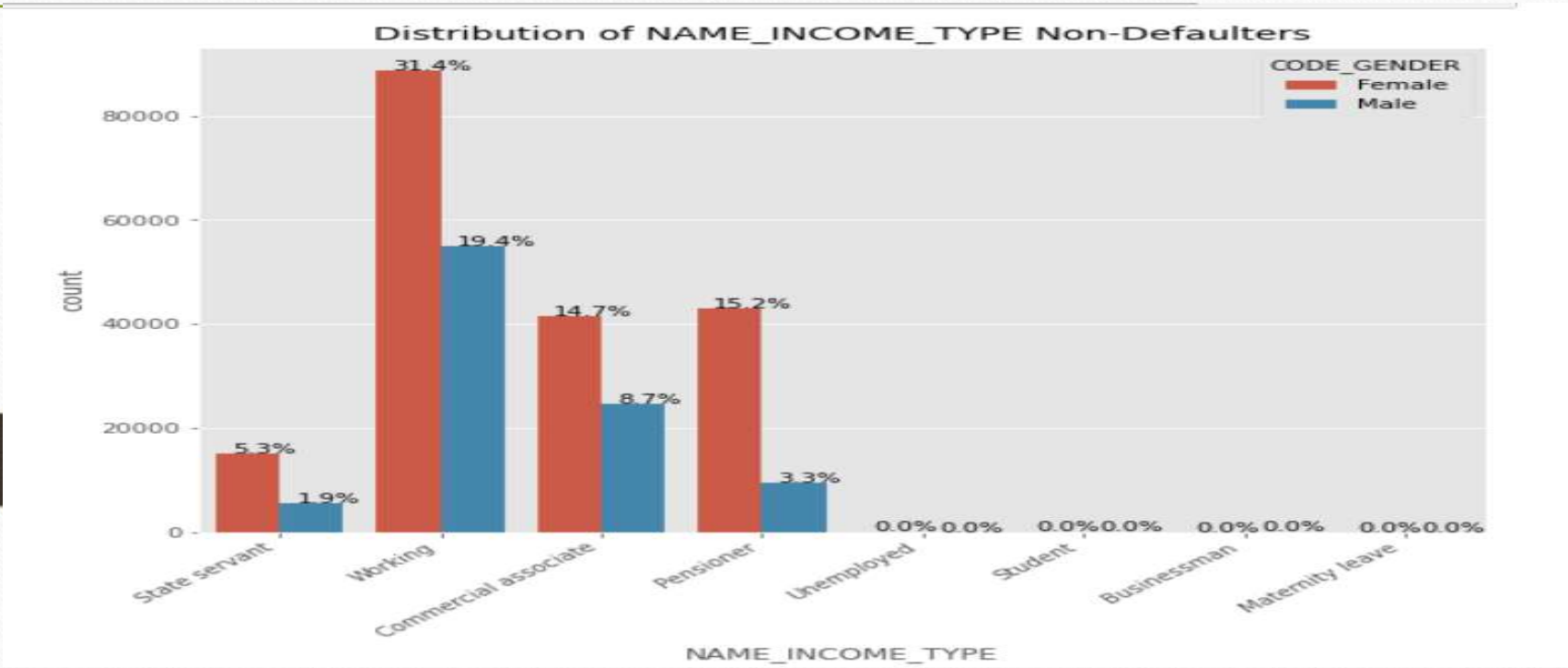
Amount of credit increases with amount of goods price.



Female are having more credit than males.

High number of credit for income type working , commercial associate , pensioner and state servant.

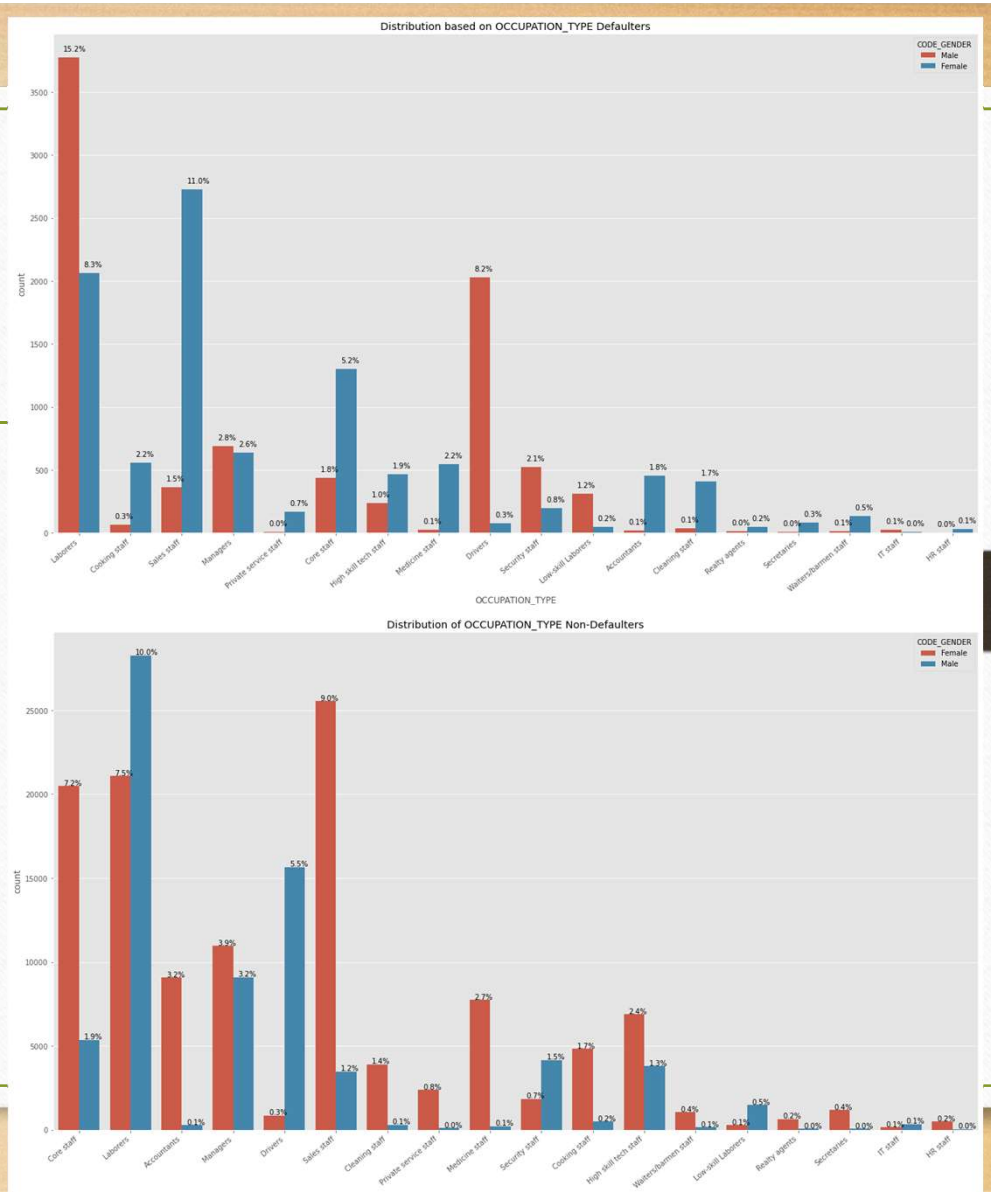
Low number of credit for income type student ,unemployed, businessman and maternity leave.



High number of credit for income type working , commercial associate , pensioner and state servant. Same as of target0

Low number of credit for income type unemployed and maternity leave

From the graph we can say that in both Defaulter and non-defaulters the males who applied for Credits are having occupation type as Laborers and most of the female who applied for credits have a occupation type as Sales Type.

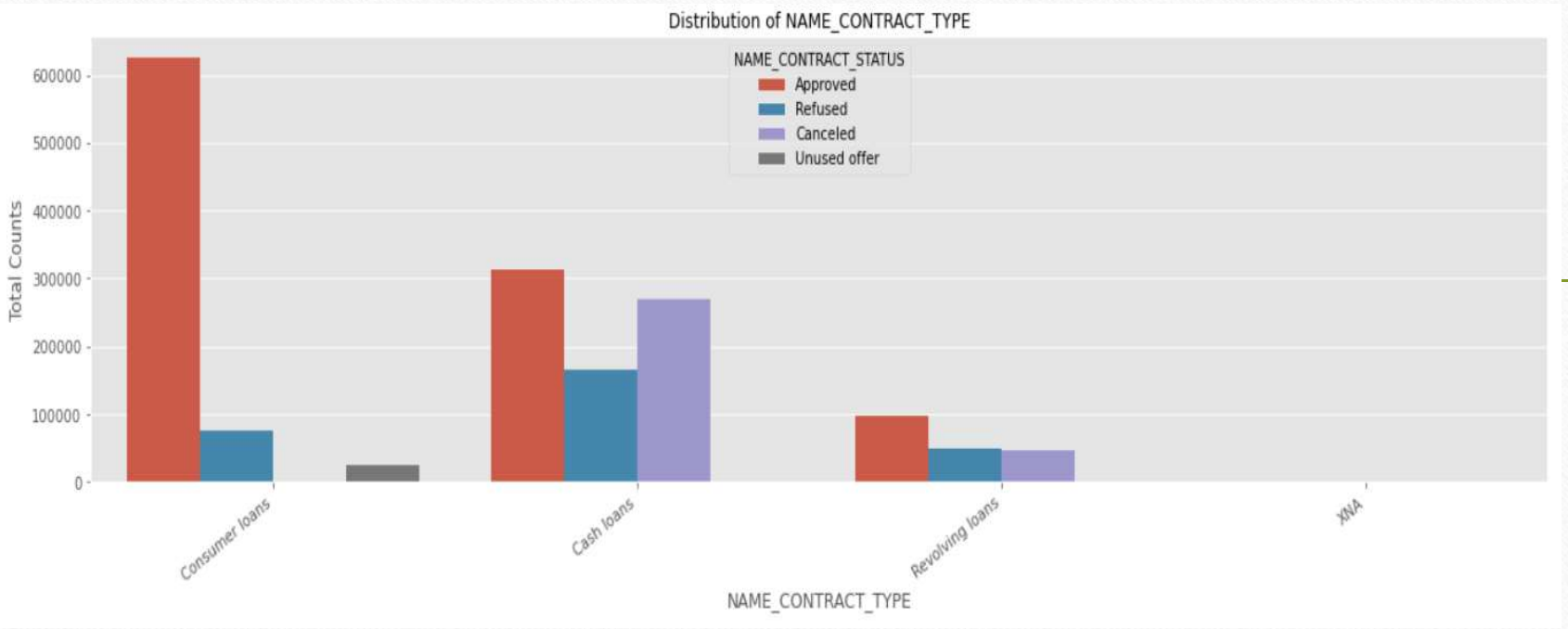


Dealing with Missing values on Previous Application :

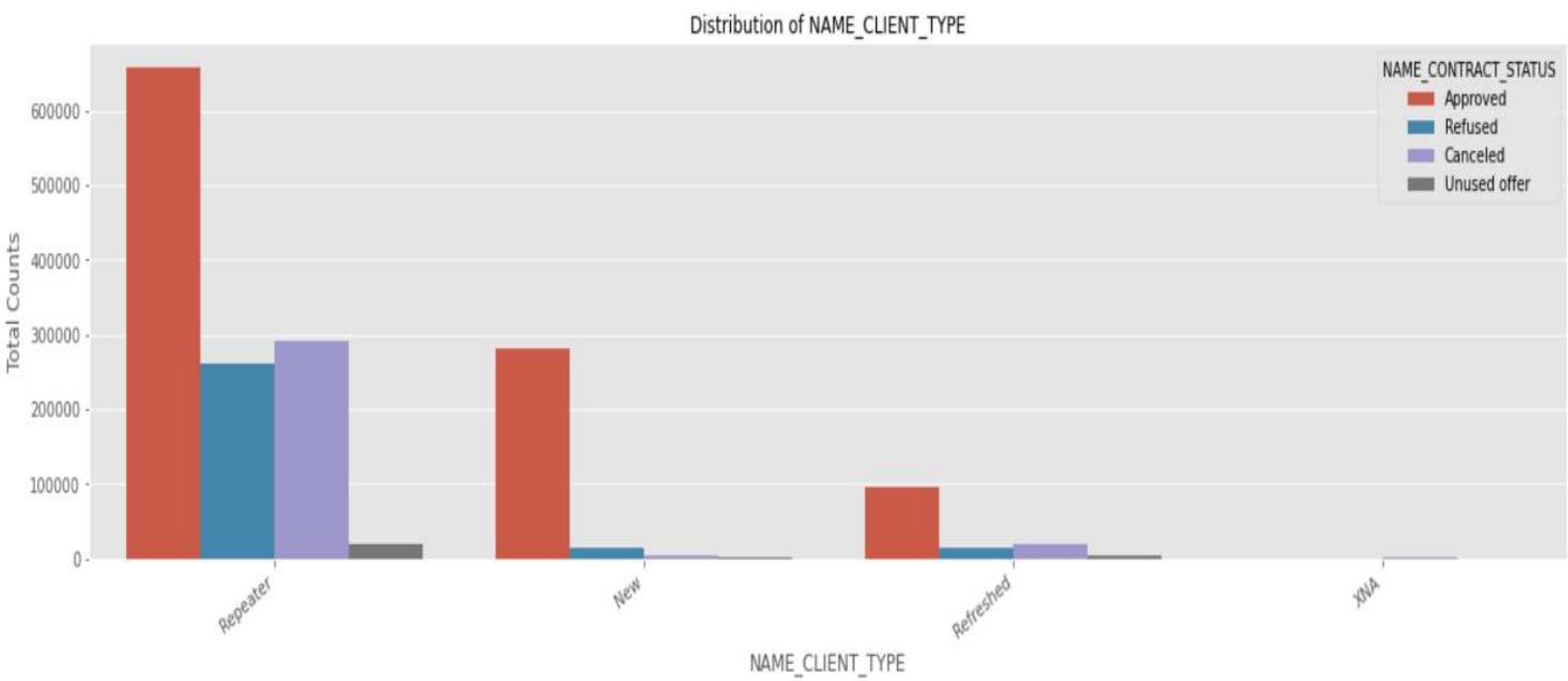
• Removed all the columns with more than 50% nulls values.

Below Analysis performed on the Previous Application :

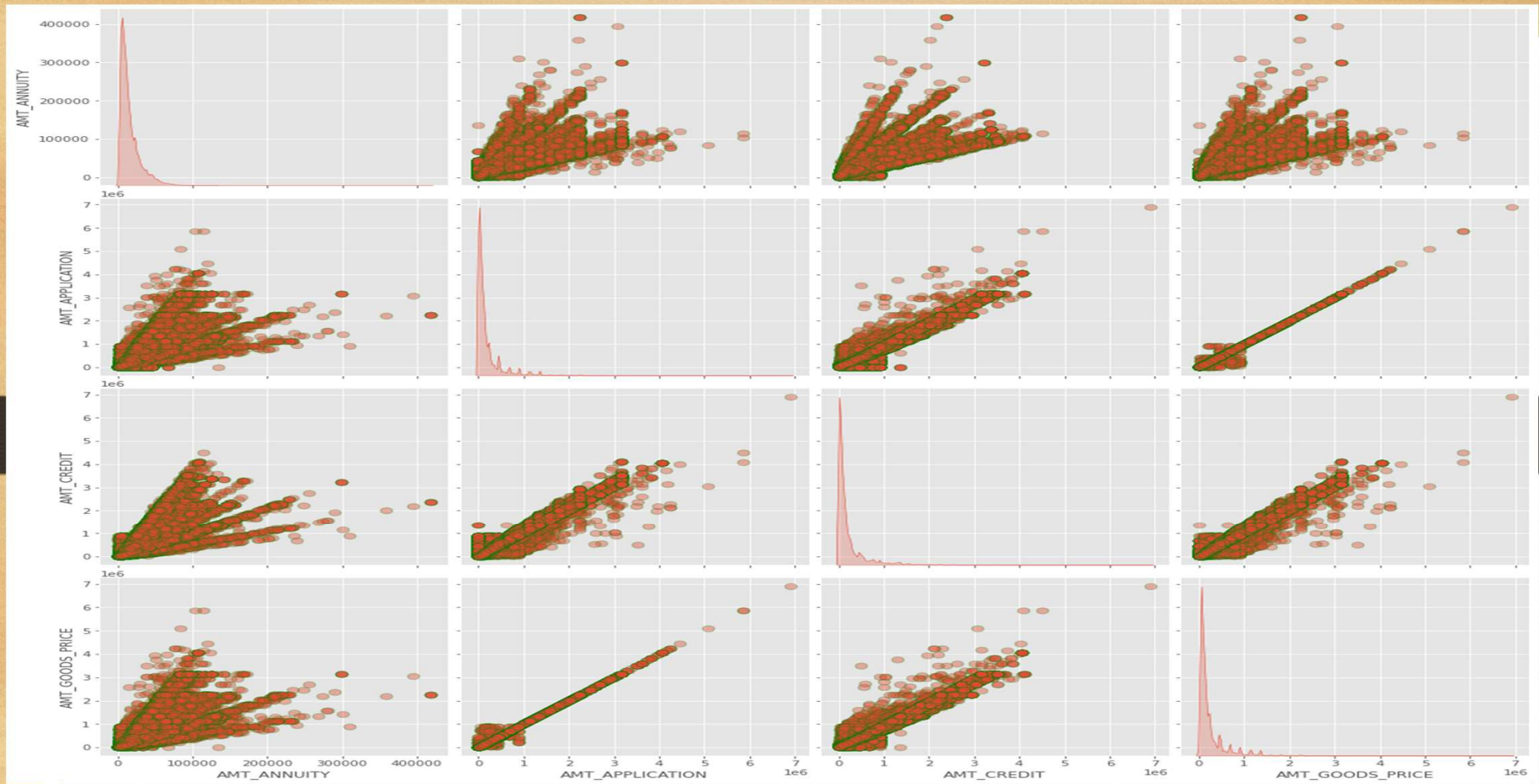
- Univariate Analysis
- Correlation check in Analysis_df0 and Analysis_df1 dataset
- Bivariate Analysis of numerical variables



From the graph we can conclude that most of the application is for Consumer loans and Cash loans, in which most of the cash loans got cancelled when compared to Consumer loans



Out of the total applications more than 60%-70% of customers are repeaters.
They also get refused most often.

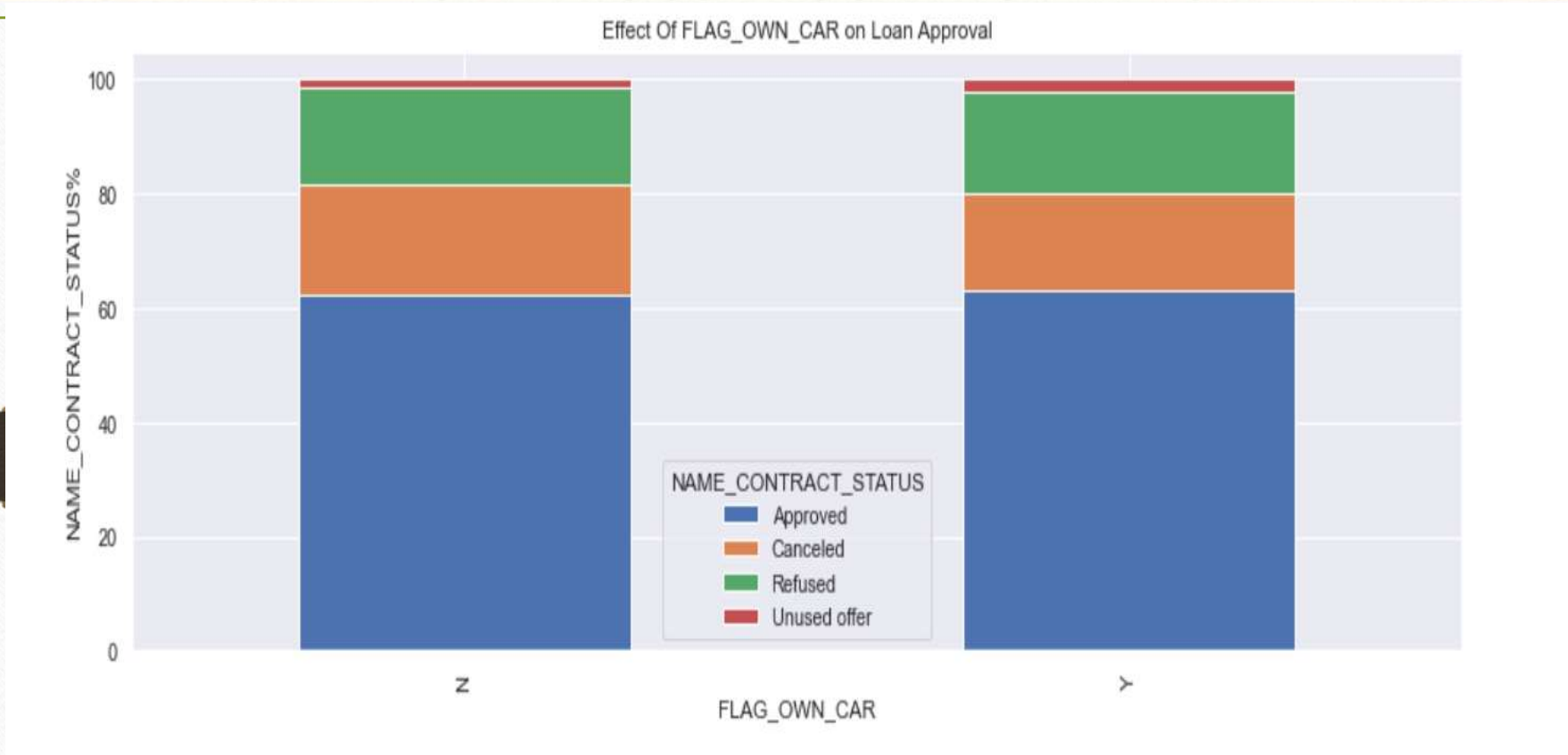

AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
1e6

The relation between highly correlated numeric variables.

- Annuity of previous application has a very high and positive supremacy on below components:
- How much credit did client asked on the previous application
- Final credit amount on the previous application that was approved by the bank
- Goods price of good that client asked for on the previous application

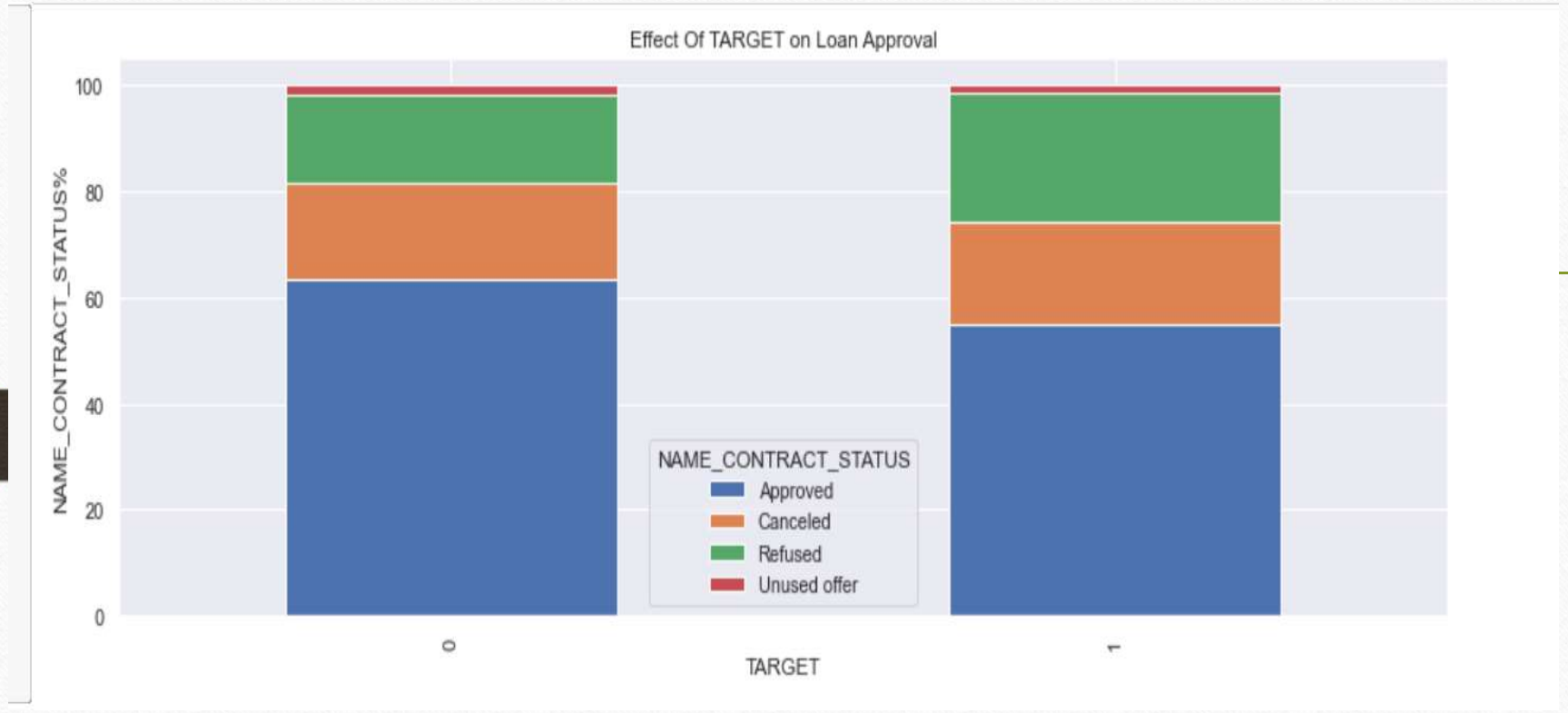
Final Data Set Preparation for Analysis

- A new Data set is formed by joining Analysis data set & Previous Application.
- Both data sets are joined on column “SK_ID_CURR”
- Bivariate Analysis performed on newly derived data set

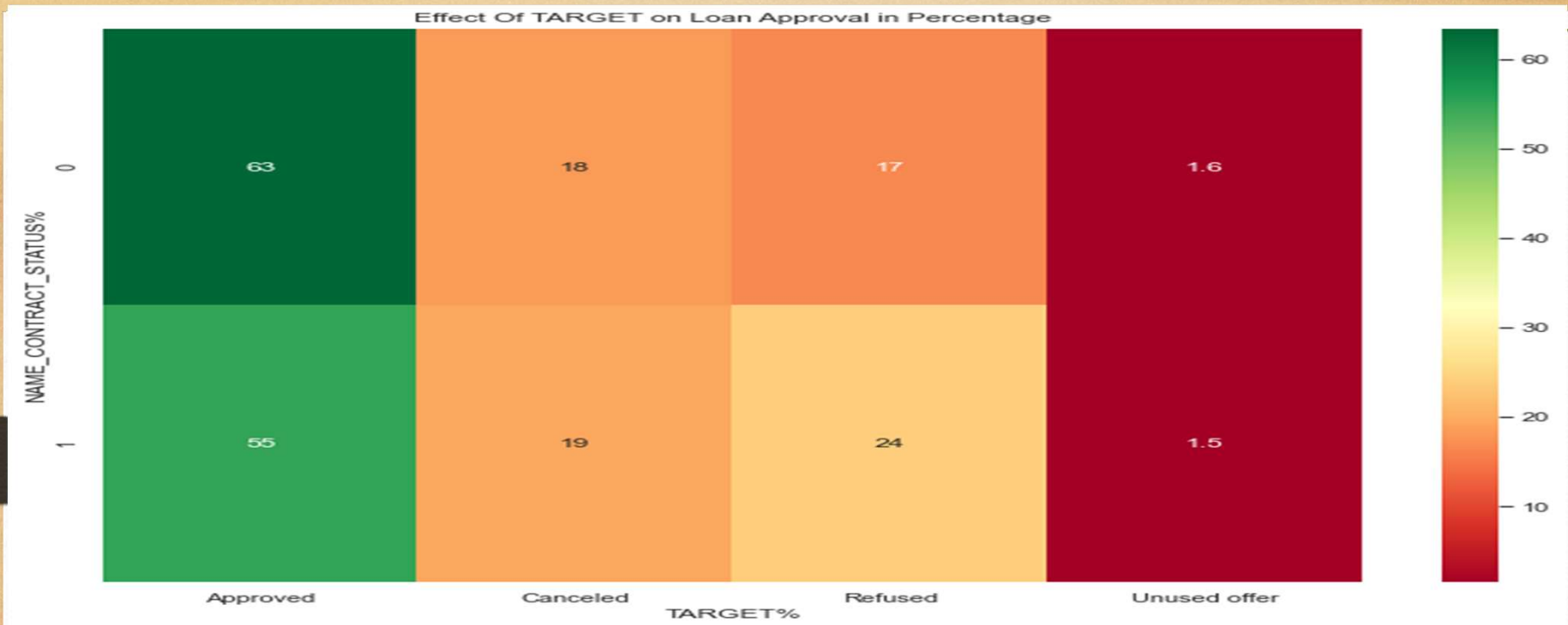


We can conclude that having a car also not affect the approval or rejection of the loan. But from the univariate graph we found having a car reduce the chance of getting defaulted.

We can't find any difference in the acceptance and rejection of loan based on Gender.
But from the above univariate graph we found female have less chance to become defaulters compare to male.

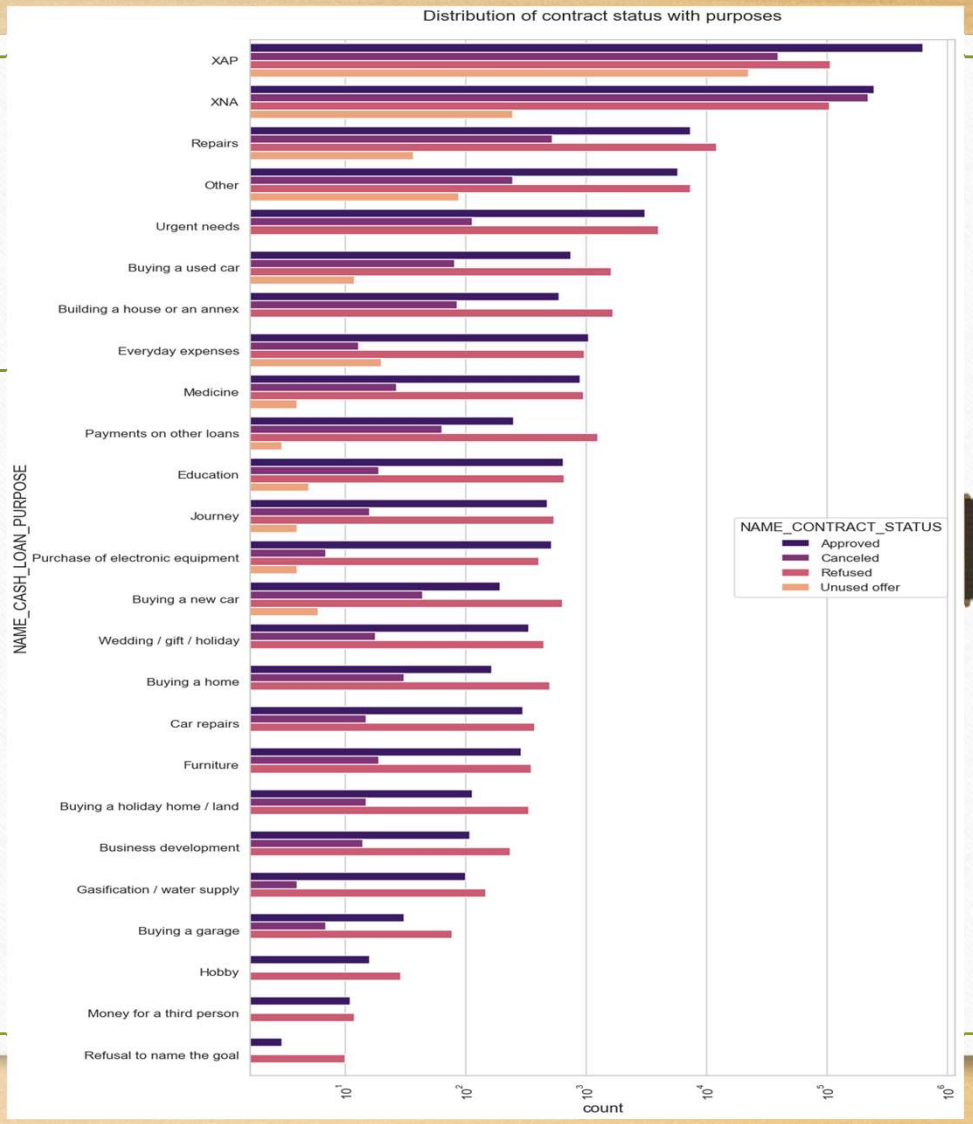


We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting



From the above heat map and bar graph with respect to Target value, we can conclude that:
non-defaulted people have more approved loans than defaulted people.
For Both the Defaulted and non-defaulted the percentage of people who cancelled and unused their loans in between the process seems to be same.
Most of the defaulted people got rejected to get the loan it shows they doesn't meet their requirement when we compare with non-defaulted people.

Most rejection of loans came from purpose 'repairs'. For education purposes we have equal number of approves and rejection Payign other loans and buying a new car is having significant higher rejection than approves.



CONCLUSION

- 1. Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- 2. Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- 3. Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.
- 4. Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.

Thank you