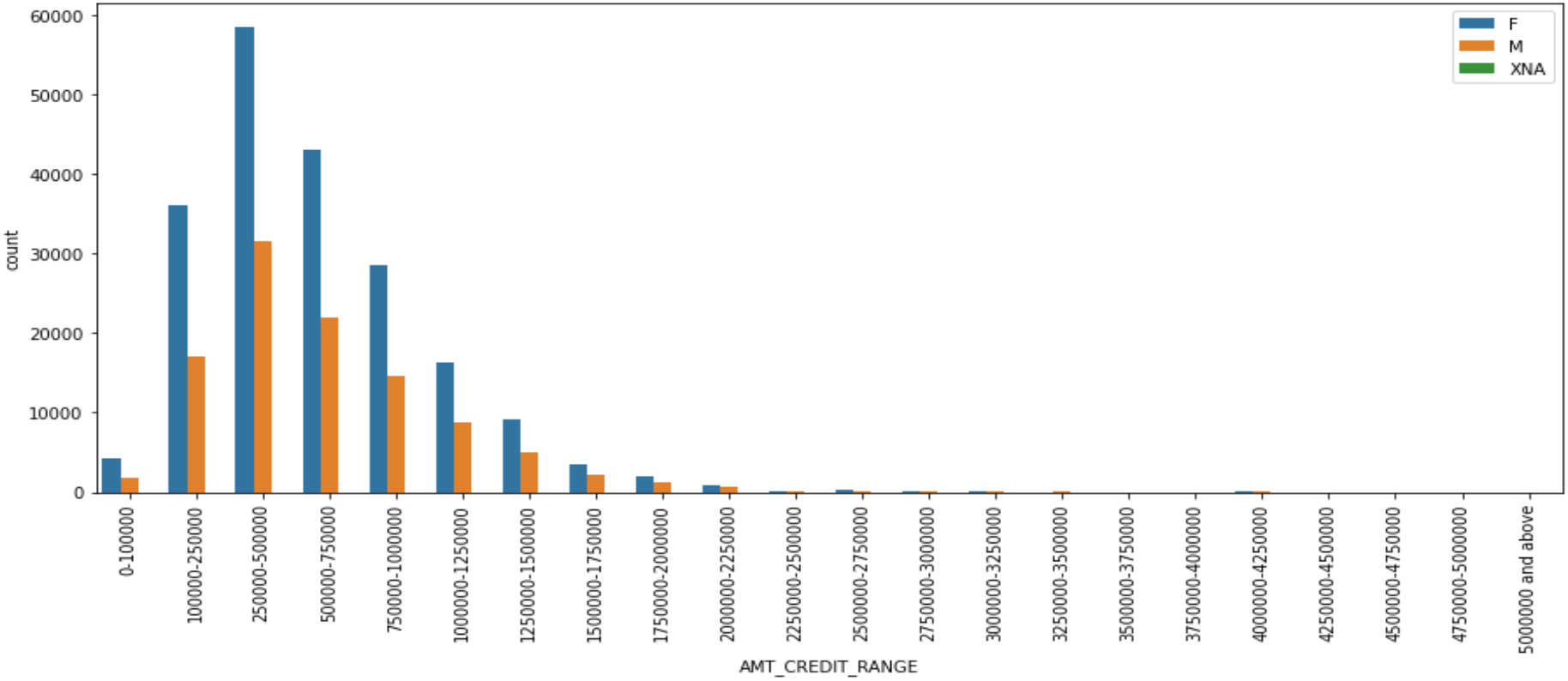
Assignment Solution

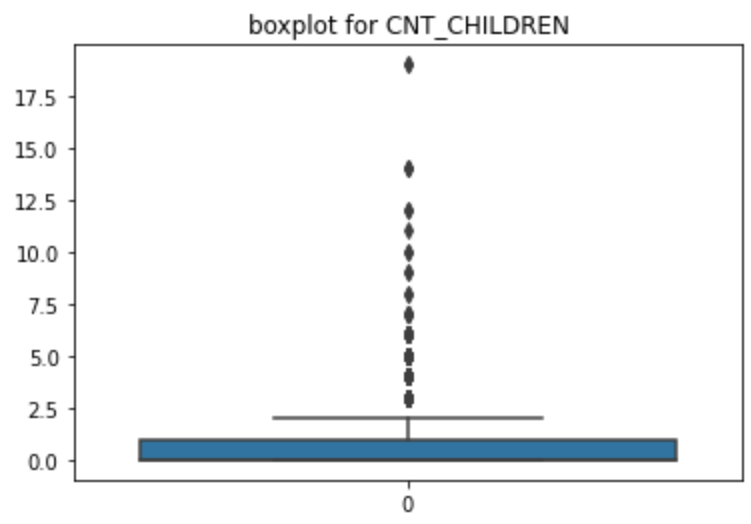
Steps Taken

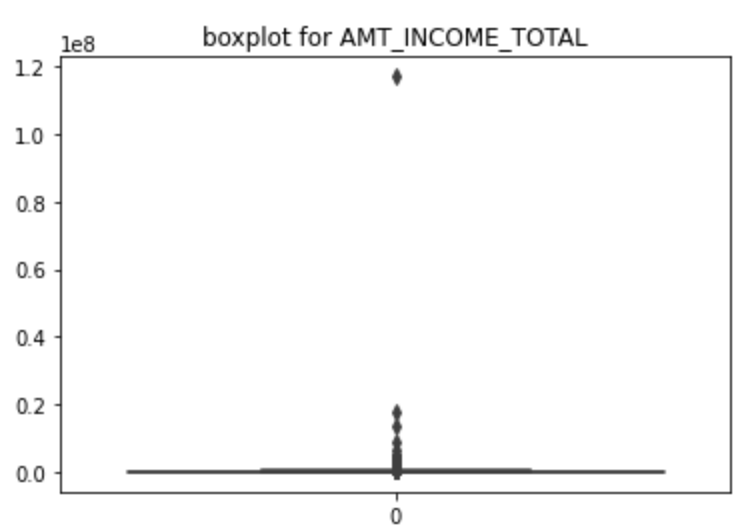
- Collect the source data from the assignment test(application_data.csv)
- Understand the columns meaning by reading columns_description.csv
- Checking the null values count on each column
- Filter the data without columns which have more than 40% of null values
- Drop the columns based on filter
- Checked the missing less values less than 15% and finding the datatypes
- Based on the dtypes, identified the columns are numerical values or categorical values

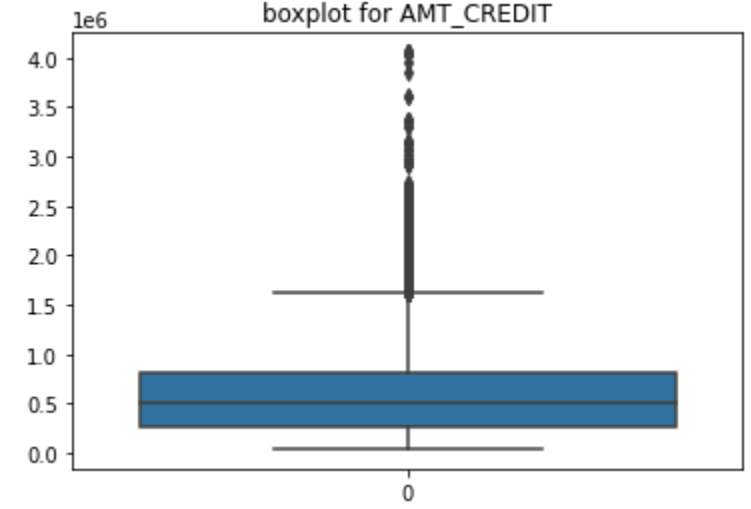
- Find the count of unique values in each column
- List down the unwanted columns by understanding the data and drop them from the source data
- Check the percentile values by using describe method
- Find out the list of categorical columns and changed their dtype to category
- SK_ID_CURR values is numeric type and it is unique for all the applicants so converted into object type
- Observed there are negative values in the data and changed to positive values by using abs() absolute function
- Changed the few columns names to understand better way
- Find the gender and change the values of XNA to Female

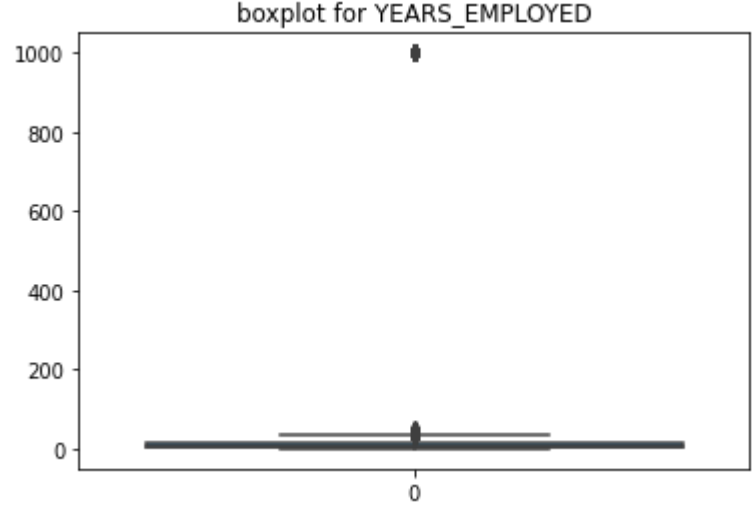


Plot between AMT_CREDIT & AMT_INCOME_TOTAL


boxplot for CNT_CHILDREN
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0

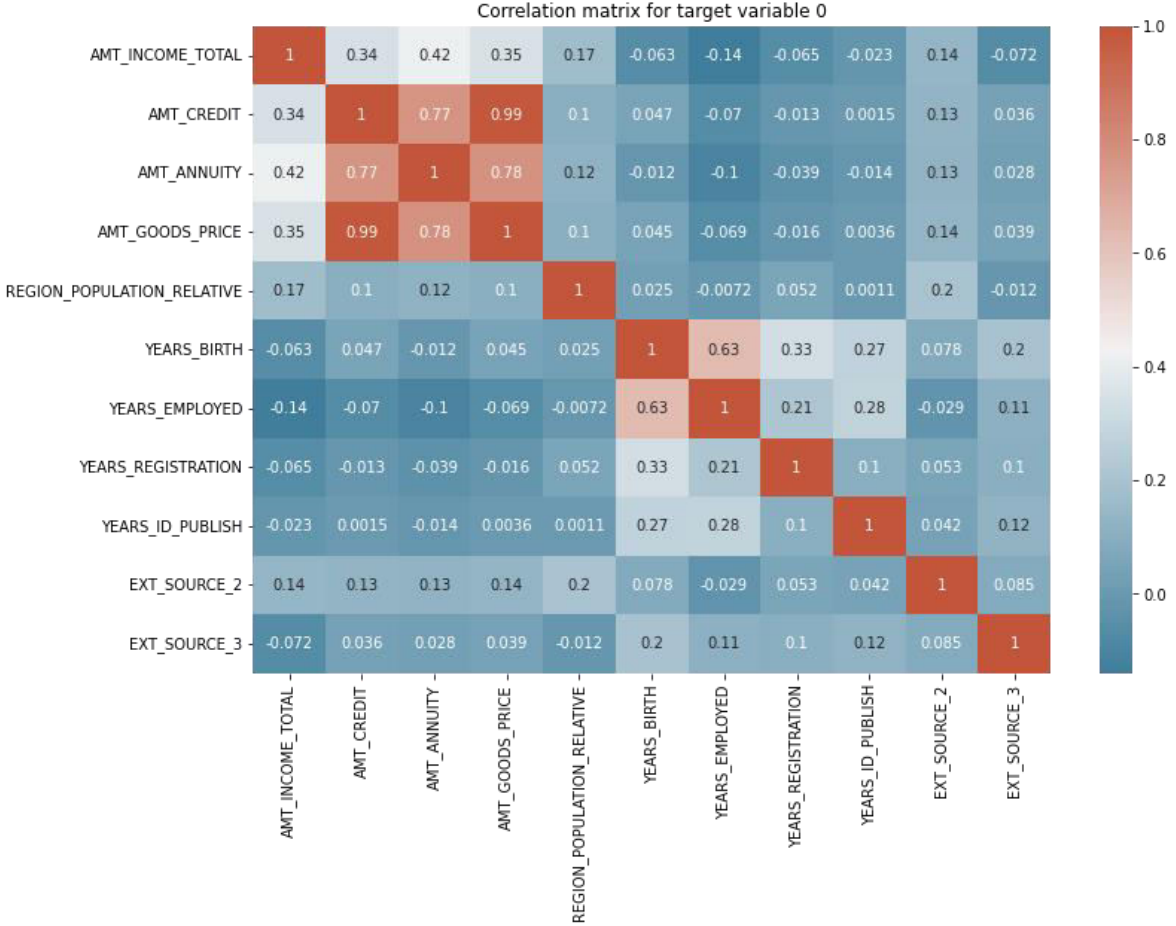

boxplot for AMT_INCOME_TOTAL
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0

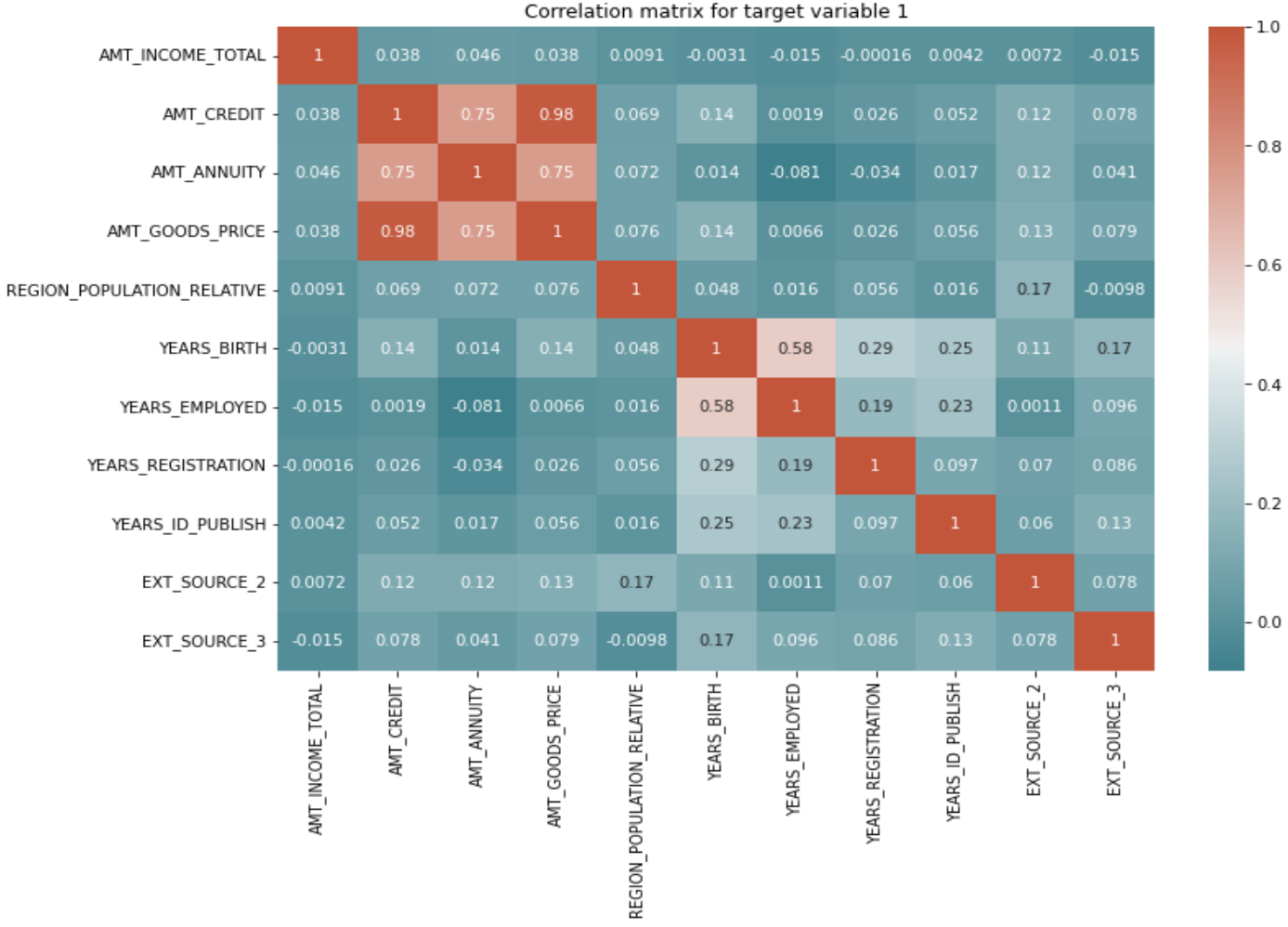

boxplot for AMT_CREDIT
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0


boxplot for YEARS_EMPLOYED
1000
800
600
400
200
0
0

Separated the final data based on Target Values

- Created appli_data_target0 dataframe for Target values zero(0)
- Created appli_data_target0 dataframe for Target values zero(0)

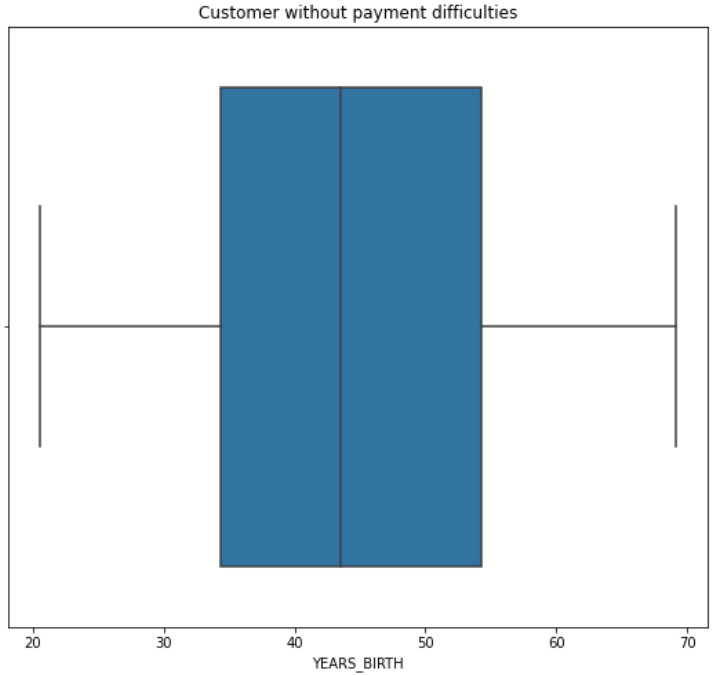


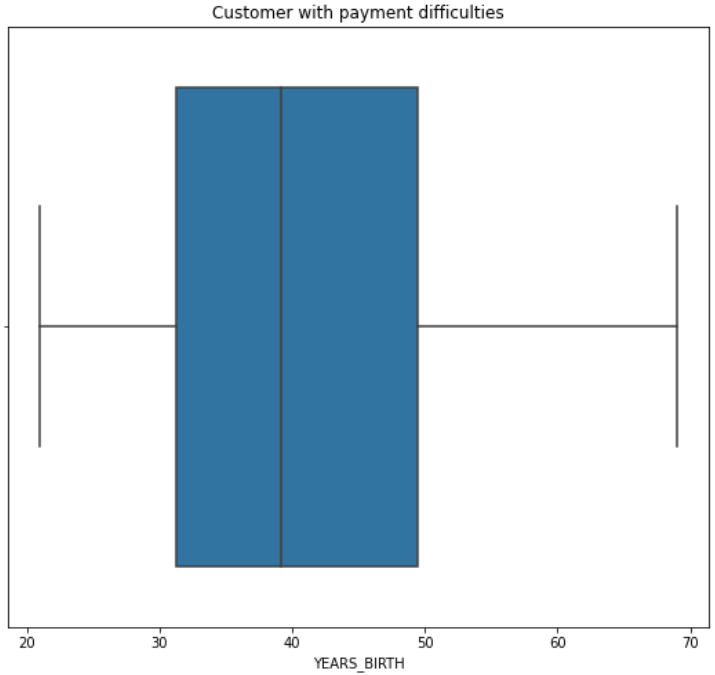

Correlation matrix for target variable 1
AMT_INCOME_TOTAL 1 0.038 0.046 0.038 0.0091 -0.0031 -0.015 -0.00016 0.0042 0.0072 -0.015
AMT_CREDIT 0.038 1 0.75 0.98 0.069 0.14 0.0019 0.026 0.052 0.12 0.078
AMT_ANNUITY 0.046 0.75 1 0.75 0.072 0.014 -0.081 -0.034 0.017 0.12 0.041
AMT_GOODS_PRICE 0.038 0.98 0.75 1 0.076 0.14 0.0066 0.026 0.056 0.13 0.079
REGION_POPULATION_RELATIVE 0.0091 0.069 0.072 0.076 1 0.048 0.016 0.056 0.016 0.17 -0.0098
YEARS_BIRTH -0.0031 0.14 0.014 0.14 0.048 1 0.58 0.29 0.25 0.11 0.17
YEARS_EMPLOYED -0.015 0.0019 -0.081 0.0066 0.016 0.58 1 0.19 0.23 0.0011 0.096
YEARS_REGISTRATION -0.00016 0.026 -0.034 0.026 0.056 0.29 0.19 1 0.097 0.07 0.086
YEARS_ID_PUBLISH 0.0042 0.052 0.017 0.056 0.016 0.25 0.23 0.097 1 0.06 0.13
EXT_SOURCE_2 0.0072 0.12 0.12 0.13 0.17 0.11 0.0011 0.07 0.06 1 0.078
EXT_SOURCE_3 -0.015 0.078 0.041 0.079 -0.0098 0.17 0.096 0.086 0.13 0.078 1
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
YEARS_BIRTH
YEARS_EMPLOYED
YEARS_REGISTRATION
YEARS_ID_PUBLISH
EXT_SOURCE_2
EXT_SOURCE_3
1.0
0.8
0.6
0.4
0.2
0.0

Checking the Univariate Analysis

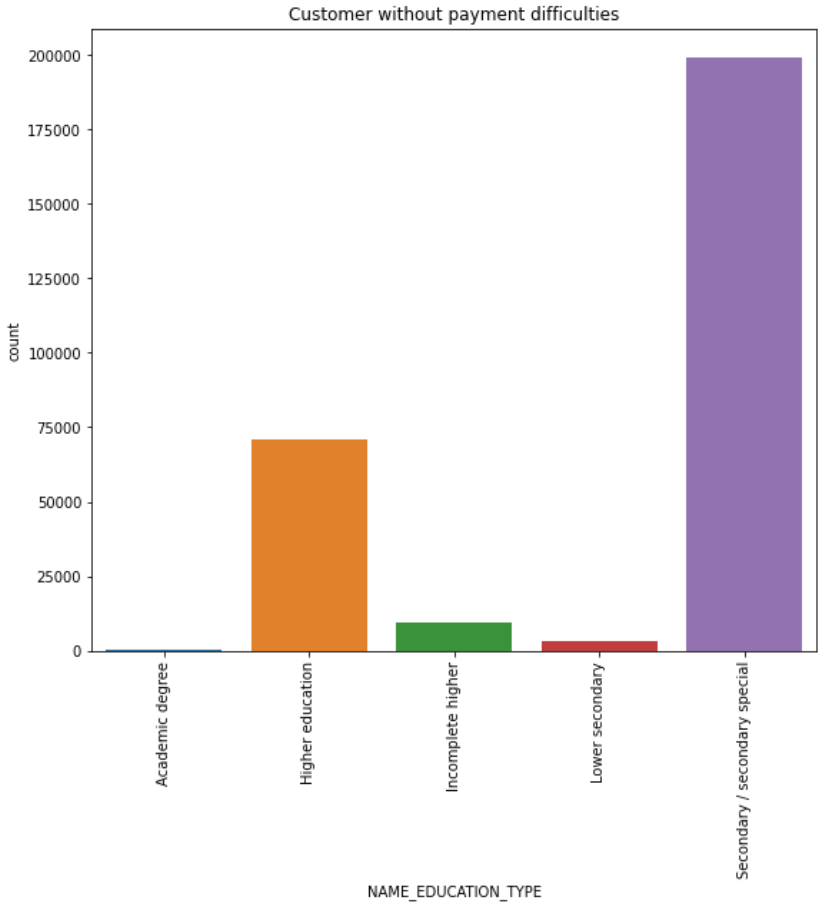
Numerical Variables

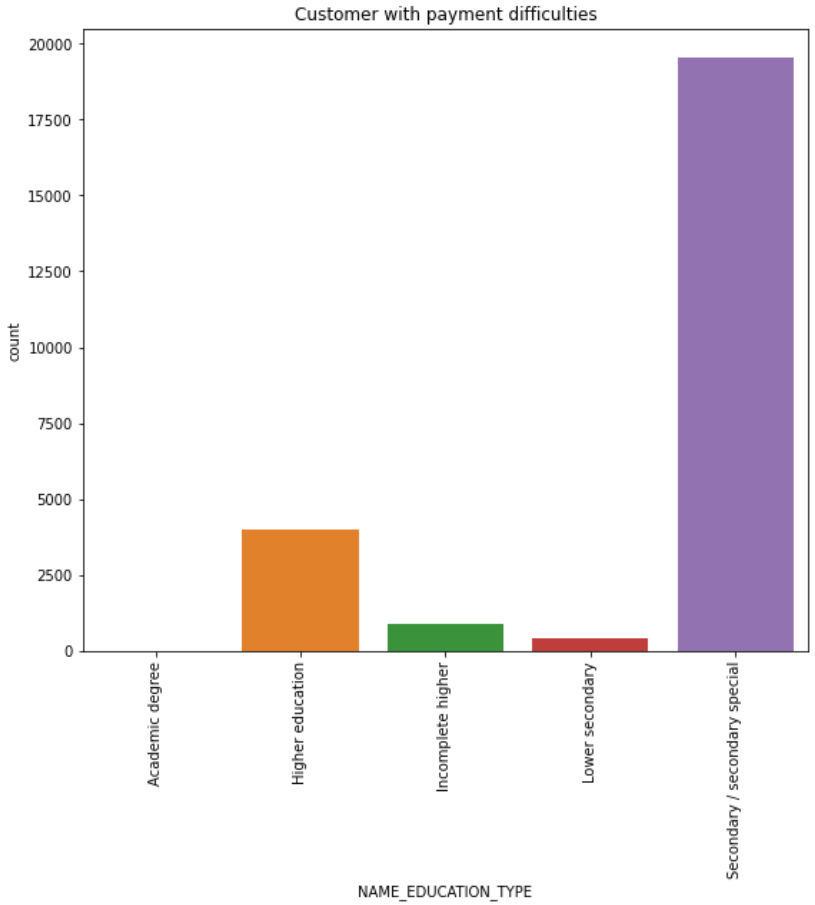
Comparing the data between Target values 0 & 1



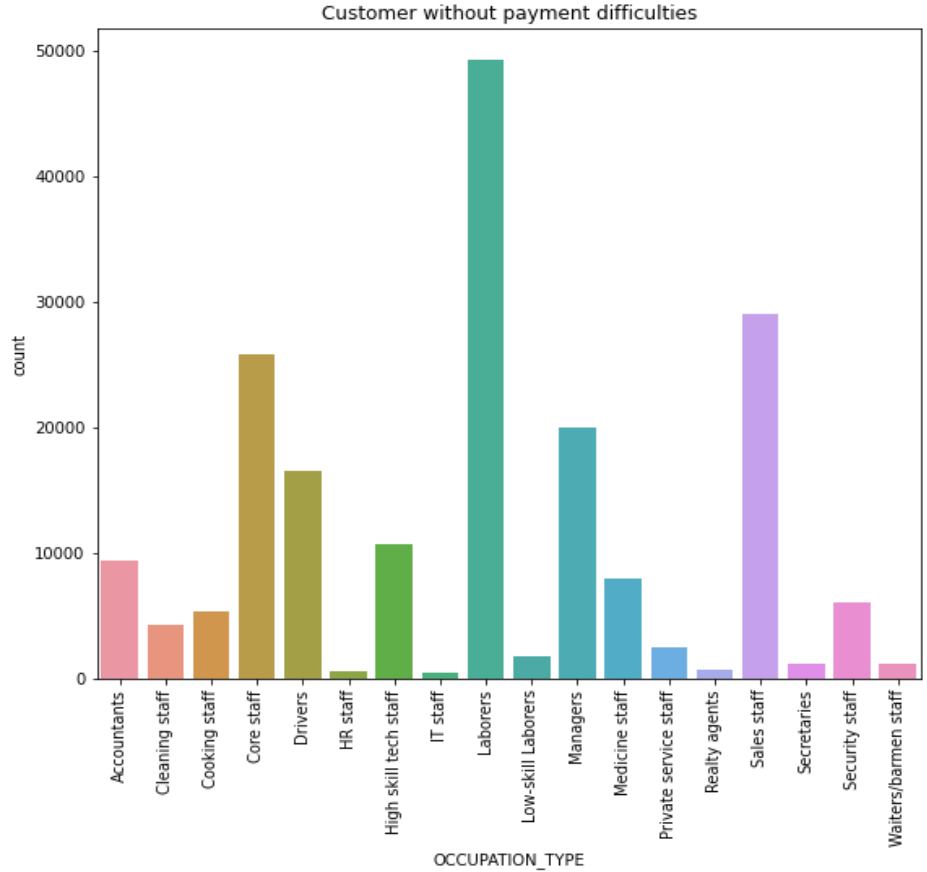


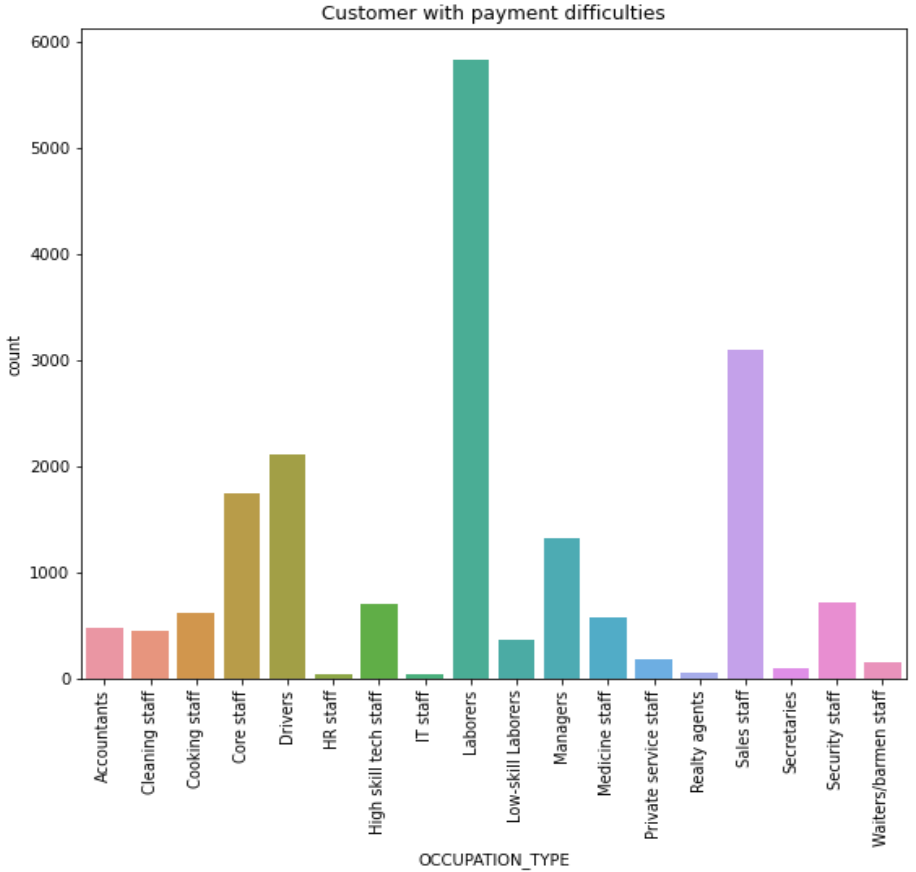
observe that customers have payment difficulty in secondary/secondary special both the cases



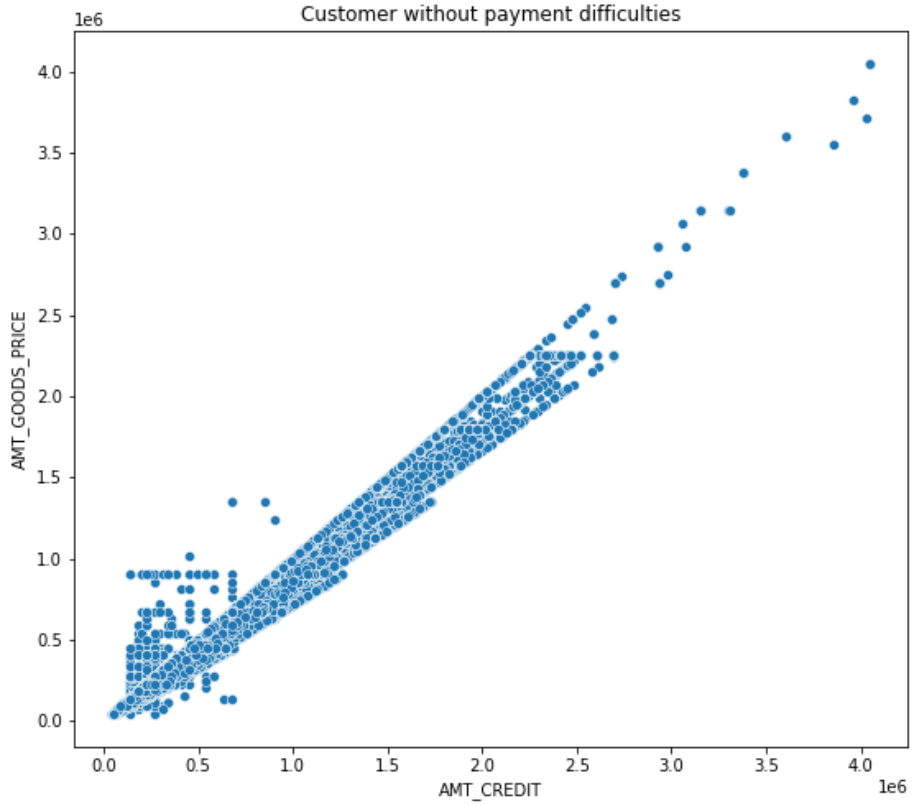


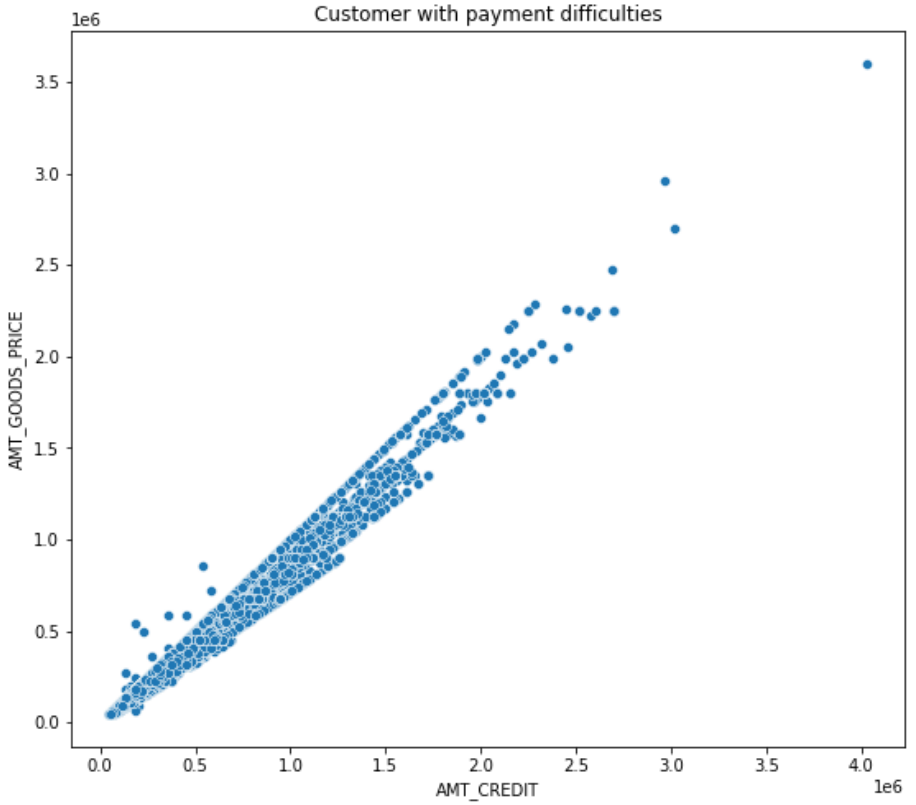
you can observe that Labours have more difficulties to repaying the loan and also find CoreStaff, SalesStaff same. closely observe that in Labours have hight without payment rate compared to with payment.



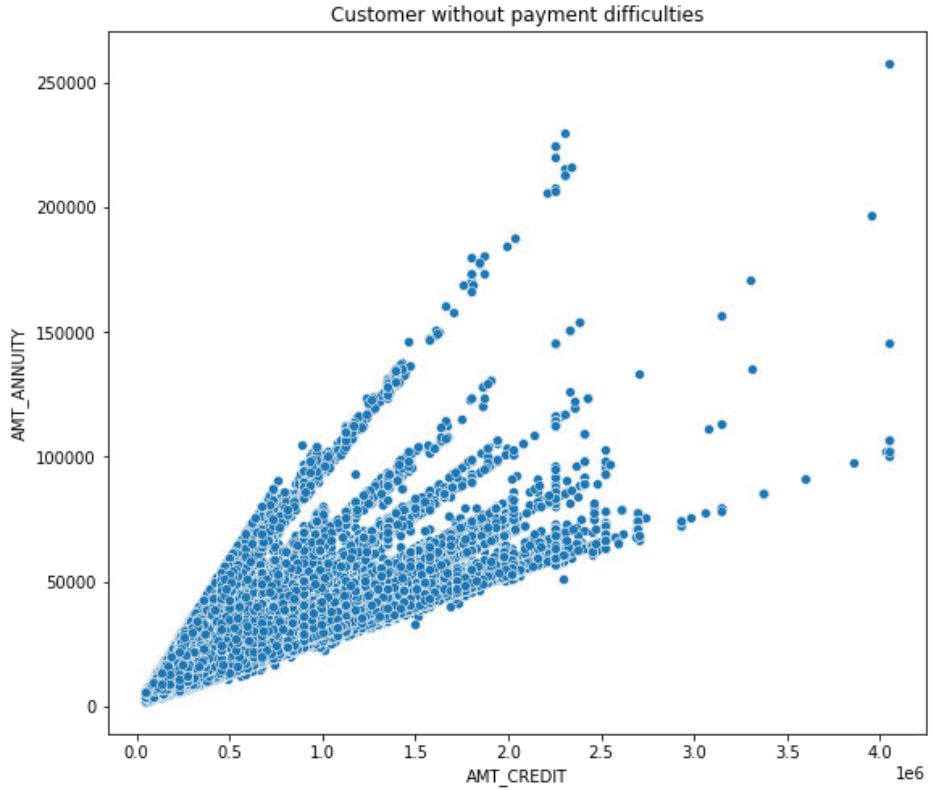


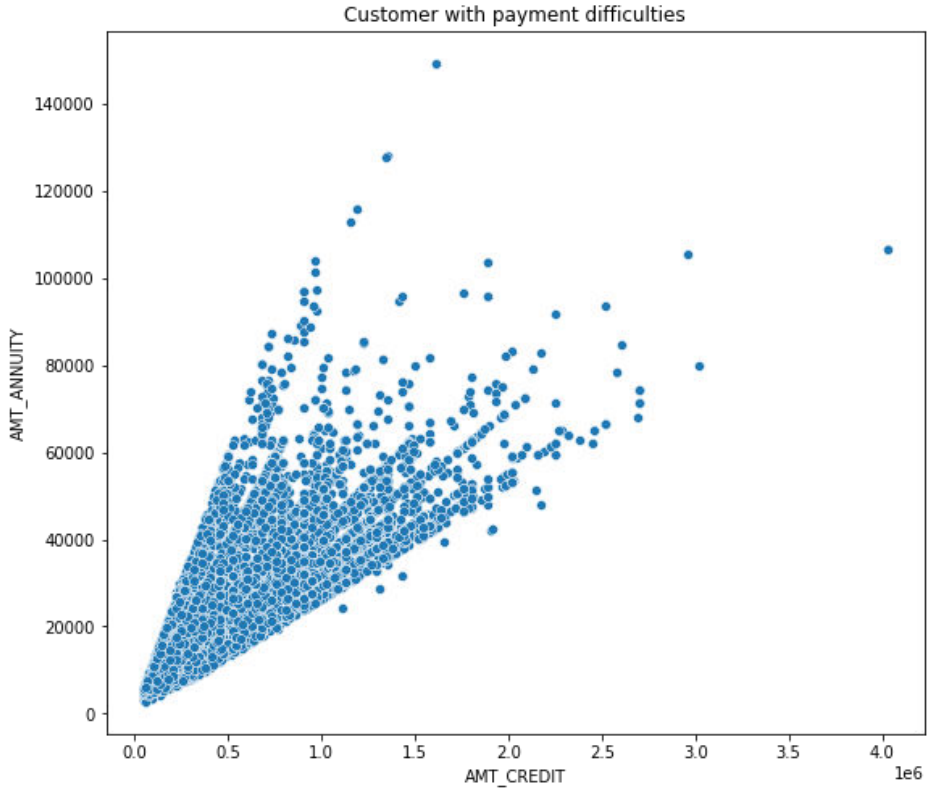
Here we have Good price is positively correlated with Credit Amount



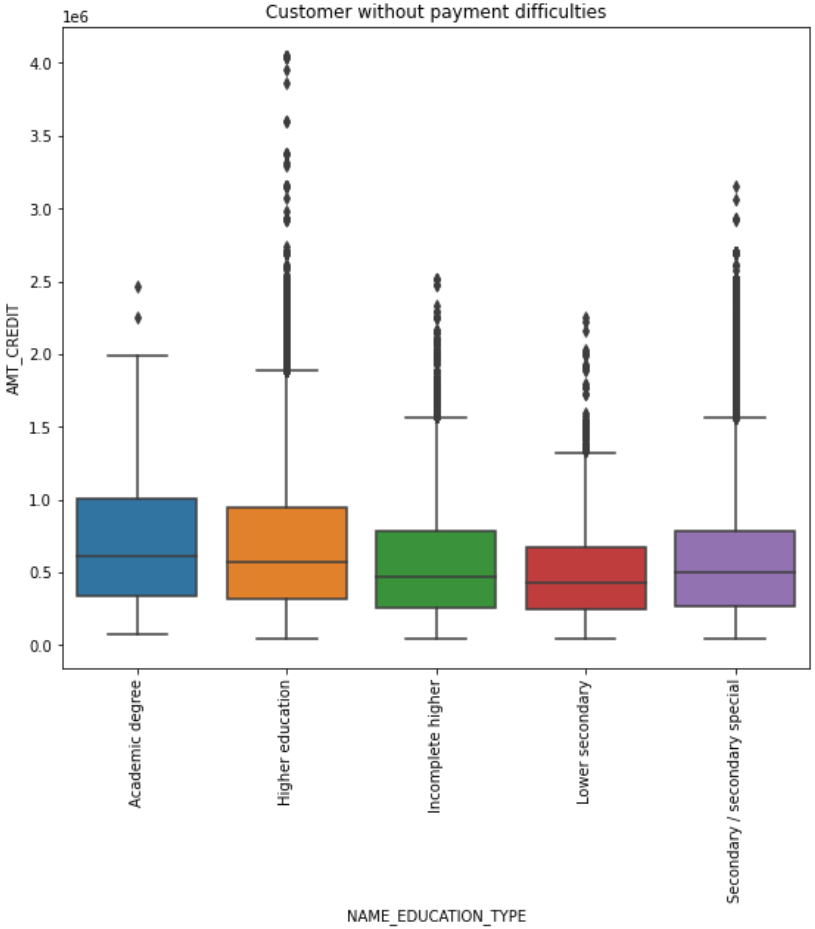


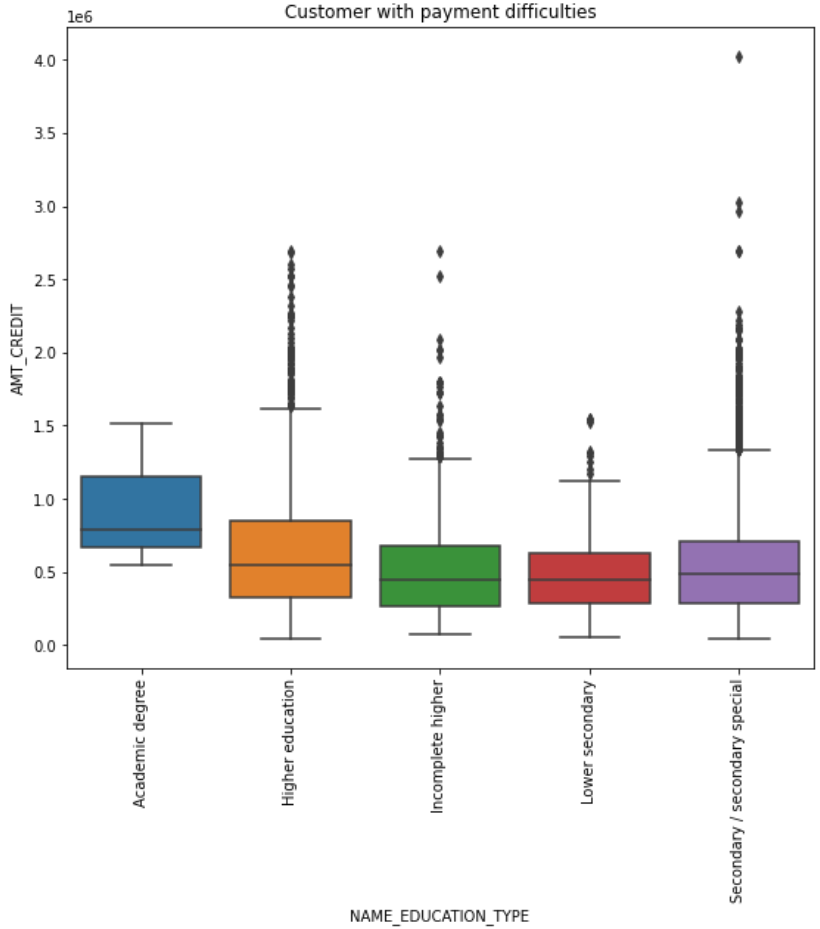
Observe that without payment difficulties take more values than with payment difficulties



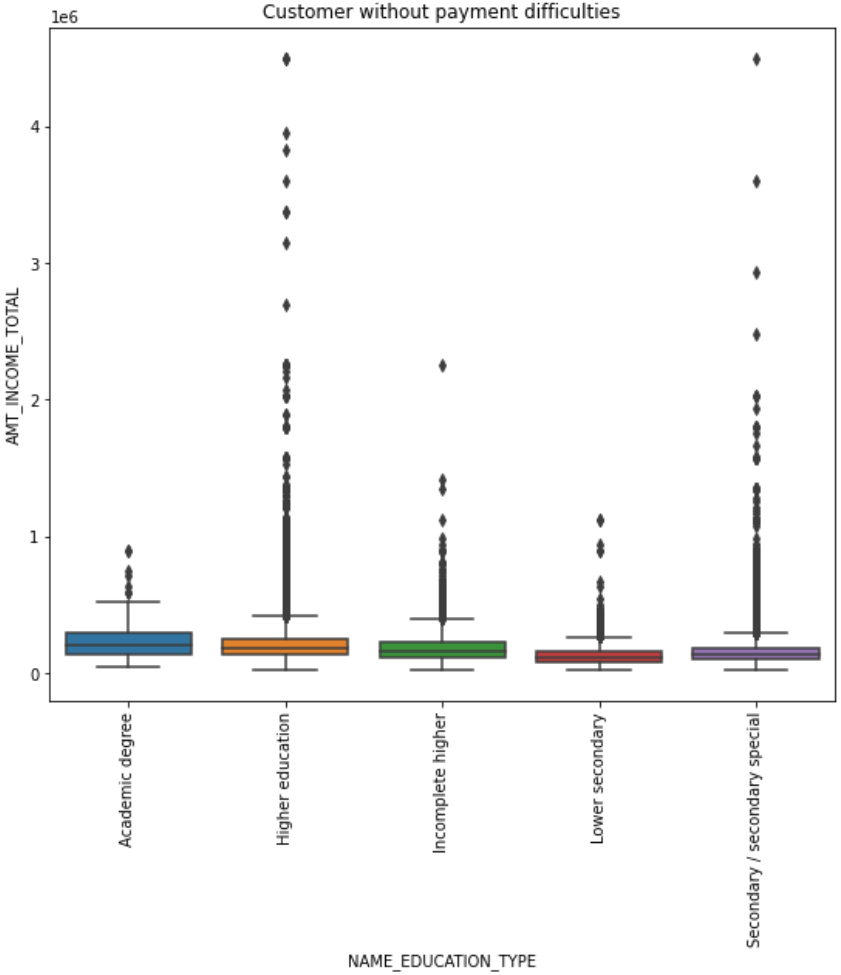


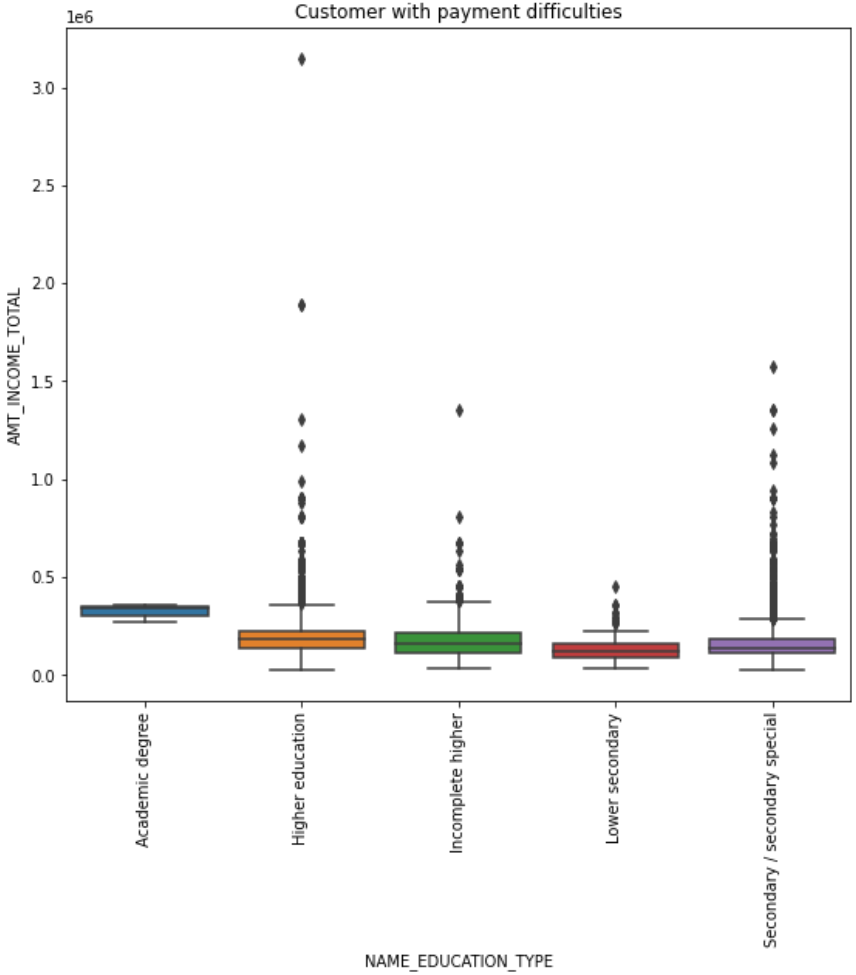
observe that Customer without payment of Academic Degree is more than customer with payment. Rest are almost same it seems



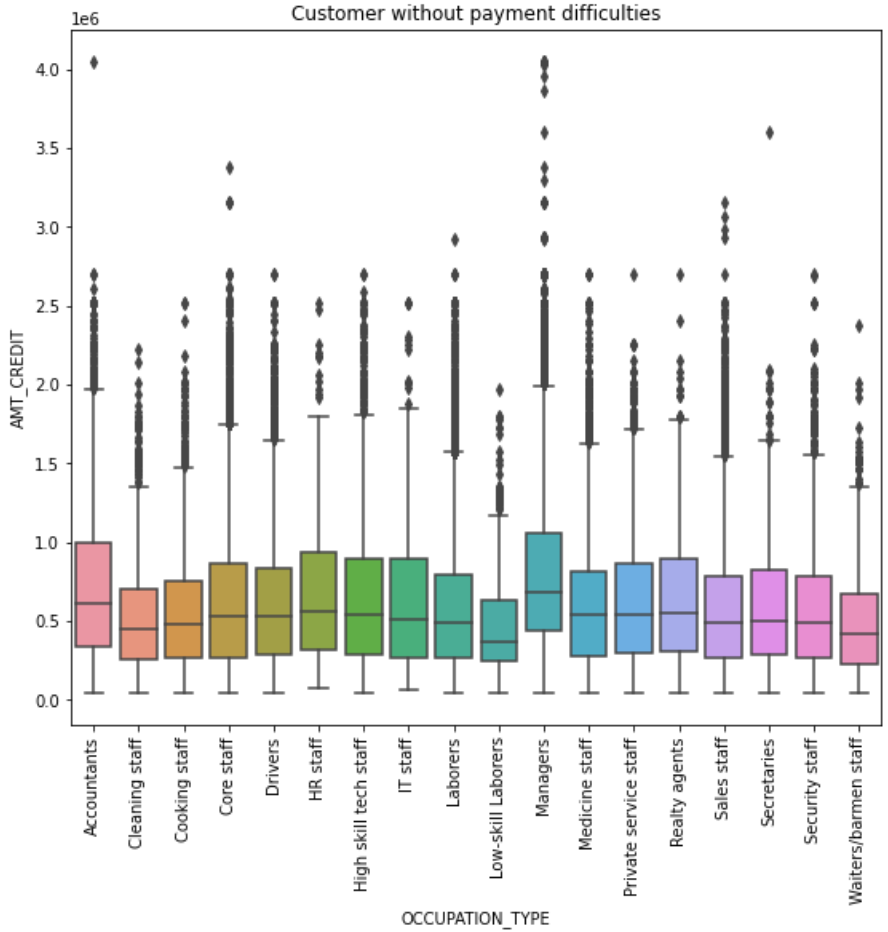


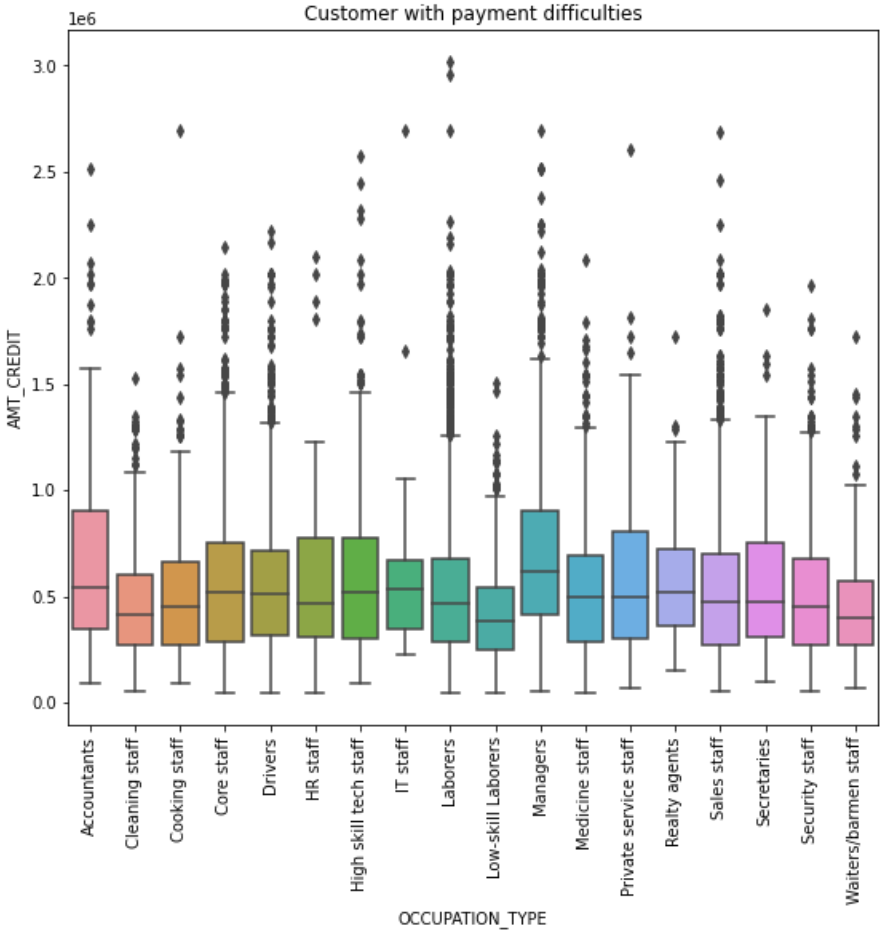
Customer without payment difficulties have more outliers than the customer with payment difficulties

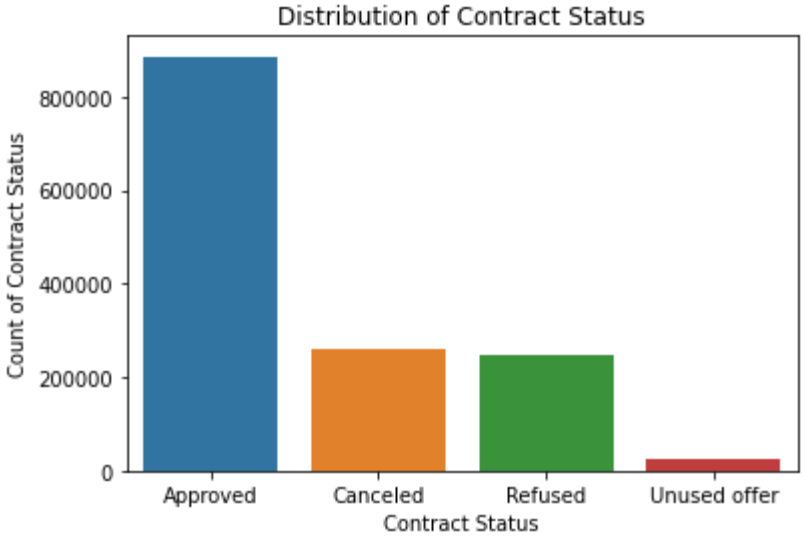




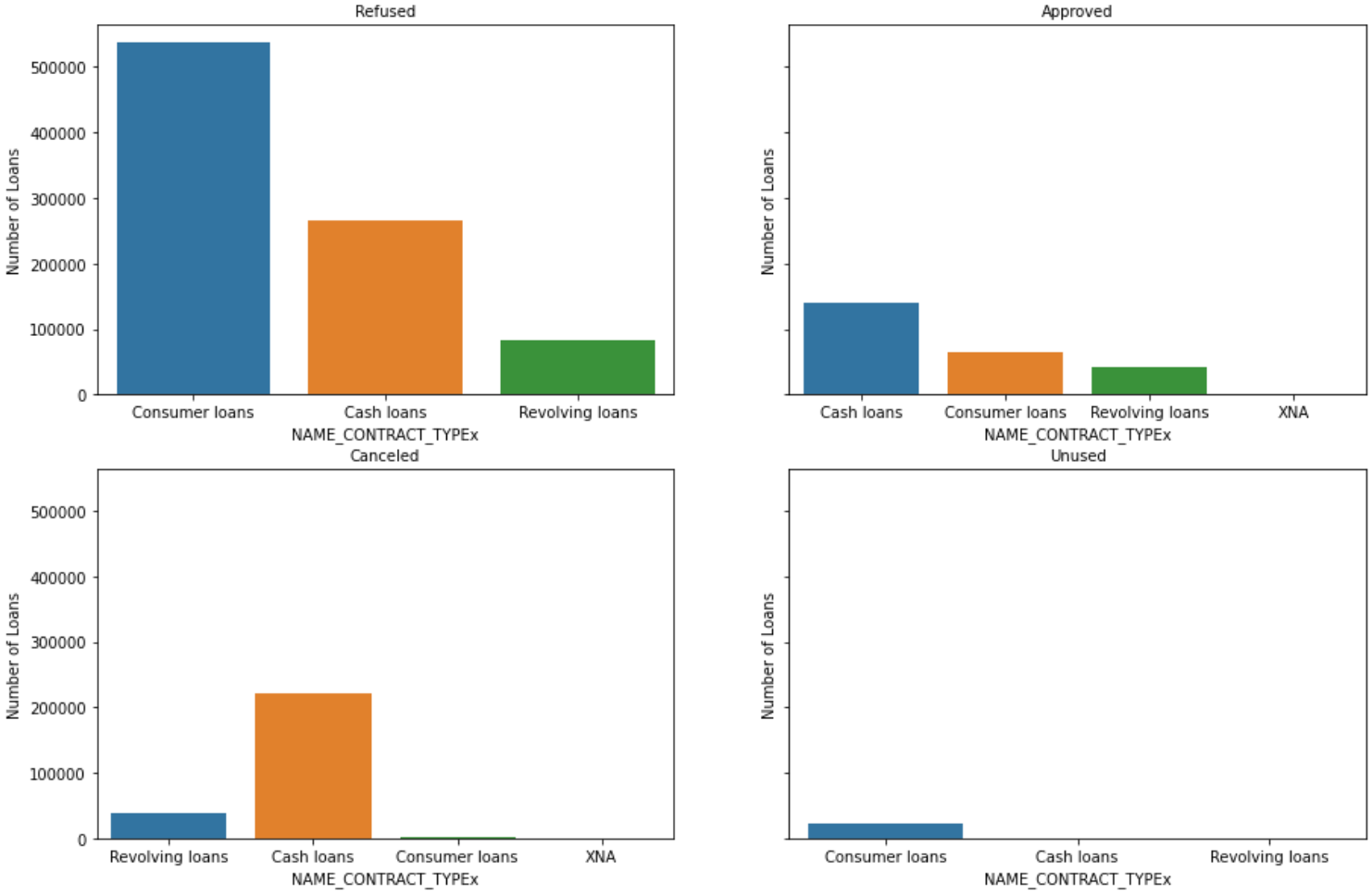
Here we can customers without payment difficulties and with pyament difficulties ranges..!!

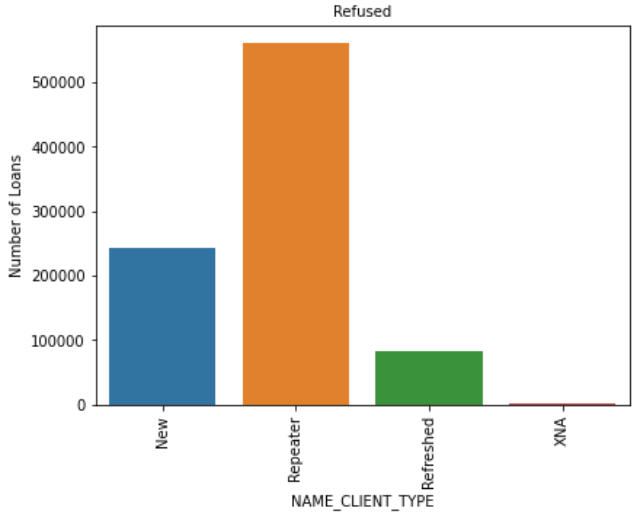


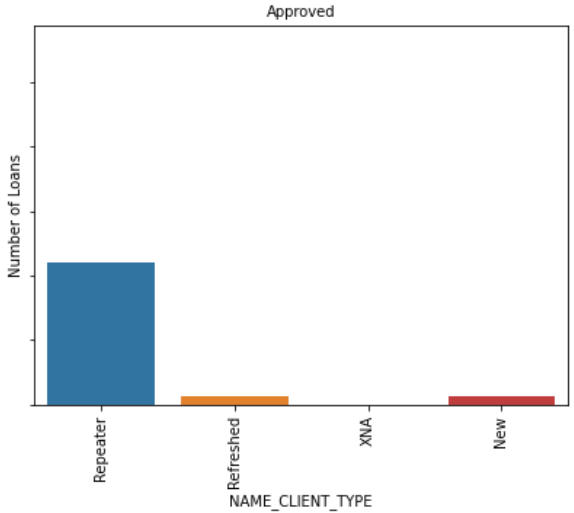


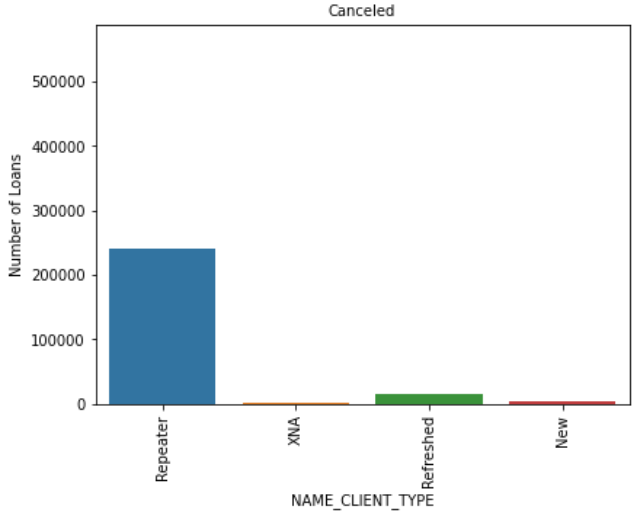


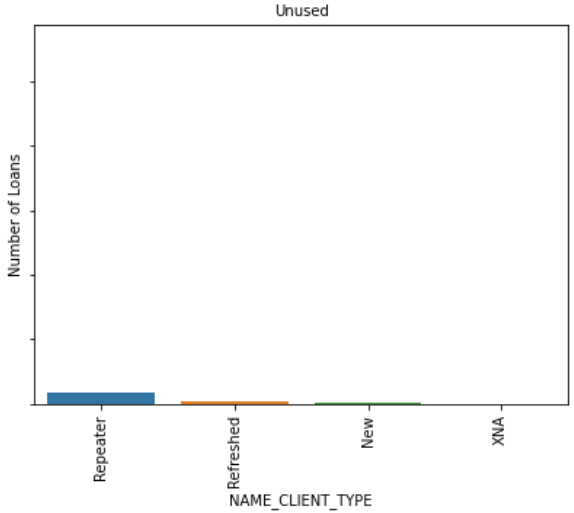
Distribution of contact status with respect to counts

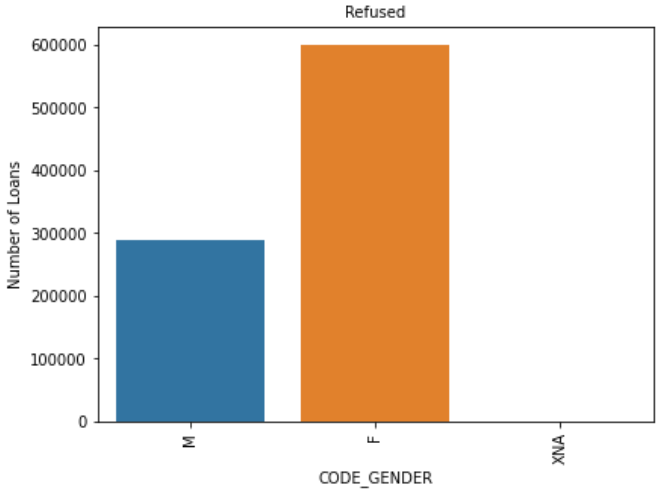

Refused
Number of Loans
500000
400000
300000
200000
100000
0
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPEx
Approved
Number of Loans
Cash loans
Consumer loans
Revolving loans
XNA
NAME_CONTRACT_TYPEx
Canceled
Number of Loans
500000
400000
300000
200000
100000
0
Revolving loans
Cash loans
Consumer loans
XNA
NAME_CONTRACT_TYPEx
Unused
Number of Loans
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPEx

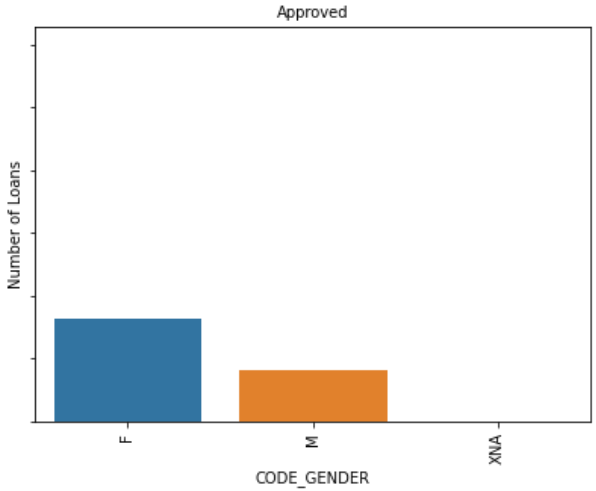

Refused
Number of Loans
500000
400000
300000
200000
100000
0
New
Repeater
Refreshed
XNA
NAME_CLIENT_TYPE

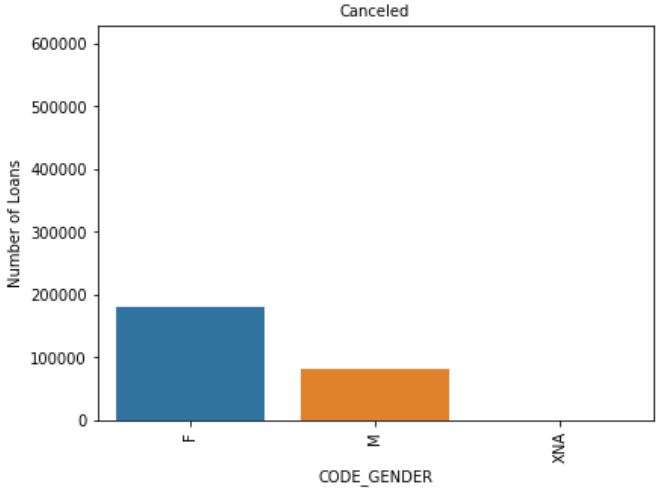

Approved
Number of Loans
Repeater
Refreshed
XNA
New
NAME_CLIENT_TYPE

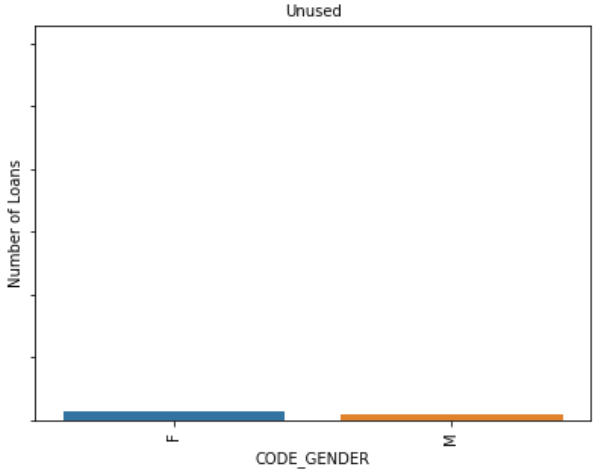

Canceled
Number of Loans
500000
400000
300000
200000
100000
0
Repeater
XNA
Refreshed
New
NAME_CLIENT_TYPE

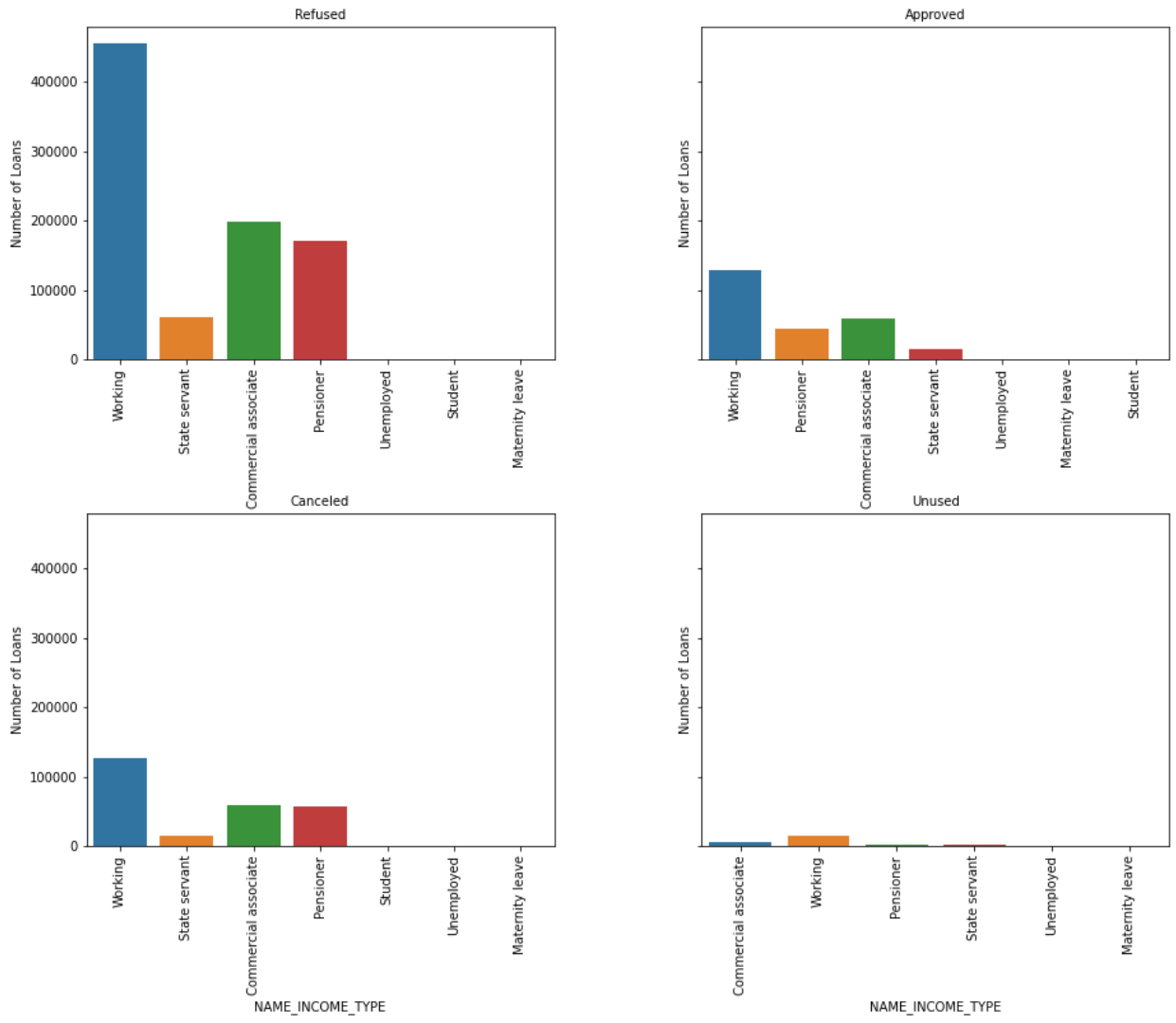

Unused
Number of Loans
Repeater
Refreshed
New
XNA
NAME_CLIENT_TYPE

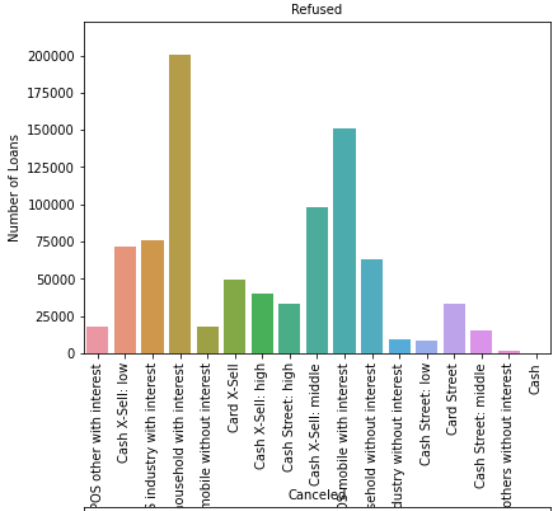

Refused
Number of Loans
600000
500000
400000
300000
200000
100000
0
M
F
XNA
CODE_GENDER

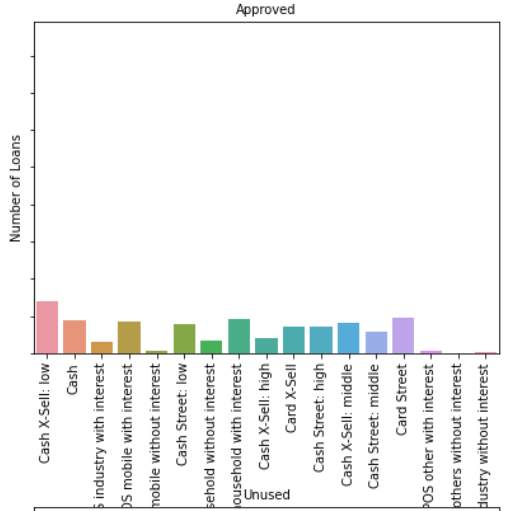

Approved
Number of Loans
F
M
XNA
CODE_GENDER


Canceled
Number of Loans
600000
500000
400000
300000
200000
100000
0
F
M
XNA
CODE_GENDER

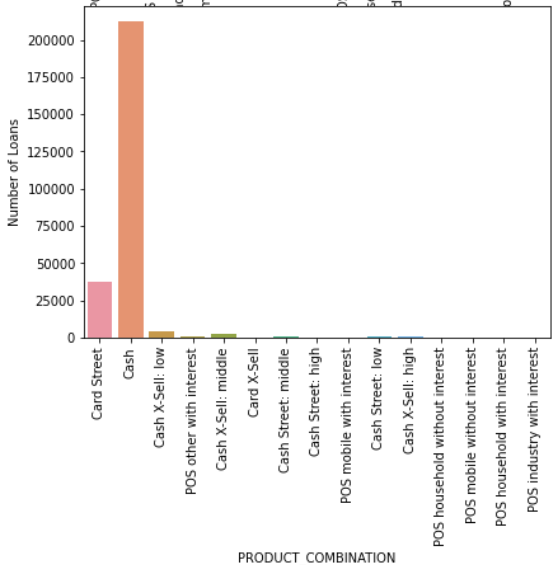

Unused
Number of Loans
F
M
CODE_GENDER


Refused
Number of Loans
400000
300000
200000
100000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Maternity leave
Approved
Number of Loans
Working
Pensioner
Commercial associate
State servant
Unemployed
Maternity leave
Student
Canceled
Number of Loans
400000
300000
200000
100000
0
Working
State servant
Commercial associate
Pensioner
Student
Unemployed
Maternity leave
NAME_INCOME_TYPE
Unused
Number of Loans
Commercial associate
Working
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE

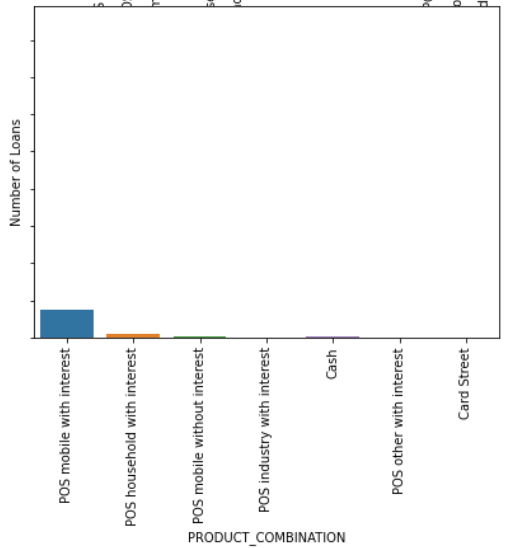




Canceled



Unused



Conclusion

- **Bank should look at contract type as a Student, Businessman, Pensioner and other type for successful payments**
- **Focus on income type as 'working' is have most non-payment records on observation on their non-payments.**
- **Check "Repair" purpose loan is having higher unsuccessful payments on time**