Exploratory Data Analysis

Bank-Loan approval Case-Study

Presented by:

Nikita Pise & Mohit Patil

(nikitapise11@gmail.com) (mohitz4418@gmail.com)

Problem Statement

There are two types of risks associated with the banks decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business of the bank.
- If the applicant is not likely to repay the loan, that is he or she is likely to default, then approving the loan may lead to financial loss for the bank.

Therefore the bank wants to know the deriving factors behind loan default.

Data Cleaning and Imputation –

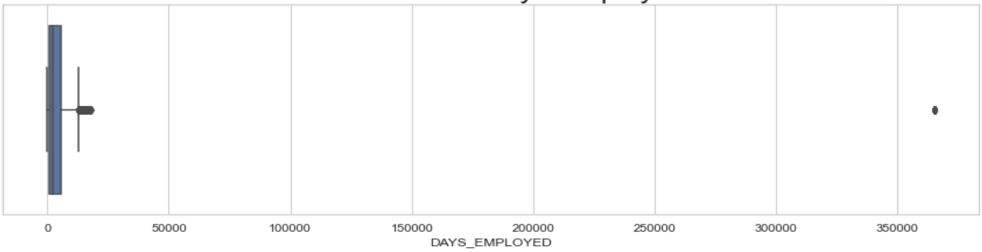
Data quality check and missing values

- There are many columns with large amount of missing values.
- Variables with more than 50% of missing values are dropped because it may affect the analysis.
- Also, we just checked the best matrix such as mean, median or mode to impute the missing values of variables having around 13% of missing values
- The data type of the variables is checked and the negative values present in the variables are converted into positive values. This step was needed because the variables such as days birth, days employed cannot be negative.

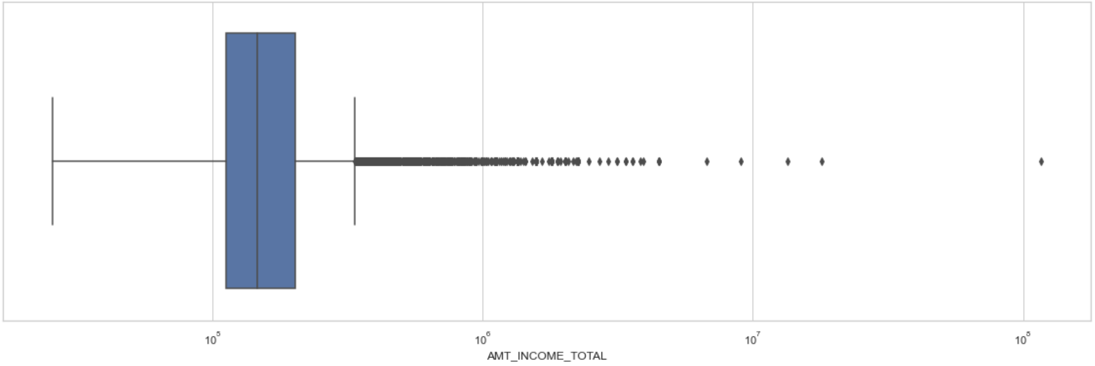
Data Cleaning and Imputation –

Numerical variables – Checked the presence of outliers

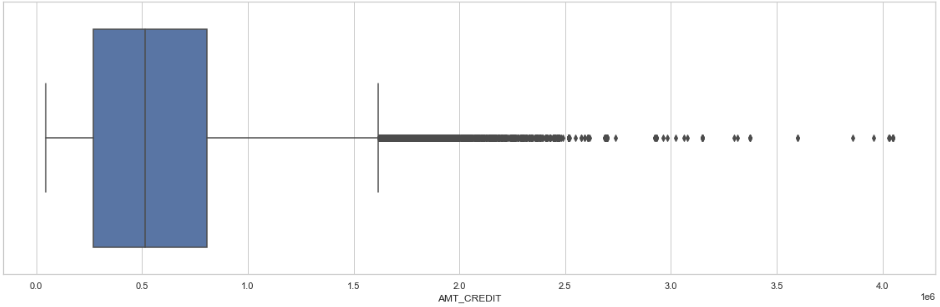
Distribution of Days Employed



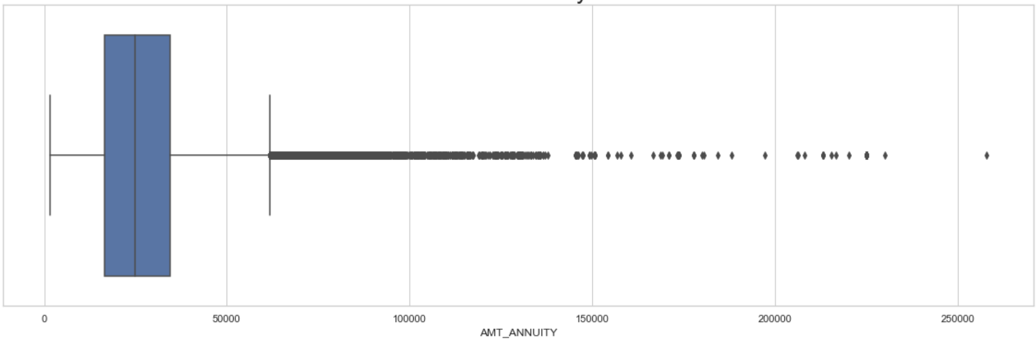
Distribution of Income amount



Distribution of Credit Amount



Distribution of Annuity Amount

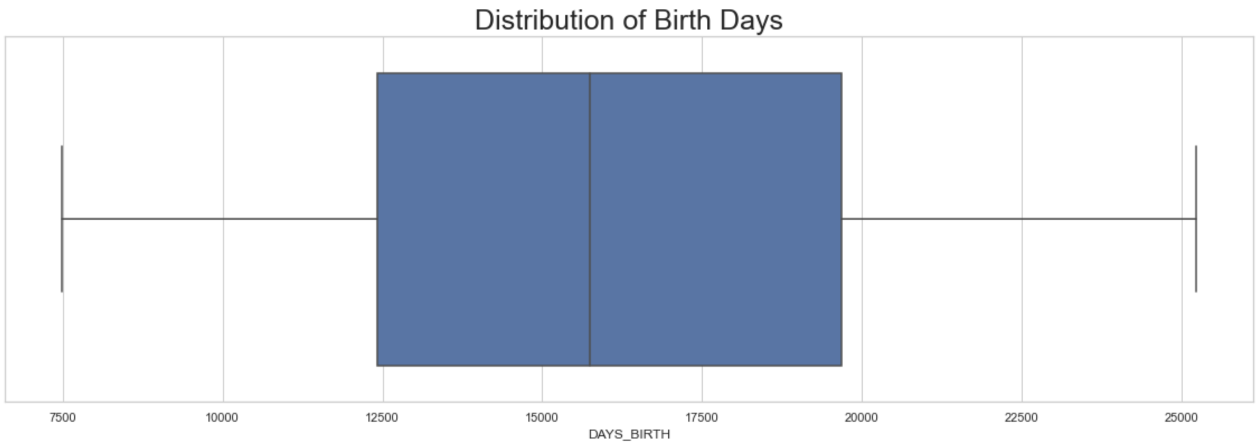


Inferences / Observation :-

There are so many outliers present in these variables. Hence , the best metric to impute these values is with their respective median.

Data Cleaning and Imputation –

Numerical variables – Checked the presence of outliers



Inferences / Observation :-

- There are no outliers in Days Birth. Hence, the best metric to impute the missing values is its mean.
- All clients have age approximately between 32 to 55.

Binning of continuous variables–

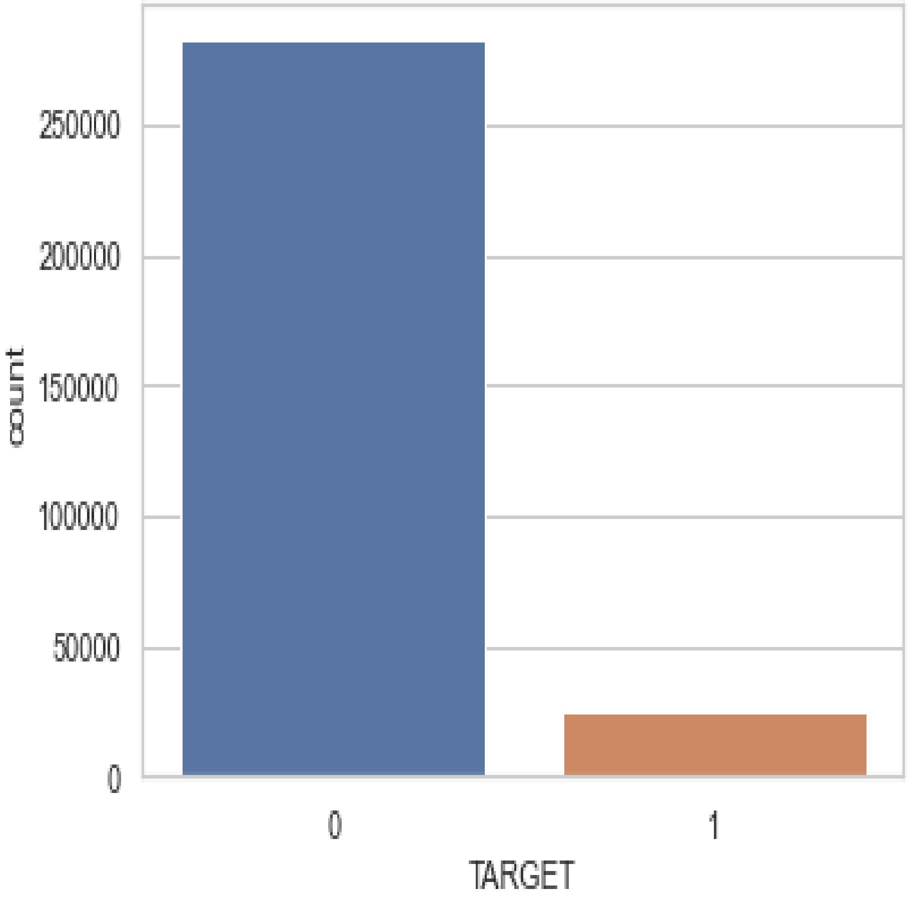
- Variable ‘DAYS_BIRTH’ is grouped into different categories such as baby, child, Young Adults, Middle-age Adults, Old-age Adults, Above 59 - Below 89, Above 89 depending on their age.
- Also, variable ‘AMT_INCOME_TOTAL’ is grouped into different categories such as extremely low, low, moderate and high depending on their income.
- We have some unnecessary columns in the dataframe. lets drop them and filter data frame.

Analysis–

The imbalance percentage is checked

- From count plot we can observe that the target variable is highly Imbalanced.

- Where in TARGET variable for value 0 it has 91.92% and for value 1 it is 8.07%.



Divide the data into two sets

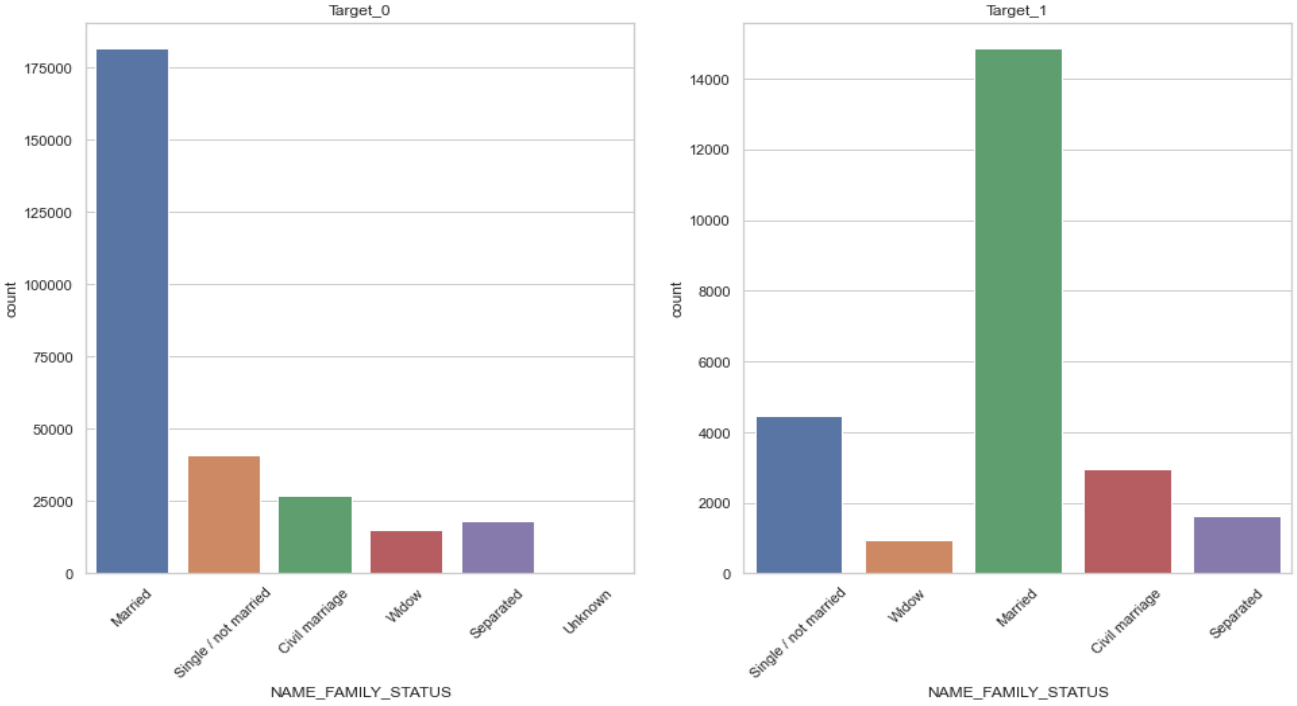
- The data is converted into two deta sets that is ‘Target_0 ‘ = 0 and ‘Target_1’ = 1.

Analysis–

Univariate analysis for categorical variables for both 0 and 1.

- Univariate analysis are carried out on 6 categorical variables i.e. NAME_CONTACT_TYPE, CODE_GENDER, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_FAMILY_STATUS and INCOME_CATEGORY.
- In most of the categories the rate of getting defaulted ranges between 5 to 10%. For each factor.

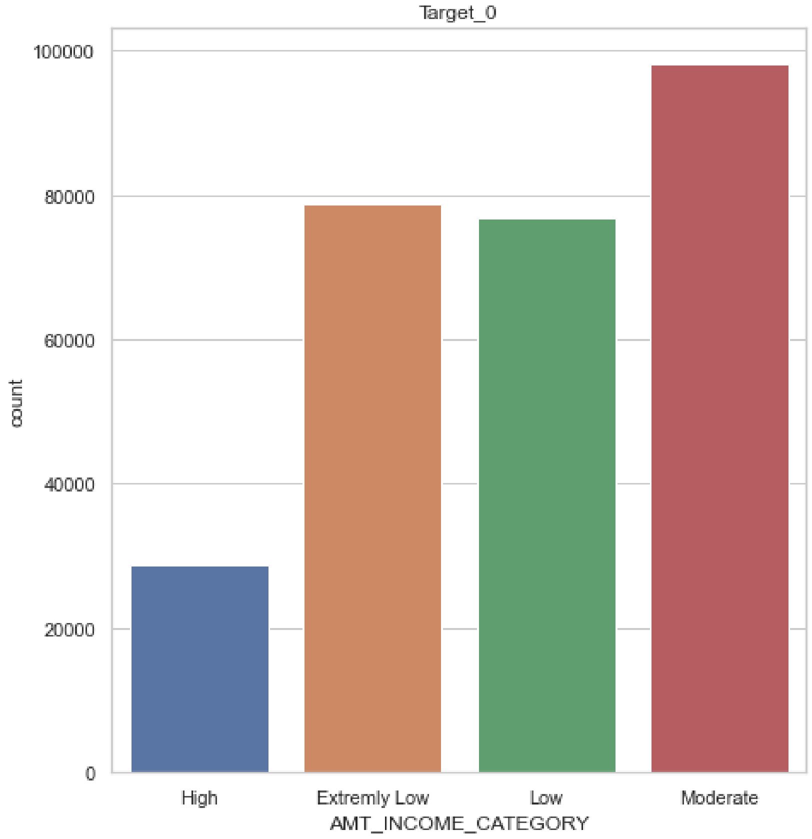
- From the graph we can observe that married, separated and widow have less chances of getting defaulted. Single people have more chances of getting defaulted followed by civil marriage.

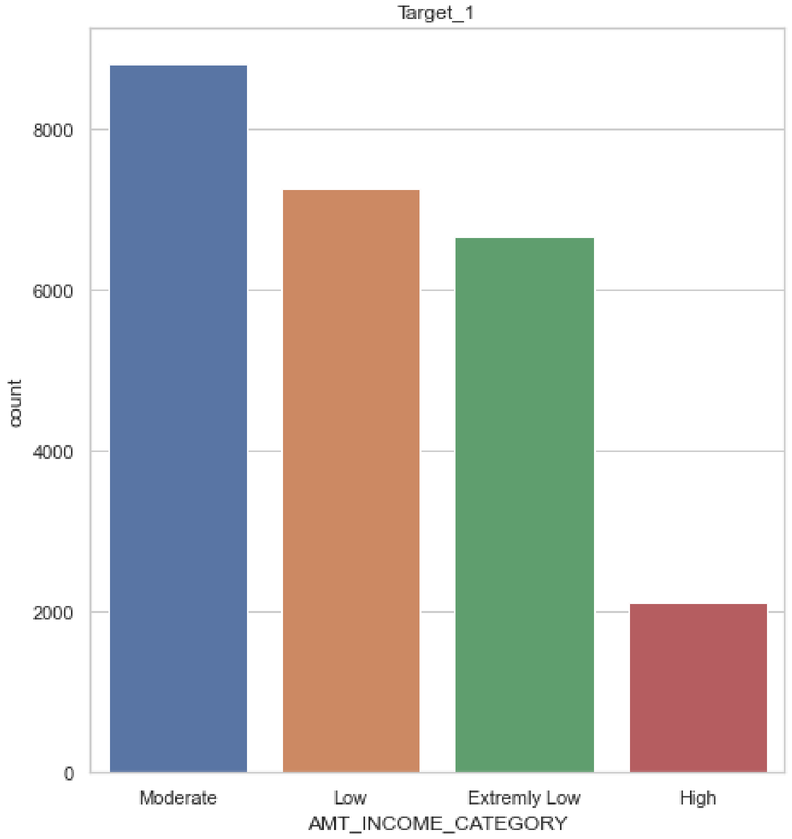


Analysis–

Univariate analysis for categorical variables for both 0 and 1.

- From all the income categories the chances of loan getting defaulted is around 9% except the high income category.
- People with high income have the least chances of getting defaulted whereas the people with low ,extremely low and moderate have almost same chances of getting defaulted

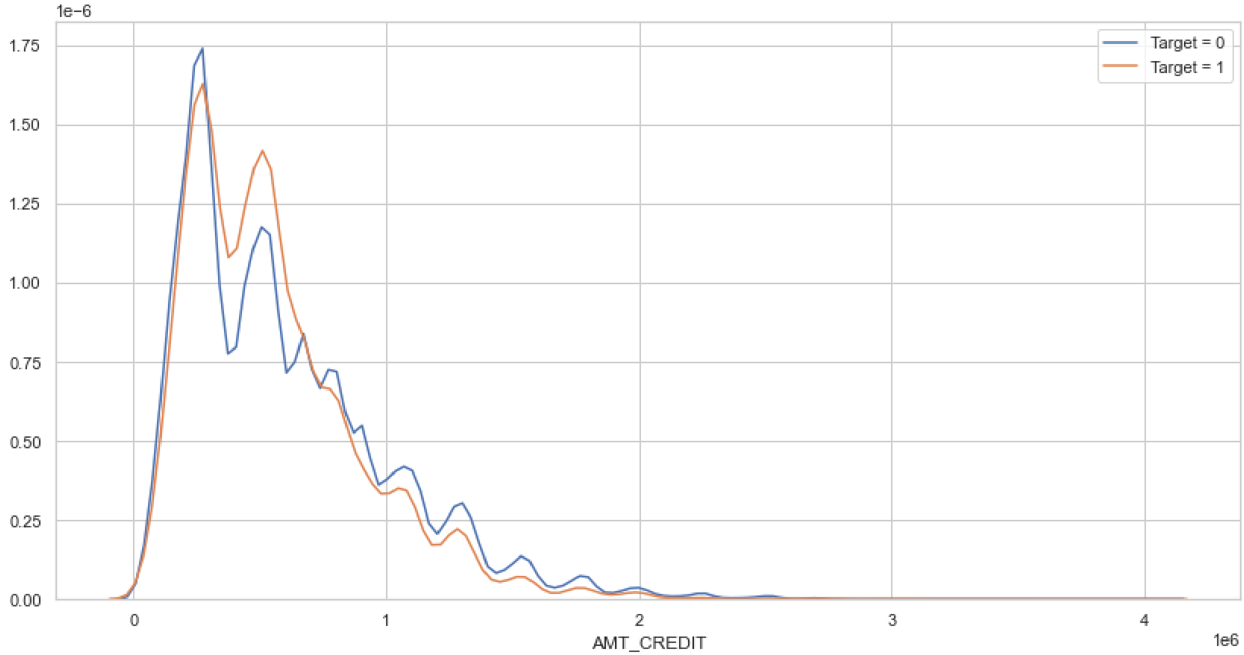


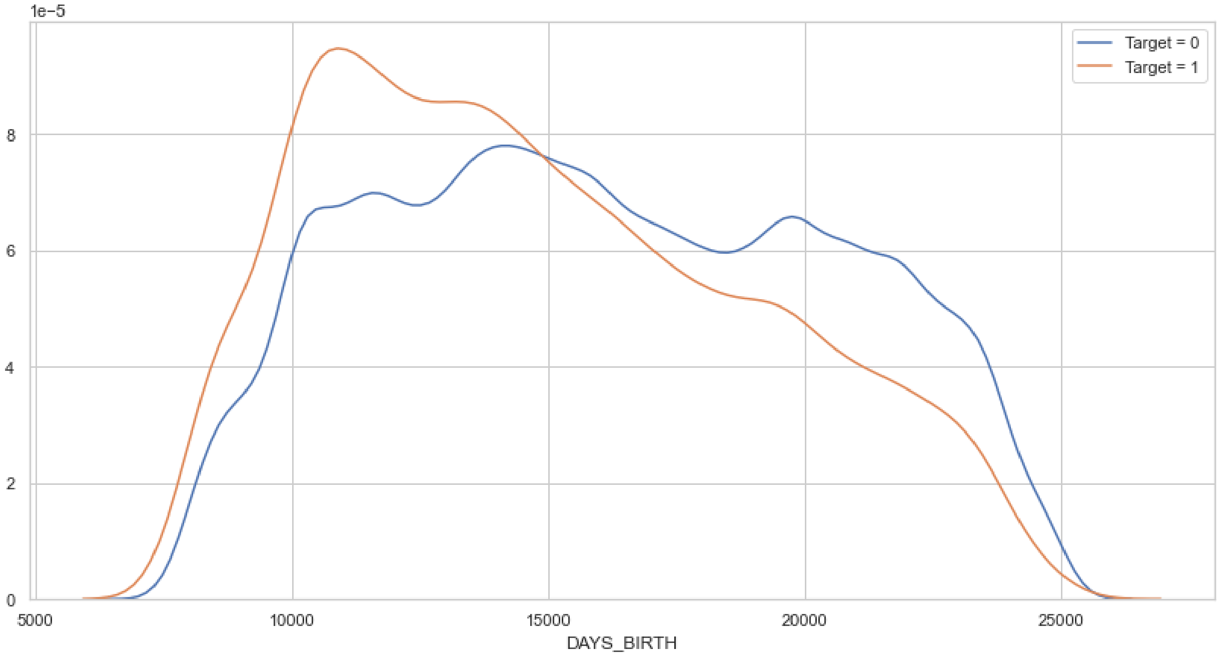


Analysis–

Univariate analysis for numerical variables for both 0 and 1.

The significant difference is seen in variables ‘AMT_CREDIT’, ‘DAYS_BIRTH’ and ‘AMT_GOODS_PRICE’





- From the variable ‘AMT_CREDIT’ we can observe that lesser the credit amount higher is the risk getting defaulted
- For higher credit amounts we have less risk of getting defaulted

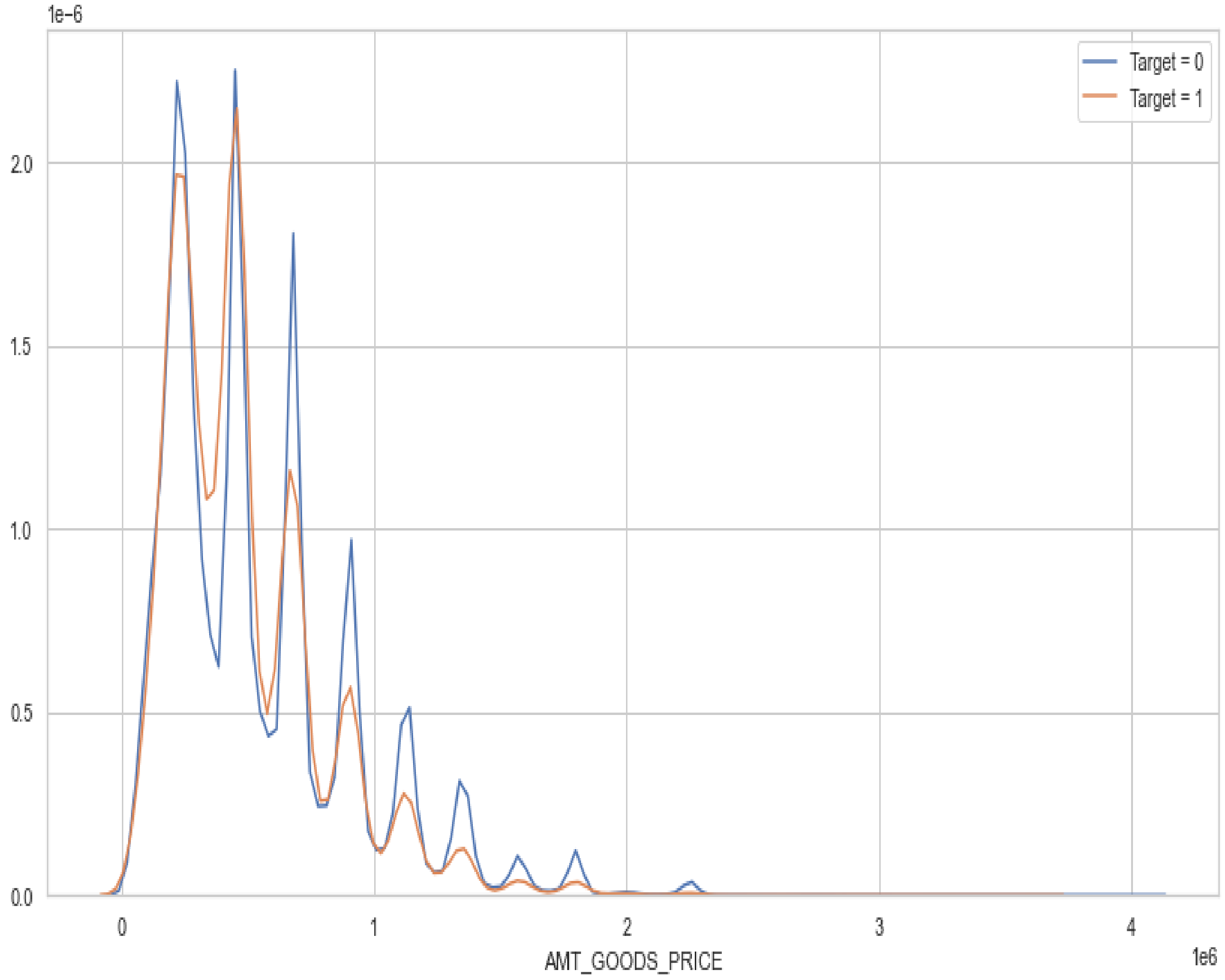
- Higher default rate is seen among the age group below 40. Whereas between the age group 40 to 70 the default rate is less.
- Hence, the credit loss can be reduced by issuing the loan to people having age above 40.

Analysis–

Univariate analysis for numerical variables for both 0 and 1.

The significant difference is seen in variables ‘AMT CREDIT’. ‘DAYS BIRTH’ and ‘AMT_GOODS_PRICE’

- Goods having price greater than 500000 are associated with less default risk
- Goods having price less than 500000 have more chances of getting default.



Analysis–

Correlation

Check the variables with highest correlation are the same in both files or not.

Top columns with high value of correlation in Target_1 are:

AMT_GOODS_PRICE = AMT_CREDIT

AMT_GOODS_PRICE = AMT_ANNUITY

AT_ANNUITY = AMT_CREDIT

Top columns with high value of correlation in Target_0 are:

AMT_GOODS_PRICE = AMT_CREDIT

AMT_GOODS_PRICE = AMT_ANNUITY

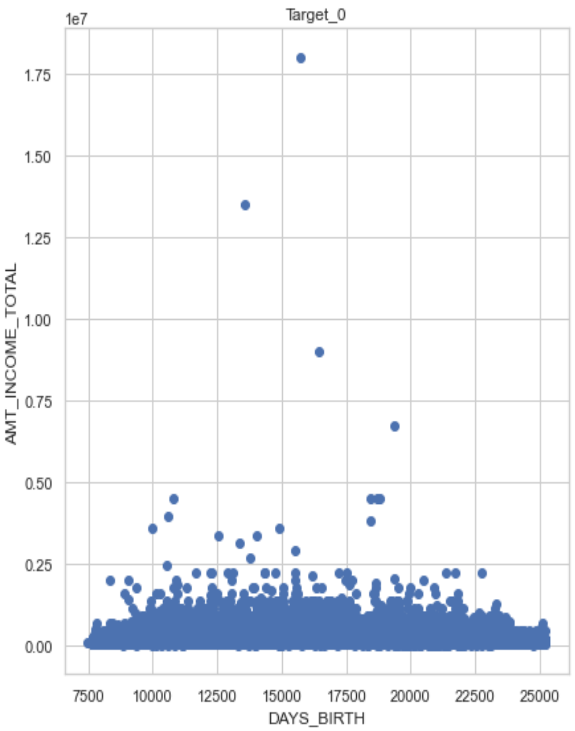
AT_ANNUITY = AMT_CREDIT

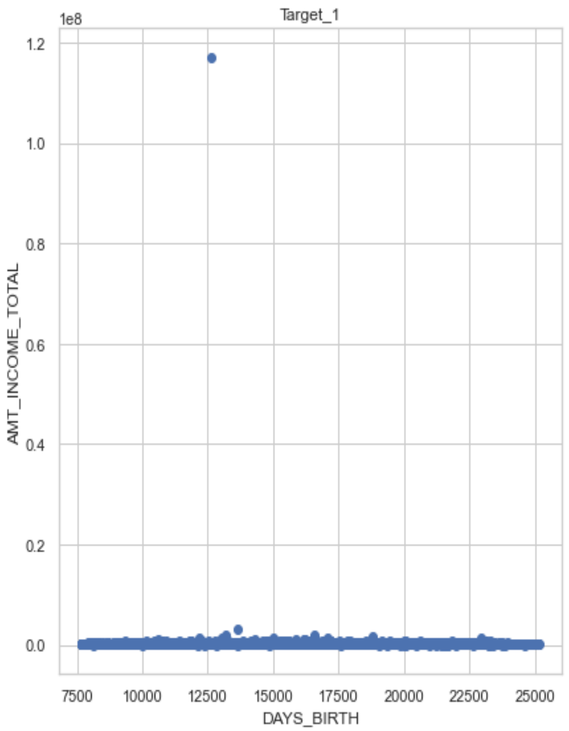
So, both Target_0 and Target_1 have same correlation variable

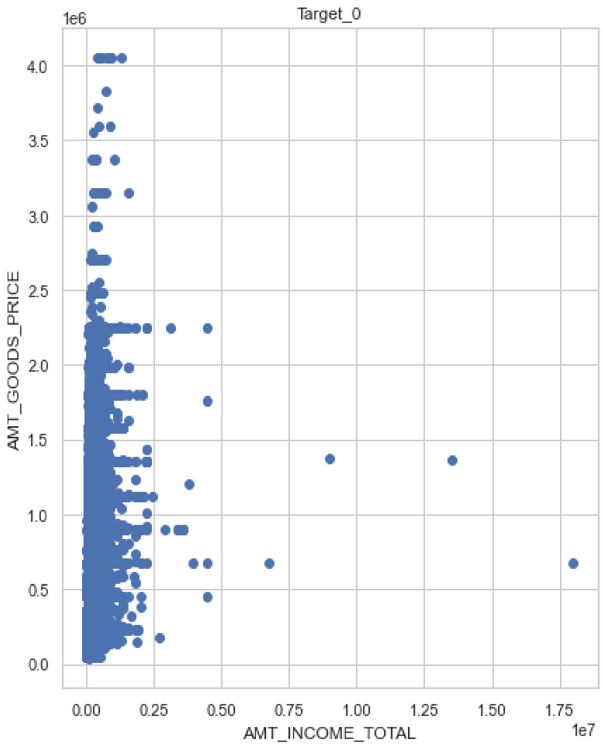
Analysis–

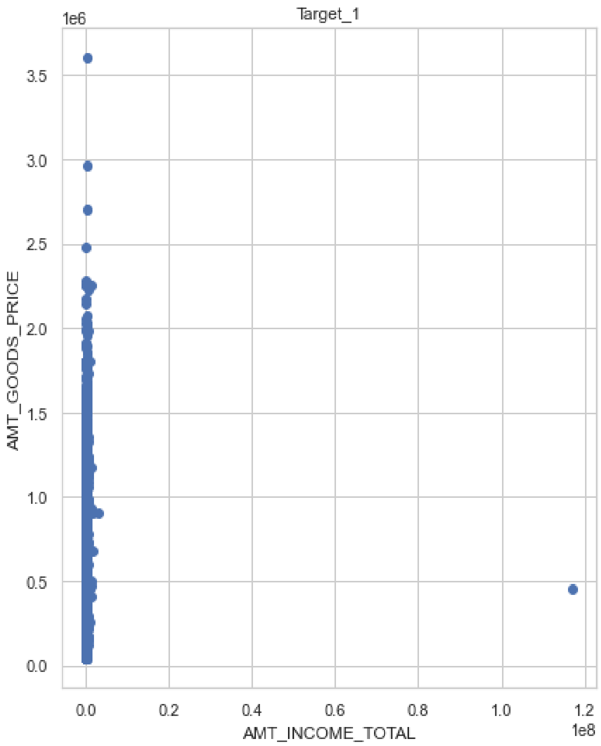
a. Bivariante Analysais for Continuos - Continuos variables.

The significant difference is seen in variables ‘AMT_INCOME_TOTAL’-‘DAYS_BIRTH’, and ‘AMT_INCOME_TOTAL’-‘AMT_GOODS_PRICE’









• From both the graphs and comparison we can observe that majority of the applicants are above the age 25 and found to have the total income less than 25 lakhs. Applicants with higher income that is greater than 10 lakhs have less risk of default.

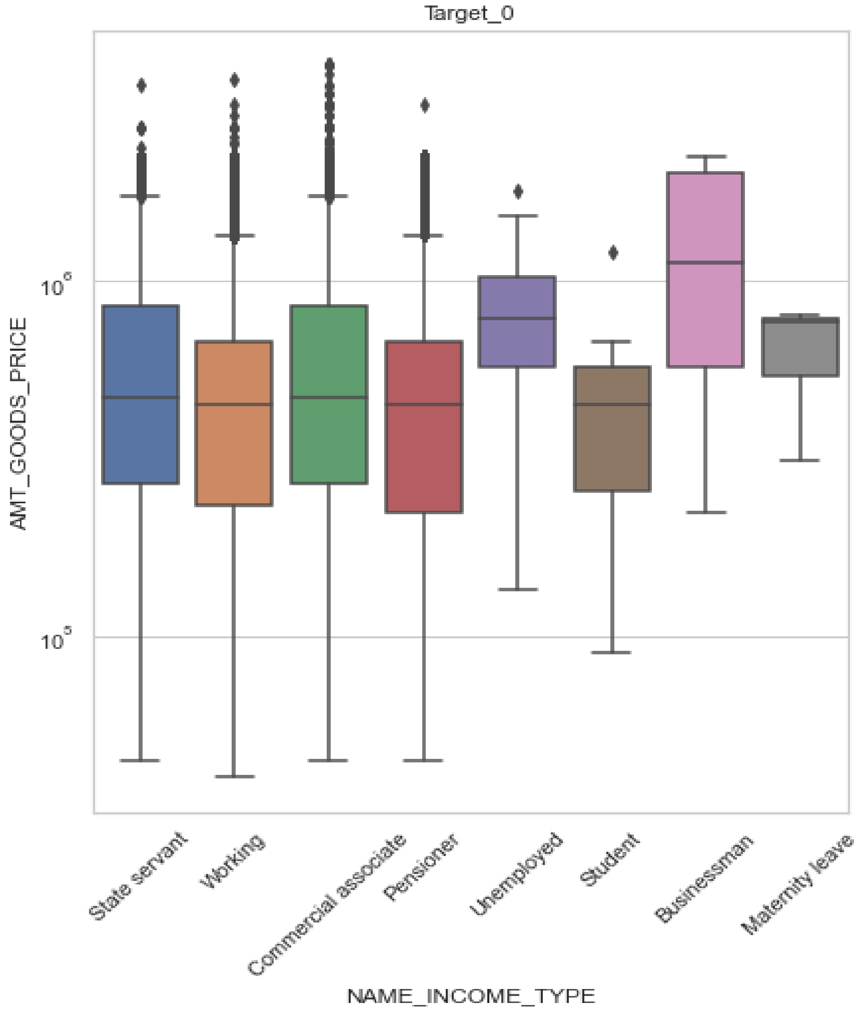
•Whereas, the applicants with income less than 10 lakhs have more chances of being defaulted irrespective of their age .

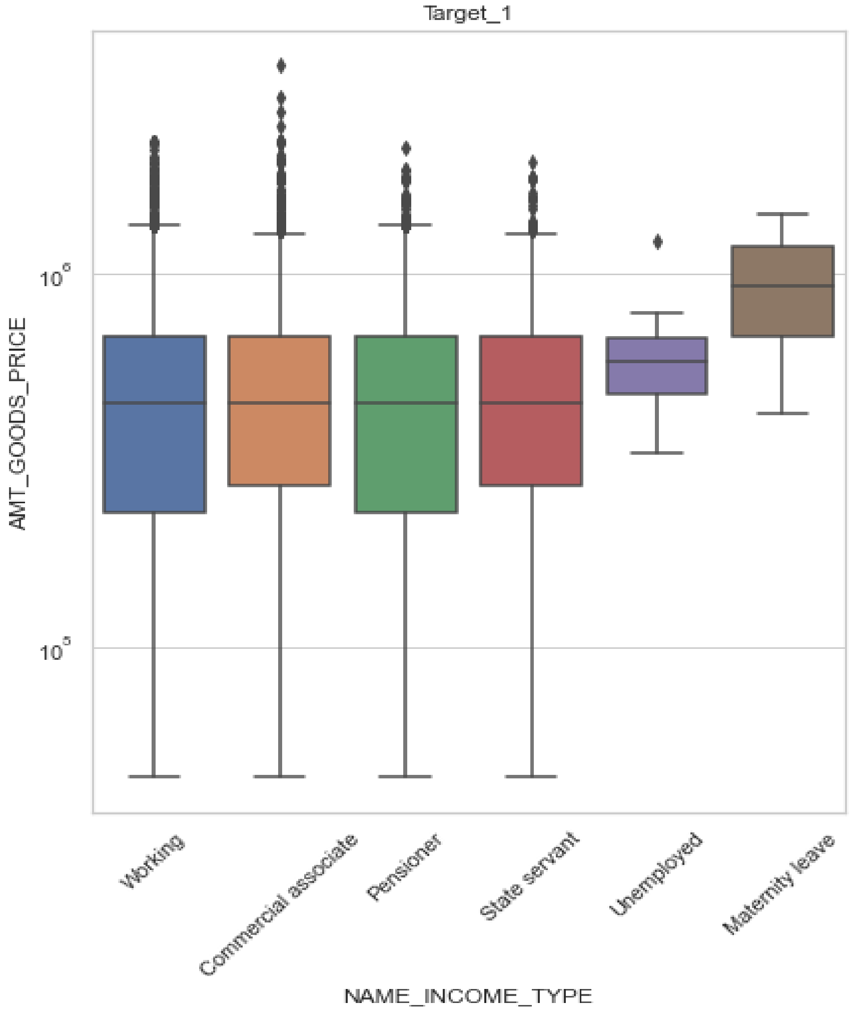
Analysis–

b. Bivariate Analysis for Continuous - Categorical variables..

The significant difference is seen in variables ‘NAME_INCOME_TYPE’- ‘AMT_GOODS_PRICE’ AND ‘NAME_EDUCATION_TYPE’-’AMT_CREDIT’

- In Target_0 clients with income type as Businessmam have maximum consumer loans price of goods for which the loan is given, so it has higher chances that the clients having income type as Businessman has more probability than other to be defaulter.
- In Target_1 clients with all income type have same range of goods price but only for income type as maternity leave has higher goods price. the clients with maternity leave have more chances of not paying the loan.
- The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.





Analysis–

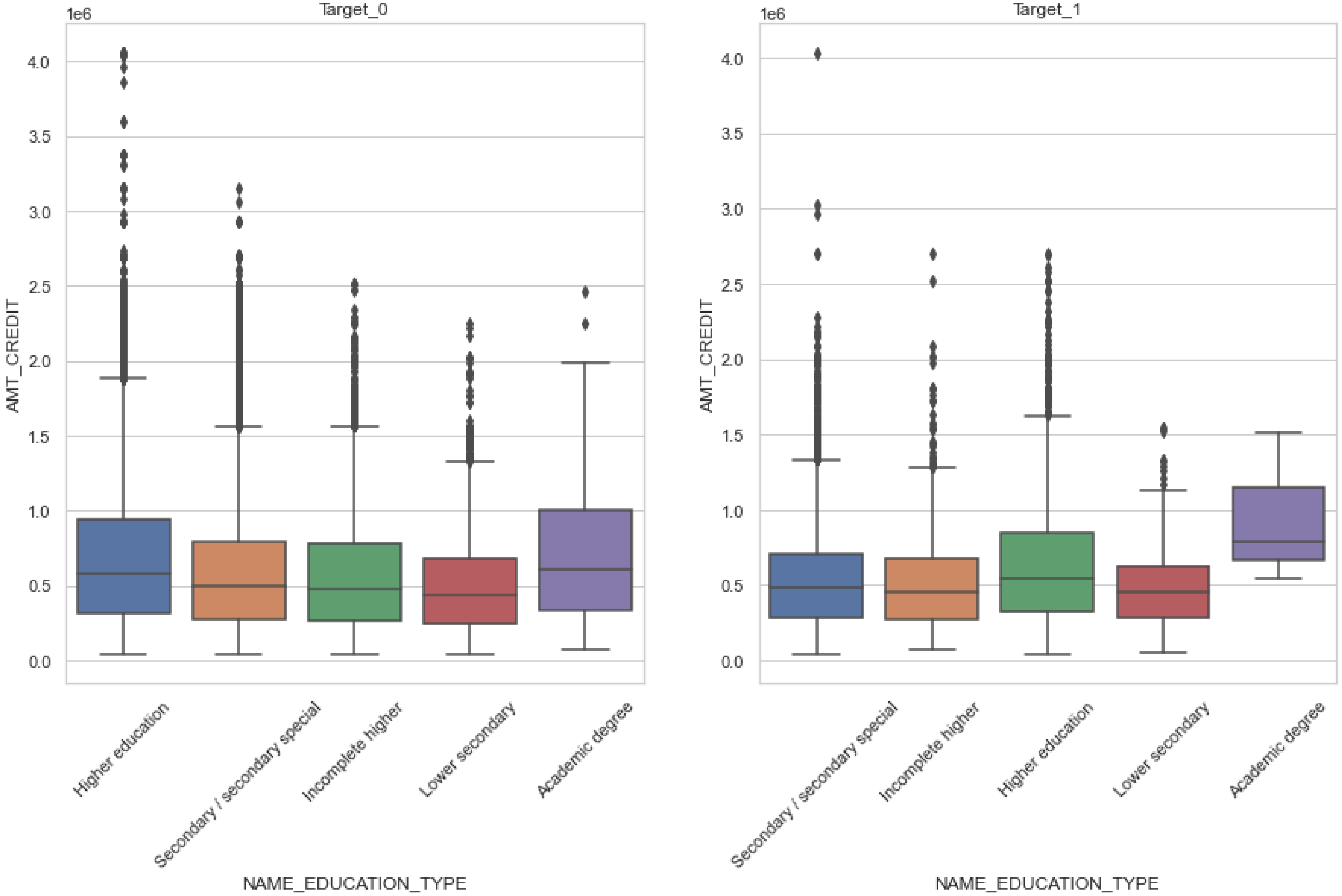
b. Bivariate Analysis for Continuous - Categorical variables..

The significant difference is seen in variables ‘NAME_INCOME_TYPE’- ‘AMT_GOODS_PRICE’ AND ‘NAME_EDUCATION_TYPE’- ’AMT_CREDIT’

- In both Target_0 and Target_1 the clients having education type as Academic degree have higher chances to get more credit amount of loan.

- Majority of the clients have Secondary secondary special education, followed by clients with Higher education. Only a very small number is having an academic degree.

- The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% not-repayment rate.

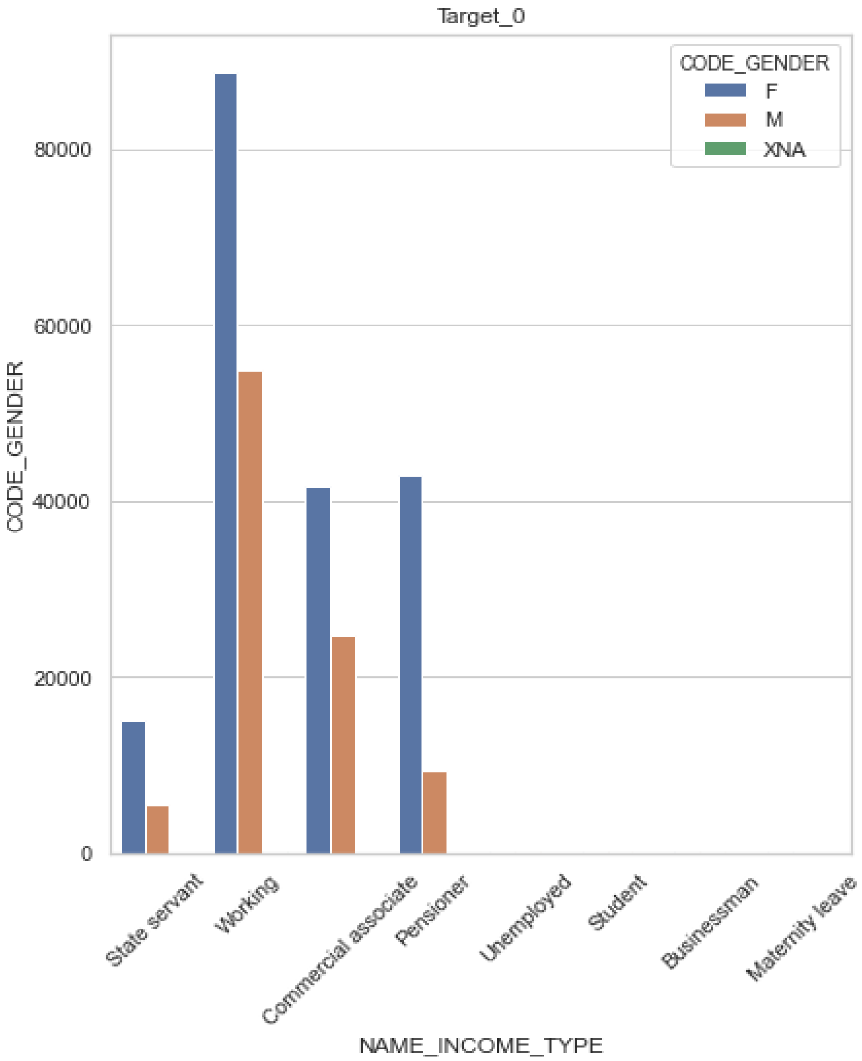


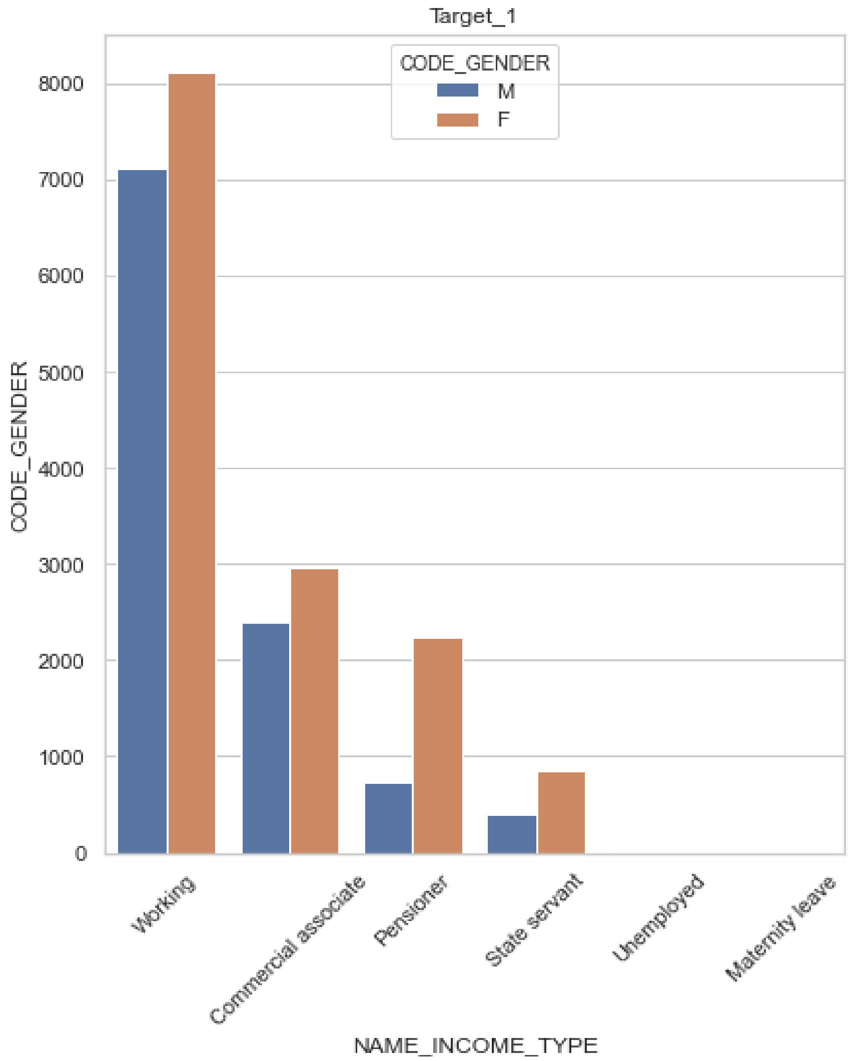
Analysis–

c. Bivariate Analysis for Categorical- Categorical variables.

The significant difference is seen in variables ‘NAME_INCOME_TYPE’- ‘CODE_GENDER’ and ‘NAME_EDUCATION_TYPE’-’NAME_FAMILY_STATUS’.

- Target_0 and Target_1 both have maximum clients with income type as working.
- Most of applicants for loans are getting income from Working sector, followed by Commercial associate, Pensioner and State servant.
- The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.





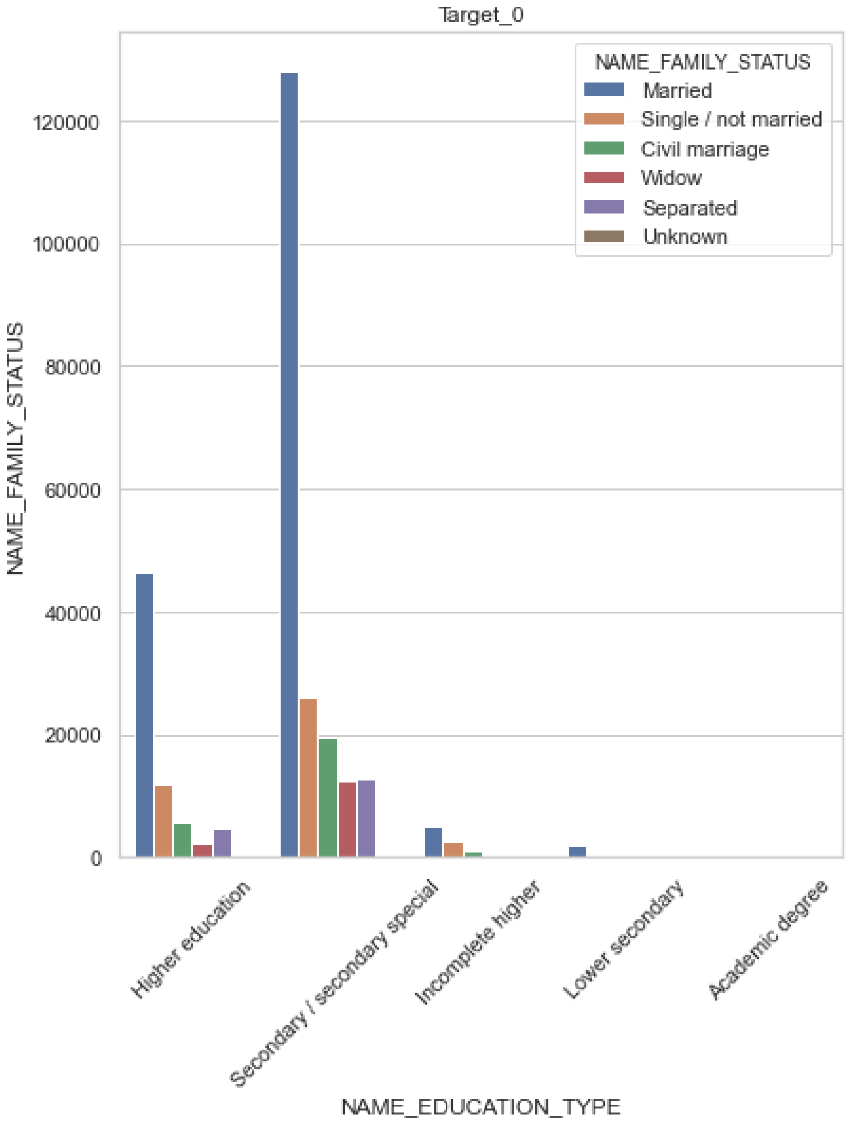
Analysis–

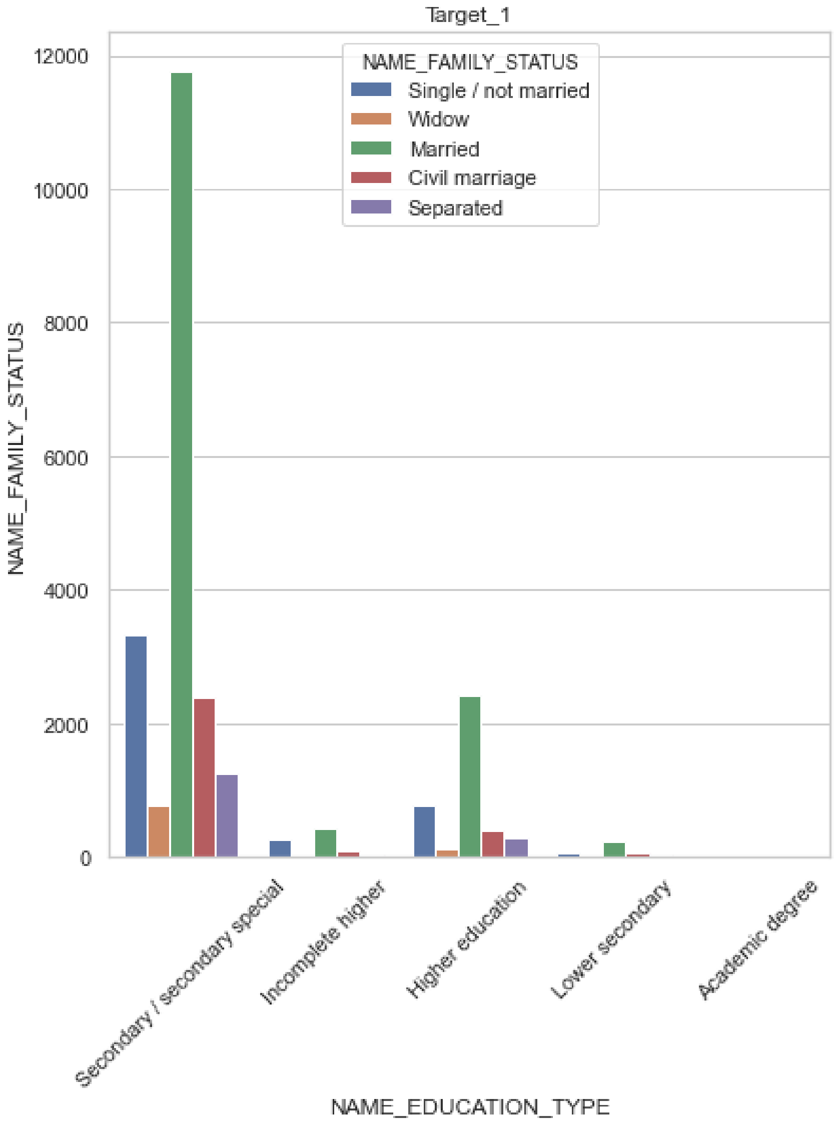
c. Bivariate Analysis for Categorical- Categorical variables.

- Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number of applicants are having an academic degree.

- Majority of the clients have Secondary / secondary special education in which most of the applicants have family status as Married in both Target_0 and Target_1.

- The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% not-repayment rate.



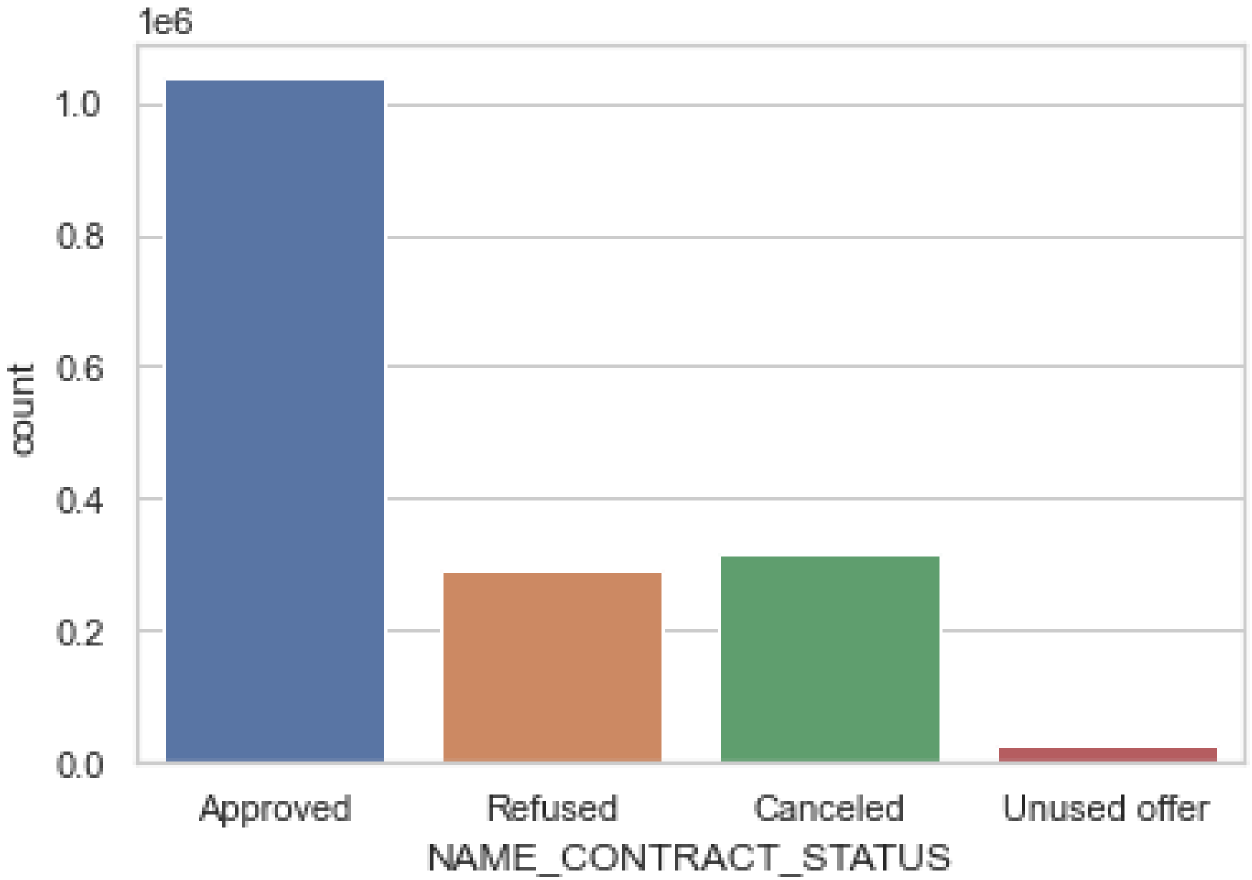


Analysis on previous _application.csv

Checking Imbalance Percentage.

- From countplot we can observe that the data Imbalanced.

- Where NAME_CONTRACT_STATUS variable for value "Approved" it has 62.07% , for value "Refused" it is 18.93% , for value "Canceled" it is 17.40% and for value "Unused offer" it is 1.58%

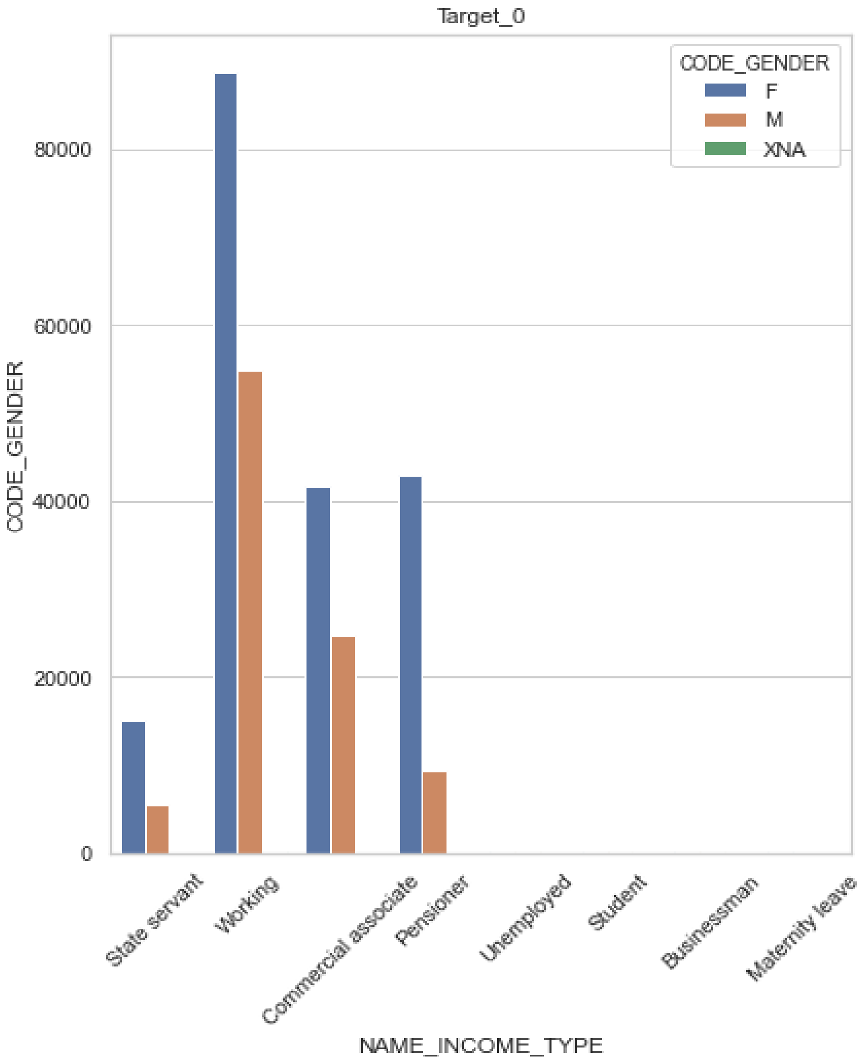


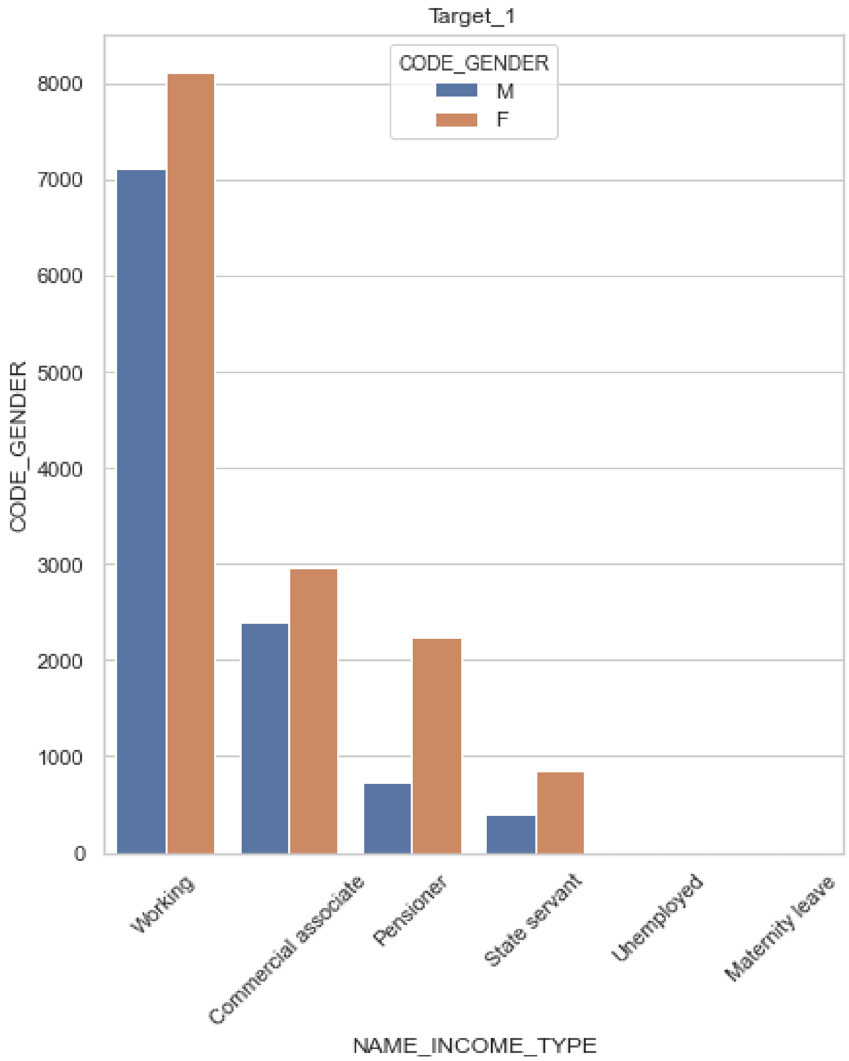
Analysis–

c. Bivariate Analysis for Categorical- Categorical variables.

The significant difference is seen in variables ‘NAME_INCOME_TYPE’- ‘CODE_GENDER’ and ‘NAME_EDUCATION_TYPE’-’NAME_FAMILY_STATUS’.

- Target_0 and Target_1 both have maximum clients with income type as working.
- Most of applicants for loans are getting income from Working sector, followed by Commercial associate, Pensioner and State servant.
- The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.





Analysis–

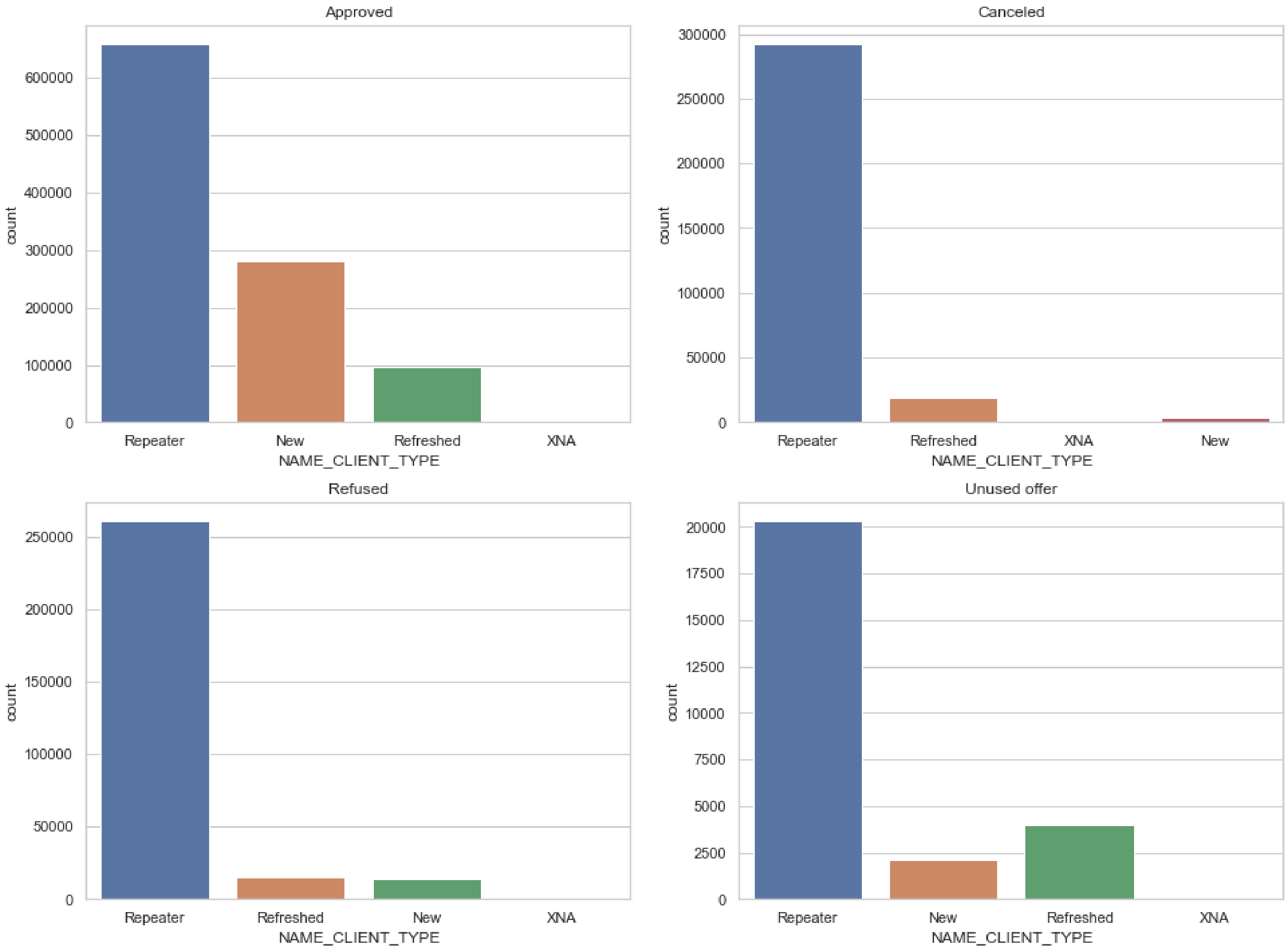
a. Univariate analysis for categorical variables.

The significant difference is seen in variables ‘NAME_CLIENT_TYPE’ and ‘NAME_CONTRACT_TYPE’

- Mostly approved application belongs to repeater client category followed by new applications.

- Half amount of repeater client applications are canceled;

- Repeater client applications refused approx 250K times.



Analysis–

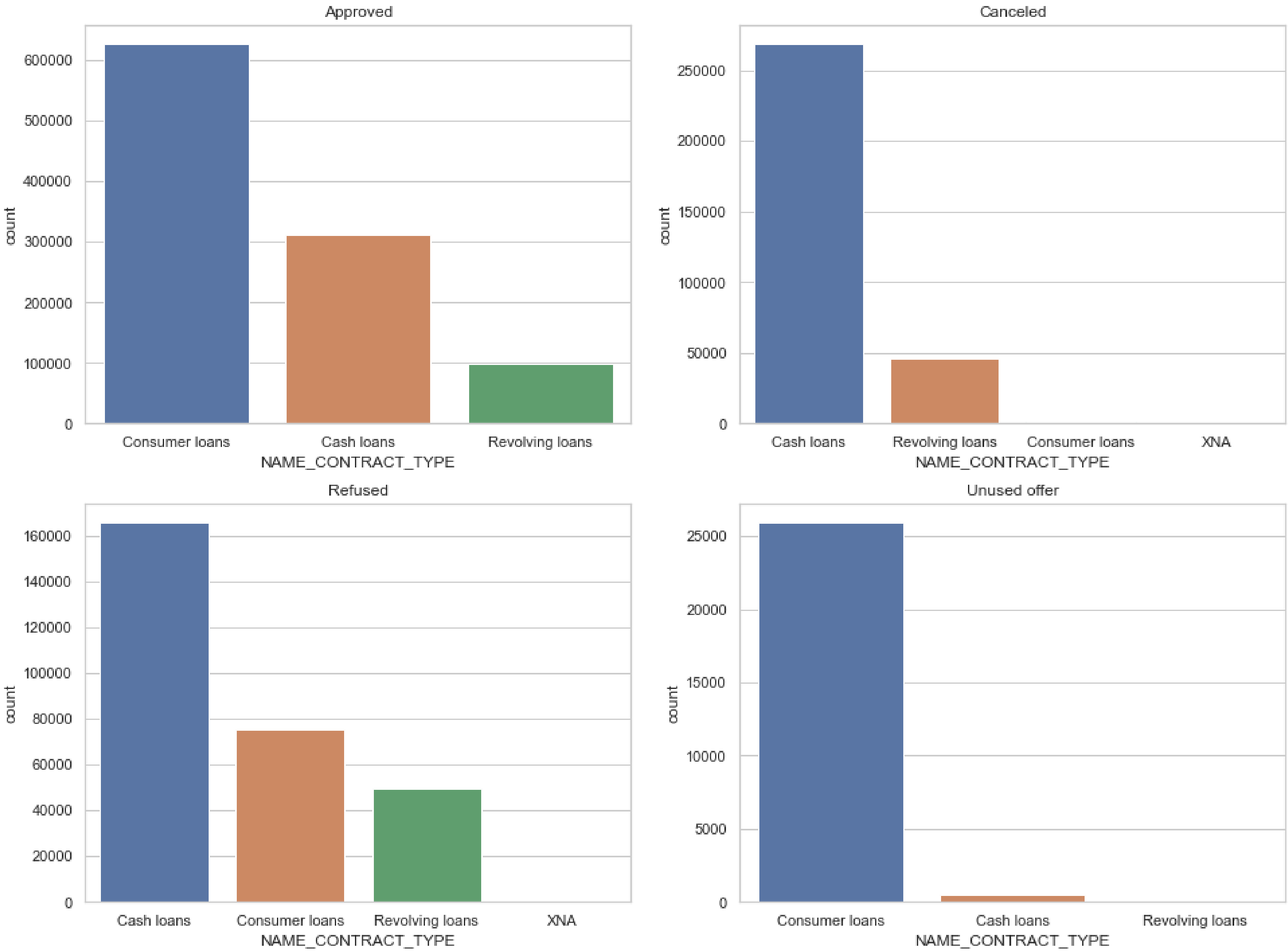
a. Univariate analysis for categorical variables.

The significant difference is seen in variables ‘NAME_CLIENT_TYPE’ and ‘NAME_CONTRACT_TYPE’

- Most approved application are from category Consumer loans followed by Cash loans and Revolving loans.

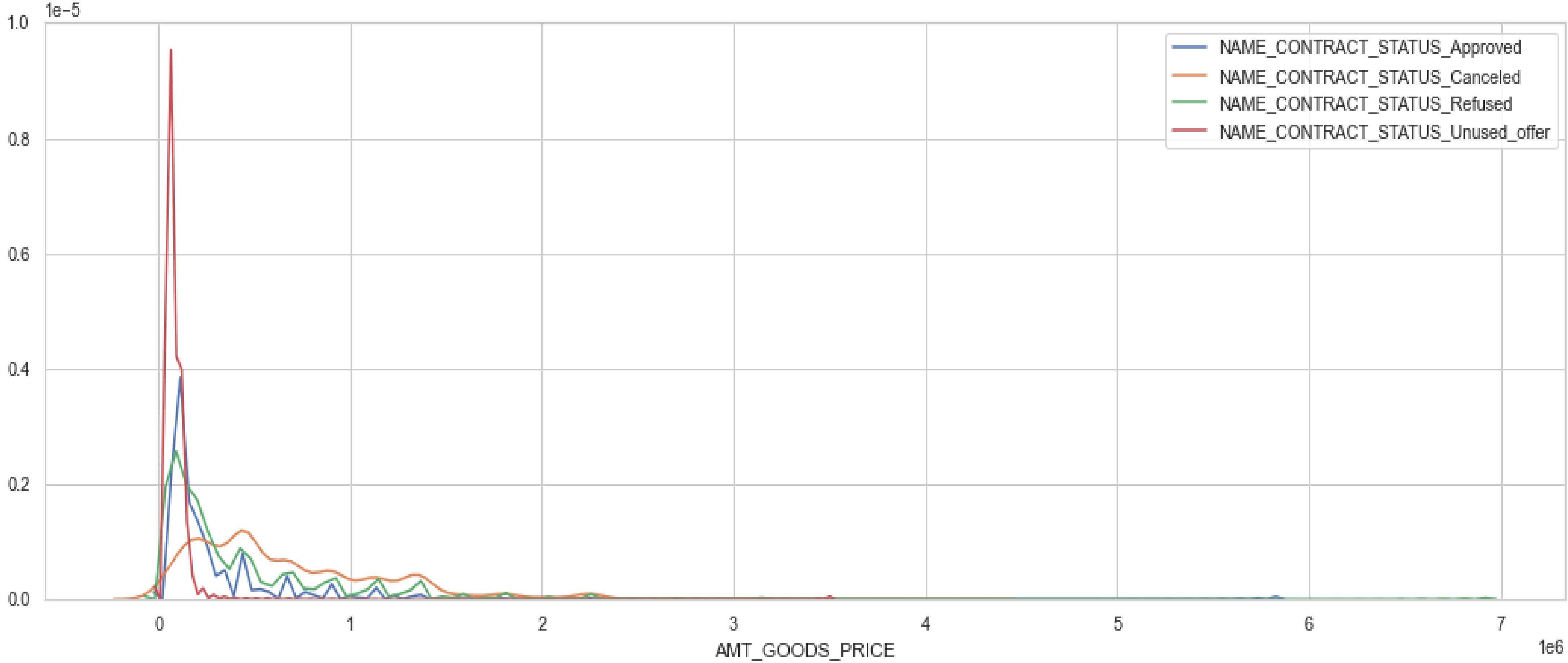
- Maximum application canceled and refused for a cash loans applications.

- Consumer loans have maximum unused offers followed by cash loans



Analysis–

a. Univariate analysis for numerical variables.

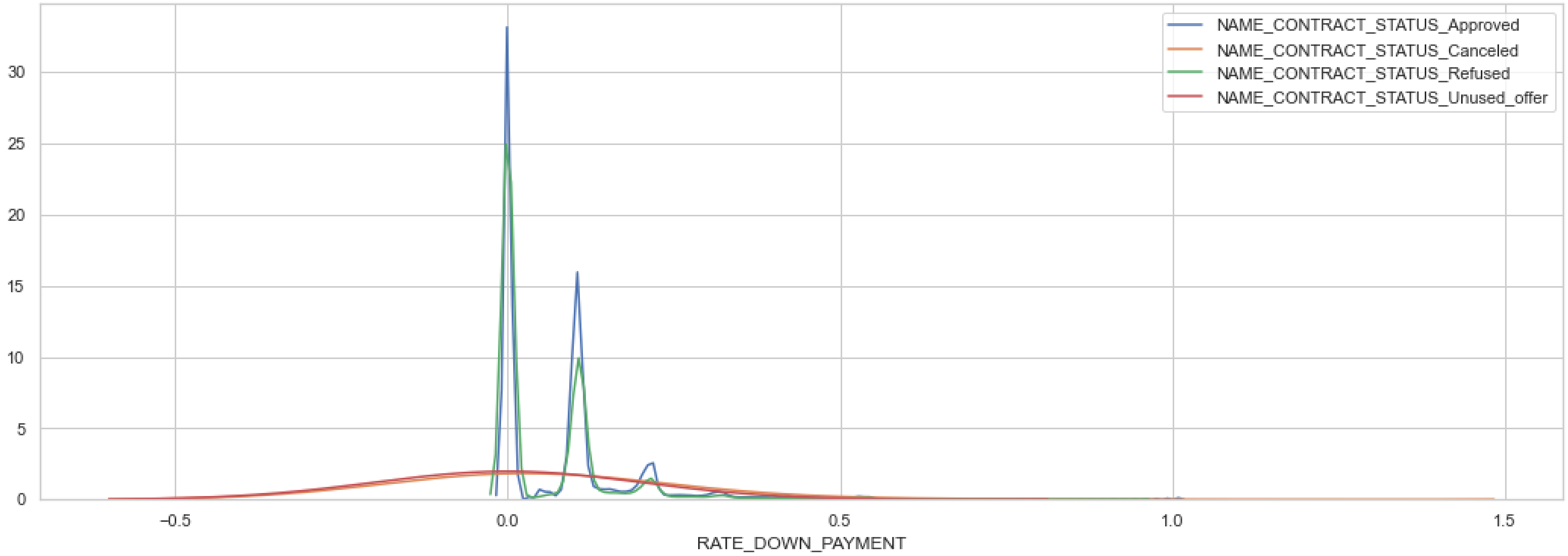


- From the graph we can see that for goods price less than 5 lakhs we have most unused offers.
- The rate of contract status being cancelled is more as compared to the status being approved.

More number of contracts are cancelled with goods price less than 10 lakhs however the rate of contract getting cancelled is more than getting approved or refused.

Analysis–

a. Univariate analysis for numerical variables.

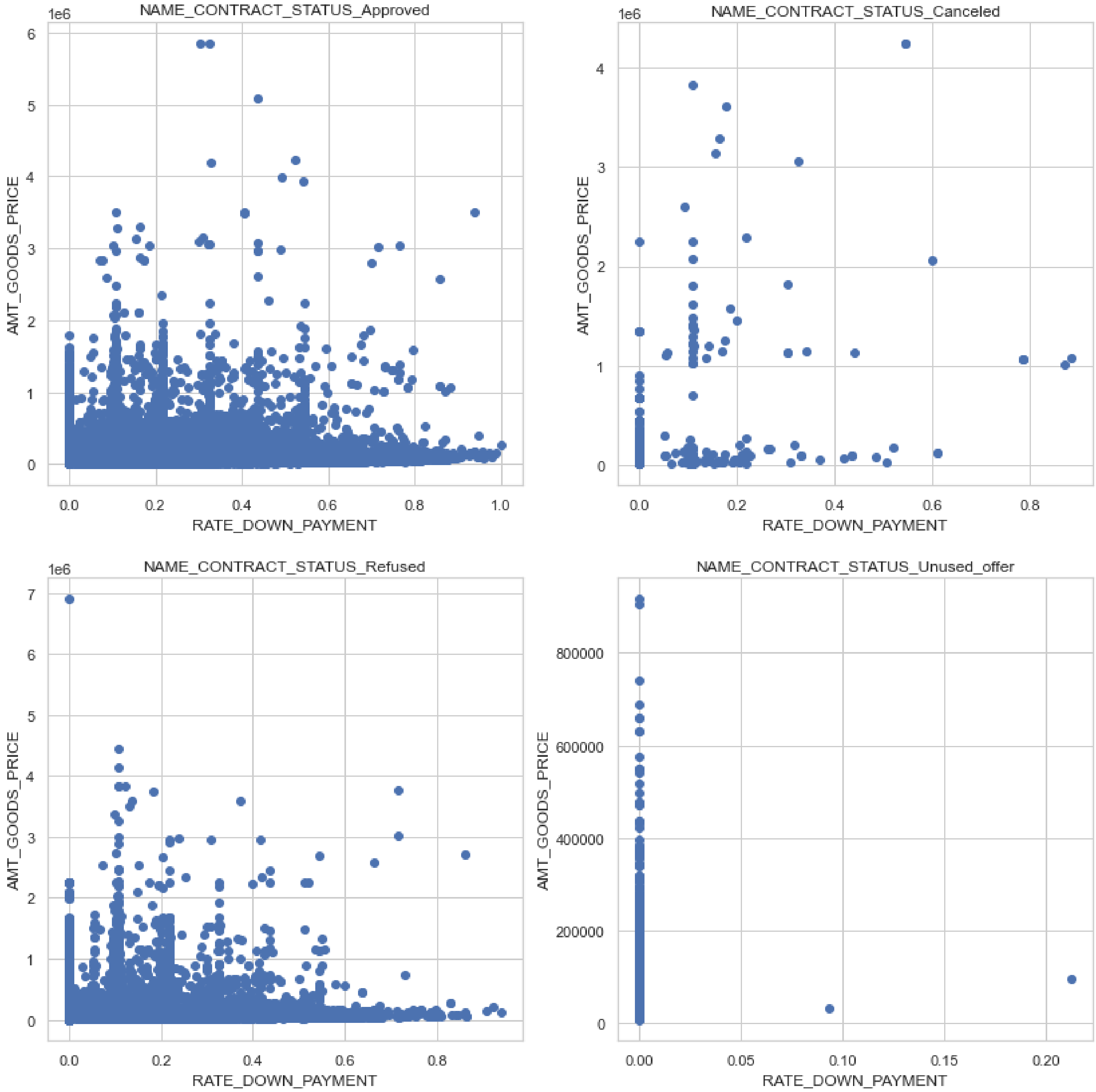


- From the above graph we can see that the chances of loan getting approved are more than it getting refused.
- More is the rate of down payment less is the chance of loan getting refused.

Analysis–

b. Bivariate analysis for continuous-continuous variables

- The chances of approving and refusing the clients application are likely to be same but mostly the application get approved on certain Rate of down payment.

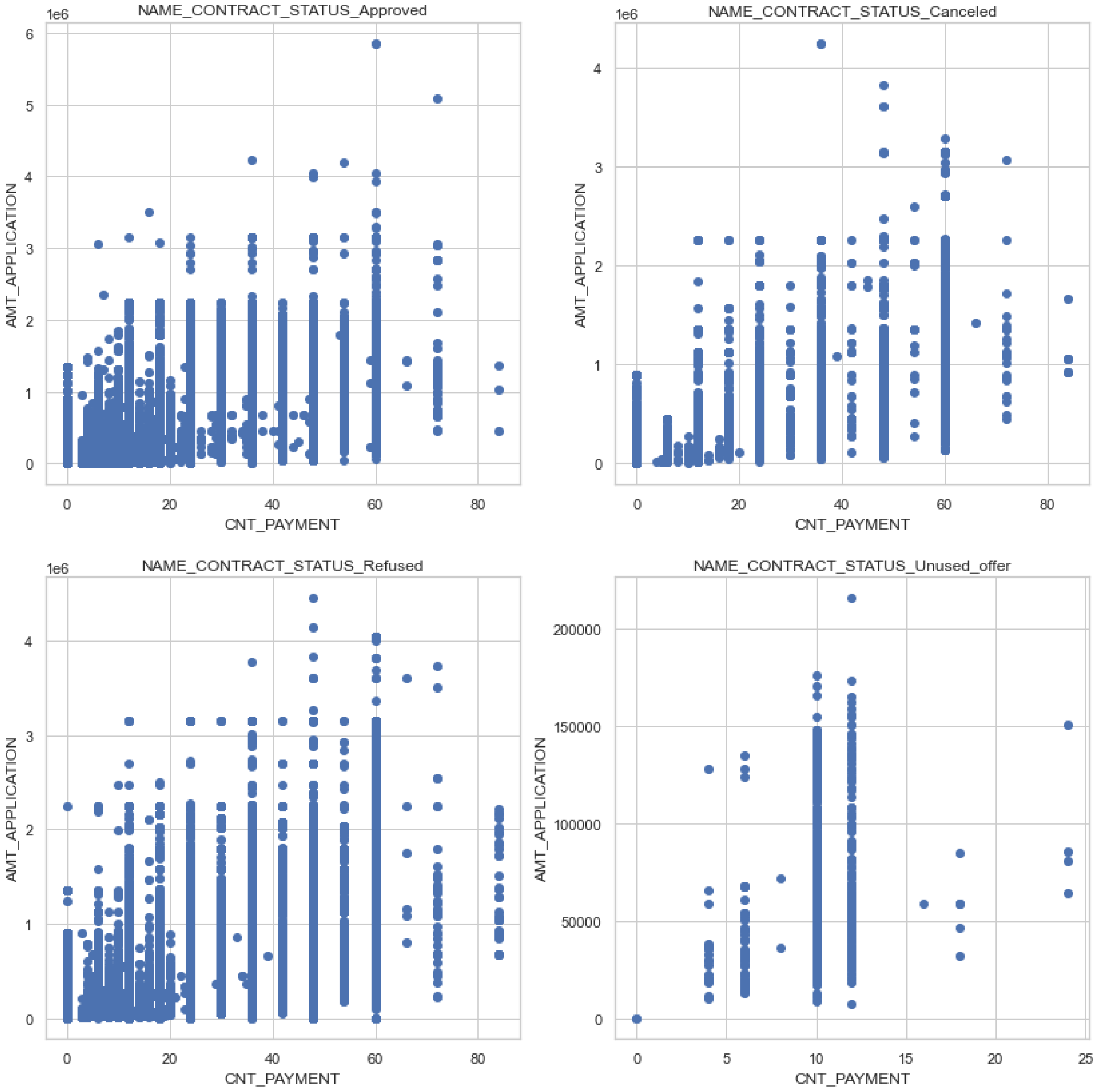


Analysis–

b. Bivariate analysis for continuous-continuous variables

- Applications with Term of previous credit 20, 40, 60 are more likely to approved.

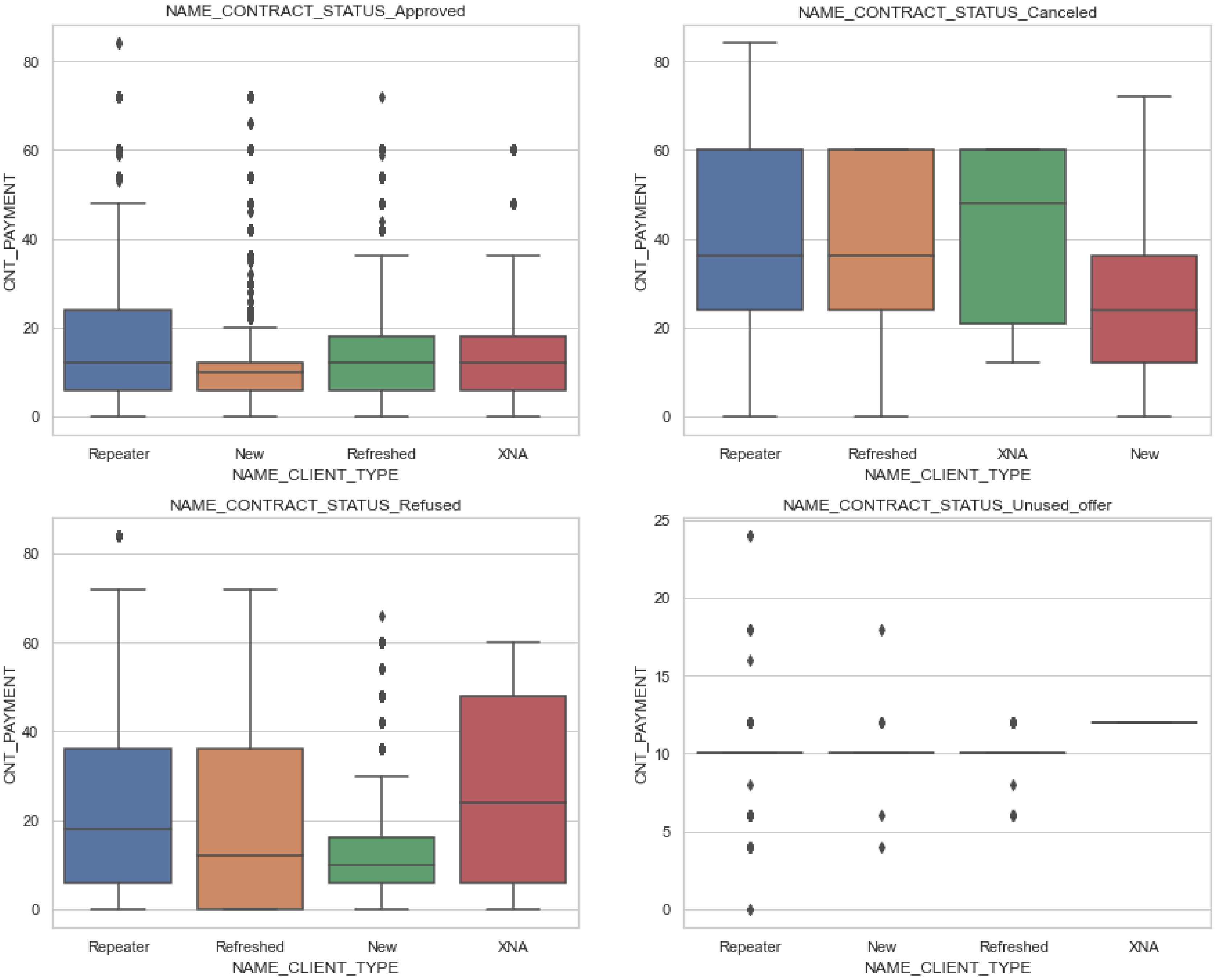
- Clients application in between 0 to 20 term are more likely to be approved and refused.



Analysis–

C. Bivariate analysis for continuous-categorical variables.

- Most of the previous applications have client type Repeater (~1M), just over 200K are New and ~100K are Refreshed.
- In terms of default percent for current applications of clients with history of previous applications, current clients with previous applications have values of percent of defaults ranging from 8.5%, 8.25% and 7% corresponding to client types in the past New, Repeater and Refreshed, respectively.



NAME_CONTRACT_STATUS_Approved

NAME_CONTRACT_STATUS_Canceled

NAME_CONTRACT_STATUS_Refused

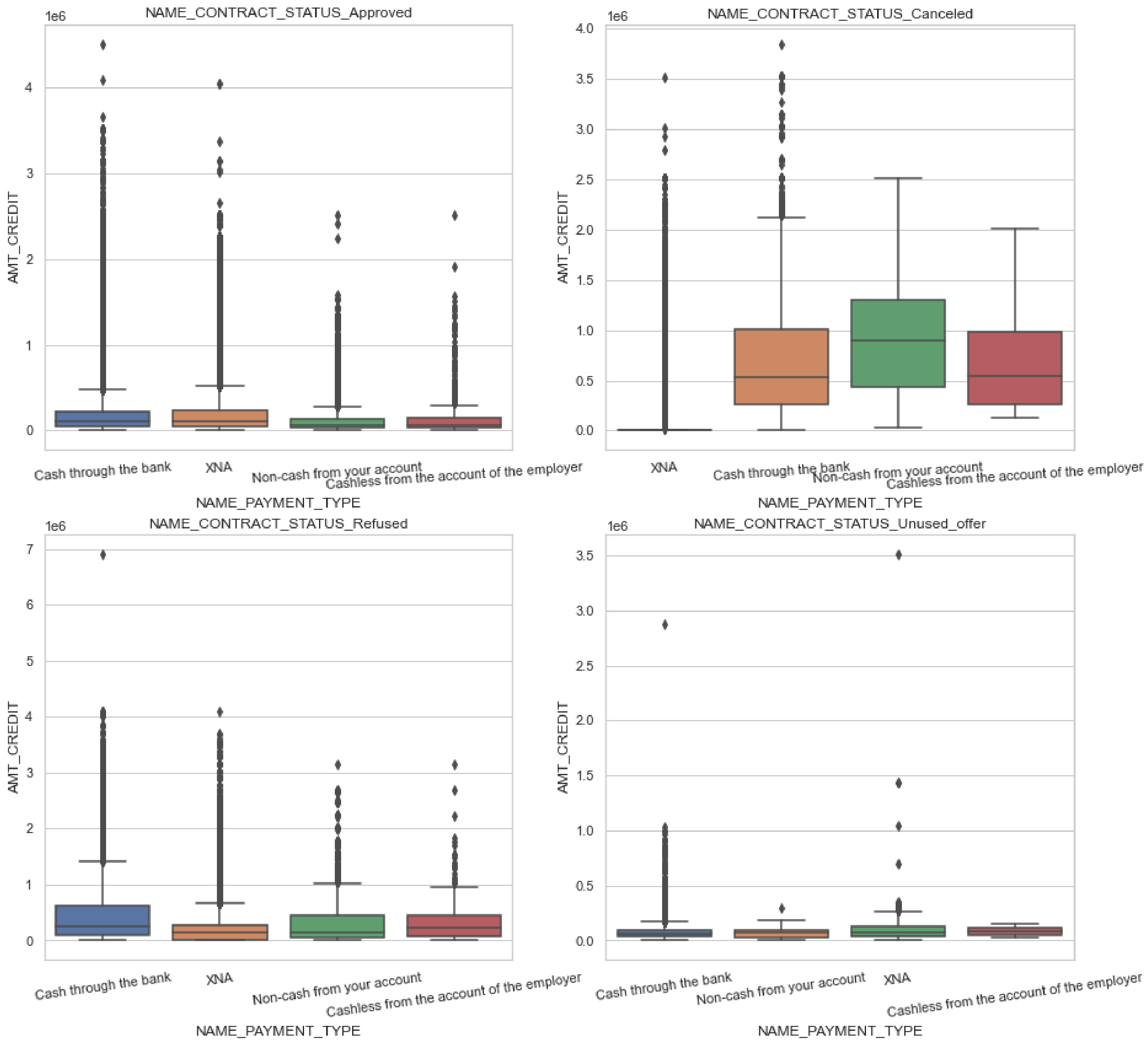
NAME_CONTRACT_STATUS_Unused_offer

Analysis–

C. Bivariate analysis for continuous-categorical variables.

•Most of the previous applications were paid with Cash through the bank (~850K).

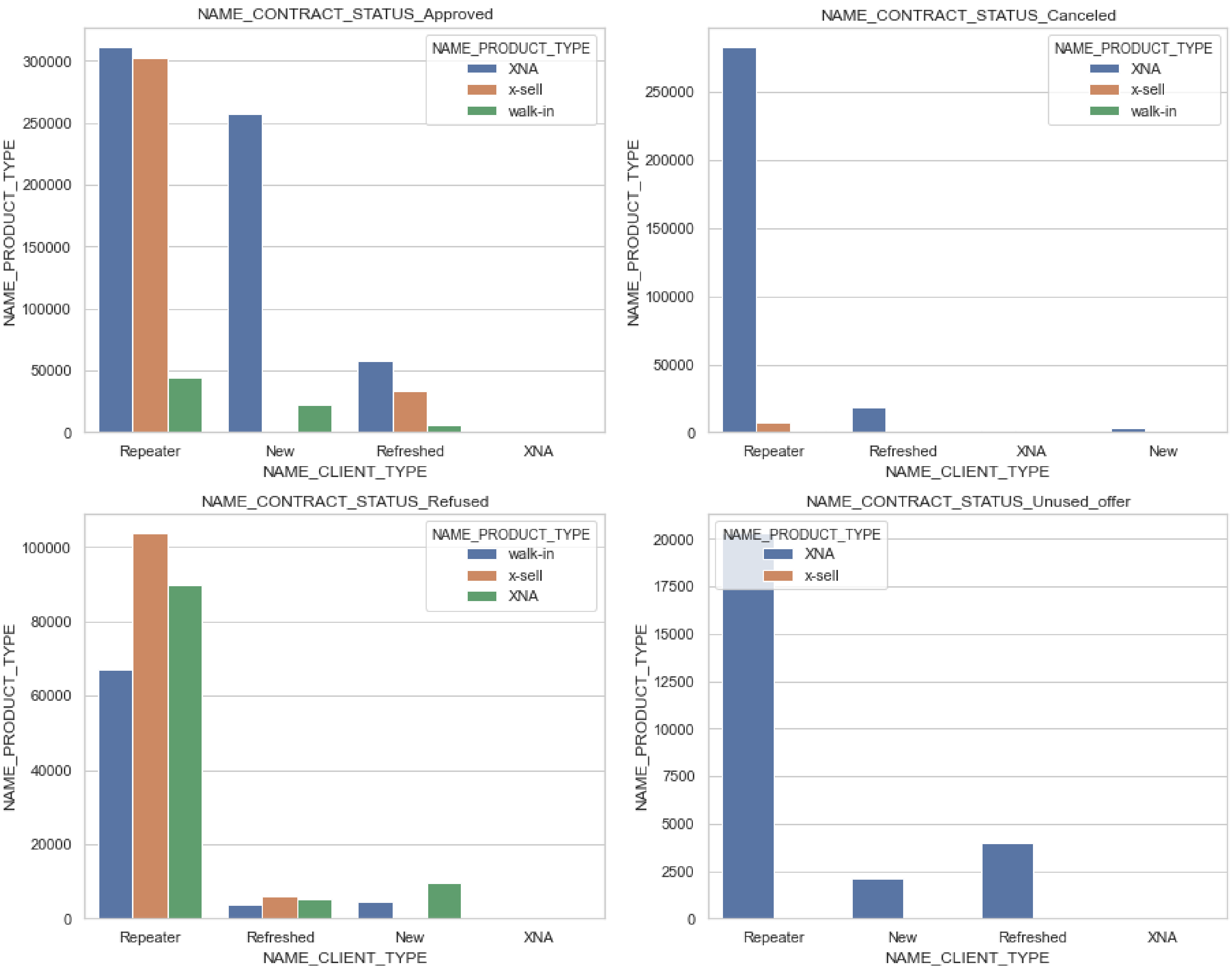
•Payments using Non-cash from your account or Cashless from the account of the employer are much rare. These three types of payments in previous applications results in almost the same percent of defaults for current clients (~8% each).



Analysis–

d. Bivariate analysis for categorical-categorical variable.

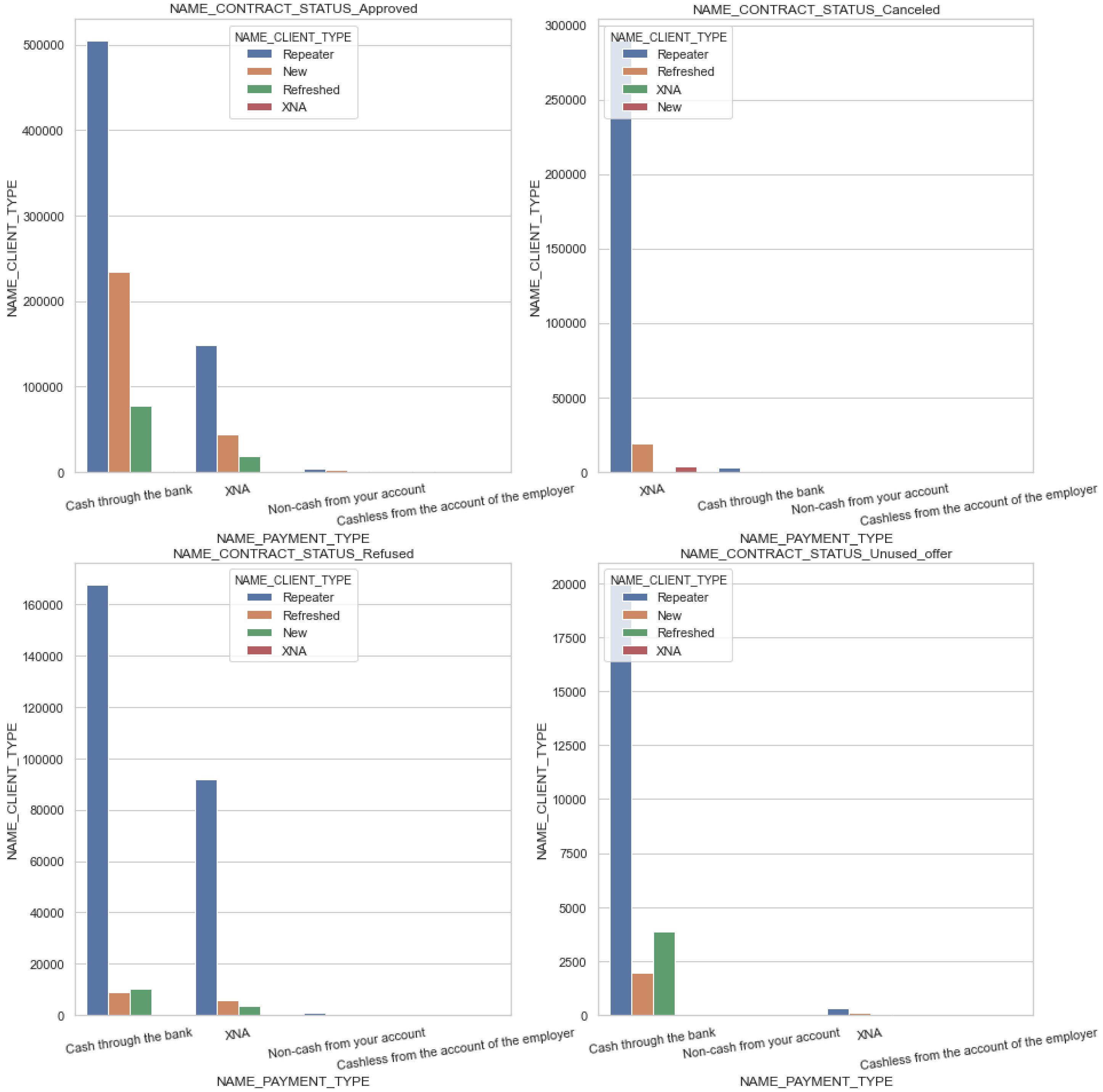
- The repeater applicant has maximum approved rate.
- Most of the previous applications have client type Repeater (~1M), just over 200K are New and ~100K are Refreshed.
- In terms of default percent for current applications of clients with history of previous applications, current clients with previous applications have values of percent c defaults ranging from from 8.5%, 8.25% and 7% corresponding to clien types in the past New, Repeater and Refreshed, respectively



Analysis–

d. Bivariate analysis for categorical-categorical variable.

- Repeater applicant has payment type as cash through the bank maximum time.
- Most of the previous applications were paid with Cash through the bank (~850K).
- Payments using Non-cash from your account or Cashless from the account of the employer are much rare. These three types of payments in previous applications results in allmost the same percent of defaults for current clients (~8% each).



Summary

Final words

Based on analysis, defining the results and conclusion.

- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.
- The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.
- In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest
- The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% not-repayment rate.
- Contract type Revolving loans are just a small fraction (10%) from the total number of loans; in the same time, a larger amount of Revolving loans, comparing with their frequency, are not repaid.
- The clients that owns a car are almost a half of the ones that doesn't own one. The clients that owns a car are less likely to not repay a car that the ones that own. Both categories have not-repayment rates around 8%.
- The clients that owns real estate are more than double of the ones that doesn't own. Both categories (owning real estate or not owning) have not-repayment rates less than 8%.
- Maximum clients fail to repay the loan if they apply for the Cash Loan and has client type as Unaccompanied. most defaulter are from this category.
- Bank should focus on client who has income type as "Working" they are more likely to pay the loans
- Bank should focus on client who have higher education as they are less likely to be a defaulter.