CREDIT EDA CASE STUDY

PROBLEM STATEMENT:

We have data which contains information about the loan application at the time of applying for the loan. It contains two types of scenarios:

→The client with payment difficulties: he/she had a late payment of more than X days on at least one of the first Y instalments of the loan in our sample,

→All other cases: All other cases when the payment is paid on time.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

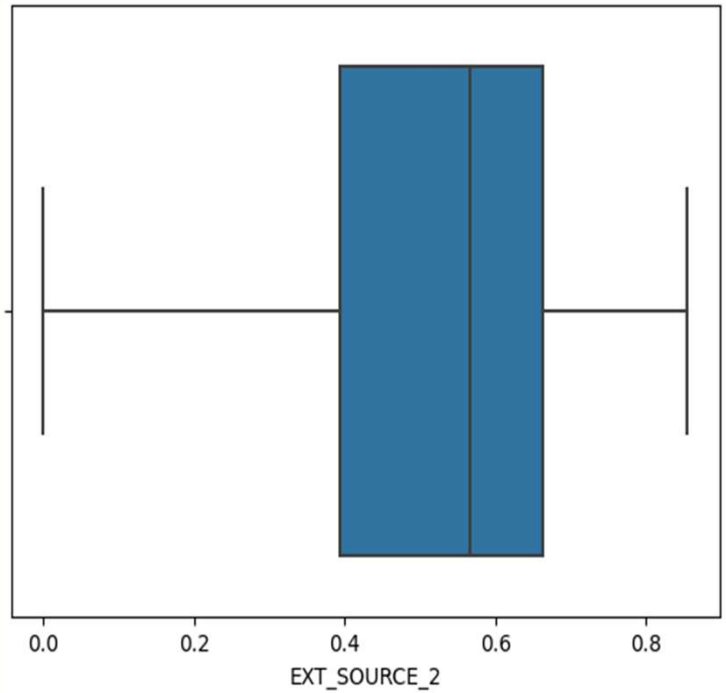
Hence, In this presentation we look into some of the factors which are the cause of default and what can done to avoid the risk for the bank.

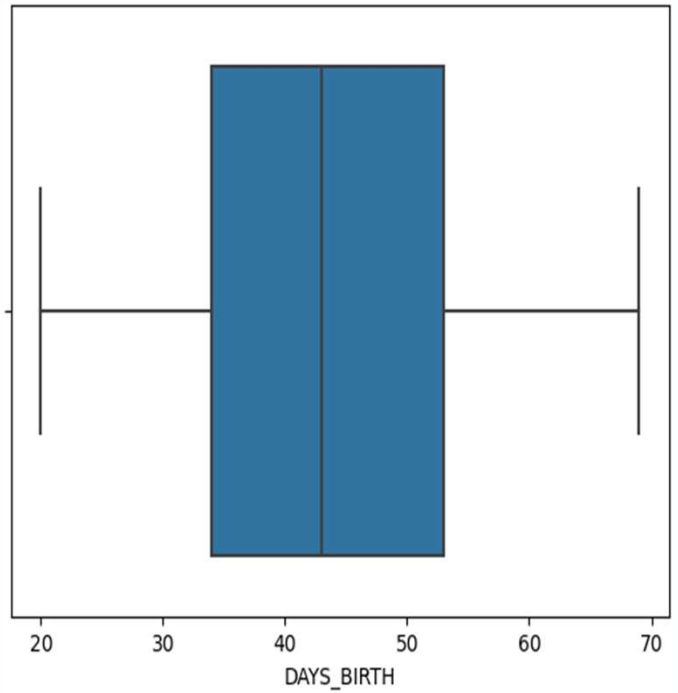
THE STEPS WE FOLLOW TO ANALYSE THE DATA:

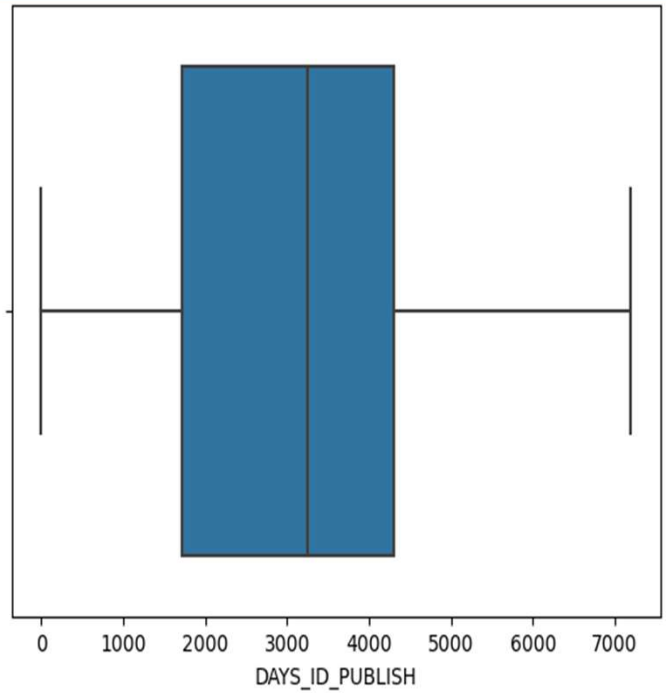
- We first import the dataset to jupyter notebook to analyse the data using Python programming language.
- The dataset has 122 columns and 307511 rows. Hence.
- first we check for missing values in the dataset and remove the columns with missing values more than 40 percent.
- Now we replace the values in columns with missing values less than 15 percent with standardised value(median or mode).
- As some of the columns have negative values which are not valid hence let's change the values to positive values.
- There are also some columns where some values are mentioned as 'XNA' which means 'Not Available. We correct these values accordingly.

- Assuming certain values based on data available and replacing the values accordingly.
- Using binning to make continuous data easier to analyse.
- Removing rarely used or unwanted columns for analysis.
- After Cleaning and preparation of data.
- We plot boxplot of numeric columns to find out outliers of the data to gain new insights in data.
- Calculate the Imbalance of the data available.
- We than perform Univariate and Bivariate analysis on merging Current application and previous application to derive results and reach conclusion.
- Assumptions made:So from the data we analysed, we can conclude that Pensioner value is approximately equal to null values in ORGANIZATION_TYPE column. So the value is not Missing At Random. Hence, imputing null values of OCCUPATION_TYPE with 'Pensioner' as most of the null values for OCCUPATION_TYPE we found is also ‘Pensioner’ for most frequent null values, will give us insights into dataset.

OUTLIER ANALYSIS: Below are some of the graphs from the outlier analysis we performed on the numeric columns to gain insights on individual columns.

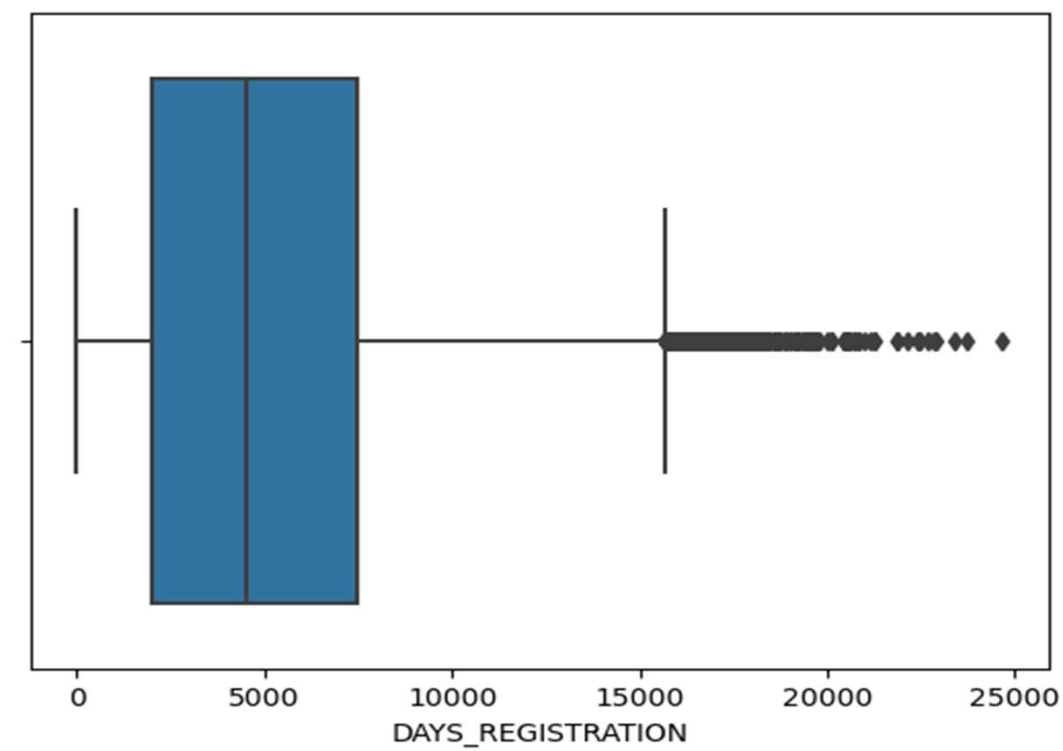


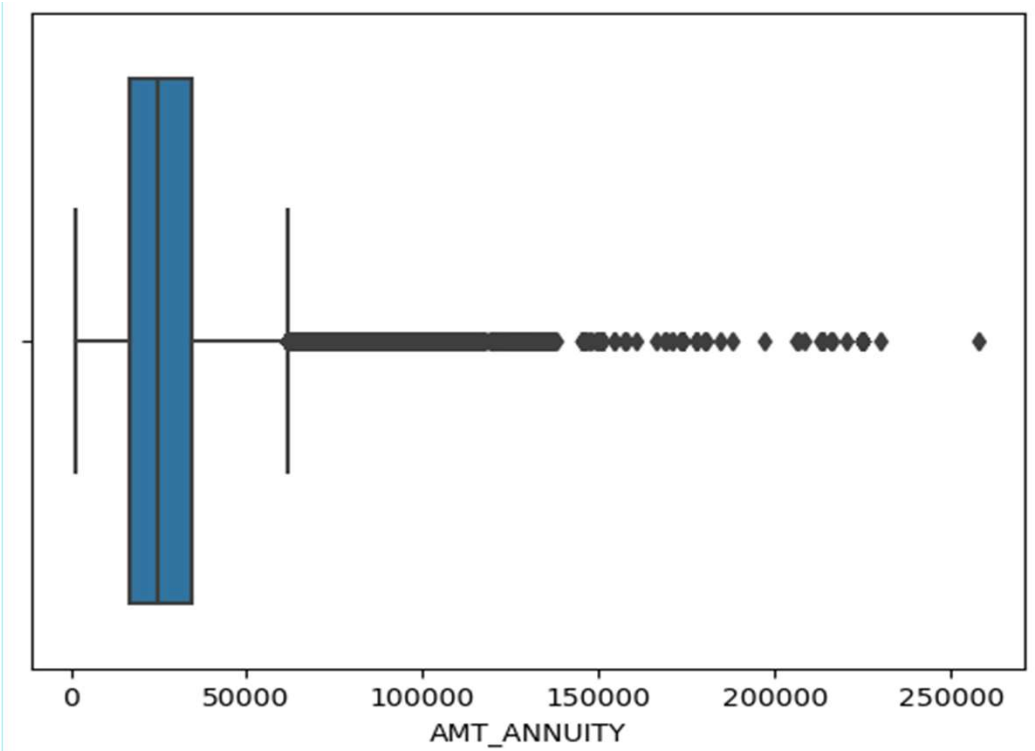




Insights from the graph:

→'EXT_SOURCE_2','DAYS_BIRTH' and 'DAYS_ID_PUBLISH’ have no outliers, hence we have perfect data for these columns.

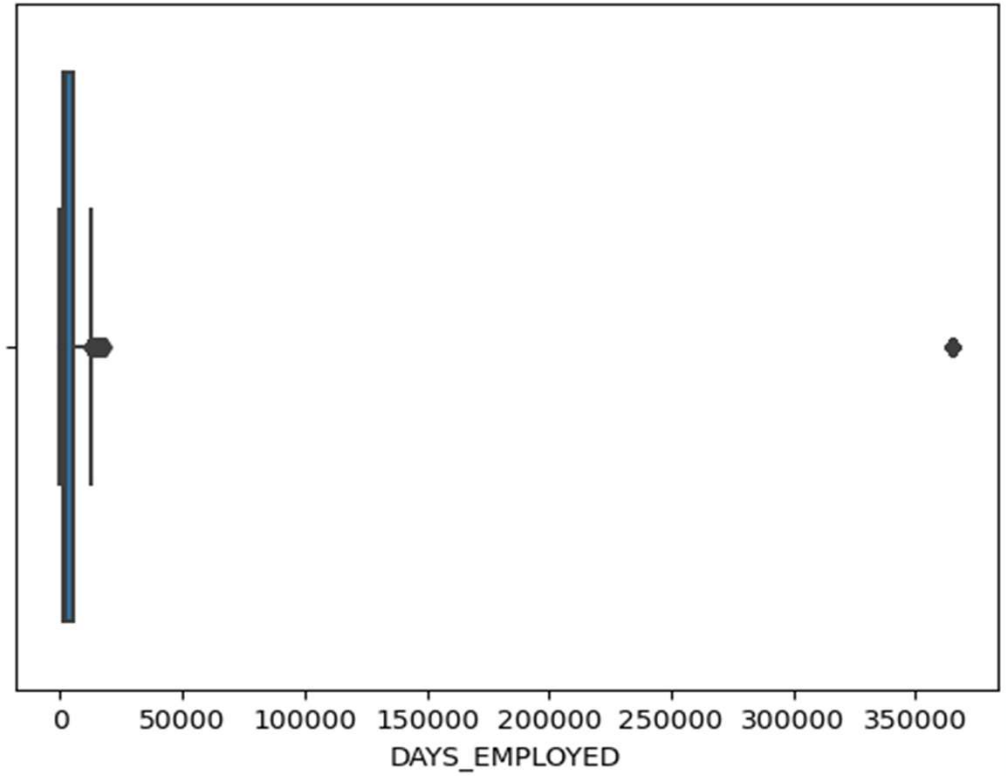


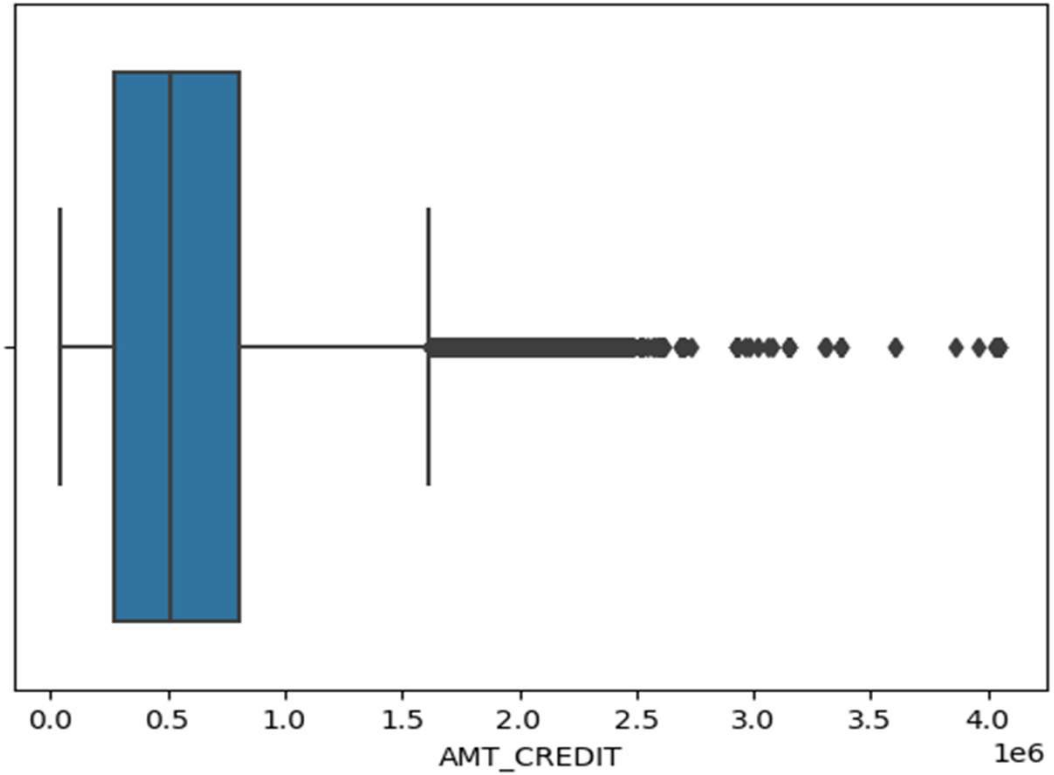


Insights from the graph:

→'DAYS_REGISTRATION' is larger as compared to the First quartile and has large number of outliers.

→The third quartile of 'AMT_ANNUITY' is slightly larger than the First quartile and there is a large number of outliers which means most of the annuity clients are from third quartile.

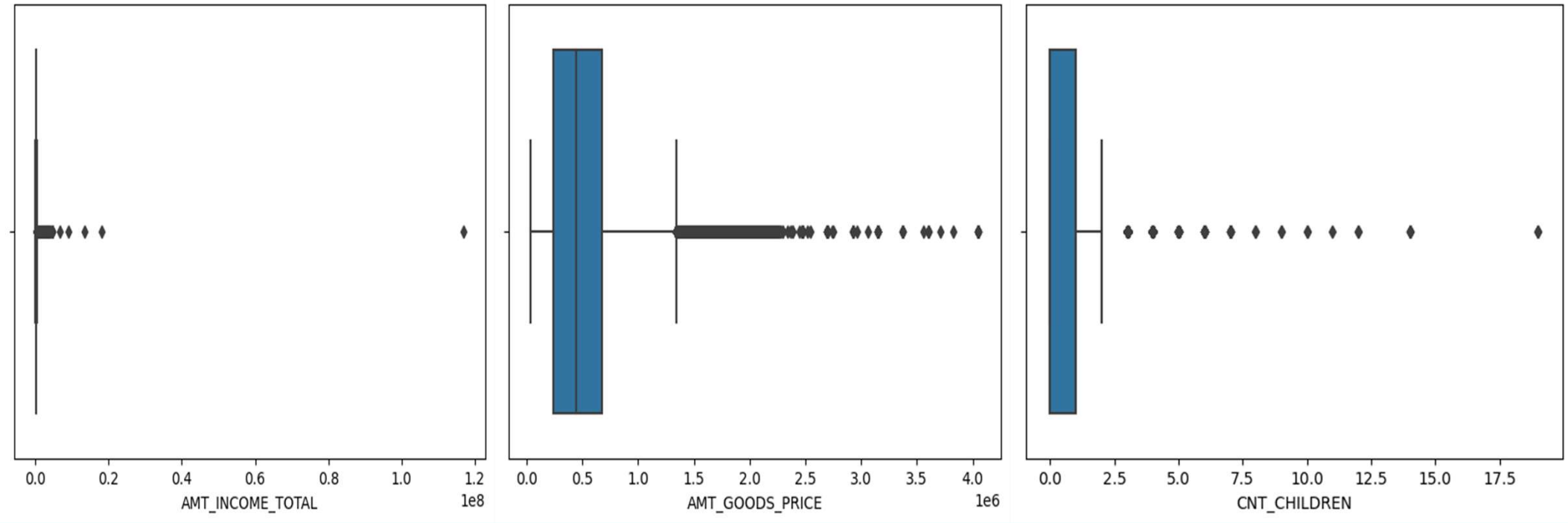




Insights from the graph:

→In 'DAYS EMPLOYED' some outliers are present below 25000 and a outlier is present above 350000.

→Third quartile of AMT_CREDIT is larger as compared to the First quartile which means that most of the Credit amount of the loan customers are present in the third quartile and there are large number of outliers present in AMT_CREDIT.



Insights from the graph:

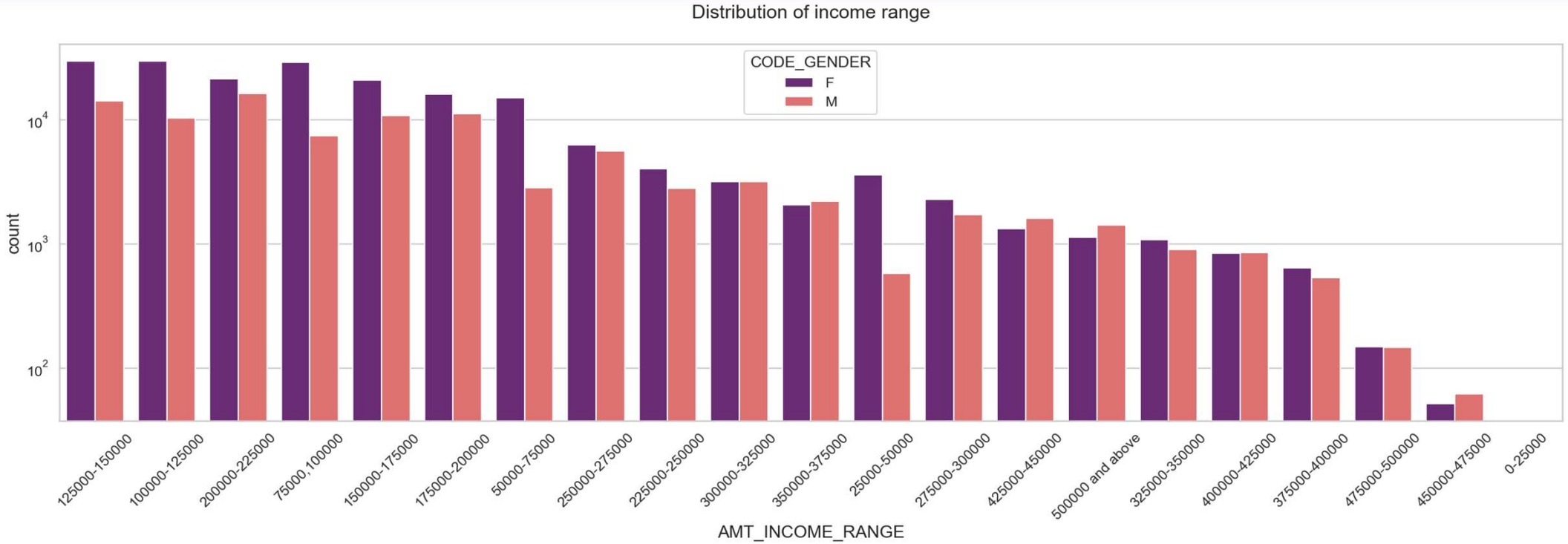
→AMT_INCOME_TOTAL' has slim Inter Quartile Range but has Outliers.

→'AMT_GOODS_PRICE's third quartile is slightly larger than first quartile with large number of outliers.

→Outliers in 'CNT_CHILDREN's show that client has more than 5 children.

CATEGORICAL UNIVARIATE ANALYSIS FOR TARGET 0 (CUSTOMERS WITH NON-PAYMENT DIFFICULTIES)

DISTRIBUTION OF INCOME RANGE:



Insights from the graph:

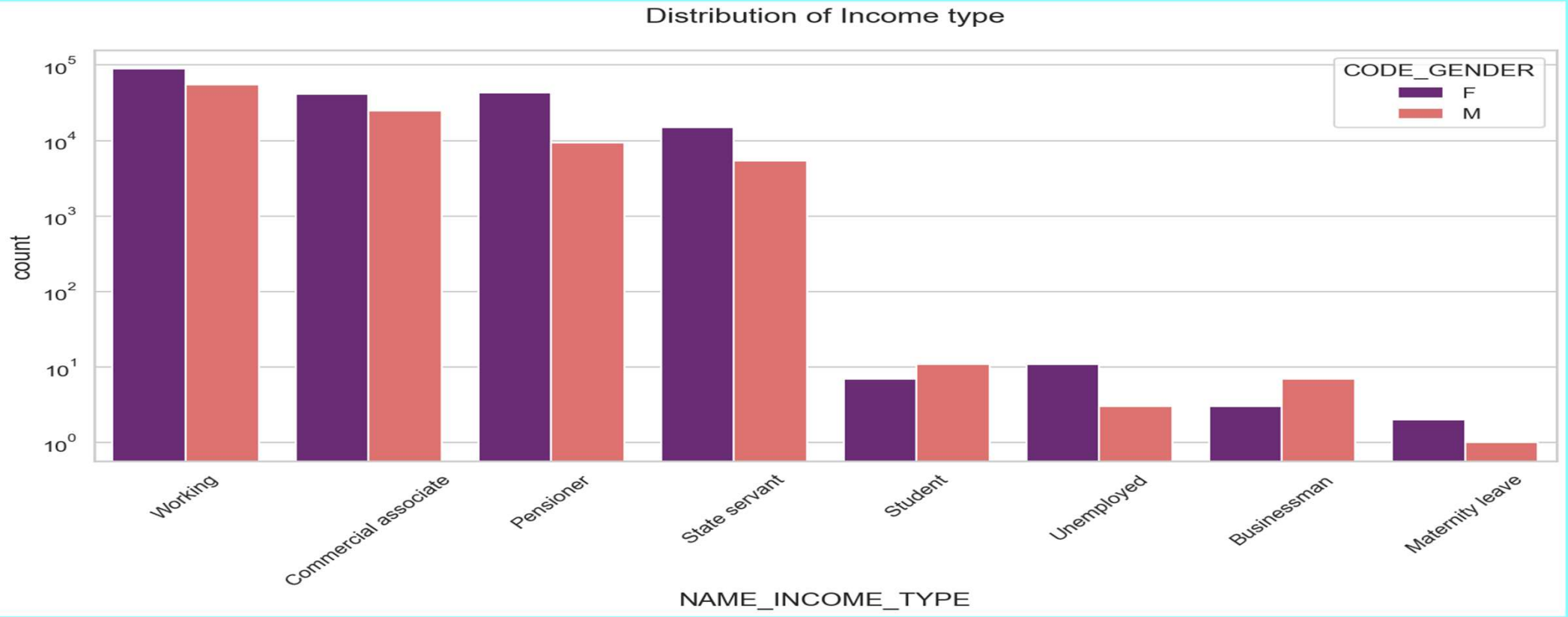
→Female counts are higher than male.

→Income range from 100000 to 200000 is having more number of credits.

→This graph show that females are more than male in having credits for that range.

→Very less count for income range 400000 and above.

DISTRIBUTION OF INCOME TYPE

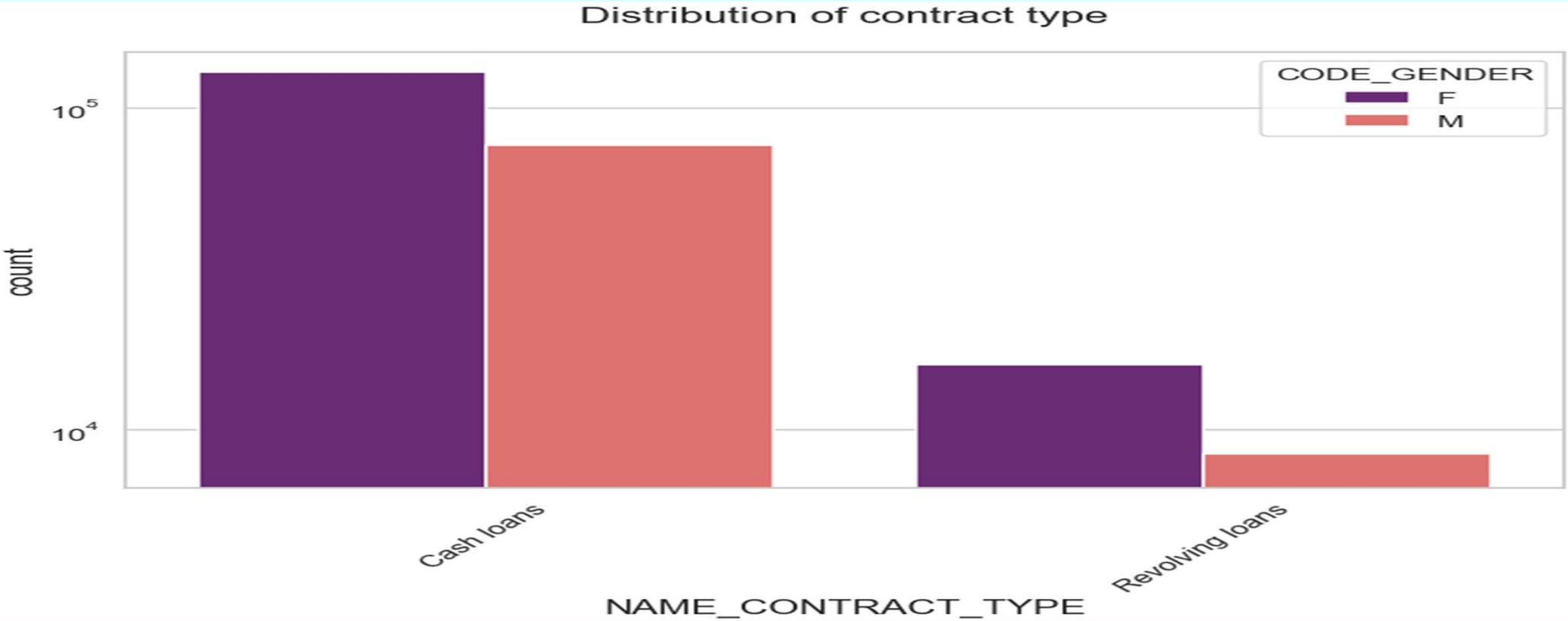


Points to be concluded from the graph on the right.

→For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than others. And Females customers have high credits compared to male customers.

→Less number of credits for income type ‘student’ ,’pensioner’, ‘Businessman’ and ‘Maternity leave’.

DISTRIBUTION FOR CONTRACT TYPE

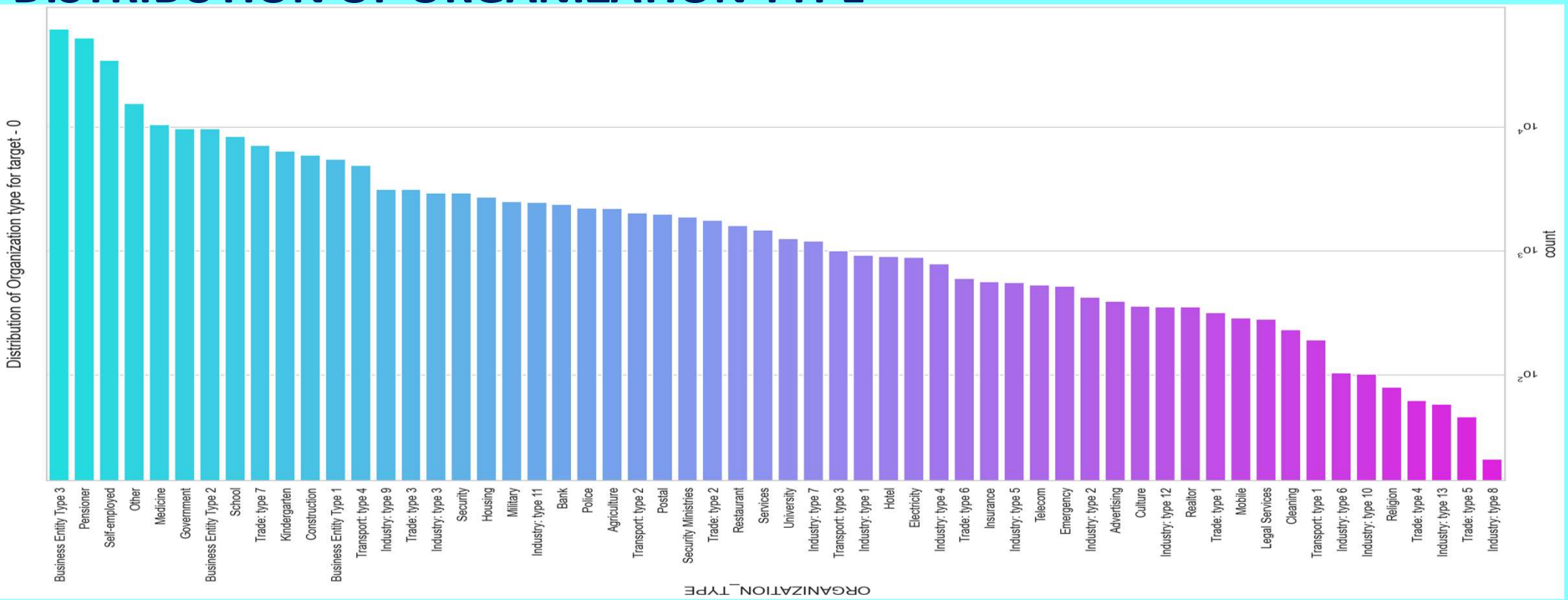


Insights from graph:

→For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.

→For this also Female is leading for applying credits.

DISTRIBUTION OF ORGANIZATION TYPE



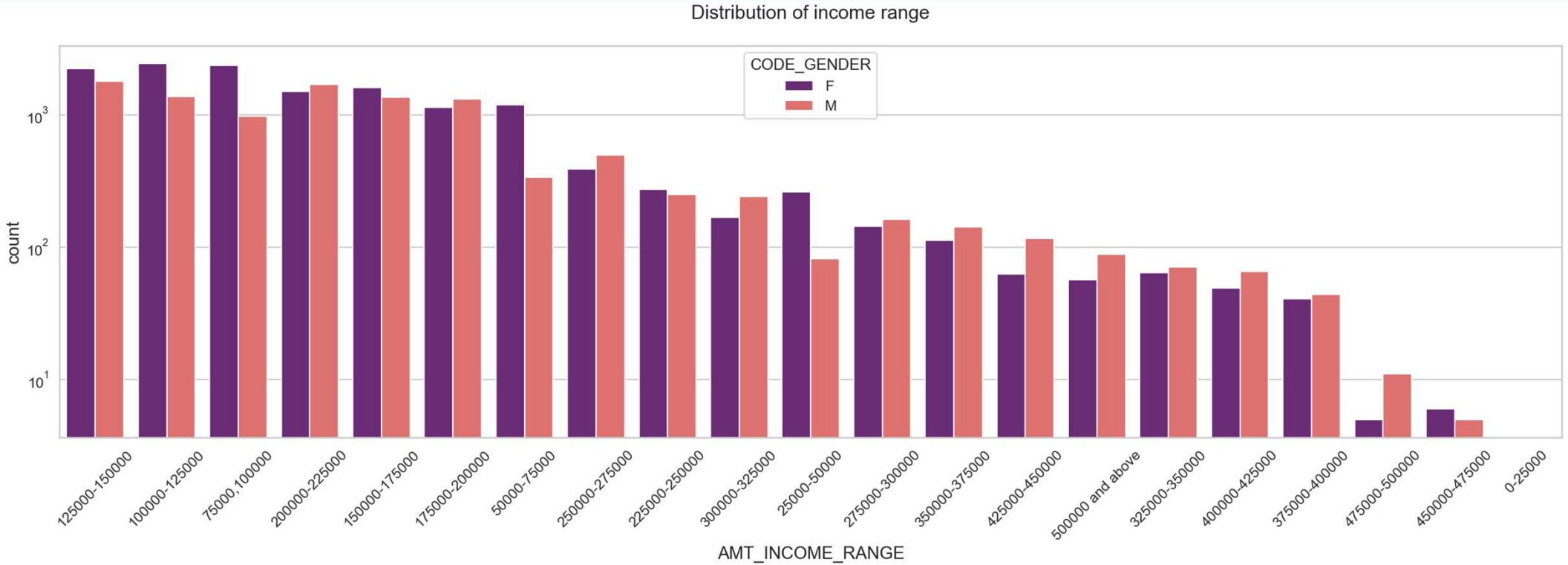
Insights from graph:

→Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.

→Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

CATEGORICAL UNIVARIATE ANALYSIS FOR TARGET 1 (CUSTOMERS WITH PAYMENT DIFFICULTIES)

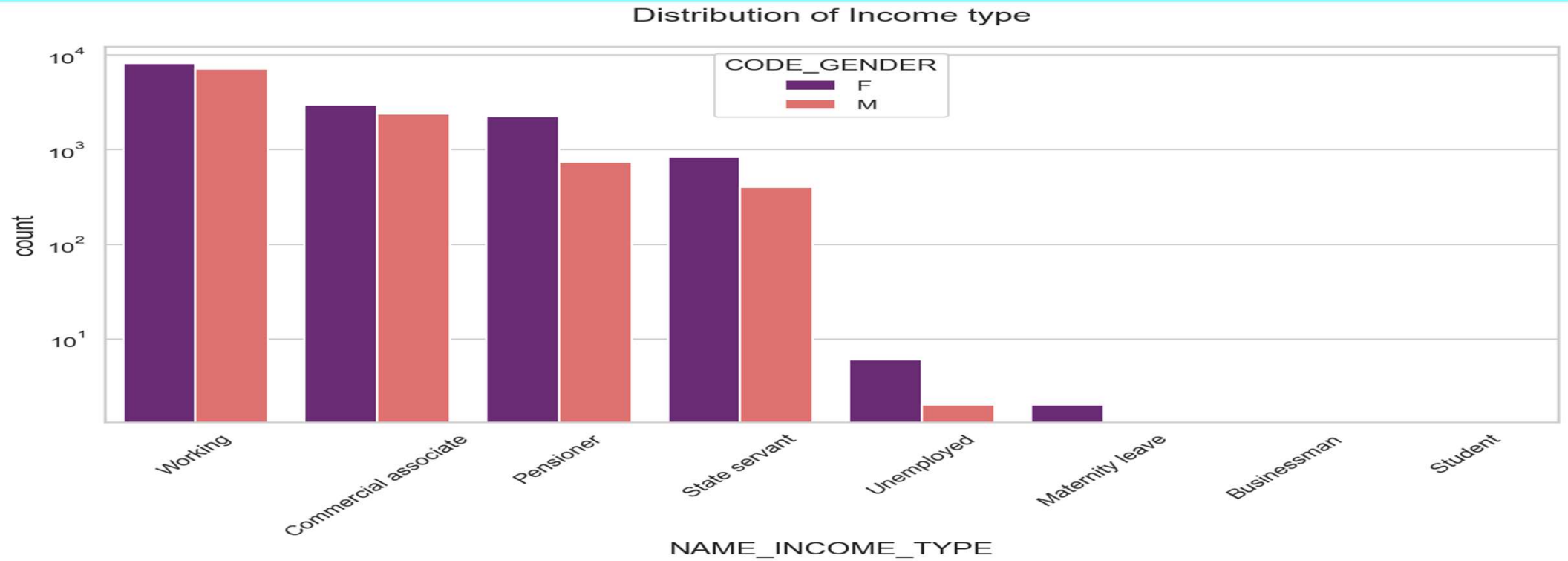
DISTRIBUTION OF INCOME RANGE



Insights from graph:

- Male counts are higher than female.
- Income range from 100000 to 200000 is has high number of credits.

DISTRIBUTION OF INCOME TYPE

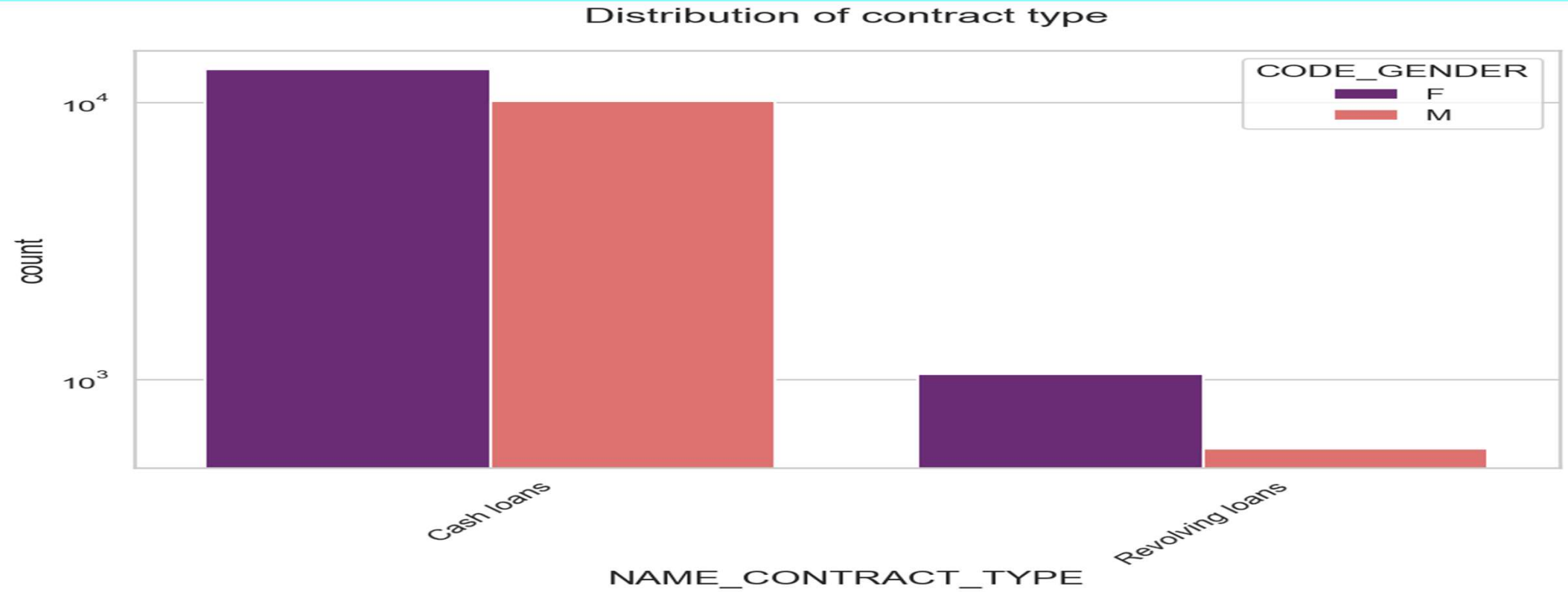


Insights from the graph:

→For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave and For Females there are having more number of credits than male.

→Less number of credits for income type ‘Maternity leave’ and For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.

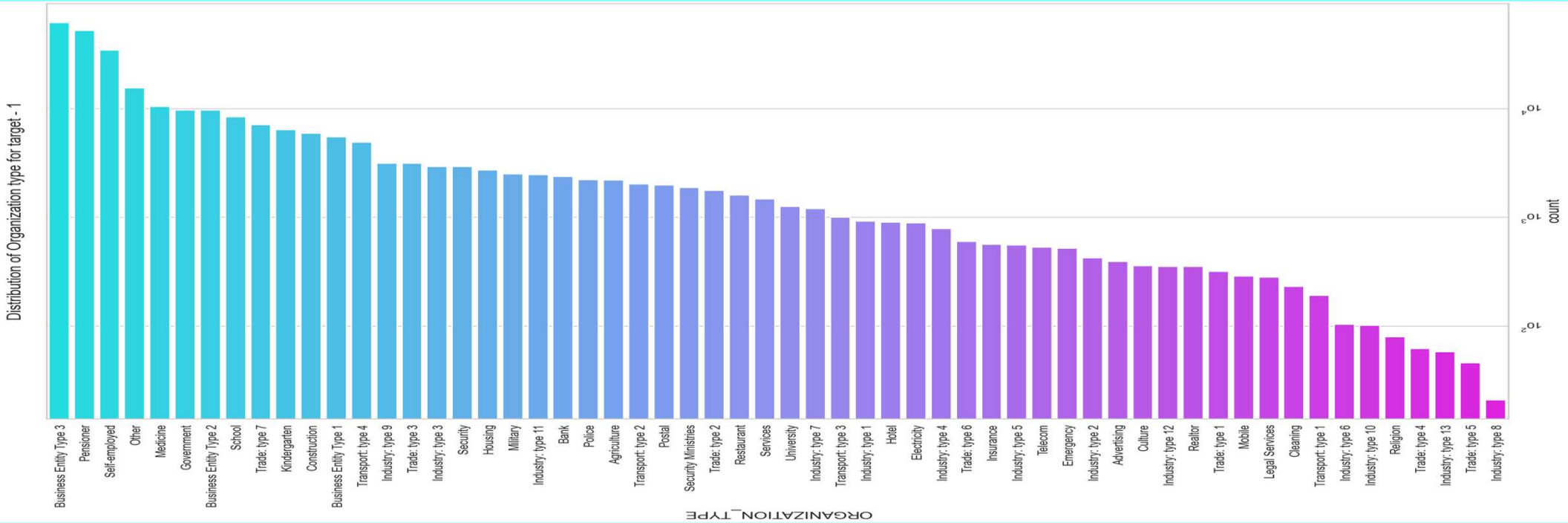
DISTRIBUTION FOR CONTRACT TYPE



Insights from graph:

→For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.

DISTRIBUTION OF ORGANIZATION TYPE

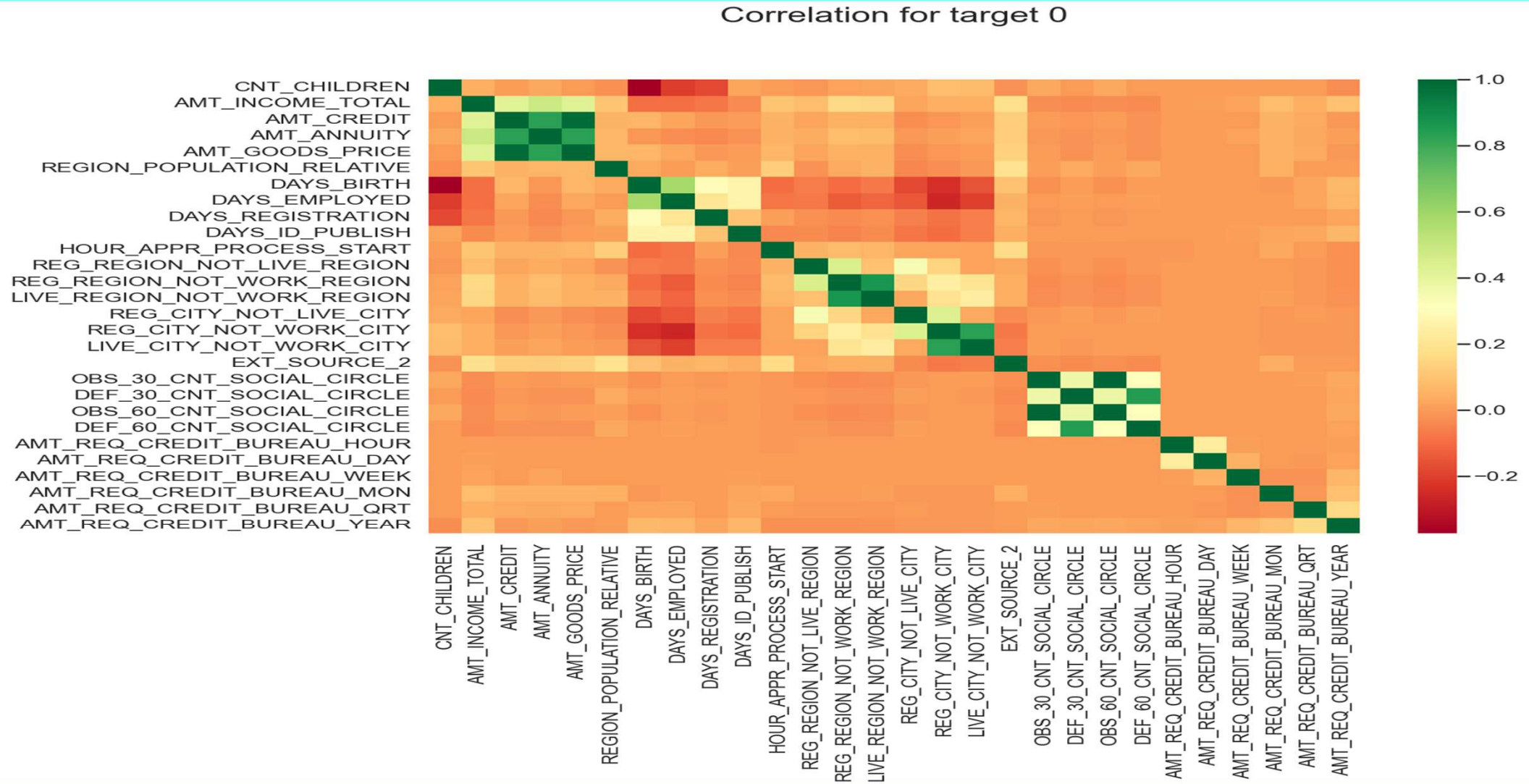


Insights from graph:

→Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Medicine’ and ‘Government’.

→Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4,Same as type 0 in distribution of organization type.

CORRELATION OF TARGET 0



CORRELATION FOR TARGET 0

Insights from graph:

→Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.

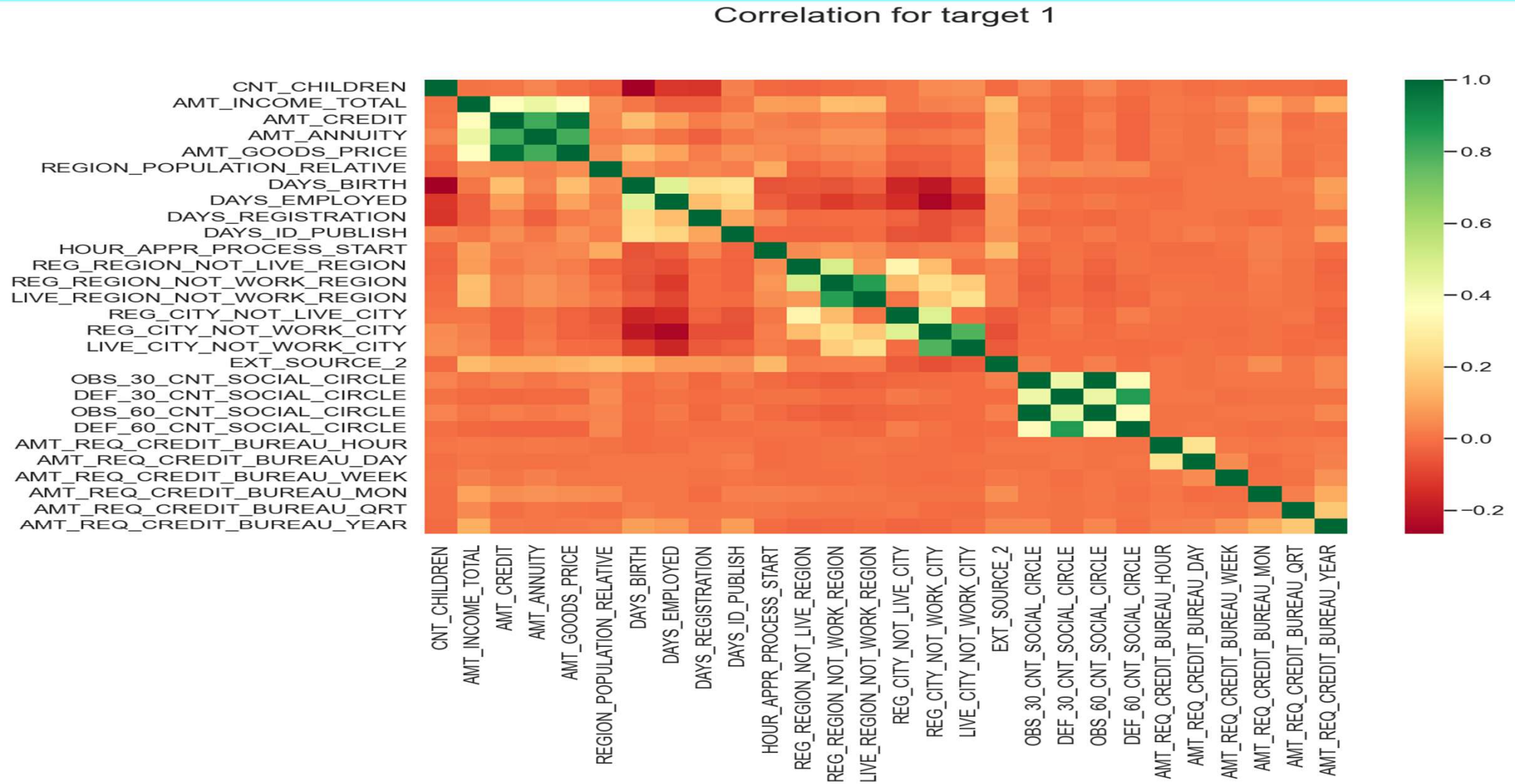
→Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.

→Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.

→Credit amount is higher to densely populated area.

→The income is also higher in densely populated area.

CORRELATION OF TARGET 1



CORRELATION OF TARGET 1 (CUSTOMERS WITH NON-PAYMENT DIFFICULTIES)

Insights from graph:

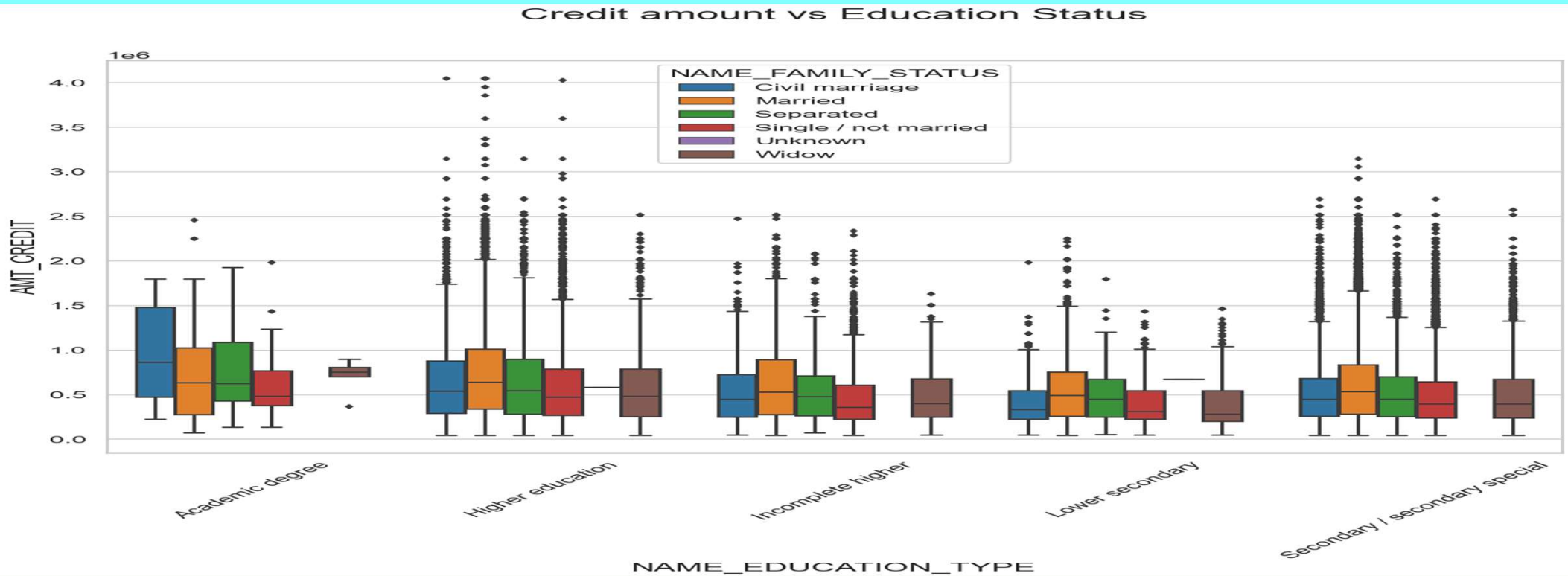
This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different:

→The client's permanent address does not match contact address are having less children and vice-versa

→The client's permanent address does not match work address are having less children and vice-versa

BIVARIATE ANALYSIS FOR TARGET 0

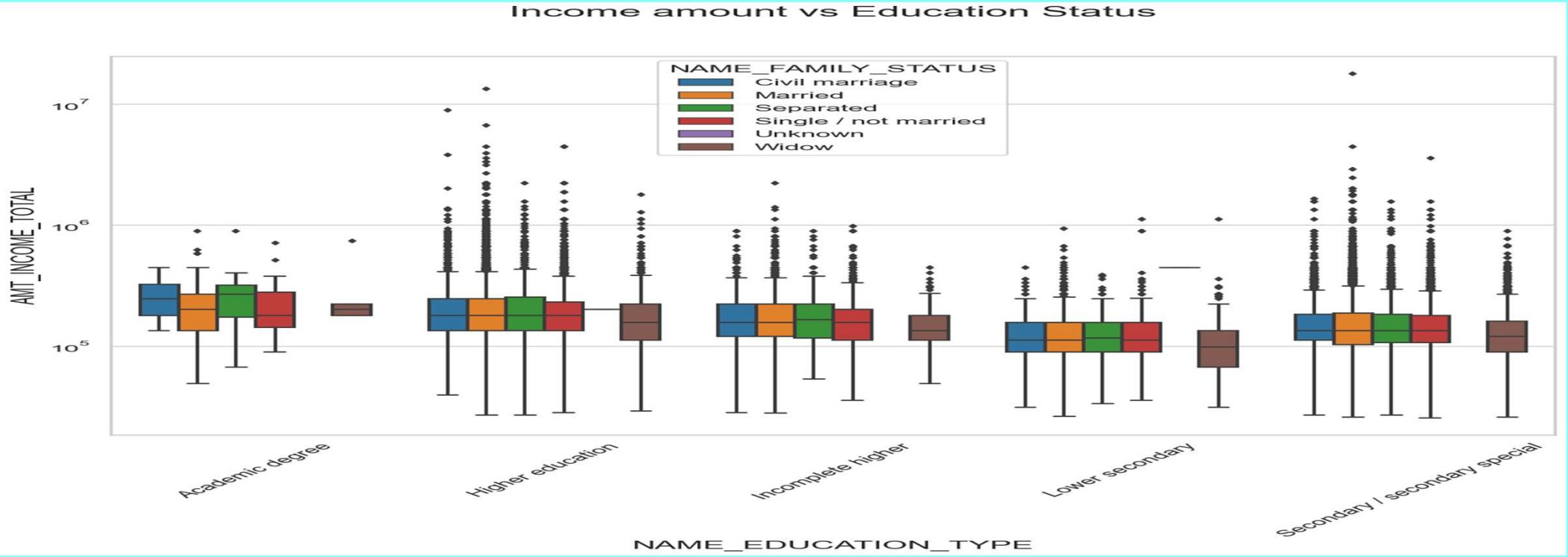
CREDIT AMOUNT VS EDUCATION STATUS



Insights from the graph.

- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.

INCOME AMOUNT VS EDUCATION STATUS



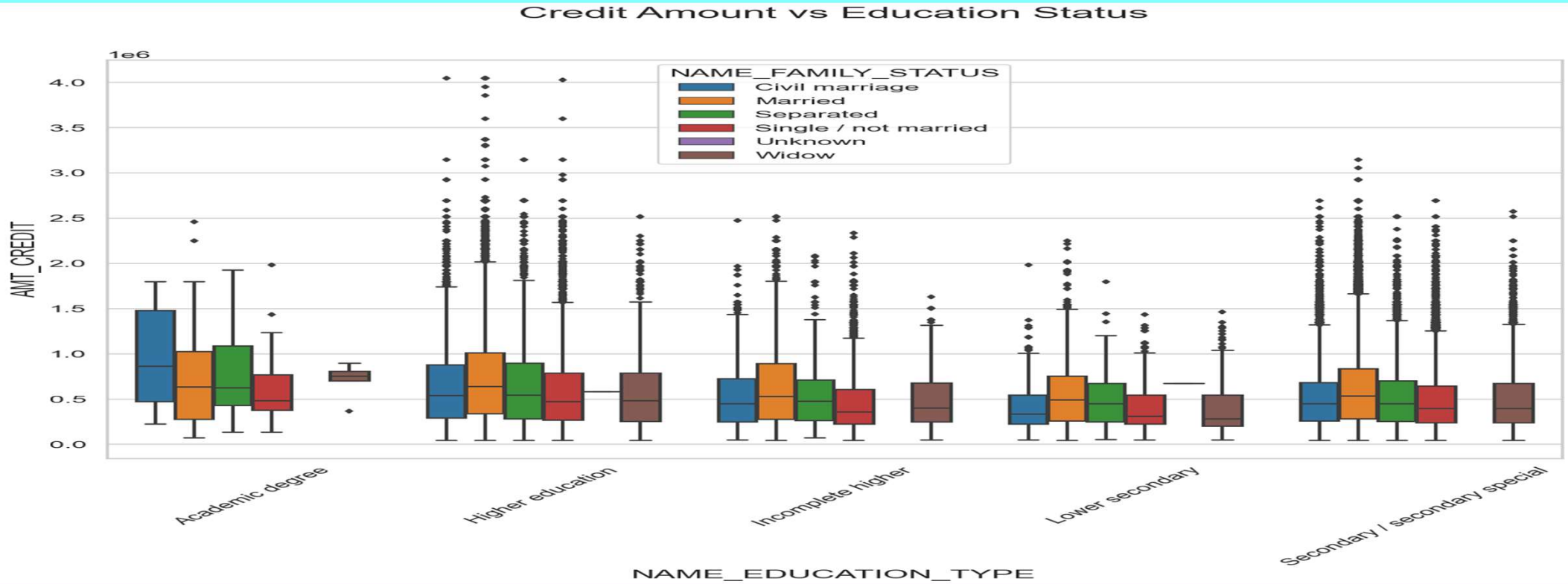
Insights from the graph:

→For Education type 'Higher education', income amount mean is mostly equal with family status. It does contain many outliers.

→Academic degree has less outliers but their income amount is little higher that Higher education.Lower secondary of civil marriage family status have less income amount than others.

BIVARIATE ANALYSIS FOR TARGET1

CREDIT AMOUNT VS EDUCATION STATUS



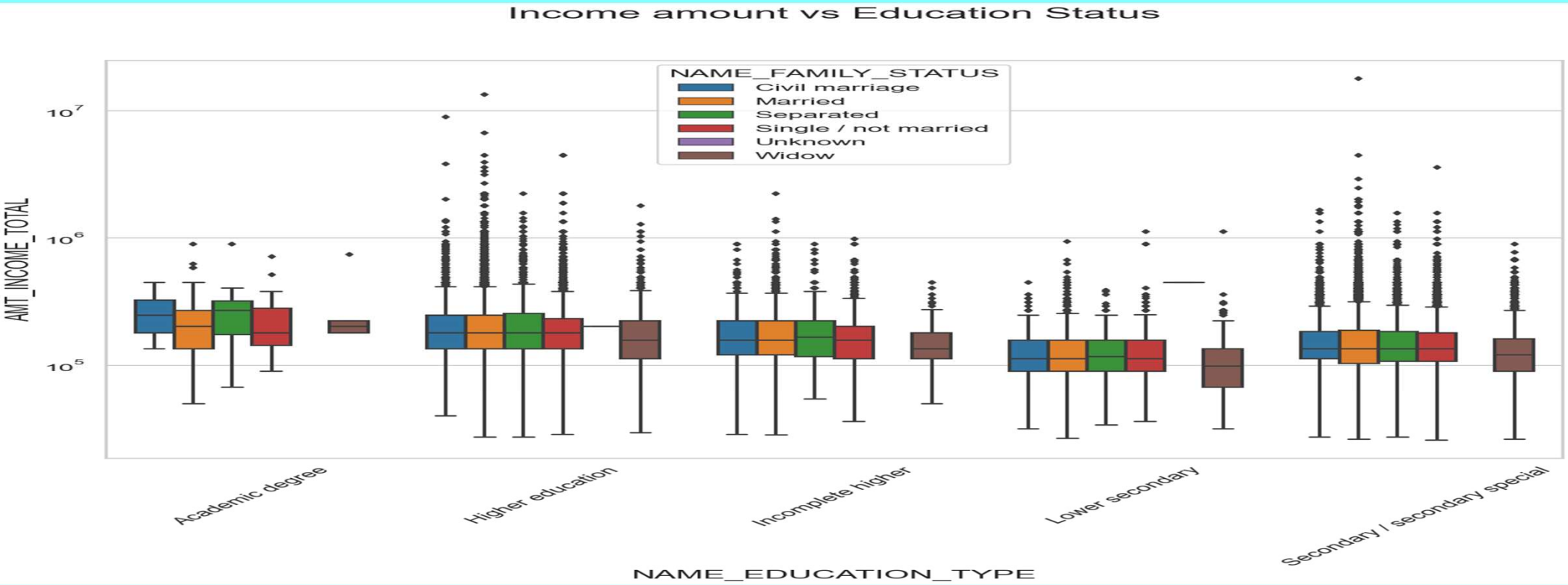
Insights from graph:

→Quite similar from Target 0, we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.

→Most of the outliers are from Education type 'Higher education' and 'Secondary'.

→Civil marriage for Academic degree is having most of the credits in the third quartile.

INCOME AMOUNT VS EDUCATION STATUS



Insights from the graph:

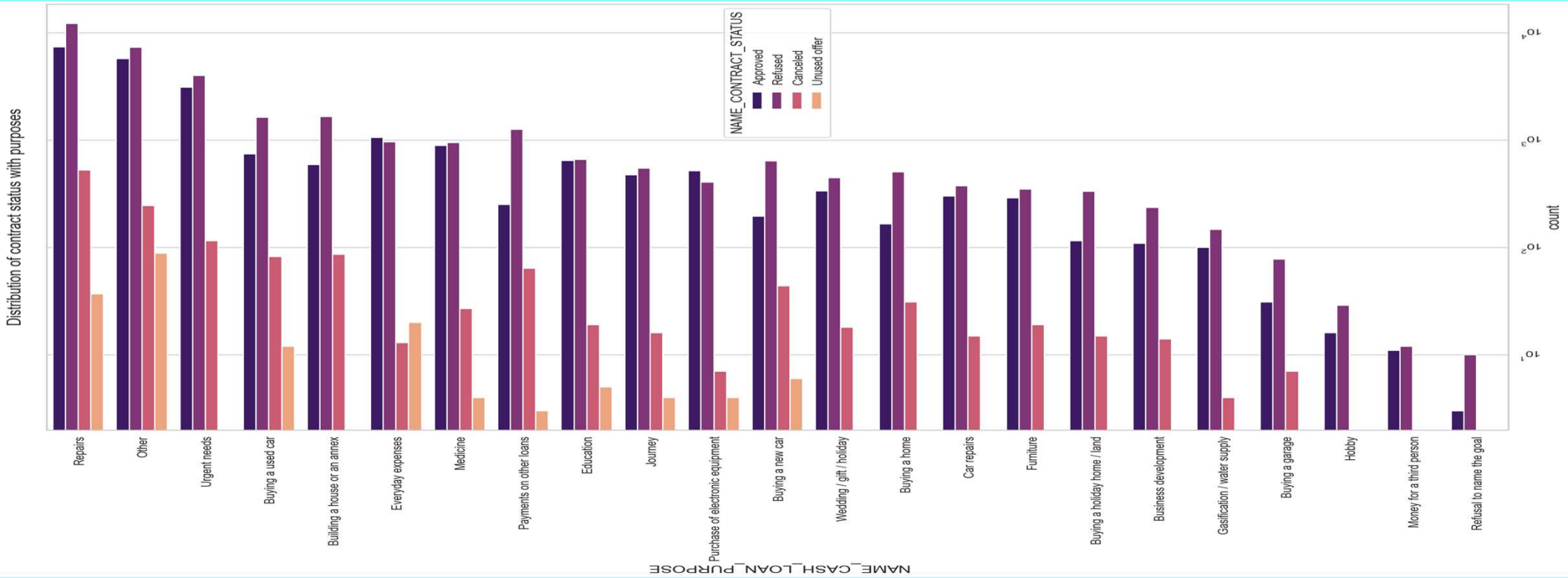
→They have some similarity with Target0, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status.

→Academic degree has less outliers but their income amount is little higher than Higher education.

→Lower secondary have less income amount than others.

UNIVARIATE ANALYSIS AFTER MERGING PREVIOUS APPLICATION DATA AND CURRENT DATA

DISTRIBUTION OF CONTRACT STATUS WITH PURPOSES



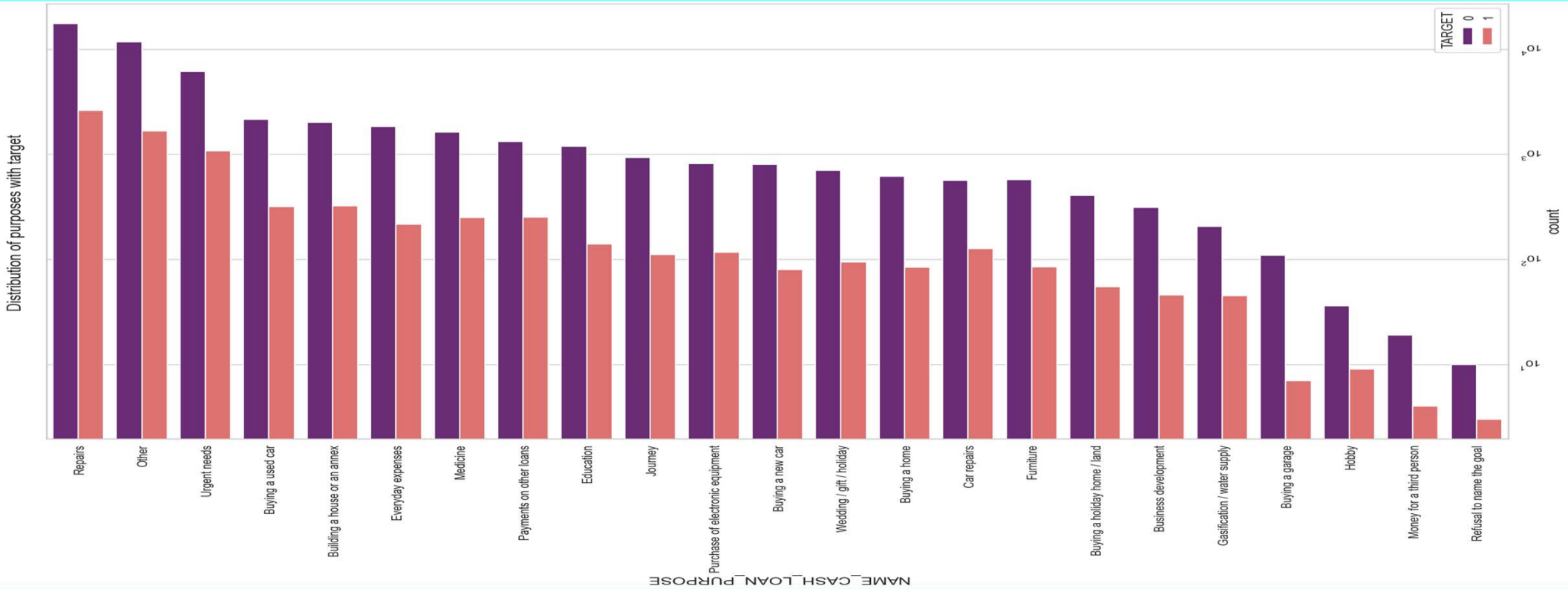
Insights from the graph:

→Most rejection of loans came from purpose 'repairs'.

→For education purposes we have equal number of approves and rejection

→Paying other loans and buying a new car is having significant higher rejection than approves.

DISTRIBUTION OF PURPOSES WITH TARGET



Insights from the graph:

→Loan purposes with 'Repairs' are facing more difficulties in payment on time.

→There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes for which the client is having for minimal payment difficulties.

PERFORMING BIVARIATE ANALYSIS

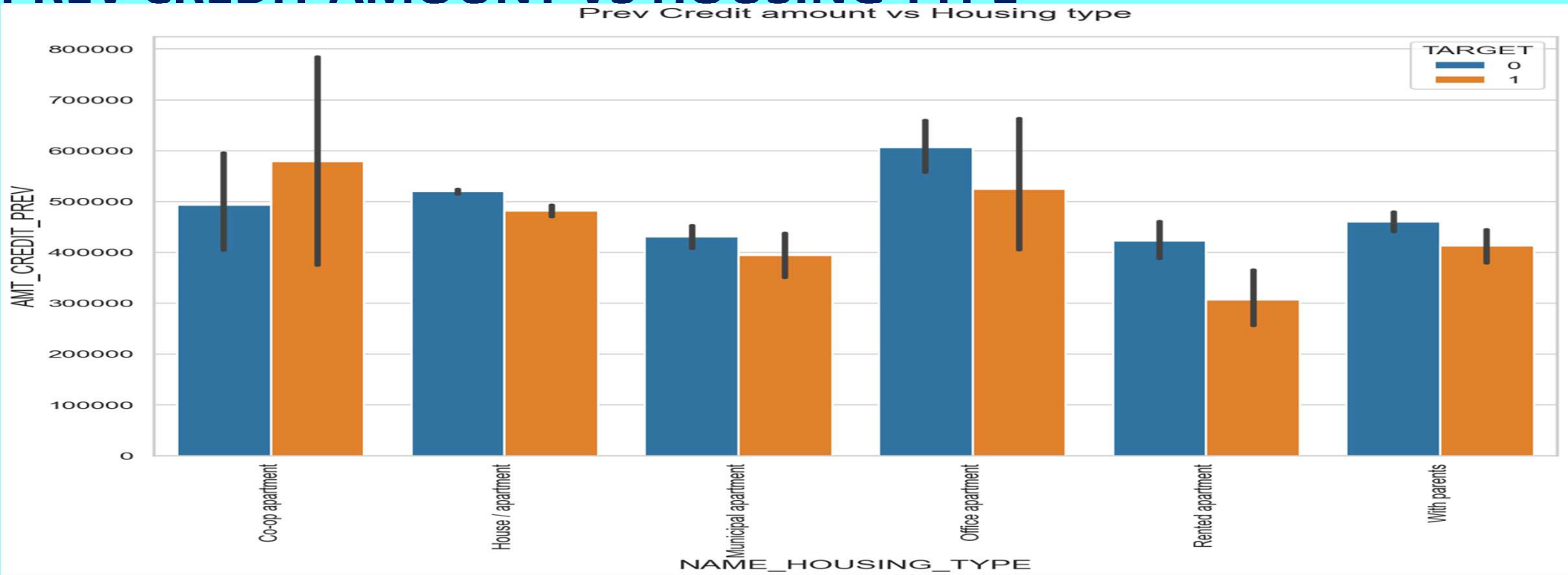
Prev Credit amount vs Loan Purpose
AMT_CREDIT_PREV
10^6
10^5
NAME_INCOME_TYPE
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
Other
Payments on other loans
Buying a used car
Repairs
Education
Buying a new car
Everyday expenses
Medicine
Car repairs
Urgent needs
Buying a holiday home / land
Building a house or an annex
Furniture
Journey
Purchase of electronic equipment
Wedding / gift / holiday
Buying a home
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE

PREVIOUS CREDIT AMOUNT VS LOAN PURPOSE

Insights from the graph:

- The credit amount for Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher.
- Income type for state servants have a significantly high amount of credit applied.
- Money for a third person or a Hobby has less credits applied.

PREV CREDIT AMOUNT VS HOUSING TYPE



Insights from the graph:

→Here for Housing type and office apartment is having higher credit for target 0 and co-op apartment is having higher credit for target 1.

→So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.

→Bank can focus mostly on housing type with parents or House\apartment or municipal apartment for successful payments.

These are the insights we have gained from the EDA analysis:

- Male customers have defaulted more compared to female customers.
- Customers with Academic degree has less defaults compared to customers with Lower Secondary & Secondary education.
- Customers who are Student and Businessmen have no defaults but Clients who are either at Maternity leave, or Unemployed default a lot.
- Customers above age of 50 have low probability of defaulting and young people who are in age group of 20-40 as have higher probability of defaulting.
- Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff have high default rate.
- When the credit amount goes beyond 3 lakhs, there is an increase in defaulters.
- Also with loan purpose for ‘Repair’ has higher number of unsuccessful payments.

CONCLUSION(and recommendations):

→The customers who have delayed their payment can be negotiated with for higher rate of interest or loan amount can be reduced to lower the risk of loss for bank.

→Favouring of Female customer to Male customer will also reduce the risk of defaulting as female customers have higher credits compared to male customers.

→focusing more on contract type ‘Student’ ,and ‘Businessman’ with housing type other than ‘Co-op apartment’ will also increase chance of successful payments.

→Banks should focus less on income type just ‘Working’ as they have defaulted the most in the past.

→loan purpose for ‘Repair’ should be avoided as it's payments have defaulted in the past.

→Bank can focus mostly on housing type 'with parents' or 'House\appartment' or 'miuncipal apartment' as they have made successful payments.

Thank you