Credit EDA Case Study

BY DARSHIL SHAH

INTRODUCTION

- This case study aims to give an idea of applying EDA in a real business scenario. In this case study we apply the techniques learnt in the EDA module to:
  - Develop a basic understanding of risk analytics in banking and financial services and
  - Understand how data is used to minimize the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING - 1

- The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. We have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.
- When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company or
  - If the applicant is likely to default, then approving the loan may lead to a financial loss for the company

BUSINESS UNDERSTANDING – 2

The data given contains the 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 late payment more than X days on at least one of the first Y installments of the loan and
- **All other cases**: All other cases when the payment is paid on time

When a client applies for a loan, there are four types of decisions that could be taken by the client/company):

1. **Approved**: The Company has approved loan Application
2. **Cancelled**: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan
3. **Refused**: The company had rejected the loan.
4. **Unused offer**: Loan has been cancelled by the client but on different stages of the process

BUSINESS OBJECTIVES

- The aim is to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending at a higher interest rate, etc. Identification of such applicants using EDA is the aim of this case study.

- To understand the factors behind loan default i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

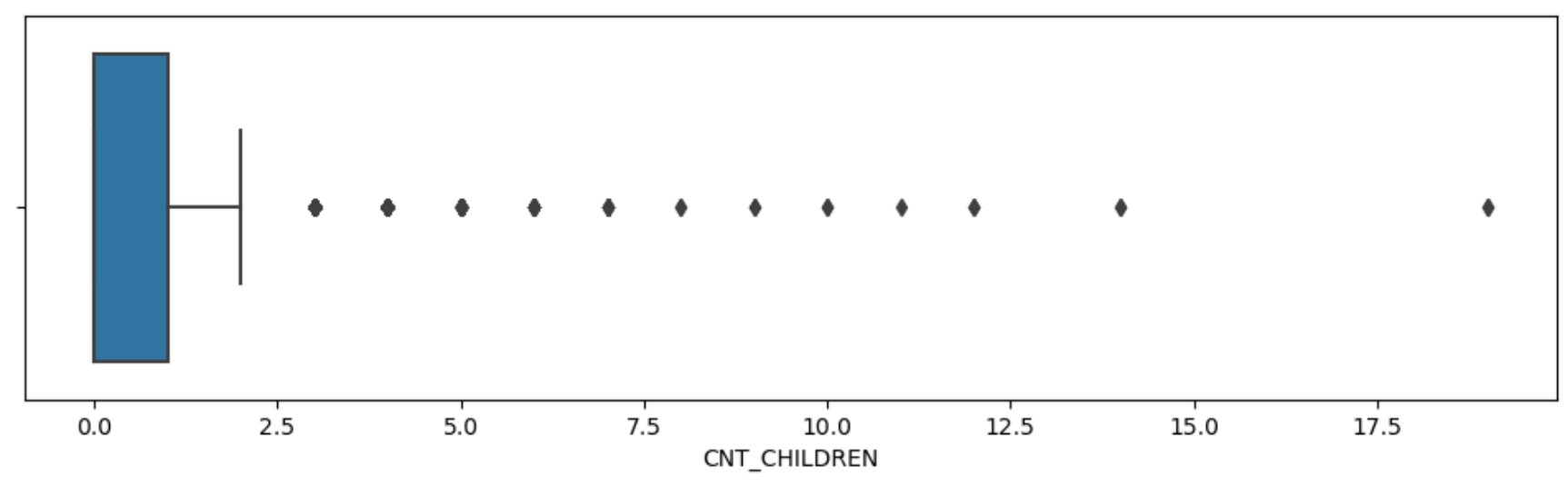
Analysis of information of the client at the time of application

Analysis of Imbalance for 'TARGET'

We have imbalance in `TARGET` variable based on the % of observations

- `Target 1' represents client with payment difficulties. This is only 8.07% of the data
- `Target 0' represents all other cases than Target 1. This is 91.93% of the data

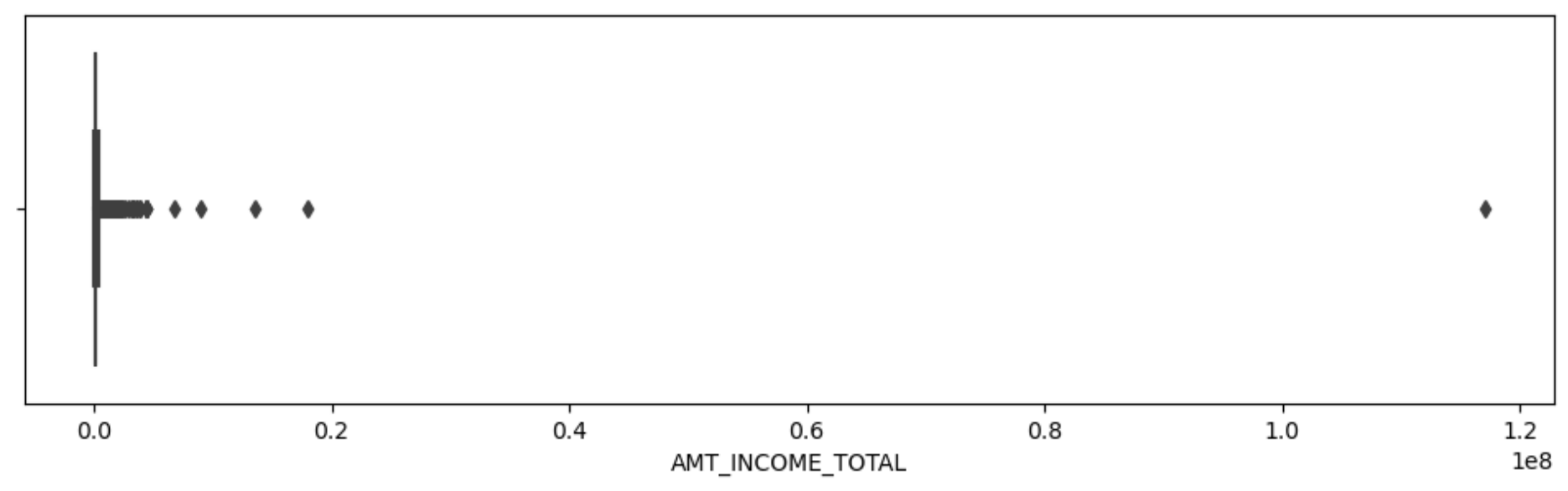
BOX PLOT FOR 'CNT_CHILDREN'



Observation:

- 1st quartile is missing for CNT_CHILDREN which means most of the data are present in the 1st quartile.
- It show the values above 2.5 as being outliers

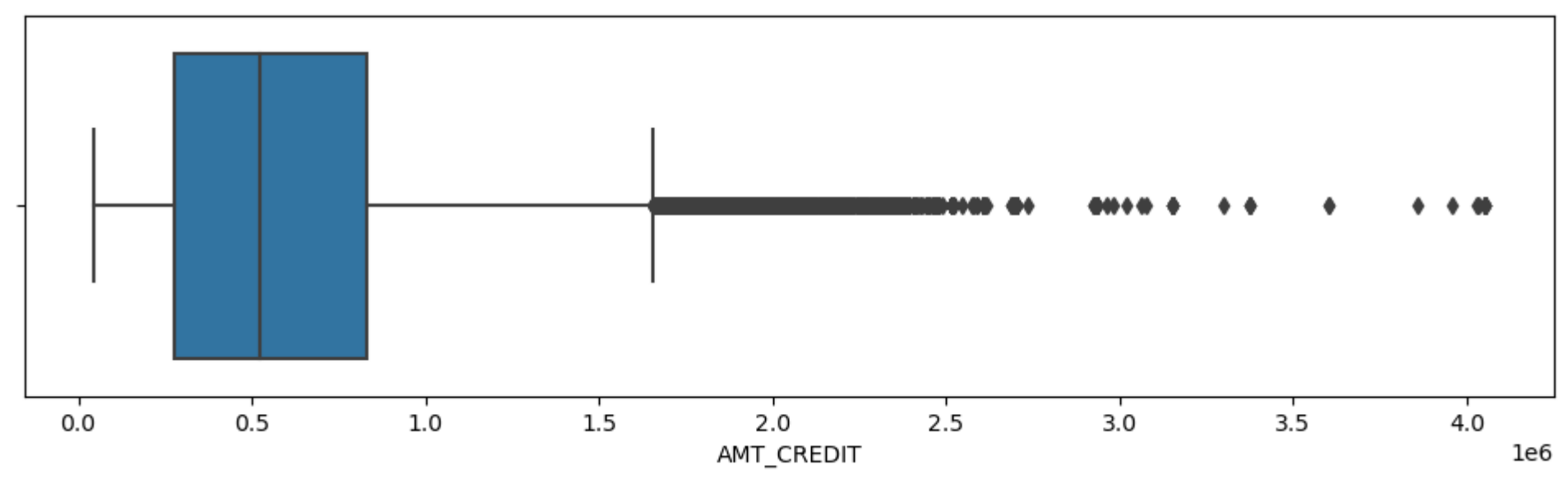
BOX PLOT FOR 'AMT_INCOME_TOTAL'



Observation:

- In AMT_INCOME_TOTAL only single high value data point is present as outlier

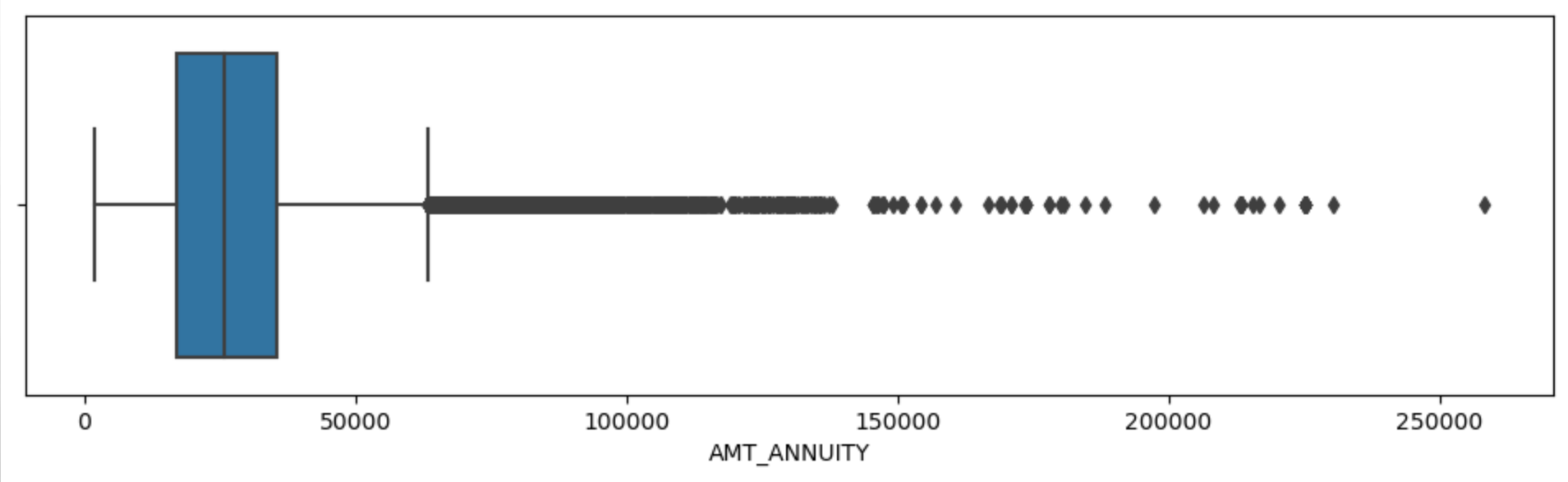
BOX PLOT FOR 'AMT_CREDIT'



Observation:

- AMT_CREDIT has little bit more outliers

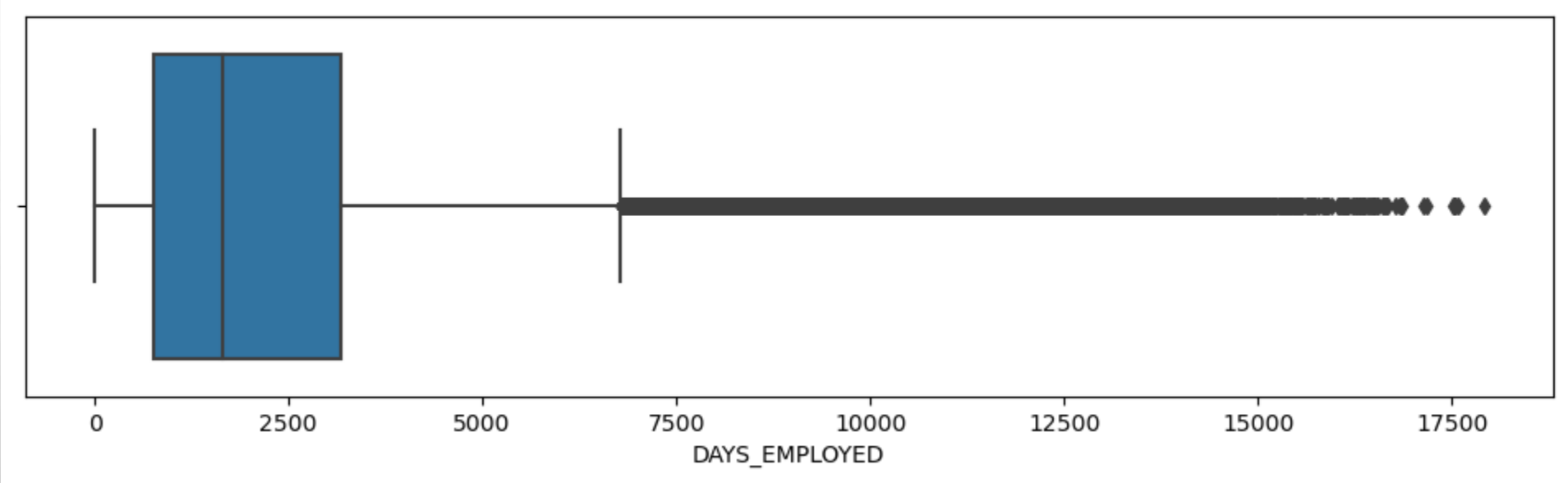
BOX PLOT FOR 'AMT_ANNUITY'



Observation:

- 1st quartiles and 3rd quartile for AMT_ANNUITY is moved towards first quartile.
- Applicants with `AMT_ANNUITY` above approx. 60,000 are outliers

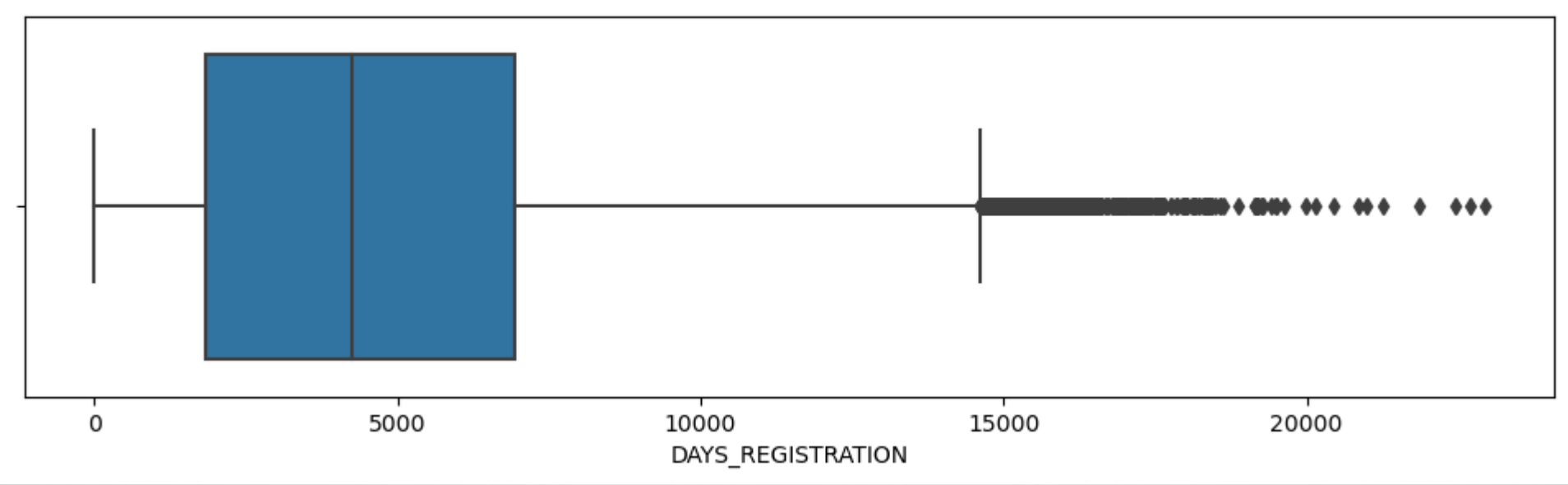
BOX PLOT FOR 'DAYS_EMPLOYED'



Observation:

- 1st quartiles and 3rd quartile for DAYS_EMPLOYED is moved towards first quartile.

BOX PLOT FOR 'DAYS_REGISTRATION'



Observation:

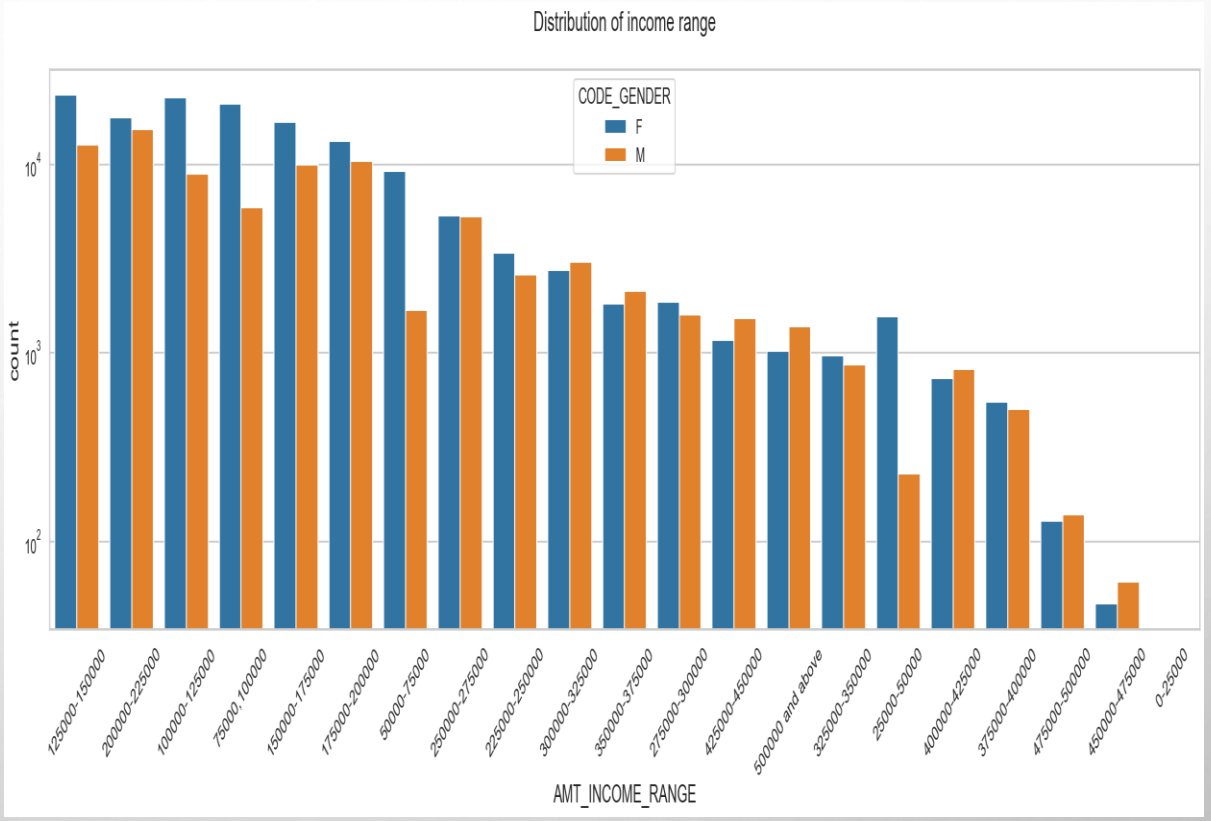
- 1st quartiles and 3rd quartile for DAYS_EMPLOYED is moved towards first quartile.

Categorical Univariate Analysis for Target 0

DISTRIBUTION OF INCOME RANGE

Points to be concluded from the graph:

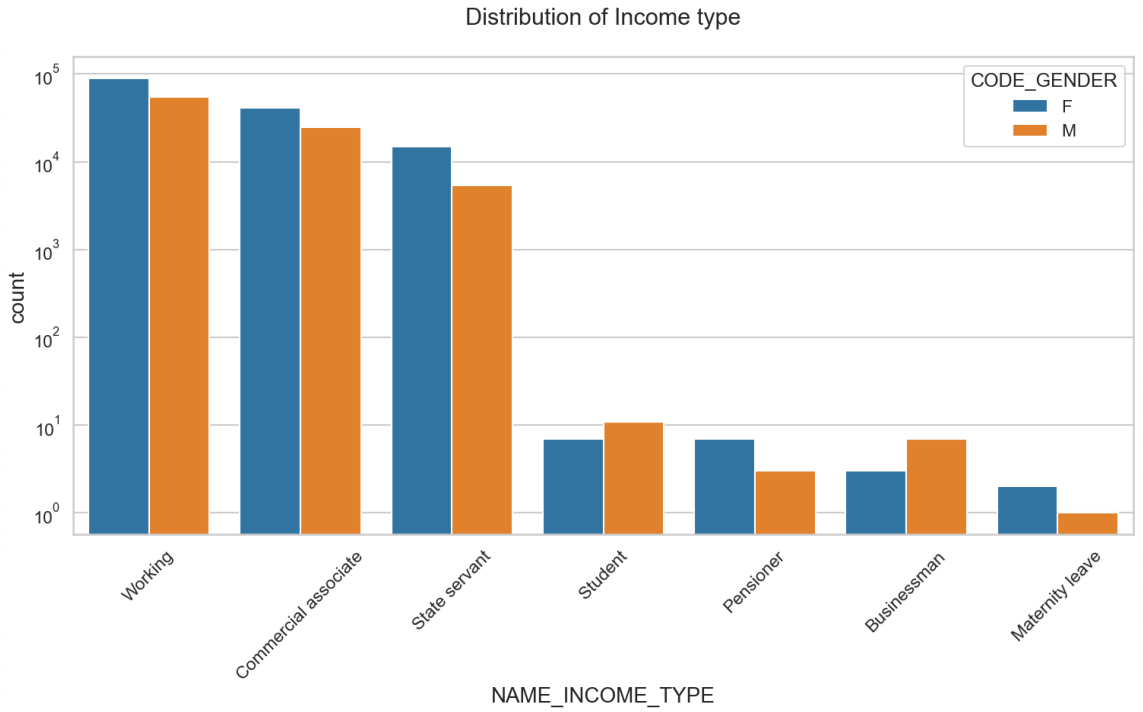
- Income range from 1,00,000 to 2,00,000 is having more number of credits.
- Females are more than male in having credits for that range.
- Very less count for income range 400000 and above.



DISTRIBUTION OF INCOME TYPE

Points to be concluded from the graph

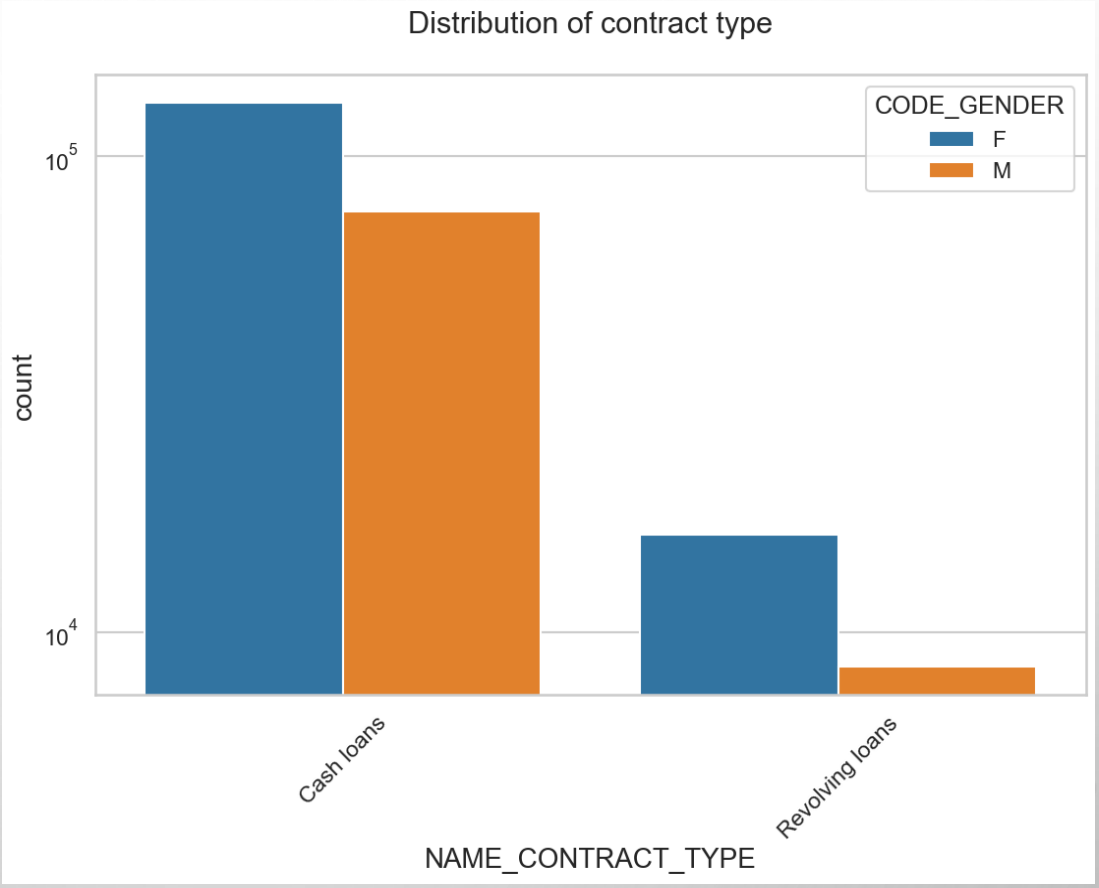
- For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than others.
- For this Females are having more number of credits than male.



DISTRIBUTION FOR CONTRACT TYPE

Points to be concluded from the graph

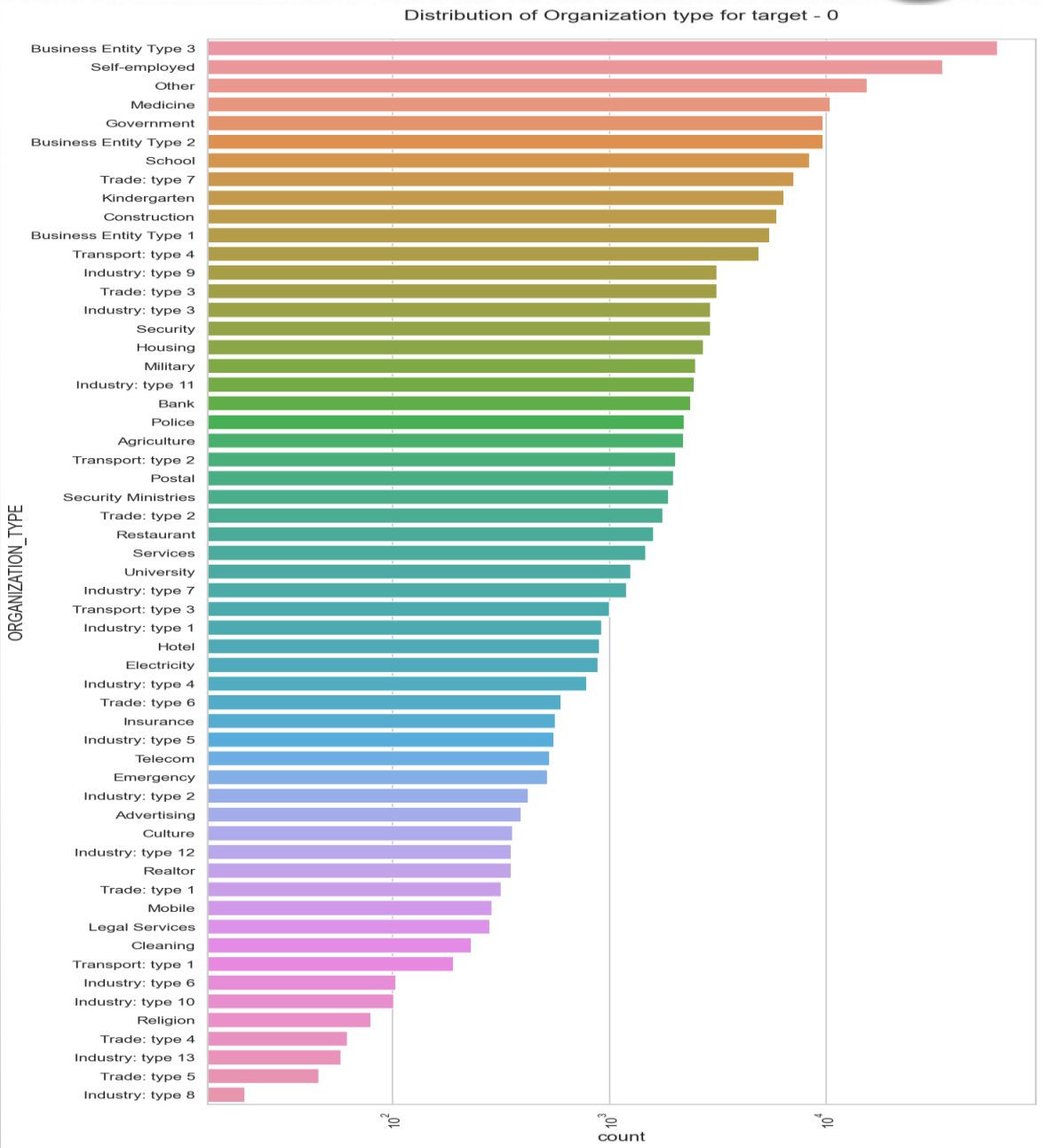
- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
- For this also Female is leading for applying credits



DISTRIBUTION OF ORGANISATION TYPE

Points to be concluded from the graph

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

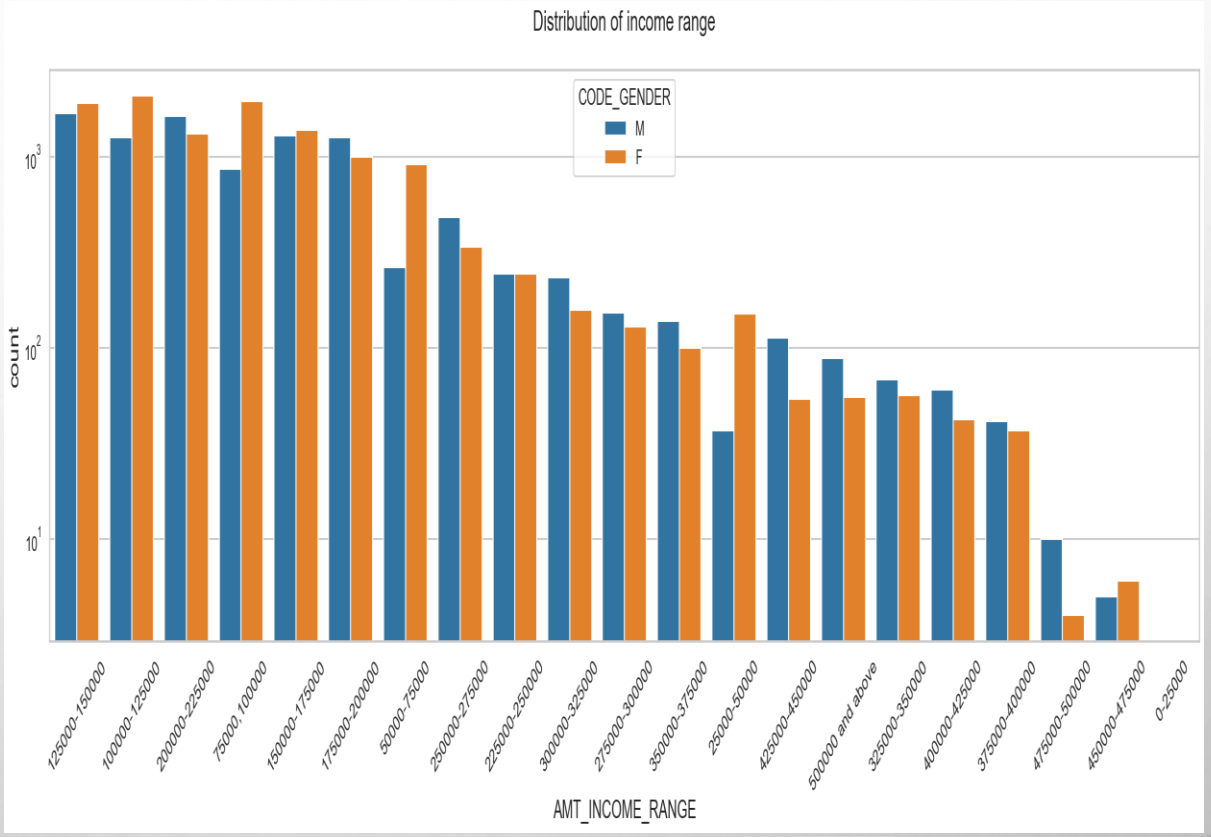


Categorical Univariate Analysis for Target 1

DISTRIBUTION OF INCOME RANGE

Points to be concluded from the graph

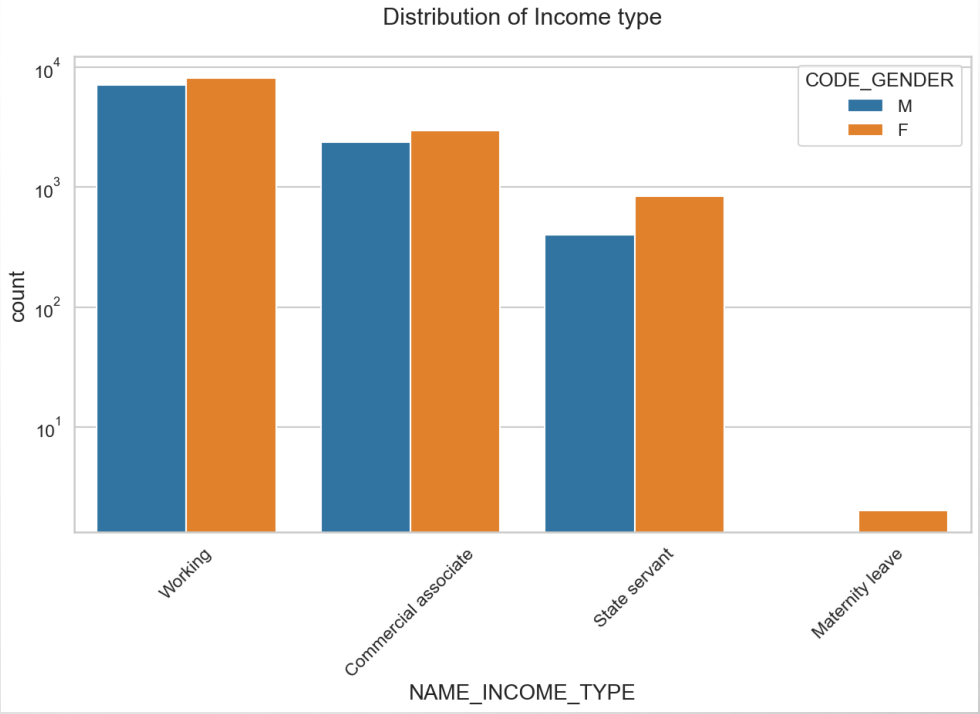
- Male counts are higher than female.
- Income range from 100000 to 200000 is having more number of credits.
- Very less count for income range 400000 and above.



DISTRIBUTION OF INCOME TYPE

Points to be concluded from the graph

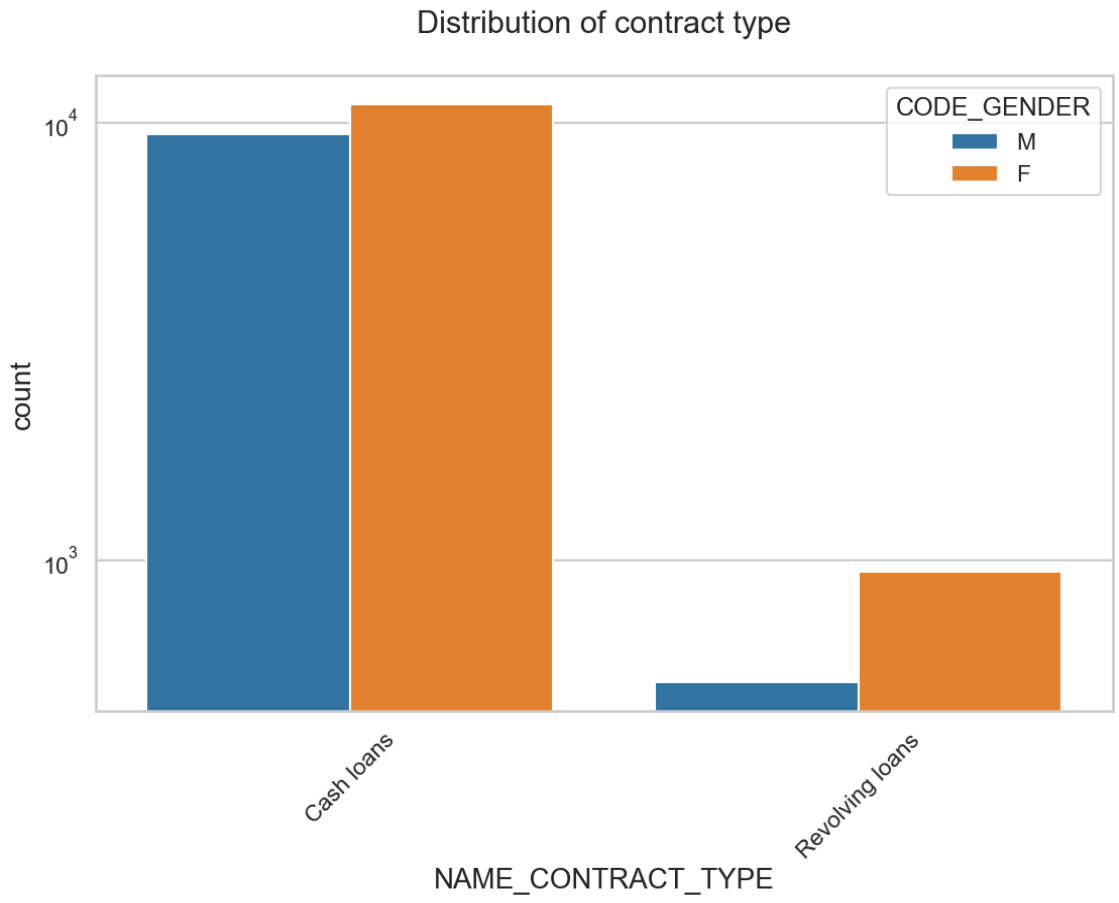
- For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave.
- Females are having more number of credits than male.
- For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.



DISTRIBUTION FOR CONTRACT TYPE

Points to be concluded from the graph

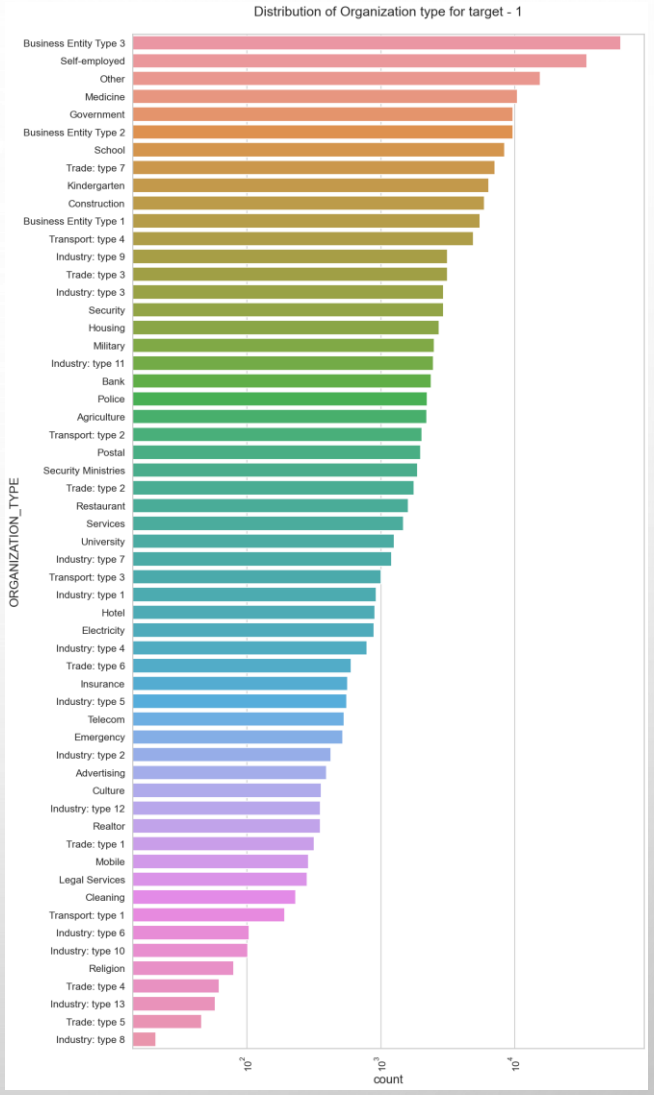
- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
- For this also Female is leading for applying credits.
- For type 1 : there are majorly Female for Revolving loans.



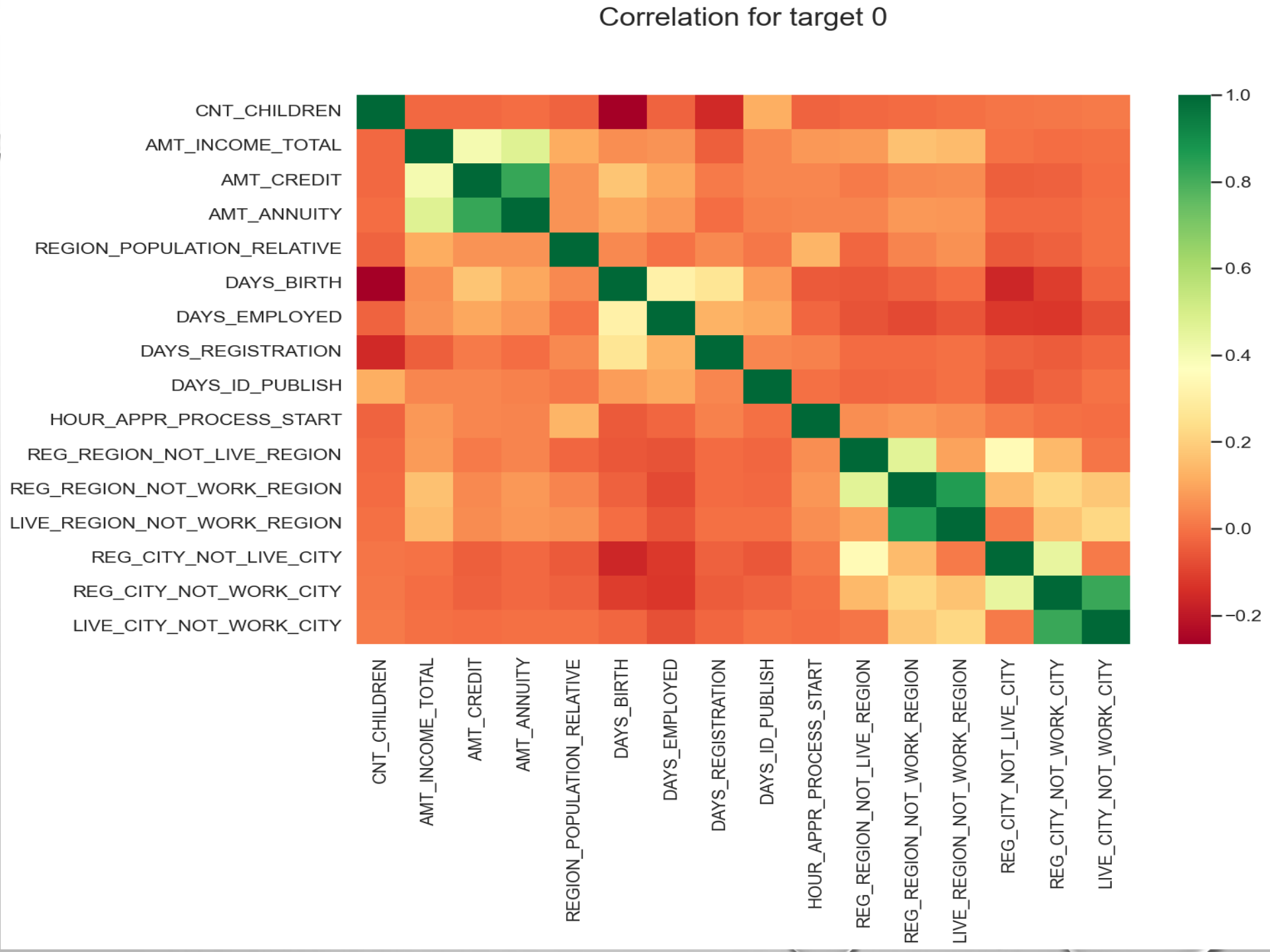
DISTRIBUTION OF ORGANISATION TYPE

Points to be concluded from the 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

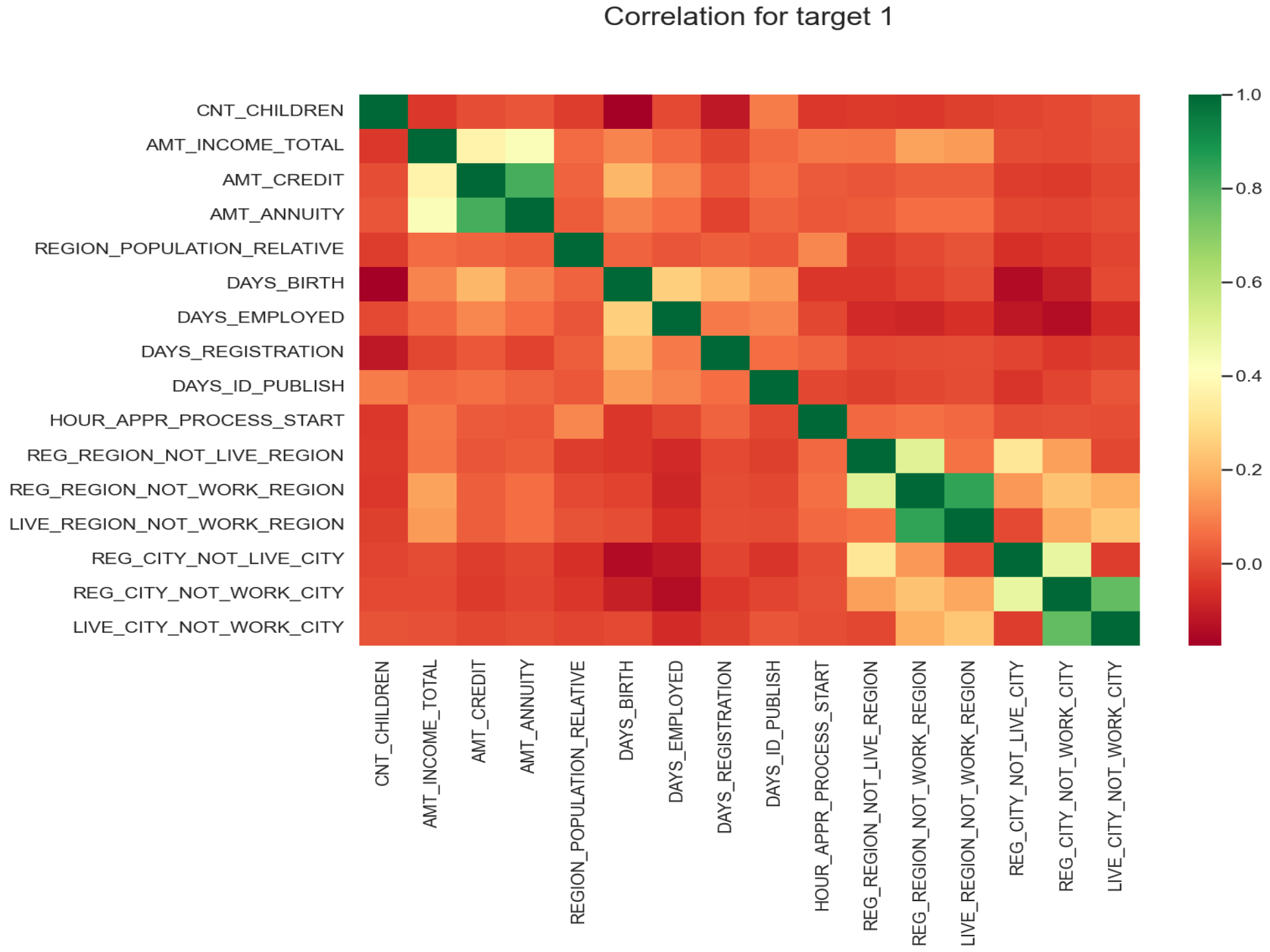

Correlation for target 0
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

CORRELATION FOR TARGET 0

Points to be concluded from the graph presented before.

- 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 directly proportional to population density, which means Credit amount is higher for high densely populated area and vice-versa.
- Income amount is directly proportional to population density, which means Income amount is higher for high densely populated area and vice-versa.

Correlation of Target 1


Correlation for target 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
1.0
0.8
0.6
0.4
0.2
0.0

CORRELATION OF TARGET 1

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

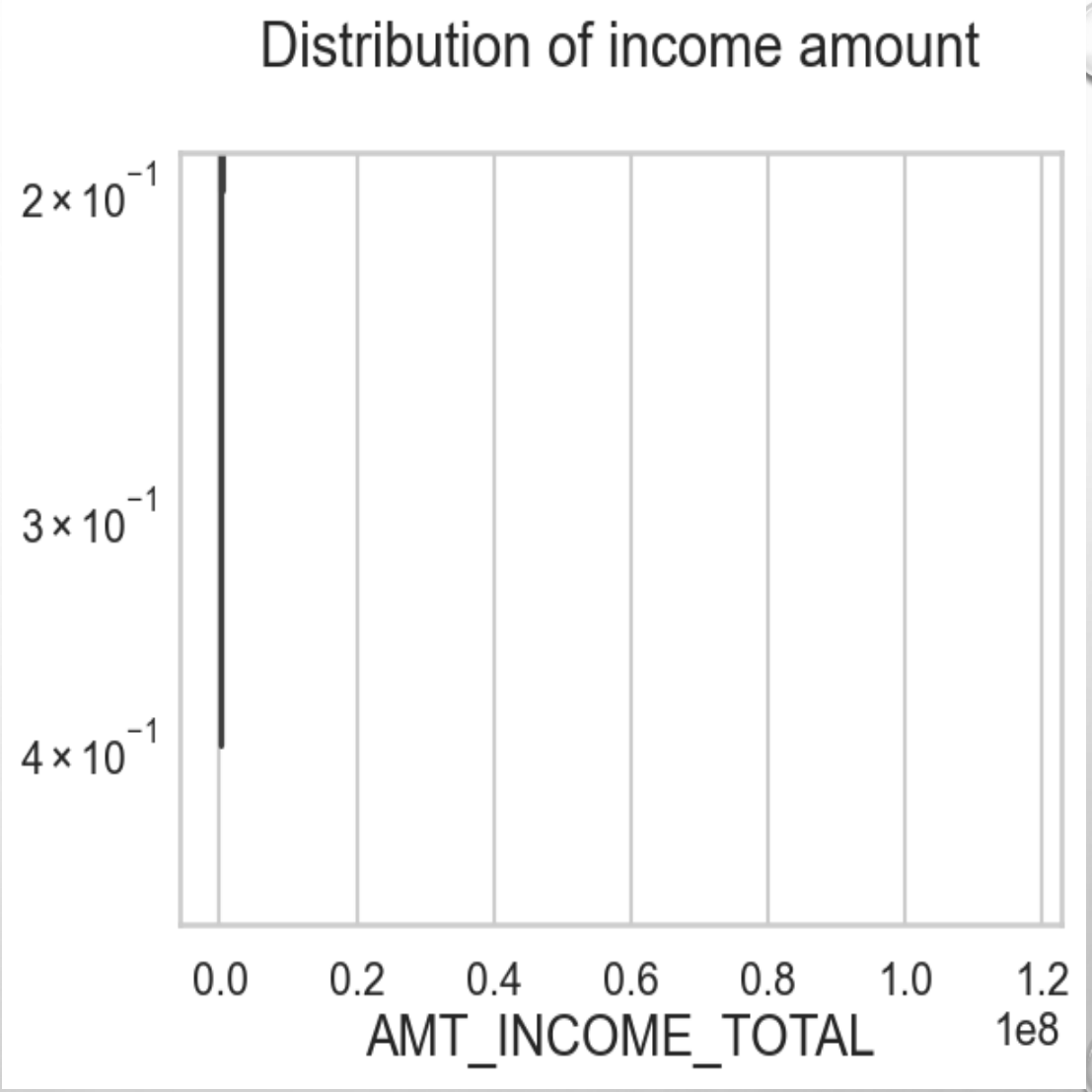
- The client's permanent address does not match contact address are having less children and vice-versa
- The client's permanent address does not match work address are having less children and vice-versa

Categorical Univariate Analysis for Variables Target 0

BOXPLOT FOR INCOME AMOUNT

Few points can be concluded from the graph.

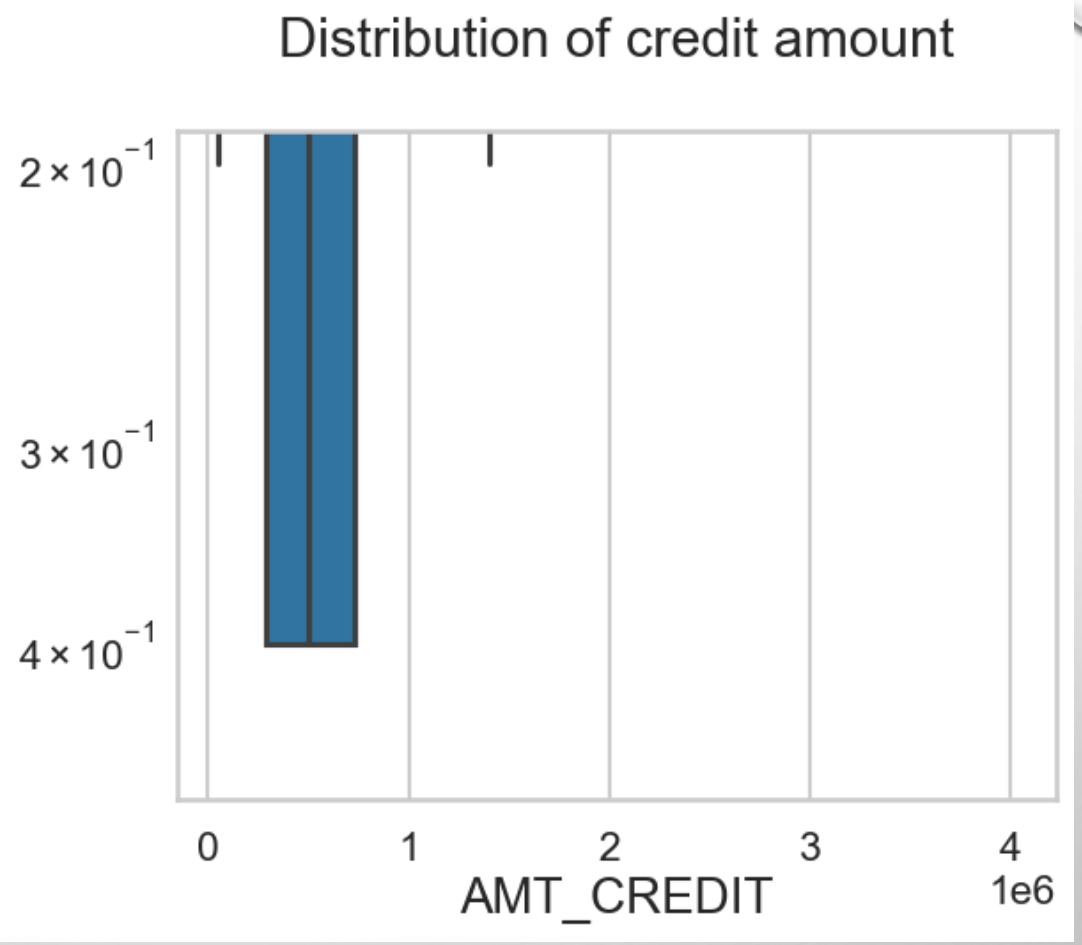
- Some outliers are noticed in income amount.
- The third quartiles is very slim for income amount.



BOXPLOT FOR CREDIT AMOUNT

Few points can be concluded from the graph.

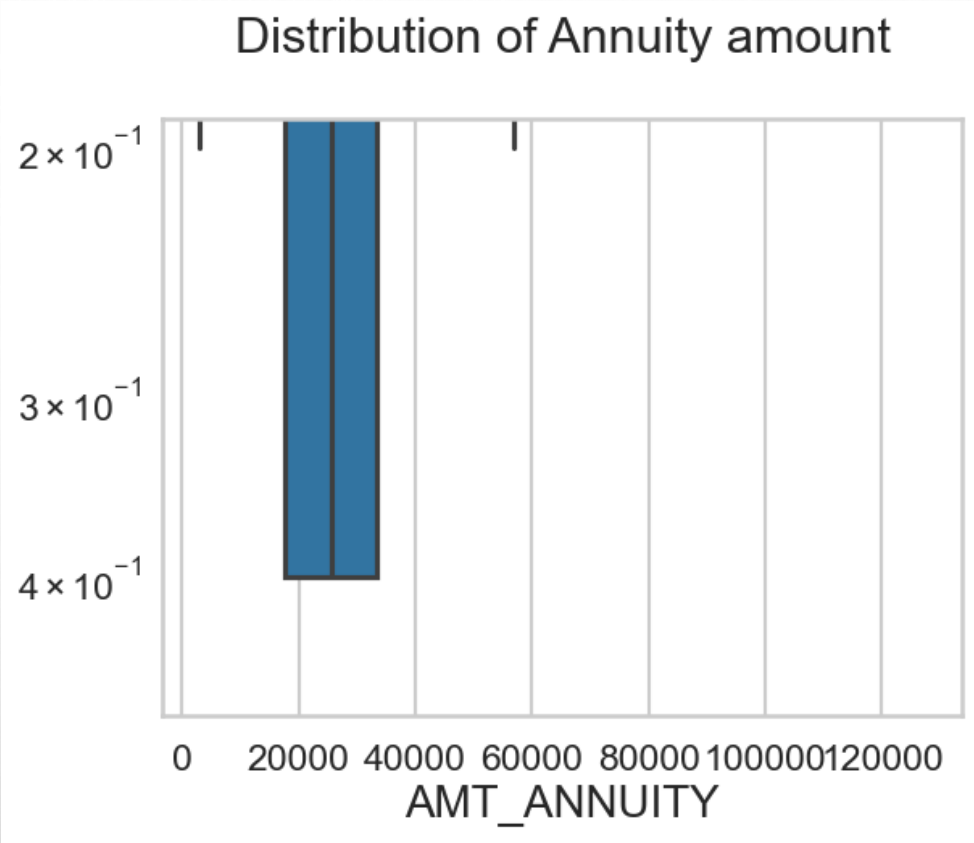
- Some outliers are noticed in credit amount.
- The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.



BOXPLOT FOR ANNUITY AMOUNT

Few points can be concluded from the graph.

- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

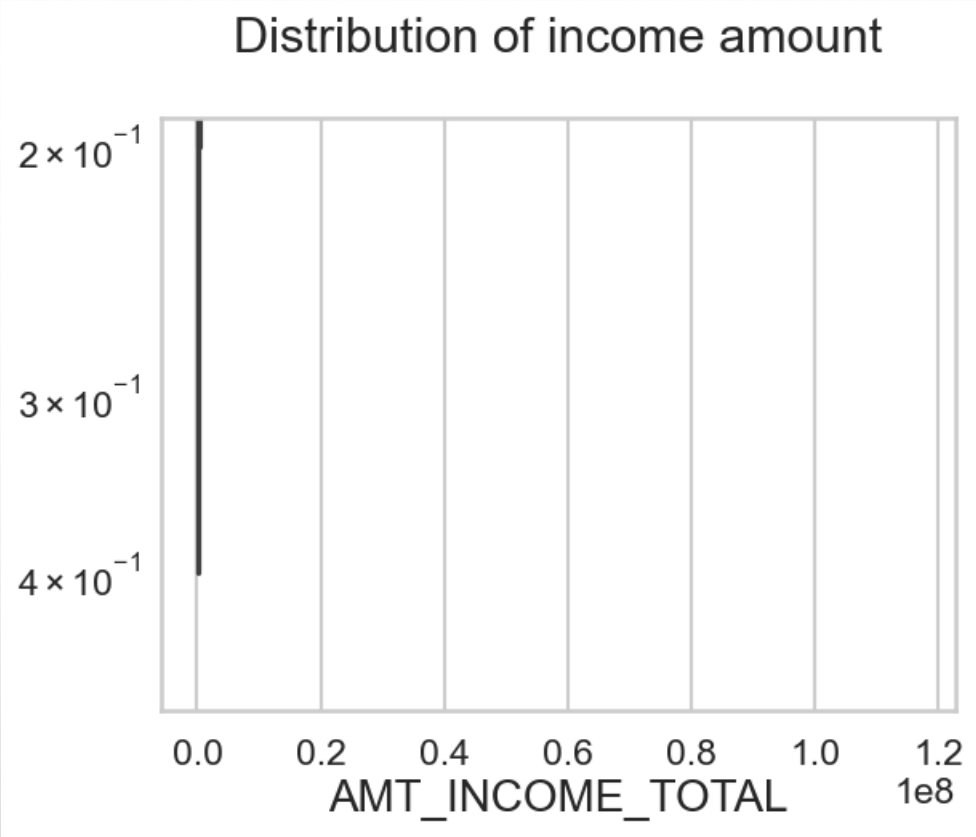


Categorical Univariate Analysis for Variables Target 1

BOXPLOT FOR INCOME AMOUNT

Few points can be concluded from the graph.

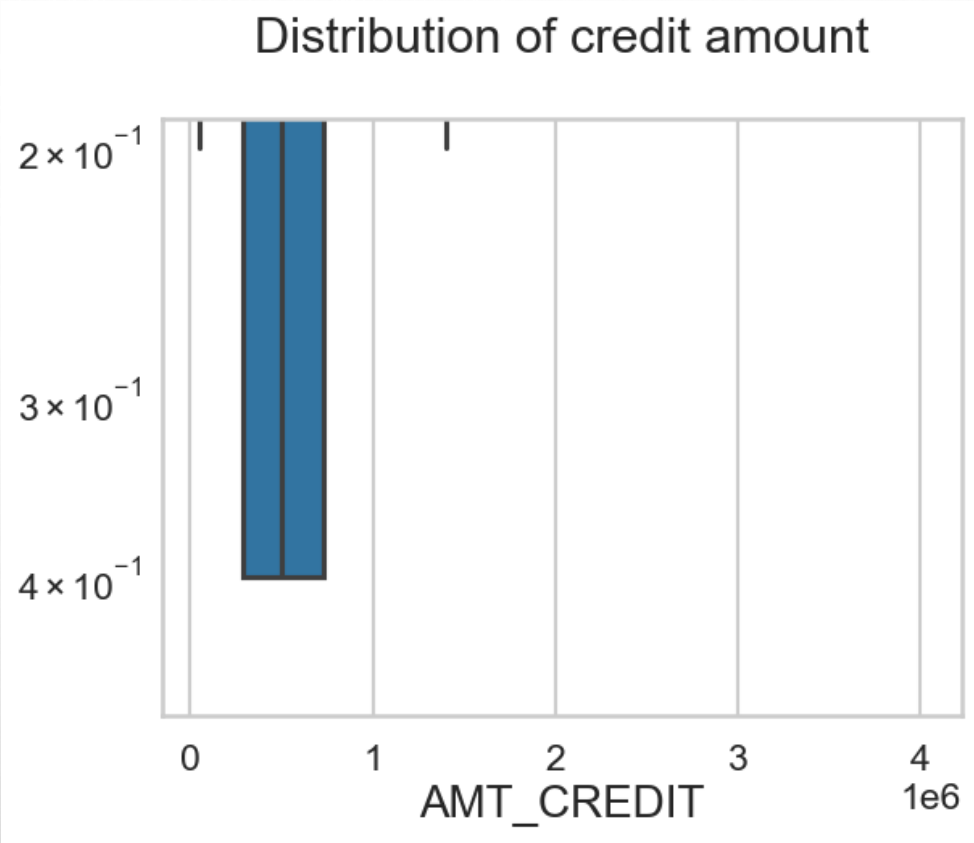
- Some outliers are noticed in income amount.
- The third quartiles is very slim for income amount.
- Most of the clients of income are present in first quartile.



BOXPLOT FOR CREDIT AMOUNT

Few points can be concluded from the graph.

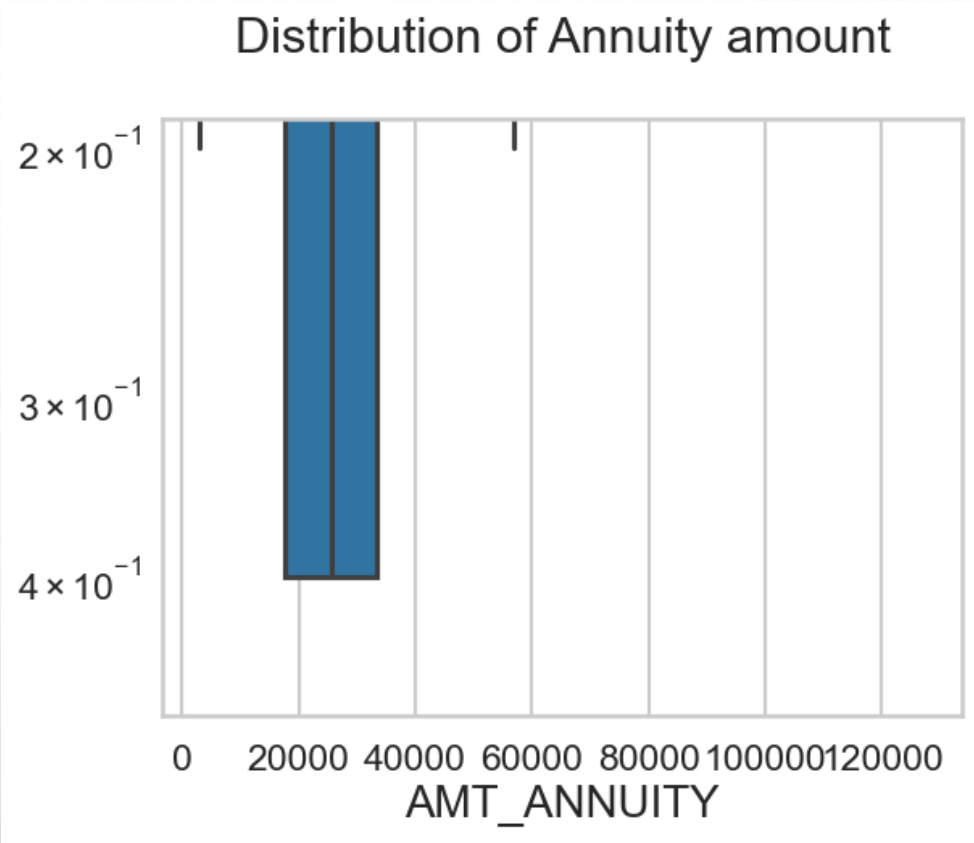
- Some outliers are noticed in credit amount.
- The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.



BOXPLOT FOR ANNUITY AMOUNT

Few points can be concluded from the graph.

- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.

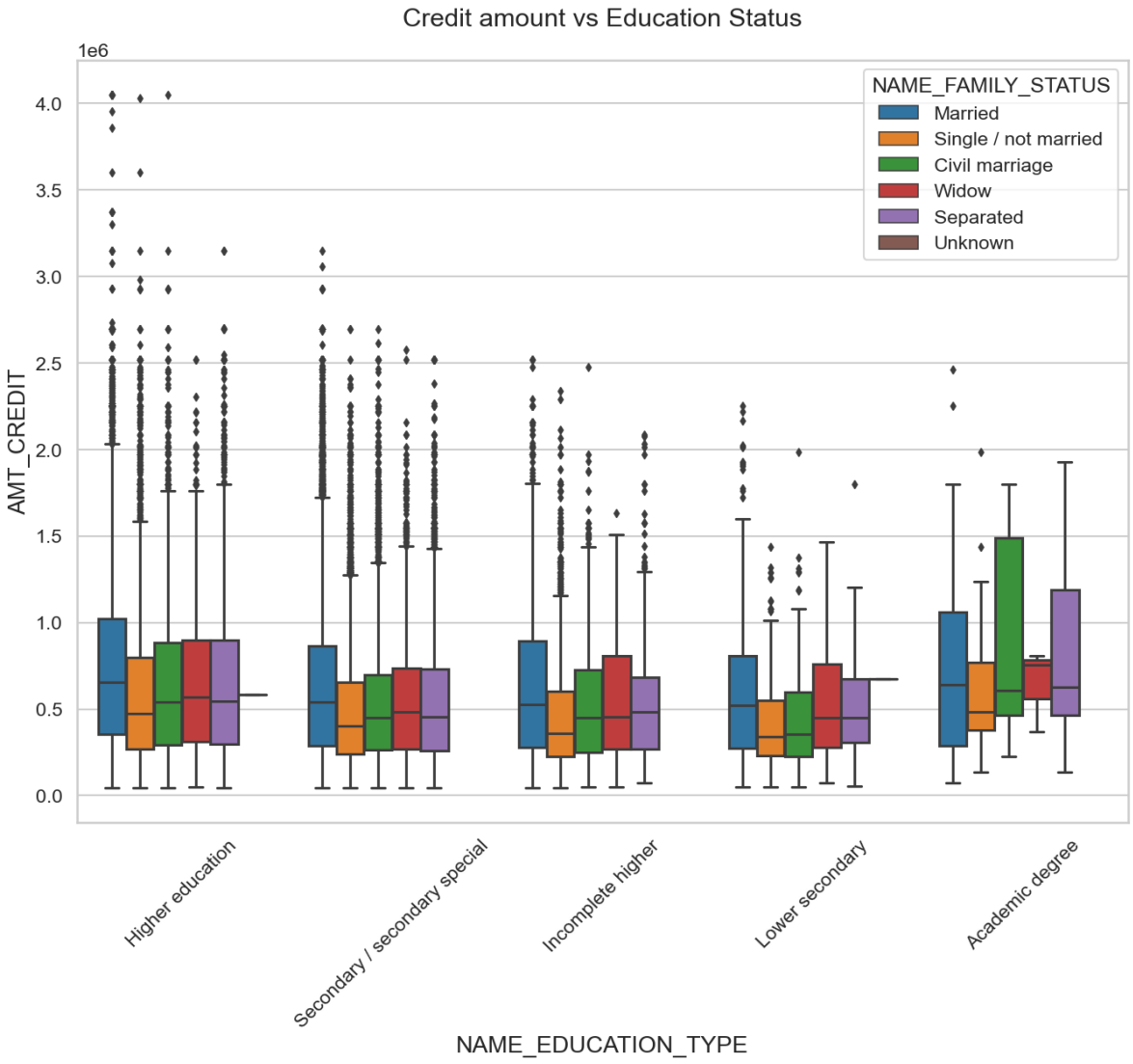


BIVARIATE ANALYSIS FOR TARGET 0

CREDIT AMOUNT VS EDUCATION STATUS

Few points can be concluded from the graph.

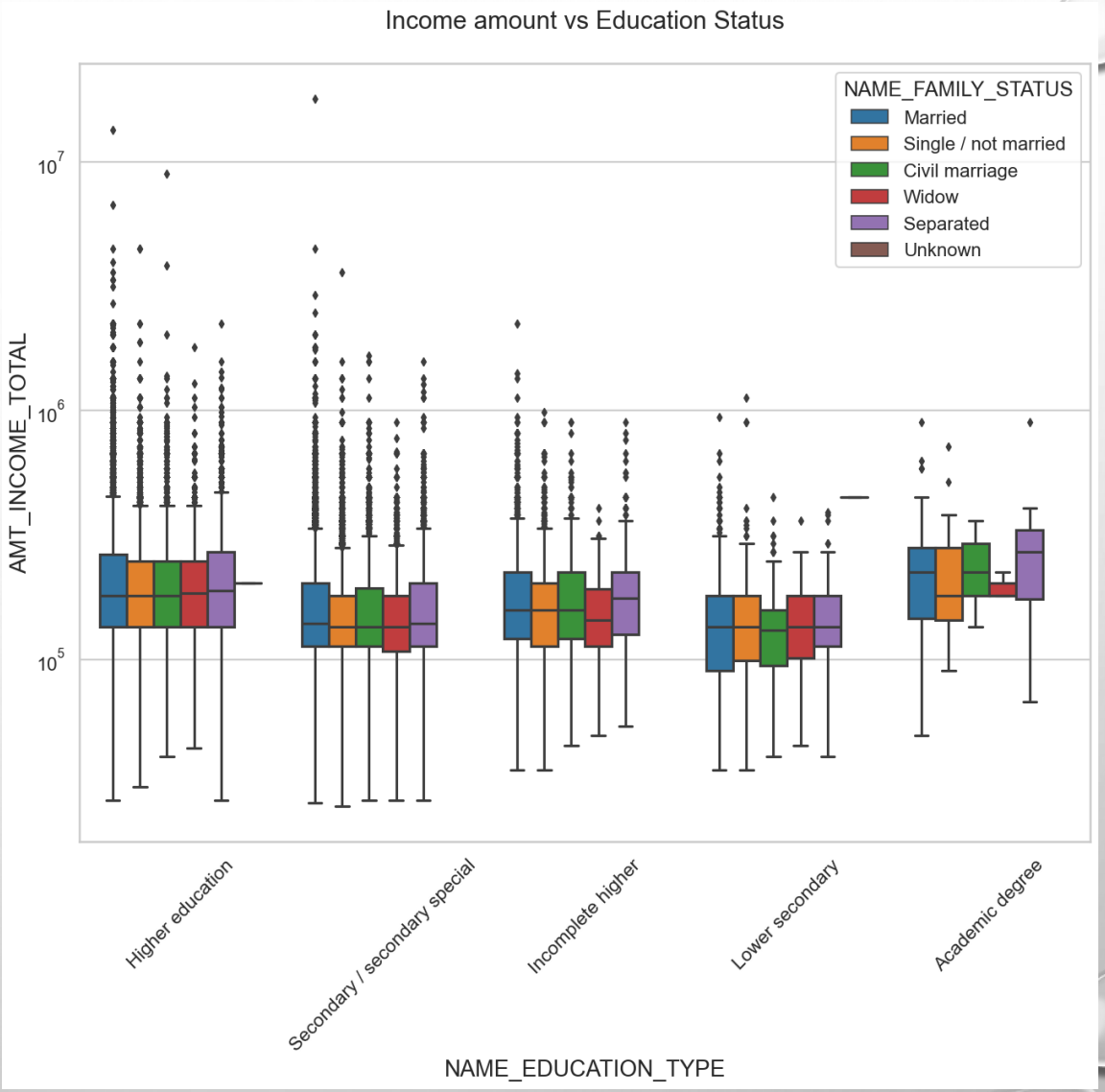
- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
- Civil marriage for Academic degree is having most of the credits in the third quartile.



INCOME AMOUNT VS EDUCATION STATUS

Few points can be concluded from the graph.

- For Education type 'Higher education' the income amount mean is mostly equal with family status. It does contain many outliers.
- Less outlier are having for Academic degree but they are having the income amount is little higher that Higher education.
- Lower secondary of civil marriage family status are have less income amount than others.

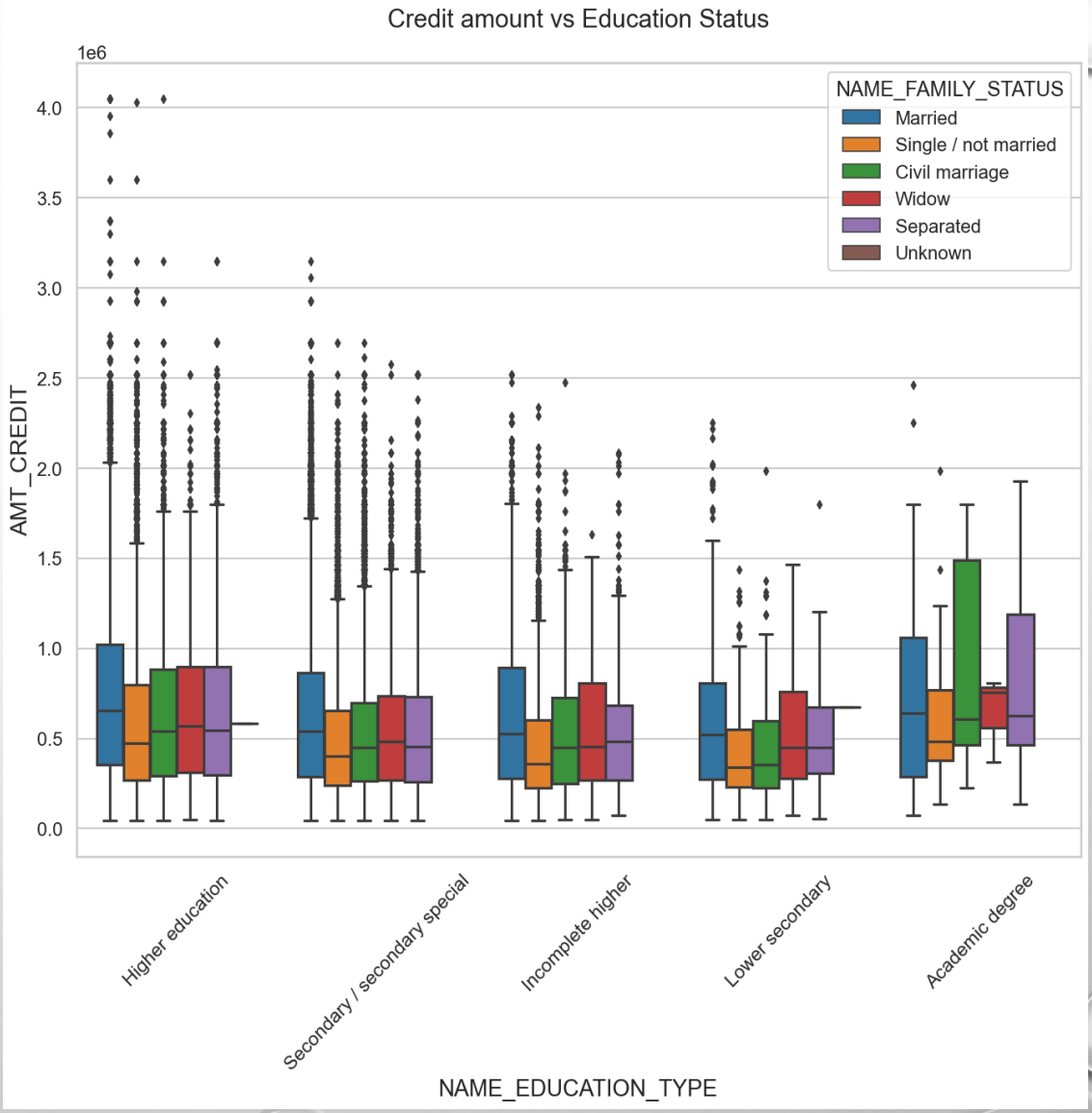


Bivariate Analysis for Target 1

CREDIT AMOUNT VS EDUCATION STATUS

Few points can be concluded from the graph.

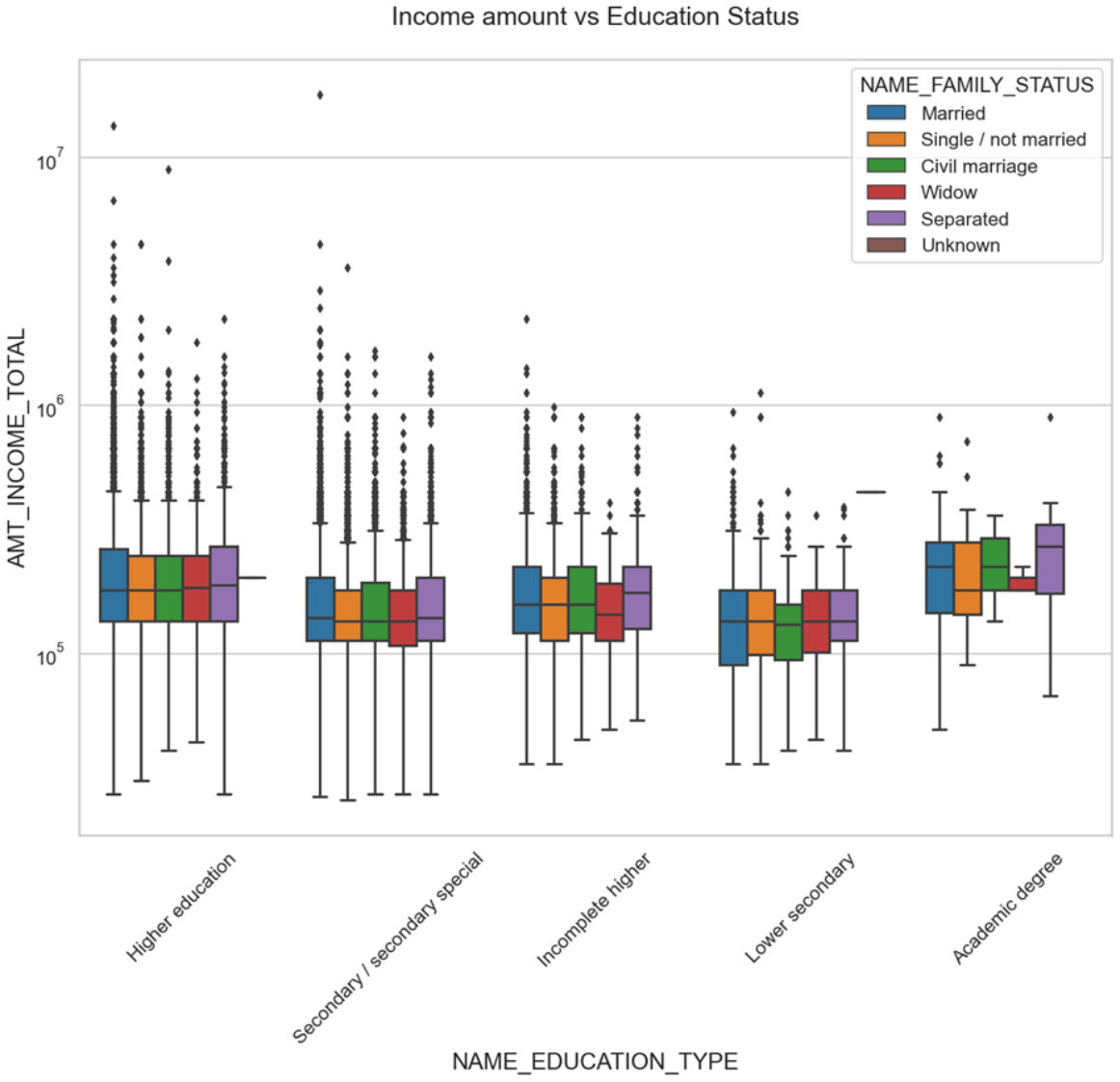
- Quite similar from Target 0, we can say that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Most of the outliers are from Education type 'Higher education' and 'Secondary'.
- Civil marriage for Academic degree is having most of the credits in the third quartile.



INCOME AMOUNT VS EDUCATION STATUS

Few points can be concluded from the graph.

- Have some similarity with Target 0, From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status.
- Less outlier are having for Academic degree but there income amount is little higher that Higher education.
- Lower secondary are have less income amount than others

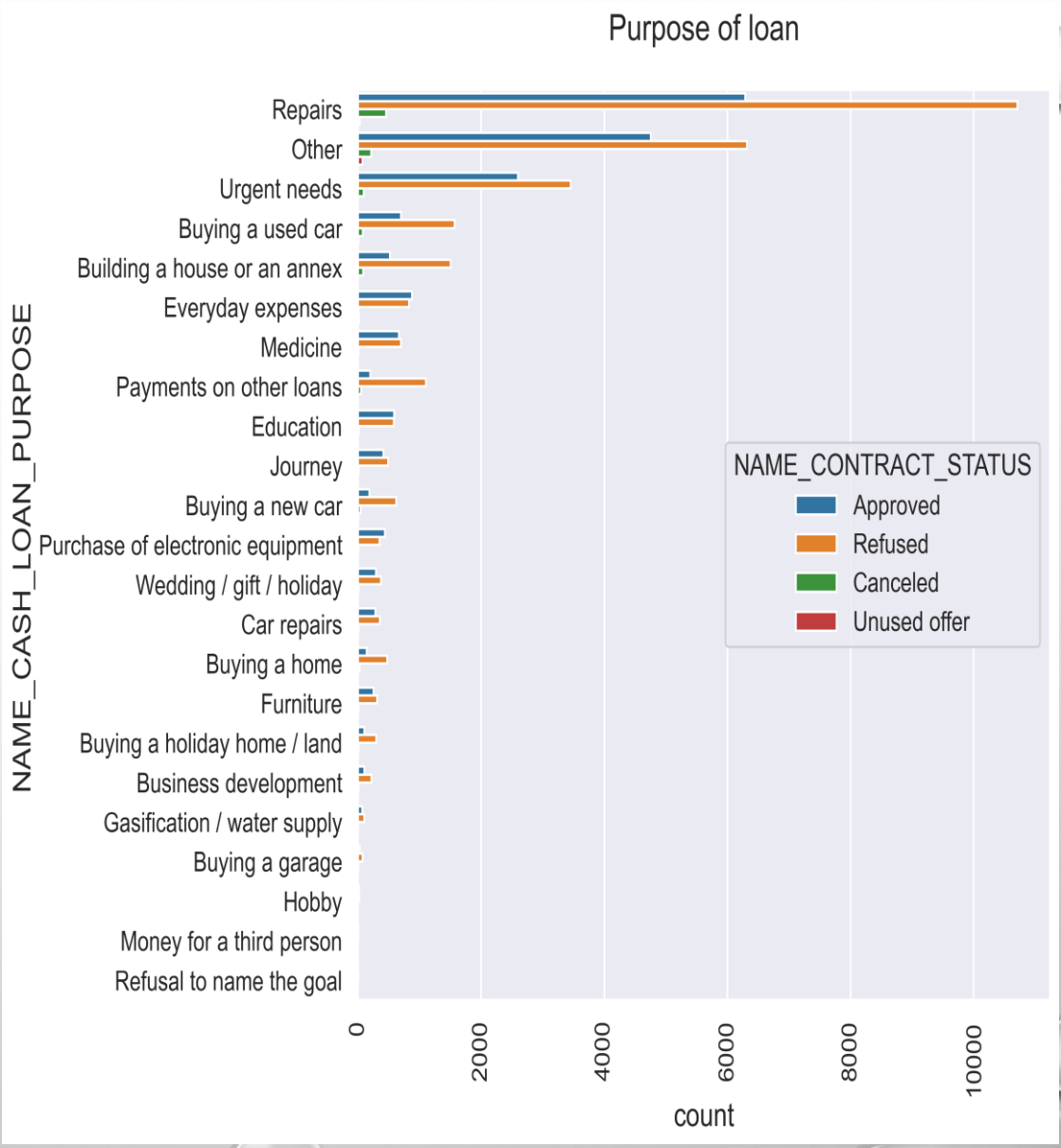


Univariate analysis after merging previous data

PURPOSE OF LOAN

Few points can be concluded from the graph

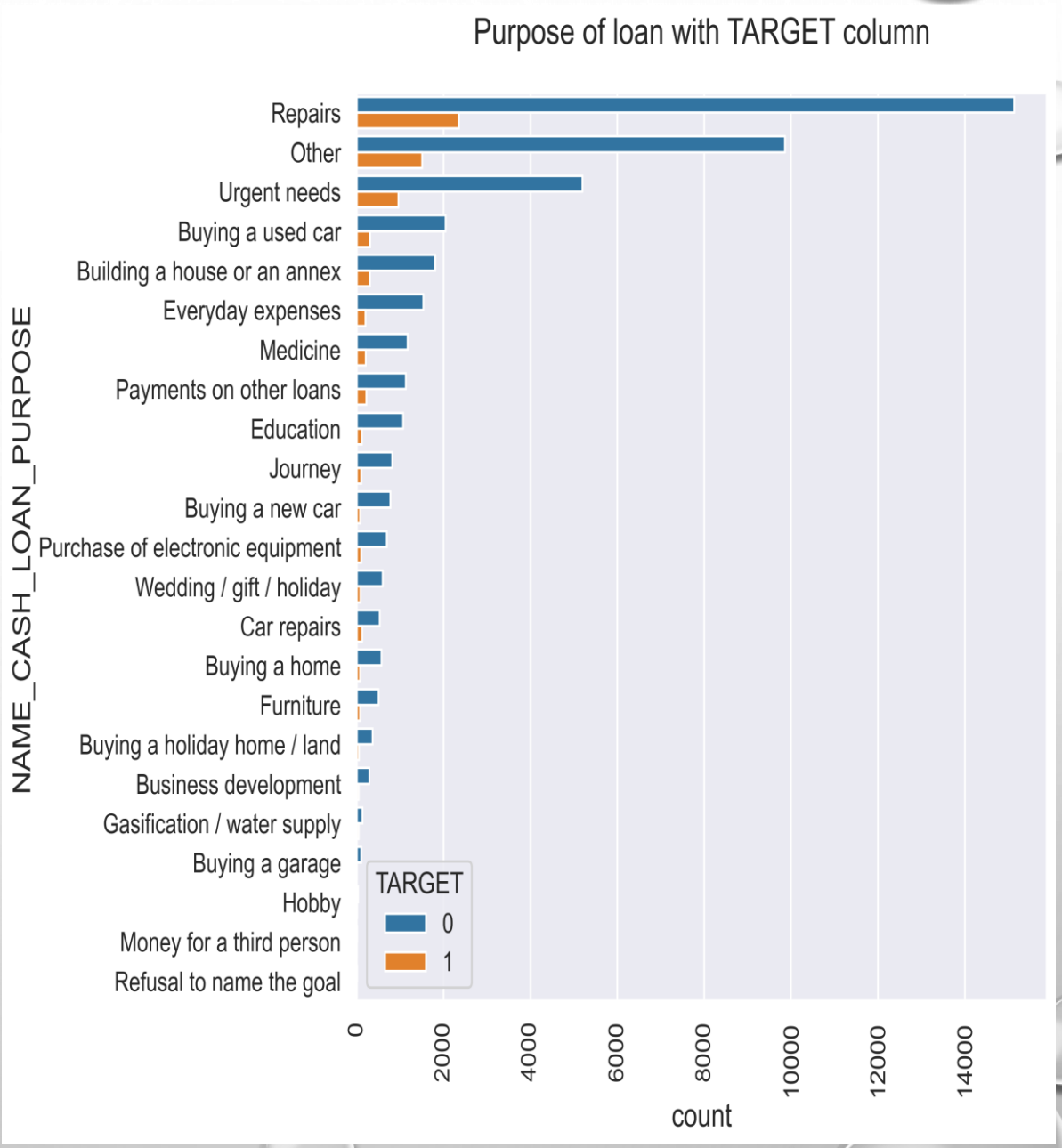
- Most of loan rejection was from 'repairs



PURPOSE OF LOAN WITH TARGET COLUMN

Few points can be concluded from the graph

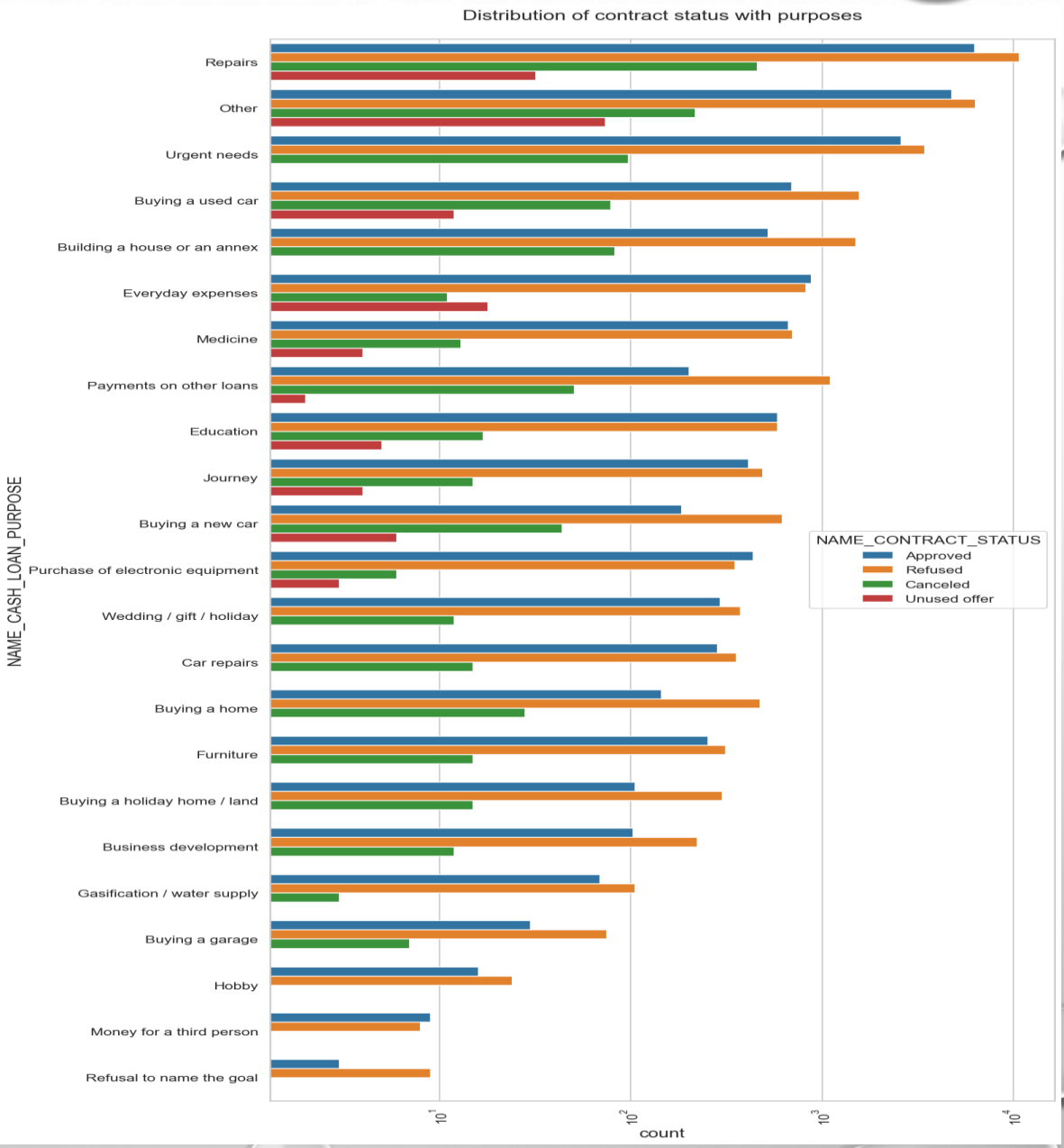
- Most of loan rejection was from 'repairs



DISTRIBUTION OF CONTRACT STATUS WITH PURPOSES

Few points can be concluded from the graph.

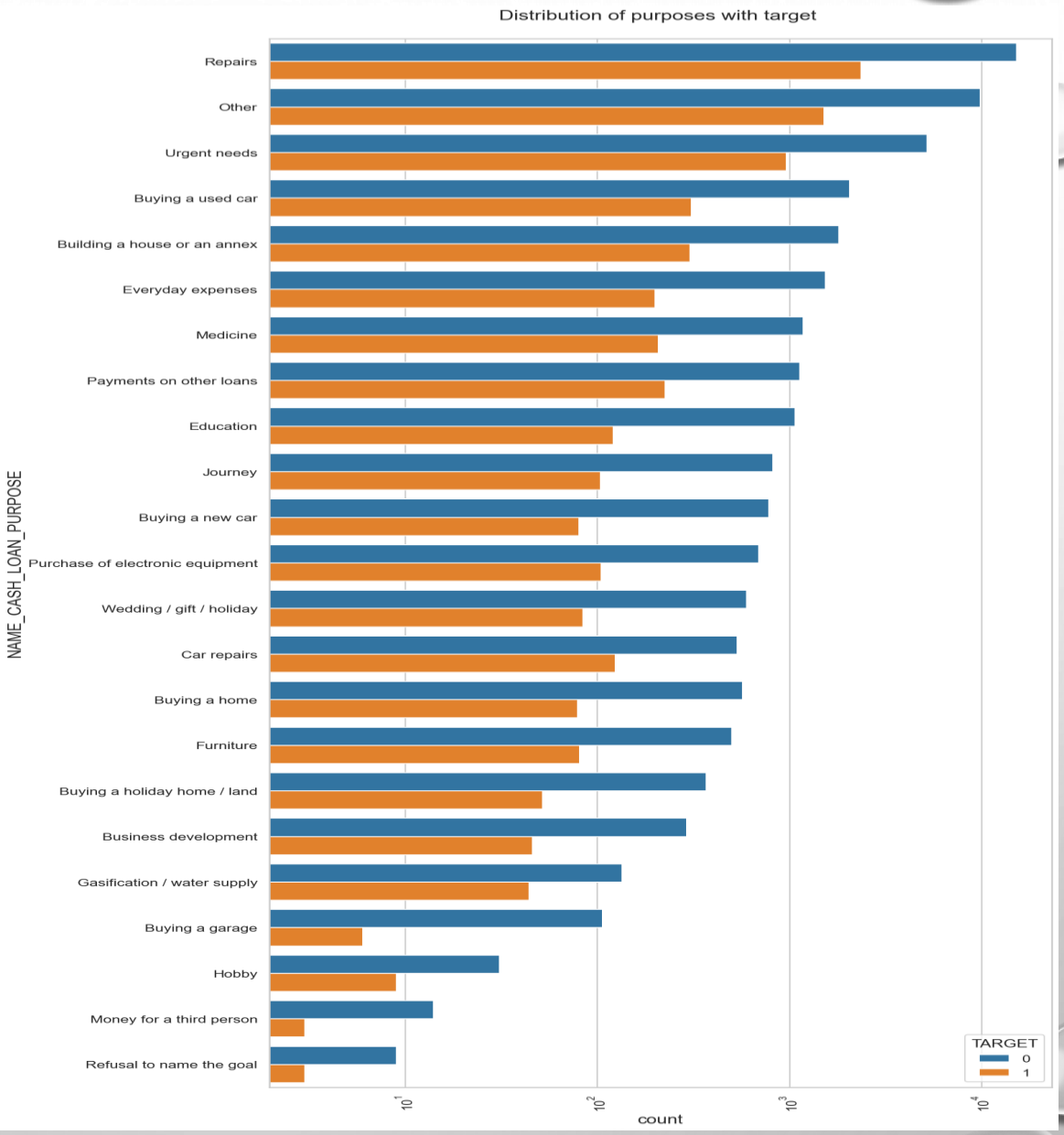
- Most rejection of loans came from purpose 'repairs'.
- For education purposes we have equal number of approves and rejection
- Paying other loans and buying a new car is having significant higher rejection than approves.



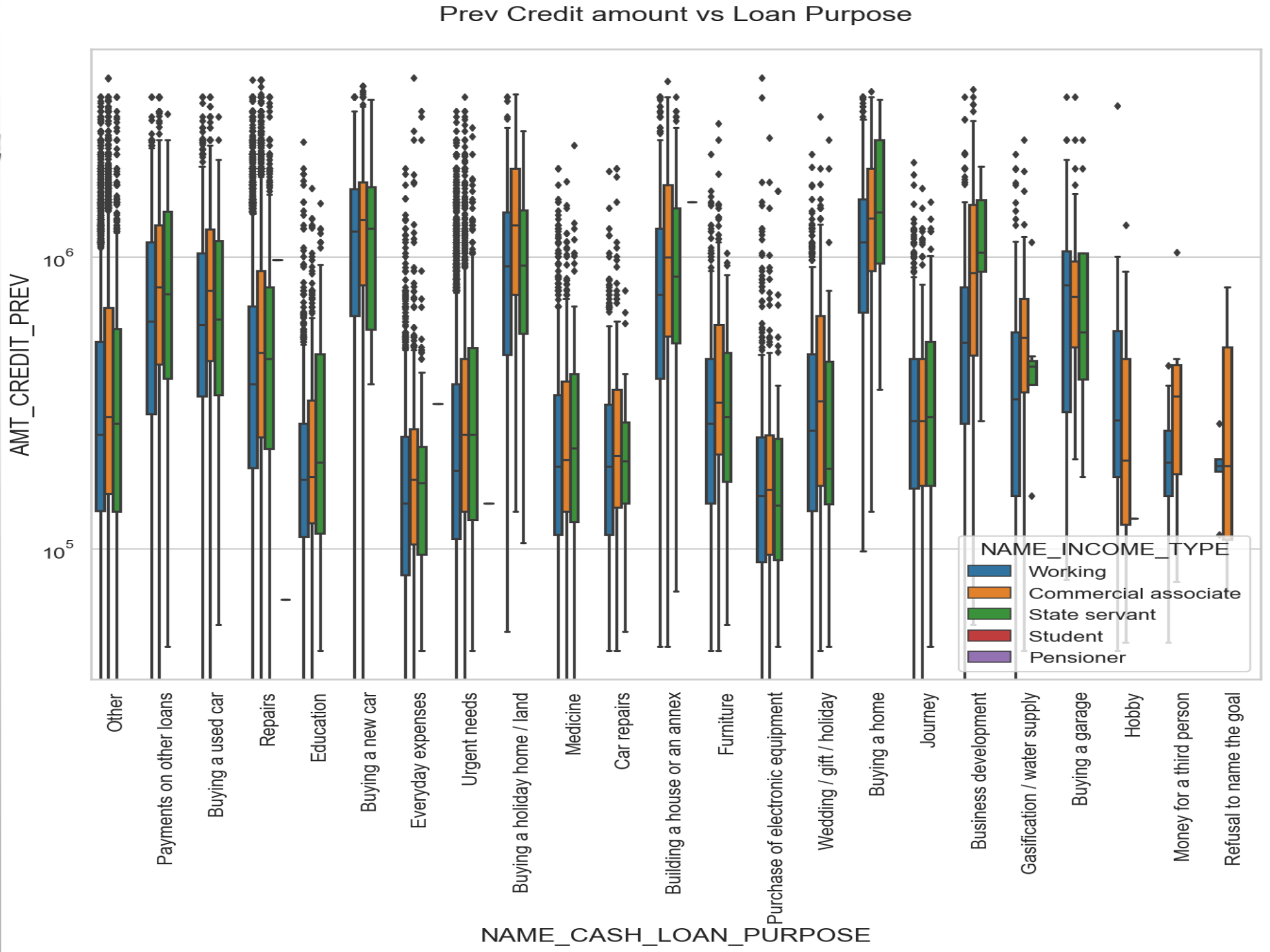
DISTRIBUTION OF PURPOSES WITH TARGET

Few points can be concluded 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.



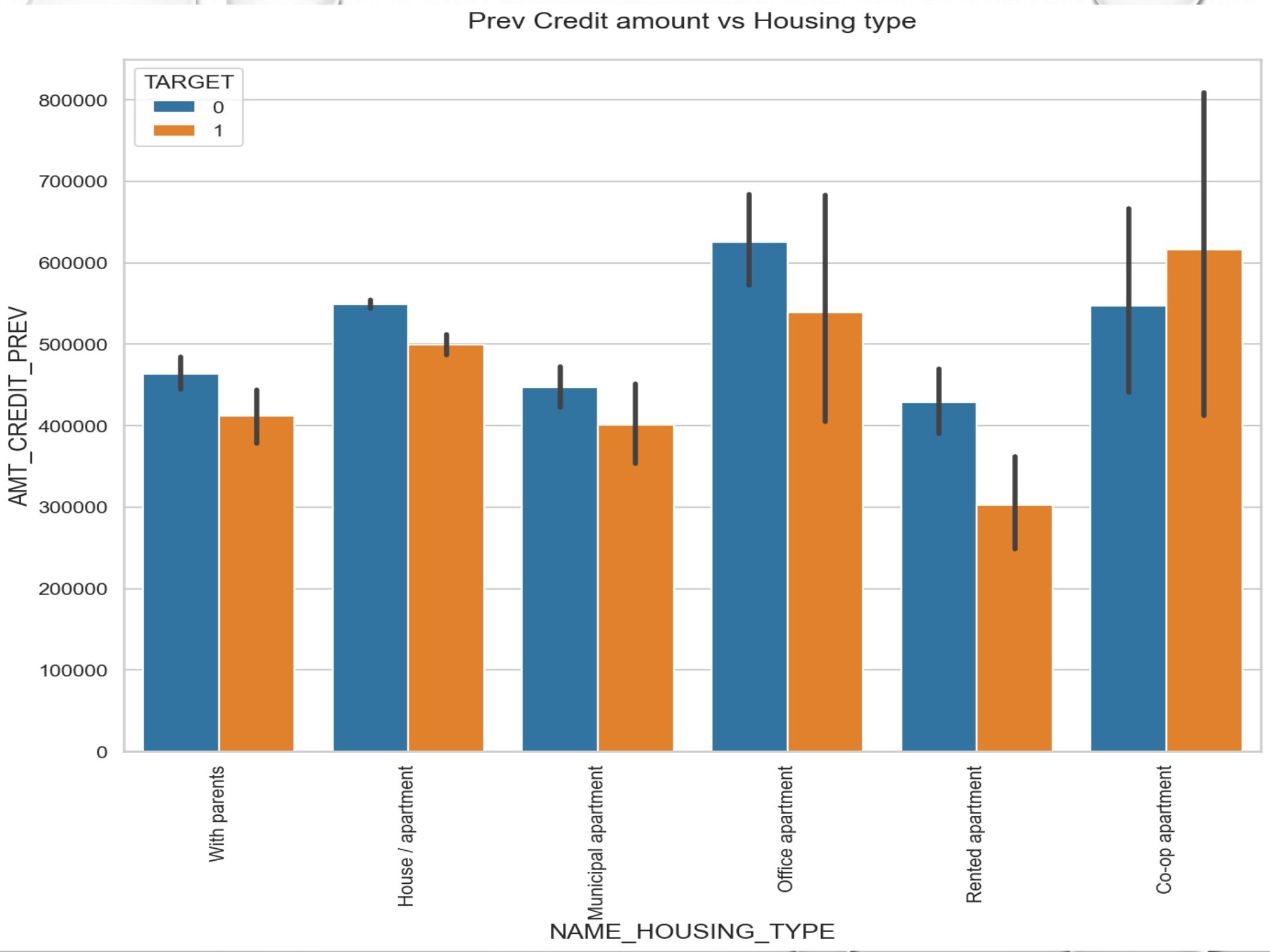
Performing Bivariate Analysis


Prev Credit amount vs Loan Purpose
AMT_CREDIT_PREV
10^6
10^5
NAME_INCOME_TYPE
Working
Commercial associate
State servant
Student
Pensioner
Other
Payments on other loans
Buying a used car
Repairs
Education
Buying a new car
Everyday expenses
Urgent needs
Buying a holiday home / land
Medicine
Car repairs
Building a house or an annex
Furniture
Purchase of electronic equipment
Wedding / gift / holiday
Buying a home
Journey
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

From the previous graph we can conclude the below points:

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


Prev Credit amount vs Housing type
TARGET
0
1
AMT_CREDIT_PREV
800000
700000
600000
500000
400000
300000
200000
100000
0
With parents
House / apartment
Municipal apartment
Office apartment
Rented apartment
Co-op apartment
NAME_HOUSING_TYPE

PREVIOUS CREDIT AMOUNT VS HOUSING TYPE

Few points can be concluded from the graph.

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

CONCLUSION

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

Thank You