Credit EDA Case Study

Presented By:

Saurabh Mudgal

Shashi Ranjan Kumar

Introduction

This case study aims to give you an idea of applying EDA in a real business scenario.

In this case study, apart from applying the techniques, we have also developed 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

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Due to which, some consumers use it as their advantage by becoming a defaulter.

Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze 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.

If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial e company.

The given data 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 instalments of the loan in our sample,

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

Business Objective

- This case study aims 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 (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.

- In other words, 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 utilize this knowledge for its portfolio and risk assessment.

Steps Involved

Data understanding and preparation

- Data Sourcing
- Data Inspection
- Inspecting the null values

Data Cleaning and Manipulation

- Check data types
- Conversions from negative to positive
- Outliers

Data Analysis

- Data imbalance
- Categories Target 0 and target 1
- Correlation matrix analysis
- Univariate Analysis and Bivariate Analysis for numerical variable for target 0 and target 1
- Load the previous data
- Join previous and application data

Handling Missing Data

- In the process of data cleaning and manipulation we dropped the columns having null values greater than 45%.

- There were few columns like: 'AMT_ANNUITY', 'DAYS_LAST_PHONE_CHANGE' which had very less null values comparatively to others , these columns were imputed mean and median.

- Since some of the columns having null value percentage below 45 and seemed to be less relevant to showcase the analysis, so below columns were dropped from the data frame:

- 'EXT_SOURCE_2', 'AMT_GOODS_PRICE', 'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE', 'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE'

- CODE_GENDER was having third value as XNA , which was converted to F(Female), since count for female applicants is higher than male.

Change in Datatype and Negative values

- Some of the variables were changed into numeric Data type:

- 'TARGET','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', 'LIVE_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY'

- Below columns had negative values which were converted into positive values using abs() function:

- 'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH'

Data Imbalance

- Target 1 ,is categorized for the clients who are having difficulties with payments.

- Target 0 , is categorized for clients who don't have any difficulties in repay the credit.

- Target 0 has the maximum values which is 10.55 %

CATEGORICAL UNIVERSAL ANALYSIS FOR TARGET 0

Conclusion for AMT_INCOME_RANGE from Target 0 based on CODE_GENDER:

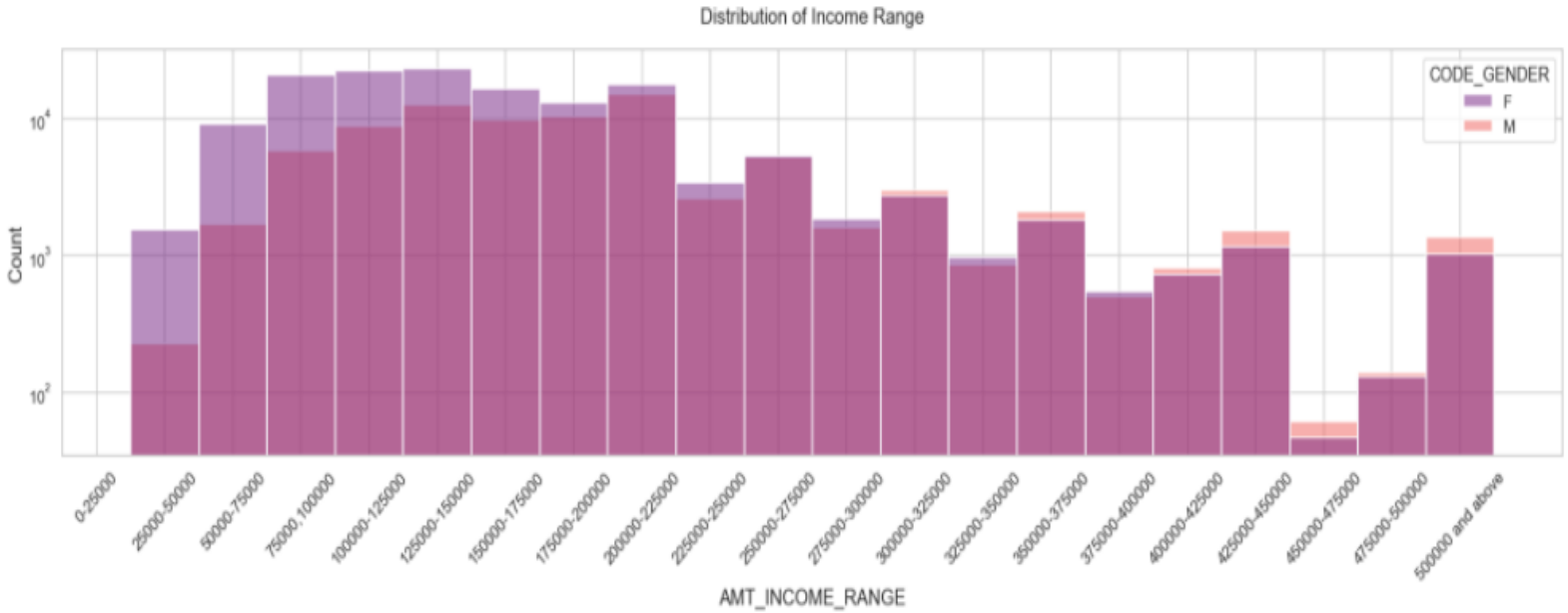
1.Female counts higher than male.

2.Income range from 100000 to 225000 is having a greater number of credits.

3.Plot clearly shows that females are having more credits than males for the same ranges.

4.On and above 400000 credit count is very less.

5.For income range 25000-50000 females have way more credits than male.



Conclusion for AMT_CREDIT_RANGE from Target 0 based on CODE_GENDER:

1.Females have more credit than males for each credit range.

2.for range 700000-750000 credit count is very less.

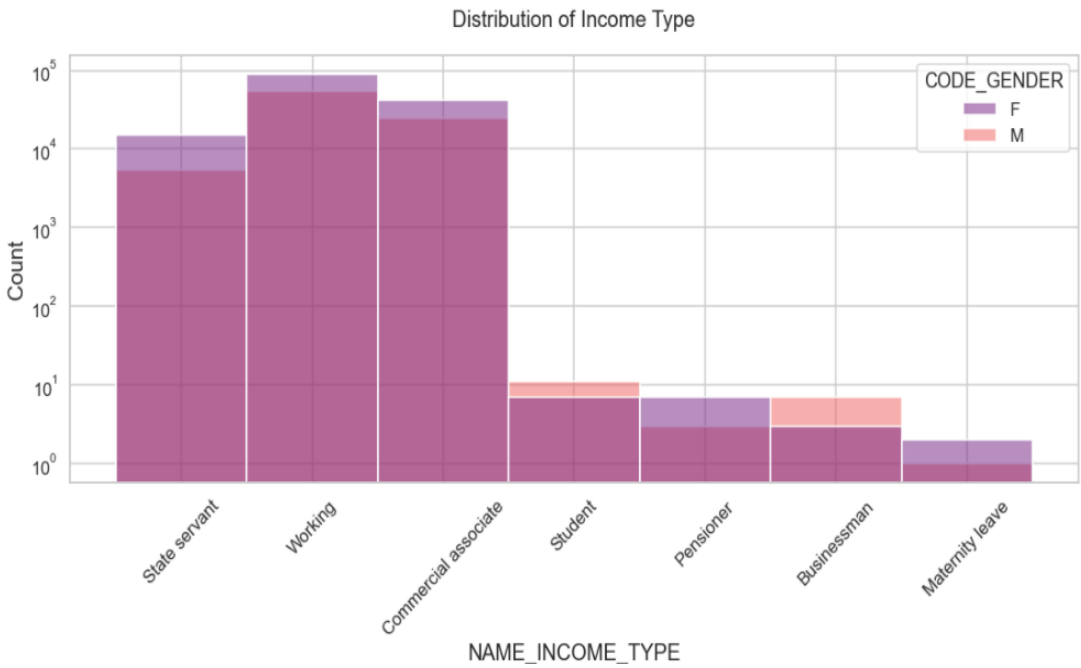
Conclusion for NAME_INCOME_TYPE from Target 0 based on CODE_GENDER:

1.There are clearly three types which are on top in the category or have the higher credit count as compare to other categories.

2.Here also, Females are having more credits in these top three income types.

3.Student income type is having more credit than Businessman , pensioner and Maternity leaves.

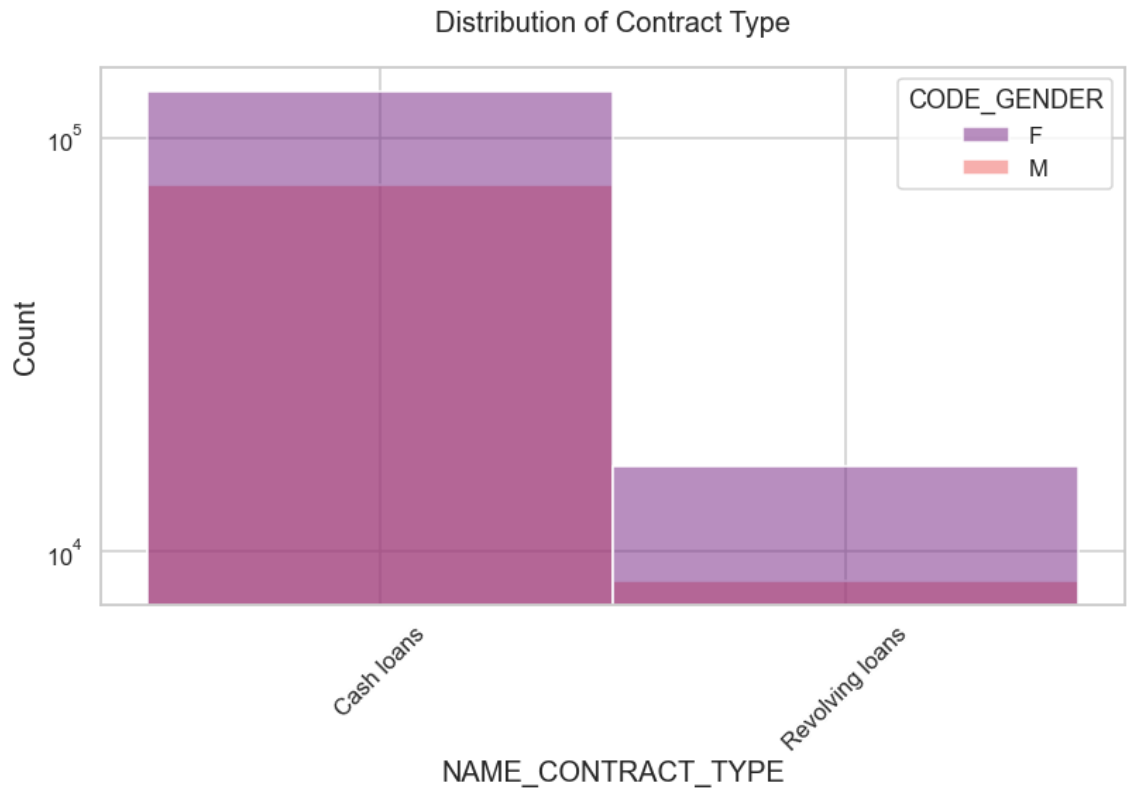
4.There are only two categories where male have more income than female i.e., 'Student', 'Businessman'.



Conclusion for NAME_CONTRACT_TYPE from Target 0 based on CODE_GENDER:

1.'Cash loans' has more credits than 'Revolving loans' contract type.

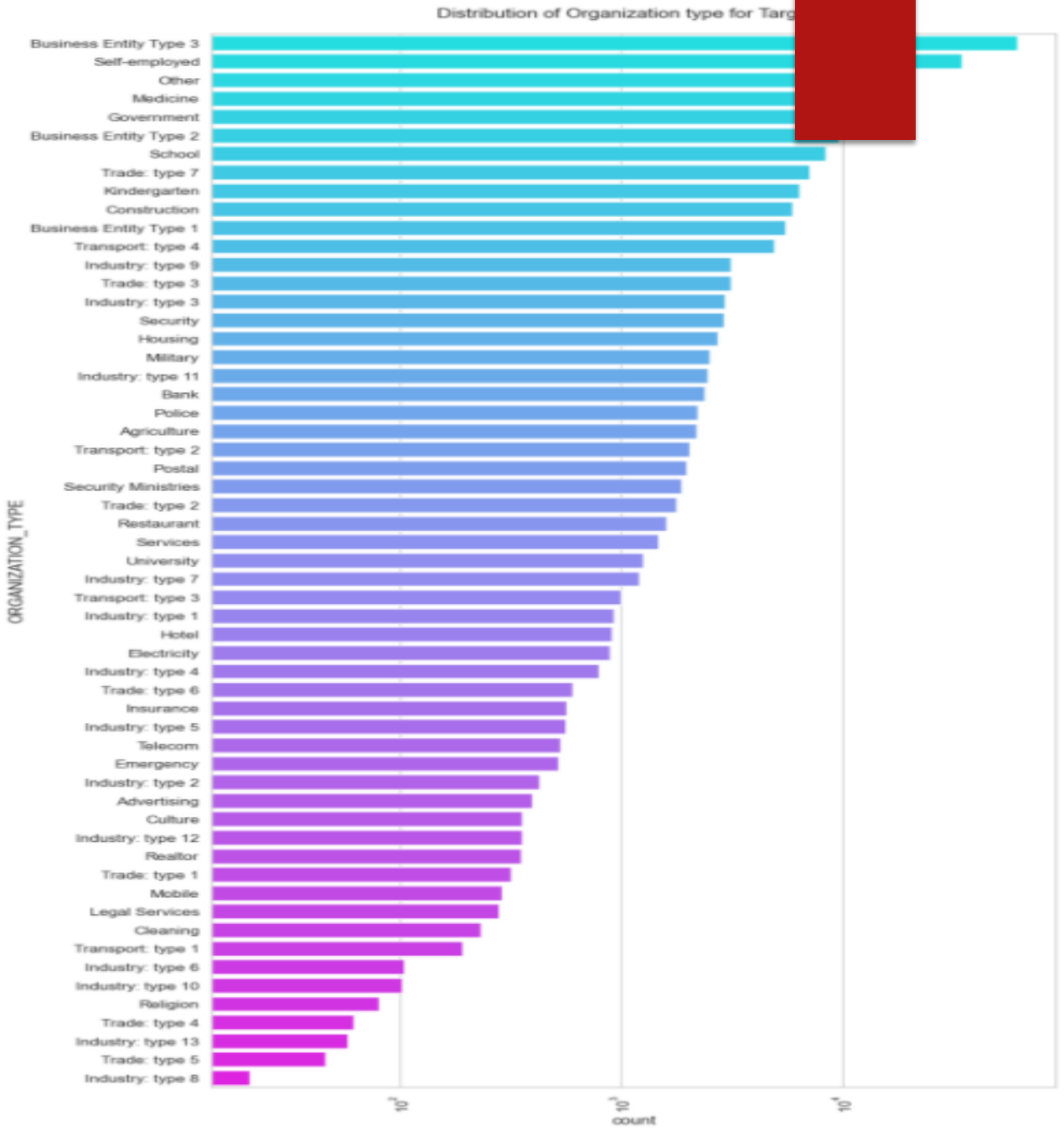
2. Also, female category is leading here as well.



Conclusion for ORGANIZATION_TYPE from Target 0:

1.Top 5 categories are 'Business entity Type 3' 'Self employed' , 'Other' , Medicine' and 'Government' which have applied for the loan.

2.Categories which have applied very less are Industry type 8, type 13 and Trade type 5, type4 and Religion.



CORRELATION FOR TARGET 0 & TARGET 1

Conclusion for Target 0 Correlation:

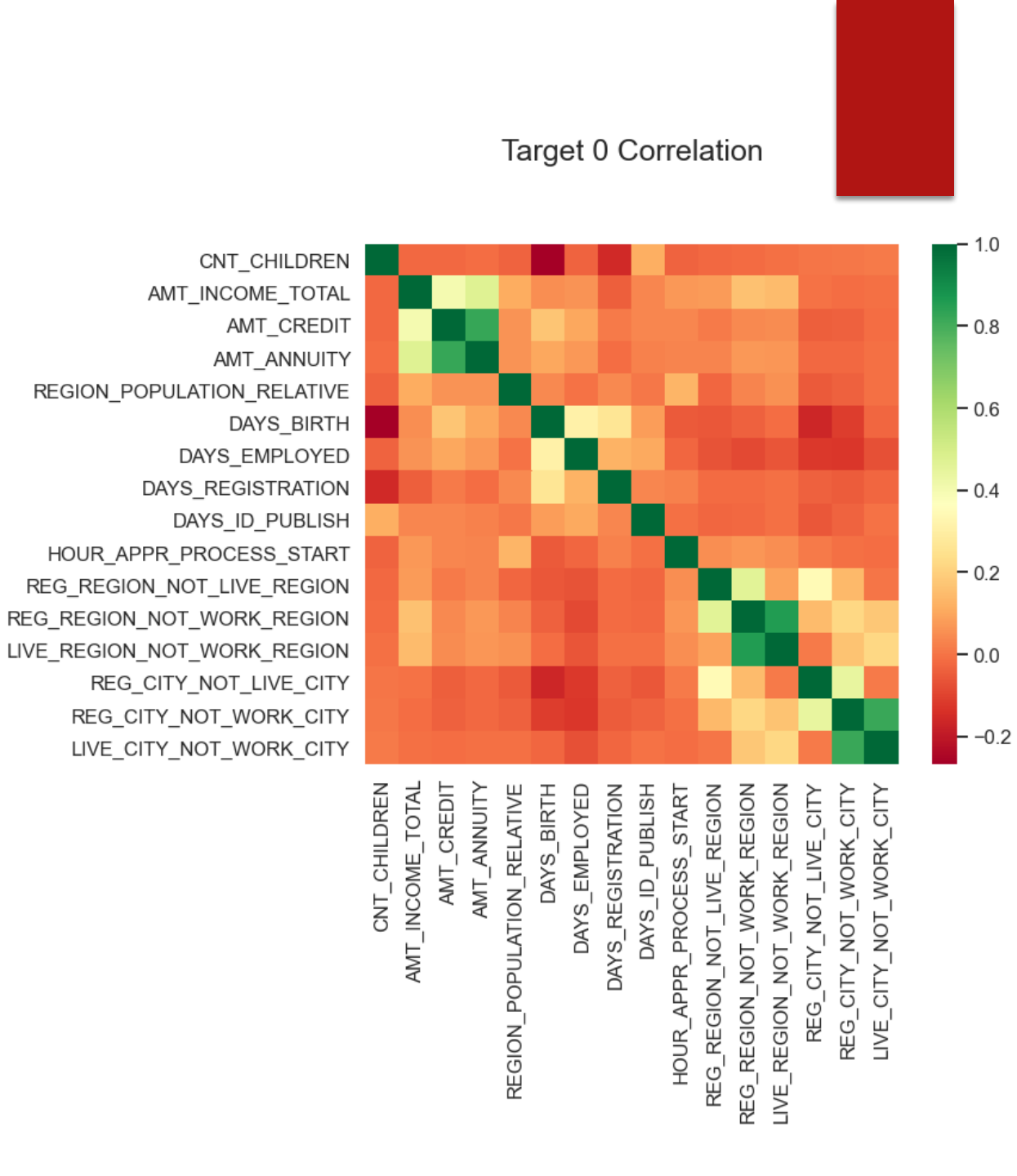
1.Credit amount is inversely proportional to the number of children client have, that means Credit amount is higher for less children count client have and vice-versa.

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

3.Customers with less children have in densely populated area.

4.Credit amount is higher to densely populated area.

5.The income is also higher in densely populated area.



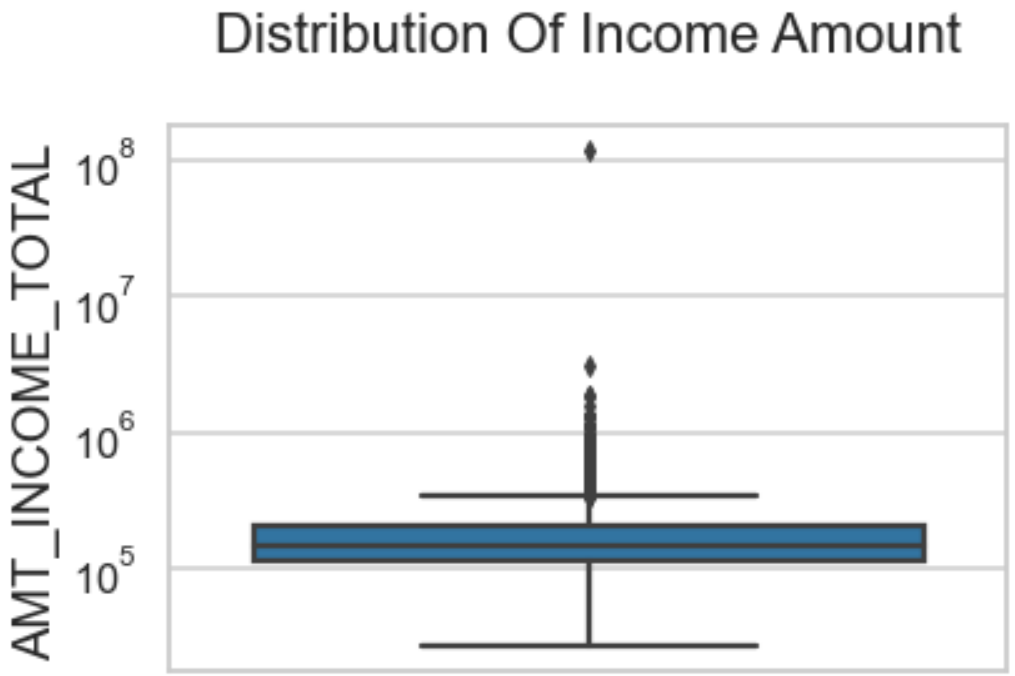
Conclusion for Target 1 Correlation:

1. Customers having less children, their permanent address does not match work address and vice-versa.
2. Customers having less children, their permanent address does not match contact address and vice-versa.

HANDLING OUTLIERS

Conclusion for AMT_INCOME_TOTAL from Target 0:

1. Third quartile is very narrow and has very less values compare to other quartiles.
2. There are some outliers as well in income amount.

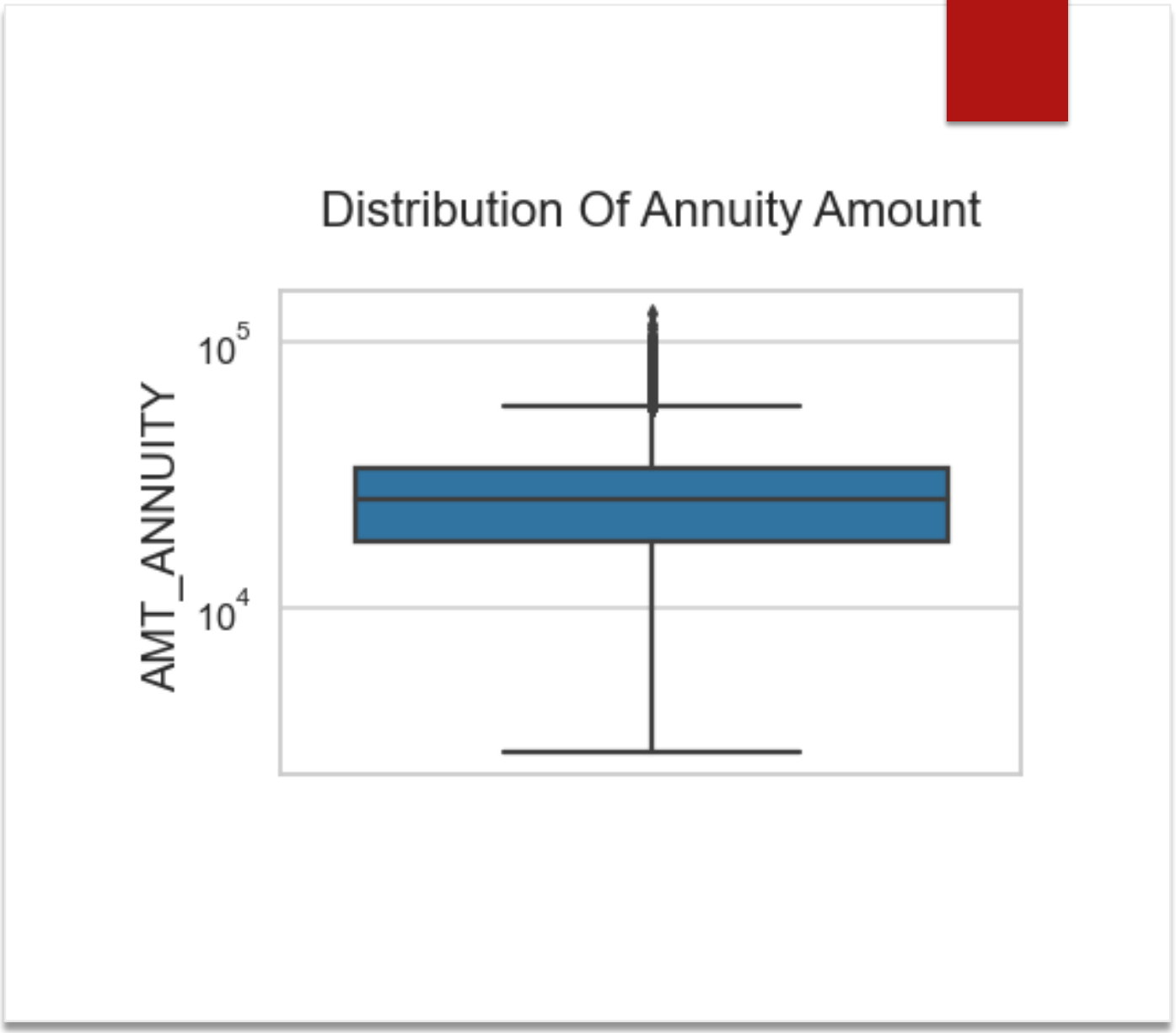


Conclusion for AMT_CREDIT from Target 0:

1. Credit amount also has some outliers in it.
2. As per the box plot first quartile is bigger than 75th quartile for credit amount which means loan of most people lie in the first quartile.

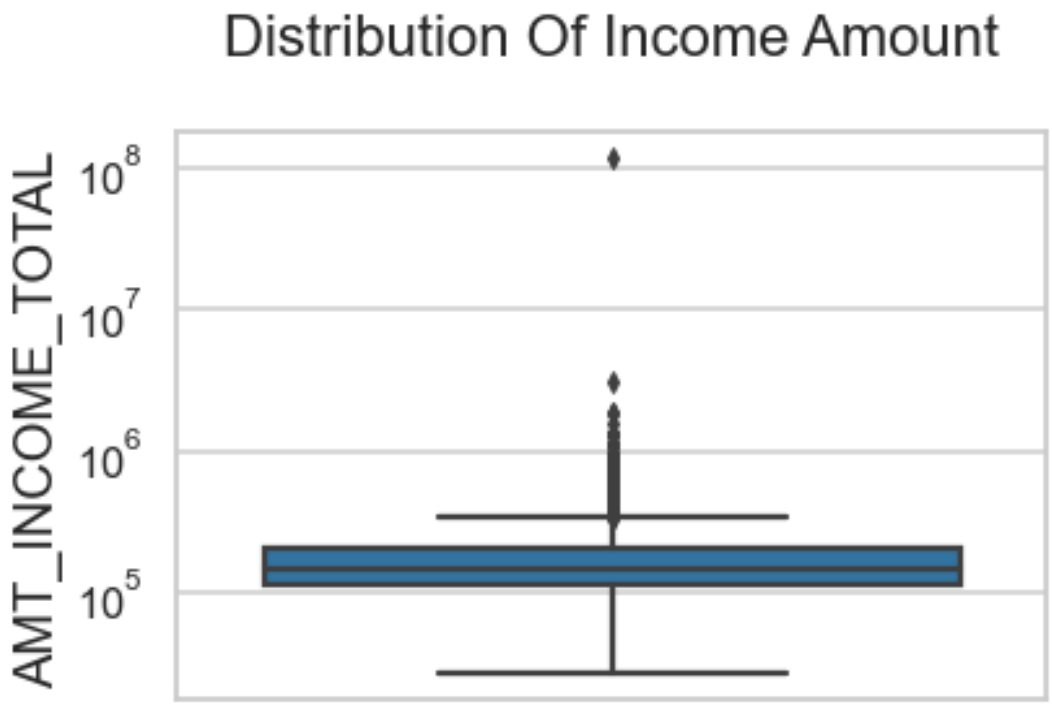
Conclusion for AMT_ANNUITY from Target 0 :

1. Annuity amount also has some outliers.
2. First quartile is way bigger than third quartile which means maximum number of clients are from first quartile.



Conclusion for AMT_INCOME_TOTAL from Target 0 :

1. Income amount has some outliers.
2. First quartile is greater than third quartile.
3. Maximum income clients are present in first quartile.

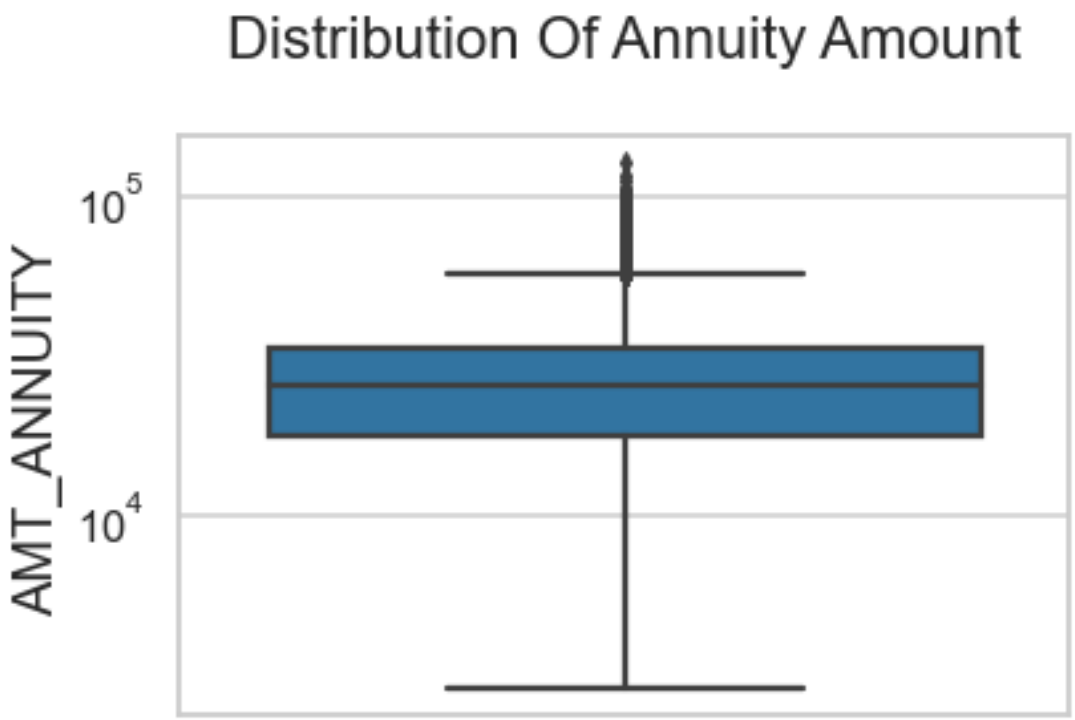


Conclusion for AMT_CREDIT from Target 1:

1. Credit amount has some outliers.
2. First quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.

Conclusion for AMT_ANNUITY from Target 1:

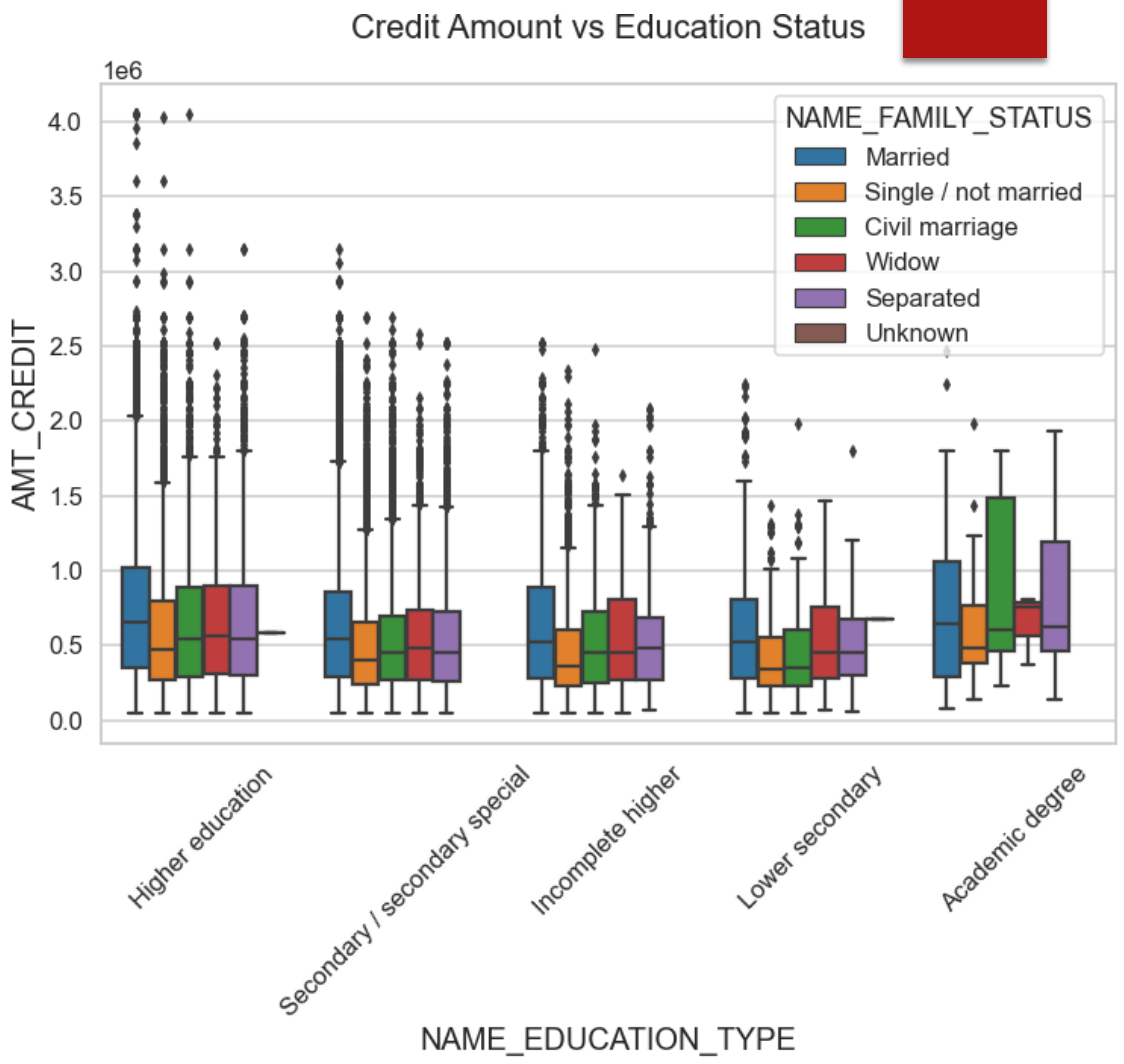
1. Annuity amount has some outliers.
2. First quartile is bigger than third quartile for annuity amount which means most of the annuity clients are present in first quartile.



BIVARIATE ANALYSIS

Conclusion for Credit amount v/s Education Status Target 0 :

1. Clients with Academic degree and with Family Status of 'Civil marriage' , 'Married', 'Separated' are having higher credits than others.
2. Clients with Higher education and with Family Status of 'Married' ,'Single', 'Civil Marriage' are having more outliers.
3. Civil Marriage for Academic degree is having most of the credits in the third quartile.
4. Widow for Academic degree is having least number of credits in third quartile.



Conclusion for Income Amount v/s Education Status from Target 0 :

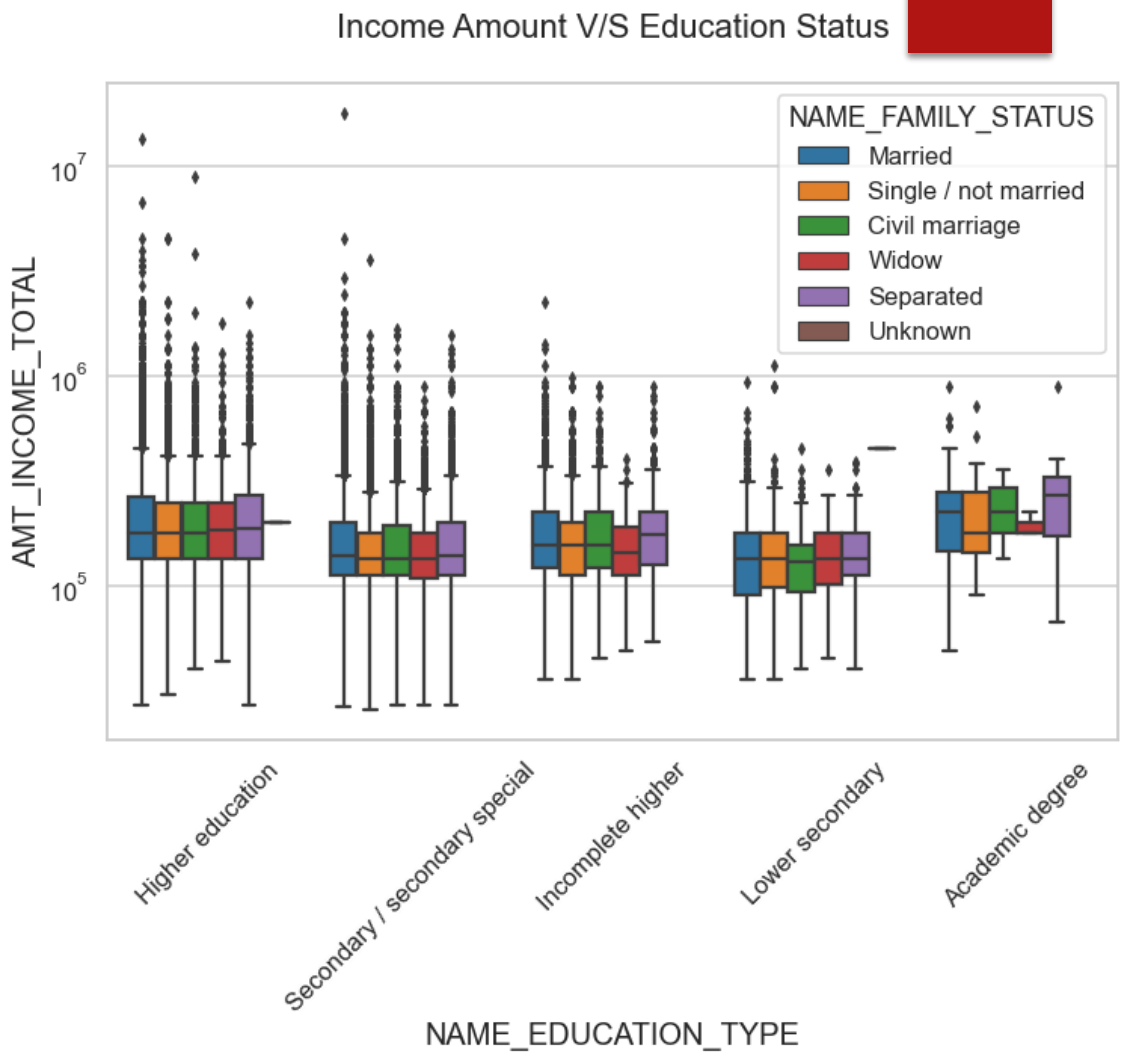
1.Credit amount is inversely proportional to the number of children client have, that means Credit amount is higher for less children count client have and vice-versa.

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

3.Customers with less children have in densely populated area.

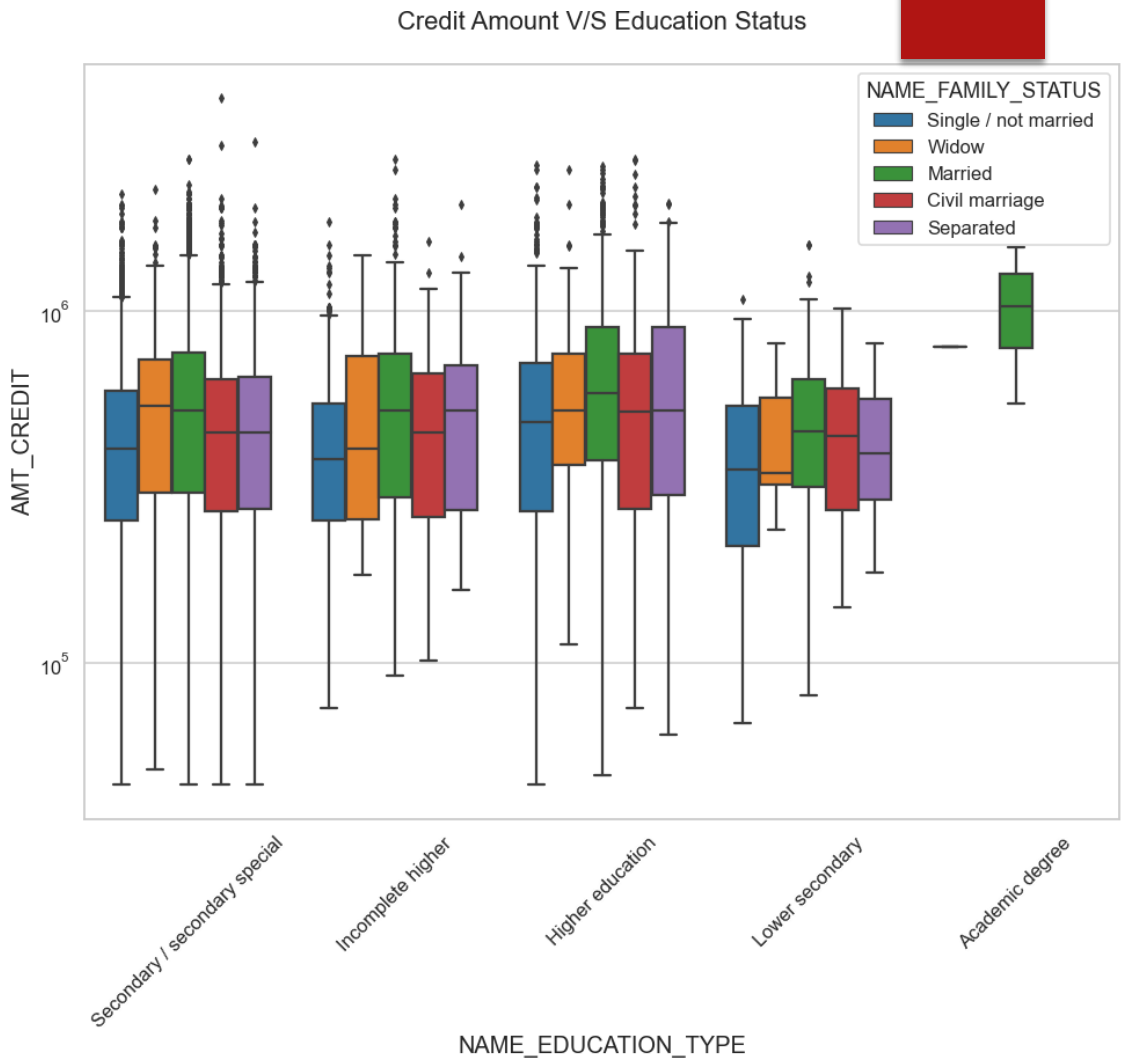
4.Credit amount is higher to densely populated area.

5.The income is also higher in densely populated area.



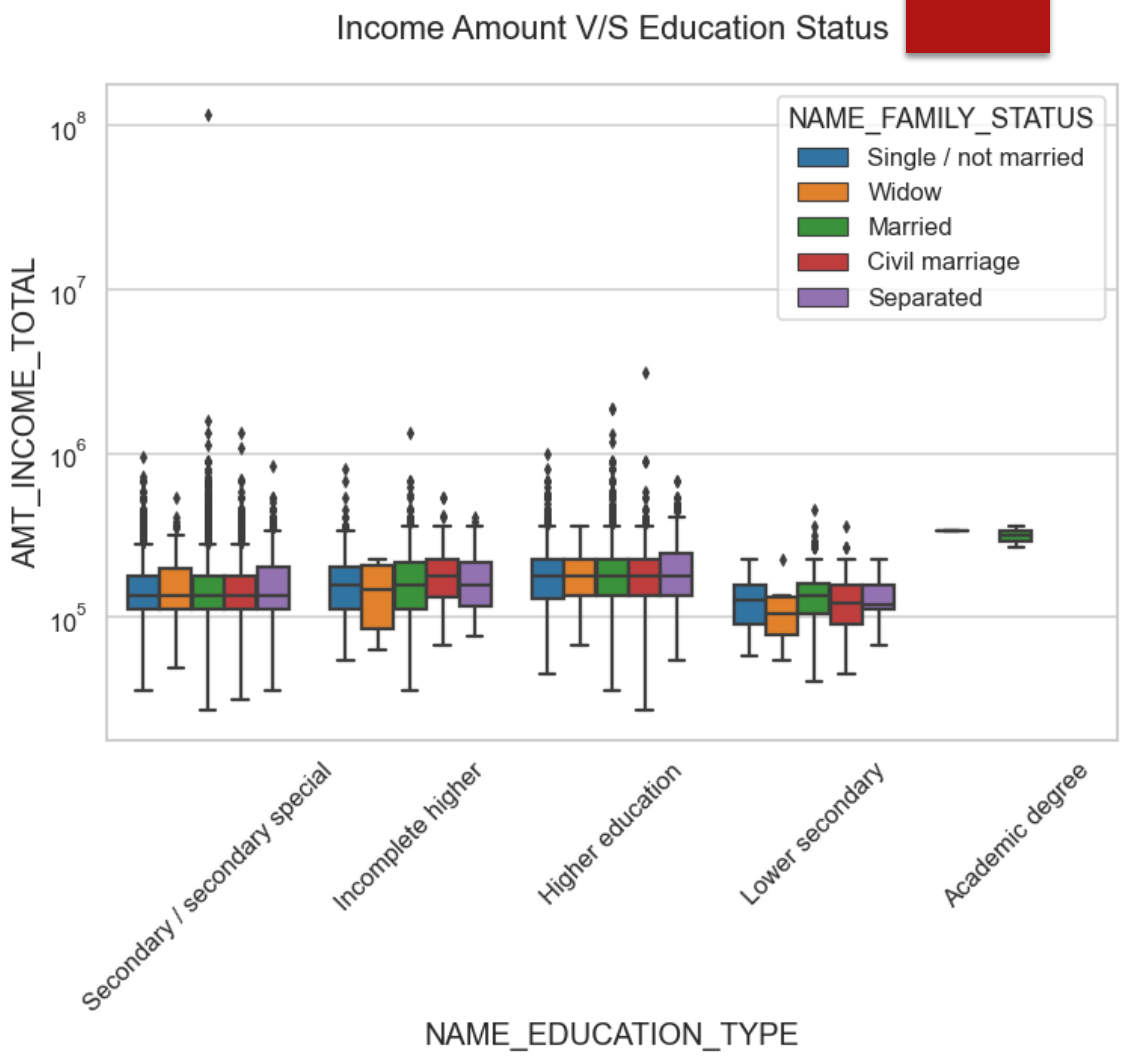
Conclusion for Credit Amount v/s Education Status from Target 1:

1. For Academic degree only 'Married' Family Status has some significant and highest credits.
2. Higher Education, Secondary Education , Incomplete Higher are the top 3 categories with outliers.
3. For Higher Education 'Married' Family status has maximum credits in first quartile.



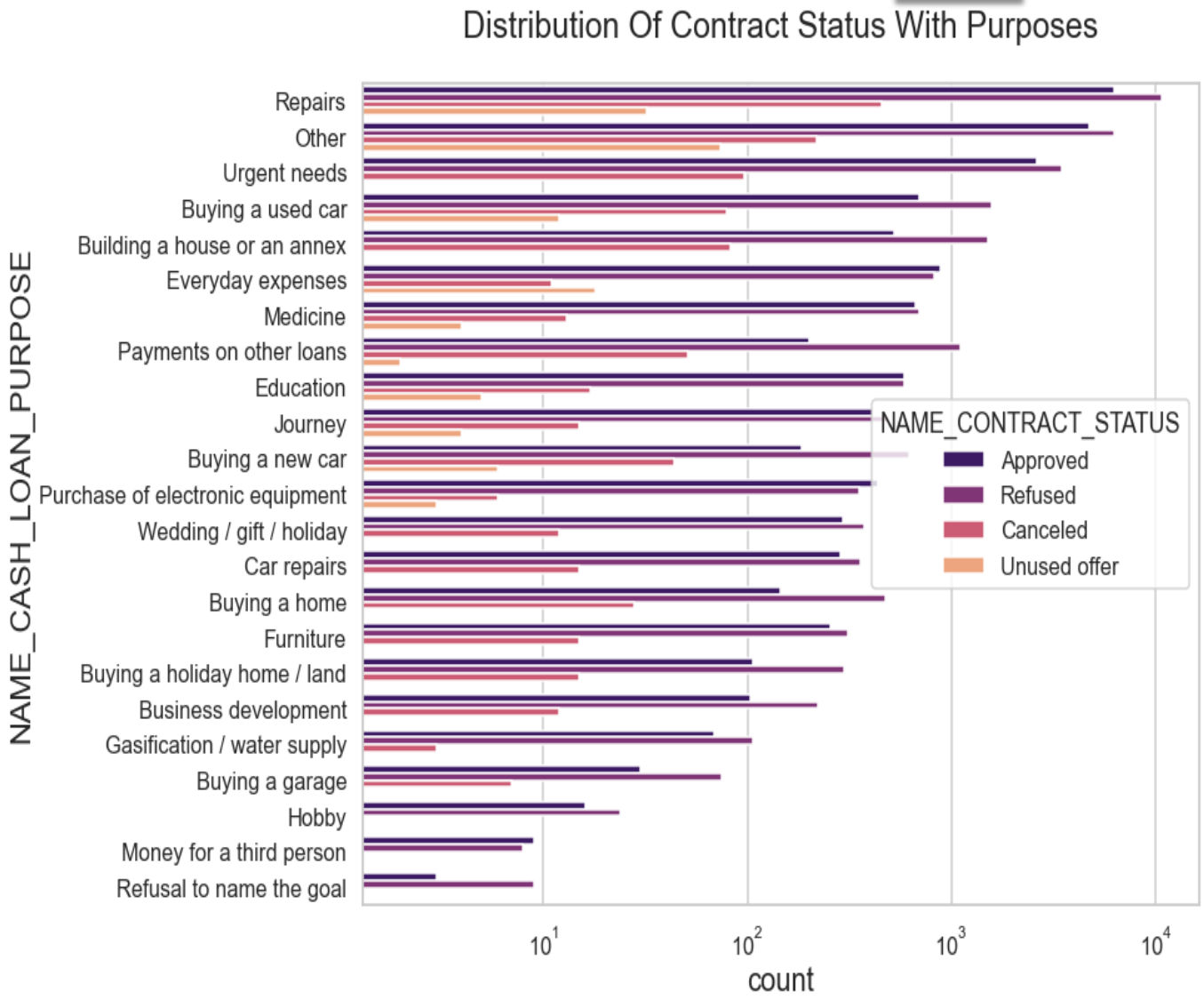
Conclusion for Income Amount v/s Education Status from Target 1:

1. Total income also has outliers.
2. Higher education each family status has the equal income.
3. Higher education has the same mean for each family status.
4. Secondary education also has the same mean for each family status.



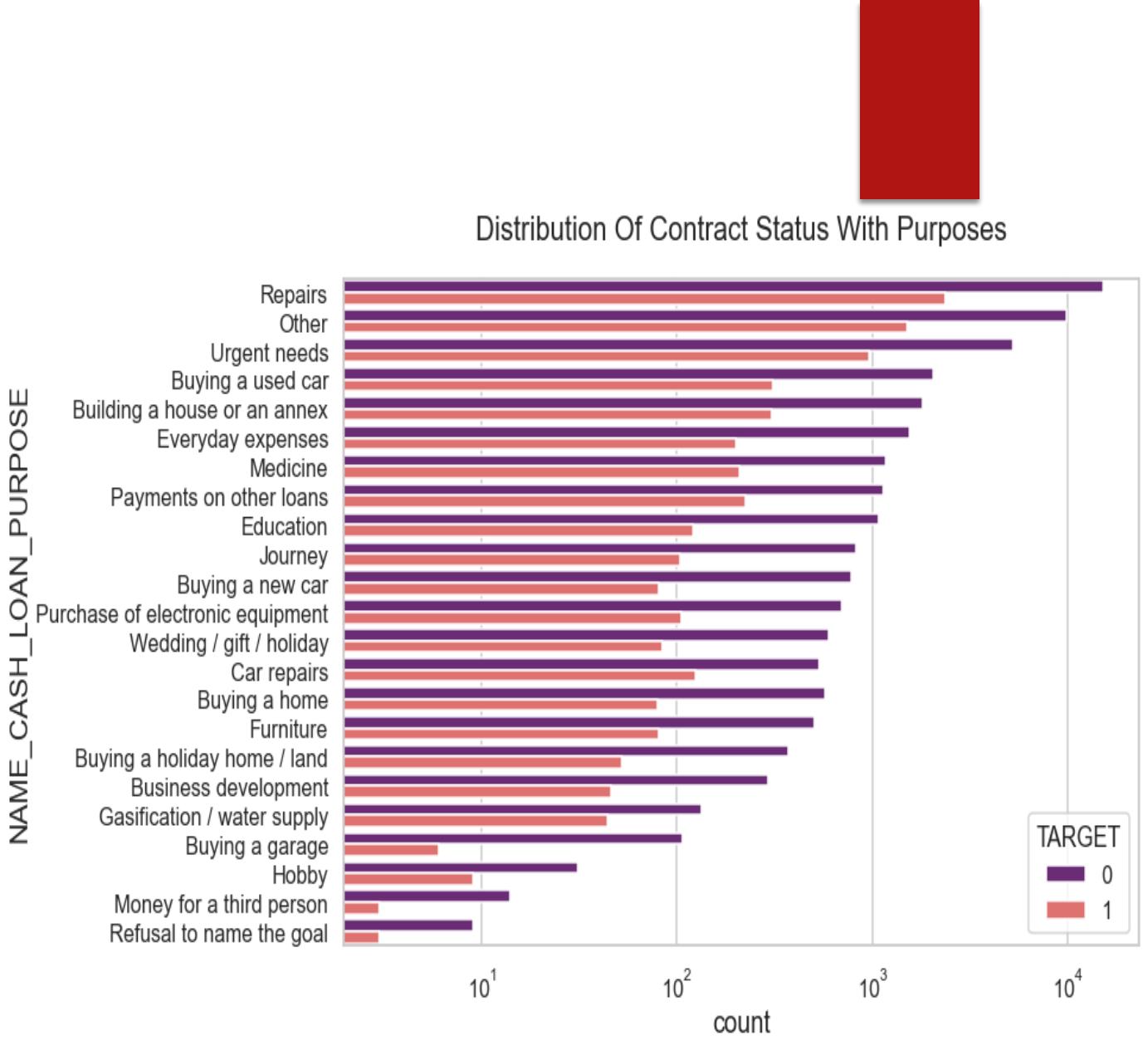
Conclusion based on Contract Status of ‘Purposes of Credit’ from previous application history and current application:

1. 'Payments on other loans' has most refused rate and 'Buying a new car' also has most refused rate than approved.
2. 'Education' purposes has the same approved and refused rate.



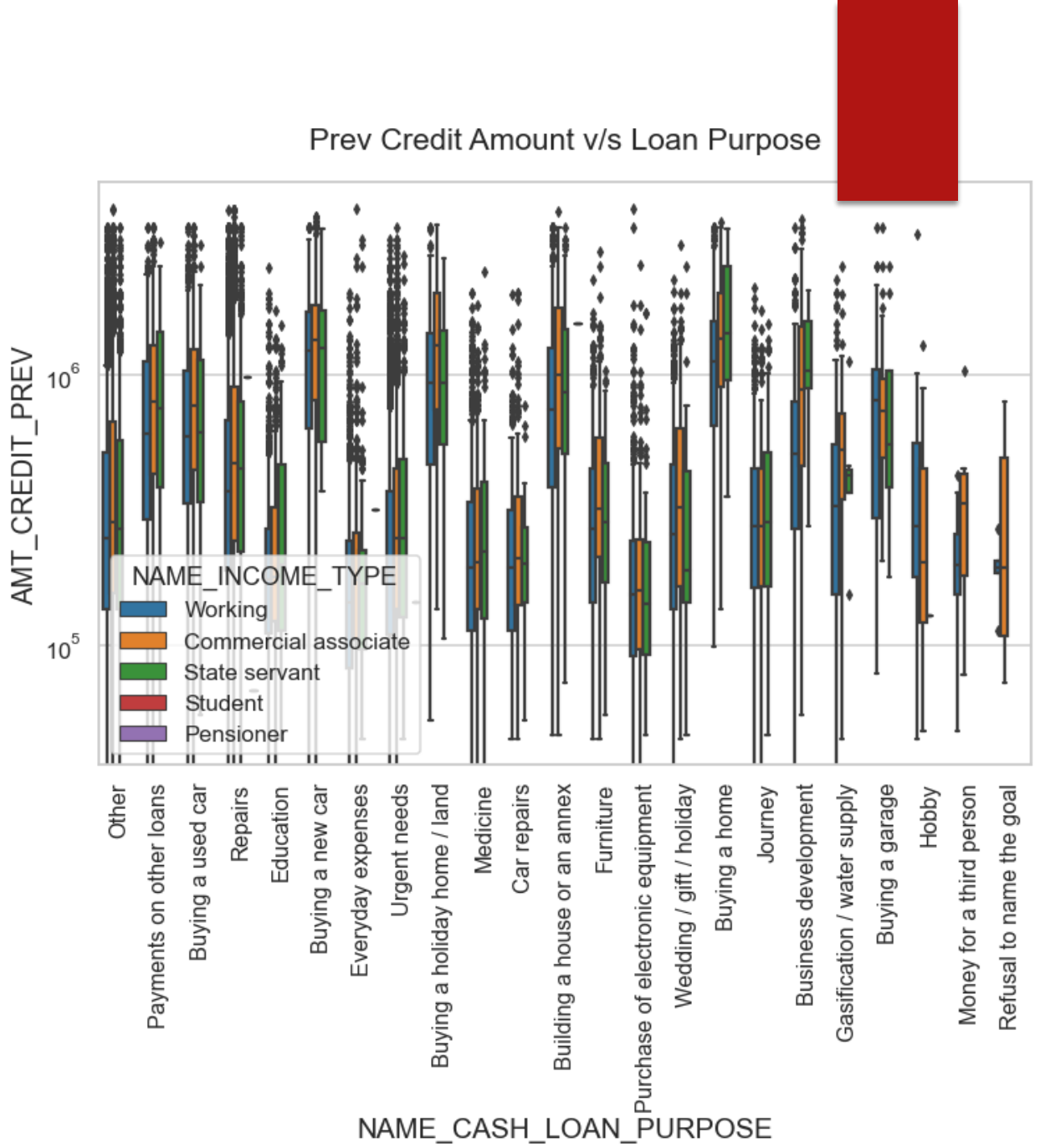
Conclusion based on Target0 and Target1 of ‘Purposes of Credit’ from previous application history and current application:

1. Loan purposes with 'Repairs' are facing more difficulties in payment on time.
2. 'Buying a garage', 'Business Development', 'Buying Land', ''Buying a new car', 'Education', are the loan purposes which are having minimal difficulties in payment.



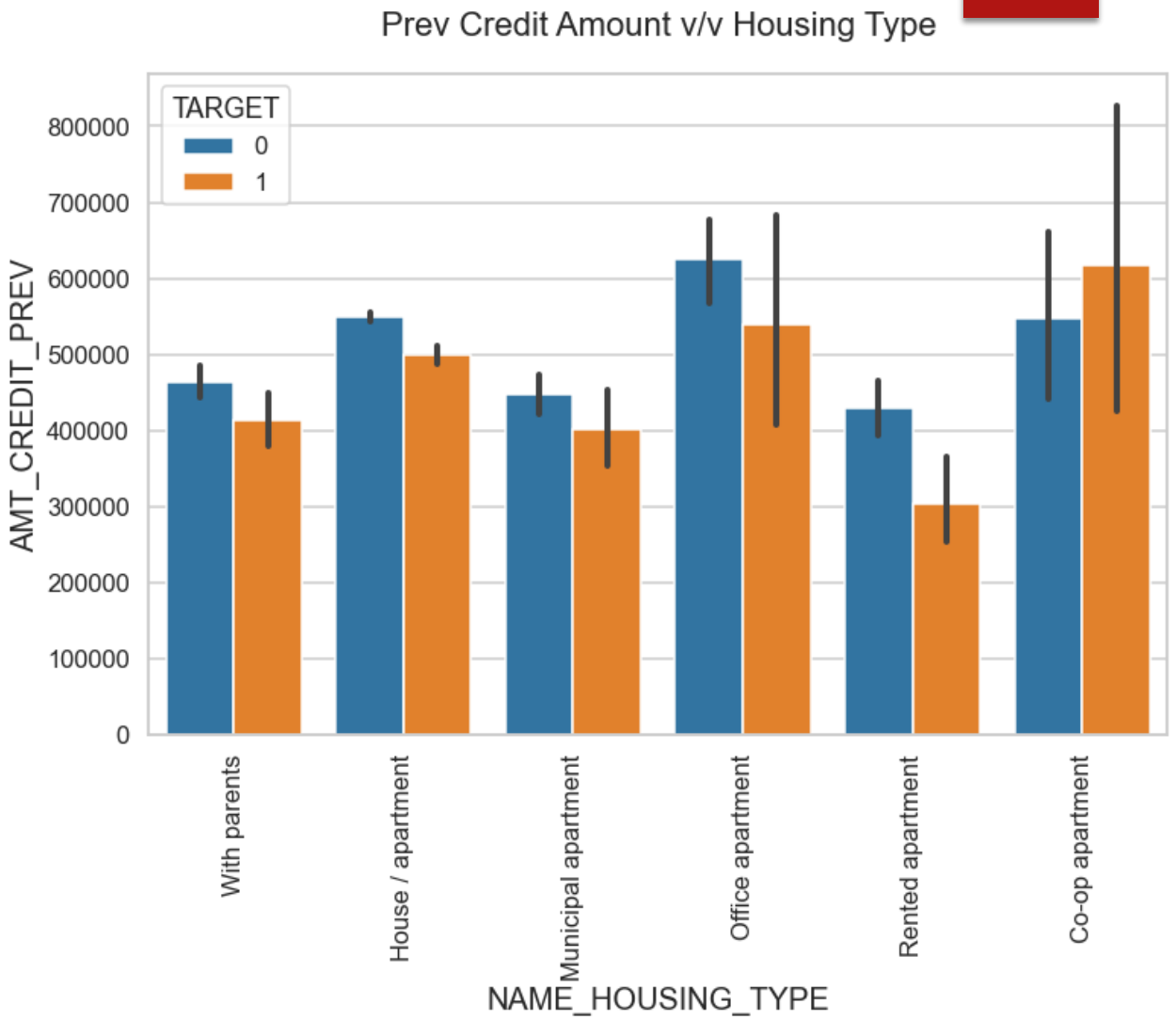
Conclusion Previous amount credit v/s Loan purposes :

1. 'Servant' income type category seems to have applied in significant amount.
2. 'Money for a third person' has less credit applied for.
3. 'Buying a land', 'Buying a home', 'Building a home', 'Buying a car', have higher credit amount loan.



Conclusion on Previous Credit amount and Housing Type:

1. 'Co-op apartment' shows difficulties in repay the loan amount as the bar of Target 1 is high, so bank needs to be careful about this category.
2. 'Office apartment' shows good loan repayment response as bar of Target0 is high , so bank should focus more on this category.



Conclusion:

1. 1. Loan purpose 'Repair' has higher number of unsuccessful payments on time.
2. 2. Bank should focus more on "student', 'Pensioner', 'Businessman' with housing type other than 'Co-op apartment for successful payments.
3. 3. Bank needs to be extra careful on income type 'Working' as they are having the greatest number of unsuccessful payments.
4. 4. Customer from housing type 'With Parents' since they are having minimum number of unsuccessful payments.

THANKS! ☺