Credit EDA – Case Study

By-Sarthak Mishra

Introduction

This Case Study aims to give you an idea of applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also 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. Suppose you work for a consumer finance company which specialises in lending various types of loans to urban customers. You have to use EDA to anayse 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 loss for the company.

Business Understanding -2

- The data given below 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.

- 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 or in some cases due to a higher risk of the client he received worse pricing which he did not want.

3. **Refused:** The company had rejected the loan (because the client does not meet their requirements etc.).

4. **Unused offer:** Loan has been cancelled by the client but on different stages of the process.

In this case study, we will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

Business Objectives

- 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.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics -understanding the types of variables and their significance should be enough).

Data Understanding

- This dataset has 3 files as explained below:

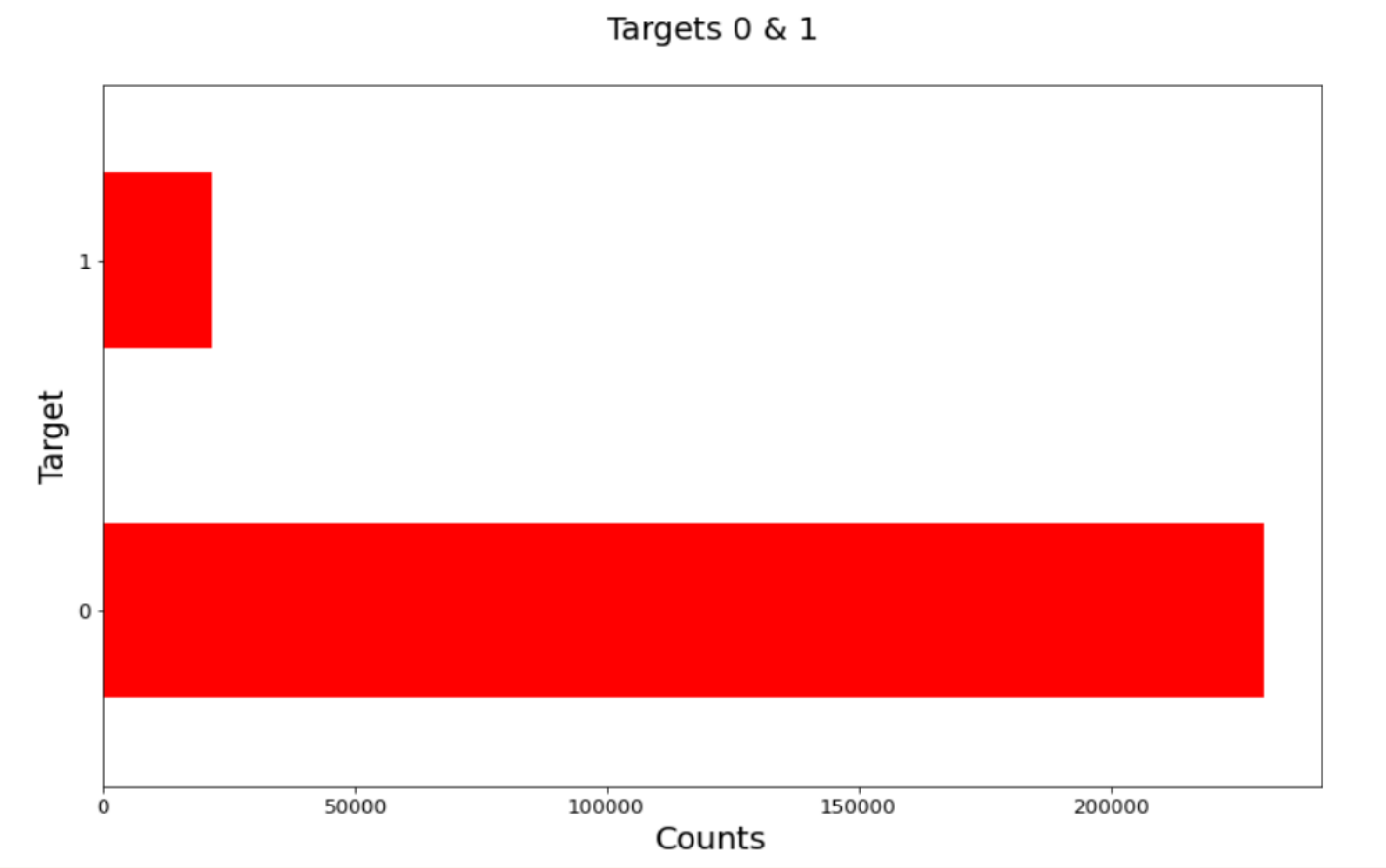
*1. 'application_data.csv'*contains all the information of the client at the time of application.The data is aboutwhether a**client has payment difficulties.**

*2. 'previous_application.csv'*contains information about the client's previous loan data. It contains the data whether the previous application had been**Approved, Cancelled, Refused or Unused offer.**

*3. 'columns_description.csv'*isdata dictionary which describes the meaning of the variables.

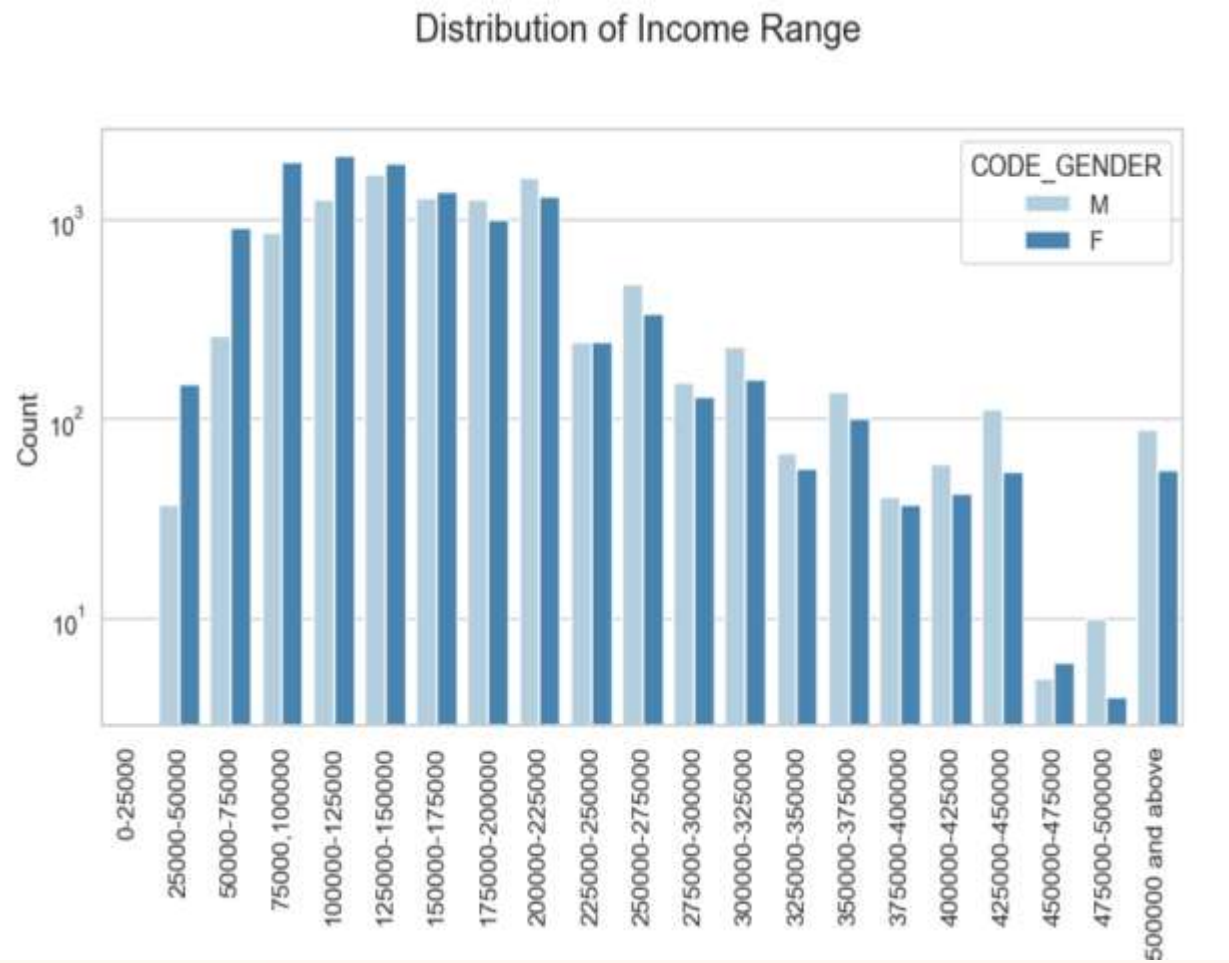
Analysis

Target



- Here, 'Target = 0' means the people those who are non-defaulters
- Here , 'Target = 1' means the people those who are defaulters

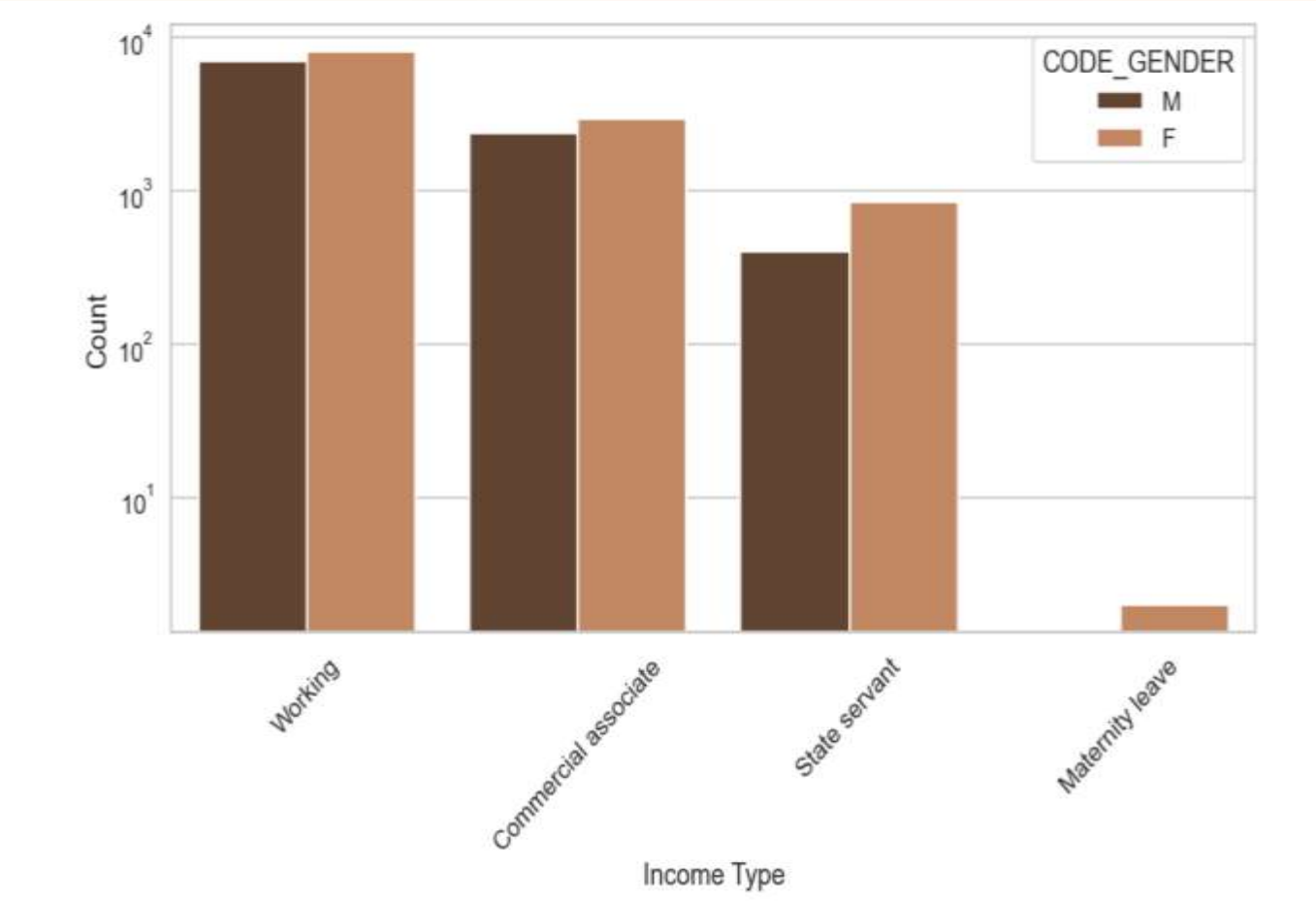
Distribution of Income Range



Conclusion from the graph

1.Male counts are higher than female.
2.Income range from 100000 to 200000 is having more number of credits.
3.This graph show that males are more than female in having credits for that range.
4.Very less count for income range 400000 and above.

Distribution of Income Type



Conclusion from the graph

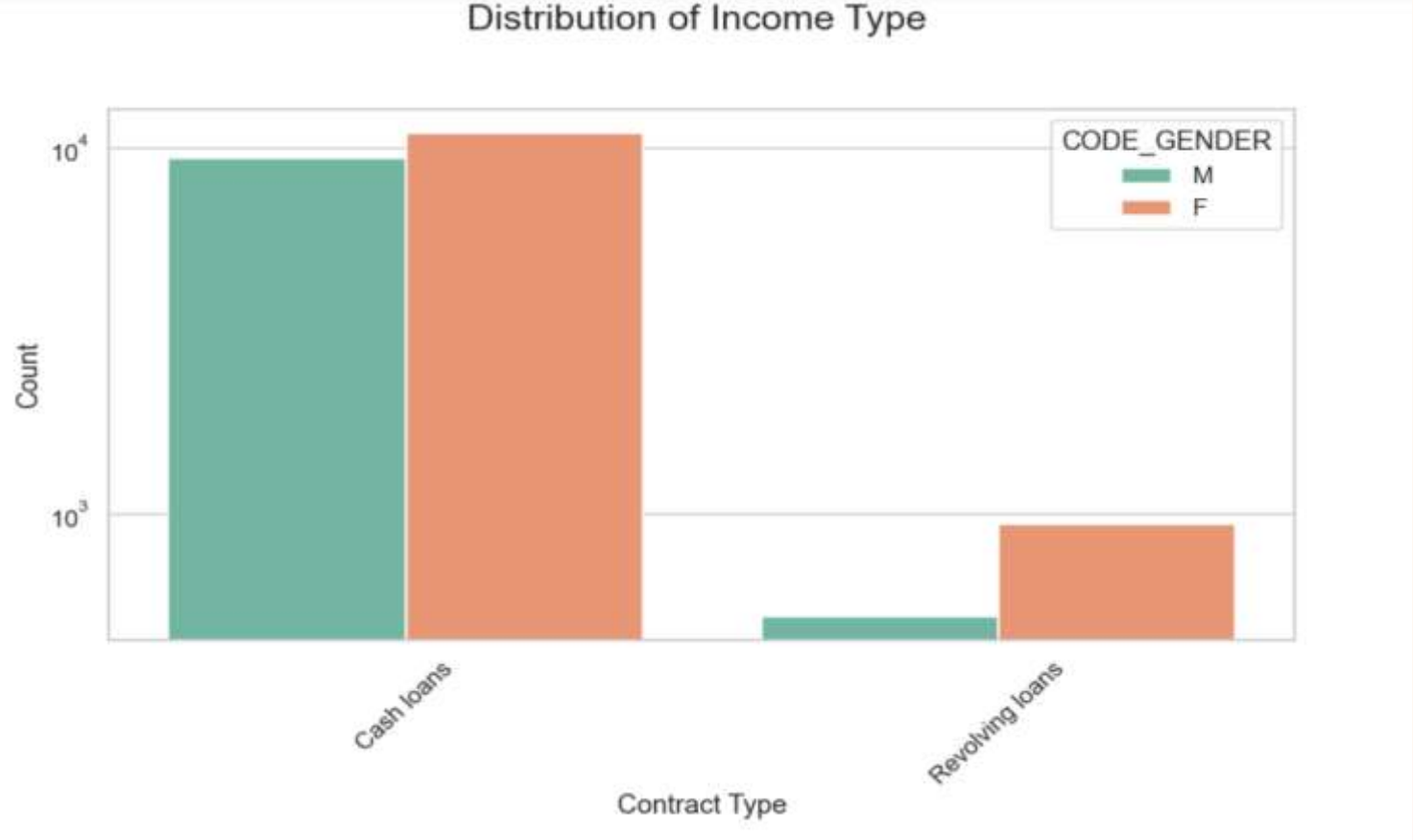
1.For income type ‘working’, ’commercial associate’, and ‘State Servant’ the number of credits are higher than other i.e. ‘Maternity leave.

2.For this Females are having more number of credits than male.

3.Less number of credits for income type ‘Maternity leave’.

4.For type 1: There is no income type for ‘student’ , ’pensioner’ and ‘Businessman’ which means they don’t do any late payments.

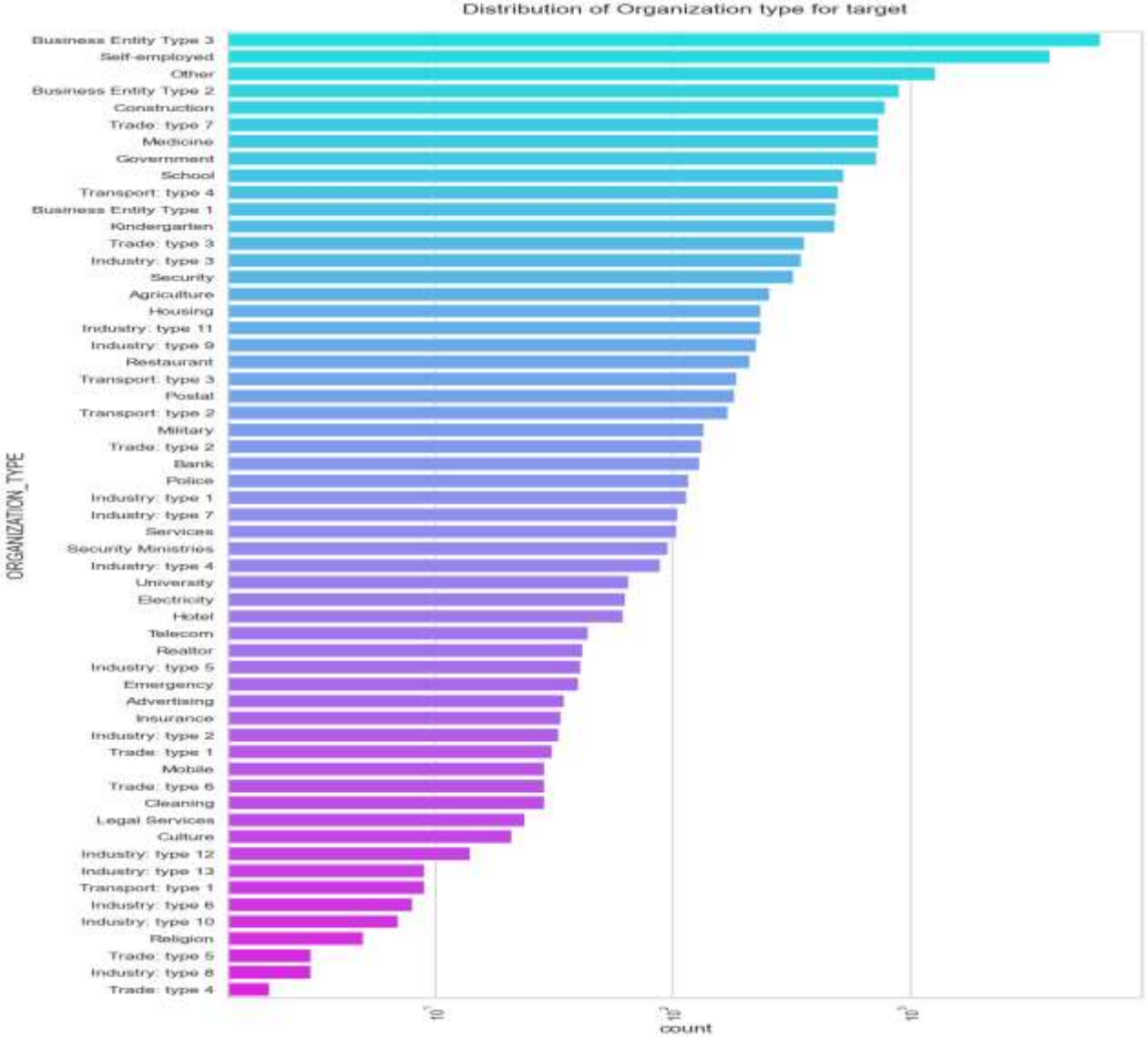
Distribution of Income Type



Conclusion from the graph

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

Distribution of Organization Type for Target



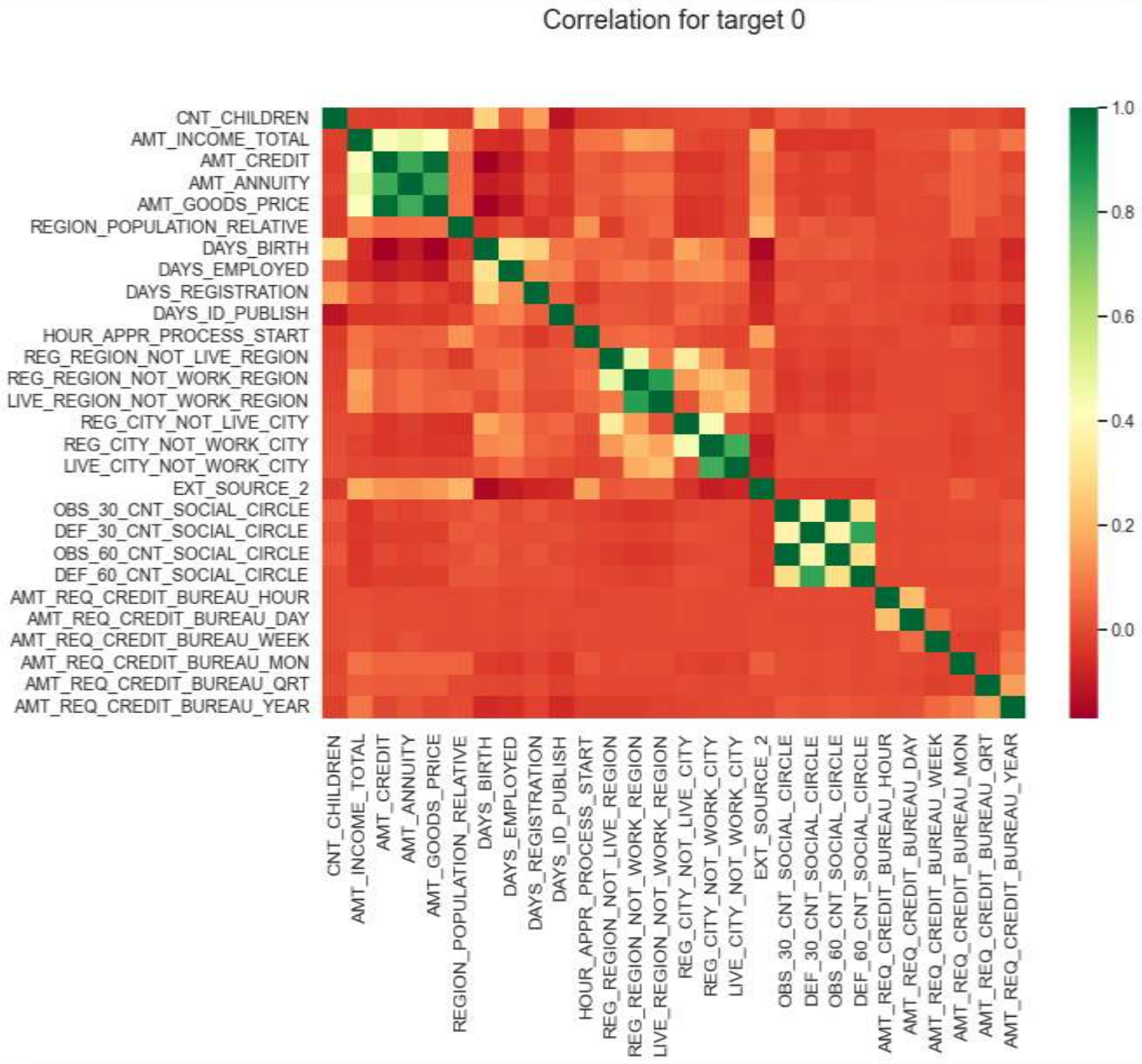
Conclusion from the graph

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

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

3.Same as type 0 in distribution of organization type.

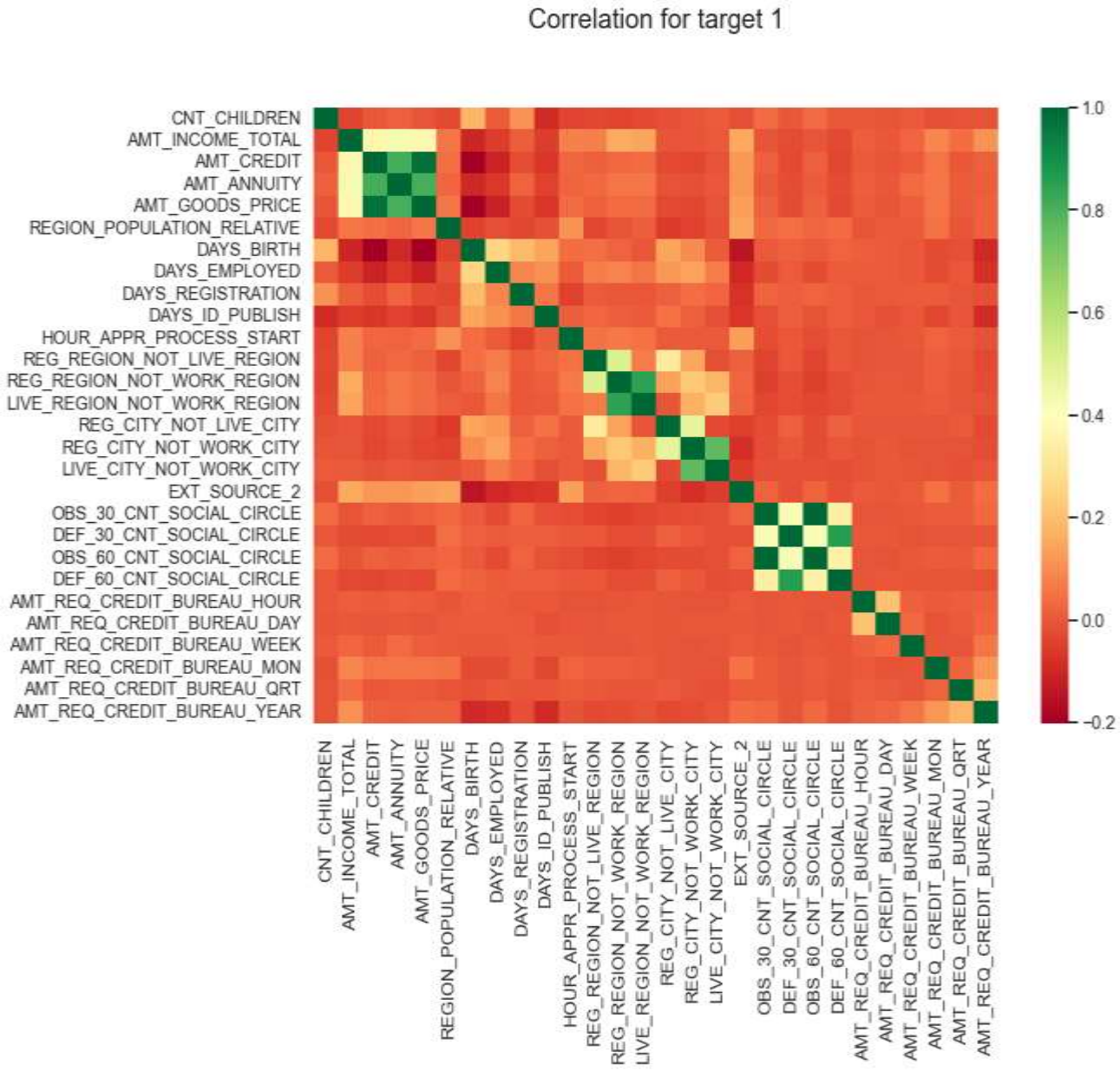
Correlation for Target 0



Conclusion from the graph

1.Credit amount is inversely proportional to the date of birth, which means Credit amount is higher for low age and vice-versa.
2.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.
3.Income amount is inversely proportional to the number of children client have, means more income for less children client have and vice-versa.
4.less children client have in densely populated area.
5.Credit amount is higher to densely populated area.
6.The income is also higher in densely populated area.

Correlation for Target 1

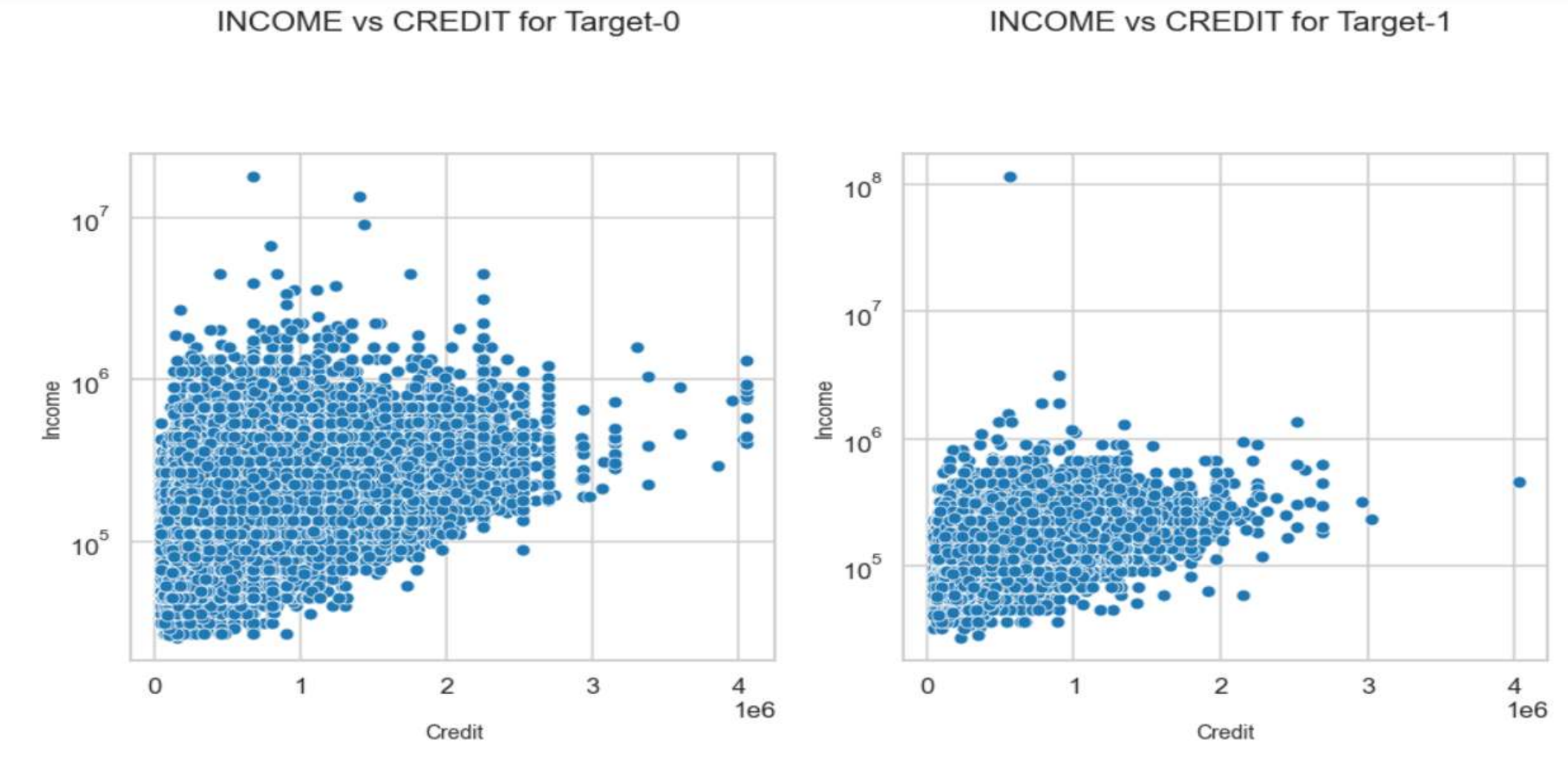


Conclusion from the graph

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

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

Bivariate Analysis of the numrical columns



Bivariate Analysis of the numrical columns

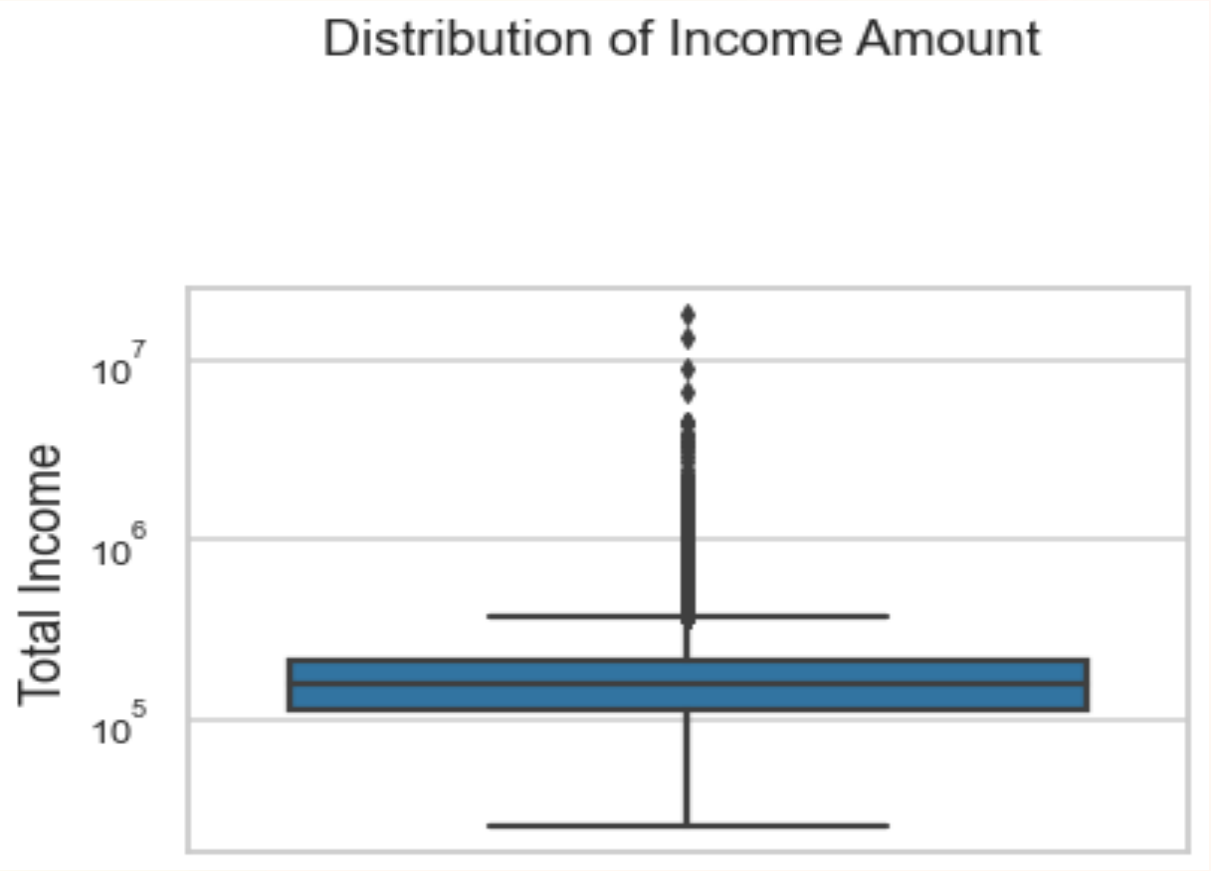
Conclusion from the graph

With the scatter plot, we can determine that AMT CREDIT and AMT GOODS PRICE are highly correlated, which means if increase in goods price the credit increased directly and vice versa

Univariate Analysis

Outliers

For Target-0



Conclusion from the graph

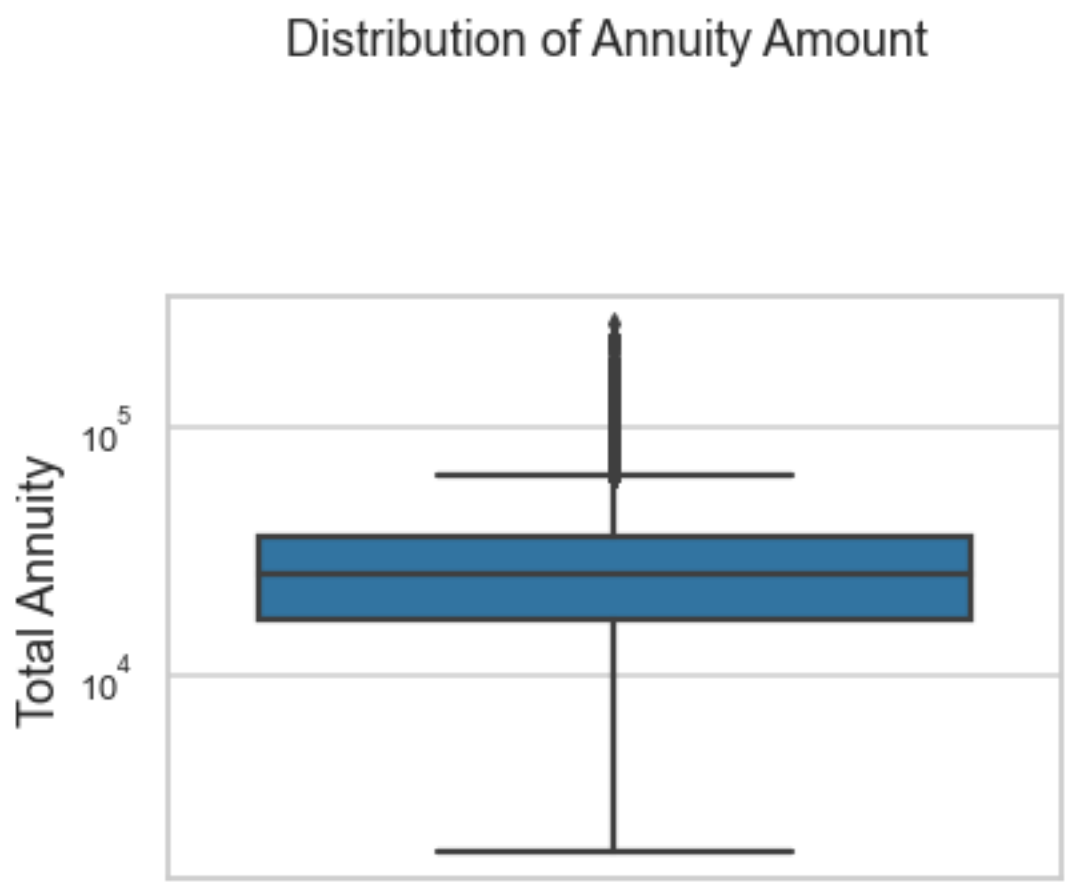
1.There seems to be an equal distribution of the income amount of the clients.
2.Also, some of the outliers present in the dataset.

For Target-0

Conclusion from the graph

1. The first quartile is bigger than the third quartile, that means most of the client credit lies in the first quartile.
2. There Seems some outliers in the Credit boxplot.

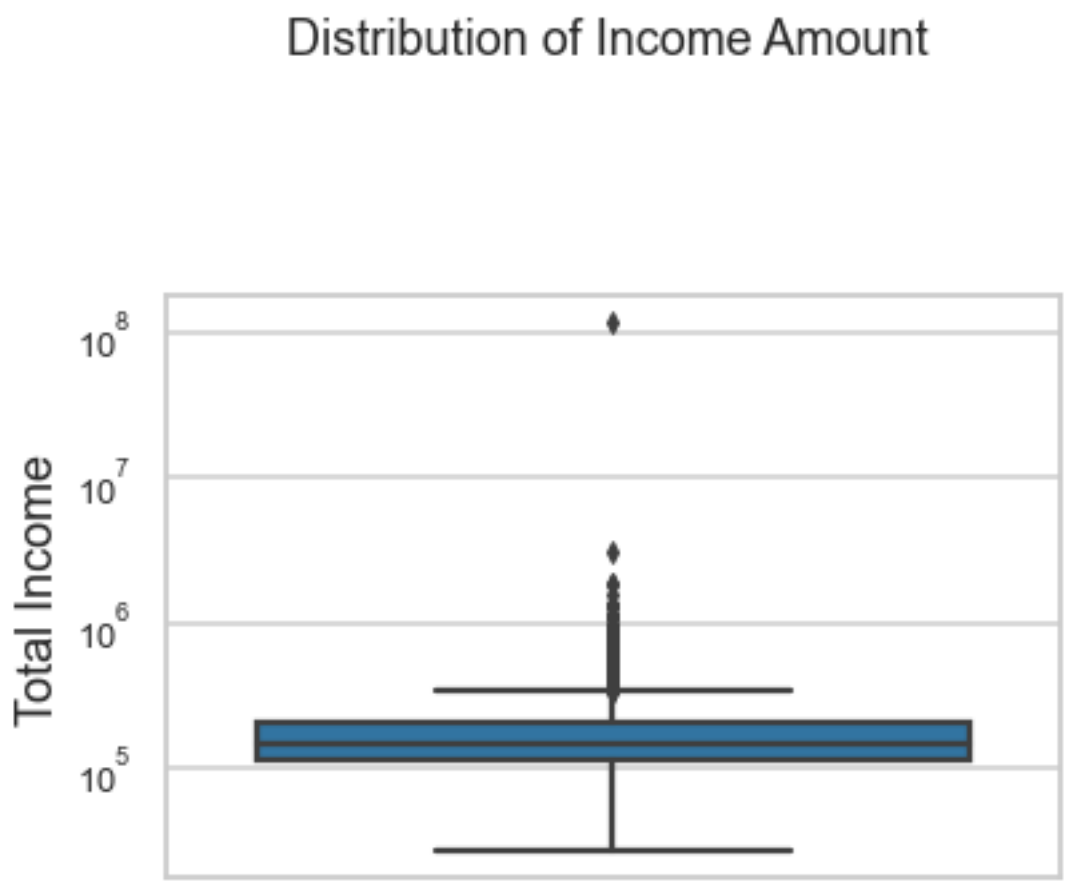
For Target-0



Conclusion from the graph

1.The first quartile is bigger than the third quartile.
2.There seems some outliers in the Annuity boxplot.

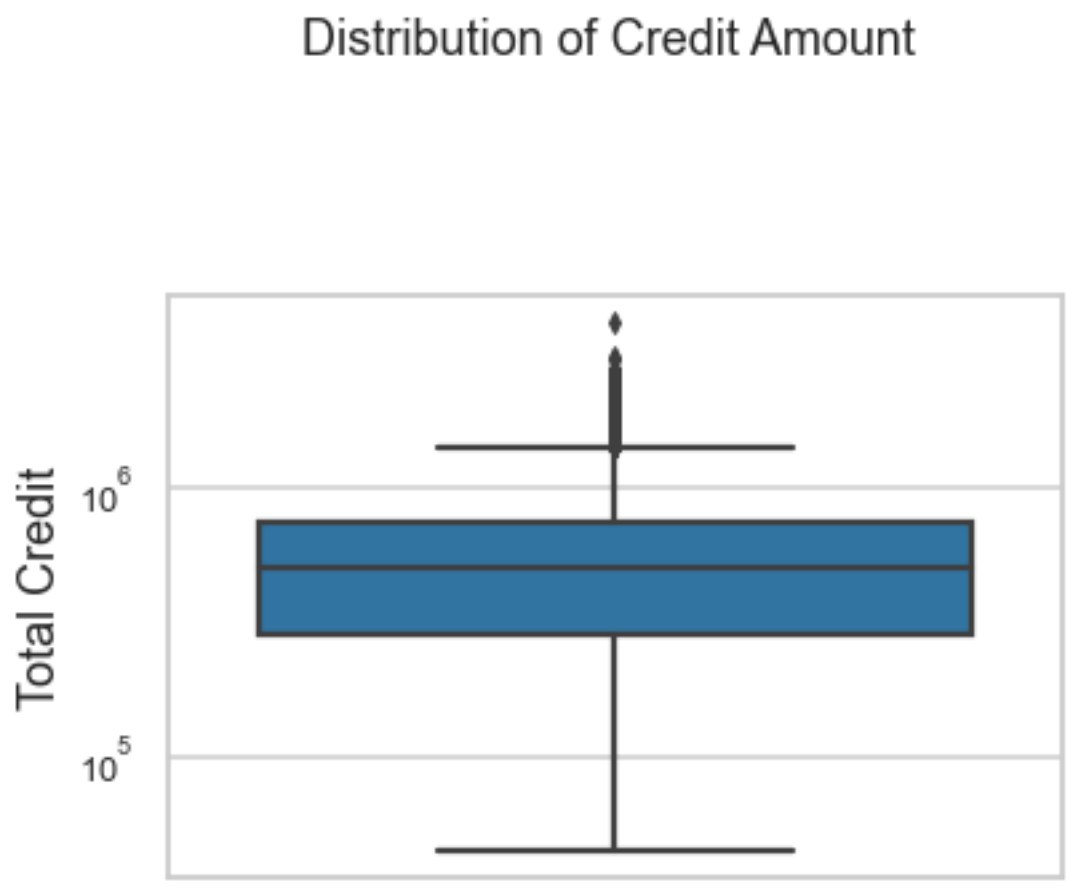
For Target-1



Conclusion from the graph

1.There seems a significant outlier in the Income dataset.
2.Most of the income of the client lies in the third quartile.

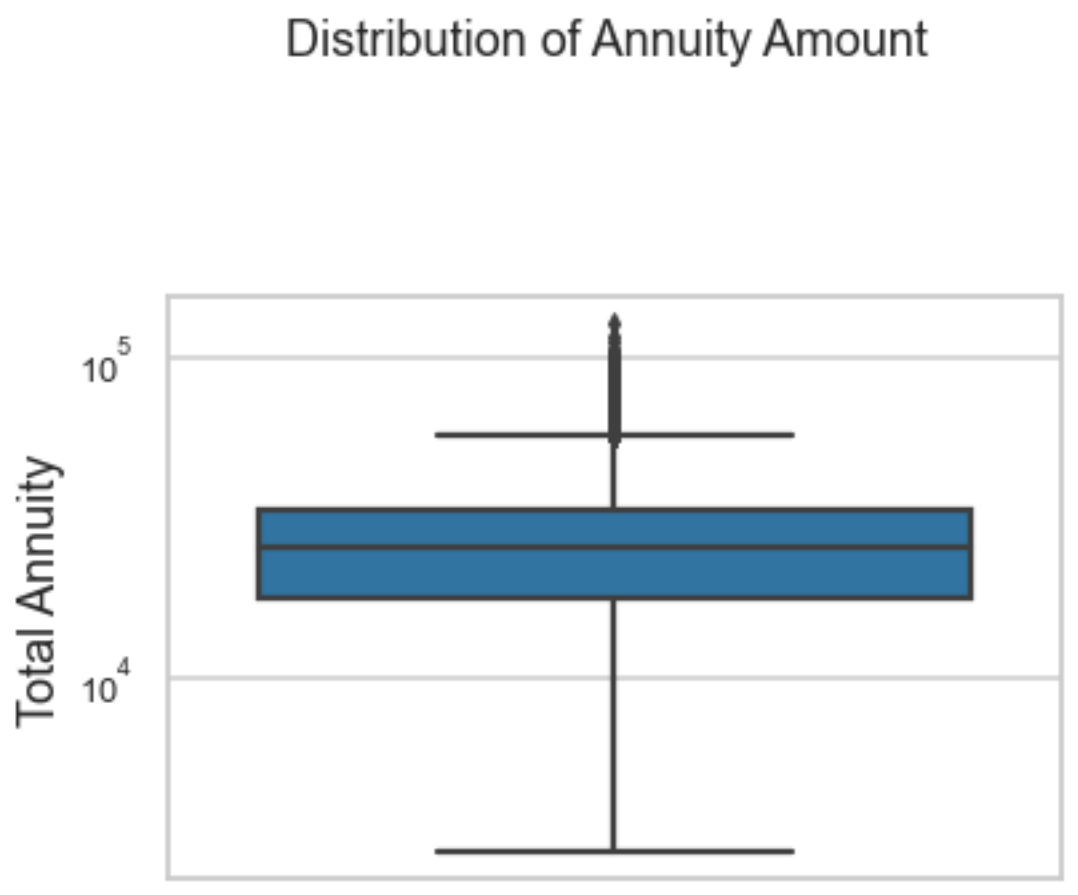
For Target-1



Conclusion from the graph

1.The first quartile is bigger than the third quartile, that means most of the client credit lies in the first quartile.
2.There seems some outliers in the credit boxplot.

For Target-1



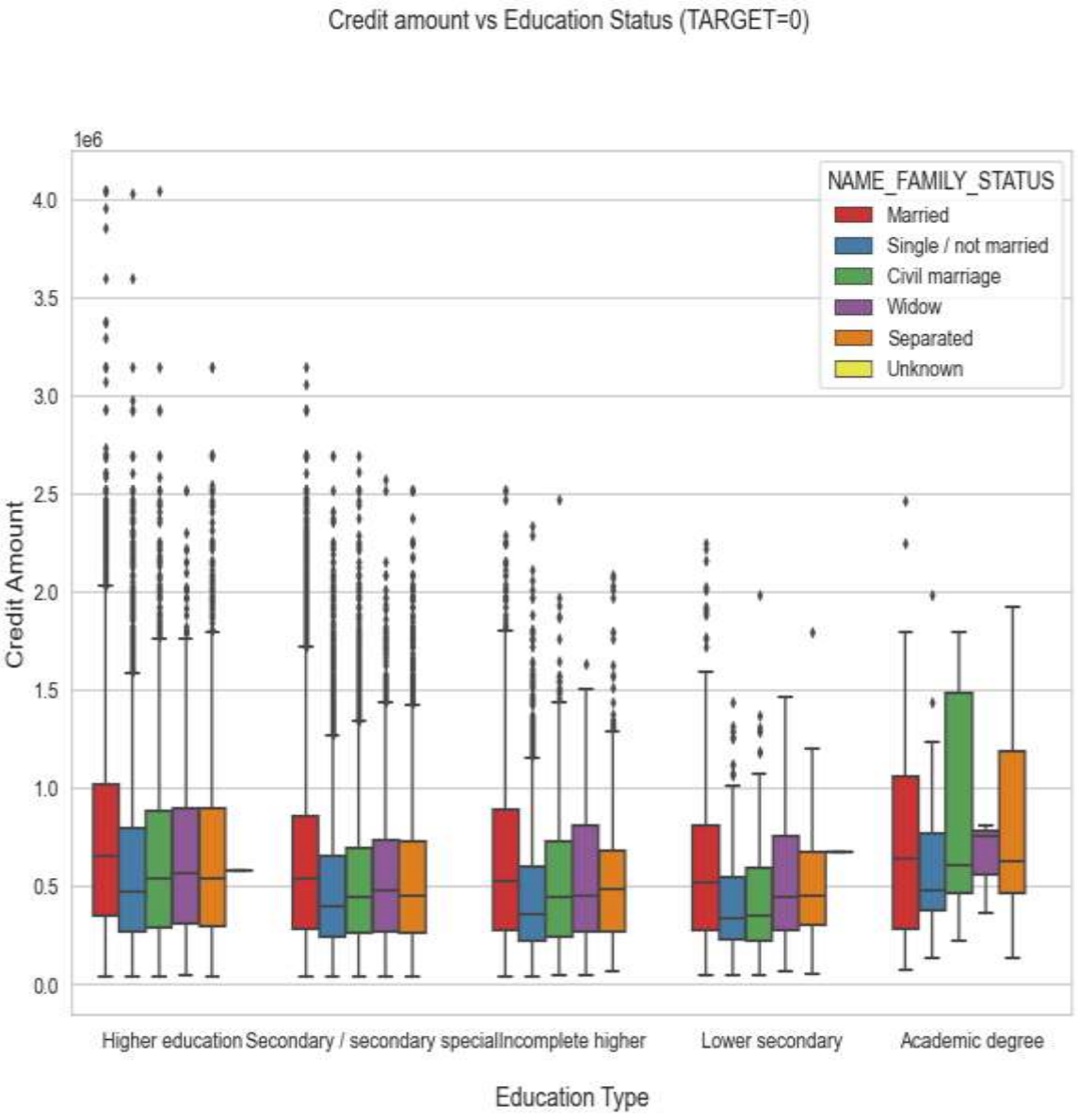
Conclusion from the graph

1. The first quartile is bigger than the third quartile.
2. There seems some outliers in the Annuity boxplot

Multivariate Analysis

Outliers

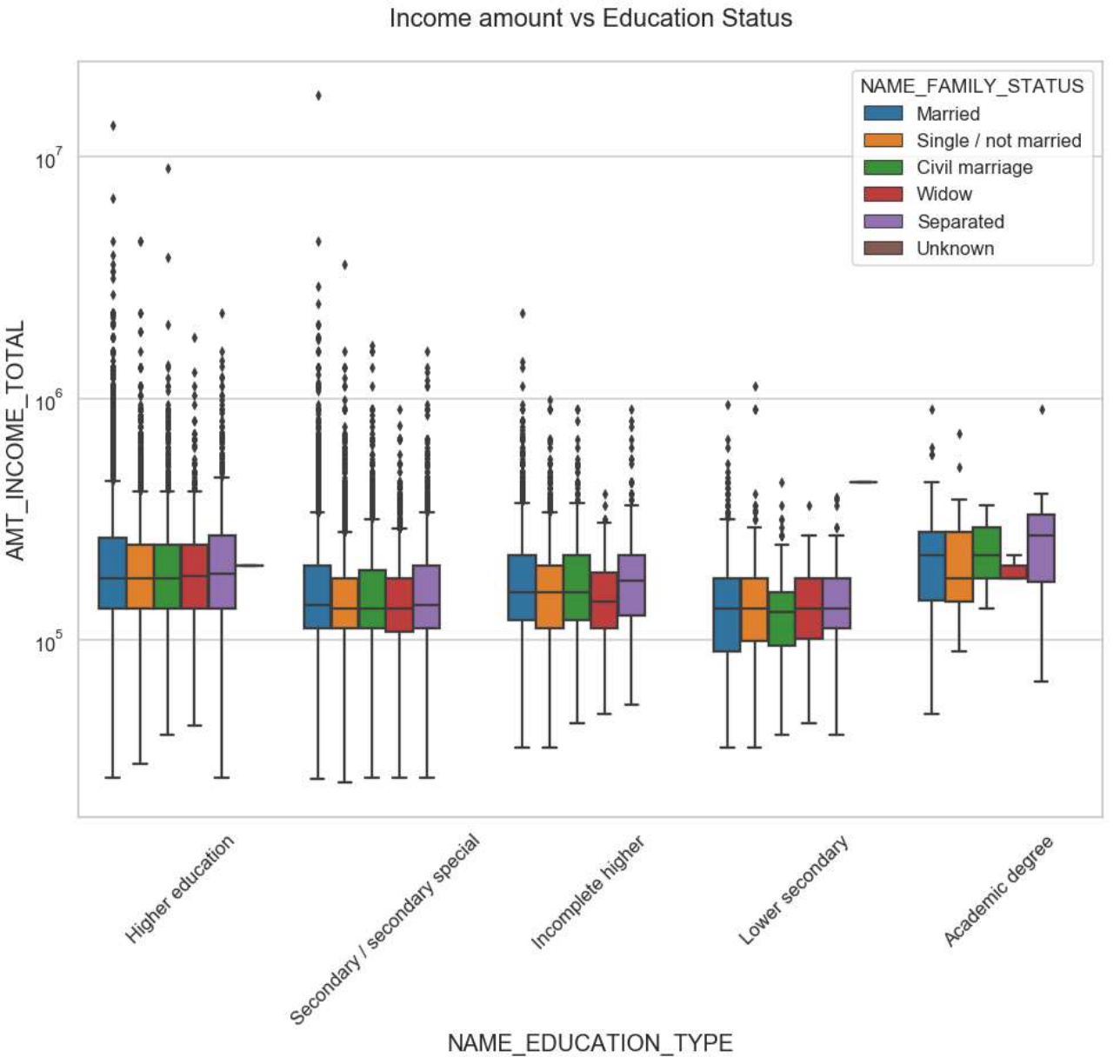
For Target-0



Conclusion from the graph

From the above box plot we can conclude that Family status of 'civil marriage', 'marriage' and 'separated of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'marriage', 'single' and ‘civil marriage’ are having more outers Civil marriage for Academic degree is having most of the credits in the third quartile.

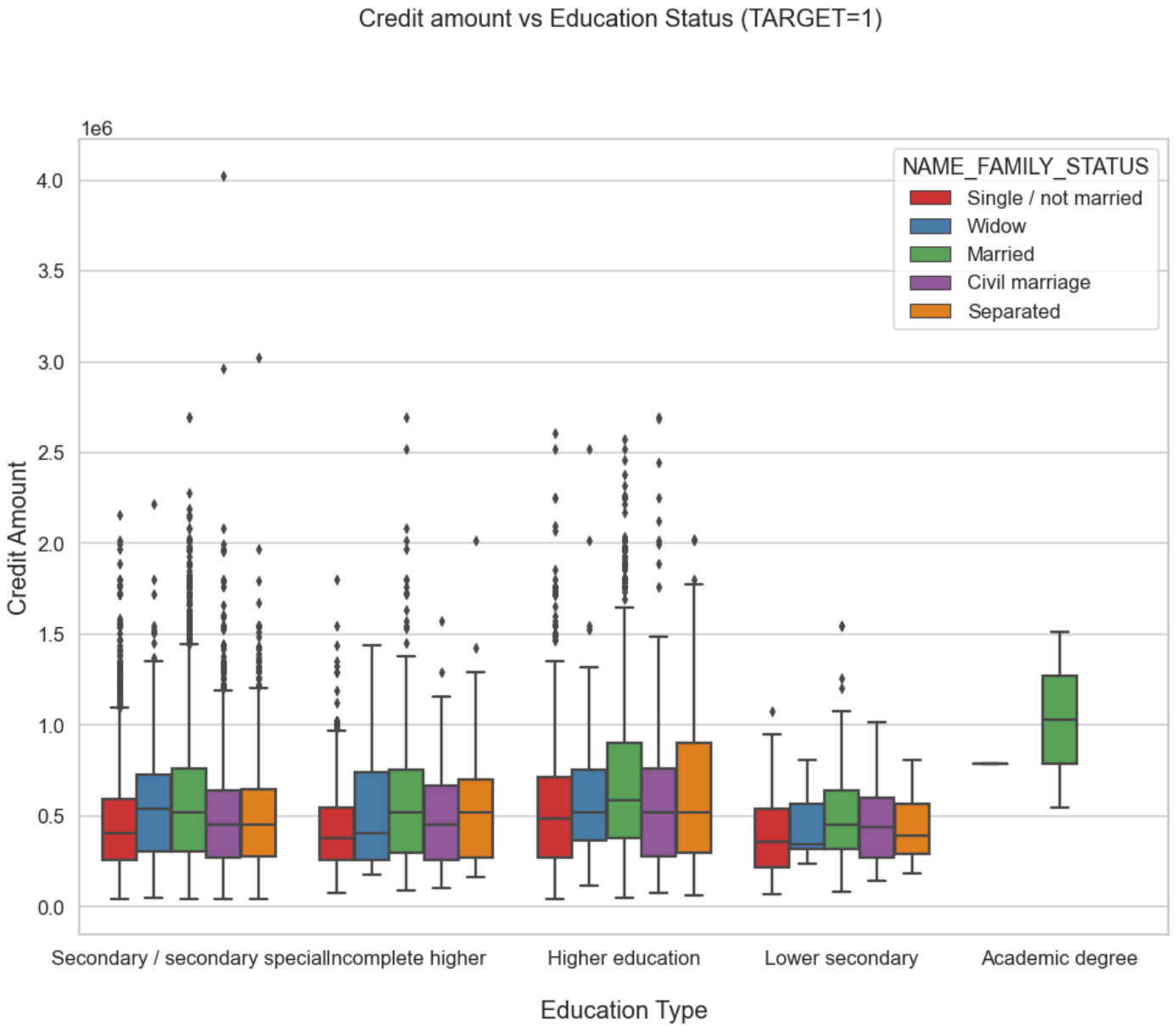
For Target-0



Conclusion from the graph

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

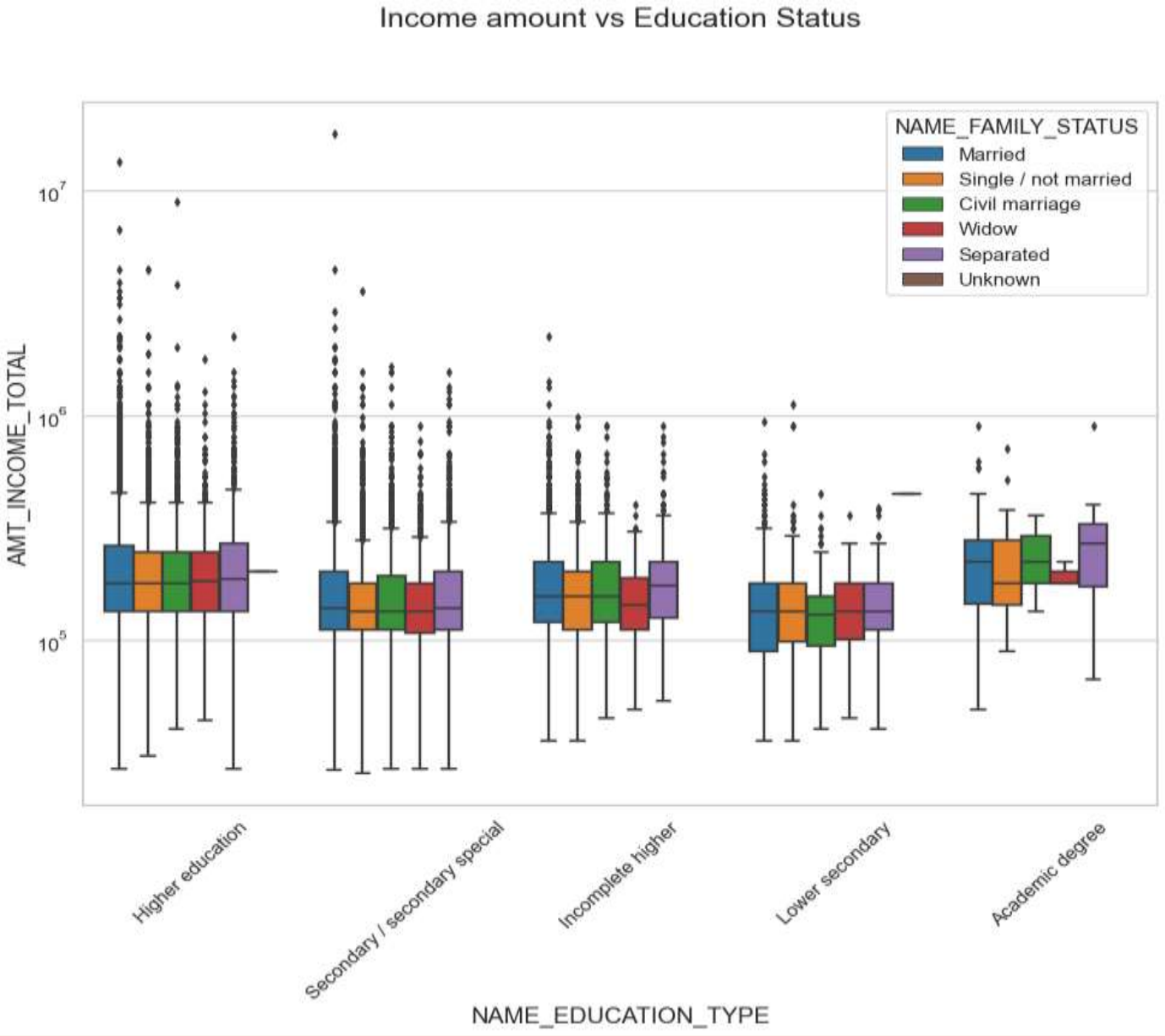
For Target-1



Conclusion from the graph

Quite similar with Target 0 From the above box plot 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.

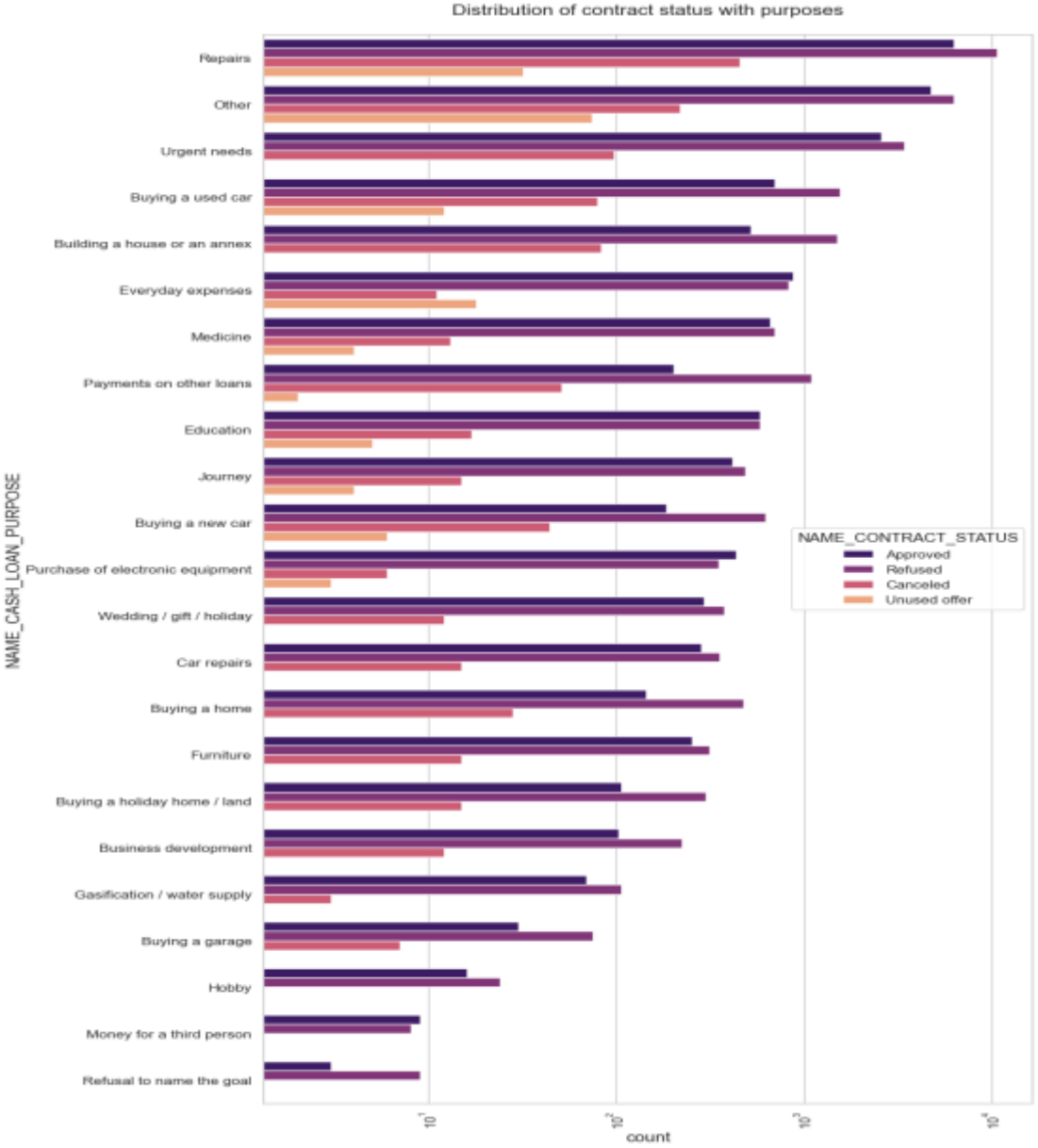
For Target-1



Conclusion from the graph

Have some similarity with Target0, 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.

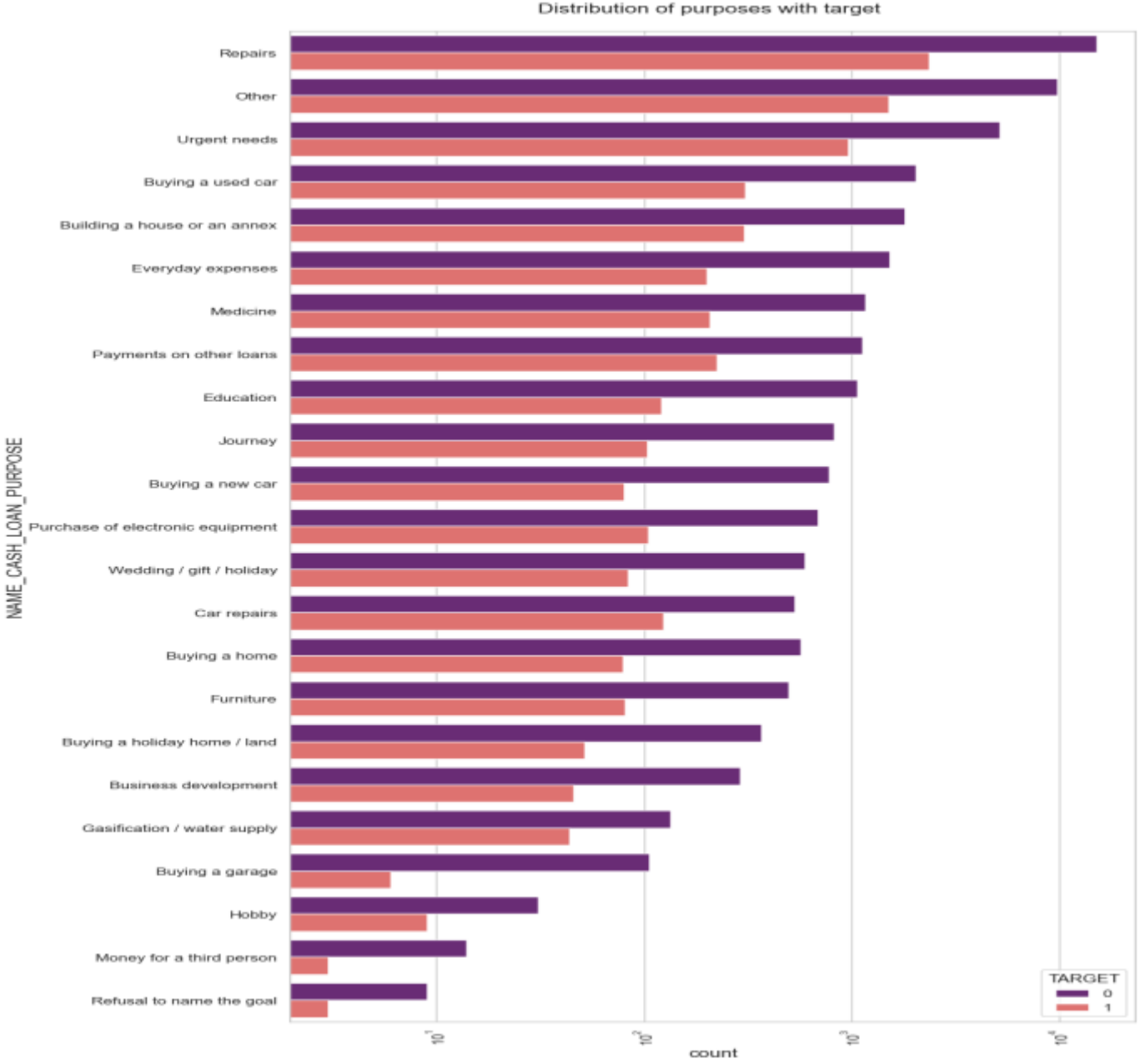
Univariate analysis



Conclusion from the graph

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

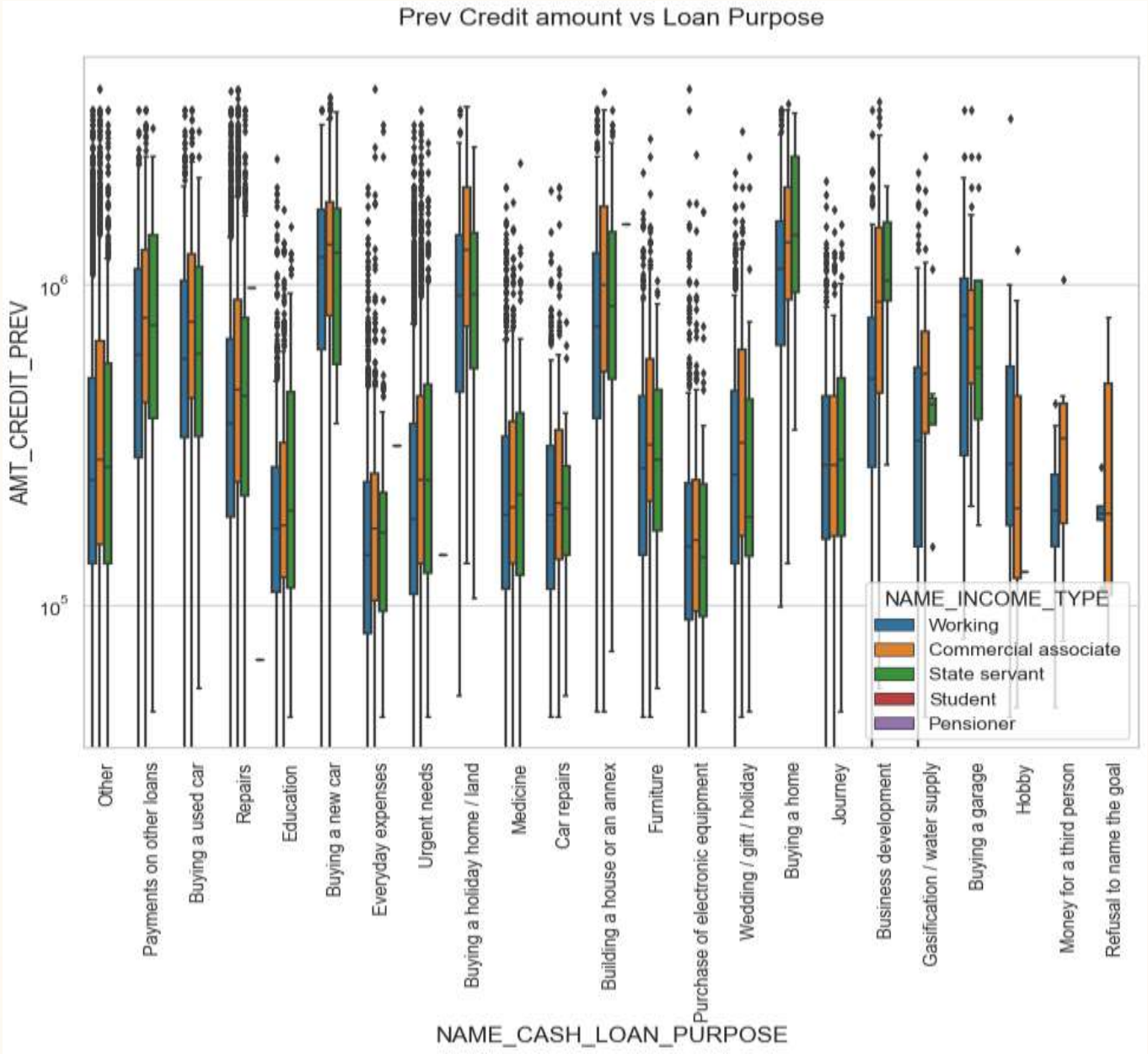
Univariate analysis



Conclusion from the graph

1.Loan purposes with 'Repairs' are facing more difficulties in payment on time.
2.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'
3.Hence we can focus on these purposes for which the client is having for minimal payment difficulties.

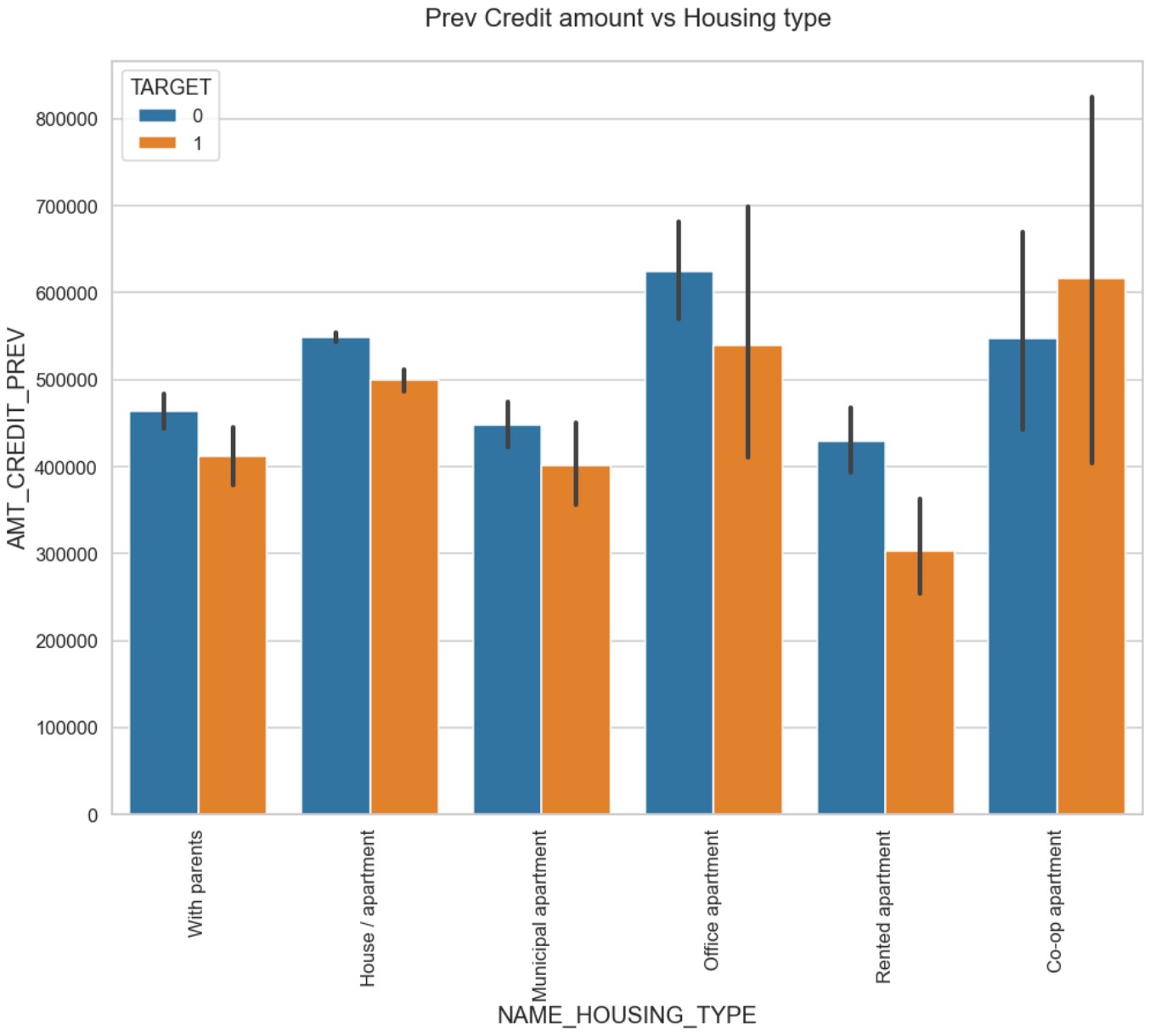
Bivariate analysis



Conclusion from the graph

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

Bivariate analysis



Conclusion 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

1. Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.

2. Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.

3. Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.

4. Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.

5. Working people especially female employees are the best target for the loans.