Credit EDA Case Study

Submitted By:

Shreya Aron

BUSINESS OBJECTIVE

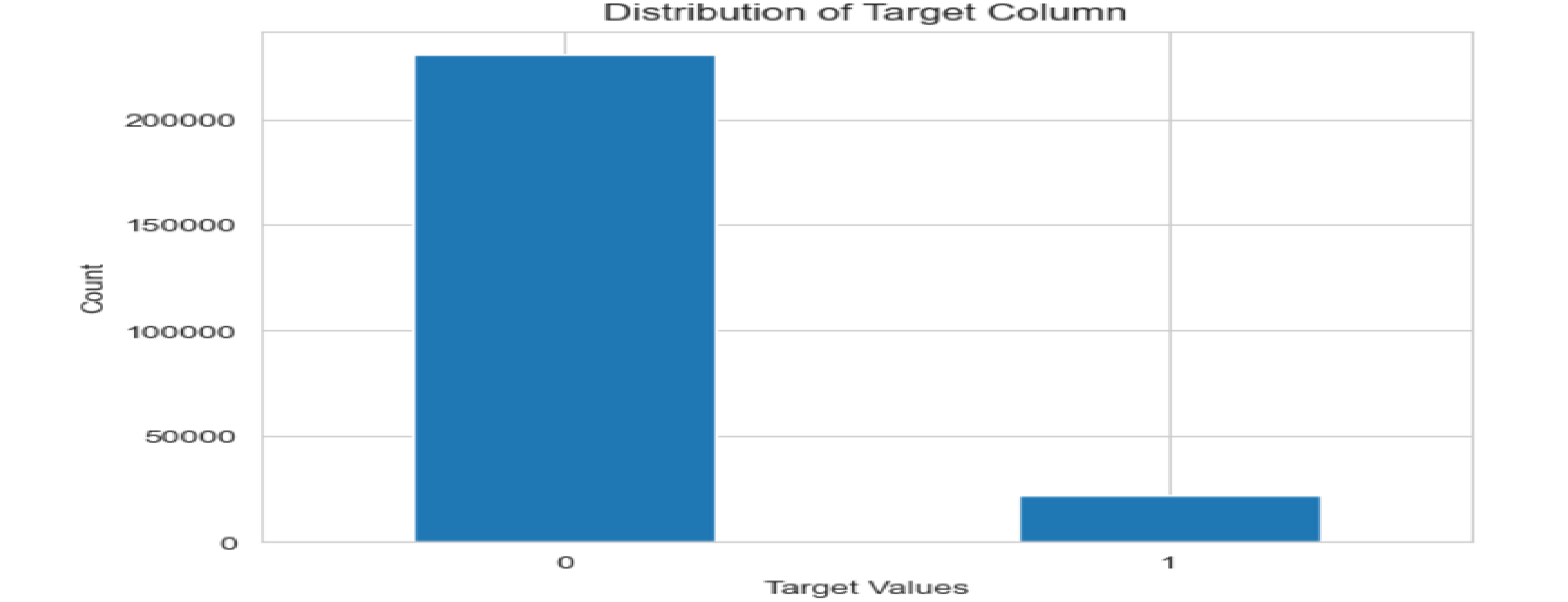
- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers 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 REQUIRED FOR ANALYSIS

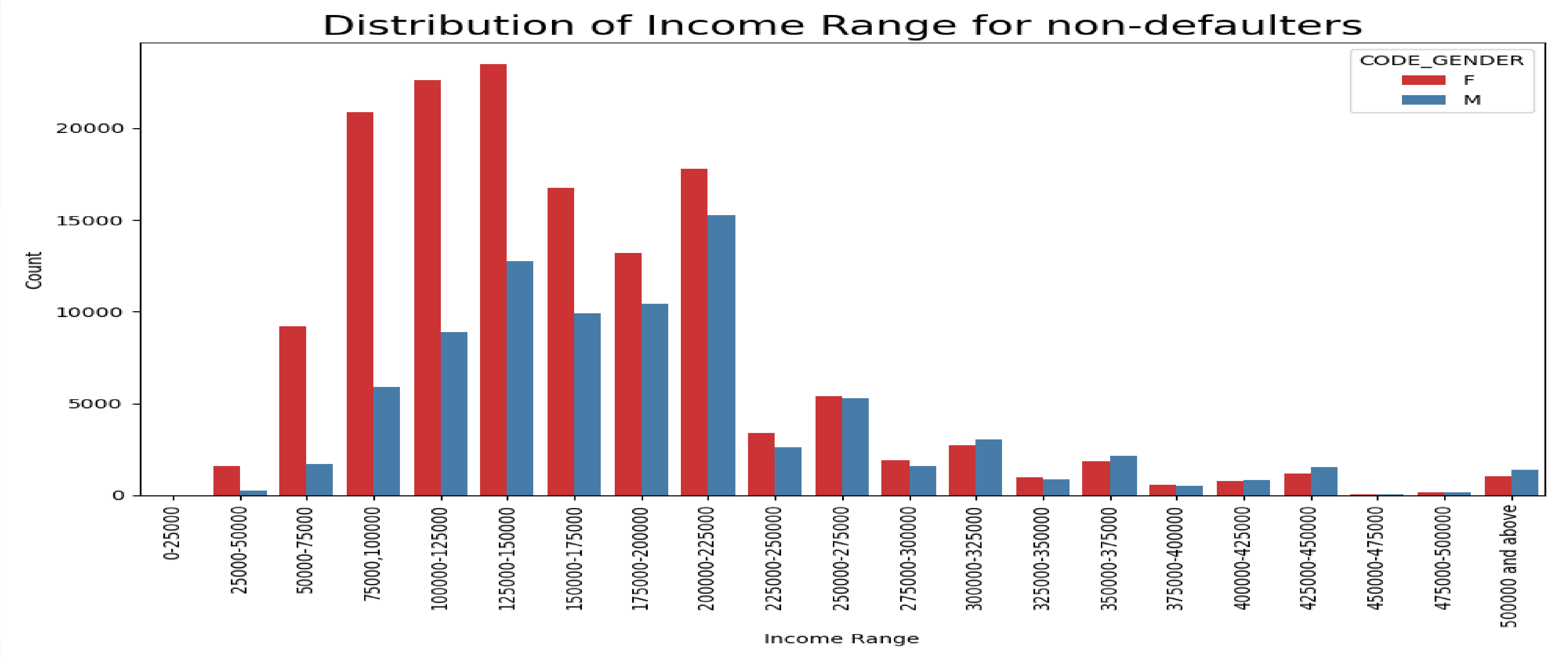
- Firstly, **understand the data** and **check the missing value** percentage of data. If missing values > 30 then drop the columns because considering them would result in incorrect analysis.
- **Impute the outliers** and **missing values** with mean, median and mode values in numerical and categorical columns.
- **Fix the data types**.
- **Analyse the individual datasets** by performing univariate analysis, bivariate and multivariate analysis.
- **Merge the datasets** and then perform combined analysis.

IMBALANCE RATIO



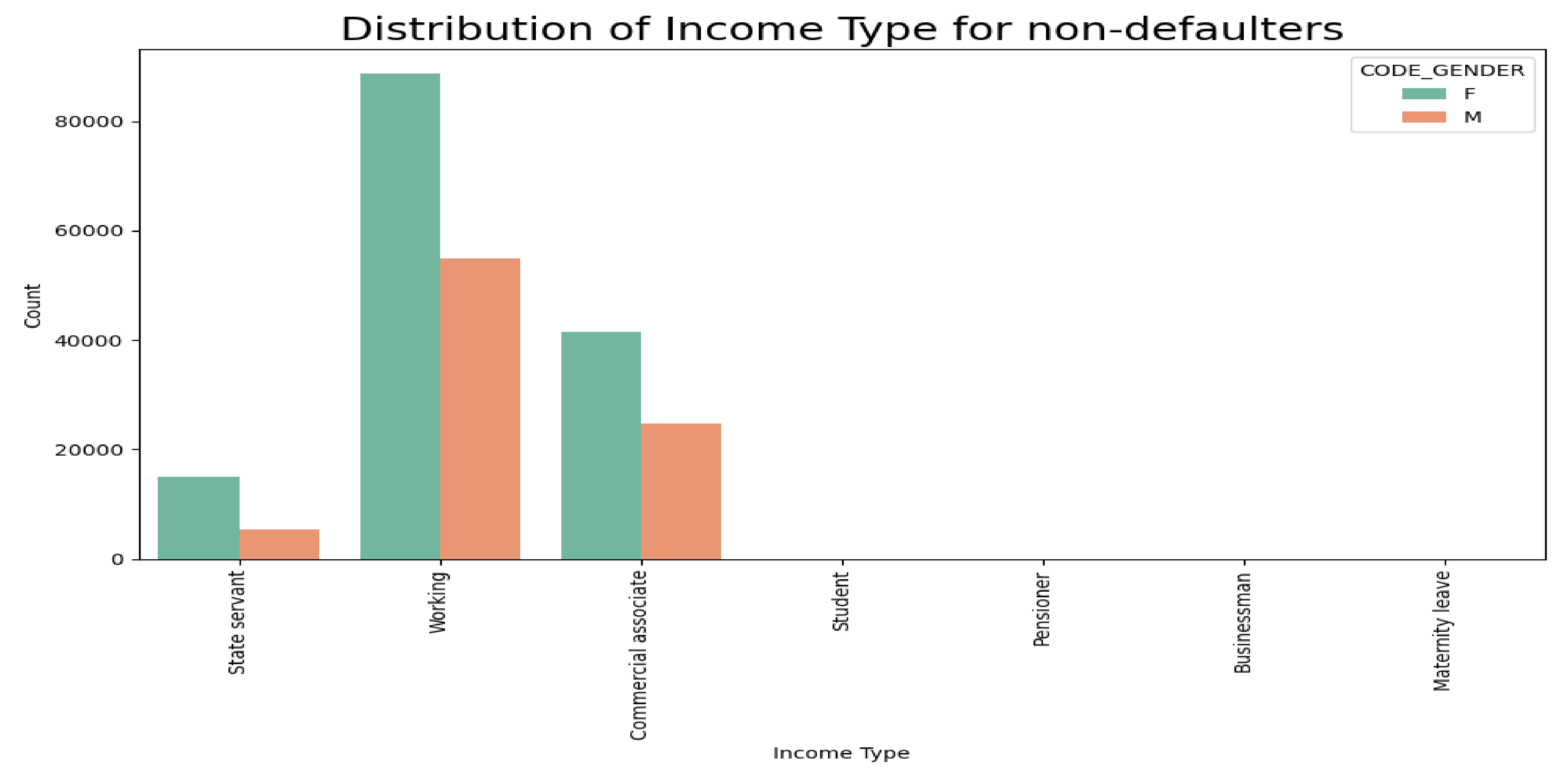
- The data after cleanup is highly imbalanced. The total count for non-defaulters i.e. target_0 is 230302 and for defaulters i.e. target_1 is 21835.
- Therefore, for every 1 instance of target_1(defaulter) there are approximately 10.55 instances of target_0(non-defaulters).

Categorical Univariate Analysis for Target 0



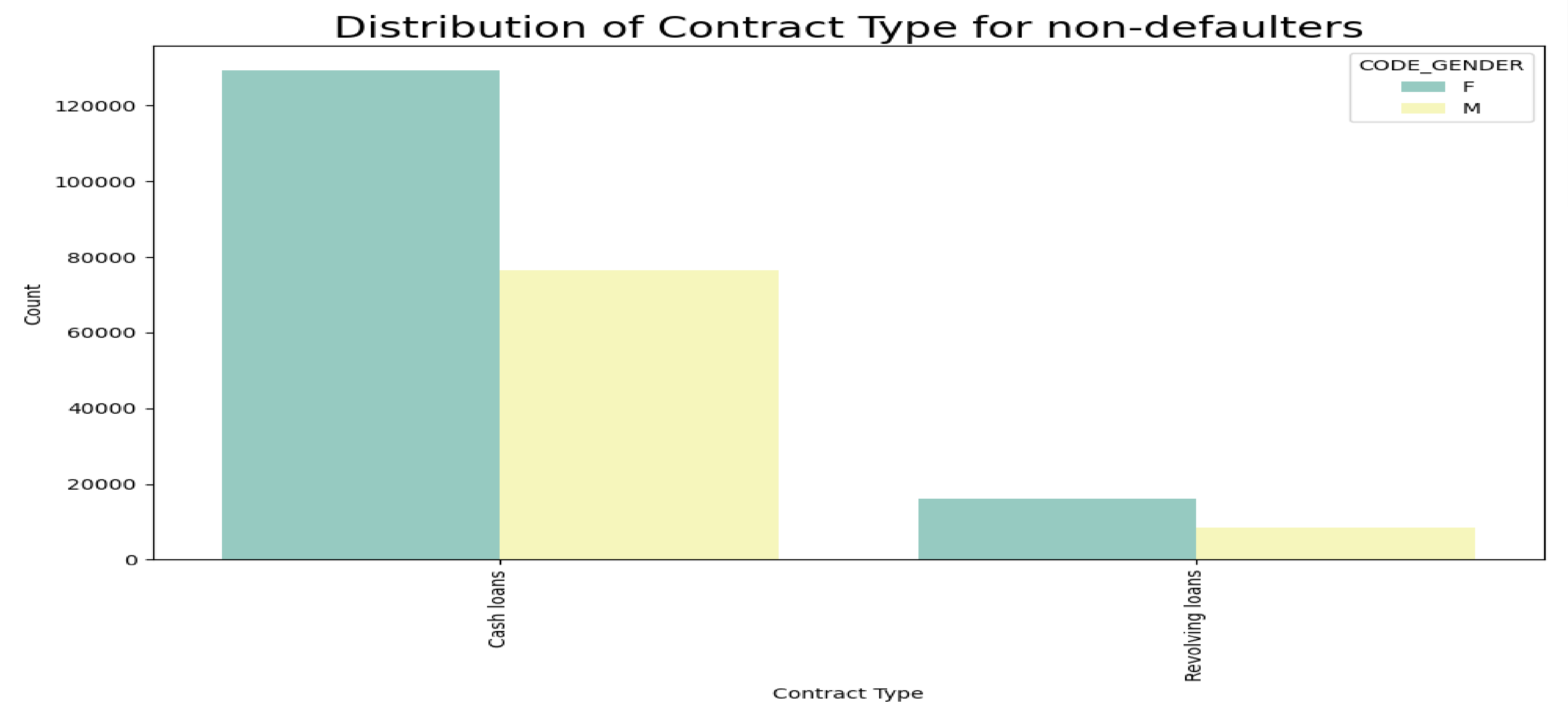
In the graph we can see that,

- For most of the income ranges, female non-defaulters are more.
- Income range from 125000-150000 is having maximum credits.
- Least count is for range 450000-475000.



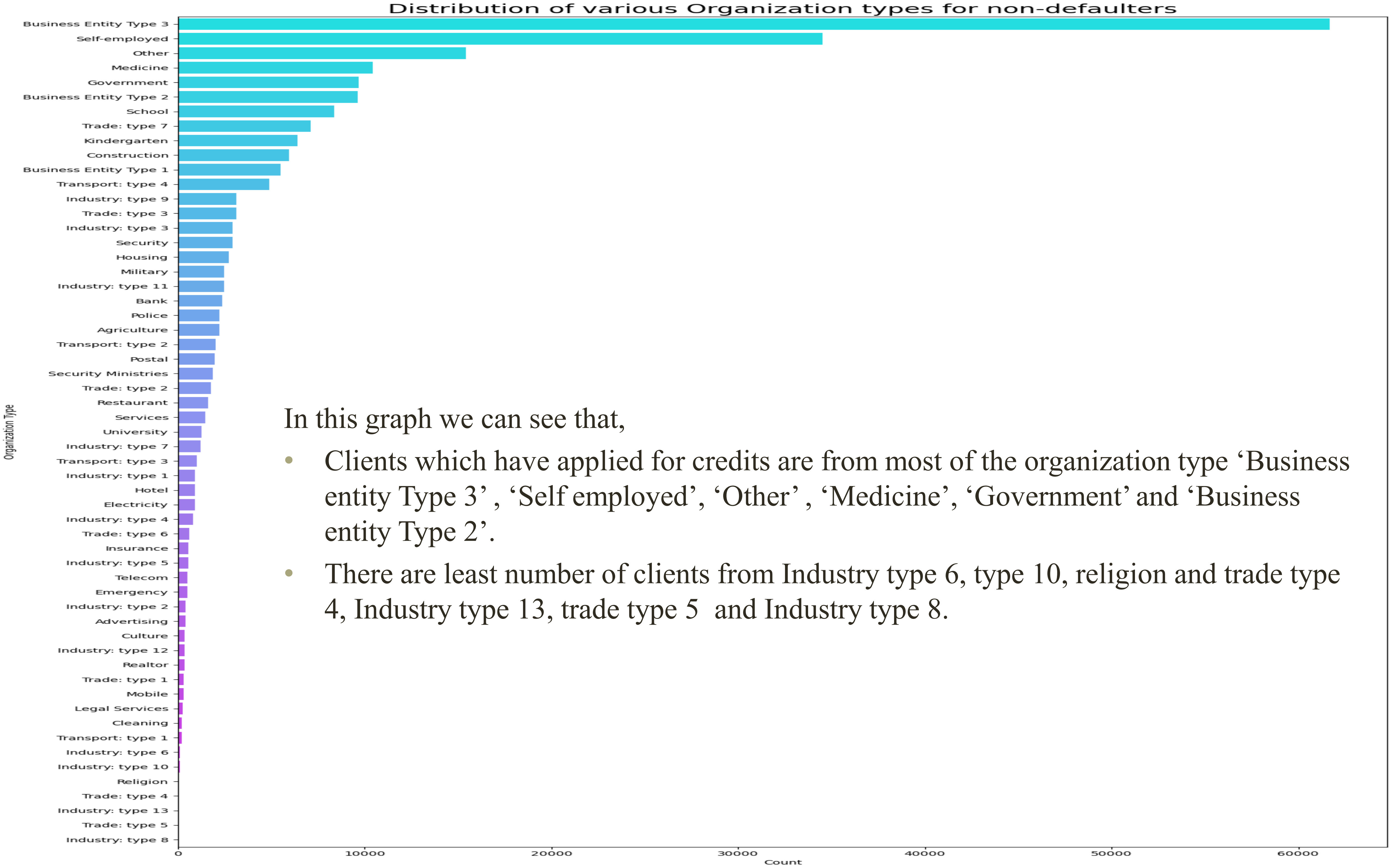
In the graph we can see that,

- Working women have maximum credits than others.
- ‘State Servant’, 'Working' and 'Commercial Associate‘ have more credit counts compared to others.



In the graph we can see that,

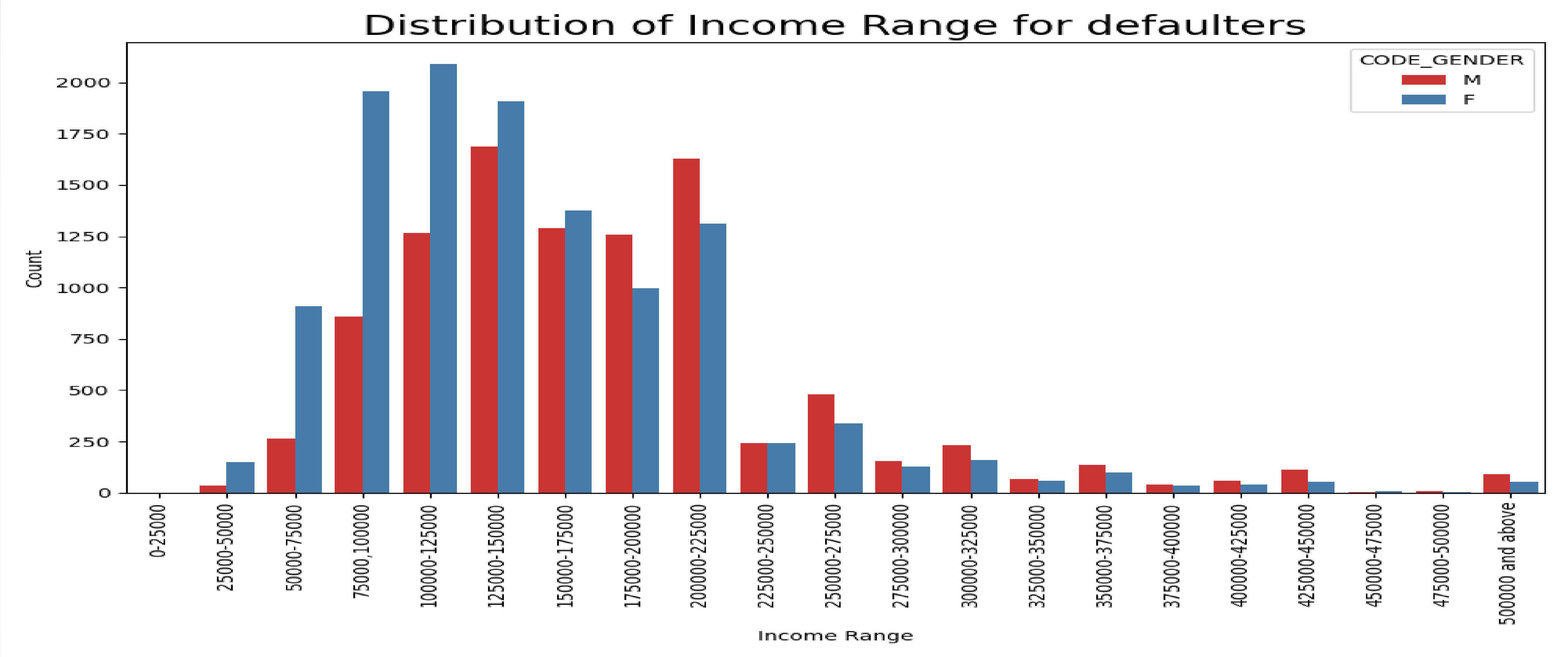
- It seems that cash loans is having higher number of credits than ‘Revolving loans’ contract type.
- Also, female applies more for Credit.



In this graph we can see that,

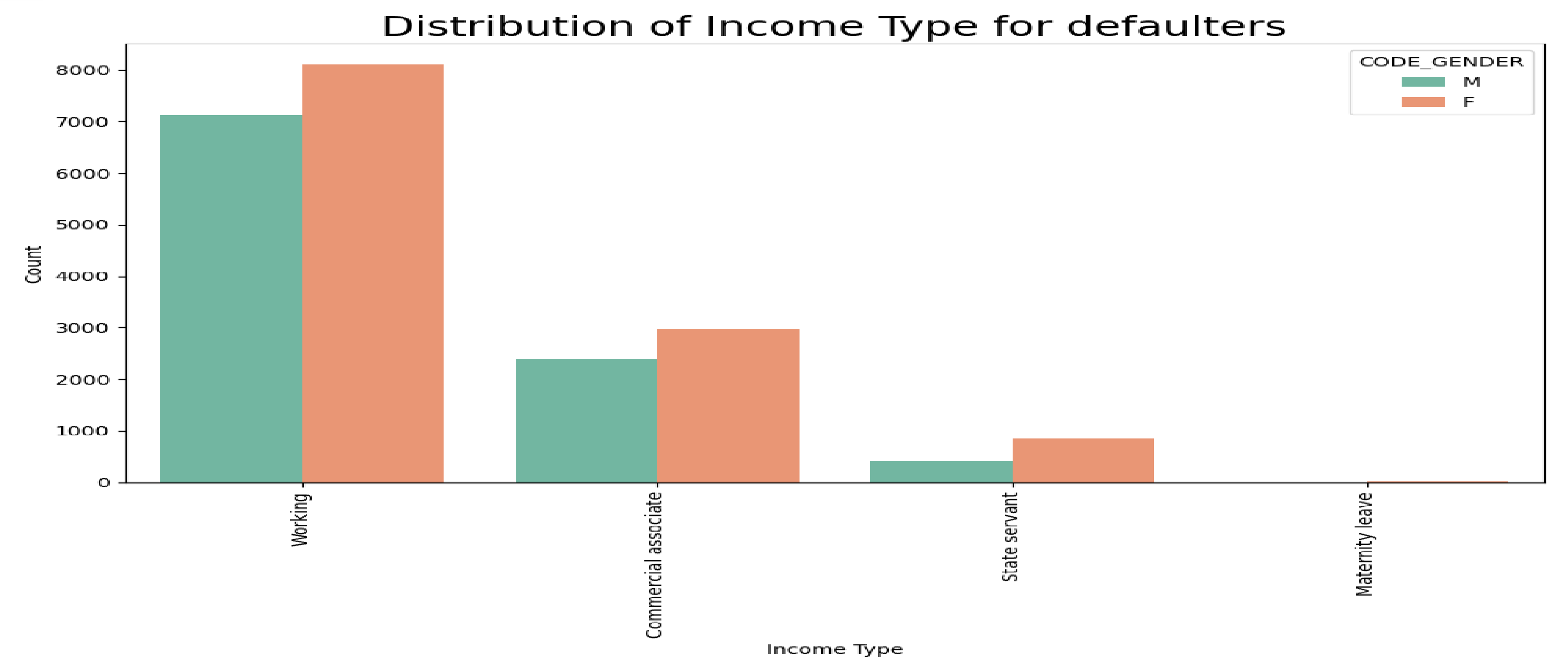
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’, ‘Other’ , ‘Medicine’, ‘Government’ and ‘Business entity Type 2’.
- There are least number of clients from Industry type 6, type 10, religion and trade type 4, Industry type 13, trade type 5 and Industry type 8.

Categorical Univariate Analysis for Target 1



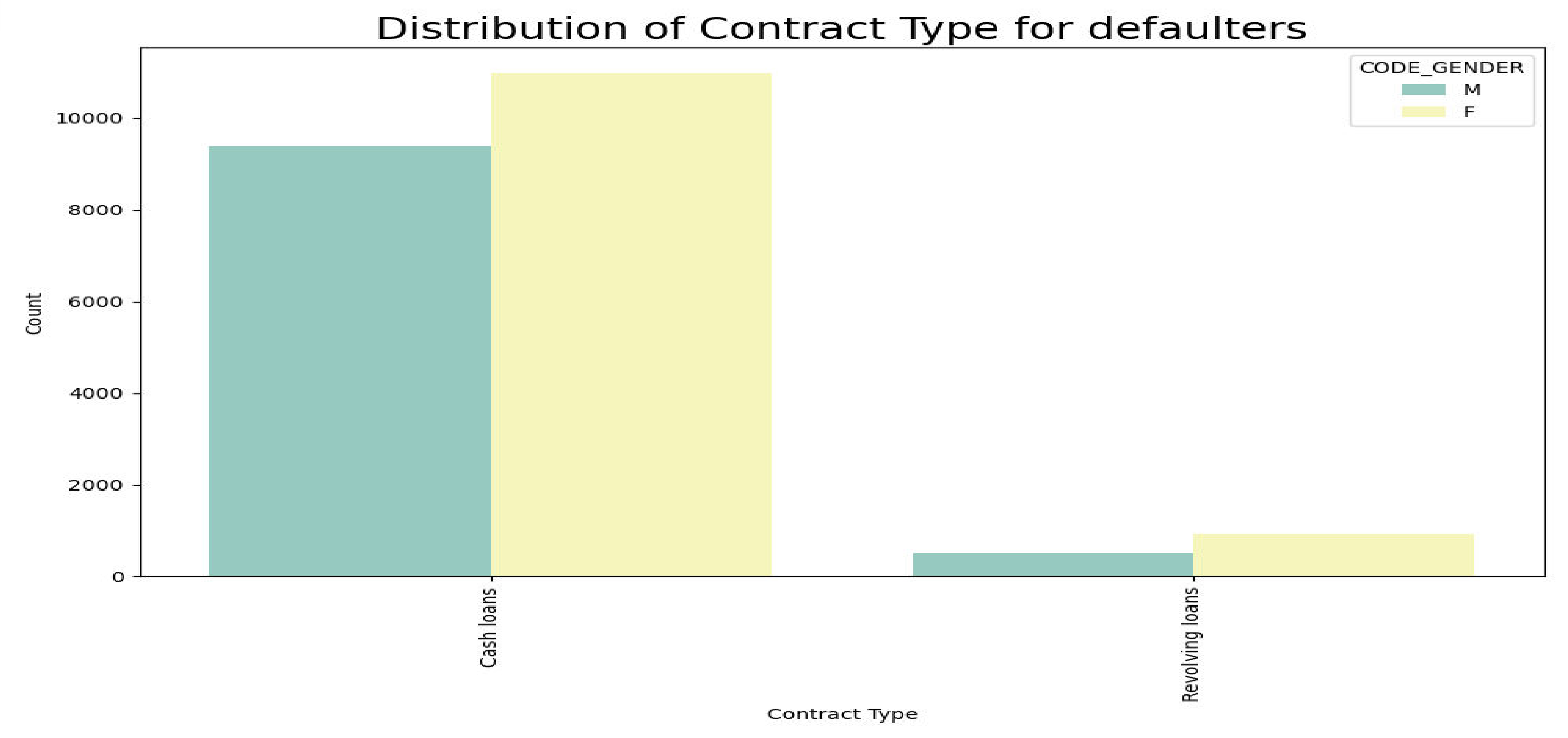
In this graph we can see that,

- Male counts are higher.
- Income range from 100000 to 200000 is having more number of credits.
- For range 225000-250000 both male and female have same credits.
- There is least count for income range 450000-500000.



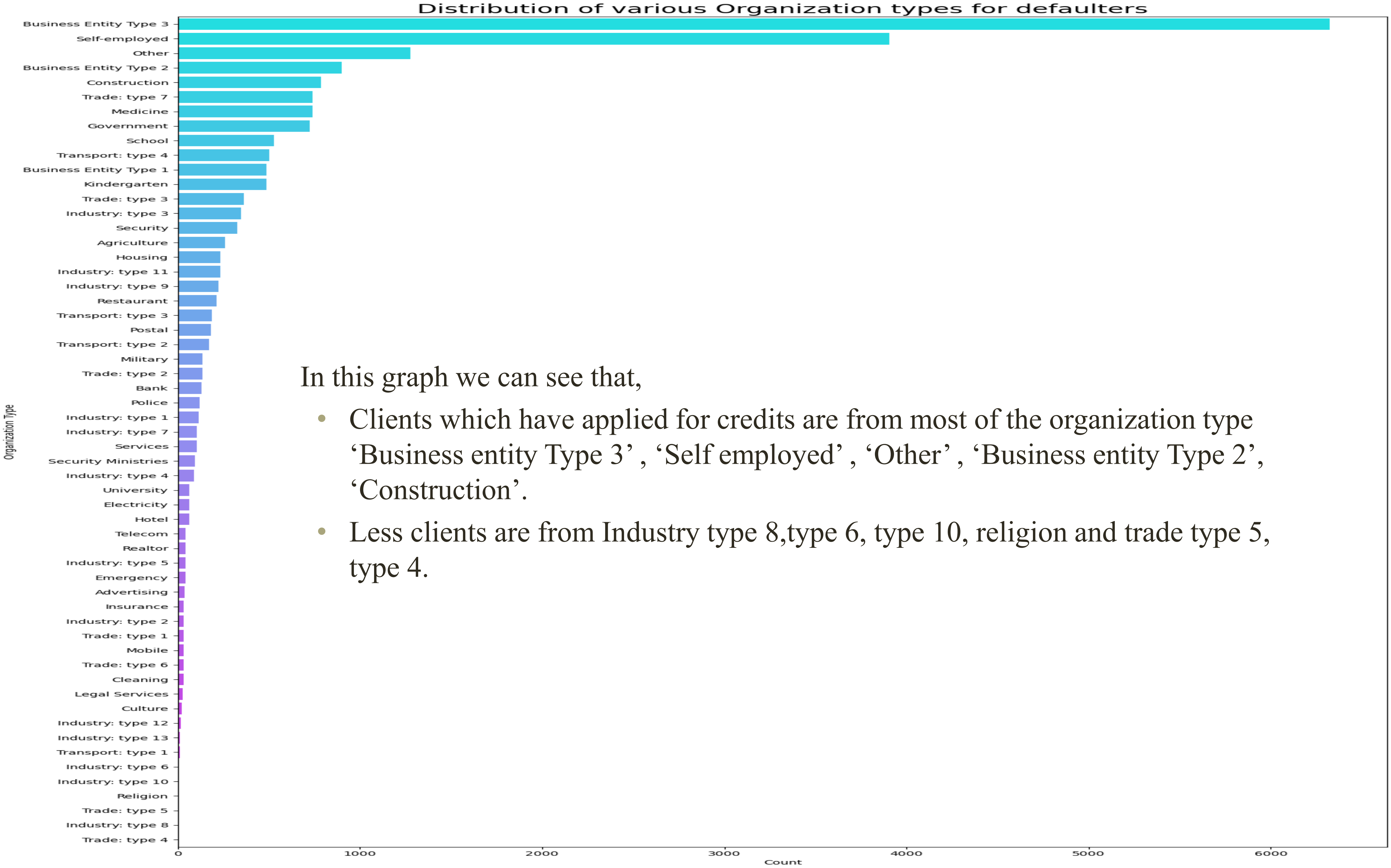
In this graph we can see that,

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



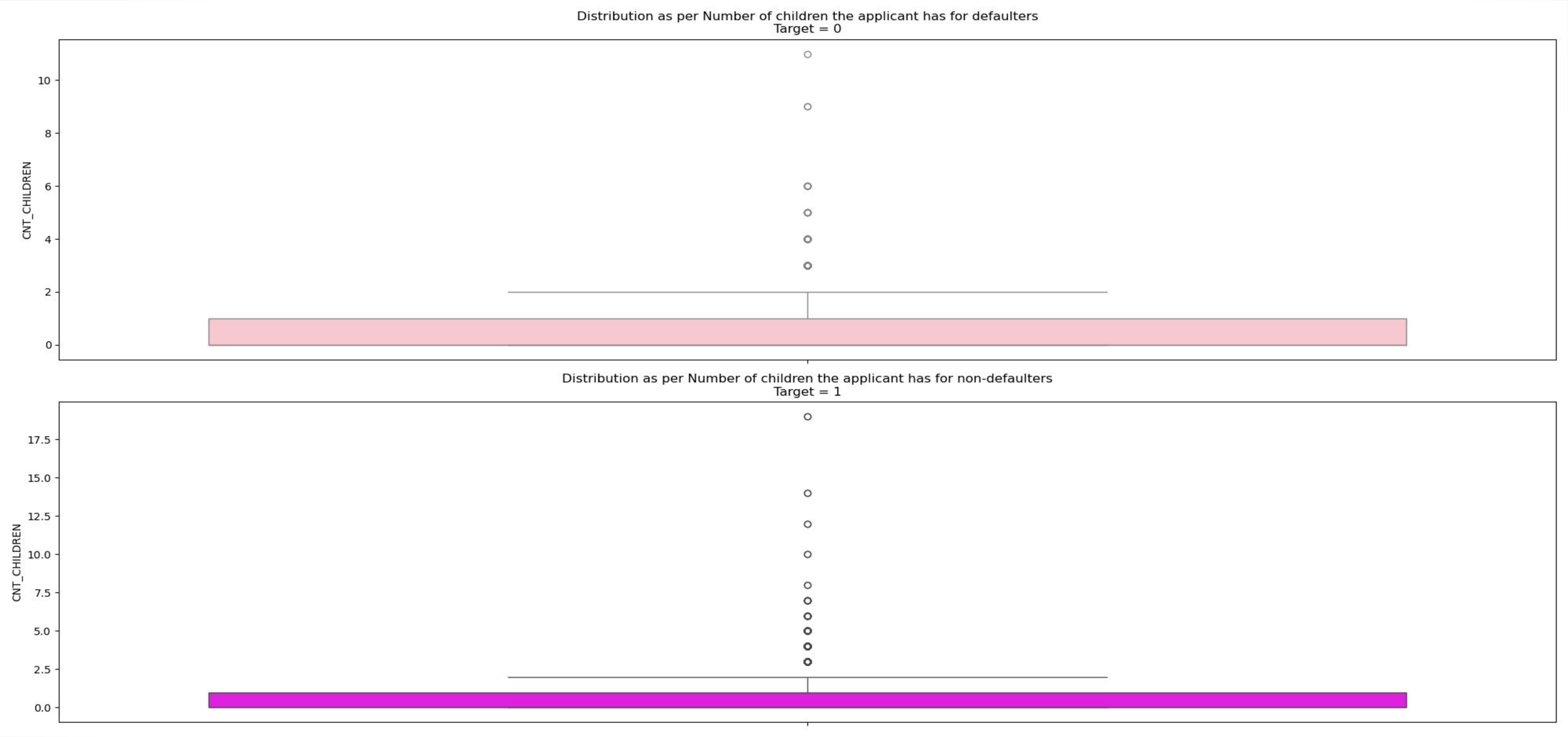
In this graph we can see that,

- Contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’.
- For this also, females are leading for applying credits.



In this graph we can see that,

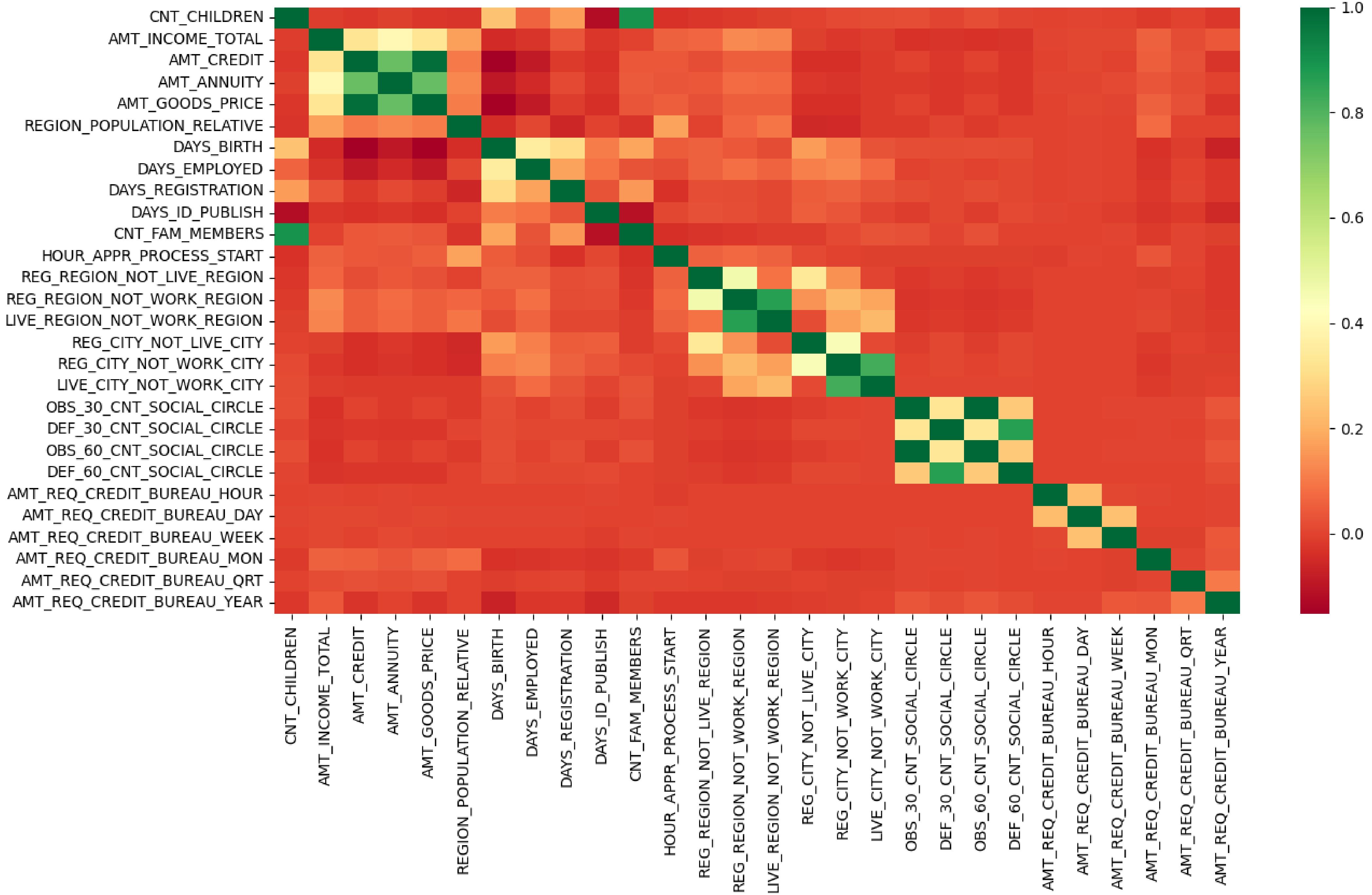
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’ , ‘Other’ , ‘Business entity Type 2’, ‘Construction’.
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.



In this graph we can see that,

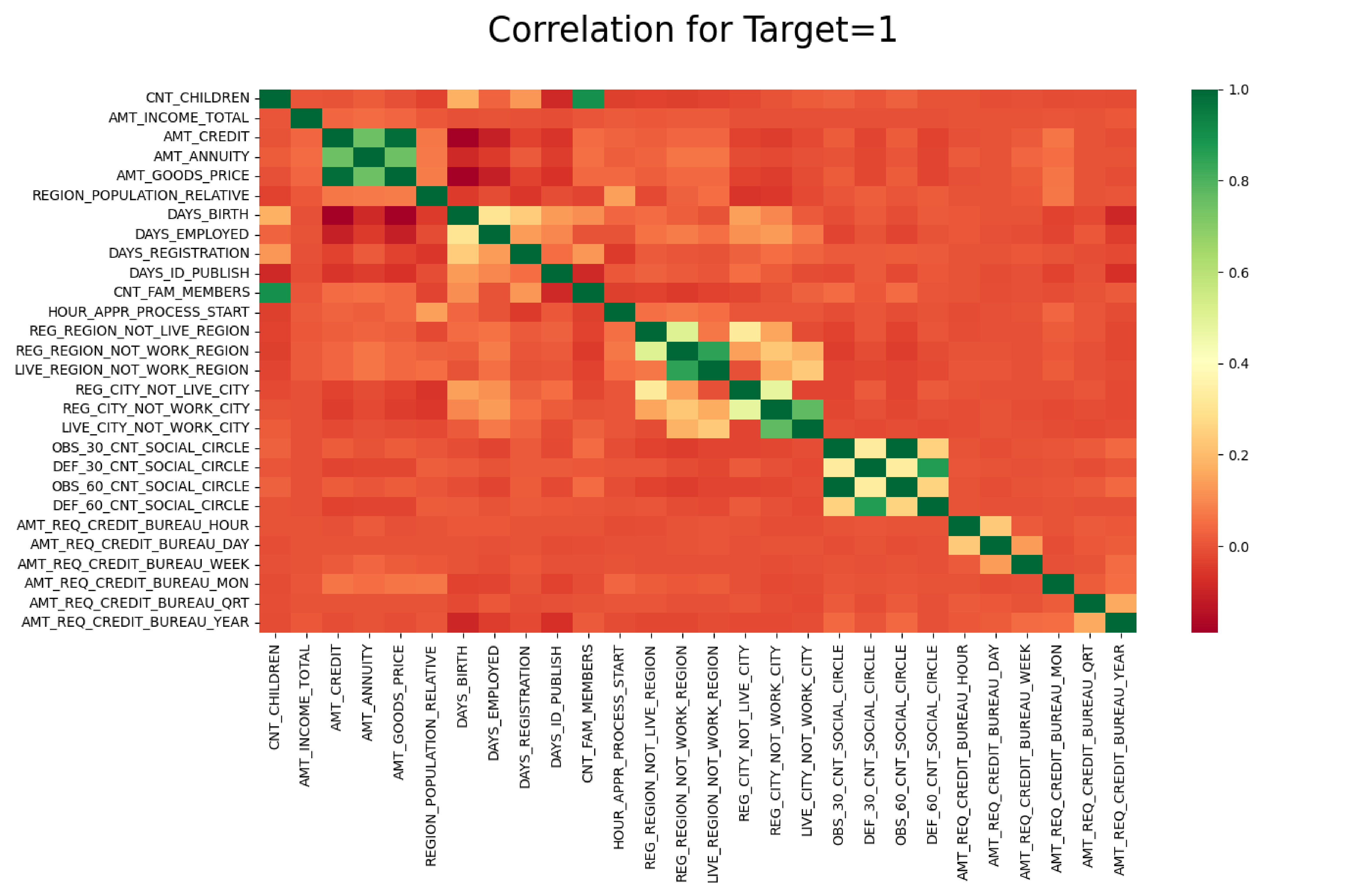
- It is observed that people with no children or 1 child take maximum loans.
- There are applicants with more than 10 children which we would consider as outliers.

Correlation for Target=0



Observations from Heatmap for Target 0

- 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.
- Less children client have in densely populated area.
- Credit amount is higher to densely populated area.
- The income is also higher in densely populated area.

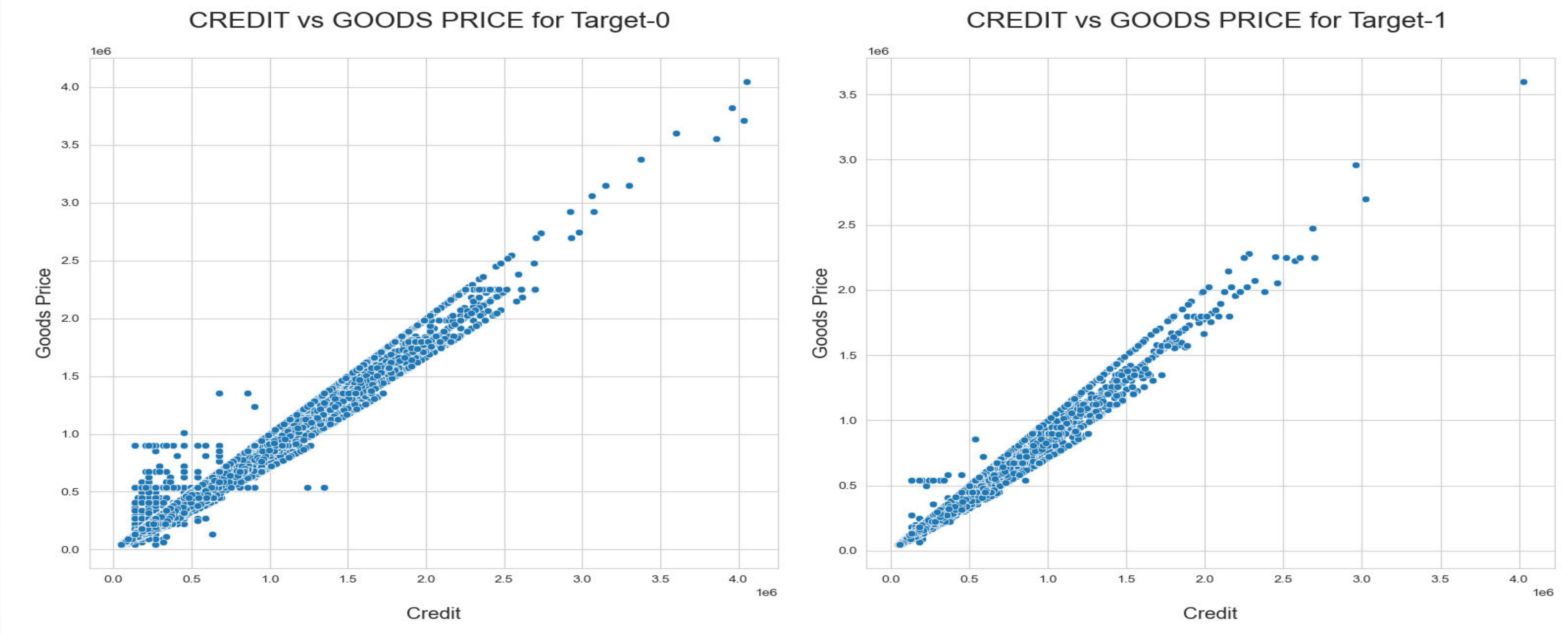

Correlation for Target=1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
CNT_FAM_MEMBERS
HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
1.0
0.8
0.6
0.4
0.2
0.0

Observations from Heatmap for Target 1

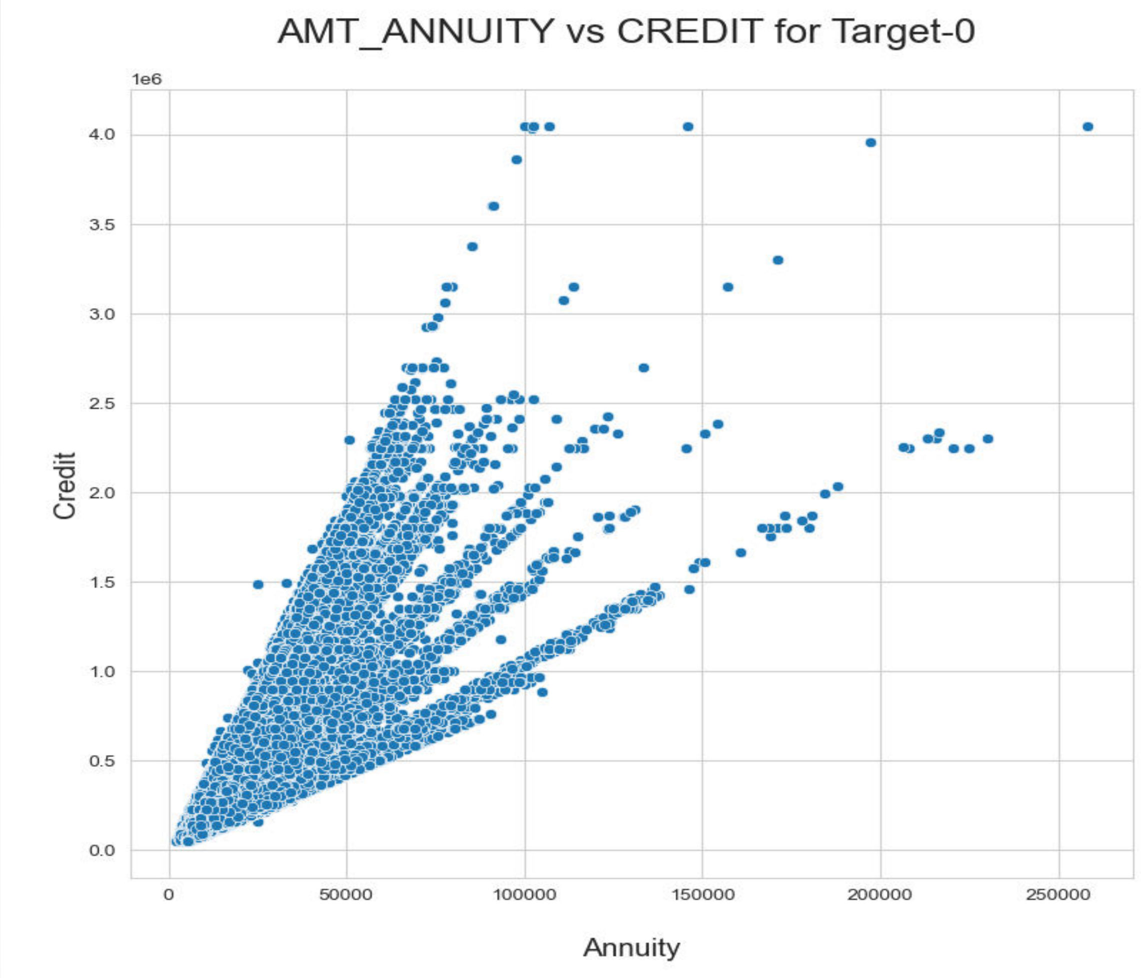
This heatmap for Target 1 is also having quite a same observation just like Target 0. But few points are different which are listed below:

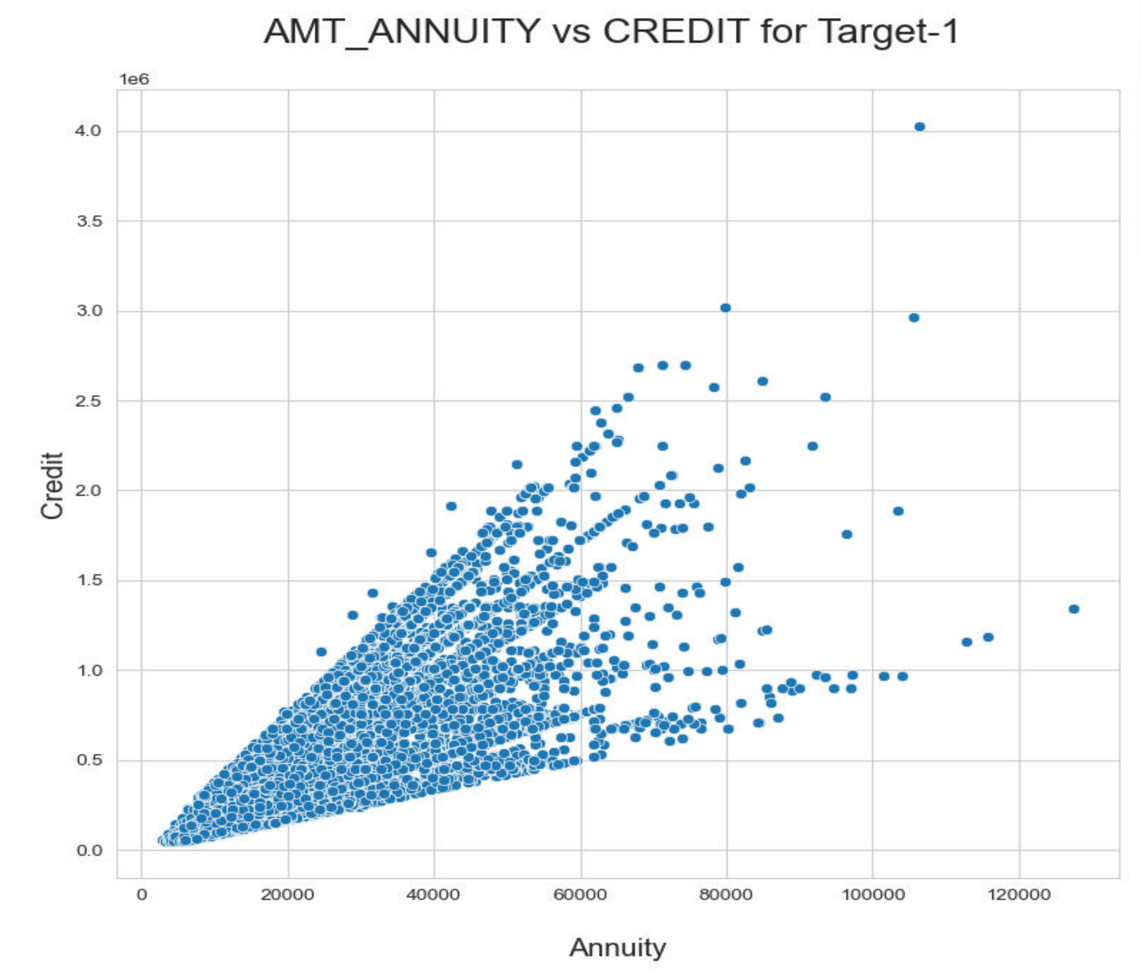
- The client's permanent address does not match contact address are having less children and vice-versa.
- The client's permanent address does not match work address are having less children and vice-versa.

Numerical Bivariate Analysis



For both target_0 and target_1 we can see that, AMT CREDIT and AMT GOODS PRICE are highly correlated, which means if there is an increase in goods price, the credit is increased directly and vice versa.

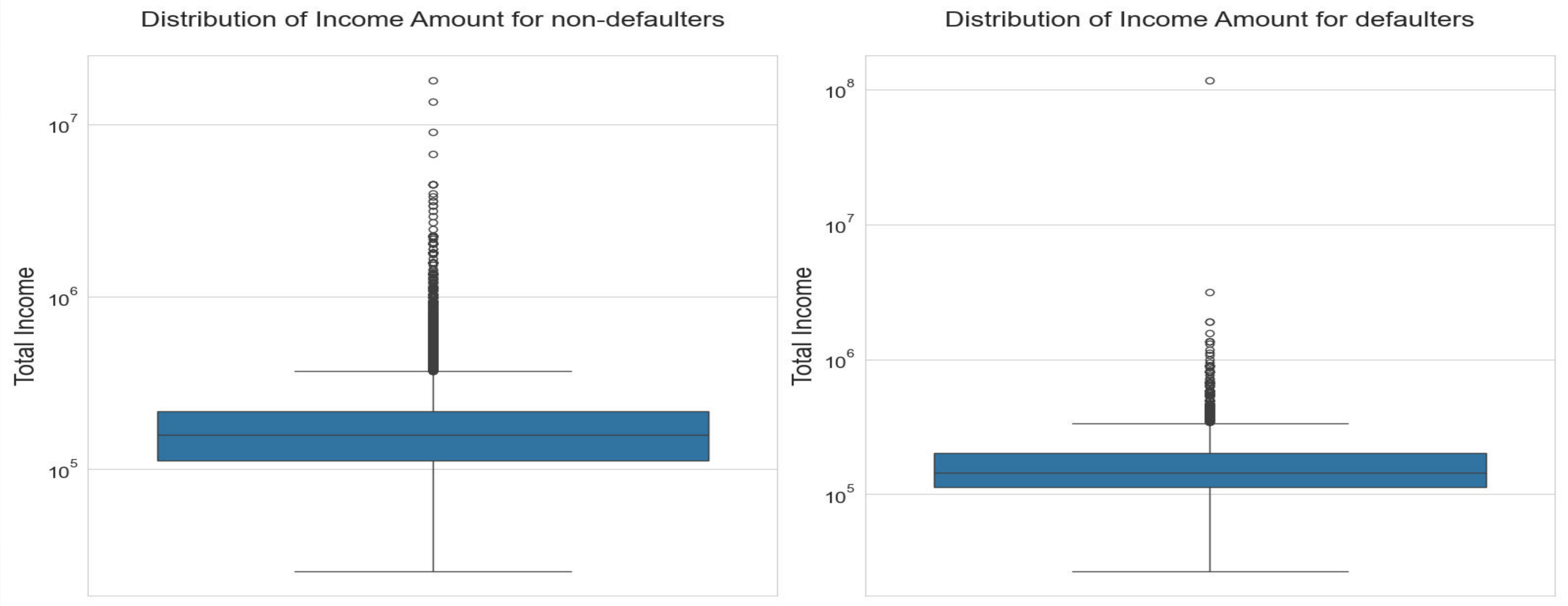




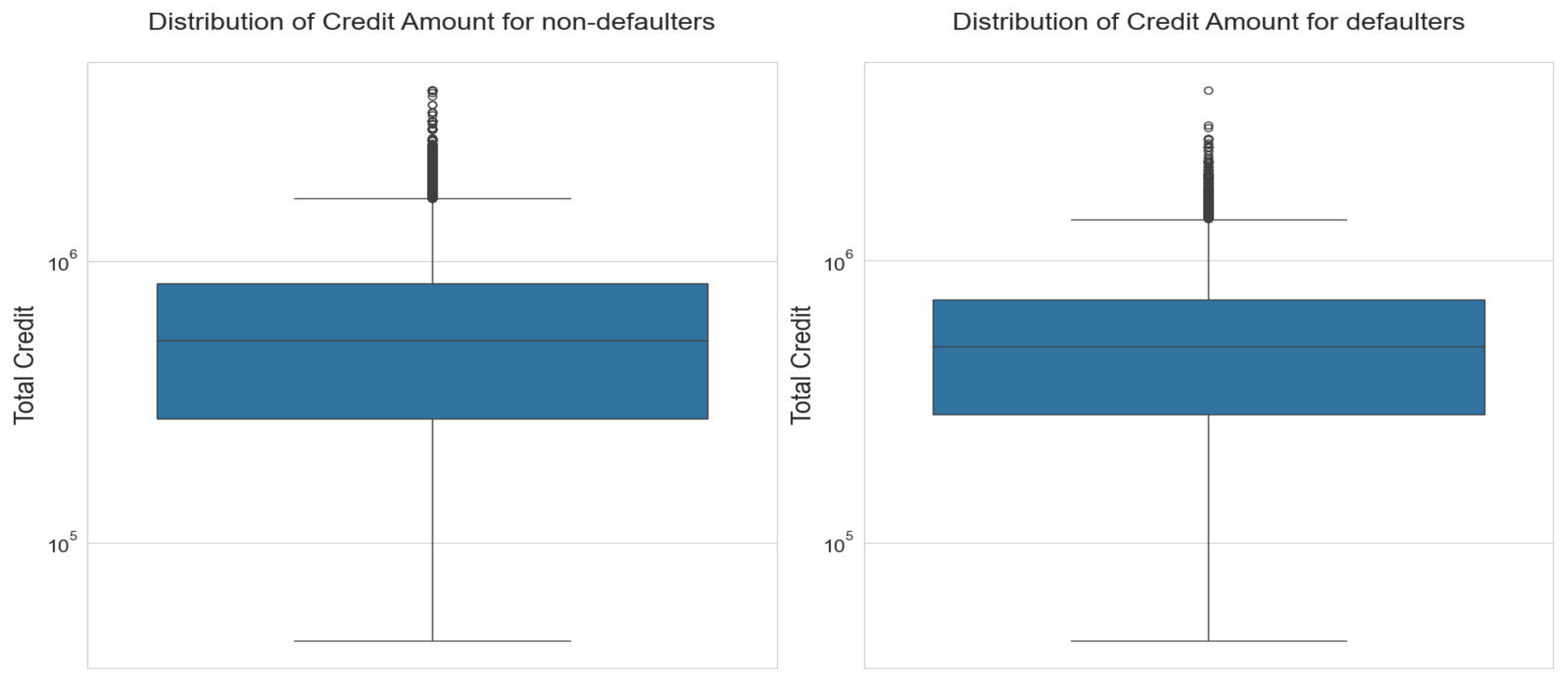
For both target_0 and target_1 we can see that, AMT_ANNUITY and AMT_CREDIT have strong positive correlation. This means that as Annuity amount increases, Credit amount also goes higher.

Finding Outliers

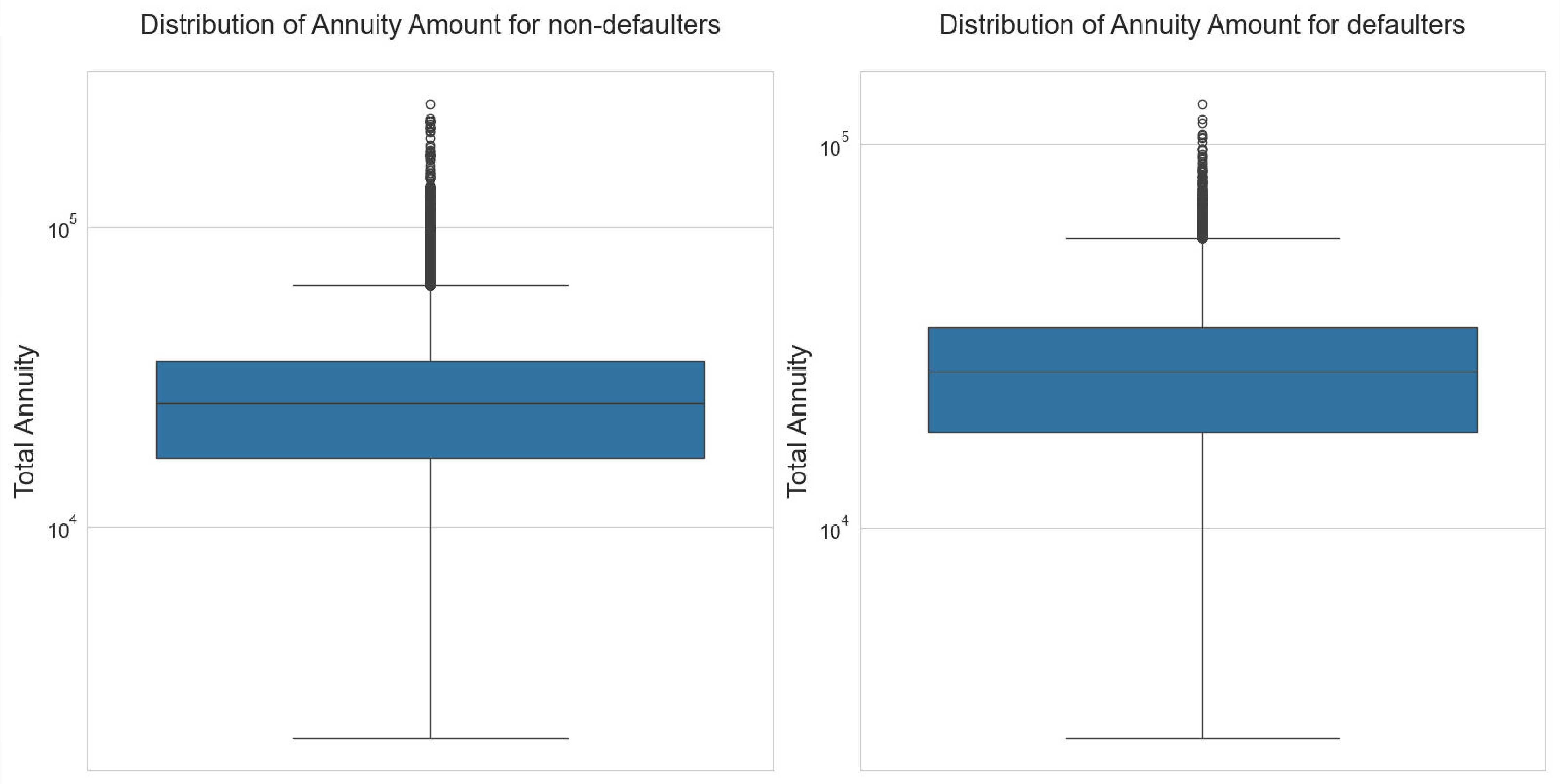
Univariate Analysis for Target: 0 &1



- For both target_0 and target_1, there seems to be an equal distribution of the Income amount of the clients.
- There are some outliers present in the dataset.

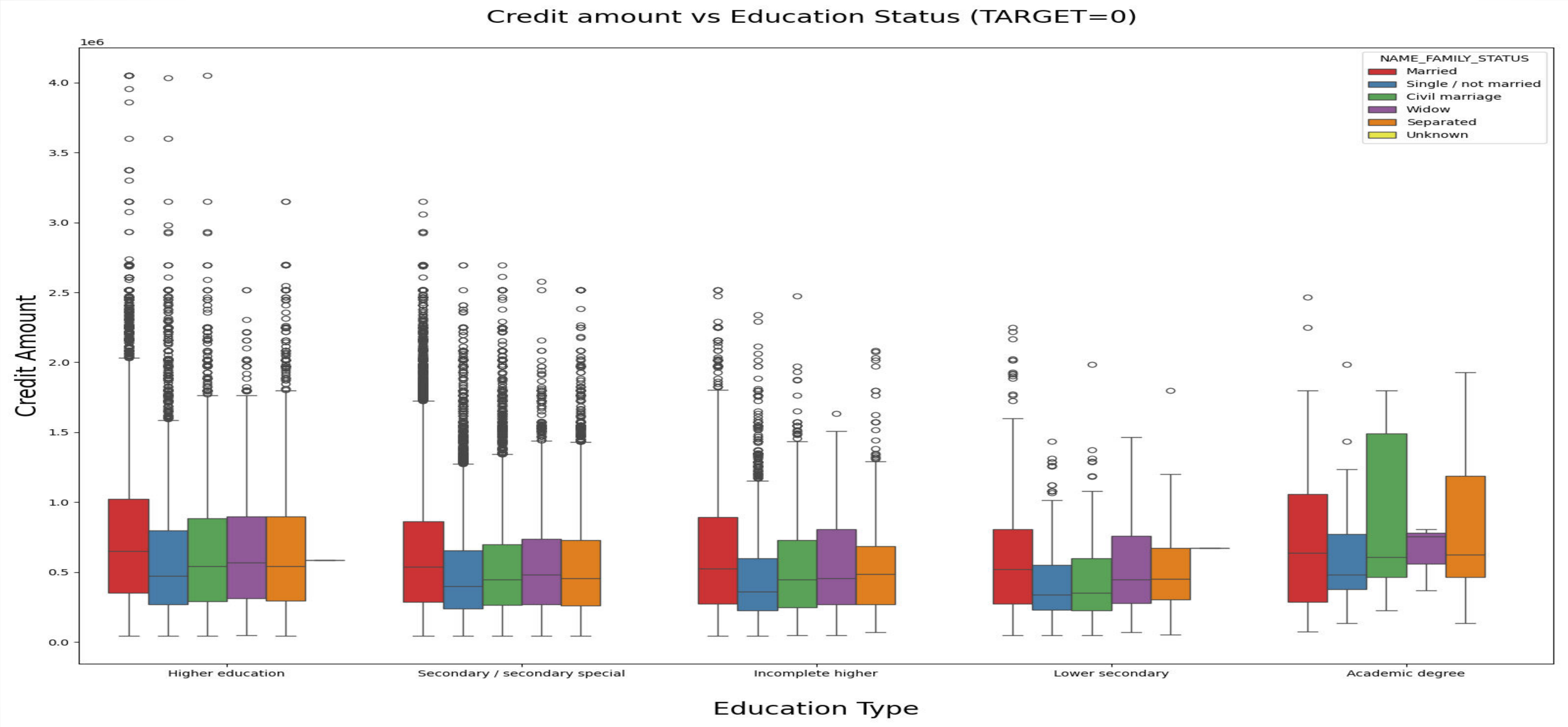


- For both target_0 and target_1, the first quartile is bigger than the third quartile, that means most of the client credit lies in the first quartile.
- There seems some outliers in the Credit boxplot.



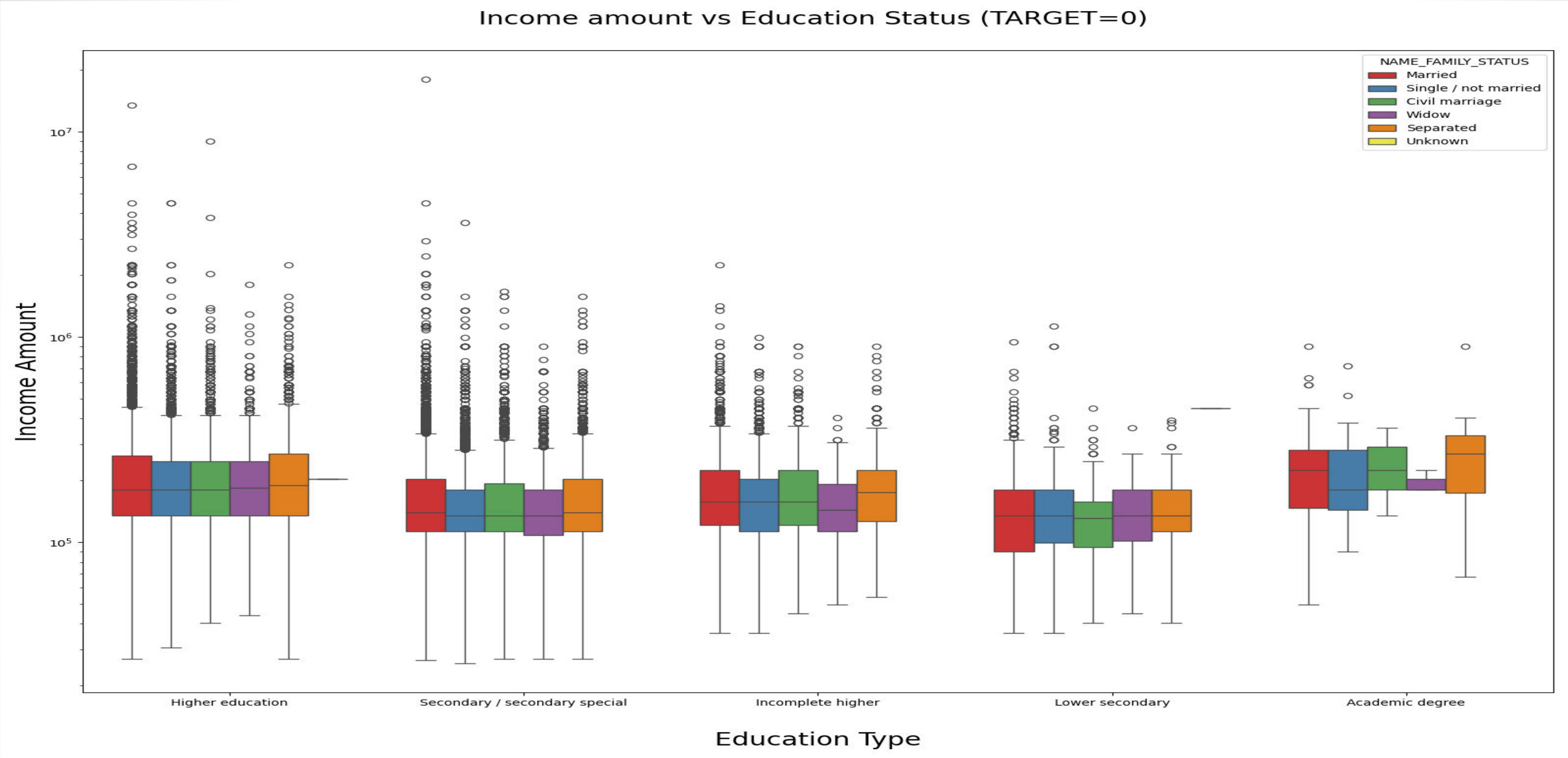
- For both target_0 and target_1, the first quartile is bigger than the third quartile.
- There seems some outliers in the Annuity boxplot.

Multivariate Analysis for Target: 0 &1



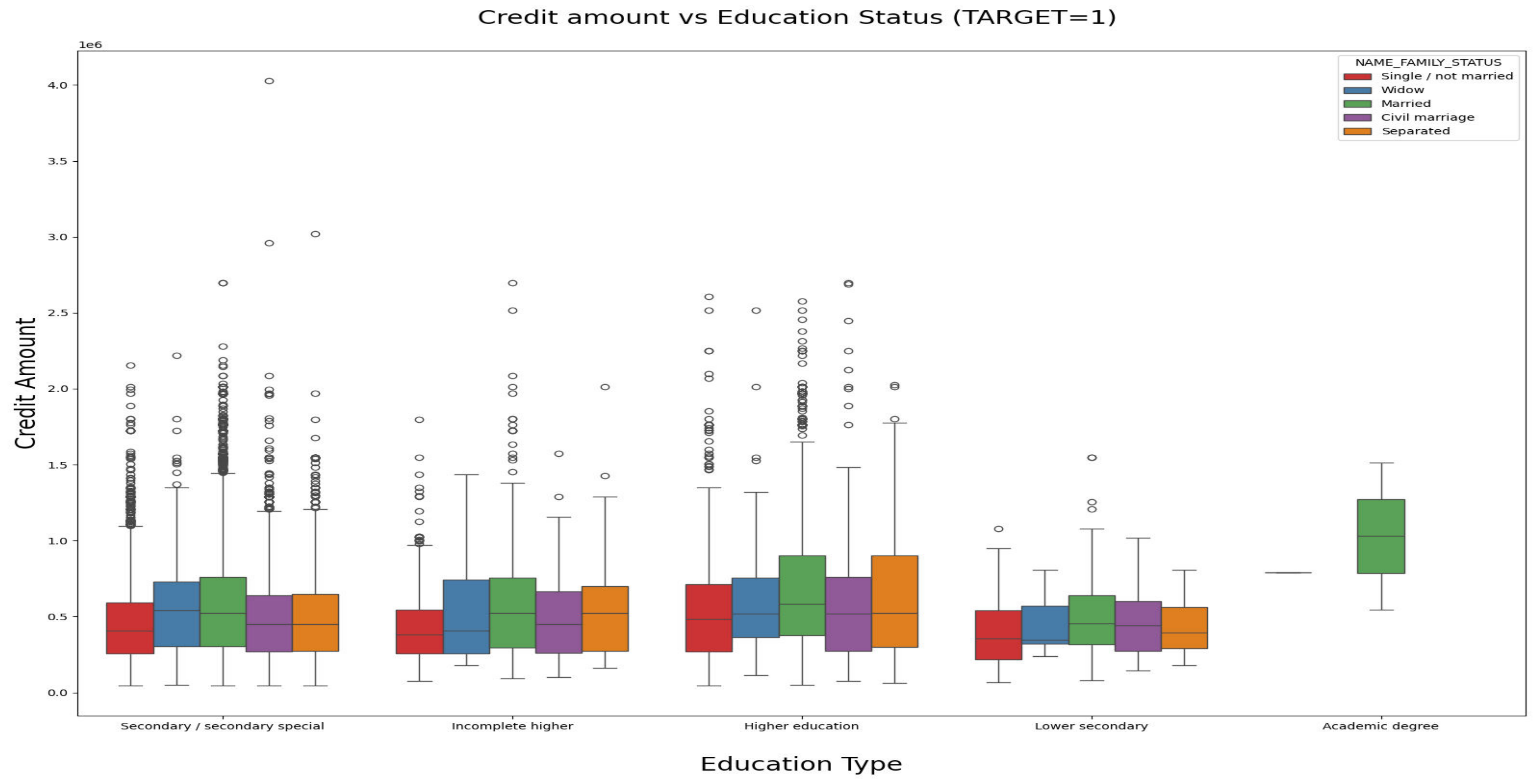
Conclusion from the graph:

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



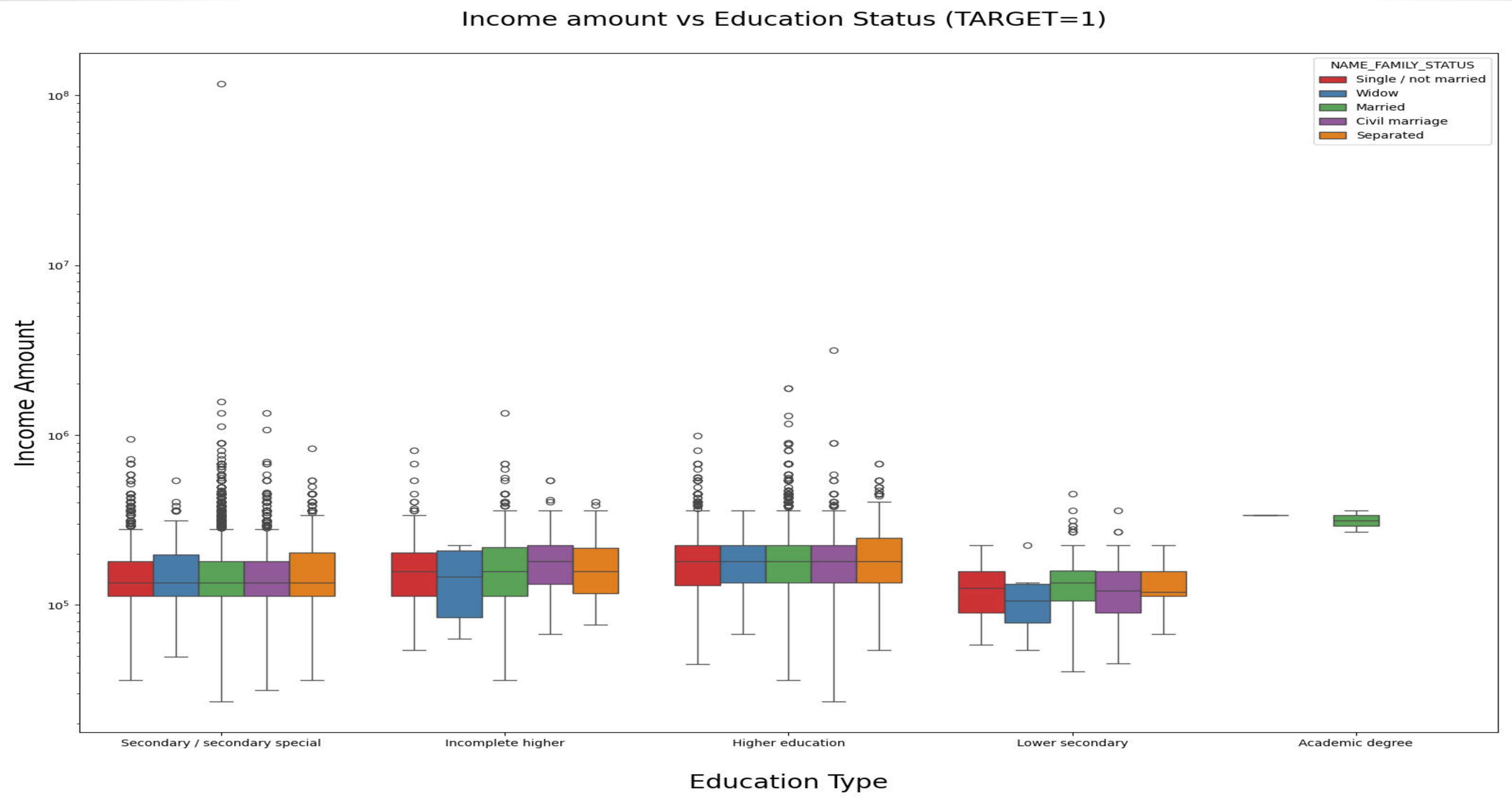
Conclusion from the graph:

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



Conclusion from the graph:

- 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.



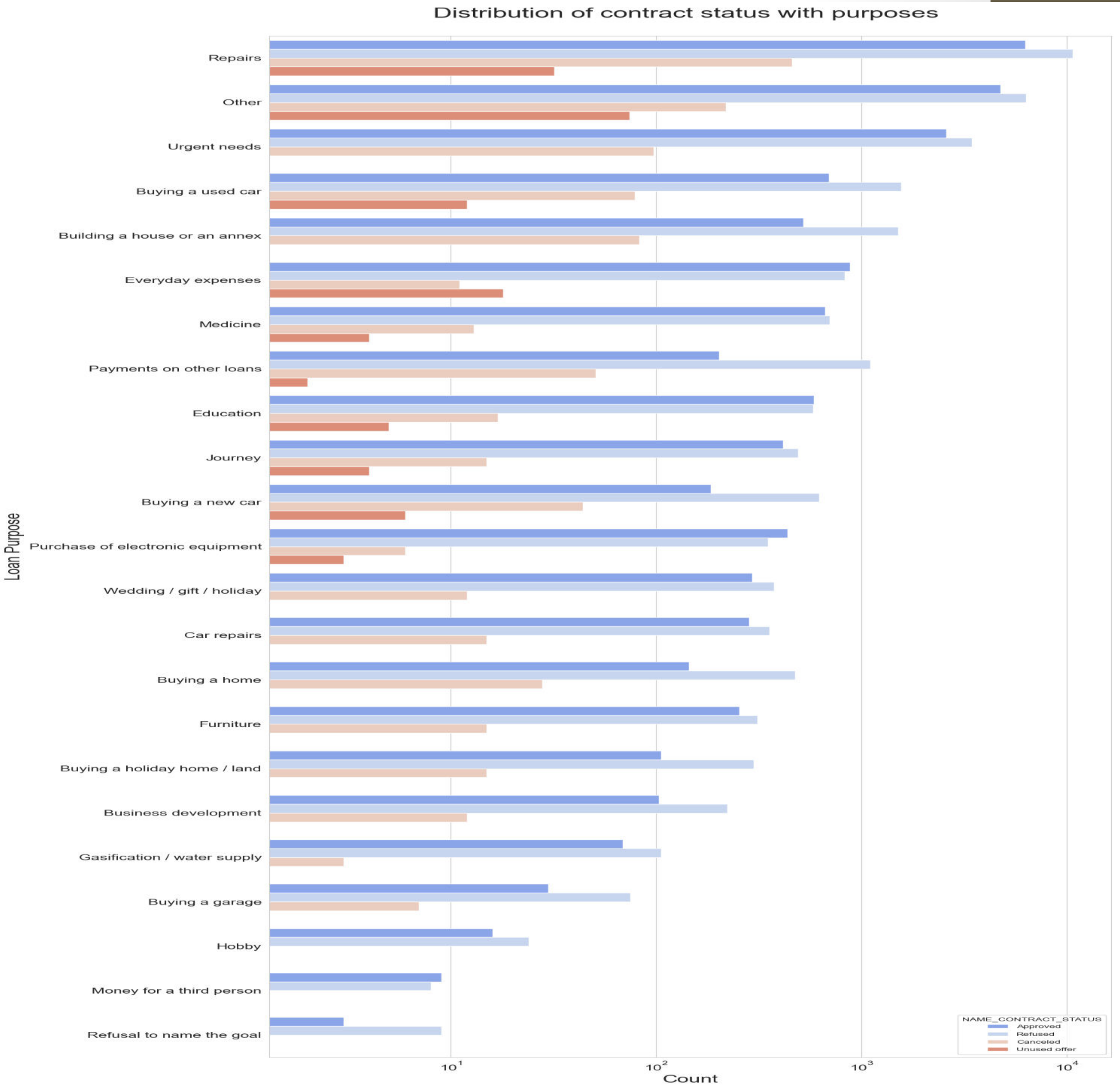
Conclusion from the graph:

- From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status.
- Lower secondary have less income amount than others.

Performing Univariate & Bivariate analysis after merging datasets

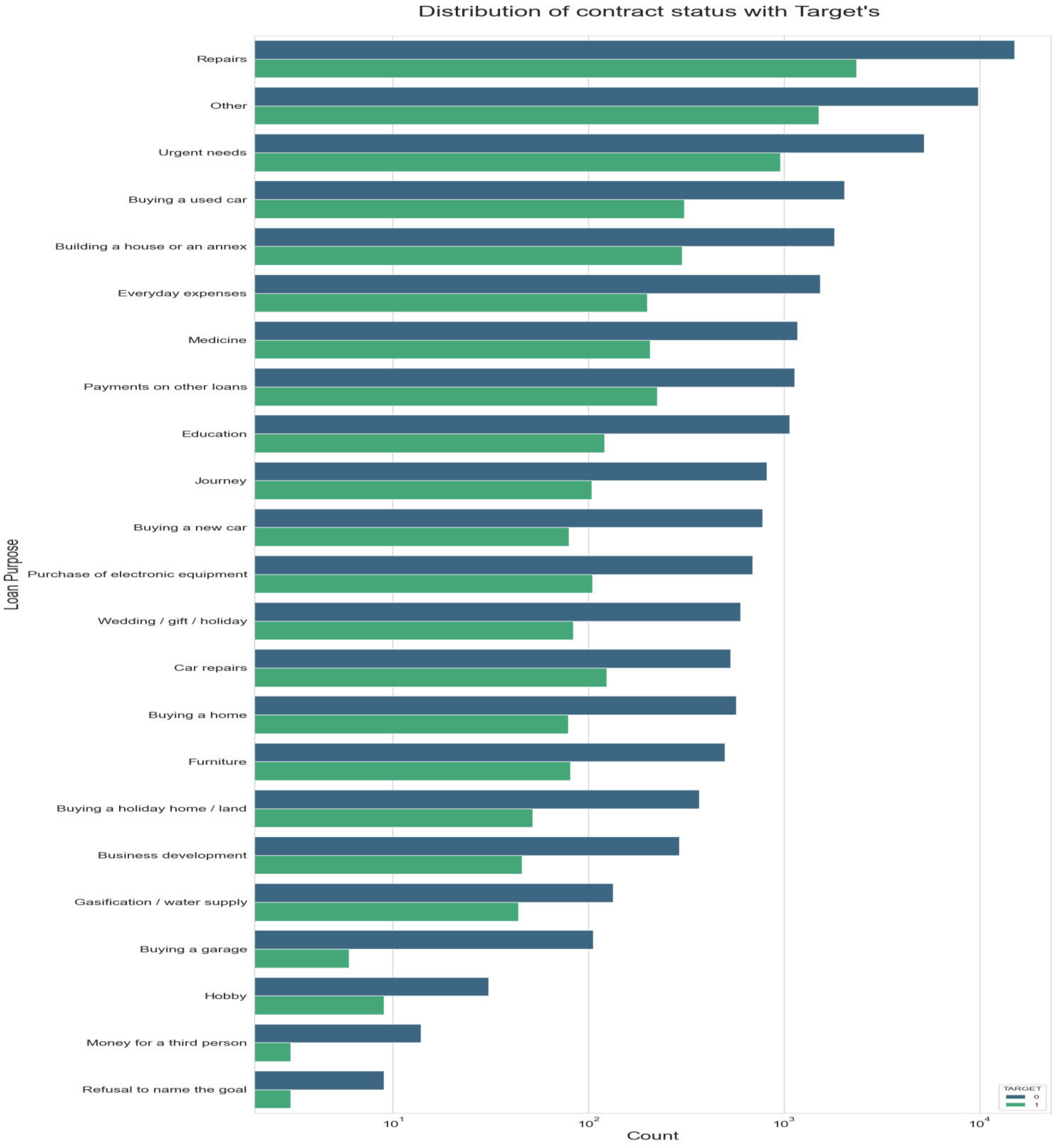
Conclusion from the graph,

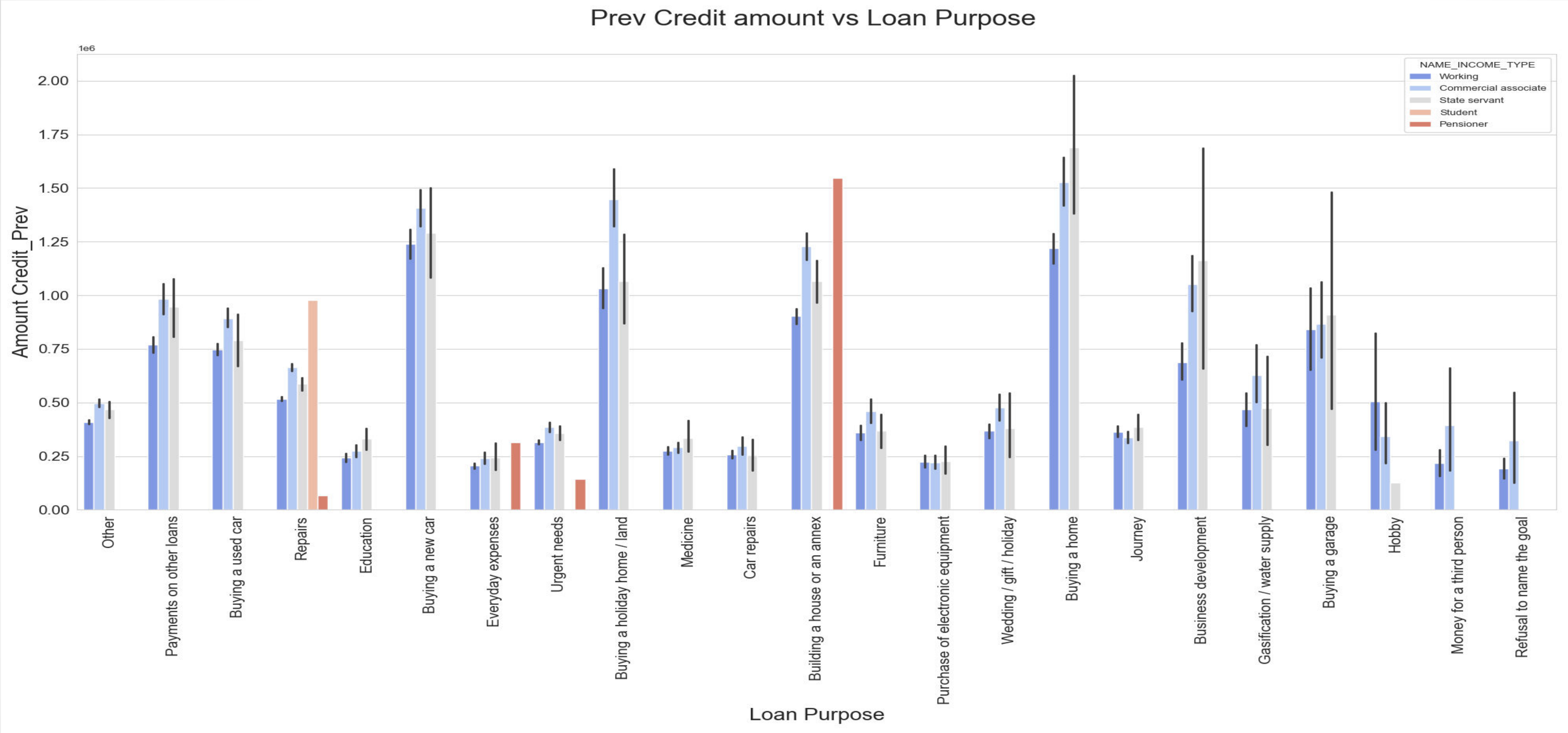
- Most rejection of loans came from purpose 'Repairs'.
- For education purposes we have equal number of approves and rejection.
- Paying other loans and buying a new car is having significant higher rejection than approves.



Conclusion from the graph,

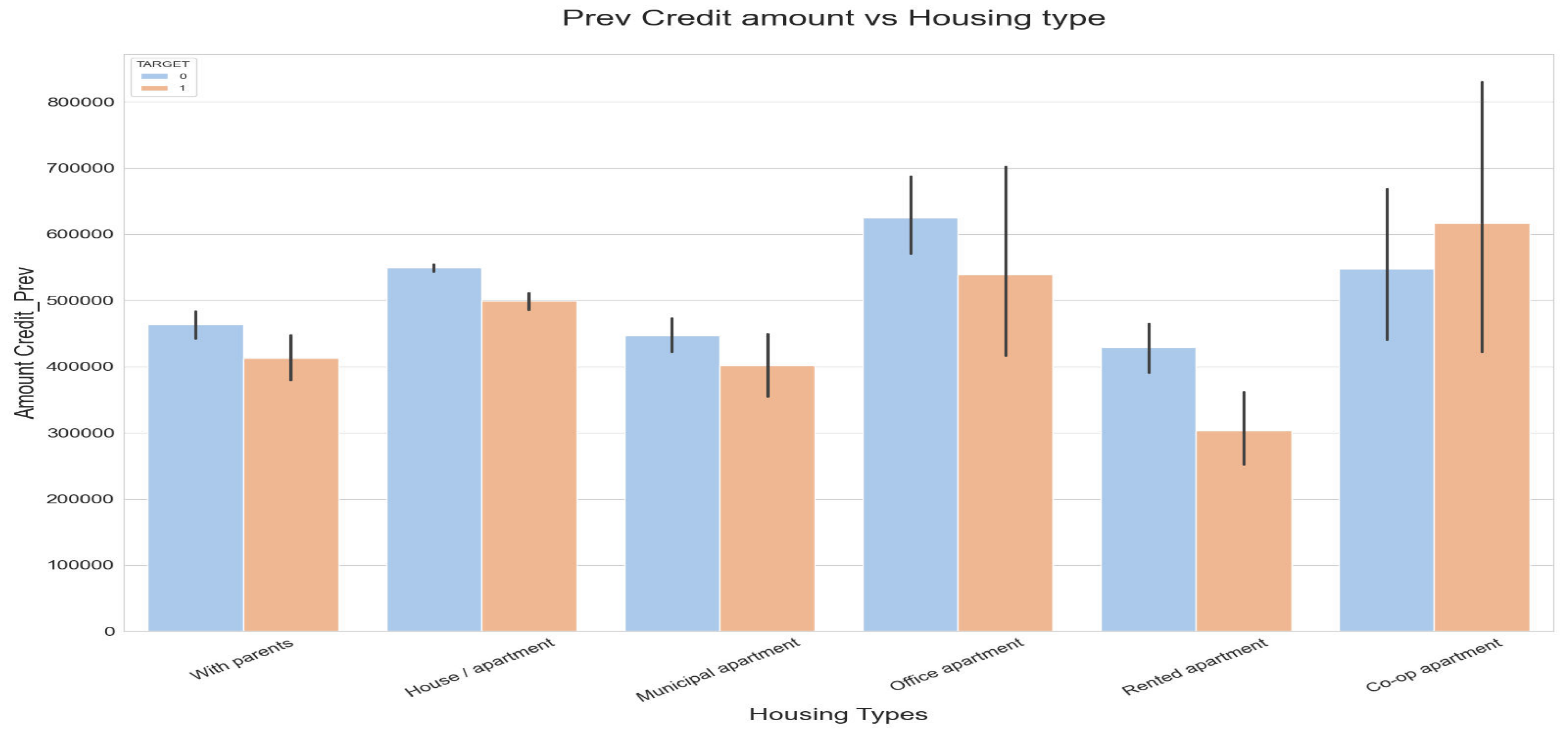
- Loan purposes with 'Repairs' are facing more difficulties in payment on time.
- There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business development', 'Buying land', 'Buying a new car' and 'Education' Hence we can focus on these purposes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





Conclusion from the graph,

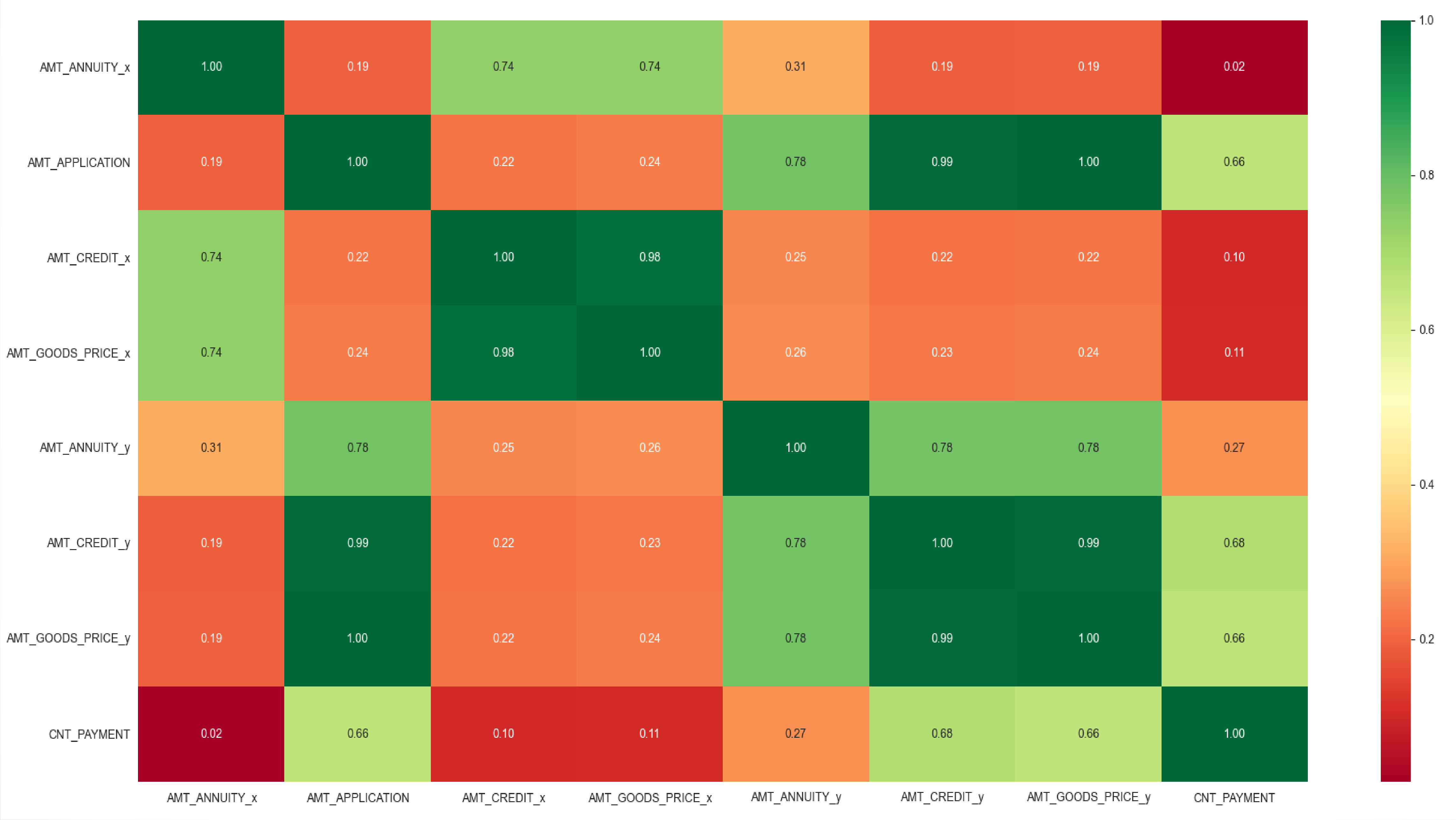
- The credit amount of Loan purposes like 'Buying a home', 'Buying a land', 'Buying a new car' and 'Building a house' is higher.
- Income type of state servants have a significant amount of credit applied
- Money for third person or a Hobby is having less credits applied for.



Conclusion from the graph,

- 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.

Heatmap for Merged data



	AMT_ANNUITY_x	AMT_APPLICATION	AMT_CREDIT_x	AMT_GOODS_PRICE_x	AMT_ANNUITY_y	AMT_CREDIT_y	AMT_GOODS_PRICE_y	CNT_PAYMENT
AMT_ANNUITY_x	1.00	0.19	0.74	0.74	0.31	0.19	0.19	0.02
AMT_APPLICATION	0.19	1.00	0.22	0.24	0.78	0.99	1.00	0.66
AMT_CREDIT_x	0.74	0.22	1.00	0.98	0.25	0.22	0.22	0.10
AMT_GOODS_PRICE_x	0.74	0.24	0.98	1.00	0.26	0.23	0.24	0.11
AMT_ANNUITY_y	0.31	0.78	0.25	0.26	1.00	0.78	0.78	0.27
AMT_CREDIT_y	0.19	0.99	0.22	0.23	0.78	1.00	0.99	0.68
AMT_GOODS_PRICE_y	0.19	1.00	0.22	0.24	0.78	0.99	1.00	0.66
CNT_PAYMENT	0.02	0.66	0.10	0.11	0.27	0.68	0.66	1.00

Observations from Heatmap

- AMT_APPLICATION has a high correlation with AMT_ANNUITY_y, AMT_CREDIT_y, AMT_GOODS_PRICE_y and decent correlation with CNT_PAYMENT.
- AMT_GOODS_PRICE_y has a high correlation with AMT_ANNUITY_y, AMT_CREDIT_y, AMT_APPLICATION and decent correlation with CNT_PAYMENT.
- AMT_CREDIT_y has a high correlation with AMT_ANNUITY_y, AMT_GOODS_PRICE_y and decent correlation with CNT_PAYMENT.
- AMT_ANNUITY_y has a high correlation with AMT_GOODS_PRICE_y, AMT_CREDIT_y.
- AMT_ANNUITY_x has a high correlation with AMT_GOODS_PRICE_x ,AMT_CREDIT_x.
- AMT_CREDIT_x has a high correlation with AMT_GOODS_PRICE_x.

Conclusion from this loan data analysis

- Banks should approve loans more for Office apartment housing type as there are less payment difficulties.
- Banks should provide loans to ‘Repairs’ & ‘Others’ category in loan purpose.
- Banks should provide loans to the ‘Business Entity Type-3’ and ‘Self-Employed’ people.
- ‘Working’ people especially female employers are the best to target for the loans.
- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing type other than ‘Co-op apartment’ for successful payments.
- Loan purpose ‘Repair’ is having higher number of unsuccessful payments.
- Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.
- Approved clients in their previous applications and senior citizens in all categories would be more preferable.

THANK YOU