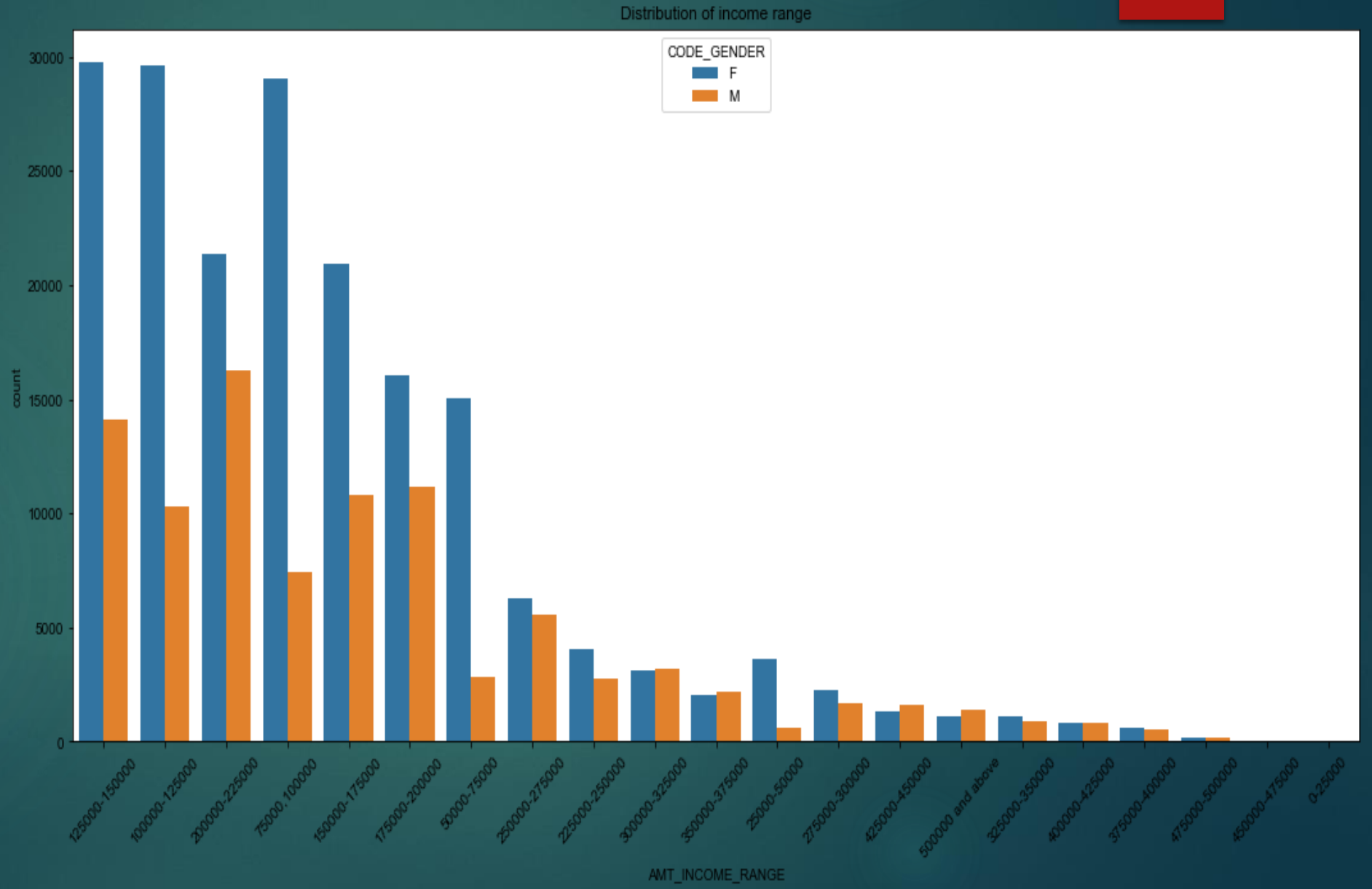
Univariate Analysis for Clients with No Difficulties (target 0)

Income Range Distribution

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

2.Graph depicts that females are more than male in having credits for the range.

3.Very less count for the income range of 400000 and above.

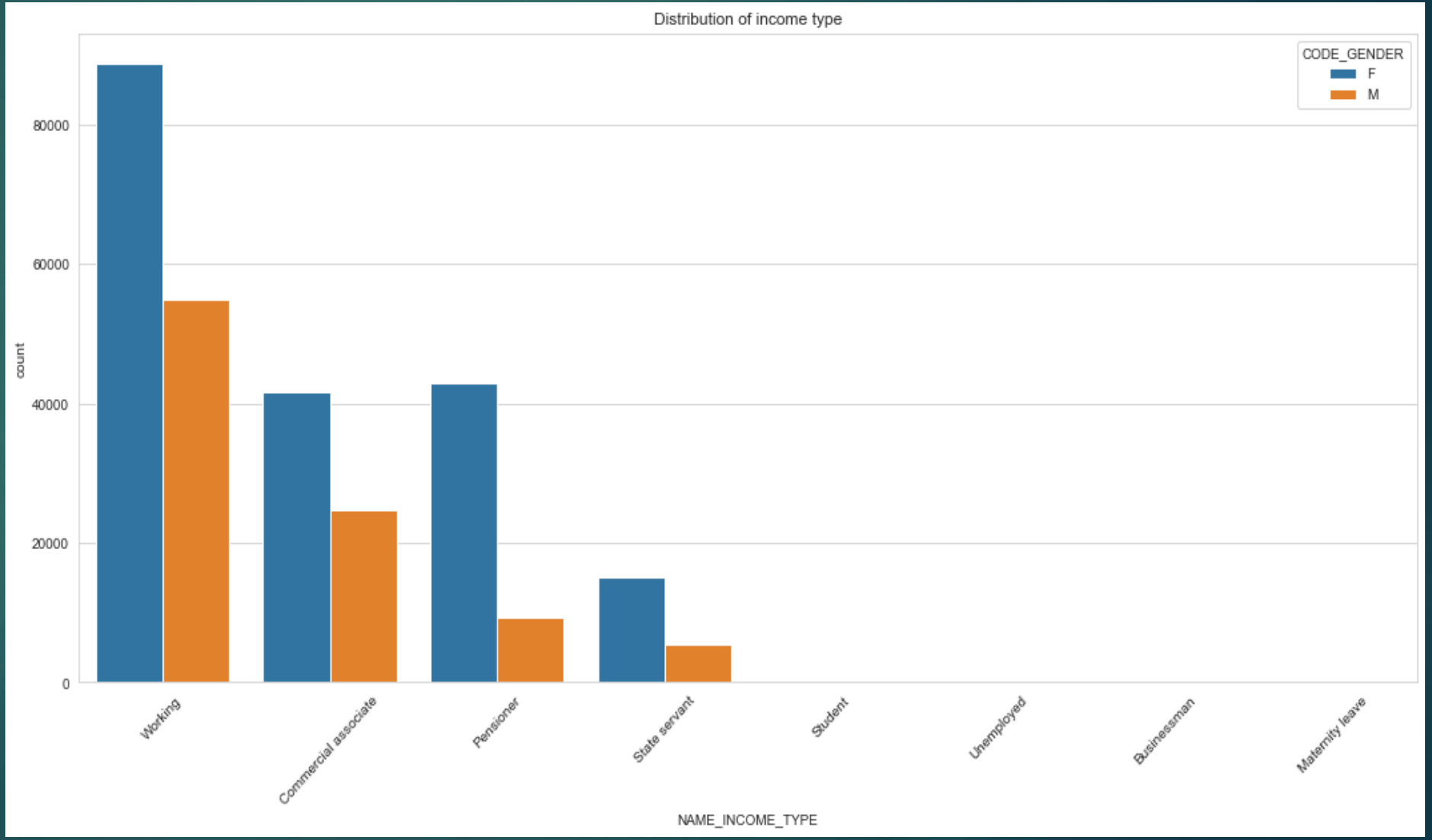


Income Type Distribution

1. For income type working, commercial associate, and State Servant the number of credits are higher than others.

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

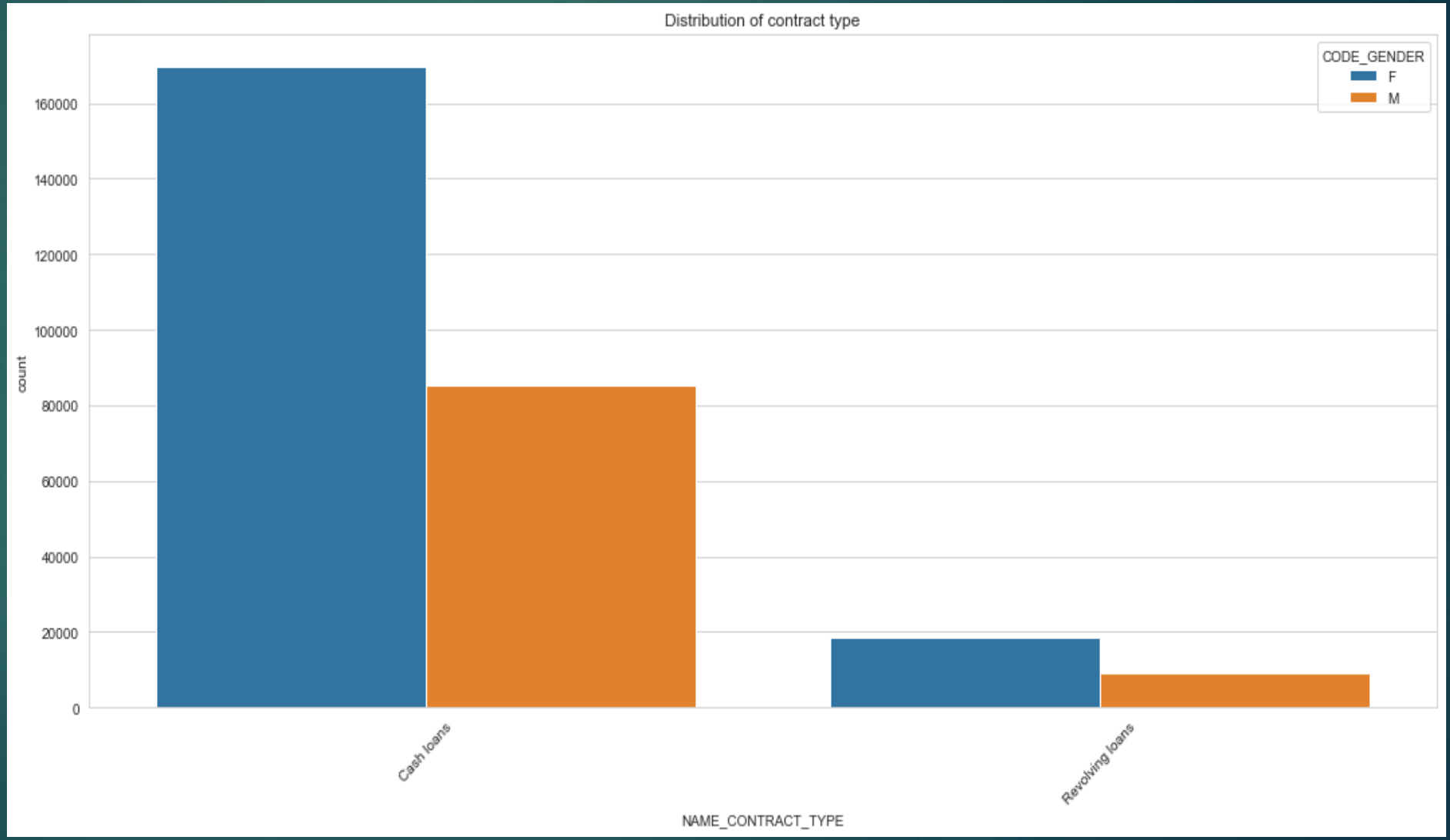
3. Less number of credits for income type student ,pensioner, Businessman and Maternity leave.



Distribution of Contract Type

1. For contract type cash loans is having higher number of credits than Revolving loans.

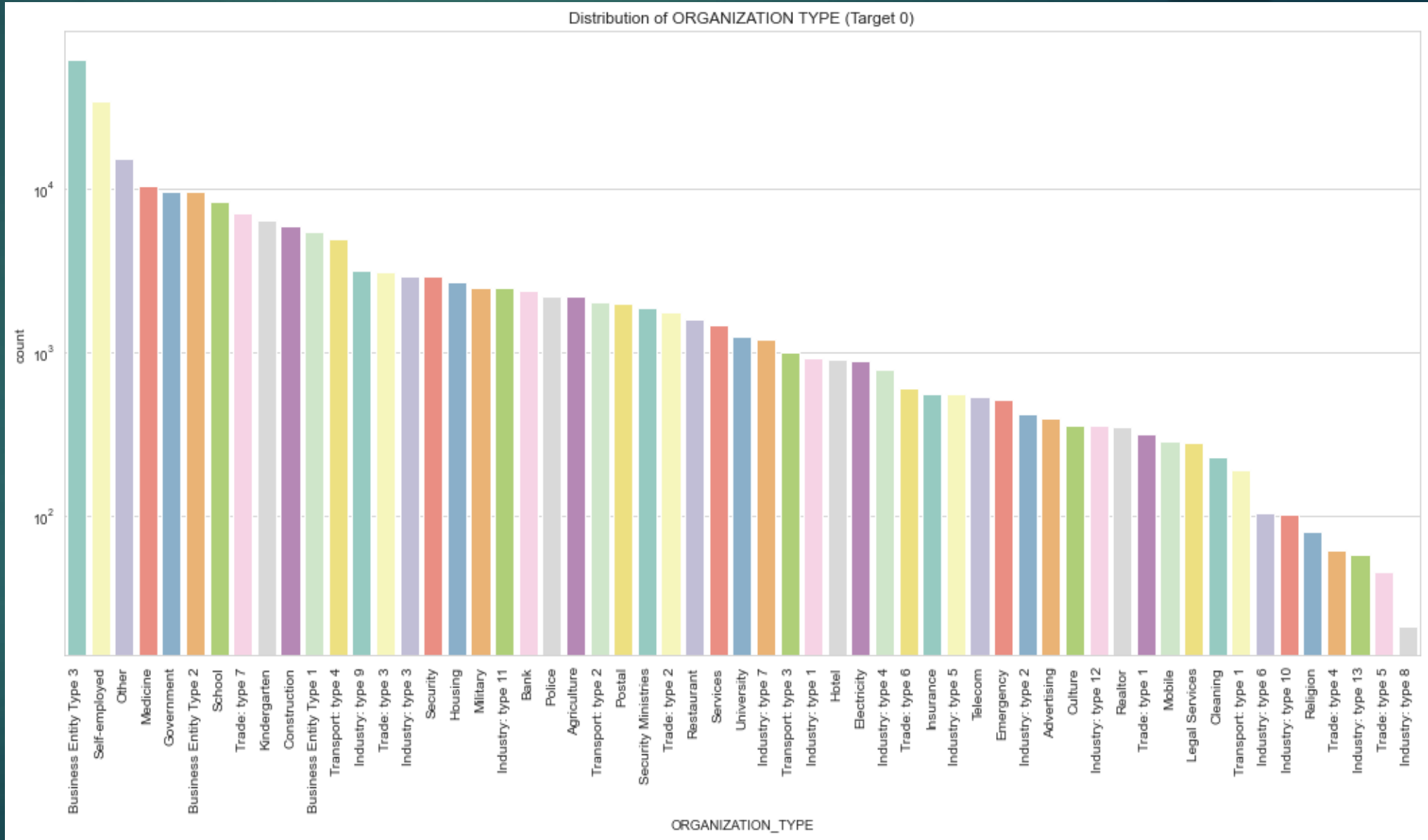
2. Here Females are leading for applying credits.



Distribution of Organization Type

1. Clients who have applied for credits are mostly from organization type Business entity Type 3 , Self employed, Other , Medicine and Government.

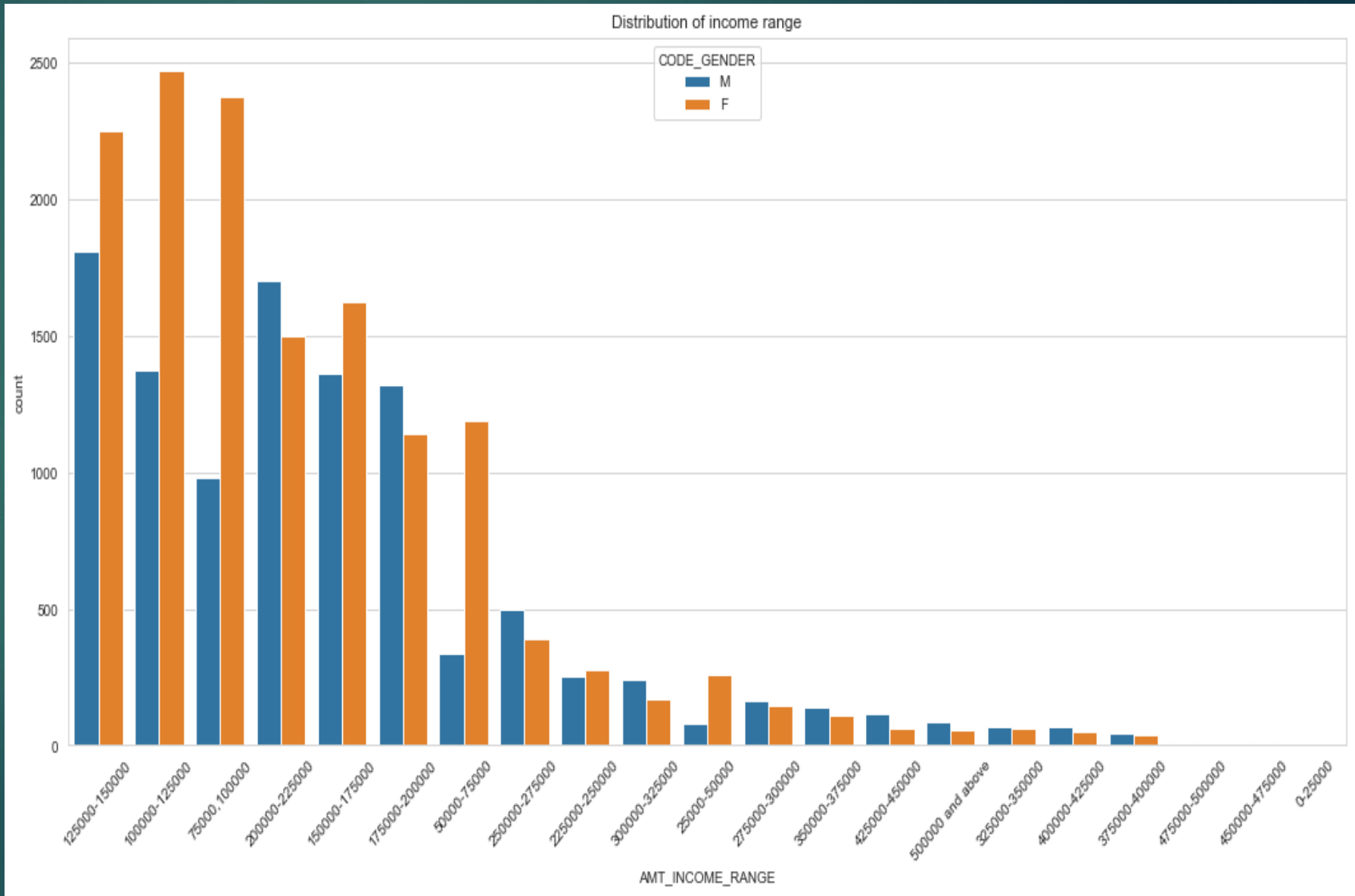
2. Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.



Univariate Analysis for Clients with Difficulties (target 1)

Distribution of Income Range

1. In some ranges male counts are higher than females.
2. Income range from 100000 to 200000 is having more number of credits.
3. Very less count for income range of 400000 and above.

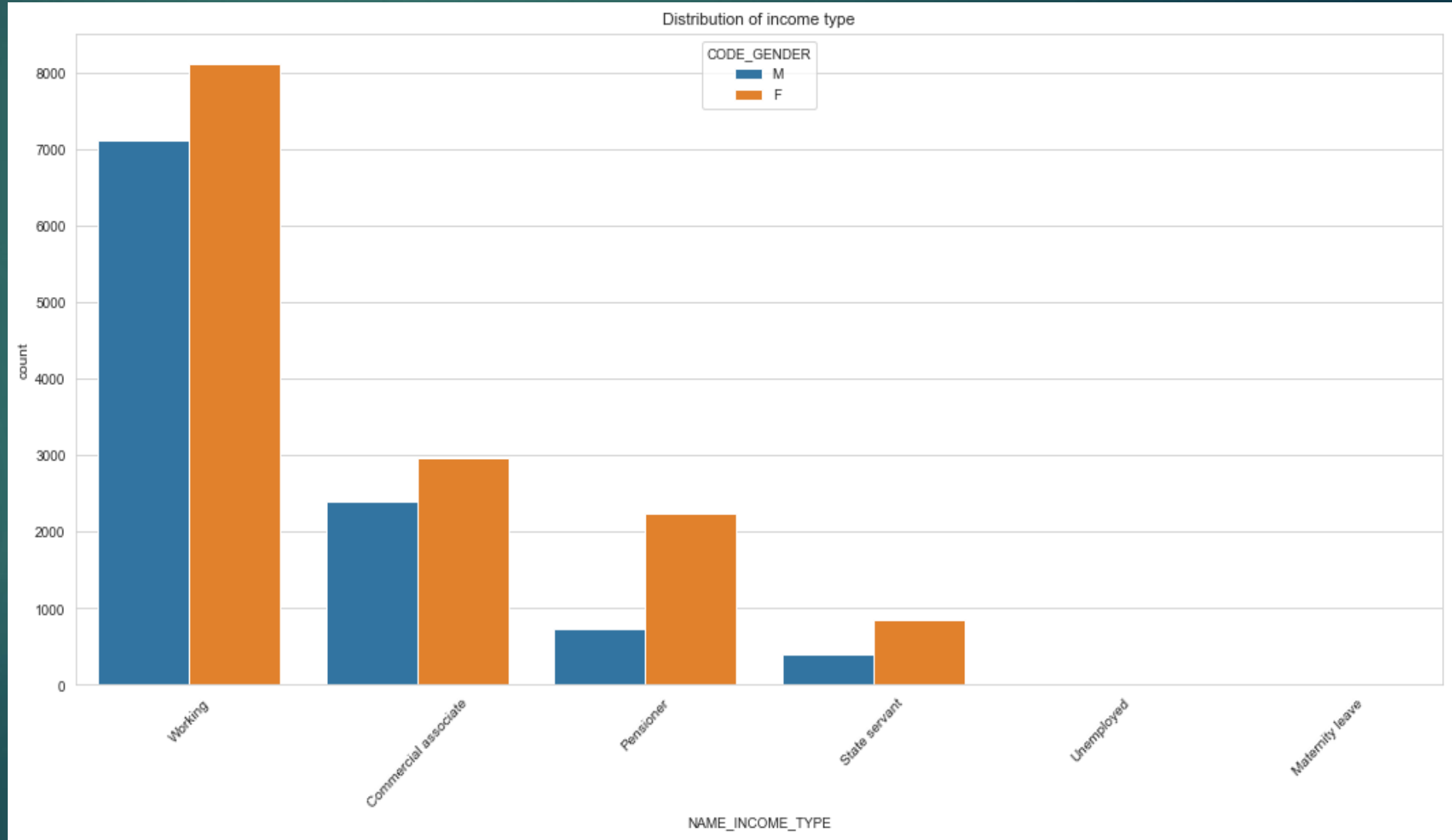


Distribution of Income Type

1. For income type working, commercial associate, and Pensioner the number of credits are higher than others.

2. Here also Females are having more number of credits than male.

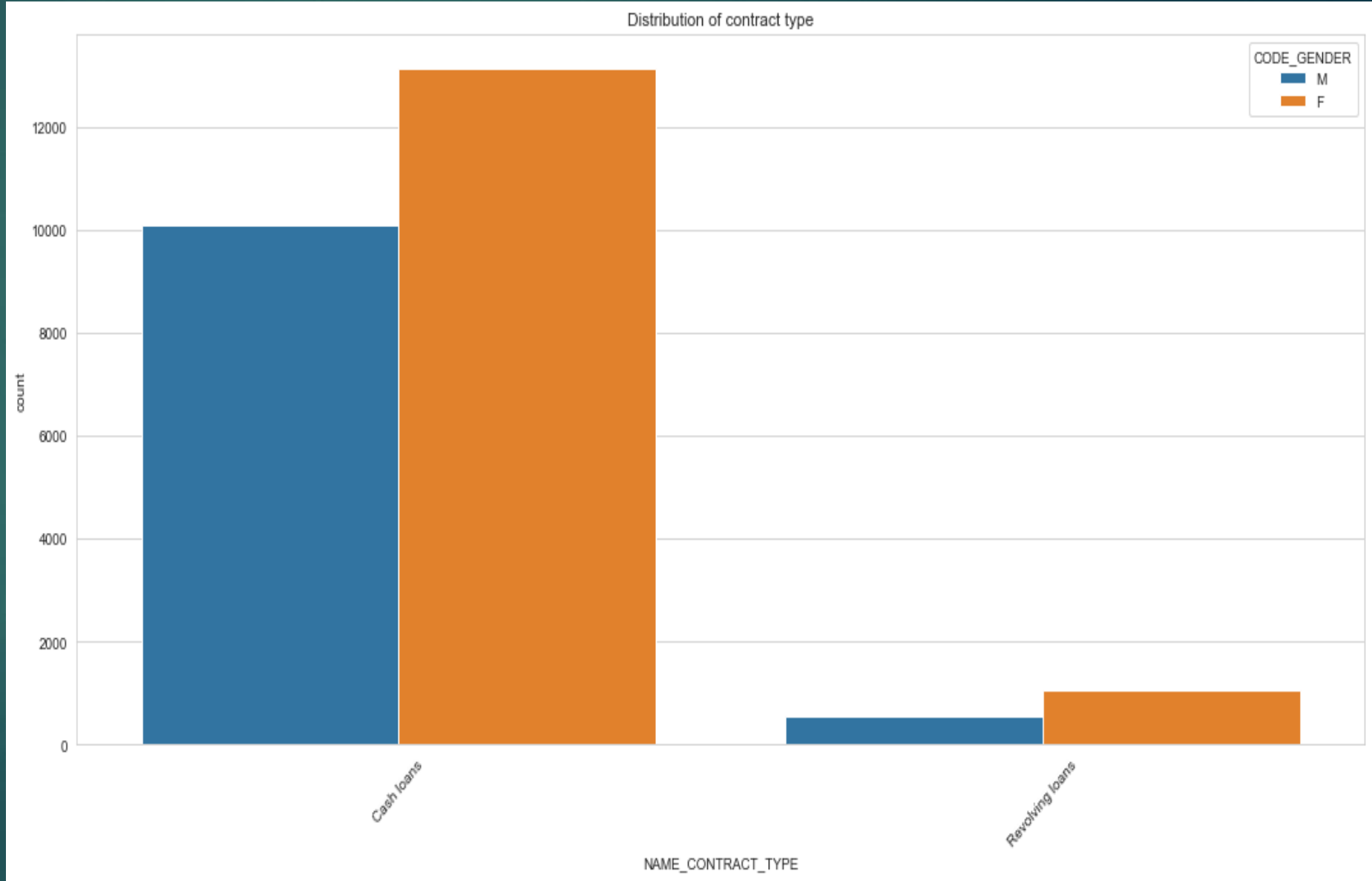
3. Less number of credits for income type Maternity leave and 'Unemployed'.



Distribution of Contract Type

1. Contract type cash loans is have higher number of credits than Revolving loans.

2. Here also Female is leading for applying credits.

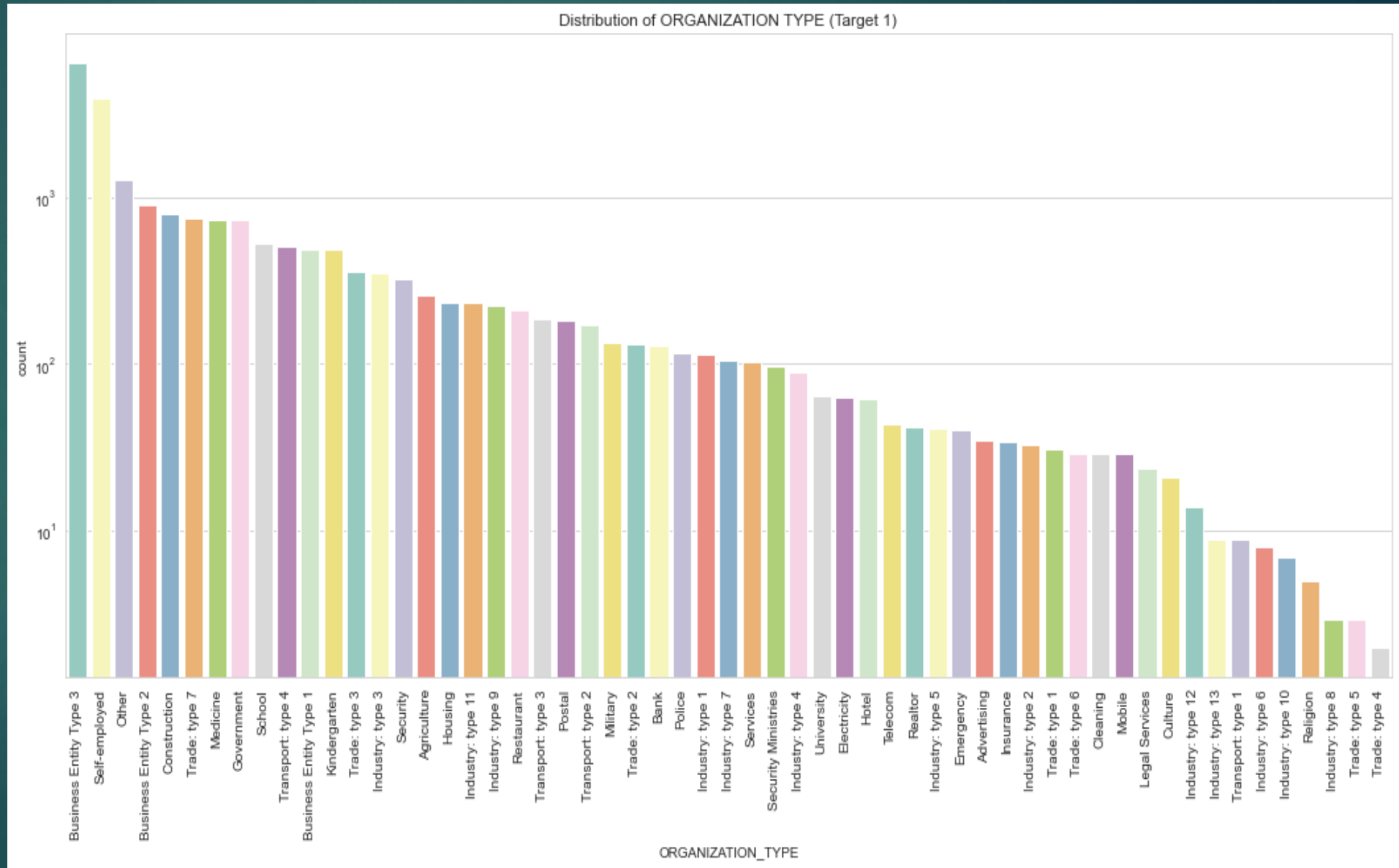


Distribution of Organization Type

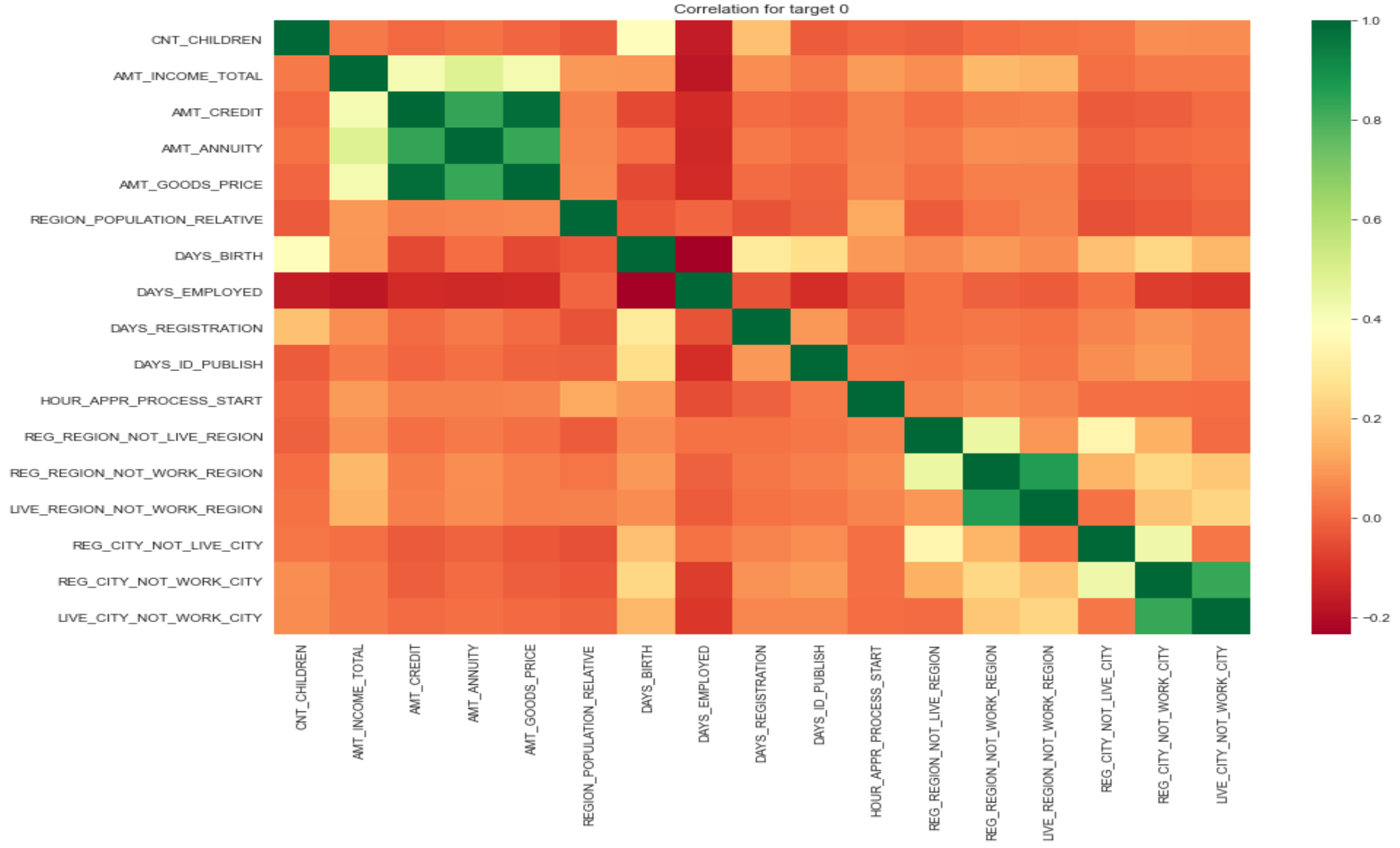
1. Clients which have applied for credits are mostly from the organization type Business entity Type 3 , Self employed , Other , 'Business Entity Type 2', 'Construction'.

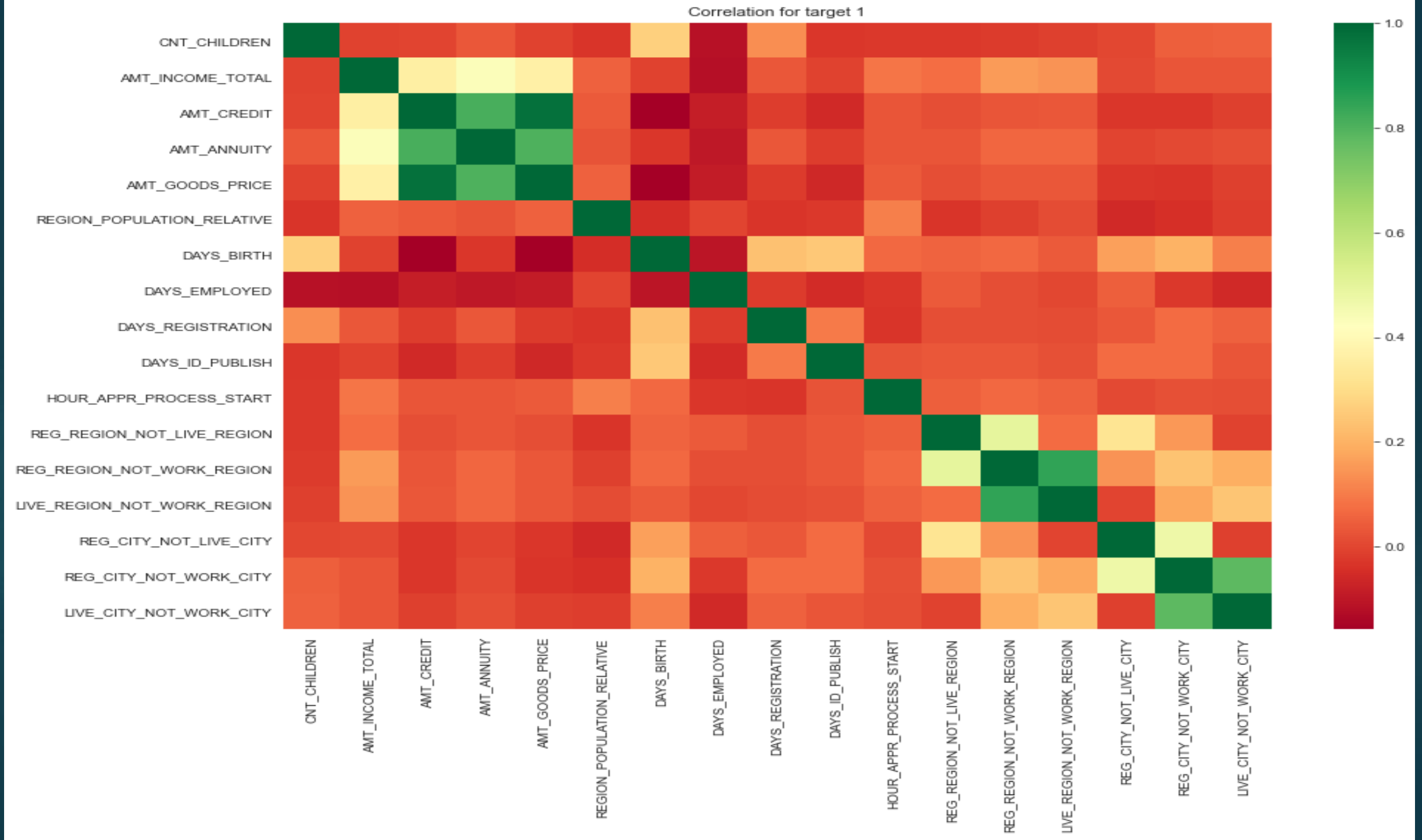
2. Less clients are from Trade type 4, type5, Industry type 8,type 10, type 6 and Religion.

3. Similar to distribution of target0.



Correlation for target0 (No Difficulties)


Correlation for target 0
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
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 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
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

Inferences from heatmap of target0 :

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

Inferences from heatmap of target1:

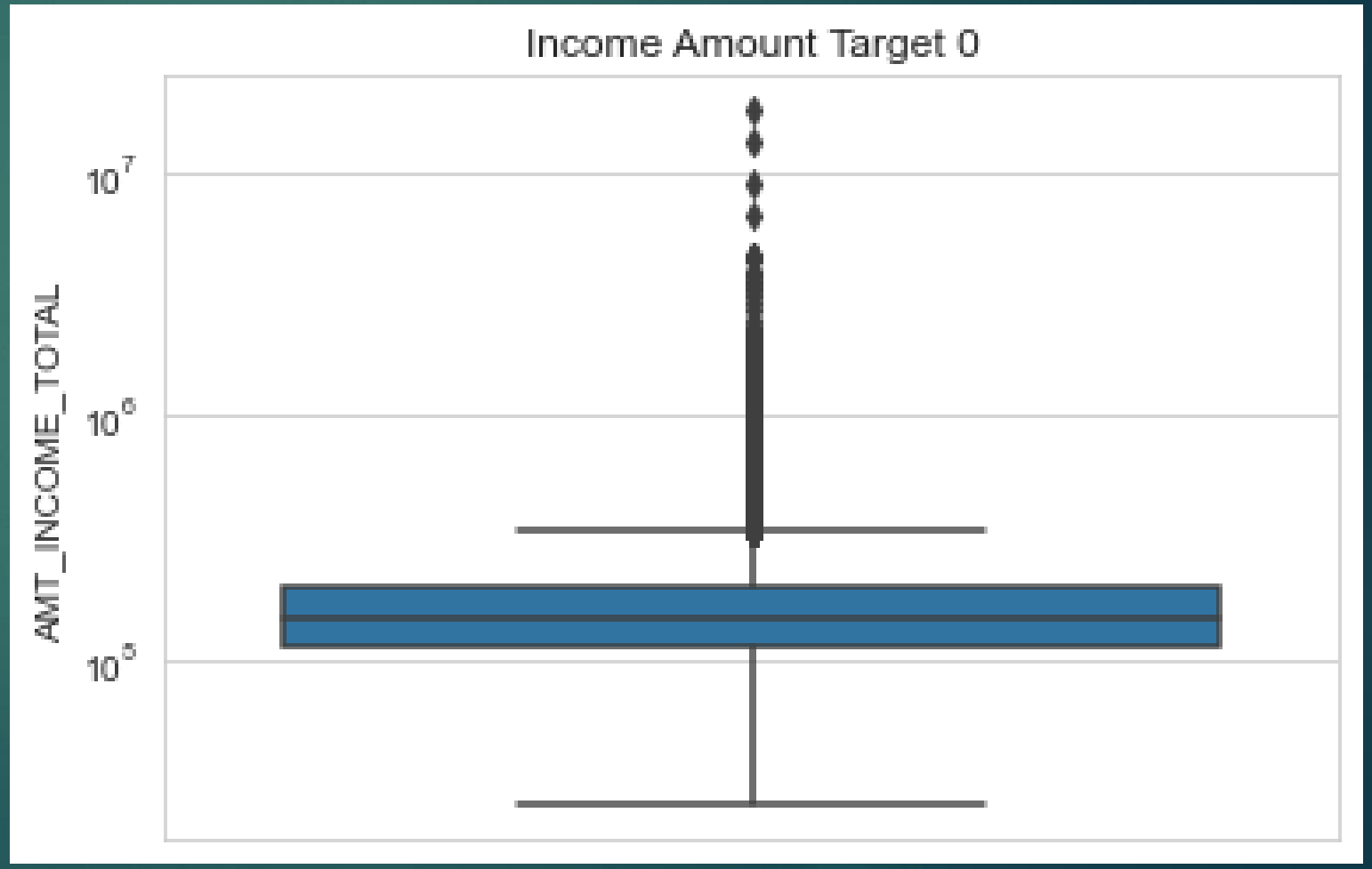
- The client's permanent address does not match contact address are having less children and vice-versa
- The client's permanent address does not match work address are having less children and vice-versa

Outlier detection for target0 variables

Boxplot of Total Income Amount

1. Some outliers are noticed in total income amount.

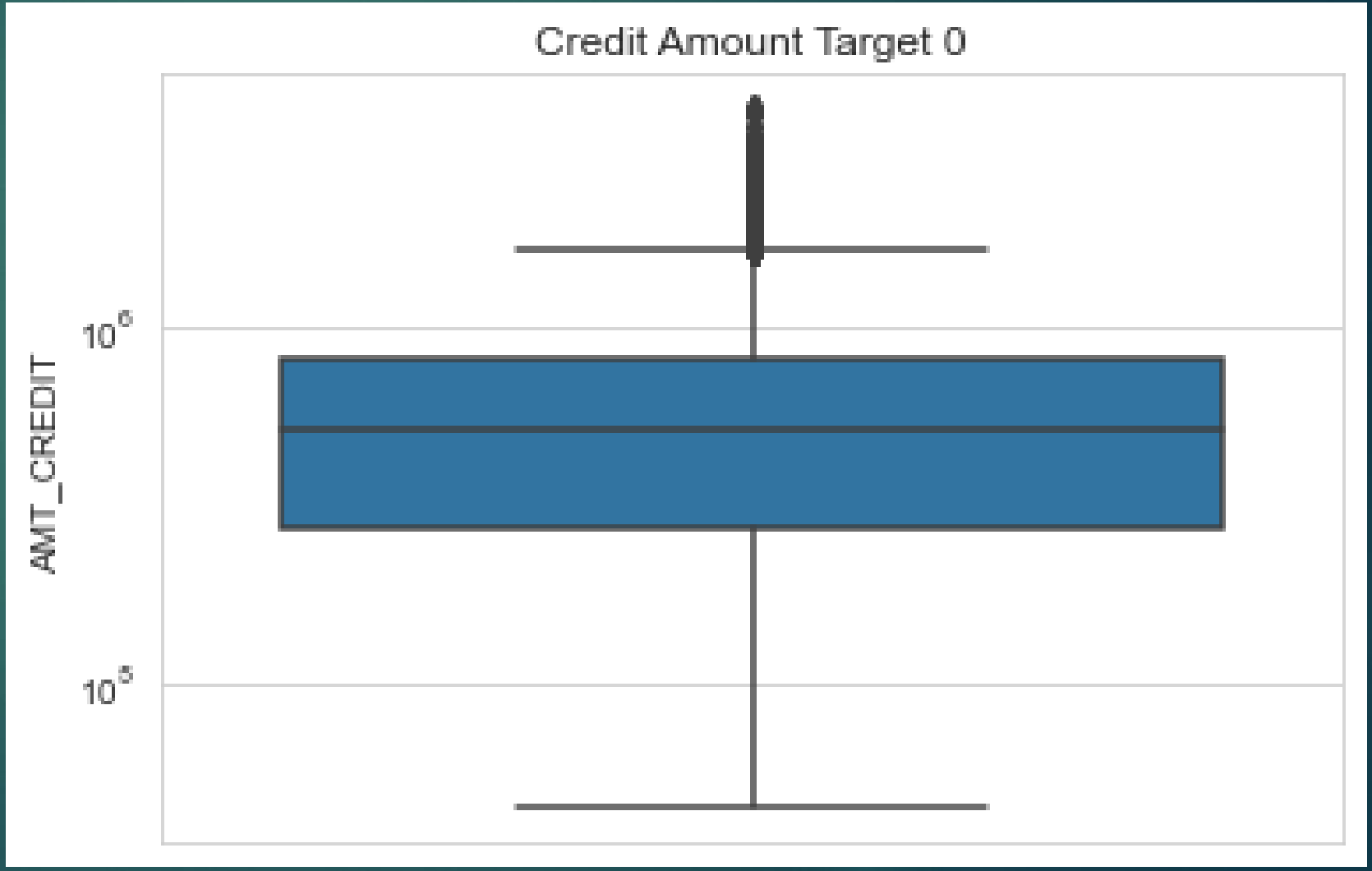
2. The third quartiles is very slim for income amount.



Boxplot of Credit Amount

1. Some outliers are noticed in credit amount.

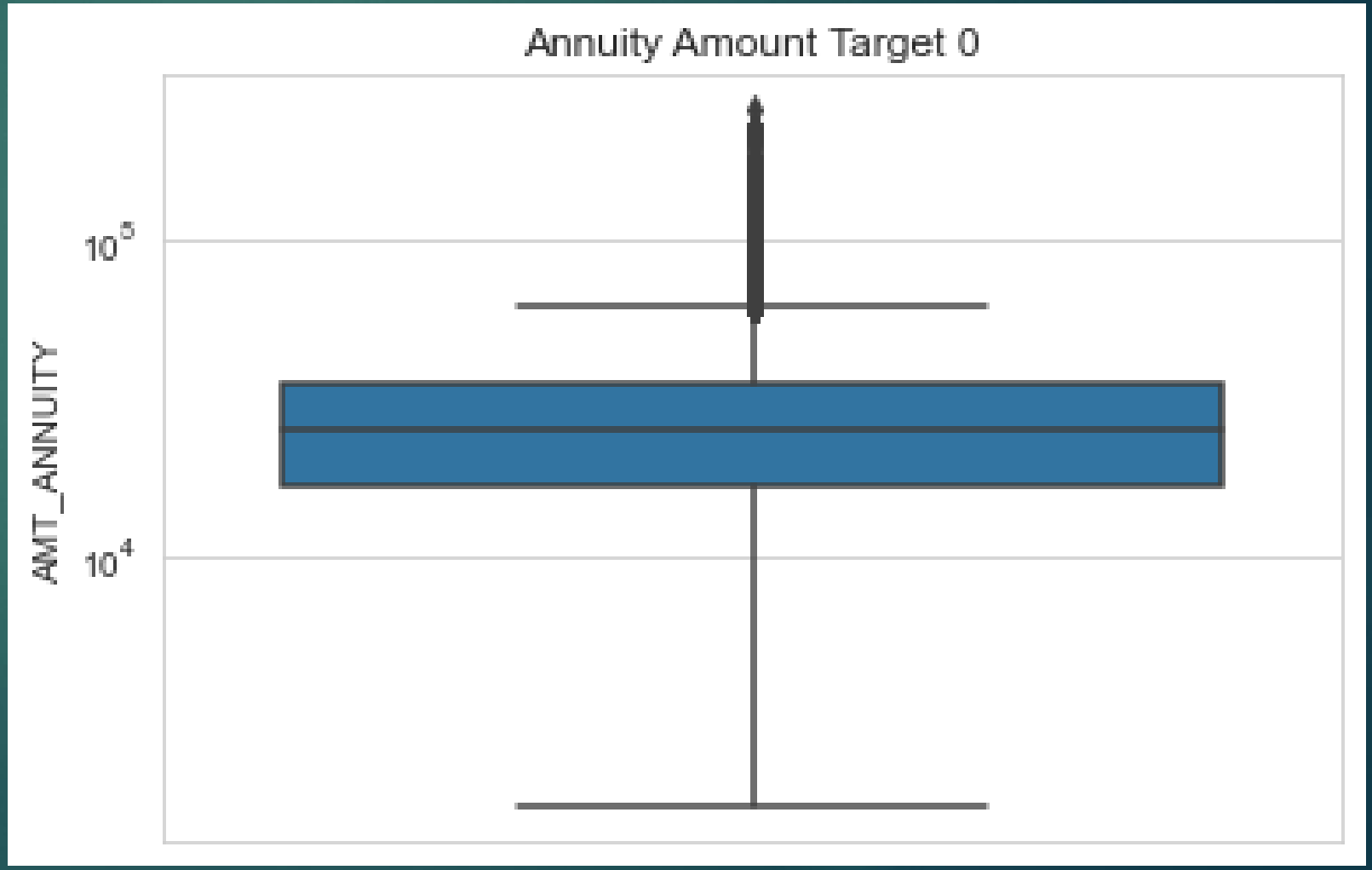
2. The first quartile is bigger than third quartile which means most of the credits of clients are present in the first quartile.



Boxplot of Annuity Amount

1. Some outliers are noticed in annuity amount.

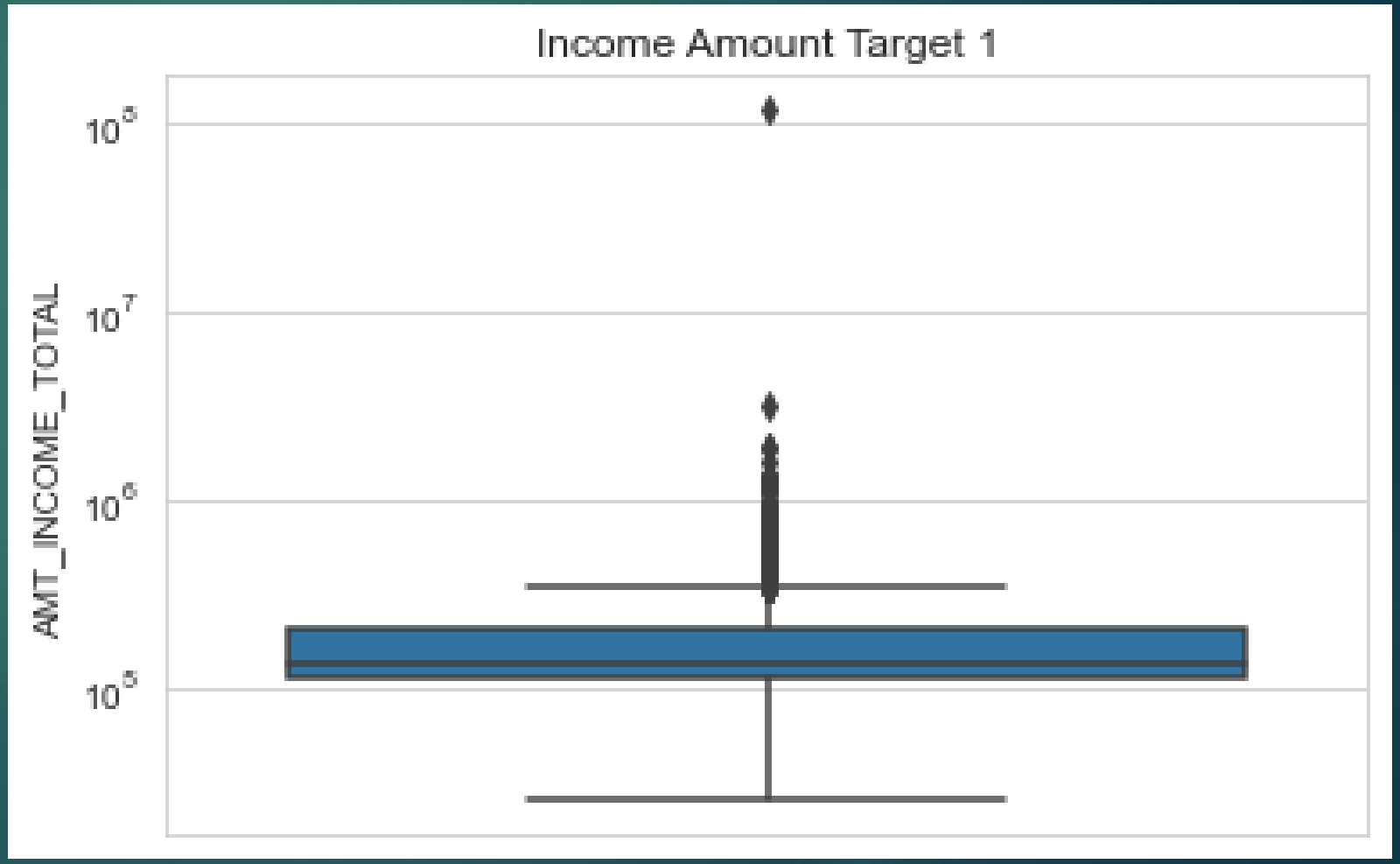
2. The first quartile is bigger than third quartile which means most of the annuity clients are from first quartile.



Outlier detection for target1 variables

Boxplot for Total Income Amount

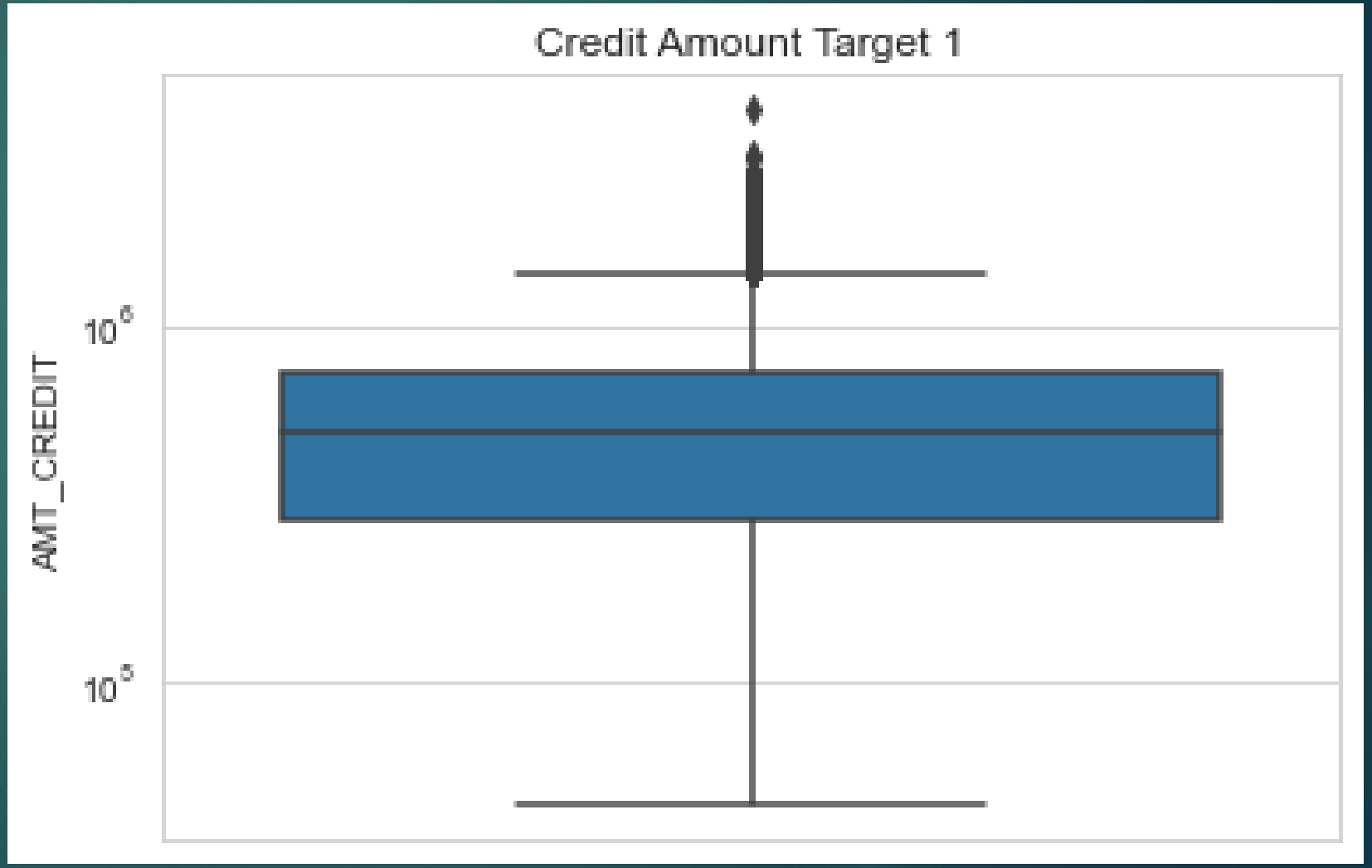
1. Some outliers are noticed in income amount.
2. The third quartiles is very slim for income amount.
3. Most of the clients of income are present in first quartile.



Boxplot of Credit Amount

1. Some outliers are noticed in credit amount.

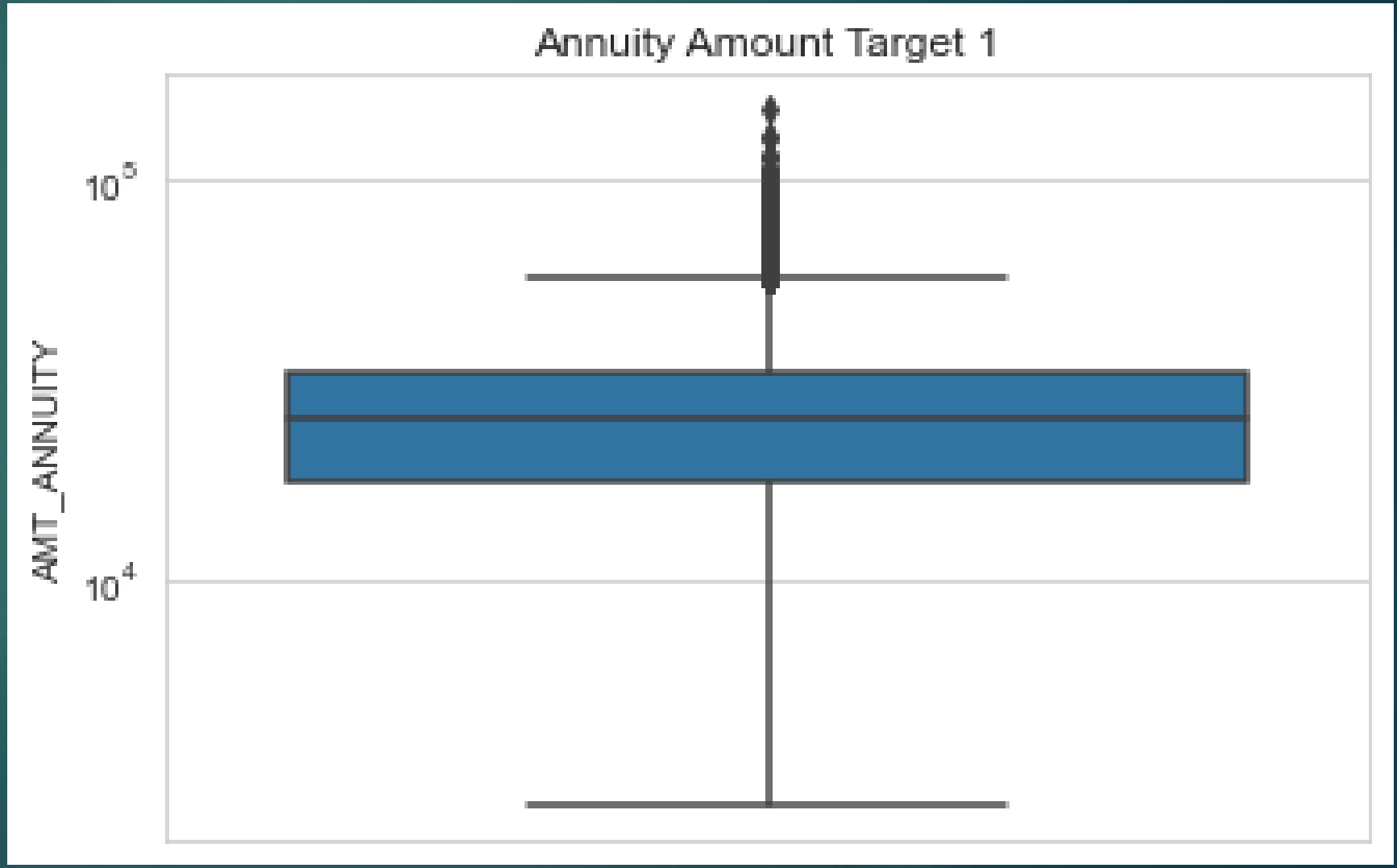
2. The first quartile is bigger than third quartile for credit amount which means most of the credits of clients are present in the first quartile.



Boxplot of Annuity Amount

1. Some outliers are noticed in annuity amount.

2. The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.



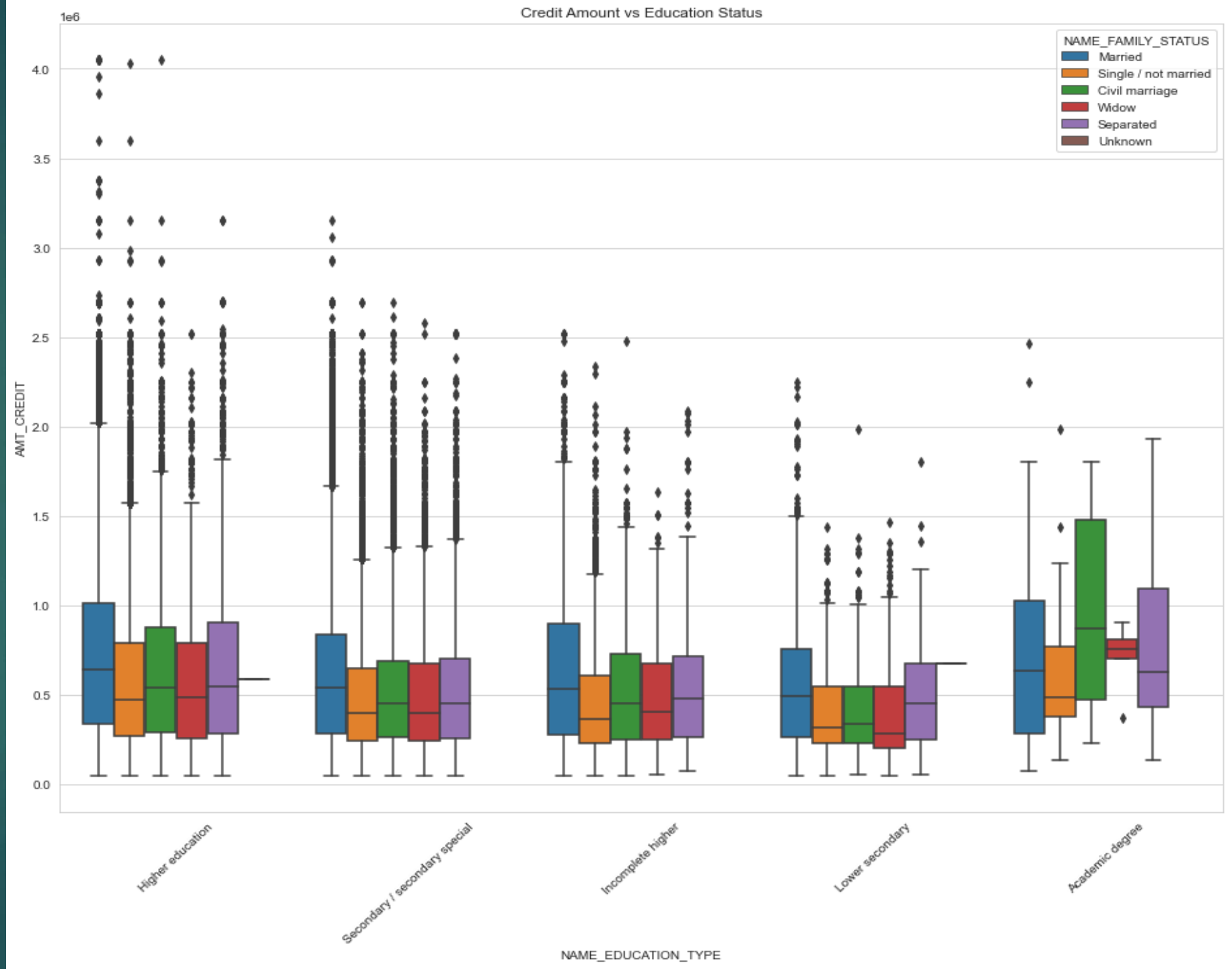
Bivariate analysis of target0

Boxplot analysis of Credit Amount and Education Status

1. In Higher education - 'marriage', 'single' and 'civil marriage' are having more outliers.

2. In Academic degree - 'civil marriage', 'marriage' and 'separated' are having higher number of credits than others.

3. Civil marriage in Academic degree have most of the credits in the third quartile.



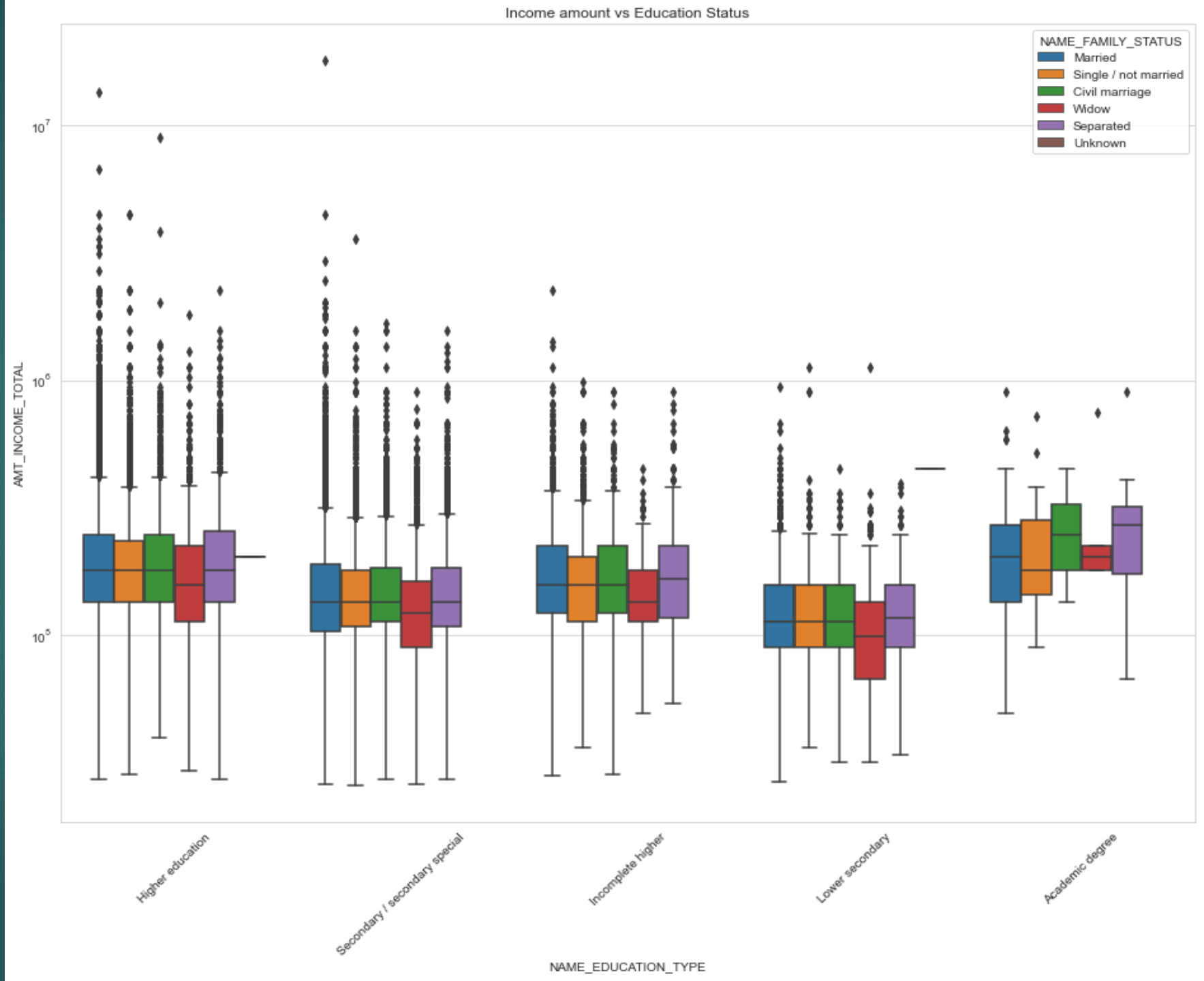
Boxplot of Income Amount and Education Status

1. In Higher education the Income amount is mostly equal with each family status

2. Income amount for Academic degree holders is higher than Higher education holders.

3. Less outliers in Academic degree and most outliers in Higher education and Secondary education.

4. Lowest Income amount for Lower secondary holders.



Bivariate analysis of target1

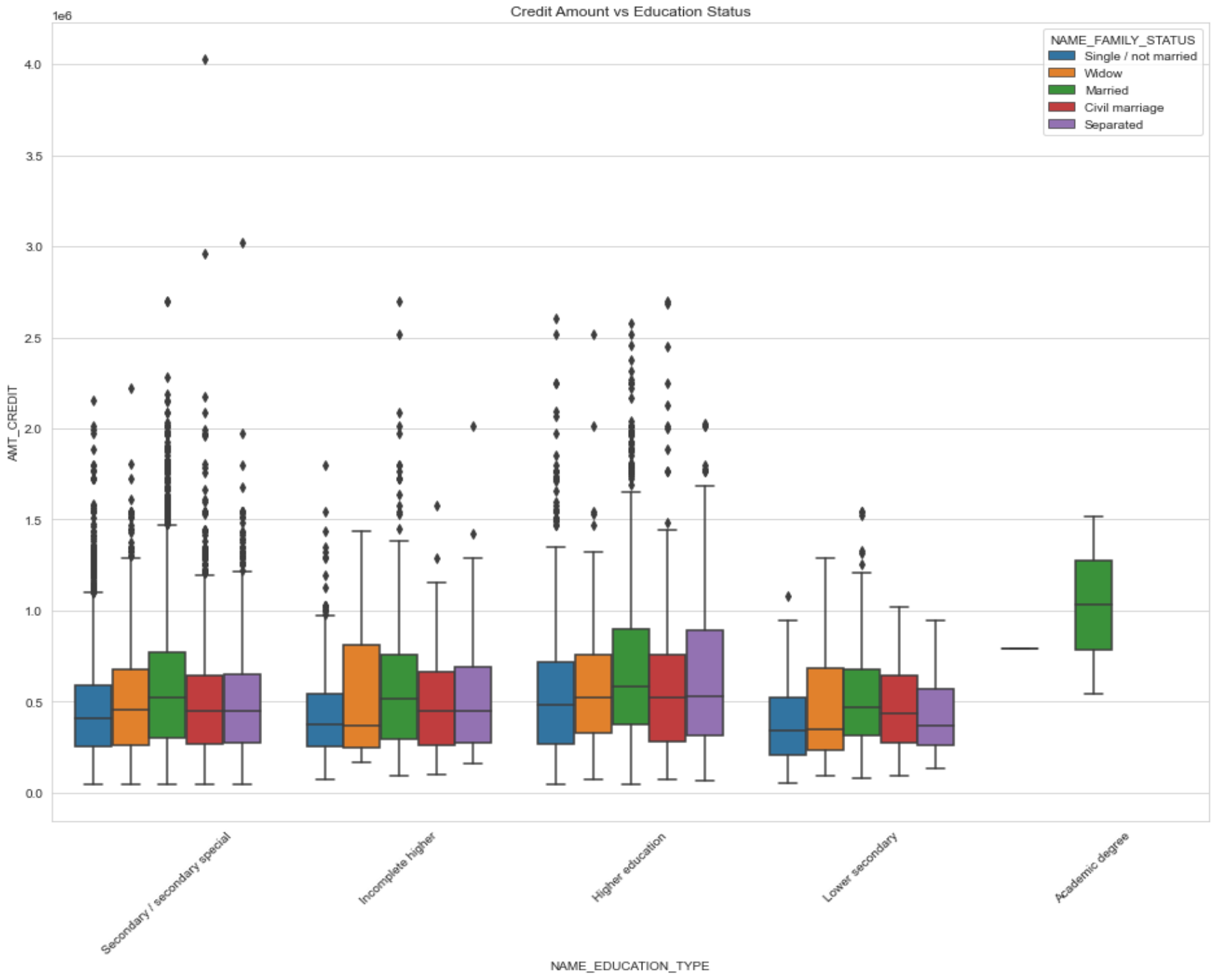
Boxplot analysis of Credit Amount and Education Status

1. Academic degree holders seems to have higher credit amount, though other family status is not available.

2. More outliers can be seen in Secondary education holders.

3. 'Separated' and 'Married' clients have more credit amount in third quartile.

4. Lower secondary holders have less credit amount comparing to others.

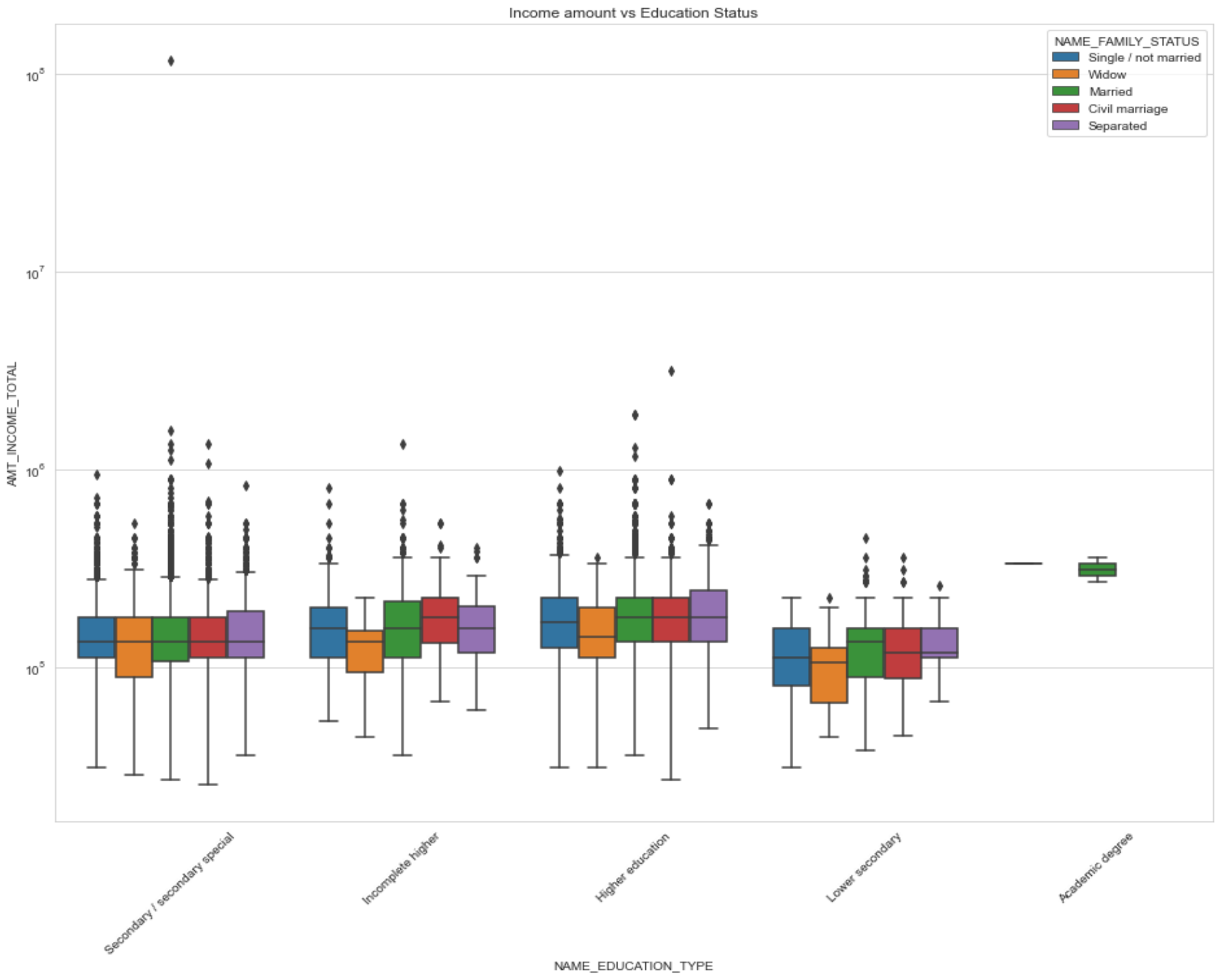


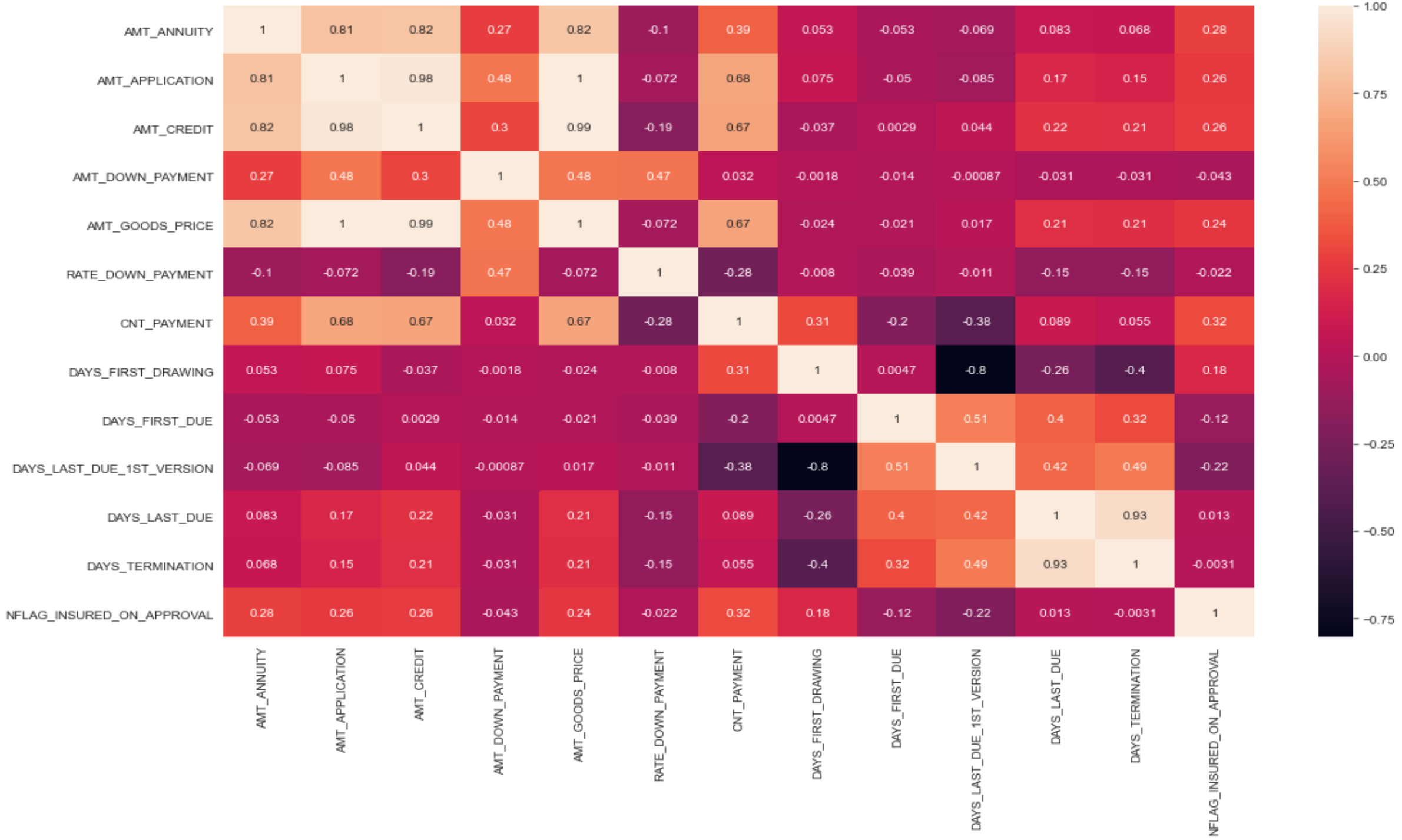
Boxplot of Income Amount and Education Status

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

2. Less outlier are having for Academic degree but there income amount is little higher that Higher education.

3. Lower secondary are have less income amount than others




AMT_ANNUITY 1 0.81 0.82 0.27 0.82 -0.1 0.39 0.053 -0.053 -0.069 0.083 0.068 0.28
AMT_APPLICATION 0.81 1 0.98 0.48 1 -0.072 0.68 0.075 -0.05 -0.085 0.17 0.15 0.26
AMT_CREDIT 0.82 0.98 1 0.3 0.99 -0.19 0.67 -0.037 0.0029 0.044 0.22 0.21 0.26
AMT_DOWN_PAYMENT 0.27 0.48 0.3 1 0.48 0.47 0.032 -0.0018 -0.014 -0.00087 -0.031 -0.031 -0.043
AMT_GOODS_PRICE 0.82 1 0.99 0.48 1 -0.072 0.67 -0.024 -0.021 0.017 0.21 0.21 0.24
RATE_DOWN_PAYMENT -0.1 -0.072 -0.19 0.47 -0.072 1 -0.28 -0.008 -0.039 -0.011 -0.15 -0.15 -0.022
CNT_PAYMENT 0.39 0.68 0.67 0.032 0.67 -0.28 1 0.31 -0.2 -0.38 0.089 0.055 0.32
DAYS_FIRST_DRAWING 0.053 0.075 -0.037 -0.0018 -0.024 -0.008 0.31 1 0.0047 -0.8 -0.26 -0.4 0.18
DAYS_FIRST_DUE -0.053 -0.05 0.0029 -0.014 -0.021 -0.039 -0.2 0.0047 1 0.51 0.4 0.32 -0.12
DAYS_LAST_DUE_1ST_VERSION -0.069 -0.085 0.044 -0.00087 0.017 -0.011 -0.38 -0.8 0.51 1 0.42 0.49 -0.22
DAYS_LAST_DUE 0.083 0.17 0.22 -0.031 0.21 -0.15 0.089 -0.26 0.4 0.42 1 0.93 0.013
DAYS_TERMINATION 0.068 0.15 0.21 -0.031 0.21 -0.15 0.055 -0.4 0.32 0.49 0.93 1 -0.0031
NFLAG_INSURED_ON_APPROVAL 0.28 0.26 0.26 -0.043 0.24 -0.022 0.32 0.18 -0.12 -0.22 0.013 -0.0031 1
AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_DOWN_PAYMENT
AMT_GOODS_PRICE
RATE_DOWN_PAYMENT
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

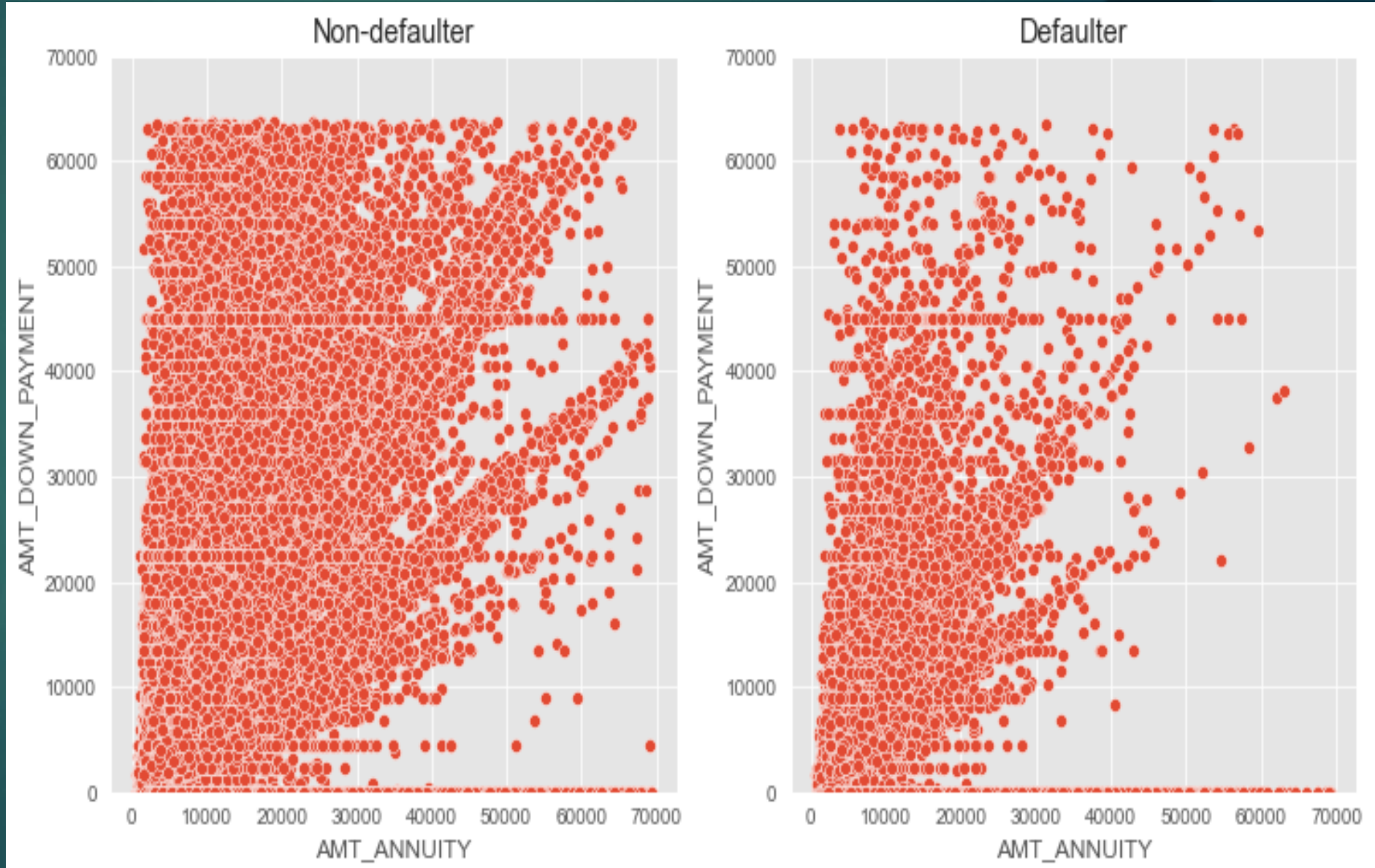
Inferences from before given heatmap:

- 'DAYS_LAST_DUE' and 'DAYS_TERMINATION' are highly correlated because they are showing constant value 1 on the plot.
- 'DAYS_FIRST_DRAWING' and 'DAYS_LAST_DUE_1st_VERSION' have high negative correlation.
- 'AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE' are highly correlated.

Scatterplot of Down Payment and Annuity amount for defaulter v/s non-defaulter

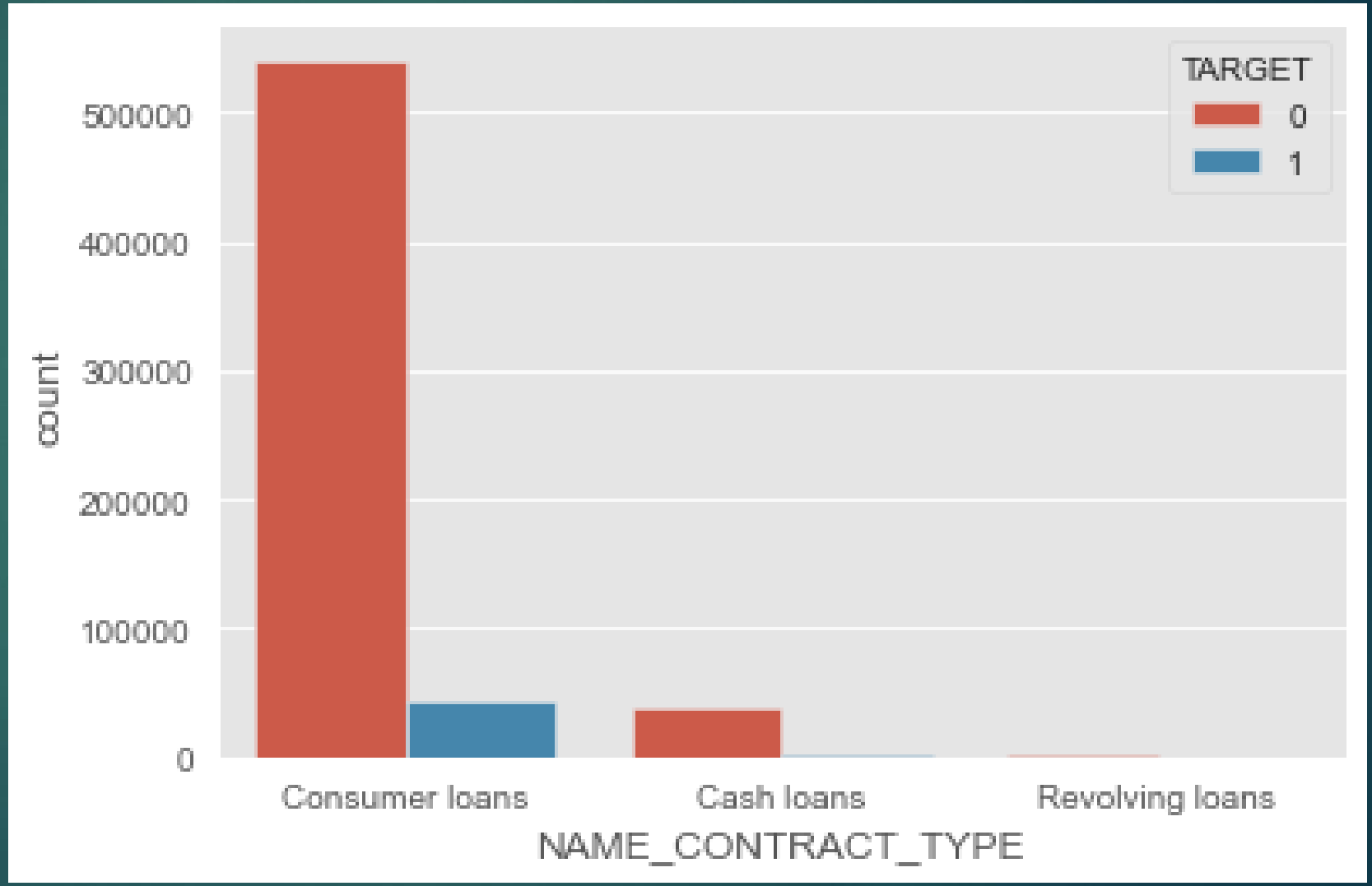
1. If you see the data insights for down payment , the defaulter cases are much less.

2. Number of defaulters are less found in previous application data for larger amount of annuity.



Count plot of Contract Type

- Highest number of loans are applied for Consumer Loans.



Count plot of Rejection Reasons

- As you can see in the above plot, 'SCO', 'LIMIT' and 'HC' are the most common reason of rejection for the loan application.

Now we will discuss the data collected from different value wise defaulter as shown in the notebook

- From name goods category we can see that Highest percentage of default cases are for the applicants who previously applied for Insurance and Vehicles.
- Similarly from name portfolio we can say that Cards defaulter rate is highest.
- From name product type all the walk-in applicants 9% defaulted in current loan.
- 15% loan application defaulted for AP+ (Cash Loan) this insight is from channel type.
- From seller industry we can say that In seller Industry "Auto technology" has highest rate of defaulter and MLM partners has lowest number of defaulters
- Defaulter percentage is highest where NAME_YIELD_GROUP is not known.
- Finally we found that highest percentage of default cases is for Card Street.

Conclusion

Insights about given dataset.

- There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.
- Feature columns with 50% or more missing data can be dropped.

Recommended step for given dataset.

- Following columns should be converted to integer. DAYS_FIRST_DRAWING float64 DAYS_FIRST_DUE float64 DAYS_LAST_DUE_1ST_VERSION float64 DAYS_LAST_DUE float64 DAYS_TERMINATION float64.
- We can covert this NFLAG_INSURED_ON_APPROVAL float64 column into integer column because it contains only 0 and 1.

Details of different value wise defaulters and important features.

- We can see that 7% of the previously approved loan applicants that defaulted in current loan.
- Total 90 % of the previously refused loan applicants that were able to pay current loan.
- the most common reason of rejection are 'SCO', 'LIMIT' and 'HC'.
- Most of the people did not request insurance during previous loan application.
- For "Cards" defaulter percentage is highest (17%). 'NAME_PORTFOLIO' is an important feature for analyzing 'TARGET' variable.
- 15% loan application defaulted for AP+ (Cash Loan). 'CHANNEL_TYPE' is an important feature for analyzing 'TARGET' variable.
- Highest percentage (17%) of default cases is for 'Card Street'.