

Credit EDA Case Study

By-
Naveen B Gotur
DS C54



Problem Statement

- **Background**

- When a consumer applies for a loan from a loan providing company that lends money to urban customers, processes loan application by verifying the applicant's ability to repay the loan.

- **Business Objective**

- This case study tries to find trends that show whether a client has trouble making their payments, which may be used to decide whether to grant the loan, reduce its size, charge riskier applicants a higher interest rate, etc.
- As a result, only those customers who can repay the loan won't be turned away. This case study seeks to identify such applicants by means of EDA.



Understanding the data

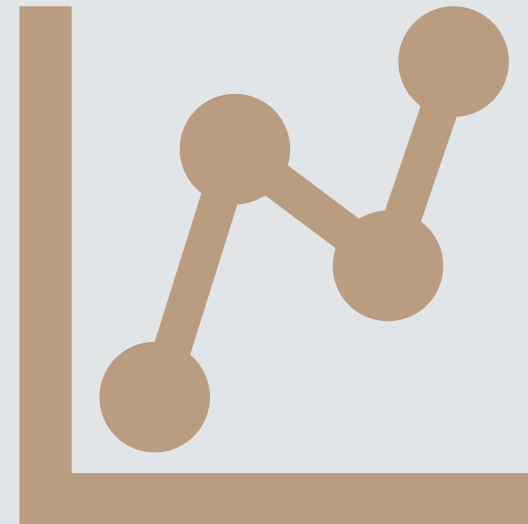
- **Missing value check**

- **For Application Dataset**

- Null values that make up more than 19% of the data should be removed since they can produce outliers and lead to incorrect analysis.

- **For Previous Application**

- Null values that make up more than 20% of the data should be removed since they can produce outliers and lead to incorrect analysis.



Cont...

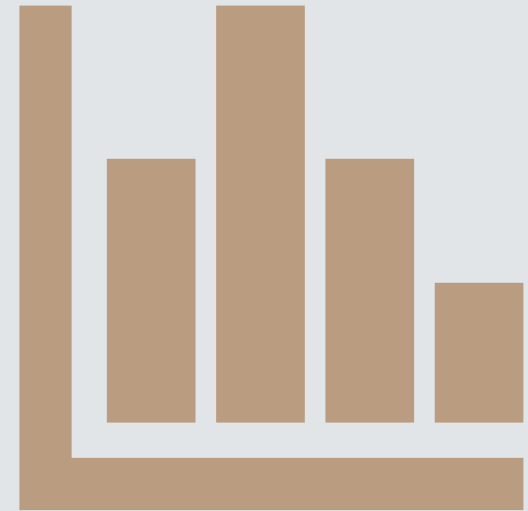
- **Imputing the missing values**

- **For Application Dataset**

- The dataset's mean, median, and mode values were used to impute values for the categorical and numerical columns.

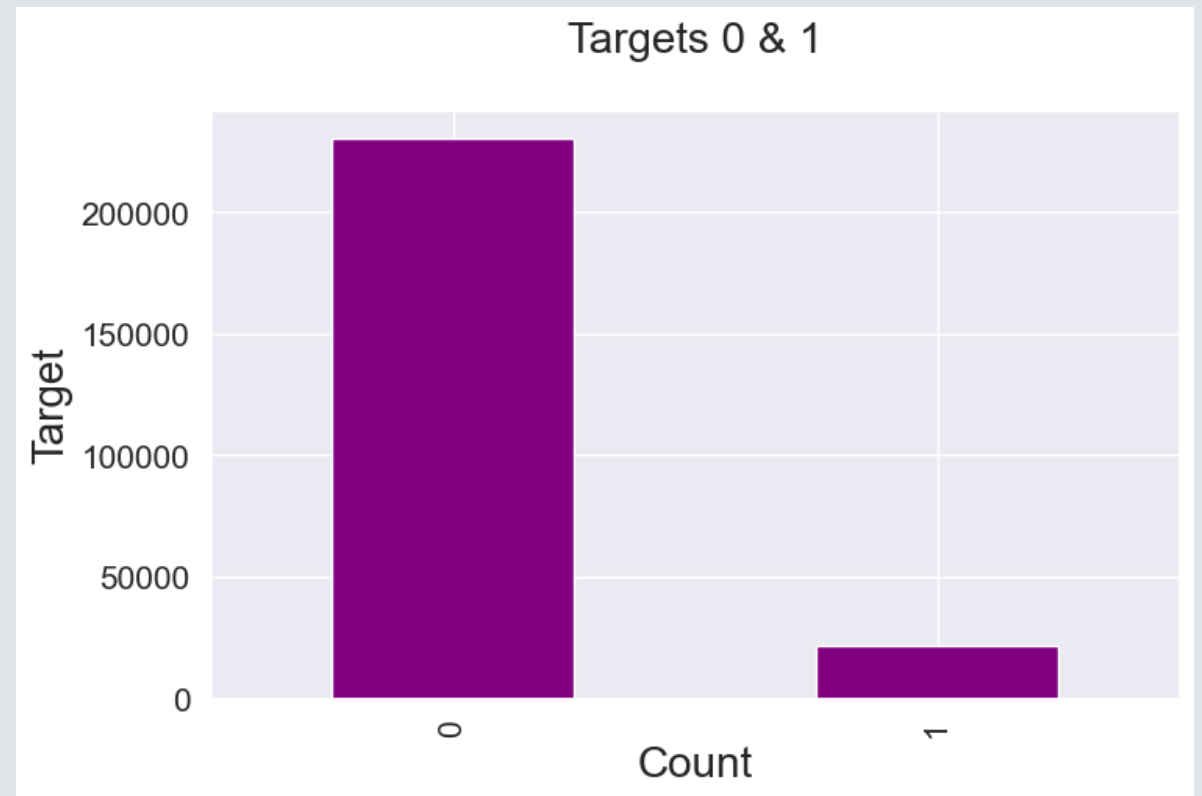
- **For Previous Application**

- There are no such values which needs to corrected.



Imbalance Ratio

After data cleaning, there is a significant imbalance between the data for loan defaulters (TARGET = 1) and the remaining data for non-defaulters (TARGET = 0).

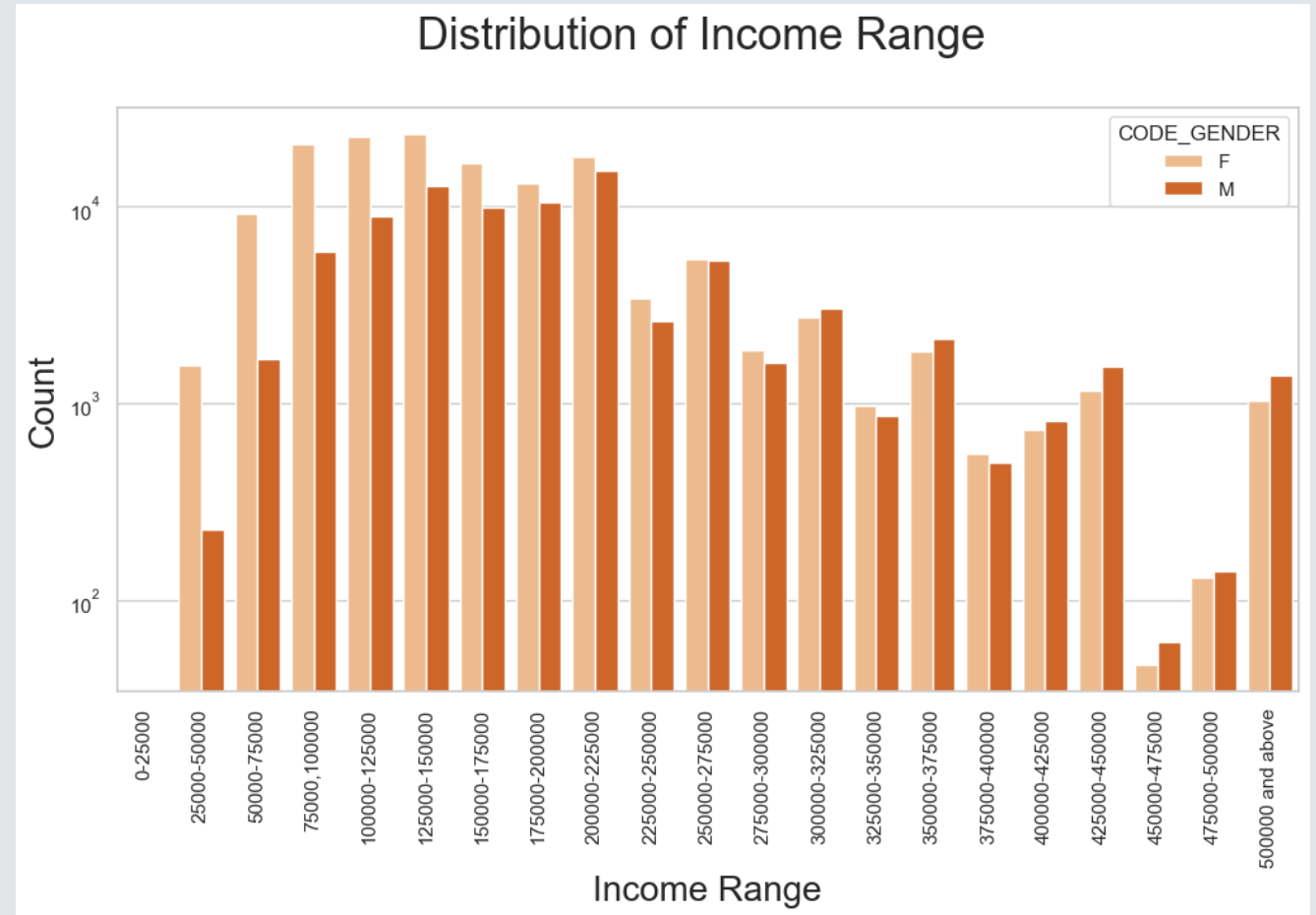




Categorical Univariate Analysis for Target 0

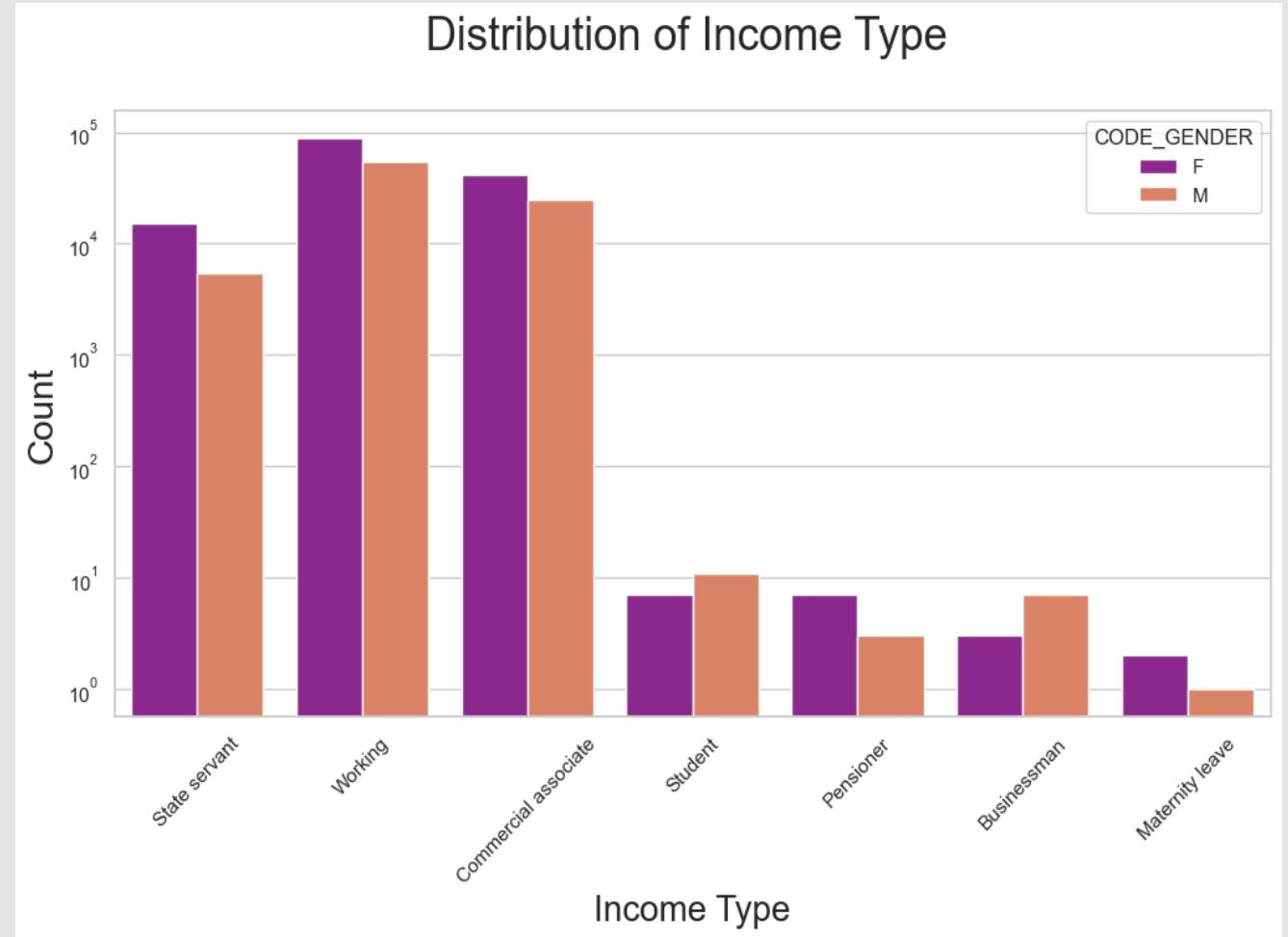
□ Conclusion from the graph:

- Income range from 125000 to 150000 is having the highest number of credits.
- Least number of credits from range 450000-475000.
- It seems that the females are more than male in having credit for range: 125000 to 150000.



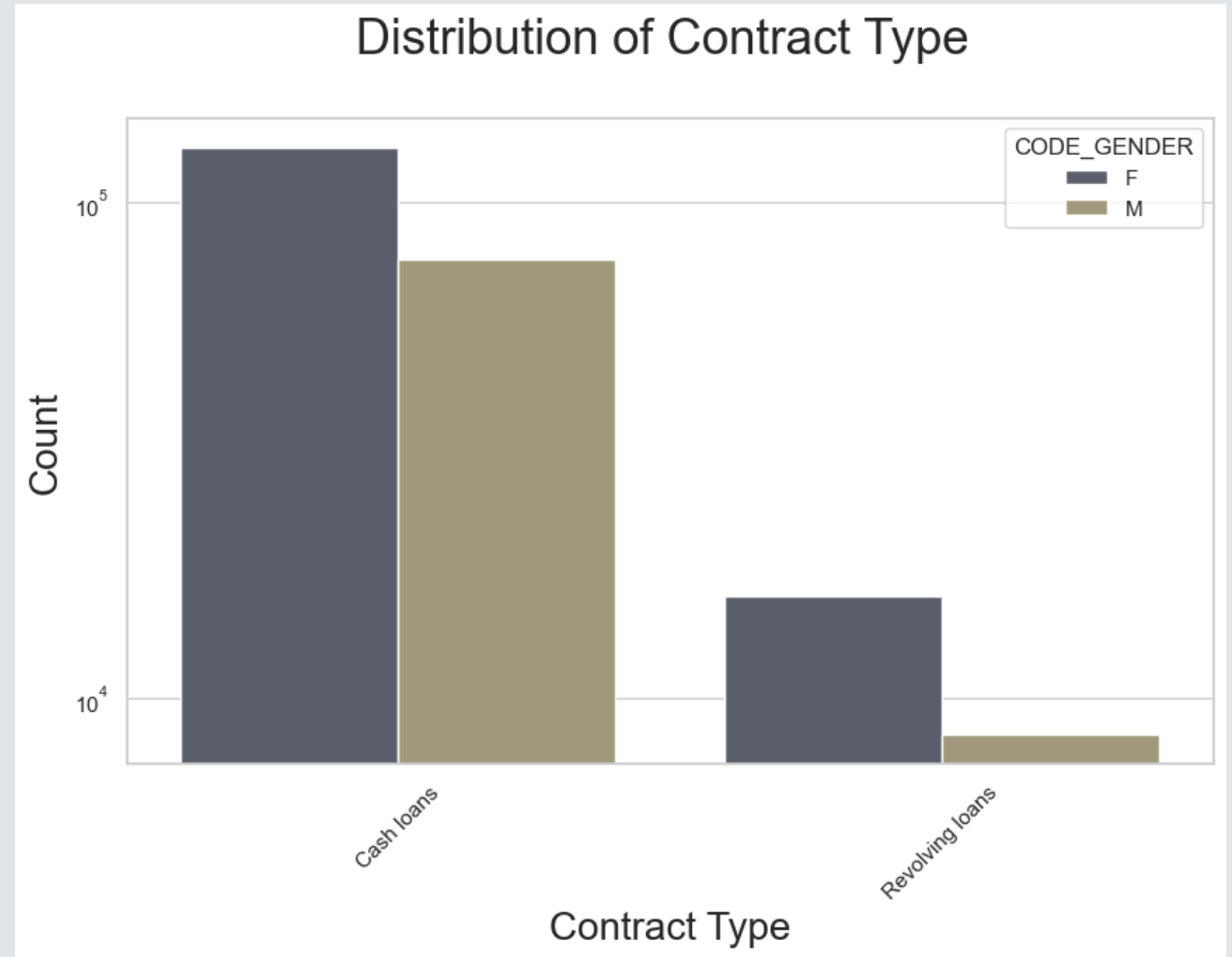
❑ Conclusion from the graph:

- 'Working' Women appear to have more credit than others.
- 'State Servant', 'Working' and 'Commercial Associate' appear to have higher credit counts than the others.
- Women in 'Maternity leave' has less credit in comparison to others.



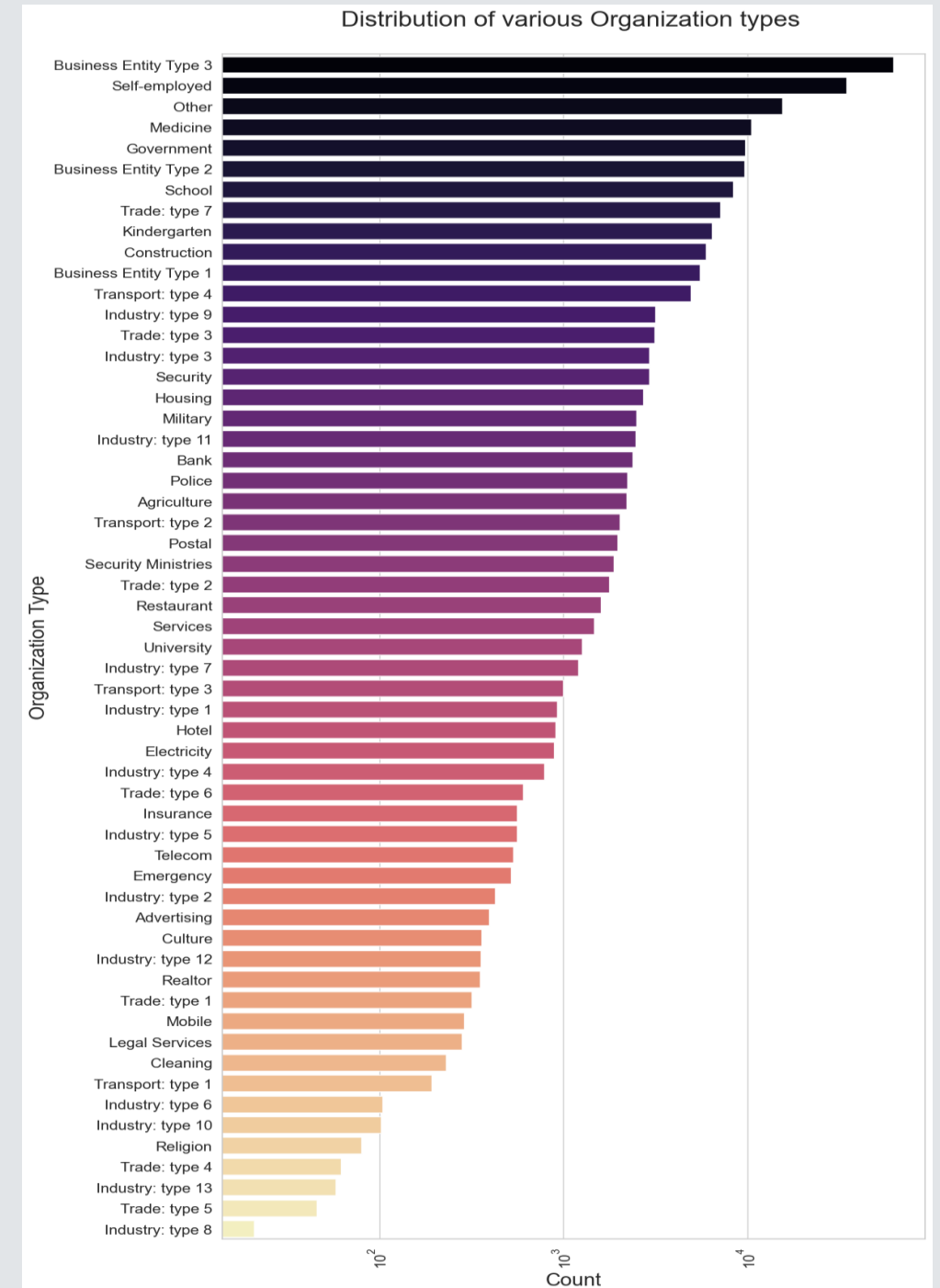
□ Conclusion from the graph:

- It appears that the contract type for 'cash loans' has more credits than the contract type for 'revolving loans'
- In addition, Female apply more for Credit.



□ Conclusion from the graph:

- The majority of the clients who have requested credit are from the organization types 'Business entity Type 3', 'Self employed', 'Other' and 'Medicine'.
- Less clients are from 'Industry type 6', 'Industry type 10', 'Industry type 13', 'Industry type 8', 'Religion', 'Trade type 4', 'Trade type 5'.

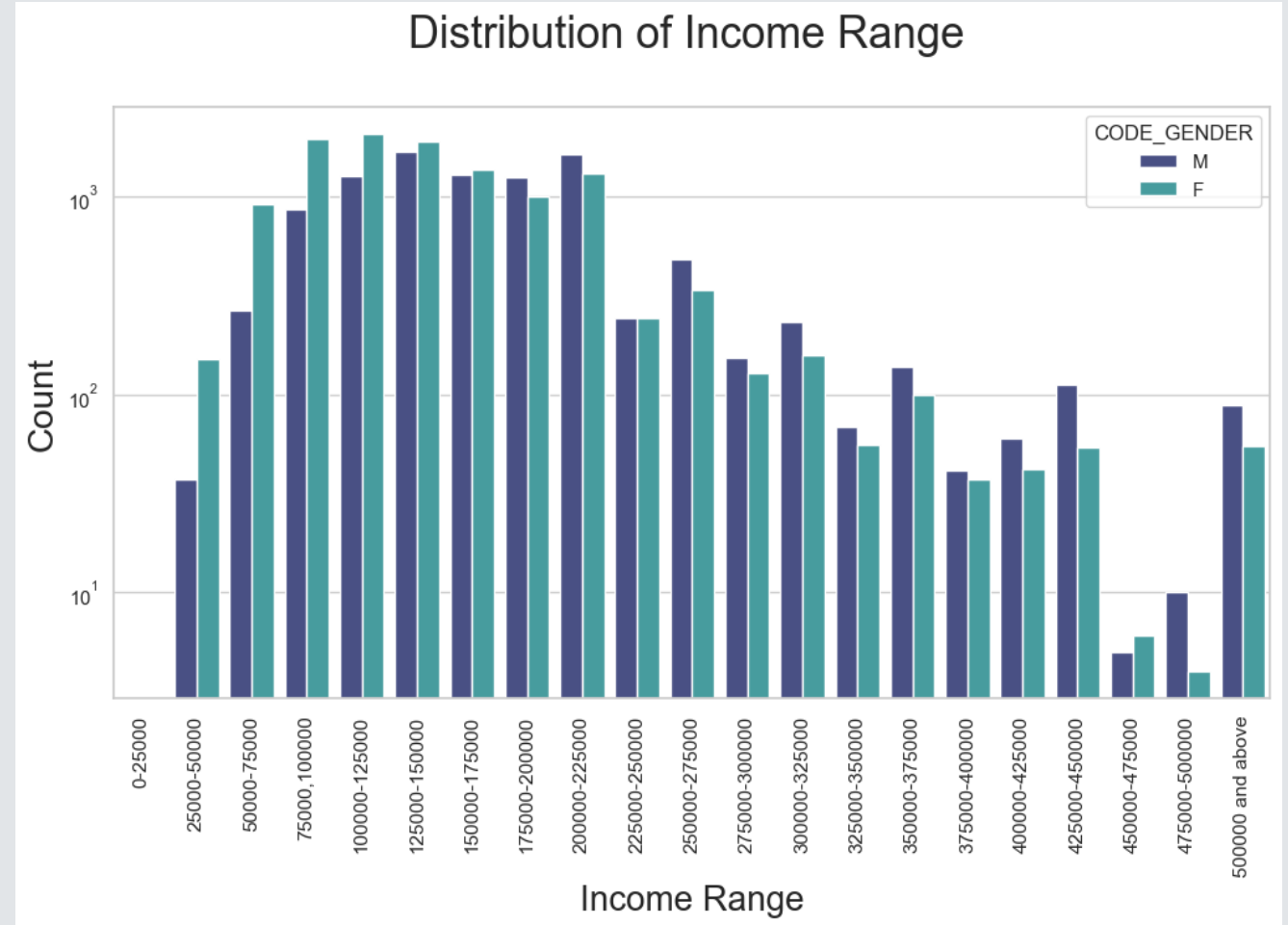




Categorical Univariate Analysis for Target 1

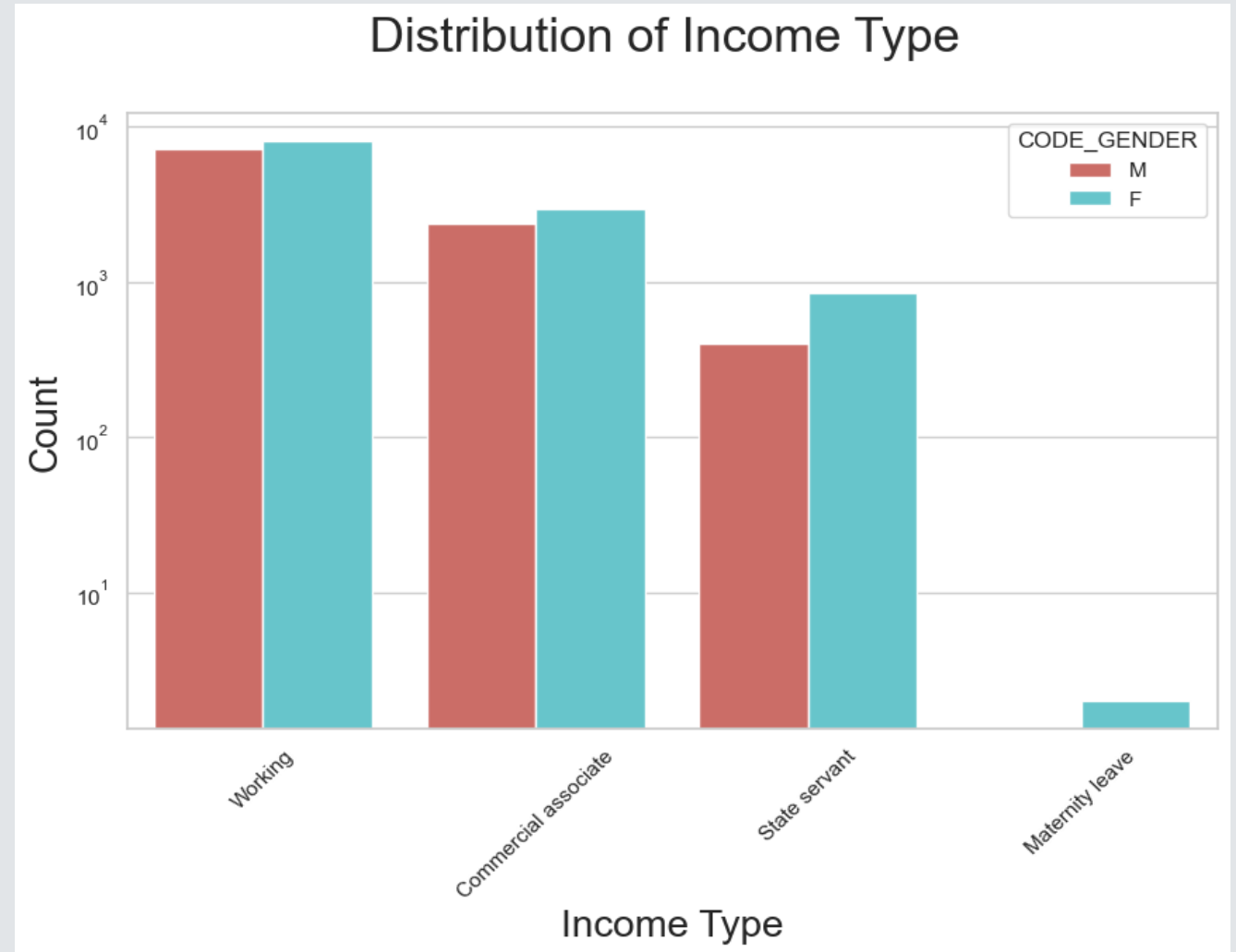
❑ Conclusion from the graph:

- The number of Male Counts is higher.
- Income range between 100000 to 200000 is having more number of credits as compared to others.
- Less count for income range between 450000-475000.



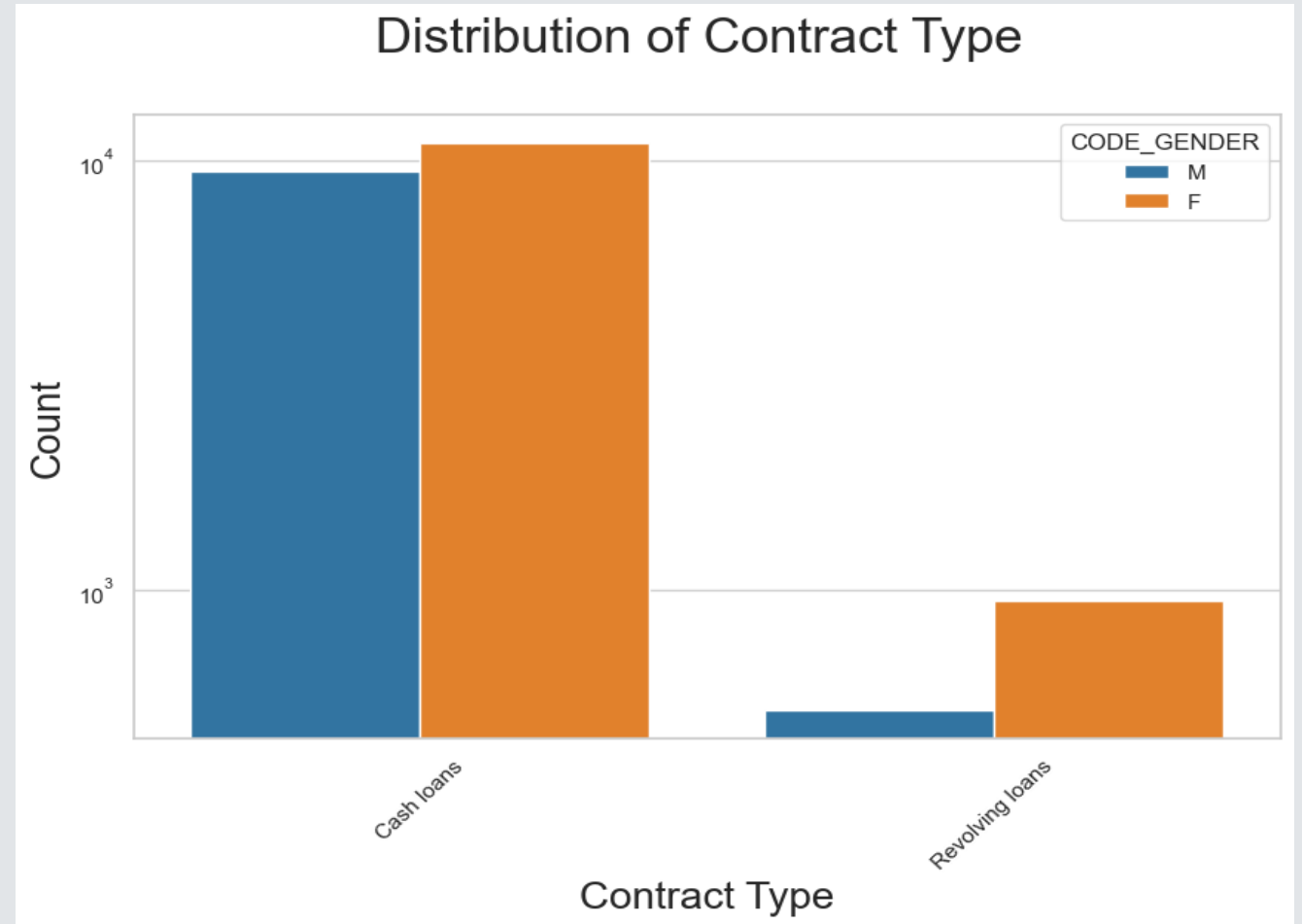
❑ Conclusion from the graph:

- The number of credits is higher for income types such as 'working', 'commercial associate', and 'State Servant' than other i.e. 'Maternity leave'.
- In this Females are having more number of credits than male.
- Income type 'Maternity leave' have less number of credits.



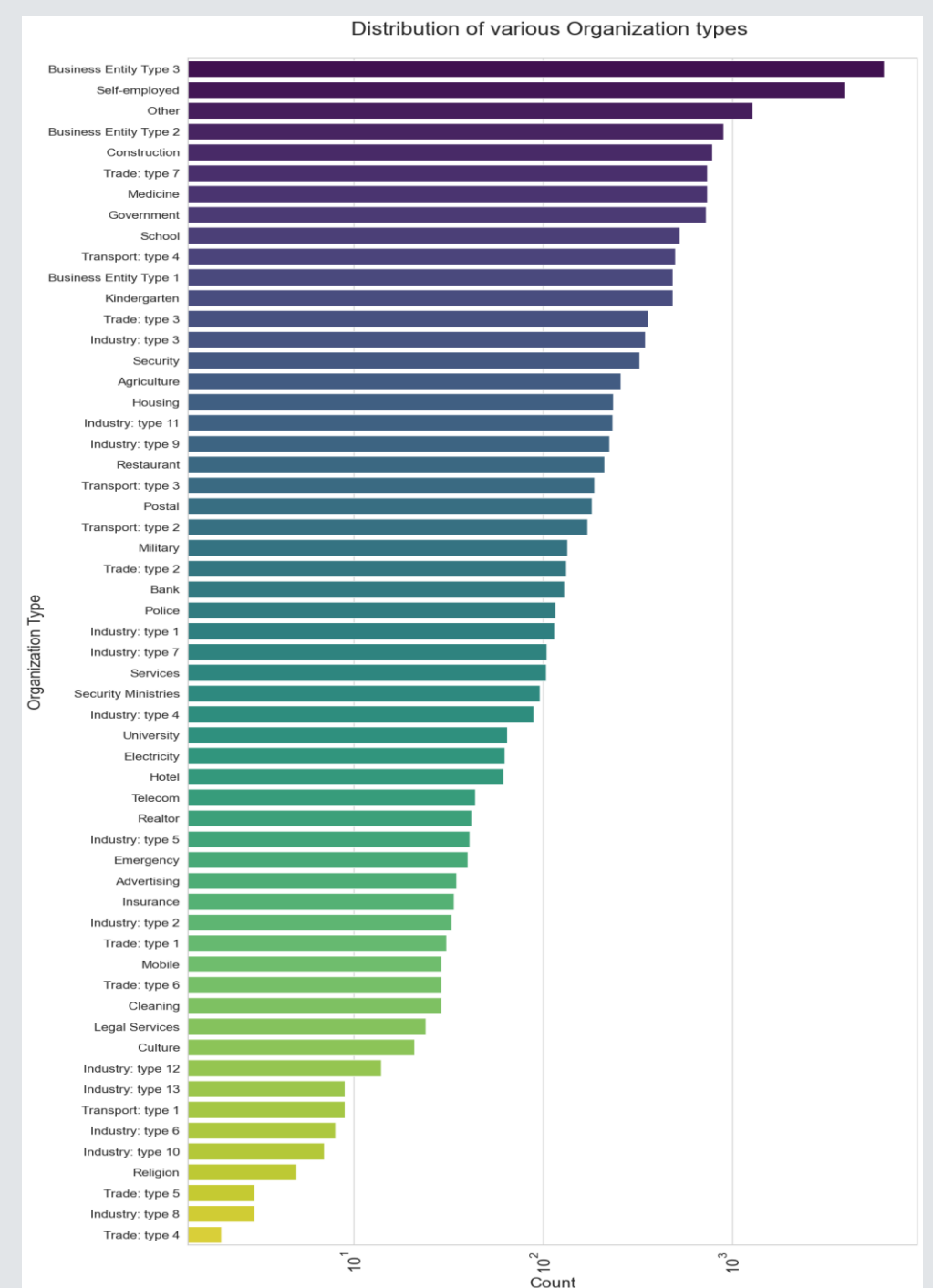
❑ Conclusion from the graph:

- 'Cash loans' contract type, have a greater number of credits than 'Revolving loans'.
- In addition, Females are leading for applying credits.

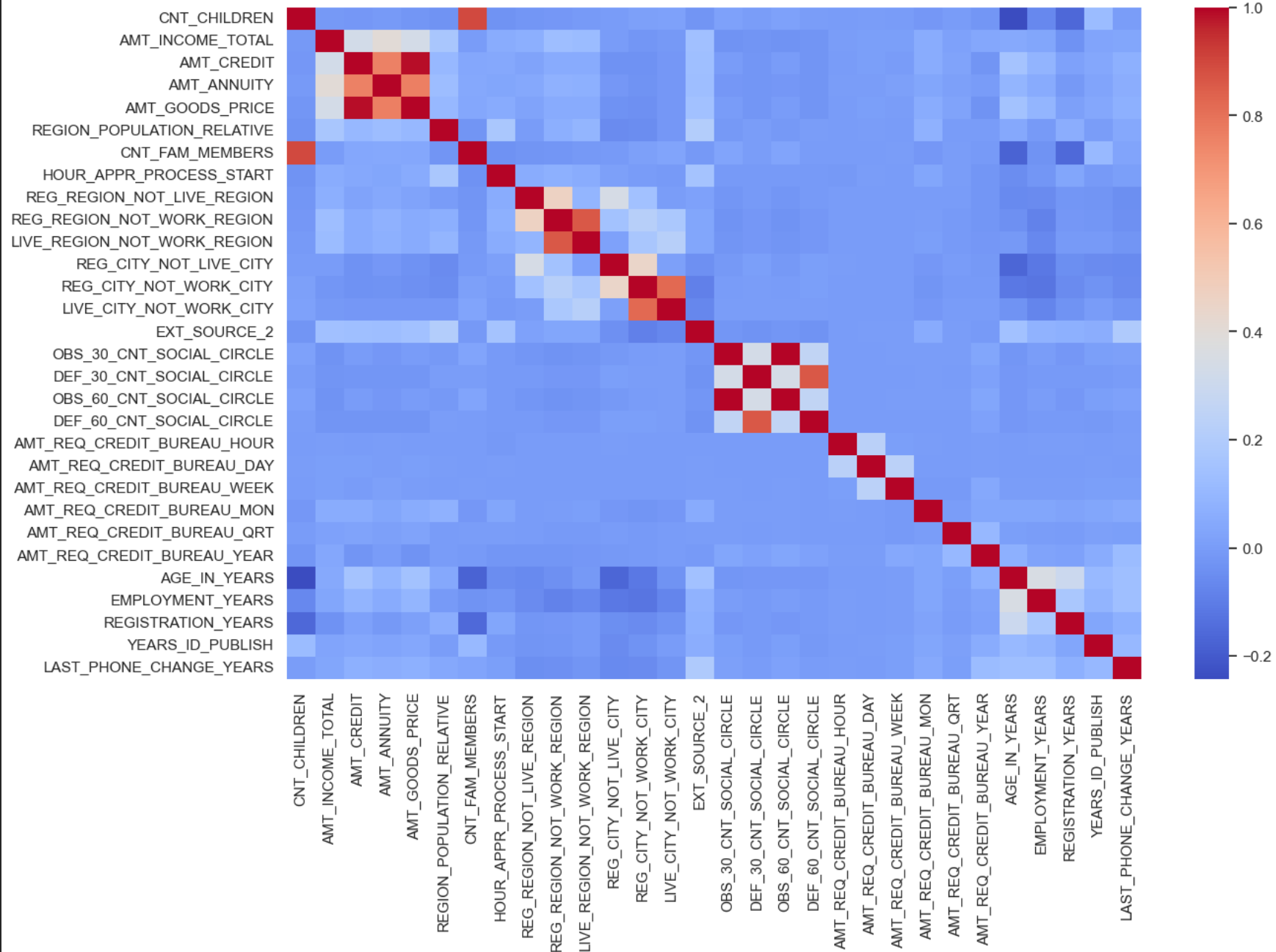


□ Conclusion from the graph:

- The majority of the clients who have requested credit are from the organisation types 'Business entity Type 3', 'Self employed', 'Other' and 'Medicine'.
- Less clients are from 'Industry type 6', 'Industry type 10', 'Industry type 8', 'Religion', 'Trade type 4', 'Trade type 5'.
- Same as in distribution of organization type.



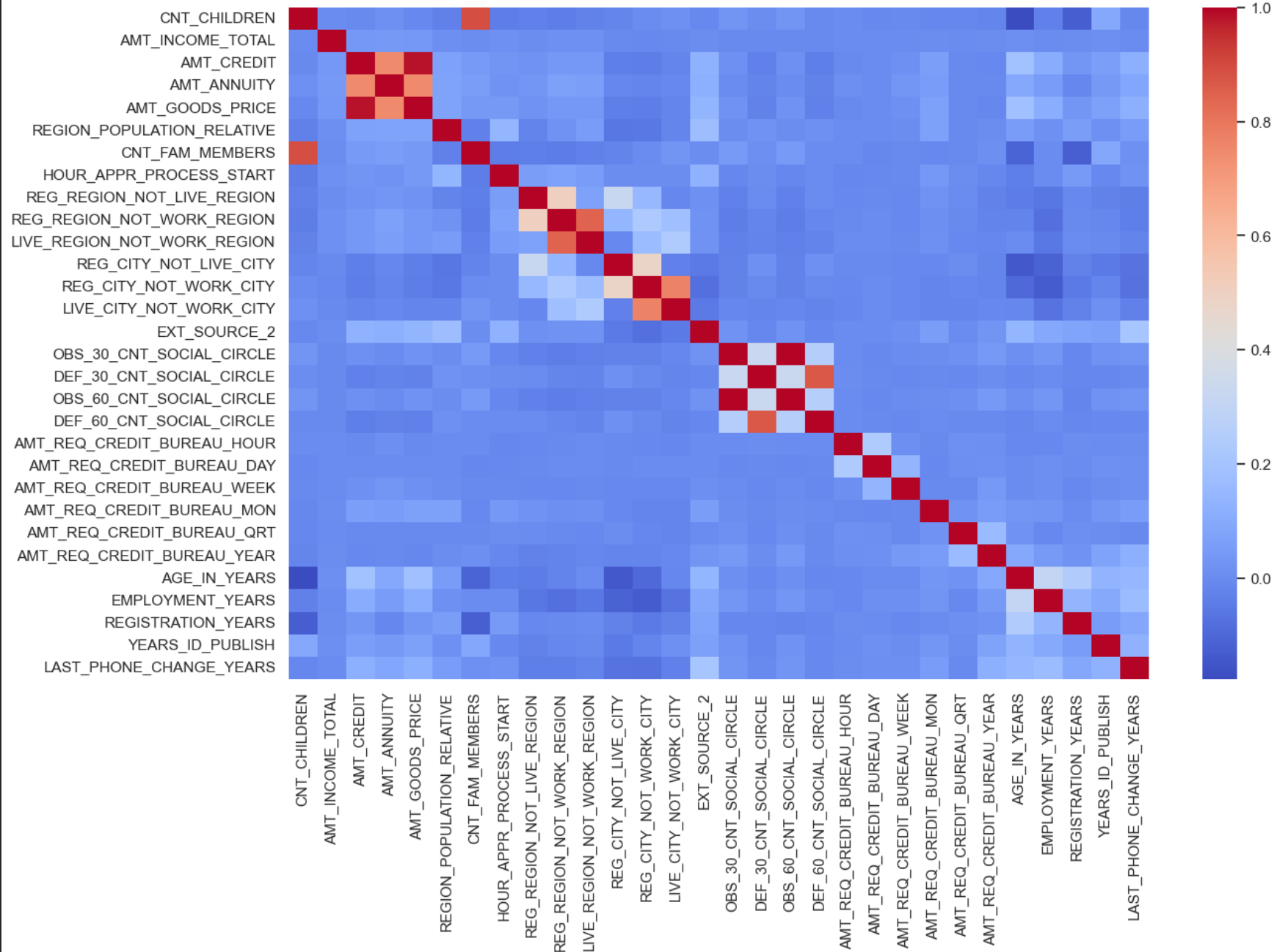
Correlation for Target=0



Conclusion from the correlation graph:

- Credit amount and birthdate are inversely related, credit amount is higher for younger people and vice versa.
- The quantity of credit is inversely correlated with the number of children a client has, therefore the amount of credit is larger when the client has fewer children and vice versa.
- The quantity of income is inversely correlated with the number of children a customer has, therefore a client with fewer children will have more income than a client with more children.
- In a densely populated location, clients have fewer children.
- Credit amount is given in greater amounts to heavily inhabited areas.
- In areas with high population density, income is also higher.

Correlation for Target=1



Conclusion from the correlation graph:

- Same as the target=0 heatmap above, adding some other points from this heatmap.
- The client's whose permanent address does not match contact address are having less children and vice-versa.
- The client's whose permanent address does not match work address are having less children and vice-versa.

Top 10 Correlations for Target = 0

- ❖ High correlations being observed at client's social surroundings 30 and 60 DPD.
- ❖ As the Goods Price increases, Credit increases as well.

REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.860421
LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.860421
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.861454
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.861454
CNT_CHILDREN	CNT_FAM_MEMBERS	0.893276
CNT_FAM_MEMBERS	CNT_CHILDREN	0.893276
AMT_GOODS_PRICE	AMT_CREDIT	0.986315
AMT_CREDIT	AMT_GOODS_PRICE	0.986315
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998491
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998491

dtype: float64

Top 10 Correlations for Target = 1

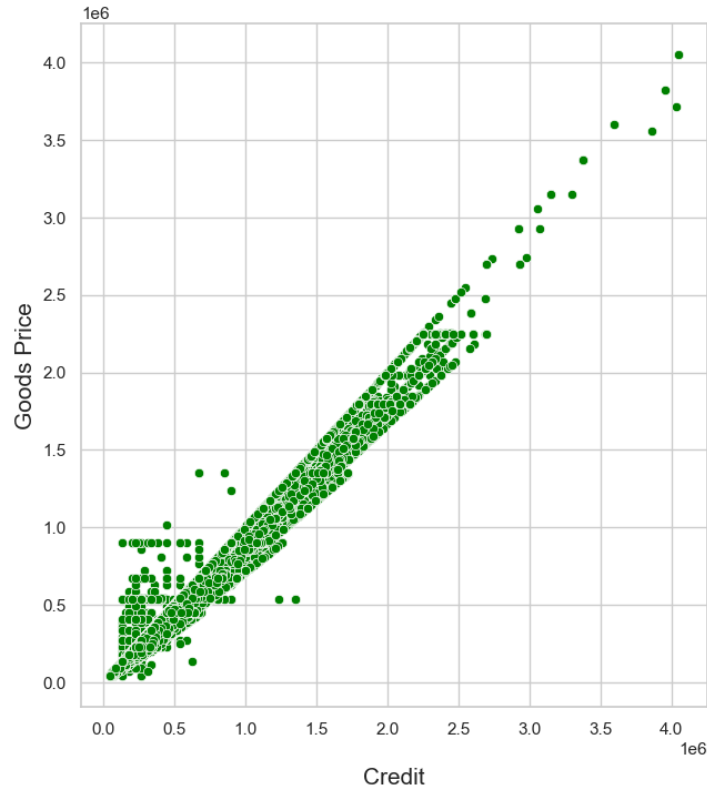
- ❖ High correlations being observed at client's social surroundings 30 and 60 DPD.
- ❖ As the Goods Price increases, Credit increases as well.

LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.846872
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.846872
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.867963
DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.867963
CNT_FAM_MEMBERS	CNT_CHILDREN	0.893829
CNT_CHILDREN	CNT_FAM_MEMBERS	0.893829
AMT_CREDIT	AMT_GOODS_PRICE	0.982239
AMT_GOODS_PRICE	AMT_CREDIT	0.982239
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.998288
OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998288
dtype: float64		

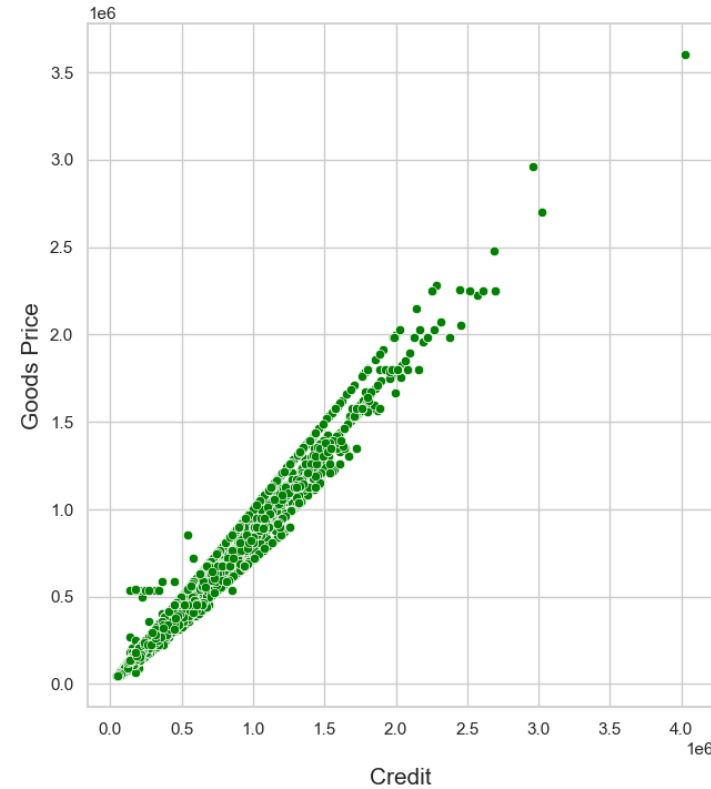
Numerical Bivariate Analysis



CREDIT vs GOODS PRICE for Target-0



CREDIT vs GOODS PRICE for Target-1



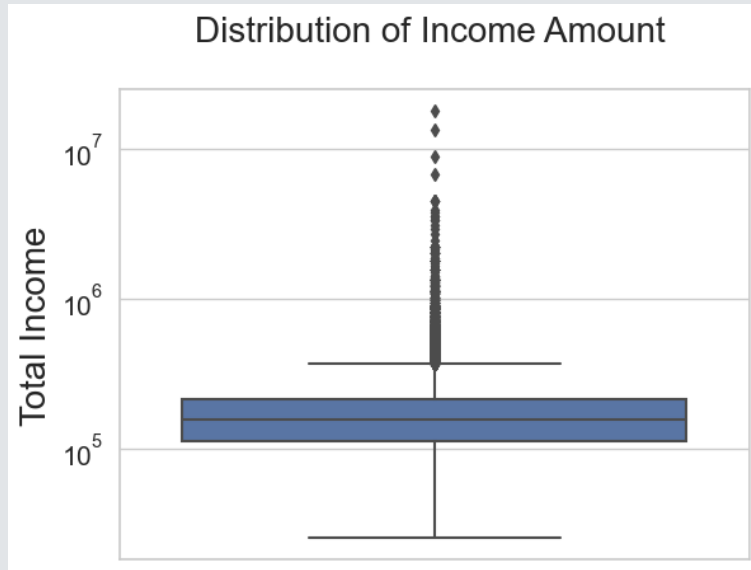
□ Conclusion from the graph:

- Using the scatter plot, we can see that there is a strong correlation between AMT CREDIT and AMT GOODS PRICE, meaning that when goods prices rise credit also rises and vice versa.

Finding Outliers

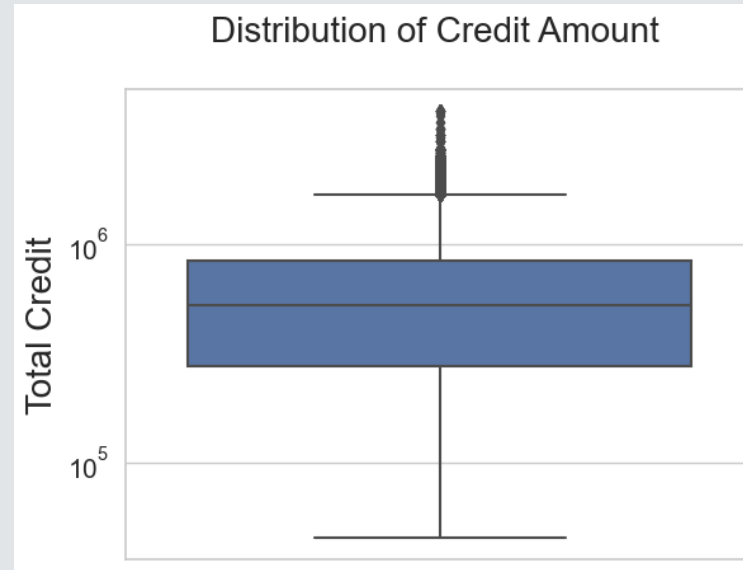
- Univariate Analysis





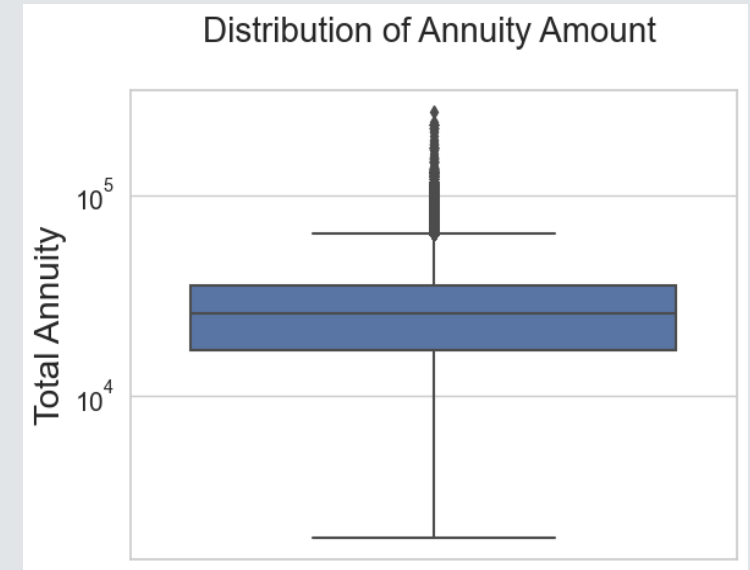
❑ Conclusion from the graph:

- The clients Income Amounts appears to be equally distributed.
- Also, some outliers are present in the dataset.



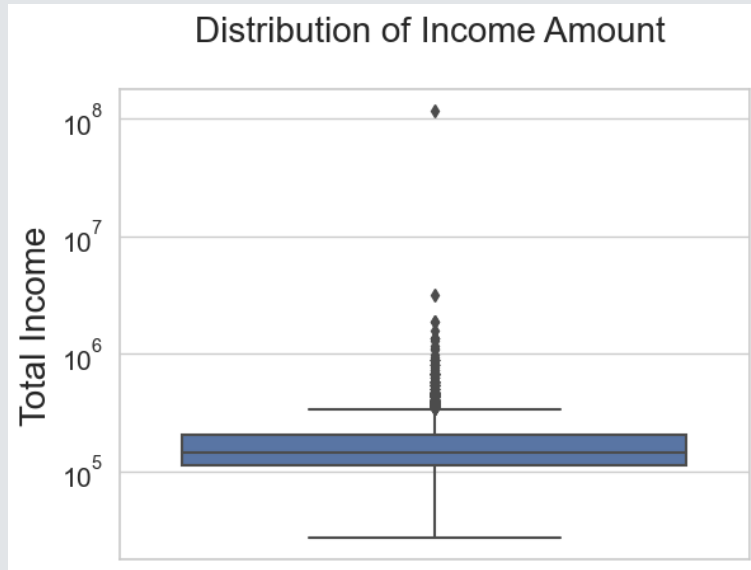
❑ Conclusion from the graph:

- Since the first quartile is larger than the third, the first quartile contains the majority of the client credit.
- Also, some outliers are present in the Credit boxplot.



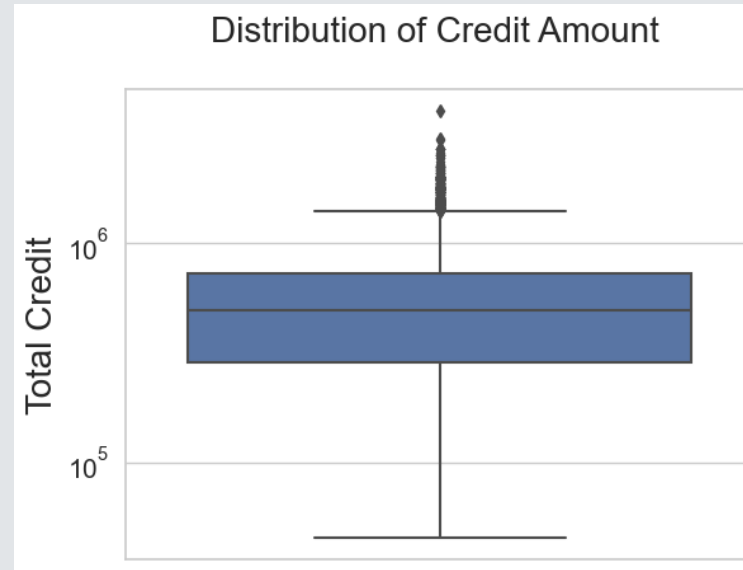
❑ Conclusion from the graph:

- The first quartile is bigger than the third quartile.
- Also, some outliers are present in the Annuity boxplot.



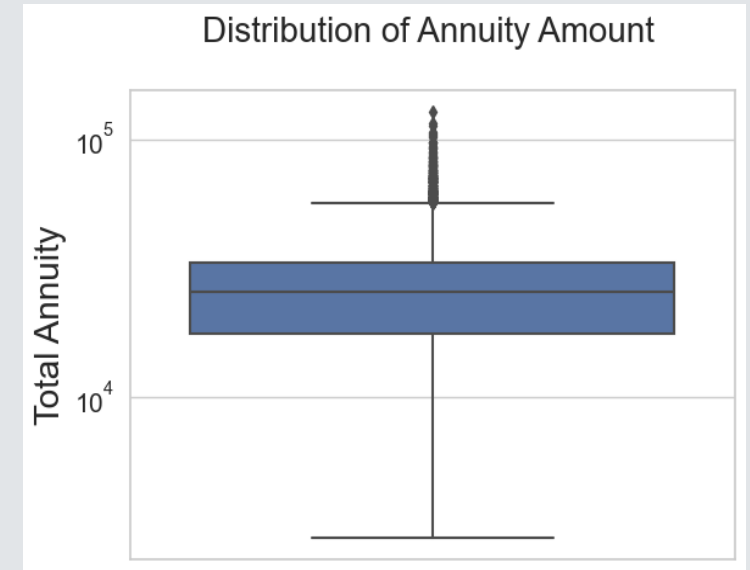
❑ Conclusion from the graph:

- The Income dataset appears to have a significant outlier.
- The client's income is located in the third quartile.



❑ Conclusion from the graph:

- Since the first quartile is greater than the third, the first quartile contains the majority of the client credit.
- Also, some outliers are present in the Credit boxplot.



❑ Conclusion from the graph:

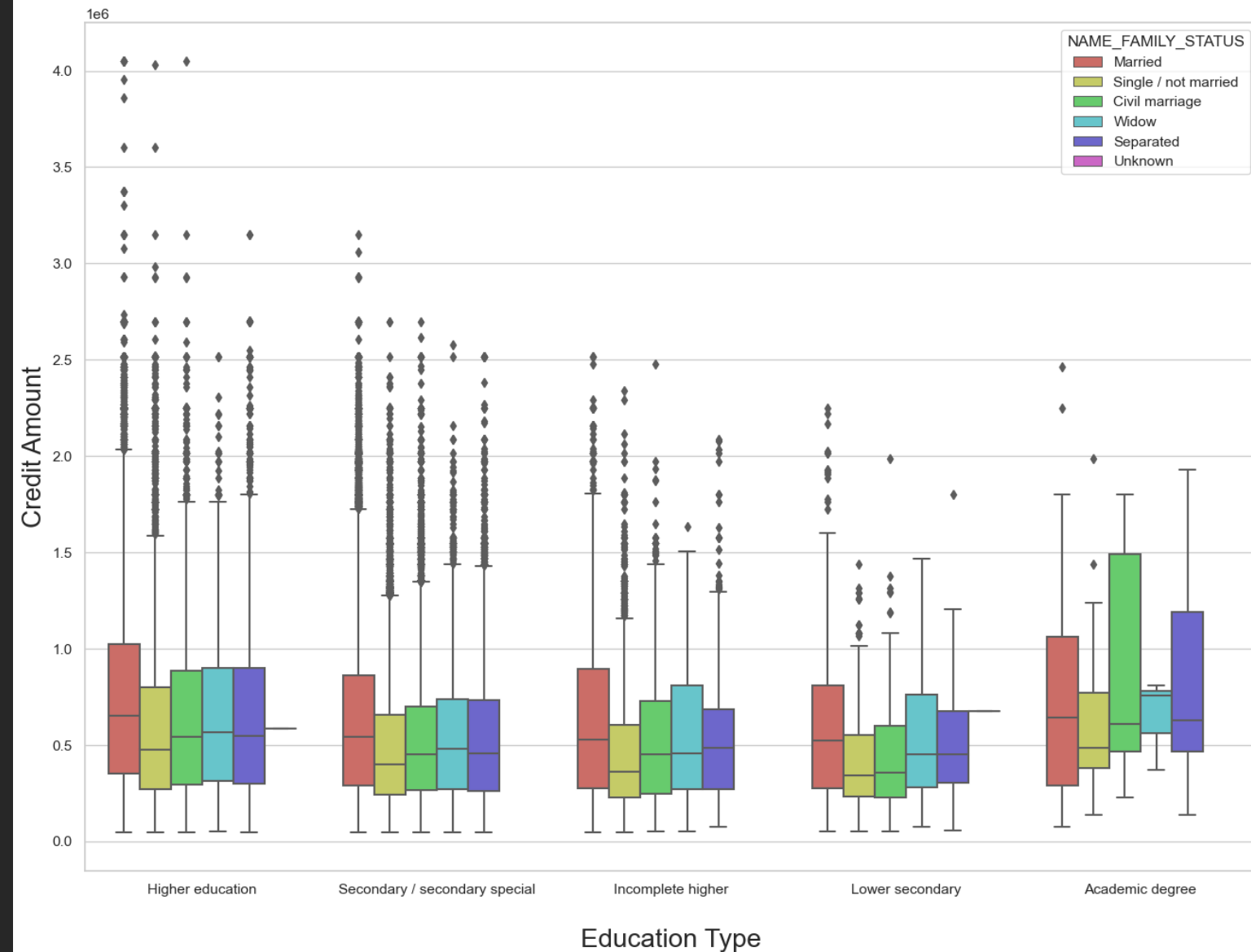
- The first quartile exceeds the third quartile in size.
- Also, some outliers are present in the Annuity boxplot.

Finding Outliers

- Multivariate Analysis



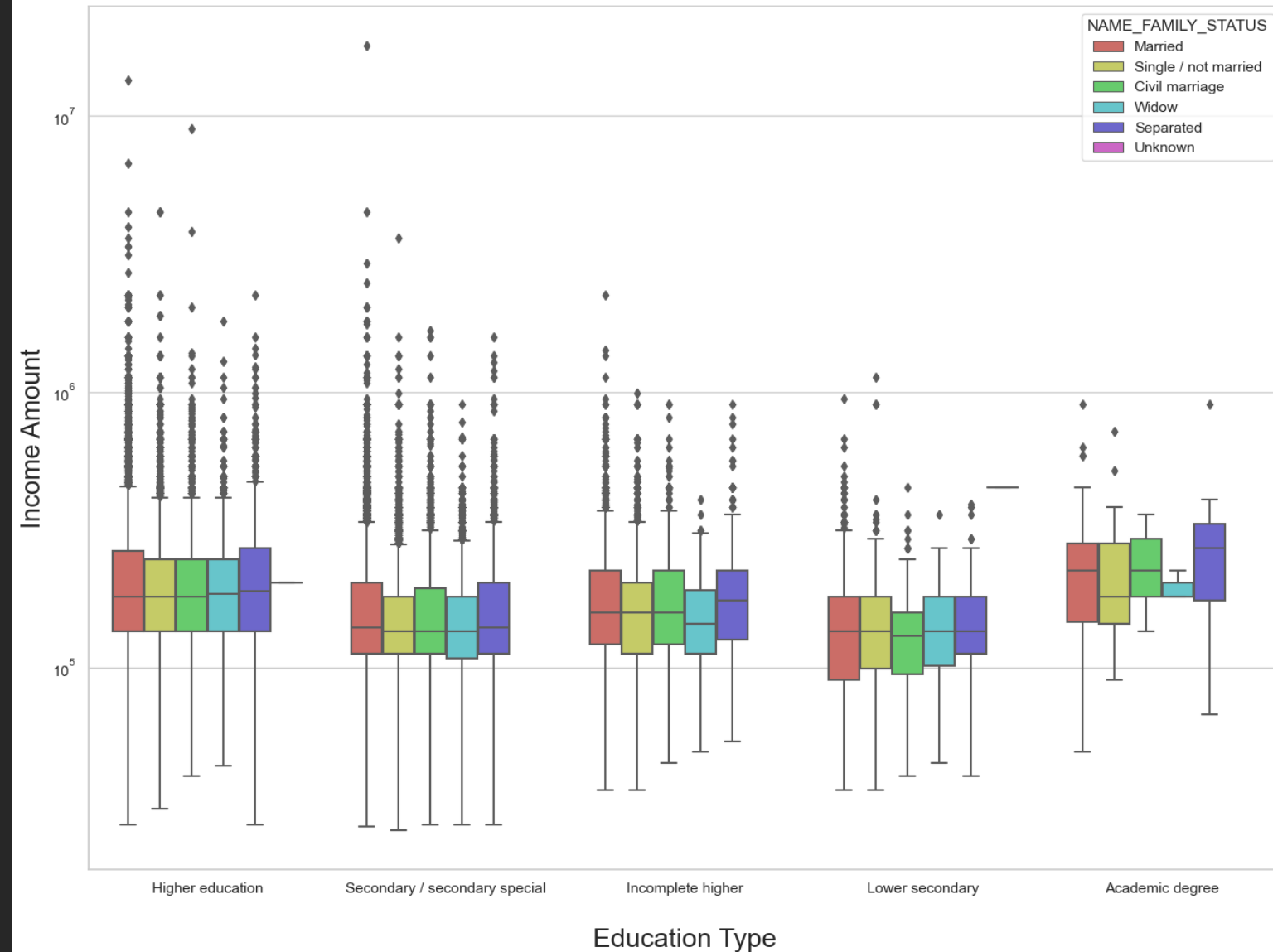
Credit amount vs Education Status (TARGET=0)



□ Conclusion from the graph:

- We can infer from the above box plot that academic degree holders with family statuses of "Civil marriage," "Marriage," and "Separated" have more credits than those with other statuses.
- Additionally, there are more outliers in higher education in the categories of "Marriage," "Single," and "Civil marriage."
- Most of the credits for civil unions for academic degrees are in the third quartile.

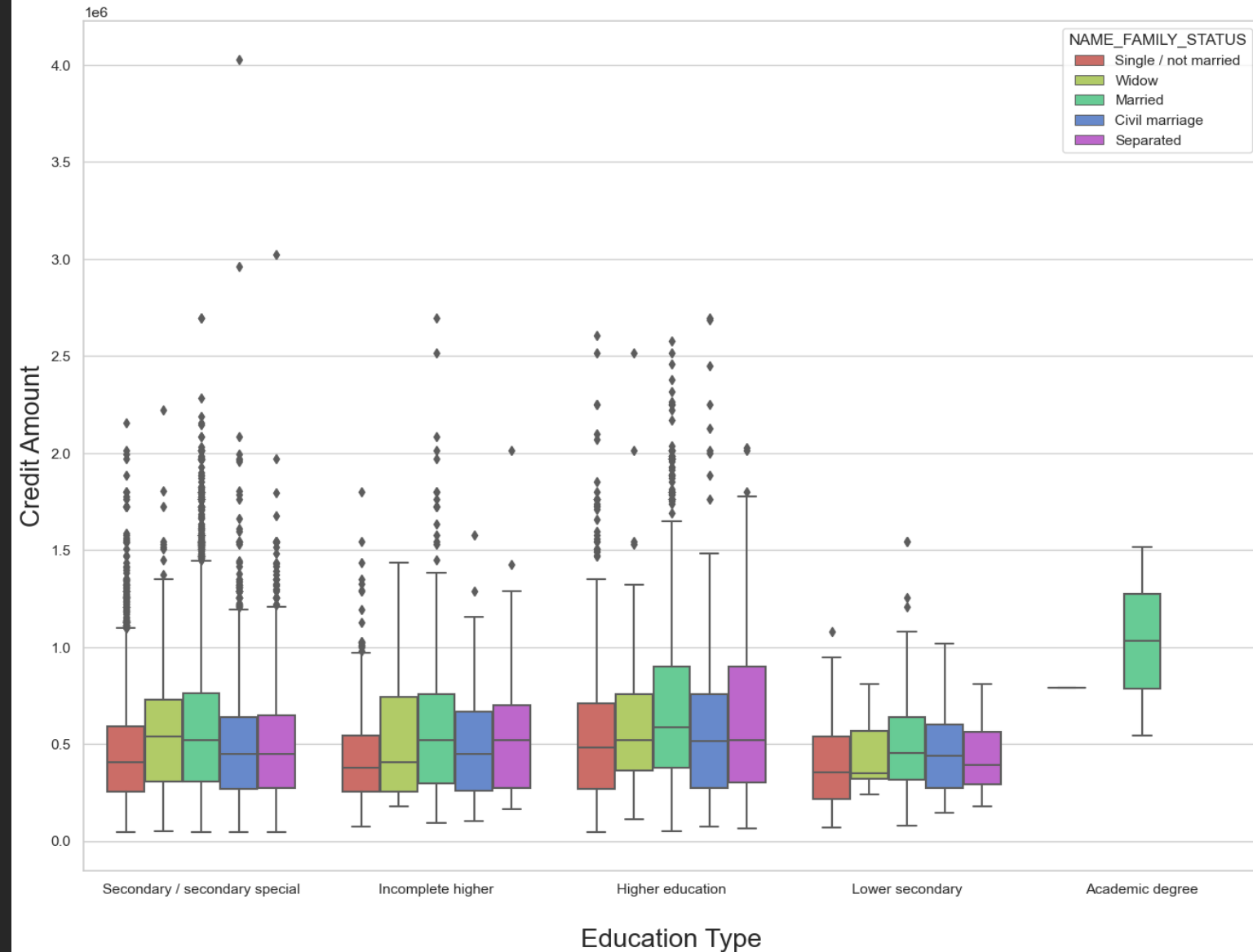
Income amount vs Education Status (TARGET=0)



□ Conclusion from the graph:

- According to the boxplot above, the salary amount and family status are largely equivalent for the education type "Higher education." There are a lot of outliers in it.
- Less outlier are having for Academic degree, but their income amount is little higher than Higher education.
- Lower secondary of civil marriage family status are have less income amount than others.

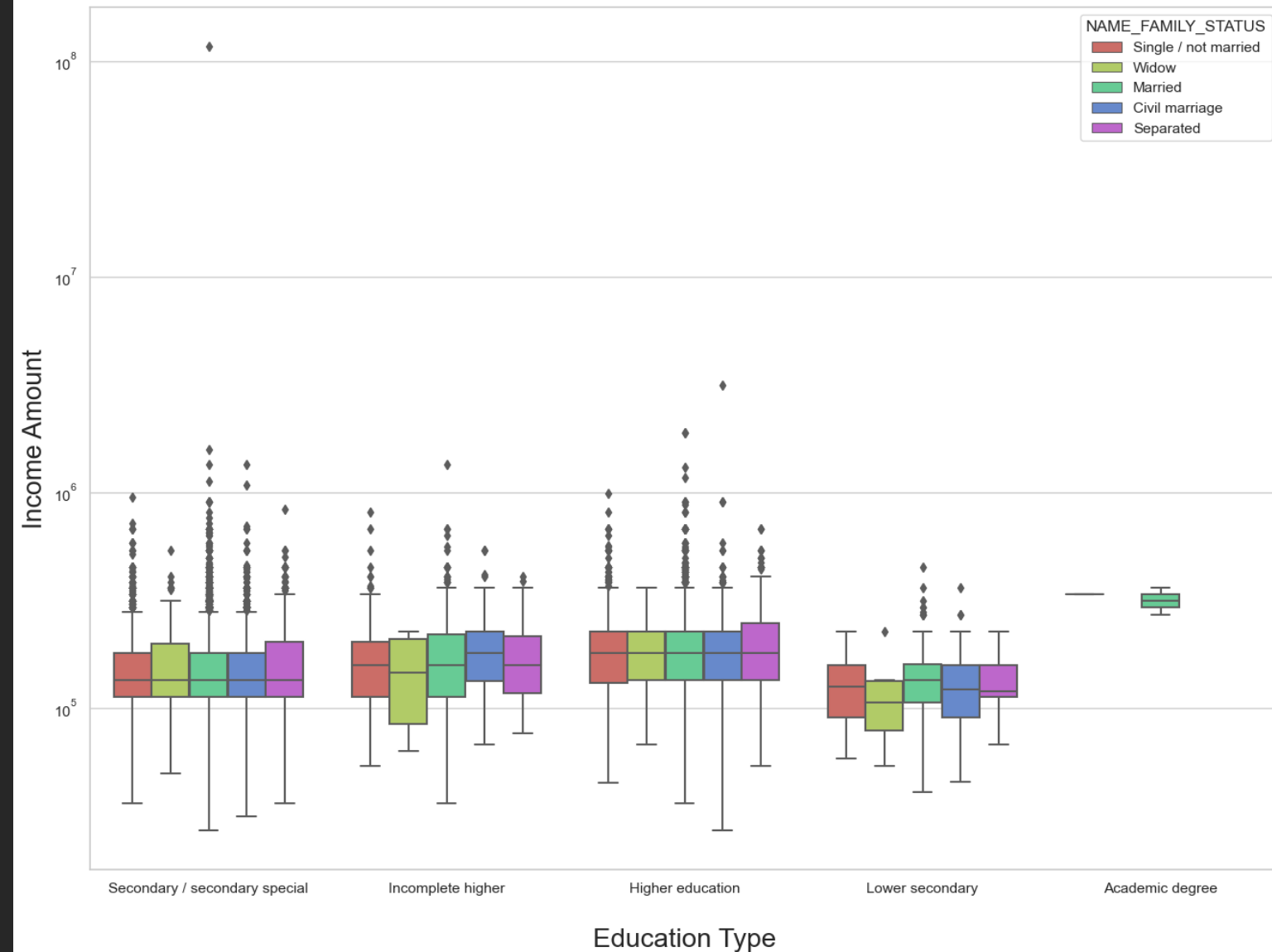
Credit amount vs Education Status (TARGET=1)



Conclusion from the graph:

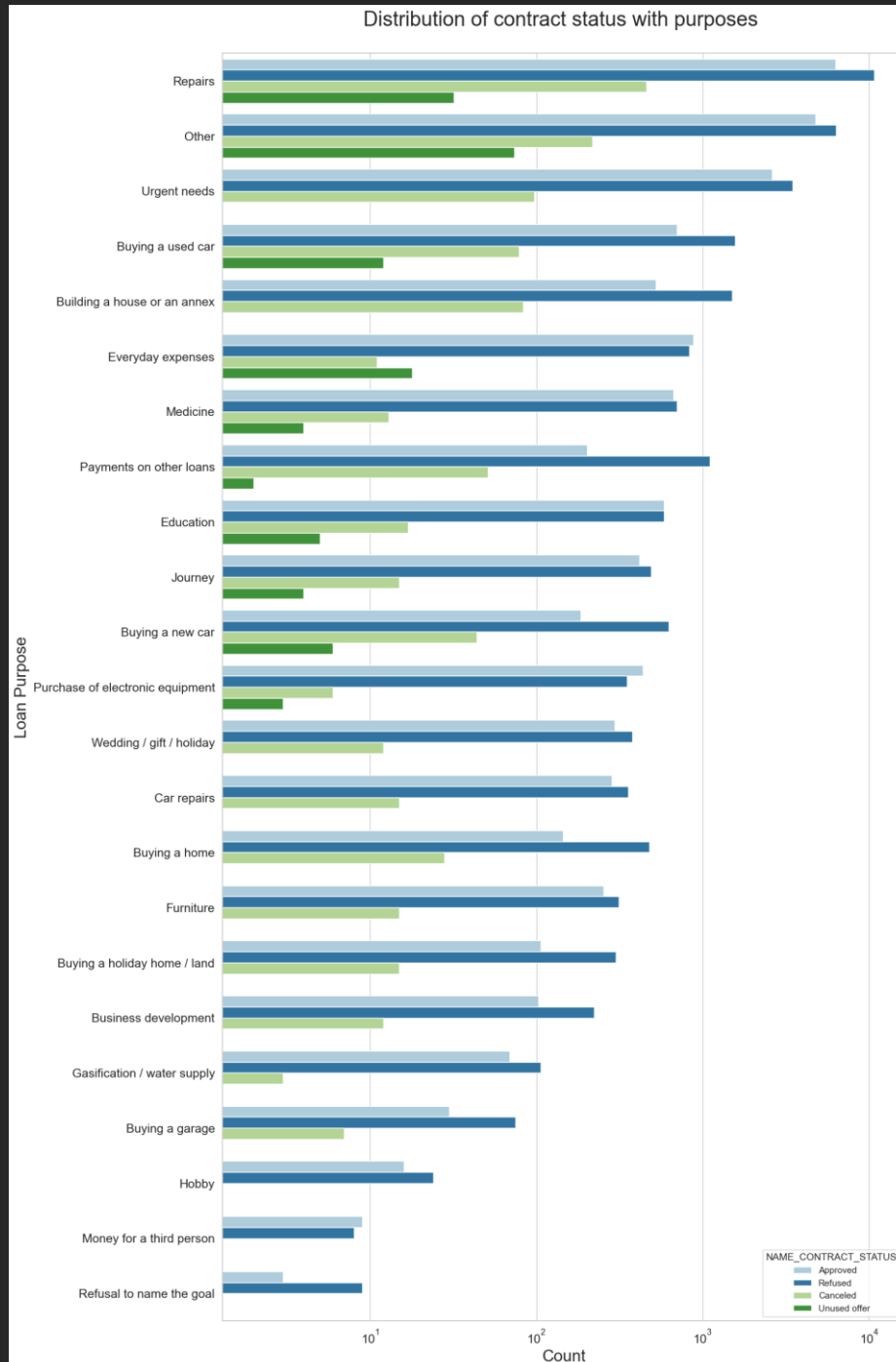
- According to the above box plot, Academic degree education with family statuses of "Civil marriage," "Marriage," and "Separated" have more credits than those with other statuses.
- Higher education and secondary make up the majority of the outliers in the data.
- Most of the credits for civil marriage for Academic degrees are in the third quartile.

Income amount vs Education Status (TARGET=1)



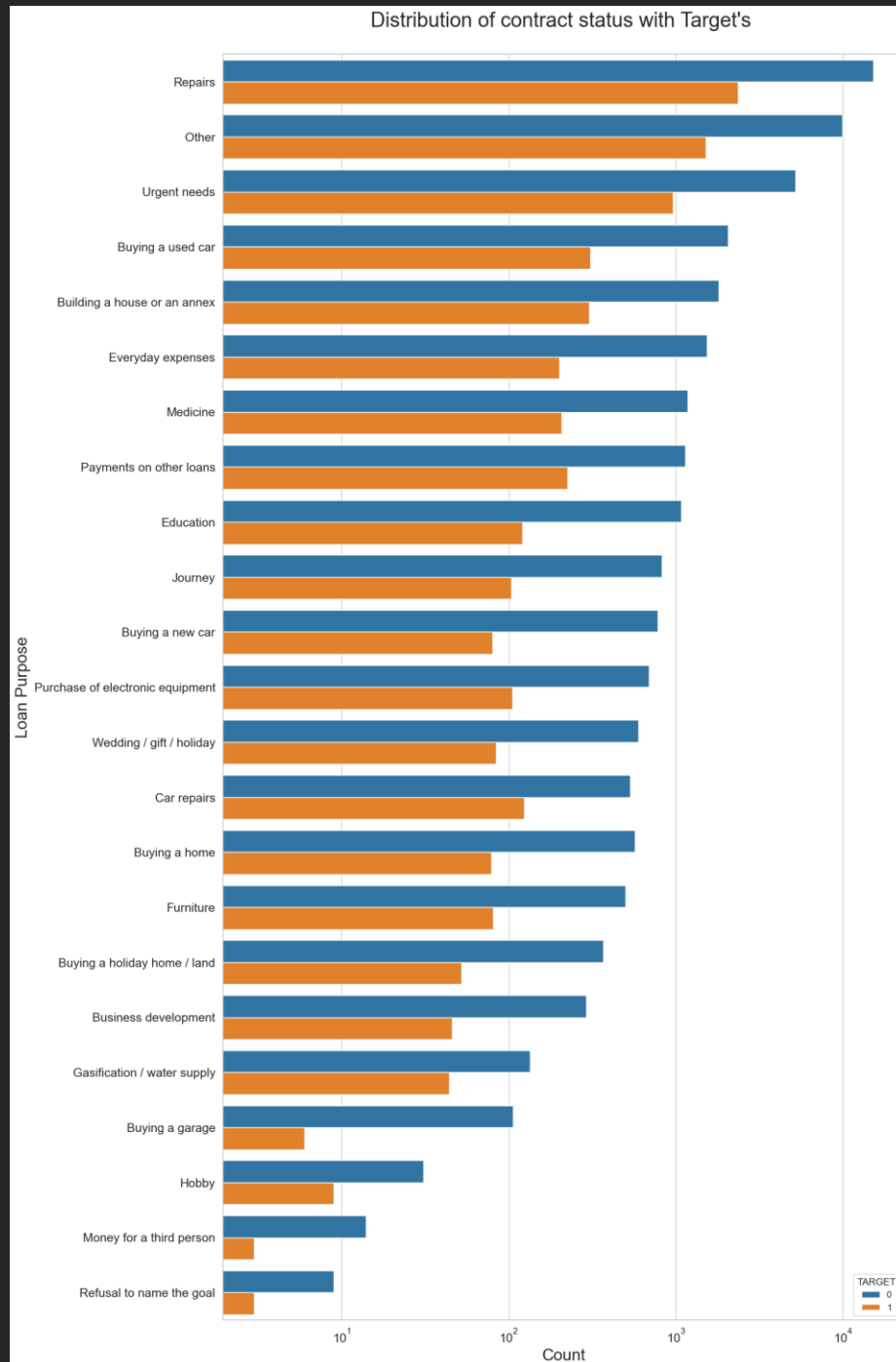
□ Conclusion from the graph:

- According to the boxplot above, the salary amount and family status are largely equivalent for the education type "Higher education."
- Less outliers are there in Academic degree, yet their salary is a little higher than that of those with higher education.
- Lower secondary have less income amount than others.



□ Conclusion from the graph:

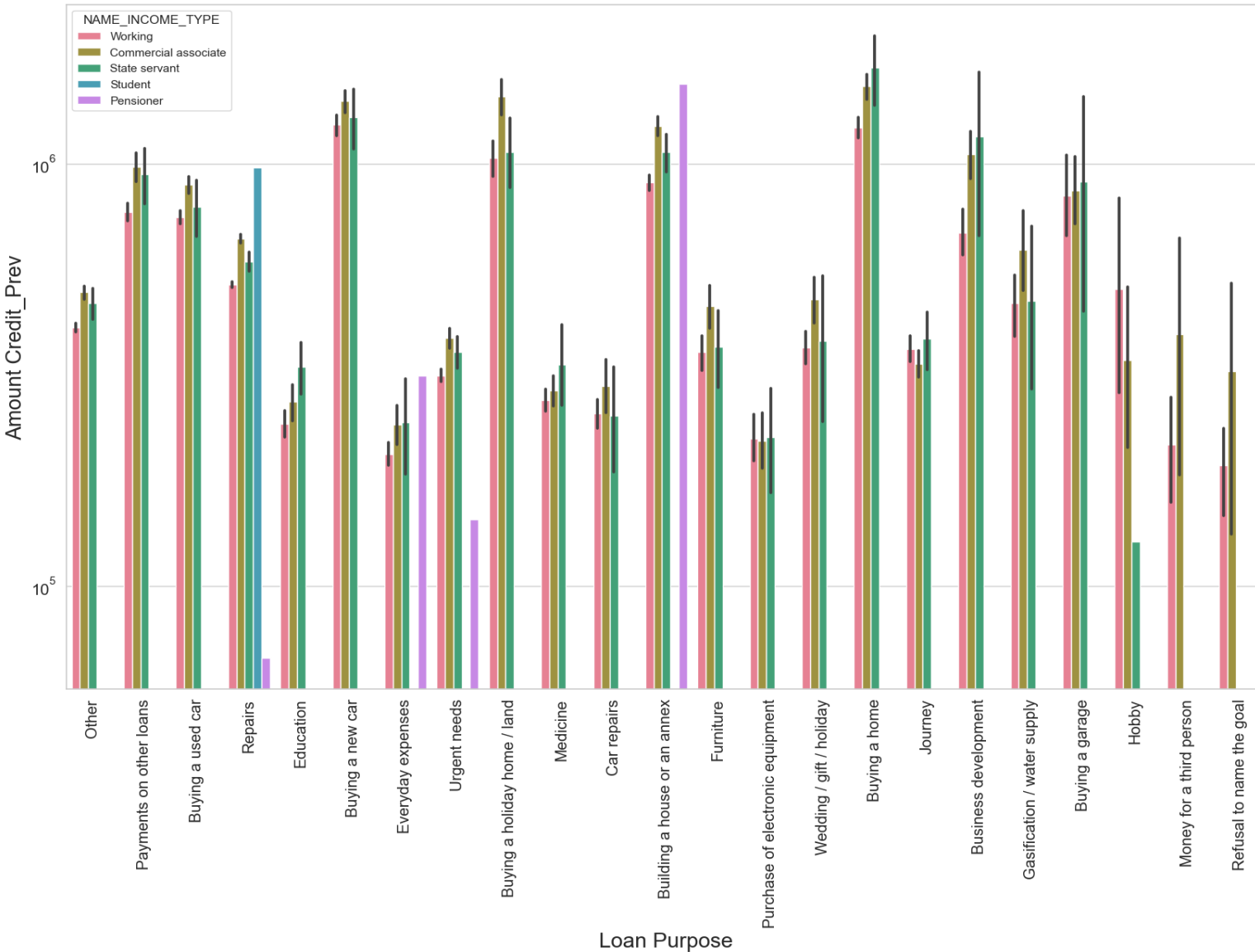
- The majority of loan rejections were for the "Repairs".
- We have an equal amount of approvals and rejections for educational purposes.
- The rejection rate for paying off other debt and getting a new car is significantly greater than the approval rate.



□ Conclusion from the graph:

- Loan purposes with "Repairs" are having difficulties making timely payments.
- There are few places where loan payment is significant higher than facing difficulties. They are 'Buying a garage', 'Business developemt', 'Buying land', 'Buying a new car' and 'Education'. Therefore, we can concentrate on these goals for which the client is experiencing just minor payment issues.

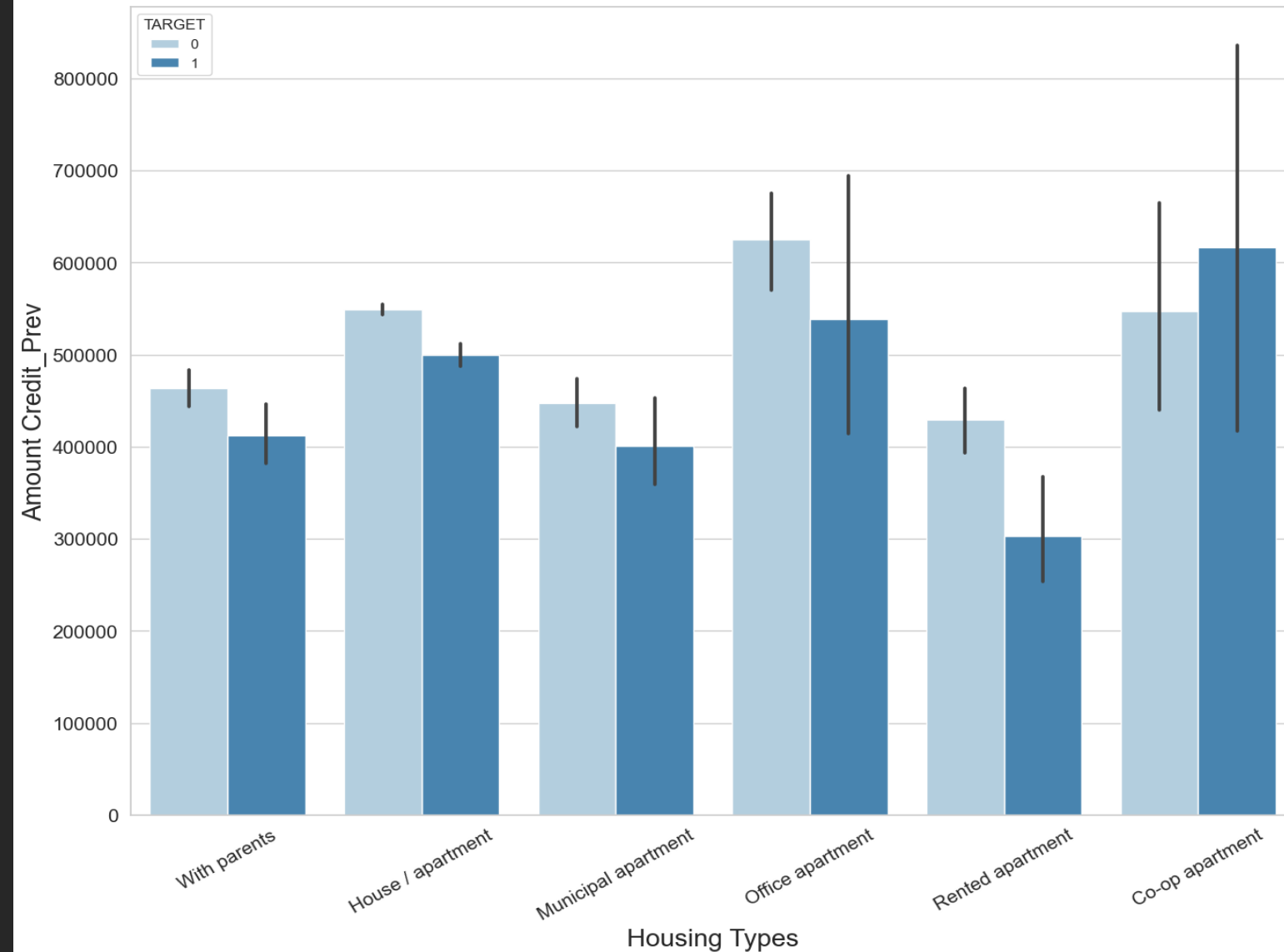
Prev Credit amount vs Loan Purpose



Conclusion from the graph:

- For loan objectives such as "buying a home," "buying a land," "buying a new car," and "building a house," the credit amount is larger.
- Income type of state servants have a significant amount of credit applied.
- Less credits are requested when money is needed for a hobby or third party.

Prev Credit amount vs Housing type



□ Conclusion from the graph:

- Here, the office apartment has a higher credit of target=0 while the co-op apartment has a higher target=1.
- We can therefore draw the conclusion that a bank should refrain from lending to the co-op apartment as they are having difficulties in payment.
- For successful payments, the bank can concentrate mostly on housing types such as those with parents, Houses\apartments, or municipal apartments.

Conclusion on the analysis of the data



Banks should approve loans more for Office apartment, Co-Op apartment housing type as there are less payment difficulties.



Banks should provide loans to 'Repairs' & 'Others' purposes.



Banks should provide loans to the 'Business Entity Type-3' and 'Self-Employed' persons.



'Working' people especially female employers are the best to target for the loans.