Credit EDA Case Study

Mohammad Tanvir

Problem Statement

- **Background**

  - A loan providing company which lends loans to the urban customers, processes loan application by verifying their capability to re-pay the loan.

- **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.

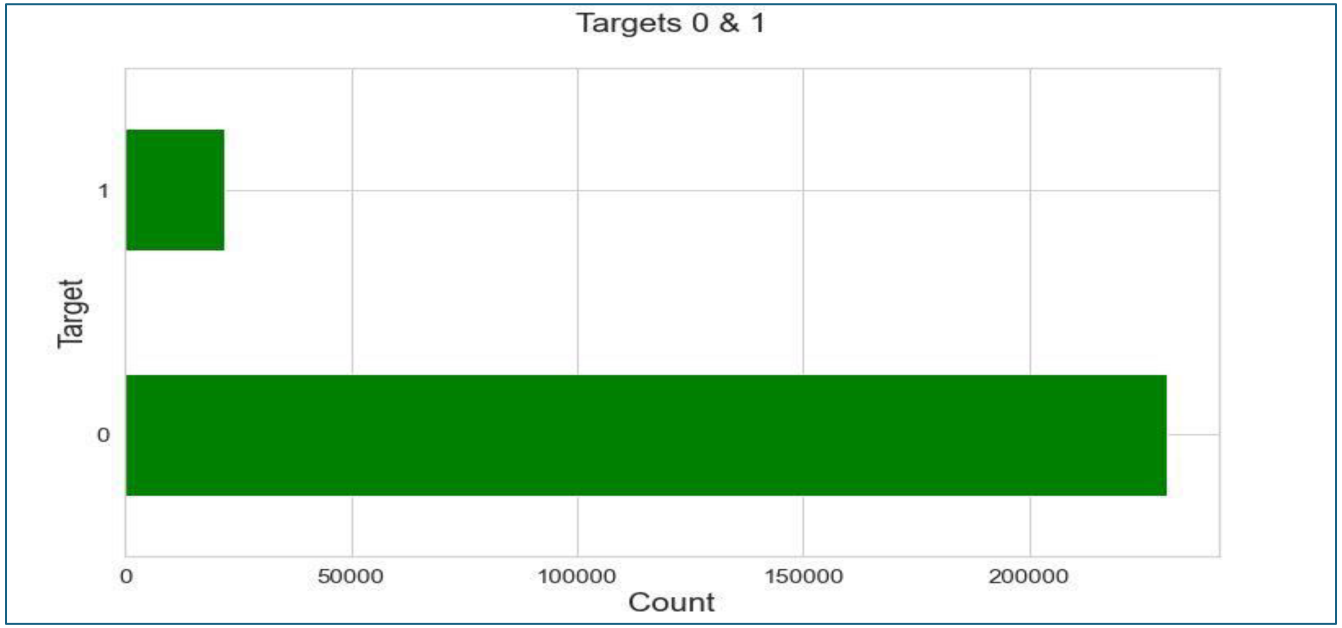
Understanding the data

- **<u>Missing value check</u>**
  - **For Application Dataset**
    - Removal of null values in data which are present more than 19%, as it can give outliers, which can result in incorrect analysis.
  - **For Previous Application**
    - Removal of null values in data which are present more than 20%, as it can give outliers, which can result in incorrect analysis.

Understanding the data

- **Imputing the missing values**
  - **For Application Dataset**
    - Imputed values in the categorical and numerical columns with the help of mean , median and mode values in the dataset.
  - **For Previous Application**
    - There are no such values which needs to corrected.

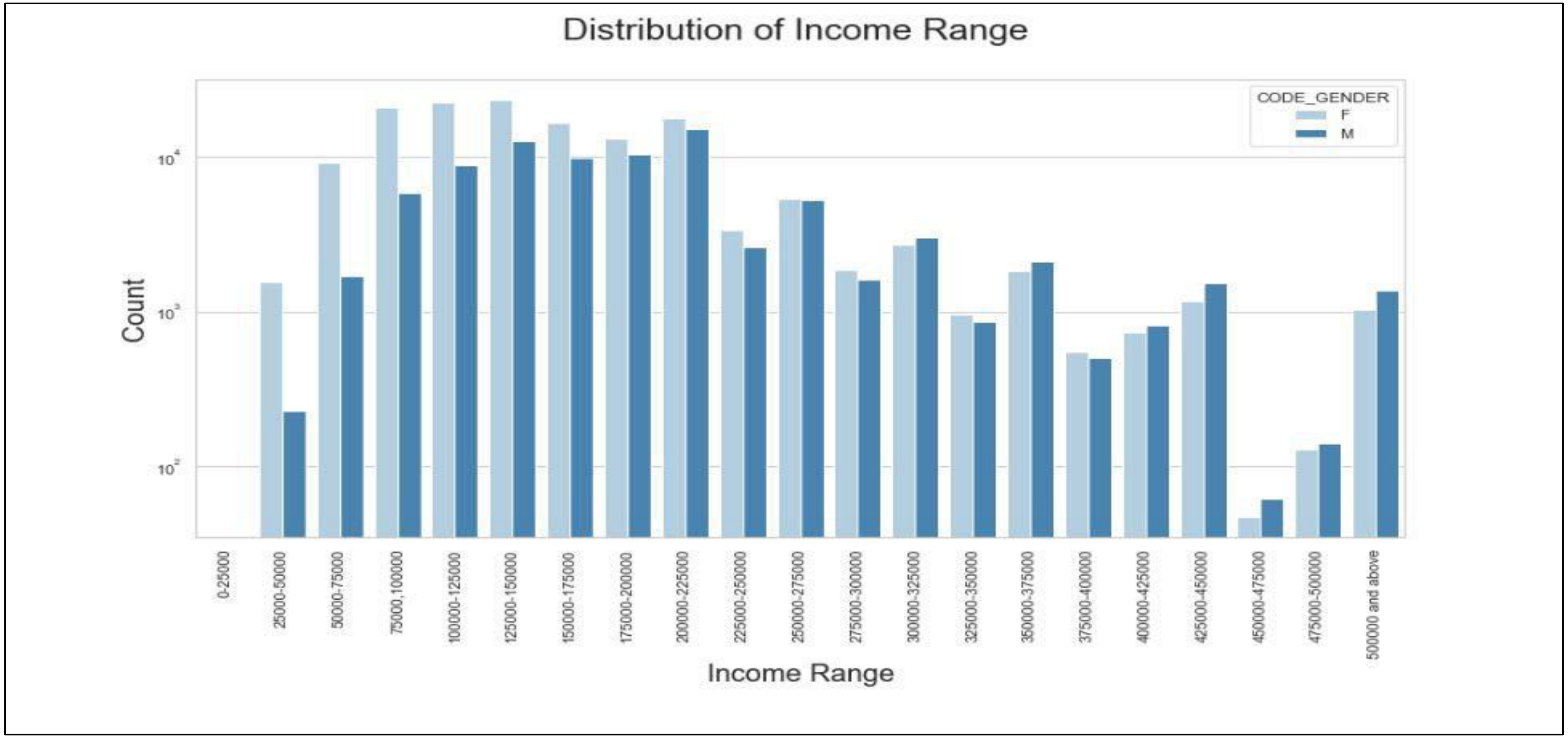
Imbalance Ratio



Plot: **1**

The data after cleanup is highly imbalanced around data for loan defaulters (TARGET = 1) and remaining data for non-defaulters (TARGET = 0).

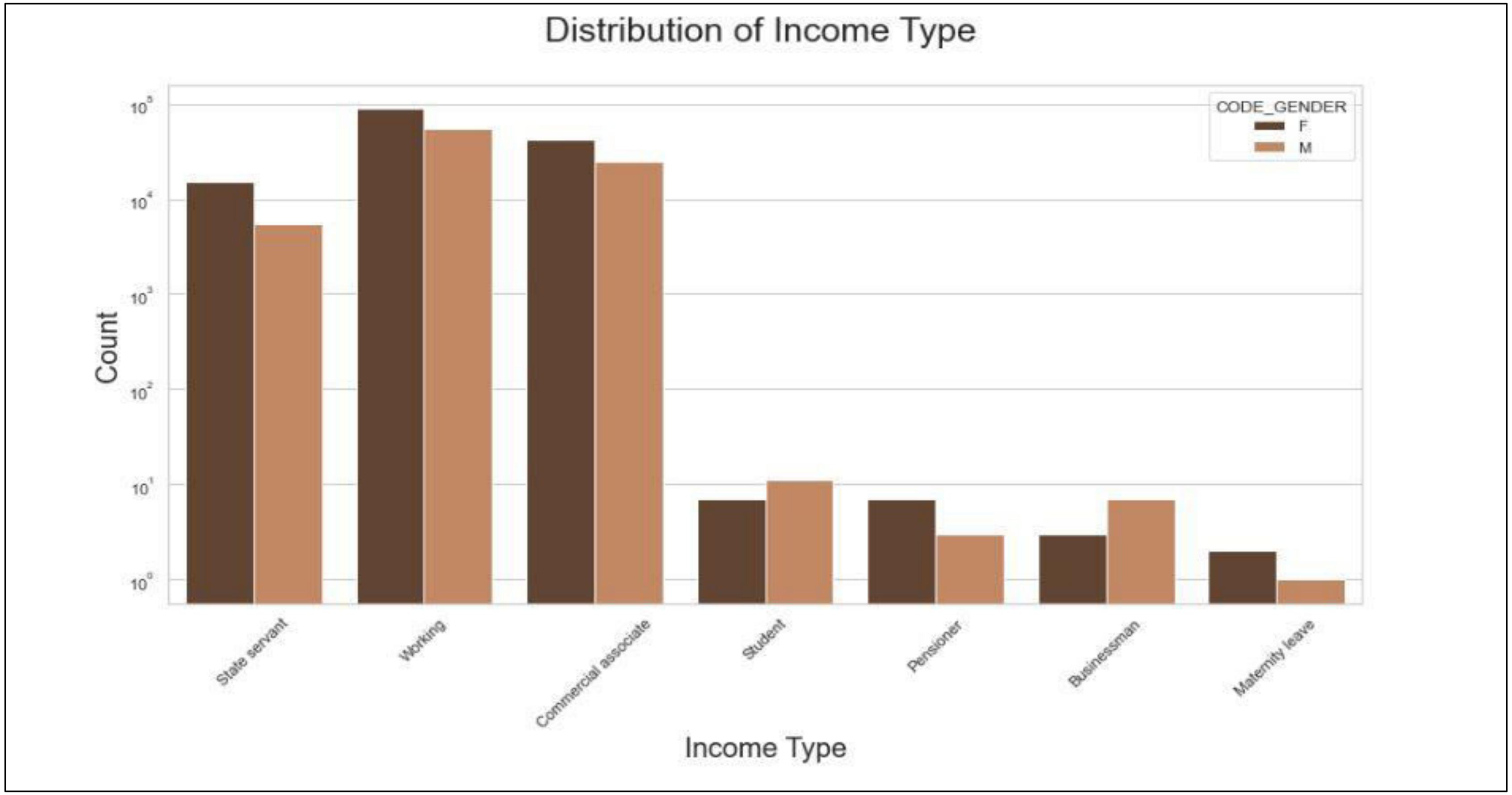
Categorical Univariate Analysis for Target 0



Plot: 2

Conclusion from the graph:

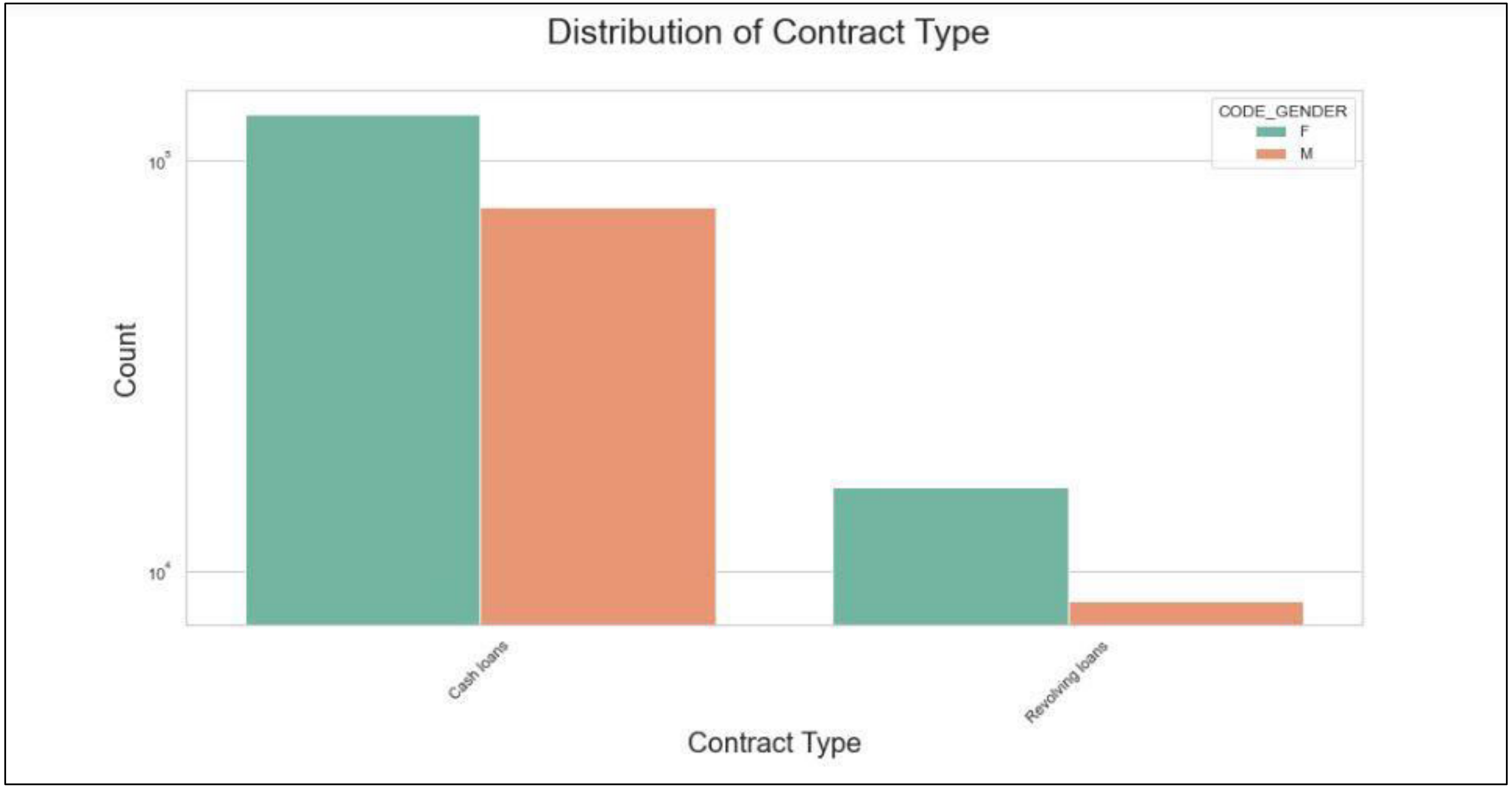
- Income range from 125000 to 150000 is having more number of credits.
- Very less count from range 450000-475000.
- It seems that the females are more than male in having credit for range:125000 to 150000.



Plot: 3

Conclusion from the graph:

- It seems that working women have most credit than others.
- It seems that 'State Servant', ‘Working' and 'Commercial Associate‘ have more credit counts compared to others.
- It seems Women in 'Maternity leave' has less credit in comparison to others.



Plot: 4

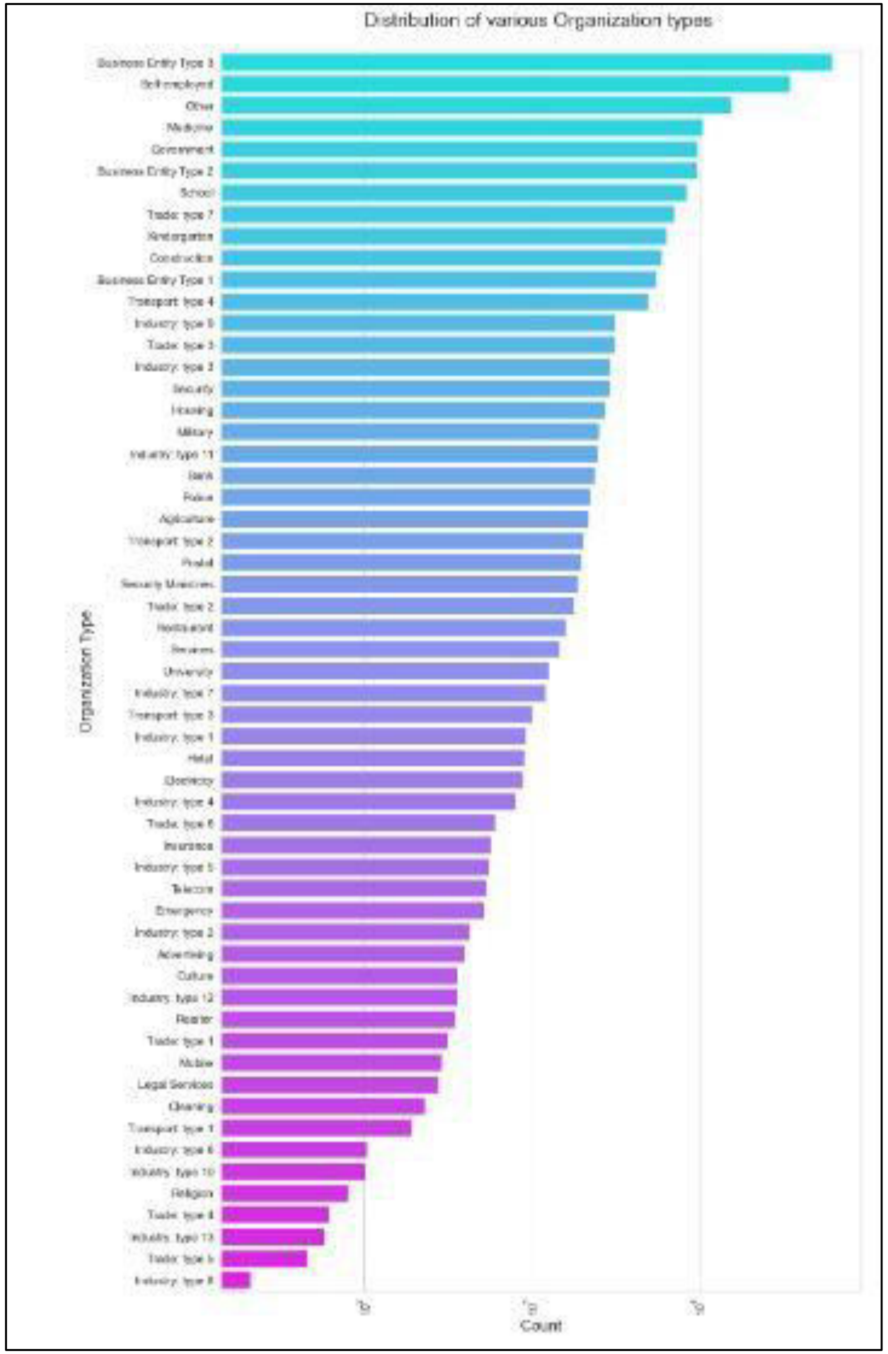
Conclusion from the graph:

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

Conclusions from the graph:

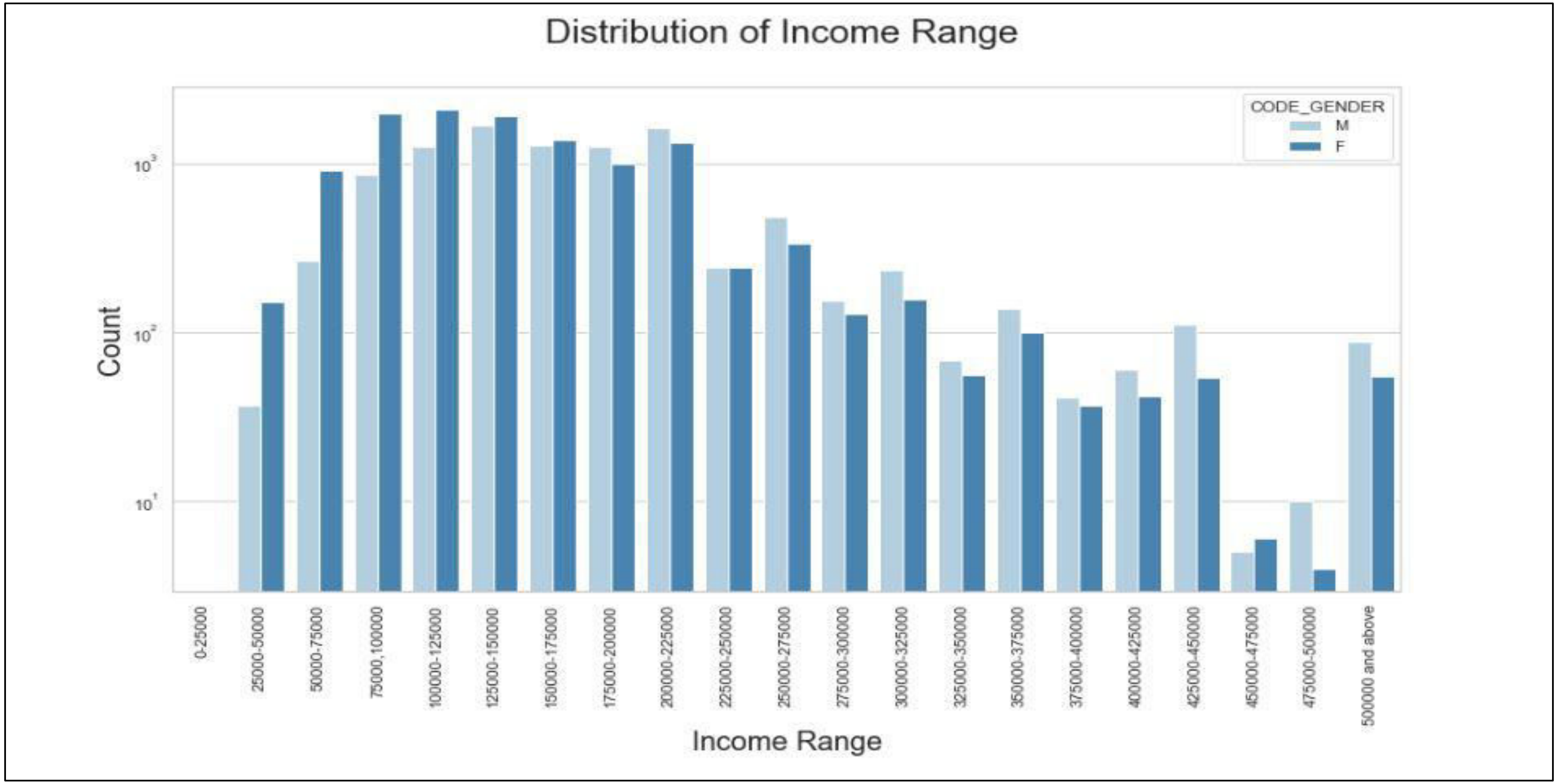
- Clients which have applied for credits are from most of the organization type ‘Business entity Type 3’ , ‘Self employed’, ‘Other’ , ‘Medicine’ and ‘Government’.

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



Plot: 5

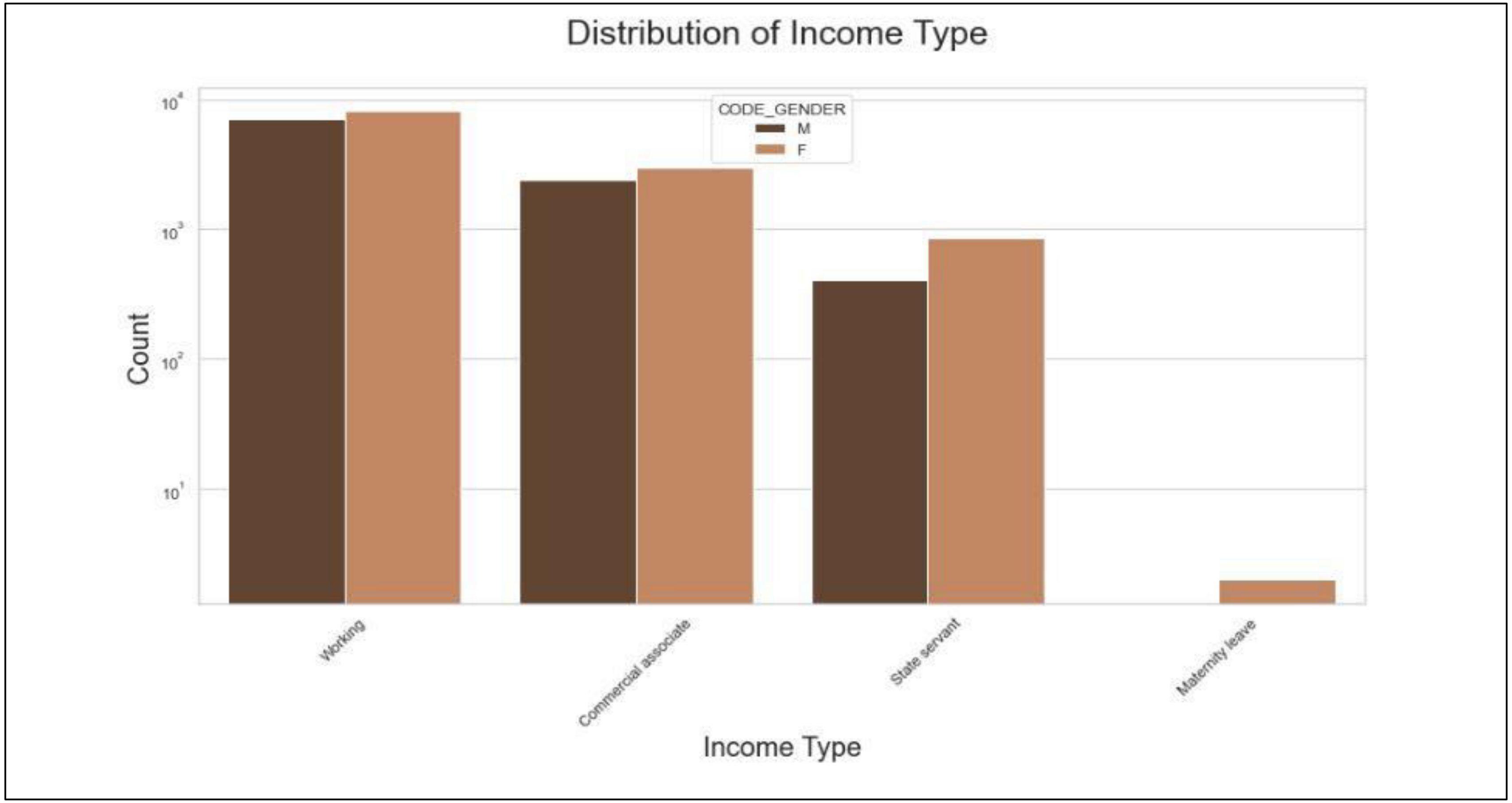
Categorical Univariate Analysis for Target 1



Plot: 6

Conclusions from the graph:

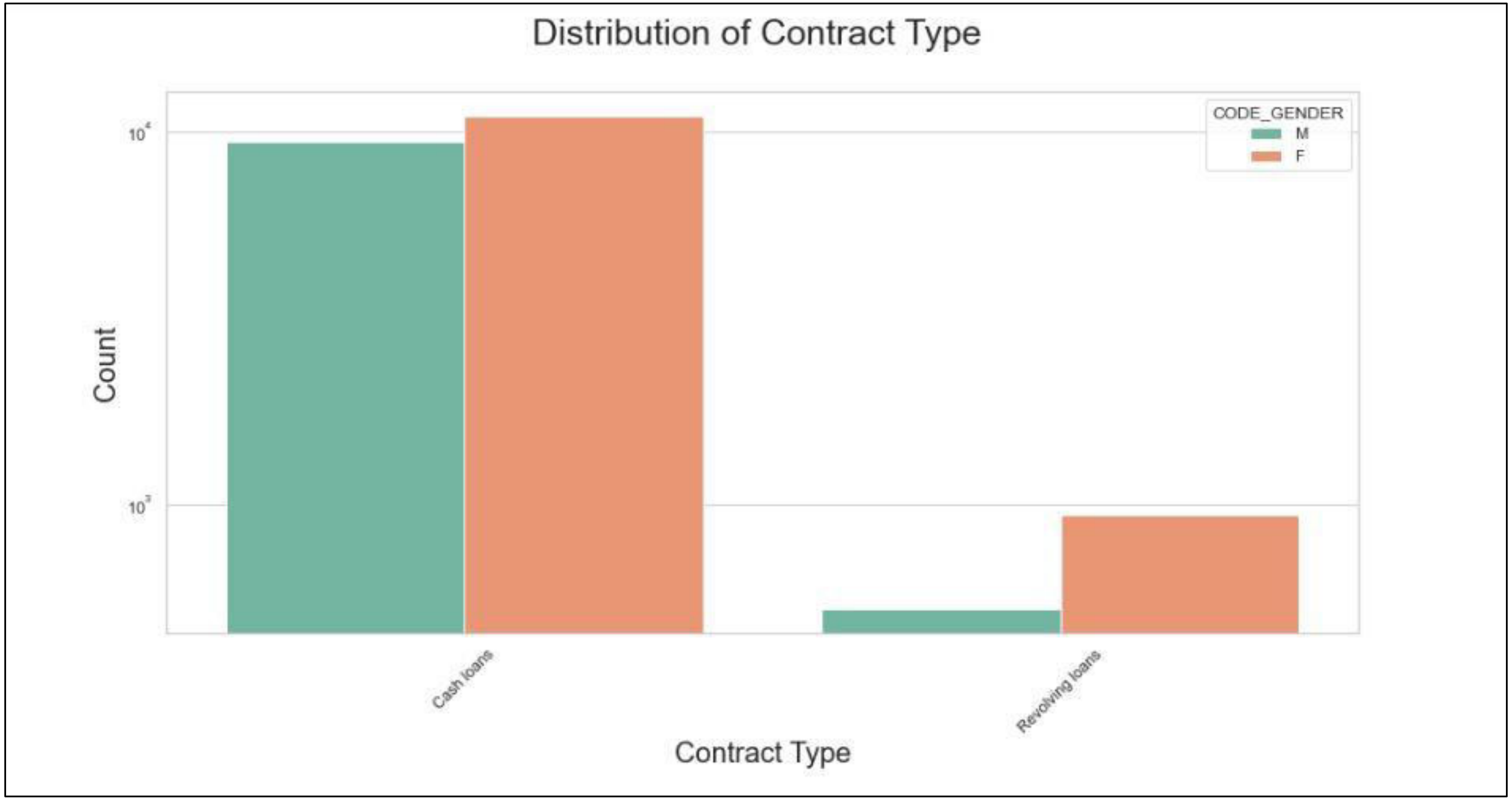
- Male Counts are higher.
- Income range from 100000 to 200000 is having more number of credits.
- Less count for income range 450000-475000.



Plot: 7

Conclusions from the graph:

- For income type 'working', 'commercial associate', and 'State Servant' the number of credits are higher than other i.e. 'Maternity leave.
- For this Females are having more number of credits than male.
- Less number of credits for income type 'Maternity leave'.



Plot: 8

Conclusions from the graph:

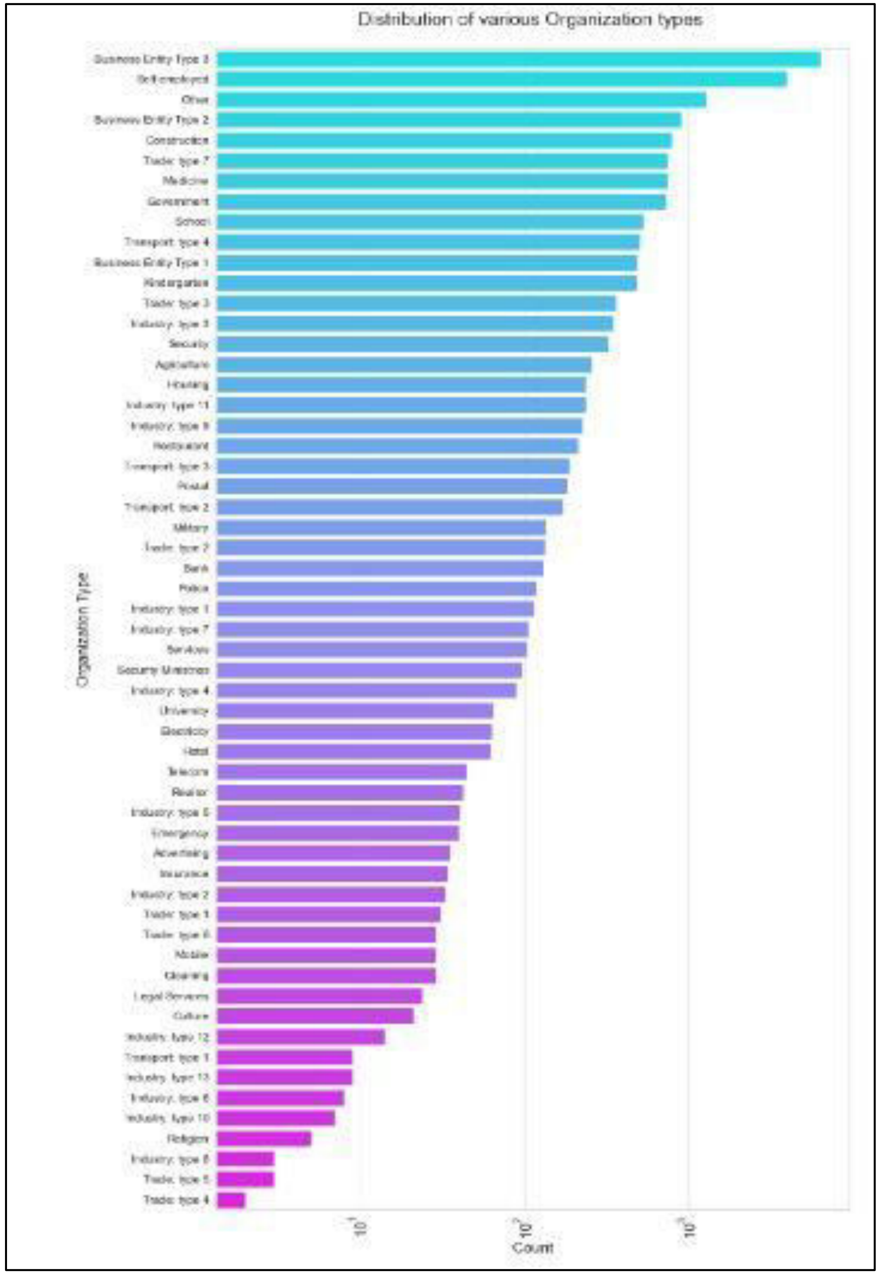
- For contract type ‘cash loans’ is having higher number of credits than ‘Revolving loans’ contract type.
- For this also Female is leading for applying credits.

Conclusions from the graph:

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

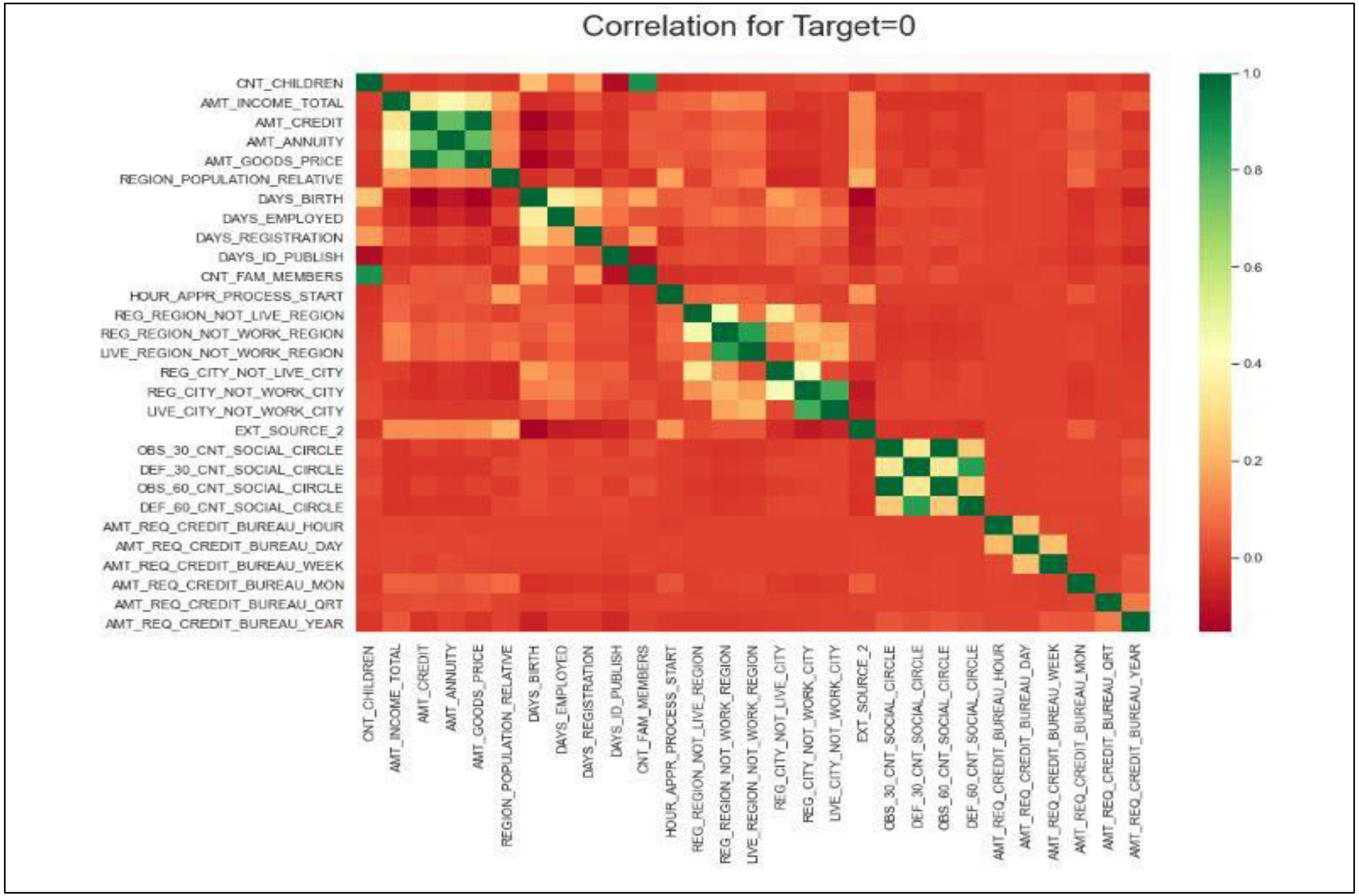
- Less clients are from Industry type 8,type 6, type 10, religion and trade type 5, type 4.

- Same as type 0 in distribution of organization type.



Plot: 9

Correlation for target = 0



Plot: 10

In Continuation…

In Continuation…

- ***Conclusions from the correlation graph:***

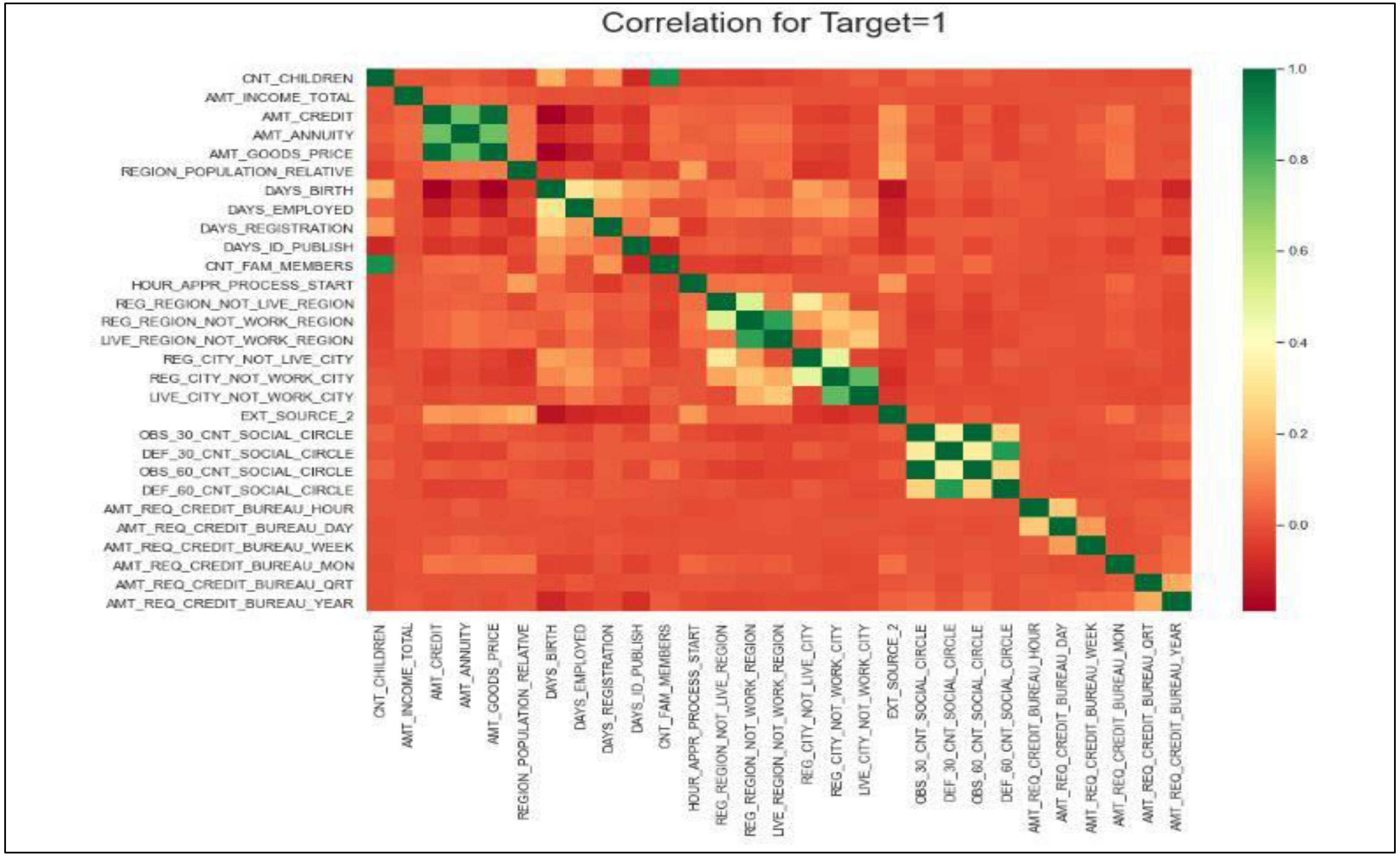
❖ Credit amount is inversely proportional to the number of children client have, means Credit amount is higher for less children count client have and vice-versa.

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

❖ Credit amount is higher to densely populated area.

❖ The income is also higher in densely populated area.

Correlation for target = 1



Plot: 11

In Continuation..

In Continuation ..

- ***Conclusions from the graph:***

- Same like the target=0 heatmap above, adding some other points from this heatmap.

- 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.

Top 10 Correlations for Target = 0

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_FAM_MEMBERS	CNT_CHILDREN	0.893276
CNT_CHILDREN	CNT_FAM_MEMBERS	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		

Plot: 12

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

Top 10 Correlations for Target = 1

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

dtype: float64

Plot: 13

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

Numerical Bivariate Analysis

Plot: 13

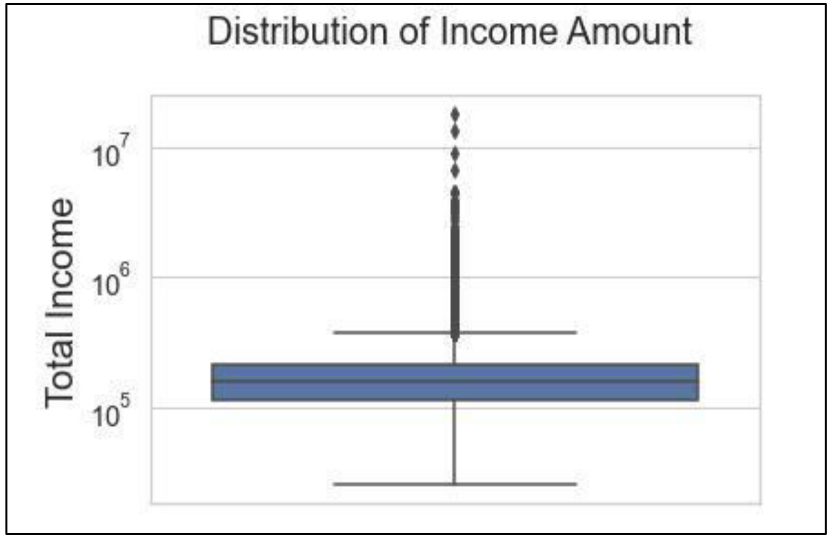
Conclusions from the graph:

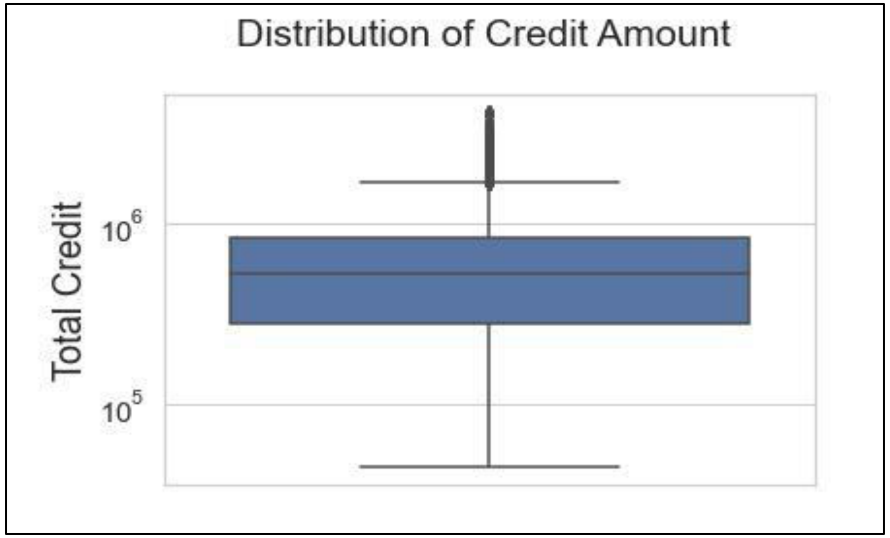
With the scatter plot, we can determine that AMT CREDIT and AMT GOODS PRICE are highly correlated, which means if increase in goods price, the credit increased directly and vice versa.

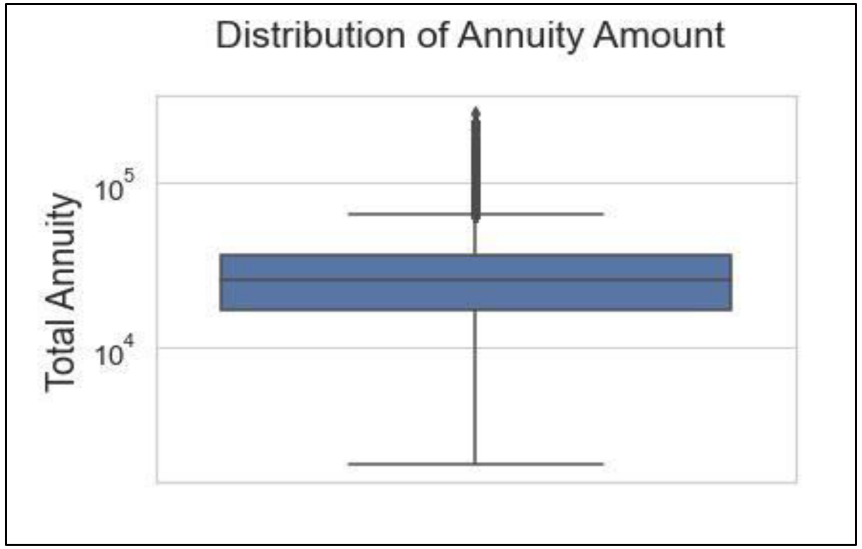
Finding Outliers

Univariate Analysis

Outliers for Target = 0

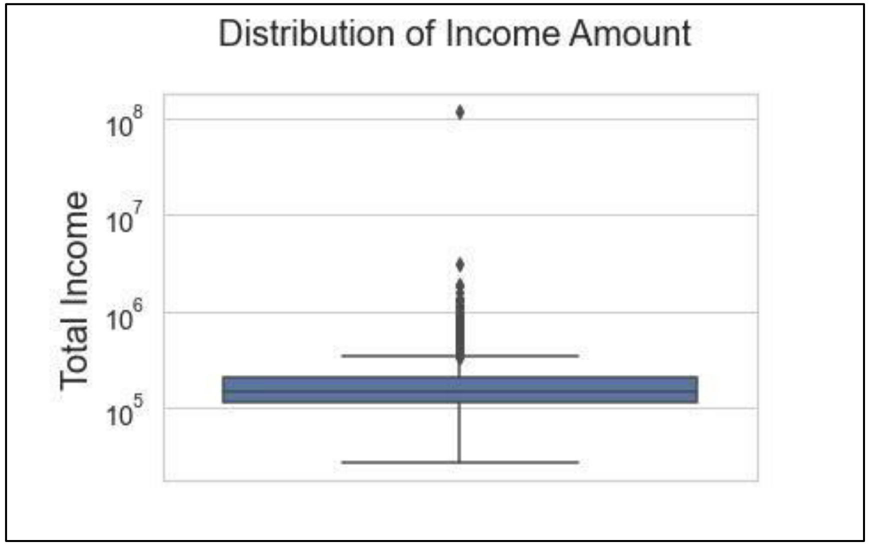


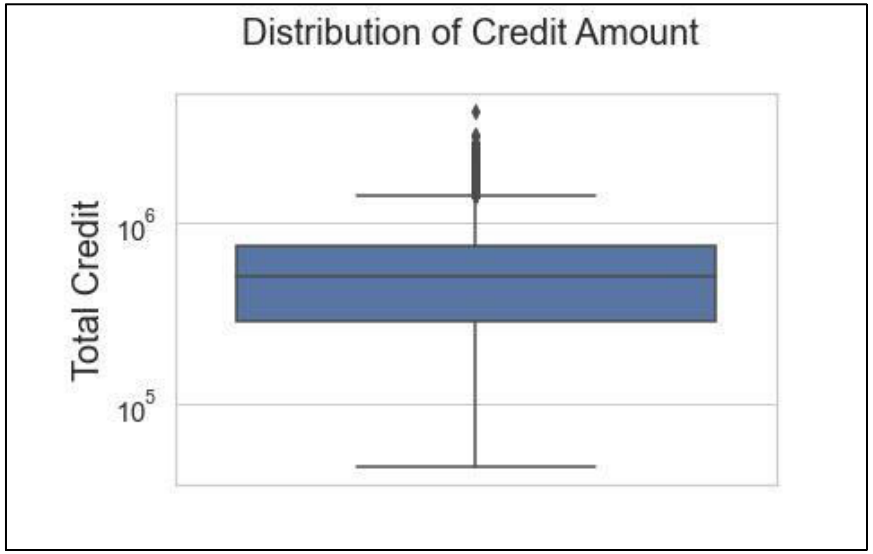


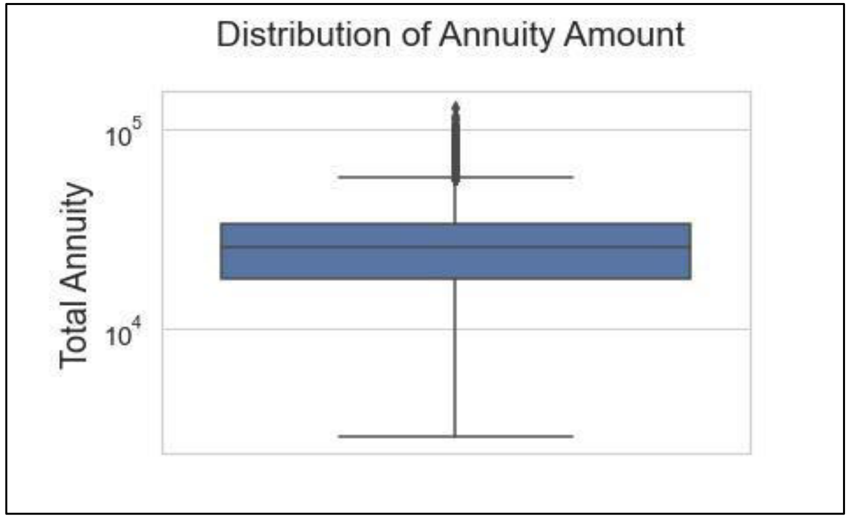


Plot: 14

Outliers for Target = 1



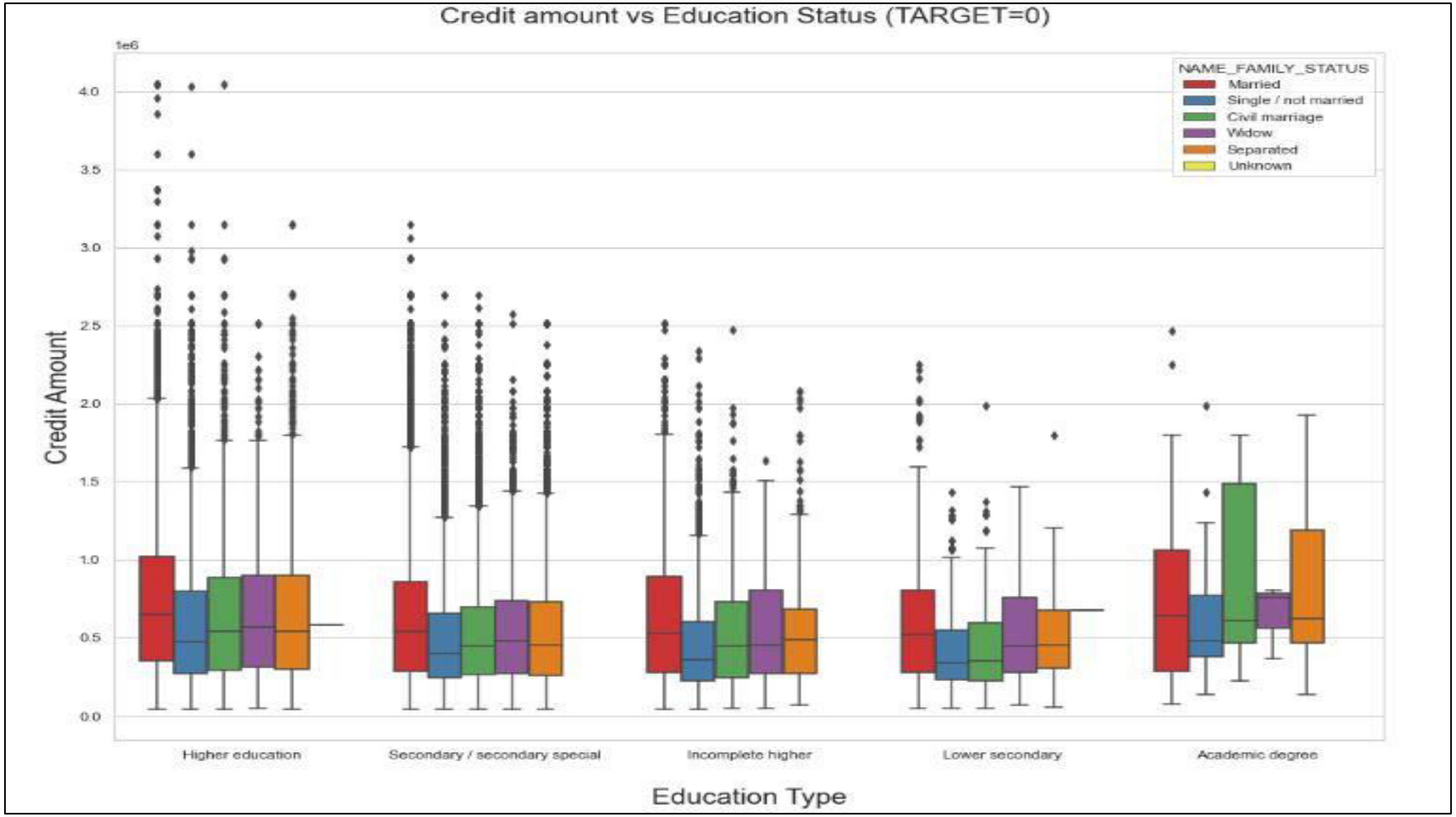




Plot: 15

Finding Outliers

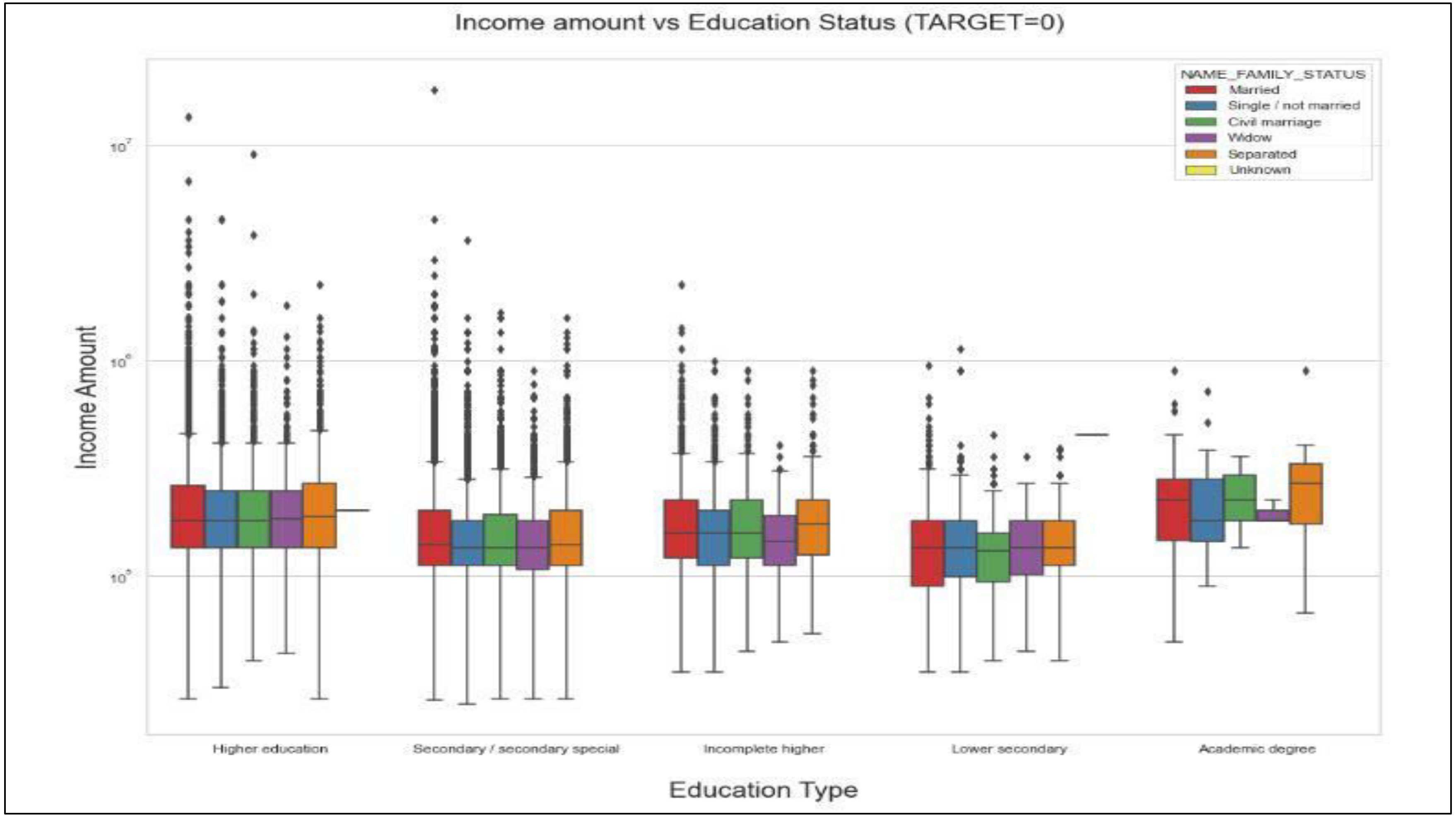
Multivariate Analysis



Plot: 16

Conclusions from the graph:

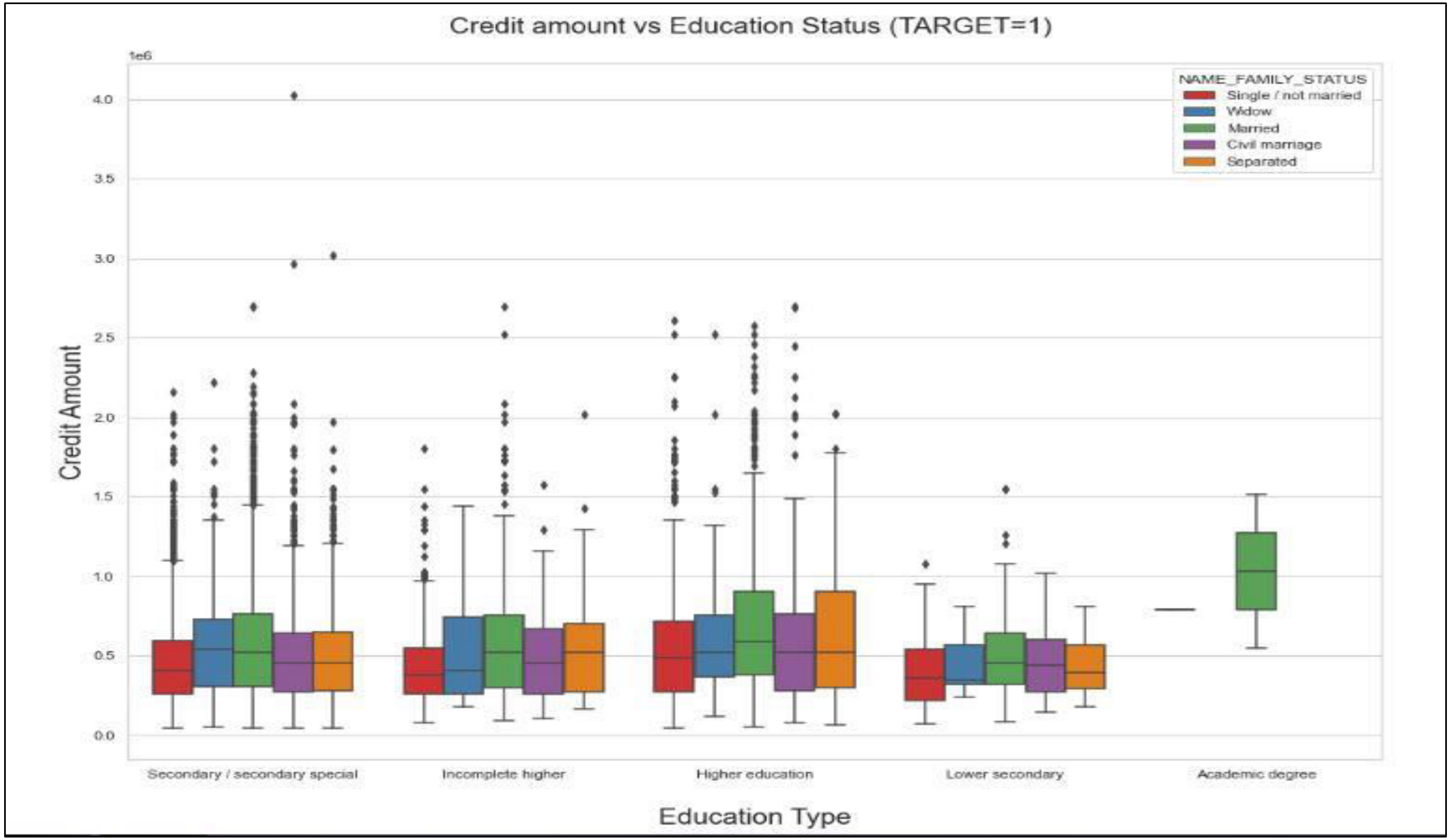
From the above box plot we can conclude that Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others. Also, higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers. Civil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



Plot: 17

Conclusions from the graph:

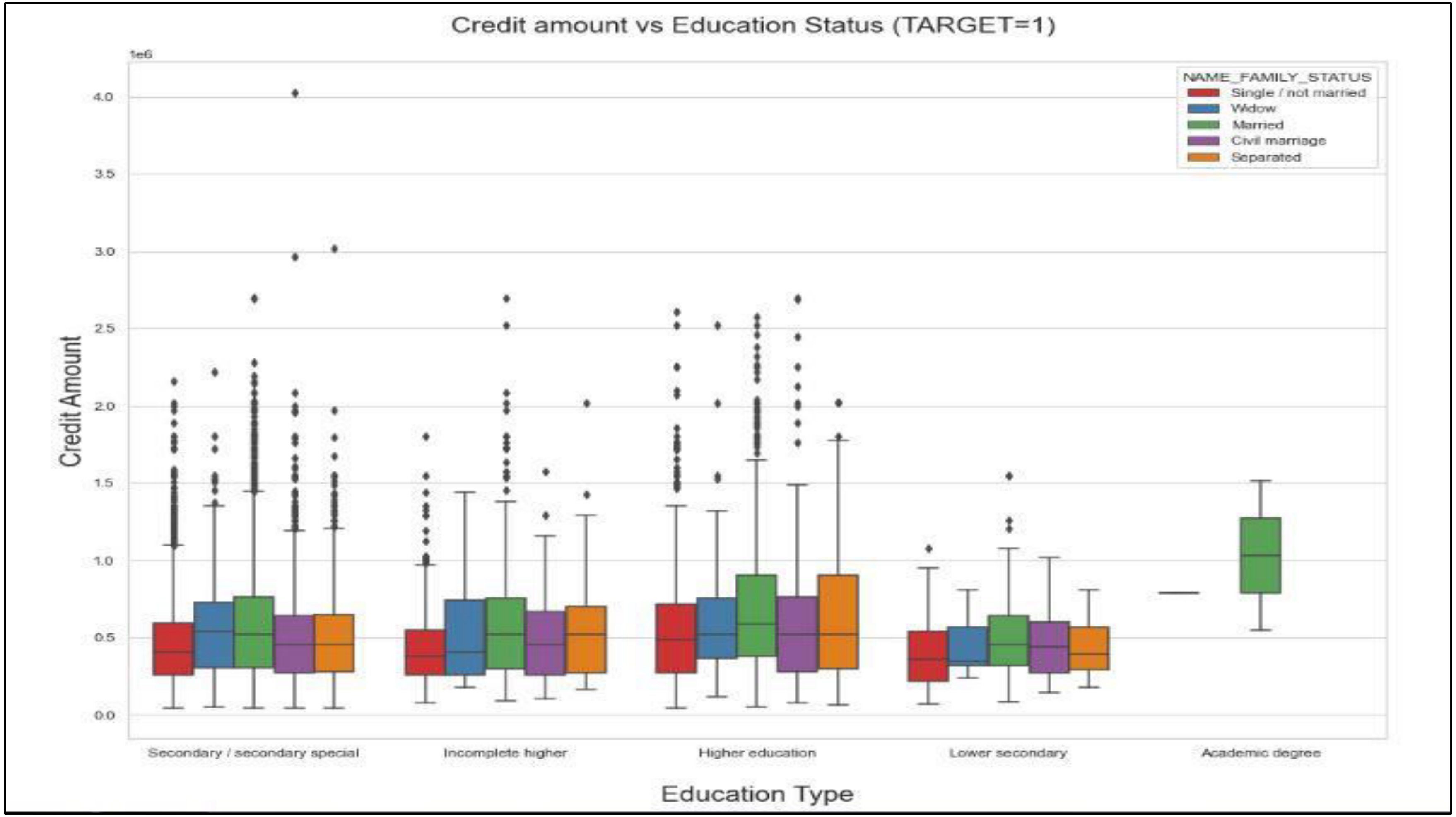
From above boxplot for Education type 'Higher education' the income amount is mostly equal with family status. It does contain many outliers. Less outlier are having for Academic degree but there income amount is little higher that Higher education. Lower secondary of civil marriage family status are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



Plot: 18

Conclusions from the graph:

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.



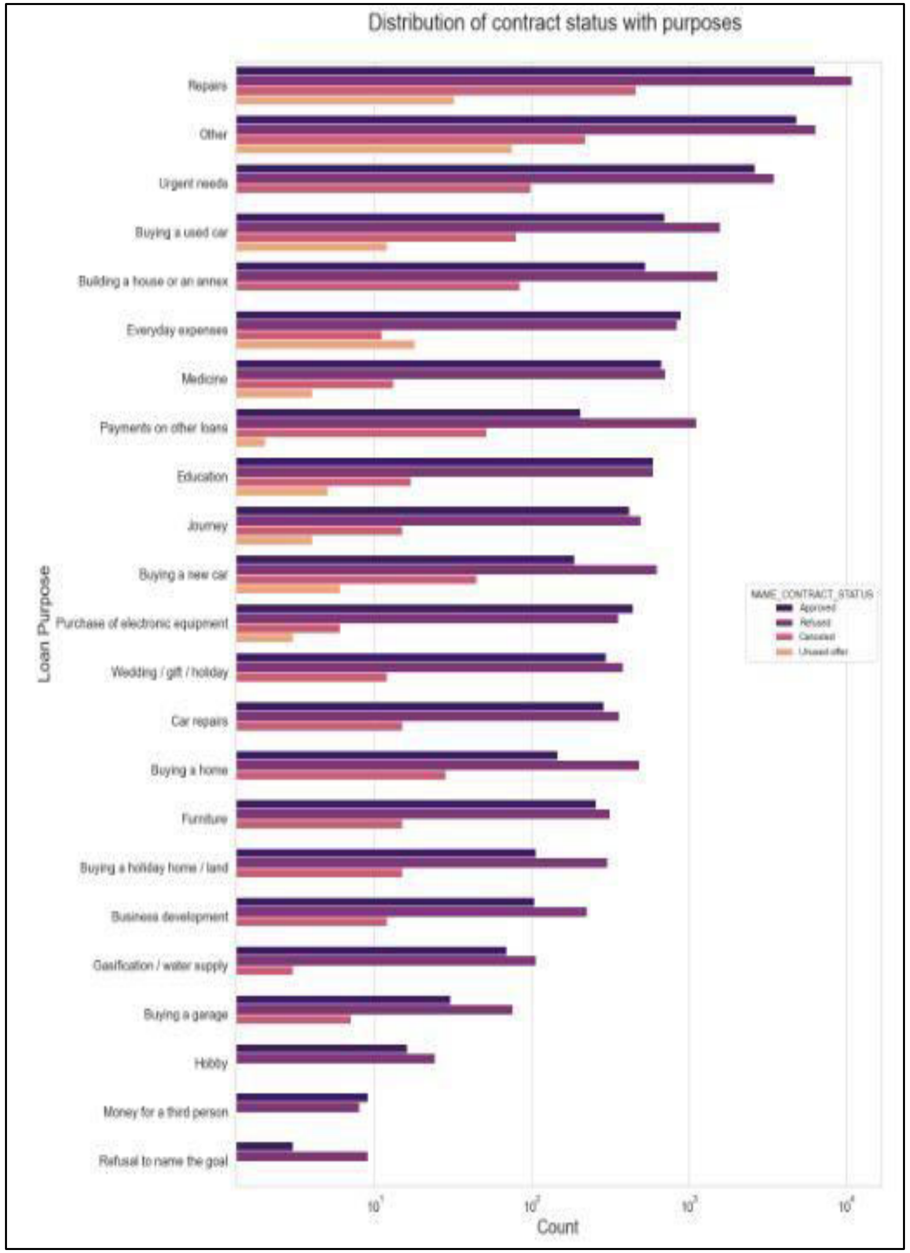
Plot: 19

Conclusions from the graph:

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.

Conclusions 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.

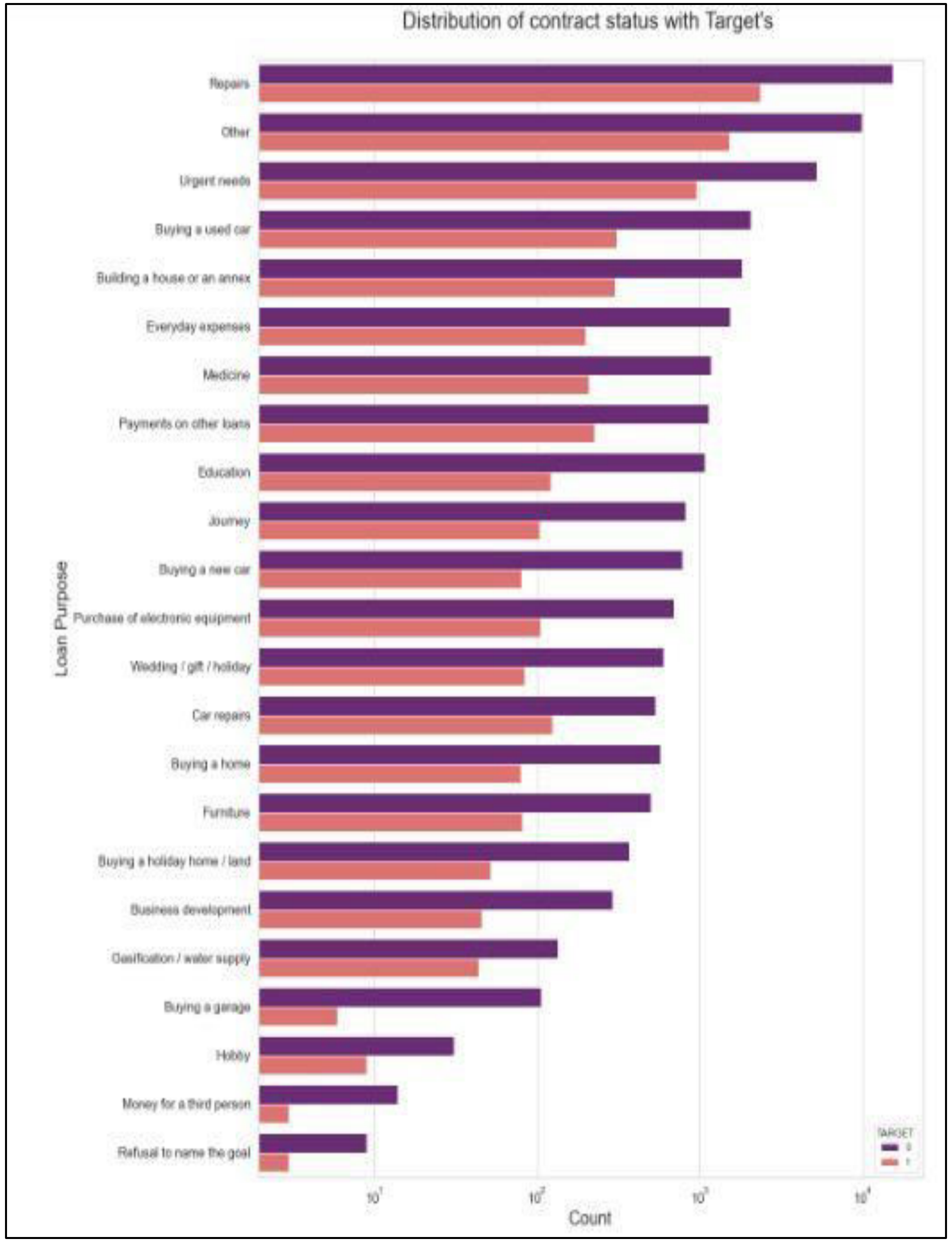


Plot: 20

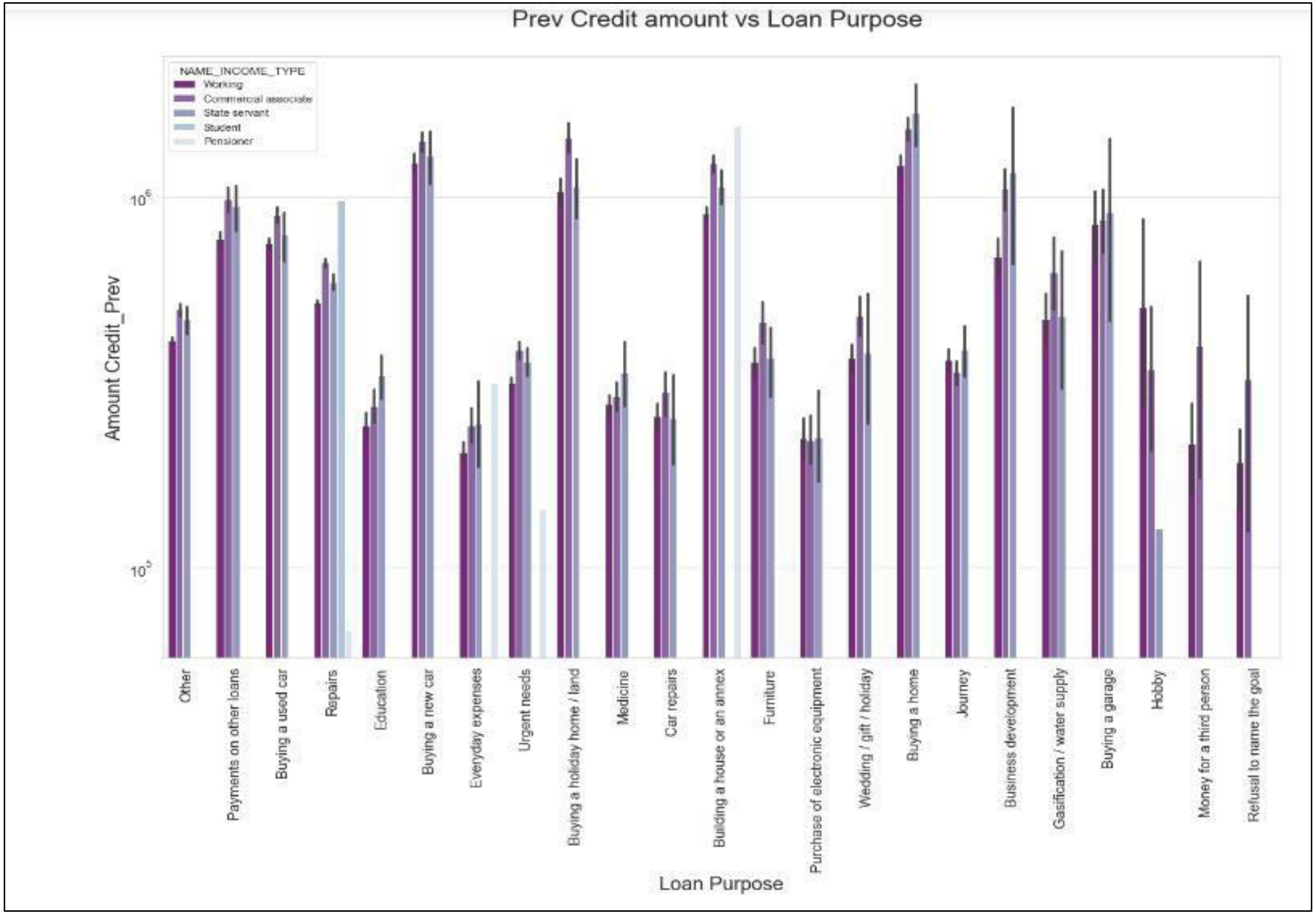
Conclusions from the graph:

- 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.



Plot: 21



Plot: 22

In Continuation..

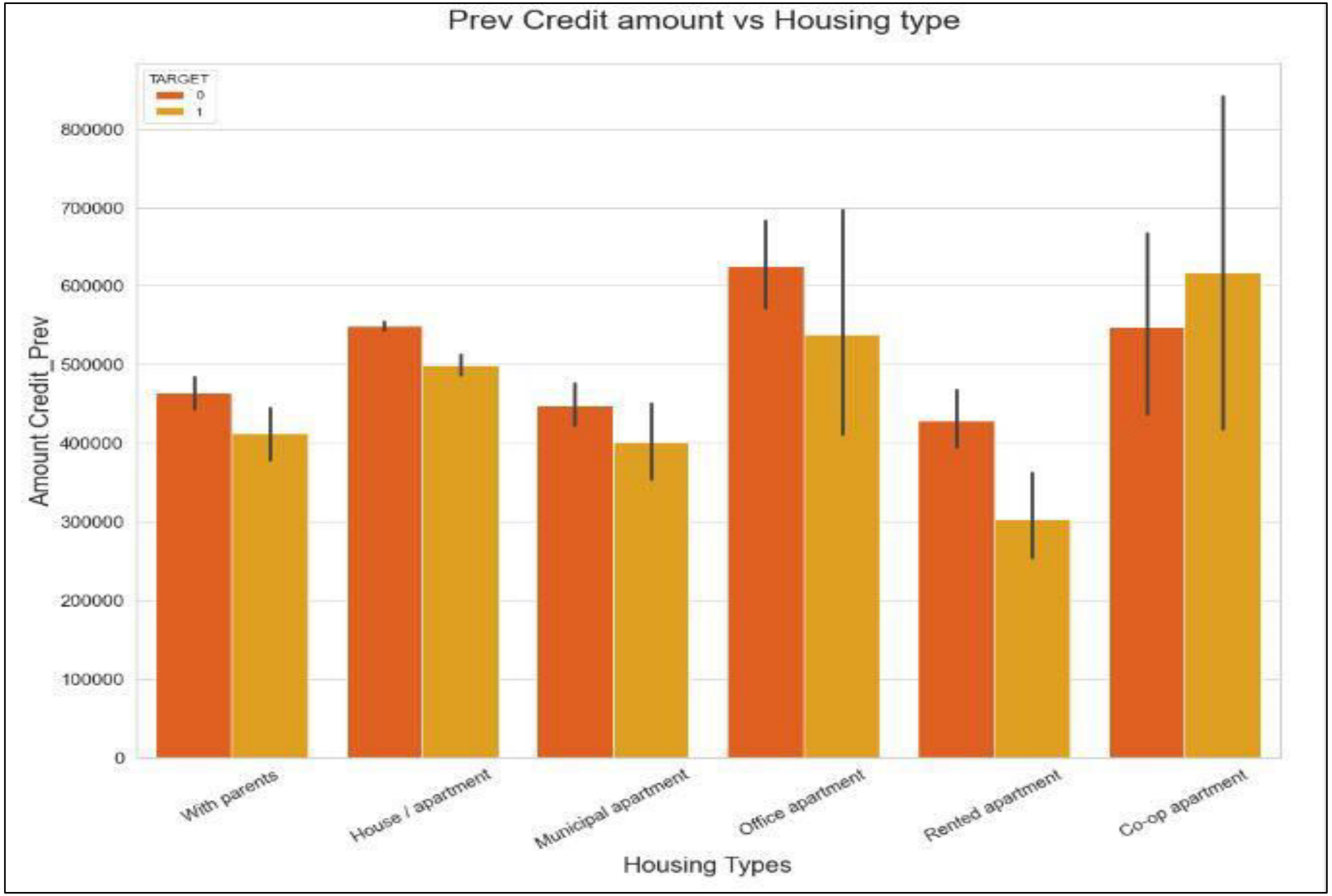
In Continuation ..

- ***Conclusions 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.



Plot: 22

In Continuation..

In Continuation ..

- ***Conclusions 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 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.