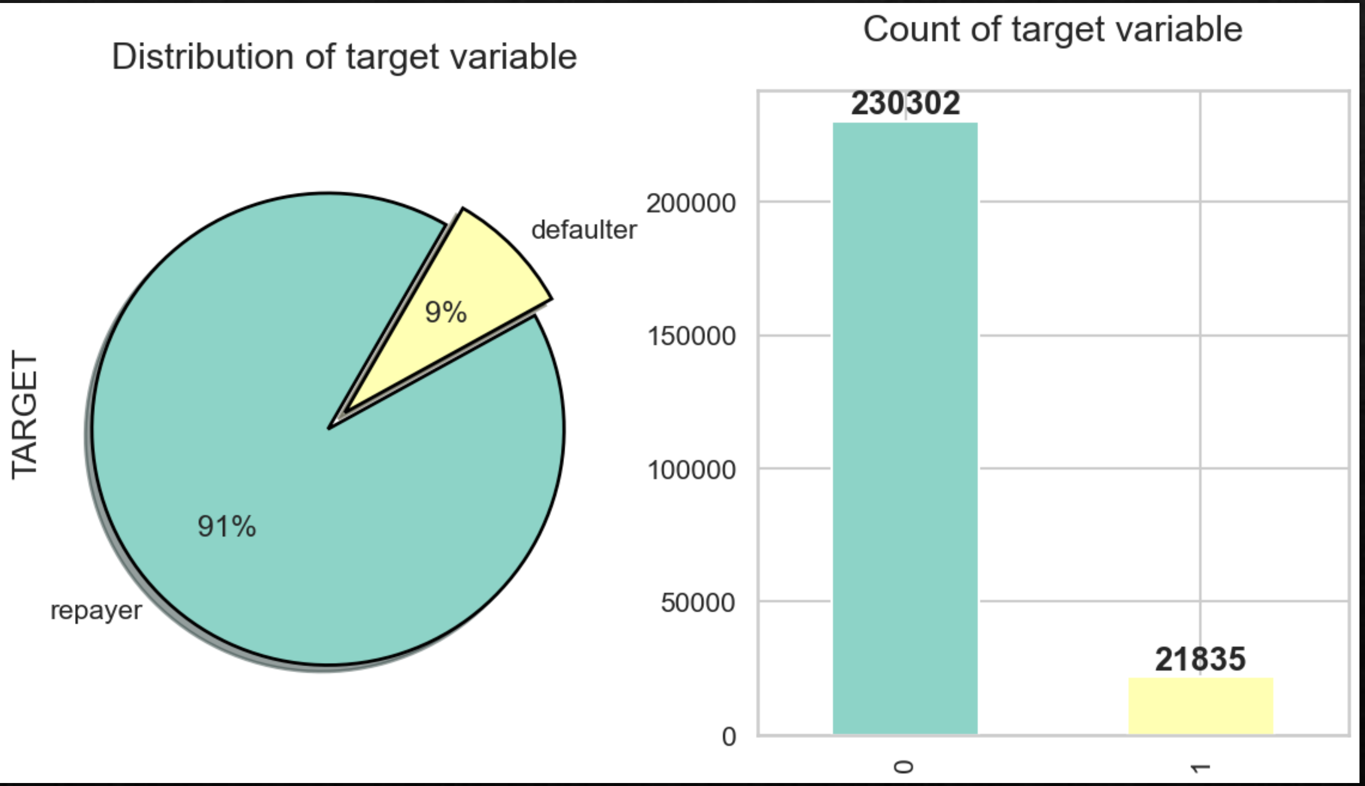
Credit EDA Assignment

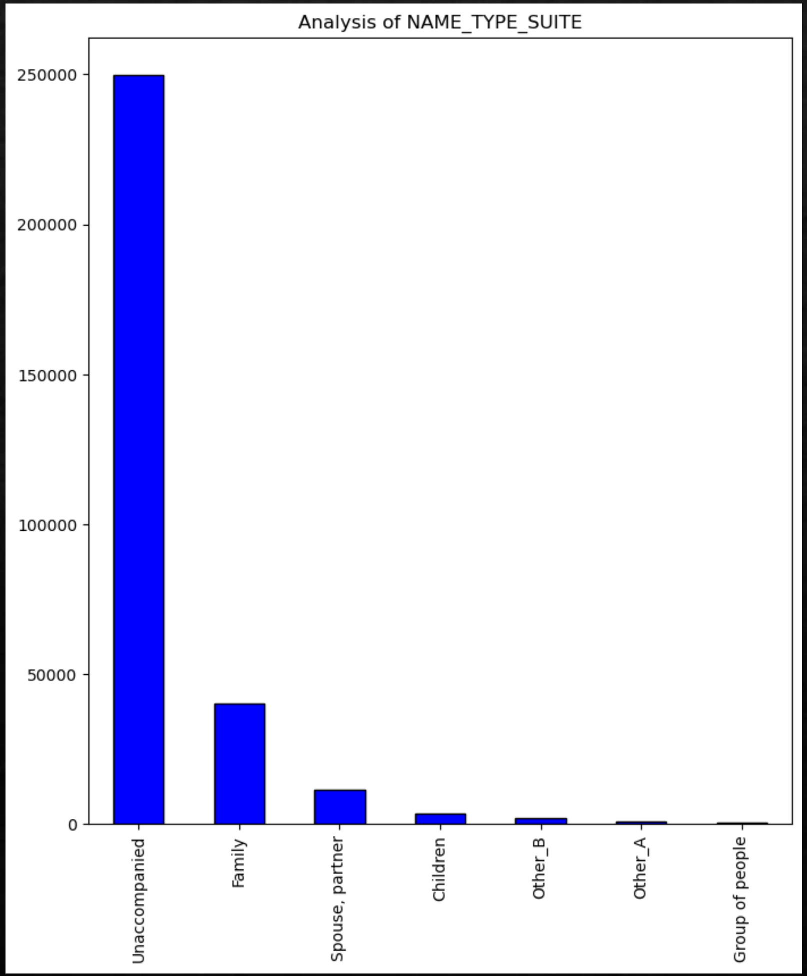
Sree Raghuma Y

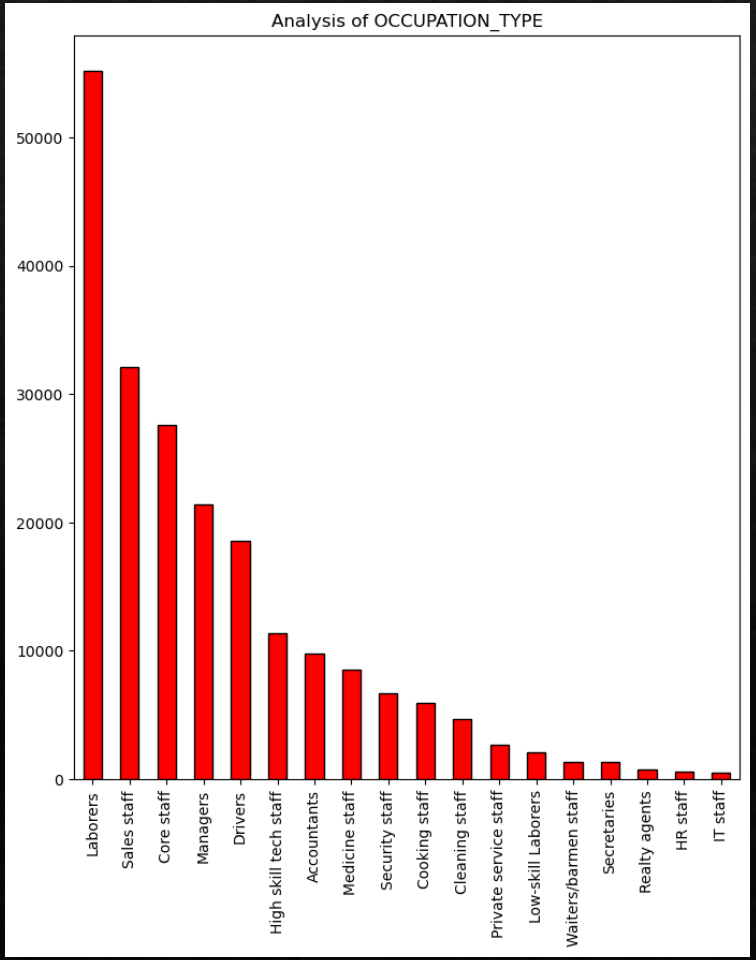
Imbalance in Target Variable

- Target Data is highly imbalanced with a ratio of 92:8. Most of the loans were paid back on time (target = 0).
- Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

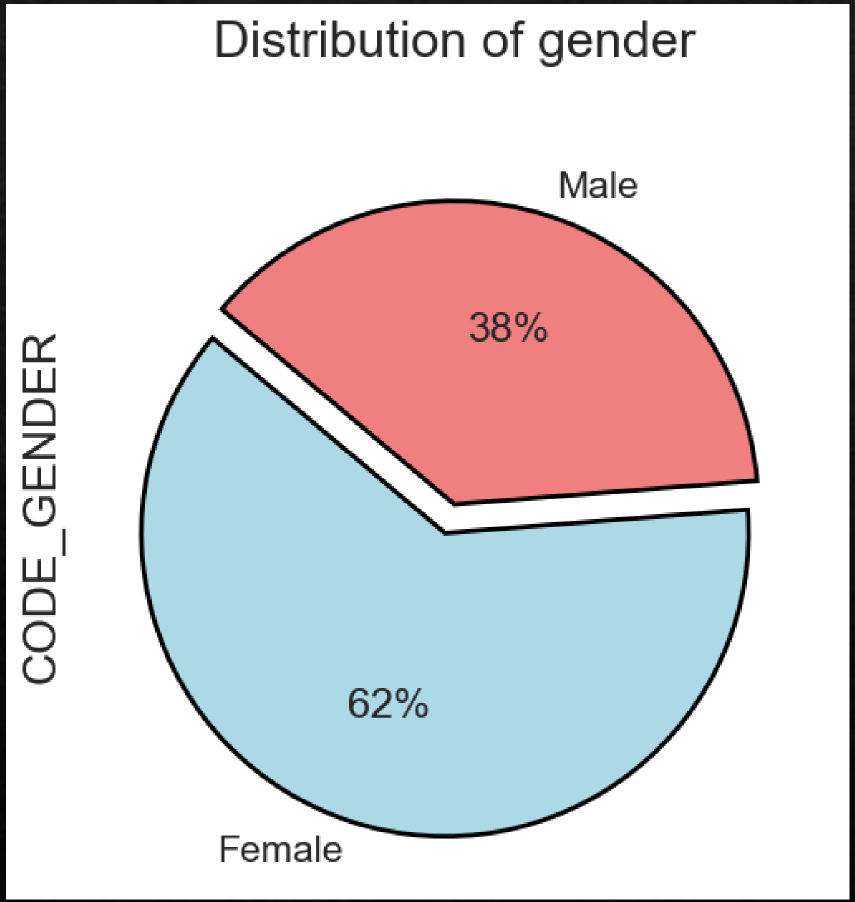


Analysis on Occupation_Type and Name_type_suite





Analysis on Gender



Univariant Analysis on Categorical Data

Target – 0 (With No Payment Difficulties)

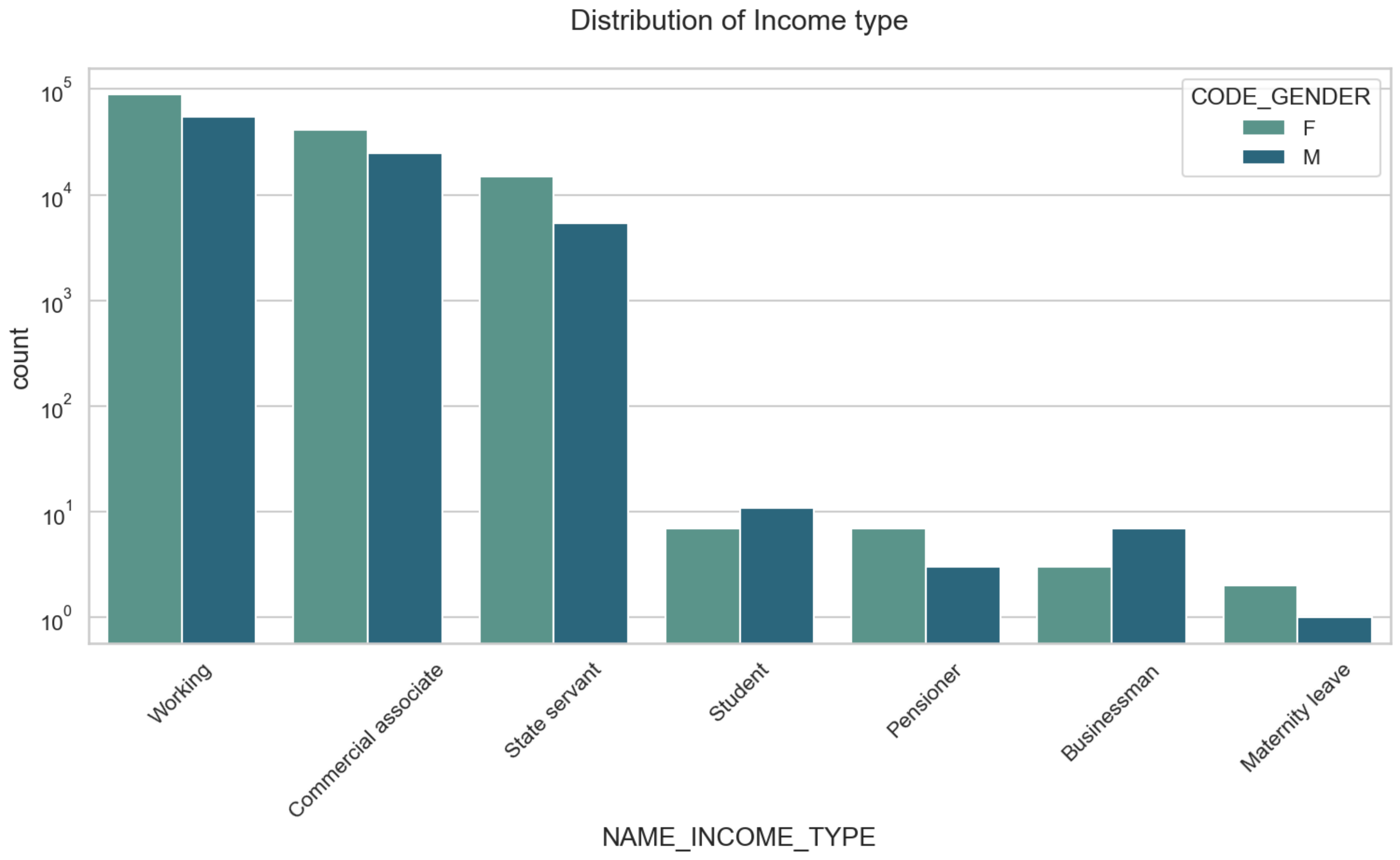
Analysis on Income Type

Here are the key observations drawn from the graph:

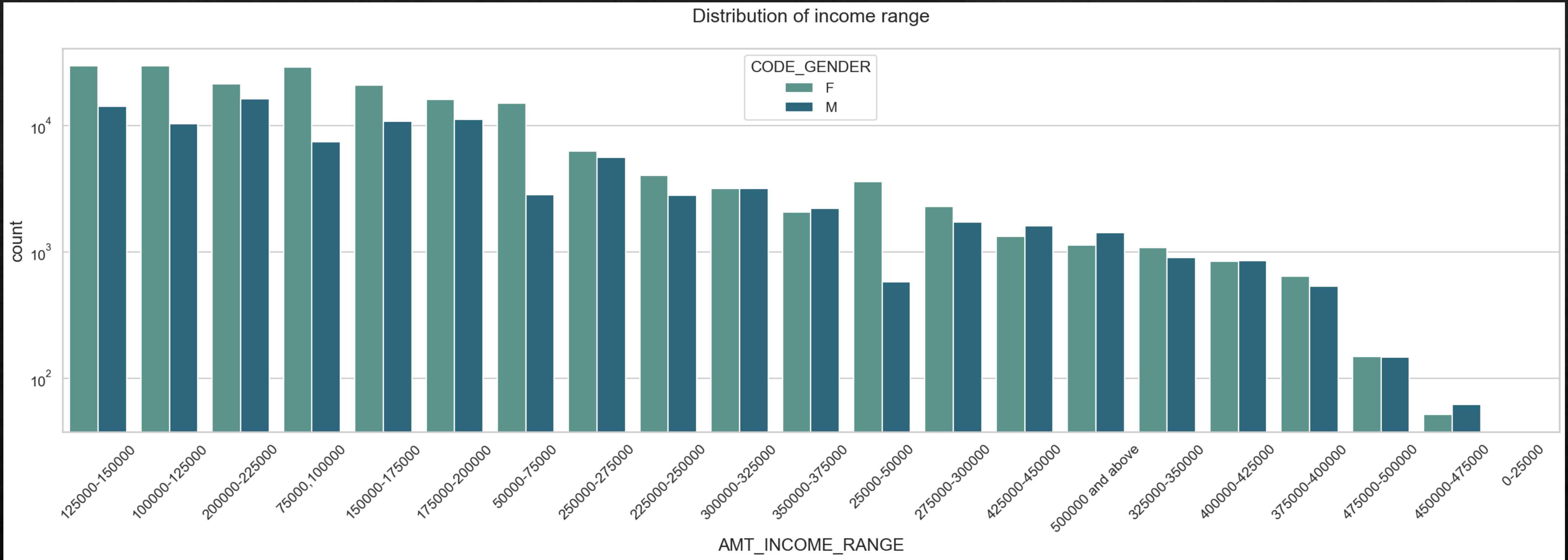
1. The number of credits is notably higher for income types 'working,' 'commercial associate,' and 'State Servant' compared to the other categories.

2. Within these income types, females hold a greater number of credits than males.

3. Conversely, there is a lower count of credits associated with income types such as 'student,' 'pensioner,' 'Businessman,' and 'Maternity leave.'



Analysis on Income Range



Key Takeaways from the Graph:

1. Females outnumber males.

2. The income range between 100,000 and 200,000 exhibits the highest number of credit cases with no payment difficulties.

3. The graph underscores that females dominate in the number of credits within this income range.

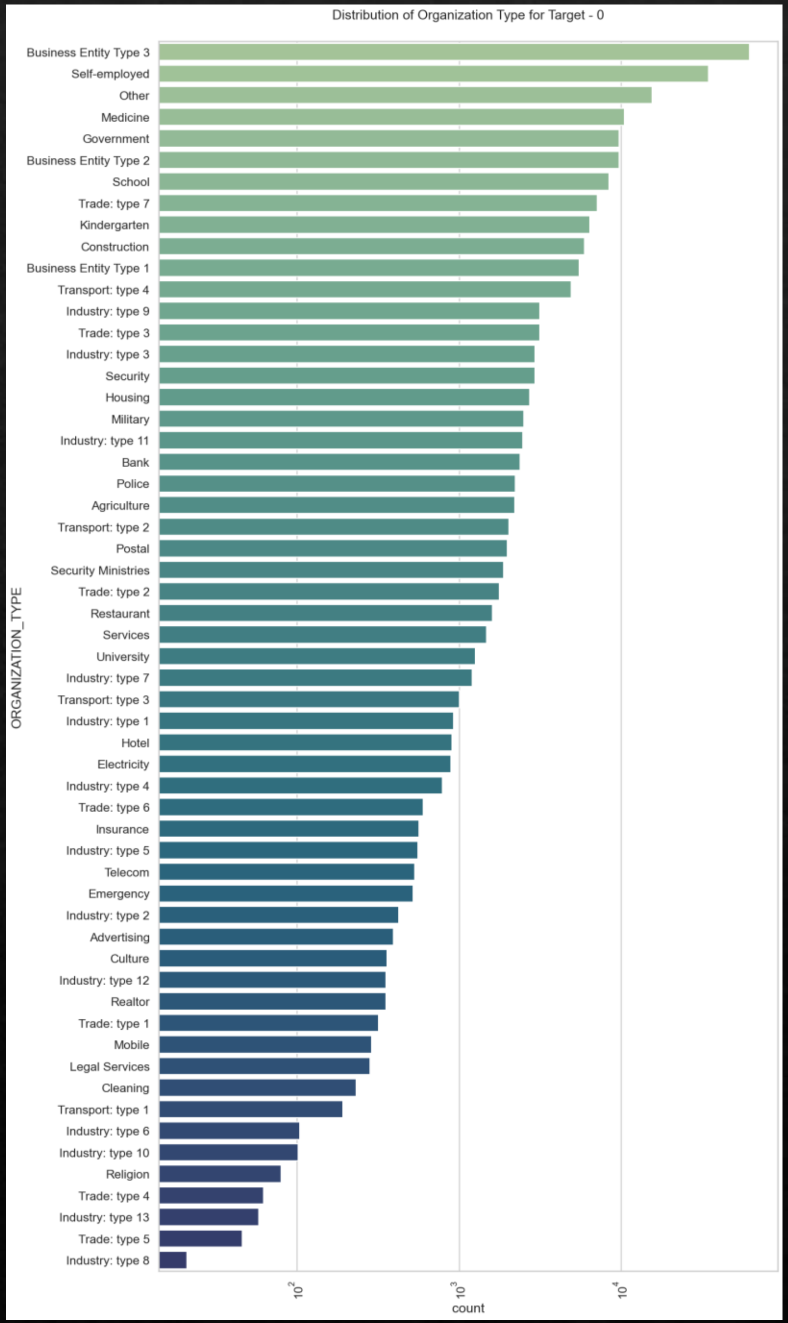
4. There is a notably low count of credits for individuals with an income range of 400,000 and above.

Distribution of Occupation Type

Here are the key findings from the graph:

1. Clients who have applied for credits represent a majority of organization types, including 'Business entity Type 3,' 'Self-employed,' 'Other,' 'Medicine,' and 'Government.'

2. Conversely, there are fewer clients from organization types such as 'Industry type 8,' 'type 6,' 'type 10,' 'religion and trade type 5,' and 'type 4.'

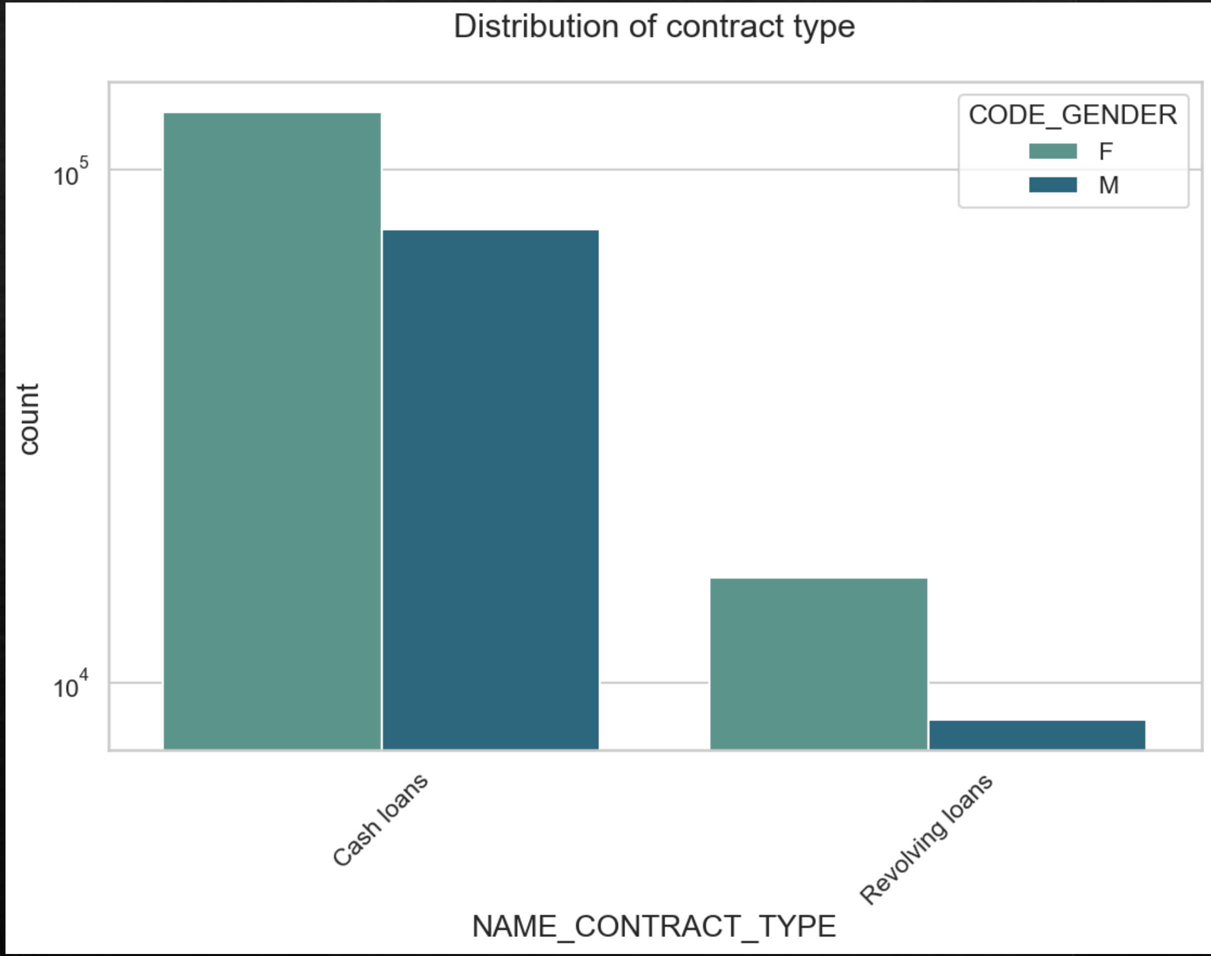


Distribution of Contract Type

Here are the key observations from the graph:

1. The 'cash loans' contract type has a higher number of credits compared to the 'Revolving loans' contract type.

2. Additionally, within both contract types, females lead in applying for credits



Univariant Analysis on Categorical Data

Target – 1 (With Payment Difficulties)

Analysis on Income Type

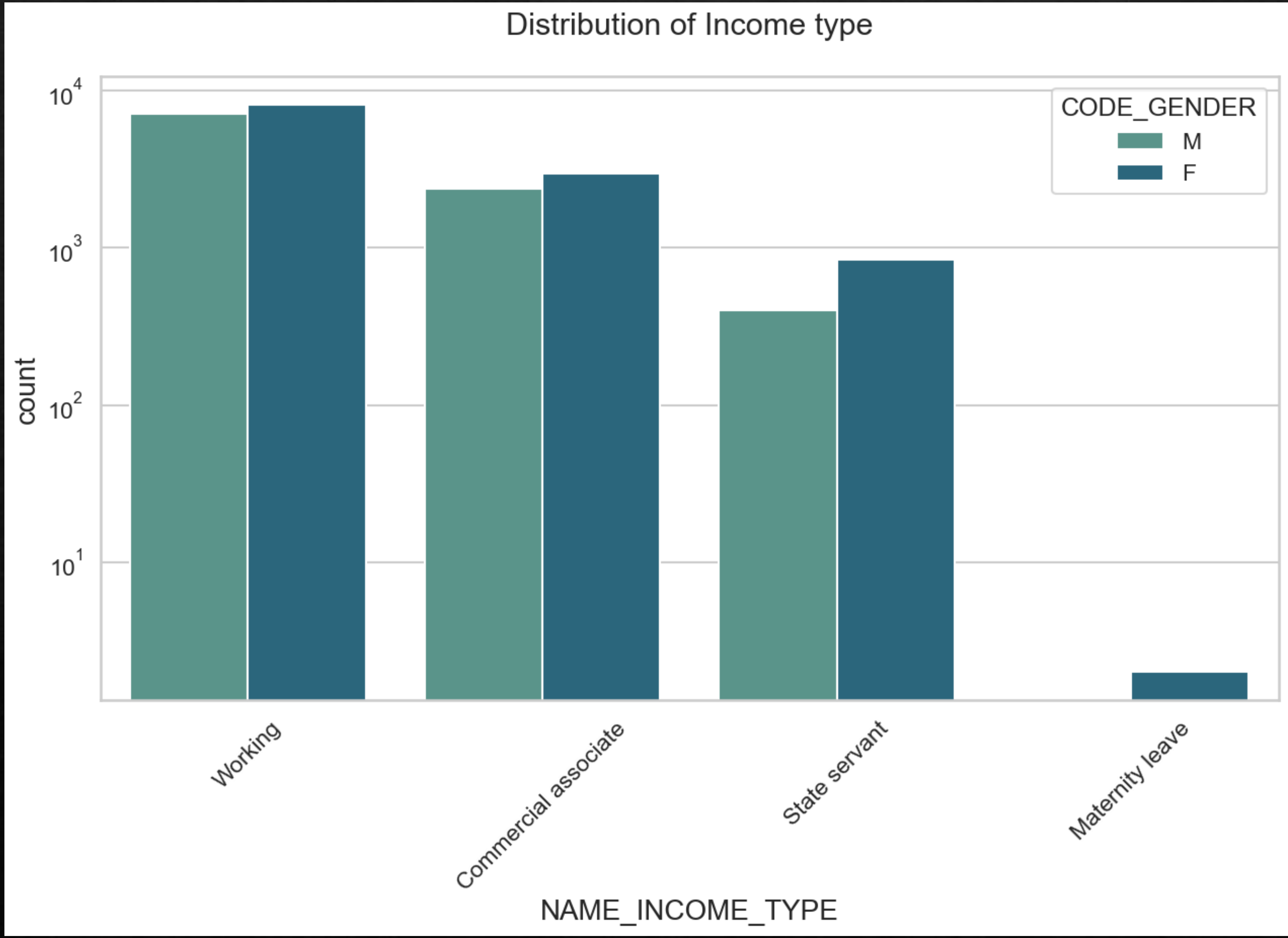
Here are the key conclusions drawn from the above graph:

1. The number of credits is notably higher for income types 'working,' 'commercial associate,' and 'State Servant' compared to 'Maternity leave.'

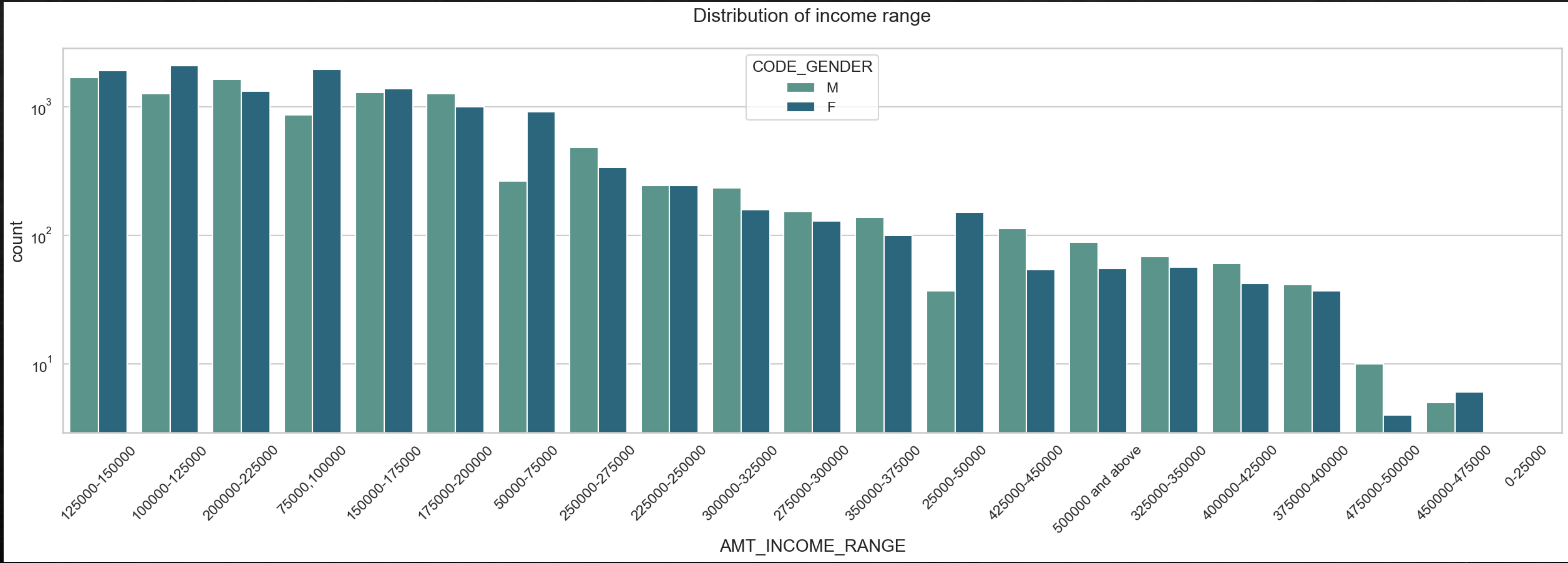
2. Within these income types, females have a greater number of credits than males.

3. Conversely, there is a lower count of credits associated with the income type 'Maternity leave.'

4. For organization type 1, there are no instances of 'student,' 'pensioner,' and 'Businessman' income types, indicating that they do not have any late payments.



Analysis on Income Range



Here are the key takeaways from the graph:

1. The count of males is greater than that of females.

2. The income range from 100,000 to 200,000 exhibits a higher number of credits.

3. Within this income range, the graph illustrates that males outnumber females in terms of credits.

4. Conversely, there is a notably low count of credits for individuals with an income range of 400,000 and above.

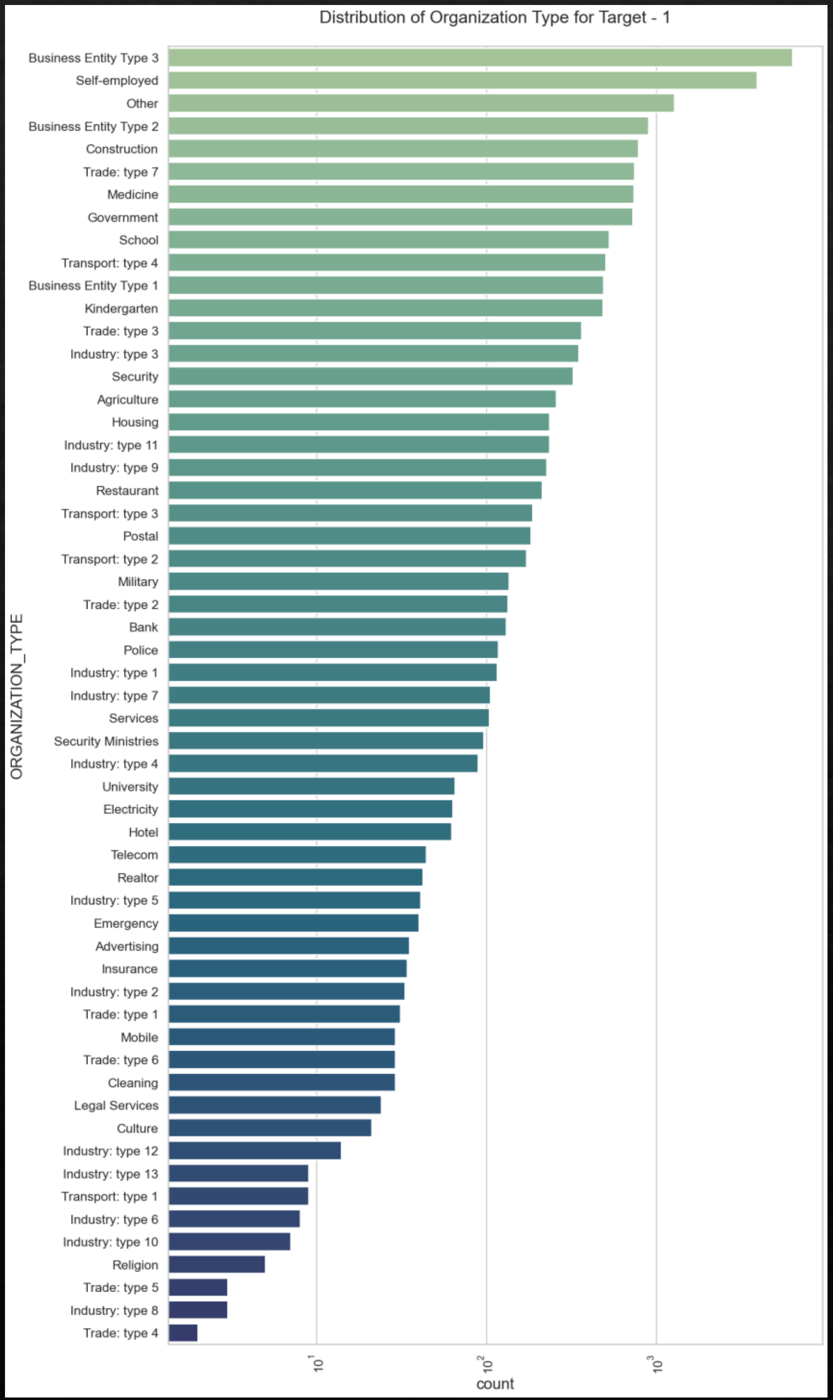
Distribution of Occupation Type

Here are the key findings from the above graph:

1. Clients who have applied for credits come from various organization types, including 'Business entity Type 3,' 'Self-employed,' 'Other,' 'Medicine,' and 'Government.'

2. Conversely, there are fewer clients from organization types such as 'Industry type 8,' 'type 6,' 'type 10,' 'religion and trade type 5,' and 'type 4.'

3. The distribution of organization type for type 0 is similar to the mentioned pattern.



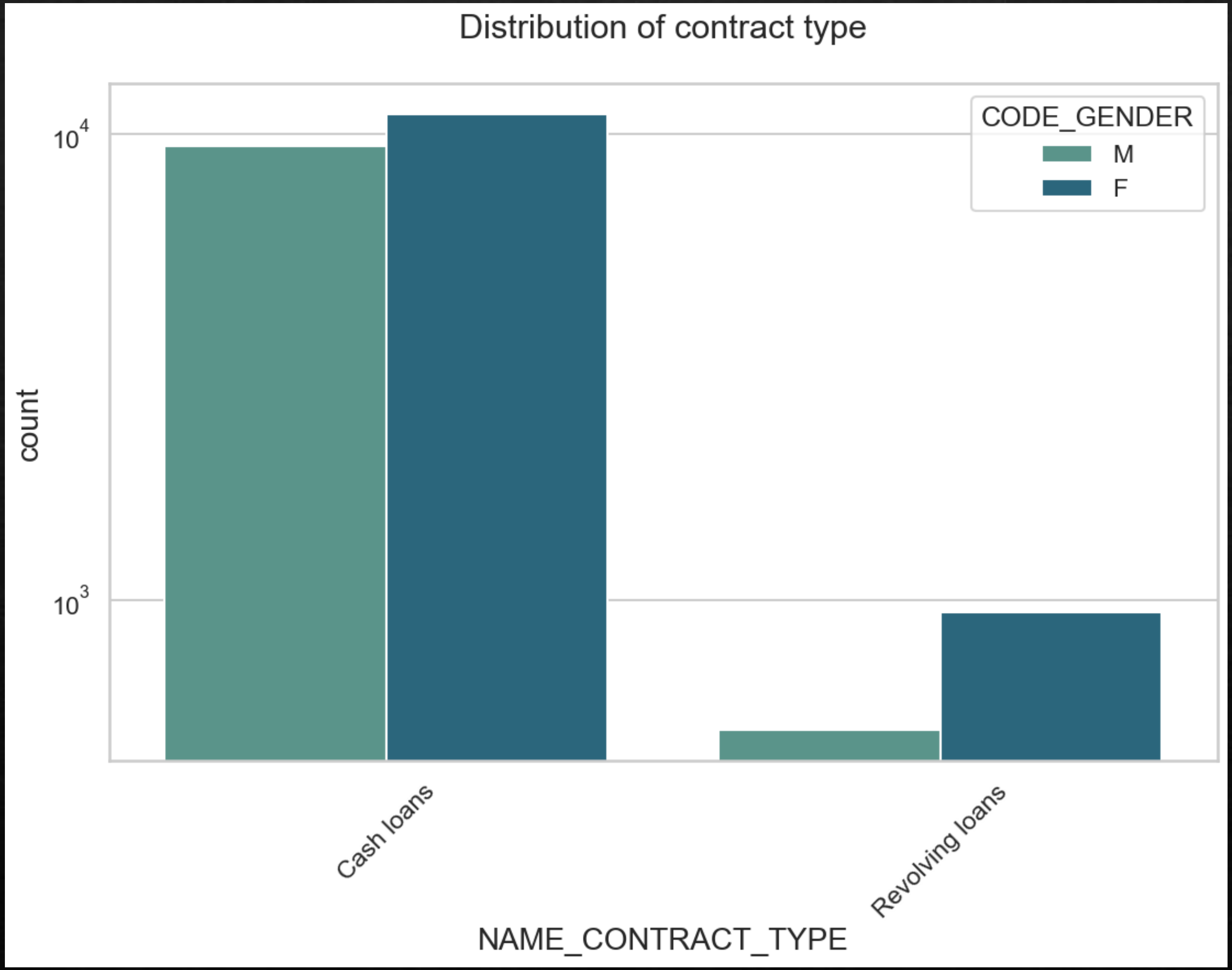
Distribution of Contract Type

Here are the key observations from the above graph:

1. The 'cash loans' contract type has a higher number of credits compared to the 'Revolving loans' contract type.

2. Additionally, within both contract types, females lead in applying for credits.

3. For organization type 1, there are only female applicants for 'Revolving loans.'



Univariant Analysis on Numerical Variable

Target – 0 (With no Payment Difficulties)

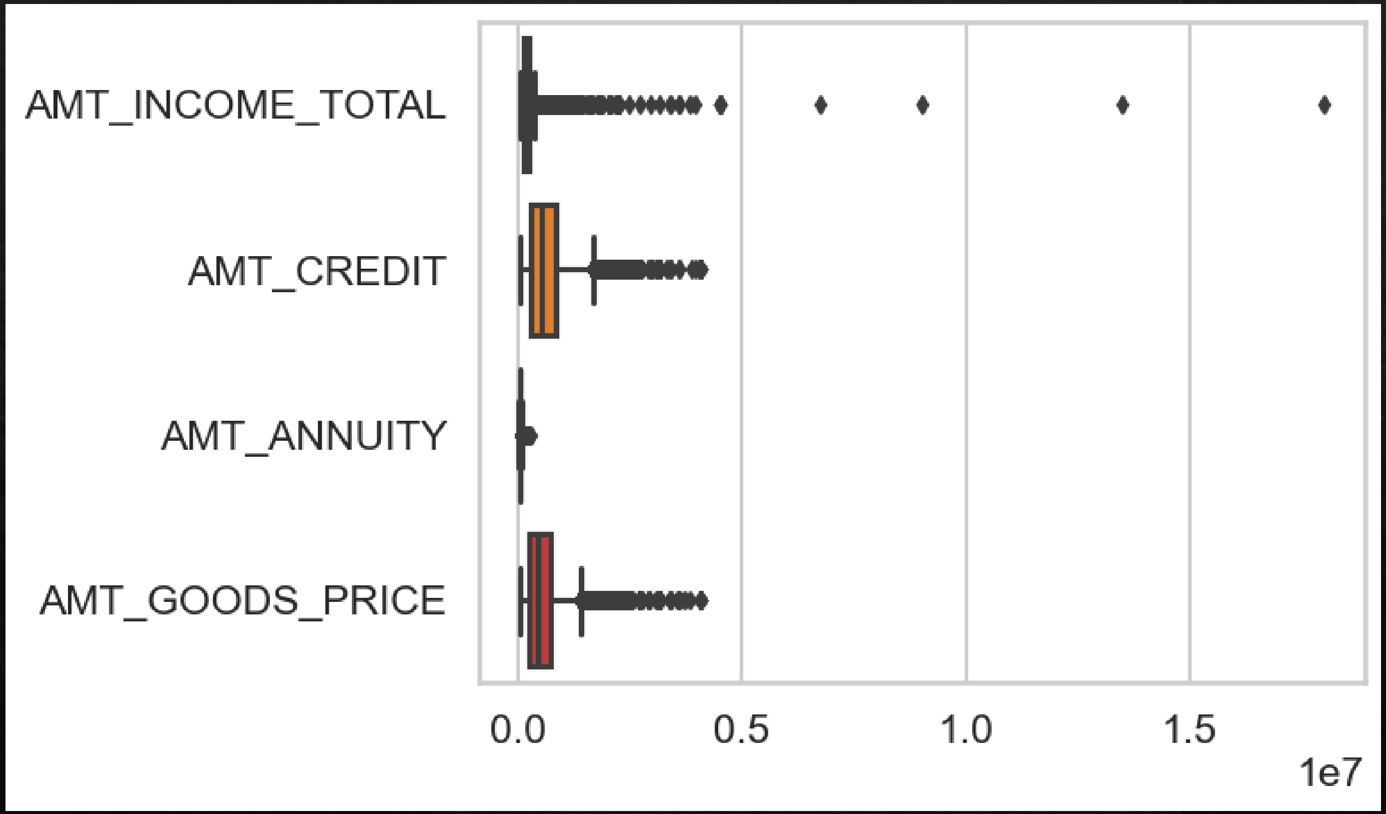
Distribution of Numeric Variables

Here are the key findings from the above graph:

1. Some outliers are noticed in income amount.

2. The third quartiles is very slim for income amount and Amt Annuity

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



Meaning of Variables:

1. AMT_INCOME_TOTAL -Income of the client

2. AMT_CREDIT - Credit amount of the loan

3. AMT_ANNUITY - Loan annuity

4. AMT_GOODS_PRICE - For consumer loans it is the price of the goods for which the loan is given

Univariant Analysis on Numerical Variable

Target – 1 (With Payment Difficulties)

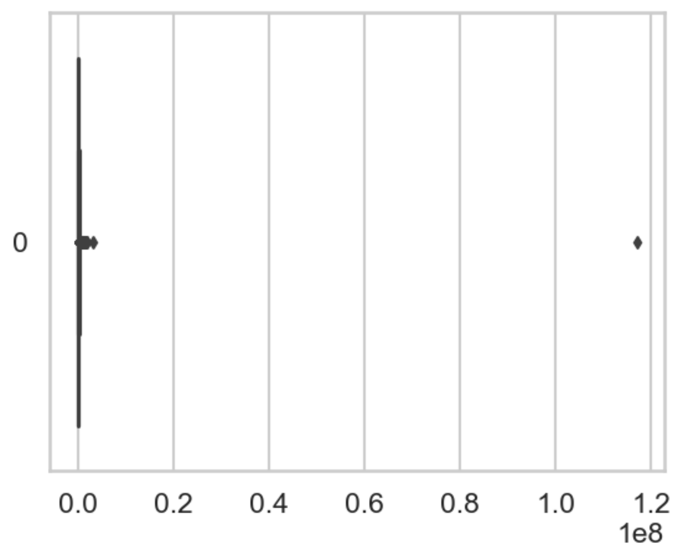
Distribution of Numeric Variables

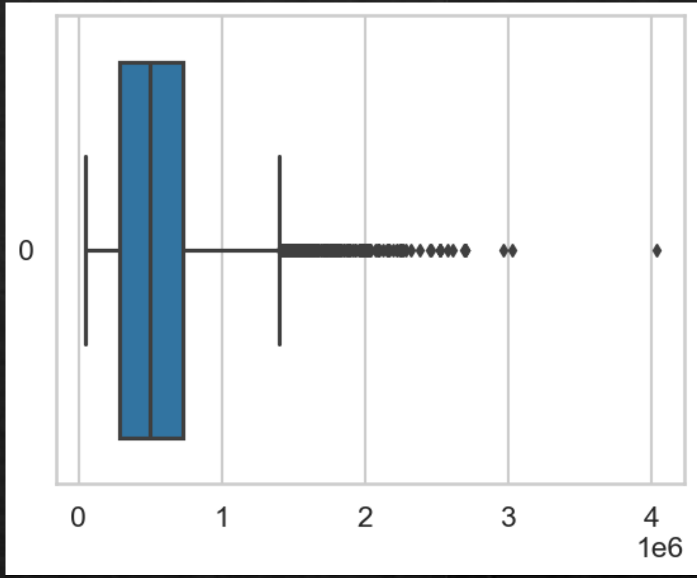
Few points can be concluded from the graph above.

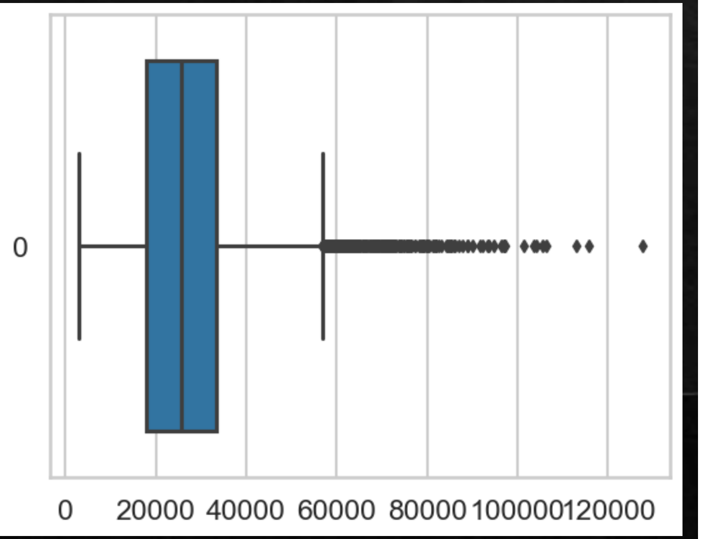
1. Some outliers are noticed on all Numerical variables and for income amount there is one univariant outlier observed.

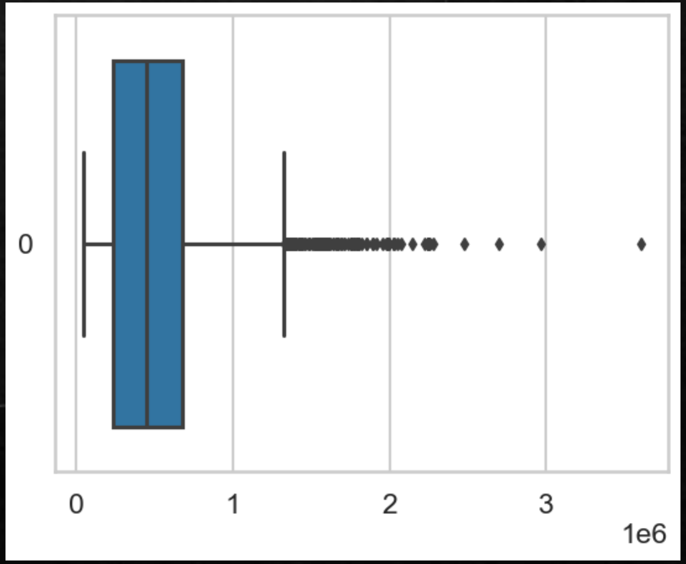
2. quartiles is very slim for income amount.

3. Most of the clients of income are present in first quartile for all Numeric Variables



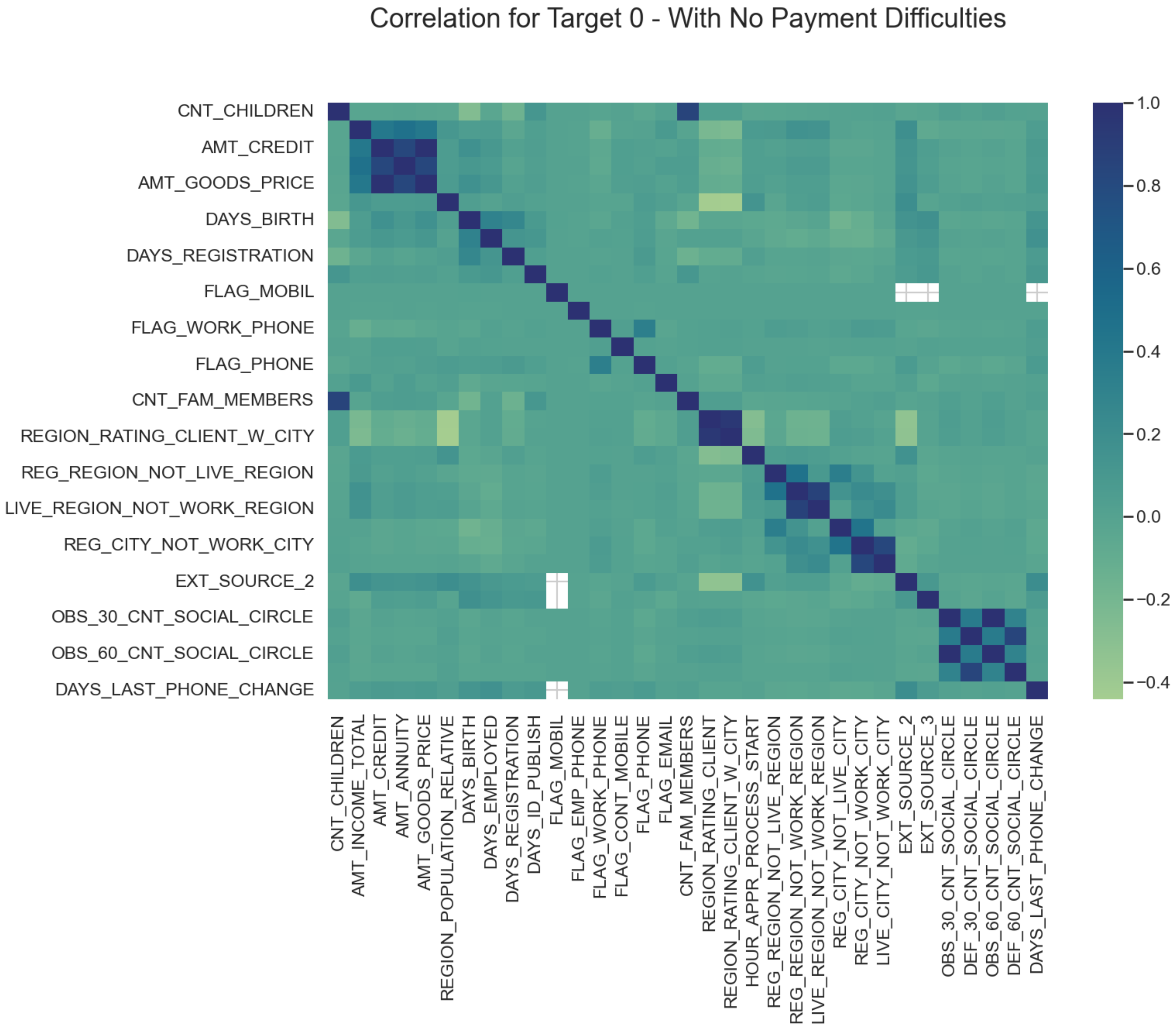






Correlation Between Numerical Variables

Target – 0 (With no Payment Difficulties)


Correlation for Target 0 - With No Payment Difficulties
CNT_CHILDREN
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_REGISTRATION
FLAG_MOBIL
FLAG_WORK_PHONE
FLAG_PHONE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT_W_CITY
REG_REGION_NOT_LIVE_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_WORK_CITY
EXT_SOURCE_2
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
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
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Based on the correlation heatmap observations you provided, here is a summary of the key points:

1. Credit amount is inversely proportional to the date of birth, indicating that younger clients tend to have higher credit amounts, while older clients have lower credit amounts.

2. Credit amount is inversely proportional to the number of children clients have. Clients with fewer children tend to have higher credit amounts, whereas those with more children have lower credit amounts.

3. Income amount is inversely proportional to the number of children clients have. Clients with fewer children tend to have higher incomes, while those with more children have lower incomes.

4. There is a trend of clients with fewer children residing in densely populated areas.

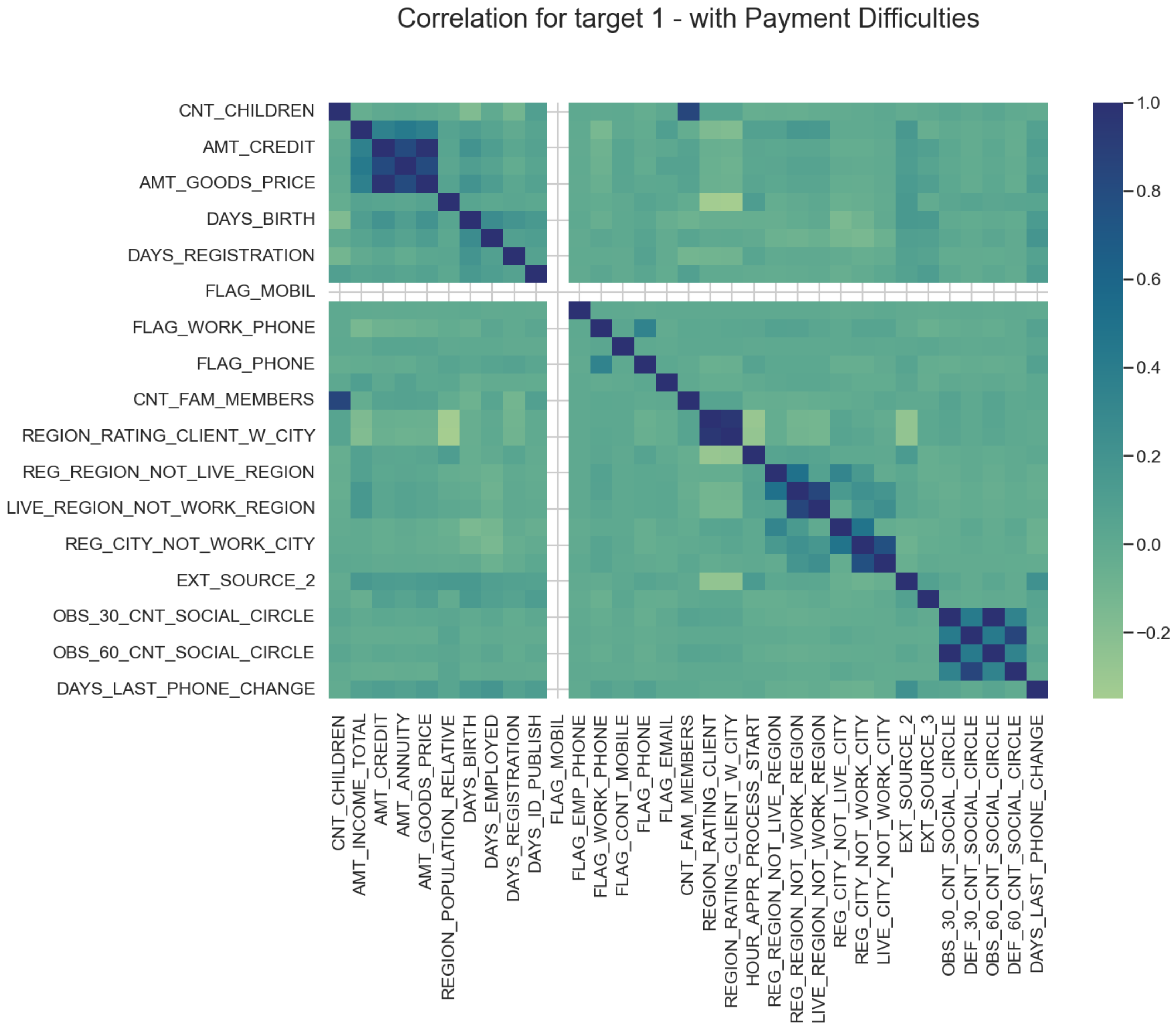
5. Credit amount tends to be higher for clients residing in densely populated areas.

6. Income levels are also higher for clients living in densely populated areas.

These observations highlight various relationships between credit amount, age, number of children, income, and population density within the dataset.

Correlation Between Numerical Variables

Target – 1 (With Payment Difficulties)



Correlation for target 1 - with Payment Difficulties
CNT_CHILDREN
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_BIRTH
DAYS_REGISTRATION
FLAG_MOBIL
FLAG_WORK_PHONE
FLAG_PHONE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT_W_CITY
REG_REGION_NOT_LIVE_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_WORK_CITY
EXT_SOURCE_2
OBS_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
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
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

The heat map for Target 1 shares several similarities with the observations for Target 0, indicating similar trends in certain aspects. However, there are also some distinct differences noted for Target 1:

1. Clients whose permanent address does not match their contact address tend to have fewer children, and vice versa.

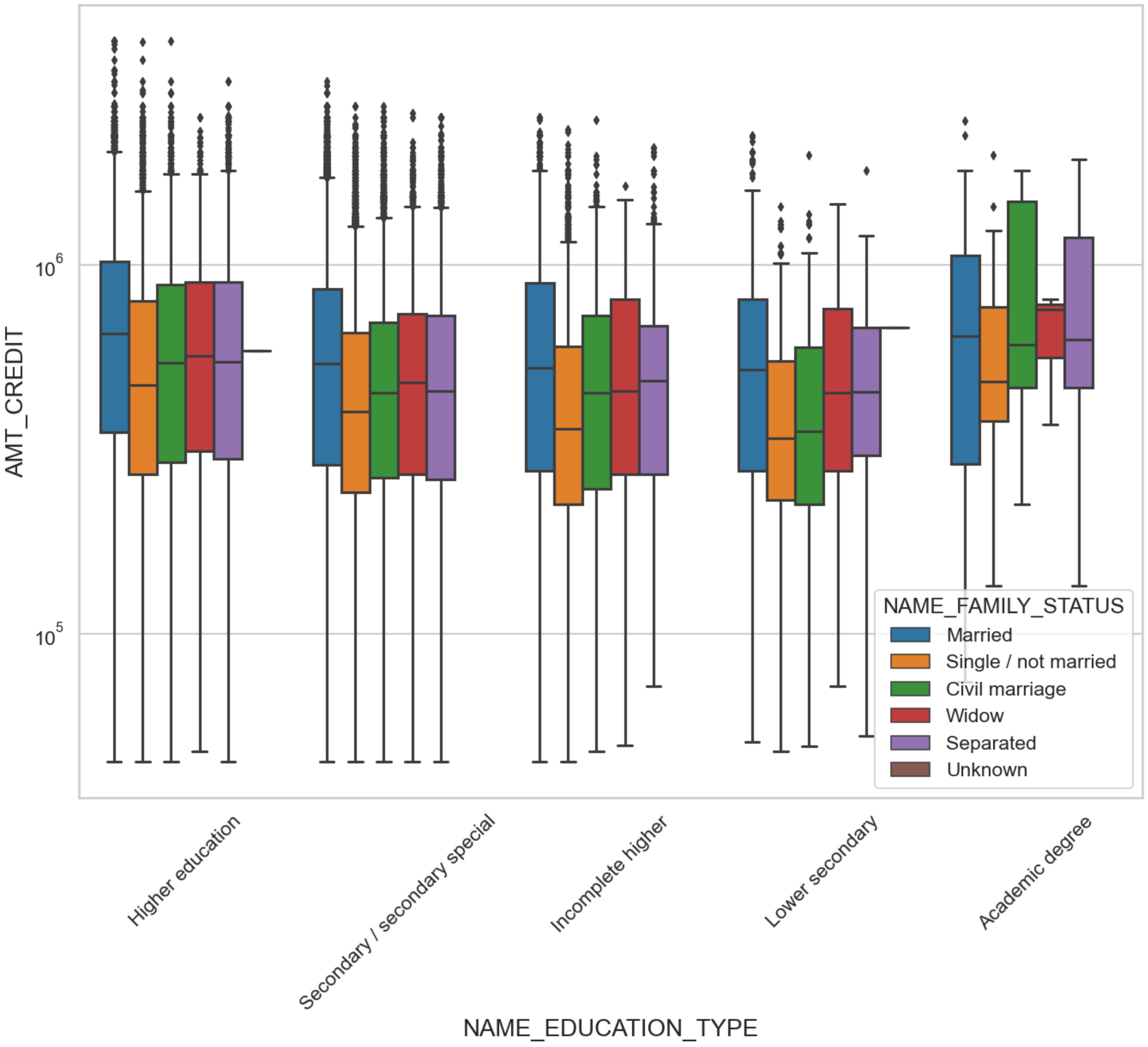
2. Similarly, clients whose permanent address does not match their work address also tend to have fewer children, and vice versa.

These differences suggest that there may be specific factors or patterns associated with Target 1 that differ from those observed for Target 0 in the dataset.

Bivariant Analysis

Target – 0 (With no Payment Difficulties)

Credit Amount Vs Education Status



Based on the information provided in the box plot analysis, we can draw the following conclusions:

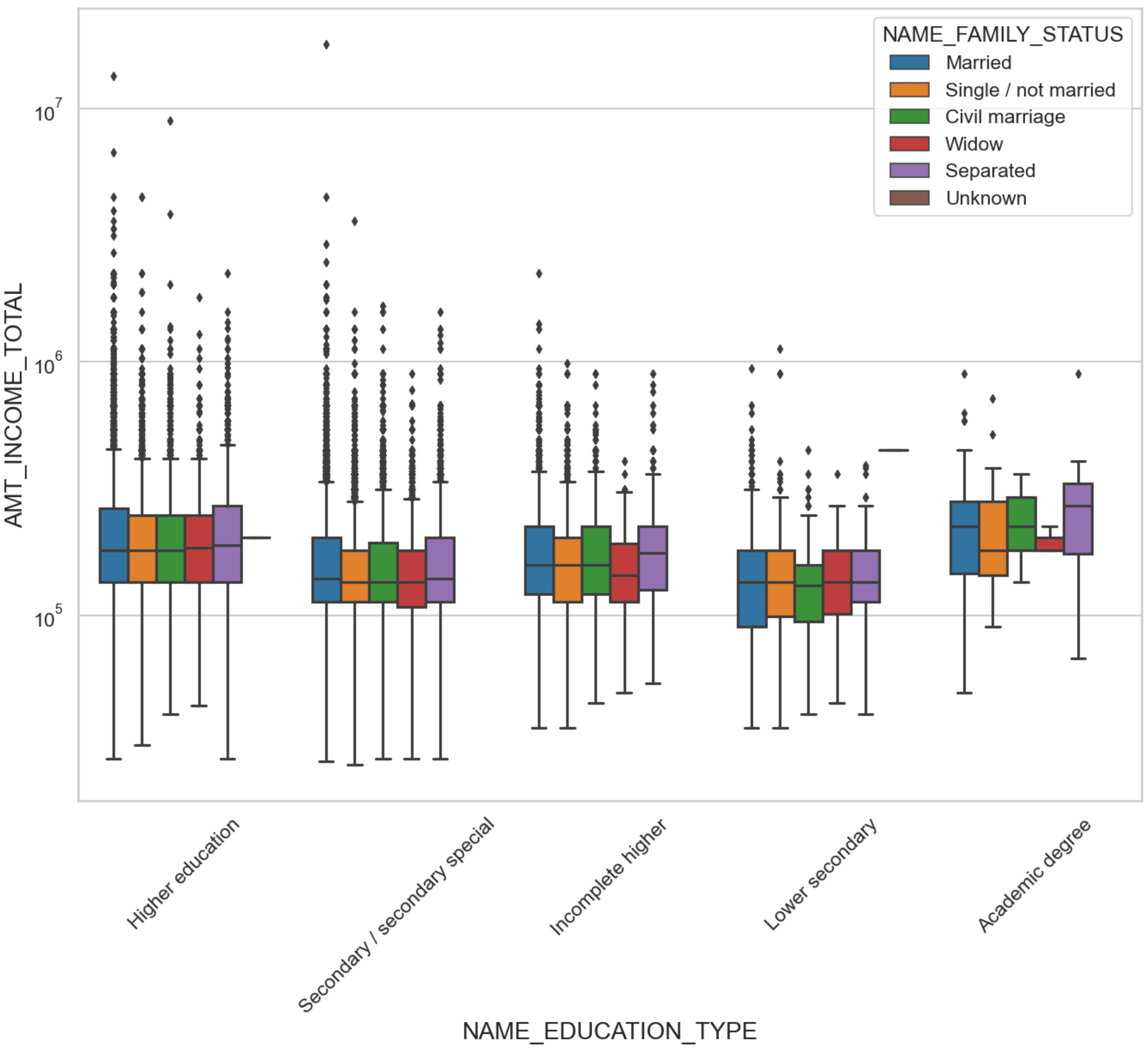
1. Family status categories such as 'civil marriage,' 'marriage,' and 'separated' tend to have a higher number of credits compared to other family status categories within the Academic degree education group.

2. Among the higher education levels, specifically for family status categories 'marriage,' 'single,' and 'civil marriage,' there are more outliers. This suggests that there may be some unique or extreme cases within these categories in terms of the number of credits.

3. In the case of 'civil marriage' within the Academic degree education group, most of the credits fall within the third quartile, indicating that this family status category has a substantial number of credits in the upper range of values.

These conclusions provide insights into the relationships between family status, education level, and the number of credits, as well as the presence of outliers and quartile distribution within the dataset.

Income Amount Vs Education Status



Based on the information provided from the boxplot analysis, we can make the following observations:

1. For the education type 'Higher education,' the income amount tends to be fairly consistent across different family status categories. This suggests that individuals with 'Higher education' generally have similar income levels, regardless of their family status.

2. The 'Higher education' group contains many outliers, indicating that there are some individuals with significantly higher or lower incomes within this category.

3. In contrast, for the 'Academic degree' education type, there are fewer outliers, and the income amount tends to be slightly higher compared to 'Higher education.' This suggests that individuals with 'Academic degree' education may have slightly higher incomes on average.

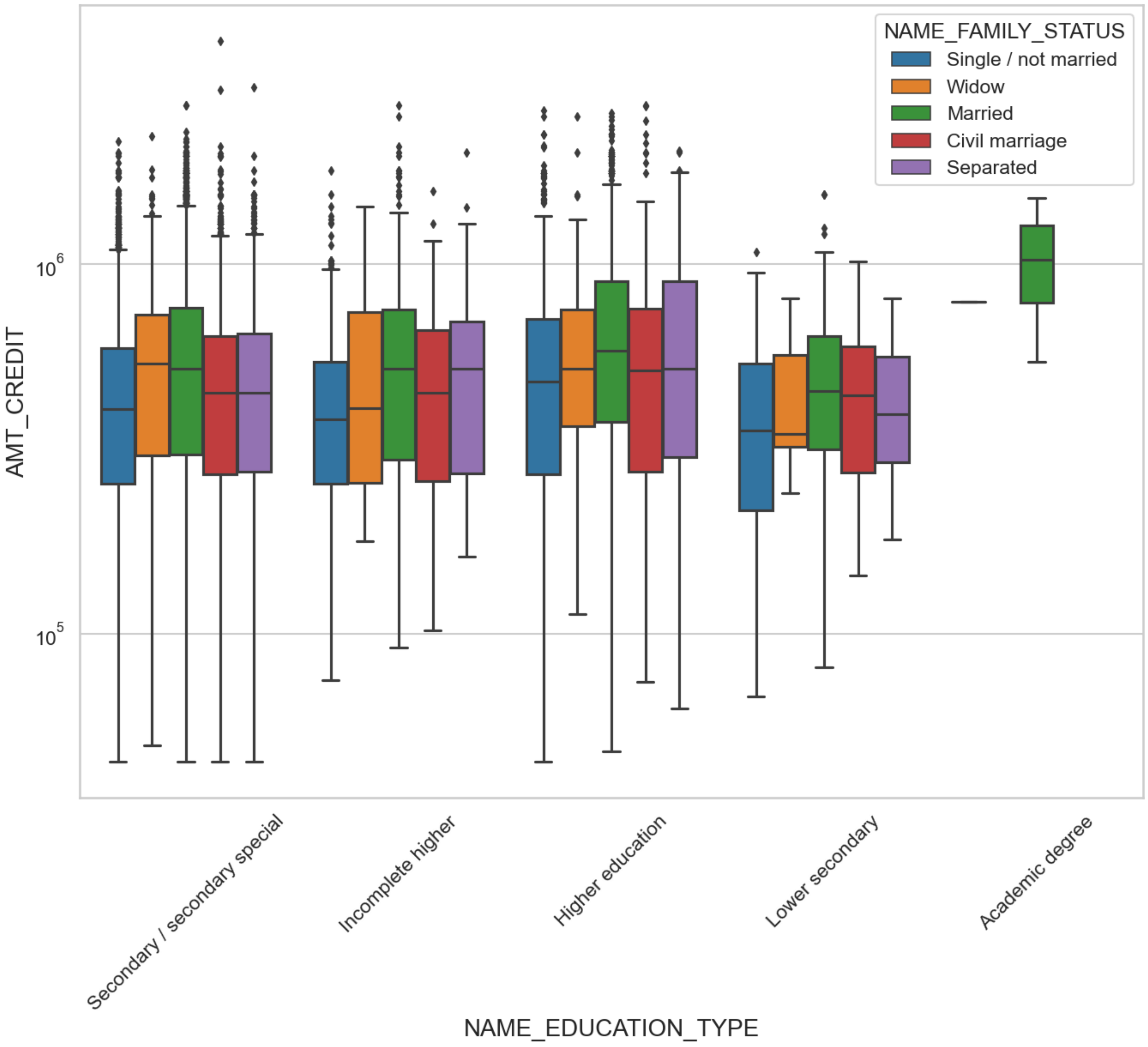
4. Within the 'Civil marriage' family status category and the 'Lower secondary' education type, there is a lower income amount compared to other combinations of family status and education types. This indicates that individuals with 'Lower secondary' education and 'Civil marriage' family status tend to have lower incomes.

These observations provide insights into the relationships between education type, family status, and income levels, as well as the presence of outliers within the dataset.

Bivariant Analysis

Target – 1 (With Payment Difficulties)

Credit Amount Vs Education Status



Based on the information provided in the box plot analysis, we can draw the following conclusions:

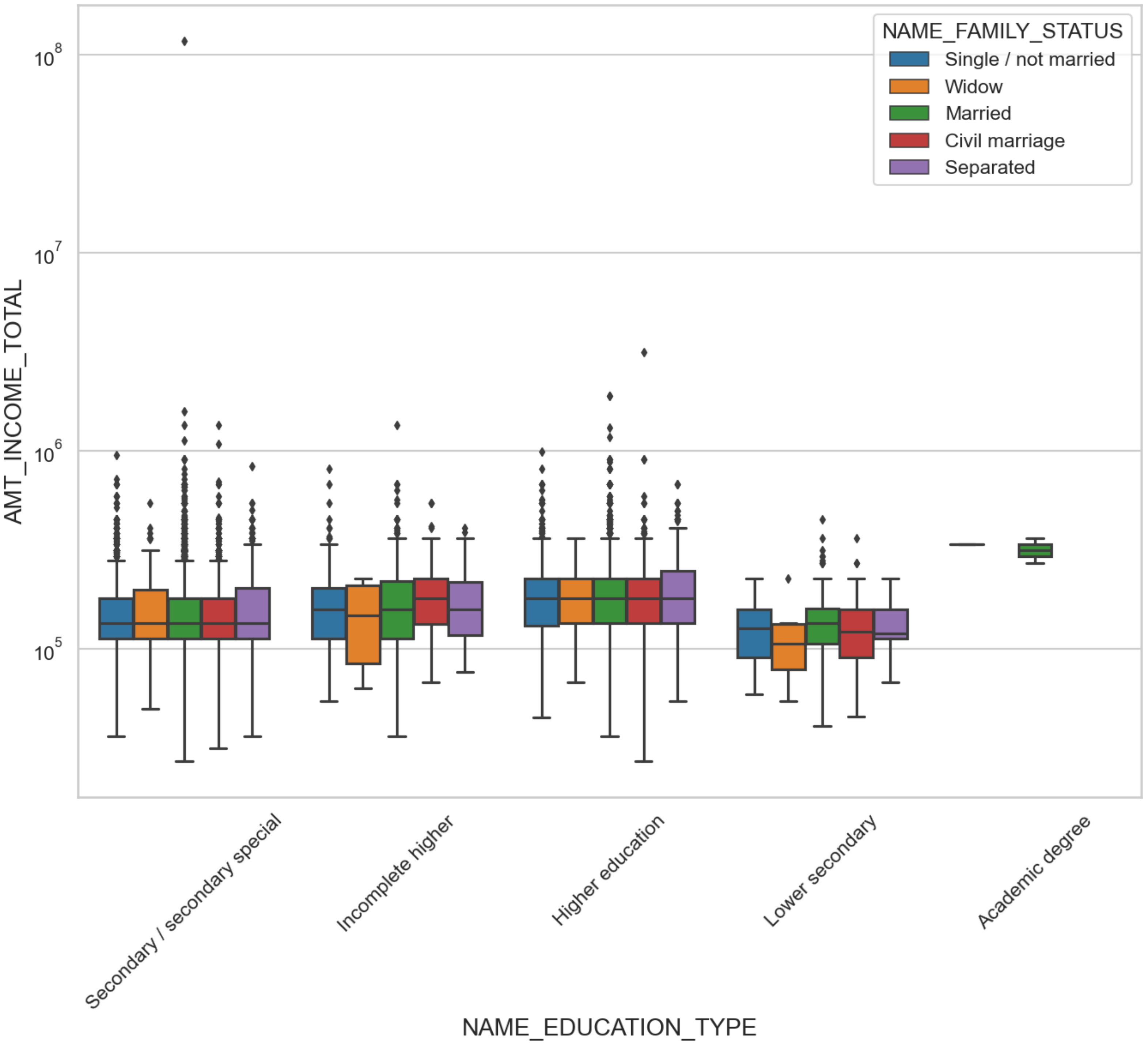
1. Family status categories such as 'civil marriage,' 'marriage,' and 'separated' tend to have a higher number of credits compared to other family status categories within the Higher education group.

2. Among the higher education levels, specifically for family status categories 'marriage,' 'single,' and 'civil marriage,' there are more outliers. This suggests that there may be some unique or extreme cases within these categories in terms of the number of credits.

3. For Academic degree Education group, Single status has very narrow Quatelies and Maried has higher number of credits compared to others.

These conclusions provide insights into the relationships between family status, education level, and the number of credits, as well as the presence of outliers and quartile distribution within the dataset.

Income Amount Vs Education Status



Based on the information provided from the boxplot analysis, we can make the following observations:

1. For the education type 'Higher education,' the income amount tends to be fairly consistent across different family status categories. This suggests that individuals with 'Higher education' generally have similar income levels, regardless of their family status.

2. The 'Higher education' group contains many outliers, indicating that there are some individuals with significantly higher or lower incomes within this category.

3. In contrast, for the 'Academic degree' education type, there are no outliers, and the income amount tends to be slightly higher compared to 'Higher education.' This suggests that individuals with 'Academic degree' education may have slightly higher incomes on average.

4. Within the 'Civil marriage' family status category and the 'Lower secondary' education type, there is a lower income amount compared to other combinations of family status and education types. This indicates that individuals with 'Lower secondary' education and 'Civil marriage' family status tend to have lower incomes.

These observations provide insights into the relationships between education type, family status, and income levels, as well as the presence of outliers within the dataset.

Univariate analysis after merging previous data

Distribution of contract status with purposes

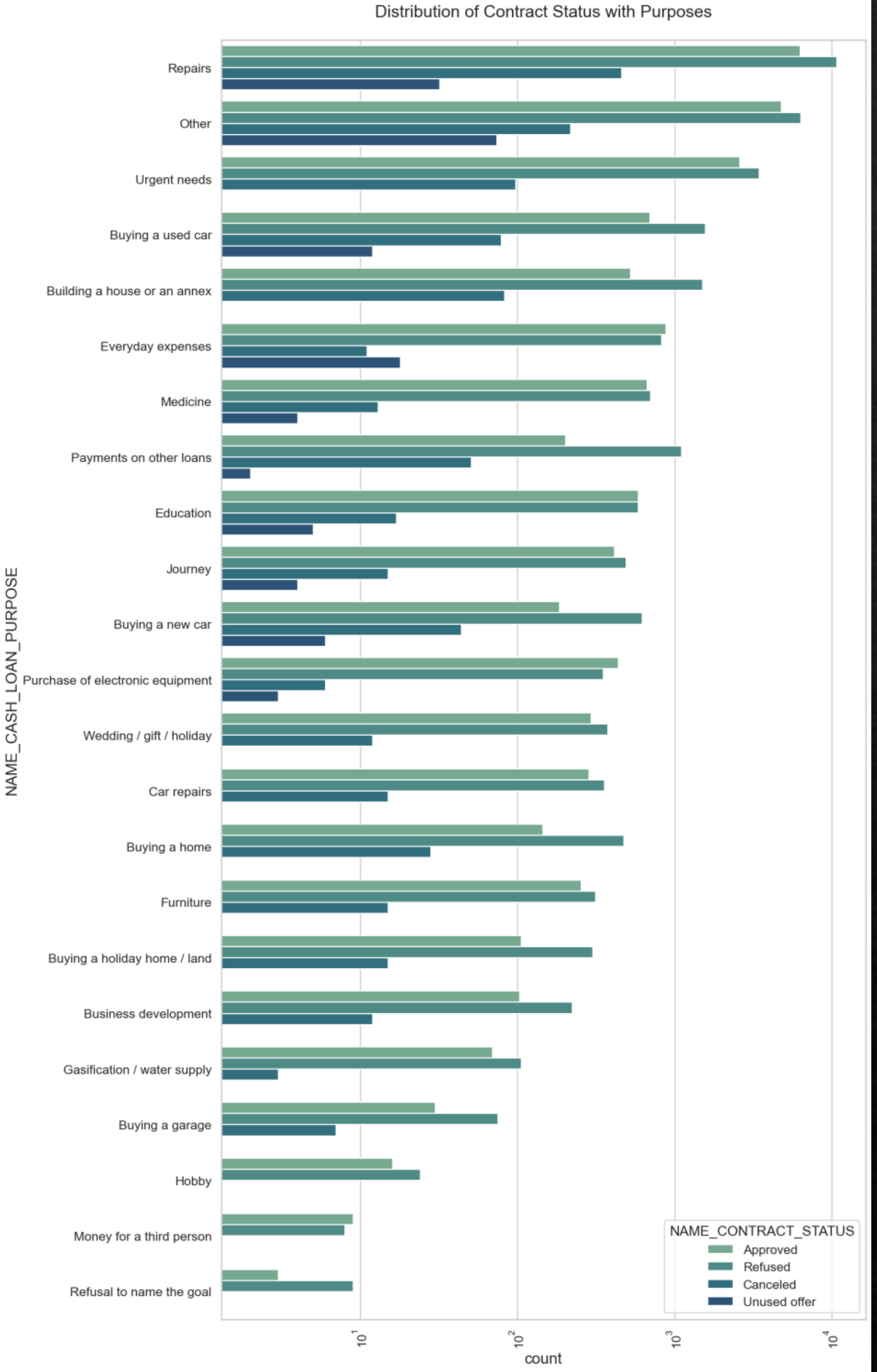
Here are the key conclusions drawn from the above plot:

1. The most rejected loan applications came from the purpose of 'repairs.' This suggests that loan applications for repair-related expenses are more likely to be rejected compared to other purposes.

2. For loan applications related to education purposes, the number of approvals and rejections is approximately equal. This indicates that loan applications for education purposes have a relatively balanced approval rate.

3. Loan applications for the purpose of 'paying other loans' and 'buying a new car' have a significantly higher rate of rejection compared to approvals. This implies that applicants seeking loans for these purposes are more likely to face rejection than approval.

4. These conclusions provide insights into the loan approval and rejection patterns based on the specified purposes, helping to identify trends and potential areas of concern.



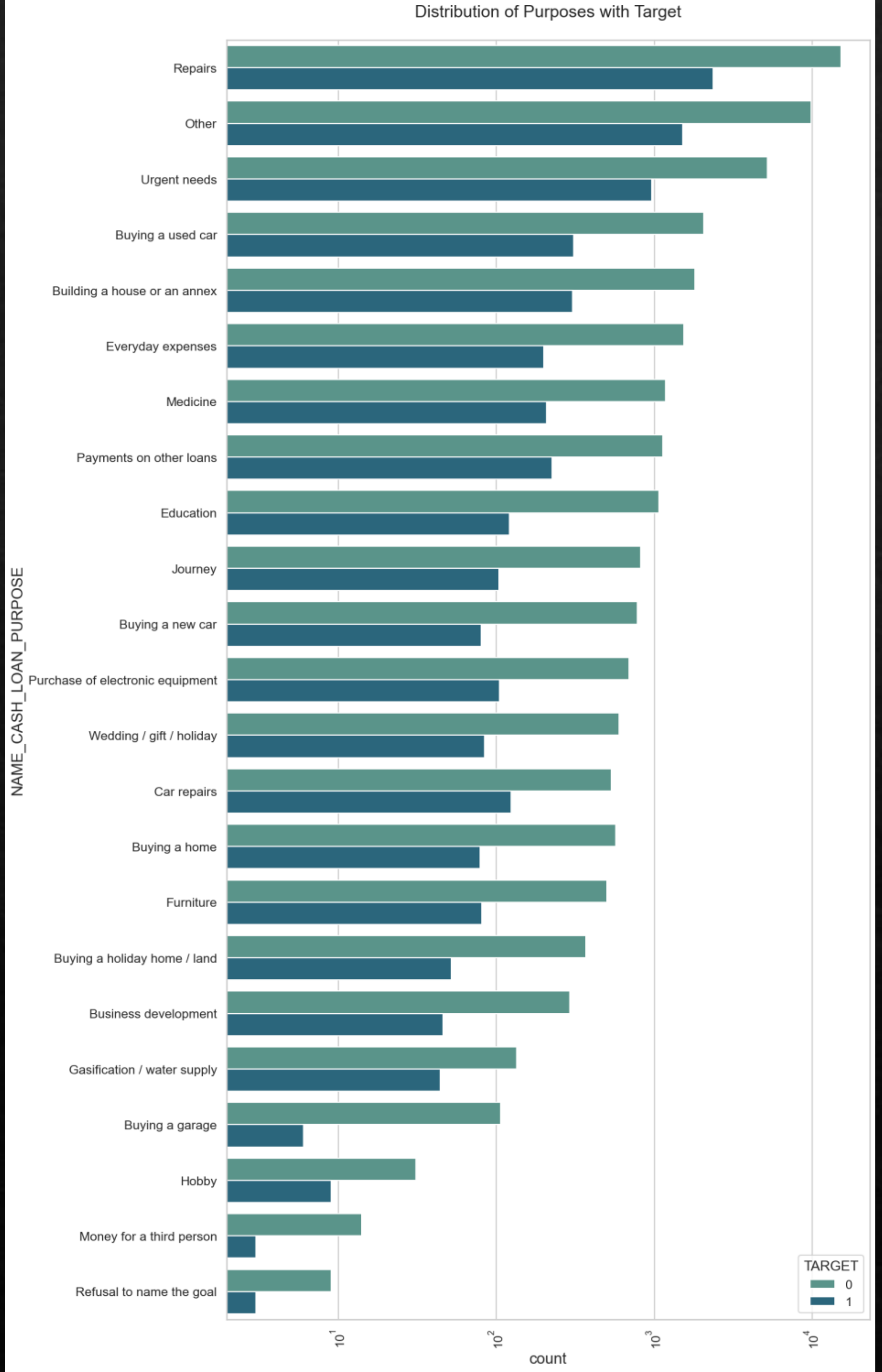
Distribution of purposes with target

Based on the information provided from the plot, we can draw the following conclusions:

1. Loan applicants with the purpose of 'Repairs' tend to face more difficulties in making payments on time. This suggests that loans taken for repair-related expenses have a higher likelihood of payment difficulties.

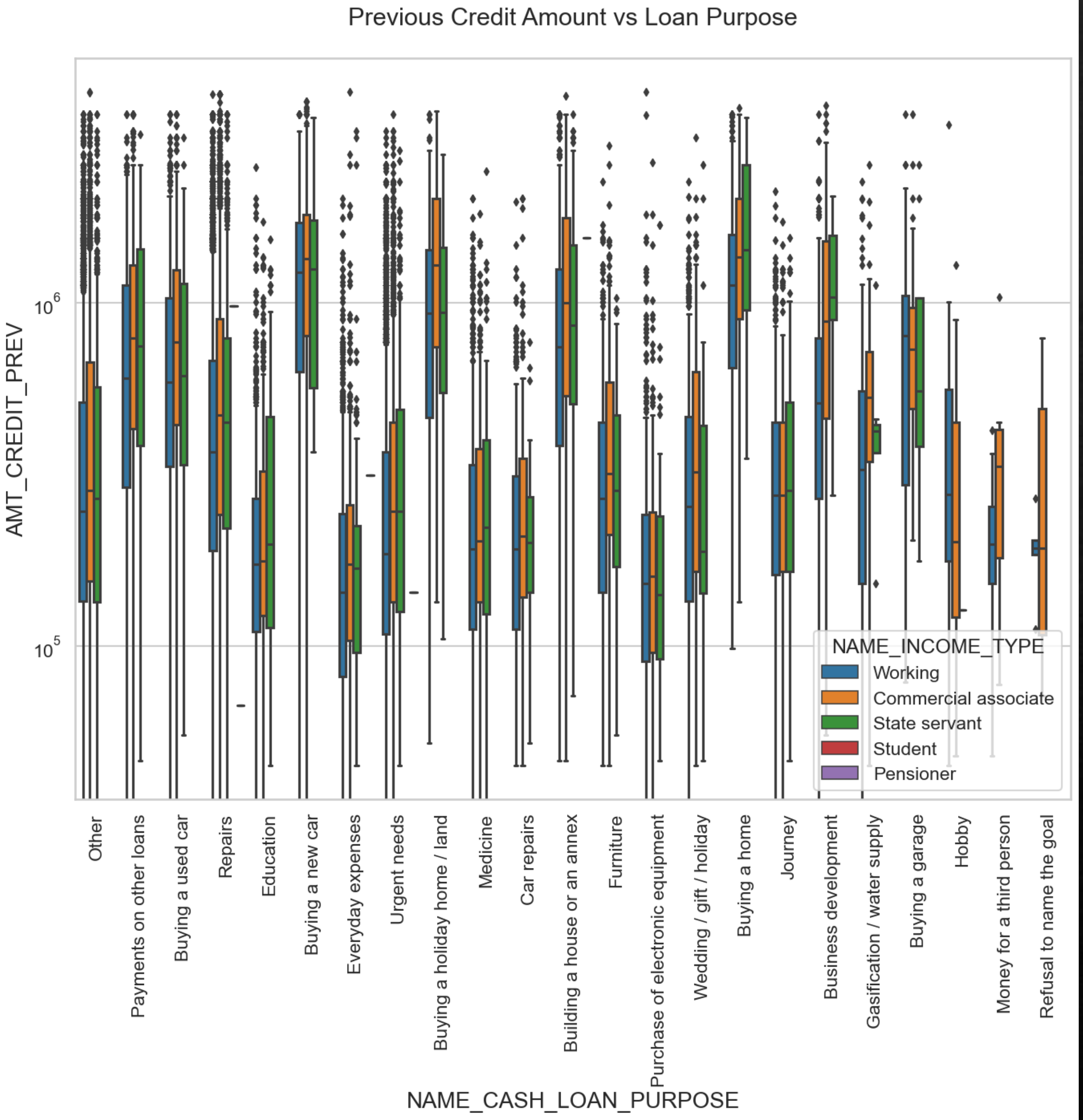
2. Conversely, there are specific loan purposes where the payment is significantly higher than facing difficulties. These purposes include 'Buying a garage,' 'Business development,' 'Buying land,' 'Buying a new car,' and 'Education.' Clients who take loans for these purposes are more likely to make payments on time and have minimal payment difficulties.

These conclusions highlight the varying payment performance associated with different loan purposes, suggesting areas of focus for better payment outcomes.



Bivariate analysis after merging previous data

Prev Credit amount vs Loan Purpose



Based on the information provided in the analysis, we can draw the following conclusions:

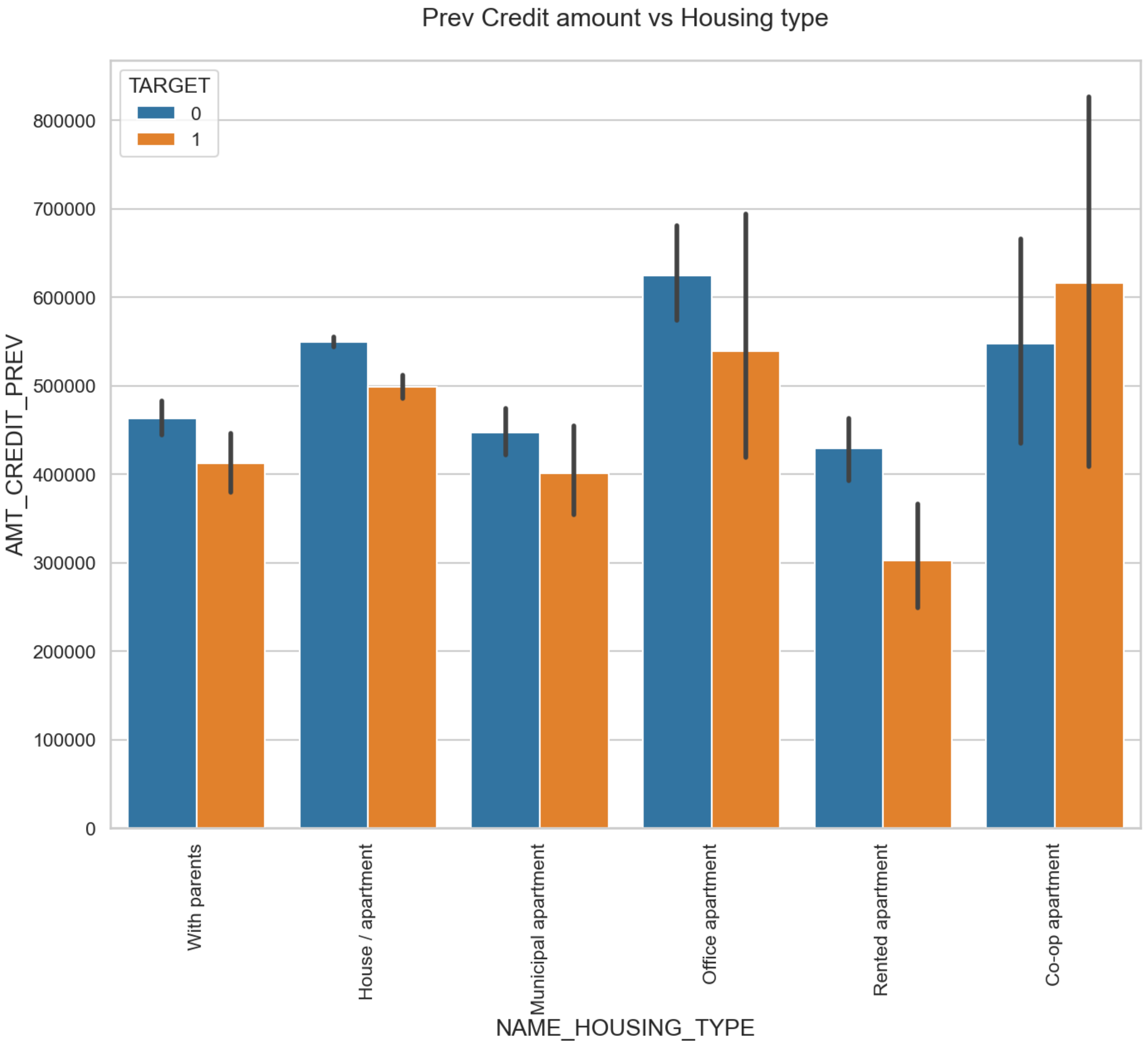
1. Loan purposes such as 'Buying a home,' 'Buying land,' 'Buying a new car,' and 'Building a house' tend to have higher credit amounts. This suggests that clients seeking loans for these purposes require larger loan amounts.

2. Income type 'State servants' have a significant number of credit applications. This indicates that individuals with the 'State servants' income type are more likely to apply for credits.

3. Loan purposes related to 'Money for a third person' or 'Hobby' have fewer credit applications. This implies that these purposes are associated with a lower number of credit applicants compared to other purposes.

These conclusions provide insights into the relationships between loan purposes, income types, and credit amounts, helping to identify patterns and trends within the dataset.

Prev Credit amount vs Housing Type



Based on the analysis provided, we can draw the following conclusions:

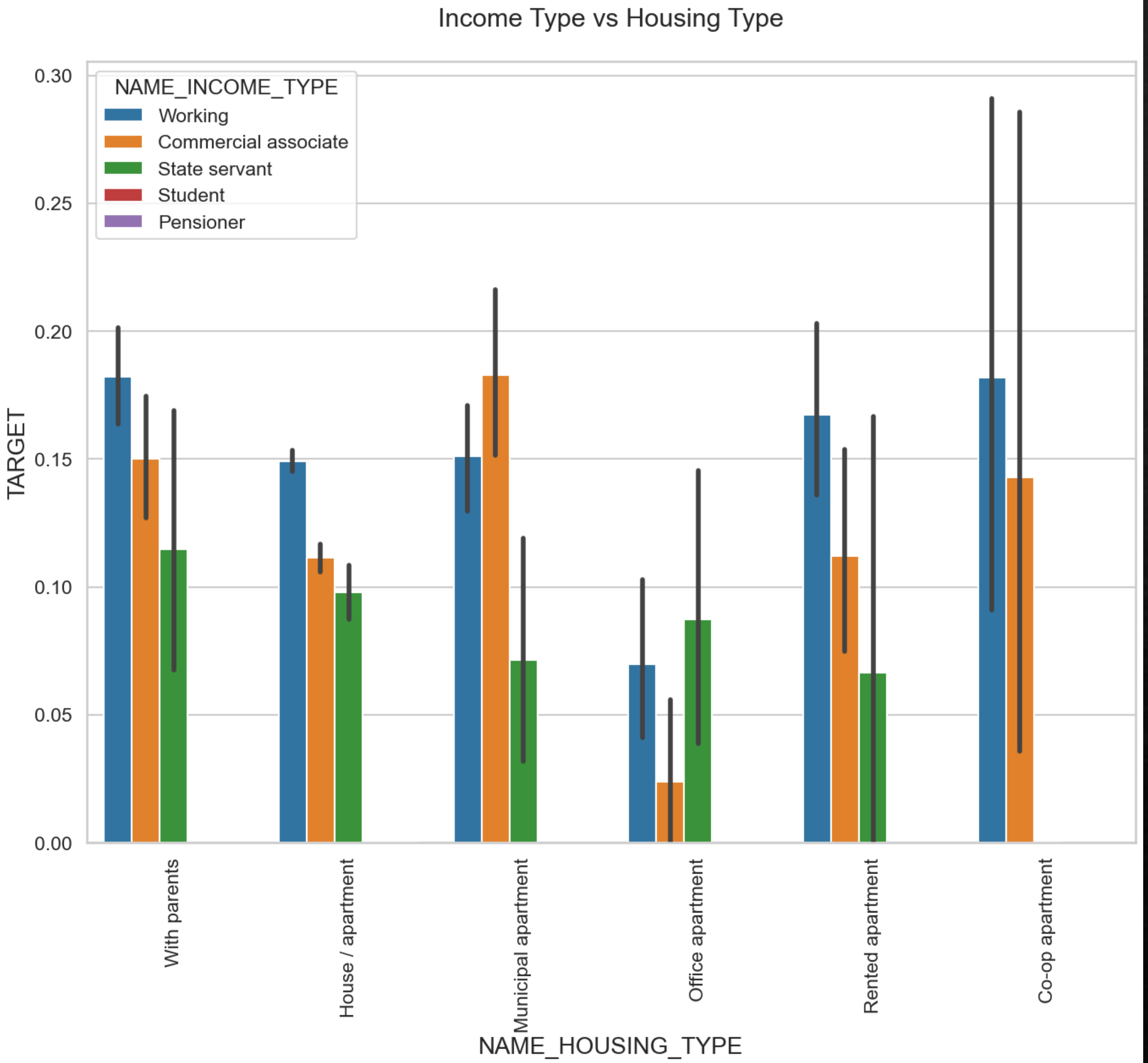
1. For the 'Housing type' variable, 'Office apartment' is associated with a higher number of credits for target 0 (successful payments), while 'Co-op apartment' is associated with a higher number of credits for target 1 (payment difficulties).

2. Based on this information, it is advisable for the bank to exercise caution when considering loan applications for the 'Co-op apartment' housing type, as it appears to have a higher likelihood of payment difficulties.

3. Conversely, the bank can focus more on housing types such as 'With parents,' 'House/apartment,' or 'Municipal apartment,' as these housing types are associated with successful payments (target 0).

These conclusions provide guidance to the bank in making informed decisions about loan applications based on the applicant's housing type and the likelihood of successful payments.

Income Type Vs Housing Type WRT Target



CONCLUSION

Based on the analysis, the following conclusions can be drawn:

1. Banks should be cautious with contract types 'Student,' 'Pensioner,' and 'Businessman' who have housing types as Office Apartments and co-apartment for loan approvals, as these combinations are associated with more unsuccessful payments.

2. Banks should Prioritize with clients with income type 'Working,' as they appear to have the highest number of successful payments. As they have a regular income.

3. Loan applications with the purpose of 'Repair' have a higher number of unsuccessful payments on time. Banks should assess such applications carefully and consider additional risk factors.

4. Banks may benefit from targeting clients with housing types such as 'With parents,' as this category has the lowest number of unsuccessful payments. Attracting more clients from this housing type may lead to a higher rate of successful payments.

These conclusions provide insights into factors that can influence the success of loan payments and help banks make informed decisions regarding loan approvals and risk assessment.