CREDIT EDA CASE STUDY

BY MOHAMMED OWAIS KHAN

PROBLEM STATEMENT

Understanding driver variables behind loan default and finding out which variables are strong indicators of a future default.

This is achieved by using EDA (exploratory data analysis) on the bank data, wherein we analyze the pattern present in data and help bank in mitigating two types of risk associated with loan approvals:-

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. He/she is likely to default, then approving the loan may lead to a financial loss for the company.

STEPS

1. Data understanding and sourcing

2.Check for data quality issues and binning

3.Check for data imbalance and univariate, segmented univariate & bivariate analysis, correlation

4.Merging of application data with previous application data

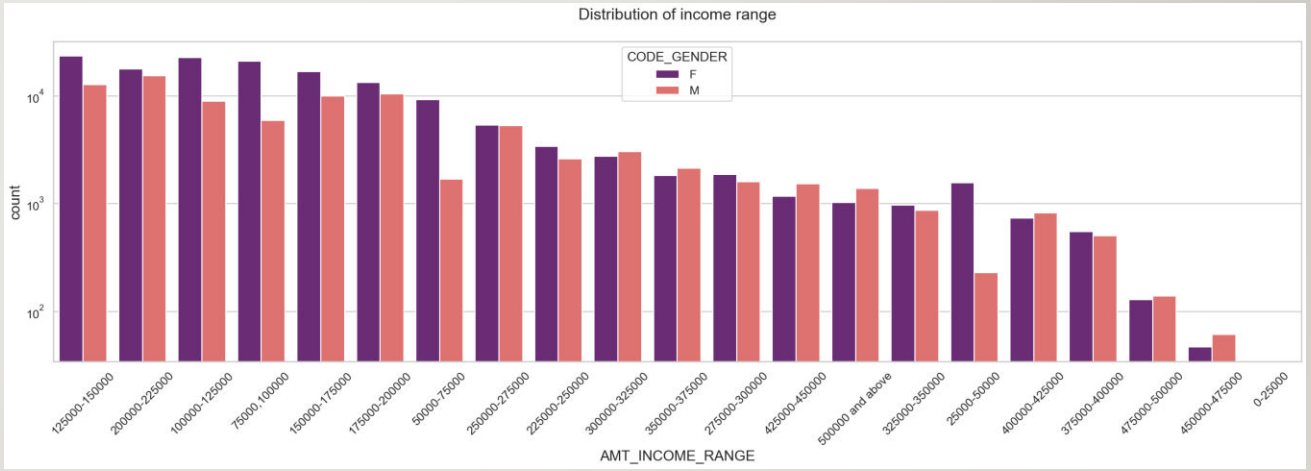
5.Data analysis by univariate, segmented univariate , bivariate analysis and correlation

6.Recommendations and risks

CATEGORICAL UNIVARIATE ANALYSIS FOR TARGET 0

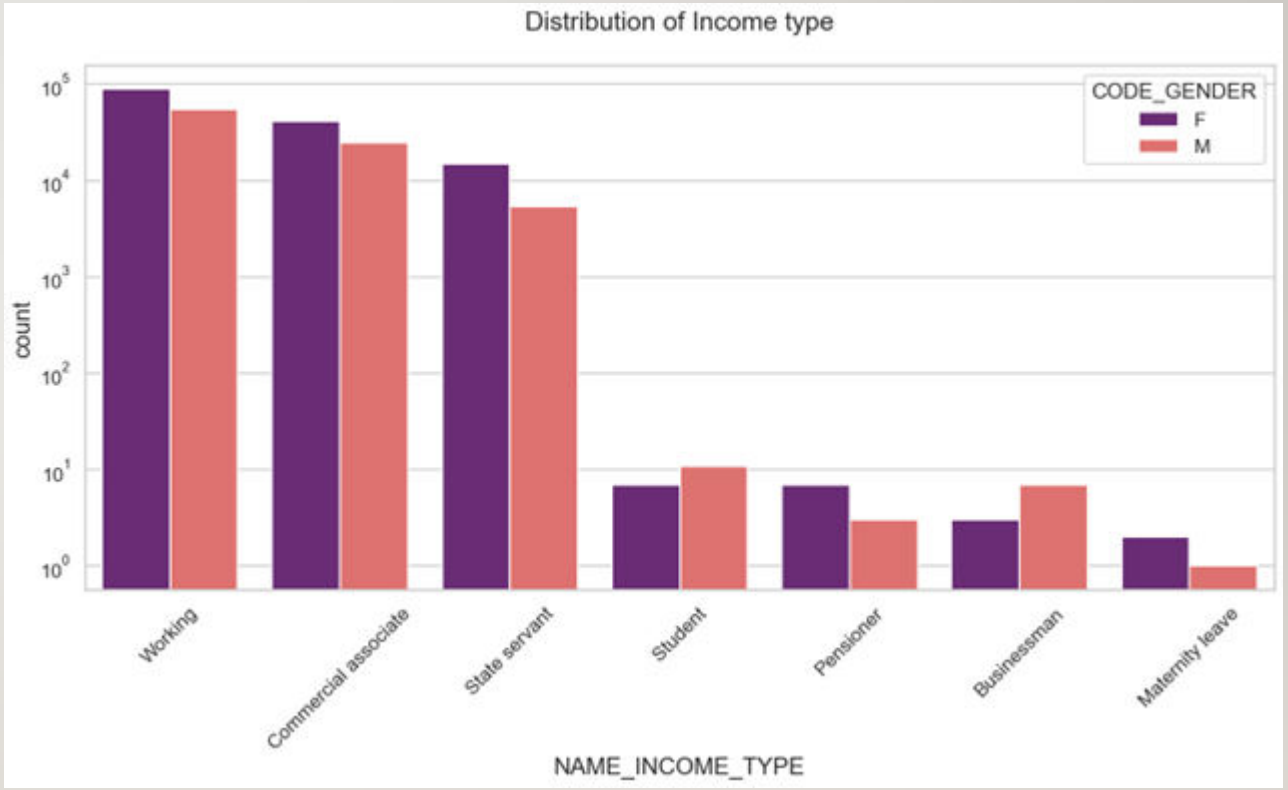
DISTRIBUTION OF INCOME RANGE

- Male numbers are lower than female counts.
- The income level between 100,000 and 200,000 has the most credits.
- This graph demonstrates that more women than men have credits in that range.
- For incomes of 400,000 and above, very few people are considered.



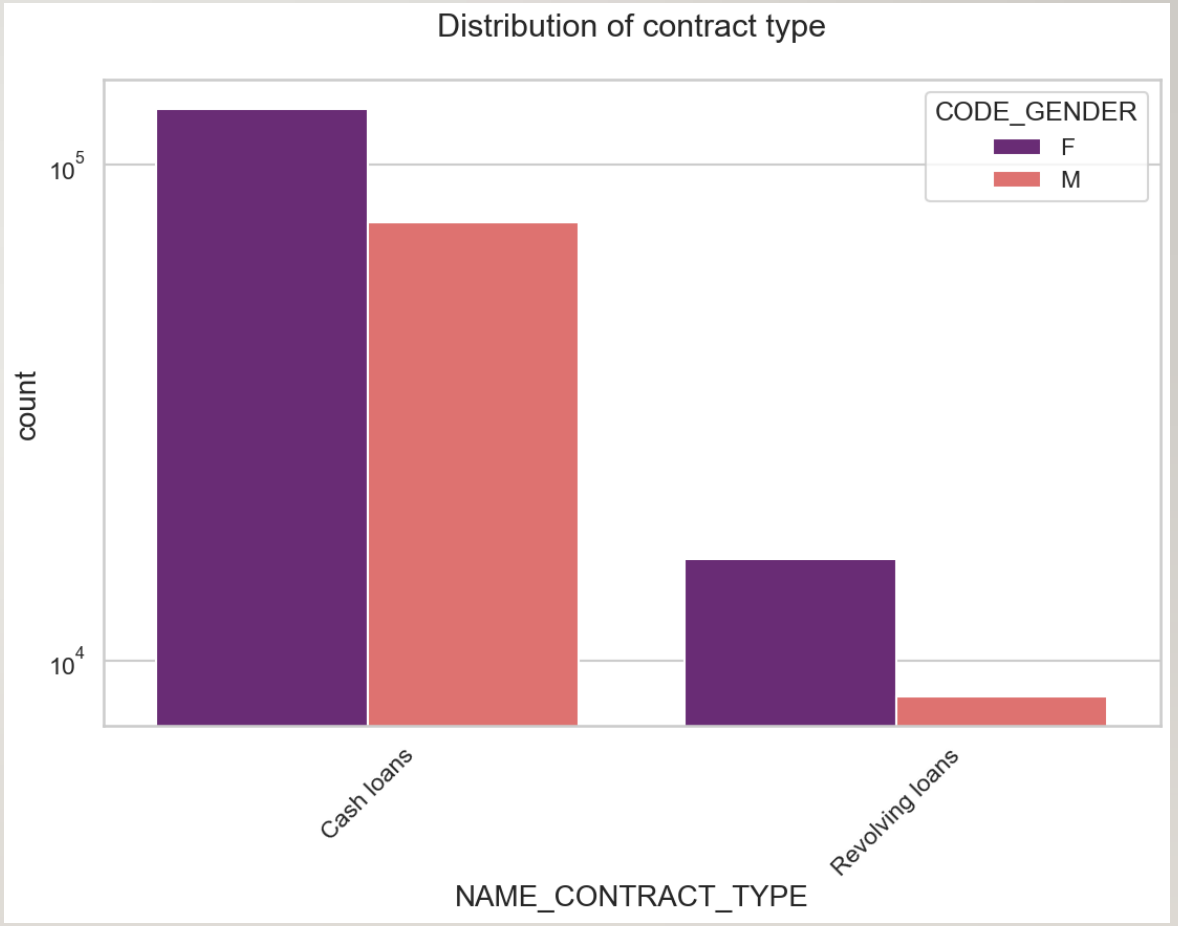
DISTRIBUTION OF INCOME TYPE

- There are more credits available for income types "working," "commercial associate," and "state servant" than for the others.
- For this reason, women have more credits than men do.
- Less credits are available for income categories such "student," "pensioner," "businessman," and "maternity leave."



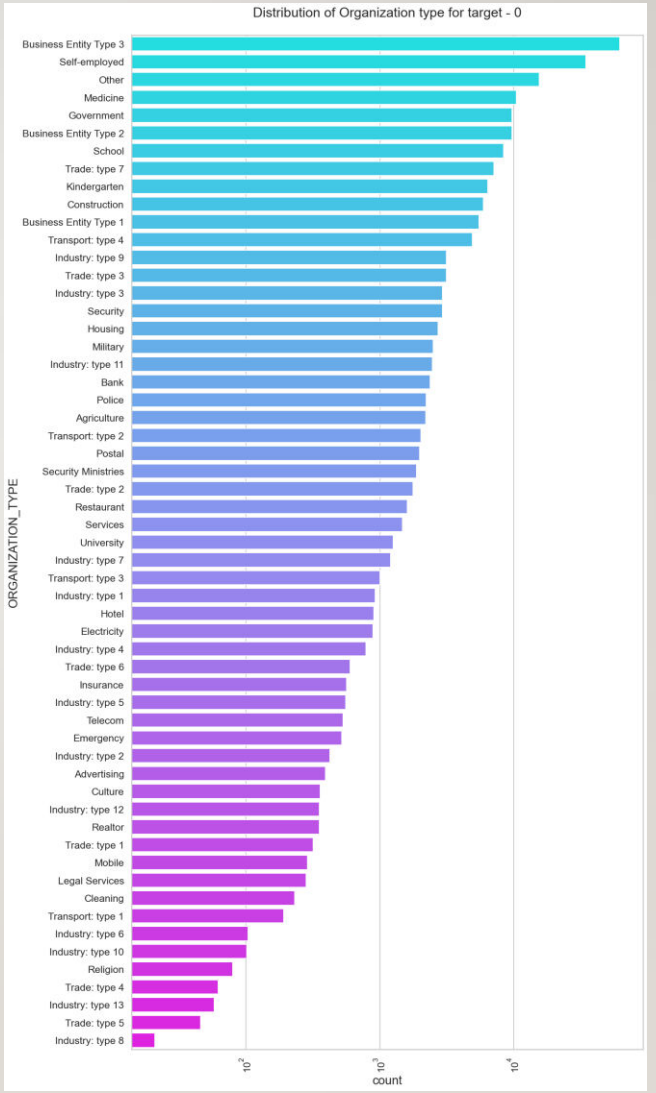
DISTRIBUTION FOR CONTRACT TYPE

- Cash loans, as a contract type, have a greater number of credits than revolving loans.
- Additionally, women are more likely than men to apply for credit.



DISTRIBUTION OF ORGANIZATION TYPE

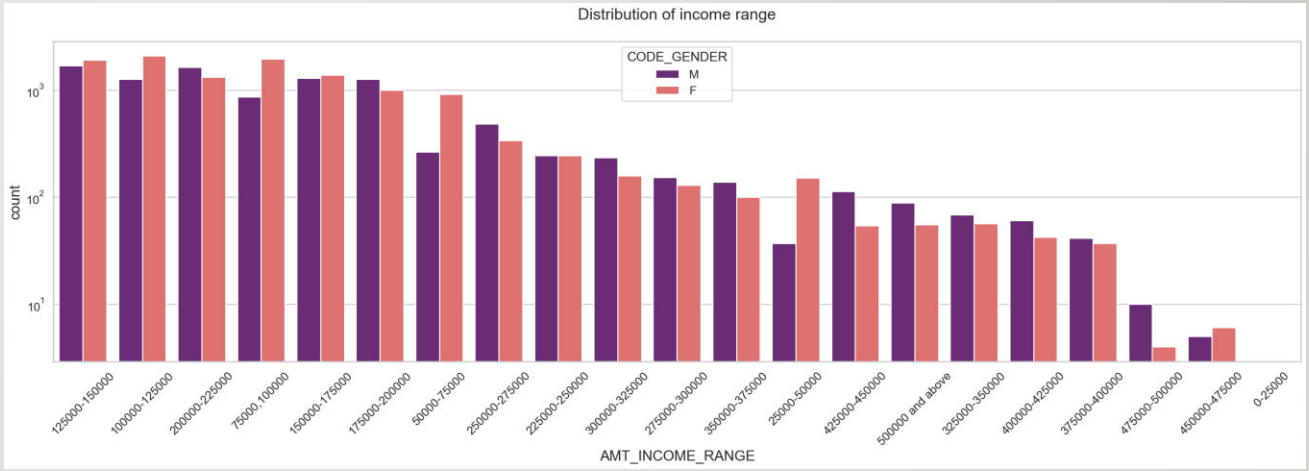
- The majority of the clients who have requested credit are from the organization types "Business entity Type 3," "Self employed," "Other," "Medicine," and "Government."
- Less customers come from industry types 8, 6, 10, religion, and type 5 and type 4 trade.



CATEGORICAL UNIVARIATE ANALYSIS FOR TARGET 1

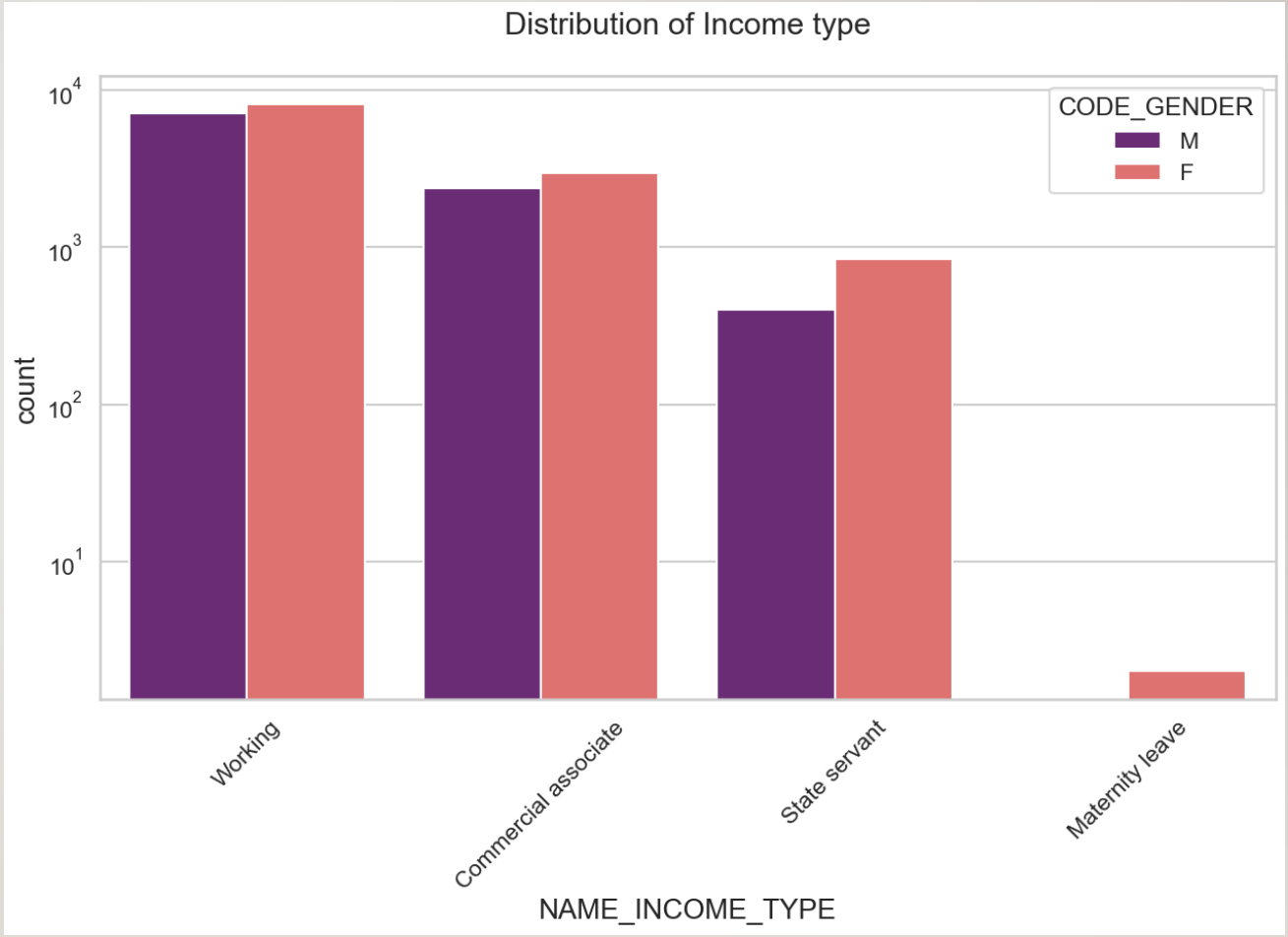
DISTRIBUTION OF INCOME RANGE

- Male counts are higher than female.
- The income level between 100,000 and 200,000 has the most credits.
- According to this graph, men have more credits in that range than women do.
- For incomes of 400,000 and above, very few people are considered



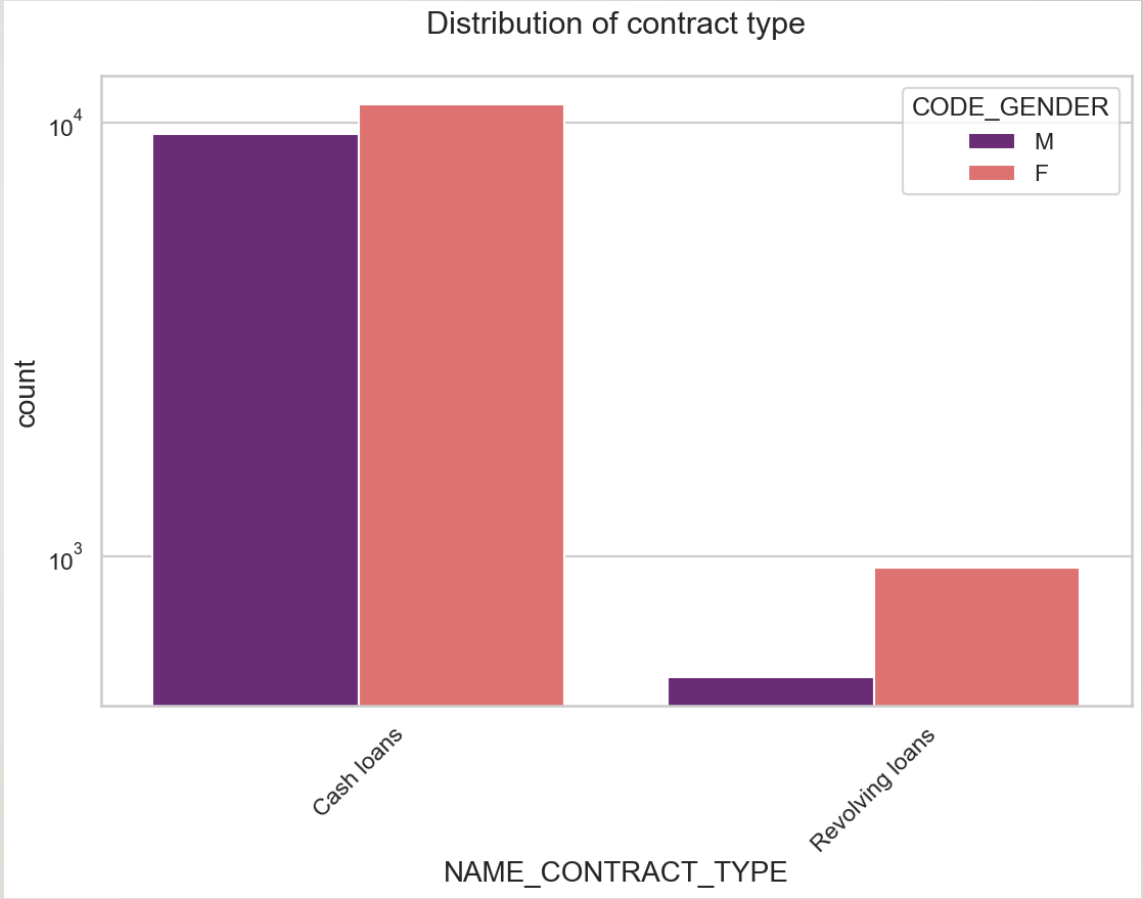
DISTRIBUTION OF INCOME TYPE

- The number of credits is higher for income types such as "working," "commercial associate," and "State Servant" than for other types, such as "Maternity leave."
- For this reason, women have more credits than men do.
- 'Maternity leave' income category receives fewer credits.
- For type 1, "student," "pensioner," and "businessman" have no income types, which suggests they don't practice late payments.



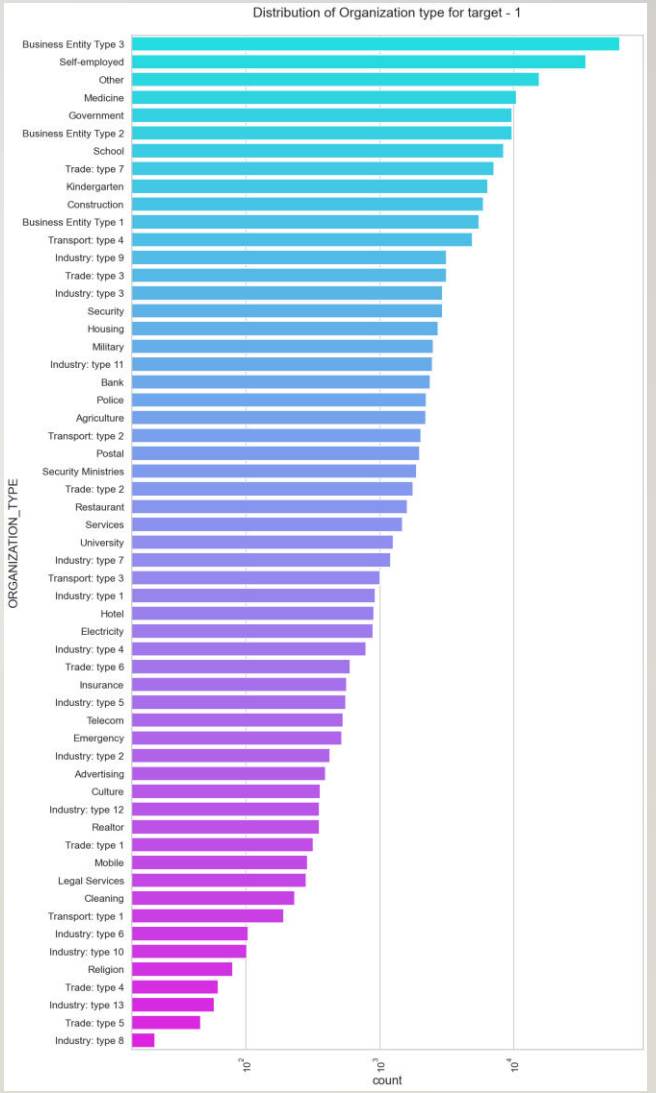
DISTRIBUTION FOR CONTRACT TYPE

- Cash loans, as a contract type, have a greater number of credits than revolving loans.
- Additionally, women are more likely than men to apply for credit.
- There are only female revolving loans for type 1.



DISTRIBUTION OF ORGANIZATION TYPE

- The majority of the clients who have requested credit are from the organization types "Business entity Type 3," "Self employed," "Other," "Medicine," and "Government."
- Less customers come from industry types 8, 6, 10, religion, and type 5 and type 4 trade.
- In the distribution of organization type, the same as type 0.



CORRELATION OF TARGET 0

CORRELATION OF TARGET 0

- Since credit amount and date of birth are inversely related, which means credit amount is higher for low age and vice-versa.
- Credit amount is inversely proportional to the number of children client have, which 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 has, therefore a client with fewer children will have more income than a client with more children.
- In a heavily populated location, clients have fewer children.
- Credit is given in greater amounts to heavily inhabited areas.
- In areas with high population density, income is also higher.

Correlation for target 0
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
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

CORRELATION OF TARGET I

CORRELATION OF TARGET 1

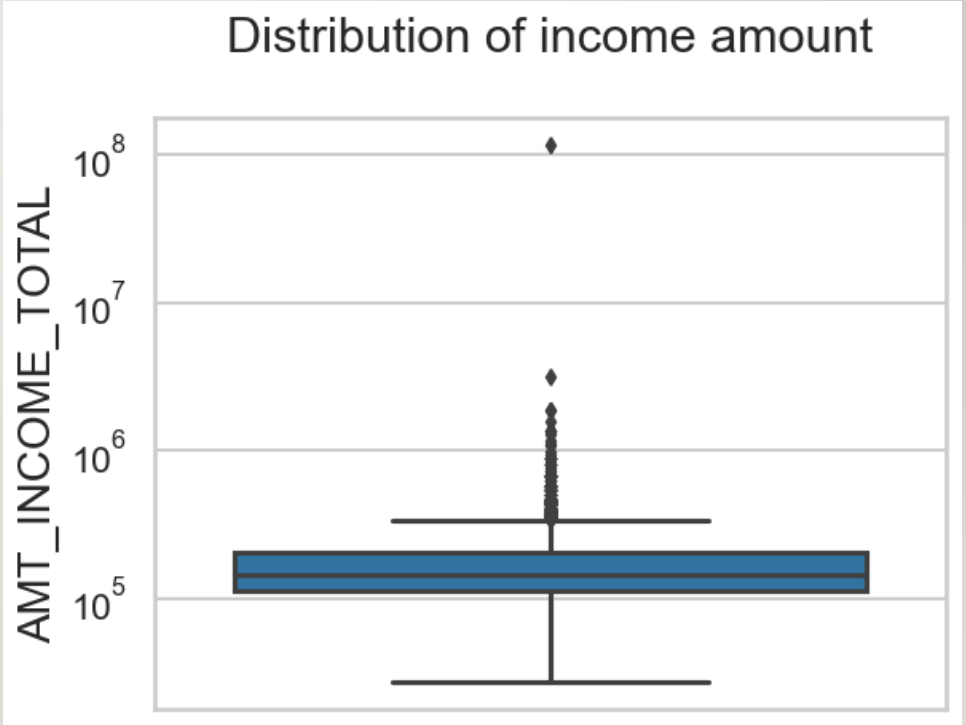
- There are fewer children being born at the client's permanent address than at their contact address, and vice versa.
- Less children are being born to clients whose permanent addresses differ from their work addresses, and vice versa

Correlation for target 1
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
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

CATEGORICAL UNIVARIATE ANALYSIS FOR VARIABLES TARGET 0

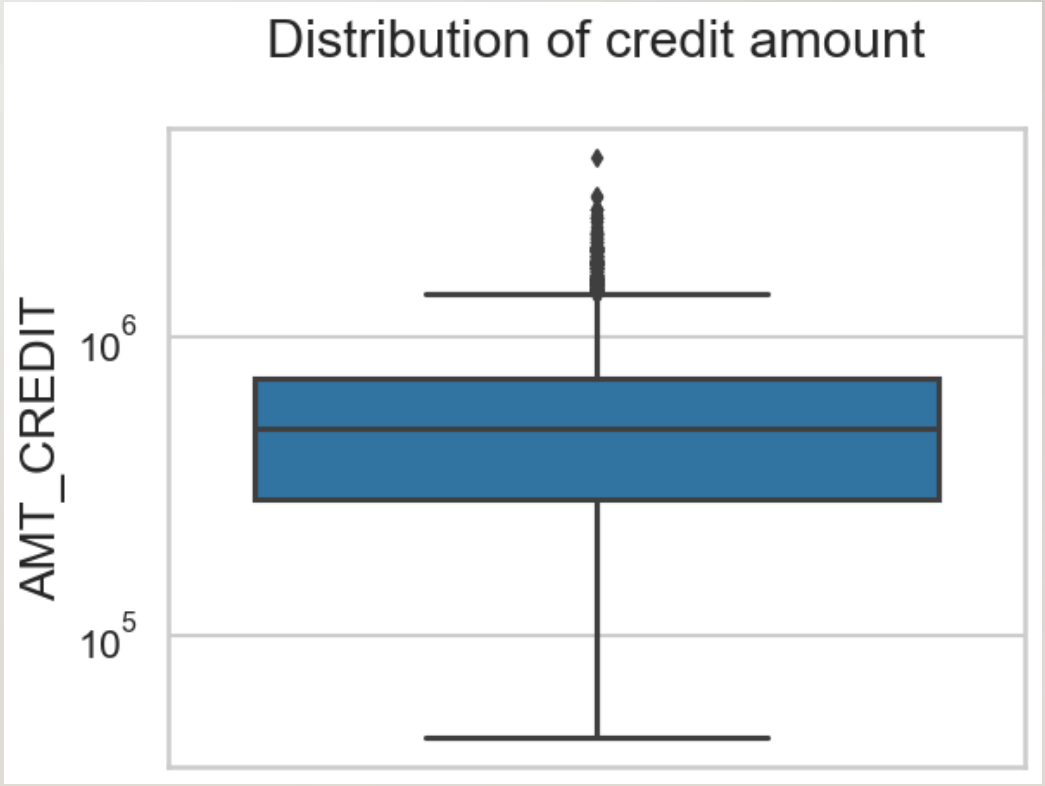
BOXPLOT FOR INCOME AMOUNT

- Some outliers are noticed in income amount.
- The third quartiles is very slim for income amount.



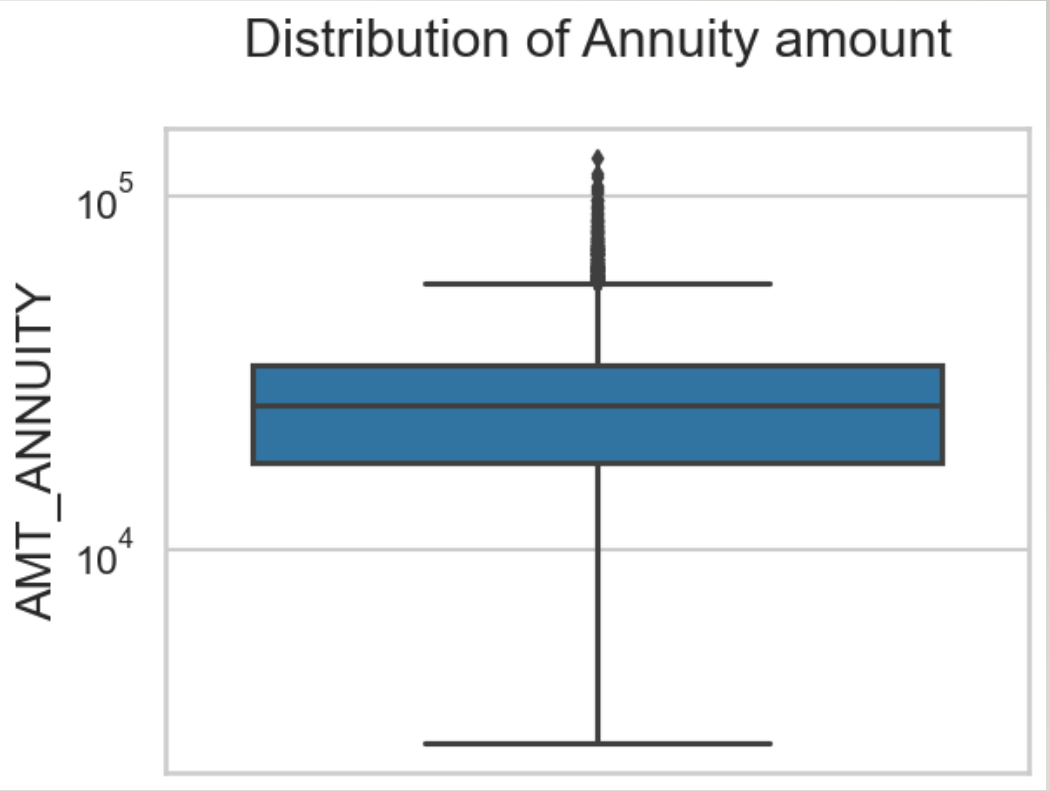
BOXPLOT FOR CREDIT AMOUNT

- Some outliers are noticed in credit amount.
- 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 FOR ANNUITY AMOUNT

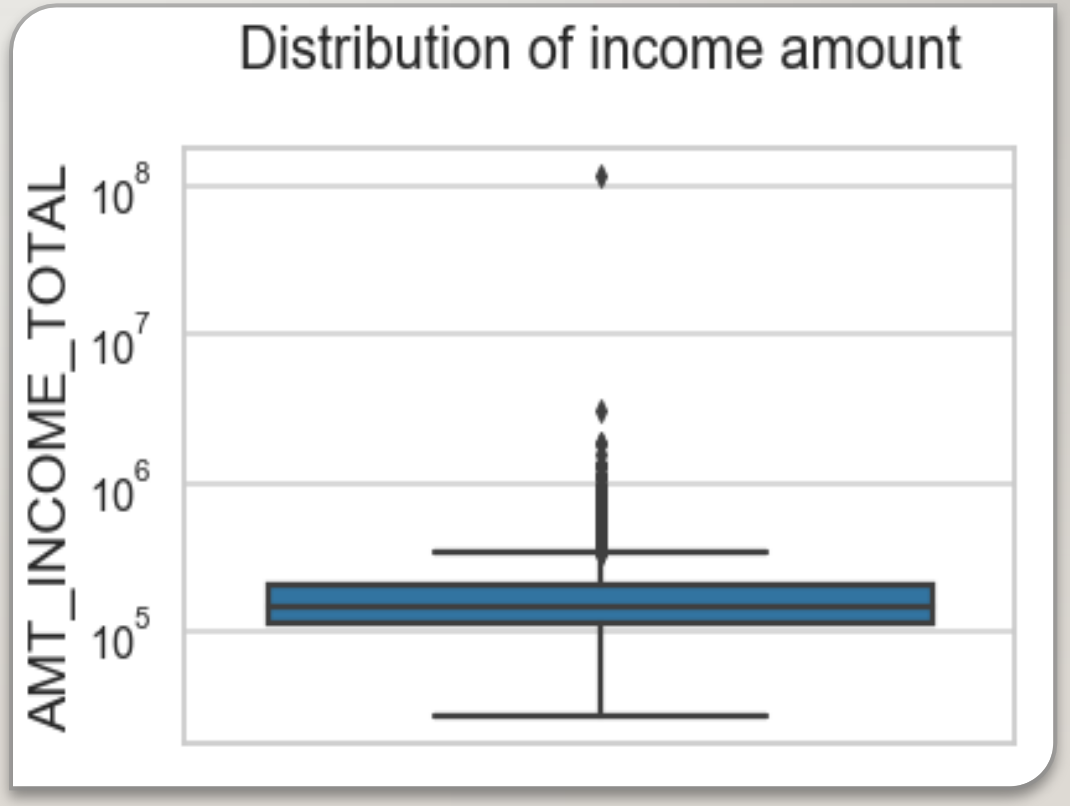
- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.



CATEGORICAL UNIVARIATE ANALYSIS FOR VARIABLES TARGET 1

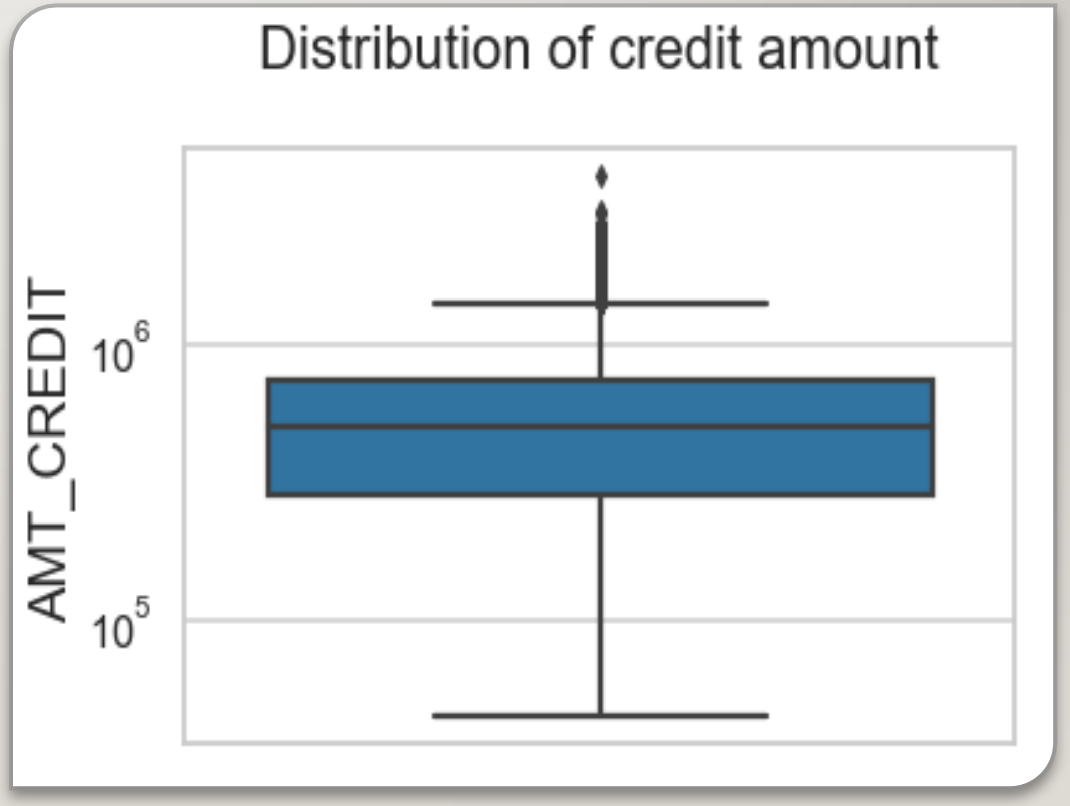
BOXPLOT FOR INCOME AMOUNT

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



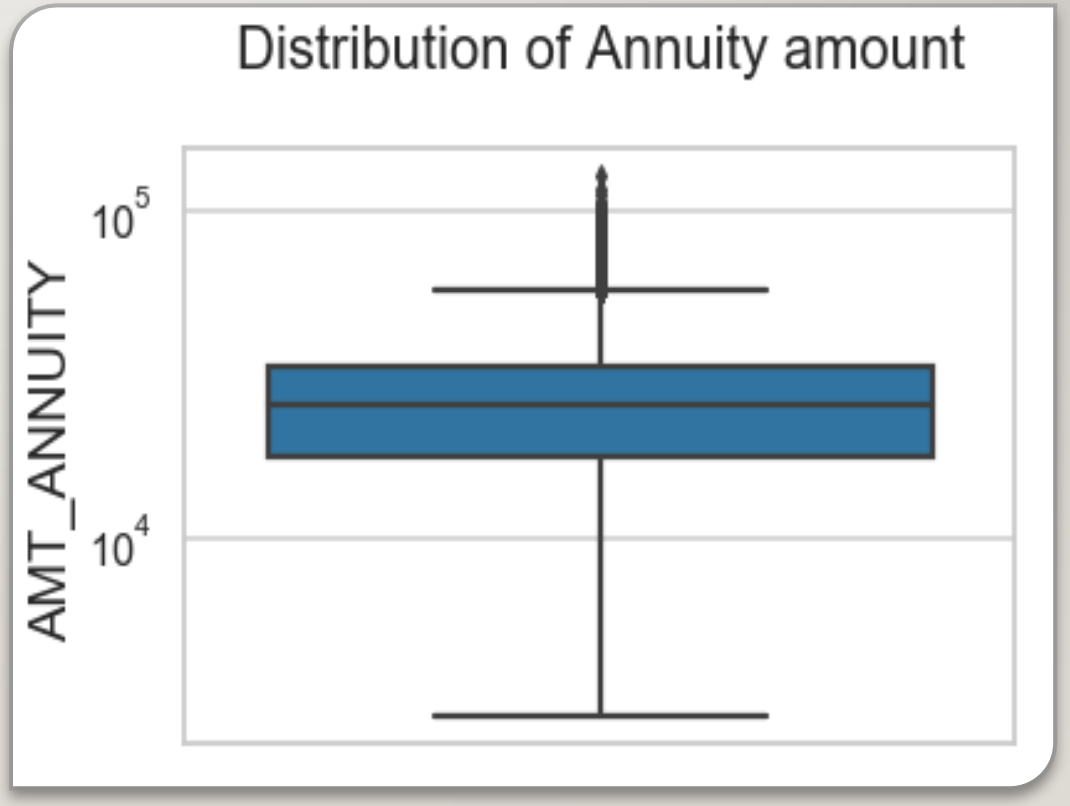
BOXPLOT FOR CREDIT AMOUNT

- Some outliers are noticed in credit amount.
- 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 FOR ANNUITY AMOUNT

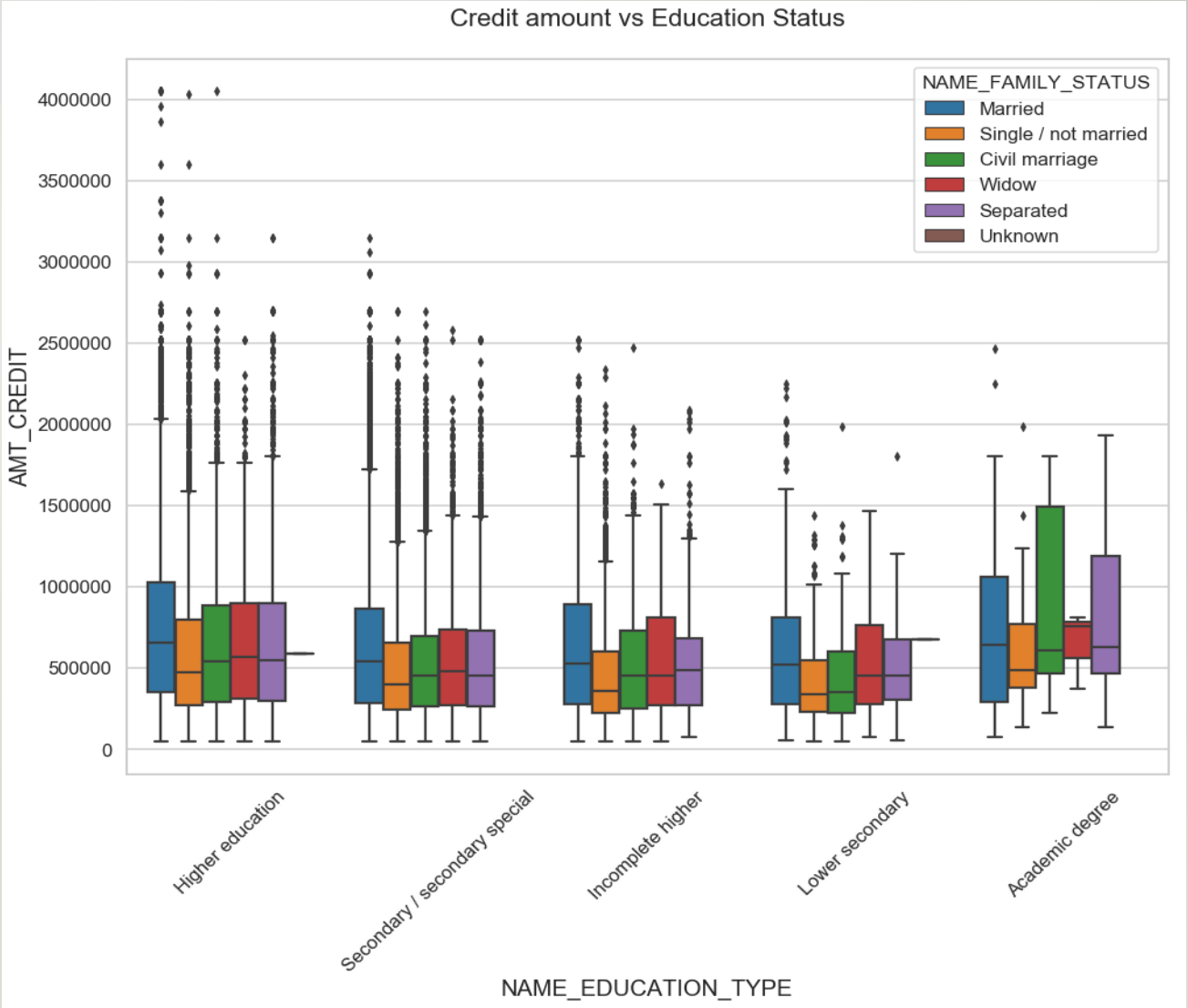
- Some outliers are noticed in annuity amount.
- The first quartile is bigger than third quartile for annuity amount which means most of the annuity clients are from first quartile.



BIVARIATE ANALYSIS FOR TYPE 0

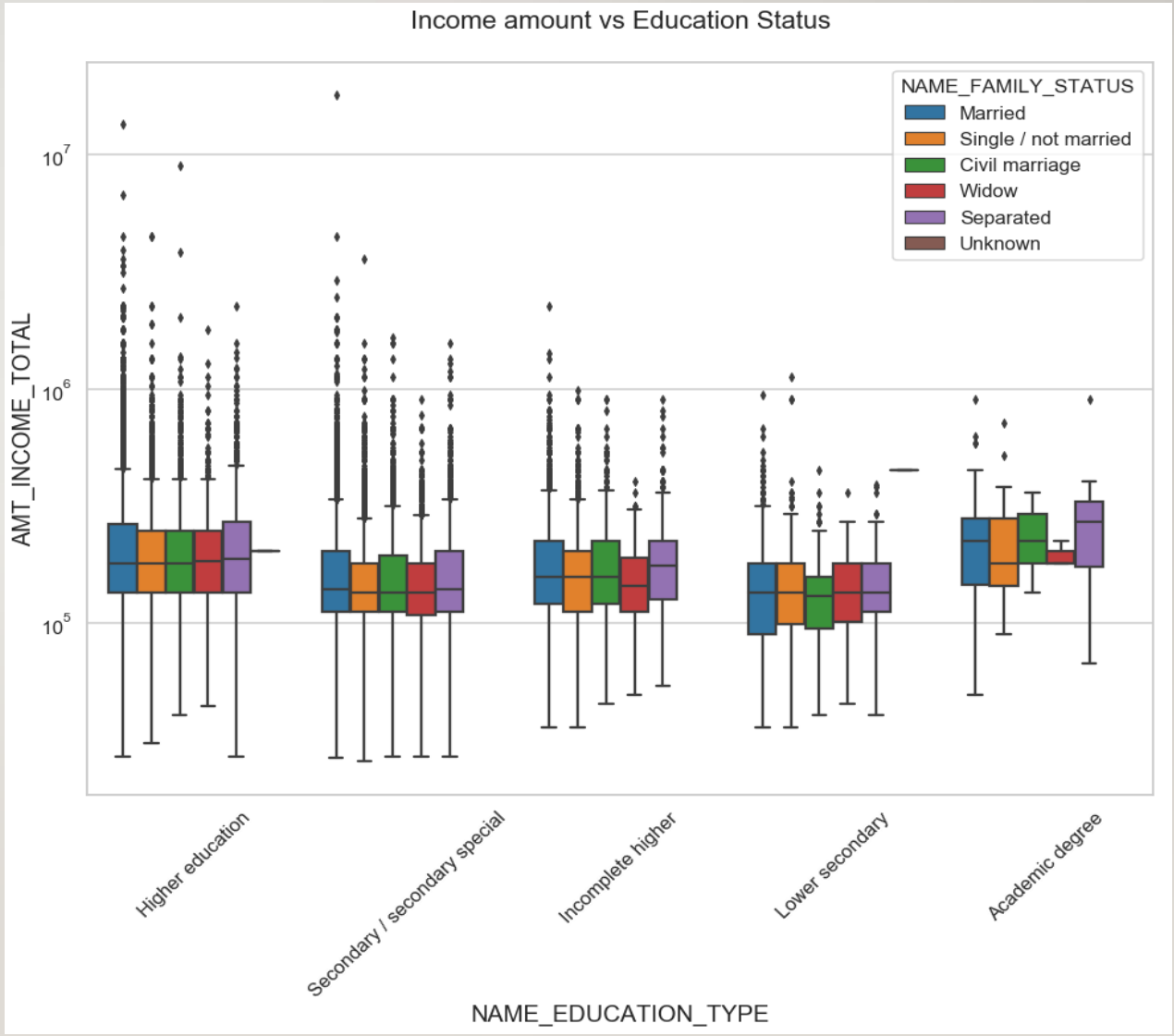
CREDIT AMOUNT VS EDUCATION STATUS

- Family status of 'civil marriage', 'marriage' and 'separated' of academic degree education are having higher number of credits than others.
- 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.



INCOME AMOUNT VS EDUCATION STATUS

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



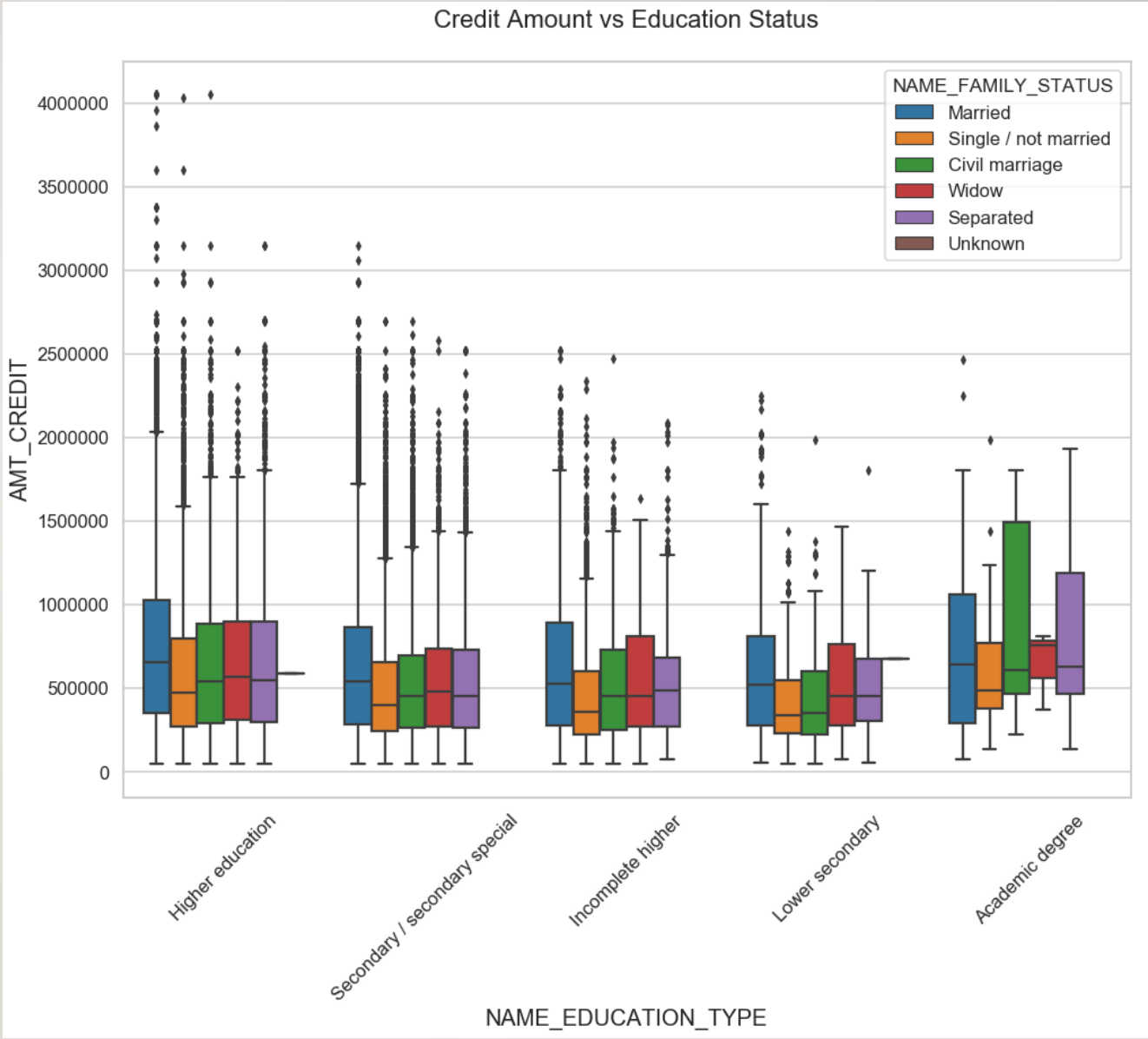
BIVARIATE ANALYSIS FOR TYPE 1

CREDIT AMOUNT VS EDUCATION STATUS

Target 0 is very similar to this. According to the above box plot, academic degree holders with family statuses of "civil marriage," "marriage," and "separated" have more credits than those with other statuses.

Higher education and secondary education make up the majority of the outliers in the data.

Most of the credits for civil unions for academic degrees are in the third quartile.

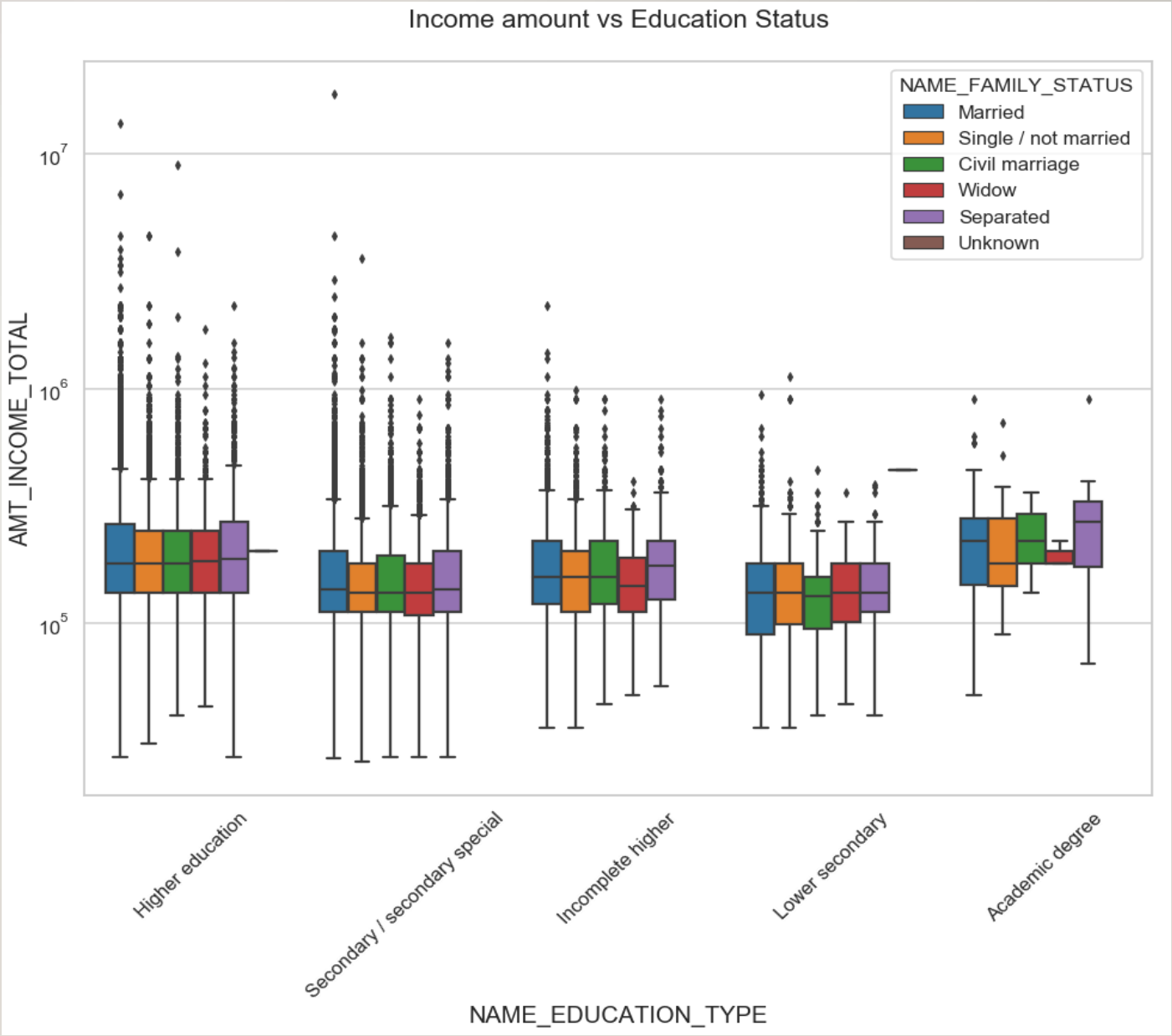


INCOME AMOUNT VS EDUCATION STATUS

Have some similarities to Target 0, and according to the boxplot above for the education category "Higher education," income levels are often comparable to family circumstances.

Less outliers have academic degrees, yet their salary is a little greater than that of those with higher education.

Less money is made than at higher secondary schools.



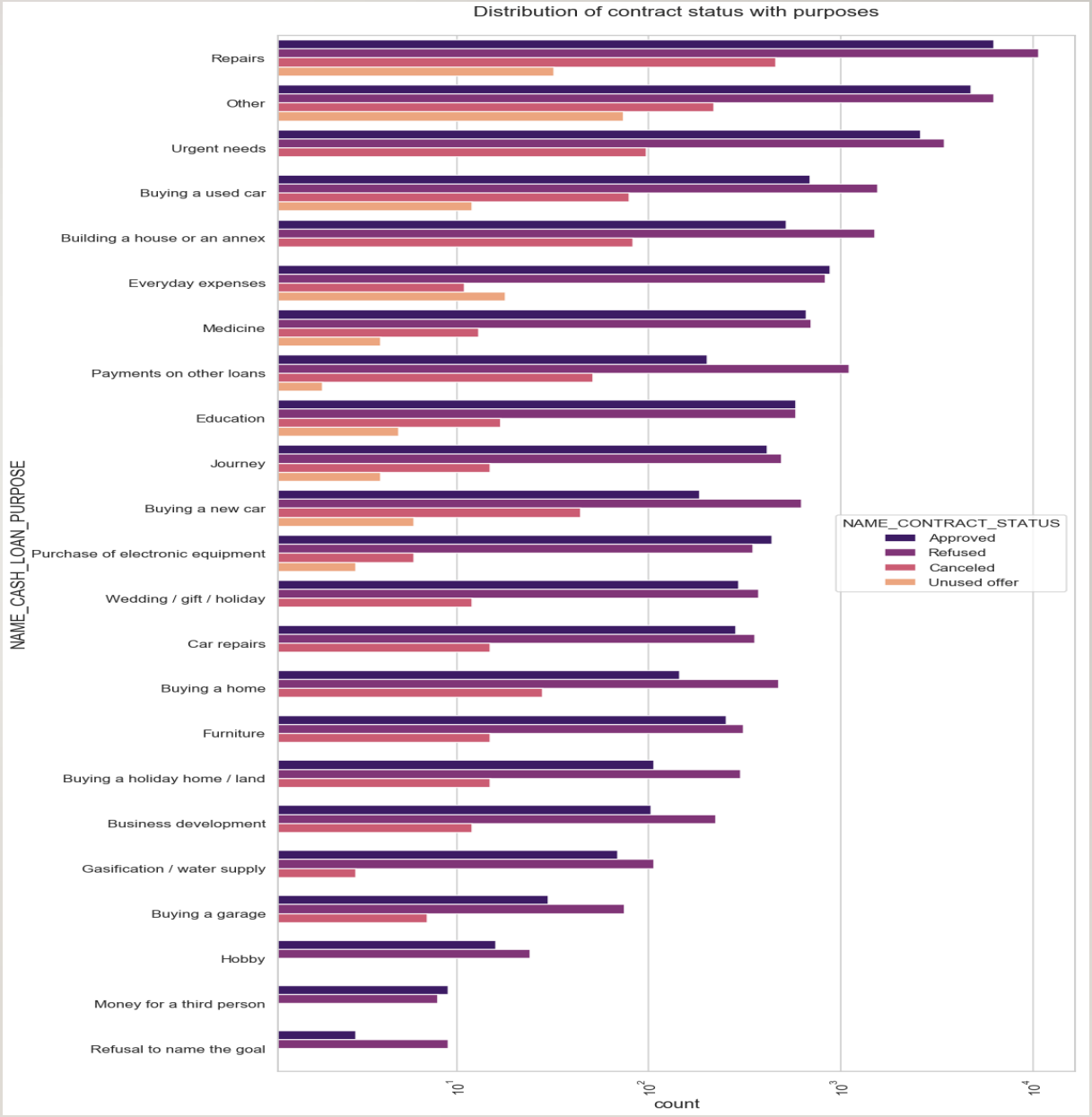
UNIVARIATE ANALYSIS AFTER MERGING PREVIOUS DATA

DISTRIBUTION OF CONTRACT STATUS WITH PURPOSES

1.The majority of loan rejections were for 'repairs'.

2.We have an equal amount of approvals and rejections for educational purposes.

3.Purchase of new cars and other loans are being rejected much more frequently than they are approved.

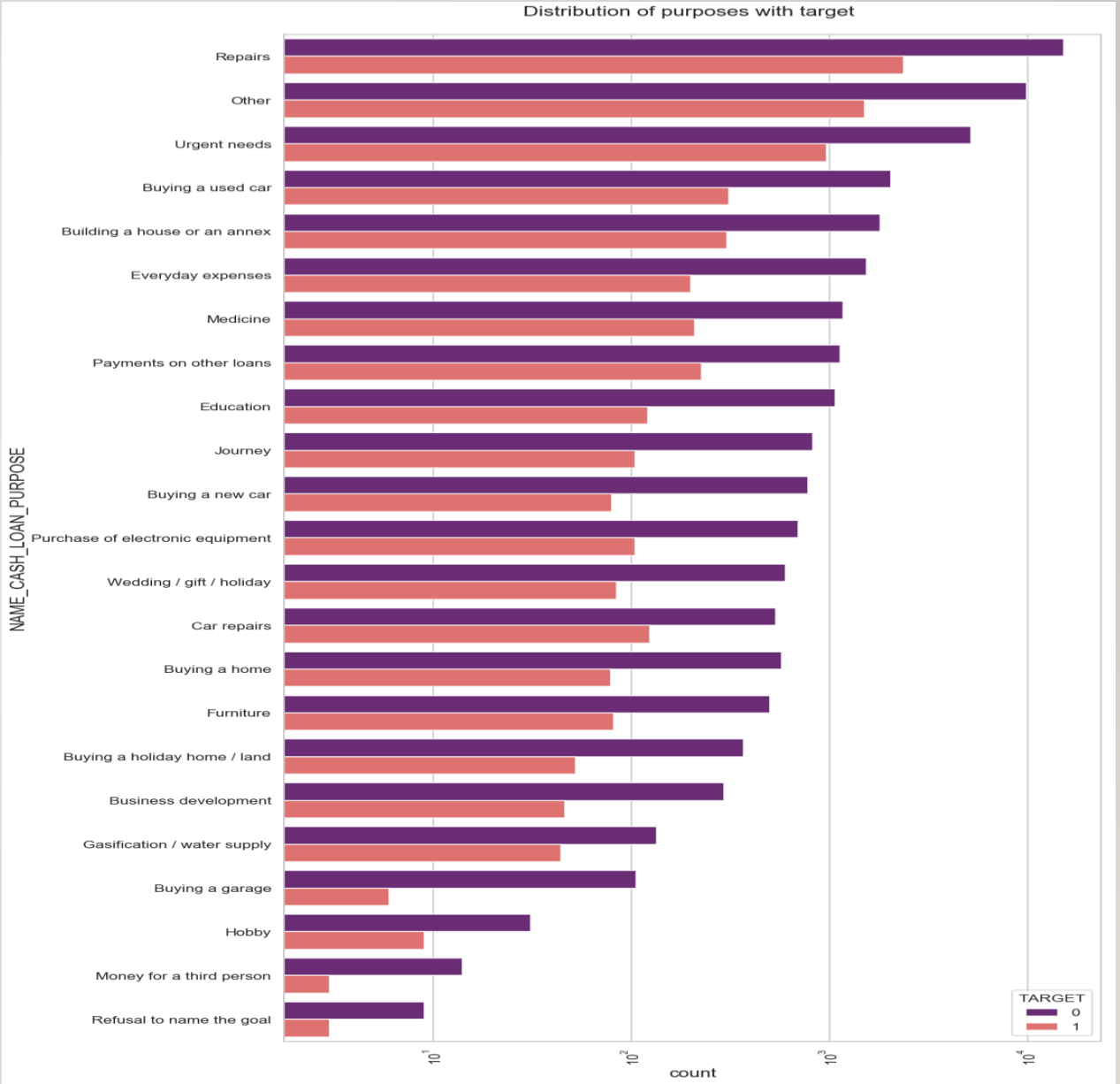


DISTRIBUTION OF PURPOSES WITH TARGET

1.Loan purposes including "Repairs" have more difficulty being paid on schedule.

2.The cost of a loan is rarely significantly higher than when one is struggling. They are "buying a garage," "buying a new car," "buying land," "business development," and "education."

Therefore, we can concentrate on these goals for which the client is experiencing just minor payment issues.



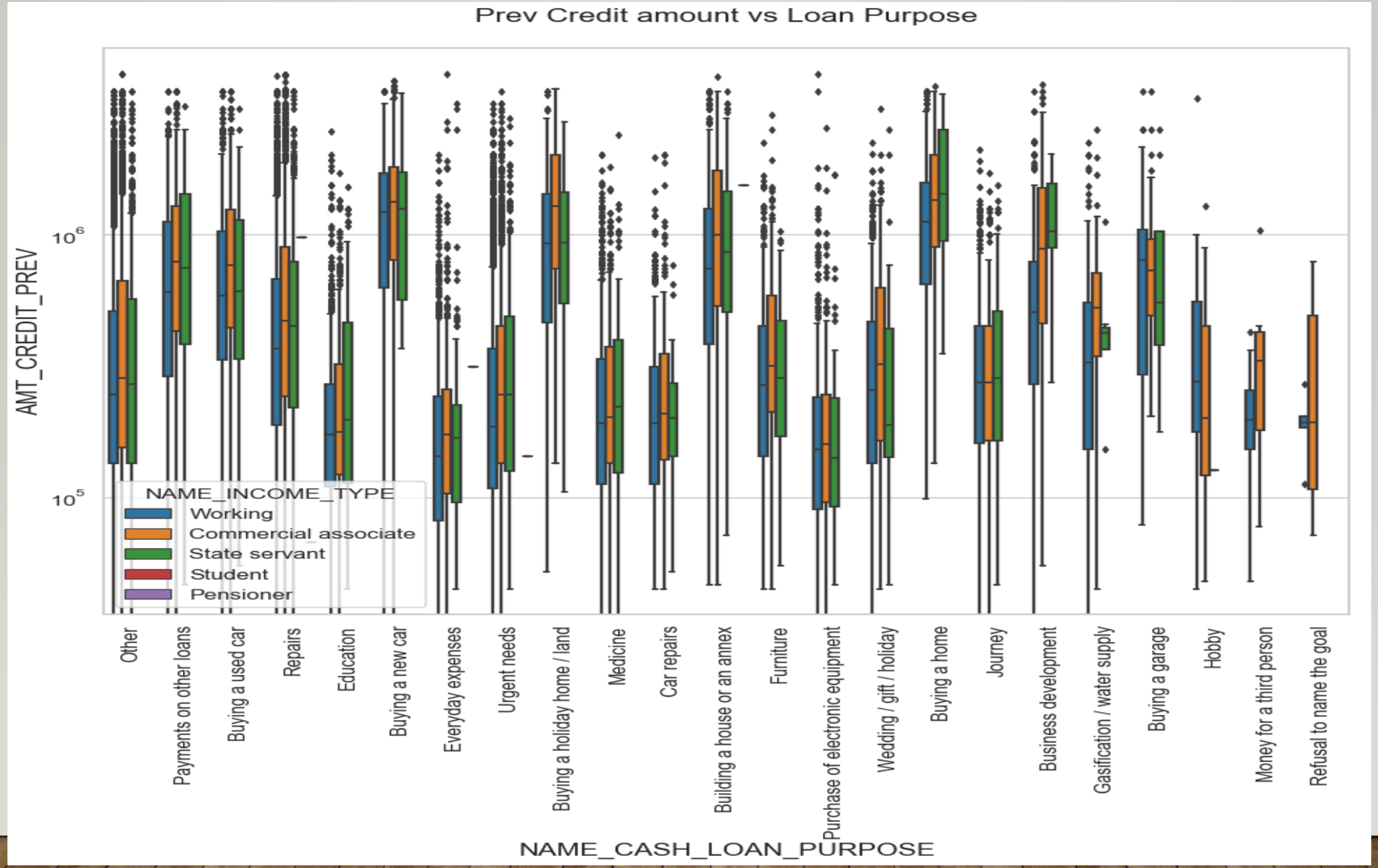
PERFORMING BIVARIATE ANALYSIS

PREV CREDIT AMOUNT VS LOAN PURPOSE

1.For loan objectives such as "buying a home," "buying a land," "buying a new car," and "building a house," the credit amount is larger.

2.State employees' income types receive a substantial amount of credit.

3.Less credits are requested when money is needed for a hobby or third party.

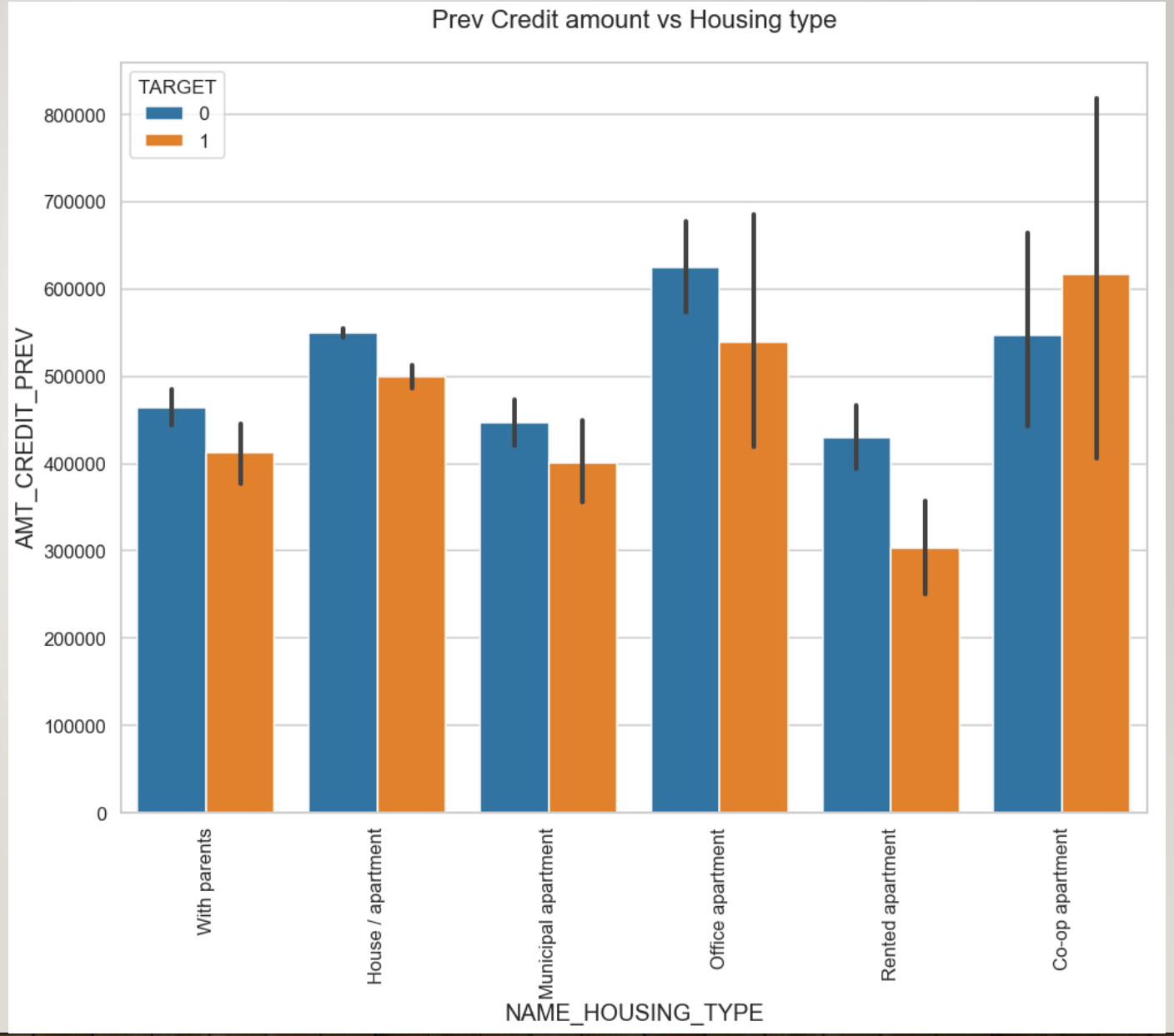

Prev Credit amount vs Loan Purpose
AMT_CREDIT_PREV
10^6
10^5
NAME_INCOME_TYPE
Working
Commercial associate
State servant
Student
Pensioner
Other
Payments on other loans
Buying a used car
Repairs
Education
Buying a new car
Everyday expenses
Urgent needs
Buying a holiday home / land
Medicine
Car repairs
Building a house or an annex
Furniture
Purchase of electronic equipment
Wedding / gift / holiday
Buying a home
Journey
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE

PREV CREDIT AMOUNT VS HOUSING TYPE

Here, the office apartment has a higher target credit of 0 while the co-op apartment has a higher target credit of 1.

We can therefore draw the conclusion that a bank should refrain from lending to the co-op flat housing type because of their payment issues.

For successful payments, the bank can concentrate mostly on dwelling types such as those with parents, houses, apartments, or municipal apartments.



CONCLUSION

- Banks should focus more on contract type ‘student’ ,’pensioner’ and ‘businessman’ with housing ‘type other than ‘co-op apartment’ for successful payments.
- Banks should focus less on income type ‘working’ as they are having most number of unsuccessful payments.
- Also with loan purpose ‘repair’ is having higher number of unsuccessful payments on time.
- Get as much as clients from housing type ‘with parents’ as they are having least number of unsuccessful payments.

THANK YOU