Exploratory Data Analysis (EDA) on Loan Applications

A case study in Risk Analytics in Banking and Financial Services

BY NEHA MOHD SHAHID

PROBLEM STATMENT

When a company issues a loan, there are two main risks:

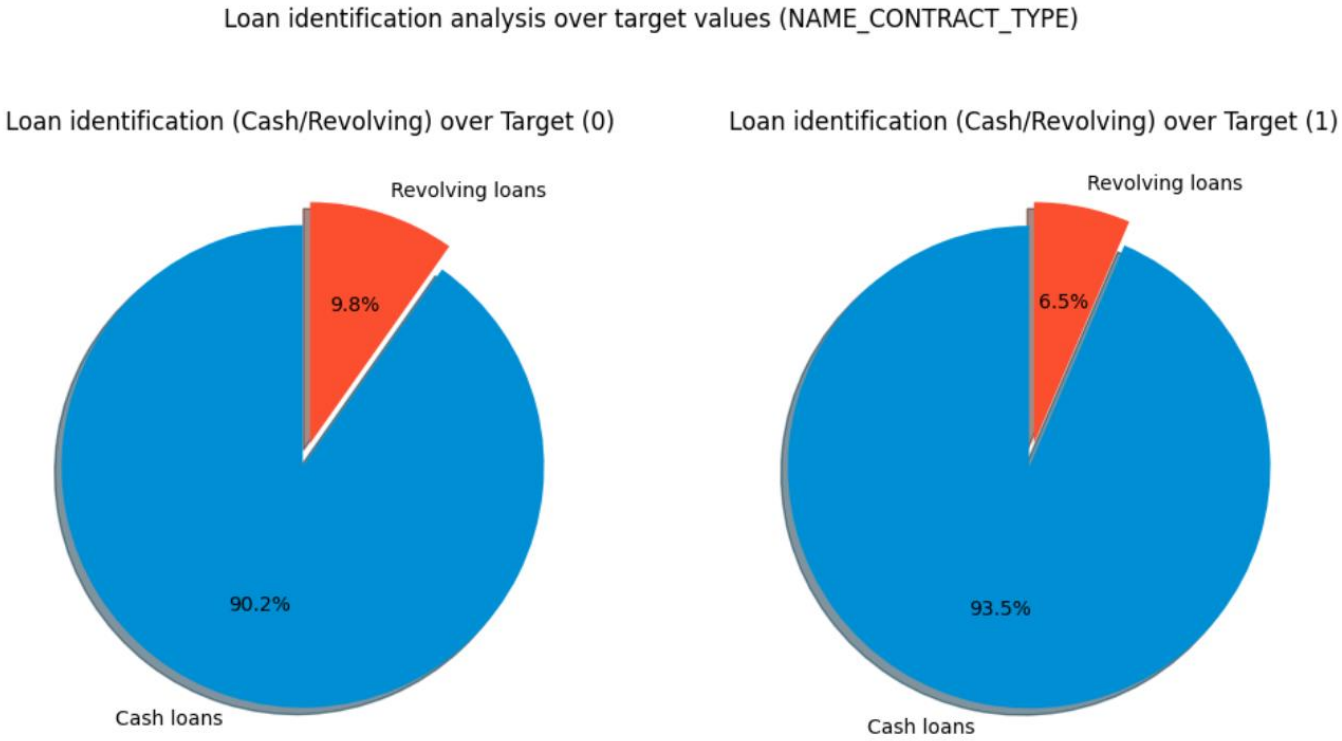
1. Approving a Loan for a Risky Applicant: If the applicant is likely to default, the company could incur financial losses.
2. Rejecting a Loan for a Safe Applicant: If the applicant is likely to repay but the loan is denied, the company misses out on potential business.

By conducting Exploratory Data Analysis (EDA) on the company's data, we can analyze patterns and help mitigating these risks.

Univariate Analysis

Analysis of Loan Types in Relation to Default Risk

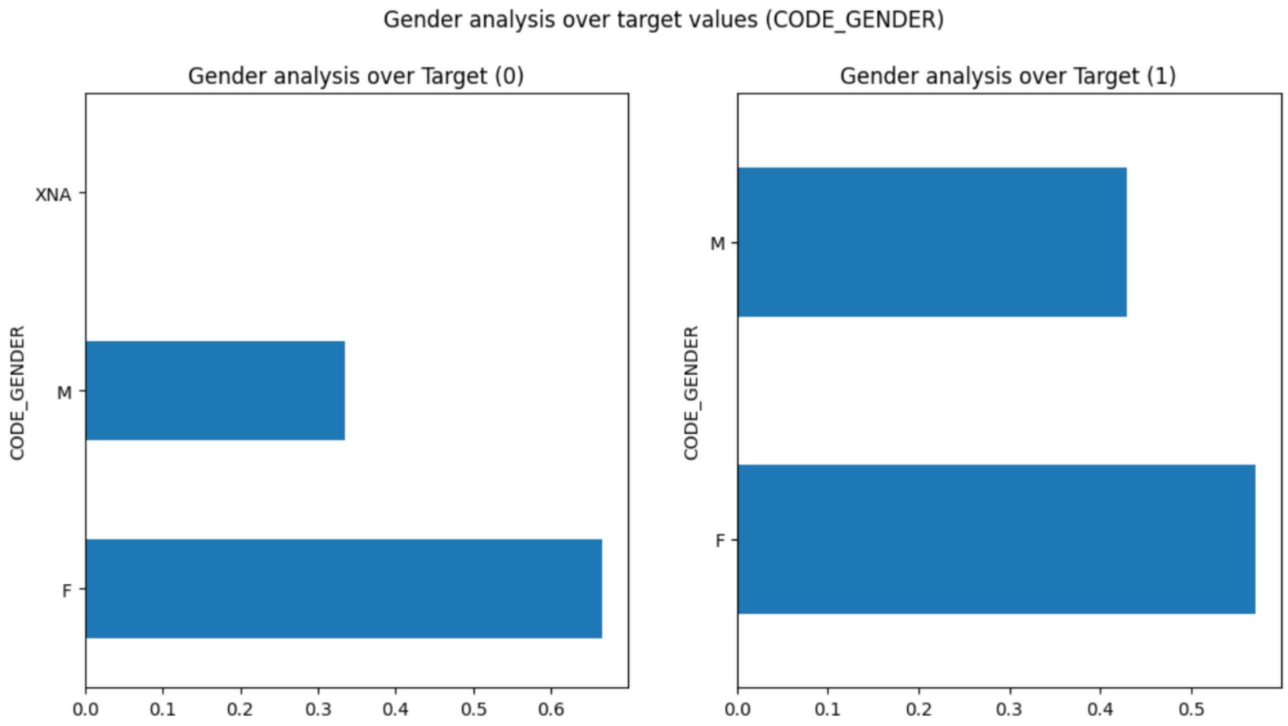
- Cash loans are the majority in both categories, making up 90.2% of non-defaulted loans and 93.5% of defaulted loans.
- The higher percentage of revolving loans in the non-defaulted category suggests that these loans are less likely to default compared to cash loans.
- The decrease in the proportion of revolving loans in the defaulted category (from 9.8% to 6.5%) may indicate that revolving loans are associated with better repayment behavior or lower default risk.



Gender Analysis in Relation to Loan Default Risk

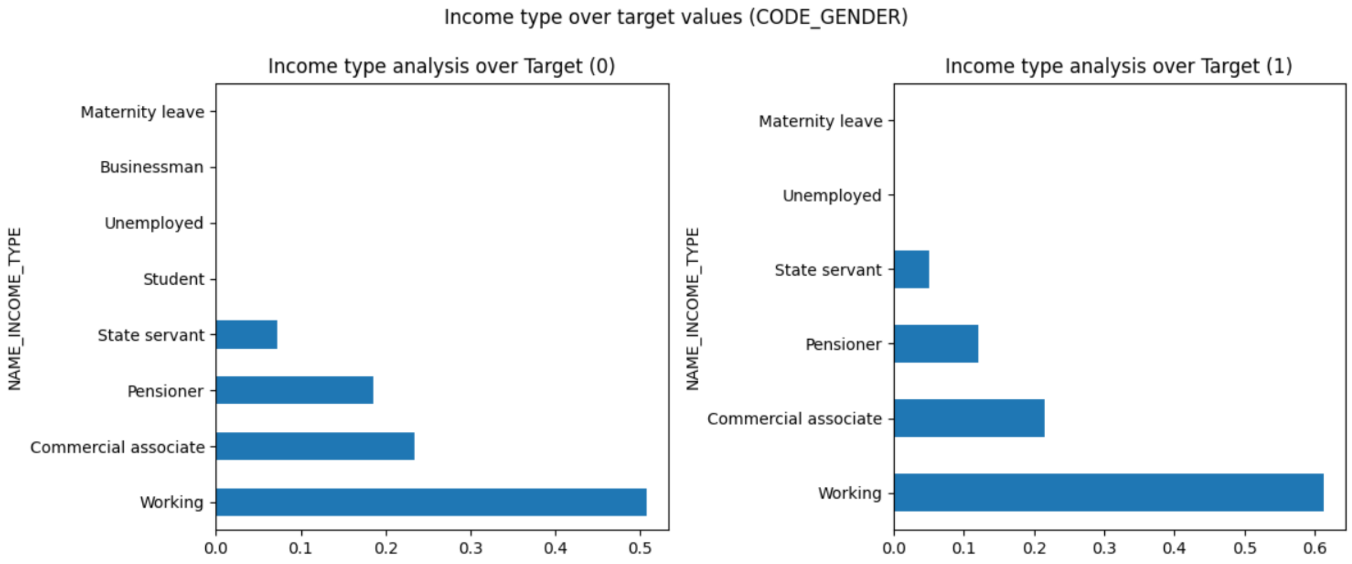
Female borrowers tend to default less frequently compared to their male counterparts as their proportion decreases from 65% in the non-defaulted category to 55% in the defaulted category.

The data suggests that gender may be a relevant factor in assessing the risk of loan default. Female borrowers appear to be slightly lower risk, while male borrowers show a higher tendency toward default.



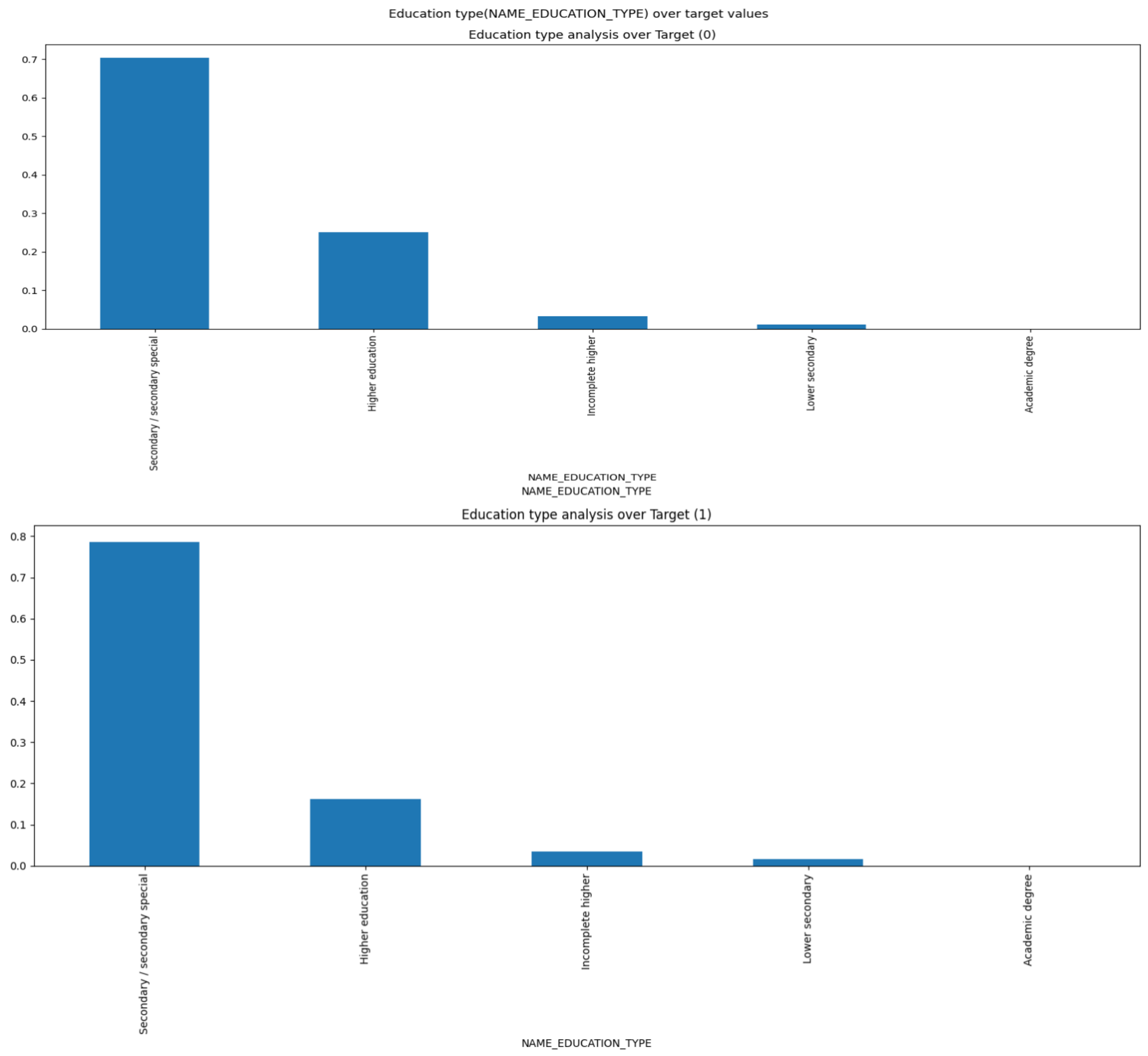
Income Type in Relation to Loan Default Risk

- **Working individuals** proportion is slightly higher among defaulted loans (around 60%) compared to non-defaulted loans (around 50%). It indicates they are more likely to default.
- Pensioners tend to default less frequently compared to other target values as their slightly proportion decreases from 22% in the non-defaulted category to 18% in the defaulted category.
- Pensioners are more likely to pay the loans



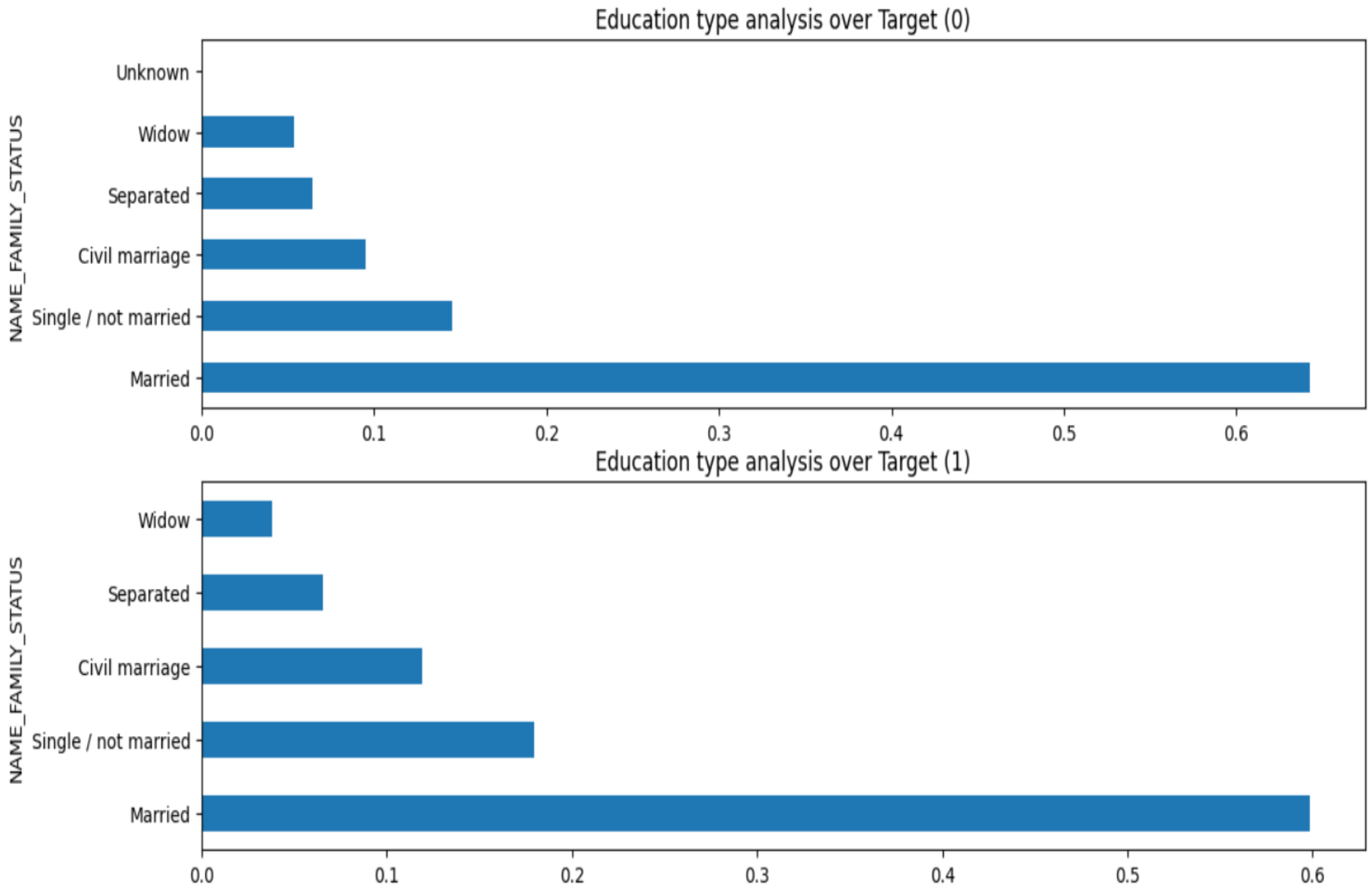
Education Type and Payment Difficulties

- Clients with "Secondary / Secondary Special" education make up the majority in both target groups. However, this proportion is slightly higher in clients with payment difficulties (Target 1).
- Higher education clients represent a smaller proportion in both groups, with a more noticeable presence in clients without payment difficulties (Target 0).
- Clients with higher education are more likely to repay loans



Family Status and Payment Difficulties

- **Married clients** are the most common in both groups (Target 0 and Target 1), but slightly more prevalent among those without payment difficulties (Target 0).
- **Single/not married clients** are the second-largest group in both targets, with a higher proportion in clients with payment difficulties (Target 1).
- Widowed clients are small minority, but are most likely to repay loans

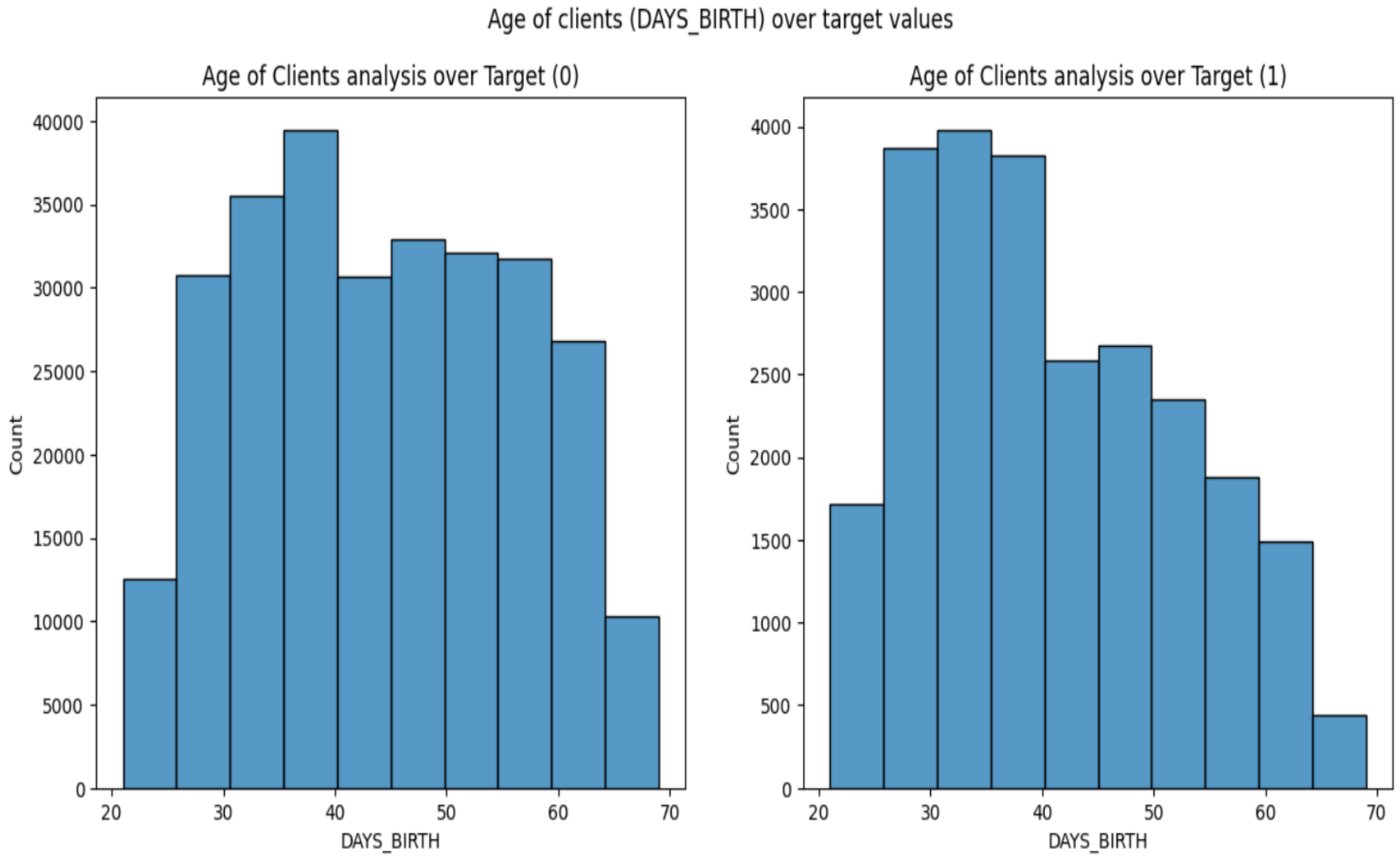


Age of Clients and Payment Difficulties

As the age of clients increases from 40 years, the proportion decreases in the non-defaulted category to the defaulted category.

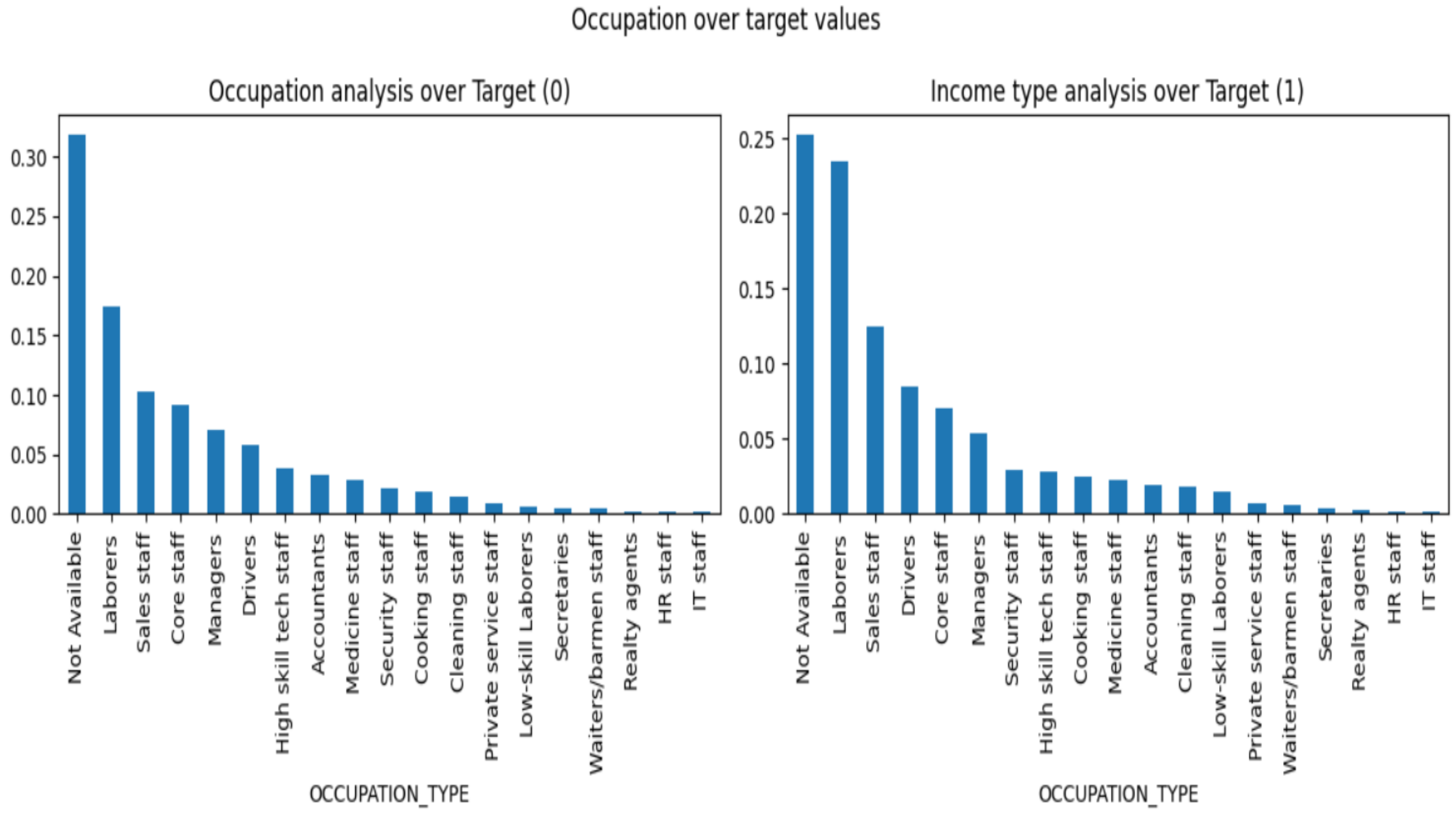
It indicates that

- Older clients (greater than 40 yrs of age) are most likely to repay loans.
- Younger clients are most likely to default on loan payments



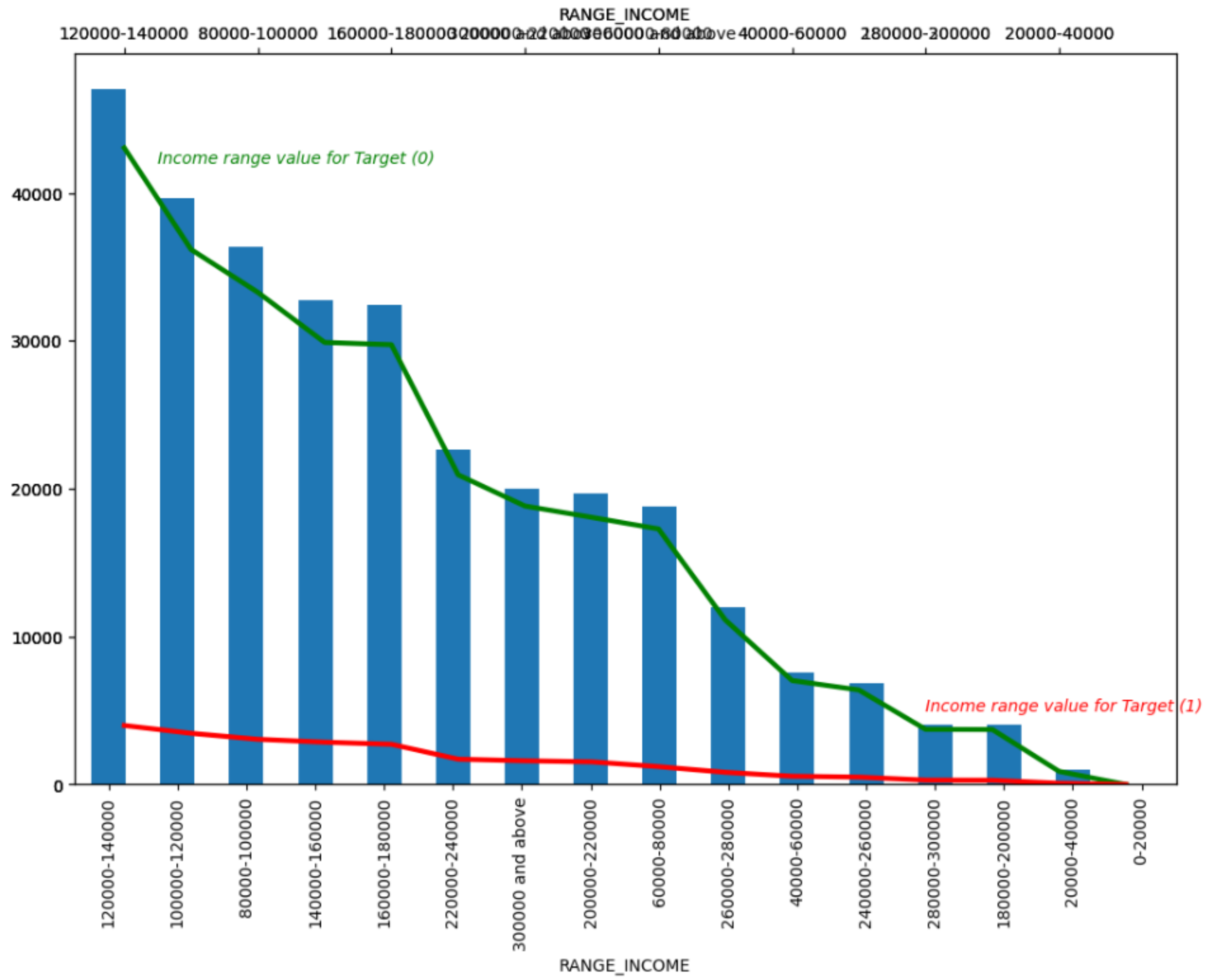
Occupation in Relation to Loan Default Risk

- Clients with occupation as laborers are most likely to default on loan payments.
- Sales staff are also most likely to default on payments.
- HR and IT staff are less likely to apply for loan



Range Income in Relation to Loan Default Risk

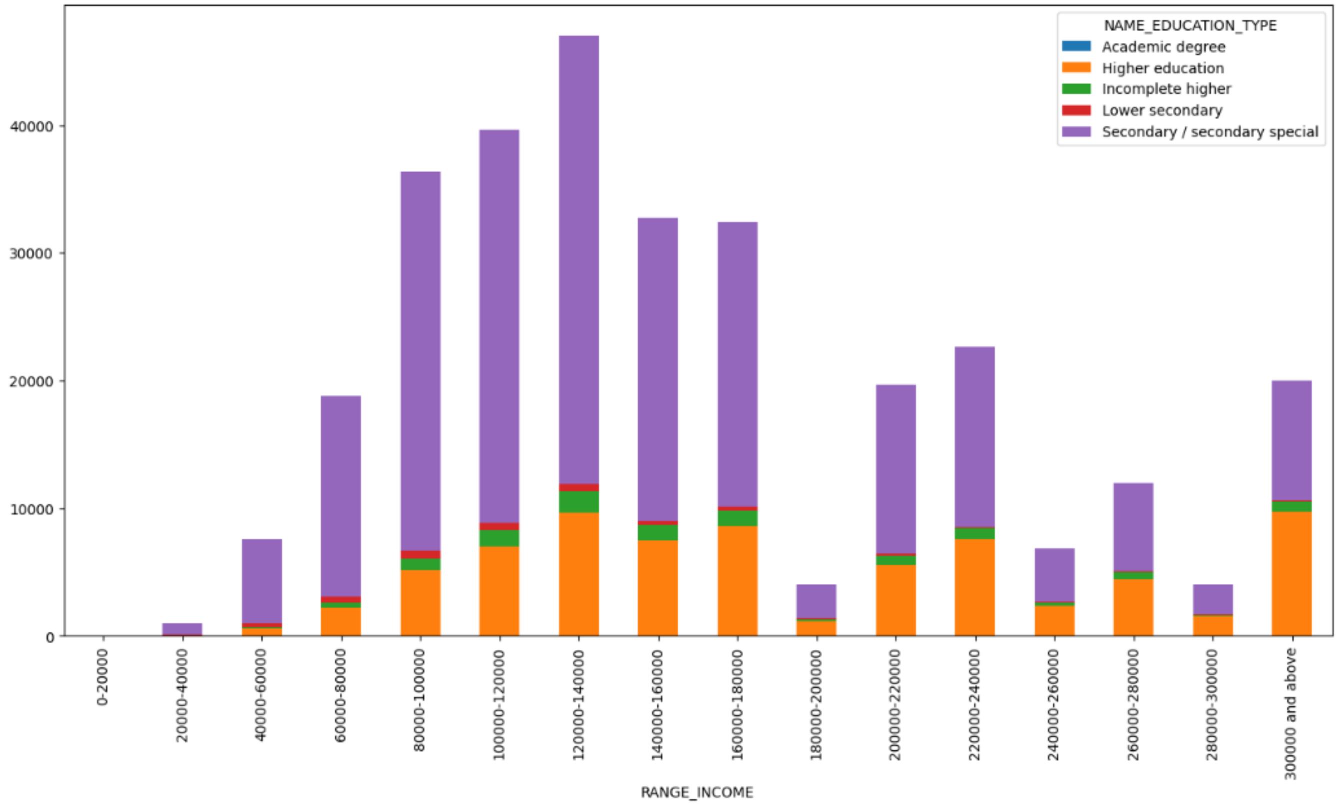
- Clients with higher income range brackets greater than 240000 are most likely to repay the loans.
- Clients with moderate income range bracket i.e., b/w 80000 to 140000 are more likely to default loans.
- Clients with low-income ranges are less likely to apply for loans.



Bivariate Analysis

Education Type VS Income Range

- Clients with low income ranges i.e., less than 80000 have very high percent of clients with secondary level education.
- Clients with mid income range have a balance of education with secondary and higher education.
- Clients with high income range have relatively higher percent of higher education holders.



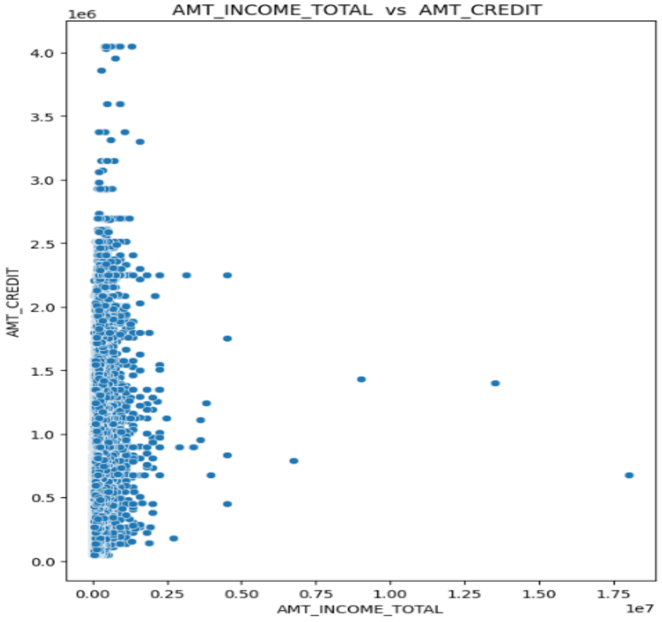
Income vs Credit, Goods price vs Credit

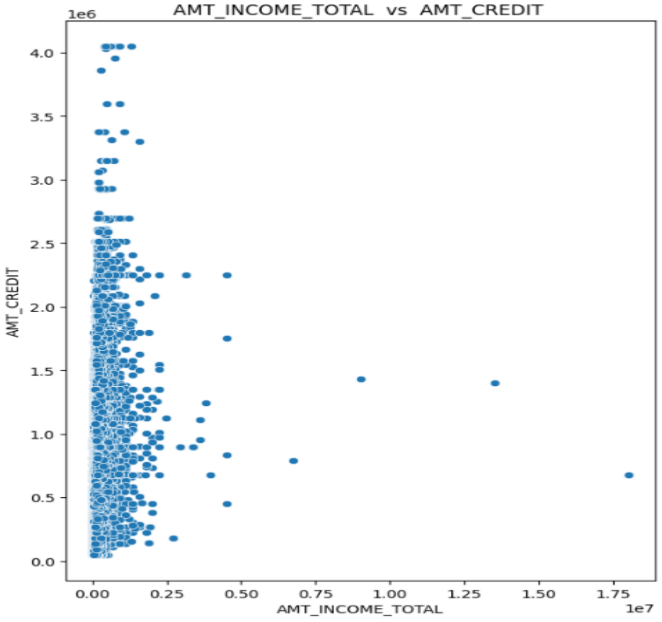
- Analysis inference of income and credit amount.

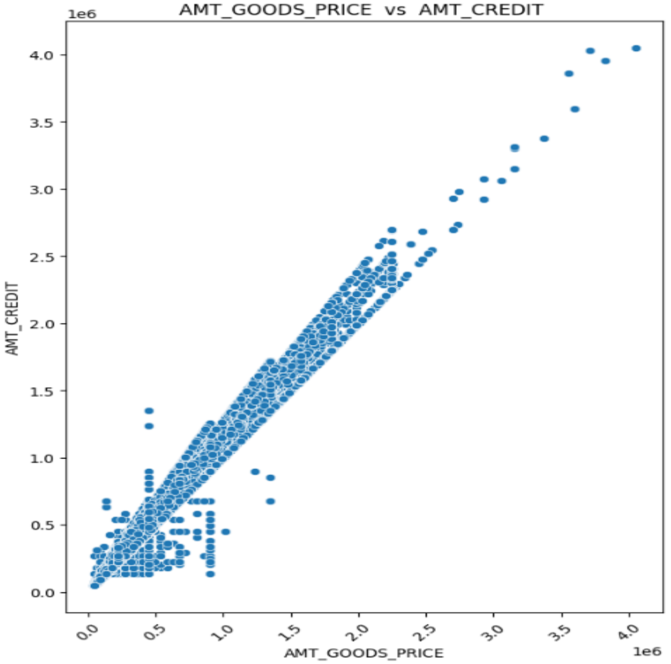
Those who have paid the loan amount on/within time are more likely to get higher credits than those who didn't pay/did late payments.

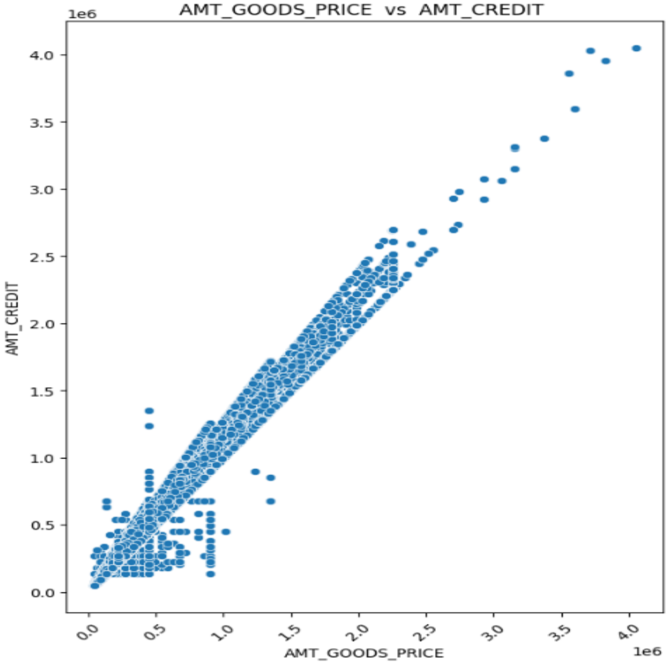
- Analysis inference of goods price and credit amount.

People who have higher goods price and have made payments on time have higher credits than those with higher goods price but didn't pay loan.



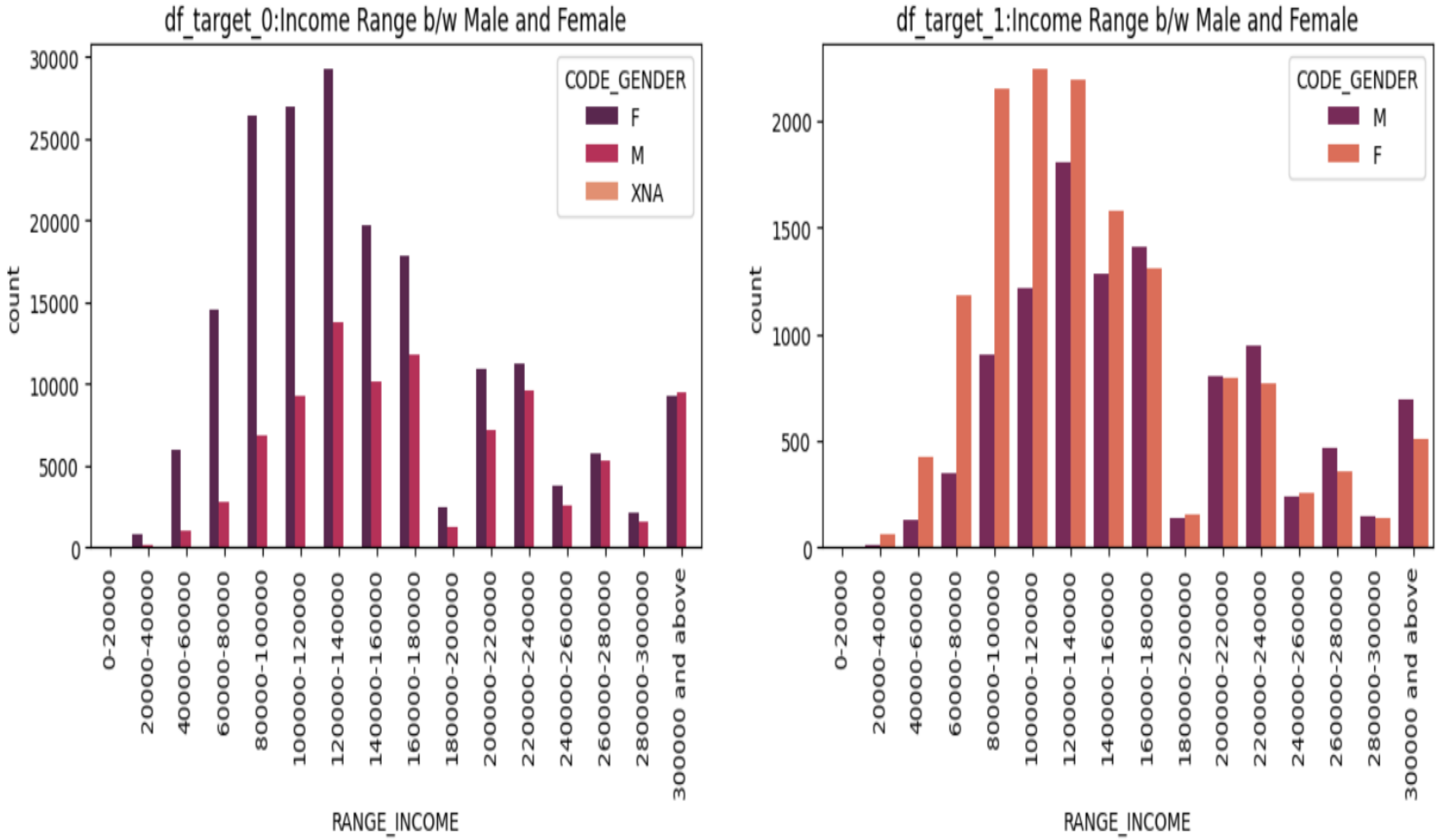






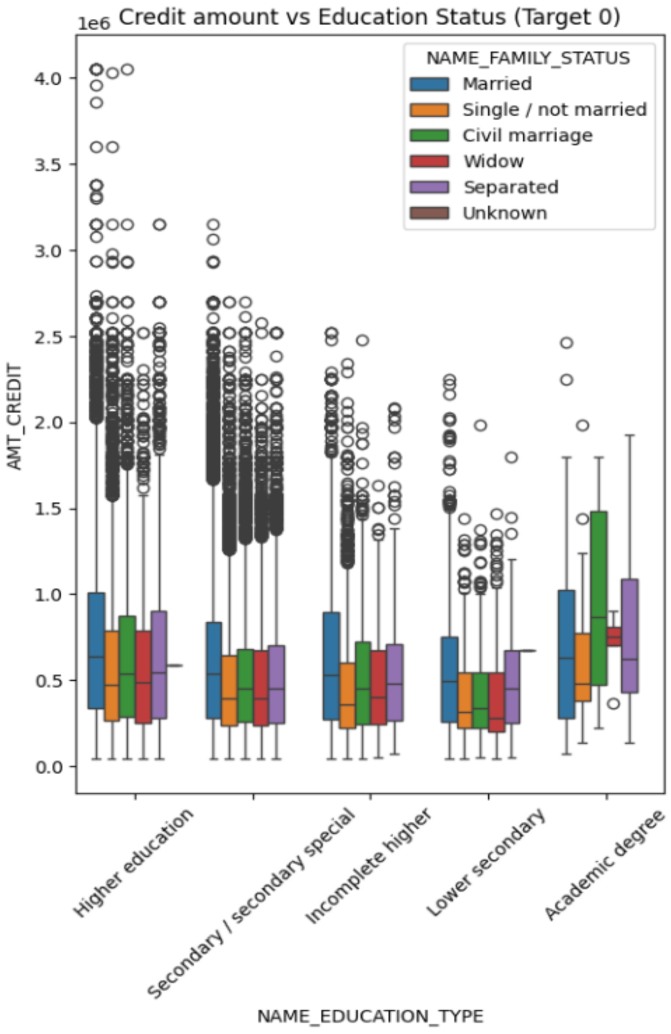
Income Range vs Gender

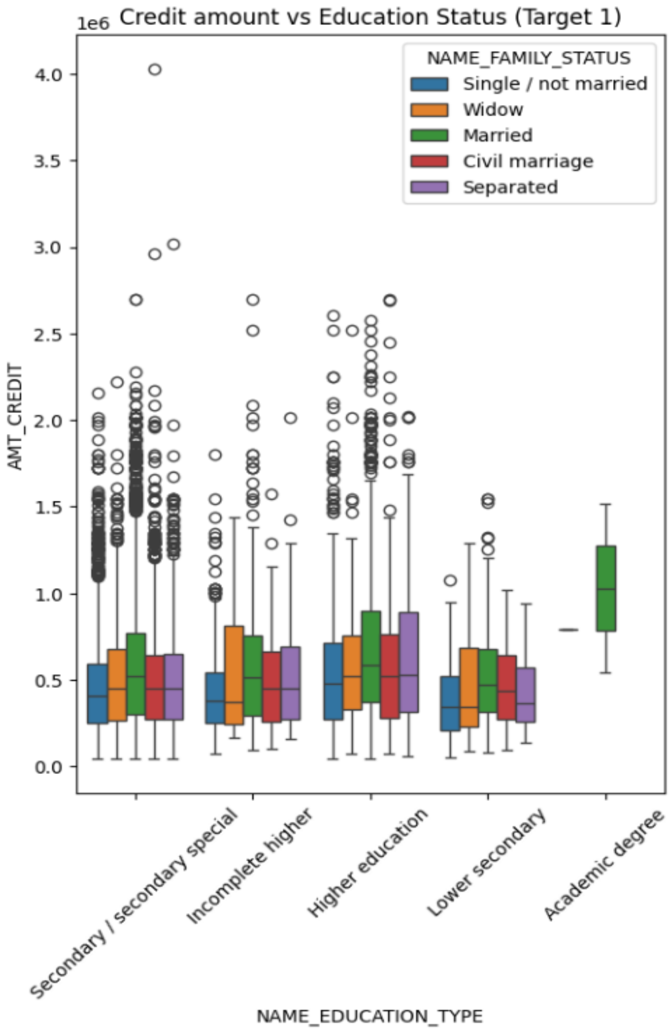
- In both graph, females dominate the income ranges up to 200,000, suggesting that they are the primary earners in these brackets.
- Females, especially in the lower to mid-income ranges, appear more likely to repay loans, possibly due to stable incomes and fewer high-risk loans.
- Males are less represented in lower and mid-income ranges but show a slight increase in the highest income bracket (300,000 and above), particularly in ‘df_target_1’.
- Males, particularly in higher income brackets, have a strong repayment capacity due to their selective loan-taking behavior, often linked to higher incomes and strategic borrowing.



Credit amount vs Education Status

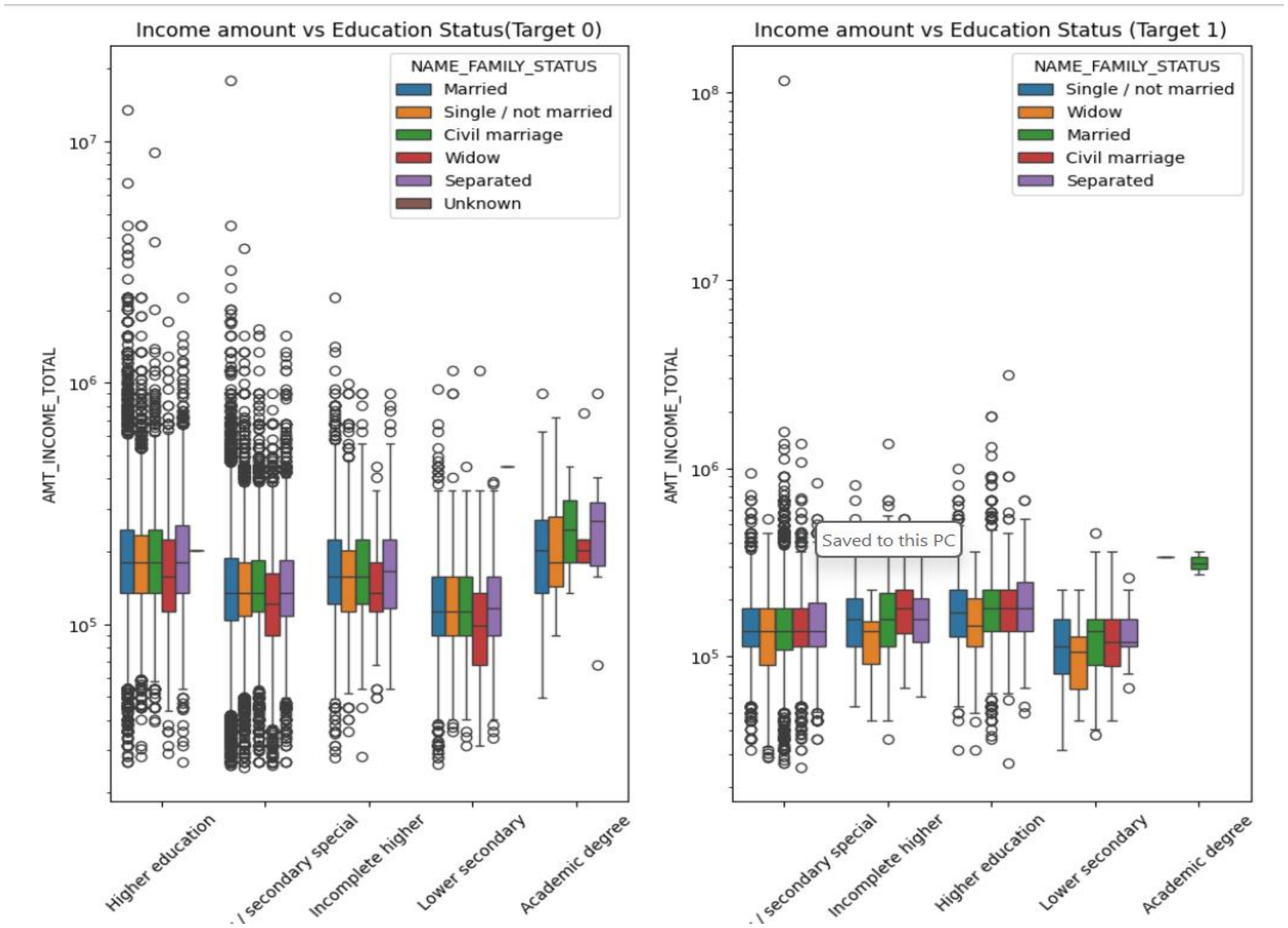
- Family status of 'civil marriage', 'marriage' and 'separated' of Academic degree education are having higher number of credits than others.
- Civil marriage for Academic degree is having most of the credits in the third quartile.
- Higher education of family status of 'marriage', 'single' and 'civil marriage' are having more outliers.
- Some of the highly educated, married person are having credits higher than those who have done lower secondary education.
- Those with higher education have higher credits and are more likely to make payments on time



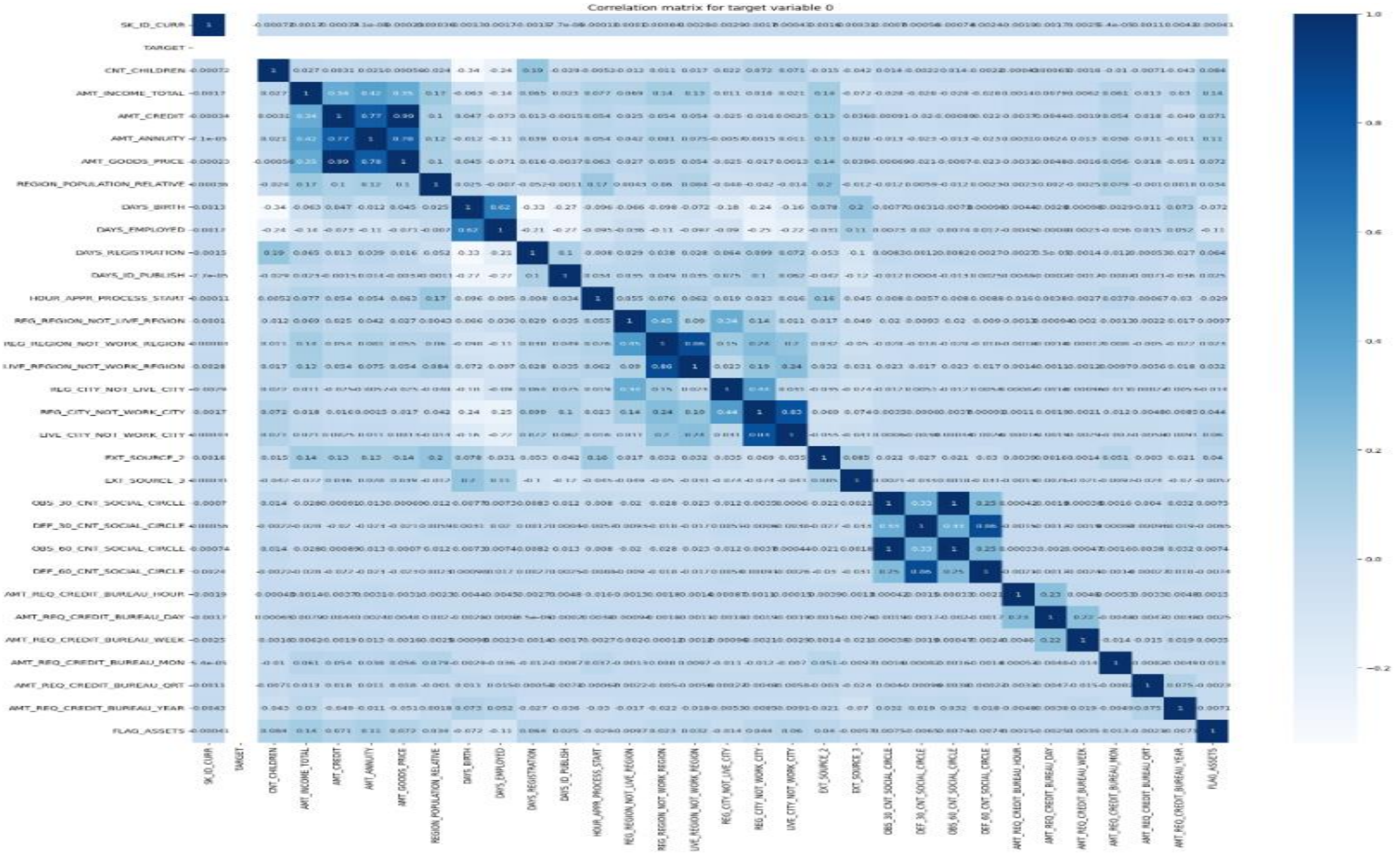


Income amount vs Education Status

- 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
- We can see that People with higher education have higher income and don't have difficulties in making loan payment.
- People with higher education who have lesser income are unable to pay the loan. Hence we can conclude that, people with Higher income are most likely to make payments.

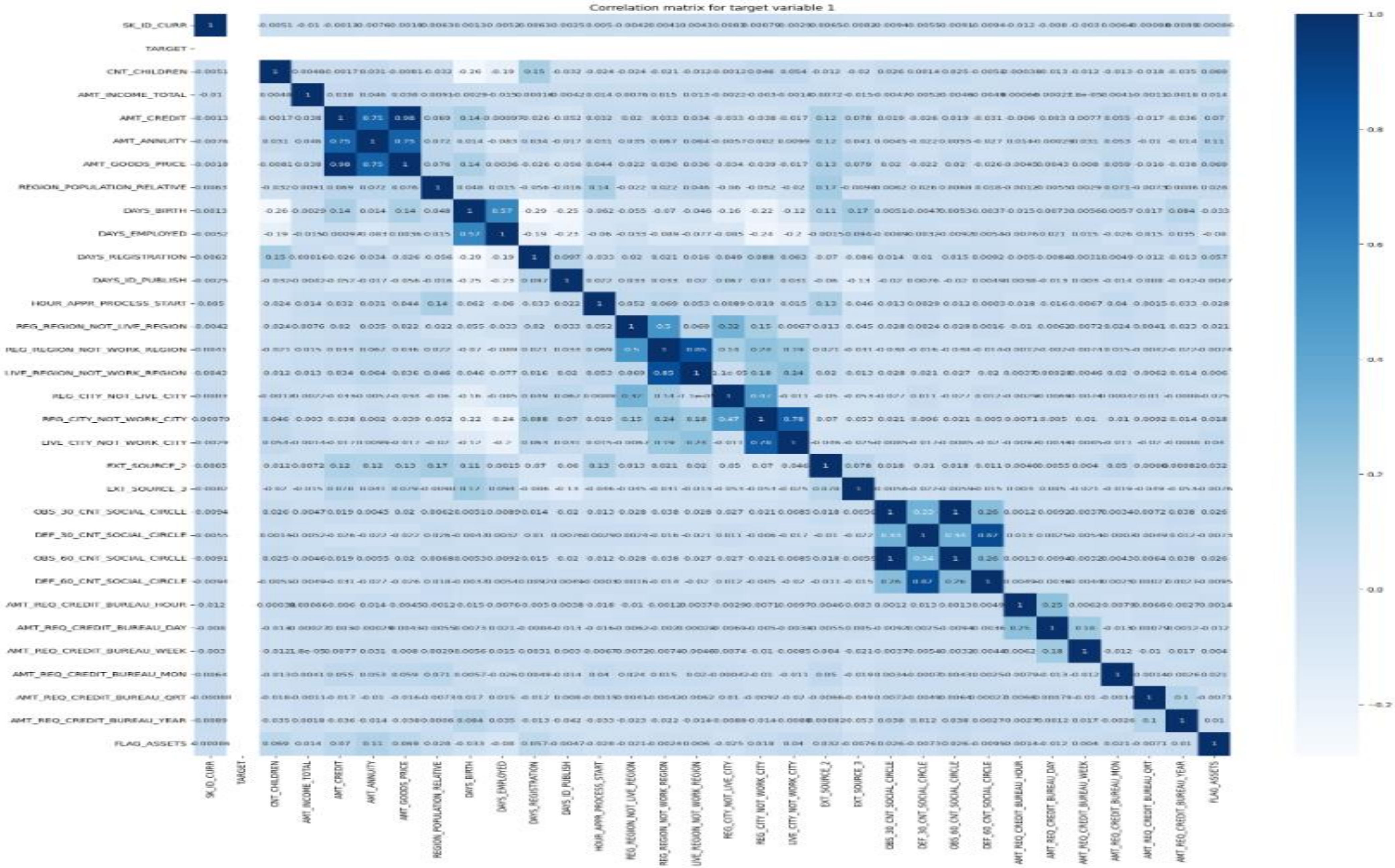


Corelation Analysis


Correlation matrix for target variable 0
SK_ID_CURR
TARGET
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
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
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
FLAG_ASSETS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Correlation For Target 0

- There is an inverse relationship between credit amount and age (DAYS_BIRTH), indicating younger clients are more likely to have higher credit amounts.
- Credit amounts tend to be higher for clients with fewer children, suggesting a negative correlation.
- There’s an inverse relationship between income amount and the number of children, implying higher incomes for clients with fewer children.
- Clients in densely populated areas tend to have fewer children, as indicated by the negative correlation between REGION_POPULATION_RELATIVE and the number of children.
- The correlation shows that credit amounts are generally higher in densely populated areas, suggesting more credit activity in these regions.
- Income tends to be higher in densely populated areas, aligning with economic opportunities often available in such locations.


Correlation matrix for target variable 1
SK_ID_CURR
TARGET
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
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
AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK
AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
FLAG_ASSETS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Correlation For Target 1

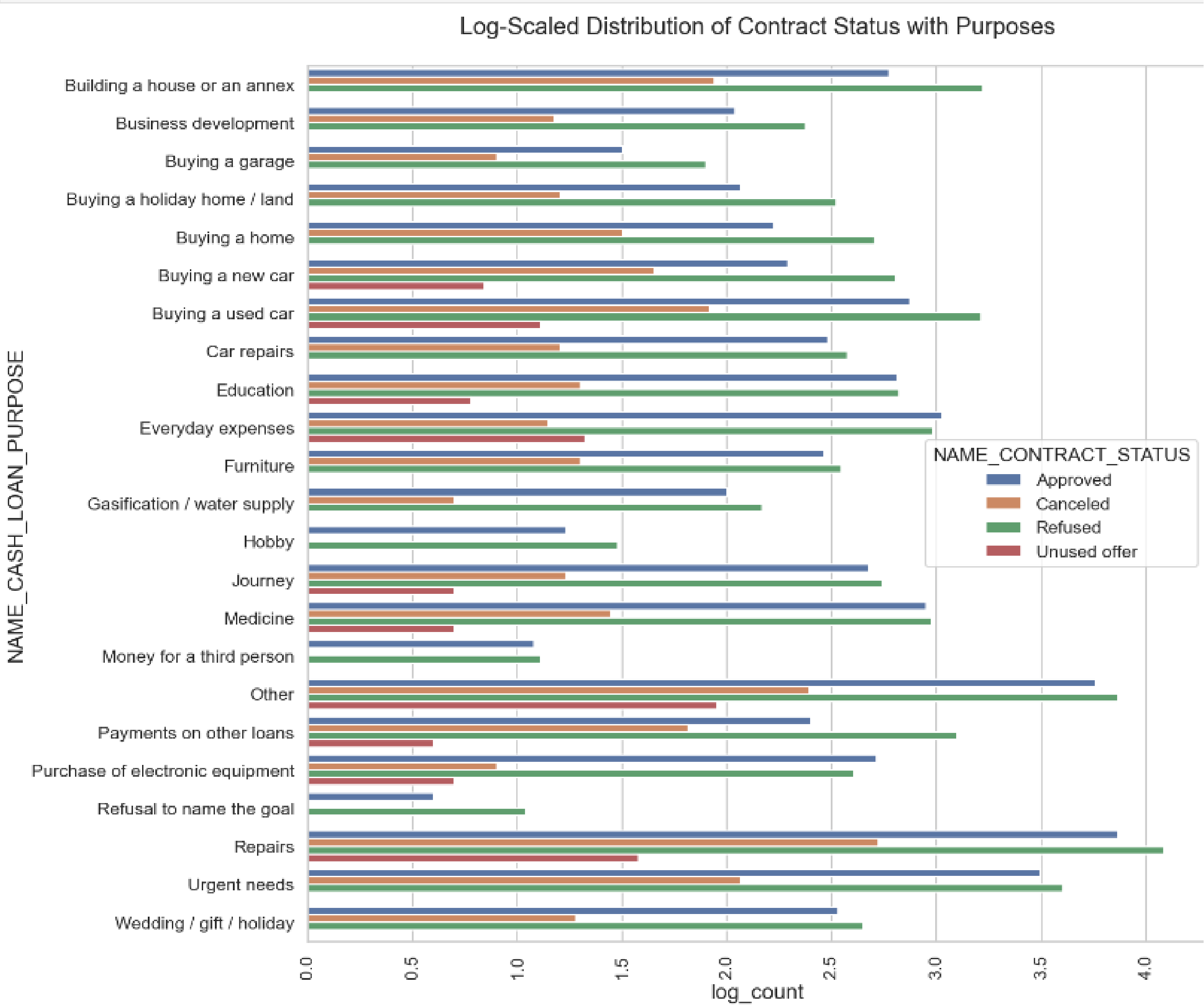
This heat map for Target 1 is also having quite a same observation just like Target 0. But for few points are different. They are listed below.

- Clients whose permanent address does not match their contact address tend to have fewer children, highlighting mobility or lifestyle differences.
- There's a similar observation where clients whose permanent address does not match their work address also tend to have fewer children, which could reflect work-related relocations.

Analysis after merging previous data

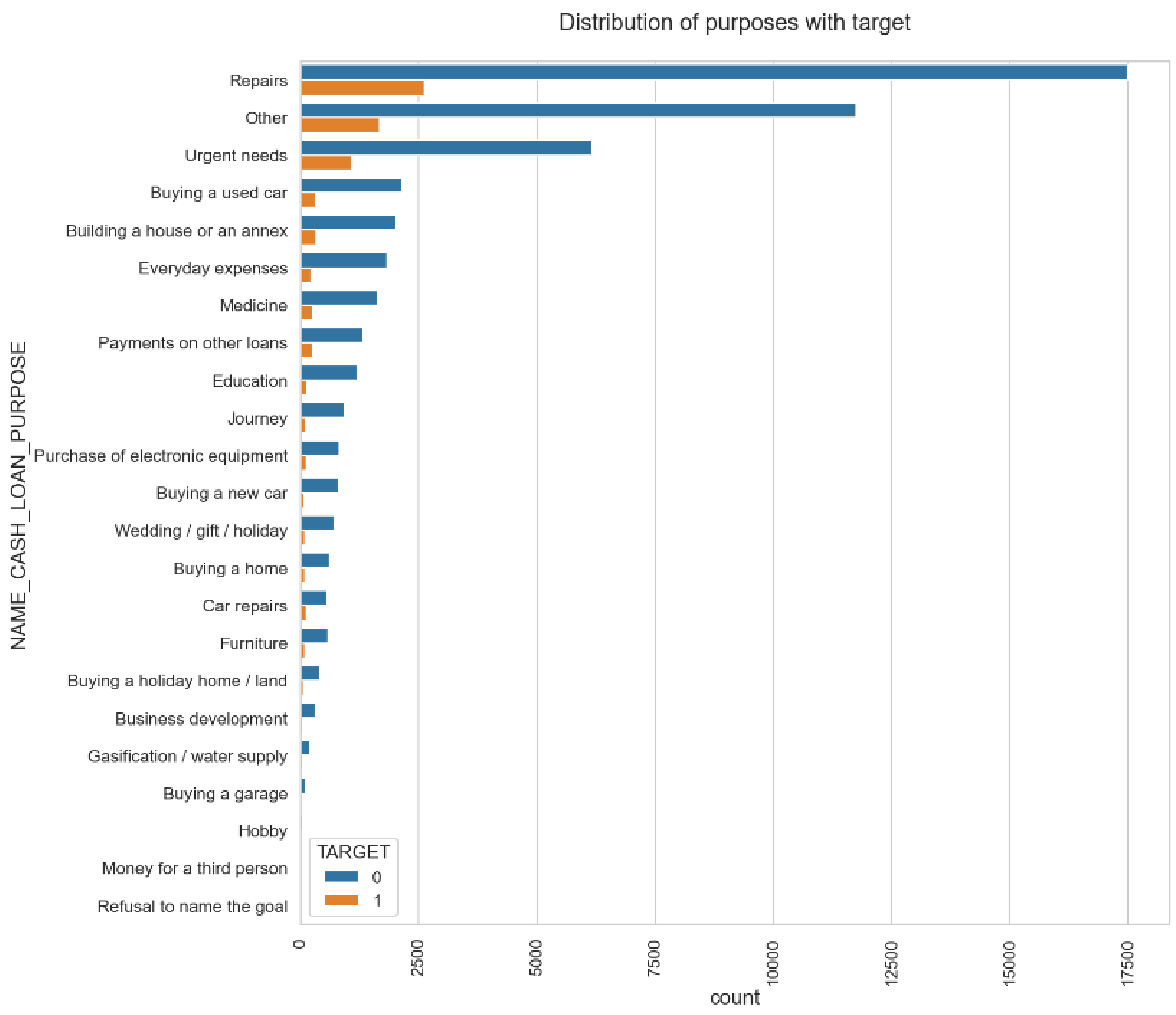
Distribution of contract status with purposes

- Most rejection of loans came from purpose 'Repairs'.
- For education purposes we can see equal number of approves and rejection.
- Purchase of Electronic equipment has higher approved than rejection.



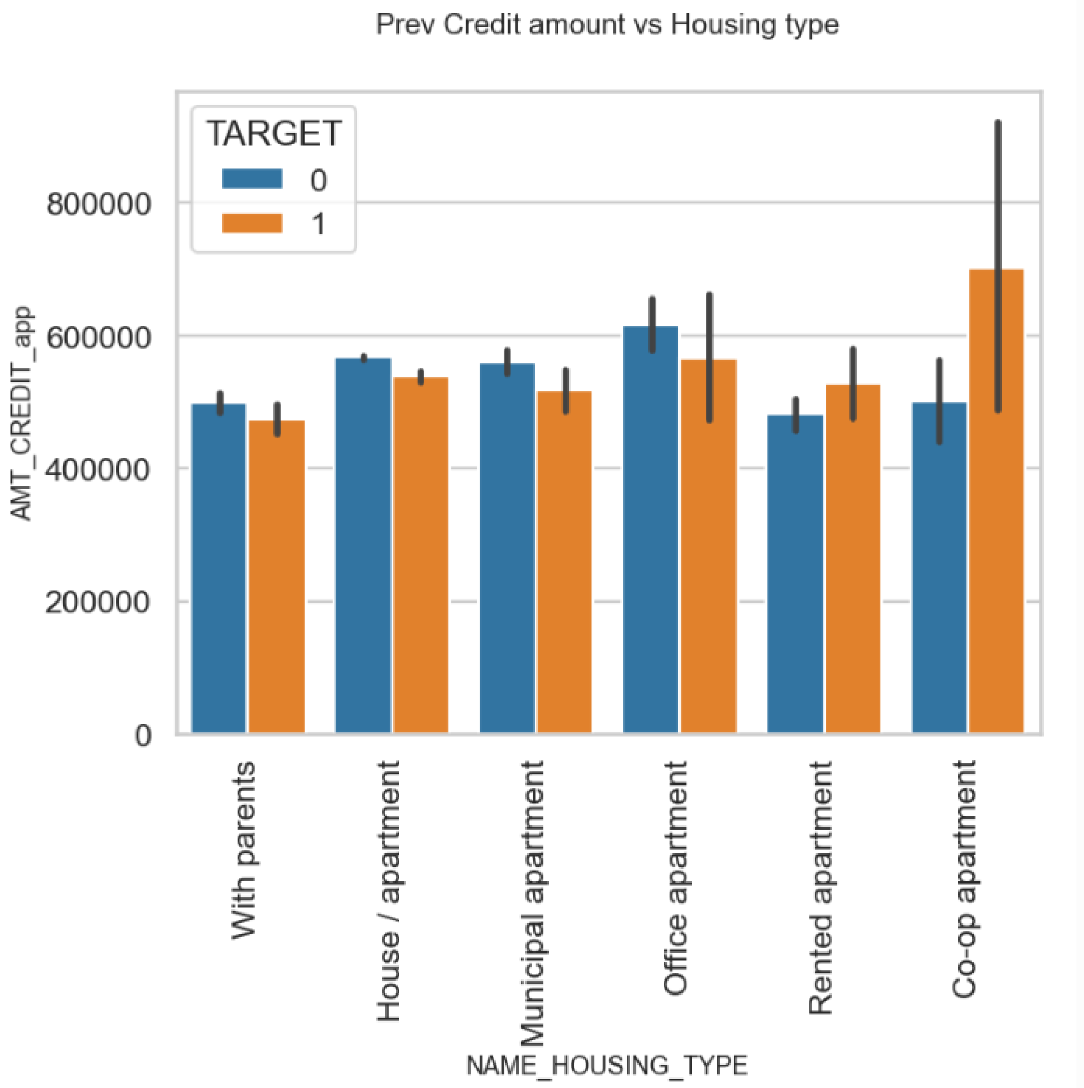
Distribution of purposes with target

Loan purposes with 'Repairs' are facing more difficulties in payment on time. loan payment is significant higher than facing difficulties.



PREV CREDIT AMOUNT VS HOUSING TYPE

- Here for Housing type, office apartment is having higher credit for target = 0 and co-op apartment is having higher credit for 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

- The loan purpose ‘Repair’ is having higher number of unsuccessful payments on time hence bank should avoid giving loan for this.
- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Pensioners & people with Higher income are most likely to make payments.
- People with housing type ‘With parents’ can be targeted as they are having least number of unsuccessful payments.

Thank You