EDA CASE STUDY

Bhagyashree P. Dhole
Pratibha S. Chougule

PROBLEM STATEMENT - I

Two types of risks are associated with the bank's decision:

- **Target0** = If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

- **Target1**= 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.

Data analysis is done in Python using Jupiter Notebook

THE DATA GIVEN CONTAINS THE INFORMATION ABOUT THE LOAN APPLICATION

It contains **two** types of scenarios:

- **The 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,

- **All other cases:** All other cases when the payment is paid on time.

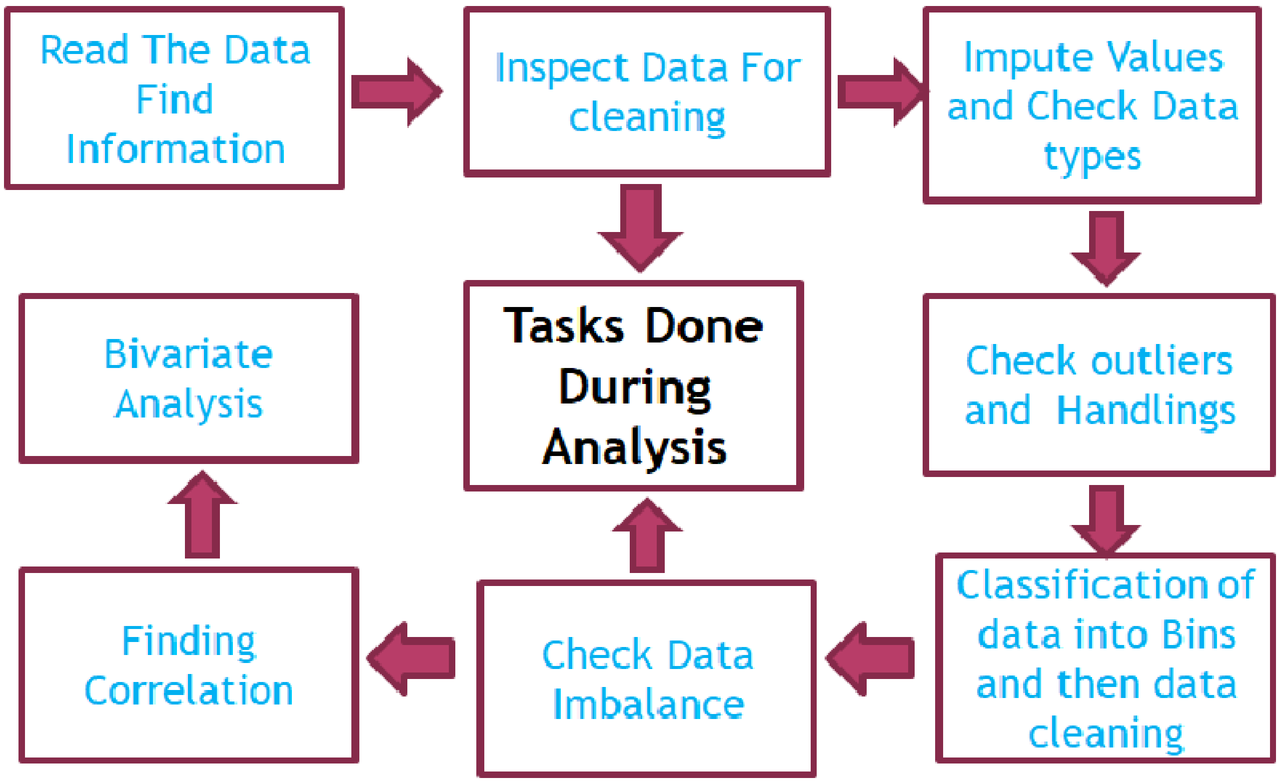
WHAT DECISIONS TAKEN BY THE COMPANY WHEN A CLIENT APPLIES FOR A LOAN ??

- **Approved**: The Company has approved loan Application
- **Cancelled**: The client cancelled the application sometime during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk
- **Refused**: The company had rejected the loan
- **Unused offer**: Loan has been cancelled by the client but on different stages of the process.

WHAT ANALYSIS DONE IN CASE STUDY?

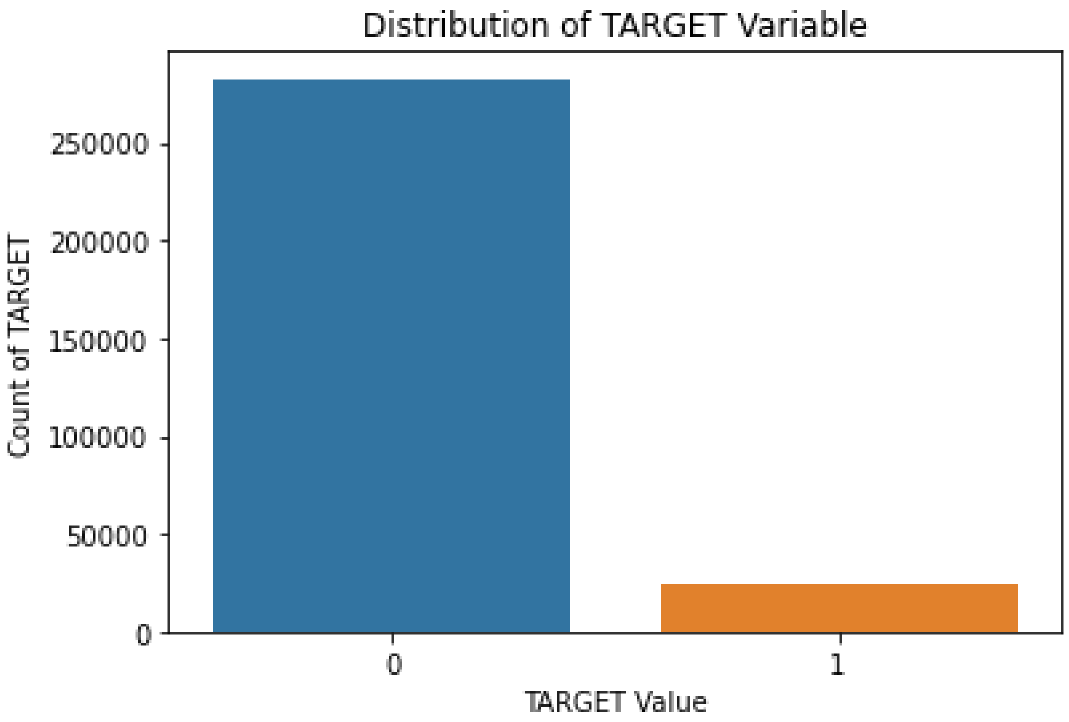
- Check Missing Values
- Check Outliers
- Top 10 Correlation for clients with payment difficulties
- Finding Most relevant correlation.

TASKS IN ANALYSIS



CHECKING DISTRIBUTION OF TARGET VARIABLE

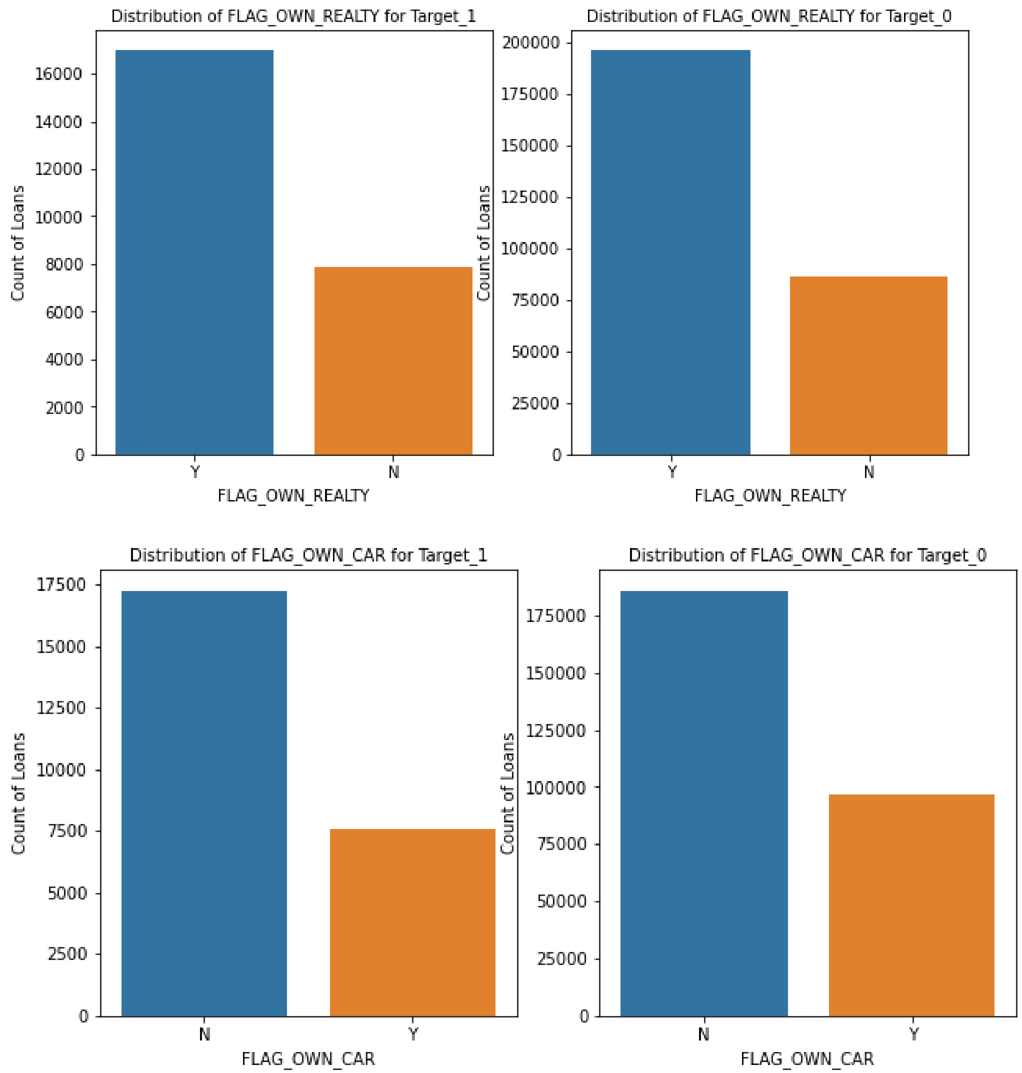
- Defaulters Percentage in Total



PROPORTION OF DEFAULTERS BY GENDER

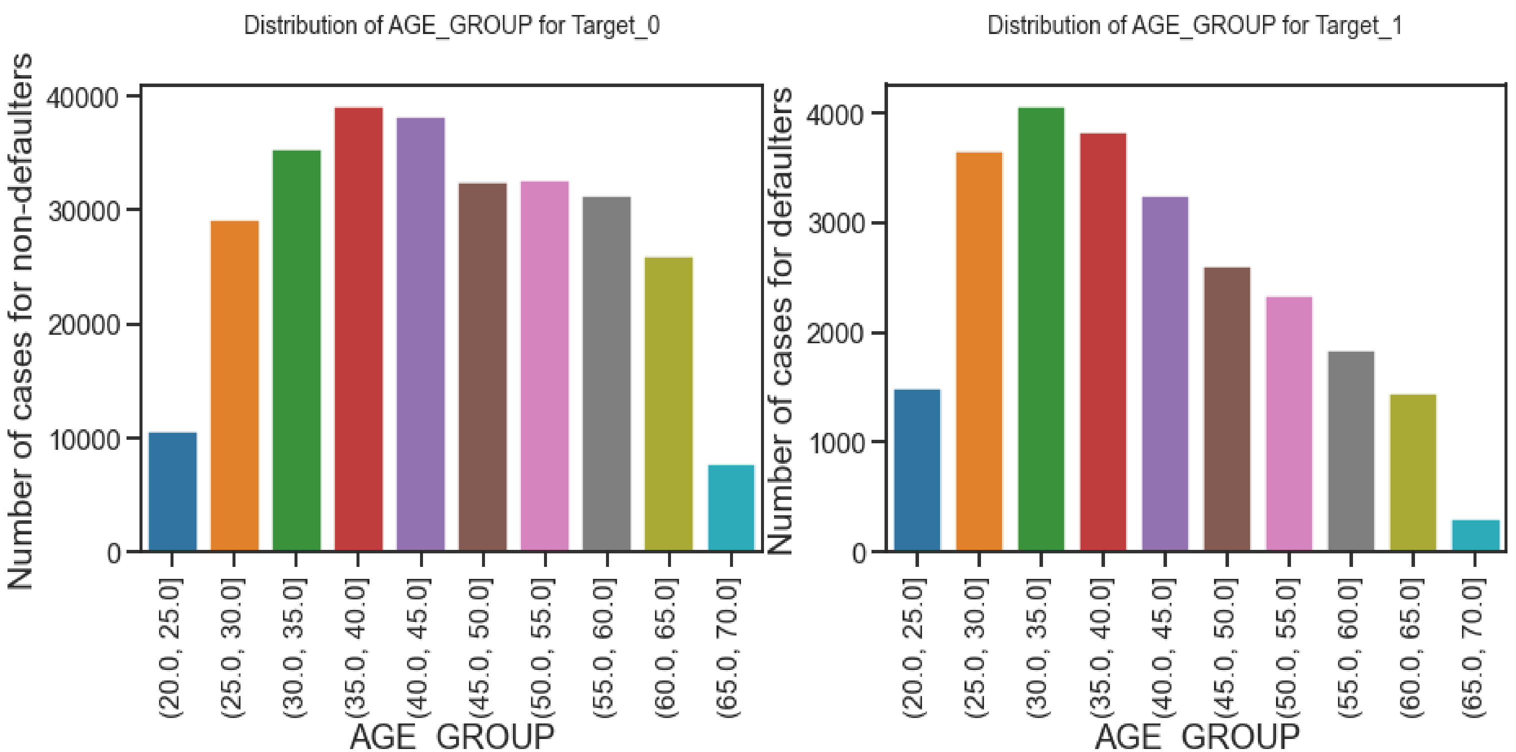
- We observe that the number of Females taking loans is much higher than the number of Males for both Target = 0 and Target = 1

PROPORTION OF DEFAULTERS BY GENDER FOR OWN CAR

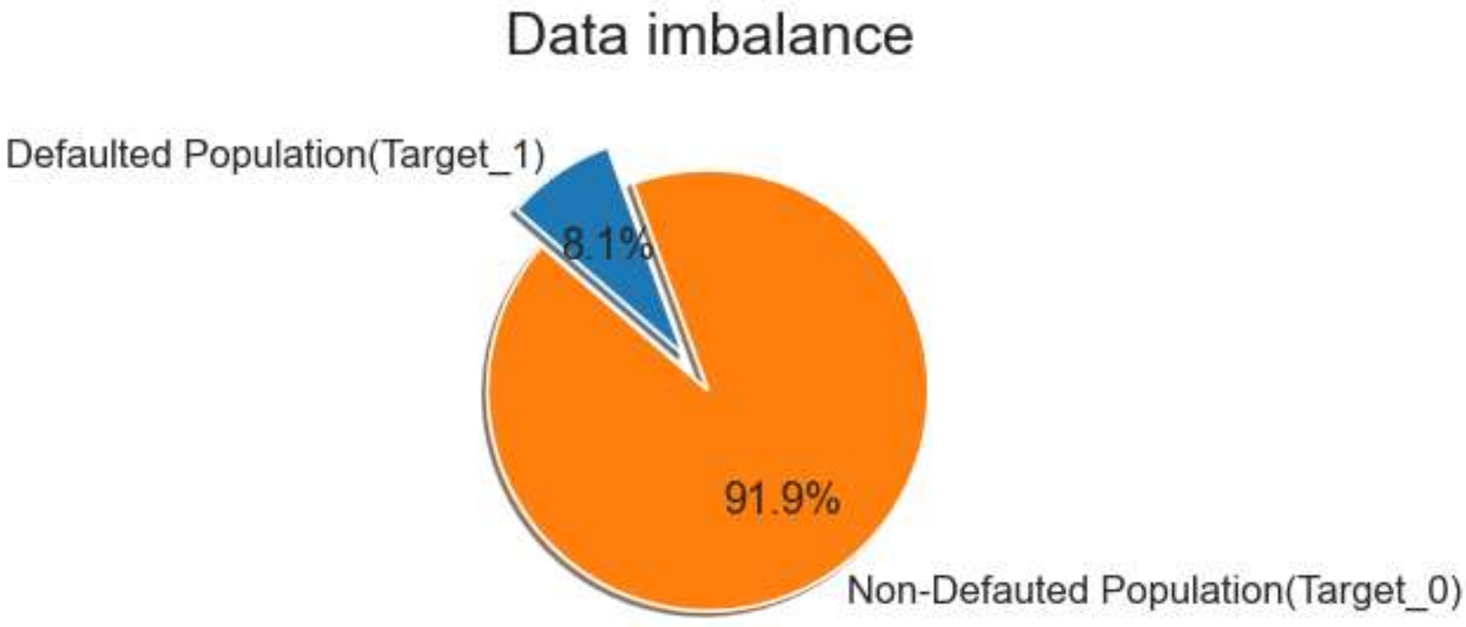


- most people applying for loan do not own a car, Own car Defaults is less.

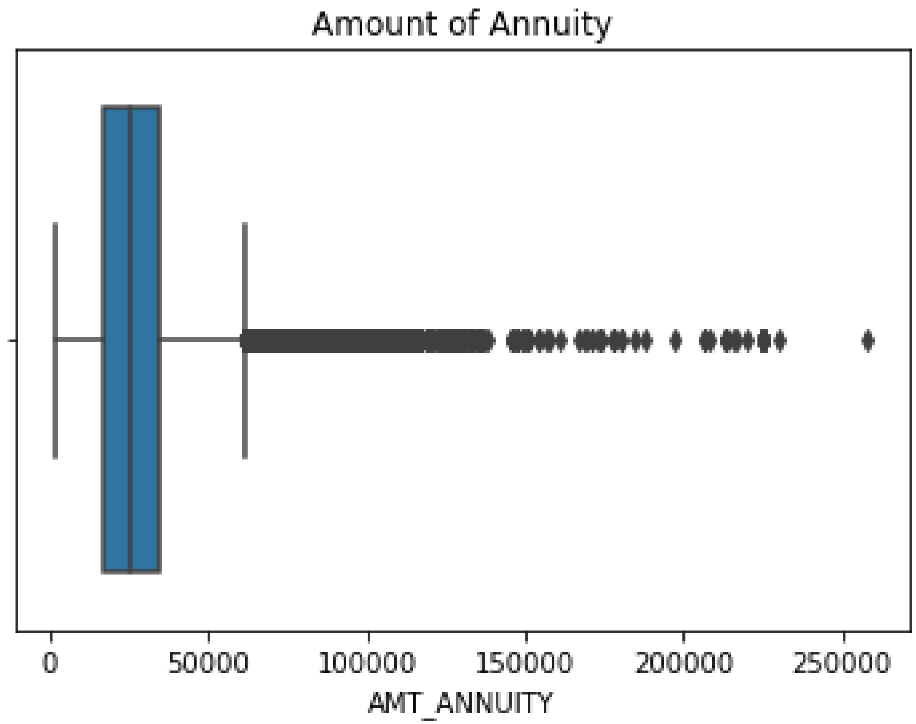
DISTRIBUTION OF AGE GROUP



DATA IMBALANCE

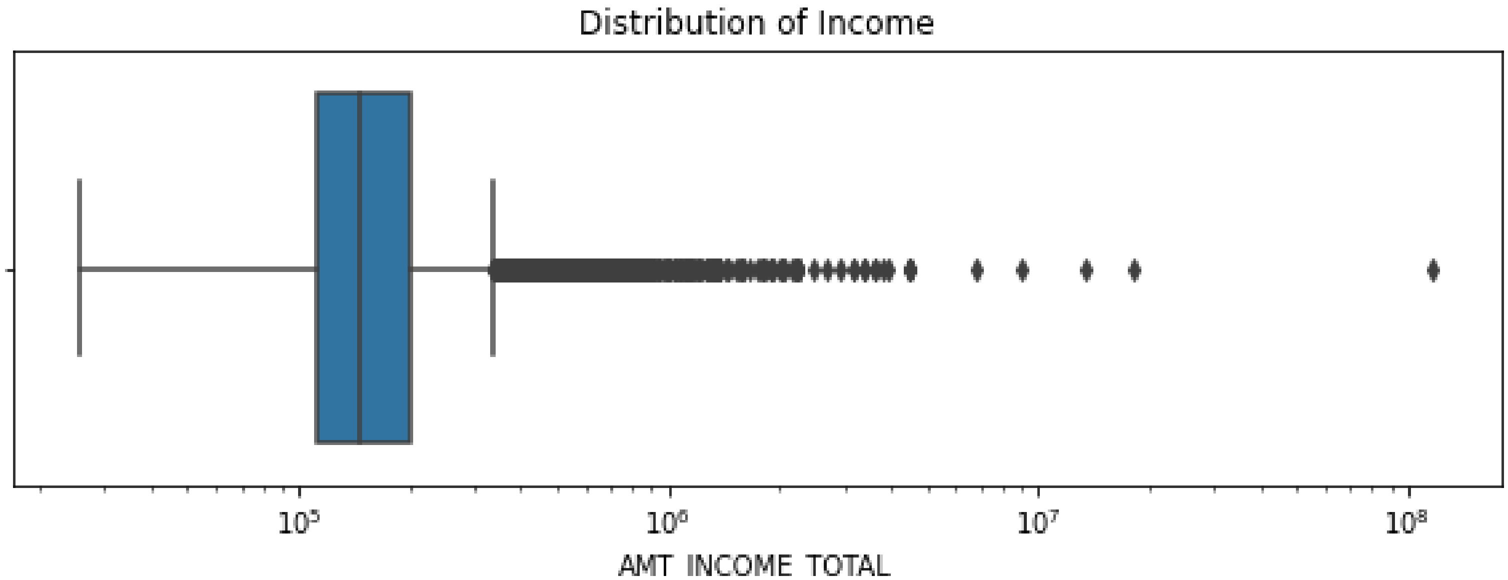


AMOUNT OF ANNUITY

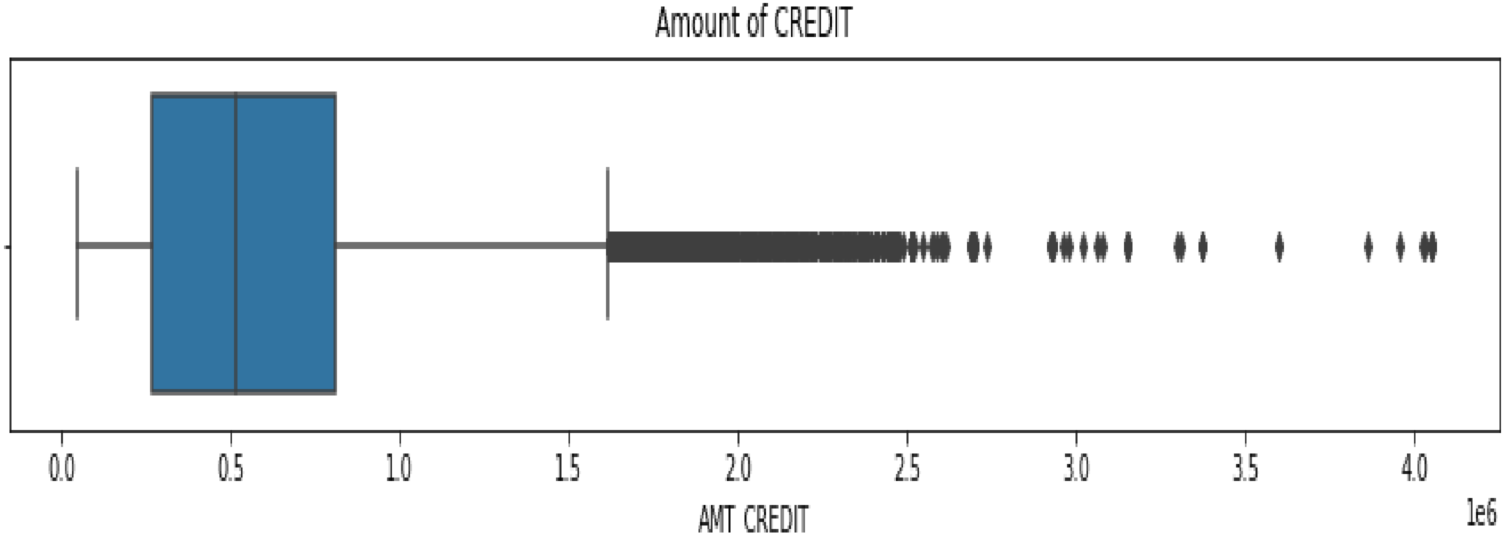


DISTRIBUTION OF INCOME

- In 'AMT_INCOME_TOTAL' column, We can see that there are outlier values at 1.17x10^8. Sometimes, it is beneficial to look into the quantiles instead of the box plot, mean or median.
  2. Quantile may give you a fair idea about the outliers. If there is a huge difference between the maximum value and the 95th or 99th quantiles, then there are outliers in the data set.

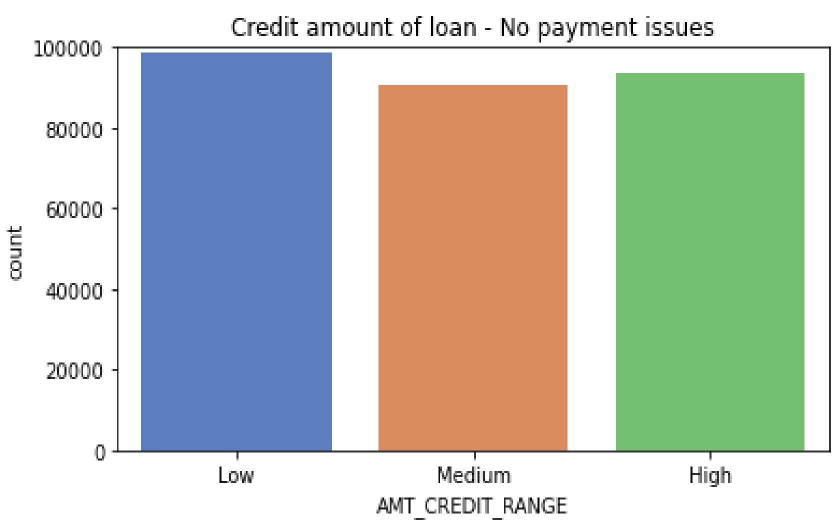


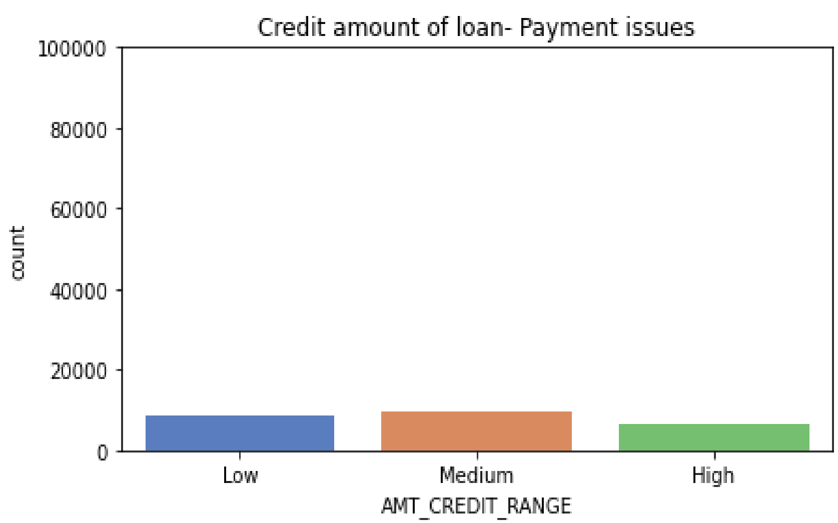
AMOUNT OF CREDIT



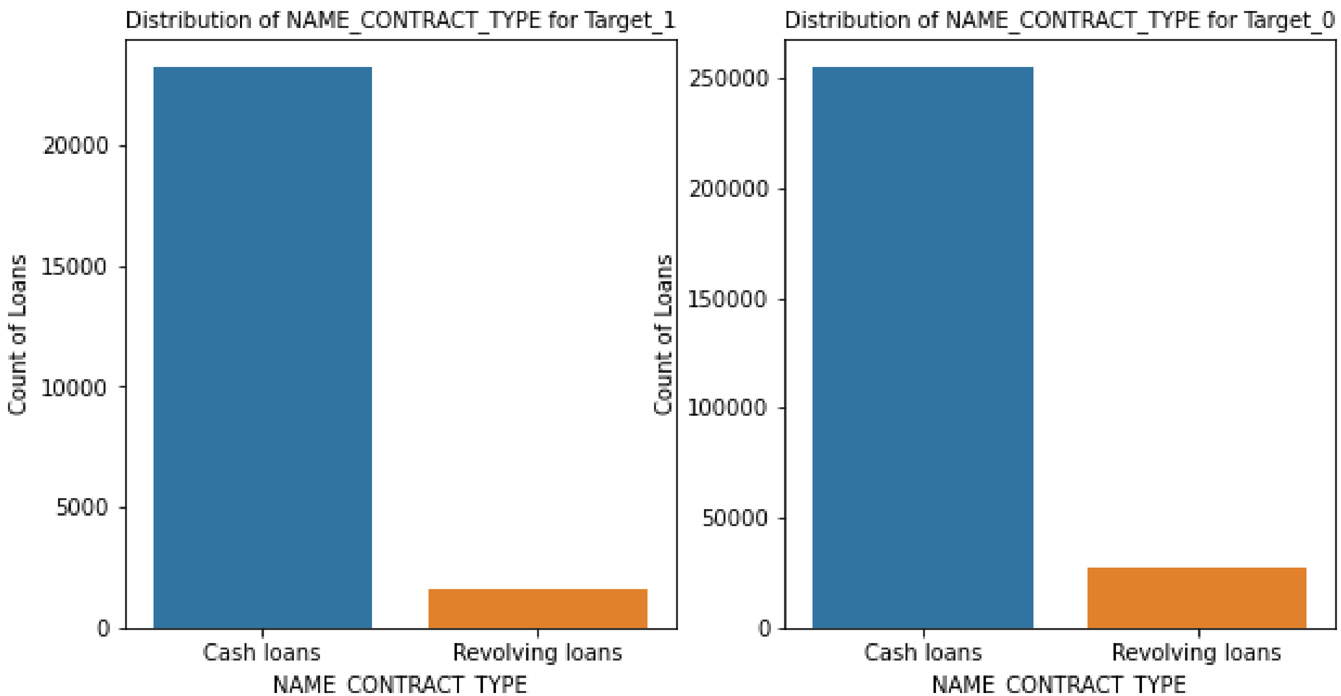
CREDIT AMOUNT PAYMENT AND ISSUES

- Customers with less credit and most likely to make payment.
- Customers having Medium and High credit can also be considered while lending the loan



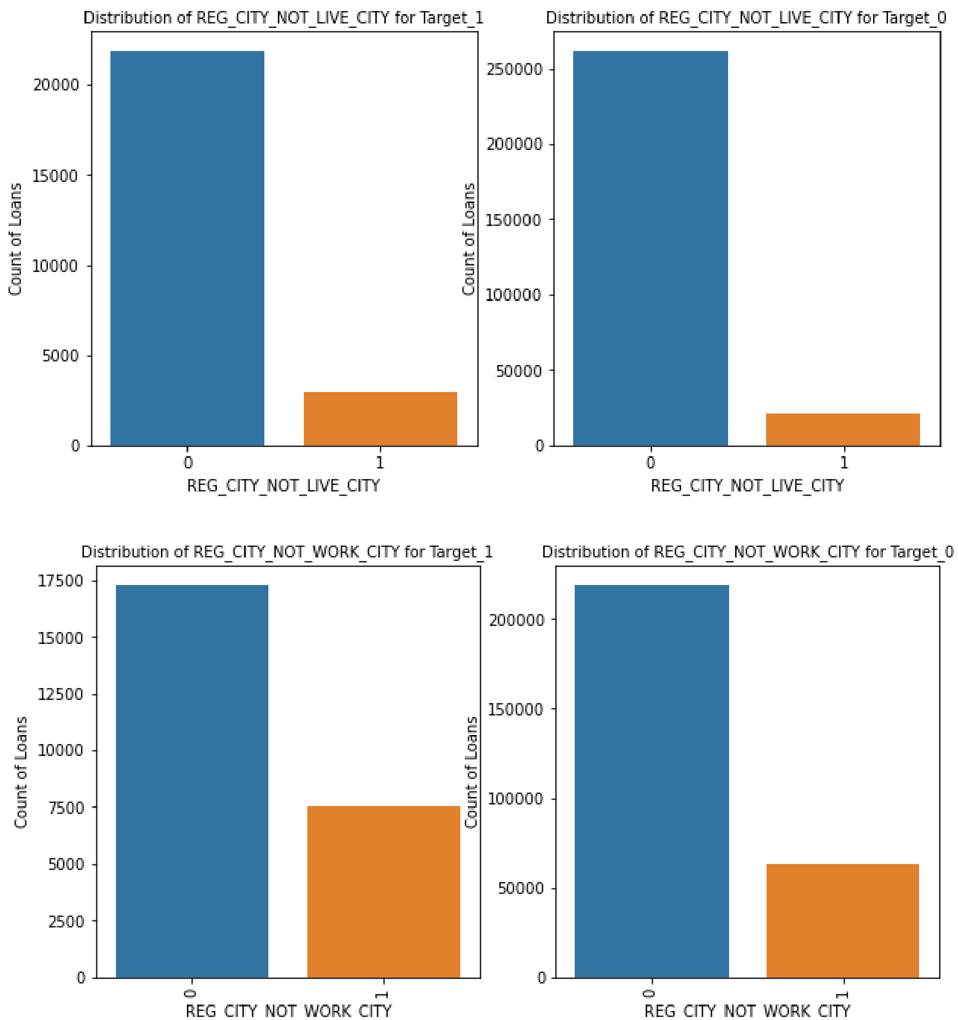


PROPORTION OF CASH LOANS AND REVOLVING LOANS



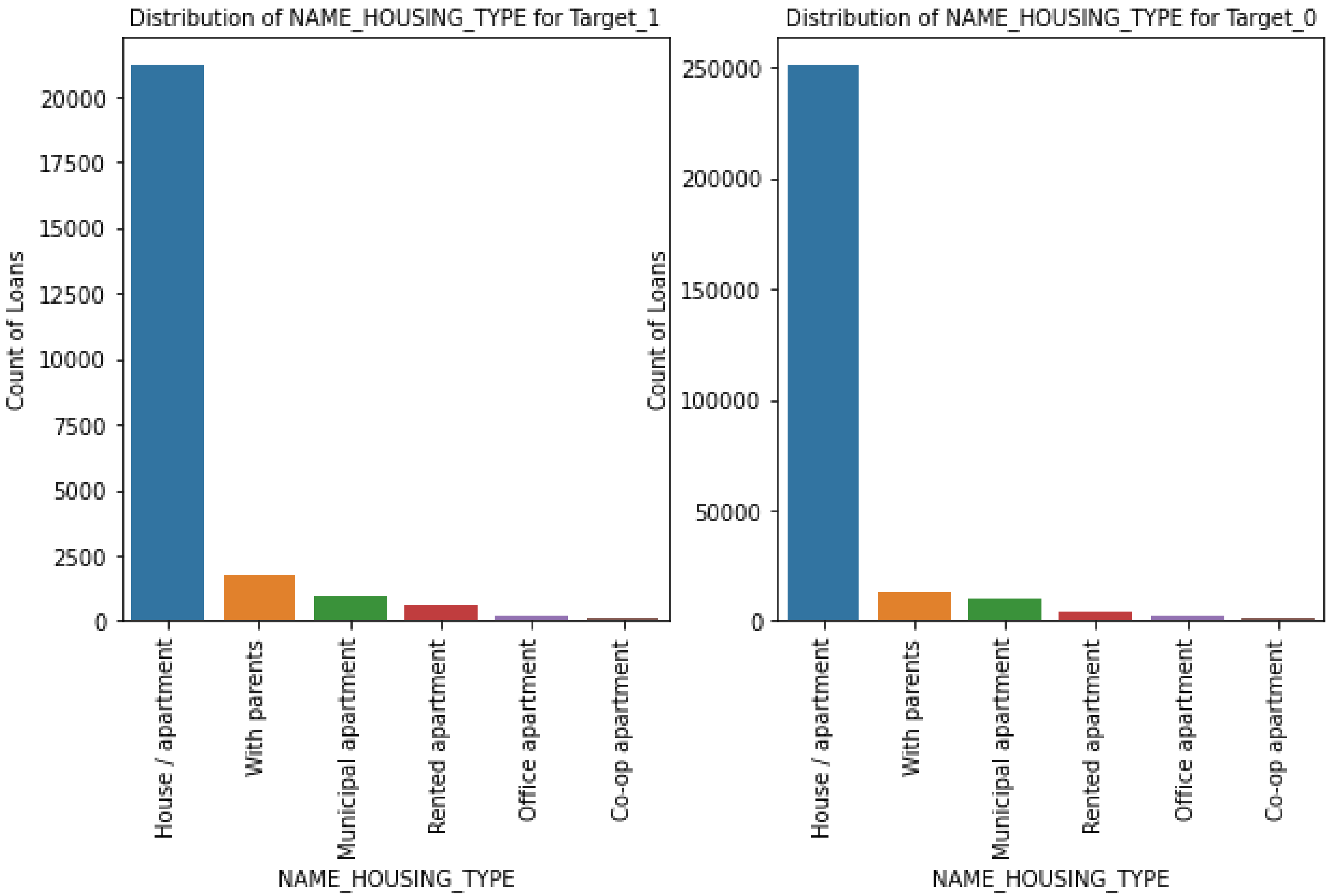
- From this information, we see this is an imbalanced dataset. There are far more loans that were repaid on time than loans that were not repaid. More than 25000 loans were repaid, Less than 5000 loans were not repaid. We observe that the number of Cash loans is much higher than the number of Revolving for both Target = 0 and Target = 1loans

DISTRIBUTION OF REG CITY AND LIVE CITY

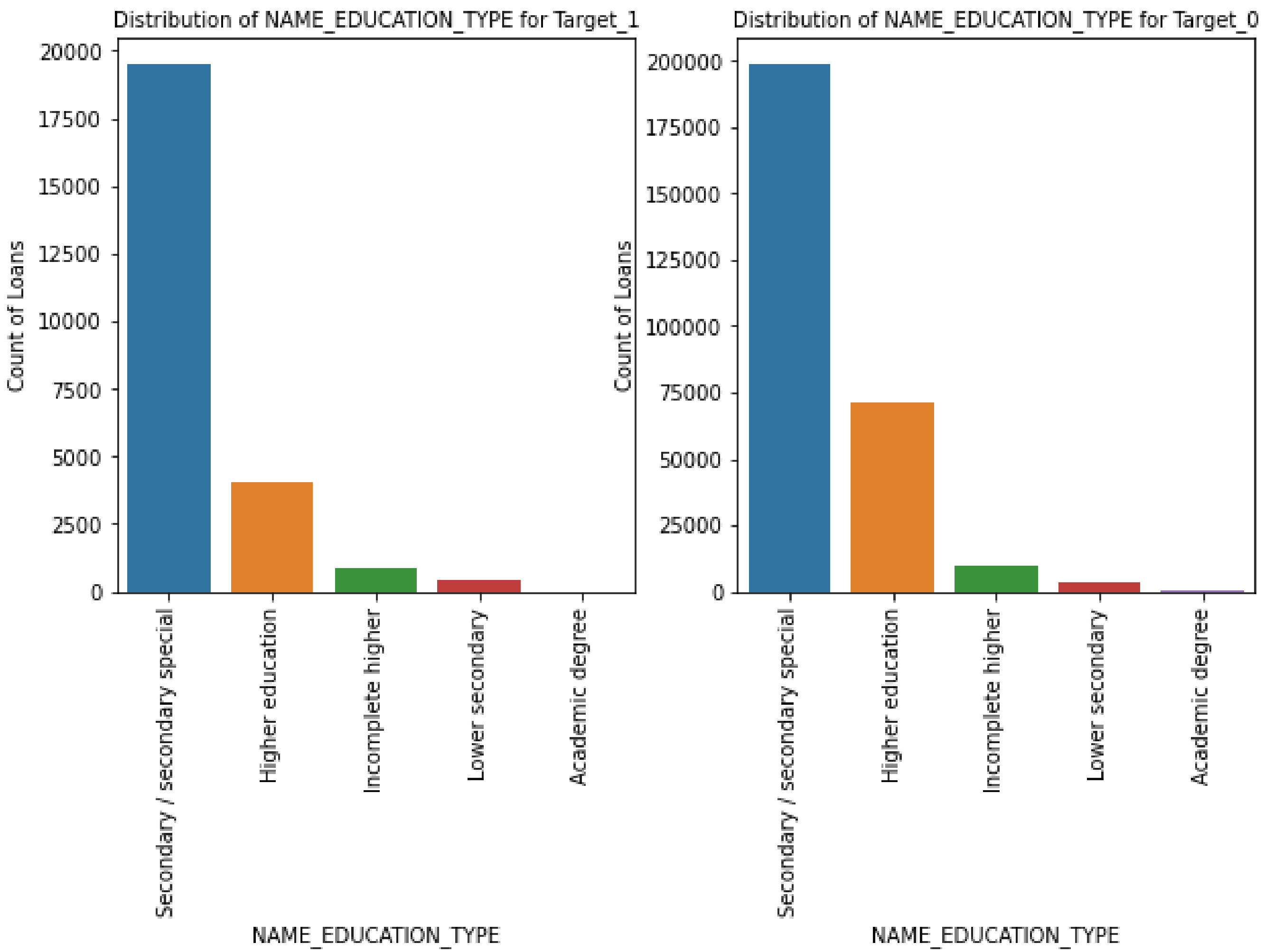


- We observe that the Ratio of people whose Registration City is not the same as live city or work city is higher in case of defaulters are compared to defaulters.

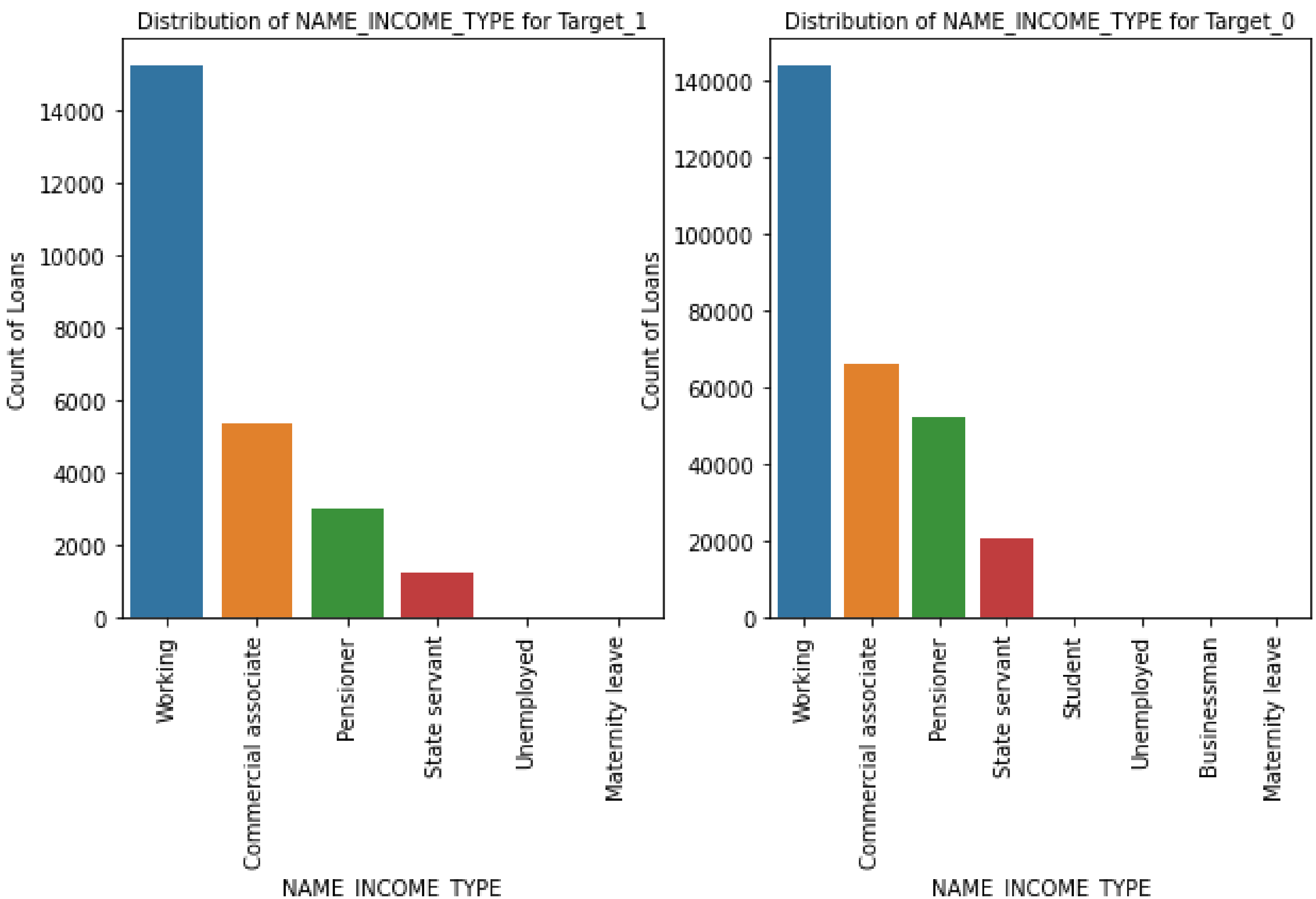
PROPORTION OF DAFAULTERS BY HOUSING TYPE



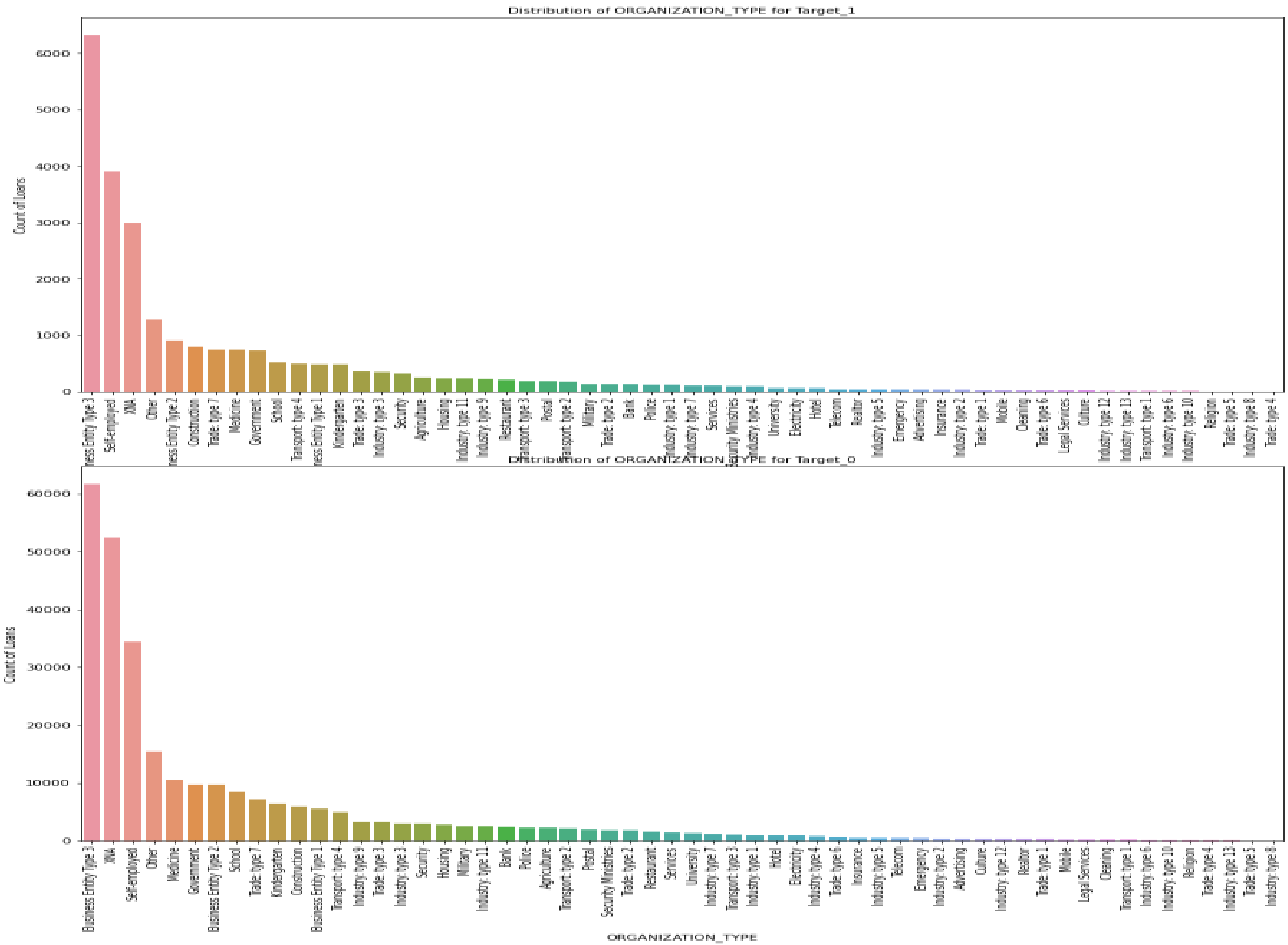
PROPORTION OF DEFAULTERS BY EDUCATION



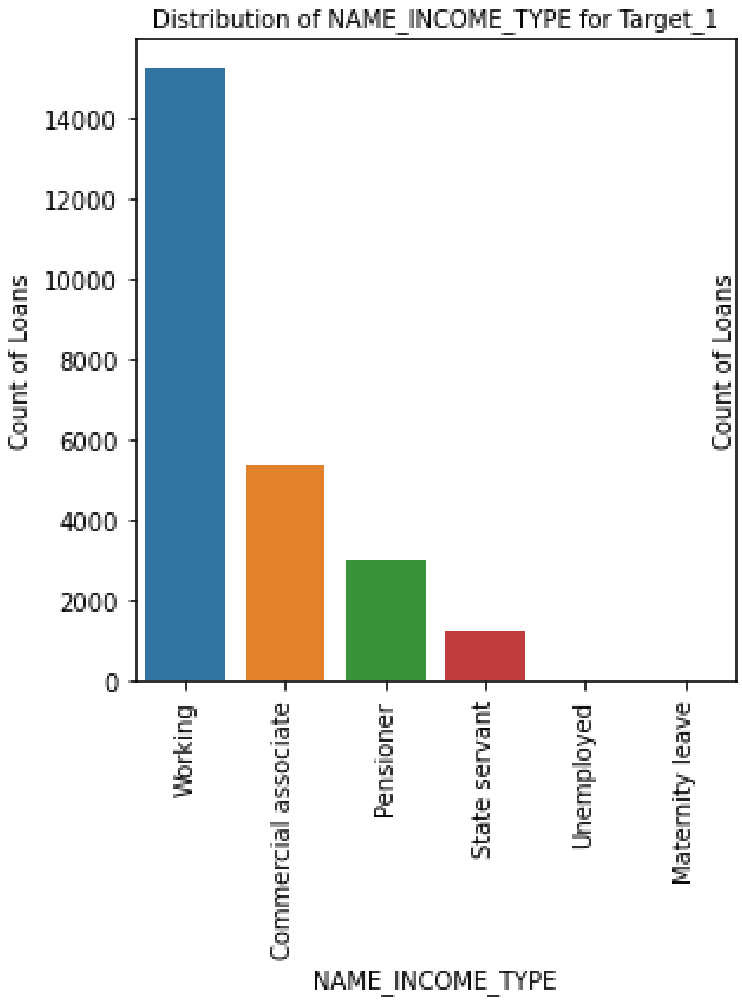
PROPORTION OF DEFAULTERS BY INCOME

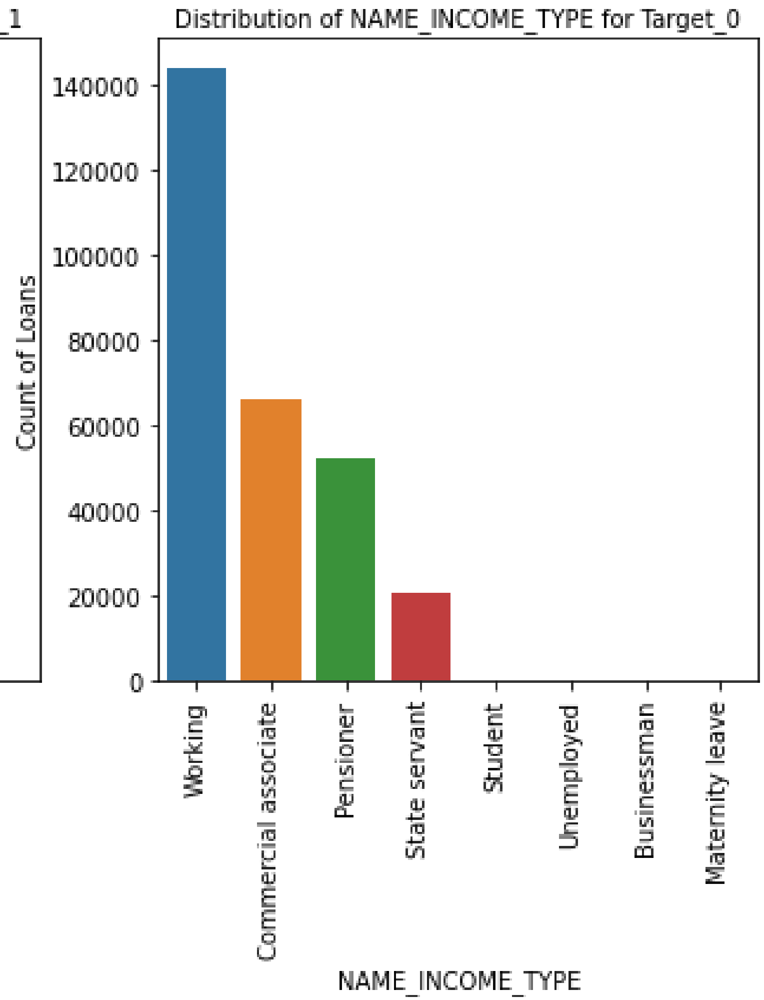


PROPORTION OF DEFAULTERS BY ORGANIZATION_TYPE

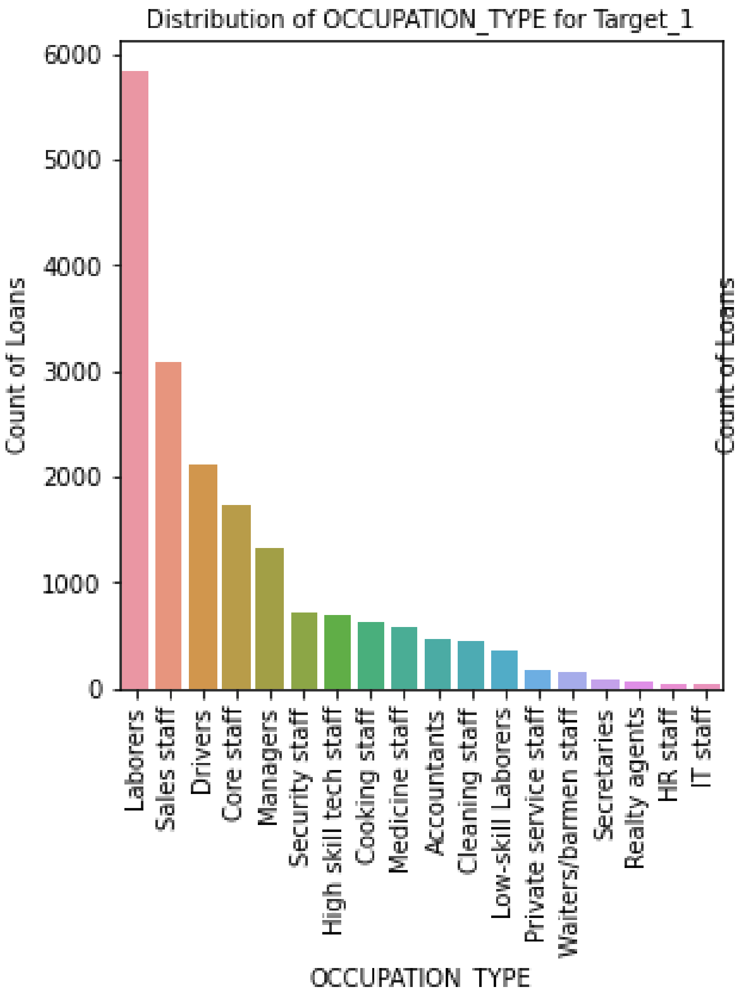


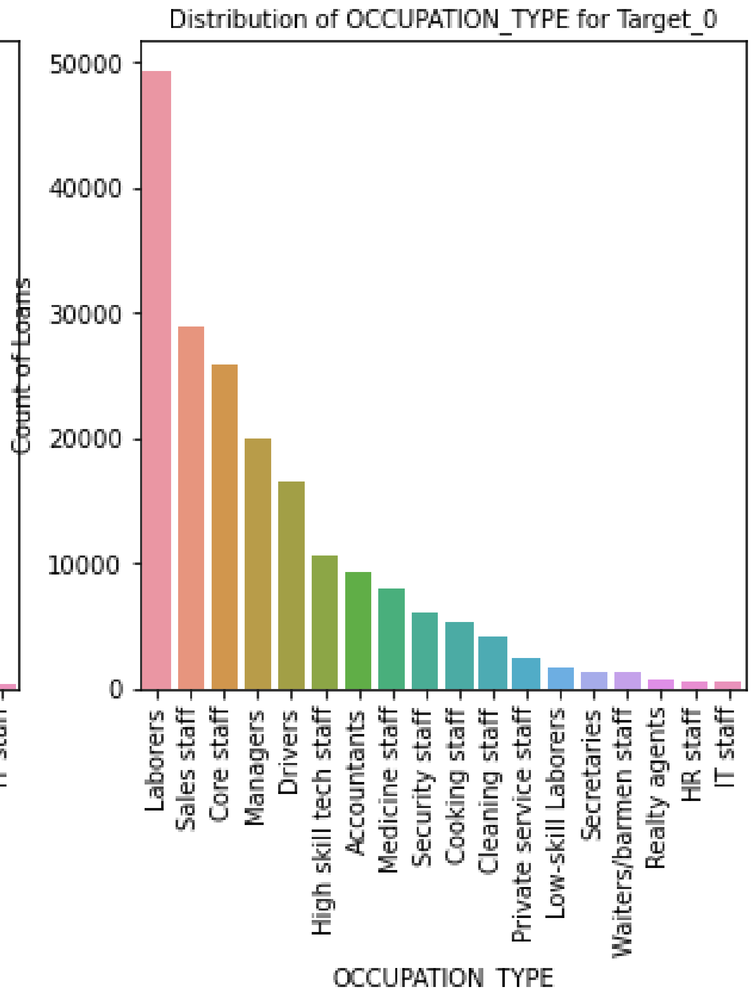
PROPORTION OF DEFAULTERS BY INCOME_TYPE



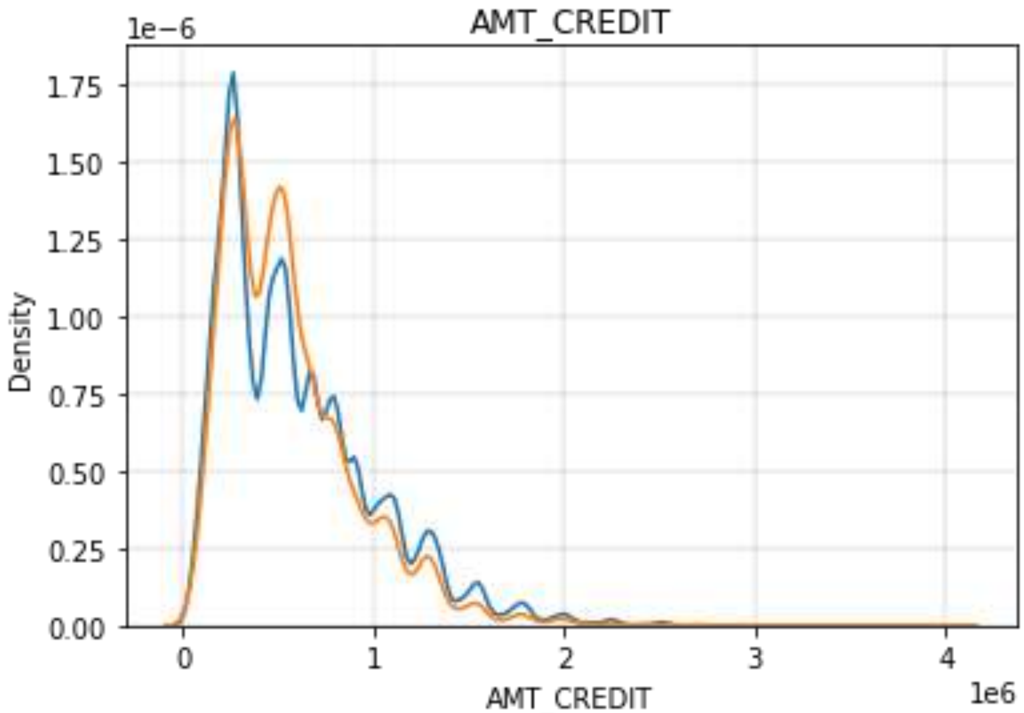


PROPORTION OF DEFAULTERS BY OCCUPATION_TYPE

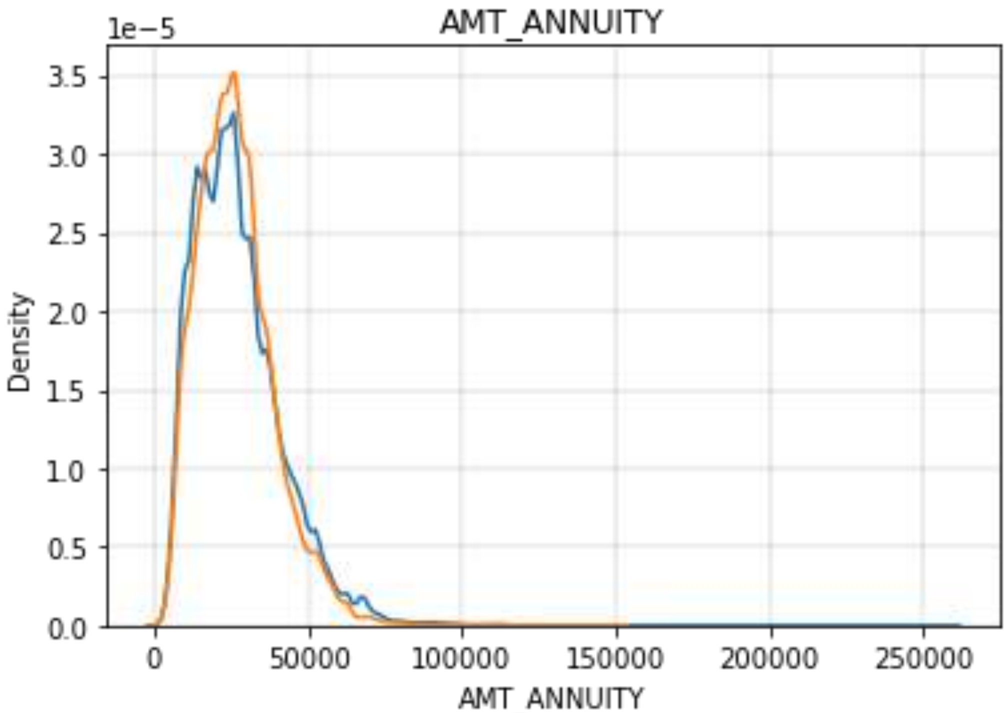




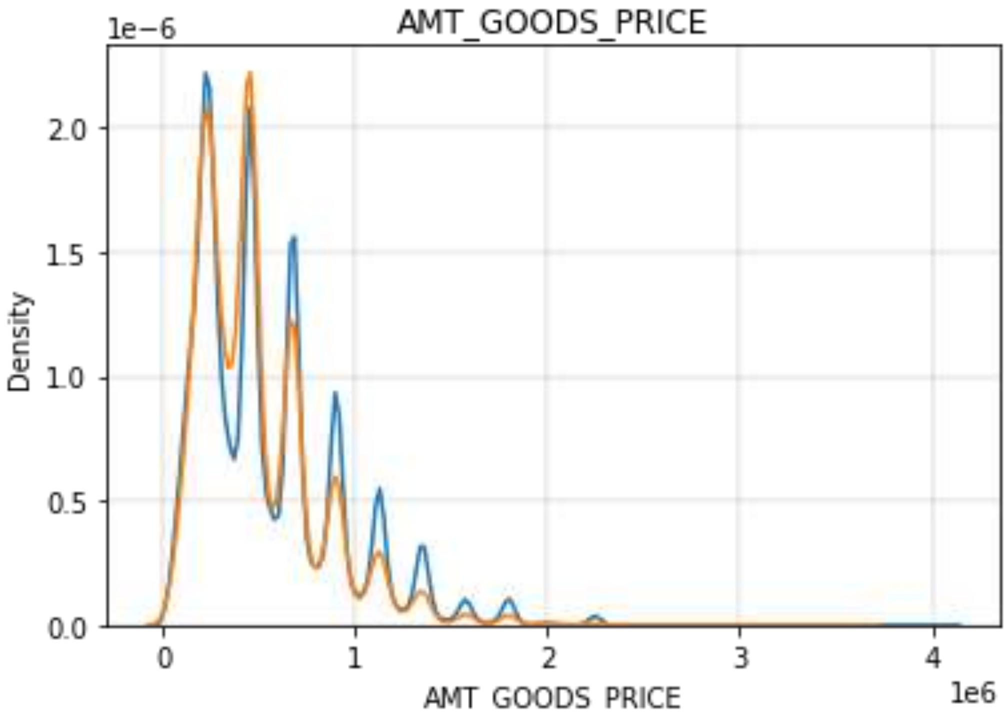
DEFAULTERS BY AMOUNT_CREDIT



DEFAULTERS BY AMOUNT_ANNUITY



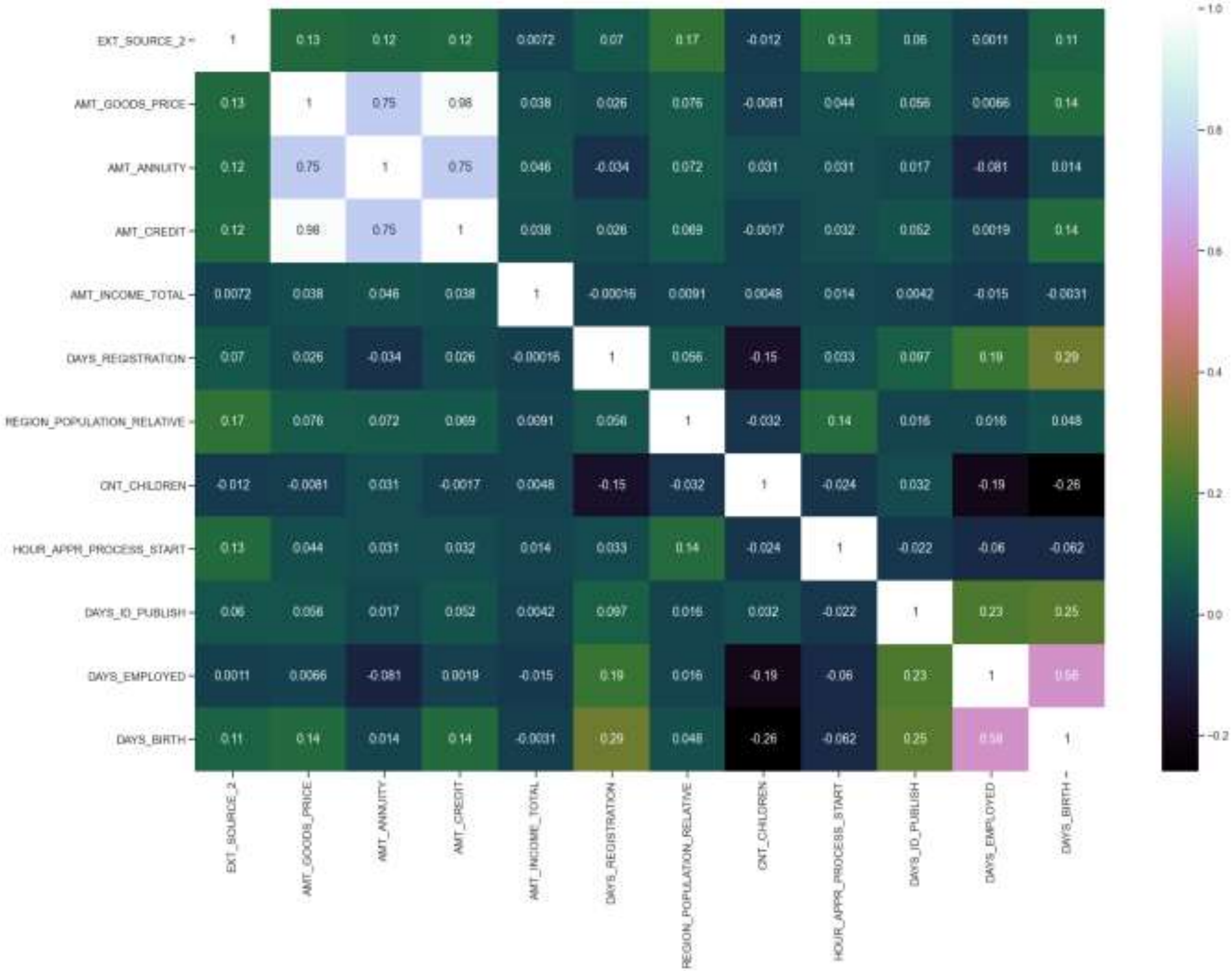
DEFAULTERS BY AMOUNT_GOODS PRIZE



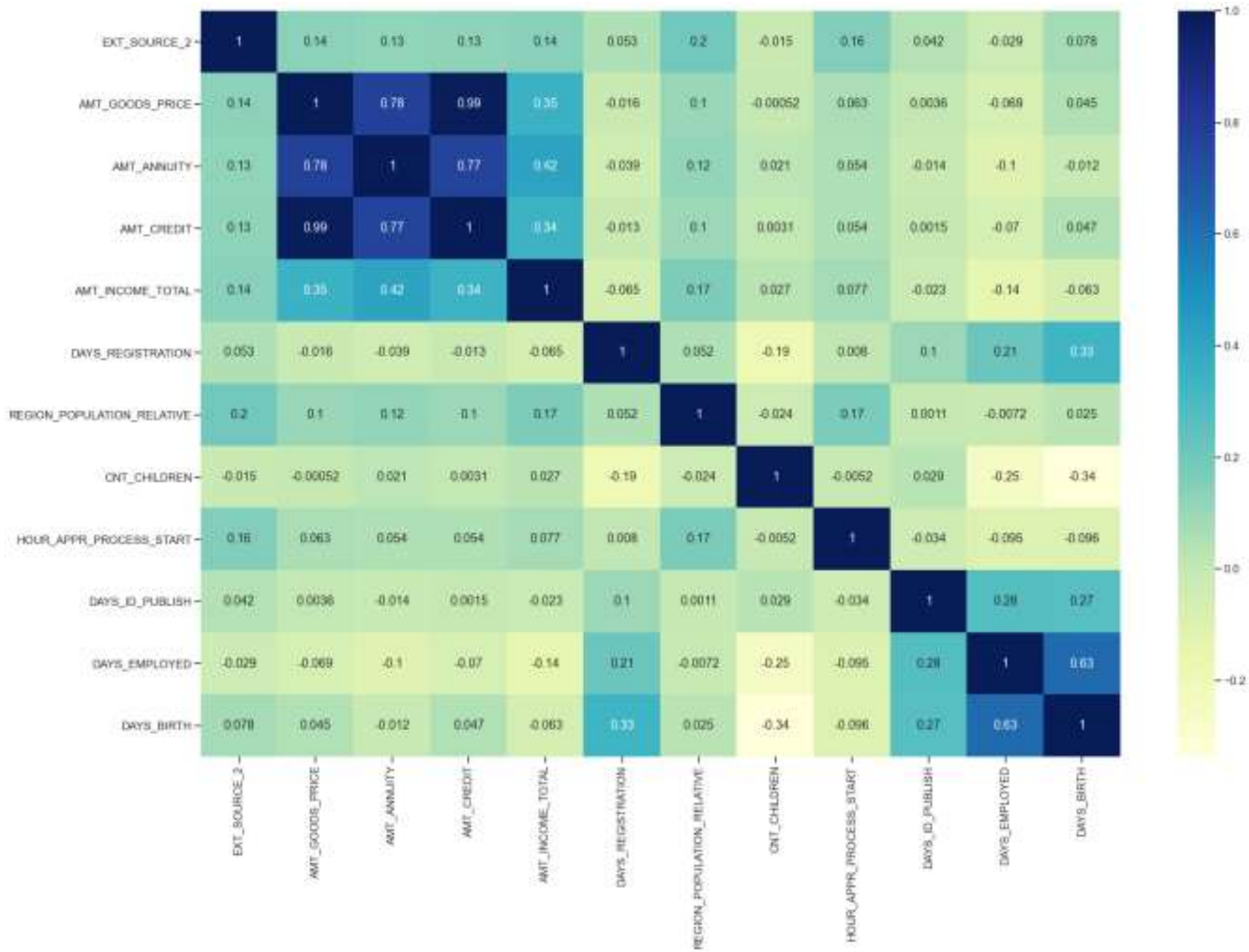
FROM ALL DEFAULTERS PLOT WE CAN CONCLUDE THAT

- **Female** customers pay loan amount **on time** and banks **can target** more female customers for **lending loan**.
- **Working customers** can be targeted to lend loans as they have **higher percentage** of making payments **on time.**
- Customers with **secondary education are most likely to make payments** when compared to customers with academic degree.
- **Married** customers have paid loan amount **on time** when compared to widows.
- Customers **owning House/apartment** are **most likely** to make payments on time compared to those living in CO-OP apartment.
- **Labors** have **high repayment** percentage. Hence bakes can think of lending small amount loans to them.

CORRELATION

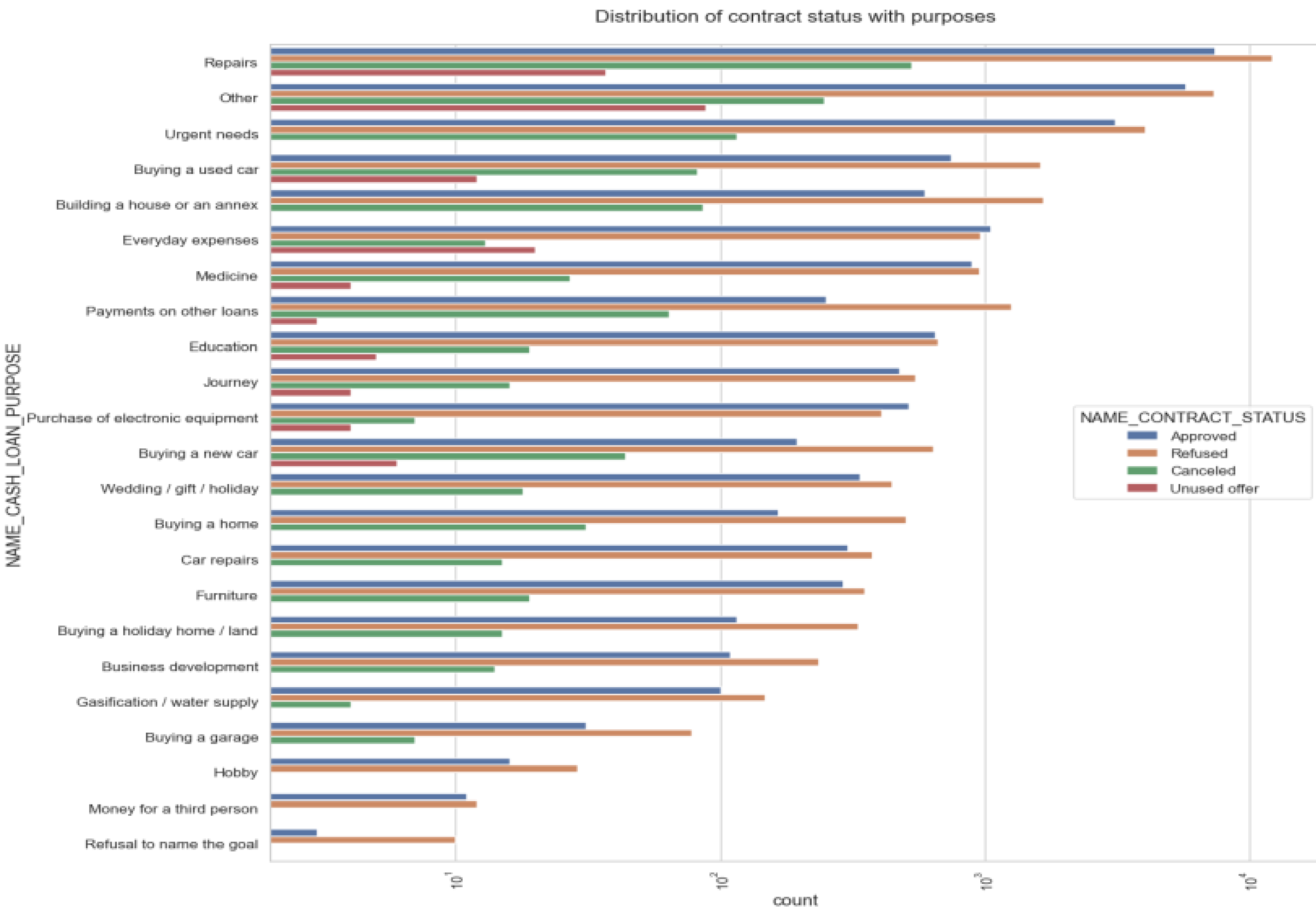


CORRELATION MATRIX

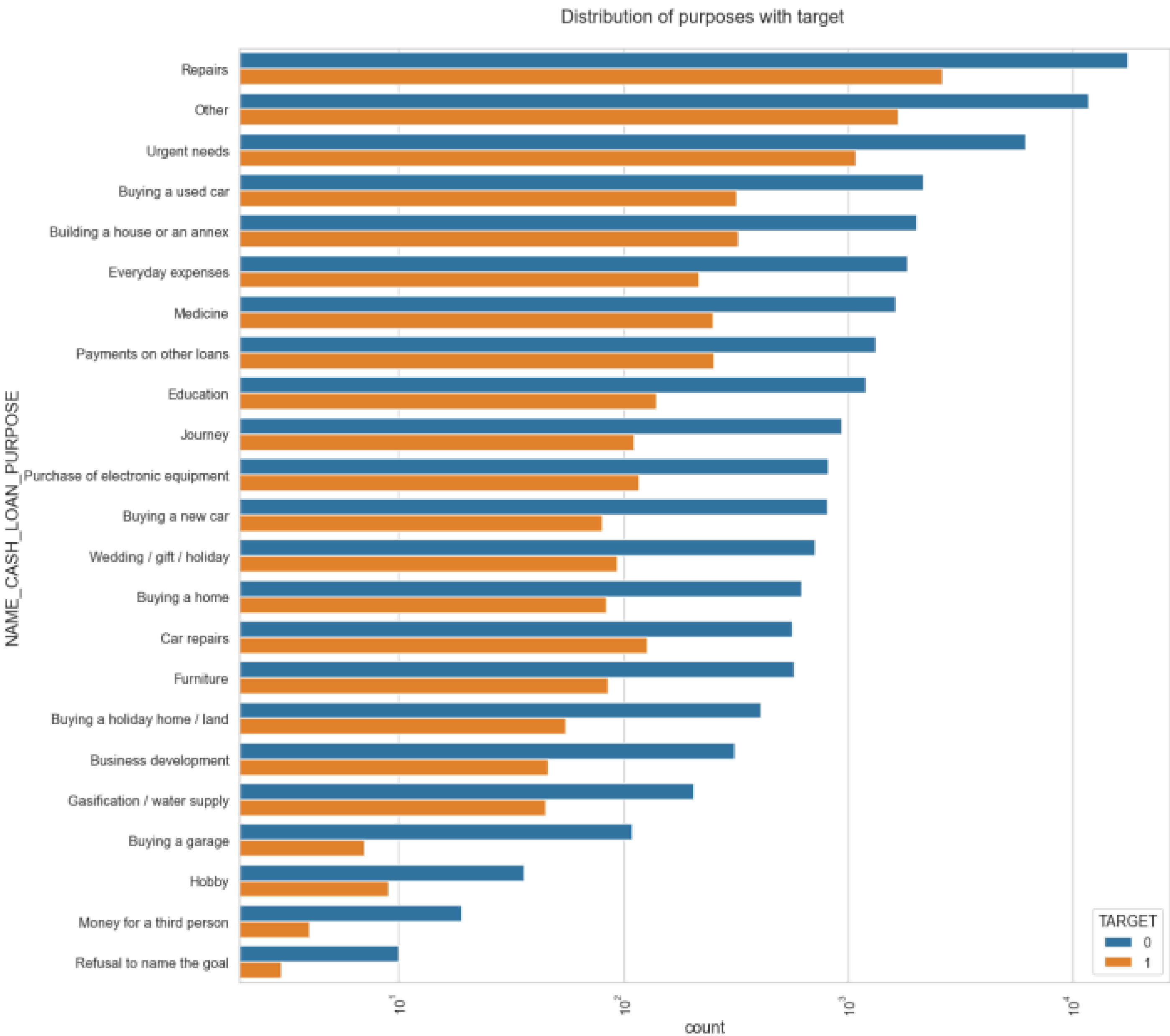


UNIVARIATE ANALYSIS

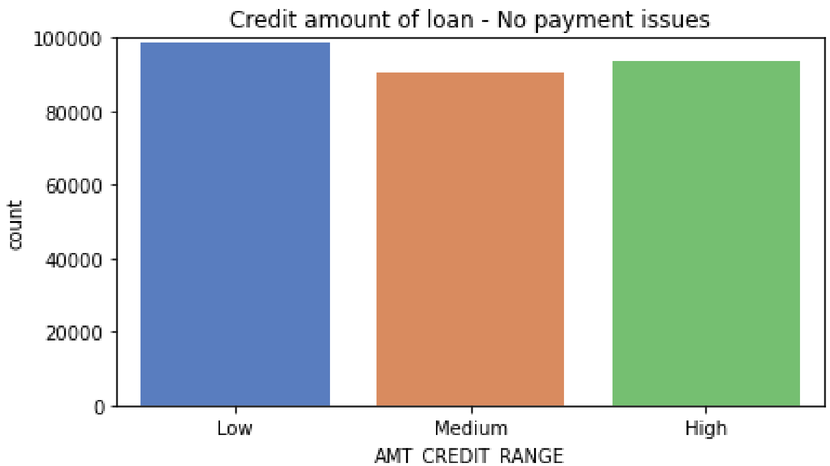
DISTRIBUTION OF CONTRACT STATUS WITH PURPOSE

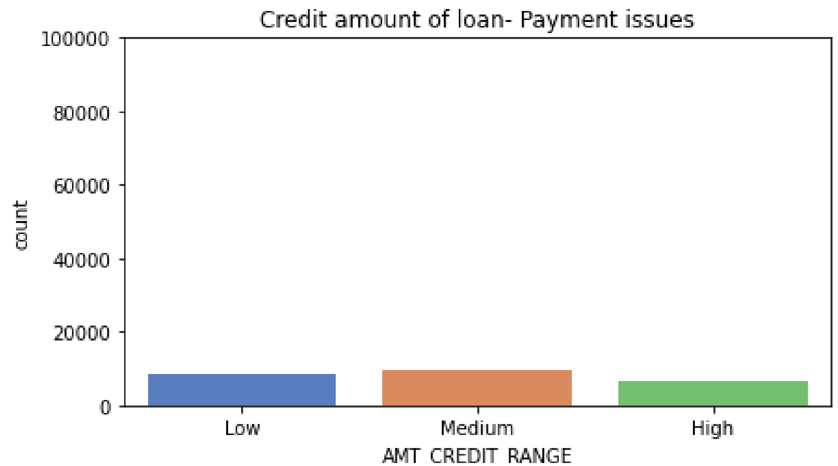


DISTRIBUTION OF PURPOSES WITH TARGET



UNIVARIATE ANALYSIS FOR TARGET 0 AND TARGET 1

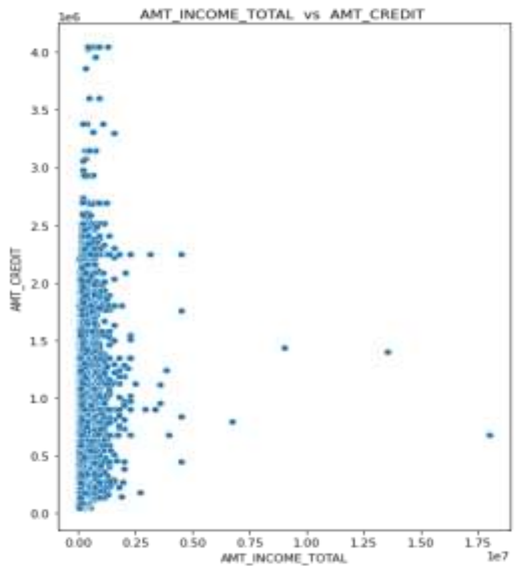


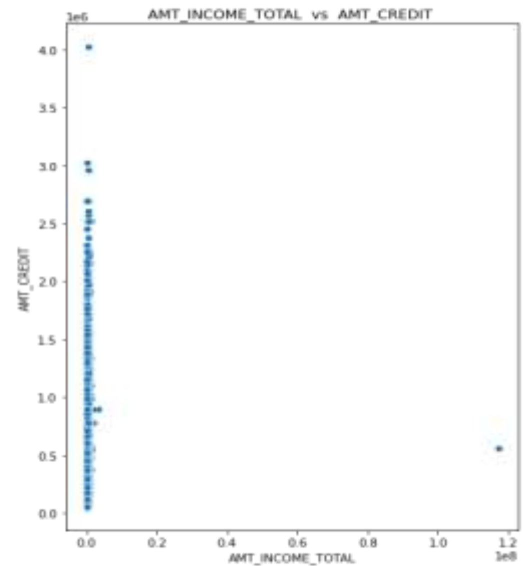


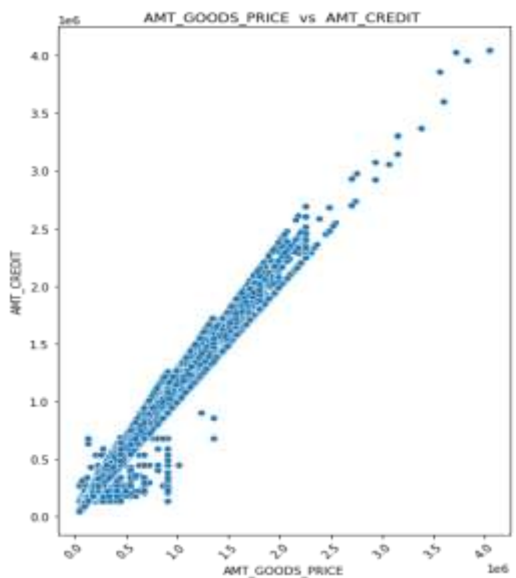
- Customers with less credit and most likely to make payment.
- Customers having medium and high credit can also be considered while lending the loan

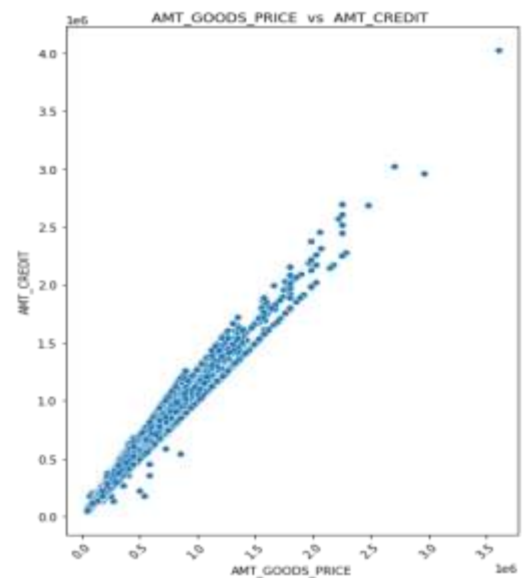
BIVARIATE ANALYSIS FOR TARGET 0 AND TARGET 1

BIVARIATE ANALYSIS FOR TARGET 0 AND TARGET 1









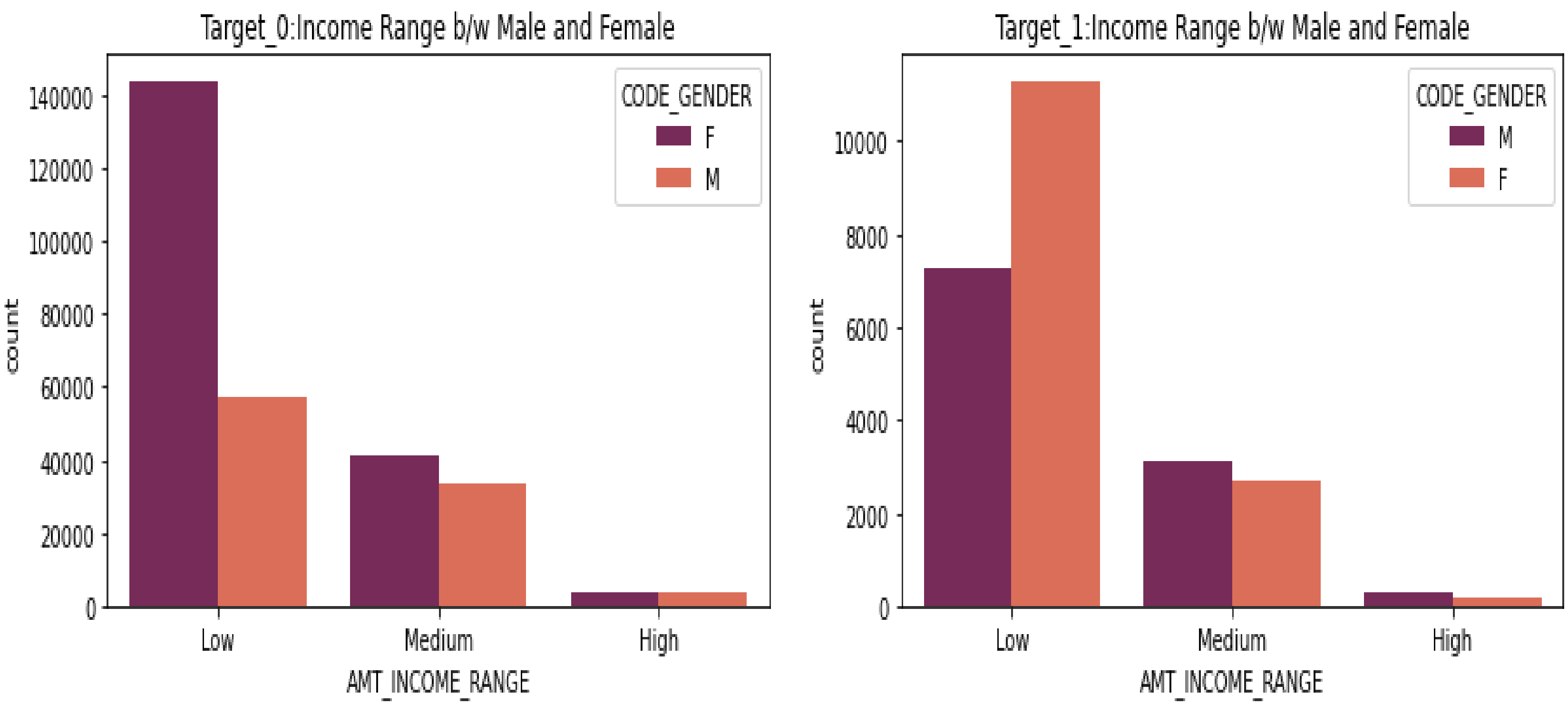
BIVARIATE ANALYSIS:

Those who have paid the **loan amount on/within time** are more likely to get **higher credits** than those who didnt pay/did late payments.

People who have higher goods price and have made payments on time have **higher credits** than those with higher goods price but didnt pay loan.

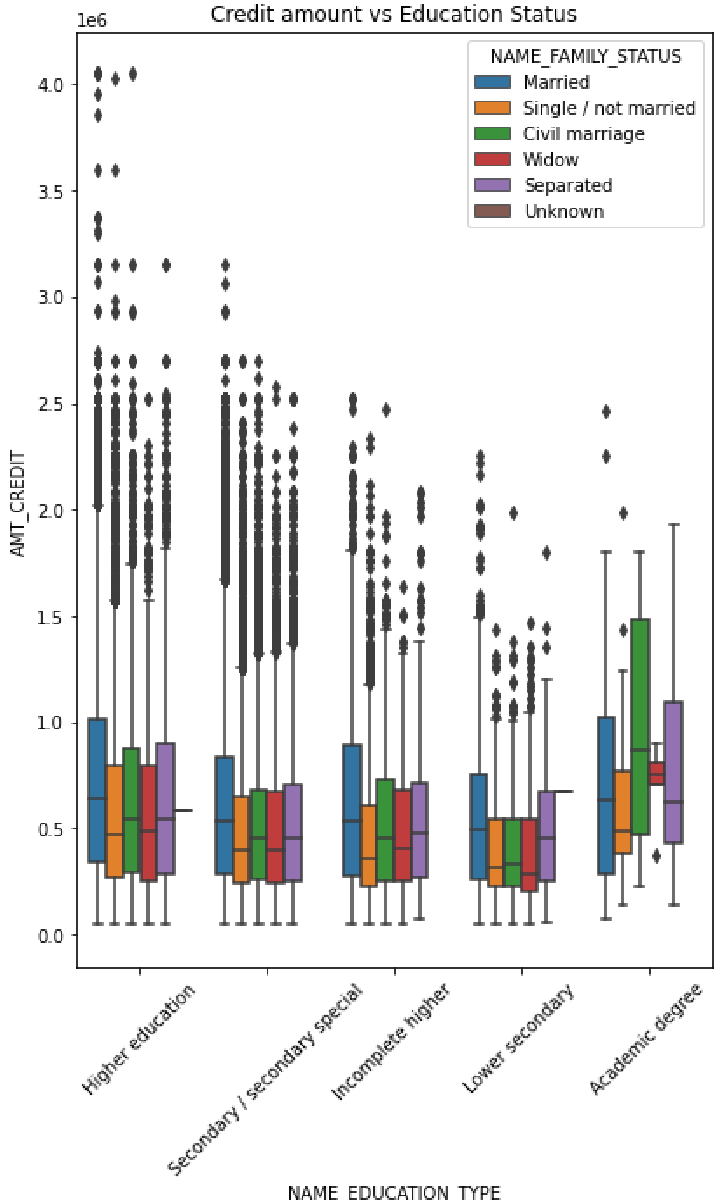
NUMERICAL CATEGORICAL ANALYSIS

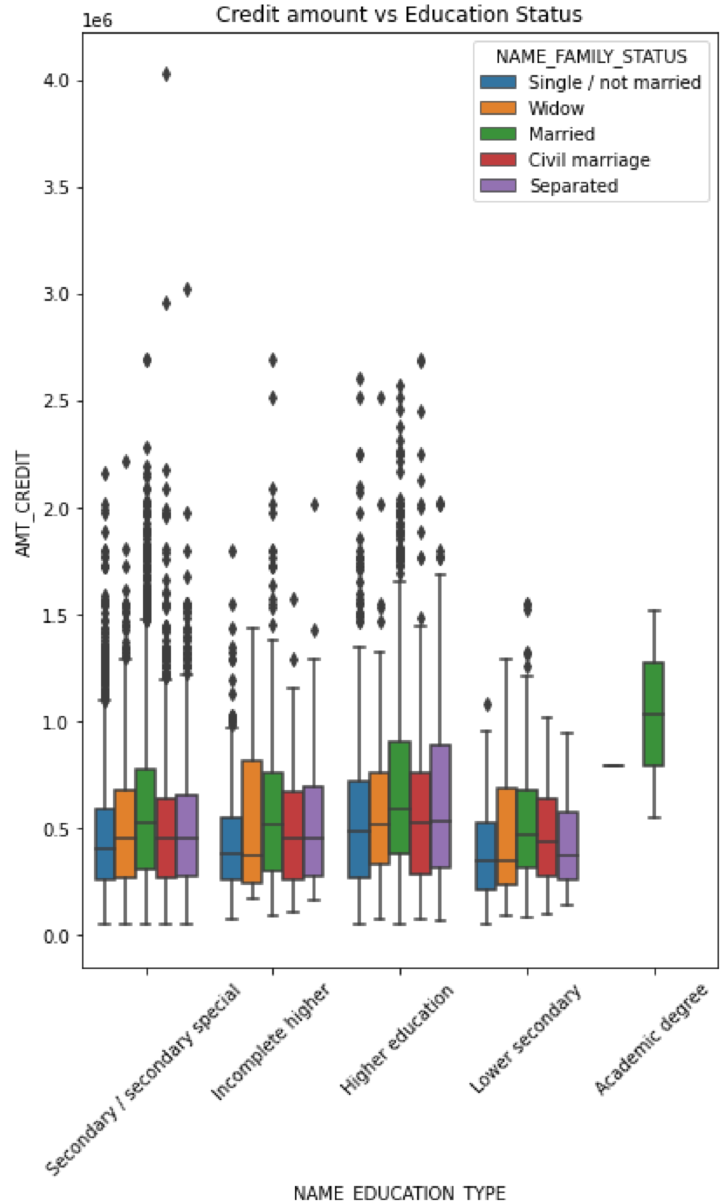
Income range- Gender



We can see that Females with low income don't have any payment issues.

CREDIT AMOUNT VS EDUCATION STATUS

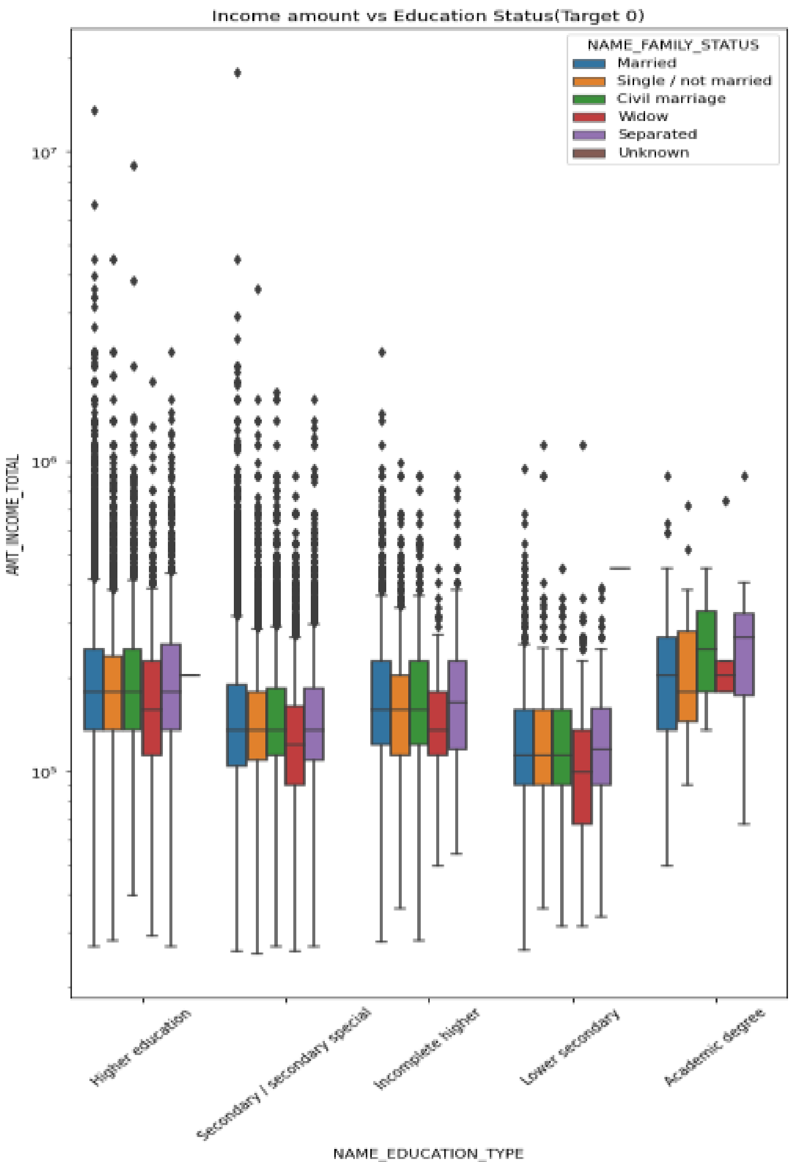


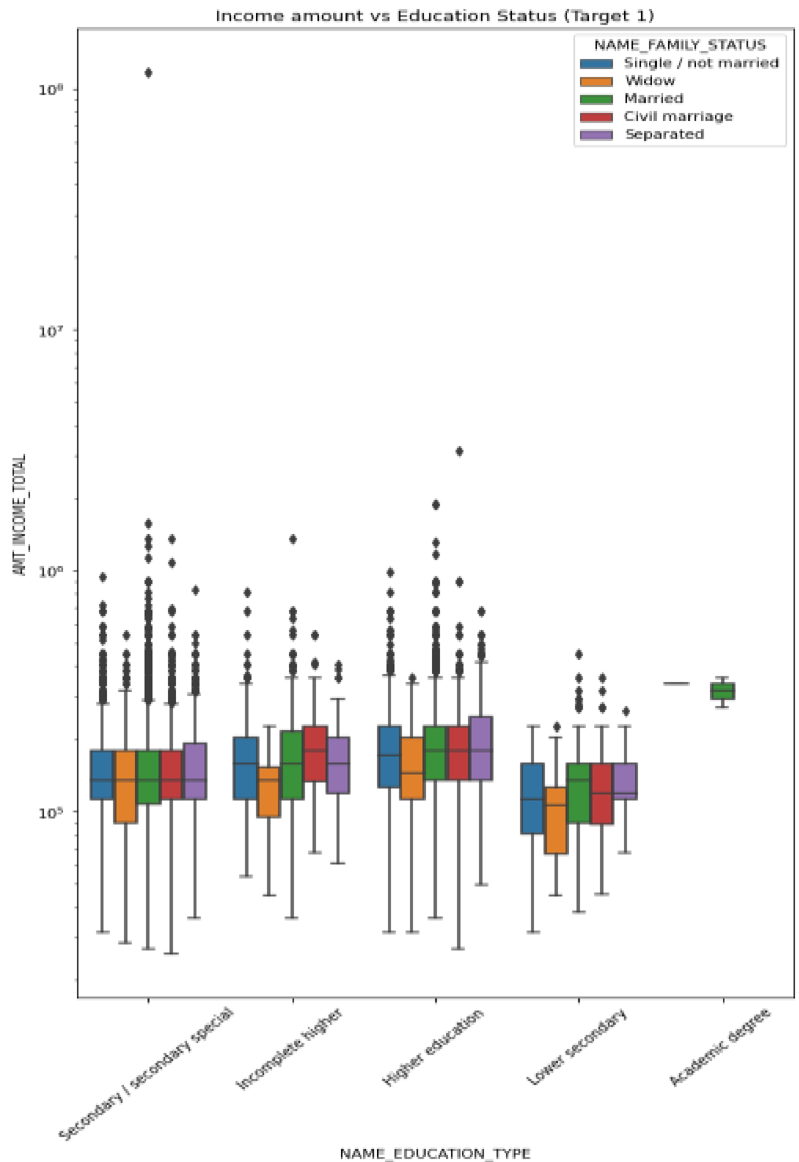


DISRIBUTION OFCREDIT AMOUNT VS EDUCATION STATUS

- 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.
- More number of outliers are seen in higher education.
- The people with secondary and secndary special education are less likely to make payments on time

INCOME VS EDUCATION STATUS

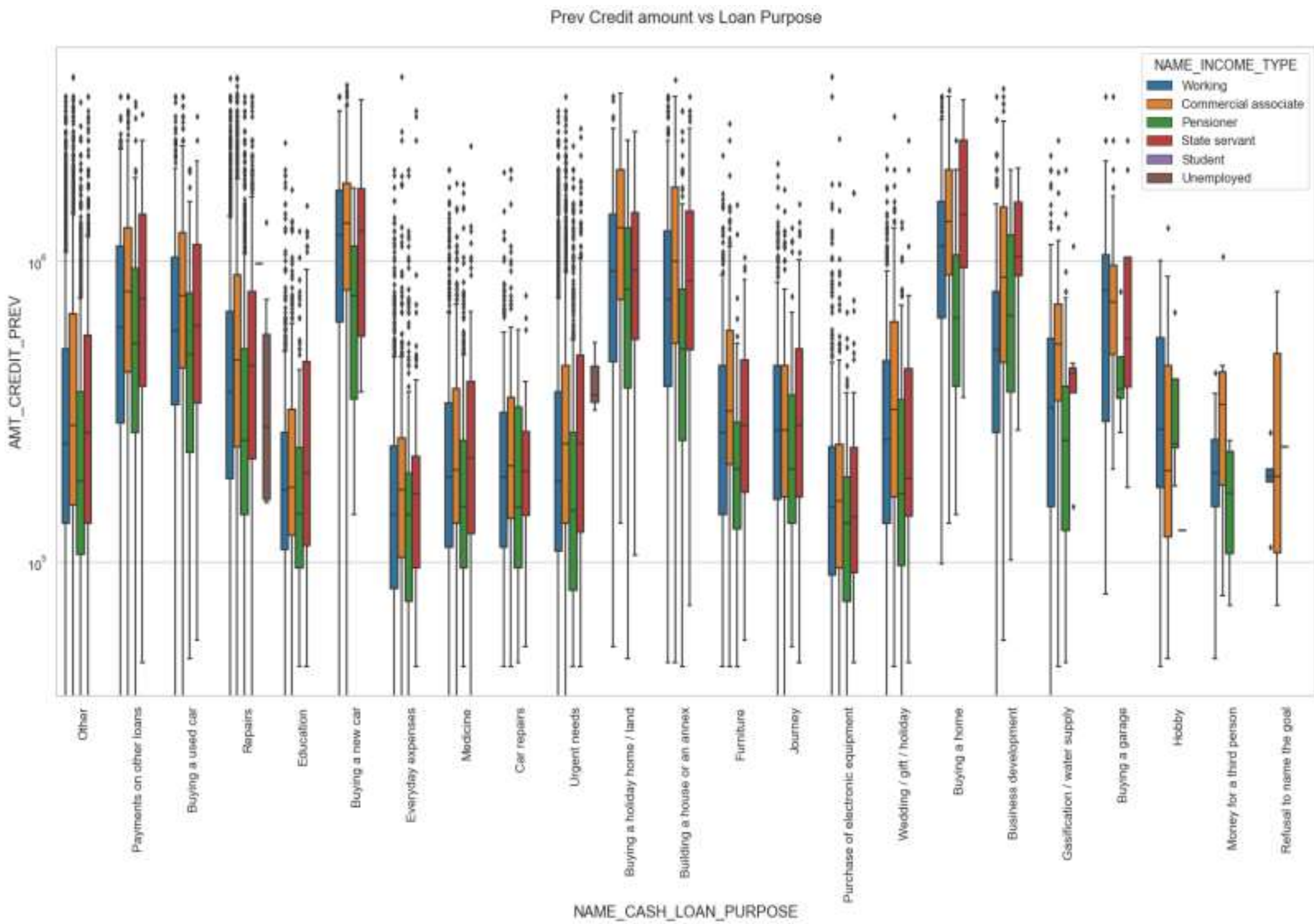




CONCLUSION FROM INCOME VS EDUCATION STATUS

1. we can see that Higher education has many outliers.
2. People with higher education have higher income and don’t have difficulties in making loan payment.
3. People with higher education who ave lesser income are unable to pay the loan.
4. Hence we can conclude that ,people with Higher income are most likely to make payments

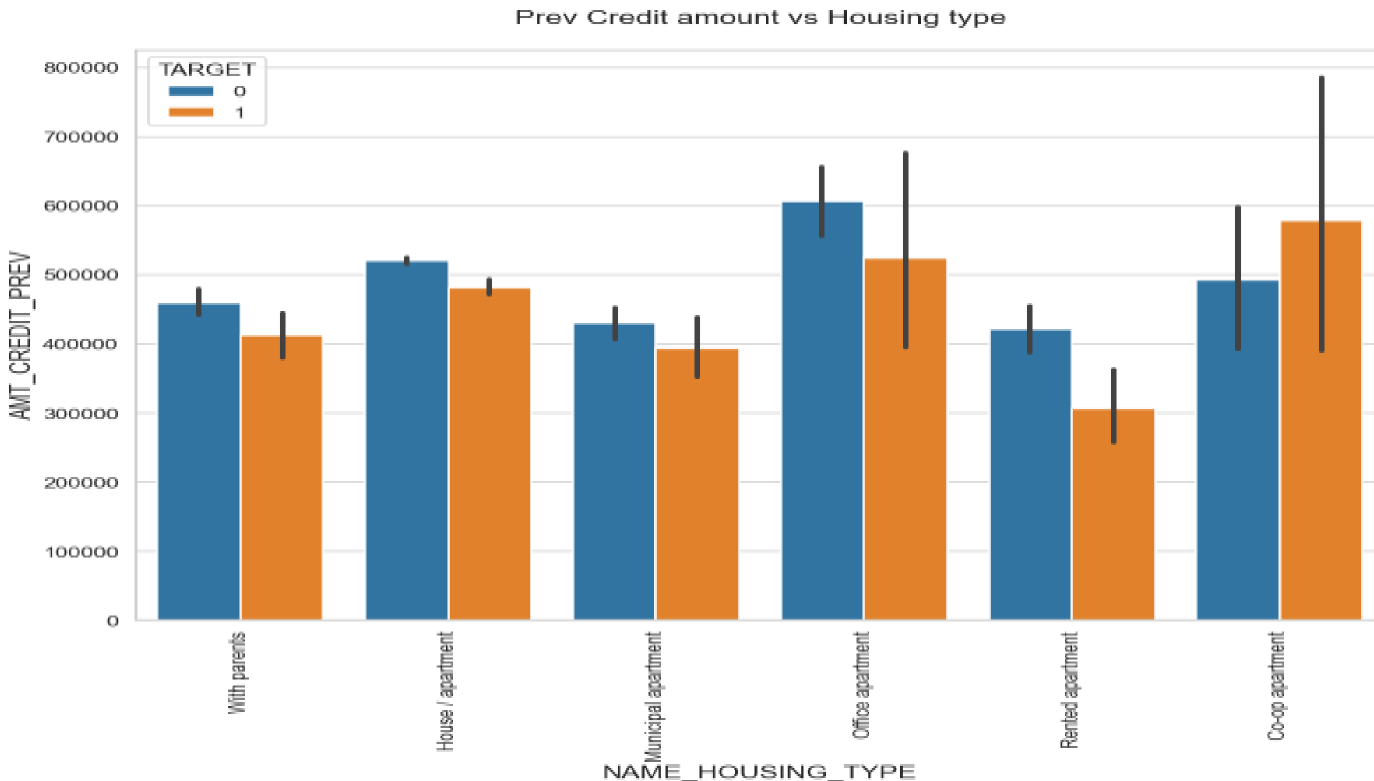
PREV CREDIT AMOUNT VS LOAN PURPOSE



CONCLUSSION FROM -PREV CREDIT AMOUNT VS LOAN PURPOSE

- From the above we can conclude some points- 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.

CREDIT AMOUNT PREV VS HOUSING TYPE



Here for Housing type, office appartment 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\appartment or miuncipal appartment for successful payments

- 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