Credit EDA Case Study

- KUNWAR URJASWIT

Problem Statement

The data given below contains the information about the loan application at the time of applying for the loan. 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.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Solution Overall Approach

We have used the below approach for deriving the insights :

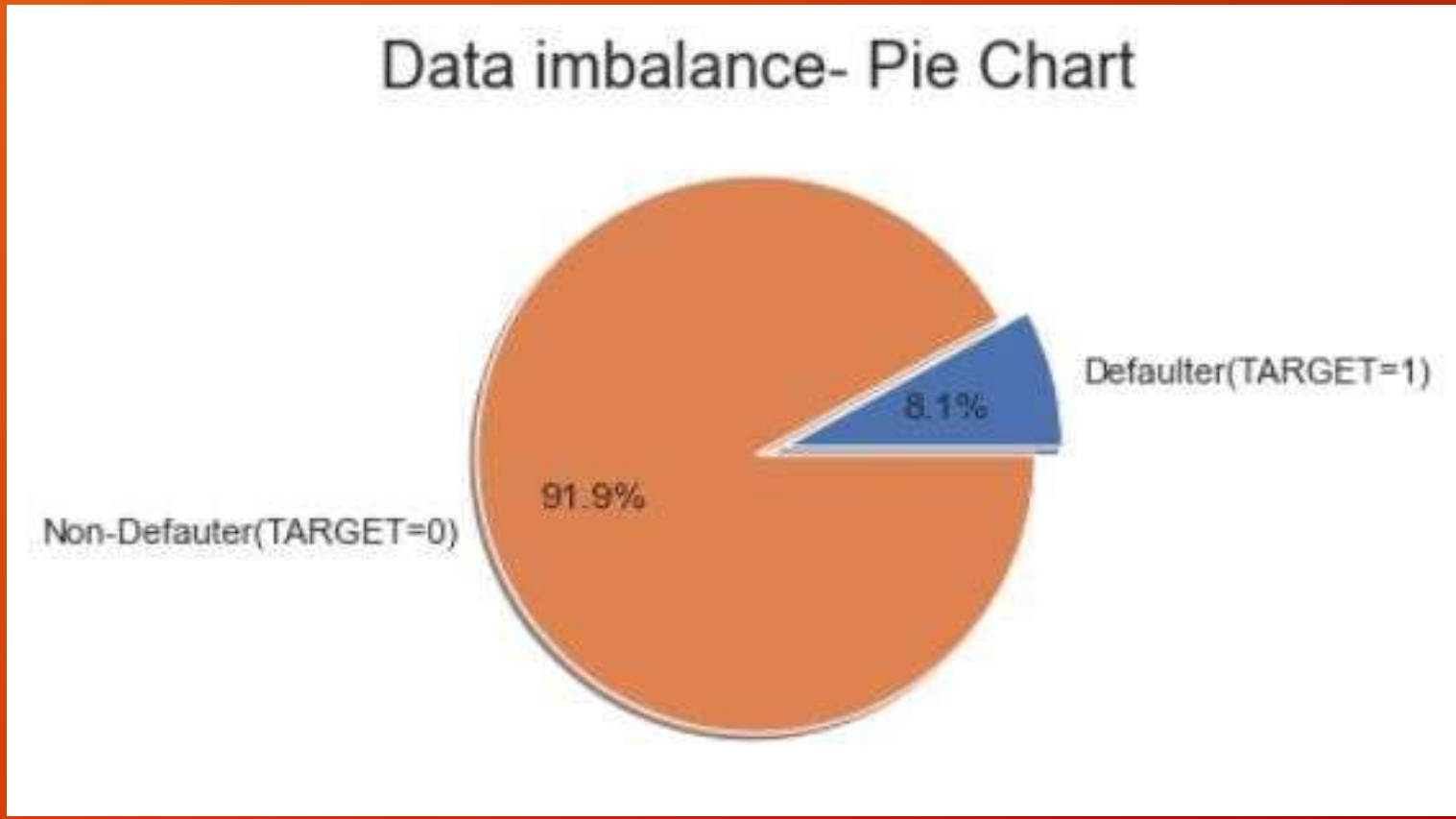
- The required libraries for data cleaning and visualisation were imported.
- We methodologically did **data cleaning** for rows & columns. wherever necessary median & mode values were used for replacement.
- Columns with majority of data missing were dropped.
- Columns providing no relevant data pertaining to problem statement were carefully selected & dropped.
- **Outliers** were methodologically identified and handled wherever possible.
- **Data imbalance** was checked.
- Created new columns as per the requirements
- **Univariate/Bivariate Analysis** of the relevant Categorical/numerical were done wrt. to Target variable and insights are inferred.
- Current and Previous application data are **merged** to derive meaningful insights based on bank Approval loan status .

Data Imbalance

Data imbalance was visible when the data was bifurcated w.r.t Target variable.

Ratio of Data Imbalance

11.39



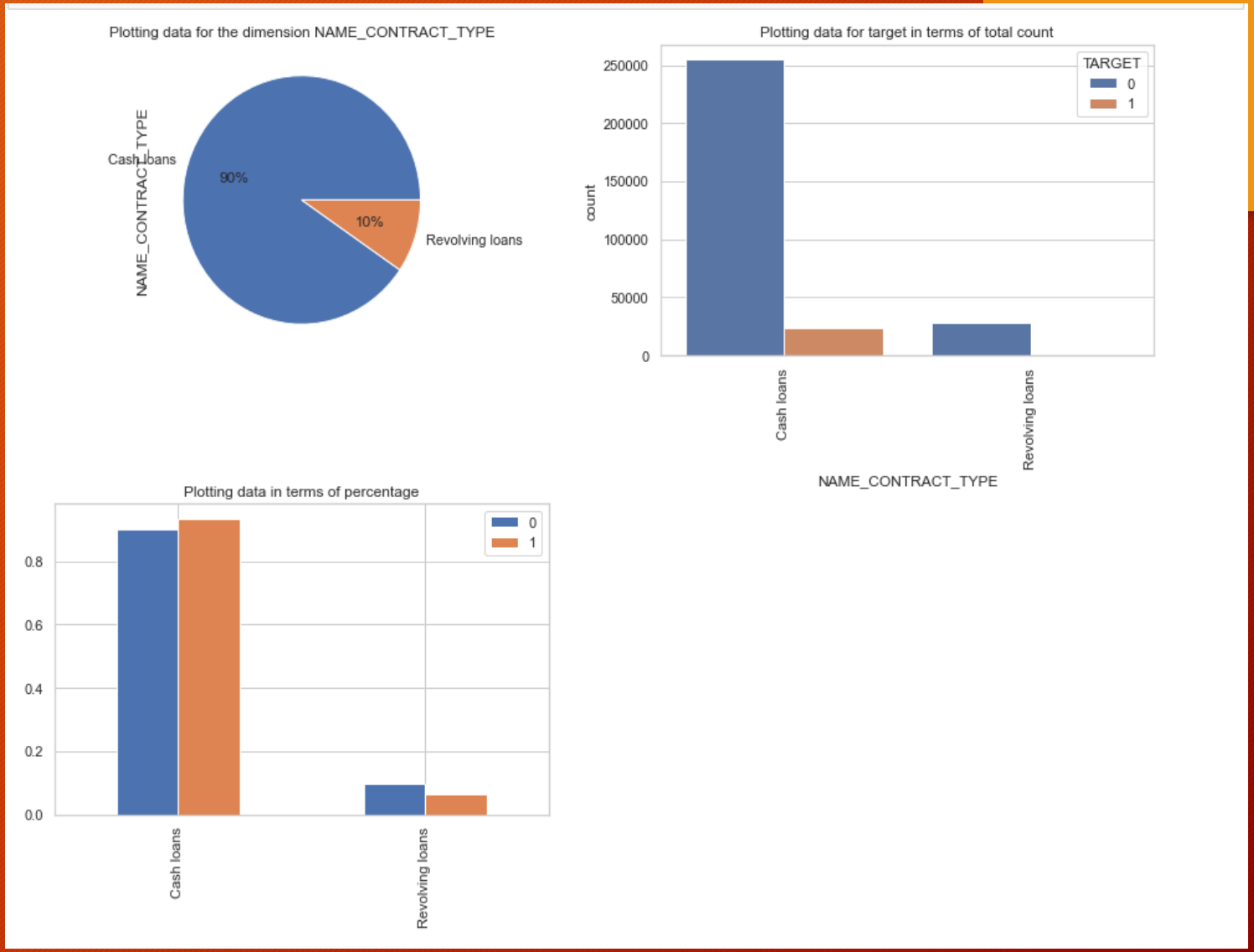
Univariate/Bivariate Analysis of Categorical columns

Bank is primarily providing 2 types of loans :Cash & Revolving .

Majority of the loans are cash loans

Analysis w.r.t Target Variable

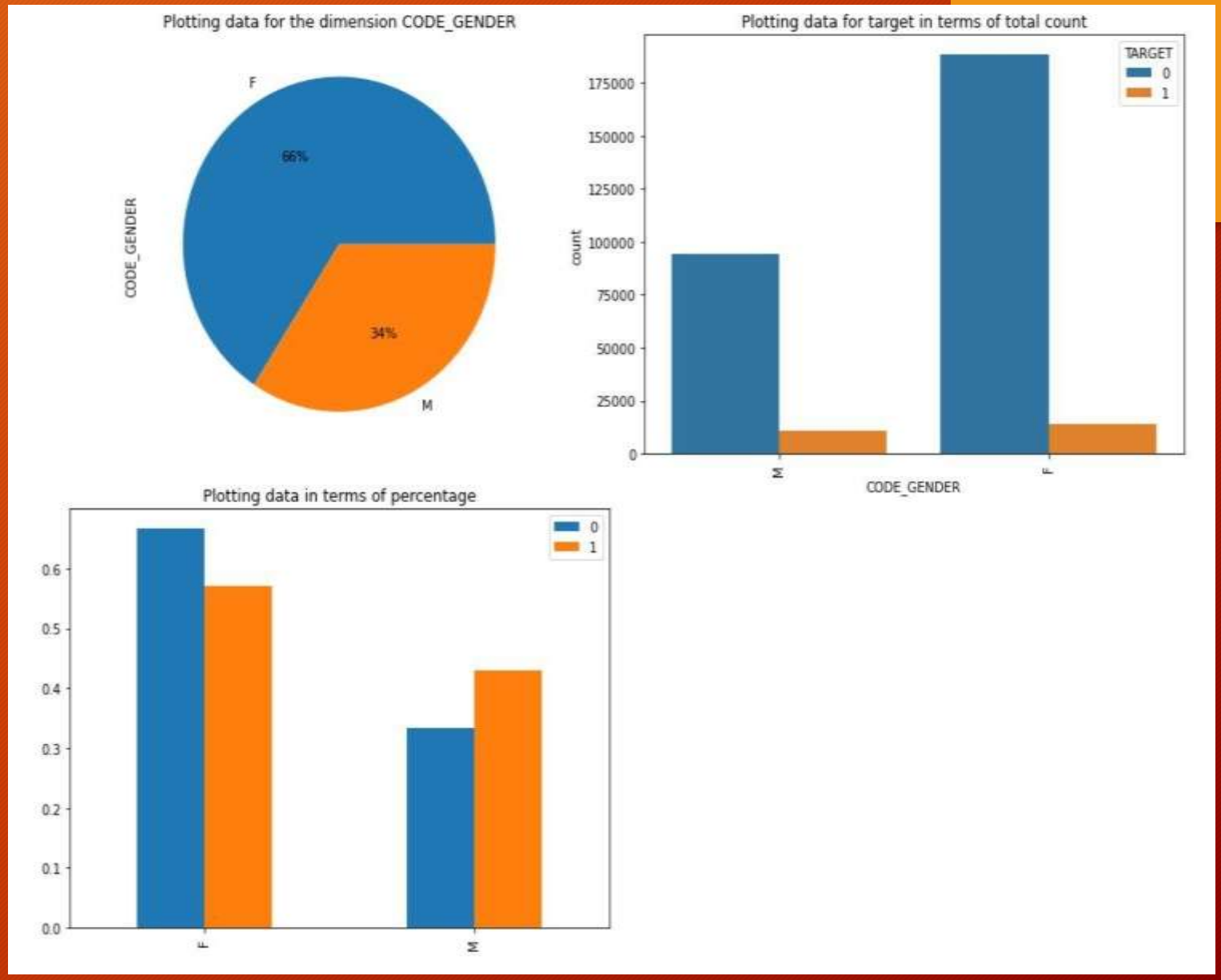
- Clients taking Cash loans are having more difficulty in returning the loan than the clients who have taken Revolving loan.



Females are opting for more loans than males.

Analysis w.r.t Target Variable

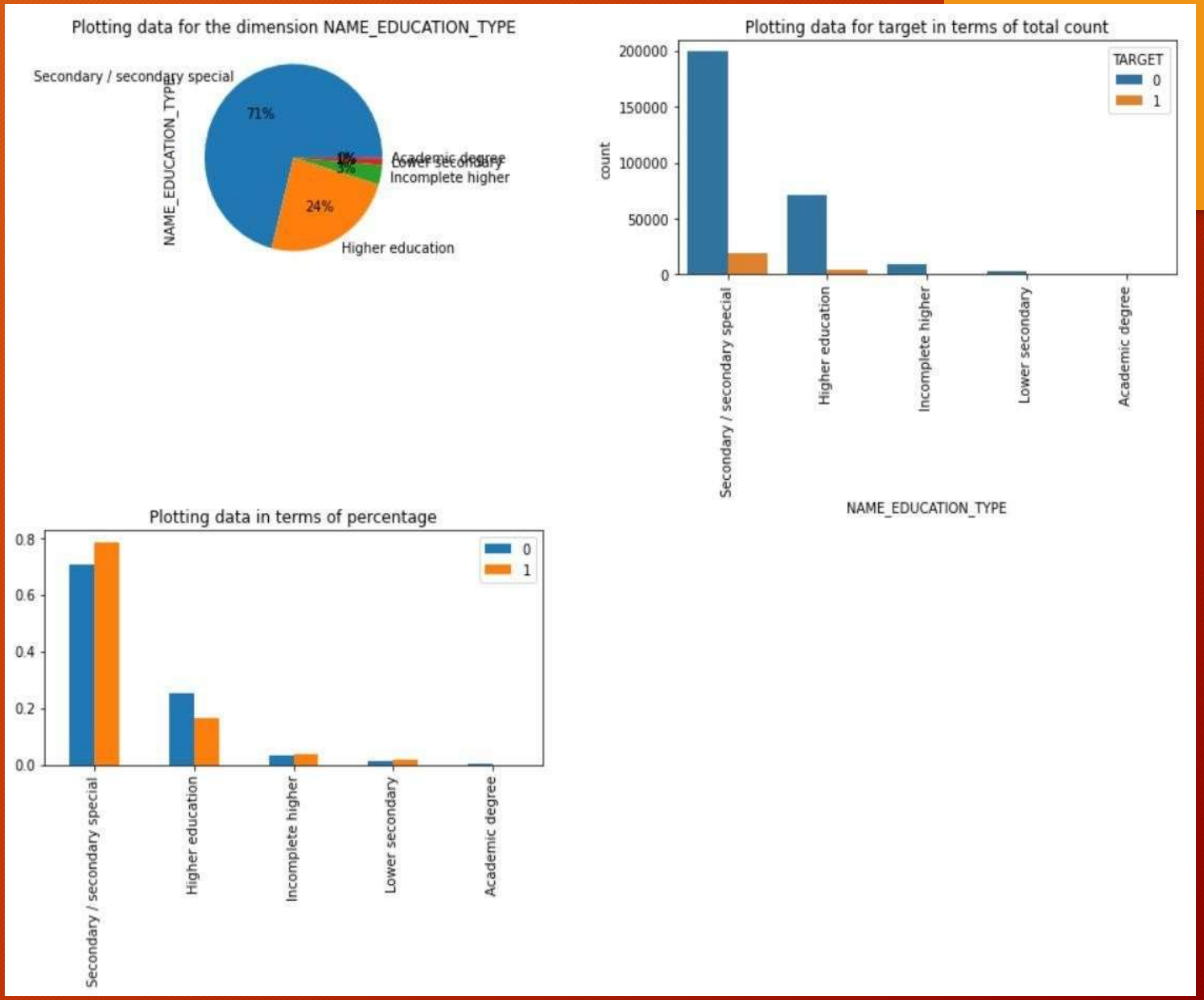
- Males are facing more payment difficulties than Females.



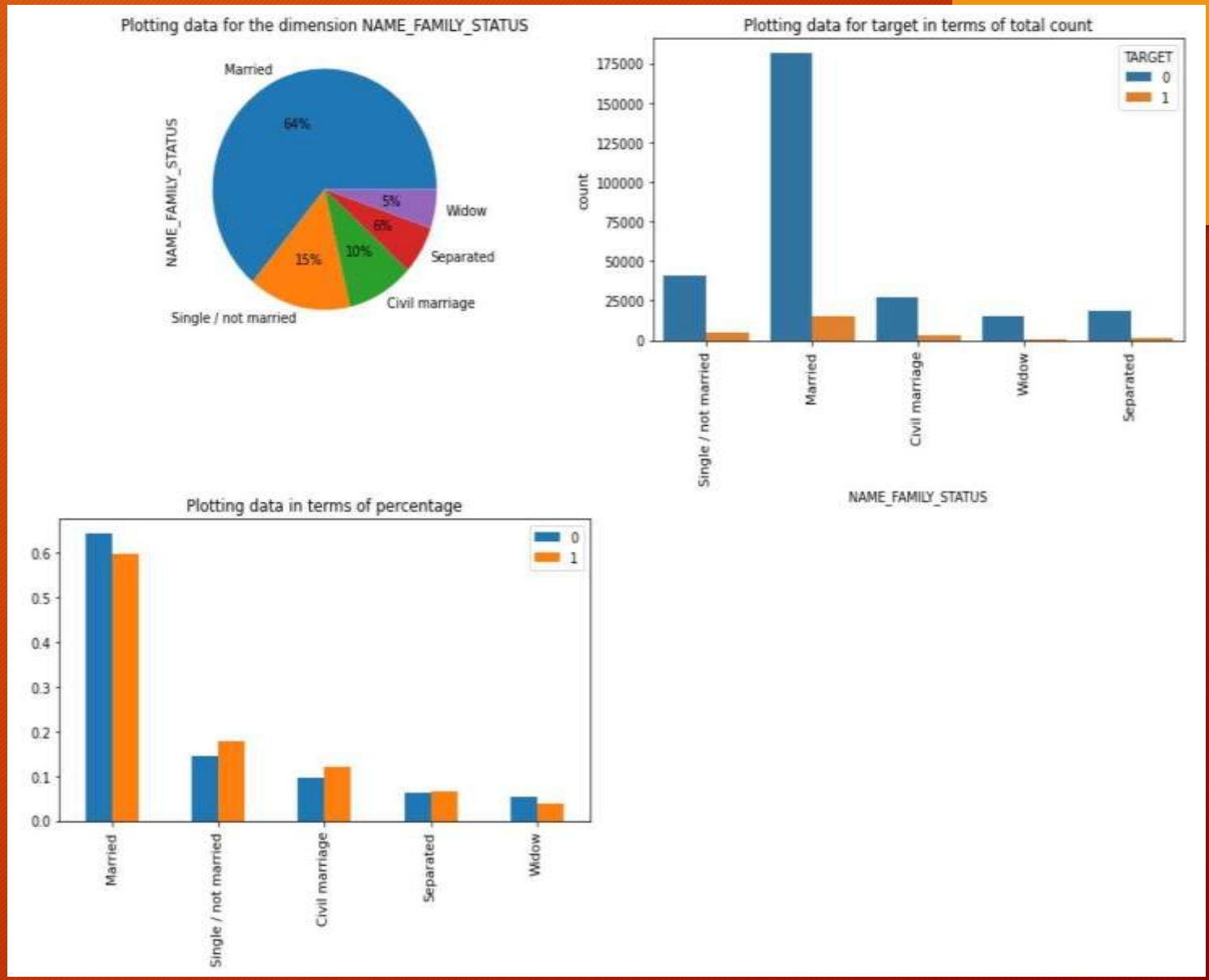
Most loans are provided to Secondary/Secondary Special and Higher education people.

Analysis w.r.t Target Variable :

- People with Higher education has less difficulty in repaying the loan in comparison to people with Secondary education.
- It can be inferred that better education helps in getting better paying jobs.



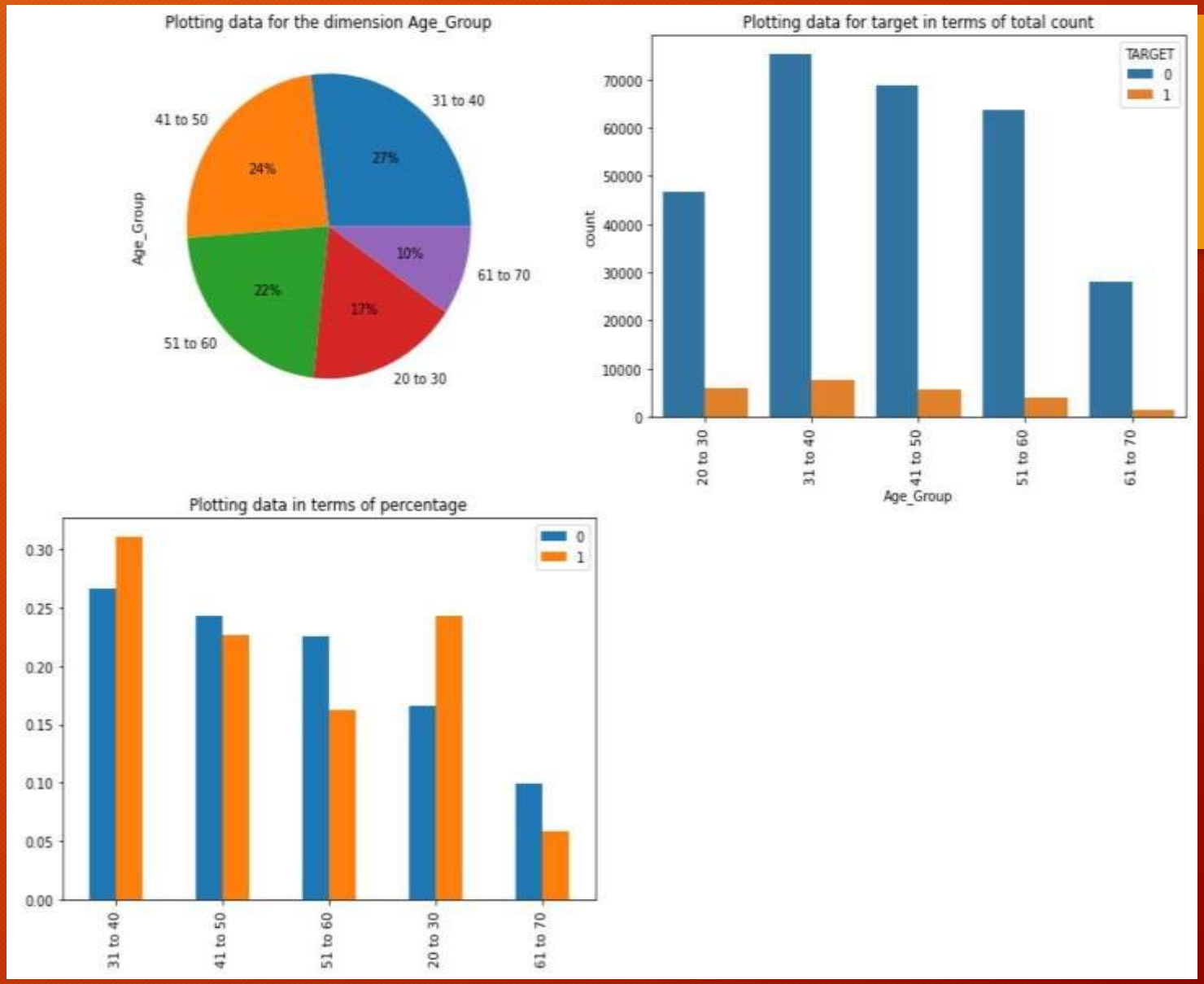
Majority of the loan is allocated to married people.



Analysis w.r.t Target Variable :

- Married people are finding it easier to repay their loans wrt. Single people.
  - This may be due to their better financial stability & management or married people may have dual source of income.
- Widowed people are also finding it less difficult to repay their loans.

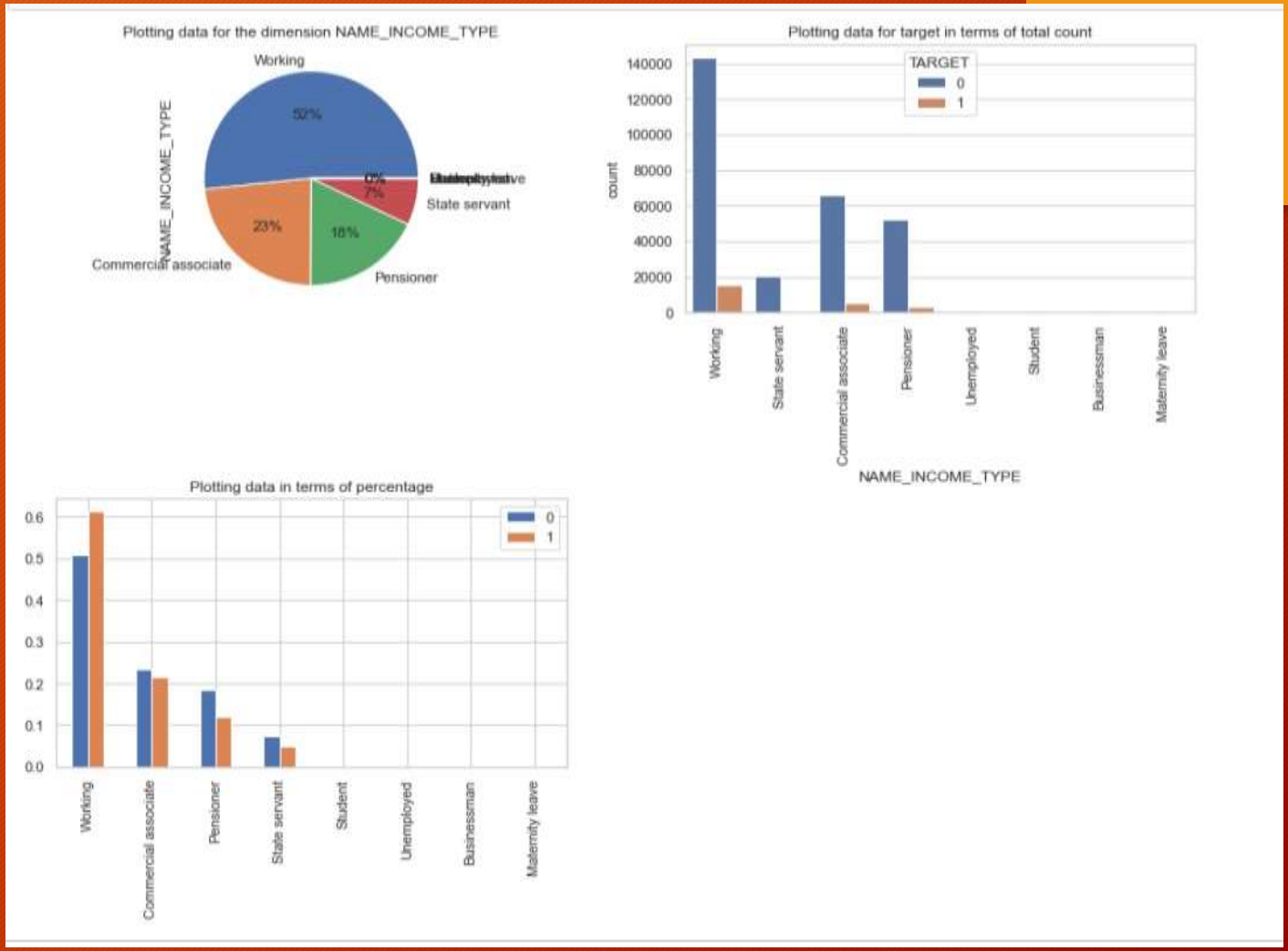
People in 20-30 age group are allocated less loans along with people 61-70 age.



Analysis w.r.t Target Variable :

- Middle aged (40+) & Senior citizens(60+) are facing less difficulties in repaying loans,may be due to better financial stability.
- People between 20-40 are facing more difficulties in repaying,may be due to their less salary.

Majority of loan is applied by Working people category



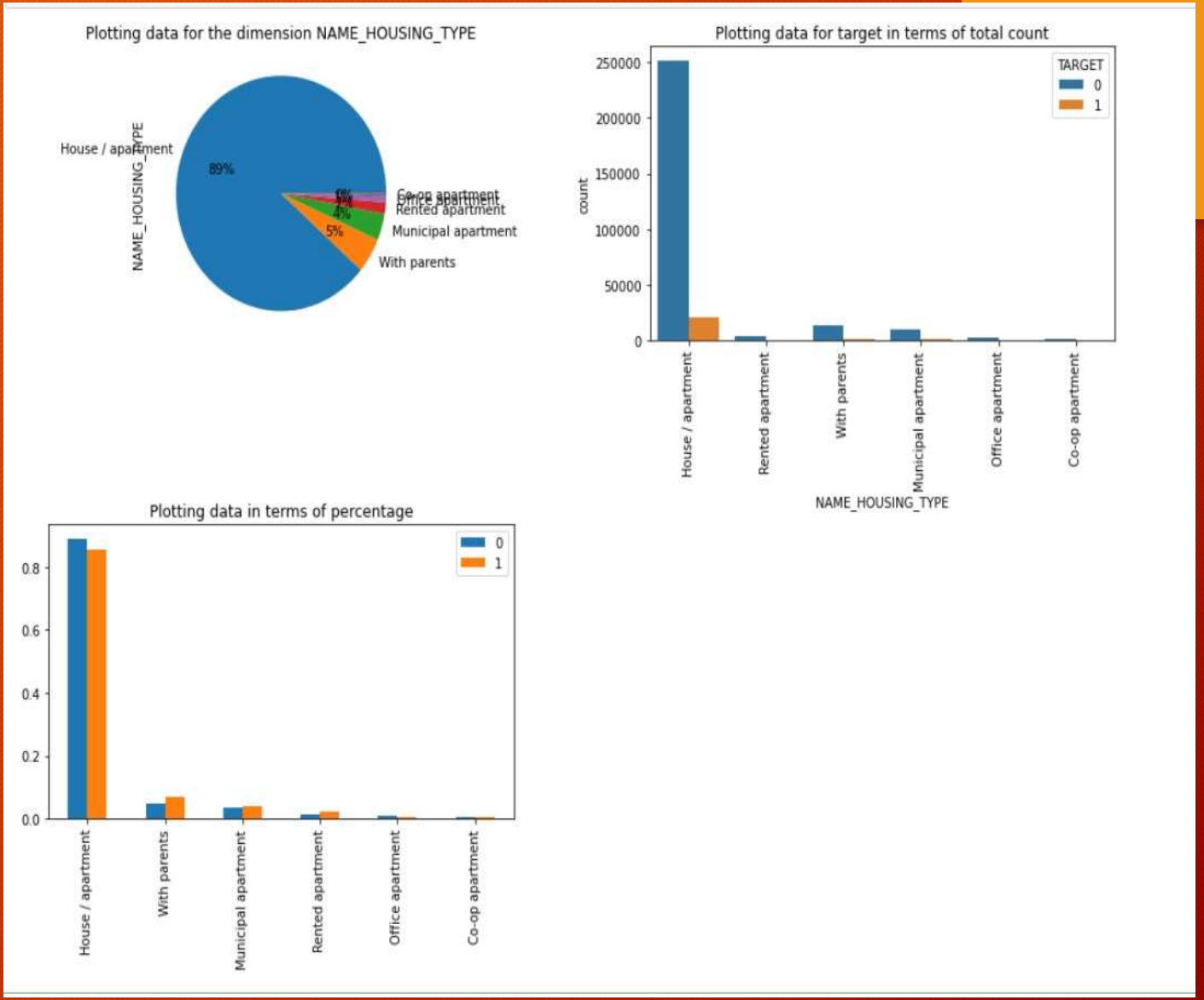
Analysis w.r.t Target Variable :

- It can be inferred that working people are facing more difficulties in comparison to other income type guys in repaying the loan.

People living in apartments or houses are opting more for more loans.

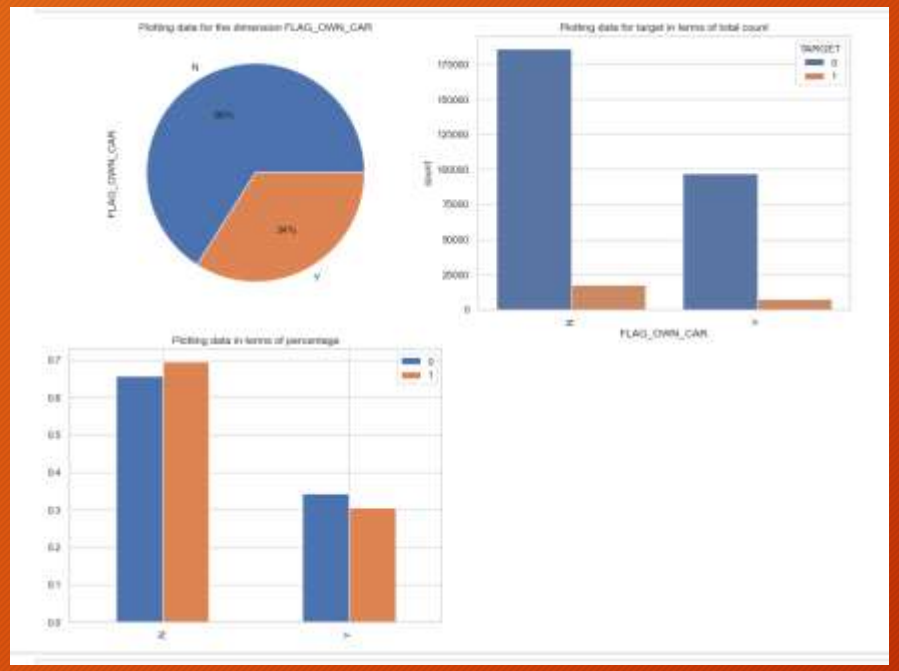
Analysis w.r.t Target Variable :

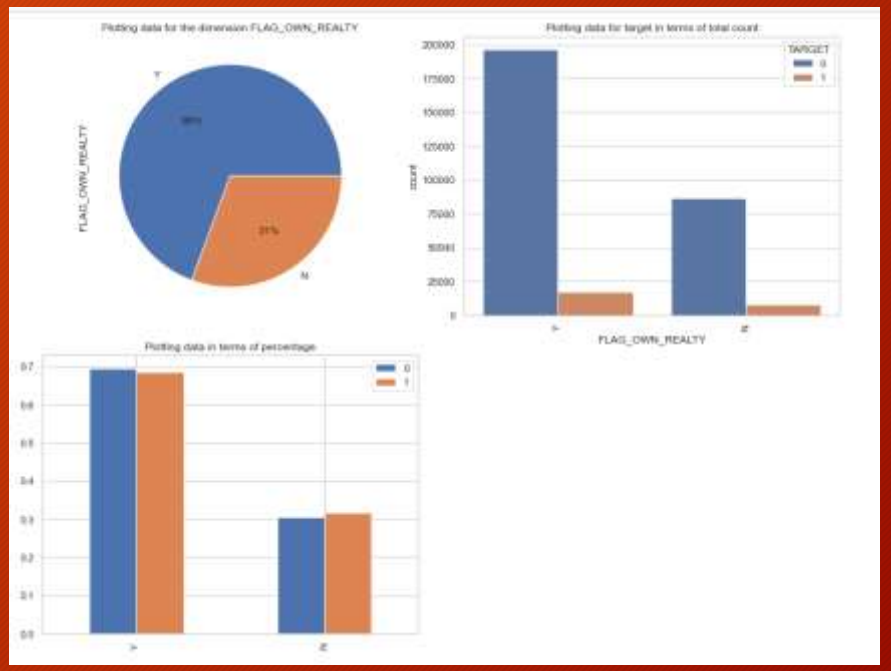
- Clients living with parents are facing more difficulties in repaying the loan. Probably their medical/other expenses are more.



People already possessing a car/real estate property are facing comparatively less difficulties in repaying the loan.

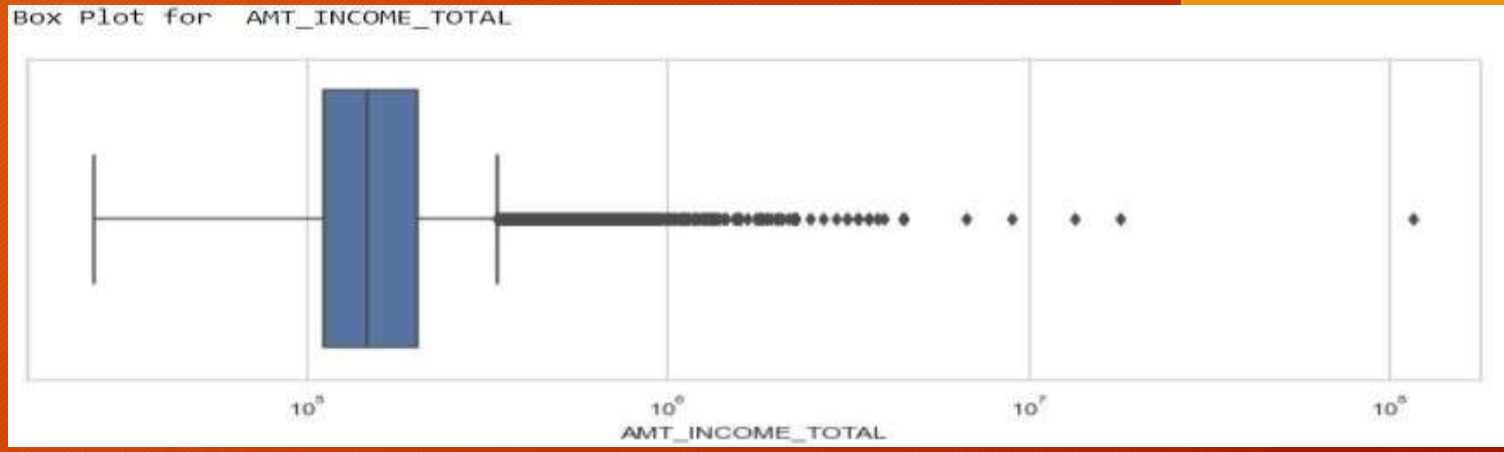
It can be inferred that they are already in a good financial position and able to pay the instalments on time.





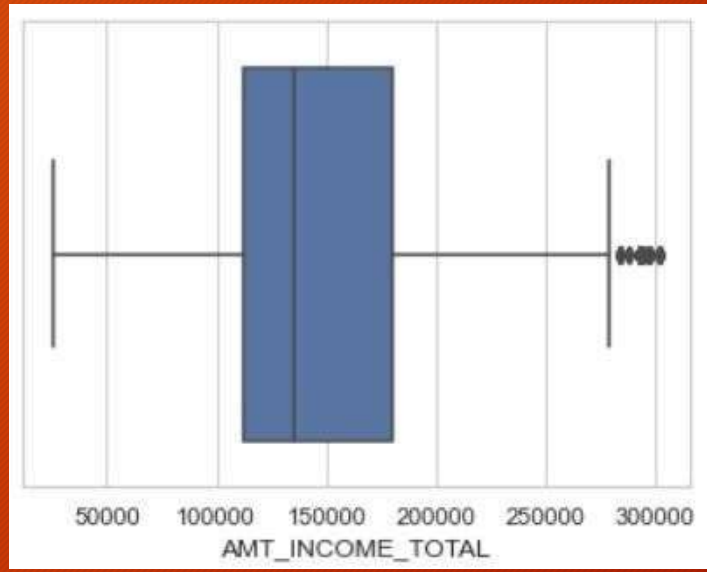
Univariate/Bivariate Analysis and Outliers for Numerical columns

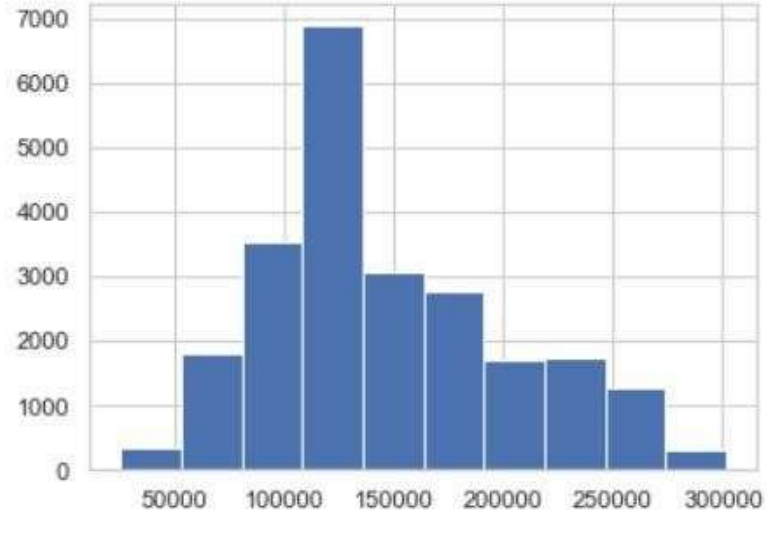
The graph depicts that the Income column has lot of outliers



After handling the outliers it can be deduced that :

- Clients having income between 100000 to 150000 is the segment where more clients are facing repayment difficulties.

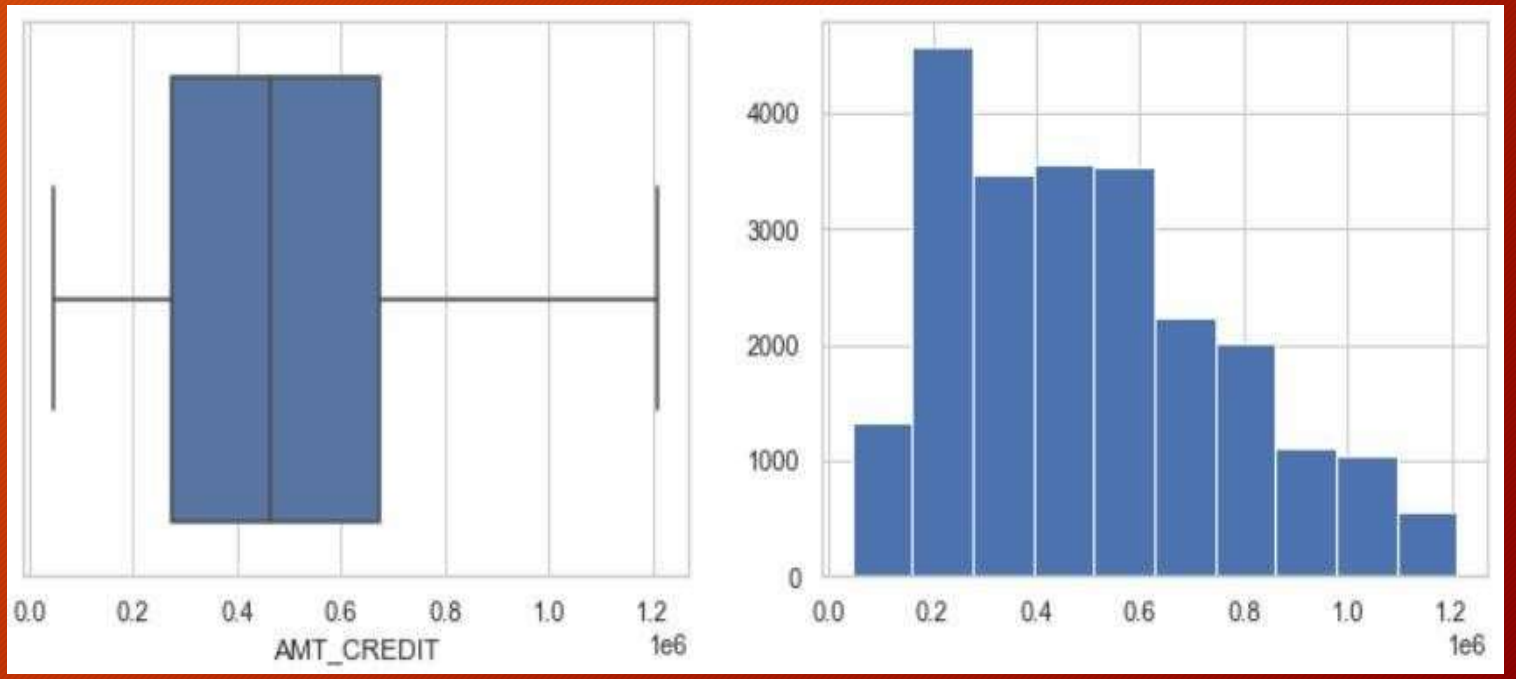




The graph depicts that the Income column has lot of outliers

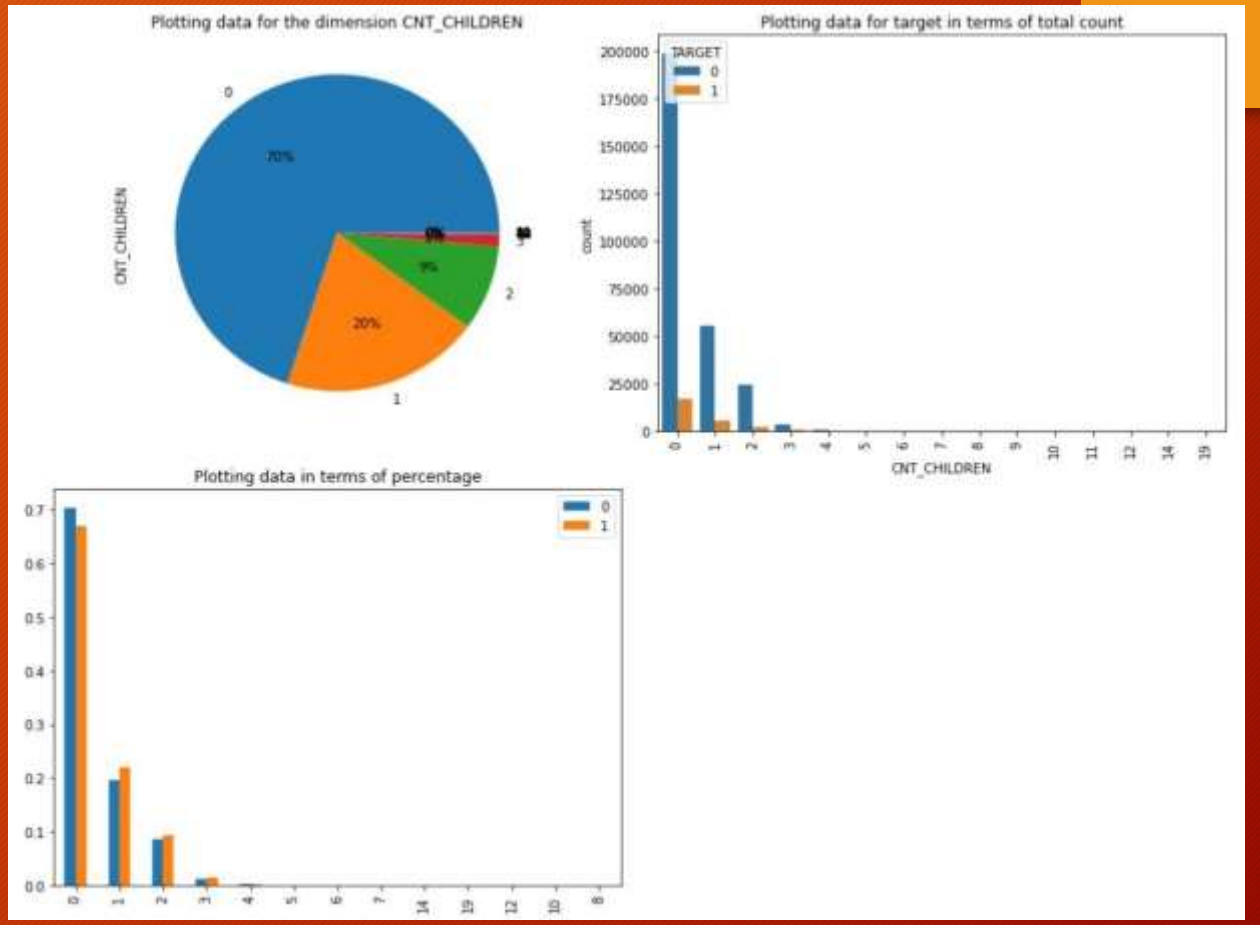
After handling the outliers it can be deduced that :

- Around 55 % of the population have loan ranging from 200000 to 600000.



Approximately 70% of the clients don’t have children:

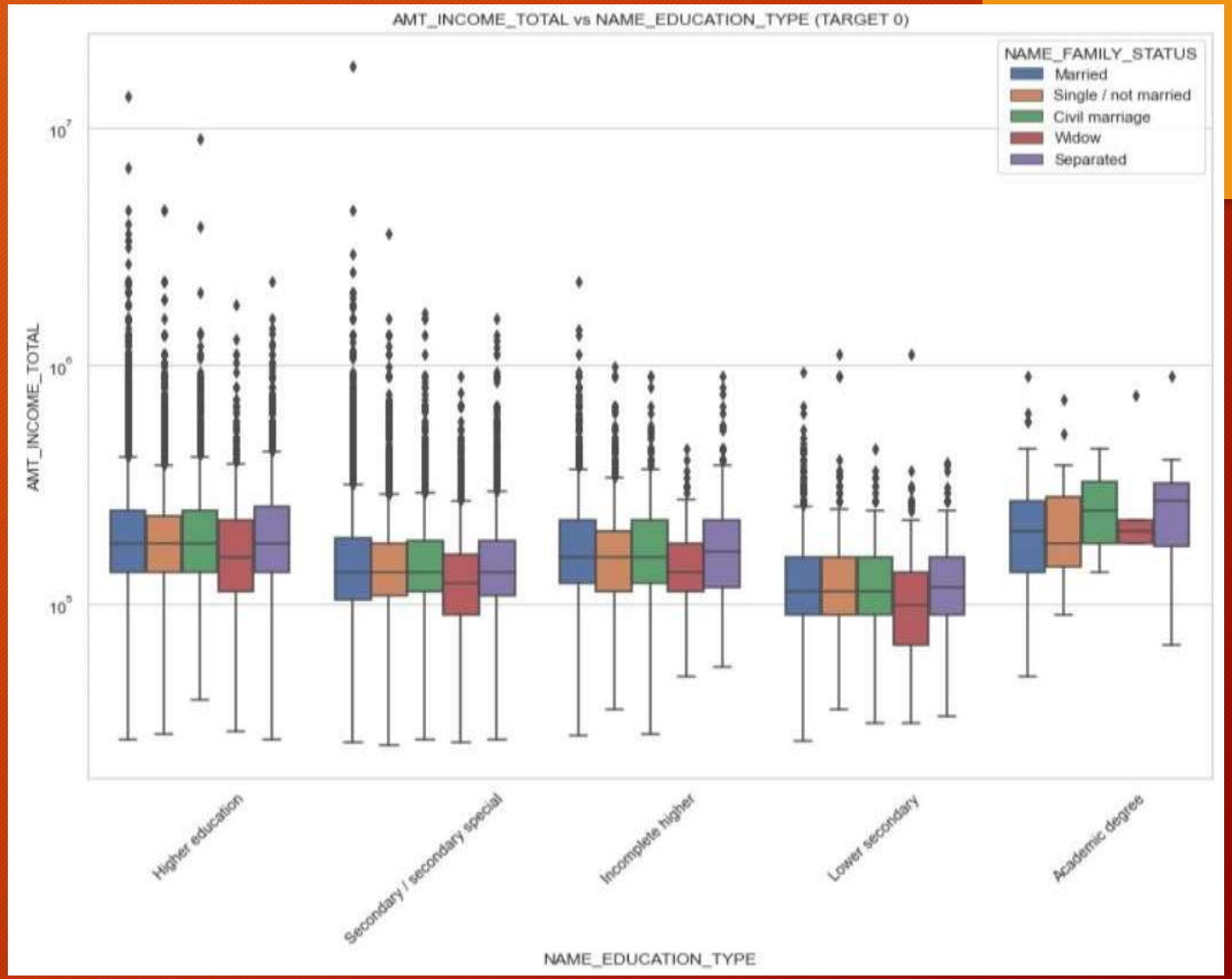
- People with no children have low chance to default.
- But as the number of children increases,people have increased difficulty in repaying the loan.



We can see that people in the beginning phase of their career i.e. experience between 1 -5 years face paying difficulties. And the experienced people once well settled and with good salary don't face much paying difficulties.

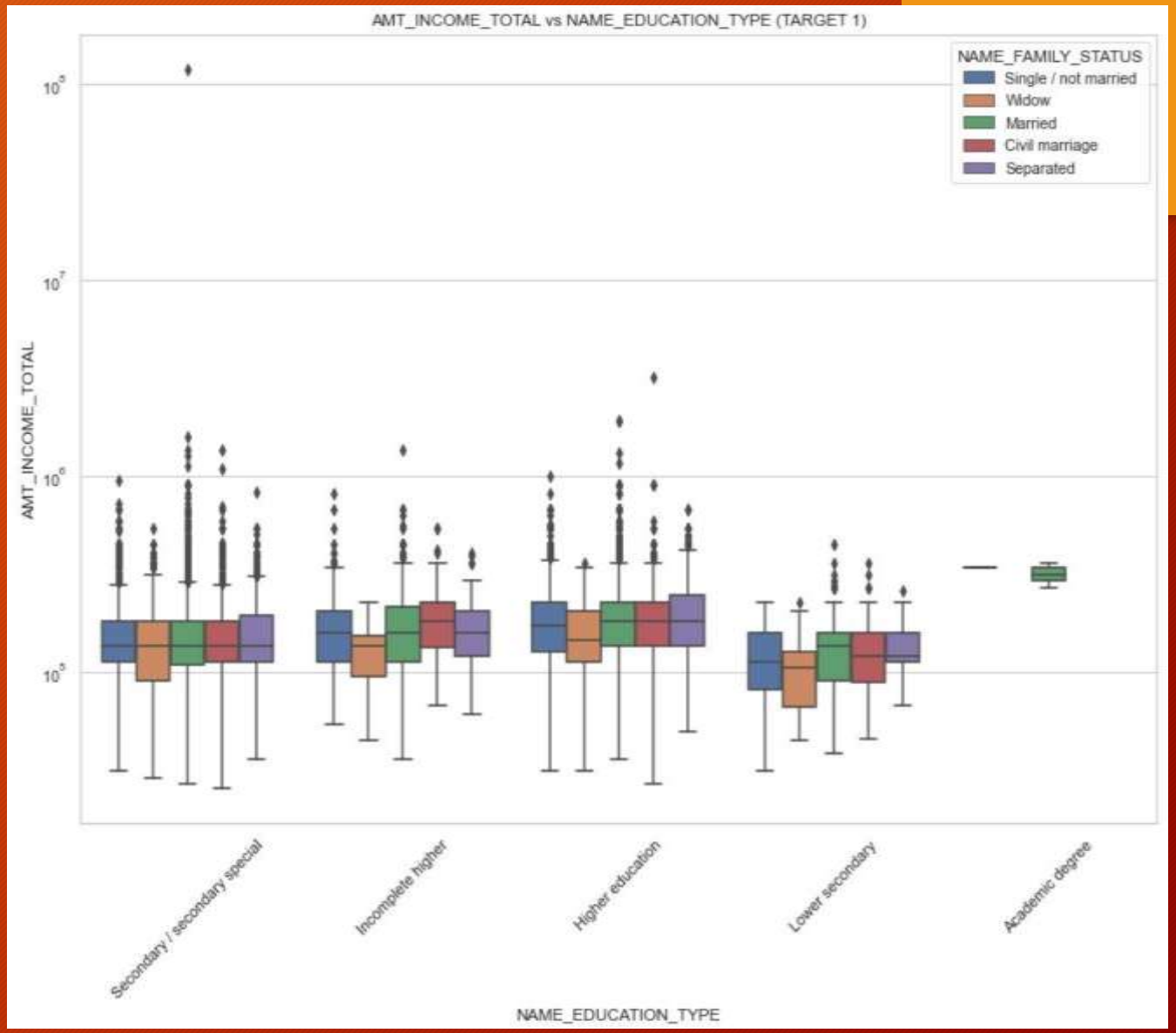
Multivariate Analysis

For non-defaulters : Education Type with Higher education and 'Secondary/secondary special' Amount income total is mostly equal among the family status. These two education type has many outliers as well. Academic degree education type have less outliers and also their amount income total seems to be little on higher side.

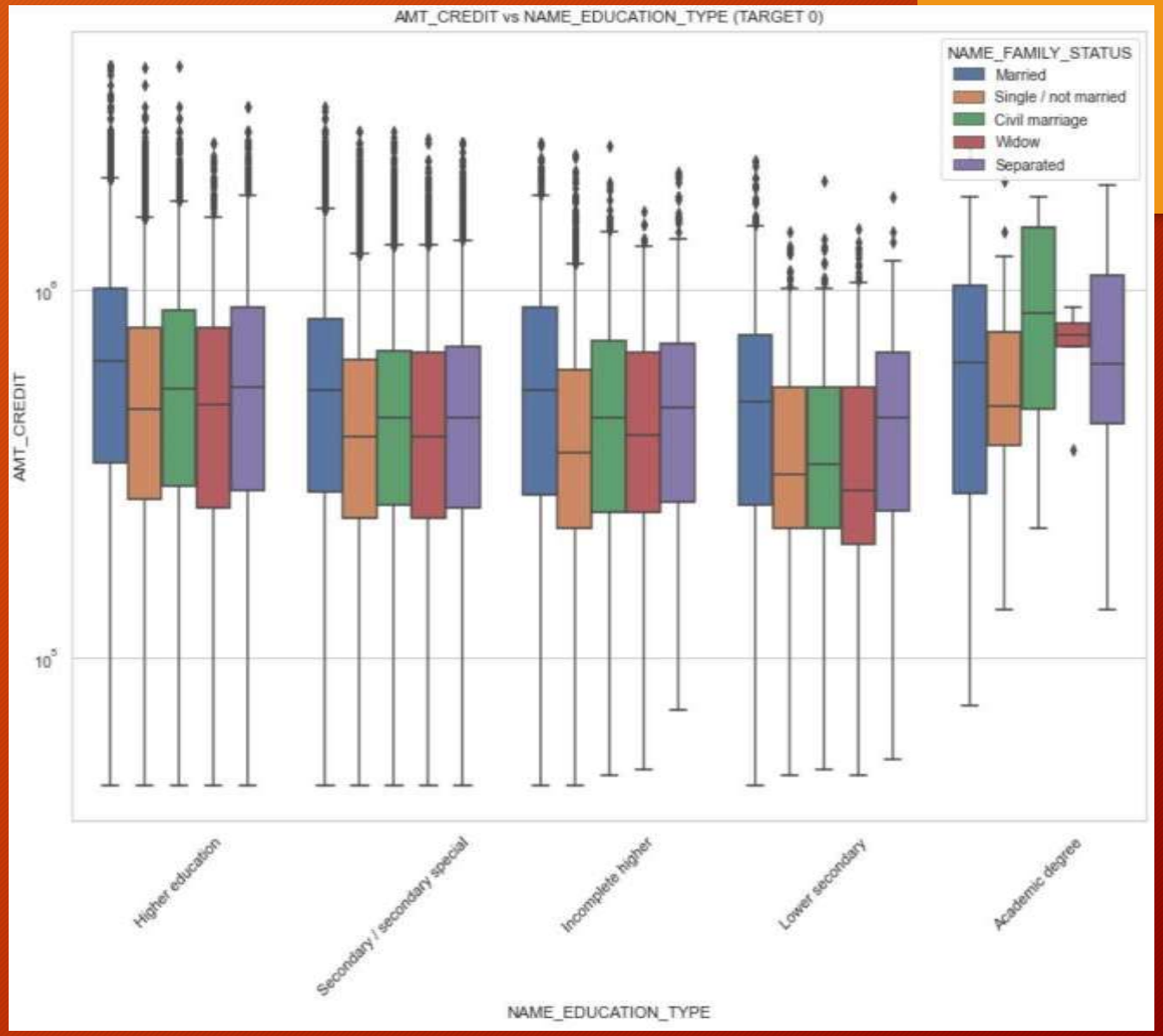


For Defaulters : Most of the outliers are from Education type 'Secondary/secondary special', 'Incomplete higher' and 'Higher education'. Very few outliers observed for Lower secondary and Academic.

Single, civil marriage and separated family status group has almost similar income amount under Secondary/secondary special status.



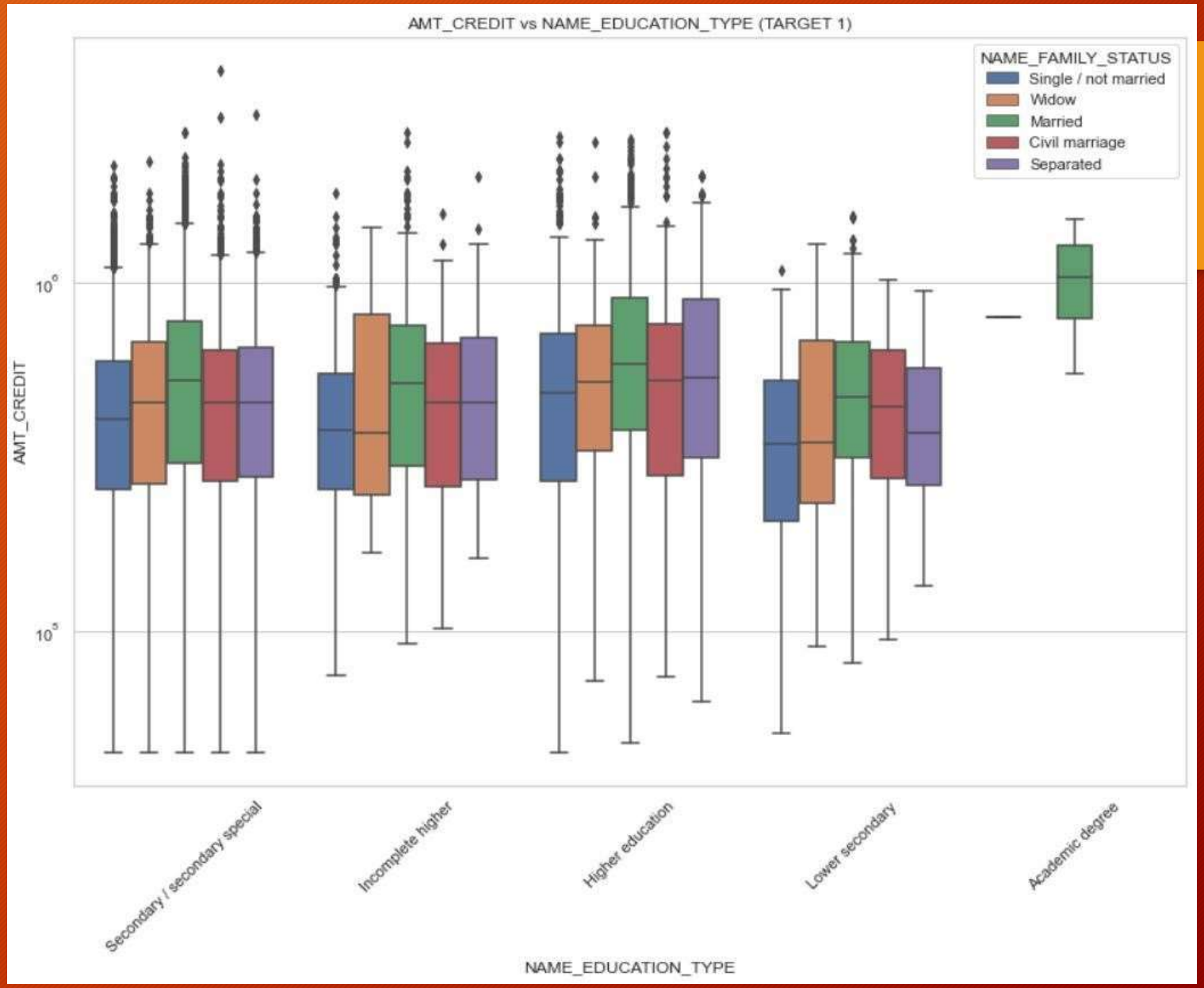
For Non Defaulters : Academic degree education type have less outliers and also family status of Married, civil marriage and separated have higher Amount Credits. Whereas, Higher education and Secondary/secondary special group have many outliers



For Defaulters : Most of the outliers are from Education type 'Secondary/secondary special', 'Incomplete higher' and 'Higher education'.

Very few outliers observed for Lower secondary and Academic.

Single, civil marriage and separated family status group has almost similar amount credit under Secondary/secondary special status.

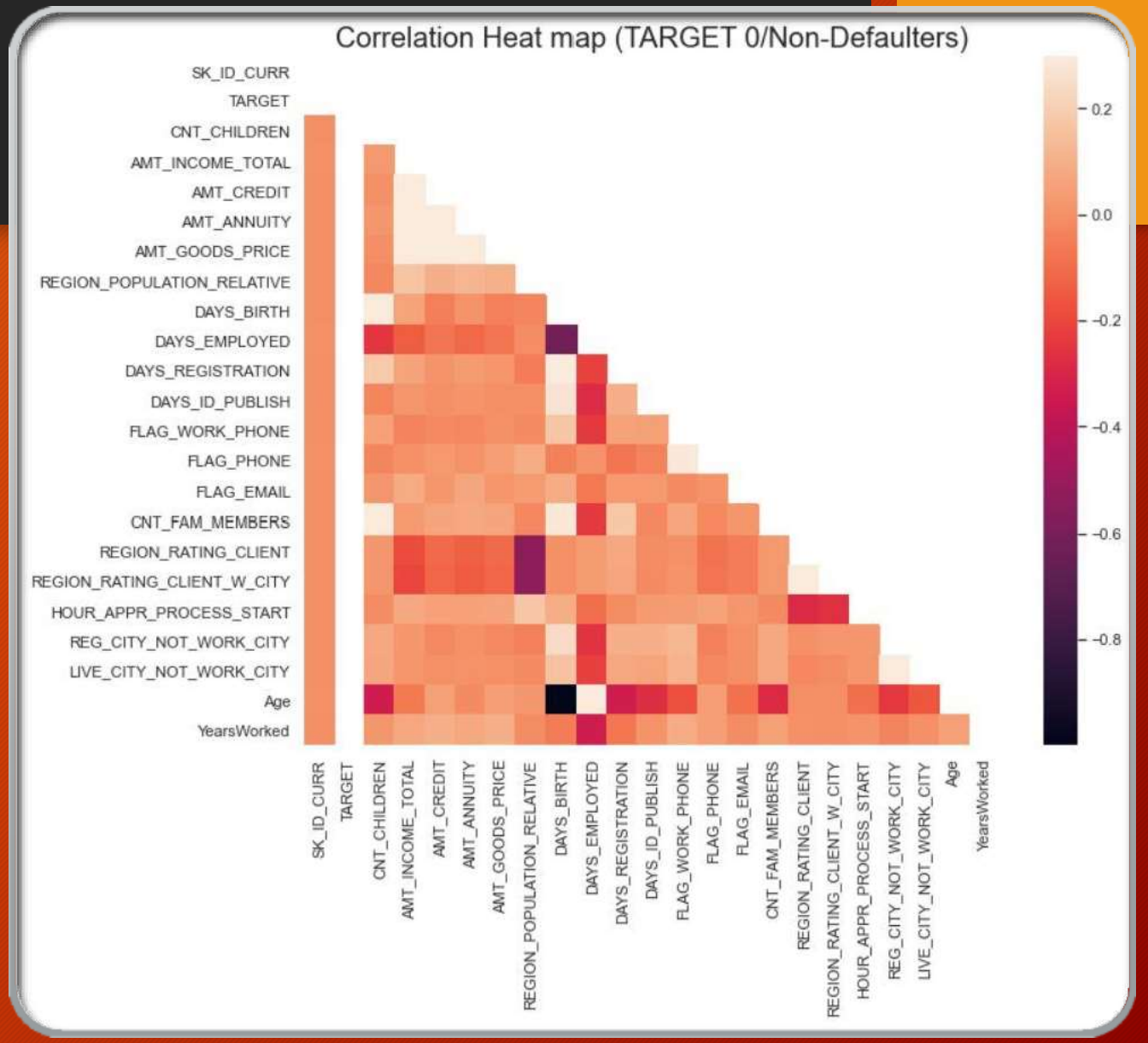


Correlation

Correlation in Non Defaulters :

Top 10 correlation (Positive and Negative) :

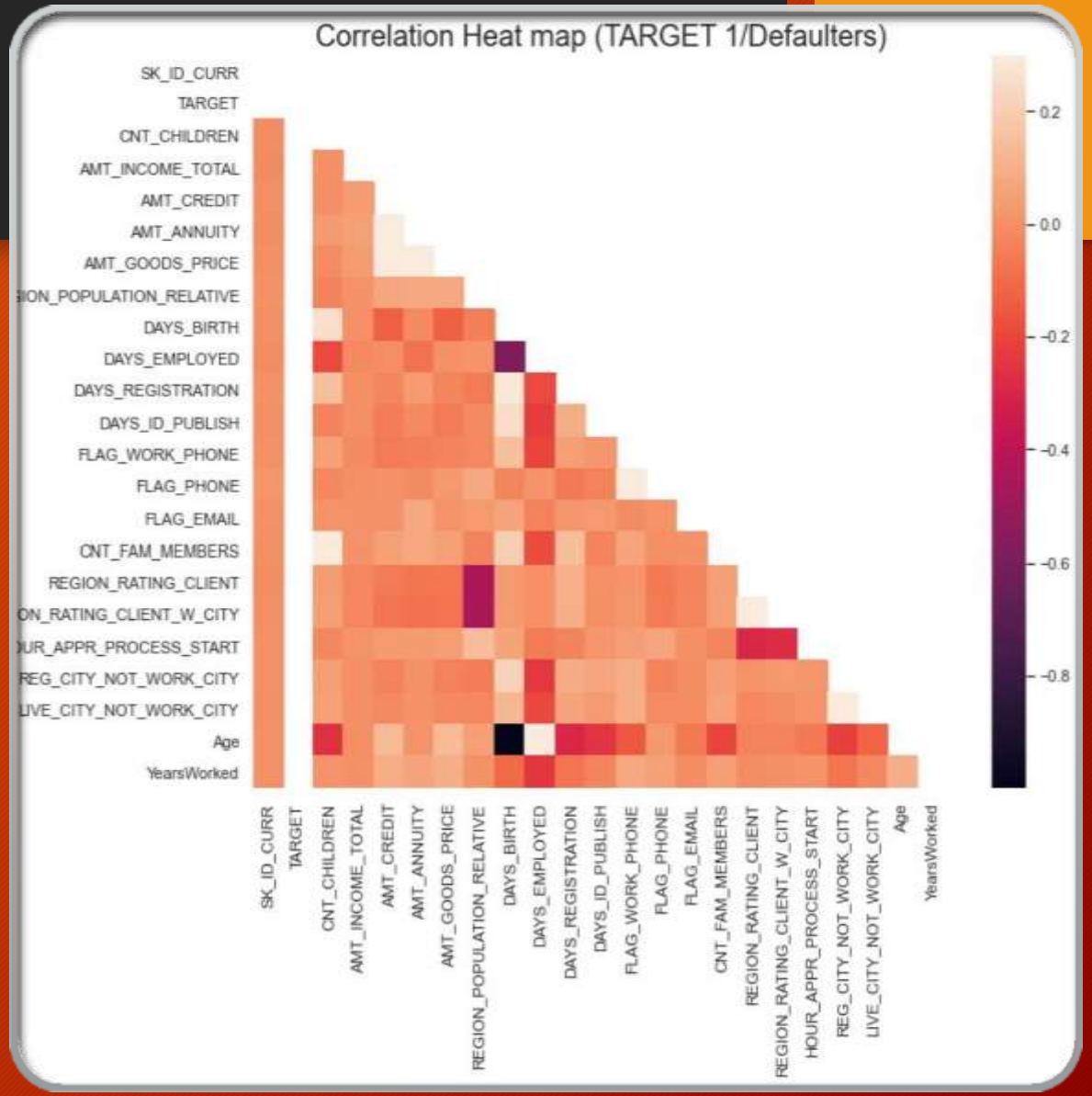
- DAYS_BIRTH---Age (**99.97 %**)
- AMT_CREDIT---AMT_GOODS_PRICE (**98.70 %**)
- REGION_RATING_CLIENT_W_CITY---REGION_RATING_CLIENT (**95.01 %**)
- CNT_FAM_MEMBERS---CNT_CHILDREN (**87.86 %**)
- LIVE_CITY_NOT_WORK_CITY ---REG_CITY_NOT_WORK_CITY (**83.04 %**)
- AMT_GOODS_PRICE---AMT_ANNUITY (**77.64 %**)
- AMT_CREDIT---AMT_ANNUITY (**77.13 %**)
- DAYS_BIRTH---DAYS_EMPLOYED (**61.80 %**)
- DAYS_EMPLOYED---Age (**61.80 %**)
- REGION_POPULATION_RELATIVE---REGION_RATING_CLIENT (**53.90 %**)



Correlation in Defaulters :

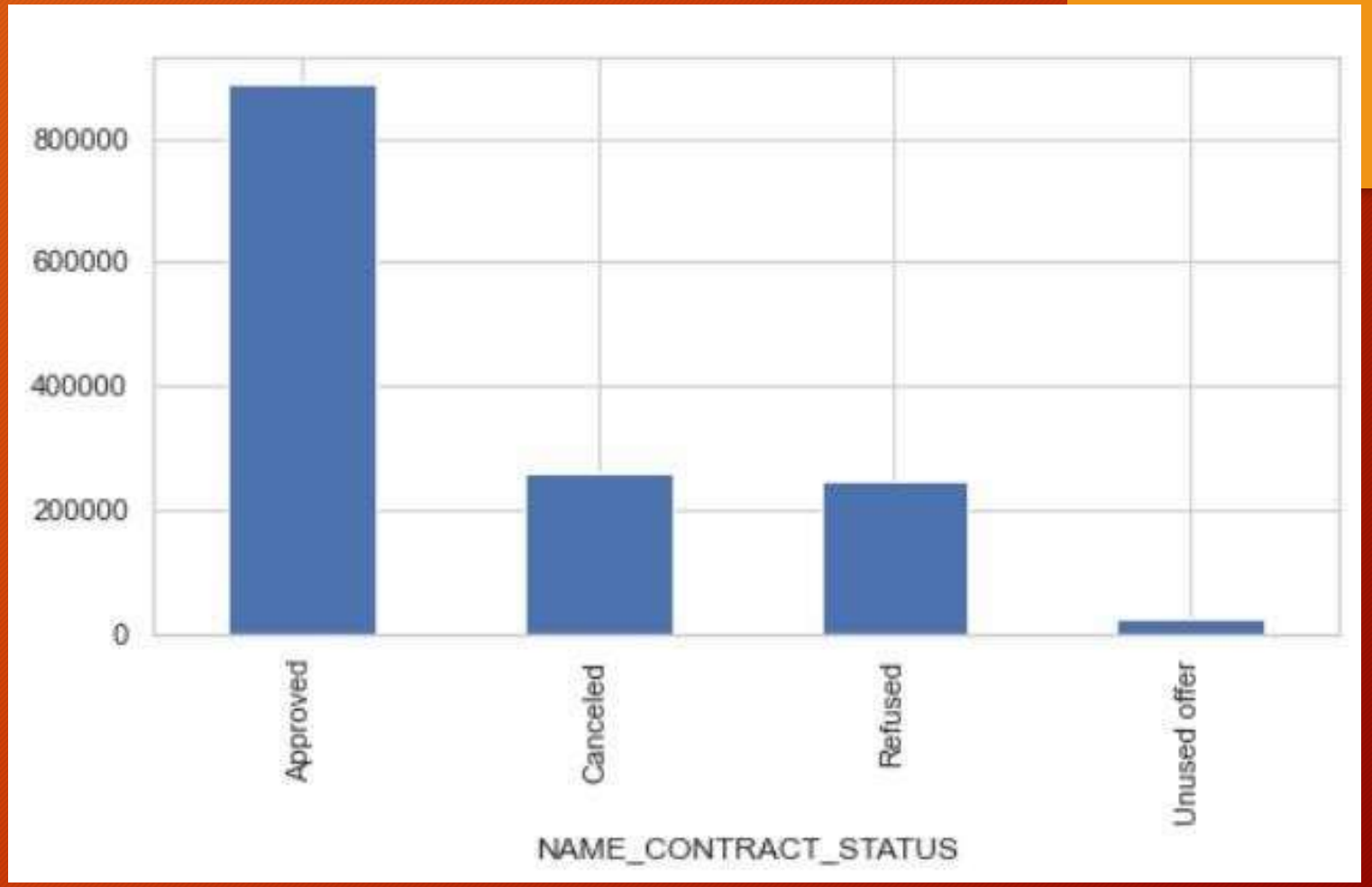
Top 10 correlation (Positive and Negative) :

- DAYS_BIRTH---Age (**99.97 %**)
- AMT_CREDIT---AMT_GOODS_PRICE (**98.28 %**)
- REGION_RATING_CLIENT_W_CITY---REGION_RATING_CLIE NT (**95.66 %**)
- CNT_FAM_MEMBERS---CNT_CHILDREN (**88.55 %**)
- REG_CITY_NOT_WORK_CITY---LIVE_CITY_NOT_WORK_CITY (**77.85 %**)
- AMT_ANNUITY---AMT_GOODS_PRICE (**75.23 %**)
- AMT_ANNUITY---AMT_CREDIT (**75.22 %**)
- Age---DAYS_EMPLOYED (**57.53 %**)
- DAYS_BIRTH---DAYS_EMPLOYED (**57.51 %**)
- REGION_POPULATION_RELATIVE---REGION_RATING_CLIENT_W_CITY (**44.70 %**)



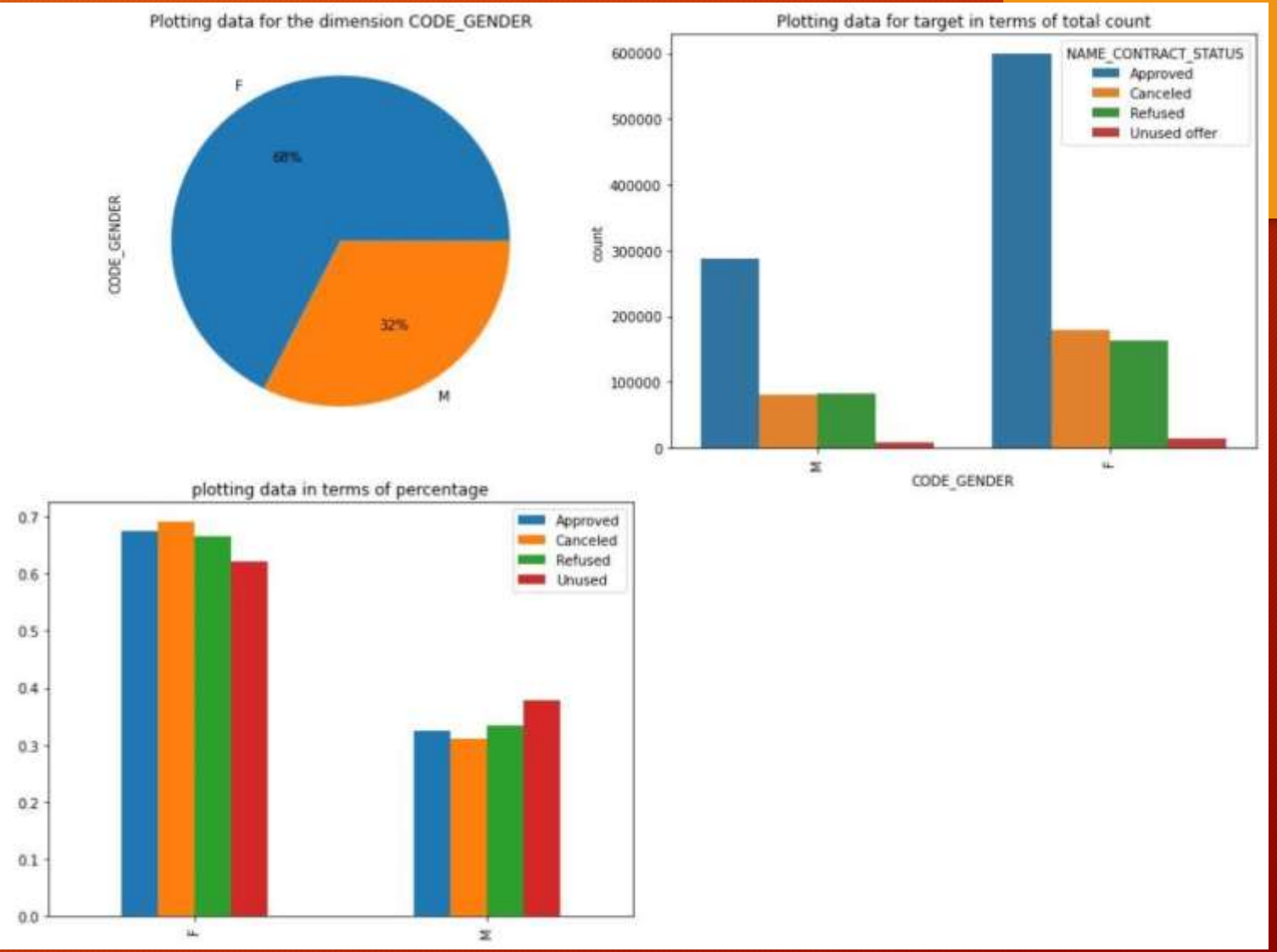
Merged Data Analysis

Across complete loan data, Approved loans outnumbered Canceled, Refused and Unused offer.



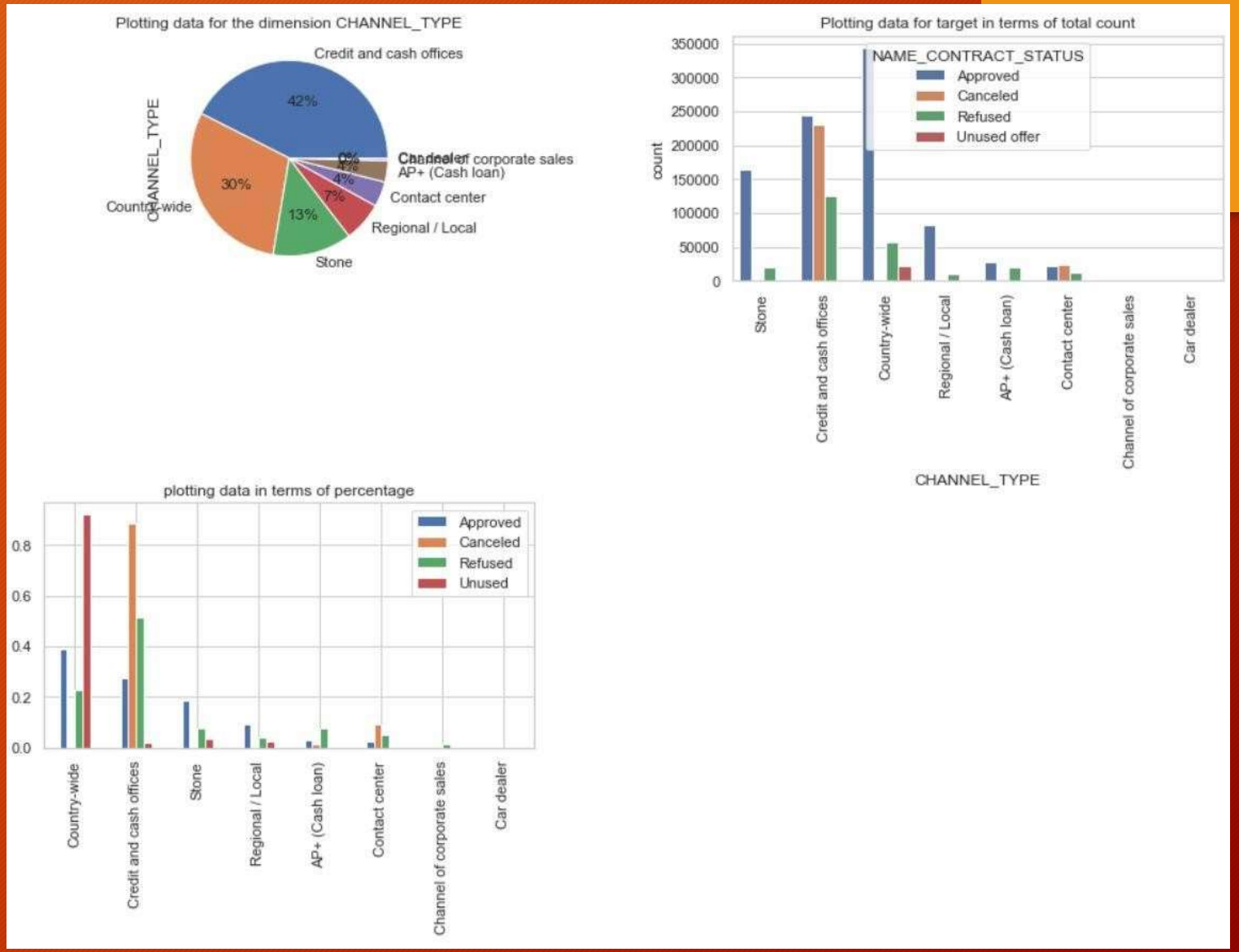
Overall, female seems to have more number of application and also the Approval percentage is high for them.

Even, the percentage for Canceled, refused and Unused status are on higher side for female as compared to male.

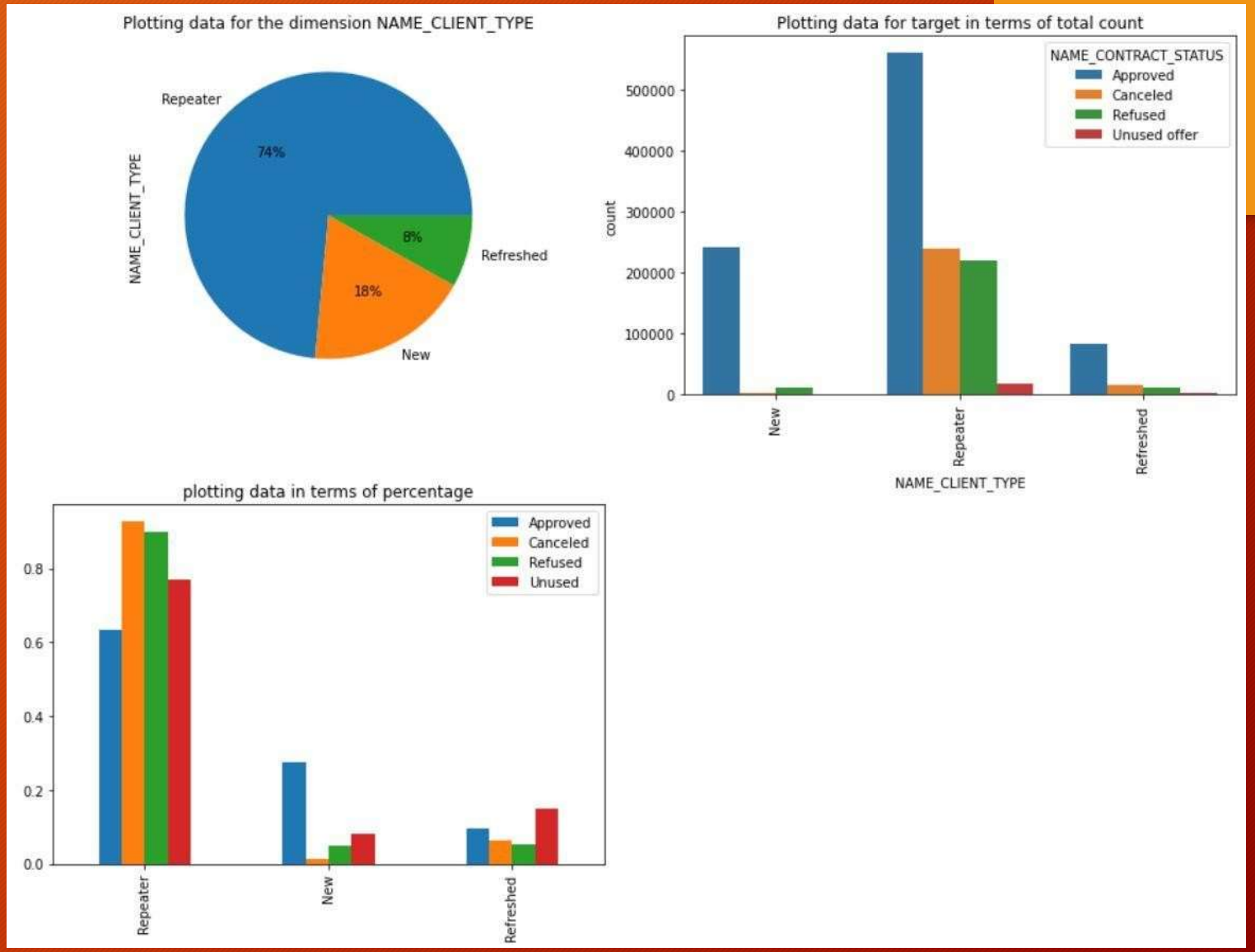


Approved loan percentage is more in Country-wide channel type.

Unused loan and Canceled loan percentage is highest among Country-wide and Credit and cash offices respectively.

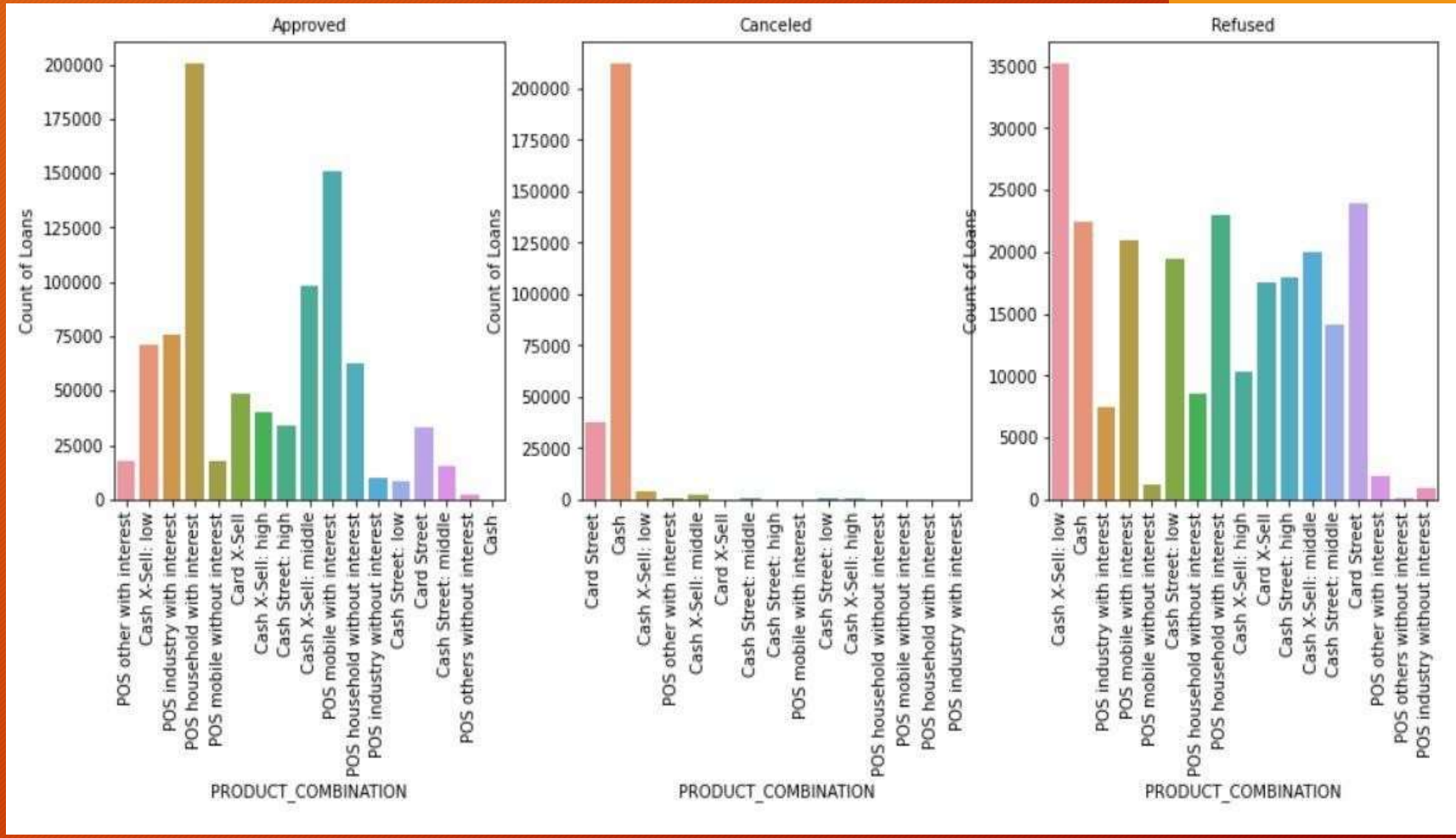


Repeater client type seems to have more percentage of Canceled and Refused loan.



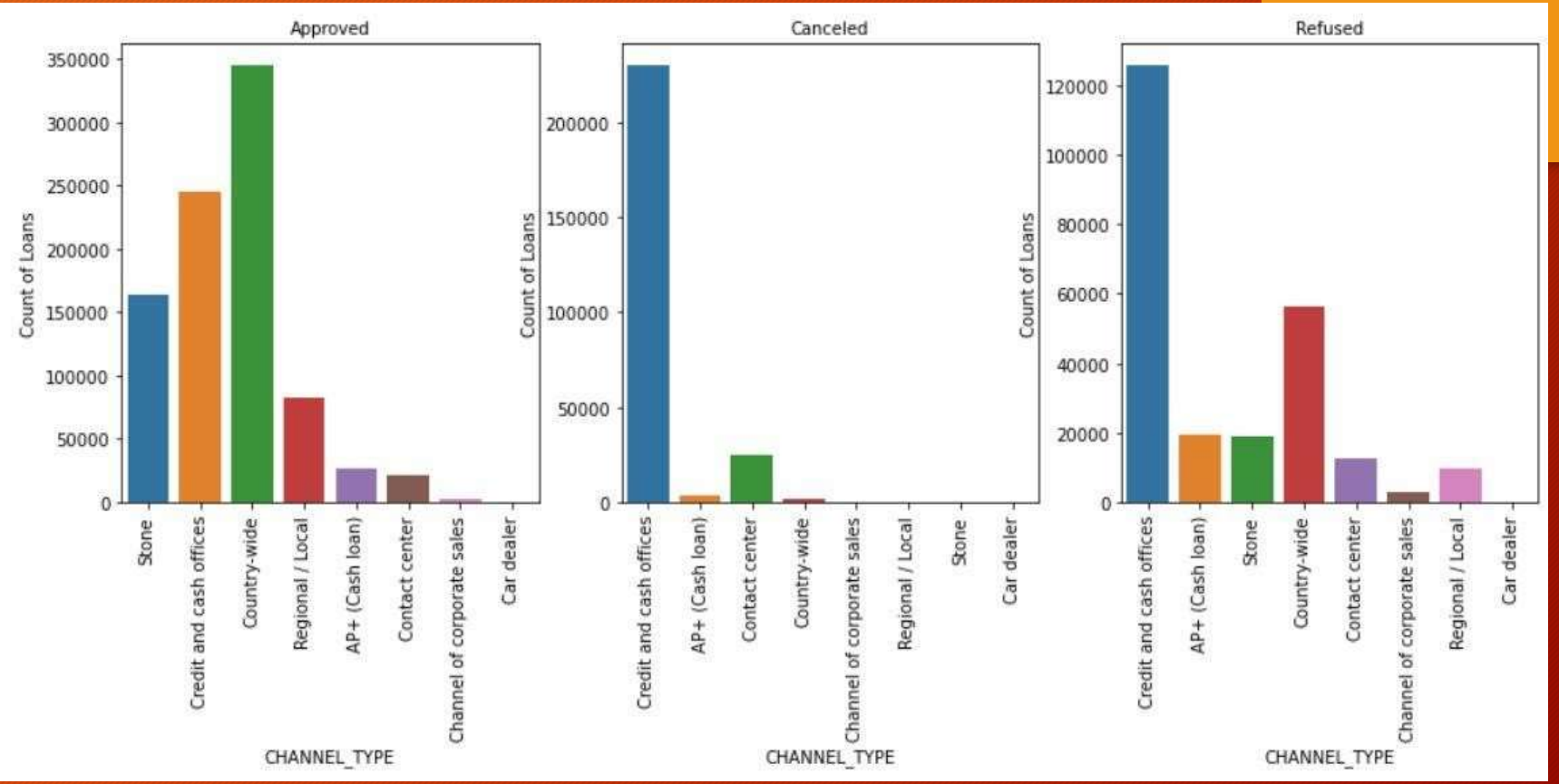
POS household with interest product combination has more number of Approved loans.

Whereas Cash product combination has more no of Cancelled loans. And, Cash X-Sell: low sees more no of Refused loans.



Country-wide channel type sees more no. of Approved loans.

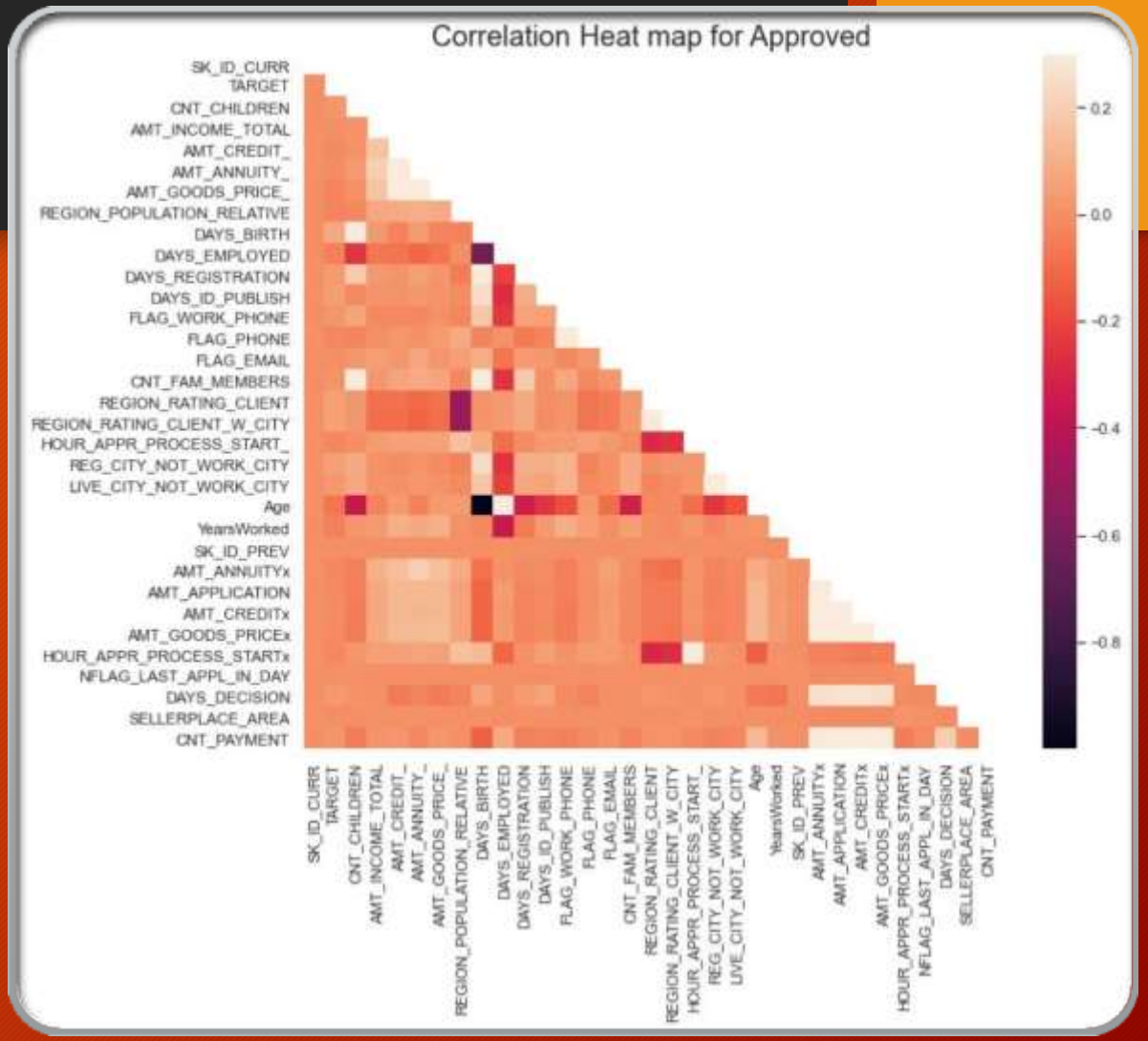
Whereas, Credit and cash offices channel type sees more number of Canceled and Refused loans.



Correlation for approved loans:

Top 10 correlation (Positive and Negative)

- DAYS_BIRTH---Age (**99.97 %**)
- AMT_CREDITx---AMT_GOODS _PRICEx (**99.33 %**)
- AMT_GOODS_PRICE_---AMT_CREDIT_ (**98.65 %**)
- AMT_CREDITx---AMT_APPLICATION (**96.18 %**)
- REGION_RATING_CLIENT---REGION_RATING_CLIENT_W_CITY (**94.26 %**)
- CNT_FAM_MEMBERS---CNT_CHILDREN (**88.29 %**)
- REG_CITY_NOT_WORK_CITY---LIVE_CITY_NOT_WORK_CITY (**83.48 %**)
- AMT_GOODS_PRICEx---AMT_ANNUITYx (**83.13 %**)
- AMT_CREDITx---AMT_ANNUITYx (**82.65 %**)
- AMT_ANNUITYx---AMT_APPLICA TION (**81.45 %**)



Key Indicators of Default

- Previously refused loans
- Client’s age
- Client’s family status
- Contract Type
- Client’s Gender
- Client’s Education level
- Loan Amount

Recommendations

- Banks should focus on the client from age group of 41 to 70 as they have more probability of being financial stable and shows less paying difficulties.
- 20-30 age group face much difficulty in repaying,so can be avoided.
- Banks should focus more on education type 'Higher education' and avoid Secondary/secondary special, incomplete higher or lower secondary as they have more difficulty in repaying.
- Avoid income type 'Working' clients as they have high percentage of paying difficulties. Instead focus on Commercial associates, pensioner and State servant.
- Revolving loans have a better chance to be returned.
- Provide more loans to females,as they have less difficulty in returning the loan than men.
- Give preference to widowed & married people.
- Car, property owners find it less difficult to repay the loan,they must be preferred.
- Focus on clients from housing type ‘House/apartment’ as they are having less paying difficulties.
- Bank should focus 'Country-wide' channel type sees more no of Approved loans. Whereas, Credit and cash offices channel type sees more number of Canceled and Refused loans.

Conclusion

After a methodical Exploratory Data Analysis for the bank,it can be concluded that the bank should adhere to the recommendations and keep an eye at the key indicators. This way the bank will stay away from probable defaulters and can net financially safe potential customers.

Thank You