CREDIT EDA CASE STUDY

- Talloji Kishore

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 :

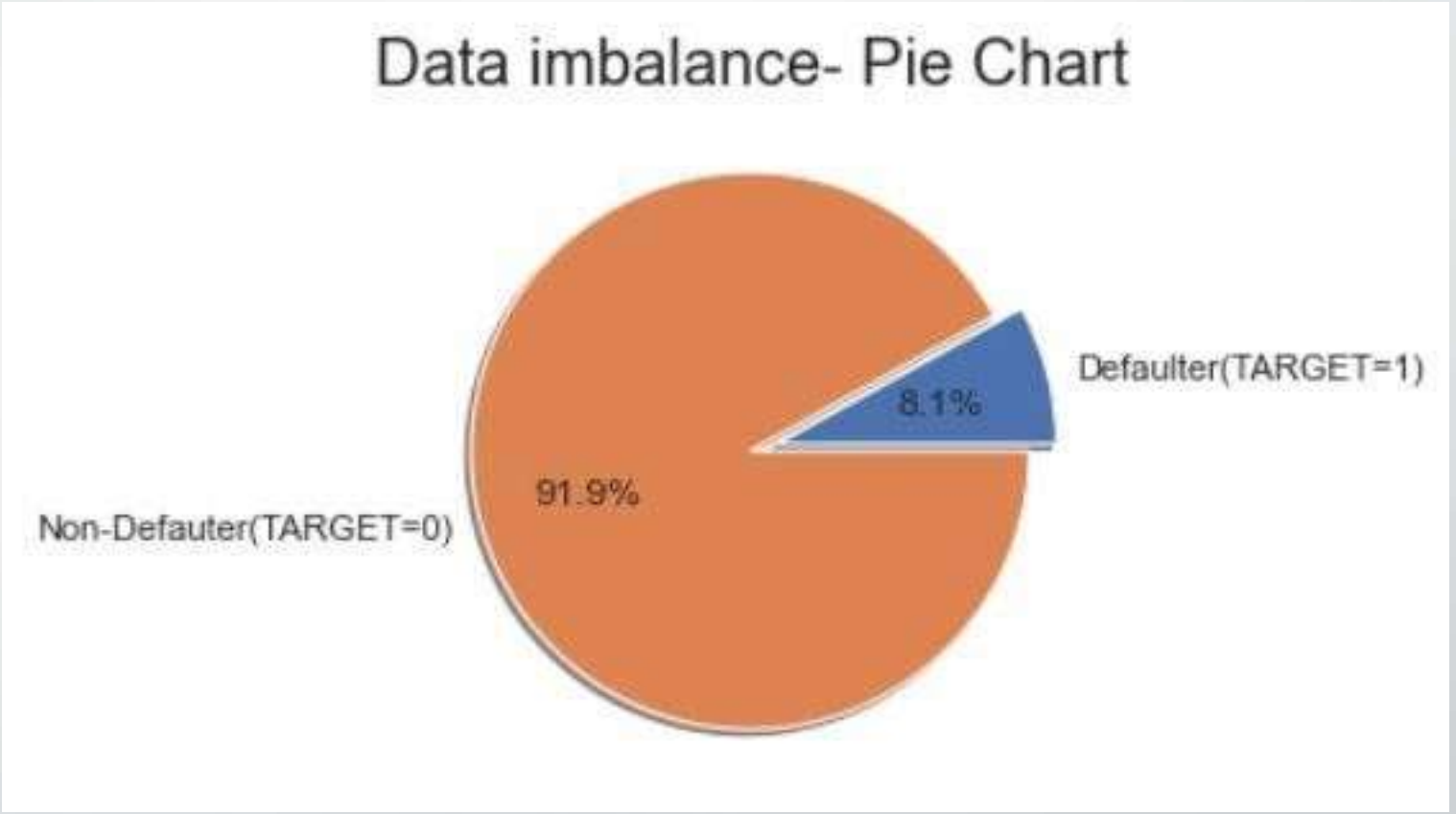
- The required libraries needed for data cleansing and visualisation are imported.
- We have done the data cleansing for columns wherever necessary and dropped the columns with majority of data as NA. Outliers are identified and handled wherever possible. Data imbalance is checked.
- Created new columns as per the requirements
- Univariate/Bivariate Analysis of the relevant Categorical/numerical is done and insights are derived
- Current and Previous application data is done to derive insights based on bank Approval loan status .

DATA IMBALANCE

CLEARLY WHEN THE DATA HAS BEEN SEPARATED TO 2 COMPONENTS W.R.T TARGET VARIABLE THERE WAS A DATA IMBALANCE

Ratio of Data Imbalance

11.39

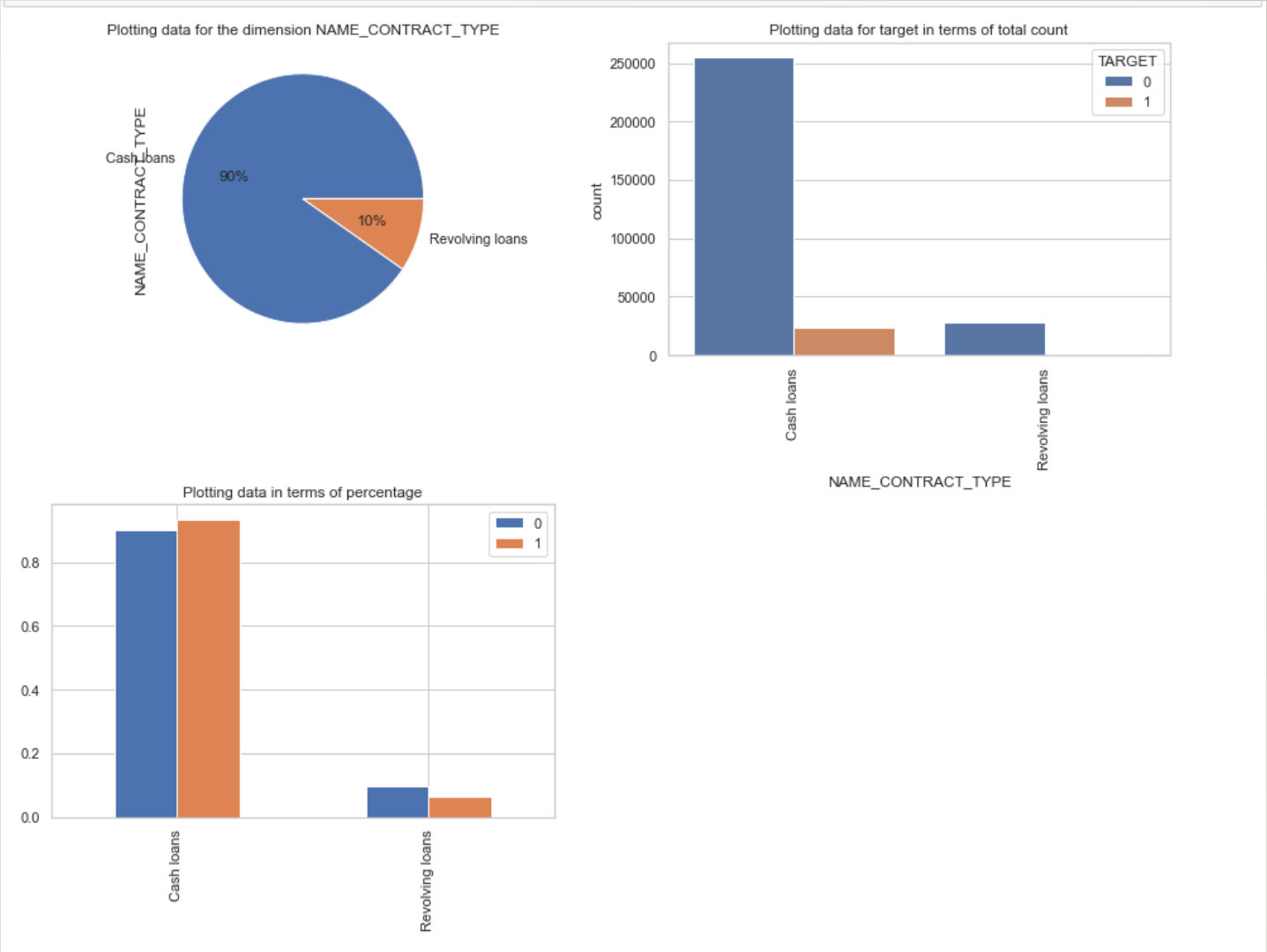


UNIVARIATE/BIVARIATE ANALYSIS OF CATEGORICAL COLUMNS

Bank is primarily providing 2 types of loans : Cash & Revolving . Majority of the loans are cash loans

Analysis w.r.t Target Variable :

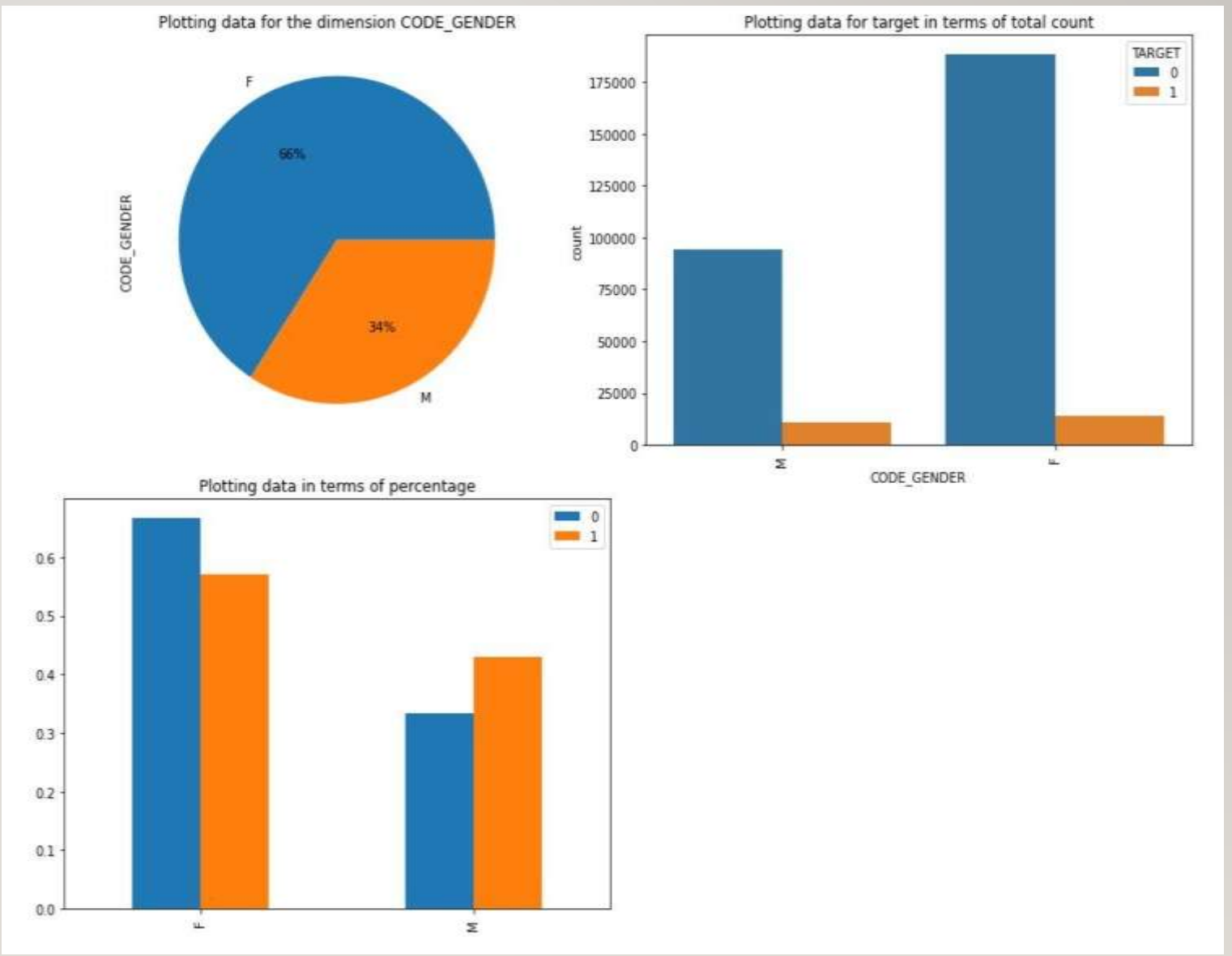
Comparatively Cash loans taken by clients are facing bit more difficulties than revolving loans



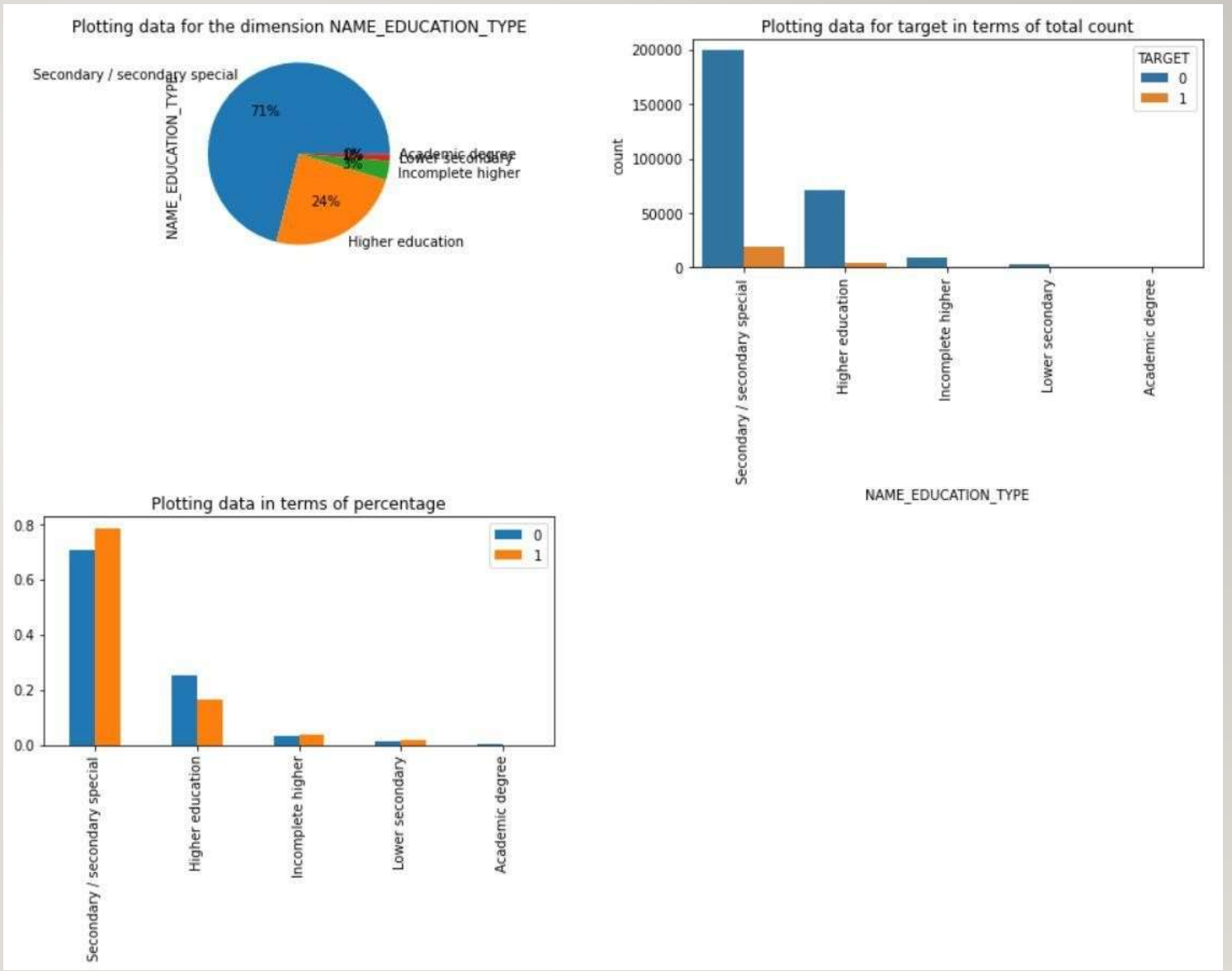
Clearly Females are opting for more loans than males

Analysis w.r.t Target Variable :

Males are having more payment difficulties than compare to Females



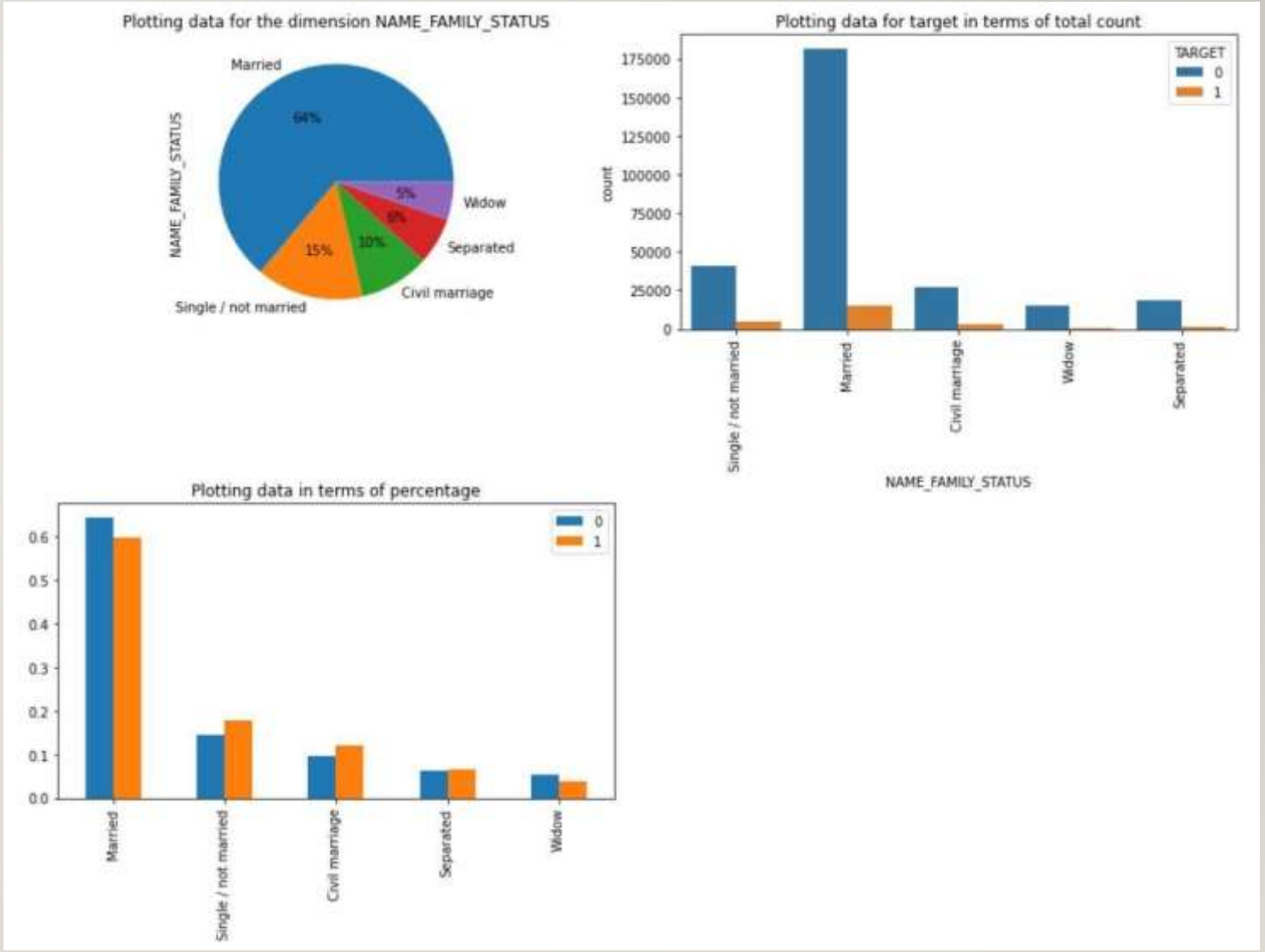
Most loans are provided to Secondary/Secondary Special and Higher education people.



Analysis w.r.t Target Variable :

- Higher education people have less paying difficulties as compared to Secondary/secondary special.
- A valid insight as better education gives better jobs and standard of living

More than half of the loans are being taken by married people.



Analysis w.r.t Target Variable :

Single/not-married people facing more paying difficulties when compared to Married people.

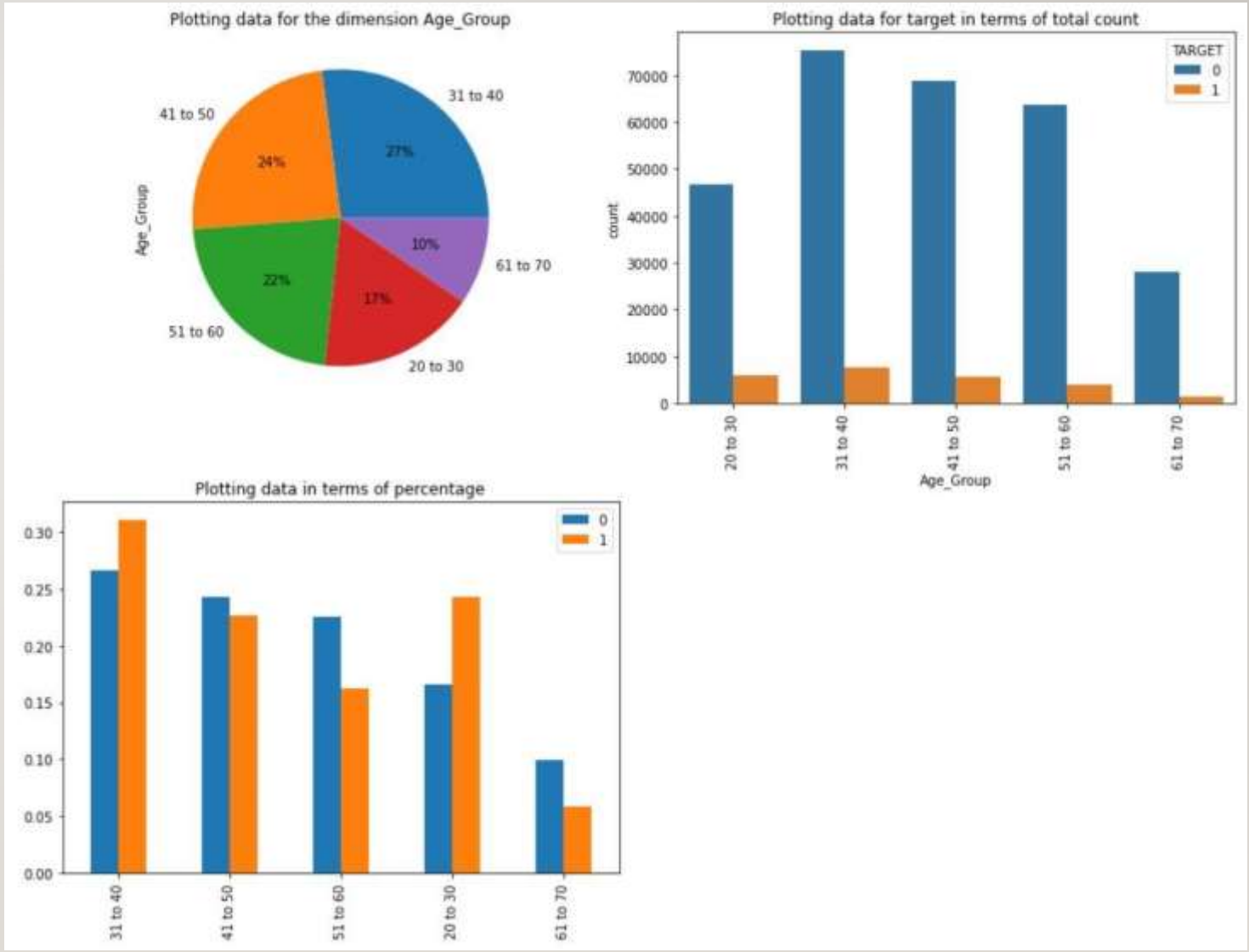
This could be due to married people may have dual source of income.

Apart from young age and old age ,almost all Age Groups are having same % of loans being taken

Old age people definitely do not opt for taking loans as its difficult to repay with their savings and retiral funds

Analysis w.r.t Target Variable :

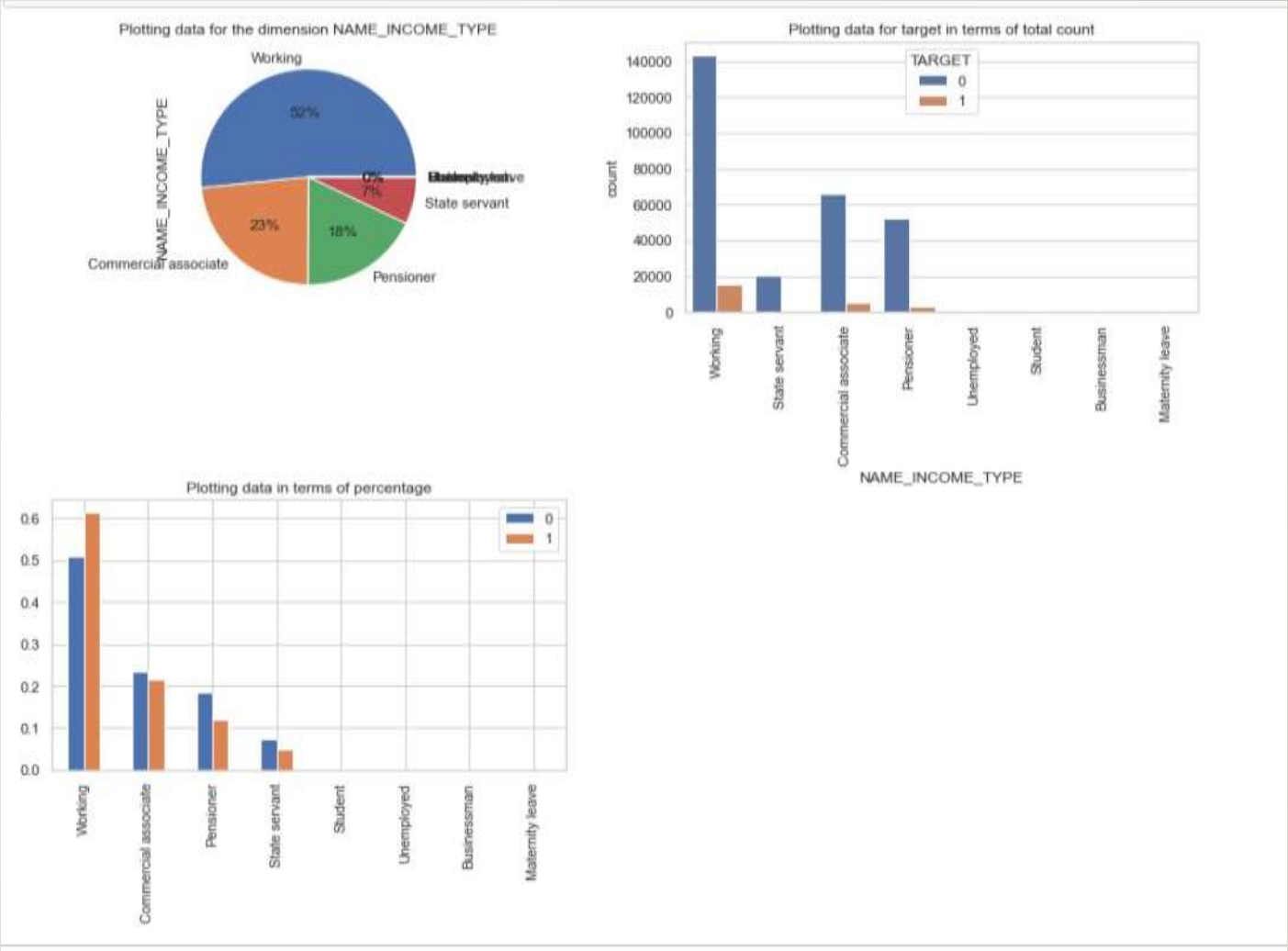
Old age/Middle age people are having less payment difficulties i.e. in age group between 41 to 70 Whereas Age group between 20 to 30 are facing payment difficulties.



Half of the loans applied by clients are from working people

Analysis w.r.t Target Variable :

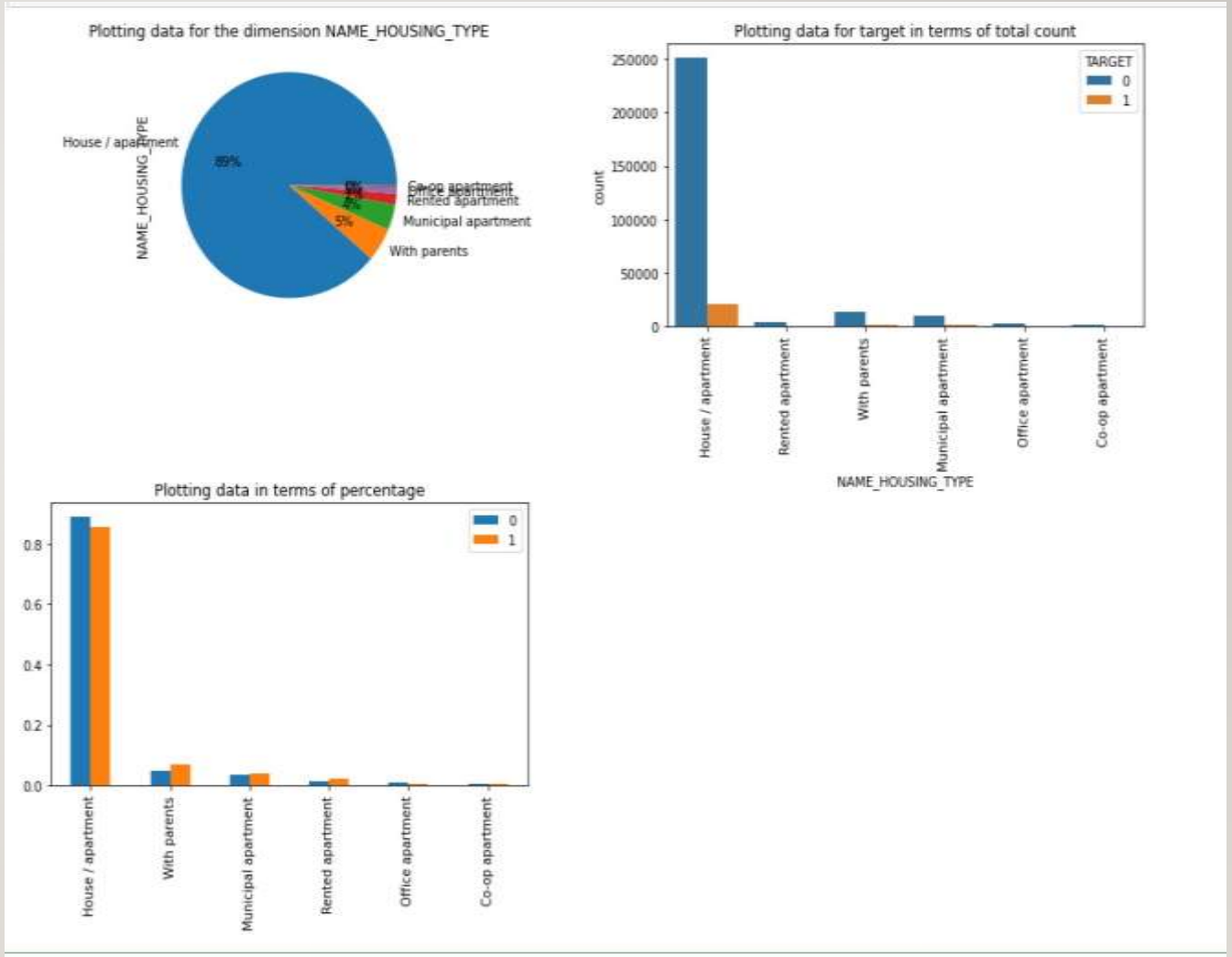
Its clearly depicted that working people are facing more difficulties when compared to other income type guys



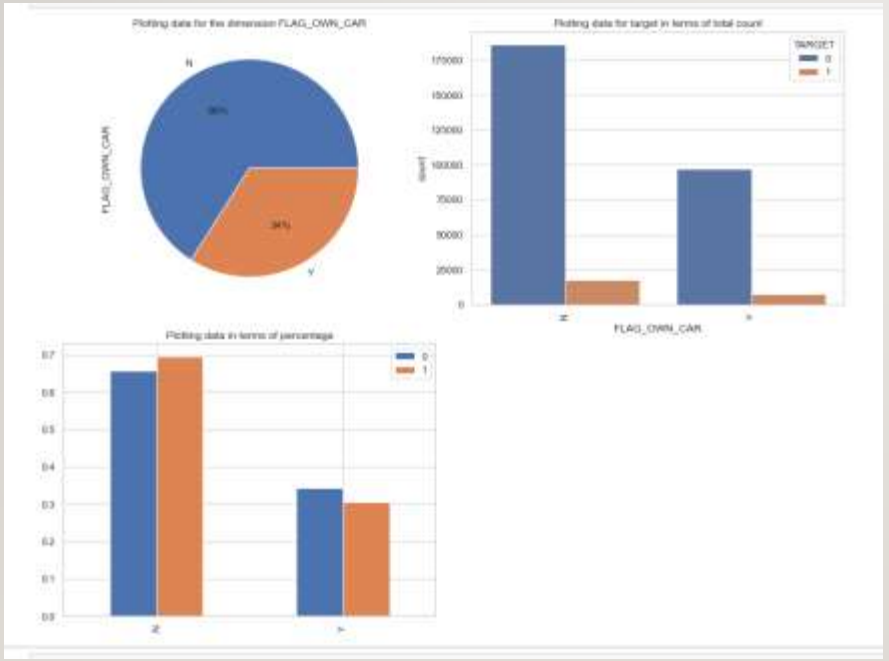
Loans are being taken mostly by people living in apartments or houses

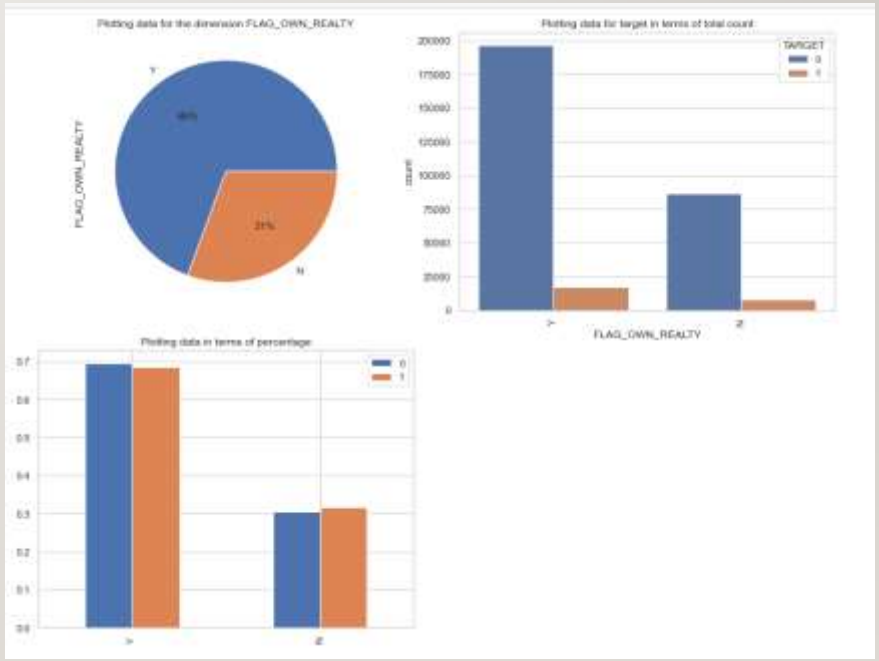
Analysis w.r.t Target Variable :

Clients living with parents are facing bit more difficulties . Probably their income needs to be shared to their medical/other expenses



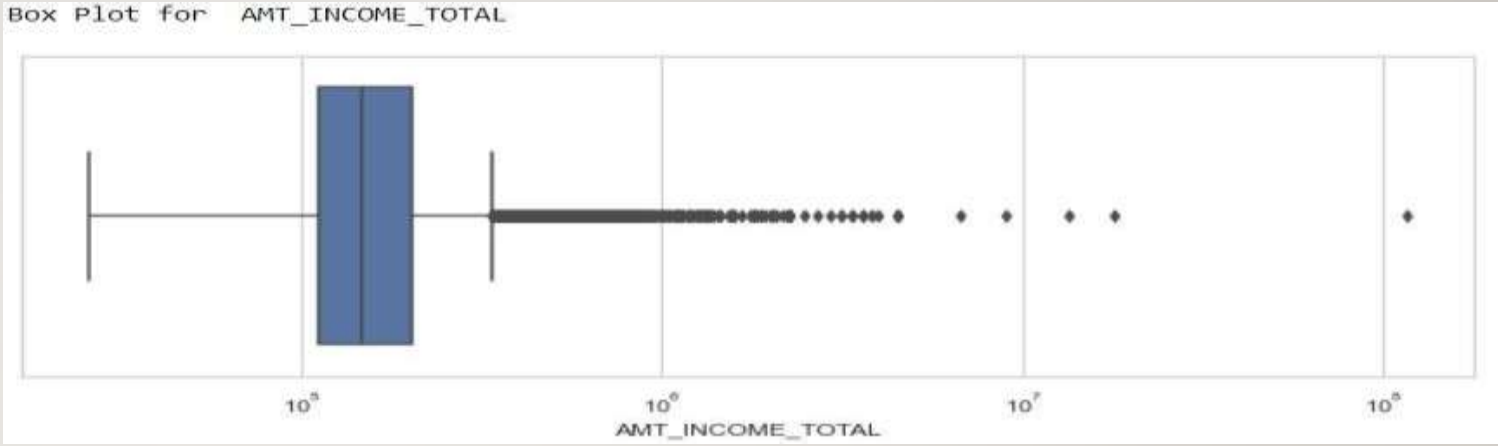
As per below Graphs clients already possessing a car/real estate property are having comparatively less difficulties. It indicates that they are already in a good financial position and able to pay the instalments on time





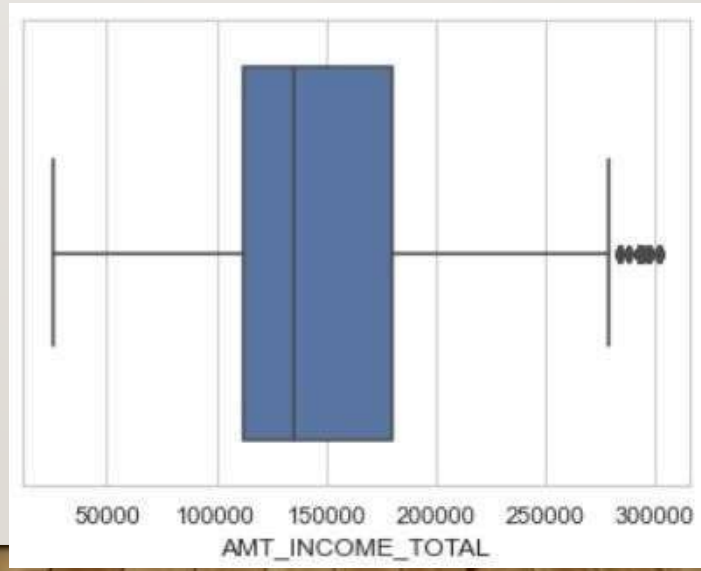
UNIVARIATE/BIVARIATE ANALYSIS AND OUTLIERS FOR NUMERICAL COLUMNS

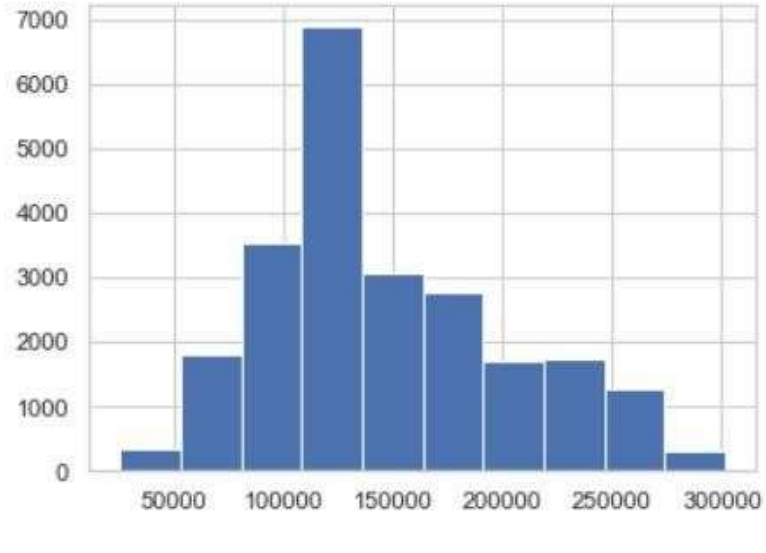
Clearly it depicts the Income column has lot of outliers



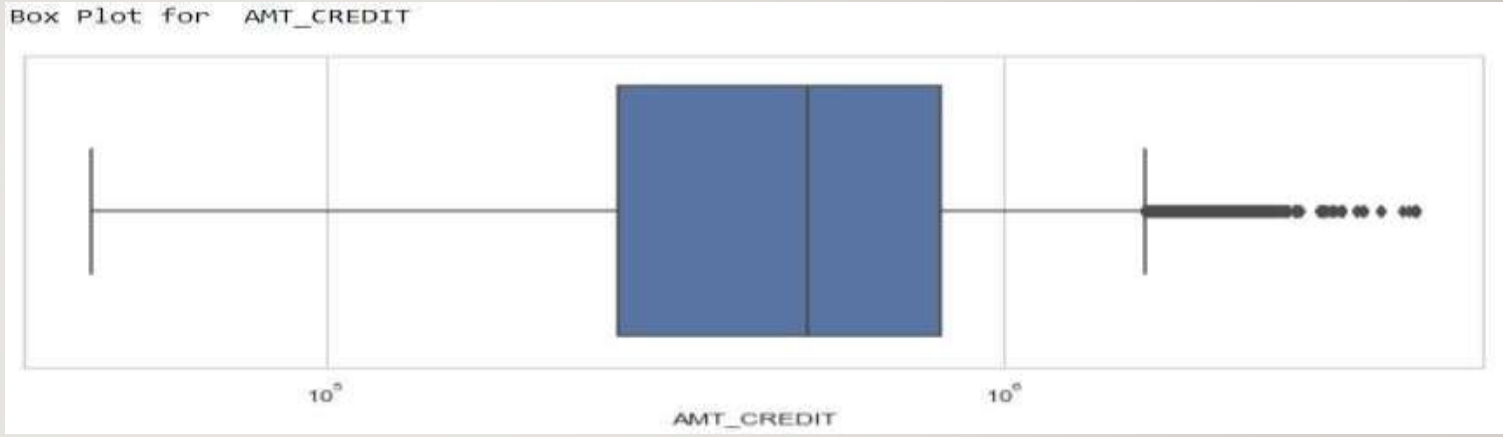
After handling the outliers and histogram is plotted we can depict below:

Clients having income between 100000 to 150000 are the segment with more clients facing payment difficulties



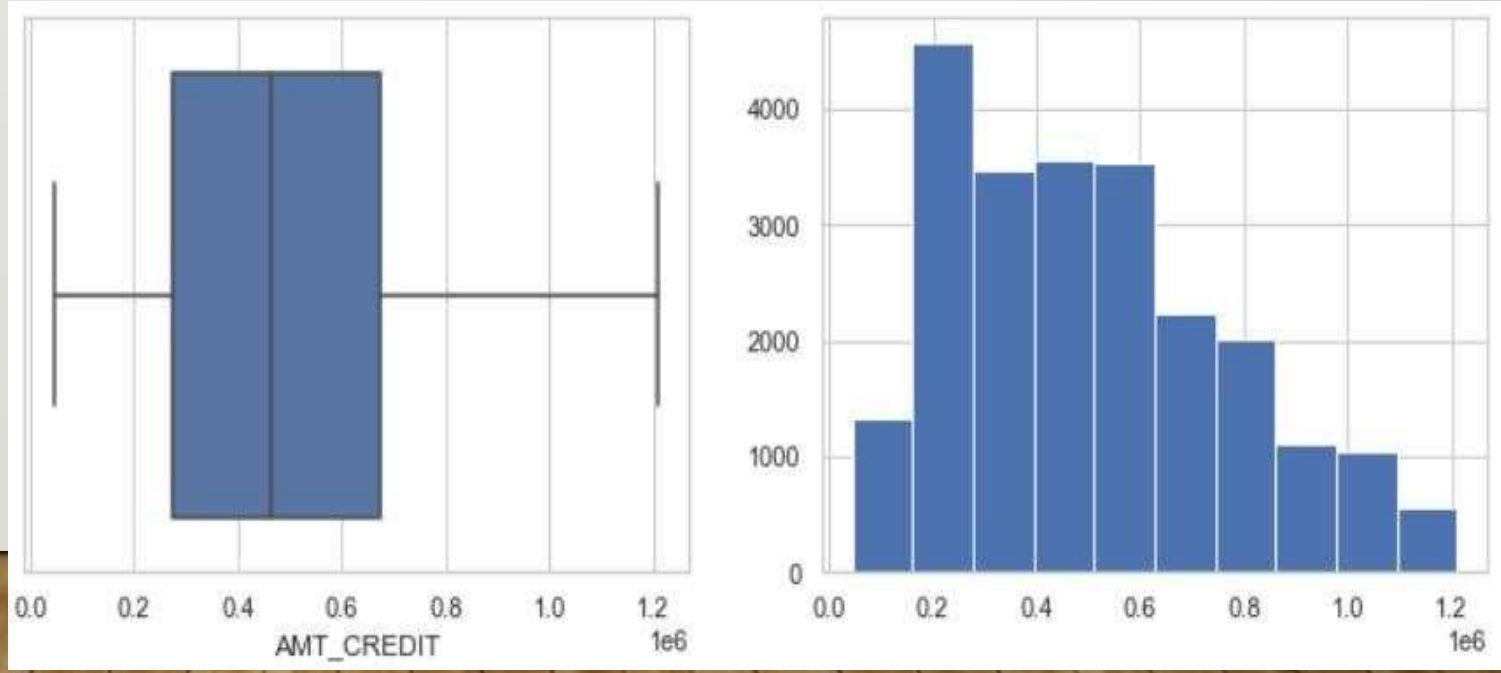


Clearly it depicts the Loan Amount column has lot of outliers



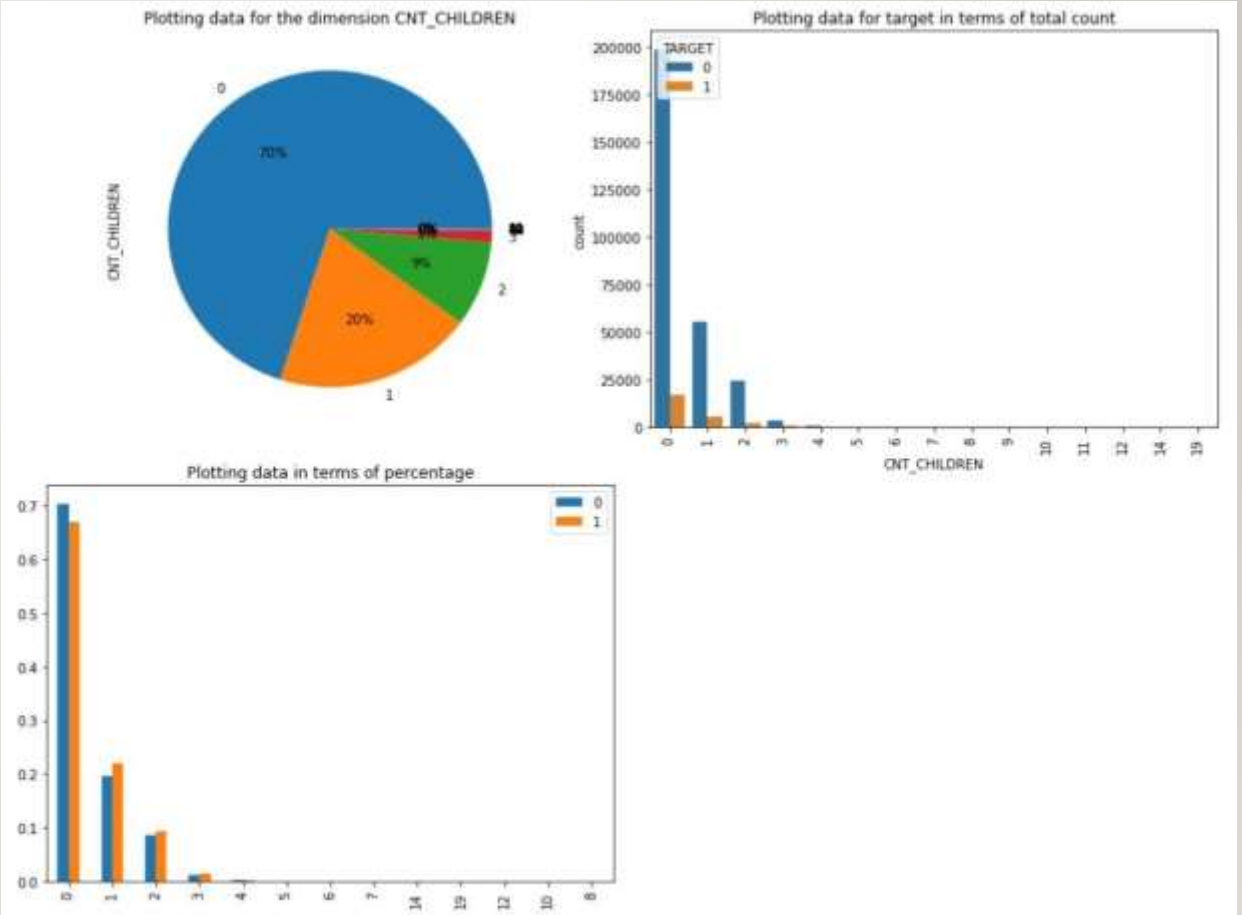
After handling the outliers and histogram is plotted we can depict below:

Around 55 % of the population have amount credit ranges from 200000 to 600000.

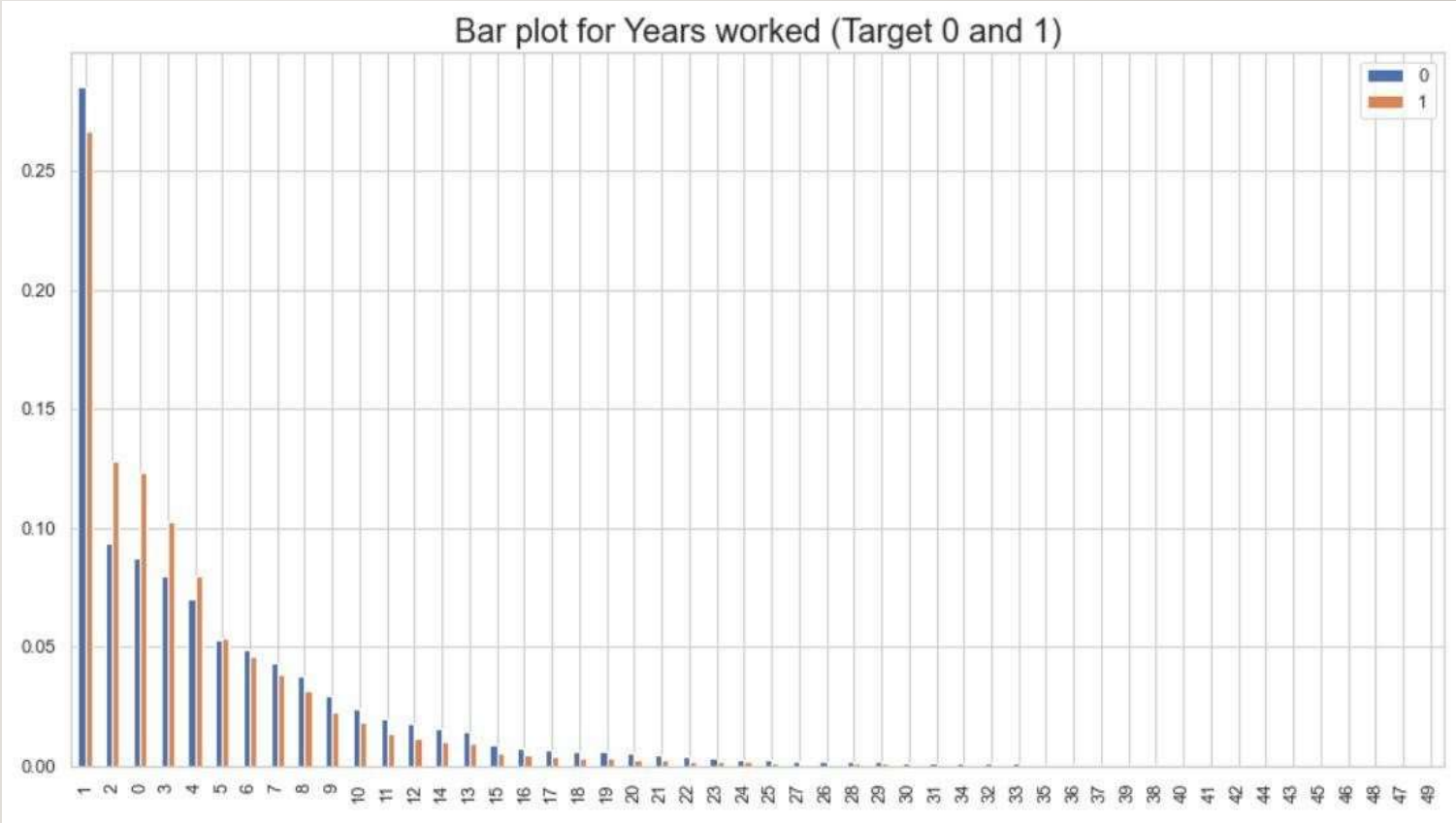


70% of the clients don’t have children

People who has no children don't face paying difficulties. Whereas, we can see difficulties arise as the no of children increases.

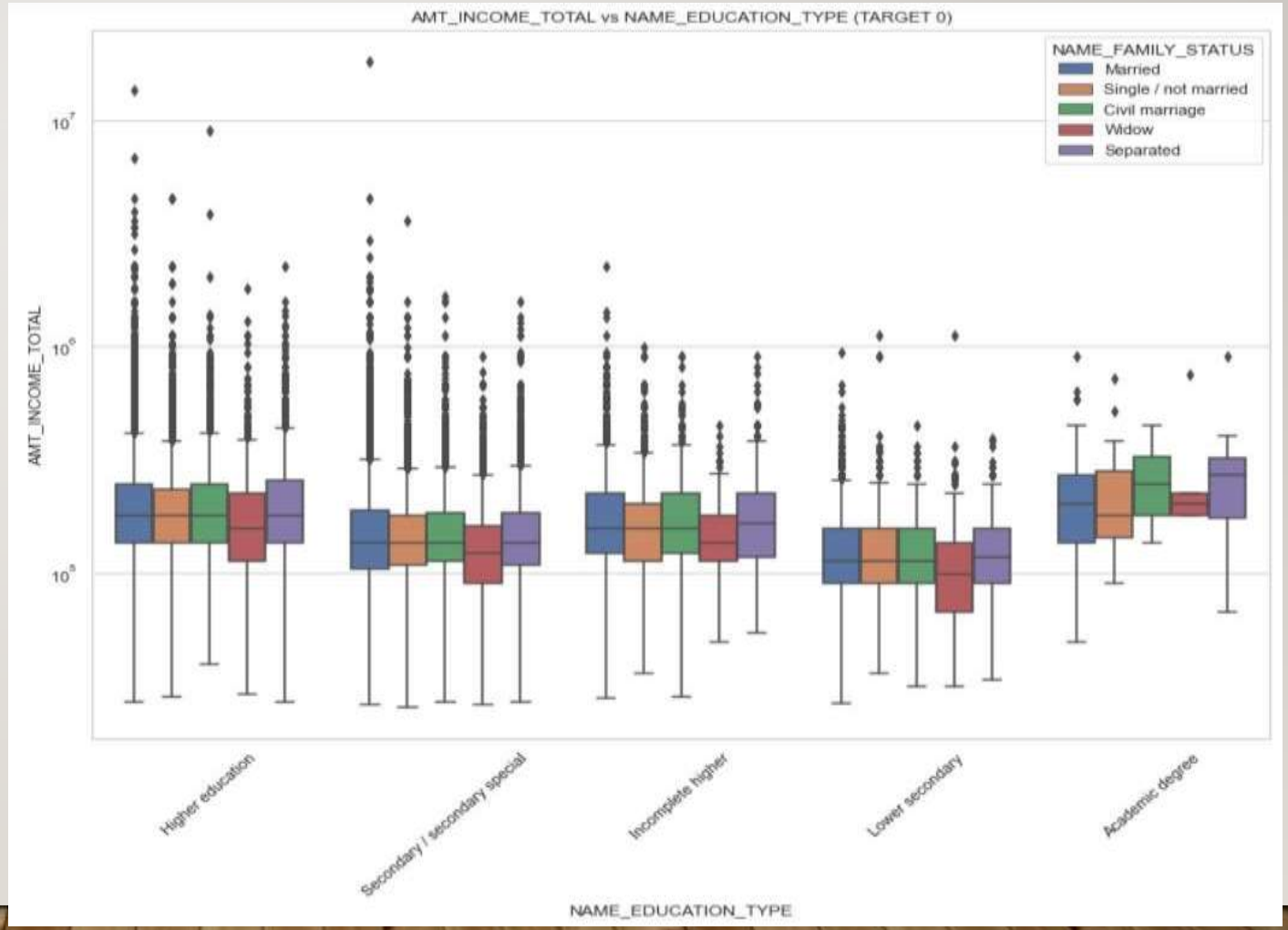


We can see that people in the beginning phase of their career i.e. experience between 1 -5 years face paying difficulties. And the experienced people once well settled and with good salary don't face much paying difficulties.

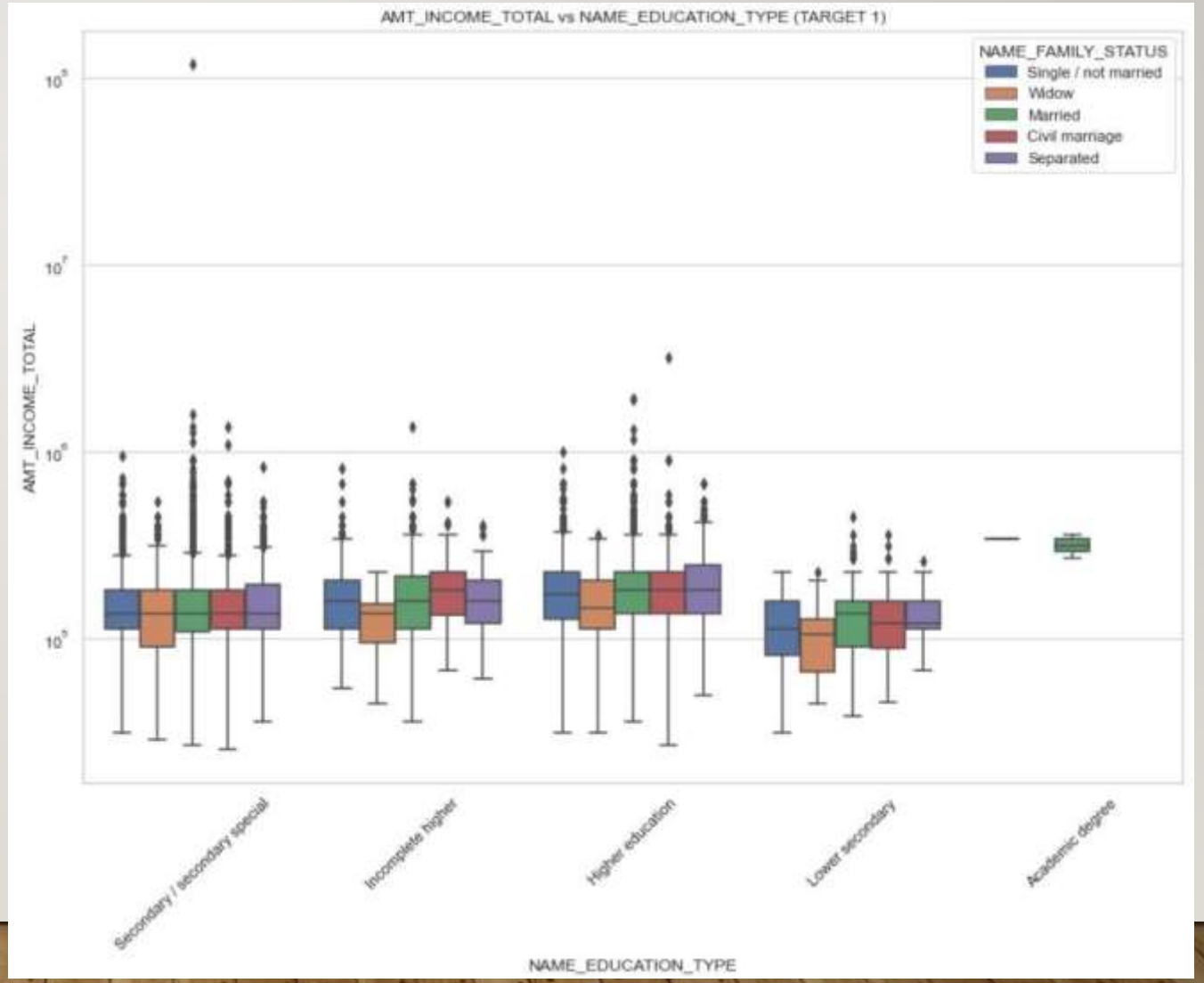


MULTIVARIATE ANALYSIS

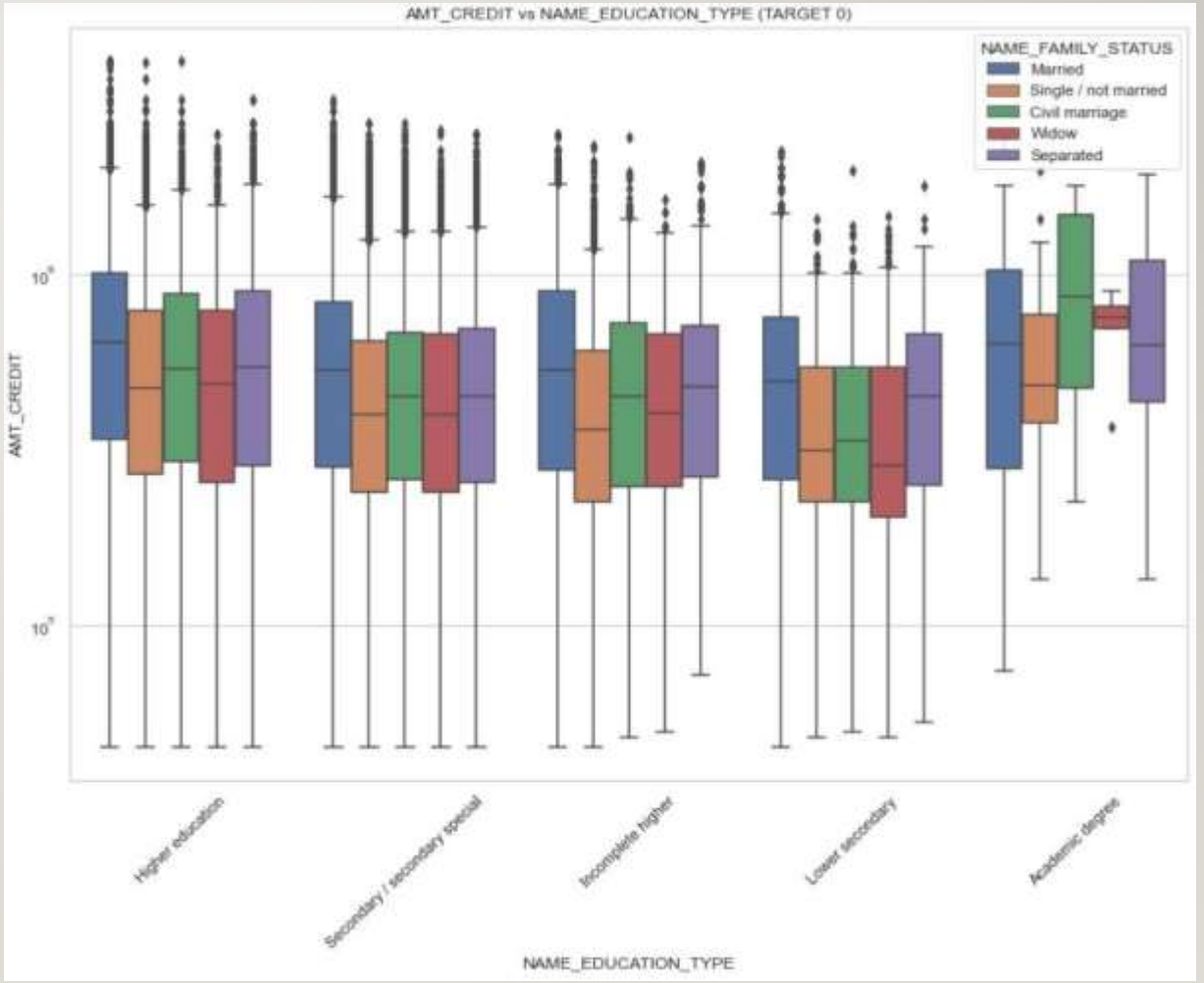
For non-defaulters : Education Type with Higher education and 'Secondary/secondary special' Amount income total is mostly equal among the family status. These two education type has many outliers as well. Academic degree education type have less outliers and also their amount income total seems to be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



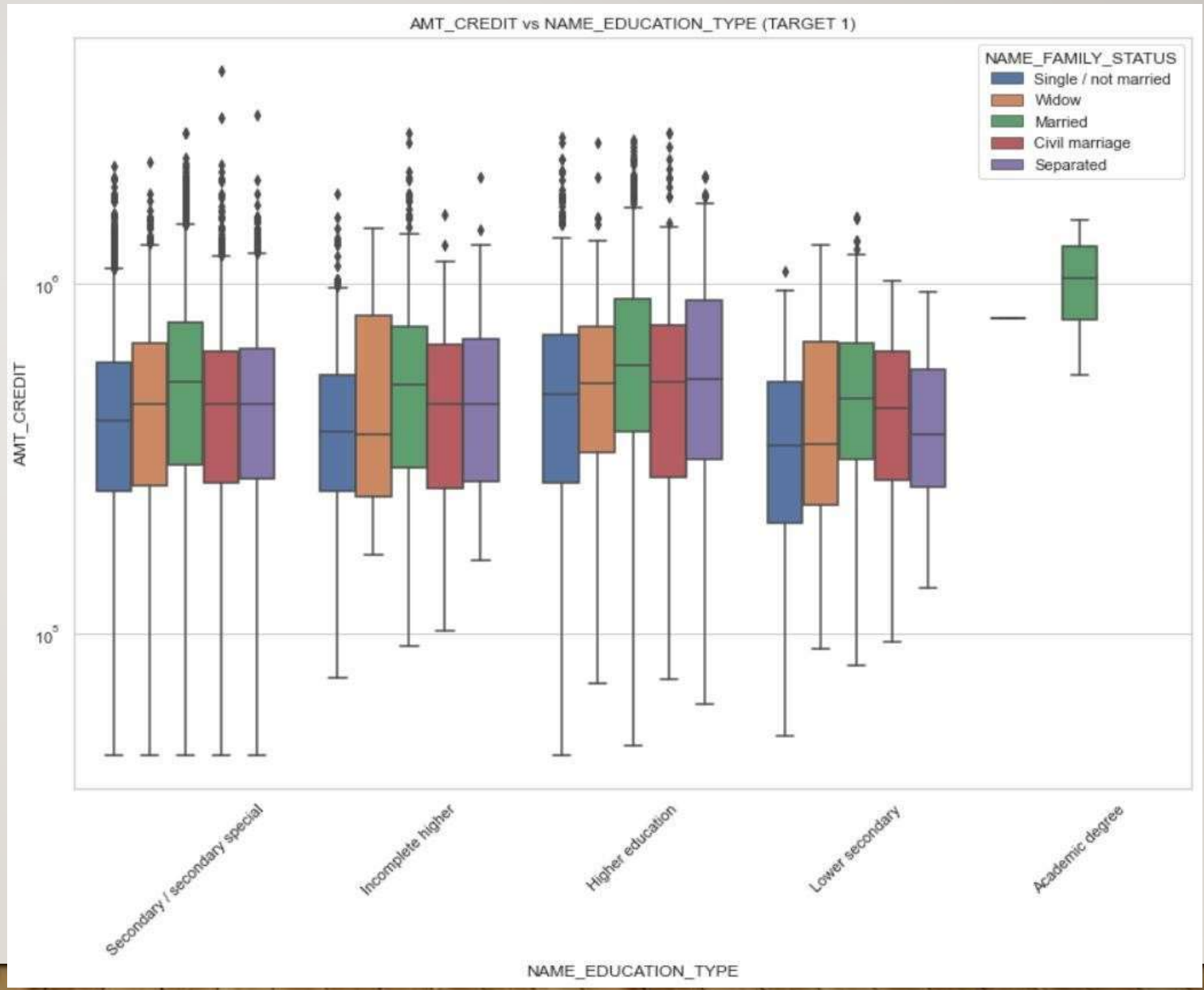
For Defaulters : Most of the outliers are from Education type 'Secondary/secondary special', 'Incomplete higher' and 'Higher education'. Very few outliers observed for Lower secondary and Academic. Single, civil marriage and separated family status group has almost similar income amount under Secondary/secondary 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.

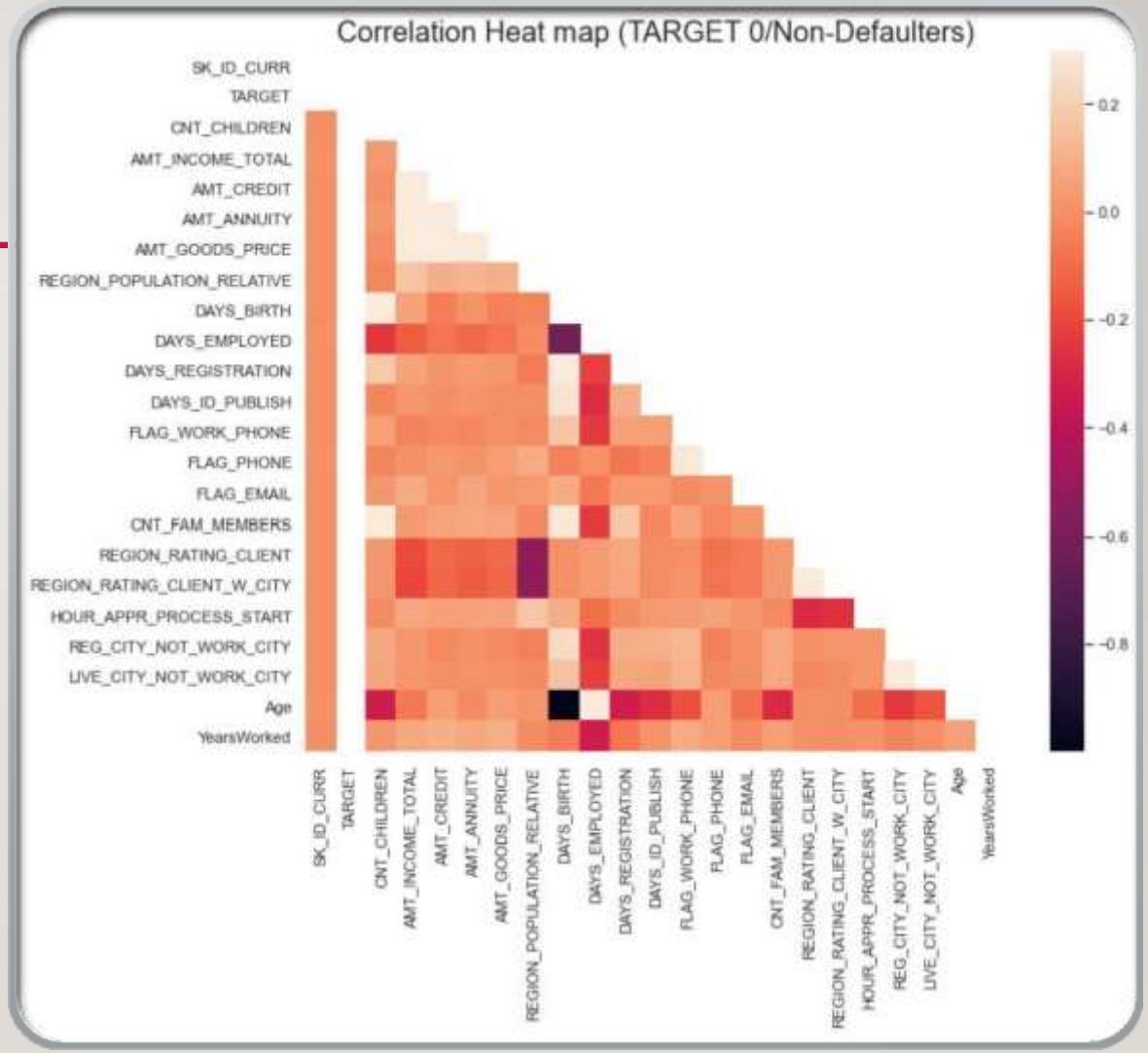


CORRELATION

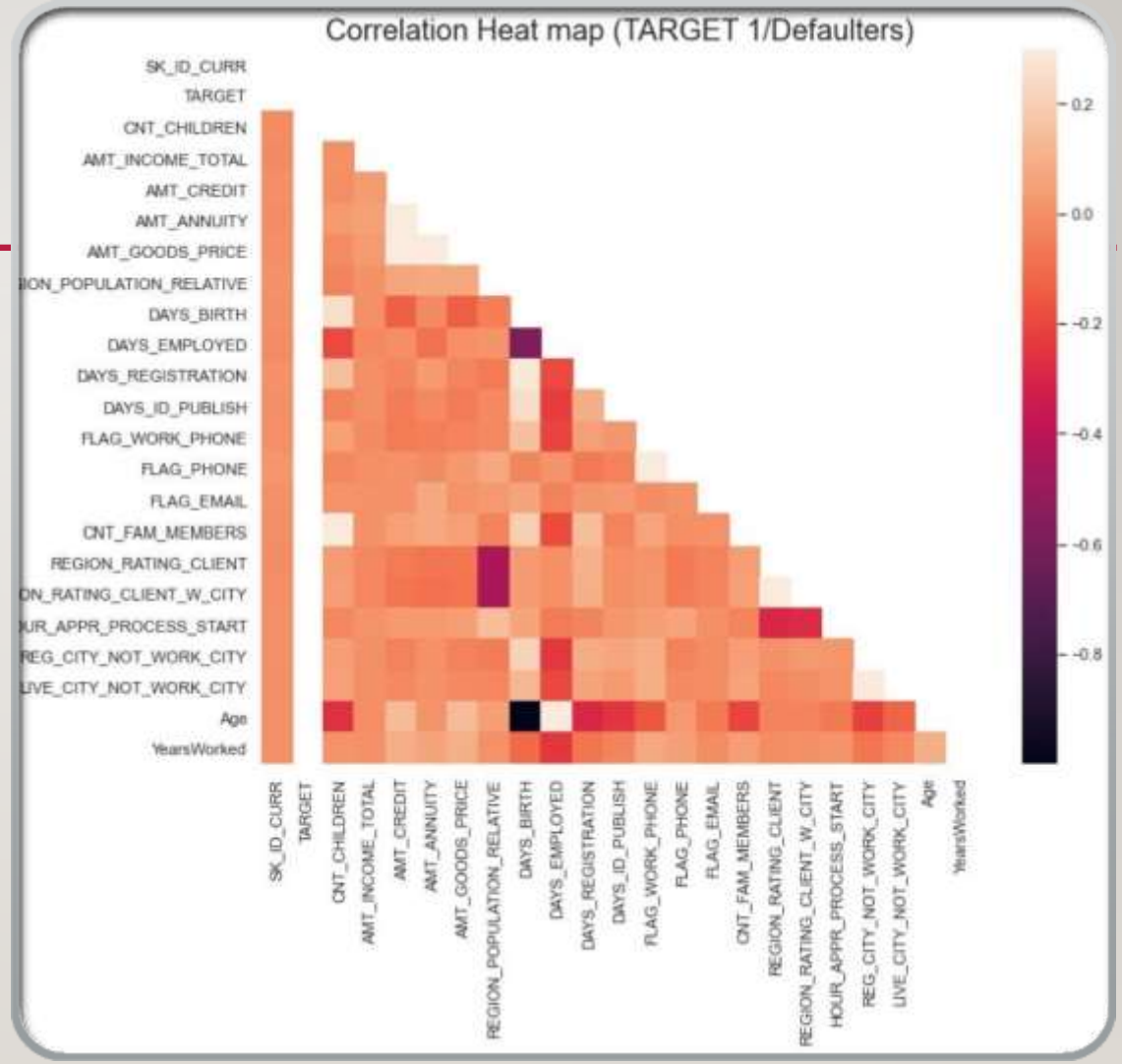
CORRELATION IN NON DEFAULTERS :

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 :

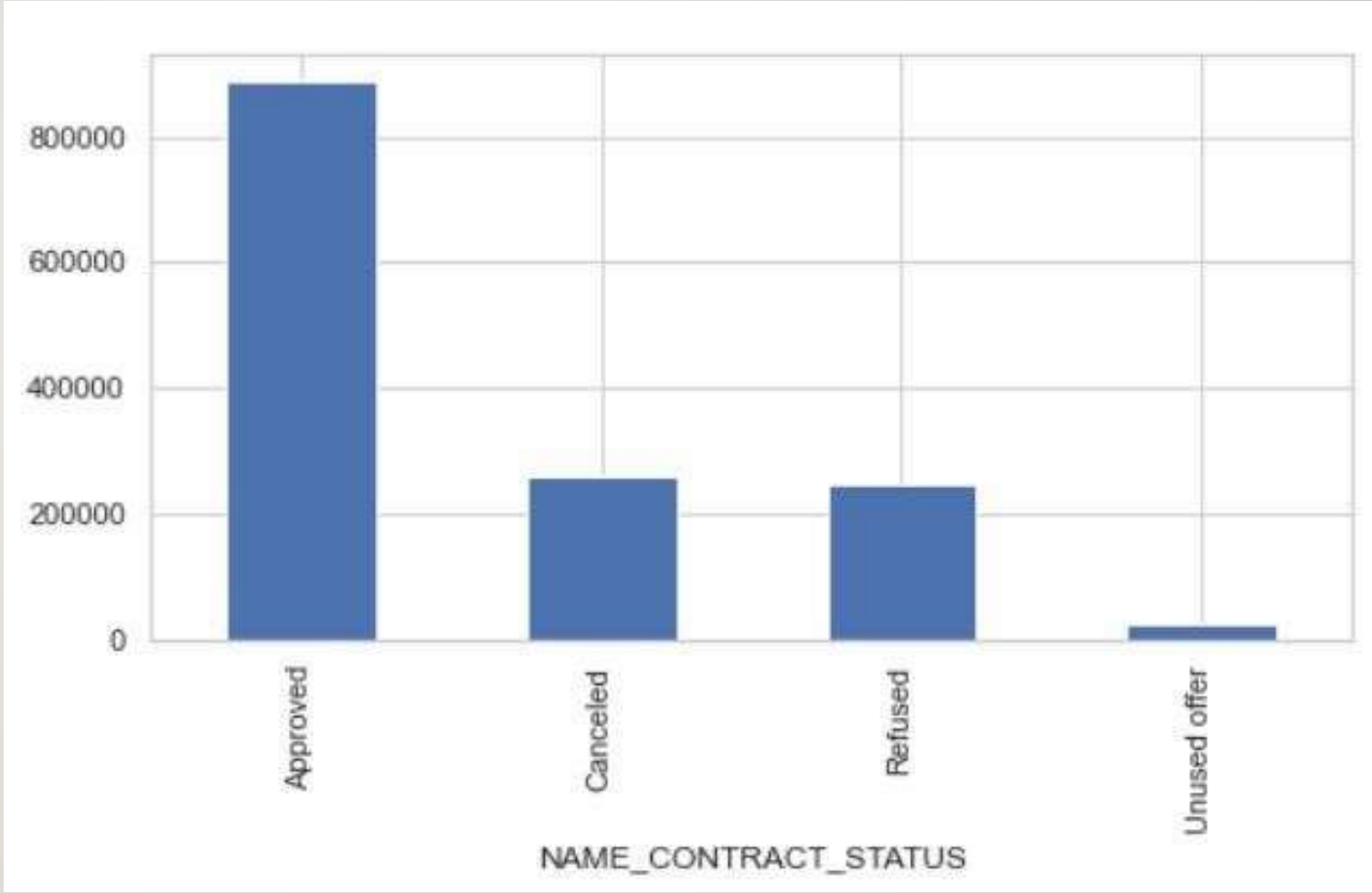


Top 10 correlation (Positive and Negative) :

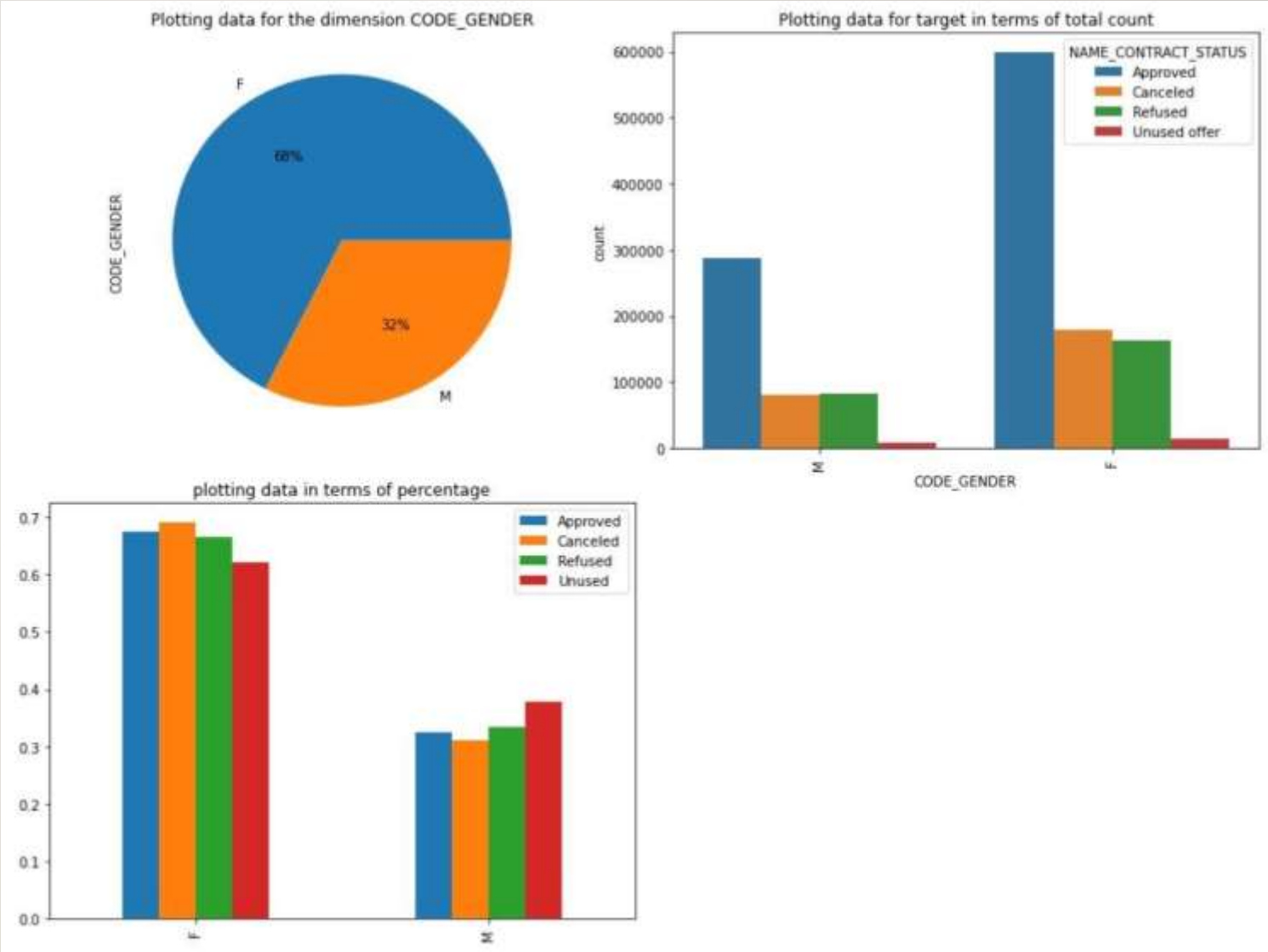
- DAYS_BIRTH---Age (**99.97 %**)
- AMT_CREDIT---AMT_GOODS_PRICE (**98.28 %**)
- REGION_RATING_CLIENT_W_CITY---REGION_RATING_CLIENT (**95.66 %**)
- CNT_FAM_MEMBERS---CNT_CHILDREN (**88.55 %**)
- REG_CITY_NOT_WORK_CITY---LIVE_CITY_NOT_WORK_CITY (**77.85 %**)
- AMT_ANNUITY---AMT_GOODS_PRICE (**75.23 %**)
- 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

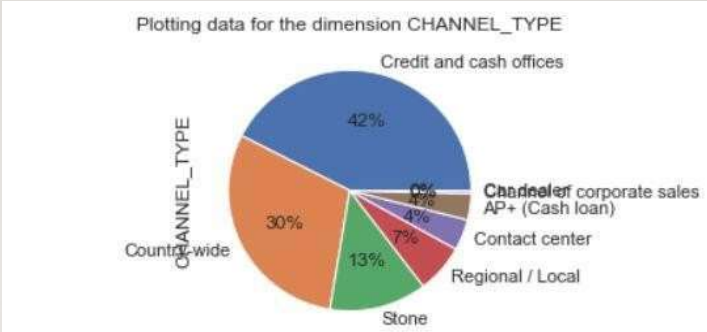
Across complete loan data, Approved loans outnumbered Canceled, Refused and Unused offer.

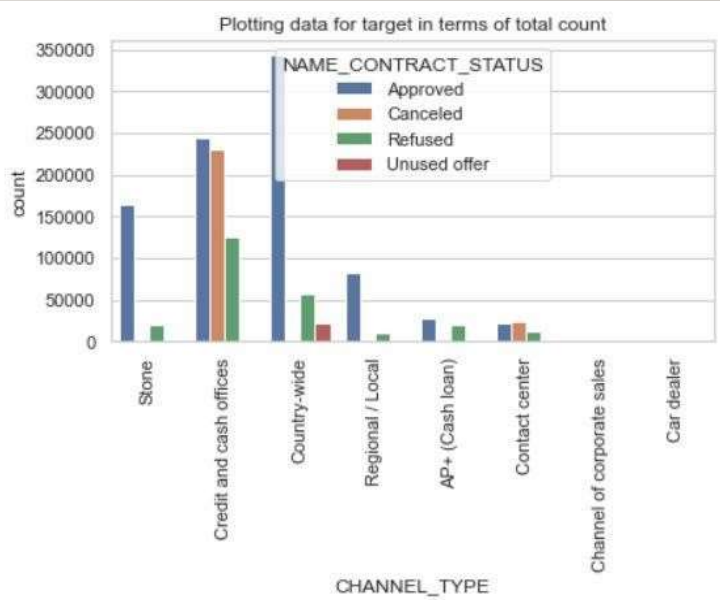


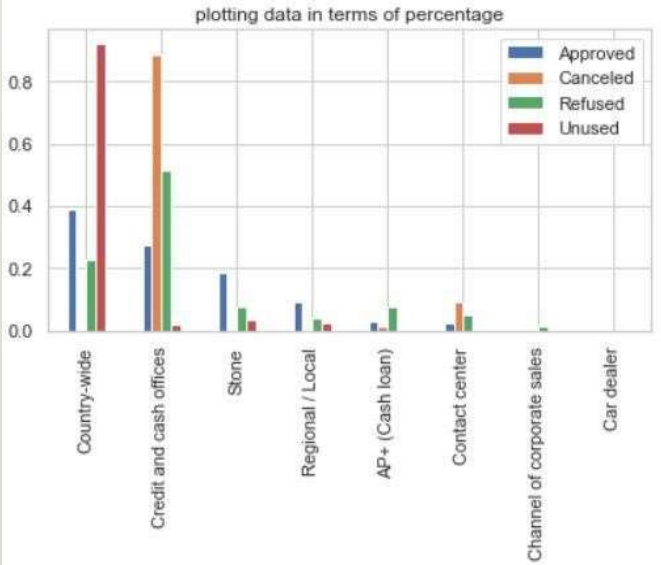
Overall, female seems to have more number of application and also the Approval percentage is high for them. Even, the percentage for Canceled, refused and Unused status are on higher side for female as compared to male.



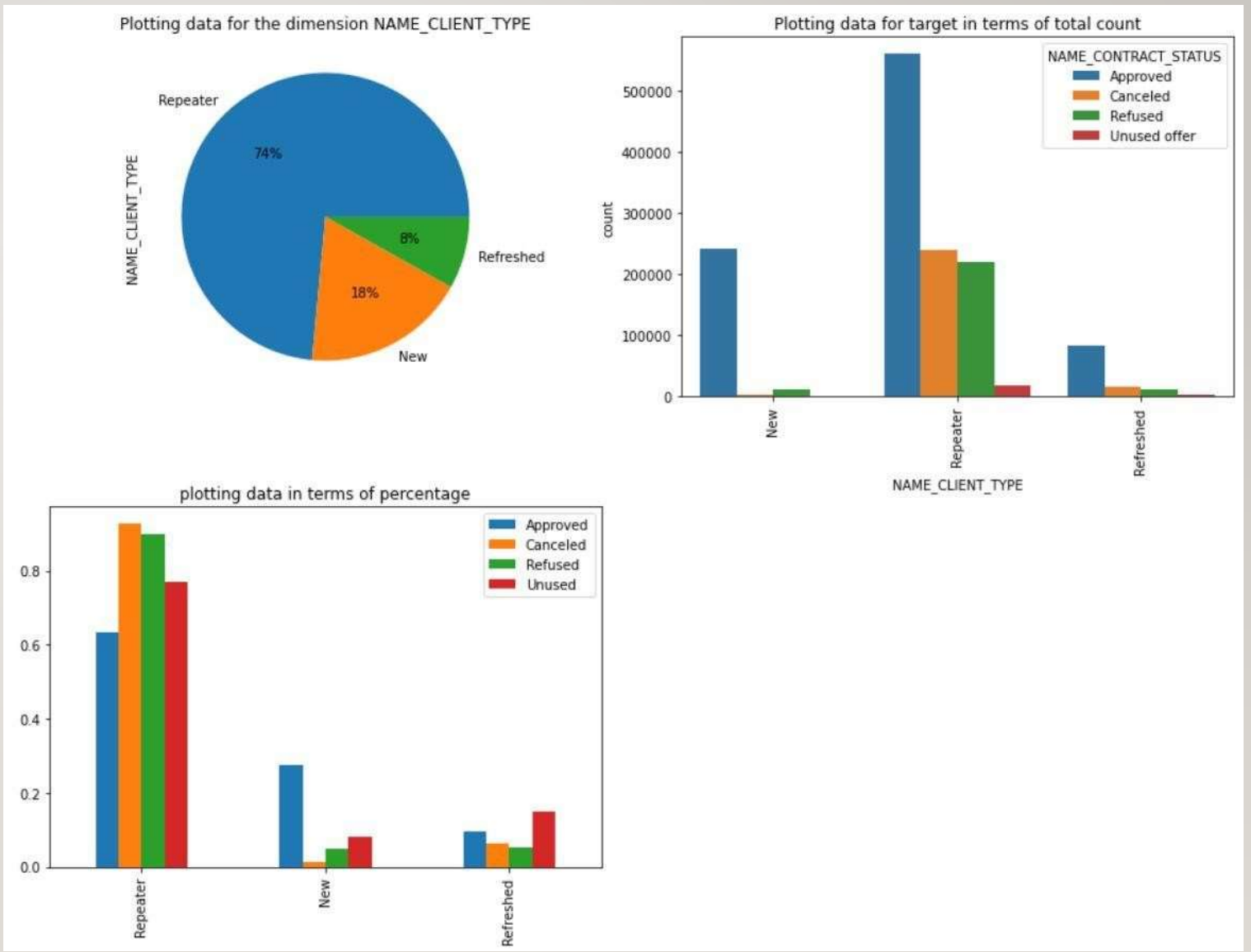
Approved loan percentage is more in Country-wide channel type. Unused loan and Canceled loan percentage is highest among Country-wide and Credit and cash offices respectively.



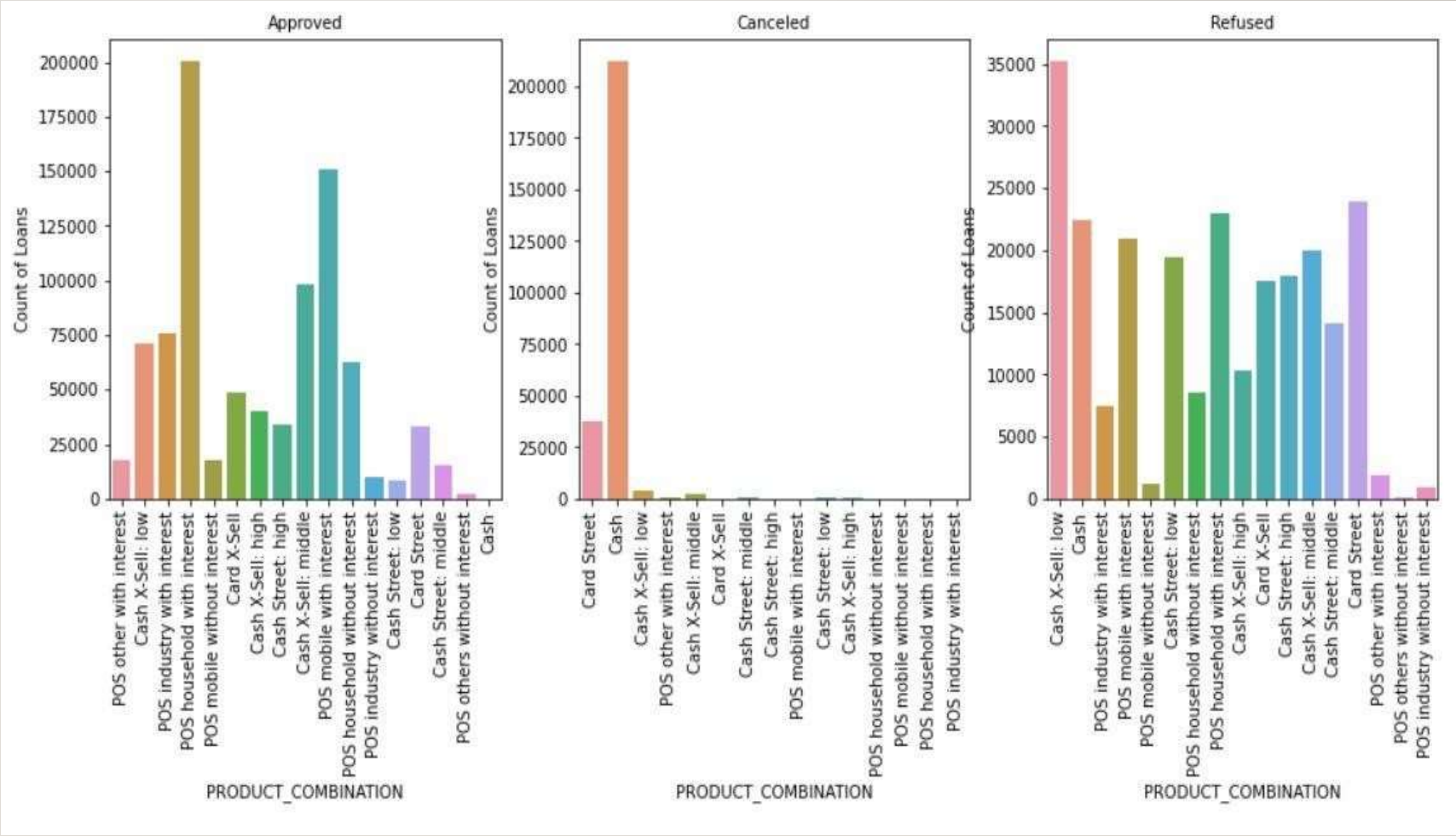




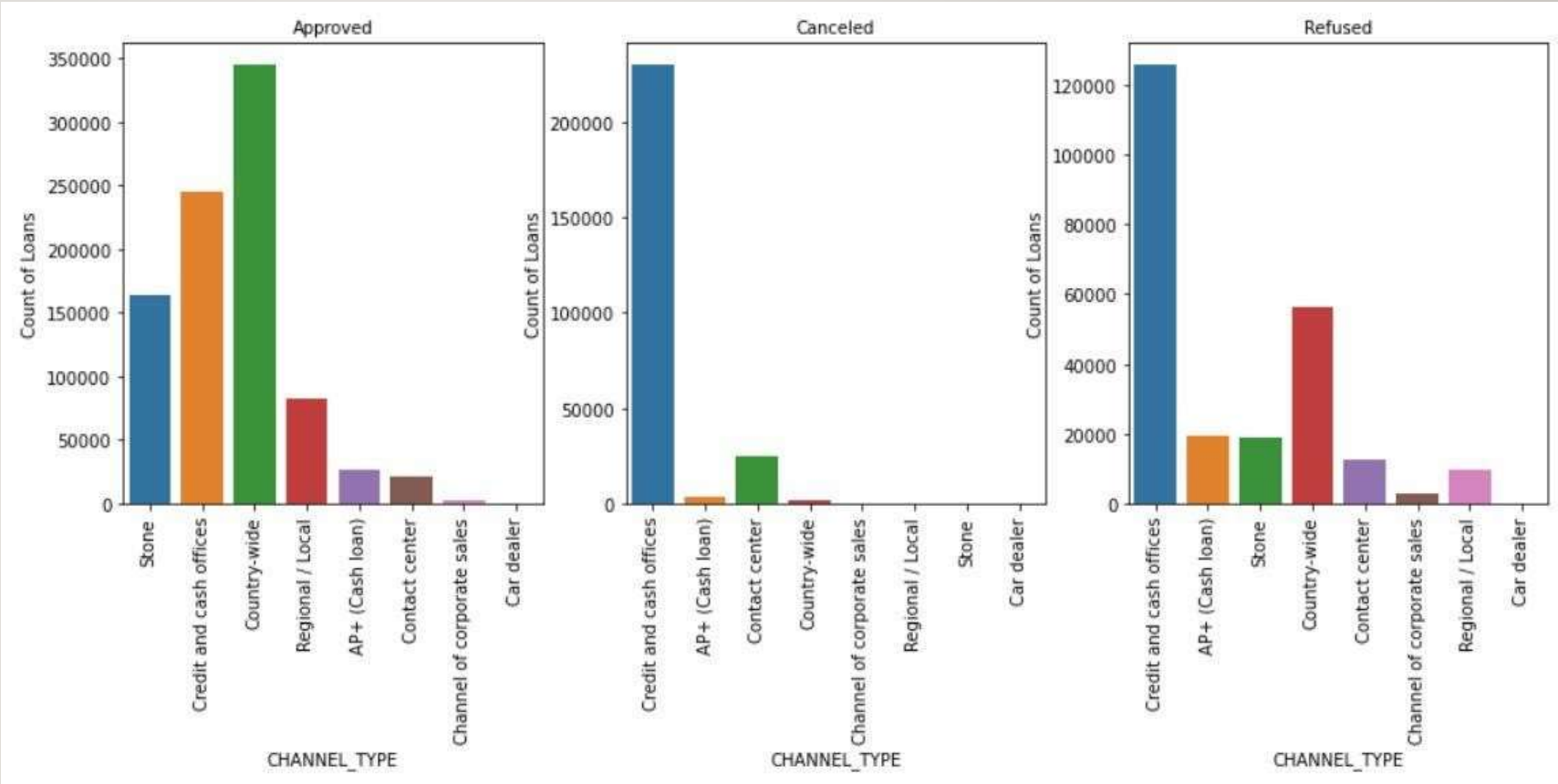
Repeater client type seems to have more percentage of Canceled and Refused loan.



POS household with interest product combination has more number of Approved loans. Whereas Cash product combination has more no of Canceled loans. And, Cash X-Sell: low sees more no of Refused loans.



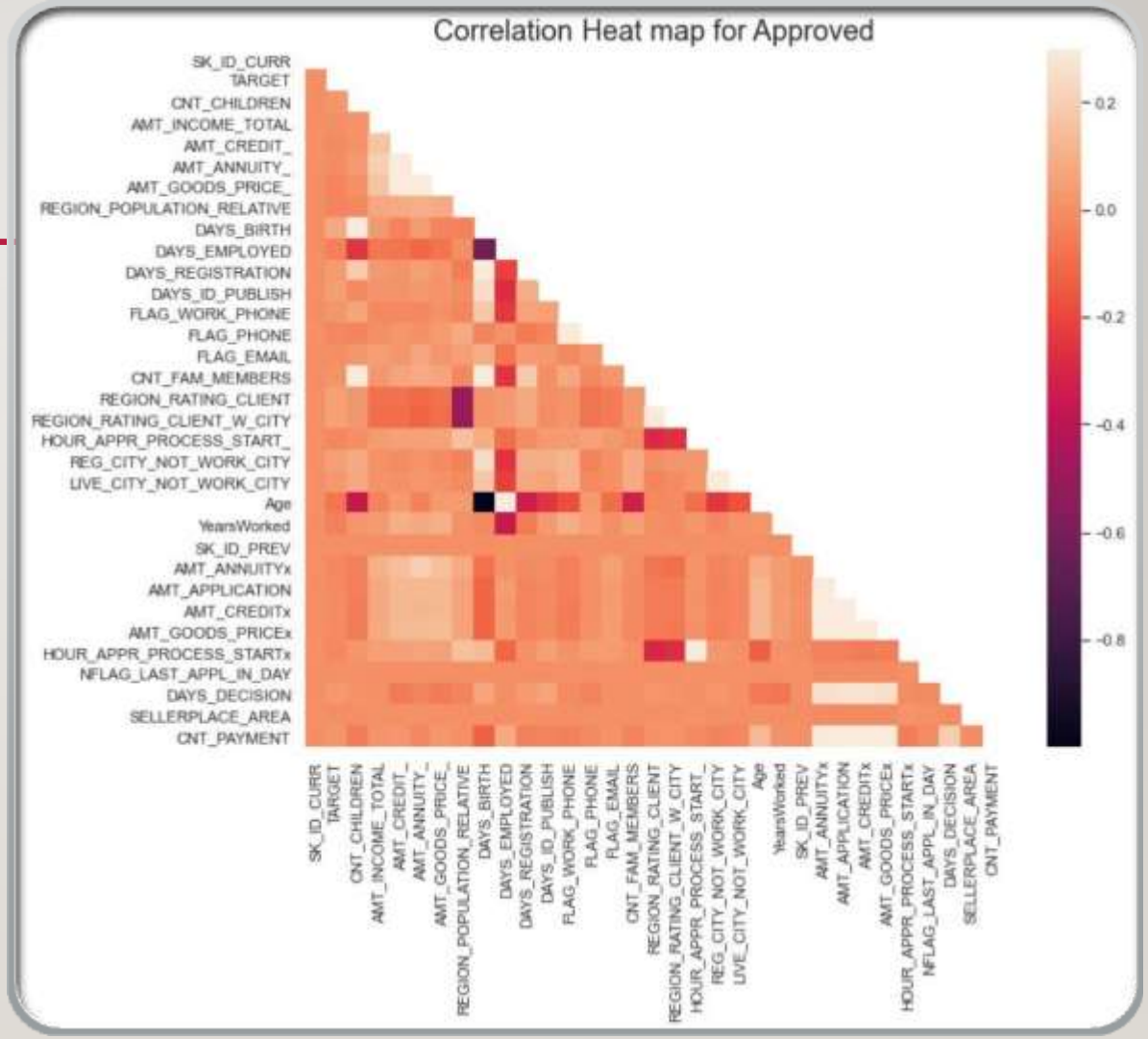
Country-wide channel type sees more no of Approved loans. Whereas, Credit and cash offices channel type sees more number of Canceled and Refused loans.



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 %**)



Conclusion

- Banks should focus more on education type 'Higher education' and avoid Secondary/secondary special, incomplete higher or lower secondary as they face paying difficulties.
- Avoid income type of 'Working' clients as they have high percentage of paying difficulties. Instead focus on Commercial associate, pensioner and State servant.
- 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.
- Banks should focus on the client from age group of 41 to 70 as they will be financial stable and shows less paying difficulties.

THANK YOU