CREDIT EDA CASE STUDY

BY - RAHUL HALLIMANI
AND
NIKHIL ALOK

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 .

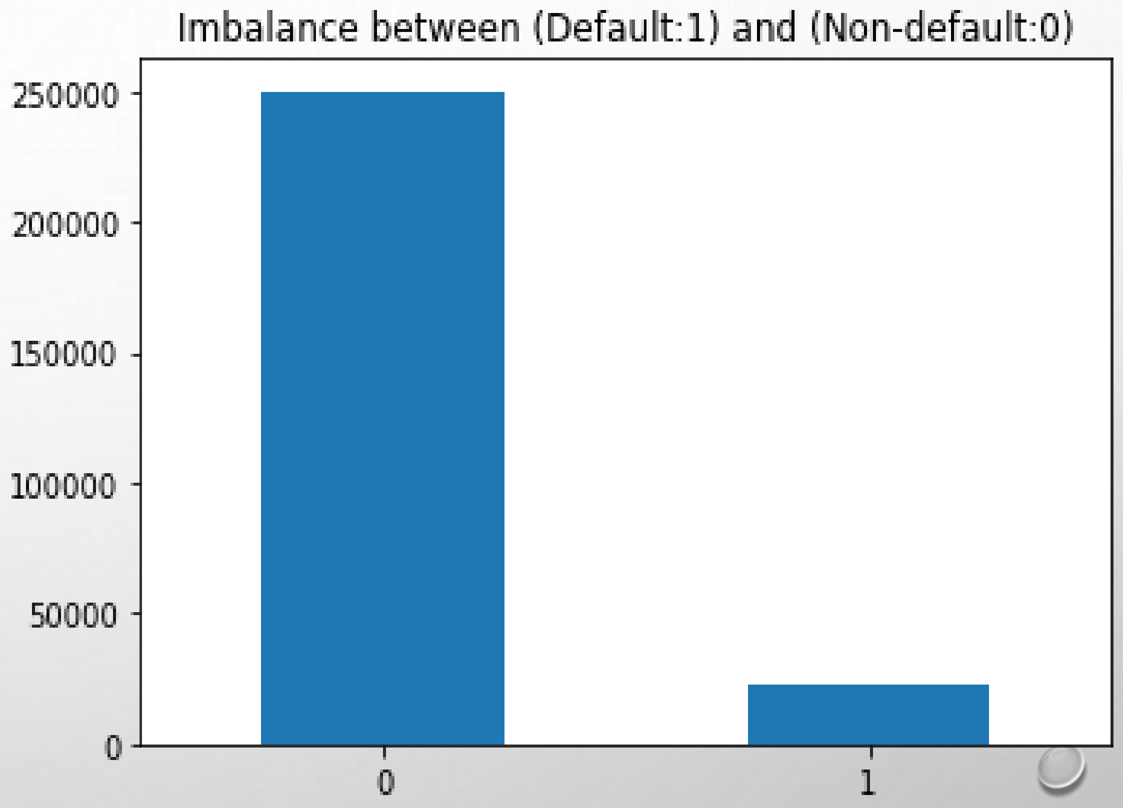
Data Imbalance

Inference:

Clearly when the data has been separated to 2 components w.r.t Target variable there was a data imbalance

The Imbalance ratio is **10.98.**

(i.e. number of Non-defaulters is **10.98** times of the number of Defaulters)



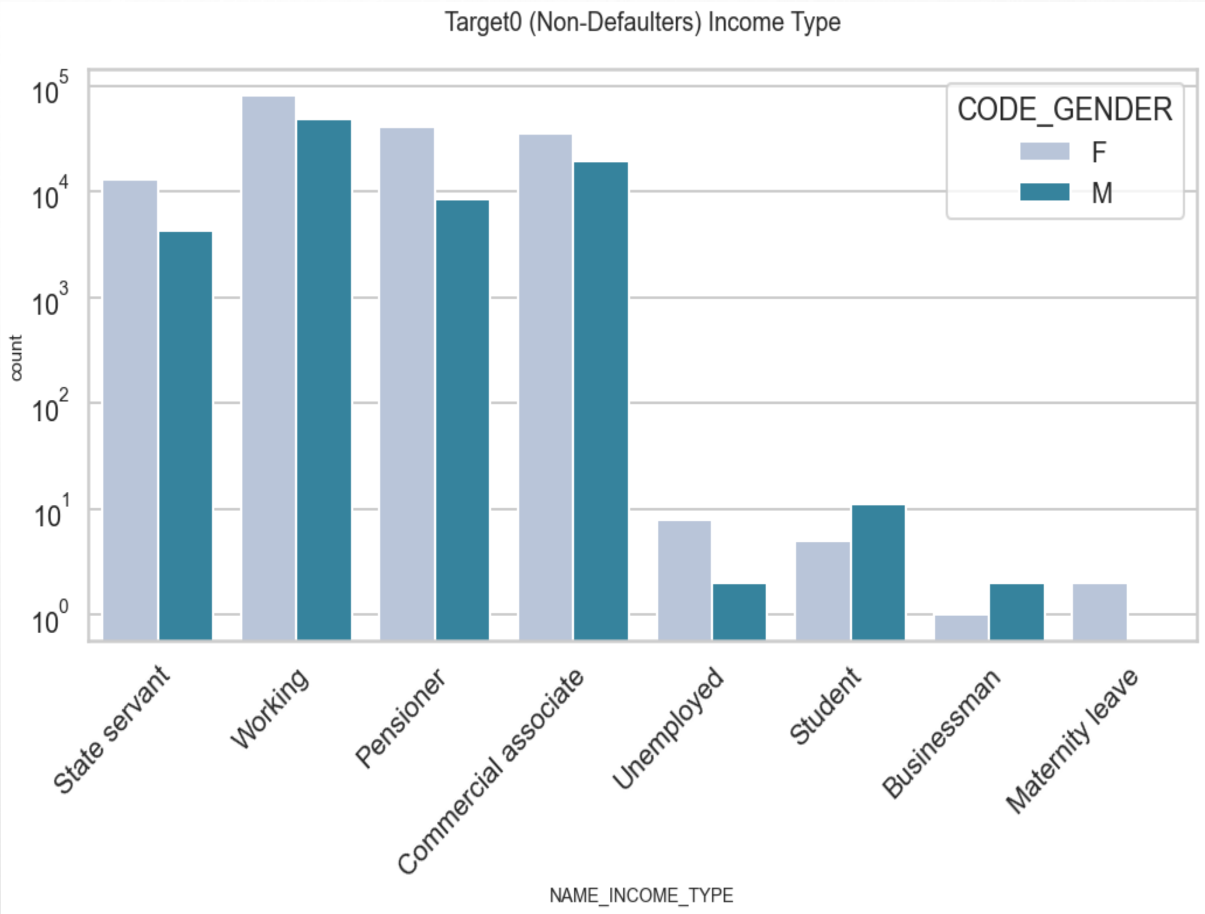
Univariate analysis for categories

Inference:
Points to be concluded from the above graph for Target=0 (Non-Defaulters):-

•Female are more in number than males in almost all the income type except student and businessman.

•For Maternity leave, the graph shows only females which is expected.
•High number for income type state servant, working , commercial associate and pensioner.

•Low number for income type student, unemployed, businessman and maternity leave.



Inference:

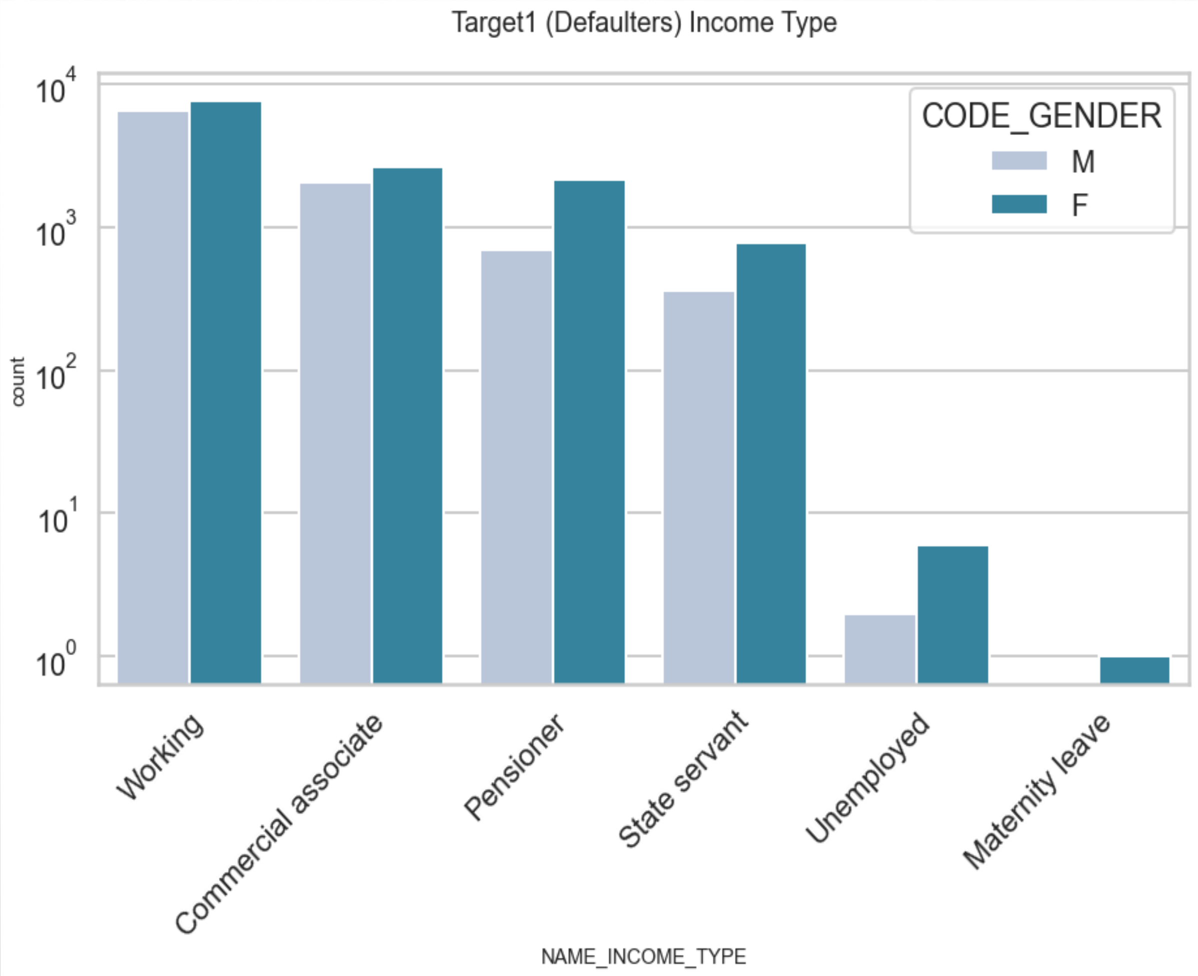
Points to be concluded from the above graph for Target=1 (Defaulters):-

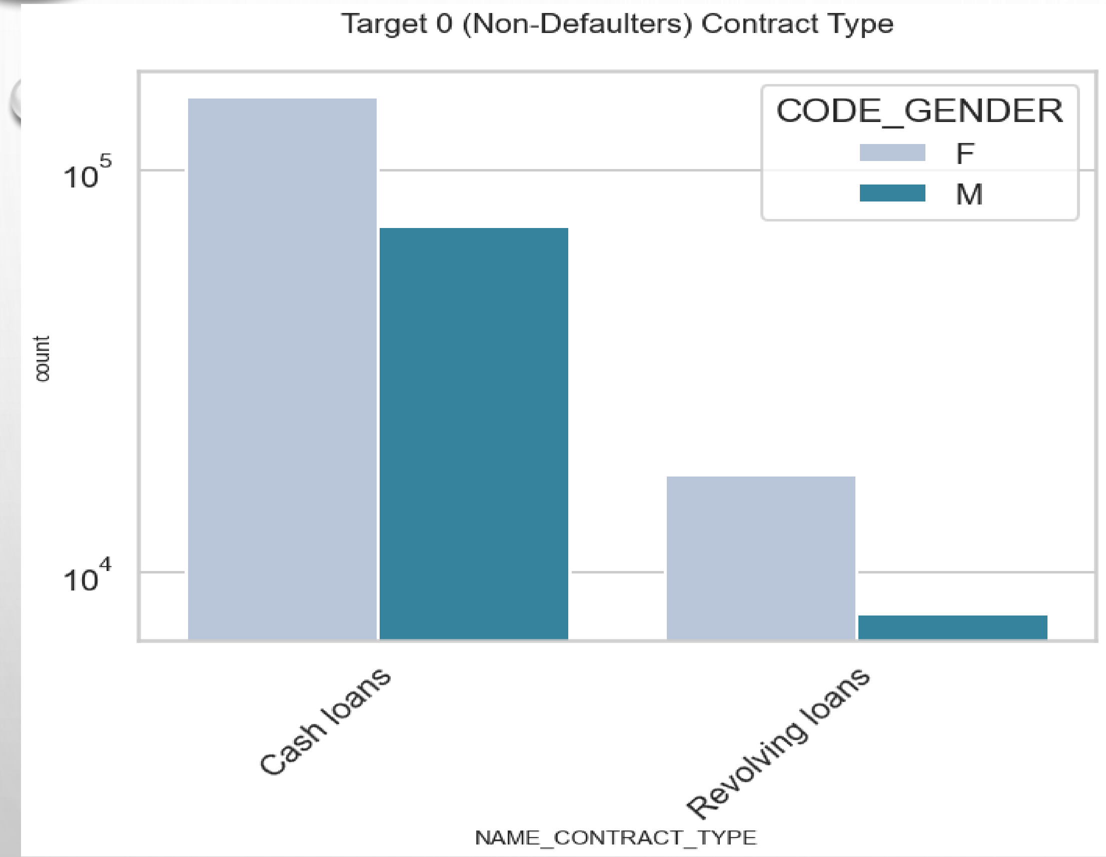
•Females are more in number than males in all the income type.

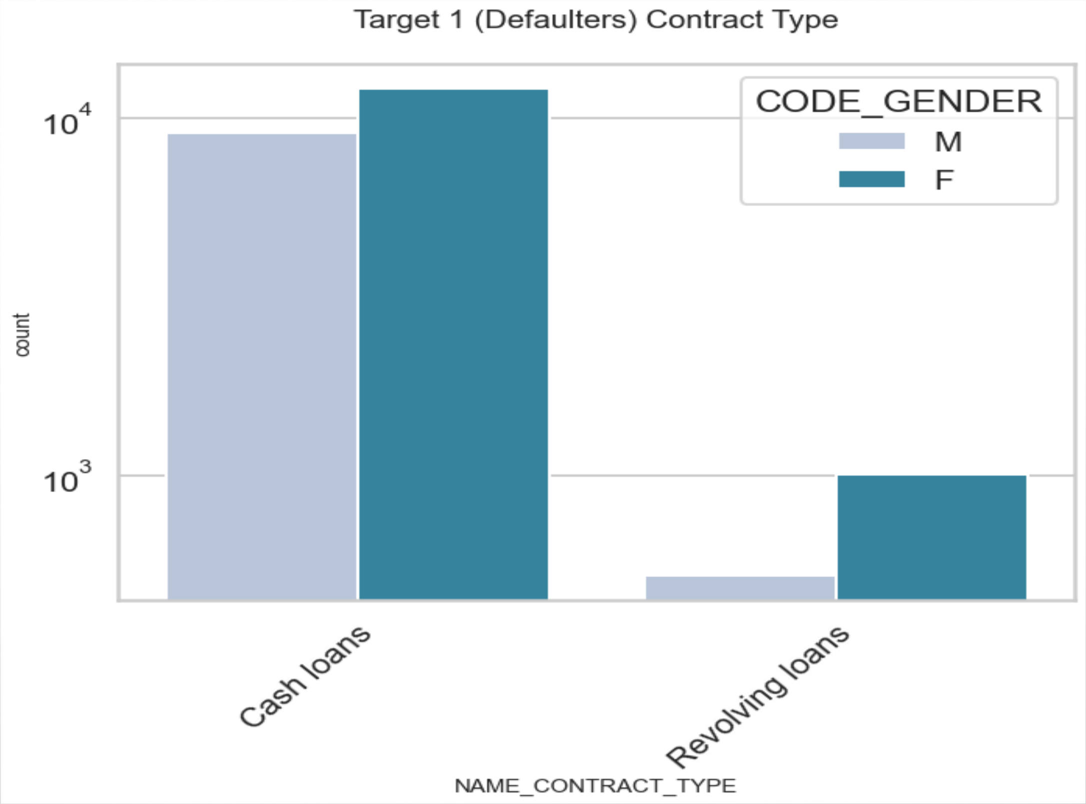
•High number of Defaulters for income type working , commercial associate , pensioner and state servant.

•Low number of Defaulters for income type unemployed and maternity leave.

•One very important thing we notice if we compare this graph(Defaulters) with that of previous one(Non-Defaulters), is that there are no 'Student' or 'Businessman' among the Defaulters.







Inference:

Points to be concluded from the above graph for Target=0 (Non-Defaulters):-

•Cash Loan contracts have a higher number of credit than revolving loan contracts.

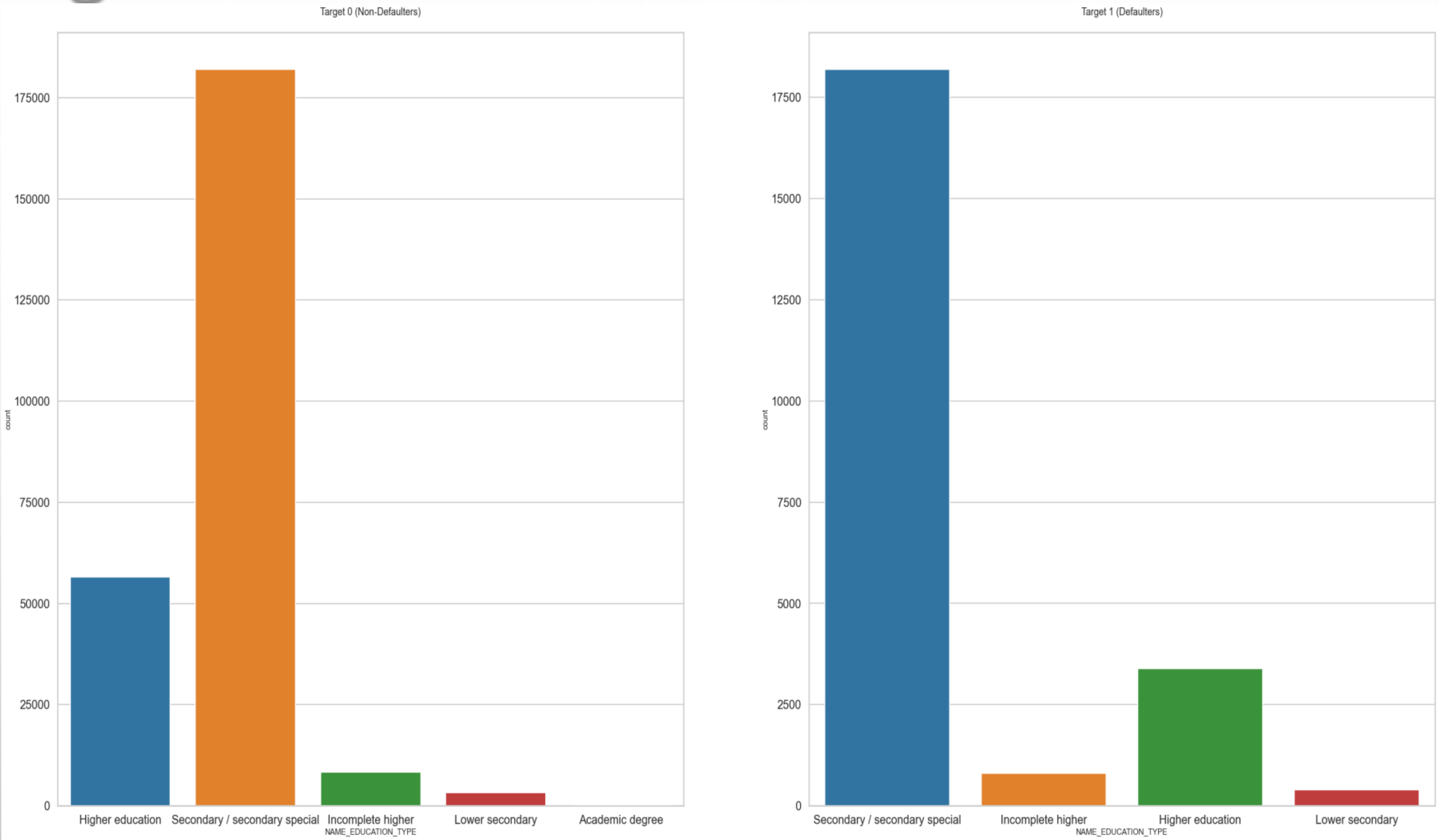
•Count of female is more.

Inference:

Points to be concluded from the above graph for Target = 1 (Defaulters):-

1.Cash Loan contracts have a higher number of Defaulters than revolving loan contracts

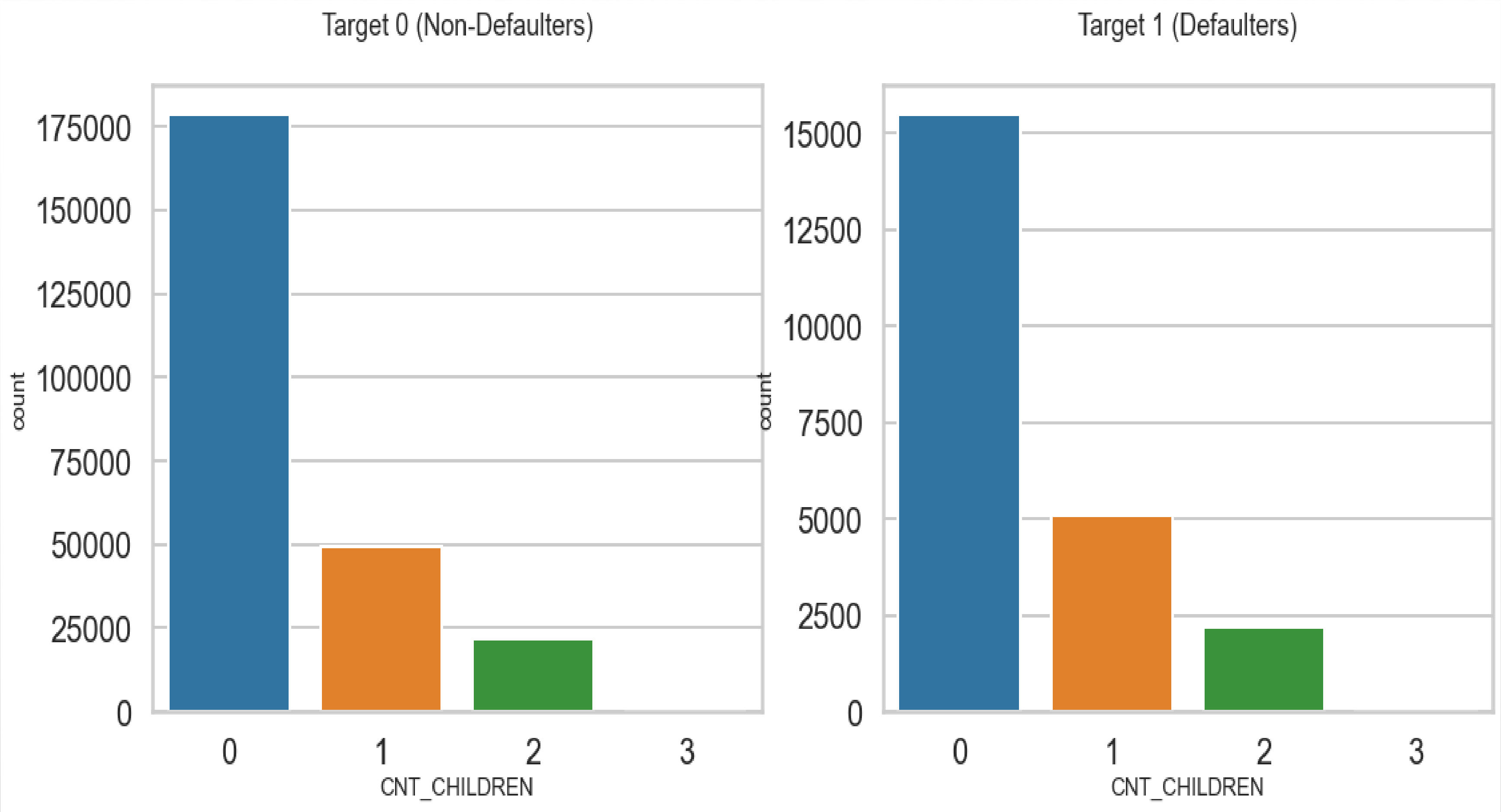
2.Females are more in number.



Inference:

Points to be concluded from the above graph:-

•We can see that people with secondary education are highest in number among both defaulters and non-defaulters. This could be because a high majority of population is secondary educated.



Inference:

Points to be concluded from the above graph:-

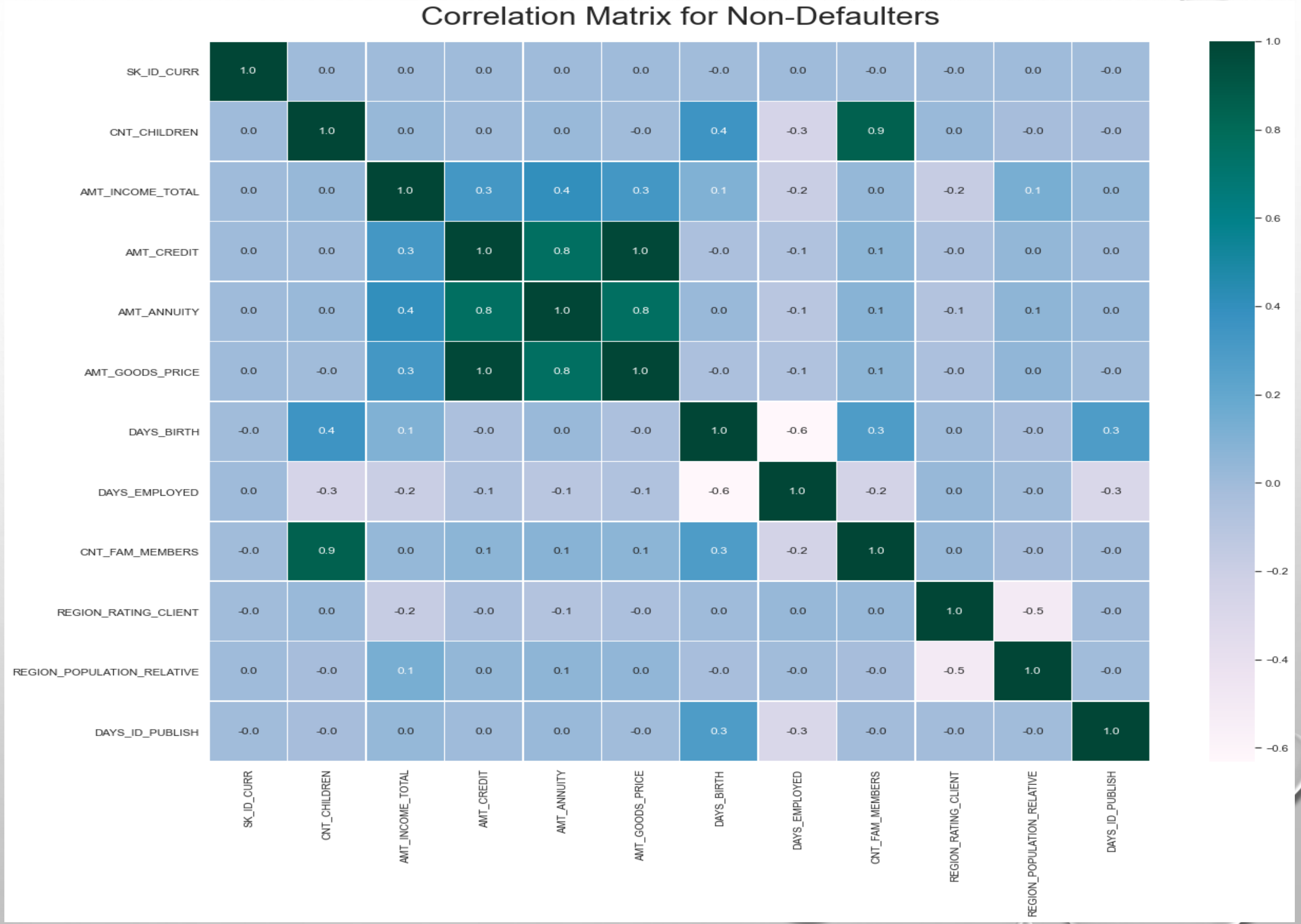
•We can see that lesser number of children maximizes the chances of both being a defaulter as well as a non-defaulter. So we cannot conclude anything significant from this exploration.

Bivariate Analysis

Inference:

These columns have high correlation values for Target 0:-

- AMT_GOODS_PRICE and AMT_CREDIT
- AMT_ANNUITY and AMT_CREDIT
- AMT_ANNUITY and AMT_GOODS_PRICE
- CNT_FAM_MEMBERS and CNT_CHILDREN
- AMT_ANNUITY and AMT_INCOME_TOTAL
- AMT_INCOME_TOTAL and AMT_GOODS_PRICE



Inference:

Both for Target 0 and Target 1 these columns have high correlation values:-

•AMT_GOODS_PRICE and AMT_CREDIT

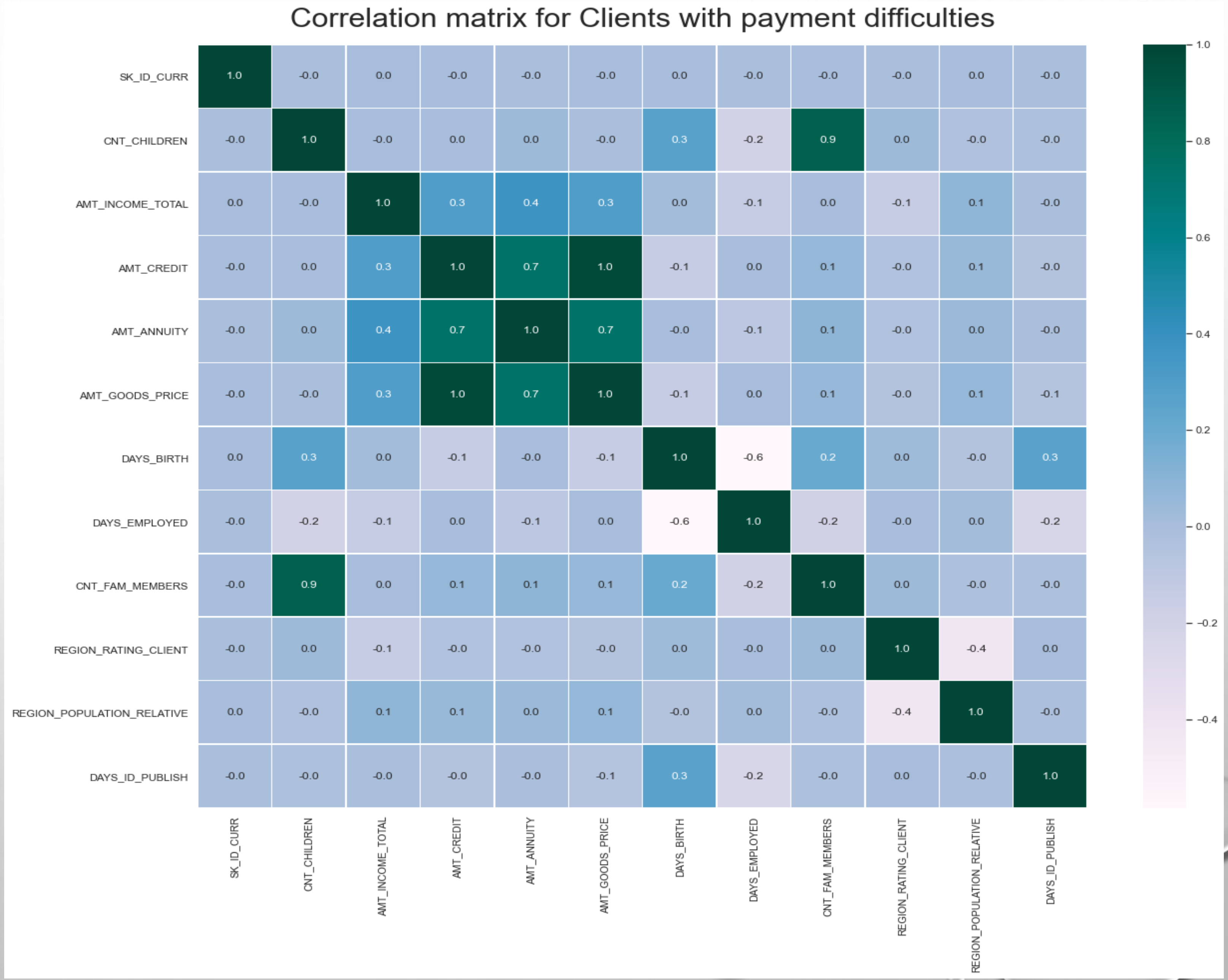
•AMT_ANNUITY and AMT_CREDIT

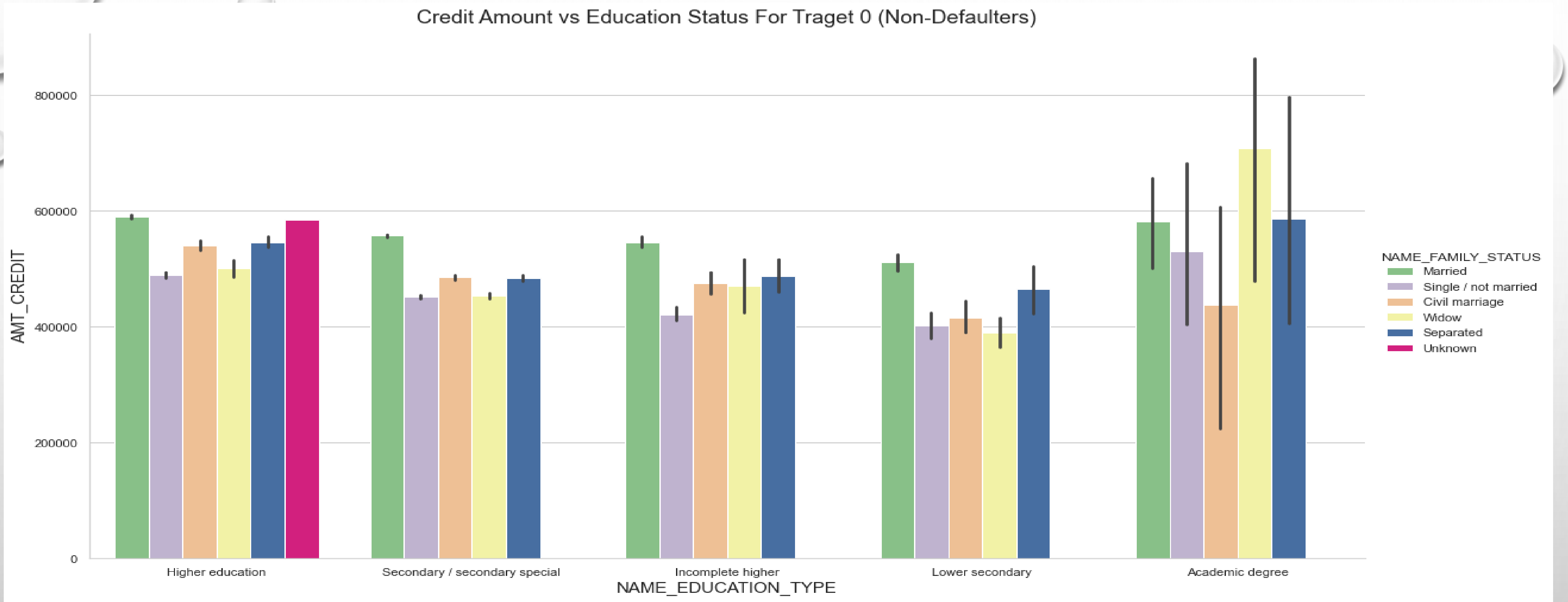
•AMT_ANNUITY and AMT_GOODS_PRICE

•CNT_FAM_MEMBER and CNT_CHILDREN

•AMT_ANNUITY and AMT_INCOME_TOTAL

•AMT_INCOME_TOTAL and AMT_GOODS_PRICE





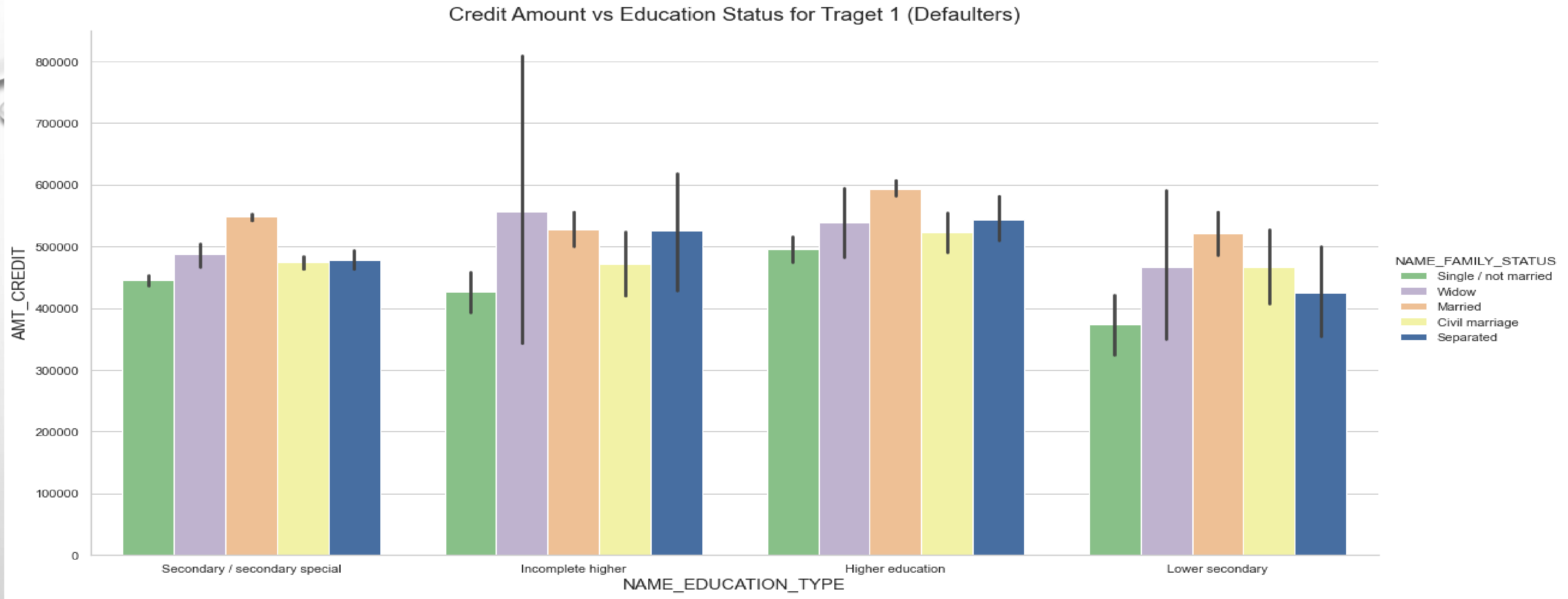
Inference:

Points to be concluded from the above graph for target = 0 (Non-Defaulters):-

•Customers holding academic degree have greater credit amount, Civil marriage segment being the highest among them.

•Lower educated customers tends to have lower credit amount, Widows being the lowest among them

•Married customers in almost all education segment except lower secondary and academic degrees have a higher credit amount.



Inference:

Points to be concluded from the above graph for target = 1 (Defaulters):-

- One very important difference we see from the previous graph is that there's no Academic Degree holders among Defaulters.
- Accross all education segment married customer tends to have higher credit amount.
- Customers holding lower education tends to have a lower credit amount but among them married ones take higher credits.

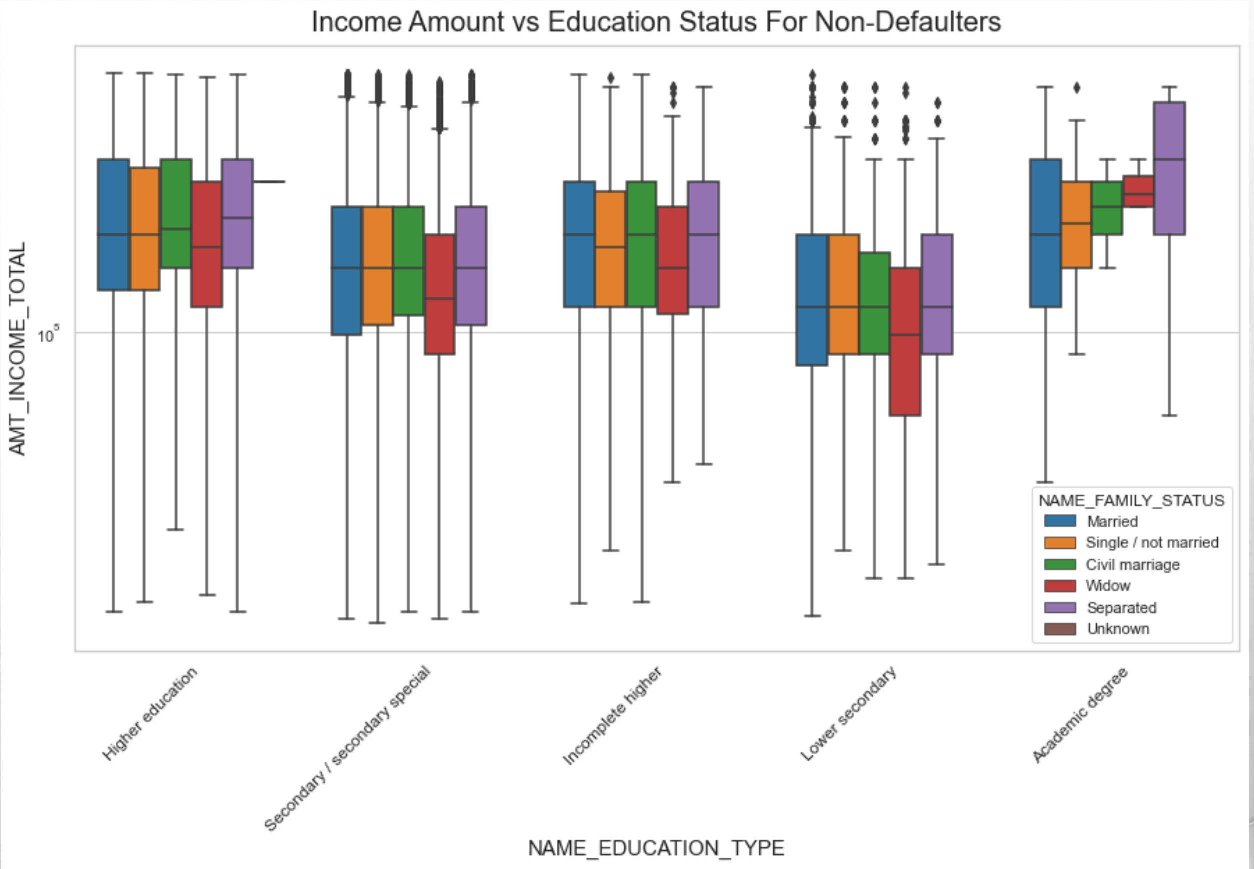
Inference:
Points to be concluded from the above graph for target = 0 (Non-Defaulters):-

•For Education type 'Lower secondary' and 'Secondary/secondary special', the median income amount is almost same across family status except for widows whose income is significantly lower in the group.

•For Higher Education and Academic degree holders, there is almost no outliers.

•Academic degree holders have highest income followed by Higher education.

•Lower secondary education type has the lowest income; and among them widows are the lowest.

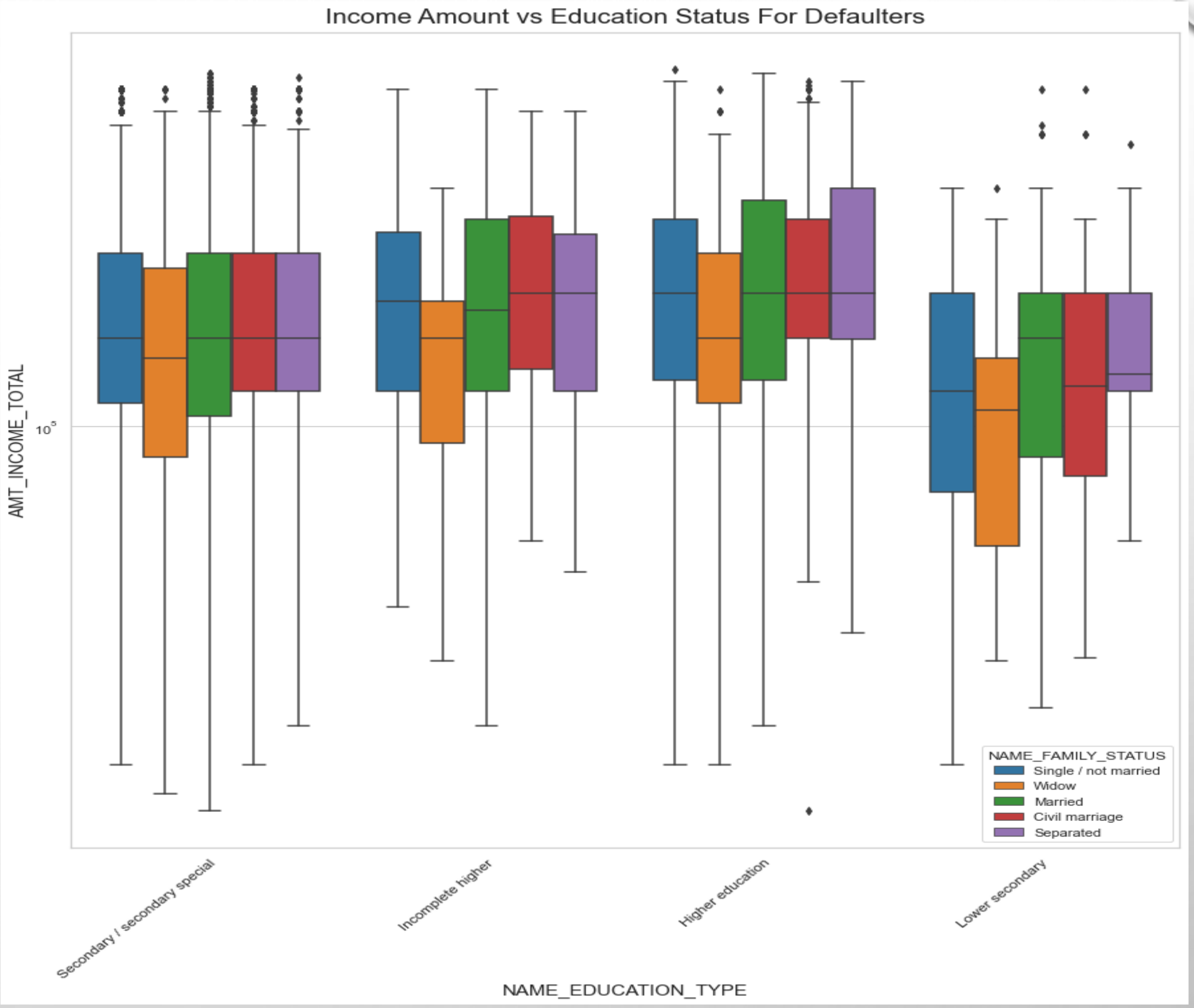


Inference:
Points to be concluded from the above graph for target = 1 (Defaulters):-

•Comparing with previous graph, we clearly notice that there is no Academic Degree holders among Defaulters

•Among Defaulters, Higher education and Incomplete Higher education have very similar income pattern.

•We don't see a very strong difference in the income of defaulter except for Lower secondary education type which is on the lower side.



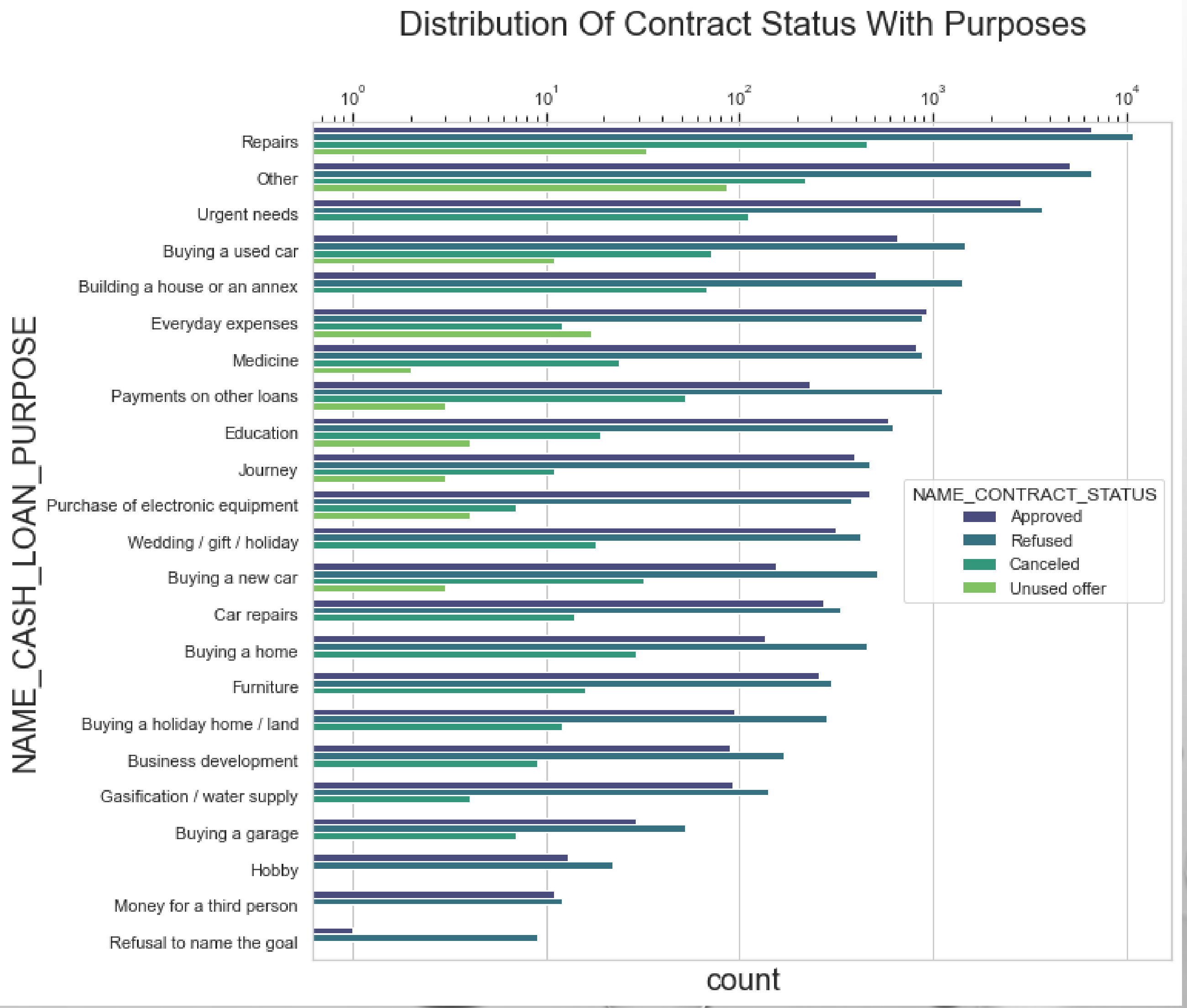
Previous Application Data
Univariate Analysis

Inference:

We can conclude the below points from the graph:-

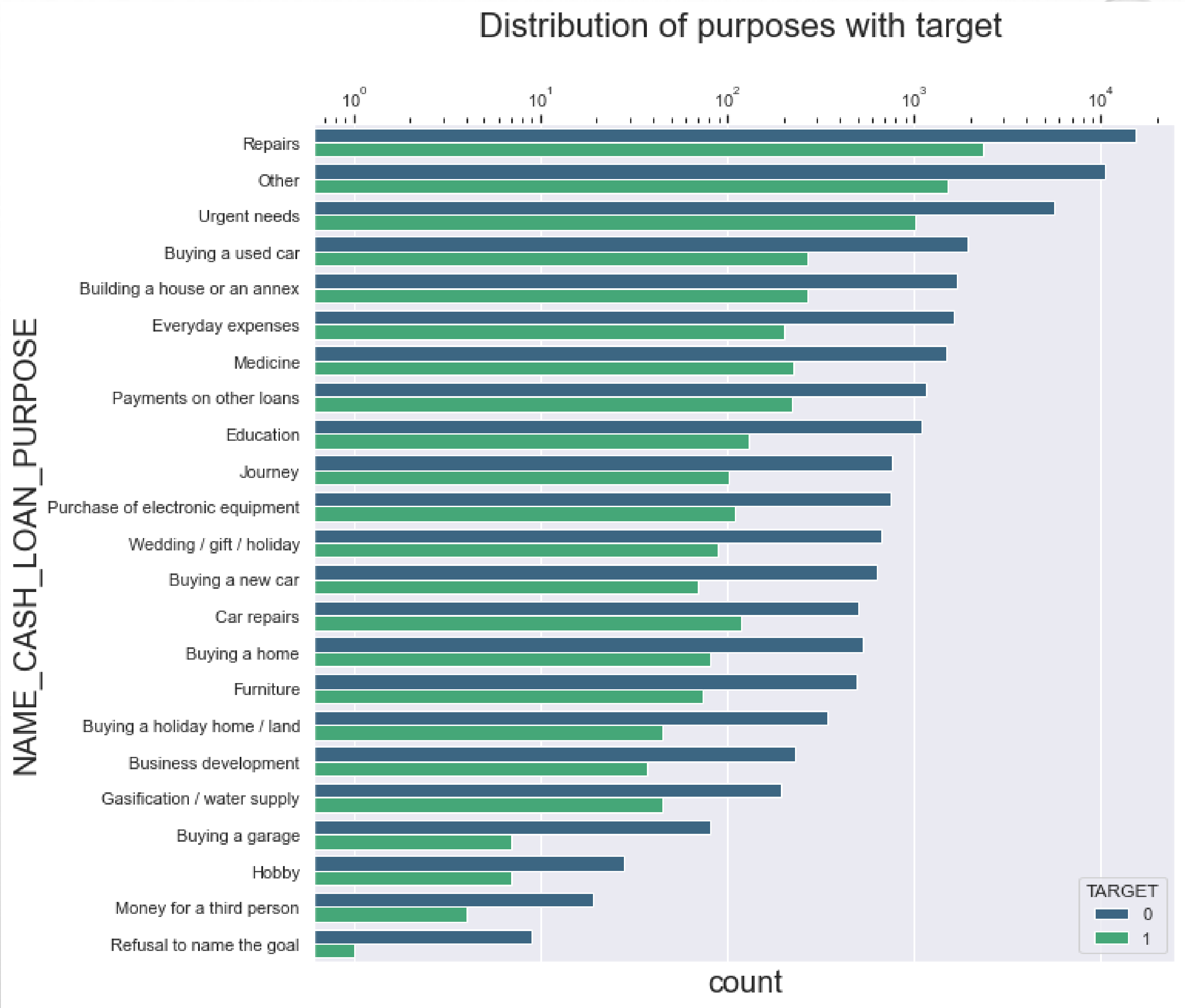
•Most rejection of loans came from purpose 'Repairs'
.

•We have almost equal number of approves and rejection for Medicine, Every day expenses and education purposes.



Inference:

•We can conclude from above plot that Loan purposes with 'Repairs', 'Urgent need' and 'Others' are facing more difficulties in payment on time.



Bivariate Analysis

Inference:

•Here for Housing type, office appartment is having higher credit of target 0 and co-op apartment is having higher credit of target 1.

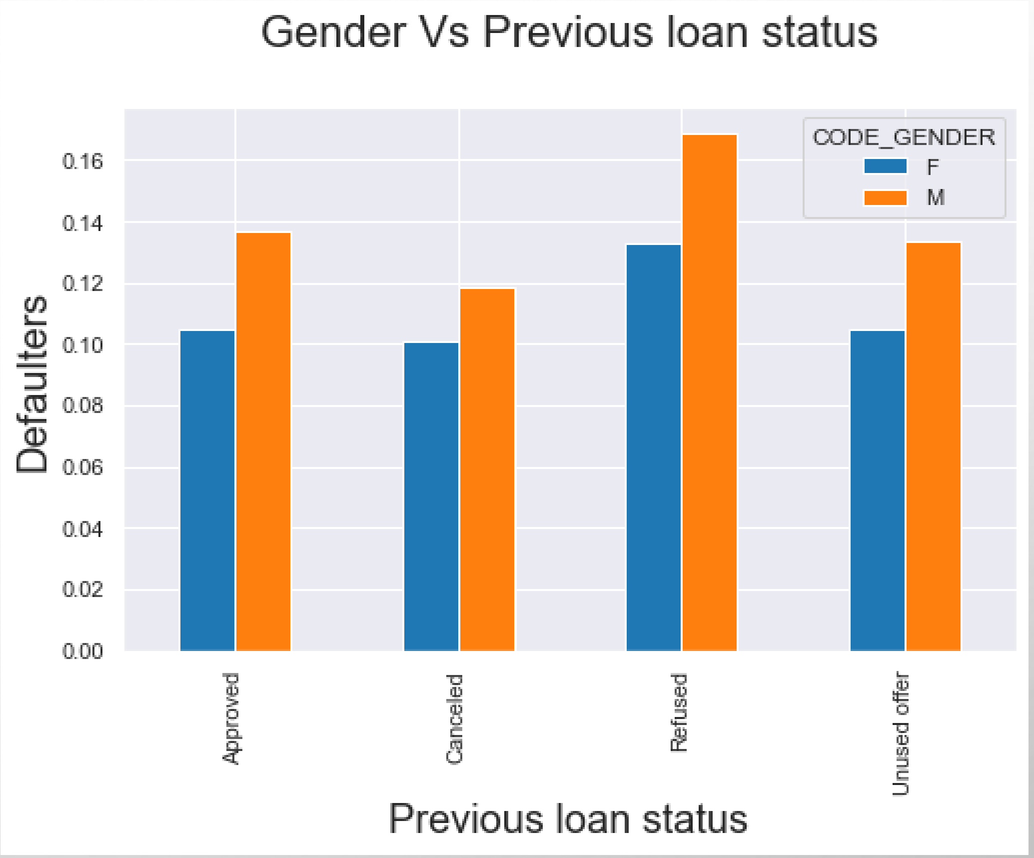
•So, we can conclude that bank should avoid giving loans to the housing type of co-op apartment as they are having difficulties in payment.

•Bank can focus mostly on housing type with parents or House \ appartment or muncipal appartment for successful payments.

Previous loan Status Vs Current Defaulters Plot

Inference:

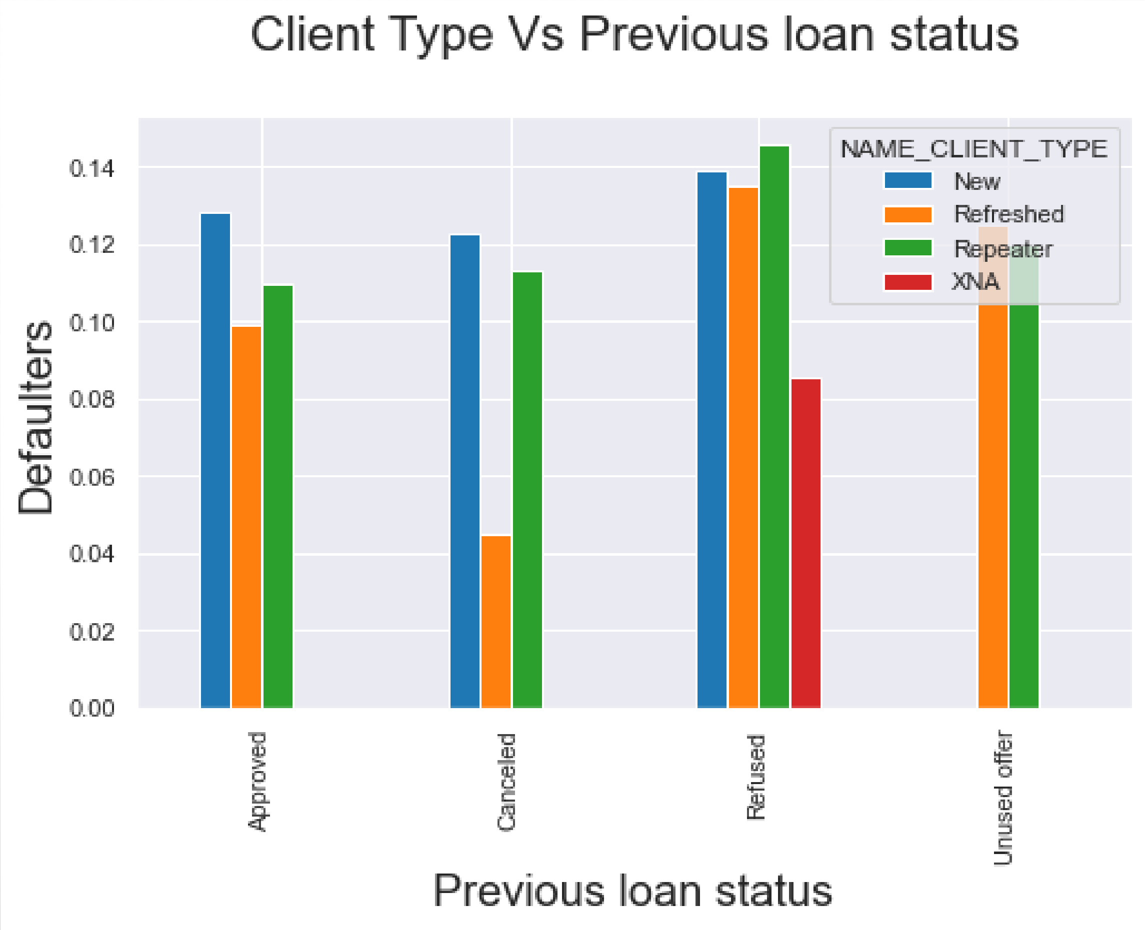
•Male clients are more defaulted than female client. Also, previously refused customer are more defaulted in current application.



Client Type Vs Previous loan status plot

Inference:

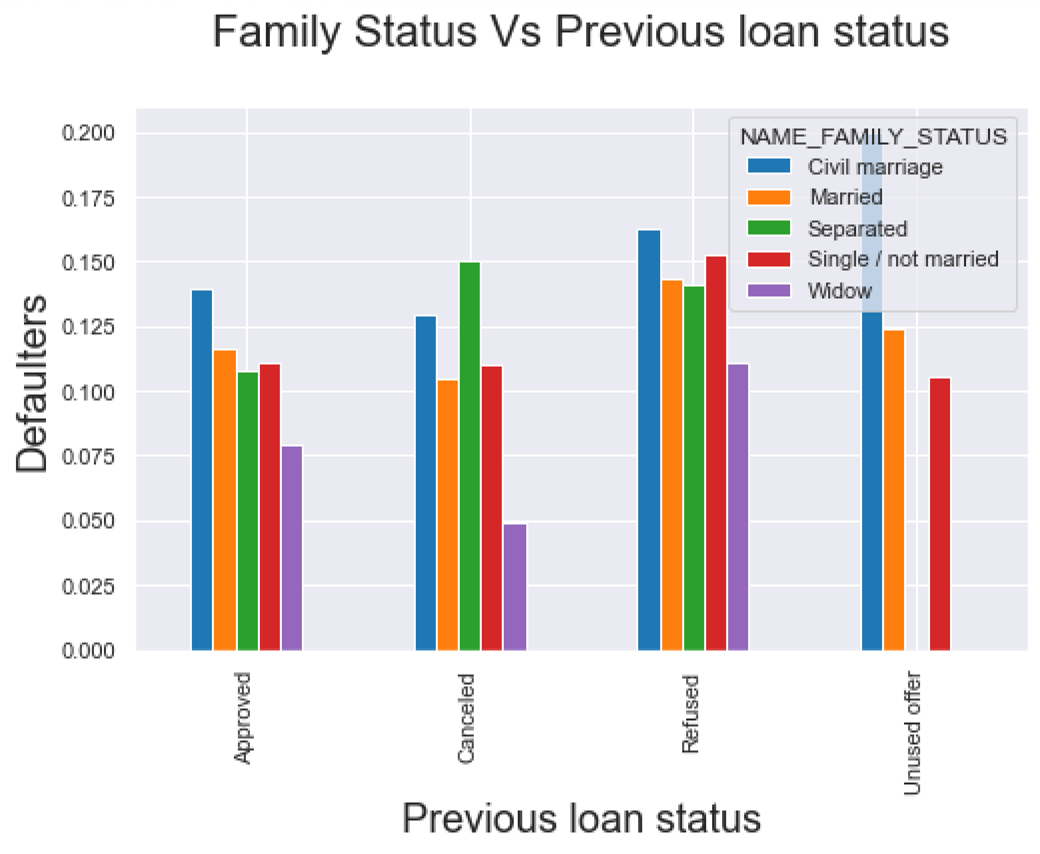
•Previously refused New, refreshed and repeater clients have defaulted more.



Family Status Vs Previous loan status plot

Inference:

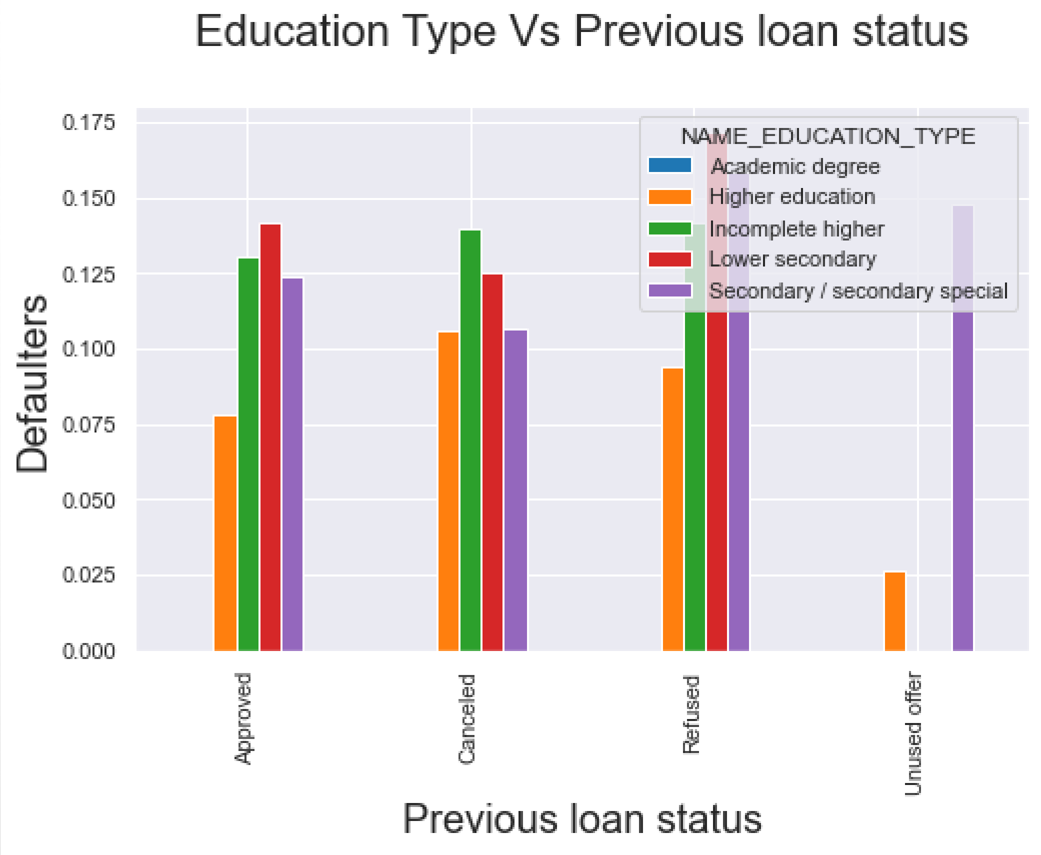
•Client who did civil marriage with previously unused loan offers are more defaulted currently



Education status Vs Previous loan status plot

Inference:

•Previously refused people with lower secondary education and secondary/secondary special are more defaulted in current application



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.

THANK YOU