Credit EDA Case Study

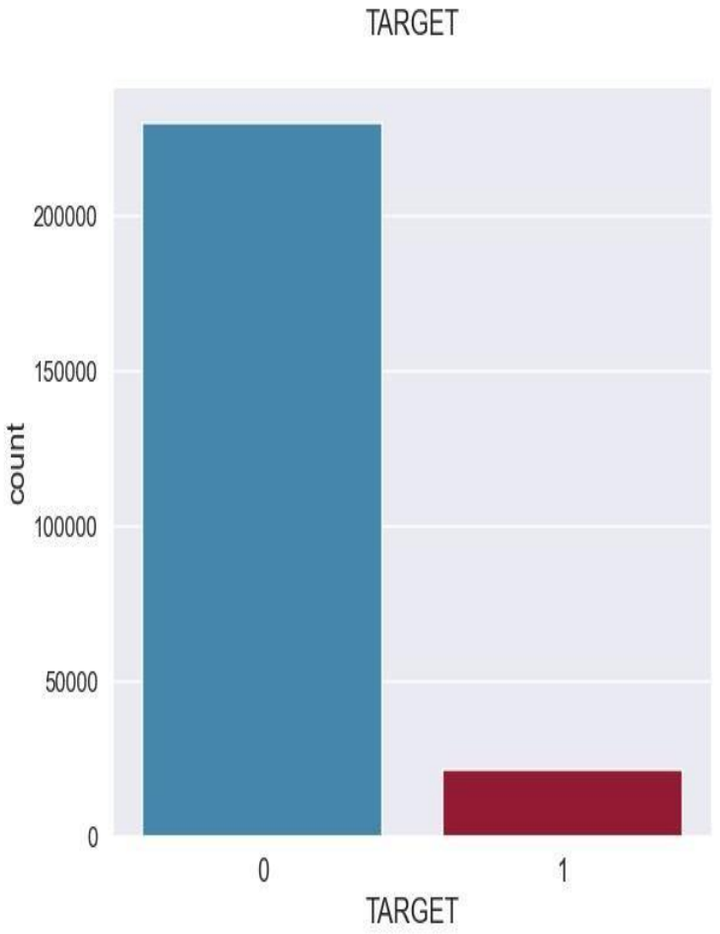
By : *Harshvardhan Pareek*

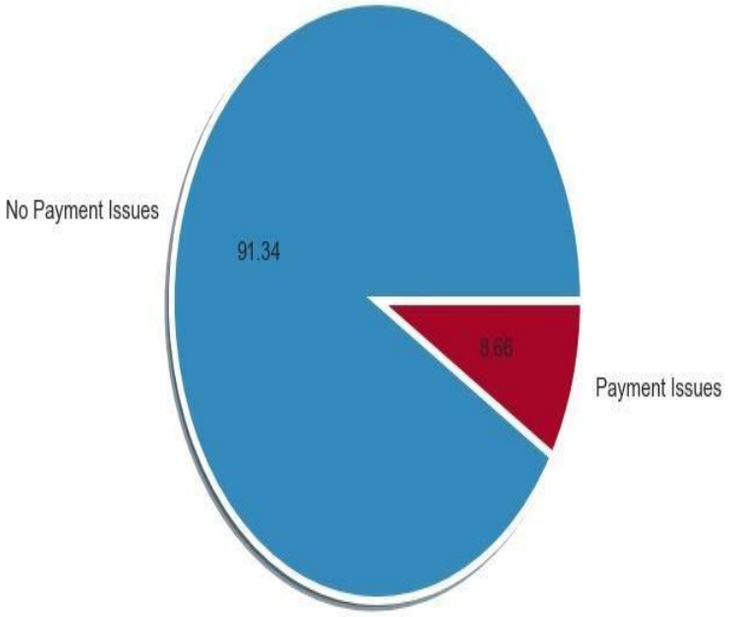
Graph for Missing Data Columns

. Occupation type having most no. of missing data when more than 50% missing data is removed

Data Imbalance Check

. It is clear from the graph that no. of loan repayer's are greater than no. of loan defaulters.





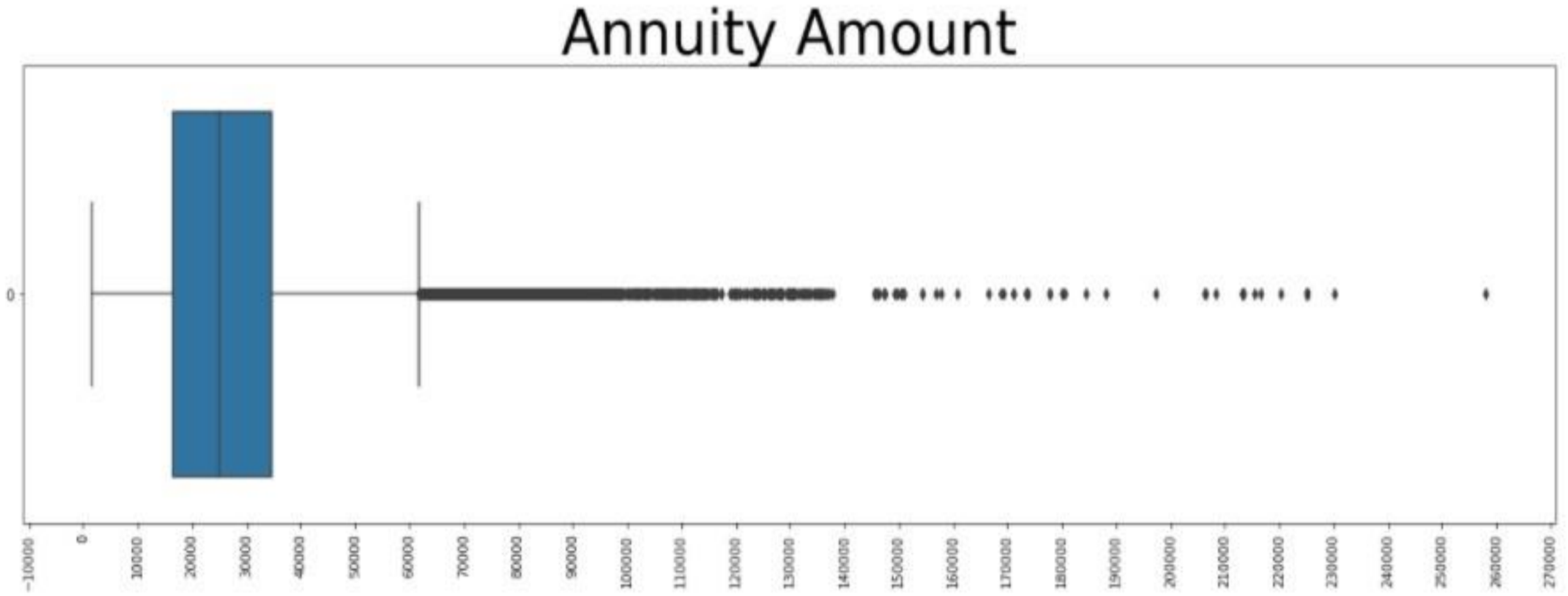
Outlier Values

Box Plot for Credit Amount

The outliers value can handled by converting into bins

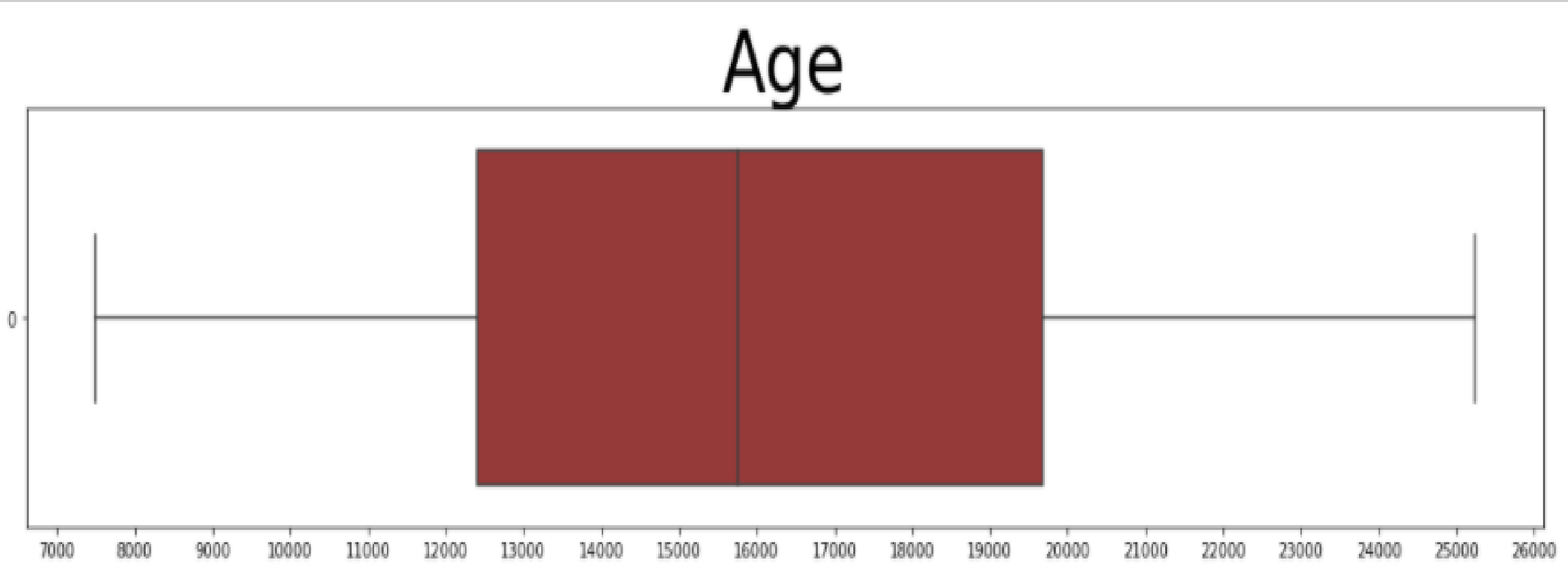
Box Plot for Annuity Amount

Amount Annuity column has some outliers value and handled it by converting into bins



Box Plot for Age

It is found that no outliers are present in DAY_BIRTH column



Box Plot for Age

Maximum value is 365243 which we need to divide from 365 to get no. of years and needs to be replaced with nearest value like 99% value.

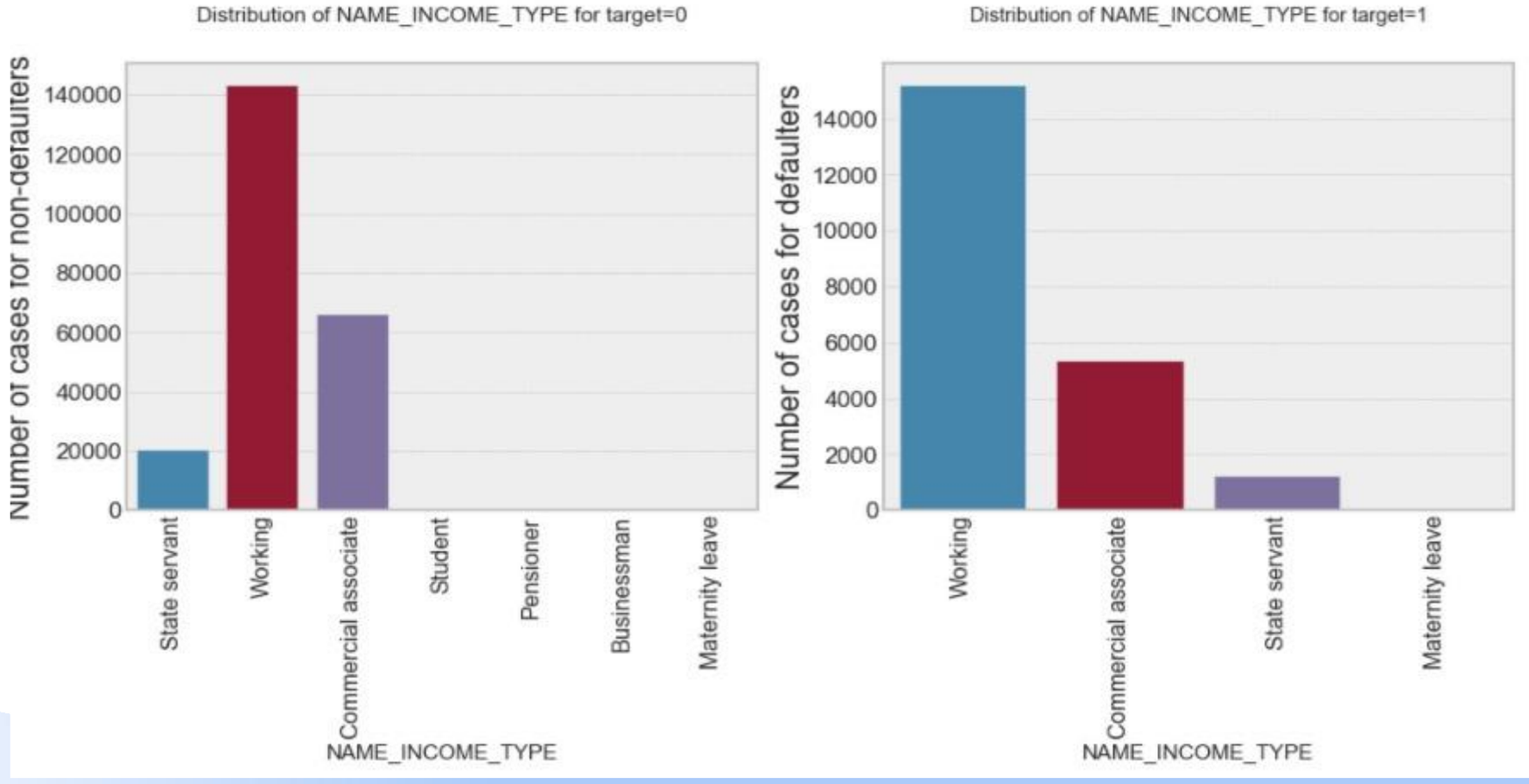
Univariate Analysis

Categorical Univariate Analysis for target =0 and target =1

. Unordered Categorical Data

Graph for Income Type

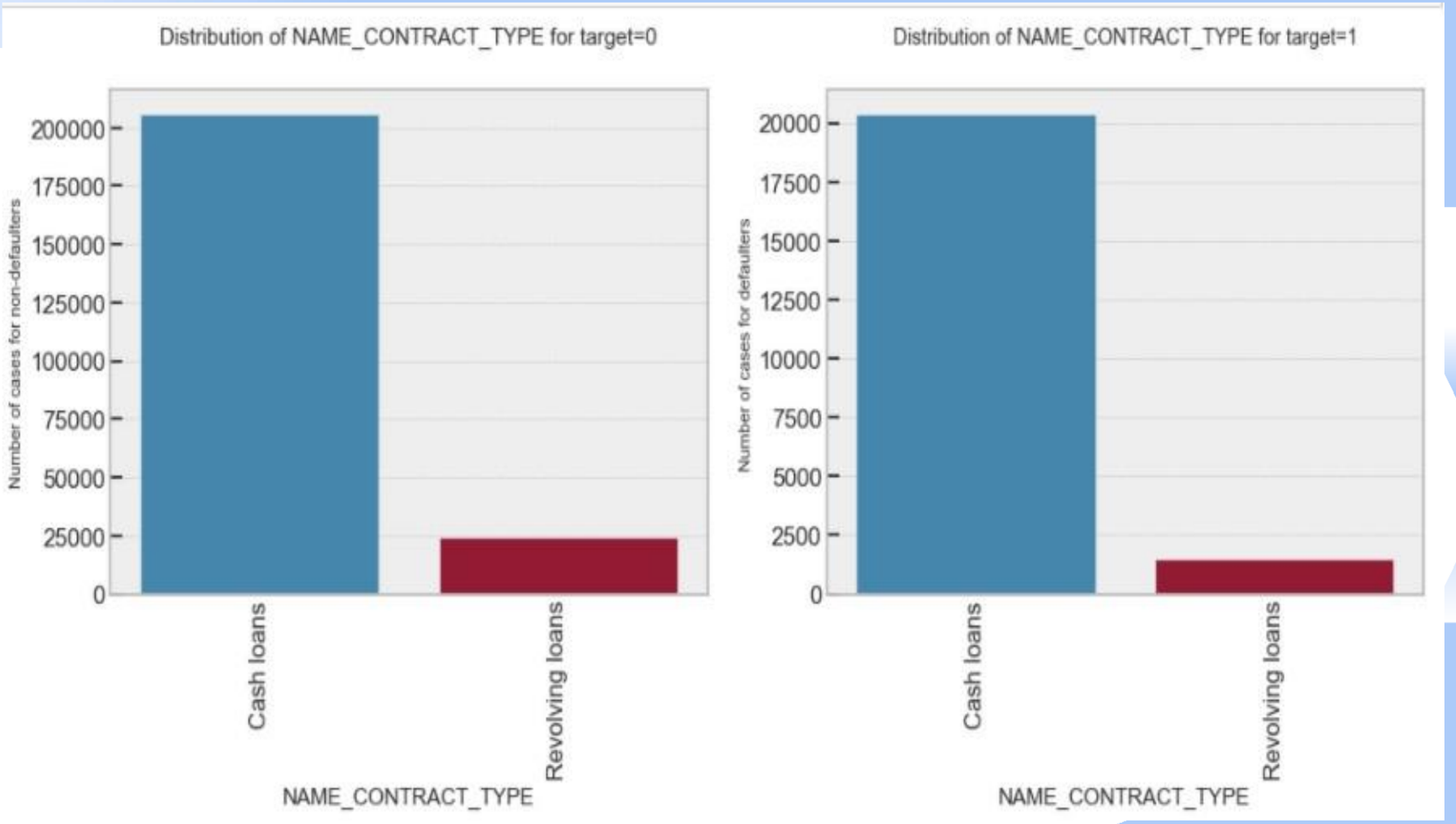
. Working category has most no. of defaulters



Graph for Contract Type

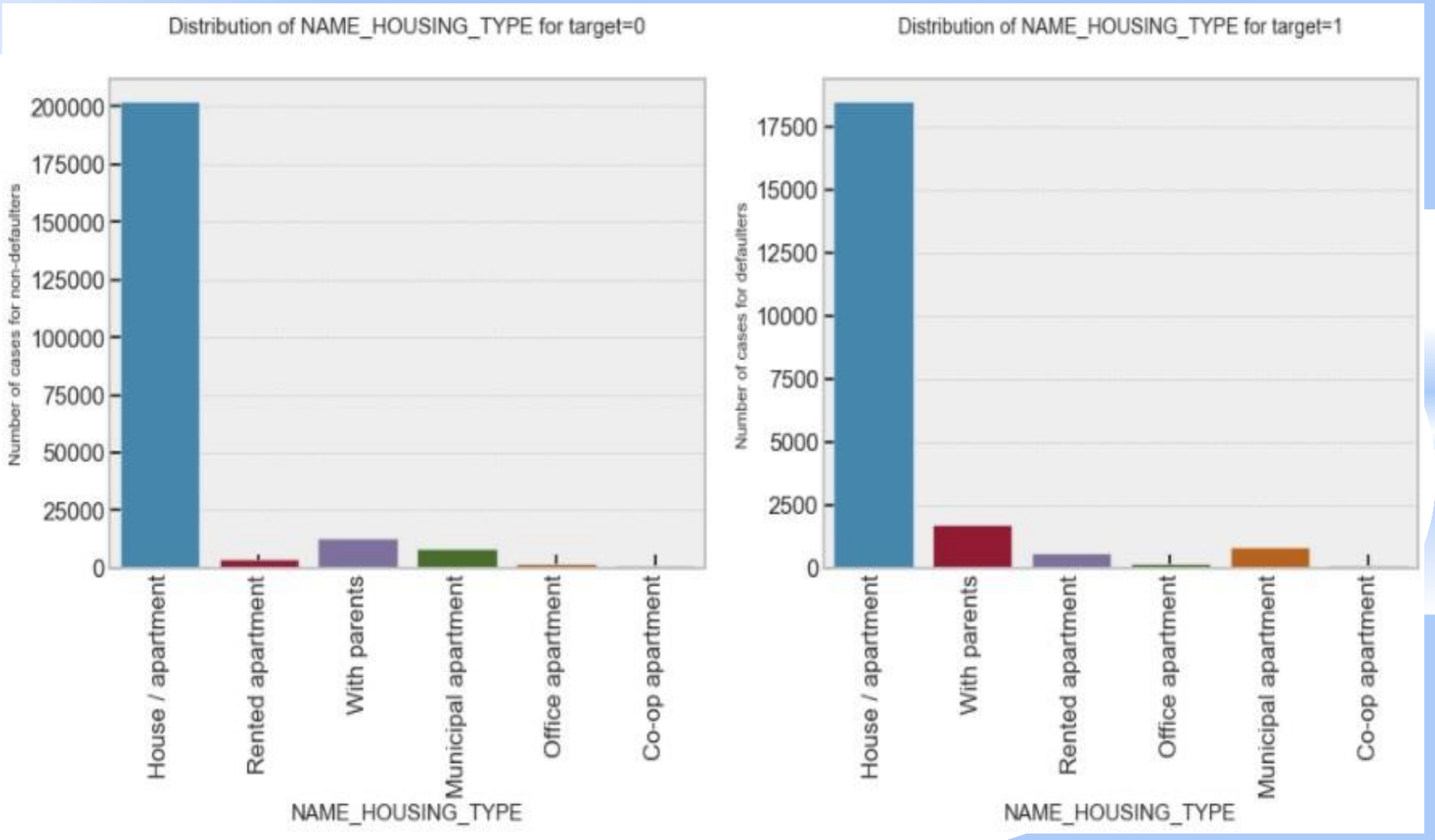
We can see the revolving loans has less distribution in number of cases for defaulters compared to non-defaulters.

This may be attributed to the Nature of revolving loan as it is considered a flexible financing tool due to its repayment and re-borrowing flexibility



Graph for Housing Type

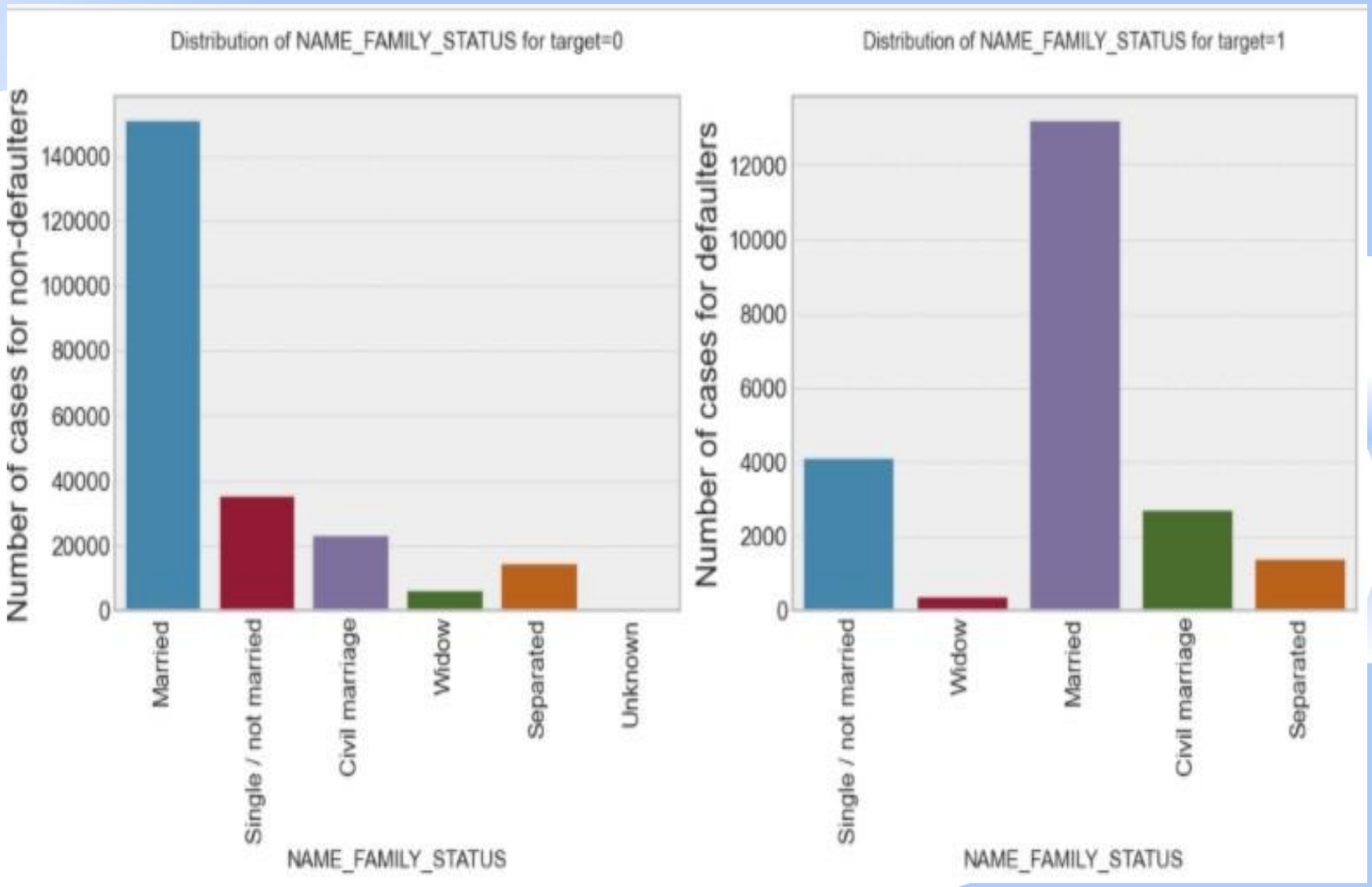
. People living with parents and in apartments show high number of defaulters as compared to non defaulters. The reason is that living with parents so less income and living in municipal apartments so more cash flow in apartment rent.



Graph for Family Status

. Single/not married category are showing high number of defaulters

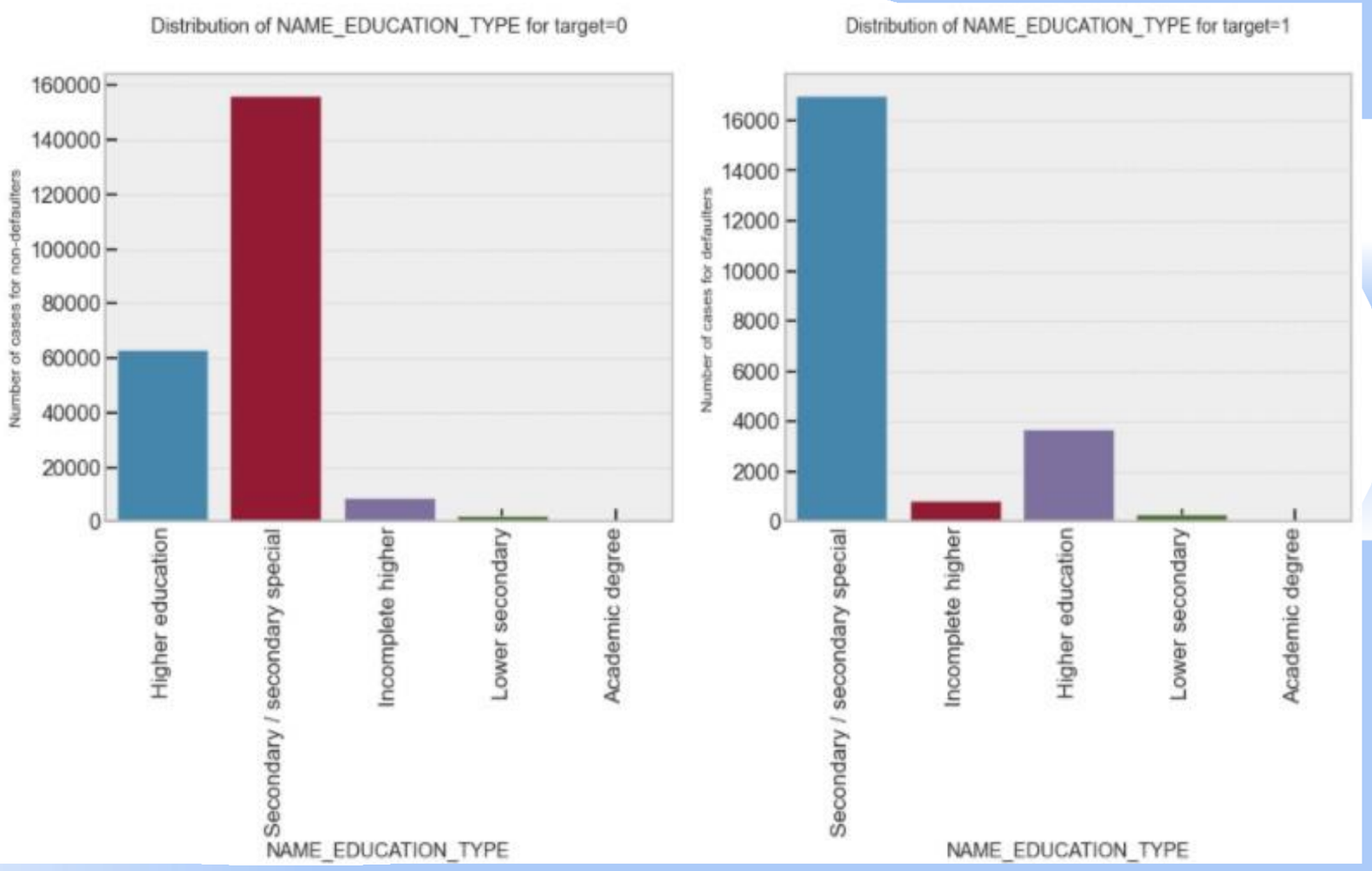
. Married category ranks the higher result in both defaulter and non-defaulter distributions.



Ordered Categorical Data

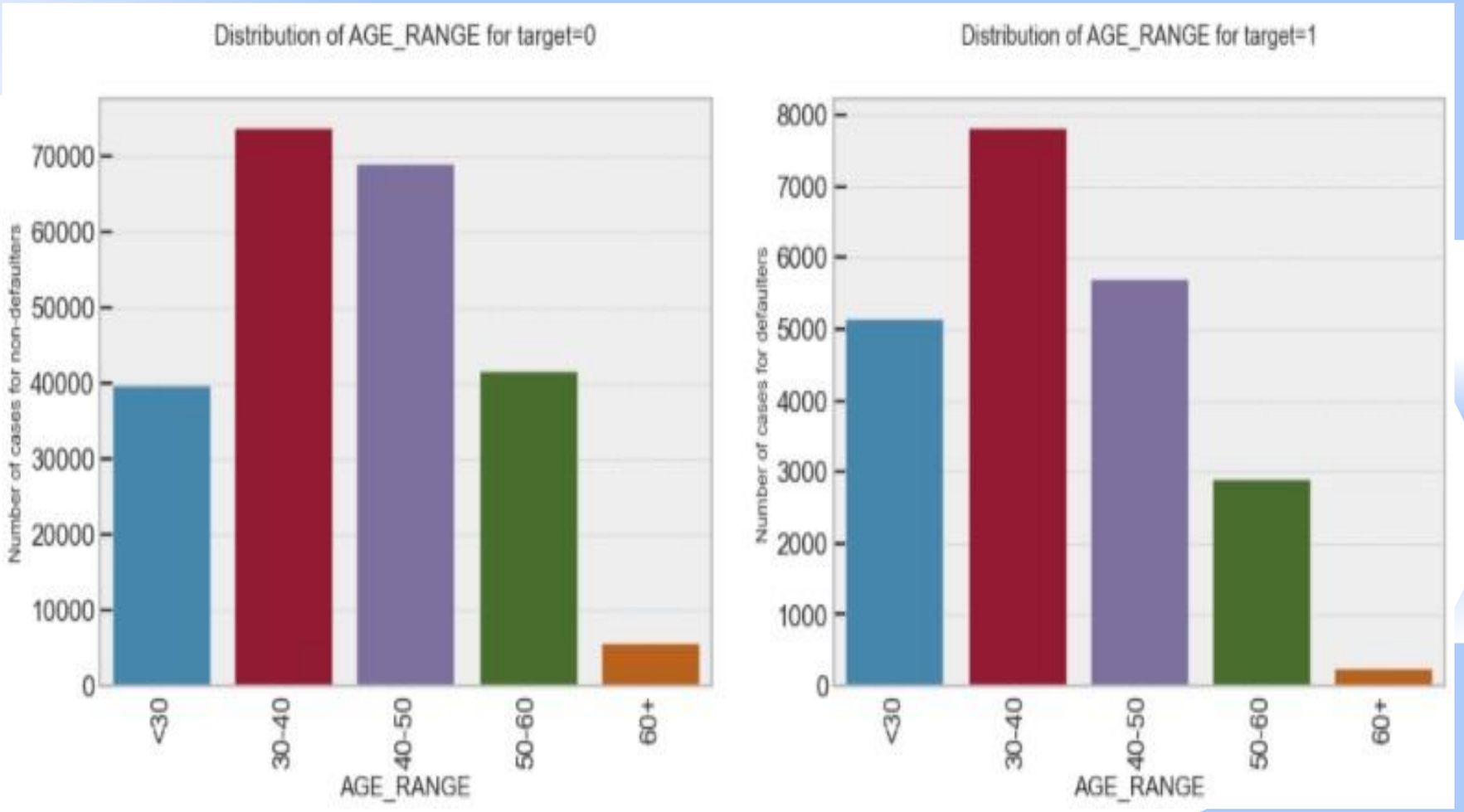
Graph for Education Type

. Higher education is more in non defaulters as compared to defaulters category. This comes to a conclusion that higher the education, more the individual is capable of paying the loan because of good salary package



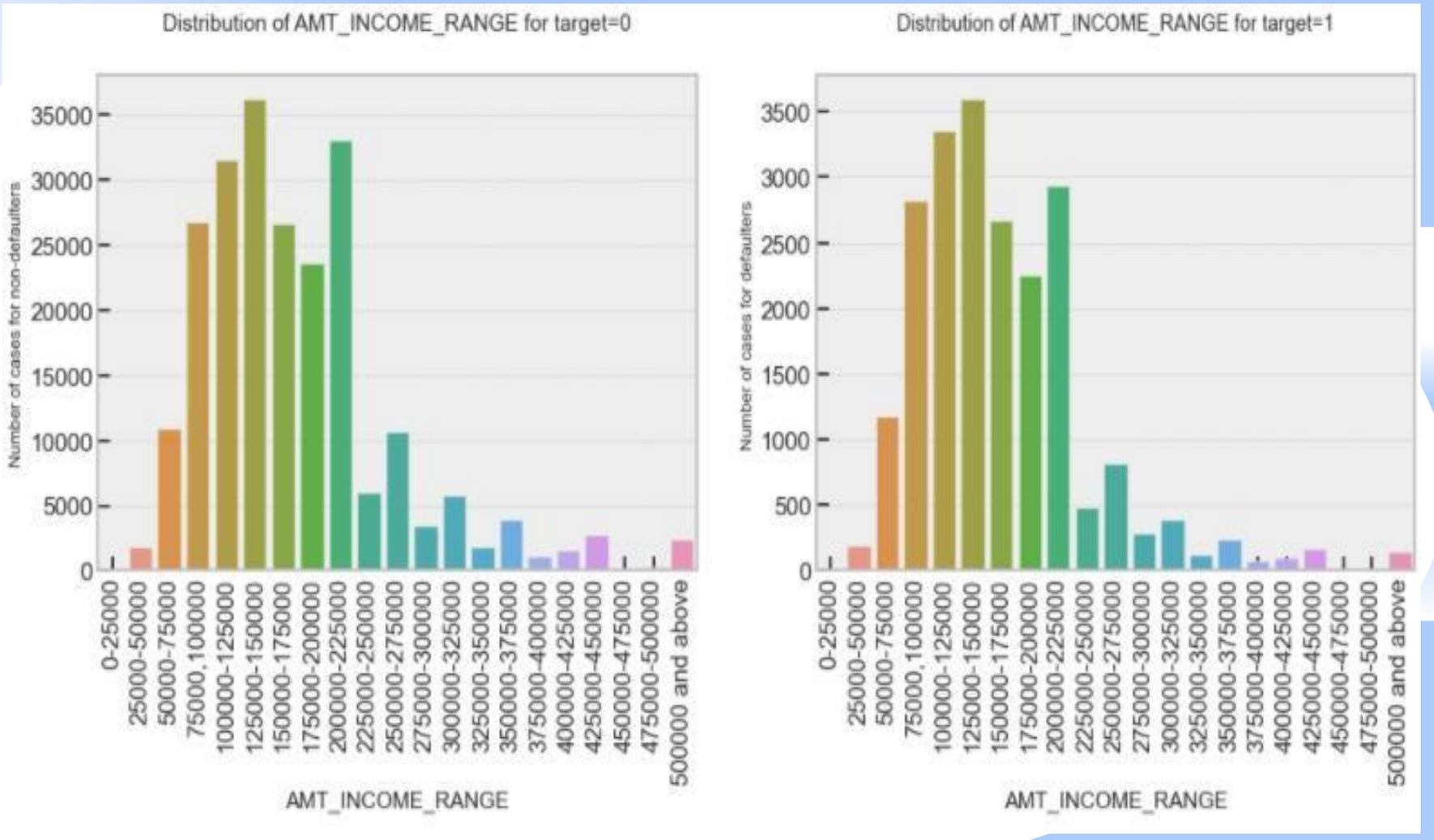
Graph for Age Range

. For age group '40-50' no of loan repayers decreased for defaulters



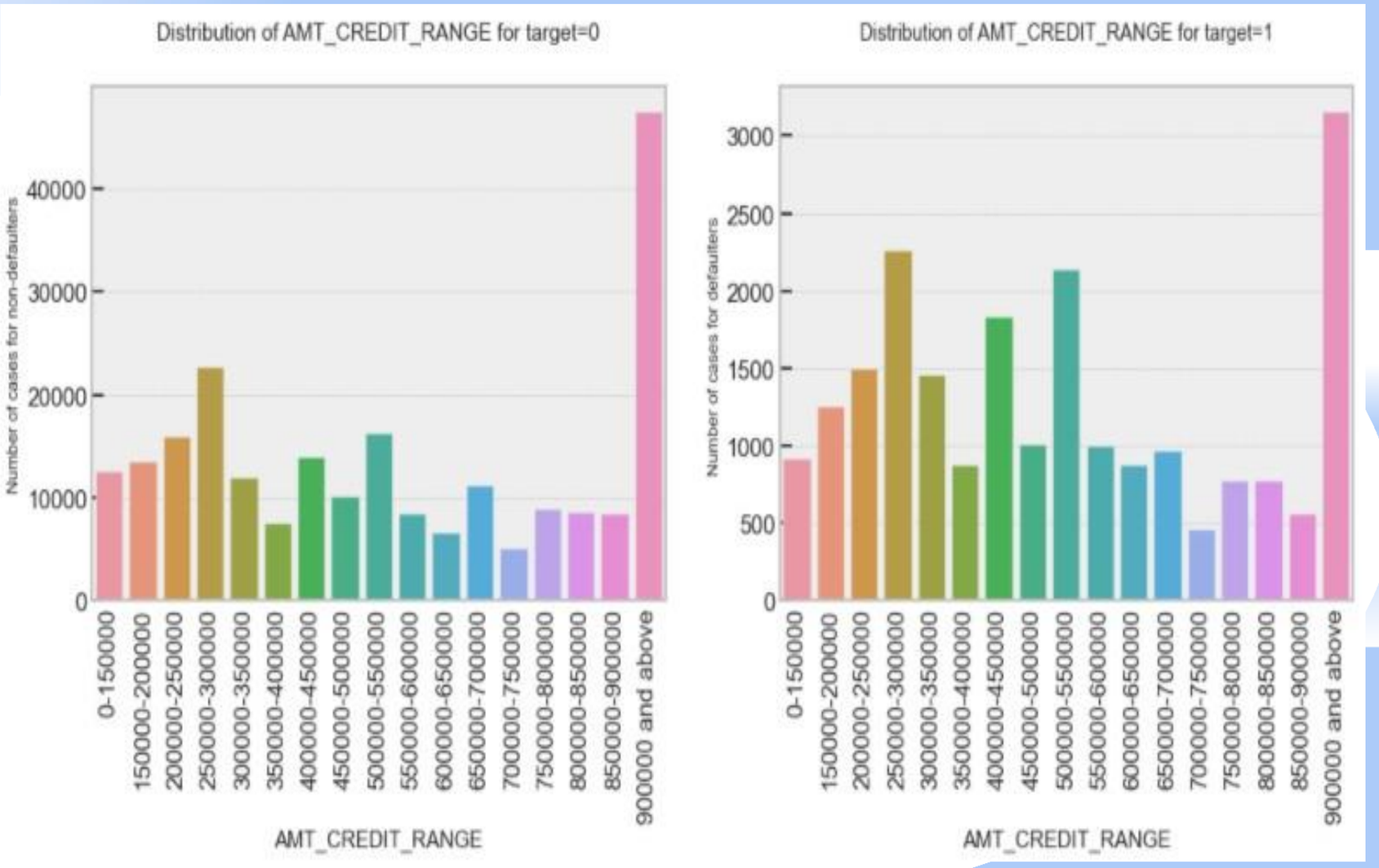
Graph for Income Range

. Income Range of 125000-150000 is maximum for both Loan Repayers & Loan Defaulters



Graph for Credit Range

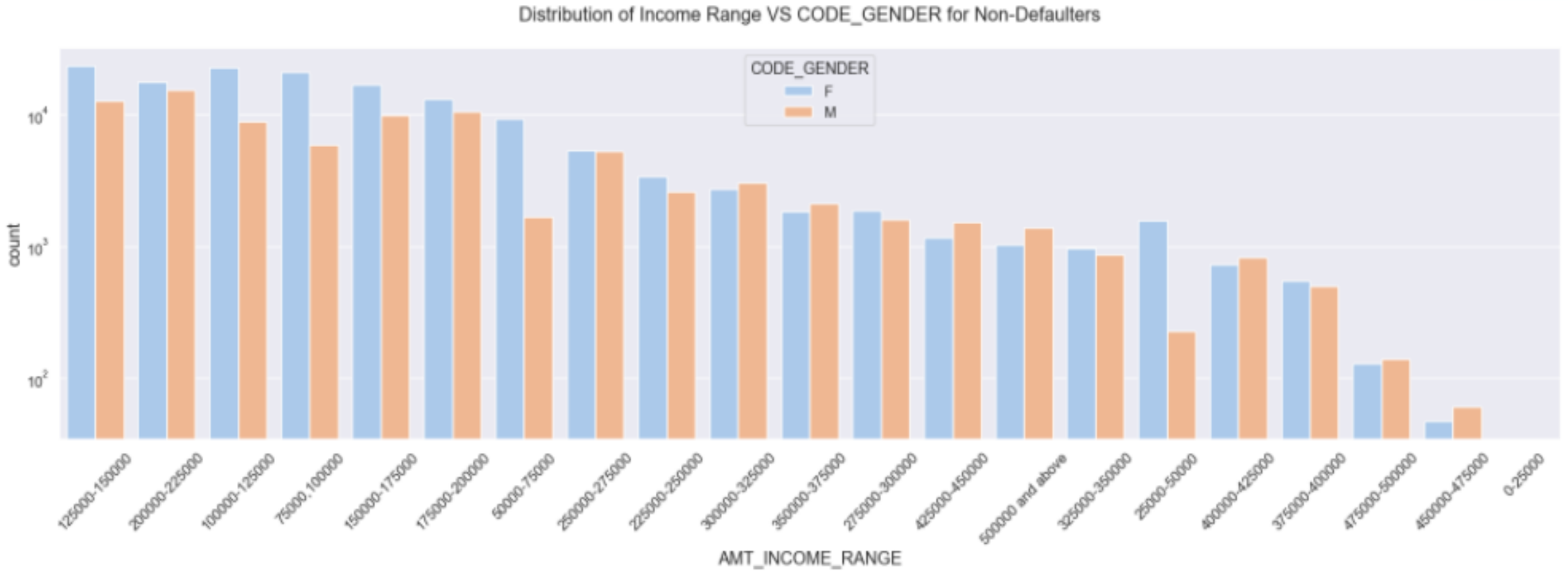
. Amount Credit Range value of 900000 and above is the highest ranked in both scenarios



. Numerical - Categorical

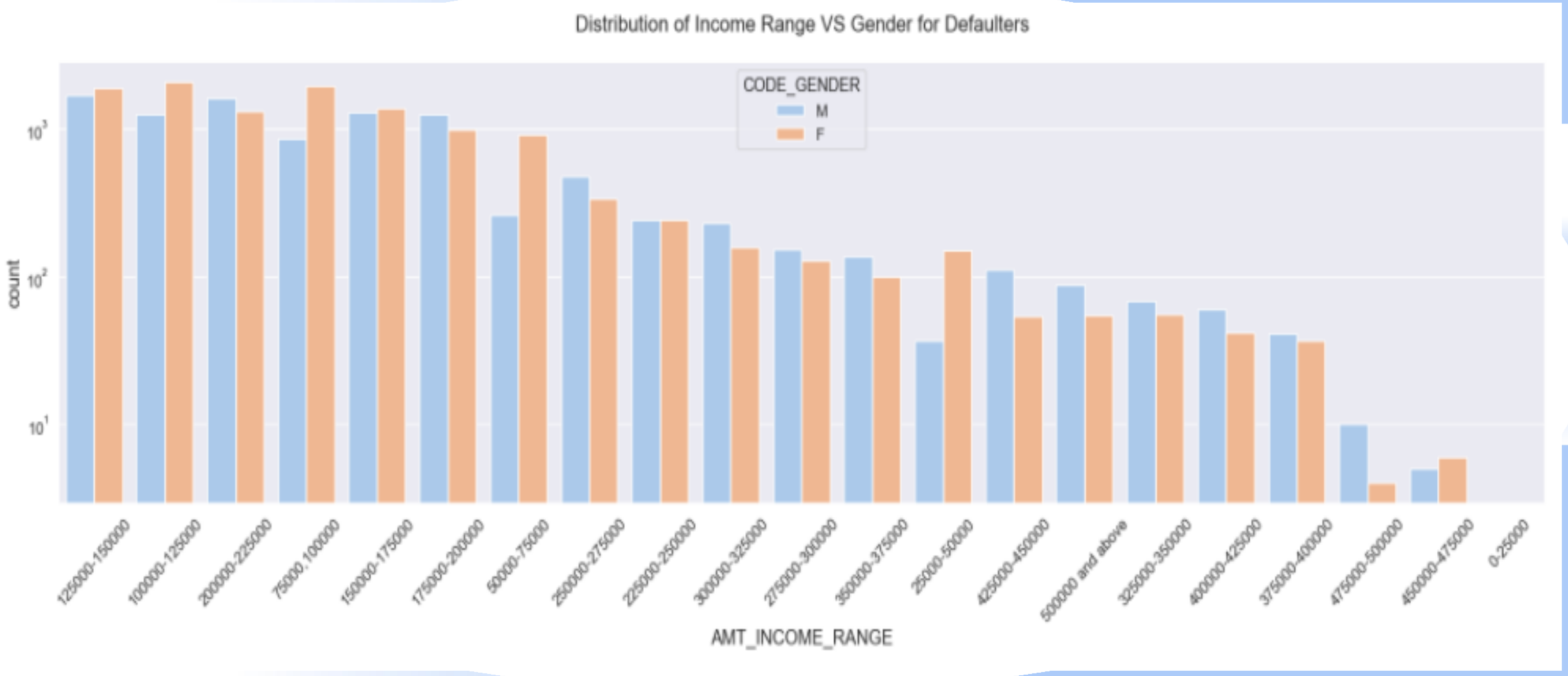
Graph for Income Range Type VS Gender for Non-defaulters

. Female numbers are mostly high in every proportion and also female credit are mostly higher than males(Loan Repayers)



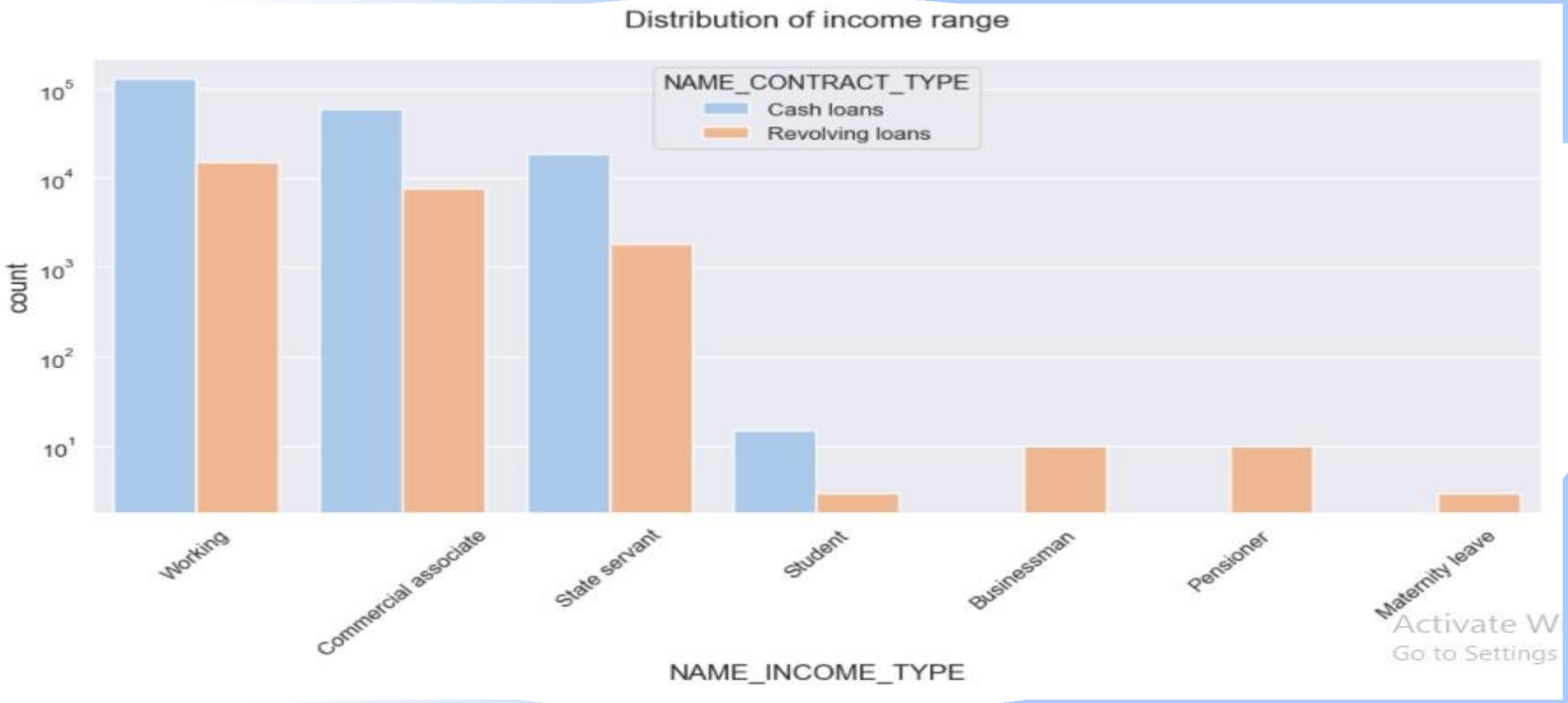
Graph for Income Range Type VS Gender for Defaulters

. **Male numbers are mostly high in every proportion and also Male credit are mostly higher than Females(Loan Defaults)**



Graph for Income Type VS Contract type for Non-defaulters

- **Working class has more number of Cash Loans**



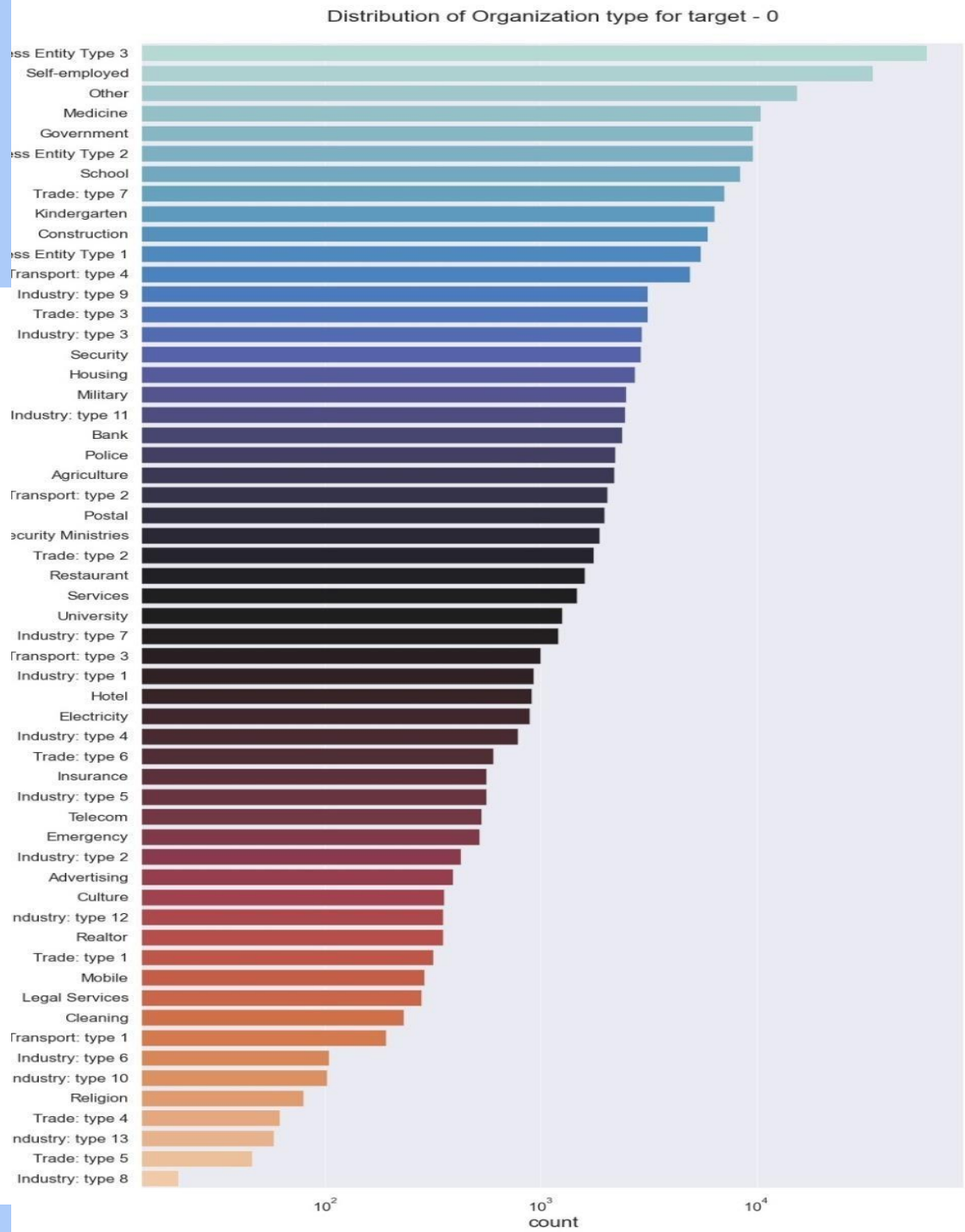
Graph for Income Type VS Contract type for Defaulters

- **Working class has more number of Cash Loans**

Distribution of Organization type for Non-defaulters

. ‘Business entity Type 3’ and ‘Self employed’ hold the most credits

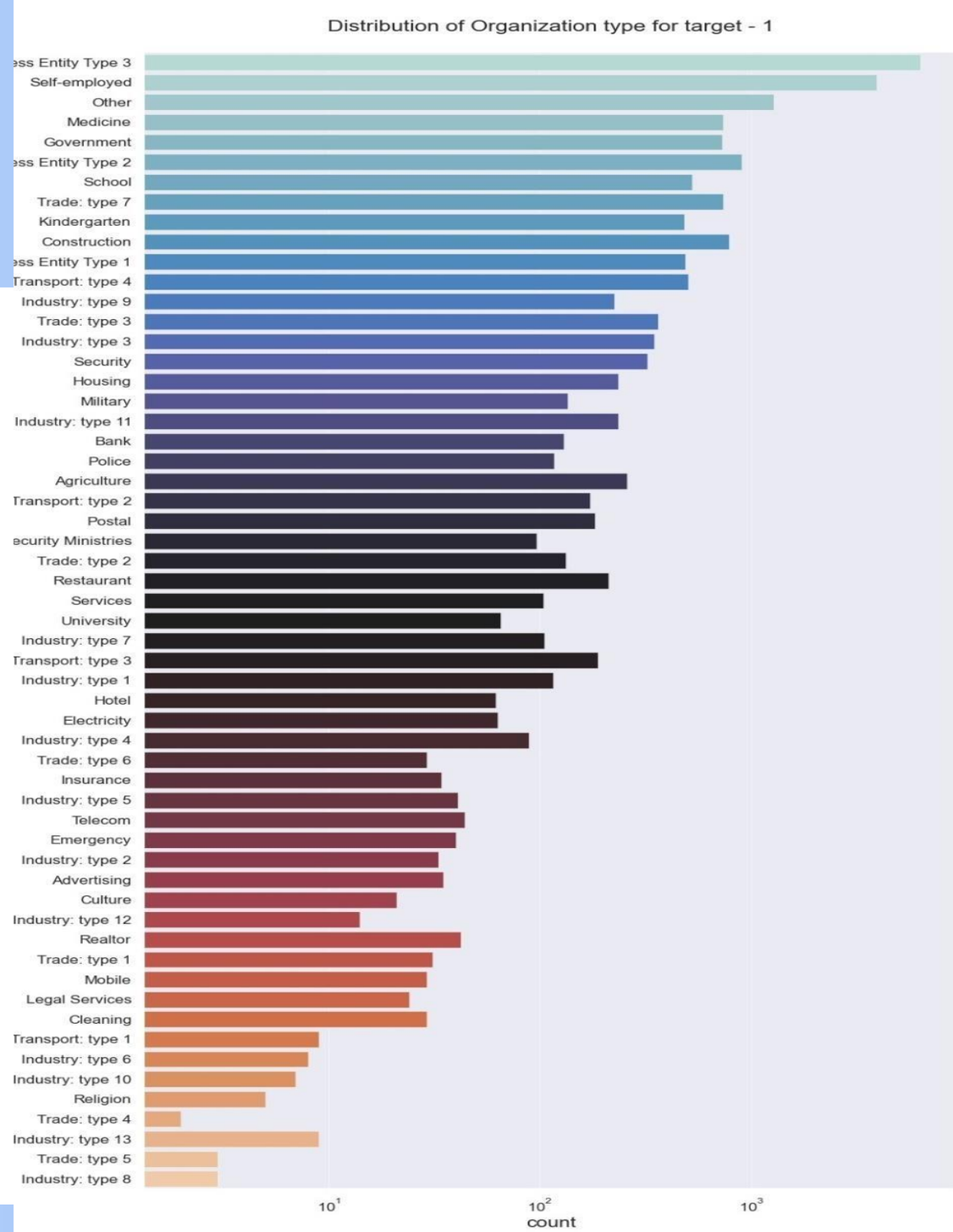
. Less credit to industry trade: type5 and Industry: type 8



Distribution of Organization type for defaulters

. Business entity Type 3’ and ‘Self employed’ hold the most credits

. Less credit to industry trade: type5 and Industry: type 8

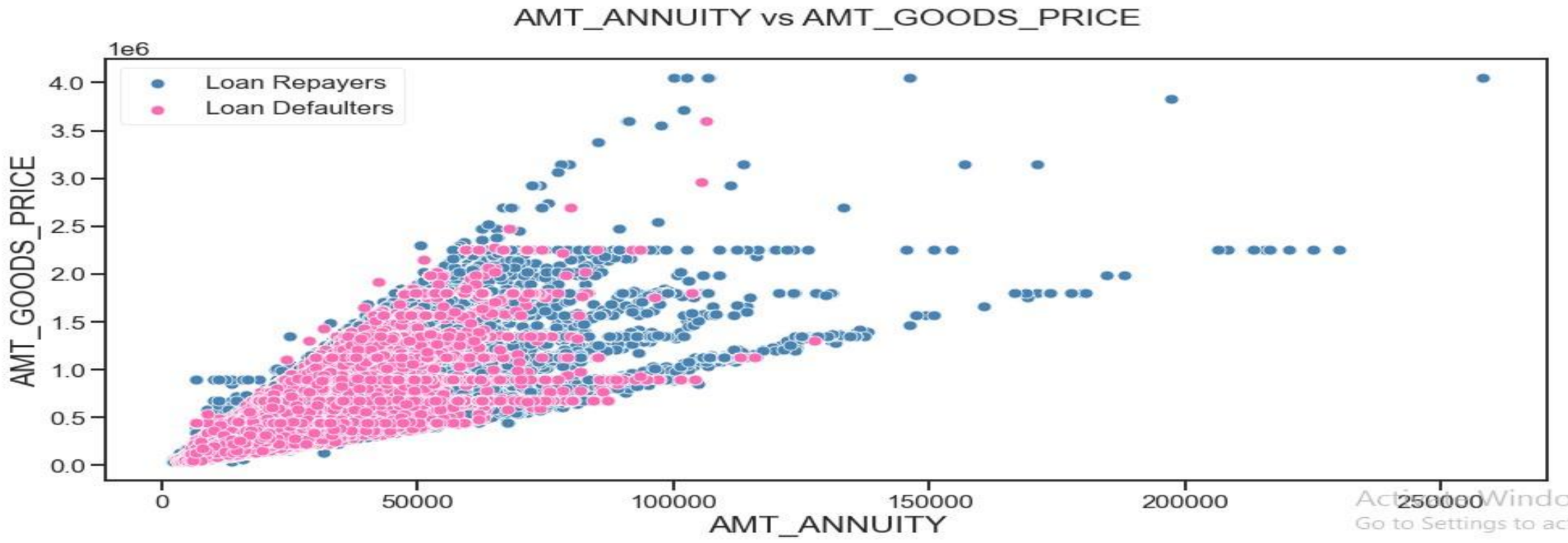


Bivariate Analysis

Numerical – Numerical Analysis

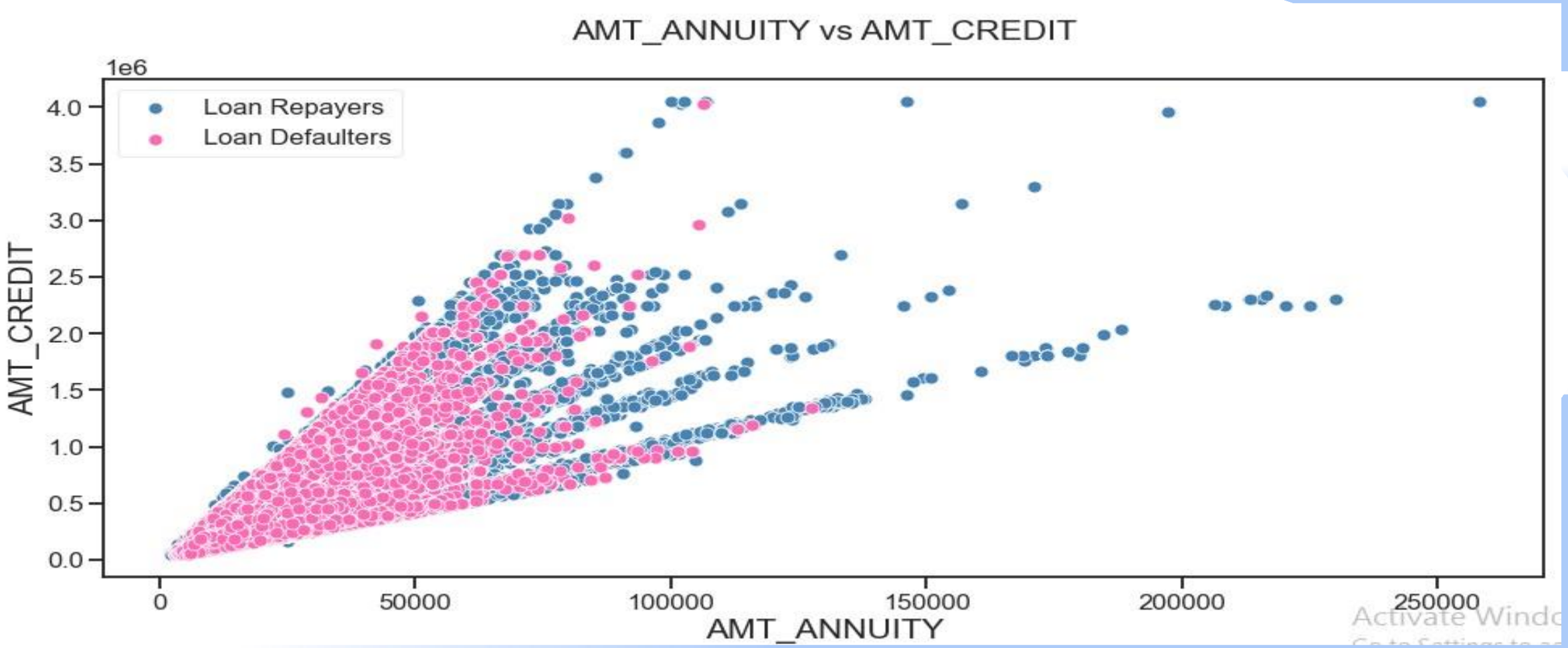
Graph for Amount Annuity VS Amount Goods Price

. Correlation between Amount Annuity & Amount Goods Price is pretty moderate but they are not thoroughly correlated because there are above par values for both the columns. Amount Annuity having values less than 70000 are prone to be in default while values above 70000 are tend to decrease in defaulters.



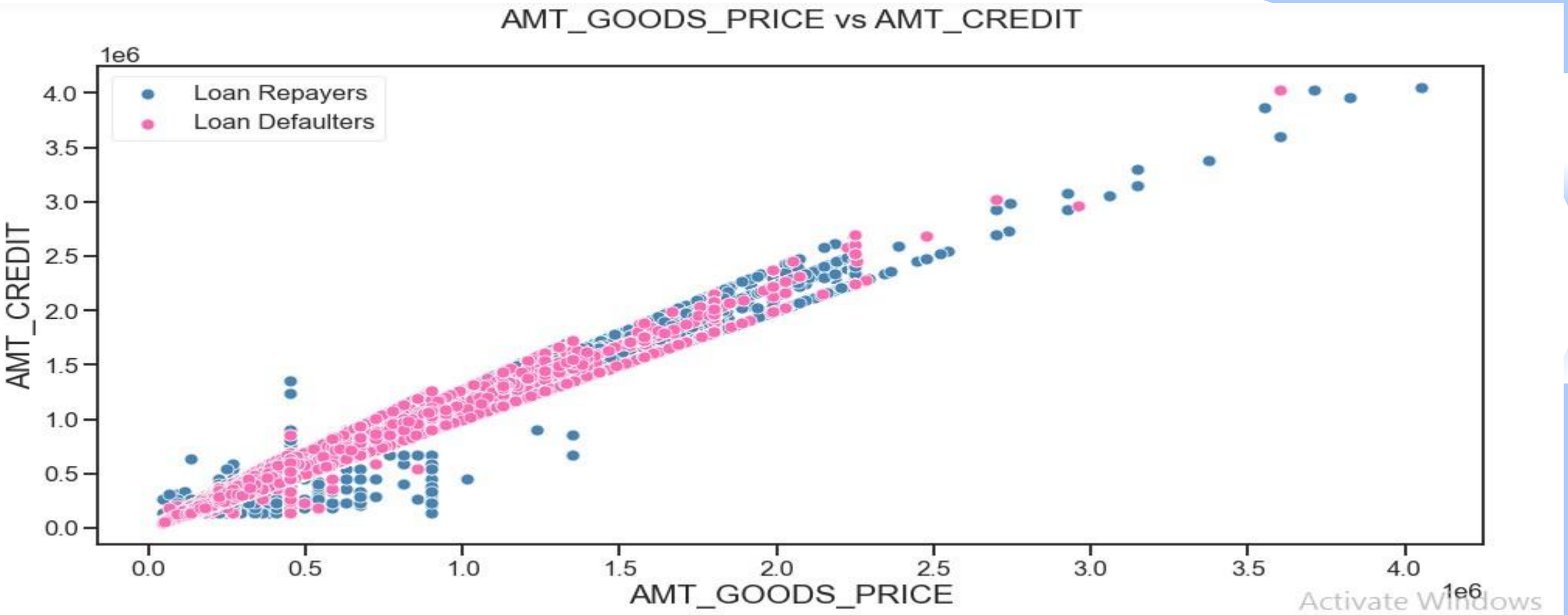
Graph for Amount Annuity VS Amount Credit

. There is fair correlation between the two columns Amount Annuity & Amount Credit. There's a decrease in defaulters when Amount Annuity increases. Most of the defaulters are having annuity values less than 60000



Graph for Amount Goods Price VS Amount Credit

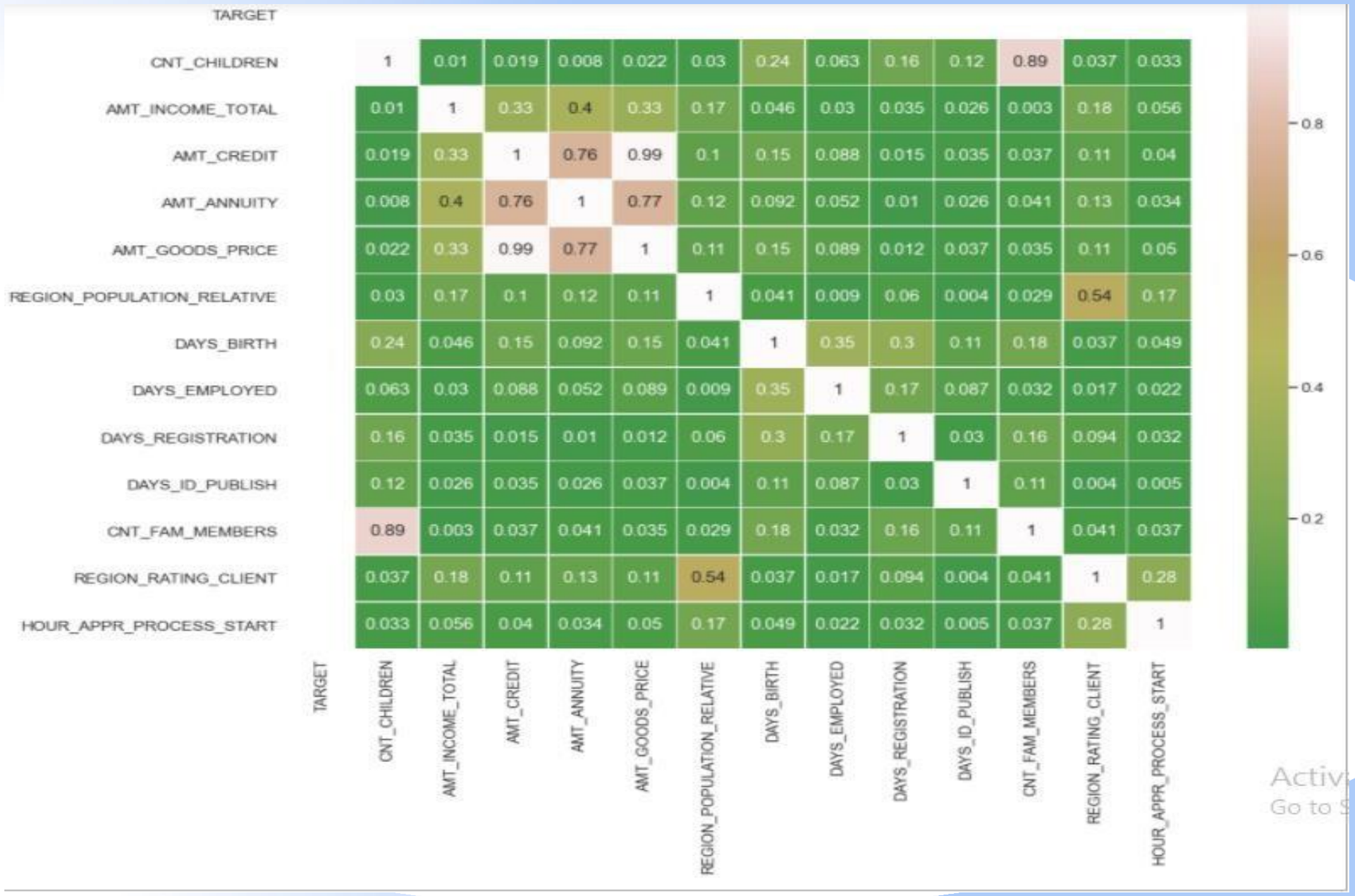
. Amount Good Price & Amount Credit have a strong correlation as clients having goods price and will repaying their loans.



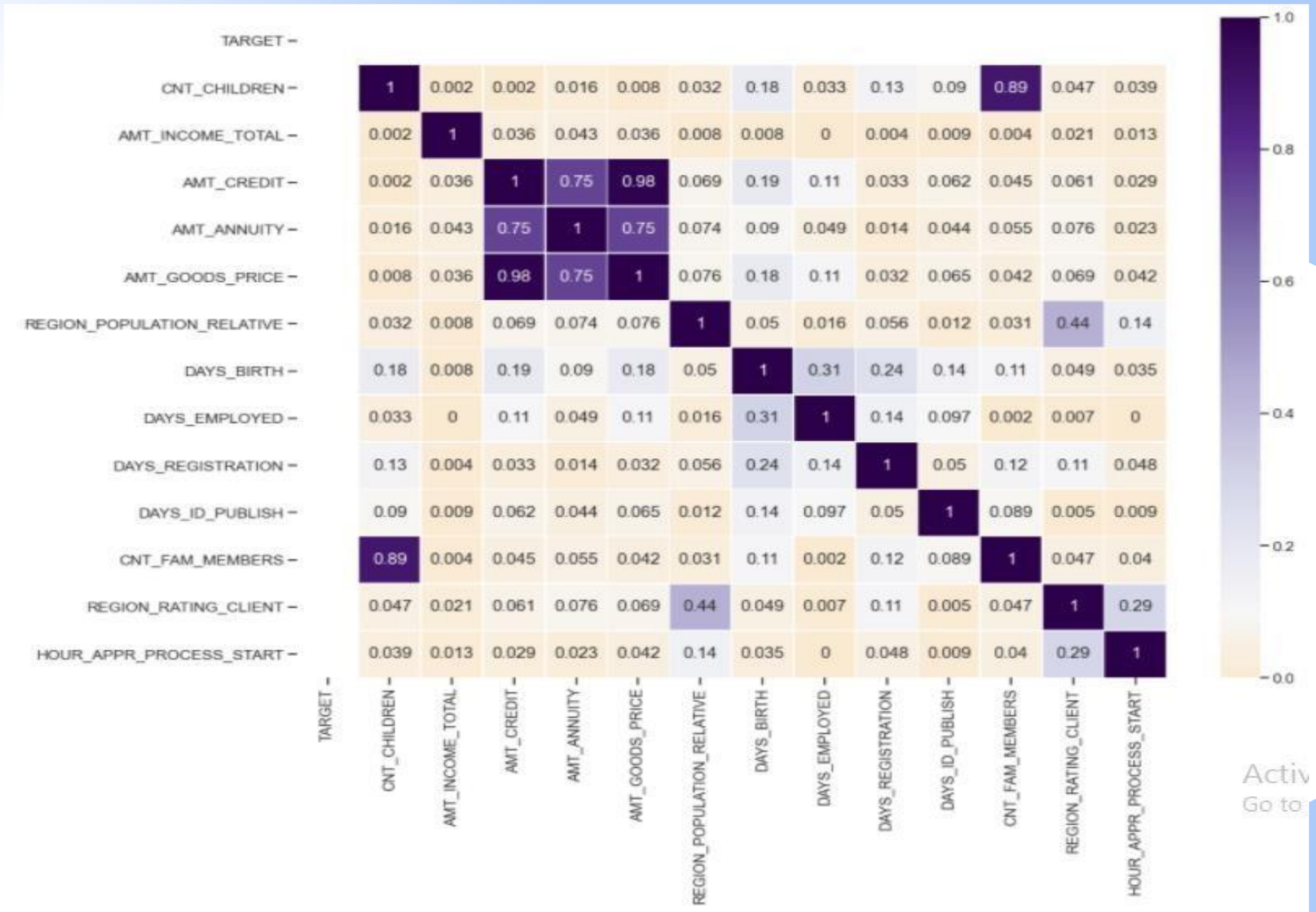
Activate Windows

Correlation or Multivariate Analysis

Heat Map to show Correlation for Non-defaulters



Heat Map to show Correlation for Defaulters



Insights from Correlation

. For target variable there is no correlation as there are empty space in the graph and NAN in the tables while comparing with other variables

. Credit amount is highly correlated with amount of goods price which is slightly different from non-defaulters i.e. Loan Repayer's

. The correlation is strong between family member and children counts, although the correlation increases for the defaulters

. The loan annuity correlation with credit amount and also with goods price has slightly reduced in defaulters(0.748) when compared to repayers(0.777)

. We can also see that repayers have high correlation in number of days employed(0.62) when compared to defaulters(0.58)

. Days birth and number of children correlation has reduced to 0.256 in defaulters when compared to 0.336 in repayers

Merged Dataset

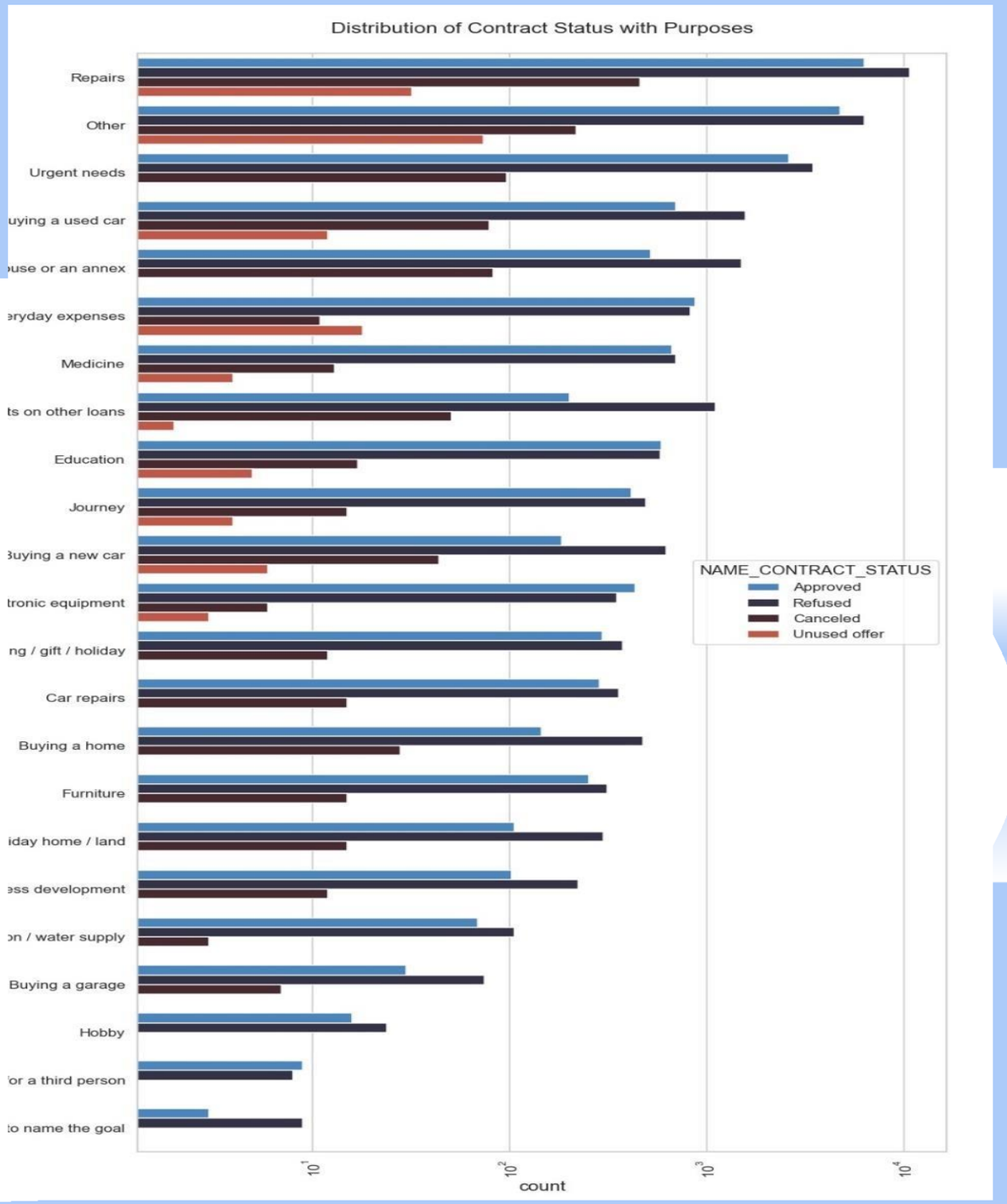
Univariate Analysis

Distribution of Cash Loans with Contract Purpose

. Repairs got most refused loans

. Education has similar outcomes for approval and rejection of loans

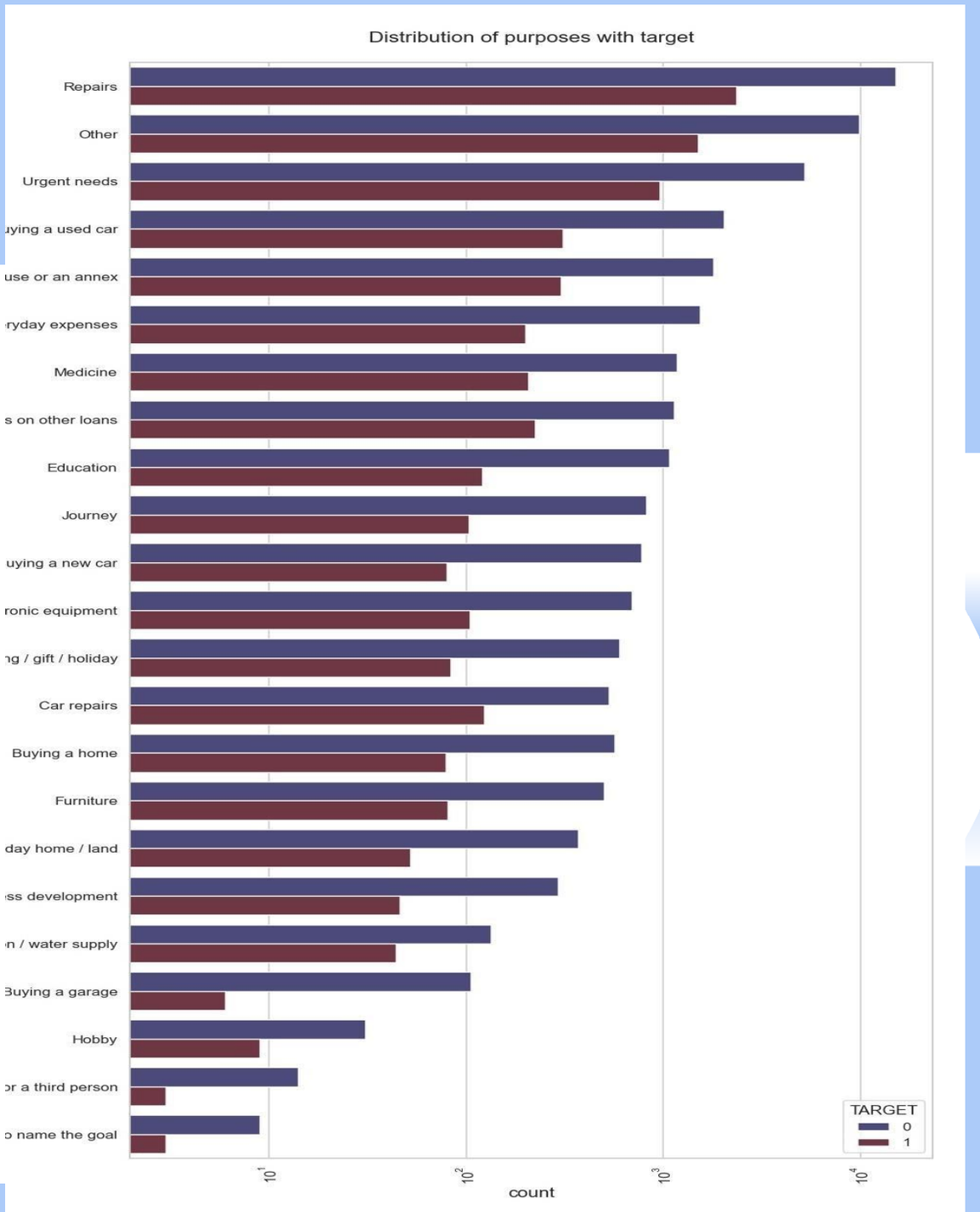
. Paying other loans and buying a new car is having significant higher rejection than approvals.



Distribution of purposes with Target

. Repairs are dealing with more difficulties in payment on time

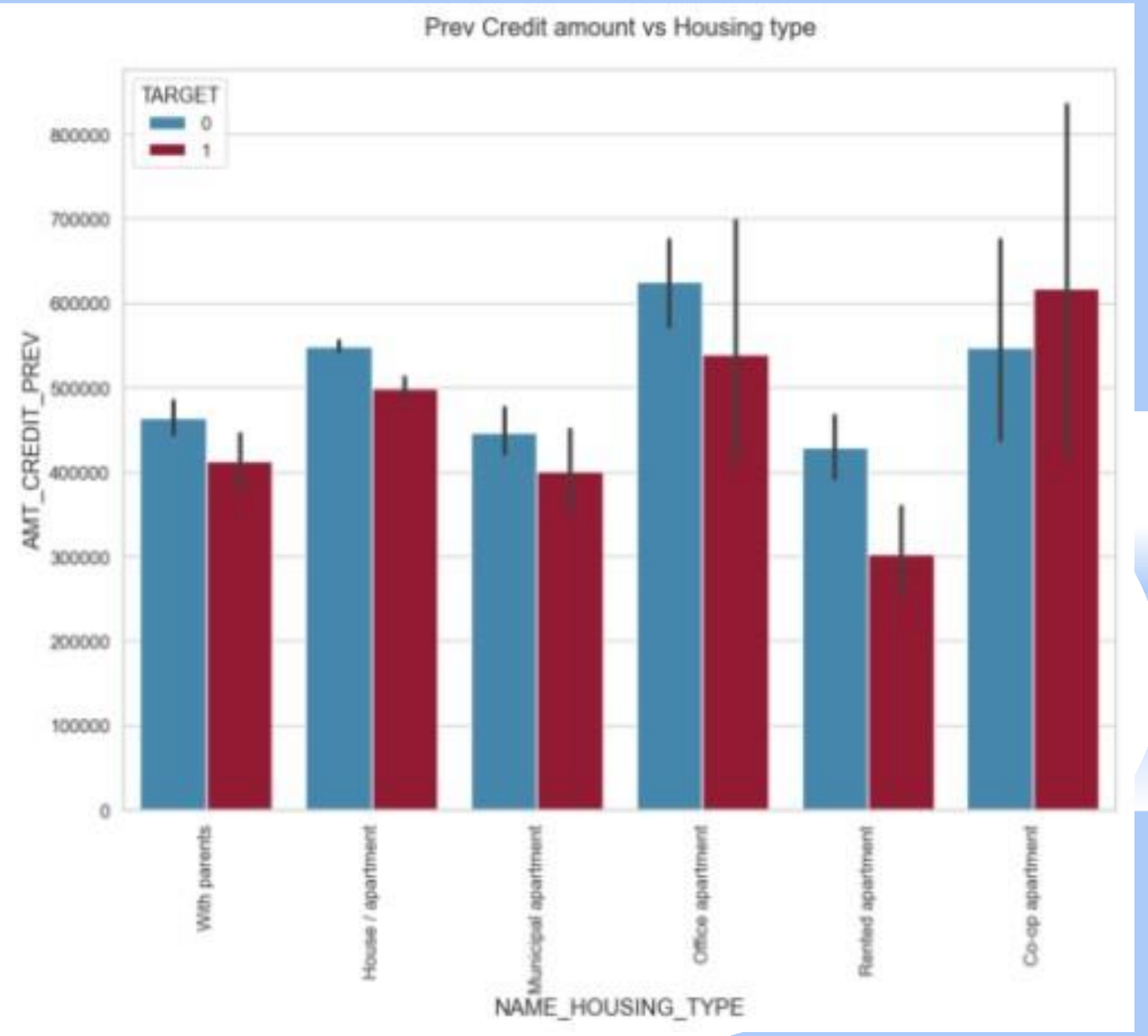
. Buying a garage, Business development, Buying land, Buying a new car and Education categories having basically higher loan payment



Bar plotting for Previous Credit Amount vs Housing type

. Office apartment is having higher credit of non-defaulters and co-op apartment is having higher credit of defaulters

. Bank can focus mostly on housing type categories like ‘with parents’ or ‘house/apartment’ or municipal apartment for successful payments



Conclusion and Recommendations

. Most number of unsuccessful payments are done from loan purpose 'Repair'.

. Very few unsuccessful payments are incorporated by housing type 'With Parents', so bank should adhere most clients from this type.

. Income Type 'Working' shows very less promise to the bank with most number of unsuccessful payments so bank should try to avoid them'

. For housing type 'Co-op apartment' ,bank should avoid giving loans as they are having difficulties in payment

Thank you