EDA CASE STUDY

Presented by: Rudri Dave

Problem Statement:

- This case study of analysing credit risks aims in helping credit card companies make wiser decisions regarding loan approvals based on an applicant’s profile.

- Such a detailed analysis, will help the company to determine potential loan applicants and also avoid financial losses by identifying applicants that are not likely to repay their loans.

- This case study will use Exploratory Data Analysis to analyse the datasets and help company in making a better business decision.

EDA STEPS

- Inspecting the datasets
- Handling missing values, nulls and outliers
- Checking for data imbalance
- Performing univariate analysis
- Performing bivariate analysis
- Combining two datasets
- Performing data analysis on the merged dataframe
- Conclusions and Recommendations

Methodolody Used

- Here, we have two data frames which we have used for data analysis.

- application_data.csv is the first dataframe which we have used as df1 in our analysis.

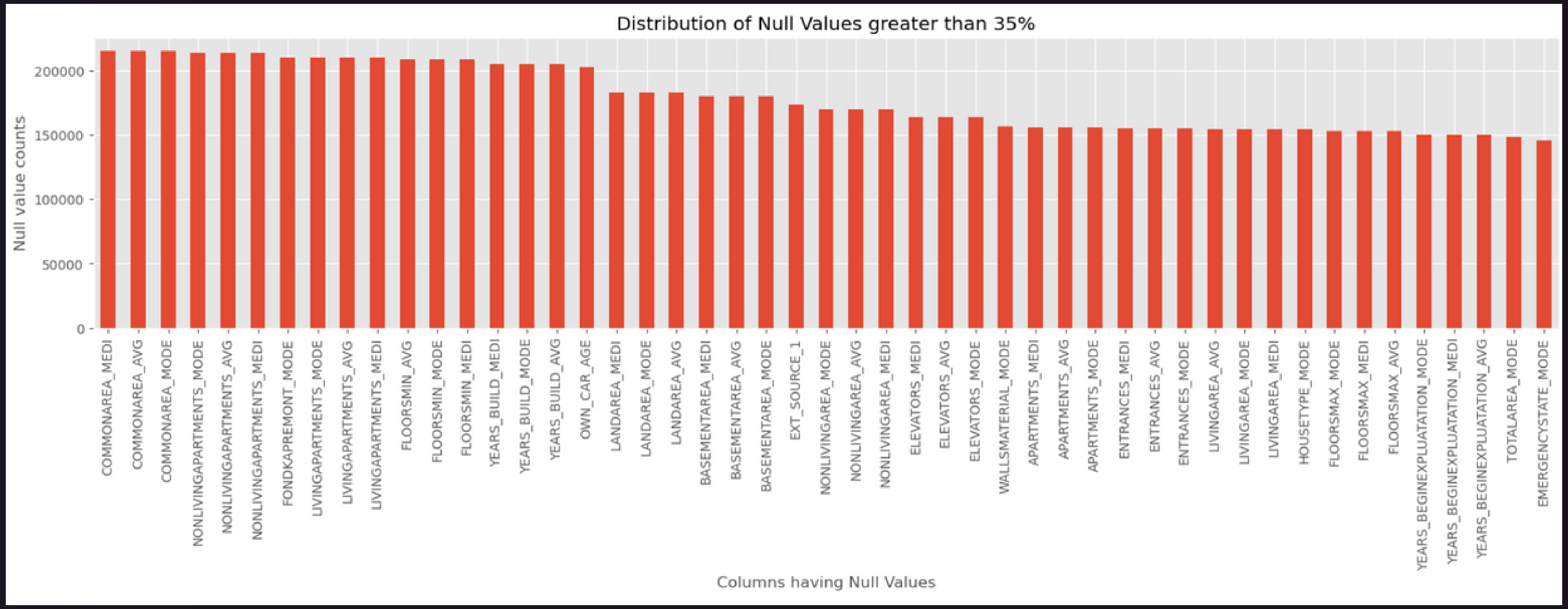
- previous_application.csv is the second data frame which we have used as df2 in our analysis.

- Post cleaning both the data frames we have merged them and have referred this newly merged data frame as df3.

Data Cleaning

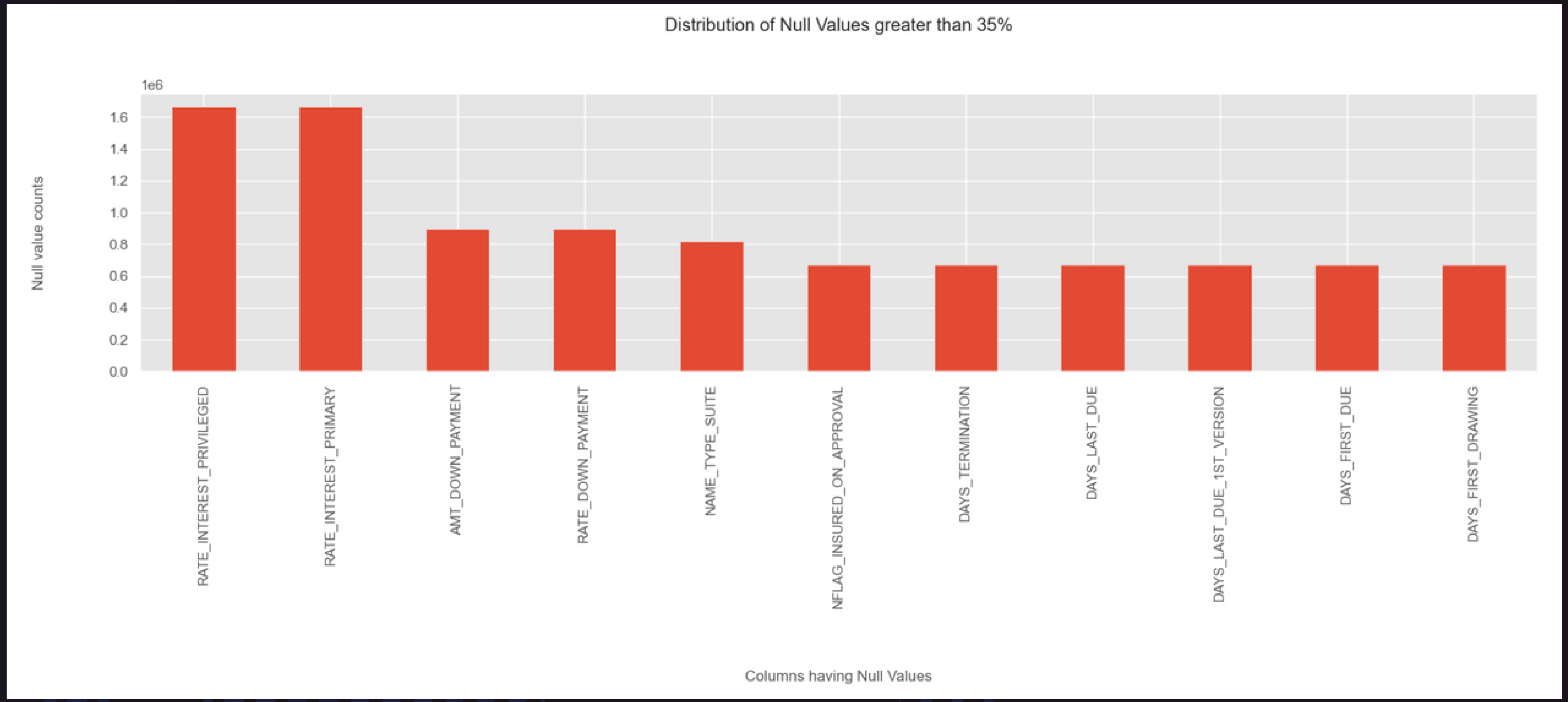
- The first plot, shows that there are 49 columns with null values greater than 35% in the first data frame application_data.csv.

- We have dropped these 49 columns with null values.



- The second plot, shows that there are 11 columns with null values greater than 35% in the second data frame previous_application.csv.

- We have removed these 11 columns with null values from second dataframe.

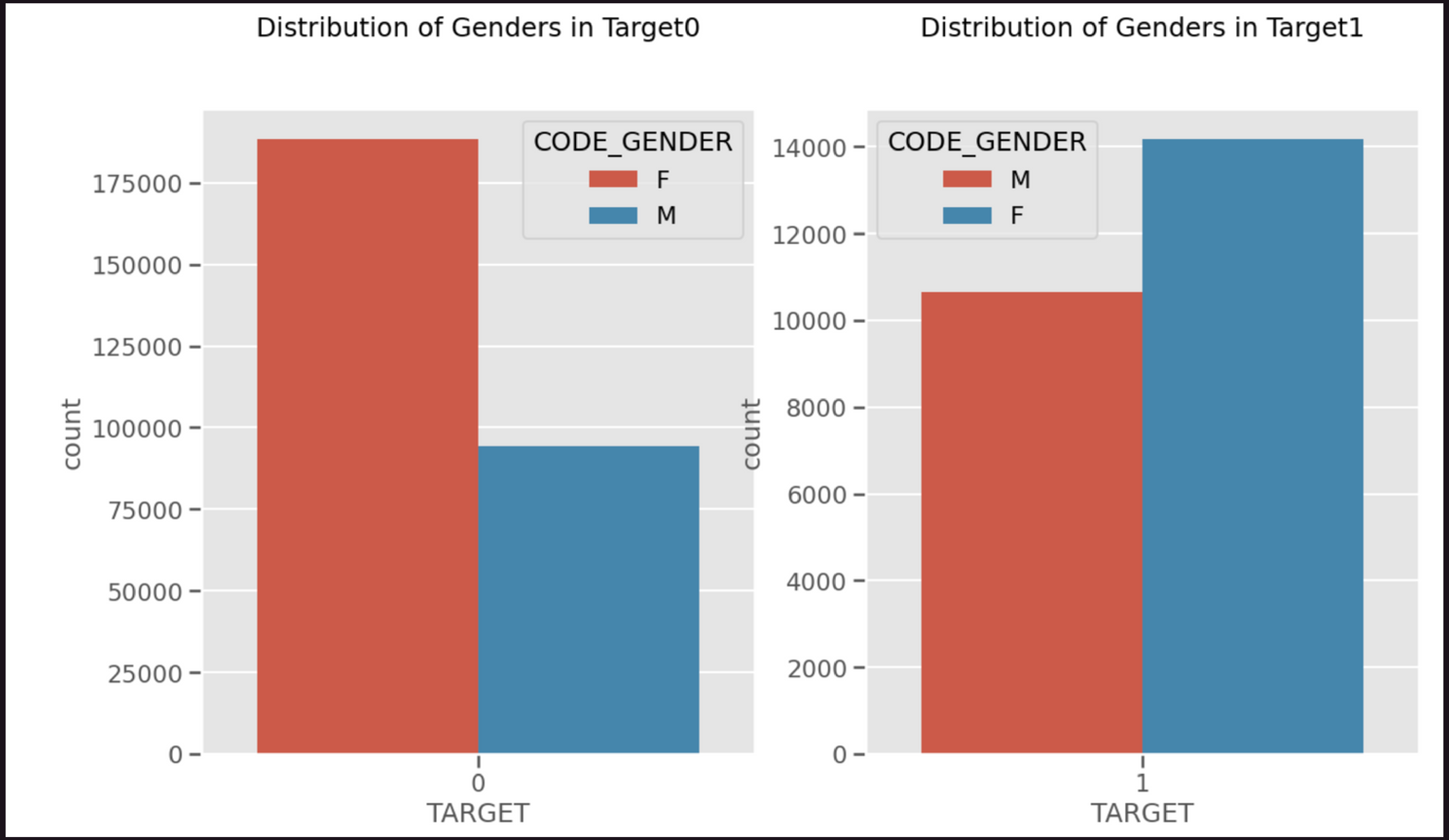


YUnivariate Analysis for Gender Distrubution wrt Target0 and Target1

paragraph text

Insights:

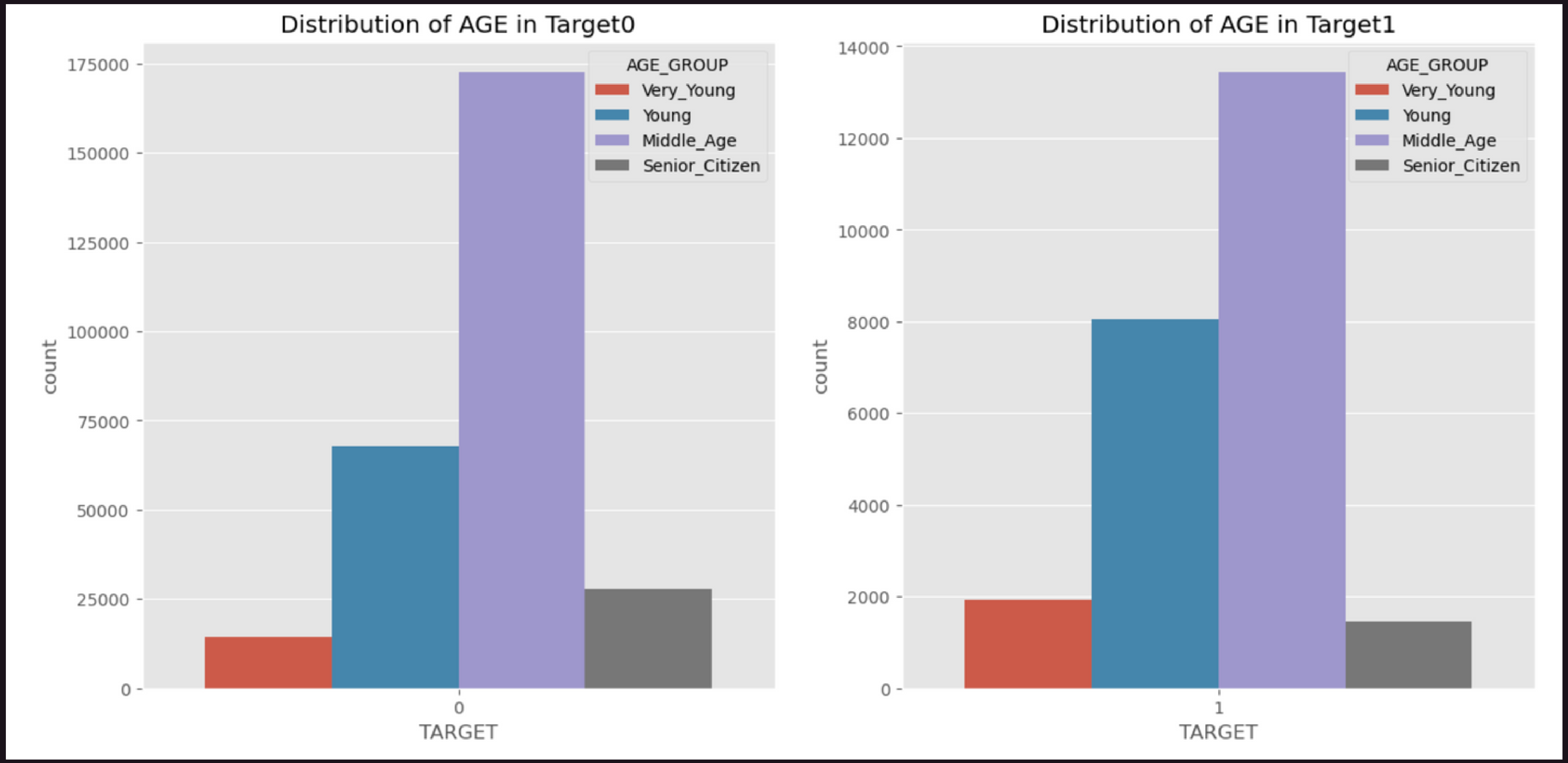
- From the distribution, we can see that more females applied for loans than males.
- Around 56% females are defaulters, while 42% males are defaulters
- Around 66% females are non-defaulters, whereas 33% males are non-defaulters



Univariate Analysis for Age Distribution wrt Target0 and Target1

Insights:

- From the distribution, we can see that the middle age groups have applied highest for loans than any other age groups.
- In addition, middle age groups is facing most difficulties.
- While, senior citizens and young age groups are facing less difficulties as compared to other groups.



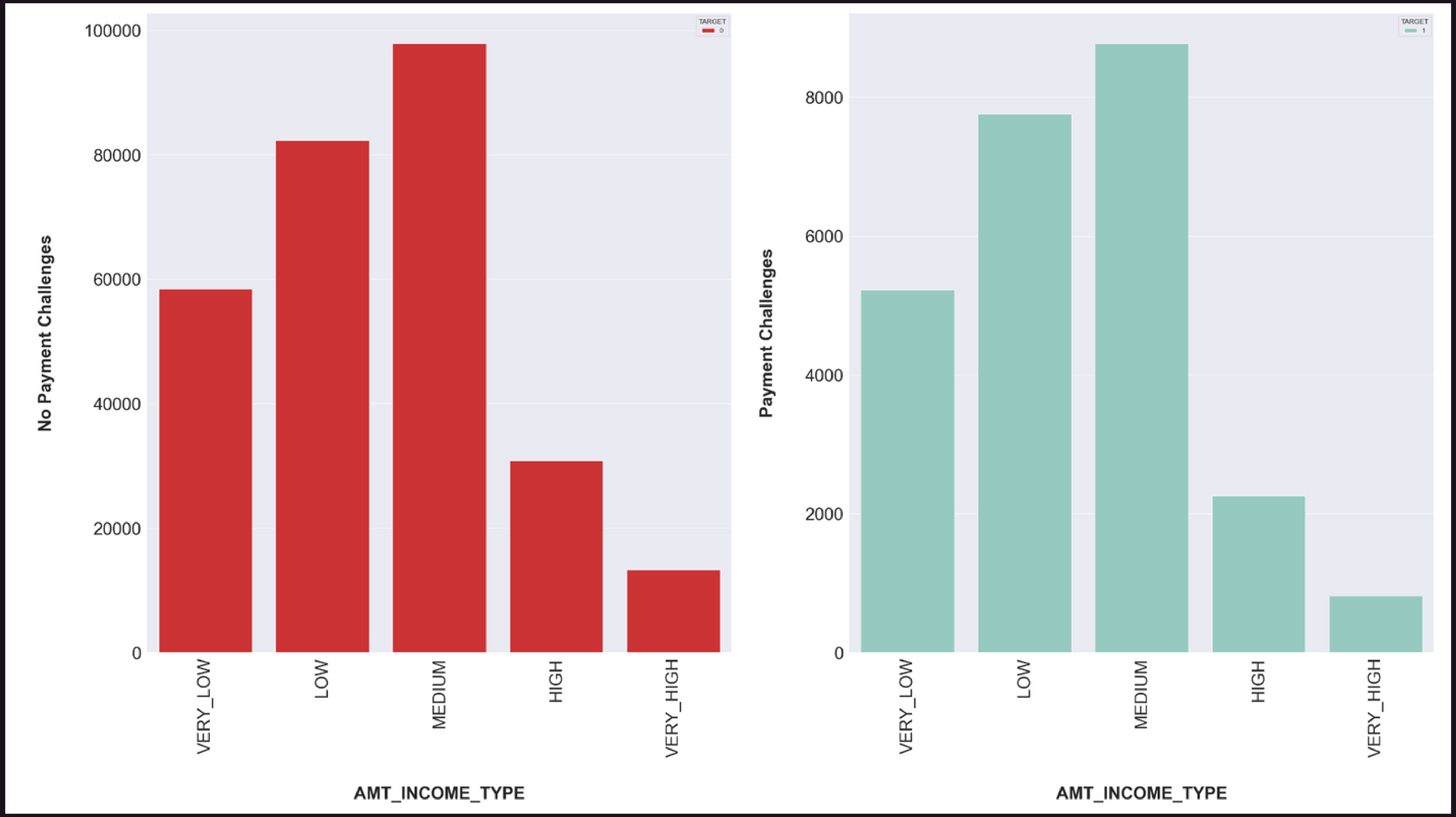
Univariate Analysis for Categorical Columnsparagraph text

Insights:

For AMT_INCOME_TYPE:

- People having medium salary ranges are more likely to apply for both defaulters and non defaulters.

- People having low income are at a higher risk of defaulters.

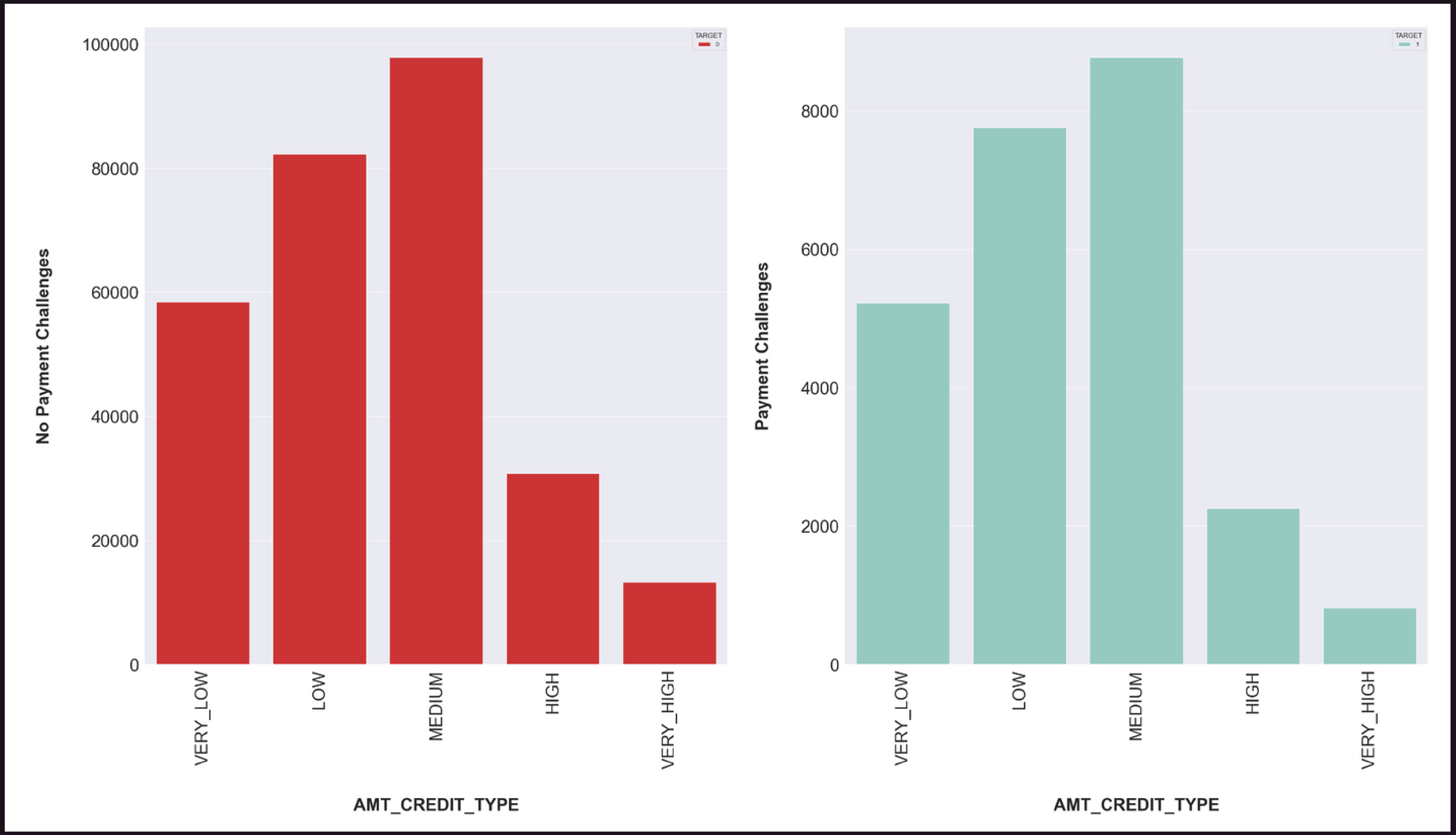


Univariate Analysis for Categorical Columnsparagraph text

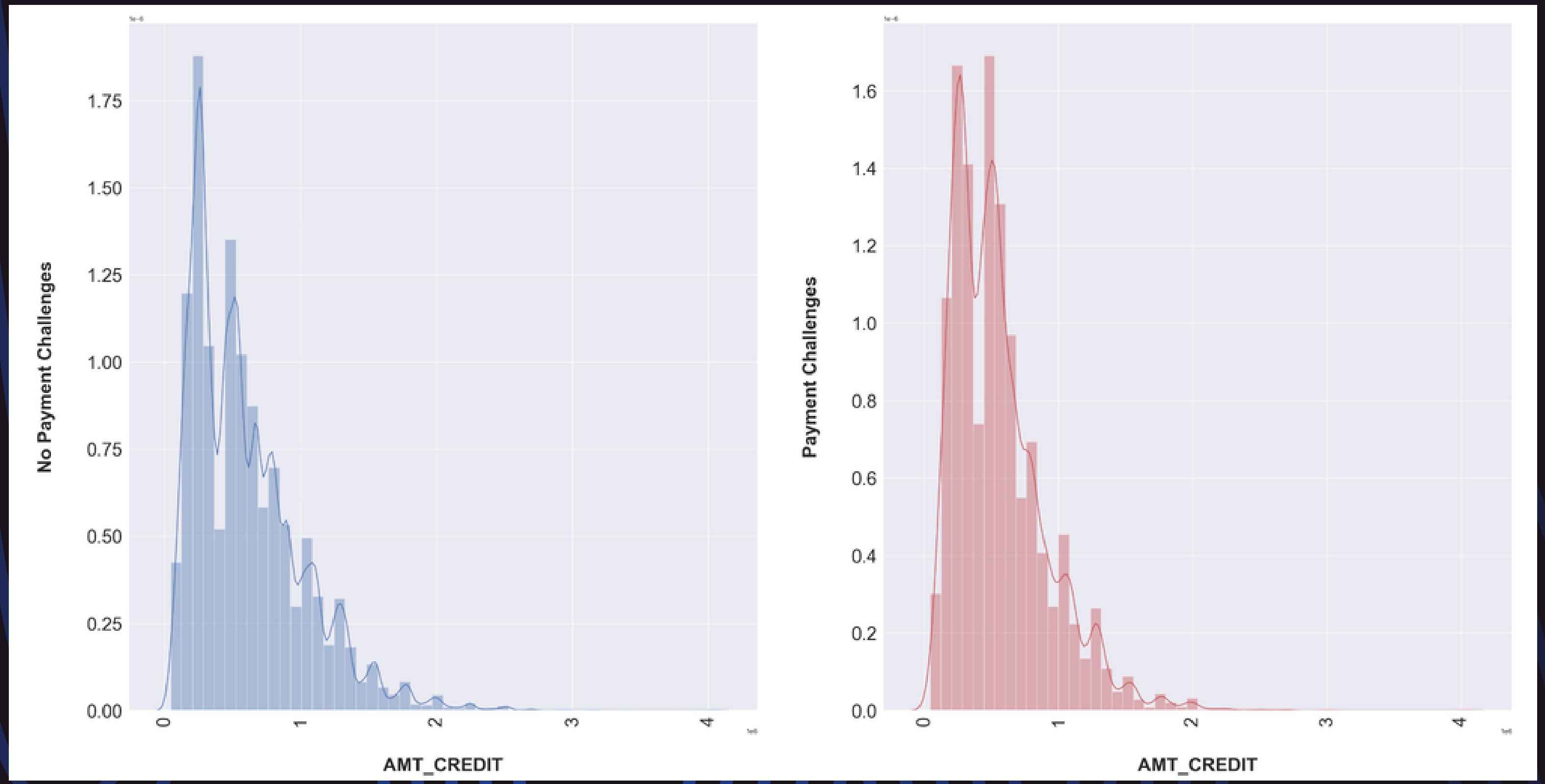
Insights:

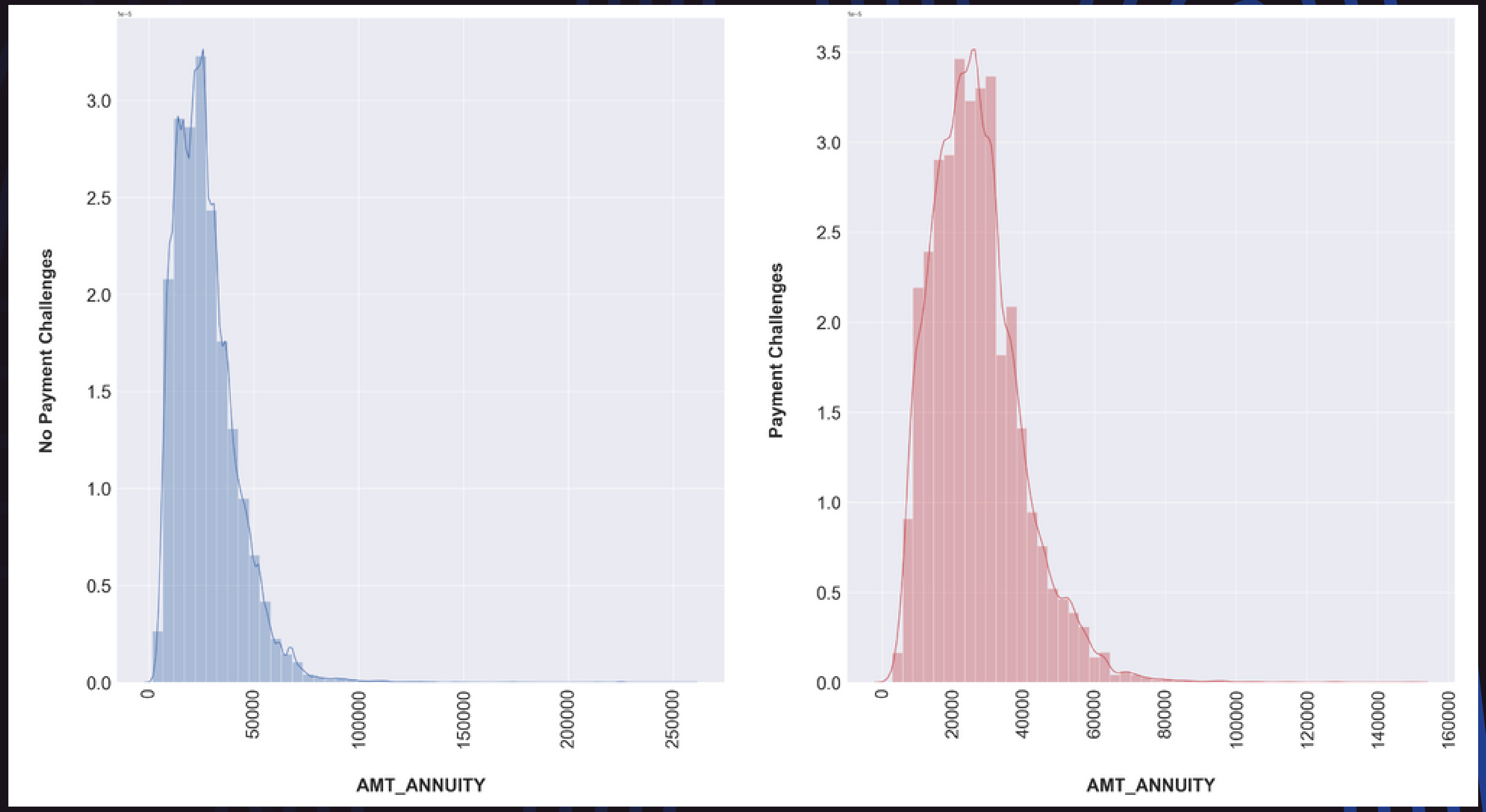
For AMT_CREDIT_TYPE:

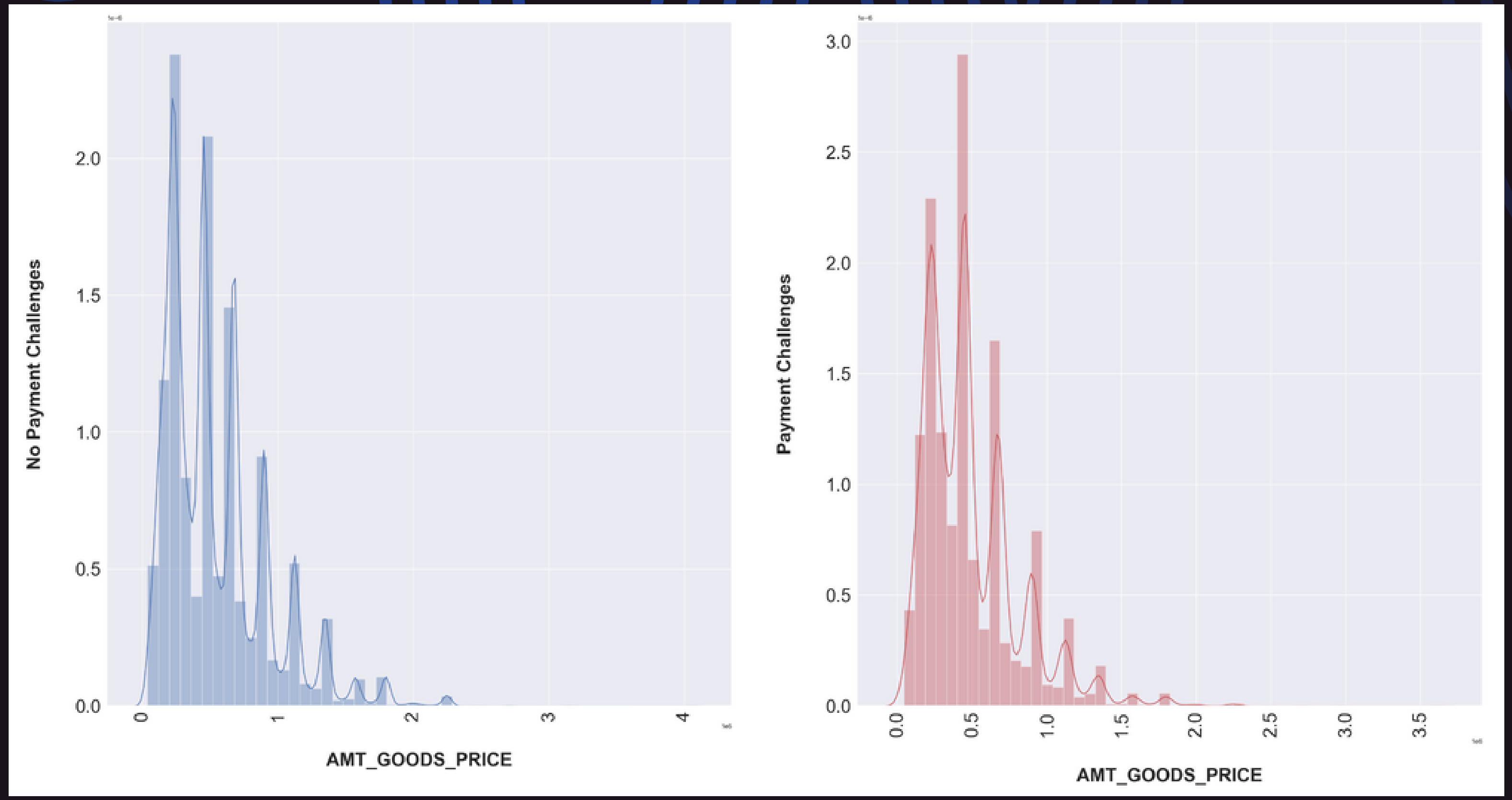
- **Majority of people applied for Medium Loan credit amount for both defaulters and non defaulters.**
- People applying for low credit amount have a high risk of default.

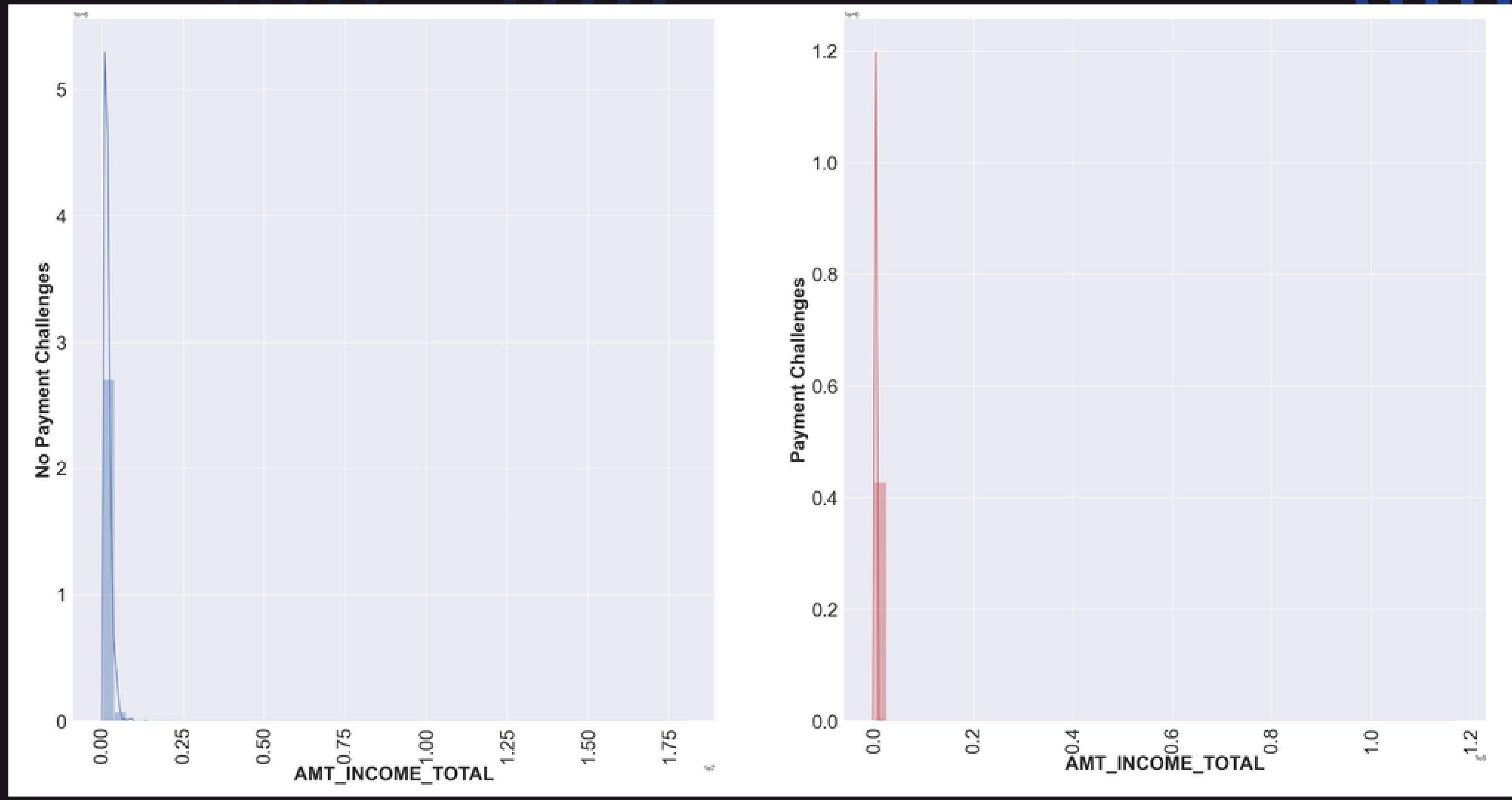


Univariate Analysis for Numerical Columns









Insights:

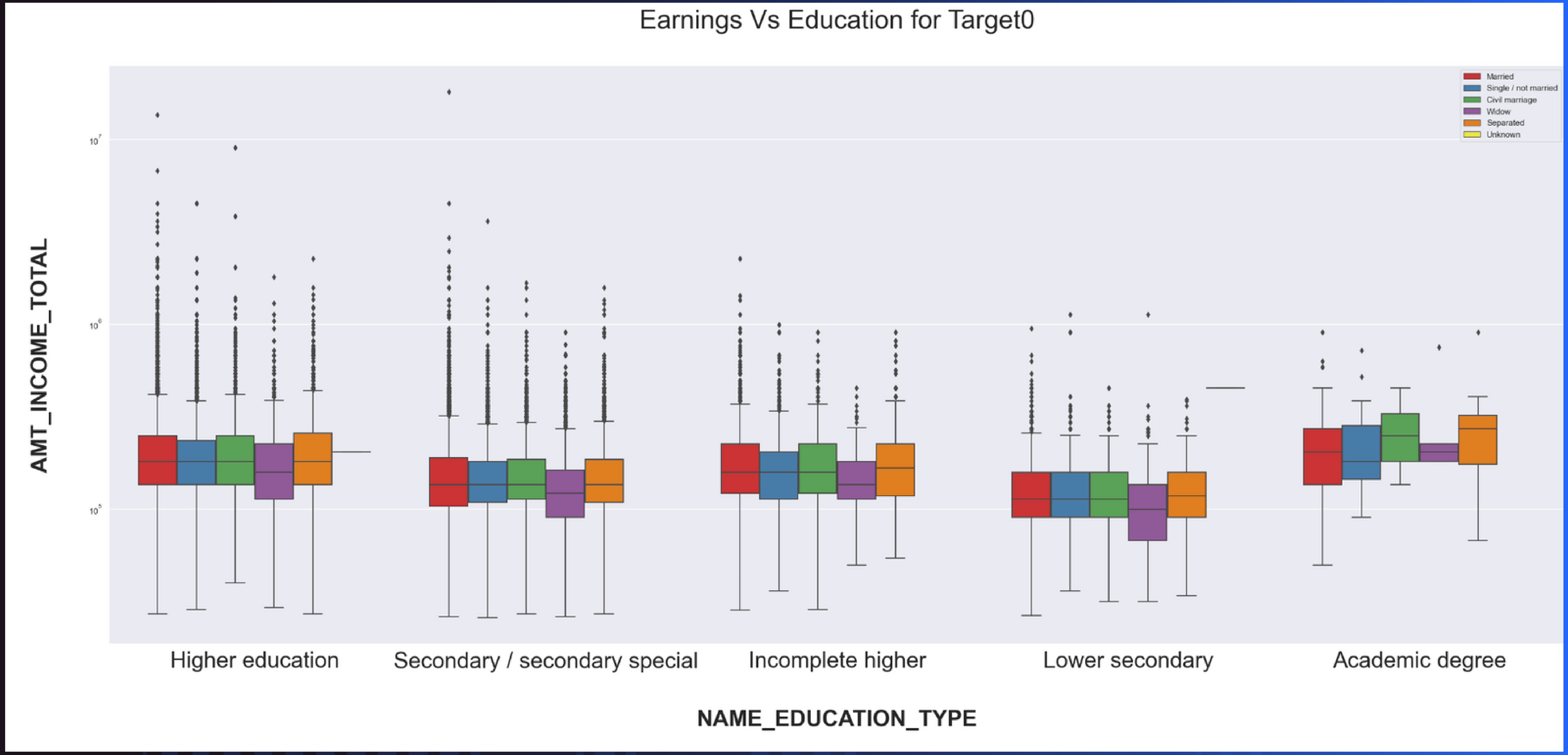
- From the four plots in the previous slide, it is clear that people fromTarget1 have staggered income as compared to Target0.
- The plots for Goods price, Annuity and Credit have identical plotshapes for Target0 and Target1.
- The plots also show people who are facing problems in paying back their loans with regards to their annual income, price of goods and loan credit amount.

Bivariate Analysis

Income Amount Vs Education Status with Payment Problems for Target0

Insights:

- From the plot, we can say that some of the people having higher education tend to have more income when compared with others.
- Also, a few people having secondary special education are earning more.
- People having higher education, secondary special education and incomplete higher education have a large number of outliers.

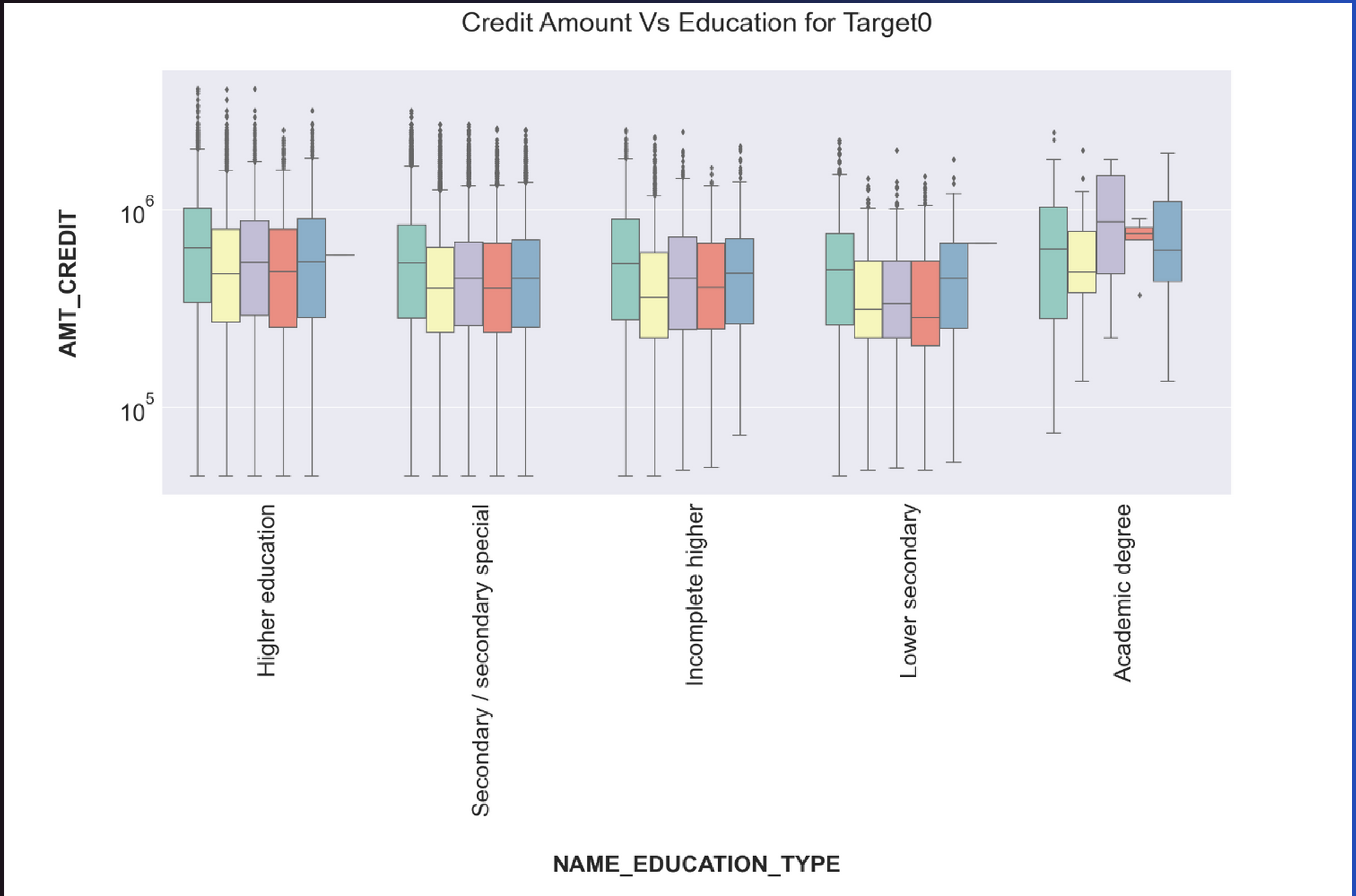


Bivariate Analysis

Credit Amount Vs Education Status for Target0

Insights:

- Except Academic degrees people from other education categories have a large number of outliers.
- Some people with Higher Education,Lower Seconday Education, and Special Education are likely to have higher credit loans.
- People with Academic Degrees and who is a widow are likey to take higher loans.

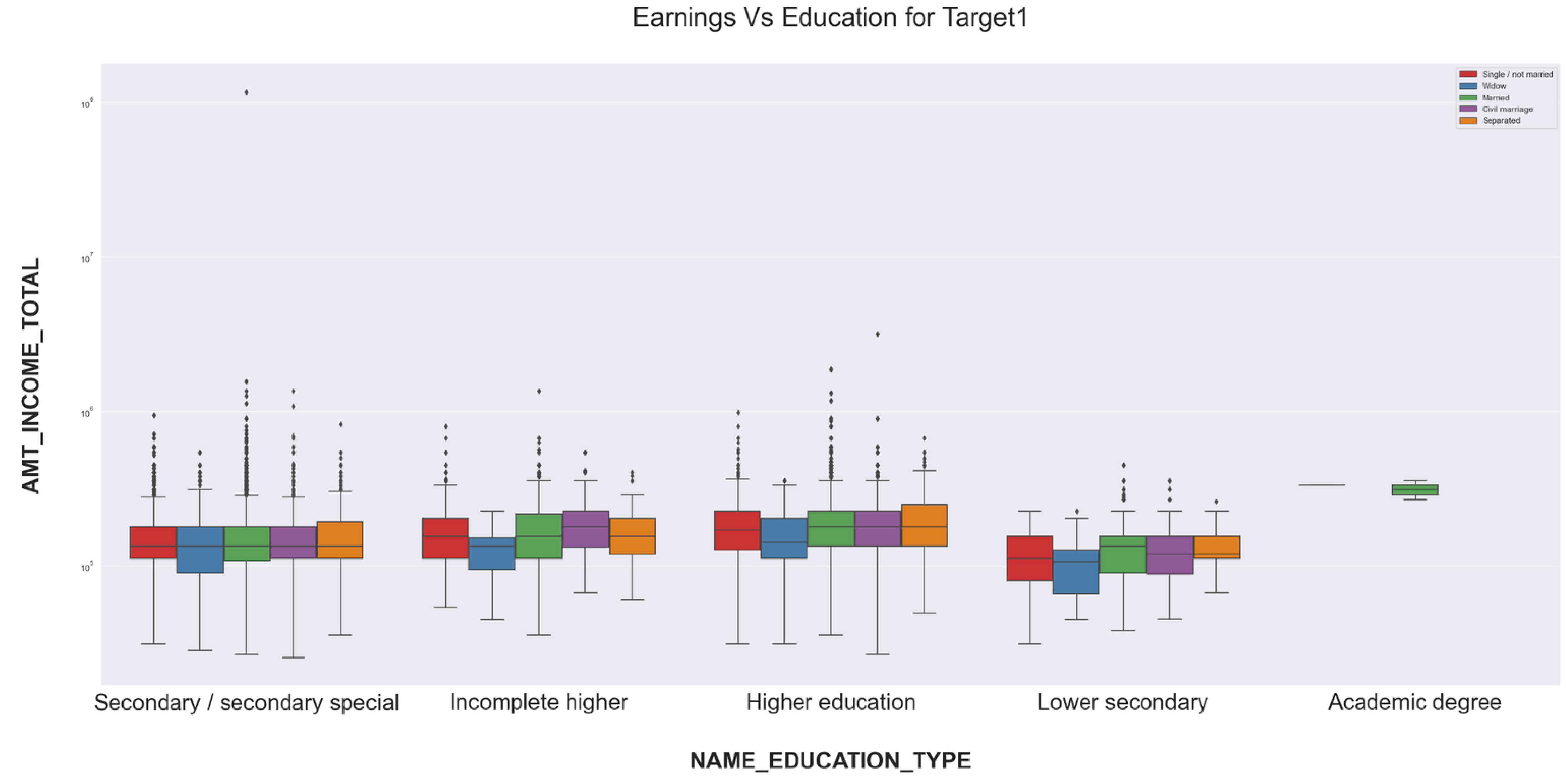


Bivariate Analysis

Income Amount Vs Education Status with Payment Problems for Target 1

Insights:

- From the distribution, we can see that married clients with an academic degree earn lesser when compared with others.
- The income of defaulters is relatively lower when compared with that of the non defaulters.

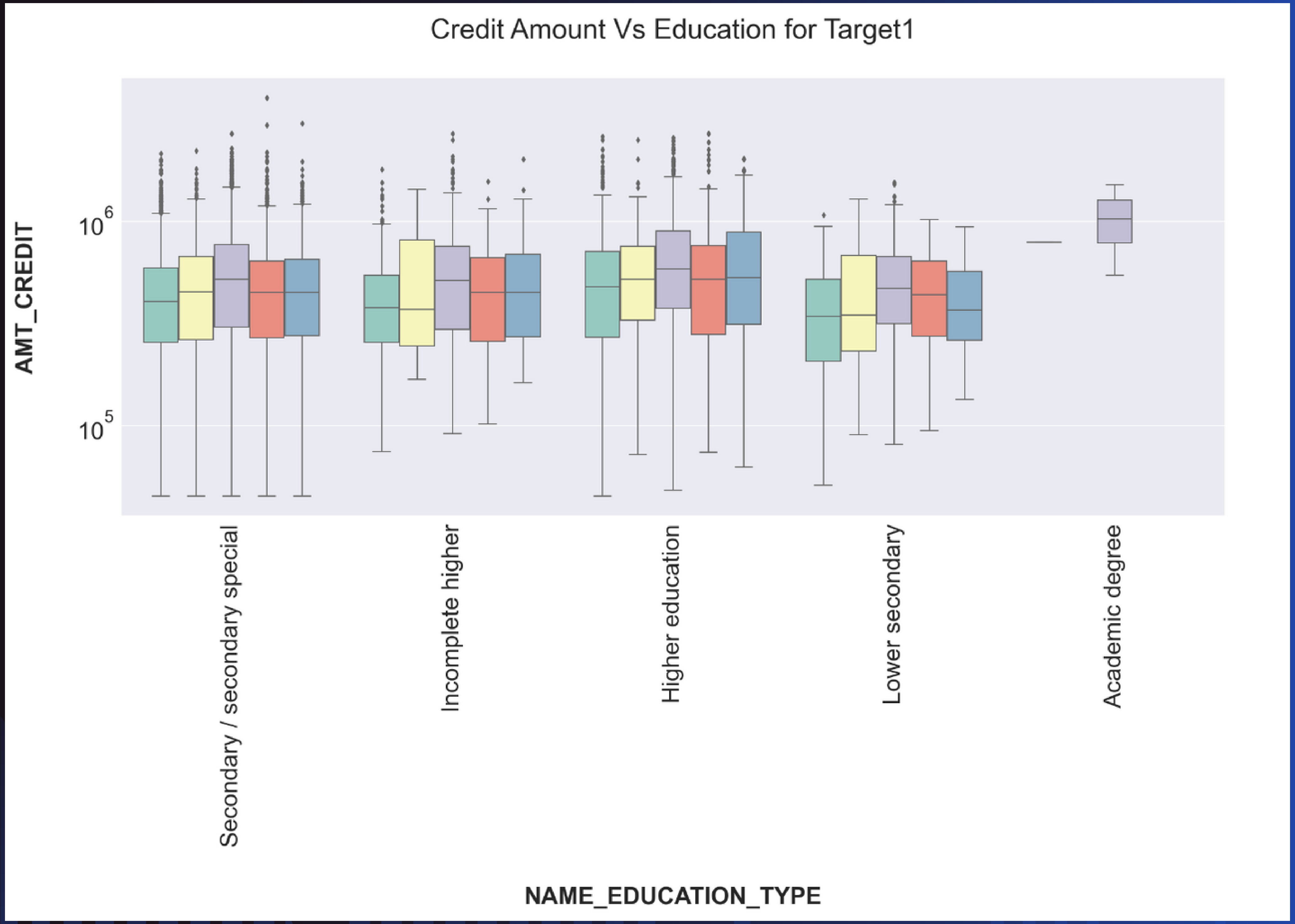


Bivariate Analysis

Credit Amount Vs Education Status for Target 1

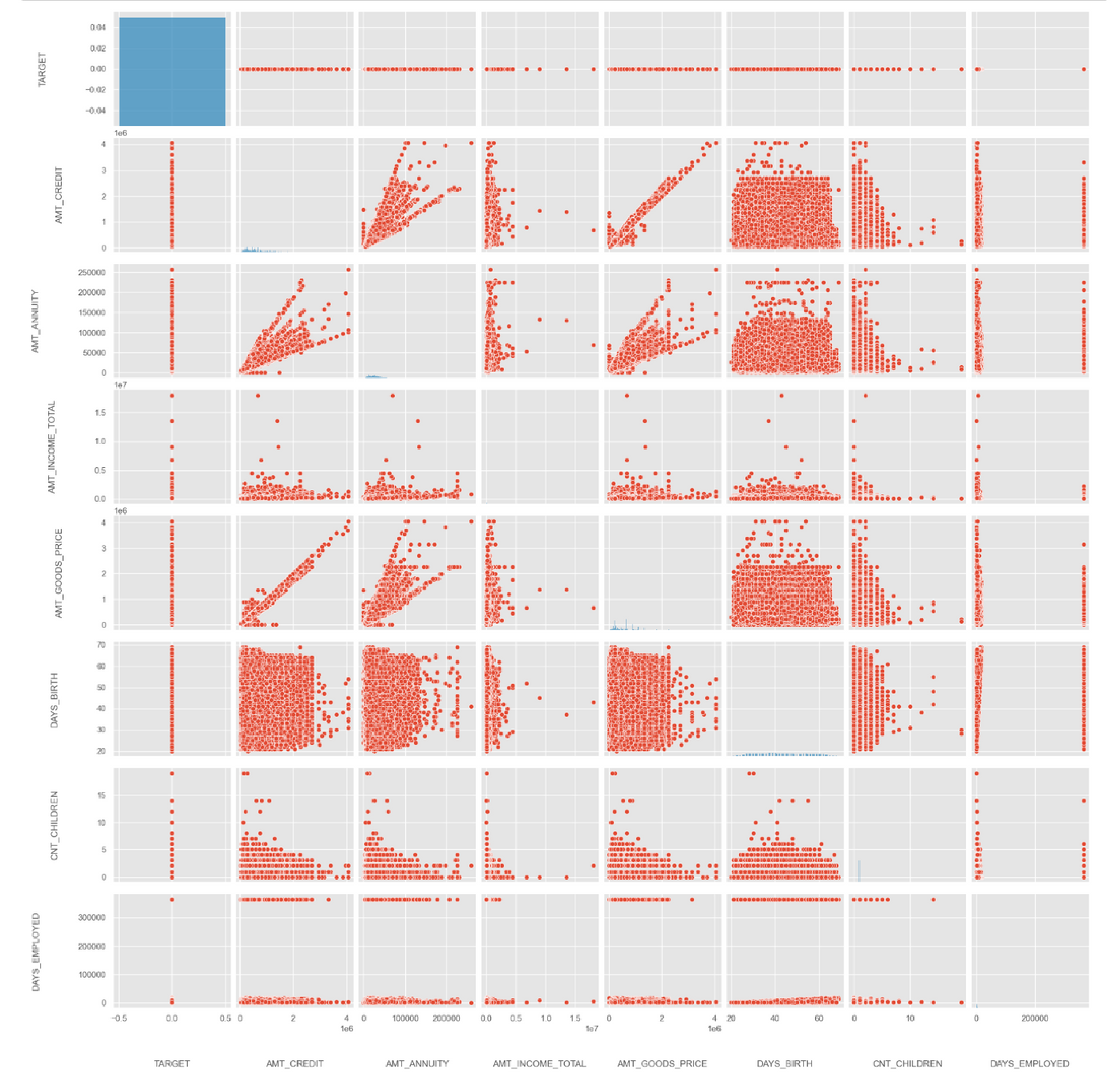
Insights:

- From the distribution, it is clear that married people with academic degrees applied for higher credit loans and have no outliers.
- Some people having Higher Education, Incomplete higher education and secondary special education tend to take higher credit loan.



Correlation Between Numerical Columns Using Pairplots

Target 0 Pairplot for No Loan Payment Challenges



Correlation Between Numerical Columns Using Pairplots

Target 1 Pairplot for Loan Payment Challenges

Pairplot Insights

- From the pairplots, it is evident that both the variables AMT_CREDIT and AMT_ANNUITY are highly corelated for both defaulters as well as non-defaulters. Thus, a spike in home price increases the EMI amount as well.

- Secondly, AMT_CREDIT and AMT_GOODS_PRICE are also correlated for both defaulters and non defaulters. Hence, as home price increases the loan amount also spikes up.

- AMT_CREDIT, AMT_ANNUITY and AMT_GOODS_PRICE and three variables are highly correlated and this might not be a good sign for defaulter detection.

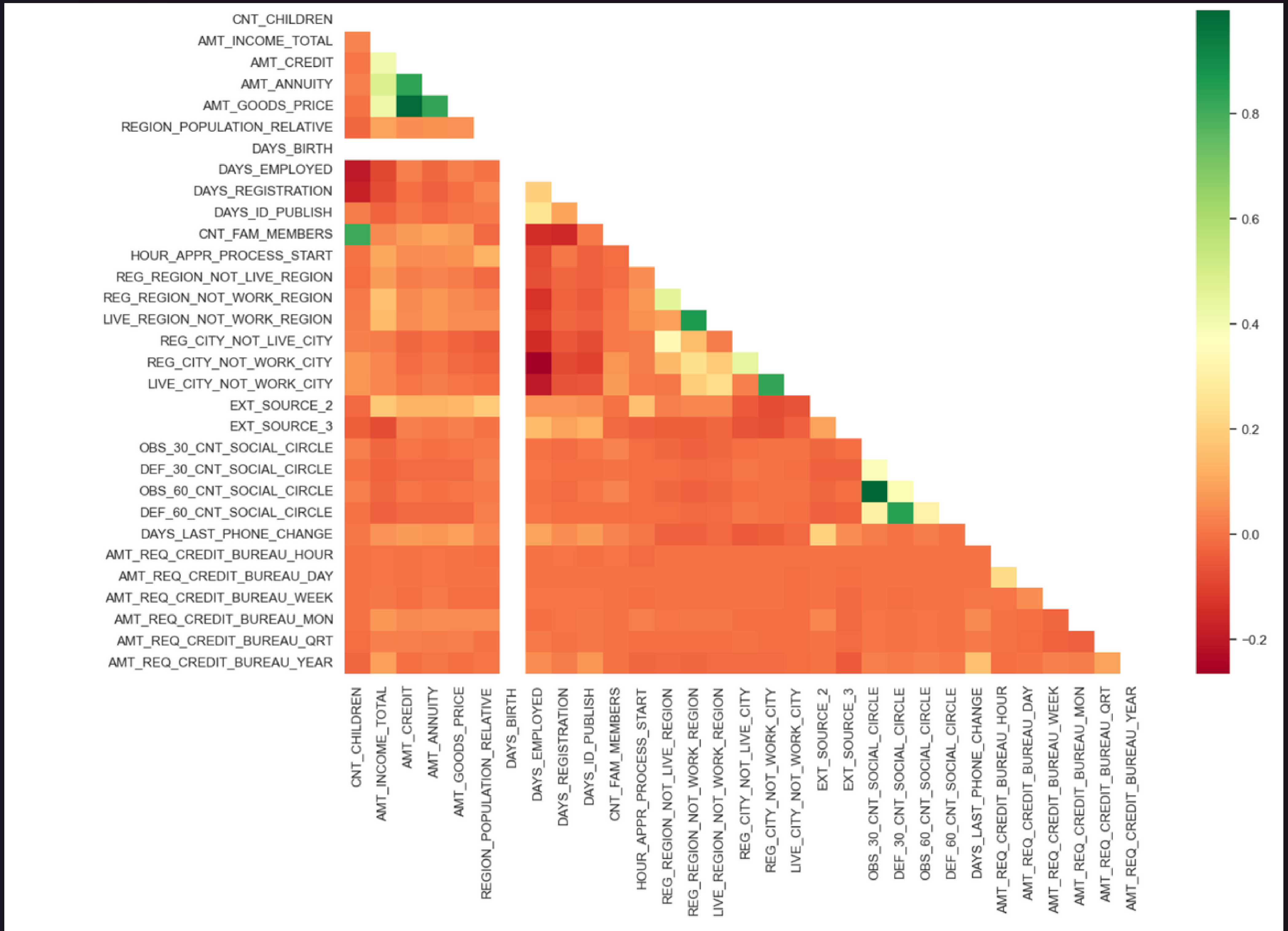
Correlation Between Numerical Columns using HeatMaps

Target 0 Correlation Insights:

- AMOUNT_CREDIT is inversely proportional to DAYS_BIRTH and CNT_CHILDREN. Thus, people from lower age groups have higher credit loans and vice versa.

- AMT_INCOME_TOTAL is inversely proportional to CNT_CHILDREN which means only fewer children have higher incomes and vice versa.

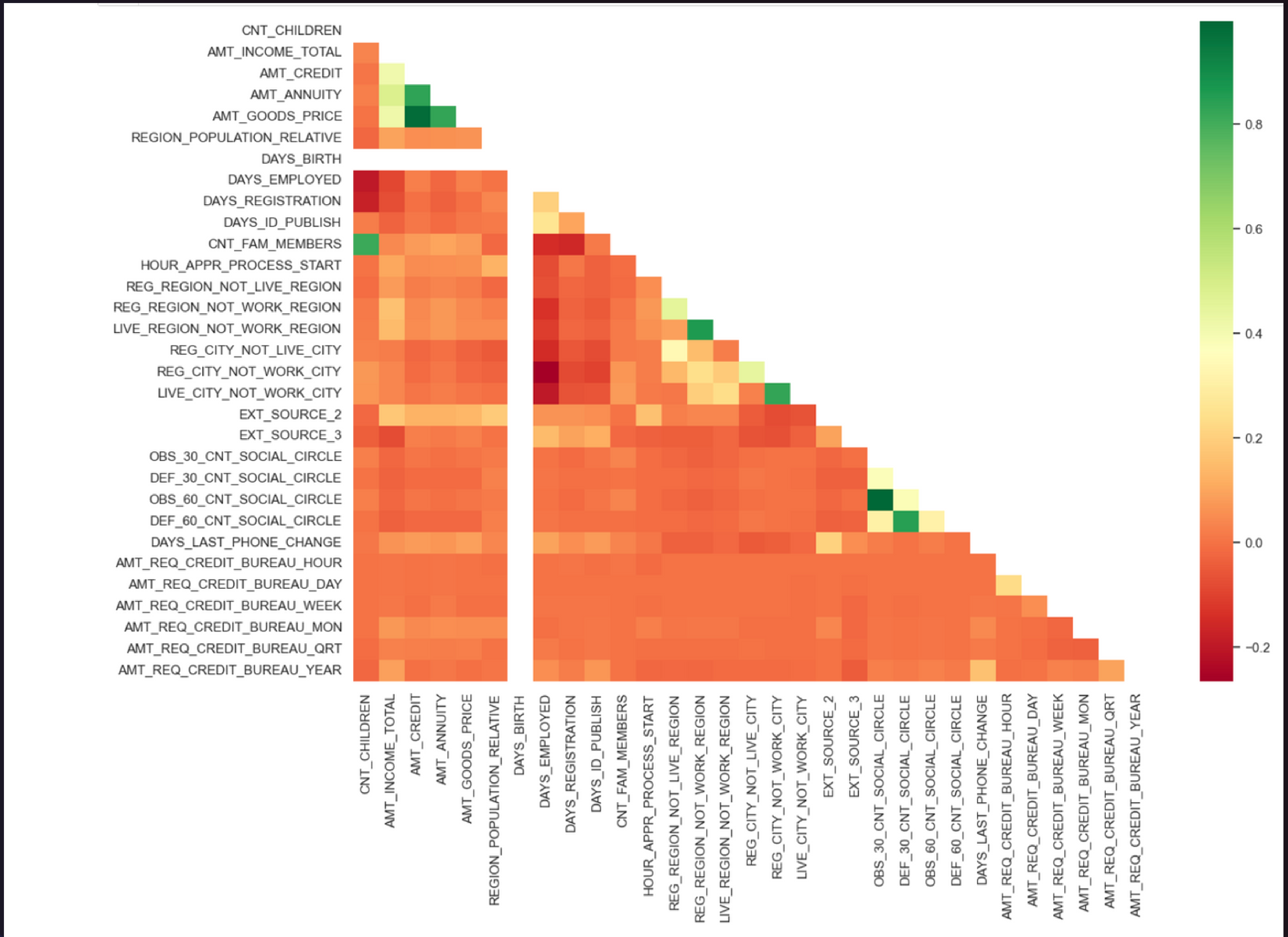
- From the heatmap, it also evident that AMT_INCOME_TOTAL and AMT_CREDIT are highly populated regions.



Correlation Between Numerical Columns using HeatMaps

Target 1 Correlation Insights:

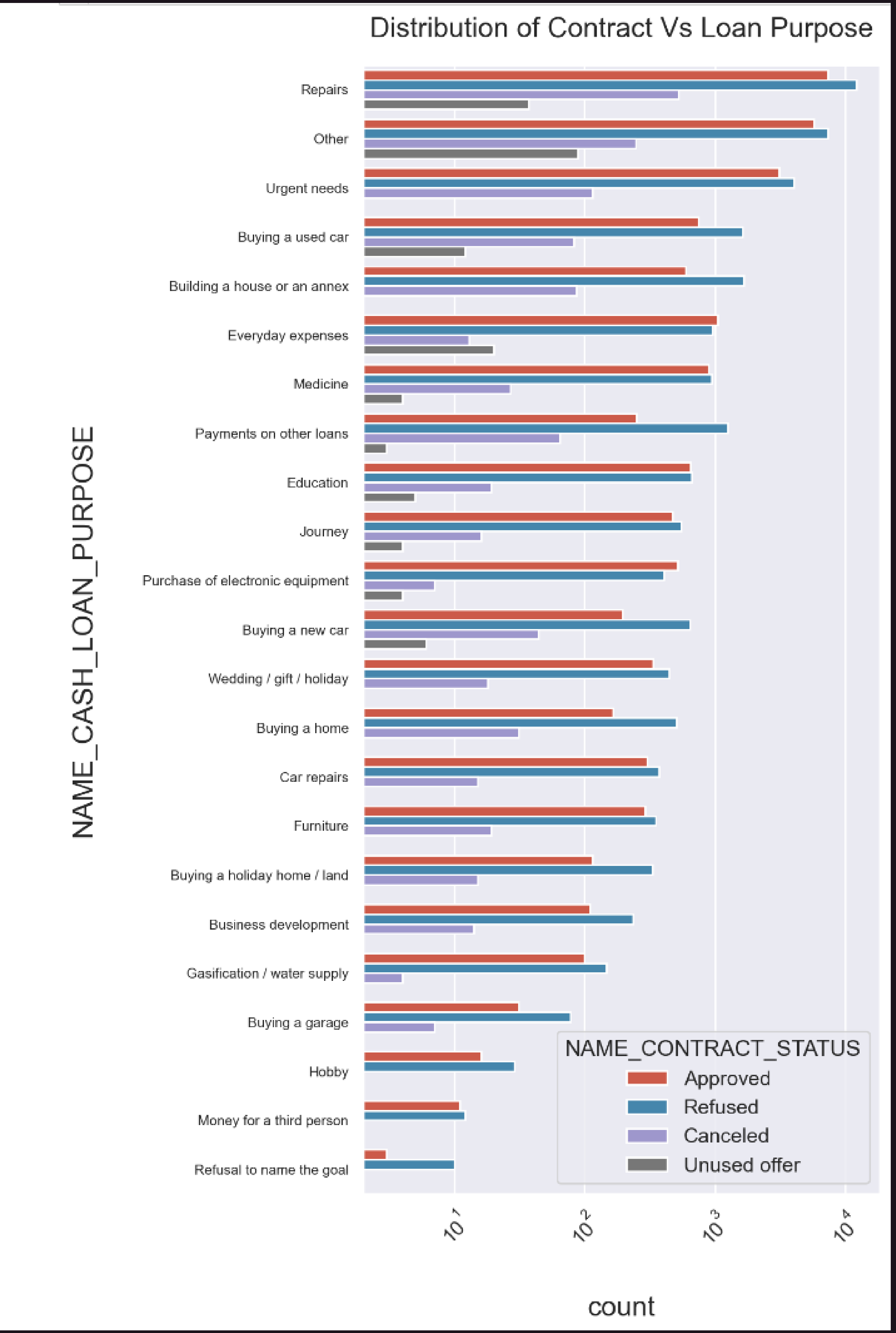
- The heatmap for Target 1 has quite similar observations as compared to Target 0.
- However, from the distribution it is clear that in Target 1 the person's permanent address does not match the contact address and are having lesser children.
- Similarly, the person's permanent address does not match the work address and are having lesser children.



Loan Distribution and Purposes

Contract Vs Loan Purpose Distribution Insights:

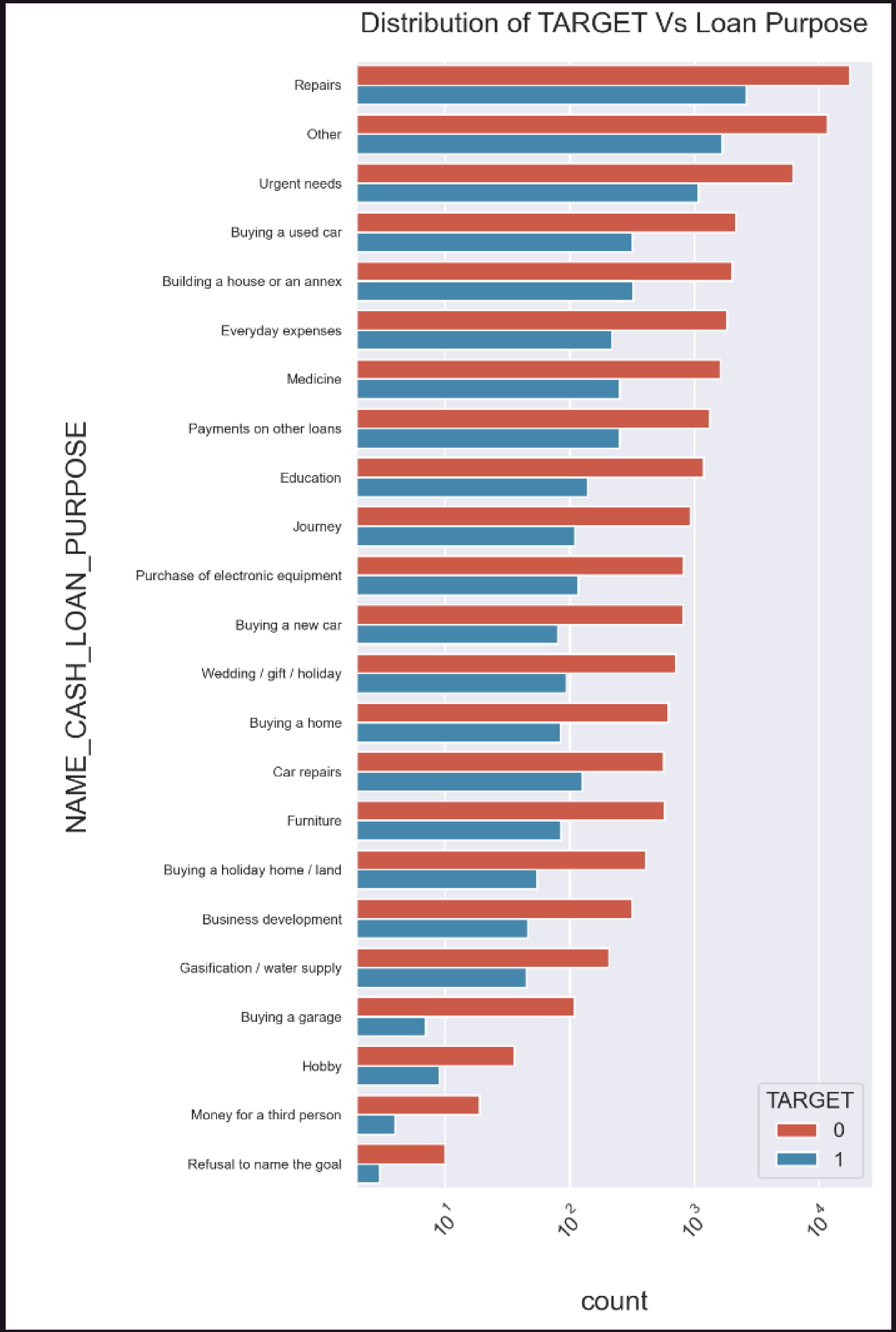
- From the distribution, it is evident that maximum loan rejections came from "Repair" loan purpose.
- We see that "Medicine" and "Educational" purposes have the same number of loan approvals and rejections.
- "Buying a new Car" and "Paying other loans" categories are having significantly higher number of rejections than that of approvals.



Loan Distribution and Purposes

Target Vs Loan Purpose Distribution Insights:

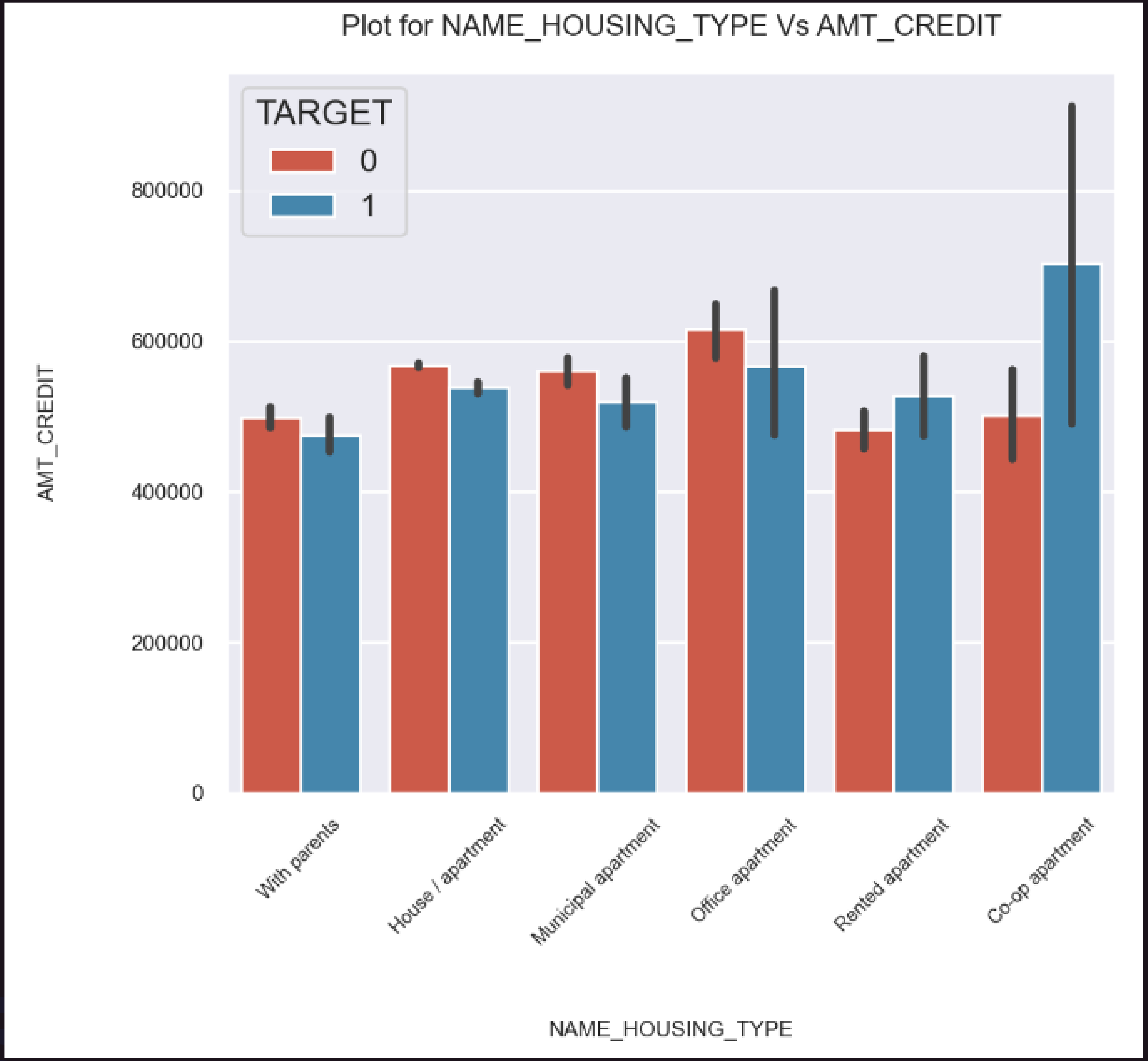
- From the distribution, we can see that for both the Targets "Repairs" purpose is having the highest number of loan rejections.
- While comparing the "Medicine" and "Educational" purposes it is clear that "Medicine" category is having people with more number of loan repayment challenges as compared with people in "Education" category.
- People falling under "Buying used car" and "Building Purpose" are having equal ratios in terms of Payment Difficulties.



Distribution of AMT_CREDIT Vs NAME_HOSUING_TYPE

Insights:

- From the barplot, we can see that, office apartment is having higher credit amount for target0 and co-op apartment is having credit for target 1.
- From this we can conclude that banks should avoid giving loans to people having housing type co-op apartment as they are having most difficulties in paying the loan.
- Banks should mainly focus on people having housing types as "With Parents" ans "Municipal Apartment" for achieving successfull loan payments.



Conclusions :

- Pensioners have lesser loan payment problems as comapred with working professionals.

- We observed a decline in the percentage of widowed and married women with Loan Payment Problems and a spike in the percentage of single and civil married with Payment challenges.

- Furthermore, it is very clear that people having completed their Higher Education are having lower loan payment challenges when compared with people having secondary special qualifications.

- "Low skilled labours" are driving factors for loan defaulters as they have maximum percentage of payment difficulties around 18%.

- Similarly, people falling under "Lower Secondary" education are also major contributors towards loan defaulters having maximum percentage of payment problems around 11%

Recommendations :

- In order to achieve a higher success rate for successful payments, banks should majorly target pensioners, businessmen and contract type students.

- Banks should avoid granting loans to people having housing types as "Co-Op Apartments" because they have the maximum credit loan payment difficulties.

- Banks should also be careful while giving loans to people falling under "Working" category as they have higher loan payment difficulties.

- Also, people with loan purpose "Repairs" are having the maximum number of unsuccessful loan payments on team, this banks should be vigilant.

- Banks should focus more on housing type as "With Parents" and "Apartments" as they have higher successful loan payments.

THANK YOU!