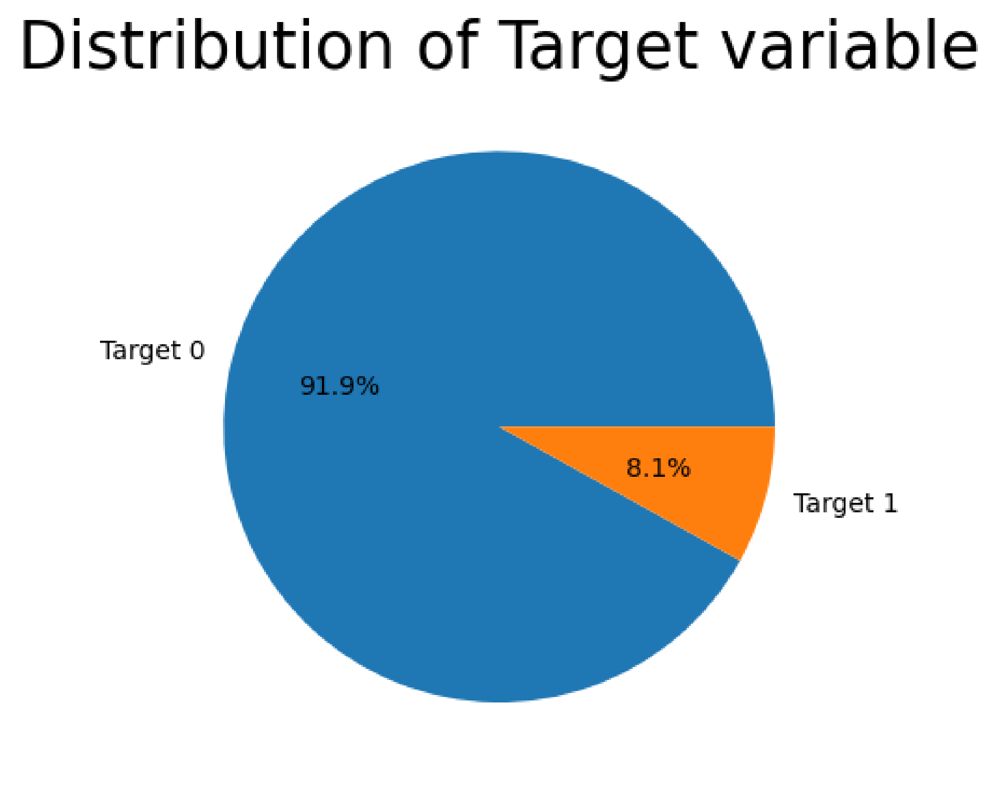
CREDIT EDA CASE STUDY

PRAVEEN KUMAR V

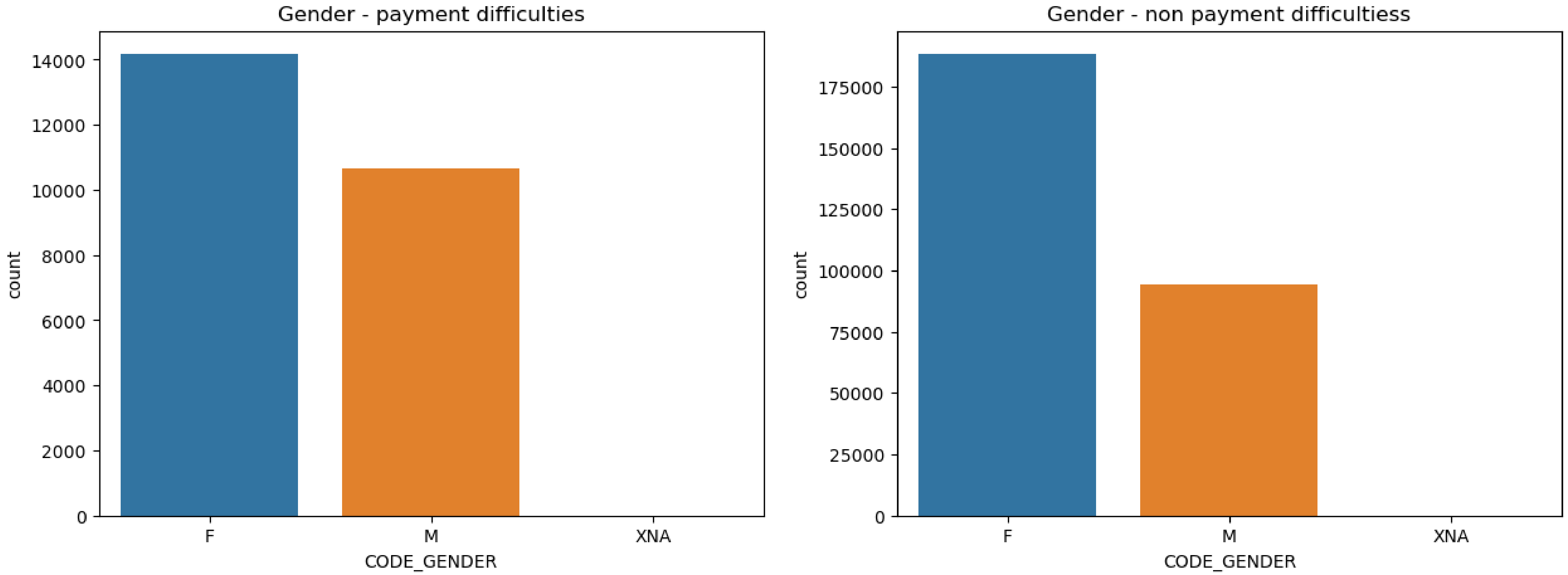
DATA IMBALNCE

Data Frame Target variable has highly imbalanced data. We will separately analyze the data based on the target variable for better understanding.



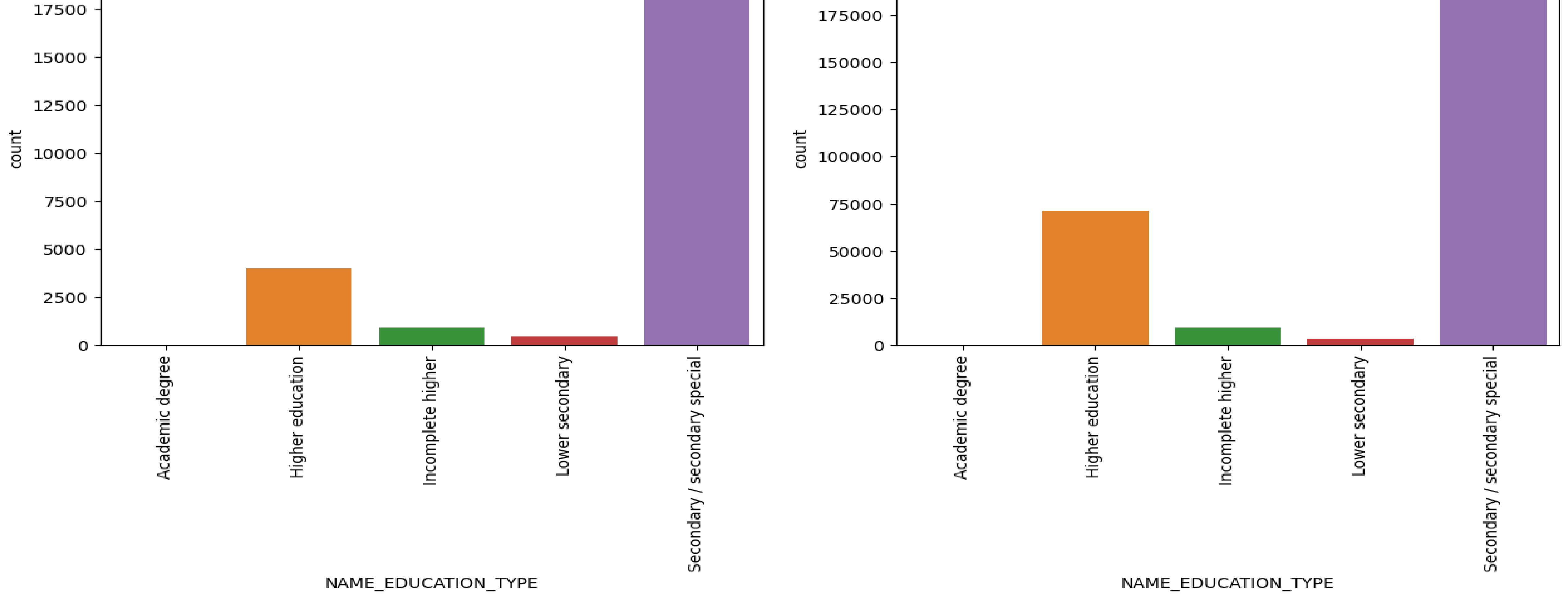
Univariate Analysis for Categorical variables

APPLICATION DATA – ANALYSIS FOR BOTH TARGETS



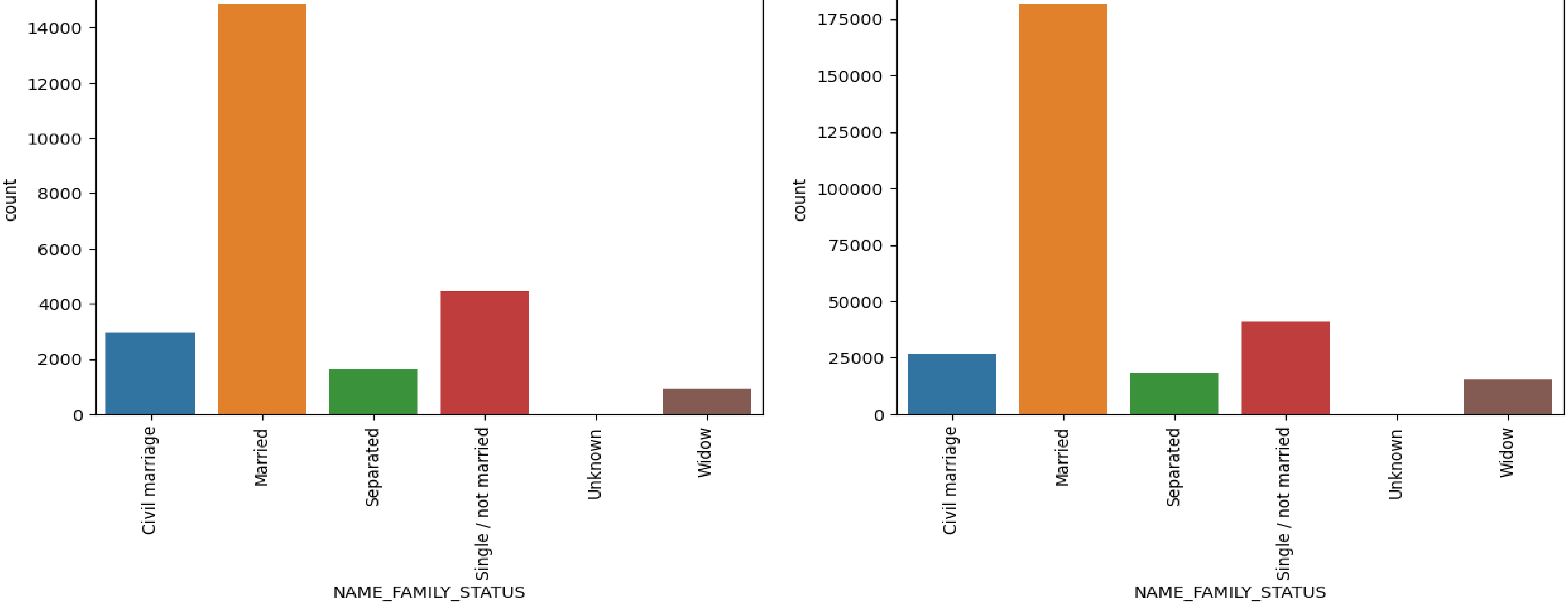
Gender distribution for Target 1 & 0

It is clearly visible that Female(F) applicants are more in both Targets. But the proportion of Male(M) is higher in first chart – indicating that there are more Male defaulters out of total mail applicants.



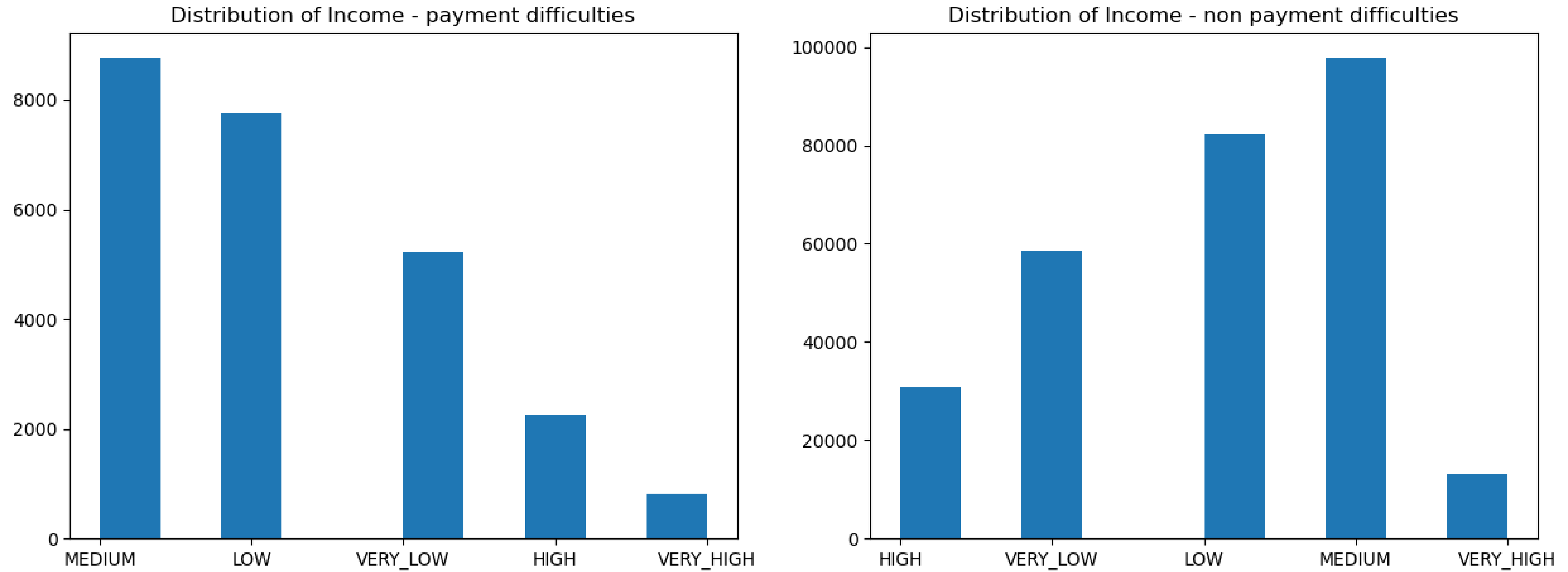
Distribution of Education (Target 1 & 0)

Higher education segment has higher proportion in Target 0 indicating that they are less likely to default. Secondary education applicants are more likely to apply for loan.



Distribution of Family Status (Target 1 & 0)

Single/ not married category has bigger proportion in Target 1 indicating default. Followed by Civil marriage. Widow are less likely to default.

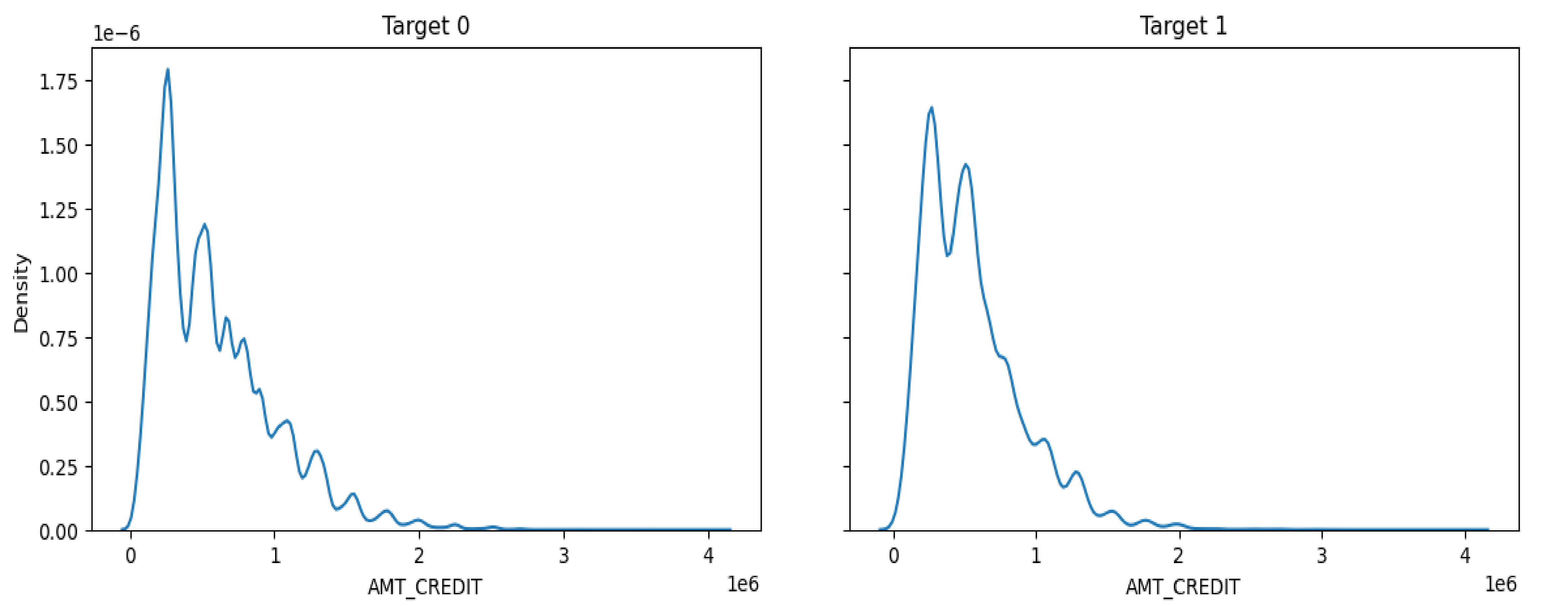


Distribution of Income(Target 1 & 0)

There are high number of applicants with Medium Income. Very High income and High income groups are less likely to be defaulters.

Univariate Analysis for Continuous variables

APPLICATION DATA – ANALYSIS FOR BOTH TARGETS

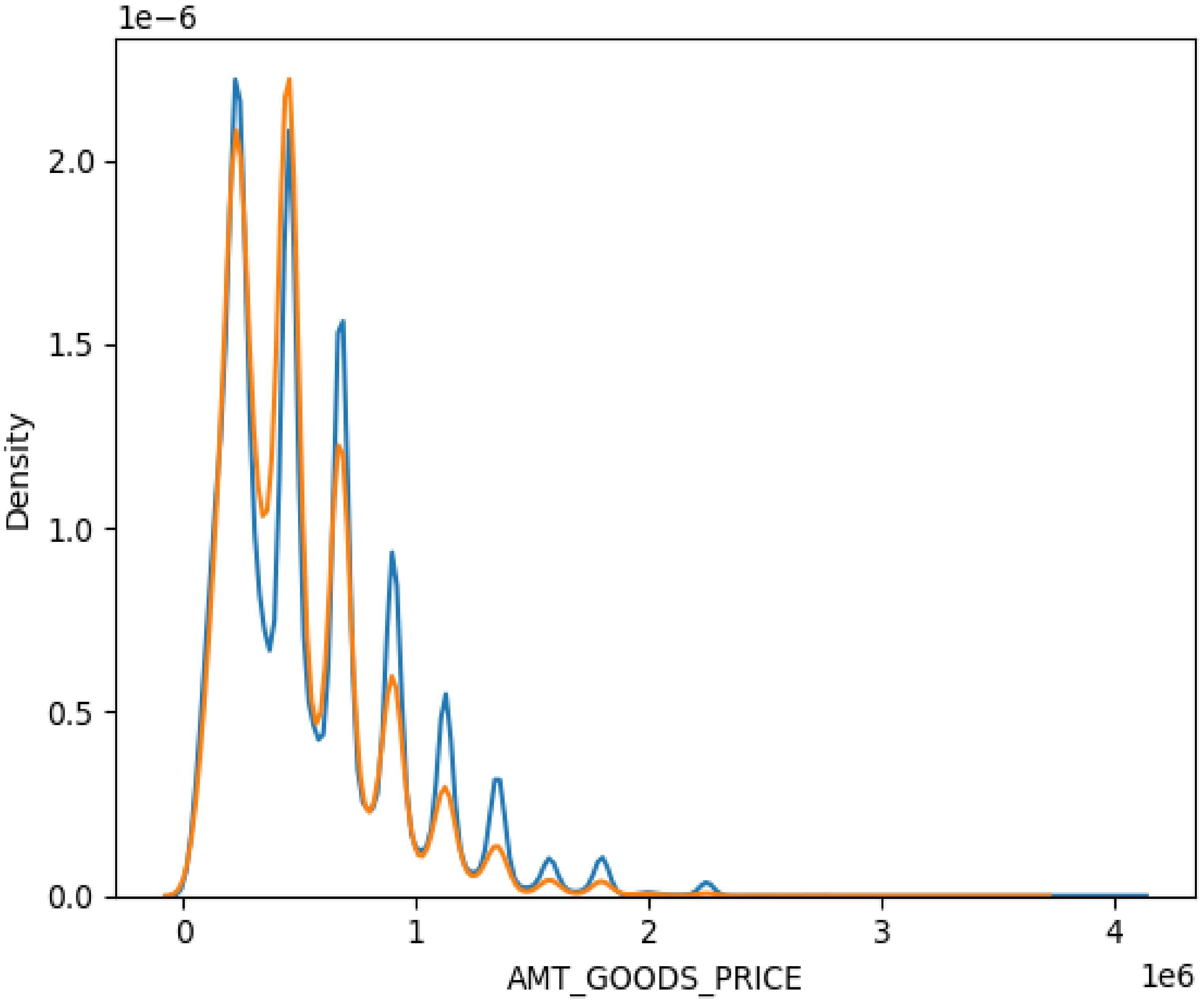


Distribution of Amt_Credit against Density

There is higher default when credit amount is less. But as the amount increases, risk of default reduces. There is almost zero default after amount crosses 20,00,000.

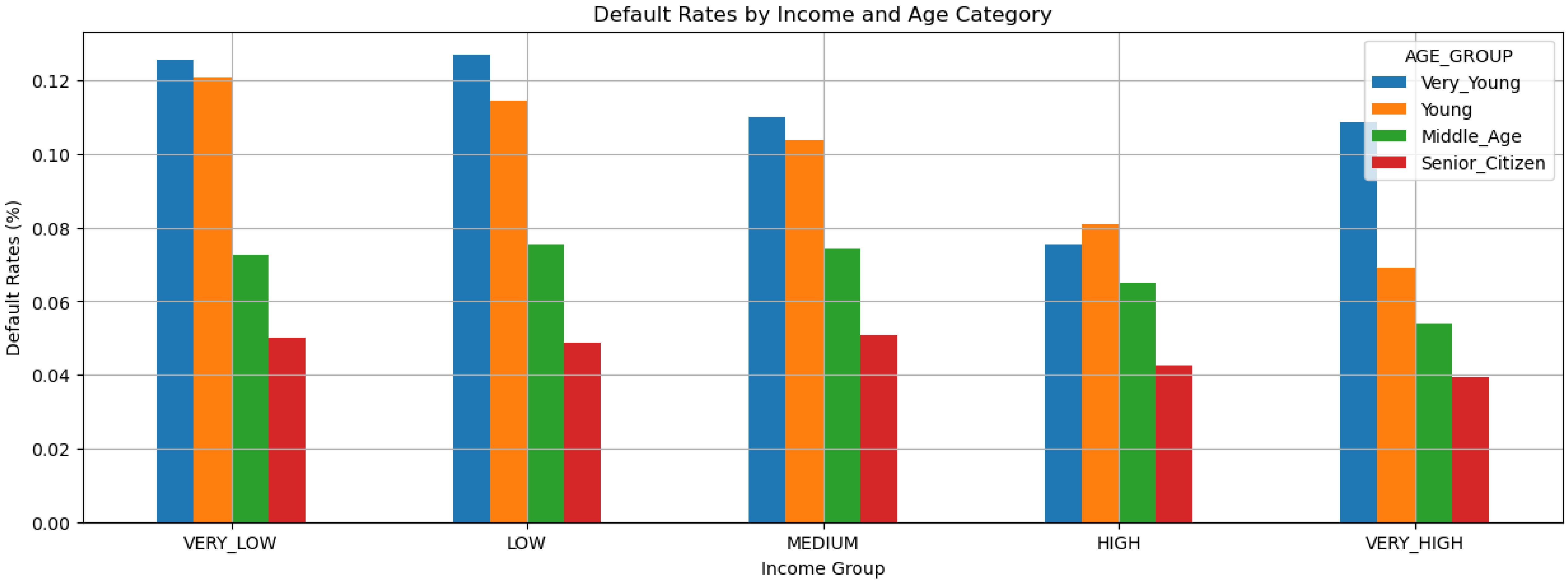
Density Distribution of Amount of Goods Price

- Blue representing Target 0 and Orange representing Target 1
- There is less default for higher goods prices.
- We can clearly observe that once Amt crosses 1, there are very less spikes of density of Target 1.



Univariate Segmented Analysis

APPLICATION DATA – ANALYSIS FOR BOTH TARGETS

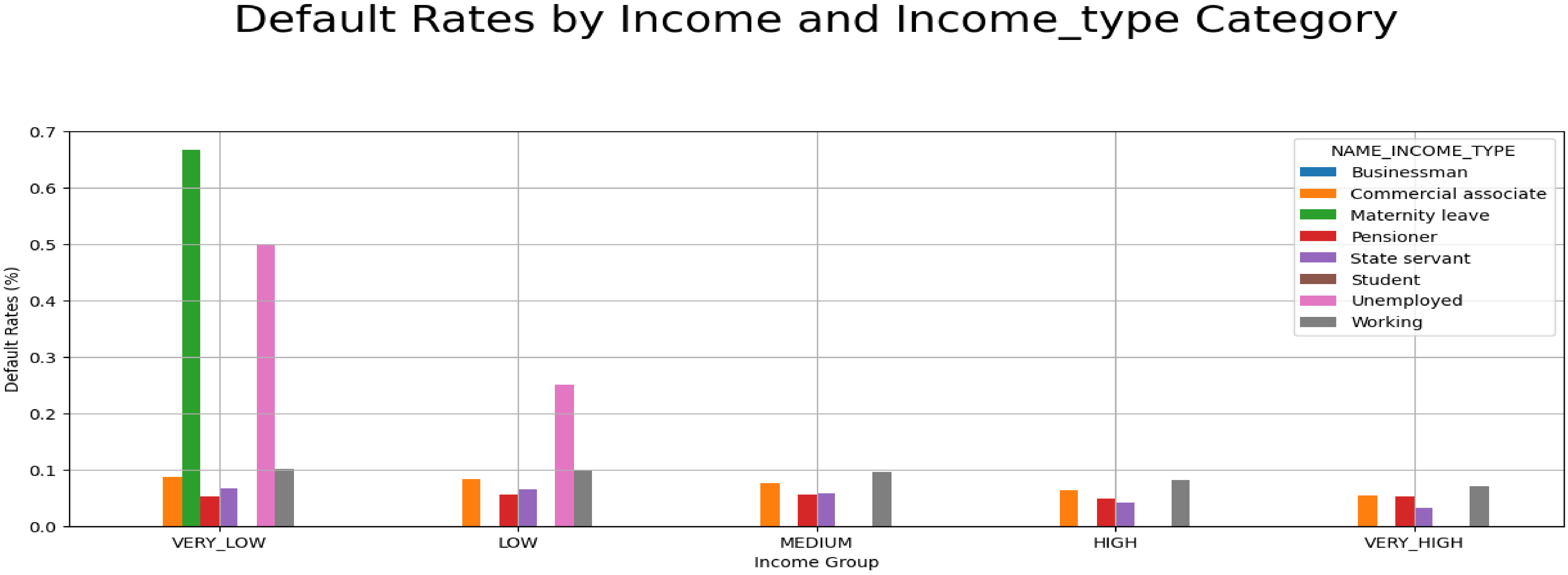


Default rates by Income and Age category

As the income level increases, there is less default rates. Very young and Very high income category is an exception, which has very high default rate.

Bivariate Analysis for Continuous variables

APPLICATION DATA – ANALYSIS FOR BOTH TARGETS

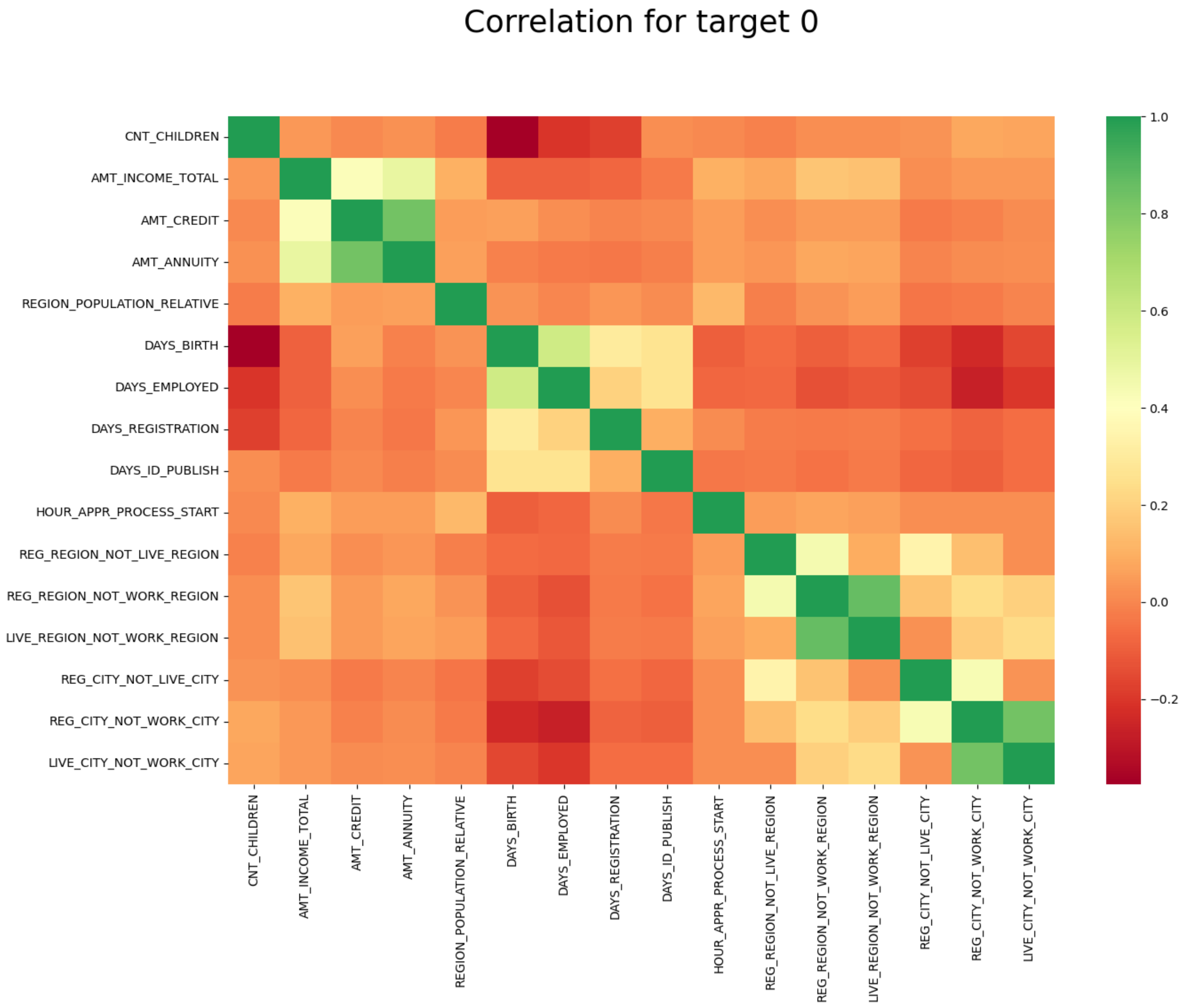


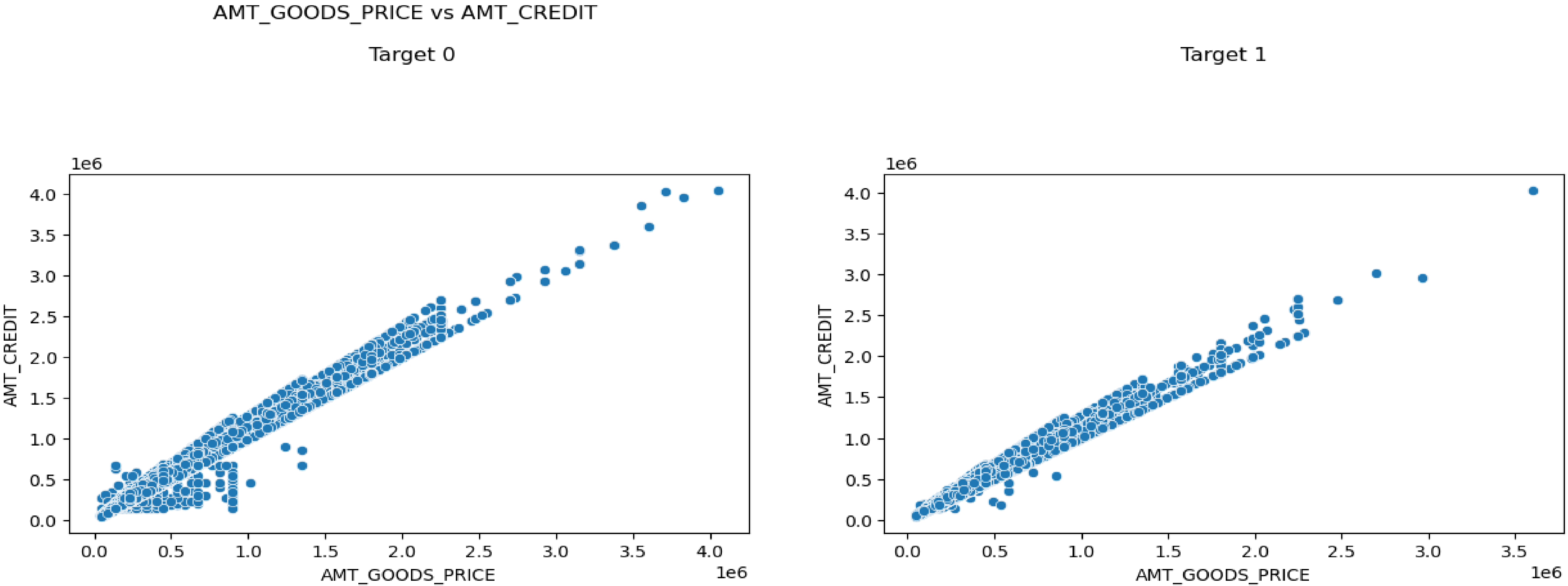
Default rates by Income and Income_type

Maternity leave present only in very low income has highest default rate. Unemployed has high default rate in very low and low. Pensioner also has high default rate in comparison.

Correlation for Target 0

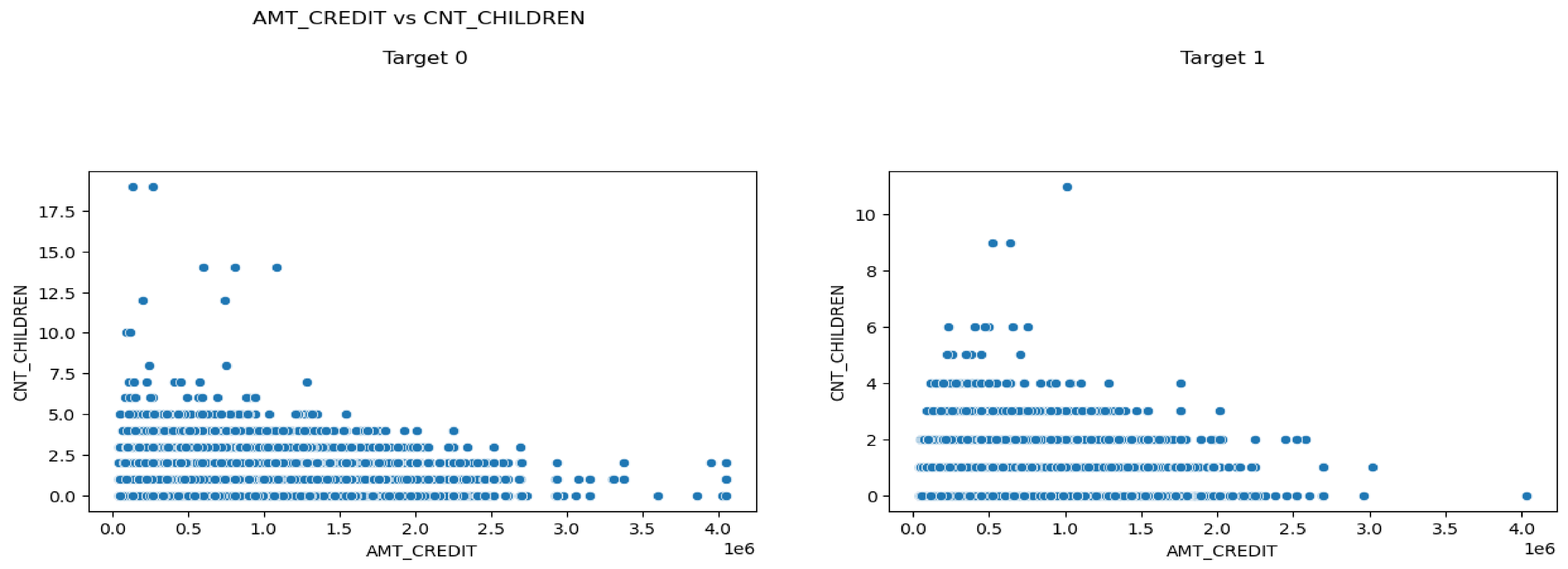
- Amt_income has high correlation with Amt_credit and Amt_annuity
- Days_birth has correlation with days_ employed as expected.
- All variables with region parameter has correlation with each other





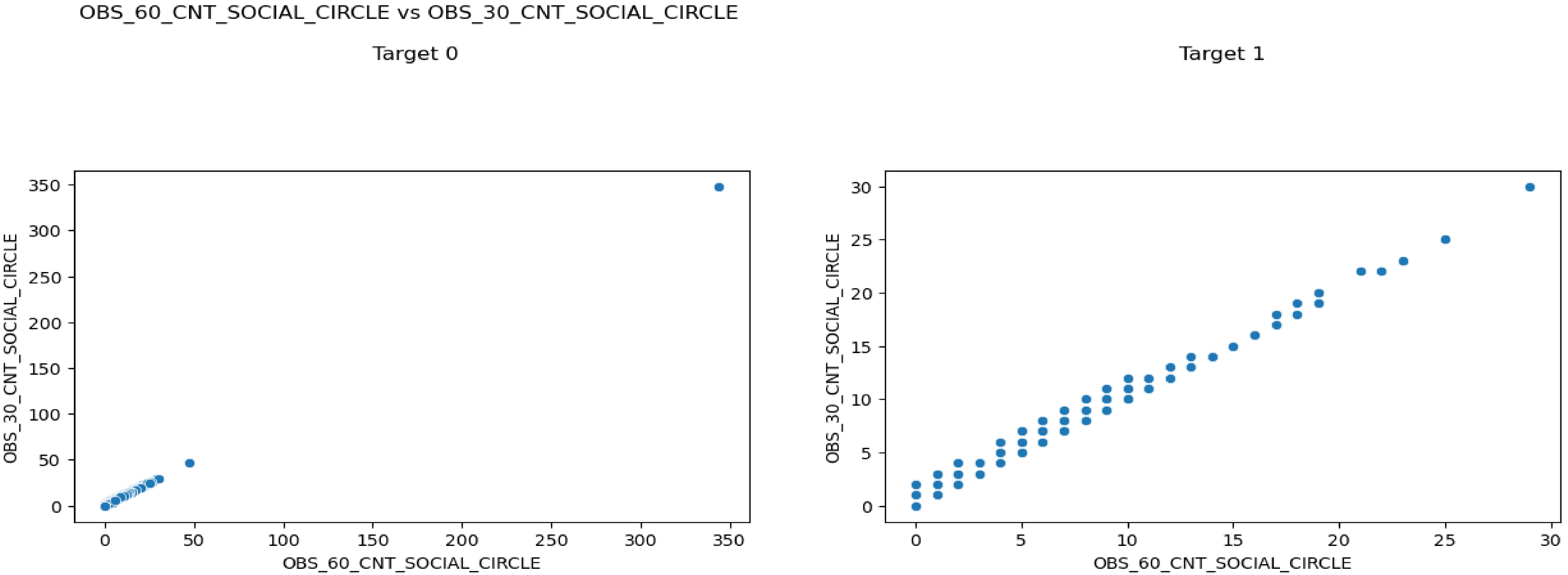
Correlation through scatter plots

Amount of credit is plotted against Goods price amount. Positive correlation between these. As the Credit amount increases, goods price also increases. Weakened correlation in Target 1 could be because of less data points.



Correlation through scatter plot

Relation is inversely proportional. As the count of children increases, there is decreasing correlation with higher credit amount. If the outlier values were removed, it would be even more definite.

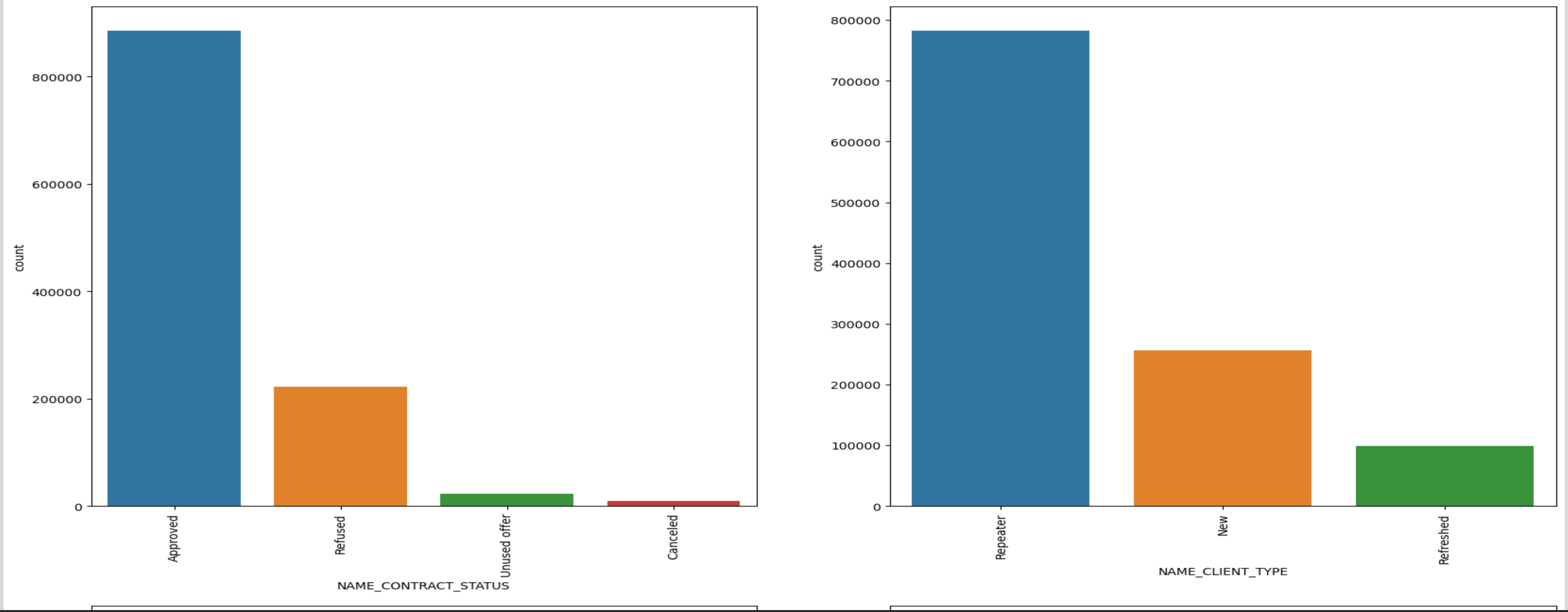


Observable defaults in social surroundings

We can clearly see how there is high correlation present only in target 1. It indicates that these social surrounding defaults are present in default applications.

Univariate, Bivariate & Multivariate Analysis

MERGED DATA FRAME – APPLICATION DATA IS COMBINED WITH PREVIOUS DATA

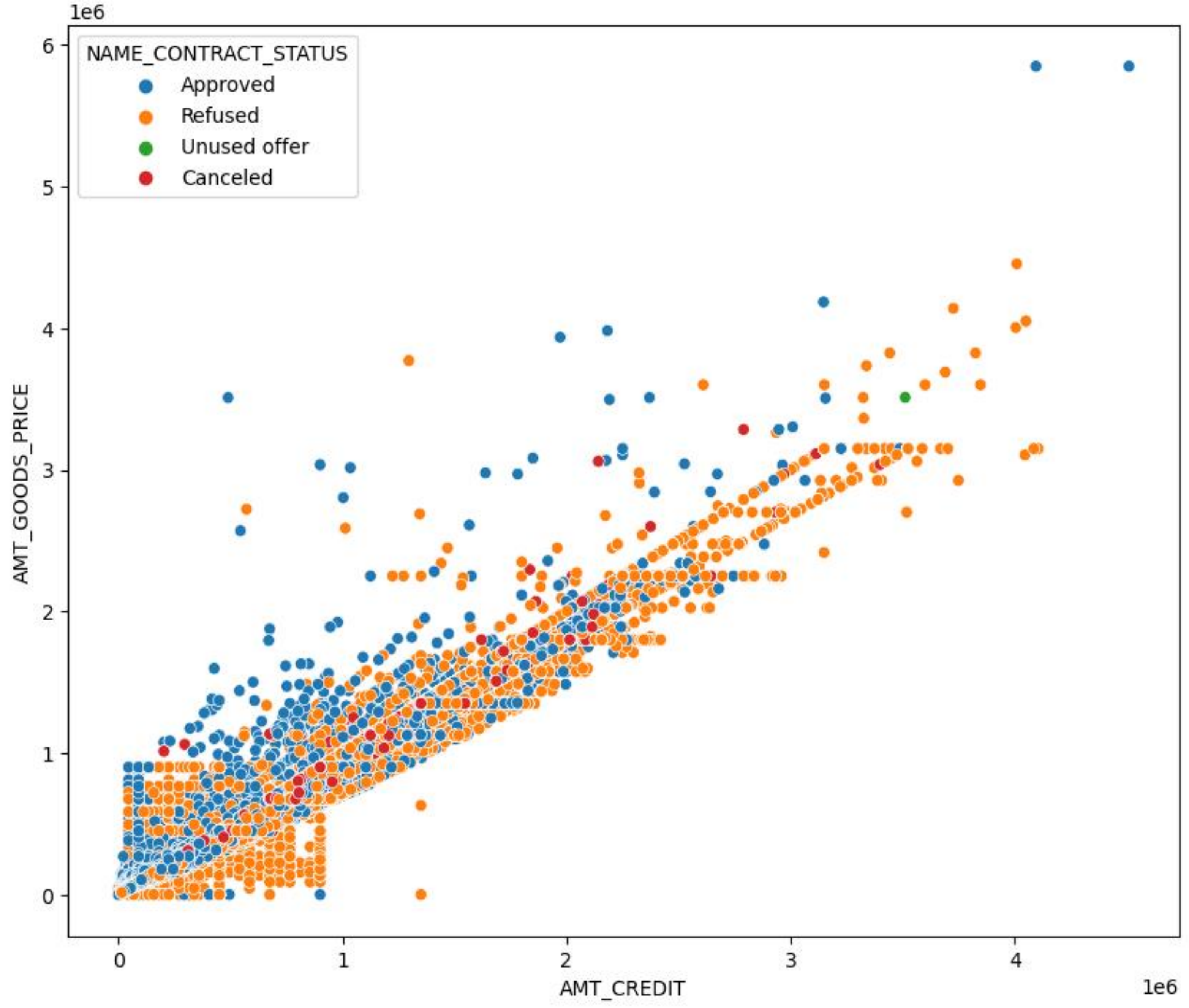


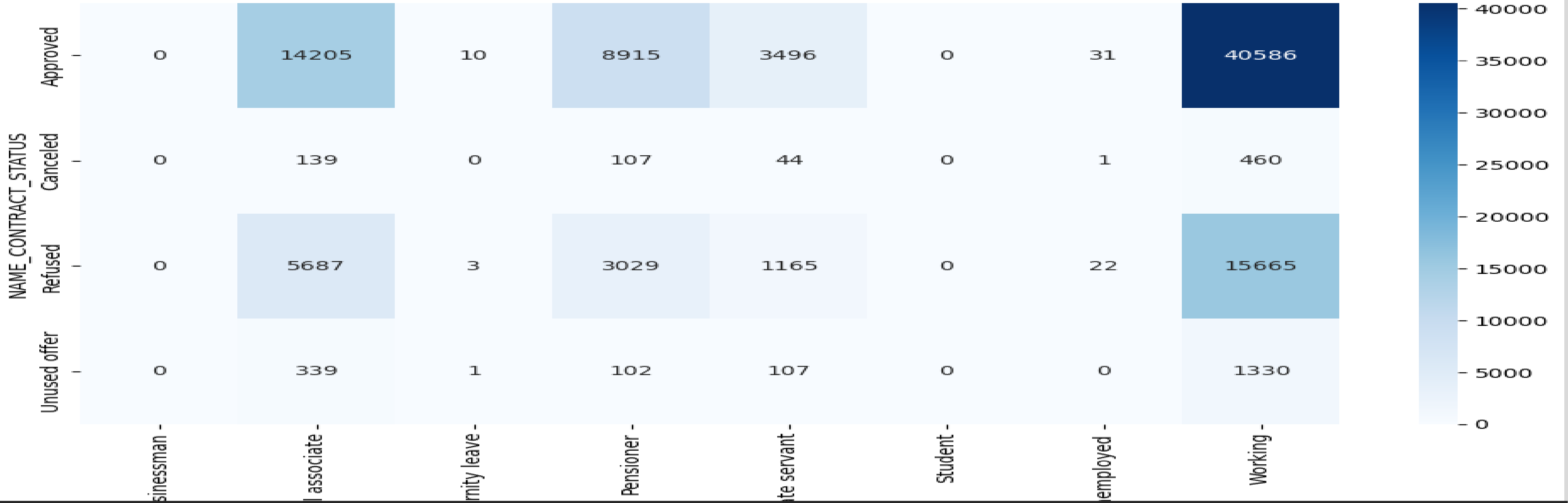
Contract_status and Client_Type

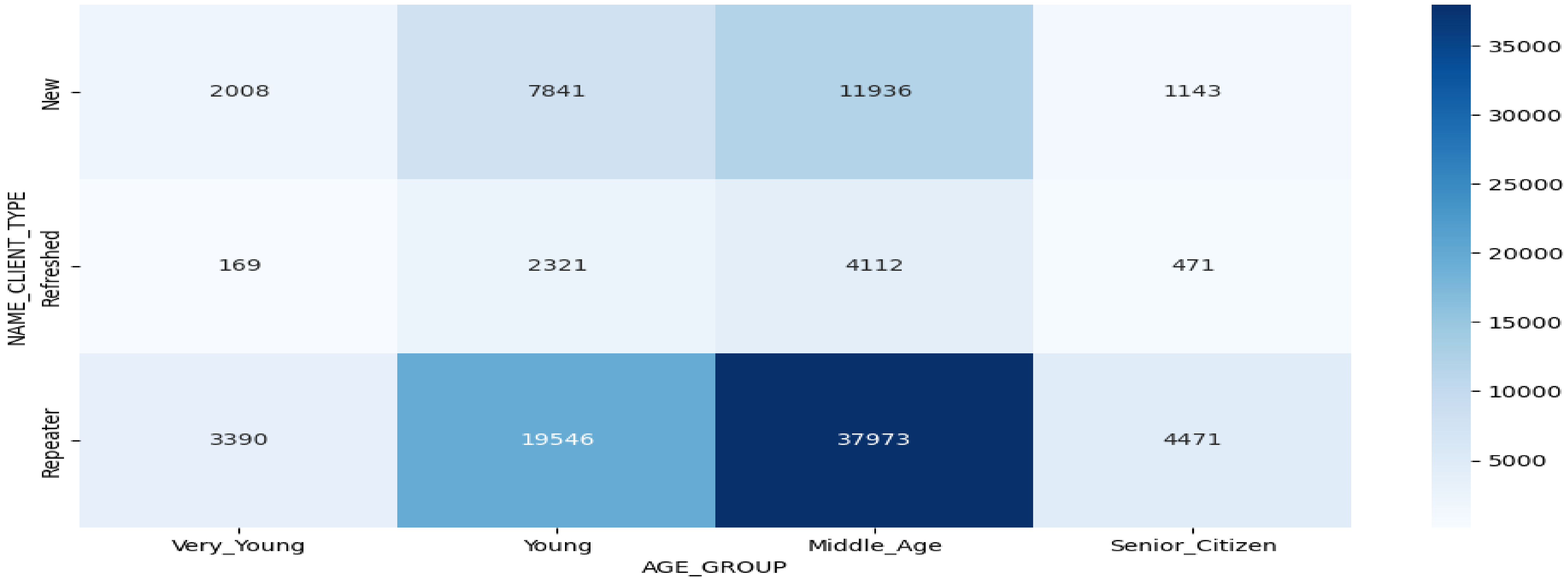
There are highest applicants with approved status, followed by Refused and unused. There are highest applicants with Repeater client type followed by New and Refreshed.

Scatter plot of Contract status across Credit & Goods

- There is less approvals as the amount of credit and goods increases.
- There are more refusals , in comparison to approval as the amounts increase.
- Increase in goods price is driving the approvals at the dense bottom of this highly positively correlated plot between both amounts.




NAME_CONTRACT_STATUS
Approved
Canceled
Refused
Unused offer
0
14205
10
8915
3496
0
31
40586
0
139
0
107
44
0
1
460
0
5687
3
3029
1165
0
22
15665
0
339
1
102
107
0
0
1330
sinessman
l associate
rnity leave
Pensioner
ate servant
Student
employed
Working
40000
35000
30000
25000
20000
15000
10000
5000
0

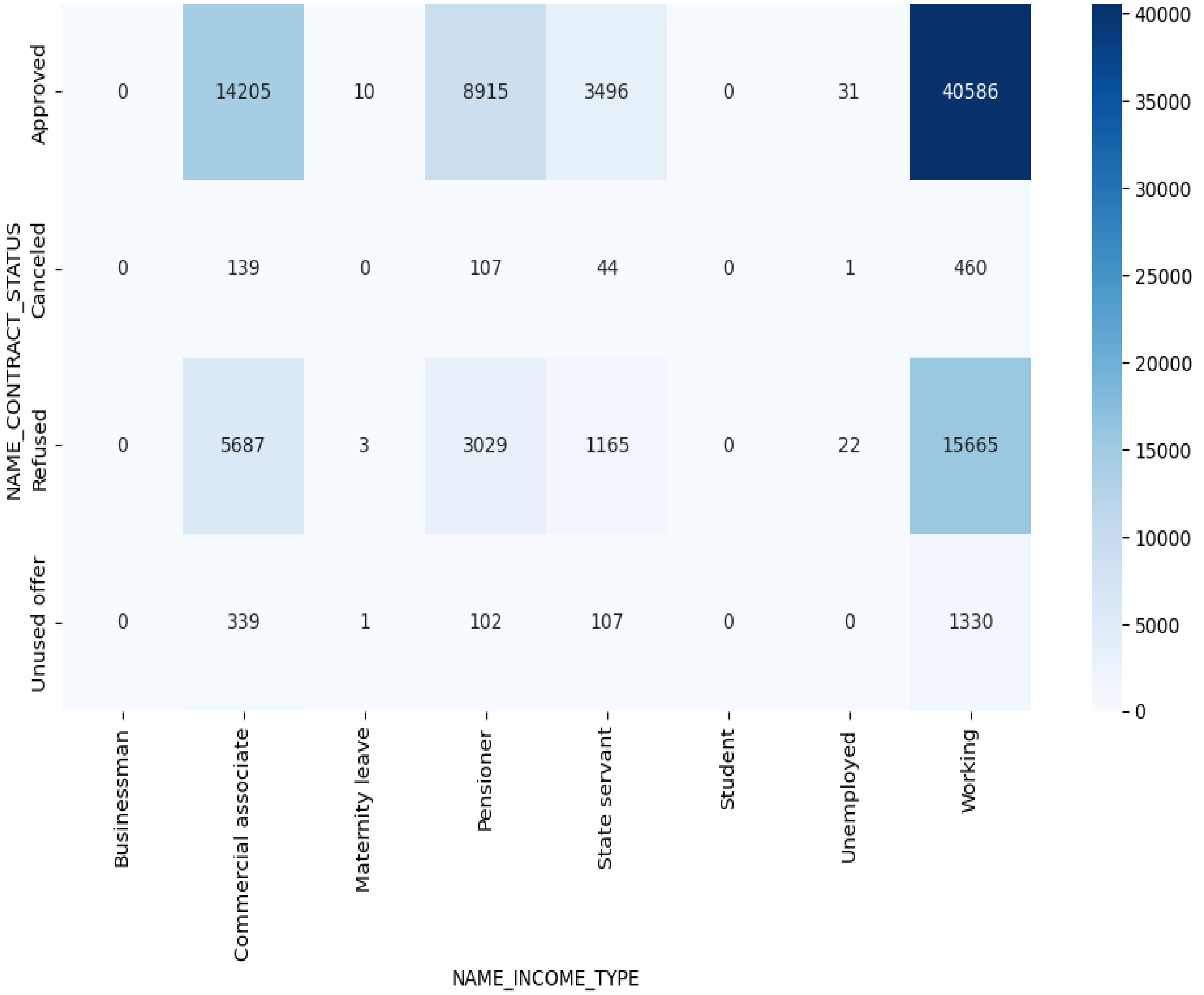


Heat map – Client type vs Age group

Since target values are 1 & 0. Heatmap numbers reflect the correlation of Default. Middle age Repeater client more prone to default. Young age New less riskier to default relative to Young age Repeater. Refreshed has least amount of defaulters.

Heatmap of variables over default cases

- 15,665 cases of previously Refused applicants defaulting currently.
- Working applicants have been approved the highest previously .



NAME_CONTRACT_STATUS	Businessman	Commercial associate	Maternity leave	Pensioner	State servant	Student	Unemployed	Working
Approved	0	14205	10	8915	3496	0	31	40586
Canceled	0	139	0	107	44	0	1	460
Refused	0	5687	3	3029	1165	0	22	15665
Unused offer	0	339	1	102	107	0	0	1330

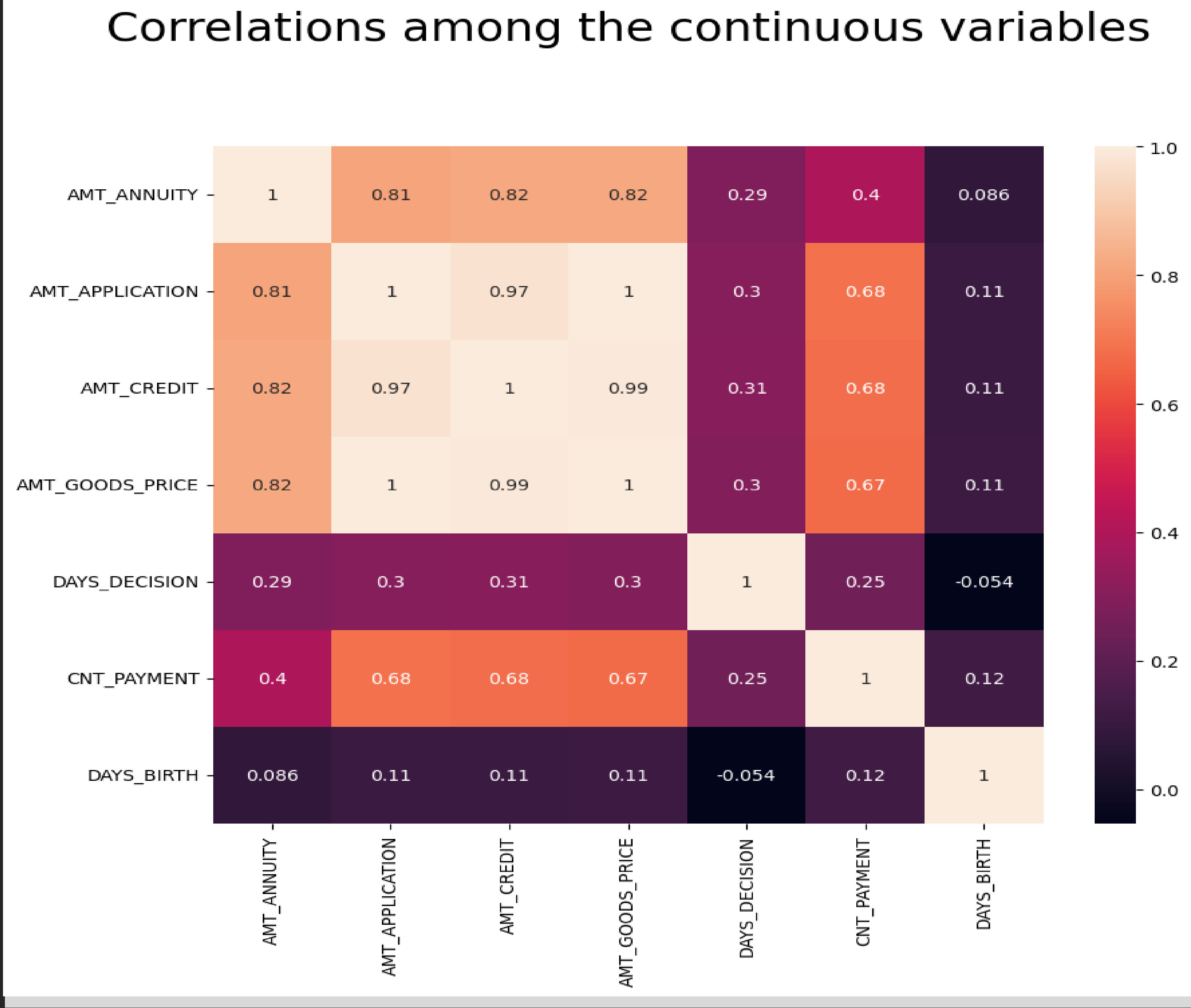
Correlation among continuous

- Days_Birth has correlation with all other variables.

- As the Amt_Credit, Amt_application and Amt_goods_price increase Count of payments have also increased.

- All the amount categories have strong correlation with each other.

- Days_decision has negative correlation with Days_birth.



Correlations among the continuous variables

	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	CNT_PAYMENT	DAYS_BIRTH
AMT_ANNUITY	1	0.81	0.82	0.82	0.29	0.4	0.086
AMT_APPLICATION	0.81	1	0.97	1	0.3	0.68	0.11
AMT_CREDIT	0.82	0.97	1	0.99	0.31	0.68	0.11
AMT_GOODS_PRICE	0.82	1	0.99	1	0.3	0.67	0.11
DAYS_DECISION	0.29	0.3	0.31	0.3	1	0.25	-0.054
CNT_PAYMENT	0.4	0.68	0.68	0.67	0.25	1	0.12
DAYS_BIRTH	0.086	0.11	0.11	0.11	-0.054	0.12	1

SUMMARY

Factors driving the default :

- Middle income + Male + Working + Business Type 3 + No Car/Realty Ownership
- There are also segments within each who have not defaulted, therefore

Suggestions to avoid refusing credible applications:

- Prioritizing new applications to repeated ones.
- Prioritizing and encouraging female applicants.
- Careful processing of application when Amt credit is high.
- Segmenting refused/cancelled applicants for thorough processing.