CREDIT EDA CASE STUDY

MANISH & RATNESH

PROBLEM STATEMENT

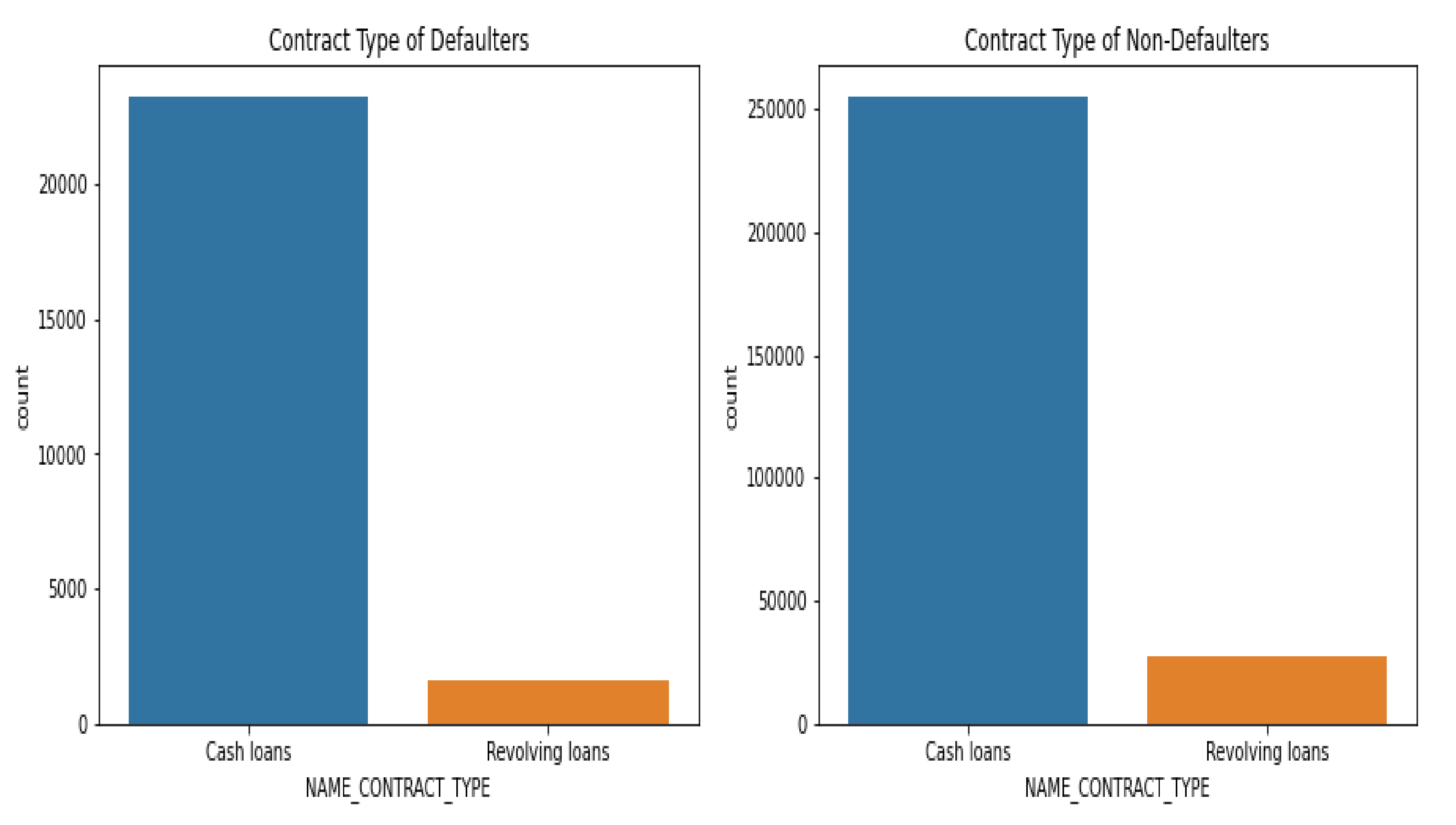
There are two types of risks associated with the Bank 'decision.

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
2. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

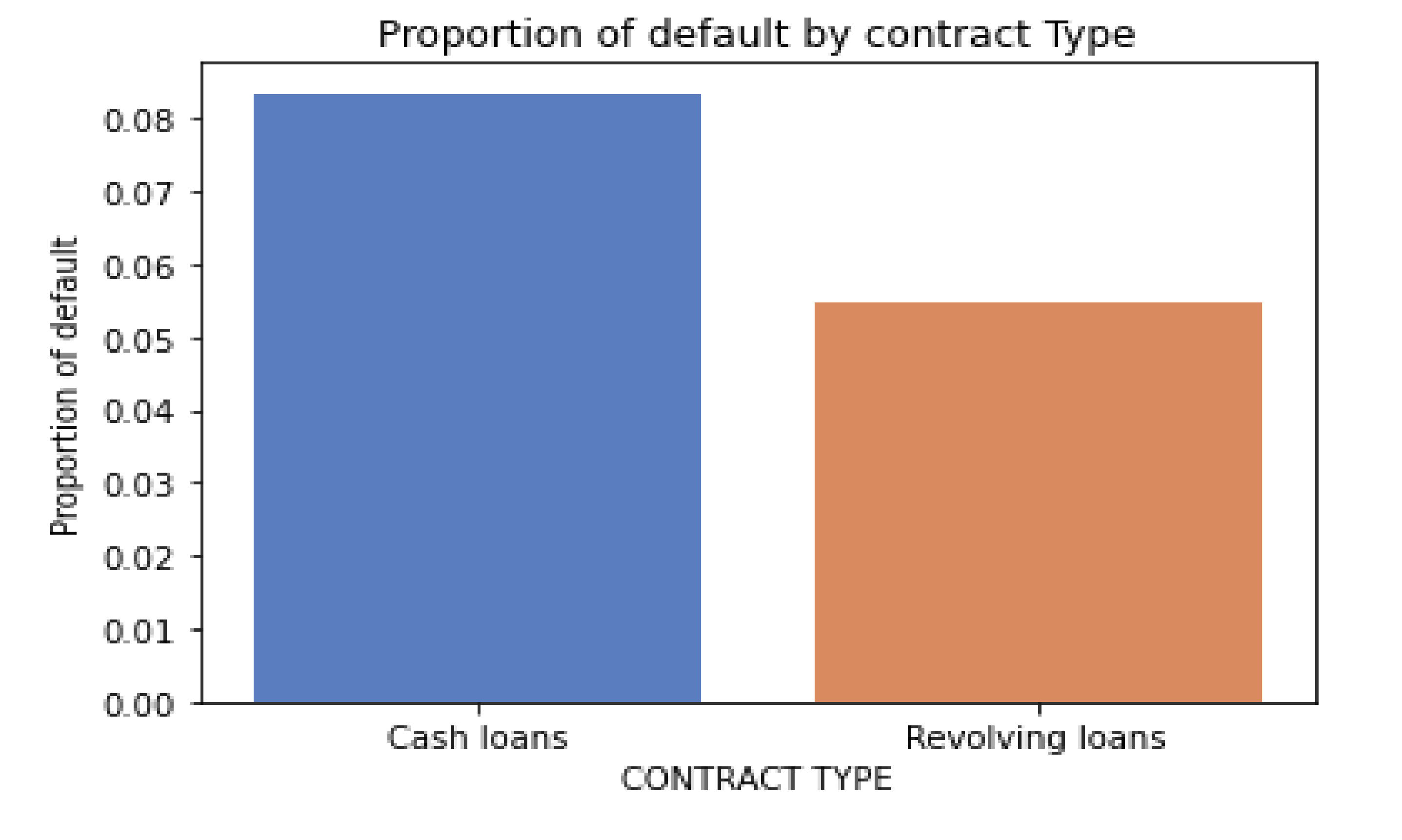
STEPS

1. Understanding the given data
2. Identify the data quality issue
3. Check for the data imbalance
4. Merging the data.
5. Data Analysis using univariate, bivariate and multivariate.

CONTRACT TYPE

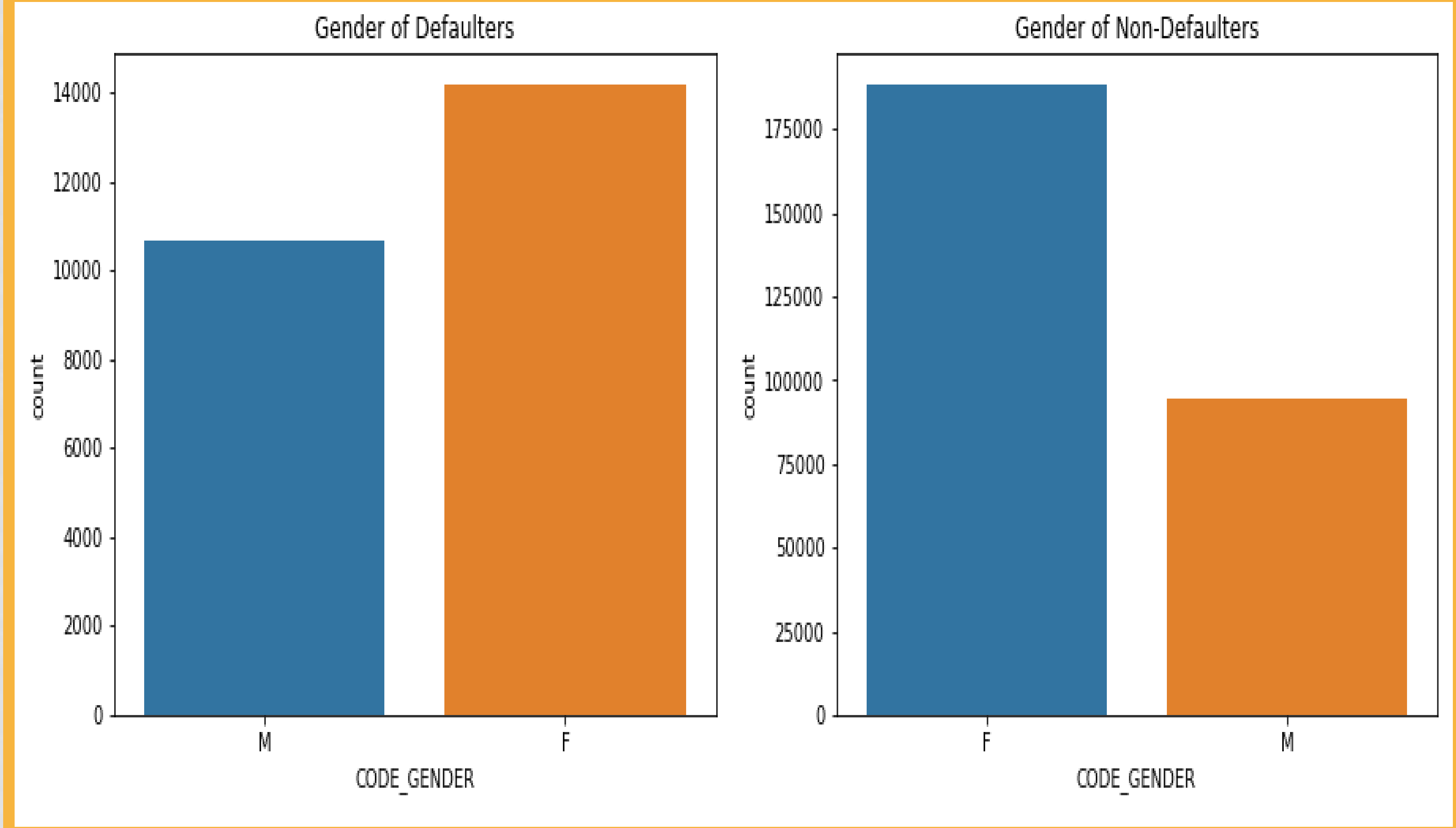


CONTRACT TYPE

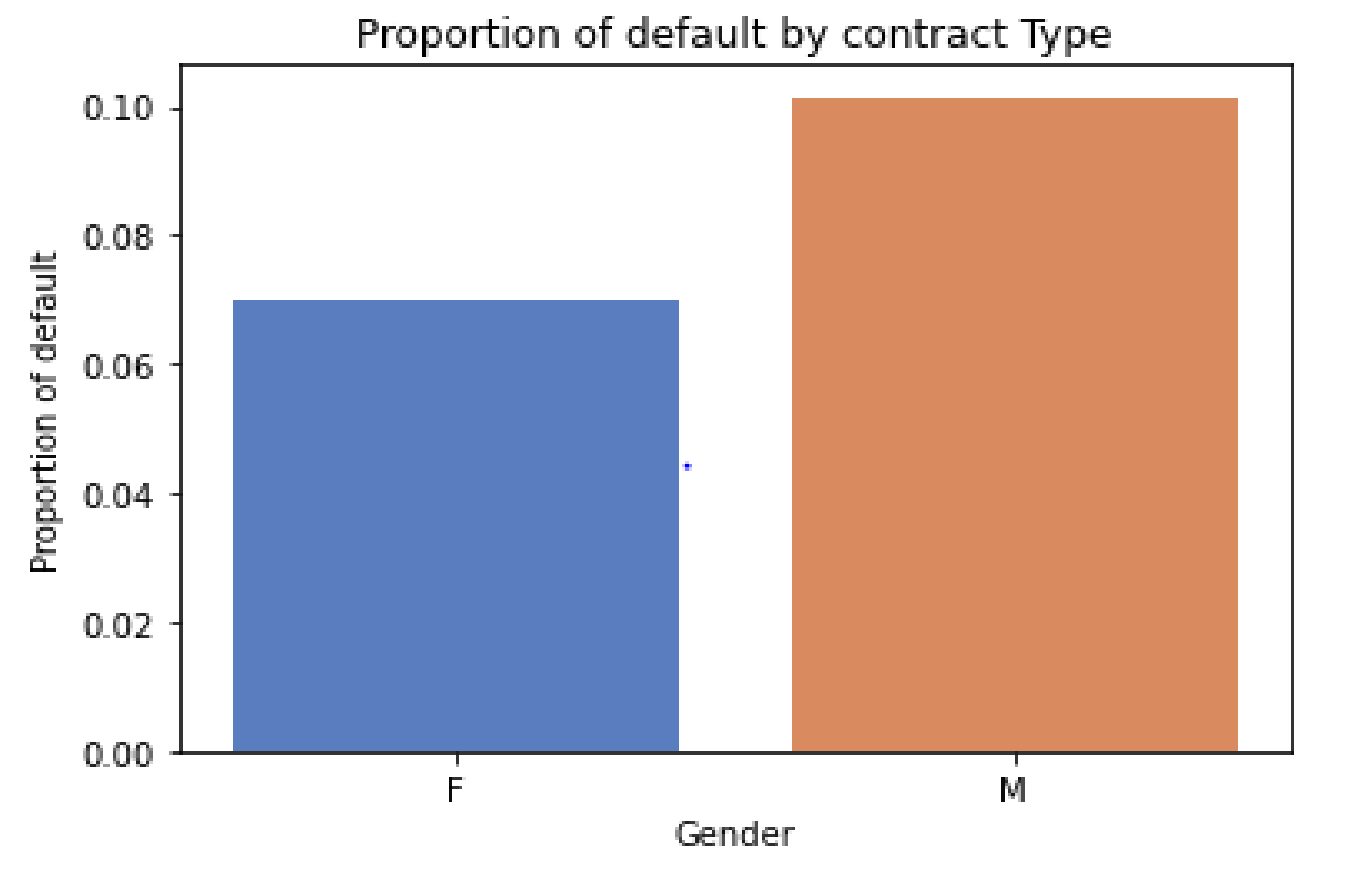


Number of Cash loans are higher in both default/ non-default cases than Revolving loans. The proportion of default is higher in cash loan category compared to the revolving loans

GENDER

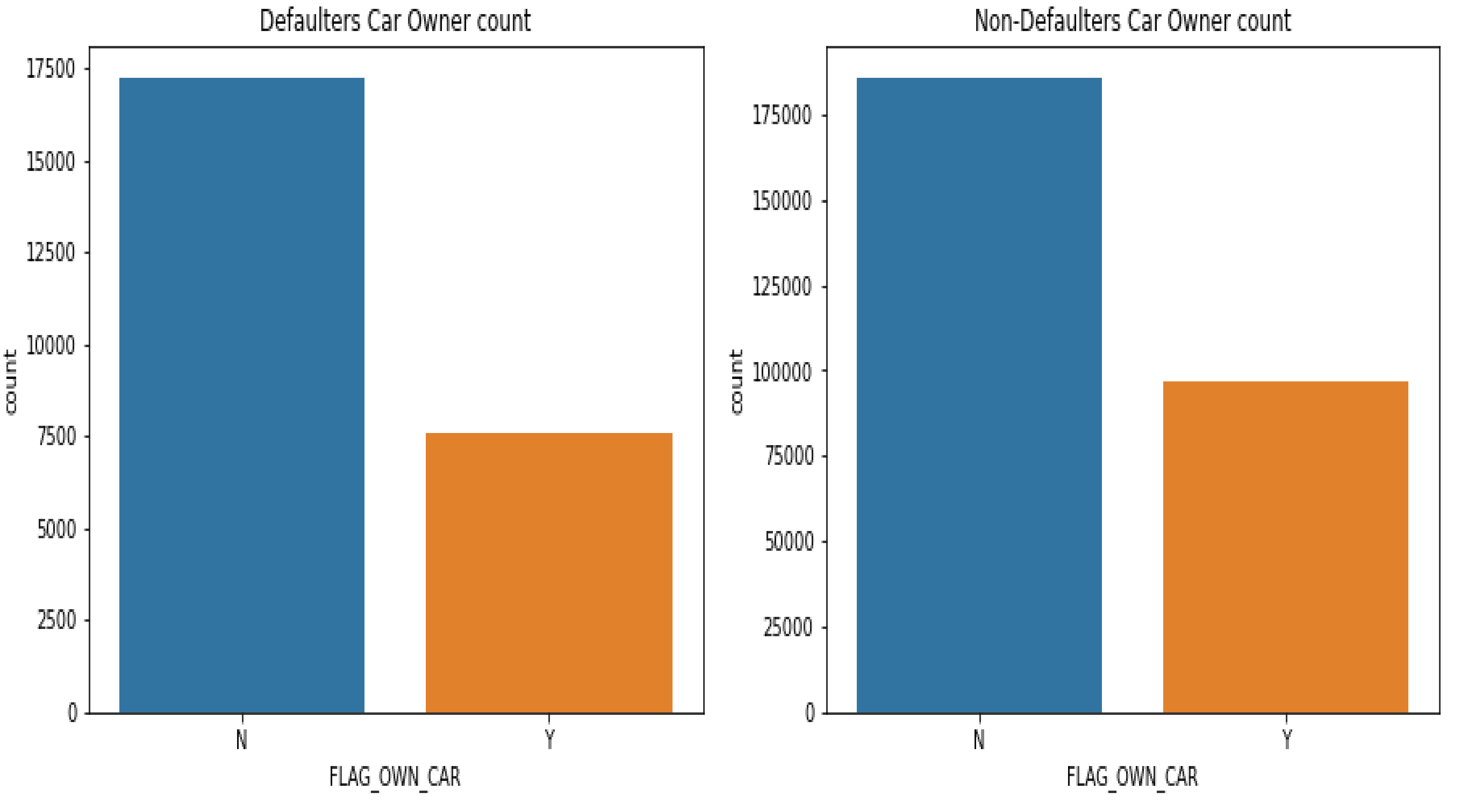

Gender of Defaulters
count
14000
12000
10000
8000
6000
4000
2000
0
M
F
CODE_GENDER
Gender of Non-Defaulters
count
175000
150000
125000
100000
75000
50000
25000
0
F
M
CODE_GENDER

GENDER

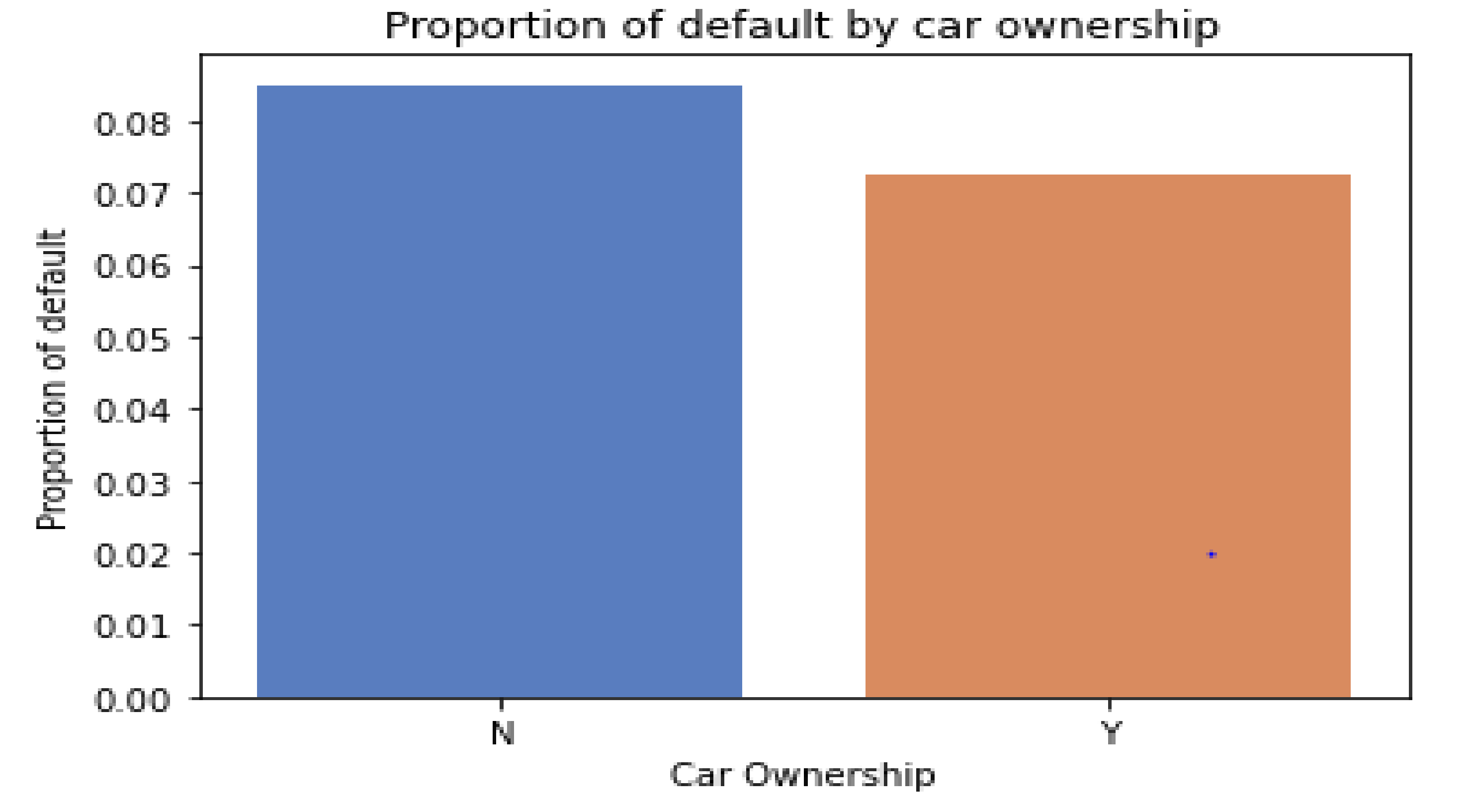


Number of Cash loans are higher in both default/ non-default cases than Revolving loans. The proportion of default is higher in cash loan category compared to the revolving loans

CAR OWNERSHIP STATUS

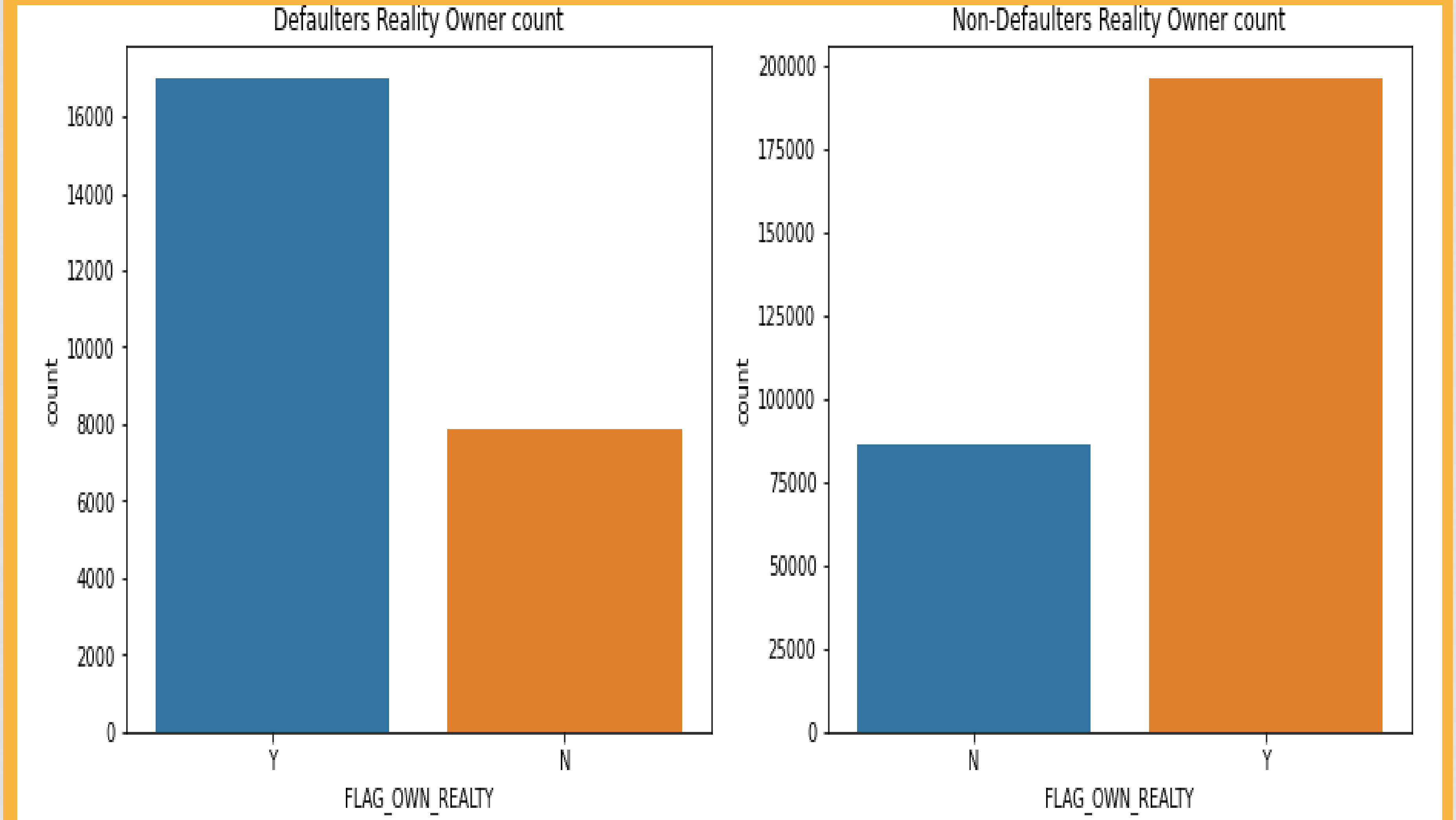


CAR OWNERSHIP STATUS

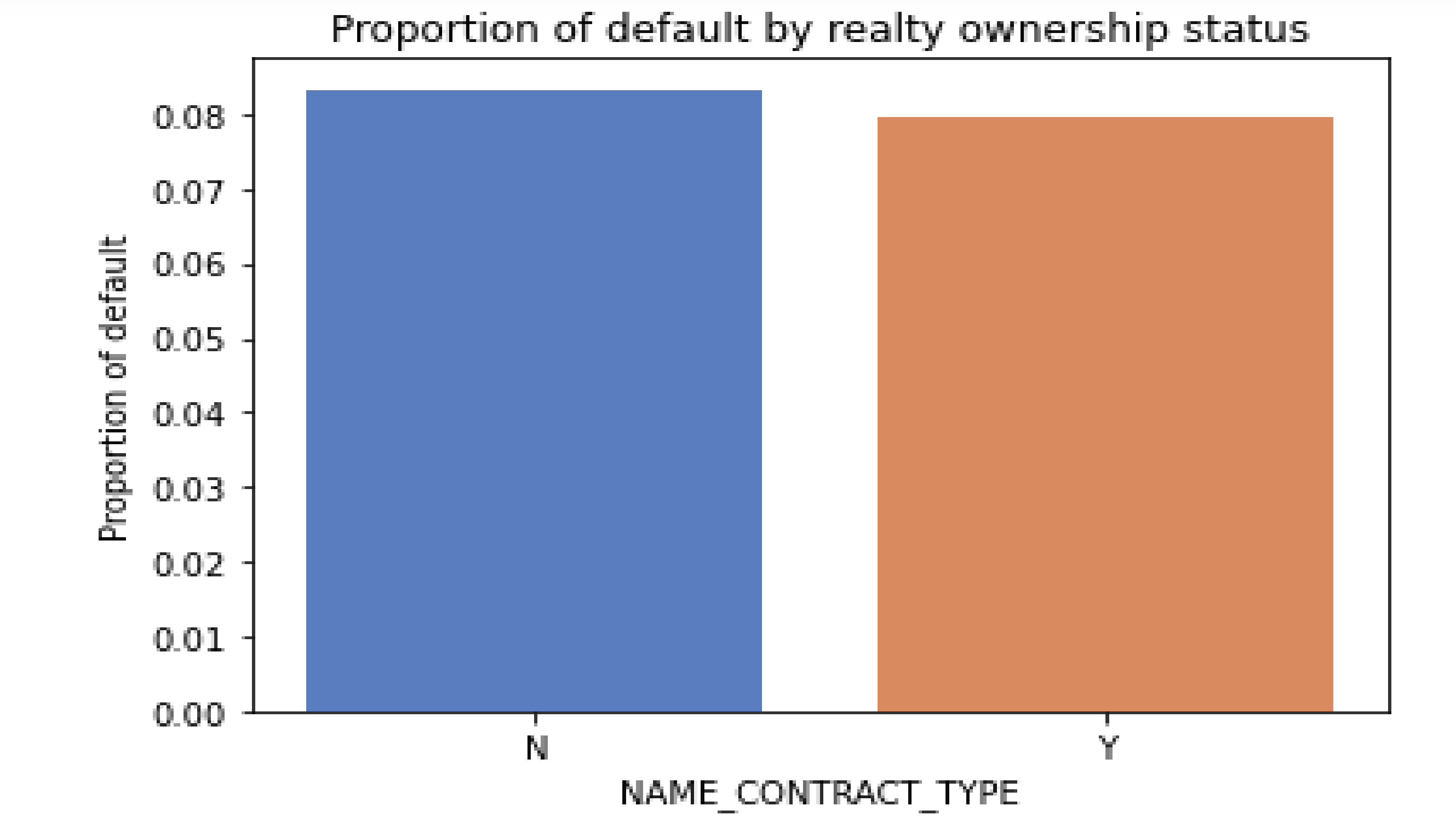


Default/Non-default both the cases, the count of client who do not own car is higher than who owns a car. The proportion of default is higher for non-car owners relative to the car owners.

REALTY OWNERSHIP STATUS

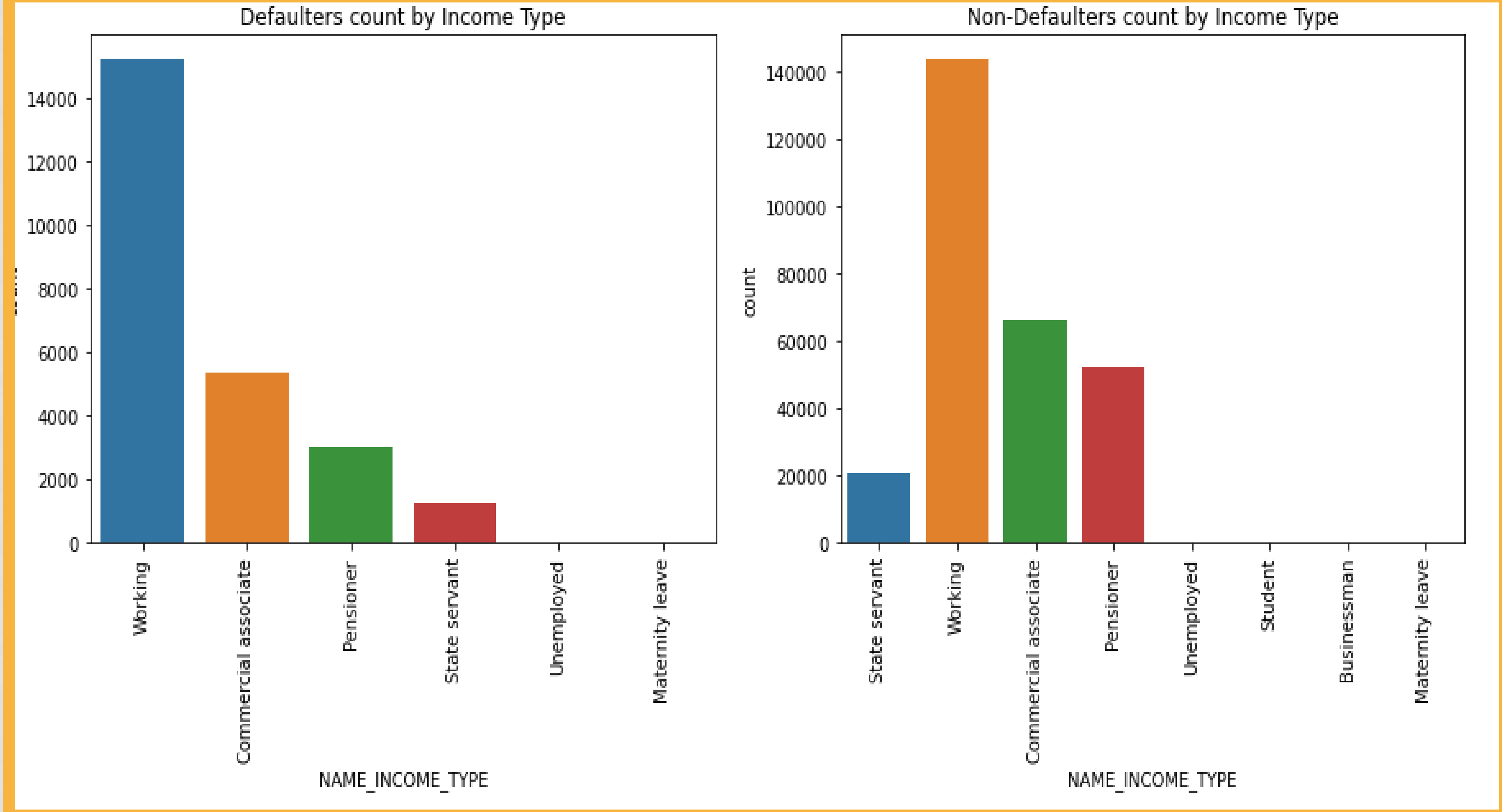


REALTY OWNERSHIP STATUSTYPE

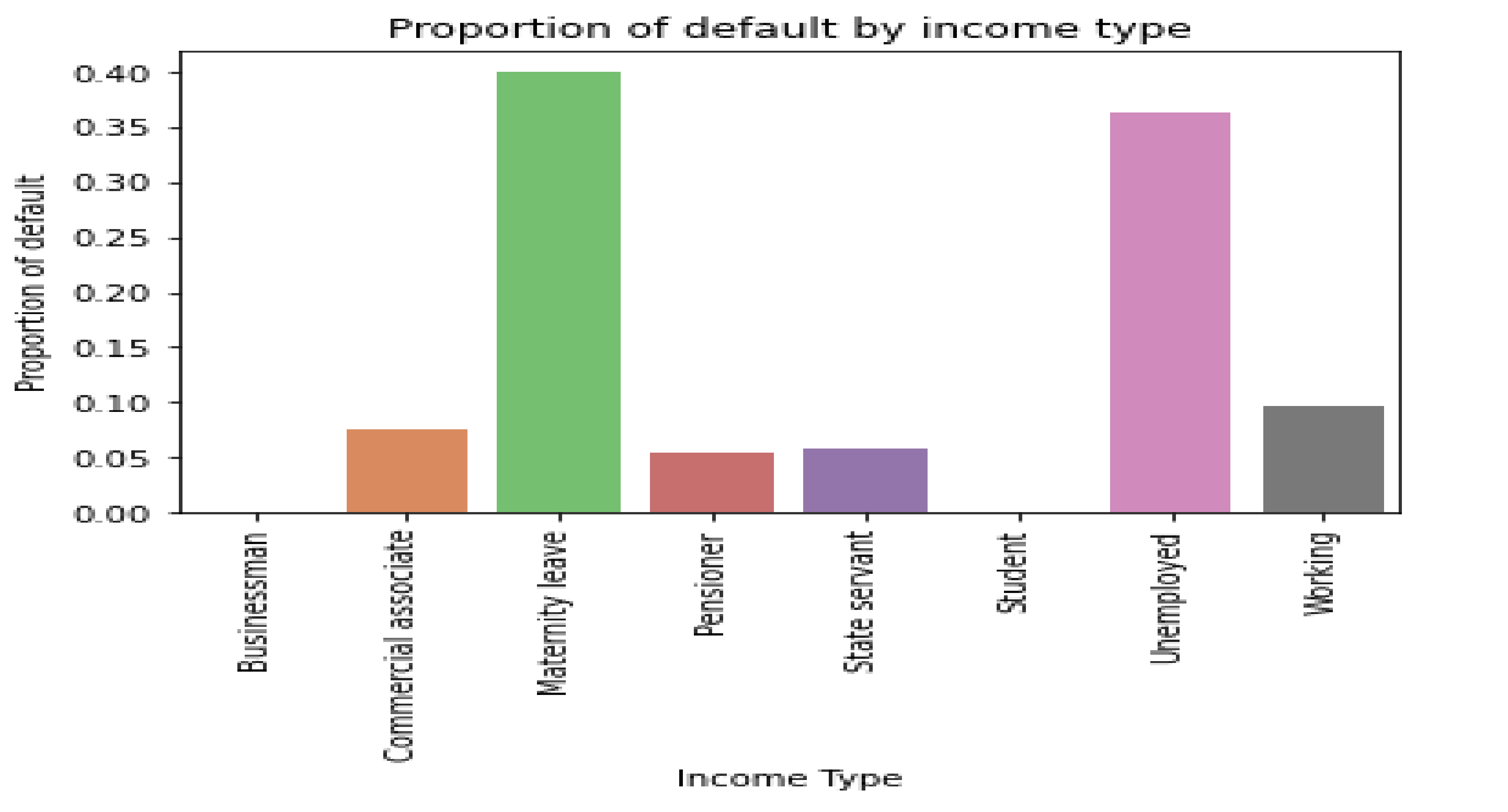


Default/Non-default both the cases, the count of client who do own real estate is higher than who has not. Applicants with no realty ownership has a higher propensity to default than the clients who own real estate

INCOME TYPE

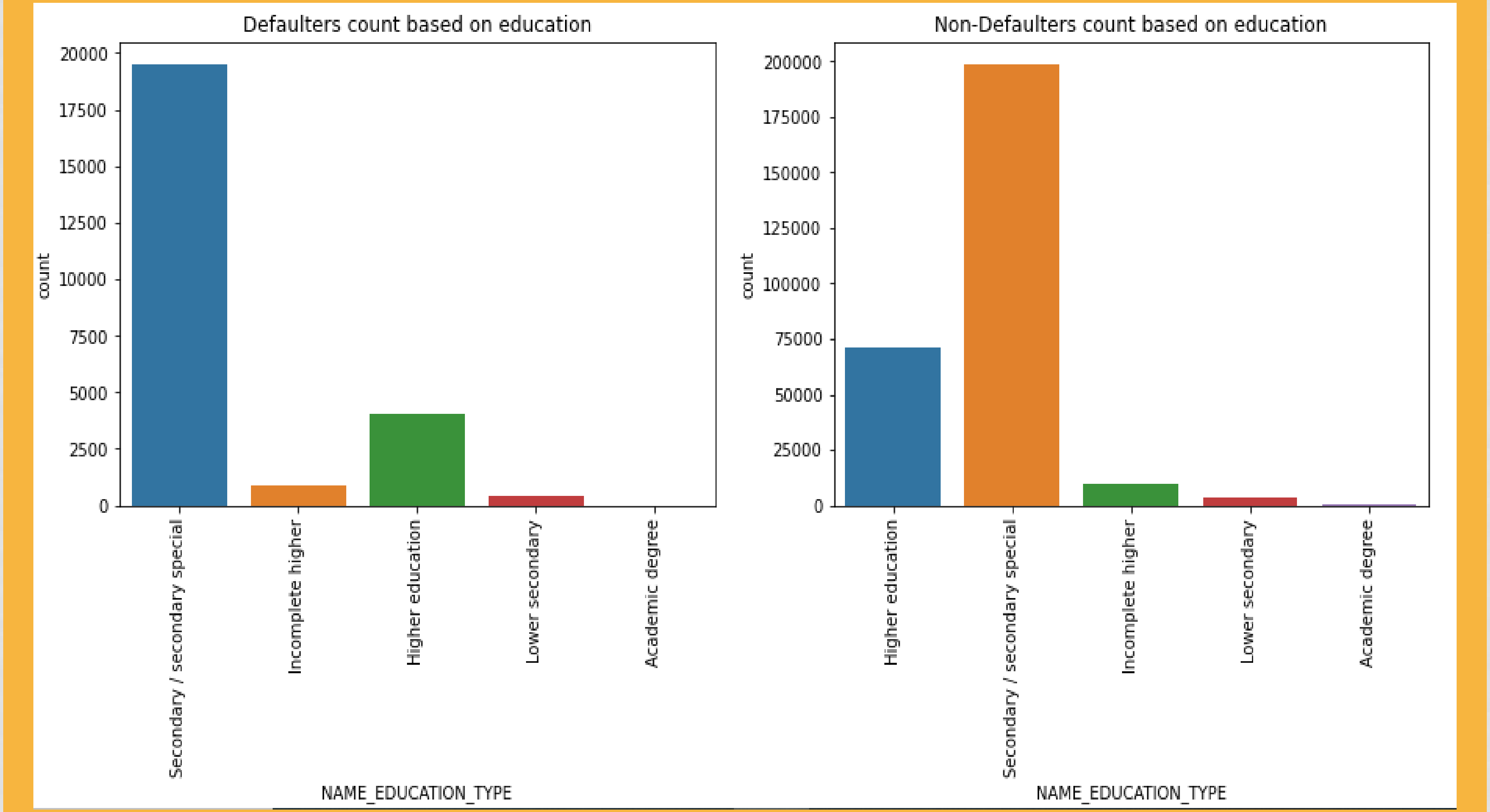

Defaulters count by Income Type
14000
12000
10000
8000
6000
4000
2000
0
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
NAME_INCOME_TYPE
Non-Defaulters count by Income Type
count
140000
120000
100000
80000
60000
40000
20000
0
State servant
Working
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE

INCOME TYPE

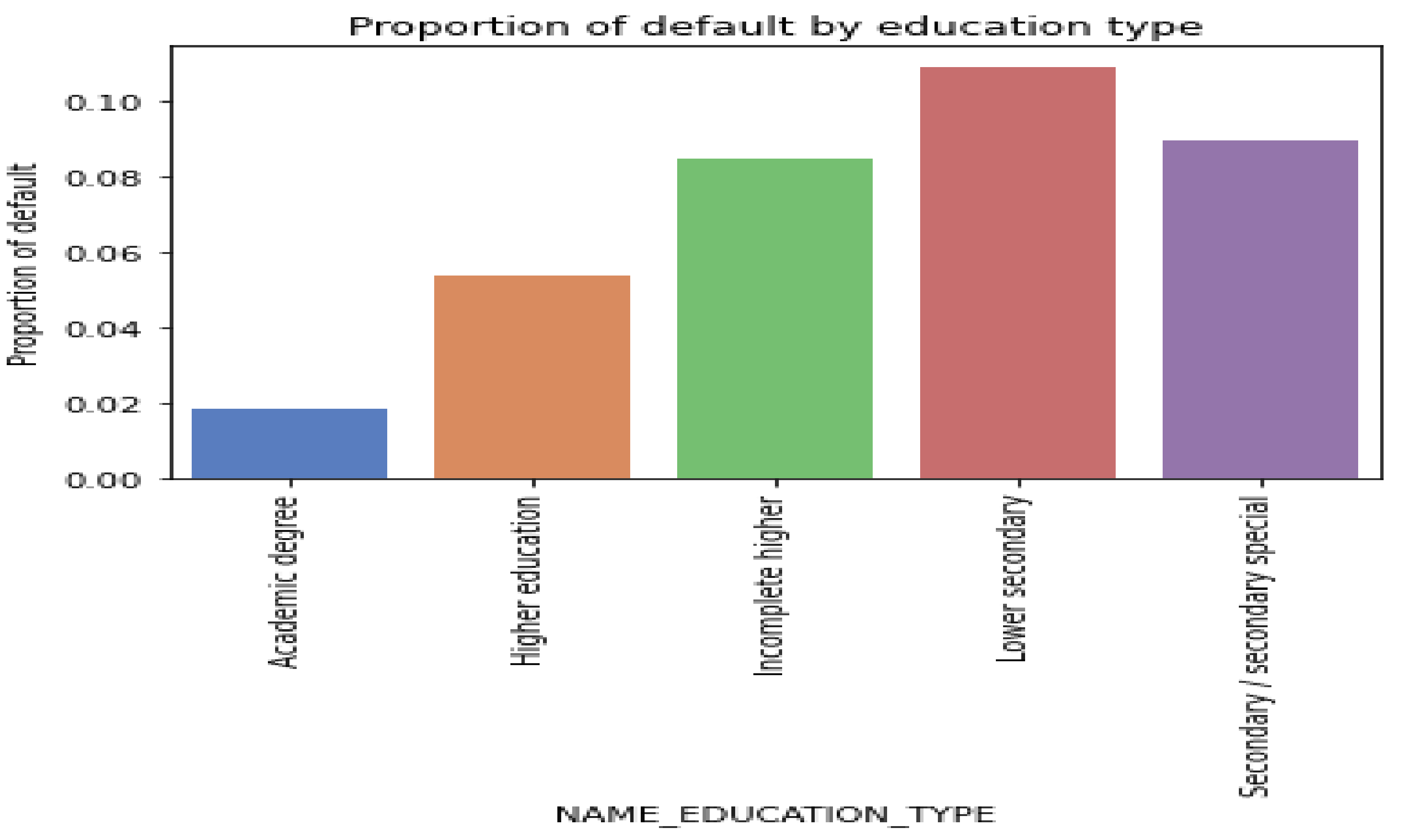


Majority of the applicants are from working, commercial associate, pensioners and state servants. The remaining categories of income types are very small. The proportion of default is high among the working and the commercial associates. It is relatively lower for the pensioner and state servant.

EDUCATION TYPE

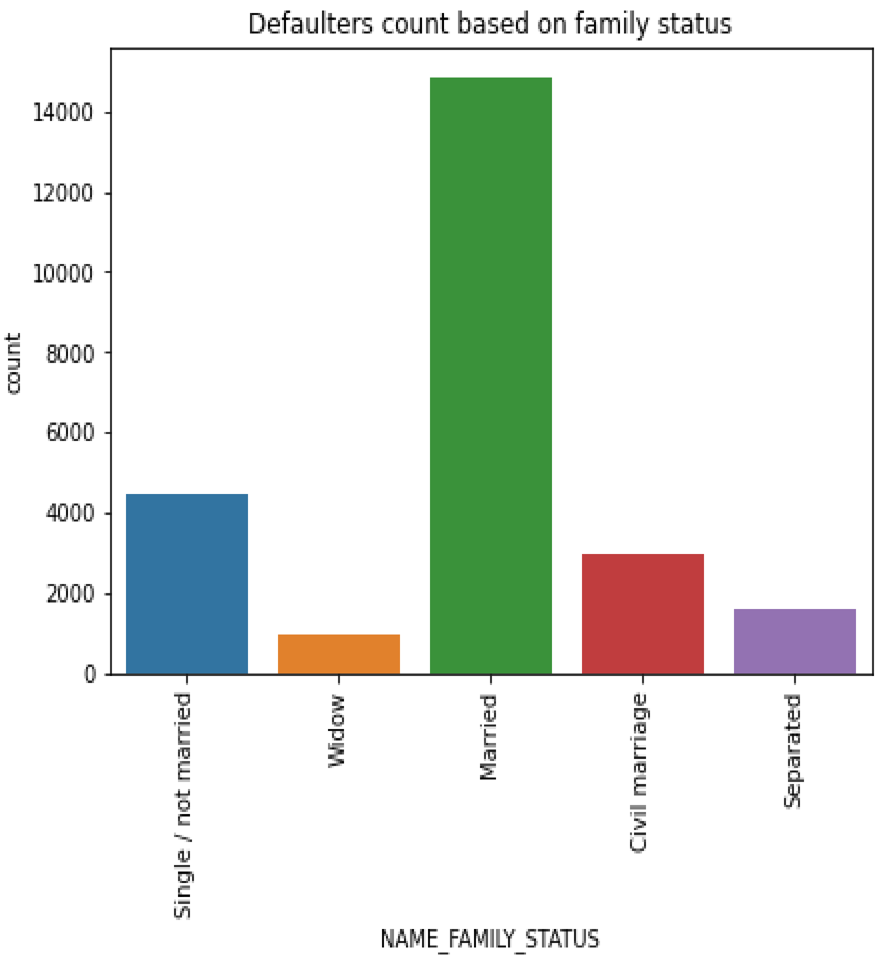


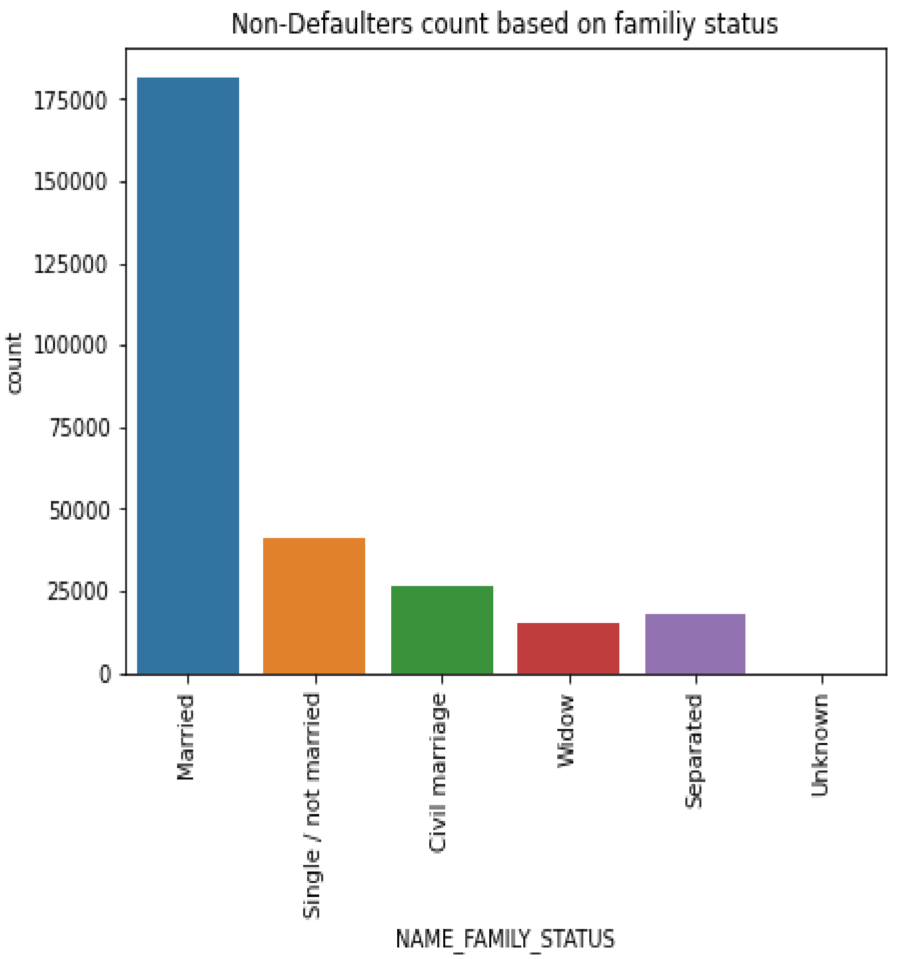
EDUCATION TYPE



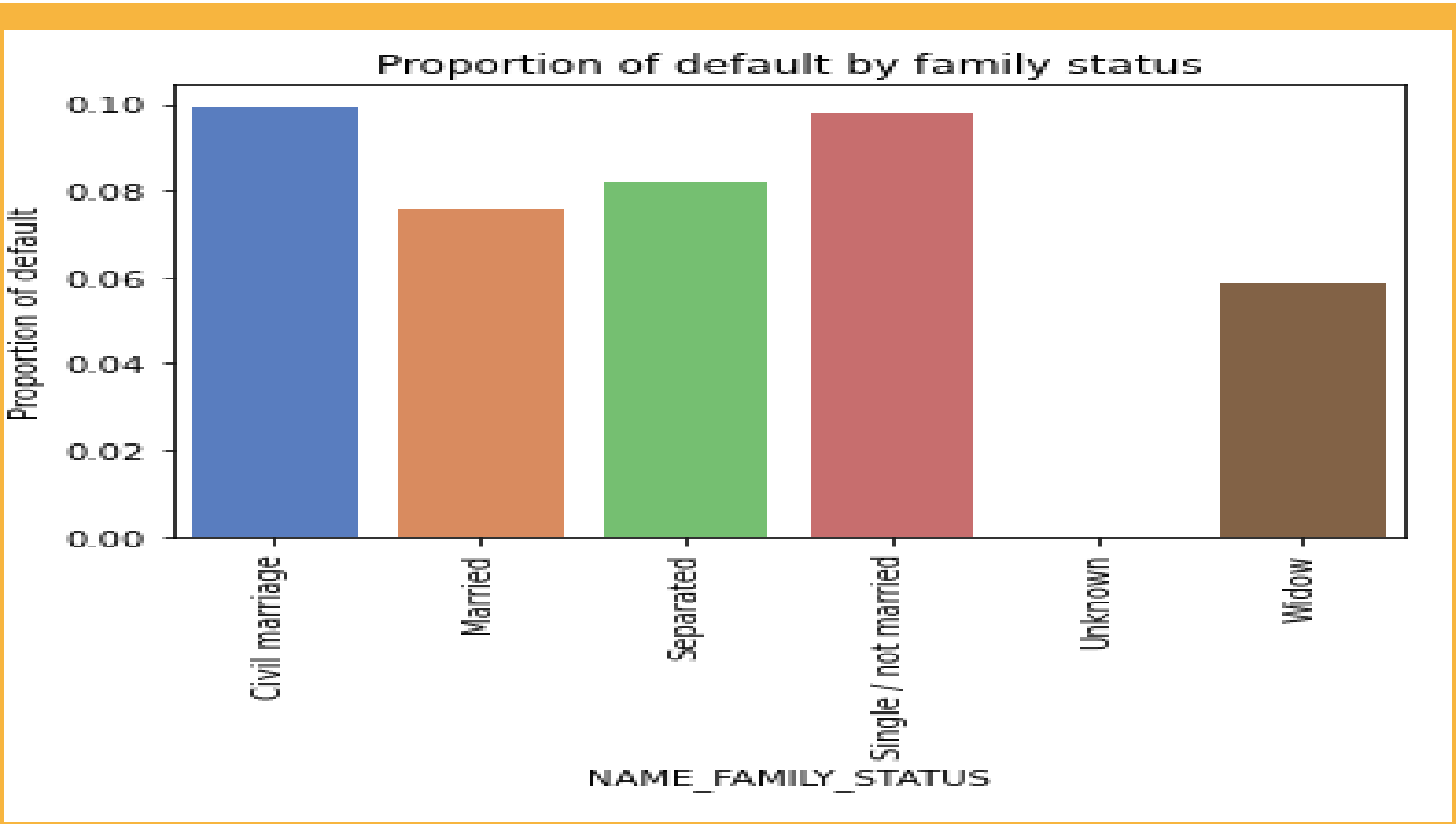
Applicants with secondary and higher secondary education are among the highest defaulters as well as not defaulters. Whereas, applicants with academic degrees are the smallest group of applicants that have applied for the loan and applicants from this background has no recorded of default. From the above figure, we see that a distinct pattern emerges. The chances of default is lower as the education level of the applicants increases.

FAMILY STATUS


Defaulters count based on family status

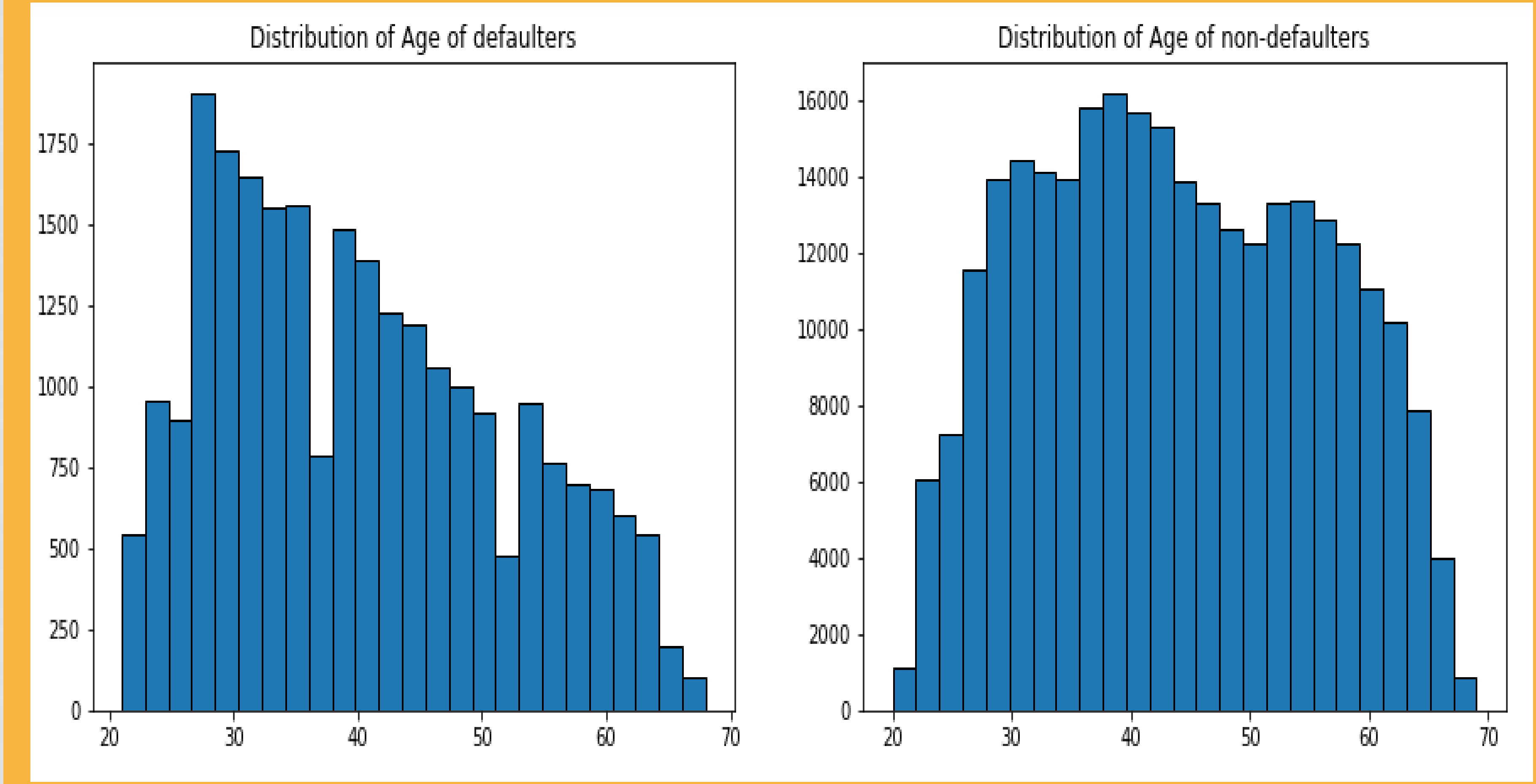

Non-Defaulters count based on familiy status

FAMILY STATUS

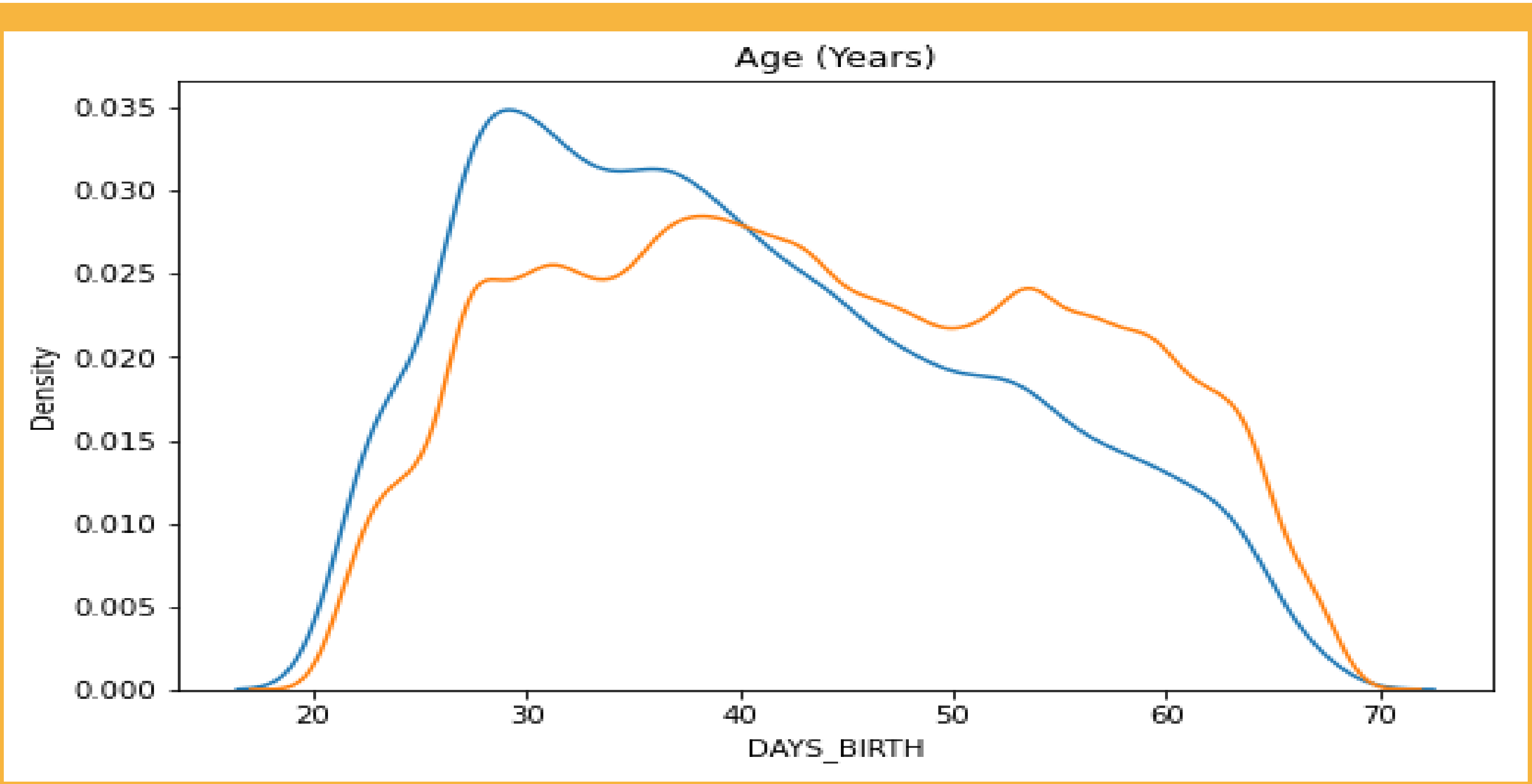


Applicants who are married are among the highest number of defaulters and non-defaulters. Whereas, widows are the lowest number of defaulters and non-defaulters. The proportion of default is the highest among the applicants who are in civil marriage category followed by applicants who are single.

AGE

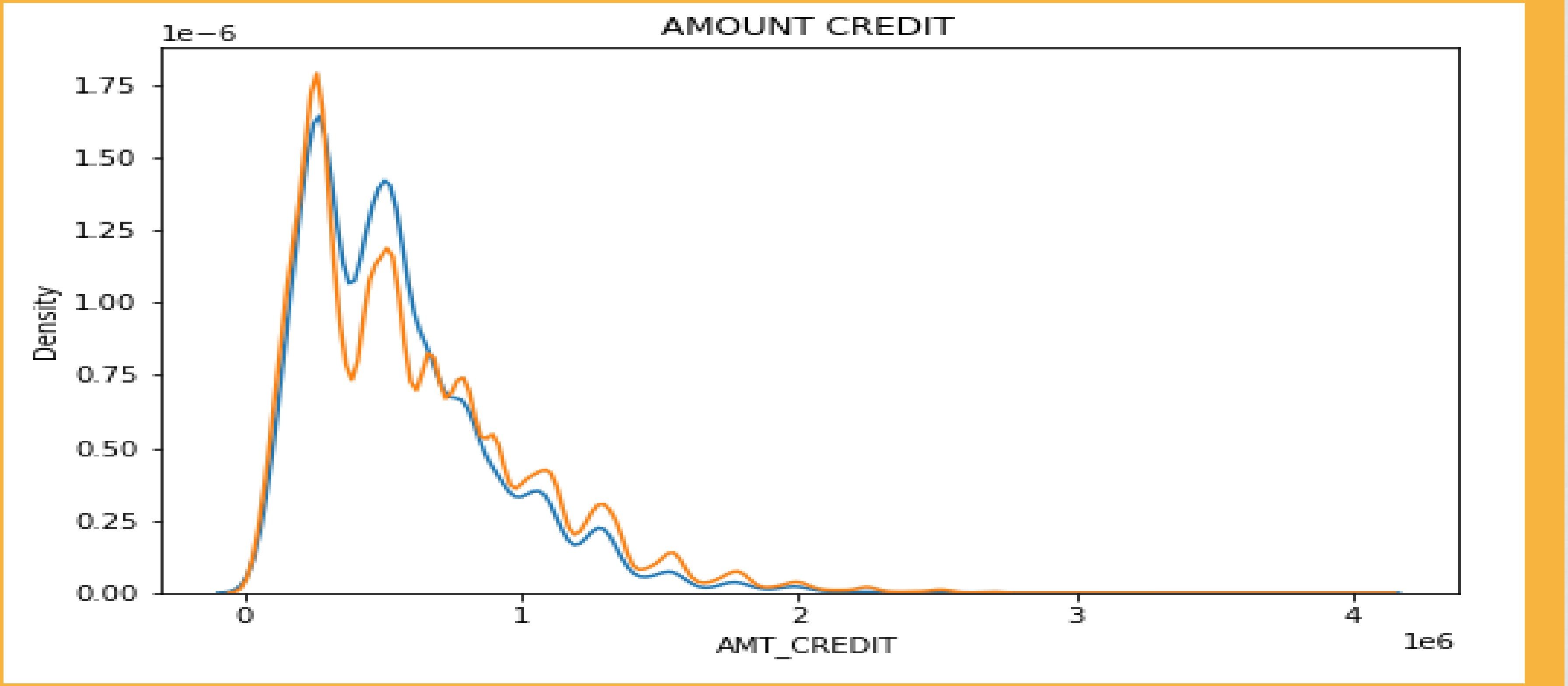

Distribution of Age of defaulters
1750
1500
1250
1000
750
500
250
0
20
30
40
50
60
70
Distribution of Age of non-defaulters
16000
14000
12000
10000
8000
6000
4000
2000
0
20
30
40
50
60
70

AGE



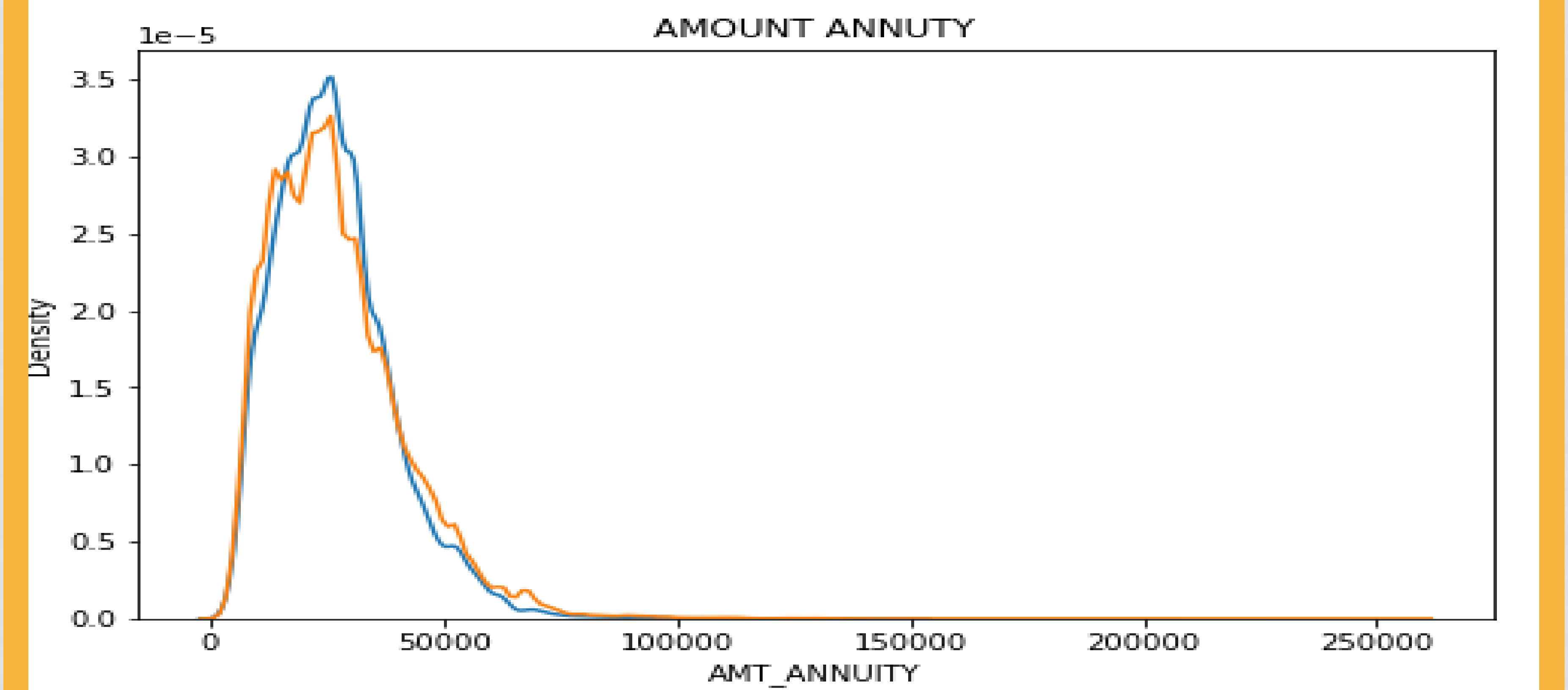
Around 29 years to 40 years people are more defaulters. There is high chance to be defaulted of the young people. Non-defaulted people are almost equally distributed.

AMOUNT CREDIT



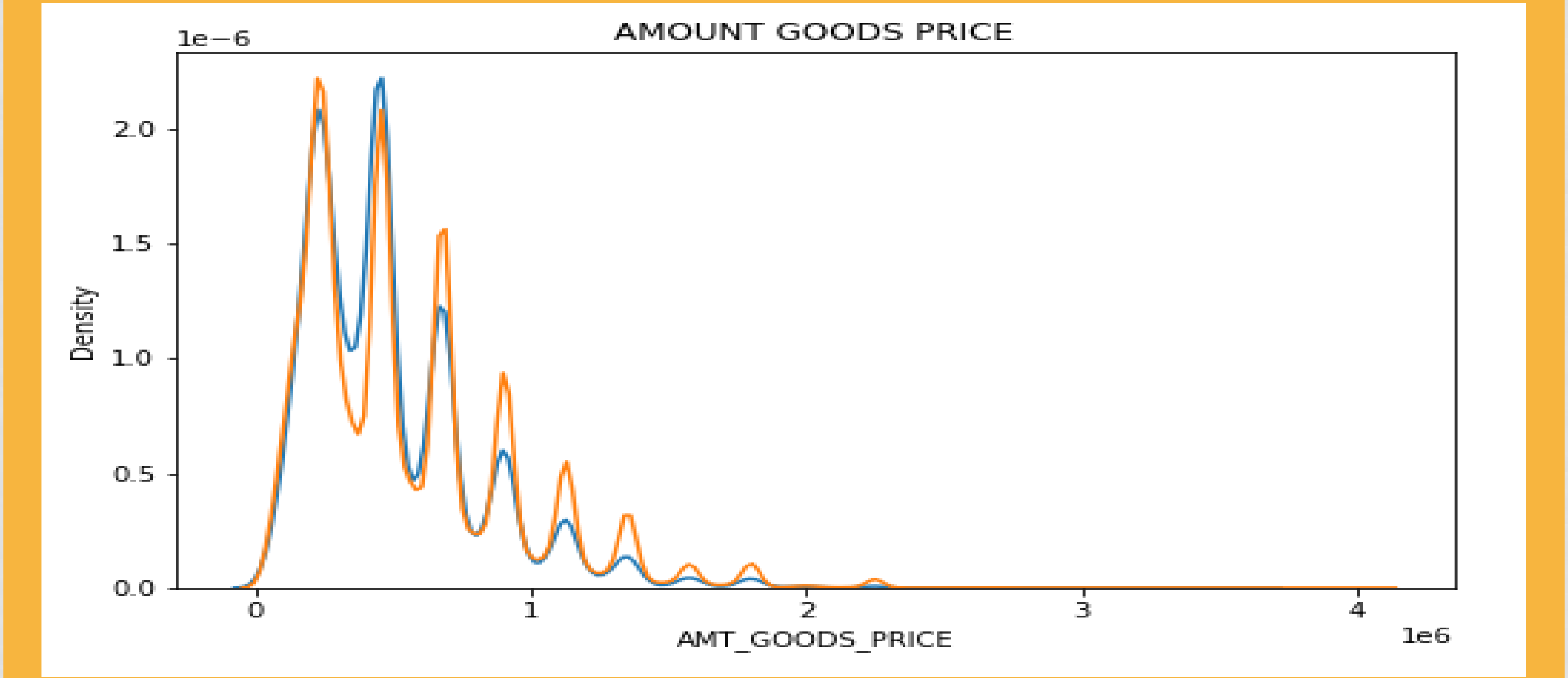
the lesser loan credit amount, the higher the default chances. We can do bivariate analysis with Occupation Type to find out more insight.

AMOUNT ANNUITY



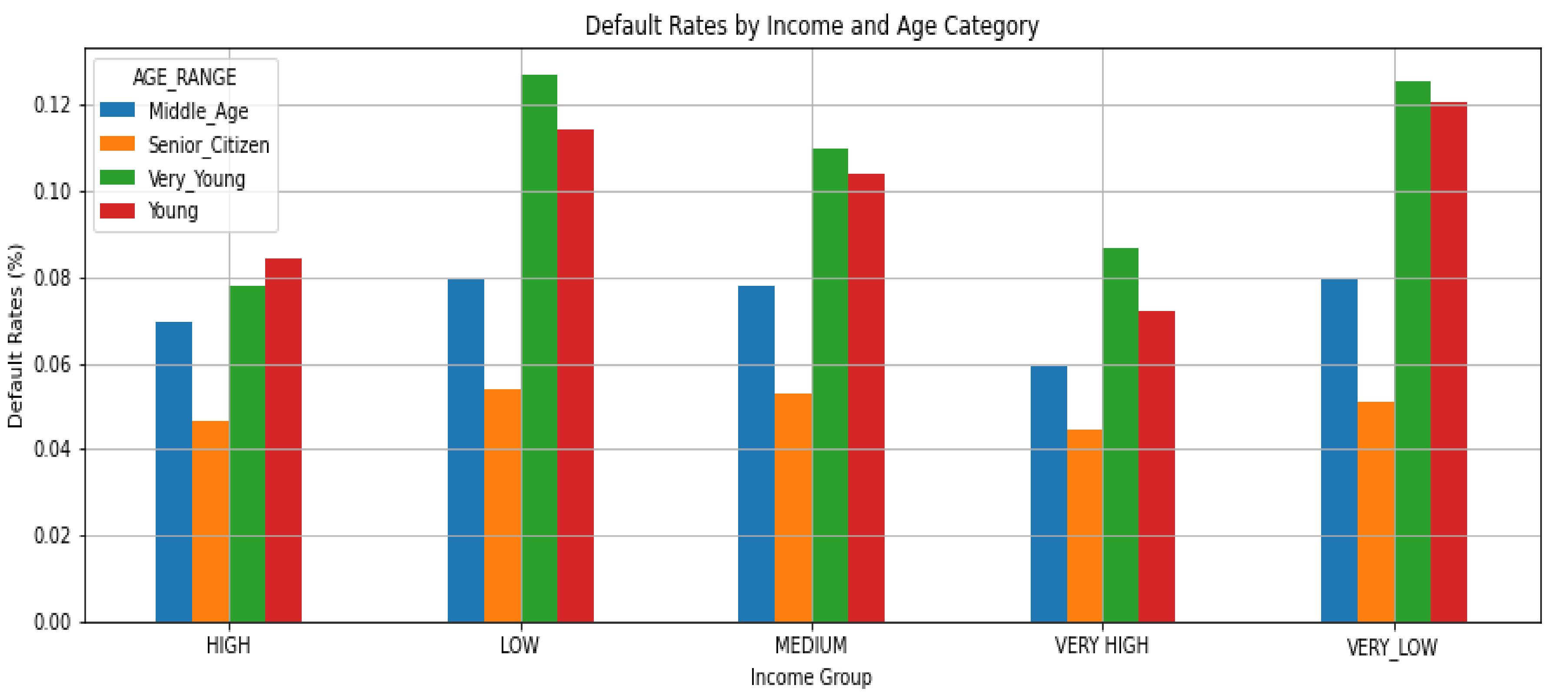
the same pattern in both the default and non-default. The loan annuity is mostly concentrated within 10000 to 40000 range in both the cases.

AMOUNT GOODS PRICE TYPE



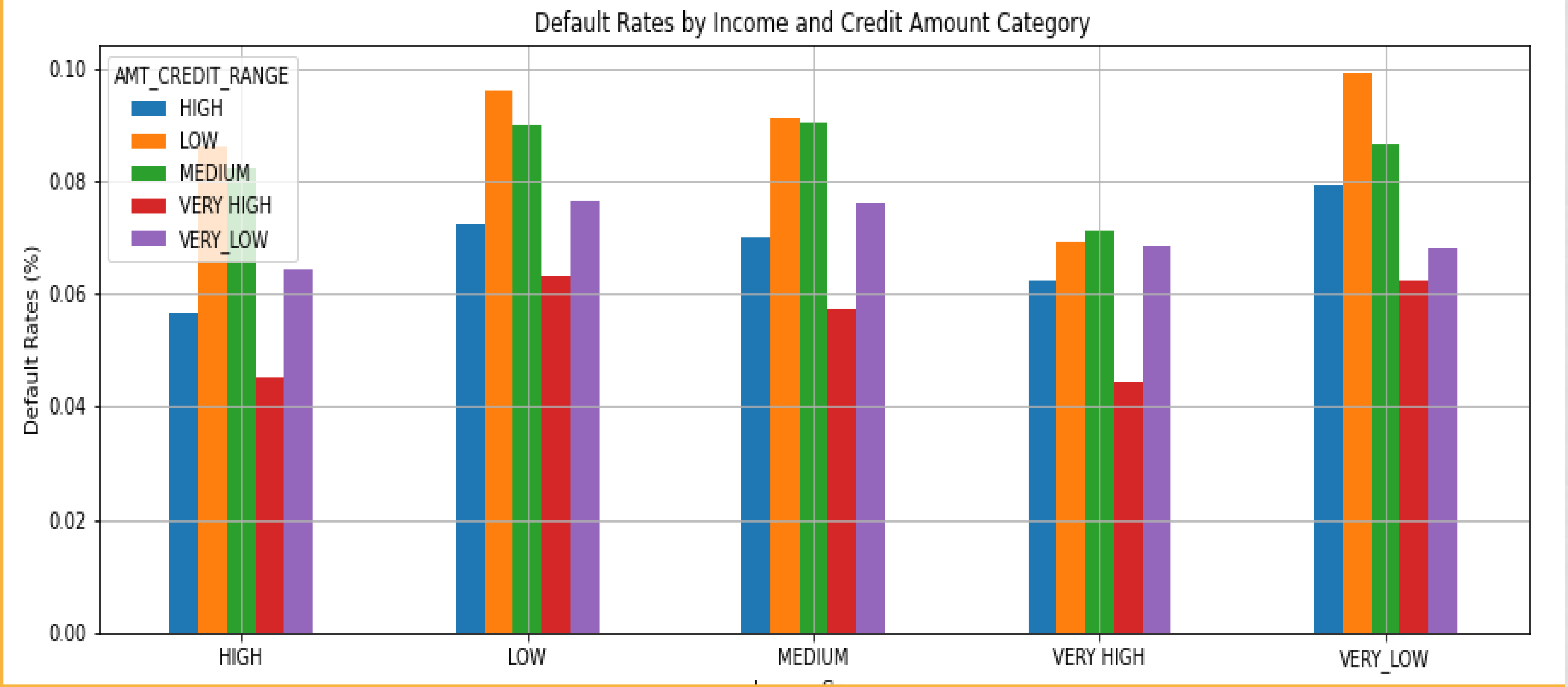
Both the curves are following the similar frequency distribution. We can see some spikes from 150000 to 220000, then around 500000 price. At this range people are more defaulted and higher the goods price, people are becoming the less defaulted. We can infer that, rich people are buying costly product and thus they are becoming less defaulted.

DEFAULT RATES BY INCOME AND AGE CATEGORY



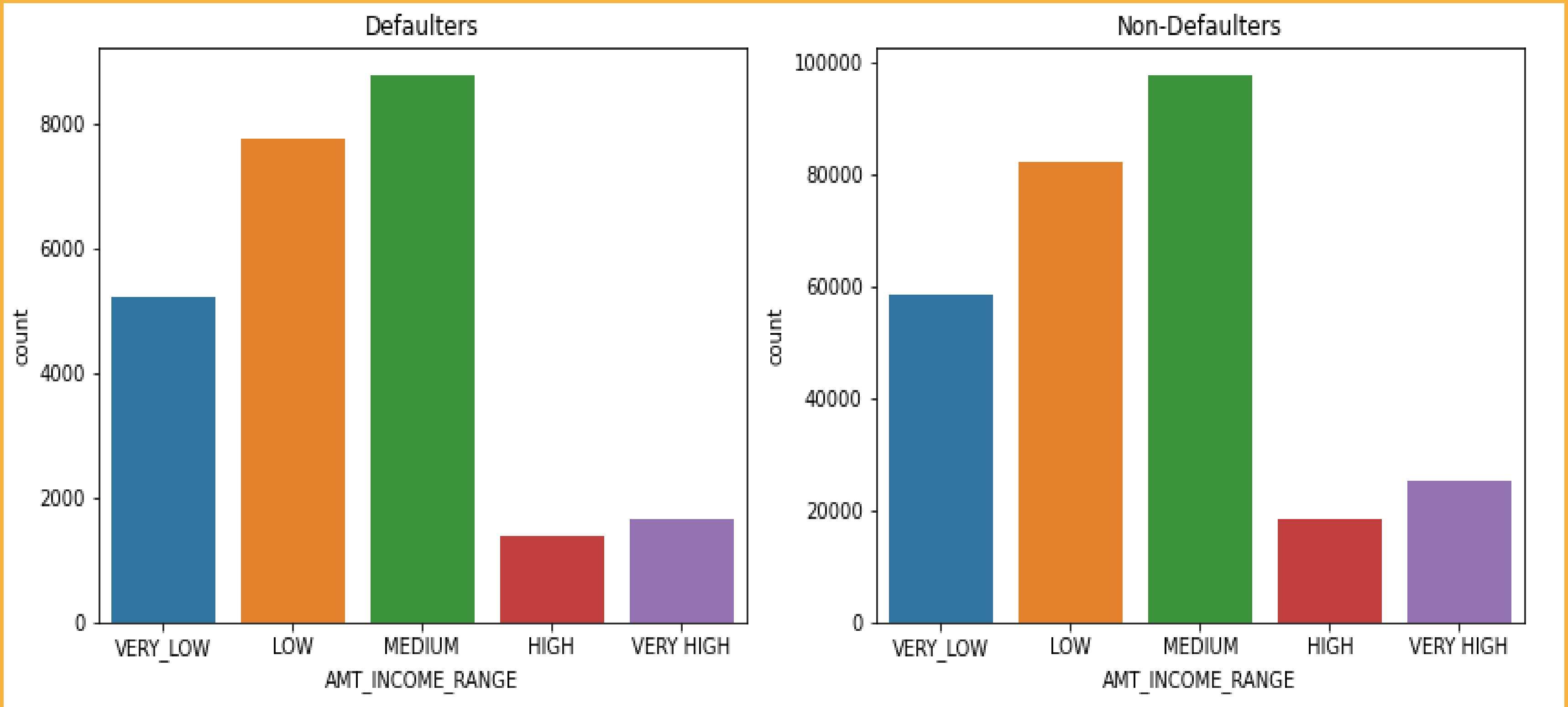
We see from the above diagram that irrespective of the income groups, the chances of default decreases as the age of the applicants increases.

DEFAULT RATES BY INCOME AND CREDIT AMOUNT CATEGORY



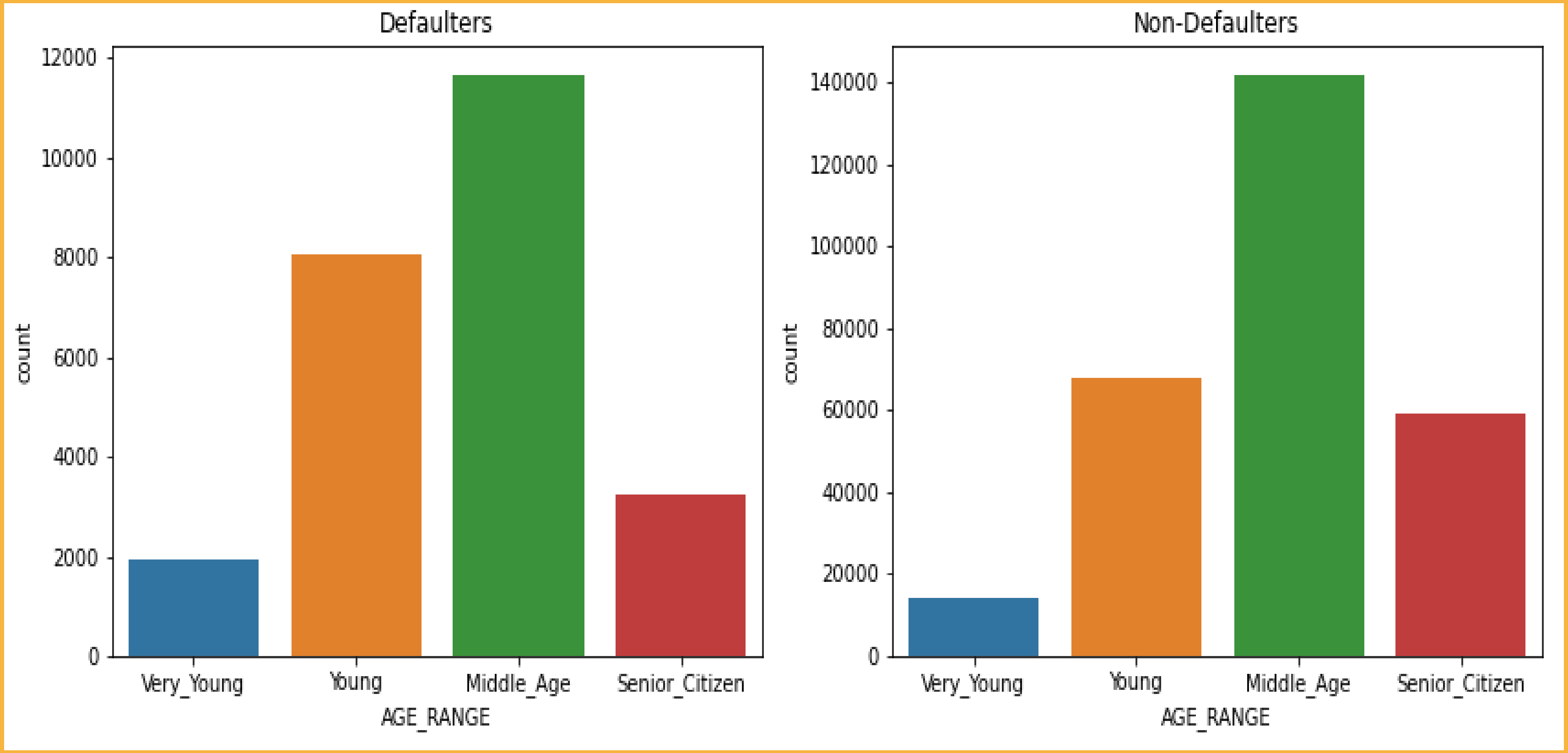
From the above plot, we find that irrespective of the income group, the chances of default increases as the credit amount increases. Also if we compare credit amount categories by different income groups, then the default rates for all the three credit amount categories are lower in the high income group relative to the medium and low income groups.

AMOUNT INCOME RANGE TYPE



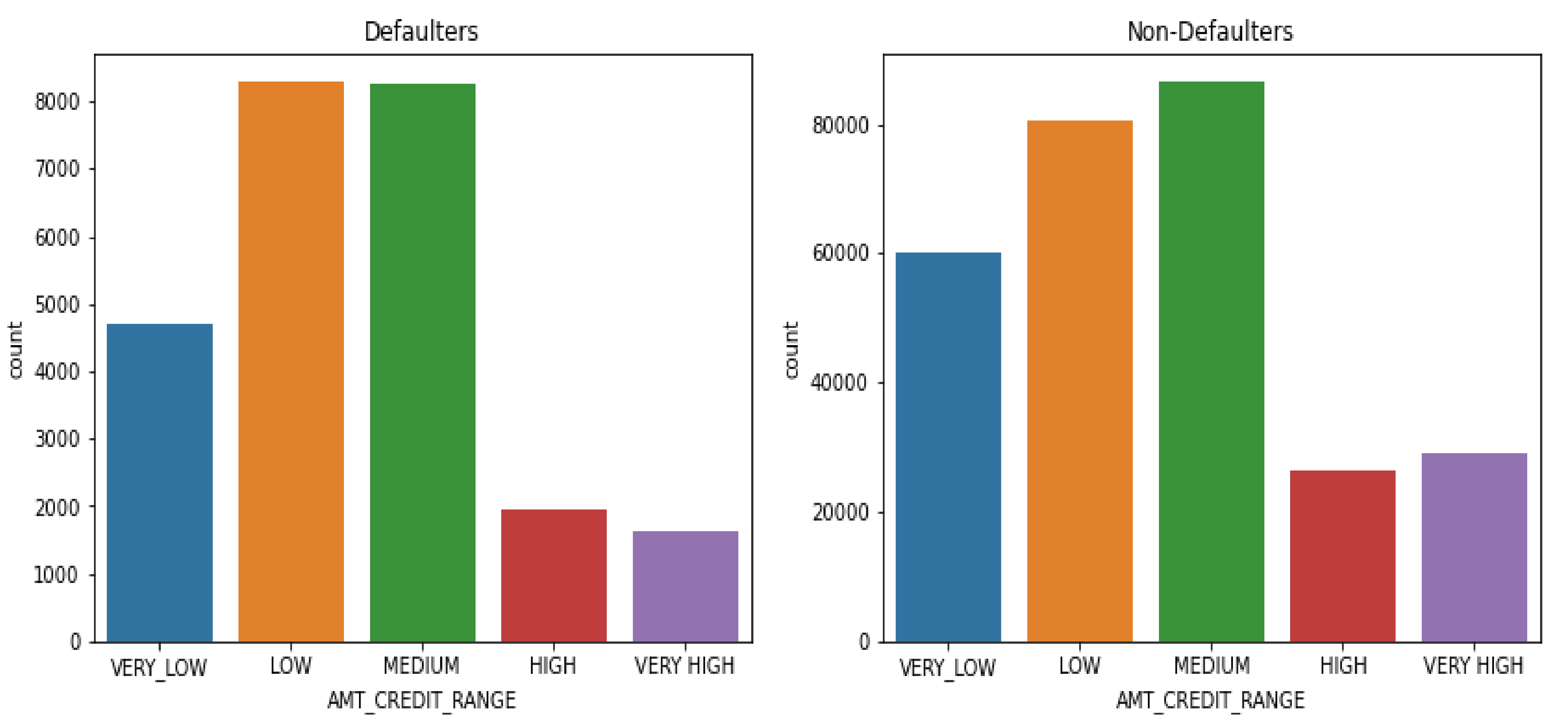
Low income group has more defaulter followed by high income group.

AGE RANGE



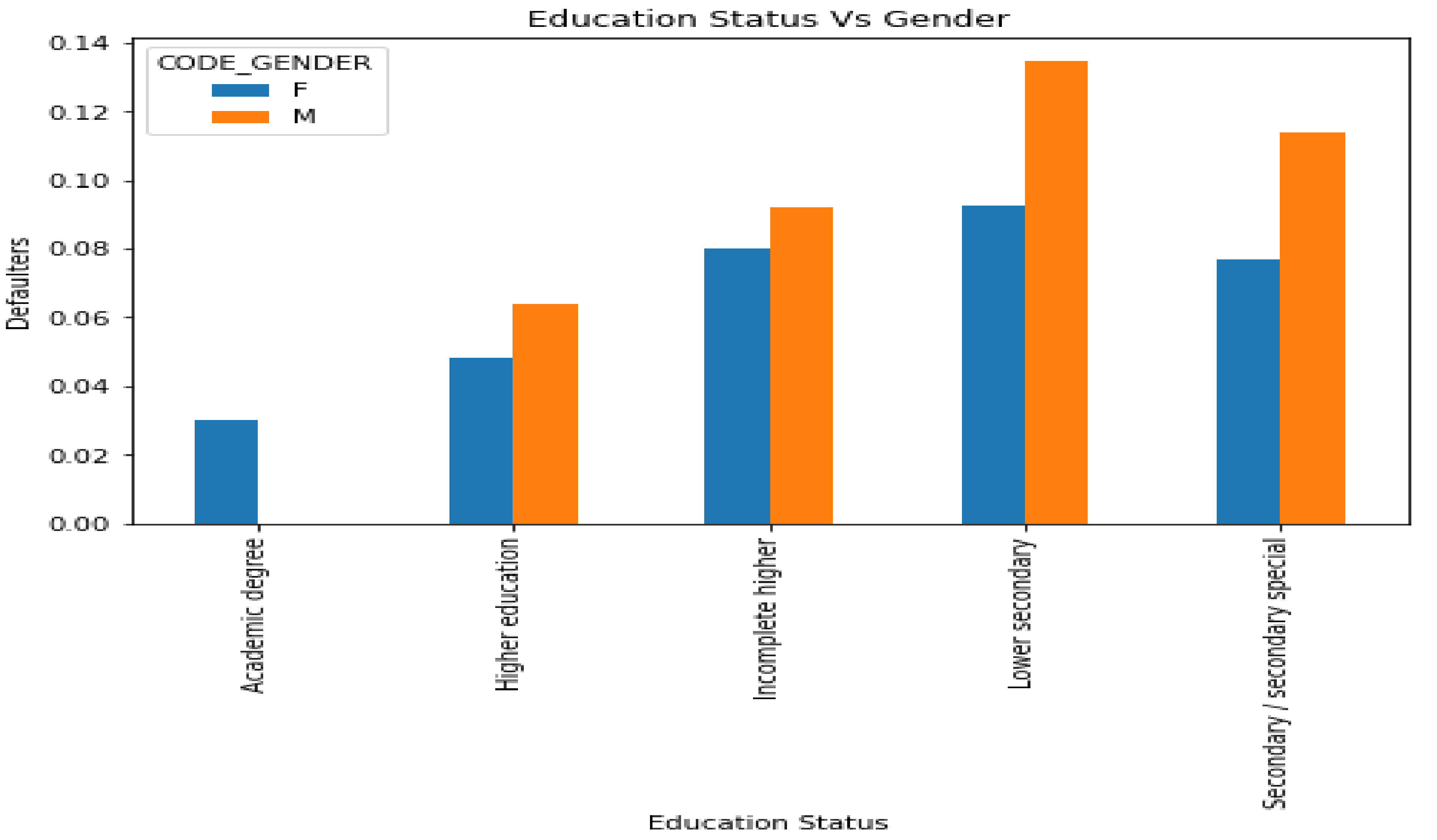
Mid age (35-55) age group of people are more likely to be defaulted followed by the young people.

AMOUNT CREDIT RANGE



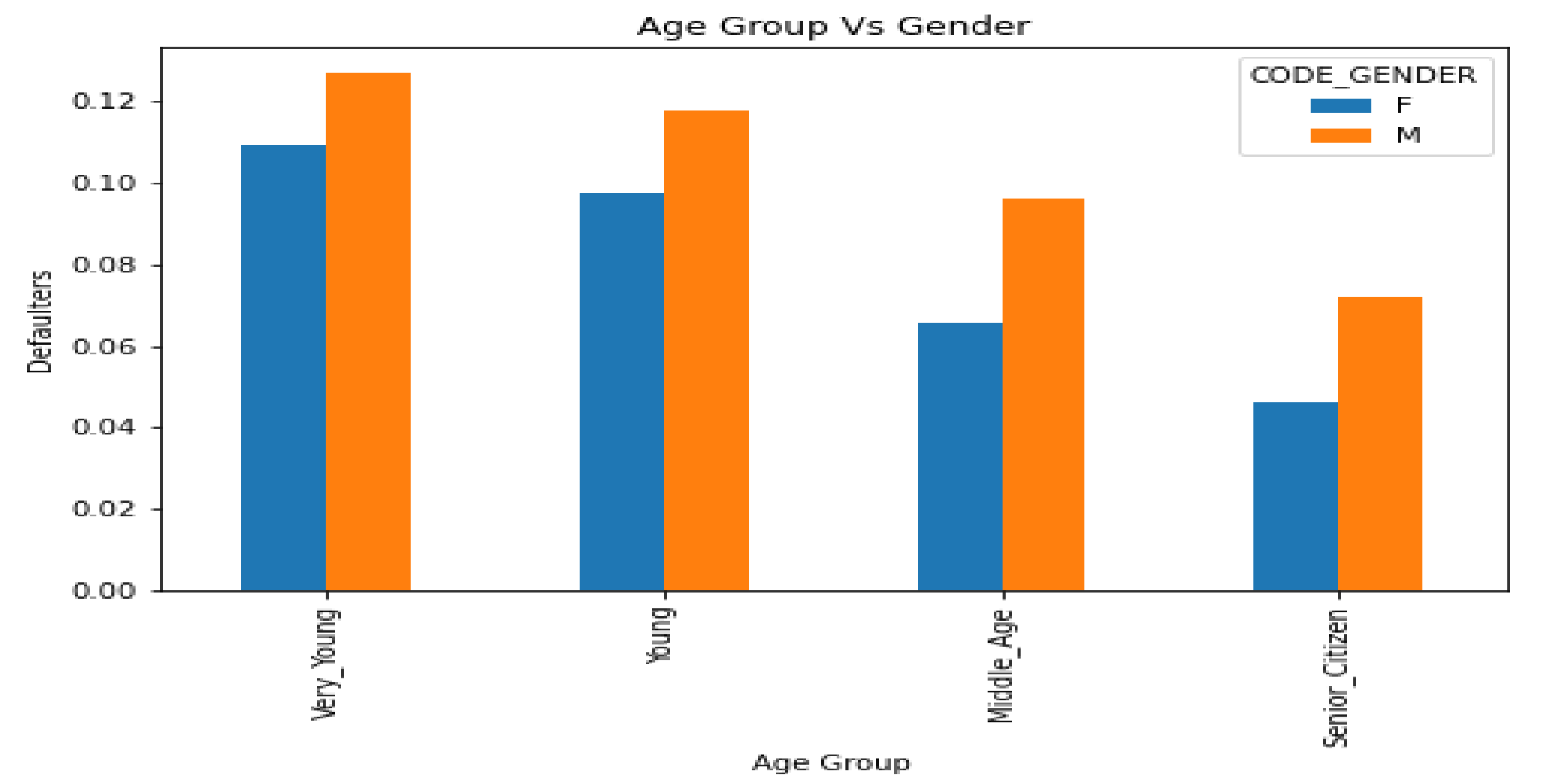
Low category of loan amount credited people are more likely to be defaulted than high amount loan credit.

EDUCATION STATUS



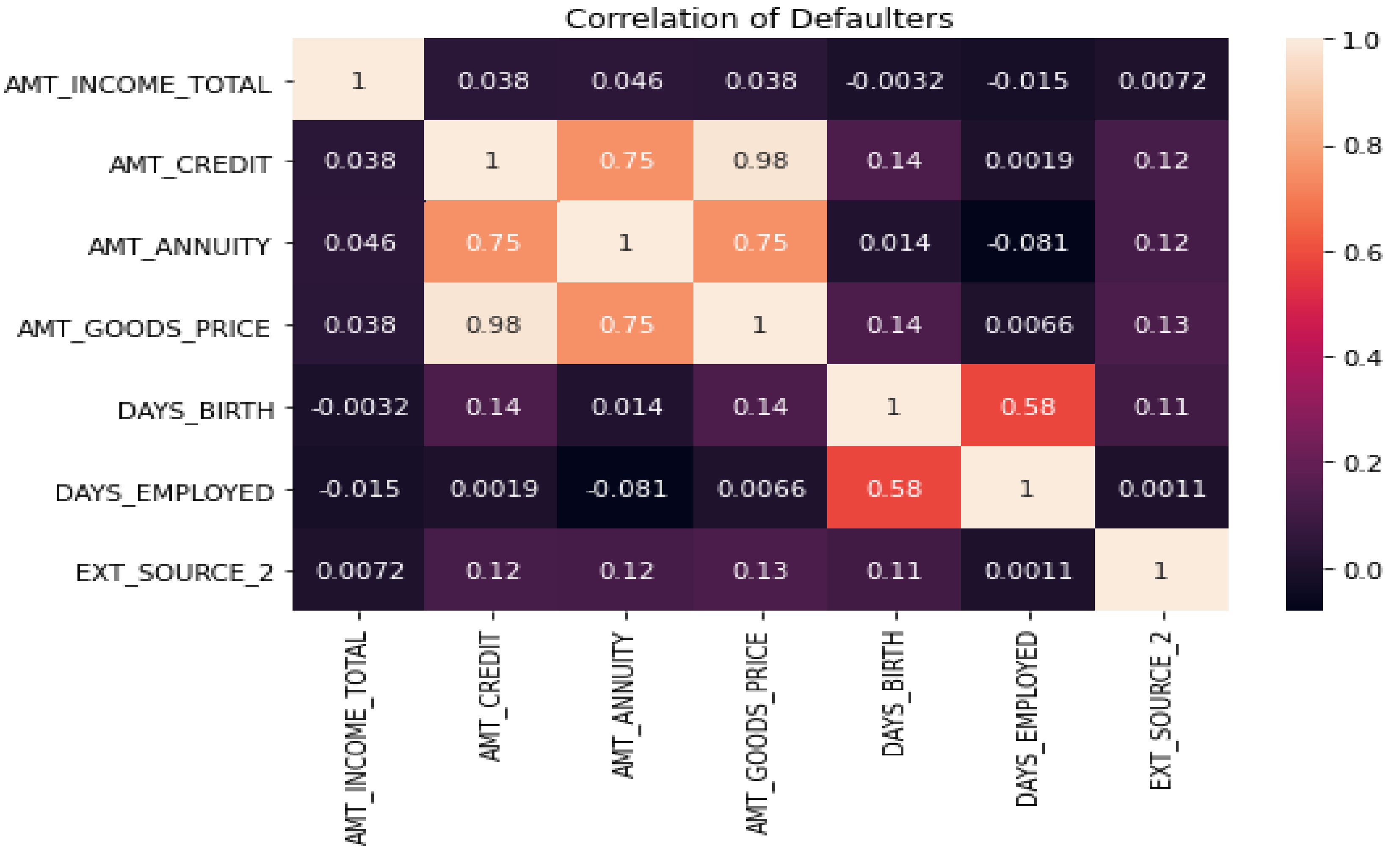
Male with lower secondary education are more defaulted followed by Secondary/secondary special education.

AGE GROUP VS. GENDER



Young male clients are more in number to be defaulted.

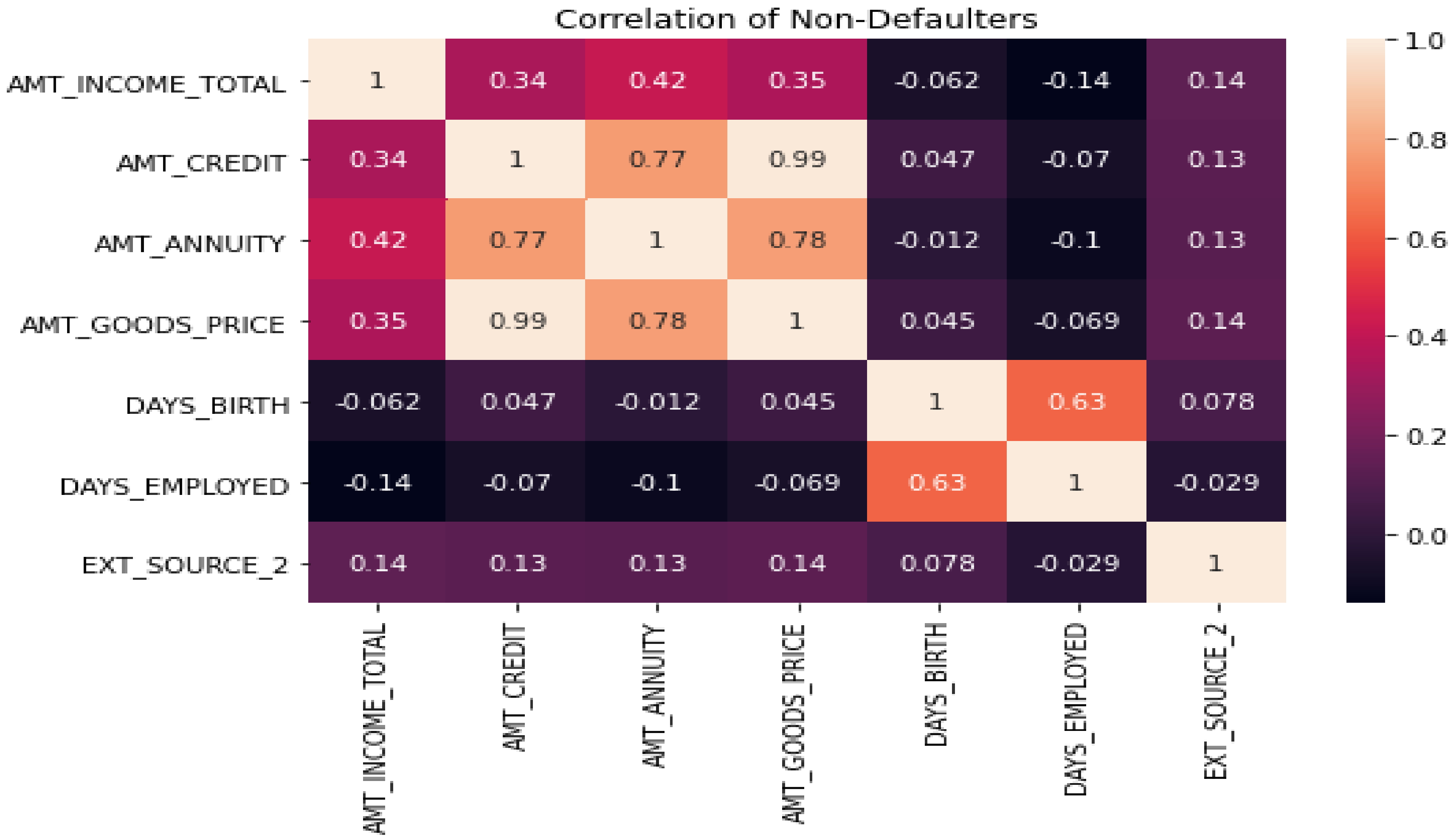
CORRELATION OF DEFAULTERS



	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	EXT_SOURCE_2
AMT_INCOME_TOTAL	1	0.038	0.046	0.038	-0.0032	-0.015	0.0072
AMT_CREDIT	0.038	1	0.75	0.98	0.14	0.0019	0.12
AMT_ANNUITY	0.046	0.75	1	0.75	0.014	-0.081	0.12
AMT_GOODS_PRICE	0.038	0.98	0.75	1	0.14	0.0066	0.13
DAYS_BIRTH	-0.0032	0.14	0.014	0.14	1	0.58	0.11
DAYS_EMPLOYED	-0.015	0.0019	-0.081	0.0066	0.58	1	0.0011
EXT_SOURCE_2	0.0072	0.12	0.12	0.13	0.11	0.0011	1

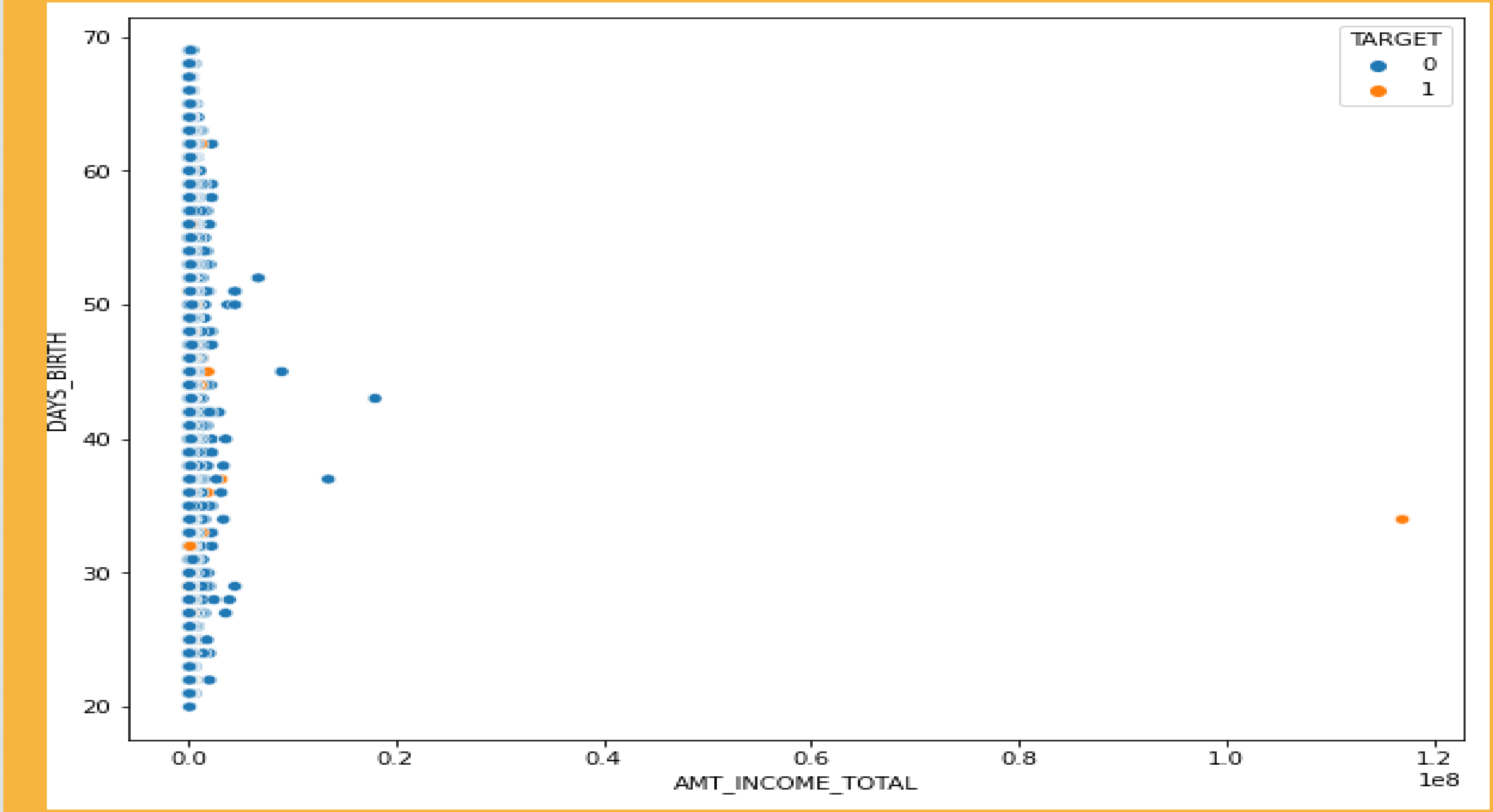
We can see that GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are highly correlated. External Rating is highly correlated with all DAYS_BIRTH(Age), GOODS_PRICE, AMT_CREDIT.

CORRELATION OF NON-DEFAULTERS



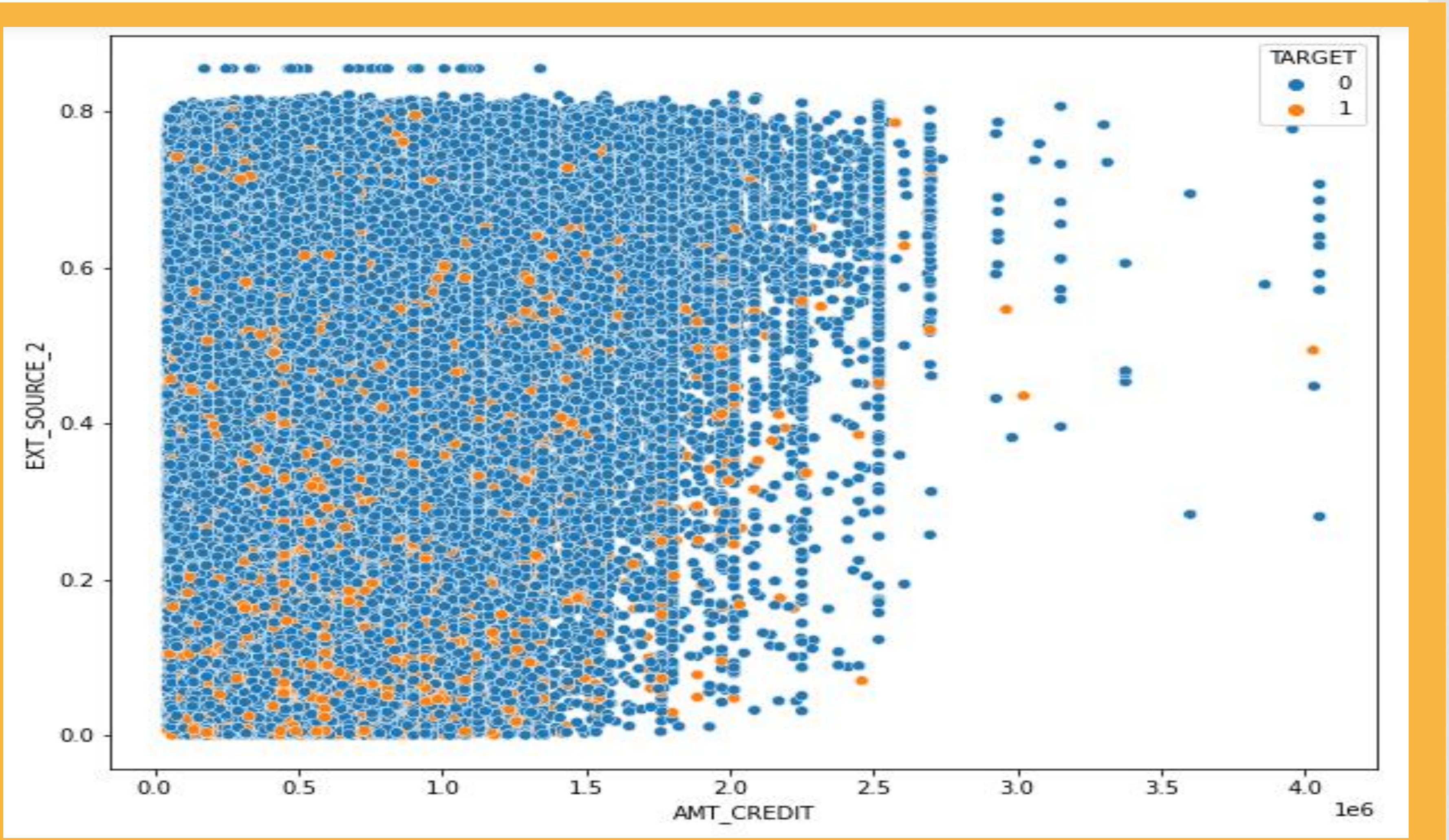
We can see that GOODS_PRICE and AMT_CREDIT, AMT_ANNUTY and AMT_AMT_CREDIT are moderately correlated with each other. External Rating is highly correlated with all DAYS_BIRTH(Age), GOODS_PRICE, AMT_CREDIT.

AGE AND INCOME



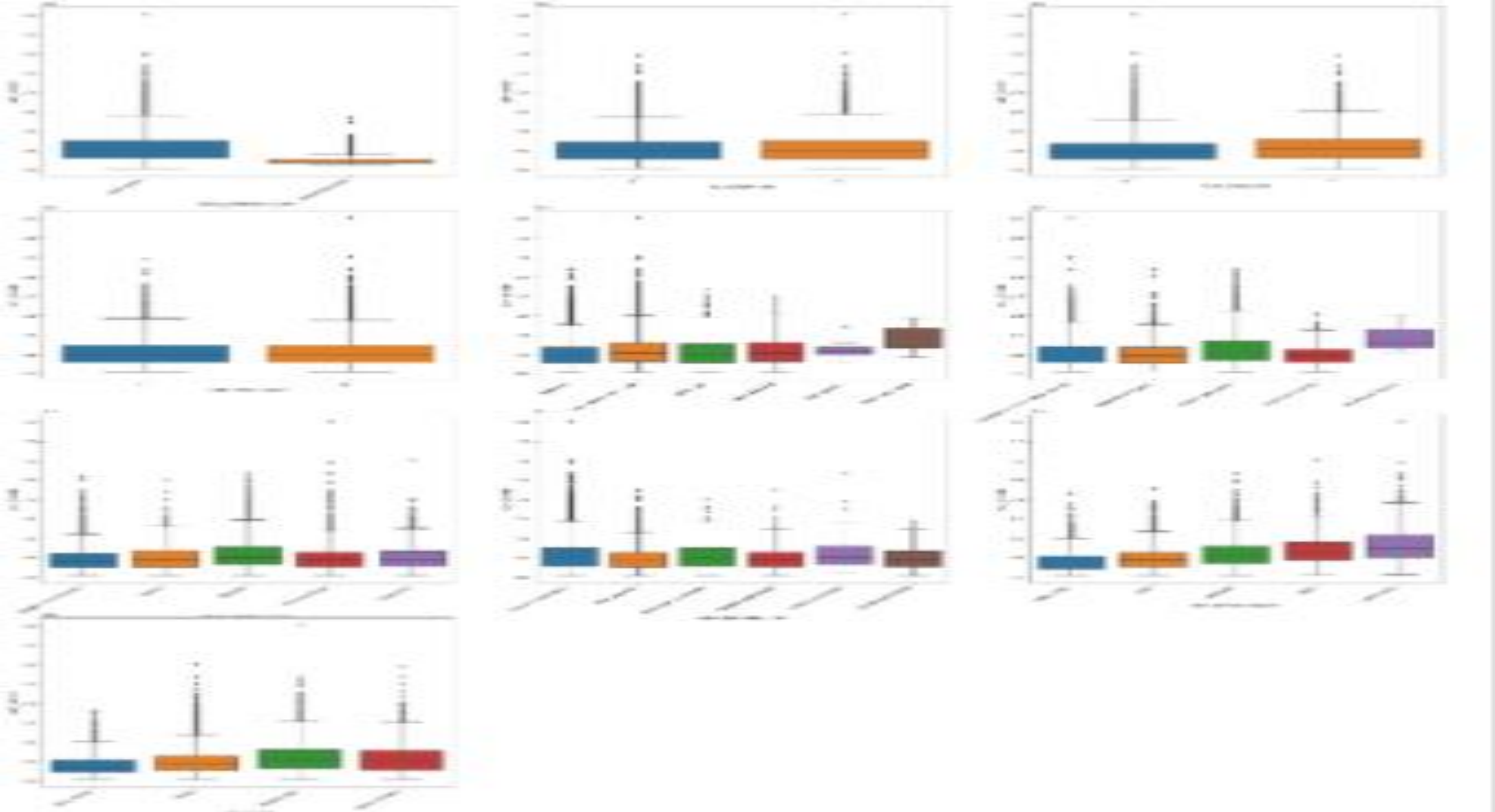
From the above figure, we see that the number of default applications are concentrated more when the days of birth i.e. age is lower, irrespective of the income

LOAN CREDIT AMOUNT AND RATING



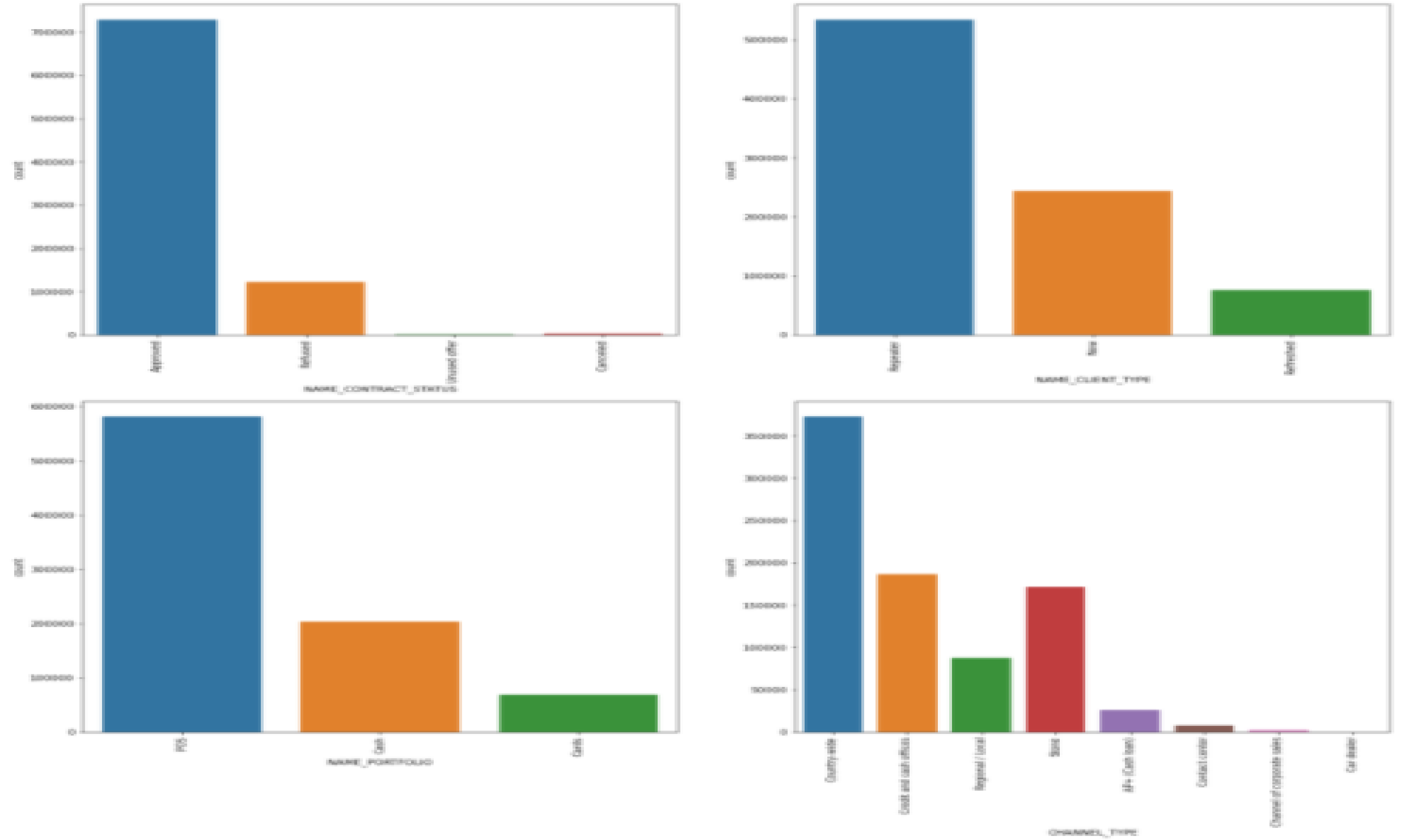
From the above plot, we cannot get much insight as the data is scattered across the plot. However, we can see some concentration of defaulters near the low rating region between 0.0 to 3.0.

CATEGORICAL COLUMNS



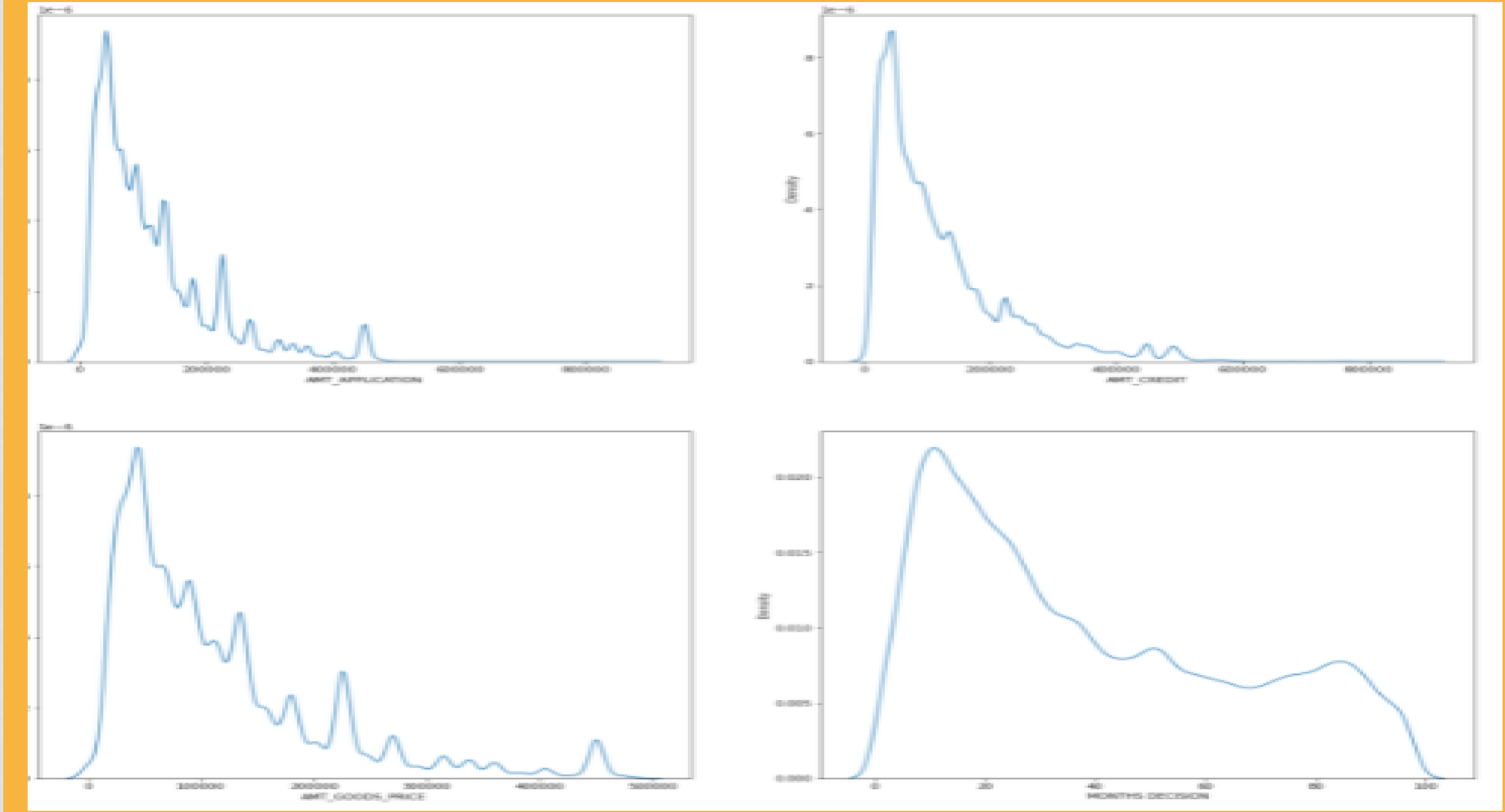
1. Cash loans are more credited.
2. Those who are female and own car they got little more number of loans
3. State servant got more number of loans
4. Higher education got more loans
5. Married people got more number of loans
6. Clients who are living in municipal apartment, got more number of loans
7. High income group people got more loans.
8. 8Mig age people got more number of loans.

CATEGORICAL COLUMNS



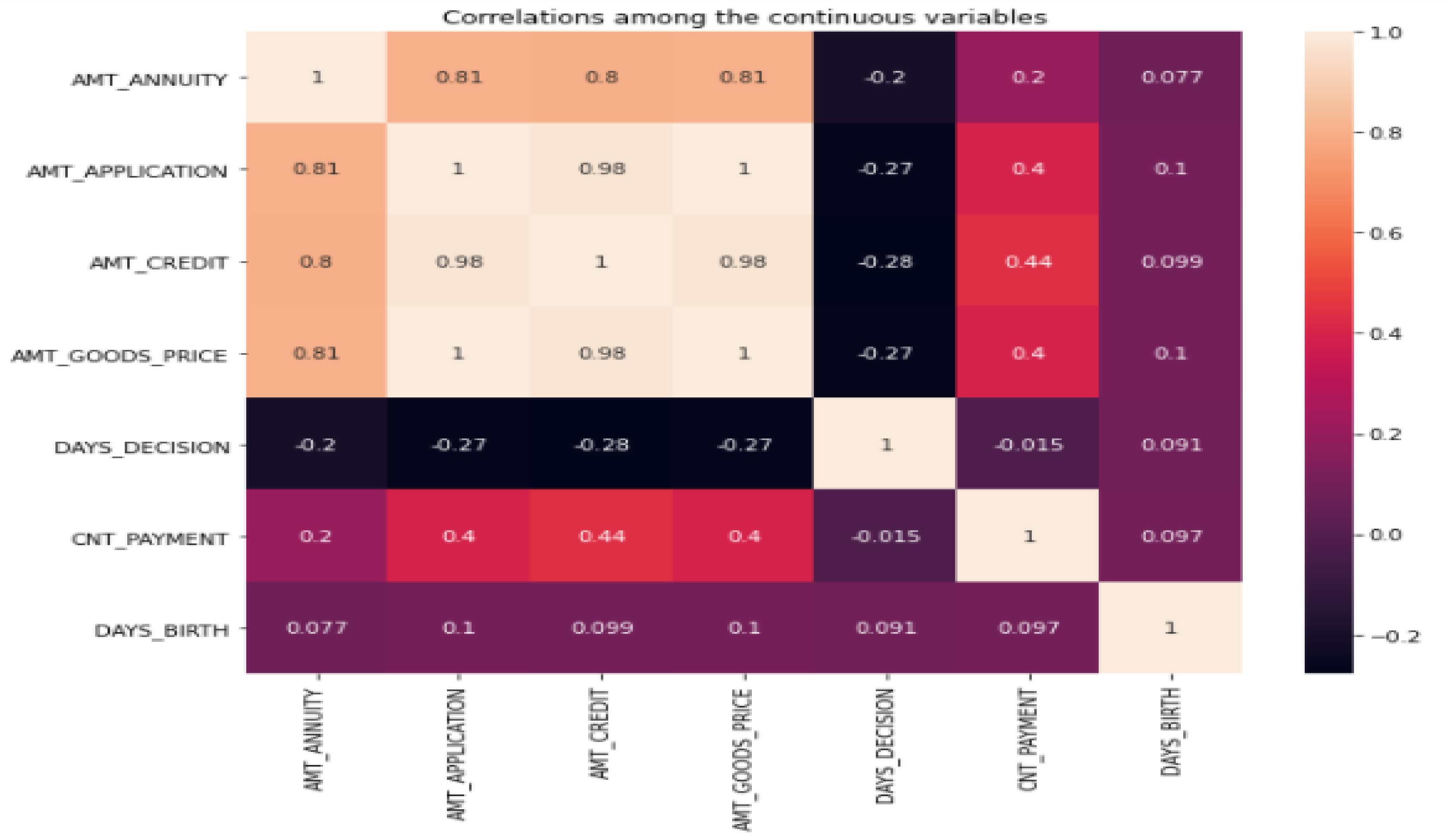
1. Approved loan status is huge than rejected or canceled.
2. Repeater clients are highest in number than new client.
3. POS loans are highest rather than cash loans.
4. Country-wide channel type is the most used channel followed by Credit and cash offers.

CONTINUOUS COLUMNS



1. Most of the loan application amount were below 500000, we can see a huge spike around 100000 amount.
2. Amount credited, is also following the pattern of loan application. We already saw that most of the application was approved in previous plots.
3. Amount of the goods price is also following the same distribution like application amount and amount credited. Because, based on the price of the goods, the loan was approved and amount was credited.
4. Most of the applications decision took around 10 to 30 months.

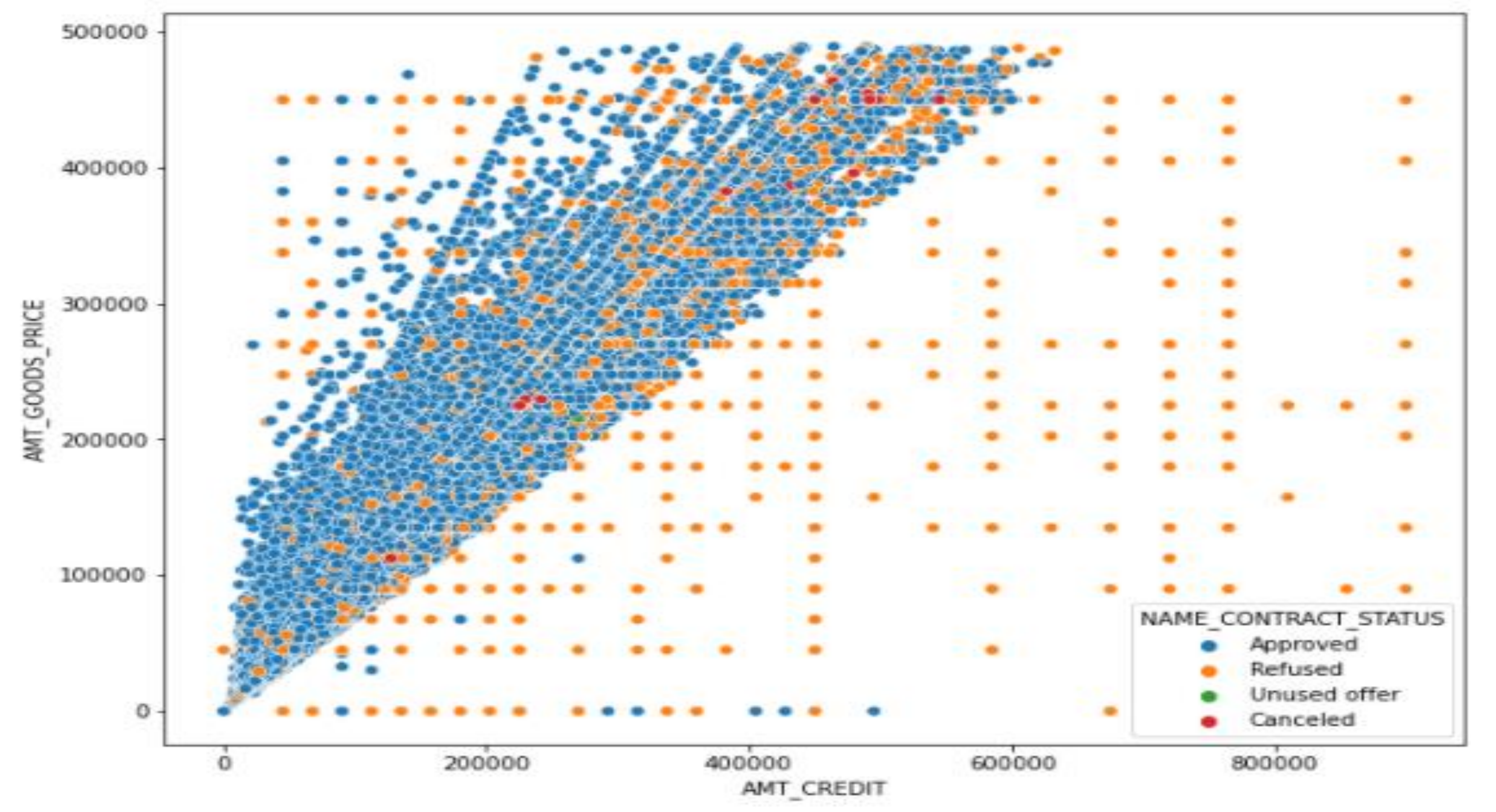
CORRELATIONS-CONTINUOUS VARIABLES



	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_GOODS_PRICE	DAYS_DECISION	CNT_PAYMENT	DAYS_BIRTH
AMT_ANNUITY	1	0.81	0.8	0.81	-0.2	0.2	0.077
AMT_APPLICATION	0.81	1	0.98	1	-0.27	0.4	0.1
AMT_CREDIT	0.8	0.98	1	0.98	-0.28	0.44	0.099
AMT_GOODS_PRICE	0.81	1	0.98	1	-0.27	0.4	0.1
DAYS_DECISION	-0.2	-0.27	-0.28	-0.27	1	-0.015	0.091
CNT_PAYMENT	0.2	0.4	0.44	0.4	-0.015	1	0.097
DAYS_BIRTH	0.077	0.1	0.099	0.1	0.091	0.097	1

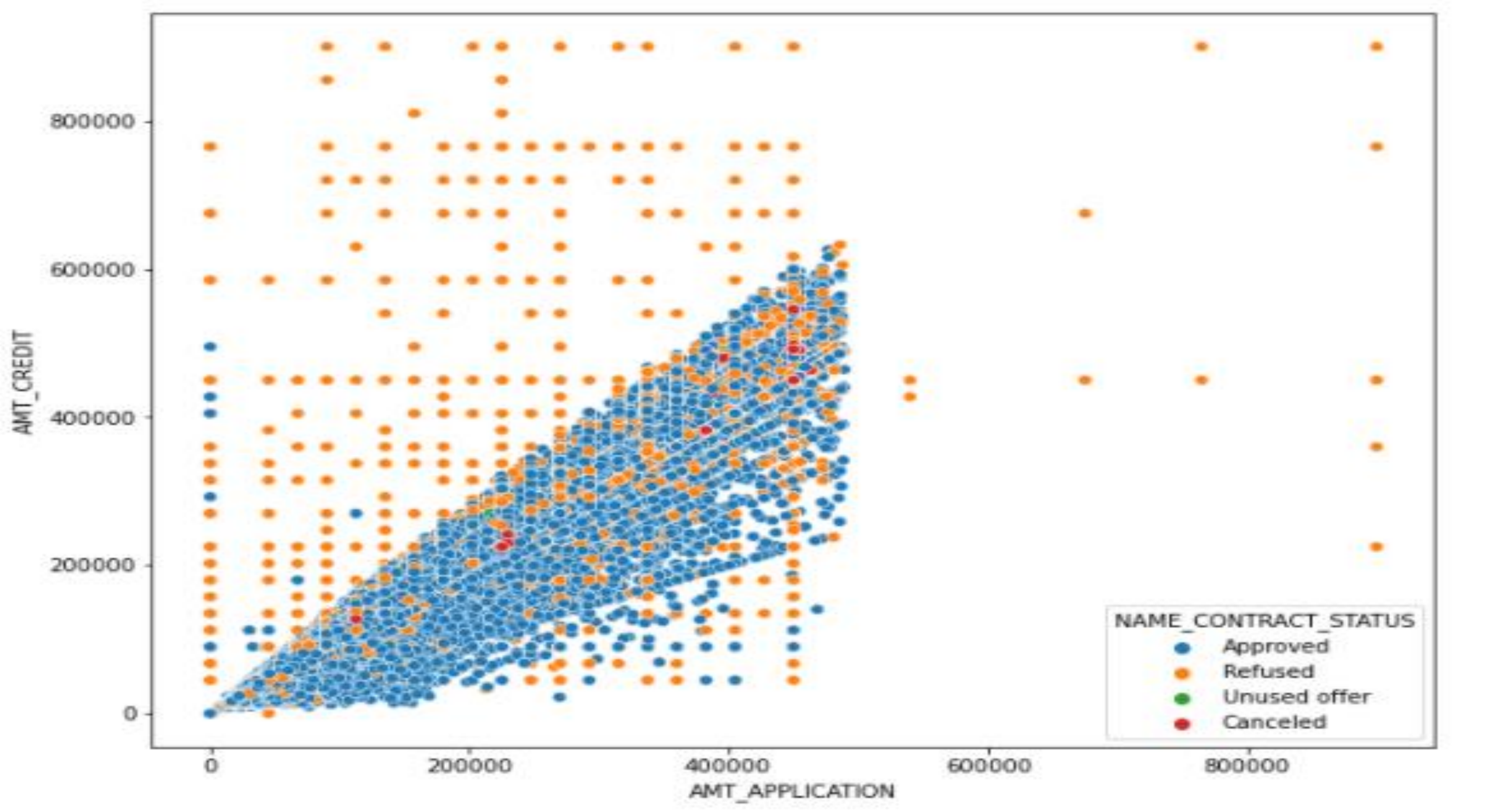
There are strong correlations between below variables DAYS_BIRTH(AGE) is correlated with all the variables AMT_APPLICATION is correlated with AMT_ANNUTY, AMT_AMT_CREDIT, AMT_GOODS_PRICE

CONTINUOUS COLUMNS



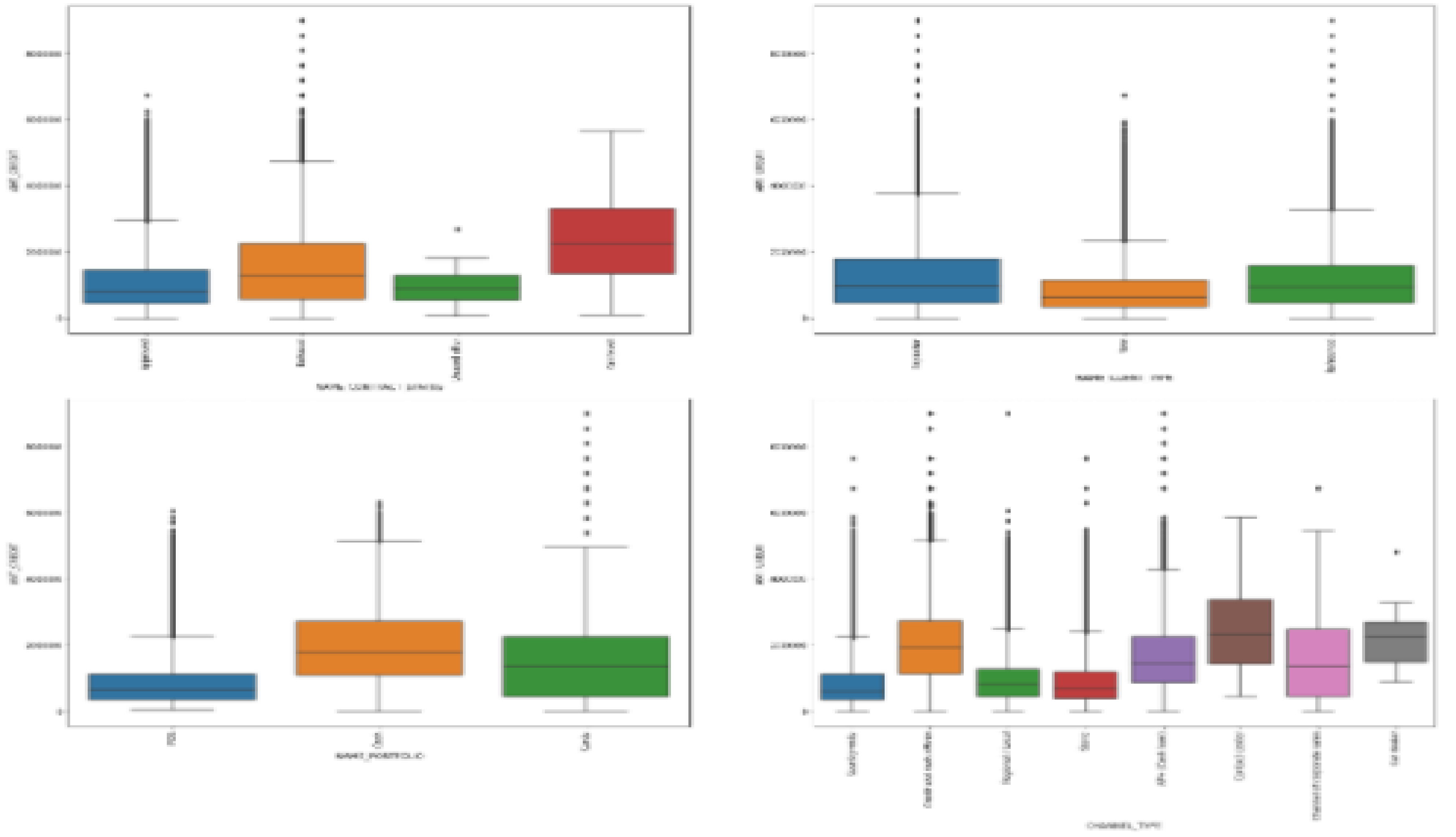
AMT_GOODS_PRICE and AMT_CREDIT are posotively correlated and mostly concentrated near the lower region. High AMT_CREDIT loans are most likely to be refused.

CONTINUOUS COLUMNS



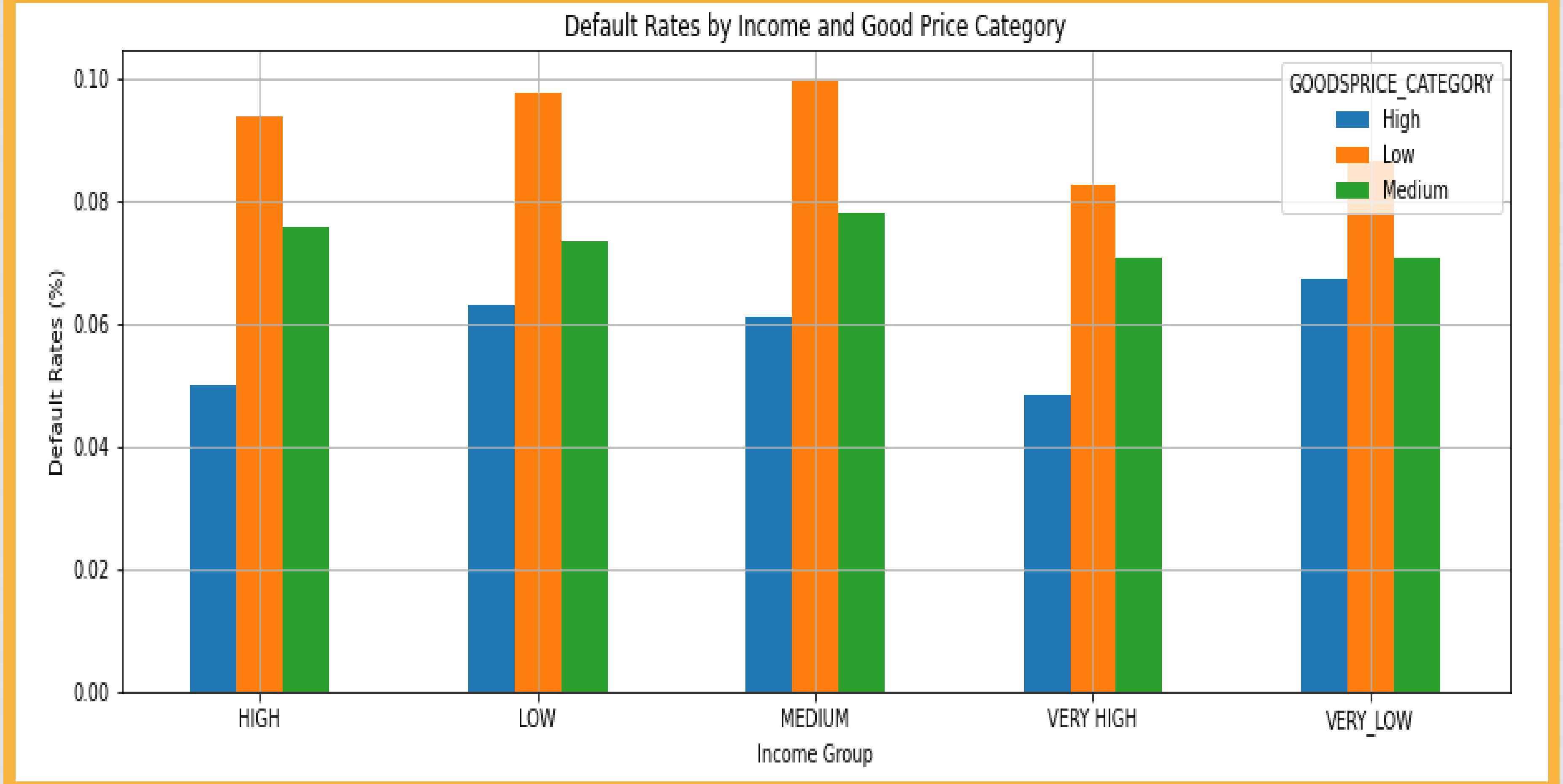
Credit amount and the application is highly correlated.

ORGANIZATION TYPE



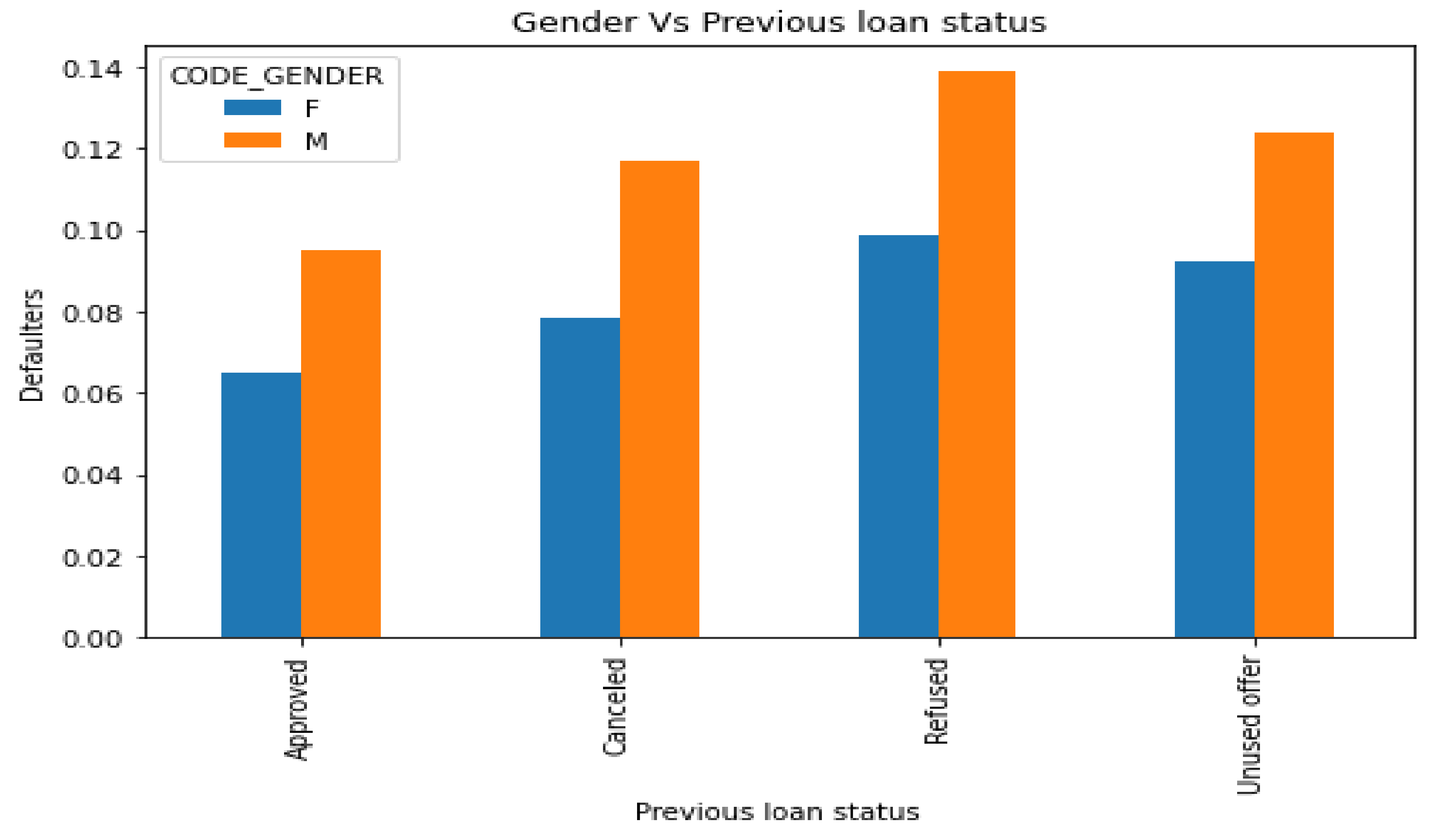
1. Most of the amount credit was cancelled in status 2. Repeater client got more loan credit 3. Cash loan got more credited. 4. Through the contact center channel, more loan got credited.

DEFAULT RATES-INCOME AND PRICE



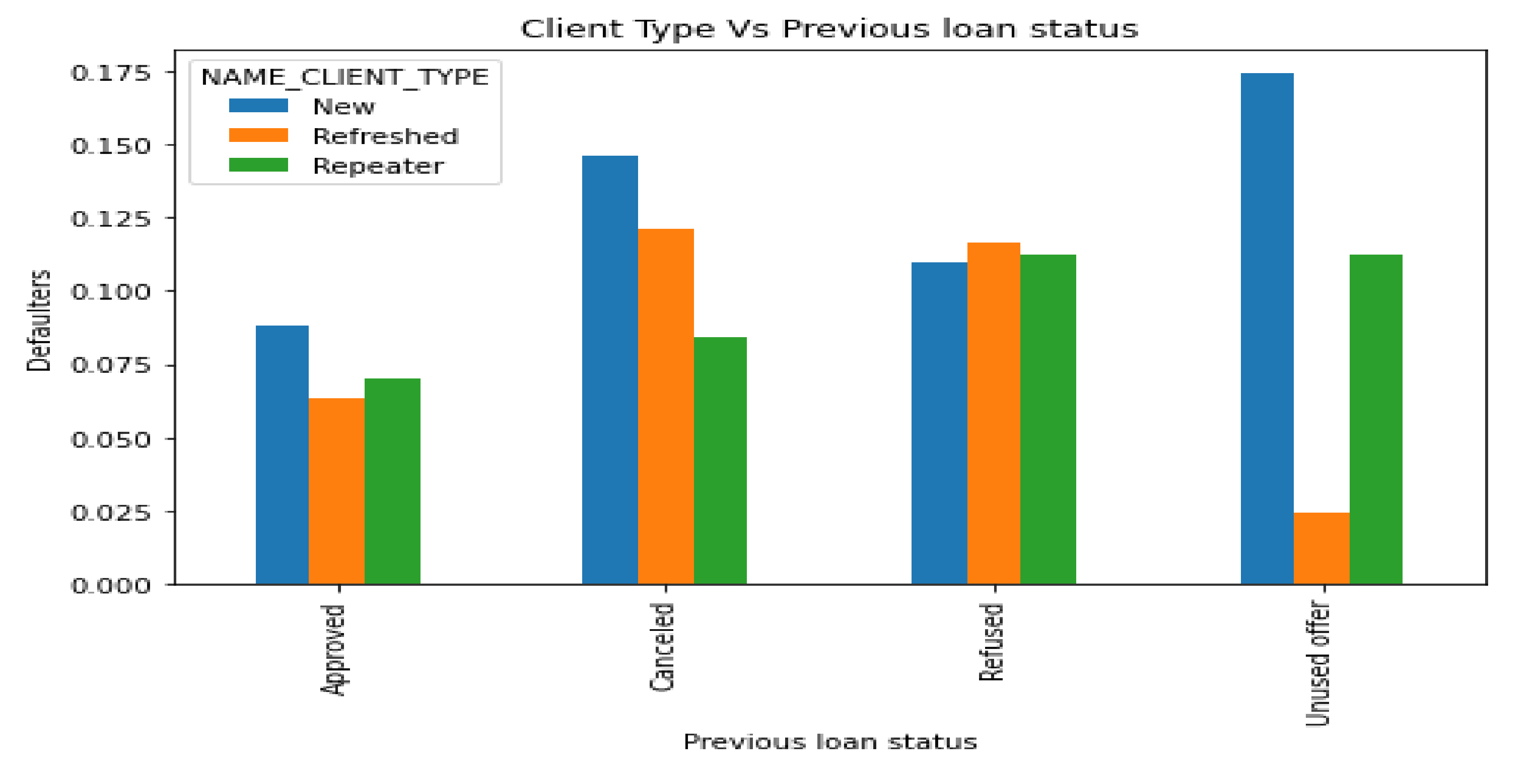
From the above analysis, we find that irrespective of the income groups, the lowest price of the good has the highest chances of default.Interestingly, the highest price category of goods has the lowest probability of default for all the income groups.

GENDER VS. PREVIOUS LOAN STATUS



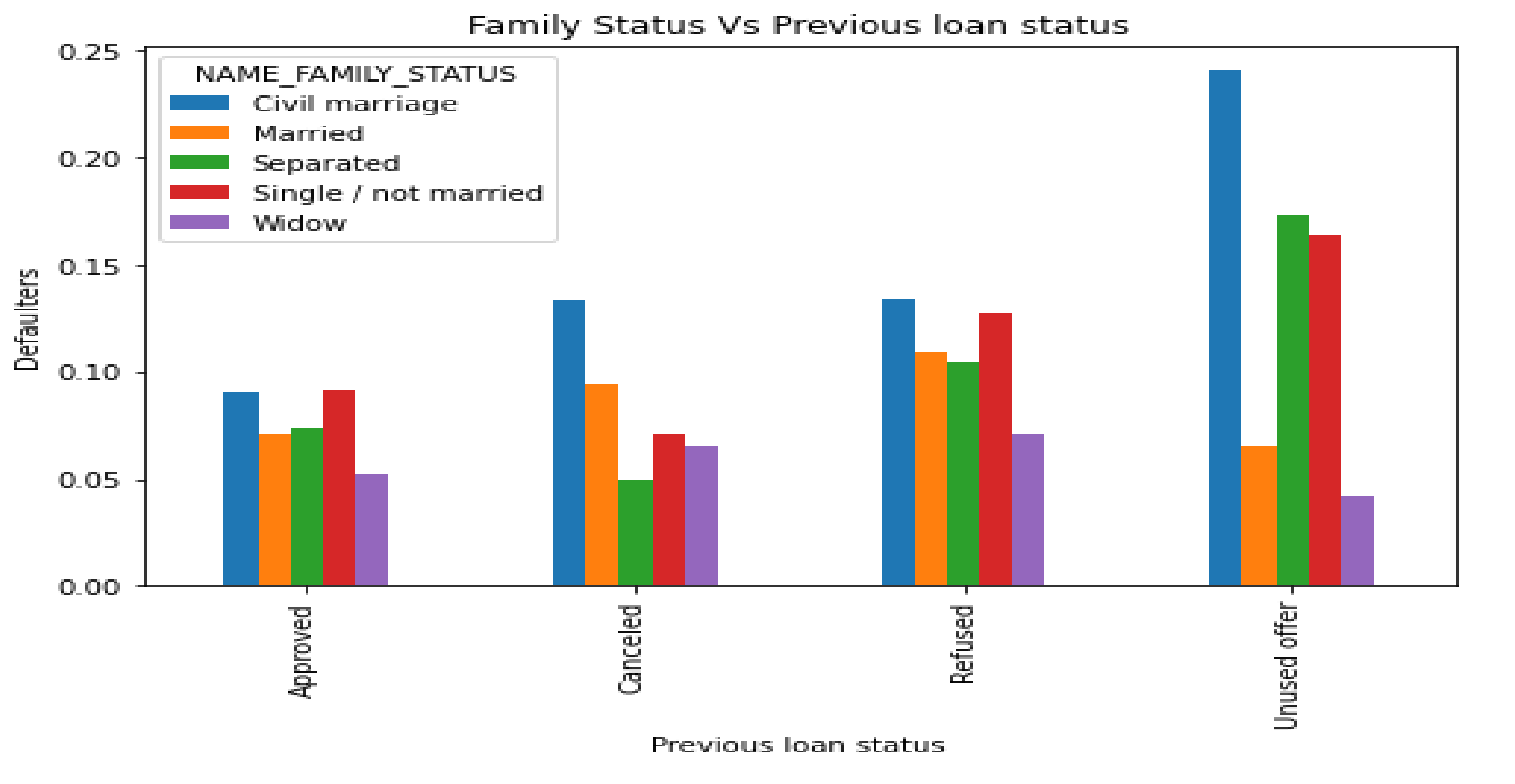
Male clients are more defaulted than female client. Also, previously refused customer are more defaulted in current application.

CLIENT TYPE VS. PREVIOUS LOAN STATUS



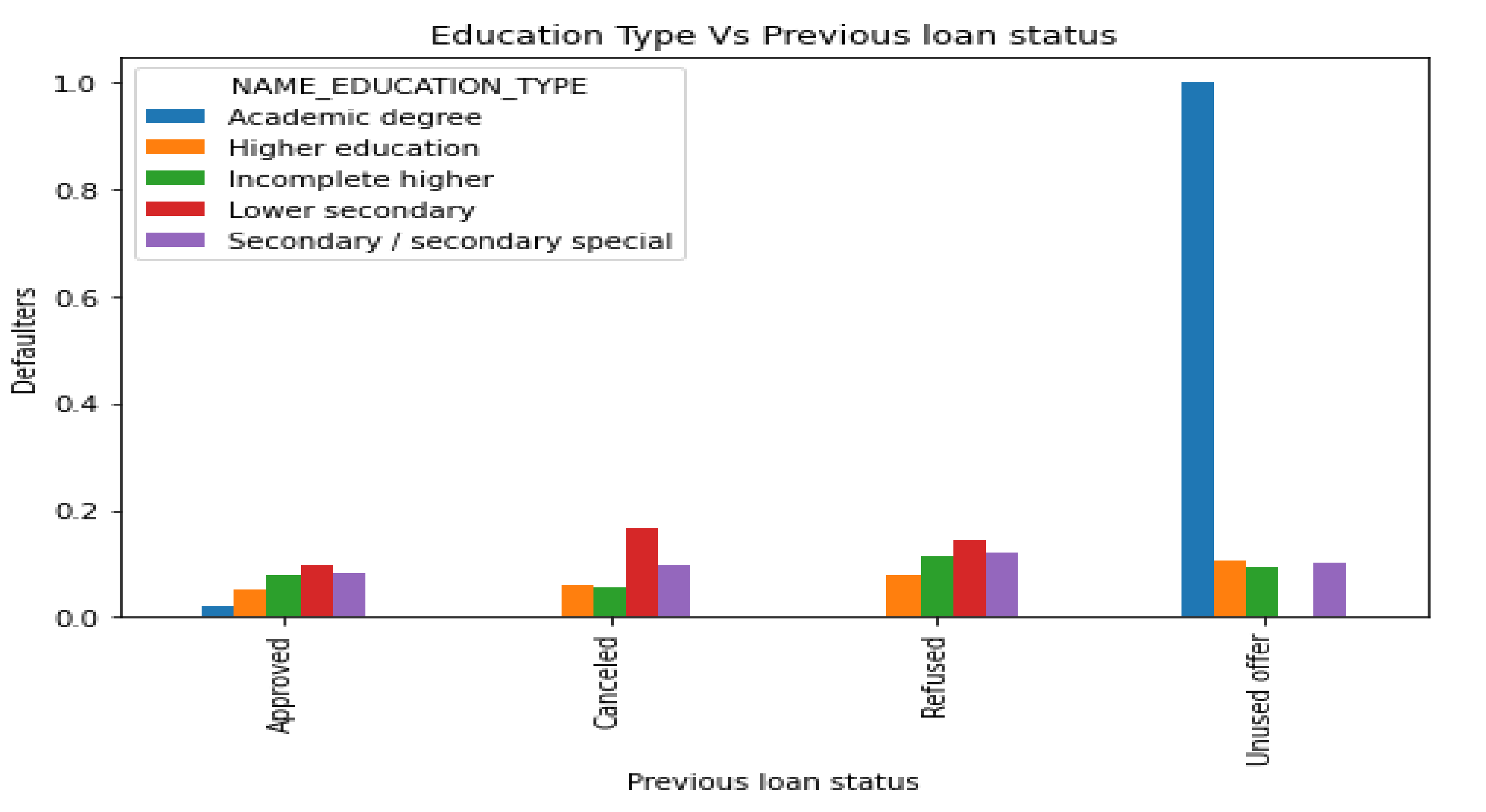
Previously cancelled New and Refreshed clients are more defaulted than repeater clients

FAMILY STATUS VS. PREVIOUS LOAN STATUS



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

ORGANIZATION TYPE



Previously refused people with lower secondary education are more defaulted in current application.

SUMMARY-I

After analyzing the datasets, there are few attributes of a client with which the bank would be able to identify if they will repay the loan or not. The analysis is consisted as below with the contributing factors and categorization:

Factors whether an applicant will be Repayer:

1. Academic degree has less defaults.
2. Student and Businessmen have no defaults.
3. RATING 1 is safer.
4. Clients with Trade Type(ORGANIZATION_TYPE) 4 and 5 and Industry type 8 have defaulted less than 3%
5. People above age of 50 have low probability of defaulting
6. Clients with 40+ year experience having less than 1% default rate
7. Applicant with Income more than 700,000 are less likely to default
8. Loans bought for Hobby, Buying garage are being repayed mostly.
9. People with zero to two children tend to repay the loans.

SUMMARY-II

Factors whether an applicant will be Defaulter:

1. Men are at relatively higher default rate
2. People who have civil marriage or who are single default a lot.
3. People with Lower Secondary & Secondary education
4. Clients who are either at Maternity leave OR Unemployed default a lot.
5. People who live in Rating 3 has highest defaults.
6. Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
7. Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.
8. Avoid young people who are in age group of 20-40 as they have higher probability of defaulting.
9. People who have less than 5 years of employment have high default rate.
10. Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected.
11. When the credit amount goes beyond 3M, there is an increase in defaulters.

THANK YOU !!!