Credit Analysis EDA

Yuthika Satheesh

Problem Statement

Two types of risks are associated with the bank’s decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
- 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.

Hence, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Analysis for missing values

- We find that there are around 40 columns with more than 50% missing values. Since most of the data is not available in these columns its better not to consider them for analysis.

Imputation for missing values less than 13%

AMT_ANNUITY: Since we have very low count of missing values in this column, we can impute them by mean.

NAME_TYPE_SUIT: We can impute the missing values in this column by mode i.e., Unaccompanied

CNT_FAM_MEMBERS: We can impute the missing values in this column by 0 since this column has very low count of missing values and 0 looks more natural.

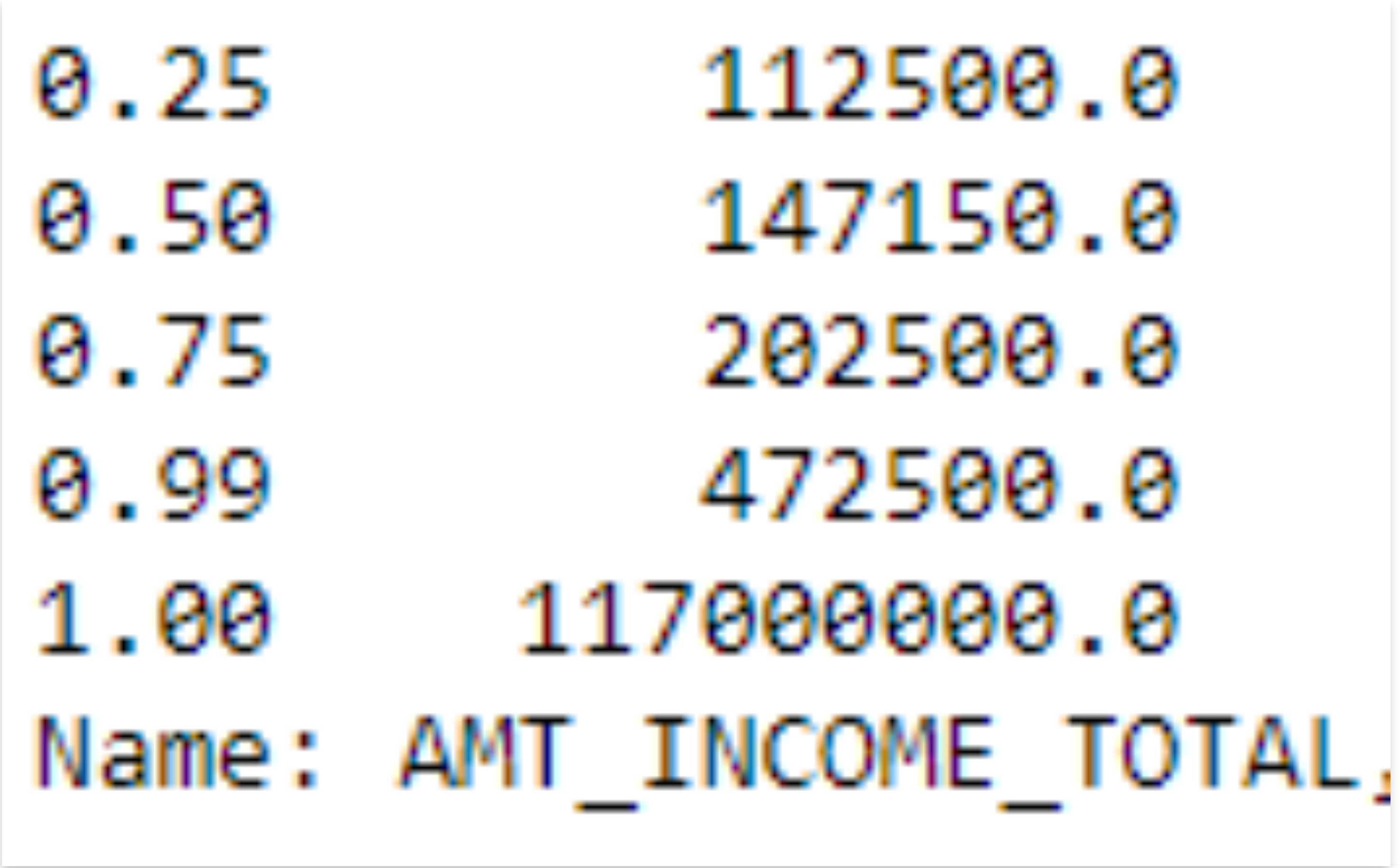
OCCUPATION_TYPE: We can impute the missing values in this column by a new category i.e., Others

DAYS_LAST_PHONE_CHANGE: Only one record had null value, so replaced it with 0.

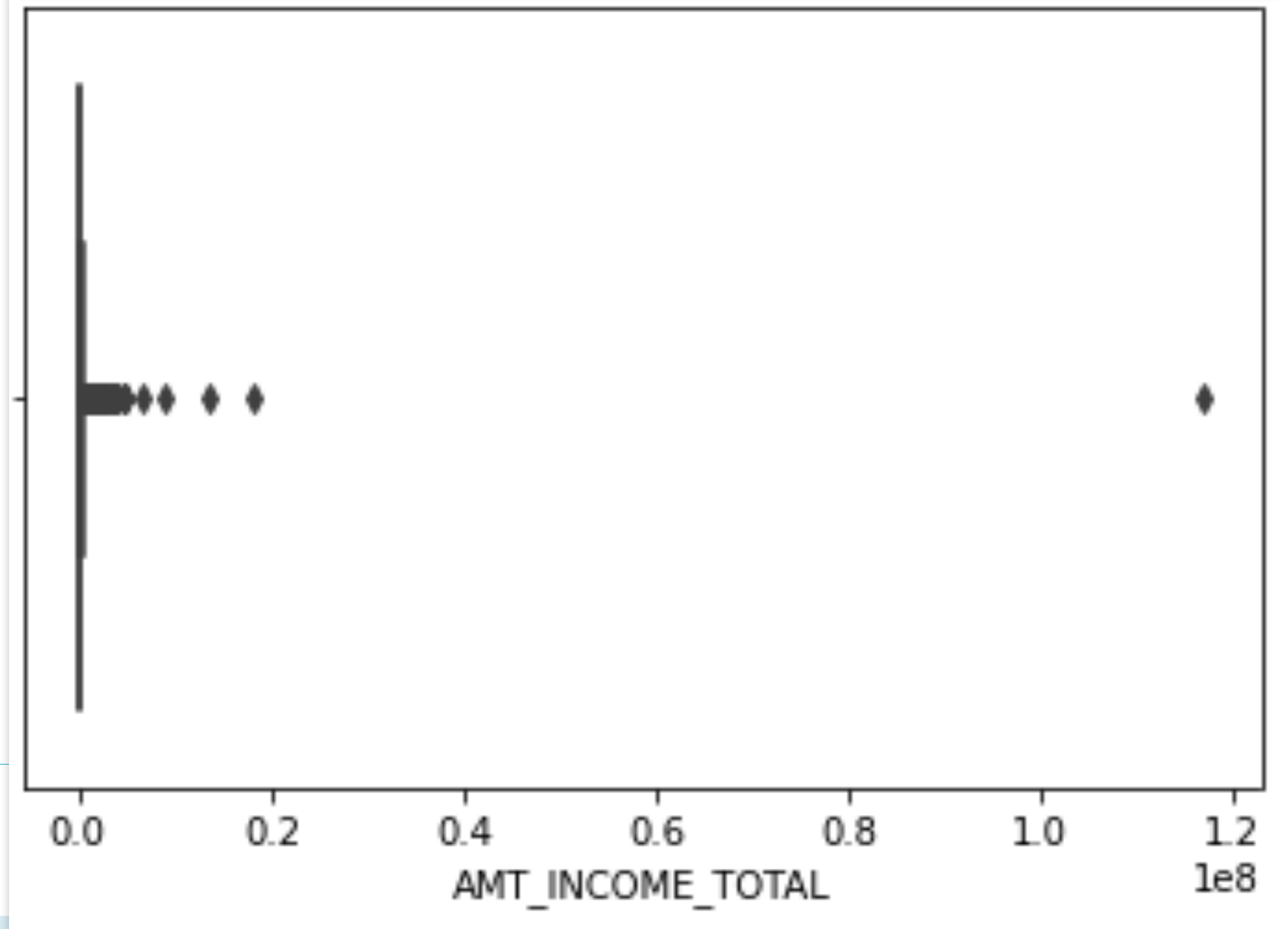
Analysis of Numerical columns for outliers

Annual Income

- From the boxplot and the difference between 0.99 & 1.0 Inter-quartile range we can clearly notice that there are outliers in this column.
- Hence, we decide to remove the outliers i.e., the values greater than 0.99 IQR from the column.

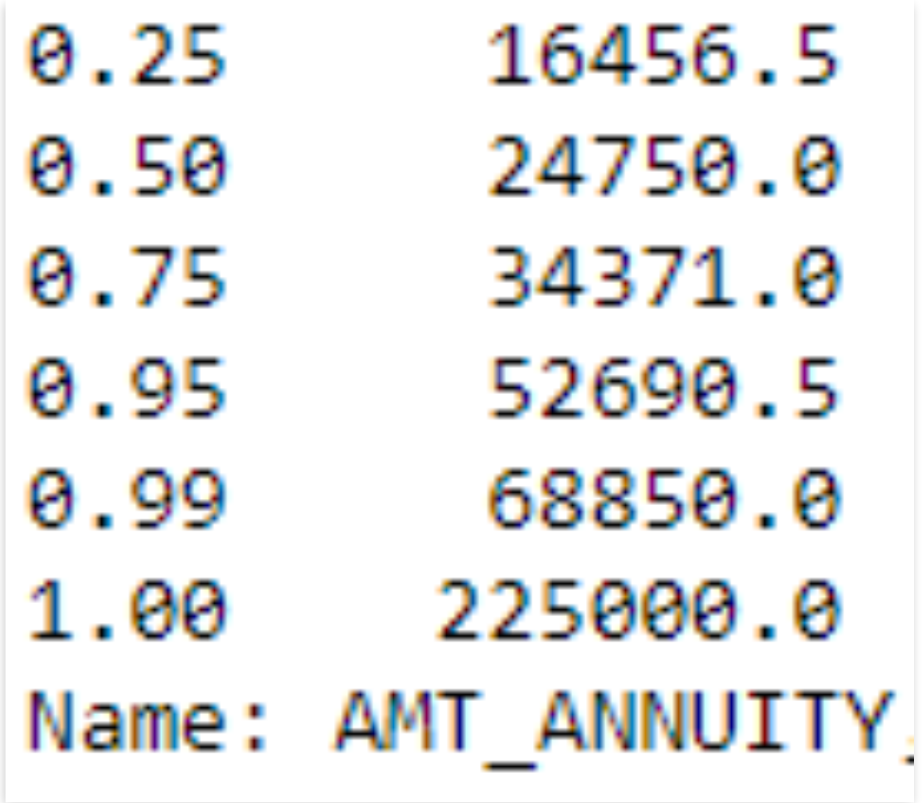


```
0.25       112500.0
0.50       147150.0
0.75       202500.0
0.99       472500.0
1.00    117000000.0
Name: AMT_INCOME_TOTAL,
```

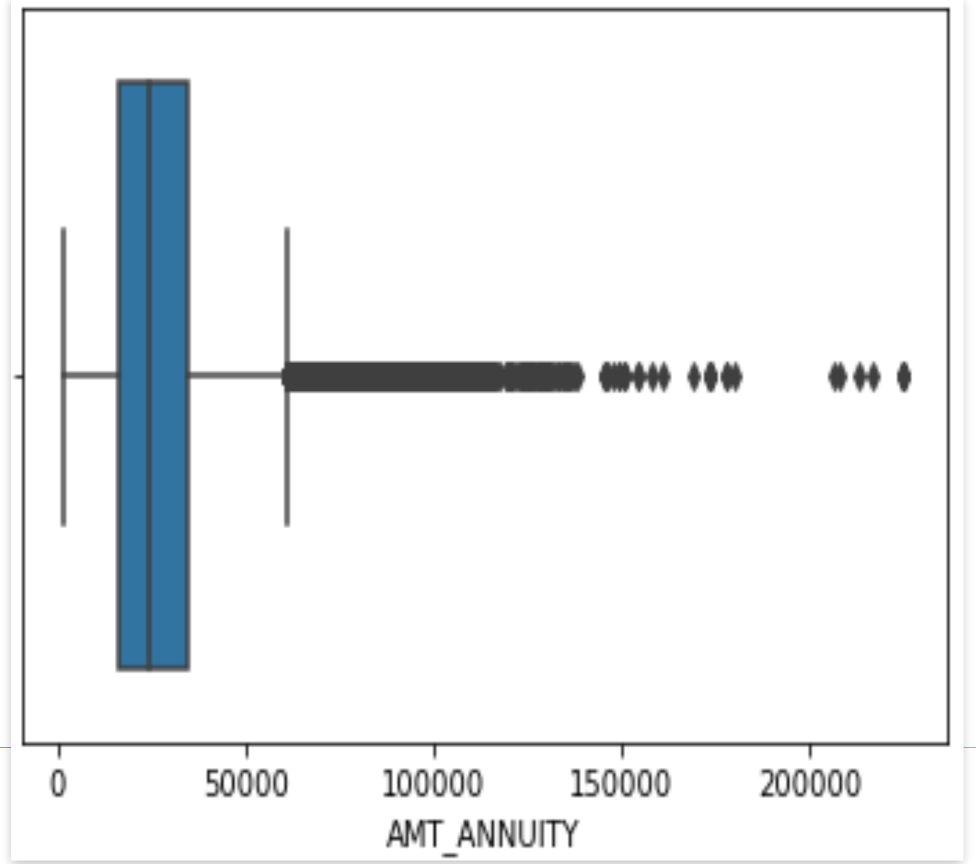


Loan Annuity Amount

- From the boxplot and the difference between 0.99 &1.0 Inter-quartile range we can clearly notice that there are outliers in this column.
- Hence, we decide to remove the outliers i.e., the values greater than 0.99 IQR from the column.

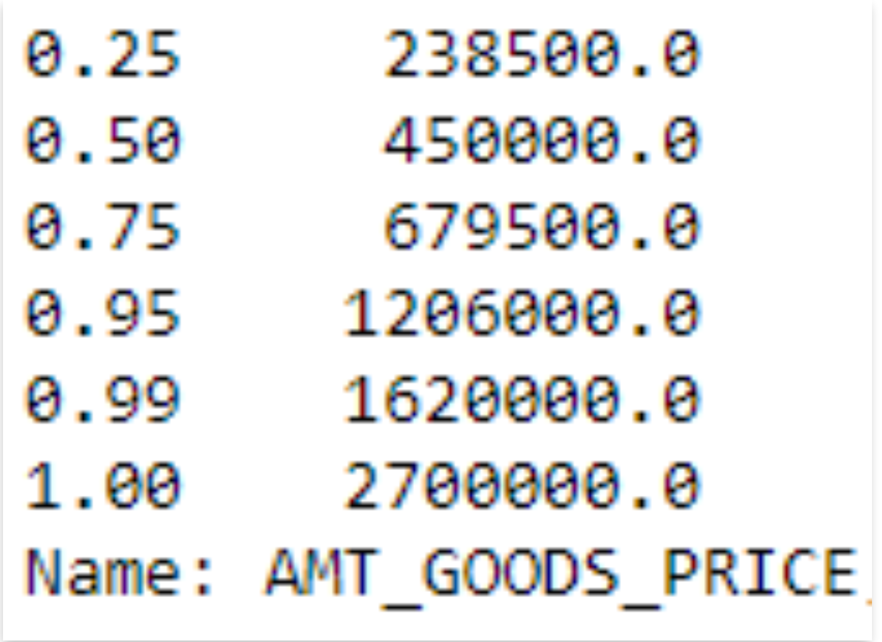


```
0.25      16456.5
0.50      24750.0
0.75      34371.0
0.95      52690.5
0.99      68850.0
1.00     225000.0
Name: AMT_ANNUITY,
```

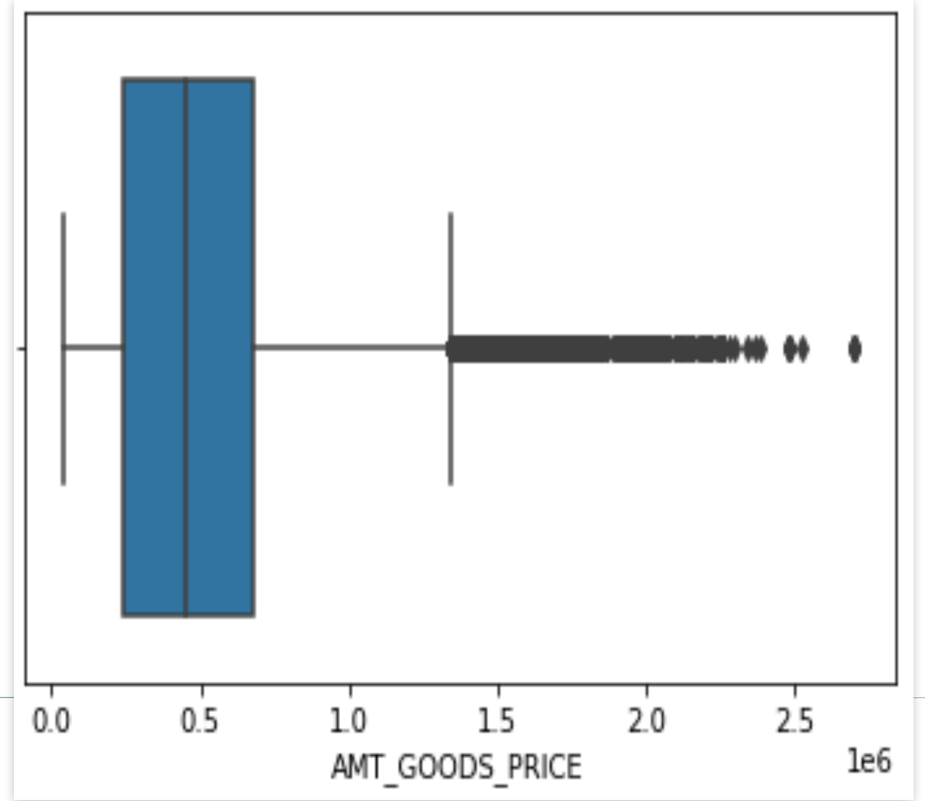


Goods Price

- From the boxplot and the difference between 0.99 &1.0 Inter-quartile range we can clearly notice that there are outliers in this column.
- Hence, we decide to remove the outliers i.e., the values greater than 0.99 IQR from the column.

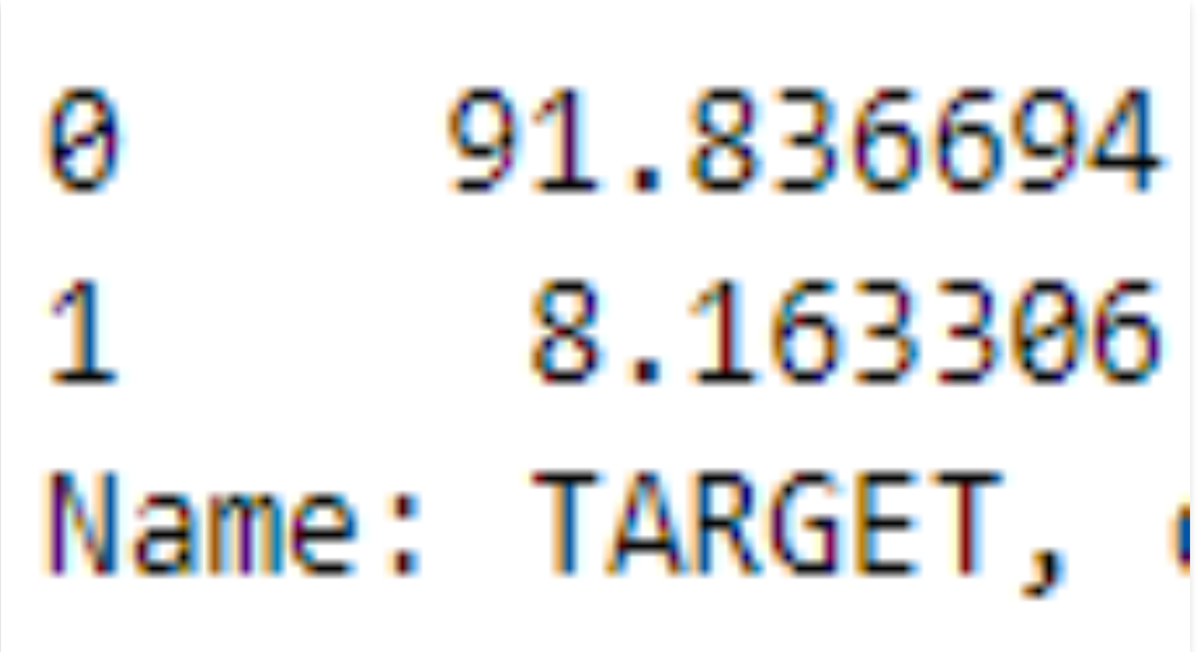


```
0.25      238500.0
0.50      450000.0
0.75      679500.0
0.95     1206000.0
0.99     1620000.0
1.00     2700000.0
Name: AMT_GOODS_PRICE
```

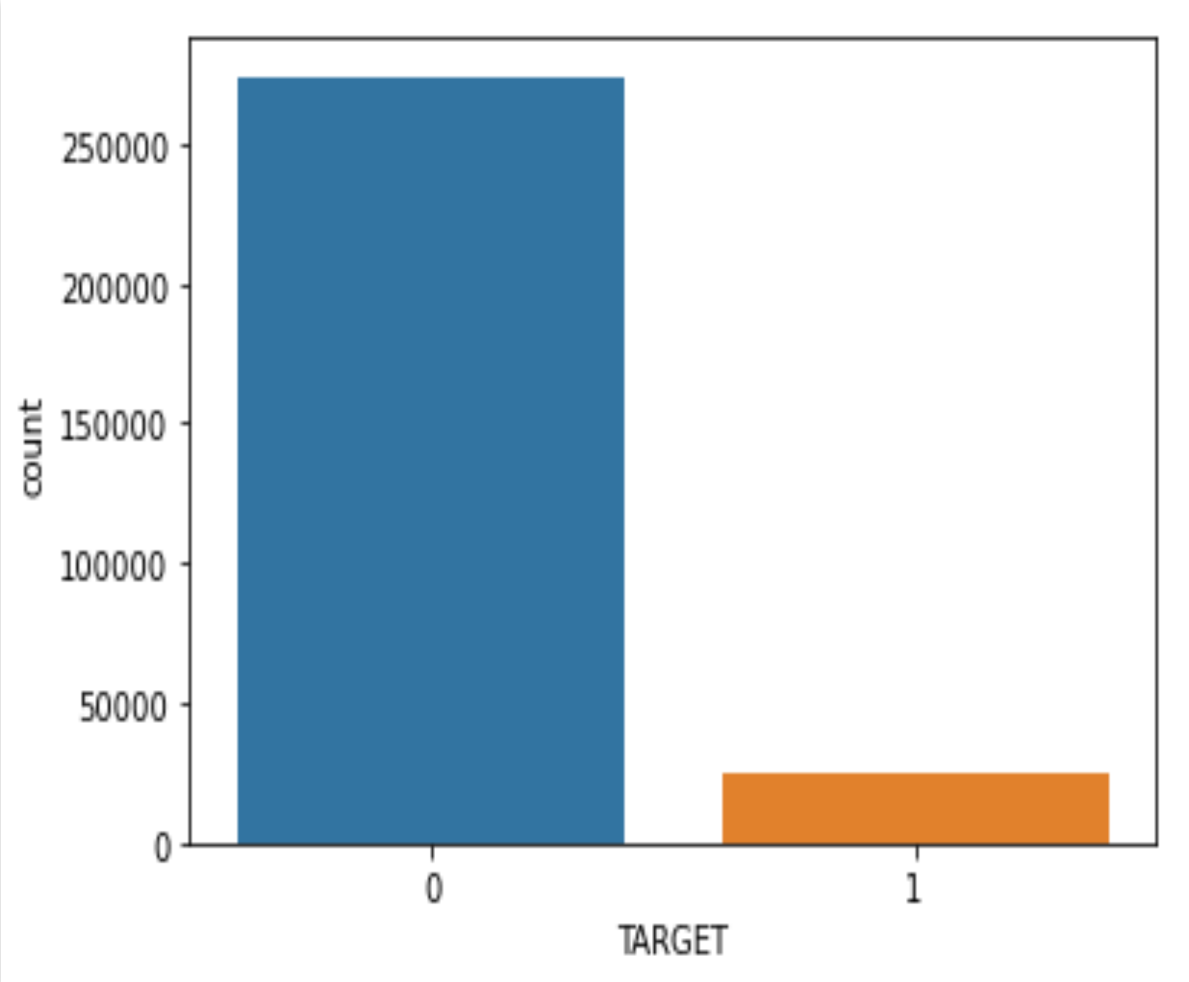


Data Imbalance

- From the above count plot and the value counts of the TARGET column, we can clearly state that there is data imbalance.
- Imbalance Ratio is 18%
- It means there will be a huge difference in density for both the plots.

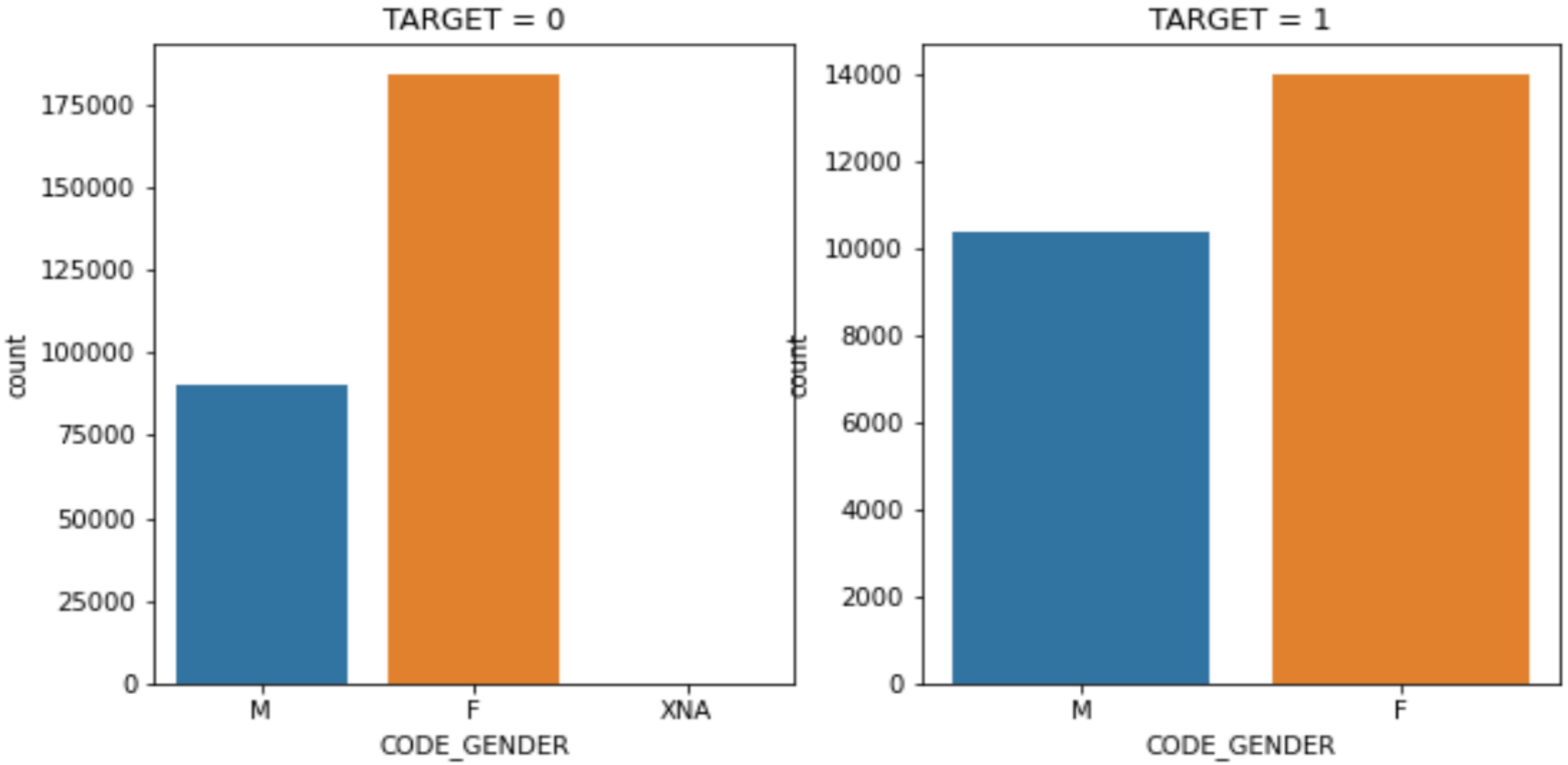


```
0     91.836694
1      8.163306
Name: TARGET, 
```



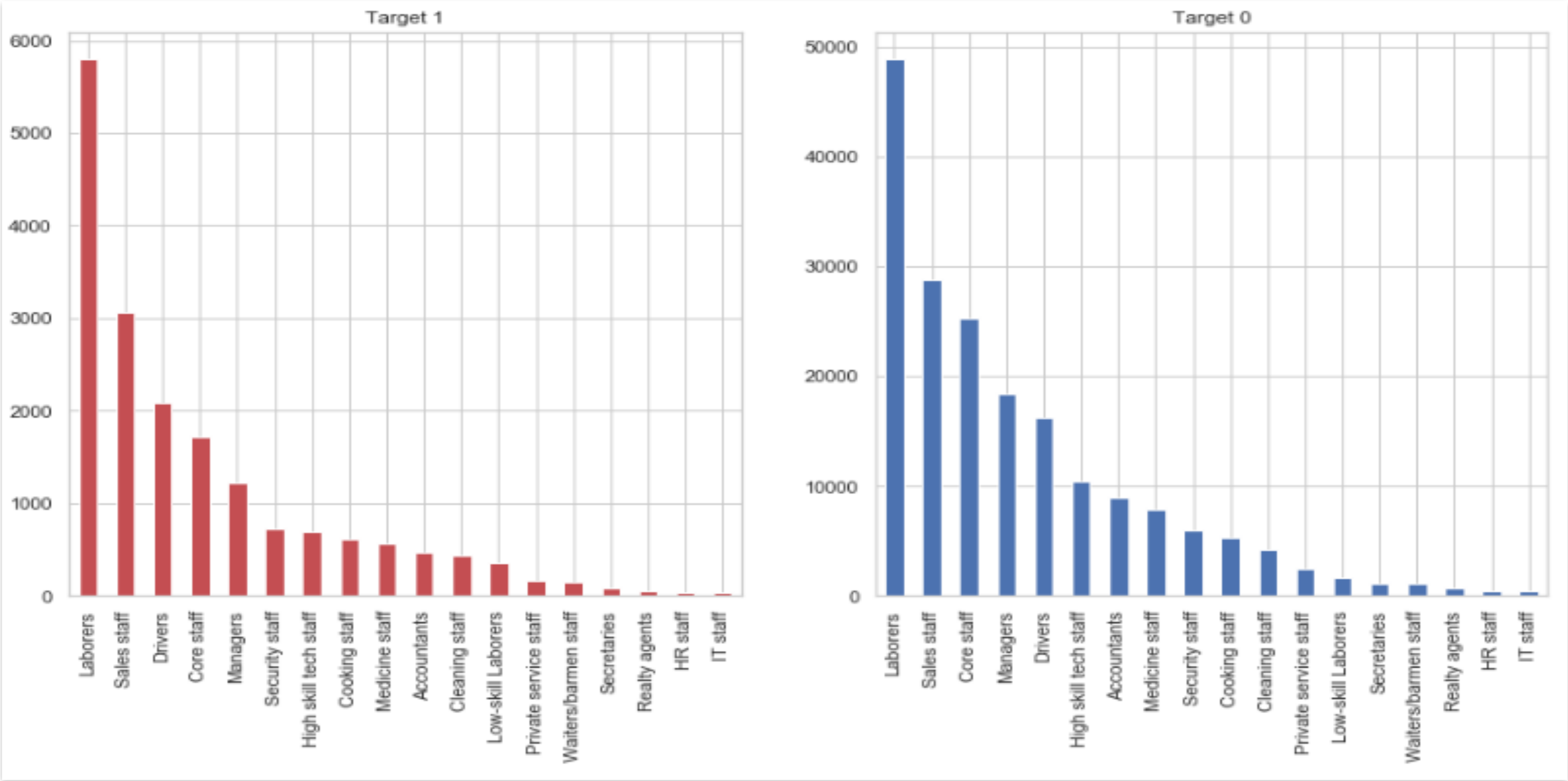
Univariate Analysis

Type of Gender



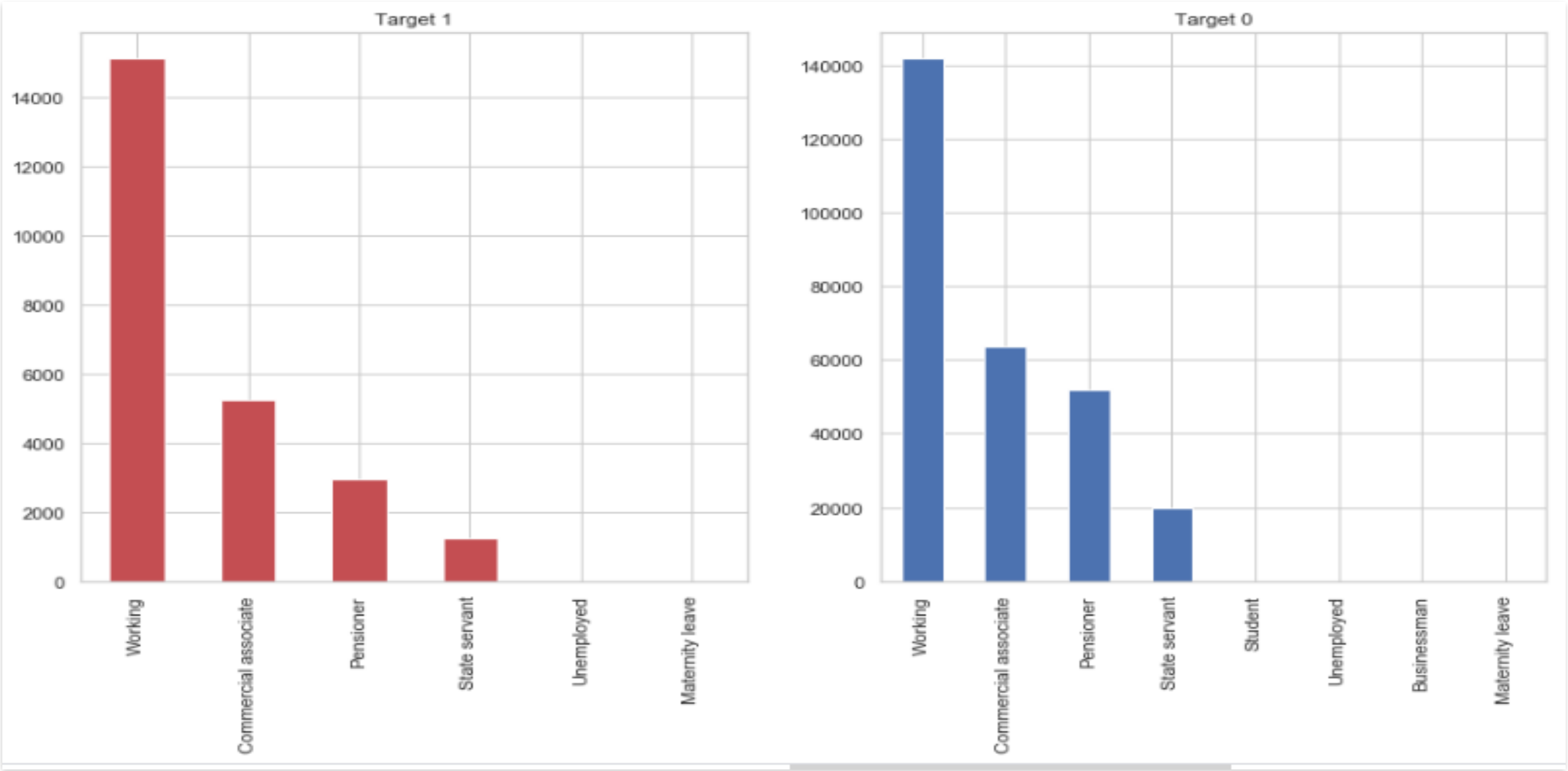
We see that number of Females take more loans when compared to number of Males.
7% of Female applicants are defaulters.
10.45% of Male applicants are defaulters.

Type of Occupation



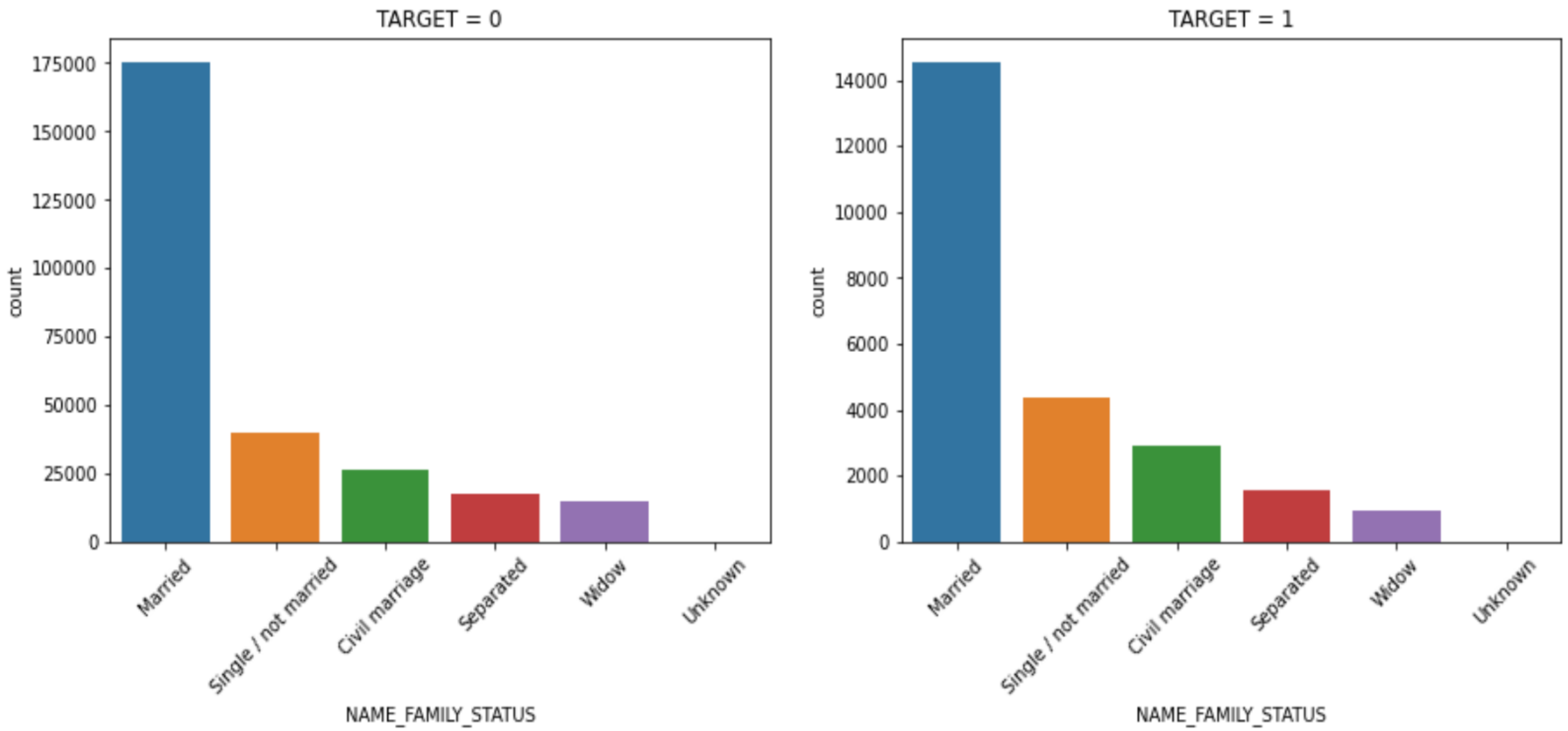
We can see that most of the loans are taken by Labourers followed by Sales Staff, Core Staff etc. and the defaulters to non defaulters also follow the same sequence.

Type of Income



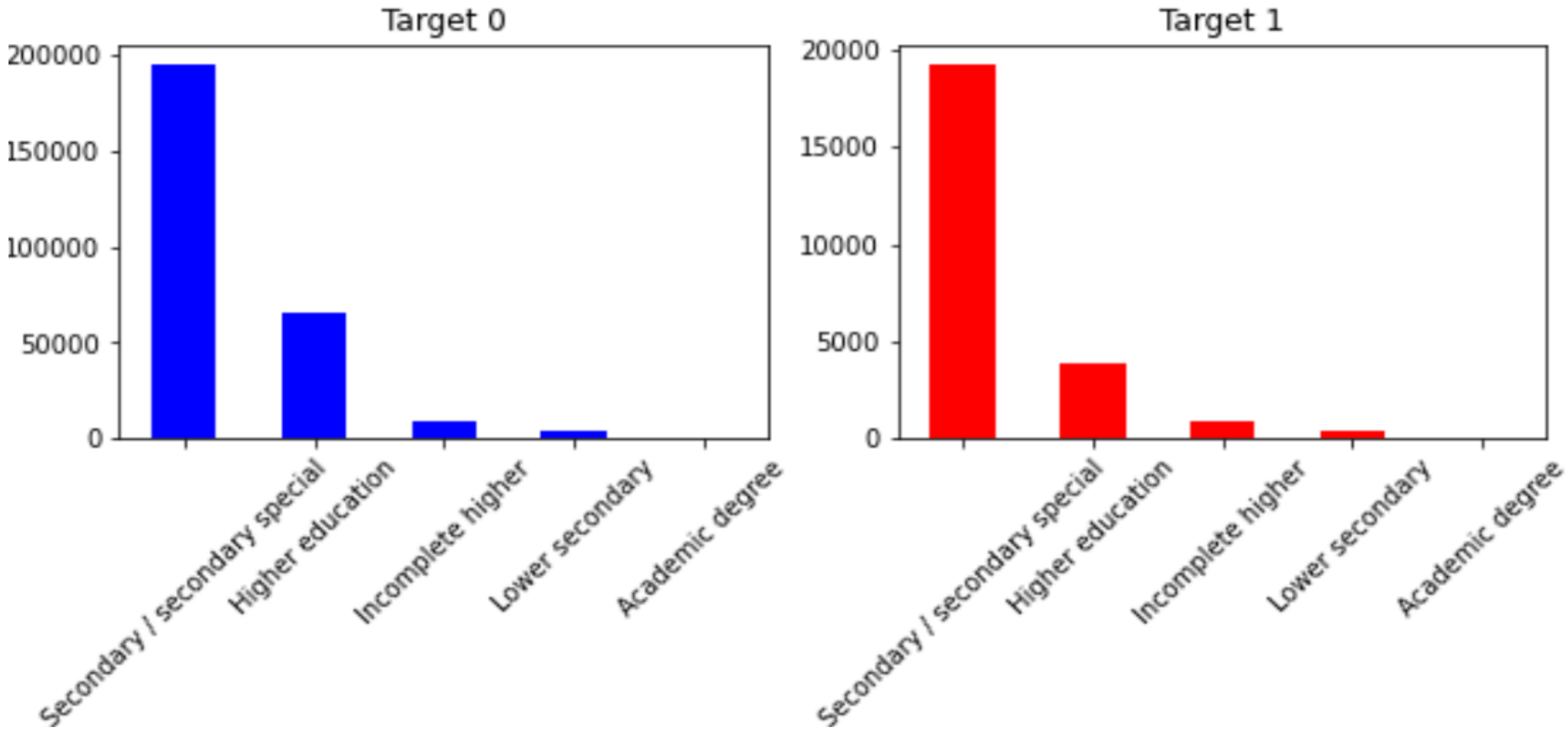
We see that working professionals take most of the loans.

Family Status



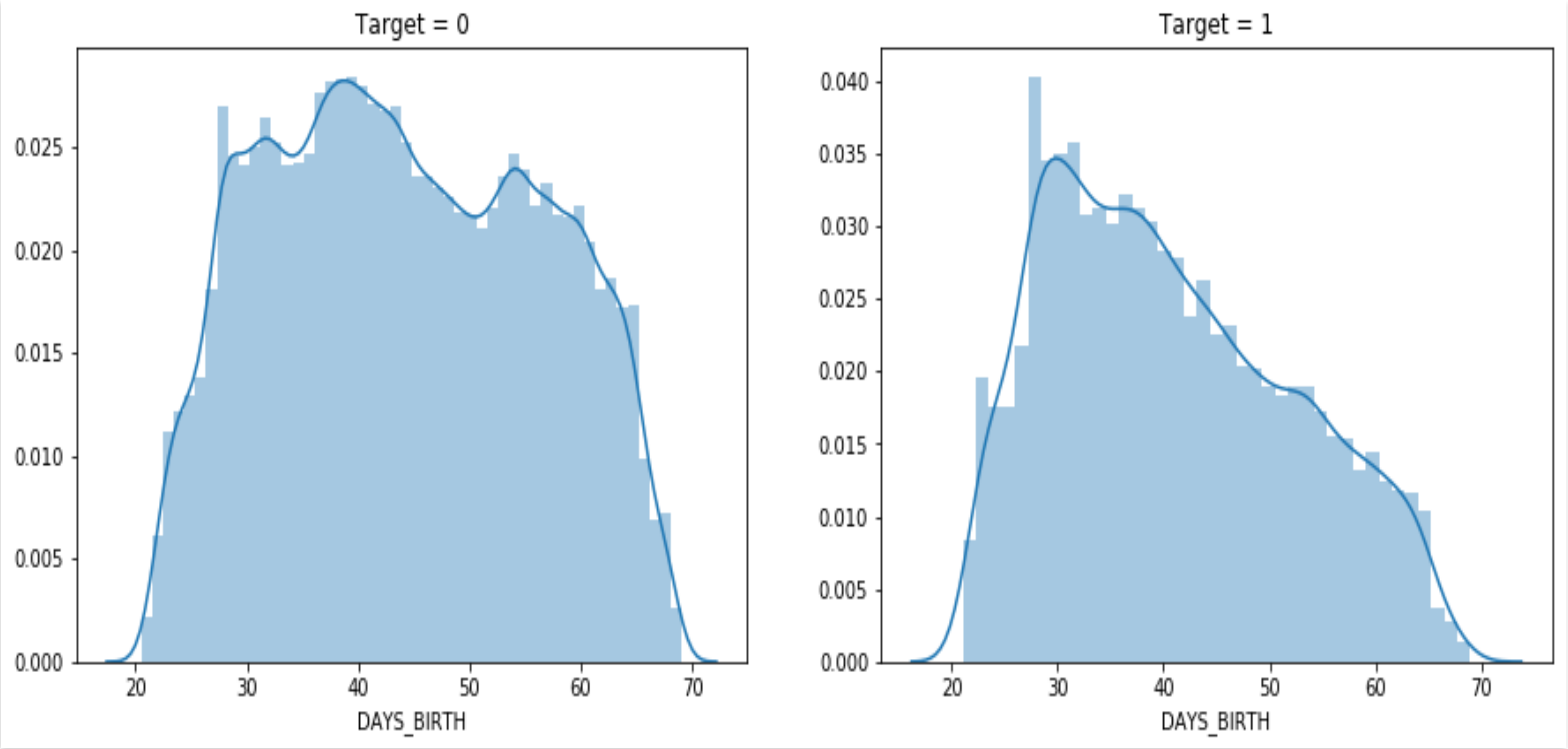
We also notice that most of the loans are taken by clients who are married.

Type of Education



We see that clients with Secondary Education take most number of loans.

Age of Applicants (in years)

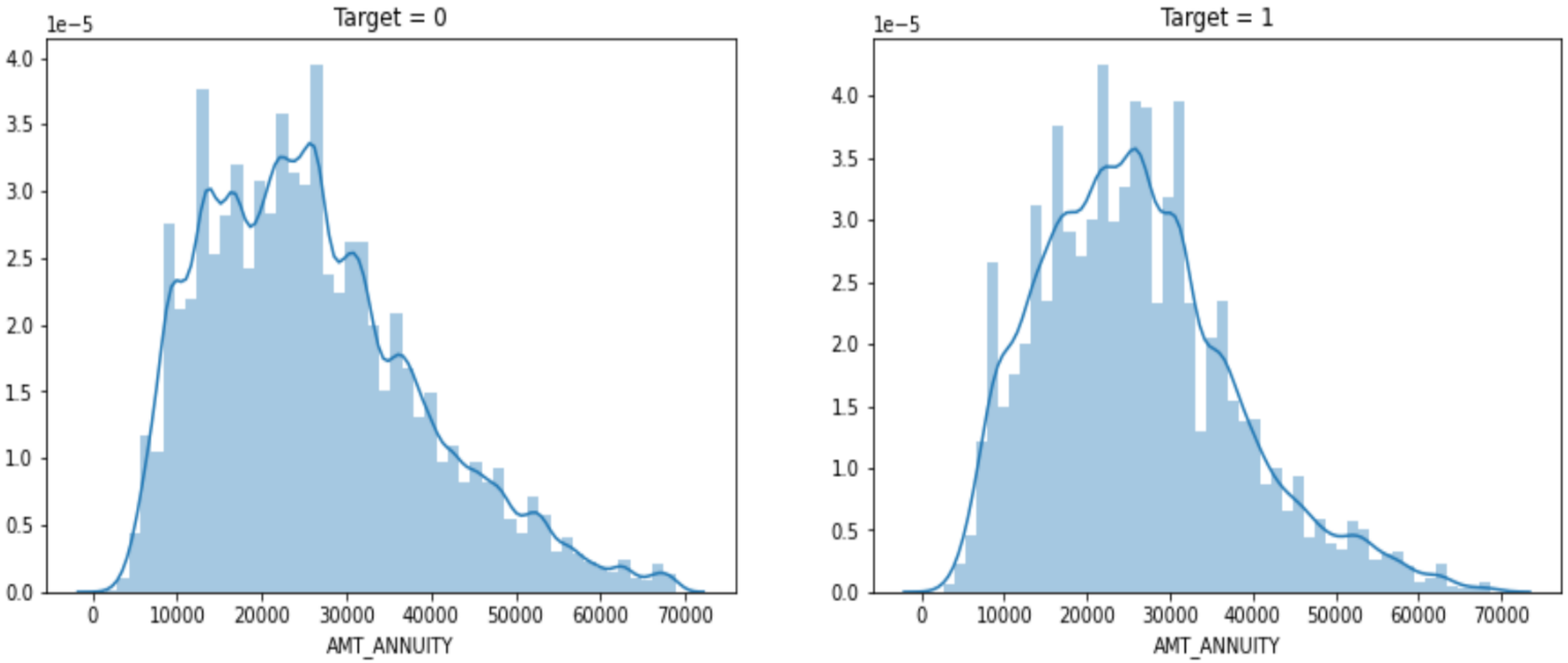


We notice that the age of applicants in the case of non-defaulters are normally distributed whereas they are left skewed in the case of defaulters.

Applicants around 30 years of age are likely to be Defaulters.

But as the age increases number of Defaulters tend to decrease.

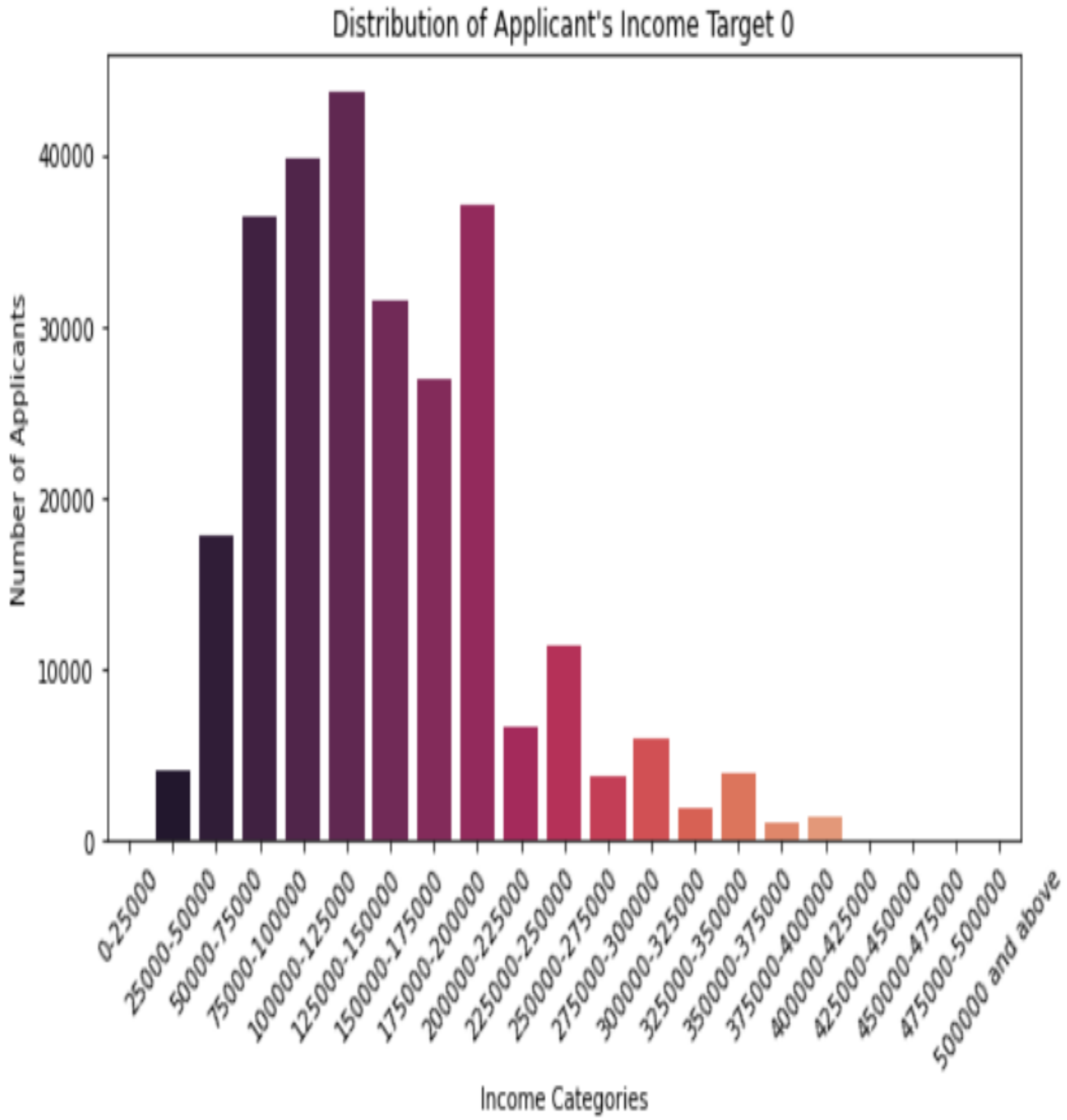
Loan Annuity Distribution

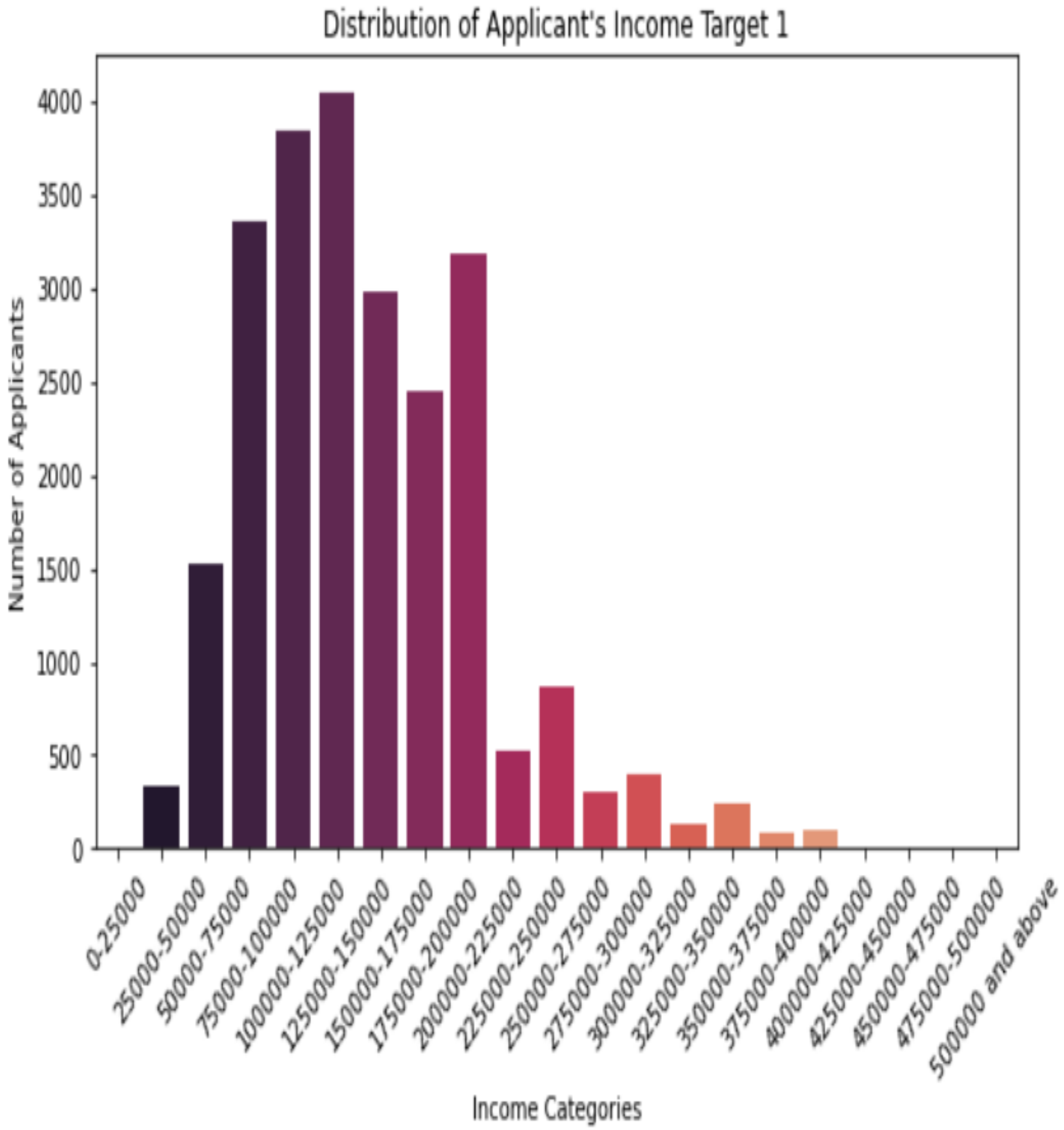


We notice that the Annuity amount data is more skewed towards left in the case of non-defaulters.

Around 70% of people with low loan annuity are bound to pay their loan on time.

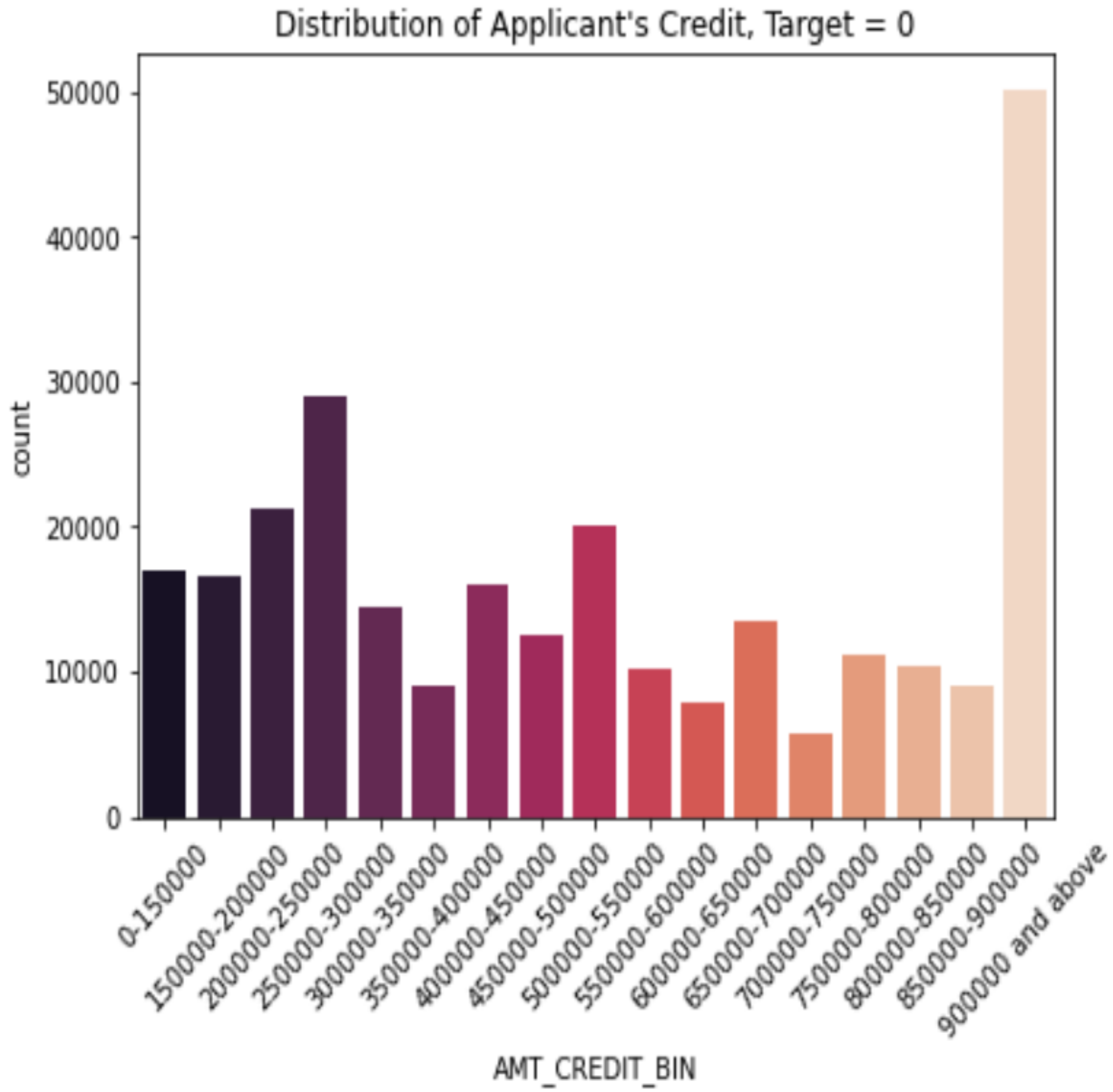
Income Distribution

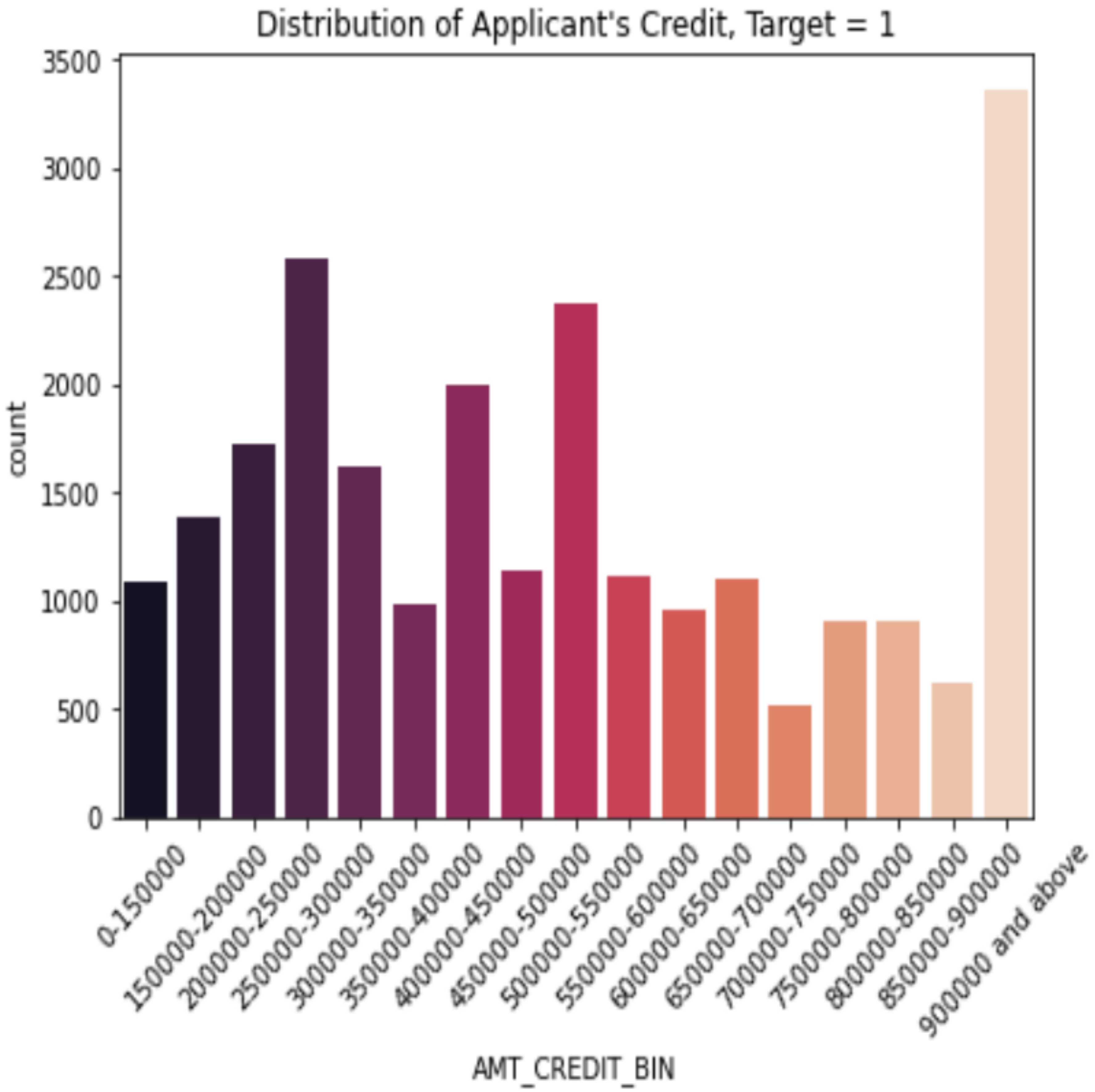




Most of the applicants lie between 50k and 225k income range.

Credit Distribution

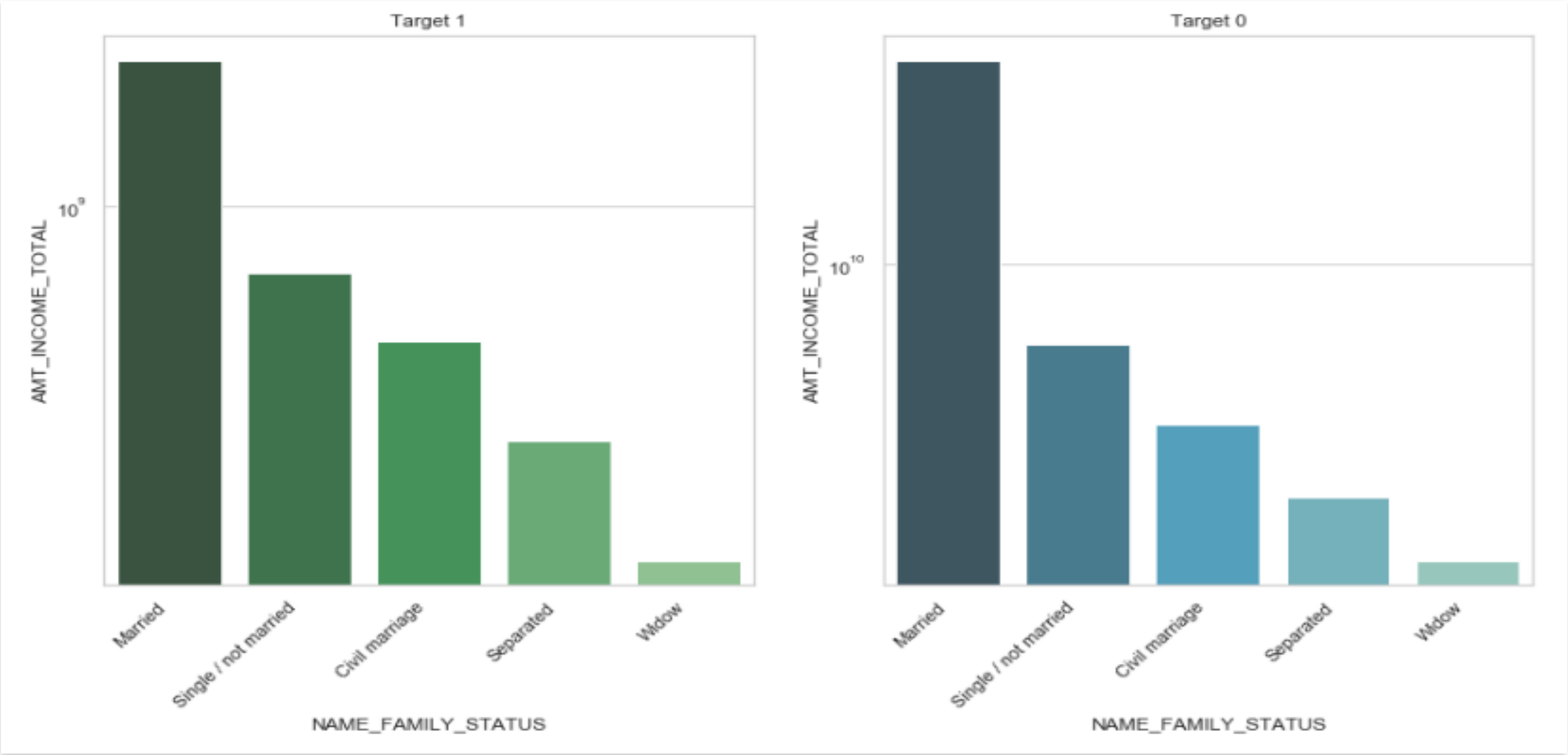




In both categories, most applicants have taken the credit in the range of 250k to 300k.

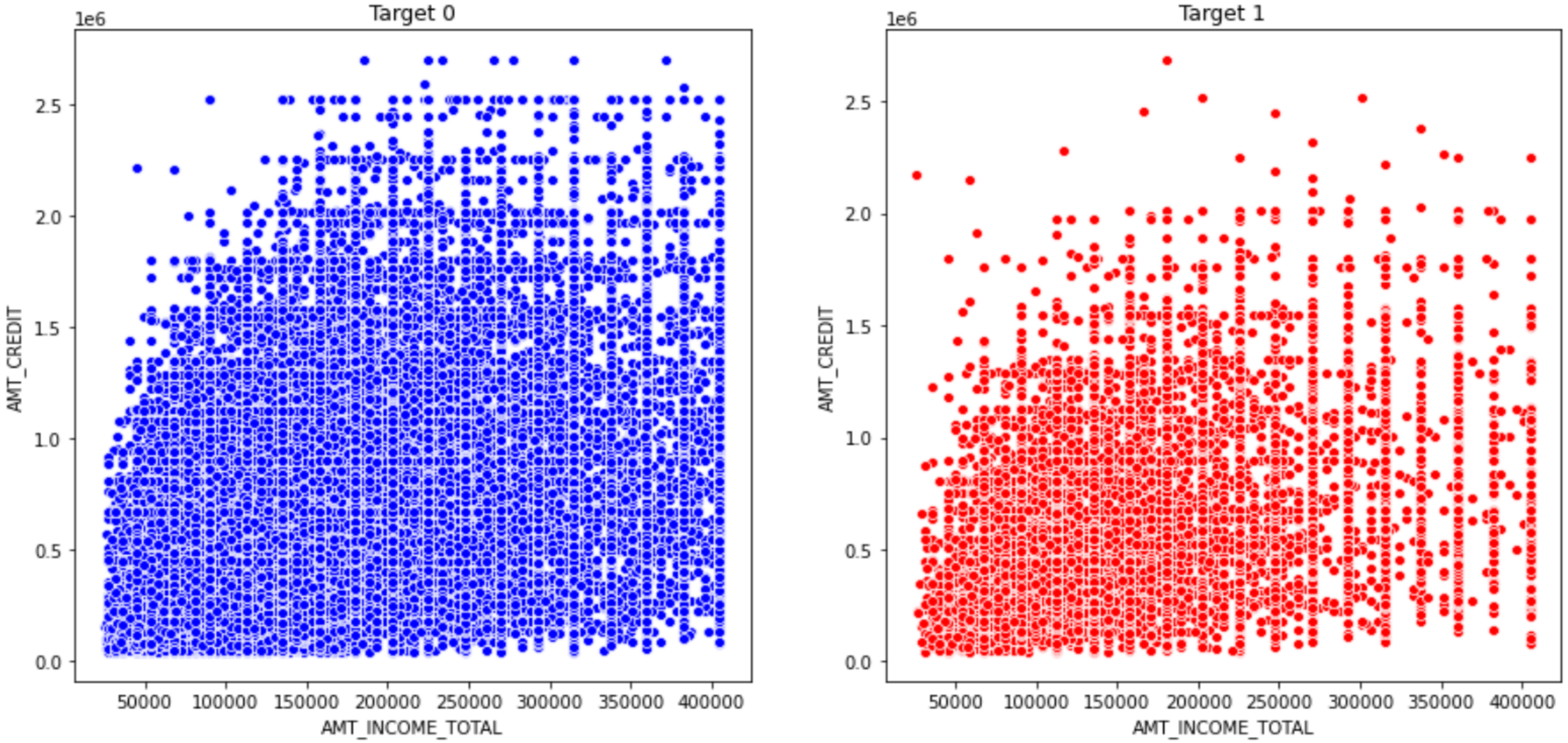
Bivariate Analysis

AMT_INCOME_TOTAL



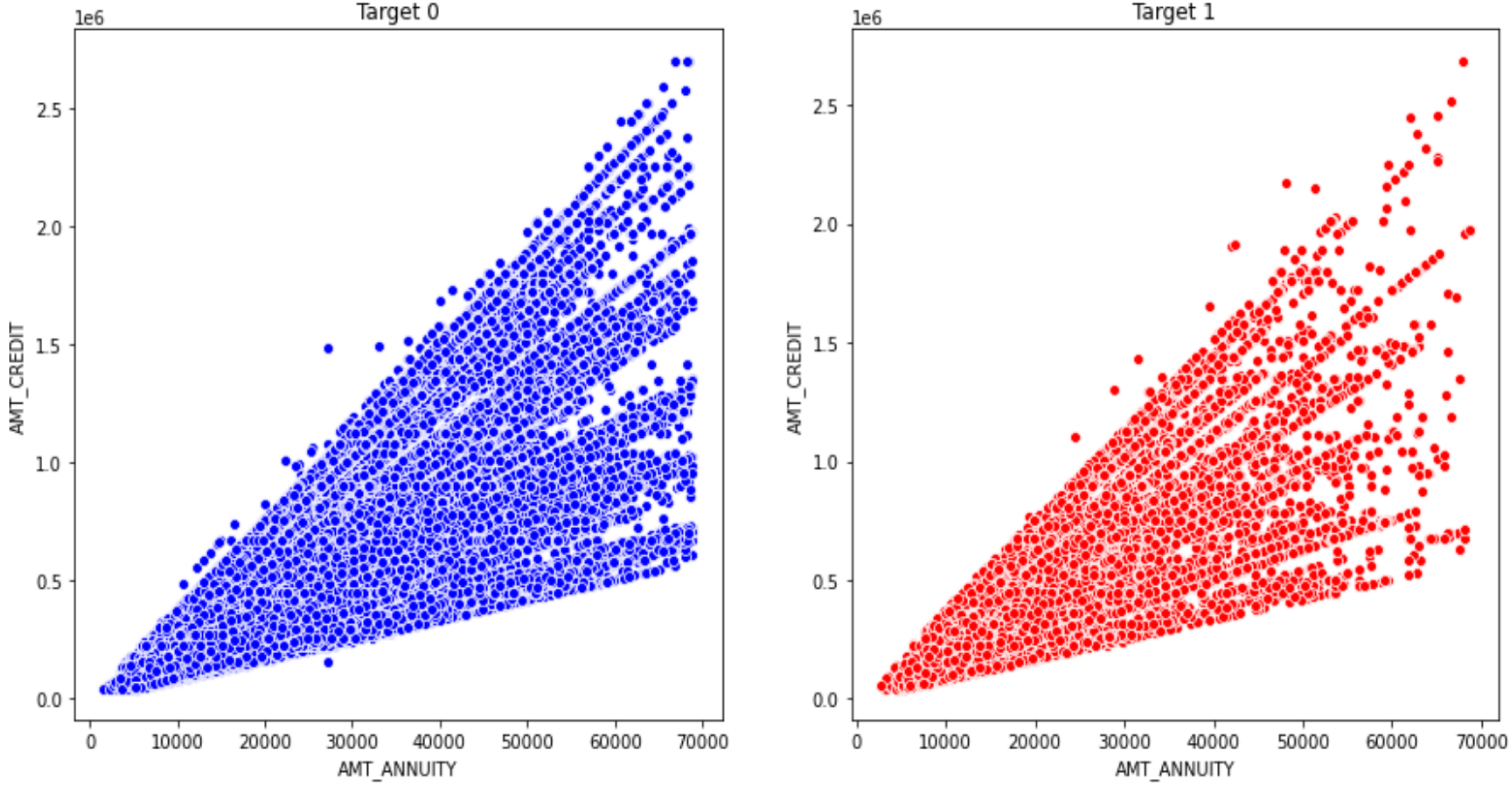
Most of the loans are taken by Married customers whereas widowers do not take much loans.

Income vs Credit



In the case of defaulters [Target 1] ,
maximum loan given to people with income till 250k is around 130k.

Distribution of Loan Annuity vs Credit



Loan Annuity vs Credit

For the same credit amount - some applicants have lower AMT_ANNUITY, others have higher. This can be due to two things :

1. Risk Involved
2. Tenure

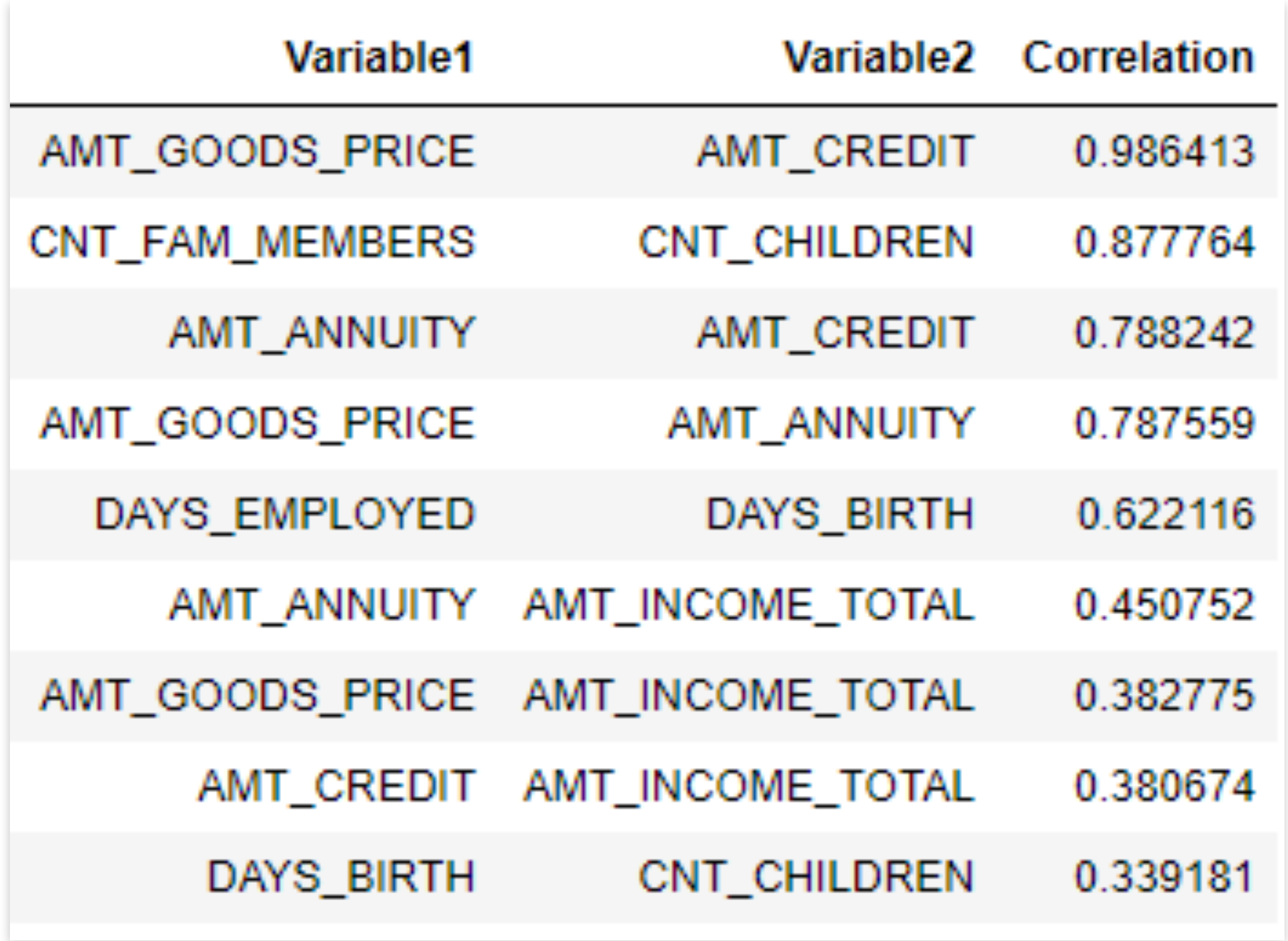
The upper edge of the cone has the applicants with the best possible AMT_ANNUITY amounts. This also marks the lowest AMT_ANNUITY one could get for a particular AMT_CREDIT.

The lower edge of the cone has the applicants who were assumed to be at a greater risk of being a defaulter. This also marks the highest AMT_ANNUITY bank can recieve from the applicants.
We notice that the applicants with higher risk are likely to be defaulters.

After comparing the two data sets for Target = 1 and Target = 0, we find that the sets of correlated columns are same for both the datasets.

AMT_GOODS_PRICE is highly correlated to AMT_CREDIT in both the datasets.

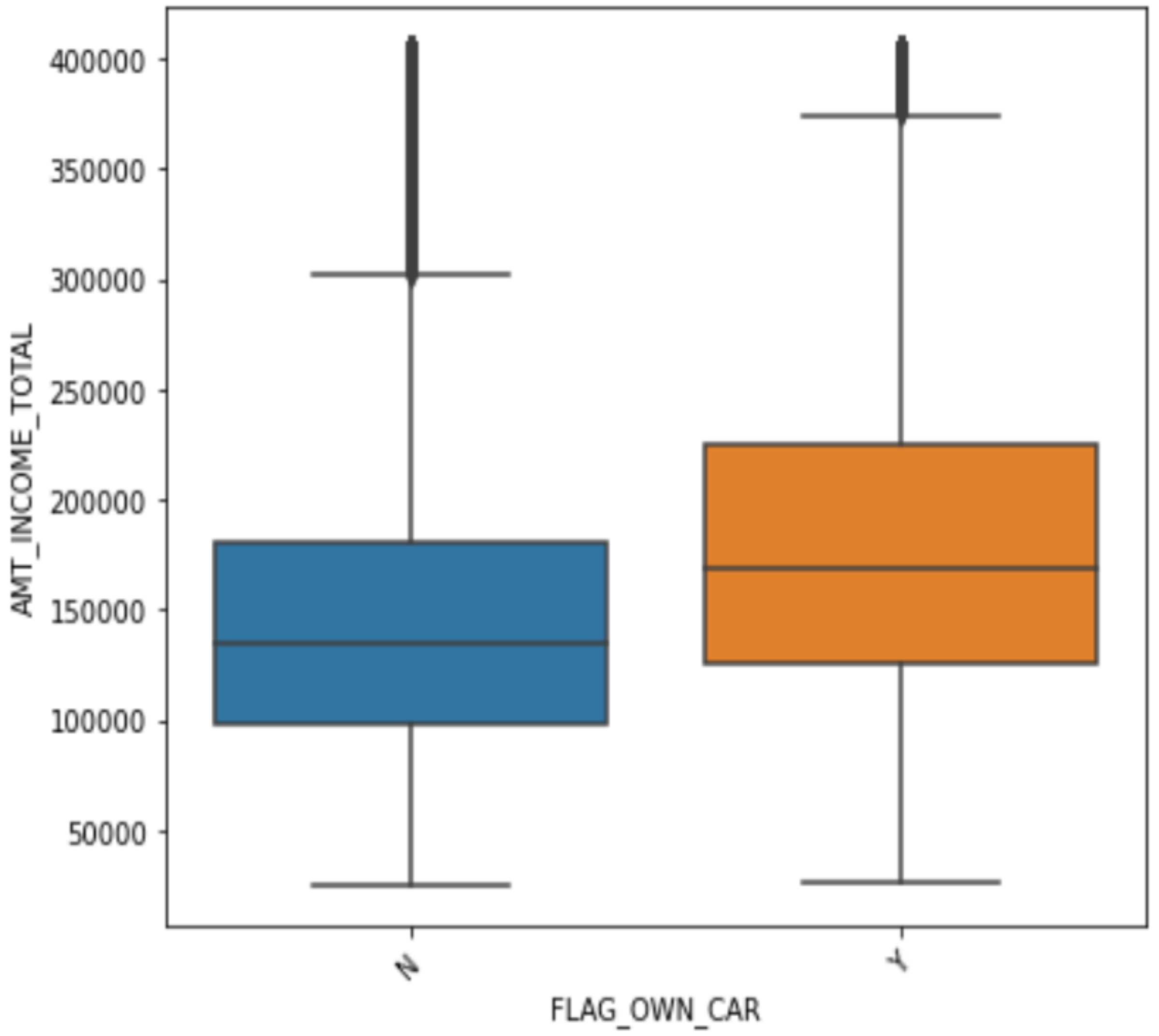
Similarly CNT_FAM_MEMBERS is highly correlated to CNT_CHILDREN and so on.

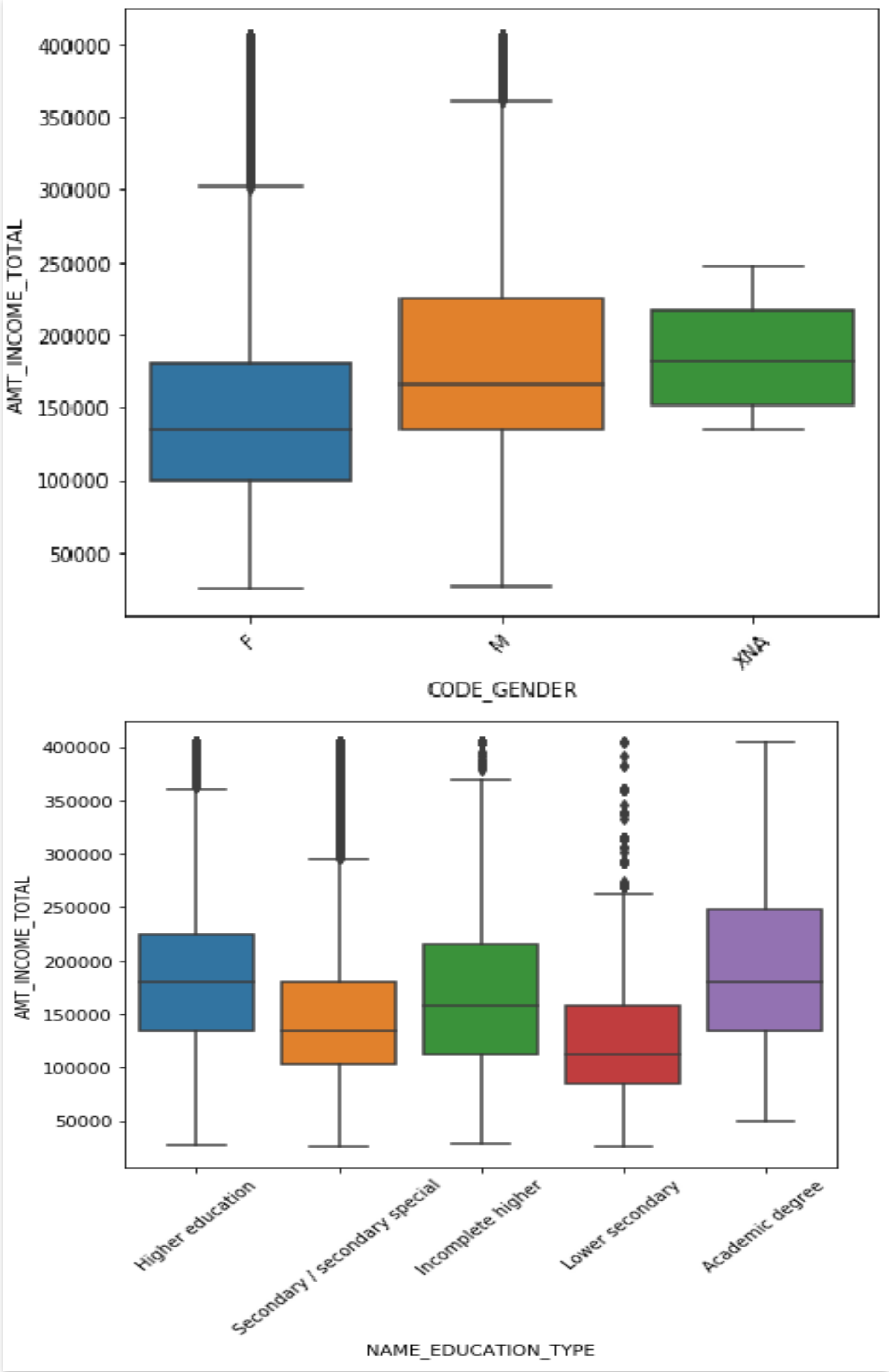


Variable1	Variable2	Correlation
AMT_GOODS_PRICE	AMT_CREDIT	0.986413
CNT_FAM_MEMBERS	CNT_CHILDREN	0.877764
AMT_ANNUITY	AMT_CREDIT	0.788242
AMT_GOODS_PRICE	AMT_ANNUITY	0.787559
DAYS_EMPLOYED	DAYS_BIRTH	0.622116
AMT_ANNUITY	AMT_INCOME_TOTAL	0.450752
AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.382775
AMT_CREDIT	AMT_INCOME_TOTAL	0.380674
DAYS_BIRTH	CNT_CHILDREN	0.339181

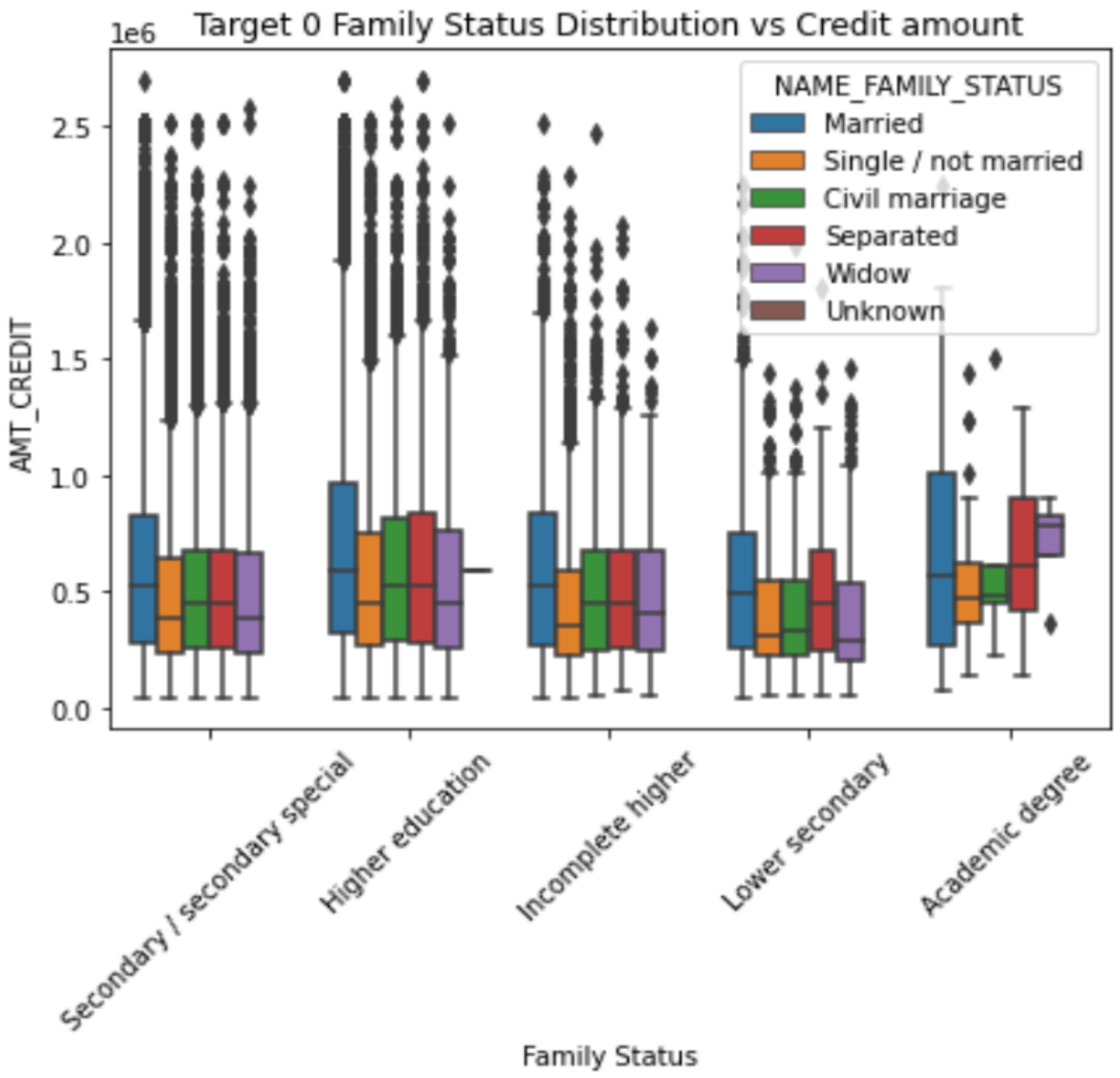
Comparative study of total Income and other categories

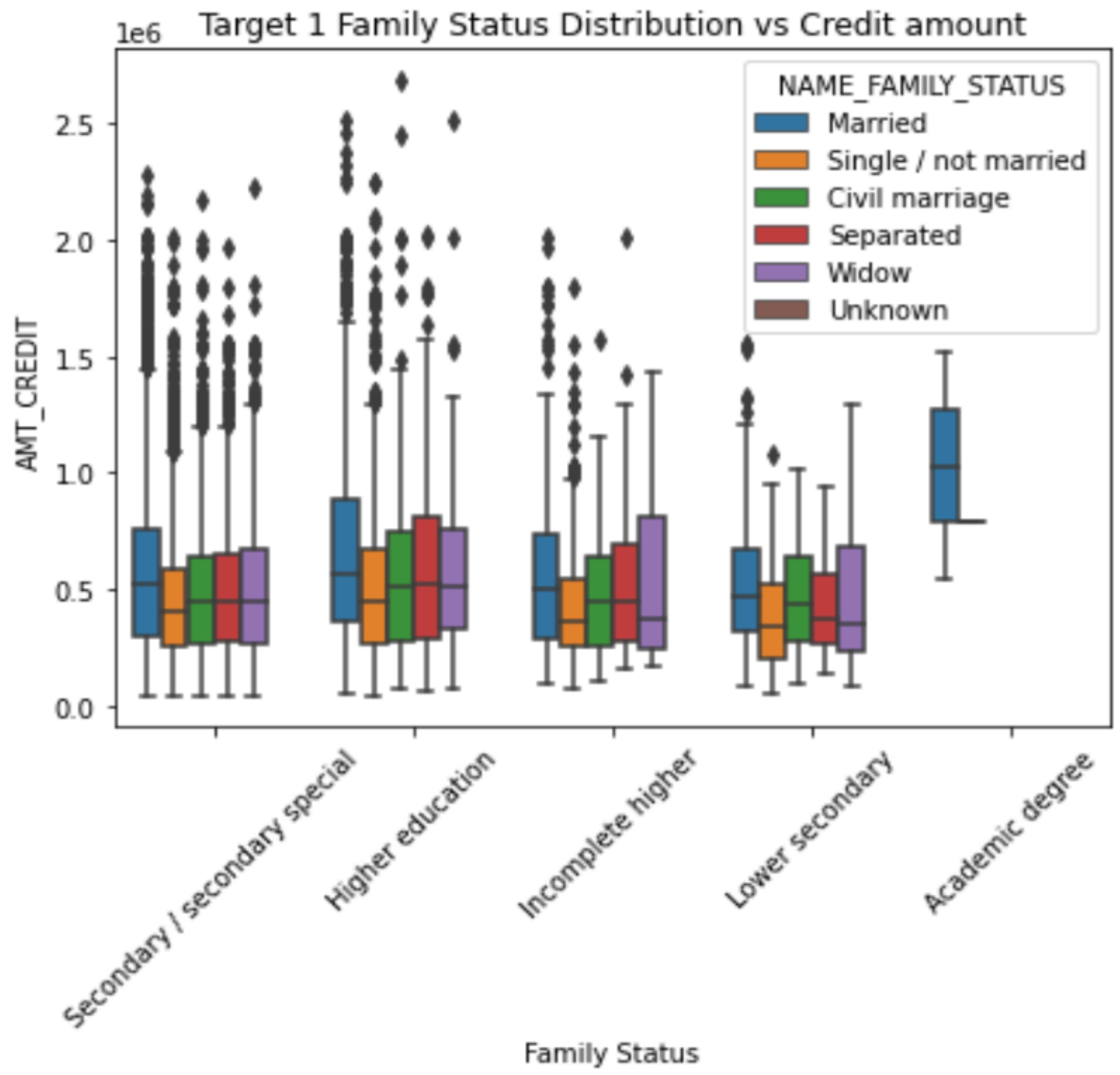
- Females have comparatively lower income than Males and Others.
- People with comparatively higher income own the cars.
- There is no significant of income on owning a house or a flat.
- Clients with Higher education and Academic degree have comparatively higher income than others.
- There is no significant of income on Family Status.



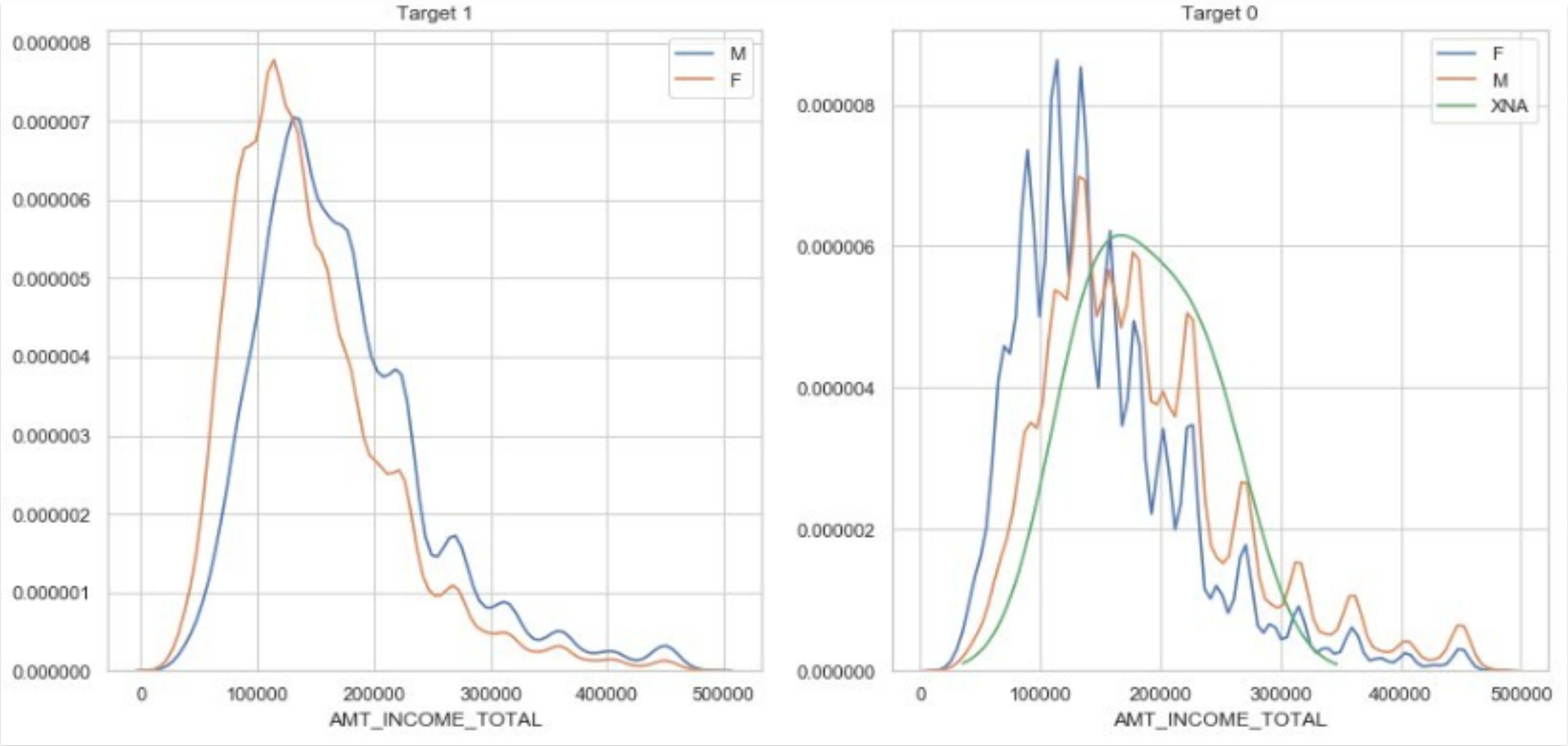


Family Status Distribution vs Credit amount





It is clearly visible that married applicants are more likely to be defaulters than other categories in Family Status

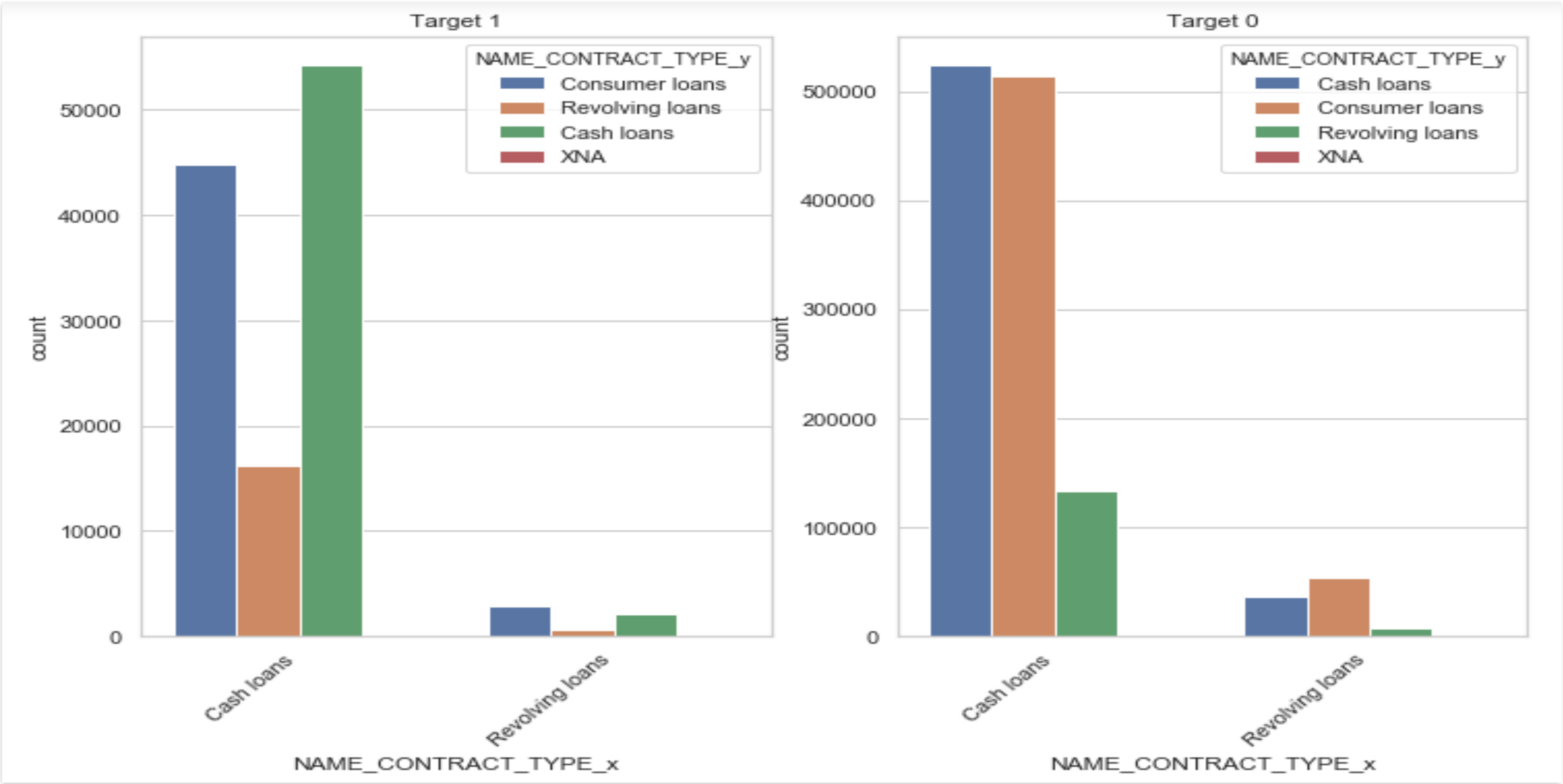


We notice that the clients with income between 50000 & 250000 take most of the loans and most of the defaulters are also from the same range across all genders.

Previous Application Data Processing

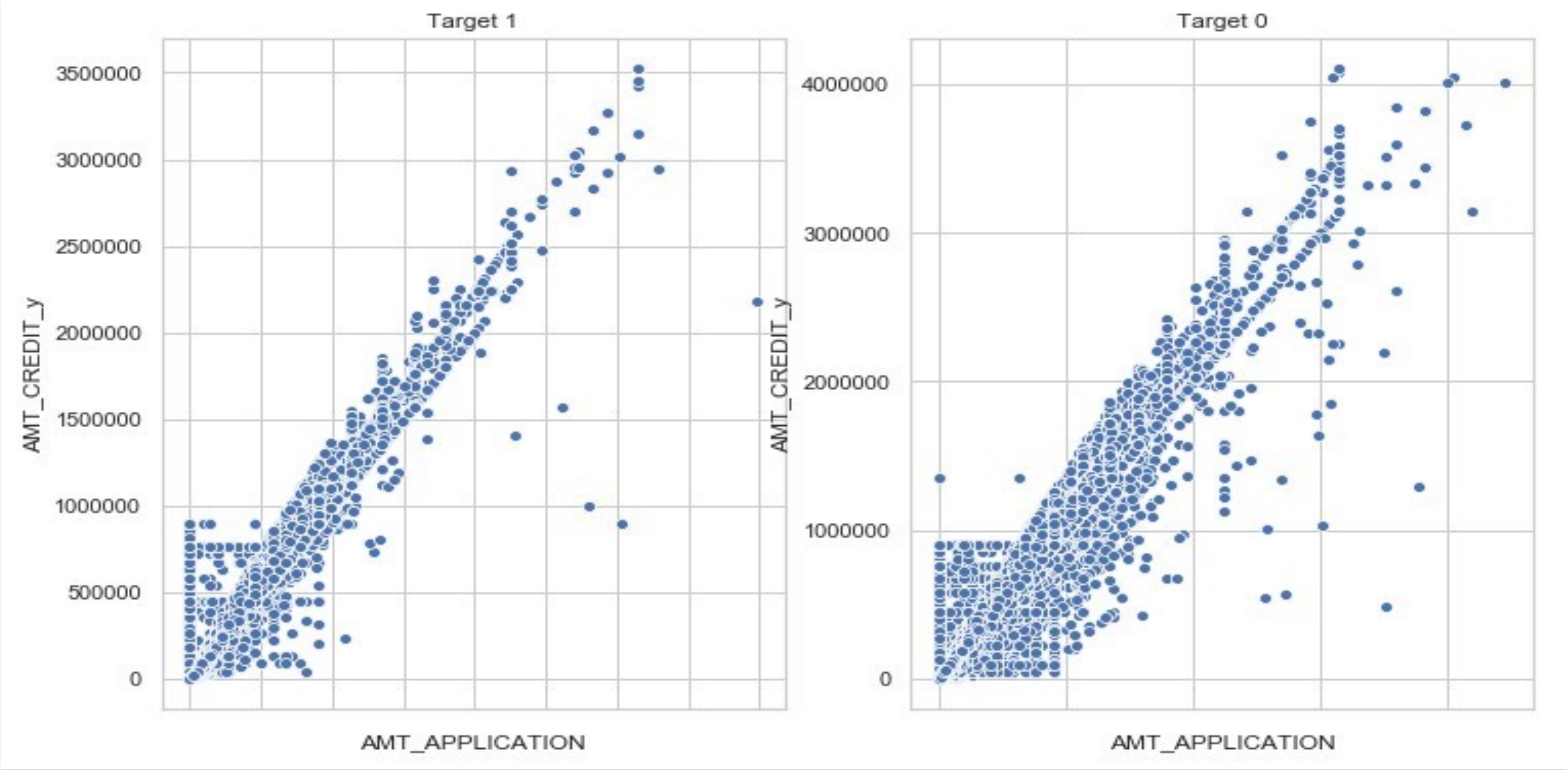
Now we will do the analysis for the Target variable based on the information given to us in the previous application data

Shift in contract type from previous to current application



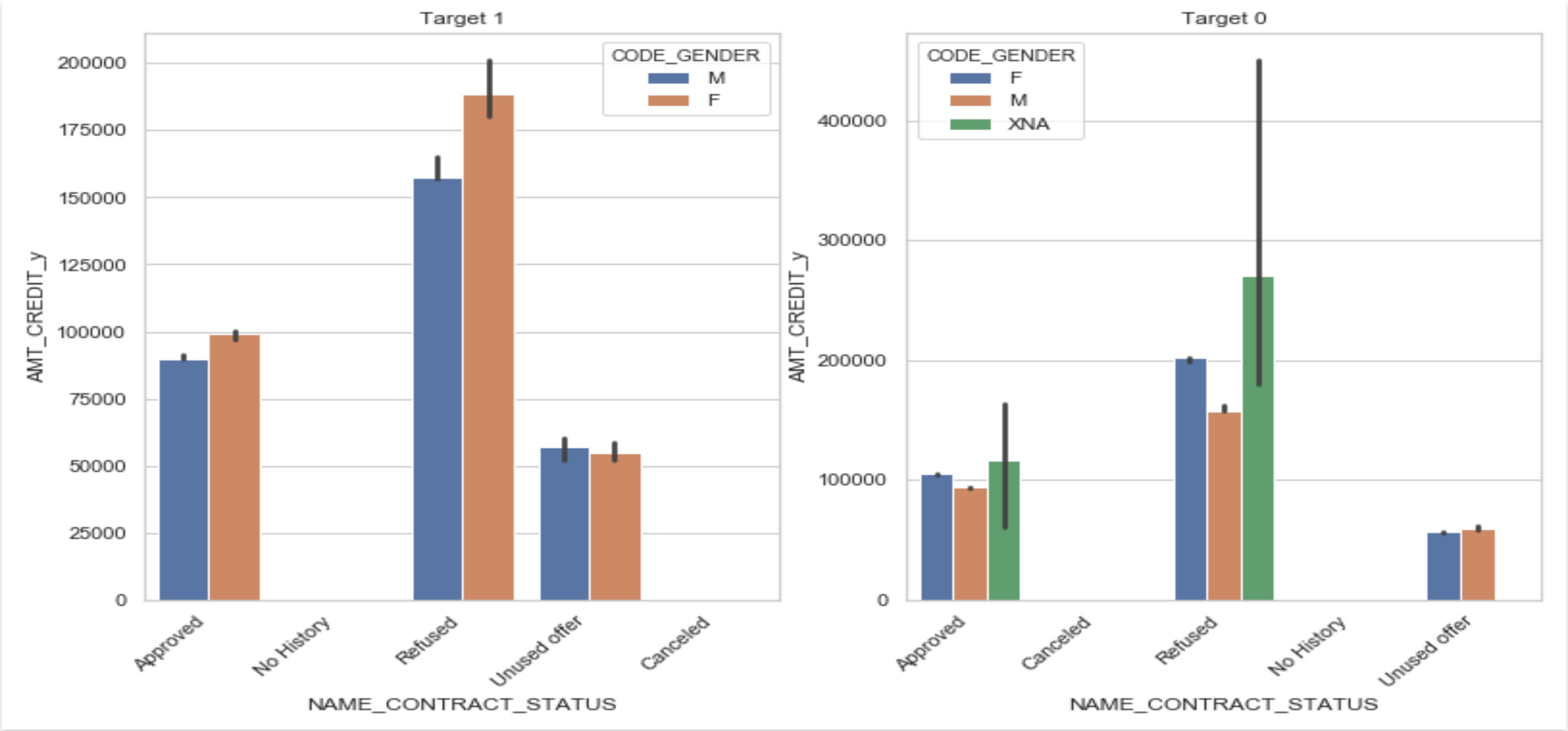
The two categories – Consumer loans and XNA have been shifted to Cash and Revolving loans

Checking the variation between credit amount and the amount for which the client initially applied in the previous application.

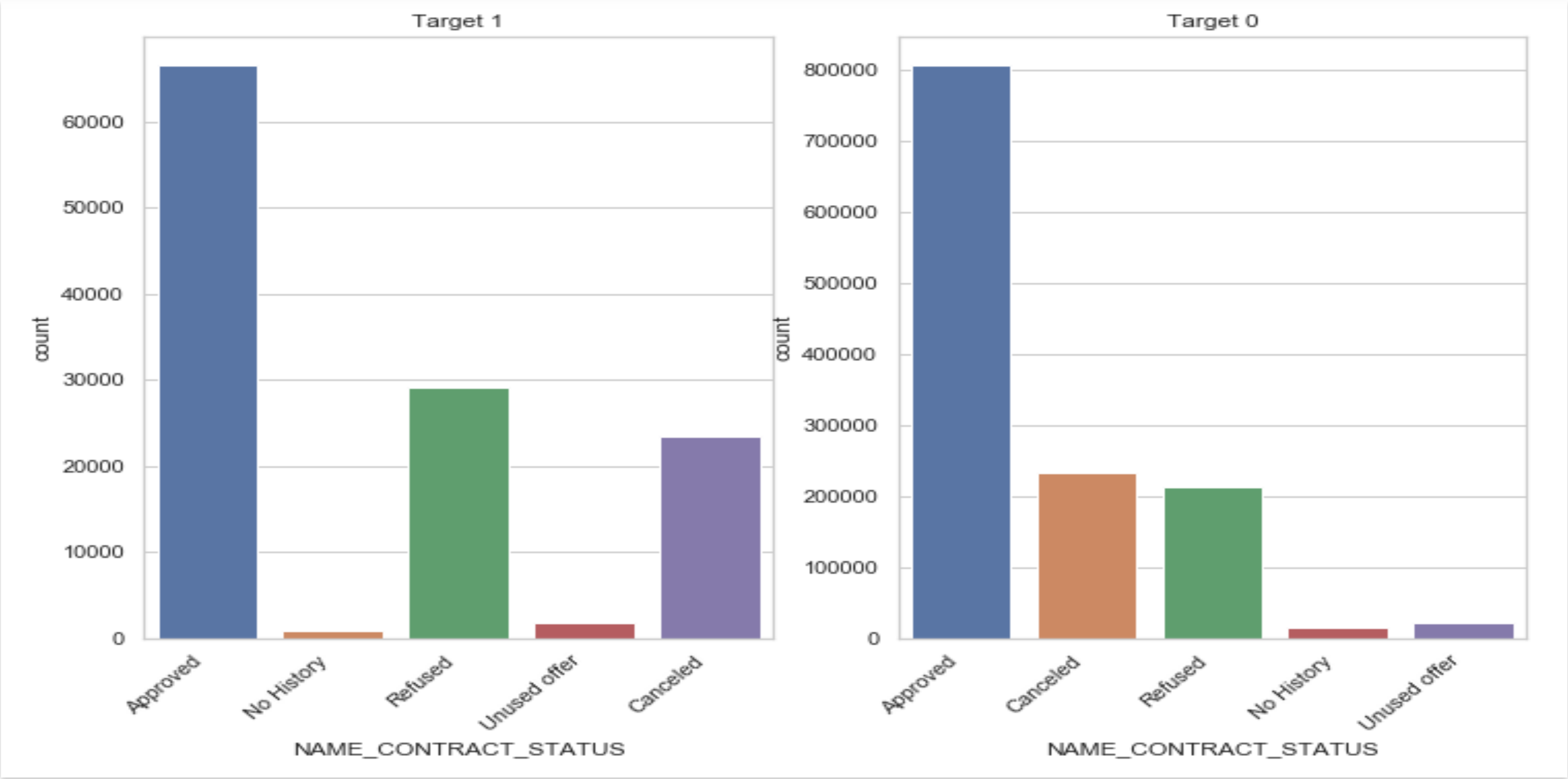


The graph is more scattered in case of non defaulters.

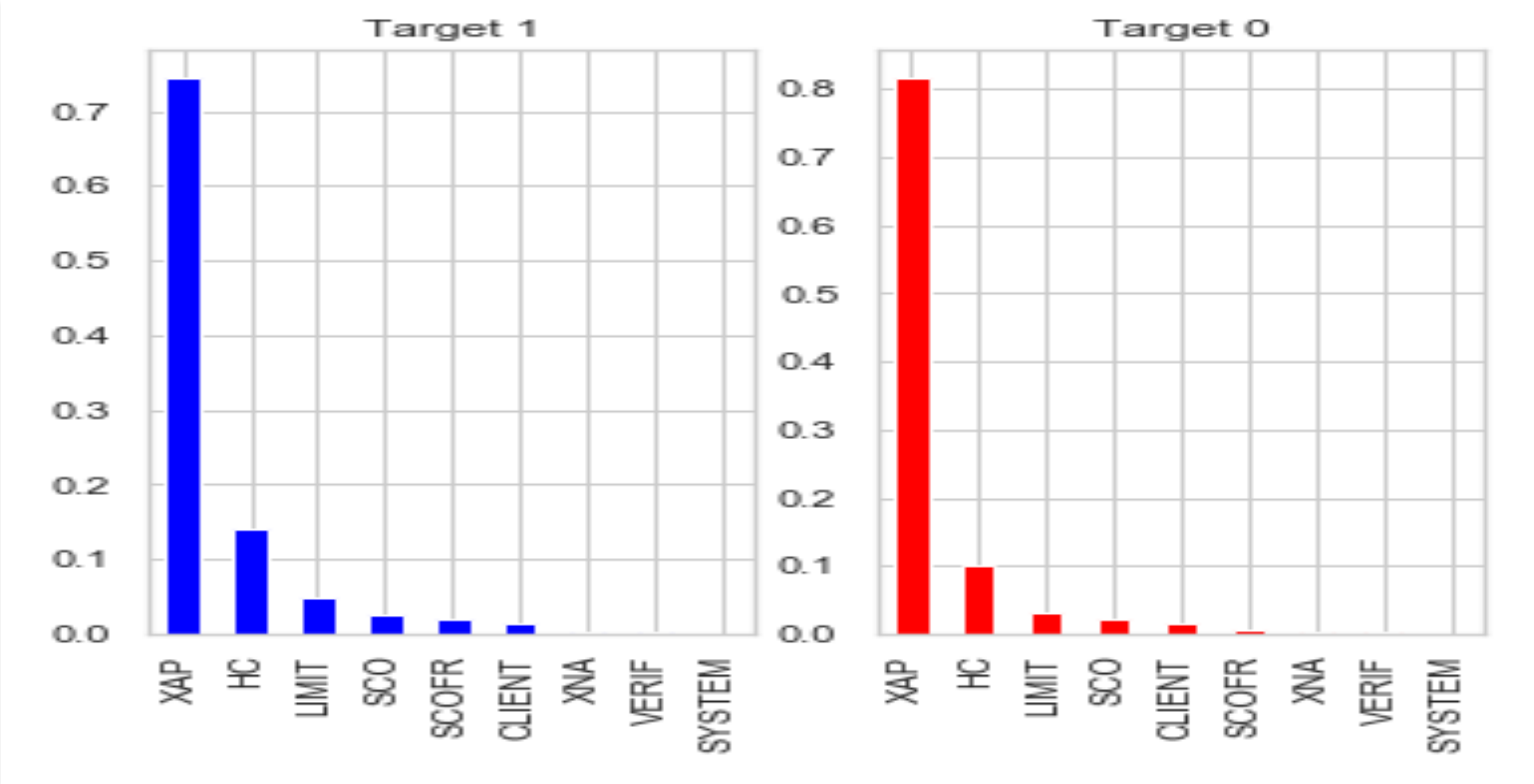
Previous application contract status and median credit amount by gender.



Customers in current application with previous application contract/loan status. Loan seems to be approved for majority of the defaulters which seem to have caused a loss to the company. But also majority of the non- defaulters have their loans sanctioned.



Checking the cause of rejection in previous applications. And it has been same in both the cases.


Target 1
XAP
HC
LIMIT
SCO
SCOFR
CLIENT
XNA
VERIF
SYSTEM
Target 0
XAP
HC
LIMIT
SCO
CLIENT
SCOFR
XNA
VERIF
SYSTEM