CREDIT EDA

ASSIGNMENT

IPSITA PAL

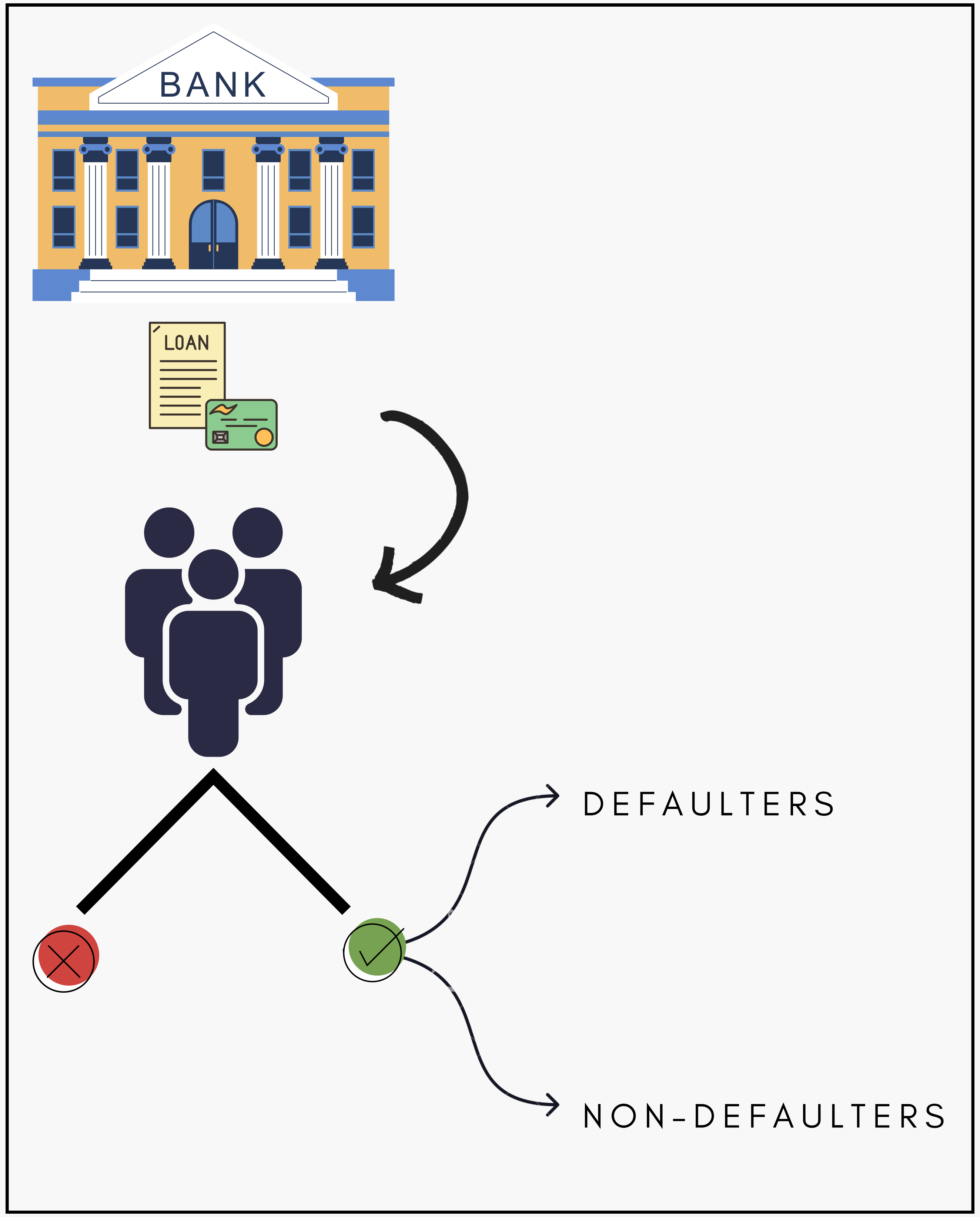
upGrad

upGrad & IIITB | Data Science Program - December 2022

BATCH ID 3459

DS51

TABLE OF CONTENTS



BUSINESS OBJECTIVE

Companies that provide loans find it difficult to provide loans to people because their credit history is insufficient or non-existent. For this reason, some consumers take advantage of it by becoming a defaulter.

When the company receives a loan application, it must make a loan approval decision based on the profile of the applicant. There are two kinds of risks associated with a bank's decision:

If the applicant is likely to repay the loan, failure to approve the loan leads to business loss for the business.

If the applicant is not likely to pay the loan, ie. is likely to be in breach of its obligations, approval of the loan may result in a financial loss to the business.

PROBLEM STATEMENT

When a client applies for a loan, they are able to make four types of decisions:

1. Approved: The loan application was approved by the corporation.
2. Cancelled: Client canceled application during approval. Either the client changed her/his mind about the loan or in some cases due to a higher risk of the client, he received worse pricing which he did not want.
3. Refusal: The company refused the loan (because the client is not meeting his needs, etc.).
4. Unused offer: The loan was cancelled by the client but at various stages in the process.

Understand how consumer and lending attributes influence default patterns.
The company wants to understand the factors (or driver variables) behind the default, i.e.. the variables which are strong indicators of default.

So they can use those metrics in the future.

DATA UNDERSTANDING

The data given below contains the information about the loan application at the time of applying for the loan. It contains two types of scenarios:

1.The client with payment difficulties: he/she had late payment more than X days on at least one of the first Y instalments of the loan in our sample.
2.All other cases: All other cases when the payment is paid on time.

Datasets Provided:

1. '*application_data.csv*' contains all the information of the client at the time of application.
The data is about whether a client has payment difficulties.

2. '*previous_application.csv*' contains information about the client's previous loan data. It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

3. '*columns_description.csv*' is data dictionary which describes the meaning of the variables.

EDA STEPS:

1 IMPORT ALL THE NECESSARY LIBRARIES

2 LOADING THE DATASETS PROVIDED

3 DATA UNDERSTANDING & CHECKING NULL VALUES

4 FIXING NULL VALUES AND DROPPING UNNECESSARY COLUMNS

5 OUTLIER OBSERVATIONS

6 BINNING VARIABLES

7 DATA ANALYSIS

A DATA IMBALANCE PERCENTAGE

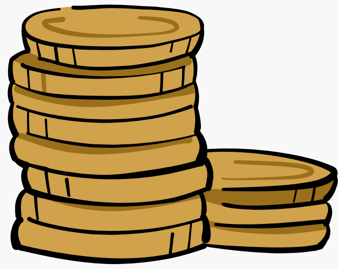
B UNIVARIATE ANALYSIS: CATEGORICAL AND NUMERIC

C BIVARIATE ANALYSIS: CATEGORICAL AND NUMERIC

D CORRELATION MATRIX

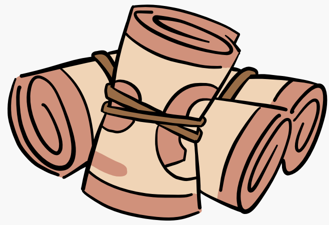
8 MERGING BOTH DATA SET AND FURTHER ANALYSIS

9 REMARKS

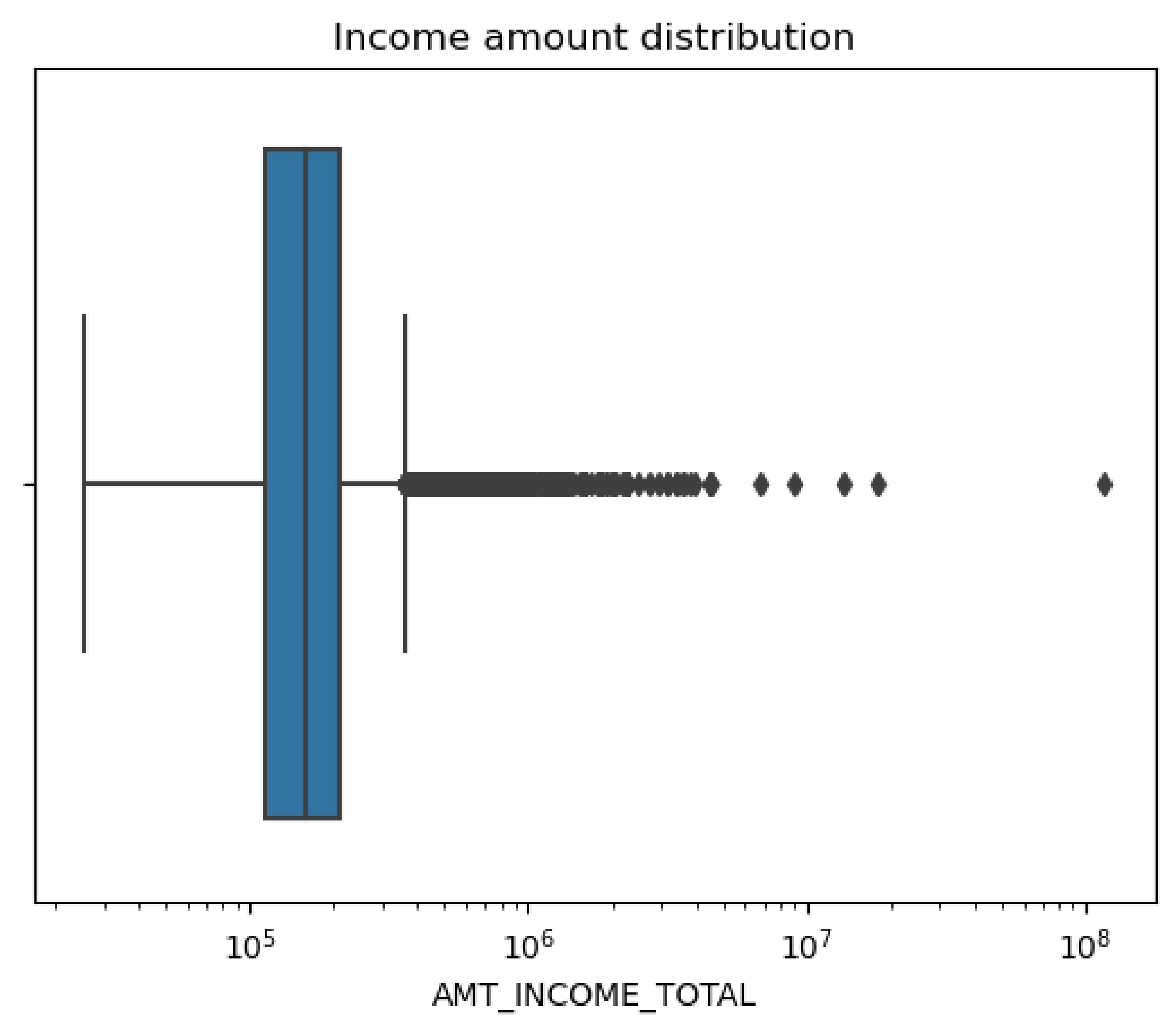


EDA APPROACH:

1. All necessary libraries were imported.
2. Both the data sets were load, Initially Application dataset was used for cleaning and analysis.
3. 'Application dataset' shape: (307511, 122); There are dtypes: float64(65), int64(41), object(16).
4. After checking for missing values unnecessary columns were dropped.[Columns with high missing percentage >=35%]
5. Remaining columns were imputed based on whether they were categorical or numerical, and these values were imputed with median or mode.
6. After initial cleaning the shape of the 'Application dataset':(307509, 25)
7. Many columns consisted of -ve values, these columns were corrected, with working on 'XNA'.
8. Numerical columns were analyzed for outliers, where either values were caped, or imputed.
9. Few continuous columns variables were put into bins.
10. Data imbalance was present i.e., TARGET column has 8.65% of 1's which means 8% clients have payment difficulties and 91.92% are having no difficulties.
11. Data was divided into two datasets : Target_1 and Target_0
12. Further analysis was performed-
    a. Univariate analysis: categorical and numeric
    b. Bivariate analysis: categorical and numeric
    c. Correlation matrix
13. Both the datasets were merged and further analysis were performed.
14. Ended the EDA process with Remarks.



HANDLING OUTLIERS: AMT_INCOME_TOTAL

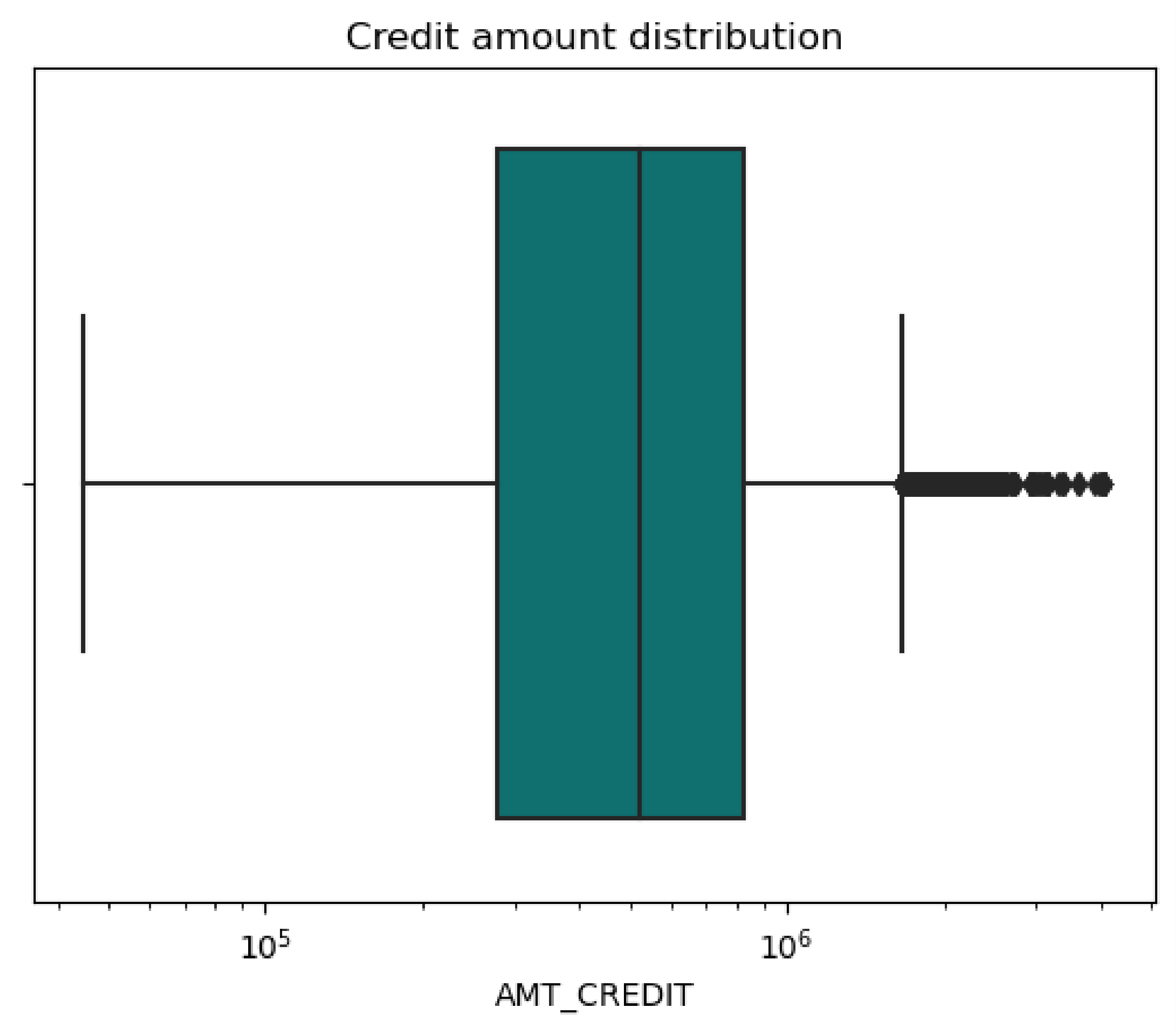


There are outliers present in this columns, might be due to the total income vary from person to person. The quantile ranges states that there are lot of difference between even 99th , 95th quantiles hence cause lot of outliers in the data set.

As the median value and max value have huge difference.

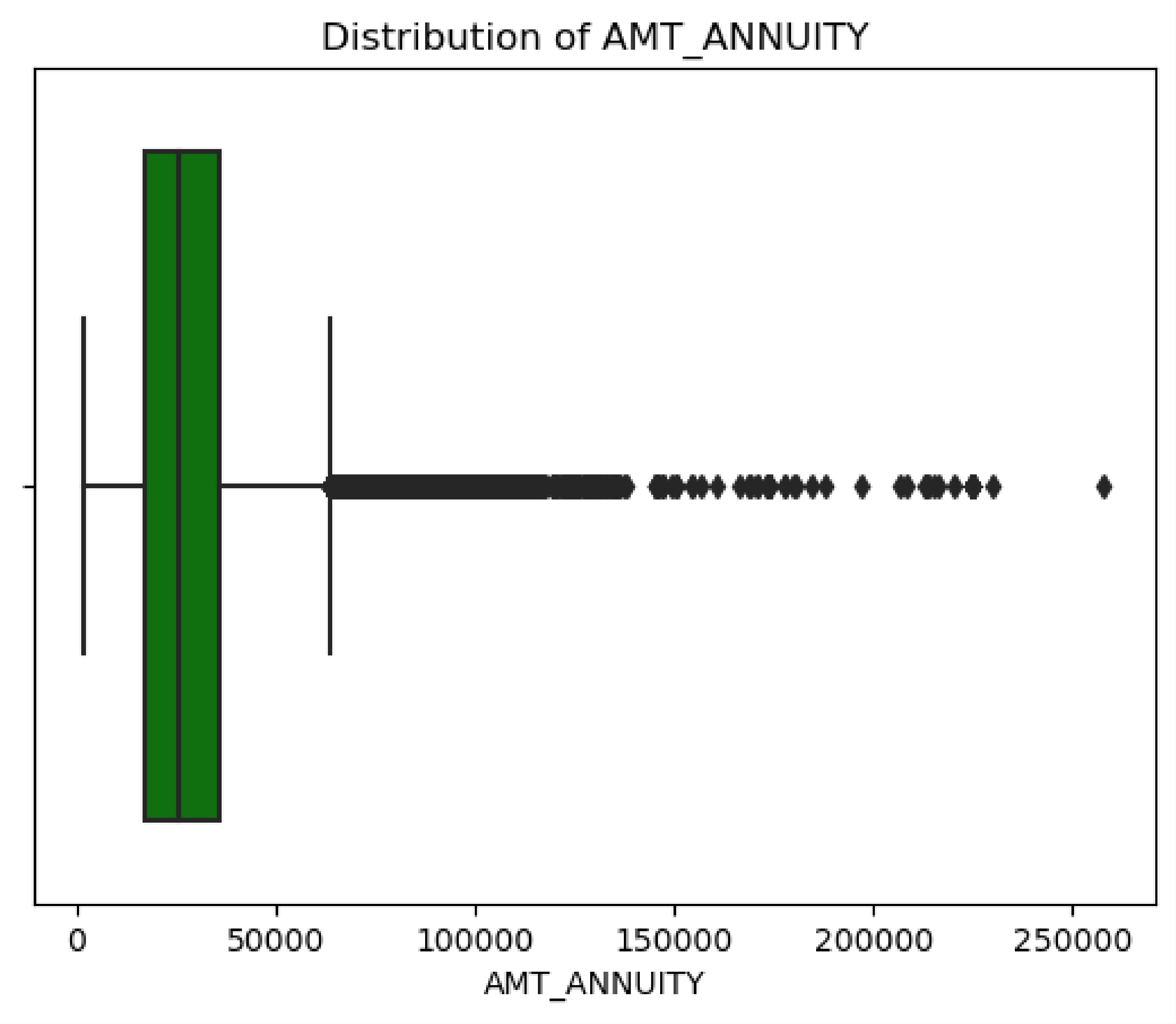
We can cap the outliers here.

HANDLING OUTLIERS: AMT_CREDIT



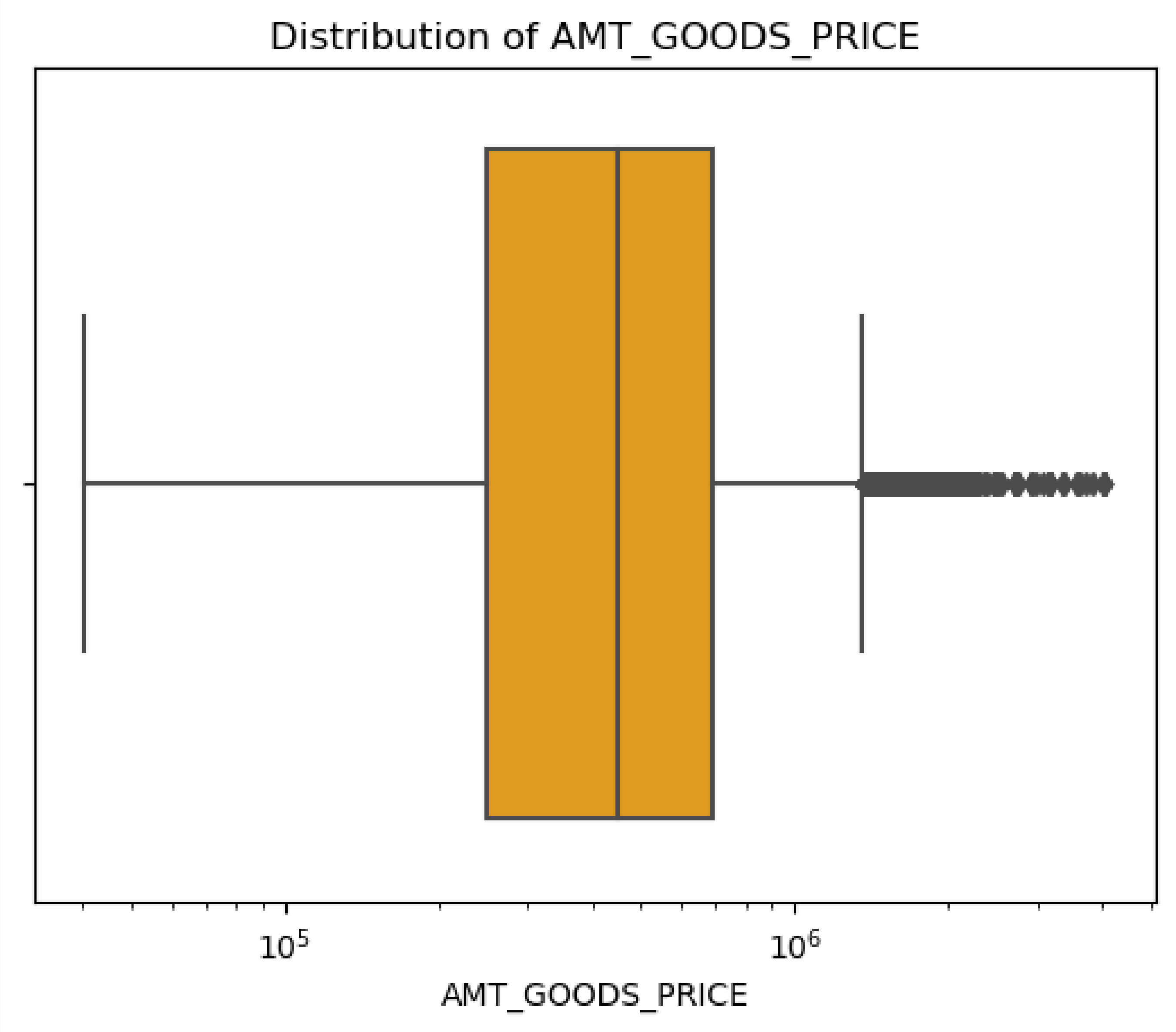
AMT_CREDIT columns has outliers, there is huge difference between 99th percentile and median which indicates the reason for outliers.
These outliers may be due to amount credited varied from customer to customers.

HANDLING OUTLIERS: AMT_ANNUITY



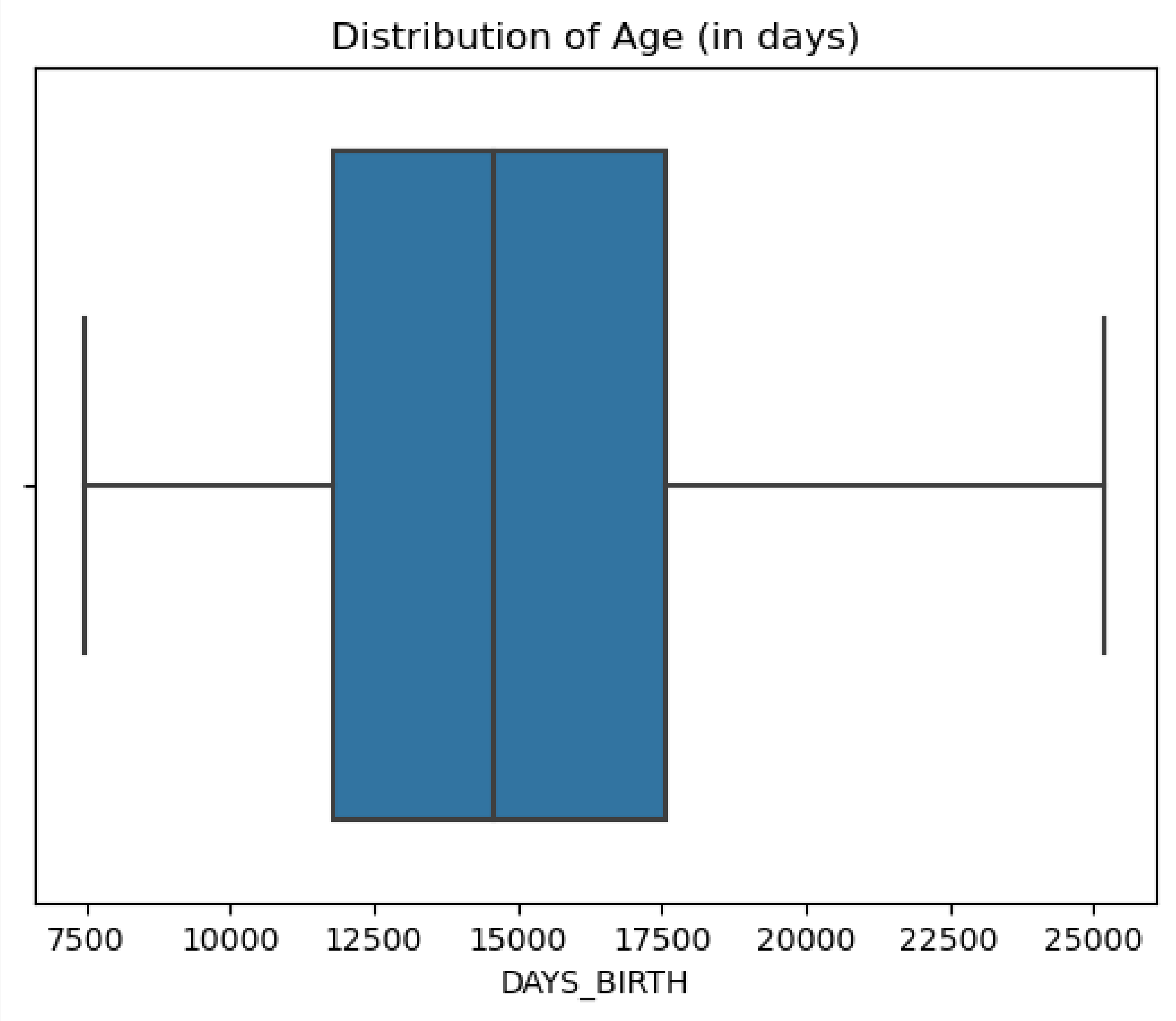
AMT_ANNUITY has outliers, but the due to not much difference in mean and median, we can either ignore or impute the outliers with median

HANDLING OUTLIERS: AMT_GOODS_PRICE



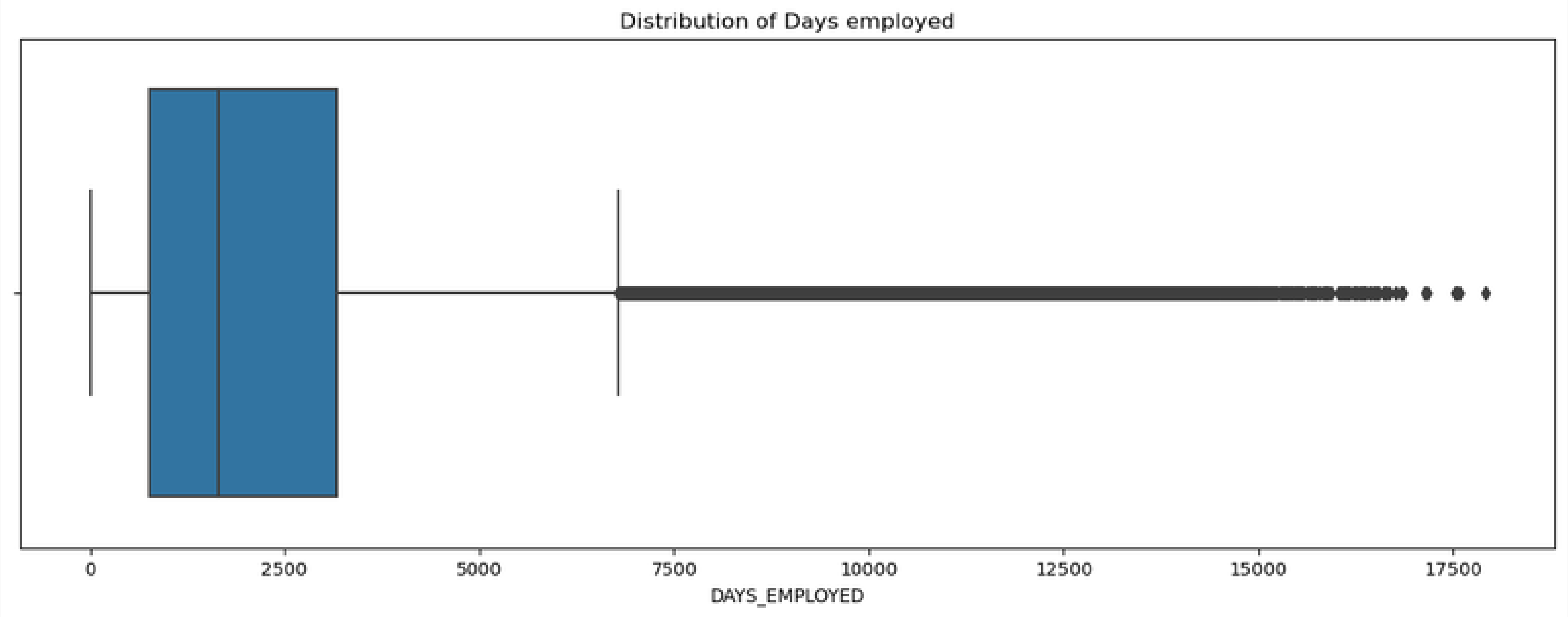
AMT_GOODS_PRICE values lies in certain range,also, has less outliers compartively.
As this column values indicates consumer loans it is the price of the goods for which the loan is given, hence indicating that the loan varied from customer to customer.

HANDLING OUTLIERS: DAYS_BIRTH



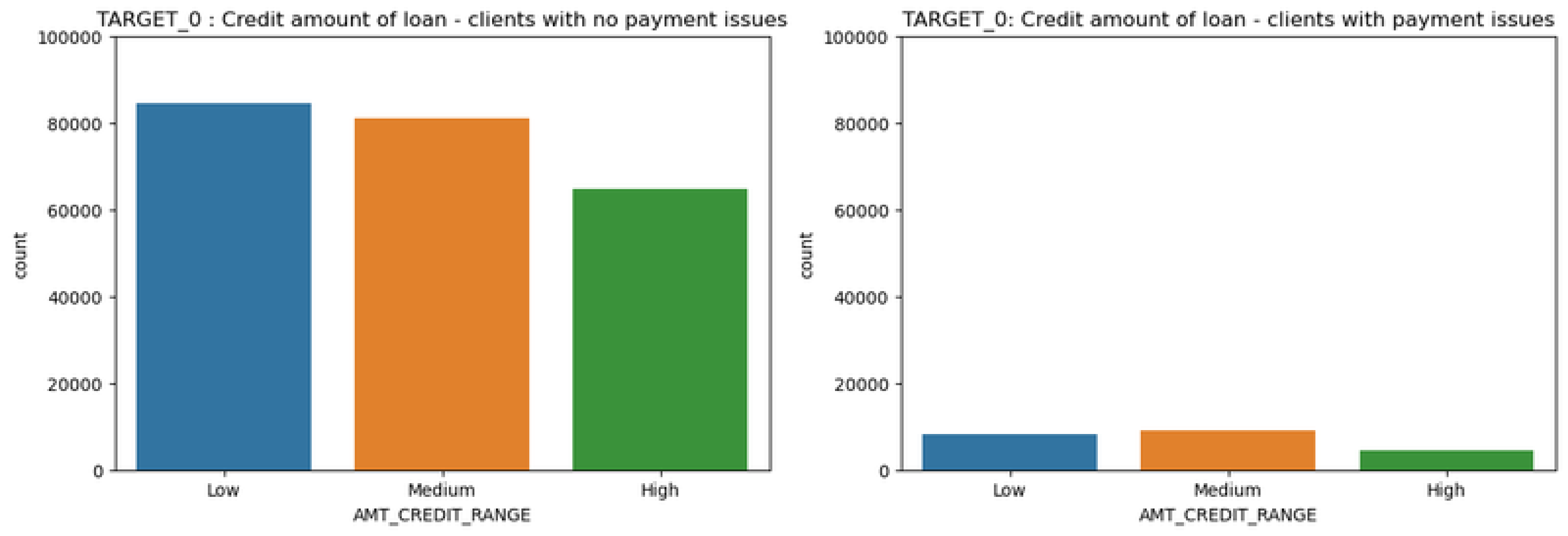
DAYS_BIRTH column -the mean and median doesn't have much difference, also there are no outliers present

HANDLING OUTLIERS: DAYS_EMPLOYED



There are very few outliers, The values differes from person to person.

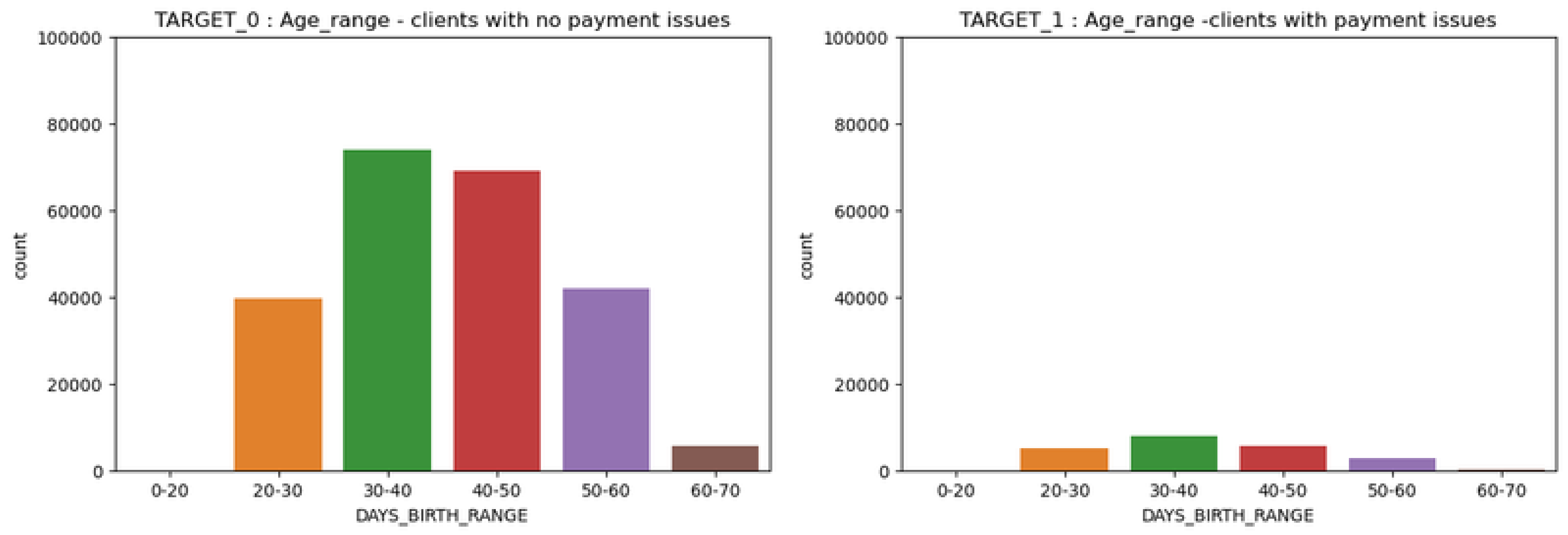
NUMERICAL UNIVARIATE ANALYSIS CREDIT_AMOUNT RANGE



Clients with low and high credit are most likely to make payments comparatively to clients who fall in medium range group. whereas the clients in the category with payment issues have almost the same response irrespective of credit amount.

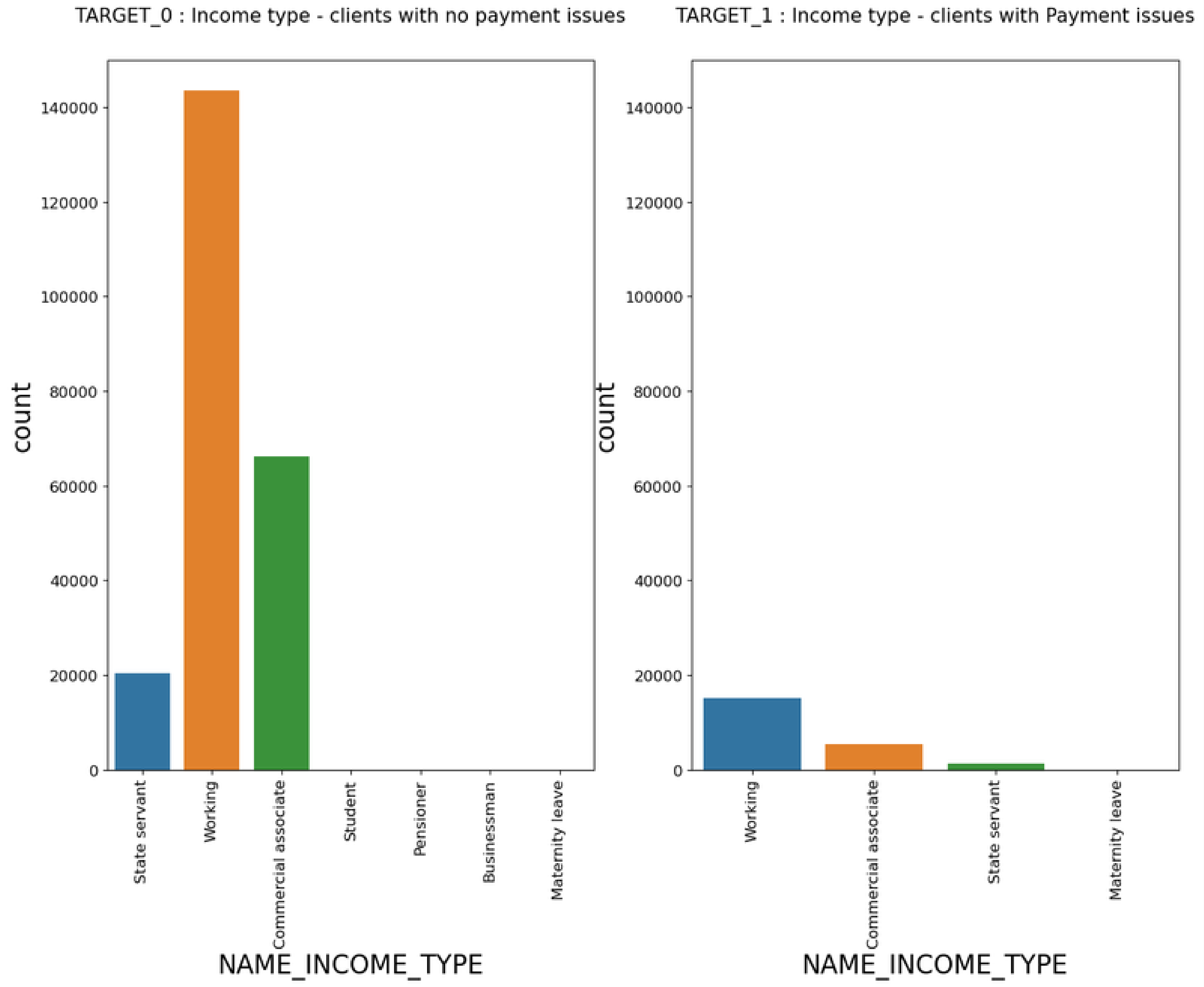
NUMERICAL UNIVARIATE ANALYSIS

AGE WITH RESPECTIVE TO PAYMENT ISSUES



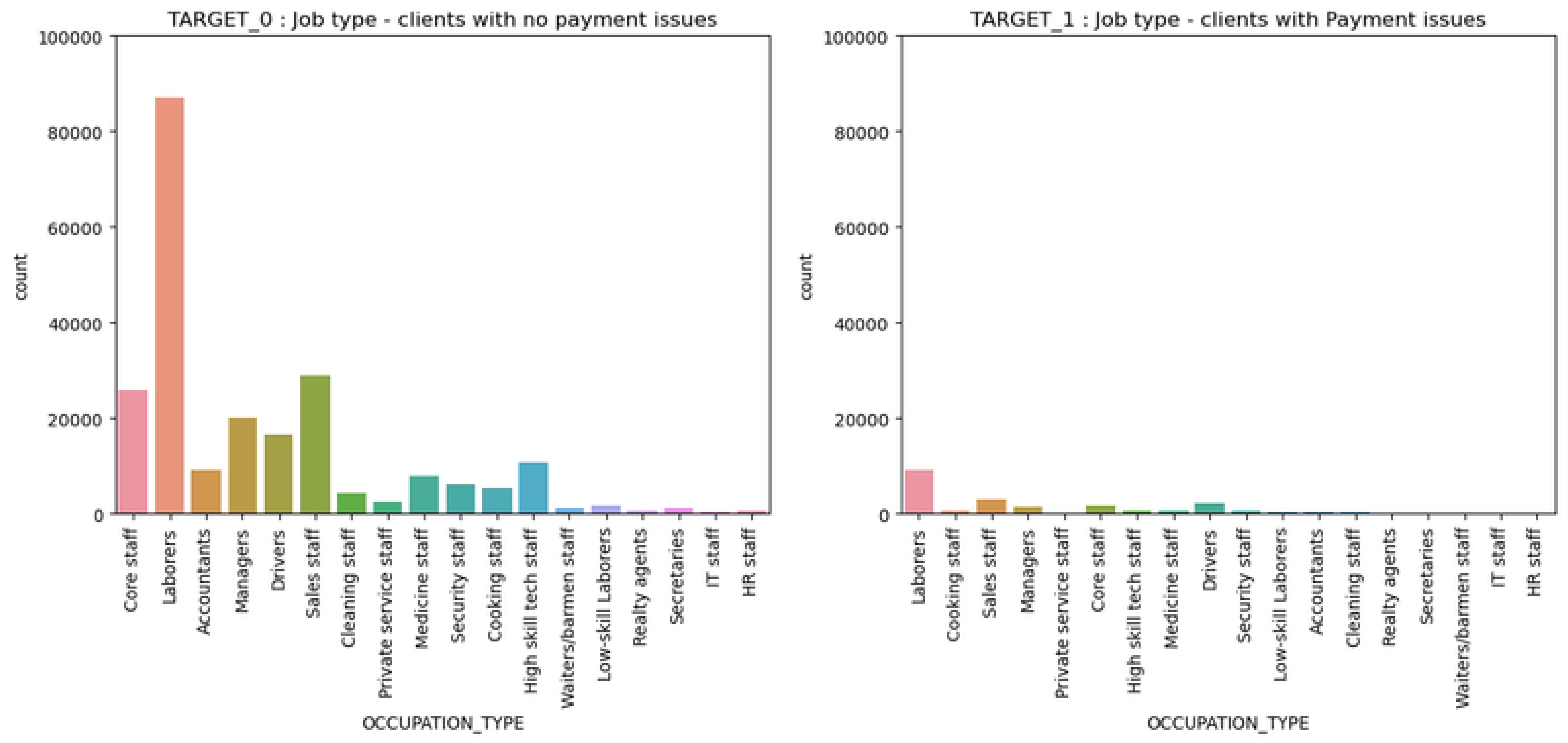
Clients with age range 30-50 are most likely to make payments on time comparatively to clients who fall in other age groups whereas the clients age range 30-50 with payment issues also might be the one who doesn't make payment on time.
With this 20-60 age can be considered the age group to lend loans, but banks should also take other categories also in consideration.

CATEGORICAL UNIVARIATE VARIABLE ANALYSIS ON NAME INCOME TYPE



Working and commericial associate professional make payments on time comparetively to clients who fall in other categories.

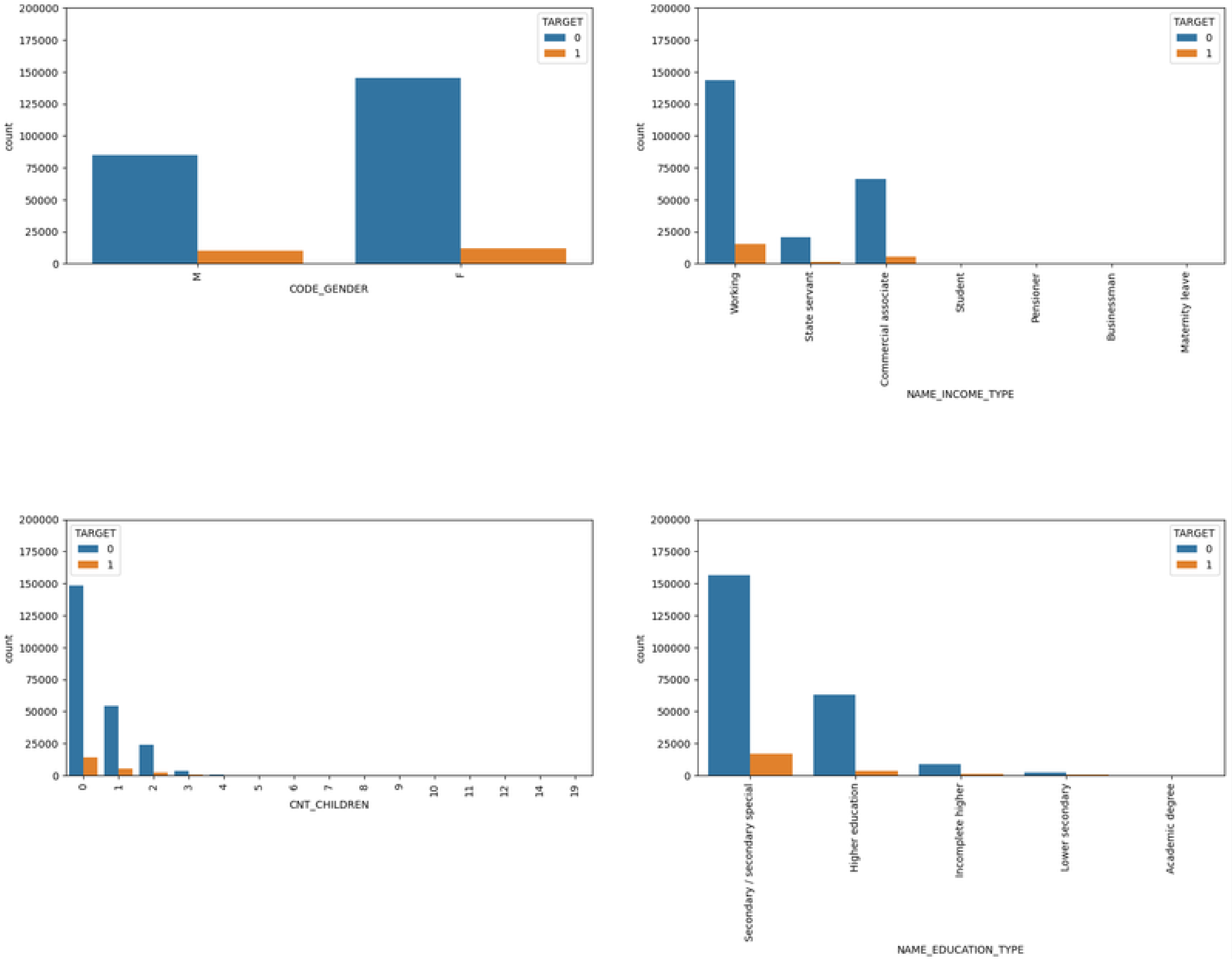
CATEGORICAL UNIVARIATE VARIABLE ANALYSIS ON OCCUPATION TYPE



Labourers make payments on time comparetively to clients who fall in other categories.Whereas HR staff are complete opposite and delay in payments

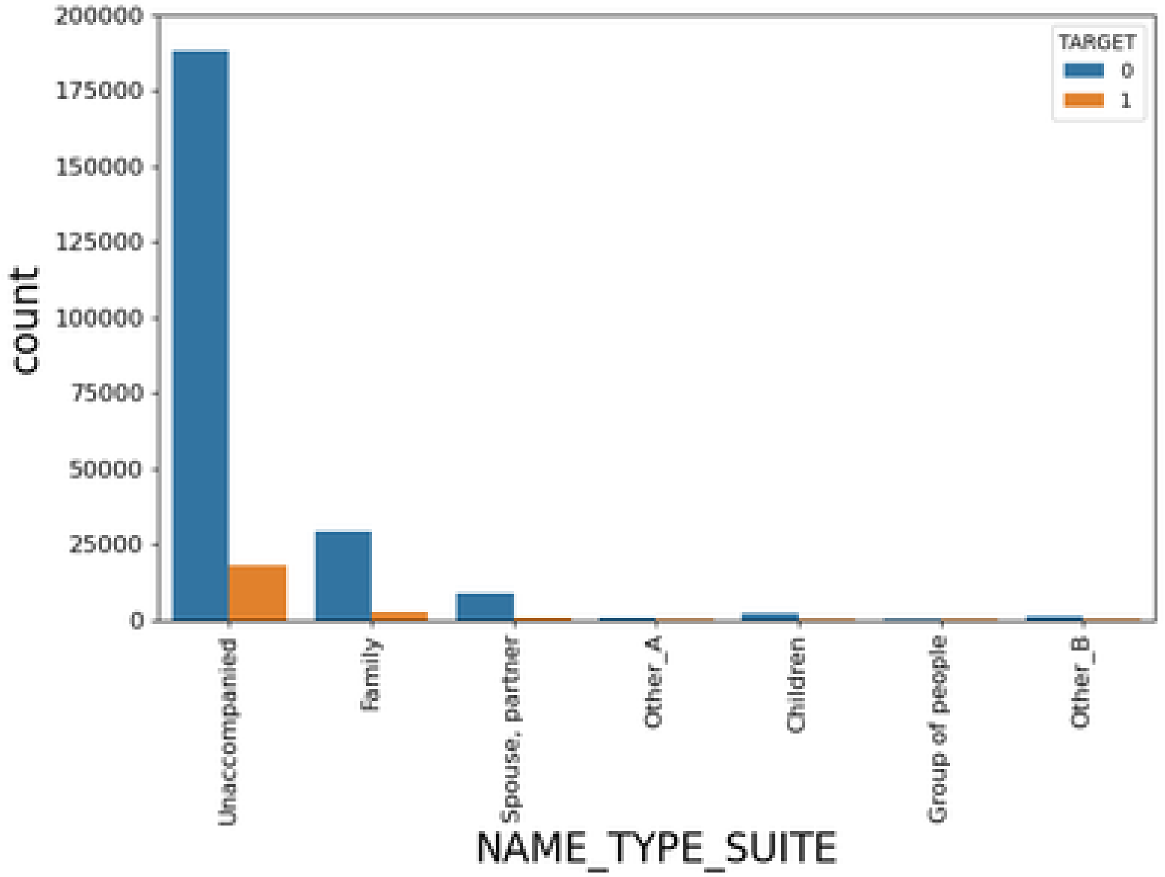
CATEGORICAL ANALYSIS

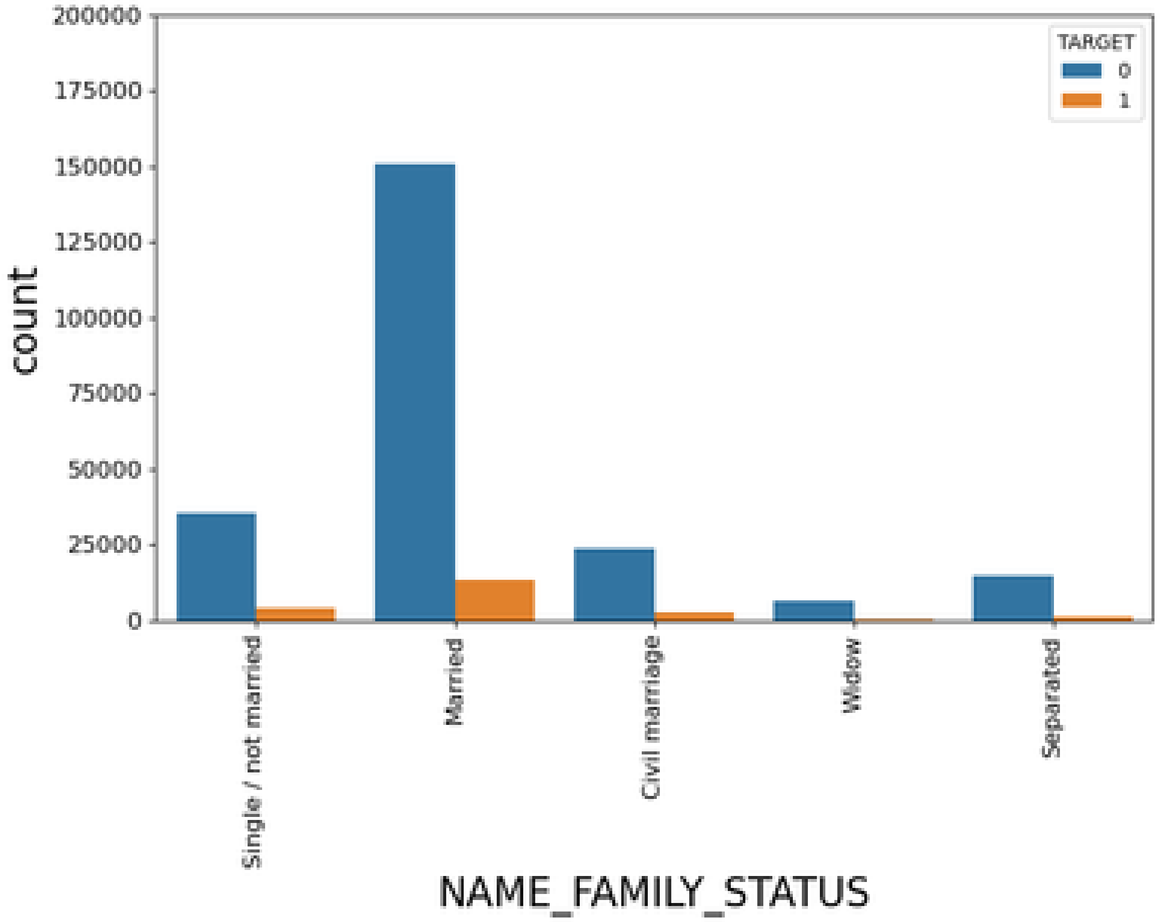
```
COLUMNS_1 = ['CODE_GENDER',
 'NAME_INCOME_TYPE',
'CNT_CHILDREN',
'NAME_EDUCATION_TYPE']
```

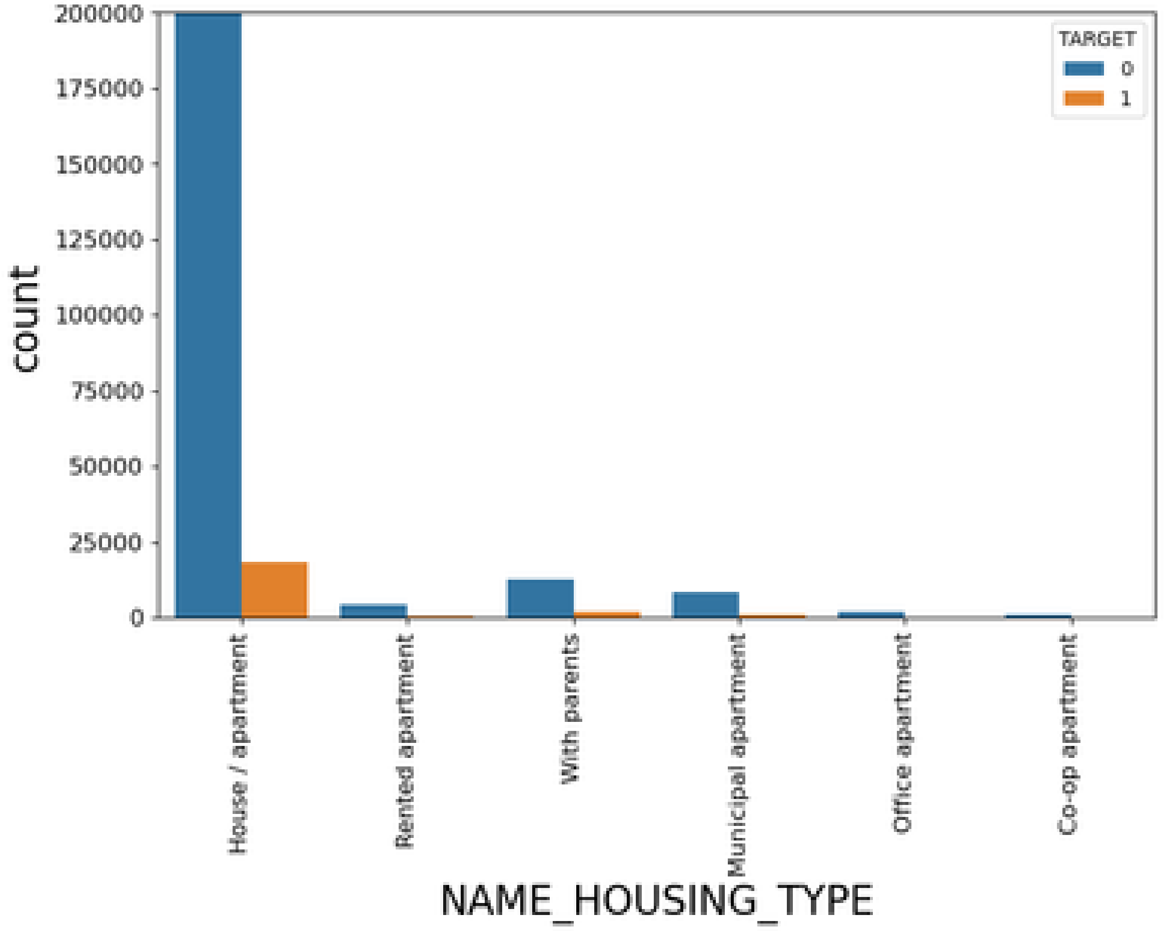


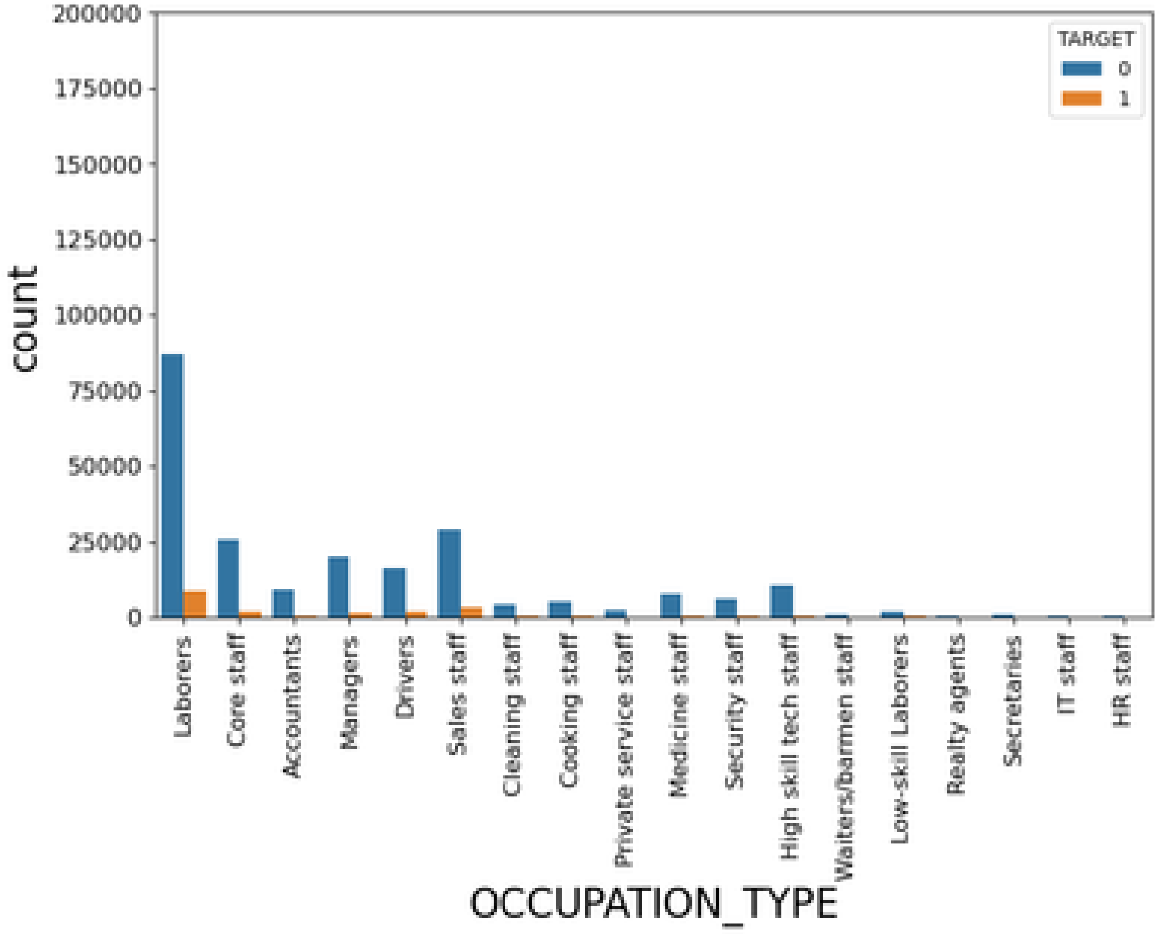
CATEGORICAL ANALYSIS

COLUMNS_2 =['NAME_TYPE_SUITE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE']









1.['CODE_GENDER', 'NAME_INCOME_TYPE', 'CNT_CHILDREN', 'NAME_EDUCATION_TYPE']

CODE_GENDER: Female clients pay loan faster comparatively to men.

NAME_INCOME_TYPE: Working clients should be given preference to lend loans comparatively to other clients.

CNT_CHILDREN: Clients with no children, or single clients pay loan faster than clients with kids.

NAME_EDUCATION_TYPE: Clients with higher education, i.e. with secondary educations should be preferred.

2.['NAME_TYPE_SUITE','NAME_FAMILY_STATUS','NAME_HOUSING_TYPE','OCCUPATION_TYPE']

NAME_TYPE_SUITE: Clients who were unaccompanied were the one who payed loan faster then other groups.

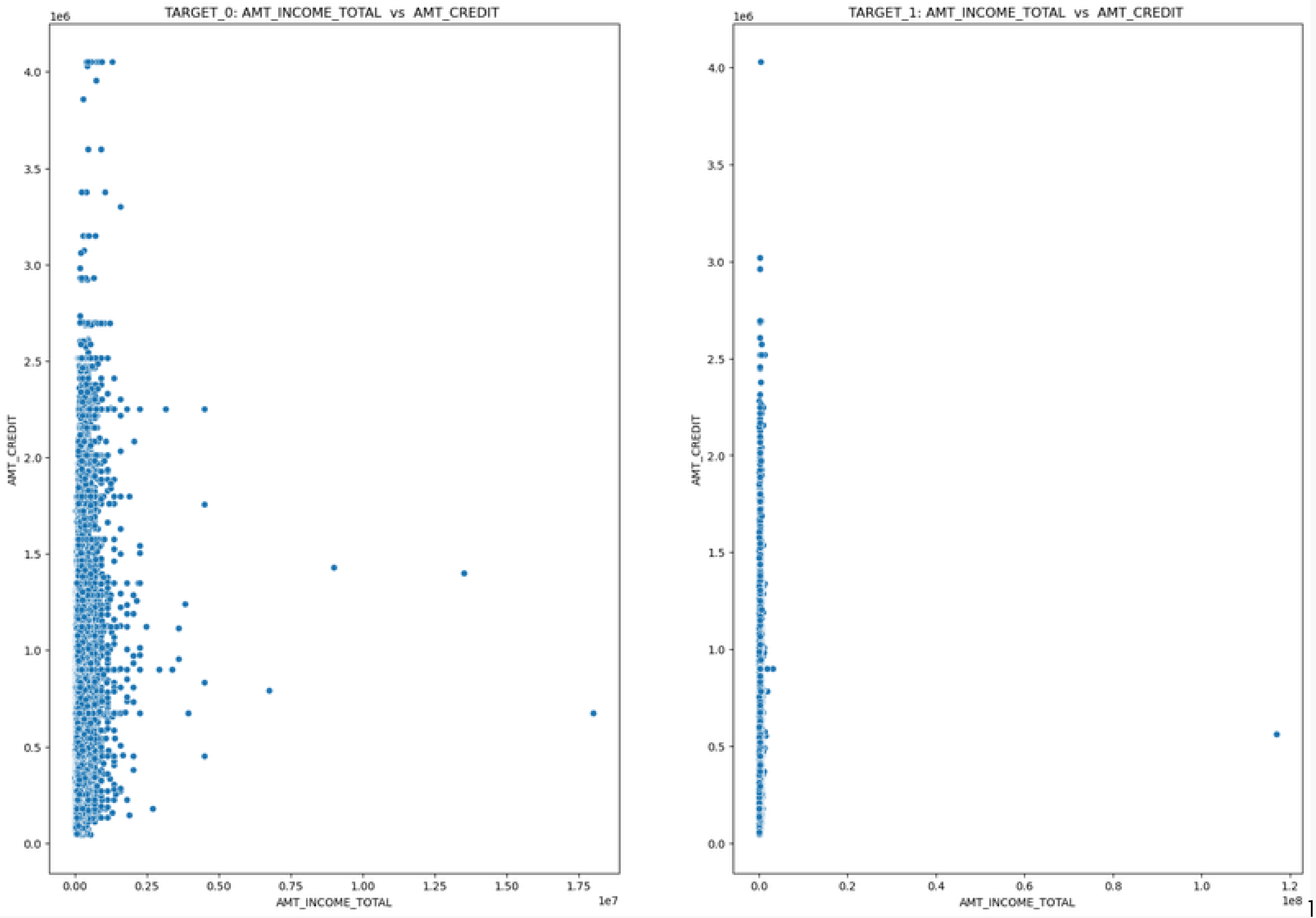
NAME_FAMILY_STATUS: Clients , especially Married had payed loan on time.

NAME_HOUSING_TYPE: Clients with house/apartment had payed loan on time whereas the co-op apartment were the opposite

OCCUPATION_TYPE: Laborers have payed loan on time comparatively to other groups whereas HR staff delayed on loan payment.

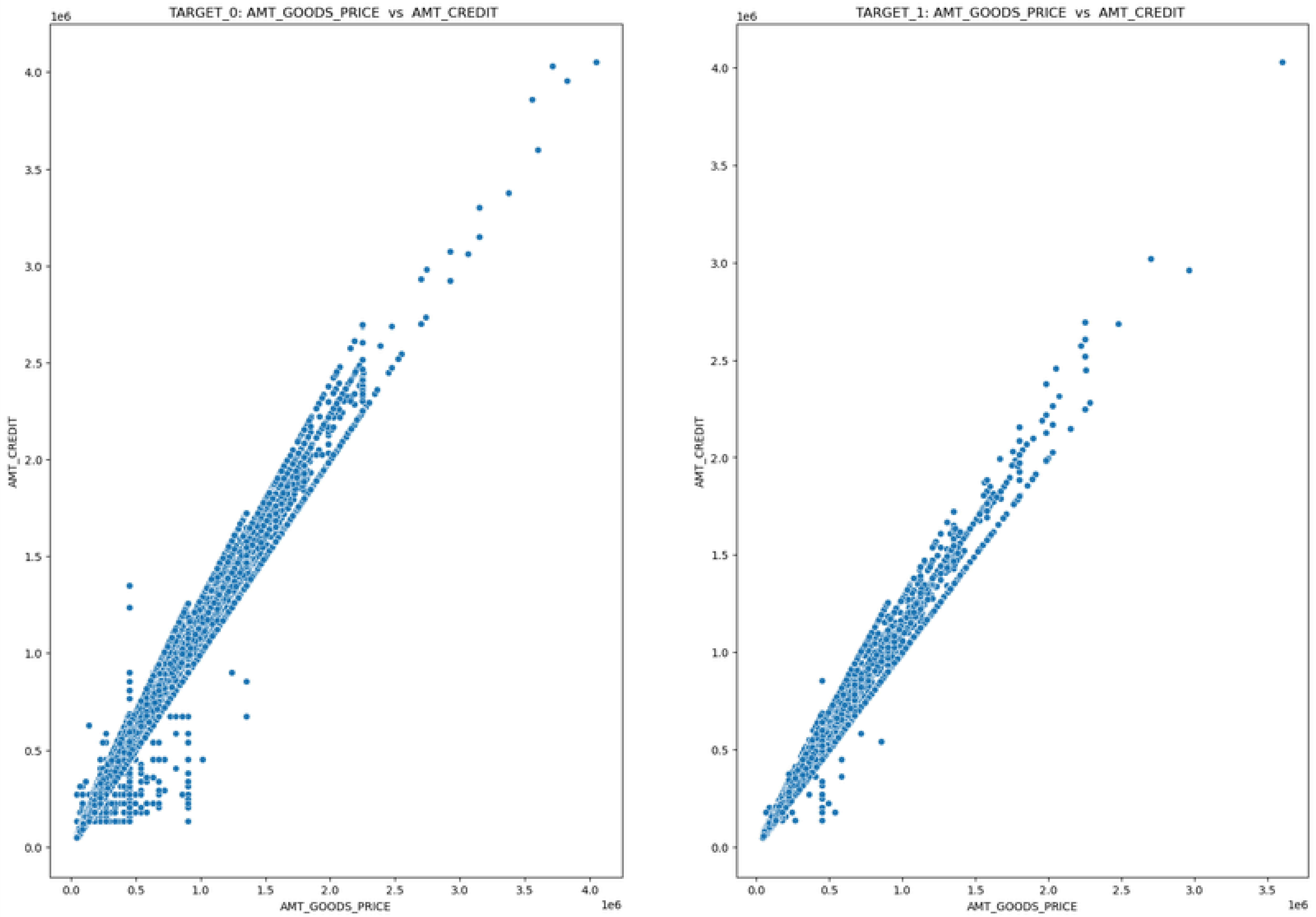
BIVARIATE ANALYSIS

INCOME VS CREDIT



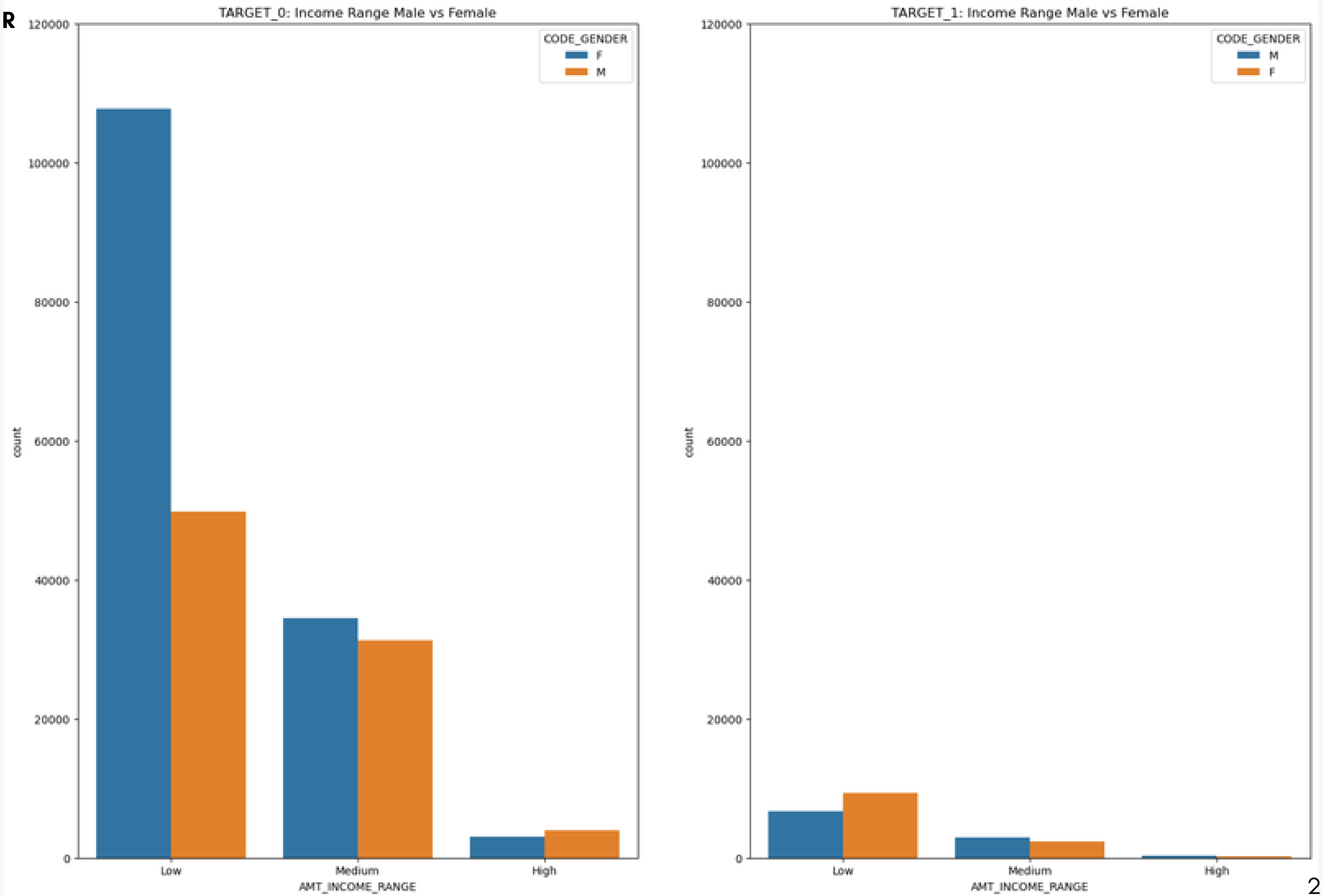
BIVARIATE ANALYSIS

GOODS PRICE VS CREDIT



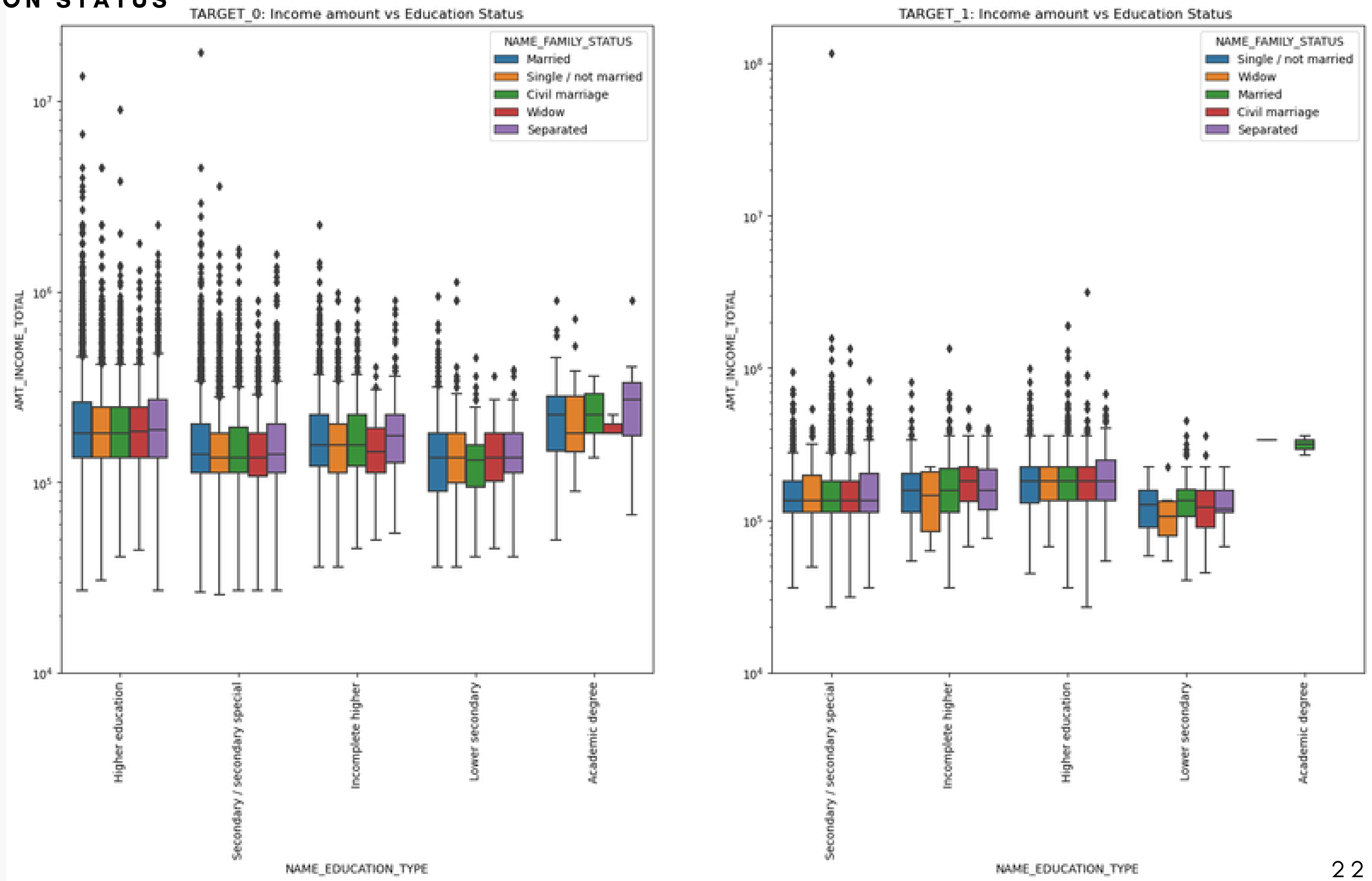
BIVARIATE ANALYSIS

INCOME RANGE- GENDER



BIVARIATE ANALYSIS

INCOME VS EDUCATION STATUS



BIVARIATE ANALYSIS

1.Income vs Credit : The income of a client plays an huge role on the amount credited to the clients, as they tend to do timely payments.

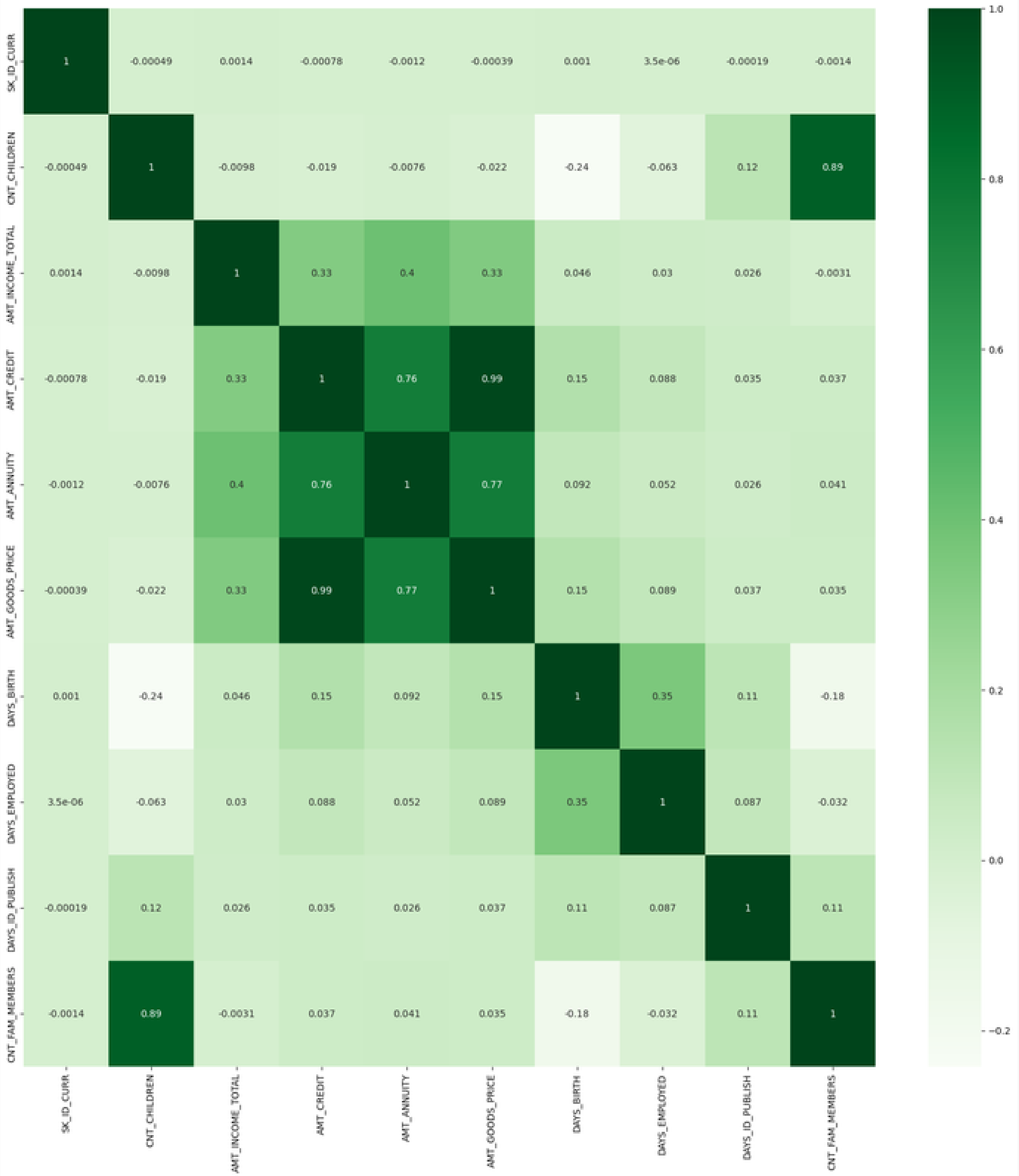
2.Goods price vs Credit : Also the goods price and the crdit seems to have a trend, and the clients have higher credits who was able to pay on time verses the deflauters groups.

3.Income range- Gender: Females in the low income range don't have payment issues.

4.Credit amount vs Education Status : highly educated, married person are having credits higher than those who have done lower secondary eduction, they tend to make payments on time comparatively to others. (More outliers are also seen in this category)

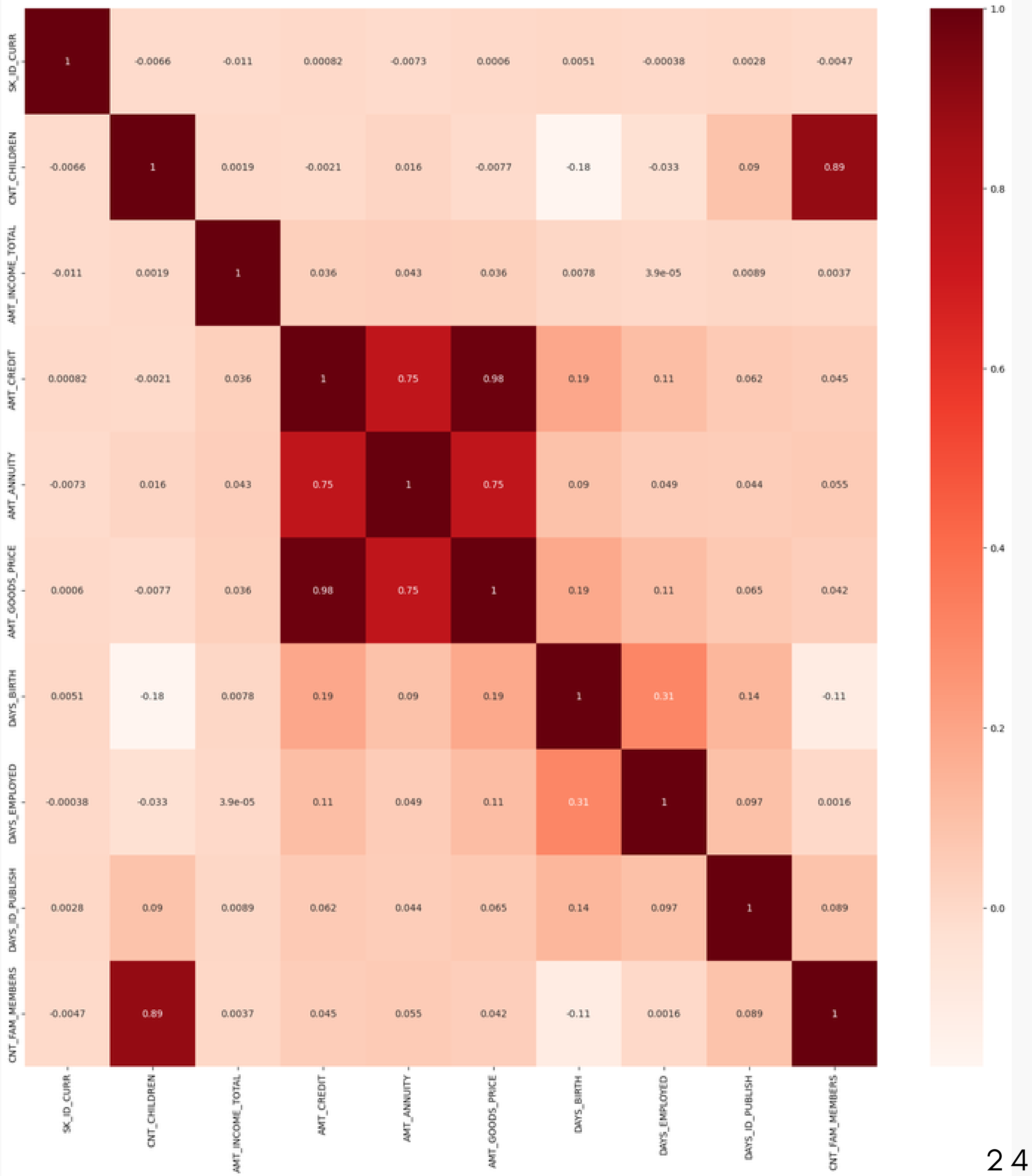
5.Income vs Education Status : Higher the eduation, higer the income, hence no difficuties for loan payments

CORRELATION MATRIX FOR TARGET 0 APPLICATION DATA



	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
SK_ID_CURR	1	-0.00049	0.0014	-0.00078	-0.0012	-0.00039	0.001	3.5e-06	-0.00019	-0.0014
CNT_CHILDREN	-0.00049	1	-0.0098	-0.019	-0.0076	-0.022	-0.24	-0.063	0.12	0.89
AMT_INCOME_TOTAL	0.0014	-0.0098	1	0.33	0.4	0.33	0.046	0.03	0.026	-0.0031
AMT_CREDIT	-0.00078	-0.019	0.33	1	0.76	0.99	0.15	0.088	0.035	0.037
AMT_ANNUITY	-0.0012	-0.0076	0.4	0.76	1	0.77	0.092	0.052	0.026	0.041
AMT_GOODS_PRICE	-0.00039	-0.022	0.33	0.99	0.77	1	0.15	0.089	0.037	0.035
DAYS_BIRTH	0.001	-0.24	0.046	0.15	0.092	0.15	1	0.35	0.11	-0.18
DAYS_EMPLOYED	3.5e-06	-0.063	0.03	0.088	0.052	0.089	0.35	1	0.087	-0.032
DAYS_ID_PUBLISH	-0.00019	0.12	0.026	0.035	0.026	0.037	0.11	0.087	1	0.11
CNT_FAM_MEMBERS	-0.0014	0.89	-0.0031	0.037	0.041	0.035	-0.18	-0.032	0.11	1

CORRELATION MATRIX FOR TARGET 1 APPLICATION DATA



	SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS
SK_ID_CURR	1	-0.0066	-0.011	0.00082	-0.0073	0.0006	0.0051	-0.00038	0.0028	-0.0047
CNT_CHILDREN	-0.0066	1	0.0019	-0.0021	0.016	-0.0077	-0.18	-0.033	0.09	0.89
AMT_INCOME_TOTAL	-0.011	0.0019	1	0.036	0.043	0.036	0.0078	3.9e-05	0.0089	0.0037
AMT_CREDIT	0.00082	-0.0021	0.036	1	0.75	0.98	0.19	0.11	0.062	0.045
AMT_ANNUITY	-0.0073	0.016	0.043	0.75	1	0.75	0.09	0.049	0.044	0.055
AMT_GOODS_PRICE	0.0006	-0.0077	0.036	0.98	0.75	1	0.19	0.11	0.065	0.042
DAYS_BIRTH	0.0051	-0.18	0.0078	0.19	0.09	0.19	1	0.31	0.14	-0.11
DAYS_EMPLOYED	-0.00038	-0.033	3.9e-05	0.11	0.049	0.11	0.31	1	0.097	0.0016
DAYS_ID_PUBLISH	0.0028	0.09	0.0089	0.062	0.044	0.065	0.14	0.097	1	0.089
CNT_FAM_MEMBERS	-0.0047	0.89	0.0037	0.045	0.055	0.042	-0.11	0.0016	0.089	1

CORRELATION MATRIX

TARGET_0

The columns which are highly correlated with Non-Defaulters are

'CNT_CHILDREN' and 'CNT_FAM_MEMBERS', 'AMT_CREDIT' and 'AMT_GOODS_PRICE','AMT_CREDIT' and 'AMT_ANNUITY','AMT_ANNUITY'and 'AMT_GOODS_PRICE','AMT_INCOME_TOTAL', hence these parameters can be used to determined for non-defaulters

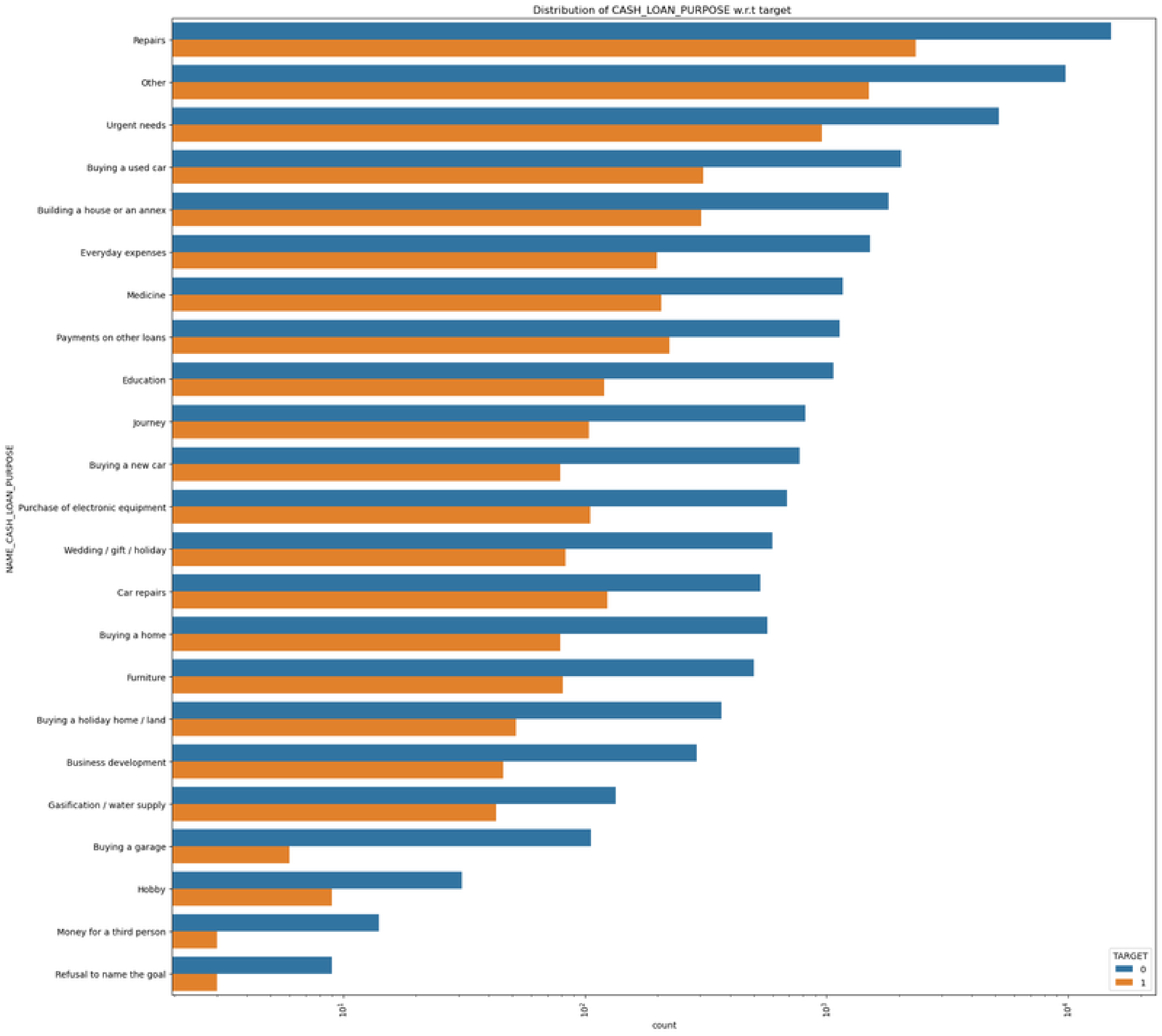
TARGET_1

The columns which are highly correlated with Defaulters are also similar to non-defaulters.

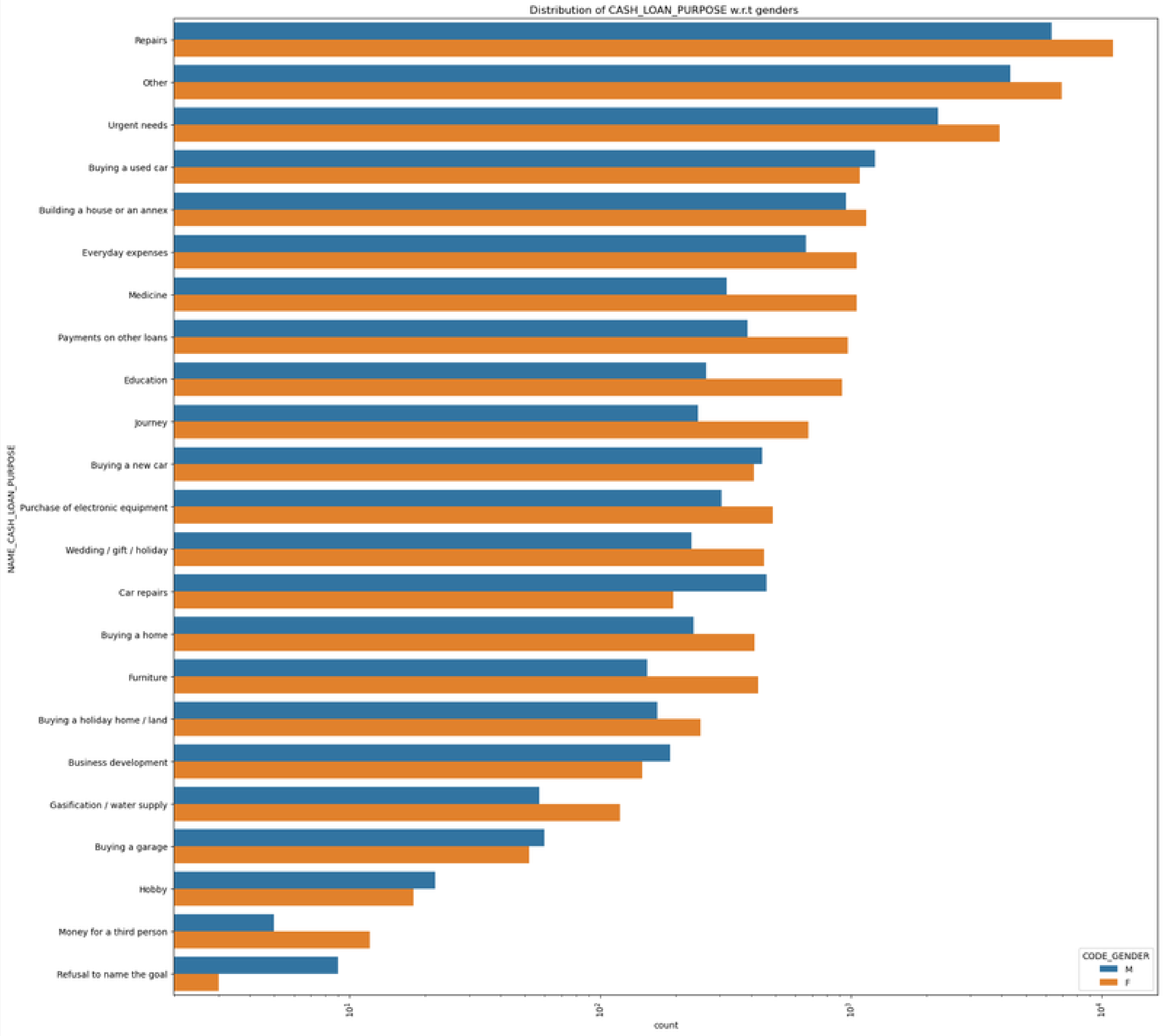
'CNT_CHILDREN' and 'CNT_FAM_MEMBERS', 'AMT_CREDIT' and 'AMT_GOODS_PRICE','AMT_CREDIT' and 'AMT_ANNUITY','AMT_ANNUITY' and 'AMT_GOODS_PRICE','AMT_INCOME_TOTAL', hence these parameters can be used to determined for defaulters as well.

UNIVARIATE ANALYSIS

DISTRIBUTION OF CONTRACT STATUS ACROSS TARGET



DISTRIBUTION OF CONTRACT STATUS ACROSS GENDERS



UNIVARIATE ANALYSIS

1.Distribution of CASH_LOAN_PURPOSE w.r.t target: The loans were more credited for Repair purpose

2.Distribution of CASH_LOAN_PURPOSE w.r.t genders: The loans were more prefered to female clients

3.Distribution of CASH_LOAN_PURPOSE w.r.t CONTRACT_STATUS:

a. Repairs had all three categories

b. Most approved loans were for Repairs and least for refusal

c. Most refused loans were for Repairs and least for money for a third person.

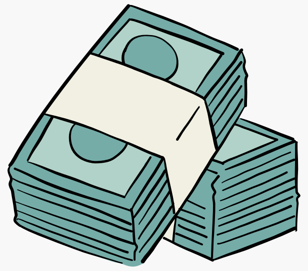
d. Most cancelled loans were for Repairs and least for money for refusal.

e. Education loans were approved and cancelled at the same amount of count

4.Distribution of CASH_LOAN_PURPOSE w.r.t NAME_INCOME_TYPE: Working professionals should be preferred more than students & pensioners

5.Distribution of CASH_LOAN_PURPOSE w.r.t NAME_EDUCATION_TYPE:Higher education should be preferred more than Academic degree

REMARKS:



1. Company should lend loans to clients by checking out mutiple parameters, to escape from loss from lending money to defaulters and not lending money to valid clients.
2. Female working clients can be given more preferences over Male as they tend to make loan payment on time.
3. According to given datasets, more loans were given for Repair Purpose, these category also falls on loan payment on time.
4. Clients who want loans for a third person , or refusal to name of the goal can be ignored or given loans with higher interest rates or less amount
5. Credit amount, goods price are related, where the higher the goods price, the credit amount also increases.

6.. Working clients, with higher income, and higher education can be categorized as non defaulters.

7. Company should be very careful while lending loans , because if certain parameters favors Non- defaulters , they might also turn up to be defaulters, hence previous_application data should be cross verified before lending any old clients with loan. For new clients they can prefer either to increase the interest rate or reduce the loan amount.

Thank You

IPSITA PAL

upGrad

upGrad & IIITB | Data Science Program - December 2022

BATCH ID 3459

DS51