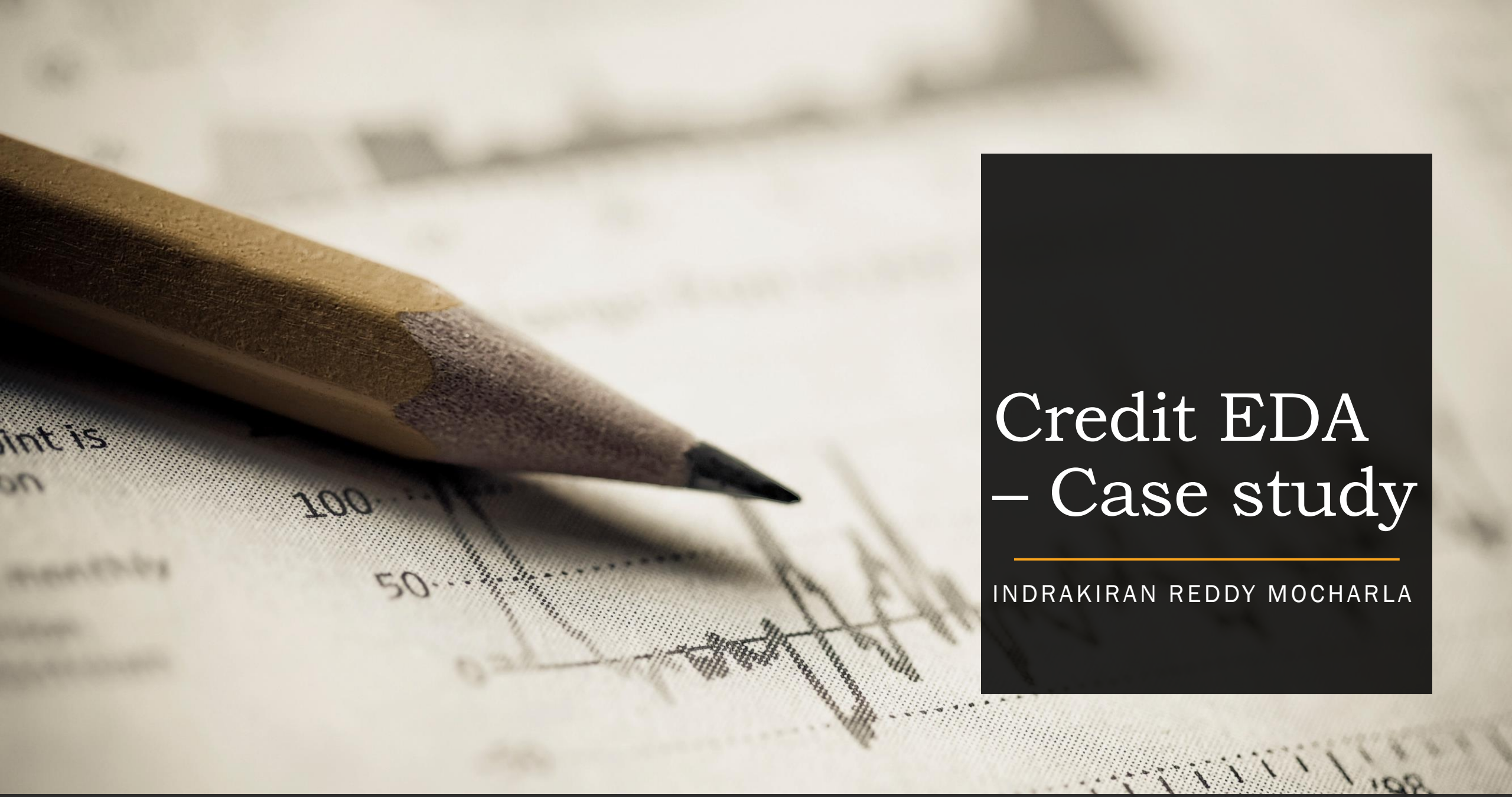


Credit EDA – Case study

INDRAKIRAN REDDY MOCHARLA

Table of contents

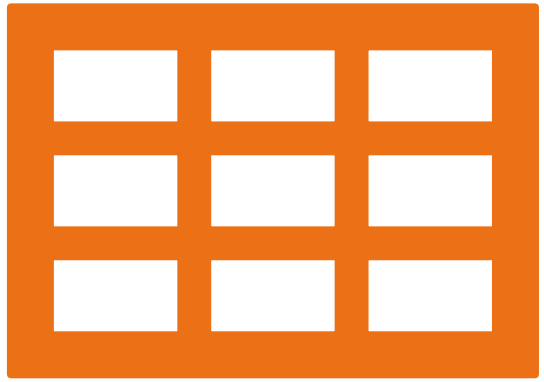
Problem statement - I

- This case study aims to identify patterns which indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. This will ensure that the consumers capable of repaying the loan are not rejected. Identification of such applicants using EDA is the aim of this case study.
- In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.
- To develop your understanding of the domain, you are advised to independently research a little about risk analytics - understanding the types of variables and their significance should be enough).

Problem statement - II

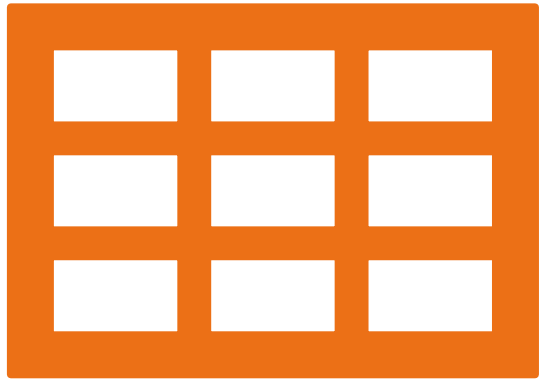
- Present the overall approach of the analysis in a presentation. Mention the problem statement and the analysis approach briefly.
- Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)
- Identify if there are outliers in the dataset. Also, mention why do you think it is an outlier. Again, remember that for this exercise, it is not necessary to remove any data points.
- Identify if there is data imbalance in the data. Find the ratio of data imbalance.
- Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.
- Find the top 10 correlation for the **Client with payment difficulties** and **all other cases** (Target variable). Note that you have to find the top correlation by segmenting the data frame w.r.t to the target variable and then find the top correlation for each of the segmented data and find if any insight is there. Say, there are 5+1(target) variables in a dataset: **Var1, Var2, Var3, Var4, Var5, Target**. And if you have to find top 3 correlation, it can be: Var1 & Var2, Var2 & Var3, Var1 & Var3. Target variable will not feature in this correlation as it is a categorical variable and not a continuous variable which is increasing or decreasing.
- Include visualisations and summarise the most important results in the presentation. You are free to choose the graphs which explain the numerical/categorical variables. Insights should explain why the variable is important for differentiating the **clients with payment difficulties with all other cases.**

Application Dataset

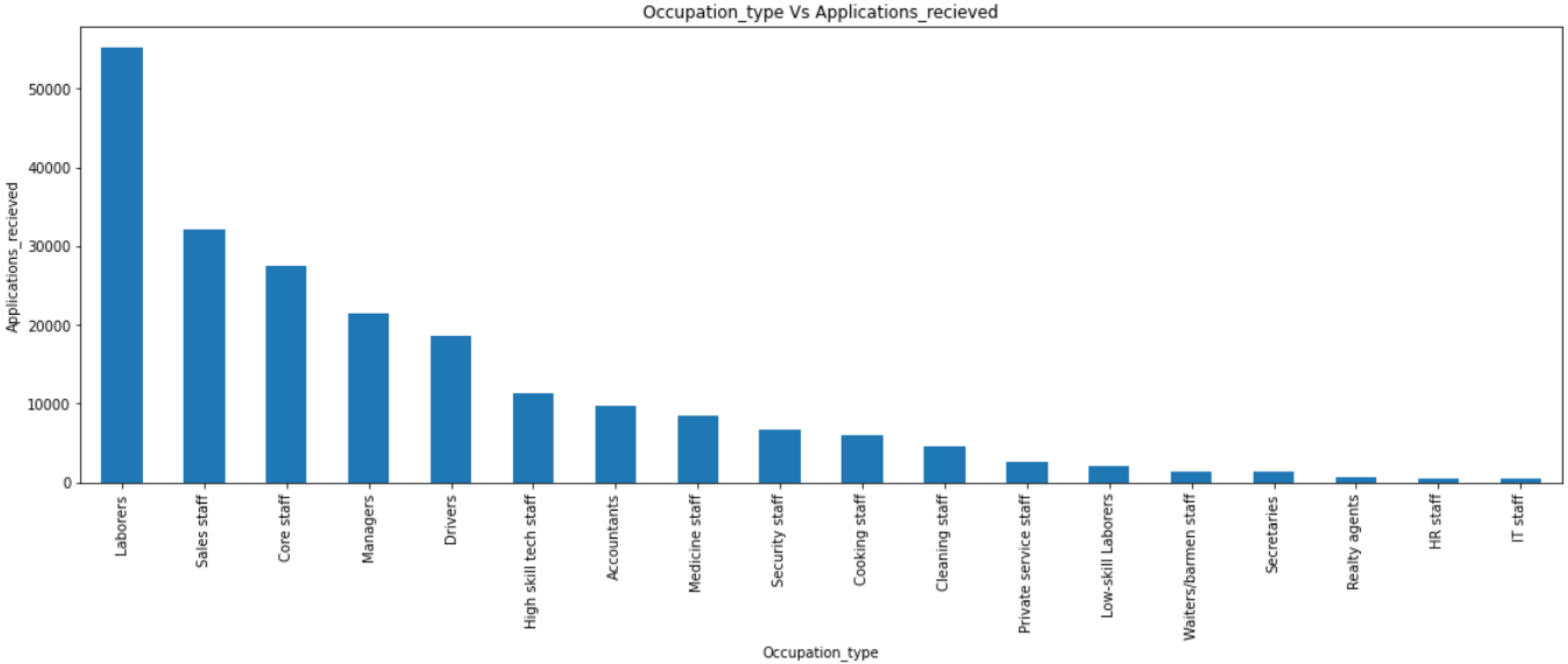


- Verified the dimensions of dataset(i.e, no.of rows and columns)
- Dropping the columns having null values greater than 40%.
- Also dropping insignificant columns which are not useful in decision making
- Imputed missing values of some useful columns with median value.
- Done Outlier analysis for both categorical variables and continuous variables.
- Computed Imbalance percentage

Application Dataset



- Divided the dataset is divided into two different datasets
  - Target 1 - Client facing some payment difficulties
  - Target 0 - all other clients
- Created bins for amount income total and credit amount columns
- Done Univariate analysis separately for both Target 1 dataset and Target 0 dataset
- Done Bivariate analysis separately for both Target 1 dataset and Target 0 dataset
- Correlation between the numerical variables

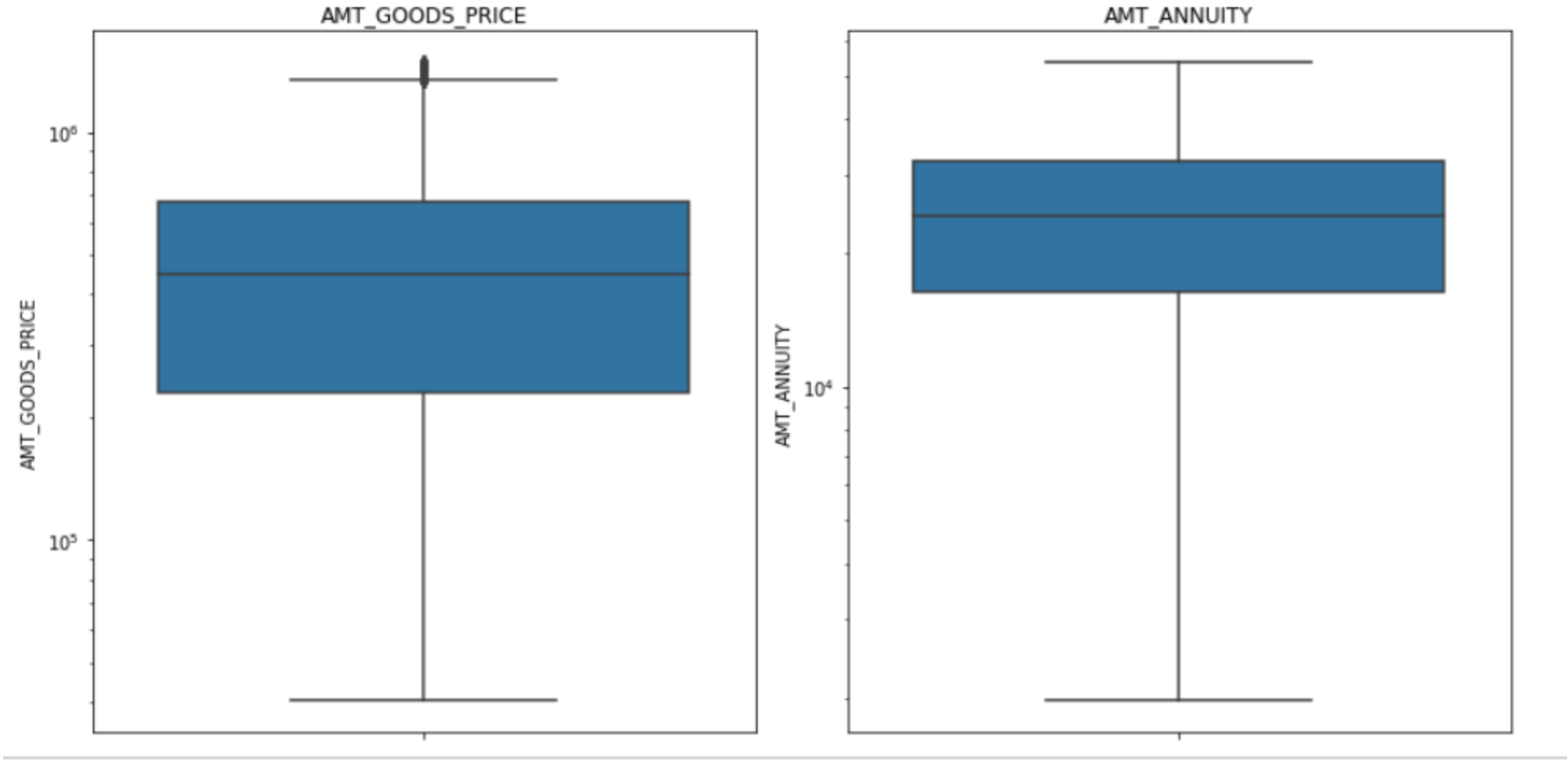


Inference:

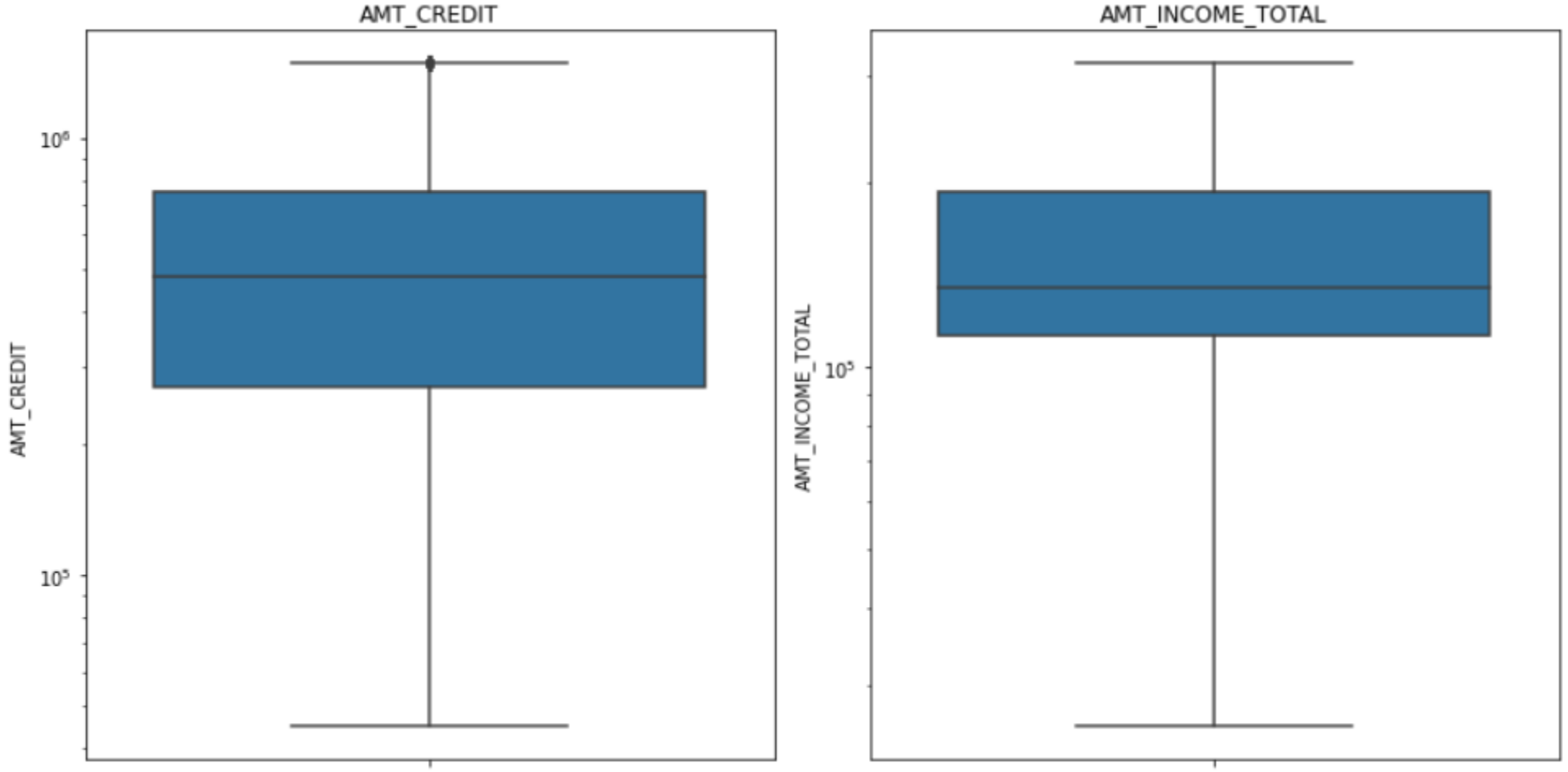
- Low-cadre jobs like Laborers, sales staff and core staff etc, have applied more when compared to High-cadre jobs

Outlier analysis

Numerical columns

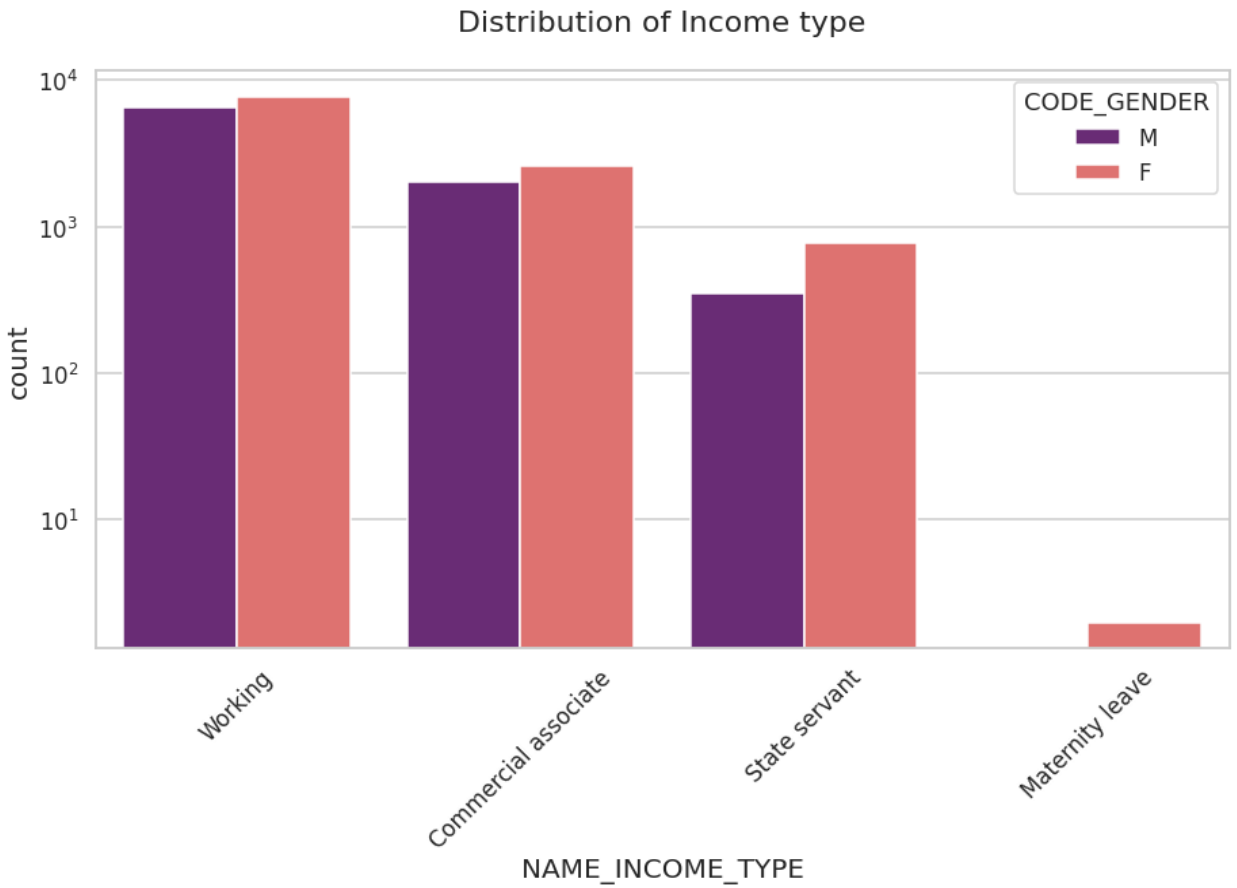


Numerical columns

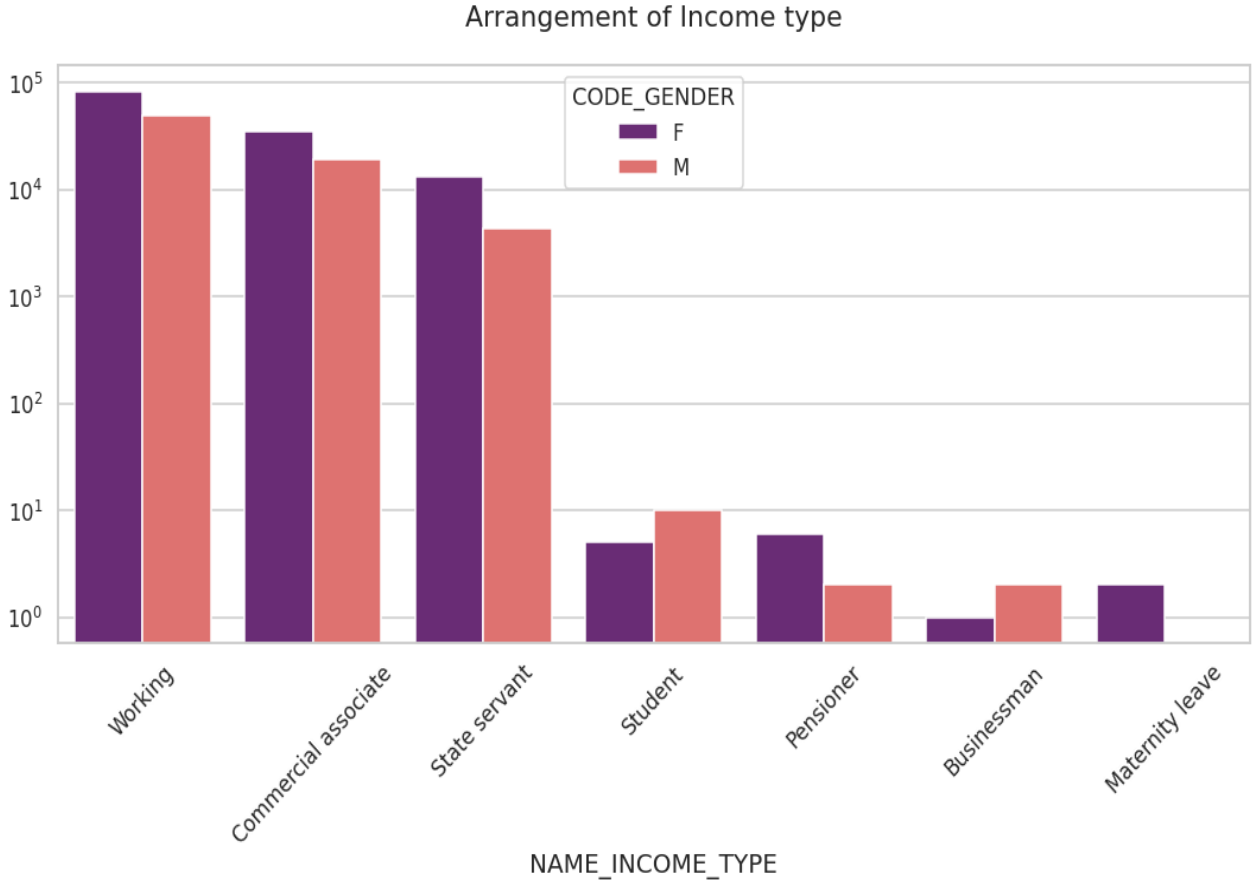


Univariate analysis

For Target=1 (Client facing payment difficulties)



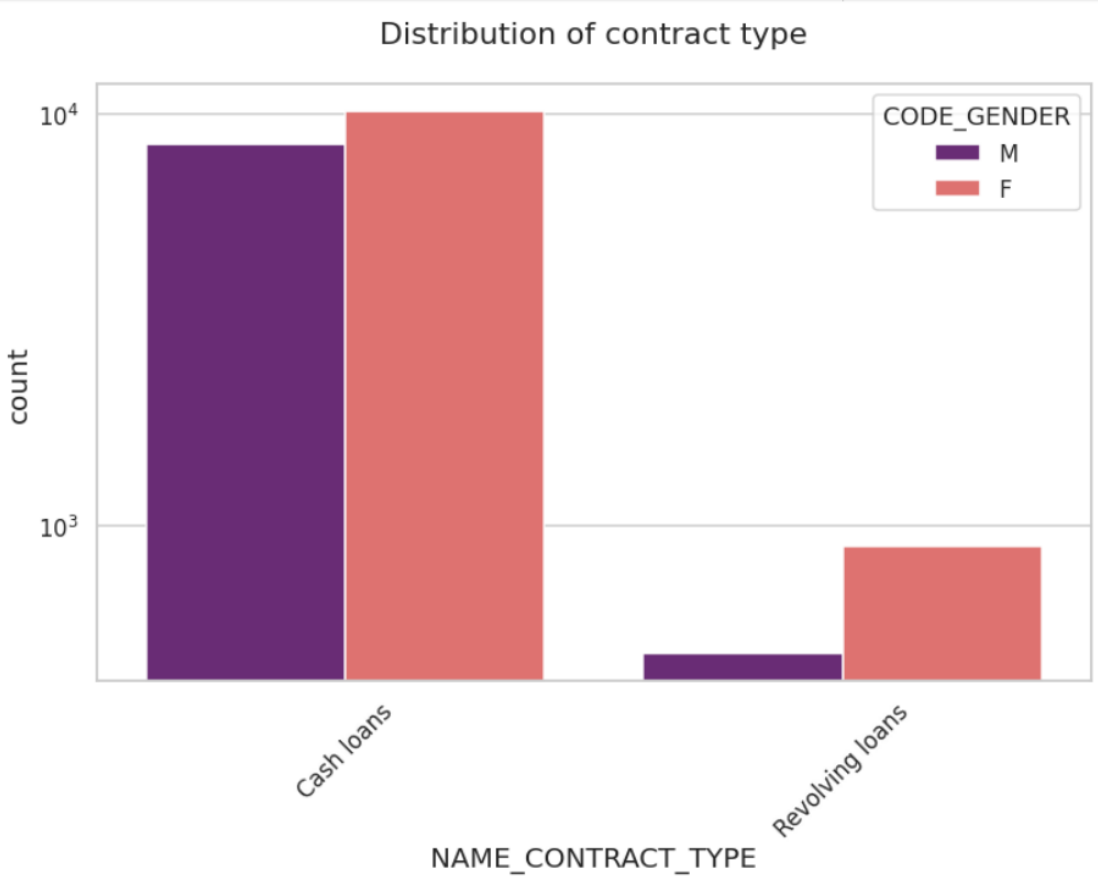
For Target=0 (All other clients)



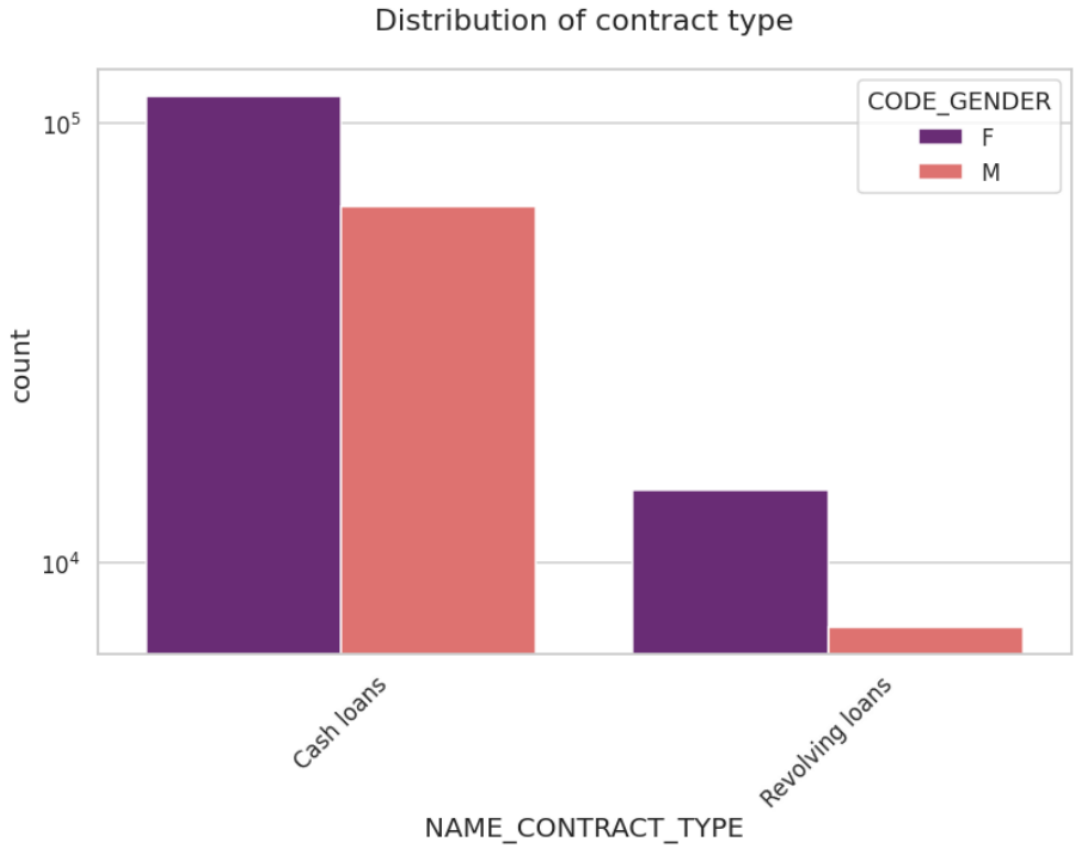
Inference:

- Female application counts is more when compared to male application counts
- Working professional applicants count is more when compared to non-working professional applicants

For Target=1 (Client facing payment difficulties)



For Target=0 (All other clients)

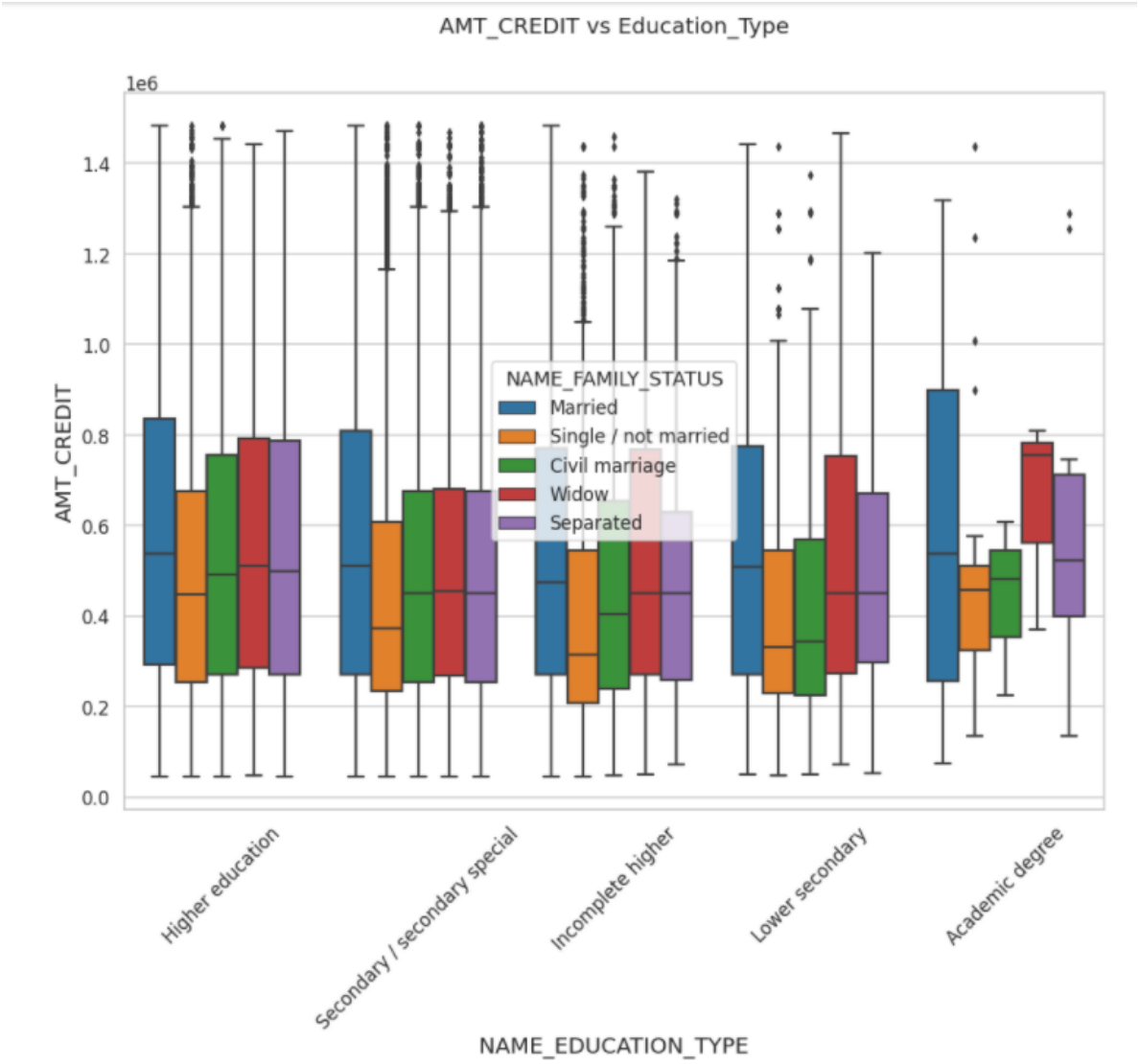


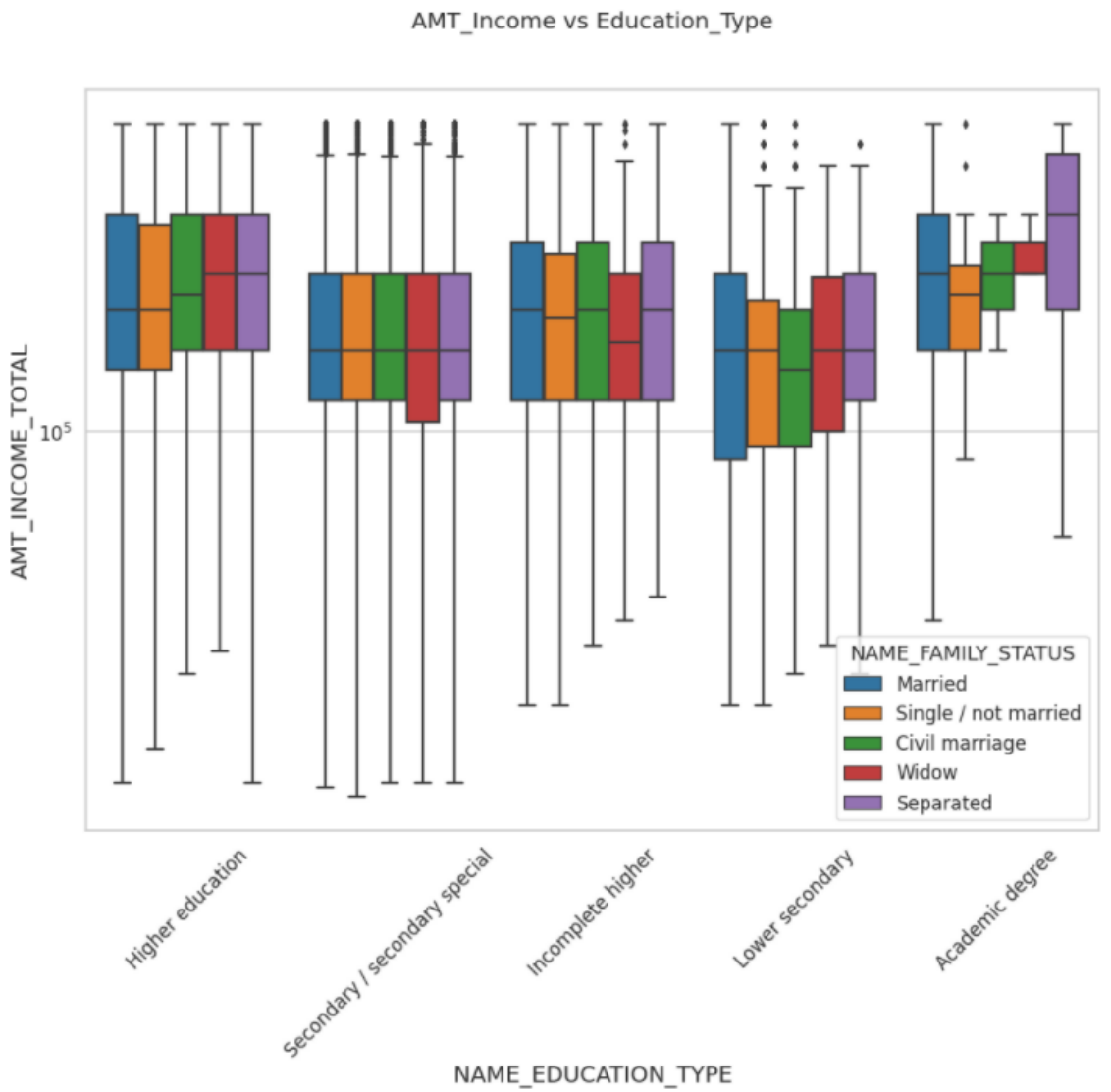
Inference:

- Female application counts is more when compared to male application counts
- Cash loan applications is more when compared to Revolving loans

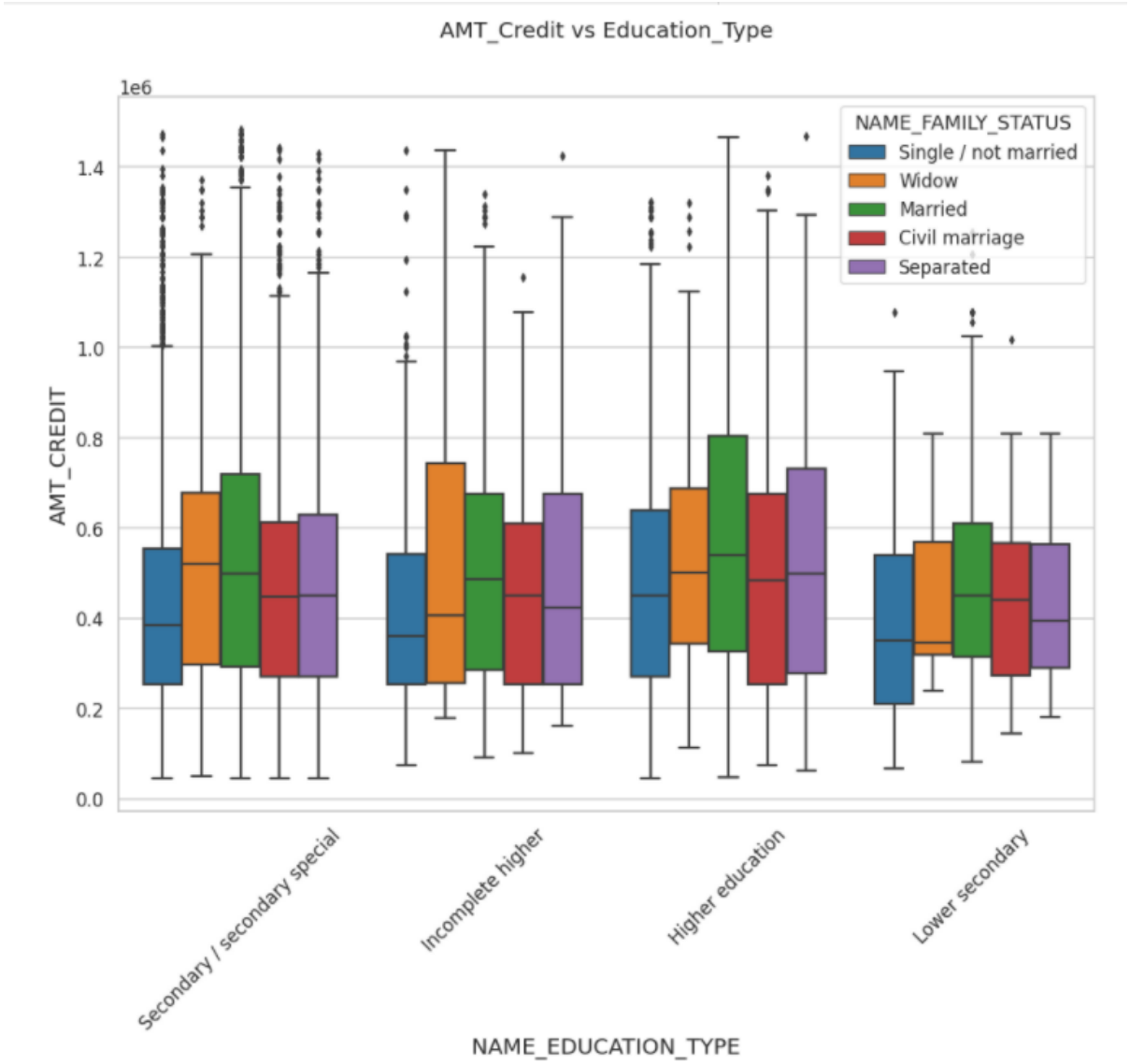
Bivariate analysis

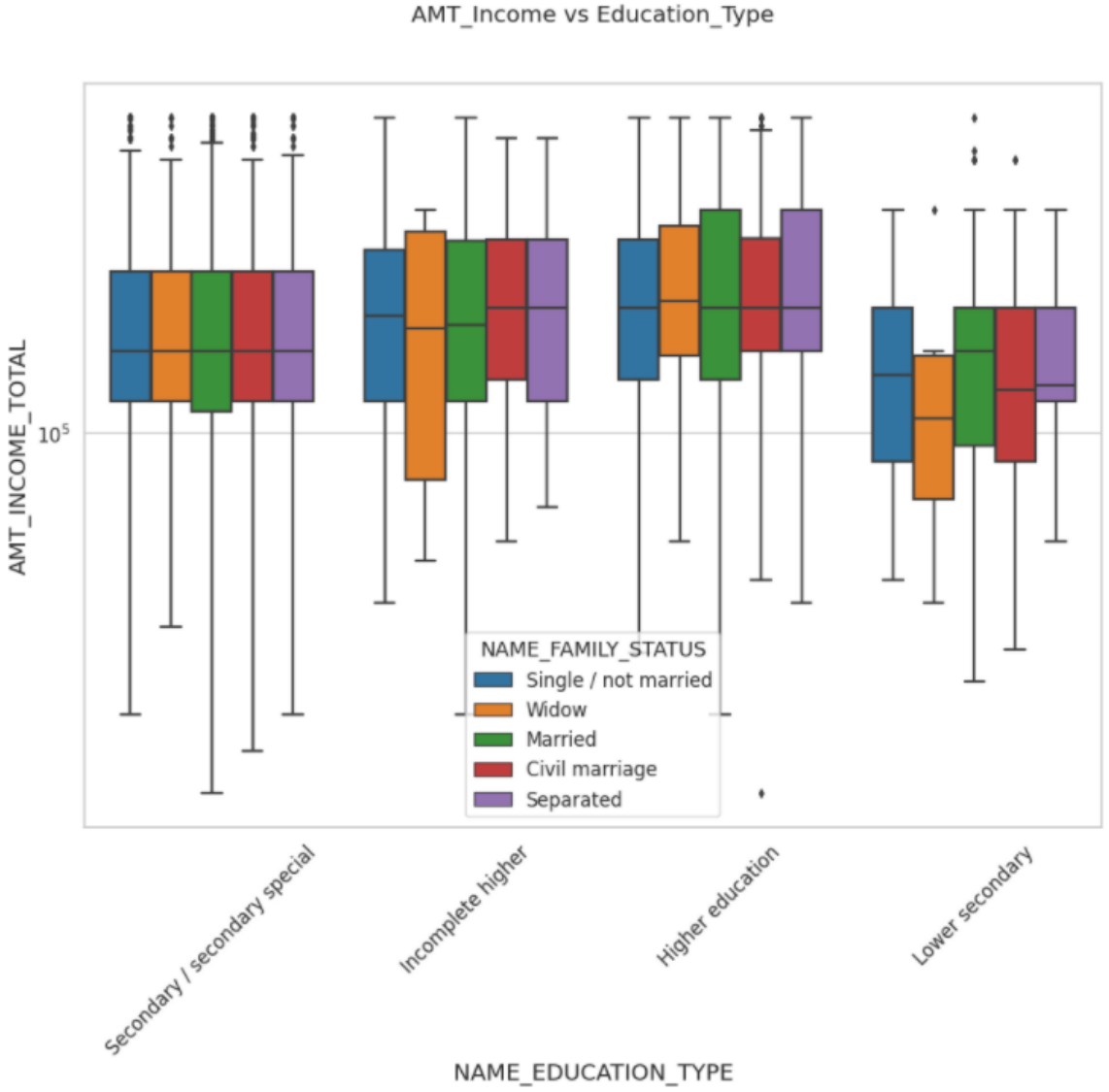
For Target=0 (All other clients)



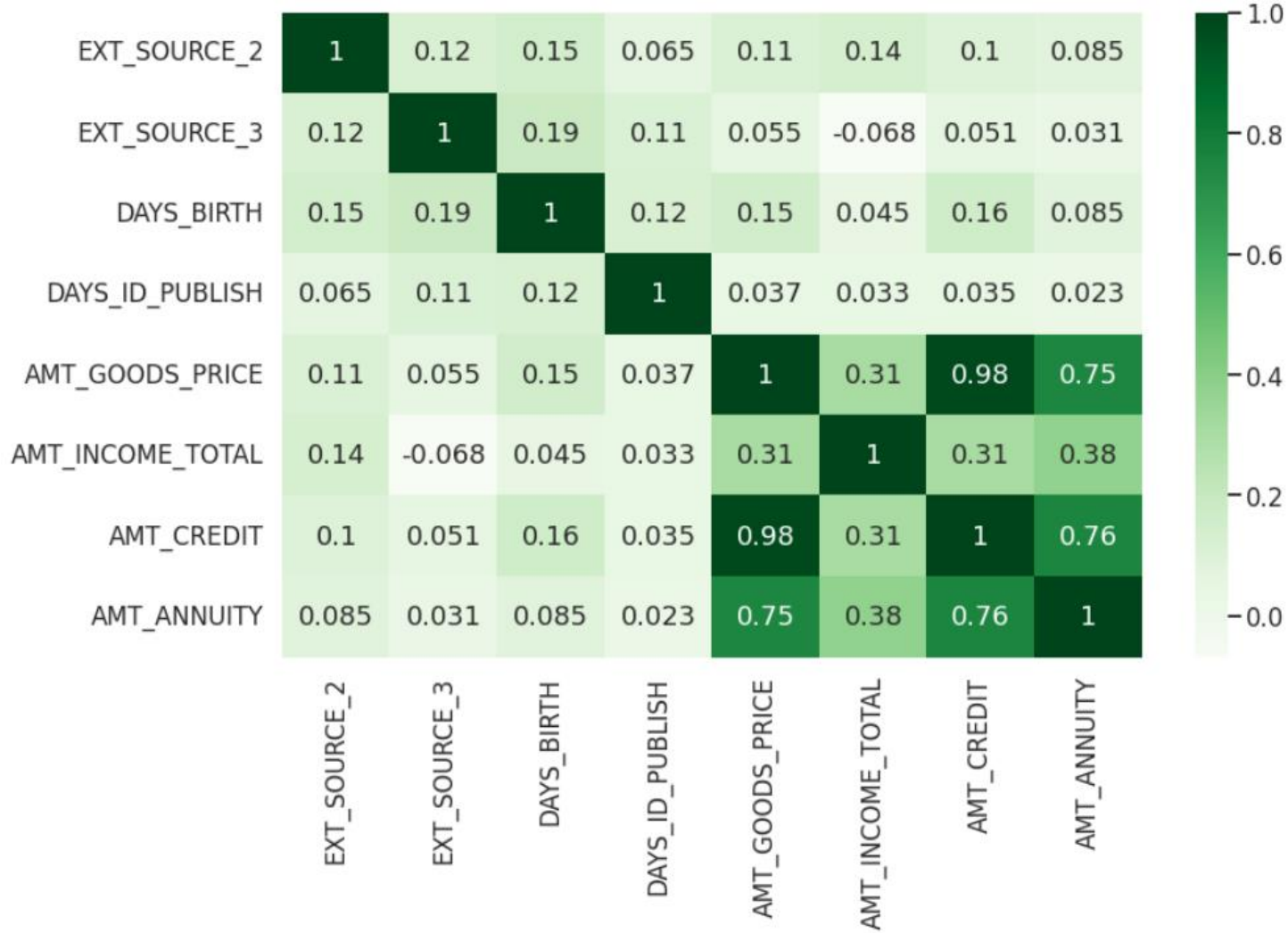


For Target=1 (Client facing payment difficulties)

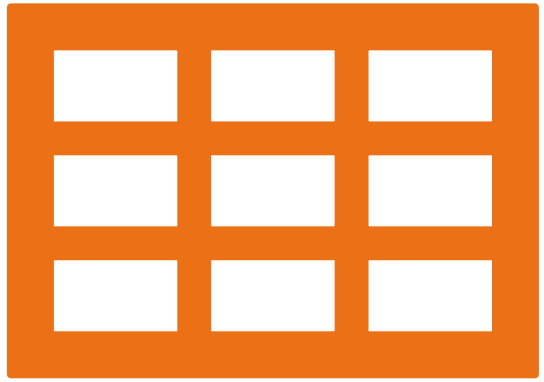




Correlation

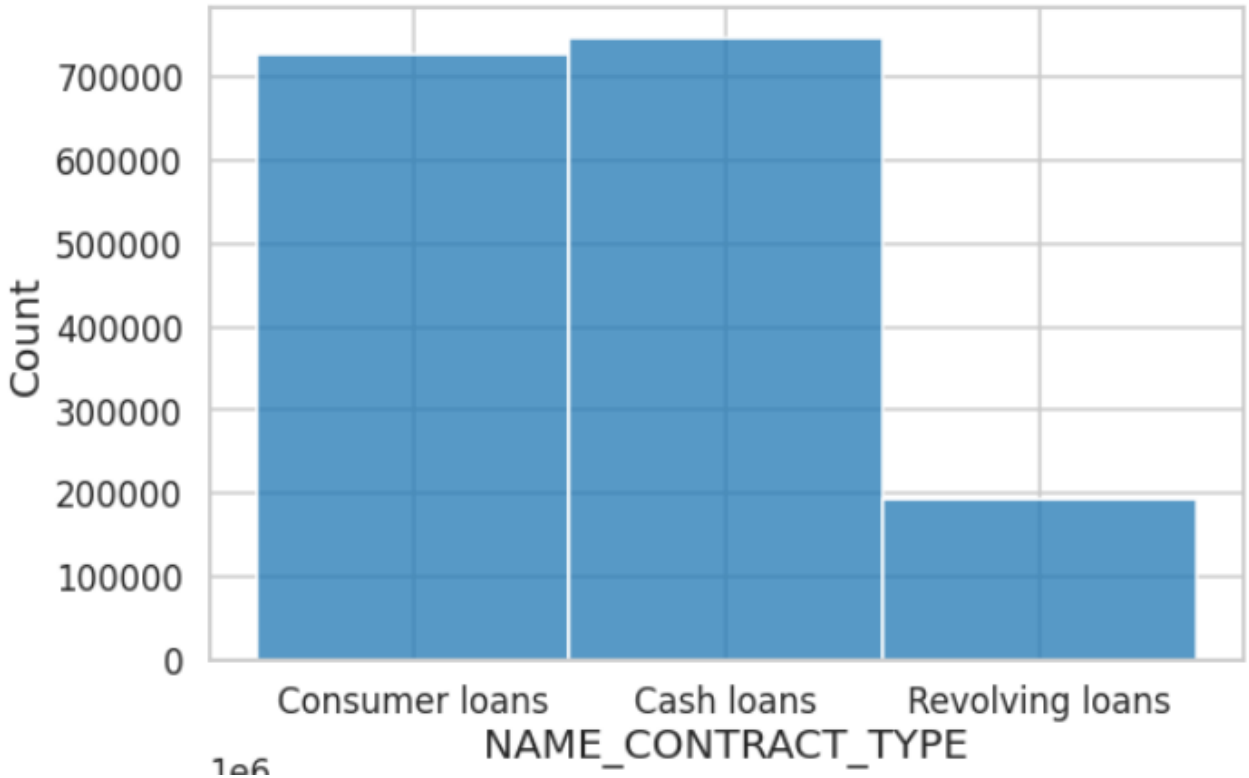

EXT_SOURCE_2 1 0.12 0.15 0.065 0.11 0.14 0.1 0.085
EXT_SOURCE_3 0.12 1 0.19 0.11 0.055 -0.068 0.051 0.031
DAYS_BIRTH 0.15 0.19 1 0.12 0.15 0.045 0.16 0.085
DAYS_ID_PUBLISH 0.065 0.11 0.12 1 0.037 0.033 0.035 0.023
AMT_GOODS_PRICE 0.11 0.055 0.15 0.037 1 0.31 0.98 0.75
AMT_INCOME_TOTAL 0.14 -0.068 0.045 0.033 0.31 1 0.31 0.38
AMT_CREDIT 0.1 0.051 0.16 0.035 0.98 0.31 1 0.76
AMT_ANNUITY 0.085 0.031 0.085 0.023 0.75 0.38 0.76 1
EXT_SOURCE_2
EXT_SOURCE_3
DAYS_BIRTH
DAYS_ID_PUBLISH
AMT_GOODS_PRICE
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
1.0
0.8
0.6
0.4
0.2
0.0

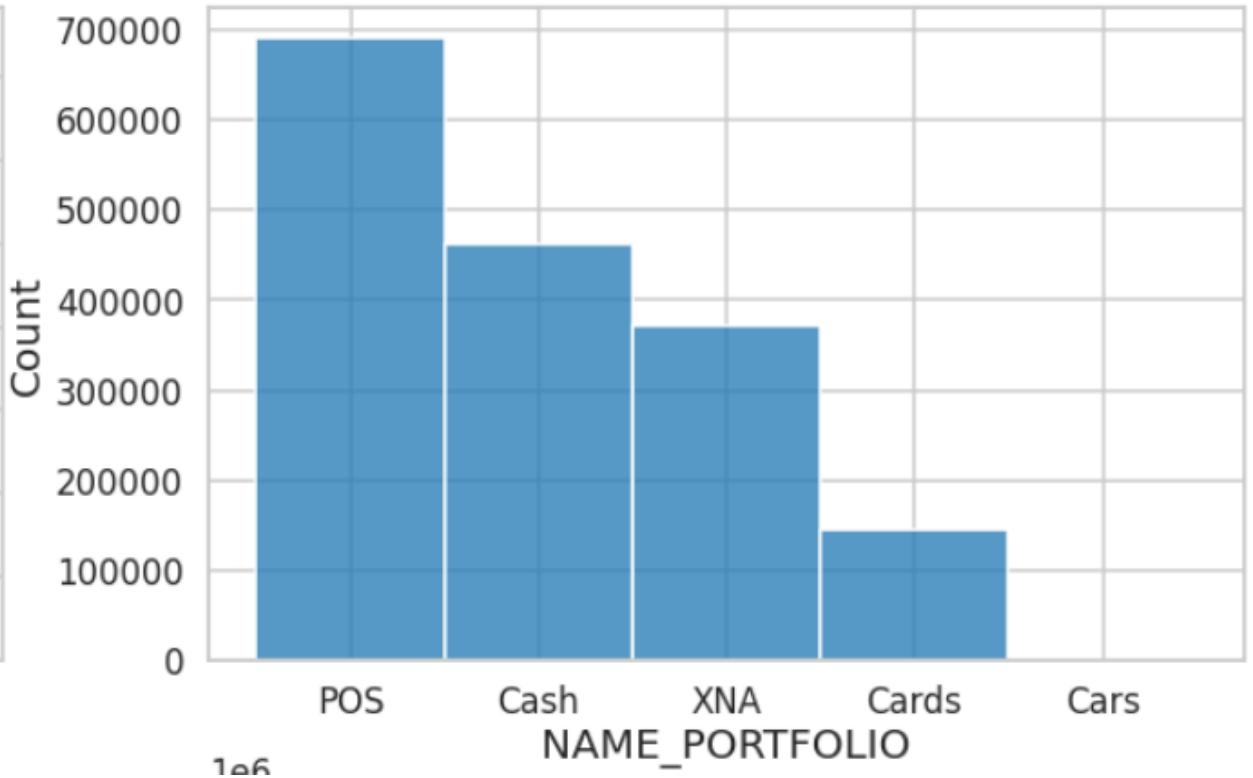
Previous application Dataset

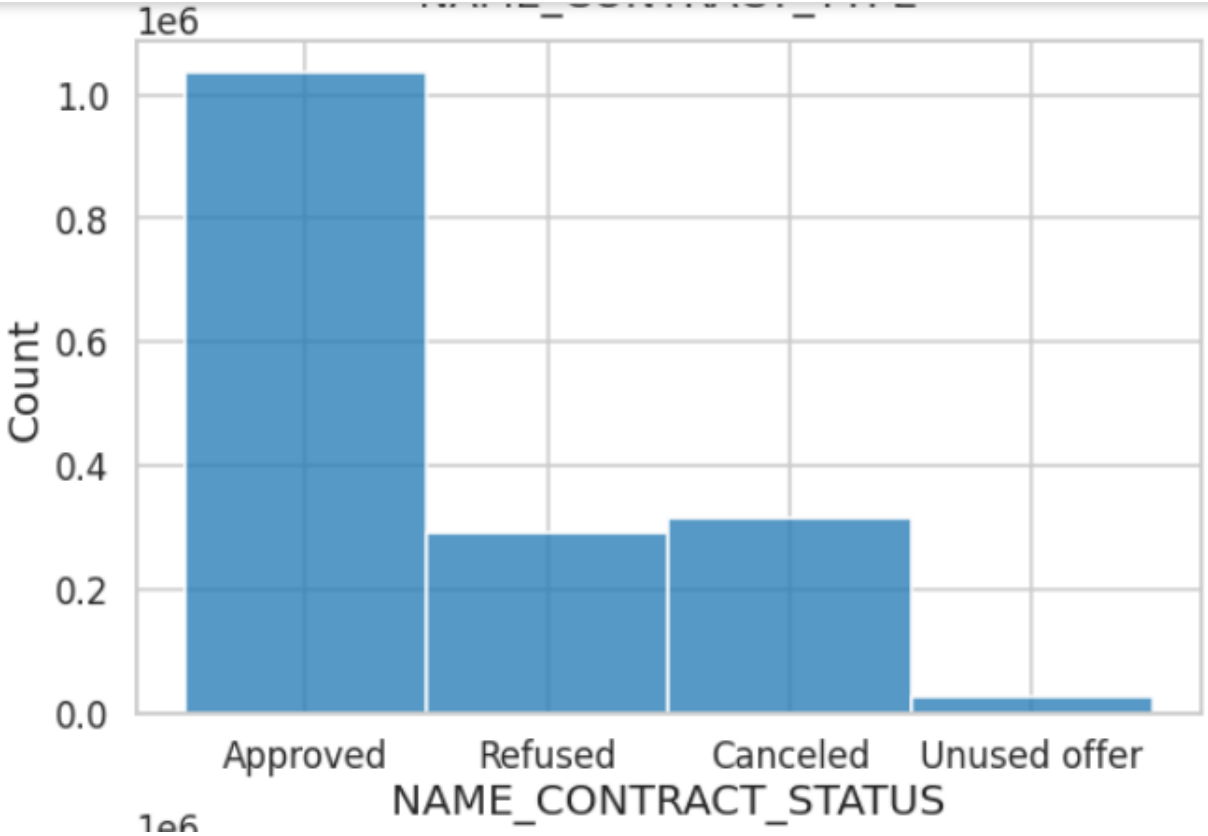


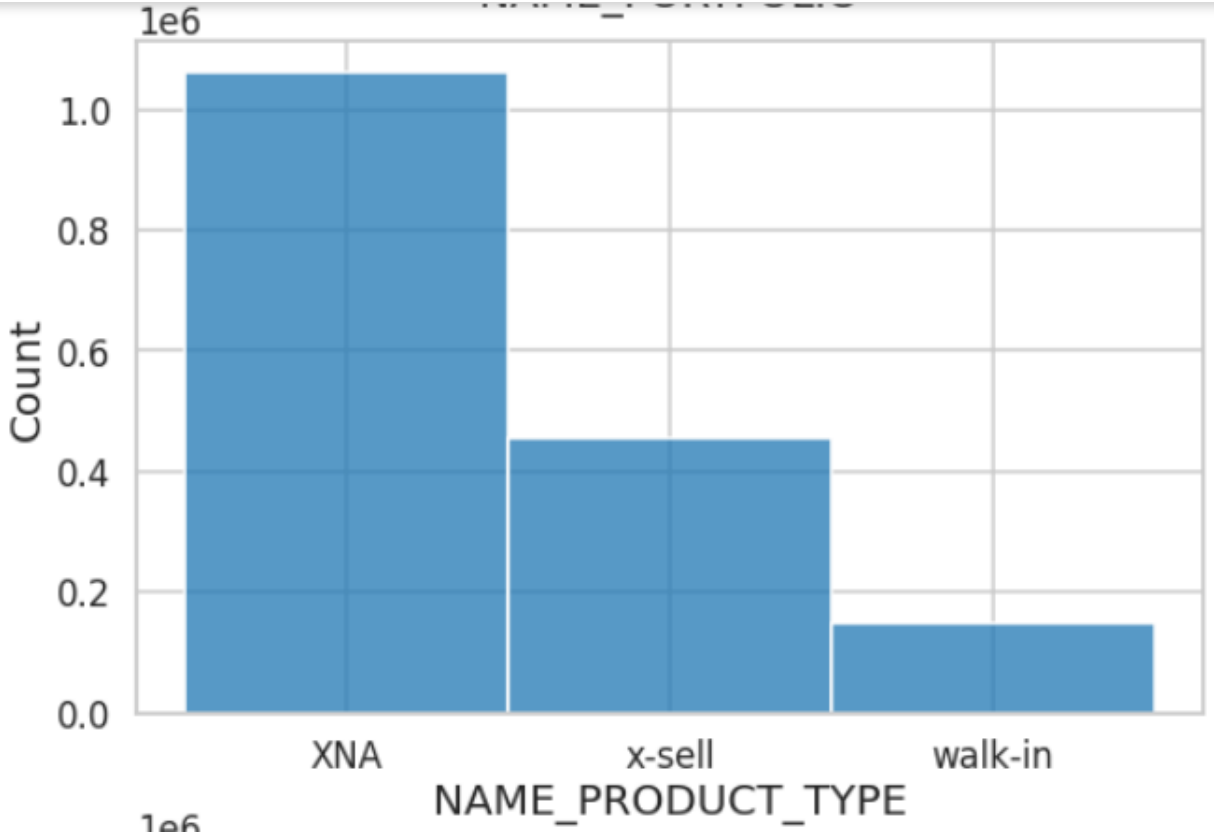
- Verified the dimensions of dataset(i.e, no.of rows and columns)
- Dropping the columns having null values greater than 40%.
- Also dropping insignificant columns which are not useful in decision making
- Imputed missing values of some useful columns with median value.
- Handling missing values.
- Done Univariate analysis
- Correlation chart

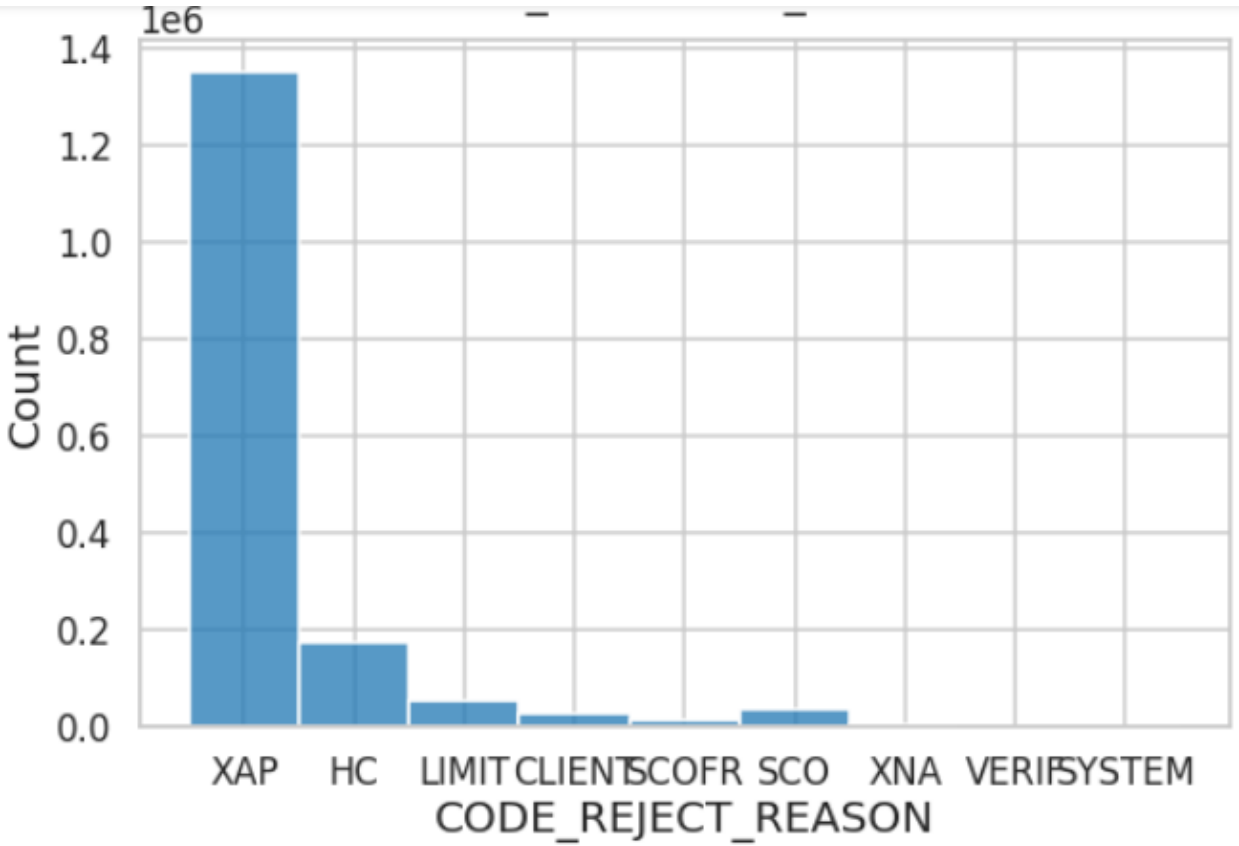
Univariate analysis

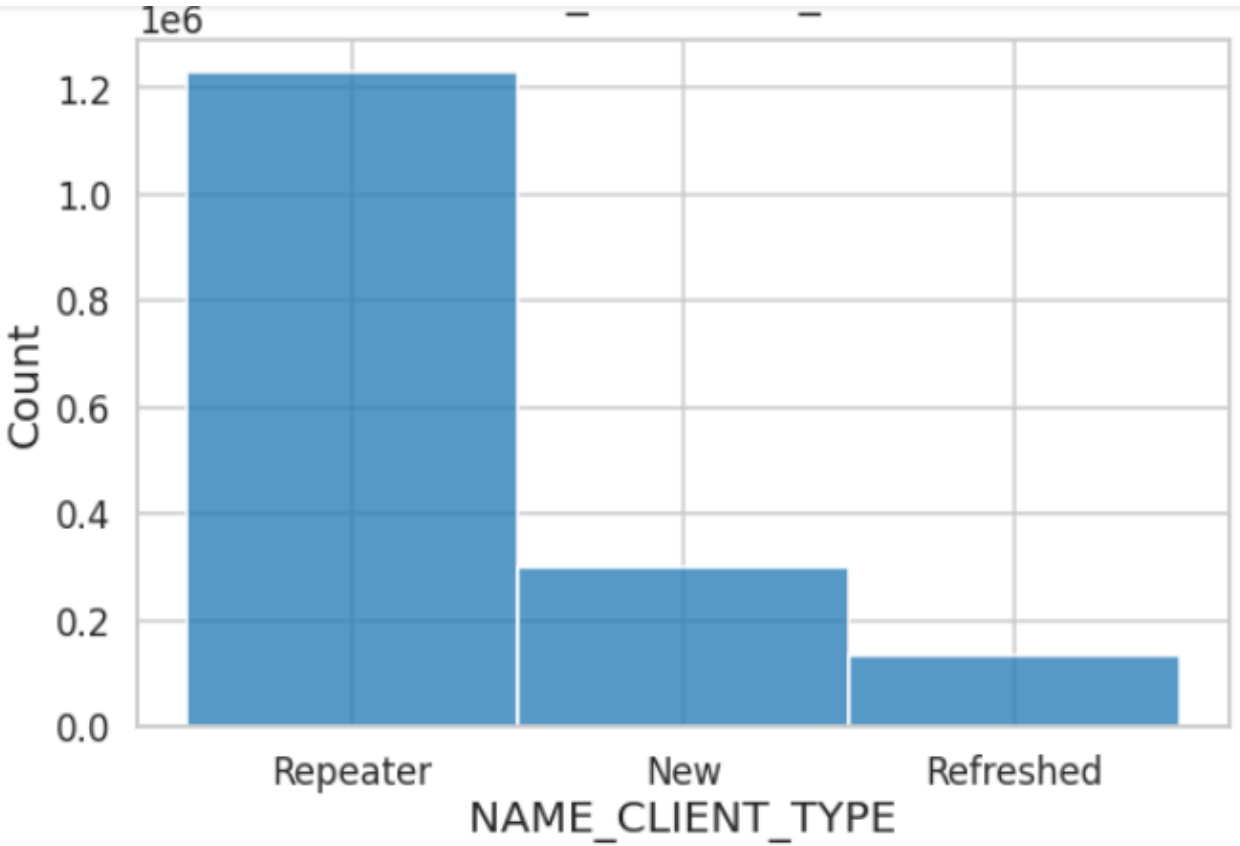

700000
600000
500000
400000
300000
200000
100000
0
Count
Consumer loans
Cash loans
Revolving loans
NAME_CONTRACT_TYPE
1e6

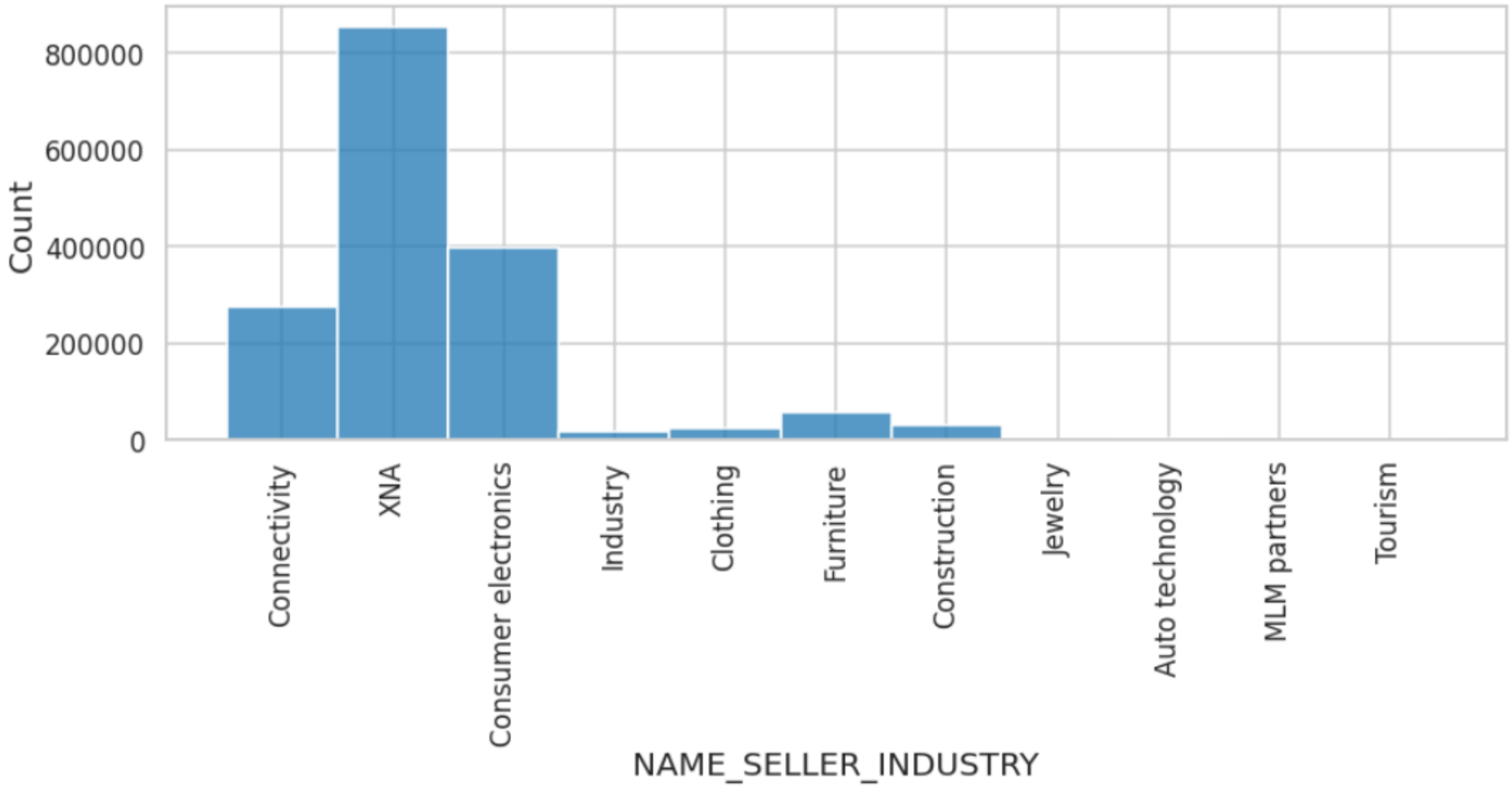

700000
600000
500000
400000
300000
200000
100000
0
Count
POS
Cash
XNA
Cards
Cars
NAME_PORTFOLIO
1e6

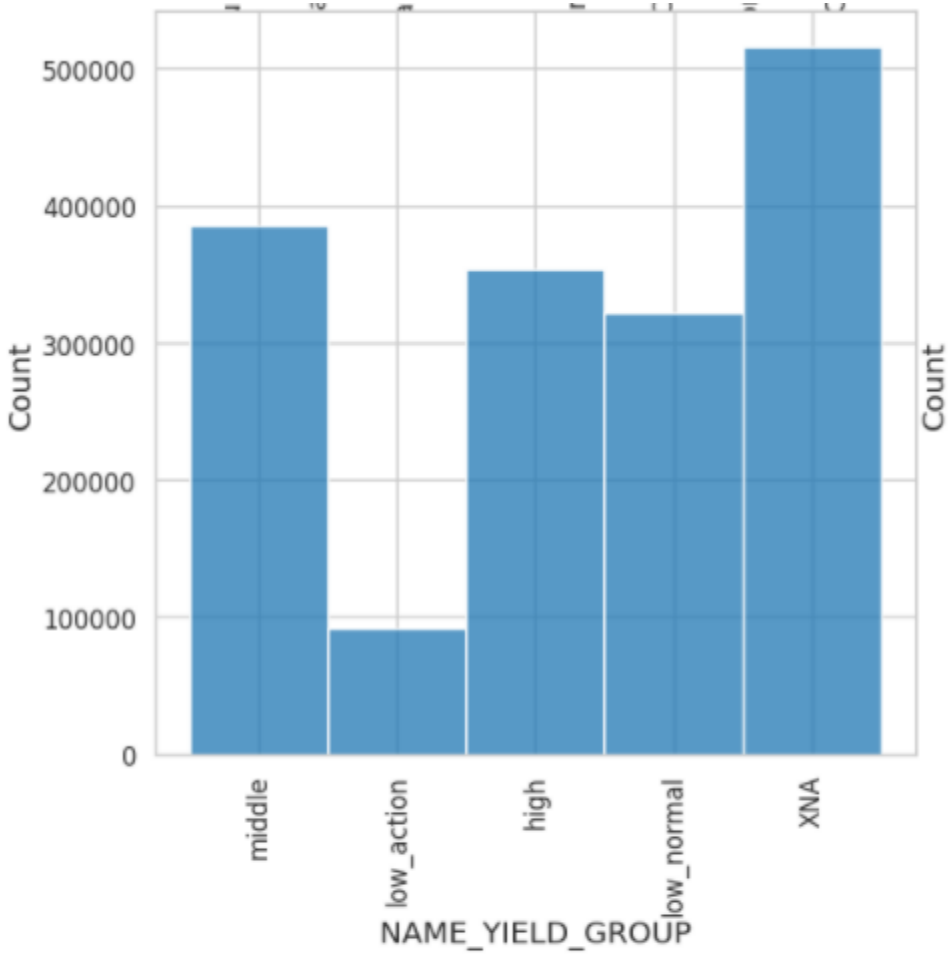

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS

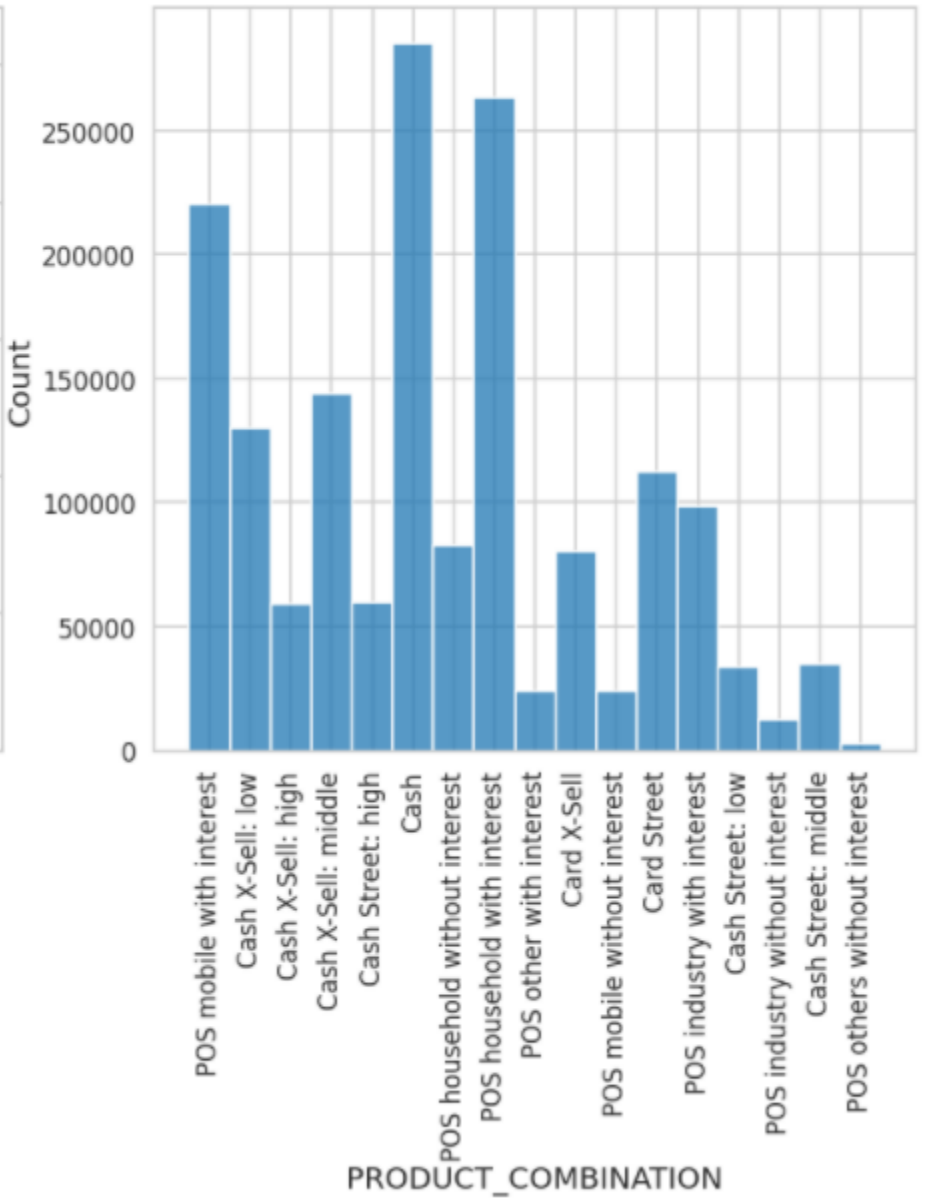

1e6
1.0
0.8
0.6
0.4
0.2
0.0
Count
XNA
x-sell
walk-in
NAME_PRODUCT_TYPE


1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
XAP
HC
LIMIT
CLIENT
SCOFR
SCO
XNA
VERIF
SYSTEM
CODE_REJECT_REASON


1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Count
Repeater
New
Refreshed
NAME_CLIENT_TYPE


800000
600000
400000
200000
0
Count
Connectivity
XNA
Consumer electronics
Industry
Clothing
Furniture
Construction
Jewelry
Auto technology
MLM partners
Tourism
NAME_SELLER_INDUSTRY

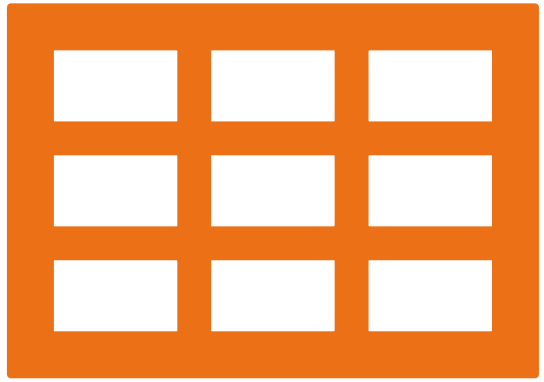

Count
500000
400000
300000
200000
100000
0
middle
low_action
high
low_normal
XNA
NAME_YIELD_GROUP


Count
250000
200000
150000
100000
50000
0
POS mobile with interest
Cash X-Sell: low
Cash X-Sell: high
Cash X-Sell: middle
Cash Street: high
Cash
POS household without interest
POS household with interest
POS other with interest
Card X-Sell
POS mobile without interest
Card Street
POS industry with interest
Cash Street: low
POS industry without interest
Cash Street: middle
POS others without interest
PRODUCT_COMBINATION

Correlation

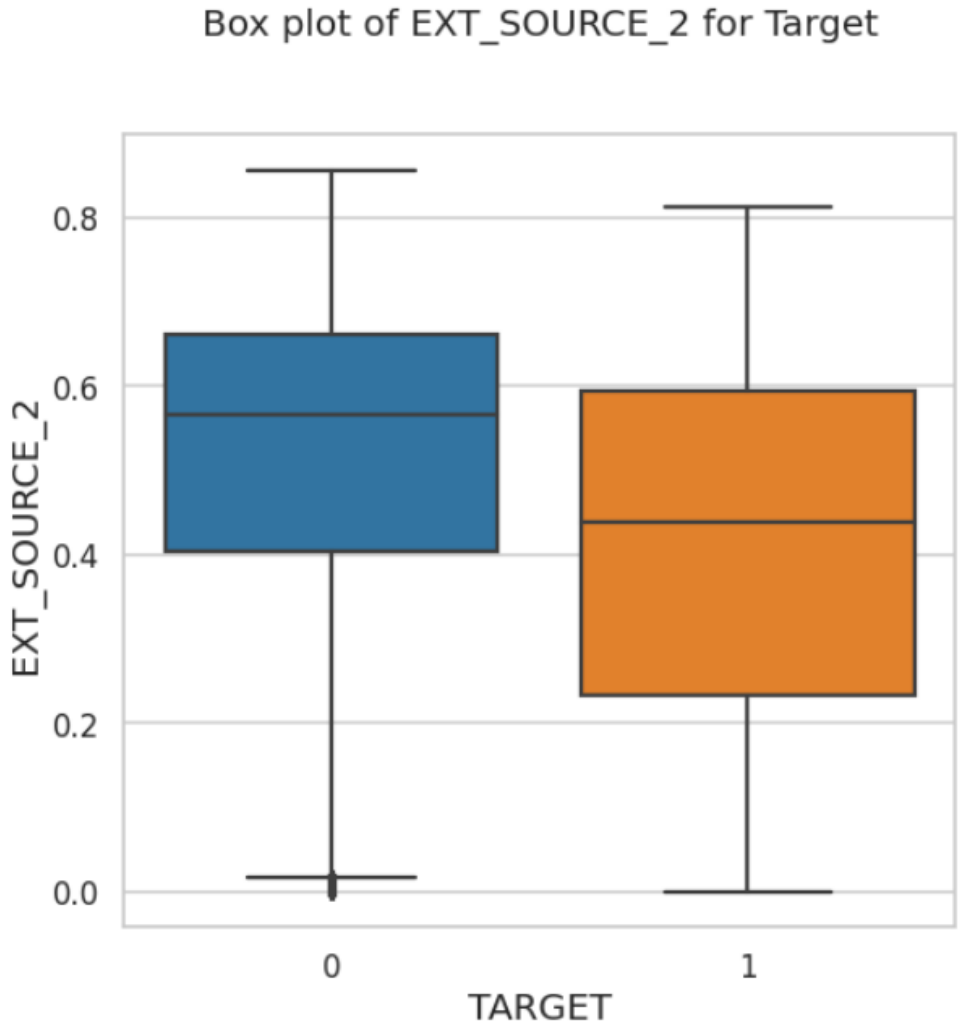

AMT_ANNUITY
AMT_APPLICATION
AMT_CREDIT
AMT_GOODS_PRICE
DAYS_DECISION
1
0.26
0.32
0.82
-0.19
0.26
1
0.76
0.17
0.03
0.32
0.76
1
0.21
0.011
0.82
0.17
0.21
1
-0.19
-0.19
0.03
0.011
-0.19
1
1.0
0.8
0.6
0.4
0.2
0.0

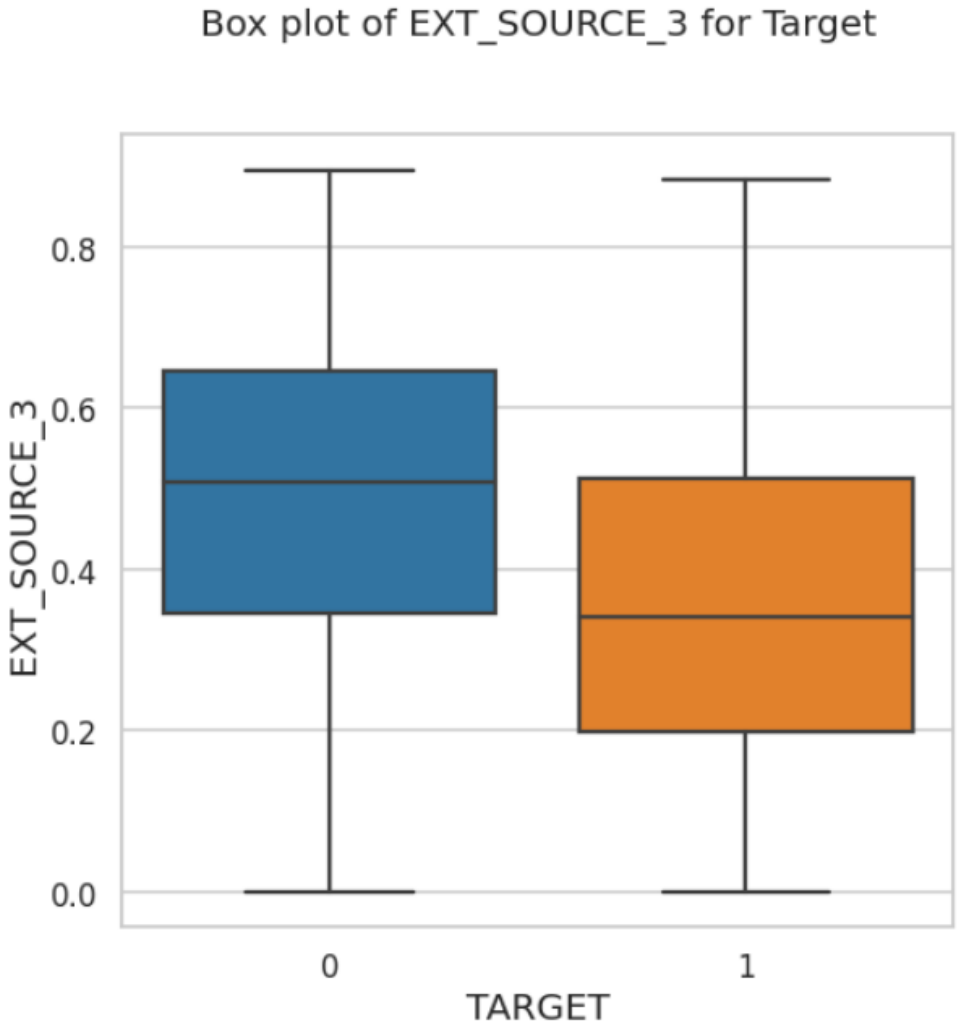
Merging Application dataset and Previous application dataset



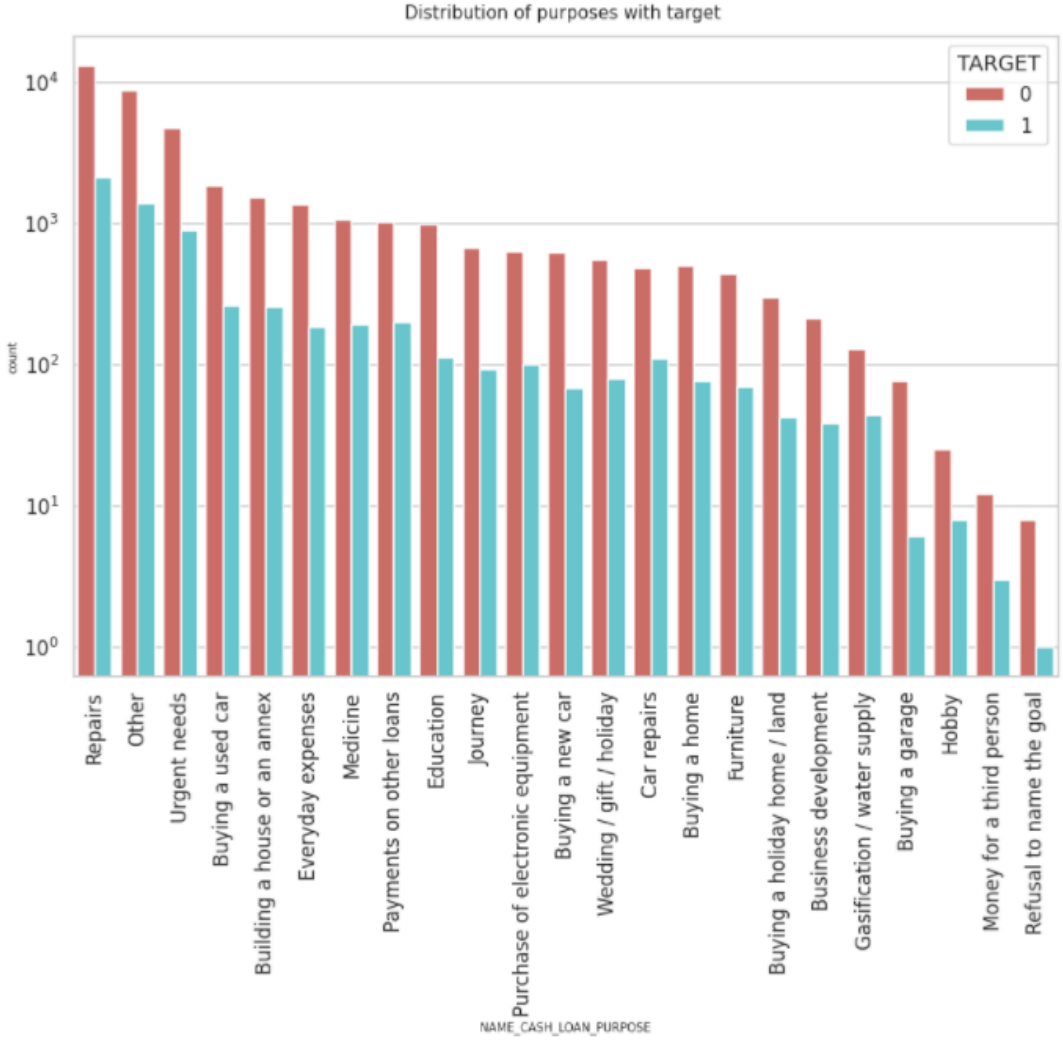
- Verified the dimensions of dataset(i.e, no.of rows and columns)
- Also dropping insignificant columns which are not useful in decision making
- Imputed missing values of some useful columns with median value.
- Handling missing values.
- Segmentation analysis
- Correlation chart

EXT_SOURCE_2 Vs TARGET VALUE & EXT_SOURCE_3 Vs TARGET VALUE

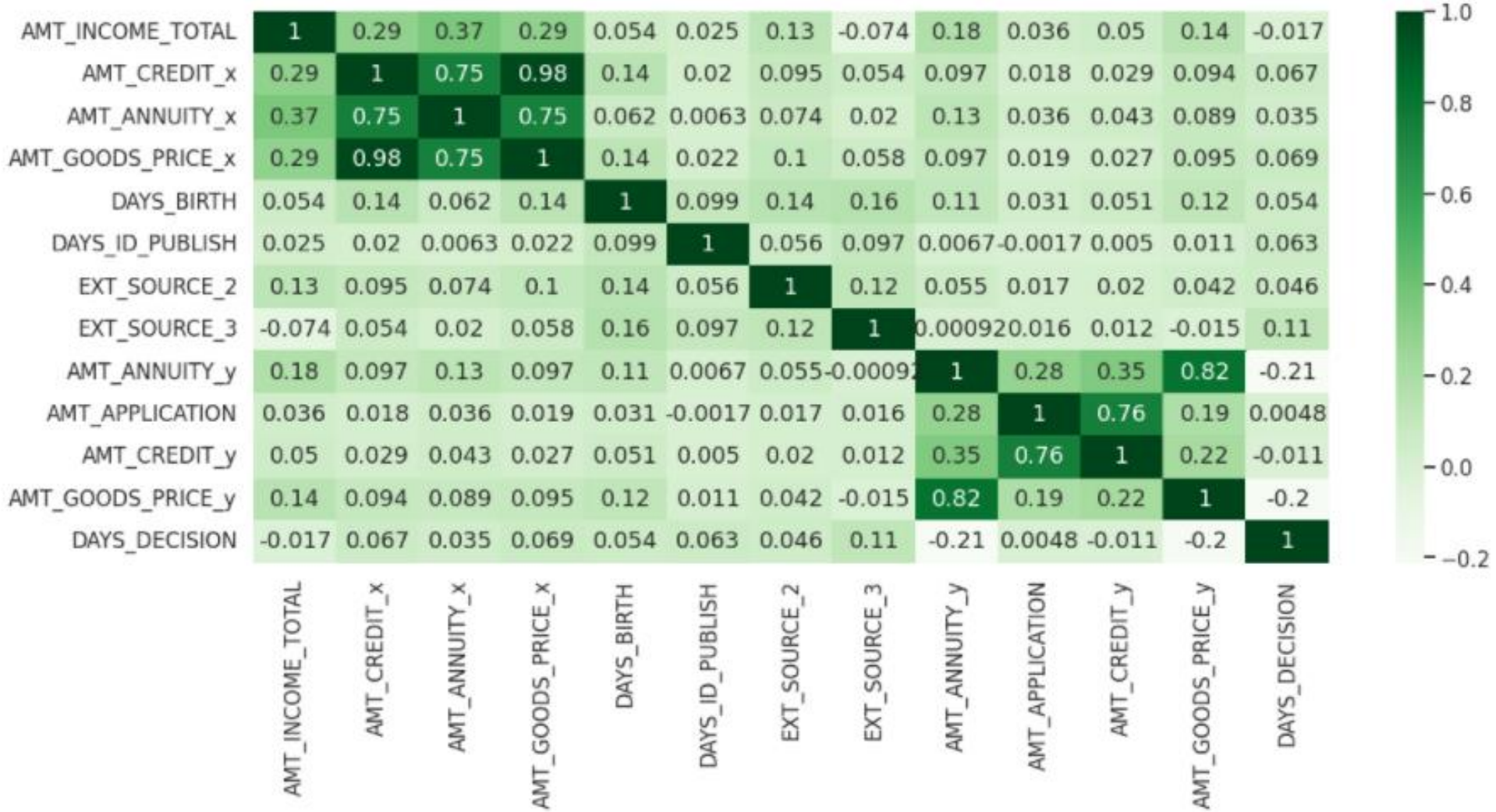




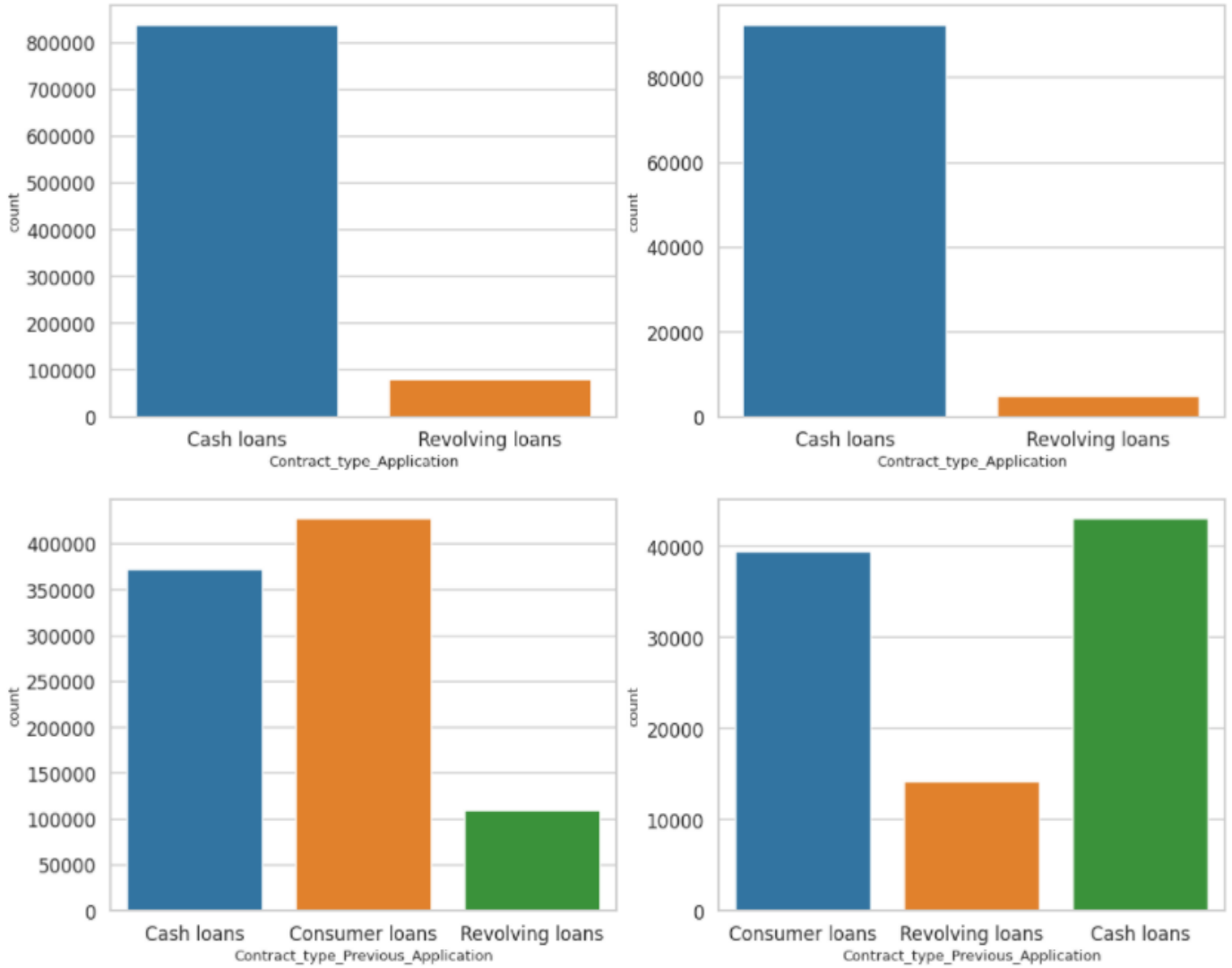
TARGET VALUE Vs CASH_LOAN_PURPOSE



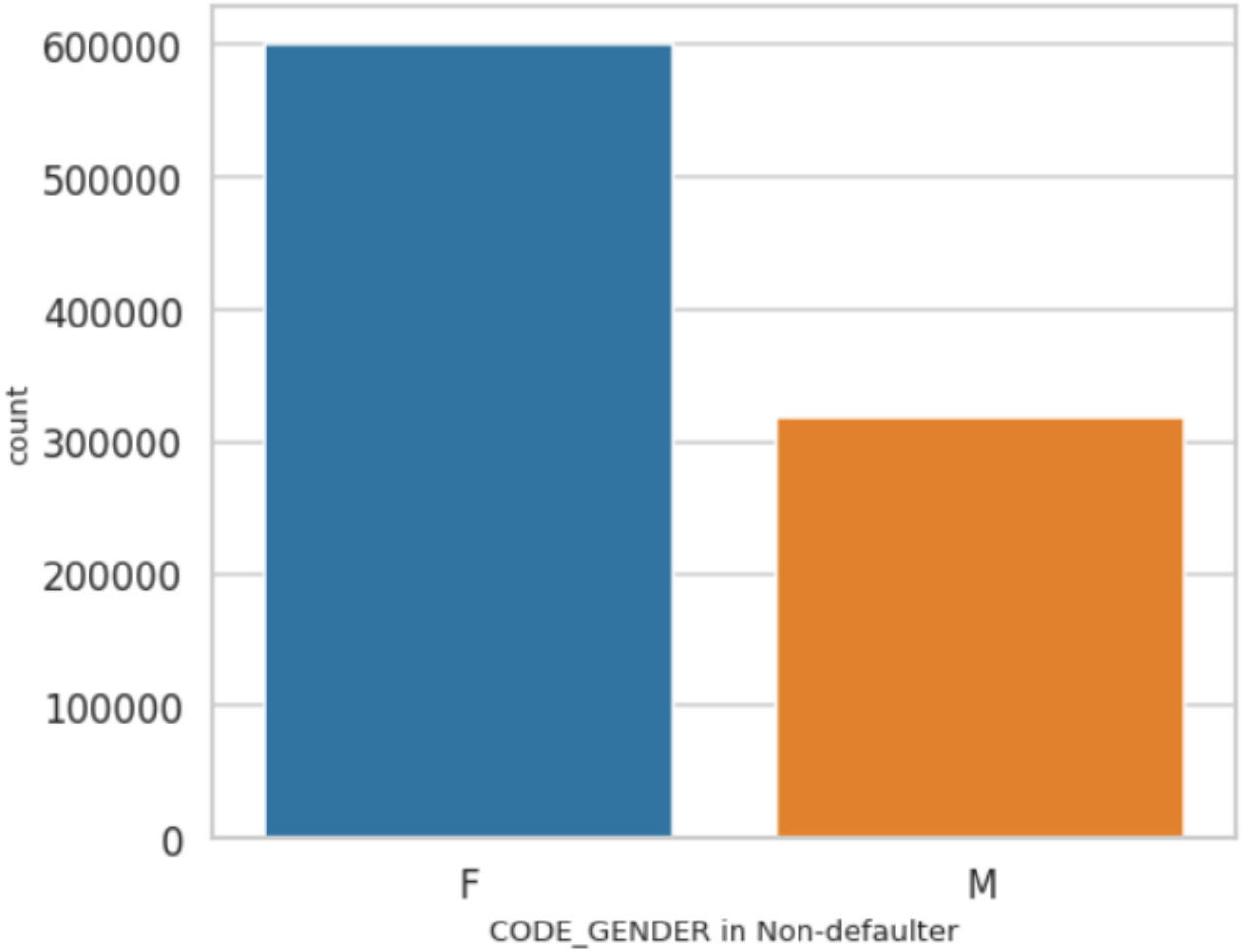
CORRELATION CHART

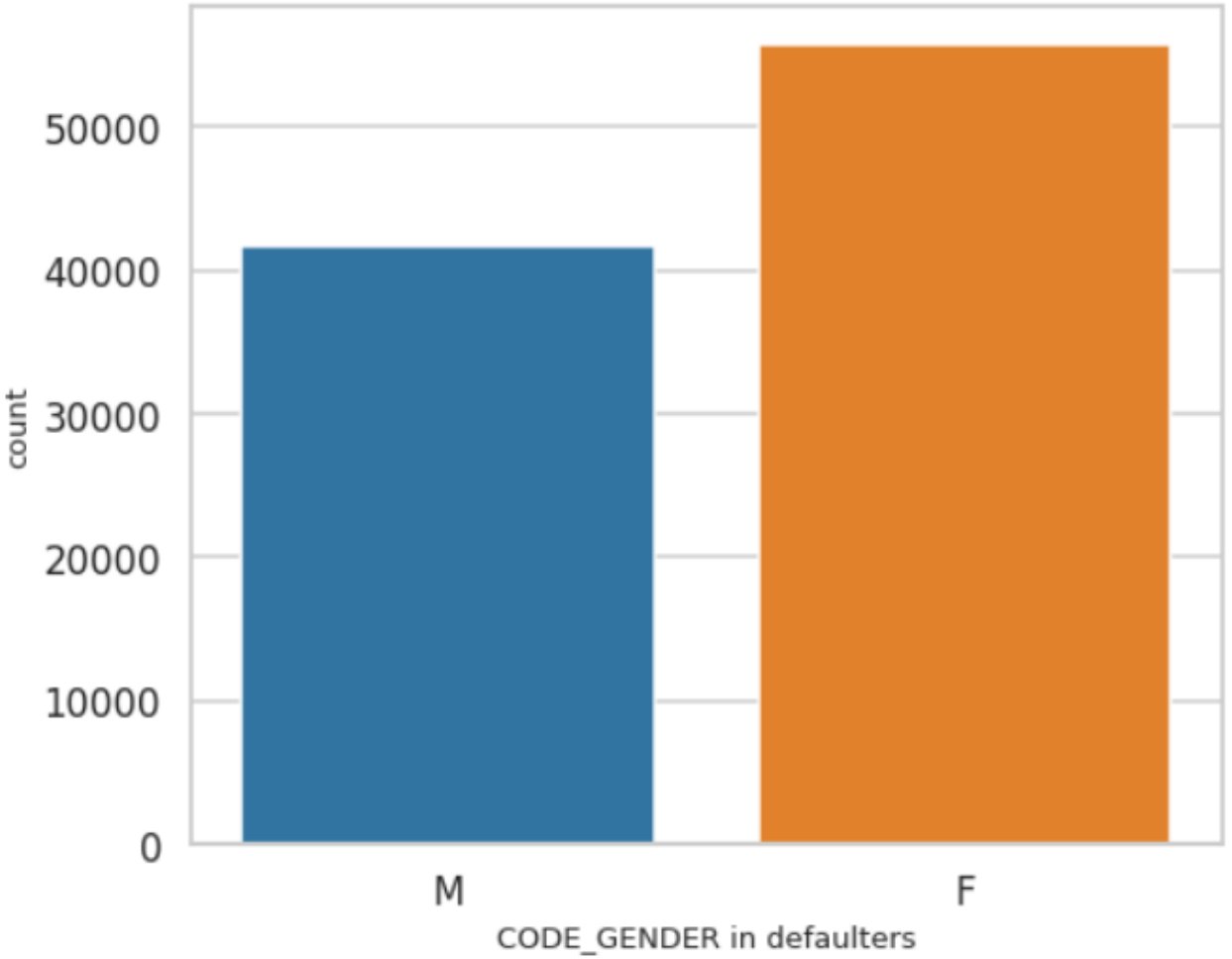


NAME_CONTRACT_TYPE for Non-Defaulters (left) and Defaulters(Right)

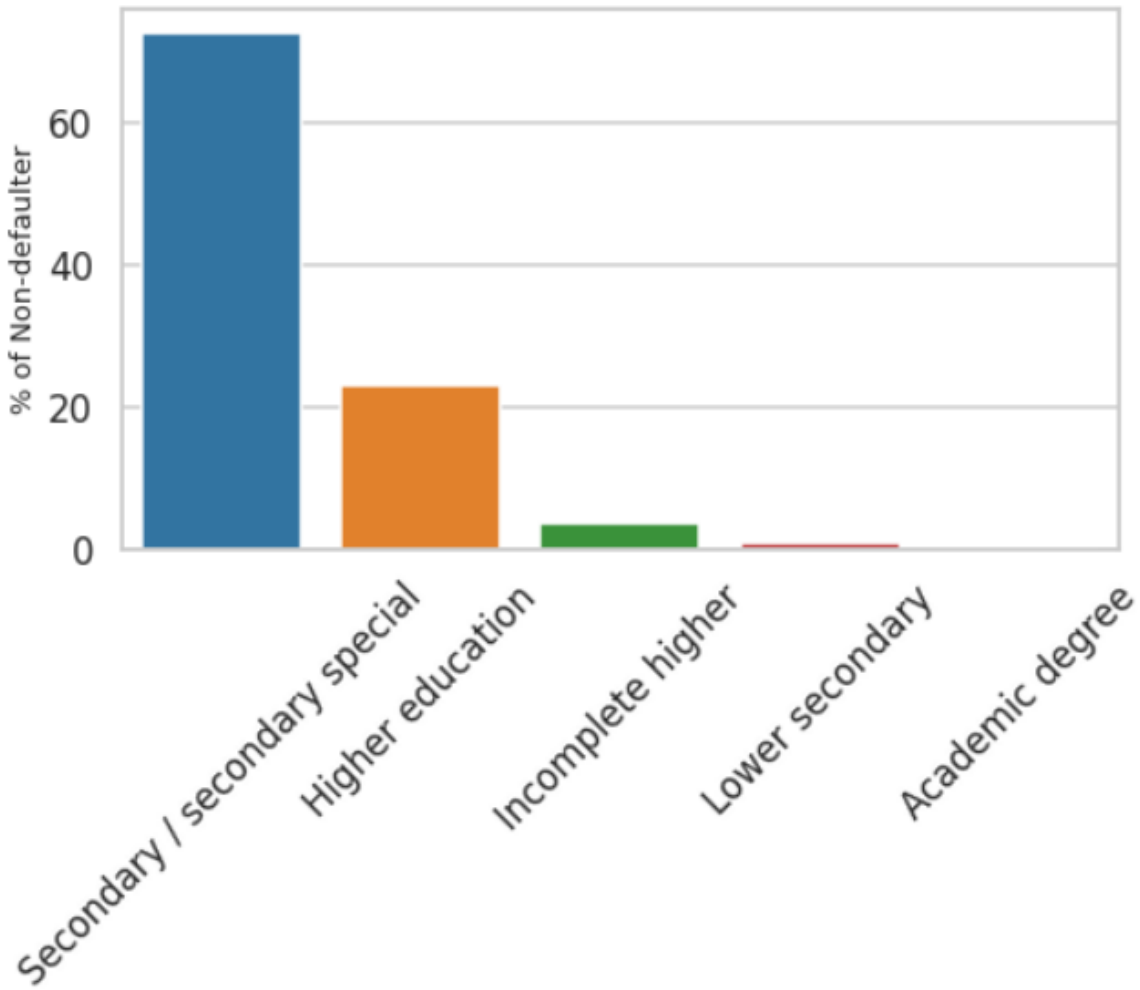


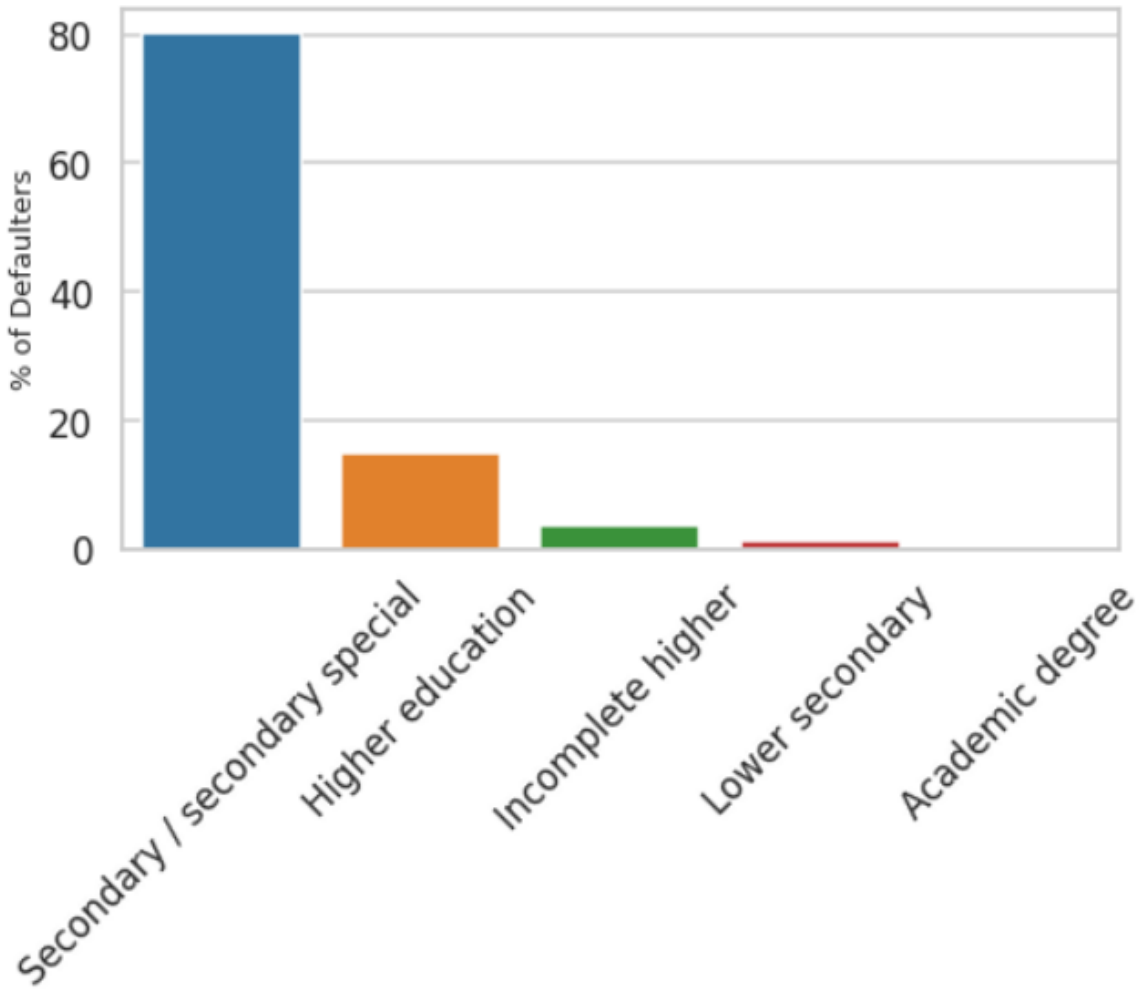
CODE_GENDER for Non-Defaulters (left) and Defaulters(Right)

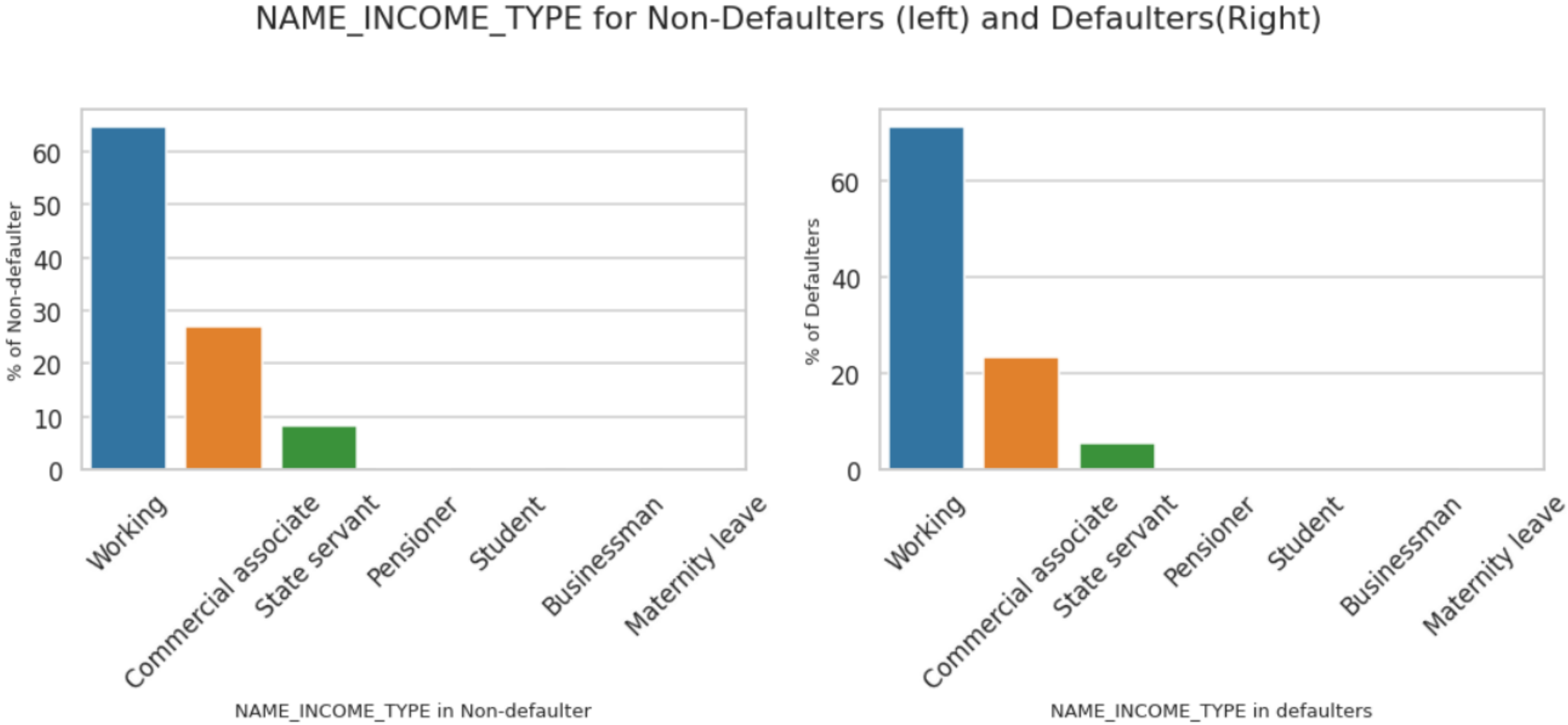




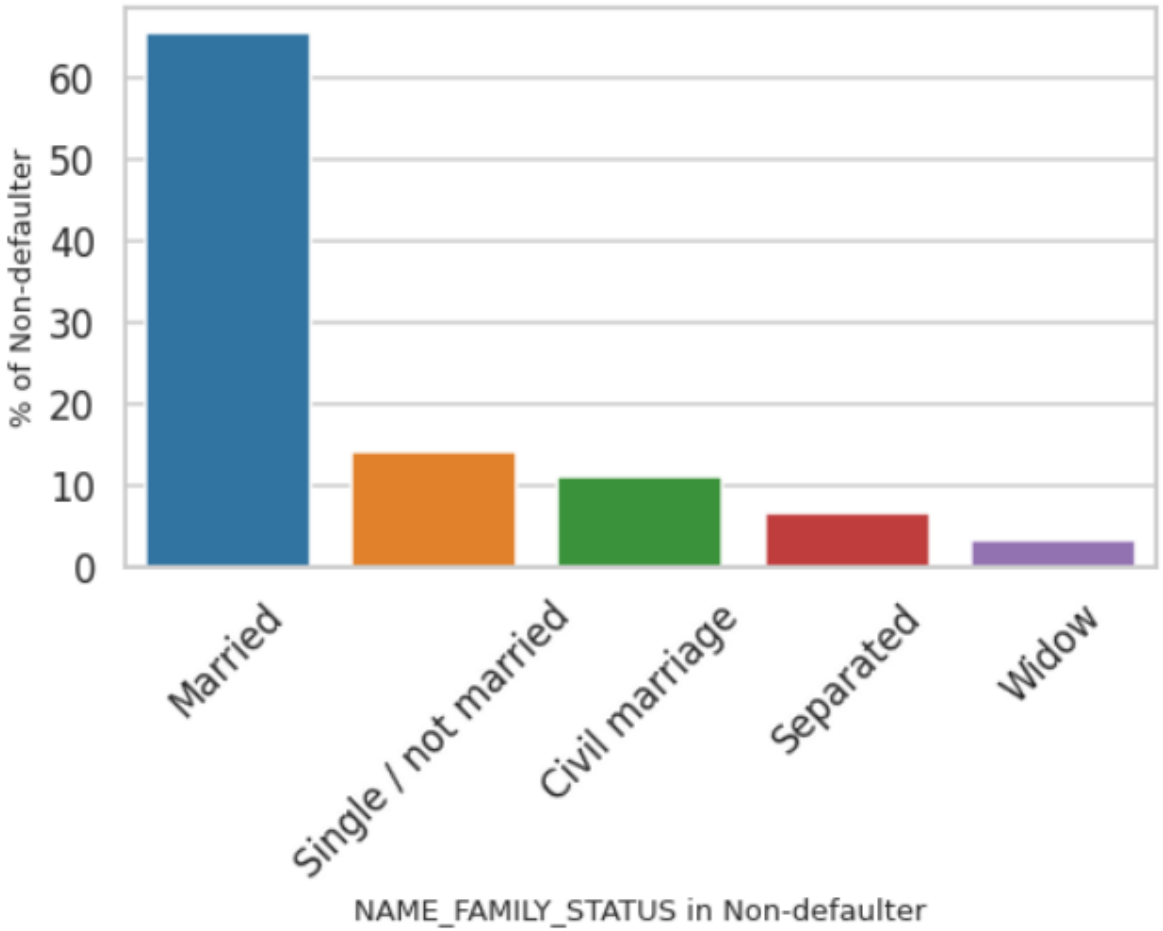
NAME_EDUCATION_TYPE for Non-Defaulters (left) and Defaulters(Right)

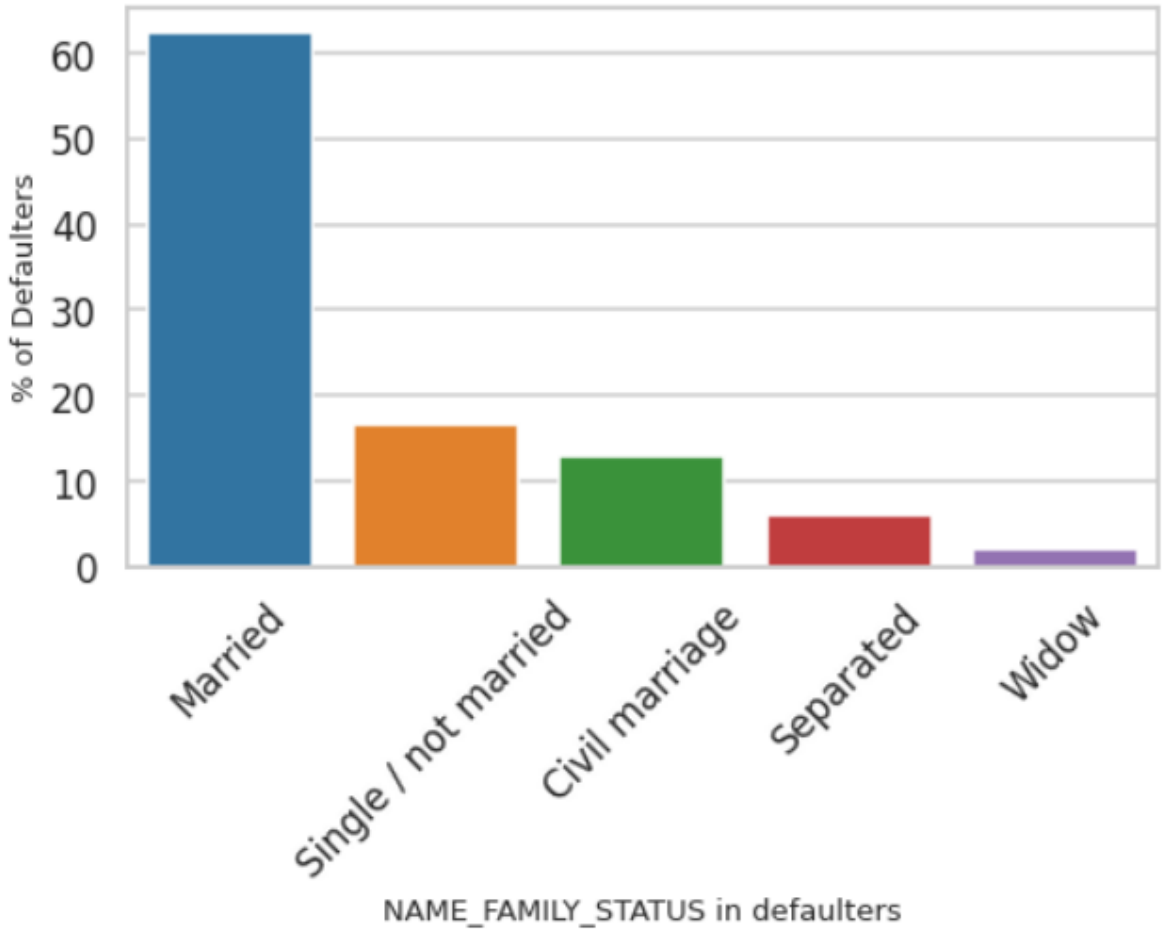




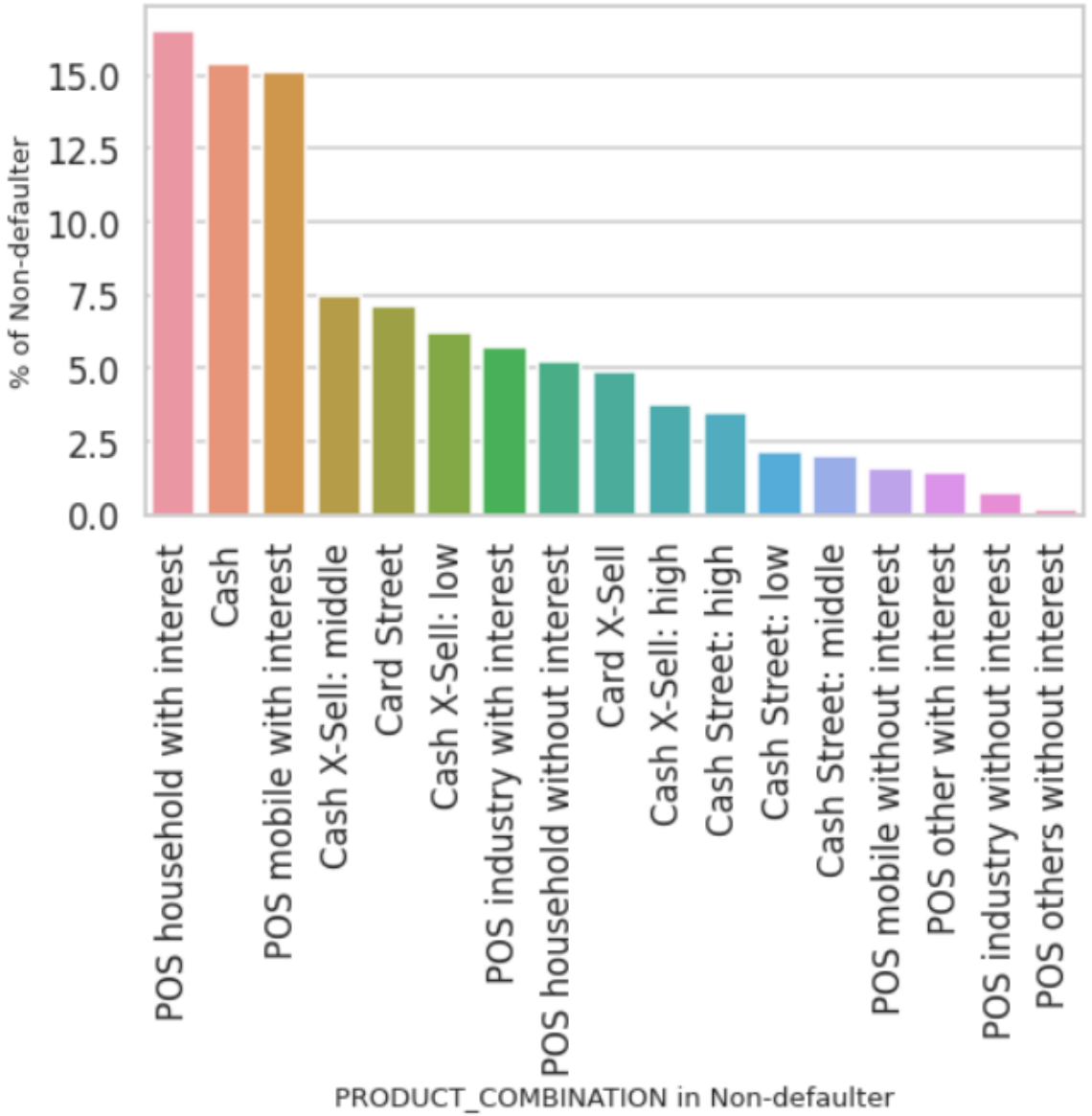

NAME_INCOME_TYPE for Non-Defaulters (left) and Defaulters(Right)
% of Non-defaulter
0
10
20
30
40
50
60
Working
Commercial associate
State servant
Pensioner
Student
Businessman
Maternity leave
NAME_INCOME_TYPE in Non-defaulter
% of Defaulters
0
20
40
60
Working
Commercial associate
State servant
Pensioner
Student
Businessman
Maternity leave
NAME_INCOME_TYPE in defaulters

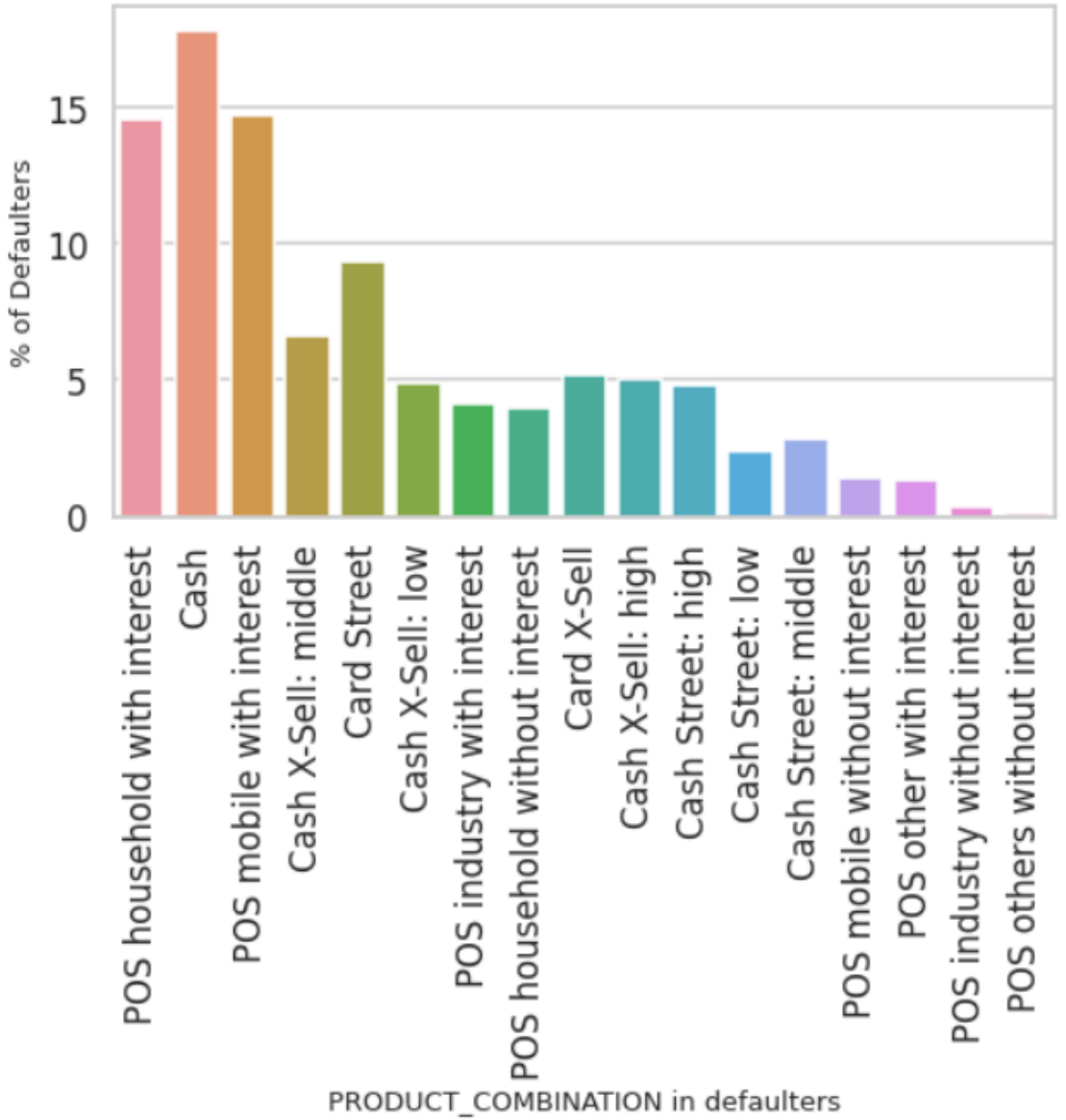
NAME_FAMILY_STATUS for Non-Defaulters (left) and Defaulters(Right)



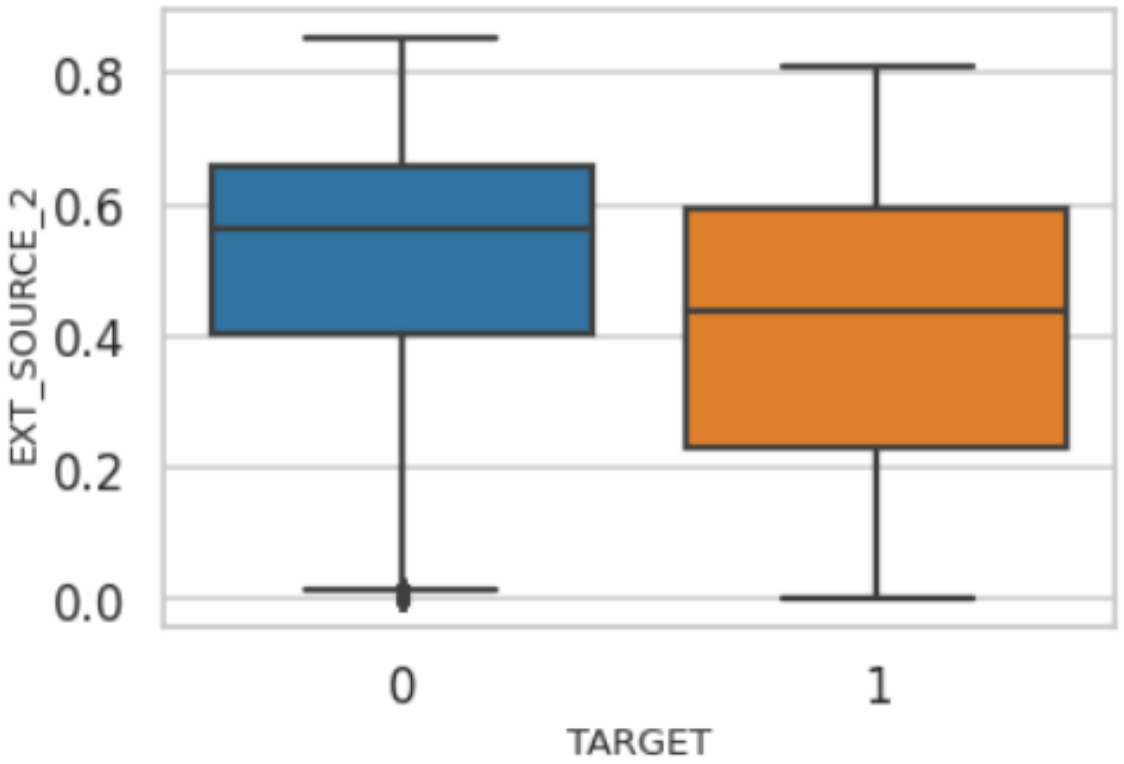


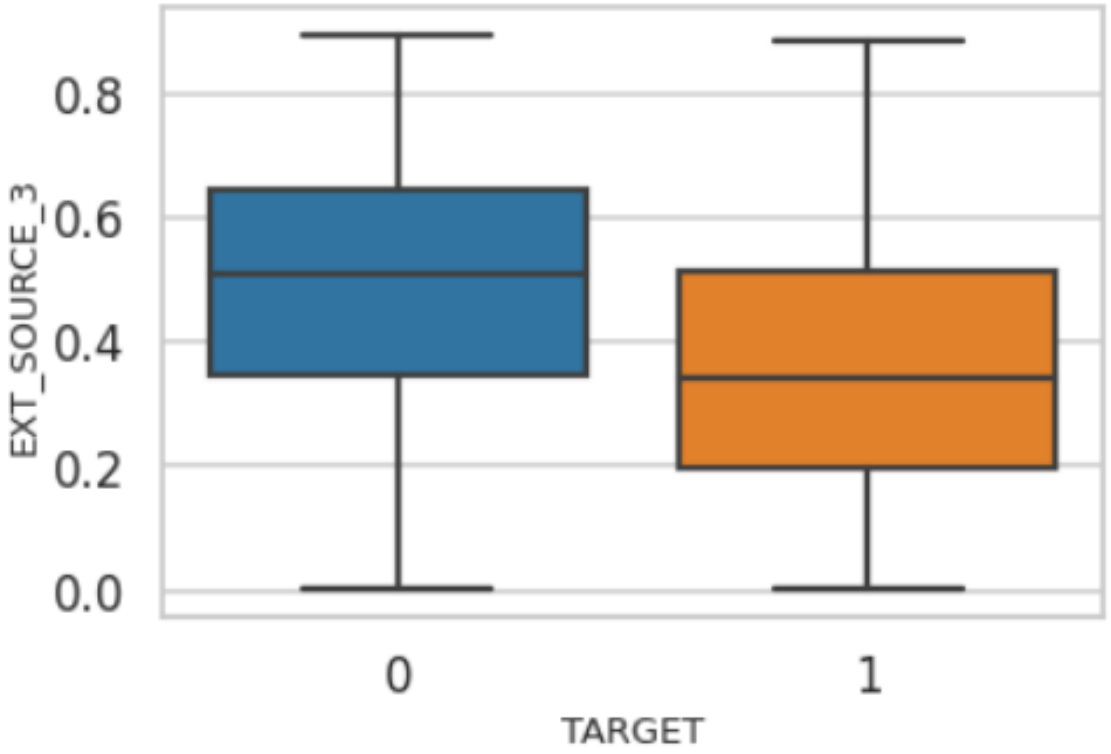
PRODUCT_COMBINATION for Non-Defaulters (left) and Defaulters(Right)

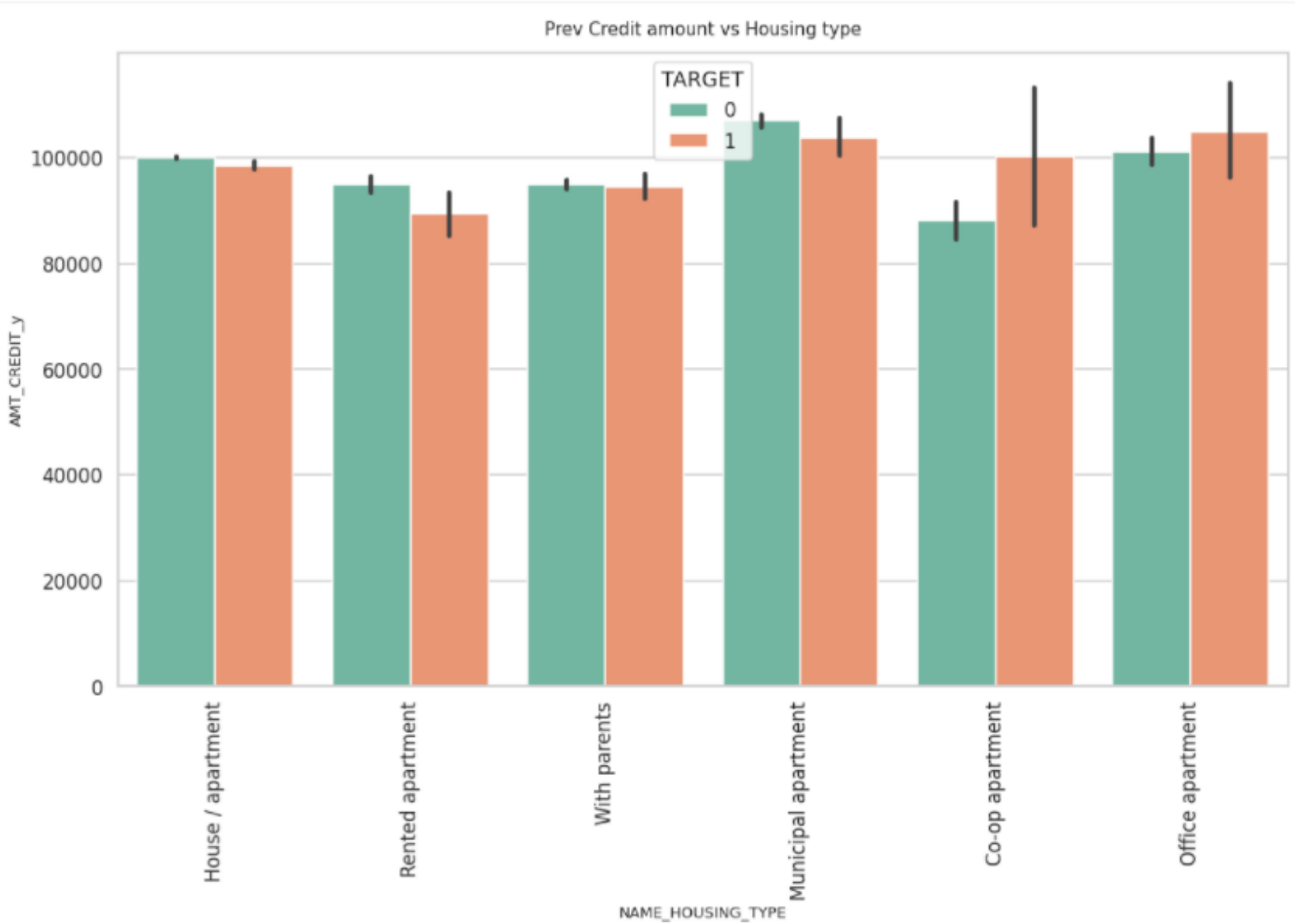




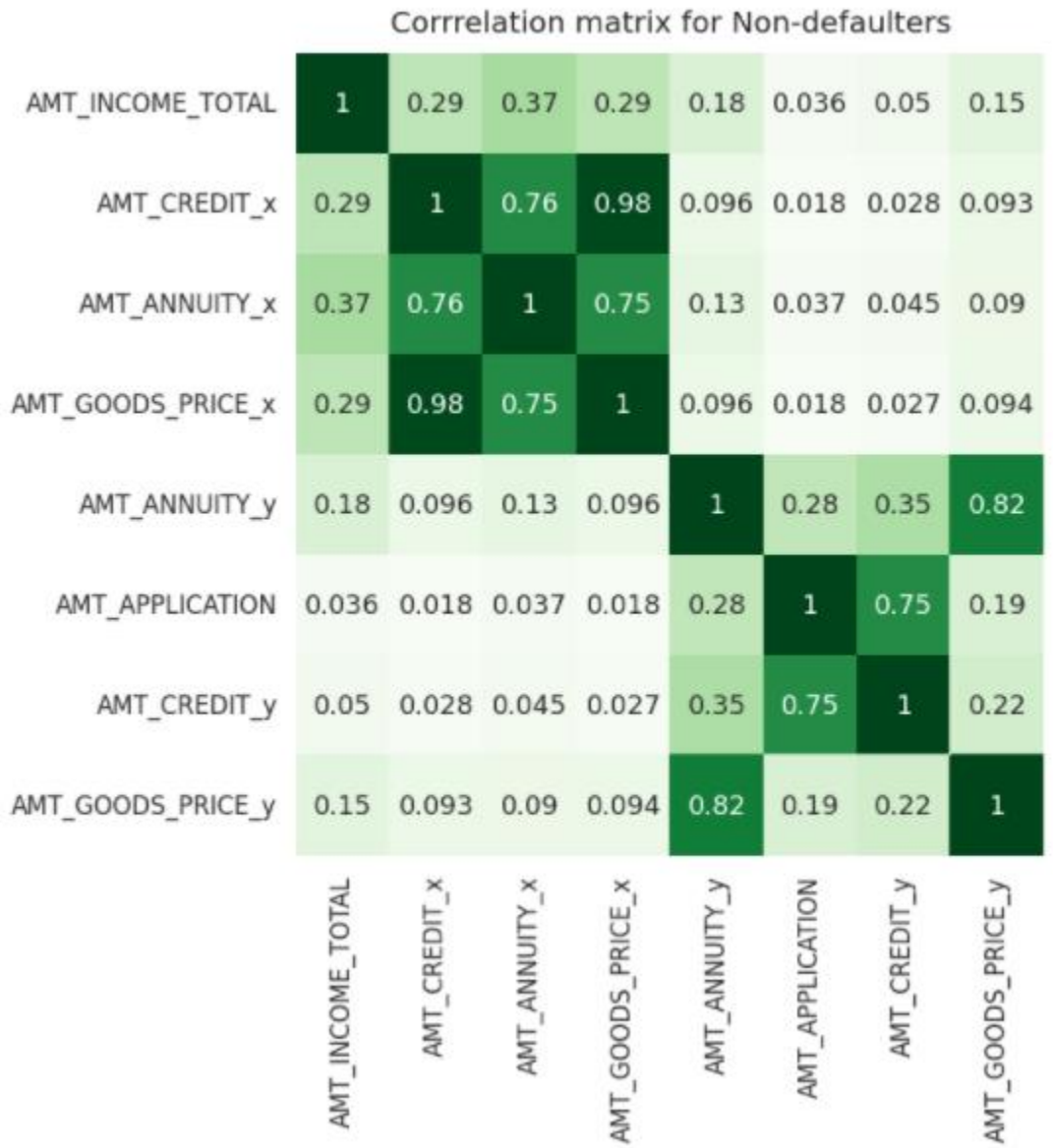
SEGMENTED NUMERICAL ANALYSIS

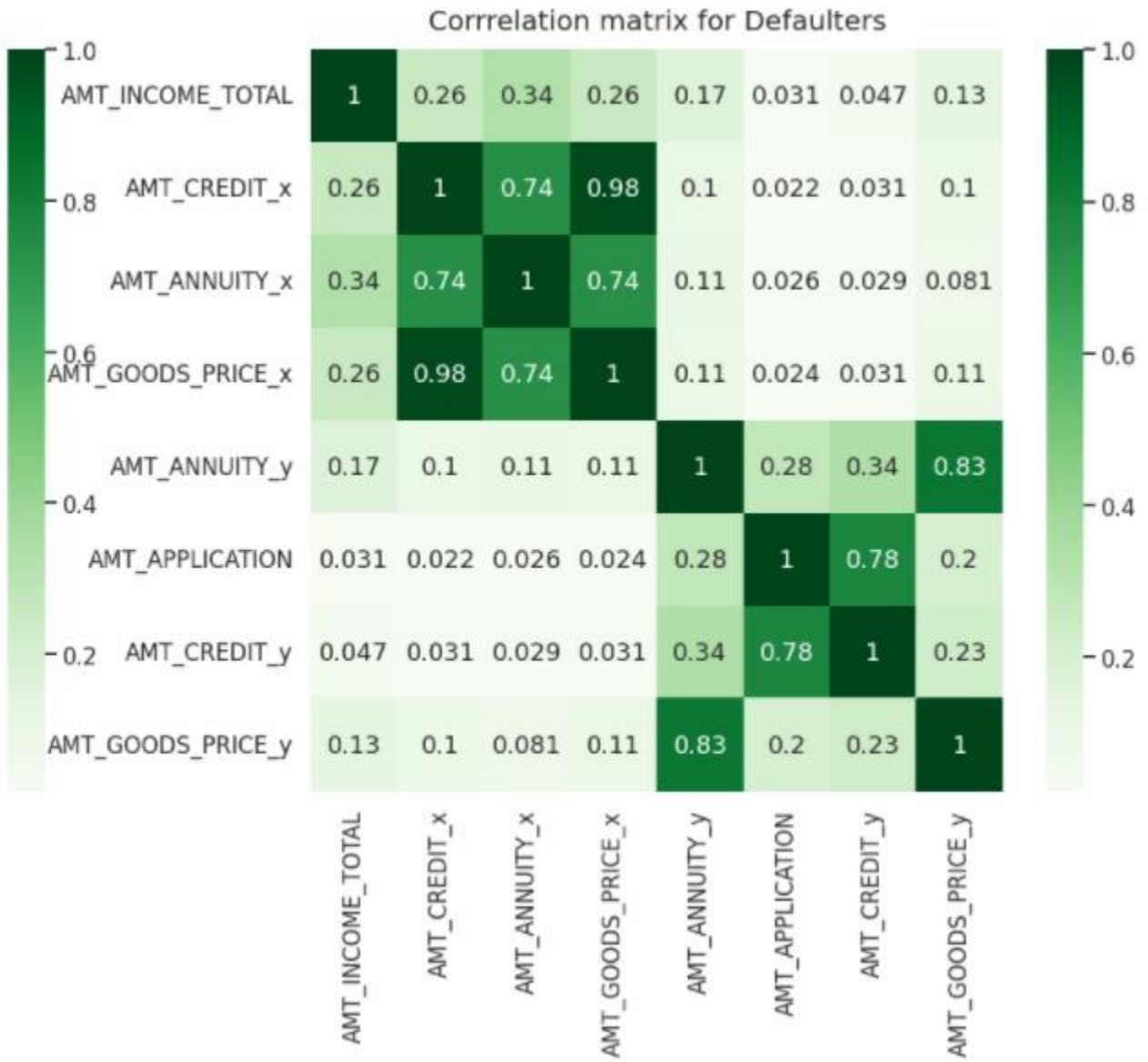





Prev Credit amount vs Housing type
TARGET
0
1
AMT_CREDIT_y
100000
80000
60000
40000
20000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Co-op apartment
Office apartment
NAME_HOUSING_TYPE

CONTINUOUS VARIABLES CORRELATION





OBSERVATIONS

Observation from Heat Map

•AMT CREDIT is closely associated with AMT GOODS PRICE. i.e. the loan credit amount is proportional to the price of the goods for which the loan is sought.

•AMT CREDIT and AMT ANNUITY have a strong relationship. (The loan annuity and the loan credit amount are exactly proportionate.)

•AMT GOODS PRICE and AMT ANNUITY have a strong relationship.

Observation from correlation matrix

•With the Categorical columns removed, we can see that the variables below are substantially connected for Non-defaulters.

1.AMT_ANNUITY_x & AMT_CREDIT_x

2.AMT_GOODS_PRICE_y & AMT_ANNUITY_y

3.AMT_GOODS_PRICE_x & AMT_CREDIT_x

CONCLUSIONS

- Banks should consider lending to those who have a high credit score from external scoring firms.
- Loan requests for the intention of 'repairs' are having more trouble making timely payments.
- The percentage of defaulters among "labourers" and "drivers" is much higher. And "sales personnel," "core staff," "security staff," "kitchen staff," "cleaning staff," and "private sector workers" are all moderately high.
- Because there are many defaulters with this combination, the bank should scrutinise their application when giving loans to persons living in co-op flats in Region 3.
- In the case of pensioners, the percentage of defaulters is much lower, and in the case of state employees, it is moderately lower. Before authorising the loan, other categories must be thoroughly examined.
- Clients with a lower secondary education are the most likely to fail when their previous loans are revoked or refused

THANK YOU