EDA Case Study

Group Members:
- Mamta Mittal
- Sneha Boora

Abstract

Business Understanding:

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. 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.

Business Object:

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.

Goals of Data Analysis:

To understand the driving factors for loan default, identifying patterns which indicate if the client has difficulty paying installments using EDA.

Data Understanding

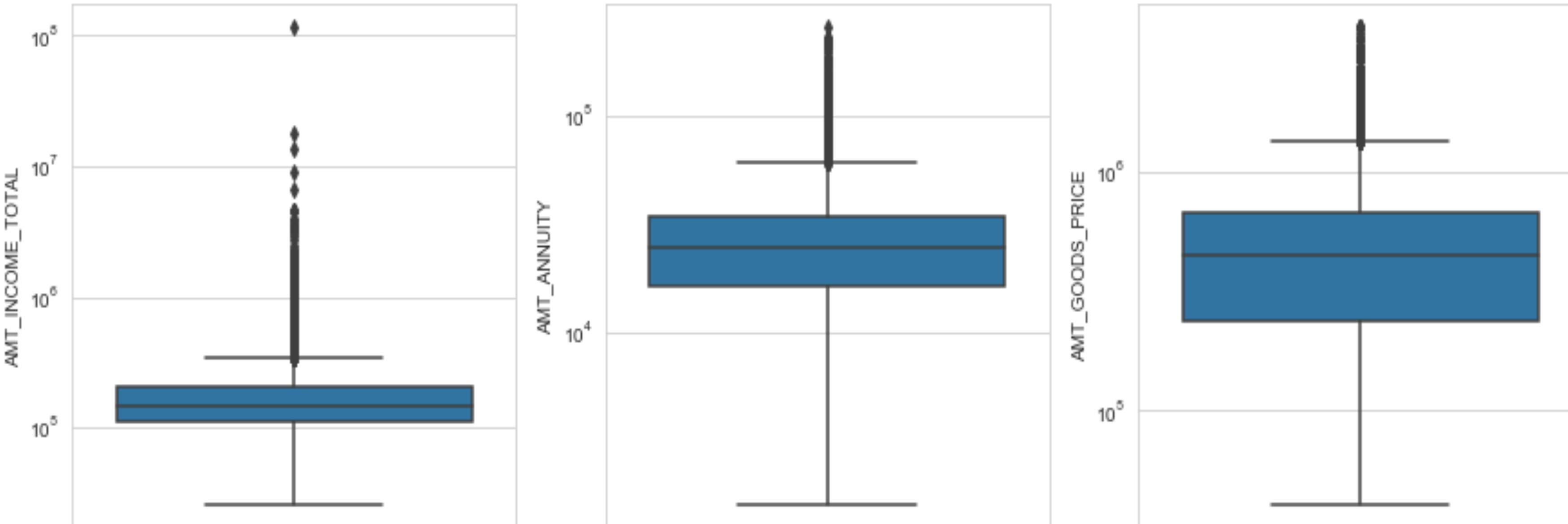
*1. '**application_data.csv**'* contains all the information of the client at the time of application.
The data is about whether a **client has payment difficulties.**

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

Approach

1. Read the dataset
2. Clean the dataset(Drop unwanted rows and columns, check data type,)
3. Identify the outliers
4. Analysis the DF_APP_DATA and get the insight from the analysis of categorical variable and numeric variable by plotting required plots(univariate and Bivariate)
5. Analysis the DF_PREV_APP and get the insight from the analysis of categorical variable and numeric variable by plotting required plots(univariate and Bivariate)
6. Merge both the data set and analyse the impact of previous application data on the Current application by checking it against TARGET variable.
7. Find the major factors for Bank which can help in Credit risk analysis.

Outliers



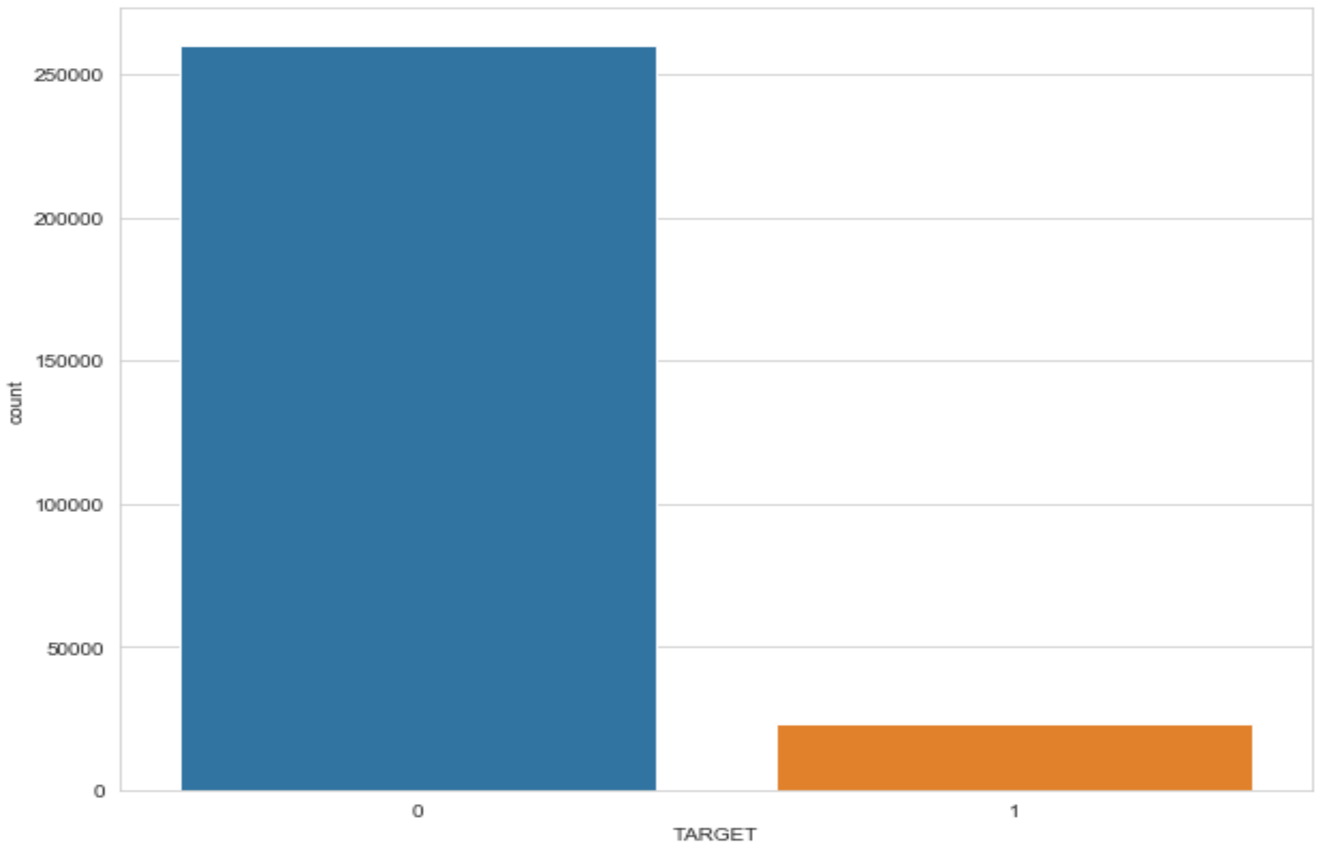
Observation:

•From the above box plot for AMT_INCOME_TOTAL/AMT_ANNUITY/AMT_GOODS_PRICE we can observe that there are lots of outliers in the data.

•Outliers are the extremely High and low values. Due to which data is skewed.

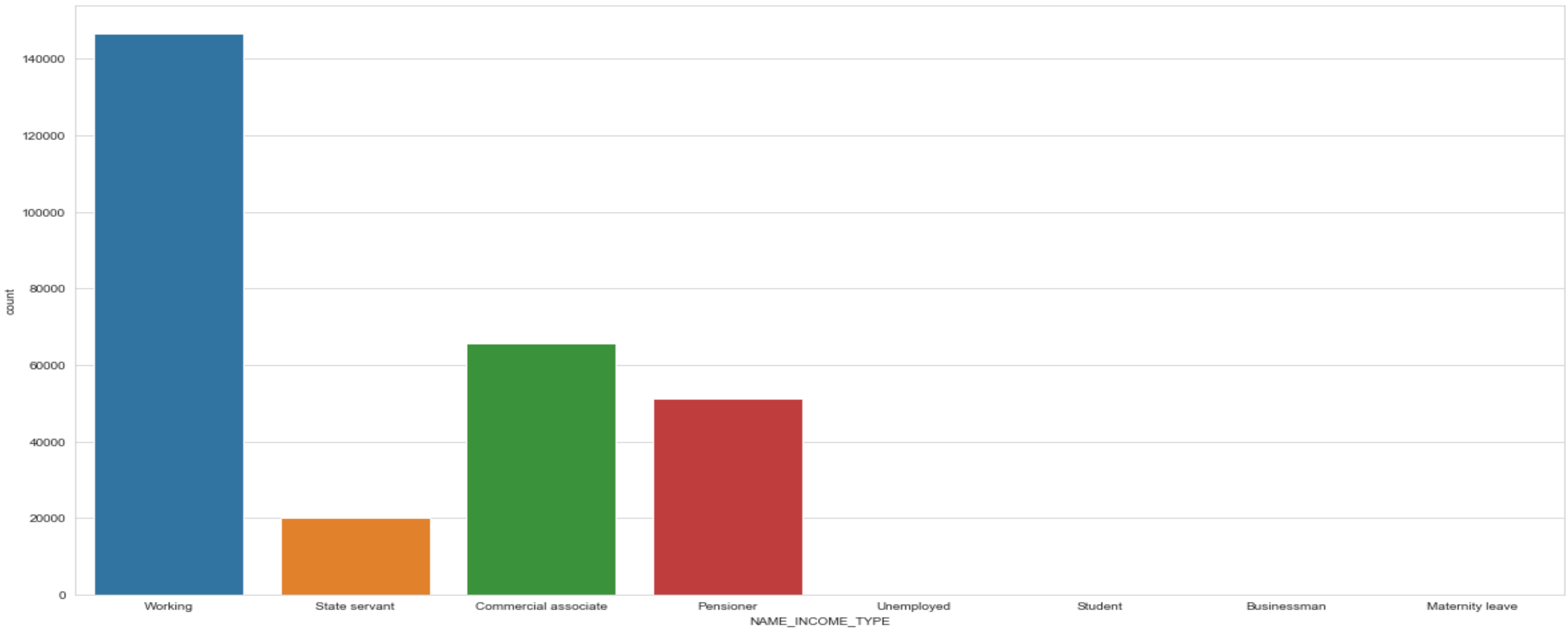
•Using DF_APP_DATA_NUM_COL.describe() metric we saw which columns have outliers by checking difference in mean, median values.

Data Imbalance of ‘TARGET’ variable

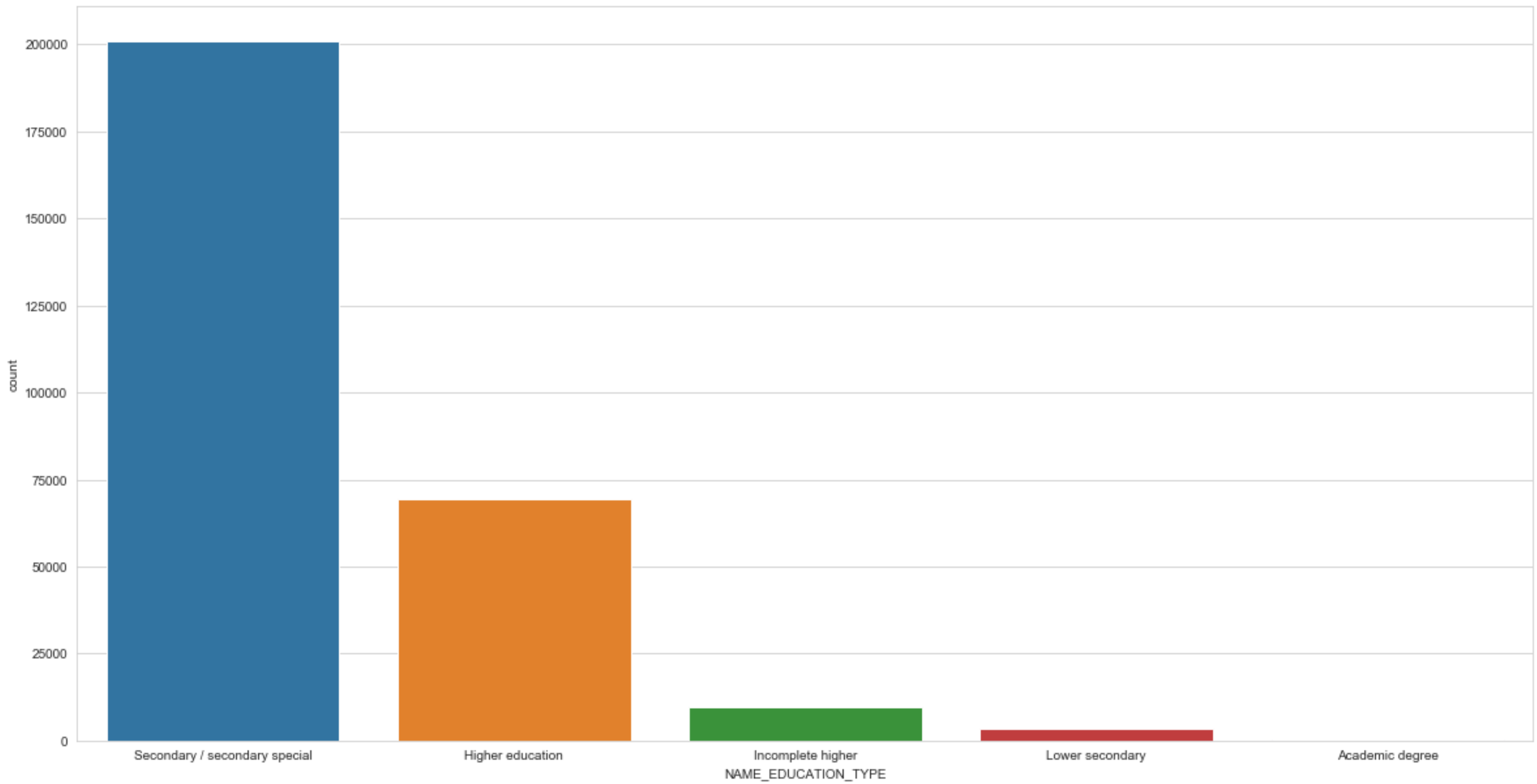


Ratio of imbalance is found out to be 11.375961202376791

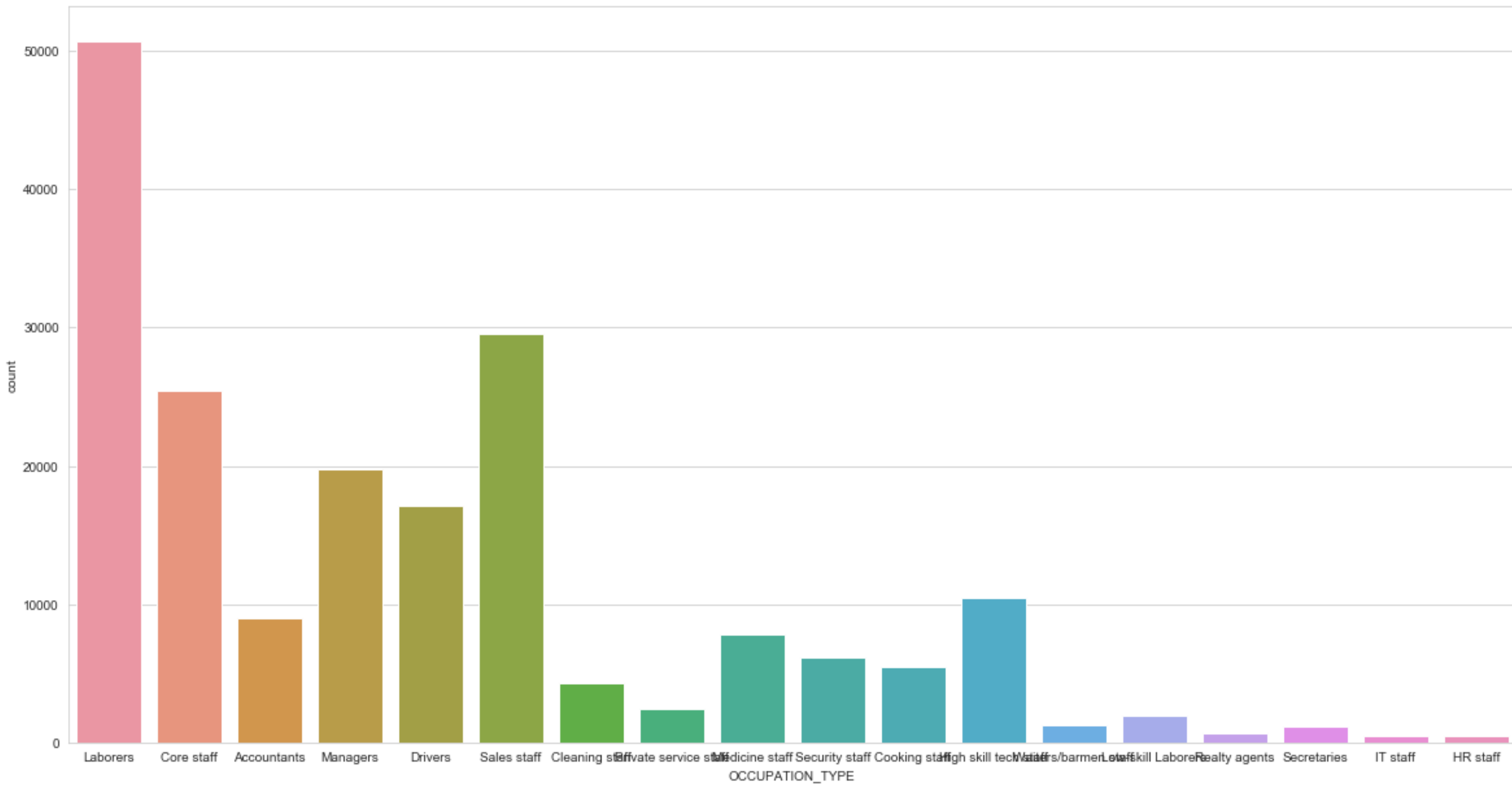
Uni-variate Analysis



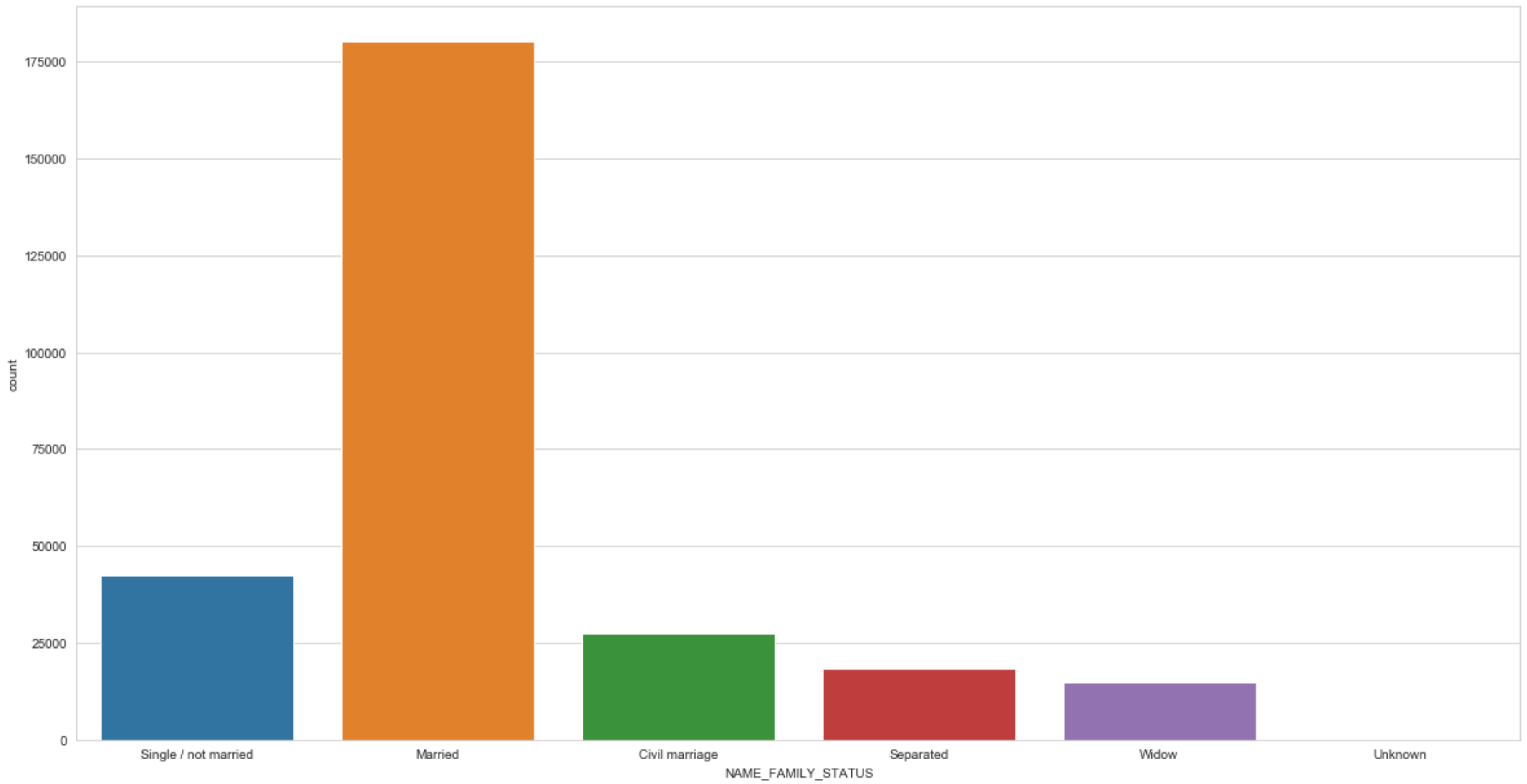
Observation (NAME_INCOME_TYPE): No. of people opting for Loan for income categories Working/ State Servent /Commercial Associate/Pensioner are significant as compare to other categories



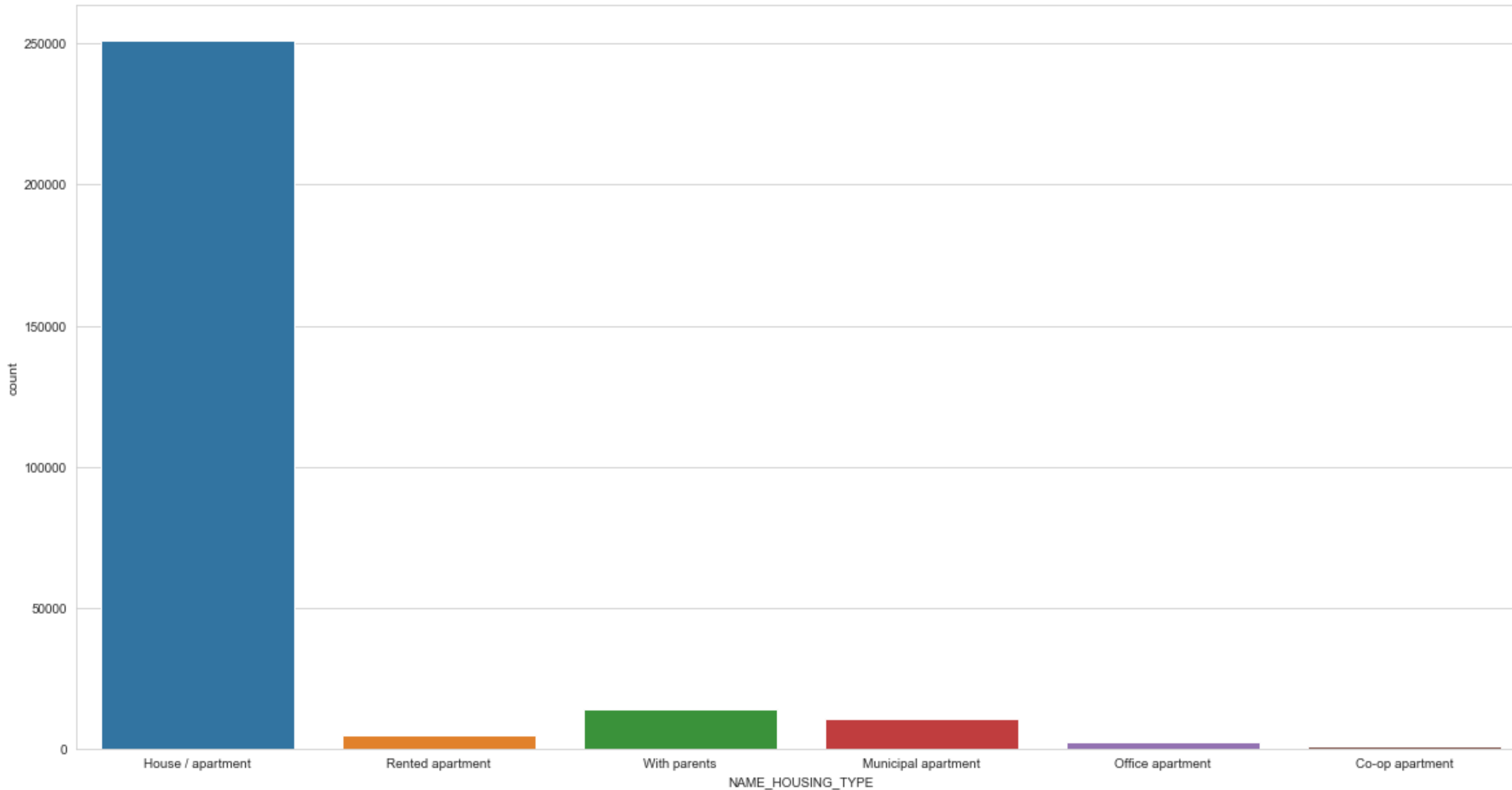
Observation(NAME_EDUCATION_TYPE) : No. of people opting for Loan for Education type categories Seconary/Higher are significant as compare to other categories.



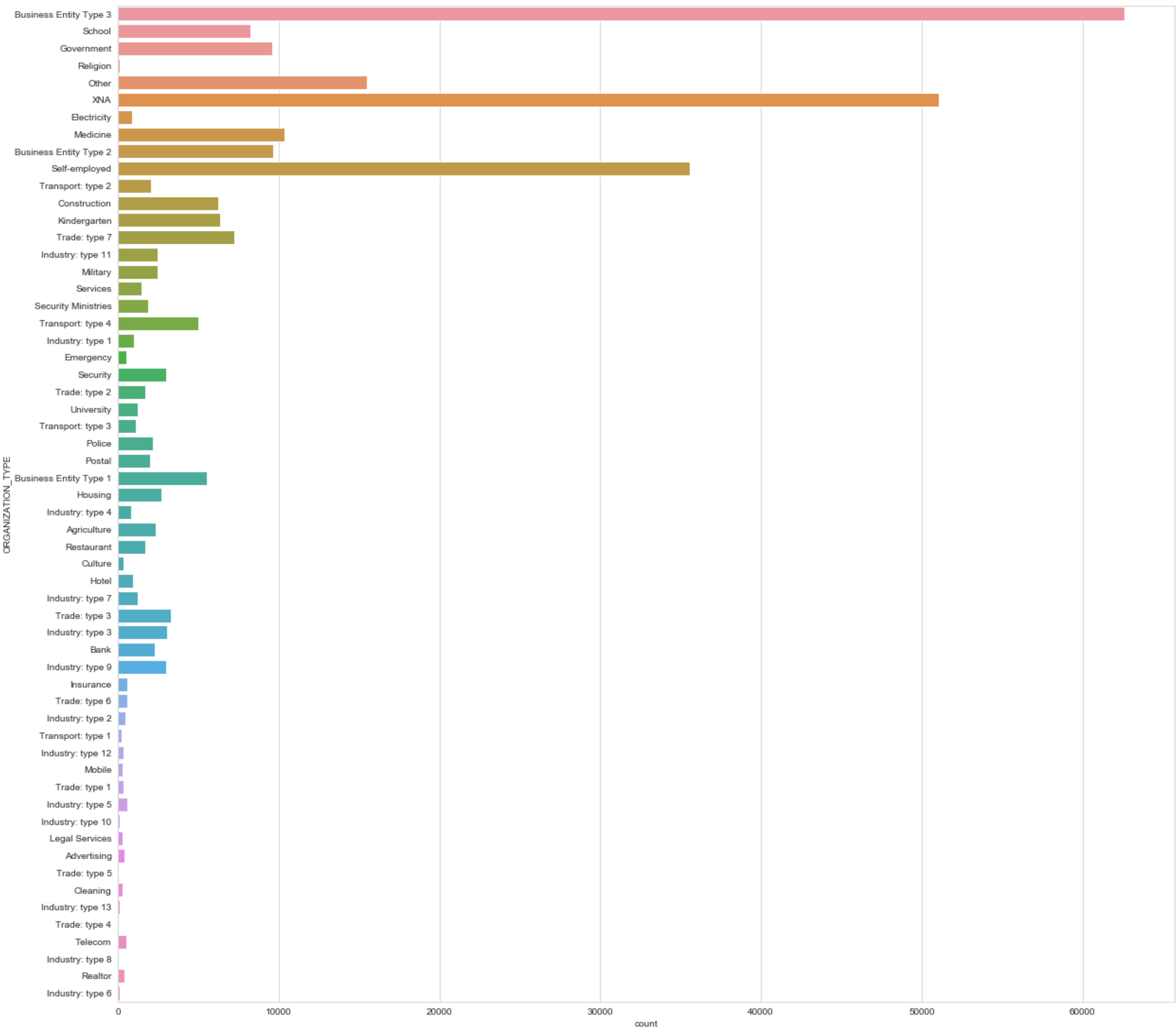
Observation(OCCUPATION_TYPE) : No. of people opting for Loan for occupation categories Laborers/Core Staff/Sales Staff are significant as compare to other categories.



Observation(NAME_FAMILY_STATUS) : No. of Married people applying for loan is much higher then other categories.

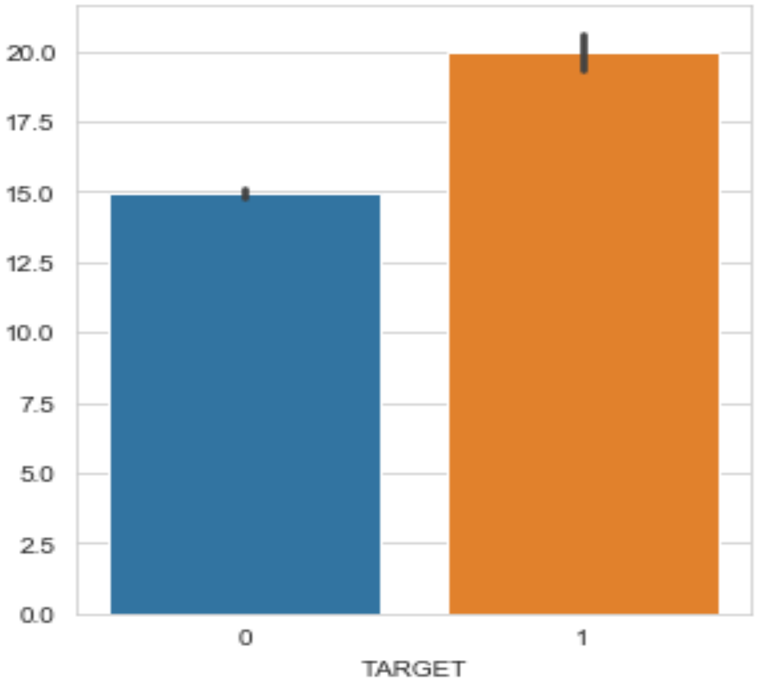


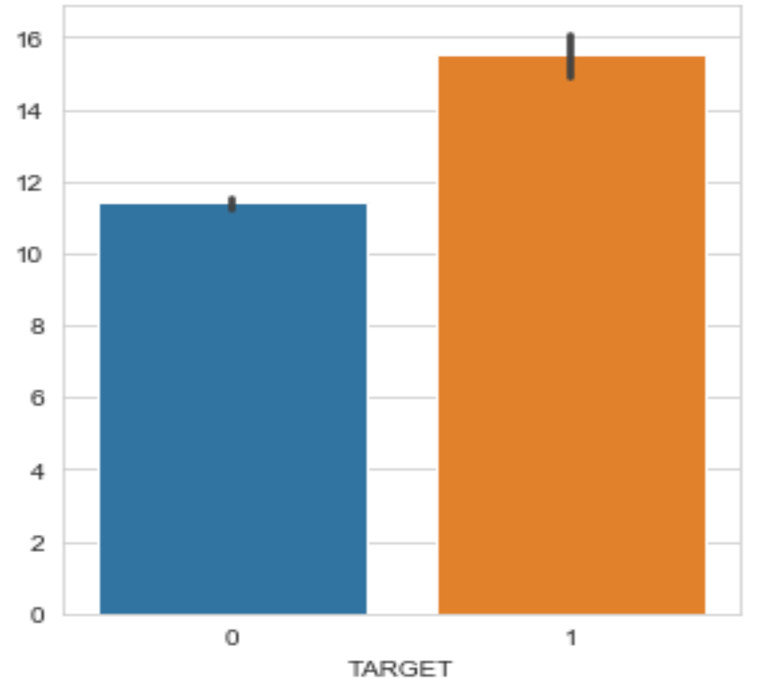
Observation(NAME_HOUSING_TYPE) : No. of people staying in Apartment are applying for loan is much higher then other categories.



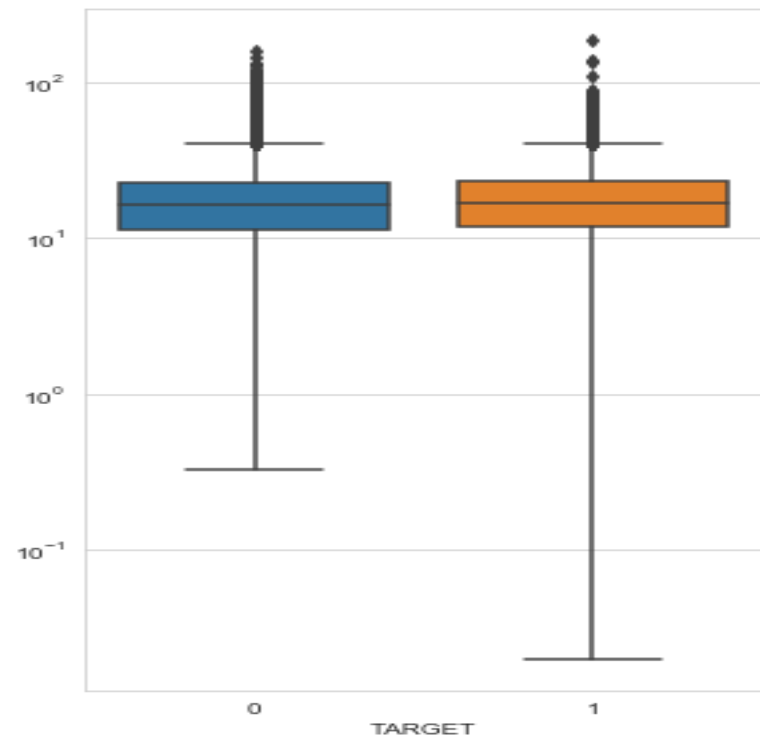
Observation(ORGANIZATION_TYPE) : people in this category "Self Employed and Business entity type 3 " Apply for more loan

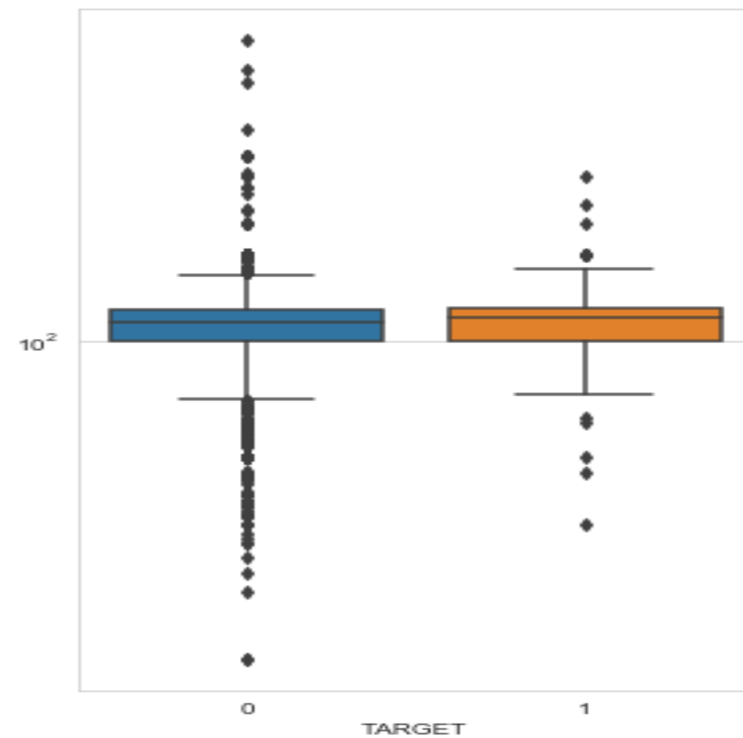
Bi-variate Analysis



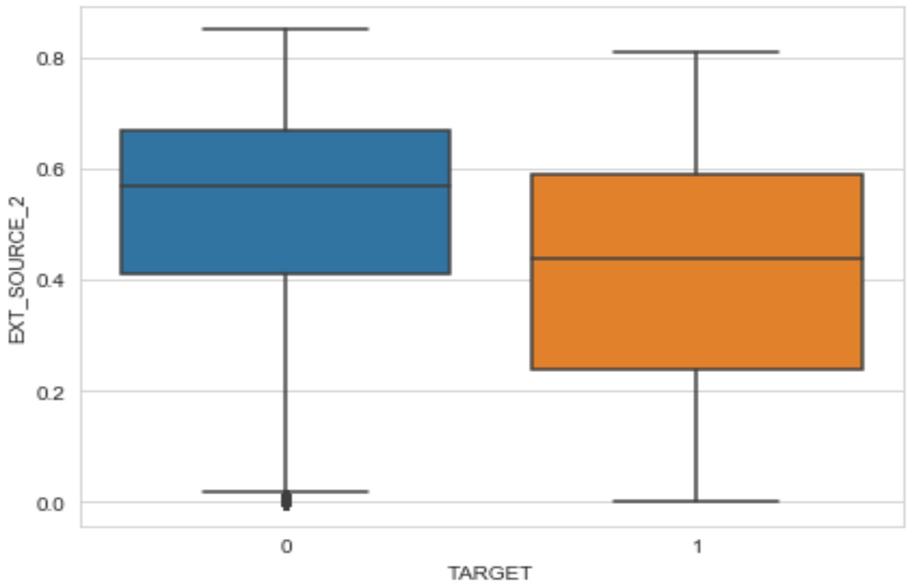


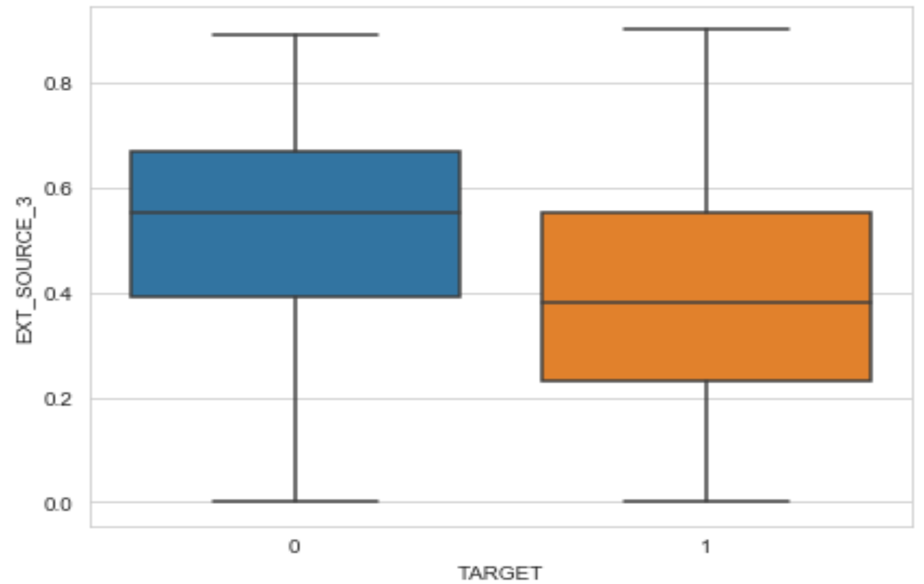
Observation (OBS_30_CNT_SOCIAL_CIRCLE/OBS_60_CNT_SOCIAL_CIRCLE):
As there are many people in the Social circle who are defaulter , Chance of defaulting a person increase because social circle has an impact on persons habits.
Hence this is one of the factor which influence the default case

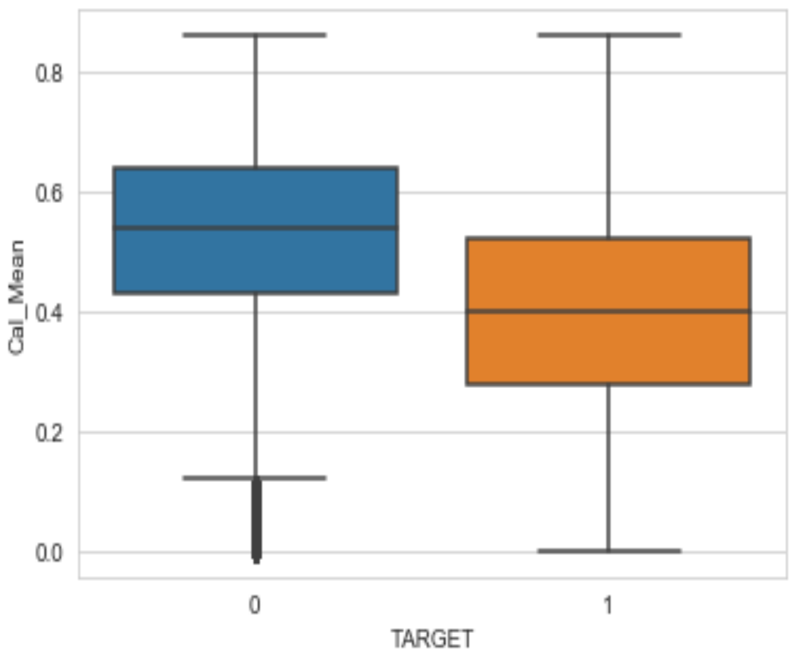




Observation:
(AMT_ANNUITY/AMT_INCOME_TOTAL/AMT_CREDIT/AMT_GOODS_PRICE) : All These 4 columns does not seem to have an Impact on default cases



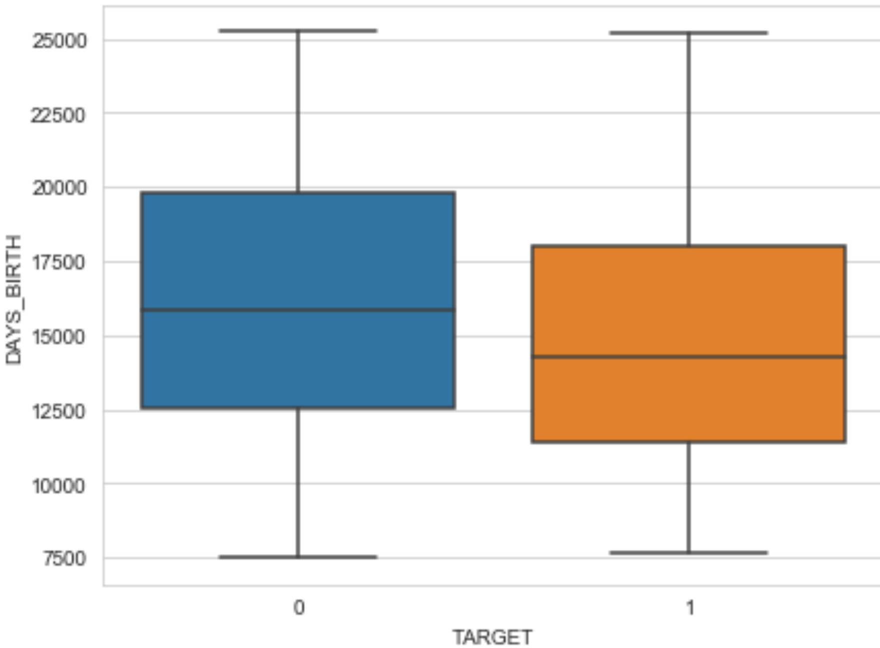


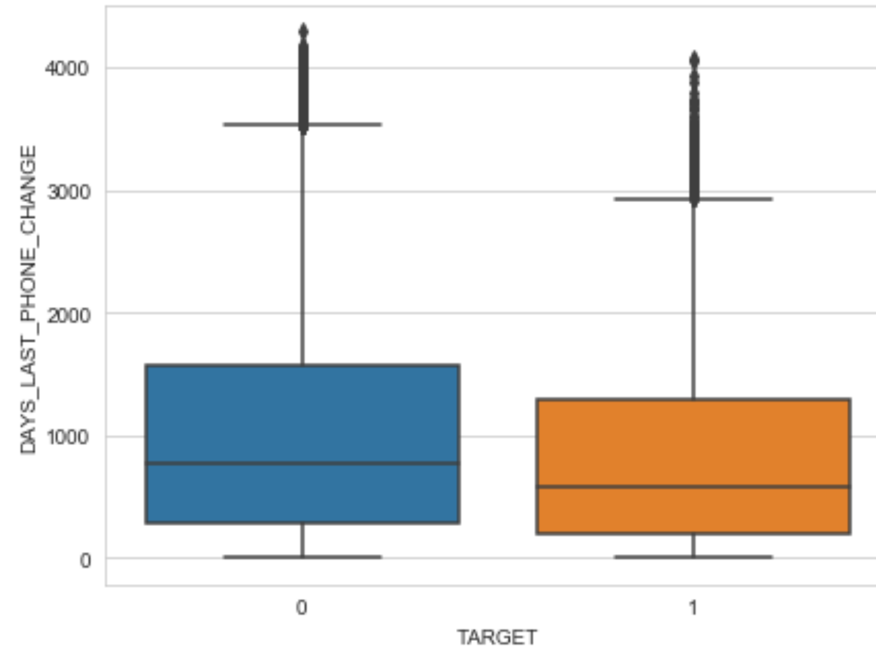


Observation:

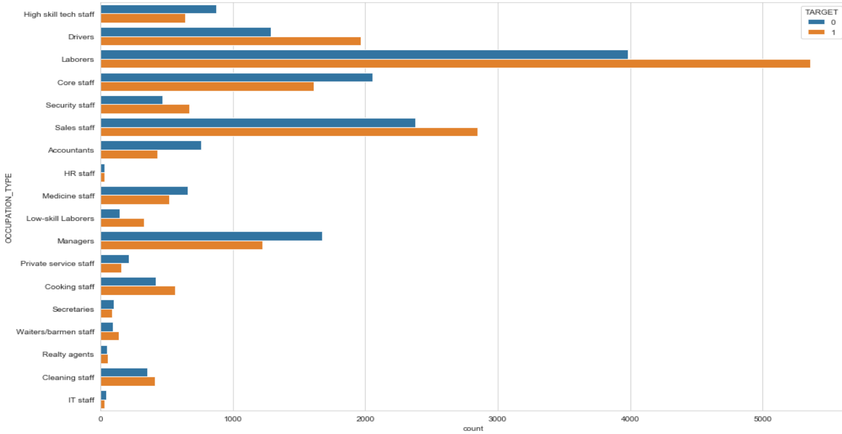
From All the above plots we can observe that EXT_SOURCE_1/2/3 has a huge impact on defaulter.

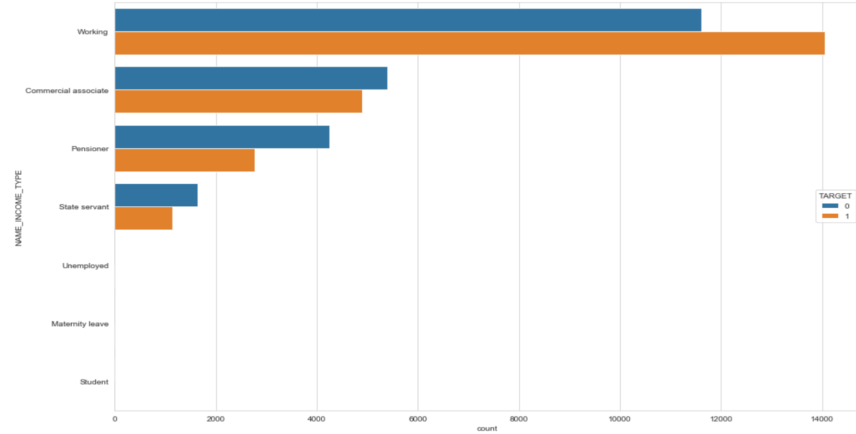
As EXT_SOURCE is the credit rating provided by the third party. Based on the Credit rating Bank may decide if the loan can be given to the customer. Customer with lower credit rating are more likely to default.

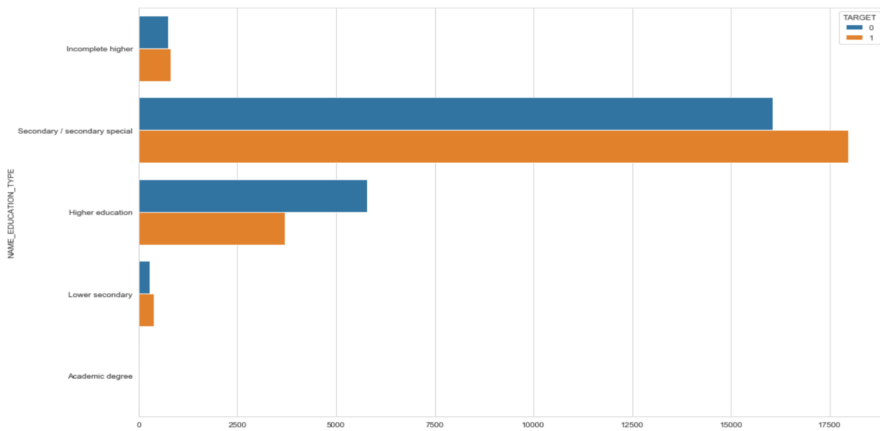


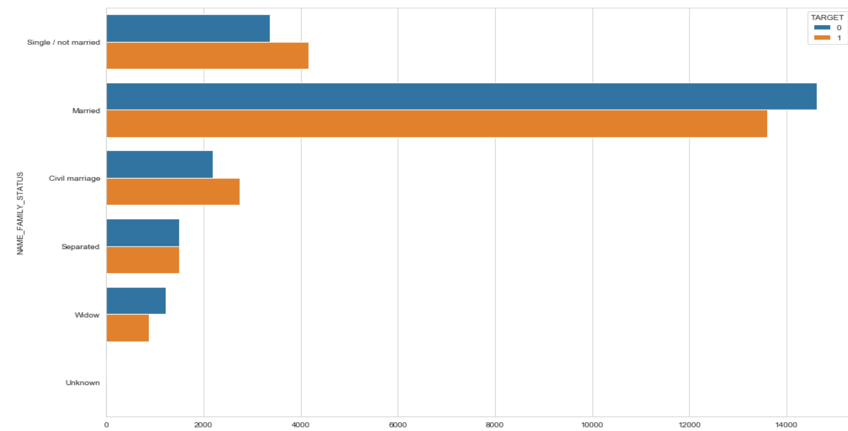


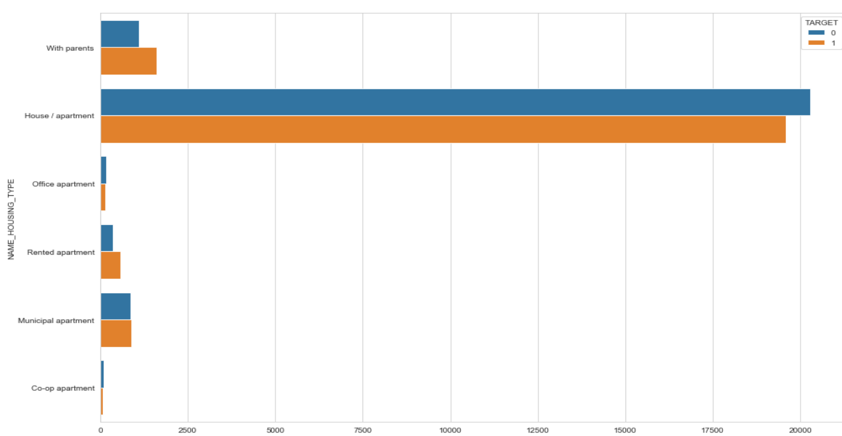
Observations(DAYS_BIRTH / DAYS_LAST_PHONE_CHANGE**):**
From All the above plots we can observe that DAYS_BIRTH : Higher the age(may be Old age), more likely the person will default
DAYS_LAST_PHONE_CHANGE: if the phone number has been changed recently, chances of a person to default increases
We may consider these 2 columns as itmay have impact on defaulter.

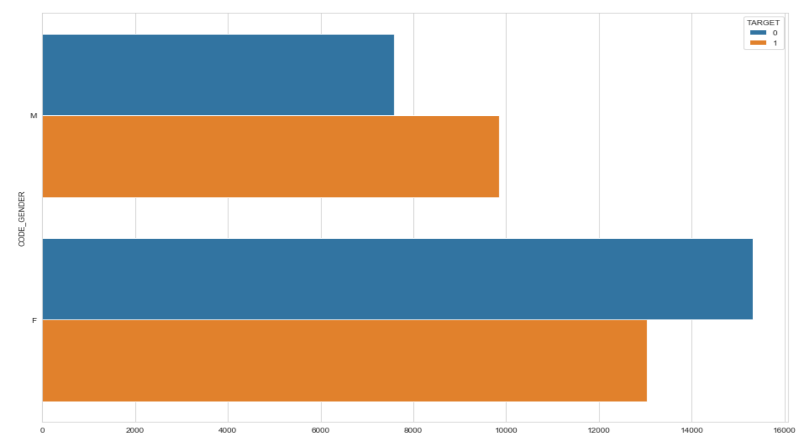

OCCUPATION_TYPE
High skill tech staff
Drivers
Laborers
Core staff
Security staff
Sales staff
Accountants
HR staff
Medicine staff
Low-skill Laborers
Managers
Private service staff
Cooking staff
Secretaries
Waiters/barmen staff
Realty agents
Cleaning staff
IT staff
TARGET
0
1
0
1000
2000
3000
4000
5000
count


NAME_INCOME_TYPE
Working
Commercial associate
Pensioner
State servant
Unemployed
Maternity leave
Student
TARGET
0
1
0
2000
4000
6000
8000
10000
12000
14000
count


NAME_EDUCATION_TYPE
Incomplete higher
Secondary / secondary special
Higher education
Lower secondary
Academic degree
TARGET
0
1
0
2500
5000
7500
10000
12500
15000
17500
count


NAME_FAMILY_STATUS
Single / not married
Married
Civil marriage
Separated
Widow
Unknown
TARGET
0
1
0
2000
4000
6000
8000
10000
12000
14000
count

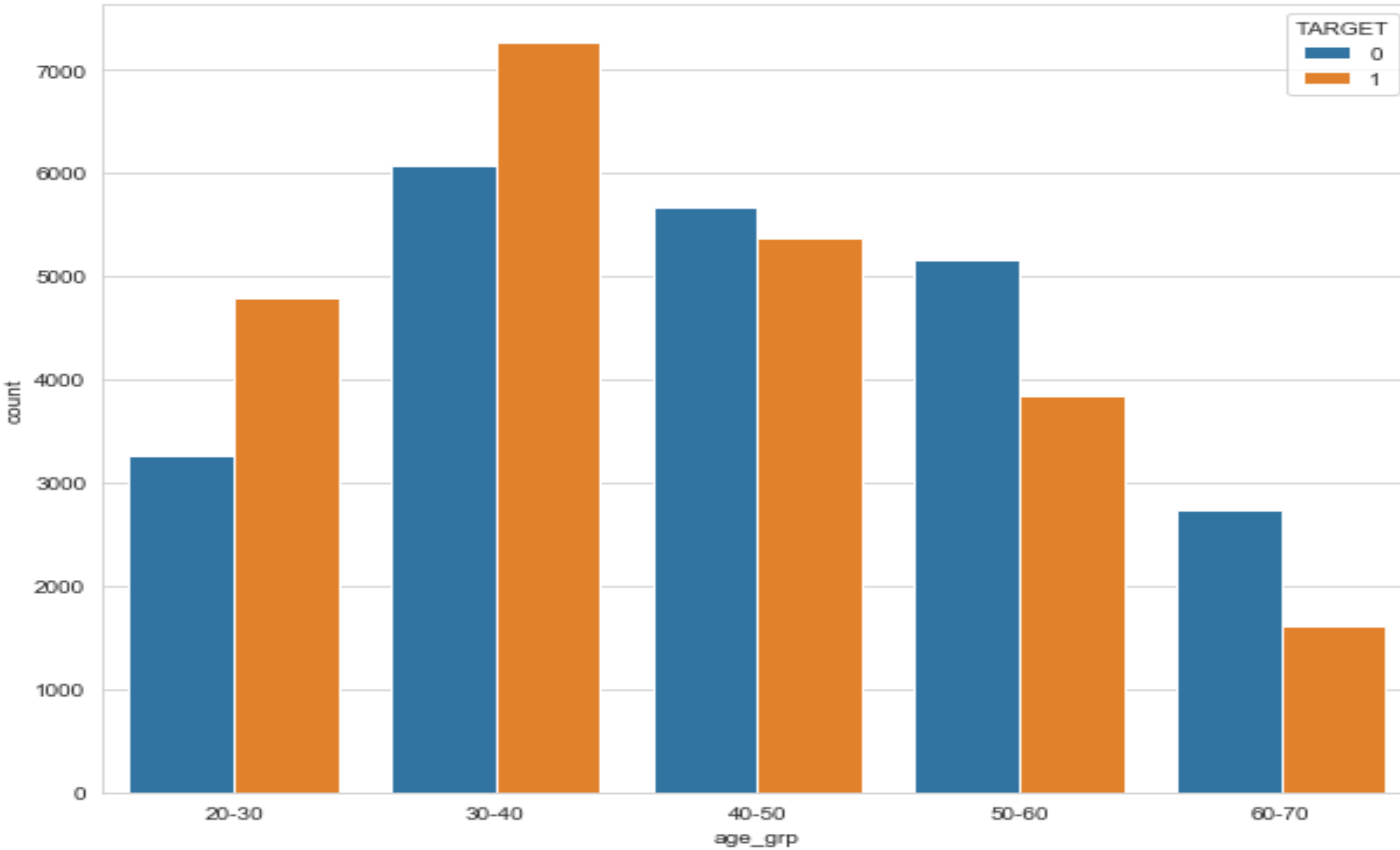

NAME_HOUSING_TYPE
With parents
House / apartment
Office apartment
Rented apartment
Municipal apartment
Co-op apartment
TARGET
0
1
0
2500
5000
7500
10000
12500
15000
17500
20000
count


CODE_GENDER
M
F
TARGET
0
1
0
2000
4000
6000
8000
10000
12000
14000
16000
count

Observation:
Following fields are significant

- CODE_GENDER
  - 'FeMale' having much more % population with NO payment difficulty
- NAME_INCOME_TYPE
  - with values "State servant" and "Pensioner" having much more % population with no payment difficulty
- NAME_EDUCATION_TYPE
  - "Higher education" having much more % population with NO payment difficulty
  - "Lower secondary" and "Secondary" having much more % population with payment difficulty
- NAME_HOUSING_TYPE
  - "With parents" and "Rented apartment" having much more % population with payment difficulty
- OCCUPATION_TYPE
  - "Security staff", "Laborers", "Cooking staff", "Drivers" having much more % in population with payment difficulty
  - "Core staff","High skill tech staff","Accountants" having much more % in population with NO payment difficulty
- ORGANIZATION_TYPE
  - "Construction", "Industry: type 3" having much more % in population with payment difficulty
  - "School" having much more % in population with NO payment difficulty

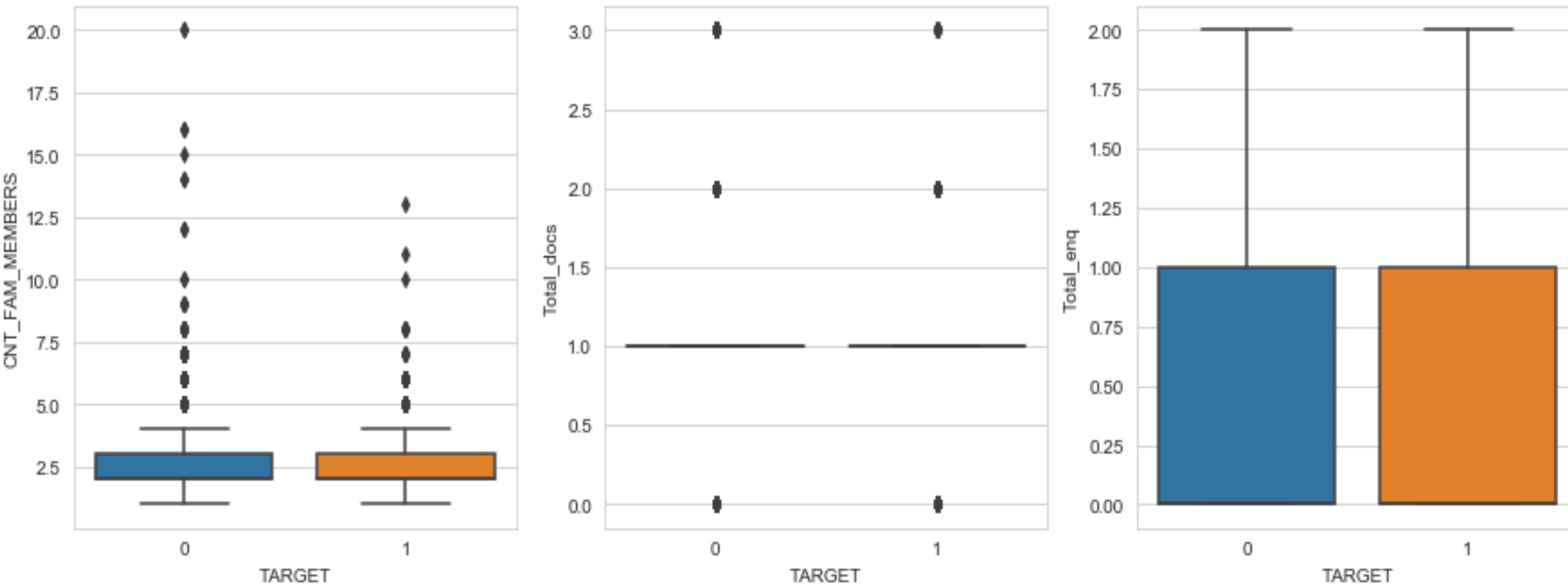
Analysis of ‘age_grp’ variable which is derived by binning DAYS_OF_BIRTH



Observation:

age_grp: 20-30/30-40 having much more % population with payment difficulty.

Few other variables:



Observation: NO Impact

CNT_FAM_MEMBERS/Total_docs/Total_enq/DAYS_REGISTRATION/DAYS_EMPLOYED/REG_REGION_NOT_LIVE_REGION

Previous application - Approach

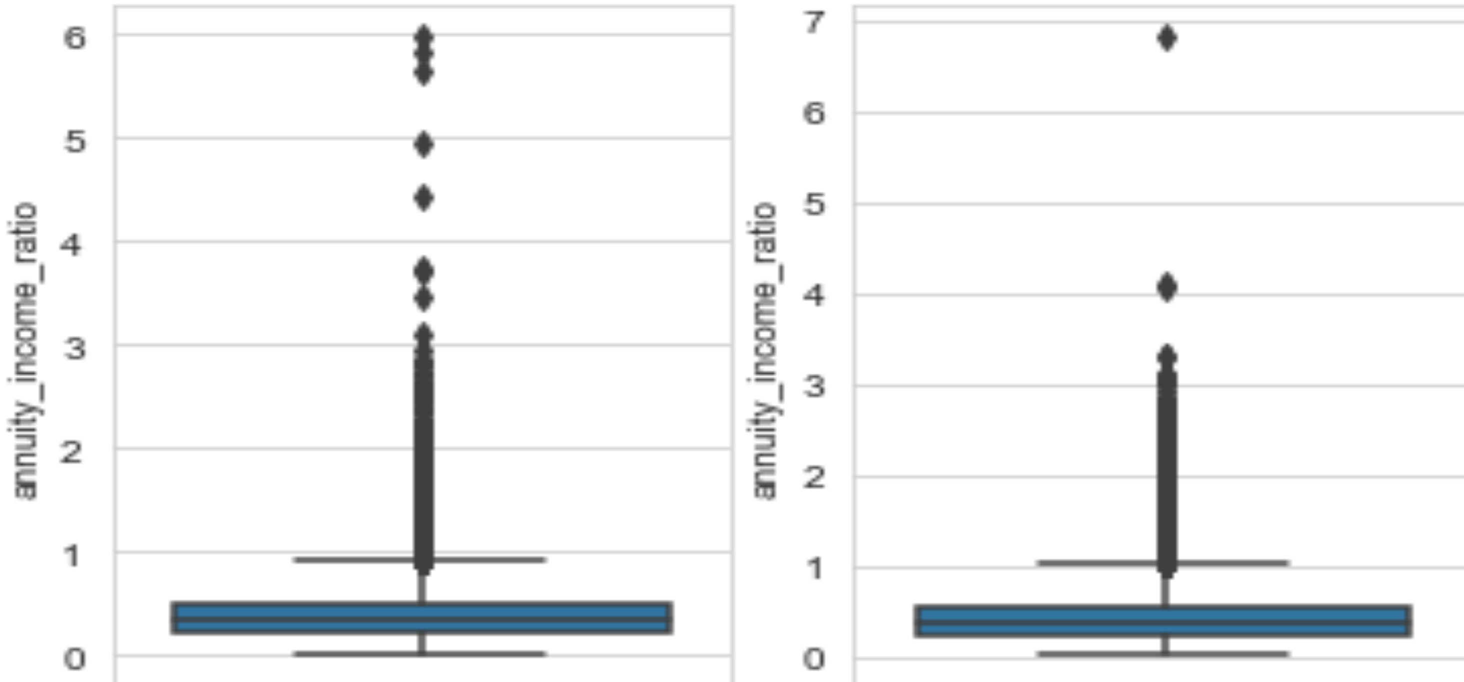
- Sum the annuity for all previous applications where the payment is still due
    - filter where DAYS_TERMINATION is still +ve i.e. loan is still active

- Add the outstanding annuity to current application annuity

- Find the annuity to income ratio

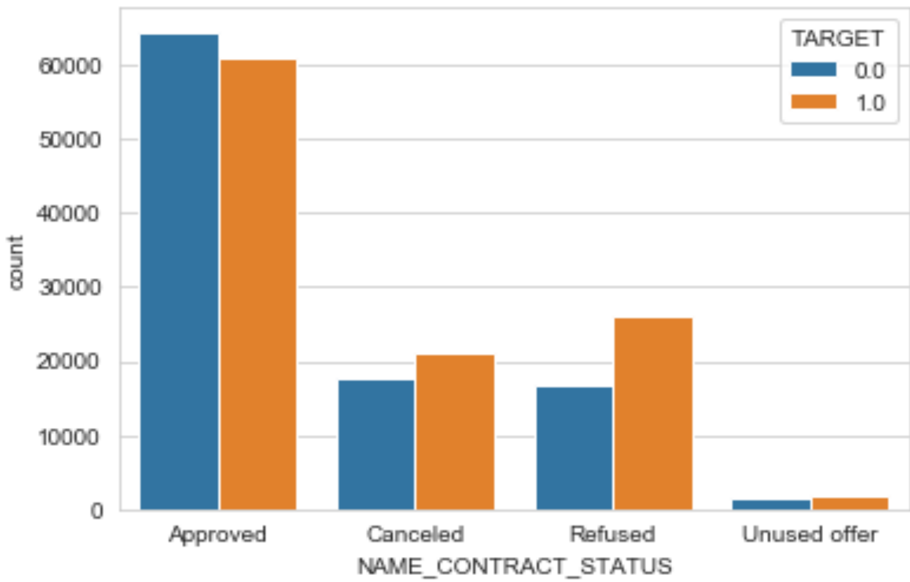
- Compare between Target 0 and target 1 population and see if any significant difference

Further Analysis

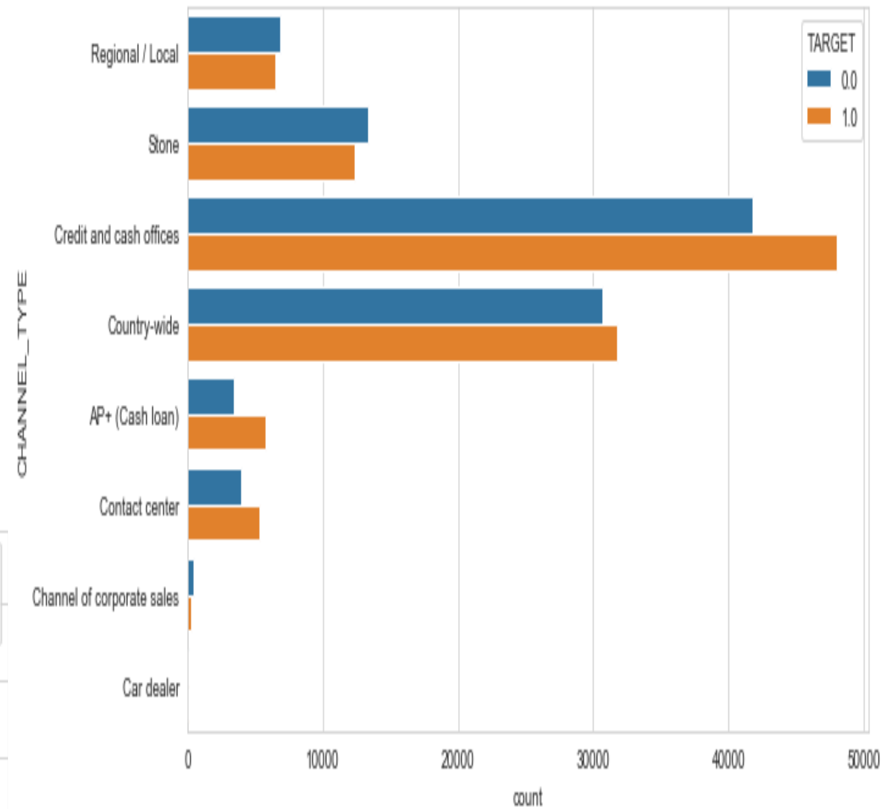


Conclusion : No Significance

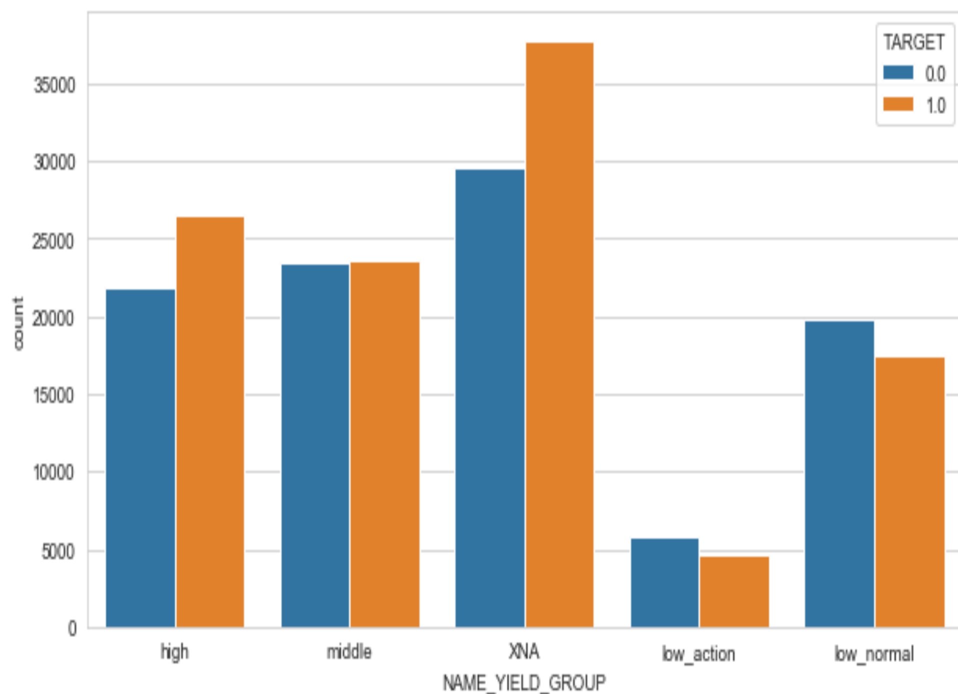
As we can see the mean, median and inter-quartile ranges are almost similar between Target 0 and Target1. So it can be concluded that Annuity and Income does not have any significant difference for Target 0 and Target 1 population



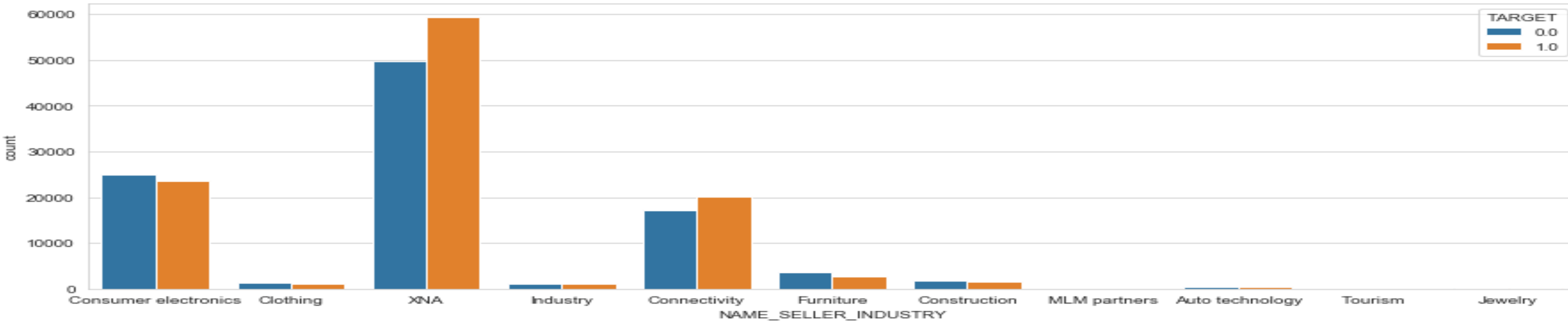
TARGET
0.0
1.0
count
60000
50000
40000
30000
20000
10000
0
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

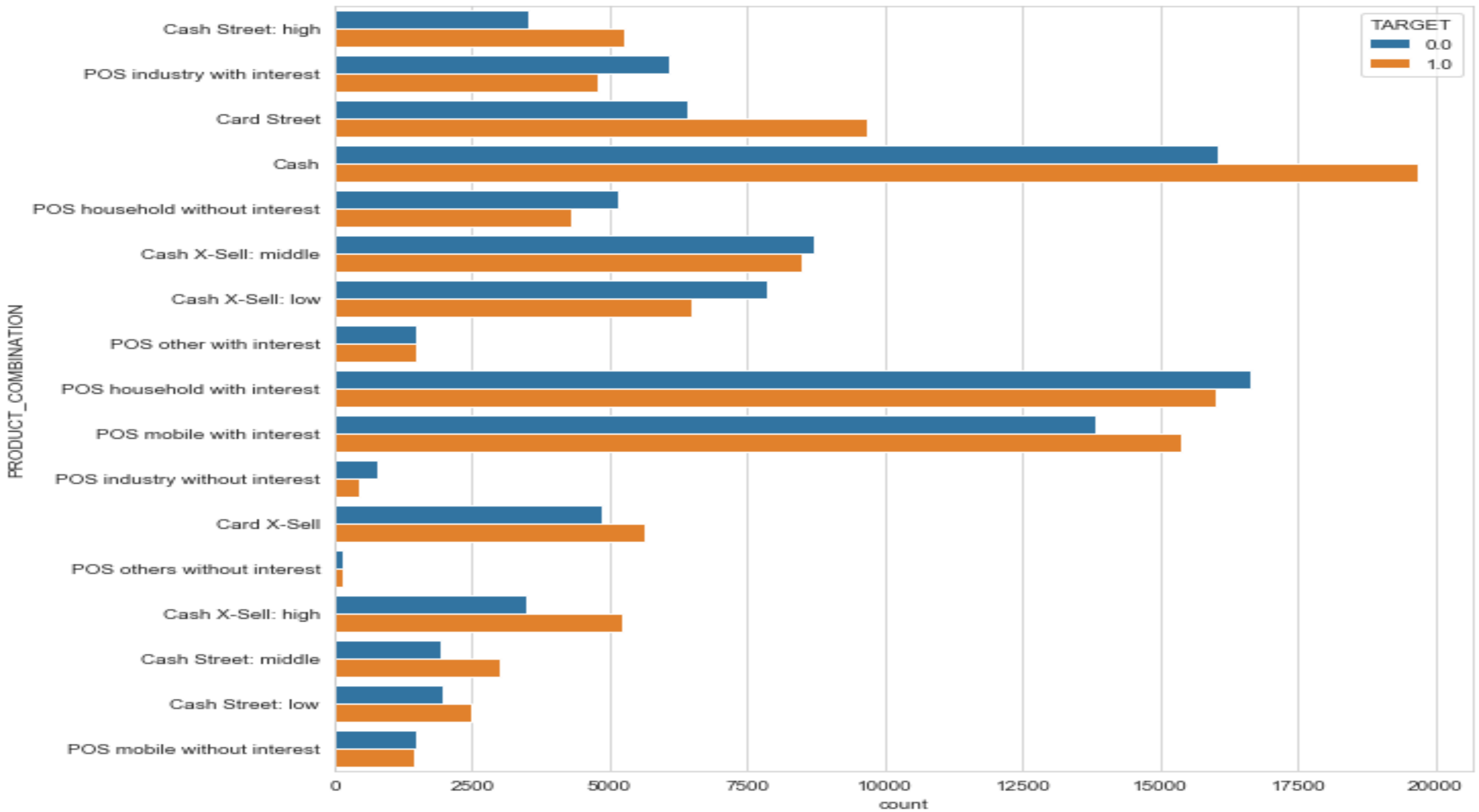


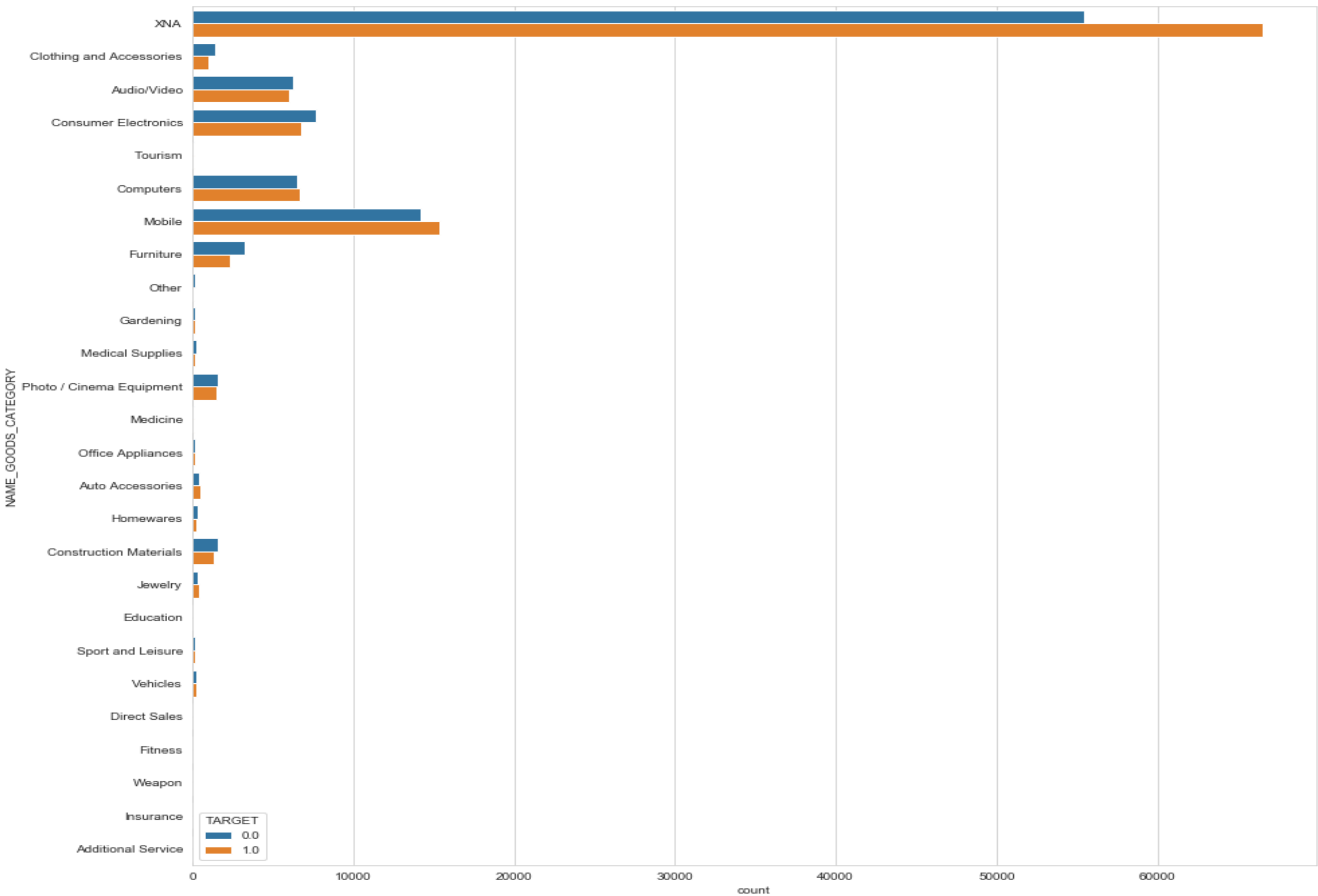
TARGET
0.0
1.0
Regional / Local
Stone
Credit and cash offices
Country-wide
AP+ (Cash loan)
Contact center
Channel of corporate sales
Car dealer
CHANNEL_TYPE
0
10000
20000
30000
40000
50000
count



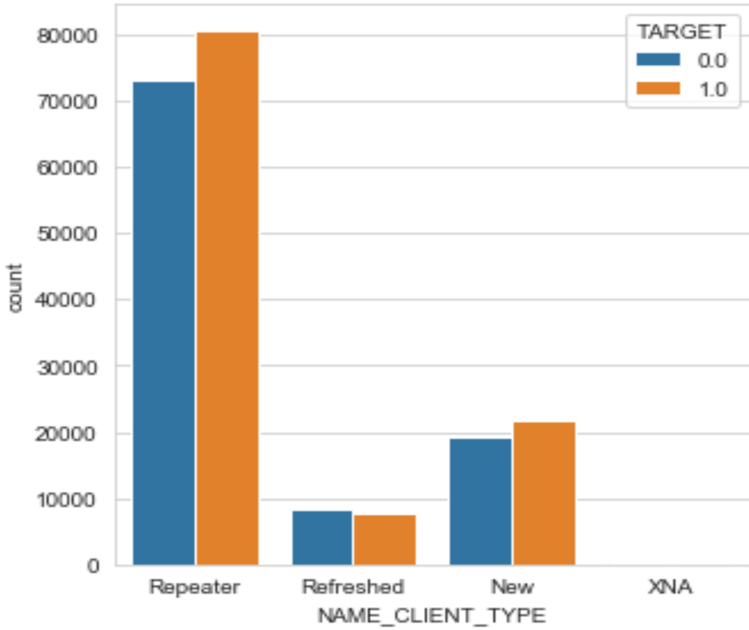
TARGET
0.0
1.0
count
35000
30000
25000
20000
15000
10000
5000
0
high
middle
XNA
low_action
low_normal
NAME_YIELD_GROUP

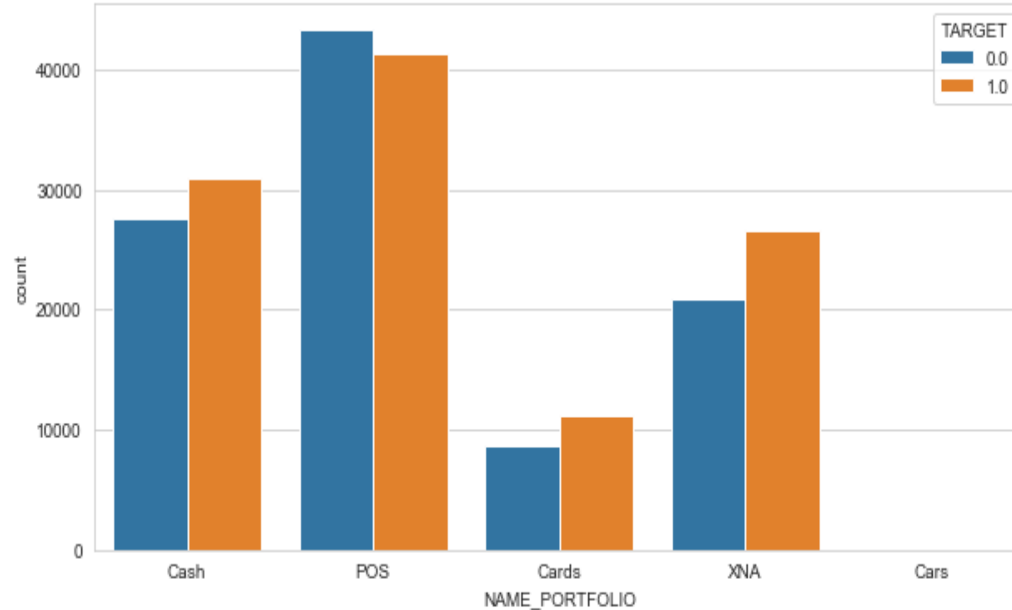

TARGET
0.0
1.0
count
60000
50000
40000
30000
20000
10000
0
Consumer electronics
Clothing
XNA
Industry
Connectivity
Furniture
Construction
MLM partners
Auto technology
Tourism
Jewelry
NAME_SELLER_INDUSTRY

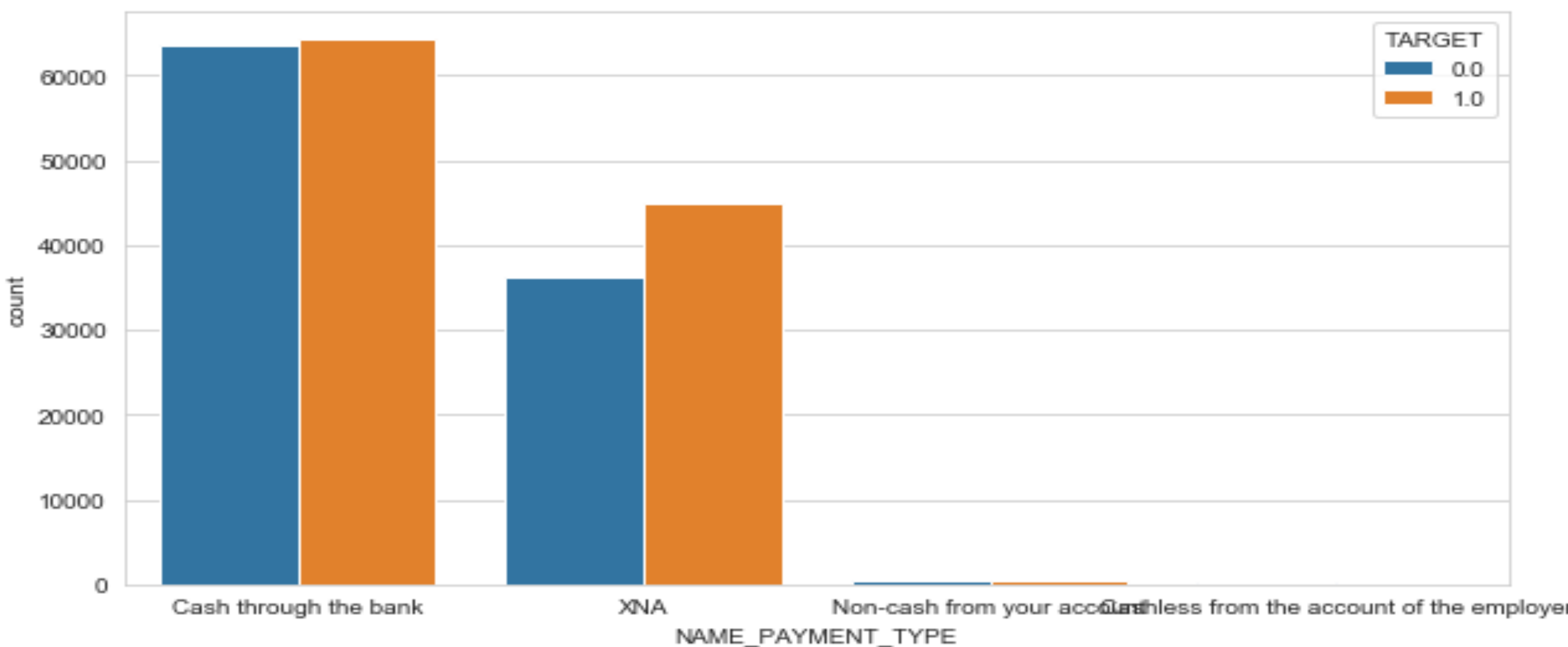

TARGET
0.0
1.0
PRODUCT_COMBINATION
Cash Street: high
POS industry with interest
Card Street
Cash
POS household without interest
Cash X-Sell: middle
Cash X-Sell: low
POS other with interest
POS household with interest
POS mobile with interest
POS industry without interest
Card X-Sell
POS others without interest
Cash X-Sell: high
Cash Street: middle
Cash Street: low
POS mobile without interest
0
2500
5000
7500
10000
12500
15000
17500
20000
count

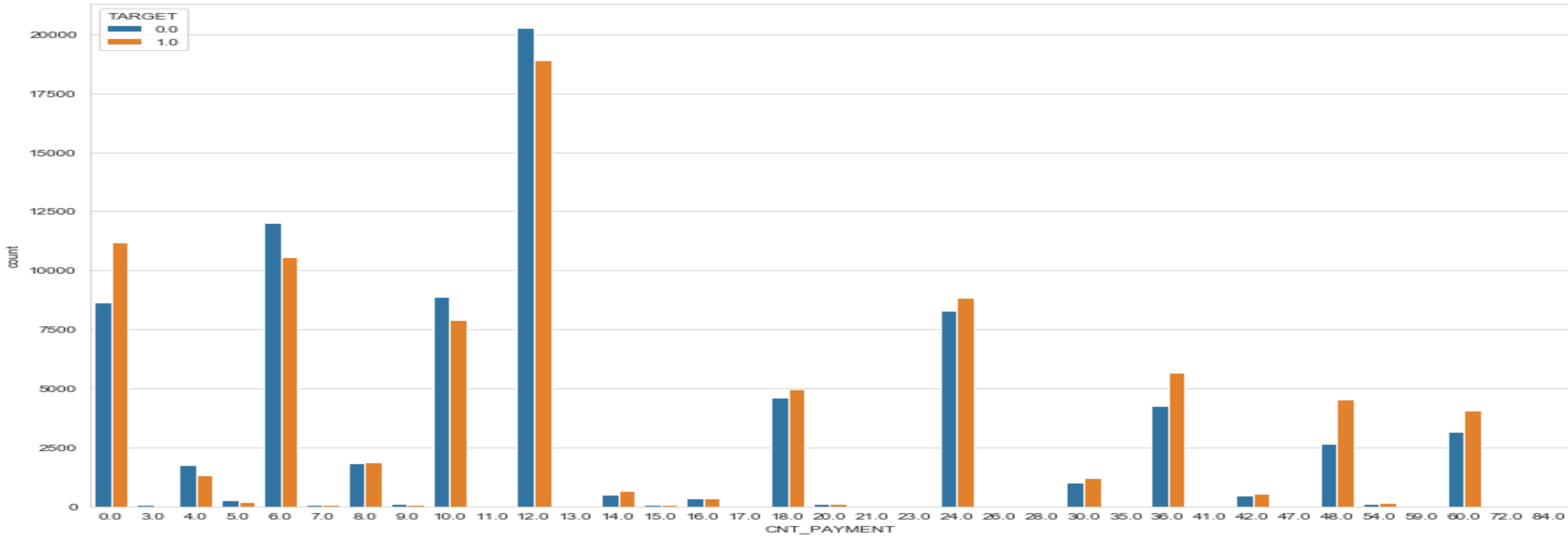


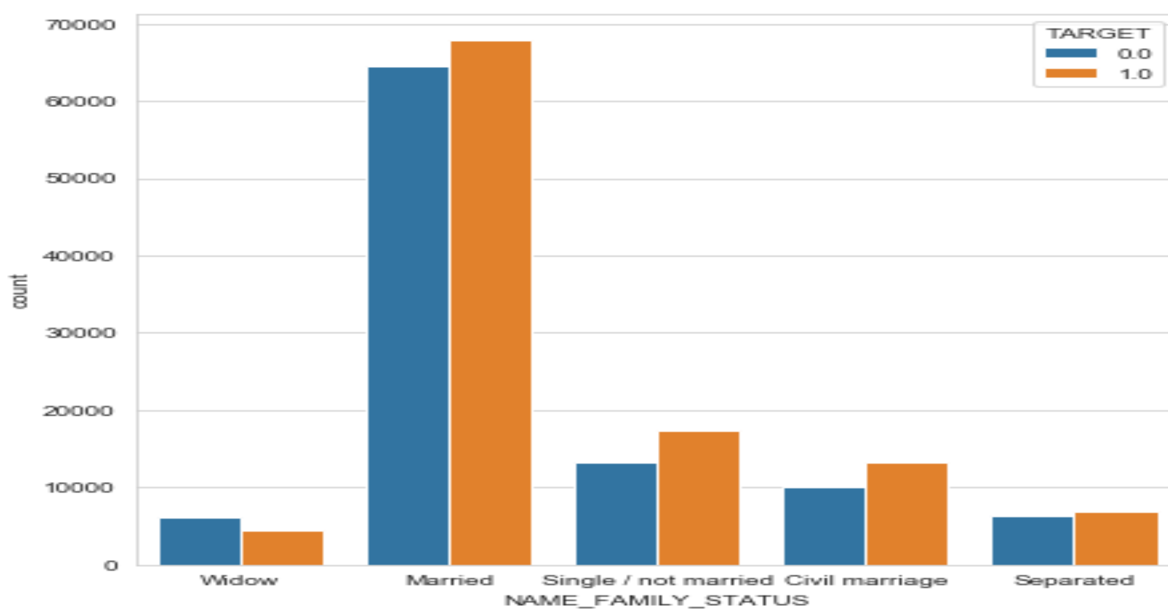
XNA
Clothing and Accessories
Audio/Video
Consumer Electronics
Tourism
Computers
Mobile
Furniture
Other
Gardening
Medical Supplies
Photo / Cinema Equipment
Medicine
Office Appliances
Auto Accessories
Homewares
Construction Materials
Jewelry
Education
Sport and Leisure
Vehicles
Direct Sales
Fitness
Weapon
Insurance
Additional Service
NAME_GOODS_CATEGORY
TARGET
0.0
1.0
0
10000
20000
30000
40000
50000
60000
count


TARGET
0.0
1.0
count
0
10000
20000
30000
40000
50000
60000
70000
80000
Repeater
Refreshed
New
XNA
NAME_CLIENT_TYPE


TARGET
0.0
1.0
count
0
10000
20000
30000
40000
Cash
POS
Cards
XNA
Cars
NAME_PORTFOLIO


TARGET
0.0
1.0
count
0
10000
20000
30000
40000
50000
60000
Cash through the bank
XNA
Non-cash from your account
Cashless from the account of the employer
NAME_PAYMENT_TYPE


TARGET
0.0
1.0
20000
17500
15000
12500
10000
7500
5000
2500
0
count
0.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 20.0 21.0 23.0 24.0 26.0 28.0 30.0 35.0 36.0 41.0 42.0 47.0 48.0 54.0 59.0 60.0 72.0 84.0
CNT_PAYMENT


TARGET
0.0
1.0
70000
60000
50000
40000
30000
20000
10000
0
count
Widow
Married
Single / not married
Civil marriage
Separated
NAME_FAMILY_STATUS

Observation:

Following fields are significant

NAME_CONTRACT_STATUS

'Refused' having much more % population with payment difficulty. Which is correct. Loan should be refused to customer with payment difficulties

NAME_YIELD_GROUP

'High interest rate' having much more % population with payment difficulty

CHANNEL_TYPE

with values "Credit and cash Office/contact center/ AP+(cash loan)" having much more % population with payment difficulty

NAME_SELLER_INDUSTRY

"Connectivity" having much more % population with payment difficulty

PRODUCT_COMBINATION

"CASH/Card Street/POS Mobile with interest" having much more % population with payment difficulty

NAME_GOODS_CATEGORY

"Mobile" having much more % population with payment difficulty

NAME_CLIENT_TYPE

"Repeater" having much more % in population with payment difficulty

NAME_PAYMENT_TYPE

No Significant Impact

NAME_PORTFOLIO

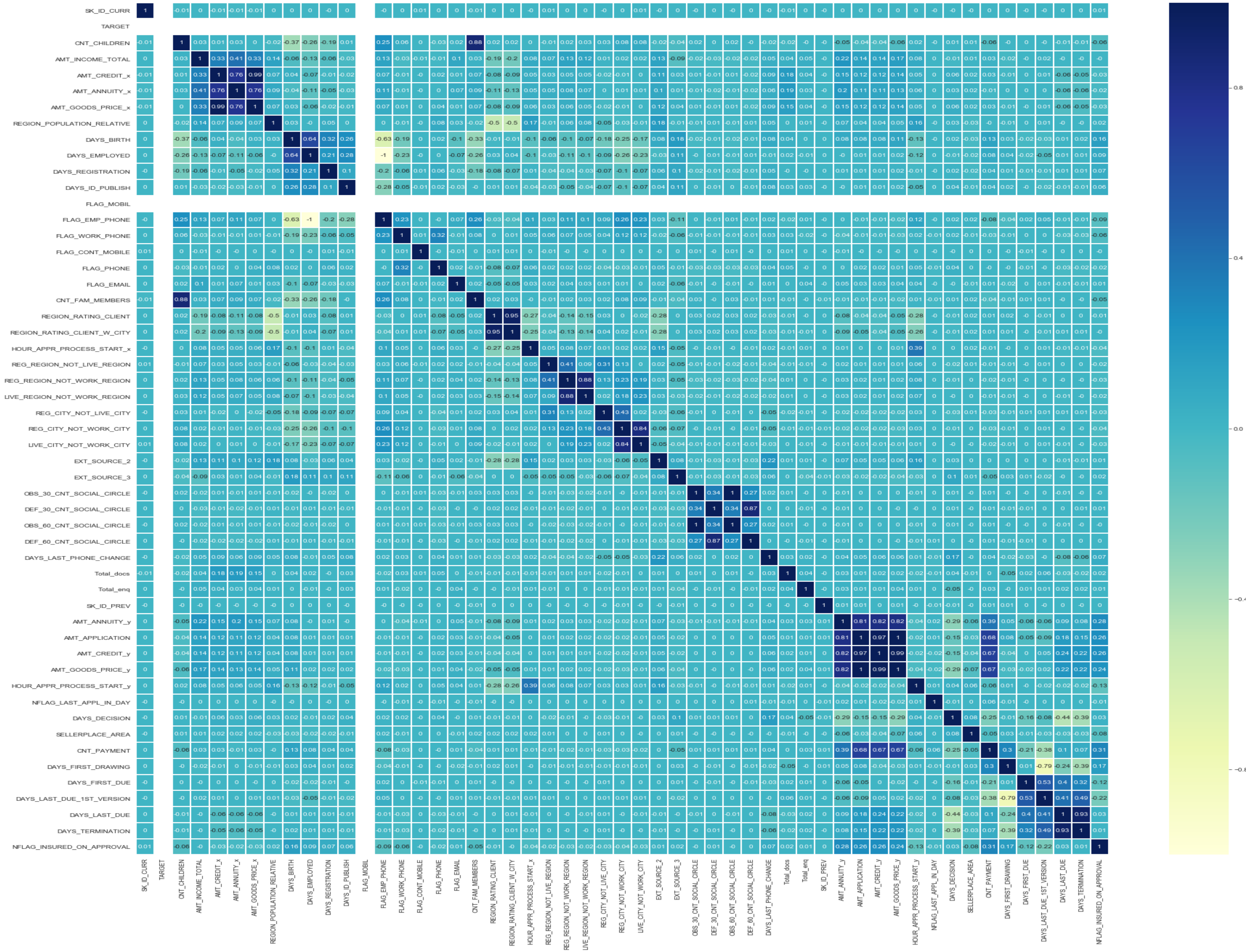
"Cash/Cards" having much more % in population with payment difficulty

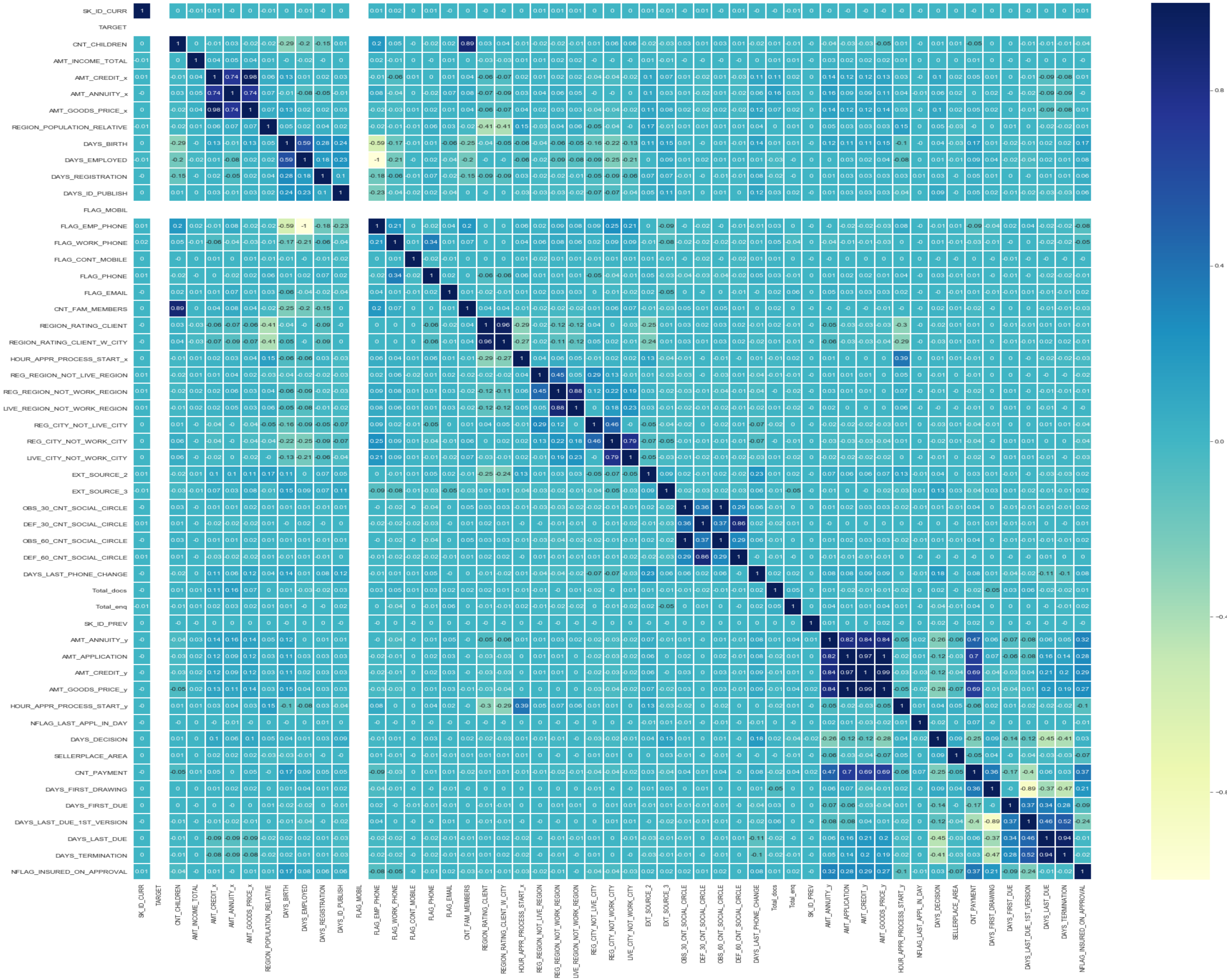
CNT_PAYMENT

'long Term loan on previous application' having much more % in population with payment difficulty

NAME_FAMILY_STATUS

'Single/Not Married/Civil_Marriage' having much more % in population with payment difficulty


SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT_x
AMT_ANNUITY_x
AMT_GOODS_PRICE_x
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START_x
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
Total_docs
Total_enq
SK_ID_PREV
AMT_ANNUITY_y
AMT_APPLICATION
AMT_CREDIT_y
AMT_GOODS_PRICE_y
HOUR_APPR_PROCESS_START_y
NFLAG_LAST_APPL_IN_DAY
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
0.8
0.4
0.0
-0.4
-0.8


SK_ID_CURR
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT_x
AMT_ANNUITY_x
AMT_GOODS_PRICE_x
REGION_POPULATION_RELATIVE
DAYS_BIRTH
DAYS_EMPLOYED
DAYS_REGISTRATION
DAYS_ID_PUBLISH
FLAG_MOBIL
FLAG_EMP_PHONE
FLAG_WORK_PHONE
FLAG_CONT_MOBILE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
HOUR_APPR_PROCESS_START_x
REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION
LIVE_REGION_NOT_WORK_REGION
REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY
EXT_SOURCE_2
EXT_SOURCE_3
OBS_30_CNT_SOCIAL_CIRCLE
DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE
DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE
Total_docs
Total_enq
SK_ID_PREV
AMT_ANNUITY_y
AMT_APPLICATION
AMT_CREDIT_y
AMT_GOODS_PRICE_y
HOUR_APPR_PROCESS_START_y
NFLAG_LAST_APPL_IN_DAY
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
0.8
0.4
0.0
-0.4
-0.8

Below are the Top 10 variables which are correlated.

Observation: df_merged_target0

AMT_CREDIT and AMT_GOODS_PRICE (.99)
AMT_CREDIT and AMT_APPLICATION (.97)
REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY(.95)
DAYS_LAST_DUE and DAYS_TERMINATION (.93)
CNT_FAM_MEMBERS and CNT_CHILDREN(.89)
REG_REGION_NOT_LIVE_REGION and LIVE_REGION_NOT_WORK_REGION(.88)
DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE(.87)
REG_CITY_NOT_WORK_CITY and REG_CITY_NOT_LIVE_CITY(.84)
AMT_ANNUITY and AMT_GOODS_PRICE(.82)
AMT_ANNUITY and AMT_CREDIT(.82)
AMT_ANNUITY and AMT_APPLICATION(.81)

Observation: df_merged_target1

AMT_CREDIT and AMT_GOODS_PRICE(.99)
AMT_CREDIT and AMT_APPLICATION(.97)
REGION_RATING_CLIENT and REGION_RATING_CLIENT_W_CITY(.96)
DAYS_LAST_DUE and DAYS_TERMINATION(.94)
CNT_FAM_MEMBERS and CNT_CHILDREN(.89)
REG_REGION_NOT_LIVE_REGION and LIVE_REGION_NOT_WORK_REGION(.88)
DEF_30_CNT_SOCIAL_CIRCLE and DEF_60_CNT_SOCIAL_CIRCLE(.86)
AMT_ANNUITY and AMT_GOODS_PRICE(.84)
AMT_ANNUITY and AMT_CREDIT(.84)
AMT_ANNUITY and AMT_APPLICATION(.82)
REG_CITY_NOT_WORK_CITY and REG_CITY_NOT_LIVE_CITY(.79)

Summary

- Social Circle, Credit rating(EXT_SOURCE), Occupation type, Education type, Family Status, Income type, Tenure of the loan are the factors to be considered during Credit risk assessment

- As we observed correlation among variables is almost SAME for Target 0 and Target 1