Project on Exploratory Data Analysis

Credit EDA Assignment

By: Bhupendra Gautam

Design

- Problem Statement: 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:

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.
3. 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.

Analysis Approach:

To analyze the dataset provided in case study, I have opted following steps First of all, I have analyzed both the data sets one by one.

1. Step1: Load the application data set in Jupyter notebook and perform basic metadata checks(Like shape, describe, info, value_counts, head) for the basic data understanding.

2. Then I have find missing value % in each column.

3. Decided threshold 50% , and exclude the data which contains more than 50% missing counts.

4. Then I have replaced missing values in some of the columns with mean, median and mode.

5. For numeric values if I found outlier I have replaced with median, If outlier is there I have replaced with mean.

6. For categorical values: Replaced with mode.

7. After identifying and replacing a few missing values, I have performed Univariate and Bi-variate analysis with the help of Bar graphs, Boxplots, distribution plots and find some observations on current data set. I have divided the data set into 2 part because of data imbalance and performed analysis for defaulters and Non Defaulters on various parameter.

Then I have performed analysis for both the Target variables defaulters and non defaulters separately.

8. Repeated the process for the previous data as well.

9. Merged the both data set again and performed the analysis.

Target variable for Application dataset – TARGET **[1-for non default, 0-default]**

Target variable for Previous dataset - NAME_CONTRACT_STATUS

To proceed with loan application I have analyzed the following parameters thoroughly:

NAME_EDUCATION_TYPE, AMT_INCOME_TOTAL, DAYS_BIRTH, AMT_CREDIT, DAYS_EMPLOYED, AMT_ANNUITY, NAME_INCOME_TYPE, CODE_GENDER, NAME_HOUSING_TYPE,

Observations/Predictions:

1. Less family members got more approval.(Nuclear Families, Having 2-3 members)

2. Mostly, repeater application got approved for again loan grant.

3. For Married applicant more loan approval is there then any other marital status

4. Those who are having secondary/ secondary special education level gets loan approval more than any other education level.

5. Highest number of loan approval for the working people, than commercial associate and pensioners. least approval for the state servants.

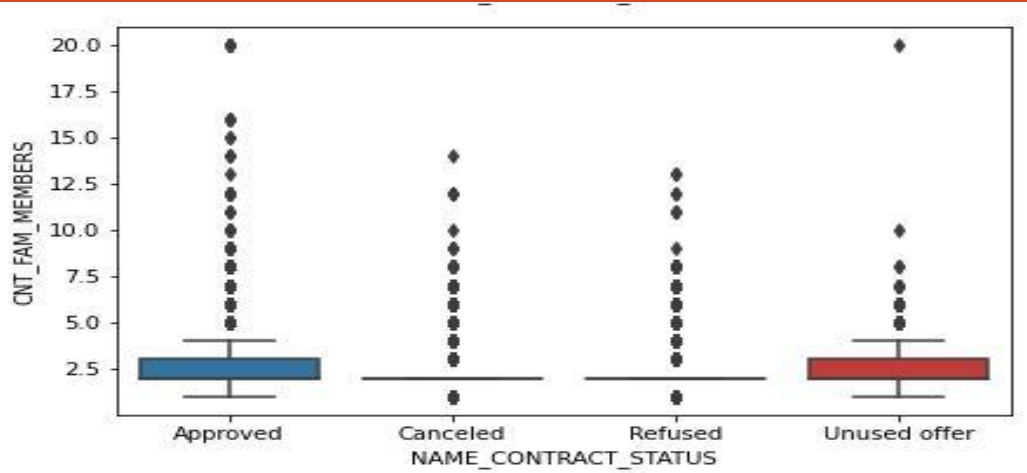
6. from above graph we can say that working people take more loans Pensioners are high is number for the non defaulter cases. while income source and the defaulters both categories are higher in count for working people

7. Those who do not have any car, takes more loan and falls under the category of defaulters.

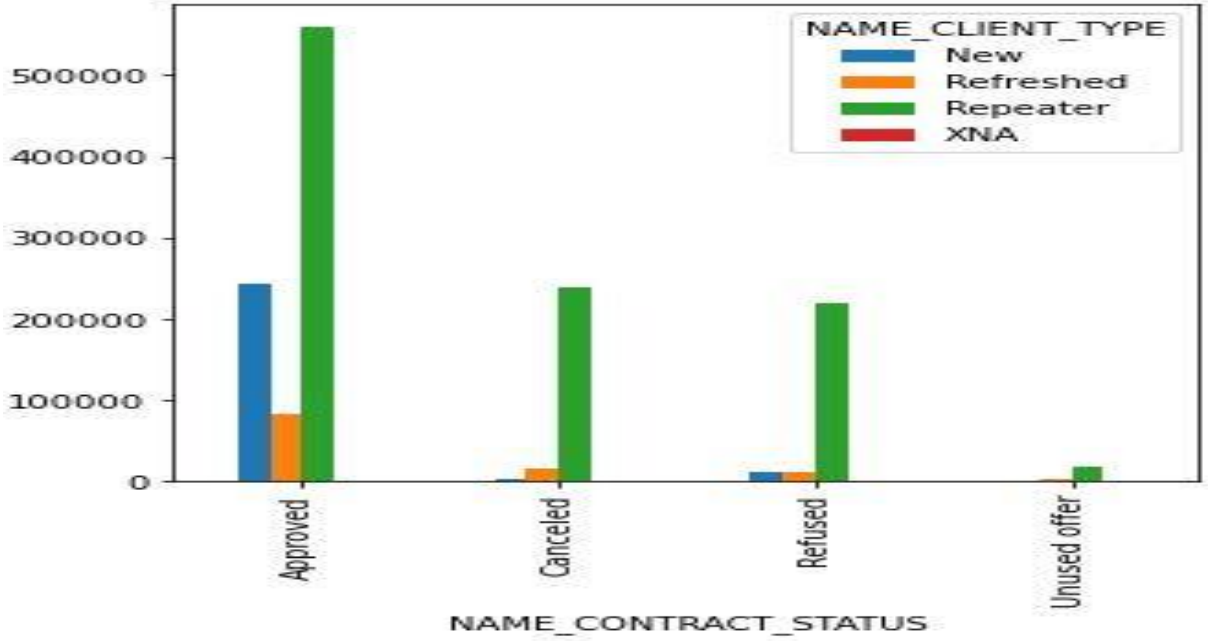
8. Those who falls under the age bracketed of 27 years to 41 years, Takes more loans and maximum comes under defaluters

9. People those who are having Medium total income are more likely falls under the category of defaulters with compare to high and low Total income.

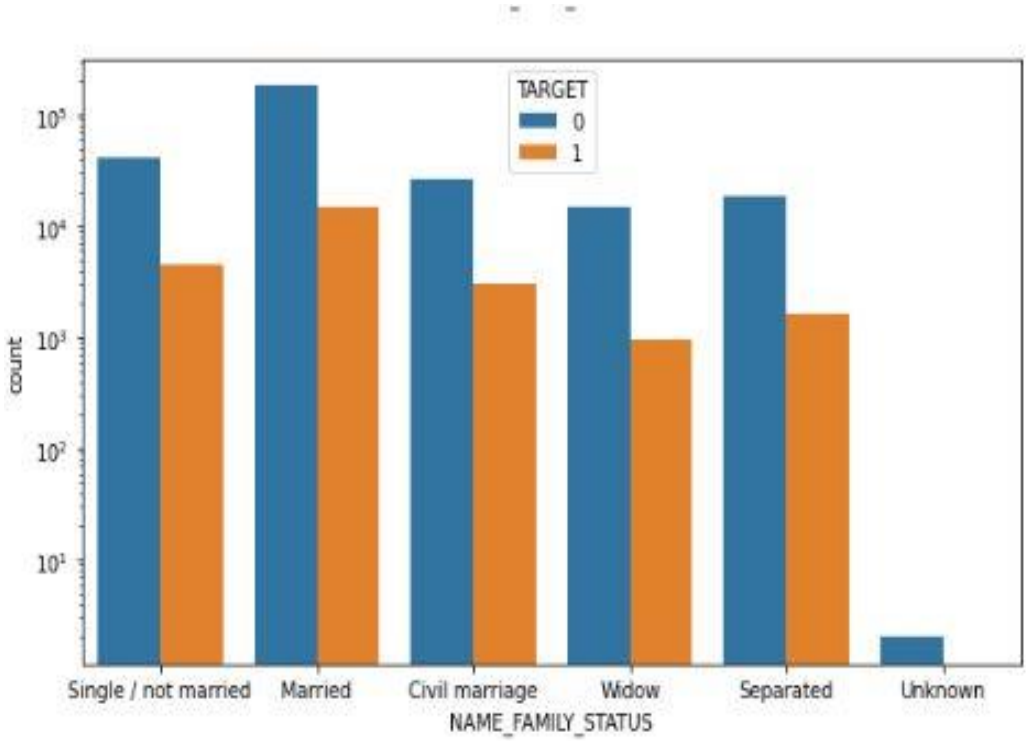
1. Less family members got more approval.(Nuclear Families, Having 2-3 members)

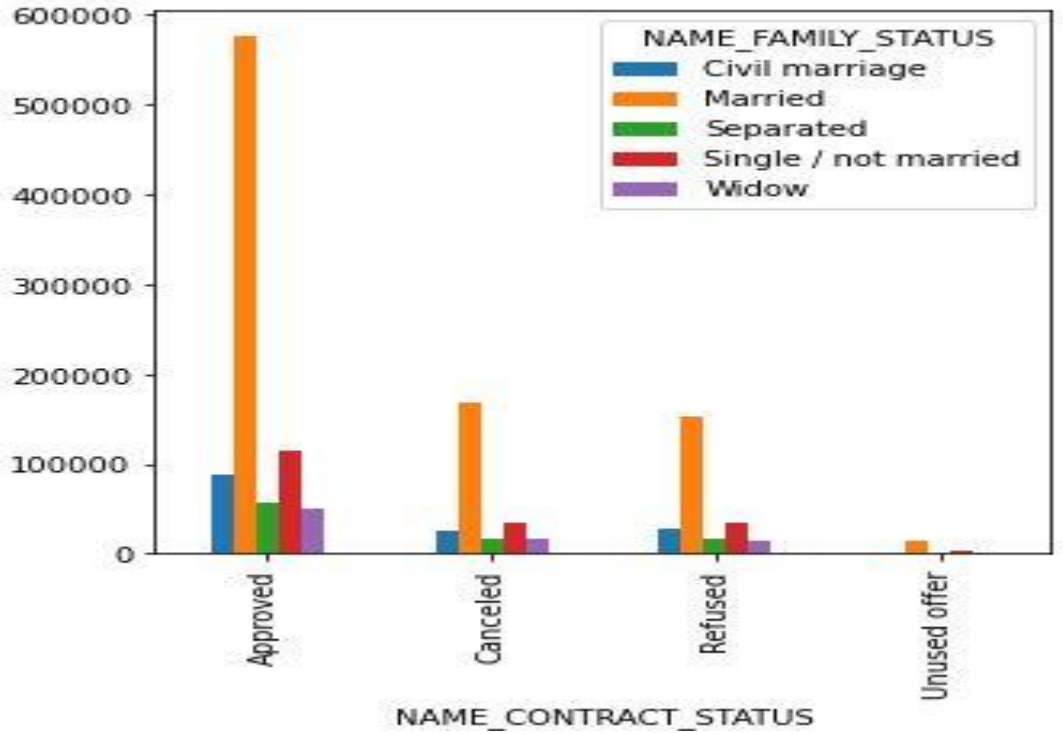


2. Mostly, repeater application got approved for again loan grant.

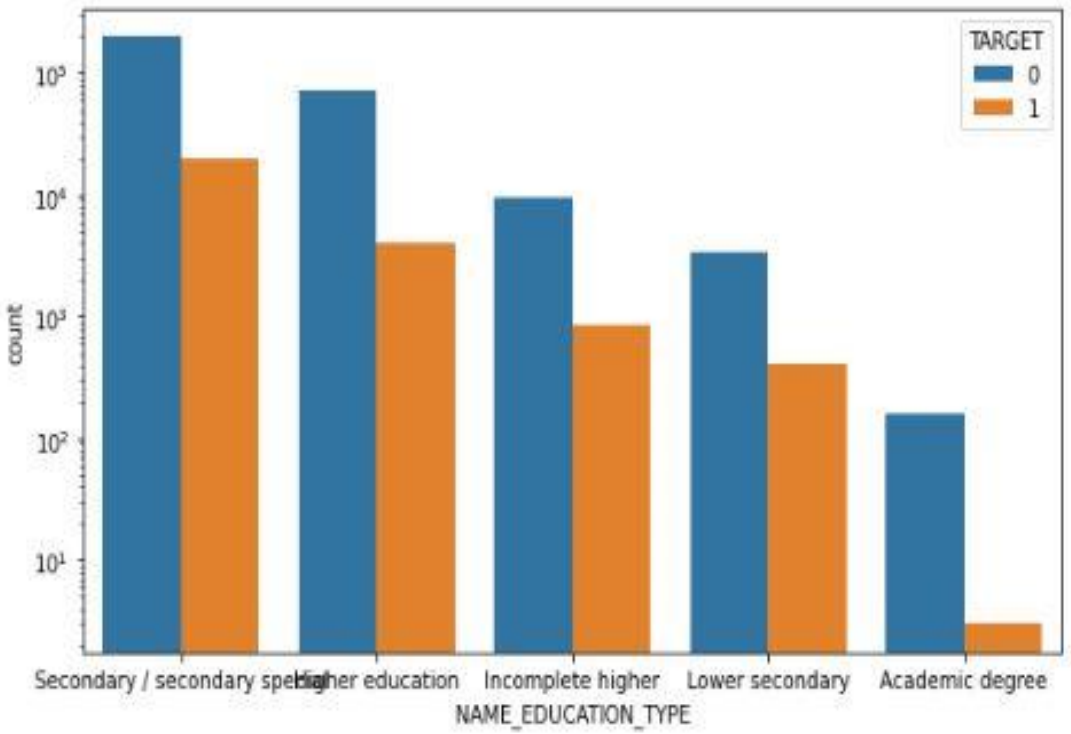


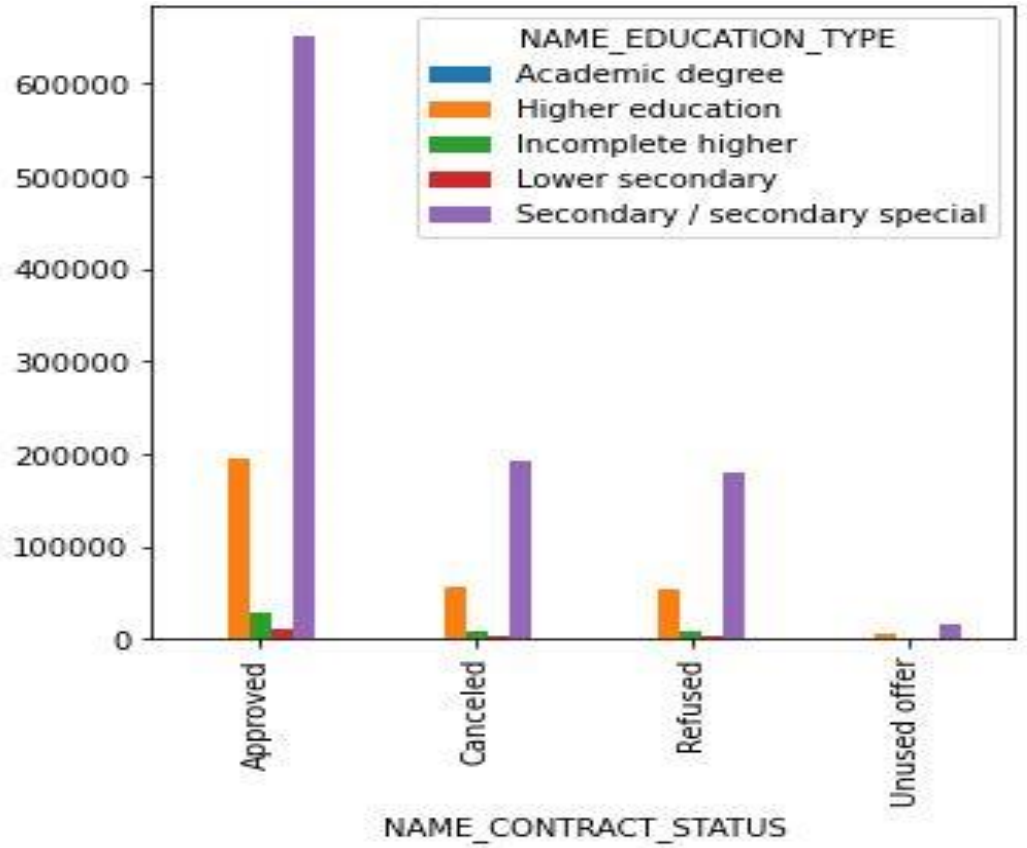
3. For Married applicant more loan approval is there then any other marital status



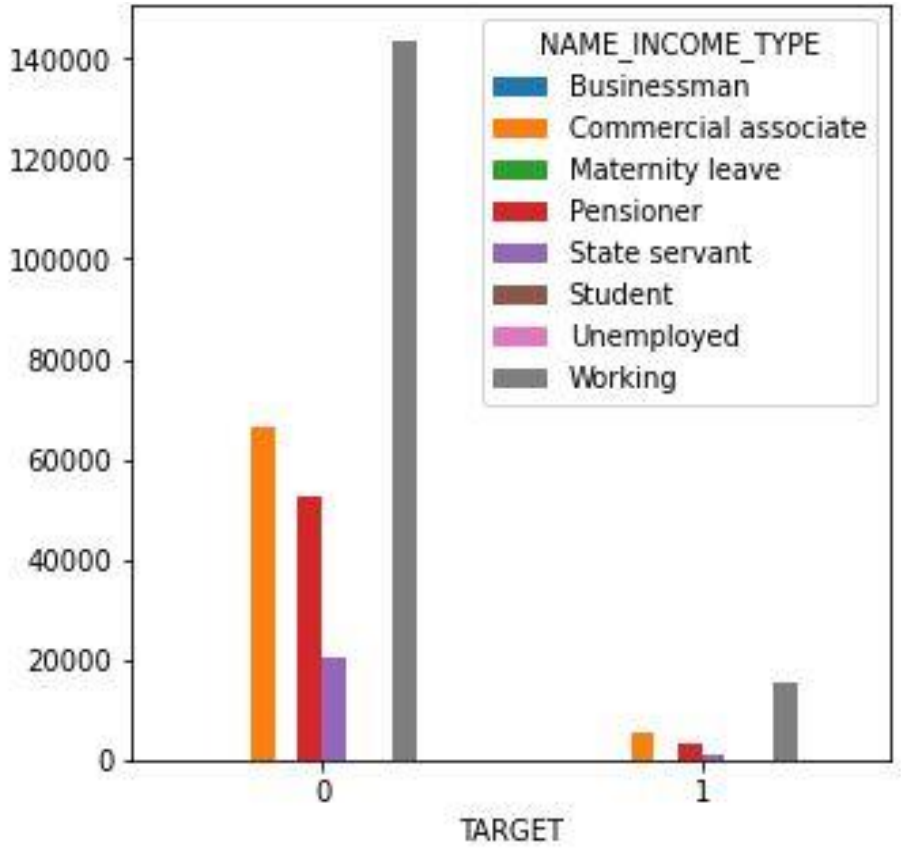


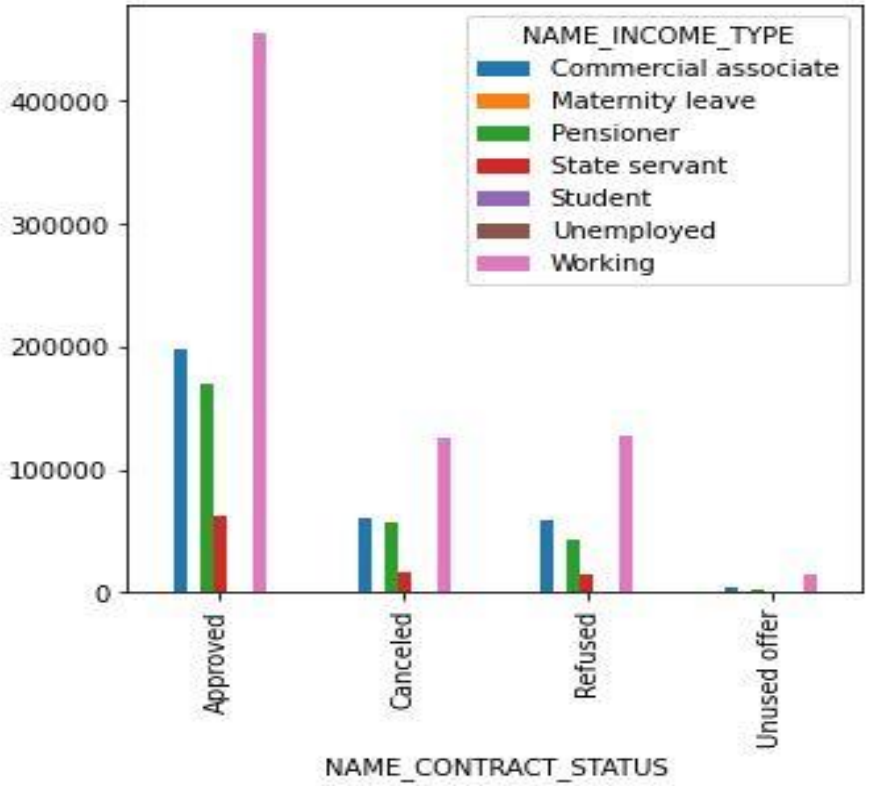
4. Those who are having secondary/ secondary special education level gets loan approval more than any other education level.



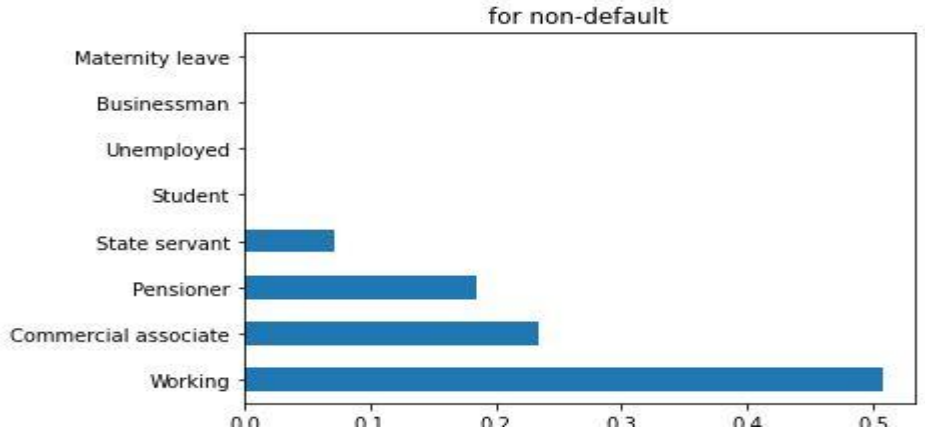


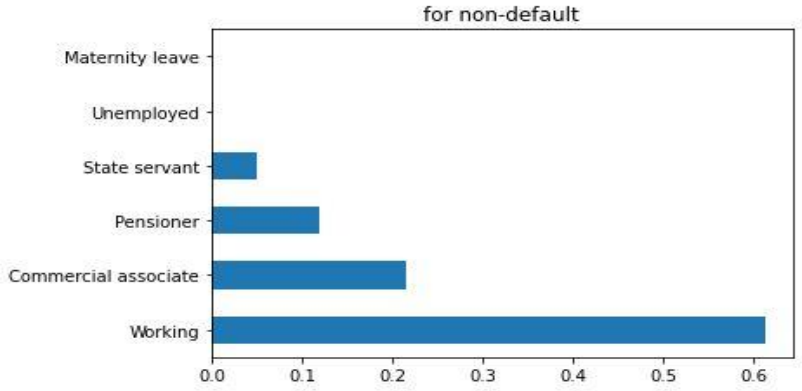
5. Highest number of loan approval for the working people, than commercial associate and pensioners. least approval for the state servants.



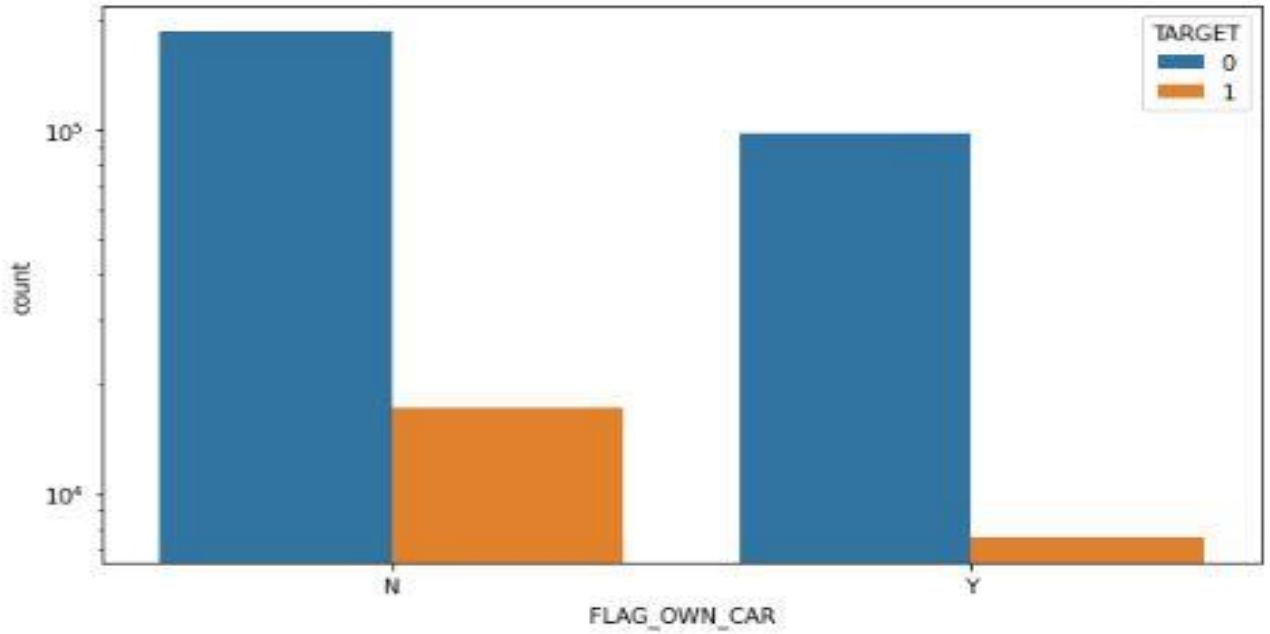


6. from above graph we can say that working people take more loans Pensioners are high is number for the non defaulter cases. while income source and the defaulters both categories are higher in count for working people.

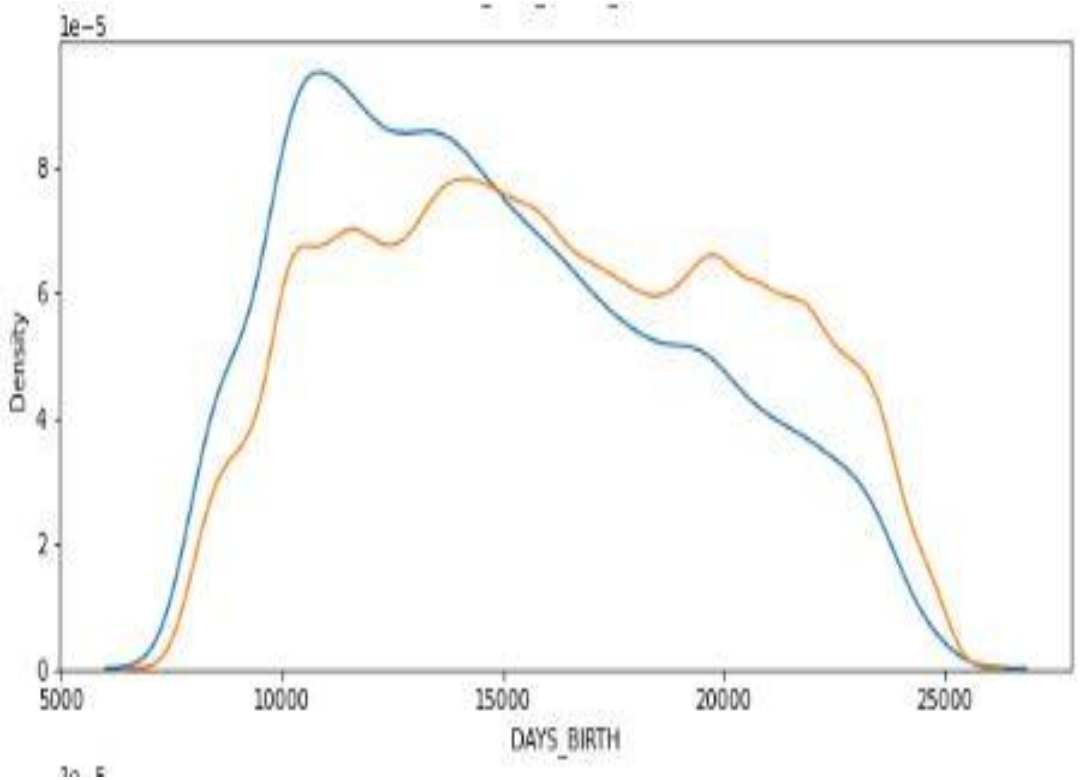


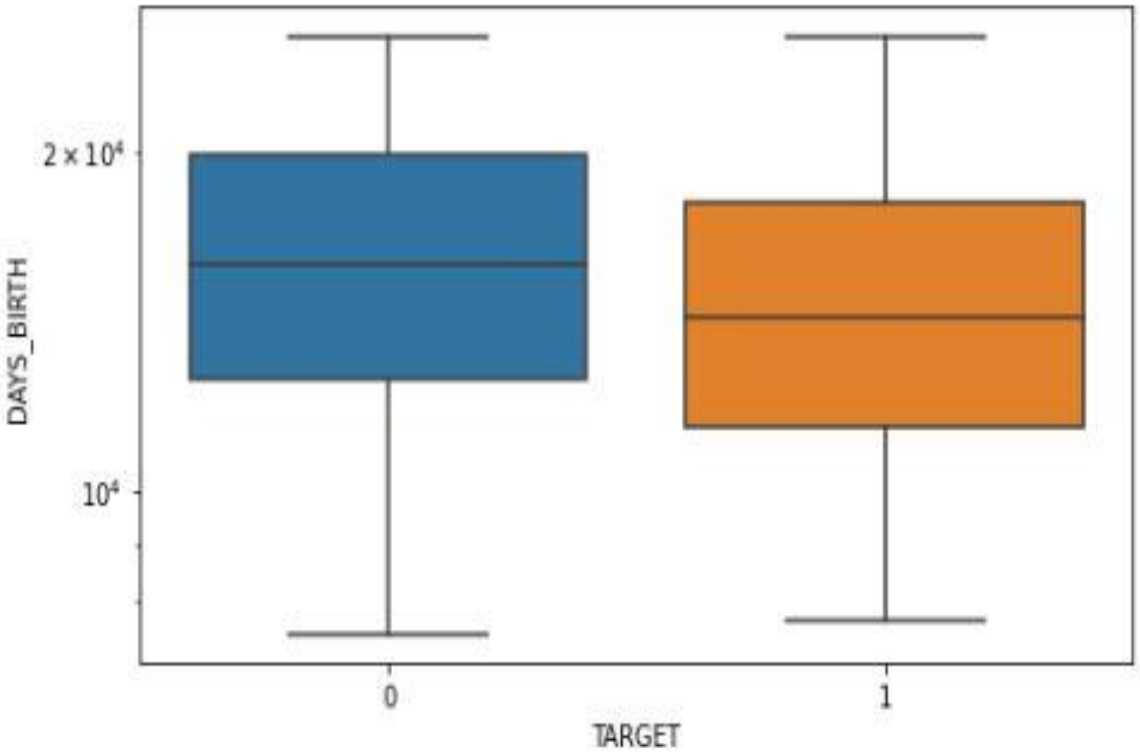


7. Those who do not have any car, takes more loan and falls under the category of defaulters.

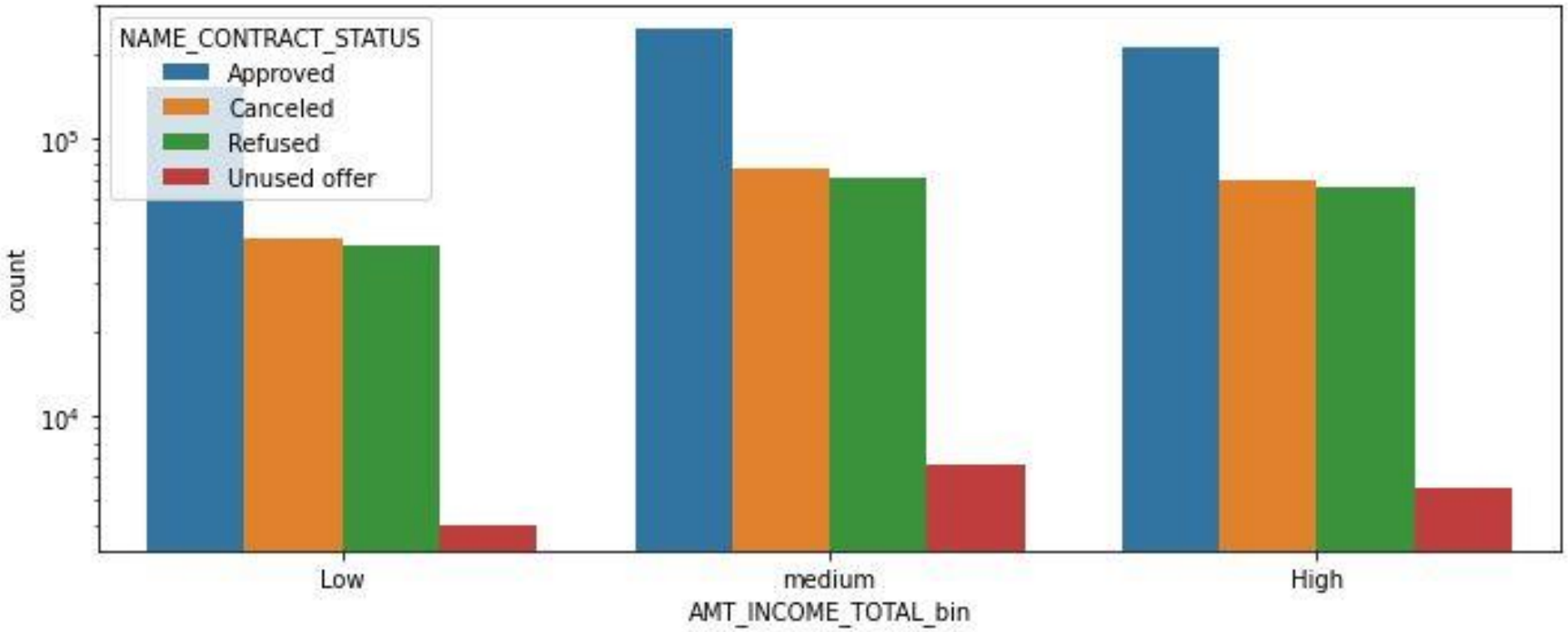


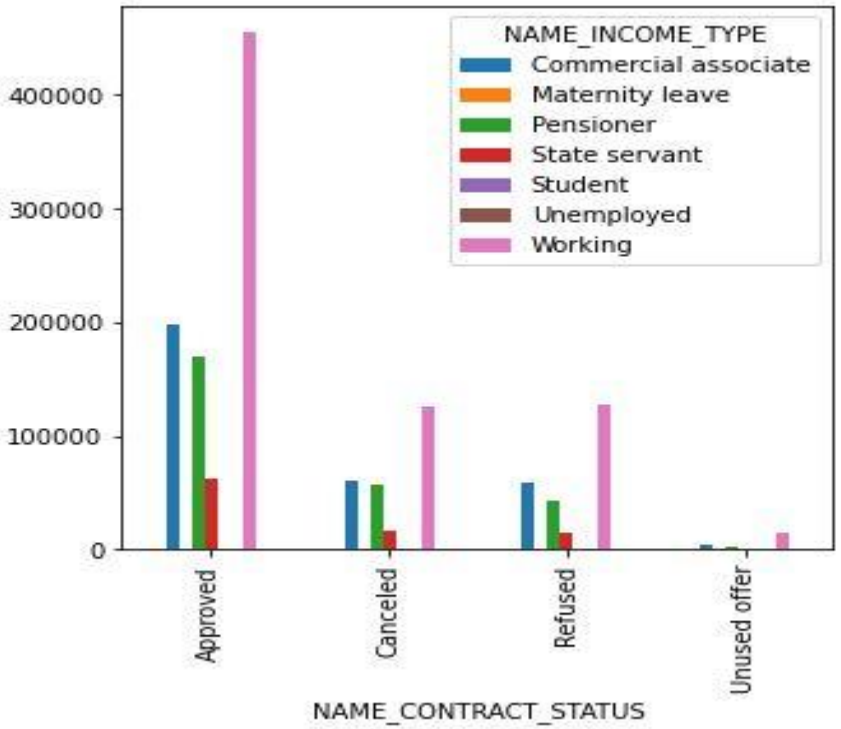
8. Those who falls under the age bracketed of 27 years to 41 years, Takes more loans and maximum comes under defaulters.

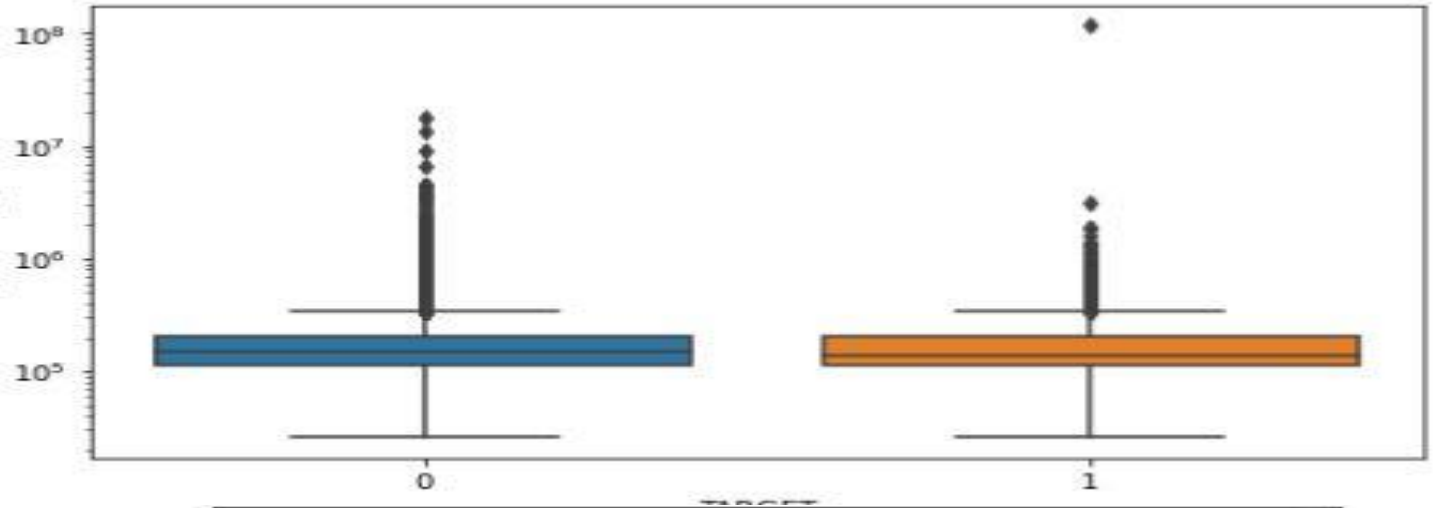


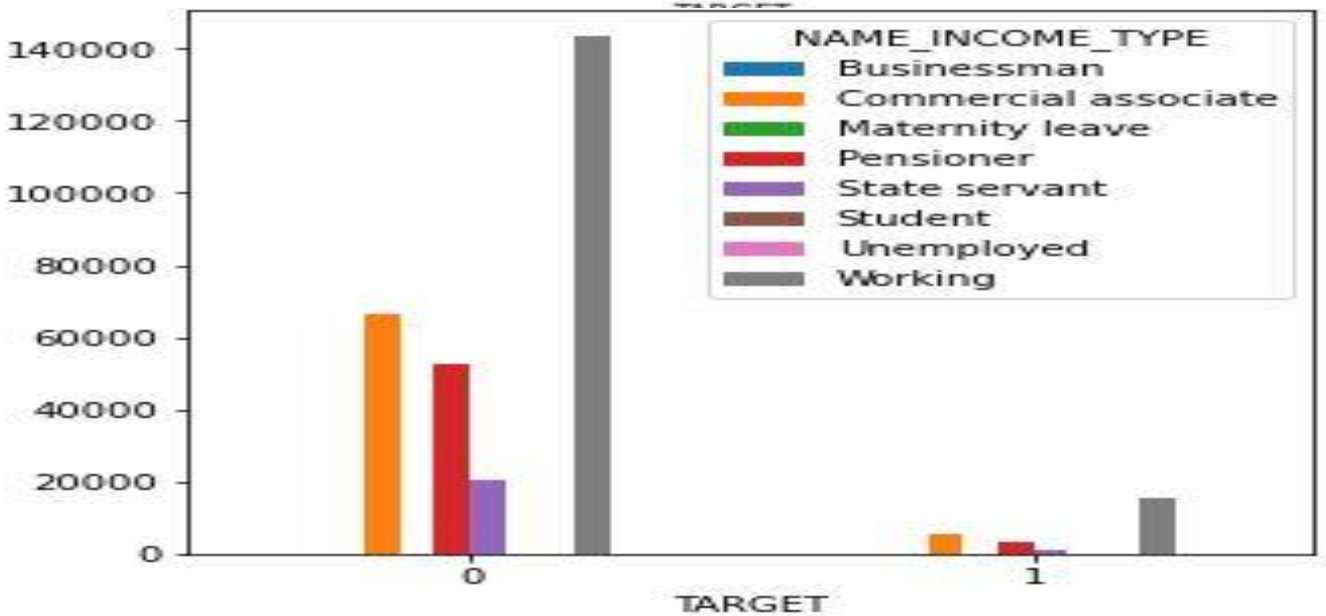


9. People those who are having Medium total income are more likely falls under the category of defaulters with compare to high and low Total income.

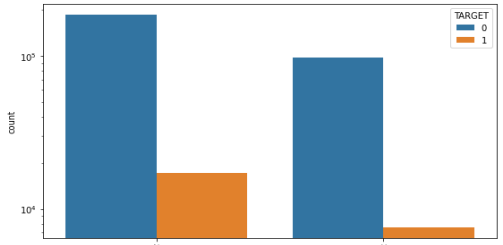


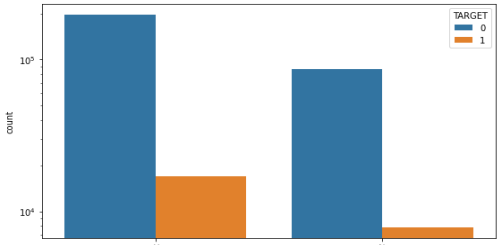

NAME_INCOME_TYPE
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
400000
300000
200000
100000
0
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

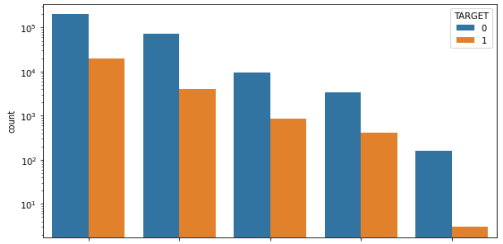

10^8
10^7
10^6
10^5
0
1

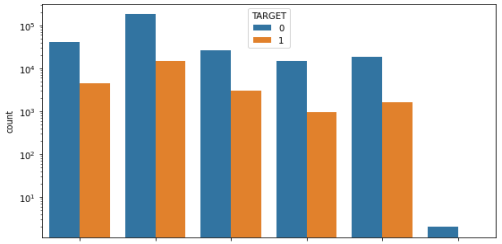

NAME_INCOME_TYPE
Businessman
Commercial associate
Maternity leave
Pensioner
State servant
Student
Unemployed
Working
140000
120000
100000
80000
60000
40000
20000
0
0
1
TARGET

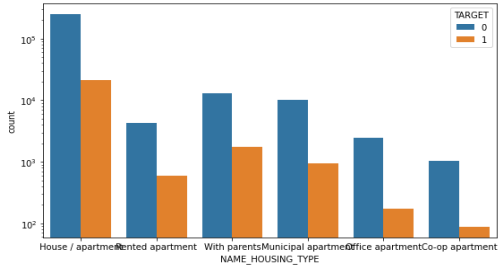
Extra graphs

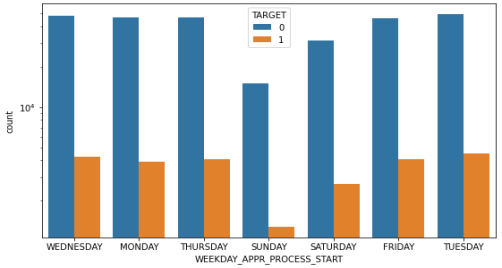


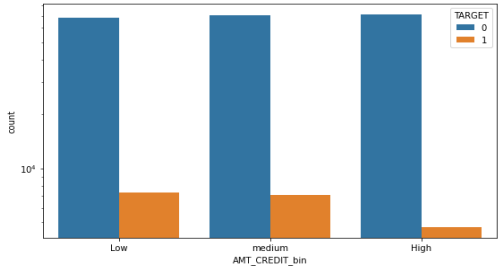


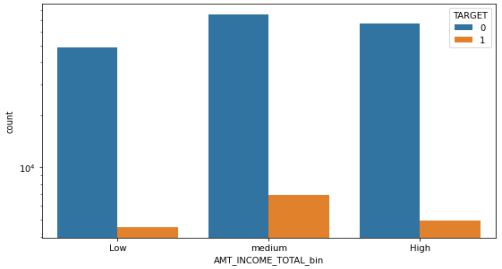


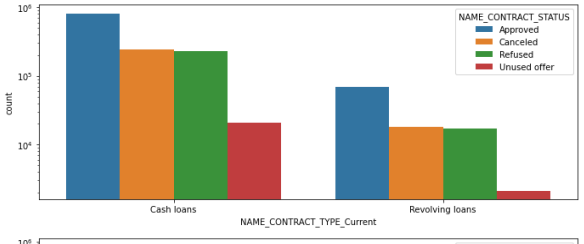


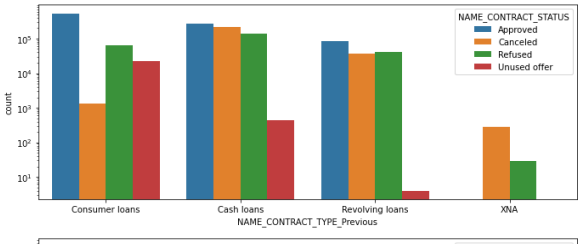


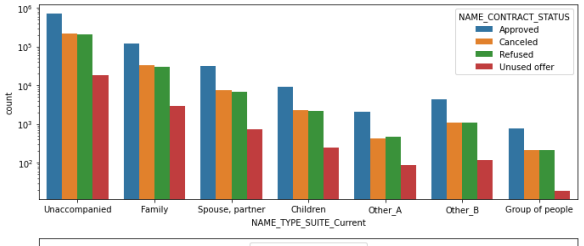


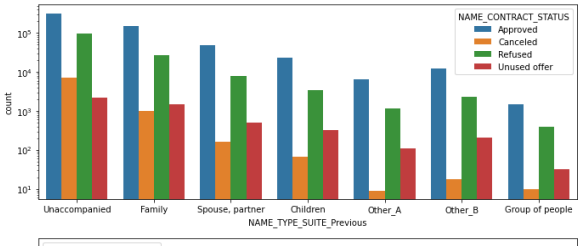


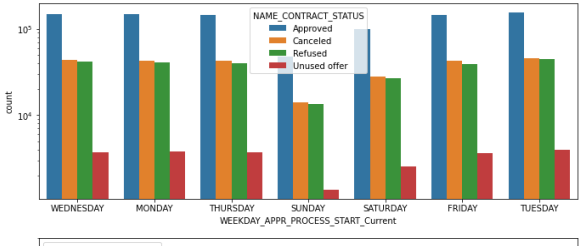


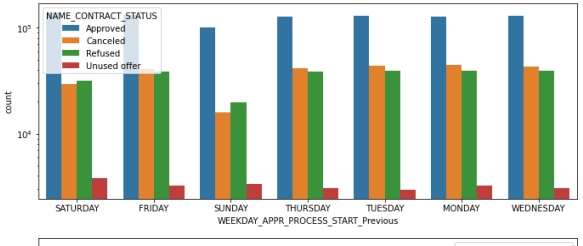


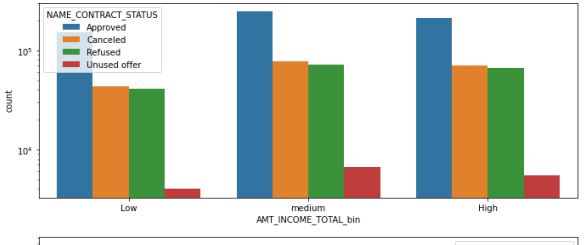


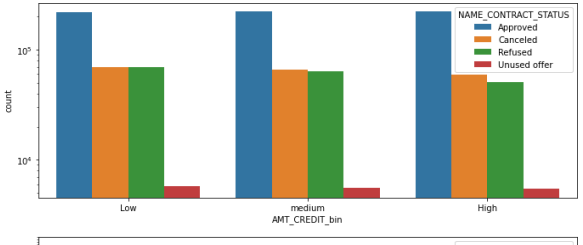


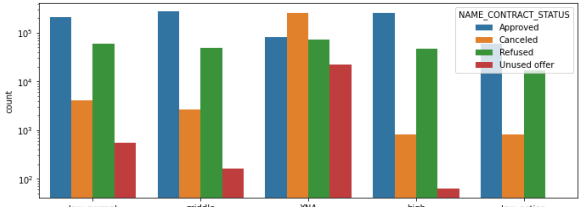


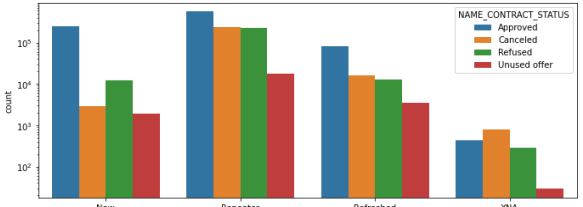


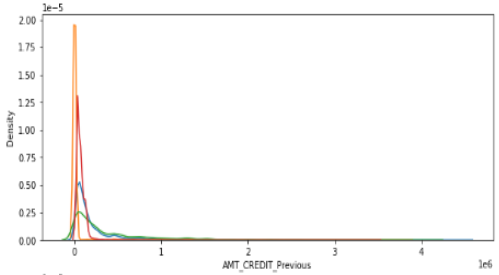


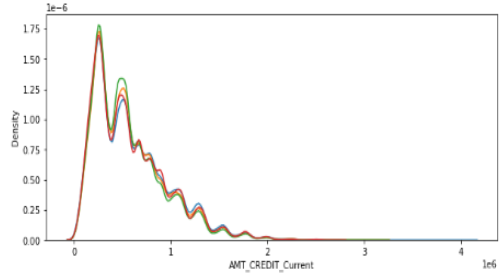


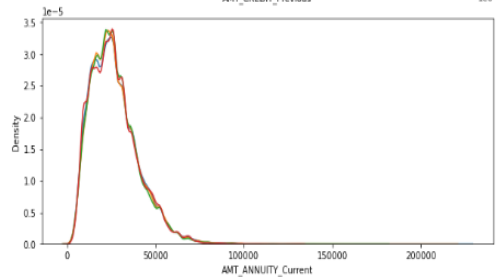


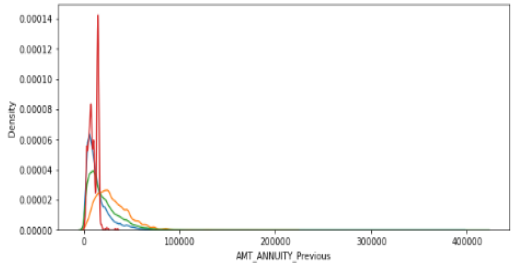


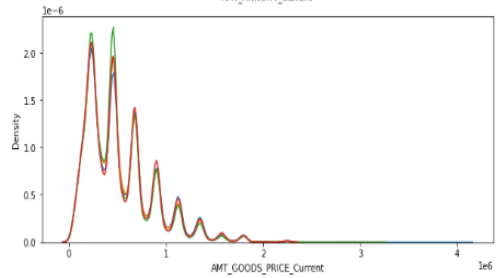


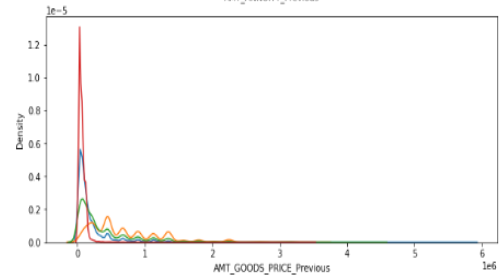

1e−5
Density
AMT_CREDIT_Previous
1e6

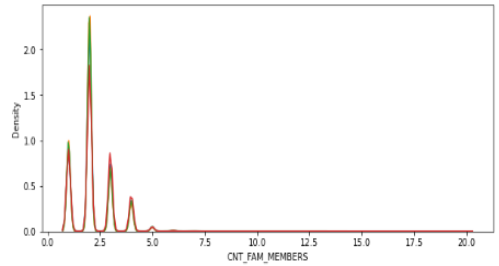

1e−6
Density
AMT_CREDIT_Current
1e6

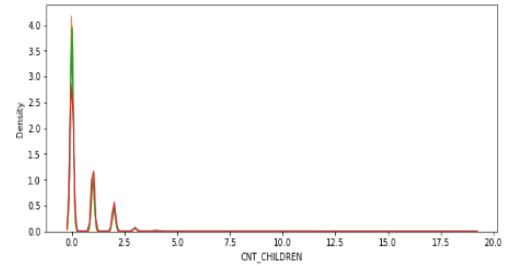

1e−5
Density
AMT_ANNUITY_Current

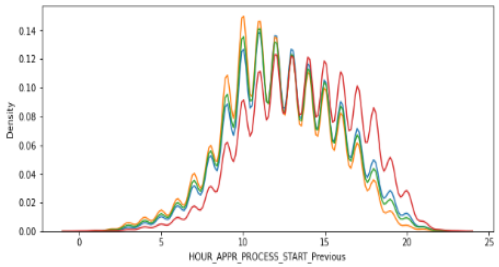

Density
AMT_ANNUITY_Previous

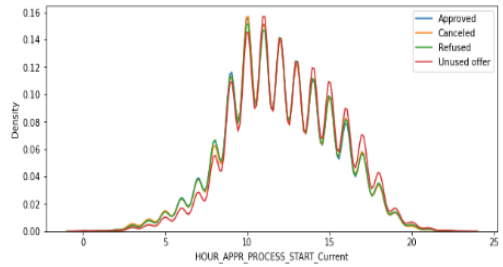

1e−6
Density
AMT_GOODS_PRICE_Current
1e6

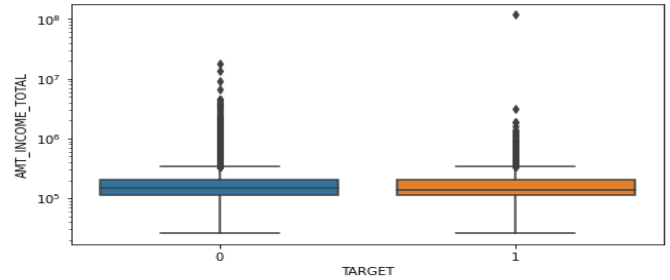

1e−5
Density
AMT_GOODS_PRICE_Previous
1e6

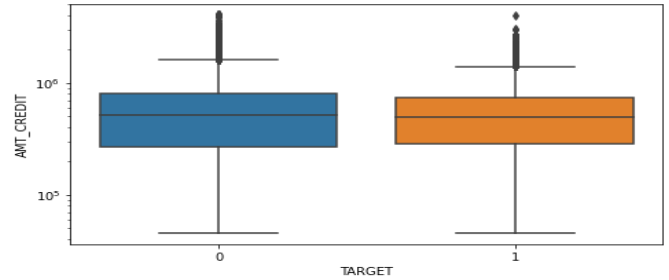

Density
CNT_FAM_MEMBERS

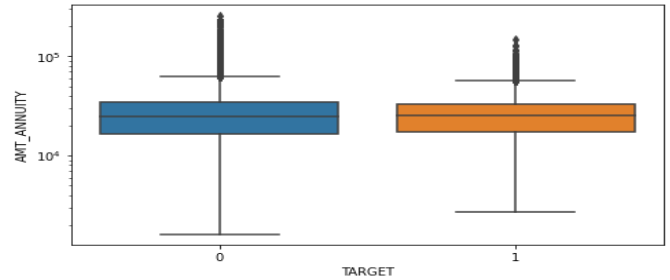

Density
CNT_CHILDREN

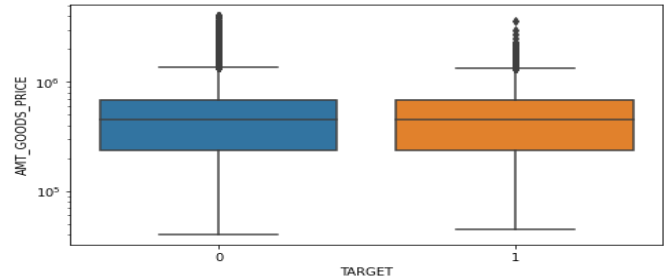

Density
HOUR_APPR_PROCESS_START_Previous

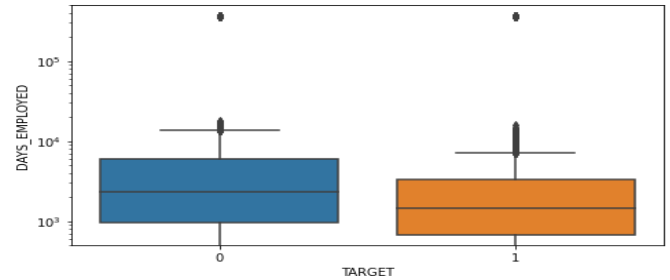

Approved
Canceled
Refused
Unused offer
Density
HOUR_APPR_PROCESS_START_Current

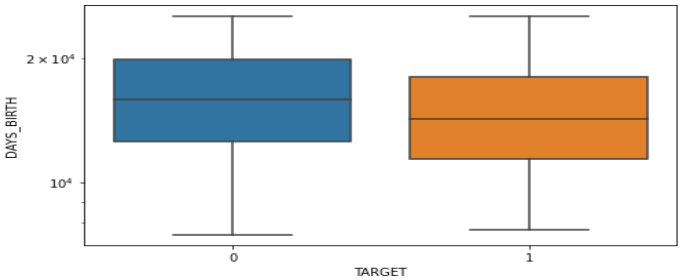

AMT_INCOME_TOTAL
TARGET

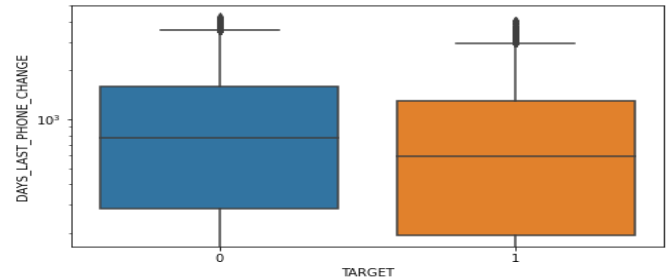

AMT_CREDIT
TARGET

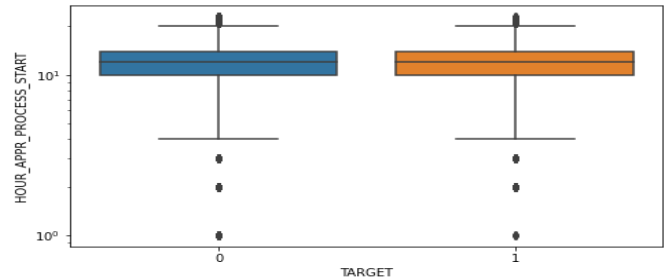

AMT_ANNUITY
TARGET

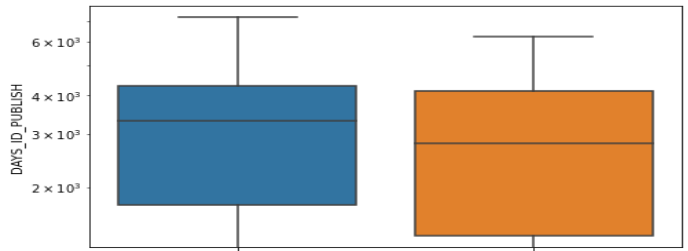

AMT_GOODS_PRICE
TARGET

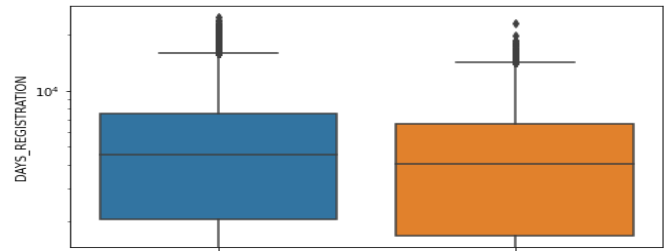

DAYS_EMPLOYED
TARGET

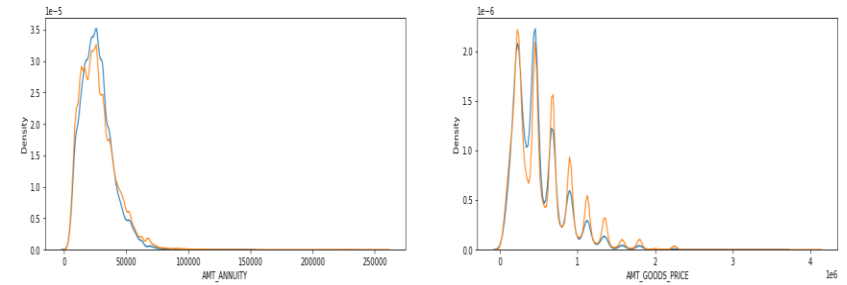

DAYS_BIRTH
TARGET


DAYS_LAST_PHONE_CHANGE
TARGET

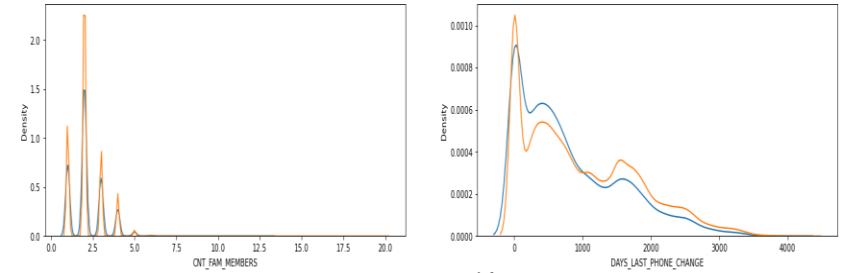

HOUR_APPR_PROCESS_START
TARGET


DAYS_ID_PUBLISH

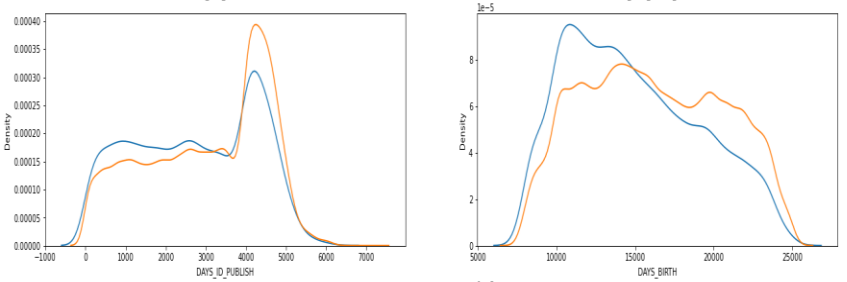

DAYS_REGISTRATION


1e−5
Density
AMT_ANNUITY

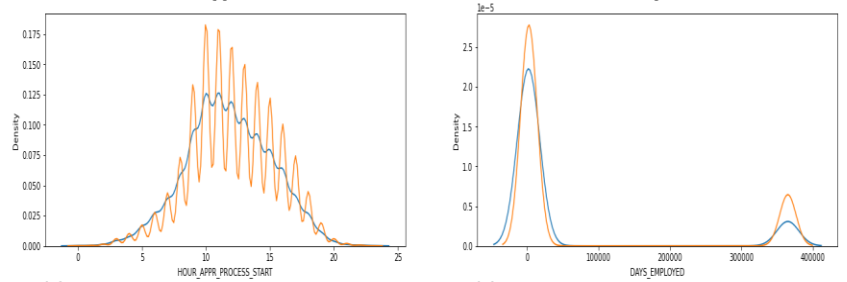
1e−6
Density
AMT_GOODS_PRICE
1e6


Density
CNT_FAM_MEMBERS

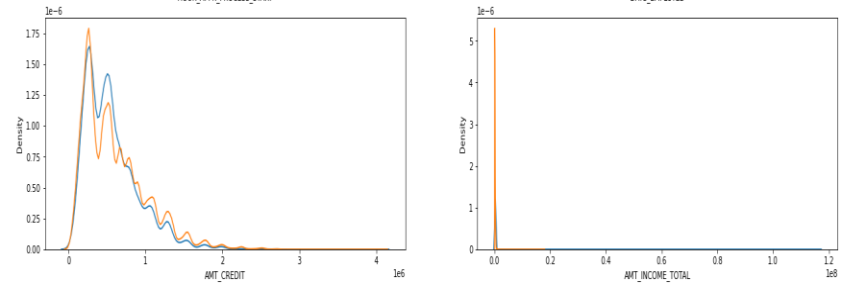
Density
DAYS_LAST_PHONE_CHANGE


Density
DAYS_ID_PUBLISH

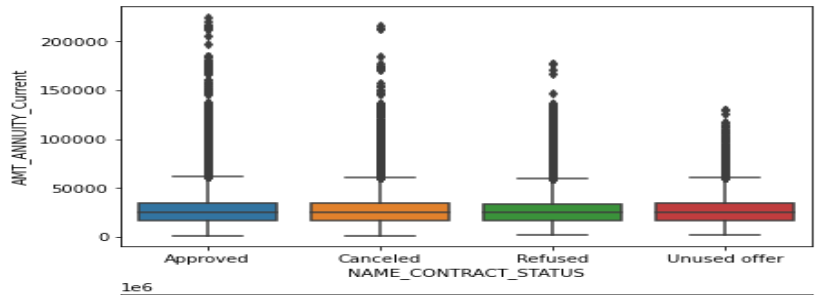
1e−5
Density
DAYS_BIRTH

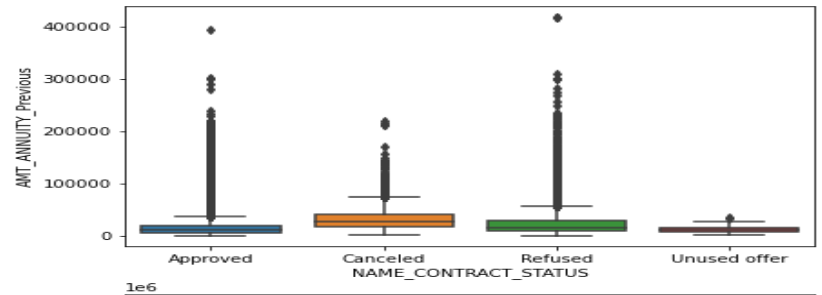

Density
HOUR_APPR_PROCESS_START

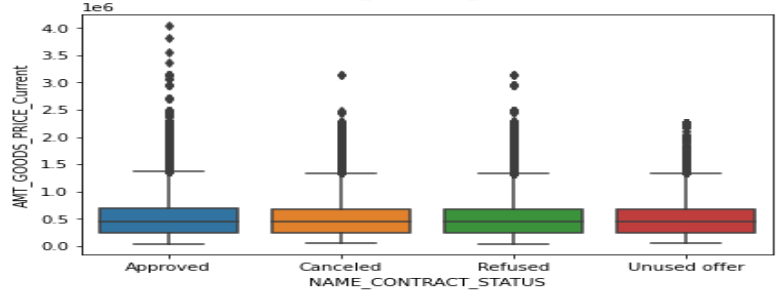
1e−5
Density
DAYS_EMPLOYED

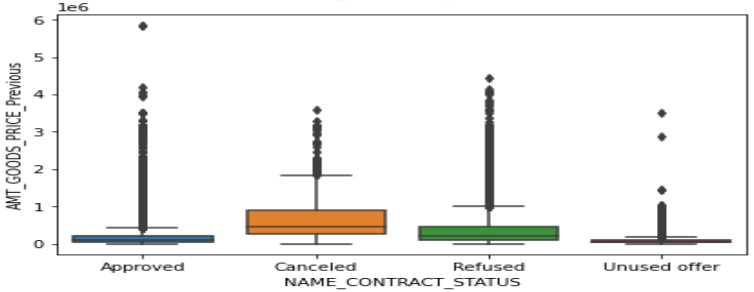

1e−6
Density
AMT_CREDIT
1e6

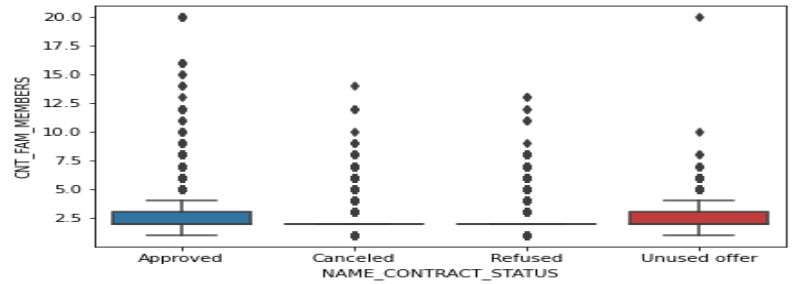
1e−6
Density
AMT_INCOME_TOTAL
1e8

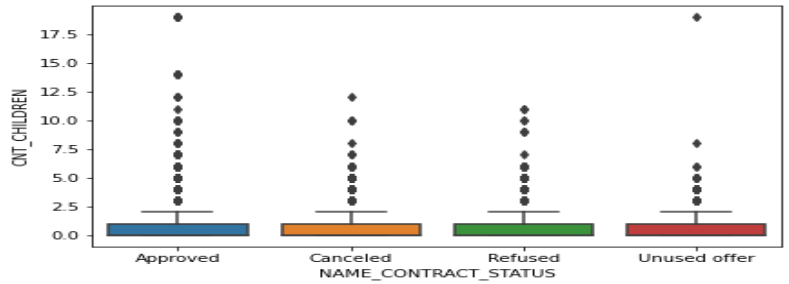

AMT_ANNUITY_Current
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

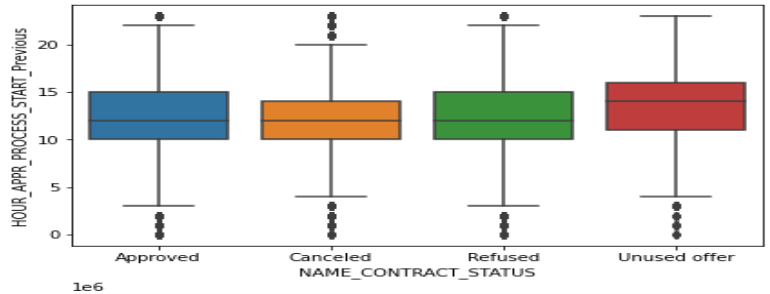

AMT_ANNUITY_Previous
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

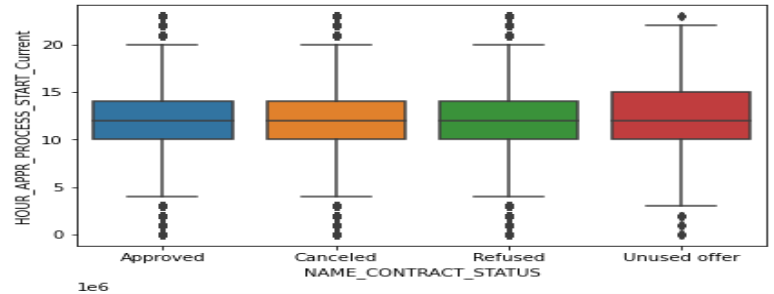

1e6
AMT_GOODS_PRICE_Current
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

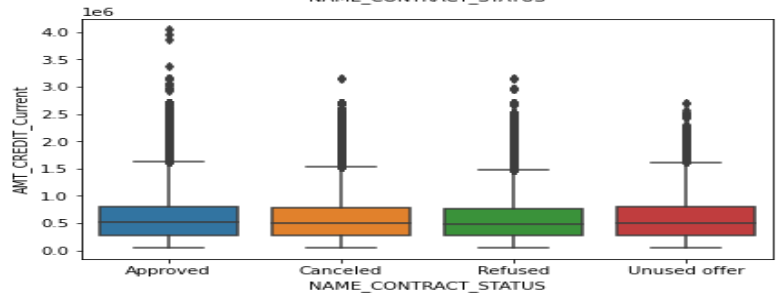

1e6
AMT_GOODS_PRICE_Previous
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

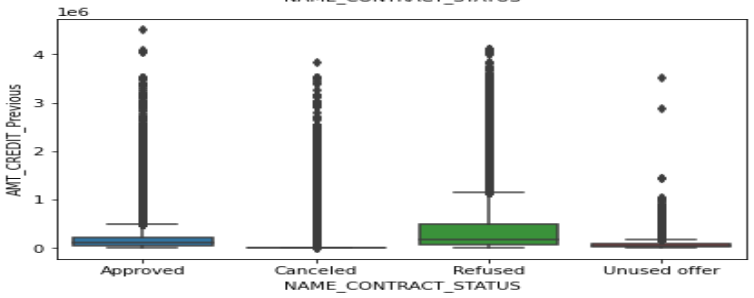

CNT_FAM_MEMBERS
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS


CNT_CHILDREN
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS


HOUR_APPR_PROCESS_START_Previous
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS


HOUR_APPR_PROCESS_START_Current
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS


1e6
AMT_CREDIT_Current
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS


1e6
AMT_CREDIT_Previous
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

Thank You