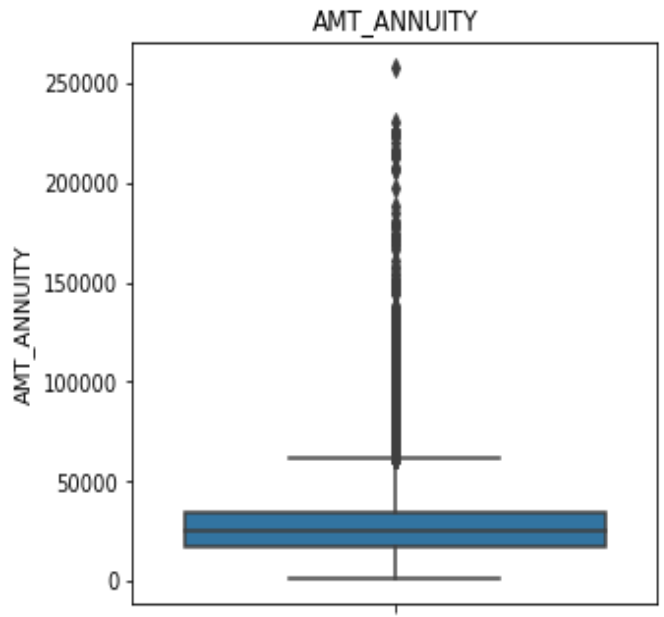
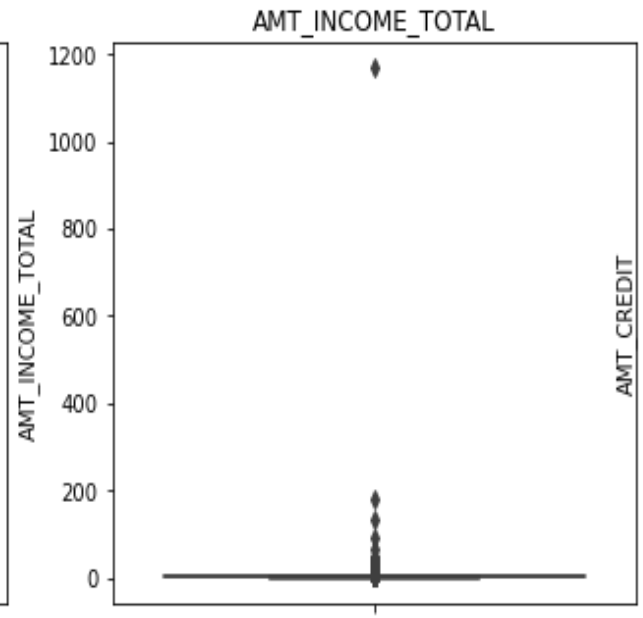
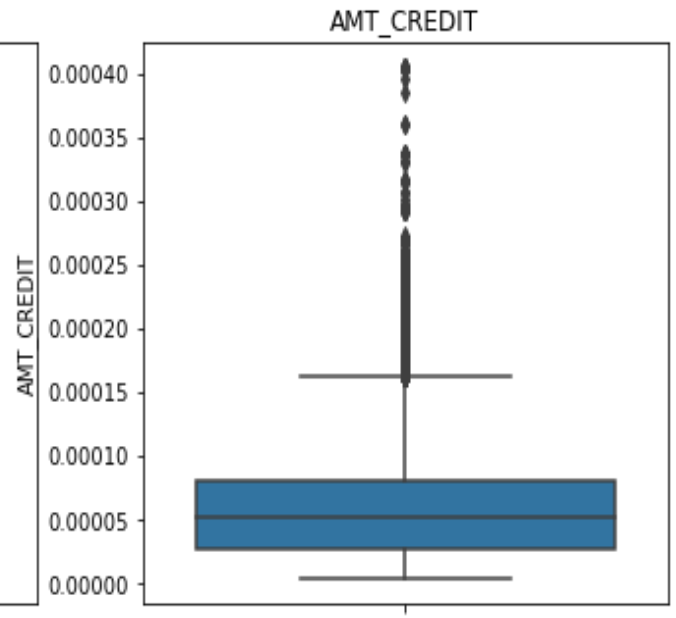
EDA Credit Assignment

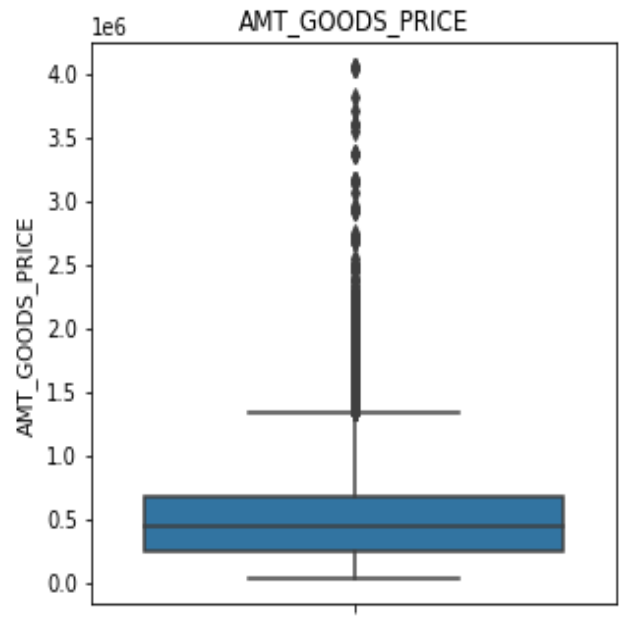
BY-ANUJA MORDE.

Outliers in application dataset

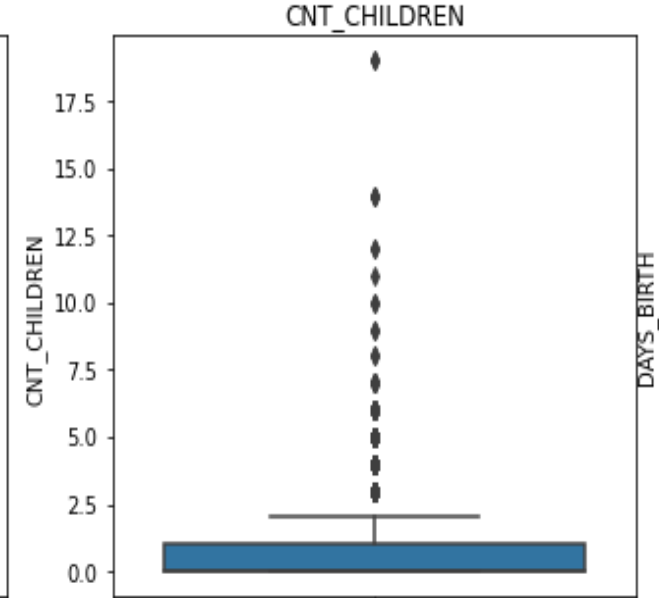

AMT_ANNUITY
AMT_ANNUITY
250000
200000
150000
100000
50000
0

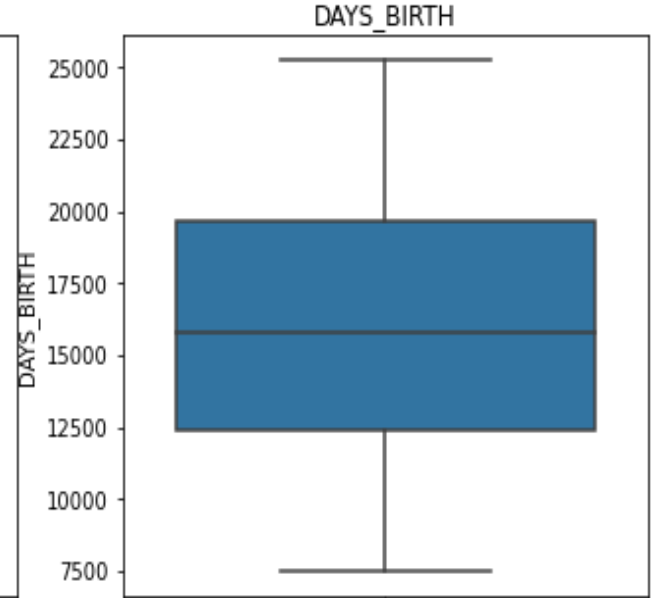

AMT_INCOME_TOTAL
AMT_INCOME_TOTAL
1200
1000
800
600
400
200
0


AMT_CREDIT
AMT_CREDIT
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000


AMT_GOODS_PRICE
1e6
AMT_GOODS_PRICE
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

DAYS_EMPLOYED
DAYS_EMPLOYED
350000
300000
250000
200000
150000
100000
50000
0

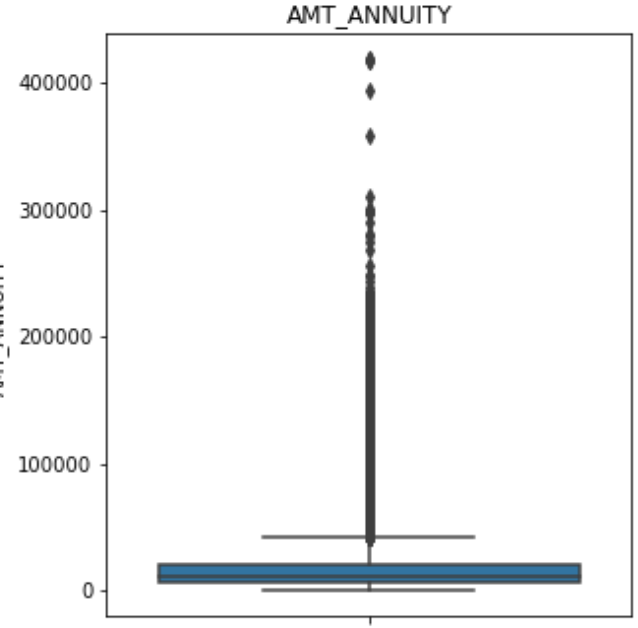

CNT_CHILDREN
CNT_CHILDREN
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

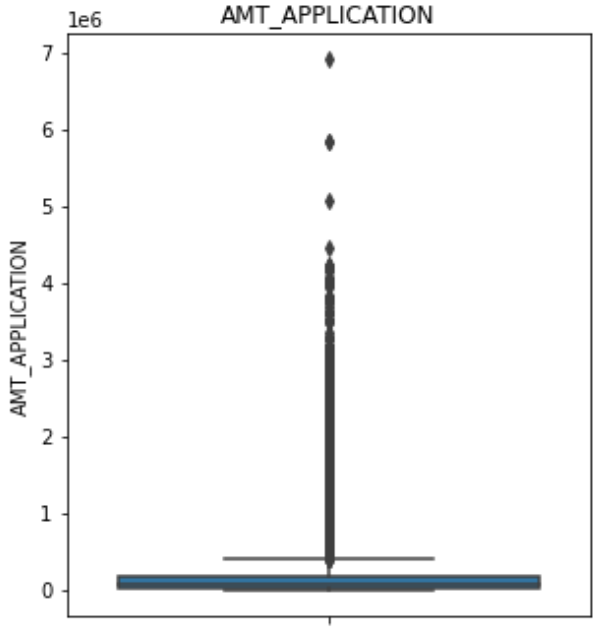

DAYS_BIRTH
DAYS_BIRTH
25000
22500
20000
17500
15000
12500
10000
7500

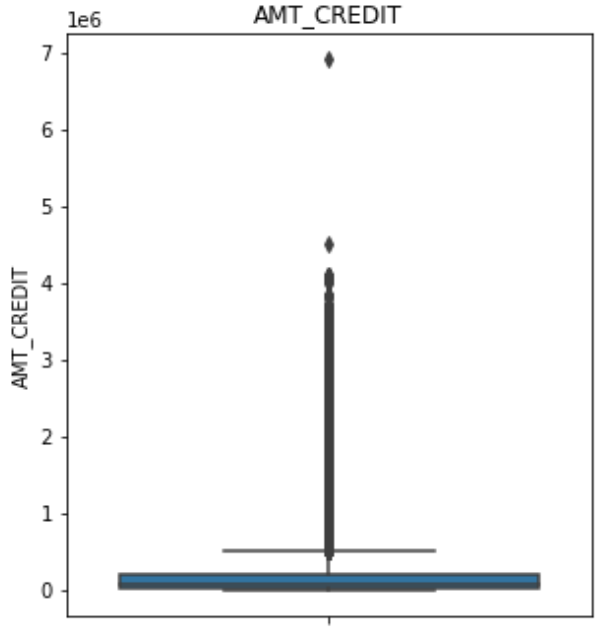
Inference on outliers of application dataset

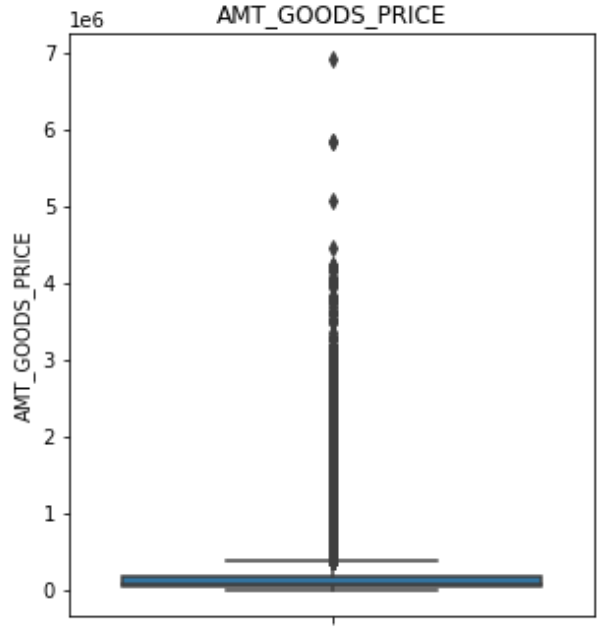
- AMT_ANNUITY,AMT_CREDIT,AMT_GOODS_PRICE,CNT_CHILDREN has outliers.
- DAYS_EMPLOYED entry is incorrect as it has outliers having value 36452 days
- AMT_INCOME_TOTAL has outliers in huge, it states that applicants has high income.
- DAYS_BIRTH has no outliers.

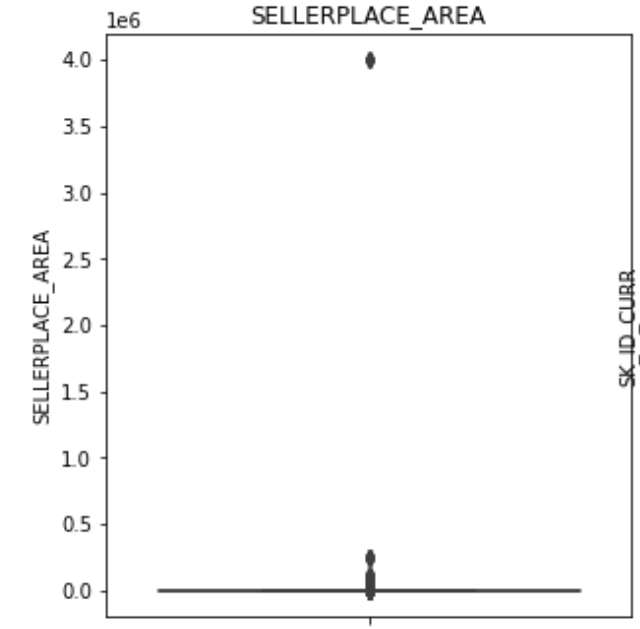
Outliers in Previous data set

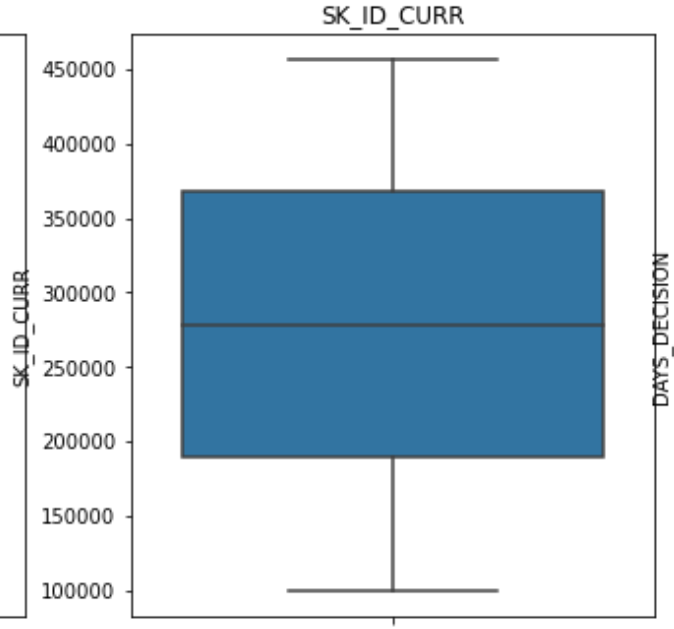

AMT_ANNUITY
400000
300000
200000
100000
0

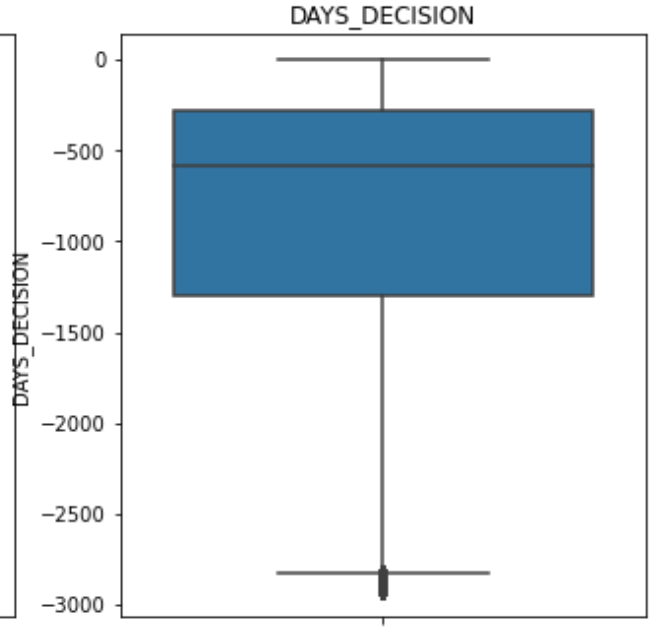

AMT_APPLICATION
1e6
7
6
5
4
3
2
1
0

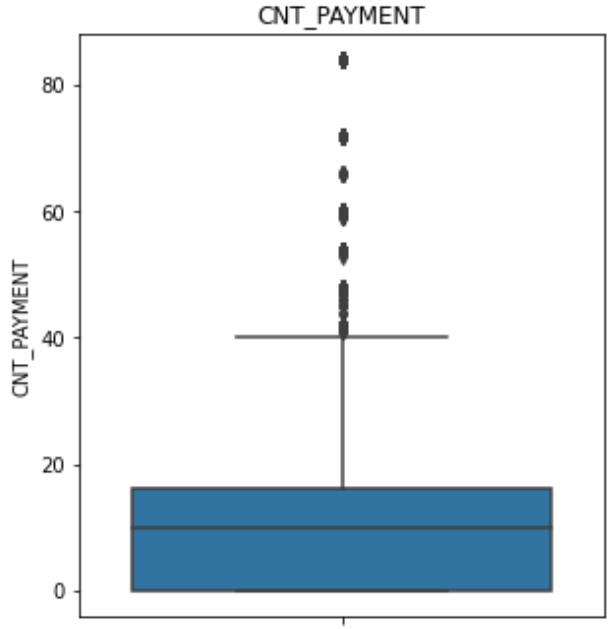

AMT_CREDIT
1e6
7
6
5
4
3
2
1
0


AMT_GOODS_PRICE
1e6
7
6
5
4
3
2
1
0


SELLERPLACE_AREA
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0


SK_ID_CURR
450000
400000
350000
300000
250000
200000
150000
100000


DAYS_DECISION
0
−500
−1000
−1500
−2000
−2500
−3000

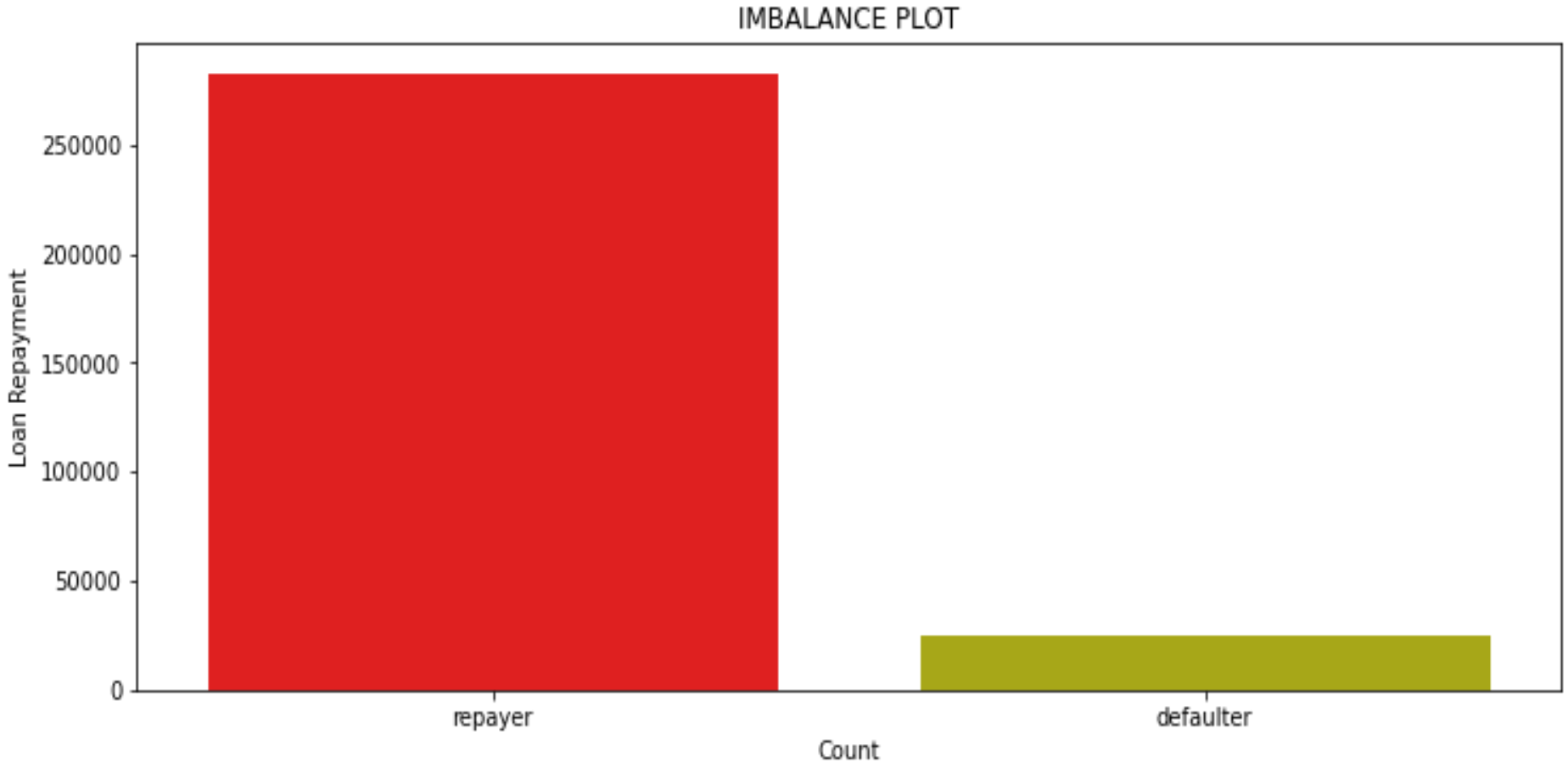

CNT_PAYMENT
80
60
40
20
0

Inference on outliers of previous dataset

- CNT_PAYMENT has less outliers

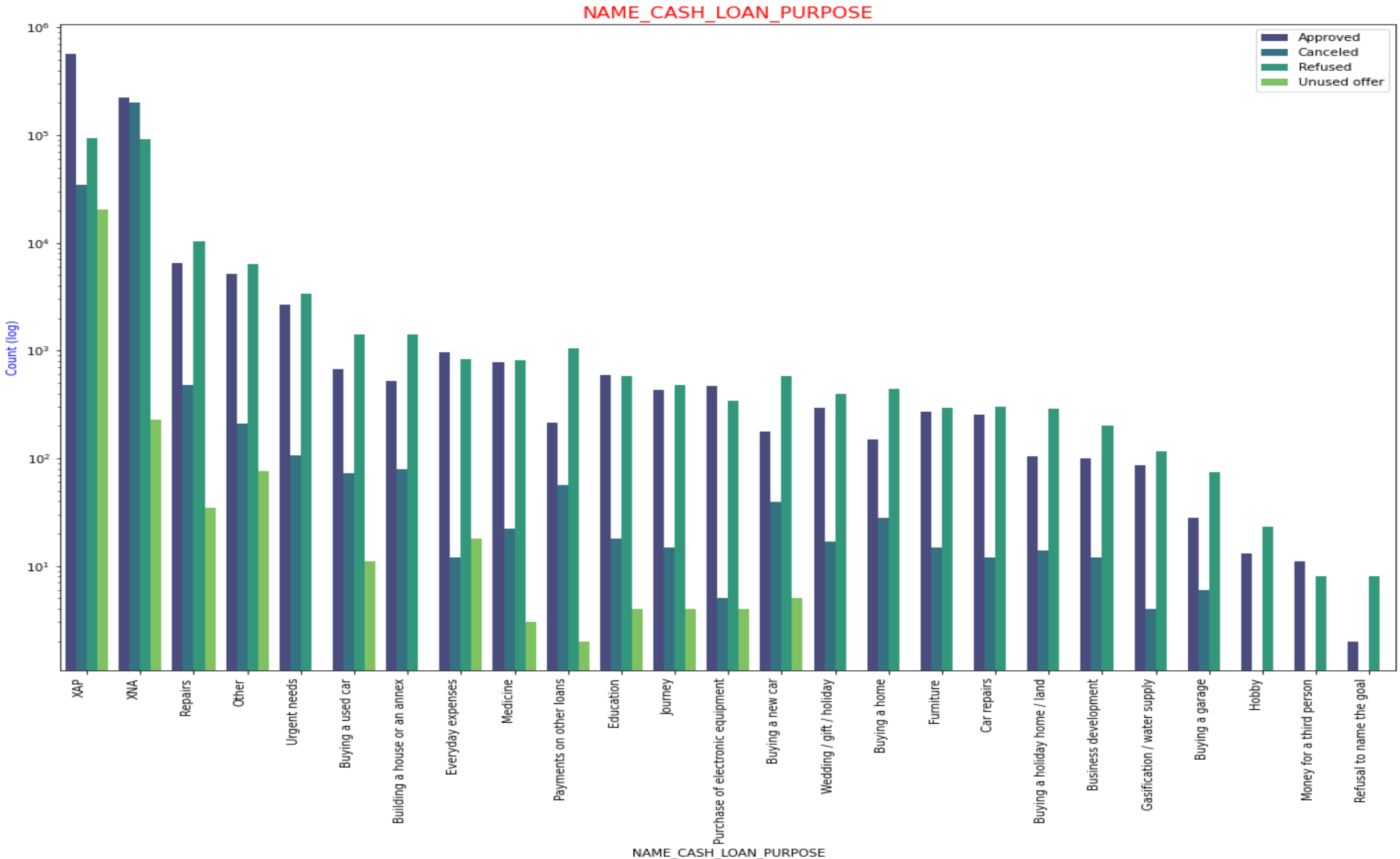
- AMT_APPLICATION,AMT_CREDIT,AMT_ANNUITY,SELLER_PRICE, GOOD_PRICE has huge number of outliers.

Imbalance ratio of repayer and defaulter



- Ratio of imbalance in percentage of REPAYER is 91.93%
- Ratio of imbalance in percentage of DEFAULTER is 8.07%

Analysis after Merging Data

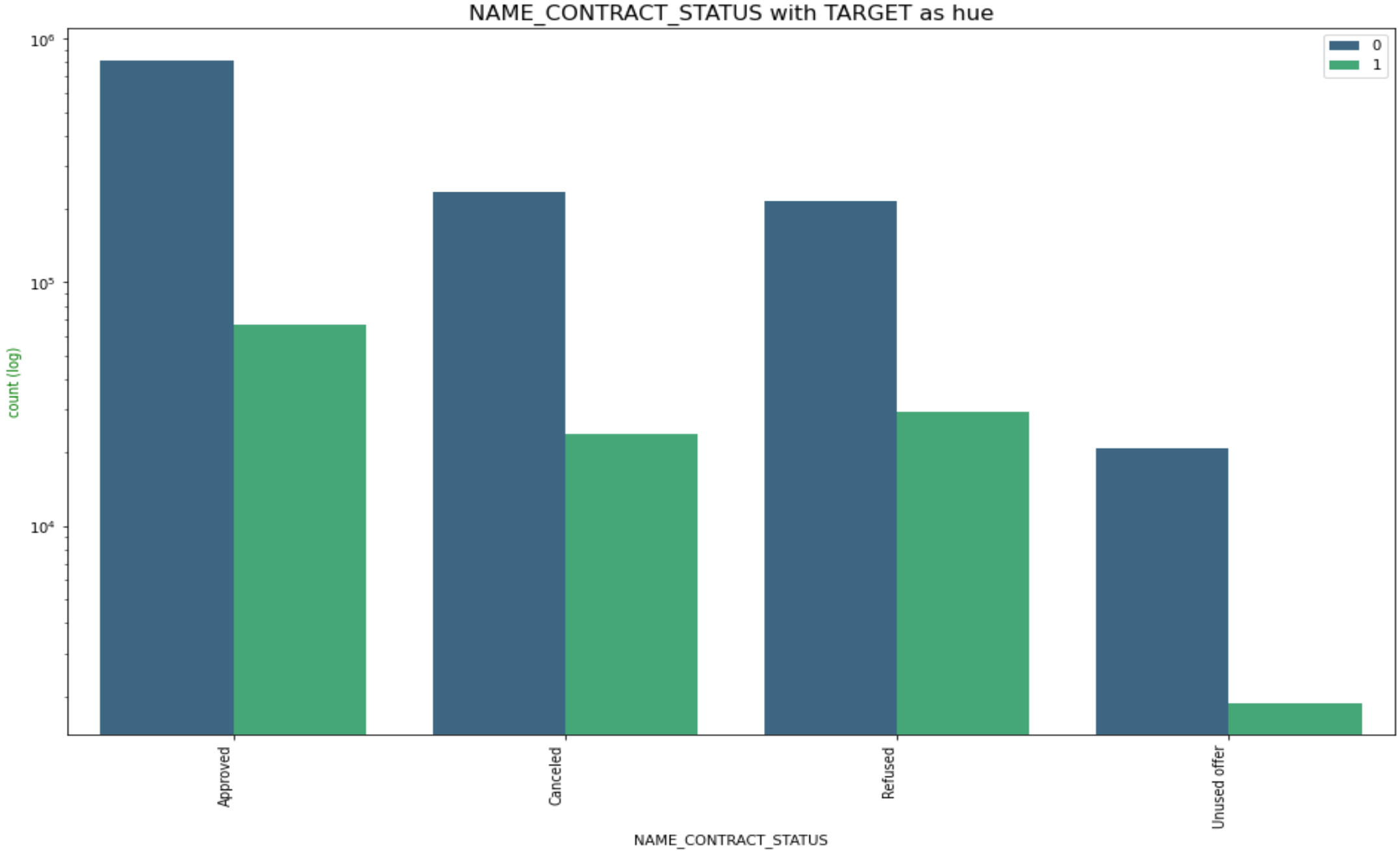

NAME_CASH_LOAN_PURPOSE
Approved
Canceled
Refused
Unused offer
Count (log)
10^6
10^5
10^4
10^3
10^2
10^1
XAP
XNA
Repairs
Other
Urgent needs
Buying a used car
Building a house or an annex
Everyday expenses
Medicine
Payments on other loans
Education
Journey
Purchase of electronic equipment
Buying a new car
Wedding / gift / holiday
Buying a home
Furniture
Car repairs
Buying a holiday home / land
Business development
Gasification / water supply
Buying a garage
Hobby
Money for a third person
Refusal to name the goal
NAME_CASH_LOAN_PURPOSE

Inference on NAME_CASH_LOAN_PURPOSE

- Highest default rate is seen when loan is taken for the purpose of repairs.

- It ha high number of unknown values.

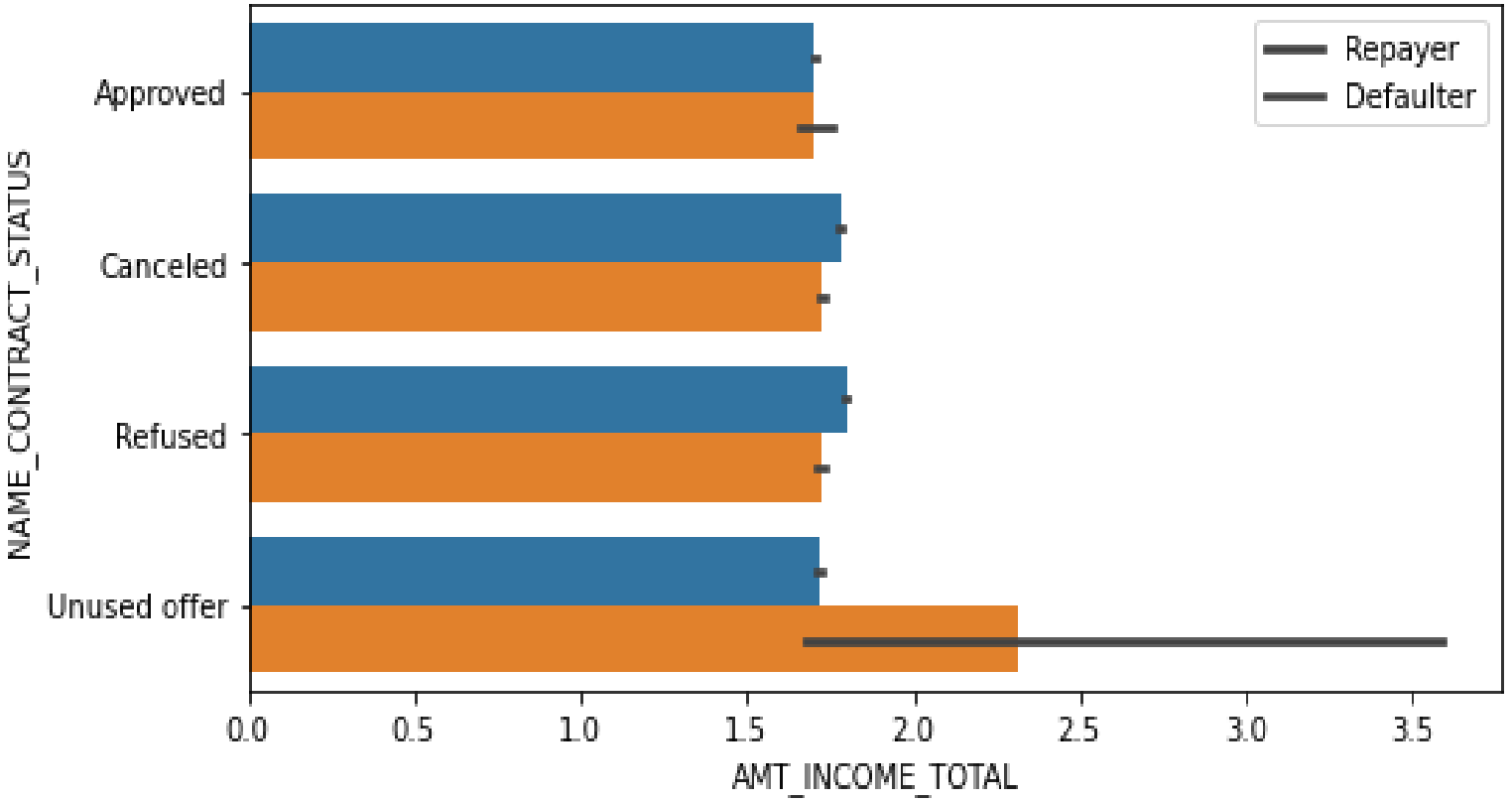
- A high number of application has been rejected by bank or refused.


NAME_CONTRACT_STATUS with TARGET as hue
0
1
count (log)
Approved
Canceled
Refused
Unused offer
NAME_CONTRACT_STATUS

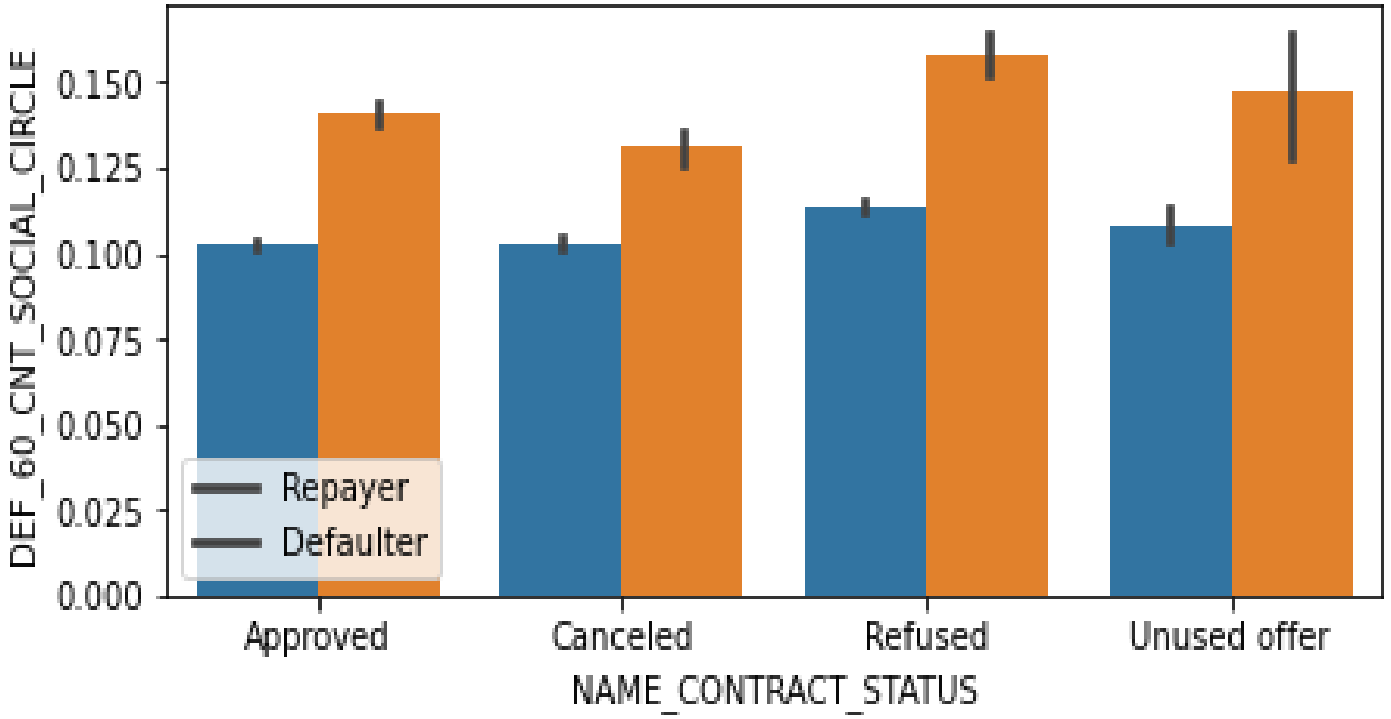
Inference on NAME_CONTRACT_STATUS

- The rate of earlier cancelled client is 90%.

- The 88% rate of earlier refused loan has repayed the loan.



- The above plot says people who has earlier not used offer have defaulted even though income is high.



- The above plot shows who defaulted between CONTRACT_STATUS and SOCIAL_CIRCLE.

CONCLUSION

- There are few attributes of customers that bank would be able to decide if they will repay the loan or not.

  - Factors whether customers will be able to repay the loan.

- Days Employed- customers who has more than 50 years of experience.
- Name income type – businessmen are less tend to default the loan.
- Name education type- degrees has less defaults.

  - Factors whether customers will not be able to repay(defaulters)

- Occupation type- labours , drivers and customers with low income.
- Amt good price- credit amount exceeds 4m.
- Days birth- young people should be excluded for loans.

THANK-YOU