ovf0oipqg

August 5, 2023

0.0.1 IMPORT LIBRERIES

[47]:
```
import numpy as np
import pandas as pd


import matplotlib.pyplot as plt
import seaborn as sns


import warnings
warnings.filterwarnings(action='ignore')

pd.set_option("display.max_rows",None)
pd.set_option("display.max_columns",None)
```

0.0.2 LOAD THE FILE

[48]:
```
df1=pd.read_csv("application_data.csv")
df2=pd.read_csv("previous_application.csv")
```

[49]:
```
df1.head()
```

[49]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

   AMT_GOODS_PRICE NAME_TYPE_SUITE NAME_INCOME_TYPE  \
```

```
0        351000.0   Unaccompanied          Working
1       1129500.0          Family    State servant
2        135000.0   Unaccompanied          Working
3        297000.0   Unaccompanied          Working
4        513000.0   Unaccompanied          Working

            NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  \
0  Secondary / secondary special  Single / not married  House / apartment
1               Higher education               Married  House / apartment
2  Secondary / secondary special  Single / not married  House / apartment
3  Secondary / secondary special        Civil marriage  House / apartment
4  Secondary / secondary special  Single / not married  House / apartment

   REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  \
0                    0.018801       -9461           -637            -3648.0
1                    0.003541      -16765          -1188            -1186.0
2                    0.010032      -19046           -225            -4260.0
3                    0.008019      -19005          -3039            -9833.0
4                    0.028663      -19932          -3038            -4311.0

   DAYS_ID_PUBLISH  OWN_CAR_AGE  FLAG_MOBIL  FLAG_EMP_PHONE  FLAG_WORK_PHONE  \
0            -2120          NaN           1               1                0
1             -291          NaN           1               1                0
2            -2531         26.0           1               1                1
3            -2437          NaN           1               1                0
4            -3458          NaN           1               1                0

   FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS  \
0                 1           1           0        Laborers              1.0
1                 1           1           0      Core staff              2.0
2                 1           1           0        Laborers              1.0
3                 1           0           0        Laborers              2.0
4                 1           0           0      Core staff              1.0

   REGION_RATING_CLIENT  REGION_RATING_CLIENT_W_CITY  \
0                     2                            2
1                     1                            1
2                     2                            2
3                     2                            2
4                     2                            2

  WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  \
0                  WEDNESDAY                       10
1                     MONDAY                       11
2                     MONDAY                        9
3                  WEDNESDAY                       17
4                   THURSDAY                       11
```

```
   REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION  \
0                           0                           0
1                           0                           0
2                           0                           0
3                           0                           0
4                           0                           0

   LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY  \
0                            0                       0
1                            0                       0
2                            0                       0
3                            0                       0
4                            0                       0

   REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY       ORGANIZATION_TYPE  \
0                       0                        0  Business Entity Type 3
1                       0                        0                  School
2                       0                        0              Government
3                       0                        0  Business Entity Type 3
4                       1                        1                Religion

   EXT_SOURCE_1  EXT_SOURCE_2  EXT_SOURCE_3  APARTMENTS_AVG  BASEMENTAREA_AVG  \
0      0.083037      0.262949      0.139376          0.0247            0.0369
1      0.311267      0.622246           NaN          0.0959            0.0529
2           NaN      0.555912      0.729567             NaN               NaN
3           NaN      0.650442           NaN             NaN               NaN
4           NaN      0.322738           NaN             NaN               NaN

   YEARS_BEGINEXPLUATATION_AVG  YEARS_BUILD_AVG  COMMONAREA_AVG  \
0                       0.9722           0.6192          0.0143
1                       0.9851           0.7960          0.0605
2                          NaN              NaN             NaN
3                          NaN              NaN             NaN
4                          NaN              NaN             NaN

   ELEVATORS_AVG  ENTRANCES_AVG  FLOORSMAX_AVG  FLOORSMIN_AVG  LANDAREA_AVG  \
0           0.00         0.0690         0.0833         0.1250        0.0369
1           0.08         0.0345         0.2917         0.3333        0.0130
2            NaN            NaN            NaN            NaN           NaN
3            NaN            NaN            NaN            NaN           NaN
4            NaN            NaN            NaN            NaN           NaN

   LIVINGAPARTMENTS_AVG  LIVINGAREA_AVG  NONLIVINGAPARTMENTS_AVG  \
0                0.0202          0.0190                   0.0000
1                0.0773          0.0549                   0.0039
2                   NaN             NaN                      NaN
```

```
3                  NaN             NaN                         NaN
4                  NaN             NaN                         NaN

   NONLIVINGAREA_AVG  APARTMENTS_MODE  BASEMENTAREA_MODE  \
0             0.0000           0.0252             0.0383
1             0.0098           0.0924             0.0538
2                NaN              NaN                NaN
3                NaN              NaN                NaN
4                NaN              NaN                NaN

   YEARS_BEGINEXPLUATATION_MODE  YEARS_BUILD_MODE  COMMONAREA_MODE  \
0                        0.9722            0.6341           0.0144
1                        0.9851            0.8040           0.0497
2                           NaN               NaN              NaN
3                           NaN               NaN              NaN
4                           NaN               NaN              NaN

   ELEVATORS_MODE  ENTRANCES_MODE  FLOORSMAX_MODE  FLOORSMIN_MODE  \
0          0.0000          0.0690          0.0833          0.1250
1          0.0806          0.0345          0.2917          0.3333
2             NaN             NaN             NaN             NaN
3             NaN             NaN             NaN             NaN
4             NaN             NaN             NaN             NaN

   LANDAREA_MODE  LIVINGAPARTMENTS_MODE  LIVINGAREA_MODE  \
0         0.0377                  0.022           0.0198
1         0.0128                  0.079           0.0554
2            NaN                    NaN              NaN
3            NaN                    NaN              NaN
4            NaN                    NaN              NaN

   NONLIVINGAPARTMENTS_MODE  NONLIVINGAREA_MODE  APARTMENTS_MEDI  \
0                       0.0                 0.0           0.0250
1                       0.0                 0.0           0.0968
2                       NaN                 NaN              NaN
3                       NaN                 NaN              NaN
4                       NaN                 NaN              NaN

   BASEMENTAREA_MEDI  YEARS_BEGINEXPLUATATION_MEDI  YEARS_BUILD_MEDI  \
0             0.0369                        0.9722            0.6243
1             0.0529                        0.9851            0.7987
2                NaN                           NaN               NaN
3                NaN                           NaN               NaN
4                NaN                           NaN               NaN

   COMMONAREA_MEDI  ELEVATORS_MEDI  ENTRANCES_MEDI  FLOORSMAX_MEDI  \
0           0.0144            0.00          0.0690          0.0833
```

```
1          0.0608           0.08          0.0345          0.2917
2             NaN            NaN             NaN             NaN
3             NaN            NaN             NaN             NaN
4             NaN            NaN             NaN             NaN

   FLOORSMIN_MEDI  LANDAREA_MEDI  LIVINGAPARTMENTS_MEDI  LIVINGAREA_MEDI  \
0          0.1250         0.0375                 0.0205           0.0193
1          0.3333         0.0132                 0.0787           0.0558
2             NaN            NaN                    NaN              NaN
3             NaN            NaN                    NaN              NaN
4             NaN            NaN                    NaN              NaN

   NONLIVINGAPARTMENTS_MEDI  NONLIVINGAREA_MEDI FONDKAPREMONT_MODE  \
0                    0.0000                0.00   reg oper account
1                    0.0039                0.01   reg oper account
2                       NaN                 NaN                NaN
3                       NaN                 NaN                NaN
4                       NaN                 NaN                NaN

   HOUSETYPE_MODE  TOTALAREA_MODE WALLSMATERIAL_MODE EMERGENCYSTATE_MODE  \
0  block of flats          0.0149       Stone, brick                  No
1  block of flats          0.0714              Block                  No
2             NaN             NaN                NaN                 NaN
3             NaN             NaN                NaN                 NaN
4             NaN             NaN                NaN                 NaN

   OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       0.0                       0.0
3                       2.0                       0.0
4                       0.0                       0.0

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
0                       2.0                       2.0                 -1134.0
1                       1.0                       0.0                  -828.0
2                       0.0                       0.0                  -815.0
3                       2.0                       0.0                  -617.0
4                       0.0                       0.0                 -1106.0

   FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  FLAG_DOCUMENT_4  FLAG_DOCUMENT_5  \
0                0                1                0                0
1                0                1                0                0
2                0                0                0                0
3                0                1                0                0
4                0                0                0                0
```

```
   FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  FLAG_DOCUMENT_8  FLAG_DOCUMENT_9  \
0                0                0                0                0   
1                0                0                0                0   
2                0                0                0                0   
3                0                0                0                0   
4                0                0                1                0   

   FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  \
0                 0                 0                 0                 0   
1                 0                 0                 0                 0   
2                 0                 0                 0                 0   
3                 0                 0                 0                 0   
4                 0                 0                 0                 0   

   FLAG_DOCUMENT_14  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16  FLAG_DOCUMENT_17  \
0                 0                 0                 0                 0   
1                 0                 0                 0                 0   
2                 0                 0                 0                 0   
3                 0                 0                 0                 0   
4                 0                 0                 0                 0   

   FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  \
0                 0                 0                 0                 0   
1                 0                 0                 0                 0   
2                 0                 0                 0                 0   
3                 0                 0                 0                 0   
4                 0                 0                 0                 0   

   AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
0                         0.0                        0.0   
1                         0.0                        0.0   
2                         0.0                        0.0   
3                         NaN                        NaN   
4                         0.0                        0.0   

   AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                         0.0                        0.0   
1                         0.0                        0.0   
2                         0.0                        0.0   
3                         NaN                        NaN   
4                         0.0                        0.0   

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR  
0                        0.0                         1.0  
1                        0.0                         0.0  
2                        0.0                         0.0  
3                        NaN                         NaN  
```

```
4                        0.0                         0.0
```

[50]:
```python
df2.head()
```

[50]:
```
   SK_ID_PREV  SK_ID_CURR  TARGET NAME_CONTRACT_TYPE  AMT_ANNUITY  \
0     2030495      271877     0.0     Consumer loans     1730.430
1     2802425      108129     0.0         Cash loans    25188.615
2     2523466      122040     0.0         Cash loans    15060.735
3     2819243      176158     0.0         Cash loans    47041.335
4     1784265      202054     0.0         Cash loans    31924.395

   AMT_APPLICATION  AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE  \
0          17145.0     17145.0               0.0          17145.0
1         607500.0    679671.0               NaN         607500.0
2         112500.0    136444.5               NaN         112500.0
3         450000.0    470790.0               NaN         450000.0
4         337500.0    404055.0               NaN         337500.0

  WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  \
0                   SATURDAY                       15
1                   THURSDAY                       11
2                    TUESDAY                       11
3                     MONDAY                        7
4                   THURSDAY                        9

  FLAG_LAST_APPL_PER_CONTRACT  NFLAG_LAST_APPL_IN_DAY  RATE_DOWN_PAYMENT  \
0                           Y                       1                0.0
1                           Y                       1                NaN
2                           Y                       1                NaN
3                           Y                       1                NaN
4                           Y                       1                NaN

   RATE_INTEREST_PRIMARY  RATE_INTEREST_PRIVILEGED NAME_CASH_LOAN_PURPOSE  \
0               0.182832                  0.867336                    XAP
1                    NaN                       NaN                    XNA
2                    NaN                       NaN                    XNA
3                    NaN                       NaN                    XNA
4                    NaN                       NaN                Repairs

  NAME_CONTRACT_STATUS  DAYS_DECISION      NAME_PAYMENT_TYPE  \
0             Approved            -73  Cash through the bank
1             Approved           -164                    XNA
2             Approved           -301  Cash through the bank
3             Approved           -512  Cash through the bank
4              Refused           -781  Cash through the bank

  CODE_REJECT_REASON  NAME_TYPE_SUITE NAME_CLIENT_TYPE NAME_GOODS_CATEGORY  \
```

```
0                XAP              NaN        Repeater             Mobile
1                XAP    Unaccompanied        Repeater                XNA
2                XAP  Spouse, partner        Repeater                XNA
3                XAP              NaN        Repeater                XNA
4                 HC              NaN        Repeater                XNA

  NAME_PORTFOLIO NAME_PRODUCT_TYPE             CHANNEL_TYPE  SELLERPLACE_AREA  \
0            POS               XNA             Country-wide                35
1           Cash            x-sell           Contact center                -1
2           Cash            x-sell  Credit and cash offices                -1
3           Cash            x-sell  Credit and cash offices                -1
4           Cash           walk-in  Credit and cash offices                -1

  NAME_SELLER_INDUSTRY  CNT_PAYMENT NAME_YIELD_GROUP  \
0         Connectivity         12.0           middle
1                  XNA         36.0       low_action
2                  XNA         12.0             high
3                  XNA         12.0           middle
4                  XNA         24.0             high

       PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE  \
0  POS mobile with interest            365243.0           -42.0
1          Cash X-Sell: low            365243.0          -134.0
2         Cash X-Sell: high            365243.0          -271.0
3       Cash X-Sell: middle            365243.0          -482.0
4         Cash Street: high                 NaN             NaN

   DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION  \
0                      300.0          -42.0             -37.0
1                      916.0       365243.0          365243.0
2                       59.0       365243.0          365243.0
3                     -152.0         -182.0            -177.0
4                        NaN            NaN               NaN

   NFLAG_INSURED_ON_APPROVAL
0                        0.0
1                        1.0
2                        1.0
3                        1.0
4                        NaN
```

[51]:
```
df1.shape
```

[51]: (307511, 122)

[52]:
```
df2.shape
```

[52]: (1048575, 38)

[53]:
```
df1.columns
```

[53]:
```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY',
       …
       'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
       'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR',
       'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
       'AMT_REQ_CREDIT_BUREAU_YEAR'],
      dtype='object', length=122)
```

[54]:
```
df2.columns
```

[54]:
```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE',
       'AMT_ANNUITY', 'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT',
       'AMT_GOODS_PRICE', 'WEEKDAY_APPR_PROCESS_START',
       'HOUR_APPR_PROCESS_START', 'FLAG_LAST_APPL_PER_CONTRACT',
       'NFLAG_LAST_APPL_IN_DAY', 'RATE_DOWN_PAYMENT', 'RATE_INTEREST_PRIMARY',
       'RATE_INTEREST_PRIVILEGED', 'NAME_CASH_LOAN_PURPOSE',
       'NAME_CONTRACT_STATUS', 'DAYS_DECISION', 'NAME_PAYMENT_TYPE',
       'CODE_REJECT_REASON', 'NAME_TYPE_SUITE', 'NAME_CLIENT_TYPE',
       'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
       'CHANNEL_TYPE', 'SELLERPLACE_AREA', 'NAME_SELLER_INDUSTRY',
       'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION',
       'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION',
       'DAYS_LAST_DUE', 'DAYS_TERMINATION', 'NFLAG_INSURED_ON_APPROVAL'],
      dtype='object')
```

0.0.3 DATA CLEANING

MANAGING MISSING VALUES

[55]:
```
100*df1.isnull().mean()
```

[55]:
```
SK_ID_CURR                    0.000000
TARGET                        0.000000
NAME_CONTRACT_TYPE            0.000000
CODE_GENDER                   0.000000
FLAG_OWN_CAR                  0.000000
FLAG_OWN_REALTY               0.000000
CNT_CHILDREN                  0.000000
AMT_INCOME_TOTAL              0.000000
AMT_CREDIT                    0.000000
AMT_ANNUITY                   0.003902
```

AMT_GOODS_PRICE	0.090403
NAME_TYPE_SUITE	0.420148
NAME_INCOME_TYPE	0.000000
NAME_EDUCATION_TYPE	0.000000
NAME_FAMILY_STATUS	0.000000
NAME_HOUSING_TYPE	0.000000
REGION_POPULATION_RELATIVE	0.000000
DAYS_BIRTH	0.000000
DAYS_EMPLOYED	0.000000
DAYS_REGISTRATION	0.000000
DAYS_ID_PUBLISH	0.000000
OWN_CAR_AGE	65.990810
FLAG_MOBIL	0.000000
FLAG_EMP_PHONE	0.000000
FLAG_WORK_PHONE	0.000000
FLAG_CONT_MOBILE	0.000000
FLAG_PHONE	0.000000
FLAG_EMAIL	0.000000
OCCUPATION_TYPE	31.345545
CNT_FAM_MEMBERS	0.000650
REGION_RATING_CLIENT	0.000000
REGION_RATING_CLIENT_W_CITY	0.000000
WEEKDAY_APPR_PROCESS_START	0.000000
HOUR_APPR_PROCESS_START	0.000000
REG_REGION_NOT_LIVE_REGION	0.000000
REG_REGION_NOT_WORK_REGION	0.000000
LIVE_REGION_NOT_WORK_REGION	0.000000
REG_CITY_NOT_LIVE_CITY	0.000000
REG_CITY_NOT_WORK_CITY	0.000000
LIVE_CITY_NOT_WORK_CITY	0.000000
ORGANIZATION_TYPE	0.000000
EXT_SOURCE_1	56.381073
EXT_SOURCE_2	0.214626
EXT_SOURCE_3	19.825307
APARTMENTS_AVG	50.749729
BASEMENTAREA_AVG	58.515956
YEARS_BEGINEXPLUATATION_AVG	48.781019
YEARS_BUILD_AVG	66.497784
COMMONAREA_AVG	69.872297
ELEVATORS_AVG	53.295980
ENTRANCES_AVG	50.348768
FLOORSMAX_AVG	49.760822
FLOORSMIN_AVG	67.848630
LANDAREA_AVG	59.376738
LIVINGAPARTMENTS_AVG	68.354953
LIVINGAREA_AVG	50.193326
NONLIVINGAPARTMENTS_AVG	69.432963

```
NONLIVINGAREA_AVG              55.179164
APARTMENTS_MODE                50.749729
BASEMENTAREA_MODE              58.515956
YEARS_BEGINEXPLUATATION_MODE   48.781019
YEARS_BUILD_MODE               66.497784
COMMONAREA_MODE                69.872297
ELEVATORS_MODE                 53.295980
ENTRANCES_MODE                 50.348768
FLOORSMAX_MODE                 49.760822
FLOORSMIN_MODE                 67.848630
LANDAREA_MODE                  59.376738
LIVINGAPARTMENTS_MODE          68.354953
LIVINGAREA_MODE                50.193326
NONLIVINGAPARTMENTS_MODE       69.432963
NONLIVINGAREA_MODE             55.179164
APARTMENTS_MEDI                50.749729
BASEMENTAREA_MEDI              58.515956
YEARS_BEGINEXPLUATATION_MEDI   48.781019
YEARS_BUILD_MEDI               66.497784
COMMONAREA_MEDI                69.872297
ELEVATORS_MEDI                 53.295980
ENTRANCES_MEDI                 50.348768
FLOORSMAX_MEDI                 49.760822
FLOORSMIN_MEDI                 67.848630
LANDAREA_MEDI                  59.376738
LIVINGAPARTMENTS_MEDI          68.354953
LIVINGAREA_MEDI                50.193326
NONLIVINGAPARTMENTS_MEDI       69.432963
NONLIVINGAREA_MEDI             55.179164
FONDKAPREMONT_MODE             68.386172
HOUSETYPE_MODE                 50.176091
TOTALAREA_MODE                 48.268517
WALLSMATERIAL_MODE             50.840783
EMERGENCYSTATE_MODE            47.398304
OBS_30_CNT_SOCIAL_CIRCLE        0.332021
DEF_30_CNT_SOCIAL_CIRCLE        0.332021
OBS_60_CNT_SOCIAL_CIRCLE        0.332021
DEF_60_CNT_SOCIAL_CIRCLE        0.332021
DAYS_LAST_PHONE_CHANGE          0.000325
FLAG_DOCUMENT_2                 0.000000
FLAG_DOCUMENT_3                 0.000000
FLAG_DOCUMENT_4                 0.000000
FLAG_DOCUMENT_5                 0.000000
FLAG_DOCUMENT_6                 0.000000
FLAG_DOCUMENT_7                 0.000000
FLAG_DOCUMENT_8                 0.000000
FLAG_DOCUMENT_9                 0.000000
```

```
FLAG_DOCUMENT_10                 0.000000
FLAG_DOCUMENT_11                 0.000000
FLAG_DOCUMENT_12                 0.000000
FLAG_DOCUMENT_13                 0.000000
FLAG_DOCUMENT_14                 0.000000
FLAG_DOCUMENT_15                 0.000000
FLAG_DOCUMENT_16                 0.000000
FLAG_DOCUMENT_17                 0.000000
FLAG_DOCUMENT_18                 0.000000
FLAG_DOCUMENT_19                 0.000000
FLAG_DOCUMENT_20                 0.000000
FLAG_DOCUMENT_21                 0.000000
AMT_REQ_CREDIT_BUREAU_HOUR      13.501631
AMT_REQ_CREDIT_BUREAU_DAY       13.501631
AMT_REQ_CREDIT_BUREAU_WEEK      13.501631
AMT_REQ_CREDIT_BUREAU_MON       13.501631
AMT_REQ_CREDIT_BUREAU_QRT       13.501631
AMT_REQ_CREDIT_BUREAU_YEAR      13.501631
dtype: float64
```

SORTING/FILTERING THE DATA FRAME

ELEMINATION OF EXTRA COLUMN FROM 1ST DATA FRAME

[56]:
```
extra_col=['AMT_GOODS_PRICE','NAME_TYPE_SUITE','REGION_POPULATION_RELATIVE','DAYS_REGISTRATION
 ↪'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20','FLAG_DOCUMENT_21',␣
 ↪'AMT_REQ_CREDIT_BUREAU_HOUR',
'AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_DAY','AMT_REQ_CREDIT_BUREAU_WEEK','AMT_REQ
 ↪'AMT_REQ_CREDIT_BUREAU_QRT',
'AMT_REQ_CREDIT_BUREAU_YEAR','OBS_30_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT
```

[57]:
```
df1=df1.drop(extra_col,axis=1)
```

[58]:
```
df1.shape
```

[58]:
```
(307511, 67)
```

[59]:
```
100*df1.isnull().mean()
```

[59]:
```
SK_ID_CURR                       0.000000
TARGET                           0.000000
NAME_CONTRACT_TYPE               0.000000
CODE_GENDER                      0.000000
FLAG_OWN_CAR                     0.000000
FLAG_OWN_REALTY                  0.000000
CNT_CHILDREN                     0.000000
AMT_INCOME_TOTAL                 0.000000
```

```
AMT_CREDIT                       0.000000
AMT_ANNUITY                      0.003902
NAME_INCOME_TYPE                 0.000000
NAME_EDUCATION_TYPE              0.000000
NAME_FAMILY_STATUS               0.000000
NAME_HOUSING_TYPE                0.000000
DAYS_BIRTH                       0.000000
DAYS_EMPLOYED                    0.000000
OWN_CAR_AGE                     65.990810
OCCUPATION_TYPE                 31.345545
CNT_FAM_MEMBERS                  0.000650
ORGANIZATION_TYPE                0.000000
APARTMENTS_AVG                  50.749729
BASEMENTAREA_AVG                58.515956
YEARS_BEGINEXPLUATATION_AVG     48.781019
YEARS_BUILD_AVG                 66.497784
COMMONAREA_AVG                  69.872297
ELEVATORS_AVG                   53.295980
ENTRANCES_AVG                   50.348768
FLOORSMAX_AVG                   49.760822
FLOORSMIN_AVG                   67.848630
LANDAREA_AVG                    59.376738
LIVINGAPARTMENTS_AVG            68.354953
LIVINGAREA_AVG                  50.193326
NONLIVINGAPARTMENTS_AVG         69.432963
NONLIVINGAREA_AVG               55.179164
APARTMENTS_MODE                 50.749729
BASEMENTAREA_MODE               58.515956
YEARS_BEGINEXPLUATATION_MODE    48.781019
YEARS_BUILD_MODE                66.497784
COMMONAREA_MODE                 69.872297
ELEVATORS_MODE                  53.295980
ENTRANCES_MODE                  50.348768
FLOORSMAX_MODE                  49.760822
FLOORSMIN_MODE                  67.848630
LANDAREA_MODE                   59.376738
LIVINGAPARTMENTS_MODE           68.354953
LIVINGAREA_MODE                 50.193326
NONLIVINGAPARTMENTS_MODE        69.432963
NONLIVINGAREA_MODE              55.179164
APARTMENTS_MEDI                 50.749729
BASEMENTAREA_MEDI               58.515956
YEARS_BEGINEXPLUATATION_MEDI    48.781019
YEARS_BUILD_MEDI                66.497784
COMMONAREA_MEDI                 69.872297
ELEVATORS_MEDI                  53.295980
ENTRANCES_MEDI                  50.348768
```

```
FLOORSMAX_MEDI                49.760822
FLOORSMIN_MEDI                67.848630
LANDAREA_MEDI                 59.376738
LIVINGAPARTMENTS_MEDI         68.354953
LIVINGAREA_MEDI               50.193326
NONLIVINGAPARTMENTS_MEDI      69.432963
NONLIVINGAREA_MEDI            55.179164
FONDKAPREMONT_MODE            68.386172
HOUSETYPE_MODE                50.176091
TOTALAREA_MODE                48.268517
WALLSMATERIAL_MODE            50.840783
EMERGENCYSTATE_MODE           47.398304
dtype: float64
```

[60]:
```
100*df2.isnull().mean()
```

[60]:
```
SK_ID_PREV                     0.000000
SK_ID_CURR                     0.000000
TARGET                        91.708461
NAME_CONTRACT_TYPE             0.000000
AMT_ANNUITY                   22.221491
AMT_APPLICATION                0.000000
AMT_CREDIT                     0.000000
AMT_DOWN_PAYMENT              53.348211
AMT_GOODS_PRICE               22.980235
WEEKDAY_APPR_PROCESS_START     0.000000
HOUR_APPR_PROCESS_START        0.000000
FLAG_LAST_APPL_PER_CONTRACT    0.000000
NFLAG_LAST_APPL_IN_DAY         0.000000
RATE_DOWN_PAYMENT             53.348211
RATE_INTEREST_PRIMARY         99.645137
RATE_INTEREST_PRIVILEGED      99.645137
NAME_CASH_LOAN_PURPOSE         0.000000
NAME_CONTRACT_STATUS           0.000000
DAYS_DECISION                  0.000000
NAME_PAYMENT_TYPE              0.000000
CODE_REJECT_REASON             0.000000
NAME_TYPE_SUITE               49.127626
NAME_CLIENT_TYPE               0.000000
NAME_GOODS_CATEGORY            0.000000
NAME_PORTFOLIO                 0.000000
NAME_PRODUCT_TYPE              0.000000
CHANNEL_TYPE                   0.000000
SELLERPLACE_AREA               0.000000
NAME_SELLER_INDUSTRY           0.000000
CNT_PAYMENT                   22.221205
NAME_YIELD_GROUP               0.000000
```

```
PRODUCT_COMBINATION            0.021362
DAYS_FIRST_DRAWING            40.121880
DAYS_FIRST_DUE                40.121880
DAYS_LAST_DUE_1ST_VERSION     40.121880
DAYS_LAST_DUE                 40.121880
DAYS_TERMINATION              40.121880
NFLAG_INSURED_ON_APPROVAL     40.121880
dtype: float64
```

[61]:
```
df1=df1.drop(df1.loc[:,(100*df1.isnull().mean())>35],axis=1)        ### DROP␣
↪THE COLUMNS WITH MISSING VALUES OF MORE THAN 35%.
```

[62]:
```
df1.shape
```

[62]:
```
(307511, 19)
```

[63]:
```
100*df1.isnull().mean()
```

[63]:
```
SK_ID_CURR             0.000000
TARGET                 0.000000
NAME_CONTRACT_TYPE     0.000000
CODE_GENDER            0.000000
FLAG_OWN_CAR           0.000000
FLAG_OWN_REALTY        0.000000
CNT_CHILDREN           0.000000
AMT_INCOME_TOTAL       0.000000
AMT_CREDIT             0.000000
AMT_ANNUITY            0.003902
NAME_INCOME_TYPE       0.000000
NAME_EDUCATION_TYPE    0.000000
NAME_FAMILY_STATUS     0.000000
NAME_HOUSING_TYPE      0.000000
DAYS_BIRTH             0.000000
DAYS_EMPLOYED          0.000000
OCCUPATION_TYPE       31.345545
CNT_FAM_MEMBERS        0.000000
ORGANIZATION_TYPE      0.000000
dtype: float64
```

[64]:
```
df1.info()    #### CHECKING THE DATA TYPE
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 19 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   SK_ID_CURR           307511 non-null  int64
```

```
 1   TARGET               307511 non-null  int64
 2   NAME_CONTRACT_TYPE   307511 non-null  object
 3   CODE_GENDER          307511 non-null  object
 4   FLAG_OWN_CAR         307511 non-null  object
 5   FLAG_OWN_REALTY      307511 non-null  object
 6   CNT_CHILDREN         307511 non-null  int64
 7   AMT_INCOME_TOTAL     307511 non-null  float64
 8   AMT_CREDIT           307511 non-null  float64
 9   AMT_ANNUITY          307499 non-null  float64
 10  NAME_INCOME_TYPE     307511 non-null  object
 11  NAME_EDUCATION_TYPE  307511 non-null  object
 12  NAME_FAMILY_STATUS   307511 non-null  object
 13  NAME_HOUSING_TYPE    307511 non-null  object
 14  DAYS_BIRTH           307511 non-null  int64
 15  DAYS_EMPLOYED        307511 non-null  int64
 16  OCCUPATION_TYPE      211120 non-null  object
 17  CNT_FAM_MEMBERS      307509 non-null  float64
 18  ORGANIZATION_TYPE    307511 non-null  object
dtypes: float64(4), int64(5), object(10)
memory usage: 44.6+ MB
```

REPLACING ROWS CORRESPONDING TO NULL VALUES OF "OCCUPATION_TYPE" COLUMN WITH ITS MODE VALUE INSTEAD OF DROPPING AS IT IS AN IMPACTFUL COLUMN FOR THIS ANALYSIS.

[65]:
```
df1["OCCUPATION_TYPE"].mode()
```

[65]:
```
0    Laborers
Name: OCCUPATION_TYPE, dtype: object
```

[66]:
```
df1["OCCUPATION_TYPE"]=df1["OCCUPATION_TYPE"].fillna('Laborers')
```

[67]:
```
100*df1.isnull().mean()
```

[67]:
```
SK_ID_CURR             0.000000
TARGET                 0.000000
NAME_CONTRACT_TYPE     0.000000
CODE_GENDER            0.000000
FLAG_OWN_CAR           0.000000
FLAG_OWN_REALTY        0.000000
CNT_CHILDREN           0.000000
AMT_INCOME_TOTAL       0.000000
AMT_CREDIT             0.000000
AMT_ANNUITY            0.003902
NAME_INCOME_TYPE       0.000000
NAME_EDUCATION_TYPE    0.000000
NAME_FAMILY_STATUS     0.000000
NAME_HOUSING_TYPE      0.000000
```

```
DAYS_BIRTH            0.000000
DAYS_EMPLOYED         0.000000
OCCUPATION_TYPE       0.000000
CNT_FAM_MEMBERS       0.000650
ORGANIZATION_TYPE     0.000000
dtype: float64
```

[68]:
```
df1=df1.dropna()  ## Dropping residual columns with NA values.
```

[69]:
```
df1.head()
```

[69]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

  NAME_INCOME_TYPE            NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  \
0          Working  Secondary / secondary special  Single / not married
1    State servant               Higher education               Married
2          Working  Secondary / secondary special  Single / not married
3          Working  Secondary / secondary special        Civil marriage
4          Working  Secondary / secondary special  Single / not married

   NAME_HOUSING_TYPE  DAYS_BIRTH  DAYS_EMPLOYED OCCUPATION_TYPE  \
0  House / apartment       -9461           -637        Laborers
1  House / apartment      -16765          -1188      Core staff
2  House / apartment      -19046           -225        Laborers
3  House / apartment      -19005          -3039        Laborers
4  House / apartment      -19932          -3038      Core staff

   CNT_FAM_MEMBERS       ORGANIZATION_TYPE
0              1.0  Business Entity Type 3
1              2.0                  School
2              1.0              Government
3              2.0  Business Entity Type 3
4              1.0                Religion
```

MERGING TWO DATA FRAMES TO ANALYSE THE ATTRIBUTES OF DATA FRAME TWO W.R.T. THE TARGET COLUMN OF DATA FRAME ONE CORRESPONDING TO THE COMMON VALUE OF COLUMN 'SK_ID_CURR'

[65]:
```
#df=pd.
↪merge(df1,df2[['AMT_CREDIT','AMT_DOWN_PAYMENT','RATE_INTEREST_PRIMARY','NAME_CASH_LOAN_PURP
```

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
Input In [65], in <cell line: 1>()
----> 1␣
↪df=pd.merge(df1,df2[['AMT_CREDIT','AMT_DOWN_PAYMENT','RATE_INTEREST_PRIMARY','NAME_CASH_LOA

File ~\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:122, in␣
↪merge(left, right, how, on, left_on, right_on, left_index, right_index, sort,␣
↪suffixes, copy, indicator, validate)
     90 @Substitution("\nleft : DataFrame or named Series")
     91 @Appender(_merge_doc, indents=0)
     92 def merge(
   (...)
    105     validate: str | None = None,
    106 ) -> DataFrame:
    107     op = _MergeOperation(
    108         left,
    109         right,
   (...)
    120         validate=validate,
    121     )
--> 122     return op.get_result()

File ~\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:716, in␣
↪_MergeOperation.get_result(self)
    713 if self.indicator:
    714     self.left, self.right = self._indicator_pre_merge(self.left, self.
↪right)
--> 716 join_index, left_indexer, right_indexer = self._get_join_info()
    718 llabels, rlabels = _items_overlap_with_suffix(
    719     self.left._info_axis, self.right._info_axis, self.suffixes
    720 )
    722 lindexers = {1: left_indexer} if left_indexer is not None else {}

File ~\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:967, in␣
↪_MergeOperation._get_join_info(self)
    963     join_index, right_indexer, left_indexer = _left_join_on_index(
    964         right_ax, left_ax, self.right_join_keys, sort=self.sort
    965     )
    966 else:
--> 967     (left_indexer, right_indexer) = self._get_join_indexers()
```

```
    969     if self.right_index:
    970         if len(self.left) > 0:

File ~\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:941, in
↪_MergeOperation._get_join_indexers(self)
    939 def _get_join_indexers(self) -> tuple[npt.NDArray[np.intp], npt.
↪NDArray[np.intp]]:
    940     """return the join indexers"""
--> 941     return get_join_indexers(
    942
↪        self.left_join_keys, self.right_join_keys, sort=self.sort, how=self.how
    943     )

File ~\anaconda3\lib\site-packages\pandas\core\reshape\merge.py:1509, in
↪get_join_indexers(left_keys, right_keys, sort, how, **kwargs)
   1499 join_func = {
   1500     "inner": libjoin.inner_join,
   1501     "left": libjoin.left_outer_join,
   (…)
   1505     "outer": libjoin.full_outer_join,
   1506 }[how]
   1508 # error: Cannot call function of unknown type
-> 1509 return join_func(lkey, rkey, count, **kwargs)

File ~\anaconda3\lib\site-packages\pandas\_libs\join.pyx:102, in pandas._libs.
↪join.left_outer_join()

MemoryError: Unable to allocate 7.05 GiB for an array with shape (946779431,)
↪and data type int64
```

SORTING 2ND DATA FRAME .

AS MERGING IS NOT POSSIBLE DUE TO MEMORY SHORTAGE MODIFING 2ND DATA FRAME WITH JUST NECESSARY COLUMNS AND FEWER ROWS. USING 'VLOOKUP' FUNCTION 'TARGET' COLUMNN HAS BEEN INDUCED IN 2ND DATA FRAME .

[70]:
```
df2=df2[['SK_ID_CURR','TARGET','AMT_CREDIT','AMT_DOWN_PAYMENT','RATE_INTEREST_PRIMARY','NAME_C
↪head(102588)
```

[71]:
```
df2.head()
```

[71]:

	SK_ID_CURR	TARGET	AMT_CREDIT	AMT_DOWN_PAYMENT	RATE_INTEREST_PRIMARY	\
0	271877	0.0	17145.0	0.0	0.182832	
1	108129	0.0	679671.0	NaN	NaN	
2	122040	0.0	136444.5	NaN	NaN	
3	176158	0.0	470790.0	NaN	NaN	

```
4      202054     0.0    404055.0               NaN                    NaN

  NAME_CASH_LOAN_PURPOSE NAME_CONTRACT_STATUS CODE_REJECT_REASON  \
0                    XAP             Approved                XAP
1                    XNA             Approved                XAP
2                    XNA             Approved                XAP
3                    XNA             Approved                XAP
4                Repairs              Refused                 HC

  NAME_CLIENT_TYPE
0        Repeater
1        Repeater
2        Repeater
3        Repeater
4        Repeater
```

[72]:
```
100*df2.isnull().mean()  ### CHECKING MISSING VALUES.
```

[72]:
```
SK_ID_CURR                 0.000000
TARGET                    15.258120
AMT_CREDIT                 0.000000
AMT_DOWN_PAYMENT          50.920186
RATE_INTEREST_PRIMARY     99.660779
NAME_CASH_LOAN_PURPOSE     0.000000
NAME_CONTRACT_STATUS       0.000000
CODE_REJECT_REASON         0.000000
NAME_CLIENT_TYPE           0.000000
dtype: float64
```

[73]:
```
df2=df2.drop('RATE_INTEREST_PRIMARY',axis=1) ### DROPPING THE COLUMNS WITH␣
↪MISSING VALUE MORE THAN 40%.
```

[74]:
```
100*df2.isnull().mean()
```

[74]:
```
SK_ID_CURR                 0.000000
TARGET                    15.258120
AMT_CREDIT                 0.000000
AMT_DOWN_PAYMENT          50.920186
NAME_CASH_LOAN_PURPOSE     0.000000
NAME_CONTRACT_STATUS       0.000000
CODE_REJECT_REASON         0.000000
NAME_CLIENT_TYPE           0.000000
dtype: float64
```

[75]:
```
df2["AMT_DOWN_PAYMENT"]=df2["AMT_DOWN_PAYMENT"].fillna(df2["AMT_DOWN_PAYMENT"].
↪median())  ### AS THIS COLUMN MAY HAVE IMPACT IN CASE OF LOAN DEFAULT ,␣
↪REPLACING THE 'NA' VALUES WITH ITS MEDIAN VALUE INSTEAD OF DROPPING .
```

[76]:
```
100*df2.isnull().mean()
```

[76]:
```
SK_ID_CURR               0.00000
TARGET                  15.25812
AMT_CREDIT               0.00000
AMT_DOWN_PAYMENT         0.00000
NAME_CASH_LOAN_PURPOSE   0.00000
NAME_CONTRACT_STATUS     0.00000
CODE_REJECT_REASON       0.00000
NAME_CLIENT_TYPE         0.00000
dtype: float64
```

[77]:
```
df2=df2.dropna(subset=df2.columns.values)  ### DROPPING THE ROWS CORRESPONDING␣
↪TO THE NA VALUES OF 'TARGET' COLUMN.
```

[78]:
```
100*df2.isnull().mean()
```

[78]:
```
SK_ID_CURR               0.0
TARGET                   0.0
AMT_CREDIT               0.0
AMT_DOWN_PAYMENT         0.0
NAME_CASH_LOAN_PURPOSE   0.0
NAME_CONTRACT_STATUS     0.0
CODE_REJECT_REASON       0.0
NAME_CLIENT_TYPE         0.0
dtype: float64
```

0.0.4 MANAGGING COLUMNS WITH IMPROPER DATATYPES

[79]:
```
df1['CNT_FAM_MEMBERS']=df1['CNT_FAM_MEMBERS'].astype('int64')
```

[80]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 307497 entries, 0 to 307510
Data columns (total 19 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   SK_ID_CURR          307497 non-null  int64
 1   TARGET              307497 non-null  int64
 2   NAME_CONTRACT_TYPE  307497 non-null  object
 3   CODE_GENDER         307497 non-null  object
 4   FLAG_OWN_CAR        307497 non-null  object
 5   FLAG_OWN_REALTY     307497 non-null  object
 6   CNT_CHILDREN        307497 non-null  int64
 7   AMT_INCOME_TOTAL    307497 non-null  float64
 8   AMT_CREDIT          307497 non-null  float64
```

```
 9   AMT_ANNUITY          307497 non-null  float64
 10  NAME_INCOME_TYPE     307497 non-null  object
 11  NAME_EDUCATION_TYPE  307497 non-null  object
 12  NAME_FAMILY_STATUS   307497 non-null  object
 13  NAME_HOUSING_TYPE    307497 non-null  object
 14  DAYS_BIRTH           307497 non-null  int64
 15  DAYS_EMPLOYED        307497 non-null  int64
 16  OCCUPATION_TYPE      307497 non-null  object
 17  CNT_FAM_MEMBERS      307497 non-null  int64
 18  ORGANIZATION_TYPE    307497 non-null  object
dtypes: float64(3), int64(6), object(10)
memory usage: 46.9+ MB
```

[81]:
```
df1["DAYS_BIRTH"]=df1["DAYS_BIRTH"].astype('str')
df1["DAYS_BIRTH"].head()
```

[81]:
```
0     -9461
1    -16765
2    -19046
3    -19005
4    -19932
Name: DAYS_BIRTH, dtype: object
```

[82]:
```
df1['DAYS_EMPLOYED']=df1["DAYS_EMPLOYED"].astype('str')
df1['DAYS_EMPLOYED'].head()
```

[82]:
```
0     -637
1    -1188
2     -225
3    -3039
4    -3038
Name: DAYS_EMPLOYED, dtype: object
```

STANDARDISING THE VALUES & FIXING INVALID VALUES.

[83]:
```
df1["DAYS_BIRTH"]=df1["DAYS_BIRTH"].apply(lambda x:int(x[1:]) if x[0]=='-' else␣
↪int(x[0:]))  ## REMOVING IMPROPER ENTRIES (PREFIX).
```

[84]:
```
df1["DAYS_BIRTH"].head()
```

[84]:
```
0     9461
1    16765
2    19046
3    19005
4    19932
Name: DAYS_BIRTH, dtype: int64
```

[85]:
```
df1["DAYS_BIRTH"]=df1["DAYS_BIRTH"].apply(lambda x:int(x/365))   ### CONVERTING␣
↪DAY TO YEAR
```

[86]:
```
df1.rename(columns={"DAYS_BIRTH":"AGE_YEAR"},inplace=True)       ## RENAME THE␣
↪ROW
```

[87]:
```
df1.head()
```

[87]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

  NAME_INCOME_TYPE            NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  \
0          Working  Secondary / secondary special  Single / not married
1    State servant               Higher education               Married
2          Working  Secondary / secondary special  Single / not married
3          Working  Secondary / secondary special        Civil marriage
4          Working  Secondary / secondary special  Single / not married

   NAME_HOUSING_TYPE  AGE_YEAR DAYS_EMPLOYED OCCUPATION_TYPE  CNT_FAM_MEMBERS  \
0  House / apartment        25          -637        Laborers                1
1  House / apartment        45         -1188      Core staff                2
2  House / apartment        52          -225        Laborers                1
3  House / apartment        52         -3039        Laborers                2
4  House / apartment        54         -3038      Core staff                1

        ORGANIZATION_TYPE
0  Business Entity Type 3
1                  School
2              Government
3  Business Entity Type 3
4                Religion
```

[88]:
```
df1['DAYS_EMPLOYED']=df1['DAYS_EMPLOYED'].apply(lambda x:int(x[1:]) if␣
↪x[0]=='-' else int(x[0:]))    ## REMOVING IMPROPER ENTRIES.
```

[89]:
```
df1['DAYS_EMPLOYED'].head()
```

[89]:
```
0     637
1    1188
2     225
3    3039
4    3038
Name: DAYS_EMPLOYED, dtype: int64
```

0.0.5 CHECKING DATA IMBALANCE

0.0.6 CHECKING OUTLIERS

[90]:
```
df1.describe()
```

[90]:
```
          SK_ID_CURR         TARGET   CNT_CHILDREN  AMT_INCOME_TOTAL  \
count  307497.000000  307497.000000  307497.000000      3.074970e+05
mean   278182.229433       0.080732       0.417071      1.687962e+05
std    102790.409563       0.272424       0.722132      2.371276e+05
min    100002.000000       0.000000       0.000000      2.565000e+04
25%    189150.000000       0.000000       0.000000      1.125000e+05
50%    278204.000000       0.000000       0.000000      1.468125e+05
75%    367144.000000       0.000000       1.000000      2.025000e+05
max    456255.000000       1.000000      19.000000      1.170000e+08

         AMT_CREDIT    AMT_ANNUITY       AGE_YEAR  DAYS_EMPLOYED  \
count  3.074970e+05  307497.000000  307497.000000  307497.000000
mean   5.990271e+05   27108.545347      43.436186   67727.733314
std    4.024939e+05   14493.778987      11.954639  139446.221239
min    4.500000e+04    1615.500000      20.000000       0.000000
25%    2.700000e+05   16524.000000      34.000000     933.000000
50%    5.135310e+05   24903.000000      43.000000    2219.000000
75%    8.086500e+05   34596.000000      53.000000    5707.000000
max    4.050000e+06  258025.500000      69.000000  365243.000000

       CNT_FAM_MEMBERS
count    307497.000000
mean          2.152681
std           0.910692
min           1.000000
25%           2.000000
50%           2.000000
75%           3.000000
max          20.000000
```

INCOME,CREDIT,ANNUITY,DAYS EMPLOYED AND FAMILY MEMBERS COLUMN HAVE OUTLIERS AS DIFFERENCE BETWEEN MEAN AND MEDIAN OR 75TH PERCENTILE AND MAX VALUE IS HIGHER.

[91]:
```
df2.describe()
```

[91]:

	SK_ID_CURR	TARGET	AMT_CREDIT	AMT_DOWN_PAYMENT
count	86935.000000	86935.000000	8.693500e+04	86935.000000
mean	278770.024754	0.085443	1.887524e+05	4116.914663
std	102904.637171	0.279542	3.104122e+05	13016.828275
min	100006.000000	0.000000	0.000000e+00	0.000000
25%	189421.500000	0.000000	2.609100e+04	1640.250000
50%	279190.000000	0.000000	7.912800e+04	1640.250000
75%	368097.000000	0.000000	1.978200e+05	1710.000000
max	456254.000000	1.000000	4.104351e+06	945000.000000

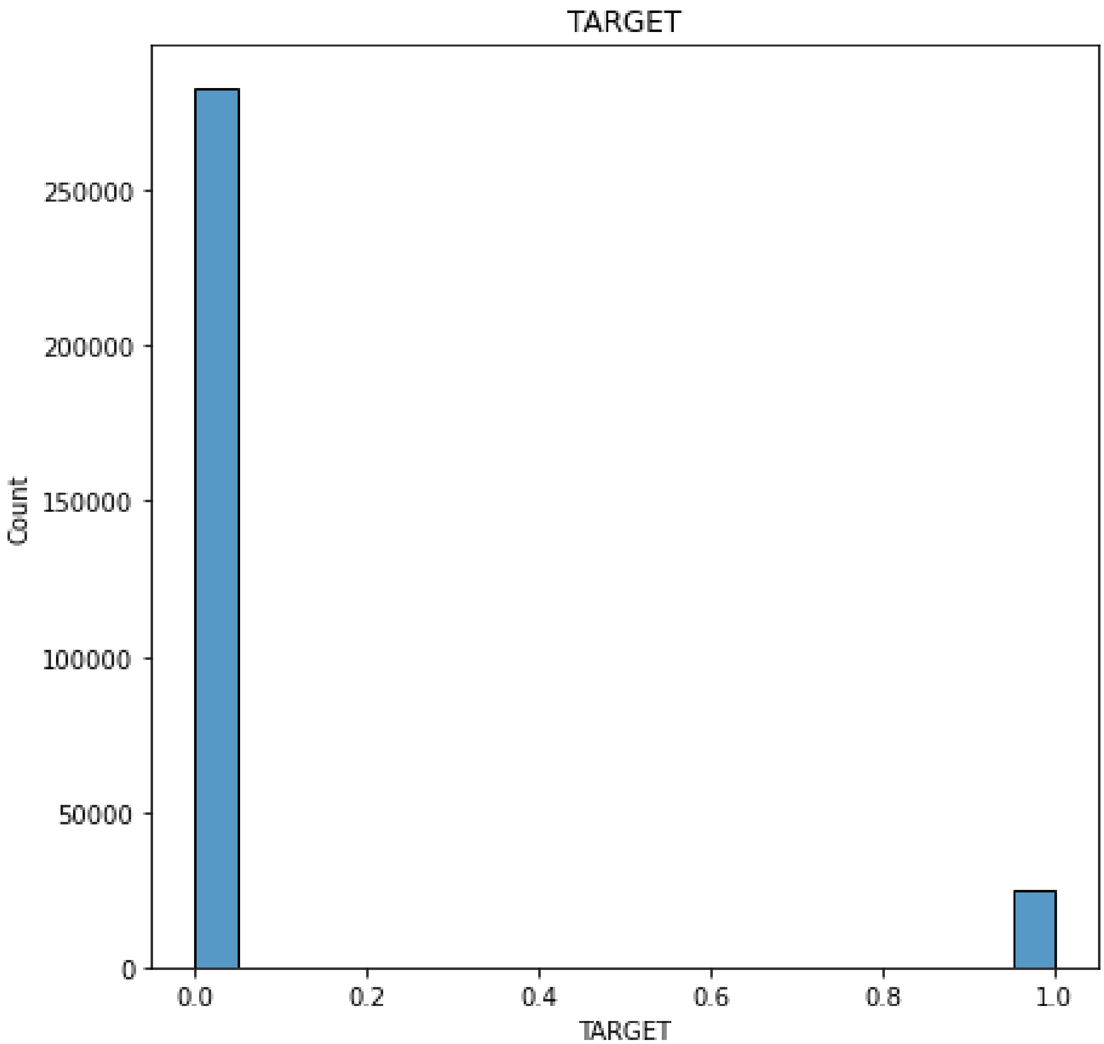
CREDIT AND DOWN PAYMENT COLUMN HAVE OUTLIERS AS DIFFERENCE BETWEEN MEAN AND MEDIAN OR 75TH PERCENTILE AND MAX VALUE IS HIGHER.

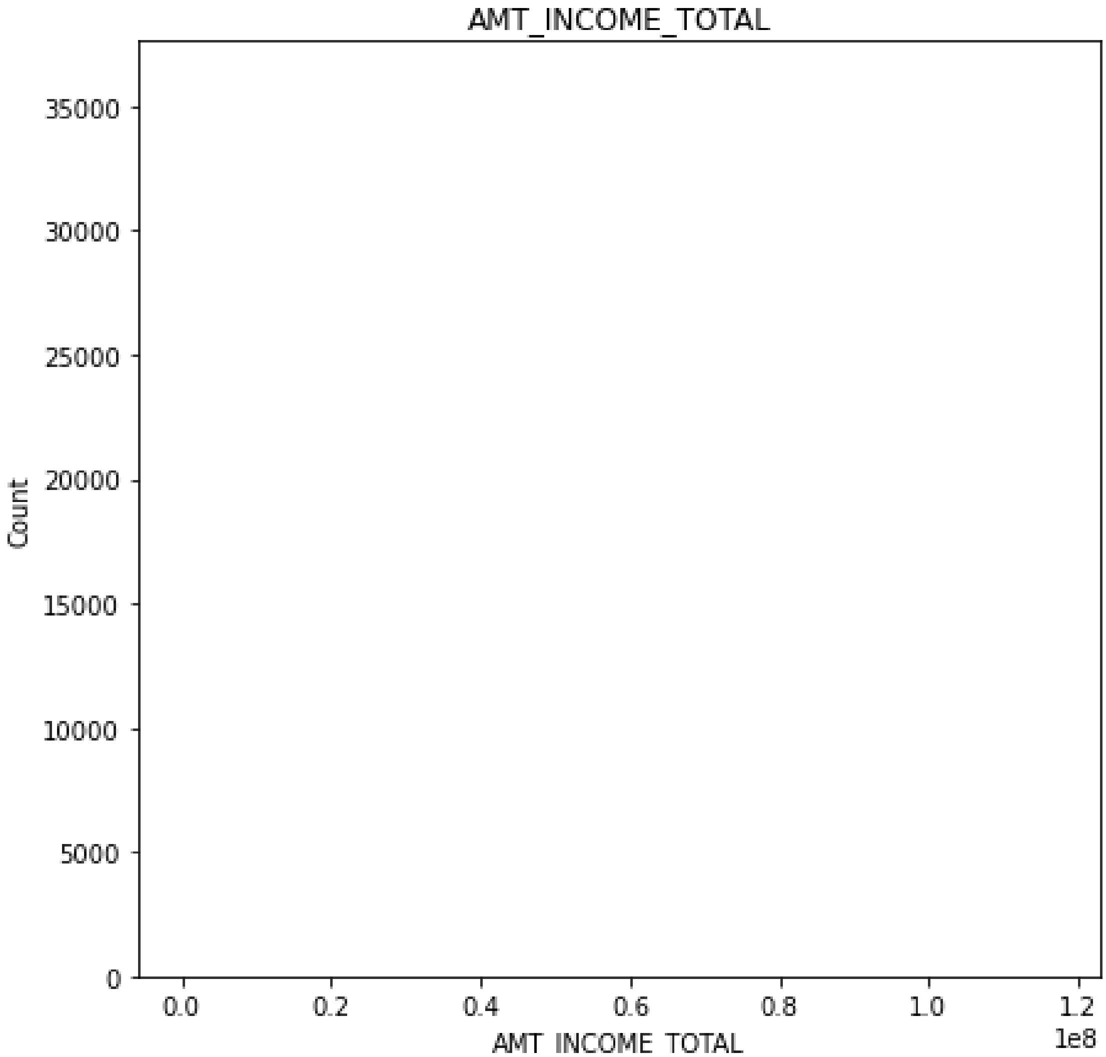
0.0.7 VISUALIZATION

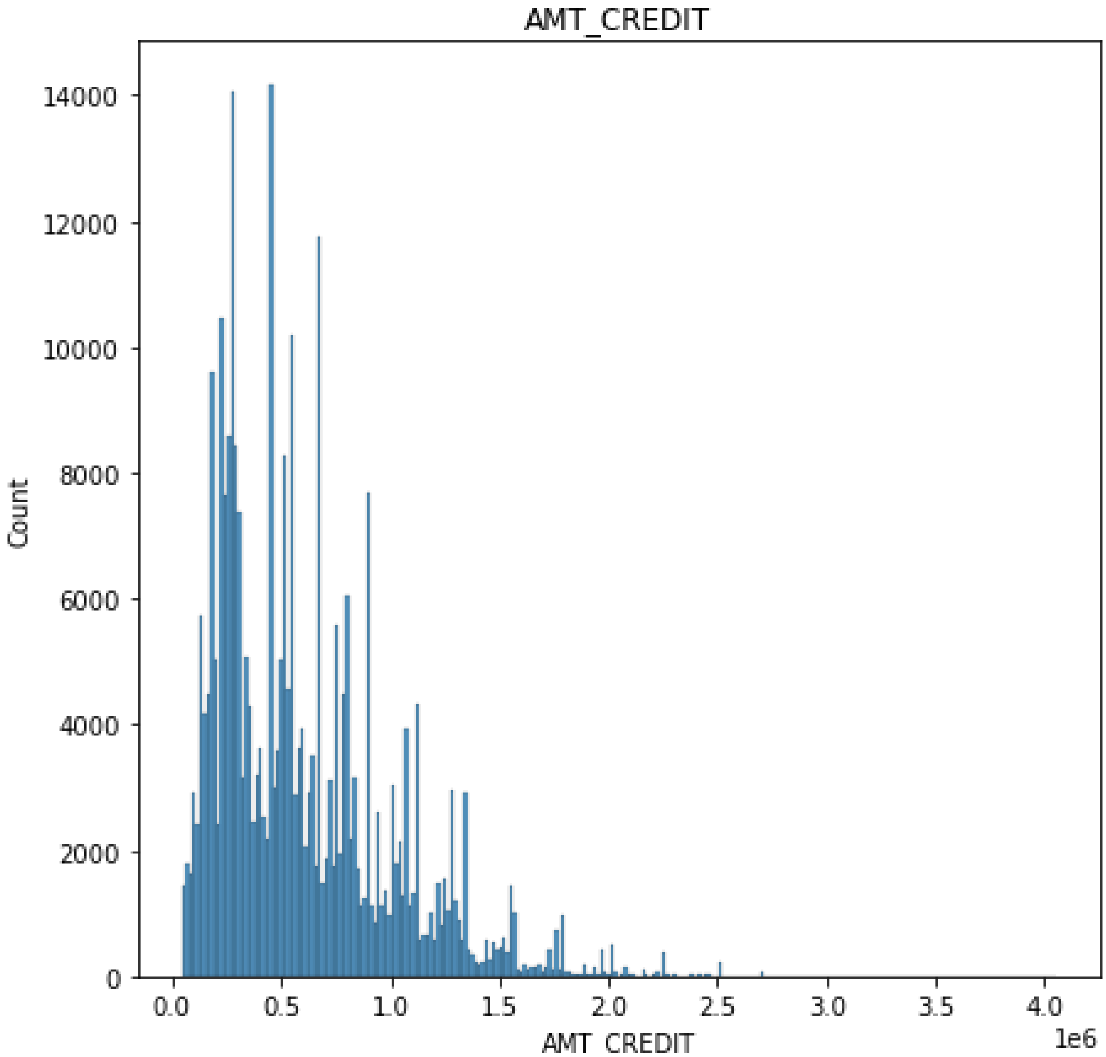
[92]:
```
categorical1=['TARGET','NAME_CONTRACT_TYPE','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY','NA
continuous1=['TARGET','AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AGE_YEAR','DAYS_EMPLOYED'
```

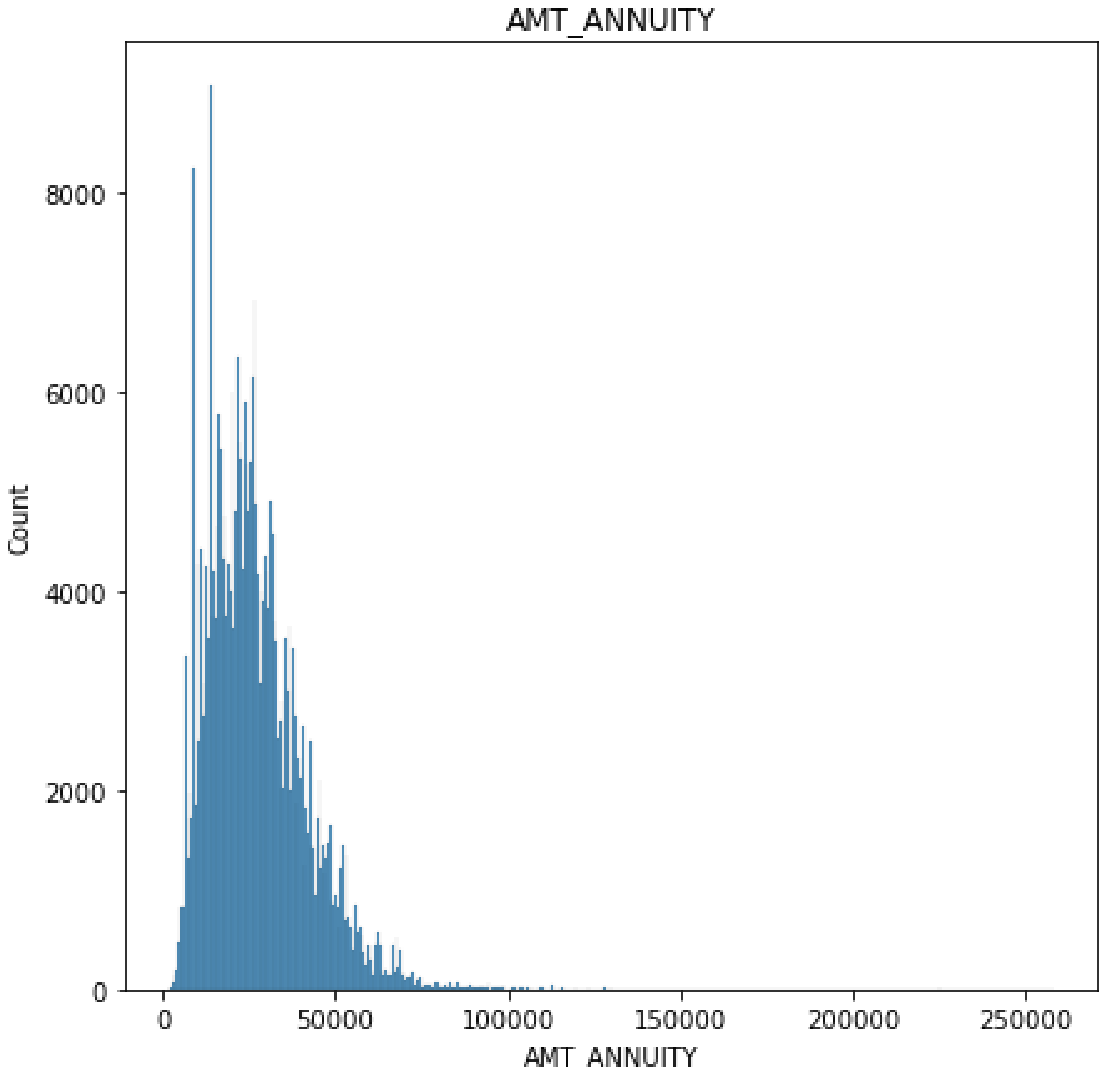
UNIVARIATE ANALYSIS OF CONTUNUOUS VARIABLES

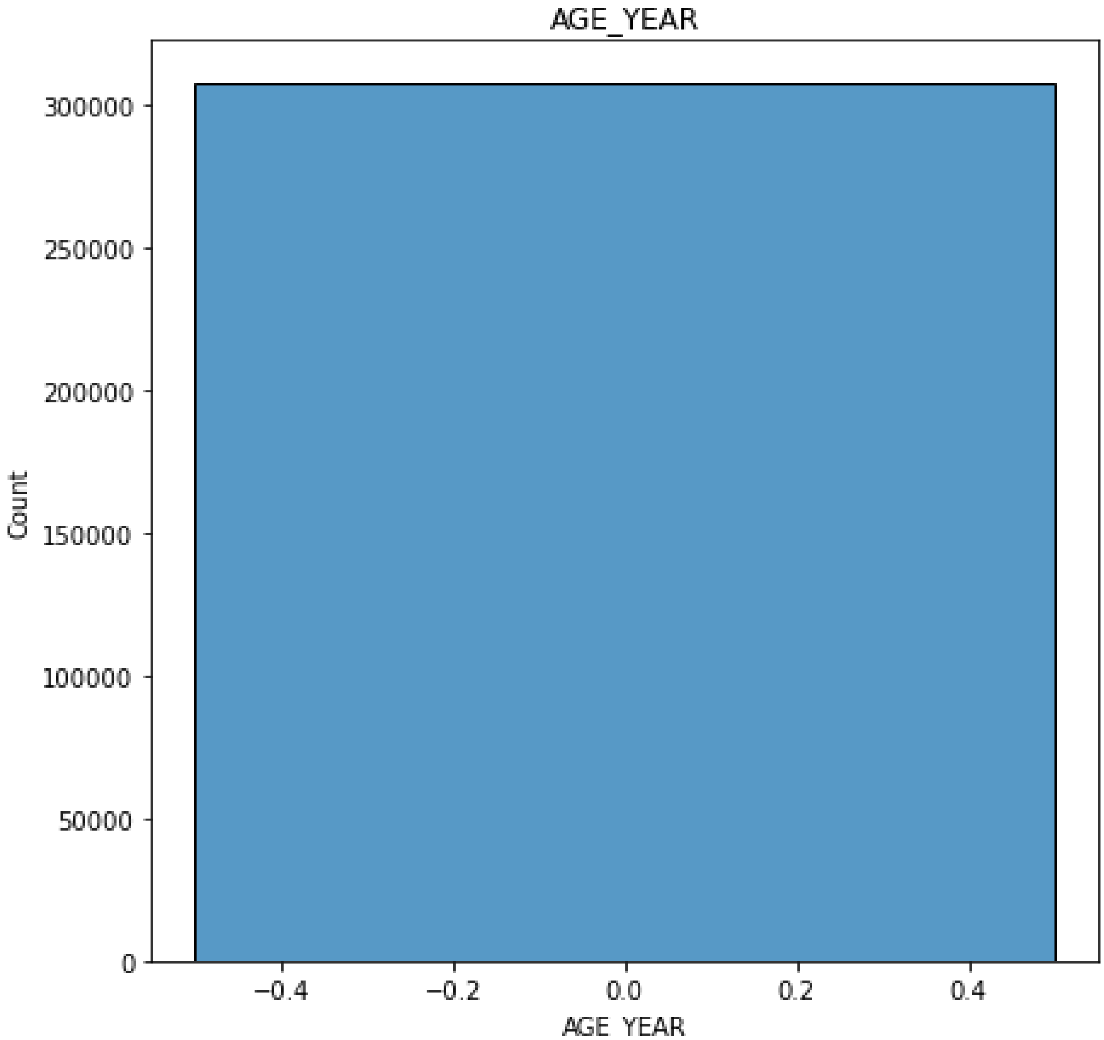
[50]:
```
for col in continuous1:
    plt.figure(figsize=[7,7])

    sns.histplot(df1[col])
    plt.title(col)
    plt.show()
```

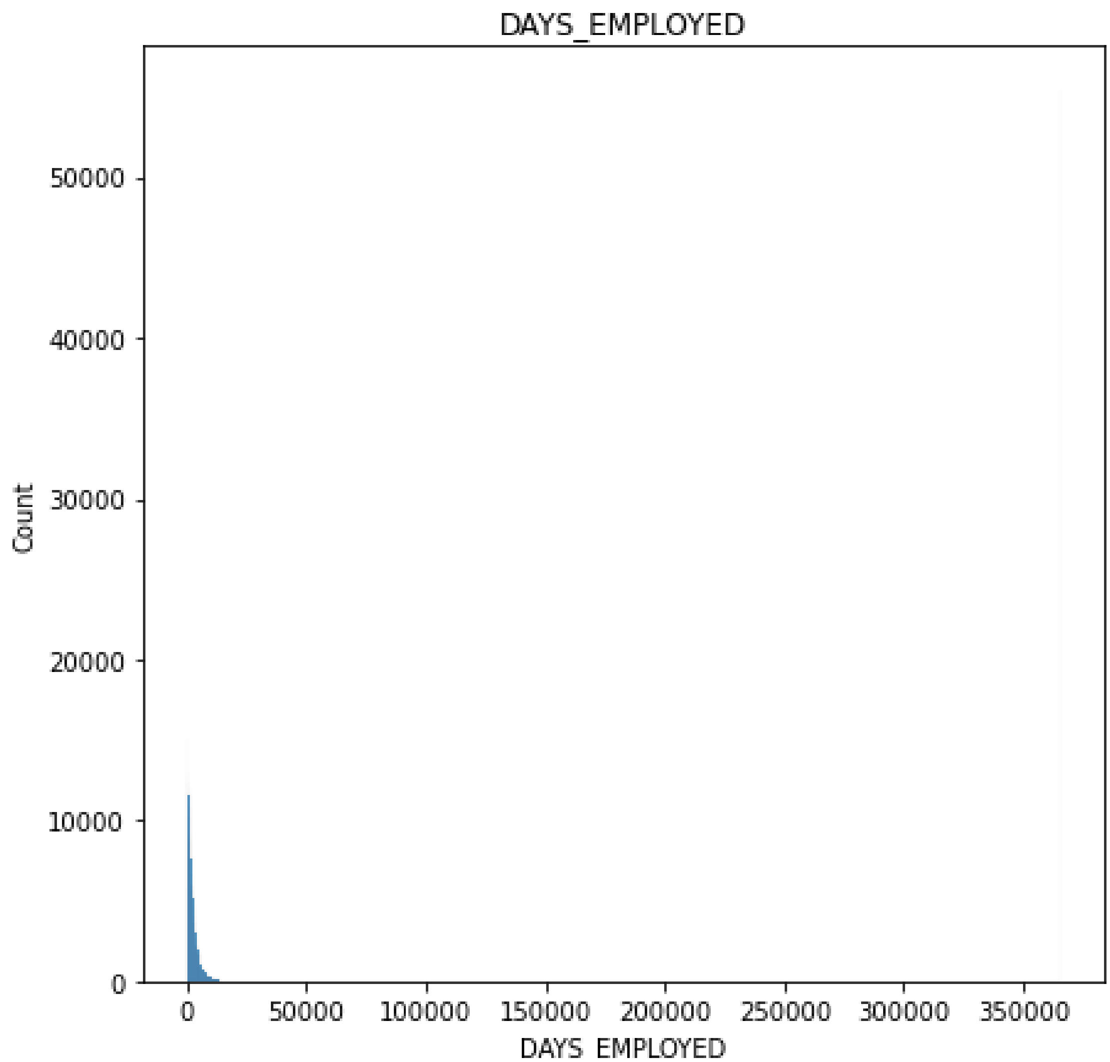

TARGET
250000
200000
150000
100000
50000
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
TARGET


AMT_INCOME_TOTAL
35000
30000
25000
20000
15000
10000
5000
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e8
AMT_INCOME_TOTAL

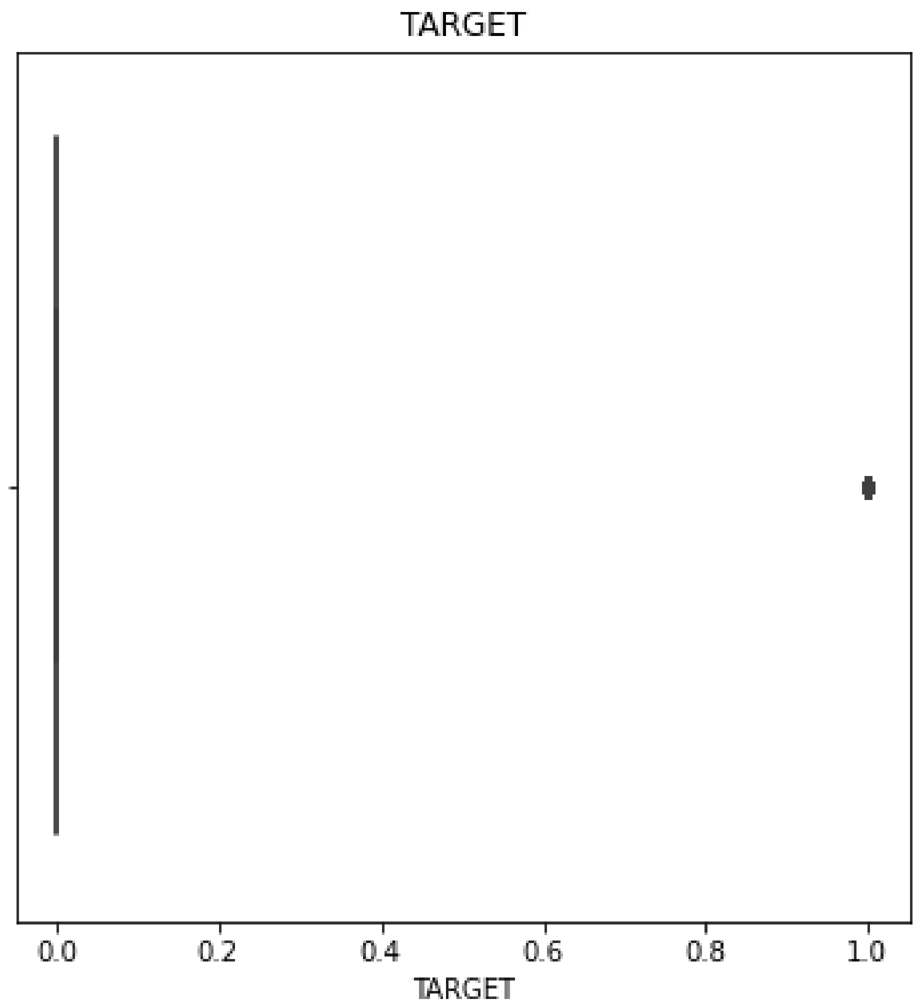

AMT_CREDIT
Count
14000
12000
10000
8000
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
AMT_CREDIT

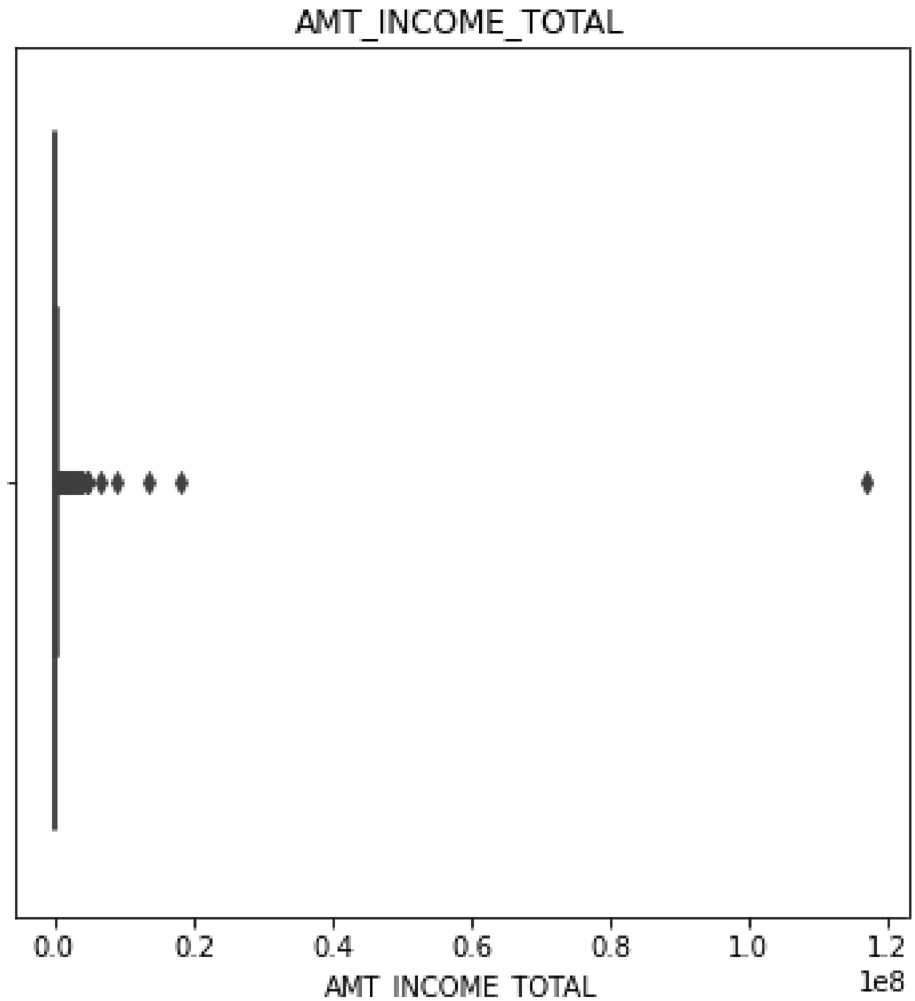

AMT_ANNUITY
Count
0
2000
4000
6000
8000
0
50000
100000
150000
200000
250000
AMT_ANNUITY

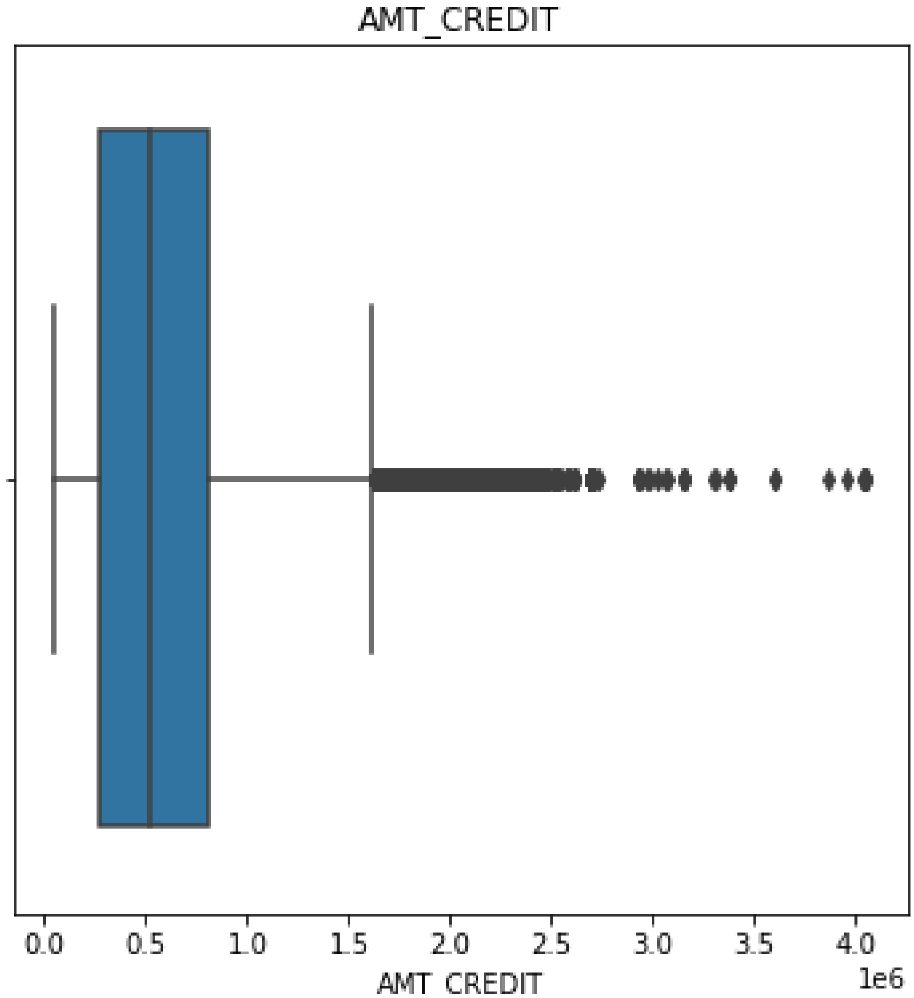

AGE_YEAR
300000
250000
200000
150000
100000
50000
0
Count
−0.4
−0.2
0.0
0.2
0.4
AGE_YEAR

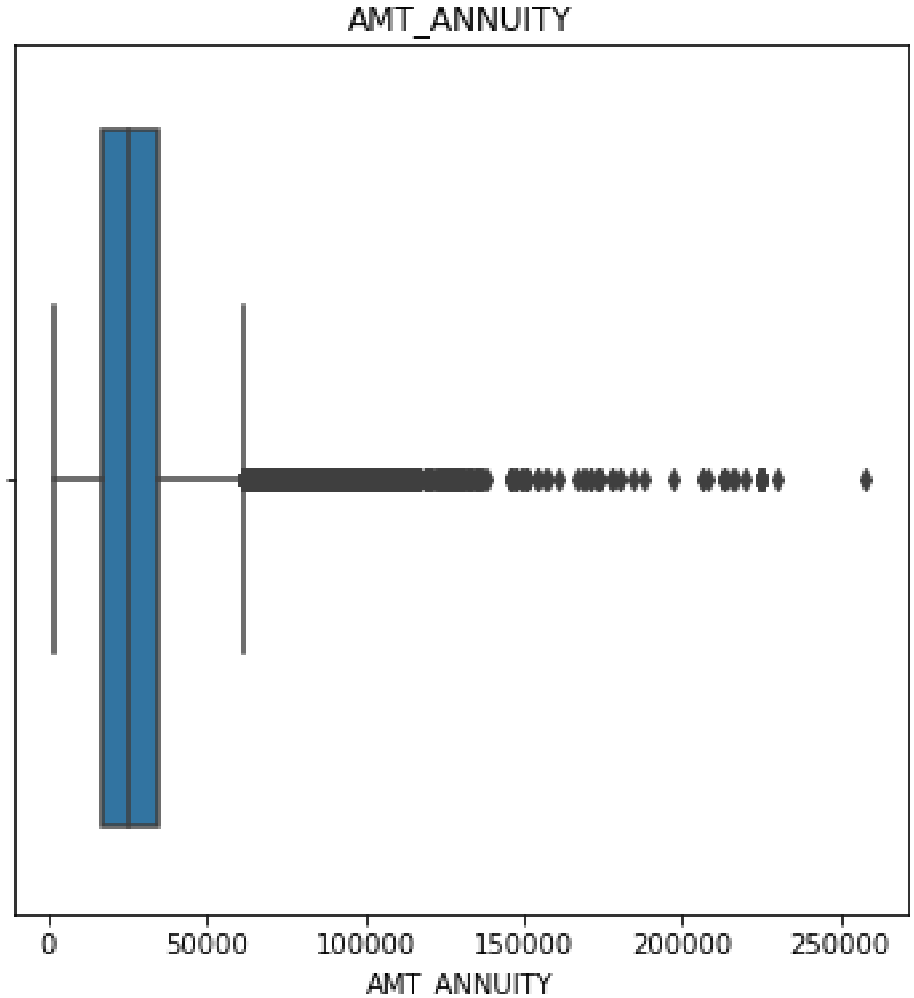


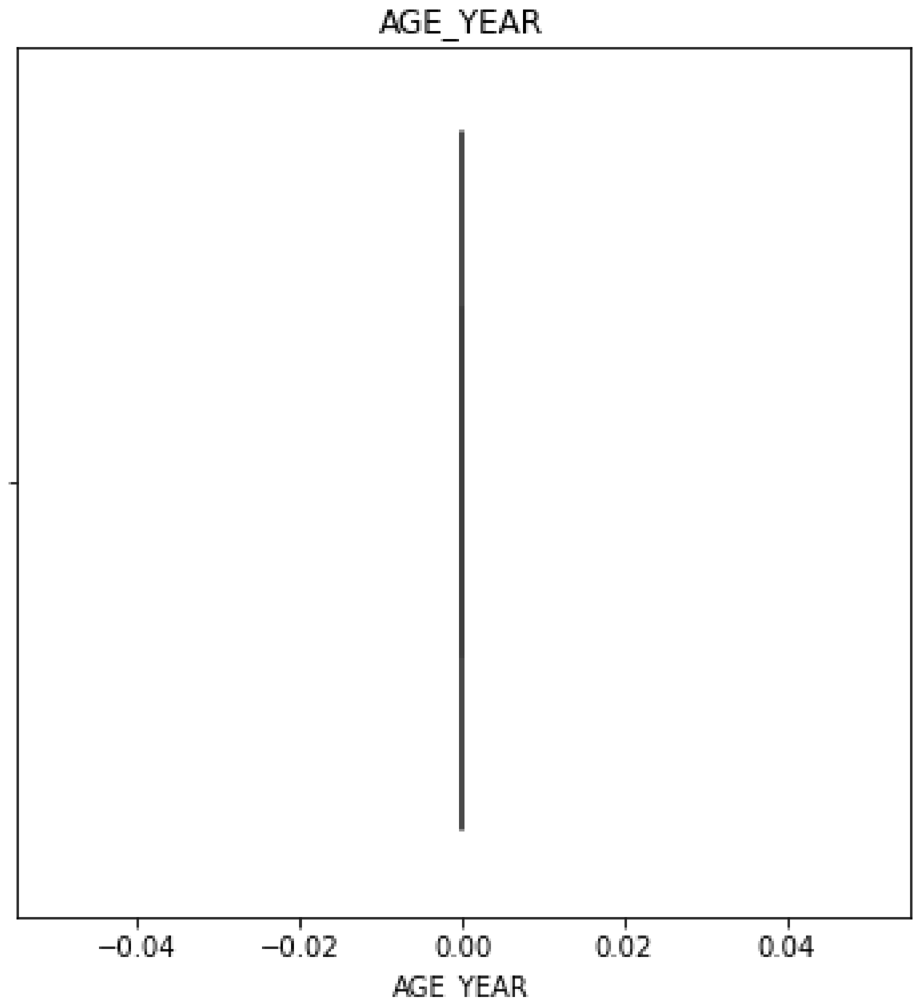
[51]:
```
for col in continuous1:
    plt.figure(figsize=[6,6])

    sns.boxplot(df1[col])
    plt.title(col)
    plt.show()
```

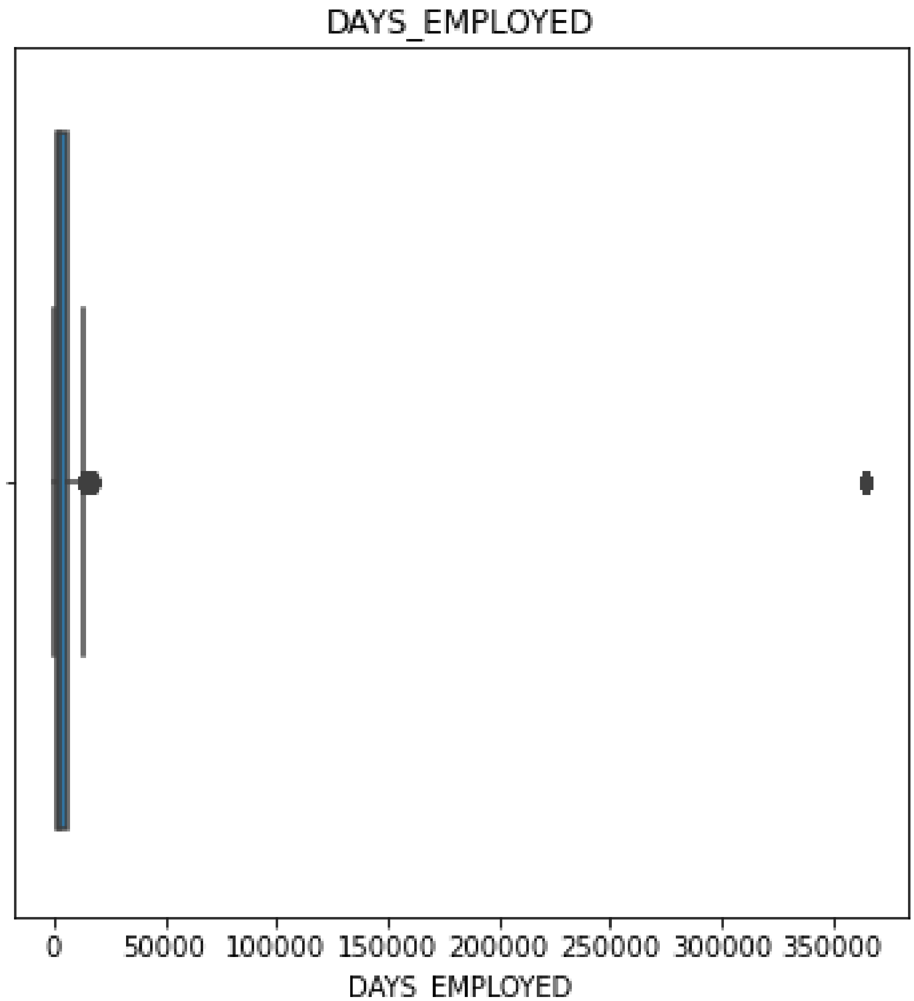

TARGET
0.0
0.2
0.4
0.6
0.8
1.0
TARGET


AMT_INCOME_TOTAL
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e8
AMT_INCOME_TOTAL


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


AMT_ANNUITY
0
50000
100000
150000
200000
250000
AMT_ANNUITY

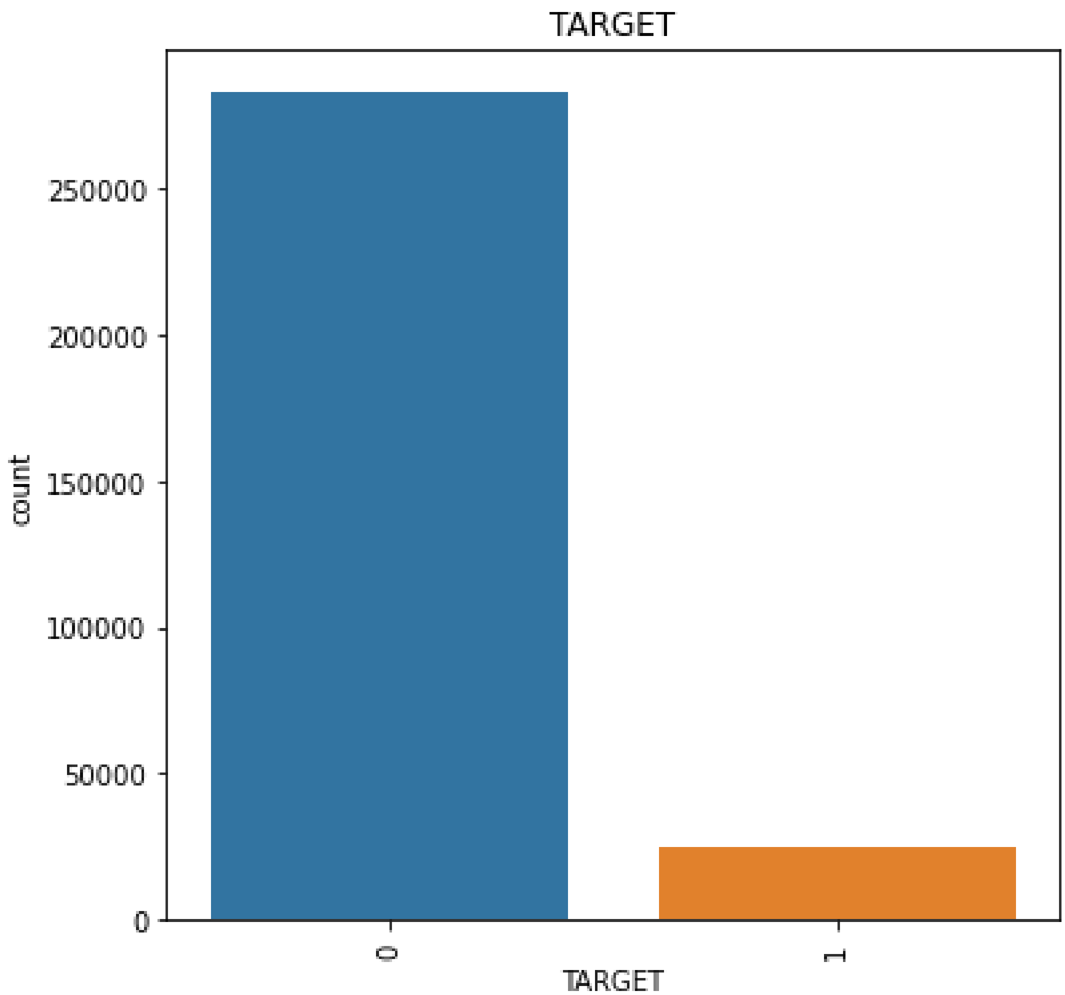

AGE_YEAR
−0.04
−0.02
0.00
0.02
0.04
AGE_YEAR

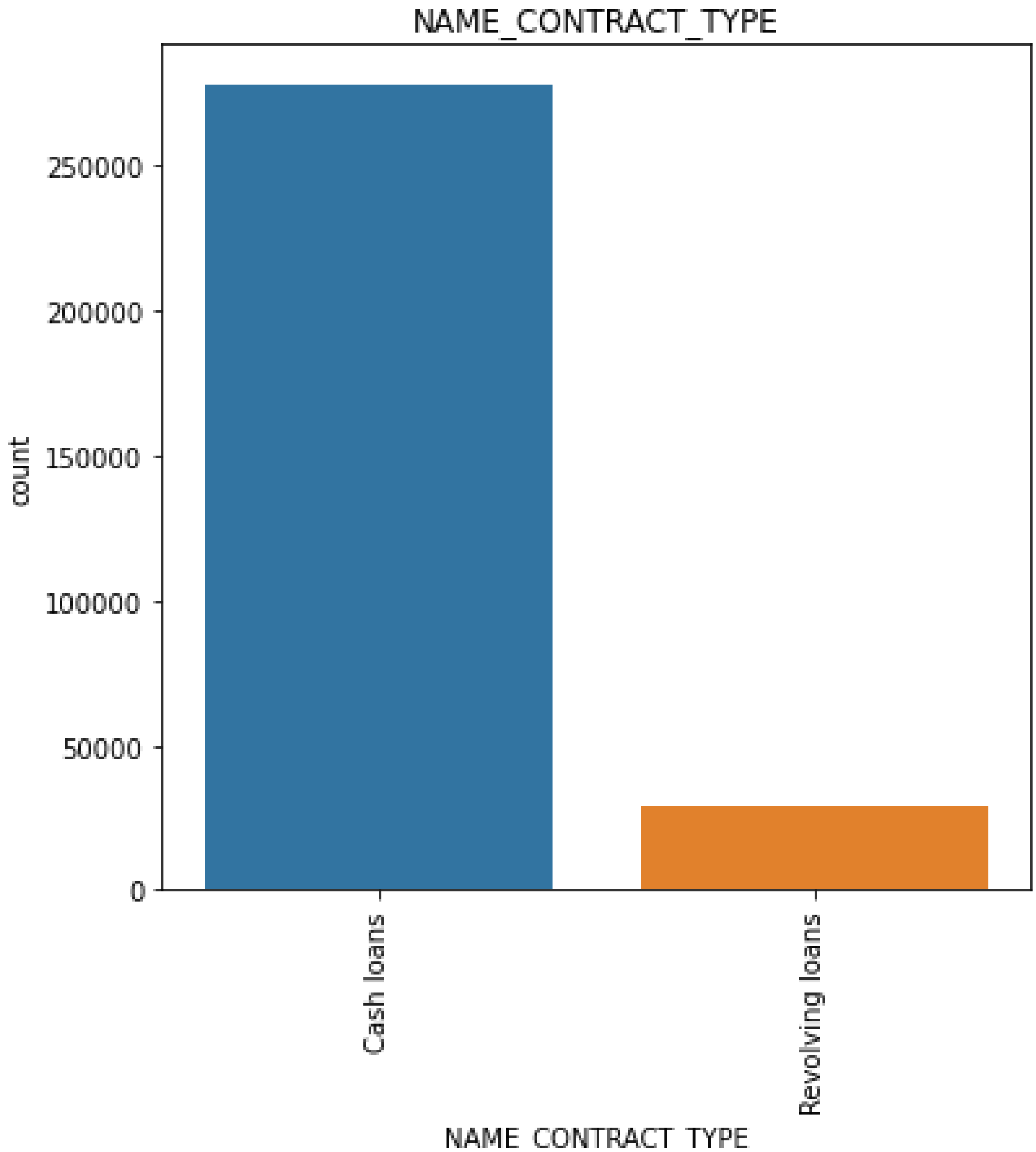


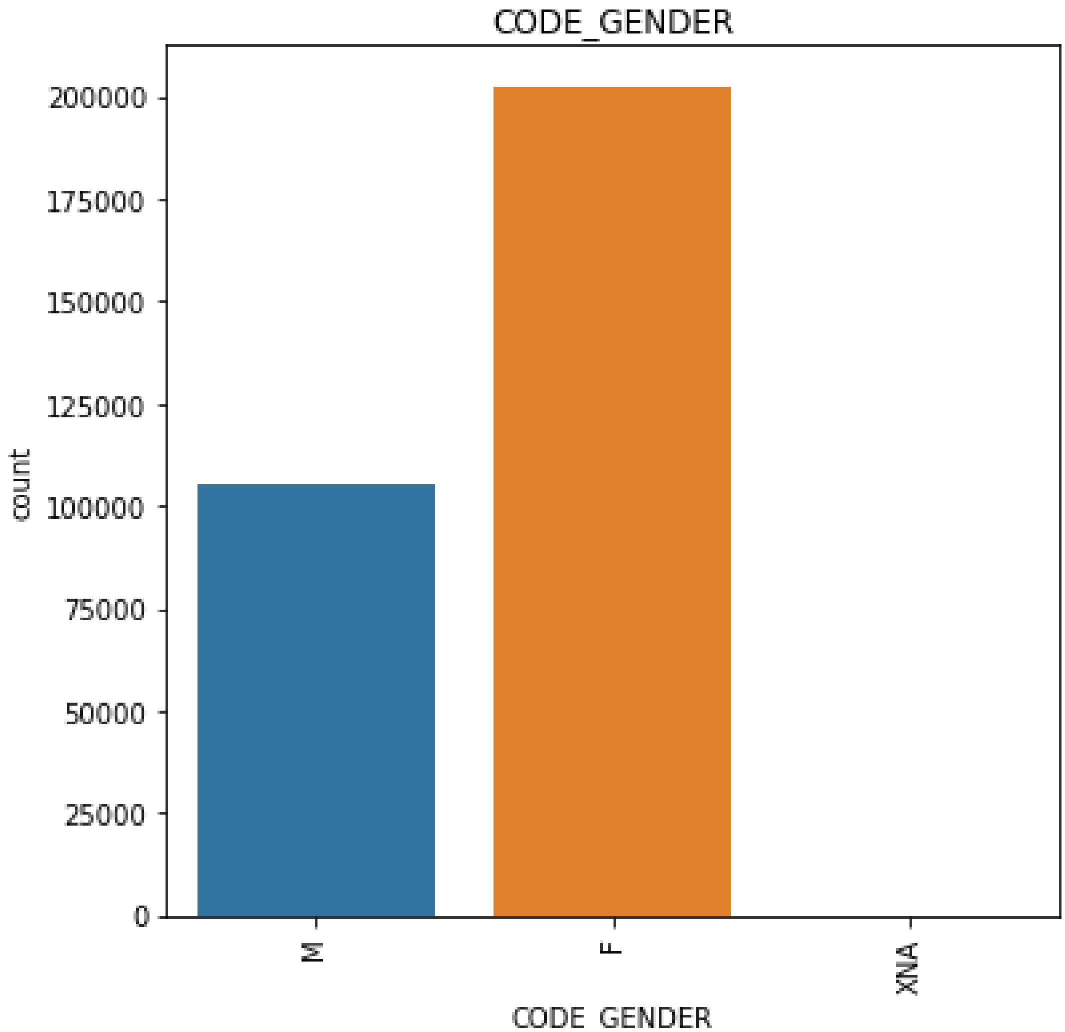
[ ]:

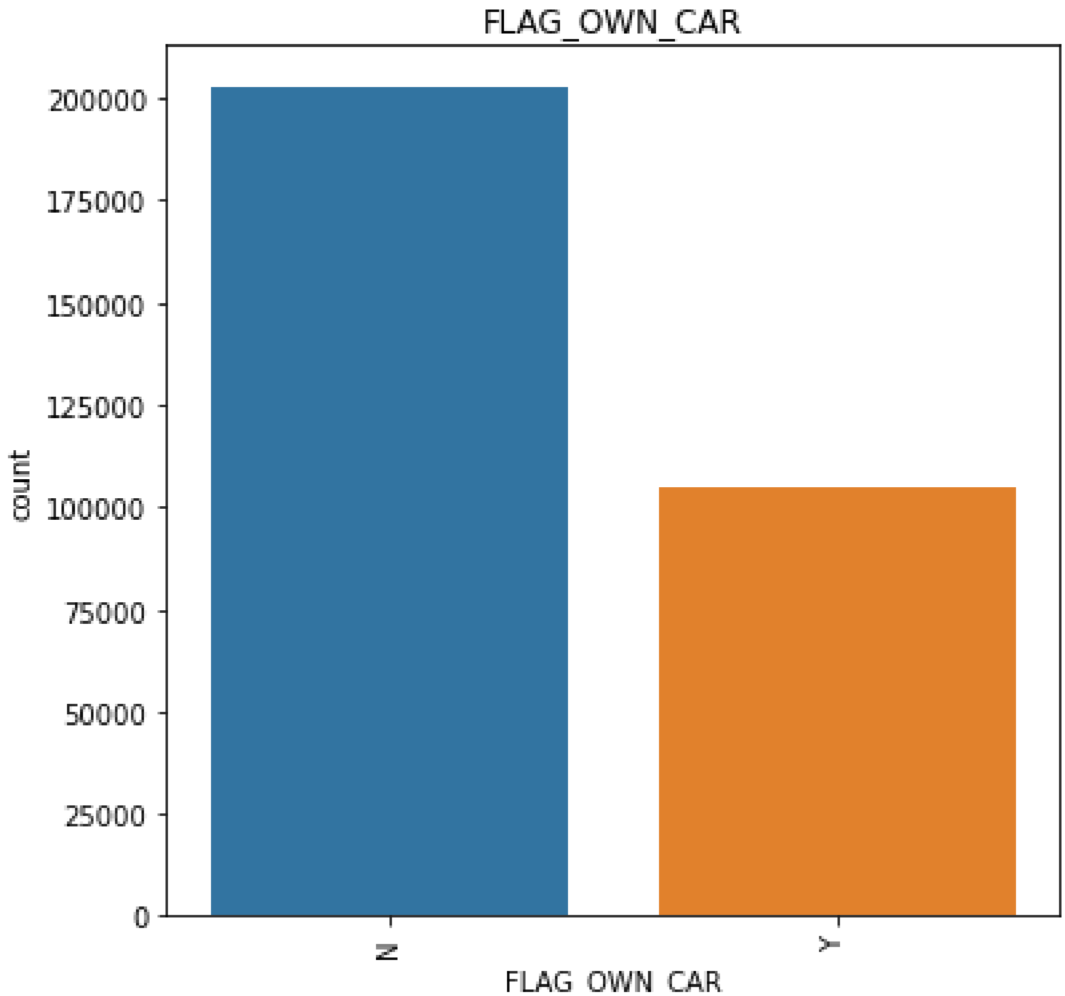
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

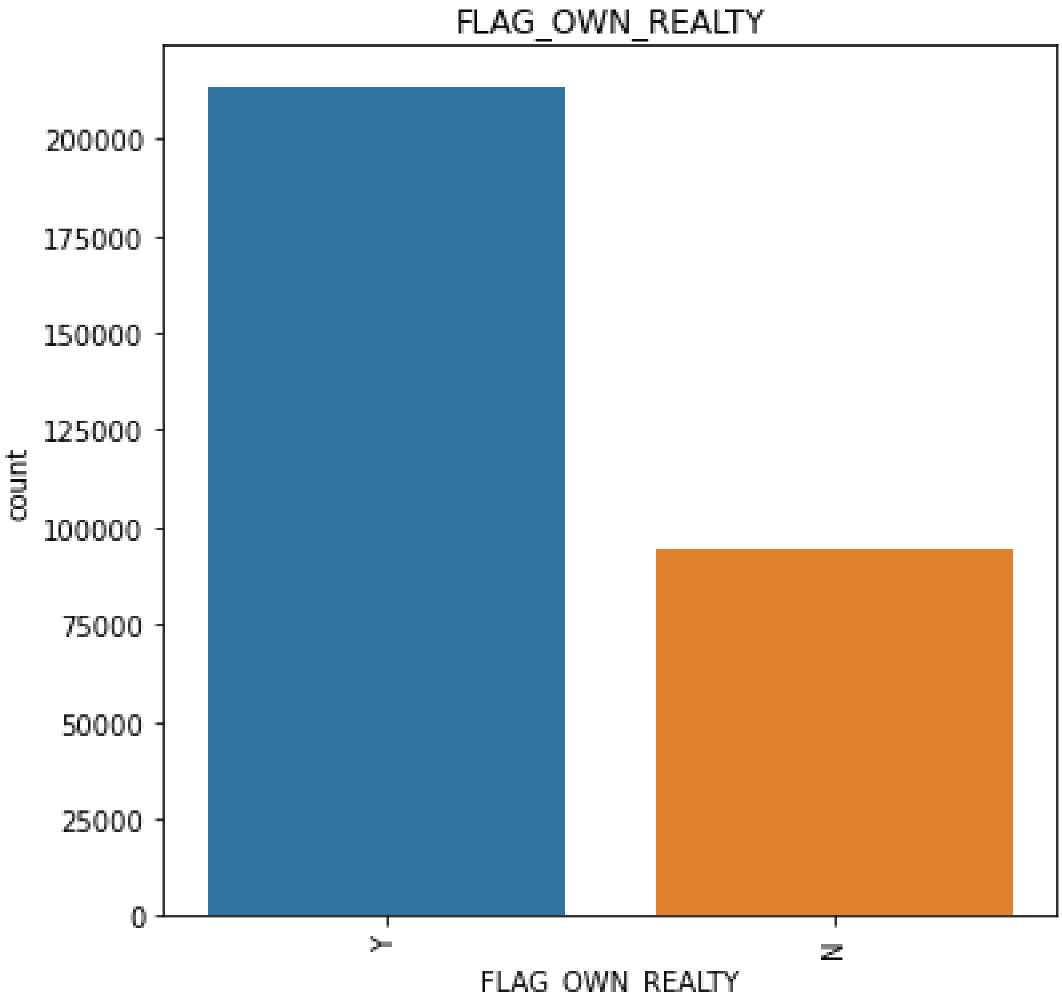
[52]:
```
for col in categorical1:
    plt.figure(figsize=[6,6])

    sns.countplot(x=df1[col])
    plt.title(col)
    plt.xticks(rotation=90)
    plt.show()
```

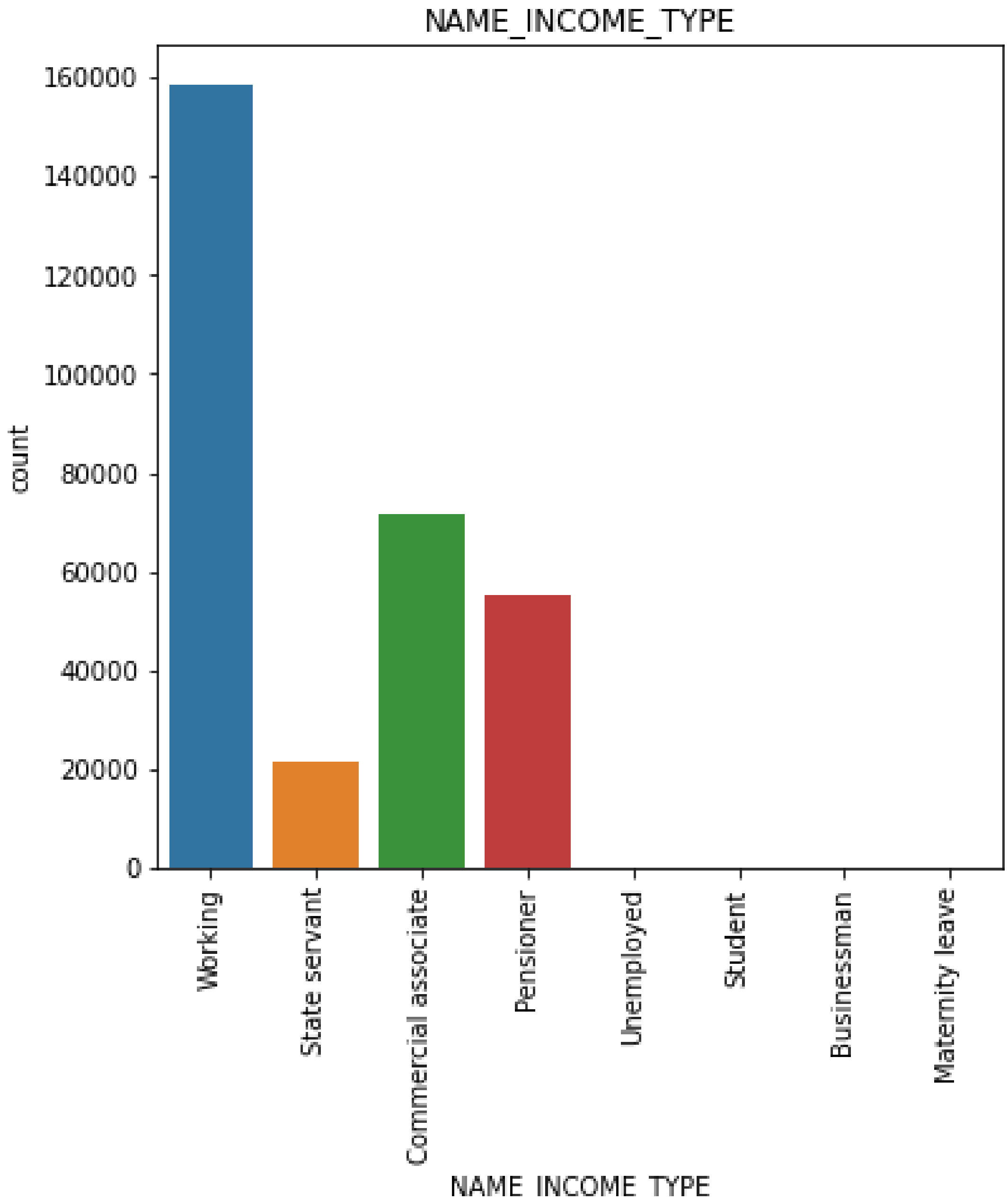

TARGET
count
250000
200000
150000
100000
50000
0
0
1
TARGET

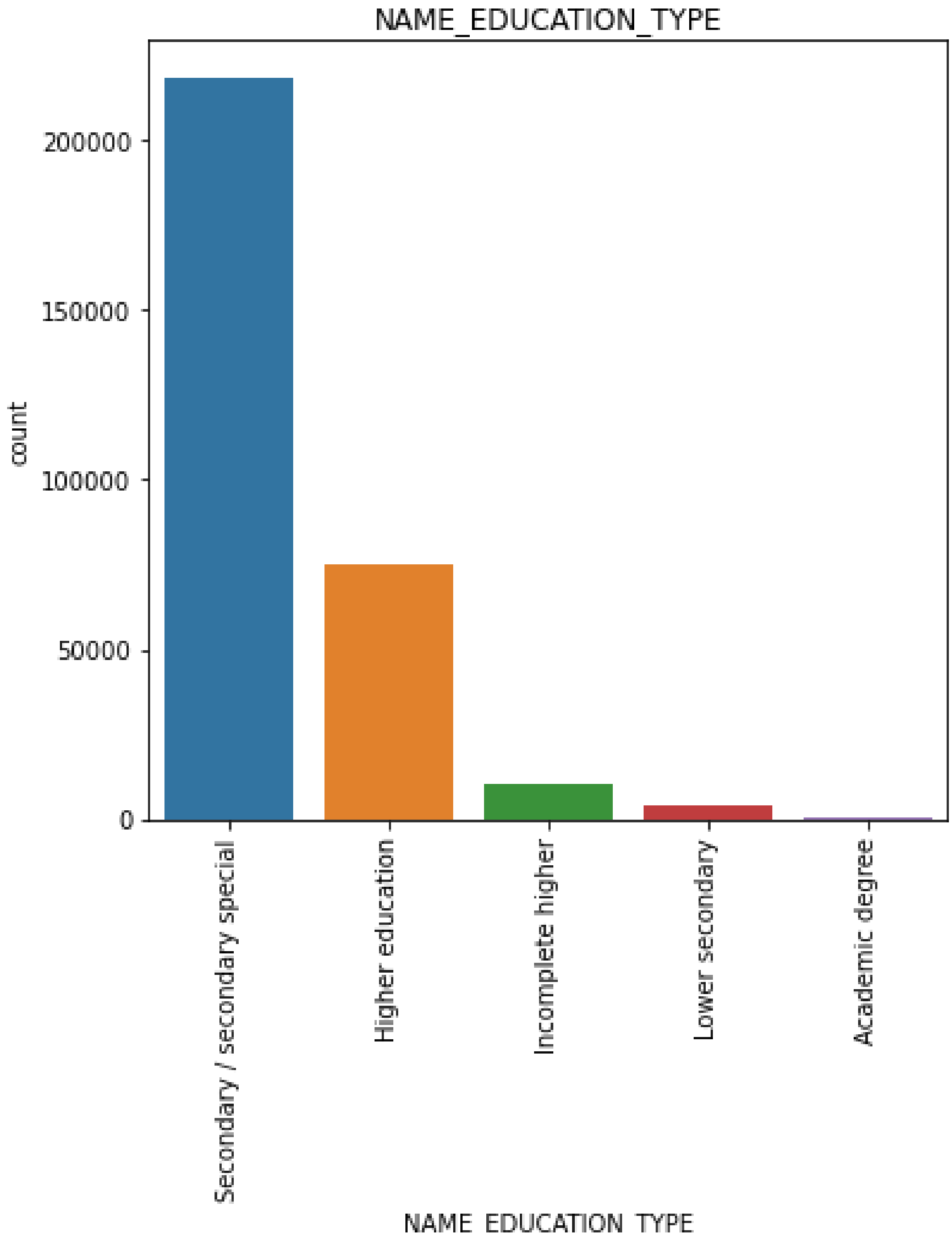

NAME_CONTRACT_TYPE
250000
200000
150000
100000
50000
0
count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

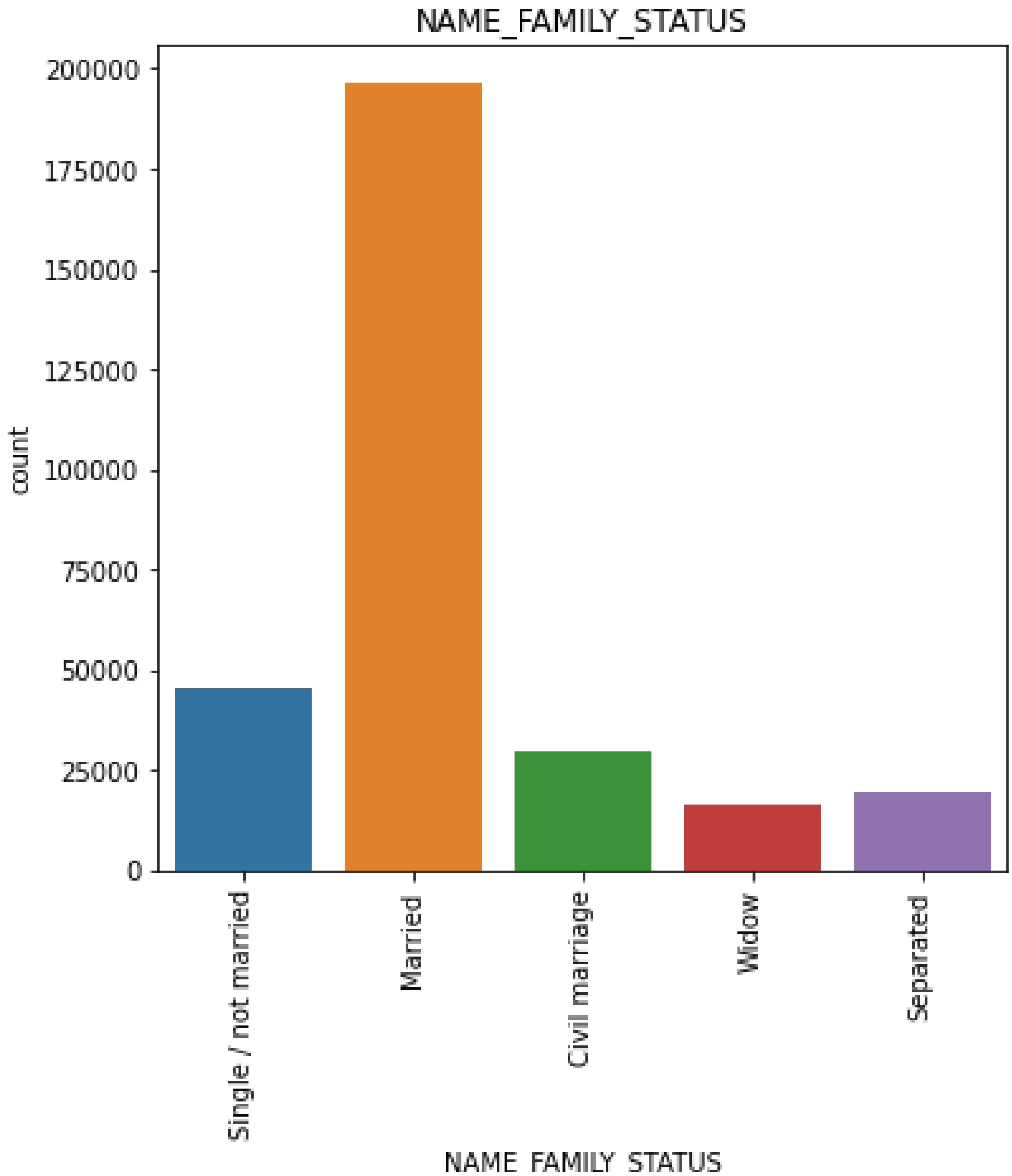

CODE_GENDER
count
200000
175000
150000
125000
100000
75000
50000
25000
0
M
F
XNA
CODE_GENDER

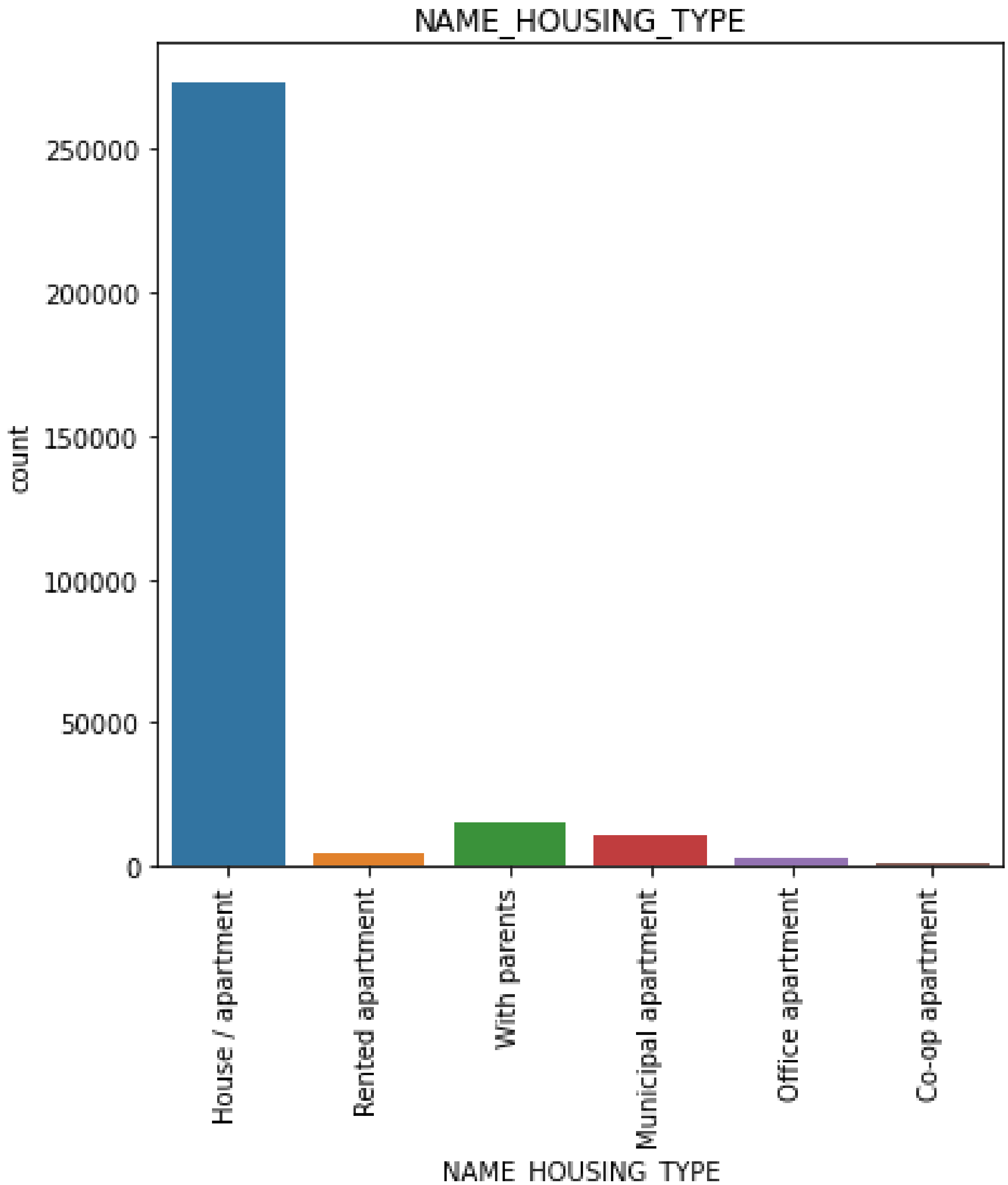

FLAG_OWN_CAR
count
200000
175000
150000
125000
100000
75000
50000
25000
0
N
Y
FLAG_OWN_CAR

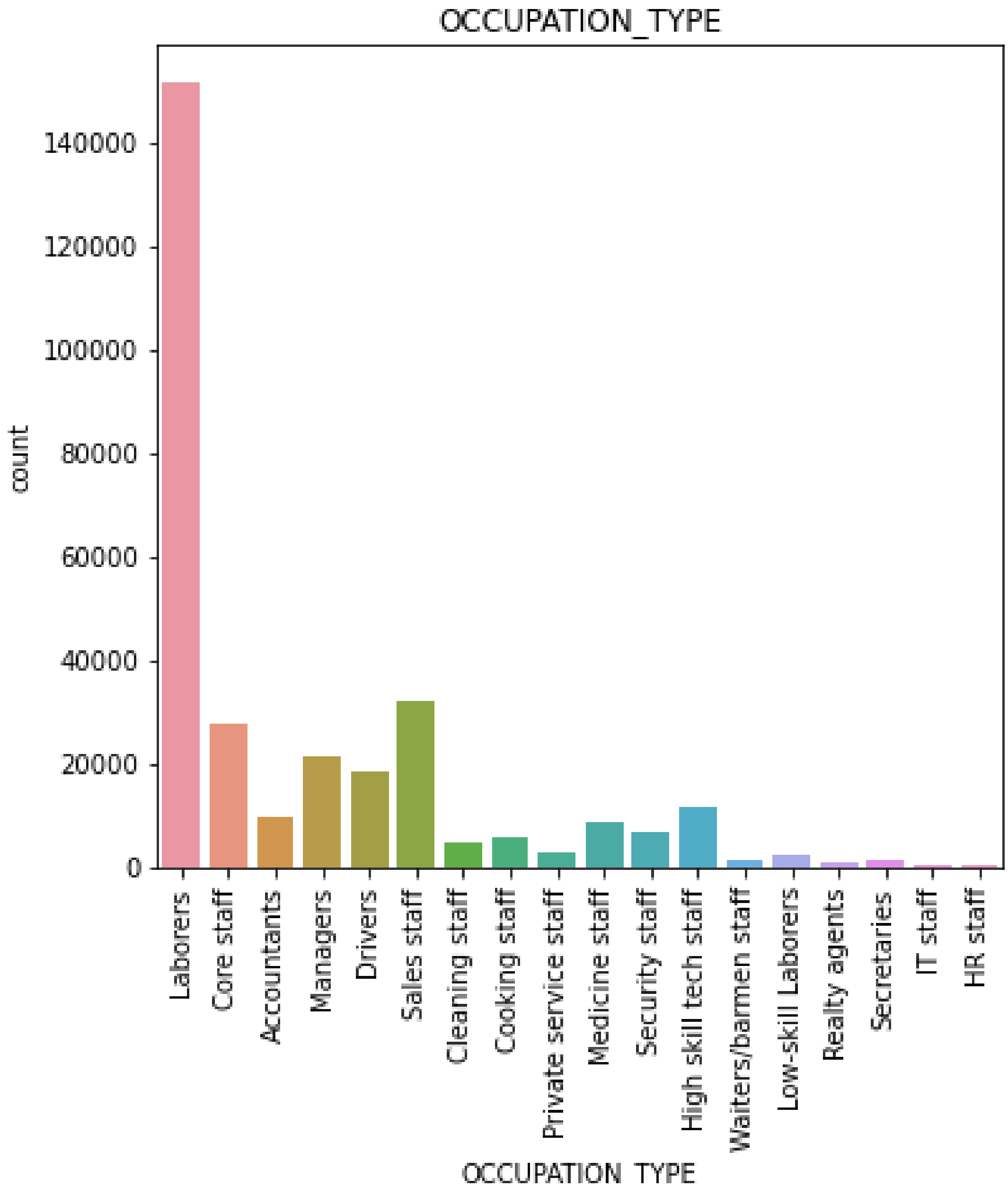

FLAG_OWN_REALTY
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Y
N
FLAG_OWN_REALTY

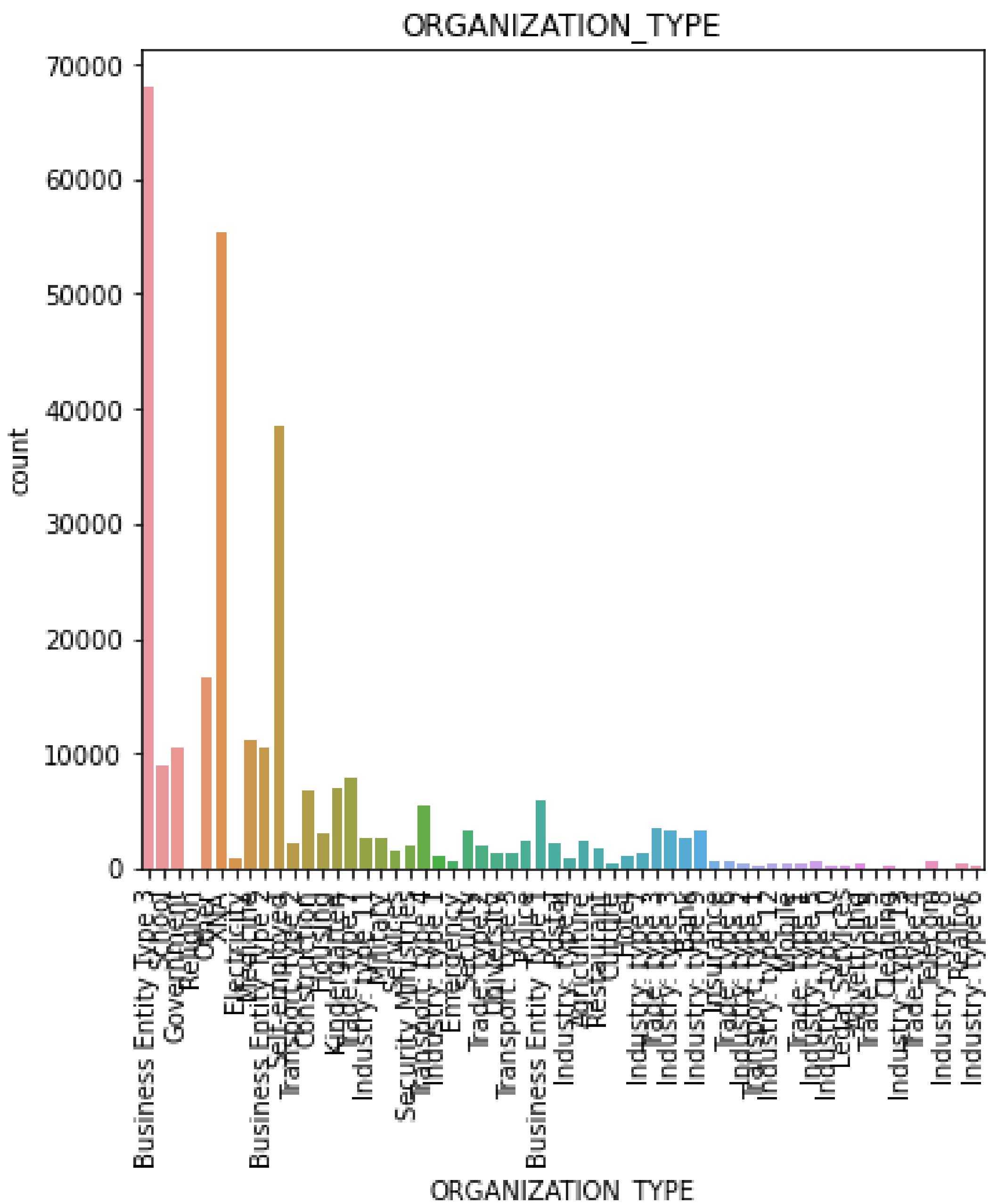

NAME_INCOME_TYPE
count
160000
140000
120000
100000
80000
60000
40000
20000
0
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


NAME_EDUCATION_TYPE
count
200000
150000
100000
50000
0
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


NAME_FAMILY_STATUS
count
200000
175000
150000
125000
100000
75000
50000
25000
0
Single / not married
Married
Civil marriage
Widow
Separated
NAME_FAMILY_STATUS


NAME_HOUSING_TYPE
count
250000
200000
150000
100000
50000
0
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE

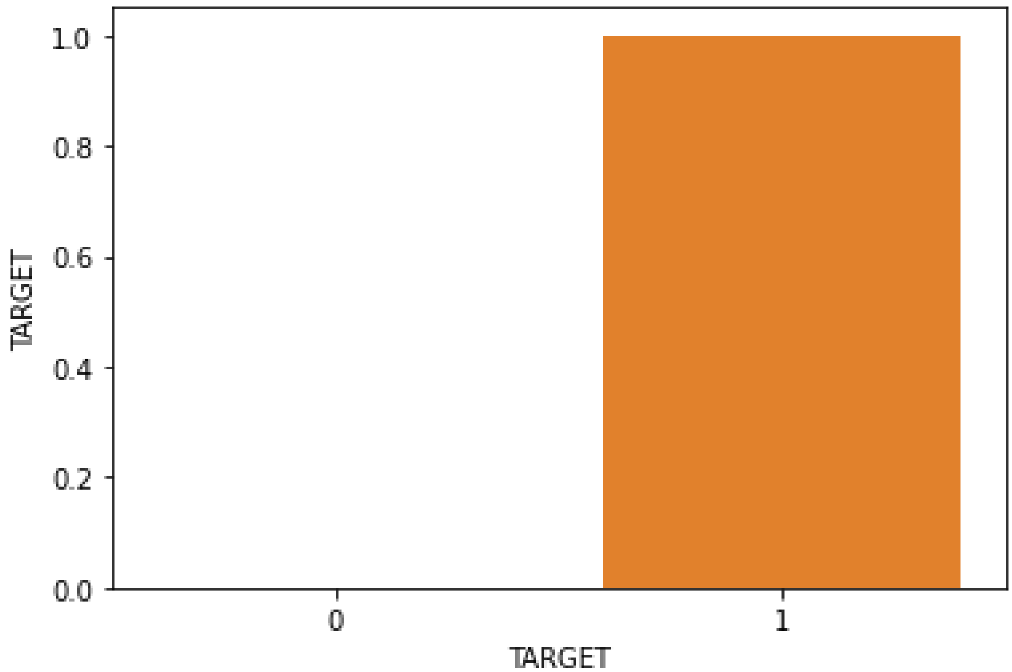

OCCUPATION_TYPE
count
0
20000
40000
60000
80000
100000
120000
140000
Laborers
Core staff
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Cooking staff
Private service staff
Medicine staff
Security staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE

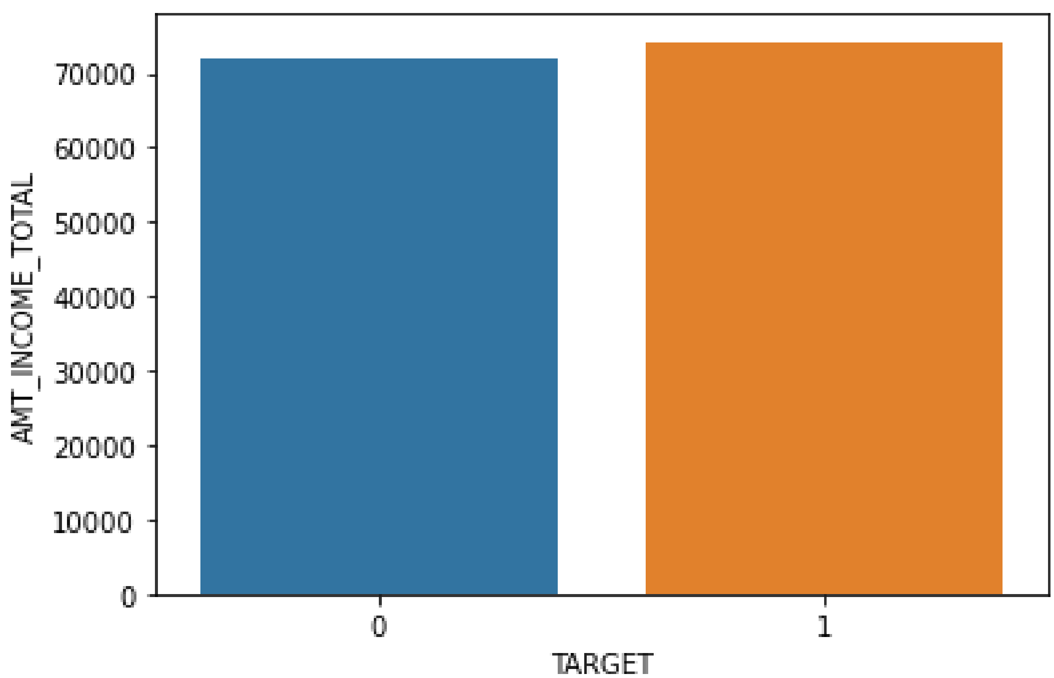


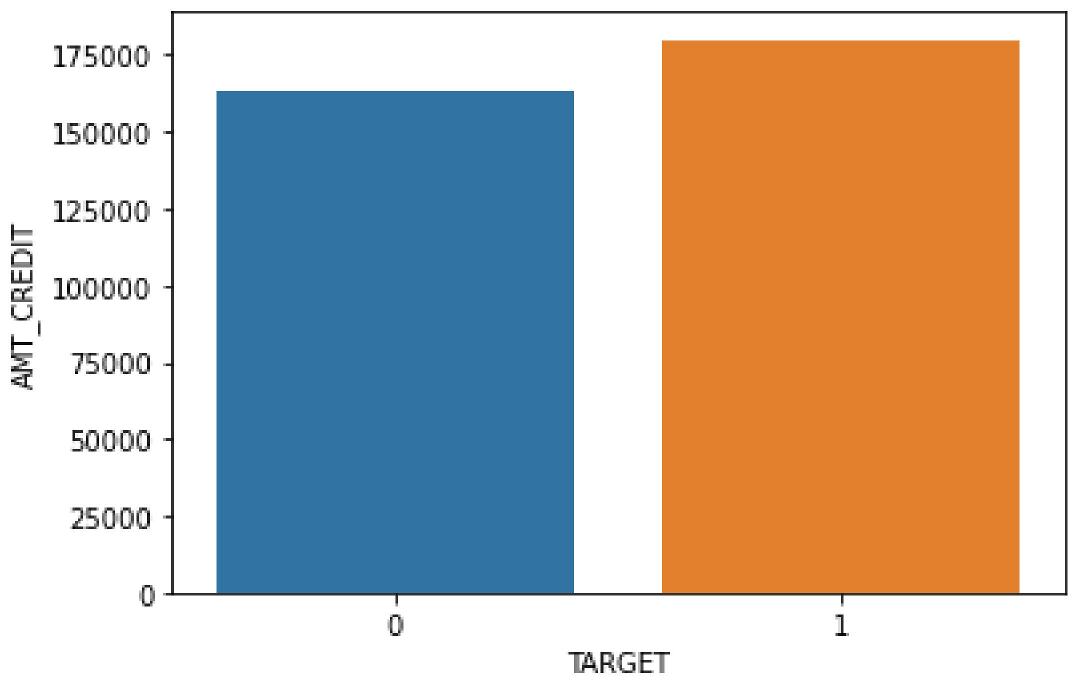
0.0.8 Bivariate Analysis

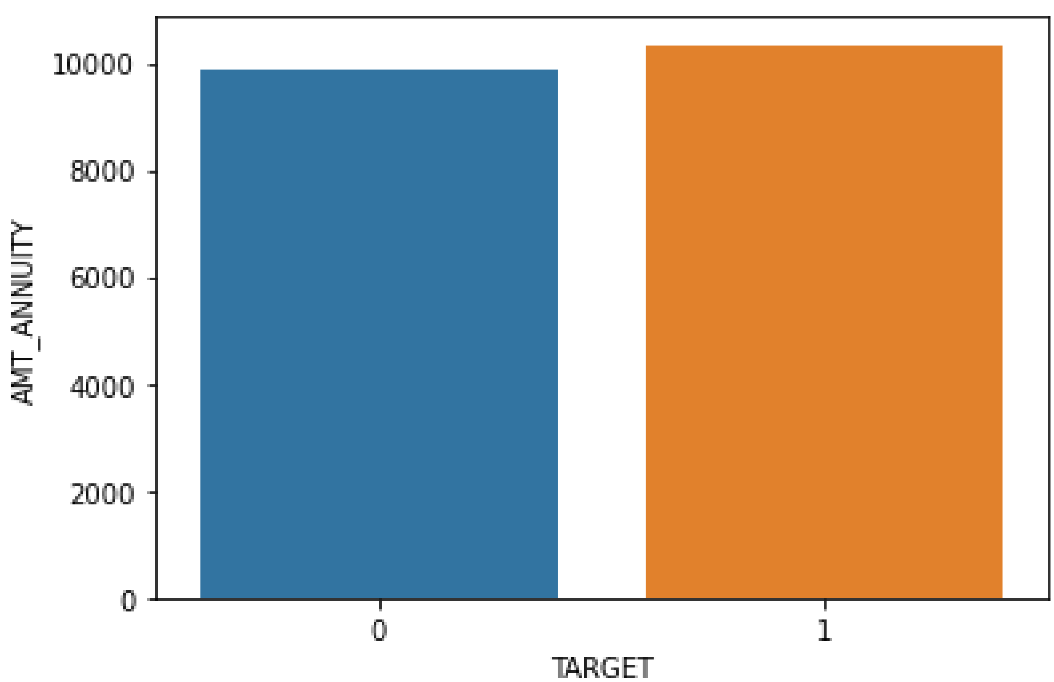
Categorical Vs Continuous

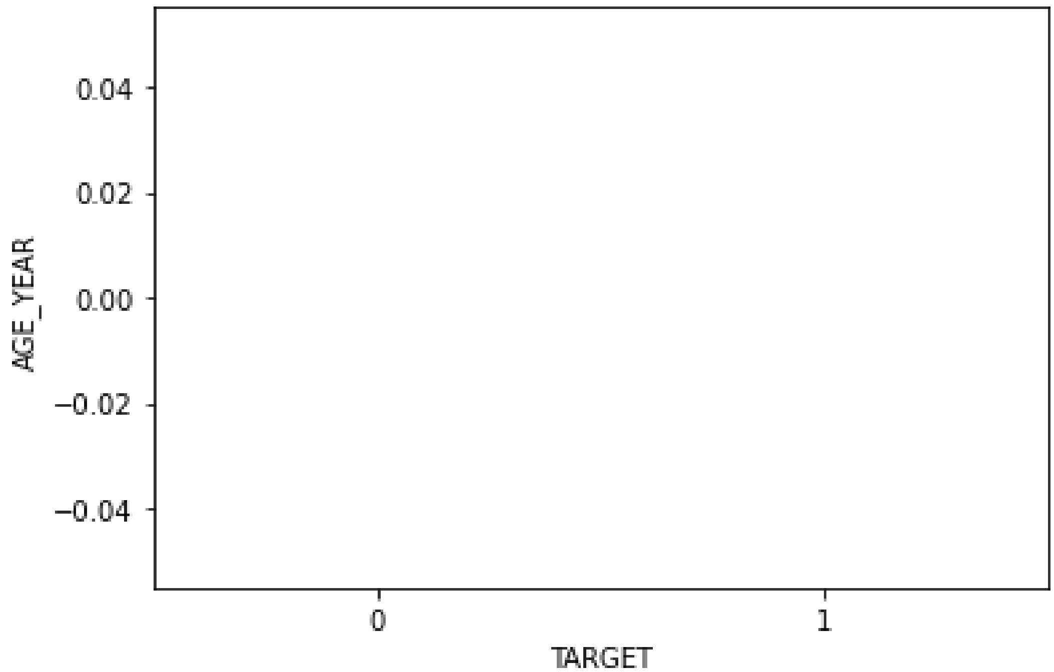
[53]:
```
for col in continuous1:
    sns.barplot(x=df1['TARGET'],y=df1[col],ci=None,estimator=lambda x:np.
 ↪quantile(x,0.075))
    plt.show()
```

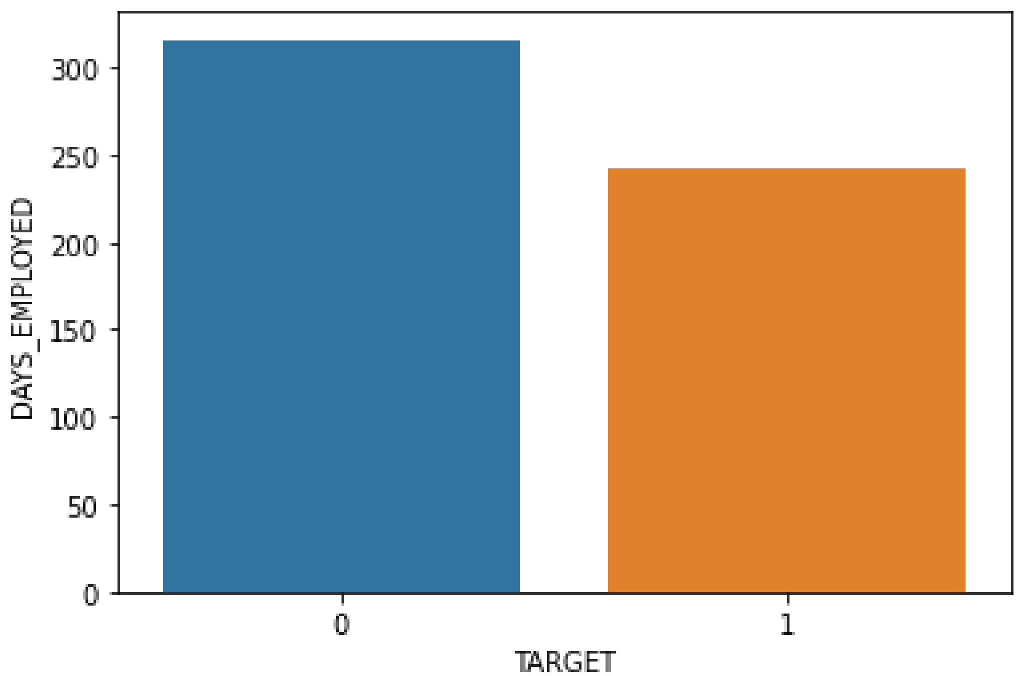

1.0
0.8
0.6
0.4
0.2
0.0
TARGET
0
1
TARGET


70000
60000
50000
40000
30000
20000
10000
0
AMT_INCOME_TOTAL
0
1
TARGET


175000
150000
125000
100000
75000
50000
25000
0
AMT_CREDIT
0
1
TARGET

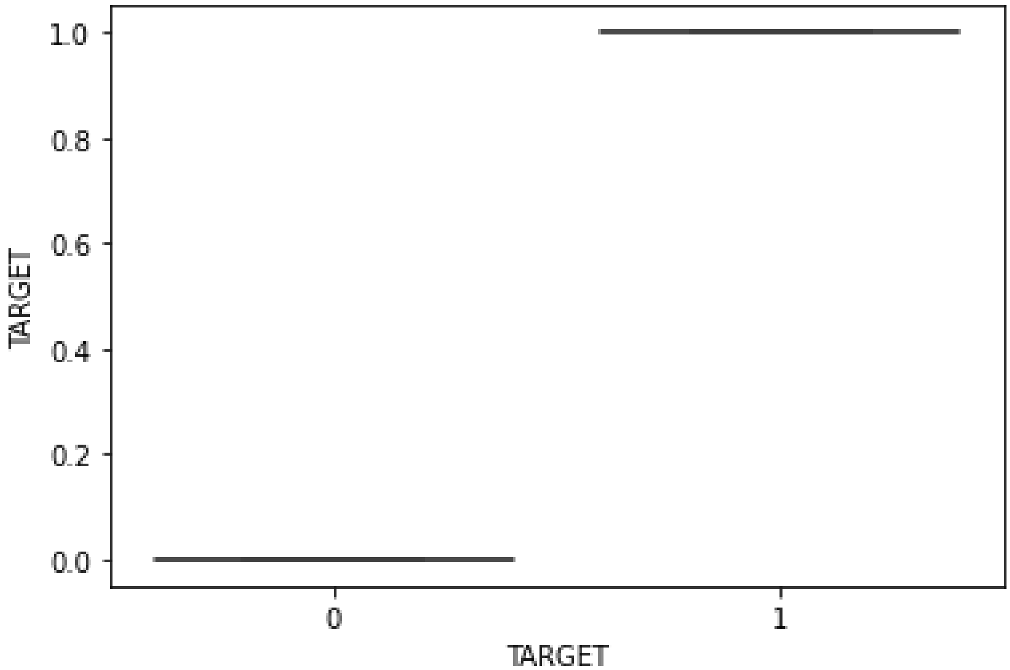

10000
8000
6000
4000
2000
0
AMT_ANNUITY
0
1
TARGET



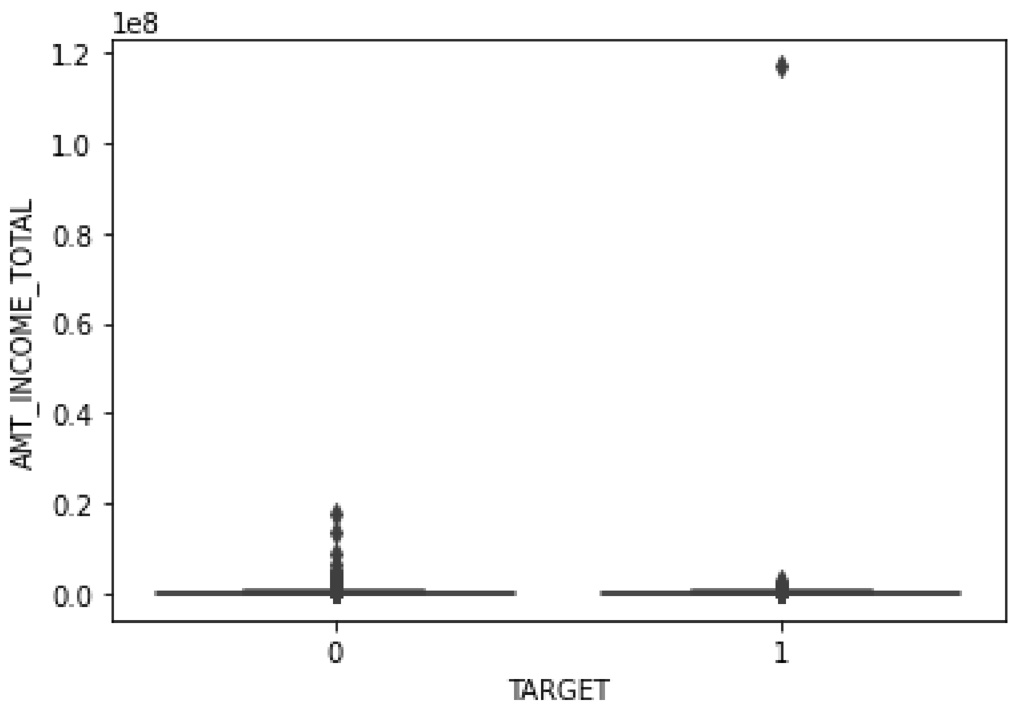


[54]:
```
for col in continuous1:
    sns.boxplot(x=df1['TARGET'],y=df1[col])
```

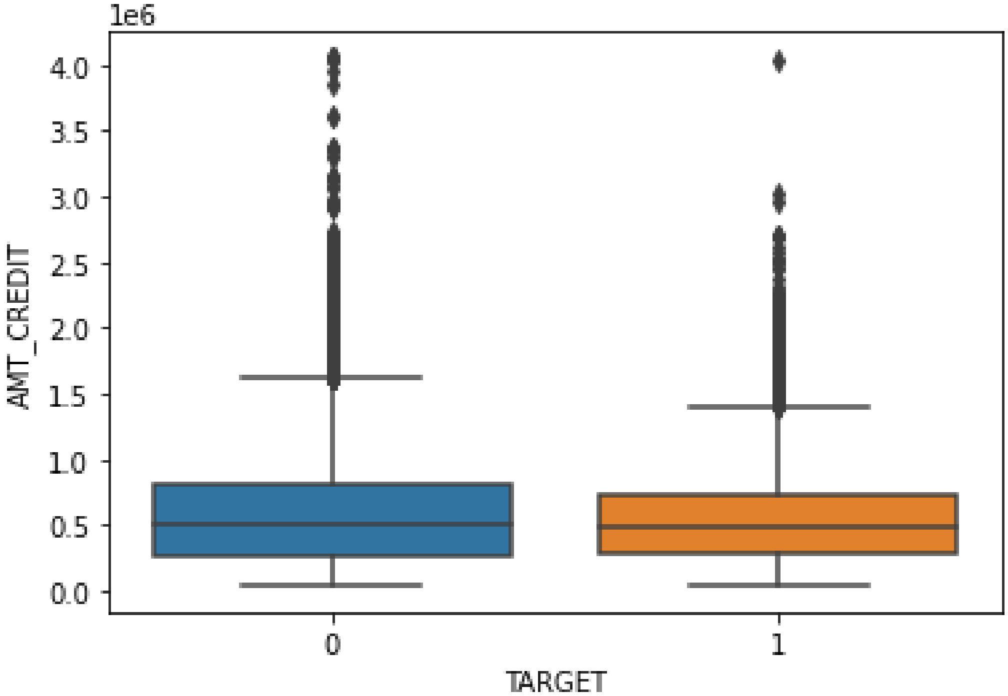
```
plt.show()
```

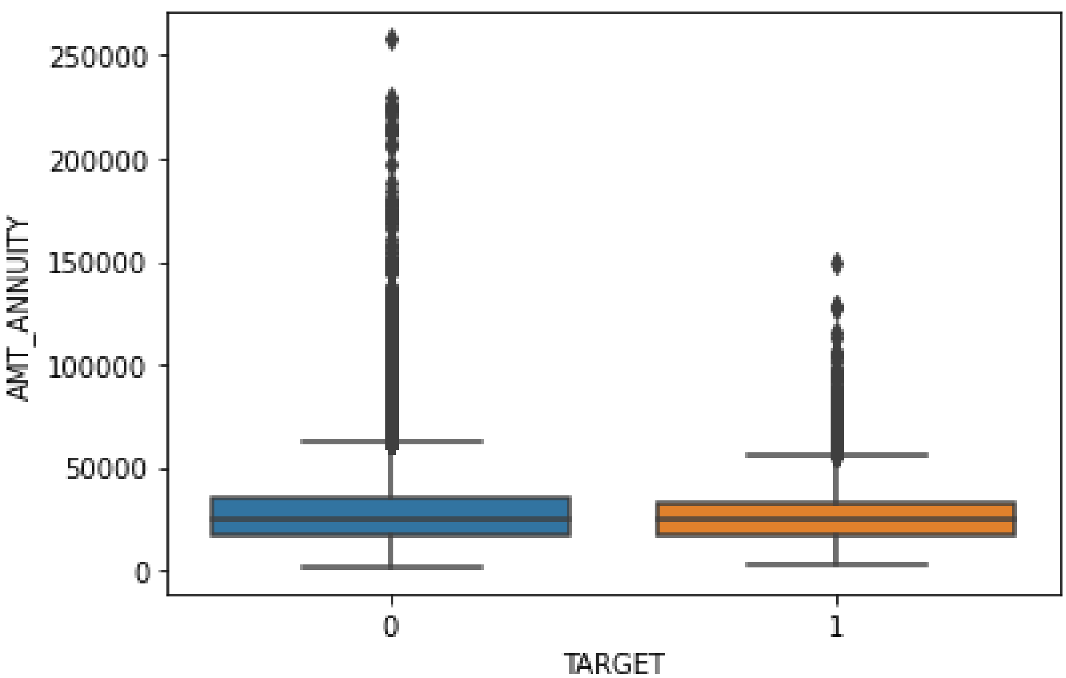


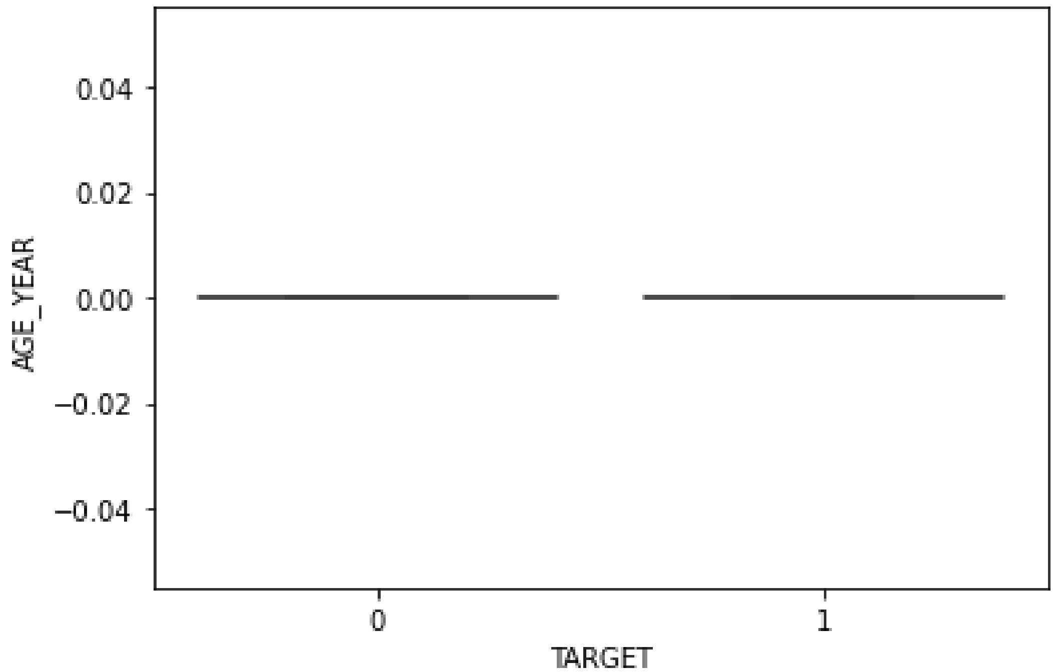
1.0
0.8
0.6
0.4
0.2
0.0
TARGET
0
1
TARGET

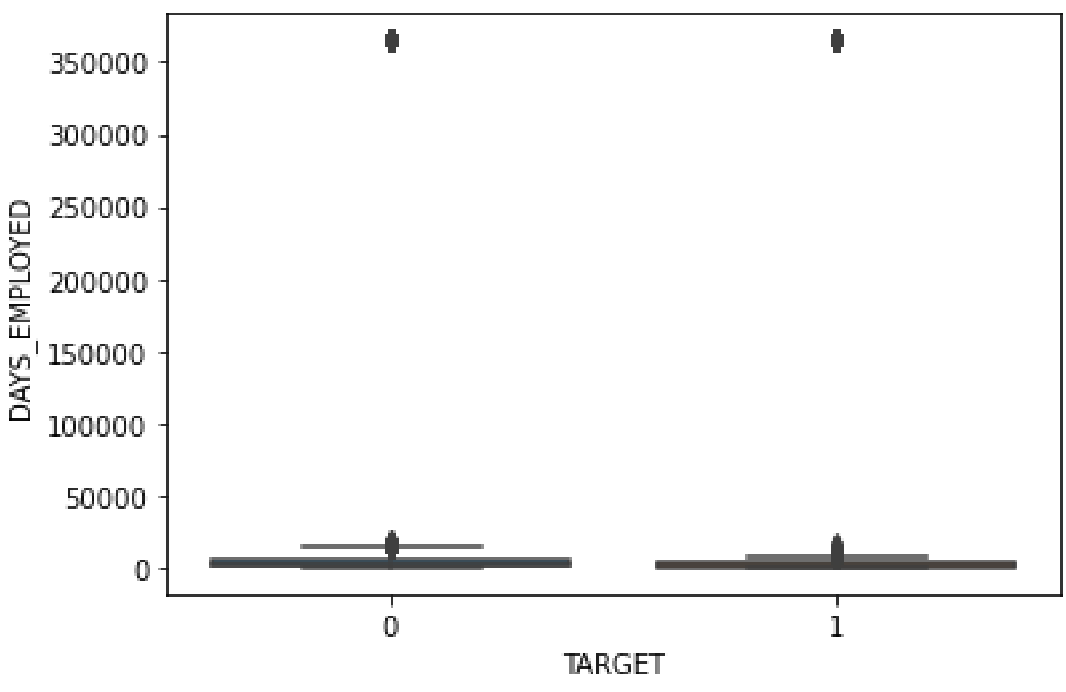


1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
AMT_INCOME_TOTAL
0
1
TARGET


1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
AMT_CREDIT
0
1
TARGET

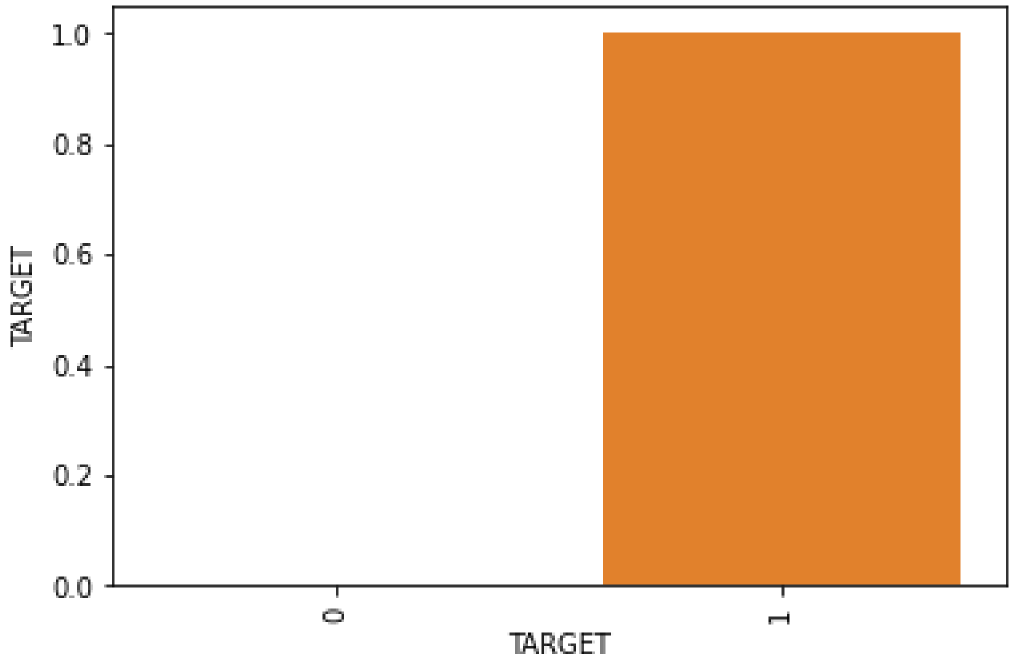

250000
200000
150000
100000
50000
0
AMT_ANNUITY
0
1
TARGET

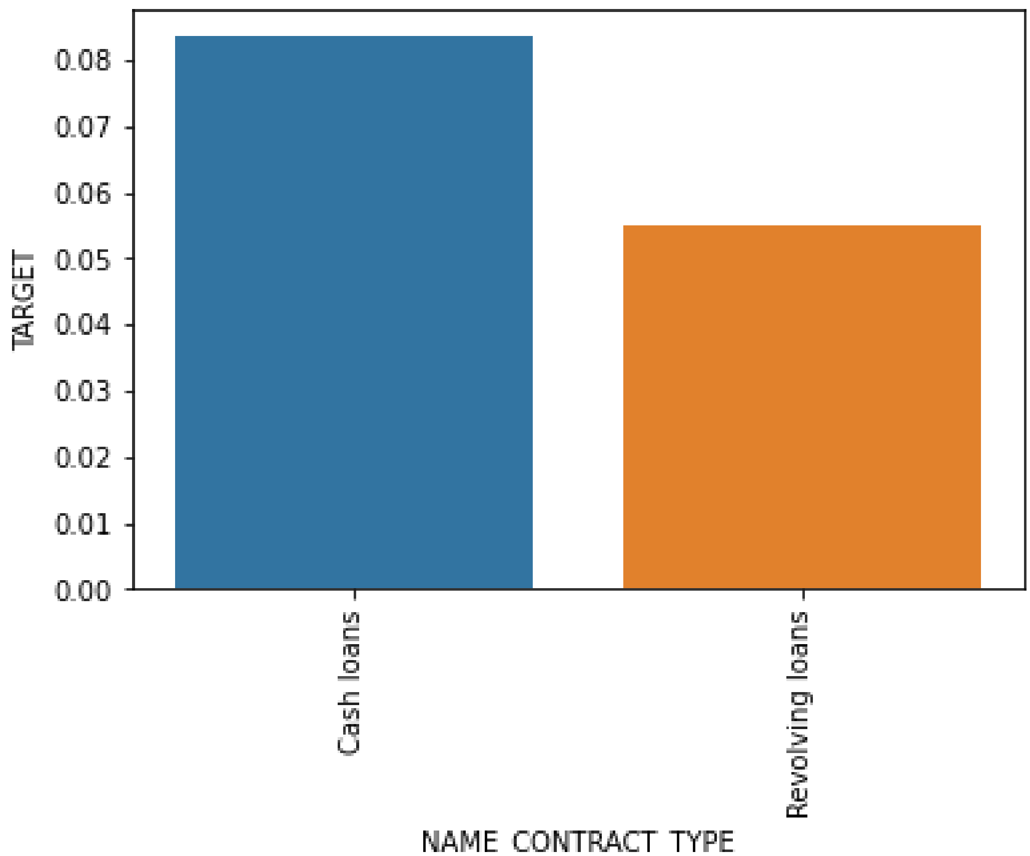

0.04
0.02
0.00
−0.02
−0.04
AGE_YEAR
0
1
TARGET


350000
300000
250000
200000
150000
100000
50000
0
DAYS_EMPLOYED
0
1
TARGET

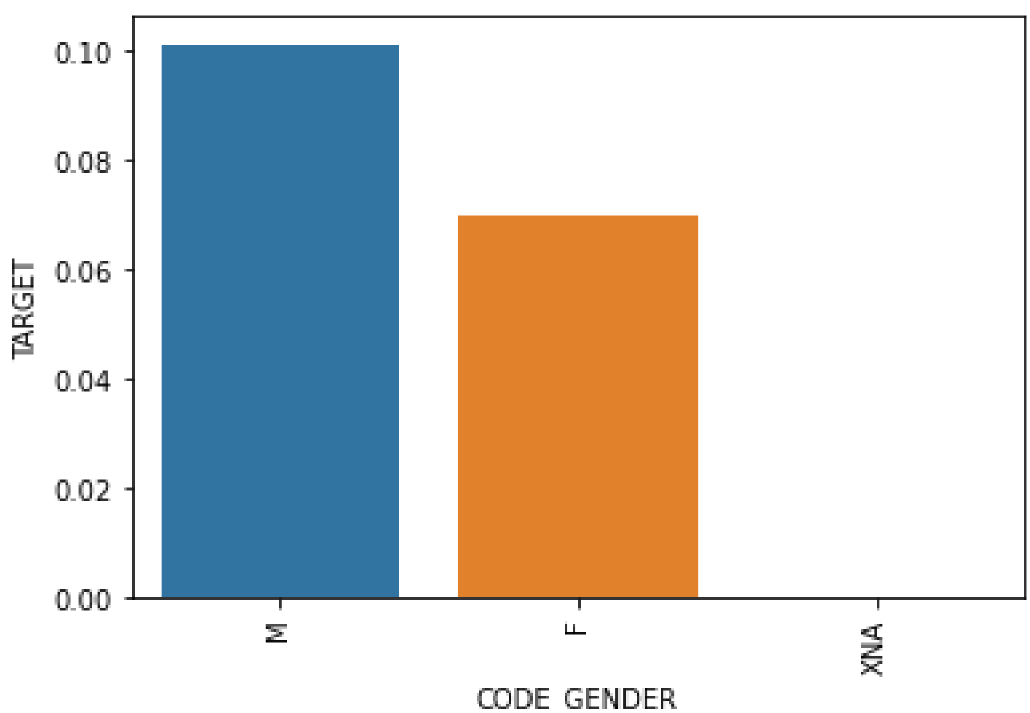
0.0.9 Categorical Vs Categorical

[55]:
```
for col in categorical1:
    sns.barplot(y=df1['TARGET'],x=df1[col],ci=None)
    plt.xticks(rotation=90)
    plt.show()
```

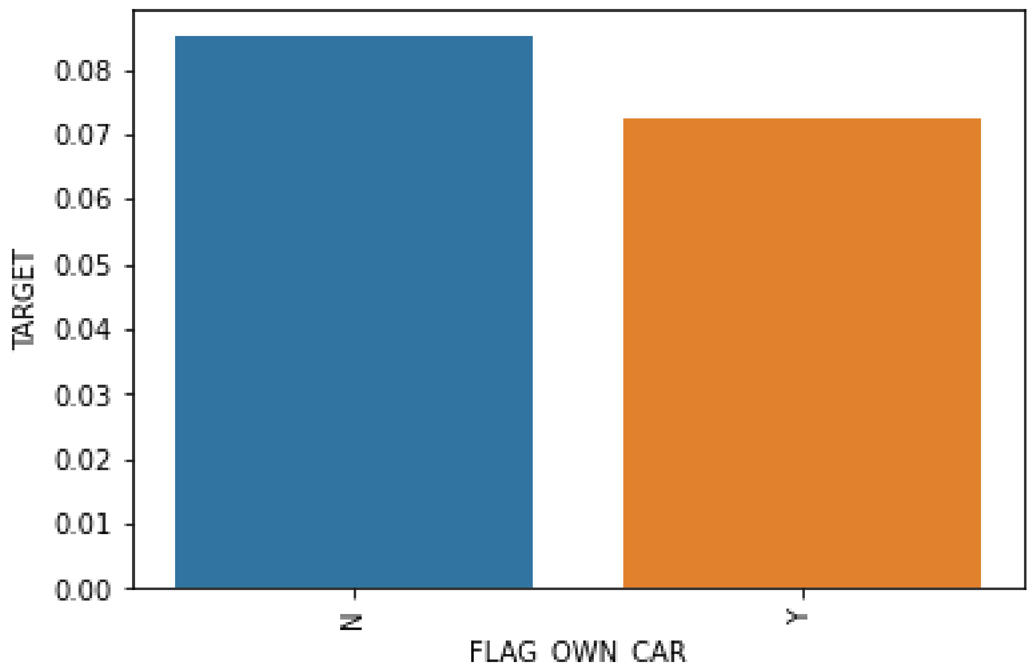




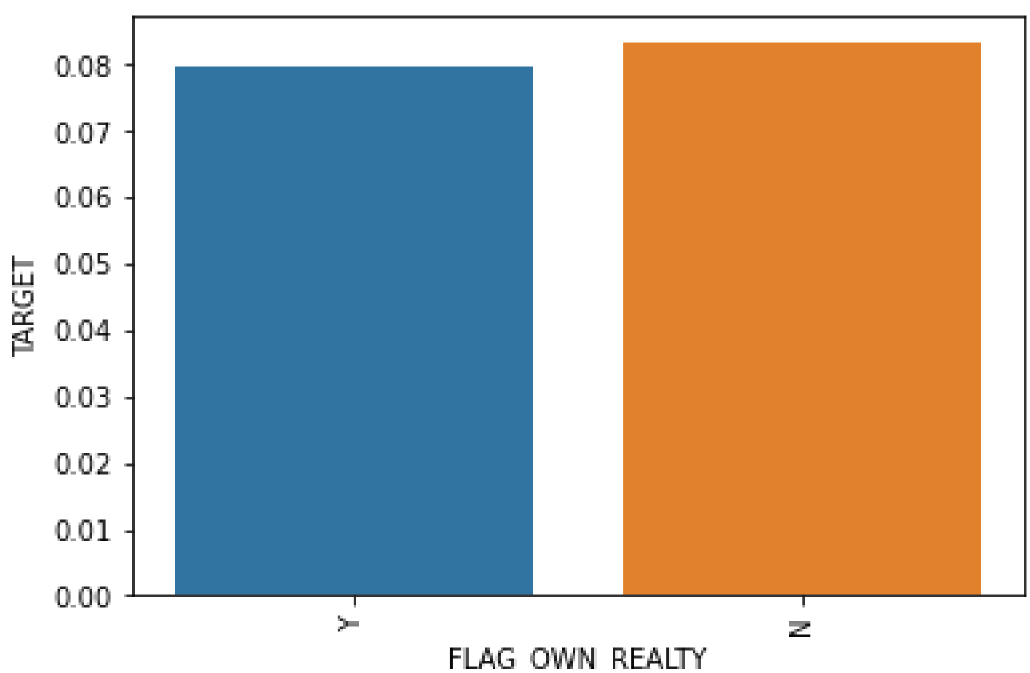
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
TARGET
Cash loans
Revolving loans
NAME_CONTRACT_TYPE



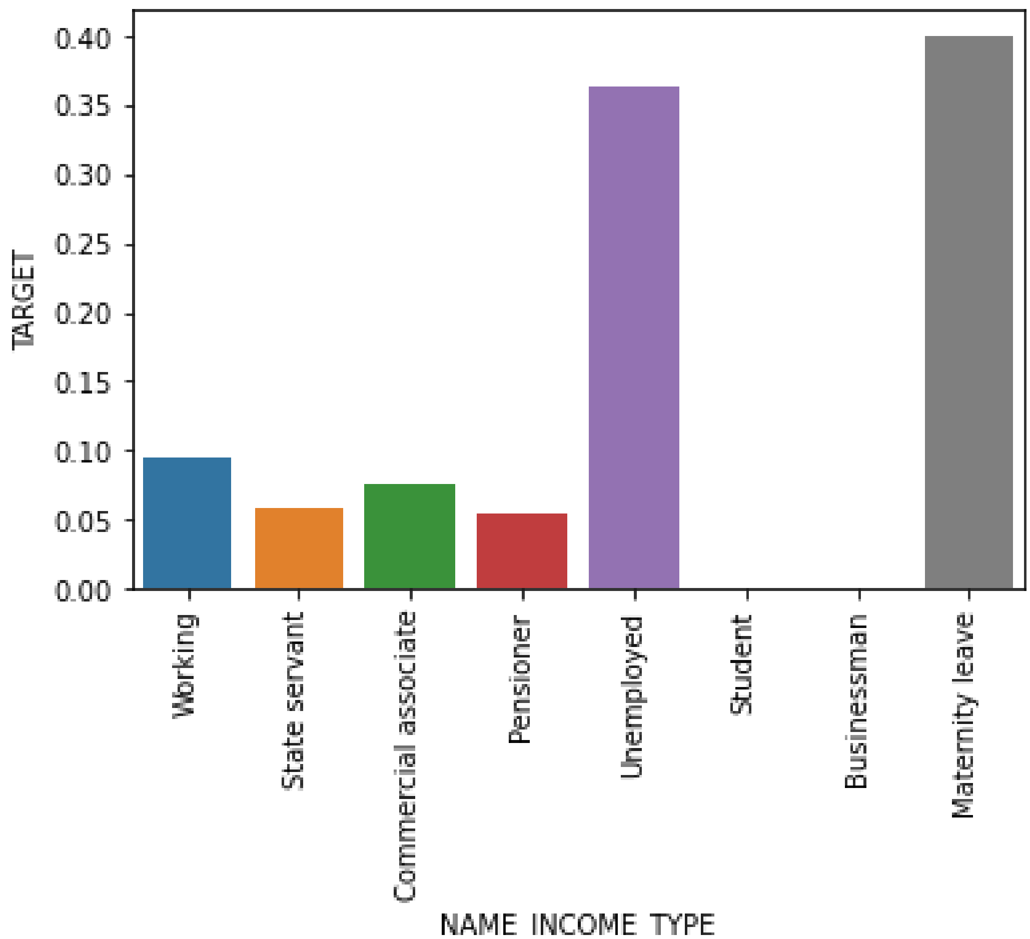
0.10
0.08
0.06
0.04
0.02
0.00
TARGET
M
F
XNA
CODE_GENDER

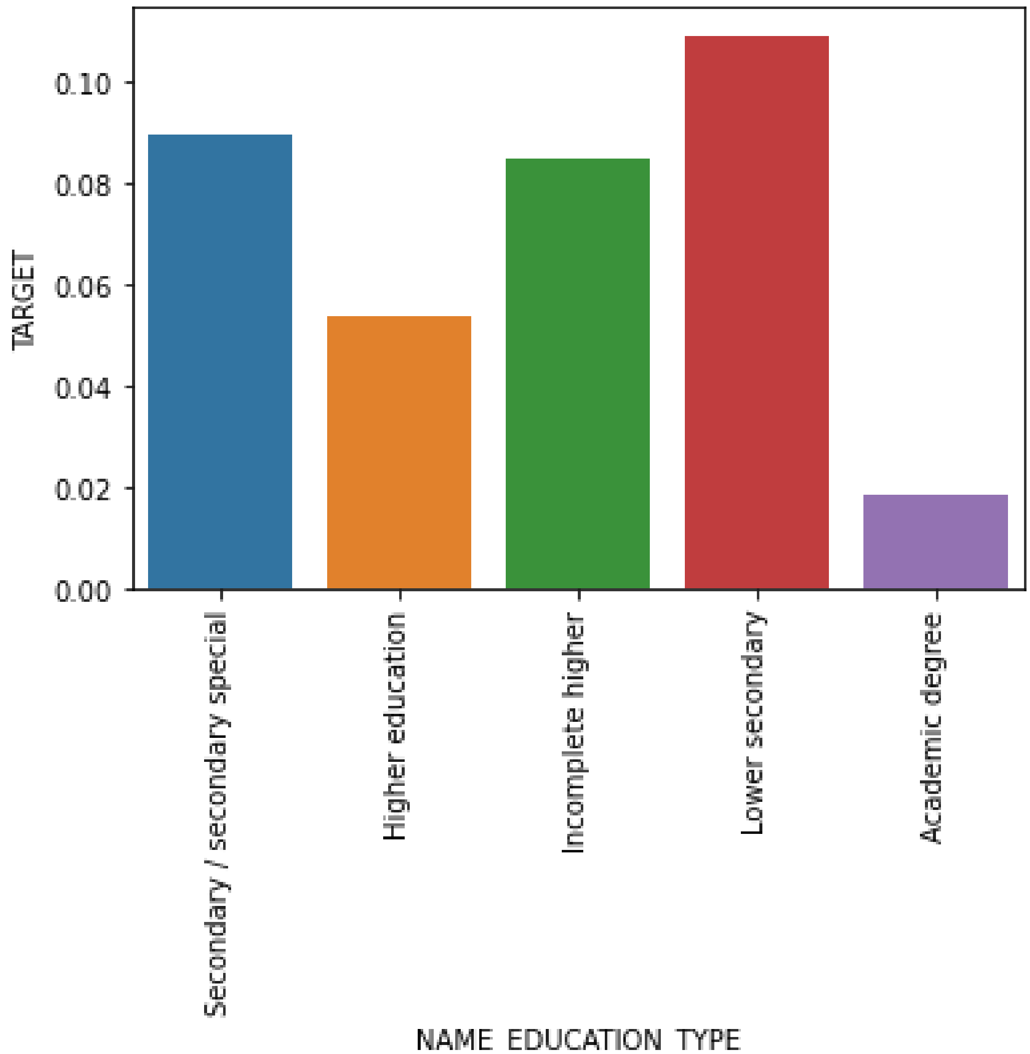


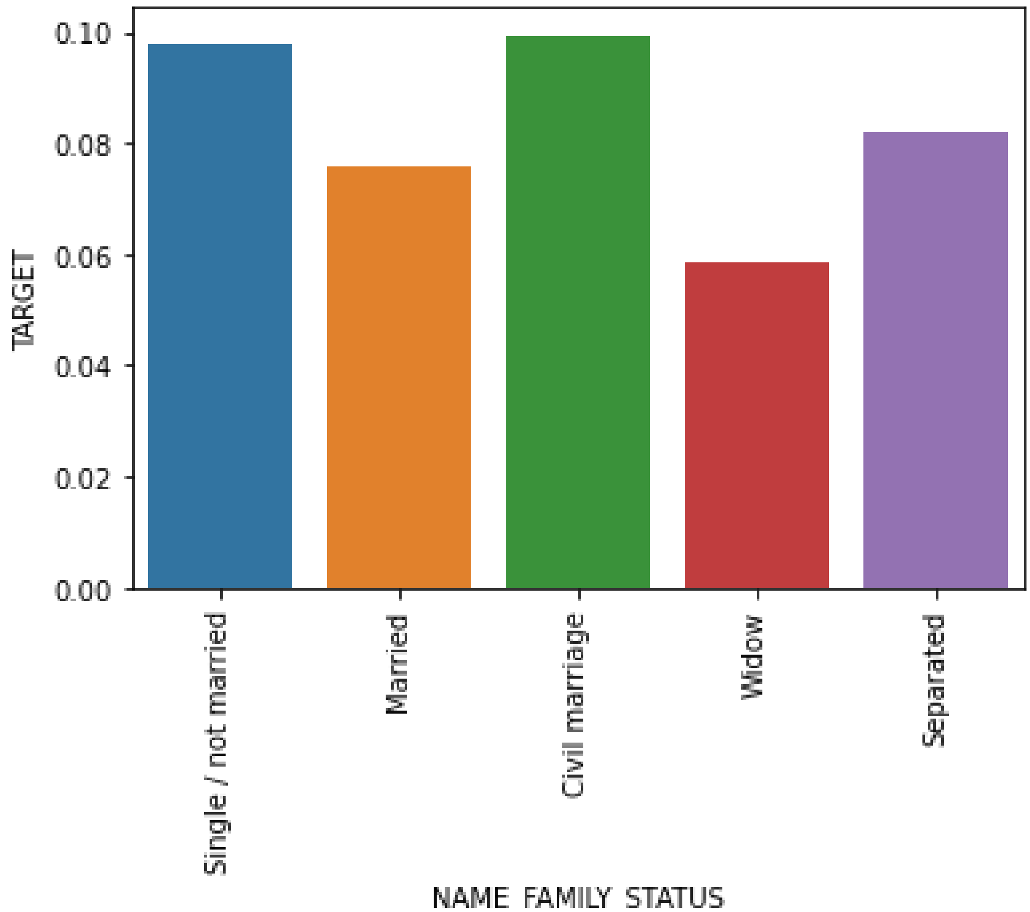
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
TARGET
N
Y
FLAG_OWN_CAR

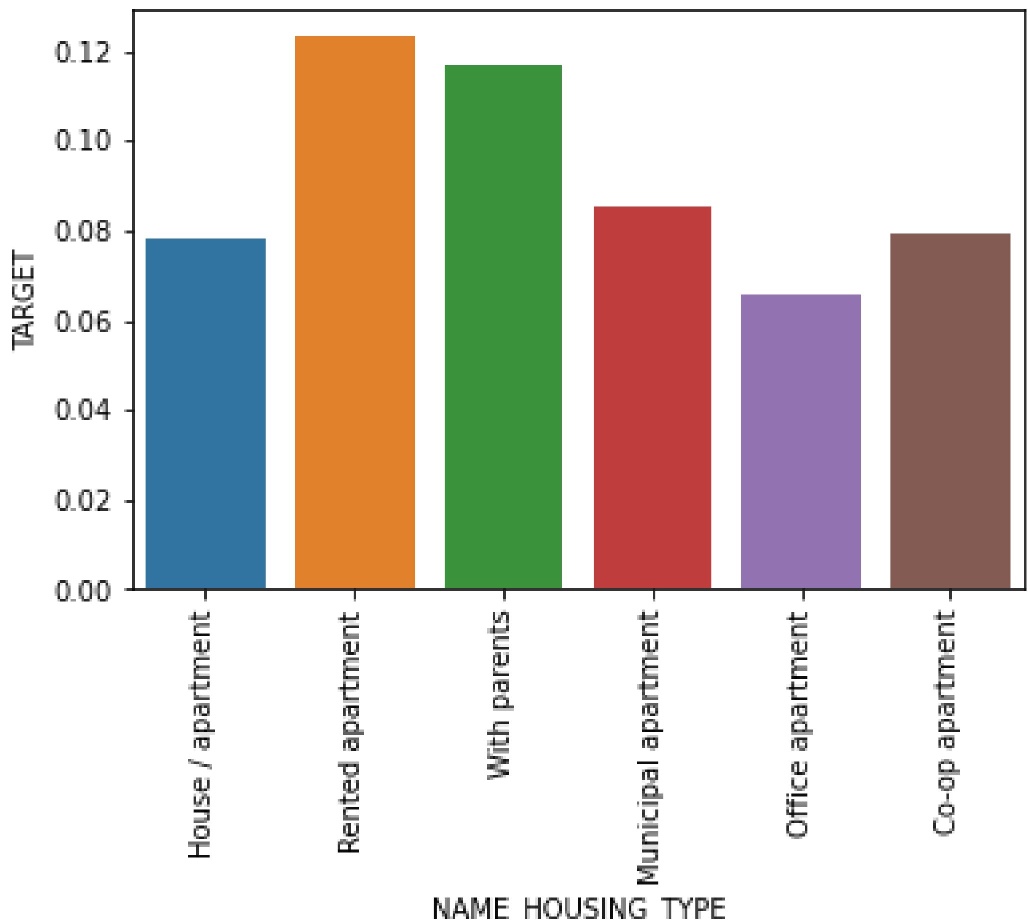


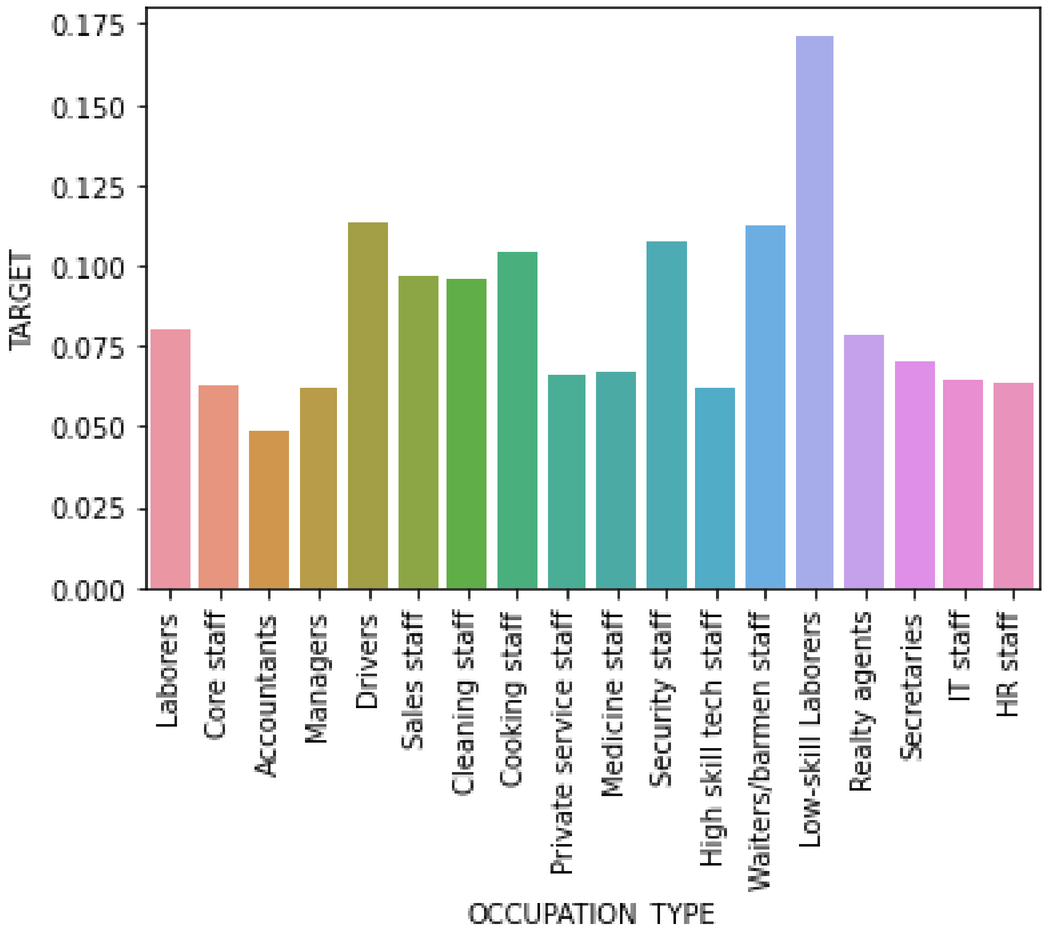
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
TARGET
Y
N
FLAG_OWN_REALTY

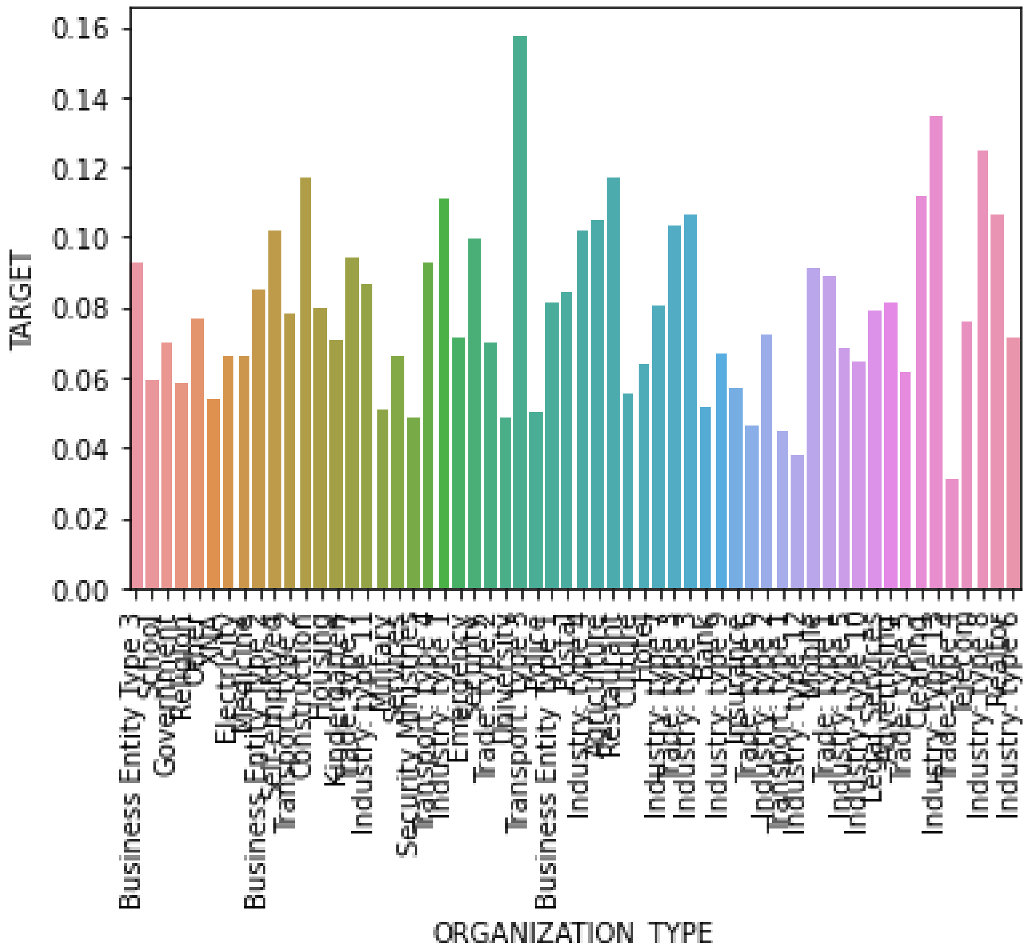

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
TARGET
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


TARGET
0.10
0.08
0.06
0.04
0.02
0.00
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


0.10
0.08
0.06
0.04
0.02
0.00
TARGET
Single / not married
Married
Civil marriage
Widow
Separated
NAME_FAMILY_STATUS


0.12
0.10
0.08
0.06
0.04
0.02
0.00
TARGET
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
TARGET
Laborers
Core staff
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Cooking staff
Private service staff
Medicine staff
Security staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE



0.1 Multivariate

[56]:
```
sns.heatmap(df1[continuous1].corr(),annot=True,cmap="RdYlGn");
```

0.1.1 VISUALIZATION FOR DATA FRAME TWO

[57]:
```
df2.head()
```

[57]:
```
   SK_ID_CURR  TARGET  AMT_CREDIT  AMT_DOWN_PAYMENT NAME_CASH_LOAN_PURPOSE  \
0      271877     0.0     17145.0              0.00                    XAP
1      108129     0.0    679671.0           1640.25                    XNA
2      122040     0.0    136444.5           1640.25                    XNA
3      176158     0.0    470790.0           1640.25                    XNA
4      202054     0.0    404055.0           1640.25                Repairs

  NAME_CONTRACT_STATUS CODE_REJECT_REASON NAME_CLIENT_TYPE
0             Approved                XAP         Repeater
1             Approved                XAP         Repeater
2             Approved                XAP         Repeater
3             Approved                XAP         Repeater
4              Refused                 HC         Repeater
```

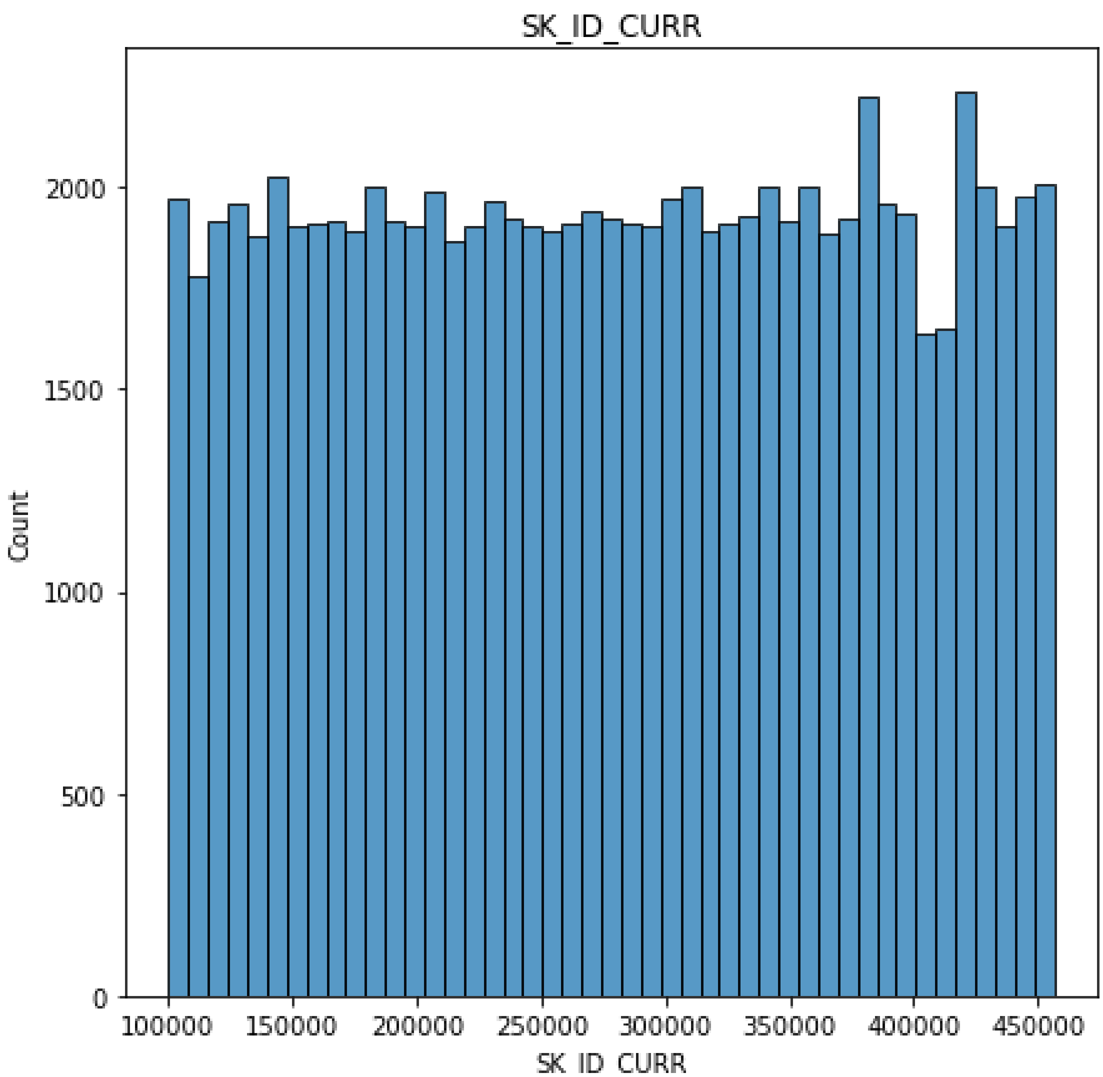
[58]:
```
df2.columns
```

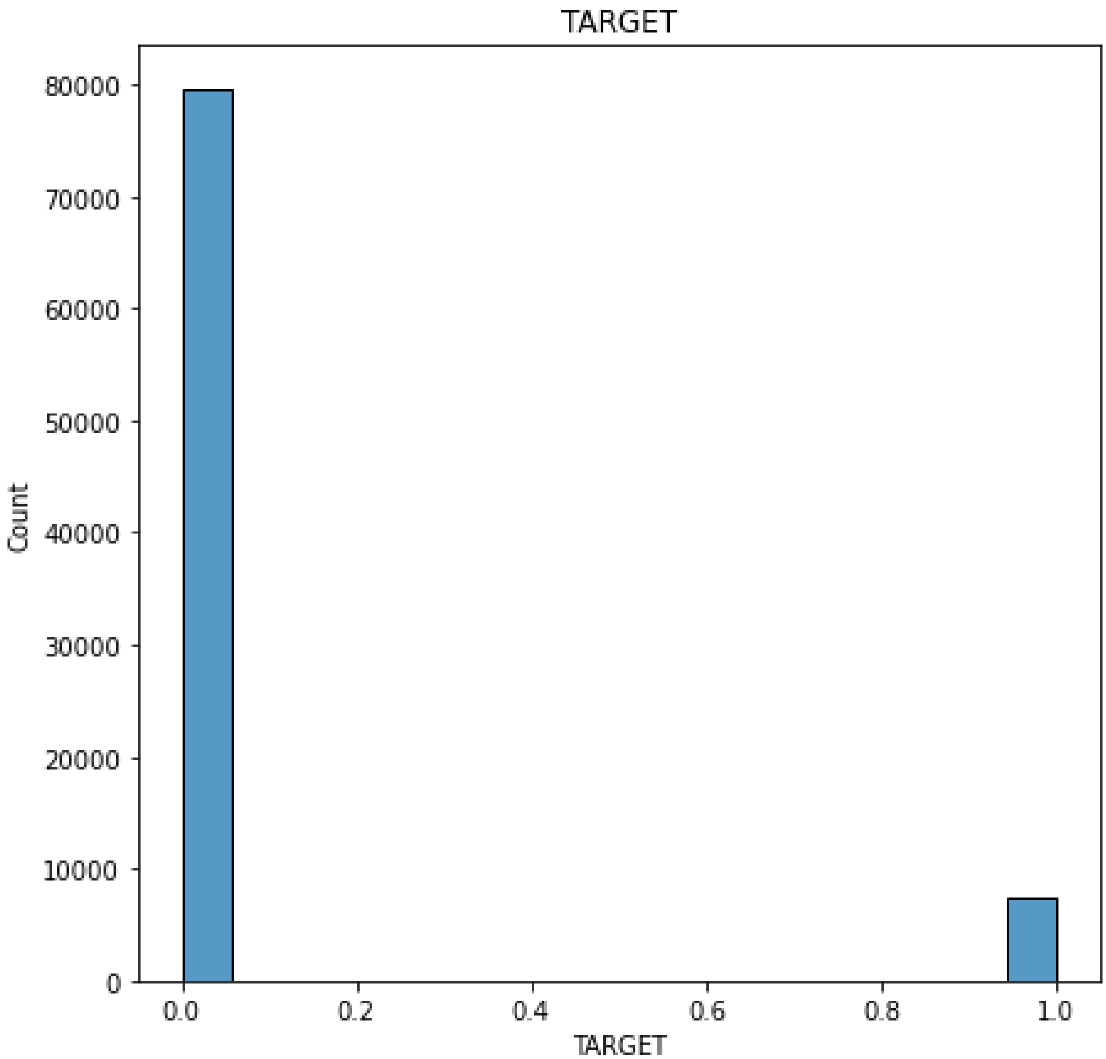
[58]:
```
Index(['SK_ID_CURR', 'TARGET', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT',
       'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS', 'CODE_REJECT_REASON',
       'NAME_CLIENT_TYPE'],
      dtype='object')
```

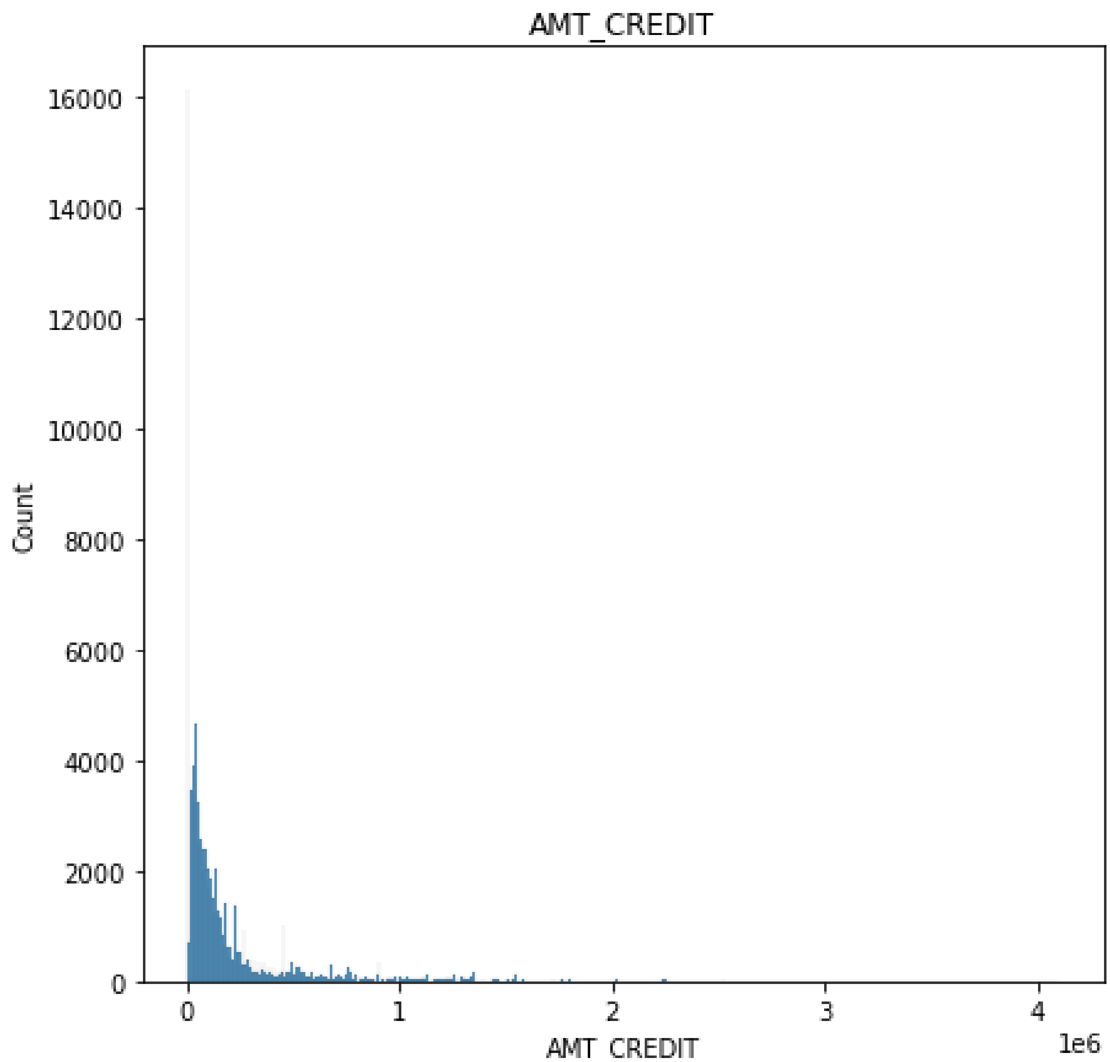
[59]:
```
categorical2=['TARGET','NAME_CASH_LOAN_PURPOSE','NAME_CONTRACT_STATUS','CODE_REJECT_REASON','N
continuous2=['SK_ID_CURR', 'TARGET', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT']
```

UNIVARIATE ANALYSIS OF CONTUNUOUS VARIABLES

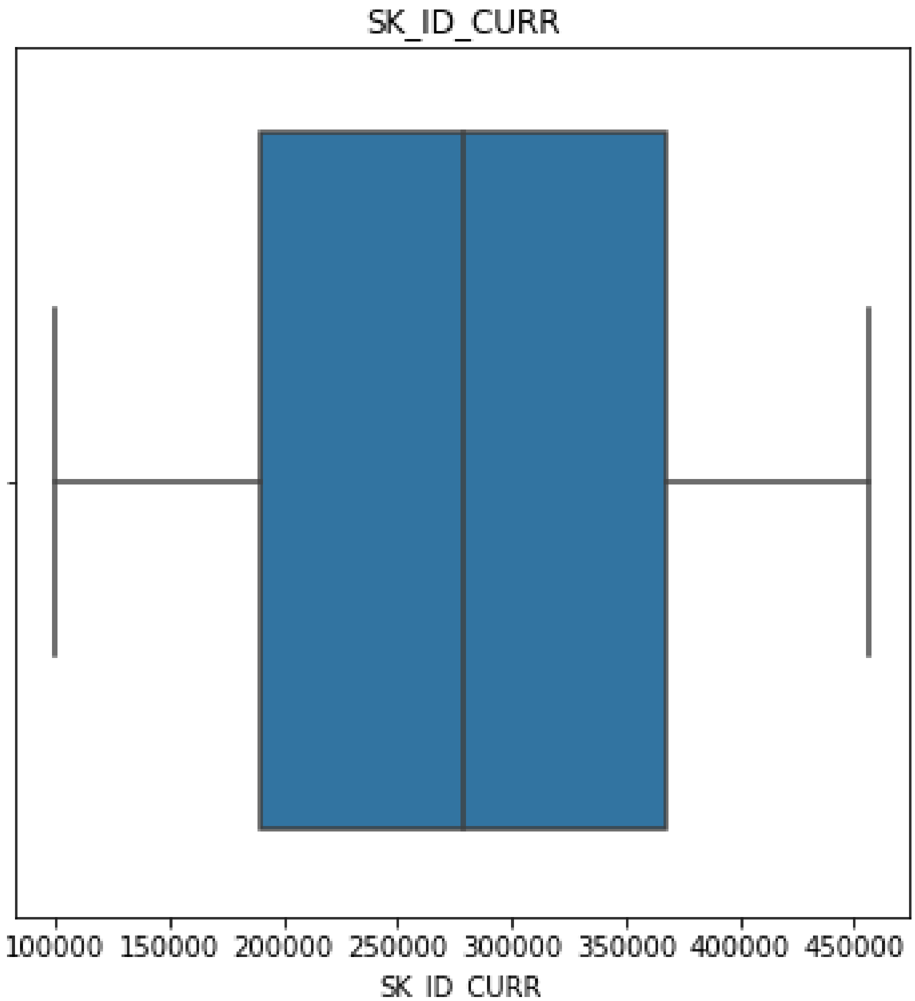
[ ]:
```
for col in continuous2:
    plt.figure(figsize=[7,7])

    sns.histplot(df2[col])
    plt.title(col)
    plt.show()
```

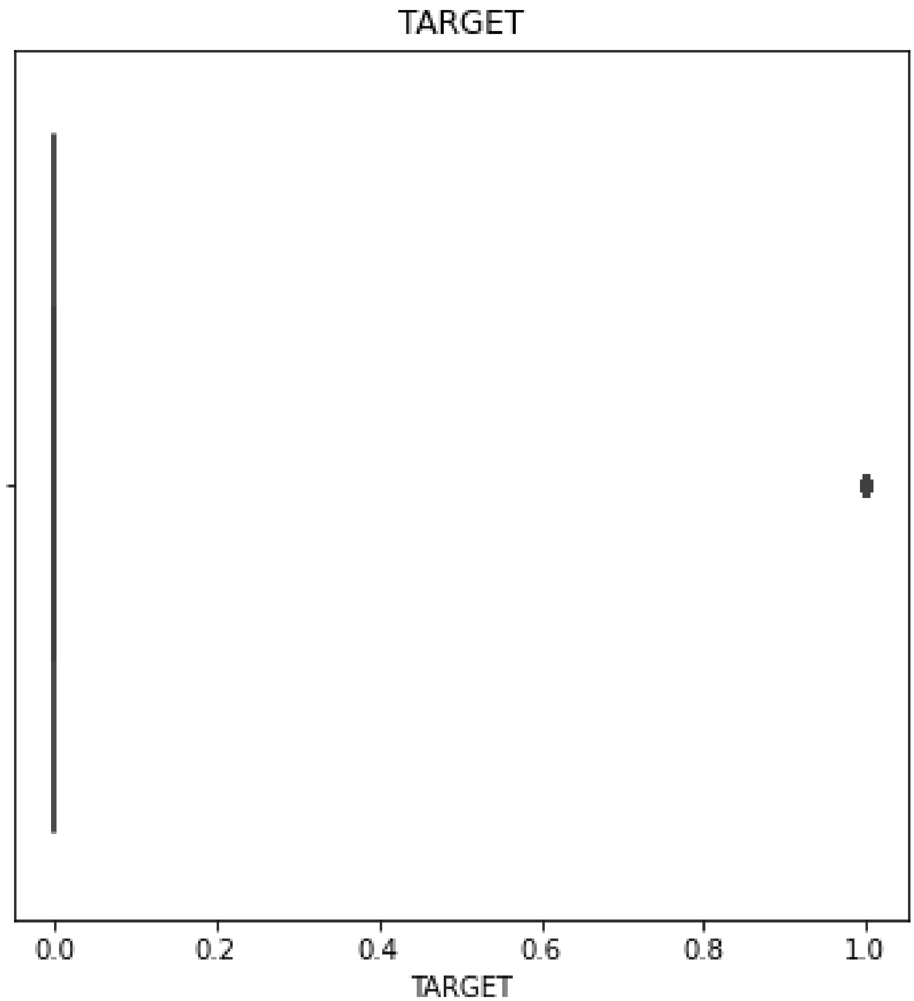


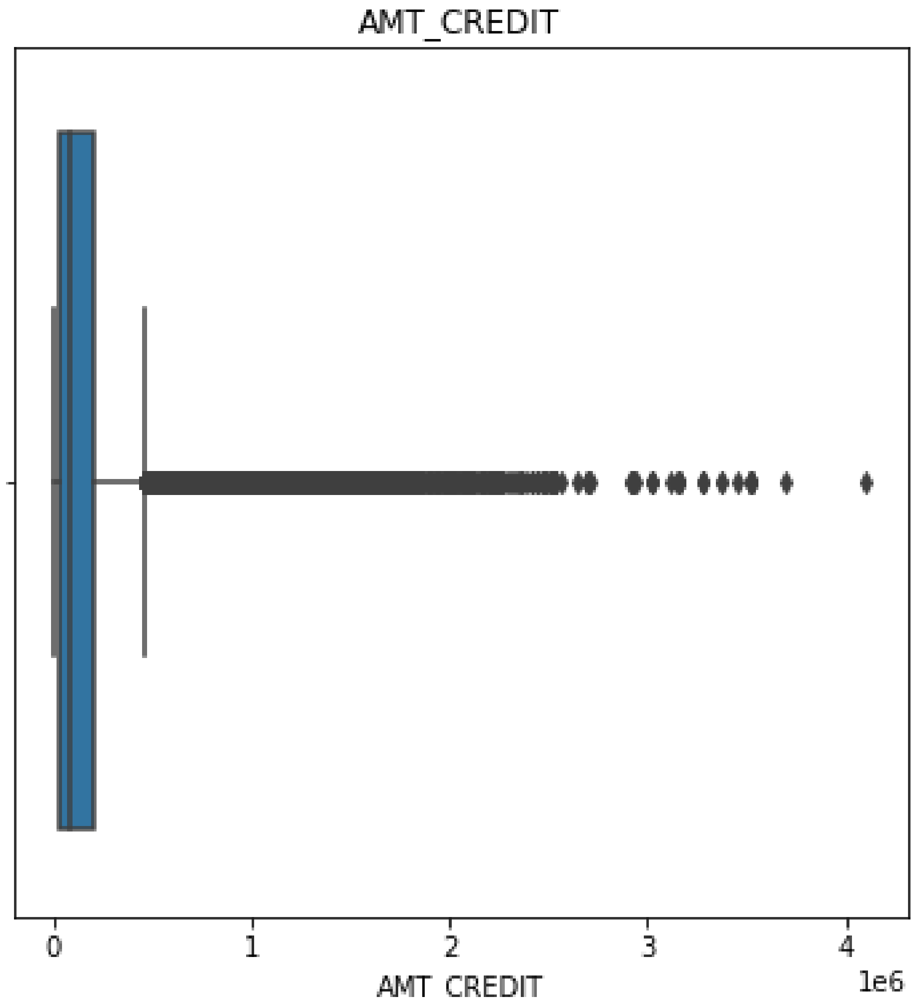

TARGET
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
TARGET

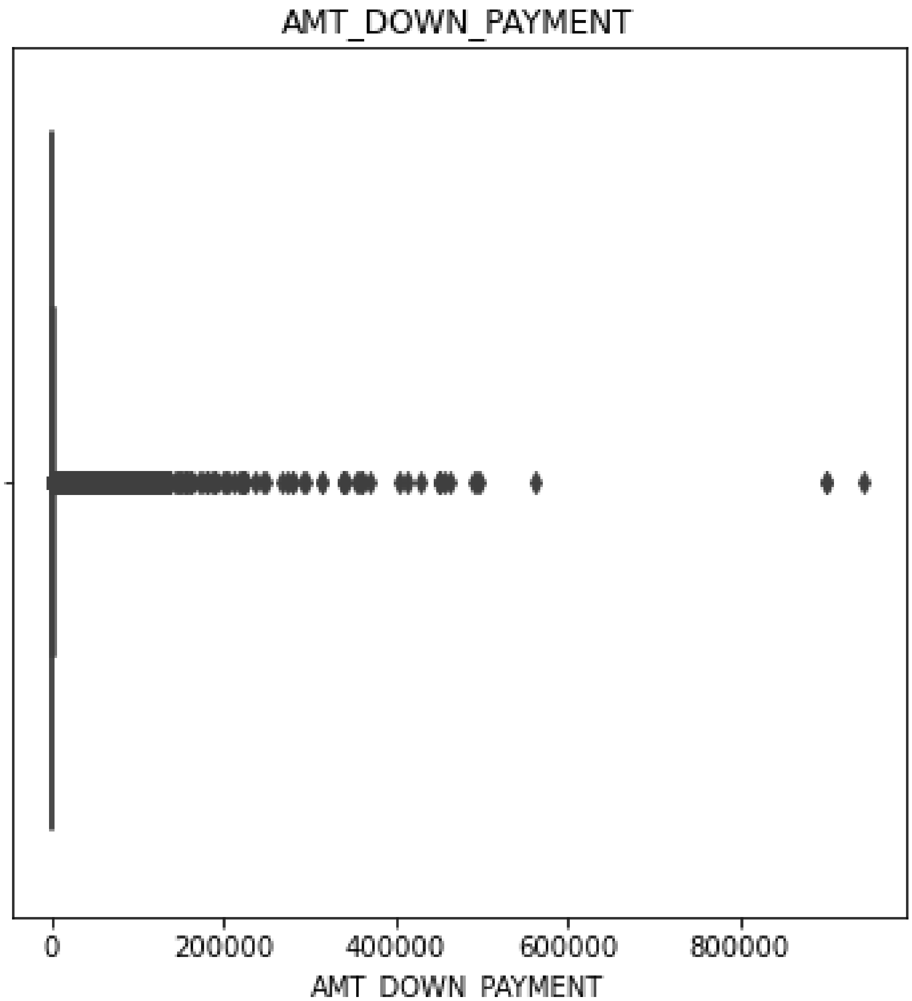


[60]:
```
for col in continuous2:
    plt.figure(figsize=[6,6])

    sns.boxplot(df2[col])
    plt.title(col)
    plt.show()
```


SK_ID_CURR
100000
150000
200000
250000
300000
350000
400000
450000
SK_ID_CURR

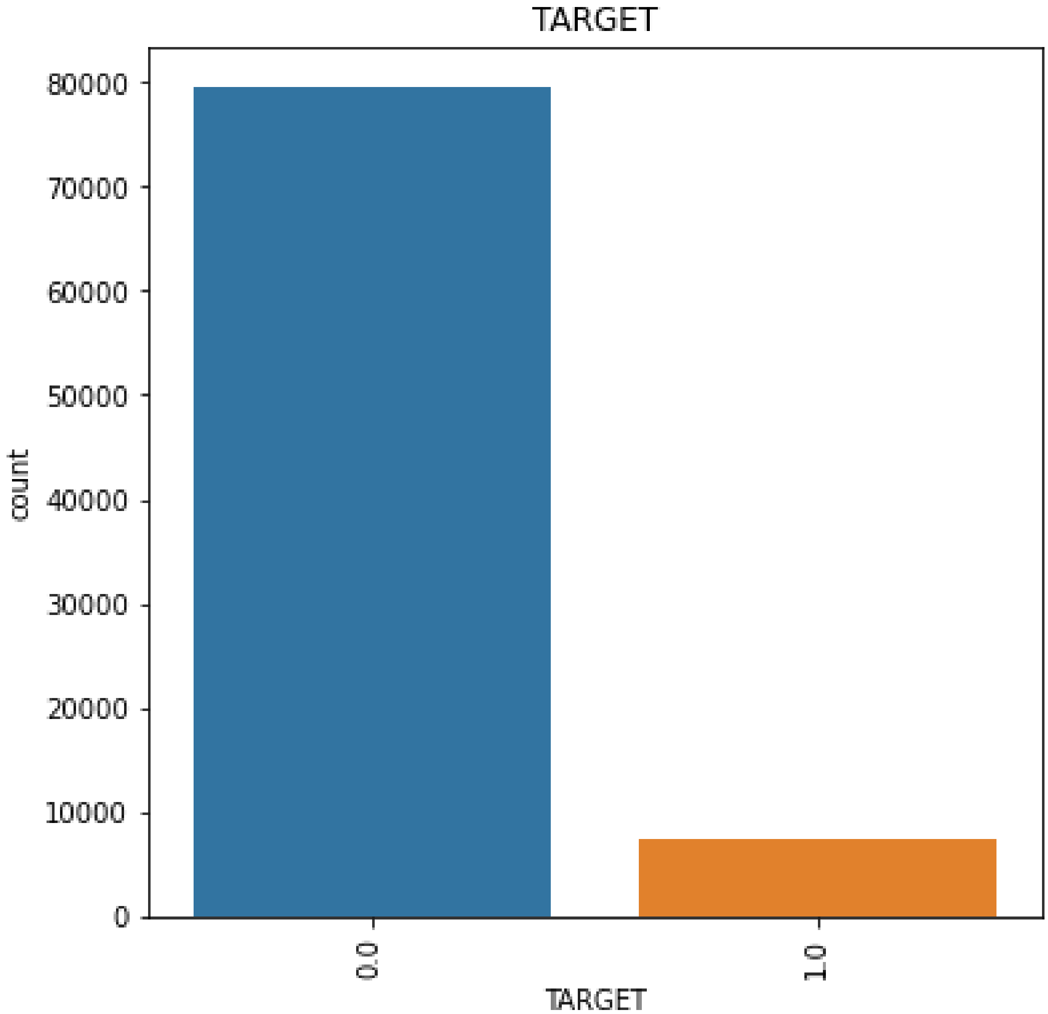

TARGET
0.0
0.2
0.4
0.6
0.8
1.0
TARGET

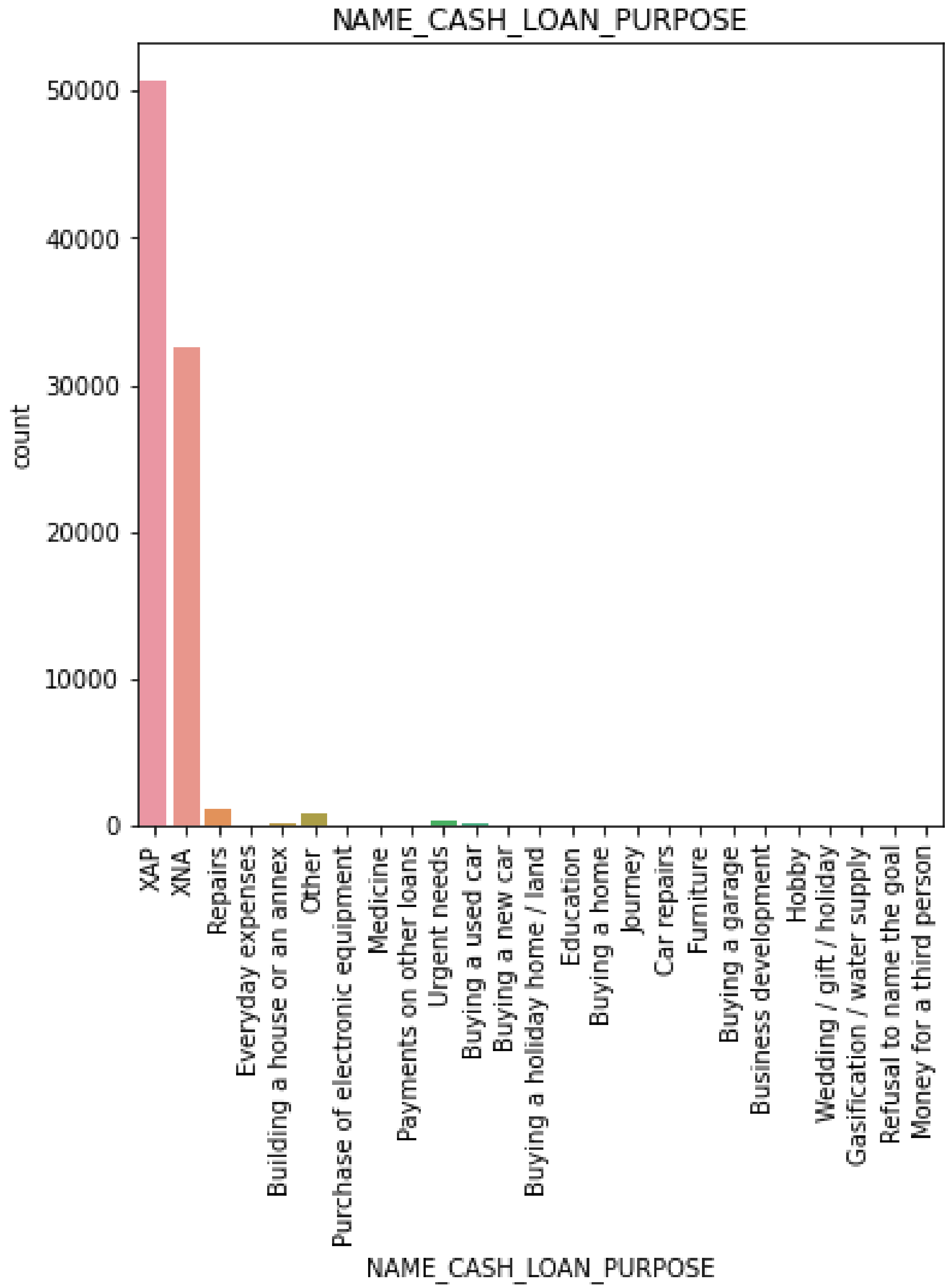

AMT_CREDIT
0
1
2
3
4
1e6
AMT_CREDIT

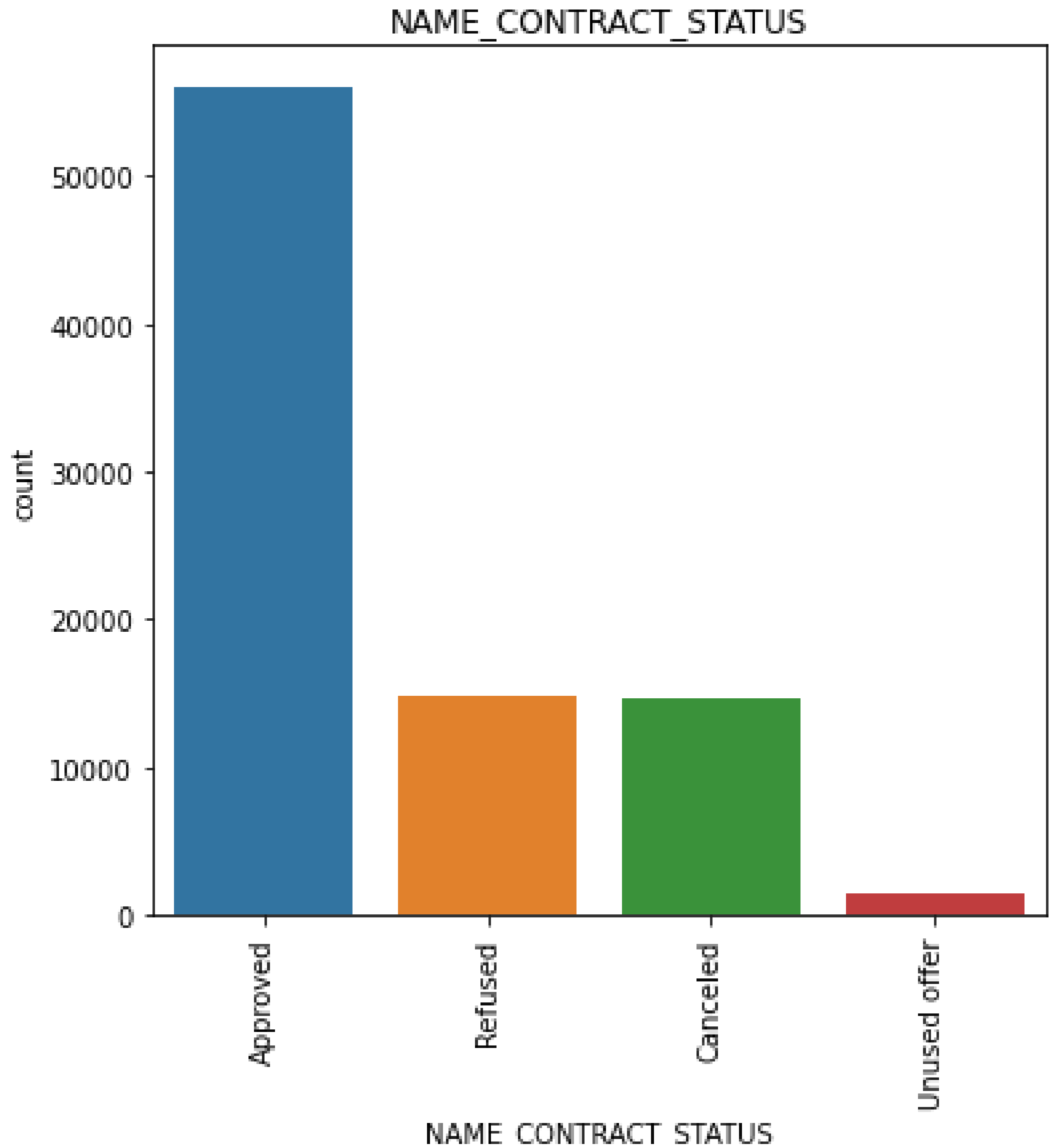


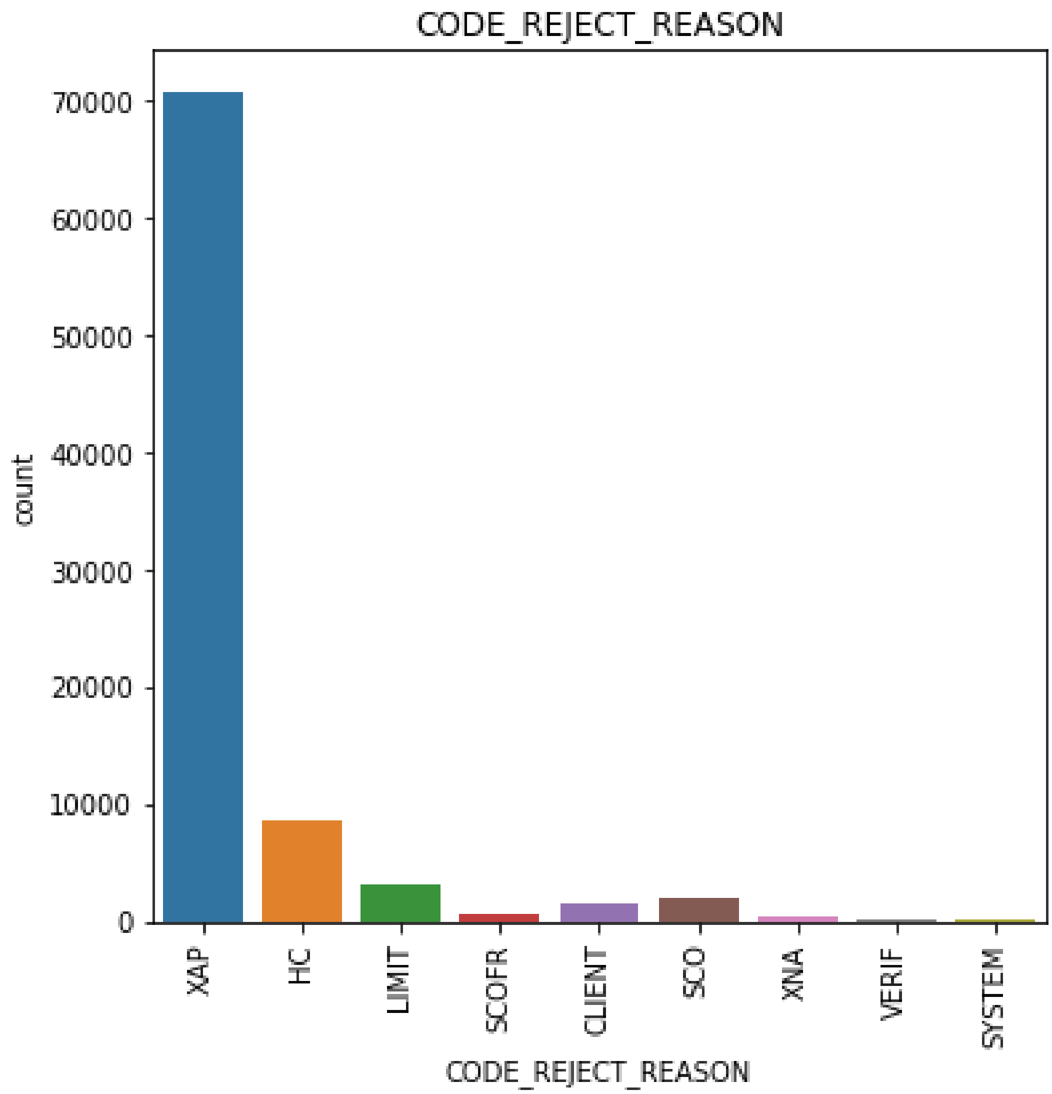
UNIVARIATE ANALYSIS OF CATEGORICAL VARIABLES

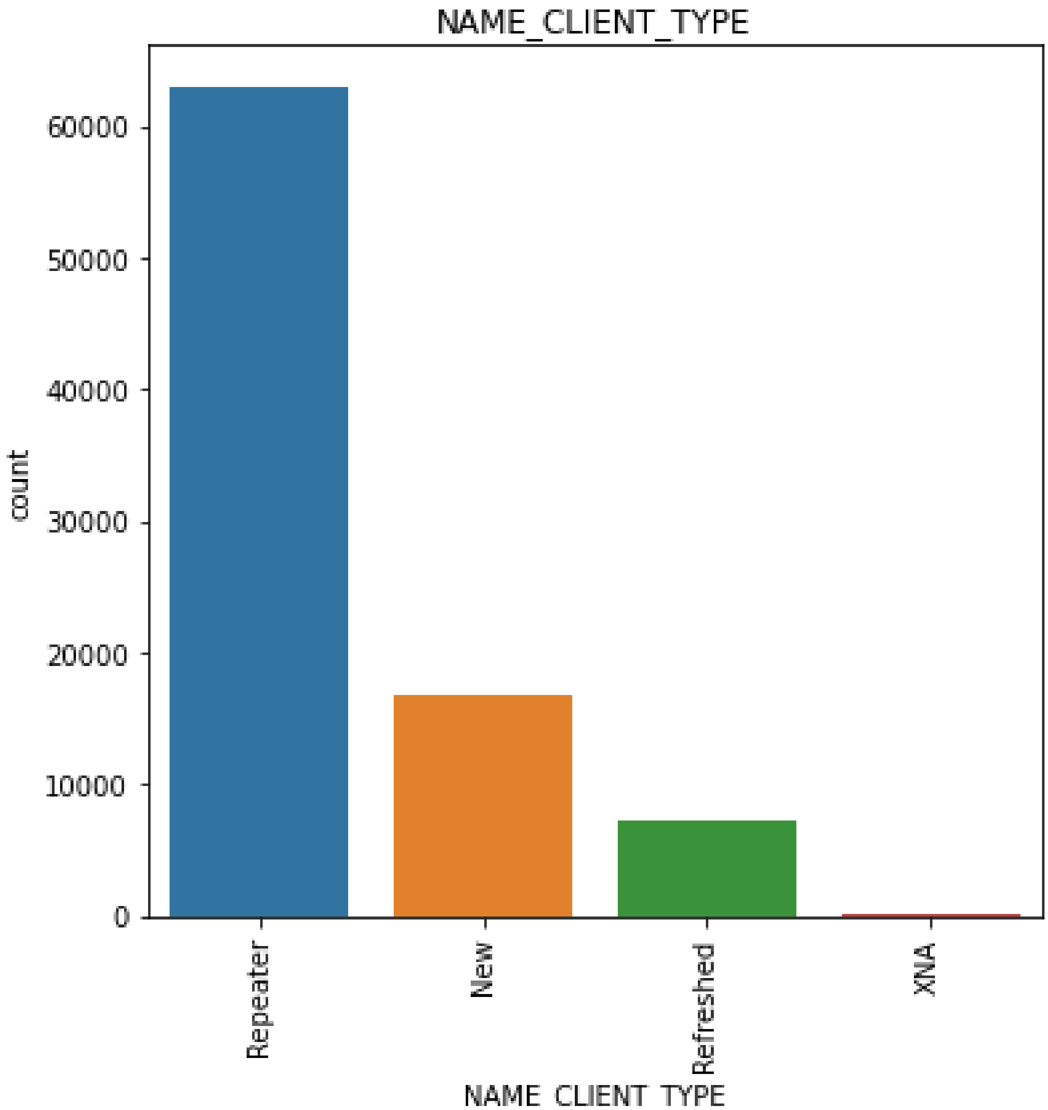
[61]:
```
for col in categorical2:
    plt.figure(figsize=[6,6])

    sns.countplot(x=df2[col])
    plt.title(col)
    plt.xticks(rotation=90)
    plt.show()
```


TARGET
80000
70000
60000
50000
40000
30000
20000
10000
0
count
0.0
1.0
TARGET


NAME_CASH_LOAN_PURPOSE
count
0
10000
20000
30000
40000
50000
XAP
XNA
Repairs
Everyday expenses
Building a house or an annex
Other
Purchase of electronic equipment
Medicine
Payments on other loans
Urgent needs
Buying a used car
Buying a new car
Buying a holiday home / land
Education
Buying a home
Journey
Car repairs
Furniture
Buying a garage
Business development
Hobby
Wedding / gift / holiday
Gasification / water supply
Refusal to name the goal
Money for a third person
NAME_CASH_LOAN_PURPOSE


NAME_CONTRACT_STATUS
count
50000
40000
30000
20000
10000
0
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS

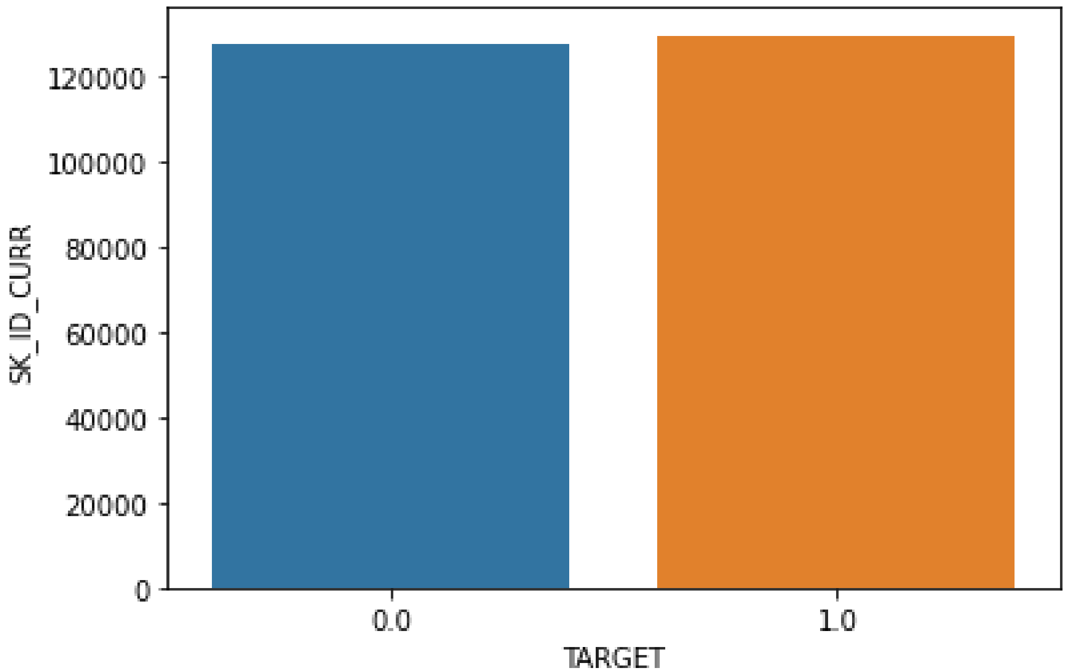

CODE_REJECT_REASON
count
70000
60000
50000
40000
30000
20000
10000
0
XAP
HC
LIMIT
SCOFR
CLIENT
SCO
XNA
VERIF
SYSTEM
CODE_REJECT_REASON



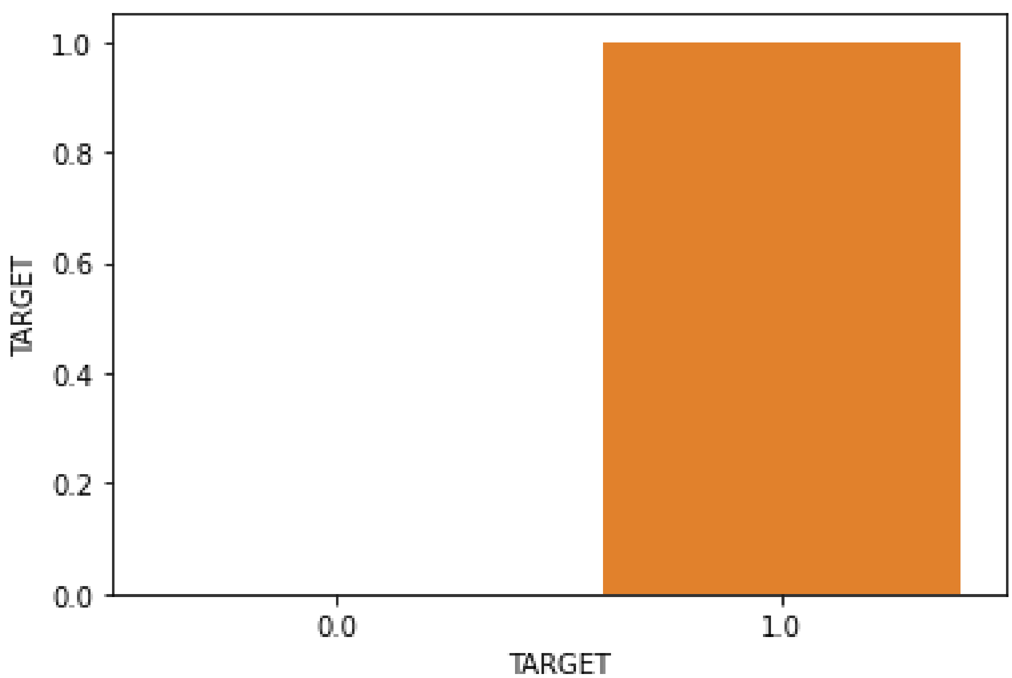
0.1.2 Bivariate Analysis

Categorical Vs Continuous

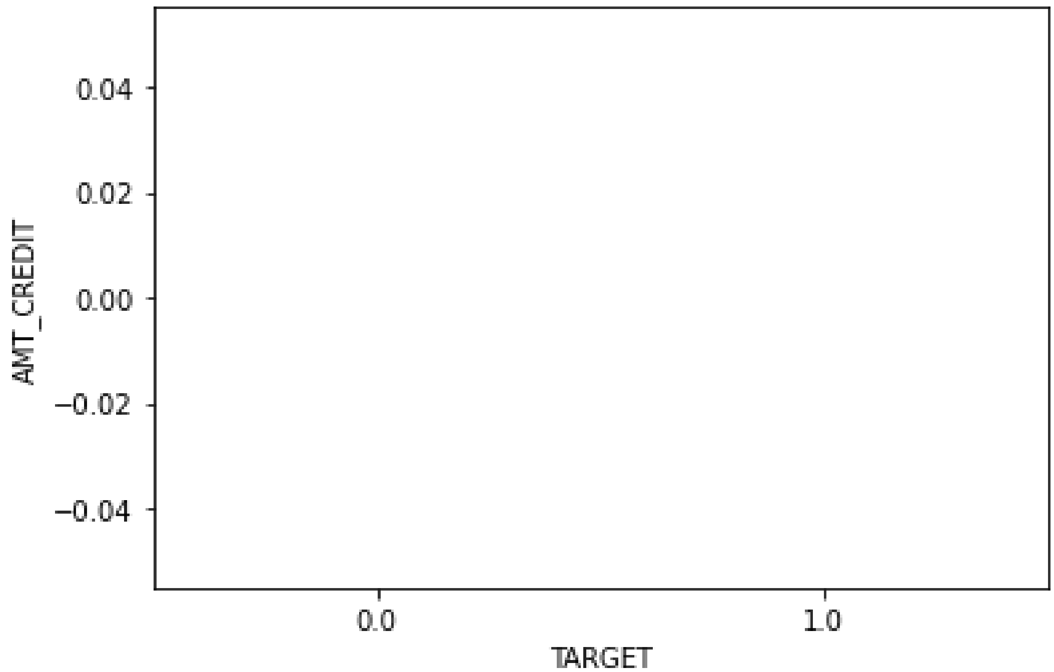
[62]:
```
for col in continuous2:
    sns.barplot(x=df2['TARGET'],y=df2[col],ci=None,estimator=lambda x:np.
 ↪quantile(x,0.075))
    plt.show()
```

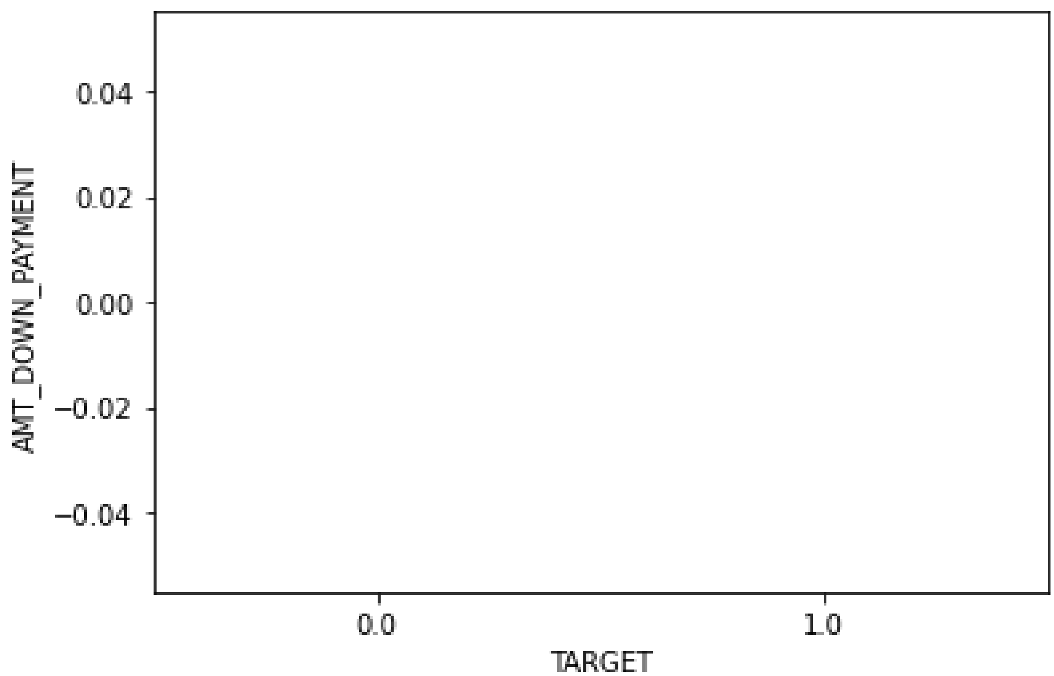


120000
100000
80000
60000
40000
20000
0
SK_ID_CURR
0.0
1.0
TARGET



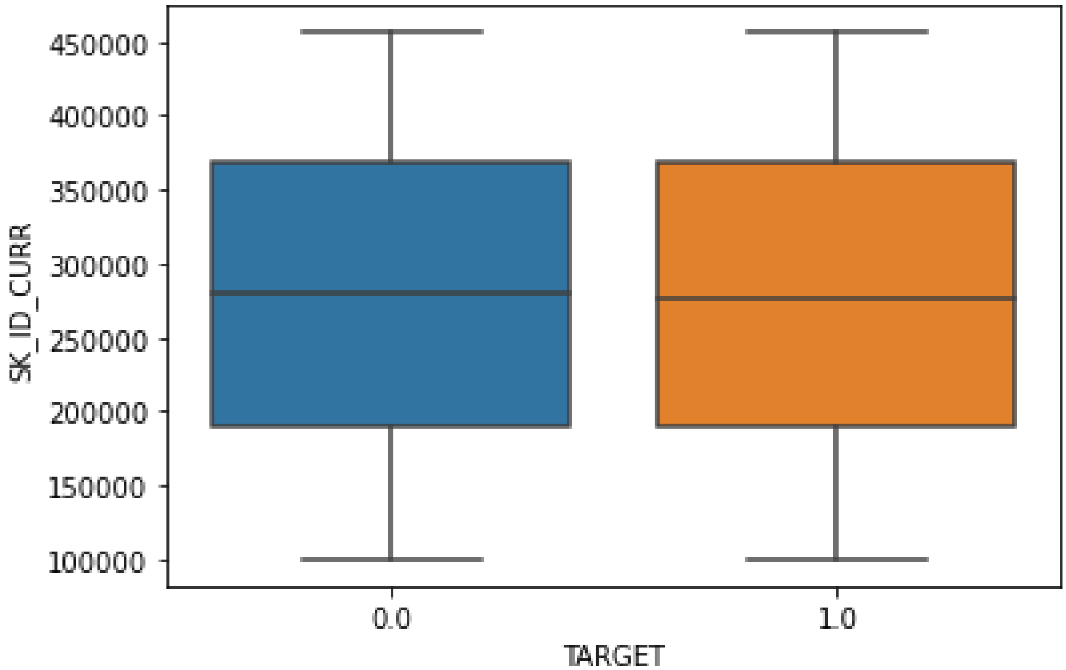
1.0
0.8
0.6
0.4
0.2
0.0
TARGET
0.0
1.0
TARGET



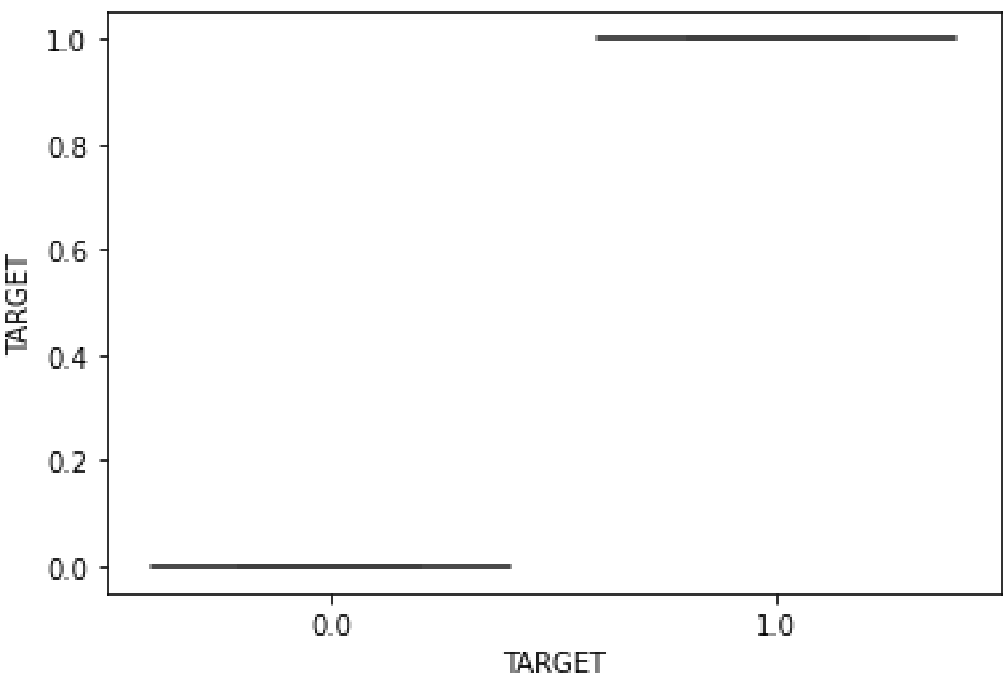


[63]:
```
for col in continuous2:
    sns.boxplot(x=df2['TARGET'],y=df2[col])
```

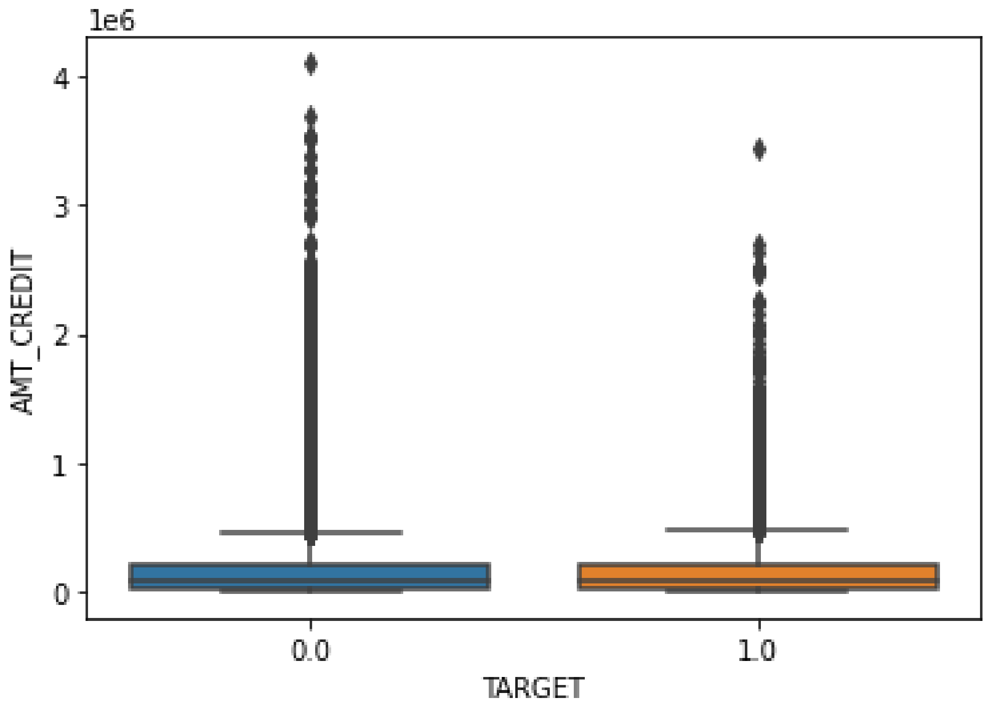
```
plt.show()
```

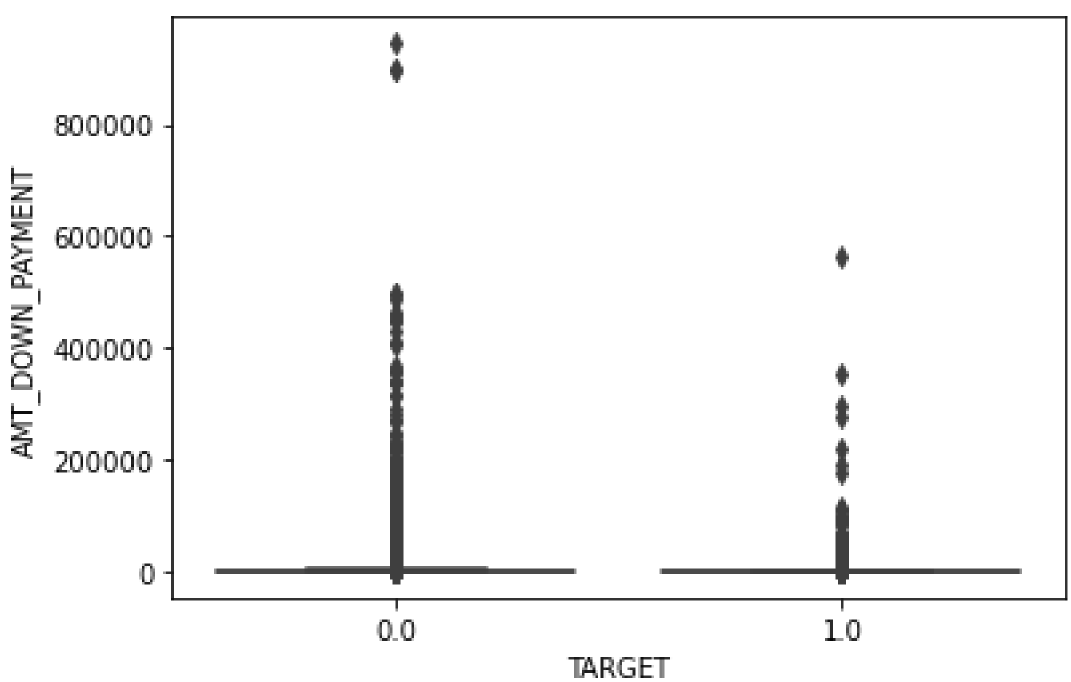


450000
400000
350000
300000
250000
200000
150000
100000
SK_ID_CURR
0.0
1.0
TARGET



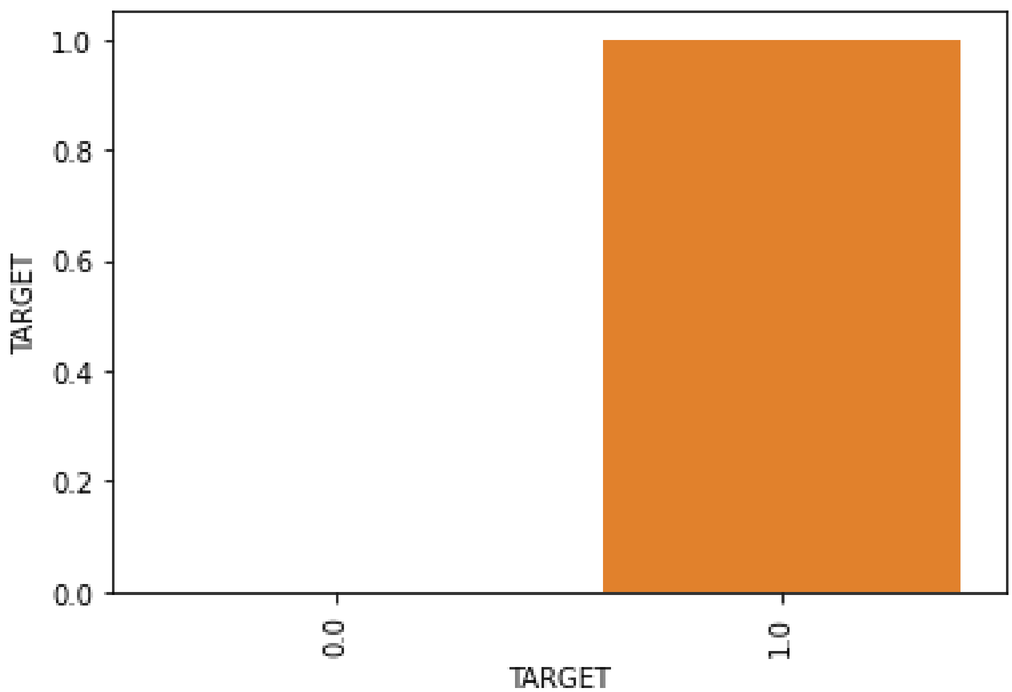
1.0
0.8
0.6
0.4
0.2
0.0
TARGET
0.0
1.0
TARGET

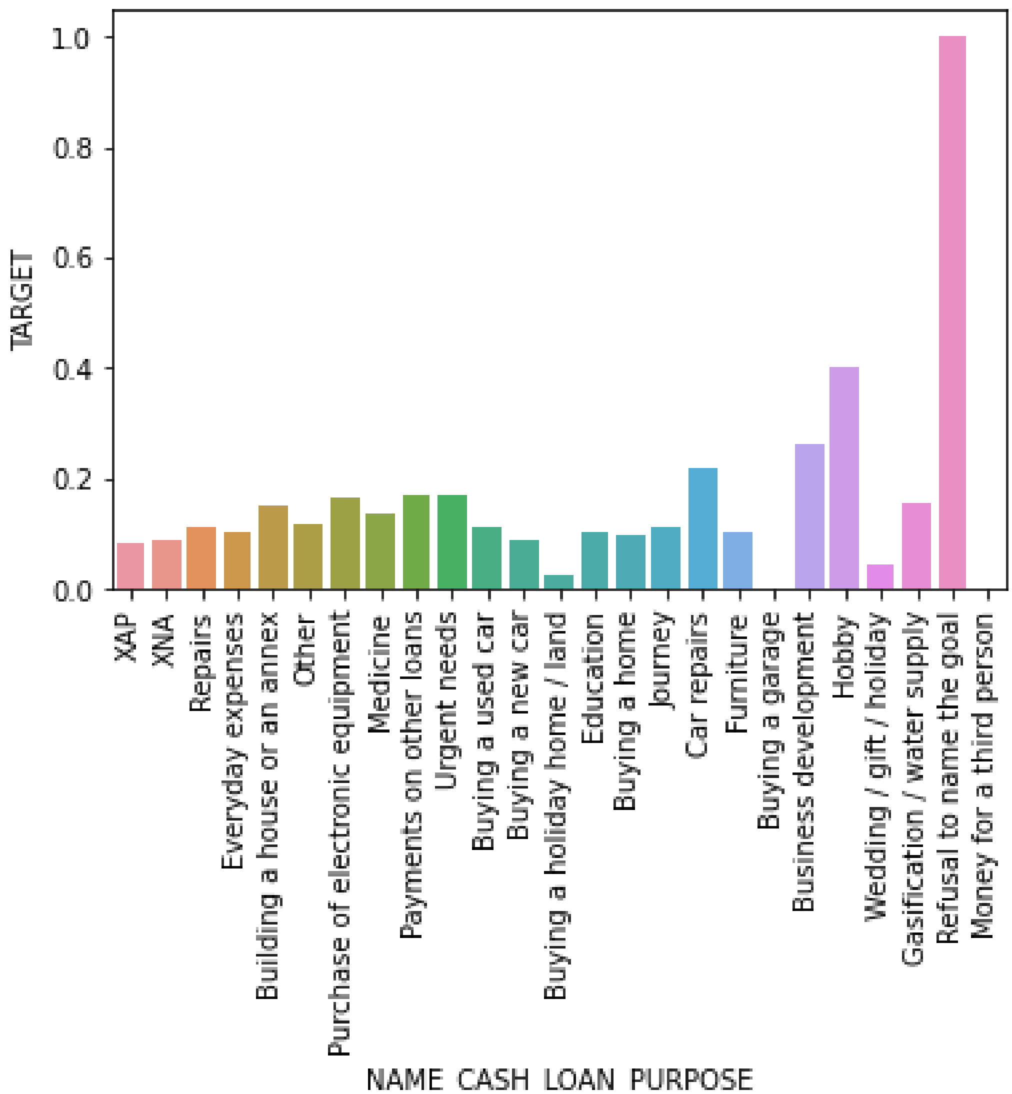


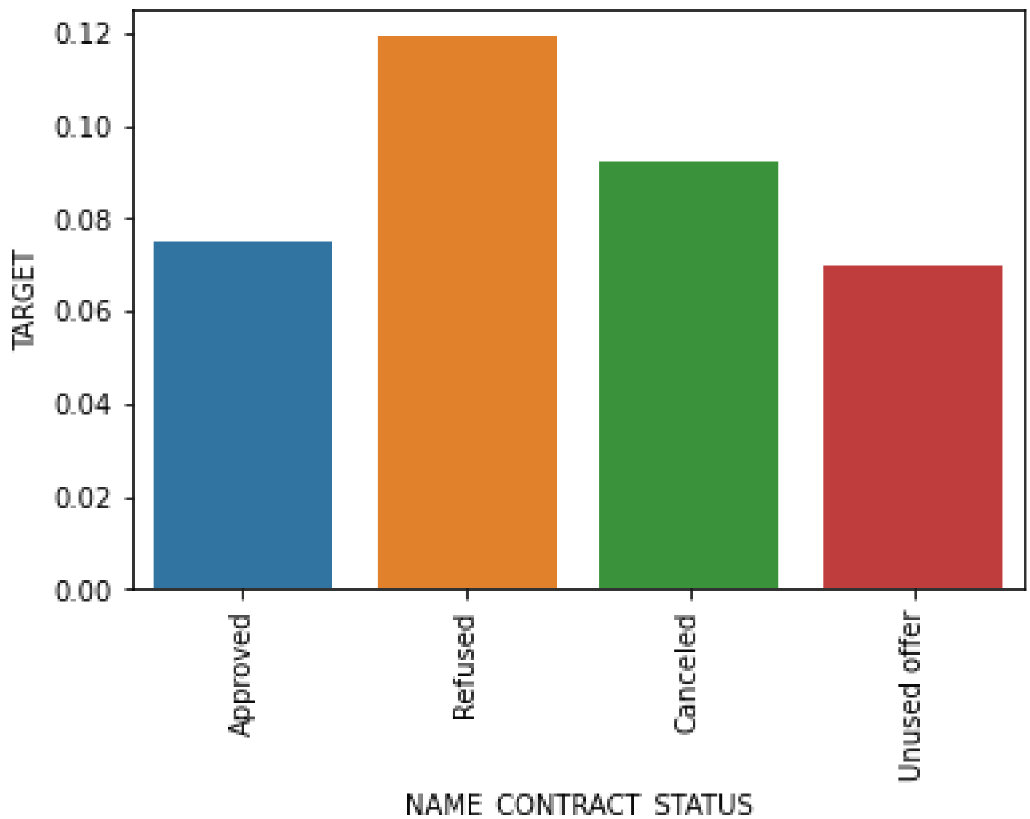


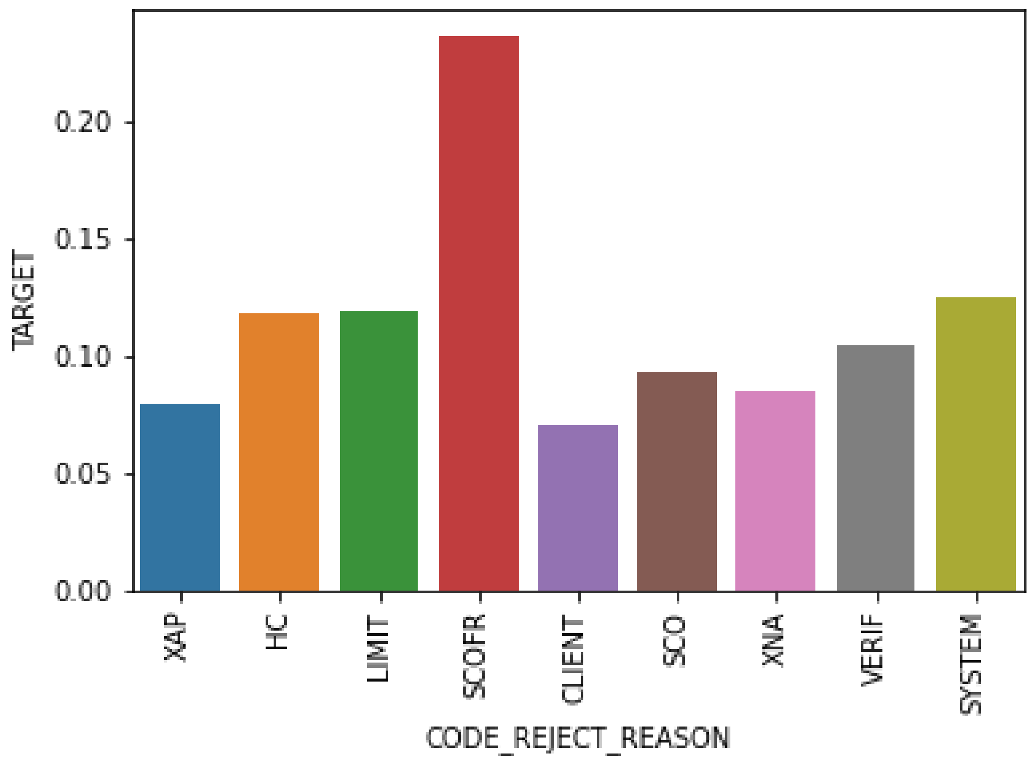
Categorical Vs Categorical

[64]:
```
for col in categorical2:
    sns.barplot(y=df2['TARGET'],x=df2[col],ci=None)
    plt.xticks(rotation=90)
    plt.show()
```

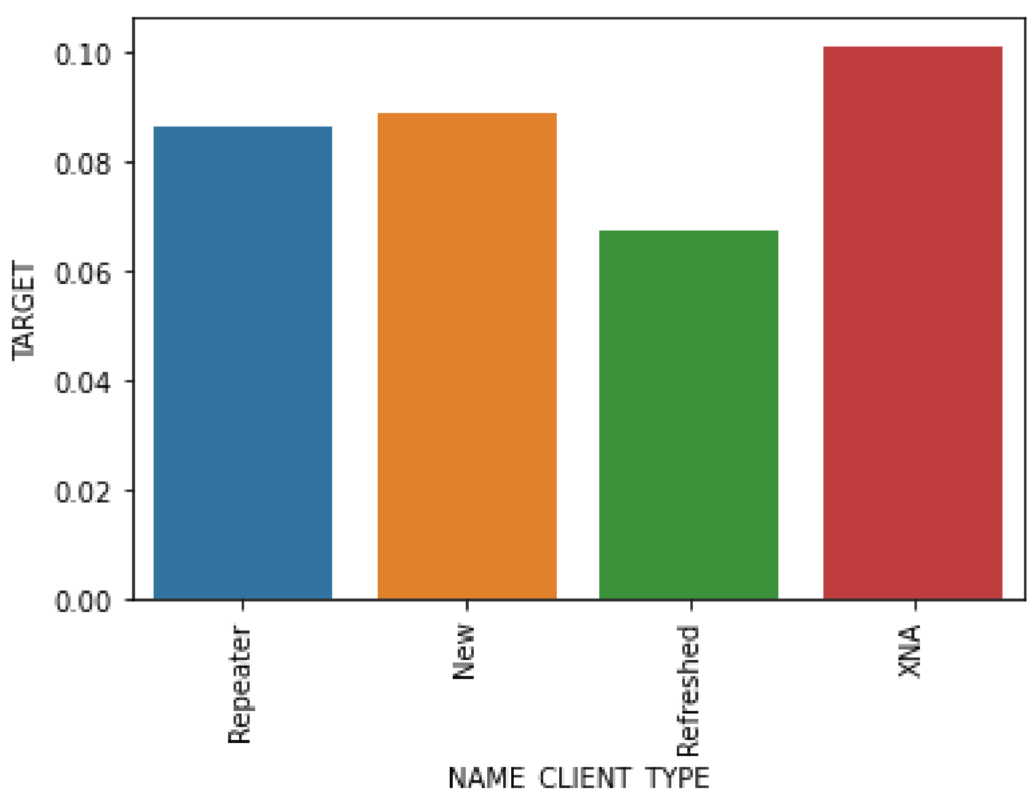



TARGET
1.0
0.8
0.6
0.4
0.2
0.0
XAP
XNA
Repairs
Everyday expenses
Building a house or an annex
Other
Purchase of electronic equipment
Medicine
Payments on other loans
Urgent needs
Buying a used car
Buying a new car
Buying a holiday home / land
Education
Buying a home
Journey
Car repairs
Furniture
Buying a garage
Business development
Hobby
Wedding / gift / holiday
Gasification / water supply
Refusal to name the goal
Money for a third person
NAME_CASH_LOAN_PURPOSE


0.12
0.10
0.08
0.06
0.04
0.02
0.00
TARGET
Approved
Refused
Canceled
Unused offer
NAME_CONTRACT_STATUS



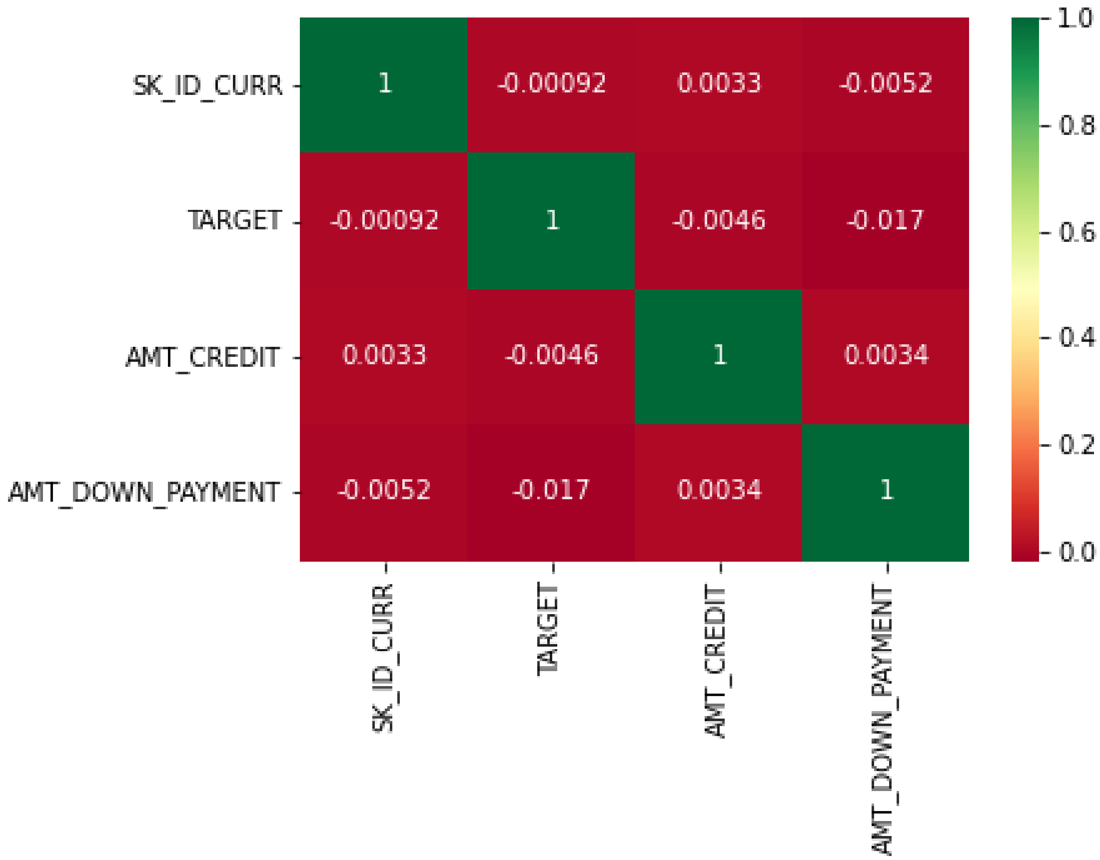
0.20
0.15
0.10
0.05
0.00
TARGET
XAP
HC
LIMIT
SCOFR
CLIENT
SCO
XNA
VERIF
SYSTEM
CODE_REJECT_REASON



0.10
0.08
0.06
0.04
0.02
0.00
TARGET
Repeater
New
Refreshed
XNA
NAME_CLIENT_TYPE

0.2 Multivariate

[65]:
```
sns.heatmap(df2[continuous2].corr(),annot=True,cmap="RdYlGn");
```



THERE ARE DATA IMBALANCE IN CASE OF TARGET VARIABLE.

[93]:
```
df1['TARGET'].value_counts()
```

[93]:
```
0    282672
1     24825
Name: TARGET, dtype: int64
```

[98]:
```
x=282672
y=24825
print("ratio is",x/y,":", y/y )
```

```
ratio is 11.386586102719033 : 1.0
```

[94]:
```
df2['TARGET'].value_counts()
```

[94]:
```
0.0    79507
1.0     7428
Name: TARGET, dtype: int64
```

[99]:
```
x=79507
y=7428
print("ratio is",x/y,":", y/y )
```

```
ratio is 10.7036887452881 : 1.0
```

[ ]:
```

```