In [62]:
```
# importing libraries
#import the warnings
import warnings
warnings.filterwarnings("ignore")
```

In [63]:
```
#import the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
pd.set_option("display.max_columns", None)
```

In [64]:
```
#read application csv
app_data = pd.read_csv("Application_data.csv")
app_data.head()
```

Out[64]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_
0	100002	1	Cash loans	M	N	Y	0	2
1	100003	0	Cash loans	F	N	N	0	2
2	100004	0	Revolving loans	M	Y	Y	0	
3	100006	0	Cash loans	F	N	Y	0	1
4	100007	0	Cash loans	M	N	Y	0	1

In [65]:
```
# Data Inspection on Application dataset
### get info and shape on dataset
app_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

In [66]:
```
## Data Quality Check
###check for percentage null values in application dataset
pd.set_option('display.max_rows' , 200)
app_data.isnull().mean()  * 100
```

Out[66]:

```
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
```

-Conclusion: columns with null values more than 47% may give wrong insights, hence will drop them

In [67]:
```
### Dropping columns with missing values greater than 47%
```

```
percentage = 47
threshold = int(((100 - percentage)/100)*app_data.shape[0] + 1)
app_df = app_data.dropna(axis=1, thresh=threshold)
app_df.head()
```

Out[67]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_
0	100002	1	Cash loans	M	N	Y	0	2
1	100003	0	Cash loans	F	N	N	0	2
2	100004	0	Revolving loans	M	Y	Y	0	
3	100006	0	Cash loans	F	N	Y	0	1
4	100007	0	Cash loans	M	N	Y	0	1

In [68]:
```
app_df.shape
```

Out[68]:
```
(307511, 73)
```

In [69]:
```
app_df.isnull().mean() * 100
```

Out[69]:

```
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
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_2                   0.214626
EXT_SOURCE_3                  19.825307
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
```

In [70]:

```
### Impute missing values
### Check missing values in application dataset before imputing
app_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 73 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  AMT_GOODS_PRICE              307233 non-null  float64
 11  NAME_TYPE_SUITE              306219 non-null  object
 12  NAME_INCOME_TYPE             307511 non-null  object
 13  NAME_EDUCATION_TYPE          307511 non-null  object
 14  NAME_FAMILY_STATUS           307511 non-null  object
 15  NAME_HOUSING_TYPE            307511 non-null  object
 16  REGION_POPULATION_RELATIVE   307511 non-null  float64
 17  DAYS_BIRTH                   307511 non-null  int64
 18  DAYS_EMPLOYED                307511 non-null  int64
 19  DAYS_REGISTRATION            307511 non-null  float64
 20  DAYS_ID_PUBLISH              307511 non-null  int64
 21  FLAG_MOBIL                   307511 non-null  int64
 22  FLAG_EMP_PHONE               307511 non-null  int64
 23  FLAG_WORK_PHONE              307511 non-null  int64
 24  FLAG_CONT_MOBILE             307511 non-null  int64
 25  FLAG_PHONE                   307511 non-null  int64
 26  FLAG_EMAIL                   307511 non-null  int64
 27  OCCUPATION_TYPE              211120 non-null  object
 28  CNT_FAM_MEMBERS              307509 non-null  float64
 29  REGION_RATING_CLIENT         307511 non-null  int64
 30  REGION_RATING_CLIENT_W_CITY  307511 non-null  int64
 31  WEEKDAY_APPR_PROCESS_START   307511 non-null  object
 32  HOUR_APPR_PROCESS_START      307511 non-null  int64
 33  REG_REGION_NOT_LIVE_REGION   307511 non-null  int64
 34  REG_REGION_NOT_WORK_REGION   307511 non-null  int64
 35  LIVE_REGION_NOT_WORK_REGION  307511 non-null  int64
 36  REG_CITY_NOT_LIVE_CITY       307511 non-null  int64
 37  REG_CITY_NOT_WORK_CITY       307511 non-null  int64
 38  LIVE_CITY_NOT_WORK_CITY      307511 non-null  int64
 39  ORGANIZATION_TYPE            307511 non-null  object
 40  EXT_SOURCE_2                 306851 non-null  float64
 41  EXT_SOURCE_3                 246546 non-null  float64
 42  OBS_30_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 43  DEF_30_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 44  OBS_60_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 45  DEF_60_CNT_SOCIAL_CIRCLE     306490 non-null  float64
 46  DAYS_LAST_PHONE_CHANGE       307510 non-null  float64
 47  FLAG_DOCUMENT_2              307511 non-null  int64
 48  FLAG_DOCUMENT_3              307511 non-null  int64
 49  FLAG_DOCUMENT_4              307511 non-null  int64
 50  FLAG_DOCUMENT_5              307511 non-null  int64
 51  FLAG_DOCUMENT_6              307511 non-null  int64
 52  FLAG_DOCUMENT_7              307511 non-null  int64
 53  FLAG_DOCUMENT_8              307511 non-null  int64
 54  FLAG_DOCUMENT_9              307511 non-null  int64
 55  FLAG_DOCUMENT_10             307511 non-null  int64
 56  FLAG_DOCUMENT_11             307511 non-null  int64
 57  FLAG_DOCUMENT_12             307511 non-null  int64
 58  FLAG_DOCUMENT_13             307511 non-null  int64
 59  FLAG_DOCUMENT_14             307511 non-null  int64
 60  FLAG_DOCUMENT_15             307511 non-null  int64
 61  FLAG_DOCUMENT_16             307511 non-null  int64
 62  FLAG_DOCUMENT_17             307511 non-null  int64
 63  FLAG_DOCUMENT_18             307511 non-null  int64
 64  FLAG_DOCUMENT_19             307511 non-null  int64
 65  FLAG_DOCUMENT_20             307511 non-null  int64
 66  FLAG_DOCUMENT_21             307511 non-null  int64
 67  AMT_REQ_CREDIT_BUREAU_HOUR   265992 non-null  float64
 68  AMT_REQ_CREDIT_BUREAU_DAY    265992 non-null  float64
 69  AMT_REQ_CREDIT_BUREAU_WEEK   265992 non-null  float64
 70  AMT_REQ_CREDIT_BUREAU_MON    265992 non-null  float64
 71  AMT_REQ_CREDIT_BUREAU_QRT    265992 non-null  float64
 72  AMT_REQ_CREDIT_BUREAU_YEAR   265992 non-null  float64
dtypes: float64(20), int64(41), object(12)
memory usage: 171.3+ MB
```

OCCUPATION_TYPE column has 31% missing values, since its a categorical column,imputing the missing values with a unknown or others value

In [71]:
```
app_df.OCCUPATION_TYPE.isnull().mean() * 100
```

Out[71]:
```
31.345545362604916
```

In [72]:
```
app_df.OCCUPATION_TYPE.value_counts(normalize=True)*100
```

Out[72]:
```
Laborers                 26.139636
Sales staff              15.205570
Core staff               13.058924
Managers                 10.122679
Drivers                   8.811576
High skill tech staff     5.390299
Accountants               4.648067
Medicine staff            4.043672
Security staff            3.183498
Cooking staff             2.816408
Cleaning staff            2.203960
Private service staff     1.256158
Low-skill Laborers        0.991379
Waiters/barmen staff      0.638499
Secretaries               0.618132
Realty agents             0.355722
HR staff                  0.266673
IT staff                  0.249147
Name: OCCUPATION_TYPE, dtype: float64
```

In [73]:
```
app_df.OCCUPATION_TYPE.fillna("Others", inplace = True)
```

In [74]:
```
app_df.OCCUPATION_TYPE.isnull().mean() *100
```

Out[74]:
```
0.0
```

In [75]:
```
app_df.OCCUPATION_TYPE.value_counts(normalize=True)*100
```

Out[75]:
```
Others                   31.345545
Laborers                 17.946025
Sales staff              10.439301
Core staff                8.965533
Managers                  6.949670
Drivers                   6.049540
High skill tech staff     3.700681
Accountants               3.191105
Medicine staff            2.776161
Security staff            2.185613
Cooking staff             1.933589
Cleaning staff            1.513117
Private service staff     0.862408
Low-skill Laborers        0.680626
Waiters/barmen staff      0.438358
Secretaries               0.424375
Realty agents             0.244219
HR staff                  0.183083
IT staff                  0.171051
Name: OCCUPATION_TYPE, dtype: float64
```

EXT_SOURCE_3 columns has 19% missing values

In [76]:
```
app_df.EXT_SOURCE_3.isnull().mean() *100
```

Out[76]:
```
19.825307062186393
```

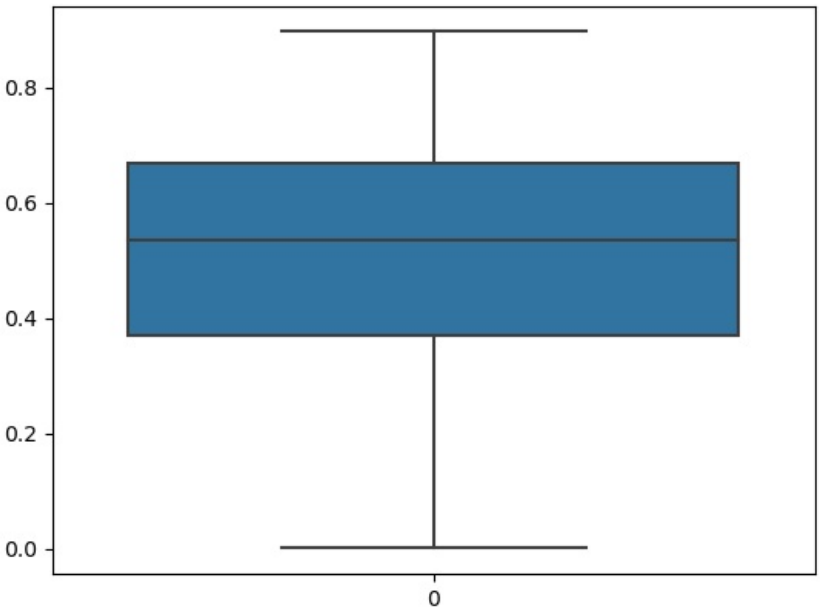
In [77]:
```
app_df.EXT_SOURCE_3.value_counts(normalize=True)*100
```

Out[77]:
```
0.746300    0.592182
0.713631    0.533369
0.694093    0.517550
0.670652    0.483074
0.652897    0.468067
              ...
0.021492    0.000406
0.019468    0.000406
0.023062    0.000406
0.014556    0.000406
0.043227    0.000406
Name: EXT_SOURCE_3, Length: 814, dtype: float64
```

In [78]:
```
app_df.EXT_SOURCE_3.describe()
```

Out[78]:
```
count    246546.000000
mean          0.510853
std           0.194844
min           0.000527
25%           0.370650
50%           0.535276
75%           0.669057
max           0.896010
Name: EXT_SOURCE_3, dtype: float64
```

In [79]:
```
sns.boxplot(app_df.EXT_SOURCE_3)
plt.show()
```



In [80]:
```
app_df.EXT_SOURCE_3.fillna(app_df.EXT_SOURCE_3.median(), inplace = True)
```

In [81]:
```
app_df.EXT_SOURCE_3.isnull().mean() *100
```

Out[81]:
```
0.0
```

In [82]:
```
app_df.EXT_SOURCE_3.value_counts(normalize=True)*100
```

Out[82]:
```
0.535276    20.080908
0.746300     0.474780
0.713631     0.427627
0.694093     0.414945
0.670652     0.387303
               ...
0.021492     0.000325
0.019468     0.000325
0.023062     0.000325
0.014556     0.000325
0.043227     0.000325
Name: EXT_SOURCE_3, Length: 814, dtype: float64
```

-conclusion: since its a numerical columns with no outliers and there is not much difference between mean and median. Hence we can impute with mean or median

In [83]:
```
null_cols = list(app_df.columns[app_df.isna().any()])
len(null_cols)
```

Out[83]:
```
16
```

In [23]:
```
app_df.isnull().mean()* 100
```

Out[23]:
```
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
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                0.000000
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_2                   0.214626
EXT_SOURCE_3                   0.000000
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
```

Handling missing values in columns with 13% null values

In [84]:
```
app_df.AMT_REQ_CREDIT_BUREAU_HOUR.value_counts(normalize=True)*100
```

Out[84]:
```
0.0    99.388703
1.0     0.586484
2.0     0.021053
3.0     0.003384
4.0     0.000376
Name: AMT_REQ_CREDIT_BUREAU_HOUR, dtype: float64
```

In [85]:
```
app_df.AMT_REQ_CREDIT_BUREAU_DAY.value_counts(normalize=True)*100
```

Out[85]:
```
0.0    99.440209
1.0     0.485729
2.0     0.039851
3.0     0.016918
4.0     0.009775
5.0     0.003384
6.0     0.003008
9.0     0.000752
8.0     0.000376
Name: AMT_REQ_CREDIT_BUREAU_DAY, dtype: float64
```

- conclusion: We could see that 99% of values in the columns 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 IS 0.0. Hence impute there columns with mode

In [86]:
```
cols = ["AMT_REQ_CREDIT_BUREAU_HOUR" , "AMT_REQ_CREDIT_BUREAU_DAY" ,"AMT_REQ_CREDIT_BUREAU_WEEK", "AMT_REQ_CRED
```

In [87]:
```
for col in cols:
    app_df[col].fillna(app_df[col].mode()[0], inplace = True)
```

In [28]:
```
app_df.isnull().mean()* 100
```

Out[28]:
```
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
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                0.000000
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_2                   0.214626
EXT_SOURCE_3                   0.000000
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     0.000000
AMT_REQ_CREDIT_BUREAU_DAY      0.000000
AMT_REQ_CREDIT_BUREAU_WEEK     0.000000
AMT_REQ_CREDIT_BUREAU_MON      0.000000
AMT_REQ_CREDIT_BUREAU_QRT      0.000000
AMT_REQ_CREDIT_BUREAU_YEAR     0.000000
dtype: float64
```

Handling missing values less than 1%

In [88]:
```
null_cols = list(app_df.columns[app_df.isna().any()])
len(null_cols)
```

Out[88]:
```
10
```

In [30]:
```
app_df.NAME_TYPE_SUITE.value_counts(normalize=True)*100
```

Out[30]:
```
Unaccompanied      81.159562
Family             13.111205
Spouse, partner     3.713029
Children            1.066884
Other_B             0.578018
Other_A             0.282804
Group of people     0.088499
Name: NAME_TYPE_SUITE, dtype: float64
```

In [31]:
```
app_df.EXT_SOURCE_2.value_counts(normalize=True)*100
```

Out[31]:
```
0.285898    0.234967
0.262258    0.135897
0.265256    0.111781
0.159679    0.104937
0.265312    0.099723
              ...
0.004725    0.000326
0.257313    0.000326
0.282030    0.000326
0.181540    0.000326
0.267834    0.000326
Name: EXT_SOURCE_2, Length: 119831, dtype: float64
```

In [89]:
```
app_df.OBS_30_CNT_SOCIAL_CIRCLE.value_counts(normalize=True)*100
```

Out[89]:
```
0.0      53.479722
1.0      15.916669
2.0       9.725603
3.0       6.630559
4.0       4.614506
5.0       3.116904
6.0       2.105452
7.0       1.432347
8.0       0.968058
9.0       0.653529
10.0      0.448954
11.0      0.277986
12.0      0.212731
13.0      0.134099
14.0      0.084179
15.0      0.054162
16.0      0.043395
17.0      0.028712
18.0      0.015009
19.0      0.014356
20.0      0.009788
21.0      0.009462
22.0      0.007178
23.0      0.004894
25.0      0.003589
24.0      0.003589
27.0      0.001631
26.0      0.000979
30.0      0.000653
28.0      0.000326
29.0      0.000326
47.0      0.000326
348.0     0.000326
Name: OBS_30_CNT_SOCIAL_CIRCLE, dtype: float64
```

-conclusion:

- for catagorical columns impute the missing values with mode
- for numerical columns imputing the missing values with median

In [90]:
```
app_df.NAME_TYPE_SUITE.fillna(app_df.NAME_TYPE_SUITE.mode()[0], inplace = True)
```

In [91]:
```
app_df.CNT_FAM_MEMBERS.fillna(app_df.CNT_FAM_MEMBERS.mode()[0], inplace = True)
```

In [92]:
```
# Imputing numerical columns
app_df.EXT_SOURCE_2.fillna(app_df.EXT_SOURCE_2.median(), inplace = True)
app_df.AMT_GOODS_PRICE.fillna(app_df.AMT_GOODS_PRICE.median(), inplace = True)
app_df.AMT_ANNUITY.fillna(app_df.AMT_ANNUITY.median(), inplace = True)
app_df.OBS_30_CNT_SOCIAL_CIRCLE.fillna(app_df.OBS_30_CNT_SOCIAL_CIRCLE.median(), inplace = True)
app_df.DEF_30_CNT_SOCIAL_CIRCLE.fillna(app_df.DEF_30_CNT_SOCIAL_CIRCLE.median(), inplace = True)
app_df.OBS_60_CNT_SOCIAL_CIRCLE.fillna(app_df.OBS_60_CNT_SOCIAL_CIRCLE.median(), inplace = True)
app_df.DEF_60_CNT_SOCIAL_CIRCLE.fillna(app_df.DEF_60_CNT_SOCIAL_CIRCLE.median(), inplace = True)
app_df.DAYS_LAST_PHONE_CHANGE.fillna(app_df.DAYS_LAST_PHONE_CHANGE.median(), inplace = True)
```

In [93]:
```
null_cols = list(app_df.columns[app_df.isna().any()])
len(null_cols)
```

Out[93]:
```
0
```

In [94]:
```
app_df.isnull().mean()* 100
```

Out[94]:
```
SK_ID_CURR                    0.0
TARGET                        0.0
NAME_CONTRACT_TYPE            0.0
CODE_GENDER                   0.0
FLAG_OWN_CAR                  0.0
FLAG_OWN_REALTY               0.0
CNT_CHILDREN                  0.0
AMT_INCOME_TOTAL              0.0
AMT_CREDIT                    0.0
AMT_ANNUITY                   0.0
AMT_GOODS_PRICE               0.0
NAME_TYPE_SUITE               0.0
NAME_INCOME_TYPE              0.0
NAME_EDUCATION_TYPE           0.0
NAME_FAMILY_STATUS            0.0
NAME_HOUSING_TYPE             0.0
REGION_POPULATION_RELATIVE    0.0
DAYS_BIRTH                    0.0
DAYS_EMPLOYED                 0.0
DAYS_REGISTRATION             0.0
DAYS_ID_PUBLISH               0.0
FLAG_MOBIL                    0.0
FLAG_EMP_PHONE                0.0
FLAG_WORK_PHONE               0.0
FLAG_CONT_MOBILE              0.0
FLAG_PHONE                    0.0
FLAG_EMAIL                    0.0
OCCUPATION_TYPE               0.0
CNT_FAM_MEMBERS               0.0
REGION_RATING_CLIENT          0.0
REGION_RATING_CLIENT_W_CITY   0.0
WEEKDAY_APPR_PROCESS_START    0.0
HOUR_APPR_PROCESS_START       0.0
REG_REGION_NOT_LIVE_REGION    0.0
REG_REGION_NOT_WORK_REGION    0.0
LIVE_REGION_NOT_WORK_REGION   0.0
REG_CITY_NOT_LIVE_CITY        0.0
REG_CITY_NOT_WORK_CITY        0.0
LIVE_CITY_NOT_WORK_CITY       0.0
ORGANIZATION_TYPE             0.0
EXT_SOURCE_2                  0.0
EXT_SOURCE_3                  0.0
OBS_30_CNT_SOCIAL_CIRCLE      0.0
DEF_30_CNT_SOCIAL_CIRCLE      0.0
OBS_60_CNT_SOCIAL_CIRCLE      0.0
DEF_60_CNT_SOCIAL_CIRCLE      0.0
DAYS_LAST_PHONE_CHANGE        0.0
FLAG_DOCUMENT_2               0.0
FLAG_DOCUMENT_3               0.0
FLAG_DOCUMENT_4               0.0
FLAG_DOCUMENT_5               0.0
FLAG_DOCUMENT_6               0.0
FLAG_DOCUMENT_7               0.0
FLAG_DOCUMENT_8               0.0
FLAG_DOCUMENT_9               0.0
FLAG_DOCUMENT_10              0.0
FLAG_DOCUMENT_11              0.0
FLAG_DOCUMENT_12              0.0
FLAG_DOCUMENT_13              0.0
FLAG_DOCUMENT_14              0.0
FLAG_DOCUMENT_15              0.0
FLAG_DOCUMENT_16              0.0
FLAG_DOCUMENT_17              0.0
FLAG_DOCUMENT_18              0.0
FLAG_DOCUMENT_19              0.0
FLAG_DOCUMENT_20              0.0
FLAG_DOCUMENT_21              0.0
AMT_REQ_CREDIT_BUREAU_HOUR    0.0
AMT_REQ_CREDIT_BUREAU_DAY     0.0
AMT_REQ_CREDIT_BUREAU_WEEK    0.0
AMT_REQ_CREDIT_BUREAU_MON     0.0
AMT_REQ_CREDIT_BUREAU_QRT     0.0
AMT_REQ_CREDIT_BUREAU_YEAR    0.0
dtype: float64
```

Convert Negative values to positive in days variable so that median is not affected

In [95]:
```
app_df.DAYS_BIRTH = app_df.DAYS_BIRTH.apply(lambda x: abs(x))
app_df.DAYS_EMPLOYED= app_df.DAYS_EMPLOYED.apply(lambda x: abs(x))
app_df.DAYS_REGISTRATION = app_df.DAYS_REGISTRATION.apply(lambda x: abs(x))
app_df.DAYS_ID_PUBLISH = app_df.DAYS_ID_PUBLISH.apply(lambda x: abs(x))
app_df.DAYS_LAST_PHONE_CHANGE= app_df.DAYS_LAST_PHONE_CHANGE.apply(lambda x: abs(x))
```

Binning of continuous variables

Standardizing Days columns in years for easy binning

In [96]:
```
app_df["YEARS_BIRTH"] = app_df.DAYS_BIRTH.apply(lambda x: int(x//356))
app_df["YEARS_EMPLOYED"] = app_df.DAYS_EMPLOYED.apply(lambda x: int(x//356))
app_df["YEARS_REGISTRATION"] = app_df.DAYS_REGISTRATION.apply(lambda x: int(x//356))
app_df["YEARS_ID_PUBLISH"] = app_df.DAYS_ID_PUBLISH.apply(lambda x: int(x//356))
app_df["YEARS_LAST_PHONE_CHANGE"] = app_df.DAYS_LAST_PHONE_CHANGE.apply(lambda x: int(x//356))
```

Binning of AMT_CREDIT column

In [97]:
```
app_df.AMT_CREDIT.value_counts(normalize=True)*100
```

Out[97]:
```
450000.0     3.157285
675000.0     2.886726
225000.0     2.654214
180000.0     2.387557
270000.0     2.354713
               ...
487318.5     0.000325
630400.5     0.000325
1875276.0    0.000325
1395895.5    0.000325
1391130.0    0.000325
Name: AMT_CREDIT, Length: 5603, dtype: float64
```

In [98]:
```
app_df.AMT_CREDIT.describe()
```

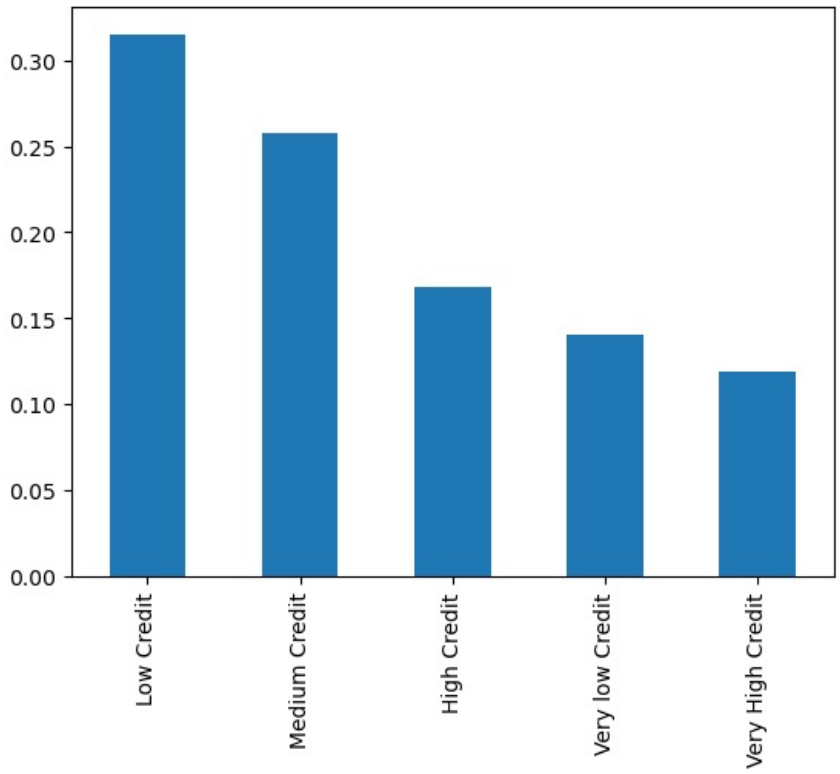
Out[98]:
```
count    3.075110e+05
mean     5.990260e+05
std      4.024908e+05
min      4.500000e+04
25%      2.700000e+05
50%      5.135310e+05
75%      8.086500e+05
max      4.050000e+06
Name: AMT_CREDIT, dtype: float64
```

In [99]:
```
app_df["AMT_CREDIT_Category"] = pd.cut(app_df.AMT_CREDIT, [0,200000,400000, 600000,800000,1000000],
                                       labels = [ "Very low Credit", "Low Credit", "Medium Credit" , "High Credi
```

In [100…
```
app_df.AMT_CREDIT_Category.value_counts(normalize=True)*100
```

Out[100]:
```
Low Credit          31.511770
Medium Credit       25.733324
High Credit         16.791314
Very low Credit     14.035088
Very High Credit    11.928504
Name: AMT_CREDIT_Category, dtype: float64
```

In [44]:
```
app_df["AMT_CREDIT_Category"].value_counts(normalize=True).plot.bar()
plt.show()
```



- Conclusion: The credit amount of the loan for amount low (2L to 4L) or very high(above 8L)

Binning YEARS_BIRTH column

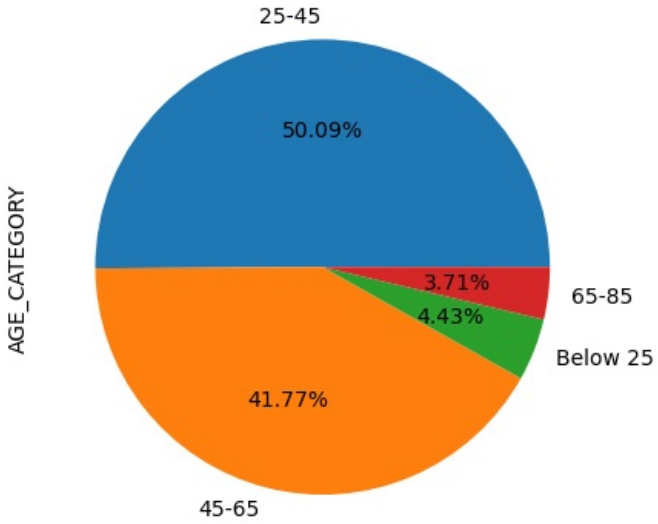
Binning YEARS_BIRTH column

In [101…
```
app_df["AGE_CATEGORY"] = pd.cut(app_df.YEARS_BIRTH, [0, 25, 45, 65, 85], labels = ["Below 25", "25-45", "45-65"
```

In [102…
```
app_df.AGE_CATEGORY.value_counts(normalize=True)*100
```

Out[102]:
```
25-45       50.094143
45-65       41.772489
Below 25     4.426834
65-85        3.706534
Name: AGE_CATEGORY, dtype: float64
```

In [47]:
```
app_df["AGE_CATEGORY"].value_counts(normalize=True).plot.pie(autopct ='%1.2f%%')
plt.show()
```



-conclusion: Most of the Applicants are between 25-45 age group

Data Imbalance Check

In [103…
```
app_df.head()
```

Out[103]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME
0	100002	1	Cash loans	M	N	Y	0	
1	100003	0	Cash loans	F	N	N	0	
2	100004	0	Revolving loans	M	Y	Y	0	
3	100006	0	Cash loans	F	N	Y	0	
4	100007	0	Cash loans	M	N	Y	0	

Dividing Application Dataset with Target Variable as 0 and 1

In [104…
```
tar_0 = app_df[app_df.TARGET == 0]
tar_1 = app_df[app_df.TARGET == 1]
```

In [105…
```
app_df.TARGET.value_counts(normalize=True)*100
```

Out[105]:
```
0    91.927118
1     8.072882
Name: TARGET, dtype: float64
```

-conclusion: 1 out of 9/10 applicants and defaulters

Univariate Analysis

In [106…
```
cat_cols = list(app_df.columns[app_df.dtypes == np.object])
num_cols = list(app_df.columns[app_df.dtypes == np.int64]) + list(app_df.columns[app_df.dtypes == np.float64])
```

In [107…
```
cat_cols
```

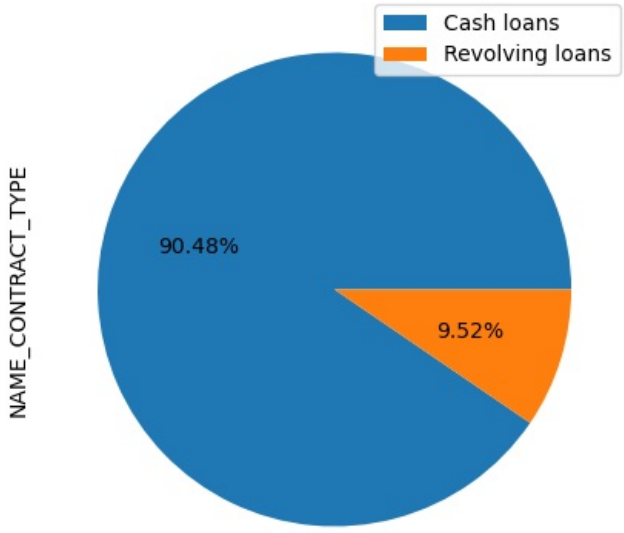
Out[107]:
```
['NAME_CONTRACT_TYPE',
 'CODE_GENDER',
 'FLAG_OWN_CAR',
 'FLAG_OWN_REALTY',
 'NAME_TYPE_SUITE',
 'NAME_INCOME_TYPE',
 'NAME_EDUCATION_TYPE',
 'NAME_FAMILY_STATUS',
 'NAME_HOUSING_TYPE',
 'OCCUPATION_TYPE',
 'WEEKDAY_APPR_PROCESS_START',
 'ORGANIZATION_TYPE']
```

In [108…
```
num_cols
```

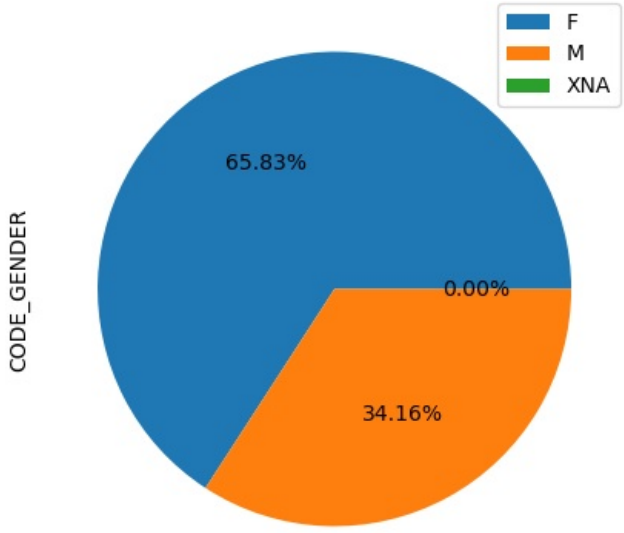
Out[108]:
```
['SK_ID_CURR',
 'TARGET',
 'CNT_CHILDREN',
 'DAYS_BIRTH',
 'DAYS_EMPLOYED',
 'DAYS_ID_PUBLISH',
 'FLAG_MOBIL',
 'FLAG_EMP_PHONE',
 'FLAG_WORK_PHONE',
 'FLAG_CONT_MOBILE',
 'FLAG_PHONE',
 'FLAG_EMAIL',
 'REGION_RATING_CLIENT',
 'REGION_RATING_CLIENT_W_CITY',
 'HOUR_APPR_PROCESS_START',
 'REG_REGION_NOT_LIVE_REGION',
 'REG_REGION_NOT_WORK_REGION',
 'LIVE_REGION_NOT_WORK_REGION',
 'REG_CITY_NOT_LIVE_CITY',
 'REG_CITY_NOT_WORK_CITY',
 'LIVE_CITY_NOT_WORK_CITY',
 'FLAG_DOCUMENT_2',
 'FLAG_DOCUMENT_3',
 'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_6',
 'FLAG_DOCUMENT_7',
 'FLAG_DOCUMENT_8',
 'FLAG_DOCUMENT_9',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_11',
 'FLAG_DOCUMENT_12',
 'FLAG_DOCUMENT_13',
 'FLAG_DOCUMENT_14',
 'FLAG_DOCUMENT_15',
 'FLAG_DOCUMENT_16',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_18',
 'FLAG_DOCUMENT_19',
 'FLAG_DOCUMENT_20',
 'FLAG_DOCUMENT_21',
 'YEARS_BIRTH',
 'YEARS_EMPLOYED',
 'YEARS_REGISTRATION',
 'YEARS_ID_PUBLISH',
 'YEARS_LAST_PHONE_CHANGE',
 'AMT_INCOME_TOTAL',
 'AMT_CREDIT',
 'AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'REGION_POPULATION_RELATIVE',
 'DAYS_REGISTRATION',
 'CNT_FAM_MEMBERS',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3',
 'OBS_30_CNT_SOCIAL_CIRCLE',
 'DEF_30_CNT_SOCIAL_CIRCLE',
 'OBS_60_CNT_SOCIAL_CIRCLE',
 'DEF_60_CNT_SOCIAL_CIRCLE',
 'DAYS_LAST_PHONE_CHANGE',
 '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']
```

In [54]:
```
for col in cat_cols:
    print(app_df[col].value_counts(normalize=True)*100)
    plt.figure(figsize=[5,5])
    app_df[col].value_counts(normalize=True).plot.pie(labeldistance= None, autopct ='%1.2f%%')
    plt.legend()
    plt.show()
```

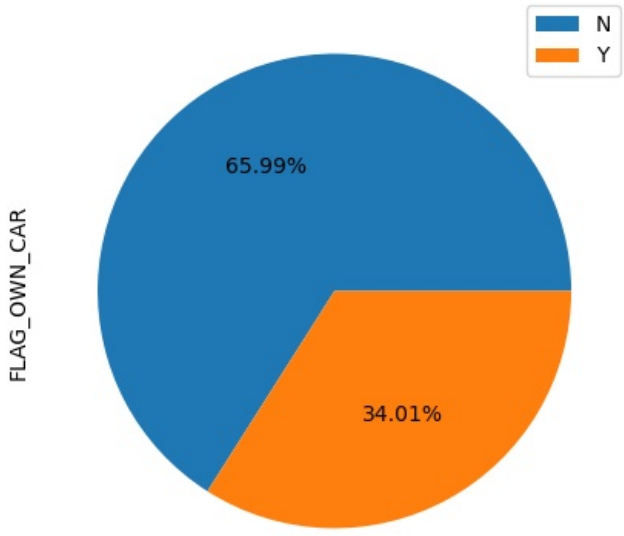
```
Cash loans         90.478715
Revolving loans     9.521285
Name: NAME_CONTRACT_TYPE, dtype: float64
```



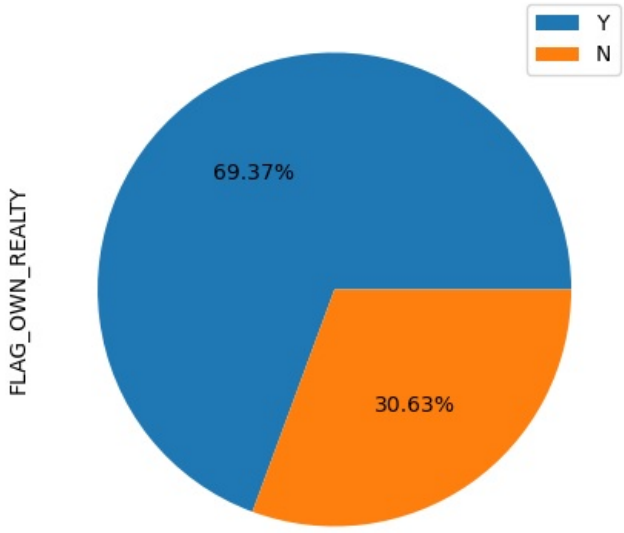

```
F      65.834393
M      34.164306
XNA     0.001301
Name: CODE_GENDER, dtype: float64
```



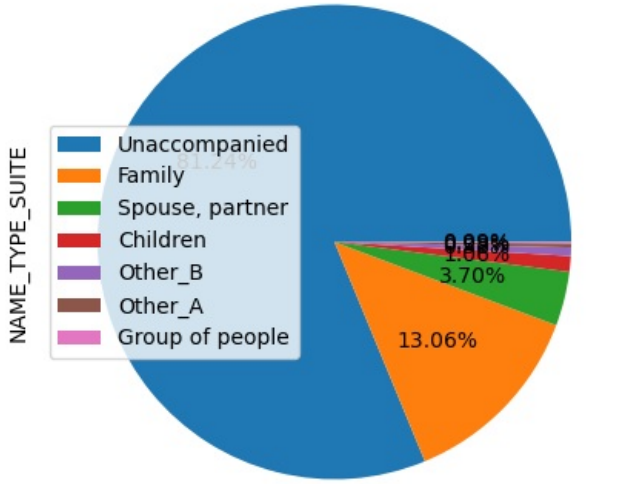

```
N    65.989184
Y    34.010816
Name: FLAG_OWN_CAR, dtype: float64
```



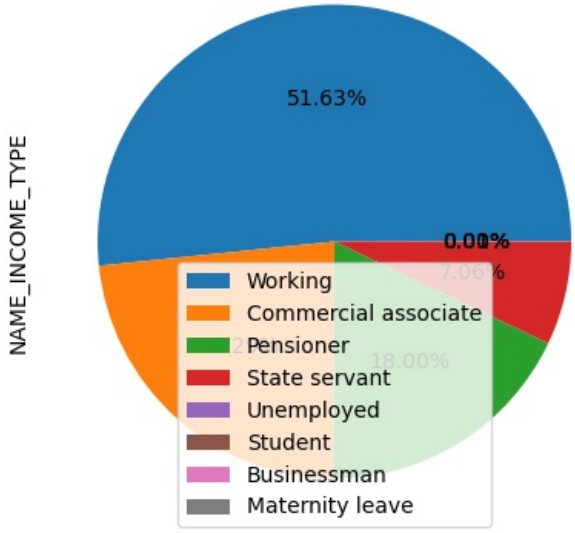
```
Y    69.367275
N    30.632725
Name: FLAG_OWN_REALTY, dtype: float64
```



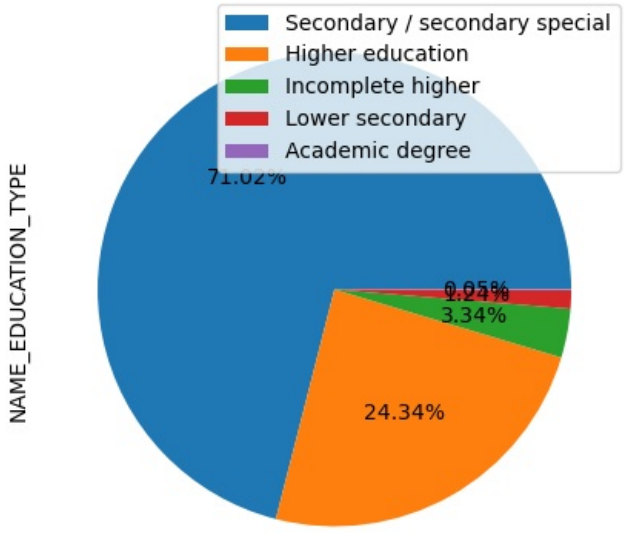

```
Unaccompanied      81.238720
Family             13.056118
Spouse, partner     3.697429
Children            1.062401
Other_B             0.575589
Other_A             0.281616
Group of people     0.088127
Name: NAME_TYPE_SUITE, dtype: float64
```



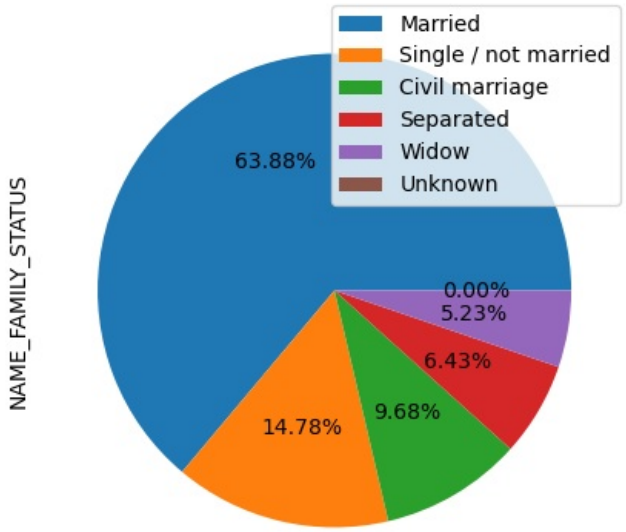

```
Working                 51.631974
Commercial associate    23.289248
Pensioner               18.003258
State servant            7.057634
Unemployed               0.007154
Student                  0.005853
Businessman              0.003252
Maternity leave          0.001626
Name: NAME_INCOME_TYPE, dtype: float64
```



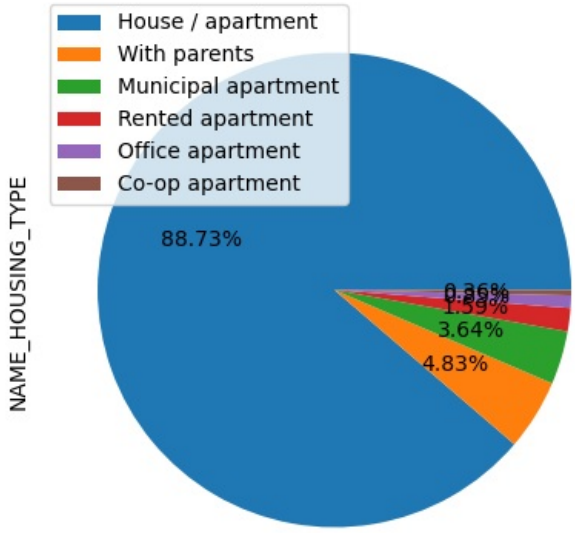

```
Secondary / secondary special    71.018923
Higher education                 24.344820
Incomplete higher                 3.341994
Lower secondary                   1.240931
Academic degree                   0.053331
Name: NAME_EDUCATION_TYPE, dtype: float64
```



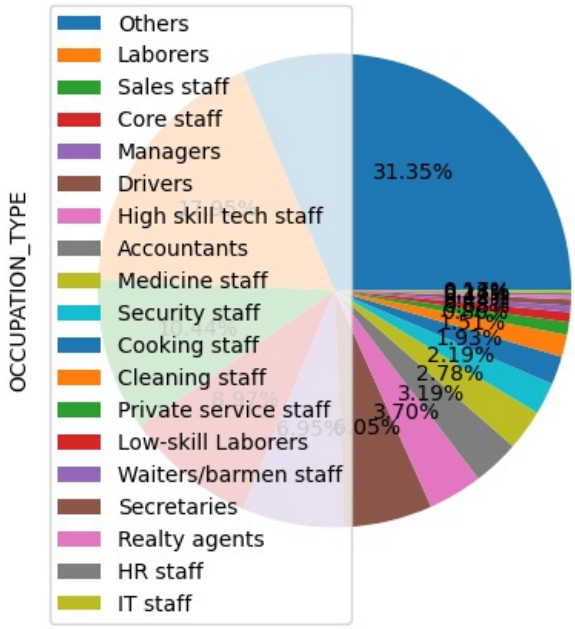

```
Married                 63.878040
Single / not married    14.778008
Civil marriage           9.682580
Separated                6.429038
Widow                    5.231683
Unknown                  0.000650
Name: NAME_FAMILY_STATUS, dtype: float64
```



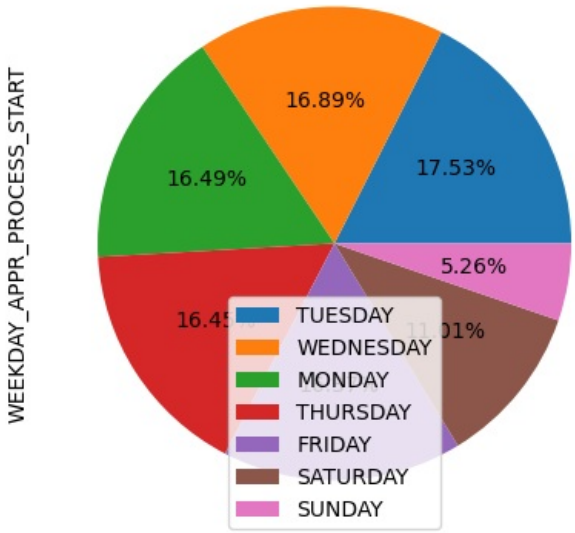

```
House / apartment      88.734387
With parents            4.825844
Municipal apartment     3.636618
Rented apartment        1.587260
Office apartment        0.851026
Co-op apartment         0.364865
Name: NAME_HOUSING_TYPE, dtype: float64
```




```
Others                 31.345545
Laborers               17.946025
Sales staff            10.439301
Core staff              8.965533
Managers                6.949670
Drivers                 6.049540
High skill tech staff   3.700681
Accountants             3.191105
Medicine staff          2.776161
Security staff          2.185613
Cooking staff           1.933589
Cleaning staff          1.513117
Private service staff   0.862408
Low-skill Laborers      0.680626
Waiters/barmen staff    0.438358
Secretaries             0.424375
Realty agents           0.244219
HR staff                0.183083
IT staff                0.171051
Name: OCCUPATION_TYPE, dtype: float64
```




```
TUESDAY      17.528153
WEDNESDAY    16.888502
MONDAY       16.491768
THURSDAY     16.451769
FRIDAY       16.369496
SATURDAY     11.008387
SUNDAY        5.261926
Name: WEEKDAY_APPR_PROCESS_START, dtype: float64
```



```
Business Entity Type 3    22.110429
XNA                       18.007161
Self-employed             12.491260
Other                      5.425172
Medicine                   3.639870
Business Entity Type 2     3.431747
Government                 3.383294
School                     2.891929
Trade: type 7              2.546576
Kindergarten               2.237318
Construction               2.185613
Business Entity Type 1     1.945947
Transport: type 4          1.755384
Trade: type 3              1.135569
Industry: type 9           1.095245
Industry: type 3           1.065978
Security                   1.055897
Housing                    0.961917
Industry: type 11          0.879318
Military                   0.856555
Bank                       0.815255
Agriculture                0.798020
Police                     0.761274
Transport: type 2          0.716722
Postal                     0.701438
Security Ministries        0.641928
Trade: type 2              0.617864
Restaurant                 0.588922
Services                   0.512177
University                 0.431529
Industry: type 7           0.425025
Transport: type 3          0.386002
Industry: type 1           0.337874
Hotel                      0.314135
Electricity                0.308932
Industry: type 4           0.285193
Trade: type 6              0.205196
Industry: type 5           0.194790
Insurance                  0.194139
Telecom                    0.187636
Emergency                  0.182107
Industry: type 2           0.148938
Advertising                0.139507
Realtor                    0.128776
Culture                    0.123248
Industry: type 12          0.119996
Trade: type 1              0.113167
Mobile                     0.103086
Legal Services             0.099183
Cleaning                   0.084550
Transport: type 1          0.065364
Industry: type 6           0.036421
Industry: type 10          0.035446
Religion                   0.027641
Industry: type 13          0.021788
Trade: type 4              0.020812
Trade: type 5              0.015934
Industry: type 8           0.007805
Name: ORGANIZATION_TYPE, dtype: float64
```

-Conclusion: insoghts on below columns

```
1.NAME_CONTRACT_TYPE-More applications have cash loans than Revolving items
2.CODE_GENDER- Number of female applicants are twice than that of male applicants
3.FLAG_OWN_CAR- Most(70%) of the applicants do not own a car
4.FLAG_OWN_REALITY- Most(70%) of applicants do not own a house
5.FLAG_OWN_SUITE- Most(81%) of applicants are unaccompanied
6.NAME_INCOME_TYPE- Most(51%) of applicants are earning their income from work
```

7.NAME_EDUCATION_TYPE- Most(71%) of applicants have completed Secondary/secondary special education
8.NAME_FAMILY_STATUS- Most(63%) of applicants are married
9.NAME_HOUSING_TYPE- Most(88%) of housing type of applicants have house/appartment
10.OCCUPATION_TYPE- most(31%) of applicants have other occupation type
11.WEEKDAY_APPR_PROCESS_START-Most of the applicants have applied loan on tuesday
12.ORGANISATION_TYPE-Most of the organisation type of employees are Business Entity Type 3

plot on Numerical columns

categorizing columns with or without flags

In [109…
```
num_cols_withoutflag = []
num_cols_withflag = []
for col in num_cols:
    if col.startswith("FLAG"):
        num_cols_withflag.append(col)
    else:
        num_cols_withoutflag.append(col)
```

In [110…
```
num_cols_withflag
```

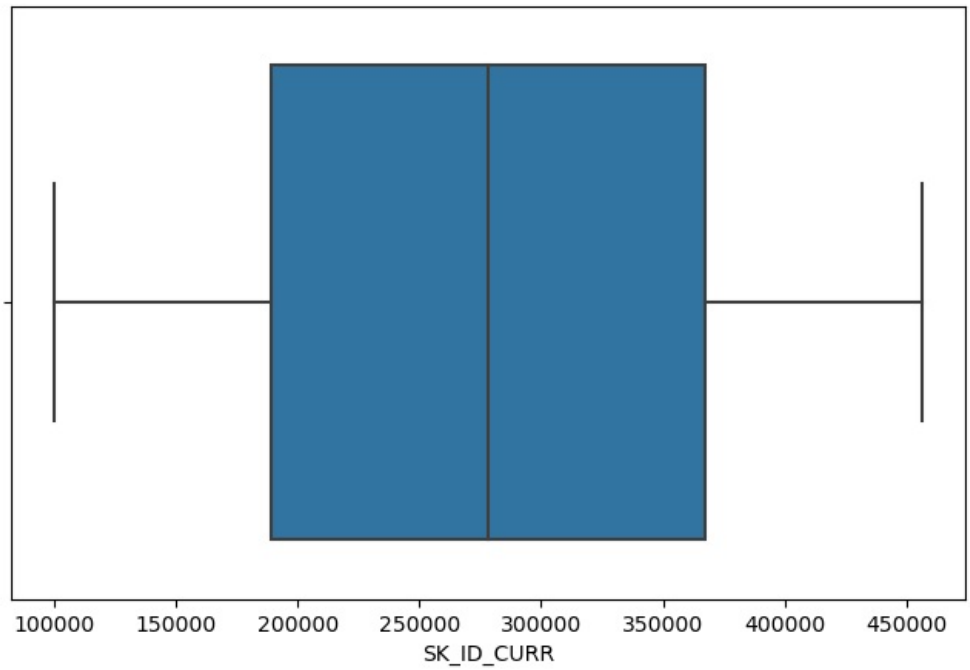
Out[110]:
```
['FLAG_MOBIL',
 'FLAG_EMP_PHONE',
 'FLAG_WORK_PHONE',
 'FLAG_CONT_MOBILE',
 'FLAG_PHONE',
 'FLAG_EMAIL',
 'FLAG_DOCUMENT_2',
 'FLAG_DOCUMENT_3',
 'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_6',
 'FLAG_DOCUMENT_7',
 'FLAG_DOCUMENT_8',
 'FLAG_DOCUMENT_9',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_11',
 'FLAG_DOCUMENT_12',
 'FLAG_DOCUMENT_13',
 'FLAG_DOCUMENT_14',
 'FLAG_DOCUMENT_15',
 'FLAG_DOCUMENT_16',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_18',
 'FLAG_DOCUMENT_19',
 'FLAG_DOCUMENT_20',
 'FLAG_DOCUMENT_21']
```

In [57]:
```
num_cols_withoutflag
```

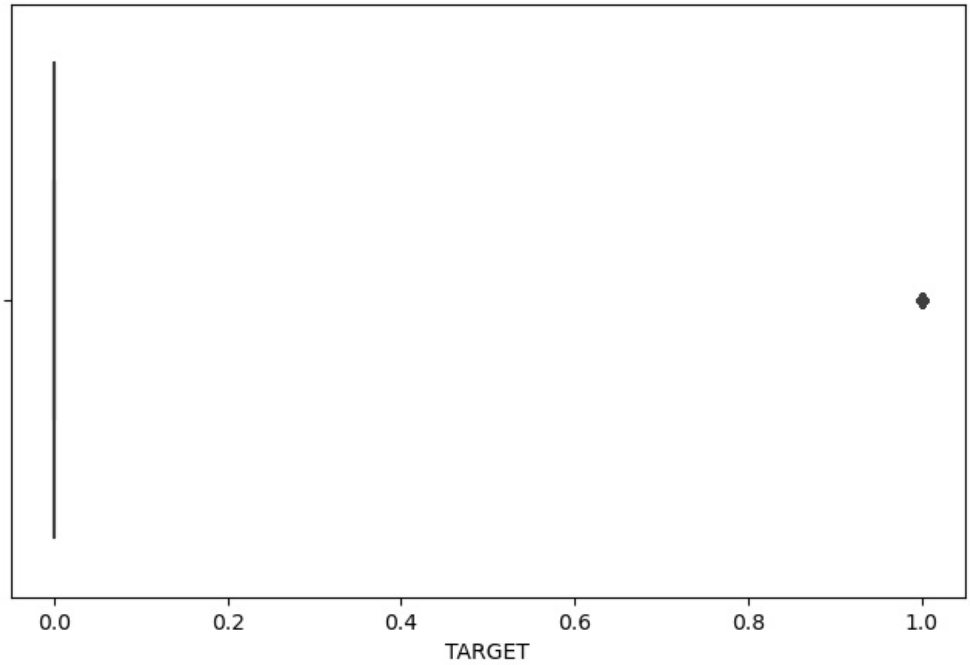
Out[57]:
```
['SK_ID_CURR',
 'TARGET',
 'CNT_CHILDREN',
 'DAYS_BIRTH',
 'DAYS_EMPLOYED',
 'DAYS_ID_PUBLISH',
 'REGION_RATING_CLIENT',
 'REGION_RATING_CLIENT_W_CITY',
 'HOUR_APPR_PROCESS_START',
 'REG_REGION_NOT_LIVE_REGION',
 'REG_REGION_NOT_WORK_REGION',
 'LIVE_REGION_NOT_WORK_REGION',
 'REG_CITY_NOT_LIVE_CITY',
 'REG_CITY_NOT_WORK_CITY',
 'LIVE_CITY_NOT_WORK_CITY',
 'YEARS_BIRTH',
 'YEARS_EMPLOYED',
 'YEARS_REGISTRATION',
 'YEARS_ID_PUBLISH',
 'YEARS_LAST_PHONE_CHANGE',
 'AMT_INCOME_TOTAL',
 'AMT_CREDIT',
 'AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'REGION_POPULATION_RELATIVE',
 'DAYS_REGISTRATION',
 'CNT_FAM_MEMBERS',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3',
 'OBS_30_CNT_SOCIAL_CIRCLE',
 'DEF_30_CNT_SOCIAL_CIRCLE',
 'OBS_60_CNT_SOCIAL_CIRCLE',
 'DEF_60_CNT_SOCIAL_CIRCLE',
 'DAYS_LAST_PHONE_CHANGE',
 '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']
```

In [111…
```
for col in num_cols_withoutflag:
    print(app_df[col].describe())
    plt.figure(figsize = [8,5])
    sns.boxplot(data=app_df , x=col)
    plt.show()
    print("----------------------")
```

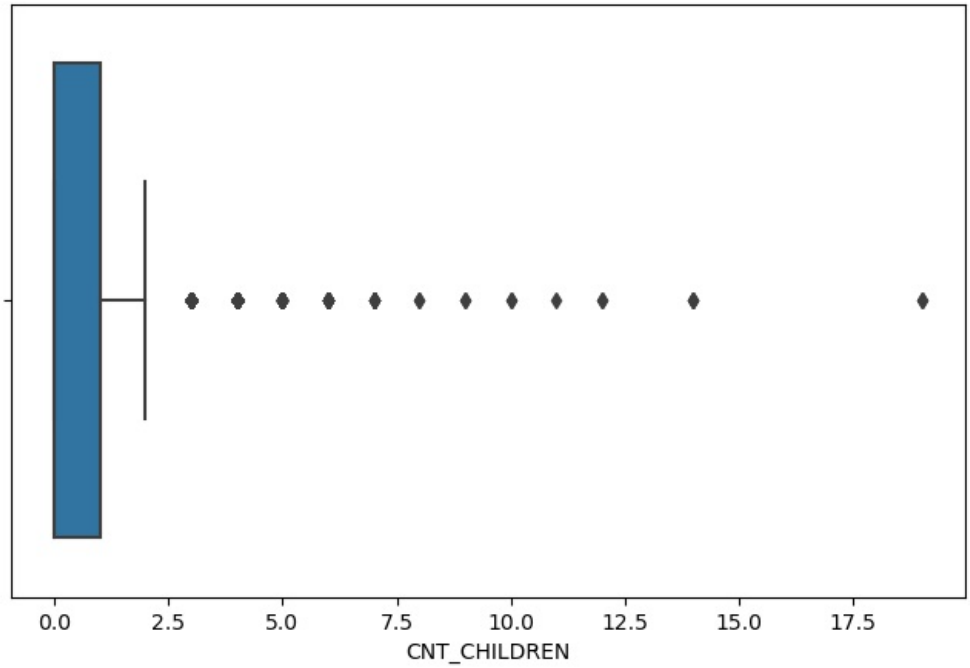
```
count    307511.000000
mean     278180.518577
std      102790.175348
min      100002.000000
25%      189145.500000
50%      278202.000000
75%      367142.500000
max      456255.000000
Name: SK_ID_CURR, dtype: float64
```



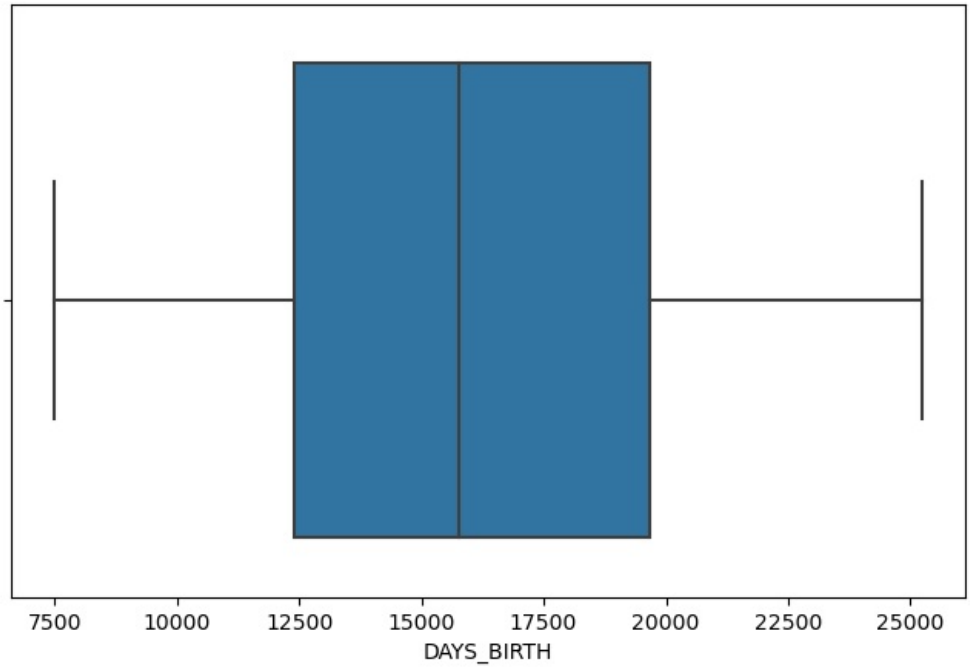
```
----------
count    307511.000000
mean          0.080729
std           0.272419
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: TARGET, dtype: float64
```



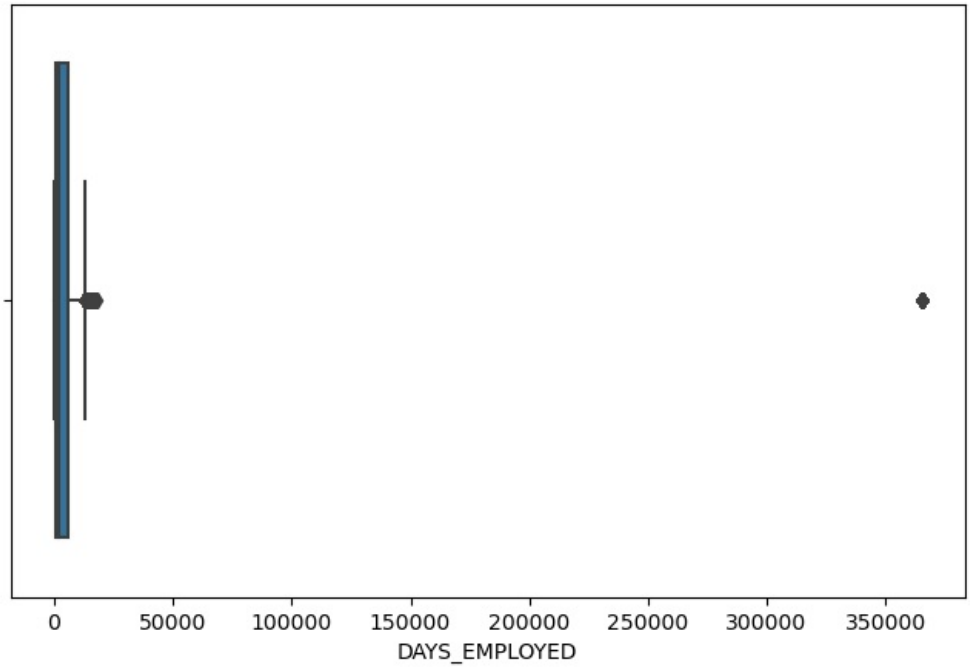

```
----------
count    307511.000000
mean          0.417052
std           0.722121
min           0.000000
25%           0.000000
50%           0.000000
75%           1.000000
max          19.000000
Name: CNT_CHILDREN, dtype: float64
```



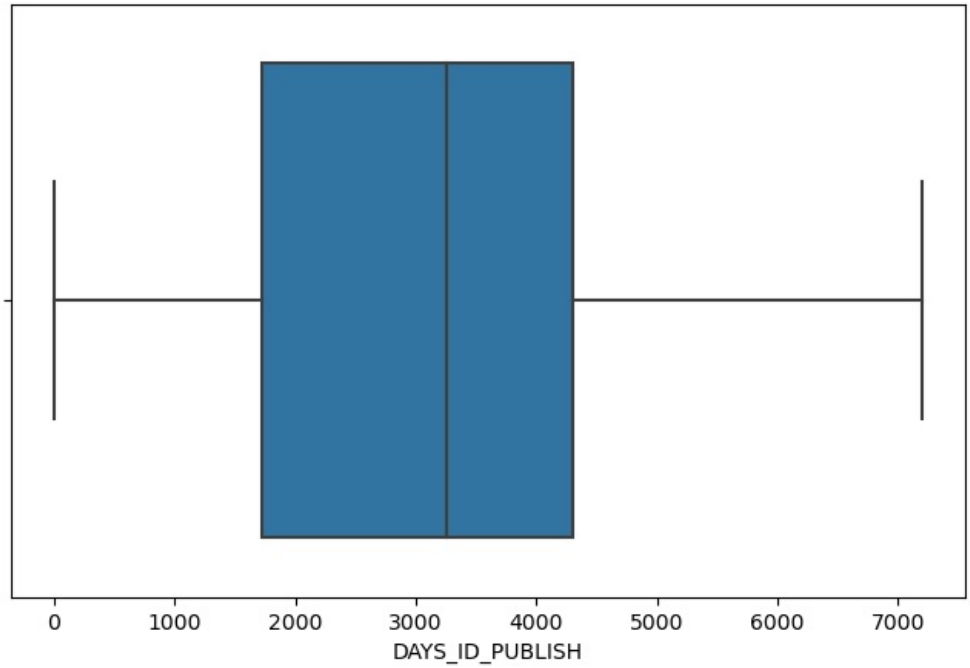

```
----------
count    307511.000000
mean      16036.995067
std        4363.988632
min        7489.000000
25%       12413.000000
50%       15750.000000
75%       19682.000000
max       25229.000000
Name: DAYS_BIRTH, dtype: float64
```



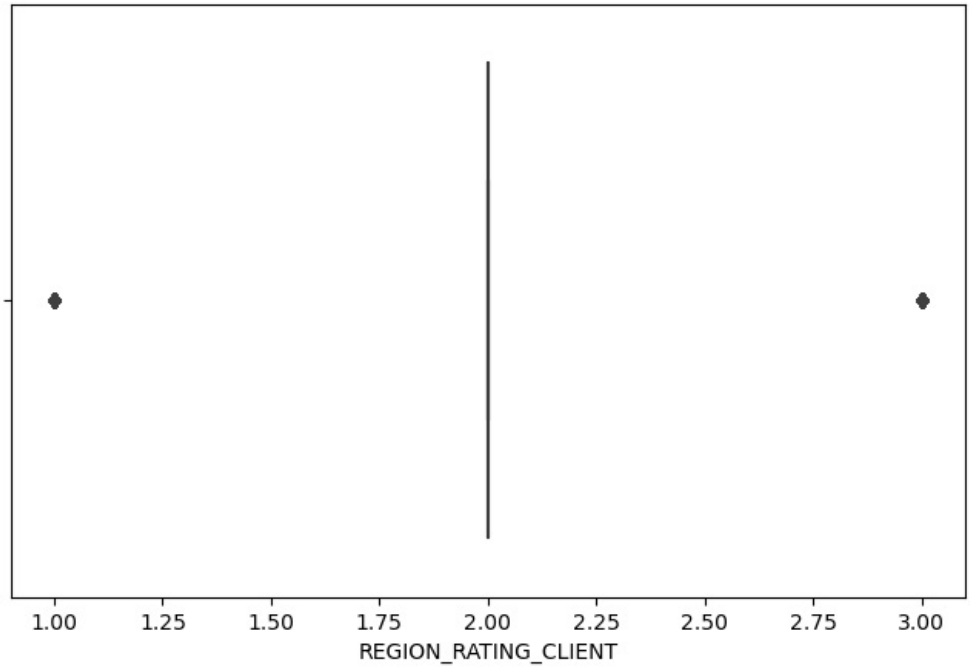

```
----------------------
count    307511.000000
mean      67724.742149
std      139443.751806
min           0.000000
25%         933.000000
50%        2219.000000
75%        5707.000000
max      365243.000000
Name: DAYS_EMPLOYED, dtype: float64
```



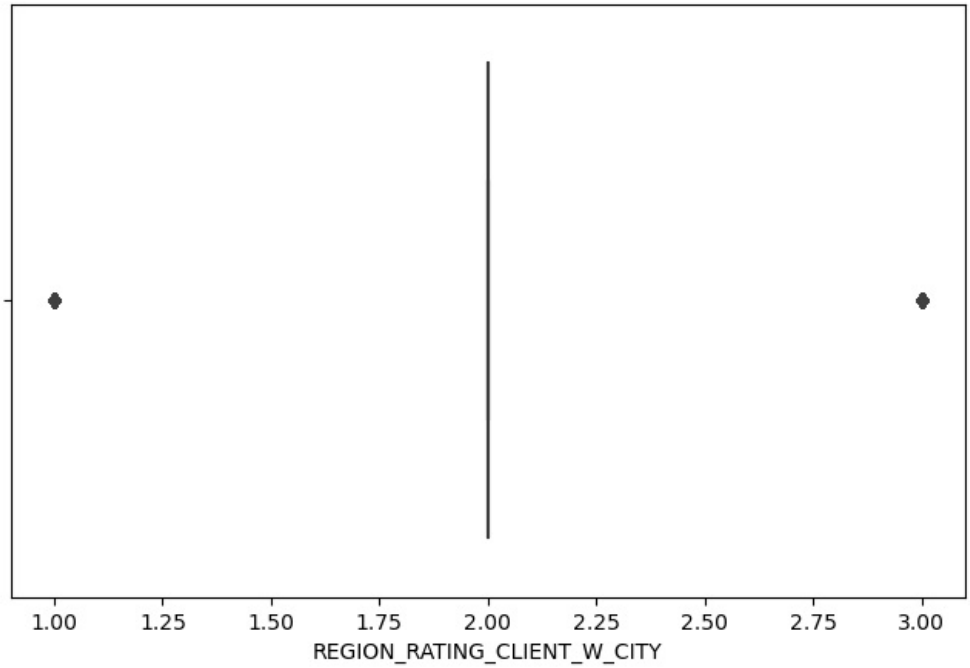

```
----------------------
count    307511.000000
mean       2994.202373
std        1509.450419
min           0.000000
25%        1720.000000
50%        3254.000000
75%        4299.000000
max        7197.000000
Name: DAYS_ID_PUBLISH, dtype: float64
```



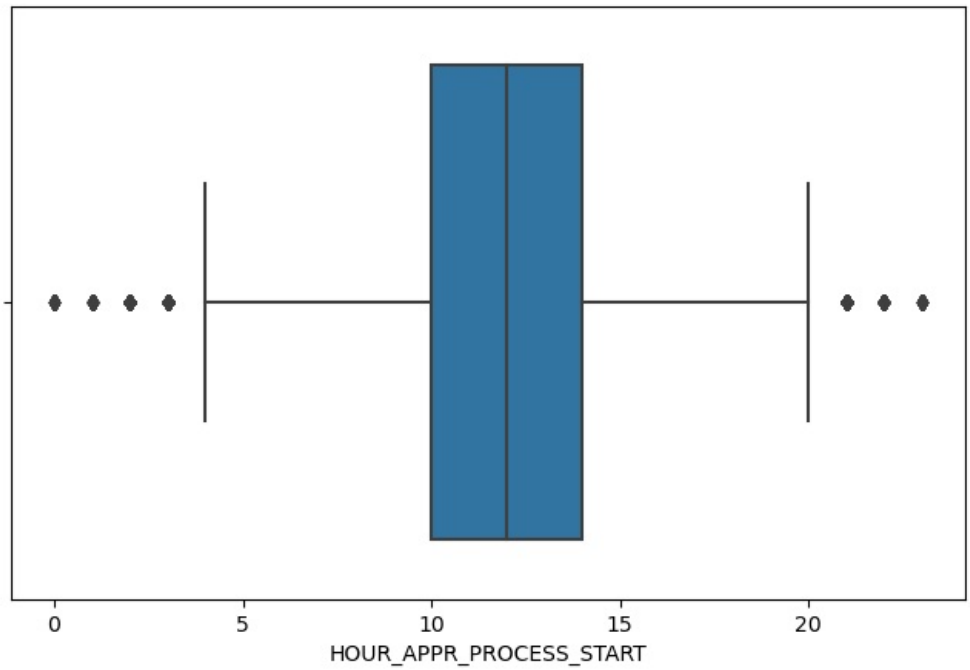

```
---------------------
count    307511.000000
mean          2.052463
std           0.509034
min           1.000000
25%           2.000000
50%           2.000000
75%           2.000000
max           3.000000
Name: REGION_RATING_CLIENT, dtype: float64
```



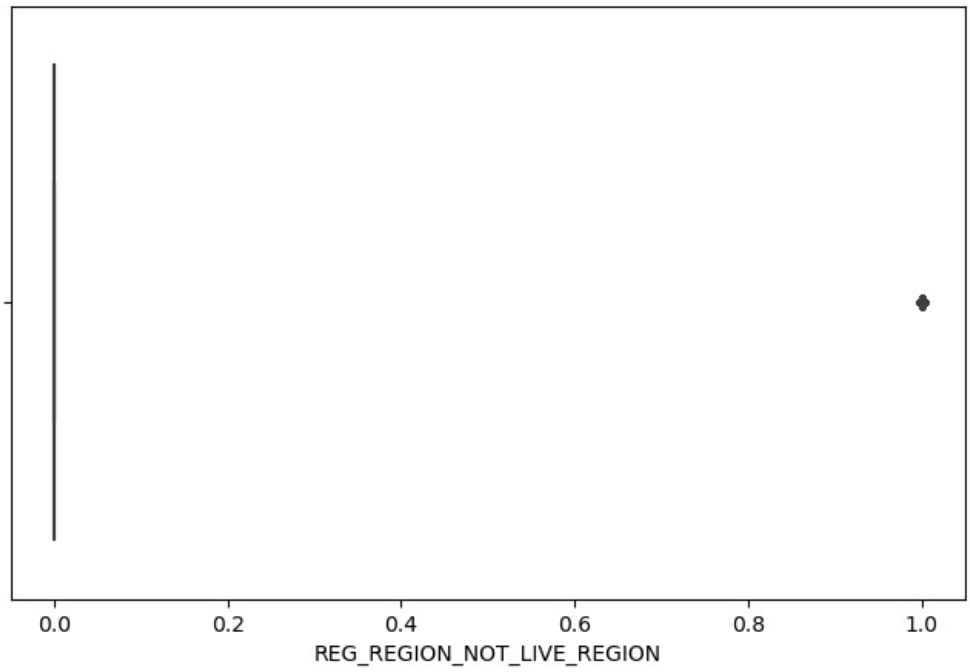

```
---------------------
count    307511.000000
mean          2.031521
std           0.502737
min           1.000000
25%           2.000000
50%           2.000000
75%           2.000000
max           3.000000
Name: REGION_RATING_CLIENT_W_CITY, dtype: float64
```



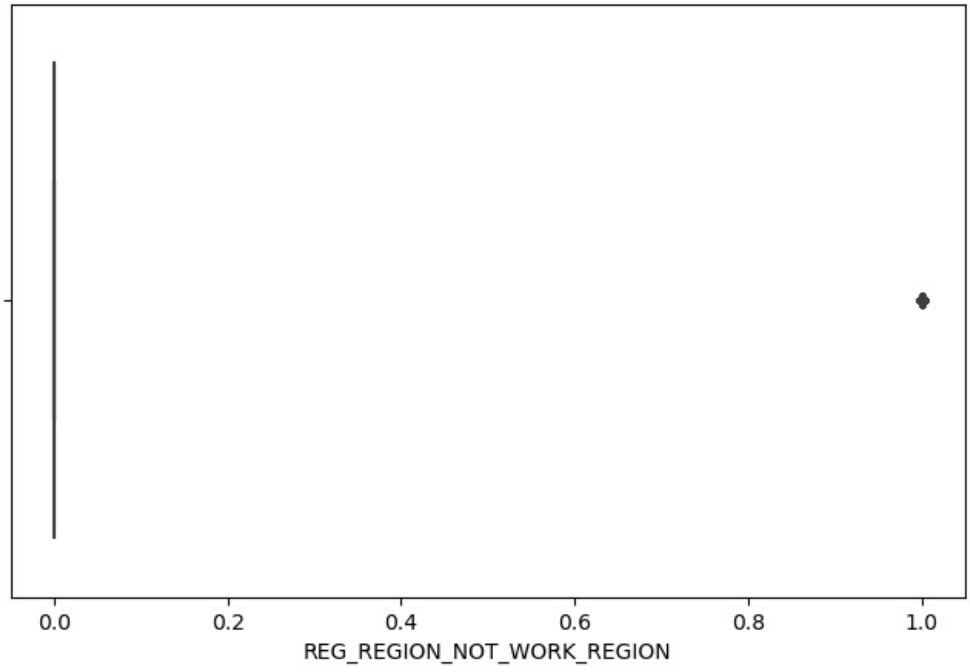

```
----------------------
count    307511.000000
mean         12.063419
std           3.265832
min           0.000000
25%          10.000000
50%          12.000000
75%          14.000000
max          23.000000
Name: HOUR_APPR_PROCESS_START, dtype: float64
```



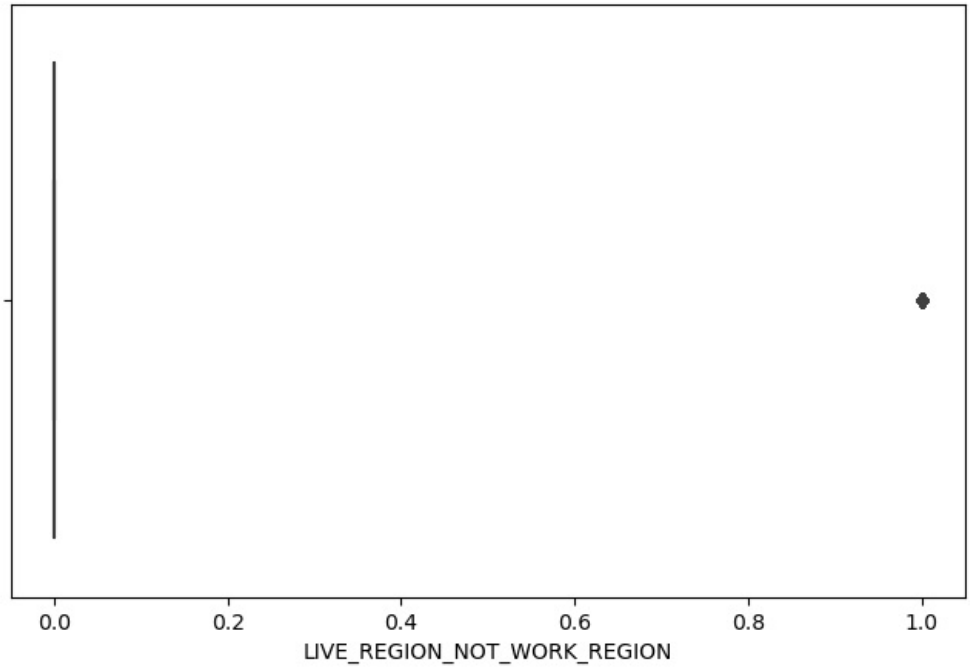

```
----------------------
count    307511.000000
mean          0.015144
std           0.122126
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: REG_REGION_NOT_LIVE_REGION, dtype: float64
```



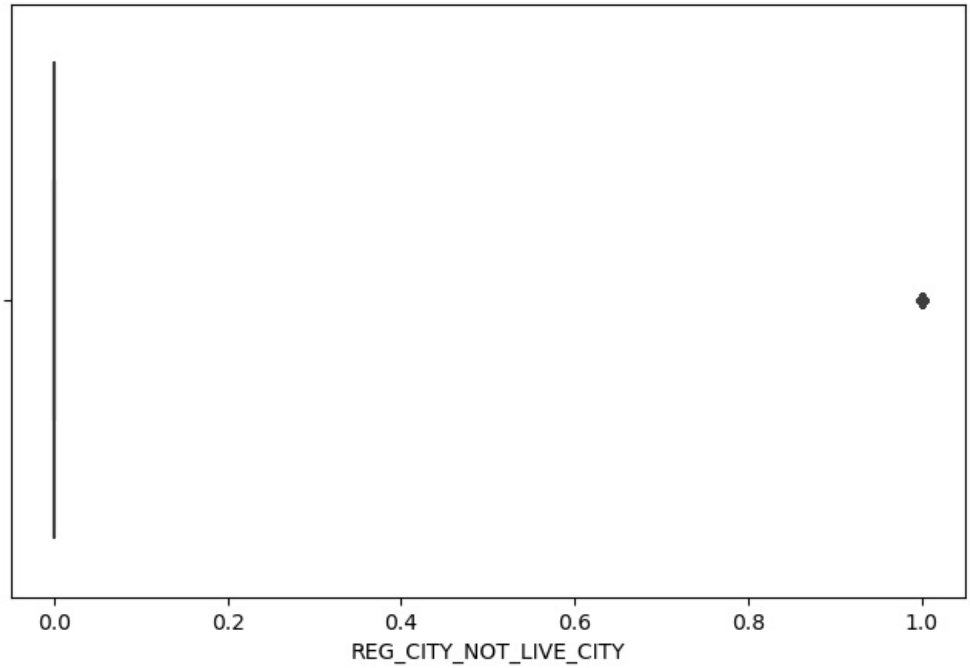

```
----------------------
count    307511.000000
mean          0.050769
std           0.219526
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: REG_REGION_NOT_WORK_REGION, dtype: float64
```



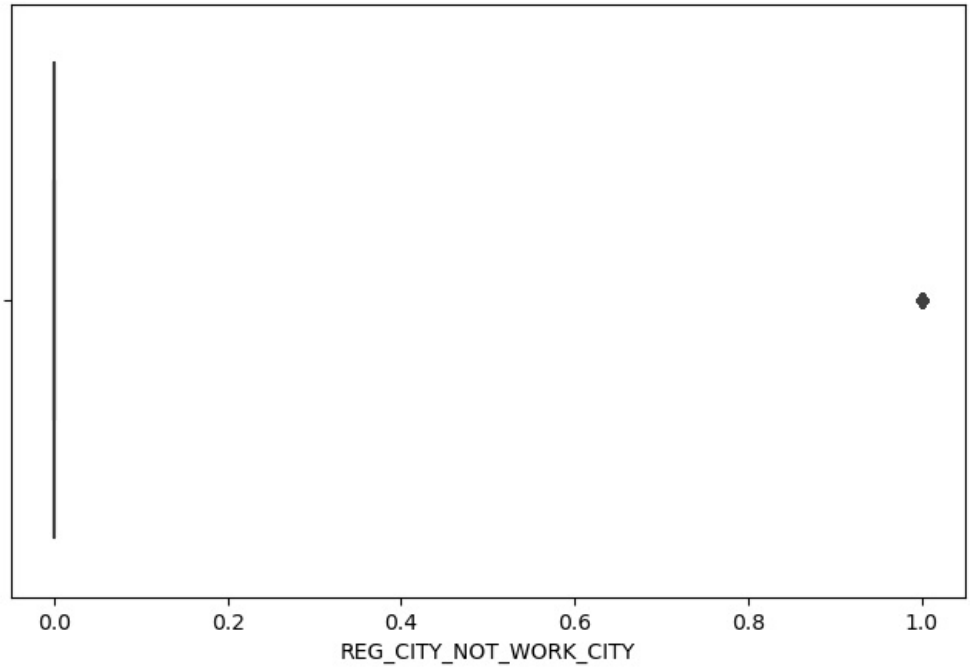

```
----------------------
count    307511.000000
mean          0.040659
std           0.197499
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: LIVE_REGION_NOT_WORK_REGION, dtype: float64
```



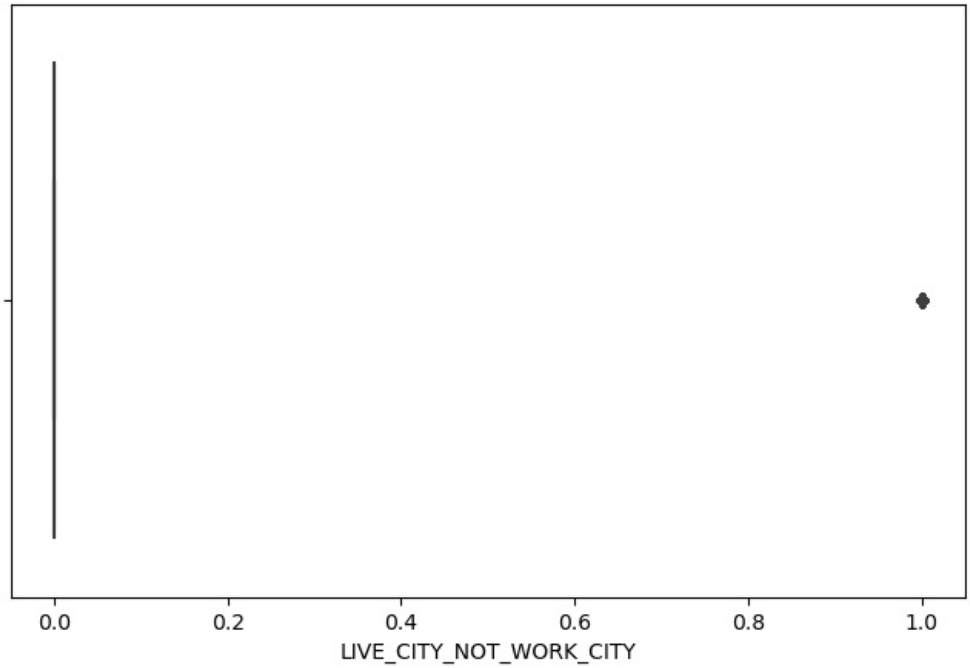

```
----------------------
count    307511.000000
mean          0.078173
std           0.268444
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: REG_CITY_NOT_LIVE_CITY, dtype: float64
```



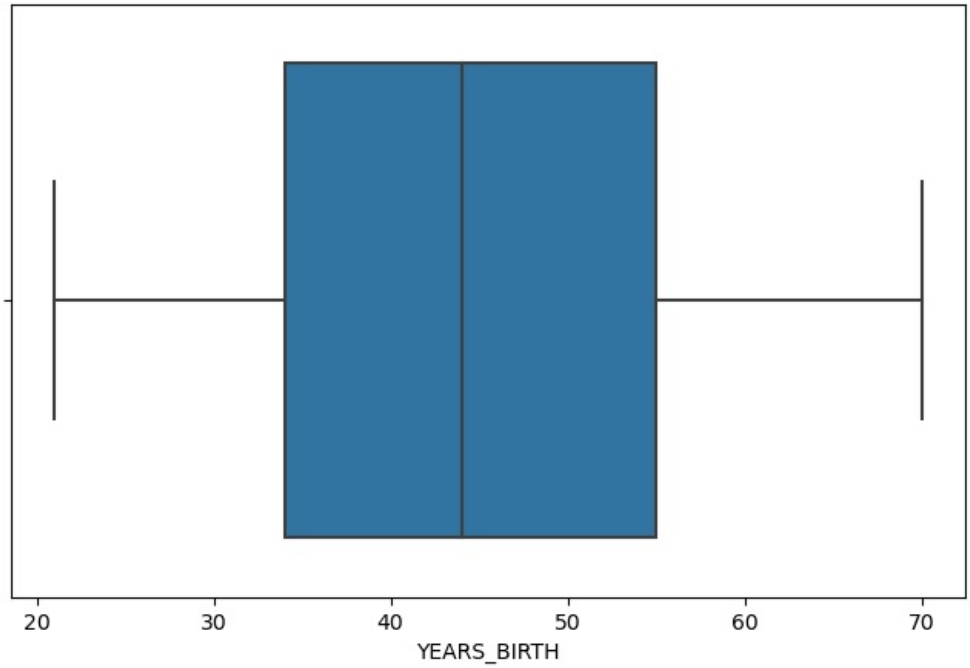

```
----------------------
count    307511.000000
mean          0.230454
std           0.421124
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: REG_CITY_NOT_WORK_CITY, dtype: float64
```



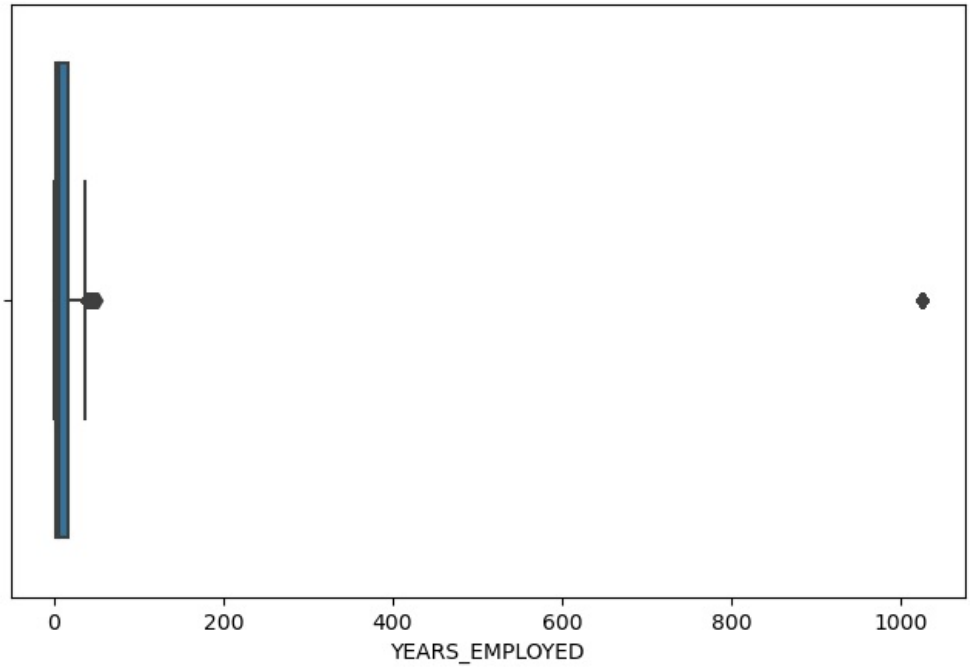

```
----------------------
count    307511.000000
mean          0.179555
std           0.383817
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           1.000000
Name: LIVE_CITY_NOT_WORK_CITY, dtype: float64
```



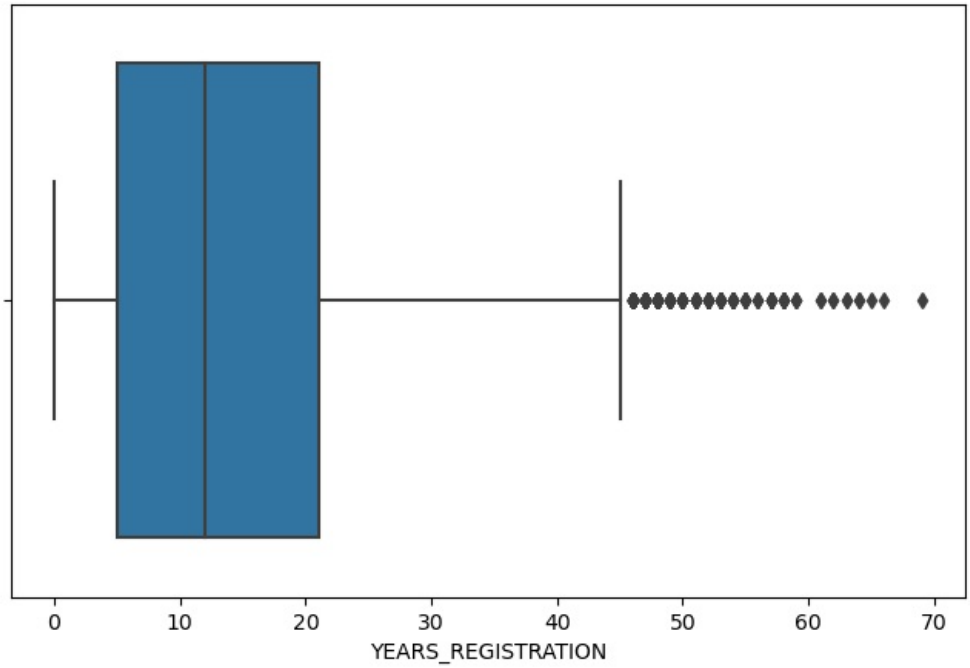

```
----------------------
count    307511.000000
mean         44.548992
std          12.263409
min          21.000000
25%          34.000000
50%          44.000000
75%          55.000000
max          70.000000
Name: YEARS_BIRTH, dtype: float64
```



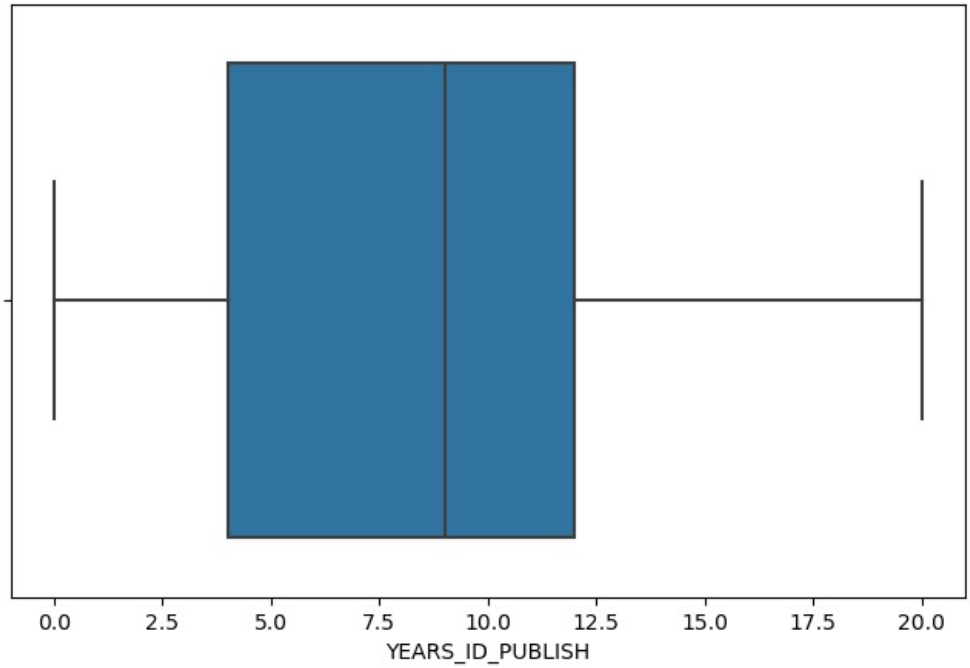

```
----------------------
count    307511.000000
mean        189.656025
std         391.517218
min           0.000000
25%           2.000000
50%           6.000000
75%          16.000000
max        1025.000000
Name: YEARS_EMPLOYED, dtype: float64
```



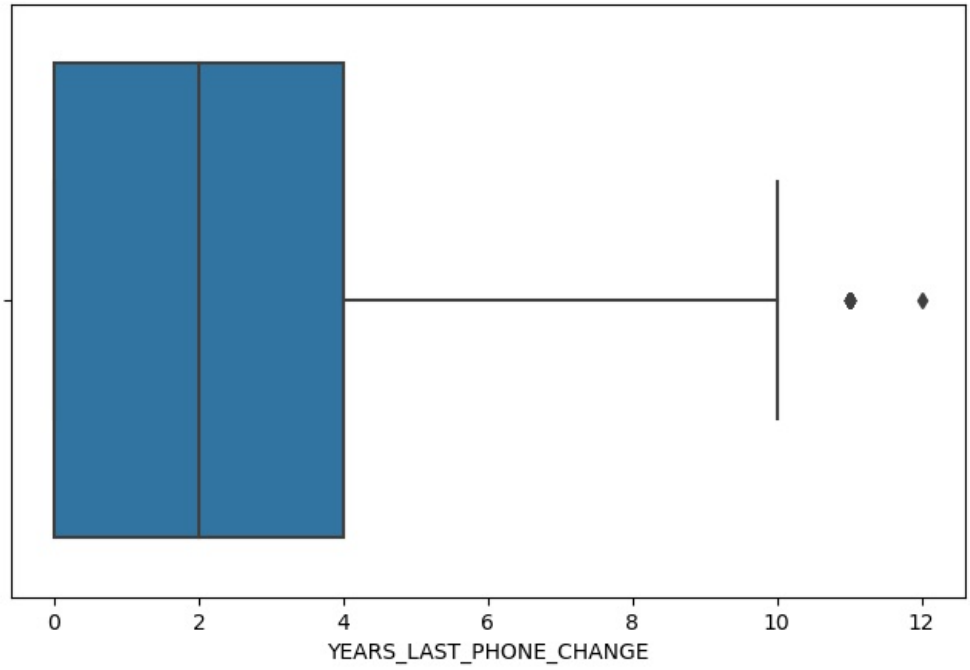

```
----------------------
count    307511.000000
mean         13.513478
std           9.891137
min           0.000000
25%           5.000000
50%          12.000000
75%          21.000000
max          69.000000
Name: YEARS_REGISTRATION, dtype: float64
```



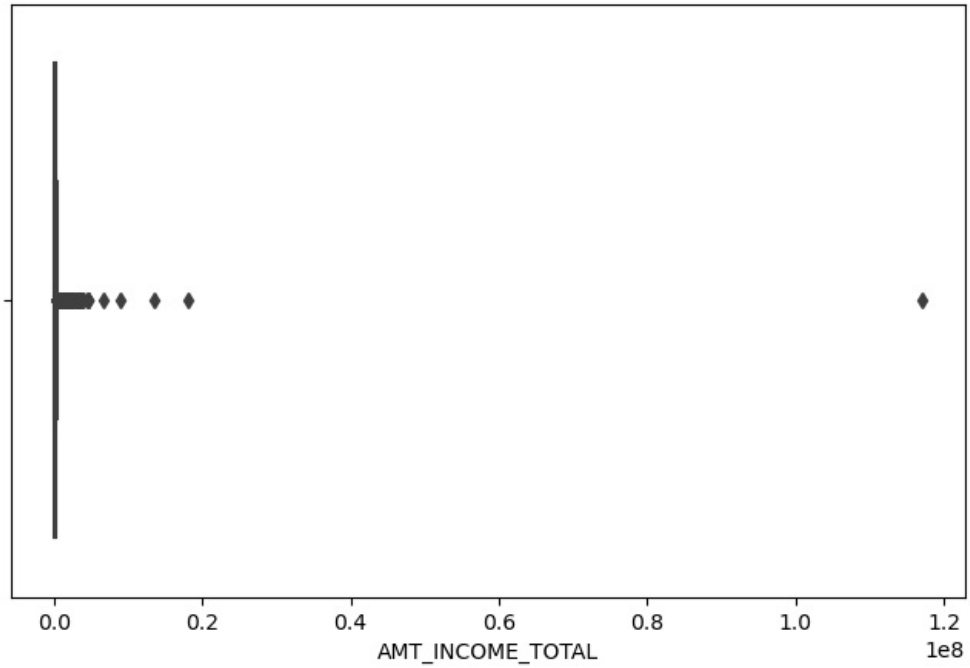

```
---------------------
count    307511.000000
mean          7.920845
std           4.238167
min           0.000000
25%           4.000000
50%           9.000000
75%          12.000000
max          20.000000
Name: YEARS_ID_PUBLISH, dtype: float64
```



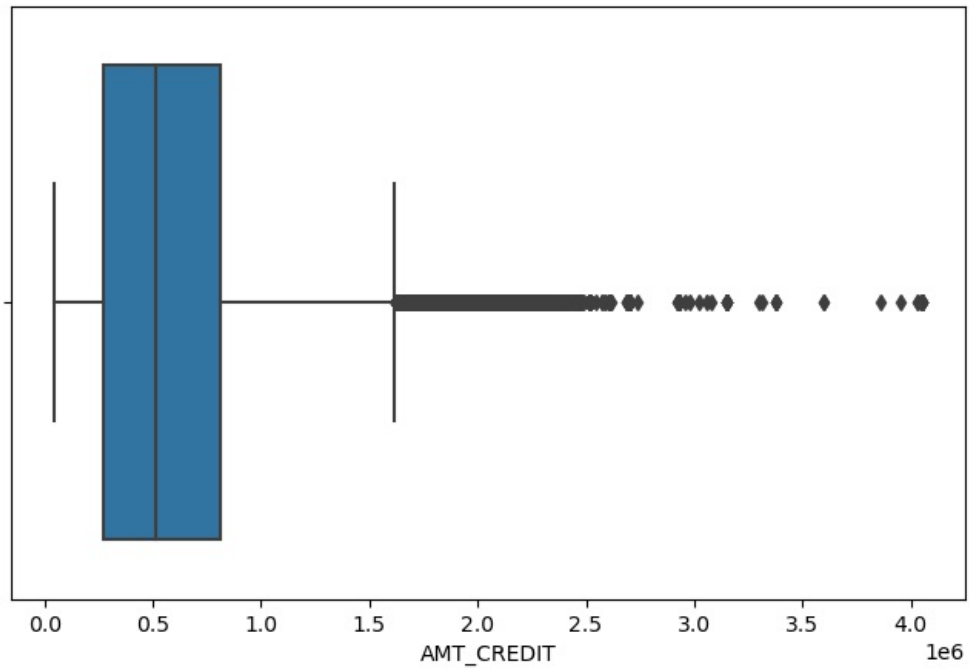

```
---------------------
count    307511.000000
mean          2.293102
std           2.249671
min           0.000000
25%           0.000000
50%           2.000000
75%           4.000000
max          12.000000
Name: YEARS_LAST_PHONE_CHANGE, dtype: float64
```



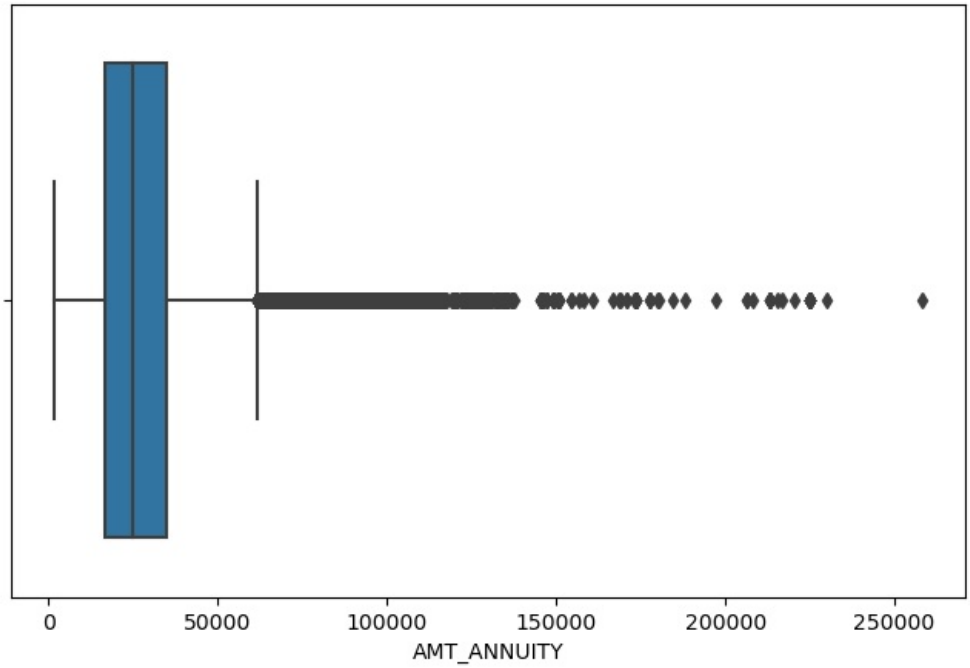

```
---------------------
count    3.075110e+05
mean     1.687979e+05
std      2.371231e+05
min      2.565000e+04
25%      1.125000e+05
50%      1.471500e+05
75%      2.025000e+05
max      1.170000e+08
Name: AMT_INCOME_TOTAL, dtype: float64
```



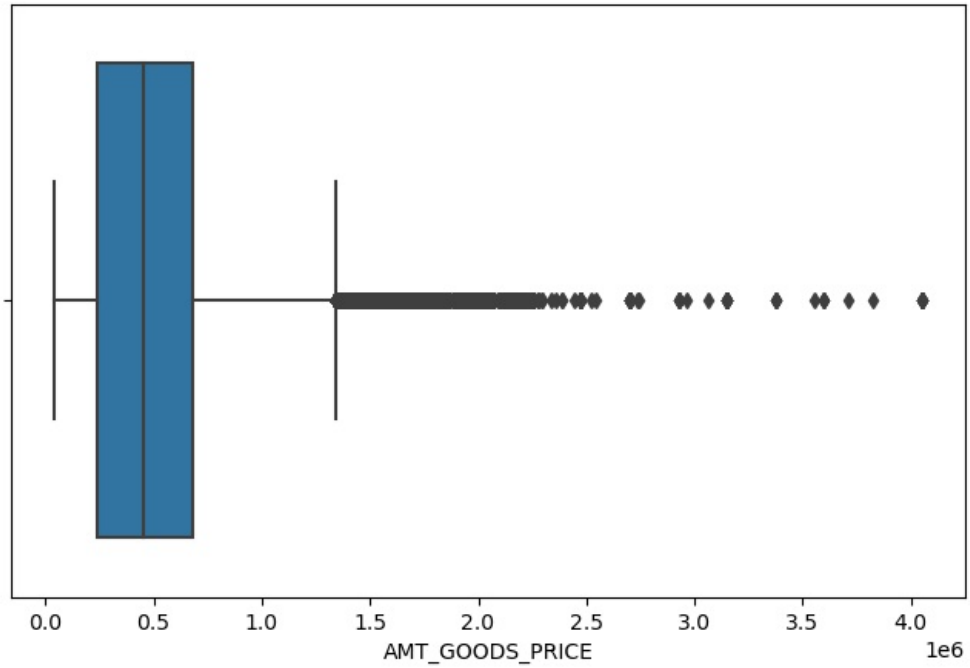

```
---------------------
count    3.075110e+05
mean     5.990260e+05
std      4.024908e+05
min      4.500000e+04
25%      2.700000e+05
50%      5.135310e+05
75%      8.086500e+05
max      4.050000e+06
Name: AMT_CREDIT, dtype: float64
```



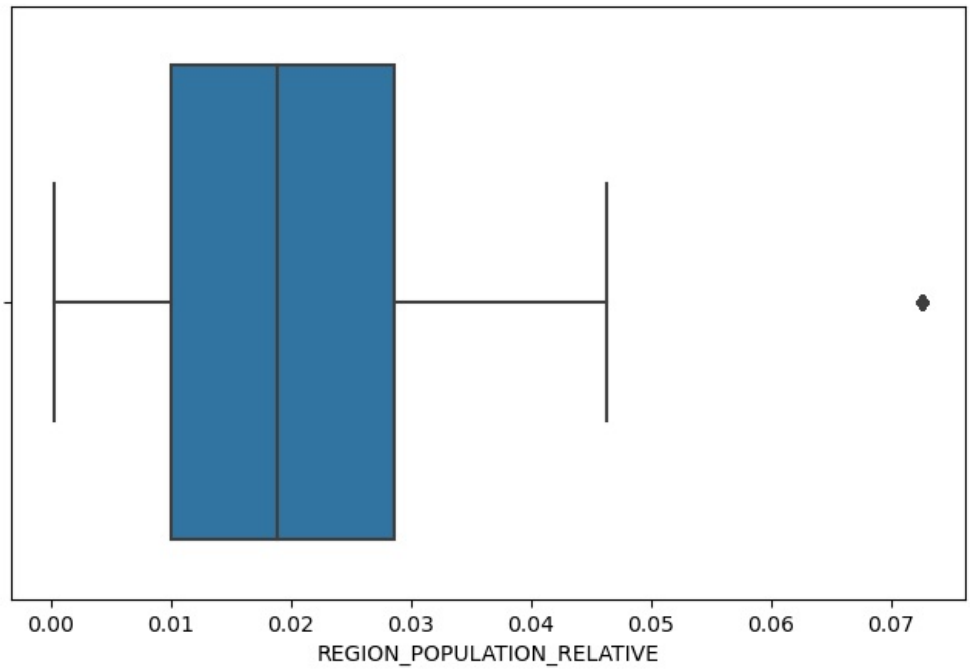

```
---------------------
count    307511.000000
mean      27108.487841
std       14493.461065
min        1615.500000
25%       16524.000000
50%       24903.000000
75%       34596.000000
max      258025.500000
Name: AMT_ANNUITY, dtype: float64
```



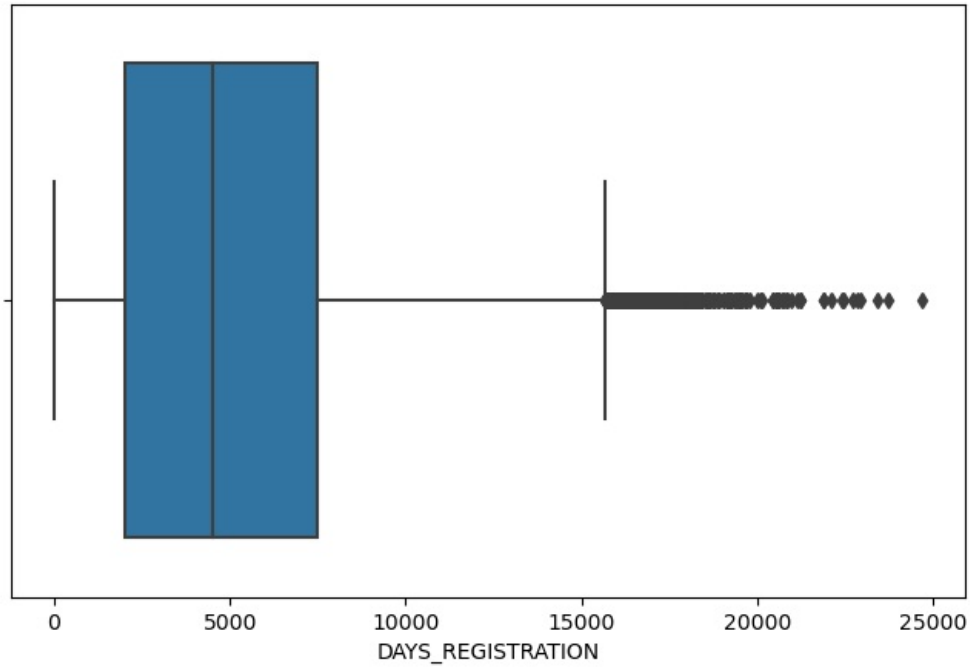

```
---------------------
count    3.075110e+05
mean     5.383163e+05
std      3.692890e+05
min      4.050000e+04
25%      2.385000e+05
50%      4.500000e+05
75%      6.795000e+05
max      4.050000e+06
Name: AMT_GOODS_PRICE, dtype: float64
```



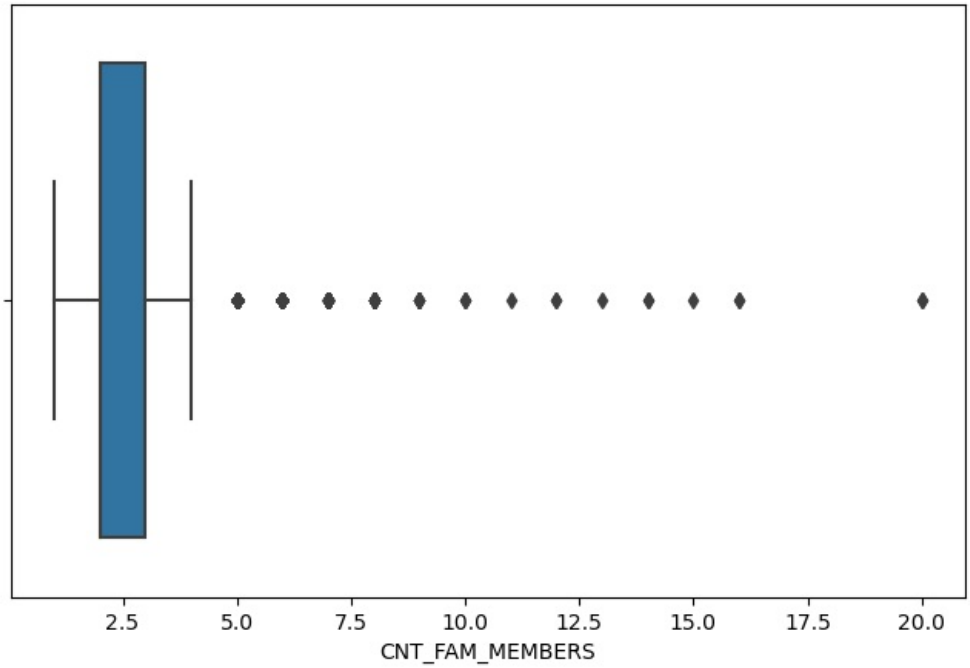

```
----------------------
count    307511.000000
mean          0.020868
std           0.013831
min           0.000290
25%           0.010006
50%           0.018850
75%           0.028663
max           0.072508
Name: REGION_POPULATION_RELATIVE, dtype: float64
```



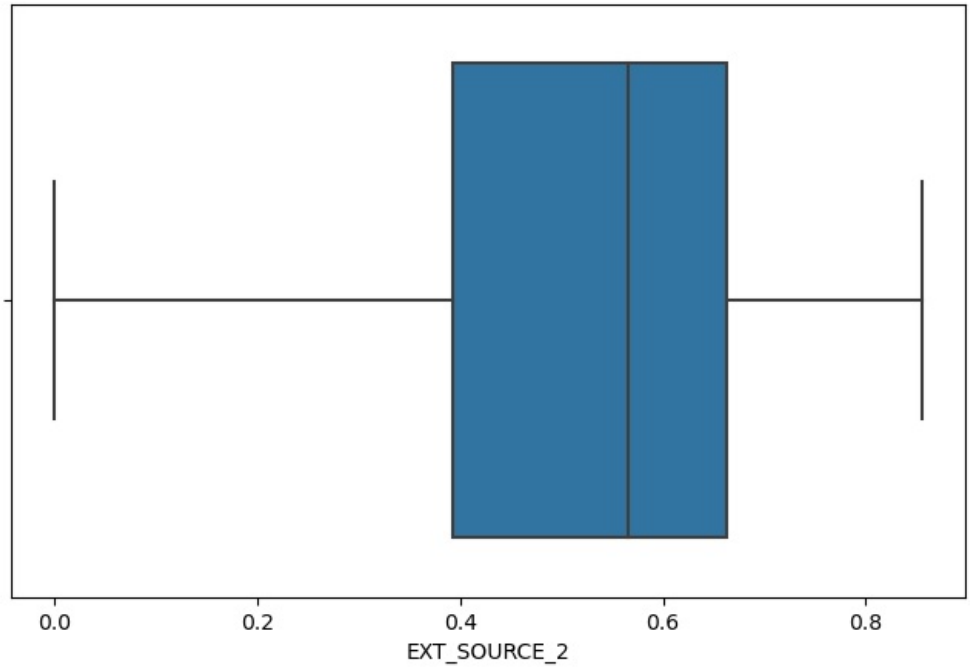

```
----------------------
count    307511.000000
mean       4986.120328
std        3522.886321
min           0.000000
25%        2010.000000
50%        4504.000000
75%        7479.500000
max       24672.000000
Name: DAYS_REGISTRATION, dtype: float64
```



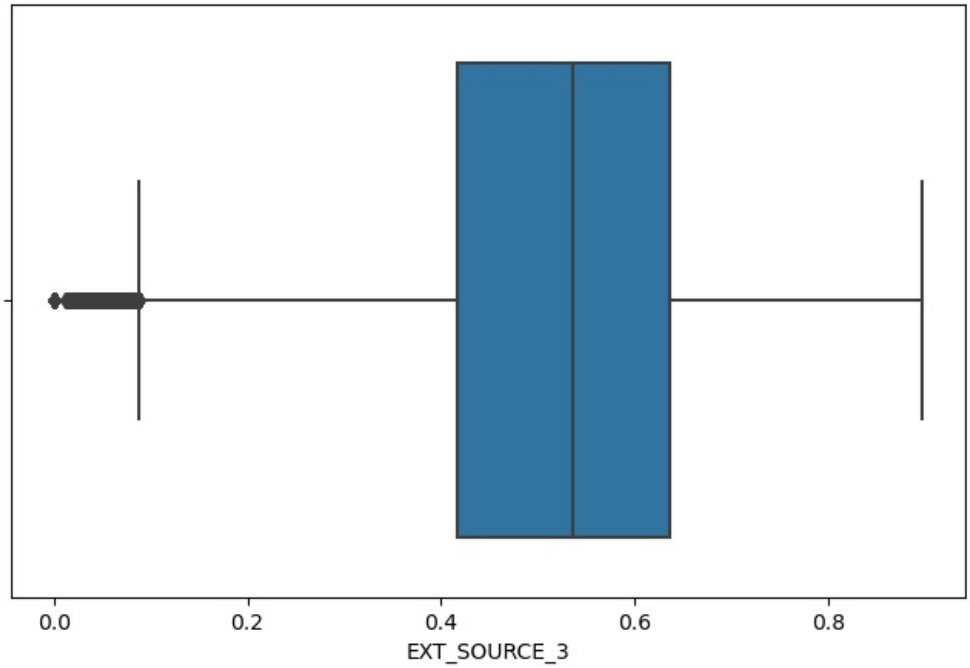

```
---------------------
count    307511.000000
mean          2.152664
std           0.910679
min           1.000000
25%           2.000000
50%           2.000000
75%           3.000000
max          20.000000
Name: CNT_FAM_MEMBERS, dtype: float64
```



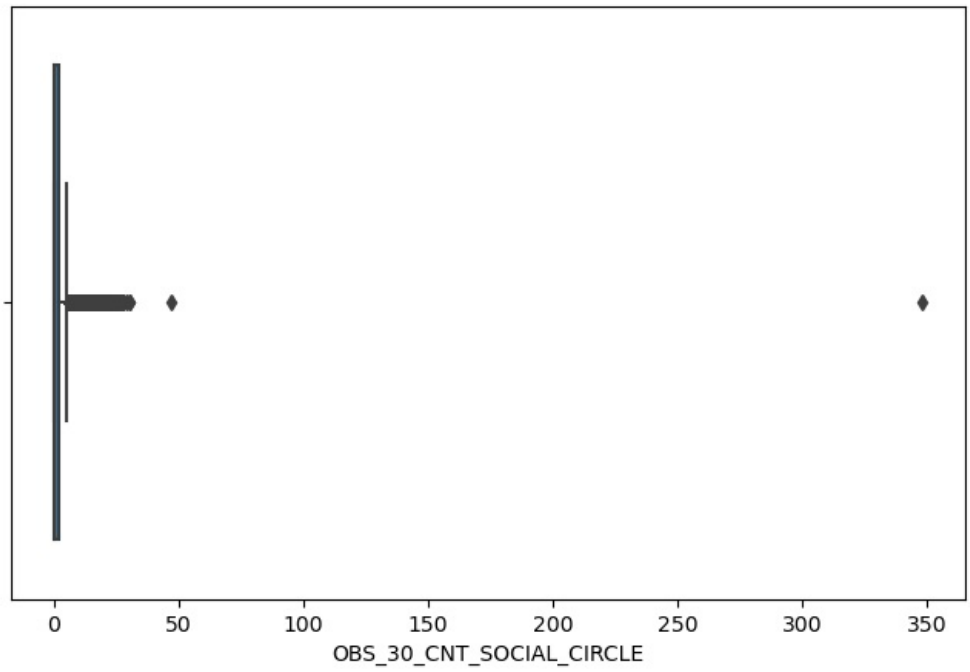

```
---------------------
count    3.075110e+05
mean     5.145034e-01
std      1.908699e-01
min      8.173617e-08
25%      3.929737e-01
50%      5.659614e-01
75%      6.634218e-01
max      8.549997e-01
Name: EXT_SOURCE_2, dtype: float64
```



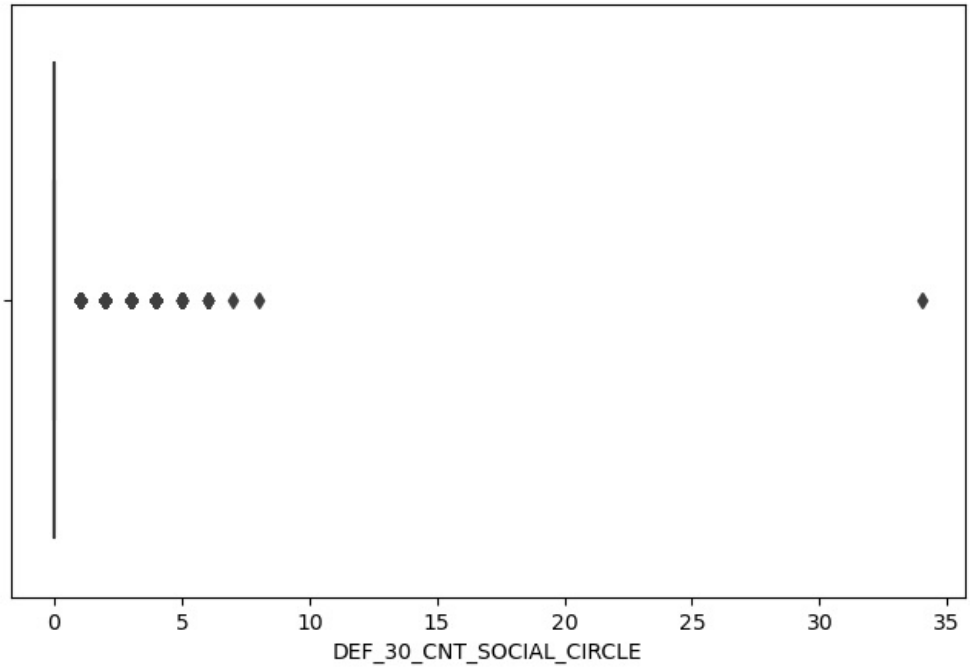

```
----------------------
count    307511.000000
mean          0.515695
std           0.174736
min           0.000527
25%           0.417100
50%           0.535276
75%           0.636376
max           0.896010
Name: EXT_SOURCE_3, dtype: float64
```



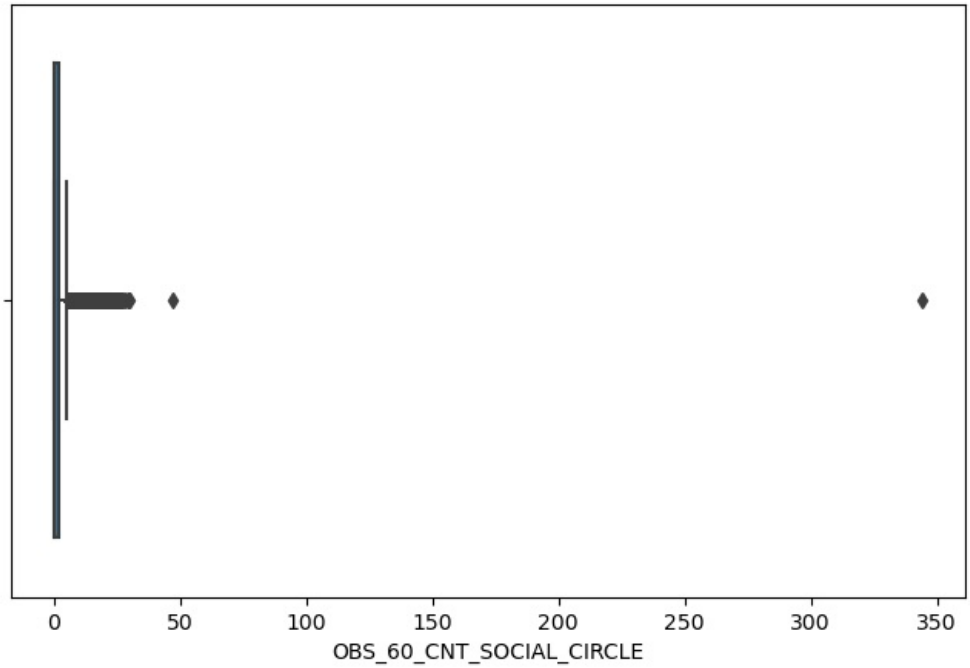

```
----------------------
count    307511.000000
mean          1.417523
std           2.398395
min           0.000000
25%           0.000000
50%           0.000000
75%           2.000000
max         348.000000
Name: OBS_30_CNT_SOCIAL_CIRCLE, dtype: float64
```



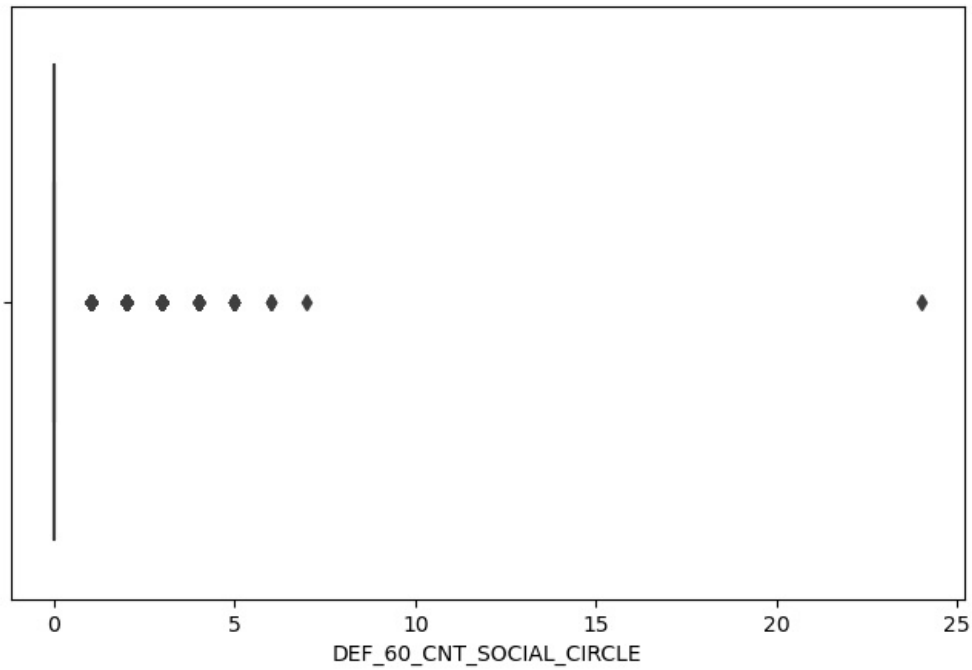

```
----------------------
count    307511.000000
mean          0.142944
std           0.446033
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          34.000000
Name: DEF_30_CNT_SOCIAL_CIRCLE, dtype: float64
```



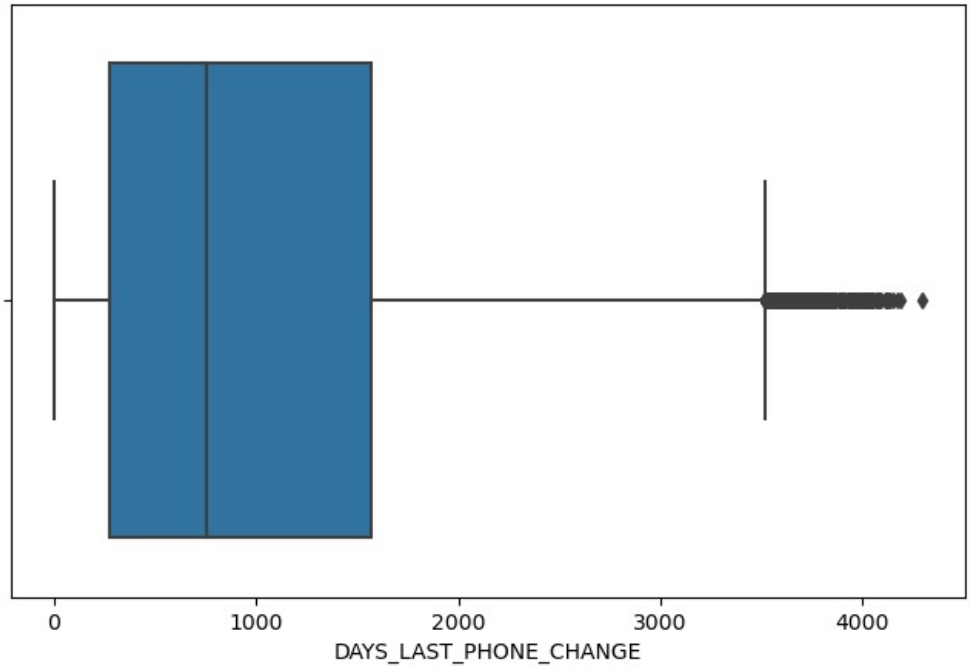

```
----------------------
count    307511.000000
mean          1.400626
std           2.377224
min           0.000000
25%           0.000000
50%           0.000000
75%           2.000000
max         344.000000
Name: OBS_60_CNT_SOCIAL_CIRCLE, dtype: float64
```



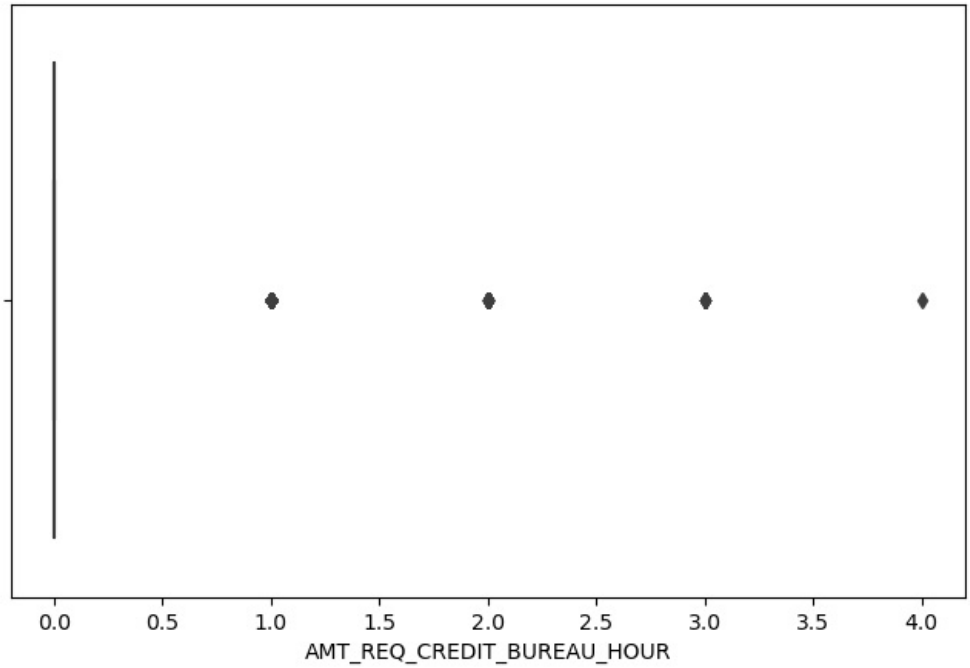

```
----------------------
count    307511.000000
mean          0.099717
std           0.361735
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          24.000000
Name: DEF_60_CNT_SOCIAL_CIRCLE, dtype: float64
```



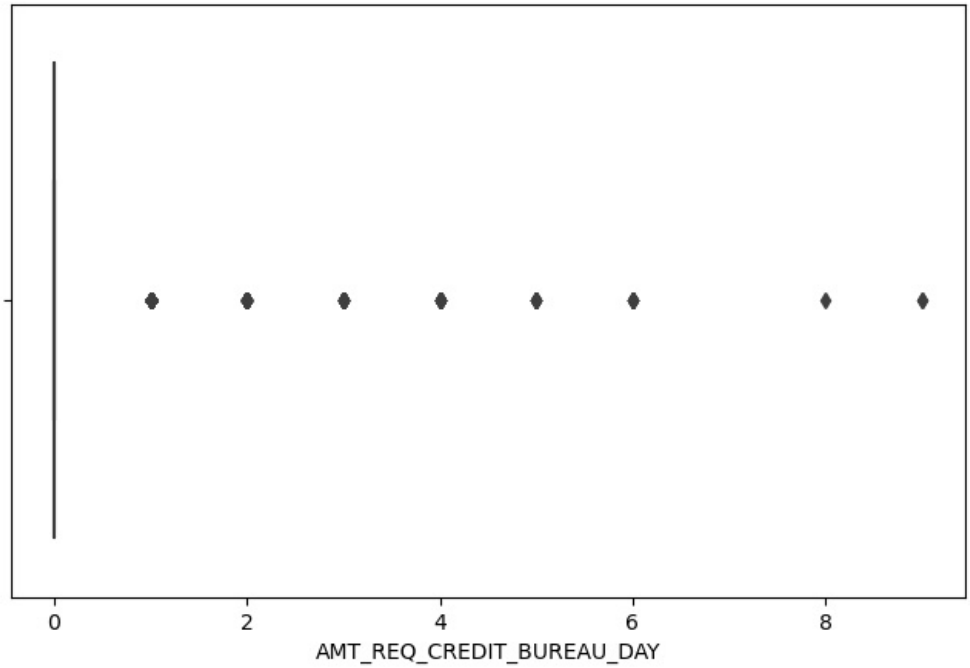

```
----------------------
count    307511.000000
mean        962.858119
std         826.807226
min           0.000000
25%         274.000000
50%         757.000000
75%        1570.000000
max        4292.000000
Name: DAYS_LAST_PHONE_CHANGE, dtype: float64
```



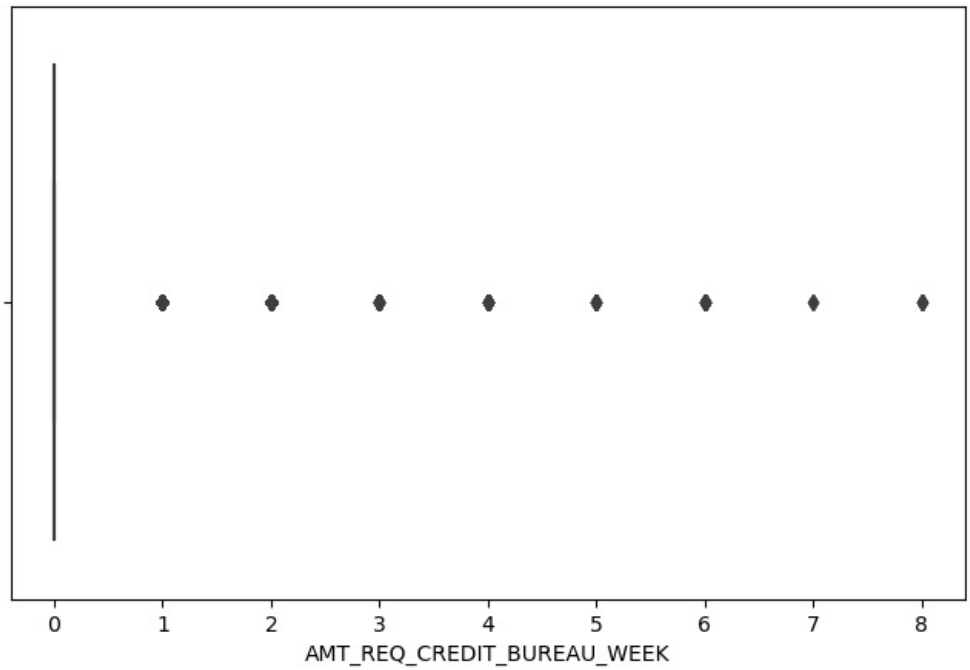

```
----------------------
count    307511.000000
mean          0.005538
std           0.078014
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           4.000000
Name: AMT_REQ_CREDIT_BUREAU_HOUR, dtype: float64
```



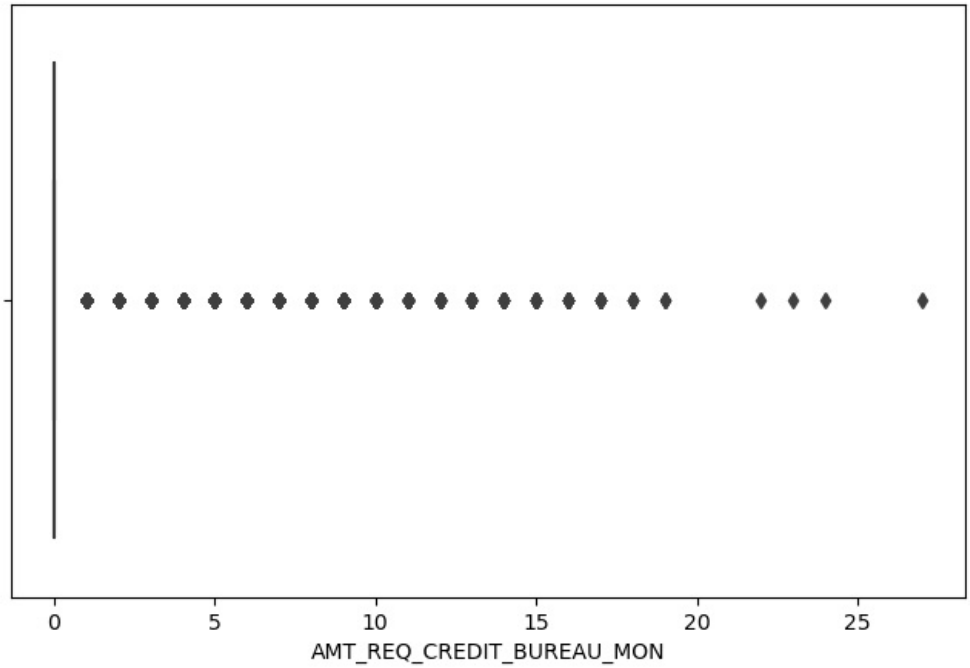

```
----------------------
count    307511.000000
mean          0.006055
std           0.103037
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           9.000000
Name: AMT_REQ_CREDIT_BUREAU_DAY, dtype: float64
```



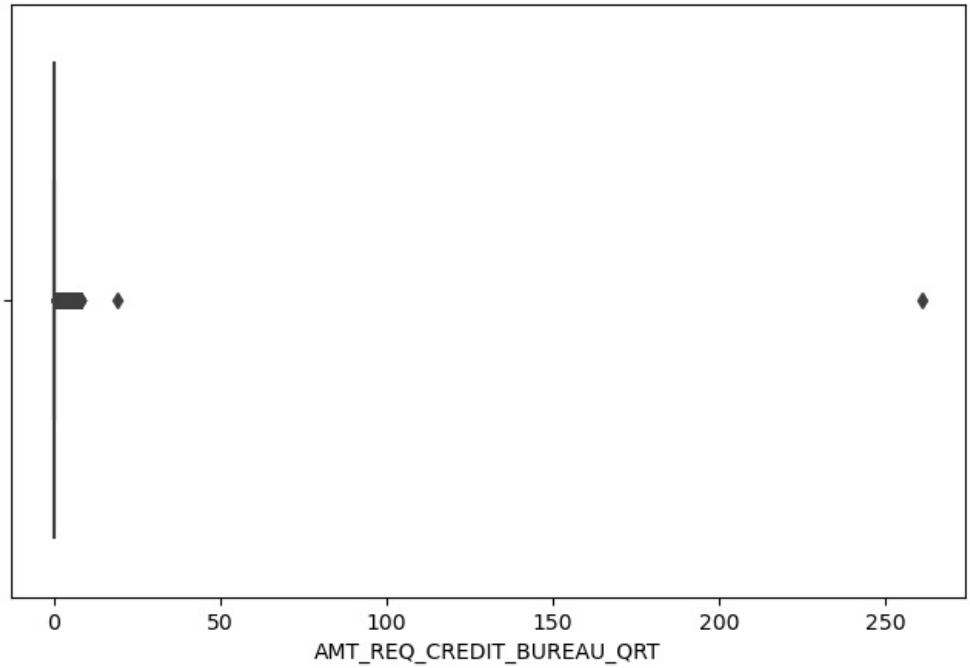

```
---------------------
count    307511.000000
mean          0.029723
std           0.190728
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           8.000000
Name: AMT_REQ_CREDIT_BUREAU_WEEK, dtype: float64
```



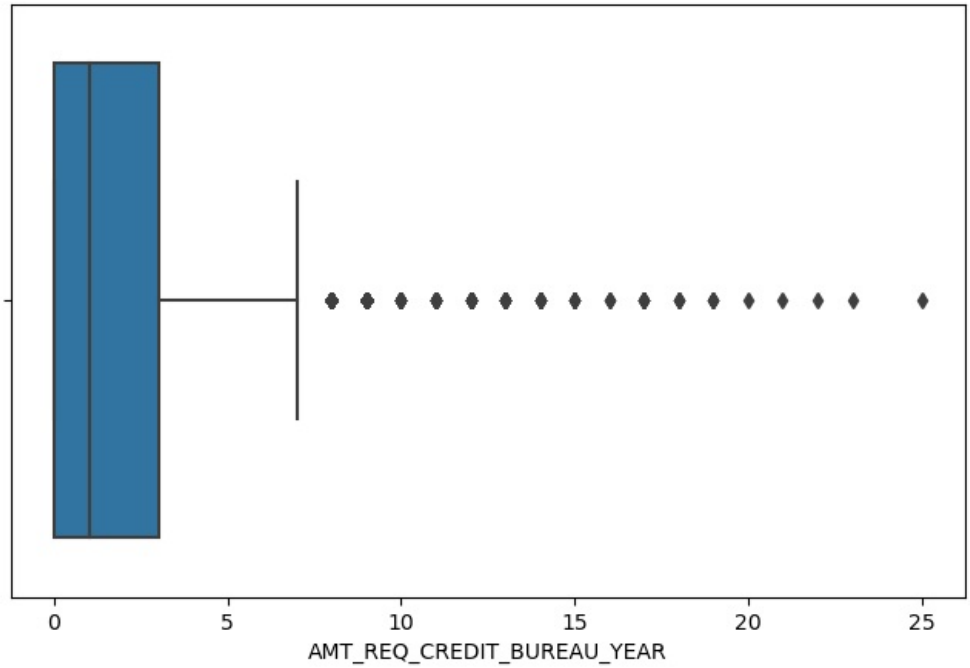

```
---------------------
count    307511.000000
mean          0.231293
std           0.856810
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max          27.000000
Name: AMT_REQ_CREDIT_BUREAU_MON, dtype: float64
```




```
---------------------
count    307511.000000
mean          0.229631
std           0.744059
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max         261.000000
Name: AMT_REQ_CREDIT_BUREAU_QRT, dtype: float64
```




```
---------------------
count    307511.000000
mean          1.643447
std           1.855821
min           0.000000
25%           0.000000
50%           1.000000
75%           3.000000
max          25.000000
Name: AMT_REQ_CREDIT_BUREAU_YEAR, dtype: float64
```

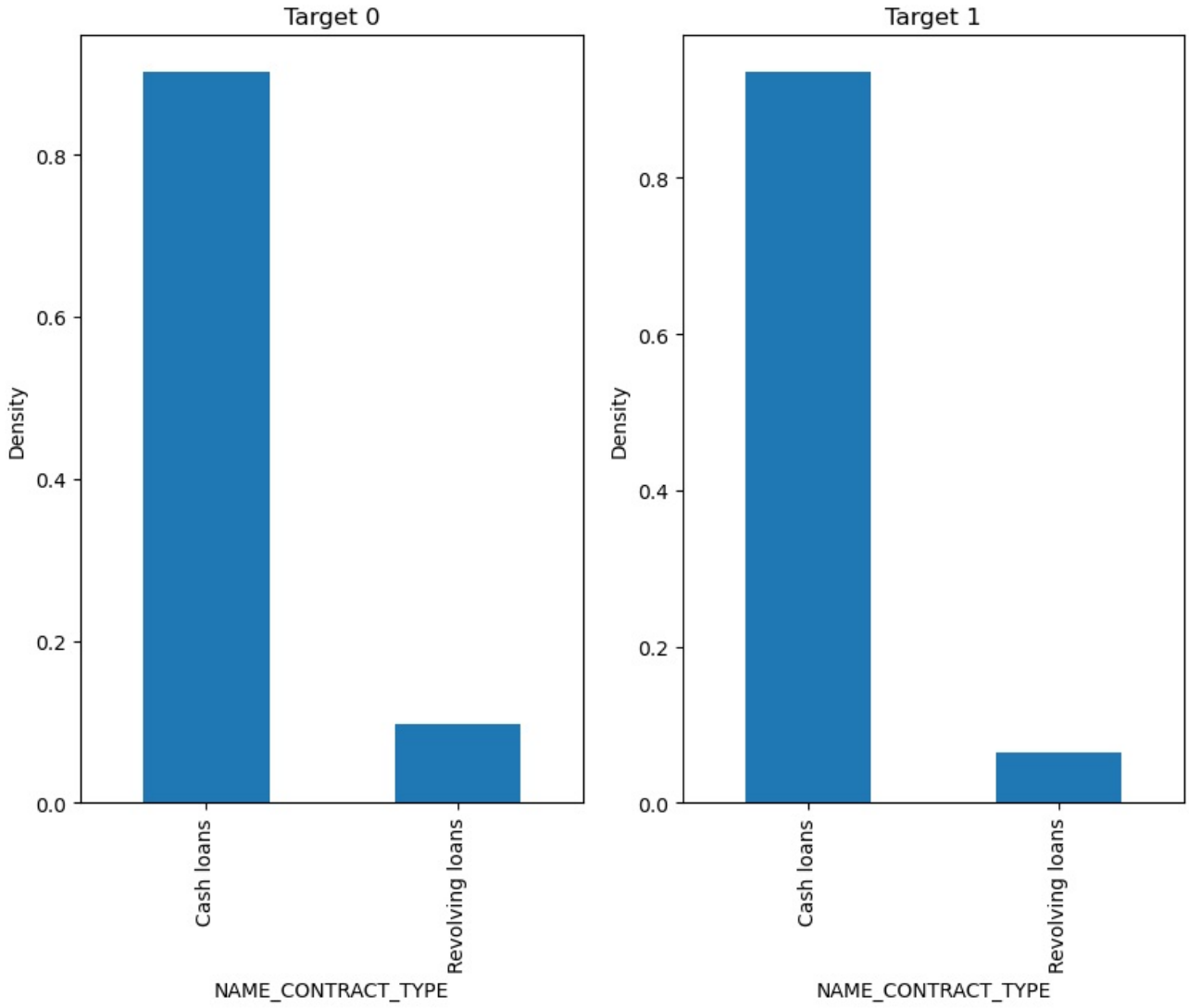


Univariate Analysis on columns With Target 0 and 1

In [112…

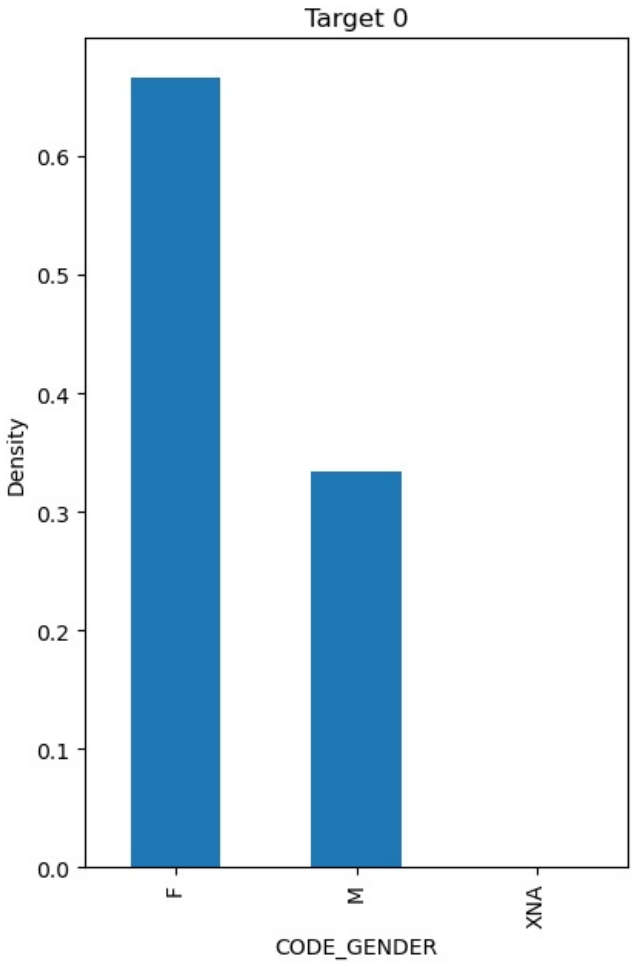
```
for col in cat_cols:
    print(f"plot on {col} for Target 0 and 1")
    plt.figure(figsize=[10,7])
    plt.subplot(1,2,1)
    tar_0[col].value_counts(normalize=True).plot.bar()
    plt.title("Target 0")
    plt.xlabel(col)
    plt.ylabel("Density")
    plt.subplot(1,2,2)
    tar_1[col].value_counts(normalize=True).plot.bar()
    plt.title("Target 1")
    plt.xlabel(col)
    plt.ylabel("Density")
    plt.show()
    print("\n\n-----------------------\n\n")
```

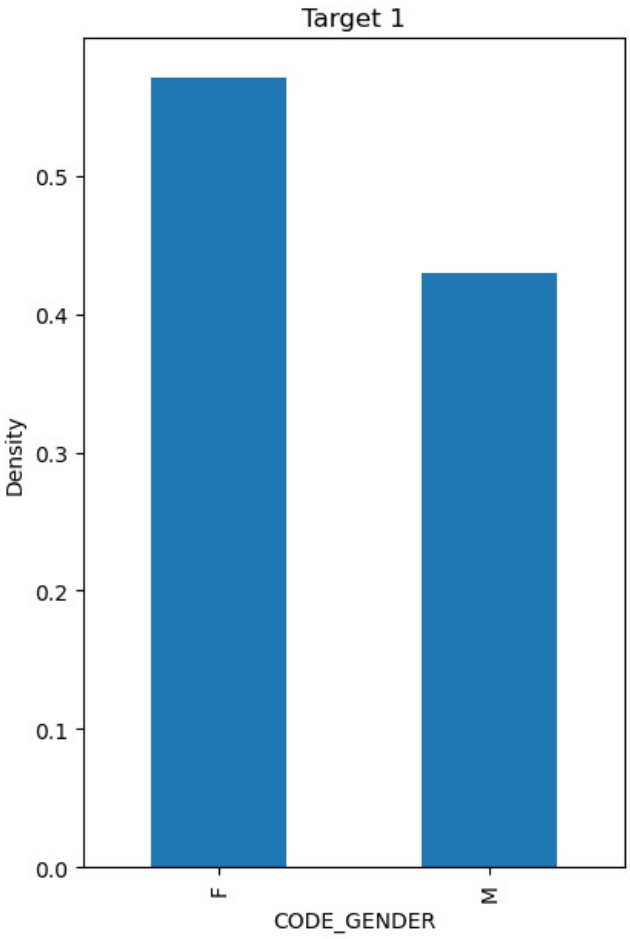
```
plot on NAME_CONTRACT_TYPE for Target 0 and 1
```



---

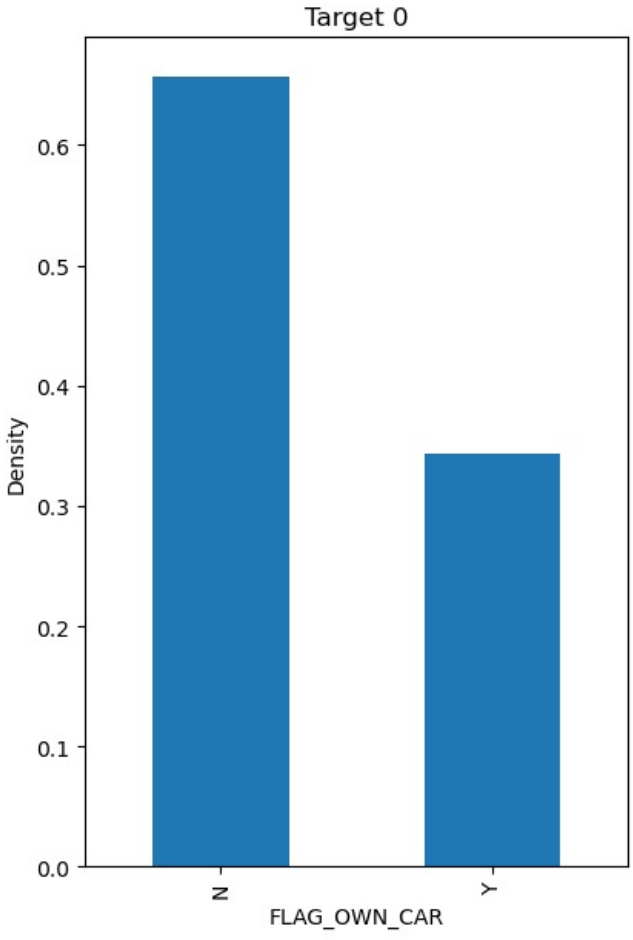
plot on CODE_GENDER for Target 0 and 1

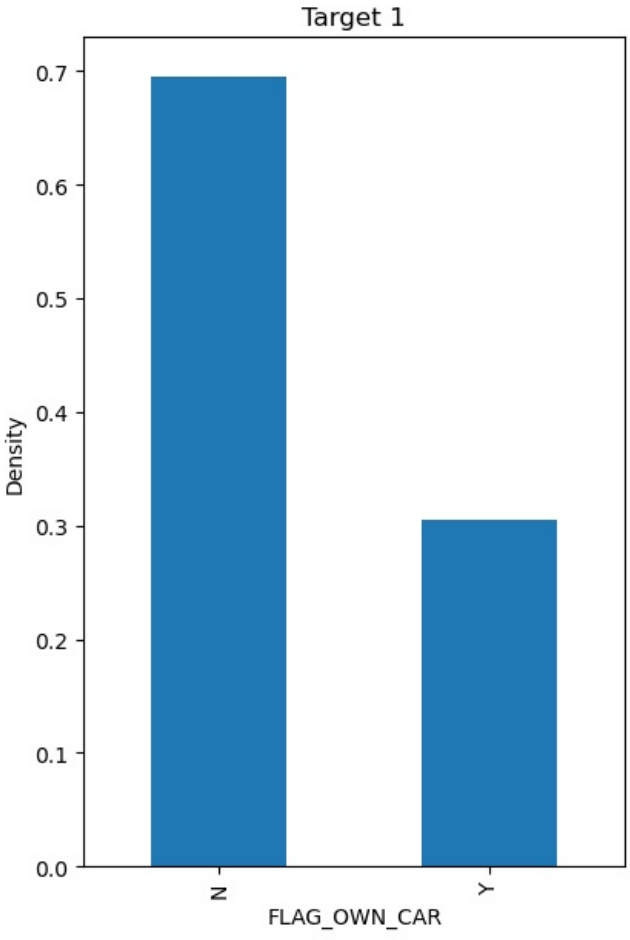




---

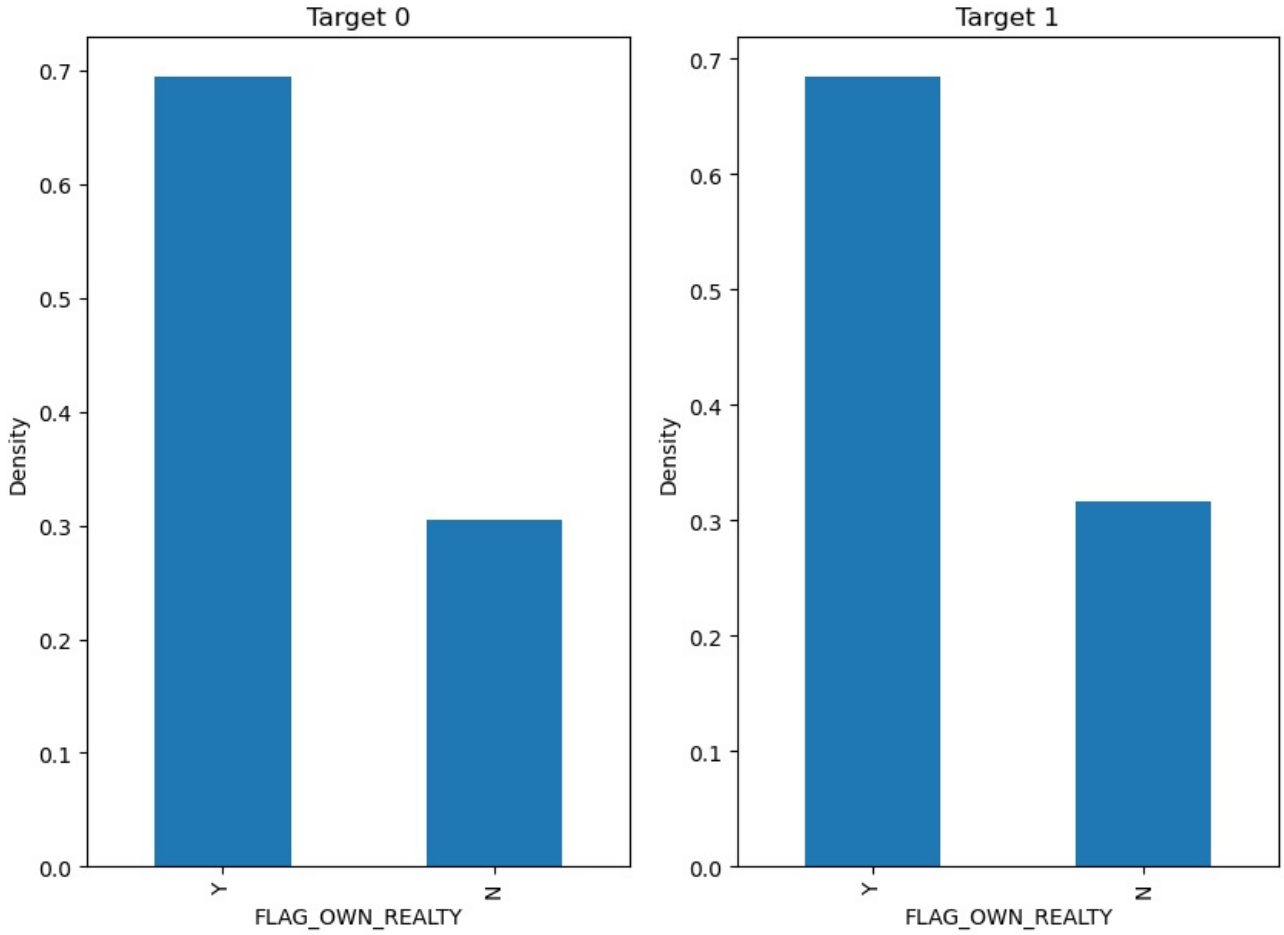
plot on FLAG_OWN_CAR for Target 0 and 1





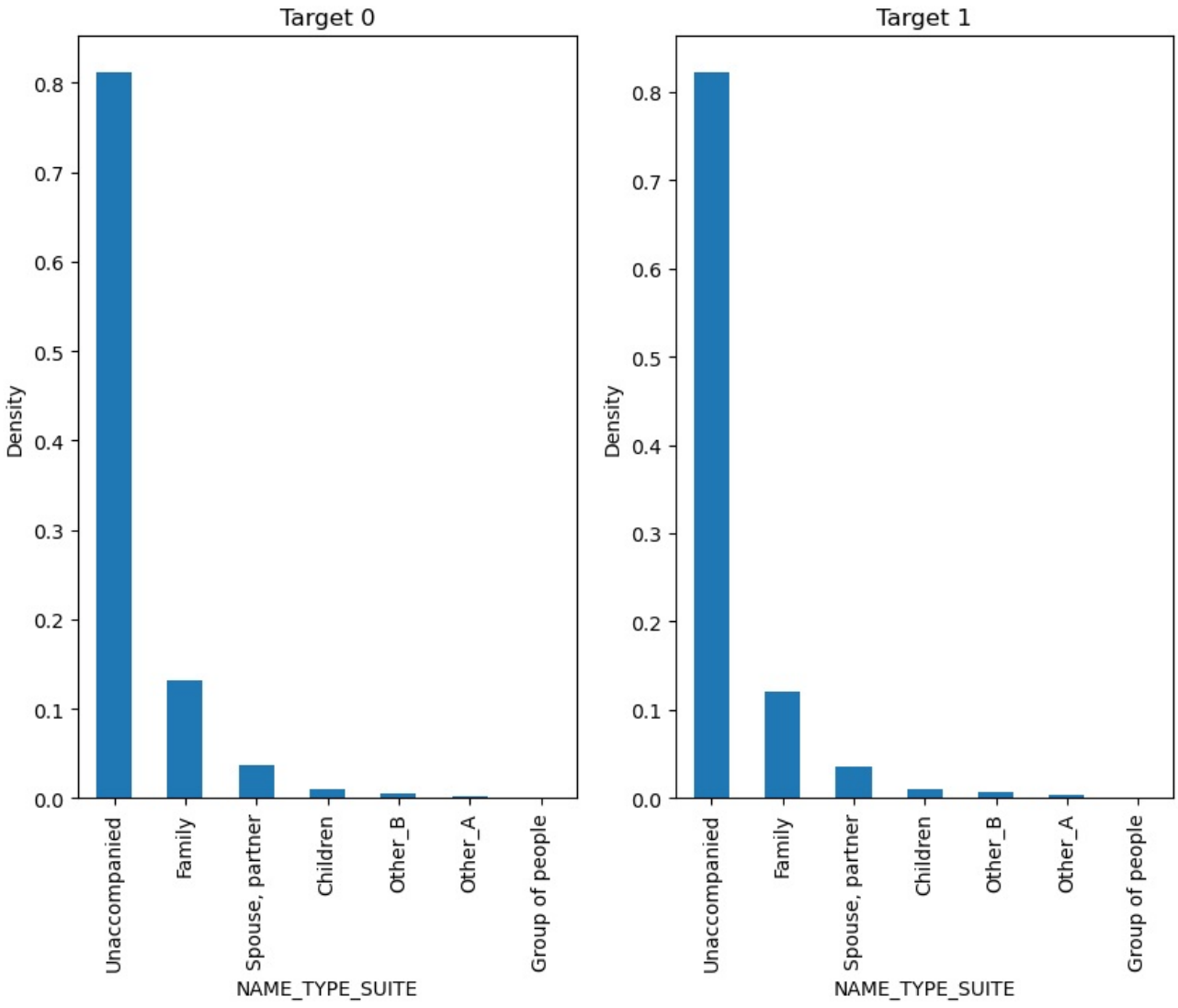
---

plot on FLAG_OWN_REALTY for Target 0 and 1



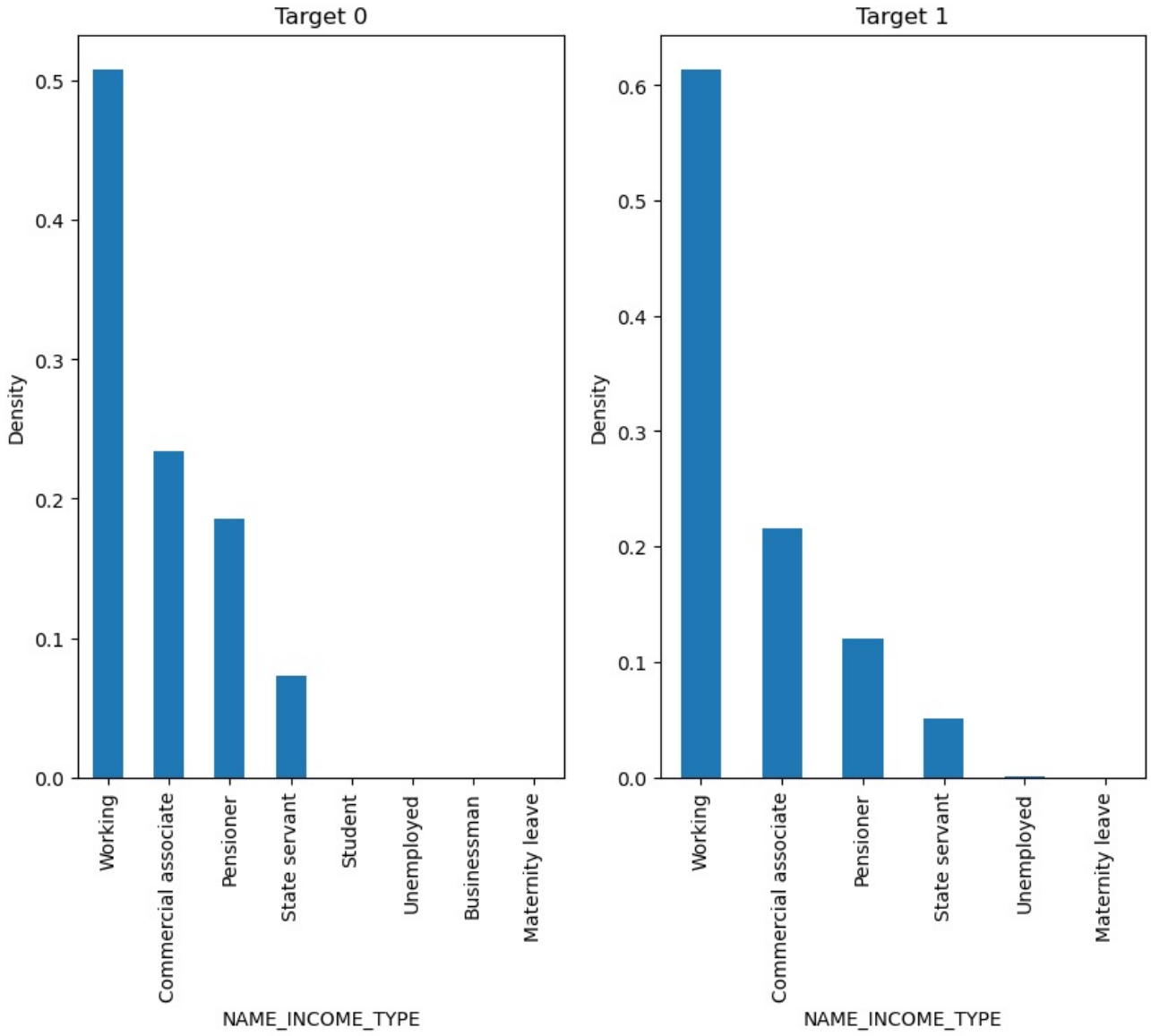
---

plot on NAME_TYPE_SUITE for Target 0 and 1


Target 0
Target 1
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Unaccompanied
Family
Spouse, partner
Children
Other_B
Other_A
Group of people
NAME_TYPE_SUITE

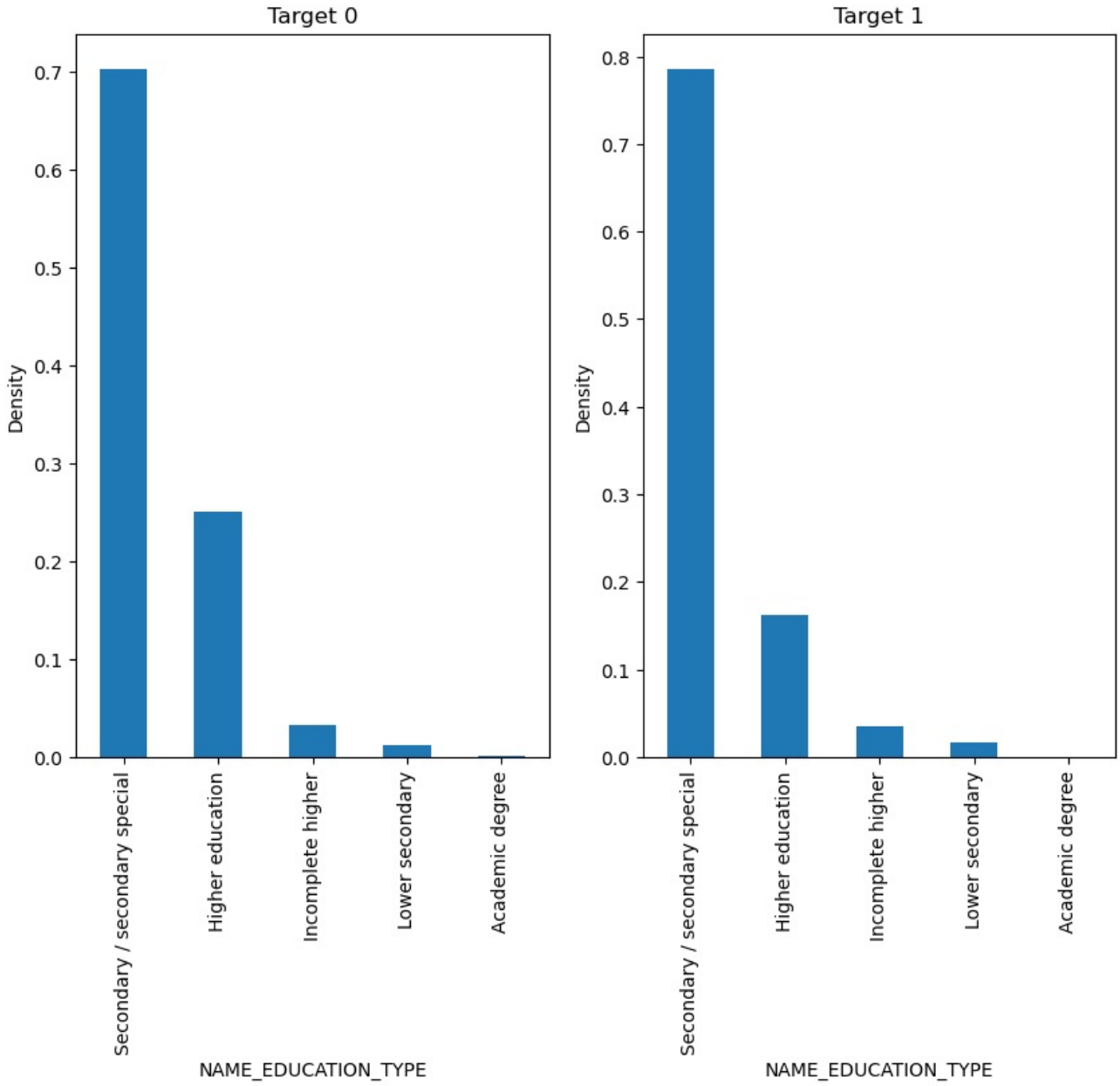
---

plot on NAME_INCOME_TYPE for Target 0 and 1



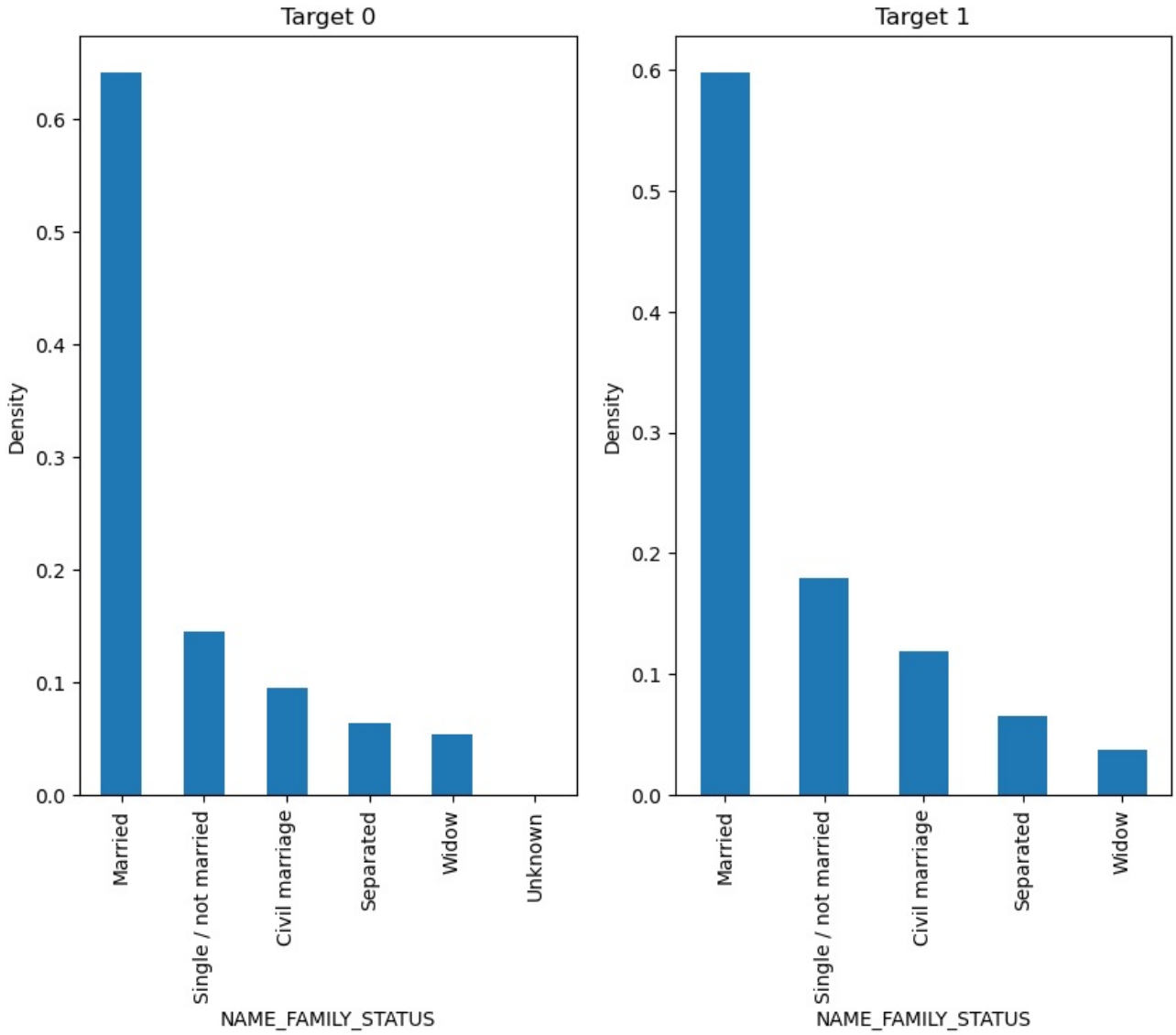
---

plot on NAME_EDUCATION_TYPE for Target 0 and 1



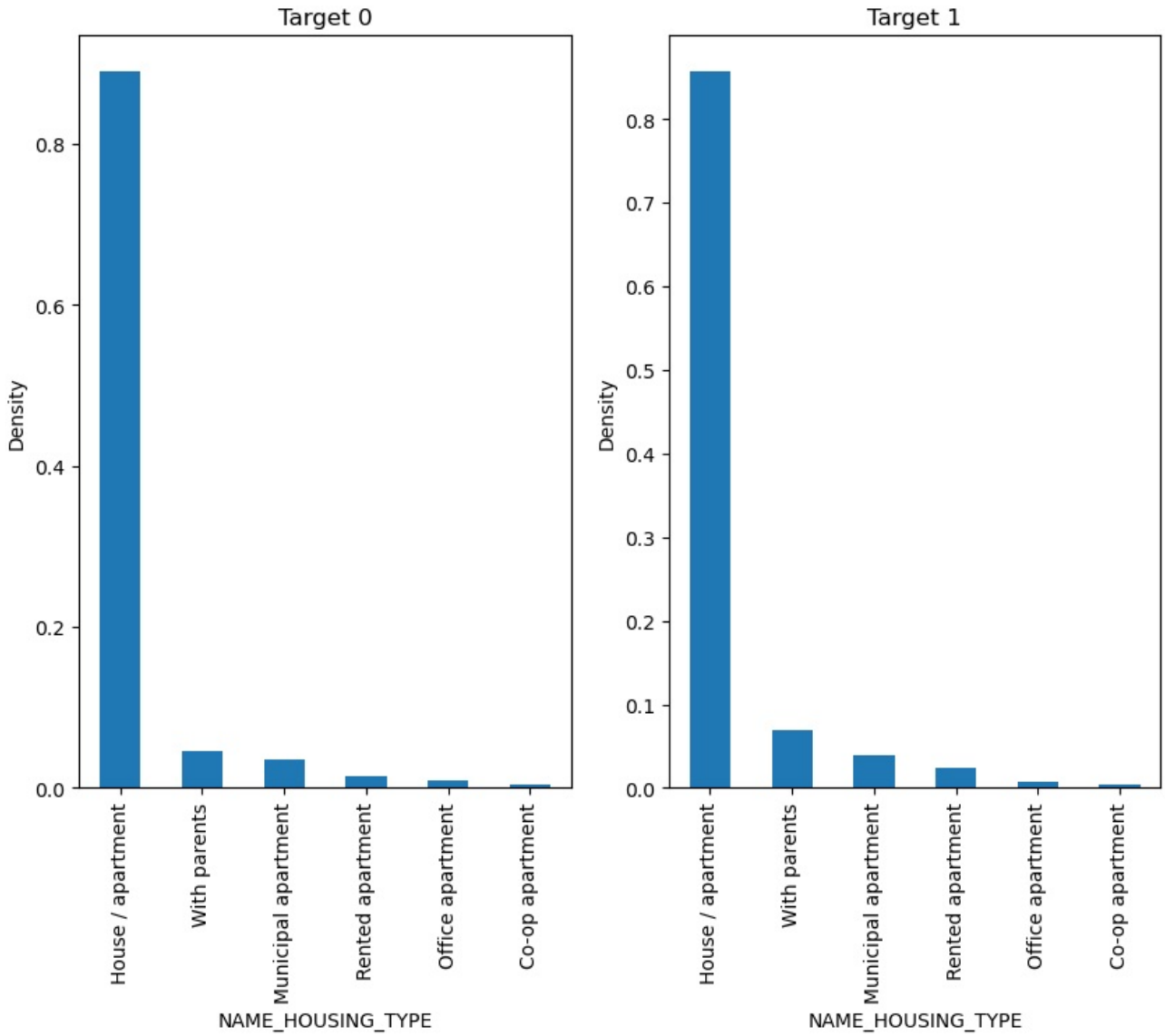
---

plot on NAME_FAMILY_STATUS for Target 0 and 1



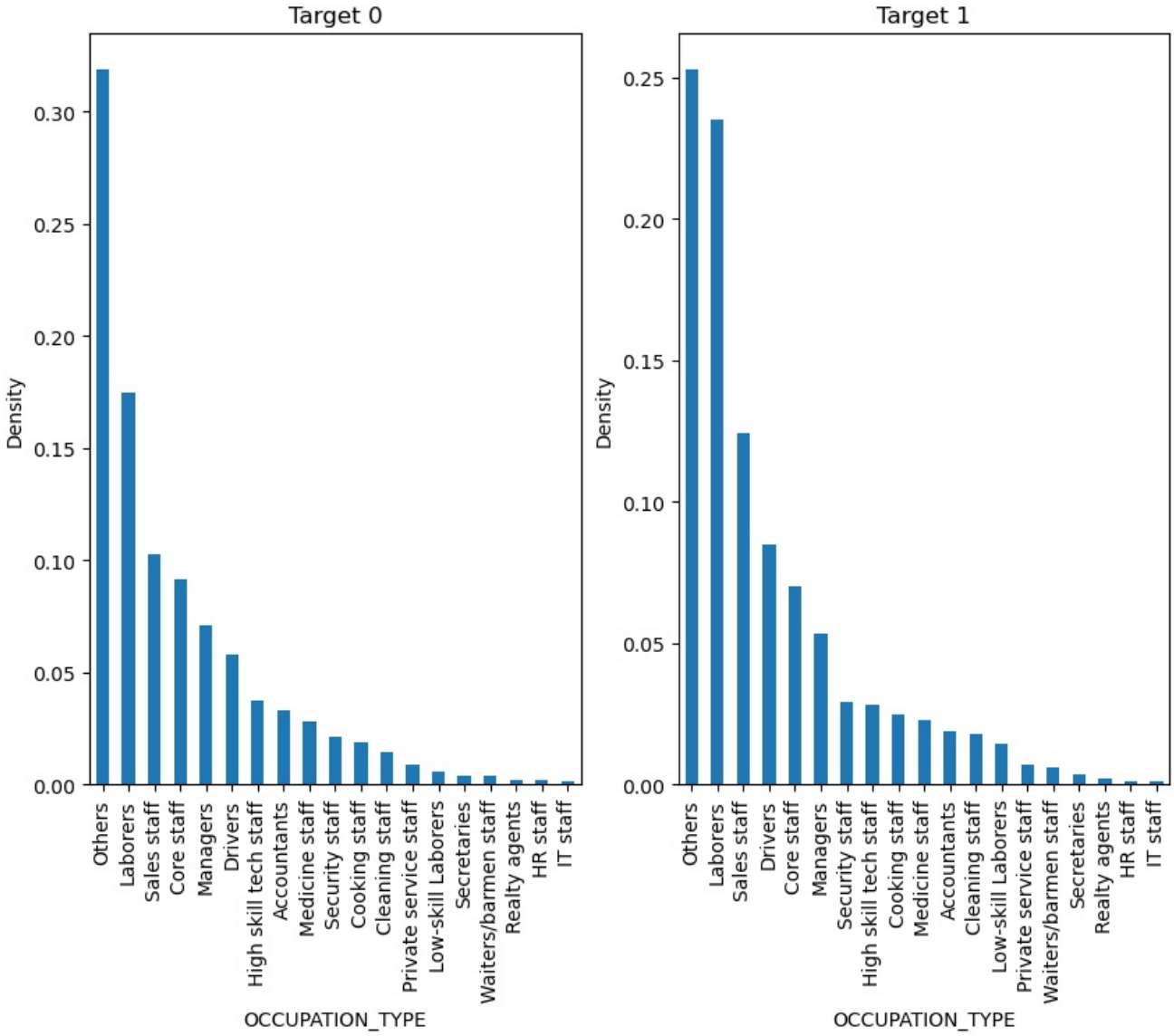
---

plot on NAME_HOUSING_TYPE for Target 0 and 1



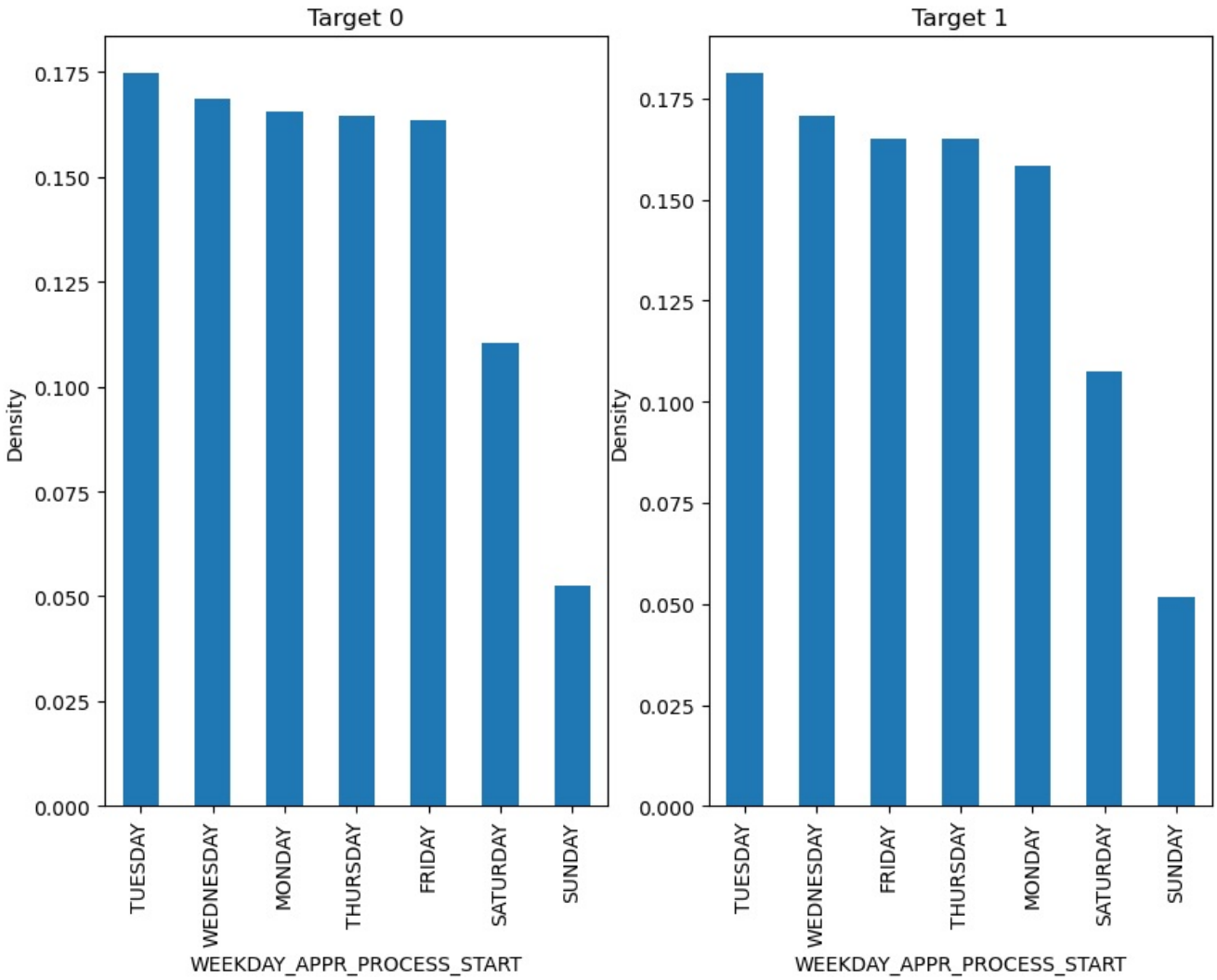
---

plot on OCCUPATION_TYPE for Target 0 and 1



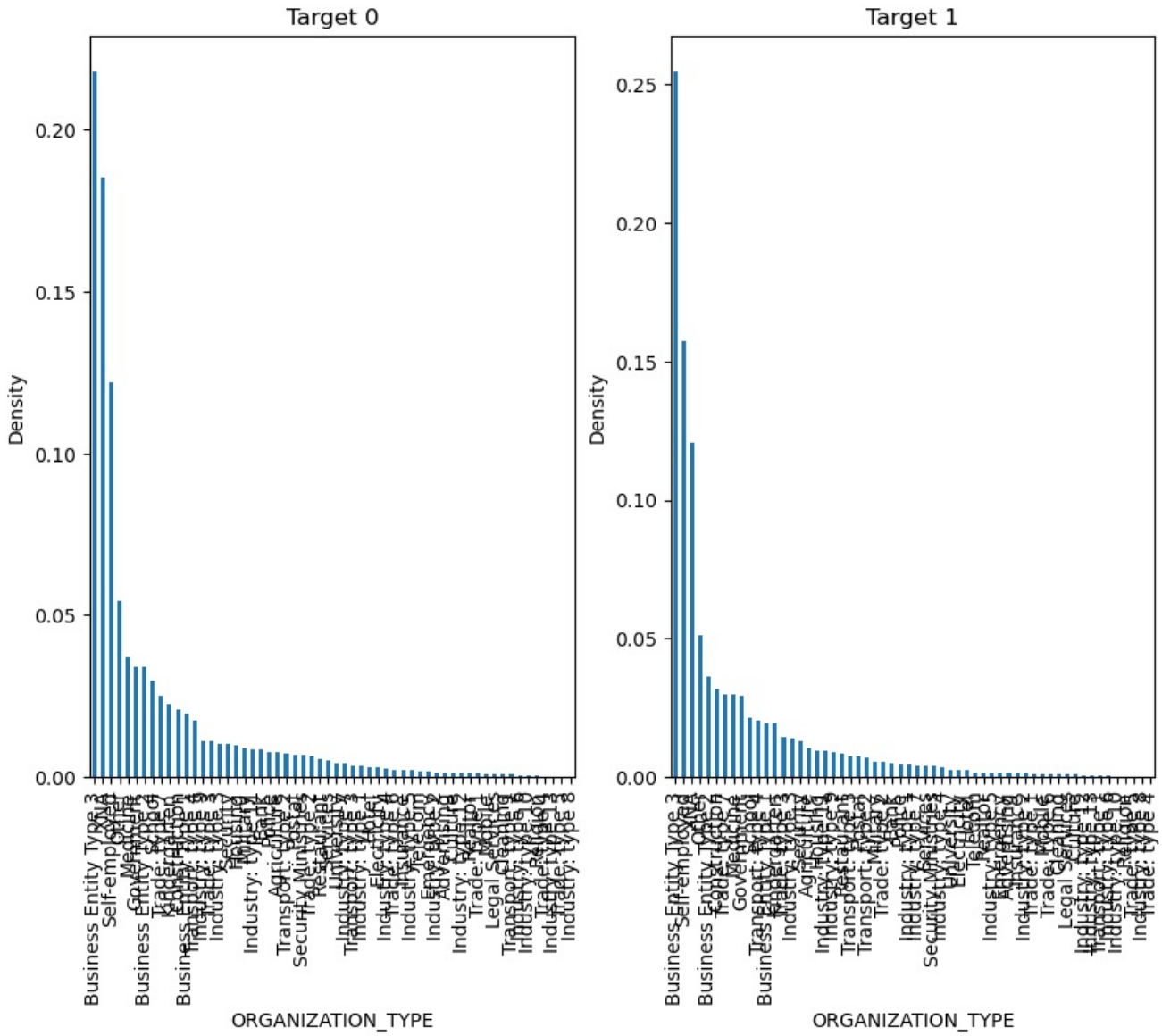
---

plot on WEEKDAY_APPR_PROCESS_START for Target 0 and 1



---

plot on ORGANIZATION_TYPE for Target 0 and 1

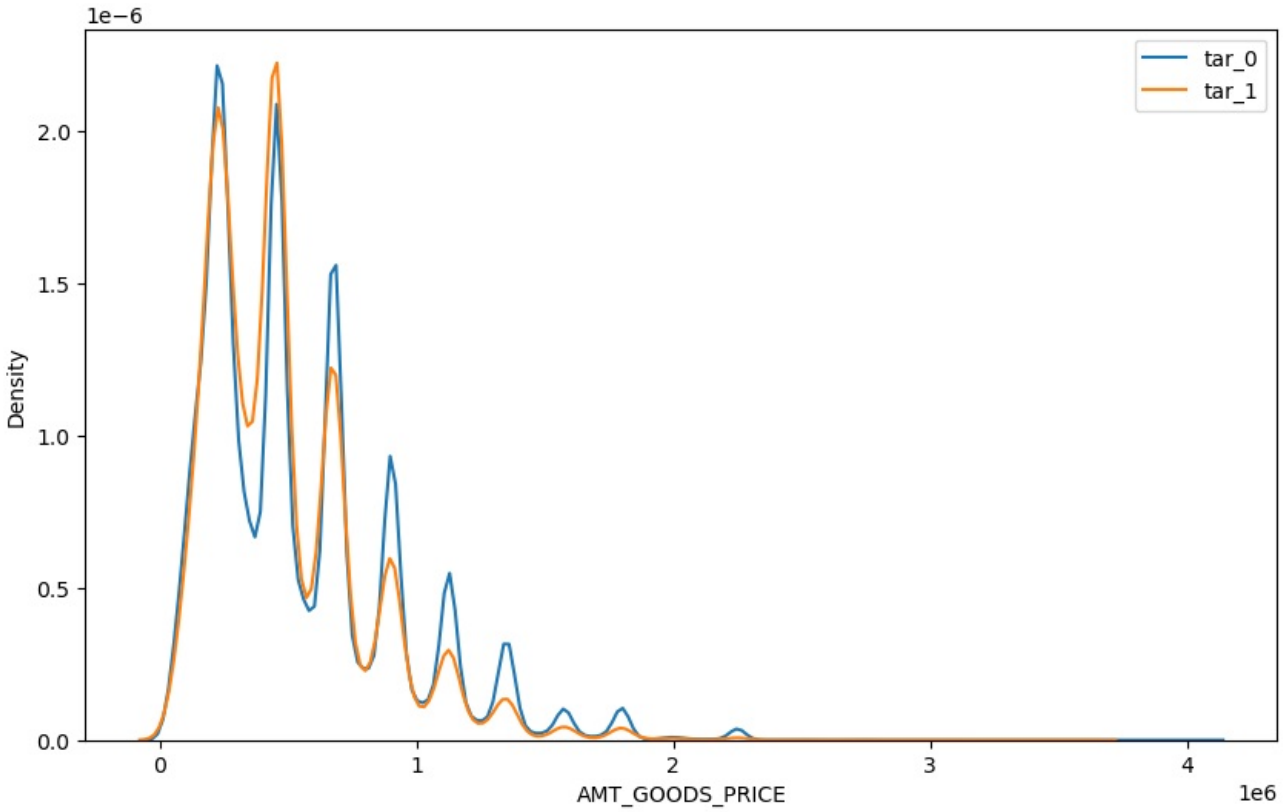


-Conclusion- 1.NAME_CONTRACT_TYPE-More applications have cash loans than Revolving loans both for target 0 and 1 2.CODE_GENDER- Number of female applicants are twice than that of male applicants both for target 0 and 1 3.FLAG_OWN_CAR- Most(70%) of the applicants do not own a car both for target 0 and 1 4.FLAG_OWN_REALITY- Most(70%) of applicants do not own a house both for target 0 and 1 5.FLAG_OWN_SUITE- Most(81%) of applicants are unaccompanied both for target 0 and 1 6.NAME_INCOME_TYPE- Most(51%) of applicants are earning their income from work both for target 0 and 1 7.NAME_EDUCATION_TYPE- Most(71%) of applicants have completed Secondary/secondary special education both for target 0 and 1 8.NAME_FAMILY_STATUS- Most(63%) of applicants are married both for target 0 and 1 9.NAME_HOUSING_TYPE- Most(88%) of housing type of applicants have house/appartment both for target 0 and 1 10.OCCUPATION_TYPE- most(31%) of applicants have other occupation type and non defaulters and laborere,sales staff,Drivers and core staff are not able to pay loan on time 11.WEEKDAY_APPR_PROCESS_START-Most of the applicants have applied loan on tuesday and the least on sunday 12.ORGANISATION_TYPE-Most of the organisation type of employees are Business Entity Type 3, Self Employed and other organisation type

Analysis on AMT_GOODS_PRICE on Target 0 and 1

In [113…

```
plt.figure(figsize = [10,6])
sns.distplot(tar_0['AMT_GOODS_PRICE'], label='tar_0', hist=False)
sns.distplot(tar_1['AMT_GOODS_PRICE'], label='tar_1', hist=False)
plt.legend()
plt.show()
```



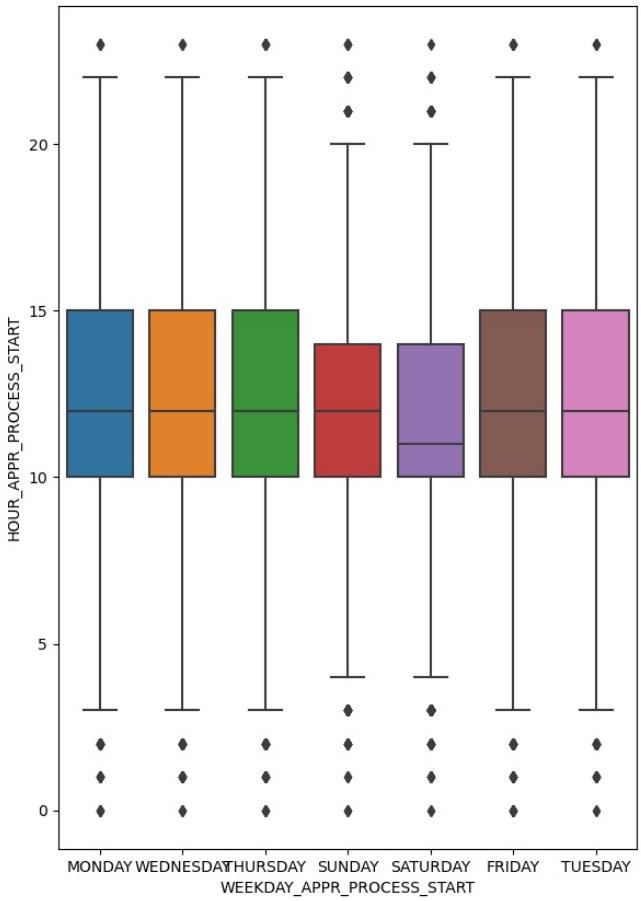
conclusion- The price of goods for which loans is given has the same variation for Target 0 and 1

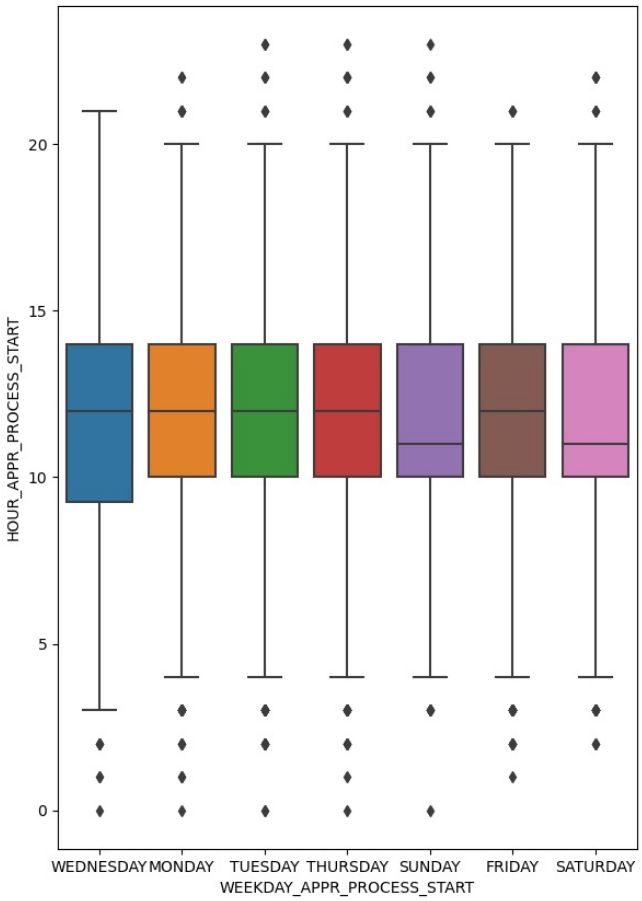
Bivariate and Multivariate Analysis

Bivariate Analysis between WEEKDAY_APPR_PROCESS_START vs HOUR_APPR_PROCESS_START

In [115…

```
plt.figure(figsize=(15,10))

plt.subplot(1,2,1)
sns.boxplot(x='WEEKDAY_APPR_PROCESS_START', y = 'HOUR_APPR_PROCESS_START' , data= tar_0)
plt.subplot(1,2,2)
sns.boxplot(x='WEEKDAY_APPR_PROCESS_START', y = 'HOUR_APPR_PROCESS_START' , data= tar_1)
plt.show()
```

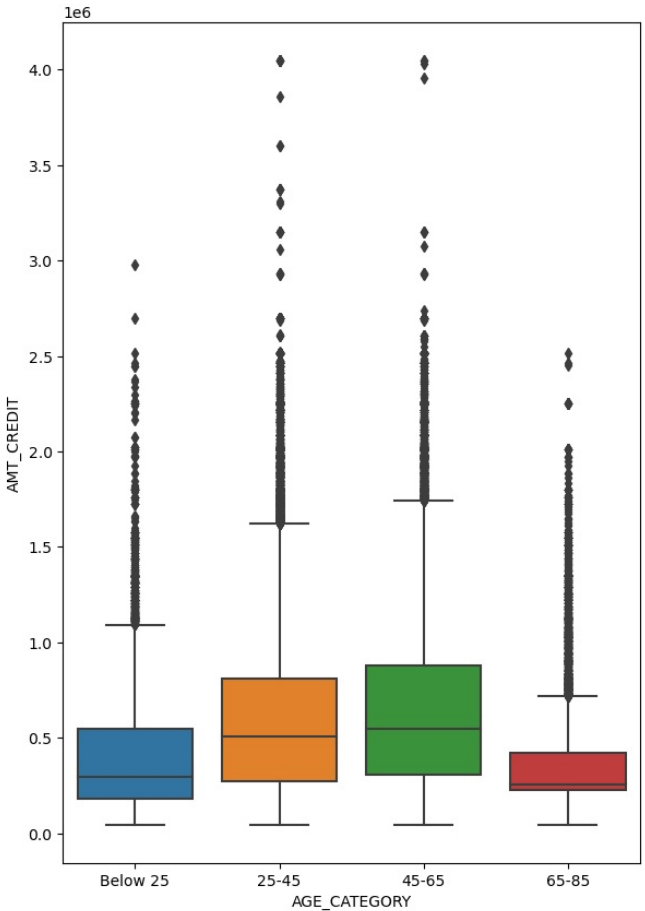


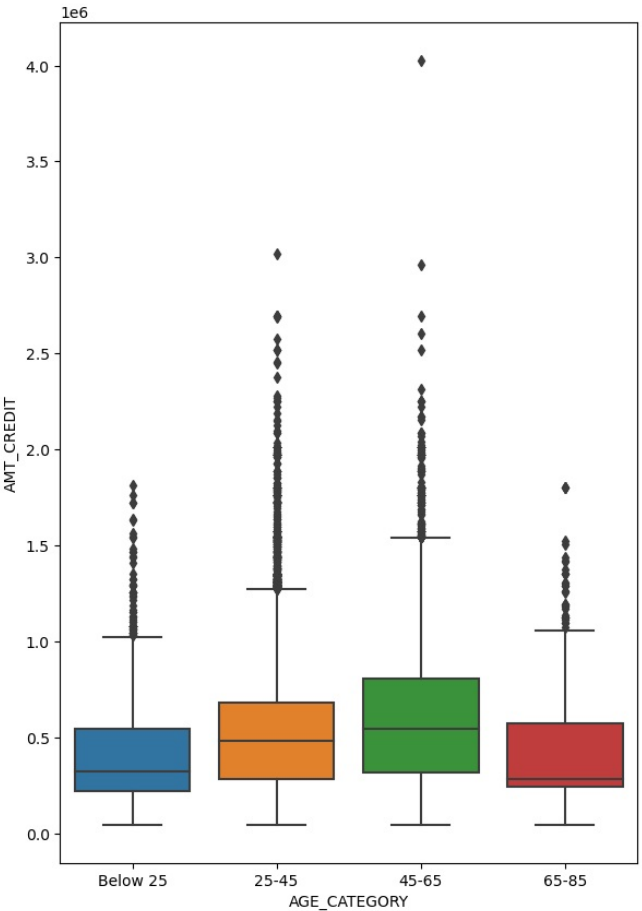


Bivariate Analysis between AGE_CATEGORY Vs AMT_CREDIT

In [118…

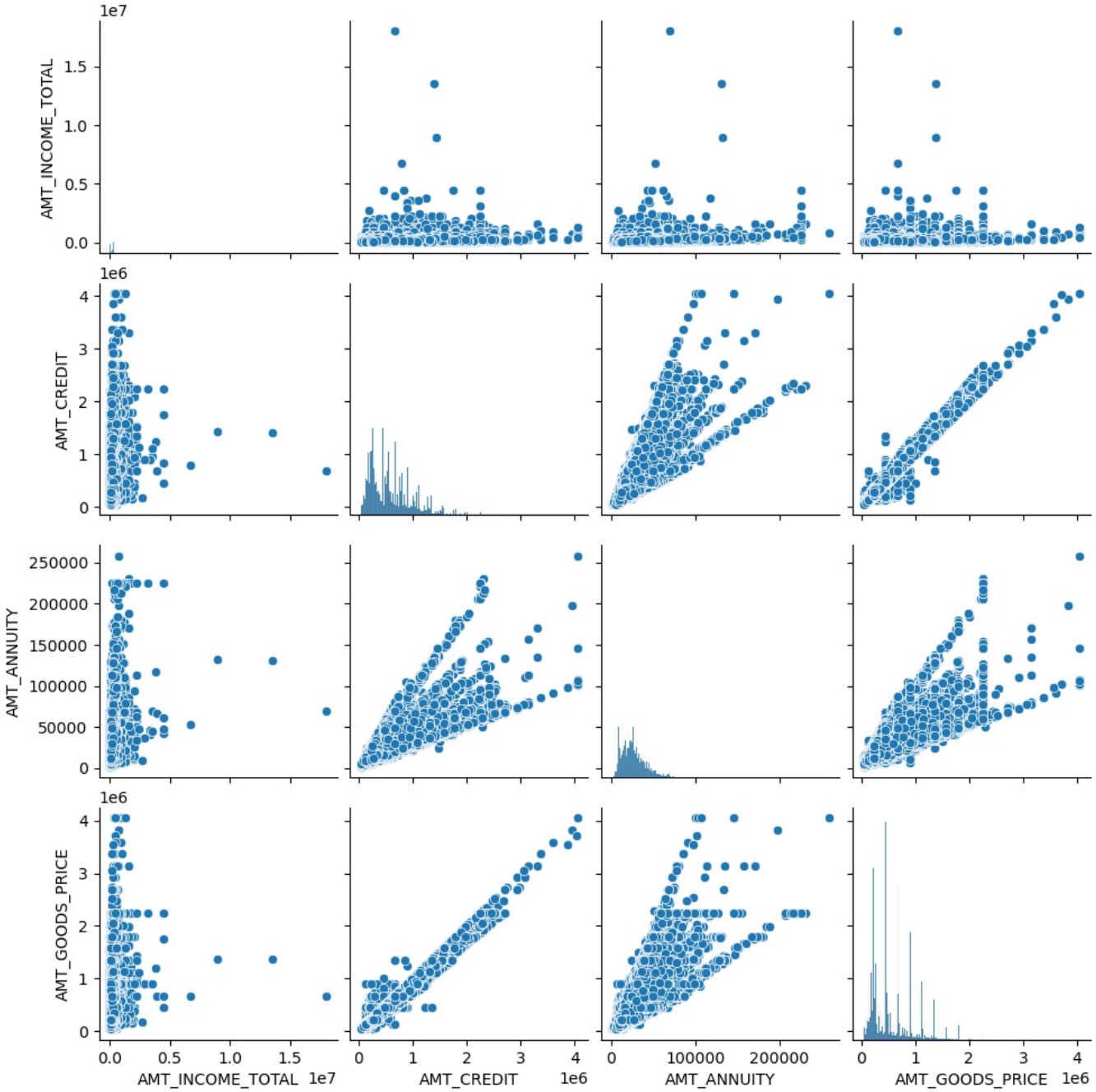
```
plt.figure(figsize=(15,10))

plt.subplot(1,2,1)
sns.boxplot(x='AGE_CATEGORY', y = 'AMT_CREDIT' , data= tar_0)
plt.subplot(1,2,2)
sns.boxplot(x='AGE_CATEGORY', y = 'AMT_CREDIT' , data= tar_1)
plt.show()
```





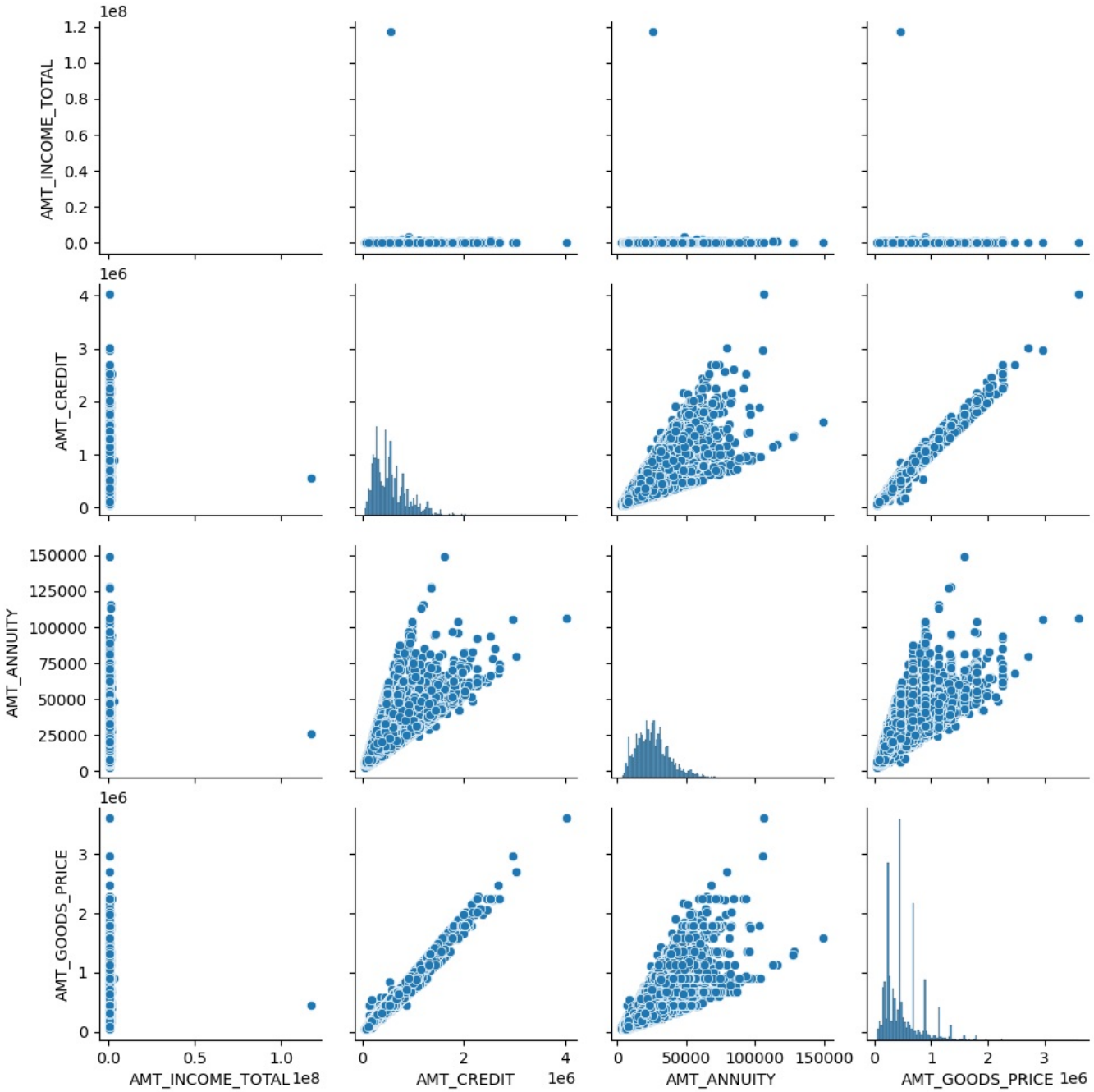
Pair plot of Amount Columns for Target 0

In [122…
```
sns.pairplot(tar_0[["AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY", "AMT_GOODS_PRICE"]])

plt.show()
```



Pair plot of Amount Columns for Target 1

In [123…
```
sns.pairplot(tar_1[["AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY", "AMT_GOODS_PRICE"]])

plt.show()
```



Co-relation between numerical columns

In [125…
```
corr_data = app_df[["AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY", "AMT_GOODS_PRICE", "YEARS_BIRTH", "YEARS_E
corr_data.head()
```

Out[125]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YE
0	202500.0	406597.5	24700.5	351000.0	26	1	10	
1	270000.0	1293502.5	35698.5	1129500.0	47	3	3	
2	67500.0	135000.0	6750.0	135000.0	53	0	11	
3	135000.0	312682.5	29686.5	297000.0	53	8	27	
4	121500.0	513000.0	21865.5	513000.0	55	8	12	

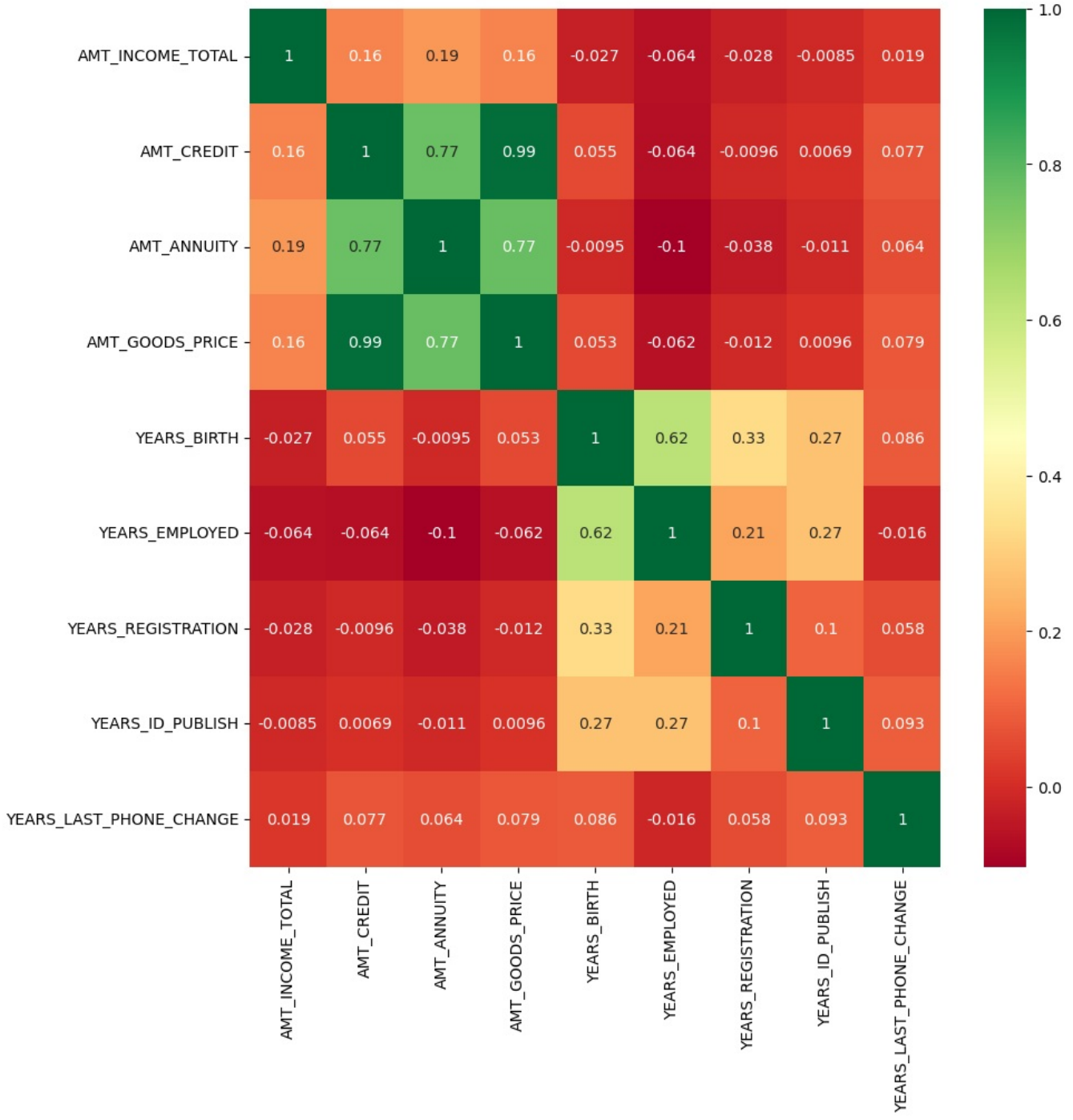
In [126…
```
corr_data.corr()
```

Out[126]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	YEARS_BIRTH	YEARS_EMPLOYEI
AMT_INCOME_TOTAL	1.000000	0.156870	0.191657	0.159632	-0.027239	-0.06383
AMT_CREDIT	0.156870	1.000000	0.770127	0.986734	0.055373	-0.06432
AMT_ANNUITY	0.191657	0.770127	1.000000	0.774837	-0.009519	-0.10284
AMT_GOODS_PRICE	0.159632	0.986734	0.774837	1.000000	0.053449	-0.06219
YEARS_BIRTH	-0.027239	0.055373	-0.009519	0.053449	1.000000	0.62374
YEARS_EMPLOYED	-0.063837	-0.064321	-0.102849	-0.062193	0.623745	1.00000
YEARS_REGISTRATION	-0.027882	-0.009590	-0.038487	-0.011518	0.331856	0.21465
YEARS_ID_PUBLISH	-0.008459	0.006942	-0.011376	0.009647	0.272054	0.27464
YEARS_LAST_PHONE_CHANGE	0.018571	0.077257	0.064494	0.079349	0.086317	-0.01642

In [128…

```
plt.figure(figsize=(10,10))
sns.heatmap(corr_data.corr(),annot=True, cmap="RdYlGn")
plt.show()
```



Split the numerical variables based on Target 0 and 1 to find the co-relation

In [129…

```
corr_data_0 = tar_0[["AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY", "AMT_GOODS_PRICE", "YEARS_BIRTH", "YEARS_
```

```
corr_data_0.head()
```

Out[129]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	YE
1	270000.0	1293502.5	35698.5	1129500.0	47	3	3	
2	67500.0	135000.0	6750.0	135000.0	53	0	11	
3	135000.0	312682.5	29686.5	297000.0	53	8	27	
4	121500.0	513000.0	21865.5	513000.0	55	8	12	
5	99000.0	490495.5	27517.5	454500.0	47	4	13	

In [130…

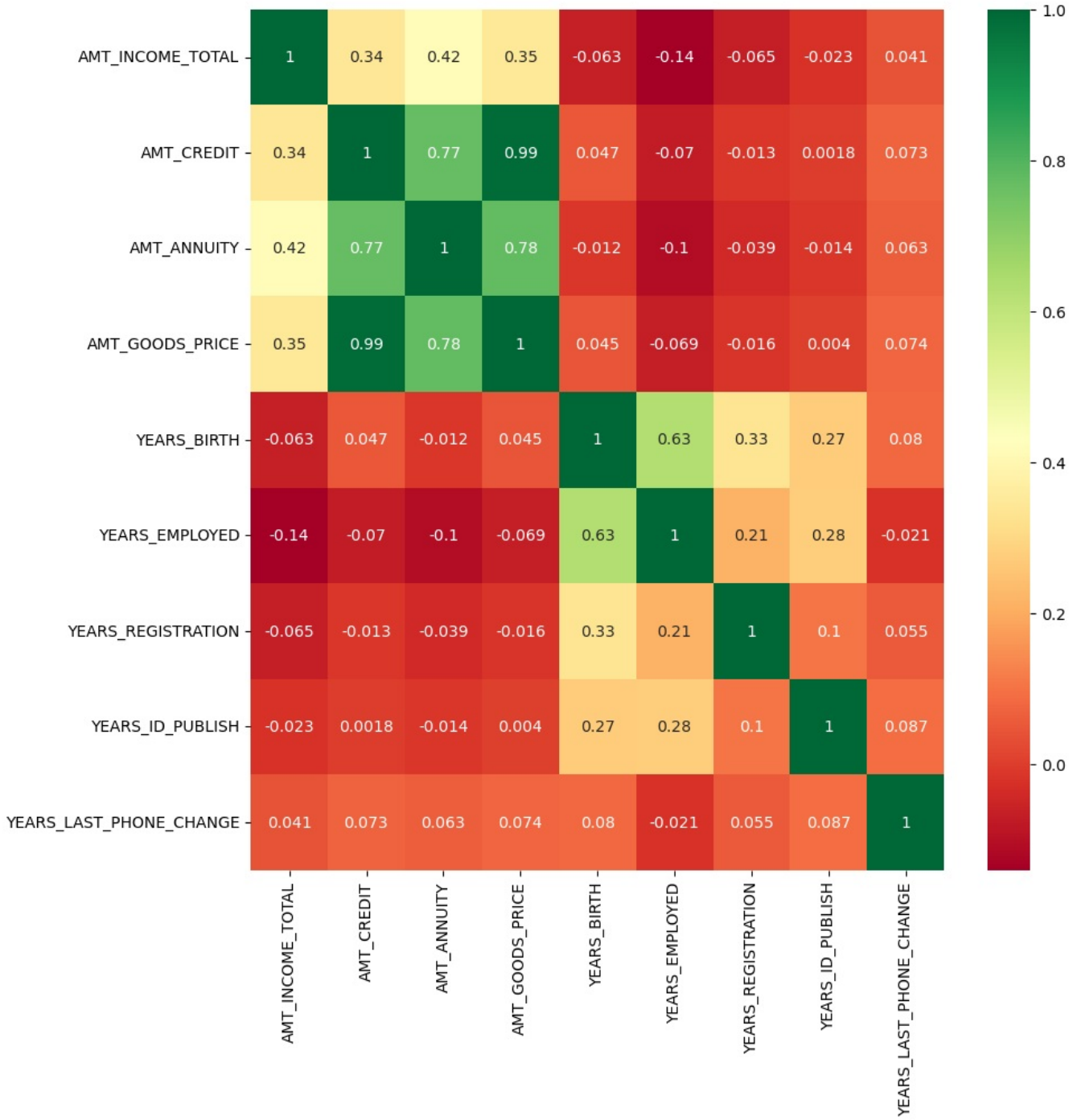
```
corr_data_1 = tar_1[["AMT_INCOME_TOTAL", "AMT_CREDIT", "AMT_ANNUITY", "AMT_GOODS_PRICE", "YEARS_BIRTH", "YEARS_
corr_data_1.head()
```

Out[130]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	YEARS_BIRTH	YEARS_EMPLOYED	YEARS_REGISTRATION	Y
0	202500.0	406597.5	24700.5	351000.0	26	1	10	
26	112500.0	979992.0	27076.5	702000.0	52	7	18	
40	202500.0	1193580.0	35028.0	855000.0	49	3	3	
42	135000.0	288873.0	16258.5	238500.0	37	10	0	
81	81000.0	252000.0	14593.5	252000.0	69	1025	15	

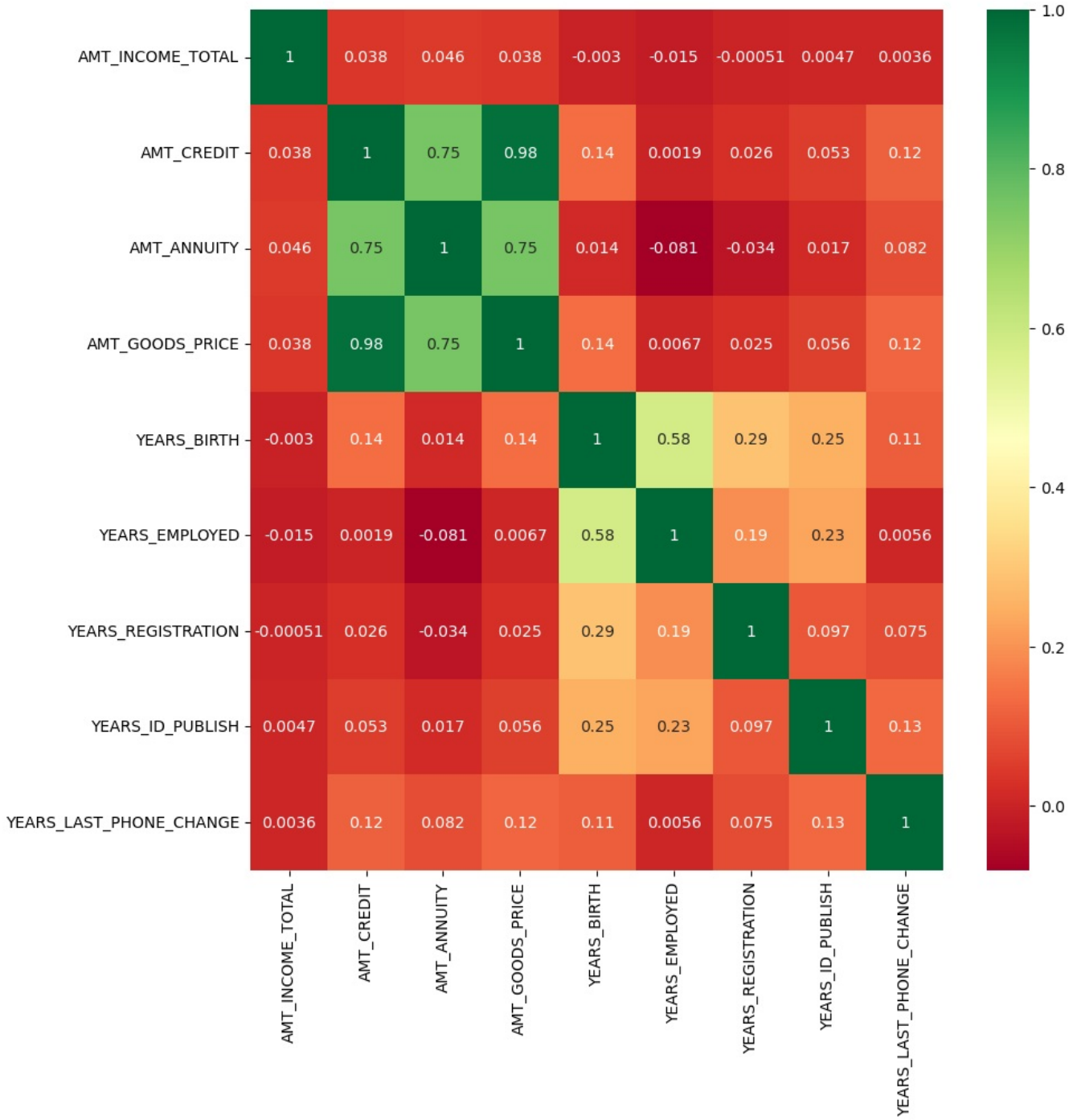
In [132…

```
plt.figure(figsize=(10,10))
sns.heatmap(corr_data_0.corr(),annot=True, cmap="RdYlGn")
plt.show()
```



In [133…

```
plt.figure(figsize=(10,10))
sns.heatmap(corr_data_1.corr(),annot=True, cmap="RdYlGn")
plt.show()
```



Read Previous Application csv

In [135…
```
papp_data = pd.read_csv("previous_application.csv")
papp_data.head()
```

Out[135]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME
0	100002	1	Cash loans	M	N	Y	0	
1	100003	0	Cash loans	F	N	N	0	
2	100004	0	Revolving loans	M	Y	Y	0	
3	100006	0	Cash loans	F	N	Y	0	
4	100007	0	Cash loans	M	N	Y	0	

Data inspection on previous application dataset

get info and shape on dataset

In [144…
```
papp_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

In [145…
```
papp_data.shape
```

Out[145]:
```
(307511, 122)
```

Data quality check

check for percentage null values in application dataset

In [146…
```
papp_data.isnull().mean()*100
```

Out[146]:
```
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
```

In [147…
```
percentage = 49
threshold_p = int(((100 - percentage)/100)*papp_data.shape[0] + 1)
papp_df = app_data.dropna(axis=1, thresh=threshold_p)
papp_df.head()
```

Out[147]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME
0	100002	1	Cash loans	M	N	Y	0	
1	100003	0	Cash loans	F	N	N	0	
2	100004	0	Revolving loans	M	Y	Y	0	
3	100006	0	Cash loans	F	N	Y	0	
4	100007	0	Cash loans	M	N	Y	0	

In [148…
```
papp_df.shape
```

Out[148]:
```
(307511, 78)
```

Impute missing values

check the dtype of missing values in dataset before imputing values

In [151…
```
for col in papp_df.columns:
    if papp_df[col].dtypes == np.int64 or papp_df[col].dtypes == np.float64:
        papp_df[col] = papp_df[col].apply(lambda x: abs)
```

validate if any null values present in datadset

In [152…
```
null_cols = list(papp_df.columns[papp_df.isna().any()])
len(null_cols)
```

Out[152]: 3

In [153…
```
papp_df.isnull().mean()*100
```

Out[153]:
```
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                    0.000000
AMT_GOODS_PRICE                0.000000
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
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                0.000000
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_2                   0.000000
EXT_SOURCE_3                   0.000000
YEARS_BEGINEXPLUATATION_AVG    0.000000
YEARS_BEGINEXPLUATATION_MODE   0.000000
YEARS_BEGINEXPLUATATION_MEDI   0.000000
TOTALAREA_MODE                 0.000000
EMERGENCYSTATE_MODE           47.398304
OBS_30_CNT_SOCIAL_CIRCLE       0.000000
DEF_30_CNT_SOCIAL_CIRCLE       0.000000
OBS_60_CNT_SOCIAL_CIRCLE       0.000000
DEF_60_CNT_SOCIAL_CIRCLE       0.000000
DAYS_LAST_PHONE_CHANGE         0.000000
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     0.000000
AMT_REQ_CREDIT_BUREAU_DAY      0.000000
AMT_REQ_CREDIT_BUREAU_WEEK     0.000000
AMT_REQ_CREDIT_BUREAU_MON      0.000000
AMT_REQ_CREDIT_BUREAU_QRT      0.000000
AMT_REQ_CREDIT_BUREAU_YEAR     0.000000
dtype: float64
```

Binning of continuos variables

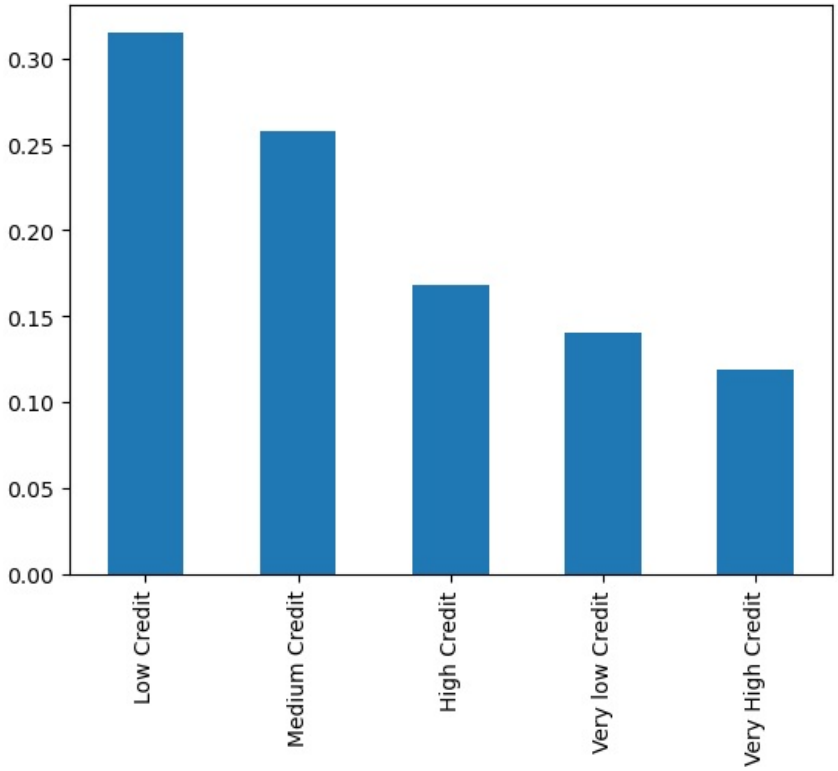
Binning AMT_CREDIT column

In [154…
```
papp_df.AMT_CREDIT.describe()
```

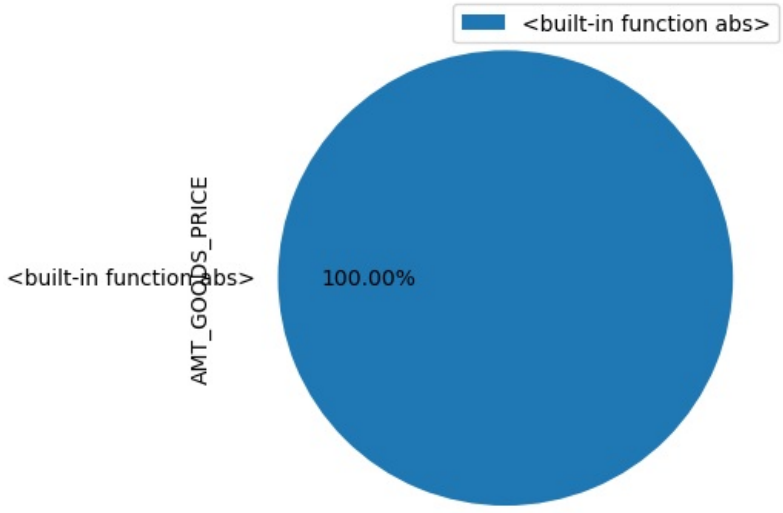
Out[154]:
```
count                      307511
unique                          1
top       <built-in function abs>
freq                       307511
Name: AMT_CREDIT, dtype: object
```

In [155…
```
papp_df["AMT_CREDIT_Category"] = pd.cut(app_df.AMT_CREDIT, [0,200000,400000, 600000,800000,1000000],
                                        labels = [ "Very low Credit", "Low Credit", "Medium Credit" , "High Credi
```

In [156…
```
papp_df["AMT_CREDIT_Category"].value_counts(normalize=True).plot.bar()
plt.show()
```



In [162…
```
papp_df["AMT_GOODS_PRICE"].value_counts(normalize=True).plot.pie(autopct ='%1.2f%%')
plt.legend()
plt.show()
```



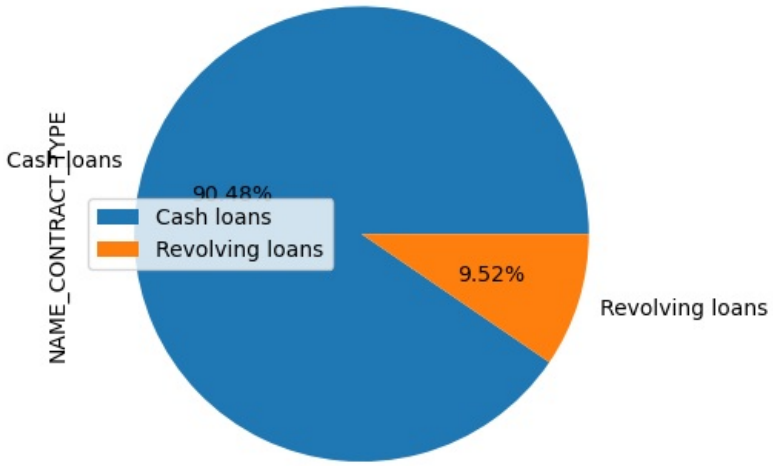
In [ ]:
```
### Data imbalance check
## Dividing application dataset with
```

In [171…
```
approved = papp_df[papp_df.NAME_CONTRACT_TYPE == "Approved"]
cancelled = papp_df[papp_df.NAME_CONTRACT_TYPE == "cancelled"]
refused = papp_df[papp_df.NAME_CONTRACT_TYPE == "refused"]
unusual = papp_df[papp_df.NAME_CONTRACT_TYPE == "unused offer"]
```

In [170…
```
papp_df.NAME_CONTRACT_TYPE.value_counts(normalize= True)*100
```

Out[170]:
```
Cash loans         90.478715
Revolving loans     9.521285
Name: NAME_CONTRACT_TYPE, dtype: float64
```

In [167…
```
papp_df.NAME_CONTRACT_TYPE.value_counts(normalize=True).plot.pie(autopct ='%1.2f%%')
plt.legend()
plt.show()
```

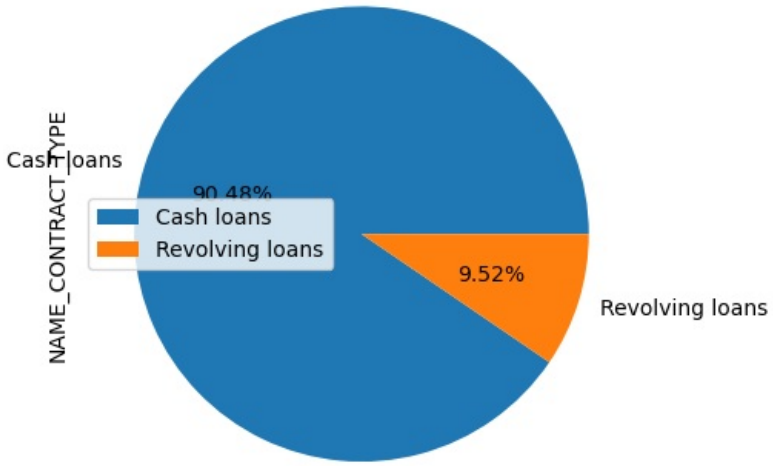


In [172…
```
approved = papp_df[papp_df.NAME_CONTRACT_TYPE == "Approved"]
cancelled = papp_df[papp_df.NAME_CONTRACT_TYPE == "cancelled"]
refused = papp_df[papp_df.NAME_CONTRACT_TYPE == "refused"]
unusual = papp_df[papp_df.NAME_CONTRACT_TYPE == "unused offer"]
```

In [173…
```
papp_df.NAME_CONTRACT_TYPE.value_counts(normalize= True)*100
```

Out[173]:
```
Cash loans         90.478715
Revolving loans     9.521285
Name: NAME_CONTRACT_TYPE, dtype: float64
```

In [174…
```
papp_df.NAME_CONTRACT_TYPE.value_counts(normalize=True).plot.pie(autopct ='%1.2f%%')
plt.legend()
plt.show()
```



In [ ]: