EDA PROJECT CASE STUDY Abdella Abdella

January 18, 2024

[3]:
```
##importing the useful libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[6]:
```
import warnings
warnings.filterwarnings("ignore")
```

[8]:
```
df = pd.read_csv("application_data.csv")
```

[10]:
```
df.head()
```

[10]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

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

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

  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
```

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

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

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                        0.0                        1.0
1                        0.0                        0.0
2                        0.0                        0.0
3                        NaN                        NaN
4                        0.0                        0.0

[5 rows x 122 columns]
```

[12]:
```
pd.set_option('display.max_columns',500)
pd.set_option("display.max_rows",500)
df.head()
```

[12]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

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

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

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

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

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

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

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

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

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

```
4                           0                           0

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

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

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

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

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

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

   NONLIVINGAREA_AVG  APARTMENTS_MODE  BASEMENTAREA_MODE  \
0             0.0000           0.0252             0.0383
1             0.0098           0.0924             0.0538
```

```
2                NaN              NaN                NaN
3                NaN              NaN                NaN
4                NaN              NaN                NaN

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

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

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

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

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

   COMMONAREA_MEDI  ELEVATORS_MEDI  ENTRANCES_MEDI  FLOORSMAX_MEDI  \
0           0.0144            0.00          0.0690          0.0833
1           0.0608            0.08          0.0345          0.2917
2              NaN             NaN             NaN             NaN
3              NaN             NaN             NaN             NaN
4              NaN             NaN             NaN             NaN

   FLOORSMIN_MEDI  LANDAREA_MEDI  LIVINGAPARTMENTS_MEDI  LIVINGAREA_MEDI  \
```

```
0          0.1250         0.0375                 0.0205           0.0193
1          0.3333         0.0132                 0.0787           0.0558
2             NaN            NaN                    NaN              NaN
3             NaN            NaN                    NaN              NaN
4             NaN            NaN                    NaN              NaN

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

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

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

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

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

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

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

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

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

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

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

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                        0.0                        1.0
1                        0.0                        0.0
2                        0.0                        0.0
3                        NaN                        NaN
4                        0.0                        0.0
```

[14]:
```
df.shape
```

[14]: (307511, 122)

[20]:
```
df.columns[df.isnull().any()]
```

[20]:
```
Index(['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'OWN_CAR_AGE',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_1', 'EXT_SOURCE_2',
       'EXT_SOURCE_3', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG',
       'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG',
       'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG',
       'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG',
       'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE',
       'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE',
       'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE',
       'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE',
       'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE',
       'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI',
       'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI',
       'ENTRANCES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI',
       'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI',
       'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
       'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE',
       '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'],
      dtype='object')
```

[22]:
```
len(df.columns[df.isnull().any()])
```

[22]:
```
67
```

[24]:
```
df.isnull().sum()
```

[24]:
```
SK_ID_CURR                       0
TARGET                           0
NAME_CONTRACT_TYPE               0
CODE_GENDER                      0
FLAG_OWN_CAR                     0
FLAG_OWN_REALTY                  0
CNT_CHILDREN                     0
AMT_INCOME_TOTAL                 0
AMT_CREDIT                       0
AMT_ANNUITY                     12
AMT_GOODS_PRICE                278
NAME_TYPE_SUITE               1292
NAME_INCOME_TYPE                 0
NAME_EDUCATION_TYPE              0
```

NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
REGION_POPULATION_RELATIVE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
DAYS_REGISTRATION	0
DAYS_ID_PUBLISH	0
OWN_CAR_AGE	202929
FLAG_MOBIL	0
FLAG_EMP_PHONE	0
FLAG_WORK_PHONE	0
FLAG_CONT_MOBILE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	96391
CNT_FAM_MEMBERS	2
REGION_RATING_CLIENT	0
REGION_RATING_CLIENT_W_CITY	0
WEEKDAY_APPR_PROCESS_START	0
HOUR_APPR_PROCESS_START	0
REG_REGION_NOT_LIVE_REGION	0
REG_REGION_NOT_WORK_REGION	0
LIVE_REGION_NOT_WORK_REGION	0
REG_CITY_NOT_LIVE_CITY	0
REG_CITY_NOT_WORK_CITY	0
LIVE_CITY_NOT_WORK_CITY	0
ORGANIZATION_TYPE	0
EXT_SOURCE_1	173378
EXT_SOURCE_2	660
EXT_SOURCE_3	60965
APARTMENTS_AVG	156061
BASEMENTAREA_AVG	179943
YEARS_BEGINEXPLUATATION_AVG	150007
YEARS_BUILD_AVG	204488
COMMONAREA_AVG	214865
ELEVATORS_AVG	163891
ENTRANCES_AVG	154828
FLOORSMAX_AVG	153020
FLOORSMIN_AVG	208642
LANDAREA_AVG	182590
LIVINGAPARTMENTS_AVG	210199
LIVINGAREA_AVG	154350
NONLIVINGAPARTMENTS_AVG	213514
NONLIVINGAREA_AVG	169682
APARTMENTS_MODE	156061
BASEMENTAREA_MODE	179943
YEARS_BEGINEXPLUATATION_MODE	150007

YEARS_BUILD_MODE	204488
COMMONAREA_MODE	214865
ELEVATORS_MODE	163891
ENTRANCES_MODE	154828
FLOORSMAX_MODE	153020
FLOORSMIN_MODE	208642
LANDAREA_MODE	182590
LIVINGAPARTMENTS_MODE	210199
LIVINGAREA_MODE	154350
NONLIVINGAPARTMENTS_MODE	213514
NONLIVINGAREA_MODE	169682
APARTMENTS_MEDI	156061
BASEMENTAREA_MEDI	179943
YEARS_BEGINEXPLUATATION_MEDI	150007
YEARS_BUILD_MEDI	204488
COMMONAREA_MEDI	214865
ELEVATORS_MEDI	163891
ENTRANCES_MEDI	154828
FLOORSMAX_MEDI	153020
FLOORSMIN_MEDI	208642
LANDAREA_MEDI	182590
LIVINGAPARTMENTS_MEDI	210199
LIVINGAREA_MEDI	154350
NONLIVINGAPARTMENTS_MEDI	213514
NONLIVINGAREA_MEDI	169682
FONDKAPREMONT_MODE	210295
HOUSETYPE_MODE	154297
TOTALAREA_MODE	148431
WALLSMATERIAL_MODE	156341
EMERGENCYSTATE_MODE	145755
OBS_30_CNT_SOCIAL_CIRCLE	1021
DEF_30_CNT_SOCIAL_CIRCLE	1021
OBS_60_CNT_SOCIAL_CIRCLE	1021
DEF_60_CNT_SOCIAL_CIRCLE	1021
DAYS_LAST_PHONE_CHANGE	1
FLAG_DOCUMENT_2	0
FLAG_DOCUMENT_3	0
FLAG_DOCUMENT_4	0
FLAG_DOCUMENT_5	0
FLAG_DOCUMENT_6	0
FLAG_DOCUMENT_7	0
FLAG_DOCUMENT_8	0
FLAG_DOCUMENT_9	0
FLAG_DOCUMENT_10	0
FLAG_DOCUMENT_11	0
FLAG_DOCUMENT_12	0
FLAG_DOCUMENT_13	0

```
FLAG_DOCUMENT_14                  0
FLAG_DOCUMENT_15                  0
FLAG_DOCUMENT_16                  0
FLAG_DOCUMENT_17                  0
FLAG_DOCUMENT_18                  0
FLAG_DOCUMENT_19                  0
FLAG_DOCUMENT_20                  0
FLAG_DOCUMENT_21                  0
AMT_REQ_CREDIT_BUREAU_HOUR    41519
AMT_REQ_CREDIT_BUREAU_DAY     41519
AMT_REQ_CREDIT_BUREAU_WEEK    41519
AMT_REQ_CREDIT_BUREAU_MON     41519
AMT_REQ_CREDIT_BUREAU_QRT     41519
AMT_REQ_CREDIT_BUREAU_YEAR    41519
dtype: int64
```

[26]:
```
missing_cols = 100*df.isnull().mean()
missing_cols
```

[26]:
```
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
```

[28]:
```
missing_cols_40 = missing_cols[missing_cols.values > 40].index.to_list()
missing_cols_40
```

[28]:
```
['OWN_CAR_AGE',
 'EXT_SOURCE_1',
 'APARTMENTS_AVG',
 'BASEMENTAREA_AVG',
 'YEARS_BEGINEXPLUATATION_AVG',
 'YEARS_BUILD_AVG',
 'COMMONAREA_AVG',
 'ELEVATORS_AVG',
 'ENTRANCES_AVG',
 'FLOORSMAX_AVG',
 'FLOORSMIN_AVG',
 'LANDAREA_AVG',
 'LIVINGAPARTMENTS_AVG',
 'LIVINGAREA_AVG',
 'NONLIVINGAPARTMENTS_AVG',
 'NONLIVINGAREA_AVG',
 'APARTMENTS_MODE',
 'BASEMENTAREA_MODE',
 'YEARS_BEGINEXPLUATATION_MODE',
 'YEARS_BUILD_MODE',
 'COMMONAREA_MODE',
 'ELEVATORS_MODE',
 'ENTRANCES_MODE',
 'FLOORSMAX_MODE',
 'FLOORSMIN_MODE',
 'LANDAREA_MODE',
 'LIVINGAPARTMENTS_MODE',
 'LIVINGAREA_MODE',
 'NONLIVINGAPARTMENTS_MODE',
 'NONLIVINGAREA_MODE',
 'APARTMENTS_MEDI',
 'BASEMENTAREA_MEDI',
 'YEARS_BEGINEXPLUATATION_MEDI',
 'YEARS_BUILD_MEDI',
 'COMMONAREA_MEDI',
 'ELEVATORS_MEDI',
 'ENTRANCES_MEDI',
 'FLOORSMAX_MEDI',
 'FLOORSMIN_MEDI',
 'LANDAREA_MEDI',
 'LIVINGAPARTMENTS_MEDI',
 'LIVINGAREA_MEDI',
```

```
 'NONLIVINGAPARTMENTS_MEDI',
 'NONLIVINGAREA_MEDI',
 'FONDKAPREMONT_MODE',
 'HOUSETYPE_MODE',
 'TOTALAREA_MODE',
 'WALLSMATERIAL_MODE',
 'EMERGENCYSTATE_MODE']
```

[30]:
```
df = df.drop(labels = missing_cols_40, axis = 1)
```

[32]:
```
100*df.isnull().mean().sort_values(ascending = False)
```

[32]:
```
OCCUPATION_TYPE               31.345545
EXT_SOURCE_3                  19.825307
AMT_REQ_CREDIT_BUREAU_YEAR    13.501631
AMT_REQ_CREDIT_BUREAU_QRT     13.501631
AMT_REQ_CREDIT_BUREAU_MON     13.501631
AMT_REQ_CREDIT_BUREAU_WEEK    13.501631
AMT_REQ_CREDIT_BUREAU_DAY     13.501631
AMT_REQ_CREDIT_BUREAU_HOUR    13.501631
NAME_TYPE_SUITE                0.420148
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
EXT_SOURCE_2                   0.214626
AMT_GOODS_PRICE                0.090403
AMT_ANNUITY                    0.003902
CNT_FAM_MEMBERS                0.000650
DAYS_LAST_PHONE_CHANGE         0.000325
FLAG_DOCUMENT_17               0.000000
FLAG_DOCUMENT_18               0.000000
FLAG_DOCUMENT_21               0.000000
FLAG_DOCUMENT_20               0.000000
FLAG_DOCUMENT_19               0.000000
FLAG_DOCUMENT_2                0.000000
FLAG_DOCUMENT_3                0.000000
FLAG_DOCUMENT_4                0.000000
FLAG_DOCUMENT_5                0.000000
FLAG_DOCUMENT_16               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
ORGANIZATION_TYPE              0.000000
```

```
FLAG_DOCUMENT_13               0.000000
FLAG_DOCUMENT_14               0.000000
FLAG_DOCUMENT_15               0.000000
FLAG_DOCUMENT_12               0.000000
SK_ID_CURR                     0.000000
LIVE_CITY_NOT_WORK_CITY        0.000000
DAYS_REGISTRATION              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
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_ID_PUBLISH                0.000000
REG_CITY_NOT_WORK_CITY         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
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
TARGET                         0.000000
REG_CITY_NOT_LIVE_CITY         0.000000
dtype: float64
```

[34]:
```
## Listing and dropping the columns which are not much relavant for the analysis

cols_not_relevant = [
 'FLAG_EMAIL',
 'FLAG_DOCUMENT_6',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_7',
```

```
 'FLAG_CONT_MOBILE',
 'FLAG_DOCUMENT_13',
 'FLAG_DOCUMENT_15',
 'FLAG_DOCUMENT_19',
 'FLAG_WORK_PHONE',
 'FLAG_DOCUMENT_21',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_3',
 'FLAG_DOCUMENT_16',
 'FLAG_DOCUMENT_2',
 'FLAG_DOCUMENT_8',
 'REGION_RATING_CLIENT_W_CITY',
 'FLAG_DOCUMENT_20',
'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_18',
 'REGION_RATING_CLIENT',
 'FLAG_EMP_PHONE',
 'FLAG_MOBIL',
 'FLAG_DOCUMENT_12',
 'CNT_FAM_MEMBERS',
 'FLAG_PHONE',
 'FLAG_DOCUMENT_14',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_11',
 'FLAG_DOCUMENT_9',
 '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',
"NAME_TYPE_SUITE",
"DAYS_ID_PUBLISH",
"WEEKDAY_APPR_PROCESS_START",
"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",
"DAYS_LAST_PHONE_CHANGE"]



df.drop(labels=cols_not_relevant,axis=1,inplace=True)
```

[36]:
```
100*df.isnull().mean().sort_values(ascending = False)
```

[36]:
```
OCCUPATION_TYPE               31.345545
EXT_SOURCE_3                  19.825307
AMT_REQ_CREDIT_BUREAU_HOUR    13.501631
DEF_60_CNT_SOCIAL_CIRCLE       0.332021
OBS_60_CNT_SOCIAL_CIRCLE       0.332021
DEF_30_CNT_SOCIAL_CIRCLE       0.332021
OBS_30_CNT_SOCIAL_CIRCLE       0.332021
EXT_SOURCE_2                   0.214626
AMT_GOODS_PRICE                0.090403
AMT_ANNUITY                    0.003902
REGION_POPULATION_RELATIVE     0.000000
ORGANIZATION_TYPE              0.000000
DAYS_REGISTRATION              0.000000
DAYS_EMPLOYED                  0.000000
DAYS_BIRTH                     0.000000
SK_ID_CURR                     0.000000
TARGET                         0.000000
NAME_FAMILY_STATUS             0.000000
NAME_EDUCATION_TYPE            0.000000
NAME_INCOME_TYPE               0.000000
AMT_CREDIT                     0.000000
AMT_INCOME_TOTAL               0.000000
CNT_CHILDREN                   0.000000
FLAG_OWN_REALTY                0.000000
FLAG_OWN_CAR                   0.000000
CODE_GENDER                    0.000000
NAME_CONTRACT_TYPE             0.000000
NAME_HOUSING_TYPE              0.000000
dtype: float64
```

[38]:
```
mode = df.OCCUPATION_TYPE.mode()[0]
mode
```

[38]:
```
'Laborers'
```

[40]:
```
df["OCCUPATION_TYPE"].fillna(mode, inplace = True)
```

[42]:
```
df["EXT_SOURCE_3"] = df["EXT_SOURCE_3"].fillna(df["EXT_SOURCE_3"].median())
df["EXT_SOURCE_2"] = df["EXT_SOURCE_2"].fillna(df["EXT_SOURCE_2"].median())
```

[44]:
```
df["AMT_ANNUITY"] = df["AMT_ANNUITY"].fillna(df["AMT_ANNUITY"].median())
df["AMT_GOODS_PRICE"] = df["AMT_GOODS_PRICE"].fillna(df["AMT_GOODS_PRICE"].
 ↪median())
```

[46]:
```
df["OBS_30_CNT_SOCIAL_CIRCLE"] = df["OBS_30_CNT_SOCIAL_CIRCLE"].
 ↪fillna(df["OBS_30_CNT_SOCIAL_CIRCLE"].median())
```

```
df["DEF_30_CNT_SOCIAL_CIRCLE"] = df["DEF_30_CNT_SOCIAL_CIRCLE"].
 ↪fillna(df["DEF_30_CNT_SOCIAL_CIRCLE"].median())
df["OBS_60_CNT_SOCIAL_CIRCLE"] = df["OBS_60_CNT_SOCIAL_CIRCLE"].
 ↪fillna(df["OBS_60_CNT_SOCIAL_CIRCLE"].median())
df["DEF_60_CNT_SOCIAL_CIRCLE"] = df["DEF_60_CNT_SOCIAL_CIRCLE"].
 ↪fillna(df["DEF_60_CNT_SOCIAL_CIRCLE"].median())
```

[48]:
```
df["AMT_REQ_CREDIT_BUREAU_HOUR"] = df["AMT_REQ_CREDIT_BUREAU_HOUR"].
 ↪fillna(df["AMT_REQ_CREDIT_BUREAU_HOUR"].median())
```

[50]:
```
df.isnull().sum()
```

[50]:
```
SK_ID_CURR                    0
TARGET                        0
NAME_CONTRACT_TYPE            0
CODE_GENDER                   0
FLAG_OWN_CAR                  0
FLAG_OWN_REALTY               0
CNT_CHILDREN                  0
AMT_INCOME_TOTAL              0
AMT_CREDIT                    0
AMT_ANNUITY                   0
AMT_GOODS_PRICE               0
NAME_INCOME_TYPE              0
NAME_EDUCATION_TYPE           0
NAME_FAMILY_STATUS            0
NAME_HOUSING_TYPE             0
REGION_POPULATION_RELATIVE    0
DAYS_BIRTH                    0
DAYS_EMPLOYED                 0
DAYS_REGISTRATION             0
OCCUPATION_TYPE               0
ORGANIZATION_TYPE             0
EXT_SOURCE_2                  0
EXT_SOURCE_3                  0
OBS_30_CNT_SOCIAL_CIRCLE      0
DEF_30_CNT_SOCIAL_CIRCLE      0
OBS_60_CNT_SOCIAL_CIRCLE      0
DEF_60_CNT_SOCIAL_CIRCLE      0
AMT_REQ_CREDIT_BUREAU_HOUR    0
dtype: int64
```

[52]:
```
df.nunique()
```

[52]:
```
SK_ID_CURR                  307511
TARGET                           2
NAME_CONTRACT_TYPE               2
```

```
CODE_GENDER                      3
FLAG_OWN_CAR                     2
FLAG_OWN_REALTY                  2
CNT_CHILDREN                    15
AMT_INCOME_TOTAL              2548
AMT_CREDIT                    5603
AMT_ANNUITY                  13672
AMT_GOODS_PRICE               1002
NAME_INCOME_TYPE                 8
NAME_EDUCATION_TYPE              5
NAME_FAMILY_STATUS               6
NAME_HOUSING_TYPE                6
REGION_POPULATION_RELATIVE      81
DAYS_BIRTH                   17460
DAYS_EMPLOYED                12574
DAYS_REGISTRATION            15688
OCCUPATION_TYPE                 18
ORGANIZATION_TYPE               58
EXT_SOURCE_2                119828
EXT_SOURCE_3                   814
OBS_30_CNT_SOCIAL_CIRCLE        33
DEF_30_CNT_SOCIAL_CIRCLE        10
OBS_60_CNT_SOCIAL_CIRCLE        33
DEF_60_CNT_SOCIAL_CIRCLE         9
AMT_REQ_CREDIT_BUREAU_HOUR       5
dtype: int64
```

[54]:
```
df.CODE_GENDER.value_counts()
```

[54]:
```
CODE_GENDER
F      202448
M      105059
XNA         4
Name: count, dtype: int64
```

[56]:
```
mode = df.CODE_GENDER.mode()[0]
df["CODE_GENDER"] = df["CODE_GENDER"].replace("XNA", mode)
```

[58]:
```
df.CODE_GENDER.value_counts()
```

[58]:
```
CODE_GENDER
F    202452
M    105059
Name: count, dtype: int64
```

[60]:
```
df.ORGANIZATION_TYPE.value_counts()
```

[60]:
```
ORGANIZATION_TYPE
Business Entity Type 3    67992
XNA                       55374
Self-employed             38412
Other                     16683
Medicine                  11193
Business Entity Type 2    10553
Government                10404
School                     8893
Trade: type 7              7831
Kindergarten               6880
Construction               6721
Business Entity Type 1     5984
Transport: type 4          5398
Trade: type 3              3492
Industry: type 9           3368
Industry: type 3           3278
Security                   3247
Housing                    2958
Industry: type 11          2704
Military                   2634
Bank                       2507
Agriculture                2454
Police                     2341
Transport: type 2          2204
Postal                     2157
Security Ministries        1974
Trade: type 2              1900
Restaurant                 1811
Services                   1575
University                 1327
Industry: type 7           1307
Transport: type 3          1187
Industry: type 1           1039
Hotel                       966
Electricity                 950
Industry: type 4            877
Trade: type 6               631
Industry: type 5            599
Insurance                   597
Telecom                     577
Emergency                   560
Industry: type 2            458
Advertising                 429
Realtor                     396
Culture                     379
Industry: type 12           369
```

```
Trade: type 1                348
Mobile                       317
Legal Services               305
Cleaning                     260
Transport: type 1            201
Industry: type 6             112
Industry: type 10            109
Religion                      85
Industry: type 13             67
Trade: type 4                 64
Trade: type 5                 49
Industry: type 8              24
Name: count, dtype: int64
```

[62]:
```
df=df.drop(df.loc[df['ORGANIZATION_TYPE']=='XNA'].index)
df[df['ORGANIZATION_TYPE']=='XNA'].shape
df.reset_index(inplace = True, drop = True)
```

[64]:
```
## Creating a new column EXT_SOURCE_SCORE by taking the average of EXT_SOURCE_2 and EXT_SOURCE_3 for a more simplified analysis.

df['EXT_SOURCE_SCORE'] = round(((df['EXT_SOURCE_2'] + df['EXT_SOURCE_2'])/2),2)

## Dropping columns EXT_SOURCE_2 and EXT_SOURCE_3

df.drop(['EXT_SOURCE_2','EXT_SOURCE_3'],axis=1,inplace=True)
```

[66]:
```
df.head()
```

[66]:

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

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

	AMT_GOODS_PRICE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE
0	351000.0	Working	Secondary / secondary special
1	1129500.0	State servant	Higher education
2	135000.0	Working	Secondary / secondary special

```
3         297000.0          Working  Secondary / secondary special
4         513000.0          Working  Secondary / secondary special

     NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  \
0  Single / not married  House / apartment                    0.018801
1               Married  House / apartment                    0.003541
2  Single / not married  House / apartment                    0.010032
3       Civil marriage  House / apartment                    0.008019
4  Single / not married  House / apartment                    0.028663

   DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION OCCUPATION_TYPE  \
0       -9461           -637            -3648.0        Laborers
1      -16765          -1188            -1186.0      Core staff
2      -19046           -225            -4260.0        Laborers
3      -19005          -3039            -9833.0        Laborers
4      -19932          -3038            -4311.0      Core staff

        ORGANIZATION_TYPE  OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0  Business Entity Type 3                       2.0                       2.0
1                  School                       1.0                       0.0
2              Government                       0.0                       0.0
3  Business Entity Type 3                       2.0                       0.0
4                Religion                       0.0                       0.0

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       0.0                       0.0
3                       2.0                       0.0
4                       0.0                       0.0

   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE
0                         0.0              0.26
1                         0.0              0.62
2                         0.0              0.56
3                         0.0              0.65
4                         0.0              0.32
```

[74]:
```
## Changing those values to absolute values

df[["DAYS_BIRTH","DAYS_EMPLOYED", "DAYS_REGISTRATION"]] =␣
 ↪abs(df[["DAYS_BIRTH","DAYS_EMPLOYED", "DAYS_REGISTRATION"]])
```

[76]:
```
df["AGE"] = (df.DAYS_BIRTH)/365
df["YEARS_EMPLOYED"] = (df.DAYS_EMPLOYED)/365
```

[78]:
```
df.drop(columns=['DAYS_BIRTH','DAYS_EMPLOYED'],inplace=True)
```

[80]:
```
df.head()
```

[80]:

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

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

	AMT_GOODS_PRICE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE
0	351000.0	Working	Secondary / secondary special
1	1129500.0	State servant	Higher education
2	135000.0	Working	Secondary / secondary special
3	297000.0	Working	Secondary / secondary special
4	513000.0	Working	Secondary / secondary special

	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE
0	Single / not married	House / apartment	0.018801
1	Married	House / apartment	0.003541
2	Single / not married	House / apartment	0.010032
3	Civil marriage	House / apartment	0.008019
4	Single / not married	House / apartment	0.028663

	DAYS_REGISTRATION	OCCUPATION_TYPE	ORGANIZATION_TYPE
0	3648.0	Laborers	Business Entity Type 3
1	1186.0	Core staff	School
2	4260.0	Laborers	Government
3	9833.0	Laborers	Business Entity Type 3
4	4311.0	Core staff	Religion

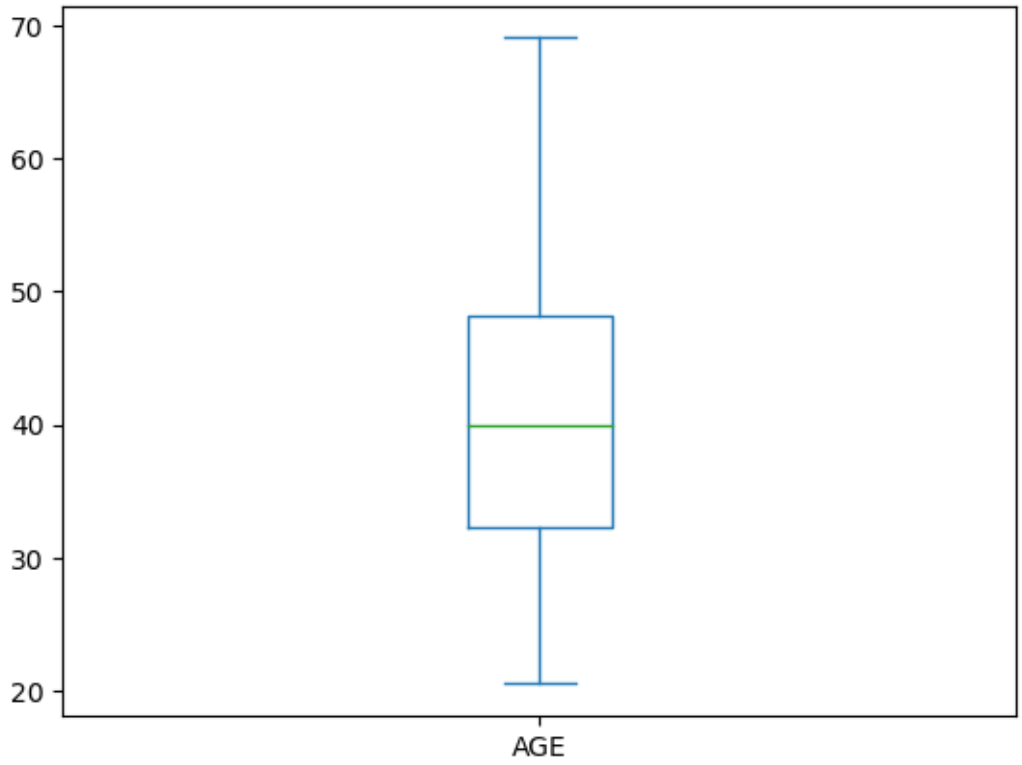
	OBS_30_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE
0	2.0	2.0
1	1.0	0.0
2	0.0	0.0
3	2.0	0.0
4	0.0	0.0

	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE
0	2.0	2.0
1	1.0	0.0

```
2                         0.0                         0.0
3                         2.0                         0.0
4                         0.0                         0.0

   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE        AGE  YEARS_EMPLOYED
0                         0.0              0.26  25.920548        1.745205
1                         0.0              0.62  45.931507        3.254795
2                         0.0              0.56  52.180822        0.616438
3                         0.0              0.65  52.068493        8.326027
4                         0.0              0.32  54.608219        8.323288
```

[82]:
```
df.AGE.plot.box()
```

[82]:
```
<Axes: >
```



[84]:
```
df["AGE_GROUP"] = pd.cut(df.AGE, [20,30,40,50,60,999],labels = ["20-30",␣
↪"30-40", "40-50", "50-60", "60+"])
```

[100]:
```
df.head()
```

[100]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
```

```
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

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

   AMT_GOODS_PRICE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE  \
0         351000.0          Working  Secondary / secondary special
1        1129500.0    State servant               Higher education
2         135000.0          Working  Secondary / secondary special
3         297000.0          Working  Secondary / secondary special
4         513000.0          Working  Secondary / secondary special

     NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  \
0  Single / not married  House / apartment                    0.018801
1               Married  House / apartment                    0.003541
2  Single / not married  House / apartment                    0.010032
3       Civil marriage  House / apartment                    0.008019
4  Single / not married  House / apartment                    0.028663

   DAYS_REGISTRATION OCCUPATION_TYPE       ORGANIZATION_TYPE  \
0             3648.0        Laborers  Business Entity Type 3
1             1186.0      Core staff                  School
2             4260.0        Laborers              Government
3             9833.0        Laborers  Business Entity Type 3
4             4311.0      Core staff                Religion

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

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       0.0                       0.0
3                       2.0                       0.0
4                       0.0                       0.0
```

```
   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE        AGE  YEARS_EMPLOYED  \
0                         0.0              0.26  25.920548        1.745205
1                         0.0              0.62  45.931507        3.254795
2                         0.0              0.56  52.180822        0.616438
3                         0.0              0.65  52.068493        8.326027
4                         0.0              0.32  54.608219        8.323288

  AGE_GROUP
0     20-30
1     40-50
2     50-60
3     50-60
4     50-60
```

[102]:
```
bins = [0,25000,50000,100000,200000,300000,500000,10000000000]
salary = ['0-25000',␣
 ↪'25000-50000','50000-100000','100000-200000','200000-300000','300000-500000',"500000+"]

df['INCOME_RANGE'] = pd.cut(df['AMT_INCOME_TOTAL'],bins,labels=salary)
```

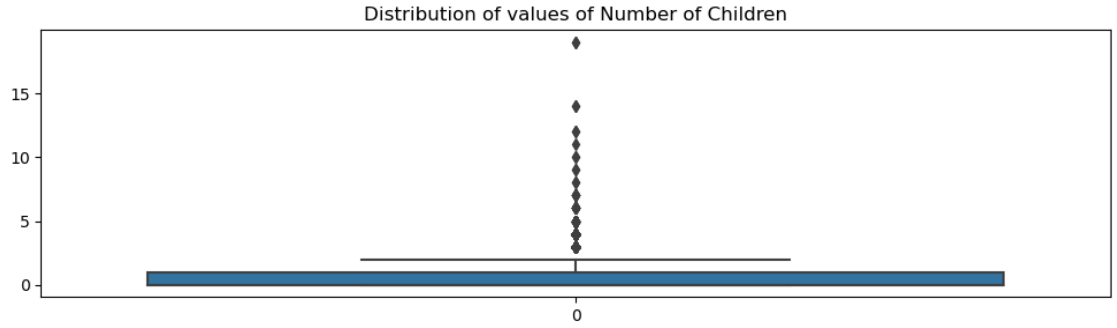
[104]:
```
bins = [0,50000,100000,200000,300000,500000,750000,1000000,1000000000]
amounts =␣
 ↪["0-50000","50000-100000","100000-200000","200000-300000","300000-500000","500000-750000","
df['CREDIT_RANGE']=pd.cut(df['AMT_CREDIT'],bins=bins,labels=amounts)
```

[106]:
```
df.head()
```

[106]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

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

   AMT_GOODS_PRICE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE  \
0         351000.0          Working  Secondary / secondary special
1        1129500.0    State servant               Higher education
2         135000.0          Working  Secondary / secondary special
3         297000.0          Working  Secondary / secondary special
4         513000.0          Working  Secondary / secondary special
```

```
     NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  \
0  Single / not married  House / apartment                    0.018801
1               Married  House / apartment                    0.003541
2  Single / not married  House / apartment                    0.010032
3       Civil marriage  House / apartment                    0.008019
4  Single / not married  House / apartment                    0.028663

   DAYS_REGISTRATION OCCUPATION_TYPE       ORGANIZATION_TYPE  \
0             3648.0        Laborers  Business Entity Type 3
1             1186.0      Core staff                  School
2             4260.0        Laborers              Government
3             9833.0        Laborers  Business Entity Type 3
4             4311.0      Core staff                Religion

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

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       0.0                       0.0
3                       2.0                       0.0
4                       0.0                       0.0

   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE        AGE  YEARS_EMPLOYED  \
0                         0.0              0.26  25.920548        1.745205
1                         0.0              0.62  45.931507        3.254795
2                         0.0              0.56  52.180822        0.616438
3                         0.0              0.65  52.068493        8.326027
4                         0.0              0.32  54.608219        8.323288

  AGE_GROUP   INCOME_RANGE   CREDIT_RANGE
0     20-30  200000-300000  300000-500000
1     40-50  200000-300000       1000000+
2     50-60   50000-100000  100000-200000
3     50-60  100000-200000  300000-500000
4     50-60  100000-200000  500000-750000
```

[108]:
```
plt.figure(figsize = (12,3))
ax = sns.boxplot(df['CNT_CHILDREN'])
ax.set(title ="Distribution of values of Number of Children")
plt.show()
```



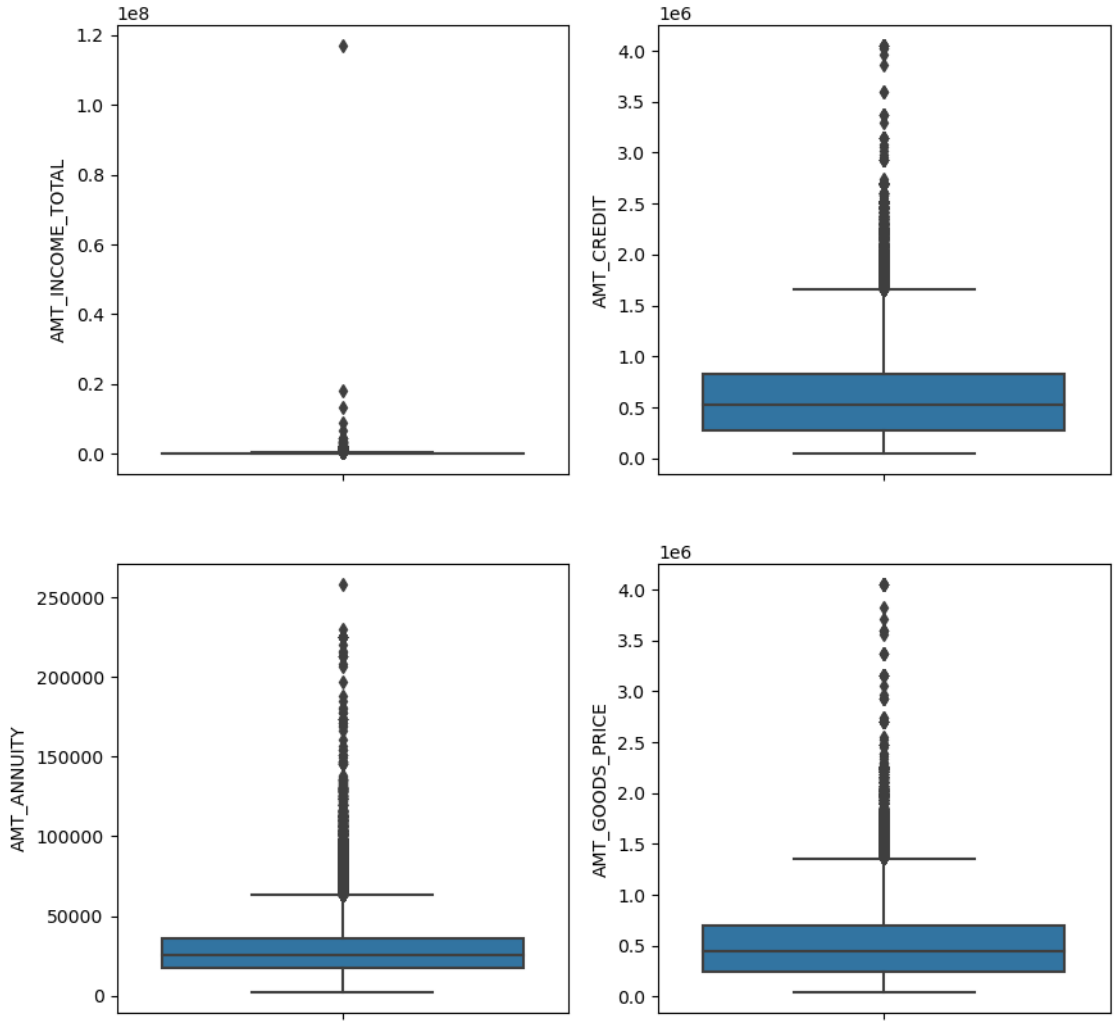
[110]:
```
df['CNT_CHILDREN'].value_counts()
```

[110]:
```
CNT_CHILDREN
0     161911
1      59698
2      26365
3       3629
4        414
5         81
6         19
7          7
8          2
9          2
12         2
10         2
19         2
14         2
11         1
Name: count, dtype: int64
```

[112]:
```
df= df[df['CNT_CHILDREN']<=10]
```

[114]:
```
ax = sns.boxplot(y=df['YEARS_EMPLOYED'])
ax.set(title = 'Distribution of values of "Years Employed"')
plt.show()
```

[116]:
```
col_names = ["AMT_INCOME_TOTAL", "AMT_CREDIT","AMT_ANNUITY", "AMT_GOODS_PRICE"]

fig,axes = plt.subplots(ncols=2,nrows=2,figsize=(10,10))
count=0
for i in range(0,2):
    for j in range(0,2):
        sns.boxplot(y=df[col_names[count]],ax=axes[i,j])
        count+=1
plt.show()
```



[118]:
```
df=df[df['AMT_INCOME_TOTAL']<df['AMT_INCOME_TOTAL'].max()]
df=df[df['AMT_ANNUITY']<df['AMT_ANNUITY'].max()]
```

[120]:
```
df.head()
```

[120]:

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

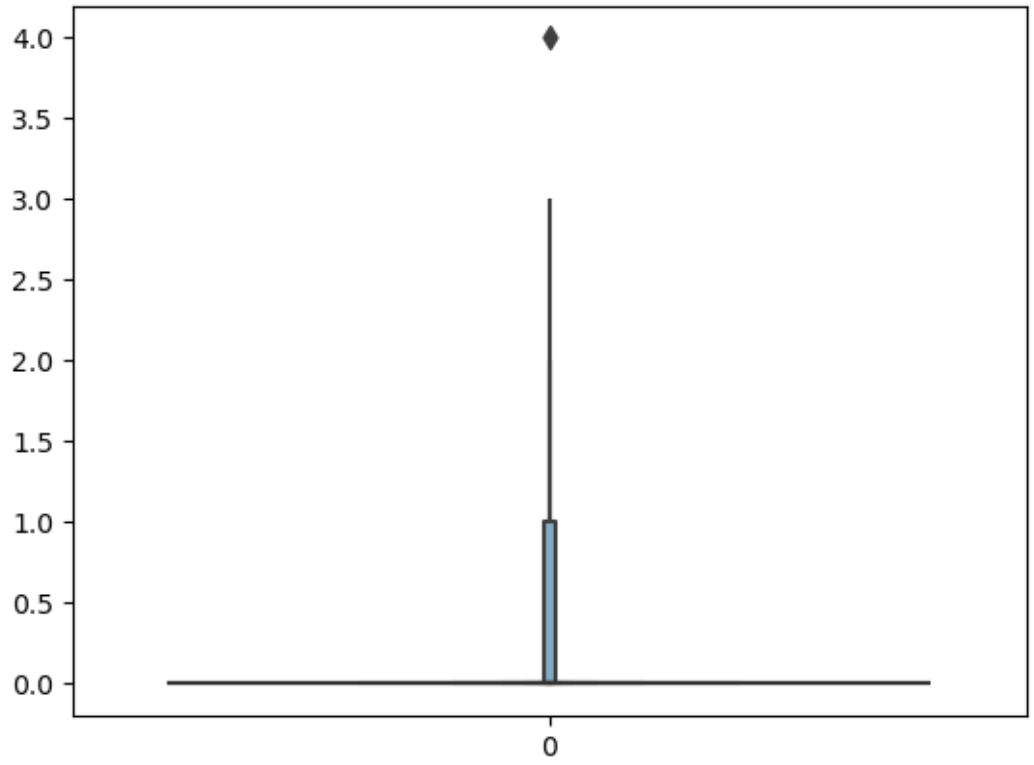
	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	\
0	Y	0	202500.0	406597.5	24700.5	
1	N	0	270000.0	1293502.5	35698.5	

```
2              Y             0          67500.0    135000.0       6750.0
3              Y             0         135000.0    312682.5      29686.5
4              Y             0         121500.0    513000.0      21865.5

   AMT_GOODS_PRICE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE  \
0         351000.0          Working  Secondary / secondary special
1        1129500.0    State servant               Higher education
2         135000.0          Working  Secondary / secondary special
3         297000.0          Working  Secondary / secondary special
4         513000.0          Working  Secondary / secondary special

     NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  \
0  Single / not married  House / apartment                    0.018801
1               Married  House / apartment                    0.003541
2  Single / not married  House / apartment                    0.010032
3        Civil marriage  House / apartment                    0.008019
4  Single / not married  House / apartment                    0.028663

   DAYS_REGISTRATION OCCUPATION_TYPE       ORGANIZATION_TYPE  \
0             3648.0        Laborers  Business Entity Type 3
1             1186.0      Core staff                  School
2             4260.0        Laborers              Government
3             9833.0        Laborers  Business Entity Type 3
4             4311.0      Core staff                Religion

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

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       0.0                       0.0
3                       2.0                       0.0
4                       0.0                       0.0

   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE        AGE  YEARS_EMPLOYED  \
0                         0.0              0.26  25.920548        1.745205
1                         0.0              0.62  45.931507        3.254795
2                         0.0              0.56  52.180822        0.616438
3                         0.0              0.65  52.068493        8.326027
4                         0.0              0.32  54.608219        8.323288

  AGE_GROUP   INCOME_RANGE   CREDIT_RANGE
```

```
0      20-30  200000-300000  300000-500000
1      40-50  200000-300000        1000000+
2      50-60   50000-100000  100000-200000
3      50-60  100000-200000  300000-500000
4      50-60  100000-200000  500000-750000
```

[122]:
```
sns.boxenplot(df.AMT_REQ_CREDIT_BUREAU_HOUR)
```

[122]:
```
<Axes: >
```



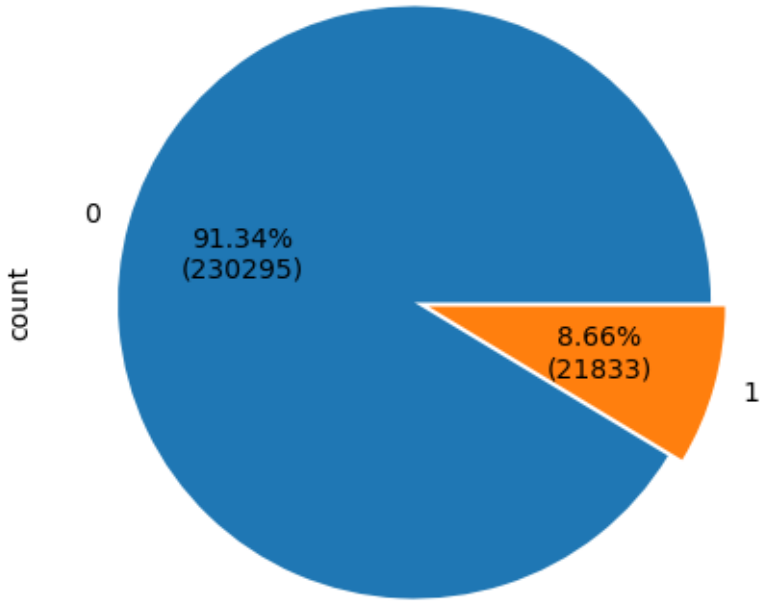
[124]:
```
df.AMT_REQ_CREDIT_BUREAU_HOUR.value_counts()
```

[124]:
```
AMT_REQ_CREDIT_BUREAU_HOUR
0.0    250755
1.0      1318
2.0        46
3.0         8
4.0         1
Name: count, dtype: int64
```

[126]:
```
df['TARGET'].value_counts(normalize = True).sort_values(ascending = False)
```

[126]:
```
TARGET
0    0.913405
1    0.086595
Name: proportion, dtype: float64
```

[128]:
```
total = len(df["TARGET"])
explode = [0, 0.05]

def format_func(x):
    return '{:.2f}%\n({:.0f})'.format(x, total*x/100)

plt.figure(figsize = [5, 5])
df["TARGET"].value_counts().plot.pie(autopct = format_func, explode = explode)

plt.show()
```



[130]:
```
round(len(df[df.TARGET == 0])/len(df[df.TARGET == 1]),2)
```

[130]:
```
10.55
```

[132]:
```
df1 = df[df['TARGET'] == 1] ## Clients with payment difficulties
df0 = df[df['TARGET'] == 0] ## All other clients
```

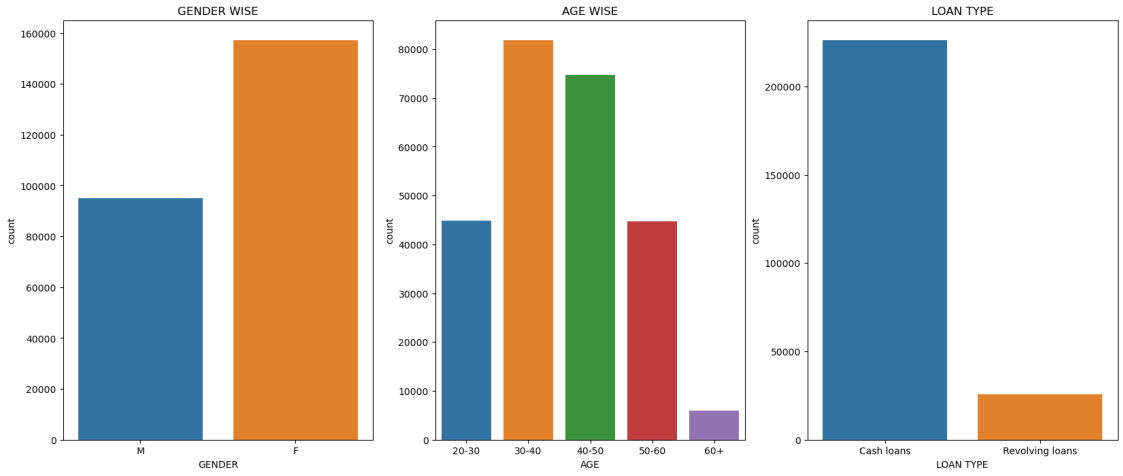
[134]:
```
df.nunique()
```

[134]:
```
SK_ID_CURR                  252128
TARGET                           2
NAME_CONTRACT_TYPE               2
CODE_GENDER                      2
FLAG_OWN_CAR                     2
FLAG_OWN_REALTY                  2
CNT_CHILDREN                    11
AMT_INCOME_TOTAL              2265
AMT_CREDIT                    5331
AMT_ANNUITY                  13189
AMT_GOODS_PRICE                894
NAME_INCOME_TYPE                 7
NAME_EDUCATION_TYPE              5
NAME_FAMILY_STATUS               6
NAME_HOUSING_TYPE                6
REGION_POPULATION_RELATIVE      81
DAYS_REGISTRATION            14419
OCCUPATION_TYPE                 18
ORGANIZATION_TYPE               57
OBS_30_CNT_SOCIAL_CIRCLE        32
DEF_30_CNT_SOCIAL_CIRCLE         9
OBS_60_CNT_SOCIAL_CIRCLE        32
DEF_60_CNT_SOCIAL_CIRCLE         8
AMT_REQ_CREDIT_BUREAU_HOUR       5
EXT_SOURCE_SCORE                84
AGE                          16513
YEARS_EMPLOYED               12573
AGE_GROUP                        5
INCOME_RANGE                     6
CREDIT_RANGE                     8
dtype: int64
```

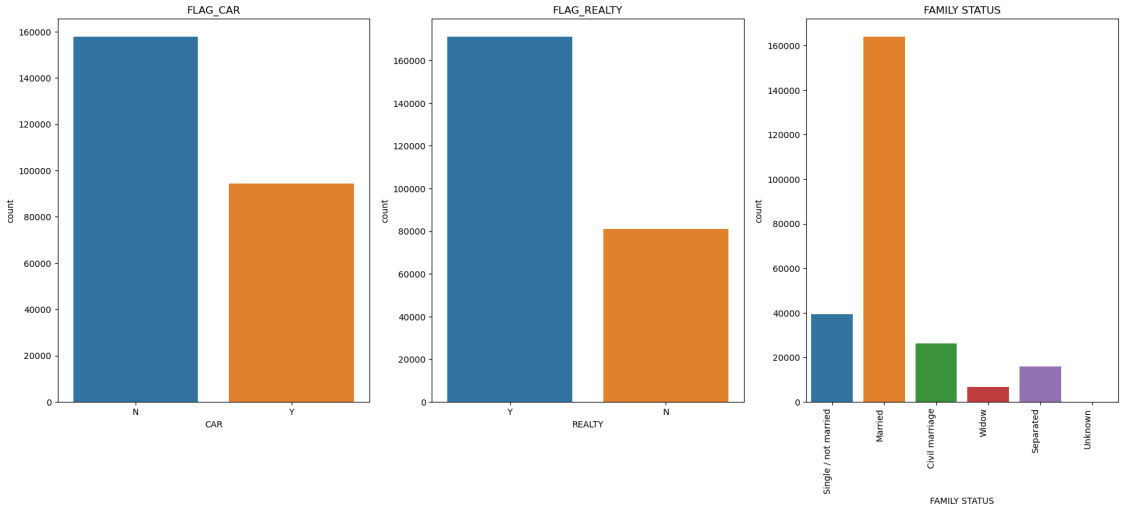
[136]:
```
plt.figure(figsize=(20,8))

plt.subplot(1,3,1)
ax = sns.countplot(x = 'CODE_GENDER',data=df)
ax.set(title = 'GENDER WISE')
ax.set(xlabel='GENDER')

plt.subplot(1,3,2)
ax = sns.countplot(x = 'AGE_GROUP',data=df)
ax.set(title='AGE WISE')
ax.set(xlabel='AGE')

plt.subplot(1,3,3)
```

```
ax = sns.countplot(x = 'NAME_CONTRACT_TYPE',data=df)
ax.set(title = 'LOAN TYPE')
ax.set(xlabel='LOAN TYPE')
```

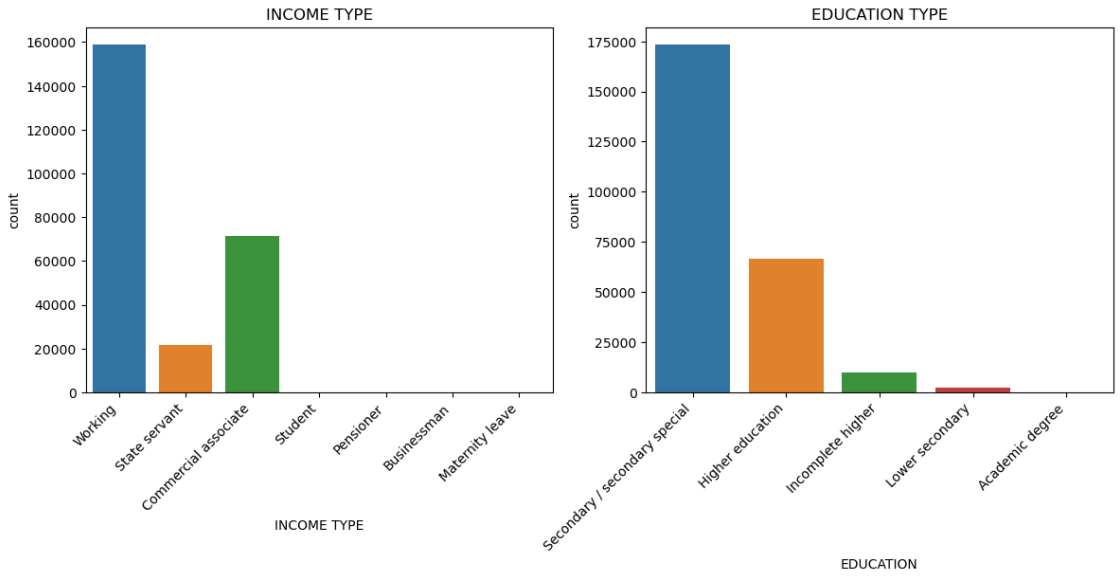
[136]: [Text(0.5, 0, 'LOAN TYPE')]



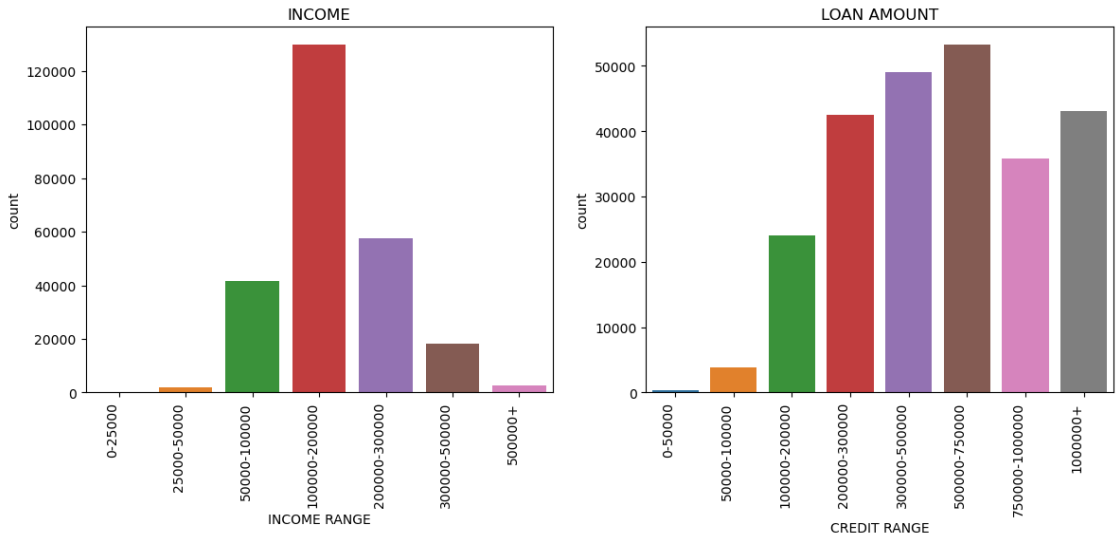
[138]:
```
plt.figure(figsize=(22,8))

plt.subplot(1,3,1)
ax = sns.countplot(x = 'FLAG_OWN_CAR',data=df)
ax.set(title = 'FLAG_CAR')
ax.set(xlabel='CAR')

plt.subplot(1,3,2)
ax = sns.countplot(x = 'FLAG_OWN_REALTY',data=df)
ax.set(title='FLAG_REALTY')
ax.set(xlabel='REALTY')

plt.subplot(1,3,3)
ax = sns.countplot(x = 'NAME_FAMILY_STATUS',data=df)
ax.set(title = 'FAMILY STATUS')
ax.set(xlabel='FAMILY STATUS')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



[140]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'NAME_INCOME_TYPE',data=df)
ax.set(title = 'INCOME TYPE')
ax.set(xlabel='INCOME TYPE')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 45,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'NAME_EDUCATION_TYPE',data=df)
ax.set(title = 'EDUCATION TYPE')
ax.set(xlabel='EDUCATION')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 45,␣
 ↪horizontalalignment='right')
```

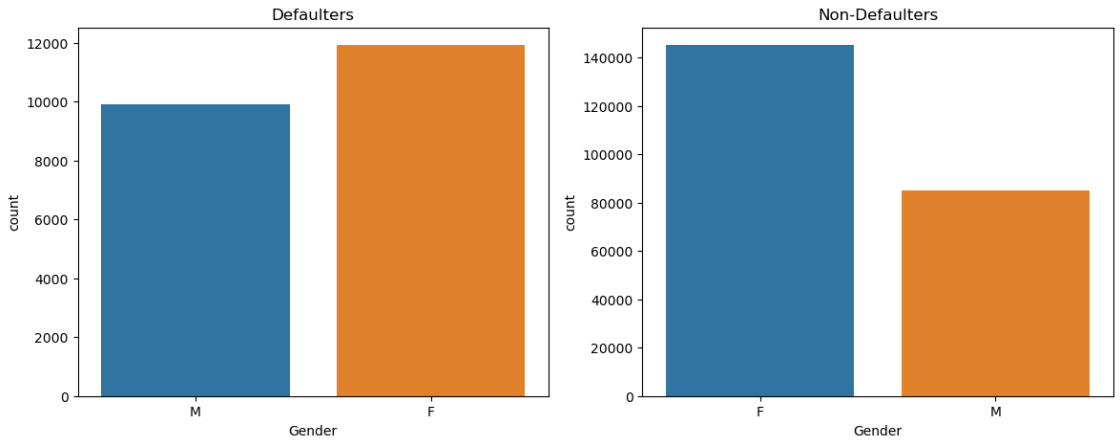


[146]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'INCOME_RANGE',data=df)
ax.set(title = 'INCOME')
ax.set(xlabel='INCOME RANGE')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'CREDIT_RANGE',data=df)
ax.set(title = 'LOAN AMOUNT')
ax.set(xlabel='CREDIT RANGE')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



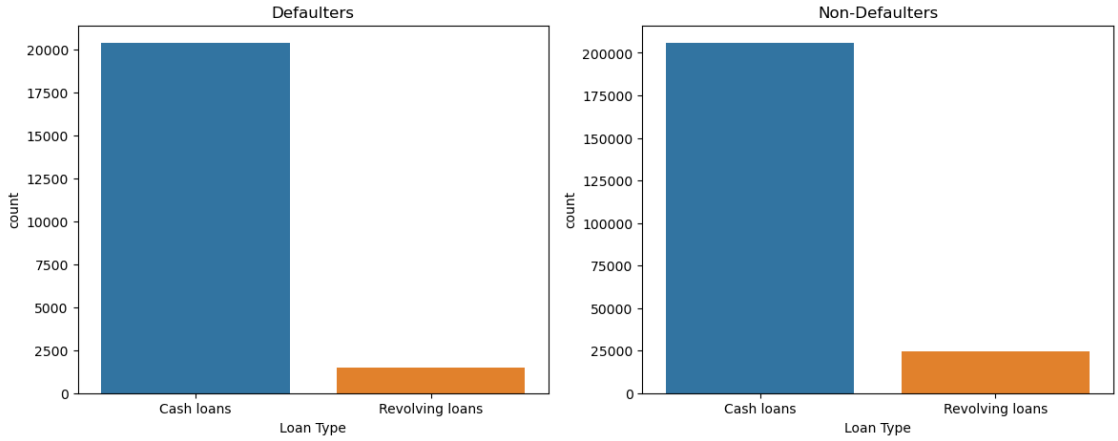
[148]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'CODE_GENDER',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Gender')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'CODE_GENDER',data=df0)
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Gender')
```

[148]:
```
[Text(0.5, 0, 'Gender')]
```



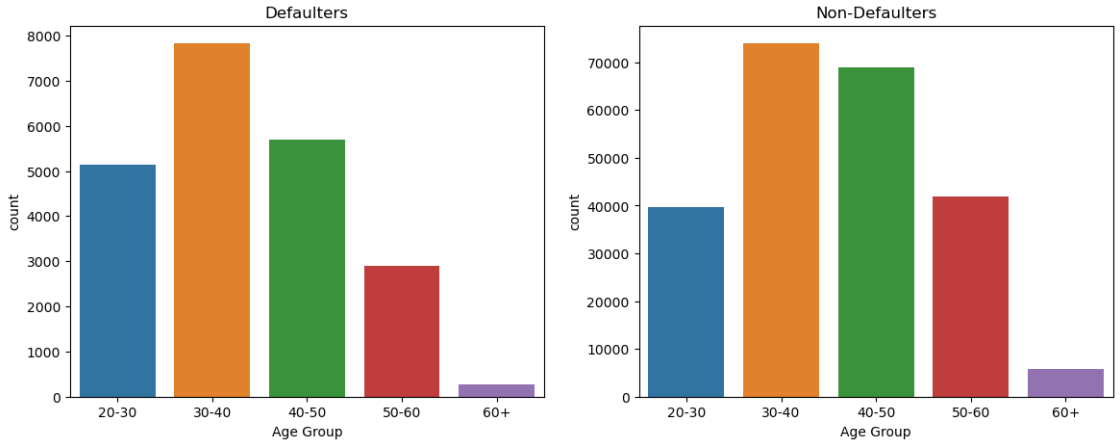
[152]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'NAME_CONTRACT_TYPE',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Loan Type')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'NAME_CONTRACT_TYPE',data=df0)
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Loan Type')
```

[152]:
```
[Text(0.5, 0, 'Loan Type')]
```



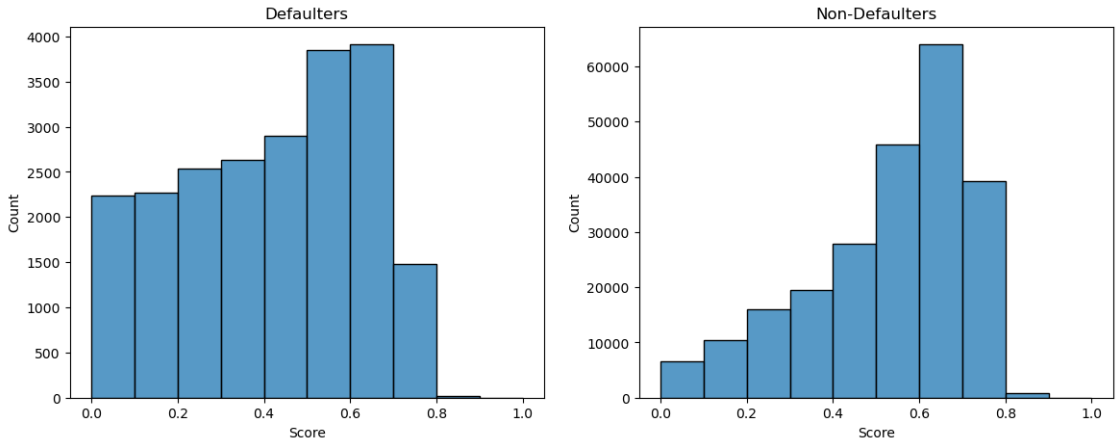
[154]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'AGE_GROUP',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Age Group')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'AGE_GROUP',data=df0)
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Age Group')
```

[154]:
```
[Text(0.5, 0, 'Age Group')]
```

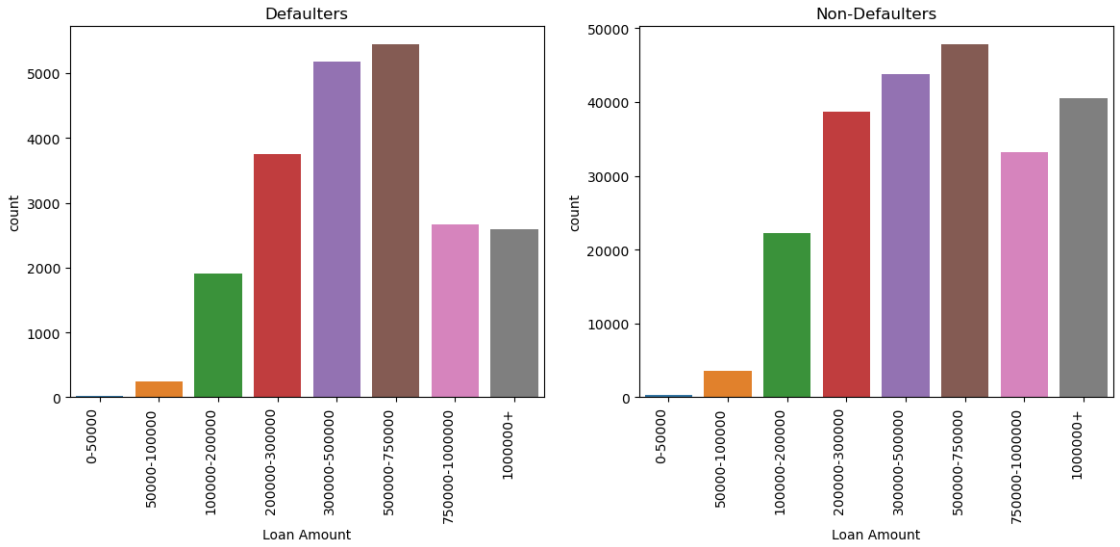


[156]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.histplot(df1.EXT_SOURCE_SCORE, bins = ([0,0.1,.2,.3,.4,.5,.6,.7,.8,.
 ↪9,1]))
ax.set(title = 'Defaulters')
ax.set(xlabel='Score')

plt.subplot(1,2,2)
ax = sns.histplot(df0.EXT_SOURCE_SCORE, bins = ([0,0.1,.2,.3,.4,.5,.6,.7,.8,.
 ↪9,1]))
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Score')
```

[156]:
```
[Text(0.5, 0, 'Score')]
```

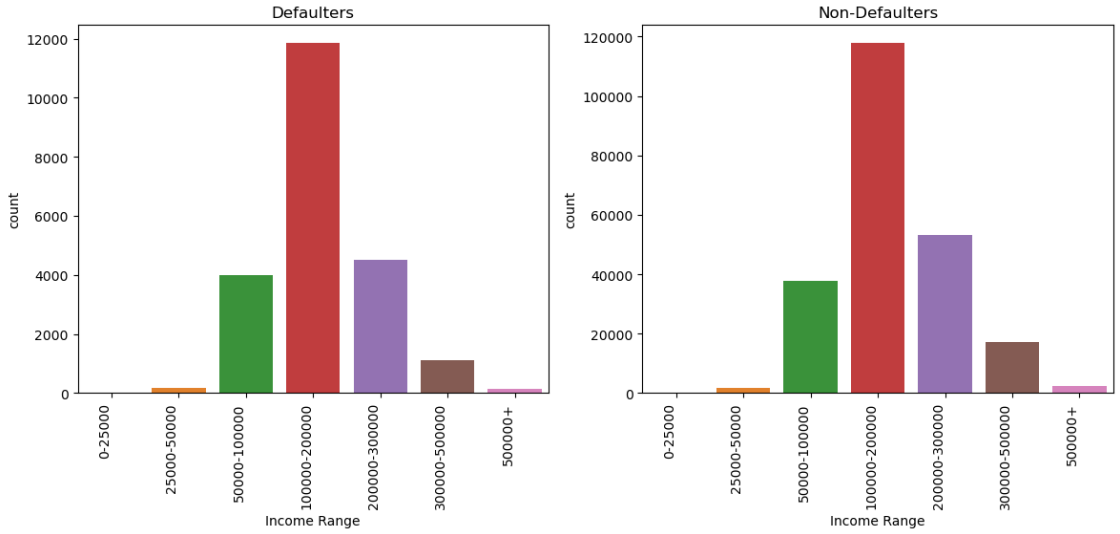


[158]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'CREDIT_RANGE',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Loan Amount')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'CREDIT_RANGE',data=df0)
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Loan Amount')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```

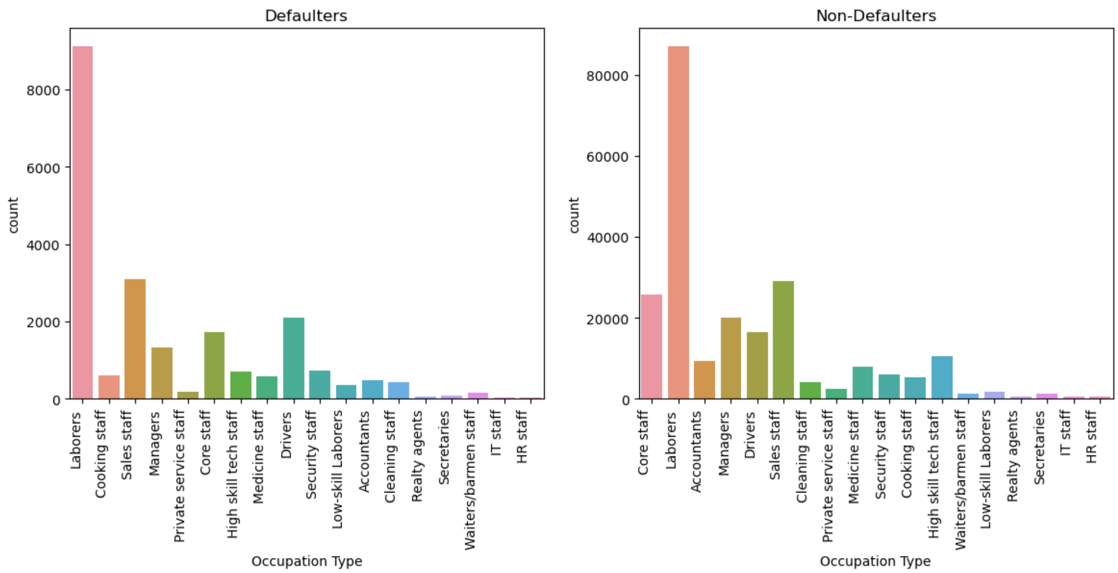


[160]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'INCOME_RANGE',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Income Range')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'INCOME_RANGE',data=df0)
ax.set(title = 'Non-Defaulters')
```

```
ax.set(xlabel='Income Range')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



[162]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'OCCUPATION_TYPE',data=df1)
ax.set(title = 'Defaulters')
ax.set(xlabel='Occupation Type')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'OCCUPATION_TYPE',data=df0)
ax.set(title = 'Non-Defaulters')
ax.set(xlabel='Occupation Type')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



[164]:
```python
cols =
 ['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','AGE','EXT_SOURCE_SCORE','
```

[166]:
```python
df1_correlation = df1[cols]
df1_correlation.head()
```

[166]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	AGE	\
0	202500.0	406597.5	24700.5	351000.0	25.920548	
23	112500.0	979992.0	27076.5	702000.0	51.298630	
36	202500.0	1193580.0	35028.0	855000.0	47.895890	
38	135000.0	288873.0	16258.5	238500.0	36.668493	
82	315000.0	953460.0	64107.0	900000.0	27.942466	

	EXT_SOURCE_SCORE	YEARS_EMPLOYED
0	0.26	1.745205
23	0.55	7.200000
36	0.31	3.457534
38	0.67	9.854795
82	0.43	5.520548

[168]:
```python
df1_correlation.corr()
```

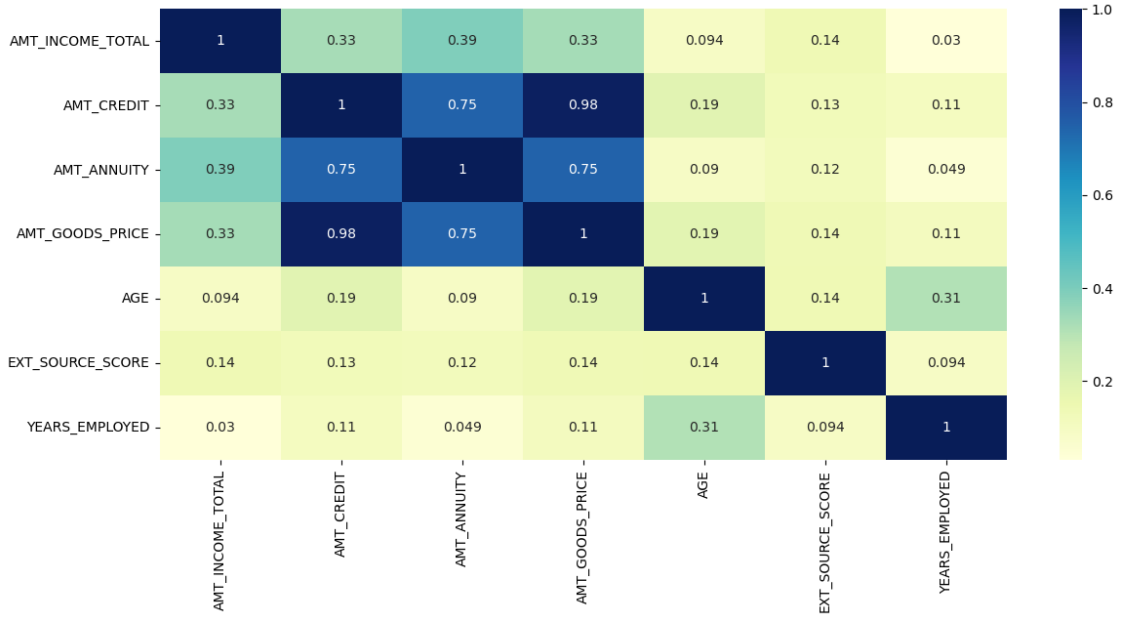
[168]:

	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	\
AMT_INCOME_TOTAL	1.000000	0.327046	0.392914	0.331053	
AMT_CREDIT	0.327046	1.000000	0.748696	0.982463	
AMT_ANNUITY	0.392914	0.748696	1.000000	0.748928	
AMT_GOODS_PRICE	0.331053	0.982463	0.748928	1.000000	
AGE	0.093690	0.189468	0.090460	0.185402	

```
EXT_SOURCE_SCORE          0.139757    0.126002     0.119630          0.136740
YEARS_EMPLOYED            0.029602    0.106062     0.049071          0.111948

                              AGE  EXT_SOURCE_SCORE  YEARS_EMPLOYED
AMT_INCOME_TOTAL         0.093690          0.139757        0.029602
AMT_CREDIT               0.189468          0.126002        0.106062
AMT_ANNUITY              0.090460          0.119630        0.049071
AMT_GOODS_PRICE          0.185402          0.136740        0.111948
AGE                      1.000000          0.142524        0.307061
EXT_SOURCE_SCORE         0.142524          1.000000        0.093797
YEARS_EMPLOYED           0.307061          0.093797        1.000000
```

[176]:
```
plt.figure(figsize=(14,6))
sns.heatmap(df1_correlation.corr(),cmap="YlGnBu",annot=True,cbar = True)
```

[176]: <Axes: >



[172]:
```
df0_correlation = df0[cols]
df0_correlation.head()
```

[172]:
```
   AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE        AGE  \
1          270000.0   1293502.5      35698.5        1129500.0  45.931507
2           67500.0    135000.0       6750.0         135000.0  52.180822
3          135000.0    312682.5      29686.5         297000.0  52.068493
4          121500.0    513000.0      21865.5         513000.0  54.608219
5           99000.0    490495.5      27517.5         454500.0  46.413699
```

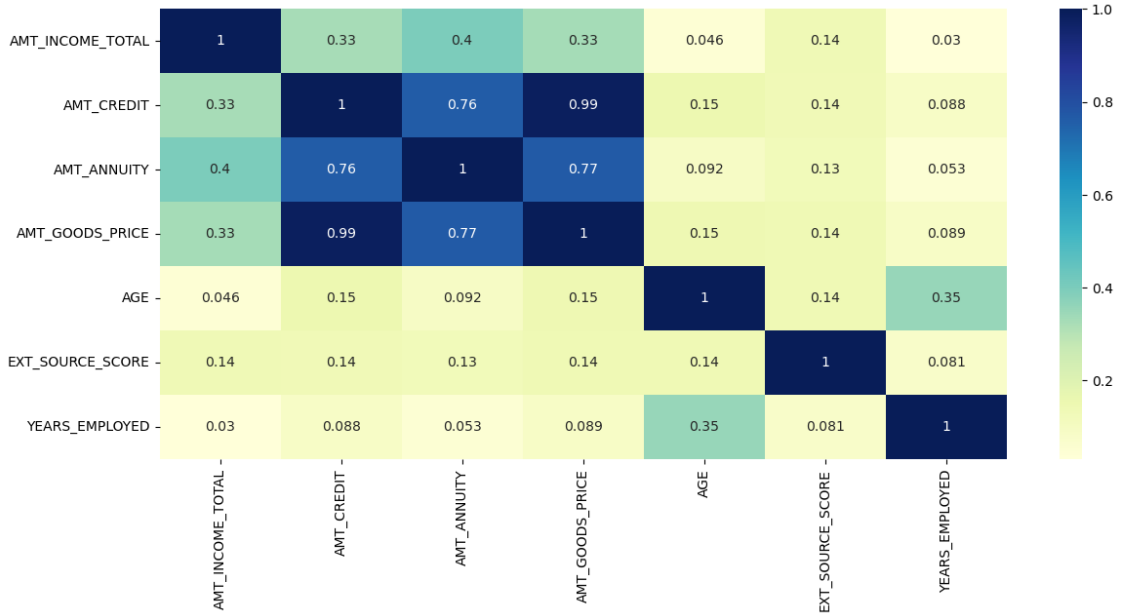
```
   EXT_SOURCE_SCORE  YEARS_EMPLOYED
1              0.62        3.254795
2              0.56        0.616438
3              0.65        8.326027
4              0.32        8.323288
5              0.35        4.350685
```

[178]:
```
df0_correlation.corr()
```

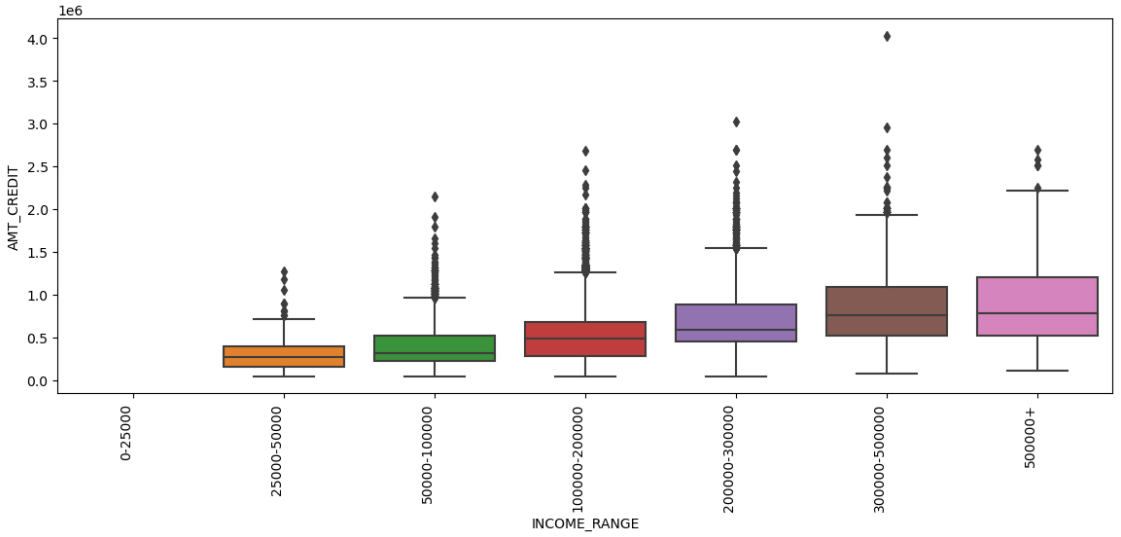
[178]:
```
                  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE  \
AMT_INCOME_TOTAL          1.000000    0.326033     0.400625         0.333162
AMT_CREDIT                0.326033    1.000000     0.762054         0.986468
AMT_ANNUITY               0.400625    0.762054     1.000000         0.766577
AMT_GOODS_PRICE           0.333162    0.986468     0.766577         1.000000
AGE                       0.045540    0.152667     0.091599         0.146866
EXT_SOURCE_SCORE          0.137539    0.135709     0.126708         0.142115
YEARS_EMPLOYED            0.030120    0.087546     0.052559         0.088666

                        AGE  EXT_SOURCE_SCORE  YEARS_EMPLOYED
AMT_INCOME_TOTAL   0.045540          0.137539        0.030120
AMT_CREDIT         0.152667          0.135709        0.087546
AMT_ANNUITY        0.091599          0.126708        0.052559
AMT_GOODS_PRICE    0.146866          0.142115        0.088666
AGE                1.000000          0.142070        0.352664
EXT_SOURCE_SCORE   0.142070          1.000000        0.081418
YEARS_EMPLOYED     0.352664          0.081418        1.000000
```

[180]:
```
plt.figure(figsize=(14,6))
sns.heatmap(df0_correlation.corr(),cmap="YlGnBu",annot=True,cbar = True)
```

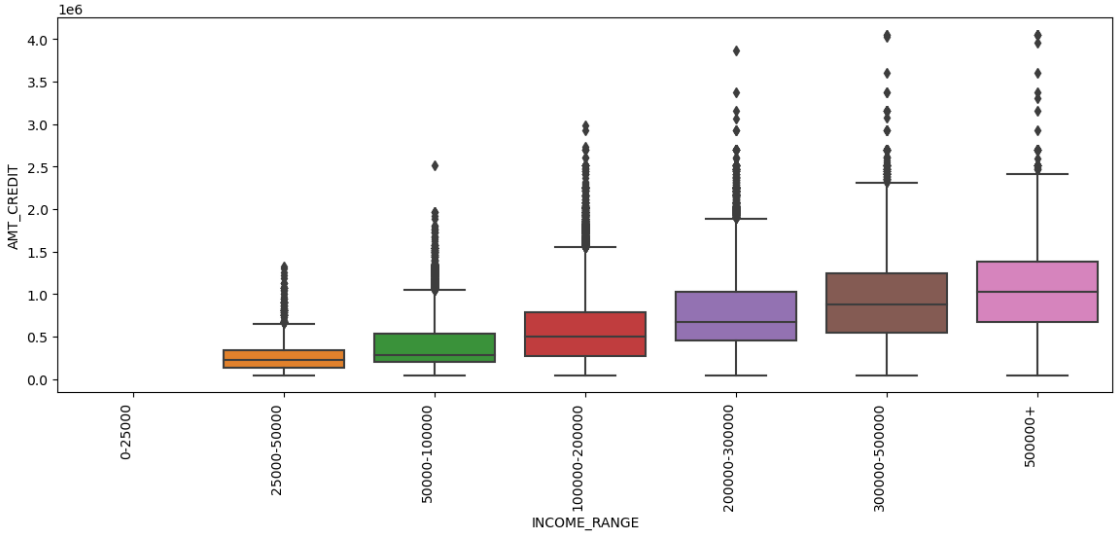
[180]:
```
<Axes: >
```



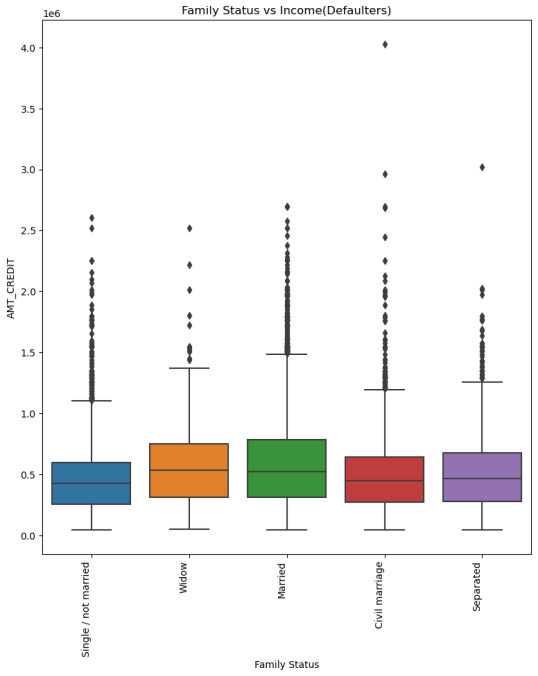
[182]:
```
plt.figure(figsize=(14,5))
ax = sns.boxplot(df1, x="INCOME_RANGE", y="AMT_CREDIT")
plt.title=("Income Range VS Credit Amount (Defaulters)")
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
plt.show()
```

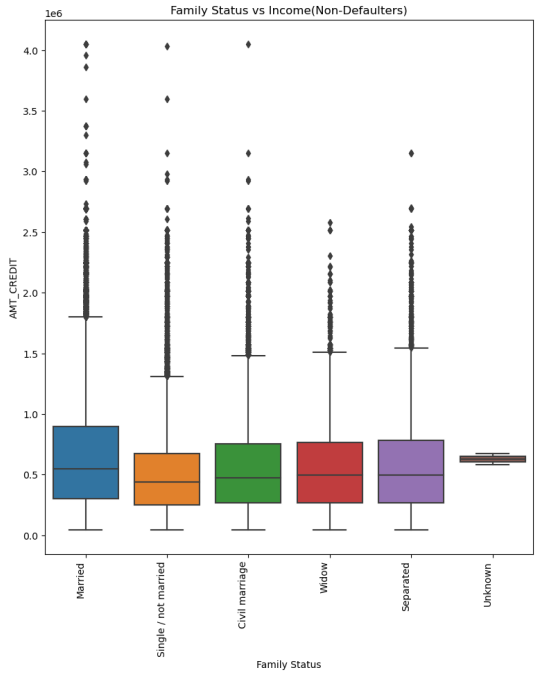


[244]:
```
plt.figure(figsize=(14,5))
ax = sns.boxplot(df0, x="INCOME_RANGE", y="AMT_CREDIT")
plt.title=("Income Range VS Credit Amount (Non-Defaulters)")
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
plt.show()
```

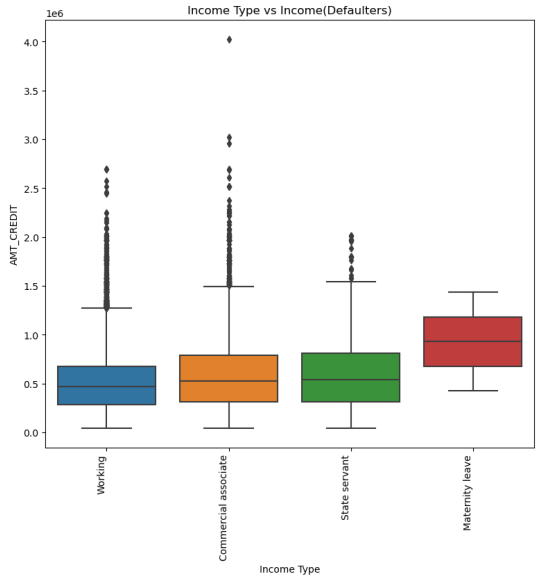


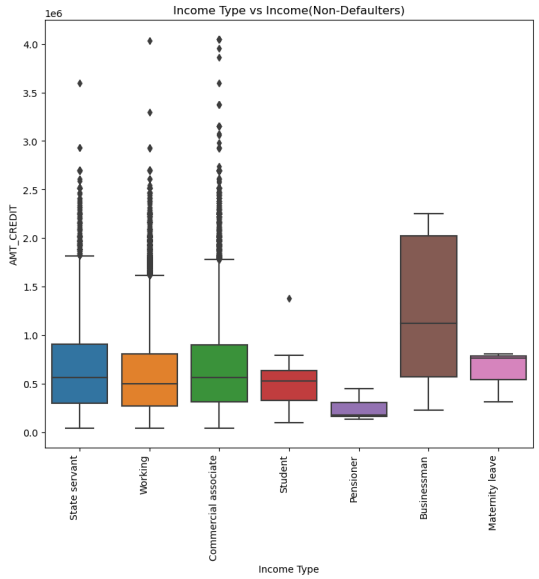
[246]:
```
plt.figure(figsize=(20,10))

plt.subplot(1,2,1)
ax = sns.boxplot(df1, x="NAME_FAMILY_STATUS", y="AMT_CREDIT")
ax.set(title = "Family Status vs Income(Defaulters)")
ax.set(xlabel='Family Status')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.boxplot(df0, x="NAME_FAMILY_STATUS", y="AMT_CREDIT")
ax.set(title = "Family Status vs Income(Non-Defaulters)")
ax.set(xlabel='Family Status')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



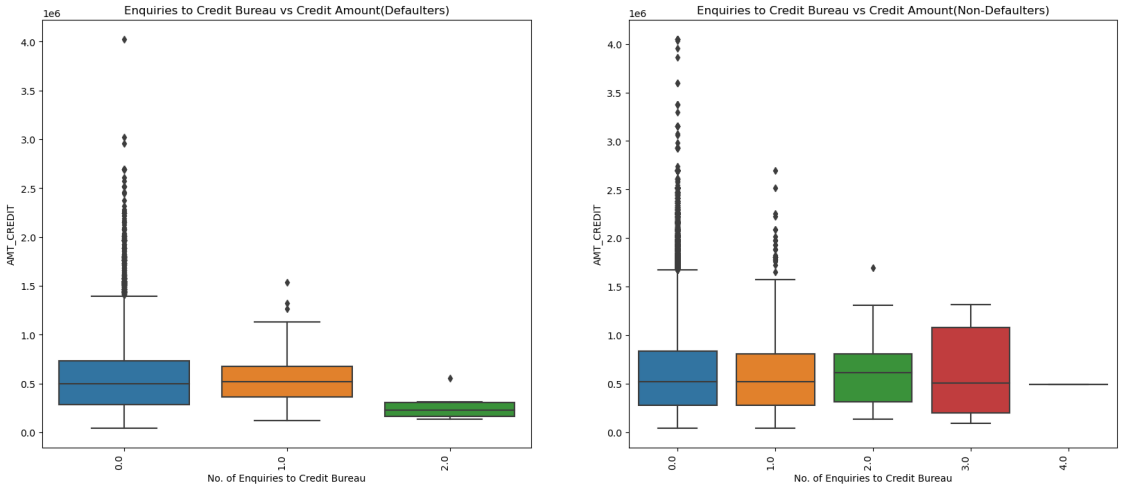


[248]:
```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
ax = sns.boxplot(df1, x="NAME_INCOME_TYPE", y="AMT_CREDIT")
ax.set(title = "Income Type vs Income(Defaulters)")
ax.set(xlabel='Income Type')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.boxplot(df0, x="NAME_INCOME_TYPE", y="AMT_CREDIT")
ax.set(title = "Income Type vs Income(Non-Defaulters)")
ax.set(xlabel='Income Type')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```

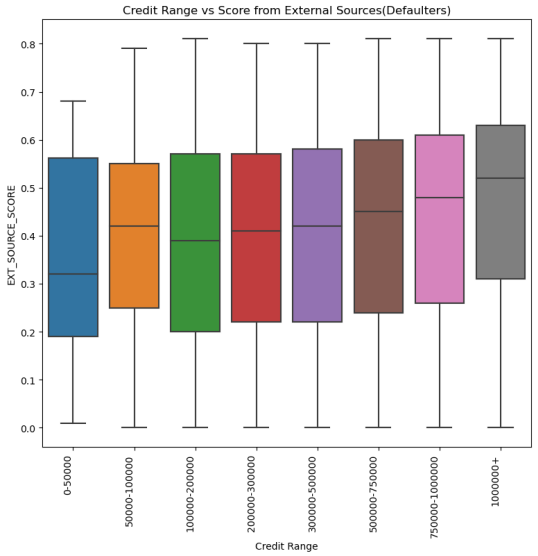


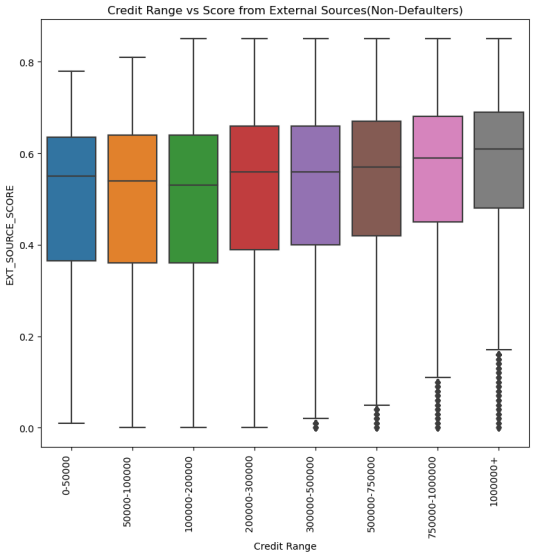


[250]:
```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
ax = sns.boxplot(df1, x="AMT_REQ_CREDIT_BUREAU_HOUR", y="AMT_CREDIT")
ax.set(title = "Enquiries to Credit Bureau vs Credit Amount(Defaulters)")
ax.set(xlabel='No. of Enquiries to Credit Bureau')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.boxplot(df0, x="AMT_REQ_CREDIT_BUREAU_HOUR", y="AMT_CREDIT")
ax.set(title = "Enquiries to Credit Bureau vs Credit Amount(Non-Defaulters)")
ax.set(xlabel='No. of Enquiries to Credit Bureau')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```



[252]:
```
plt.figure(figsize=(20,8))

plt.subplot(1,2,1)
ax = sns.boxplot(df1, x="CREDIT_RANGE", y="EXT_SOURCE_SCORE")
ax.set(title = "Credit Range vs Score from External Sources(Defaulters)")
ax.set(xlabel='Credit Range')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')

plt.subplot(1,2,2)
ax = sns.boxplot(df0, x="CREDIT_RANGE", y="EXT_SOURCE_SCORE")
ax.set(title = "Credit Range vs Score from External Sources(Non-Defaulters)")
ax.set(xlabel='Credit Range')
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```





[184]:
```
df_prev = pd.read_csv("previous_application.csv")
```

[186]:
```
df_prev.head()
```

[186]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	HOUR_APPR_PROCESS_START	FLAG_LAST_APPL_PER_CONTRACT	NFLAG_LAST_APPL_IN_DAY	RATE_DOWN_PAYMENT	RATE_INTEREST_PRIMARY
0	2030495	271877	Consumer loans	1730.430	17145.0	17145.0	0.0	17145.0	SATURDAY	15	Y	1	0.0	0.182832
1	2802425	108129	Cash loans	25188.615	607500.0	679671.0	NaN	607500.0	THURSDAY	11	Y	1	NaN	NaN
2	2523466	122040	Cash loans	15060.735	112500.0	136444.5	NaN	112500.0	TUESDAY	11	Y			
3	2819243	176158	Cash loans	47041.335	450000.0	470790.0	NaN	450000.0	MONDAY	7	Y			
4	1784265	202054	Cash loans	31924.395	337500.0	404055.0	NaN	337500.0	THURSDAY	9	Y			

```
2                       1                NaN                     NaN
3                       1                NaN                     NaN
4                       1                NaN                     NaN

   RATE_INTEREST_PRIVILEGED NAME_CASH_LOAN_PURPOSE NAME_CONTRACT_STATUS  \
0                  0.867336                   XAP             Approved
1                       NaN                   XNA             Approved
2                       NaN                   XNA             Approved
3                       NaN                   XNA             Approved
4                       NaN               Repairs              Refused

   DAYS_DECISION      NAME_PAYMENT_TYPE CODE_REJECT_REASON  NAME_TYPE_SUITE  \
0            -73  Cash through the bank               XAP              NaN
1           -164                    XNA               XAP    Unaccompanied
2           -301  Cash through the bank               XAP  Spouse, partner
3           -512  Cash through the bank               XAP              NaN
4           -781  Cash through the bank                HC              NaN

  NAME_CLIENT_TYPE NAME_GOODS_CATEGORY NAME_PORTFOLIO NAME_PRODUCT_TYPE  \
0         Repeater              Mobile            POS               XNA
1         Repeater                 XNA           Cash            x-sell
2         Repeater                 XNA           Cash            x-sell
3         Repeater                 XNA           Cash            x-sell
4         Repeater                 XNA           Cash           walk-in

             CHANNEL_TYPE  SELLERPLACE_AREA NAME_SELLER_INDUSTRY  \
0            Country-wide                35         Connectivity
1          Contact center                -1                  XNA
2  Credit and cash offices                -1                  XNA
3  Credit and cash offices                -1                  XNA
4  Credit and cash offices                -1                  XNA

   CNT_PAYMENT NAME_YIELD_GROUP       PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  \
0         12.0           middle  POS mobile with interest            365243.0
1         36.0       low_action          Cash X-Sell: low            365243.0
2         12.0             high         Cash X-Sell: high            365243.0
3         12.0           middle       Cash X-Sell: middle            365243.0
4         24.0             high         Cash Street: high                 NaN

   DAYS_FIRST_DUE  DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION  \
0           -42.0                      300.0          -42.0             -37.0
1          -134.0                      916.0       365243.0          365243.0
2          -271.0                       59.0       365243.0          365243.0
3          -482.0                     -152.0         -182.0            -177.0
4             NaN                        NaN            NaN               NaN

   NFLAG_INSURED_ON_APPROVAL
```

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

[188]:
```
df_prev.shape
```

[188]:
```
(1048575, 37)
```

[194]:
```
df_prev.isnull().sum()
```

[194]:
```
SK_ID_PREV                       0
SK_ID_CURR                       0
NAME_CONTRACT_TYPE               0
AMT_ANNUITY                 233009
AMT_APPLICATION                  0
AMT_CREDIT                       0
AMT_DOWN_PAYMENT            559396
AMT_GOODS_PRICE             240965
WEEKDAY_APPR_PROCESS_START       0
HOUR_APPR_PROCESS_START          0
FLAG_LAST_APPL_PER_CONTRACT      0
NFLAG_LAST_APPL_IN_DAY           0
RATE_DOWN_PAYMENT           559396
RATE_INTEREST_PRIMARY      1044854
RATE_INTEREST_PRIVILEGED   1044854
NAME_CASH_LOAN_PURPOSE           0
NAME_CONTRACT_STATUS             0
DAYS_DECISION                    0
NAME_PAYMENT_TYPE                0
CODE_REJECT_REASON               0
NAME_TYPE_SUITE             515140
NAME_CLIENT_TYPE                 0
NAME_GOODS_CATEGORY              0
NAME_PORTFOLIO                   0
NAME_PRODUCT_TYPE                0
CHANNEL_TYPE                     0
SELLERPLACE_AREA                 0
NAME_SELLER_INDUSTRY             0
CNT_PAYMENT                 233006
NAME_YIELD_GROUP                 0
PRODUCT_COMBINATION            224
DAYS_FIRST_DRAWING          420708
DAYS_FIRST_DUE              420708
DAYS_LAST_DUE_1ST_VERSION   420708
DAYS_LAST_DUE               420708
```

```
DAYS_TERMINATION                420708
NFLAG_INSURED_ON_APPROVAL       420708
dtype: int64
```

[196]:
```
cols_missing = 100*df_prev.isnull().mean()
cols_missing
```

[196]:
```
SK_ID_PREV                     0.000000
SK_ID_CURR                     0.000000
NAME_CONTRACT_TYPE             0.000000
AMT_ANNUITY                   22.221491
AMT_APPLICATION                0.000000
AMT_CREDIT                     0.000000
AMT_DOWN_PAYMENT              53.348211
AMT_GOODS_PRICE               22.980235
WEEKDAY_APPR_PROCESS_START     0.000000
HOUR_APPR_PROCESS_START        0.000000
FLAG_LAST_APPL_PER_CONTRACT    0.000000
NFLAG_LAST_APPL_IN_DAY         0.000000
RATE_DOWN_PAYMENT             53.348211
RATE_INTEREST_PRIMARY         99.645137
RATE_INTEREST_PRIVILEGED      99.645137
NAME_CASH_LOAN_PURPOSE         0.000000
NAME_CONTRACT_STATUS           0.000000
DAYS_DECISION                  0.000000
NAME_PAYMENT_TYPE              0.000000
CODE_REJECT_REASON             0.000000
NAME_TYPE_SUITE               49.127626
NAME_CLIENT_TYPE               0.000000
NAME_GOODS_CATEGORY            0.000000
NAME_PORTFOLIO                 0.000000
NAME_PRODUCT_TYPE              0.000000
CHANNEL_TYPE                   0.000000
SELLERPLACE_AREA               0.000000
NAME_SELLER_INDUSTRY           0.000000
CNT_PAYMENT                   22.221205
NAME_YIELD_GROUP               0.000000
PRODUCT_COMBINATION            0.021362
DAYS_FIRST_DRAWING            40.121880
DAYS_FIRST_DUE                40.121880
DAYS_LAST_DUE_1ST_VERSION     40.121880
DAYS_LAST_DUE                 40.121880
DAYS_TERMINATION              40.121880
NFLAG_INSURED_ON_APPROVAL     40.121880
dtype: float64
```

[198]:
```
missing_cols_40 = cols_missing[cols_missing.values > 40].index.to_list()
missing_cols_40
```

[198]:
```
['AMT_DOWN_PAYMENT',
 'RATE_DOWN_PAYMENT',
 'RATE_INTEREST_PRIMARY',
 'RATE_INTEREST_PRIVILEGED',
 'NAME_TYPE_SUITE',
 'DAYS_FIRST_DRAWING',
 'DAYS_FIRST_DUE',
 'DAYS_LAST_DUE_1ST_VERSION',
 'DAYS_LAST_DUE',
 'DAYS_TERMINATION',
 'NFLAG_INSURED_ON_APPROVAL']
```

[200]:
```
df_prev = df_prev.drop(labels = missing_cols_40, axis = 1)
```

[202]:
```
df_prev['NAME_CASH_LOAN_PURPOSE'].value_counts(normalize = True)
```

[202]:
```
NAME_CASH_LOAN_PURPOSE
XAP                                 0.555720
XNA                                 0.402944
Repairs                             0.014083
Other                               0.009216
Urgent needs                        0.005063
Buying a used car                   0.001701
Building a house or an annex        0.001604
Everyday expenses                   0.001418
Medicine                            0.001337
Payments on other loans             0.001154
Education                           0.000930
Journey                             0.000710
Purchase of electronic equipment    0.000614
Buying a new car                    0.000607
Wedding / gift / holiday            0.000566
Buying a home                       0.000521
Car repairs                         0.000470
Furniture                           0.000453
Buying a holiday home / land        0.000319
Business development                0.000242
Gasification / water supply         0.000194
Buying a garage                     0.000080
Hobby                               0.000030
Money for a third person            0.000015
Refusal to name the goal            0.000008
Name: proportion, dtype: float64
```

[204]:
```
df_prev['DAYS_DECISION'].value_counts()
```

[204]:
```
DAYS_DECISION
-238     1502
-245     1498
-210     1469
-224     1444
-196     1444
         …
-2915      99
-2893      98
-2909      98
-2902      96
-2894      93
Name: count, Length: 2921, dtype: int64
```

[206]:
```
df_prev['DAYS_DECISION']= df_prev['DAYS_DECISION'].abs()
```

[208]:
```
df_prev['DAYS_DECISION'].value_counts()
```

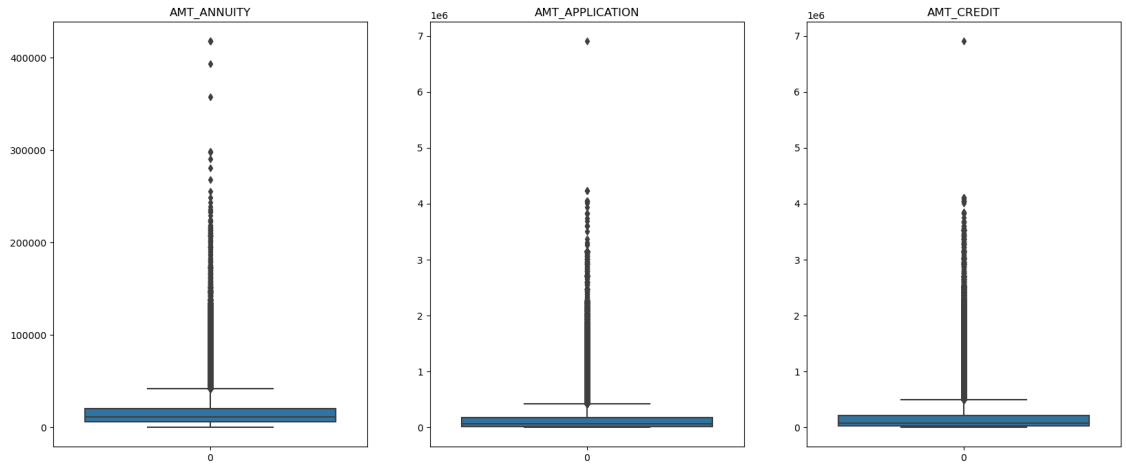
[208]:
```
DAYS_DECISION
238     1502
245     1498
210     1469
224     1444
196     1444
        …
2915      99
2893      98
2909      98
2902      96
2894      93
Name: count, Length: 2921, dtype: int64
```

[212]:
```
100*df_prev.NAME_CLIENT_TYPE.value_counts(normalize = True)
```

[212]:
```
NAME_CLIENT_TYPE
Repeater     73.619627
New          18.118876
Refreshed     8.145435
XNA           0.116062
Name: proportion, dtype: float64
```

[214]:
```
plt.figure(figsize=(20,8))

plt.subplot(1,3,1)
ax = sns.boxplot(df_prev.AMT_ANNUITY)
```

```
ax.set(title = "AMT_ANNUITY")

plt.subplot(1,3,2)
ax = sns.boxplot(df_prev.AMT_APPLICATION)
ax.set(title = "AMT_APPLICATION")

plt.subplot(1,3,3)
ax = sns.boxplot(df_prev.AMT_CREDIT)
ax.set(title = "AMT_CREDIT")
plt.show()
```



[216]:
```
df_prev.head()
```

[216]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	\
0	2030495	271877	Consumer loans	1730.430	17145.0	
1	2802425	108129	Cash loans	25188.615	607500.0	
2	2523466	122040	Cash loans	15060.735	112500.0	
3	2819243	176158	Cash loans	47041.335	450000.0	
4	1784265	202054	Cash loans	31924.395	337500.0	

	AMT_CREDIT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START	\
0	17145.0	17145.0	SATURDAY	
1	679671.0	607500.0	THURSDAY	
2	136444.5	112500.0	TUESDAY	
3	470790.0	450000.0	MONDAY	
4	404055.0	337500.0	THURSDAY	

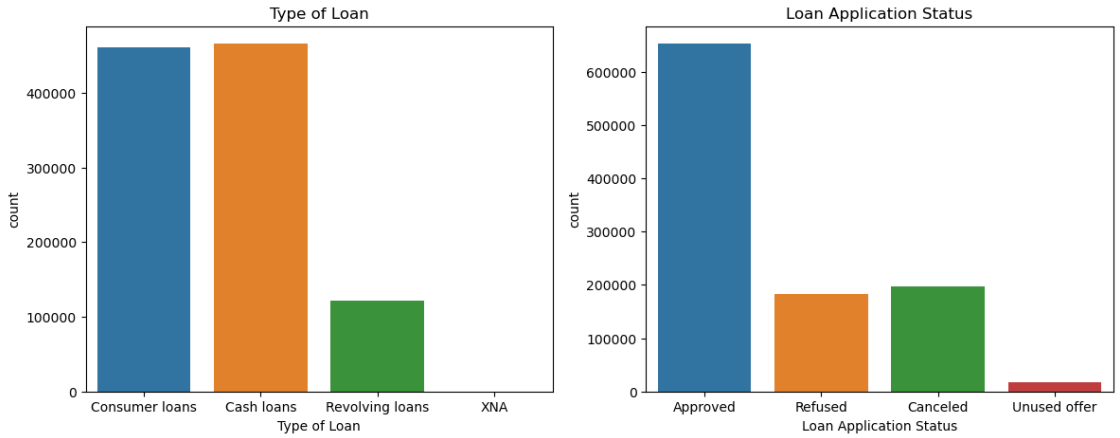
	HOUR_APPR_PROCESS_START	FLAG_LAST_APPL_PER_CONTRACT	\
0	15	Y	
1	11	Y	
2	11	Y	

```
3                       7                          Y
4                       9                          Y

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

   DAYS_DECISION      NAME_PAYMENT_TYPE CODE_REJECT_REASON NAME_CLIENT_TYPE  \
0             73  Cash through the bank               XAP         Repeater
1            164                    XNA               XAP         Repeater
2            301  Cash through the bank               XAP         Repeater
3            512  Cash through the bank               XAP         Repeater
4            781  Cash through the bank                HC         Repeater

  NAME_GOODS_CATEGORY NAME_PORTFOLIO NAME_PRODUCT_TYPE  \
0              Mobile            POS               XNA
1                 XNA           Cash            x-sell
2                 XNA           Cash            x-sell
3                 XNA           Cash            x-sell
4                 XNA           Cash           walk-in

             CHANNEL_TYPE  SELLERPLACE_AREA NAME_SELLER_INDUSTRY  \
0            Country-wide                35         Connectivity
1          Contact center                -1                  XNA
2  Credit and cash offices                -1                  XNA
3  Credit and cash offices                -1                  XNA
4  Credit and cash offices                -1                  XNA

   CNT_PAYMENT NAME_YIELD_GROUP       PRODUCT_COMBINATION
0         12.0           middle  POS mobile with interest
1         36.0       low_action          Cash X-Sell: low
2         12.0             high         Cash X-Sell: high
3         12.0           middle       Cash X-Sell: middle
4         24.0             high         Cash Street: high
```

[218]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'NAME_CONTRACT_TYPE',data=df_prev)
ax.set(title = "Type of Loan")
ax.set(xlabel='Type of Loan')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'NAME_CONTRACT_STATUS',data=df_prev)
```

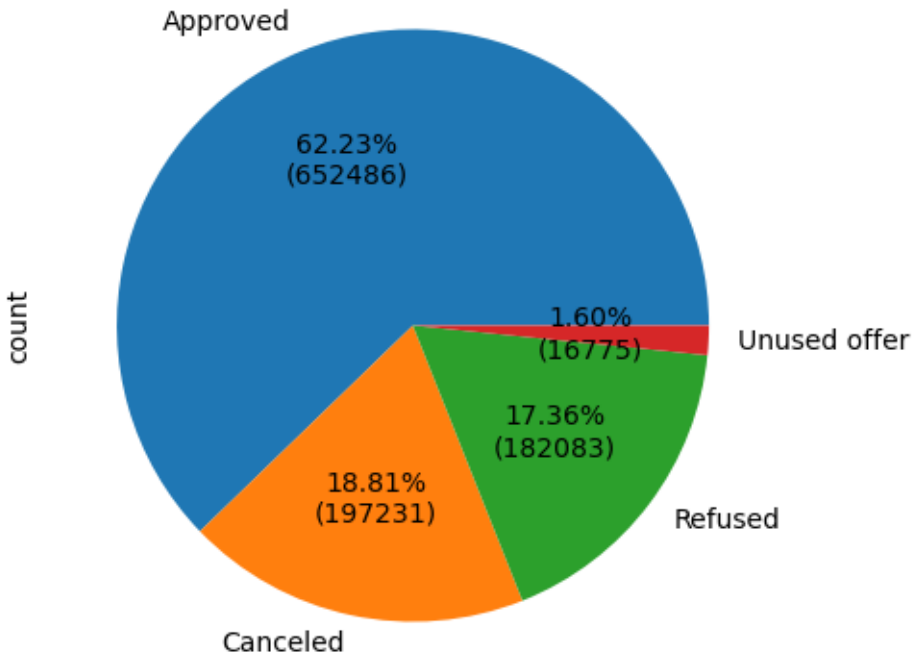
```
ax.set(title = "Loan Application Status")
ax.set(xlabel='Loan Application Status')
```

[218]: [Text(0.5, 0, 'Loan Application Status')]

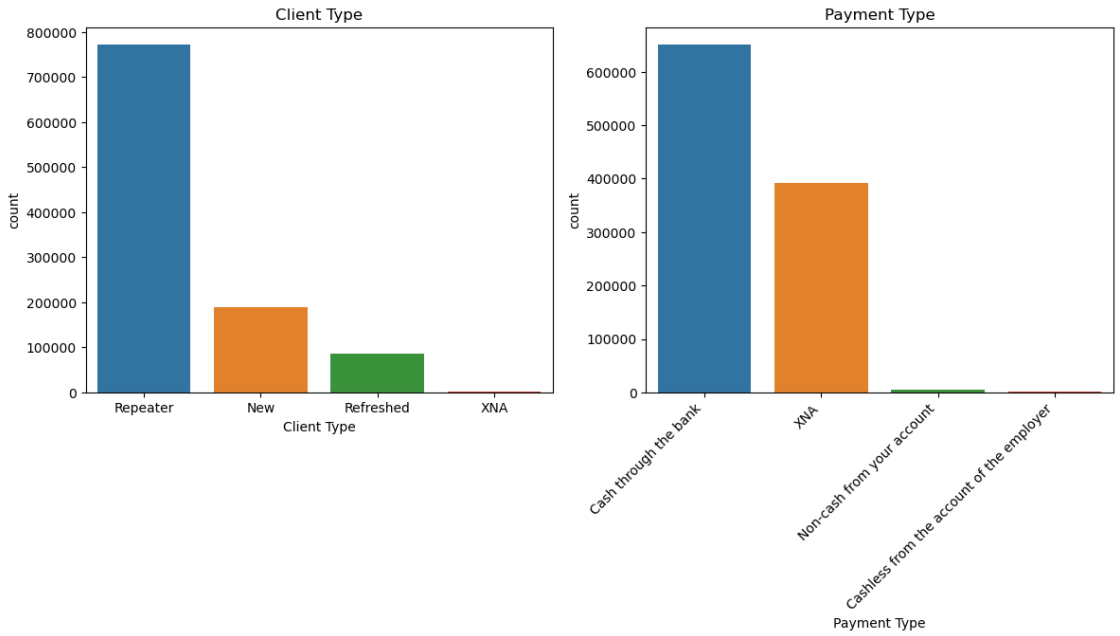


[220]:
```
plt.figure(figsize= (14,14))
total = len(df_prev["NAME_CONTRACT_STATUS"])

def format_func(x):
    return '{:.2f}%\n({:.0f})'.format(x, total*x/100)

plt.figure(figsize = [5, 5])
df_prev["NAME_CONTRACT_STATUS"].value_counts().plot.pie(autopct = format_func)

plt.show()
```

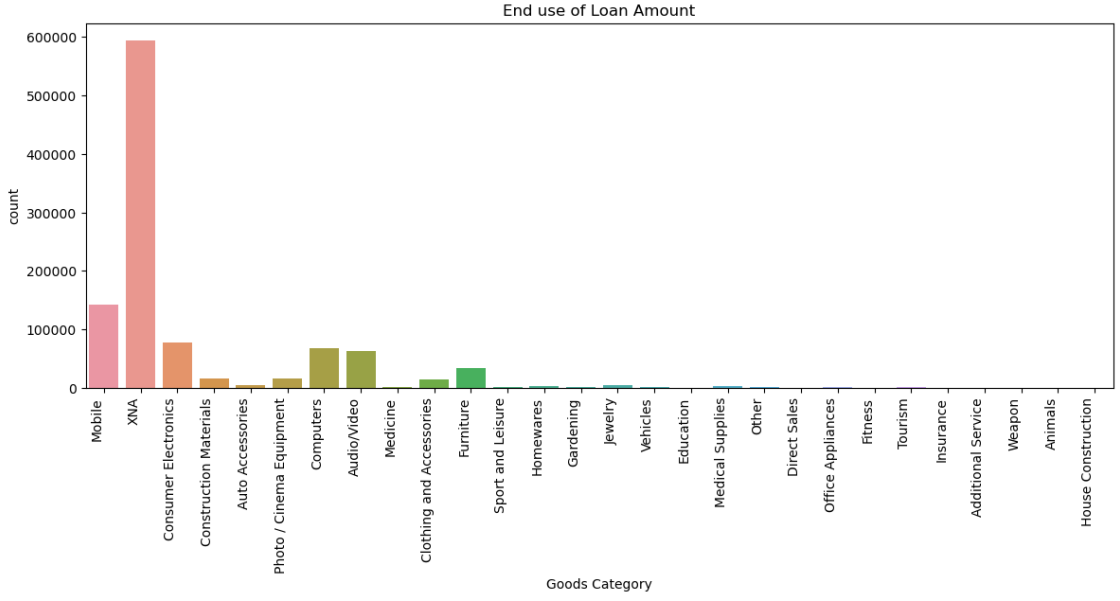
```
<Figure size 1400x1400 with 0 Axes>
```



[222]:
```
plt.figure(figsize=(14,5))

plt.subplot(1,2,1)
ax = sns.countplot(x = 'NAME_CLIENT_TYPE',data=df_prev)
ax.set(title = "Client Type")
ax.set(xlabel='Client Type')

plt.subplot(1,2,2)
ax = sns.countplot(x = 'NAME_PAYMENT_TYPE',data=df_prev)
ax.set(xlabel='Payment Type')
ax.set(title = "Payment Type")
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 45,␣
 ↪horizontalalignment='right')
```

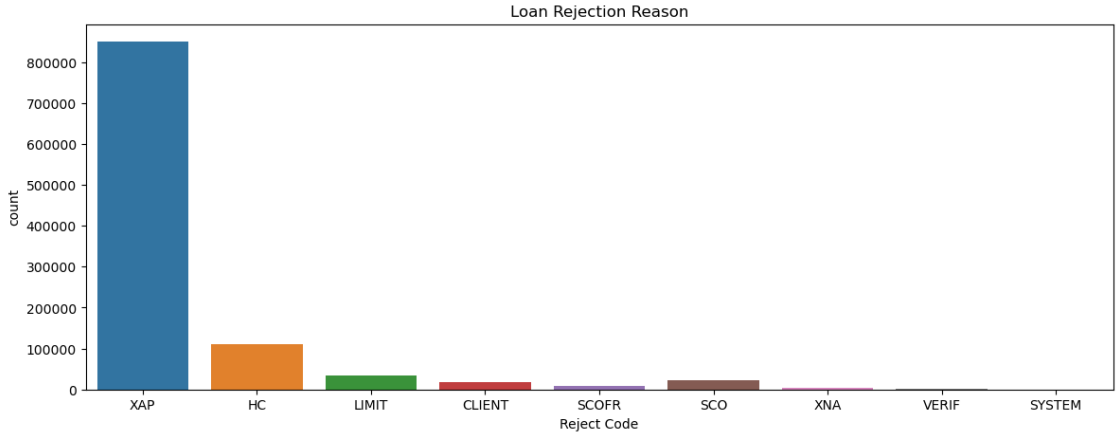


[224]:
```
plt.figure(figsize=(14,5))
ax = sns.countplot(x = 'NAME_GOODS_CATEGORY',data=df_prev)
ax.set(xlabel='Goods Category')
ax.set(title = "End use of Loan Amount")
temp = ax.set_xticklabels(ax.get_xticklabels(), rotation = 90,␣
 ↪horizontalalignment='right')
```

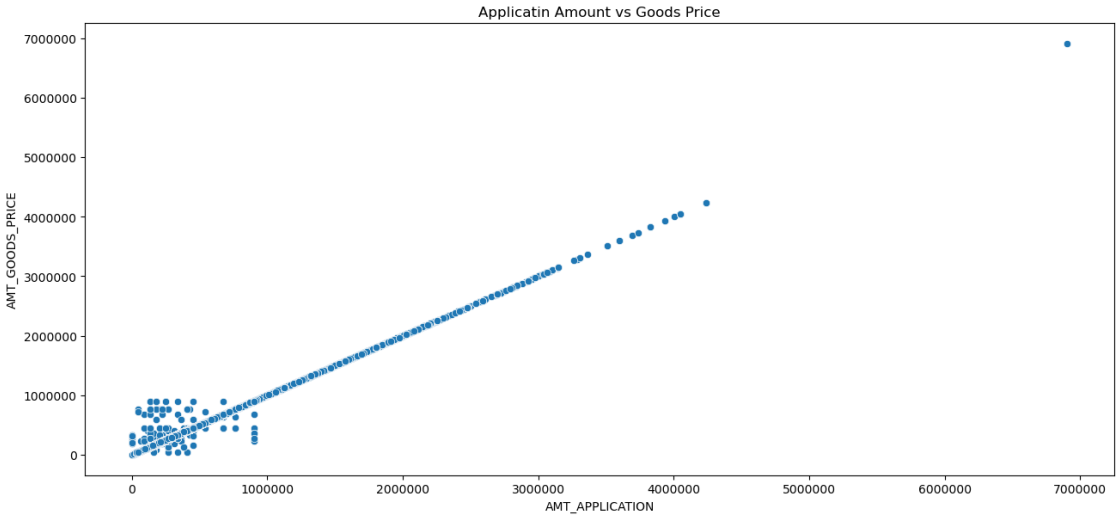


[226]:
```
plt.figure(figsize=(14,5))
ax = sns.countplot(x = 'CODE_REJECT_REASON',data=df_prev)
ax.set(xlabel='Reject Code')
ax.set(title = "Loan Rejection Reason")
```

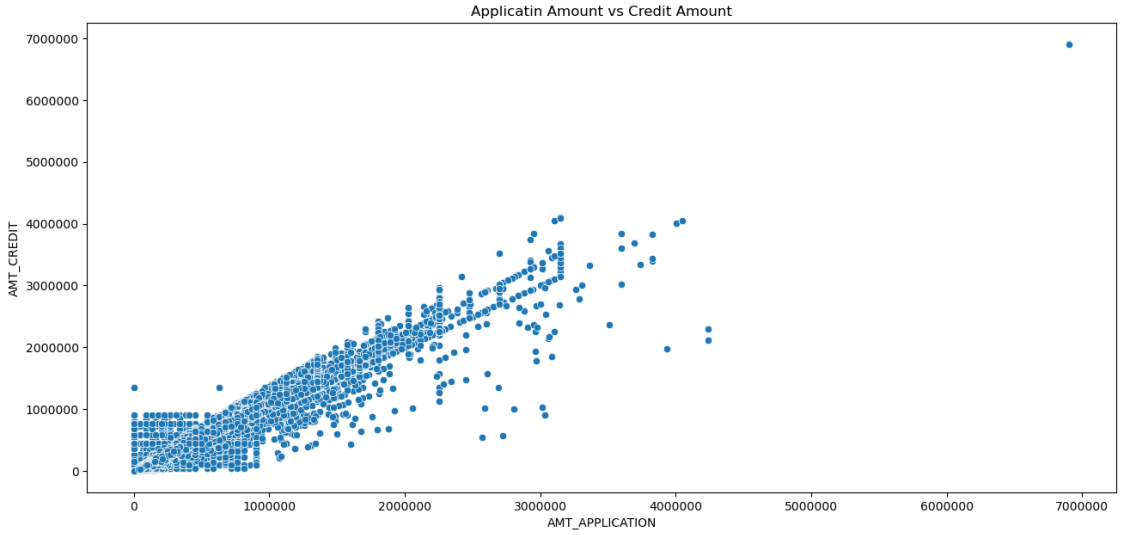
[226]:
```
[Text(0.5, 1.0, 'Loan Rejection Reason')]
```



[228]:
```
plt.figure(figsize = (14,7))
ax = sns.scatterplot(data = df_prev, x ="AMT_APPLICATION", y =␣
↪"AMT_GOODS_PRICE" )
plt.ticklabel_format(style='plain', axis='x')
plt.ticklabel_format(style='plain', axis='y')
ax.set(title = "Applicatin Amount vs Goods Price")
plt.tight_layout(pad = 4)
plt.show()
```



[230]:
```
plt.figure(figsize = (14,7))
ax = sns.scatterplot(data = df_prev, x ="AMT_APPLICATION", y = "AMT_CREDIT" )
plt.ticklabel_format(style='plain', axis='x')
plt.ticklabel_format(style='plain', axis='y')
plt.tight_layout(pad = 4)
ax.set(title = "Applicatin Amount vs Credit Amount")
plt.show()
```



[232]:
```
df_merge = df.merge(df_prev,  right_on='SK_ID_CURR',left_on='SK_ID_CURR',␣
↪how='inner')
```

[234]:
```
df_merge.shape
```

[234]:
```
(715525, 55)
```

[236]:
```
df_merge.head()
```

[236]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE_x	CODE_GENDER	FLAG_OWN_CAR	\
0	100002	1	Cash loans	M	N	
1	100003	0	Cash loans	F	N	
2	100003	0	Cash loans	F	N	
3	100004	0	Revolving loans	M	Y	
4	100006	0	Cash loans	F	N	

	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT_x	\
0	Y	0	202500.0	406597.5	
1	N	0	270000.0	1293502.5	
2	N	0	270000.0	1293502.5	

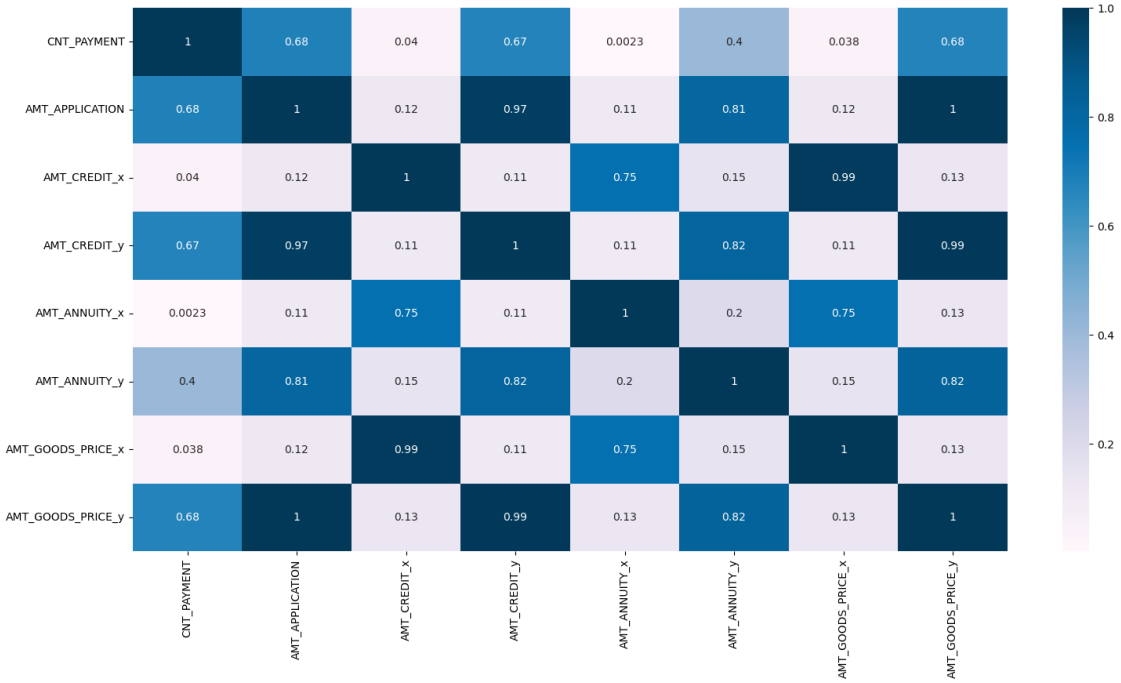
```
3              Y             0           67500.0      135000.0
4              Y             0          135000.0      312682.5

   AMT_ANNUITY_x  AMT_GOODS_PRICE_x NAME_INCOME_TYPE  \
0        24700.5           351000.0          Working
1        35698.5          1129500.0    State servant
2        35698.5          1129500.0    State servant
3         6750.0           135000.0          Working
4        29686.5           297000.0          Working

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

   REGION_POPULATION_RELATIVE  DAYS_REGISTRATION OCCUPATION_TYPE  \
0                    0.018801             3648.0       Laborers
1                    0.003541             1186.0      Core staff
2                    0.003541             1186.0      Core staff
3                    0.010032             4260.0       Laborers
4                    0.008019             9833.0       Laborers

        ORGANIZATION_TYPE  OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0  Business Entity Type 3                       2.0                       2.0
1                  School                       1.0                       0.0
2                  School                       1.0                       0.0
3              Government                       0.0                       0.0
4  Business Entity Type 3                       2.0                       0.0

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                       2.0                       2.0
1                       1.0                       0.0
2                       1.0                       0.0
3                       0.0                       0.0
4                       2.0                       0.0

   AMT_REQ_CREDIT_BUREAU_HOUR  EXT_SOURCE_SCORE        AGE  YEARS_EMPLOYED  \
0                         0.0              0.26  25.920548        1.745205
1                         0.0              0.62  45.931507        3.254795
2                         0.0              0.62  45.931507        3.254795
3                         0.0              0.56  52.180822        0.616438
4                         0.0              0.65  52.068493        8.326027

  AGE_GROUP   INCOME_RANGE   CREDIT_RANGE  SK_ID_PREV NAME_CONTRACT_TYPE_y  \
0     20-30  200000-300000  300000-500000     1038818       Consumer loans
```

```
1     40-50  200000-300000       1000000+     1810518           Cash loans
2     40-50  200000-300000       1000000+     2636178       Consumer loans
3     50-60   50000-100000  100000-200000     1564014       Consumer loans
4     50-60  100000-200000  300000-500000     2078043           Cash loans

   AMT_ANNUITY_y  AMT_APPLICATION  AMT_CREDIT_y  AMT_GOODS_PRICE_y  \
0       9251.775         179055.0      179055.0           179055.0
1      98356.995         900000.0     1035882.0           900000.0
2      64567.665         337500.0      348637.5           337500.0
3       5357.250          24282.0       20106.0            24282.0
4      24246.000         675000.0      675000.0           675000.0

  WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  \
0                   SATURDAY                        9
1                     FRIDAY                       12
2                     SUNDAY                       17
3                     FRIDAY                        5
4                   THURSDAY                       15

  FLAG_LAST_APPL_PER_CONTRACT  NFLAG_LAST_APPL_IN_DAY NAME_CASH_LOAN_PURPOSE  \
0                           Y                       1                    XAP
1                           Y                       1                    XNA
2                           Y                       1                    XAP
3                           Y                       1                    XAP
4                           Y                       1                    XNA

  NAME_CONTRACT_STATUS  DAYS_DECISION      NAME_PAYMENT_TYPE  \
0             Approved            606                    XNA
1             Approved            746                    XNA
2             Approved            828  Cash through the bank
3             Approved            815  Cash through the bank
4             Approved            181  Cash through the bank

  CODE_REJECT_REASON NAME_CLIENT_TYPE NAME_GOODS_CATEGORY NAME_PORTFOLIO  \
0                XAP             New           Vehicles            POS
1                XAP        Repeater                XNA           Cash
2                XAP       Refreshed          Furniture            POS
3                XAP             New             Mobile            POS
4                XAP        Repeater                XNA           Cash

  NAME_PRODUCT_TYPE             CHANNEL_TYPE  SELLERPLACE_AREA  \
0               XNA                   Stone               500
1           x-sell  Credit and cash offices                -1
2               XNA                   Stone              1400
3               XNA         Regional / Local                30
4           x-sell  Credit and cash offices                -1
```

```
  NAME_SELLER_INDUSTRY  CNT_PAYMENT NAME_YIELD_GROUP  \
0      Auto technology         24.0       low_normal
1                  XNA         12.0       low_normal
2            Furniture          6.0           middle
3         Connectivity          4.0           middle
4                  XNA         48.0       low_normal

           PRODUCT_COMBINATION
0      POS other with interest
1              Cash X-Sell: low
2   POS industry with interest
3  POS mobile without interest
4              Cash X-Sell: low
```

[238]:
```
corr_df_merge = df_merge[ ["CNT_PAYMENT",␣
↪"AMT_APPLICATION","AMT_CREDIT_x","AMT_CREDIT_y","AMT_ANNUITY_x",
                           "AMT_ANNUITY_y",␣
↪"AMT_GOODS_PRICE_x","AMT_GOODS_PRICE_y"]   ].corr()
```

[242]:
```
plt.figure(figsize = (18,9))
sns.heatmap(data = corr_df_merge,cmap = "PuBu", annot = True,cbar = True)
plt.show()
```



[ ]: