```
#importing Library

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

#calling file or data

app = pd.read_csv('1application_data.csv')

#top 5 rows

app.head()
```

```
   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]
```

```
app.shape
```

```
(307511, 122)
```

```
#creating data frame for sum of null values in columns , asceending

msng_info =
pd.DataFrame(app.isnull().sum().sort_values()).reset_index()

msng_info
```

```
                            index       0
0                      SK_ID_CURR       0
1         HOUR_APPR_PROCESS_START       0
2      REG_REGION_NOT_WORK_REGION       0
3     LIVE_REGION_NOT_WORK_REGION       0
4          REG_CITY_NOT_LIVE_CITY       0
..                            ...     ...
117     NONLIVINGAPARTMENTS_MEDI  213514
118     NONLIVINGAPARTMENTS_MODE  213514
119               COMMONAREA_MODE  214865
120                COMMONAREA_AVG  214865
121               COMMONAREA_MEDI  214865

[122 rows x 2 columns]
```

```
#renaming columns

msng_info=msng_info.rename(columns={'index': 'col_name',
0:'null_count'})

msng_info
```

```
                         col_name  null_count
0                     SK_ID_CURR           0
1        HOUR_APPR_PROCESS_START           0
2     REG_REGION_NOT_WORK_REGION           0
3    LIVE_REGION_NOT_WORK_REGION           0
4         REG_CITY_NOT_LIVE_CITY           0
..                           ...         ...
117     NONLIVINGAPARTMENTS_MEDI      213514
118     NONLIVINGAPARTMENTS_MODE      213514
119              COMMONAREA_MODE      214865
120               COMMONAREA_AVG      214865
121              COMMONAREA_MEDI      214865

[122 rows x 2 columns]
```

```
#finding null percent count

msng_info['msng_pct']= msng_info['null_count']/app.shape[0]*100

msng_info['msng_pct']
```

```
0       0.000000
1       0.000000
2       0.000000
3       0.000000
4       0.000000
         ...
117    69.432963
118    69.432963
119    69.872297
120    69.872297
121    69.872297
Name: msng_pct, Length: 122, dtype: float64
```

```
#Export dataframe in excel file

msng_info.to_excel('EDA.xlsx')

#Store the names of columns in your dataset that have missing values
exceeding or equal to 40% of the total number of rows.

msng_col=msng_info[msng_info['msng_pct']>=40]['col_name'].to_list()
msng_col
len(msng_col)
```

```
49
```

```
#drop columns who have null values greater or equals to 40%

app_msng_rmvd = app.drop(labels = msng_col,axis = 1)
app_msng_rmvd.shape
```

```
(307511, 73)
```

```
#Data overview and Feature selection

flag_col=[]
for col in app_msng_rmvd.columns:
    if col.startswith("FLAG_"):
        flag_col.append(col)

len(flag_col)
```

```
28
```

```
flag_col
```

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

```
#calling first 5 entries
app_msng_rmvd[flag_col].head()
```

```
  FLAG_OWN_CAR FLAG_OWN_REALTY  FLAG_MOBIL  FLAG_EMP_PHONE
FLAG_WORK_PHONE  \
0            N               Y           1               1
0
```

```
1            N               N           1                1
0
2            Y               Y           1                1
1
3            N               Y           1                1
0
4            N               Y           1                1
0

   FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL  FLAG_DOCUMENT_2
FLAG_DOCUMENT_3  \
0                 1           1           0                0
1
1                 1           1           0                0
1
2                 1           1           0                0
0
3                 1           0           0                0
1
4                 1           0           0                0
0

   ...  FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  FLAG_DOCUMENT_14  \
0  ...                 0                 0                 0
1  ...                 0                 0                 0
2  ...                 0                 0                 0
3  ...                 0                 0                 0
4  ...                 0                 0                 0

   FLAG_DOCUMENT_15  FLAG_DOCUMENT_16  FLAG_DOCUMENT_17
FLAG_DOCUMENT_18  \
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_19  FLAG_DOCUMENT_20  FLAG_DOCUMENT_21
0                 0                 0                0
1                 0                 0                0
2                 0                 0                0
3                 0                 0                0
4                 0                 0                0

[5 rows x 28 columns]
```

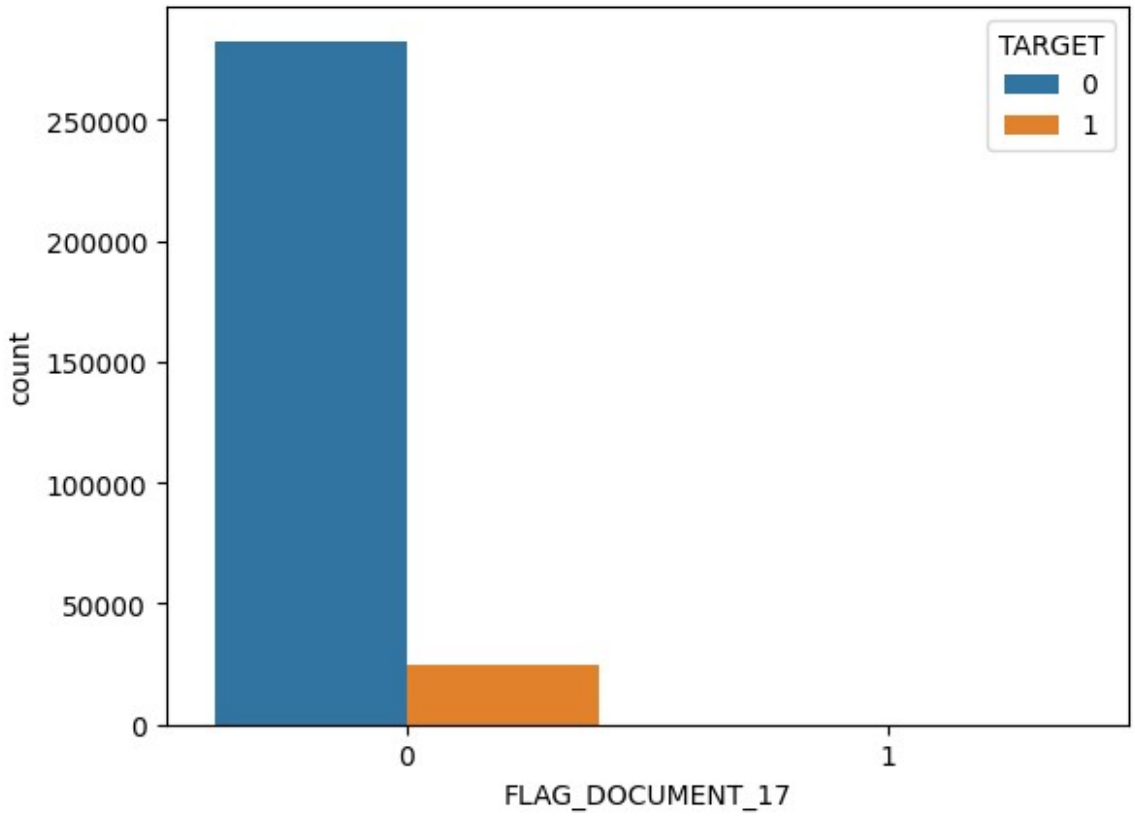
```
#code extracts a subset of columns from the DataFrame and target value
flag_tgt_col = app_msng_rmvd[flag_col+['TARGET']]

flag_tgt_col
```

```
       FLAG_OWN_CAR FLAG_OWN_REALTY  FLAG_MOBIL  FLAG_EMP_PHONE  \
0                 N               Y           1               1
1                 N               N           1               1
2                 Y               Y           1               1
3                 N               Y           1               1
4                 N               Y           1               1
...             ...             ...         ...             ...
307506            N               N           1               1
307507            N               Y           1               0
307508            N               Y           1               1
307509            N               Y           1               1
307510            N               N           1               1

        FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL  \
0                     0                 1           1           0
1                     0                 1           1           0
2                     1                 1           1           0
3                     0                 1           0           0
4                     0                 1           0           0
...                 ...               ...         ...         ...
307506                0                 1           0           0
307507                0                 1           1           0
307508                0                 1           0           1
307509                0                 1           0           0
307510                1                 1           1           0

        FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  ...  FLAG_DOCUMENT_13  \
0                     0                1  ...                 0
1                     0                1  ...                 0
2                     0                0  ...                 0
3                     0                1  ...                 0
4                     0                0  ...                 0
...                 ...              ...  ...               ...
307506                0                0  ...                 0
307507                0                1  ...                 0
307508                0                1  ...                 0
307509                0                1  ...                 0
307510                0                1  ...                 0

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

```
...                    ...                 ...                 ...
307506                   0                   0                   0
307507                   0                   0                   0
307508                   0                   0                   0
307509                   0                   0                   0
307510                   0                   0                   0

        FLAG_DOCUMENT_17  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  \
0                        0                   0                   0
1                        0                   0                   0
2                        0                   0                   0
3                        0                   0                   0
4                        0                   0                   0
...                    ...                 ...                 ...
307506                   0                   0                   0
307507                   0                   0                   0
307508                   0                   0                   0
307509                   0                   0                   0
307510                   0                   0                   0

        FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  TARGET
0                        0                   0       1
1                        0                   0       0
2                        0                   0       0
3                        0                   0       0
4                        0                   0       0
...                    ...                 ...     ...
307506                   0                   0       0
307507                   0                   0       0
307508                   0                   0       0
307509                   0                   0       1
307510                   0                   0       0

[307511 rows x 29 columns]
```

```
# lets find whether there are any patterns or differences in its
distribution based on the target category
sns.countplot(data =flag_tgt_col,x='FLAG_DOCUMENT_17', hue = 'TARGET')
```

```
<Axes: xlabel='FLAG_DOCUMENT_17', ylabel='count'>
```




```
#plotiing chart and explore the relationships between different
features and the target variable in a dataset.
plt.figure(figsize = (20,20))

for i, col in enumerate(flag_col):
    plt.subplot(7,4,i+1)
    sns.countplot(data=flag_tgt_col, x=col,hue='TARGET')
```


```
flg_corr=['FLAG_MOBIL' ,
'FLAG_EMP_PHONE','FLAG_WORK_PHONE','FLAG_CONT_MOBILE','FLAG_PHONE','FL
AG_EMAIL','TARGET']
flag_corr_df = app_msng_rmvd[flg_corr]
flag_corr_df
```

	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE
0	1	1	0	1
1	1	1	0	1
2	1	1	1	1

```
3                1                1                0                1

4                1                1                0                1

...            ...              ...              ...              ...

307506           1                1                0                1

307507           1                0                0                1

307508           1                1                0                1

307509           1                1                0                1

307510           1                1                1                1


        FLAG_PHONE  FLAG_EMAIL  TARGET
0                1           0       1
1                1           0       0
2                1           0       0
3                0           0       0
4                0           0       0
...            ...         ...     ...
307506           0           0       0
307507           1           0       0
307508           0           1       0
307509           0           0       1
307510           1           0       0

[307511 rows x 7 columns]
```

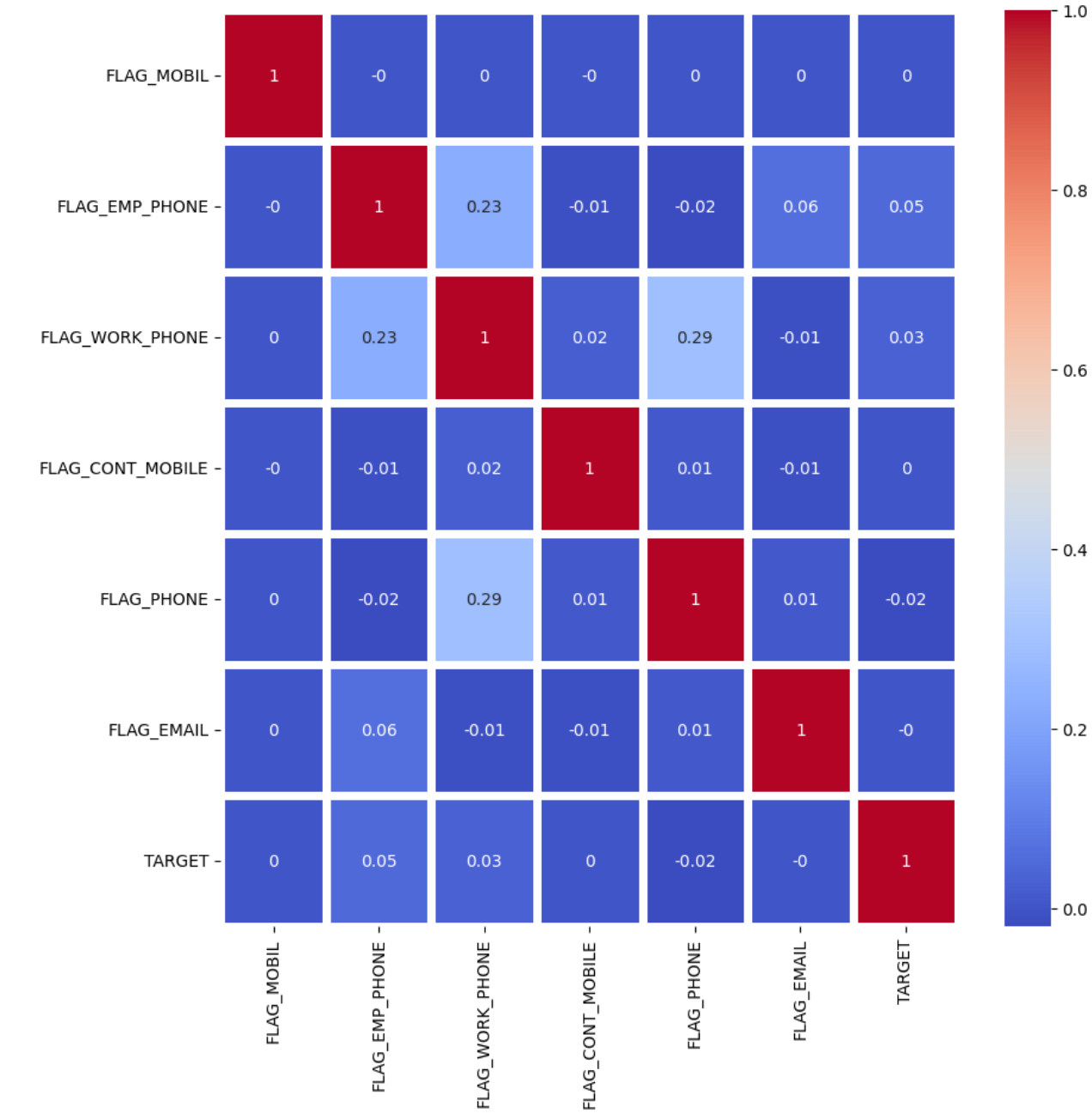
```
#rounding the correlation coefficients to two decimal places
corr_df = round(flag_corr_df.corr(), 2)
corr_df
```

```
                   FLAG_MOBIL  FLAG_EMP_PHONE  FLAG_WORK_PHONE  \
FLAG_MOBIL                1.0           -0.00             0.00
FLAG_EMP_PHONE           -0.0            1.00             0.23
FLAG_WORK_PHONE           0.0            0.23             1.00
FLAG_CONT_MOBILE         -0.0           -0.01             0.02
FLAG_PHONE                0.0           -0.02             0.29
FLAG_EMAIL                0.0            0.06            -0.01
TARGET                    0.0            0.05             0.03

                   FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL  TARGET
FLAG_MOBIL                    -0.00        0.00        0.00    0.00
FLAG_EMP_PHONE                -0.01       -0.02        0.06    0.05
FLAG_WORK_PHONE                0.02        0.29       -0.01    0.03
FLAG_CONT_MOBILE               1.00        0.01       -0.01    0.00
FLAG_PHONE                     0.01        1.00        0.01   -0.02
```

```
FLAG_EMAIL                   -0.01         0.01         1.00    -0.00
TARGET                        0.00        -0.02        -0.00     1.00
```

```
plt.figure(figsize=(10,10))
sns.heatmap(corr_df,cmap = 'coolwarm', linewidths=5, annot = True)
```

```
<Axes: >
```




```
#Dropping flag_col dataframe where fla_col consissts of all columns
that have name starting with flag
```

```
app_flag_rmvd = app_msng_rmvd.drop(labels =flag_col,axis=1)
app_flag_rmvd.head()
```

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

   AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE
NAME_TYPE_SUITE  \
0          202500.0    406597.5      24700.5         351000.0
Unaccompanied
1          270000.0   1293502.5      35698.5        1129500.0
Family
2           67500.0    135000.0       6750.0         135000.0
Unaccompanied
3          135000.0    312682.5      29686.5         297000.0
Unaccompanied
4          121500.0    513000.0      21865.5         513000.0
Unaccompanied

   ... DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE  \
0  ...                      2.0                      2.0
1  ...                      0.0                      1.0
2  ...                      0.0                      0.0
3  ...                      0.0                      2.0
4  ...                      0.0                      0.0

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

   AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
0                        0.0                         0.0
1                        0.0                         0.0
2                        0.0                         0.0
3                        NaN                         NaN
4                        0.0                         0.0
```

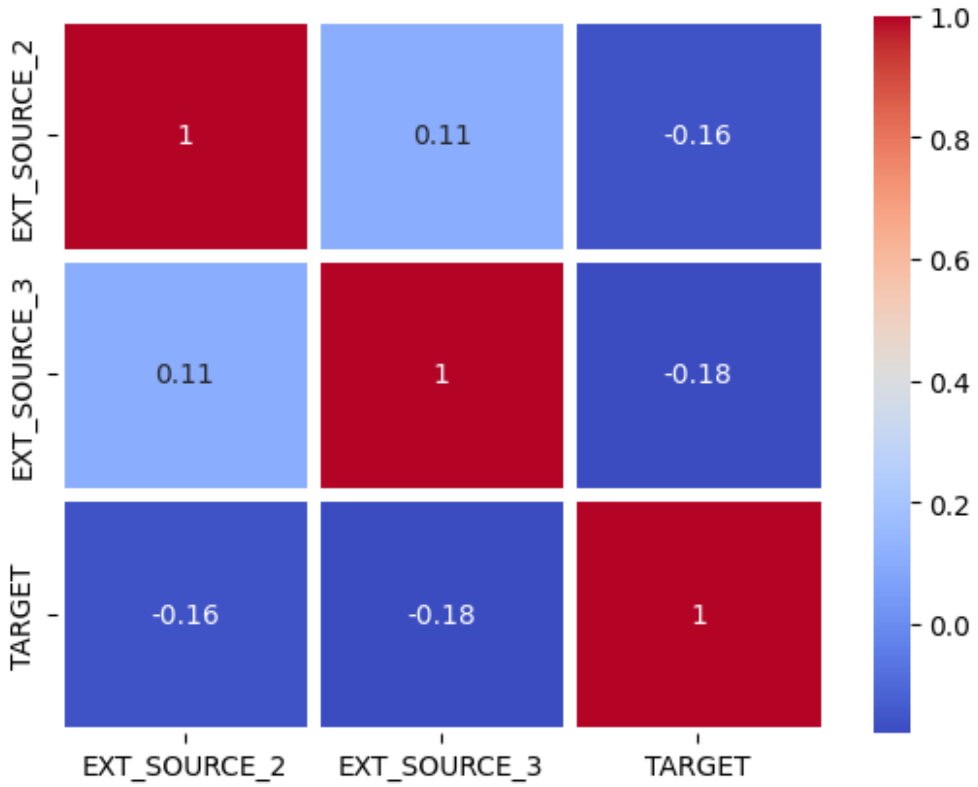
```
   AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                        0.0                        0.0
1                        0.0                        0.0
2                        0.0                        0.0
3                        NaN                        NaN
4                        0.0                        0.0

  AMT_REQ_CREDIT_BUREAU_YEAR
0                        1.0
1                        0.0
2                        0.0
3                        NaN
4                        0.0

[5 rows x 45 columns]
```

```
app_flag_rmvd[[]]
```

```
Empty DataFrame
Columns: []
Index: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34,
35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51,
52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68,
69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85,
86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, ...]

[307511 rows x 0 columns]
```

```
#creates a heatmap to visualize the correlation matrix between
'EXT_SOURCE_2', 'EXT_SOURCE_3', and 'TARGET'
# annot parameter helps with the adding numeric note to each cell
sns.heatmap(round(app_flag_rmvd[['EXT_SOURCE_2','EXT_SOURCE_3','TARGET
']].corr(),2),cmap='coolwarm',linewidth=5,annot= True)
```

```
<Axes: >
```




```
#dropping the columns having correlation less
app_score_col_rmvd =
app_flag_rmvd.drop(['EXT_SOURCE_2','EXT_SOURCE_3'],axis=1)

app_score_col_rmvd.shape
```

```
(307511, 43)
```

```
#to find the high percentage of missing data
app_score_col_rmvd.isnull().sum().sort_values()/app_score_col_rmvd.sha
pe[0]
```

```
SK_ID_CURR                      0.000000
ORGANIZATION_TYPE               0.000000
LIVE_CITY_NOT_WORK_CITY         0.000000
REG_CITY_NOT_WORK_CITY          0.000000
REG_CITY_NOT_LIVE_CITY          0.000000
LIVE_REGION_NOT_WORK_REGION     0.000000
REG_REGION_NOT_WORK_REGION      0.000000
REG_REGION_NOT_LIVE_REGION      0.000000
HOUR_APPR_PROCESS_START         0.000000
WEEKDAY_APPR_PROCESS_START      0.000000
REGION_RATING_CLIENT_W_CITY     0.000000
DAYS_ID_PUBLISH                 0.000000
DAYS_REGISTRATION               0.000000
DAYS_EMPLOYED                   0.000000
```

```
DAYS_BIRTH                    0.000000
REGION_RATING_CLIENT          0.000000
NAME_HOUSING_TYPE             0.000000
TARGET                        0.000000
NAME_CONTRACT_TYPE            0.000000
REGION_POPULATION_RELATIVE    0.000000
CNT_CHILDREN                  0.000000
AMT_INCOME_TOTAL              0.000000
AMT_CREDIT                    0.000000
CODE_GENDER                   0.000000
NAME_INCOME_TYPE              0.000000
NAME_EDUCATION_TYPE           0.000000
NAME_FAMILY_STATUS            0.000000
DAYS_LAST_PHONE_CHANGE        0.000003
CNT_FAM_MEMBERS               0.000007
AMT_ANNUITY                   0.000039
AMT_GOODS_PRICE               0.000904
DEF_60_CNT_SOCIAL_CIRCLE      0.003320
OBS_60_CNT_SOCIAL_CIRCLE      0.003320
DEF_30_CNT_SOCIAL_CIRCLE      0.003320
OBS_30_CNT_SOCIAL_CIRCLE      0.003320
NAME_TYPE_SUITE               0.004201
AMT_REQ_CREDIT_BUREAU_QRT     0.135016
AMT_REQ_CREDIT_BUREAU_HOUR    0.135016
AMT_REQ_CREDIT_BUREAU_DAY     0.135016
AMT_REQ_CREDIT_BUREAU_WEEK    0.135016
AMT_REQ_CREDIT_BUREAU_MON     0.135016
AMT_REQ_CREDIT_BUREAU_YEAR    0.135016
OCCUPATION_TYPE               0.313455
dtype: float64
```

```
#Missing Imputation

#finding mode
app_score_col_rmvd['CNT_FAM_MEMBERS'].mode()
```

```
0    2.0
Name: CNT_FAM_MEMBERS, dtype: float64
```

```
#this line of code is filling null values with the most frequent values(mode)
app_score_col_rmvd['CNT_FAM_MEMBERS']=app_score_col_rmvd['CNT_FAM_MEMBERS'].fillna(app_score_col_rmvd['CNT_FAM_MEMBERS'].mode()[0])

app_score_col_rmvd['CNT_FAM_MEMBERS'].isnull().sum()
```

```
0
```

```
#occupation type column details
app_score_col_rmvd.groupby(['OCCUPATION_TYPE']).size().sort_values()
```

```
OCCUPATION_TYPE
IT staff                   526
HR staff                   563
Realty agents              751
Secretaries               1305
Waiters/barmen staff      1348
Low-skill Laborers        2093
Private service staff     2652
Cleaning staff            4653
Cooking staff             5946
Security staff            6721
Medicine staff            8537
Accountants               9813
High skill tech staff    11380
Drivers                  18603
Managers                 21371
Core staff               27570
Sales staff              32102
Laborers                 55186
dtype: int64
```

```
#most frquest values in occupation type column
df =app_score_col_rmvd['OCCUPATION_TYPE'].mode()

#fillin na values in occupation type column with mode values
app_score_col_rmvd['OCCUPATION_TYPE']=app_score_col_rmvd['OCCUPATION_T
YPE'].fillna(df.mode()[0])

#finding null values in occupation type column
app_score_col_rmvd['OCCUPATION_TYPE'].isnull().sum()
```

```
0
```

```
#finding null values in occupation type column
app_score_col_rmvd['NAME_TYPE_SUITE'].isnull().sum()
```

```
1292
```

```
df1=app_score_col_rmvd['NAME_TYPE_SUITE'].mode()

#fillin na values in NAME_TYPE_SUITE column with mode values
app_score_col_rmvd['NAME_TYPE_SUITE'] =
app_score_col_rmvd['NAME_TYPE_SUITE'].fillna(df1.mode()[0])

app_score_col_rmvd['NAME_TYPE_SUITE'].isnull().sum()
```

```
0
```

```
#details in AMT_ANNUITY column
app_score_col_rmvd['AMT_ANNUITY'].describe()
```

```
count    307499.000000
mean      27108.573909
std       14493.737315
min        1615.500000
25%       16524.000000
50%       24903.000000
75%       34596.000000
max      258025.500000
Name: AMT_ANNUITY, dtype: float64
```

```
df3 = app_score_col_rmvd['AMT_ANNUITY'].mean()

#filling null vales in AMT_ANNUITY column with mean values
app_score_col_rmvd['AMT_ANNUITY'] =
app_score_col_rmvd['AMT_ANNUITY'].fillna(df3.mean())

app_score_col_rmvd['AMT_ANNUITY'].isnull().sum()
```

```
0
```

```
#sum of  null values in DEF_60_CNT_SOCIAL_CIRCLE columns
app_score_col_rmvd['DEF_60_CNT_SOCIAL_CIRCLE'].isnull().sum()
```

```
1021
```

```
app_score_col_rmvd['AMT_REQ_CREDIT_BUREAU_HOUR'].describe()
```

```
count    265992.000000
mean          0.006402
std           0.083849
min           0.000000
25%           0.000000
50%           0.000000
75%           0.000000
max           4.000000
Name: AMT_REQ_CREDIT_BUREAU_HOUR, dtype: float64
```

```
#creating a data frame having all columns starts with
AMT_REQ_CREDIT_BUREAU
amt_req_col = []

for col in app_score_col_rmvd.columns:
    if col.startswith("AMT_REQ_CREDIT_BUREAU"):
        amt_req_col.append(col)

#all columns name starts with AMT_REQ_CREDIT_BUREAU
amt_req_col
```

```
['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']

#filling null values in the all column starts with
AMT_REQ_CREDIT_BUREAU with median values
for col in amt_req_col:
    app_score_col_rmvd[col] =
app_score_col_rmvd[col].fillna((app_score_col_rmvd[col].median()))

app_score_col_rmvd[col].isnull().sum()

0

app_score_col_rmvd['AMT_GOODS_PRICE'].isnull().sum()

278

app_score_col_rmvd['AMT_GOODS_PRICE'].describe()

count    3.072330e+05
mean     5.383962e+05
std      3.694465e+05
min      4.050000e+04
25%      2.385000e+05
50%      4.500000e+05
75%      6.795000e+05
max      4.050000e+06
Name: AMT_GOODS_PRICE, dtype: float64

app_score_col_rmvd['AMT_GOODS_PRICE'].agg(['min','max','median'])

min         40500.0
max       4050000.0
median     450000.0
Name: AMT_GOODS_PRICE, dtype: float64

app_score_col_rmvd['AMT_GOODS_PRICE'].mean()

538396.2074288895

app_score_col_rmvd['AMT_GOODS_PRICE'] =
app_score_col_rmvd['AMT_GOODS_PRICE'].fillna((app_score_col_rmvd['AMT_
GOODS_PRICE'].median()))

app_score_col_rmvd['AMT_GOODS_PRICE'].isnull().sum()

0

#value modification
#all columns starts with DAYS in a one data frame
days_col=[]

for col in app_score_col_rmvd.columns:
```

```
    if col.startswith("DAYS"):
        days_col.append(col)

days_col
```

```
['DAYS_BIRTH',
 'DAYS_EMPLOYED',
 'DAYS_REGISTRATION',
 'DAYS_ID_PUBLISH',
 'DAYS_LAST_PHONE_CHANGE']
```

```
# loop iterates through each column specified in the days_col list and
replaces the values in each column with their absolute values.
for col in days_col:
    app_score_col_rmvd[col]= abs( app_score_col_rmvd[col])


app_score_col_rmvd['DAYS_BIRTH']
```

```
0          9461
1         16765
2         19046
3         19005
4         19932
          ...
307506     9327
307507    20775
307508    14966
307509    11961
307510    16856
Name: DAYS_BIRTH, Length: 307511, dtype: int64
```

```
app_score_col_rmvd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 43 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   CNT_CHILDREN                 307511 non-null  int64
 5   AMT_INCOME_TOTAL             307511 non-null  float64
 6   AMT_CREDIT                   307511 non-null  float64
 7   AMT_ANNUITY                  307511 non-null  float64
 8   AMT_GOODS_PRICE              307511 non-null  float64
 9   NAME_TYPE_SUITE              307511 non-null  object
 10  NAME_INCOME_TYPE             307511 non-null  object
 11  NAME_EDUCATION_TYPE          307511 non-null  object
```

```
 12  NAME_FAMILY_STATUS            307511 non-null  object
 13  NAME_HOUSING_TYPE             307511 non-null  object
 14  REGION_POPULATION_RELATIVE    307511 non-null  float64
 15  DAYS_BIRTH                    307511 non-null  int64
 16  DAYS_EMPLOYED                 307511 non-null  int64
 17  DAYS_REGISTRATION             307511 non-null  float64
 18  DAYS_ID_PUBLISH               307511 non-null  int64
 19  OCCUPATION_TYPE               307511 non-null  object
 20  CNT_FAM_MEMBERS               307511 non-null  float64
 21  REGION_RATING_CLIENT          307511 non-null  int64
 22  REGION_RATING_CLIENT_W_CITY   307511 non-null  int64
 23  WEEKDAY_APPR_PROCESS_START    307511 non-null  object
 24  HOUR_APPR_PROCESS_START       307511 non-null  int64
 25  REG_REGION_NOT_LIVE_REGION    307511 non-null  int64
 26  REG_REGION_NOT_WORK_REGION    307511 non-null  int64
 27  LIVE_REGION_NOT_WORK_REGION   307511 non-null  int64
 28  REG_CITY_NOT_LIVE_CITY        307511 non-null  int64
 29  REG_CITY_NOT_WORK_CITY        307511 non-null  int64
 30  LIVE_CITY_NOT_WORK_CITY       307511 non-null  int64
 31  ORGANIZATION_TYPE             307511 non-null  object
 32  OBS_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 33  DEF_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 34  OBS_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 35  DEF_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 36  DAYS_LAST_PHONE_CHANGE        307510 non-null  float64
 37  AMT_REQ_CREDIT_BUREAU_HOUR    307511 non-null  float64
 38  AMT_REQ_CREDIT_BUREAU_DAY     307511 non-null  float64
 39  AMT_REQ_CREDIT_BUREAU_WEEK    307511 non-null  float64
 40  AMT_REQ_CREDIT_BUREAU_MON     307511 non-null  float64
 41  AMT_REQ_CREDIT_BUREAU_QRT     307511 non-null  float64
 42  AMT_REQ_CREDIT_BUREAU_YEAR    307511 non-null  float64
dtypes: float64(18), int64(15), object(10)
memory usage: 100.9+ MB
```

```
#fnding no. of unique values in every column
app_score_col_rmvd.nunique().sort_values()
```

```
LIVE_REGION_NOT_WORK_REGION         2
TARGET                              2
NAME_CONTRACT_TYPE                  2
REG_REGION_NOT_LIVE_REGION          2
REG_CITY_NOT_LIVE_CITY              2
REG_CITY_NOT_WORK_CITY              2
LIVE_CITY_NOT_WORK_CITY             2
REG_REGION_NOT_WORK_REGION          2
REGION_RATING_CLIENT_W_CITY         3
REGION_RATING_CLIENT                3
CODE_GENDER                         3
NAME_EDUCATION_TYPE                 5
AMT_REQ_CREDIT_BUREAU_HOUR          5
```

```
NAME_HOUSING_TYPE                     6
NAME_FAMILY_STATUS                    6
WEEKDAY_APPR_PROCESS_START            7
NAME_TYPE_SUITE                       7
NAME_INCOME_TYPE                      8
AMT_REQ_CREDIT_BUREAU_DAY             9
DEF_60_CNT_SOCIAL_CIRCLE              9
AMT_REQ_CREDIT_BUREAU_WEEK            9
DEF_30_CNT_SOCIAL_CIRCLE             10
AMT_REQ_CREDIT_BUREAU_QRT            11
CNT_CHILDREN                         15
CNT_FAM_MEMBERS                      17
OCCUPATION_TYPE                      18
HOUR_APPR_PROCESS_START              24
AMT_REQ_CREDIT_BUREAU_MON            24
AMT_REQ_CREDIT_BUREAU_YEAR           25
OBS_30_CNT_SOCIAL_CIRCLE             33
OBS_60_CNT_SOCIAL_CIRCLE             33
ORGANIZATION_TYPE                    58
REGION_POPULATION_RELATIVE           81
AMT_GOODS_PRICE                    1002
AMT_INCOME_TOTAL                   2548
DAYS_LAST_PHONE_CHANGE             3773
AMT_CREDIT                         5603
DAYS_ID_PUBLISH                    6168
DAYS_EMPLOYED                     12574
AMT_ANNUITY                       13673
DAYS_REGISTRATION                 15688
DAYS_BIRTH                        17460
SK_ID_CURR                       307511
dtype: int64
```

```
#unique no. in AMT_GOODS_PRICE column
app_score_col_rmvd['AMT_GOODS_PRICE'].unique()
```

```
array([ 351000. , 1129500. ,  135000. , ...,  453465. ,  143977.5,
        743863.5])
```

```
app_score_col_rmvd['AMT_GOODS_PRICE'].describe()
```

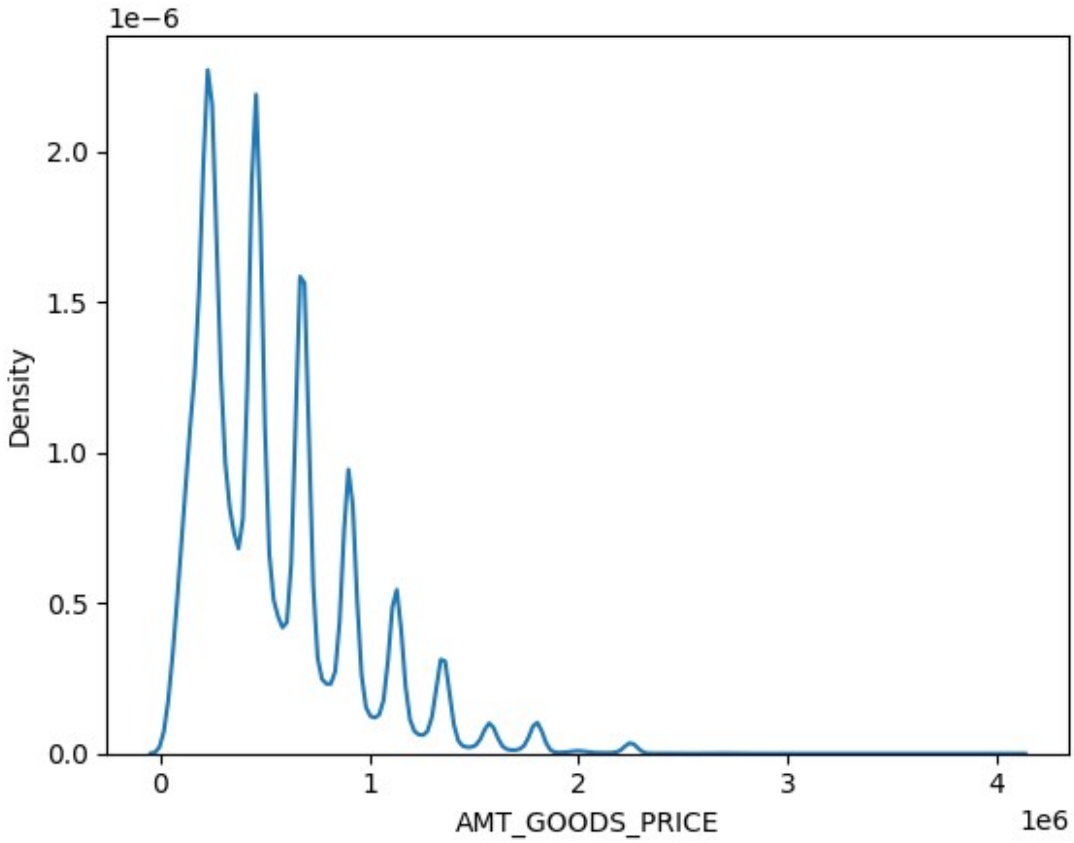
```
count    3.075110e+05
mean     5.383163e+05
std      3.692890e+05
min      4.050000e+04
25%      2.385000e+05
50%      4.500000e+05
75%      6.795000e+05
max      4.050000e+06
Name: AMT_GOODS_PRICE, dtype: float64
```

```
#outlier detectionn and treatement
app_score_col_rmvd['AMT_GOODS_PRICE'].agg(['min','max','median'])
```

```
min         40500.0
max       4050000.0
median     450000.0
Name: AMT_GOODS_PRICE, dtype: float64
```

```
sns.kdeplot(data =app_score_col_rmvd,x = 'AMT_GOODS_PRICE')
```

```
<Axes: xlabel='AMT_GOODS_PRICE', ylabel='Density'>
```




```
#binning the variables and creating bins
app_score_col_rmvd['AMT_GOODS_PRICE'].quantile([0.1,0.2,0.3,0.4,0.5,0.
6,0.7,0.8,0.9])
```

```
0.1     180000.0
0.2     225000.0
0.3     270000.0
0.4     378000.0
0.5     450000.0
0.6     522000.0
```

```
0.7     675000.0
0.8     814500.0
0.9    1093500.0
Name: AMT_GOODS_PRICE, dtype: float64
```

```
bins=[0,10000,20000,30000,40000,50000,60000,70000,80000,90000,4050000]

ranges = ['0-100k','100k-200k','200k-300k','300k-400k','400k-
500k','500k-600k','600k-700k','700k-800k','800k-900k','Above 900k']

app_score_col_rmvd['AMT_GOODS_PRICE_RANGE'] =
pd.cut(app_score_col_rmvd['AMT_GOODS_PRICE'],bins,labels=ranges)
app_score_col_rmvd.groupby(['AMT_GOODS_PRICE_RANGE']).size()
```

```
AMT_GOODS_PRICE_RANGE
0-100k             0
100k-200k          0
200k-300k          0
300k-400k          0
400k-500k       1327
500k-600k        616
600k-700k       1624
700k-800k        542
800k-900k       3693
Above 900k    299709
dtype: int64
```

```
app_score_col_rmvd['AMT_INCOME_TOTAL'].quantile([0.1,0.2,0.3,0.4,0.5,0
.6,0.7,0.8,0.9,0.99])
```

```
0.10     81000.0
0.20     99000.0
0.30    112500.0
0.40    135000.0
0.50    147150.0
0.60    162000.0
0.70    180000.0
0.80    225000.0
0.90    270000.0
0.99    472500.0
Name: AMT_INCOME_TOTAL, dtype: float64
```

```
app_score_col_rmvd['AMT_INCOME_TOTAL'].agg(['min','max','median'])
```

```
min           25650.0
max       117000000.0
median       147150.0
Name: AMT_INCOME_TOTAL, dtype: float64
```

```
bins = [0,10000,150000,200000,250000,300000,350000,400000,472500]
```

```
range = ['0-100k','100k-150k','150k-200k','200k-250k','250k-
300k','300k-350k','350k-400k','Above 400k']

app_score_col_rmvd['AMT_INCOME_TOTAL_RANGE'] =
pd.cut(app_score_col_rmvd['AMT_INCOME_TOTAL'],bins,labels = range)

app_score_col_rmvd.groupby(['AMT_INCOME_TOTAL_RANGE']).size()
```

```
AMT_INCOME_TOTAL_RANGE
0-100k             0
100k-150k     155289
150k-200k      64307
200k-250k      48137
250k-300k      17039
300k-350k       8874
350k-400k       5802
Above 400k      5049
dtype: int64
```

```
app_score_col_rmvd['AMT_CREDIT'].quantile([0.1,0.2,0.3,0.4,0.5,0.6,0.7
,0.8,0.9,0.99])
```

```
0.10     180000.0
0.20     254700.0
0.30     306306.0
0.40     432000.0
0.50     513531.0
0.60     604152.0
0.70     755190.0
0.80     900000.0
0.90    1133748.0
0.99    1854000.0
Name: AMT_CREDIT, dtype: float64
```

```
bins = [0,200000,400000,600000,800000,1000000,1854000]

ranges = ['0-200k','200k-400k','400k-600k','600k-800k','800k-
1M','Above 1M']
app_score_col_rmvd['AMT_CREDIT_RANGE'] =
pd.cut(app_score_col_rmvd['AMT_CREDIT'],bins,labels=ranges)

app_score_col_rmvd.groupby(['AMT_CREDIT_RANGE']).size()
```

```
AMT_CREDIT_RANGE
0-200k       36144
200k-400k    81151
400k-600k    66270
600k-800k    43242
800k-1M      30719
```

```
Above 1M     46910
dtype: int64
```

```
app_score_col_rmvd['AMT_ANNUITY'].quantile([0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.99])
```

```
0.10    11074.5
0.20    14701.5
0.30    18189.0
0.40    21870.0
0.50    24903.0
0.60    28062.0
0.70    32004.0
0.80    37516.5
0.90    45954.0
0.99    70006.5
Name: AMT_ANNUITY, dtype: float64
```

```
app_score_col_rmvd['AMT_ANNUITY'].max()
```

```
258025.5
```

```
bins = [0,50000,100000,150000,200000,258025.5]

ranges = ['0-50k','50k-100k','100k-150k','150k-200k','Above 200k']

app_score_col_rmvd['AMT_ANNUITY_RANGE'] = pd.cut(app_score_col_rmvd['AMT_ANNUITY'],bins,labels= ranges)

app_score_col_rmvd.groupby(['AMT_ANNUITY_RANGE']).size()
```

```
AMT_ANNUITY_RANGE
0-50k         286214
50k-100k       20792
100k-150k        437
150k-200k         32
Above 200k        36
dtype: int64
```

```
app_score_col_rmvd['AMT_ANNUITY_RANGE'].isnull().sum()
```

```
0
```

```
app_score_col_rmvd['DAYS_EMPLOYED'].agg(['min','max','median'])
```

```
min            0.0
max       365243.0
median      2219.0
Name: DAYS_EMPLOYED, dtype: float64
```

```
app_score_col_rmvd['DAYS_EMPLOYED'].quantile([0.1,0.2,0.3,0.4,0.5,.6,0.7,0.8,0.85,0.90,0.95,0.99])
```

```
0.10       392.0
0.20       749.0
0.30      1132.0
0.40      1597.0
0.50      2219.0
0.60      3032.0
0.70      4435.0
0.80      9188.0
0.85    365243.0
0.90    365243.0
0.95    365243.0
0.99    365243.0
Name: DAYS_EMPLOYED, dtype: float64
```

```
app_score_col_rmvd[app_score_col_rmvd['DAYS_EMPLOYED']<app_score_col_r
mvd['DAYS_EMPLOYED'].max()].max()['DAYS_EMPLOYED']
```

```
17912
```

```
app_score_col_rmvd['DAYS_EMPLOYED'].max()
```

```
365243
```

```
bins =
[0,1825,3650,5475,7300,9125,10950,12775,14600,16425,18250,23691,365243
]

ranges = ['0-5Y','5Y-10Y','10Y-15Y','15Y-20Y','20Y-25Y','25Y-
30Y','30Y-35Y','35Y-40Y','40Y-45Y','45Y-50Y','50Y-65Y','Above 65Y']
app_score_col_rmvd['DAYS_EMPLOYED_RANGE'] =
pd.cut(app_score_col_rmvd['DAYS_EMPLOYED'],bins,labels=ranges)
```

```
app_score_col_rmvd['DAYS_EMPLOYED_RANGE'].isnull().sum()
```

```
2
```

```
app_score_col_rmvd.groupby(['DAYS_EMPLOYED_RANGE']).size()
```

```
DAYS_EMPLOYED_RANGE
0-5Y         136309
5Y-10Y        64872
10Y-15Y       27549
15Y-20Y       10849
20Y-25Y        6243
25Y-30Y        3308
30Y-35Y        1939
35Y-40Y         832
40Y-45Y         210
45Y-50Y          24
50Y-65Y           0
```

```
Above 65Y     55374
dtype: int64
```

```
app_score_col_rmvd['DAYS_BIRTH'].agg(['min','max','median'])
```

```
min        7489.0
max       25229.0
median    15750.0
Name: DAYS_BIRTH, dtype: float64
```

```
app_score_col_rmvd['DAYS_BIRTH'].isnull().sum()
```

```
0
```

```
bins = [0,7300,10950,14600,18250,21900,25229]
ranges = ['20Y','20Y-30Y','30Y-40Y','40Y-50Y','50Y-60Y','Above 60Y']
app_score_col_rmvd['DAYS_BIRTH_RANGE'] =
pd.cut(app_score_col_rmvd['DAYS_BIRTH'],bins,labels=ranges)
```

```
app_score_col_rmvd.groupby(['DAYS_BIRTH_RANGE']).size()
```

```
DAYS_BIRTH_RANGE
20Y              0
20Y-30Y      45021
30Y-40Y      82308
40Y-50Y      76541
50Y-60Y      68062
Above 60Y    35579
dtype: int64
```

```
app_score_col_rmvd['DAYS_BIRTH_RANGE'].isnull().sum()
```

```
0
```

```
app_score_col_rmvd.info()
```

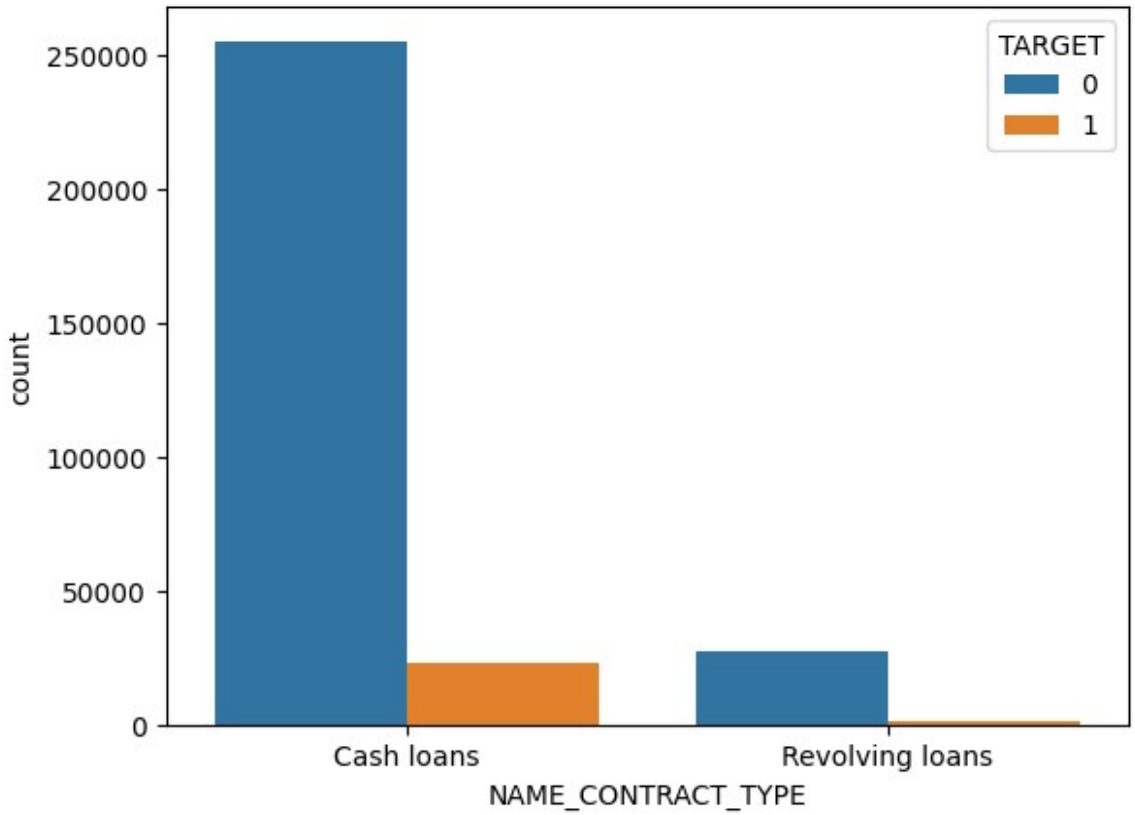
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 49 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   SK_ID_CURR                   307511 non-null  int64
 1   TARGET                       307511 non-null  int64
 2   NAME_CONTRACT_TYPE           307511 non-null  object
 3   CODE_GENDER                  307511 non-null  object
 4   CNT_CHILDREN                 307511 non-null  int64
 5   AMT_INCOME_TOTAL             307511 non-null  float64
 6   AMT_CREDIT                   307511 non-null  float64
 7   AMT_ANNUITY                  307511 non-null  float64
 8   AMT_GOODS_PRICE              307511 non-null  float64
 9   NAME_TYPE_SUITE              307511 non-null  object
 10  NAME_INCOME_TYPE             307511 non-null  object
```

```
 11  NAME_EDUCATION_TYPE           307511 non-null  object
 12  NAME_FAMILY_STATUS            307511 non-null  object
 13  NAME_HOUSING_TYPE             307511 non-null  object
 14  REGION_POPULATION_RELATIVE    307511 non-null  float64
 15  DAYS_BIRTH                    307511 non-null  int64
 16  DAYS_EMPLOYED                 307511 non-null  int64
 17  DAYS_REGISTRATION             307511 non-null  float64
 18  DAYS_ID_PUBLISH               307511 non-null  int64
 19  OCCUPATION_TYPE               307511 non-null  object
 20  CNT_FAM_MEMBERS               307511 non-null  float64
 21  REGION_RATING_CLIENT          307511 non-null  int64
 22  REGION_RATING_CLIENT_W_CITY   307511 non-null  int64
 23  WEEKDAY_APPR_PROCESS_START    307511 non-null  object
 24  HOUR_APPR_PROCESS_START       307511 non-null  int64
 25  REG_REGION_NOT_LIVE_REGION    307511 non-null  int64
 26  REG_REGION_NOT_WORK_REGION    307511 non-null  int64
 27  LIVE_REGION_NOT_WORK_REGION   307511 non-null  int64
 28  REG_CITY_NOT_LIVE_CITY        307511 non-null  int64
 29  REG_CITY_NOT_WORK_CITY        307511 non-null  int64
 30  LIVE_CITY_NOT_WORK_CITY       307511 non-null  int64
 31  ORGANIZATION_TYPE             307511 non-null  object
 32  OBS_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 33  DEF_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 34  OBS_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 35  DEF_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 36  DAYS_LAST_PHONE_CHANGE        307510 non-null  float64
 37  AMT_REQ_CREDIT_BUREAU_HOUR    307511 non-null  float64
 38  AMT_REQ_CREDIT_BUREAU_DAY     307511 non-null  float64
 39  AMT_REQ_CREDIT_BUREAU_WEEK    307511 non-null  float64
 40  AMT_REQ_CREDIT_BUREAU_MON     307511 non-null  float64
 41  AMT_REQ_CREDIT_BUREAU_QRT     307511 non-null  float64
 42  AMT_REQ_CREDIT_BUREAU_YEAR    307511 non-null  float64
 43  AMT_GOODS_PRICE_RANGE         307511 non-null  category
 44  AMT_INCOME_TOTAL_RANGE        304497 non-null  category
 45  AMT_CREDIT_RANGE              304436 non-null  category
 46  AMT_ANNUITY_RANGE             307511 non-null  category
 47  DAYS_EMPLOYED_RANGE           307509 non-null  category
 48  DAYS_BIRTH_RANGE              307511 non-null  category
dtypes: category(6), float64(18), int64(15), object(10)
memory usage: 102.6+ MB
```

```
obj_var = app_score_col_rmvd.select_dtypes(include =
['object']).columns

obj_var
```

```
Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE',
```

```
'WEEKDAY_APPR_PROCESS_START',
       'ORGANIZATION_TYPE'],
      dtype='object')

app_score_col_rmvd.groupby(['NAME_CONTRACT_TYPE']).size()

NAME_CONTRACT_TYPE
Cash loans         278232
Revolving loans     29279
dtype: int64

sns.countplot(data=app_score_col_rmvd ,
x='NAME_CONTRACT_TYPE',hue='TARGET')

<Axes: xlabel='NAME_CONTRACT_TYPE', ylabel='count'>
```



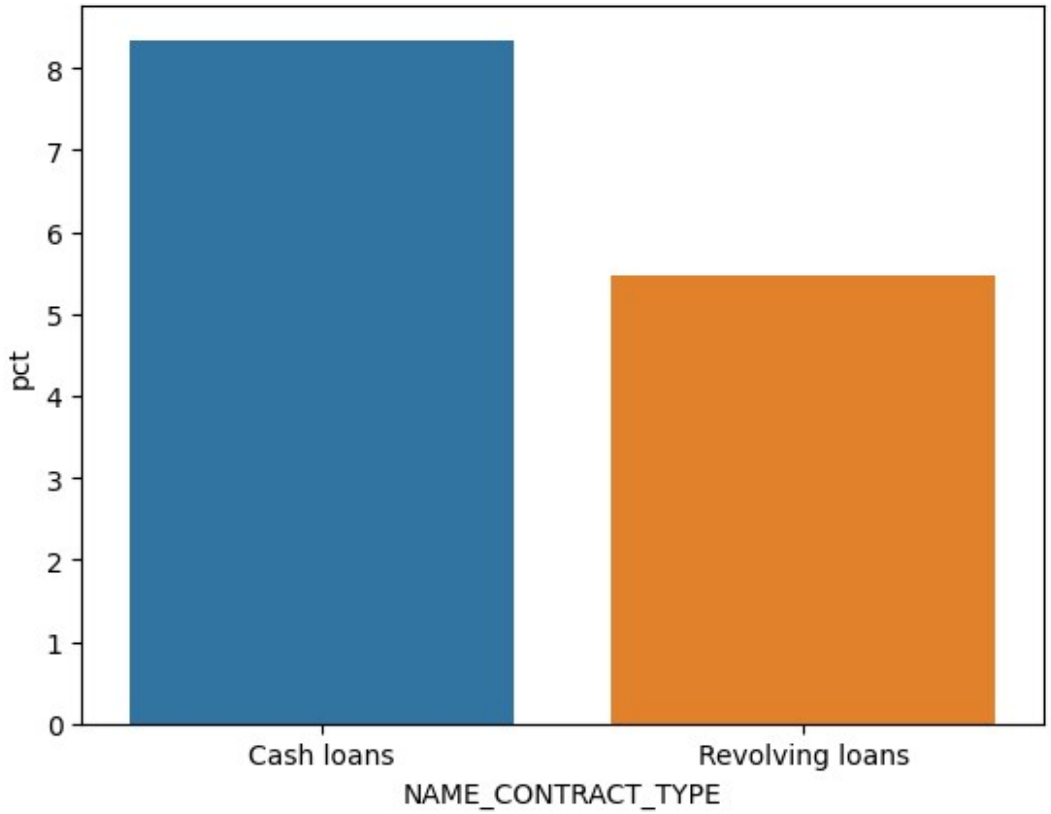

```
data_pct =
app_score_col_rmvd[['NAME_CONTRACT_TYPE','TARGET']].groupby(['NAME_CON
TRACT_TYPE'], as_index = False).mean()
data_pct

  NAME_CONTRACT_TYPE    TARGET
0         Cash loans  0.083459
1    Revolving loans  0.054783
```

```
data_pct['pct']= data_pct['TARGET']*100

data_pct['pct']
```

```
0    8.345913
1    5.478329
Name: pct, dtype: float64
```

```
sns.barplot(data=data_pct,x='NAME_CONTRACT_TYPE',y='pct')
```

```
<Axes: xlabel='NAME_CONTRACT_TYPE', ylabel='pct'>
```



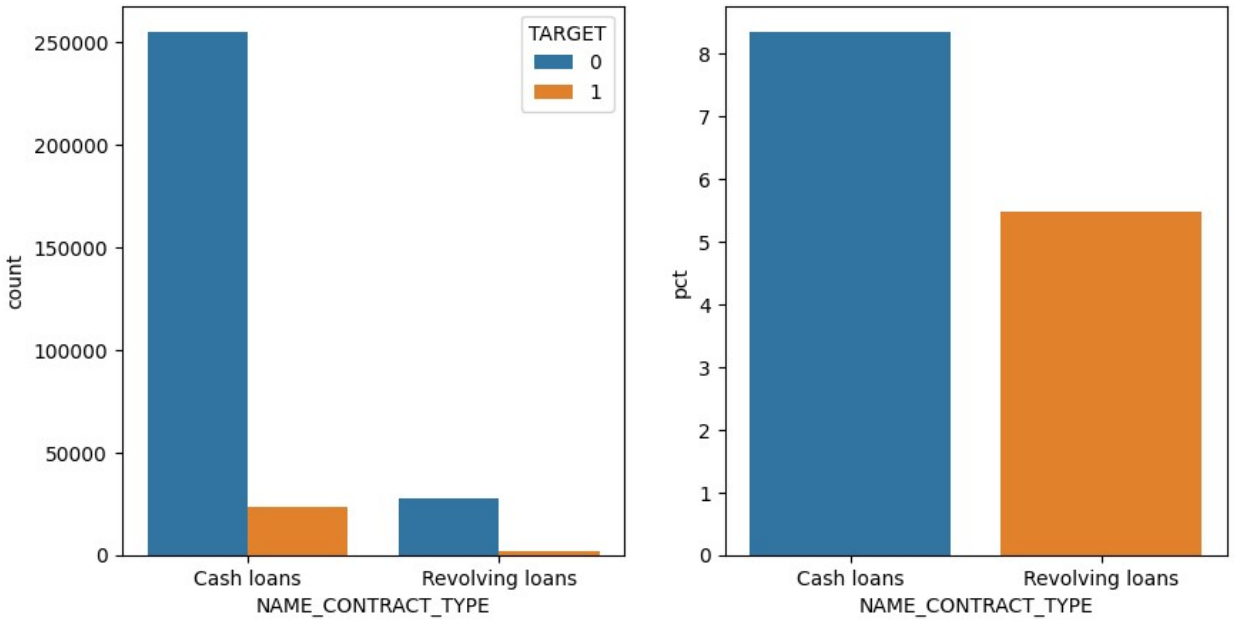

```
obj_var
```

```
Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE',
'WEEKDAY_APPR_PROCESS_START',
       'ORGANIZATION_TYPE'],
      dtype='object')
```

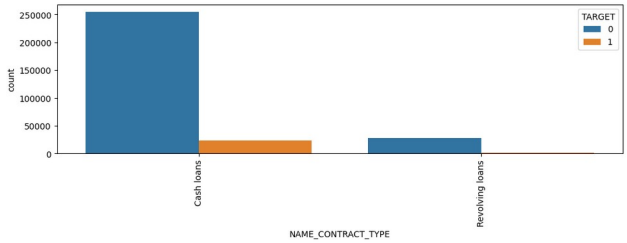
```
plt.figure(figsize=(10,5))
```

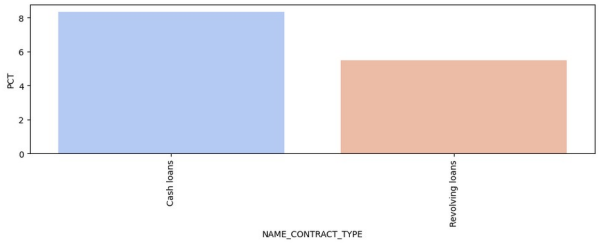
```
plt.subplot(1,2,1)
sns.countplot(data=app_score_col_rmvd,x
='NAME_CONTRACT_TYPE',hue='TARGET')

plt.subplot(1,2,2)
sns.barplot(data=data_pct,x ='NAME_CONTRACT_TYPE',y='pct')
```

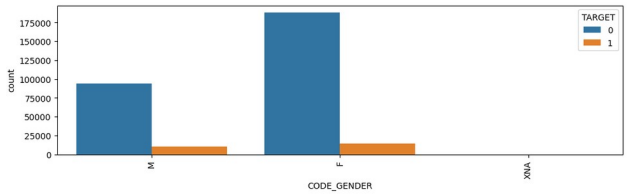
```
<Axes: xlabel='NAME_CONTRACT_TYPE', ylabel='pct'>
```

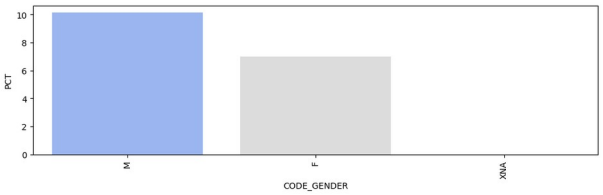


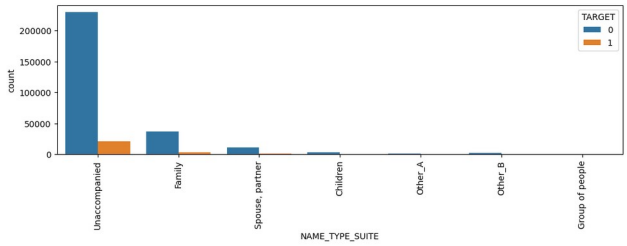

```
plt.figure(figsize=(25,60))


for i, var in enumerate(obj_var):
    data_pct =
app_score_col_rmvd[[var,'TARGET']].groupby([var],as_index=False).mean(
).sort_values(by='TARGET',ascending =False)
    data_pct['PCT'] = data_pct['TARGET']*100

    plt.subplot(10,2,i+i+1)
    plt.subplots_adjust(wspace =0.1,hspace=1)
    sns.countplot(data=app_score_col_rmvd,x=var,hue='TARGET')
    plt.xticks(rotation = 90)

    plt.subplot(10,2,i+i+2)
    sns.barplot(data=data_pct,x=var,y='PCT',palette='coolwarm')
    plt.xticks(rotation=90)
```

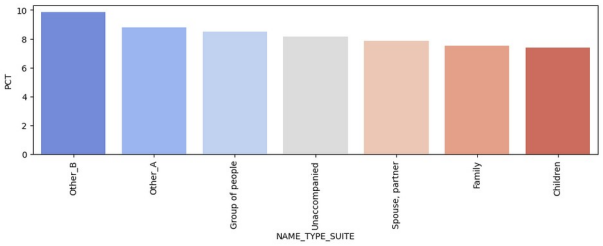

TARGET
0
1
count
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

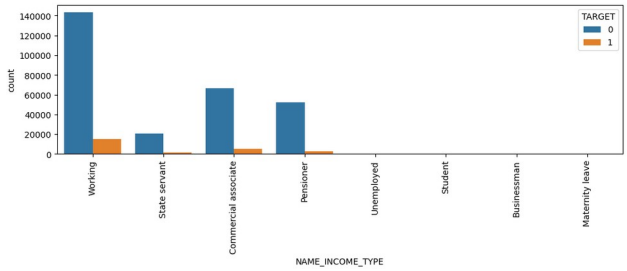

PCT
Cash loans
Revolving loans
NAME_CONTRACT_TYPE

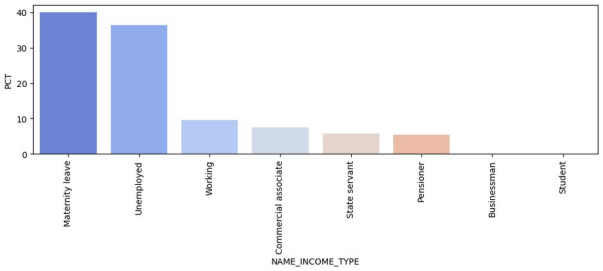

TARGET
0
1
count
M
F
XNA
CODE_GENDER

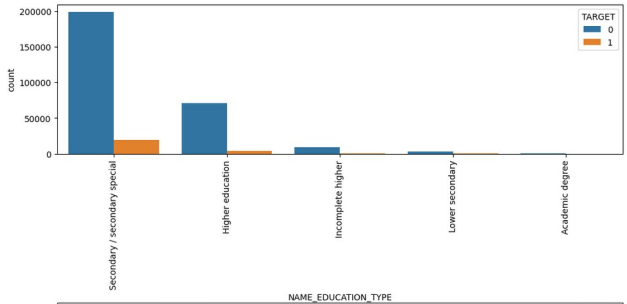

PCT
M
F
XNA
CODE_GENDER

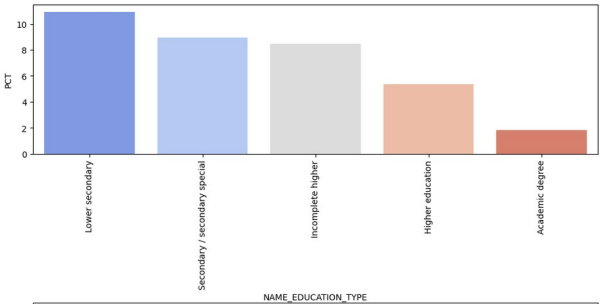

TARGET
0
1
count
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people
NAME_TYPE_SUITE

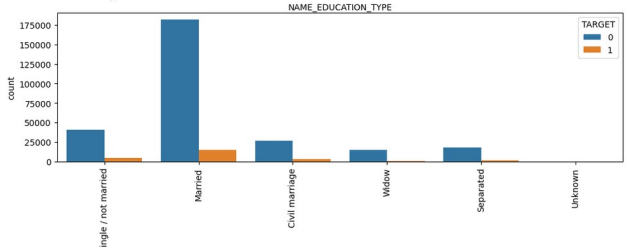

PCT
Other_B
Other_A
Group of people
Unaccompanied
Spouse, partner
Family
Children
NAME_TYPE_SUITE

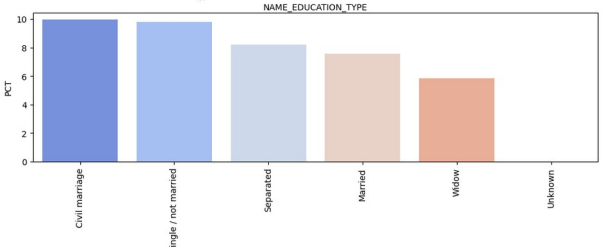

TARGET
0
1
count
Working
State servant
Commercial associate
Pensioner
Unemployed
Student
Businessman
Maternity leave
NAME_INCOME_TYPE


PCT
Maternity leave
Unemployed
Working
Commercial associate
State servant
Pensioner
Businessman
Student
NAME_INCOME_TYPE


TARGET
0
1
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE


PCT
Lower secondary
Secondary / secondary special
Incomplete higher
Higher education
Academic degree
NAME_EDUCATION_TYPE


TARGET
0
1
count
ingle / not married
Married
Civil marriage
Widow
Separated
Unknown


PCT
Civil marriage
ingle / not married
Separated
Married
Widow
Unknown

```
app_score_col_rmvd.dtypes.value_counts()
```

```
float64     18
int64       15
object      10
category     1
category     1
category     1
category     1
category     1
category     1
dtype: int64
```

```
num_var =
app_score_col_rmvd.select_dtypes(include=['float64','int64']).columns
num_cat_var =
app_score_col_rmvd.select_dtypes(include=['float64','int64','category'
]).columns
len(num_var)
```

```
33
```

```
app_score_col_rmvd[num_var].head()
```

```
   SK_ID_CURR  TARGET  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0      100002       1             0          202500.0    406597.5
1      100003       0             0          270000.0   1293502.5
2      100004       0             0           67500.0    135000.0
3      100006       0             0          135000.0    312682.5
4      100007       0             0          121500.0    513000.0

   AMT_ANNUITY  AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE
DAYS_BIRTH  \
0      24700.5         351000.0                    0.018801
9461
1      35698.5        1129500.0                    0.003541
16765
2       6750.0         135000.0                    0.010032
19046
3      29686.5         297000.0                    0.008019
19005
4      21865.5         513000.0                    0.028663
19932

   DAYS_EMPLOYED  ...  DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE  \
0            637  ...                       2.0
2.0
1           1188  ...                       0.0
1.0
2            225  ...                       0.0
```

```
0.0
3           3039  ...                         0.0
2.0
4           3038  ...                         0.0
0.0

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

   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                         0.0                        0.0
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                         0.0                        0.0
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                        0.0                        1.0
4                        0.0                        0.0

[5 rows x 33 columns]
```

```
num_data = app_score_col_rmvd[num_var]
num_data.groupby(['TARGET']).size()/num_data.shape[0]*100
```

```
TARGET
0    91.927118
1     8.072882
dtype: float64
```

```
defaulters = num_data[num_data['TARGET'] == 1].drop(['TARGET'],axis=1)
repayers = num_data[num_data['TARGET'] == 0].drop(['TARGET'],axis=1)
repayers.head()
defaulters.head()
```

```
    SK_ID_CURR  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT
AMT_ANNUITY  \
```

```
0       100002              0          202500.0    406597.5
24700.5
26      100031              0          112500.0    979992.0
27076.5
40      100047              0          202500.0   1193580.0
35028.0
42      100049              0          135000.0    288873.0
16258.5
81      100096              0           81000.0    252000.0
14593.5

    AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE  DAYS_BIRTH
DAYS_EMPLOYED  \
0          351000.0                    0.018801        9461
637
26         702000.0                    0.018029       18724
2628
40         855000.0                    0.025164       17482
1262
42         238500.0                    0.007305       13384
3597
81         252000.0                    0.028663       24794
365243

    DAYS_REGISTRATION  ...  DEF_30_CNT_SOCIAL_CIRCLE  \
0              3648.0  ...                       2.0
26             6573.0  ...                       1.0
40             1182.0  ...                       0.0
42               45.0  ...                       0.0
81             5391.0  ...                       1.0

    OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                        2.0                       2.0
26                      10.0                       0.0
40                       0.0                       0.0
42                       1.0                       0.0
81                       1.0                       1.0

    DAYS_LAST_PHONE_CHANGE  AMT_REQ_CREDIT_BUREAU_HOUR  \
0                   1134.0                         0.0
26                   161.0                         0.0
40                  1075.0                         0.0
42                  1480.0                         0.0
81                     0.0                         0.0

    AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
0                         0.0                        0.0
26                        0.0                        0.0
40                        0.0                        0.0
42                        0.0                        0.0
```

```
81                          0.0                           0.0

    AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                           0.0                         0.0
26                          0.0                         2.0
40                          2.0                         0.0
42                          0.0                         0.0
81                          0.0                         0.0

    AMT_REQ_CREDIT_BUREAU_YEAR
0                          1.0
26                         2.0
40                         4.0
42                         2.0
81                         0.0

[5 rows x 32 columns]
```

```
defaulters.head()
```

```
    SK_ID_CURR  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT
AMT_ANNUITY  \
0       100002             0          202500.0    406597.5
24700.5
26      100031             0          112500.0    979992.0
27076.5
40      100047             0          202500.0   1193580.0
35028.0
42      100049             0          135000.0    288873.0
16258.5
81      100096             0           81000.0    252000.0
14593.5

    AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE  DAYS_BIRTH
DAYS_EMPLOYED  \
0          351000.0                    0.018801        9461
637
26         702000.0                    0.018029       18724
2628
40         855000.0                    0.025164       17482
1262
42         238500.0                    0.007305       13384
3597
81         252000.0                    0.028663       24794
365243

    DAYS_REGISTRATION  ...  DEF_30_CNT_SOCIAL_CIRCLE  \
0              3648.0  ...                       2.0
26             6573.0  ...                       1.0
40             1182.0  ...                       0.0
```

```
42                45.0  ...                          0.0
81              5391.0  ...                          1.0

    OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
0                        2.0                       2.0
26                      10.0                       0.0
40                       0.0                       0.0
42                       1.0                       0.0
81                       1.0                       1.0

    DAYS_LAST_PHONE_CHANGE  AMT_REQ_CREDIT_BUREAU_HOUR  \
0                   1134.0                         0.0
26                   161.0                         0.0
40                  1075.0                         0.0
42                  1480.0                         0.0
81                     0.0                         0.0

    AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
0                         0.0                         0.0
26                        0.0                         0.0
40                        0.0                         0.0
42                        0.0                         0.0
81                        0.0                         0.0

    AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                         0.0                        0.0
26                        0.0                        2.0
40                        2.0                        0.0
42                        0.0                        0.0
81                        0.0                        0.0

    AMT_REQ_CREDIT_BUREAU_YEAR
0                          1.0
26                         2.0
40                         4.0
42                         2.0
81                         0.0

[5 rows x 32 columns]
```

```
defaulter_corr=defaulters.corr()

defaulter_corr_unstack=defaulter_corr.where(np.triu(np.ones(defaulter_
corr.shape),k=1).astype(np.bool_)).unstack().reset_index().rename(colu
mns={'level_1':'var_1',

'level_2':'var_2',

0:'corr'})
defaulter_corr_unstack['corr'] =abs(defaulter_corr_unstack['corr'])
```

```
defaulter_corr_unstack.dropna(subset=['corr']).sort_values(['corr'],as
cending=False).head(10)
```

```
                         level_0                       var_1      corr
757     OBS_60_CNT_SOCIAL_CIRCLE    OBS_30_CNT_SOCIAL_CIRCLE  0.998269
163             AMT_GOODS_PRICE                  AMT_CREDIT  0.982783
428  REGION_RATING_CLIENT_W_CITY        REGION_RATING_CLIENT  0.956637
353             CNT_FAM_MEMBERS                CNT_CHILDREN  0.885484
790     DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE  0.868994
560  LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION  0.847885
659      LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY  0.778540
164             AMT_GOODS_PRICE                 AMT_ANNUITY  0.752295
131                  AMT_ANNUITY                  AMT_CREDIT  0.752195
263                DAYS_EMPLOYED                  DAYS_BIRTH  0.582185
```

```
repayers.head()
repayers_corr = repayers.corr()

repayers_corr_unstack =
repayers_corr.where(np.triu(np.ones(repayers_corr.shape),k=1).astype(n
p.bool_)).unstack().reset_index().rename(columns={'level_1':'var_1','l
evel_2':'var_2',0:'corr'})
repayers_corr_unstack['corr'] = abs(repayers_corr_unstack['corr'])
repayers_corr_unstack.dropna(subset=['corr']).sort_values(['corr'],asc
ending=False).head(10)
```

```
                         level_0                       var_1      corr
757     OBS_60_CNT_SOCIAL_CIRCLE    OBS_30_CNT_SOCIAL_CIRCLE  0.998508
163             AMT_GOODS_PRICE                  AMT_CREDIT  0.987022
428  REGION_RATING_CLIENT_W_CITY        REGION_RATING_CLIENT  0.950149
353             CNT_FAM_MEMBERS                CNT_CHILDREN  0.878571
560  LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION  0.861861
790     DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE  0.859332
659      LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY  0.830381
164             AMT_GOODS_PRICE                 AMT_ANNUITY  0.776421
131                  AMT_ANNUITY                  AMT_CREDIT  0.771297
263                DAYS_EMPLOYED                  DAYS_BIRTH  0.626114
```

```
num_data.head()
```

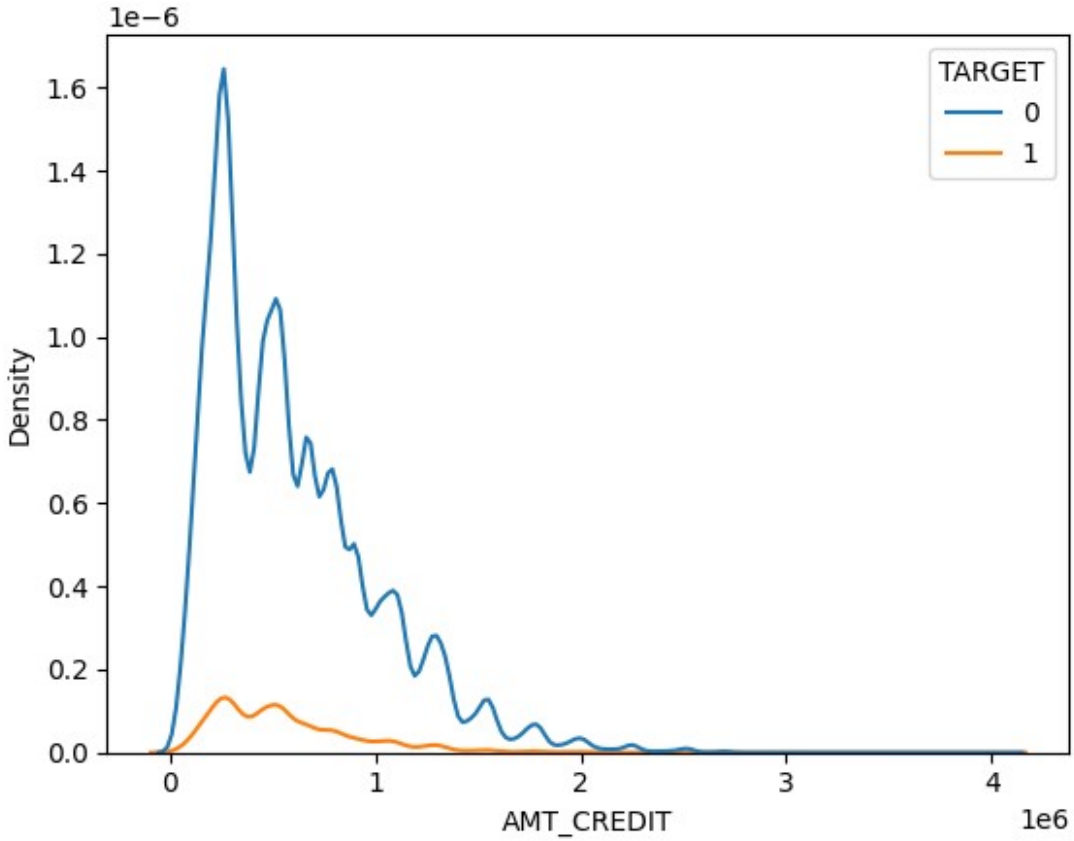
```
   SK_ID_CURR  TARGET  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0      100002       1             0          202500.0    406597.5
1      100003       0             0          270000.0   1293502.5
2      100004       0             0           67500.0    135000.0
3      100006       0             0          135000.0    312682.5
4      100007       0             0          121500.0    513000.0

   AMT_ANNUITY  AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE
DAYS_BIRTH  \
0      24700.5         351000.0                    0.018801
9461
```

```
1      35698.5        1129500.0                    0.003541
16765
2       6750.0         135000.0                    0.010032
19046
3      29686.5         297000.0                    0.008019
19005
4      21865.5         513000.0                    0.028663
19932

   DAYS_EMPLOYED  ...  DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE  \
0            637  ...                       2.0
2.0
1           1188  ...                       0.0
1.0
2            225  ...                       0.0
0.0
3           3039  ...                       0.0
2.0
4           3038  ...                       0.0
0.0

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

   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                         0.0                        0.0
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                         0.0                        0.0
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                        0.0                        1.0
4                        0.0                        0.0
```

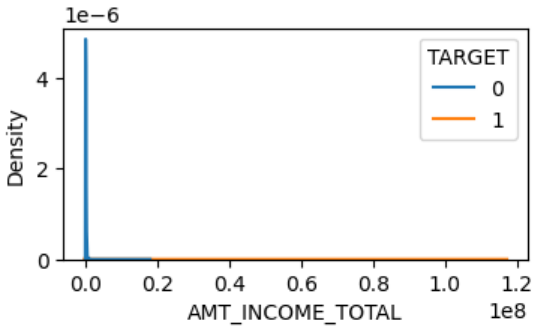
```
[5 rows x 33 columns]
```

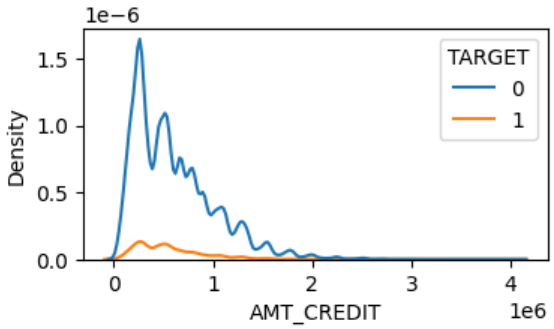
```
amt_var =
['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE']

sns.kdeplot(data=num_data,x='AMT_CREDIT',hue='TARGET')
```

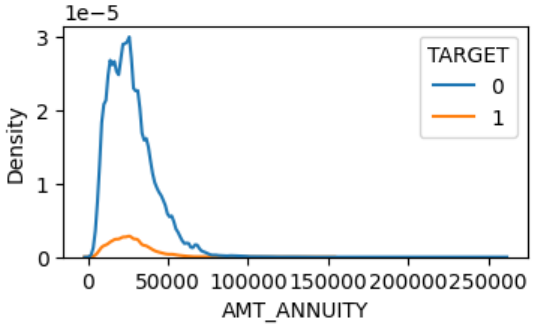
```
<Axes: xlabel='AMT_CREDIT', ylabel='Density'>
```

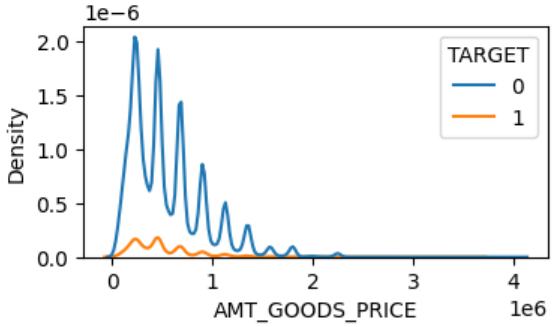



```
plt.figure(figsize=(10,5))

for i,col in enumerate(amt_var):
    plt.subplot(2,2,i+1)
    sns.kdeplot(data=num_data,x=col,hue='TARGET')
    plt.subplots_adjust(wspace=0.5,hspace=0.5)
```









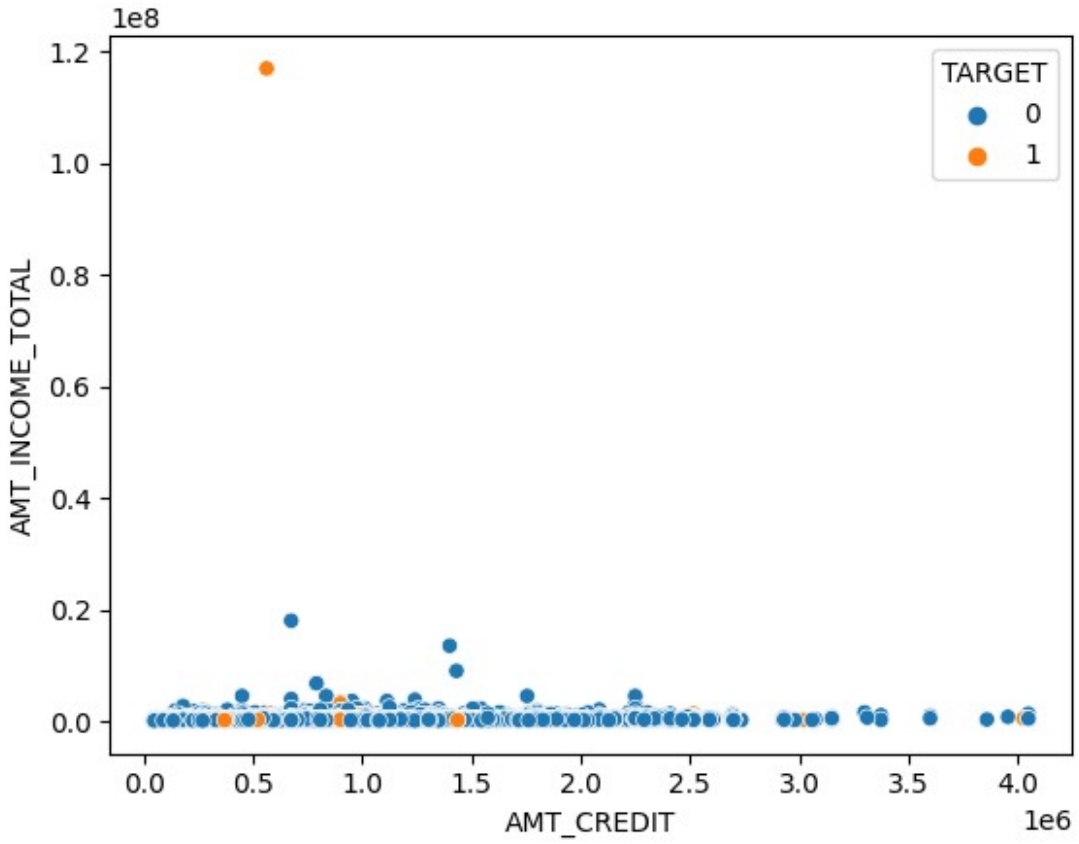

```
num_data.head()
```

```
   SK_ID_CURR  TARGET  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0      100002       1             0          202500.0    406597.5
1      100003       0             0          270000.0   1293502.5
2      100004       0             0           67500.0    135000.0
3      100006       0             0          135000.0    312682.5
4      100007       0             0          121500.0    513000.0

   AMT_ANNUITY  AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE
DAYS_BIRTH  \
0      24700.5         351000.0                    0.018801
9461
1      35698.5        1129500.0                    0.003541
16765
2       6750.0         135000.0                    0.010032
19046
3      29686.5         297000.0                    0.008019
19005
4      21865.5         513000.0                    0.028663
19932

   DAYS_EMPLOYED  ...  DEF_30_CNT_SOCIAL_CIRCLE
OBS_60_CNT_SOCIAL_CIRCLE  \
0            637  ...                       2.0
2.0
1           1188  ...                       0.0
1.0
2            225  ...                       0.0
```

```
0.0
3           3039  ...                         0.0
2.0
4           3038  ...                         0.0
0.0

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

   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                         0.0                        0.0
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                         0.0                        0.0
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                        0.0                        1.0
4                        0.0                        0.0

[5 rows x 33 columns]
```

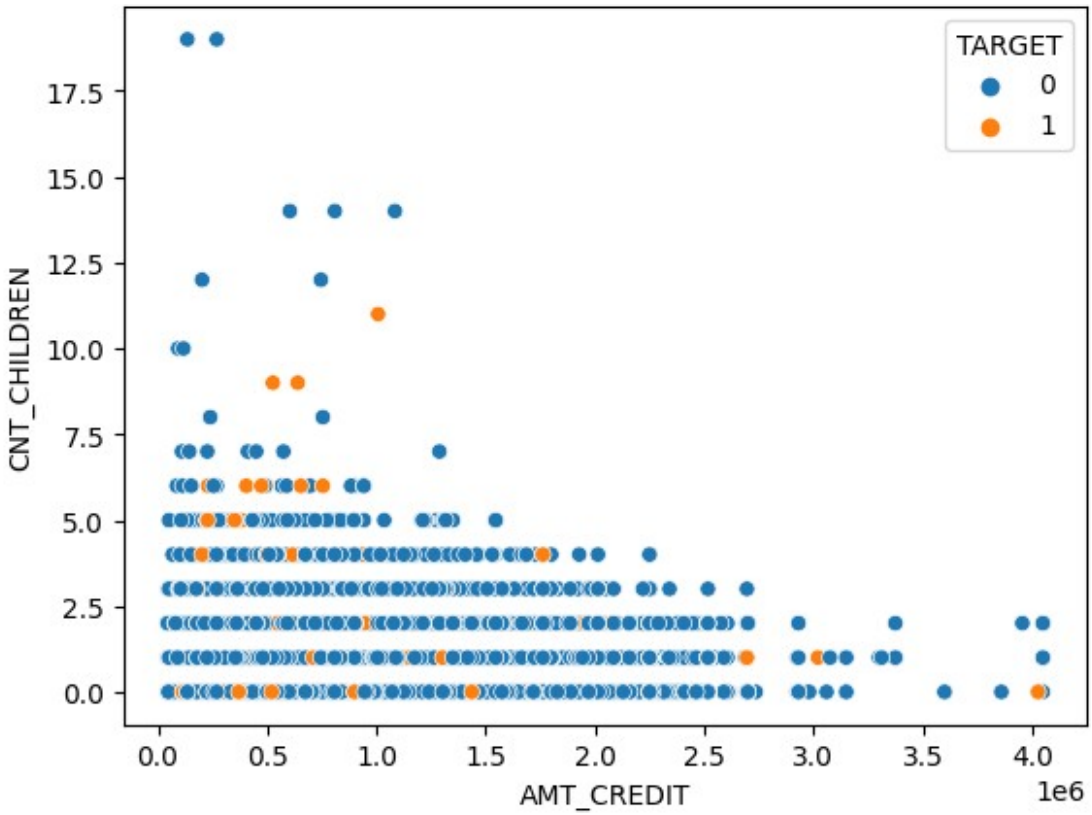
```
sns.scatterplot(data=num_data,x='AMT_CREDIT',y='AMT_INCOME_TOTAL',hue=
'TARGET')
```

```
<Axes: xlabel='AMT_CREDIT', ylabel='AMT_INCOME_TOTAL'>
```



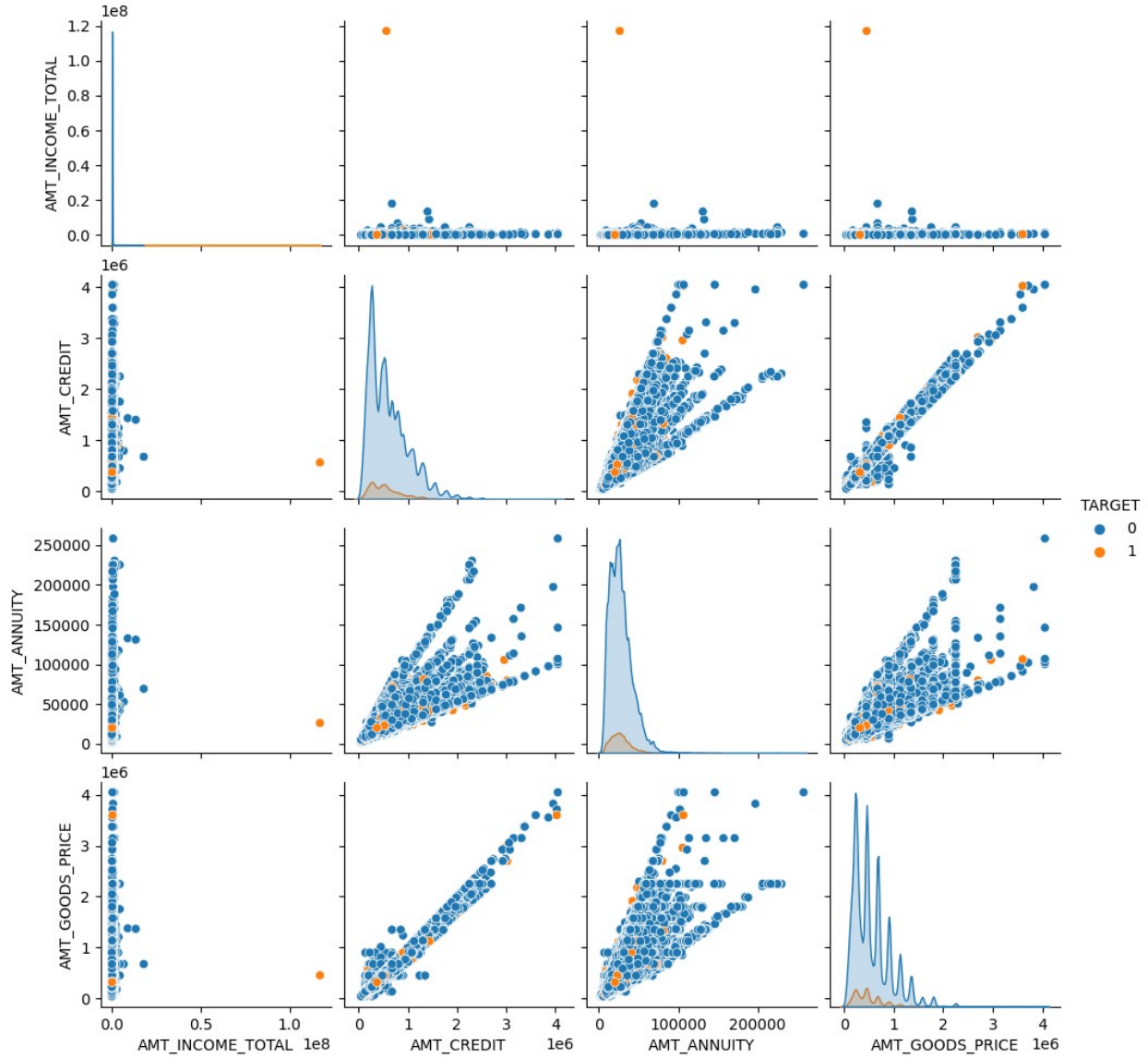

```
sns.scatterplot(data=num_data,x='AMT_CREDIT',y='CNT_CHILDREN',hue='TAR
GET')
```

```
<Axes: xlabel='AMT_CREDIT', ylabel='CNT_CHILDREN'>
```




```
amt_var1 =
num_data[['AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRI
CE' , 'TARGET']]

sns.pairplot(data=amt_var1 , hue = 'TARGET')
```

```
<seaborn.axisgrid.PairGrid at 0x1cc86002d90>
```




```
prev_app = pd.read_csv('previous_application.csv')

prev_app.head()
```

```
   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_DOWN_PAYMENT  AMT_GOODS_PRICE
WEEKDAY_APPR_PROCESS_START  \
0     17145.0               0.0          17145.0
SATURDAY
1    679671.0               NaN         607500.0
THURSDAY
2    136444.5               NaN         112500.0
TUESDAY
3    470790.0               NaN         450000.0
MONDAY
4    404055.0               NaN         337500.0
THURSDAY

   HOUR_APPR_PROCESS_START  ... NAME_SELLER_INDUSTRY  CNT_PAYMENT  \
0                       15  ...         Connectivity         12.0
1                       11  ...                  XNA         36.0
2                       11  ...                  XNA         12.0
3                        7  ...                  XNA         12.0
4                        9  ...                  XNA         24.0

   NAME_YIELD_GROUP       PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  \
0            middle  POS mobile with interest            365243.0
1        low_action          Cash X-Sell: low            365243.0
2              high         Cash X-Sell: high            365243.0
3            middle       Cash X-Sell: middle            365243.0
4              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

[5 rows x 37 columns]
```

```
null_count =
pd.DataFrame(prev_app.isnull().sum().sort_values(ascending =
False)/prev_app.shape[0]*100).reset_index().rename(columns =
{'index':'var',0: 'count_pct'})

null_count
```

	var	count_pct
0	RATE_INTEREST_PRIVILEGED	99.640366
1	RATE_INTEREST_PRIMARY	99.640366
2	AMT_DOWN_PAYMENT	53.380827
3	RATE_DOWN_PAYMENT	53.380827
4	NAME_TYPE_SUITE	49.082042
5	NFLAG_INSURED_ON_APPROVAL	40.150474
6	DAYS_TERMINATION	40.150474
7	DAYS_LAST_DUE	40.150474
8	DAYS_LAST_DUE_1ST_VERSION	40.150474
9	DAYS_FIRST_DUE	40.150474
10	DAYS_FIRST_DRAWING	40.150407
11	AMT_GOODS_PRICE	22.978818
12	AMT_ANNUITY	22.211173
13	CNT_PAYMENT	22.210971
14	PRODUCT_COMBINATION	0.020454
15	AMT_CREDIT	0.000067
16	NAME_YIELD_GROUP	0.000000
17	NAME_PORTFOLIO	0.000000
18	NAME_SELLER_INDUSTRY	0.000000
19	SELLERPLACE_AREA	0.000000
20	CHANNEL_TYPE	0.000000
21	NAME_PRODUCT_TYPE	0.000000
22	SK_ID_PREV	0.000000
23	NAME_GOODS_CATEGORY	0.000000
24	NAME_CLIENT_TYPE	0.000000
25	CODE_REJECT_REASON	0.000000
26	SK_ID_CURR	0.000000
27	DAYS_DECISION	0.000000
28	NAME_CONTRACT_STATUS	0.000000
29	NAME_CASH_LOAN_PURPOSE	0.000000
30	NFLAG_LAST_APPL_IN_DAY	0.000000
31	FLAG_LAST_APPL_PER_CONTRACT	0.000000
32	HOUR_APPR_PROCESS_START	0.000000
33	WEEKDAY_APPR_PROCESS_START	0.000000
34	AMT_APPLICATION	0.000000
35	NAME_CONTRACT_TYPE	0.000000
36	NAME_PAYMENT_TYPE	0.000000

```
var_msng_ge_40 = list(null_count[null_count['count_pct']>40]['var'])

var_msng_ge_40
```

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

nva_cols = var_msng_ge_40 +
['WEEKDAY_APPR_PROCESS_START','HOUR_APPR_PROCESS_START','FLAG_LAST_APP
L_PER_CONTRACT','NFLAG_LAST_APPL_IN_DAY']

len(nva_cols)

15

len(prev_app.columns)

37

prev_app_nva_col_rmvd = prev_app.drop(labels =nva_cols, axis = 1)

len(prev_app_nva_col_rmvd.columns)

22

prev_app_nva_col_rmvd.head()

   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 NAME_CASH_LOAN_PURPOSE
NAME_CONTRACT_STATUS  \
0     17145.0          17145.0                   XAP
Approved
1    679671.0         607500.0                   XNA
Approved
2    136444.5         112500.0                   XNA
```

```
Approved
3    470790.0          450000.0                     XNA
Approved
4    404055.0          337500.0                 Repairs
Refused

   DAYS_DECISION  ... NAME_CLIENT_TYPE NAME_GOODS_CATEGORY
NAME_PORTFOLIO  \
0            -73  ...         Repeater               Mobile
POS
1           -164  ...         Repeater                  XNA
Cash
2           -301  ...         Repeater                  XNA
Cash
3           -512  ...         Repeater                  XNA
Cash
4           -781  ...         Repeater                  XNA
Cash

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

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

[5 rows x 22 columns]
```

```
prev_app_nva_col_rmvd.isnull().sum().sort_values(ascending =
False)/prev_app_nva_col_rmvd.shape[0]*100
```

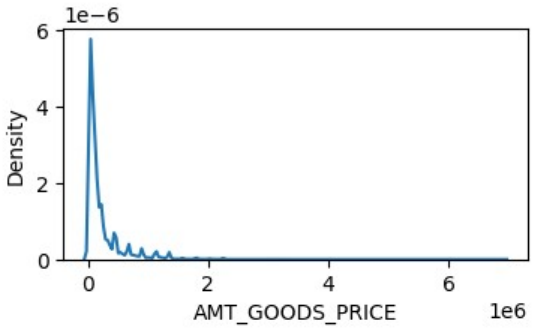
```
AMT_GOODS_PRICE           22.978818
AMT_ANNUITY               22.211173
CNT_PAYMENT               22.210971
PRODUCT_COMBINATION        0.020454
AMT_CREDIT                 0.000067
NAME_GOODS_CATEGORY        0.000000
```

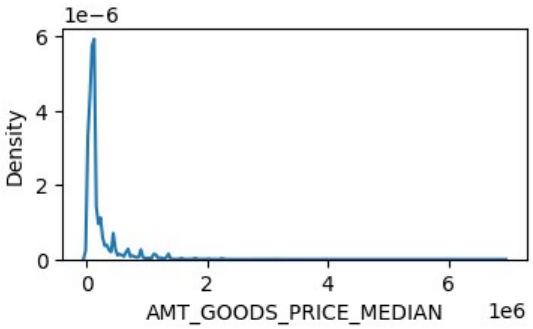
```
NAME_YIELD_GROUP            0.000000
NAME_SELLER_INDUSTRY        0.000000
SELLERPLACE_AREA            0.000000
CHANNEL_TYPE                0.000000
NAME_PRODUCT_TYPE           0.000000
NAME_PORTFOLIO              0.000000
SK_ID_PREV                  0.000000
NAME_CLIENT_TYPE            0.000000
SK_ID_CURR                  0.000000
NAME_PAYMENT_TYPE           0.000000
DAYS_DECISION               0.000000
NAME_CONTRACT_STATUS        0.000000
NAME_CASH_LOAN_PURPOSE      0.000000
AMT_APPLICATION             0.000000
NAME_CONTRACT_TYPE          0.000000
CODE_REJECT_REASON          0.000000
dtype: float64
```

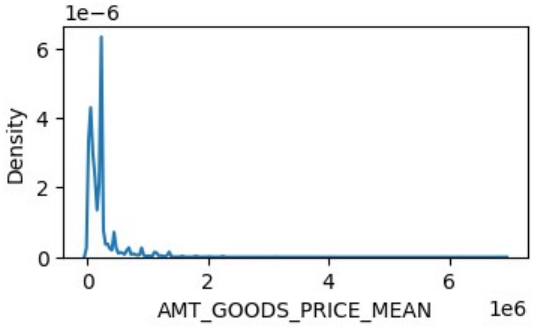
```
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].agg(func=['mean','median'])
```

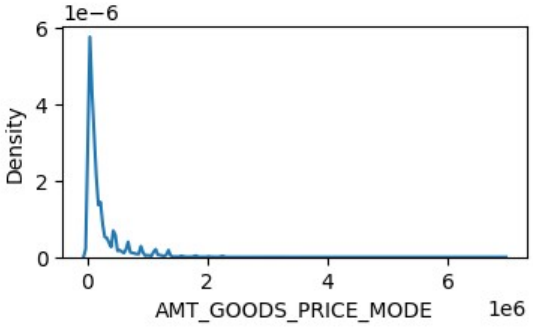
```
mean      226572.810903
median    111555.000000
Name: AMT_GOODS_PRICE, dtype: float64
```

```
prev_app_nva_col_rmvd['AMT_GOODS_PRICE_MEDIAN'] =
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].fillna(prev_app_nva_col_rmvd[
'AMT_GOODS_PRICE'].median())

prev_app_nva_col_rmvd['AMT_GOODS_PRICE_MEAN'] =
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].fillna(prev_app_nva_col_rmvd[
'AMT_GOODS_PRICE'].mean())

prev_app_nva_col_rmvd['AMT_GOODS_PRICE_MODE'] =
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].fillna(prev_app_nva_col_rmvd[
'AMT_GOODS_PRICE'].mode())

gp_cols =
['AMT_GOODS_PRICE','AMT_GOODS_PRICE_MEDIAN','AMT_GOODS_PRICE_MEAN','AM
T_GOODS_PRICE_MODE']

plt.figure(figsize=(10,5))

for i, col in enumerate(gp_cols):
    plt.subplot(2,2,i+1)
    sns.kdeplot(data =prev_app_nva_col_rmvd,x=col)
    plt.subplots_adjust(wspace=0.5,hspace=0.5)
```










```
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'] =
prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].fillna(prev_app_nva_col_rmvd[
'AMT_GOODS_PRICE'].median)

prev_app_nva_col_rmvd['AMT_GOODS_PRICE'].isnull().sum()

0

prev_app_nva_col_rmvd['AMT_ANNUITY'].agg(func =
['mean','median','max'])

mean       15901.781831
median     11250.000000
max       418058.145000
Name: AMT_ANNUITY, dtype: float64

prev_app_nva_col_rmvd['AMT_ANNUITY'] =
prev_app_nva_col_rmvd['AMT_ANNUITY'].fillna(prev_app_nva_col_rmvd['AMT
_ANNUITY'].median)

prev_app_nva_col_rmvd['AMT_ANNUITY'].isnull().sum()

0

prev_app_nva_col_rmvd['PRODUCT_COMBINATION'] =
prev_app_nva_col_rmvd['PRODUCT_COMBINATION'].fillna(prev_app_nva_col_r
mvd['PRODUCT_COMBINATION'].mode)

prev_app_nva_col_rmvd['PRODUCT_COMBINATION'].isnull().sum()
```

```
0
```

```
prev_app_nva_col_rmvd['CNT_PAYMENT'].agg(func=['mean','median','max'])
```

```
mean      16.007697
median    12.000000
max       84.000000
Name: CNT_PAYMENT, dtype: float64
```

```
prev_app_nva_col_rmvd[prev_app_nva_col_rmvd['CNT_PAYMENT'].isnull()].g
roupby(['NAME_CONTRACT_STATUS']).size().sort_values(ascending=False)
```

```
NAME_CONTRACT_STATUS
Canceled        270859
Refused          36385
Unused offer     22859
Approved             4
dtype: int64
```

```
prev_app_nva_col_rmvd['CNT_PAYMENT'] =
prev_app_nva_col_rmvd['CNT_PAYMENT'].fillna(0)
```

```
prev_app_nva_col_rmvd.isnull().sum()
```

```
SK_ID_PREV                      0
SK_ID_CURR                      0
NAME_CONTRACT_TYPE              0
AMT_ANNUITY                     0
AMT_APPLICATION                 0
AMT_CREDIT                      1
AMT_GOODS_PRICE                 0
NAME_CASH_LOAN_PURPOSE          0
NAME_CONTRACT_STATUS            0
DAYS_DECISION                   0
NAME_PAYMENT_TYPE               0
CODE_REJECT_REASON              0
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                     0
NAME_YIELD_GROUP                0
PRODUCT_COMBINATION             0
AMT_GOODS_PRICE_MEDIAN          0
AMT_GOODS_PRICE_MEAN            0
AMT_GOODS_PRICE_MODE       341519
dtype: int64
```

```
prev_app_nva_col_rmvd.columns
```

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE',
'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE',
       'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS',
'DAYS_DECISION',
       'NAME_PAYMENT_TYPE', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE',
       'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
       'CHANNEL_TYPE', 'SELLERPLACE_AREA', 'NAME_SELLER_INDUSTRY',
       'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION',
       'AMT_GOODS_PRICE_MEDIAN', 'AMT_GOODS_PRICE_MEAN',
       'AMT_GOODS_PRICE_MODE'],
      dtype='object')
```

```
prev_app_nva_col_rmvd = prev_app_nva_col_rmvd.drop(labels =
['AMT_GOODS_PRICE_MEDIAN', 'AMT_GOODS_PRICE_MEAN',
       'AMT_GOODS_PRICE_MODE'],axis =1)
```

```
prev_app_nva_col_rmvd.columns
```

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE',
'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE',
       'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS',
'DAYS_DECISION',
       'NAME_PAYMENT_TYPE', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE',
       'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
       'CHANNEL_TYPE', 'SELLERPLACE_AREA', 'NAME_SELLER_INDUSTRY',
       'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION'],
      dtype='object')
```

```
    merged_df =
pd.merge(app_score_col_rmvd,prev_app_nva_col_rmvd,how='inner',on =
'SK_ID_CURR')
```

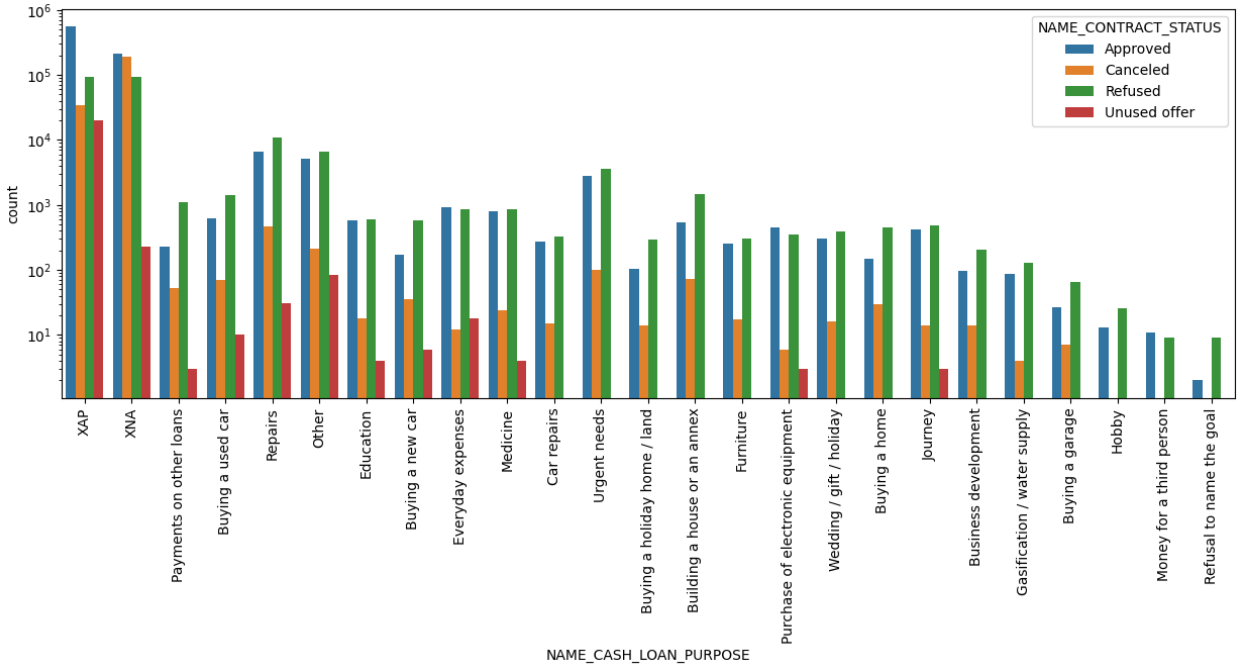
```
merged_df
```

```
         SK_ID_CURR  TARGET NAME_CONTRACT_TYPE_x CODE_GENDER
CNT_CHILDREN  \
0            100002       1           Cash loans           M
0
1            100003       0           Cash loans           F
0
2            100003       0           Cash loans           F
0
3            100003       0           Cash loans           F
0
4            100004       0      Revolving loans           M
0
...             ...     ...                  ...         ...
```

```
...
1258173      456255       0           Cash loans           F
0
1258174      456255       0           Cash loans           F
0
1258175      456255       0           Cash loans           F
0
1258176      456255       0           Cash loans           F
0
1258177      456255       0           Cash loans           F
0

         AMT_INCOME_TOTAL  AMT_CREDIT_x  AMT_ANNUITY_x
AMT_GOODS_PRICE_x  \
0                202500.0      406597.5        24700.5
351000.0
1                270000.0     1293502.5        35698.5
1129500.0
2                270000.0     1293502.5        35698.5
1129500.0
3                270000.0     1293502.5        35698.5
1129500.0
4                 67500.0      135000.0         6750.0
135000.0
...                   ...           ...            ...
...
1258173          157500.0      675000.0        49117.5
675000.0
1258174          157500.0      675000.0        49117.5
675000.0
1258175          157500.0      675000.0        49117.5
675000.0
1258176          157500.0      675000.0        49117.5
675000.0
1258177          157500.0      675000.0        49117.5
675000.0

        NAME_TYPE_SUITE  ... NAME_CLIENT_TYPE   NAME_GOODS_CATEGORY  \
0         Unaccompanied  ...              New              Vehicles
1                Family  ...         Repeater                   XNA
2                Family  ...        Refreshed             Furniture
3                Family  ...        Refreshed  Consumer Electronics
4         Unaccompanied  ...              New                Mobile
...                 ...  ...              ...                   ...
1258173   Unaccompanied  ...         Repeater                   XNA
1258174   Unaccompanied  ...         Repeater                   XNA
1258175   Unaccompanied  ...         Repeater                   XNA
1258176   Unaccompanied  ...         Repeater                   XNA
1258177   Unaccompanied  ...         Repeater             Computers
```

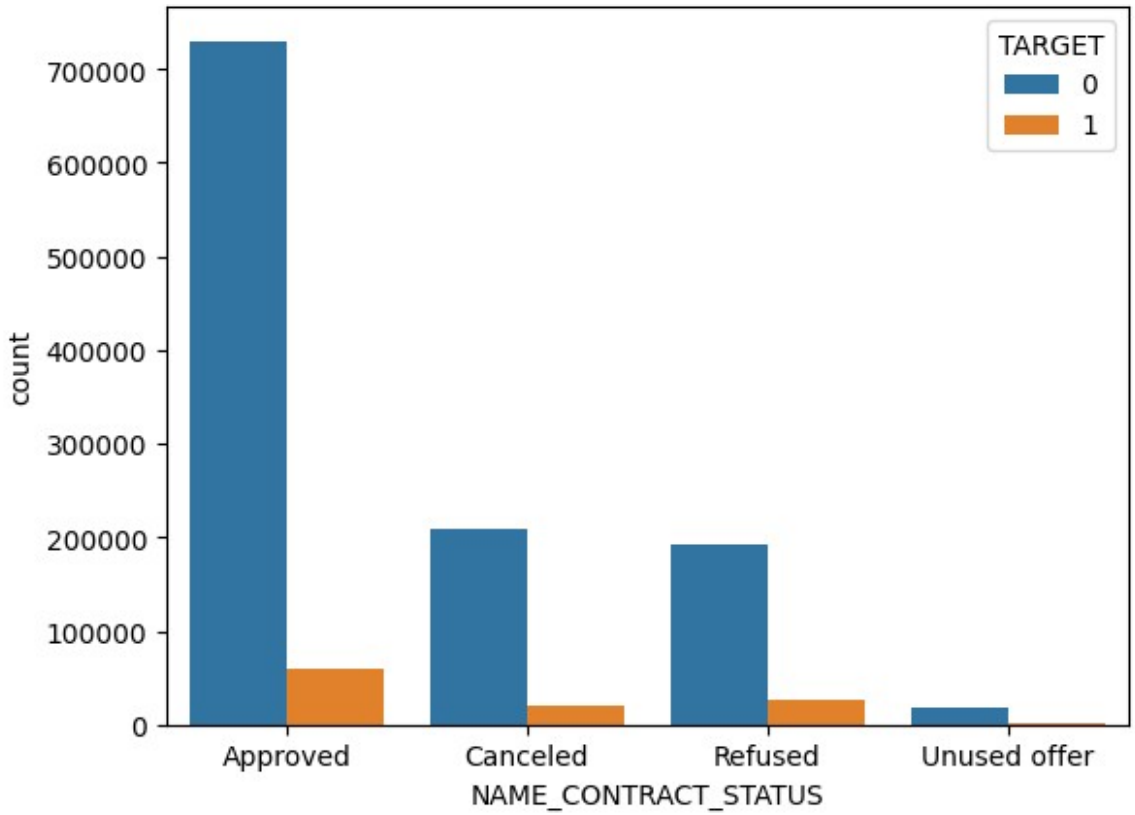
```
        NAME_PORTFOLIO NAME_PRODUCT_TYPE             CHANNEL_TYPE  \
0                  POS              XNA                    Stone
1                 Cash            x-sell  Credit and cash offices
2                  POS              XNA                    Stone
3                  POS              XNA            Country-wide
4                  POS              XNA         Regional / Local
...                ...              ...                      ...
1258173           Cash            x-sell  Credit and cash offices
1258174          Cards           walk-in            Country-wide
1258175           Cash           walk-in  Credit and cash offices
1258176           Cash            x-sell          AP+ (Cash loan)
1258177            POS              XNA            Country-wide

         SELLERPLACE_AREA  NAME_SELLER_INDUSTRY  CNT_PAYMENT  \
0                     500       Auto technology         24.0
1                      -1                   XNA         12.0
2                    1400             Furniture          6.0
3                     200  Consumer electronics         12.0
4                      30          Connectivity          4.0
...                   ...                   ...          ...
1258173                -1                   XNA         24.0
1258174                20          Connectivity          0.0
1258175                -1                   XNA         60.0
1258176                 6                   XNA         36.0
1258177                20          Connectivity          6.0

         NAME_YIELD_GROUP          PRODUCT_COMBINATION
0              low_normal      POS other with interest
1              low_normal             Cash X-Sell: low
2                  middle   POS industry with interest
3                  middle  POS household with interest
4                  middle  POS mobile without interest
...                   ...                         ...
1258173            middle         Cash X-Sell: middle
1258174               XNA                Card Street
1258175        low_normal            Cash Street: low
1258176        low_normal            Cash X-Sell: low
1258177              high     POS mobile with interest

[1258178 rows x 70 columns]
```

```
plt.figure(figsize = (15,5))
sns.countplot(data=merged_df,x='NAME_CASH_LOAN_PURPOSE',hue =
'NAME_CONTRACT_STATUS')
plt.xticks(rotation=90)
plt.yscale('log')
```




```
sns.countplot(data=merged_df,x='NAME_CONTRACT_STATUS',hue ='TARGET')
```

```
<Axes: xlabel='NAME_CONTRACT_STATUS', ylabel='count'>
```



```
merged_agg =
merged_df.groupby(['NAME_CONTRACT_STATUS','TARGET']).size().reset_inde
x().rename(columns={0:'counts'})
merged_agg
```

	NAME_CONTRACT_STATUS	TARGET	counts
0	Approved	0	730112
1	Approved	1	60046
2	Canceled	0	208407
3	Canceled	1	21137
4	Refused	0	192012
5	Refused	1	26087
6	Unused offer	0	18687
7	Unused offer	1	1690

```
sum_df = merged_agg.groupby(['NAME_CONTRACT_STATUS'])
['counts'].sum().reset_index()

sum_df
```

	NAME_CONTRACT_STATUS	counts
0	Approved	790158
1	Canceled	229544
2	Refused	218099
3	Unused offer	20377

```
 merged_agg_2 = pd.merge(merged_agg,sum_df,how
='left',on='NAME_CONTRACT_STATUS')
 merged_agg_2
```

	NAME_CONTRACT_STATUS	TARGET	counts_x	counts_y
0	Approved	0	730112	790158
1	Approved	1	60046	790158
2	Canceled	0	208407	229544
3	Canceled	1	21137	229544
4	Refused	0	192012	218099
5	Refused	1	26087	218099
6	Unused offer	0	18687	20377
7	Unused offer	1	1690	20377

```
merged_agg_2['pct'] =
round( merged_agg_2['counts_x']/merged_agg_2['counts_y']*100)
merged_agg_2
```

	NAME_CONTRACT_STATUS	TARGET	counts_x	counts_y	pct
0	Approved	0	730112	790158	92.0
1	Approved	1	60046	790158	8.0
2	Canceled	0	208407	229544	91.0
3	Canceled	1	21137	229544	9.0
4	Refused	0	192012	218099	88.0
5	Refused	1	26087	218099	12.0

```
6          Unused offer       0     18687     20377  92.0
7          Unused offer       1      1690     20377   8.0
```