In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
# seaborn is used for visual representation
```

In [2]:
```
application_data =pd.read_csv("C:/Users/Hello/Downloads/application_data.csv")
```

In [3]:
```
#understanding the data
application_data.head()
```

Out[3]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REA
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	

5 rows × 122 columns

In [4]:
```
application_data.shape
```

Out[4]:
```
(307511, 122)
```

In [5]:
```
application_data.columns
```

Out[5]:
```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY',
       ...
       'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',
       'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR',
       'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
       'AMT_REQ_CREDIT_BUREAU_YEAR'],
      dtype='object', length=122)
```

In [6]:
```
#cleaning the data
#check for the null values
null_values=(application_data.isnull().sum()/307511)*100
```

In [7]:
```
print(null_values)
```

```
SK_ID_CURR                    0.000000
TARGET                        0.000000
NAME_CONTRACT_TYPE            0.000000
CODE_GENDER                   0.000000
FLAG_OWN_CAR                  0.000000
                                ...
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
Length: 122, dtype: float64
```

In [8]:
```
drop_columns=null_values[null_values>50].index
```

In [9]:
```
application_data_filtered=application_data.drop(columns=drop_columns)
```

In [10]:
```
application_data_filtered.shape
```

Out[10]:
```
(307511, 81)
```

In [11]:
```
application_data_filtered.head()
```

Out[11]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REA
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	

5 rows × 81 columns

In [12]:
```
application_data_filtered.columns
```

Out[12]:
```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH',
       'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'FLAG_MOBIL',
       'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
       'FLAG_EMAIL', 'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS',
       'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
       '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',
       'ORGANIZATION_TYPE', 'EXT_SOURCE_2', 'EXT_SOURCE_3',
       'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG',
       'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE',
       'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_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', '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', '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')
```

In [13]:
```
drop_columns1=['FLAG_MOBIL',
       'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
       'FLAG_EMAIL','REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY','FLAG_DOCU
```

```
       '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','EXT_SOURCE_2', 'EXT_SOURCE_3',
       'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG',
       'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE',
       'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_MODE','EMERGENCY
       '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','WEEKDAY_APPR_PROCESS_START',
       'HOUR_APPR_PROCESS_START']
```

In [14]:
```
application_data_filtered=application_data_filtered.drop(columns=drop_columns1)
```

In [15]:
```
application_data_filtered.shape
```

Out[15]:
```
(307511, 35)
```

In [16]:
```
application_data_filtered.head()
```

Out[16]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REA
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	

5 rows × 35 columns

In [17]:
```
application_data_filtered.columns
```

Out[17]:
```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH',
       'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'ORGANIZATION_TYPE',
       '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')
```

In [18]:
```
change_columns=['DAYS_BIRTH','DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH'
application_data_filtered[change_columns]=application_data_filtered[change_columns].
```

In [19]:
```
((application_data_filtered.isnull().sum()/307511)*100).sort_values(ascending=True)
```

Out[19]:
```
SK_ID_CURR                    0.000000
ORGANIZATION_TYPE             0.000000
DAYS_ID_PUBLISH               0.000000
DAYS_REGISTRATION             0.000000
DAYS_EMPLOYED                 0.000000
REGION_POPULATION_RELATIVE    0.000000
NAME_HOUSING_TYPE             0.000000
NAME_FAMILY_STATUS            0.000000
NAME_EDUCATION_TYPE           0.000000
NAME_INCOME_TYPE              0.000000
DAYS_BIRTH                    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
TARGET                        0.000000
DAYS_LAST_PHONE_CHANGE        0.000325
CNT_FAM_MEMBERS               0.000650
AMT_ANNUITY                   0.003902
AMT_GOODS_PRICE               0.090403
DEF_60_CNT_SOCIAL_CIRCLE      0.332021
OBS_30_CNT_SOCIAL_CIRCLE      0.332021
DEF_30_CNT_SOCIAL_CIRCLE      0.332021
OBS_60_CNT_SOCIAL_CIRCLE      0.332021
NAME_TYPE_SUITE               0.420148
AMT_REQ_CREDIT_BUREAU_QRT    13.501631
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_YEAR   13.501631
OCCUPATION_TYPE              31.345545
dtype: float64
```

In [20]:
```
application_data_filtered['DAYS_LAST_PHONE_CHANGE'].describe()
```

Out[20]:
```
count    307510.000000
mean        962.858788
std         826.808487
min           0.000000
25%         274.000000
50%         757.000000
75%        1570.000000
max        4292.000000
Name: DAYS_LAST_PHONE_CHANGE, dtype: float64
```

In [21]:
```
application_data_filtered['DAYS_LAST_PHONE_CHANGE'].fillna(963,inplace=True)
```

In [22]:
```
application_data_filtered['DAYS_LAST_PHONE_CHANGE'].head()
```

Out[22]:
```
0    1134.0
1     828.0
2     815.0
3     617.0
4    1106.0
Name: DAYS_LAST_PHONE_CHANGE, dtype: float64
```

In [23]:
```
application_data_filtered.isnull().sum().sort_values(ascending=True)
```

Out[23]:
```
SK_ID_CURR                      0
DAYS_LAST_PHONE_CHANGE          0
ORGANIZATION_TYPE               0
DAYS_ID_PUBLISH                 0
DAYS_REGISTRATION               0
DAYS_EMPLOYED                   0
REGION_POPULATION_RELATIVE      0
NAME_HOUSING_TYPE               0
NAME_FAMILY_STATUS              0
NAME_EDUCATION_TYPE             0
NAME_INCOME_TYPE                0
DAYS_BIRTH                      0
NAME_CONTRACT_TYPE              0
AMT_CREDIT                      0
AMT_INCOME_TOTAL                0
CNT_CHILDREN                    0
TARGET                          0
FLAG_OWN_REALTY                 0
FLAG_OWN_CAR                    0
CODE_GENDER                     0
CNT_FAM_MEMBERS                 2
AMT_ANNUITY                    12
AMT_GOODS_PRICE               278
OBS_30_CNT_SOCIAL_CIRCLE     1021
DEF_30_CNT_SOCIAL_CIRCLE     1021
OBS_60_CNT_SOCIAL_CIRCLE     1021
DEF_60_CNT_SOCIAL_CIRCLE     1021
NAME_TYPE_SUITE              1292
AMT_REQ_CREDIT_BUREAU_QRT   41519
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_YEAR  41519
OCCUPATION_TYPE             96391
dtype: int64
```

In [24]:
```
application_data_filtered['CNT_FAM_MEMBERS'].describe()
```

Out[24]:
```
count    307509.000000
mean          2.152665
std           0.910682
min           1.000000
25%           2.000000
50%           2.000000
75%           3.000000
max          20.000000
Name: CNT_FAM_MEMBERS, dtype: float64
```

In [25]:
```
application_data_filtered['CNT_FAM_MEMBERS'].fillna(2,inplace=True)
```

In [26]:
```
fill_null_values=['AMT_ANNUITY','AMT_GOODS_PRICE']
application_data_filtered[fill_null_values].describe()
```

Out[26]:

	AMT_ANNUITY	AMT_GOODS_PRICE
count	307499.000000	3.072330e+05
mean	27108.573909	5.383962e+05
std	14493.737315	3.694465e+05
min	1615.500000	4.050000e+04
25%	16524.000000	2.385000e+05
50%	24903.000000	4.500000e+05
75%	34596.000000	6.795000e+05
max	258025.500000	4.050000e+06

In [27]:
```
application_data_filtered[fill_null_values].head()
```

Out[27]:

	AMT_ANNUITY	AMT_GOODS_PRICE
0	24700.5	351000.0
1	35698.5	1129500.0
2	6750.0	135000.0
3	29686.5	297000.0
4	21865.5	513000.0

In [28]:
```
application_data_filtered['AMT_GOODS_PRICE'].median()
```

Out[28]:
```
450000.0
```

In [29]:
```
application_data_filtered['AMT_ANNUITY'].fillna(27108.5,inplace=True)
```

In [30]:
```
application_data_filtered['AMT_GOODS_PRICE'].fillna(450000,inplace=True)
```

In [31]:
```
application_data_filtered.isnull().sum().sort_values(ascending=True)
```

Out[31]:
```
SK_ID_CURR                       0
DAYS_LAST_PHONE_CHANGE           0
ORGANIZATION_TYPE                0
CNT_FAM_MEMBERS                  0
DAYS_ID_PUBLISH                  0
DAYS_REGISTRATION                0
DAYS_EMPLOYED                    0
REGION_POPULATION_RELATIVE       0
NAME_HOUSING_TYPE                0
NAME_FAMILY_STATUS               0
NAME_EDUCATION_TYPE              0
NAME_INCOME_TYPE                 0
DAYS_BIRTH                       0
AMT_GOODS_PRICE                  0
TARGET                           0
CODE_GENDER                      0
FLAG_OWN_CAR                     0
FLAG_OWN_REALTY                  0
NAME_CONTRACT_TYPE               0
AMT_INCOME_TOTAL                 0
AMT_CREDIT                       0
AMT_ANNUITY                      0
CNT_CHILDREN                     0
OBS_30_CNT_SOCIAL_CIRCLE      1021
DEF_30_CNT_SOCIAL_CIRCLE      1021
OBS_60_CNT_SOCIAL_CIRCLE      1021
DEF_60_CNT_SOCIAL_CIRCLE      1021
NAME_TYPE_SUITE               1292
AMT_REQ_CREDIT_BUREAU_QRT    41519
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_YEAR   41519
OCCUPATION_TYPE              96391
dtype: int64
```

In [32]:
```
fill_null_values=['DEF_30_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','OBS_60_CNT_
                  'OBS_30_CNT_SOCIAL_CIRCLE','NAME_TYPE_SUITE','AMT_REQ_CREDIT_BUREA
                  'AMT_REQ_CREDIT_BUREAU_HOUR','AMT_REQ_CREDIT_BUREAU_WEEK','AMT_REQ
                  'AMT_REQ_CREDIT_BUREAU_QRT','OCCUPATION_TYPE']
application_data_filtered[fill_null_values].describe()
```

Out[32]:

	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	OBS_3
count	306490.000000	306490.000000	306490.000000	
mean	0.143421	0.100049	1.405292	
std	0.446698	0.362291	2.379803	
min	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	
75%	0.000000	0.000000	2.000000	
max	34.000000	24.000000	344.000000	

In [33]:
```
application_data_filtered[fill_null_values].head()
```

Out[33]:

	DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CN
0	2.0	2.0	2.0	
1	0.0	0.0	1.0	
2	0.0	0.0	0.0	
3	0.0	0.0	2.0	
4	0.0	0.0	0.0	

In [34]:
```
application_data_filtered['DEF_30_CNT_SOCIAL_CIRCLE'].fillna(0,inplace=True)
application_data_filtered['DEF_60_CNT_SOCIAL_CIRCLE'].fillna(0,inplace=True)
application_data_filtered['OBS_60_CNT_SOCIAL_CIRCLE'].fillna(1,inplace=True)
application_data_filtered['OBS_30_CNT_SOCIAL_CIRCLE'].fillna(1,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_MON'].fillna(0,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_DAY'].fillna(0,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_HOUR'].fillna(0,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_WEEK'].fillna(0,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_YEAR'].fillna(1,inplace=True)
application_data_filtered['AMT_REQ_CREDIT_BUREAU_QRT'].fillna(0,inplace=True)
```

In [35]:
```
application_data_filtered.groupby('NAME_TYPE_SUITE').size()
```

Out[35]:
```
NAME_TYPE_SUITE
Children             3267
Family              40149
Group of people       271
Other_A               866
Other_B              1770
Spouse, partner     11370
Unaccompanied      248526
dtype: int64
```

In [36]:
```
application_data_filtered['NAME_TYPE_SUITE'].fillna('Family',inplace=True)
```

In [37]:
```
application_data_filtered.groupby('OCCUPATION_TYPE').size()
```

Out[37]:
```
OCCUPATION_TYPE
Accountants              9813
Cleaning staff           4653
Cooking staff            5946
Core staff              27570
Drivers                 18603
HR staff                  563
High skill tech staff   11380
IT staff                  526
Laborers                55186
Low-skill Laborers       2093
Managers                21371
Medicine staff           8537
Private service staff    2652
Realty agents             751
Sales staff             32102
Secretaries              1305
Security staff           6721
Waiters/barmen staff     1348
dtype: int64
```

In [38]:
```
application_data_filtered['OCCUPATION_TYPE'].fillna('Laborers',inplace=True)
```

In [ ]:

In [39]:
```
application_data_filtered.isnull().sum().sort_values(ascending=True)
```

Out[39]:
```
SK_ID_CURR                    0
DAYS_REGISTRATION             0
DAYS_ID_PUBLISH               0
OCCUPATION_TYPE               0
CNT_FAM_MEMBERS               0
ORGANIZATION_TYPE             0
OBS_30_CNT_SOCIAL_CIRCLE      0
DAYS_EMPLOYED                 0
DEF_30_CNT_SOCIAL_CIRCLE      0
DEF_60_CNT_SOCIAL_CIRCLE      0
DAYS_LAST_PHONE_CHANGE        0
AMT_REQ_CREDIT_BUREAU_HOUR    0
AMT_REQ_CREDIT_BUREAU_DAY     0
AMT_REQ_CREDIT_BUREAU_WEEK    0
AMT_REQ_CREDIT_BUREAU_MON     0
OBS_60_CNT_SOCIAL_CIRCLE      0
AMT_REQ_CREDIT_BUREAU_QRT     0
DAYS_BIRTH                    0
NAME_HOUSING_TYPE             0
TARGET                        0
NAME_CONTRACT_TYPE            0
CODE_GENDER                   0
FLAG_OWN_CAR                  0
FLAG_OWN_REALTY               0
CNT_CHILDREN                  0
REGION_POPULATION_RELATIVE    0
AMT_INCOME_TOTAL              0
AMT_ANNUITY                   0
AMT_GOODS_PRICE               0
NAME_TYPE_SUITE               0
NAME_INCOME_TYPE              0
NAME_EDUCATION_TYPE           0
NAME_FAMILY_STATUS            0
AMT_CREDIT                    0
AMT_REQ_CREDIT_BUREAU_YEAR    0
dtype: int64
```

In [40]:
```
application_data_filtered.to_csv("C:/Users/Hello/Downloads/application_data_filtered
```

Previous_application data

In [41]:
```
previous_application =pd.read_csv("C:/Users/Hello/Downloads/previous_application.csv
```

In [42]:
```
previous_application.shape
```

Out[42]:
```
(1670214, 37)
```

In [43]:
```
previous_application.head()
```

Out[43]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CRE
0	2030495	271877	Consumer loans	1730.430	17145.0	171
1	2802425	108129	Cash loans	25188.615	607500.0	6796
2	2523466	122040	Cash loans	15060.735	112500.0	1364
3	2819243	176158	Cash loans	47041.335	450000.0	4707
4	1784265	202054	Cash loans	31924.395	337500.0	4040

5 rows × 37 columns

In [44]:
```
null_values2=(previous_application.isnull().sum()/1670214)*100
print(null_values2)
```

```
SK_ID_PREV                    0.000000
SK_ID_CURR                    0.000000
NAME_CONTRACT_TYPE            0.000000
AMT_ANNUITY                  22.286665
AMT_APPLICATION               0.000000
AMT_CREDIT                    0.000060
AMT_DOWN_PAYMENT             53.636480
AMT_GOODS_PRICE              23.081773
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            53.636480
RATE_INTEREST_PRIMARY        99.643698
RATE_INTEREST_PRIVILEGED     99.643698
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              49.119754
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                  22.286366
NAME_YIELD_GROUP              0.000000
PRODUCT_COMBINATION           0.020716
DAYS_FIRST_DRAWING           40.298129
DAYS_FIRST_DUE               40.298129
DAYS_LAST_DUE_1ST_VERSION    40.298129
DAYS_LAST_DUE                40.298129
DAYS_TERMINATION             40.298129
NFLAG_INSURED_ON_APPROVAL    40.298129
dtype: float64
```

In [45]:
```
drop_column=null_values2[null_values2>40].index
```

In [46]:
```
previous_application=previous_application.drop(columns=drop_column)
```

In [47]:
```
previous_application.shape
```

Out[47]: (1670214, 26)

In [48]:
```
previous_application.head()
```

Out[48]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CRE
0	2030495	271877	Consumer loans	1730.430	17145.0	171
1	2802425	108129	Cash loans	25188.615	607500.0	6796
2	2523466	122040	Cash loans	15060.735	112500.0	1364
3	2819243	176158	Cash loans	47041.335	450000.0	4707
4	1784265	202054	Cash loans	31924.395	337500.0	4040

5 rows × 26 columns

In [49]:
```
previous_application.columns
```

Out[49]:
```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE',
       'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START',
       'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY',
       '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')
```

In [50]:
```
drop_column=['WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START',
       'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY','NAME_GOODS_CATEGORY'
            ]
previous_application=previous_application.drop(columns=drop_column)
```

In [51]:
```
previous_application.head()
```

Out[51]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CRE
0	2030495	271877	Consumer loans	1730.430	17145.0	171
1	2802425	108129	Cash loans	25188.615	607500.0	6796
2	2523466	122040	Cash loans	15060.735	112500.0	1364
3	2819243	176158	Cash loans	47041.335	450000.0	4707
4	1784265	202054	Cash loans	31924.395	337500.0	4040

In [52]:
```
previous_application.isnull().sum().sort_values(ascending=True)
```

Out[52]:
```
SK_ID_PREV                   0
NAME_SELLER_INDUSTRY         0
CHANNEL_TYPE                 0
NAME_PRODUCT_TYPE            0
NAME_CLIENT_TYPE             0
CODE_REJECT_REASON           0
NAME_PAYMENT_TYPE            0
NAME_YIELD_GROUP             0
DAYS_DECISION                0
NAME_CASH_LOAN_PURPOSE       0
AMT_APPLICATION              0
NAME_CONTRACT_TYPE           0
SK_ID_CURR                   0
NAME_CONTRACT_STATUS         0
AMT_CREDIT                   1
PRODUCT_COMBINATION        346
CNT_PAYMENT             372230
AMT_ANNUITY             372235
AMT_GOODS_PRICE         385515
dtype: int64
```

In [53]:
```
fill_null=['AMT_CREDIT','PRODUCT_COMBINATION','CNT_PAYMENT','AMT_ANNUITY','AMT_GOODS
previous_application[fill_null].describe()
```

Out[53]:

	AMT_CREDIT	CNT_PAYMENT	AMT_ANNUITY	AMT_GOODS_PRICE
count	1.670213e+06	1.297984e+06	1.297979e+06	1.284699e+06
mean	1.961140e+05	1.605408e+01	1.595512e+04	2.278473e+05
std	3.185746e+05	1.456729e+01	1.478214e+04	3.153966e+05
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.416050e+04	6.000000e+00	6.321780e+03	5.084100e+04
50%	8.054100e+04	1.200000e+01	1.125000e+04	1.123200e+05
75%	2.164185e+05	2.400000e+01	2.065842e+04	2.340000e+05
max	6.905160e+06	8.400000e+01	4.180581e+05	6.905160e+06

In [54]:
```
previous_application['AMT_GOODS_PRICE'].median()
```

Out[54]:
```
112320.0
```

In [55]:
```
previous_application['AMT_CREDIT'].median()
```

Out[55]:
```
80541.0
```

In [56]:
```
previous_application['AMT_ANNUITY'].median()
```

Out[56]:
```
11250.0
```

In [57]:
```
previous_application[fill_null].head()
```

Out[57]:

	AMT_CREDIT	PRODUCT_COMBINATION	CNT_PAYMENT	AMT_ANNUITY	AMT_GOODS_PRICE
0	17145.0	POS mobile with interest	12.0	1730.430	17145.0
1	679671.0	Cash X-Sell: low	36.0	25188.615	607500.0
2	136444.5	Cash X-Sell: high	12.0	15060.735	112500.0
3	470790.0	Cash X-Sell: middle	12.0	47041.335	450000.0
4	404055.0	Cash Street: high	24.0	31924.395	337500.0

In [58]:
```
previous_application['AMT_GOODS_PRICE'].fillna(11232,inplace=True)
previous_application['CNT_PAYMENT'].fillna(16,inplace=True)
previous_application['AMT_CREDIT'].fillna(80541,inplace=True)
previous_application['AMT_ANNUITY'].fillna(11250,inplace=True)
```

In [59]:
```
previous_application.groupby('PRODUCT_COMBINATION').size().sort_values()
```

Out[59]:
```
PRODUCT_COMBINATION
POS others without interest         2555
POS industry without interest      12602
POS other with interest            23879
POS mobile without interest        24082
Cash Street: low                   33834
Cash Street: middle                34658
Cash X-Sell: high                  59301
Cash Street: high                  59639
Card X-Sell                        80582
POS household without interest     82908
POS industry with interest         98833
Card Street                       112582
Cash X-Sell: low                  130248
Cash X-Sell: middle               143883
POS mobile with interest          220670
POS household with interest       263622
Cash                              285990
dtype: int64
```

In [60]:
```
previous_application['PRODUCT_COMBINATION'].fillna('Cash',inplace=True)
```

In [61]:
```
previous_application.isnull().sum()
```

Out[61]:
```
SK_ID_PREV                0
SK_ID_CURR                0
NAME_CONTRACT_TYPE        0
AMT_ANNUITY               0
AMT_APPLICATION           0
AMT_CREDIT                0
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_PRODUCT_TYPE         0
CHANNEL_TYPE              0
NAME_SELLER_INDUSTRY      0
CNT_PAYMENT               0
NAME_YIELD_GROUP          0
PRODUCT_COMBINATION       0
dtype: int64
```

In [62]:
```
previous_application.head()
```

Out[62]:

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CRE
0	2030495	271877	Consumer loans	1730.430	17145.0	171
1	2802425	108129	Cash loans	25188.615	607500.0	6796
2	2523466	122040	Cash loans	15060.735	112500.0	1364
3	2819243	176158	Cash loans	47041.335	450000.0	4707
4	1784265	202054	Cash loans	31924.395	337500.0	4040

In [63]:
```
convert_to_positive=['DAYS_DECISION']
previous_application[convert_to_positive]=previous_application[convert_to_positive].
```

In [64]:
```
previous_application.to_csv("C:/Users/Hello/Downloads/previous_application_filtered.
```

Merge both data sets

In [65]:
```
final_data=pd.merge(application_data_filtered,previous_application,how='inner',on='S
```

In [66]:
```
final_data.head()
```

Out[66]:

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

5 rows × 53 columns

In [67]:
```
final_data.columns
```

Out[67]:
```
Index(['SK_ID_CURR', 'TARGET', 'NAME_CONTRACT_TYPE_x', 'CODE_GENDER',
       'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL',
       'AMT_CREDIT_x', 'AMT_ANNUITY_x', 'AMT_GOODS_PRICE_x', 'NAME_TYPE_SUITE',
       'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE', 'NAME_FAMILY_STATUS',
       'NAME_HOUSING_TYPE', 'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH',
       'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'ORGANIZATION_TYPE',
       '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', 'SK_ID_PREV', 'NAME_CONTRACT_TYPE_y',
       'AMT_ANNUITY_y', 'AMT_APPLICATION', 'AMT_CREDIT_y', 'AMT_GOODS_PRICE_y',
       'NAME_CASH_LOAN_PURPOSE', 'NAME_CONTRACT_STATUS', 'DAYS_DECISION',
       'NAME_PAYMENT_TYPE', 'CODE_REJECT_REASON', 'NAME_CLIENT_TYPE',
       'NAME_PRODUCT_TYPE', 'CHANNEL_TYPE', 'NAME_SELLER_INDUSTRY',
       'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION'],
      dtype='object')
```

In [68]:
```
final_data.to_csv("C:/Users/Hello/Downloads/final_data.csv",index=False)
```

Ratio Imbalance

In [69]:
```
final_data['TARGET'].value_counts()
difficulty=len(final_data[final_data['TARGET']==1])
no_difficulty=len(final_data[final_data['TARGET']==0])
imbalance_ratio=difficulty/no_difficulty
print(imbalance_ratio)
```

```
0.09475421286863811
```

Correlation

In [70]:
```
columns=['CNT_CHILDREN','CNT_FAM_MEMBERS','CNT_PAYMENT','AMT_ANNUITY_y','AMT_APPLICA
correlation=final_data[columns].corr()
```

In [71]:
```
print(correlation)
```

	CNT_CHILDREN	CNT_FAM_MEMBERS	CNT_PAYMENT	AMT_ANNUITY_y	\
CNT_CHILDREN	1.000000	0.879224	-0.049161	-0.032608	
CNT_FAM_MEMBERS	0.879224	1.000000	-0.030302	-0.004577	
CNT_PAYMENT	-0.049161	-0.030302	1.000000	0.394190	
AMT_ANNUITY_y	-0.032608	-0.004577	0.394190	1.000000	
AMT_APPLICATION	-0.034168	-0.005809	0.650920	0.806458	
AMT_CREDIT_y	-0.034860	-0.005233	0.641663	0.812972	
AMT_GOODS_PRICE_y	-0.034768	-0.006154	0.652509	0.808430	
DAYS_BIRTH	-0.363034	-0.326241	0.110095	0.071977	
DAYS_EMPLOYED	-0.249912	-0.250514	0.065006	-0.006126	
AMT_INCOME_TOTAL	0.011661	0.014119	0.017729	0.099077	

	AMT_APPLICATION	AMT_CREDIT_y	AMT_GOODS_PRICE_y	\
CNT_CHILDREN	-0.034168	-0.034860	-0.034768	
CNT_FAM_MEMBERS	-0.005809	-0.005233	-0.006154	
CNT_PAYMENT	0.650920	0.641663	0.652509	
AMT_ANNUITY_y	0.806458	0.812972	0.808430	
AMT_APPLICATION	1.000000	0.975683	0.999768	
AMT_CREDIT_y	0.975683	1.000000	0.976391	
AMT_GOODS_PRICE_y	0.999768	0.976391	1.000000	
DAYS_BIRTH	0.079786	0.078078	0.080834	
DAYS_EMPLOYED	0.010754	0.005046	0.011140	
AMT_INCOME_TOTAL	0.071491	0.070651	0.072037	

	DAYS_BIRTH	DAYS_EMPLOYED	AMT_INCOME_TOTAL
CNT_CHILDREN	-0.363034	-0.249912	0.011661
CNT_FAM_MEMBERS	-0.326241	-0.250514	0.014119
CNT_PAYMENT	0.110095	0.065006	0.017729
AMT_ANNUITY_y	0.071977	-0.006126	0.099077
AMT_APPLICATION	0.079786	0.010754	0.071491
AMT_CREDIT_y	0.078078	0.005046	0.070651
AMT_GOODS_PRICE_y	0.080834	0.011140	0.072037
DAYS_BIRTH	1.000000	0.632509	-0.025717
DAYS_EMPLOYED	0.632509	1.000000	-0.067046
AMT_INCOME_TOTAL	-0.025717	-0.067046	1.000000

In [ ]: