Bank Loan Analysis

Name: NIKHIL ANAND
Email: nikhil3313@gmail.com
Task: 6 – Bank Loan Analyis
Tech Stack: Python

Overall Approach of the Analysis: The problem statement is to analyse patterns in loan application data to identify factors that indicate if a client is likely to have difficulty paying their instalments. This analysis will help in minimizing the risk of approving loans to clients who are likely to default. The approach involves conducting exploratory data analysis (EDA) on the provided datasets to gain insights into the relationship between client attributes, loan attributes, and the tendency of default.

Data Understanding: The dataset includes three files: "application_data.csv," "previous_application.csv," and "columns_description.csv." The "application_data.csv" file contains information about clients at the time of loan application. The "previous_application.csv" file provides data on the client's previous loan applications. The "columns_description.csv" file serves as a data dictionary, explaining the meaning of the variables.

Missing Data Handling: To deal with missing data, we can follow the approach of either removing columns with a significant number of missing values or replacing missing values with appropriate substitutes. In this case study, we will choose the appropriate method for handling missing data based on the nature of the variable and the analysis requirements.

Outlier Identification: To identify outliers in the dataset, we can use statistical methods such as box plots and interquartile range (IQR) analysis. Outliers can be identified as data points that fall significantly above or below the expected range. However, for this exercise, we are not required to remove any data points, so the identification of outliers will serve the purpose of understanding the data distribution.

Data Imbalance Check: To check for data imbalance, we need to examine the distribution of the target variable. If there is a significant difference in the number of instances between different categories of the target variable, it indicates data imbalance. We can calculate the ratio of data imbalance by dividing the number of instances in the minority class by the number of instances in the majority class.

Univariate Analysis: In the univariate analysis, we will explore each variable individually to understand its distribution and characteristics. This will involve visualizations such as histograms, bar plots, or box plots.

Top Correlations: To find the top correlations, we will segment the dataset based on the target variable (clients with payment difficulties and all other cases). Then, for each segmented dataset, we will calculate the correlations between variables. The top correlations will indicate the relationships between variables within each segment, providing insights into the factors influencing payment difficulties.

Visualization and Summarization: Visualizations will be used throughout the analysis to present the numerical and categorical variables effectively. The most important results, including significant variables, relationships, and patterns, will be summarized in the final presentation to provide meaningful insights into minimizing the risk of loan default.

Here is a report on my findings in the dataset given for the 6th project using python-

Dataset info:

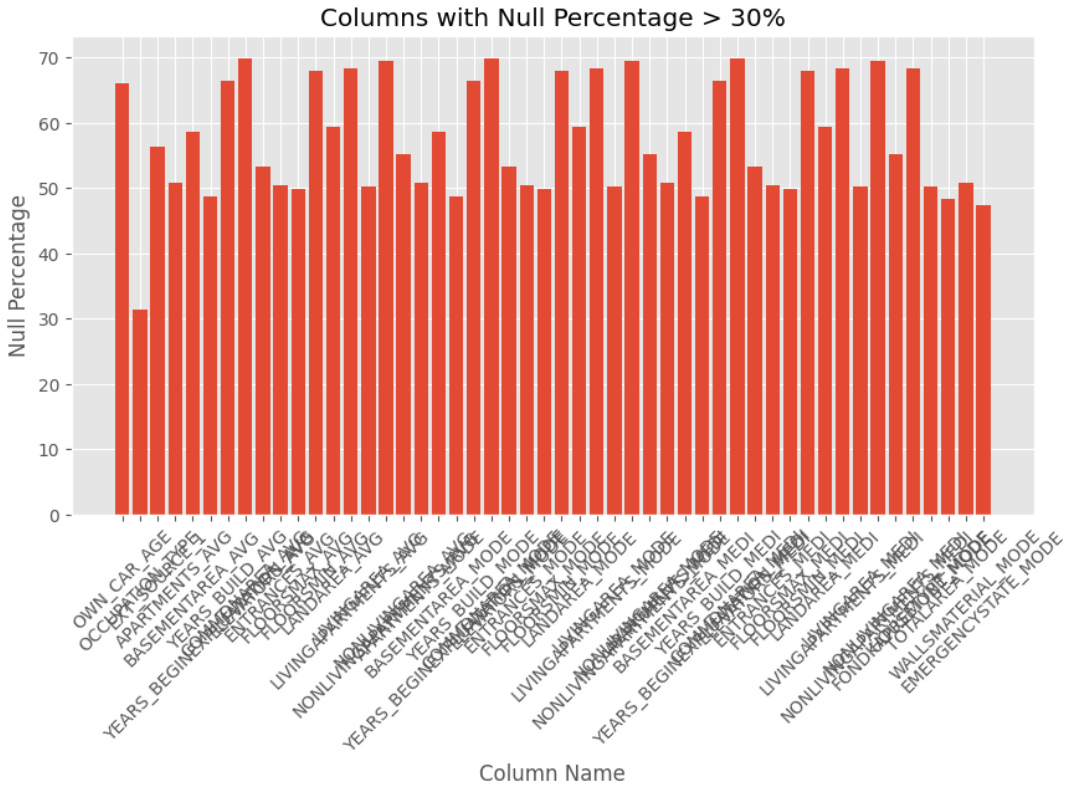
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
 #    Column                       Non-Null Count   Dtype
---   ------                       --------------   -----
 0    SK_ID_CURR                   307511 non-null  int64
 1    TARGET                       307511 non-null  int64
 2    NAME_CONTRACT_TYPE           307511 non-null  object
 3    CODE_GENDER                  307511 non-null  object
 4    FLAG_OWN_CAR                 307511 non-null  object
 5    FLAG_OWN_REALTY              307511 non-null  object
 6    CNT_CHILDREN                 307511 non-null  int64
 7    AMT_INCOME_TOTAL             307511 non-null  float64
 8    AMT_CREDIT                   307511 non-null  float64
 9    AMT_ANNUITY                  307499 non-null  float64
 10   AMT_GOODS_PRICE              307233 non-null  float64
 11   NAME_TYPE_SUITE              306219 non-null  object
 12   NAME_INCOME_TYPE             307511 non-null  object
 13   NAME_EDUCATION_TYPE          307511 non-null  object
 14   NAME_FAMILY_STATUS           307511 non-null  object
 15   NAME_HOUSING_TYPE            307511 non-null  object
 16   REGION_POPULATION_RELATIVE   307511 non-null  float64
 17   DAYS_BIRTH                   307511 non-null  int64
 18   DAYS_EMPLOYED                307511 non-null  int64
 19   DAYS_REGISTRATION            307511 non-null  float64
...
 120  AMT_REQ_CREDIT_BUREAU_QRT    265992 non-null  float64
 121  AMT_REQ_CREDIT_BUREAU_YEAR   265992 non-null  float64
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

Columns with Missing values in Dataset :

```
Index(['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'OWN_CAR_AGE',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_1', 'EXT_SOURCE_2',
       'EXT_SOURCE_3', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG',
       'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG',
       'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG',
       'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG',
       'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE',
       'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE',
       'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE',
       'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE',
       'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE',
       'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI',
       'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI',
       'ENTRANCES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI',
       'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI',
       'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
       'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE',
       'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
       'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
       'DAYS_LAST_PHONE_CHANGE', 'AMT_REQ_CREDIT_BUREAU_HOUR',
       'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
       'AMT_REQ_CREDIT_BUREAU_YEAR'],
      dtype='object')
```

// There are 67 columns having one or more NULL values

Columns with NULL values >30% in Dataset :



// Dropping these columns from dataset

Remaining columns with missing values in Dataset :

	column_name	null_count	null_percentage
43	EXT_SOURCE_3	60965	19.83
116	AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.50
117	AMT_REQ_CREDIT_BUREAU_DAY	41519	13.50
118	AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.50
119	AMT_REQ_CREDIT_BUREAU_MON	41519	13.50
120	AMT_REQ_CREDIT_BUREAU_QRT	41519	13.50
121	AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.50
11	NAME_TYPE_SUITE	1292	0.42
91	OBS_30_CNT_SOCIAL_CIRCLE	1021	0.33
92	DEF_30_CNT_SOCIAL_CIRCLE	1021	0.33
93	OBS_60_CNT_SOCIAL_CIRCLE	1021	0.33
94	DEF_60_CNT_SOCIAL_CIRCLE	1021	0.33
42	EXT_SOURCE_2	660	0.21
10	AMT_GOODS_PRICE	278	0.09

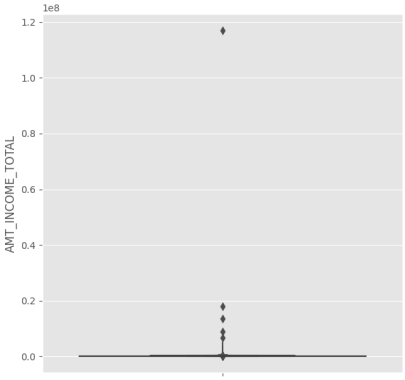
//Replacing them with median/mode for analysis

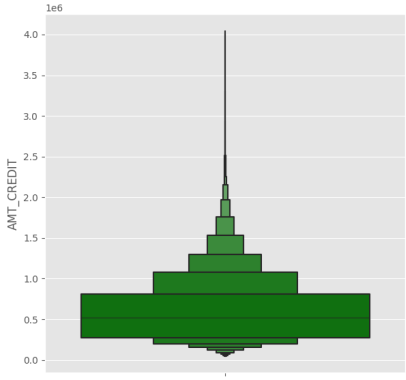
Correcting date data in Dataset :

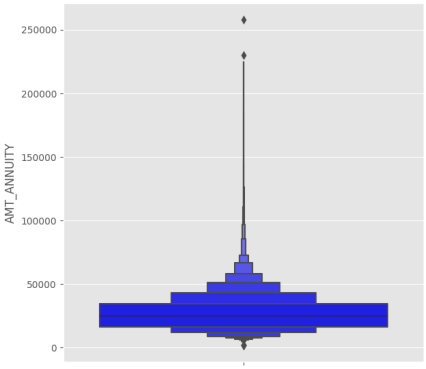
	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	DAYS_LAST_PHONE_CHANGE
0	9461	637	3648.00	2120	1134.00
1	16765	1188	1186.00	291	828.00
2	19046	225	4260.00	2531	815.00
3	19005	3039	9833.00	2437	617.00
4	19932	3038	4311.00	3458	1106.00
...	...	...	...	...	...
307506	9327	236	8456.00	1982	273.00
307507	20775	365243	4388.00	4090	0.00
307508	14966	7921	6737.00	5150	1909.00
307509	11961	4786	2562.00	931	322.00
307510	16856	1262	5128.00	410	787.00

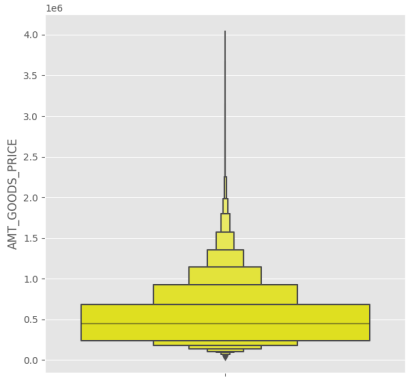
307511 rows × 5 columns

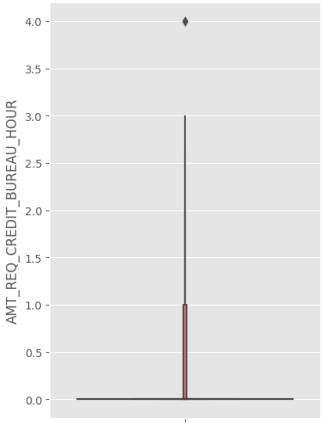
Identifying Outliers (through boxplots) and removing them from dataset :

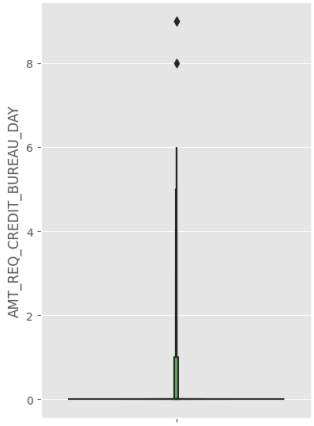


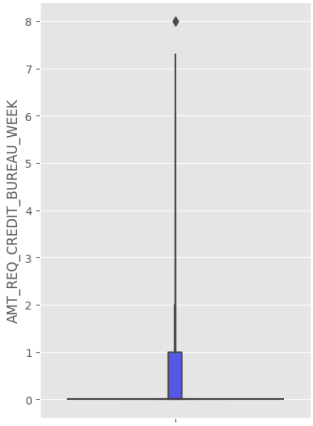


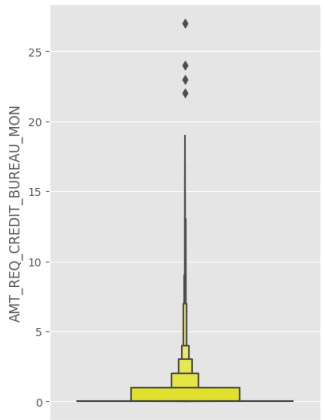


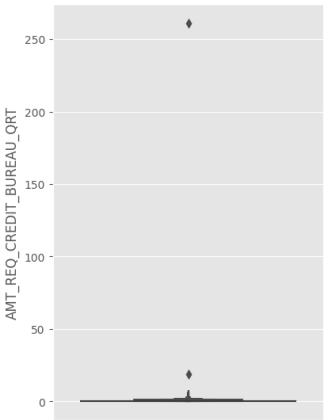


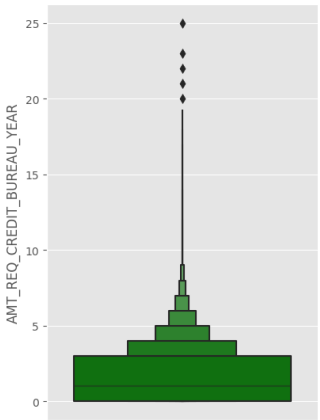




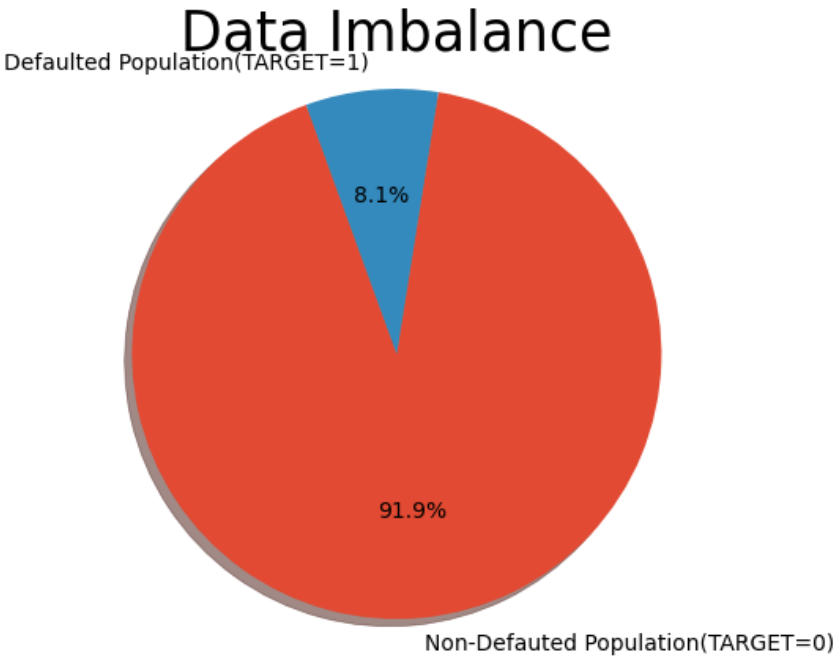




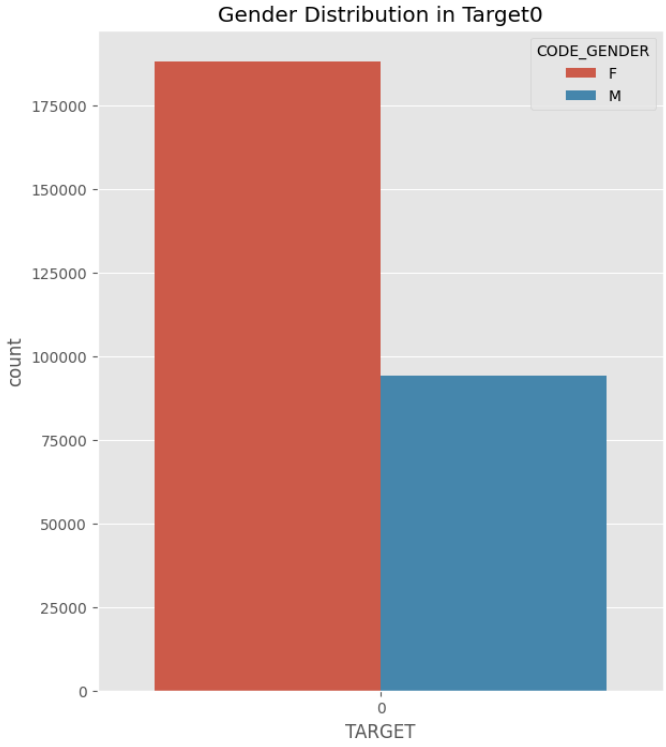


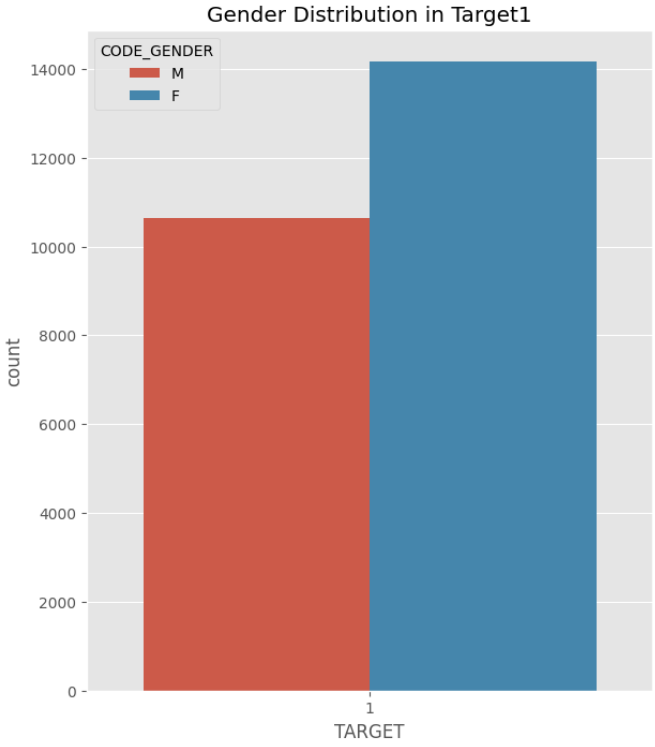


Data Imbalance = 11.390328430384848



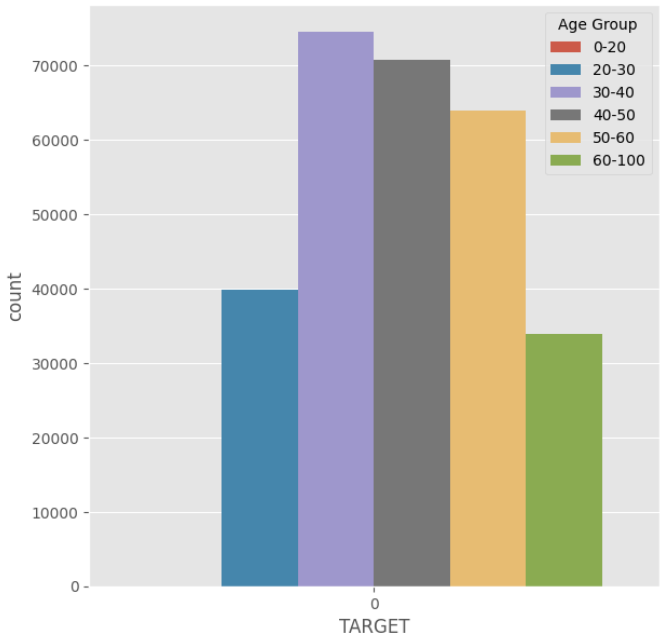
Univariate Analysis (on gender and age) :

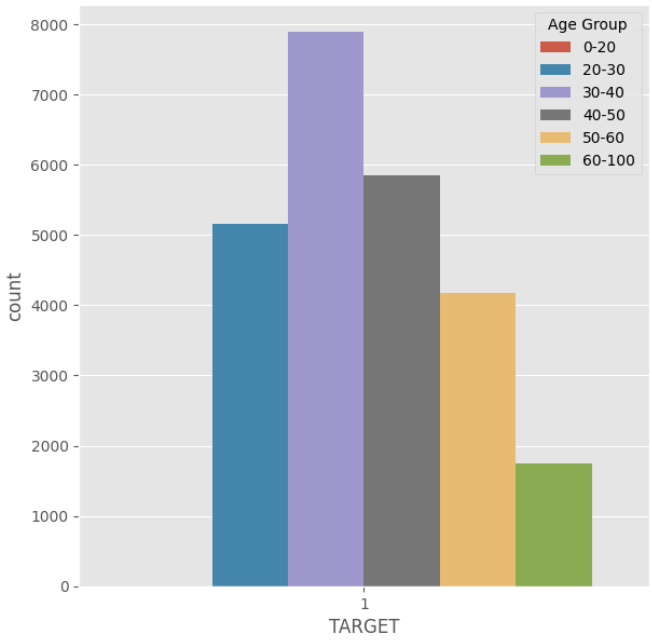




Insights -

- It seems like Female clients applied higher than male clients for loan 66.6%
- Female clients are non-defaulters while 33.4% male clients are non-defaulters.
- 57% Female clients are defaulters while 42% male clients are defaulters.





Insights -

- Middle Age(30-50) the group seems to applied higher than any other age group.
- Senior Citizens(60-100) and Very young(19-25) age group facing paying difficulties less as compared to other age groups.

// Analysing Categorical Values:

Insights-

- Most of clients who have applied for loan belong to Working Income Type.
- Most of clients with Secondary/Secondary Special education type have applied for the loan.
- Most of clients who are have applied for loan are married.
- Most of the Clients who have applied for the loan have their own house/apartment.
- Female applied for loan more than males.
- Most clients who applied most for loan have work experience between 0-5 years have.

Top 10 Correlation :

// in default category

	VAR1	VAR2	CORRELATION	CORR_ABS
746	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
202	AMT_GOODS_PRICE	AMT_CREDIT	0.99	0.99
332	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88	0.88
475	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86	0.86
780	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86	0.86
577	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83	0.83
203	AMT_GOODS_PRICE	AMT_ANNUITY	0.78	0.78
169	AMT_ANNUITY	AMT_CREDIT	0.77	0.77
441	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.45	0.45
543	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.44	0.44

// in non-default category

	VAR1	VAR2	CORRELATION	CORR_ABS
746	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	1.00
202	AMT_GOODS_PRICE	AMT_CREDIT	0.98	0.98
332	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88	0.88
780	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.87	0.87
475	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.85	0.85
577	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.78	0.78
203	AMT_GOODS_PRICE	AMT_ANNUITY	0.75	0.75
169	AMT_ANNUITY	AMT_CREDIT	0.75	0.75
441	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.50	0.50
543	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.47	0.47

Insight-

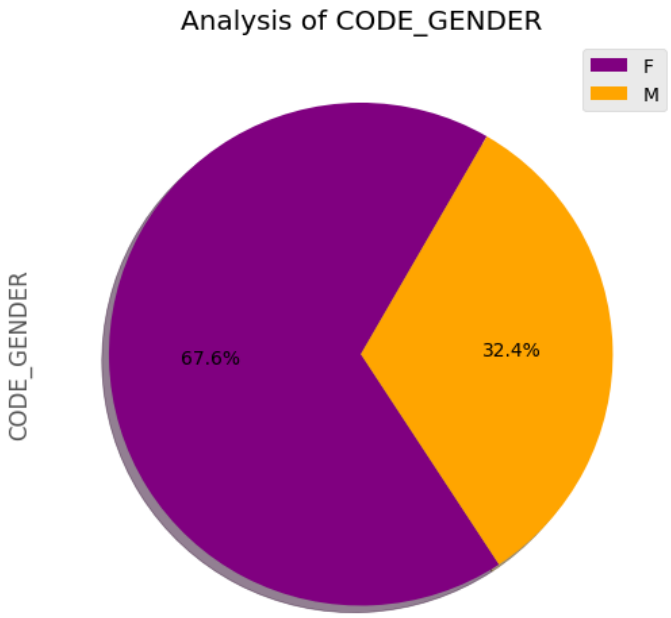
- Top 10 correlations are almost at the same level in both the Default and Non Default population

Previous dataset :

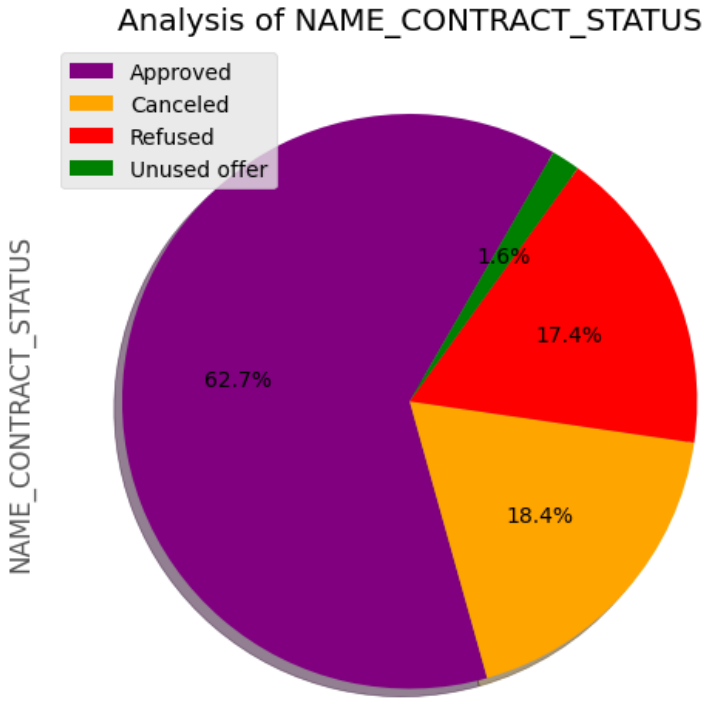
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
...
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

// Dropping columns not needed and dealing with missing/null values
// Then merging previous dataset with current dataset

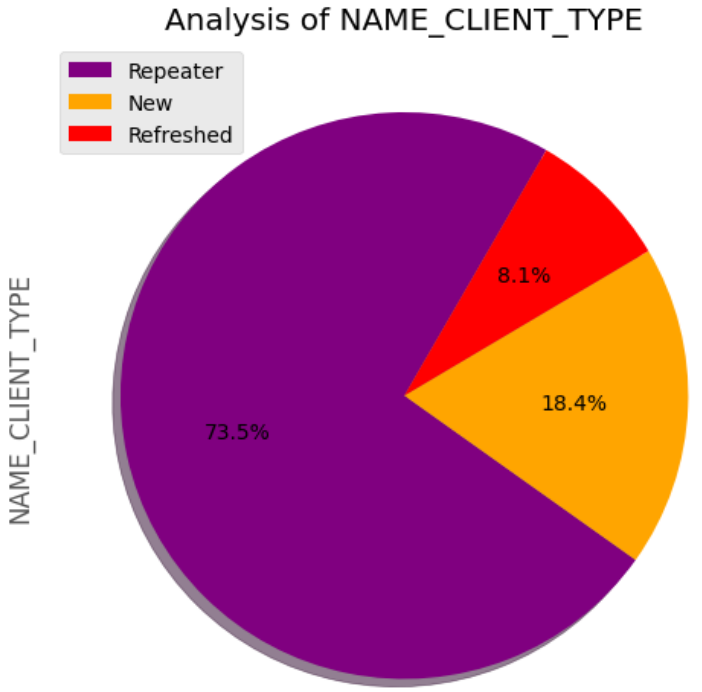
Insights :



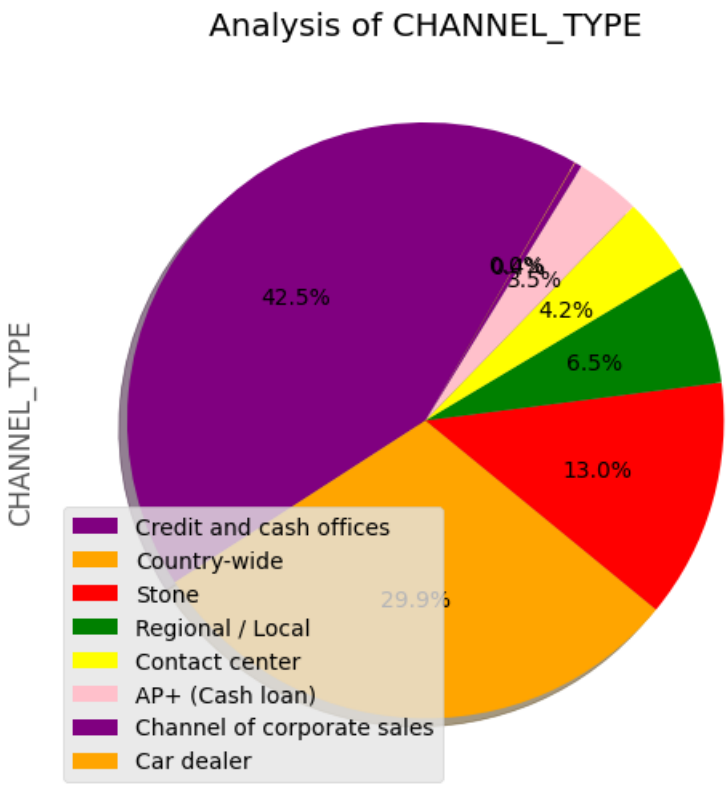
- *Approved percentage of loans provided to females is more as compared to refused percentage.*



- *Approved loan status is the highest among all loan applications*
- *Cancelled loan status is the second highest among all loan applications*



- *Repeater client type is the highest among all loan applications*
- *New client type is the second highest among all loan applications*




```
# Country-wide Channel type is the highest among all loan applications
# Credit and cash offices is the second highest Channel Type among all loan applications
```

←----------→

CODE FOR ANALYSIS -

ppspnlzcw

June 2, 2023

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from plotly.subplots import make_subplots
import plotly.graph_objects as go
plt.style.use('ggplot')
from plotly.subplots import make_subplots
import plotly.io as pio
pio.renderers.default = 'iframe'


pd.set_option('display.max.rows',130)
pd.set_option('display.max.columns',130)
pd.set_option('float_format', '{:.2f}'.format)
```

[2]:
```
# Importing dataset 1
df = pd.read_csv("application_data.csv")
```

[3]:
```
# Checking few records from the dataframe
df.head()
```

[3]:

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	100002	1	Cash loans	M	N	Y	0	202500.00	406597.50	24700.50
1	100003	0	Cash loans	F	N	N	0	270000.00	1293502.50	35698.50
2	100004	0	Revolving loans	M	Y	Y	0	67500.00	135000.00	6750.00
3	100006	0	Cash loans	F	N	Y	0	135000.00	312682.50	29686.50
4	100007	0	Cash loans	M	N	Y	0	121500.00	513000.00	21865.50

```
   AMT_GOODS_PRICE NAME_TYPE_SUITE NAME_INCOME_TYPE  \
0        351000.00   Unaccompanied          Working
1       1129500.00          Family    State servant
2        135000.00   Unaccompanied          Working
3        297000.00   Unaccompanied          Working
4        513000.00   Unaccompanied          Working

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

   REGION_POPULATION_RELATIVE  DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  \
0                        0.02       -9461           -637           -3648.00
1                        0.00      -16765          -1188           -1186.00
2                        0.01      -19046           -225           -4260.00
3                        0.01      -19005          -3039           -9833.00
4                        0.03      -19932          -3038           -4311.00

   DAYS_ID_PUBLISH  OWN_CAR_AGE  FLAG_MOBIL  FLAG_EMP_PHONE  FLAG_WORK_PHONE  \
0            -2120          NaN           1               1                0
1             -291          NaN           1               1                0
2            -2531        26.00           1               1                1
3            -2437          NaN           1               1                0
4            -3458          NaN           1               1                0

   FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS  \
0                 1           1           0       Laborers             1.00
1                 1           1           0     Core staff             2.00
2                 1           1           0       Laborers             1.00
3                 1           0           0       Laborers             2.00
4                 1           0           0     Core staff             1.00

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

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

```
4                  THURSDAY                        11

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

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

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

   EXT_SOURCE_1  EXT_SOURCE_2  EXT_SOURCE_3  APARTMENTS_AVG  BASEMENTAREA_AVG  \
0          0.08          0.26          0.14            0.02              0.04
1          0.31          0.62           NaN            0.10              0.05
2           NaN          0.56          0.73             NaN               NaN
3           NaN          0.65           NaN             NaN               NaN
4           NaN          0.32           NaN             NaN               NaN

   YEARS_BEGINEXPLUATATION_AVG  YEARS_BUILD_AVG  COMMONAREA_AVG  \
0                         0.97             0.62            0.01
1                         0.99             0.80            0.06
2                          NaN              NaN             NaN
3                          NaN              NaN             NaN
4                          NaN              NaN             NaN

   ELEVATORS_AVG  ENTRANCES_AVG  FLOORSMAX_AVG  FLOORSMIN_AVG  LANDAREA_AVG  \
0           0.00           0.07           0.08           0.12          0.04
1           0.08           0.03           0.29           0.33          0.01
2            NaN            NaN            NaN            NaN           NaN
3            NaN            NaN            NaN            NaN           NaN
4            NaN            NaN            NaN            NaN           NaN

   LIVINGAPARTMENTS_AVG  LIVINGAREA_AVG  NONLIVINGAPARTMENTS_AVG  \
0                  0.02            0.02                     0.00
1                  0.08            0.05                     0.00
```

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

   NONLIVINGAREA_AVG  APARTMENTS_MODE  BASEMENTAREA_MODE  \
0               0.00             0.03               0.04
1               0.01             0.09               0.05
2                NaN              NaN                NaN
3                NaN              NaN                NaN
4                NaN              NaN                NaN

   YEARS_BEGINEXPLUATATION_MODE  YEARS_BUILD_MODE  COMMONAREA_MODE  \
0                          0.97              0.63             0.01
1                          0.99              0.80             0.05
2                           NaN               NaN              NaN
3                           NaN               NaN              NaN
4                           NaN               NaN              NaN

   ELEVATORS_MODE  ENTRANCES_MODE  FLOORSMAX_MODE  FLOORSMIN_MODE  \
0            0.00            0.07            0.08            0.12
1            0.08            0.03            0.29            0.33
2             NaN             NaN             NaN             NaN
3             NaN             NaN             NaN             NaN
4             NaN             NaN             NaN             NaN

   LANDAREA_MODE  LIVINGAPARTMENTS_MODE  LIVINGAREA_MODE  \
0           0.04                   0.02             0.02
1           0.01                   0.08             0.06
2            NaN                    NaN              NaN
3            NaN                    NaN              NaN
4            NaN                    NaN              NaN

   NONLIVINGAPARTMENTS_MODE  NONLIVINGAREA_MODE  APARTMENTS_MEDI  \
0                      0.00                0.00             0.03
1                      0.00                0.00             0.10
2                       NaN                 NaN              NaN
3                       NaN                 NaN              NaN
4                       NaN                 NaN              NaN

   BASEMENTAREA_MEDI  YEARS_BEGINEXPLUATATION_MEDI  YEARS_BUILD_MEDI  \
0               0.04                          0.97              0.62
1               0.05                          0.99              0.80
2                NaN                           NaN               NaN
3                NaN                           NaN               NaN
4                NaN                           NaN               NaN

   COMMONAREA_MEDI  ELEVATORS_MEDI  ENTRANCES_MEDI  FLOORSMAX_MEDI  \
```

```
0             0.01            0.00            0.07            0.08
1             0.06            0.08            0.03            0.29
2              NaN             NaN             NaN             NaN
3              NaN             NaN             NaN             NaN
4              NaN             NaN             NaN             NaN

   FLOORSMIN_MEDI  LANDAREA_MEDI  LIVINGAPARTMENTS_MEDI  LIVINGAREA_MEDI  \
0            0.12           0.04                   0.02             0.02
1            0.33           0.01                   0.08             0.06
2             NaN            NaN                    NaN              NaN
3             NaN            NaN                    NaN              NaN
4             NaN            NaN                    NaN              NaN

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

   HOUSETYPE_MODE  TOTALAREA_MODE WALLSMATERIAL_MODE EMERGENCYSTATE_MODE  \
0  block of flats            0.01       Stone, brick                  No
1  block of flats            0.07              Block                  No
2             NaN             NaN                NaN                 NaN
3             NaN             NaN                NaN                 NaN
4             NaN             NaN                NaN                 NaN

   OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0                      2.00                      2.00
1                      1.00                      0.00
2                      0.00                      0.00
3                      2.00                      0.00
4                      0.00                      0.00

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
0                      2.00                      2.00                -1134.00
1                      1.00                      0.00                 -828.00
2                      0.00                      0.00                 -815.00
3                      2.00                      0.00                 -617.00
4                      0.00                      0.00                -1106.00

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

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

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

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

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

   AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
0                        0.00                       0.00
1                        0.00                       0.00
2                        0.00                       0.00
3                         NaN                        NaN
4                        0.00                       0.00

   AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                        0.00                       0.00
1                        0.00                       0.00
2                        0.00                       0.00
3                         NaN                        NaN
4                        0.00                       0.00

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                       0.00                        1.00
1                       0.00                        0.00
2                       0.00                        0.00
```

```
3                              NaN                              NaN
4                             0.00                             0.00
```

[4]:
```
df.info(verbose = True,null_counts = True)
```

```
/var/folders/px/544lvycn58z65rdg800zkv9h0000gn/T/ipykernel_64675/3066153830.py:1
: FutureWarning:

null_counts is deprecated. Use show_counts instead


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Data columns (total 122 columns):
 #    Column                       Non-Null Count   Dtype
---   ------                       --------------   -----
 0    SK_ID_CURR                   307511 non-null  int64
 1    TARGET                       307511 non-null  int64
 2    NAME_CONTRACT_TYPE           307511 non-null  object
 3    CODE_GENDER                  307511 non-null  object
 4    FLAG_OWN_CAR                 307511 non-null  object
 5    FLAG_OWN_REALTY              307511 non-null  object
 6    CNT_CHILDREN                 307511 non-null  int64
 7    AMT_INCOME_TOTAL             307511 non-null  float64
 8    AMT_CREDIT                   307511 non-null  float64
 9    AMT_ANNUITY                  307499 non-null  float64
 10   AMT_GOODS_PRICE              307233 non-null  float64
 11   NAME_TYPE_SUITE              306219 non-null  object
 12   NAME_INCOME_TYPE             307511 non-null  object
 13   NAME_EDUCATION_TYPE          307511 non-null  object
 14   NAME_FAMILY_STATUS           307511 non-null  object
 15   NAME_HOUSING_TYPE            307511 non-null  object
 16   REGION_POPULATION_RELATIVE   307511 non-null  float64
 17   DAYS_BIRTH                   307511 non-null  int64
 18   DAYS_EMPLOYED                307511 non-null  int64
 19   DAYS_REGISTRATION            307511 non-null  float64
 20   DAYS_ID_PUBLISH              307511 non-null  int64
 21   OWN_CAR_AGE                  104582 non-null  float64
 22   FLAG_MOBIL                   307511 non-null  int64
 23   FLAG_EMP_PHONE               307511 non-null  int64
 24   FLAG_WORK_PHONE              307511 non-null  int64
 25   FLAG_CONT_MOBILE             307511 non-null  int64
 26   FLAG_PHONE                   307511 non-null  int64
 27   FLAG_EMAIL                   307511 non-null  int64
 28   OCCUPATION_TYPE              211120 non-null  object
 29   CNT_FAM_MEMBERS              307509 non-null  float64
 30   REGION_RATING_CLIENT         307511 non-null  int64
 31   REGION_RATING_CLIENT_W_CITY  307511 non-null  int64
```

```
32   WEEKDAY_APPR_PROCESS_START    307511 non-null  object
33   HOUR_APPR_PROCESS_START       307511 non-null  int64
34   REG_REGION_NOT_LIVE_REGION    307511 non-null  int64
35   REG_REGION_NOT_WORK_REGION    307511 non-null  int64
36   LIVE_REGION_NOT_WORK_REGION   307511 non-null  int64
37   REG_CITY_NOT_LIVE_CITY        307511 non-null  int64
38   REG_CITY_NOT_WORK_CITY        307511 non-null  int64
39   LIVE_CITY_NOT_WORK_CITY       307511 non-null  int64
40   ORGANIZATION_TYPE             307511 non-null  object
41   EXT_SOURCE_1                  134133 non-null  float64
42   EXT_SOURCE_2                  306851 non-null  float64
43   EXT_SOURCE_3                  246546 non-null  float64
44   APARTMENTS_AVG                151450 non-null  float64
45   BASEMENTAREA_AVG              127568 non-null  float64
46   YEARS_BEGINEXPLUATATION_AVG   157504 non-null  float64
47   YEARS_BUILD_AVG               103023 non-null  float64
48   COMMONAREA_AVG                92646 non-null   float64
49   ELEVATORS_AVG                 143620 non-null  float64
50   ENTRANCES_AVG                 152683 non-null  float64
51   FLOORSMAX_AVG                 154491 non-null  float64
52   FLOORSMIN_AVG                 98869 non-null   float64
53   LANDAREA_AVG                  124921 non-null  float64
54   LIVINGAPARTMENTS_AVG          97312 non-null   float64
55   LIVINGAREA_AVG                153161 non-null  float64
56   NONLIVINGAPARTMENTS_AVG       93997 non-null   float64
57   NONLIVINGAREA_AVG             137829 non-null  float64
58   APARTMENTS_MODE               151450 non-null  float64
59   BASEMENTAREA_MODE             127568 non-null  float64
60   YEARS_BEGINEXPLUATATION_MODE  157504 non-null  float64
61   YEARS_BUILD_MODE              103023 non-null  float64
62   COMMONAREA_MODE               92646 non-null   float64
63   ELEVATORS_MODE                143620 non-null  float64
64   ENTRANCES_MODE                152683 non-null  float64
65   FLOORSMAX_MODE                154491 non-null  float64
66   FLOORSMIN_MODE                98869 non-null   float64
67   LANDAREA_MODE                 124921 non-null  float64
68   LIVINGAPARTMENTS_MODE         97312 non-null   float64
69   LIVINGAREA_MODE               153161 non-null  float64
70   NONLIVINGAPARTMENTS_MODE      93997 non-null   float64
71   NONLIVINGAREA_MODE            137829 non-null  float64
72   APARTMENTS_MEDI               151450 non-null  float64
73   BASEMENTAREA_MEDI             127568 non-null  float64
74   YEARS_BEGINEXPLUATATION_MEDI  157504 non-null  float64
75   YEARS_BUILD_MEDI              103023 non-null  float64
76   COMMONAREA_MEDI               92646 non-null   float64
77   ELEVATORS_MEDI                143620 non-null  float64
78   ENTRANCES_MEDI                152683 non-null  float64
79   FLOORSMAX_MEDI                154491 non-null  float64
```

```
 80   FLOORSMIN_MEDI                98869 non-null   float64
 81   LANDAREA_MEDI                 124921 non-null  float64
 82   LIVINGAPARTMENTS_MEDI         97312 non-null   float64
 83   LIVINGAREA_MEDI               153161 non-null  float64
 84   NONLIVINGAPARTMENTS_MEDI      93997 non-null   float64
 85   NONLIVINGAREA_MEDI            137829 non-null  float64
 86   FONDKAPREMONT_MODE            97216 non-null   object
 87   HOUSETYPE_MODE                153214 non-null  object
 88   TOTALAREA_MODE                159080 non-null  float64
 89   WALLSMATERIAL_MODE            151170 non-null  object
 90   EMERGENCYSTATE_MODE           161756 non-null  object
 91   OBS_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 92   DEF_30_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 93   OBS_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 94   DEF_60_CNT_SOCIAL_CIRCLE      306490 non-null  float64
 95   DAYS_LAST_PHONE_CHANGE        307510 non-null  float64
 96   FLAG_DOCUMENT_2               307511 non-null  int64
 97   FLAG_DOCUMENT_3               307511 non-null  int64
 98   FLAG_DOCUMENT_4               307511 non-null  int64
 99   FLAG_DOCUMENT_5               307511 non-null  int64
 100  FLAG_DOCUMENT_6               307511 non-null  int64
 101  FLAG_DOCUMENT_7               307511 non-null  int64
 102  FLAG_DOCUMENT_8               307511 non-null  int64
 103  FLAG_DOCUMENT_9               307511 non-null  int64
 104  FLAG_DOCUMENT_10              307511 non-null  int64
 105  FLAG_DOCUMENT_11              307511 non-null  int64
 106  FLAG_DOCUMENT_12              307511 non-null  int64
 107  FLAG_DOCUMENT_13              307511 non-null  int64
 108  FLAG_DOCUMENT_14              307511 non-null  int64
 109  FLAG_DOCUMENT_15              307511 non-null  int64
 110  FLAG_DOCUMENT_16              307511 non-null  int64
 111  FLAG_DOCUMENT_17              307511 non-null  int64
 112  FLAG_DOCUMENT_18              307511 non-null  int64
 113  FLAG_DOCUMENT_19              307511 non-null  int64
 114  FLAG_DOCUMENT_20              307511 non-null  int64
 115  FLAG_DOCUMENT_21              307511 non-null  int64
 116  AMT_REQ_CREDIT_BUREAU_HOUR    265992 non-null  float64
 117  AMT_REQ_CREDIT_BUREAU_DAY     265992 non-null  float64
 118  AMT_REQ_CREDIT_BUREAU_WEEK    265992 non-null  float64
 119  AMT_REQ_CREDIT_BUREAU_MON     265992 non-null  float64
 120  AMT_REQ_CREDIT_BUREAU_QRT     265992 non-null  float64
 121  AMT_REQ_CREDIT_BUREAU_YEAR    265992 non-null  float64
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB
```

[5]:
```
df.describe()
```

[5]:

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
count	307511.00	307511.00	307511.00	307511.00	307511.00
mean	278180.52	0.08	0.42	168797.92	599026.00
std	102790.18	0.27	0.72	237123.15	402490.78
min	100002.00	0.00	0.00	25650.00	45000.00
25%	189145.50	0.00	0.00	112500.00	270000.00
50%	278202.00	0.00	0.00	147150.00	513531.00
75%	367142.50	0.00	1.00	202500.00	808650.00
max	456255.00	1.00	19.00	117000000.00	4050000.00

	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH
count	307499.00	307233.00	307511.00	307511.00
mean	27108.57	538396.21	0.02	-16037.00
std	14493.74	369446.46	0.01	4363.99
min	1615.50	40500.00	0.00	-25229.00
25%	16524.00	238500.00	0.01	-19682.00
50%	24903.00	450000.00	0.02	-15750.00
75%	34596.00	679500.00	0.03	-12413.00
max	258025.50	4050000.00	0.07	-7489.00

	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	OWN_CAR_AGE
count	307511.00	307511.00	307511.00	104582.00
mean	63815.05	-4986.12	-2994.20	12.06
std	141275.77	3522.89	1509.45	11.94
min	-17912.00	-24672.00	-7197.00	0.00
25%	-2760.00	-7479.50	-4299.00	5.00
50%	-1213.00	-4504.00	-3254.00	9.00
75%	-289.00	-2010.00	-1720.00	15.00
max	365243.00	0.00	0.00	91.00

	FLAG_MOBIL	FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE
count	307511.00	307511.00	307511.00	307511.00
mean	1.00	0.82	0.20	1.00
std	0.00	0.38	0.40	0.04
min	0.00	0.00	0.00	0.00
25%	1.00	1.00	0.00	1.00
50%	1.00	1.00	0.00	1.00
75%	1.00	1.00	0.00	1.00
max	1.00	1.00	1.00	1.00

	FLAG_PHONE	FLAG_EMAIL	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
count	307511.00	307511.00	307509.00	307511.00
mean	0.28	0.06	2.15	2.05
std	0.45	0.23	0.91	0.51
min	0.00	0.00	1.00	1.00
25%	0.00	0.00	2.00	2.00
50%	0.00	0.00	2.00	2.00

```
75%          1.00        0.00             3.00                  2.00
max          1.00        1.00            20.00                  3.00

       REGION_RATING_CLIENT_W_CITY  HOUR_APPR_PROCESS_START  \
count                    307511.00                307511.00
mean                          2.03                    12.06
std                           0.50                     3.27
min                           1.00                     0.00
25%                           2.00                    10.00
50%                           2.00                    12.00
75%                           2.00                    14.00
max                           3.00                    23.00

       REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION  \
count                   307511.00                   307511.00
mean                         0.02                        0.05
std                          0.12                        0.22
min                          0.00                        0.00
25%                          0.00                        0.00
50%                          0.00                        0.00
75%                          0.00                        0.00
max                          1.00                        1.00

       LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY  \
count                   307511.00               307511.00
mean                         0.04                    0.08
std                          0.20                    0.27
min                          0.00                    0.00
25%                          0.00                    0.00
50%                          0.00                    0.00
75%                          0.00                    0.00
max                          1.00                    1.00

       REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY  EXT_SOURCE_1  \
count               307511.00                307511.00     134133.00
mean                    0.23                     0.18          0.50
std                     0.42                     0.38          0.21
min                     0.00                     0.00          0.01
25%                     0.00                     0.00          0.33
50%                     0.00                     0.00          0.51
75%                     0.00                     0.00          0.68
max                     1.00                     1.00          0.96

       EXT_SOURCE_2  EXT_SOURCE_3  APARTMENTS_AVG  BASEMENTAREA_AVG  \
count     306851.00     246546.00       151450.00         127568.00
mean           0.51          0.51            0.12              0.09
std            0.19          0.19            0.11              0.08
```

min	0.00	0.00	0.00	0.00
25%	0.39	0.37	0.06	0.04
50%	0.57	0.54	0.09	0.08
75%	0.66	0.67	0.15	0.11
max	0.85	0.90	1.00	1.00

	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	\
count	157504.00	103023.00	92646.00	
mean	0.98	0.75	0.04	
std	0.06	0.11	0.08	
min	0.00	0.00	0.00	
25%	0.98	0.69	0.01	
50%	0.98	0.76	0.02	
75%	0.99	0.82	0.05	
max	1.00	1.00	1.00	

	ELEVATORS_AVG	ENTRANCES_AVG	FLOORSMAX_AVG	FLOORSMIN_AVG	\
count	143620.00	152683.00	154491.00	98869.00	
mean	0.08	0.15	0.23	0.23	
std	0.13	0.10	0.14	0.16	
min	0.00	0.00	0.00	0.00	
25%	0.00	0.07	0.17	0.08	
50%	0.00	0.14	0.17	0.21	
75%	0.12	0.21	0.33	0.38	
max	1.00	1.00	1.00	1.00	

	LANDAREA_AVG	LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG	\
count	124921.00	97312.00	153161.00	
mean	0.07	0.10	0.11	
std	0.08	0.09	0.11	
min	0.00	0.00	0.00	
25%	0.02	0.05	0.05	
50%	0.05	0.08	0.07	
75%	0.09	0.12	0.13	
max	1.00	1.00	1.00	

	NONLIVINGAPARTMENTS_AVG	NONLIVINGAREA_AVG	APARTMENTS_MODE	\
count	93997.00	137829.00	151450.00	
mean	0.01	0.03	0.11	
std	0.05	0.07	0.11	
min	0.00	0.00	0.00	
25%	0.00	0.00	0.05	
50%	0.00	0.00	0.08	
75%	0.00	0.03	0.14	
max	1.00	1.00	1.00	

	BASEMENTAREA_MODE	YEARS_BEGINEXPLUATATION_MODE	YEARS_BUILD_MODE	\

count	127568.00	157504.00	103023.00
mean	0.09	0.98	0.76
std	0.08	0.06	0.11
min	0.00	0.00	0.00
25%	0.04	0.98	0.70
50%	0.07	0.98	0.76
75%	0.11	0.99	0.82
max	1.00	1.00	1.00

	COMMONAREA_MODE	ELEVATORS_MODE	ENTRANCES_MODE	FLOORSMAX_MODE	\
count	92646.00	143620.00	152683.00	154491.00	
mean	0.04	0.07	0.15	0.22	
std	0.07	0.13	0.10	0.14	
min	0.00	0.00	0.00	0.00	
25%	0.01	0.00	0.07	0.17	
50%	0.02	0.00	0.14	0.17	
75%	0.05	0.12	0.21	0.33	
max	1.00	1.00	1.00	1.00	

	FLOORSMIN_MODE	LANDAREA_MODE	LIVINGAPARTMENTS_MODE	LIVINGAREA_MODE	\
count	98869.00	124921.00	97312.00	153161.00	
mean	0.23	0.06	0.11	0.11	
std	0.16	0.08	0.10	0.11	
min	0.00	0.00	0.00	0.00	
25%	0.08	0.02	0.05	0.04	
50%	0.21	0.05	0.08	0.07	
75%	0.38	0.08	0.13	0.13	
max	1.00	1.00	1.00	1.00	

	NONLIVINGAPARTMENTS_MODE	NONLIVINGAREA_MODE	APARTMENTS_MEDI	\
count	93997.00	137829.00	151450.00	
mean	0.01	0.03	0.12	
std	0.05	0.07	0.11	
min	0.00	0.00	0.00	
25%	0.00	0.00	0.06	
50%	0.00	0.00	0.09	
75%	0.00	0.02	0.15	
max	1.00	1.00	1.00	

	BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MEDI	YEARS_BUILD_MEDI	\
count	127568.00	157504.00	103023.00	
mean	0.09	0.98	0.76	
std	0.08	0.06	0.11	
min	0.00	0.00	0.00	
25%	0.04	0.98	0.69	
50%	0.08	0.98	0.76	
75%	0.11	0.99	0.83	

```
max                1.00                          1.00             1.00

       COMMONAREA_MEDI  ELEVATORS_MEDI  ENTRANCES_MEDI  FLOORSMAX_MEDI  \
count         92646.00       143620.00       152683.00       154491.00
mean              0.04            0.08            0.15            0.23
std               0.08            0.13            0.10            0.15
min               0.00            0.00            0.00            0.00
25%               0.01            0.00            0.07            0.17
50%               0.02            0.00            0.14            0.17
75%               0.05            0.12            0.21            0.33
max               1.00            1.00            1.00            1.00

       FLOORSMIN_MEDI  LANDAREA_MEDI  LIVINGAPARTMENTS_MEDI  LIVINGAREA_MEDI  \
count        98869.00      124921.00               97312.00        153161.00
mean             0.23           0.07                   0.10             0.11
std              0.16           0.08                   0.09             0.11
min              0.00           0.00                   0.00             0.00
25%              0.08           0.02                   0.05             0.05
50%              0.21           0.05                   0.08             0.07
75%              0.38           0.09                   0.12             0.13
max              1.00           1.00                   1.00             1.00

       NONLIVINGAPARTMENTS_MEDI  NONLIVINGAREA_MEDI  TOTALAREA_MODE  \
count                  93997.00           137829.00       159080.00
mean                       0.01                0.03            0.10
std                        0.05                0.07            0.11
min                        0.00                0.00            0.00
25%                        0.00                0.00            0.04
50%                        0.00                0.00            0.07
75%                        0.00                0.03            0.13
max                        1.00                1.00            1.00

       OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
count                 306490.00                 306490.00
mean                       1.42                      0.14
std                        2.40                      0.45
min                        0.00                      0.00
25%                        0.00                      0.00
50%                        0.00                      0.00
75%                        2.00                      0.00
max                      348.00                     34.00

       OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  \
count                 306490.00                 306490.00
mean                       1.41                      0.10
std                        2.38                      0.36
min                        0.00                      0.00
```

```
25%                      0.00                    0.00
50%                      0.00                    0.00
75%                      2.00                    0.00
max                    344.00                   24.00

       DAYS_LAST_PHONE_CHANGE  FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  \
count               307510.00        307511.00        307511.00
mean                  -962.86             0.00             0.71
std                    826.81             0.01             0.45
min                  -4292.00             0.00             0.00
25%                  -1570.00             0.00             0.00
50%                   -757.00             0.00             1.00
75%                   -274.00             0.00             1.00
max                      0.00             1.00             1.00

       FLAG_DOCUMENT_4  FLAG_DOCUMENT_5  FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  \
count        307511.00        307511.00        307511.00        307511.00
mean              0.00             0.02             0.09             0.00
std               0.01             0.12             0.28             0.01
min               0.00             0.00             0.00             0.00
25%               0.00             0.00             0.00             0.00
50%               0.00             0.00             0.00             0.00
75%               0.00             0.00             0.00             0.00
max               1.00             1.00             1.00             1.00

       FLAG_DOCUMENT_8  FLAG_DOCUMENT_9  FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  \
count        307511.00        307511.00         307511.00         307511.00
mean              0.08             0.00              0.00              0.00
std               0.27             0.06              0.00              0.06
min               0.00             0.00              0.00              0.00
25%               0.00             0.00              0.00              0.00
50%               0.00             0.00              0.00              0.00
75%               0.00             0.00              0.00              0.00
max               1.00             1.00              1.00              1.00

       FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  FLAG_DOCUMENT_14  FLAG_DOCUMENT_15  \
count         307511.00         307511.00         307511.00         307511.00
mean              0.00              0.00              0.00              0.00
std               0.00              0.06              0.05              0.03
min               0.00              0.00              0.00              0.00
25%               0.00              0.00              0.00              0.00
50%               0.00              0.00              0.00              0.00
75%               0.00              0.00              0.00              0.00
max               1.00              1.00              1.00              1.00

       FLAG_DOCUMENT_16  FLAG_DOCUMENT_17  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  \
count         307511.00         307511.00         307511.00         307511.00
```

mean	0.01	0.00	0.01	0.00
std	0.10	0.02	0.09	0.02
min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.00
50%	0.00	0.00	0.00	0.00
75%	0.00	0.00	0.00	0.00
max	1.00	1.00	1.00	1.00

	FLAG_DOCUMENT_20	FLAG_DOCUMENT_21	AMT_REQ_CREDIT_BUREAU_HOUR	\
count	307511.00	307511.00	265992.00	
mean	0.00	0.00	0.01	
std	0.02	0.02	0.08	
min	0.00	0.00	0.00	
25%	0.00	0.00	0.00	
50%	0.00	0.00	0.00	
75%	0.00	0.00	0.00	
max	1.00	1.00	4.00	

	AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	\
count	265992.00	265992.00	
mean	0.01	0.03	
std	0.11	0.20	
min	0.00	0.00	
25%	0.00	0.00	
50%	0.00	0.00	
75%	0.00	0.00	
max	9.00	8.00	

	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	\
count	265992.00	265992.00	
mean	0.27	0.27	
std	0.92	0.79	
min	0.00	0.00	
25%	0.00	0.00	
50%	0.00	0.00	
75%	0.00	0.00	
max	27.00	261.00	

	AMT_REQ_CREDIT_BUREAU_YEAR
count	265992.00
mean	1.90
std	1.87
min	0.00
25%	0.00
50%	1.00
75%	3.00
max	25.00

[6]:
```
# Missing values
df.isnull().values.any()
```

[6]:
```
True
```

[7]:
```
df.columns[df.isnull().any()]
```

[7]:
```
Index(['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'NAME_TYPE_SUITE', 'OWN_CAR_AGE',
       'OCCUPATION_TYPE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_1', 'EXT_SOURCE_2',
       'EXT_SOURCE_3', 'APARTMENTS_AVG', 'BASEMENTAREA_AVG',
       'YEARS_BEGINEXPLUATATION_AVG', 'YEARS_BUILD_AVG', 'COMMONAREA_AVG',
       'ELEVATORS_AVG', 'ENTRANCES_AVG', 'FLOORSMAX_AVG', 'FLOORSMIN_AVG',
       'LANDAREA_AVG', 'LIVINGAPARTMENTS_AVG', 'LIVINGAREA_AVG',
       'NONLIVINGAPARTMENTS_AVG', 'NONLIVINGAREA_AVG', 'APARTMENTS_MODE',
       'BASEMENTAREA_MODE', 'YEARS_BEGINEXPLUATATION_MODE', 'YEARS_BUILD_MODE',
       'COMMONAREA_MODE', 'ELEVATORS_MODE', 'ENTRANCES_MODE', 'FLOORSMAX_MODE',
       'FLOORSMIN_MODE', 'LANDAREA_MODE', 'LIVINGAPARTMENTS_MODE',
       'LIVINGAREA_MODE', 'NONLIVINGAPARTMENTS_MODE', 'NONLIVINGAREA_MODE',
       'APARTMENTS_MEDI', 'BASEMENTAREA_MEDI', 'YEARS_BEGINEXPLUATATION_MEDI',
       'YEARS_BUILD_MEDI', 'COMMONAREA_MEDI', 'ELEVATORS_MEDI',
       'ENTRANCES_MEDI', 'FLOORSMAX_MEDI', 'FLOORSMIN_MEDI', 'LANDAREA_MEDI',
       'LIVINGAPARTMENTS_MEDI', 'LIVINGAREA_MEDI', 'NONLIVINGAPARTMENTS_MEDI',
       'NONLIVINGAREA_MEDI', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
       'TOTALAREA_MODE', 'WALLSMATERIAL_MODE', 'EMERGENCYSTATE_MODE',
       'OBS_30_CNT_SOCIAL_CIRCLE', 'DEF_30_CNT_SOCIAL_CIRCLE',
       'OBS_60_CNT_SOCIAL_CIRCLE', 'DEF_60_CNT_SOCIAL_CIRCLE',
       'DAYS_LAST_PHONE_CHANGE', 'AMT_REQ_CREDIT_BUREAU_HOUR',
       'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',
       'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',
       'AMT_REQ_CREDIT_BUREAU_YEAR'],
      dtype='object')
```

[8]:
```
len(df.columns[df.isnull().any()])
```

[8]:
```
67
```

[9]:
```
#There are 67 columns having one or more NULL values in the dataset
```

[10]:
```
# Computing count and percentage of missing values
nullcount = df.isnull().sum()
nullpercentage = round((df.isnull().sum()/df.shape[0])*100, 2)
```

[11]:
```
null_df = pd.DataFrame({'column_name' : df.columns,'null_count' :␣
 ↪nullcount,'null_percentage': nullpercentage})
null_df.reset_index(drop = True, inplace = True)
```

[12]:
```
null_df.sort_values(by = 'null_percentage', ascending = False)
```

[12]:

	column_name	null_count	null_percentage
76	COMMONAREA_MEDI	214865	69.87
48	COMMONAREA_AVG	214865	69.87
62	COMMONAREA_MODE	214865	69.87
70	NONLIVINGAPARTMENTS_MODE	213514	69.43
56	NONLIVINGAPARTMENTS_AVG	213514	69.43
84	NONLIVINGAPARTMENTS_MEDI	213514	69.43
86	FONDKAPREMONT_MODE	210295	68.39
68	LIVINGAPARTMENTS_MODE	210199	68.35
54	LIVINGAPARTMENTS_AVG	210199	68.35
82	LIVINGAPARTMENTS_MEDI	210199	68.35
52	FLOORSMIN_AVG	208642	67.85
66	FLOORSMIN_MODE	208642	67.85
80	FLOORSMIN_MEDI	208642	67.85
75	YEARS_BUILD_MEDI	204488	66.50
61	YEARS_BUILD_MODE	204488	66.50
47	YEARS_BUILD_AVG	204488	66.50
21	OWN_CAR_AGE	202929	65.99
81	LANDAREA_MEDI	182590	59.38
67	LANDAREA_MODE	182590	59.38
53	LANDAREA_AVG	182590	59.38
73	BASEMENTAREA_MEDI	179943	58.52
45	BASEMENTAREA_AVG	179943	58.52
59	BASEMENTAREA_MODE	179943	58.52
41	EXT_SOURCE_1	173378	56.38
71	NONLIVINGAREA_MODE	169682	55.18
57	NONLIVINGAREA_AVG	169682	55.18
85	NONLIVINGAREA_MEDI	169682	55.18
77	ELEVATORS_MEDI	163891	53.30
49	ELEVATORS_AVG	163891	53.30
63	ELEVATORS_MODE	163891	53.30
89	WALLSMATERIAL_MODE	156341	50.84
72	APARTMENTS_MEDI	156061	50.75
44	APARTMENTS_AVG	156061	50.75
58	APARTMENTS_MODE	156061	50.75
78	ENTRANCES_MEDI	154828	50.35
50	ENTRANCES_AVG	154828	50.35
64	ENTRANCES_MODE	154828	50.35
55	LIVINGAREA_AVG	154350	50.19
69	LIVINGAREA_MODE	154350	50.19
83	LIVINGAREA_MEDI	154350	50.19
87	HOUSETYPE_MODE	154297	50.18
65	FLOORSMAX_MODE	153020	49.76
79	FLOORSMAX_MEDI	153020	49.76
51	FLOORSMAX_AVG	153020	49.76
60	YEARS_BEGINEXPLUATATION_MODE	150007	48.78
74	YEARS_BEGINEXPLUATATION_MEDI	150007	48.78

46	YEARS_BEGINEXPLUATATION_AVG	150007	48.78
88	TOTALAREA_MODE	148431	48.27
90	EMERGENCYSTATE_MODE	145755	47.40
28	OCCUPATION_TYPE	96391	31.35
43	EXT_SOURCE_3	60965	19.83
116	AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.50
117	AMT_REQ_CREDIT_BUREAU_DAY	41519	13.50
118	AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.50
119	AMT_REQ_CREDIT_BUREAU_MON	41519	13.50
120	AMT_REQ_CREDIT_BUREAU_QRT	41519	13.50
121	AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.50
11	NAME_TYPE_SUITE	1292	0.42
92	DEF_30_CNT_SOCIAL_CIRCLE	1021	0.33
91	OBS_30_CNT_SOCIAL_CIRCLE	1021	0.33
93	OBS_60_CNT_SOCIAL_CIRCLE	1021	0.33
94	DEF_60_CNT_SOCIAL_CIRCLE	1021	0.33
42	EXT_SOURCE_2	660	0.21
10	AMT_GOODS_PRICE	278	0.09
6	CNT_CHILDREN	0	0.00
102	FLAG_DOCUMENT_8	0	0.00
2	NAME_CONTRACT_TYPE	0	0.00
3	CODE_GENDER	0	0.00
4	FLAG_OWN_CAR	0	0.00
95	DAYS_LAST_PHONE_CHANGE	1	0.00
96	FLAG_DOCUMENT_2	0	0.00
97	FLAG_DOCUMENT_3	0	0.00
98	FLAG_DOCUMENT_4	0	0.00
99	FLAG_DOCUMENT_5	0	0.00
100	FLAG_DOCUMENT_6	0	0.00
101	FLAG_DOCUMENT_7	0	0.00
103	FLAG_DOCUMENT_9	0	0.00
115	FLAG_DOCUMENT_21	0	0.00
104	FLAG_DOCUMENT_10	0	0.00
105	FLAG_DOCUMENT_11	0	0.00
5	FLAG_OWN_REALTY	0	0.00
107	FLAG_DOCUMENT_13	0	0.00
108	FLAG_DOCUMENT_14	0	0.00
109	FLAG_DOCUMENT_15	0	0.00
110	FLAG_DOCUMENT_16	0	0.00
111	FLAG_DOCUMENT_17	0	0.00
112	FLAG_DOCUMENT_18	0	0.00
113	FLAG_DOCUMENT_19	0	0.00
114	FLAG_DOCUMENT_20	0	0.00
106	FLAG_DOCUMENT_12	0	0.00
8	AMT_CREDIT	0	0.00
7	AMT_INCOME_TOTAL	0	0.00
26	FLAG_PHONE	0	0.00

39	LIVE_CITY_NOT_WORK_CITY	0	0.00
38	REG_CITY_NOT_WORK_CITY	0	0.00
1	TARGET	0	0.00
37	REG_CITY_NOT_LIVE_CITY	0	0.00
36	LIVE_REGION_NOT_WORK_REGION	0	0.00
35	REG_REGION_NOT_WORK_REGION	0	0.00
34	REG_REGION_NOT_LIVE_REGION	0	0.00
33	HOUR_APPR_PROCESS_START	0	0.00
32	WEEKDAY_APPR_PROCESS_START	0	0.00
31	REGION_RATING_CLIENT_W_CITY	0	0.00
30	REGION_RATING_CLIENT	0	0.00
29	CNT_FAM_MEMBERS	2	0.00
27	FLAG_EMAIL	0	0.00
25	FLAG_CONT_MOBILE	0	0.00
40	ORGANIZATION_TYPE	0	0.00
24	FLAG_WORK_PHONE	0	0.00
23	FLAG_EMP_PHONE	0	0.00
22	FLAG_MOBIL	0	0.00
20	DAYS_ID_PUBLISH	0	0.00
19	DAYS_REGISTRATION	0	0.00
18	DAYS_EMPLOYED	0	0.00
17	DAYS_BIRTH	0	0.00
16	REGION_POPULATION_RELATIVE	0	0.00
15	NAME_HOUSING_TYPE	0	0.00
14	NAME_FAMILY_STATUS	0	0.00
13	NAME_EDUCATION_TYPE	0	0.00
12	NAME_INCOME_TYPE	0	0.00
9	AMT_ANNUITY	12	0.00
0	SK_ID_CURR	0	0.00

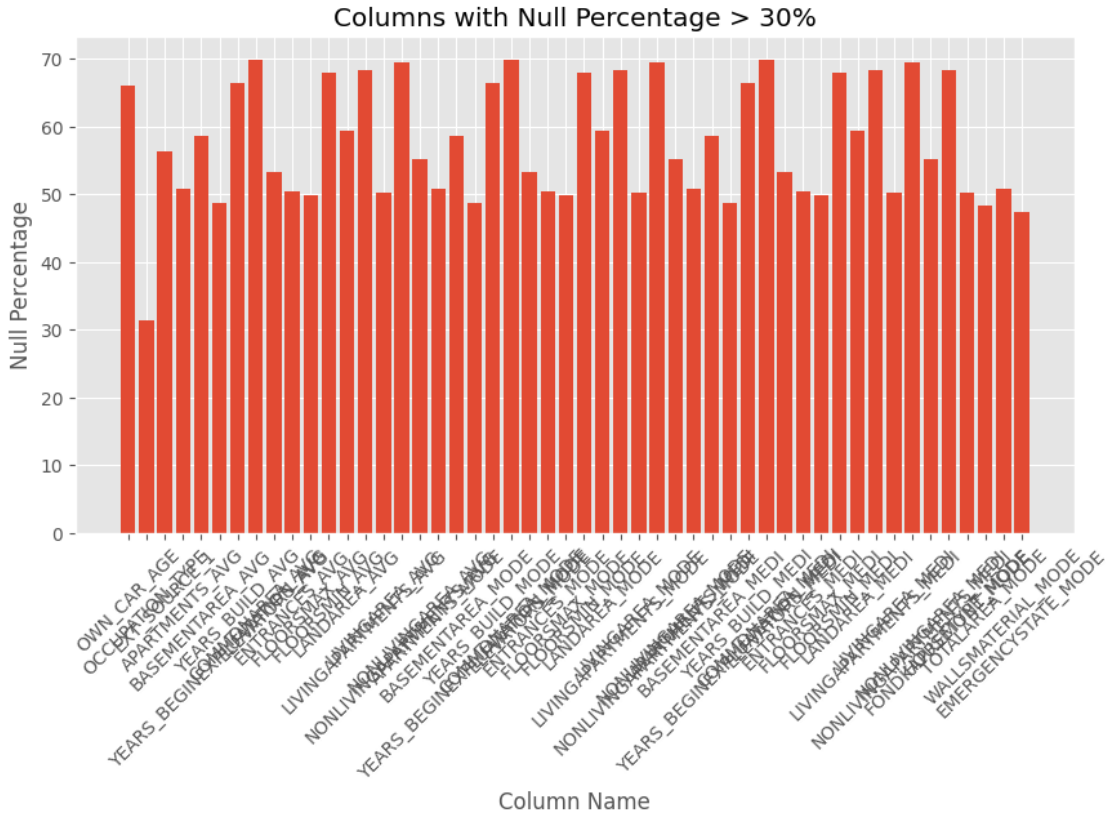
[13]:
```
# Removing columns with NULL values > 30%
columns_to_be_deleted = null_df[null_df['null_percentage'] > 30].column_name.
 ↪to_list()
columns_to_be_deleted
```

[13]:
```
['OWN_CAR_AGE',
 'OCCUPATION_TYPE',
 'EXT_SOURCE_1',
 'APARTMENTS_AVG',
 'BASEMENTAREA_AVG',
 'YEARS_BEGINEXPLUATATION_AVG',
 'YEARS_BUILD_AVG',
 'COMMONAREA_AVG',
 'ELEVATORS_AVG',
 'ENTRANCES_AVG',
 'FLOORSMAX_AVG',
 'FLOORSMIN_AVG',
```

```
 'LANDAREA_AVG',
 'LIVINGAPARTMENTS_AVG',
 'LIVINGAREA_AVG',
 'NONLIVINGAPARTMENTS_AVG',
 'NONLIVINGAREA_AVG',
 'APARTMENTS_MODE',
 'BASEMENTAREA_MODE',
 'YEARS_BEGINEXPLUATATION_MODE',
 'YEARS_BUILD_MODE',
 'COMMONAREA_MODE',
 'ELEVATORS_MODE',
 'ENTRANCES_MODE',
 'FLOORSMAX_MODE',
 'FLOORSMIN_MODE',
 'LANDAREA_MODE',
 'LIVINGAPARTMENTS_MODE',
 'LIVINGAREA_MODE',
 'NONLIVINGAPARTMENTS_MODE',
 'NONLIVINGAREA_MODE',
 'APARTMENTS_MEDI',
 'BASEMENTAREA_MEDI',
 'YEARS_BEGINEXPLUATATION_MEDI',
 'YEARS_BUILD_MEDI',
 'COMMONAREA_MEDI',
 'ELEVATORS_MEDI',
 'ENTRANCES_MEDI',
 'FLOORSMAX_MEDI',
 'FLOORSMIN_MEDI',
 'LANDAREA_MEDI',
 'LIVINGAPARTMENTS_MEDI',
 'LIVINGAREA_MEDI',
 'NONLIVINGAPARTMENTS_MEDI',
 'NONLIVINGAREA_MEDI',
 'FONDKAPREMONT_MODE',
 'HOUSETYPE_MODE',
 'TOTALAREA_MODE',
 'WALLSMATERIAL_MODE',
 'EMERGENCYSTATE_MODE']
```

[14]:
```
# Filter the corresponding null_percentage values
null_percentages = null_df[null_df['null_percentage'] > 30]['null_percentage'].
 ↪tolist()

# Plotting the graph
plt.figure(figsize=(10, 5))
plt.bar(range(len(columns_to_be_deleted)), null_percentages)
```

```
plt.xticks(range(len(columns_to_be_deleted)), columns_to_be_deleted,␣
 ↪rotation=45)
plt.xlabel('Column Name')
plt.ylabel('Null Percentage')
plt.title('Columns with Null Percentage > 30%')
plt.show()
```



[15]:
```
len(columns_to_be_deleted)
```

[15]: 50

[16]:
```
#There are totally 50 columns to be removed. Deleting them from main dataframe
```

[17]:
```
df.drop(columns = columns_to_be_deleted, inplace = True)
```

[18]:
```
df.shape
```

[18]: (307511, 72)

[19]:
```
# Checking columns with NULL values < 30%
```

```
null_df_under30 = null_df[(null_df['null_percentage'] < 30) &␣
 ↪(null_df['null_percentage'] > 0)]
```

[20]:
```
null_df_under30.sort_values(by = 'null_percentage', ascending = False)
```

[20]:

	column_name	null_count	null_percentage
43	EXT_SOURCE_3	60965	19.83
116	AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.50
117	AMT_REQ_CREDIT_BUREAU_DAY	41519	13.50
118	AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.50
119	AMT_REQ_CREDIT_BUREAU_MON	41519	13.50
120	AMT_REQ_CREDIT_BUREAU_QRT	41519	13.50
121	AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.50
11	NAME_TYPE_SUITE	1292	0.42
91	OBS_30_CNT_SOCIAL_CIRCLE	1021	0.33
92	DEF_30_CNT_SOCIAL_CIRCLE	1021	0.33
93	OBS_60_CNT_SOCIAL_CIRCLE	1021	0.33
94	DEF_60_CNT_SOCIAL_CIRCLE	1021	0.33
42	EXT_SOURCE_2	660	0.21
10	AMT_GOODS_PRICE	278	0.09

[21]:
```
# Replacing 19.83% of missing in EXT_SOURCE_3 with median

df.EXT_SOURCE_3.fillna(df.EXT_SOURCE_3.median(),inplace=True)
```

[22]:
```
# Analysis of CODE_GENDER

df['CODE_GENDER'].value_counts()
```

[22]:
```
F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64
```

[23]:
```
df['CODE_GENDER'].replace(to_replace='XNA',value=df['CODE_GENDER'].
 ↪mode()[0],inplace=True)
df['CODE_GENDER'].value_counts()
```

[23]:
```
F    202452
M    105059
Name: CODE_GENDER, dtype: int64
```

[24]:
```
# Replace 'XNA' with NaN
df = df.replace('XNA',np.NaN)
```

[25]:
```
df= df[[i for i in  df.columns if 'FLAG' not in i]]
df.head()
```

[25]:

	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	202500.00	406597.50	24700.50	351000.00	Unaccompanied
1	270000.00	1293502.50	35698.50	1129500.00	Family
2	67500.00	135000.00	6750.00	135000.00	Unaccompanied
3	135000.00	312682.50	29686.50	297000.00	Unaccompanied
4	121500.00	513000.00	21865.50	513000.00	Unaccompanied

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

	NAME_HOUSING_TYPE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED
0	House / apartment	0.02	-9461	-637
1	House / apartment	0.00	-16765	-1188
2	House / apartment	0.01	-19046	-225
3	House / apartment	0.01	-19005	-3039
4	House / apartment	0.03	-19932	-3038

	DAYS_REGISTRATION	DAYS_ID_PUBLISH	CNT_FAM_MEMBERS	REGION_RATING_CLIENT
0	-3648.00	-2120	1.00	2
1	-1186.00	-291	2.00	1
2	-4260.00	-2531	1.00	2
3	-9833.00	-2437	2.00	2
4	-4311.00	-3458	1.00	2

	REGION_RATING_CLIENT_W_CITY	WEEKDAY_APPR_PROCESS_START
0	2	WEDNESDAY
1	1	MONDAY
2	2	MONDAY
3	2	WEDNESDAY
4	2	THURSDAY

	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION
0	10	0
1	11	0
2	9	0
3	17	0

```
4                       11                             0

   REG_REGION_NOT_WORK_REGION  LIVE_REGION_NOT_WORK_REGION  \
0                           0                            0
1                           0                            0
2                           0                            0
3                           0                            0
4                           0                            0

   REG_CITY_NOT_LIVE_CITY  REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY  \
0                       0                       0                        0
1                       0                       0                        0
2                       0                       0                        0
3                       0                       0                        0
4                       0                       1                        1

        ORGANIZATION_TYPE  EXT_SOURCE_2  EXT_SOURCE_3  \
0  Business Entity Type 3          0.26          0.14
1                  School          0.62          0.54
2              Government          0.56          0.73
3  Business Entity Type 3          0.65          0.54
4                Religion          0.32          0.54

   OBS_30_CNT_SOCIAL_CIRCLE  DEF_30_CNT_SOCIAL_CIRCLE  \
0                      2.00                      2.00
1                      1.00                      0.00
2                      0.00                      0.00
3                      2.00                      0.00
4                      0.00                      0.00

   OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
0                      2.00                      2.00                -1134.00
1                      1.00                      0.00                 -828.00
2                      0.00                      0.00                 -815.00
3                      2.00                      0.00                 -617.00
4                      0.00                      0.00                -1106.00

   AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
0                        0.00                       0.00
1                        0.00                       0.00
2                        0.00                       0.00
3                         NaN                        NaN
4                        0.00                       0.00

   AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                        0.00                       0.00
1                        0.00                       0.00
```

```
2                           0.00                           0.00
3                            NaN                            NaN
4                           0.00                           0.00

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                       0.00                        1.00
1                       0.00                        0.00
2                       0.00                        0.00
3                        NaN                         NaN
4                       0.00                        0.00
```

[26]:
```
# Replace the missing values of below columns with mode
# - AMT_REQ_CREDIT_BUREAU_MONTH
# - AMT_REQ_CREDIT_BUREAU_WEEK
# - AMT_REQ_CREDIT_BUREAU_DAY
# - AMT_REQ_CREDIT_BUREAU_HOUR
# - AMT_REQ_CREDIT_BUREAU_QRT


for i in null_df_under30['column_name']:
    if 'AMT_REQ_CREDIT' in i:
        print(f'Most frequent value in {i} is : {df[i].mode()[0]}')
        print(f'Replacing the missing value with : {df[i].mode()[0]}')
        df[i].fillna(df[i].mode()[0],inplace=True)
        print(f'NULL Values in {i} after replacement : {df[i].isnull().sum()}')
        print()
```

```
Most frequent value in AMT_REQ_CREDIT_BUREAU_HOUR is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_HOUR after replacement : 0

Most frequent value in AMT_REQ_CREDIT_BUREAU_DAY is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_DAY after replacement : 0

Most frequent value in AMT_REQ_CREDIT_BUREAU_WEEK is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_WEEK after replacement : 0

Most frequent value in AMT_REQ_CREDIT_BUREAU_MON is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_MON after replacement : 0

Most frequent value in AMT_REQ_CREDIT_BUREAU_QRT is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_QRT after replacement : 0
```

```
Most frequent value in AMT_REQ_CREDIT_BUREAU_YEAR is : 0.0
Replacing the missing value with : 0.0
NULL Values in AMT_REQ_CREDIT_BUREAU_YEAR after replacement : 0
```

[27]:
```
# Replacing 0.21% of missing in EXT_SOURCE_2 with median

df.EXT_SOURCE_2.fillna(df.EXT_SOURCE_2.median(),inplace=True)
```

[28]:
```
# Replacing following columns of AMT with 0.09% missing value with median

df['AMT_ANNUITY'].fillna(df['AMT_ANNUITY'].median(),inplace=True)
df['AMT_GOODS_PRICE'].fillna(df['AMT_GOODS_PRICE'].median(),inplace=True)
```

[29]:
```
# Replacing following columns of CNT_SOCIAL_CIRCLE with 0.33% missing value␣
 ↪with mode

df.OBS_30_CNT_SOCIAL_CIRCLE.fillna( df.OBS_30_CNT_SOCIAL_CIRCLE.
 ↪mode()[0],inplace = True)
df.DEF_30_CNT_SOCIAL_CIRCLE.fillna( df.DEF_30_CNT_SOCIAL_CIRCLE.
 ↪mode()[0],inplace = True)
df.OBS_60_CNT_SOCIAL_CIRCLE.fillna( df.OBS_60_CNT_SOCIAL_CIRCLE.
 ↪mode()[0],inplace = True)
df.DEF_60_CNT_SOCIAL_CIRCLE.fillna( df.DEF_60_CNT_SOCIAL_CIRCLE.
 ↪mode()[0],inplace = True)
```

[30]:
```
# Replacing the column NAME_TYPE_SUITE with 0.42% missing value with median

df.NAME_TYPE_SUITE.fillna(df.NAME_TYPE_SUITE.mode()[0],inplace = True)
```

[31]:
```
#Removing last minimal null values

df.CNT_FAM_MEMBERS.fillna(df.CNT_FAM_MEMBERS.mode() , inplace = True)␣
 ↪#CNT_FAM_MEMBERS

df.DAYS_LAST_PHONE_CHANGE.fillna(df.DAYS_LAST_PHONE_CHANGE.mode()[0],inplace =␣
 ↪True) #DAYS_LAST_PHONE_CHANGE
```

[32]:
```
df.isna().sum()
```

[32]:
```
SK_ID_CURR                    0
TARGET                        0
NAME_CONTRACT_TYPE            0
CODE_GENDER                   0
CNT_CHILDREN                  0
AMT_INCOME_TOTAL              0
AMT_CREDIT                    0
```

```
AMT_ANNUITY                      0
AMT_GOODS_PRICE                  0
NAME_TYPE_SUITE                  0
NAME_INCOME_TYPE                 0
NAME_EDUCATION_TYPE              0
NAME_FAMILY_STATUS               0
NAME_HOUSING_TYPE                0
REGION_POPULATION_RELATIVE       0
DAYS_BIRTH                       0
DAYS_EMPLOYED                    0
DAYS_REGISTRATION                0
DAYS_ID_PUBLISH                  0
CNT_FAM_MEMBERS                  2
REGION_RATING_CLIENT             0
REGION_RATING_CLIENT_W_CITY      0
WEEKDAY_APPR_PROCESS_START       0
HOUR_APPR_PROCESS_START          0
REG_REGION_NOT_LIVE_REGION       0
REG_REGION_NOT_WORK_REGION       0
LIVE_REGION_NOT_WORK_REGION      0
REG_CITY_NOT_LIVE_CITY           0
REG_CITY_NOT_WORK_CITY           0
LIVE_CITY_NOT_WORK_CITY          0
ORGANIZATION_TYPE            55374
EXT_SOURCE_2                     0
EXT_SOURCE_3                     0
OBS_30_CNT_SOCIAL_CIRCLE         0
DEF_30_CNT_SOCIAL_CIRCLE         0
OBS_60_CNT_SOCIAL_CIRCLE         0
DEF_60_CNT_SOCIAL_CIRCLE         0
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
AMT_REQ_CREDIT_BUREAU_QRT        0
AMT_REQ_CREDIT_BUREAU_YEAR       0
dtype: int64
```

[33]:
```
# Correcting date data

days = []
for i in df.columns:
    if 'DAYS' in i:
        days.append(i)
df[days]
```

[33]:

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	\
0	-9461	-637	-3648.00	-2120	
1	-16765	-1188	-1186.00	-291	
2	-19046	-225	-4260.00	-2531	
3	-19005	-3039	-9833.00	-2437	
4	-19932	-3038	-4311.00	-3458	
…	…	…	…	…	
307506	-9327	-236	-8456.00	-1982	
307507	-20775	365243	-4388.00	-4090	
307508	-14966	-7921	-6737.00	-5150	
307509	-11961	-4786	-2562.00	-931	
307510	-16856	-1262	-5128.00	-410	

	DAYS_LAST_PHONE_CHANGE
0	-1134.00
1	-828.00
2	-815.00
3	-617.00
4	-1106.00
…	…
307506	-273.00
307507	0.00
307508	-1909.00
307509	-322.00
307510	-787.00

[307511 rows x 5 columns]

[34]:
```
df[days] = abs(df[days])
df[days]
```

[34]:

	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	\
0	9461	637	3648.00	2120	
1	16765	1188	1186.00	291	
2	19046	225	4260.00	2531	
3	19005	3039	9833.00	2437	
4	19932	3038	4311.00	3458	
…	…	…	…	…	
307506	9327	236	8456.00	1982	
307507	20775	365243	4388.00	4090	
307508	14966	7921	6737.00	5150	
307509	11961	4786	2562.00	931	
307510	16856	1262	5128.00	410	

	DAYS_LAST_PHONE_CHANGE
0	1134.00
1	828.00

```
2                       815.00
3                       617.00
4                      1106.00
…                          …
307506                  273.00
307507                    0.00
307508                 1909.00
307509                  322.00
307510                  787.00

[307511 rows x 5 columns]
```

[35]:
```
# Analysing ORGANIZATION_TYPE with NAME_INCOME_TYPE

df[['ORGANIZATION_TYPE','NAME_INCOME_TYPE']].head(10)
```

[35]:
```
         ORGANIZATION_TYPE      NAME_INCOME_TYPE
0  Business Entity Type 3               Working
1                  School         State servant
2              Government               Working
3  Business Entity Type 3               Working
4                Religion               Working
5                   Other         State servant
6  Business Entity Type 3  Commercial associate
7                   Other         State servant
8                     NaN             Pensioner
9             Electricity               Working
```

[36]:
```
df.NAME_INCOME_TYPE.value_counts()
```

[36]:
```
Working                 158774
Commercial associate     71617
Pensioner                55362
State servant            21703
Unemployed                  22
Student                     18
Businessman                 10
Maternity leave              5
Name: NAME_INCOME_TYPE, dtype: int64
```

[37]:
```
# Here we observe that wherever NAME_INCOME_TYPE is Pensioner there we have␣
 ↪null value in ORGANIZATON_TYPE column

df['ORGANIZATION_TYPE'] = df['ORGANIZATION_TYPE'].replace(np.NaN, 'Pensioner')
```

[38]:
```
#Binning, as we did binning on age group and years employed
```

```
df['YEARS_EMPLOYED']= df['DAYS_EMPLOYED']/365
df['Client_Age']= df['DAYS_BIRTH']/365
df['Age Group']=pd.cut(
    x=df['Client_Age'],
    bins=[0,20,30,40,50,60,100],
    labels=['0-20','20-30','30-40','40-50','50-60','60-100'])
df['Work Experience']=pd.cut(
    x=df['YEARS_EMPLOYED'],
    bins=[0,5,10,15,20,25,30,100],
    labels=['0-5','5-10','10-15','15-20','20-25','25-30','30-100'])
df.drop(columns=['DAYS_EMPLOYED','DAYS_BIRTH'],inplace=True)
```

[39]:
```
df[['SK_ID_CURR','Client_Age','Age Group']]
```

[39]:

	SK_ID_CURR	Client_Age	Age Group
0	100002	25.92	20-30
1	100003	45.93	40-50
2	100004	52.18	50-60
3	100006	52.07	50-60
4	100007	54.61	50-60
…	…	…	…
307506	456251	25.55	20-30
307507	456252	56.92	50-60
307508	456253	41.00	40-50
307509	456254	32.77	30-40
307510	456255	46.18	40-50

[307511 rows x 3 columns]

[40]:
```
df[['SK_ID_CURR','YEARS_EMPLOYED','Work Experience']]
```

[40]:

	SK_ID_CURR	YEARS_EMPLOYED	Work Experience
0	100002	1.75	0-5
1	100003	3.25	0-5
2	100004	0.62	0-5
3	100006	8.33	5-10
4	100007	8.32	5-10
…	…	…	…
307506	456251	0.65	0-5
307507	456252	1000.67	NaN
307508	456253	21.70	20-25
307509	456254	13.11	10-15
307510	456255	3.46	0-5

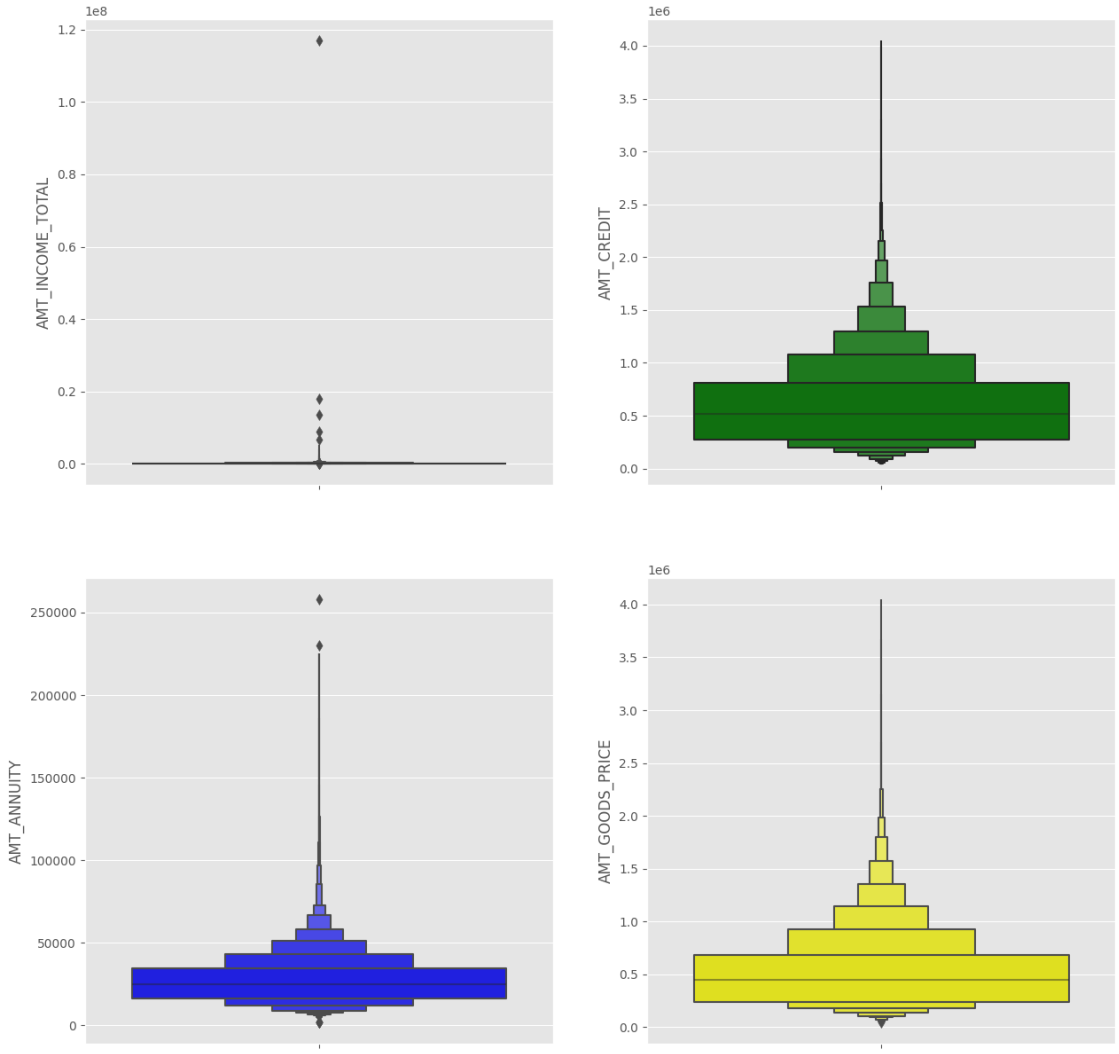
[307511 rows x 3 columns]

[41]:
```
numerical_col = df.select_dtypes(include='number').columns
len(numerical_col)
```

[41]:
```
35
```

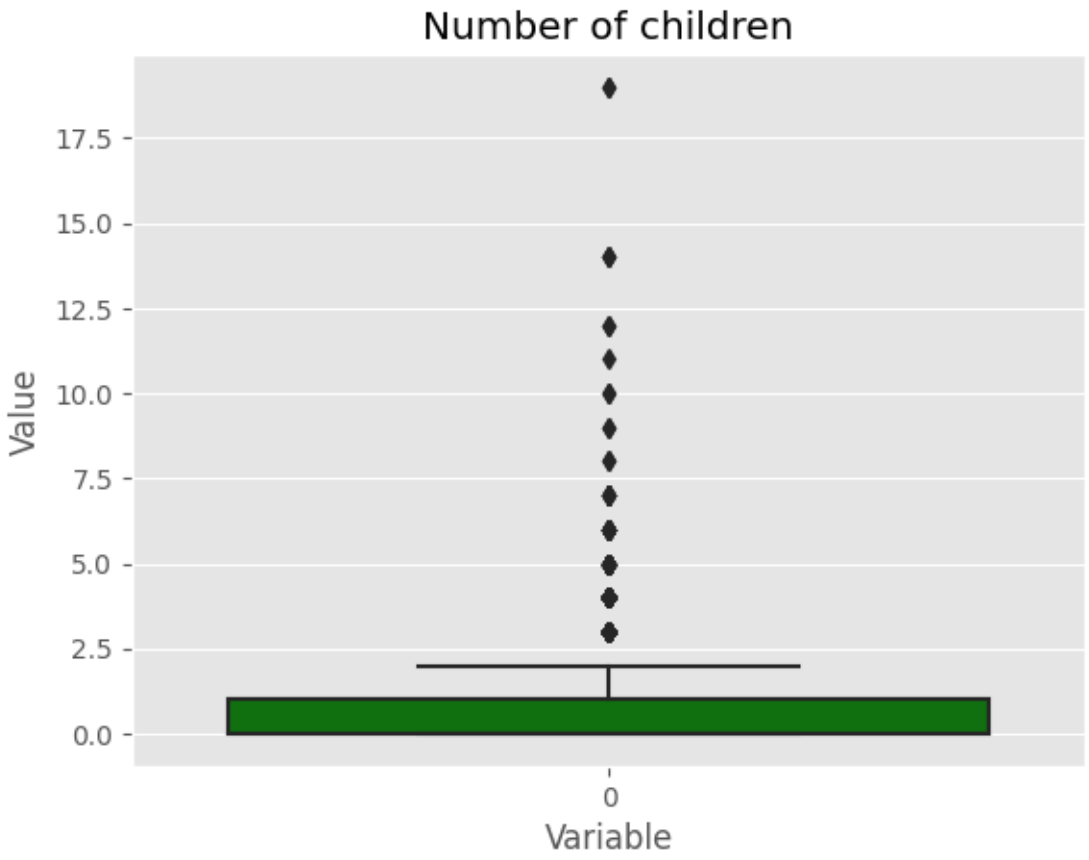
[42]:
```
# droping unwanted columns

unwanted=['REGION_RATING_CLIENT','REGION_RATING_CLIENT_W_CITY',␣
 ↪'REGION_RATING_CLIENT']
df.drop(labels=unwanted,axis=1,inplace=True)
```

[43]:
```
# Analysis Outliers

cols = ['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', 'AMT_GOODS_PRICE']
fig, axes = plt.subplots(ncols=2, nrows=2, figsize=(15, 15))
count = 0
colors = ['red', 'green', 'blue', 'yellow']

for i in range(0, 2):
    for j in range(0, 2):
        sns.boxenplot(y=df[cols[count]], ax=axes[i, j], color=colors[count])
        count += 1

plt.show()
```



[44]:
```
# Removing Outliers

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

[45]:
```
sns.boxplot(df['CNT_CHILDREN'], color='green')
plt.title("Number of children")
plt.xlabel("Variable")
plt.ylabel("Value")
plt.show()
```



[46]:
```
df['CNT_CHILDREN'].value_counts()
```

[46]:
```
0     215371
1      61118
2      26748
3       3717
4        429
5         84
6         21
7          7
14         3
8          2
9          2
12         2
10         2
19         2
11         1
Name: CNT_CHILDREN, dtype: int64
```

[47]:
```
df.shape[0]
```

[47]: 307509

[48]:
```
df= df[df['CNT_CHILDREN']<=5]
```

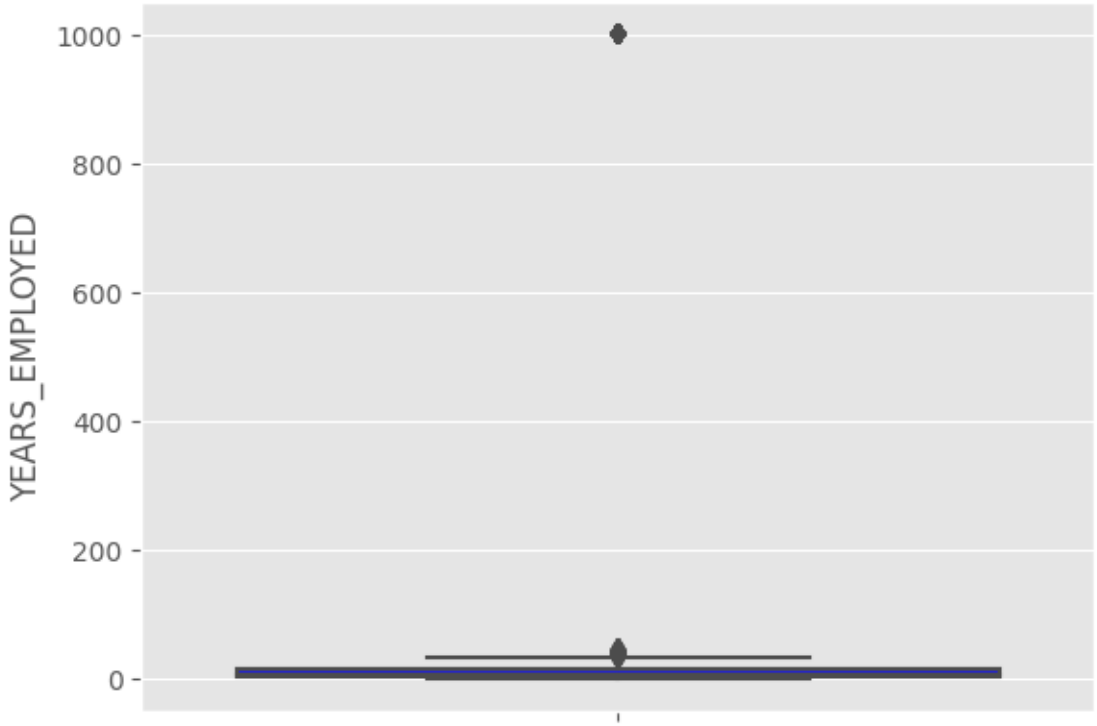
[49]:
```
df.shape[0]
```

[49]: 307467

[50]:
```
#42 values dropped where number of children are greater than 5
```

[51]:
```
sns.boxplot(y=df['YEARS_EMPLOYED'], color='blue')
```

[51]: <AxesSubplot: ylabel='YEARS_EMPLOYED'>

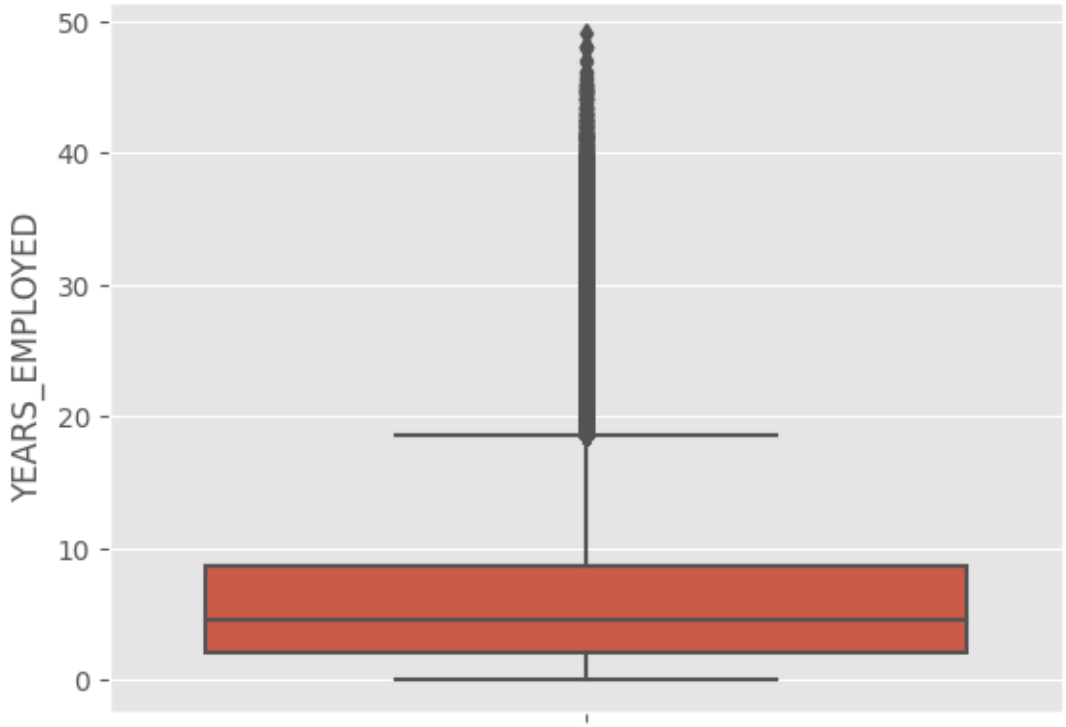


[52]:
```
df['YEARS_EMPLOYED'].value_counts()
```

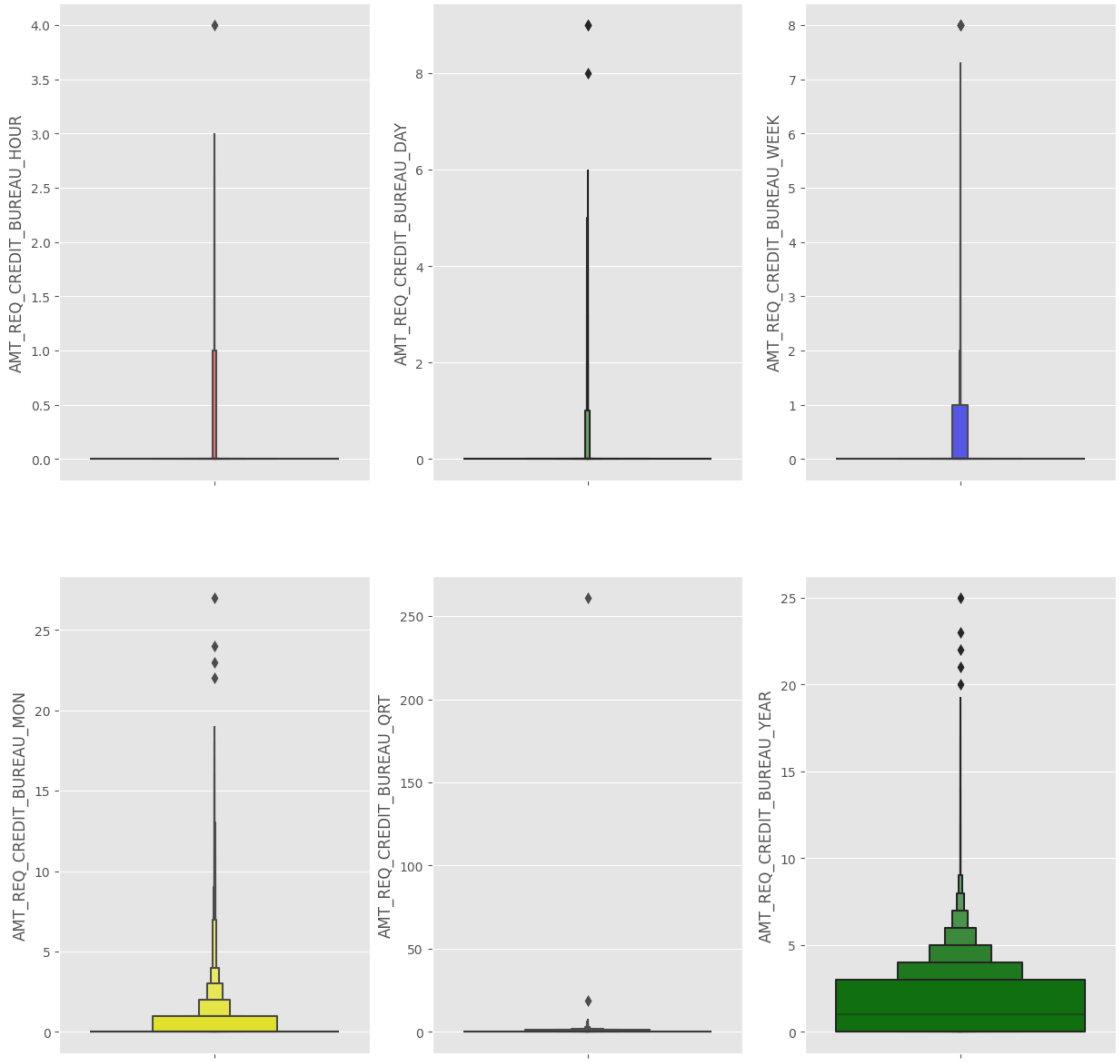
[52]:
```
1000.67    55371
0.55         156
0.61         152
0.63         151
0.55         151
```

```
            ...
23.64           1
42.99           1
30.68           1
36.19           1
23.82           1
Name: YEARS_EMPLOYED, Length: 12574, dtype: int64
```

[53]:
```
df['YEARS_EMPLOYED'][df['YEARS_EMPLOYED']>1000]=np.NaN
sns.boxplot(y=df['YEARS_EMPLOYED'])
plt.show()
```



[54]:
```
# Outliers analysis on AMT_REQ_CREDIT columns

cols = [i for i in  df.columns if 'AMT_REQ' in i]
fig,axes = plt.subplots(ncols=3,nrows=2,figsize=(15,15))
count=0
for i in range(0,2):
    for j in range(0,3):
        sns.boxenplot(y=df[cols[count]],ax=axes[i,j], color=colors[count%4])
        count+=1
plt.show()
```



[55]:
```
#removing outlier from AMT_REQ_CREDIT_BUREAU_QRT

df=df[df['AMT_REQ_CREDIT_BUREAU_QRT']<df['AMT_REQ_CREDIT_BUREAU_QRT'].max()]
```

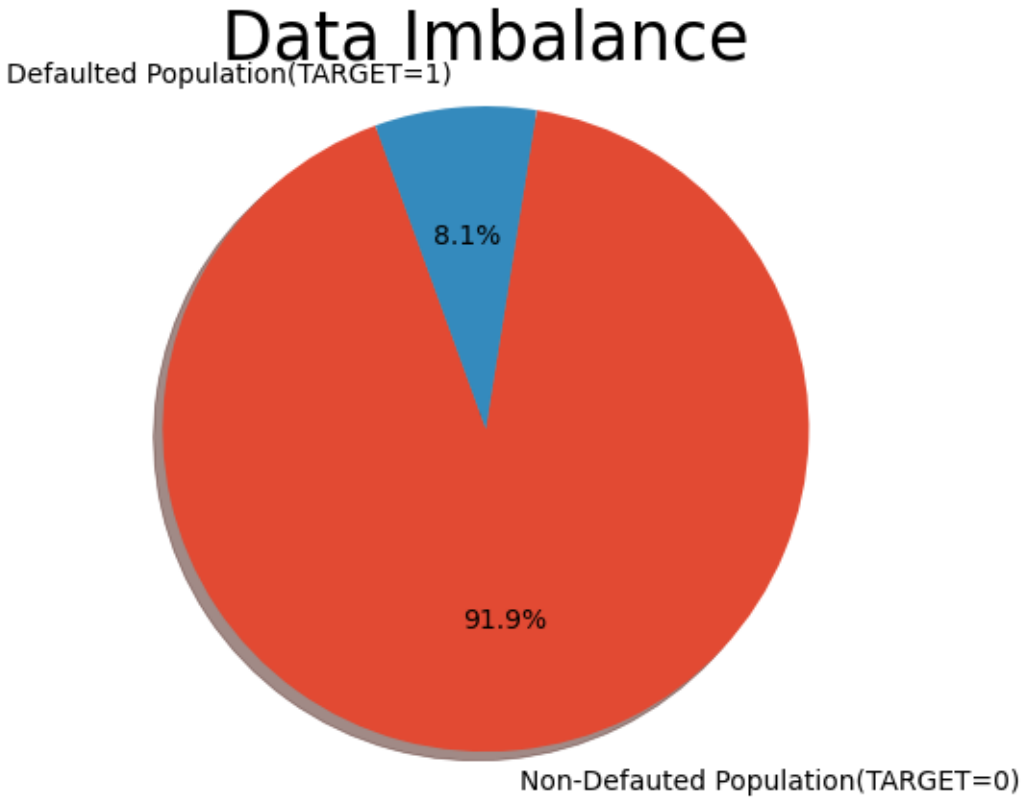
[56]:
```
# Data Imbalance

Target0 = df.loc[df["TARGET"]==0]
Target1 = df.loc[df["TARGET"]==1]
len(Target0)/len(Target1)
```

[56]: 11.390328430384848

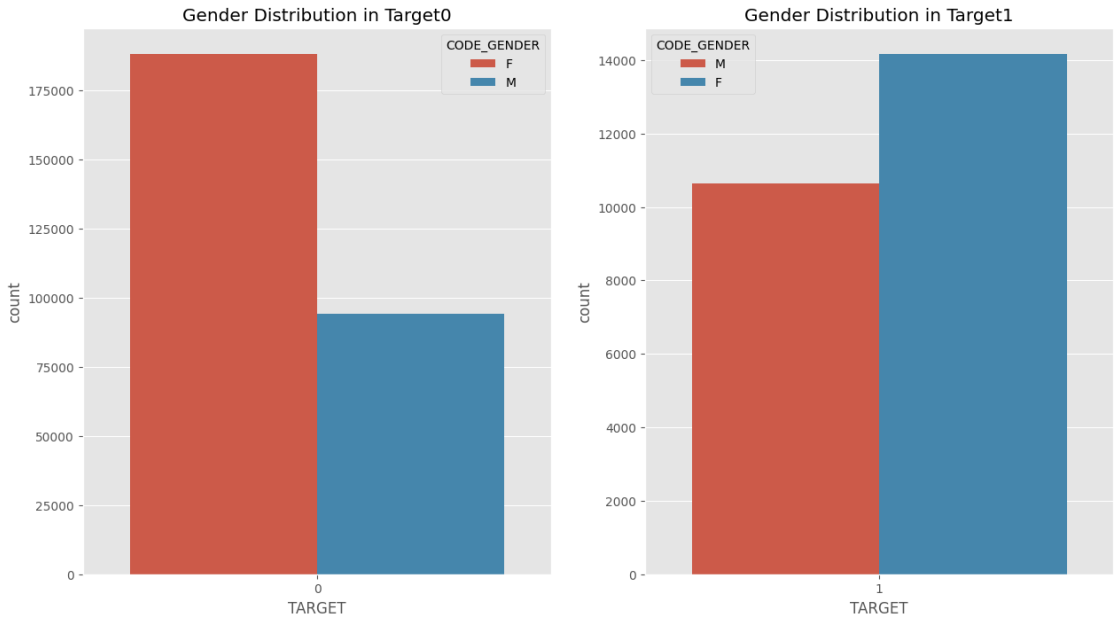
[57]:
```
# Imbalance = 11.390328430384848
```

[58]:
```
# UNIVARIATE Analysis (Gender and Age)

c0=0
c1=0
for i in df['TARGET'].values:
    if i == 0:
        c0 += 1
    else:
        c1 += 1
c0 = (c0/len(df['TARGET']))*100
c1 = (c1/len(df['TARGET']))*100
```

[59]:
```
x = ['Non-Defauted Population(TARGET=0)', 'Defaulted Population(TARGET=1)']
y = [c0, c1]
fig1, ax1 = plt.subplots()
ax1.pie(y, labels=x, autopct='%1.1f%%',
        shadow=True, startangle=110)
ax1.axis('equal')
plt.title('Data Imbalance',fontsize=25)
plt.show()
```

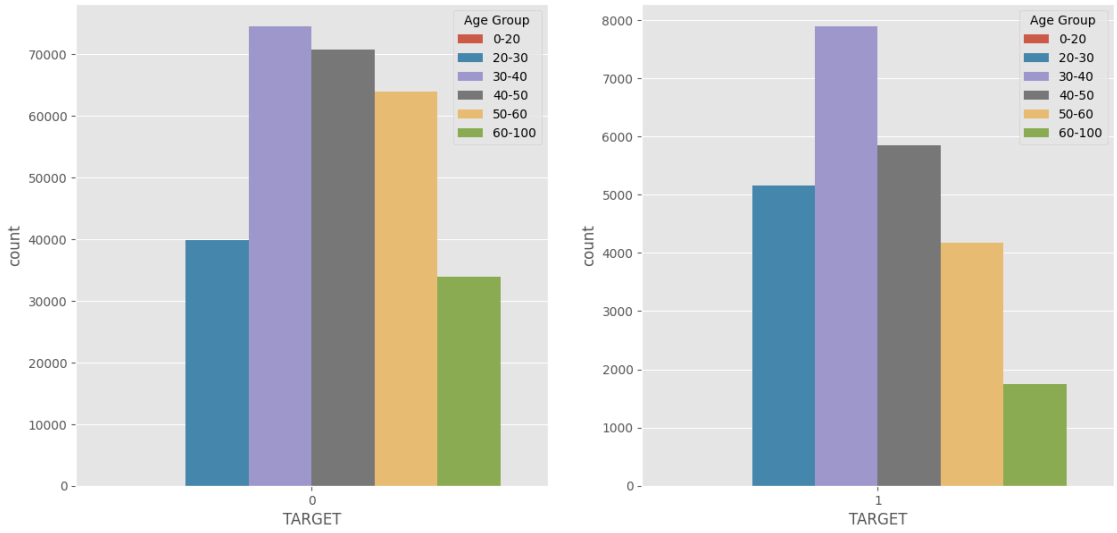


[60]:
```
plt.figure(figsize=(15, 8))

# Plot 1
plt.subplot(121)
sns.countplot(x='TARGET', hue='CODE_GENDER', data=Target0)
plt.title("Gender Distribution in Target0")

# Plot 2
plt.subplot(122)
sns.countplot(x='TARGET', hue='CODE_GENDER', data=Target1)
plt.title("Gender Distribution in Target1")

plt.show()
```



[61]:
```
# It seems like Female clients applied higher than male clients for loan 66.6%
# Female clients are non-defaulters while 33.4% male clients are non-defaulters.
# 57% Female clients are defaulters while 42% male clients are defaulters.
```

[62]:
```
plt.figure(figsize=(15,7))
plt.subplot(121)
sns.countplot(x='TARGET',hue='Age Group',data=Target0)
plt.subplot(122)
sns.countplot(x='TARGET',hue='Age Group',data=Target1)
plt.show()
```



[63]:
```
# Middle Age(35-60) the group seems to applied higher than any other age group.
# Middle Age group facing paying difficulties the most.
# Senior Citizens(60-100) and Very young(19-25) age group facing paying␣
 ↪difficulties less as compared to other age groups.
```

[ ]:
```
cols = ['Age Group','NAME_CONTRACT_TYPE',␣
 ↪'NAME_INCOME_TYPE','NAME_EDUCATION_TYPE',
       'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE','CODE_GENDER','Work␣
 ↪Experience']

#Subplot initialization
fig = make_subplots(
                    rows=4,
                    cols=2,
                    subplot_titles=cols,
                    horizontal_spacing=0.1,
                    vertical_spacing=0.13
                   )
count=0
for i in range(1,5):
    for j in range(1,3):
        fig.add_trace(go.Bar(x=df[cols[count]].value_counts().index,
                             y=df[cols[count]].value_counts(),
                             name=cols[count],
                             textposition='auto',
                             text= [str(i) + '%' for i in (df[cols[count]].
 ↪value_counts(normalize=True)*100).round(1).tolist()],
                            ),
                      row=i,col=j)
```

```
        count+=1
fig.update_layout(
                    title=dict(text = "Analyze Categorical variables (Frequency␣
 ↪/ Percentage)",x=0.5,y=0.99),
                    title_font_size=20,
                    showlegend=False,
                    width = 960,
                    height = 1600,
                  )
fig.show()
```

[65]:
```
# Insights-

# Most of clients who have applied for loan belong to Working Income Type.

# Most of clients with Secondary/Secondary Special education type have applied␣
 ↪for the loan.

# Most of clients who are have applied for loan are married.

# Most of the Clients who have applied for the loan have their own house/
 ↪apartment.

# Female applied for loan more than males.

# Most clients who applied most for loan have work experience between 0-5 years␣
 ↪have.
```

[66]:
```
# Finding TOP 10 CORRELATION in Default population

corr=Target0[Target0.columns].corr(method = 'pearson')
corr=corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool_))
top10_corr0=corr.unstack().reset_index()

# dividing into columns VAR1 & VAR2
top10_corr0.columns = ['VAR1','VAR2','CORRELATION']
top10_corr0.dropna(subset=['CORRELATION'],inplace=True)
top10_corr0['CORR_ABS']=top10_corr0['CORRELATION'].abs()
top10_corr0.sort_values('CORR_ABS', ascending=False).head(10)
```

```
/var/folders/px/544lvycn58z65rdg800zkv9h0000gn/T/ipykernel_64675/1260761502.py:3
: FutureWarning:

The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
```

[66]:

	VAR1	VAR2	CORRELATION	\
746	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	
202	AMT_GOODS_PRICE	AMT_CREDIT	0.99	
332	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88	
475	LIVE_REGION_NOT_WORK_REGION	REG_REGION_NOT_WORK_REGION	0.86	
780	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86	
577	LIVE_CITY_NOT_WORK_CITY	REG_CITY_NOT_WORK_CITY	0.83	
203	AMT_GOODS_PRICE	AMT_ANNUITY	0.78	
169	AMT_ANNUITY	AMT_CREDIT	0.77	
441	REG_REGION_NOT_WORK_REGION	REG_REGION_NOT_LIVE_REGION	0.45	
543	REG_CITY_NOT_WORK_CITY	REG_CITY_NOT_LIVE_CITY	0.44	

	CORR_ABS
746	1.00
202	0.99
332	0.88
475	0.86
780	0.86
577	0.83
203	0.78
169	0.77
441	0.45
543	0.44

[67]:
```
# Finding TOP 10 CORRELATION in Default population

corr=Target1[Target1.columns].corr(method = 'pearson')
corr=corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool_))
top10_corr1=corr.unstack().reset_index()

# dividing into columns VAR1 & VAR2
top10_corr1.columns = ['VAR1','VAR2','CORRELATION']
top10_corr1.dropna(subset=['CORRELATION'],inplace=True)
top10_corr1['CORR_ABS']=top10_corr1['CORRELATION'].abs()
top10_corr1.sort_values('CORR_ABS', ascending=False).head(10)
```

```
/var/folders/px/544lvycn58z65rdg800zkv9h0000gn/T/ipykernel_64675/3596321074.py:3
: FutureWarning:

The default value of numeric_only in DataFrame.corr is deprecated. In a future
version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
```

[67]:

	VAR1	VAR2	CORRELATION	\
746	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	1.00	
202	AMT_GOODS_PRICE	AMT_CREDIT	0.98	

```
332              CNT_FAM_MEMBERS                 CNT_CHILDREN         0.88
780     DEF_60_CNT_SOCIAL_CIRCLE    DEF_30_CNT_SOCIAL_CIRCLE         0.87
475  LIVE_REGION_NOT_WORK_REGION  REG_REGION_NOT_WORK_REGION         0.85
577      LIVE_CITY_NOT_WORK_CITY      REG_CITY_NOT_WORK_CITY         0.78
203              AMT_GOODS_PRICE                  AMT_ANNUITY         0.75
169                  AMT_ANNUITY                   AMT_CREDIT         0.75
441   REG_REGION_NOT_WORK_REGION  REG_REGION_NOT_LIVE_REGION         0.50
543       REG_CITY_NOT_WORK_CITY      REG_CITY_NOT_LIVE_CITY         0.47

     CORR_ABS
746      1.00
202      0.98
332      0.88
780      0.87
475      0.85
577      0.78
203      0.75
169      0.75
441      0.50
543      0.47
```

[68]:
```
# Top 10 correlations are almost at the same level in both the Default and Non
↪Default population
```

[69]:
```
appdata_previous = pd.read_csv("previous_application.csv");
appdata_previous.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
```

```
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB
```

[70]:
```
dfp= (appdata_previous.isnull().mean()*100).
  ↪sort_values(ascending=False)[appdata_previous.isnull().mean()*100 > 30]
appdata_previous.drop(columns = dfp.index,inplace=True)
```

[71]:
```
appdata_previous=appdata_previous.replace('XNA', np.NaN)
appdata_previous=appdata_previous.replace('XAP', np.NaN)
```

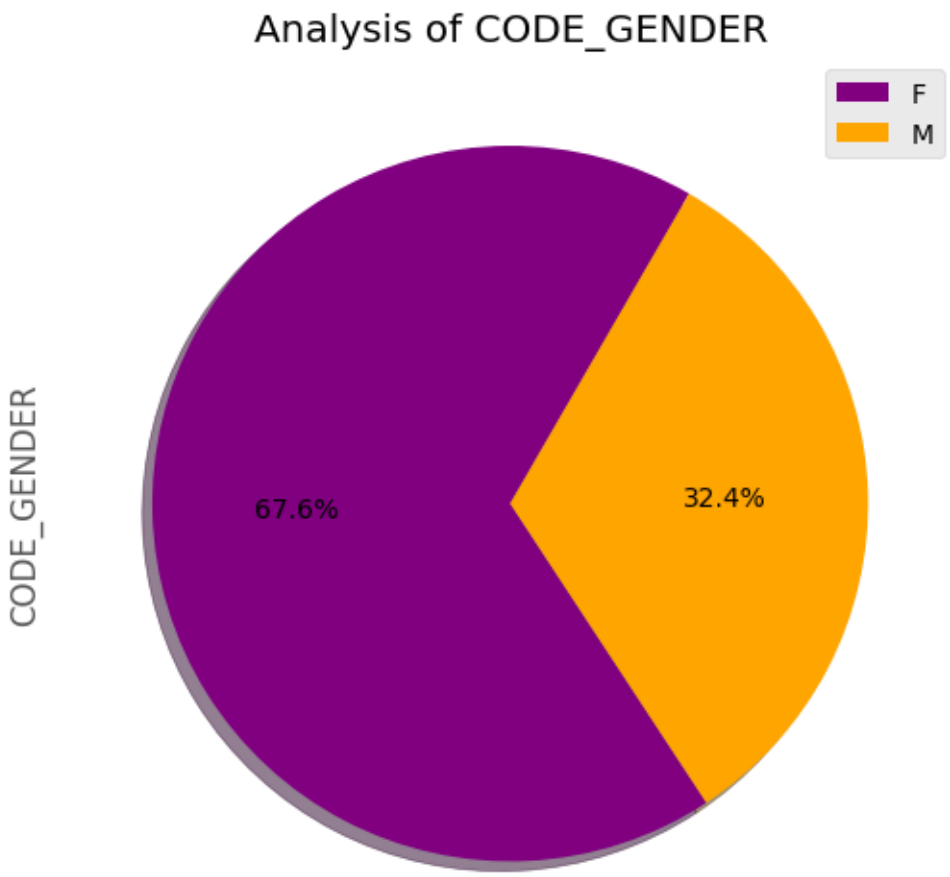
[72]:
```
appdata_merge = df.merge(appdata_previous,on='SK_ID_CURR', how='inner')
appdata_merge.shape
```

[72]:
```
(1413521, 69)
```

[73]:
```
def fn_piechart(column):
    plt.figure(figsize = [10,6])

    pltname = 'Analysis of ' + column
    plt.title(pltname)
    appdata_merge[column].value_counts().plot.pie(autopct='%1.
  ↪1f%%',shadow=True, startangle=60, colors =␣
  ↪['purple','orange','red','green','yellow','pink'], labeldistance=None)

    plt.legend()
```

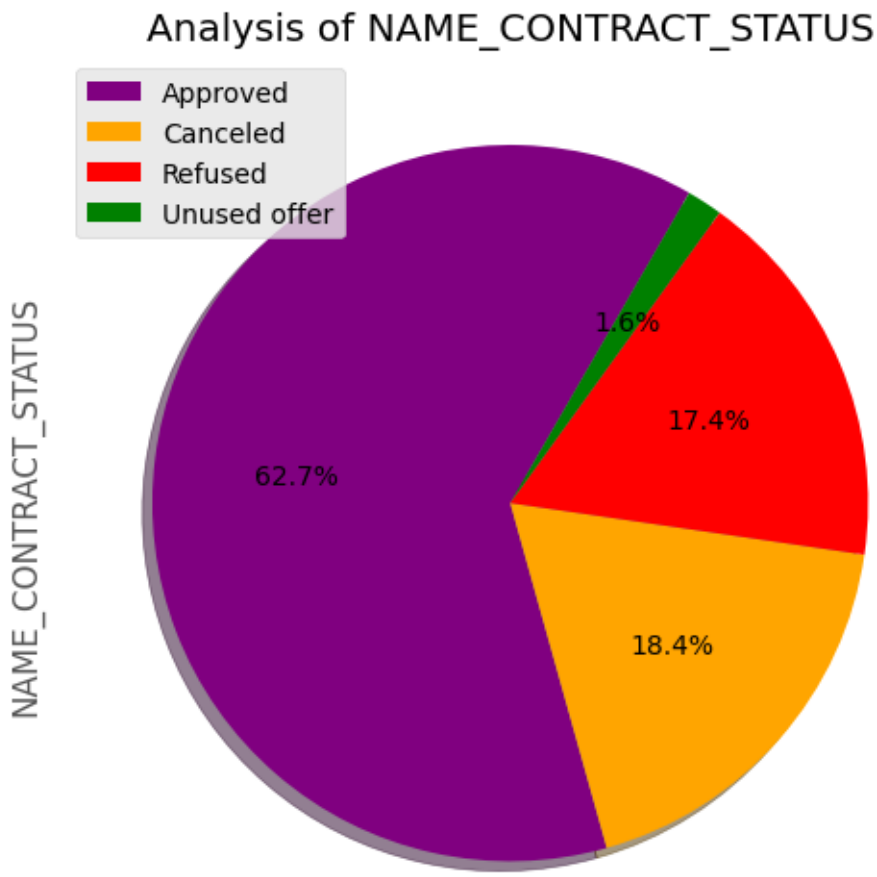
```
    plt.tight_layout(pad = 4)
    plt.show()
```

[74]:
```
fn_piechart('CODE_GENDER')
```



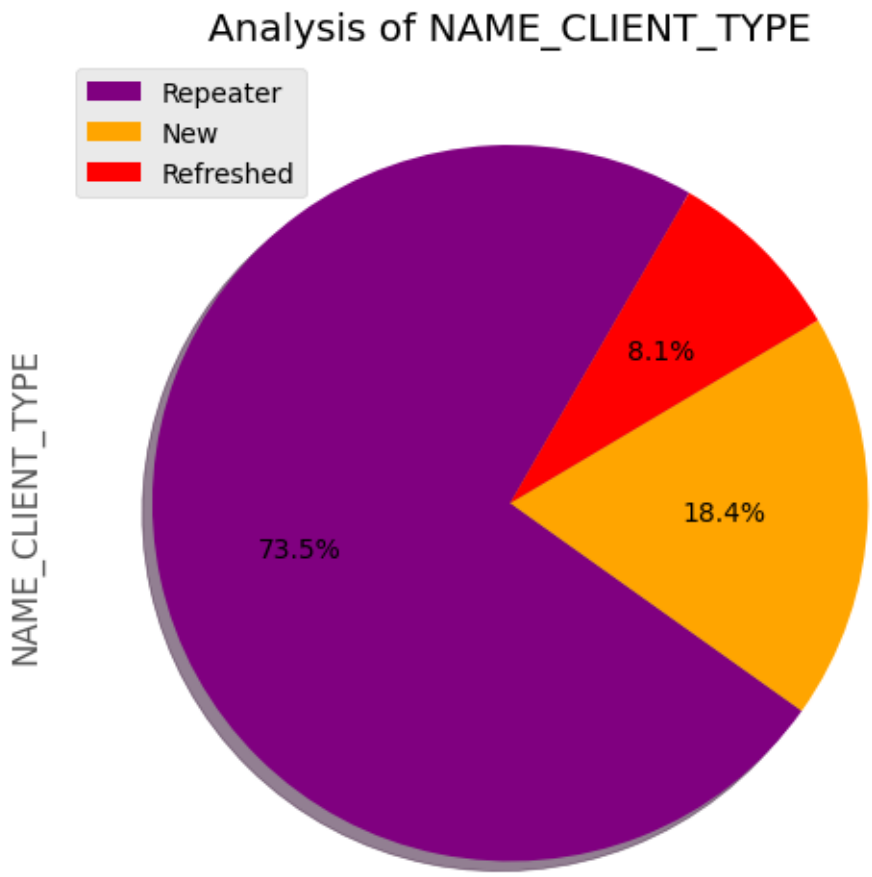
[75]:
```
# Approved percentage of loans provided to females is more as compared to␣
↪refused percentage.
```

[76]:
```
fn_piechart('NAME_CONTRACT_STATUS')
```



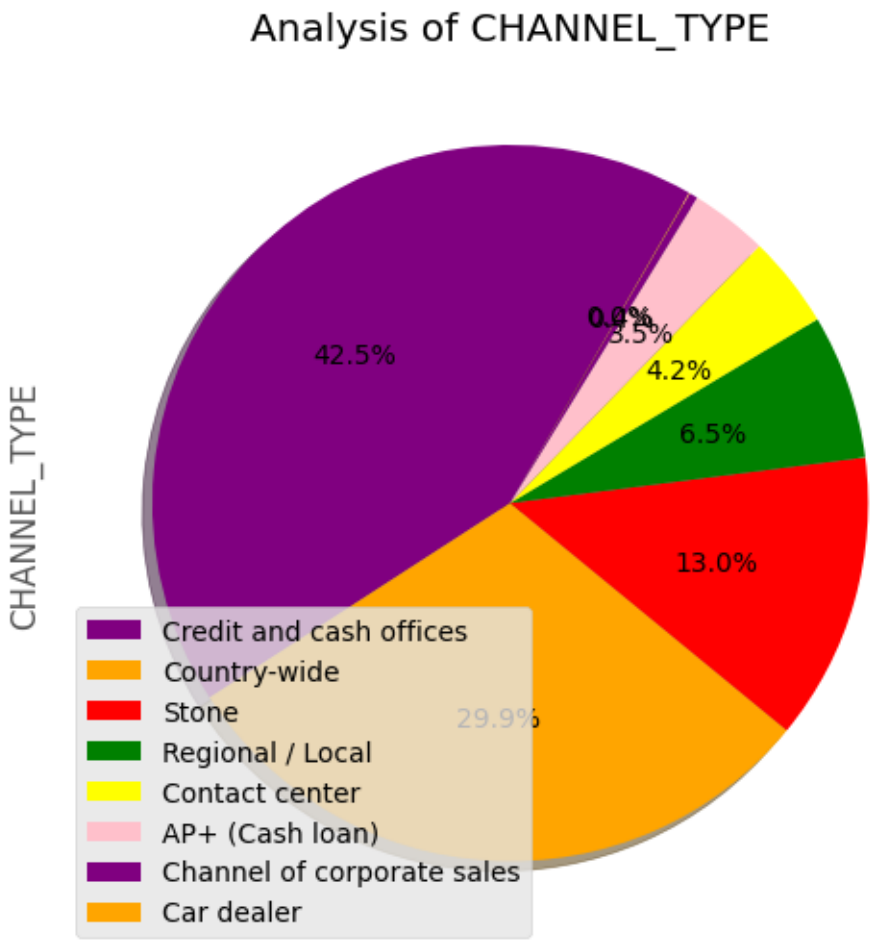
[77]:
```
# Approved loan status is the highest among all loan applications
# Cancelled loan status is the second highest among all loan applications
```

[78]:
```
fn_piechart('NAME_CLIENT_TYPE')
```



[79]:
```
# Repeater client type is the highest among all loan applications
# New client type is the second highest among all loan applications
```

[80]:
```
fn_piechart('CHANNEL_TYPE')
```



[81]:
```
# Country-wide Channel type is the highest among all loan applications
# Credit and cash offices is the second highest Channel Type among all loan␣
 ↪applications
```

[ ]:
```

```