Loan Case Study (Pankaj_Saha)

June 20, 2021

CREDIT - LOAN CASE EDA Case Study:

0.0.1 Importing Libraries and Required Files:

[1]:
```
import pandas as pd
import numpy as np
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
print('Libraries imported successfully:')
```

```
Libraries imported successfully:
```

0.0.2 Setting the max viewing dimension of rows and columns:

[2]:
```
pd.set_option('display.max_columns', 500)
pd.set_option('display.max_rows', 500)
pd.set_option('display.width', 1000)
warnings.filterwarnings('ignore')
print("Alterationn done Successfully:")
```

```
Alterationn done Successfully:
```

0.1 (1) New Application Dataset:

[3]:
```
application=pd.read_csv('application_data.csv')
application.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
AMT_GOODS_PRICE NAME_TYPE_SUITE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  DAYS_BIRTH
DAYS_EMPLOYED  DAYS_REGISTRATION  DAYS_ID_PUBLISH  OWN_CAR_AGE  FLAG_MOBIL
FLAG_EMP_PHONE  FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL
OCCUPATION_TYPE  CNT_FAM_MEMBERS  REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START
REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION
```

```
LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY  REG_CITY_NOT_WORK_CITY
LIVE_CITY_NOT_WORK_CITY       ORGANIZATION_TYPE  EXT_SOURCE_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  \
0      100002       1         Cash loans           M            N
Y            0           202500.0    406597.5      24700.5          351000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                    0.018801       -9461
-637           -3648.0            -2120          NaN           1
1               0                 1           1            0        Laborers
1.0                  2                            2
WEDNESDAY                       10                            0
0                            0                        0                        0
0  Business Entity Type 3      0.083037      0.262949      0.139376
0.0247            0.0369                      0.9722           0.6192
0.0143           0.00         0.0690         0.0833         0.1250        0.0369
1      100003       0         Cash loans           F            N
N            0           270000.0   1293502.5      35698.5         1129500.0
Family    State servant               Higher education               Married
House / apartment                    0.003541      -16765          -1188
-1186.0             -291          NaN           1               1
0               1           1           0      Core staff             2.0
1                         1                     MONDAY
11                           0                          0
0                    0                        0                        0
School      0.311267      0.622246           NaN          0.0959
0.0529                      0.9851           0.7960          0.0605
0.08          0.0345          0.2917          0.3333        0.0130
2      100004       0    Revolving loans           M            Y
Y            0            67500.0    135000.0       6750.0          135000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                    0.010032       -19046
-225           -4260.0            -2531          26.0           1
1               1                 1           1            0        Laborers
1.0                  2                            2
MONDAY                         9                            0
0                            0                        0                        0
0              Government           NaN      0.555912      0.729567
NaN              NaN                         NaN              NaN
NaN            NaN            NaN            NaN            NaN           NaN
3      100006       0         Cash loans           F            N
Y            0           135000.0    312682.5      29686.5          297000.0
Unaccompanied          Working  Secondary / secondary special        Civil
marriage  House / apartment                    0.008019       -19005
-3039            -9833.0            -2437          NaN           1
1               0                 1           0            0        Laborers
```

```
2.0                    2                             2
WEDNESDAY                        17                            0
0                            0                        0                        0
0  Business Entity Type 3           NaN      0.650442           NaN
NaN                 NaN                          NaN              NaN
NaN              NaN              NaN              NaN              NaN            NaN
4      100007       0         Cash loans           M            N
Y             0          121500.0    513000.0      21865.5         513000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                   0.028663      -19932
-3038             -4311.0             -3458          NaN           1
1                 0                 1           0           0      Core staff
1.0                    2                             2
THURSDAY                        11                            0
0                            0                        0                        1
1                Religion           NaN      0.322738           NaN
NaN                 NaN                          NaN              NaN
NaN              NaN              NaN              NaN              NaN            NaN

   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  FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  FLAG_DOCUMENT_4
FLAG_DOCUMENT_5  FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  FLAG_DOCUMENT_8
FLAG_DOCUMENT_9  FLAG_DOCUMENT_10  \
0                 0.0202          0.0190                   0.0000
0.0000            0.0252             0.0383                            0.9722
0.6341            0.0144          0.0000          0.0690          0.0833
0.1250          0.0377                 0.022           0.0198
0.0                  0.0           0.0250             0.0369
0.9722             0.6243          0.0144             0.00            0.0690
0.0833           0.1250          0.0375                 0.0205           0.0193
0.0000                0.00   reg oper account  block of flats          0.0149
Stone, brick                 No                        2.0
2.0                    2.0                       2.0                   -1134.0
0                 1              0                 0                 0
0                 0              0                 0
1                 0.0773          0.0549                   0.0039
```

0.0098 0.0924 0.0538 0.9851
0.8040 0.0497 0.0806 0.0345 0.2917
0.3333 0.0128 0.079 0.0554
0.0 0.0 0.0968 0.0529
0.9851 0.7987 0.0608 0.08 0.0345
0.2917 0.3333 0.0132 0.0787 0.0558
0.0039 0.01 reg oper account block of flats 0.0714
Block No 1.0 0.0
1.0 0.0 -828.0 0
1 0 0 0 0
0 0 0
2 NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN 0.0 0.0
0.0 0.0 -815.0 0
0 0 0 0 0
0 0 0
3 NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN 2.0 0.0
2.0 0.0 -617.0 0
1 0 0 0 0
0 0 0
4 NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN NaN NaN NaN
NaN NaN 0.0 0.0
0.0 0.0 -1106.0 0
0 0 0 0 0
1 0 0

```
   FLAG_DOCUMENT_11  FLAG_DOCUMENT_12  FLAG_DOCUMENT_13  FLAG_DOCUMENT_14
FLAG_DOCUMENT_15  FLAG_DOCUMENT_16  FLAG_DOCUMENT_17  FLAG_DOCUMENT_18
FLAG_DOCUMENT_19  FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR
AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON
AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                 0                 0                 0                 0
0                 0                 0                 0                 0
0                 0                         0.0                         0.0
0.0                        0.0                        0.0
1.0
1                 0                 0                 0                 0
0                 0                 0                 0                 0
0                 0                         0.0                         0.0
0.0                        0.0                        0.0
0.0
2                 0                 0                 0                 0
0                 0                 0                 0                 0
0                 0                         0.0                         0.0
0.0                        0.0                        0.0
0.0
3                 0                 0                 0                 0
0                 0                 0                 0                 0
0                 0                         NaN                         NaN
NaN                        NaN                        NaN
NaN
4                 0                 0                 0                 0
0                 0                 0                 0                 0
0                 0                         0.0                         0.0
0.0                        0.0                        0.0
0.0
```

0.1.1 (1.1) Getting the dimension of Rows and columns of the Application dataset:

[4]:
```
application.shape
```

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

[5]:
```
application.info()
```

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

0.1.2 (1.2) Extracting the column names and its dimensions too:

[6]:
```
col=list(application.columns)
col
```

[6]:
```
['SK_ID_CURR',
 'TARGET',
 'NAME_CONTRACT_TYPE',
 'CODE_GENDER',
 'FLAG_OWN_CAR',
 'FLAG_OWN_REALTY',
 'CNT_CHILDREN',
 'AMT_INCOME_TOTAL',
 'AMT_CREDIT',
 'AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'NAME_TYPE_SUITE',
 'NAME_INCOME_TYPE',
 'NAME_EDUCATION_TYPE',
 'NAME_FAMILY_STATUS',
 'NAME_HOUSING_TYPE',
 'REGION_POPULATION_RELATIVE',
 'DAYS_BIRTH',
 'DAYS_EMPLOYED',
 'DAYS_REGISTRATION',
 'DAYS_ID_PUBLISH',
 'OWN_CAR_AGE',
 'FLAG_MOBIL',
 'FLAG_EMP_PHONE',
 'FLAG_WORK_PHONE',
 'FLAG_CONT_MOBILE',
 'FLAG_PHONE',
 'FLAG_EMAIL',
 'OCCUPATION_TYPE',
 'CNT_FAM_MEMBERS',
 'REGION_RATING_CLIENT',
 'REGION_RATING_CLIENT_W_CITY',
 'WEEKDAY_APPR_PROCESS_START',
 'HOUR_APPR_PROCESS_START',
 'REG_REGION_NOT_LIVE_REGION',
 'REG_REGION_NOT_WORK_REGION',
 'LIVE_REGION_NOT_WORK_REGION',
 'REG_CITY_NOT_LIVE_CITY',
 'REG_CITY_NOT_WORK_CITY',
 'LIVE_CITY_NOT_WORK_CITY',
 'ORGANIZATION_TYPE',
 'EXT_SOURCE_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',
 'FLAG_DOCUMENT_2',
 'FLAG_DOCUMENT_3',
 'FLAG_DOCUMENT_4',
 'FLAG_DOCUMENT_5',
 'FLAG_DOCUMENT_6',
 'FLAG_DOCUMENT_7',
 'FLAG_DOCUMENT_8',
 'FLAG_DOCUMENT_9',
 'FLAG_DOCUMENT_10',
 'FLAG_DOCUMENT_11',
 'FLAG_DOCUMENT_12',
 'FLAG_DOCUMENT_13',
 'FLAG_DOCUMENT_14',
 'FLAG_DOCUMENT_15',
 'FLAG_DOCUMENT_16',
 'FLAG_DOCUMENT_17',
 'FLAG_DOCUMENT_18',
 'FLAG_DOCUMENT_19',
 'FLAG_DOCUMENT_20',
 'FLAG_DOCUMENT_21',
 'AMT_REQ_CREDIT_BUREAU_HOUR',
 'AMT_REQ_CREDIT_BUREAU_DAY',
 'AMT_REQ_CREDIT_BUREAU_WEEK',
 'AMT_REQ_CREDIT_BUREAU_MON',
 'AMT_REQ_CREDIT_BUREAU_QRT',
 'AMT_REQ_CREDIT_BUREAU_YEAR']
```

0.1.3 (1.3) Length of the columns stands out to be:

[7]:
```
len(col)
```

[7]:
```
122
```

0.1.4 (1.4) Checking the unique values of column (Name_Contract_Type):

[8]:
```
application['NAME_CONTRACT_TYPE'].unique()
```

[8]:
```
array(['Cash loans', 'Revolving loans'], dtype=object)
```

0.2 (2) Importing the previous application dataset:

[328]:
```
prev_app=pd.read_csv('previous_application.csv')
prev_app.head()
```

[328]:
```
   SK_ID_PREV  SK_ID_CURR NAME_CONTRACT_TYPE  AMT_ANNUITY  AMT_APPLICATION  \
AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE WEEKDAY_APPR_PROCESS_START  \
HOUR_APPR_PROCESS_START FLAG_LAST_APPL_PER_CONTRACT  NFLAG_LAST_APPL_IN_DAY  \
RATE_DOWN_PAYMENT  RATE_INTEREST_PRIMARY  RATE_INTEREST_PRIVILEGED  \
NAME_CASH_LOAN_PURPOSE NAME_CONTRACT_STATUS  DAYS_DECISION  \
NAME_PAYMENT_TYPE CODE_REJECT_REASON  NAME_TYPE_SUITE NAME_CLIENT_TYPE  \
NAME_GOODS_CATEGORY NAME_PORTFOLIO NAME_PRODUCT_TYPE            CHANNEL_TYPE  \
SELLERPLACE_AREA NAME_SELLER_INDUSTRY  CNT_PAYMENT NAME_YIELD_GROUP  \
PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE  \
DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION  \
NFLAG_INSURED_ON_APPROVAL
0      2030495      271877     Consumer loans     1730.430          17145.0
17145.0               0.0          17145.0                   SATURDAY
15                           Y                       1                0.0
0.182832                  0.867336                     XAP            Approved
-73  Cash through the bank                XAP               NaN         Repeater
Mobile             POS                XNA             Country-wide
35          Connectivity         12.0           middle  POS mobile with interest
365243.0           -42.0                      300.0          -42.0
-37.0                        0.0
1      2802425      108129         Cash loans    25188.615         607500.0
679671.0              NaN         607500.0                   THURSDAY
11                           Y                       1                NaN
NaN                       NaN                     XNA            Approved
-164                   XNA                XAP    Unaccompanied          Repeater
XNA           Cash              x-sell          Contact center              -1
XNA         36.0        low_action          Cash X-Sell: low          365243.0
-134.0                      916.0       365243.0          365243.0
1.0
2      2523466      122040         Cash loans    15060.735         112500.0
136444.5              NaN         112500.0                    TUESDAY
11                           Y                       1                NaN
NaN                       NaN                     XNA            Approved
-301   Cash through the bank                XAP  Spouse, partner          Repeater
XNA           Cash              x-sell  Credit and cash offices              -1
XNA         12.0              high          Cash X-Sell: high          365243.0
-271.0                       59.0       365243.0          365243.0
1.0
3      2819243      176158         Cash loans    47041.335         450000.0
470790.0              NaN         450000.0                     MONDAY
7                           Y                       1                NaN
NaN                       NaN                     XNA            Approved
```

```
-512  Cash through the bank               XAP              NaN          Repeater
XNA            Cash             x-sell  Credit and cash offices                 -1
XNA          12.0            middle       Cash X-Sell: middle             365243.0
-482.0                       -152.0         -182.0             -177.0
1.0
4     1784265       202054          Cash loans     31924.395          337500.0
404055.0                NaN         337500.0                      THURSDAY
9                            Y                       1                NaN
NaN                        NaN                Repairs              Refused
-781  Cash through the bank                HC              NaN          Repeater
XNA            Cash            walk-in  Credit and cash offices                 -1
XNA          24.0             high          Cash Street: high                  NaN
NaN                          NaN            NaN                NaN
NaN
```

0.2.1 (2.1) Examining the dimension of the previous application:

[10]:
```python
prev_app.shape
```

[10]:
```
(1670214, 37)
```

Rows=1670214 and Columns=37

[11]:
```python
prev_app.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
```

0.3 (3) Exploratory Data Analysis :

0.3.1 (3.1) Checking the null value percentages of Application Dataset:

(3.1.1) Finding the percentage of missing values in all columns of application:

[12]:
```
round(application.isnull().mean()*100,2).sort_values(ascending = False)
```

[12]:
```
COMMONAREA_MEDI                 69.87
COMMONAREA_AVG                  69.87
COMMONAREA_MODE                 69.87
NONLIVINGAPARTMENTS_MODE        69.43
NONLIVINGAPARTMENTS_MEDI        69.43
NONLIVINGAPARTMENTS_AVG         69.43
FONDKAPREMONT_MODE              68.39
LIVINGAPARTMENTS_MEDI           68.35
LIVINGAPARTMENTS_MODE           68.35
LIVINGAPARTMENTS_AVG            68.35
FLOORSMIN_MEDI                  67.85
FLOORSMIN_MODE                  67.85
FLOORSMIN_AVG                   67.85
YEARS_BUILD_MEDI                66.50
YEARS_BUILD_AVG                 66.50
YEARS_BUILD_MODE                66.50
OWN_CAR_AGE                     65.99
LANDAREA_MODE                   59.38
LANDAREA_AVG                    59.38
```

LANDAREA_MEDI	59.38
BASEMENTAREA_MEDI	58.52
BASEMENTAREA_AVG	58.52
BASEMENTAREA_MODE	58.52
EXT_SOURCE_1	56.38
NONLIVINGAREA_MEDI	55.18
NONLIVINGAREA_AVG	55.18
NONLIVINGAREA_MODE	55.18
ELEVATORS_MODE	53.30
ELEVATORS_AVG	53.30
ELEVATORS_MEDI	53.30
WALLSMATERIAL_MODE	50.84
APARTMENTS_MODE	50.75
APARTMENTS_AVG	50.75
APARTMENTS_MEDI	50.75
ENTRANCES_MEDI	50.35
ENTRANCES_MODE	50.35
ENTRANCES_AVG	50.35
LIVINGAREA_MEDI	50.19
LIVINGAREA_MODE	50.19
LIVINGAREA_AVG	50.19
HOUSETYPE_MODE	50.18
FLOORSMAX_MODE	49.76
FLOORSMAX_MEDI	49.76
FLOORSMAX_AVG	49.76
YEARS_BEGINEXPLUATATION_MEDI	48.78
YEARS_BEGINEXPLUATATION_AVG	48.78
YEARS_BEGINEXPLUATATION_MODE	48.78
TOTALAREA_MODE	48.27
EMERGENCYSTATE_MODE	47.40
OCCUPATION_TYPE	31.35
EXT_SOURCE_3	19.83
AMT_REQ_CREDIT_BUREAU_QRT	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	13.50
AMT_REQ_CREDIT_BUREAU_DAY	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	13.50
AMT_REQ_CREDIT_BUREAU_MON	13.50
AMT_REQ_CREDIT_BUREAU_HOUR	13.50
NAME_TYPE_SUITE	0.42
OBS_30_CNT_SOCIAL_CIRCLE	0.33
OBS_60_CNT_SOCIAL_CIRCLE	0.33
DEF_60_CNT_SOCIAL_CIRCLE	0.33
DEF_30_CNT_SOCIAL_CIRCLE	0.33
EXT_SOURCE_2	0.21
AMT_GOODS_PRICE	0.09
DAYS_ID_PUBLISH	0.00
FLAG_EMP_PHONE	0.00

FLAG_MOBIL	0.00
DAYS_EMPLOYED	0.00
FLAG_WORK_PHONE	0.00
FLAG_CONT_MOBILE	0.00
FLAG_PHONE	0.00
FLAG_EMAIL	0.00
DAYS_REGISTRATION	0.00
NAME_HOUSING_TYPE	0.00
DAYS_BIRTH	0.00
REGION_POPULATION_RELATIVE	0.00
REGION_RATING_CLIENT	0.00
NAME_FAMILY_STATUS	0.00
NAME_EDUCATION_TYPE	0.00
NAME_INCOME_TYPE	0.00
AMT_ANNUITY	0.00
AMT_CREDIT	0.00
AMT_INCOME_TOTAL	0.00
CNT_CHILDREN	0.00
FLAG_OWN_REALTY	0.00
FLAG_OWN_CAR	0.00
CODE_GENDER	0.00
NAME_CONTRACT_TYPE	0.00
TARGET	0.00
CNT_FAM_MEMBERS	0.00
REG_REGION_NOT_LIVE_REGION	0.00
REGION_RATING_CLIENT_W_CITY	0.00
FLAG_DOCUMENT_14	0.00
DAYS_LAST_PHONE_CHANGE	0.00
FLAG_DOCUMENT_2	0.00
FLAG_DOCUMENT_3	0.00
FLAG_DOCUMENT_4	0.00
FLAG_DOCUMENT_5	0.00
FLAG_DOCUMENT_6	0.00
FLAG_DOCUMENT_7	0.00
FLAG_DOCUMENT_8	0.00
FLAG_DOCUMENT_9	0.00
FLAG_DOCUMENT_10	0.00
FLAG_DOCUMENT_11	0.00
FLAG_DOCUMENT_12	0.00
FLAG_DOCUMENT_13	0.00
FLAG_DOCUMENT_15	0.00
WEEKDAY_APPR_PROCESS_START	0.00
FLAG_DOCUMENT_16	0.00
FLAG_DOCUMENT_17	0.00
FLAG_DOCUMENT_18	0.00
FLAG_DOCUMENT_19	0.00
FLAG_DOCUMENT_20	0.00

```
FLAG_DOCUMENT_21                 0.00
ORGANIZATION_TYPE                0.00
LIVE_CITY_NOT_WORK_CITY          0.00
REG_CITY_NOT_WORK_CITY           0.00
REG_CITY_NOT_LIVE_CITY           0.00
LIVE_REGION_NOT_WORK_REGION      0.00
REG_REGION_NOT_WORK_REGION       0.00
HOUR_APPR_PROCESS_START          0.00
SK_ID_CURR                       0.00
dtype: float64
```

(3.1.2) Removing all the columns of the application dataset having Null value percentage > 50% and keeping the remaining:

[13]:
```
application=application.loc[:,application.isnull().mean()<=0.5]
application.shape
```

[13]:
```
(307511, 81)
```

Previously, the number of columns was 122 and now its updated to 81.

(3.1.3) Getting the list of null values less then 15% and more than 0%:

[14]:
```
list(application.columns[(application.isnull().mean()<=0.15)&(application.
 ↪isnull().mean()>0.0)])
```

[14]:
```
['AMT_ANNUITY',
 'AMT_GOODS_PRICE',
 'NAME_TYPE_SUITE',
 'CNT_FAM_MEMBERS',
 'EXT_SOURCE_2',
 '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']
```

0.3.2 (3.2) Examining the Amt_Annuity column:

[15]:
```
application['AMT_ANNUITY'].isnull().value_counts()
```

[15]:
```
False    307499
True         12
```

```
Name: AMT_ANNUITY, dtype: int64
```

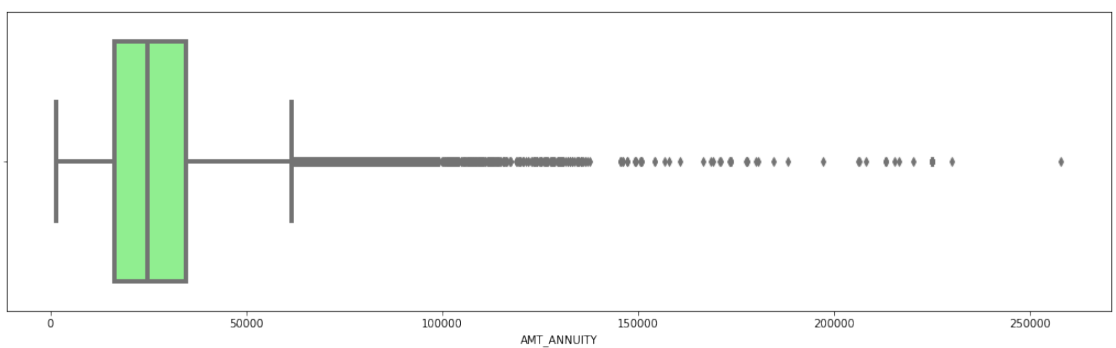
[16]:
```
application['AMT_ANNUITY'].unique()
```

[16]:
```
array([24700.5, 35698.5,  6750. , …, 71986.5, 58770. , 77809.5])
```

[17]:
```
application['AMT_ANNUITY'].value_counts()
```

[17]:
```
9000.0     6385
13500.0    5514
6750.0     2279
10125.0    2035
37800.0    1602
           …
15210.0       1
50265.0       1
73012.5       1
40558.5       1
4437.0        1
Name: AMT_ANNUITY, Length: 13672, dtype: int64
```

[18]:
```
f = plt.figure()
f.set_figwidth(18)
f.set_figheight(5)
sns.
↪boxplot(application['AMT_ANNUITY'],color='lightgreen',linewidth=4,saturation=50)
plt.show()
```



Here we can see that the column 'AMT_ANNUITY' have outliers, therefore the column can be altered using the median.

[19]:
```
print('The Median value for column: {} is: {}'.
↪format('AMT_ANNUITY',round(application['AMT_ANNUITY'].median())))
```

```
The Median value for column: AMT_ANNUITY is: 24903
```

0.3.3 (3.3) Examing the AMT_GOODS_PRICE Column:

[20]:
```
application['AMT_GOODS_PRICE'].isnull().value_counts()
```

[20]:
```
False    307233
True        278
Name: AMT_GOODS_PRICE, dtype: int64
```

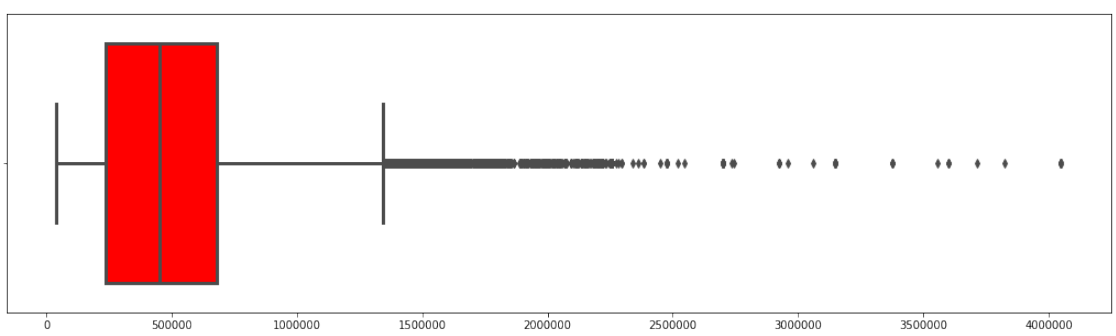
[21]:
```
application['AMT_GOODS_PRICE'].unique()
```

[21]:
```
array([ 351000. , 1129500. ,  135000. , …,  453465. ,  143977.5,
        743863.5])
```

[22]:
```
application['AMT_GOODS_PRICE'].value_counts()
```

[22]:
```
450000.0    26022
225000.0    25282
675000.0    24962
900000.0    15416
270000.0    11428
            …
705892.5        1
442062.0        1
353641.5        1
353749.5        1
738945.0        1
Name: AMT_GOODS_PRICE, Length: 1002, dtype: int64
```

[23]:
```
f = plt.figure()
f.set_figwidth(18)
f.set_figheight(5)
sns.
↪boxplot(application['AMT_GOODS_PRICE'],color='red',linewidth=3,saturation=50)
plt.show()
```



Here we can see that the column 'AMT_GOODS_PRICE' have outliers, therefore the column can be altered using the median.

[24]:
```
print('The Median value for column: {} is: {}'.
↪format('AMT_GOODS_PRICE',round(application['AMT_GOODS_PRICE'].median())))
```

```
The Median value for column: AMT_GOODS_PRICE is: 450000
```

0.3.4 (3.4) Examing the NAME_TYPE_SUITE Column:

[25]:
```
application['NAME_TYPE_SUITE'].value_counts()
```

[25]:
```
Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64
```

[26]:
```
application['NAME_TYPE_SUITE'].isnull().value_counts()
```

[26]:
```
False    306219
True       1292
Name: NAME_TYPE_SUITE, dtype: int64
```

[27]:
```
application['NAME_TYPE_SUITE'].unique()
```

[27]:
```
array(['Unaccompanied', 'Family', 'Spouse, partner', 'Children',
       'Other_A', nan, 'Other_B', 'Group of people'], dtype=object)
```

Since the column (Name_Type_Suite) holds a categoral value in it, so inorder to remove the outliers, we will update the null values with most repeated string kinda a mode(Name_Type_Suite)

[28]:
```
print('The Mode for column: {} is: {}'.
↪format('NAME_TYPE_SUITE',application['NAME_TYPE_SUITE'].mode()[0]))
```

```
The Mode for column: NAME_TYPE_SUITE is: Unaccompanied
```

0.3.5 (3.5) Examing CNT_FAM_MEMBERS column:

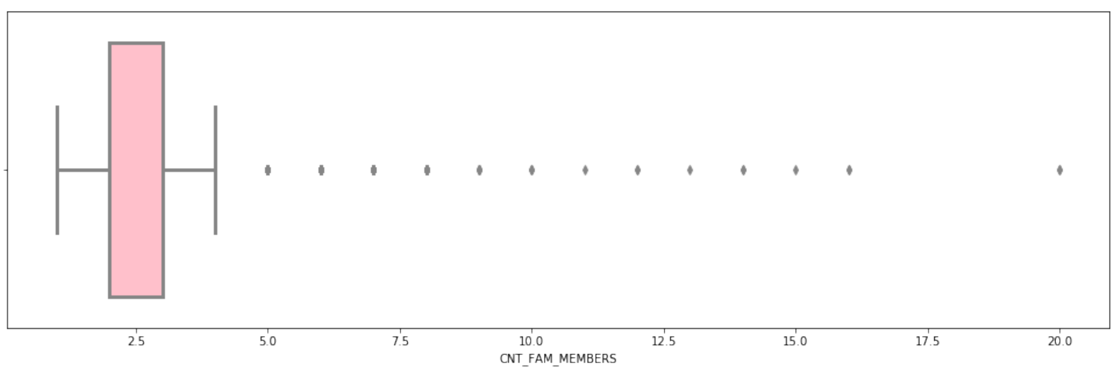
[29]:
```
application['CNT_FAM_MEMBERS'].isnull().value_counts()
```

[29]:
```
False    307509
True          2
Name: CNT_FAM_MEMBERS, dtype: int64
```

[30]:
```
application['CNT_FAM_MEMBERS'].value_counts()
```

```
[30]: 2.0      158357
      1.0       67847
      3.0       52601
      4.0       24697
      5.0        3478
      6.0         408
      7.0          81
      8.0          20
      9.0           6
      10.0          3
      14.0          2
      16.0          2
      12.0          2
      20.0          2
      11.0          1
      13.0          1
      15.0          1
      Name: CNT_FAM_MEMBERS, dtype: int64
```

```
[31]: f = plt.figure()
f.set_figwidth(17)
f.set_figheight(5)
sns.
 ↪boxplot(application['CNT_FAM_MEMBERS'],color='pink',linewidth=3,saturation=50)
plt.show()
```



Here we can see that the column 'CNT_FAM_MEMBERS' have outliers, therefore the column can be altered using the median.

```
[32]: print('The Median value for column: {} is: {}'.
 ↪format('CNT_FAM_MEMBERS',round(application['CNT_FAM_MEMBERS'].median())))
```

```
The Median value for column: CNT_FAM_MEMBERS is: 2
```

0.3.6 (3.6) EXT_SOURCE_2 column:

[33]:
```
application['EXT_SOURCE_2'].isnull().value_counts()
```

[33]:
```
False    306851
True        660
Name: EXT_SOURCE_2, dtype: int64
```

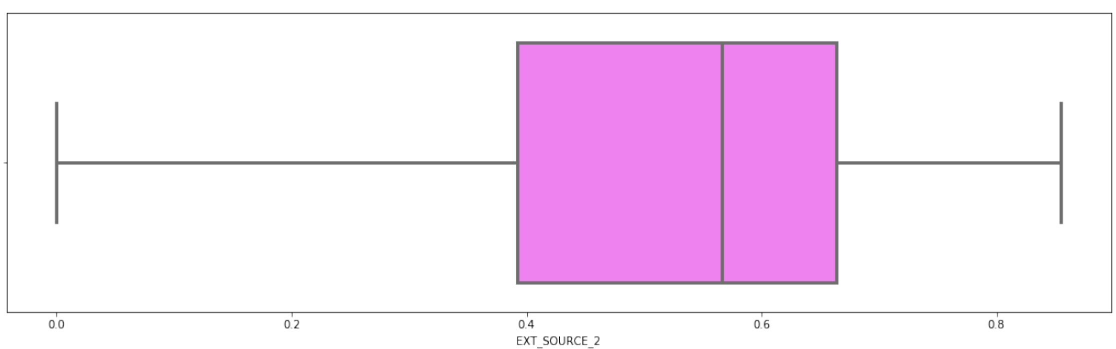
[34]:
```
application['EXT_SOURCE_2'].value_counts()
```

[34]:
```
0.285898    721
0.262258    417
0.265256    343
0.159679    322
0.265312    306
           …
0.169134      1
0.213753      1
0.057994      1
0.229146      1
0.336367      1
Name: EXT_SOURCE_2, Length: 119831, dtype: int64
```

[35]:
```
application['EXT_SOURCE_2'].unique()
```

[35]:
```
array([0.26294859, 0.62224578, 0.55591208, …, 0.13118876, 0.26448565,
       0.2678342 ])
```

[36]:
```
f = plt.figure()
f.set_figwidth(18)
f.set_figheight(5)
sns.
↪boxplot(application['EXT_SOURCE_2'],color='violet',linewidth=3,saturation=50)
plt.show()
```



Here we can see that the column 'EXT_SOURCE_2' have outliers, therefore the column can be altered using the median.

[37]:
```
print('The Median value for column: {} is: {}'.
↪format('EXT_SOURCE_2',round(application['EXT_SOURCE_2'].median())))
```

```
The Median value for column: EXT_SOURCE_2 is: 1
```

0.4 (4) Crosschecking the columns (its datatypes) of application dataset:

[77]:
```
#Checking the int type columns
A=application.select_dtypes(include='int64').columns
print(A)
print('\n')
print('length of the columns:',len(A))
```

```
Index(['SK_ID_CURR', 'TARGET', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT',
'REGION_POPULATION_RELATIVE', 'DAYS_BIRTH', 'DAYS_EMPLOYED',
'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH', 'FLAG_MOBIL', 'FLAG_EMP_PHONE',
'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE', 'FLAG_EMAIL',
'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY',
'HOUR_APPR_PROCESS_START', 'REG_REGION_NOT_LIVE_REGION',
'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION',
'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY',
'FLAG_DOCUMENT_2', 'FLAG_DOCUMENT_3', 'FLAG_DOCUMENT_4', 'FLAG_DOCUMENT_5',
'FLAG_DOCUMENT_6', 'FLAG_DOCUMENT_7', 'FLAG_DOCUMENT_8', 'FLAG_DOCUMENT_9',
'FLAG_DOCUMENT_10', 'FLAG_DOCUMENT_11', 'FLAG_DOCUMENT_12', 'FLAG_DOCUMENT_13',
'FLAG_DOCUMENT_14', 'FLAG_DOCUMENT_15', 'FLAG_DOCUMENT_16', 'FLAG_DOCUMENT_17',
'FLAG_DOCUMENT_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20', 'FLAG_DOCUMENT_21'],
dtype='object')


length of the columns: 45
```

[76]:
```
#Checking the float type columns
float=application.select_dtypes(include='float64').columns
print(float)
print('\n')
print('length of the columns:',len(float))
```

```
Index(['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2',
'EXT_SOURCE_3', 'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG',
'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE',
'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_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')


length of the columns: 23
```

[71]:
```python
#converting the float datatype of the column to the int datatype.
for i in float:
    application.loc[:,i]=application.loc[:,i].astype('int64',errors='ignore')
print("Updation Done!")
```

```
Updation Done!
```

[75]:
```python
application.select_dtypes(include='float64').columns
```

[75]:
```
Index(['AMT_ANNUITY', 'AMT_GOODS_PRICE', 'CNT_FAM_MEMBERS', 'EXT_SOURCE_2',
'EXT_SOURCE_3', 'YEARS_BEGINEXPLUATATION_AVG', 'FLOORSMAX_AVG',
'YEARS_BEGINEXPLUATATION_MODE', 'FLOORSMAX_MODE',
'YEARS_BEGINEXPLUATATION_MEDI', 'FLOORSMAX_MEDI', 'TOTALAREA_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')
```

[83]:
```python
len(list(application.select_dtypes(include='float64').columns))
```

[83]:
```
23
```

[86]:
```python
#Checking for columns as an object dtypes:
obj=application.select_dtypes('object').columns
print(obj)
print(len(obj))
```

```
Index(['NAME_CONTRACT_TYPE', 'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY',
'NAME_TYPE_SUITE', 'NAME_INCOME_TYPE', 'NAME_EDUCATION_TYPE',
'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'OCCUPATION_TYPE',
'WEEKDAY_APPR_PROCESS_START', 'ORGANIZATION_TYPE', 'EMERGENCYSTATE_MODE'],
dtype='object')
13
```

[96]:
```python
#converting the object dtypes to string dtypes:
for i in obj:
    application[i]=application[i].astype('str')
print('Updation Done!')
```

```
Updation Done!
```

[104]:
```python
application.head()
```

```
[104]:    SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR
      FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY
      AMT_GOODS_PRICE NAME_TYPE_SUITE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE
      NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  DAYS_BIRTH
      DAYS_EMPLOYED  DAYS_REGISTRATION  DAYS_ID_PUBLISH  FLAG_MOBIL  FLAG_EMP_PHONE
      FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL OCCUPATION_TYPE
      CNT_FAM_MEMBERS  REGION_RATING_CLIENT  REGION_RATING_CLIENT_W_CITY
      WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  REG_REGION_NOT_LIVE_REGION
      REG_REGION_NOT_WORK_REGION  LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY
      REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY       ORGANIZATION_TYPE
      EXT_SOURCE_2  EXT_SOURCE_3  YEARS_BEGINEXPLUATATION_AVG  FLOORSMAX_AVG
      YEARS_BEGINEXPLUATATION_MODE  FLOORSMAX_MODE  YEARS_BEGINEXPLUATATION_MEDI
      FLOORSMAX_MEDI  TOTALAREA_MODE EMERGENCYSTATE_MODE  OBS_30_CNT_SOCIAL_CIRCLE  \
      0      100002       1          Cash loans           M            N
      Y             0             202500      406597      24700.5         351000.0
      Unaccompanied          Working  Secondary / secondary special  Single / not
      married  House / apartment                           0       -9461
      -637              -3648            -2120           1               1
      0                 1           1           0        Laborers             1.0
      2                            2                  WEDNESDAY
      10                           0                            0
      0                       0                      0                         0
      Business Entity Type 3      0.262949      0.139376                     0.9722
      0.0833                      0.9722          0.0833
      0.9722          0.0833          0.0149                  No
      2.0
      1      100003       0          Cash loans           F            N
      N             0             270000     1293502      35698.5        1129500.0
      Family     State servant               Higher education               Married
      House / apartment                           0      -16765          -1188
      -1186             -291            1               1               0
      1           1           0      Core staff              2.0                    1
      1                     MONDAY                       11
      0                          0                            0
      0                       0                       0                    School
      0.622246            NaN                      0.9851          0.2917
      0.9851          0.2917                       0.9851          0.2917
      0.0714                  No                       1.0
      2      100004       0    Revolving loans           M            Y
      Y             0              67500      135000       6750.0         135000.0
      Unaccompanied          Working  Secondary / secondary special  Single / not
      married  House / apartment                           0      -19046
      -225              -4260            -2531           1               1
      1                 1           1           0        Laborers             1.0
      2                            2                     MONDAY
      9                           0                           0
      0                       0                      0                         0
```

```
Government      0.555912      0.729567                          NaN
NaN                          NaN             NaN                            NaN
NaN            NaN                 nan                         0.0
3      100006       0         Cash loans           F            N
Y             0            135000      312682      29686.5         297000.0
Unaccompanied          Working  Secondary / secondary special        Civil
marriage  House / apartment                         0      -19005
-3039              -9833            -2437            1               1
0                1           0           0       Laborers              2.0
2                           2                  WEDNESDAY
17                          0                           0
0                      0                       0                        0
Business Entity Type 3      0.650442           NaN                          NaN
NaN                          NaN             NaN                            NaN
NaN            NaN                 nan                         2.0
4      100007       0         Cash loans           M            N
Y             0            121500      513000      21865.5         513000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                         0      -19932
-3038              -4311            -3458            1               1
0                1           0           0      Core staff             1.0
2                           2                   THURSDAY
11                          0                           0
0                      0                       1                        1
Religion      0.322738           NaN                          NaN          NaN
NaN            NaN                           NaN             NaN
NaN                 nan                        0.0

   DEF_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE  FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  FLAG_DOCUMENT_4
FLAG_DOCUMENT_5  FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  FLAG_DOCUMENT_8
FLAG_DOCUMENT_9  FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  FLAG_DOCUMENT_12
FLAG_DOCUMENT_13  FLAG_DOCUMENT_14  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16
FLAG_DOCUMENT_17  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  FLAG_DOCUMENT_20
FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
0                      2.0                       2.0                       2.0
-1134.0              0                1                0                0
0                0                0                0                0
0                0                0                0                0
0                0                0                0                0
0                     0.0                         0.0
0.0                    0.0                         0.0
1.0
1                      0.0                       1.0                       0.0
-828.0               0                1                0                0
```

```
0                 0                 0                 0                  0
0                  0                  0                  0                  0
0                  0                  0                  0                  0
0                           0.0                           0.0
0.0                          0.0                          0.0
0.0
2                       0.0                         0.0                          0.0
-815.0                 0                 0                 0                 0
0                 0                 0                 0                  0
0                  0                  0                  0                  0
0                  0                  0                  0                  0
0                           0.0                           0.0
0.0                          0.0                          0.0
0.0
3                       0.0                         2.0                          0.0
-617.0                 0                 1                 0                 0
0                 0                 0                 0                  0
0                  0                  0                  0                  0
0                  0                  0                  0                  0
0                           NaN                           NaN
NaN                          NaN                          NaN
NaN
4                       0.0                         0.0                          0.0
-1106.0                 0                 0                 0                 0
0                 0                 1                 0                  0
0                  0                  0                  0                  0
0                  0                  0                  0                  0
0                           0.0                           0.0
0.0                          0.0                          0.0
0.0
```

0.5 (5) Getting the proportion of GENDER among the dataset:

[117]:
```
print(application['CODE_GENDER'].unique())
print('\n')
print(application['CODE_GENDER'].value_counts())
```

```
['M' 'F' 'XNA']


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

0.5.1 (5.1) Removing the XNA from the dataset due to its less frequency:

[121]:
```
application=application.where(application['CODE_GENDER']!='XNA')
print('XNA removed from the dataset, due to its less frequency:')
```

```
XNA removed from the dataset, due to its less frequency:
```

0.5.2 (5.2) Updated attribute of column 'CODE_GENDER':

[122]:
```
application['CODE_GENDER'].value_counts()
```

[122]:
```
F    202448
M    105059
Name: CODE_GENDER, dtype: int64
```

[123]:
```
application['CODE_GENDER'].replace(['M','F'],['Male','Female'],inplace=True)
application.head()
```

[123]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR
FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY
AMT_GOODS_PRICE NAME_TYPE_SUITE NAME_INCOME_TYPE            NAME_EDUCATION_TYPE
NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE  DAYS_BIRTH
DAYS_EMPLOYED  DAYS_REGISTRATION  DAYS_ID_PUBLISH  FLAG_MOBIL  FLAG_EMP_PHONE
FLAG_WORK_PHONE  FLAG_CONT_MOBILE  FLAG_PHONE  FLAG_EMAIL OCCUPATION_TYPE
CNT_FAM_MEMBERS  REGION_RATING_CLIENT  REGION_RATING_CLIENT_W_CITY
WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START  REG_REGION_NOT_LIVE_REGION
REG_REGION_NOT_WORK_REGION  LIVE_REGION_NOT_WORK_REGION  REG_CITY_NOT_LIVE_CITY
REG_CITY_NOT_WORK_CITY  LIVE_CITY_NOT_WORK_CITY       ORGANIZATION_TYPE
EXT_SOURCE_2  EXT_SOURCE_3  YEARS_BEGINEXPLUATATION_AVG  FLOORSMAX_AVG
YEARS_BEGINEXPLUATATION_MODE  FLOORSMAX_MODE  YEARS_BEGINEXPLUATATION_MEDI
FLOORSMAX_MEDI  TOTALAREA_MODE EMERGENCYSTATE_MODE  OBS_30_CNT_SOCIAL_CIRCLE  \
0    100002.0     1.0         Cash loans        Male            N
Y          0.0          202500.0    406597.0      24700.5          351000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                         0.0     -9461.0
-637.0             -3648.0            -2120.0         1.0             1.0
0.0                1.0         1.0         0.0        Laborers              1.0
2.0                         2.0                       WEDNESDAY
10.0                        0.0                          0.0
0.0                    0.0                    0.0                        0.0
Business Entity Type 3      0.262949      0.139376                       0.9722
0.0833                        0.9722          0.0833
0.9722          0.0833          0.0149                   No
2.0
1    100003.0     0.0         Cash loans      Female            N
N          0.0          270000.0   1293502.0      35698.5         1129500.0
Family     State servant               Higher education               Married
House / apartment                        0.0     -16765.0        -1188.0
```

```
-1186.0          -291.0        1.0             1.0              0.0
1.0        1.0        0.0      Core staff             2.0
1.0                      1.0                    MONDAY
11.0                     0.0                          0.0
0.0                    0.0                     0.0                      0.0
School      0.622246          NaN                         0.9851        0.2917
0.9851         0.2917                          0.9851         0.2917
0.0714                 No                         1.0
2    100004.0     0.0    Revolving loans        Male            Y
Y          0.0          67500.0    135000.0      6750.0         135000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                         0.0    -19046.0
-225.0           -4260.0          -2531.0         1.0             1.0
1.0              1.0         1.0         0.0        Laborers             1.0
2.0                      2.0                    MONDAY
9.0                     0.0                          0.0
0.0                    0.0                     0.0                      0.0
Government      0.555912      0.729567                        NaN
NaN                         NaN              NaN                           NaN
NaN            NaN                 nan                       0.0
3    100006.0     0.0         Cash loans      Female            N
Y          0.0          135000.0    312682.0      29686.5        297000.0
Unaccompanied          Working  Secondary / secondary special       Civil
marriage  House / apartment                         0.0    -19005.0
-3039.0           -9833.0          -2437.0         1.0             1.0
0.0              1.0         0.0         0.0        Laborers             2.0
2.0                      2.0                 WEDNESDAY
17.0                     0.0                          0.0
0.0                    0.0                     0.0                      0.0
Business Entity Type 3      0.650442          NaN                          NaN
NaN                         NaN              NaN                           NaN
NaN            NaN                 nan                       2.0
4    100007.0     0.0         Cash loans        Male            N
Y          0.0          121500.0    513000.0      21865.5        513000.0
Unaccompanied          Working  Secondary / secondary special  Single / not
married  House / apartment                         0.0    -19932.0
-3038.0           -4311.0          -3458.0         1.0             1.0
0.0              1.0         0.0         0.0      Core staff             1.0
2.0                      2.0                  THURSDAY
11.0                     0.0                          0.0
0.0                    0.0                     1.0                      1.0
Religion      0.322738          NaN                         NaN          NaN
NaN            NaN                         NaN              NaN
NaN                nan                        0.0

   DEF_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE  DEF_60_CNT_SOCIAL_CIRCLE
DAYS_LAST_PHONE_CHANGE  FLAG_DOCUMENT_2  FLAG_DOCUMENT_3  FLAG_DOCUMENT_4
```

```
FLAG_DOCUMENT_5  FLAG_DOCUMENT_6  FLAG_DOCUMENT_7  FLAG_DOCUMENT_8
FLAG_DOCUMENT_9  FLAG_DOCUMENT_10  FLAG_DOCUMENT_11  FLAG_DOCUMENT_12
FLAG_DOCUMENT_13  FLAG_DOCUMENT_14  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16
FLAG_DOCUMENT_17  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  FLAG_DOCUMENT_20
FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
AMT_REQ_CREDIT_BUREAU_YEAR
0                       2.0                         2.0                         2.0
-1134.0             0.0               1.0               0.0               0.0
0.0               0.0               0.0               0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                         0.0                         0.0
0.0                        0.0                         0.0
1.0
1                       0.0                         1.0                         0.0
-828.0              0.0               1.0               0.0               0.0
0.0               0.0               0.0               0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                         0.0                         0.0
0.0                        0.0                         0.0
0.0
2                       0.0                         0.0                         0.0
-815.0              0.0               0.0               0.0               0.0
0.0               0.0               0.0               0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                         0.0                         0.0
0.0                        0.0                         0.0
0.0
3                       0.0                         2.0                         0.0
-617.0              0.0               1.0               0.0               0.0
0.0               0.0               0.0               0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                         NaN                         NaN
NaN                        NaN                         NaN
NaN
4                       0.0                         0.0                         0.0
-1106.0             0.0               0.0               0.0               0.0
0.0               0.0               1.0               0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                0.0                0.0                0.0                0.0
0.0                         0.0                         0.0
0.0                        0.0                         0.0
0.0
```

0.6 (6) Binning variable for analysis:

[128]:
```
application['AMT_INCOME_TOTAL'].quantile([0,0.1,0.3,0.6,0.8,1])
```

[128]:
```
0.0        25650.0
0.1        81000.0
0.3       112500.0
0.6       162000.0
0.8       225000.0
1.0    117000000.0
Name: AMT_INCOME_TOTAL, dtype: float64
```

[130]:
```
#Creating A new categorical variable based on income total
application['INCOME_GROUP']=pd.qcut(application['AMT_INCOME_TOTAL'],
                                       q=[0,0.1,0.3,0.6,0.8,1],
                                       ␣
 ↪labels=['VeryLow','Low','Medium','High','VeryHigh'])
print("Done!")
```

```
Done!
```

[131]:
```
application['INCOME_GROUP'].head()
```

[131]:
```
0        High
1    VeryHigh
2     VeryLow
3      Medium
4      Medium
Name: INCOME_GROUP, dtype: category
Categories (5, object): [VeryLow < Low < Medium < High < VeryHigh]
```

0.6.1 (6.1) Binning Birth Date:

[138]:
```
application['DAYS_BIRTH'].head()
```

[138]:
```
0    -9461.0
1   -16765.0
2   -19046.0
3   -19005.0
4   -19932.0
Name: DAYS_BIRTH, dtype: float64
```

[134]:
```
#Binning DAYS_BIRTH
abs(application['DAYS_BIRTH']).quantile([0,0.1,0.3,0.6,0.8,1])
```

[134]:
```
0.0     7489.0
0.1    10284.6
0.3    13140.0
```

```
0.6    17220.0
0.8    20474.0
1.0    25229.0
Name: DAYS_BIRTH, dtype: float64
```

Since DAYS_BIRTH consist negative values, hence we will use abs to typecast it into positive value:

0.6.2 (6.2) Creating a column AGE using the Days_Birth column for future reference:

[147]:
```
application['AGE']=abs(application['DAYS_BIRTH'])//365.25
application['AGE'].head(10)
```

[147]:
```
0    25.0
1    45.0
2    52.0
3    52.0
4    54.0
5    46.0
6    37.0
7    51.0
8    55.0
9    39.0
Name: AGE, dtype: float64
```

(6.2.1) Now lets analyise the AGE dataset:

[151]:
```
application['AGE'].describe()
```

[151]:
```
count    307507.000000
mean         43.405223
std          11.945763
min          20.000000
25%          33.000000
50%          43.000000
75%          53.000000
max          69.000000
Name: AGE, dtype: float64
```

Here we can see that the min age is 20 and max age is 69 (70approx)

NOTE: Since the age is varrying from 20 to 70, we would create a bins of approx length 5 each:

[152]:
```
application['AGE_GROUP'] = pd.cut(application['AGE'],bins=np.arange(20,71,5))
#Here 20 is the starting point and 71 is the ending and the difference 5 each
```

[155]:
```
application['AGE_GROUP'].head(10)
```

[155]:
```
0    (20, 25]
1    (40, 45]
2    (50, 55]
3    (50, 55]
4    (50, 55]
5    (45, 50]
6    (35, 40]
7    (50, 55]
8    (50, 55]
9    (35, 40]
Name: AGE_GROUP, dtype: category
Categories (10, interval[int64]): [(20, 25] < (25, 30] < (30, 35] < (35, 40] …
(50, 55] < (55, 60] < (60, 65] < (65, 70]]
```

0.7 (7) Adding a new column to the application dataframe for future use:

[156]:
```
application['CREDIT_INCOME_RATIO']=round((application['AMT_CREDIT']/
↪application['AMT_INCOME_TOTAL']))
```

[157]:
```
application['CREDIT_INCOME_RATIO'].head(10)
```

[157]:
```
0    2.0
1    5.0
2    2.0
3    2.0
4    4.0
5    5.0
6    9.0
7    4.0
8    9.0
9    3.0
Name: CREDIT_INCOME_RATIO, dtype: float64
```

[158]:
```
# Getting the percentage of social circle who defaulted for 30 and 60 days each
application['SOCIAL_CIRCLE_30_DAYS_DEF_PERC']=application['DEF_30_CNT_SOCIAL_CIRCLE']/
↪application['OBS_30_CNT_SOCIAL_CIRCLE']
application['SOCIAL_CIRCLE_60_DAYS_DEF_PERC']=application['DEF_60_CNT_SOCIAL_CIRCLE']/
↪application['OBS_60_CNT_SOCIAL_CIRCLE']
```

[160]:
```
application['SOCIAL_CIRCLE_30_DAYS_DEF_PERC'].head()
```

[160]:
```
0    1.0
1    0.0
2    NaN
3    0.0
4    NaN
Name: SOCIAL_CIRCLE_30_DAYS_DEF_PERC, dtype: float64
```

[161]:
```
application['SOCIAL_CIRCLE_60_DAYS_DEF_PERC'].head()
```

[161]:
```
0    1.0
1    0.0
2    NaN
3    0.0
4    NaN
Name: SOCIAL_CIRCLE_60_DAYS_DEF_PERC, dtype: float64
```

0.8 (8) Checking for imbalance in Target attribute:

[165]:
```
application['TARGET'].value_counts()
```

[165]:
```
0.0    282682
1.0     24825
Name: TARGET, dtype: int64
```

[173]:
```
application['TARGET'].count()
```

[173]:
```
307507
```

[178]:
```
round(application['TARGET'].value_counts()[0]/application['TARGET'].count()*100)
```

[178]:
```
92.0
```

[179]:
```
round(application['TARGET'].value_counts()[1]/application['TARGET'].count()*100)
```

[179]:
```
8.0
```

NOTE: Performing the above steps using normalization:

[182]:
```
application['TARGET'].value_counts(normalize=True)*100
```

[182]:
```
0.0    91.927013
1.0     8.072987
Name: TARGET, dtype: float64
```

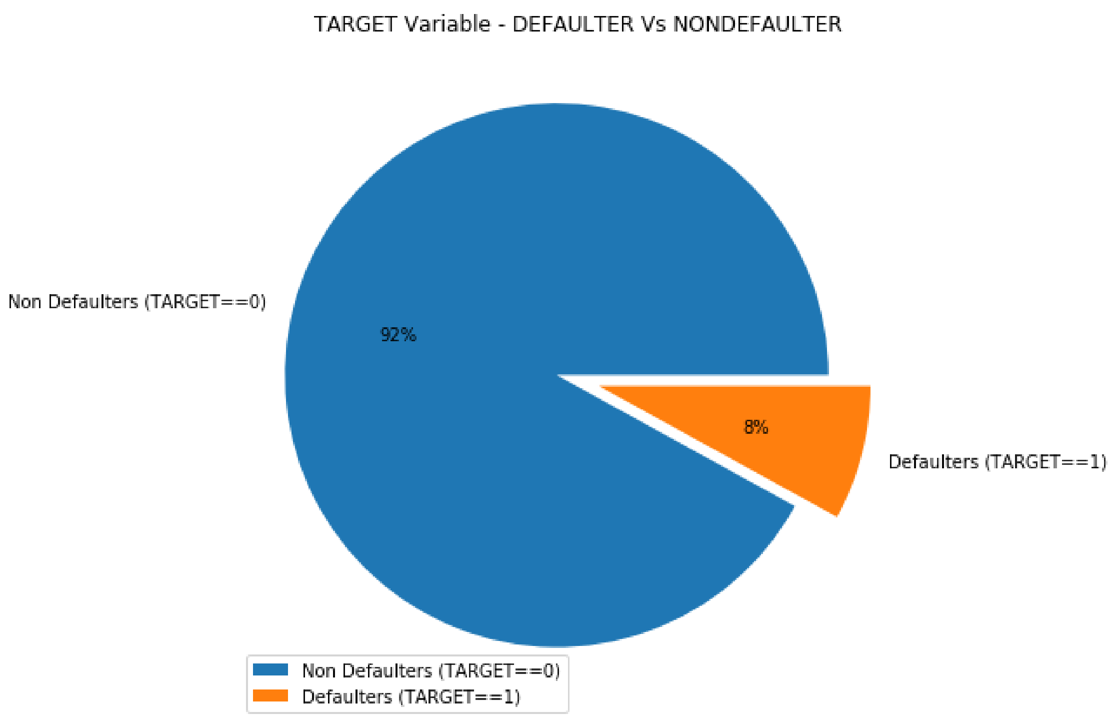
[184]:
```
A=round(application['TARGET'].value_counts(normalize=True)*100)
print(A)
```

```
0.0    92.0
1.0     8.0
Name: TARGET, dtype: float64
```

Ploting a pie chart for visualisation:

[228]:
```
f = plt.figure()
f.set_figwidth(7)
f.set_figheight(7.5)
```

```
plt.pie(A,labels=['Non Defaulters (TARGET==0)','Defaulters␣
↪(TARGET==1)'],explode=(0.08,0.08),autopct='%1.f%%')
plt.legend()
plt.title('TARGET Variable - DEFAULTER Vs NONDEFAULTER')
plt.show()
```



NOTE: Here we can visualise that approx 8% people defaulted their loan by not paying any installment, where as approx 92% people were geniuengly paying the sum:

0.9 (9) Performind descriptive analysis:

Removing the unwanted columns from the application dataset and keeping the important attributes only:

[229]:
```
FinalColumns =␣
↪['SK_ID_CURR','TARGET','CODE_GENDER','FLAG_OWN_CAR','FLAG_OWN_REALTY','INCOME_GROUP','AGE_G
'CREDIT_INCOME_RATIO','NAME_INCOME_TYPE','NAME_EDUCATION_TYPE','NAME_FAMILY_STATUS','NAME_HOUS
'DAYS_REGISTRATION','FLAG_EMAIL','OCCUPATION_TYPE',
'CNT_FAM_MEMBERS','REGION_RATING_CLIENT_W_CITY','ORGANIZATION_TYPE','SOCIAL_CIRCLE_30_DAYS_DEF
'SOCIAL_CIRCLE_60_DAYS_DEF_PERC','AMT_REQ_CREDIT_BUREAU_DAY',
'AMT_REQ_CREDIT_BUREAU_MON','AMT_REQ_CREDIT_BUREAU_QRT','NAME_CONTRACT_TYPE','AMT_ANNUITY','RE
```

[230]:
```
application_final=application[FinalColumns]
```

[232]:
```
application_final.head()
```

[232]:
```
   SK_ID_CURR  TARGET CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY INCOME_GROUP
AGE_GROUP  AMT_CREDIT  AMT_INCOME_TOTAL  CREDIT_INCOME_RATIO NAME_INCOME_TYPE
NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  DAYS_EMPLOYED
DAYS_REGISTRATION  FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS
REGION_RATING_CLIENT_W_CITY       ORGANIZATION_TYPE
SOCIAL_CIRCLE_30_DAYS_DEF_PERC  SOCIAL_CIRCLE_60_DAYS_DEF_PERC
AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
NAME_CONTRACT_TYPE  AMT_ANNUITY  REGION_RATING_CLIENT  AMT_GOODS_PRICE
0     100002.0     1.0         Male             N               Y         High
(20, 25]     406597.0          202500.0                    2.0          Working
Secondary / secondary special  Single / not married  House / apartment
-637.0              -3648.0          0.0         Laborers              1.0
2.0  Business Entity Type 3                           1.0
1.0                        0.0                           0.0
0.0          Cash loans      24700.5                   2.0          351000.0
1     100003.0     0.0       Female             N               N     VeryHigh
(40, 45]    1293502.0          270000.0                    5.0    State servant
Higher education                Married  House / apartment        -1188.0
-1186.0         0.0       Core staff             2.0
1.0                   School                          0.0
0.0                        0.0                           0.0
0.0          Cash loans      35698.5                   1.0         1129500.0
2     100004.0     0.0         Male             Y               Y      VeryLow
(50, 55]     135000.0           67500.0                    2.0          Working
Secondary / secondary special  Single / not married  House / apartment
-225.0              -4260.0          0.0         Laborers              1.0
2.0              Government                           NaN
NaN                        0.0                           0.0
0.0    Revolving loans       6750.0                   2.0          135000.0
3     100006.0     0.0       Female             N               Y       Medium
(50, 55]     312682.0          135000.0                    2.0          Working
Secondary / secondary special        Civil marriage  House / apartment
-3039.0              -9833.0          0.0         Laborers              2.0
2.0  Business Entity Type 3                           0.0
0.0                        NaN                           NaN
NaN          Cash loans      29686.5                   2.0          297000.0
4     100007.0     0.0         Male             N               Y       Medium
(50, 55]     513000.0          121500.0                    4.0          Working
Secondary / secondary special  Single / not married  House / apartment
-3038.0              -4311.0          0.0      Core staff              1.0
2.0                 Religion                          NaN
NaN                        0.0                           0.0
0.0          Cash loans      21865.5                   2.0          513000.0
```

[233]:
```
application_final.shape
```

[233]:
```
(307511, 30)
```

Dividing the application datasheet on the basis of Target Values; example (Non-defaulter and Defaulter)

[271]:
```
#dataset for non-defaulter
app_final_nondef=application_final[application['TARGET']==0]
app_final_nondef.head(10)
```

[271]:
```
    SK_ID_CURR  TARGET CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY INCOME_GROUP  \
AGE_GROUP  AMT_CREDIT  AMT_INCOME_TOTAL  CREDIT_INCOME_RATIO
NAME_INCOME_TYPE            NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS
NAME_HOUSING_TYPE  DAYS_EMPLOYED  DAYS_REGISTRATION  FLAG_EMAIL OCCUPATION_TYPE
CNT_FAM_MEMBERS  REGION_RATING_CLIENT_W_CITY       ORGANIZATION_TYPE
SOCIAL_CIRCLE_30_DAYS_DEF_PERC  SOCIAL_CIRCLE_60_DAYS_DEF_PERC
AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
NAME_CONTRACT_TYPE  AMT_ANNUITY  REGION_RATING_CLIENT  AMT_GOODS_PRICE
1      100003.0     0.0       Female            N               N     VeryHigh
(40, 45]   1293502.0          270000.0                   5.0         State
servant               Higher education               Married  House / apartment
-1188.0            -1186.0         0.0      Core staff               2.0
1.0                  School                             0.0
0.0                           0.0                           0.0
0.0          Cash loans      35698.5                    1.0       1129500.0
2      100004.0     0.0         Male            Y               Y      VeryLow
(50, 55]     135000.0           67500.0                   2.0
Working  Secondary / secondary special  Single / not married  House / apartment
-225.0            -4260.0         0.0        Laborers               1.0
2.0             Government                              NaN
NaN                           0.0                           0.0
0.0     Revolving loans       6750.0                    2.0        135000.0
3      100006.0     0.0       Female            N               Y       Medium
(50, 55]     312682.0          135000.0                   2.0
Working  Secondary / secondary special        Civil marriage  House / apartment
-3039.0            -9833.0         0.0        Laborers               2.0
2.0  Business Entity Type 3                             0.0
0.0                           NaN                           NaN
NaN          Cash loans      29686.5                    2.0        297000.0
4      100007.0     0.0         Male            N               Y       Medium
(50, 55]     513000.0          121500.0                   4.0
Working  Secondary / secondary special  Single / not married  House / apartment
-3038.0            -4311.0         0.0      Core staff               1.0
2.0               Religion                              NaN
NaN                           0.0                           0.0
0.0          Cash loans      21865.5                    2.0        513000.0
5      100008.0     0.0         Male            N               Y          Low
(45, 50]     490495.0           99000.0                   5.0         State
servant  Secondary / secondary special               Married  House / apartment
```

```
-1588.0            -4970.0         0.0       Laborers             2.0
2.0                   Other                              NaN
NaN                        0.0                          0.0
1.0         Cash loans      27517.5                     2.0         454500.0
6     100009.0      0.0       Female            Y               Y          High
(35, 40]    1560726.0           171000.0                  9.0  Commercial
associate                 Higher education               Married  House /
apartment        -3130.0            -1213.0          0.0      Accountants
3.0                          2.0  Business Entity Type 3
0.0                           0.0                          0.0
1.0                        1.0          Cash loans      41301.0
2.0         1395000.0
7     100010.0      0.0         Male            Y               Y     VeryHigh
(50, 55]    1530000.0           360000.0                  4.0          State
servant                   Higher education             Married  House / apartment
-449.0             -4597.0         0.0         Managers             2.0
3.0                   Other                              0.0
0.0                        0.0                          0.0
0.0         Cash loans      42075.0                     3.0        1530000.0
8     100011.0      0.0       Female            N               Y           Low
(50, 55]    1019610.0           112500.0                  9.0
Pensioner  Secondary / secondary special               Married  House /
apartment       365243.0            -7427.0          0.0              nan
2.0                          2.0                       XNA
0.0                           0.0                          0.0
0.0                        0.0          Cash loans      33826.5
2.0          913500.0
9     100012.0      0.0         Male            N               Y        Medium
(35, 40]     405000.0           135000.0                  3.0
Working  Secondary / secondary special  Single / not married  House / apartment
-2019.0            -14437.0         0.0       Laborers             1.0
2.0            Electricity                               0.0
0.0                        NaN                          NaN
NaN    Revolving loans      20250.0                     2.0         405000.0
10    100014.0      0.0       Female            N               Y           Low
(25, 30]     652500.0           112500.0                  6.0
Working                   Higher education             Married  House / apartment
-679.0             -4427.0         0.0      Core staff             3.0
2.0                Medicine                              NaN
NaN                        0.0                          1.0
0.0         Cash loans      21177.0                     2.0         652500.0
```

[272]:
```
app_final_nondef.shape
```

[272]:
```
(282682, 30)
```

```
[273]: A=app_final_nondef[app_final_nondef['SK_ID_CURR'].isnull()!=True]
       A.shape
```

```
[273]: (282682, 30)
```

```
[274]: A.head()
```

```
[274]:    SK_ID_CURR  TARGET CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY INCOME_GROUP
       AGE_GROUP  AMT_CREDIT  AMT_INCOME_TOTAL  CREDIT_INCOME_RATIO NAME_INCOME_TYPE
       NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  DAYS_EMPLOYED
       DAYS_REGISTRATION  FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS
       REGION_RATING_CLIENT_W_CITY       ORGANIZATION_TYPE
       SOCIAL_CIRCLE_30_DAYS_DEF_PERC  SOCIAL_CIRCLE_60_DAYS_DEF_PERC
       AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
       NAME_CONTRACT_TYPE  AMT_ANNUITY  REGION_RATING_CLIENT  AMT_GOODS_PRICE
       1     100003.0     0.0      Female             N               N     VeryHigh
       (40, 45]   1293502.0          270000.0                  5.0    State servant
       Higher education               Married  House / apartment        -1188.0
       -1186.0         0.0      Core staff             2.0
       1.0                  School                            0.0
       0.0                      0.0                           0.0
       0.0         Cash loans      35698.5                    1.0        1129500.0
       2     100004.0     0.0        Male               Y               Y      VeryLow
       (50, 55]    135000.0           67500.0                  2.0          Working
       Secondary / secondary special  Single / not married  House / apartment
       -225.0           -4260.0         0.0         Laborers              1.0
       2.0              Government                            NaN
       NaN                      0.0                           0.0
       0.0    Revolving loans       6750.0                    2.0         135000.0
       3     100006.0     0.0      Female               N               Y       Medium
       (50, 55]    312682.0          135000.0                  2.0          Working
       Secondary / secondary special        Civil marriage  House / apartment
       -3039.0           -9833.0         0.0         Laborers              2.0
       2.0  Business Entity Type 3                            0.0
       0.0                      NaN                           NaN
       NaN         Cash loans      29686.5                    2.0         297000.0
       4     100007.0     0.0        Male               N               Y       Medium
       (50, 55]    513000.0          121500.0                  4.0          Working
       Secondary / secondary special  Single / not married  House / apartment
       -3038.0           -4311.0         0.0       Core staff              1.0
       2.0                Religion                            NaN
       NaN                      0.0                           0.0
       0.0         Cash loans      21865.5                    2.0         513000.0
       5     100008.0     0.0        Male               N               Y          Low
       (45, 50]    490495.0           99000.0                  5.0    State servant
       Secondary / secondary special                Married  House / apartment
       -1588.0           -4970.0         0.0         Laborers              2.0
```

```
2.0                   Other                              NaN
NaN                        0.0                          0.0
1.0         Cash loans      27517.5                     2.0         454500.0
```

[275]:
```python
# Dataset for defaulter:
app_final_def=application_final[application_final['TARGET']==1]
app_final_def.head(10)
```

[275]:
```
     SK_ID_CURR  TARGET CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY INCOME_GROUP
AGE_GROUP  AMT_CREDIT  AMT_INCOME_TOTAL  CREDIT_INCOME_RATIO
NAME_INCOME_TYPE            NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS
NAME_HOUSING_TYPE  DAYS_EMPLOYED  DAYS_REGISTRATION  FLAG_EMAIL
OCCUPATION_TYPE  CNT_FAM_MEMBERS  REGION_RATING_CLIENT_W_CITY
ORGANIZATION_TYPE  SOCIAL_CIRCLE_30_DAYS_DEF_PERC
SOCIAL_CIRCLE_60_DAYS_DEF_PERC  AMT_REQ_CREDIT_BUREAU_DAY
AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT NAME_CONTRACT_TYPE
AMT_ANNUITY  REGION_RATING_CLIENT  AMT_GOODS_PRICE
0      100002.0     1.0         Male            N               Y         High
(20, 25]    406597.0          202500.0                 2.0
Working  Secondary / secondary special  Single / not married  House / apartment
-637.0            -3648.0         0.0               Laborers              1.0
2.0  Business Entity Type 3                           1.0
1.0                        0.0                          0.0
0.0         Cash loans      24700.5                     2.0         351000.0
26     100031.0     1.0       Female            N               Y          Low
(50, 55]    979992.0          112500.0                 9.0
Working  Secondary / secondary special                 Widow  House / apartment
-2628.0             -6573.0         0.0          Cooking staff             1.0
2.0  Business Entity Type 3                           0.1
0.0                        0.0                          0.0
2.0         Cash loans      27076.5                     3.0         702000.0
40     100047.0     1.0         Male            N               Y         High
(45, 50]   1193580.0          202500.0                 6.0  Commercial
associate  Secondary / secondary special              Married  House /
apartment        -1262.0            -1182.0         0.0               Laborers
2.0                           2.0  Business Entity Type 3
NaN                           NaN                          0.0
2.0                        0.0        Cash loans      35028.0
2.0         855000.0
42     100049.0     1.0       Female            N               N       Medium
(35, 40]    288873.0          135000.0                 2.0
Working  Secondary / secondary special        Civil marriage  House / apartment
-3597.0               -45.0         0.0            Sales staff             2.0
3.0          Self-employed                                0.0
0.0                        0.0                          0.0
0.0         Cash loans      16258.5                     3.0         238500.0
81     100096.0     1.0       Female            N               Y      VeryLow
```

```
(65, 70]    252000.0          81000.0                  3.0
Pensioner  Secondary / secondary special              Married  House /
apartment       365243.0            -5391.0         0.0                    nan
2.0                         2.0                      XNA
1.0                            1.0                           0.0
0.0                       0.0         Cash loans      14593.5
2.0         252000.0
94     100112.0     1.0         Male            Y               Y     VeryHigh
(25, 30]    953460.0         315000.0                  3.0  Commercial
associate             Incomplete higher  Single / not married       With
parents         -2015.0             -4802.0         0.0                    nan
1.0                         2.0        Industry: type 4
NaN                            NaN                           0.0
0.0                       0.0         Cash loans      64107.0
2.0         900000.0
110    100130.0     1.0       Female            N               Y       Medium
(25, 30]    723996.0         157500.0                  5.0  Commercial
associate             Incomplete higher             Separated  House /
apartment         -267.0             -387.0         0.0            Sales staff
2.0                         2.0           Trade: type 2
NaN                            NaN                           0.0
0.0                       0.0         Cash loans      30802.5
2.0         585000.0
138    100160.0     1.0         Male            N               Y     VeryHigh
(40, 45]    675000.0         292500.0                  2.0
Working              Higher education              Married  House / apartment
-200.0             -5239.0         0.0               Managers              2.0
2.0  Business Entity Type 3                              NaN
NaN                       0.0                           0.0
0.0         Cash loans      36747.0                     2.0         675000.0
154    100181.0     1.0       Female            N               Y       Medium
(45, 50]    245619.0         157500.0                  2.0
Working  Secondary / secondary special  Single / not married  House / apartment
-7676.0             -774.0         0.0  Private service staff              1.0
2.0  Business Entity Type 3                              NaN
NaN                       0.0                           0.0
0.0         Cash loans      12667.5                     2.0         166500.0
163    100192.0     1.0       Female            N               N          Low
(20, 25]    225000.0         111915.0                  2.0  Commercial
associate  Secondary / secondary special  Single / not married       With
parents          -150.0             -2570.0         0.0            Core staff
1.0                         2.0           Trade: type 3
NaN                            NaN                           0.0
0.0                       0.0         Cash loans      21037.5
2.0         225000.0
```

[276]:
```python
app_final_def.shape
```

[276]: (24825, 30)

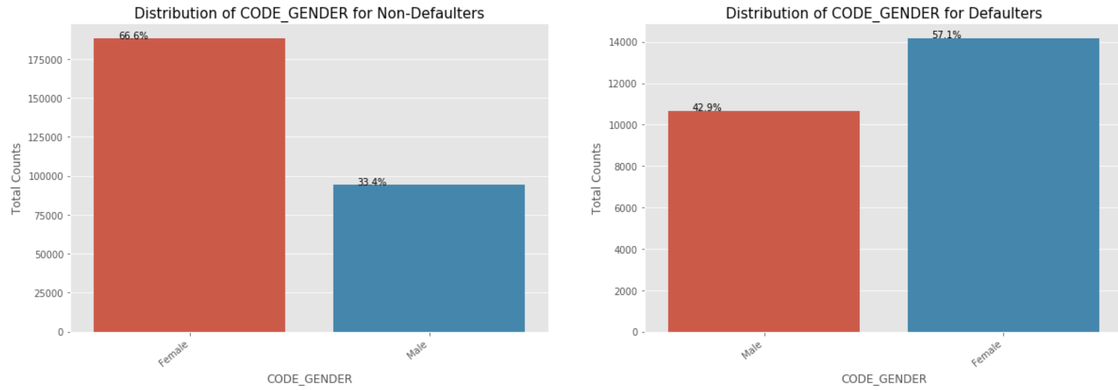
0.10 (10) Univariate Analysis:

0.10.1 (10.1) Function to plot univariate variables:

[285]:
```
def plot_univ(D):
    plt.style.use('ggplot')
    sns.despine
    fig,(ax1,ax2) = plt.subplots(1,2,figsize=(20,6))

    sns.countplot(x=D, data=app_final_nondef,ax=ax1)
    ax1.set_ylabel('Total Counts')
    ax1.set_title(f'Distribution of {D} for Non-Defaulters',fontsize=15)
    ax1.set_xticklabels(ax1.get_xticklabels(), rotation=40, ha="right")

    # Adding the normalized percentage for easier comparision between defaulter␣
↪and non-defaulter
    for p in ax1.patches:
        ax1.annotate('{:.1f}%'.format((p.get_height()/
↪len(app_final_nondef))*100), (p.get_x()+0.1, p.get_height()+50))

    sns.countplot(x=D, data=app_final_def,ax=ax2)
    ax2.set_ylabel('Total Counts')
    ax2.set_title(f'Distribution of {D} for Defaulters',fontsize=15)
    ax2.set_xticklabels(ax2.get_xticklabels(), rotation=40, ha="right")

    # Adding the normalized percentage for easier comparision between defaulter␣
↪and non-defaulter
    for p in ax2.patches:
        ax2.annotate('{:.1f}%'.format((p.get_height()/len(app_final_def))*100),␣
↪(p.get_x()+0.1, p.get_height()+50))

    plt.show()
```

0.10.2 (10.2) Plotting the CODE Gender:

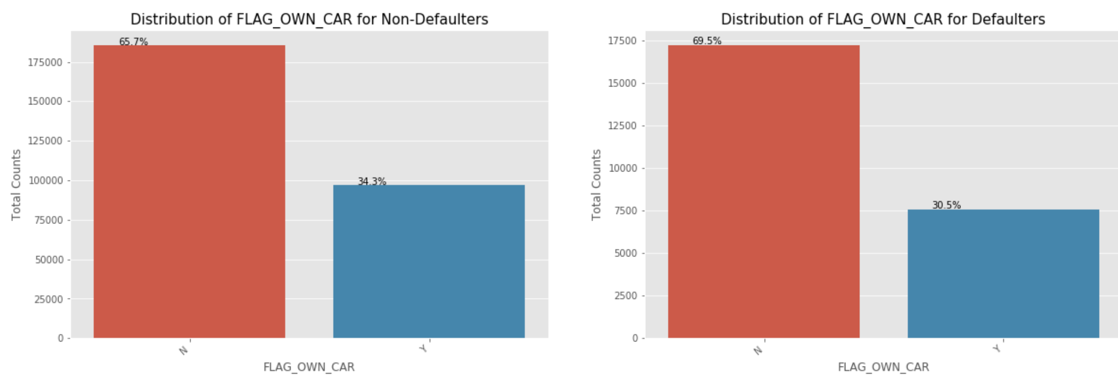
[286]:
```
plot_univ('CODE_GENDER')
```



NOTE: We can see that Female contribute 67% to the non-defaulters while 57% to the defaulters. We see more female applying for loans than males and hence the more number of female defaulters as well. **But the rate of defaulting of FEMALE is much lower compared to their MALE counterparts.**

0.10.3 (10.3) Plotting Flag_own_car:

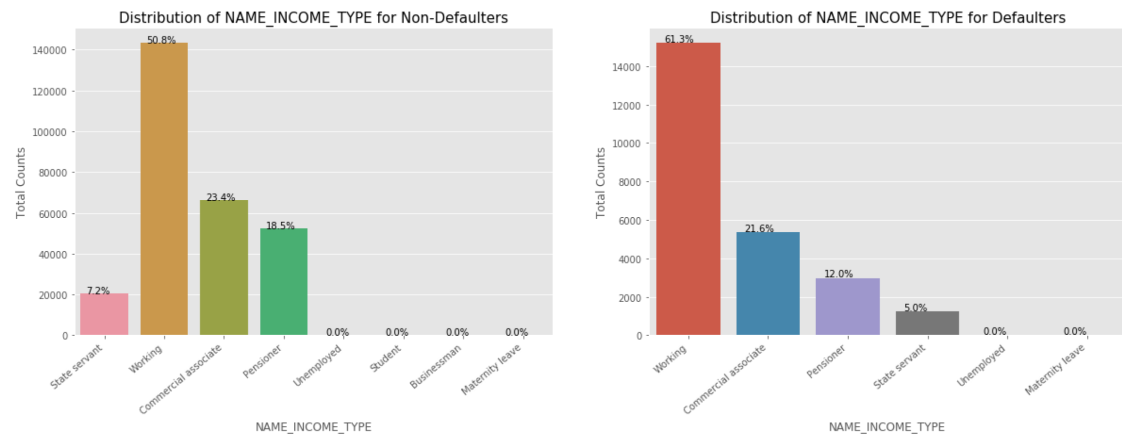
[287]:
```
plot_univ('FLAG_OWN_CAR')
```



We can see that people with cars contribute 65.7% to the non-defaulters while 69.5% to the defaulters. We can conclude that While people who have car default more often, the reason could be there are simply more people without cars Looking at the percentages in both the charts, we can conclude that the rate of default of people having car is low compared to people who don't.

0.10.4 (10.3) Plotting Name_Income_Type:

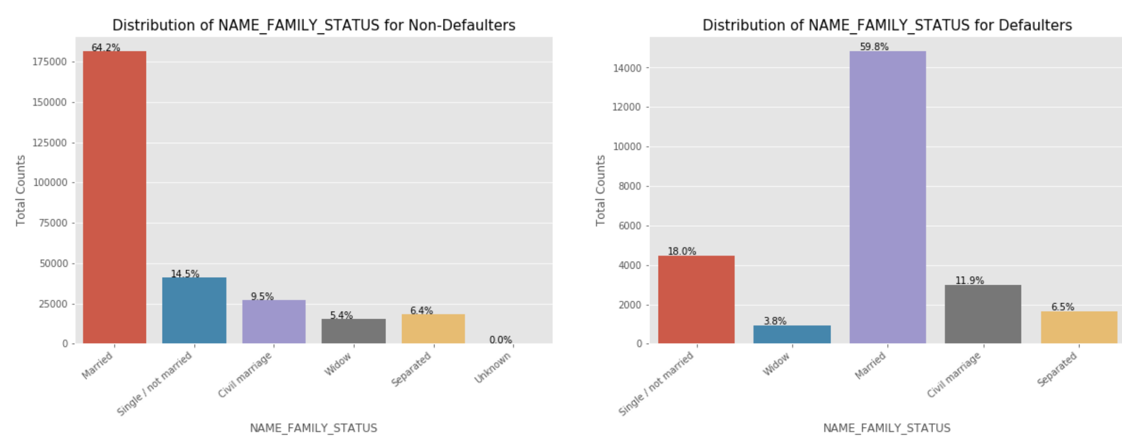
[288]:
```
plot_univ('NAME_INCOME_TYPE')
```



We can notice that the students don't default. The reason could be they are not required to pay during the time they are students. We can also see that the BusinessMen never default. Most of the loans are distributed to working class people We also see that working class people contribute 51% to non defaulters while they contribute to 61% of the defaulters. Clearly, the chances of defaulting are more in their case.

0.10.5 (10.4) Plotting Name_Income_Type:

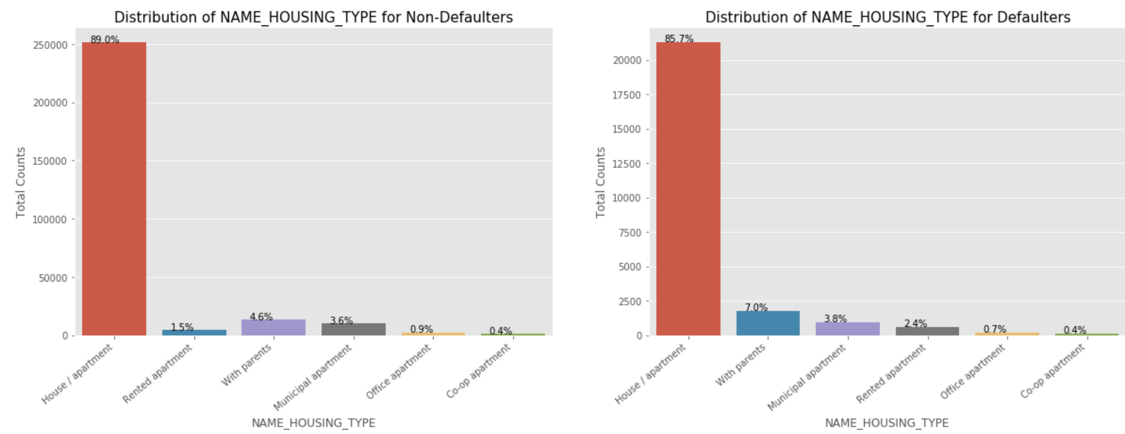
[289]:
```
plot_univ('NAME_FAMILY_STATUS')
```



Married people tend to apply for more loans comparatively. But from the graph we see that Single/non Married people contribute 14.5% to Non Defaulters and 18% to the defaulters. So there is more risk associated with them.

0.10.6 (10.5) Plotting Name_Housing_Type:

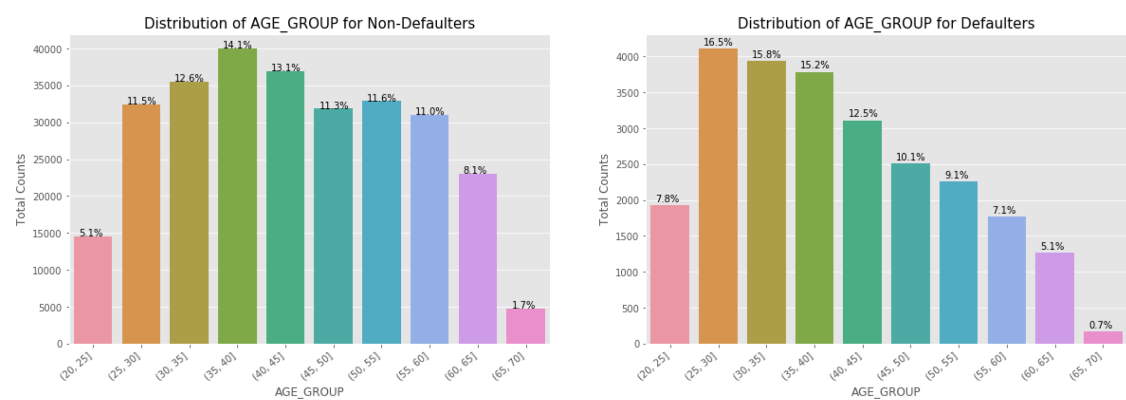
[290]:
```
plot_univ('NAME_HOUSING_TYPE')
```



It is clear from the graph that people who have House/Appartment, tend to apply for more loans. People living with parents tend to default more often when compared with others.The reason could be their living expenses are more due to their parents living with them.

0.10.7 (10.6) Plotting AGE_GROUP:

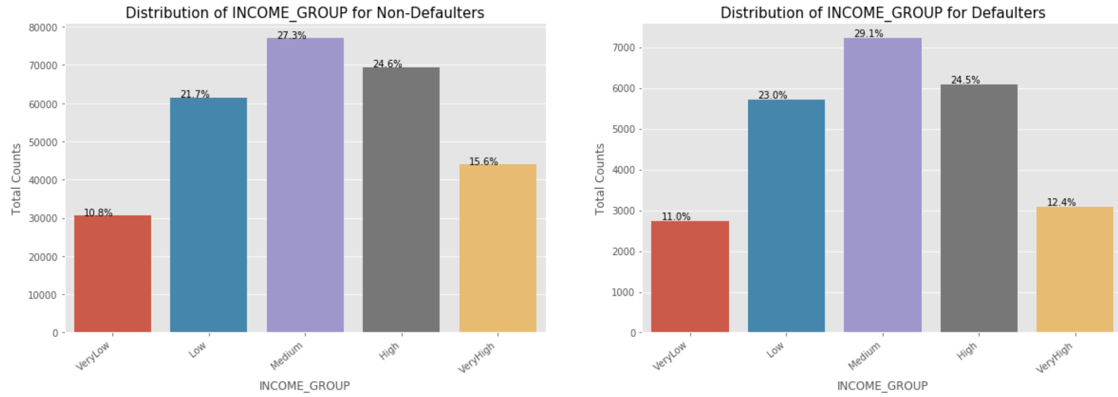
[291]:
```
plot_univ('AGE_GROUP')
```



We see that (25,30] age group tend to default more often. So they are the riskiest people to loan to. With increasing age group, people tend to default less starting from the age 25. One of the reasons could be they get employed around that age and with increasing age, their salary also increases.

0.10.8 (10.6) Plotting INCOME_GROUP:

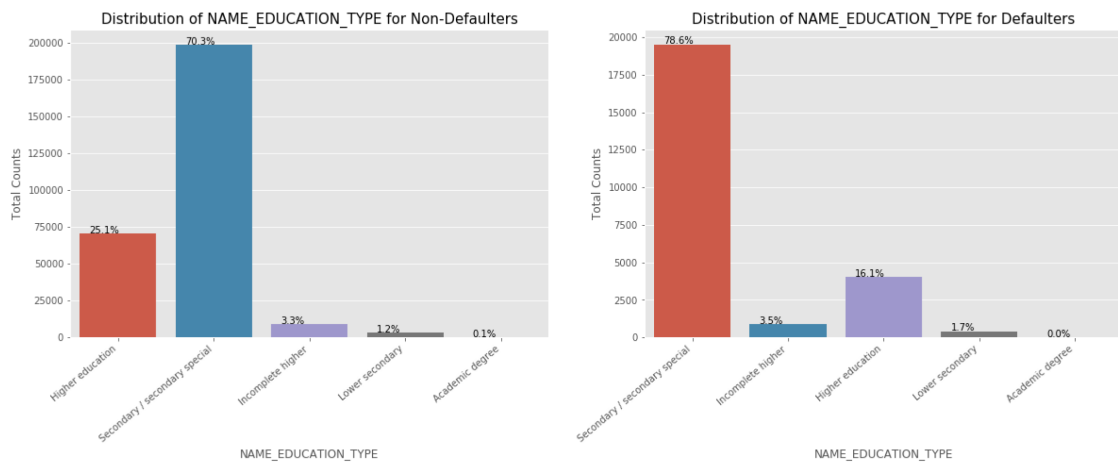
[292]:
```python
plot_univ('INCOME_GROUP')
```



The Very High income group tend to default less often. They contribute 12.4% to the total number of defaulters, while they contribute 15.6% to the Non-Defaulters.

0.10.9 (10.7) Plotting NAME_EDUCTAION_TYPE:

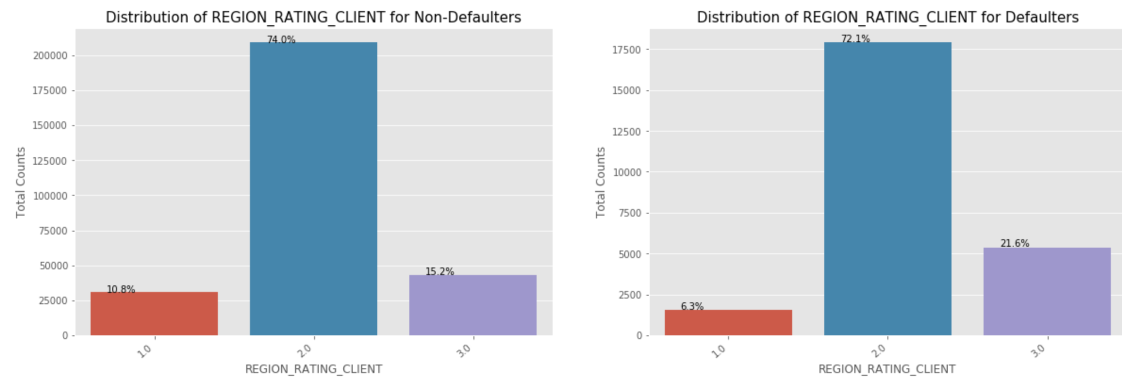
[293]:
```python
plot_univ('NAME_EDUCATION_TYPE')
```



Almost all of the Education categories are equally likely to default except for the higher educated ones who are less likely to default and secondary educated people are more likely to default

0.10.10 (10.8) Plotting REGION_RATING_CLIENT:

[294]:
```
plot_univ('REGION_RATING_CLIENT')
```



More people from second tier regions tend to apply for loans. We can infer that people living in better areas(Rating 3) tend contribute more to the defaulters by their weightage. People living in 1 rated areas

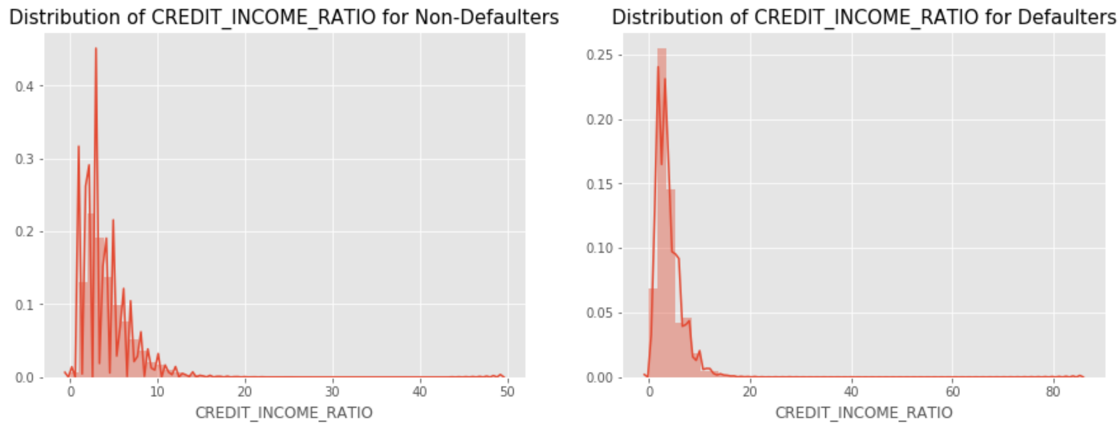
0.11 (11) univariate continuos variable analysis:

0.11.1 (11.1) FUNCTION :

[297]:
```
# function to dist plot for continuous variables
def plotunidist(D):

    plt.style.use('ggplot')
    sns.despine
    fig,(ax1,ax2) = plt.subplots(1,2,figsize=(15,5))

    sns.distplot(a=app_final_nondef[D],ax=ax1)

    ax1.set_title(f'Distribution of {D} for Non-Defaulters',fontsize=15)

    sns.distplot(a=app_final_def[D],ax=ax2)
    ax2.set_title(f'Distribution of {D} for Defaulters',fontsize=15)

    plt.show()
```

0.11.2 (11.2) Plotting the credit income ratio:

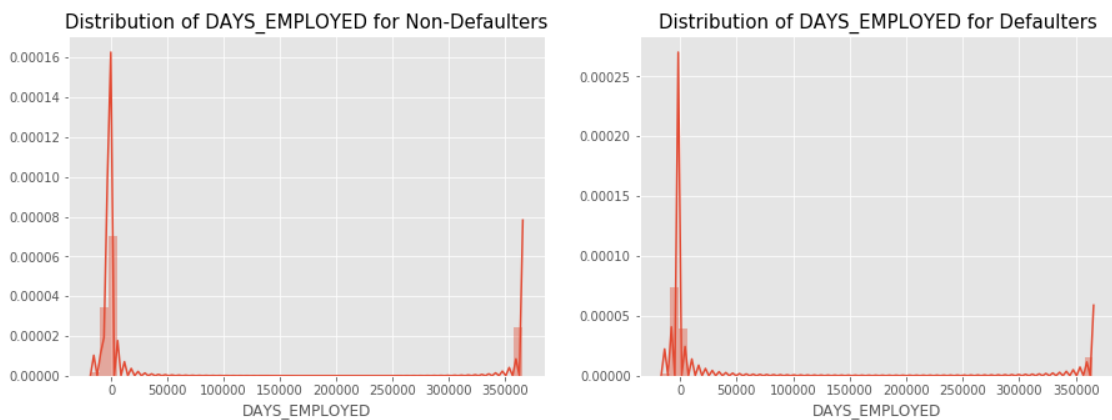
[298]:
```
plotunidist('CREDIT_INCOME_RATIO')
```



Credit income ratio the ratio of AMT_CREDIT/AMT_INCOME_TOTAL. Although there doesn't seem to be a clear distiguish between the group which defaulted vs the group which didn't when compared using the ratio, we can see that when the CREDIT_INCOME_RATIO is more than 50, people default.

0.11.3 (11.3) Plotting the DAYS_EMPLOYED:

[300]:
```
plotunidist('DAYS_EMPLOYED')
```



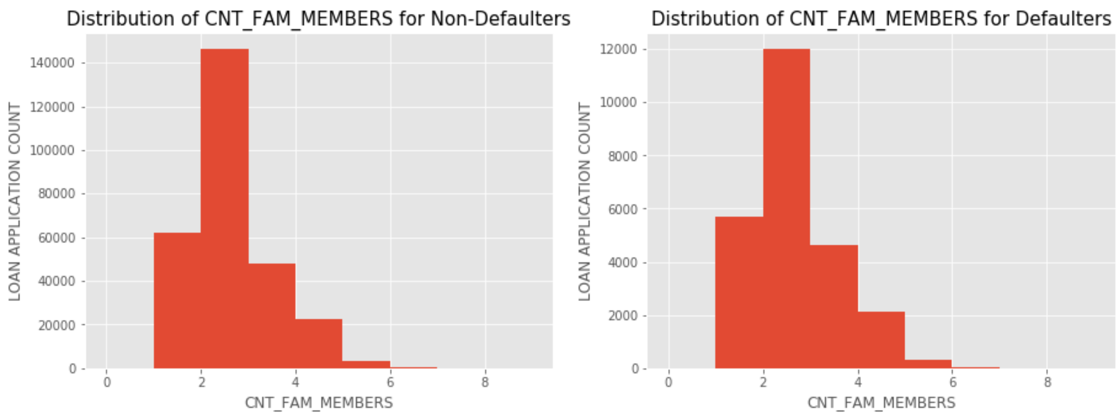
0.11.4 (11.3) Analysing the CNT_FAM_MEMBERS

[303]:
```
app_final_def['CNT_FAM_MEMBERS'].value_counts()
```

[303]:
```
2.0     12009
1.0      5675
3.0      4608
4.0      2136
```

```
5.0      327
6.0       55
7.0        6
8.0        6
11.0       1
10.0       1
13.0       1
Name: CNT_FAM_MEMBERS, dtype: int64
```

[311]:
```
plt.figure(figsize=(15,5))

plt.subplot(1, 2, 1)
app_final_nondef['CNT_FAM_MEMBERS'].plot.hist(bins=range(10))
plt.title('Distribution of CNT_FAM_MEMBERS for Non-Defaulters',fontsize=15)
plt.xlabel('CNT_FAM_MEMBERS')
plt.ylabel('LOAN APPLICATION COUNT')

plt.subplot(1, 2, 2)
app_final_def['CNT_FAM_MEMBERS'].plot.hist(bins=range(10))
plt.title(f'Distribution of CNT_FAM_MEMBERS for Defaulters',fontsize=15)
plt.xlabel('CNT_FAM_MEMBERS')
plt.ylabel('LOAN APPLICATION COUNT')

plt.show()
```



We can see that a family of 3 applies loan more often than the other families

0.12 (12) Getting the top 20 correlation of the selected columns

[313]:
```
#Getting the top 20 correlation in Non_defaulters
corr=app_final_nondef.corr()
corr_df = corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool)).
↪unstack().reset_index()
```

```
corr_df.columns=['Column1','Column2','Correlation']
corr_df.dropna(subset=['Correlation'],inplace=True)
corr_df['Abs_Correlation']=corr_df['Correlation'].abs()
corr_df = corr_df.sort_values(by=['Abs_Correlation'], ascending=False)
corr_df.head(20)
```

[313]:

```
                           Column1                        Column2  Correlation  \
Abs_Correlation
308                 AMT_GOODS_PRICE                     AMT_CREDIT     0.987253
0.987253
297            REGION_RATING_CLIENT    REGION_RATING_CLIENT_W_CITY     0.950148
0.950148
208   SOCIAL_CIRCLE_60_DAYS_DEF_PERC  SOCIAL_CIRCLE_30_DAYS_DEF_PERC     0.873003
0.873003
321                 AMT_GOODS_PRICE                    AMT_ANNUITY     0.776686
0.776686
272                     AMT_ANNUITY                     AMT_CREDIT     0.771308
0.771308
74              CREDIT_INCOME_RATIO                     AMT_CREDIT     0.648589
0.648589
310                 AMT_GOODS_PRICE            CREDIT_INCOME_RATIO     0.628749
0.628749
273                     AMT_ANNUITY               AMT_INCOME_TOTAL     0.418954
0.418954
274                     AMT_ANNUITY            CREDIT_INCOME_RATIO     0.391499
0.391499
309                 AMT_GOODS_PRICE               AMT_INCOME_TOTAL     0.349461
0.349461
56                 AMT_INCOME_TOTAL                     AMT_CREDIT     0.342801
0.342801
149                 CNT_FAM_MEMBERS                  DAYS_EMPLOYED    -0.237411
0.237411
75              CREDIT_INCOME_RATIO               AMT_INCOME_TOTAL    -0.225923
0.225923
113               DAYS_REGISTRATION                  DAYS_EMPLOYED    -0.210188
0.210188
165     REGION_RATING_CLIENT_W_CITY               AMT_INCOME_TOTAL    -0.200470
0.200470
291            REGION_RATING_CLIENT               AMT_INCOME_TOTAL    -0.186577
0.186577
150                 CNT_FAM_MEMBERS              DAYS_REGISTRATION     0.175622
0.175622
279                     AMT_ANNUITY    REGION_RATING_CLIENT_W_CITY    -0.145151
0.145151
93                    DAYS_EMPLOYED               AMT_INCOME_TOTAL    -0.141249
0.141249
303            REGION_RATING_CLIENT                    AMT_ANNUITY    -0.132126
```

0.132126

[314]:
```python
#Getting the top 20 correlation in defaulters
corr=app_final_def.corr()
corr_df = corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool)).
↪unstack().reset_index()
corr_df.columns=['Column1','Column2','Correlation']
corr_df.dropna(subset=['Correlation'],inplace=True)
corr_df['Abs_Correlation']=corr_df['Correlation'].abs()
corr_df = corr_df.sort_values(by=['Abs_Correlation'], ascending=False)
corr_df.head(20)
```

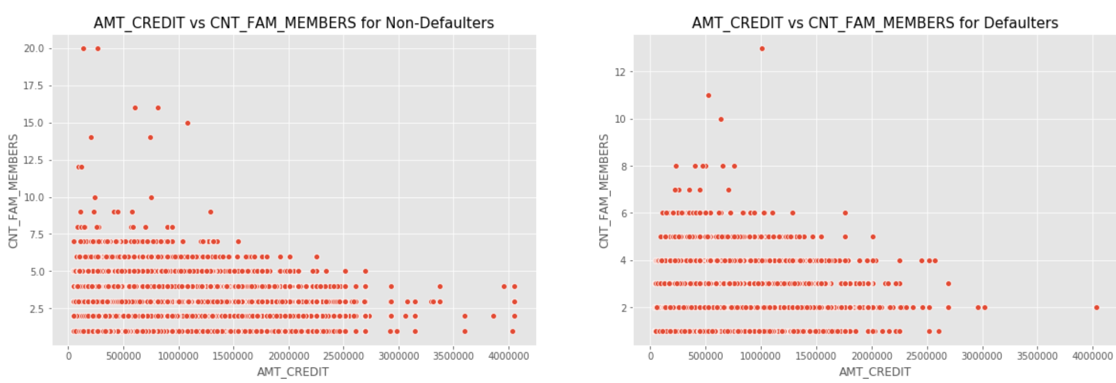
[314]:
```
                          Column1                        Column2  Correlation  \
Abs_Correlation
308                AMT_GOODS_PRICE                     AMT_CREDIT     0.983103
0.983103
297           REGION_RATING_CLIENT    REGION_RATING_CLIENT_W_CITY     0.956637
0.956637
208  SOCIAL_CIRCLE_60_DAYS_DEF_PERC  SOCIAL_CIRCLE_30_DAYS_DEF_PERC     0.874562
0.874562
321                AMT_GOODS_PRICE                    AMT_ANNUITY     0.752699
0.752699
272                    AMT_ANNUITY                     AMT_CREDIT     0.752195
0.752195
74             CREDIT_INCOME_RATIO                     AMT_CREDIT     0.639744
0.639744
310                AMT_GOODS_PRICE            CREDIT_INCOME_RATIO     0.623163
0.623163
274                    AMT_ANNUITY            CREDIT_INCOME_RATIO     0.381298
0.381298
113              DAYS_REGISTRATION                  DAYS_EMPLOYED    -0.188929
0.188929
149                CNT_FAM_MEMBERS                  DAYS_EMPLOYED    -0.186561
0.186561
150                CNT_FAM_MEMBERS              DAYS_REGISTRATION     0.145828
0.145828
94                   DAYS_EMPLOYED            CREDIT_INCOME_RATIO     0.119095
0.119095
294           REGION_RATING_CLIENT              DAYS_REGISTRATION     0.103855
0.103855
168    REGION_RATING_CLIENT_W_CITY              DAYS_REGISTRATION     0.100285
0.100285
279                    AMT_ANNUITY    REGION_RATING_CLIENT_W_CITY    -0.089291
0.089291
275                    AMT_ANNUITY                  DAYS_EMPLOYED    -0.082552
0.082552
277                    AMT_ANNUITY                     FLAG_EMAIL     0.078188
```

```
0.078188
315             AMT_GOODS_PRICE    REGION_RATING_CLIENT_W_CITY    -0.077191
0.077191
278                 AMT_ANNUITY                CNT_FAM_MEMBERS     0.075711
0.075711
303        REGION_RATING_CLIENT                    AMT_ANNUITY    -0.073784
0.073784
```

0.13 (13) Bivariate Analysis of numerical variables

[315]:
```python
# function for scatter plot for continuous variables
def plotbivar(var1,var2):

    plt.style.use('ggplot')
    sns.despine
    fig,(ax1,ax2) = plt.subplots(1,2,figsize=(20,6))

    sns.scatterplot(x=var1, y=var2,data=app_final_nondef,ax=ax1)
    ax1.set_xlabel(var1)
    ax1.set_ylabel(var2)
    ax1.set_title(f'{var1} vs {var2} for Non-Defaulters',fontsize=15)

    sns.scatterplot(x=var1, y=var2,data=app_final_def,ax=ax2)
    ax2.set_xlabel(var1)
    ax2.set_ylabel(var2)
    ax2.set_title(f'{var1} vs {var2} for Defaulters',fontsize=15)

    plt.show()
```

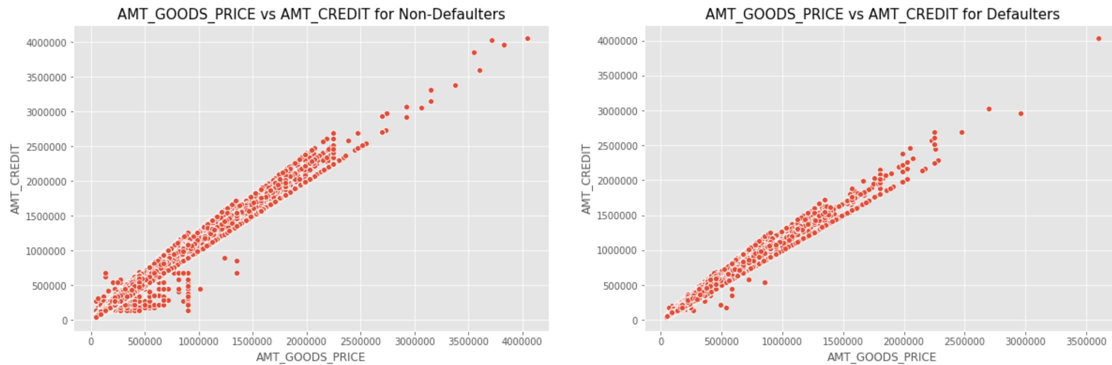
[316]:
```python
plotbivar('AMT_CREDIT','CNT_FAM_MEMBERS')
```



We can see that the density in the lower left corner is similar in both the case, so the people are equally likely to default if the family is small and the AMT_CREDIT is low. We can observe that larger families and people with larger AMT_CREDIT default less often

[317]:

```python
plotbivar('AMT_GOODS_PRICE','AMT_CREDIT')
```



1 (14) Data Analysis for previous application dataset:

[319]:

```python
prev_app.head()
```

[319]:

```
   SK_ID_PREV  SK_ID_CURR NAME_CONTRACT_TYPE  AMT_ANNUITY  AMT_APPLICATION
AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE WEEKDAY_APPR_PROCESS_START
HOUR_APPR_PROCESS_START FLAG_LAST_APPL_PER_CONTRACT  NFLAG_LAST_APPL_IN_DAY
RATE_DOWN_PAYMENT  RATE_INTEREST_PRIMARY  RATE_INTEREST_PRIVILEGED
NAME_CASH_LOAN_PURPOSE NAME_CONTRACT_STATUS  DAYS_DECISION
NAME_PAYMENT_TYPE CODE_REJECT_REASON  NAME_TYPE_SUITE NAME_CLIENT_TYPE
NAME_GOODS_CATEGORY NAME_PORTFOLIO NAME_PRODUCT_TYPE             CHANNEL_TYPE
SELLERPLACE_AREA NAME_SELLER_INDUSTRY  CNT_PAYMENT NAME_YIELD_GROUP
PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
0     2030495      271877     Consumer loans     1730.430          17145.0
17145.0                0.0           17145.0                     SATURDAY
15                           Y                       1                0.0
0.182832                 0.867336                     XAP            Approved
-73  Cash through the bank                XAP              NaN         Repeater
Mobile              POS                XNA               Country-wide
35          Connectivity          12.0          middle  POS mobile with interest
365243.0           -42.0                         300.0          -42.0
-37.0                        0.0
1     2802425      108129         Cash loans    25188.615         607500.0
679671.0               NaN           607500.0                    THURSDAY
11                           Y                       1                NaN
NaN                      NaN                     XNA             Approved
-164                   XNA                XAP     Unaccompanied         Repeater
XNA          Cash              x-sell         Contact center               -1
XNA         36.0        low_action          Cash X-Sell: low          365243.0
```

```
-134.0                      916.0       365243.0          365243.0
1.0
2     2523466      122040        Cash loans     15060.735         112500.0
136444.5              NaN         112500.0                        TUESDAY
11                           Y                        1                NaN
NaN                        NaN                     XNA             Approved
-301   Cash through the bank                XAP   Spouse, partner         Repeater
XNA           Cash            x-sell   Credit and cash offices              -1
XNA         12.0            high          Cash X-Sell: high           365243.0
-271.0                       59.0       365243.0          365243.0
1.0
3     2819243      176158        Cash loans     47041.335         450000.0
470790.0              NaN         450000.0                         MONDAY
7                           Y                        1                NaN
NaN                        NaN                     XNA             Approved
-512   Cash through the bank                XAP              NaN          Repeater
XNA           Cash            x-sell   Credit and cash offices              -1
XNA         12.0          middle       Cash X-Sell: middle          365243.0
-482.0                     -152.0         -182.0            -177.0
1.0
4     1784265      202054        Cash loans     31924.395         337500.0
404055.0              NaN         337500.0                       THURSDAY
9                           Y                        1                NaN
NaN                        NaN                 Repairs              Refused
-781   Cash through the bank                 HC              NaN          Repeater
XNA           Cash           walk-in   Credit and cash offices              -1
XNA         24.0            high          Cash Street: high                NaN
NaN                        NaN            NaN               NaN
NaN
```

[329]:
```python
# Removing all the columns with more than 50% of null values
prev_app = prev_app.loc[:,prev_app.isnull().mean()<=0.5]
prev_app.shape
```

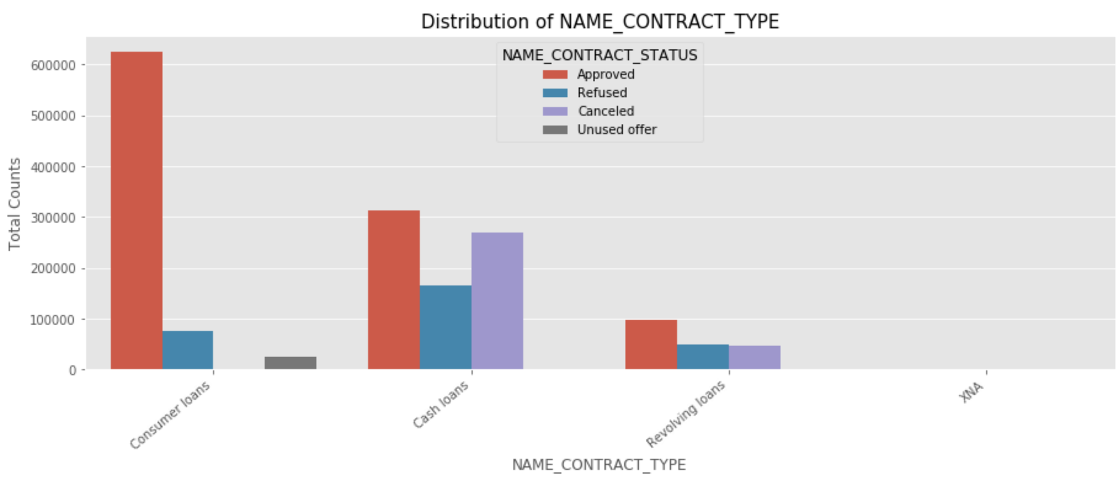
[329]: (1670214, 33)

1.1 (15) Univariate analysis

[334]:
```python
# function to count plot for categorical variables
def plot_uni(var):

    plt.style.use('ggplot')
    sns.despine
    fig,ax = plt.subplots(1,1,figsize=(15,5))

    sns.countplot(x=var, data=prev_app,ax=ax,hue='NAME_CONTRACT_STATUS')
    ax.set_ylabel('Total Counts')
```

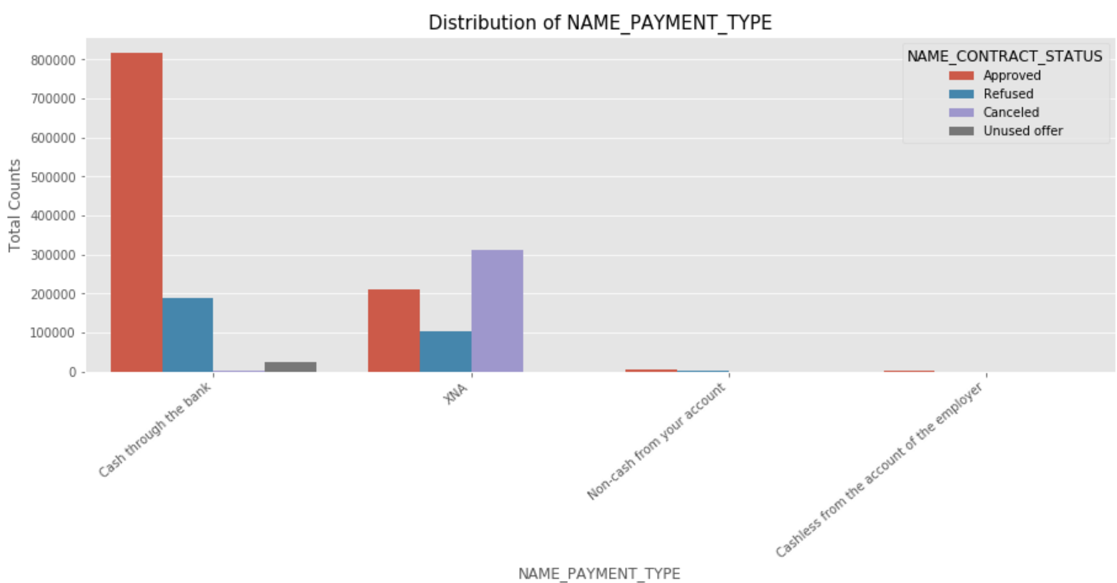
```
    ax.set_title(f'Distribution of {var}',fontsize=15)
    ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")

    plt.show()
```

[335]:
```
plot_uni('NAME_CONTRACT_TYPE')
```



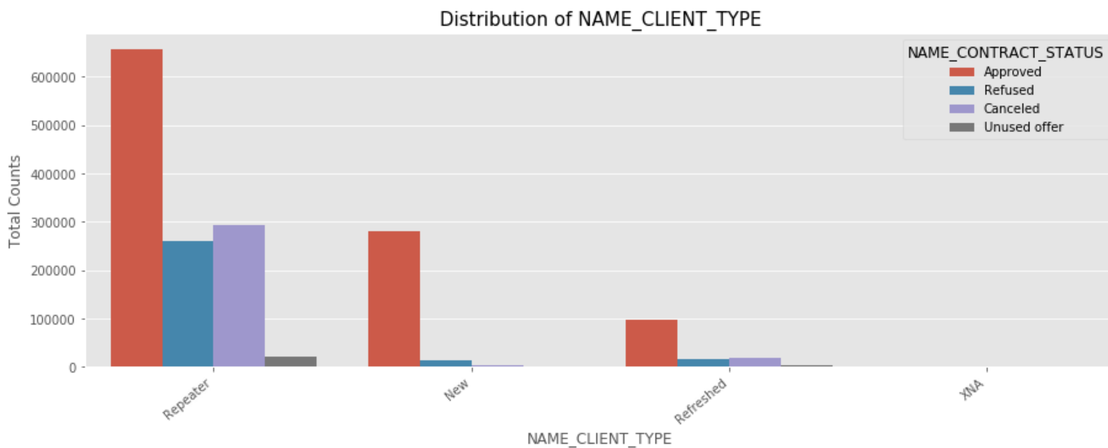
From the above chart, we can infer that, most of the applications are for 'Cash loan' and 'Consumer loan'. Although the cash loans are refused more often than others.

[336]:
```
plot_uni('NAME_PAYMENT_TYPE')
```



From the above chart, we can infer that most of the clients chose to repay the loan using the 'Cash through the bank' option We can also see that 'Non-Cash from your account' & 'Cashless from the account of the employee' options are not at all popular in terms of loan repayment amongst the customers

[337]:
```
plot_uni('NAME_CLIENT_TYPE')
```



Most of the loan applications are from repeat customers, out of the total applications 70% of customers are repeaters. They also get refused most often.

1.2 (16) Checking the correlation in the PreviousApplication dataset

[339]:
```
#Getting the top 20 correlation prev_app
corr=prev_app.corr()
corr_df = corr.where(np.triu(np.ones(corr.shape),k=1).astype(np.bool)).
↪unstack().reset_index()
corr_df.columns=['Column1','Column2','Correlation']
corr_df.dropna(subset=['Correlation'],inplace=True)
corr_df['Abs_Correlation']=corr_df['Correlation'].abs()
corr_df = corr_df.sort_values(by=['Abs_Correlation'], ascending=False)
corr_df.head(20)
```

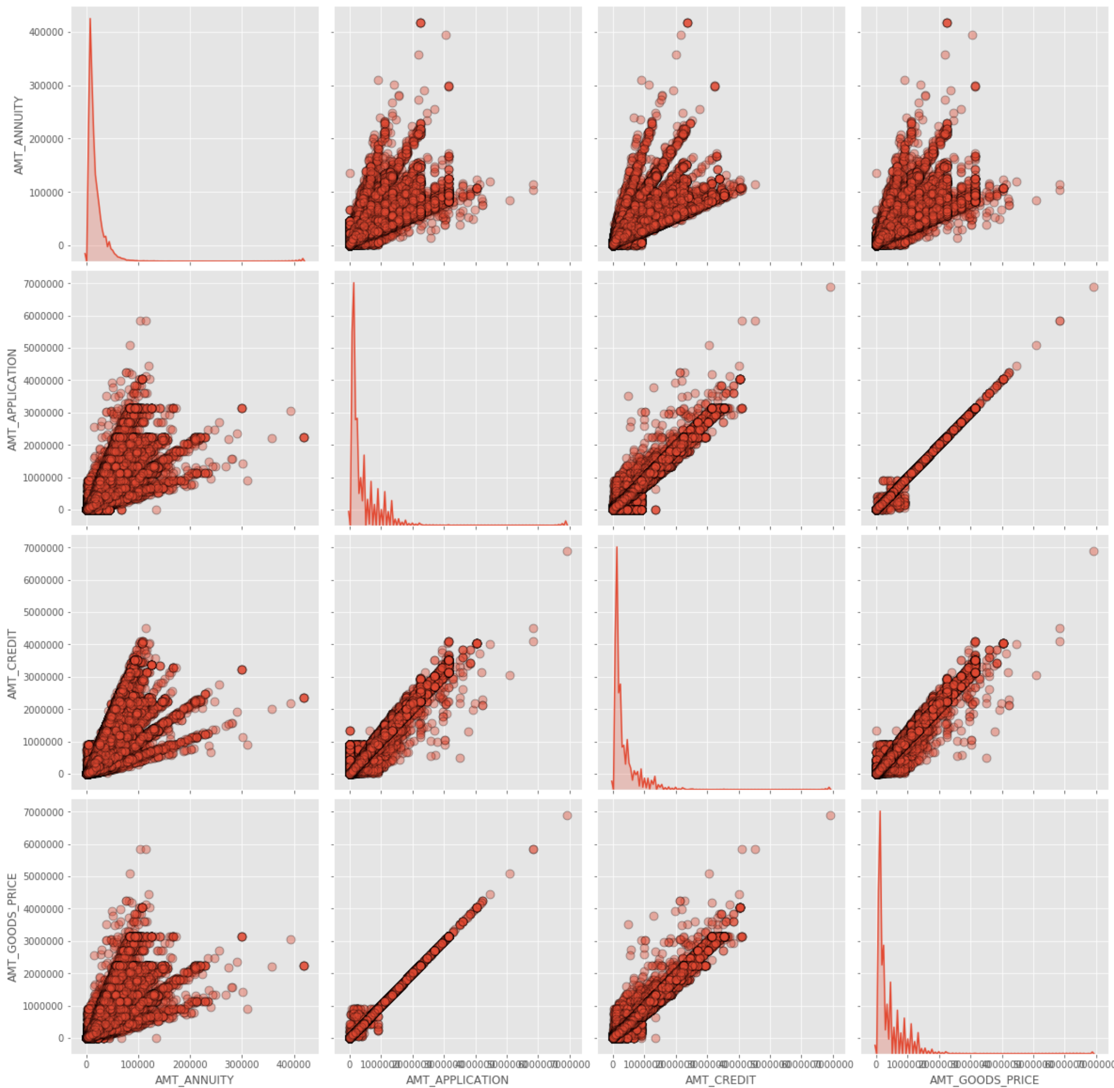
[339]:
```
                      Column1                  Column2  Correlation  \
Abs_Correlation
88             AMT_GOODS_PRICE          AMT_APPLICATION     0.999884
0.999884
89             AMT_GOODS_PRICE               AMT_CREDIT     0.993087
0.993087
71                  AMT_CREDIT          AMT_APPLICATION     0.975824
0.975824
269           DAYS_TERMINATION            DAYS_LAST_DUE     0.927990
0.927990
```

```
87              AMT_GOODS_PRICE                AMT_ANNUITY      0.820895
0.820895
70                   AMT_CREDIT                AMT_ANNUITY      0.816429
0.816429
53              AMT_APPLICATION                AMT_ANNUITY      0.808872
0.808872
232  DAYS_LAST_DUE_1ST_VERSION          DAYS_FIRST_DRAWING     -0.803494
0.803494
173                 CNT_PAYMENT            AMT_APPLICATION      0.680630
0.680630
174                 CNT_PAYMENT                 AMT_CREDIT      0.674278
0.674278
175                 CNT_PAYMENT            AMT_GOODS_PRICE      0.672129
0.672129
233  DAYS_LAST_DUE_1ST_VERSION              DAYS_FIRST_DUE      0.513949
0.513949
268            DAYS_TERMINATION  DAYS_LAST_DUE_1ST_VERSION      0.493174
0.493174
246               DAYS_LAST_DUE              DAYS_DECISION      0.448549
0.448549
251               DAYS_LAST_DUE  DAYS_LAST_DUE_1ST_VERSION      0.423462
0.423462
250               DAYS_LAST_DUE             DAYS_FIRST_DUE      0.401838
0.401838
263            DAYS_TERMINATION              DAYS_DECISION      0.400179
0.400179
266            DAYS_TERMINATION         DAYS_FIRST_DRAWING     -0.396284
0.396284
172                 CNT_PAYMENT                AMT_ANNUITY      0.394535
0.394535
231  DAYS_LAST_DUE_1ST_VERSION                CNT_PAYMENT     -0.381013
0.381013
```

1.3 (17) Using pairplot to perform bivariate analysis on numerical columns

[340]:
```
#plotting the relation between correlated highly corelated numeric vriables
plt.figure(figsize=[20,8])
sns.
 ↪pairplot(prev_app[['AMT_ANNUITY','AMT_APPLICATION','AMT_CREDIT','AMT_GOODS_PRICE','NAME_CON
             diag_kind = 'kde',
             plot_kws = {'alpha': 0.4, 's': 80, 'edgecolor': 'k'},
             size = 4)
plt.show()
```

```
<Figure size 1440x576 with 0 Axes>
```

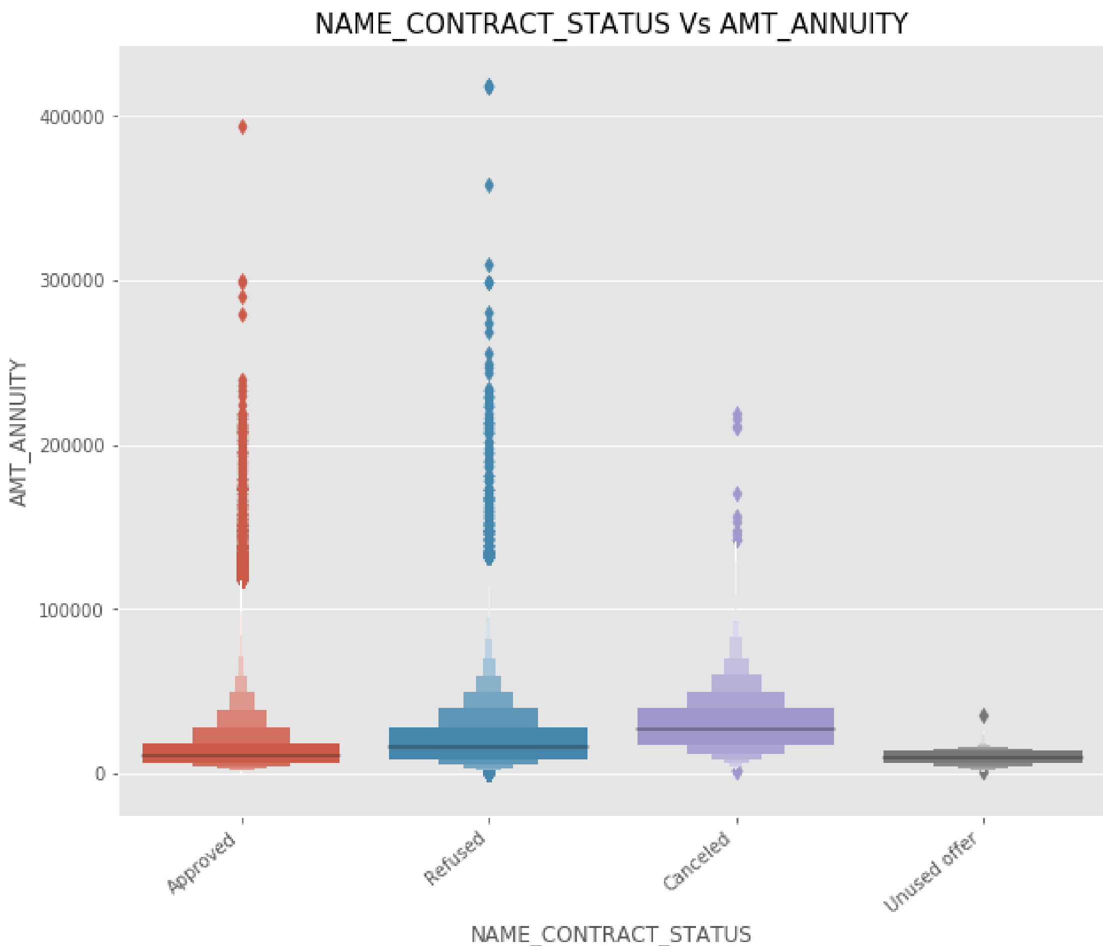


1. Annuity of previous application has a very high and positive influence over: (Increase of annuity increases below factors) (1) How much credit did client asked on the previous application (2)Final credit amount on the previous application that was approved by the bank (3) Goods price of good that client asked for on the previous application.
2. For how much credit did client ask on the previous application is highly influenced by the Goods price of good that client has asked for on the previous application
3. Final credit amount disbursed to the customer previously, after approval is highly influence by the application amount and also the goods price of good that client asked for on the previous application.

1.4 (18) Using box plot to do some more bivariate analysis on categorical vs numeric columns

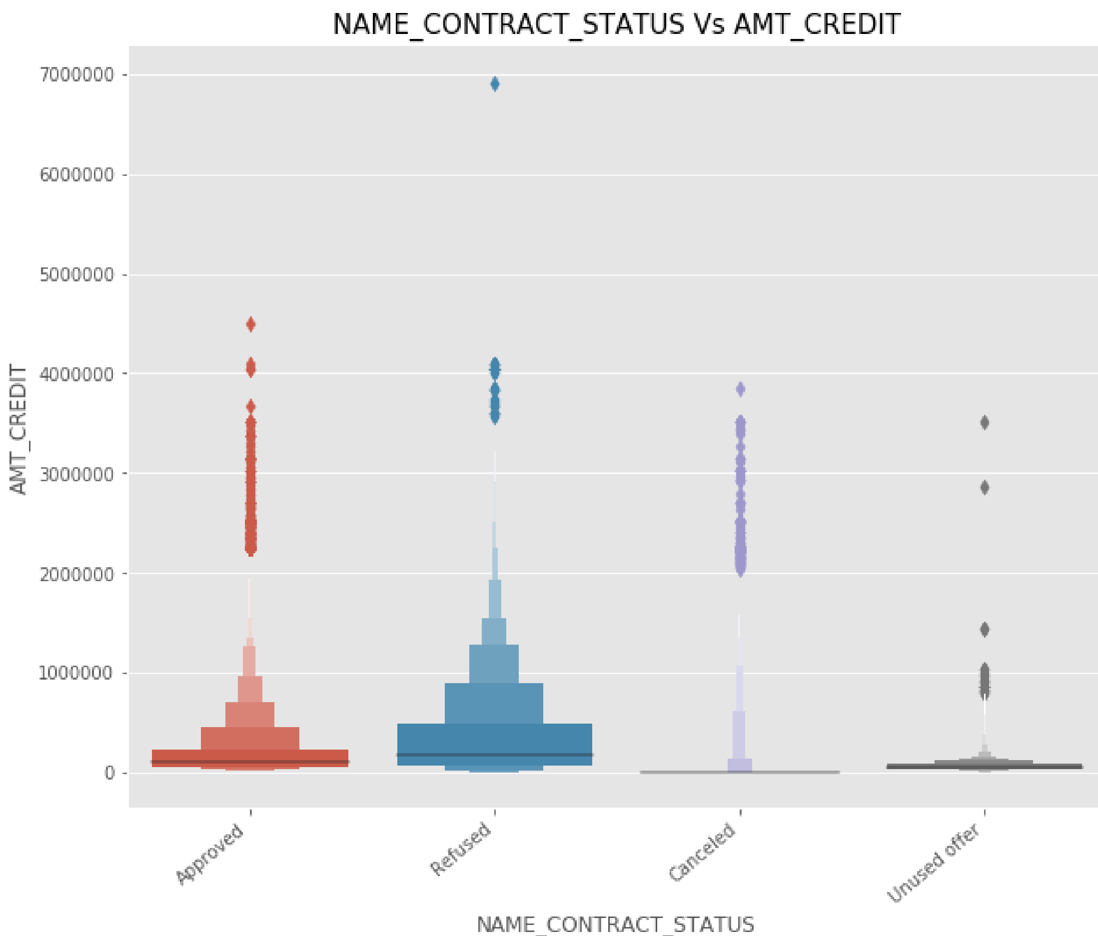
[341]:
```
#by variant analysis function
def plot_by_cat_num(cat, num):

    plt.style.use('ggplot')
    sns.despine
    fig,ax = plt.subplots(1,1,figsize=(10,8))

    sns.boxenplot(x=cat,y = num, data=prev_app)
    ax.set_ylabel(f'{num}')
    ax.set_xlabel(f'{cat}')

    ax.set_title(f'{cat} Vs {num}',fontsize=15)
    ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")

    plt.show()
```

[342]:
```
#by-varient analysis of Contract status and Annuity of previous appliction
plot_by_cat_num('NAME_CONTRACT_STATUS', 'AMT_ANNUITY')
```



From the above plot we can see that loan application for people with lower AMT_ANNUITY gets canceled or Unused most of the time. We also see that applications with too high AMT ANNUITY also got refused more often than others.

[343]:
```
#by-varient analysis of Contract status and Final credit amount disbursed to␣
↪the customer previously, after approval
plot_by_cat_num('NAME_CONTRACT_STATUS', 'AMT_CREDIT')
```



We can infer that when the AMT_CREDIT is too low, it get's cancelled/unused most of the time.

1.5 (19) Merging the files and analyzing the data

[344]:
```
## Merging the two files to do some analysis
NewLeftPrev = pd.merge(application_final,prev_app, how='left',␣
 ↪on=['SK_ID_CURR'])
```

[345]:
```
NewLeftPrev.head()
```

[345]:
```
   SK_ID_CURR  TARGET CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY INCOME_GROUP
AGE_GROUP  AMT_CREDIT_x  AMT_INCOME_TOTAL  CREDIT_INCOME_RATIO NAME_INCOME_TYPE
NAME_EDUCATION_TYPE    NAME_FAMILY_STATUS  NAME_HOUSING_TYPE  DAYS_EMPLOYED
DAYS_REGISTRATION  FLAG_EMAIL OCCUPATION_TYPE  CNT_FAM_MEMBERS
REGION_RATING_CLIENT_W_CITY        ORGANIZATION_TYPE
SOCIAL_CIRCLE_30_DAYS_DEF_PERC  SOCIAL_CIRCLE_60_DAYS_DEF_PERC
AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT
NAME_CONTRACT_TYPE_x  AMT_ANNUITY_x  REGION_RATING_CLIENT  AMT_GOODS_PRICE_x
```

```
SK_ID_PREV NAME_CONTRACT_TYPE_y  AMT_ANNUITY_y  AMT_APPLICATION  AMT_CREDIT_y
AMT_GOODS_PRICE_y WEEKDAY_APPR_PROCESS_START  HOUR_APPR_PROCESS_START
FLAG_LAST_APPL_PER_CONTRACT  NFLAG_LAST_APPL_IN_DAY NAME_CASH_LOAN_PURPOSE
NAME_CONTRACT_STATUS  DAYS_DECISION      NAME_PAYMENT_TYPE CODE_REJECT_REASON
NAME_TYPE_SUITE NAME_CLIENT_TYPE   NAME_GOODS_CATEGORY NAME_PORTFOLIO
NAME_PRODUCT_TYPE             CHANNEL_TYPE  \
0    100002.0     1.0        Male            N               Y         High
(20, 25]      406597.0         202500.0                  2.0          Working
Secondary / secondary special  Single / not married  House / apartment
-637.0          -3648.0         0.0        Laborers              1.0
2.0  Business Entity Type 3                           1.0
1.0                    0.0                        0.0
0.0           Cash loans        24700.5                  2.0           351000.0
1038818.0       Consumer loans       9251.775         179055.0      179055.0
179055.0                   SATURDAY                         9.0
Y                     1.0                   XAP             Approved
-606.0                    XNA               XAP             NaN
New              Vehicles            POS              XNA
Stone
1    100003.0     0.0      Female            N               N     VeryHigh
(40, 45]     1293502.0         270000.0                  5.0    State servant
Higher education               Married  House / apartment        -1188.0
-1186.0         0.0      Core staff             2.0
1.0                  School                         0.0
0.0                    0.0                        0.0
0.0           Cash loans        35698.5                  1.0          1129500.0
1810518.0           Cash loans      98356.995        900000.0     1035882.0
900000.0                     FRIDAY                        12.0
Y                     1.0                   XNA             Approved
-746.0                    XNA               XAP   Unaccompanied
Repeater                   XNA           Cash            x-sell  Credit and cash
offices
2    100003.0     0.0      Female            N               N     VeryHigh
(40, 45]     1293502.0         270000.0                  5.0    State servant
Higher education               Married  House / apartment        -1188.0
-1186.0         0.0      Core staff             2.0
1.0                  School                         0.0
0.0                    0.0                        0.0
0.0           Cash loans        35698.5                  1.0          1129500.0
2636178.0       Consumer loans      64567.665        337500.0      348637.5
337500.0                     SUNDAY                        17.0
Y                     1.0                   XAP             Approved
-828.0  Cash through the bank               XAP          Family
Refreshed             Furniture            POS              XNA
Stone
3    100003.0     0.0      Female            N               N     VeryHigh
(40, 45]     1293502.0         270000.0                  5.0    State servant
```

```
Higher education               Married  House / apartment        -1188.0
-1186.0         0.0      Core staff              2.0
1.0                 School                            0.0
0.0                    0.0                           0.0
0.0           Cash loans        35698.5                    1.0          1129500.0
2396755.0       Consumer loans       6737.310           68809.5       68053.5
68809.5                 SATURDAY                     15.0
Y                      1.0                   XAP             Approved
-2341.0  Cash through the bank               XAP           Family
Refreshed  Consumer Electronics           POS               XNA
Country-wide
4    100004.0     0.0         Male            Y               Y      VeryLow
(50, 55]      135000.0            67500.0                  2.0          Working
Secondary / secondary special  Single / not married  House / apartment
-225.0           -4260.0         0.0        Laborers             1.0
2.0             Government                            NaN
NaN                    0.0                           0.0
0.0      Revolving loans         6750.0                    2.0           135000.0
1564014.0       Consumer loans       5357.250           24282.0       20106.0
24282.0                   FRIDAY                      5.0
Y                      1.0                   XAP             Approved
-815.0  Cash through the bank               XAP   Unaccompanied
New                Mobile            POS                XNA         Regional /
Local

   SELLERPLACE_AREA  NAME_SELLER_INDUSTRY  CNT_PAYMENT NAME_YIELD_GROUP
PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
0             500.0       Auto technology         24.0       low_normal      POS
other with interest            365243.0          -565.0
125.0          -25.0              -17.0                       0.0
1              -1.0                   XNA         12.0       low_normal
Cash X-Sell: low            365243.0          -716.0                      -386.0
-536.0           -527.0                        1.0
2            1400.0            Furniture          6.0           middle   POS
industry with interest            365243.0          -797.0
-647.0          -647.0            -639.0                        0.0
3             200.0  Consumer electronics         12.0           middle  POS
household with interest            365243.0         -2310.0
-1980.0         -1980.0           -1976.0                         1.0
4              30.0          Connectivity          4.0           middle  POS
mobile without interest            365243.0          -784.0
-694.0          -724.0            -714.0                        0.0
```

1.5.1 (19.1) Basic checks on NewLeftPrev

[346]:
```
NewLeftPrev.shape
```

[346]:
```
(1430104, 62)
```

[347]:
```
NewLeftPrev.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1430104 entries, 0 to 1430103
Data columns (total 62 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   SK_ID_CURR                      1430100 non-null  float64
 1   TARGET                          1430100 non-null  float64
 2   CODE_GENDER                     1430100 non-null  object
 3   FLAG_OWN_CAR                    1430100 non-null  object
 4   FLAG_OWN_REALTY                 1430100 non-null  object
 5   INCOME_GROUP                    1430100 non-null  category
 6   AGE_GROUP                       1430096 non-null  category
 7   AMT_CREDIT_x                    1430100 non-null  float64
 8   AMT_INCOME_TOTAL                1430100 non-null  float64
 9   CREDIT_INCOME_RATIO             1430100 non-null  float64
 10  NAME_INCOME_TYPE                1430100 non-null  object
 11  NAME_EDUCATION_TYPE             1430100 non-null  object
 12  NAME_FAMILY_STATUS              1430100 non-null  object
 13  NAME_HOUSING_TYPE               1430100 non-null  object
 14  DAYS_EMPLOYED                   1430100 non-null  float64
 15  DAYS_REGISTRATION               1430100 non-null  float64
 16  FLAG_EMAIL                      1430100 non-null  float64
 17  OCCUPATION_TYPE                 1430100 non-null  object
 18  CNT_FAM_MEMBERS                 1430098 non-null  float64
 19  REGION_RATING_CLIENT_W_CITY     1430100 non-null  float64
 20  ORGANIZATION_TYPE               1430100 non-null  object
 21  SOCIAL_CIRCLE_30_DAYS_DEF_PERC  684767 non-null   float64
 22  SOCIAL_CIRCLE_60_DAYS_DEF_PERC  681441 non-null   float64
 23  AMT_REQ_CREDIT_BUREAU_DAY       1264288 non-null  float64
 24  AMT_REQ_CREDIT_BUREAU_MON       1264288 non-null  float64
 25  AMT_REQ_CREDIT_BUREAU_QRT       1264288 non-null  float64
 26  NAME_CONTRACT_TYPE_x            1430100 non-null  object
 27  AMT_ANNUITY_x                   1430007 non-null  float64
 28  REGION_RATING_CLIENT            1430100 non-null  float64
 29  AMT_GOODS_PRICE_x               1428881 non-null  float64
 30  SK_ID_PREV                      1413646 non-null  float64
 31  NAME_CONTRACT_TYPE_y            1413646 non-null  object
 32  AMT_ANNUITY_y                   1106438 non-null  float64
 33  AMT_APPLICATION                 1413646 non-null  float64
 34  AMT_CREDIT_y                    1413645 non-null  float64
```

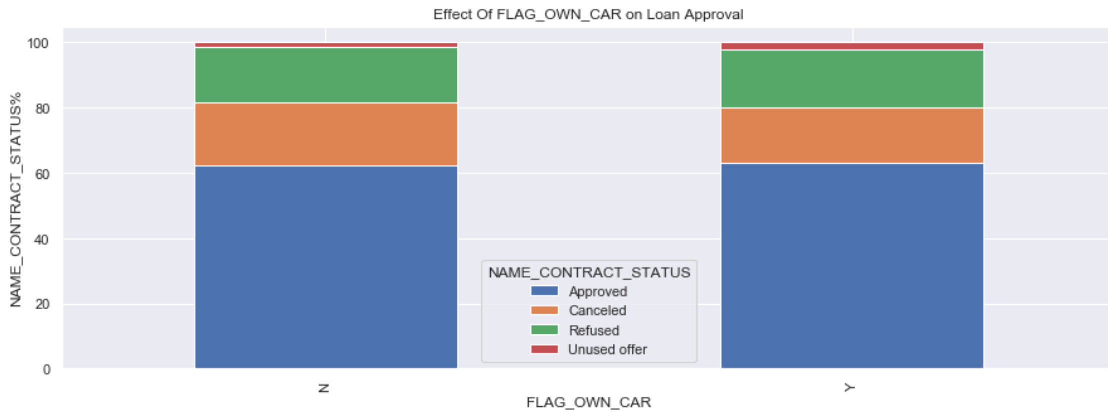
```
 35  AMT_GOODS_PRICE_y             1094130 non-null  float64
 36  WEEKDAY_APPR_PROCESS_START    1413646 non-null  object
 37  HOUR_APPR_PROCESS_START       1413646 non-null  float64
 38  FLAG_LAST_APPL_PER_CONTRACT   1413646 non-null  object
 39  NFLAG_LAST_APPL_IN_DAY        1413646 non-null  float64
 40  NAME_CASH_LOAN_PURPOSE        1413646 non-null  object
 41  NAME_CONTRACT_STATUS          1413646 non-null  object
 42  DAYS_DECISION                 1413646 non-null  float64
 43  NAME_PAYMENT_TYPE             1413646 non-null  object
 44  CODE_REJECT_REASON            1413646 non-null  object
 45  NAME_TYPE_SUITE               718992 non-null   object
 46  NAME_CLIENT_TYPE              1413646 non-null  object
 47  NAME_GOODS_CATEGORY           1413646 non-null  object
 48  NAME_PORTFOLIO                1413646 non-null  object
 49  NAME_PRODUCT_TYPE             1413646 non-null  object
 50  CHANNEL_TYPE                  1413646 non-null  object
 51  SELLERPLACE_AREA              1413646 non-null  float64
 52  NAME_SELLER_INDUSTRY          1413646 non-null  object
 53  CNT_PAYMENT                   1106443 non-null  float64
 54  NAME_YIELD_GROUP              1413646 non-null  object
 55  PRODUCT_COMBINATION           1413333 non-null  object
 56  DAYS_FIRST_DRAWING            852573 non-null   float64
 57  DAYS_FIRST_DUE                852573 non-null   float64
 58  DAYS_LAST_DUE_1ST_VERSION     852573 non-null   float64
 59  DAYS_LAST_DUE                 852573 non-null   float64
 60  DAYS_TERMINATION              852573 non-null   float64
 61  NFLAG_INSURED_ON_APPROVAL     852573 non-null   float64
dtypes: category(2), float64(34), object(26)
memory usage: 668.3+ MB
```

[348]:

```python
def plotuni_combined(Varx,Vary):
    # 100% bar chart
    plt.style.use('ggplot')
    sns.despine
    NewDat = NewLeftPrev.pivot_table(values='SK_ID_CURR',
                      index=Varx,
                      columns=Vary,
                      aggfunc='count')
    NewDat=NewDat.div(NewDat.sum(axis=1),axis='rows')*100
    sns.set()
    NewDat.plot(kind='bar',stacked=True,figsize=(15,5))
    plt.title(f'Effect Of {Varx} on Loan Approval')
    plt.xlabel(f'{Varx}')
    plt.ylabel(f'{Vary}%')
    plt.show()
```

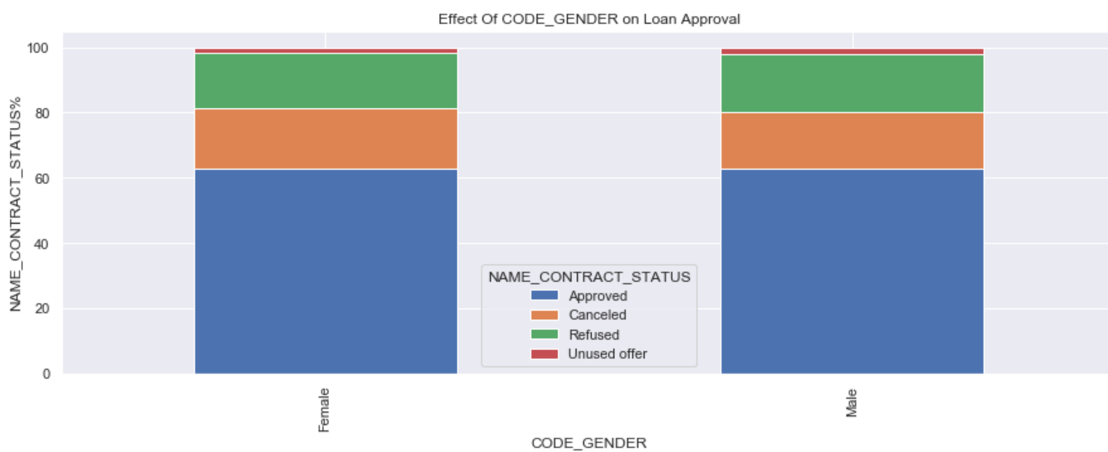
[349]:

```python
plotuni_combined('FLAG_OWN_CAR','NAME_CONTRACT_STATUS')
```



We see that car ownership doesn't have any effect on application approval or rejection. But we saw earlier that the people who has a car has lesser chances of default. The bank can add more weightage to car ownership while approving a loan amount

[350]:
```
plotuni_combined('CODE_GENDER','NAME_CONTRACT_STATUS')
```



We see that code gender doesn't have any effect on application approval or rejection. But we saw earlier that female have lesser chances of default compared to males. The bank can add more weightage to female while approving a loan amount.

[351]:
```
plotuni_combined('TARGET','NAME_CONTRACT_STATUS')
```

1.6 Target variable (0 - Non Defaulter 1 - Defaulter)

We can see that the people who were approved for a loan earlier, defaulted less often where as people who were refused a loan earlier have higher chances of defaulting.

2 The END