01_Exploration.ipynb

September 10, 2021

1 Initialisation

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

2 Exploration

[ ]:
```
! ls -lh 02_data/
```

```
total 2,5G
-rw-rw-r-- 1 adrien adrien  26M juin  26  2018 application_test.csv
-rw-rw-r-- 1 adrien adrien 159M juin  26  2018 application_train.csv
-rw-rw-r-- 1 adrien adrien 359M juin  26  2018 bureau_balance.csv
-rw-rw-r-- 1 adrien adrien 163M juin  26  2018 bureau.csv
-rw-rw-r-- 1 adrien adrien 405M juin  26  2018 credit_card_balance.csv
-rw-rw-r-- 1 adrien adrien  37K juin  26  2018
HomeCredit_columns_description.csv
-rw-rw-r-- 1 adrien adrien 690M juin  26  2018 installments_payments.csv
-rw-rw-r-- 1 adrien adrien 375M juin  26  2018 POS_CASH_balance.csv
-rw-rw-r-- 1 adrien adrien 387M juin  26  2018 previous_application.csv
-rw-rw-r-- 1 adrien adrien 524K juin  26  2018 sample_submission.csv
```

[ ]:
```
col_desc = pd.read_csv('02_data/HomeCredit_columns_description.csv',
                       index_col=0)
col_desc
```

[ ]:
```
                          Table                   Row  \
1    application_{train|test}.csv            SK_ID_CURR
2    application_{train|test}.csv                TARGET
5    application_{train|test}.csv    NAME_CONTRACT_TYPE
6    application_{train|test}.csv           CODE_GENDER
7    application_{train|test}.csv          FLAG_OWN_CAR
..                              …                     …
217     installments_payments.csv  NUM_INSTALMENT_NUMBER
218     installments_payments.csv       DAYS_INSTALMENT
```

```
219     installments_payments.csv     DAYS_ENTRY_PAYMENT
220     installments_payments.csv         AMT_INSTALMENT
221     installments_payments.csv            AMT_PAYMENT

                                         Description  \
1                             ID of loan in our sample
2    Target variable (1 - client with payment diffi…
5          Identification if loan is cash or revolving
6                                   Gender of the client
7                          Flag if the client owns a car
..                                                   …
217             On which installment we observe payment
218  When the installment of previous credit was su…
219  When was the installments of previous credit p…
220  What was the prescribed installment amount of …
221  What the client actually paid on previous cred…

                               Special
1                                  NaN
2                                  NaN
5                                  NaN
6                                  NaN
7                                  NaN
..                                   …
217                                NaN
218  time only relative to the application
219  time only relative to the application
220                                NaN
221                                NaN

[219 rows x 4 columns]
```

2.1 Tables application_{train|test}.csv

Il y a plus de 200 colonnes pour 9 tables au format csv ! Avant d'aller plus loin dans l'exploration je vais me concentrer sur les tables principales : les tables `application_{train|test}.csv`.

Je vais d'abord regarder les plus grosses corrélations avec la variable `TARGET`

[ ]:
```
app_train = pd.read_csv('02_data/application_train.csv')
app_test = pd.read_csv('02_data/application_test.csv')
app_train
```

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

```
4           100007       0         Cash loans          M            N
…            …     …                …         …         …
307506      456251       0         Cash loans          M            N
307507      456252       0         Cash loans          F            N
307508      456253       0         Cash loans          F            N
307509      456254       1         Cash loans          F            N
307510      456255       0         Cash loans          F            N

       FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0                    Y             0          202500.0    406597.5
1                    N             0          270000.0   1293502.5
2                    Y             0           67500.0    135000.0
3                    Y             0          135000.0    312682.5
4                    Y             0          121500.0    513000.0
…                 …          …               …         …
307506               N             0          157500.0    254700.0
307507               Y             0           72000.0    269550.0
307508               Y             0          153000.0    677664.0
307509               Y             0          171000.0    370107.0
307510               N             0          157500.0    675000.0

        AMT_ANNUITY  …  FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20  \
0           24700.5  …                 0                0                0
1           35698.5  …                 0                0                0
2            6750.0  …                 0                0                0
3           29686.5  …                 0                0                0
4           21865.5  …                 0                0                0
…             …  …            …              …              …
307506      27558.0  …                 0                0                0
307507      12001.5  …                 0                0                0
307508      29979.0  …                 0                0                0
307509      20205.0  …                 0                0                0
307510      49117.5  …                 0                0                0

       FLAG_DOCUMENT_21 AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0                     0                        0.0                       0.0
1                     0                        0.0                       0.0
2                     0                        0.0                       0.0
3                     0                        NaN                       NaN
4                     0                        0.0                       0.0
…                  …                        …                       …
307506                0                        NaN                       NaN
307507                0                        NaN                       NaN
307508                0                        1.0                       0.0
307509                0                        0.0                       0.0
307510                0                        0.0                       0.0
```

```
        AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                              0.0                        0.0
1                              0.0                        0.0
2                              0.0                        0.0
3                              NaN                        NaN
4                              0.0                        0.0
…                                …                          …
307506                         NaN                        NaN
307507                         NaN                        NaN
307508                         0.0                        1.0
307509                         0.0                        0.0
307510                         0.0                        2.0

        AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                             0.0                         1.0
1                             0.0                         0.0
2                             0.0                         0.0
3                             NaN                         NaN
4                             0.0                         0.0
…                               …                           …
307506                        NaN                         NaN
307507                        NaN                         NaN
307508                        0.0                         1.0
307509                        0.0                         0.0
307510                        0.0                         1.0

[307511 rows x 122 columns]
```

[ ]:
```
assert len(app_train.SK_ID_CURR.unique()) == app_train.shape[0]
assert len(app_test.SK_ID_CURR.unique()) == app_test.shape[0]

app_train.set_index('SK_ID_CURR', inplace=True)
app_test.set_index('SK_ID_CURR', inplace=True)
```

[ ]:
```
app_train
```

[ ]:
```
            TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
SK_ID_CURR
100002           1         Cash loans           M            N
100003           0         Cash loans           F            N
100004           0    Revolving loans           M            Y
100006           0         Cash loans           F            N
100007           0         Cash loans           M            N
…              …                  …         …           …
456251           0         Cash loans           M            N
456252           0         Cash loans           F            N
456253           0         Cash loans           F            N
```

```
456254          1         Cash loans           F            N
456255          0         Cash loans           F            N

           FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
SK_ID_CURR
100002                   Y             0          202500.0    406597.5
100003                   N             0          270000.0   1293502.5
100004                   Y             0           67500.0    135000.0
100006                   Y             0          135000.0    312682.5
100007                   Y             0          121500.0    513000.0
…                     …             …                …          …
456251                   N             0          157500.0    254700.0
456252                   Y             0           72000.0    269550.0
456253                   Y             0          153000.0    677664.0
456254                   Y             0          171000.0    370107.0
456255                   N             0          157500.0    675000.0

            AMT_ANNUITY  AMT_GOODS_PRICE  … FLAG_DOCUMENT_18  \
SK_ID_CURR                                …
100002          24700.5         351000.0  …                0
100003          35698.5        1129500.0  …                0
100004           6750.0         135000.0  …                0
100006          29686.5         297000.0  …                0
100007          21865.5         513000.0  …                0
…                 …               …   …              …
456251          27558.0         225000.0  …                0
456252          12001.5         225000.0  …                0
456253          29979.0         585000.0  …                0
456254          20205.0         319500.0  …                0
456255          49117.5         675000.0  …                0

           FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21  \
SK_ID_CURR
100002                    0                0                0
100003                    0                0                0
100004                    0                0                0
100006                    0                0                0
100007                    0                0                0
…                      …              …              …
456251                    0                0                0
456252                    0                0                0
456253                    0                0                0
456254                    0                0                0
456255                    0                0                0

           AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
SK_ID_CURR
```

```
100002                              0.0                         0.0
100003                              0.0                         0.0
100004                              0.0                         0.0
100006                              NaN                         NaN
100007                              0.0                         0.0
…                                 …                         …
456251                              NaN                         NaN
456252                              NaN                         NaN
456253                              1.0                         0.0
456254                              0.0                         0.0
456255                              0.0                         0.0

            AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
SK_ID_CURR
100002                               0.0                        0.0
100003                               0.0                        0.0
100004                               0.0                        0.0
100006                               NaN                        NaN
100007                               0.0                        0.0
…                                  …                        …
456251                               NaN                        NaN
456252                               NaN                        NaN
456253                               0.0                        1.0
456254                               0.0                        0.0
456255                               0.0                        2.0

            AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
SK_ID_CURR
100002                              0.0                         1.0
100003                              0.0                         0.0
100004                              0.0                         0.0
100006                              NaN                         NaN
100007                              0.0                         0.0
…                                 …                          …
456251                              NaN                         NaN
456252                              NaN                         NaN
456253                              0.0                         1.0
456254                              0.0                         0.0
456255                              0.0                         1.0

[307511 rows x 121 columns]
```

2.1.1 Encodage des colonnes textuelles

[ ]:
```
app_train.dtypes.value_counts()
```

[ ]:
```
float64    65
int64      40
object     16
dtype: int64
```

[ ]:
```
app_train.select_dtypes('object').apply(pd.Series.nunique, axis=0)
```

[ ]:
```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                7
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
OCCUPATION_TYPE               18
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
FONDKAPREMONT_MODE             4
HOUSETYPE_MODE                 3
WALLSMATERIAL_MODE             7
EMERGENCYSTATE_MODE            2
dtype: int64
```

[ ]:
```
categorical_labels = app_train.select_dtypes('object').columns.tolist()
```

[ ]:
```
for col in categorical_labels:
    null_count = app_train[col].isna().sum()
    null_perct = null_count / app_train[col].isna().count()
    if null_count != 0:
        print(col, null_count, round(null_perct, 4))
```

```
NAME_TYPE_SUITE 1292 0.0042
OCCUPATION_TYPE 96391 0.3135
FONDKAPREMONT_MODE 210295 0.6839
HOUSETYPE_MODE 154297 0.5018
WALLSMATERIAL_MODE 156341 0.5084
EMERGENCYSTATE_MODE 145755 0.474
```

[ ]:
```
app_train.OCCUPATION_TYPE.unique()
```

[ ]:
```
array(['Laborers', 'Core staff', 'Accountants', 'Managers', nan,
       'Drivers', 'Sales staff', 'Cleaning staff', 'Cooking staff',
       'Private service staff', 'Medicine staff', 'Security staff',
       'High skill tech staff', 'Waiters/barmen staff',
       'Low-skill Laborers', 'Realty agents', 'Secretaries', 'IT staff',
```

```
       'HR staff'], dtype=object)
```

[ ]:
```
col_desc.loc[col_desc.Table.eq('application_{train|test}.csv')
             & col_desc.Row.eq('OCCUPATION_TYPE')].Description.tolist()
```

[ ]:
```
['What kind of occupation does the client have']
```

[ ]:
```
app_train.OCCUPATION_TYPE.fillna('Unknown', inplace=True)
```

[ ]:
```
app_train.dropna(subset=['NAME_TYPE_SUITE'], inplace=True)
```

[ ]:
```
app_train.drop(columns=['FONDKAPREMONT_MODE',
                        'HOUSETYPE_MODE',
                        'WALLSMATERIAL_MODE',
                        'EMERGENCYSTATE_MODE'], inplace=True)
```

[ ]:
```
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer
```

[ ]:
```
app_train.WEEKDAY_APPR_PROCESS_START.unique()

weekday_codes = {
    'MONDAY': 0,
    'TUESDAY': 1,
    'WEDNESDAY': 2,
    'THURSDAY': 3,
    'FRIDAY': 4,
    'SATURDAY': 5,
    'SUNDAY': 6
}

app_train.WEEKDAY_APPR_PROCESS_START =\
    app_train.WEEKDAY_APPR_PROCESS_START.map(weekday_codes)
```

[ ]:
```
len(categorical_labels)
```

[ ]:
```
11
```

[ ]:
```
categorical_labels = app_train.select_dtypes('object').columns.tolist()
preprocessor = make_column_transformer(
    (OrdinalEncoder(), categorical_labels), remainder='passthrough'
)

app_train_encoded = preprocessor.fit_transform(app_train)
```

[ ]:
```
app_train_encoded
```

```
[ ]: array([[0., 1., 0., …, 0., 0., 1.],
        [0., 0., 0., …, 0., 0., 0.],
        [1., 1., 1., …, 0., 0., 0.],
        …,
        [0., 0., 0., …, 1., 0., 1.],
        [0., 0., 0., …, 0., 0., 0.],
        [0., 0., 0., …, 2., 0., 1.]])
```

2.2 Autres tables

```
[ ]: col_desc.loc[col_desc.Table == 'bureau_balance.csv'].Description.values
```

```
[ ]: array(['Recoded ID of Credit Bureau credit (unique coding for each application)
- use this to join to CREDIT_BUREAU table ',
       'Month of balance relative to application date (-1 means the freshest
balance date)',
       'Status of Credit Bureau loan during the month (active, closed, DPD0-30,
[C means closed, X means status unknown, 0 means no DPD, 1 means maximal did
during month between 1-30, 2 means DPD 31-60,  5 means DPD 120+ or sold or
written off ] )'],
      dtype=object)
```

```
[ ]: bureau = pd.read_csv('02_data/bureau.csv')
bureau_balance = pd.read_csv('02_data/bureau_balance.csv')
```

```
[ ]: bureau_balance.shape
```

```
[ ]: (27299925, 3)
```

```
[ ]: bureau_balance.shape[0] / 10 ** 3
```

```
[ ]: 27299.925
```

```
[ ]: bureau_balance.columns
```

```
[ ]: Index(['SK_ID_BUREAU', 'MONTHS_BALANCE', 'STATUS'], dtype='object')
```

```
[ ]: bureau.shape
```

```
[ ]: (1716428, 17)
```

```
[ ]: app_train.NAME_FAMILY_STATUS.value_counts()
```

```
[ ]: Married                 196432
Single / not married     45444
Civil marriage           29775
Separated                19770
```

```
Widow                       16088
Unknown                         2
Name: NAME_FAMILY_STATUS, dtype: int64
```

[ ]:
```
app_train.NAME_EDUCATION_TYPE.value_counts()
```

[ ]:
```
Secondary / secondary special    218391
Higher education                  74863
Incomplete higher                 10277
Lower secondary                    3816
Academic degree                     164
Name: NAME_EDUCATION_TYPE, dtype: int64
```

[ ]:
```
app_train.NAME_TYPE_SUITE.value_counts()
```

[ ]:
```
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
```

[ ]:
```
app_train['AGE'] = round(app_train['DAYS_BIRTH'] / - 365, 0).astype('int')
```

[ ]:
```
app_train.AGE
```

[ ]:
```
0         26
1         46
2         52
3         52
4         55
          ..
307506    26
307507    57
307508    41
307509    33
307510    46
Name: AGE, Length: 307511, dtype: int64
```

[ ]:
```
educ_type_dict = {
    'Lower secondary' : 0,
    'Secondary / secondary special': 1,
    'Incomplete higher': 2,
    'Higher education': 3,
    'Academic degree': 4
```

```
}

app_train.NAME_EDUCATION_TYPE.map(educ_type_dict)
```

[ ]:
```
0         1
1         3
2         1
3         1
4         1
         ..
307506    1
307507    1
307508    3
307509    1
307510    3
Name: NAME_EDUCATION_TYPE, Length: 307511, dtype: int64
```

[ ]:
```
app_train.corr()['TARGET'].sort_values()
```

[ ]:
```
EXT_SOURCE_3                  -0.178919
EXT_SOURCE_2                  -0.160472
EXT_SOURCE_1                  -0.155317
AGE                           -0.078263
DAYS_EMPLOYED                 -0.044932
                                 …
DAYS_LAST_PHONE_CHANGE         0.055218
REGION_RATING_CLIENT           0.058899
REGION_RATING_CLIENT_W_CITY    0.060893
DAYS_BIRTH                     0.078239
TARGET                         1.000000
Name: TARGET, Length: 107, dtype: float64
```

[ ]:
```
print(app_train.CODE_GENDER.unique())
print(app_train.CODE_GENDER.value_counts(normalize=True))
```

```
['M' 'F' 'XNA']
F      0.658344
M      0.341643
XNA    0.000013
Name: CODE_GENDER, dtype: float64
```

[ ]:
```
str_cols = [col for col in app_train.columns if app_train[col].dtype ==␣
 ↪'object']
```

[ ]:
```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
```

```
for col in str_cols:
    app_train[col] = le.fit_transform(app_train[col])
app_train
```

```
[ ]:         SK_ID_CURR  TARGET  NAME_CONTRACT_TYPE  CODE_GENDER  FLAG_OWN_CAR  \
0            100002       1                   0            1             0
1            100003       0                   0            0             0
2            100004       0                   1            1             1
3            100006       0                   0            0             0
4            100007       0                   0            1             0
…                 …     …                   …            …             …
307506       456251       0                   0            1             0
307507       456252       0                   0            0             0
307508       456253       0                   0            0             0
307509       456254       1                   0            0             0
307510       456255       0                   0            0             0

        FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  \
0                     1             0          202500.0    406597.5
1                     0             0          270000.0   1293502.5
2                     1             0           67500.0    135000.0
3                     1             0          135000.0    312682.5
4                     1             0          121500.0    513000.0
…                     …             …                 …           …
307506                0             0          157500.0    254700.0
307507                1             0           72000.0    269550.0
307508                1             0          153000.0    677664.0
307509                1             0          171000.0    370107.0
307510                0             0          157500.0    675000.0

        AMT_ANNUITY  …  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  \
0           24700.5  …                 0                 0
1           35698.5  …                 0                 0
2            6750.0  …                 0                 0
3           29686.5  …                 0                 0
4           21865.5  …                 0                 0
…                 …  …                 …                 …
307506      27558.0  …                 0                 0
307507      12001.5  …                 0                 0
307508      29979.0  …                 0                 0
307509      20205.0  …                 0                 0
307510      49117.5  …                 0                 0

        FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  \
0                      0                 0                         0.0
1                      0                 0                         0.0
2                      0                 0                         0.0
```

```
3                     0               0                     NaN
4                     0               0                     0.0
…                  …             …                    …
307506                0               0                     NaN
307507                0               0                     NaN
307508                0               0                     1.0
307509                0               0                     0.0
307510                0               0                     0.0

        AMT_REQ_CREDIT_BUREAU_DAY  AMT_REQ_CREDIT_BUREAU_WEEK  \
0                             0.0                         0.0
1                             0.0                         0.0
2                             0.0                         0.0
3                             NaN                         NaN
4                             0.0                         0.0
…                          …                        …
307506                        NaN                         NaN
307507                        NaN                         NaN
307508                        0.0                         0.0
307509                        0.0                         0.0
307510                        0.0                         0.0

        AMT_REQ_CREDIT_BUREAU_MON  AMT_REQ_CREDIT_BUREAU_QRT  \
0                             0.0                        0.0
1                             0.0                        0.0
2                             0.0                        0.0
3                             NaN                        NaN
4                             0.0                        0.0
…                          …                       …
307506                        NaN                        NaN
307507                        NaN                        NaN
307508                        1.0                        0.0
307509                        0.0                        0.0
307510                        2.0                        0.0

        AMT_REQ_CREDIT_BUREAU_YEAR
0                              1.0
1                              0.0
2                              0.0
3                              NaN
4                              0.0
…                           …
307506                         NaN
307507                         NaN
307508                         1.0
307509                         0.0
307510                         1.0
```

```
[307511 rows x 122 columns]
```

[ ]:
```
app_train.CODE_GENDER.value_counts()
```

[ ]:
```
0    202448
1    105059
2         4
Name: CODE_GENDER, dtype: int64
```

[ ]:
```
bureau = pd.read_csv('02_data/bureau.csv')
bureau
```

[ ]:
```
         SK_ID_CURR  SK_ID_BUREAU CREDIT_ACTIVE CREDIT_CURRENCY  DAYS_CREDIT  \
0            215354       5714462        Closed      currency 1         -497
1            215354       5714463        Active      currency 1         -208
2            215354       5714464        Active      currency 1         -203
3            215354       5714465        Active      currency 1         -203
4            215354       5714466        Active      currency 1         -629
…               …           …             …               …           …
1716423      259355       5057750        Active      currency 1          -44
1716424      100044       5057754        Closed      currency 1        -2648
1716425      100044       5057762        Closed      currency 1        -1809
1716426      246829       5057770        Closed      currency 1        -1878
1716427      246829       5057778        Closed      currency 1         -463

         CREDIT_DAY_OVERDUE  DAYS_CREDIT_ENDDATE  DAYS_ENDDATE_FACT  \
0                         0               -153.0             -153.0
1                         0               1075.0                NaN
2                         0                528.0                NaN
3                         0                  NaN                NaN
4                         0               1197.0                NaN
…                       …                   …                 …
1716423                   0                -30.0                NaN
1716424                   0              -2433.0            -2493.0
1716425                   0              -1628.0             -970.0
1716426                   0              -1513.0            -1513.0
1716427                   0                  NaN             -387.0

         AMT_CREDIT_MAX_OVERDUE  CNT_CREDIT_PROLONG  AMT_CREDIT_SUM  \
0                           NaN                   0        91323.00
1                           NaN                   0       225000.00
2                           NaN                   0       464323.50
3                           NaN                   0        90000.00
4                       77674.5                   0      2700000.00
…                           …                   …             …
1716423                     0.0                   0        11250.00
```

```
1716424                  5476.5                  0        38130.84
1716425                     NaN                  0        15570.00
1716426                     NaN                  0        36000.00
1716427                     NaN                  0        22500.00

         AMT_CREDIT_SUM_DEBT  AMT_CREDIT_SUM_LIMIT  AMT_CREDIT_SUM_OVERDUE  \
0                        0.0                   NaN                     0.0
1                   171342.0                   NaN                     0.0
2                        NaN                   NaN                     0.0
3                        NaN                   NaN                     0.0
4                        NaN                   NaN                     0.0
…                      …                   …                     …
1716423              11250.0                   0.0                     0.0
1716424                  0.0                   0.0                     0.0
1716425                  NaN                   NaN                     0.0
1716426                  0.0                   0.0                     0.0
1716427                  0.0                   NaN                     0.0

             CREDIT_TYPE  DAYS_CREDIT_UPDATE  AMT_ANNUITY
0        Consumer credit                -131          NaN
1            Credit card                 -20          NaN
2        Consumer credit                 -16          NaN
3            Credit card                 -16          NaN
4        Consumer credit                 -21          NaN
…                  …                 …          …
1716423        Microloan                 -19          NaN
1716424  Consumer credit               -2493          NaN
1716425  Consumer credit                -967          NaN
1716426  Consumer credit               -1508          NaN
1716427        Microloan                -387          NaN

[1716428 rows x 17 columns]
```

[ ]:
```
app_train.columns
```

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

[ ]:
```
bureau.columns
```

```
[ ]: Index(['SK_ID_CURR', 'SK_ID_BUREAU', 'CREDIT_ACTIVE', 'CREDIT_CURRENCY',
            'DAYS_CREDIT', 'CREDIT_DAY_OVERDUE', 'DAYS_CREDIT_ENDDATE',
            'DAYS_ENDDATE_FACT', 'AMT_CREDIT_MAX_OVERDUE', 'CNT_CREDIT_PROLONG',
            'AMT_CREDIT_SUM', 'AMT_CREDIT_SUM_DEBT', 'AMT_CREDIT_SUM_LIMIT',
            'AMT_CREDIT_SUM_OVERDUE', 'CREDIT_TYPE', 'DAYS_CREDIT_UPDATE',
            'AMT_ANNUITY'],
           dtype='object')
```

```
[ ]: col_desc.loc[col_desc.Row.eq('SK_ID_CURR') & col_desc.Table.eq('bureau.csv')].
      ↪Description.values
```

```
[ ]: array(['ID of loan in our sample - one loan in our sample can have 0,1,2 or more
     related previous credits in credit bureau '],
           dtype=object)
```

```
[ ]: bureau.shape
```

```
[ ]: (1716428, 17)
```

```
[ ]: len(bureau.SK_ID_BUREAU.unique())
```

```
[ ]: 1716428
```

```
[ ]: len(bureau.SK_ID_CURR.unique())
```

```
[ ]: 305811
```

```
[ ]: bureau[bureau.duplicated(subset=['SK_ID_CURR']) == True]
```

```
[ ]:          SK_ID_CURR  SK_ID_BUREAU CREDIT_ACTIVE CREDIT_CURRENCY  DAYS_CREDIT  \
     1           215354       5714463        Active      currency 1         -208
     2           215354       5714464        Active      currency 1         -203
     3           215354       5714465        Active      currency 1         -203
     4           215354       5714466        Active      currency 1         -629
     5           215354       5714467        Active      currency 1         -273
     …             …            …          …            …          …
     1716423     259355       5057750        Active      currency 1          -44
     1716424     100044       5057754        Closed      currency 1        -2648
     1716425     100044       5057762        Closed      currency 1        -1809
     1716426     246829       5057770        Closed      currency 1        -1878
     1716427     246829       5057778        Closed      currency 1         -463

              CREDIT_DAY_OVERDUE  DAYS_CREDIT_ENDDATE  DAYS_ENDDATE_FACT  \
     1                         0               1075.0                NaN
     2                         0                528.0                NaN
     3                         0                  NaN                NaN
     4                         0               1197.0                NaN
```

```
5                        0          27460.0              NaN
…                     …           …                …
1716423                  0            -30.0              NaN
1716424                  0          -2433.0          -2493.0
1716425                  0          -1628.0           -970.0
1716426                  0          -1513.0          -1513.0
1716427                  0              NaN           -387.0

         AMT_CREDIT_MAX_OVERDUE  CNT_CREDIT_PROLONG  AMT_CREDIT_SUM  \
1                           NaN                   0       225000.00
2                           NaN                   0       464323.50
3                           NaN                   0        90000.00
4                       77674.5                   0      2700000.00
5                           0.0                   0       180000.00
…                         …                 …              …
1716423                     0.0                   0        11250.00
1716424                  5476.5                   0        38130.84
1716425                     NaN                   0        15570.00
1716426                     NaN                   0        36000.00
1716427                     NaN                   0        22500.00

         AMT_CREDIT_SUM_DEBT  AMT_CREDIT_SUM_LIMIT  AMT_CREDIT_SUM_OVERDUE  \
1                  171342.00                   NaN                     0.0
2                        NaN                   NaN                     0.0
3                        NaN                   NaN                     0.0
4                        NaN                   NaN                     0.0
5                   71017.38             108982.62                     0.0
…                      …                   …                     …
1716423             11250.00                  0.00                     0.0
1716424                 0.00                  0.00                     0.0
1716425                  NaN                   NaN                     0.0
1716426                 0.00                  0.00                     0.0
1716427                 0.00                   NaN                     0.0

             CREDIT_TYPE  DAYS_CREDIT_UPDATE  AMT_ANNUITY
1            Credit card                 -20          NaN
2        Consumer credit                 -16          NaN
3            Credit card                 -16          NaN
4        Consumer credit                 -21          NaN
5            Credit card                 -31          NaN
…                    …                 …          …
1716423        Microloan                 -19          NaN
1716424  Consumer credit               -2493          NaN
1716425  Consumer credit                -967          NaN
1716426  Consumer credit               -1508          NaN
1716427        Microloan                -387          NaN
```

```
[1410617 rows x 17 columns]
```