01_Exploration.ipynb

September 11, 2021

1 Initialisation

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

from sklearn.preprocessing import OrdinalEncoder #, LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.compose import make_column_transformer
from sklearn.pipeline import make_pipeline

import warnings
warnings.simplefilter(action='ignore', category=UserWarning)

from FeatureNames import get_feature_names
```

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
```

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

[ ]:
```
   SK_ID_CURR  TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR  \
0      100002       1         Cash loans           M            N
1      100003       0         Cash loans           F            N
2      100004       0    Revolving loans           M            Y
3      100006       0         Cash loans           F            N
4      100007       0         Cash loans           M            N

  FLAG_OWN_REALTY  CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  \
0               Y             0          202500.0    406597.5      24700.5
1               N             0          270000.0   1293502.5      35698.5
2               Y             0           67500.0    135000.0       6750.0
3               Y             0          135000.0    312682.5      29686.5
4               Y             0          121500.0    513000.0      21865.5

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

  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY  \
0                        0.0                       0.0
1                        0.0                       0.0
2                        0.0                       0.0
3                        NaN                       NaN
4                        0.0                       0.0

   AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
0                         0.0                        0.0
1                         0.0                        0.0
2                         0.0                        0.0
3                         NaN                        NaN
4                         0.0                        0.0

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                        0.0                         1.0
```

```
1                         0.0                         0.0
2                         0.0                         0.0
3                         NaN                         NaN
4                         0.0                         0.0

[5 rows x 122 columns]
```

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

[ ]:
```
            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

           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

            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

           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
```

```
            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

            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

            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

[5 rows x 121 columns]
```

[ ]:

```
print('name_col' + '\t' + 'data_type' + '\t' + 'dimensionality' + '\t'
      + 'null_count' + '\t' + 'null_perct' + '\t'+ 'description')
for col in app_train.columns.tolist():
    column_typ = app_train[col].dtypes
    null_count = app_train[col].isna().sum()
    null_perct = null_count / app_train[col].isna().count()
    if app_train[col].dtype in ['object', 'int64']:
        dimensionality = app_train[col].nunique()
    else:
        dimensionality = np.nan
    desc = col_desc.loc[col_desc.Table.eq('application_{train|test}.csv')
                        & col_desc.Row.eq(col)].Description.tolist()[0]
    print(col + '\t'
          + str(column_typ) + '\t'
          + str(dimensionality) + '\t'
          + str(null_count) + '\t'
          + str(round(null_perct, 4) * 100) + '\t'
          + str(desc))
```

```
name_col        data_type       dimensionality  null_count      null_perct
description
```

```
TARGET  int64   2       0       0.0     Target variable (1 - client with payment
difficulties: he/she had late payment more than X days on at least one of the
first Y installments of the loan in our sample, 0 - all other cases)
NAME_CONTRACT_TYPE      object  2       0       0.0     Identification if loan
is cash or revolving
CODE_GENDER     object  3       0       0.0     Gender of the client
FLAG_OWN_CAR    object  2       0       0.0     Flag if the client owns a car
FLAG_OWN_REALTY object  2       0       0.0     Flag if client owns a house or
flat
CNT_CHILDREN    int64   15      0       0.0     Number of children the client
has
AMT_INCOME_TOTAL        float64 nan     0       0.0     Income of the client
AMT_CREDIT      float64 nan     0       0.0     Credit amount of the loan
AMT_ANNUITY     float64 nan     12      0.0     Loan annuity
AMT_GOODS_PRICE float64 nan     278     0.09    For consumer loans it is the
price of the goods for which the loan is given
NAME_TYPE_SUITE object  7       1292    0.42    Who was accompanying client when
he was applying for the loan
NAME_INCOME_TYPE        object  8       0       0.0     Clients income type
(businessman, working, maternity leave, )
NAME_EDUCATION_TYPE     object  5       0       0.0     Level of highest
education the client achieved
NAME_FAMILY_STATUS      object  6       0       0.0     Family status of the
client
NAME_HOUSING_TYPE       object  6       0       0.0     What is the housing
situation of the client (renting, living with parents, …)
REGION_POPULATION_RELATIVE      float64 nan     0       0.0     Normalized
population of region where client lives (higher number means the client lives in
more populated region)
DAYS_BIRTH      int64   17460   0       0.0     Client's age in days at the time
of application
DAYS_EMPLOYED   int64   12574   0       0.0     How many days before the
application the person started current employment
DAYS_REGISTRATION       float64 nan     0       0.0     How many days before the
application did client change his registration
DAYS_ID_PUBLISH int64   6168    0       0.0     How many days before the
application did client change the identity document with which he applied for
the loan
OWN_CAR_AGE     float64 nan     202929  65.99000000000001       Age of client's
car
FLAG_MOBIL      int64   2       0       0.0     Did client provide mobile phone
(1=YES, 0=NO)
FLAG_EMP_PHONE  int64   2       0       0.0     Did client provide work phone
(1=YES, 0=NO)
FLAG_WORK_PHONE int64   2       0       0.0     Did client provide home phone
(1=YES, 0=NO)
FLAG_CONT_MOBILE        int64   2       0       0.0     Was mobile phone
reachable (1=YES, 0=NO)
```

FLAG_PHONE   int64   2   0   0.0   Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL   int64   2   0   0.0   Did client provide email (1=YES, 0=NO)
OCCUPATION_TYPE object 18   96391   31.35   What kind of occupation does the client have
CNT_FAM_MEMBERS float64 nan   2   0.0   How many family members does client have
REGION_RATING_CLIENT   int64   3   0   0.0   Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY   int64   3   0   0.0   Our rating of the region where client lives with taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START   object 7   0   0.0   On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START int64   24   0   0.0   Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION   int64   2   0   0.0   Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
REG_REGION_NOT_WORK_REGION   int64   2   0   0.0   Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
LIVE_REGION_NOT_WORK_REGION   int64   2   0   0.0   Flag if client's contact address does not match work address (1=different, 0=same, at region level)
REG_CITY_NOT_LIVE_CITY int64   2   0   0.0   Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
REG_CITY_NOT_WORK_CITY int64   2   0   0.0   Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
LIVE_CITY_NOT_WORK_CITY int64   2   0   0.0   Flag if client's contact address does not match work address (1=different, 0=same, at city level)
ORGANIZATION_TYPE   object 58   0   0.0   Type of organization where client works
EXT_SOURCE_1   float64 nan   173378 56.379999999999995   Normalized score from external data source
EXT_SOURCE_2   float64 nan   660   0.21   Normalized score from external data source
EXT_SOURCE_3   float64 nan   60965   19.830000000000002   Normalized score from external data source
APARTMENTS_AVG float64 nan   156061 50.74999999999999   Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_AVG   float64 nan   179943 58.52   Normalized information about building where the client lives, What is average (_AVG suffix), modus

(_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_AVG float64 nan 150007 48.78 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_AVG float64 nan 204488 66.5 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_AVG float64 nan 214865 69.87 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_AVG float64 nan 163891 53.300000000000004 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_AVG float64 nan 154828 50.349999999999994 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_AVG float64 nan 153020 49.76 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_AVG float64 nan 208642 67.85 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_AVG float64 nan 182590 59.38 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_AVG float64 nan 210199 68.35 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

LIVINGAREA_AVG float64 nan 154350 50.19 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_AVG float64 nan 213514 69.43 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_AVG float64 nan 169682 55.17999999999999 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MODE float64 nan 156061 50.74999999999999 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MODE float64 nan 179943 58.52 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MODE float64 nan 150007 48.78 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MODE float64 nan 204488 66.5 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MODE float64 nan 214865 69.87 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MODE float64 nan 163891 53.300000000000004 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MODE float64 nan 154828 50.349999999999994 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common

area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MODE float64 nan 153020 49.76 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MODE float64 nan 208642 67.85 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MODE float64 nan 182590 59.38 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MODE float64 nan 210199 68.35 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MODE float64 nan 154350 50.19 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MODE float64 nan 213514 69.43 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MODE float64 nan 169682 55.17999999999999 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MEDI float64 nan 156061 50.74999999999999 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MEDI float64 nan 179943 58.52 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MEDI float64 nan 150007 48.78 Normalized

information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MEDI float64 nan 204488 66.5 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MEDI float64 nan 214865 69.87 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MEDI float64 nan 163891 53.300000000000004 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MEDI float64 nan 154828 50.349999999999994 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MEDI float64 nan 153020 49.76 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MEDI float64 nan 208642 67.85 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MEDI float64 nan 182590 59.38 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MEDI float64 nan 210199 68.35 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MEDI float64 nan 154350 50.19 Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building,

```
number of floor
NONLIVINGAPARTMENTS_MEDI        float64 nan     213514  69.43   Normalized
information about building where the client lives, What is average (_AVG
suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common
area, living area, age of building, number of elevators, number of entrances,
state of the building, number of floor
NONLIVINGAREA_MEDI      float64 nan     169682  55.17999999999999
Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size,
common area, living area, age of building, number of elevators, number of
entrances, state of the building, number of floor
FONDKAPREMONT_MODE      object  4       210295  68.39   Normalized information
about building where the client lives, What is average (_AVG suffix), modus
(_MODE suffix), median (_MEDI suffix) apartment size, common area, living area,
age of building, number of elevators, number of entrances, state of the
building, number of floor
HOUSETYPE_MODE  object  3       154297  50.18   Normalized information about
building where the client lives, What is average (_AVG suffix), modus (_MODE
suffix), median (_MEDI suffix) apartment size, common area, living area, age of
building, number of elevators, number of entrances, state of the building,
number of floor
TOTALAREA_MODE  float64 nan     148431  48.27   Normalized information about
building where the client lives, What is average (_AVG suffix), modus (_MODE
suffix), median (_MEDI suffix) apartment size, common area, living area, age of
building, number of elevators, number of entrances, state of the building,
number of floor
WALLSMATERIAL_MODE      object  7       156341  50.839999999999996
Normalized information about building where the client lives, What is average
(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size,
common area, living area, age of building, number of elevators, number of
entrances, state of the building, number of floor
EMERGENCYSTATE_MODE     object  2       145755  47.4    Normalized information
about building where the client lives, What is average (_AVG suffix), modus
(_MODE suffix), median (_MEDI suffix) apartment size, common area, living area,
age of building, number of elevators, number of entrances, state of the
building, number of floor
OBS_30_CNT_SOCIAL_CIRCLE        float64 nan     1021    0.33    How many
observation of client's social surroundings with observable 30 DPD (days past
due) default
DEF_30_CNT_SOCIAL_CIRCLE        float64 nan     1021    0.33    How many
observation of client's social surroundings defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE        float64 nan     1021    0.33    How many
observation of client's social surroundings with observable 60 DPD (days past
due) default
DEF_60_CNT_SOCIAL_CIRCLE        float64 nan     1021    0.33    How many
observation of client's social surroundings defaulted on 60 (days past due) DPD
DAYS_LAST_PHONE_CHANGE  float64 nan     1       0.0     How many days before
application did client change phone
```

```
FLAG_DOCUMENT_2 int64   2       0       0.0     Did client provide document 2
FLAG_DOCUMENT_3 int64   2       0       0.0     Did client provide document 3
FLAG_DOCUMENT_4 int64   2       0       0.0     Did client provide document 4
FLAG_DOCUMENT_5 int64   2       0       0.0     Did client provide document 5
FLAG_DOCUMENT_6 int64   2       0       0.0     Did client provide document 6
FLAG_DOCUMENT_7 int64   2       0       0.0     Did client provide document 7
FLAG_DOCUMENT_8 int64   2       0       0.0     Did client provide document 8
FLAG_DOCUMENT_9 int64   2       0       0.0     Did client provide document 9
FLAG_DOCUMENT_10        int64   2       0       0.0     Did client provide
document 10
FLAG_DOCUMENT_11        int64   2       0       0.0     Did client provide
document 11
FLAG_DOCUMENT_12        int64   2       0       0.0     Did client provide
document 12
FLAG_DOCUMENT_13        int64   2       0       0.0     Did client provide
document 13
FLAG_DOCUMENT_14        int64   2       0       0.0     Did client provide
document 14
FLAG_DOCUMENT_15        int64   2       0       0.0     Did client provide
document 15
FLAG_DOCUMENT_16        int64   2       0       0.0     Did client provide
document 16
FLAG_DOCUMENT_17        int64   2       0       0.0     Did client provide
document 17
FLAG_DOCUMENT_18        int64   2       0       0.0     Did client provide
document 18
FLAG_DOCUMENT_19        int64   2       0       0.0     Did client provide
document 19
FLAG_DOCUMENT_20        int64   2       0       0.0     Did client provide
document 20
FLAG_DOCUMENT_21        int64   2       0       0.0     Did client provide
document 21
AMT_REQ_CREDIT_BUREAU_HOUR      float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY       float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client one day before application
(excluding one hour before application)
AMT_REQ_CREDIT_BUREAU_WEEK      float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client one week before application
(excluding one day before application)
AMT_REQ_CREDIT_BUREAU_MON       float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client one month before application
(excluding one week before application)
AMT_REQ_CREDIT_BUREAU_QRT       float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client 3 month before application
(excluding one month before application)
AMT_REQ_CREDIT_BUREAU_YEAR      float64 nan     41519   13.5    Number of
enquiries to Credit Bureau about the client one day year (excluding last 3
```

```
months before application)
```

2.1.1 Encodage des colonnes textuelles

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

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

```
[ ]: print(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
```

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

```
['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']
```

[ ]:
```
desc = col_desc.loc[col_desc.Table.eq('application_{train|test}.csv')
                    & col_desc.Row.eq(col)].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.NAME_TYPE_SUITE.unique()
```

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

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

```
['Gender of the client']
```

[ ]:
```
app_train.FONDKAPREMONT_MODE
```

[ ]:
```
SK_ID_CURR
100002    reg oper account
100003    reg oper account
100004                 NaN
100006                 NaN
100007                 NaN
                 …
456251    reg oper account
456252    reg oper account
456253    reg oper account
456254                 NaN
456255                 NaN
Name: FONDKAPREMONT_MODE, Length: 307511, dtype: object
```

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

[ ]:
```
categ_const_input_prep = make_pipeline(SimpleImputer(strategy='constant',
                                                     fill_value='Unknown'),
                                       OrdinalEncoder())
```

```python
categ_const_imput_cols = ['OCCUPATION_TYPE', 'WALLSMATERIAL_MODE',
                          'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
                          'EMERGENCYSTATE_MODE']

categ_mfreq_imput_prep = make_pipeline(SimpleImputer(strategy='most_frequent'),
                                       OrdinalEncoder())

categ_mfreq_imput_cols = ['NAME_TYPE_SUITE']

weekdays = ['MONDAY', 'TUESDAY', 'WEDNESDAY', 'THURSDAY',
            'FRIDAY', 'SATURDAY', 'SUNDAY']
# * cos(2 * np.pi) / 7 # pour garder la notion de cyclicité
# * sin(2 * np.pi) / 7
educ_typ = ['Lower secondary', 'Secondary / secondary special',
            'Incomplete higher', 'Higher education', 'Academic degree']
ordinal_maps = [weekdays, educ_typ]

categ_ordinal_map_prep = make_pipeline(OrdinalEncoder(categories=ordinal_maps))

categ_ordinal_map_cols = ['WEEKDAY_APPR_PROCESS_START', 'NAME_EDUCATION_TYPE']

categ_separate_prep_cols = categ_const_imput_cols + categ_mfreq_imput_cols\
                           + categ_ordinal_map_cols

categorical_labels = app_train.select_dtypes('object').columns.tolist()
preprocessor = make_column_transformer(
    (categ_const_imput_prep, categ_const_imput_cols),
    (categ_mfreq_imput_prep, categ_mfreq_imput_cols),
    (categ_ordinal_map_prep, categ_ordinal_map_cols),
    (OrdinalEncoder(), [c for c in categorical_labels
                        if c not in categ_separate_prep_cols]),
    remainder='passthrough'
)
```

[ ]:
```python
app_train_processed = preprocessor.fit_transform(app_train)
```

[ ]:
```python
app_train.REO_.value_counts()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-57-b763811ba5f2> in <module>
----> 1 app_train.AMT_REQ_CREDIT_BUREAU_HOUR.value_counts()

/usr/bin/anaconda3/lib/python3.8/site-packages/pandas/core/generic.py in␣
 ↪__getattr__(self, name)
```

```
            if self._info_axis.
↪_can_hold_identifiers_and_holds_name(name):
                return self[name]
->          return object.__getattribute__(self, name)

    def __setattr__(self, name: str, value) -> None:

AttributeError: 'DataFrame' object has no attribute 'AMT_REQ_CREDIT_BUREAU_HOUR
```

[ ]:
```python
feature_names = get_feature_names(preprocessor)
```

[ ]:
```python
print(feature_names)
```

```
['pipeline-1__OCCUPATION_TYPE', 'pipeline-1__WALLSMATERIAL_MODE',
'pipeline-1__FONDKAPREMONT_MODE', 'pipeline-1__HOUSETYPE_MODE',
'pipeline-1__EMERGENCYSTATE_MODE', 'pipeline-2__NAME_TYPE_SUITE',
'pipeline-3__WEEKDAY_APPR_PROCESS_START', 'pipeline-3__NAME_EDUCATION_TYPE',
'ordinalencoder__NAME_CONTRACT_TYPE', 'ordinalencoder__CODE_GENDER',
'ordinalencoder__FLAG_OWN_CAR', 'ordinalencoder__FLAG_OWN_REALTY',
'ordinalencoder__NAME_INCOME_TYPE', 'ordinalencoder__NAME_FAMILY_STATUS',
'ordinalencoder__NAME_HOUSING_TYPE', 'ordinalencoder__ORGANIZATION_TYPE',
'TARGET', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY',
'AMT_GOODS_PRICE', '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', 'CNT_FAM_MEMBERS', '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', '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', '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', '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']
```

[ ]:
```python
new_feature_names = [name.split('__')[-1] for name in feature_names]
print(new_feature_names)
```

```
['OCCUPATION_TYPE', 'WALLSMATERIAL_MODE', 'FONDKAPREMONT_MODE',
'HOUSETYPE_MODE', 'EMERGENCYSTATE_MODE', 'NAME_TYPE_SUITE',
'WEEKDAY_APPR_PROCESS_START', 'NAME_EDUCATION_TYPE', 'NAME_CONTRACT_TYPE',
'CODE_GENDER', 'FLAG_OWN_CAR', 'FLAG_OWN_REALTY', 'NAME_INCOME_TYPE',
'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE', 'ORGANIZATION_TYPE', 'TARGET',
'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY',
'AMT_GOODS_PRICE', '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', 'CNT_FAM_MEMBERS', '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', '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', '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', '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']
```

[ ]:
```
df = pd.DataFrame(app_train_processed, columns = new_feature_names)
```

[ ]:
```
df
```

[ ]:
```
        OCCUPATION_TYPE  WALLSMATERIAL_MODE  FONDKAPREMONT_MODE  \
0                   8.0                 5.0                 3.0
1                   3.0                 0.0                 3.0
2                   8.0                 6.0                 0.0
3                   8.0                 6.0                 0.0
4                   3.0                 6.0                 0.0
…                   …                   …                   …
307506             14.0                 5.0                 3.0
307507             17.0                 5.0                 3.0
307508             10.0                 4.0                 3.0
307509              8.0                 5.0                 0.0
307510              8.0                 4.0                 0.0

        HOUSETYPE_MODE  EMERGENCYSTATE_MODE  NAME_TYPE_SUITE  \
0                  1.0                  0.0              6.0
1                  1.0                  0.0              1.0
2                  0.0                  1.0              6.0
3                  0.0                  1.0              6.0
4                  0.0                  1.0              6.0
…                  …                    …                …
307506             1.0                  0.0              6.0
307507             1.0                  0.0              6.0
307508             1.0                  0.0              6.0
307509             1.0                  0.0              6.0
307510             1.0                  0.0              6.0

        WEEKDAY_APPR_PROCESS_START  NAME_EDUCATION_TYPE  NAME_CONTRACT_TYPE  \
0                              2.0                  1.0                 0.0
1                              0.0                  3.0                 0.0
2                              0.0                  1.0                 1.0
3                              2.0                  1.0                 0.0
4                              3.0                  1.0                 0.0
…                              …                    …                   …
307506                         3.0                  1.0                 0.0
307507                         0.0                  1.0                 0.0
307508                         3.0                  3.0                 0.0
307509                         2.0                  1.0                 0.0
307510                         3.0                  3.0                 0.0

        CODE_GENDER  …  FLAG_DOCUMENT_18  FLAG_DOCUMENT_19  \
0               1.0  …               0.0               0.0
```

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

        FLAG_DOCUMENT_20  FLAG_DOCUMENT_21  AMT_REQ_CREDIT_BUREAU_HOUR  \
0                    0.0               0.0                         0.0
1                    0.0               0.0                         0.0
2                    0.0               0.0                         0.0
3                    0.0               0.0                         NaN
4                    0.0               0.0                         0.0
…                      …                 …                           …
307506               0.0               0.0                         NaN
307507               0.0               0.0                         NaN
307508               0.0               0.0                         1.0
307509               0.0               0.0                         0.0
307510               0.0               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 121 columns]
```

2.1.2 Étude des variables

[ ]:
```python
feat_corr = df.corr()['TARGET'].sort_values()
```

[ ]:
```python
print(feat_corr[:10])
print(feat_corr[-10:])
```

```
EXT_SOURCE_3                  -0.178919
EXT_SOURCE_2                  -0.160472
EXT_SOURCE_1                  -0.155317
NAME_EDUCATION_TYPE           -0.056872
DAYS_EMPLOYED                 -0.044932
FLOORSMAX_AVG                 -0.044003
FLOORSMAX_MEDI                -0.043768
FLOORSMAX_MODE                -0.043226
AMT_GOODS_PRICE               -0.039645
REGION_POPULATION_RELATIVE    -0.037227
Name: TARGET, dtype: float64
FLAG_EMP_PHONE                  0.045982
NAME_INCOME_TYPE                0.046829
REG_CITY_NOT_WORK_CITY          0.050994
DAYS_ID_PUBLISH                 0.051457
CODE_GENDER                     0.054692
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, dtype: float64
```

[ ]:
```
app_train.EXT_SOURCE_1.dtype
```

[ ]:
```
dtype('float64')
```

[ ]:
```
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
```

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)
```

[ ]:
```
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]
```

[ ]:
```
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]
```