home_credit_default_risk

June 6, 2018

1 Kaggle Home Credit Default Risk Competition

https://www.kaggle.com/c/home-credit-default-risk

This competition is sponsored by Home Credit, whose mission is to provide a positive and safe borrowing experience to groups of people that traditional, mainstream banks and financial institutions typically refuse to serve.

Home Credit targets a demographic that typically has no recourse but to deal with shady characters such as loan sharks when borrowing money. Many of these unbanked individuals are hard-working, well-intentioned folks who, either due to circumstances beyond their control or past mistakes, have fallen through the financial system's cracks.

Home Credit needs an algorithm that will take as inputs various financial and personal information originally taken from a loan applicant's profile, and then determine and output a probability of the applicant eventually repaying the loan. This probability will be in the range [0.0, 1.0], where 1.0 represents a 100% certainty that the applicant will repay the loan and 0.0 indicates that there is zero chance that the applicant will eventually repay. The algorithm will be tested on a set of 48,744 individuals who previously borrowed from Home Credit. A CSV file must be produced that contains one header row, and 48,744 prediction rows, where each prediction row contains both a user ID, the `SKI_ID_CURR` column, and the probability, the `TARGET` column, of that user repaying their loan. The file must be formatted as follows:

```
SK_ID_CURR,TARGET
100001,0.1
100005,0.9
100013,0.2
etc.
```

Home Credit knows which borrowers ultimately paid off their loans, and which ones eventually defaulted. A good algorithm will need to predict a high probability of repayment for the majority of borrowers who did successfully repay their loans. This algorithm will also need to predict a low probability of repayment for the majority of borrowers who eventually defaulted on their loans.

1.1 I. Data Exploration

```
In [1]: # Import libraries necessary for this project
        import numpy as np
        import pandas as pd
        from time import time
```

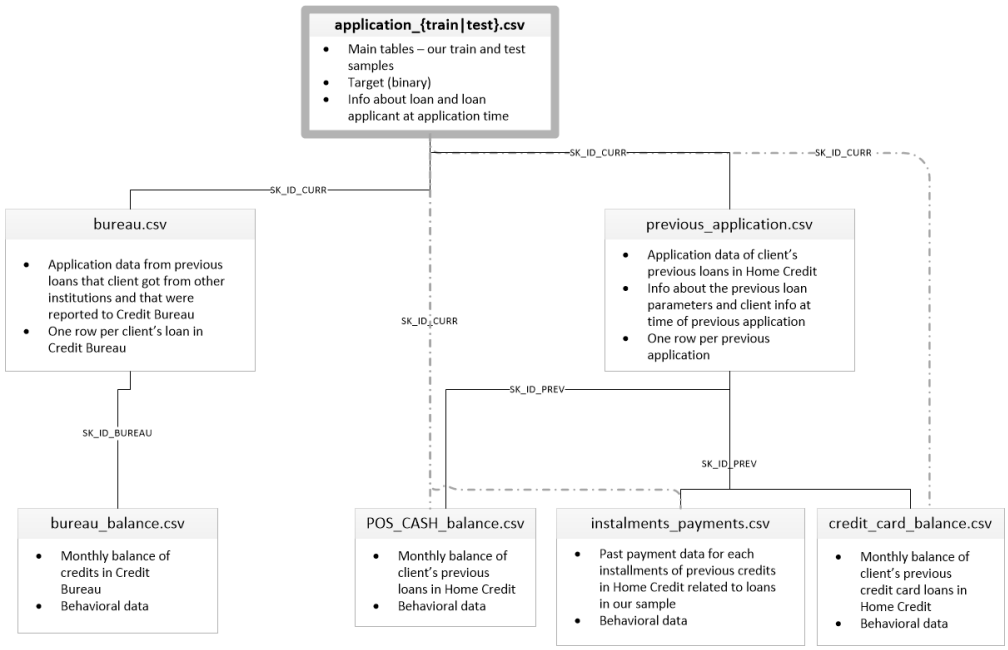
```
from IPython.display import display # Allows the use of display() for DataFrames

# Pretty display for notebooks
%matplotlib inline
```

1.1.1 Data Description

From https://www.kaggle.com/c/home-credit-default-risk/data:

1. **application_{train | test}.csv**

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- Static data for all applications. One row represents one loan in our data sample.
- **bureau.csv**
- All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
- For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
- **bureau_balance.csv**
- Monthly balances of previous credits in Credit Bureau.
- This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.
- **previous_application.csv**
- All previous applications for Home Credit loans of clients who have loans in our sample.
- There is one row for each previous application related to loans in our data sample.
- **POS_CASH_balance.csv**
- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months in which we have some history observable for the previous credits) rows.
- **installments_payments.csv**
- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.



alt text

- There is a) one row for every payment that was made plus b) one row each for missed payment.
- One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
- **credit_card_balance.csv**
- Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.

```
In [13]: # Load the data tables
         application_train_data = pd.read_csv("data/application_train.csv")
         application_test_data = pd.read_csv("data/application_test.csv")
         bureau_data = pd.read_csv("data/bureau.csv")
         bureau_balance_data = pd.read_csv("data/bureau_balance.csv")
         previous_application_data = pd.read_csv("data/previous_application.csv")
         POS_CASH_balance_data = pd.read_csv("data/POS_CASH_balance.csv")
         installments_payments_data = pd.read_csv("data/installments_payments.csv")
         credit_card_balance_data = pd.read_csv("data/credit_card_balance.csv")
```

1.1.2 1. Main Data Table (application_{train|test}.csv)

```
In [14]: # Display the first five records
         display(application_train_data.head(n=5))
```

```
   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  \
0             ...                             0                0
1             ...                             0                0
2             ...                             0                0
3             ...                             0                0
4             ...                             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

  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

   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

   AMT_REQ_CREDIT_BUREAU_YEAR
0                         1.0
1                         0.0
2                         0.0
3                         NaN
4                         0.0
```

```
[5 rows x 122 columns]
```

```
In [15]: # Total number of entries in training group
         print("Total number of entries in training group: {}".format(application_train_data.sh
```

```
Total number of entries in training group: 307511
```

```
In [16]: # Total number of entries in test group
         print("Total number of entries in test group: {}".format(application_test_data.shape[0
```

```
Total number of entries in test group: 48744
```

Main Data Table Featureset Exploration

1. **SK_ID_CURR**: ID of loan in our sample

- **TARGET**: 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**: Identification if loan is cash or revolving

- **CODE_GENDER**: Gender of the client
- **FLAG_OWN_CAR**: Flag if the client owns a car

- **FLAG_OWN_REALTY**: Flag if client owns a house or flat

- **CNT_CHILDREN**: Number of children the client has

- **AMT_INCOME_TOTAL**: Income of the client

- **AMT_CREDIT**: Credit amount of the loan
- **AMT_ANNUITY**: Loan annuity
- **AMT_GOODS_PRICE**: For consumer loans it is the price of the goods for which the loan is given

- **NAME_TYPE_SUITE**: Who was accompanying client when he was applying for the loan

- **NAME_INCOME_TYPE**: Clients income type (businessman, working, maternity leave,Ö)

- **NAME_EDUCATION_TYPE**: Level of highest education the client achieved
- **NAME_FAMILY_STATUS**: Family status of the client

- **NAME_HOUSING_TYPE**: What is the housing situation of the client (renting, living with parents, ...)

- **REGION_POPULATION_RELATIVE**: Normalized population of region where client lives (higher number means the client lives in more populated region) -- normalized
- **DAYS_BIRTH**: Client's age in days at the time of application -- time only relative to the application
- **DAYS_EMPLOYED**: How many days before the application the person started current employment -- time only relative to the application
- **DAYS_REGISTRATION**: How many days before the application did client change his registration -- time only relative to the application
- **DAYS_ID_PUBLISH**: How many days before the application did client change the identity document with which he applied for the loan -- time only relative to the application
- **OWN_CAR_AGE**: Age of client's car

- **FLAG_MOBIL**: Did client provide mobile phone (1=YES, 0=NO)
- **FLAG_EMP_PHONE**: Did client provide work phone (1=YES, 0=NO)

- **FLAG_WORK_PHONE**: Did client provide home phone (1=YES, 0=NO)

- **FLAG_CONT_MOBILE**: Was mobile phone reachable (1=YES, 0=NO)

- **FLAG_PHONE**: Did client provide home phone (1=YES, 0=NO)

- **FLAG_EMAIL**: Did client provide email (1=YES, 0=NO)

- **OCCUPATION_TYPE**: What kind of occupation does the client have
- **CNT_FAM_MEMBERS**: How many family members does client have
- **REGION_RATING_CLIENT**: Our rating of the region where client lives (1,2,3)
- **REGION_RATING_CLIENT_W_CITY**: Our rating of the region where client lives with taking city into account (1,2,3)

- **WEEKDAY_APPR_PROCESS_START**: On which day of the week did the client apply for the loan

- **HOUR_APPR_PROCESS_START**: Approximately at what hour did the client apply for the loan rounded
- **REG_REGION_NOT_LIVE_REGION**: Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)

- **REG_REGION_NOT_WORK_REGION**: Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
- **LIVE_REGION_NOT_WORK_REGION**: Flag if client's contact address does not match work address (1=different, 0=same, at region level)

- **REG_CITY_NOT_LIVE_CITY**: Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)

- **REG_CITY_NOT_WORK_CITY**: Flag if client's permanent address does not match work address (1=different, 0=same, at city level)

- **LIVE_CITY_NOT_WORK_CITY**: Flag if client's contact address does not match work address (1=different, 0=same, at city level)

- **ORGANIZATION_TYPE**: Type of organization where client works

- **EXT_SOURCE_1**: Normalized score from external data source -- normalized
- **EXT_SOURCE_2**: Normalized score from external data source -- normalized
- **EXT_SOURCE_3**: Normalized score from external data source -- normalized
- **APARTMENTS_AVG**: 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 -- normalized
- **BASEMENTAREA_AVG**: 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 -- normalized
- **YEARS_BEGINEXPLUATATION_AVG**: 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 -- normalized
- **YEARS_BUILD_AVG**: 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 -- normalized
- **COMMONAREA_AVG**: 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 -- normalized
- **ELEVATORS_AVG**: 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 -- normalized
- **ENTRANCES_AVG**: 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 -- normalized
- **FLOORSMAX_AVG**: 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 -- normalized
- **FLOORSMIN_AVG**: 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 -- normalized
- **LANDAREA_AVG**: 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 -- normalized

- **LIVINGAPARTMENTS_AVG**: 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 -- normalized
- **LIVINGAREA_AVG**: 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 -- normalized
- **NONLIVINGAPARTMENTS_AVG**: 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 -- normalized
- **NONLIVINGAREA_AVG**: 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 -- normalized
- **APARTMENTS_MODE**: 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 -- normalized
- **BASEMENTAREA_MODE**: 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 -- normalized
- **YEARS_BEGINEXPLUATATION_MODE**: 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 -- normalized
- **YEARS_BUILD_MODE**: 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 -- normalized
- **COMMONAREA_MODE**: 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 -- normalized
- **ELEVATORS_MODE**: 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 -- normalized
- **ENTRANCES_MODE**: 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 -- normalized
- **FLOORSMAX_MODE**: 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 -- normalized

- **FLOORSMIN_MODE**: 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 -- normalized
- **LANDAREA_MODE**: 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 -- normalized
- **LIVINGAPARTMENTS_MODE**: 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 -- normalized
- **LIVINGAREA_MODE**: 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 -- normalized
- **NONLIVINGAPARTMENTS_MODE**: 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 -- normalized
- **NONLIVINGAREA_MODE**: 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 -- normalized
- **APARTMENTS_MEDI**: 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 -- normalized
- **BASEMENTAREA_MEDI**: 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 -- normalized
- **YEARS_BEGINEXPLUATATION_MEDI**: 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 -- normalized
- **YEARS_BUILD_MEDI**: 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 -- normalized
- **COMMONAREA_MEDI**: 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 -- normalized
- **ELEVATORS_MEDI**: 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 -- normalized

- **ENTRANCES_MEDI**: 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 -- normalized
- **FLOORSMAX_MEDI**: 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 -- normalized
- **FLOORSMIN_MEDI**: 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 -- normalized
- **LANDAREA_MEDI**: 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 -- normalized
- **LIVINGAPARTMENTS_MEDI**: 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 -- normalized
- **LIVINGAREA_MEDI**: 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 -- normalized
- **NONLIVINGAPARTMENTS_MEDI**: 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 -- normalized
- **NONLIVINGAREA_MEDI**: 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 -- normalized
- **FONDKAPREMONT_MODE**: 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 -- normalized
- **HOUSETYPE_MODE**: 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 -- normalized
- **TOTALAREA_MODE**: 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 -- normalized
- **WALLSMATERIAL_MODE**: 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 -- normalized

- **EMERGENCYSTATE_MODE**: 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 -- normalized
- **OBS_30_CNT_SOCIAL_CIRCLE**: How many observation of client's social surroundings with observable 30 DPD (days past due) default
- **DEF_30_CNT_SOCIAL_CIRCLE**: How many observation of client's social surroundings defaulted on 30 DPD (days past due)

- **OBS_60_CNT_SOCIAL_CIRCLE**: How many observation of client's social surroundings with observable 60 DPD (days past due) default
- **DEF_60_CNT_SOCIAL_CIRCLE**: How many observation of client's social surroundings defaulted on 60 (days past due) DPD

- **DAYS_LAST_PHONE_CHANGE**: How many days before application did client change phone

- **FLAG_DOCUMENT_2**: Did client provide document 2

- **FLAG_DOCUMENT_3**: Did client provide document 3

- **FLAG_DOCUMENT_4**: Did client provide document 4

- **FLAG_DOCUMENT_5**: Did client provide document 5

- **FLAG_DOCUMENT_6**: Did client provide document 6

- **FLAG_DOCUMENT_7**: Did client provide document 7

- **FLAG_DOCUMENT_8**: Did client provide document 8

- **FLAG_DOCUMENT_9**: Did client provide document 9

- **FLAG_DOCUMENT_10**: Did client provide document 10

- **FLAG_DOCUMENT_11**: Did client provide document 11

- **FLAG_DOCUMENT_12**: Did client provide document 12

- **FLAG_DOCUMENT_13**: Did client provide document 13

- **FLAG_DOCUMENT_14**: Did client provide document 14

- **FLAG_DOCUMENT_15**: Did client provide document 15

- **FLAG_DOCUMENT_16**: Did client provide document 16

- **FLAG_DOCUMENT_17**: Did client provide document 17

- **FLAG_DOCUMENT_18**: Did client provide document 18
- **FLAG_DOCUMENT_19**: Did client provide document 19
- **FLAG_DOCUMENT_20**: Did client provide document 20
- **FLAG_DOCUMENT_21**: Did client provide document 21
- **AMT_REQ_CREDIT_BUREAU_HOUR**: Number of enquiries to Credit Bureau about the client one hour before application
- **AMT_REQ_CREDIT_BUREAU_DAY**: Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
- **AMT_REQ_CREDIT_BUREAU_WEEK**: Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
- **AMT_REQ_CREDIT_BUREAU_MON**: Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
- **AMT_REQ_CREDIT_BUREAU_QRT**: Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
- **AMT_REQ_CREDIT_BUREAU_YEAR**: Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

1.1.3 2. Bureau Data Table (bureau_balance.csv)

```
In [25]: # Display the first five records
         display(bureau_data.head(n=5))
```

```
   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

   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

   AMT_CREDIT_MAX_OVERDUE  CNT_CREDIT_PROLONG  AMT_CREDIT_SUM  \
0                     NaN                   0         91323.0
1                     NaN                   0        225000.0
2                     NaN                   0        464323.5
```

```
3                    NaN                   0         90000.0
4                77674.5                   0       2700000.0

   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

       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
```

Bureau Data Table Featureset Exploration

1. **SK_ID_CURR**: ID of loan in our sample - one loan in our sample can have 0,1,2 or more related previous credits in credit bureau -- hashed

- **SK_BUREAU_ID**: Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application) -- hashed
- **CREDIT_ACTIVE**: Status of the Credit Bureau (CB) reported credits

- **CREDIT_CURRENCY**: Recoded currency of the Credit Bureau credit -- recoded
- **DAYS_CREDIT**: How many days before current application did client apply for Credit Bureau credit -- time only relative to the application
- **CREDIT_DAY_OVERDUE**: Number of days past due on CB credit at the time of application for related loan in our sample

- **DAYS_CREDIT_ENDDATE**: Remaining duration of CB credit (in days) at the time of application in Home Credit -- time only relative to the application
- **DAYS_ENDDATE_FACT**: Days since CB credit ended at the time of application in Home Credit (only for closed credit) -- time only relative to the application
- **AMT_CREDIT_MAX_OVERDUE**: Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)
- **CNT_CREDIT_PROLONG**: How many times was the Credit Bureau credit prolonged
- **AMT_CREDIT_SUM**: Current credit amount for the Credit Bureau credit
- **AMT_CREDIT_SUM_DEBT**: Current debt on Credit Bureau credit
- **AMT_CREDIT_SUM_LIMIT**: Current credit limit of credit card reported in Credit Bureau

- **AMT_CREDIT_SUM_OVERDUE**: Current amount overdue on Credit Bureau credit

- **CREDIT_TYPE**: Type of Credit Bureau credit (Car, cash,...)
- **DAYS_CREDIT_UPDATE**: How many days before loan application did last information about the Credit Bureau credit come -- time only relative to the application

- **AMT_ANNUITY**: Annuity of the Credit Bureau credit

1.1.4 3. Bureau Balance Data Table (bureau_balance.csv)

```
In [24]: # Display the first five records
         display(bureau_balance_data.head(n=5))
```

```
   SK_ID_BUREAU  MONTHS_BALANCE STATUS
0       5715448               0      C
1       5715448              -1      C
2       5715448              -2      C
3       5715448              -3      C
4       5715448              -4      C
```

Bureau Balance Data Table Featureset Exploration

1. **SK_BUREAU_ID**: Recoded ID of Credit Bureau credit (unique coding for each application) - use this to join to CREDIT_BUREAU table -- hashed

- **MONTHS_BALANCE**: Month of balance relative to application date (-1 means the freshest balance date) -- time only relative to the application
- **STATUS**: Status of Credit Bureau loan during the month (active, closed, DPD0-30,Ö [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,Ö 5 means DPD 120+ or sold or written off])

1.1.5 4. Previous Application Data Table (previous_application.csv)

```
In [26]: # Display the first five records
         display(previous_application_data.head(n=5))
```

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

   AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE WEEKDAY_APPR_PROCESS_START  \
0     17145.0               0.0          17145.0                   SATURDAY
1    679671.0               NaN         607500.0                   THURSDAY
2    136444.5               NaN         112500.0                    TUESDAY
3    470790.0               NaN         450000.0                     MONDAY
4    404055.0               NaN         337500.0                   THURSDAY

   HOUR_APPR_PROCESS_START            ...             NAME_SELLER_INDUSTRY  \
0                       15            ...                     Connectivity
1                       11            ...                              XNA
2                       11            ...                              XNA
```

```
3                        7           ...                            XNA
4                        9           ...                            XNA

   CNT_PAYMENT  NAME_YIELD_GROUP       PRODUCT_COMBINATION  \
0         12.0            middle  POS mobile with interest
1         36.0        low_action          Cash X-Sell: low
2         12.0              high         Cash X-Sell: high
3         12.0            middle       Cash X-Sell: middle
4         24.0              high         Cash Street: high

   DAYS_FIRST_DRAWING DAYS_FIRST_DUE DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  \
0            365243.0          -42.0                     300.0          -42.0
1            365243.0         -134.0                     916.0       365243.0
2            365243.0         -271.0                      59.0       365243.0
3            365243.0         -482.0                    -152.0         -182.0
4                 NaN            NaN                       NaN            NaN

  DAYS_TERMINATION NFLAG_INSURED_ON_APPROVAL
0            -37.0                       0.0
1         365243.0                       1.0
2         365243.0                       1.0
3           -177.0                       1.0
4              NaN                       NaN

[5 rows x 37 columns]
```

Previous Application Data Table Featureset Exploration

1. **SK_ID_PREV**: ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loan applications in Home Credit, previous application could, but not necessarily have to lead to credit) -- hashed

- **SK_ID_CURR**: ID of loan in our sample -- hashed
- **NAME_CONTRACT_TYPE**: Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application

- **AMT_ANNUITY**: Annuity of previous application

- **AMT_APPLICATION**: For how much credit did client ask on the previous application

- **AMT_CREDIT**: Final credit amount on the previous application. This differs from AMT_APPLICATION in a way that the AMT_APPLICATION is the amount for which the client initially applied for, but during our approval process he could have received different amount - AMT_CREDIT

- **AMT_DOWN_PAYMENT**: Down payment on the previous application

- **AMT_GOODS_PRICE**: Goods price of good that client asked for (if applicable) on the previous application

- **WEEKDAY_APPR_PROCESS_START**: On which day of the week did the client apply for previous application

- **HOUR_APPR_PROCESS_START**: Approximately at what day hour did the client apply for the previous application -- rounded
- **FLAG_LAST_APPL_PER_CONTRACT**: Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract

- **NFLAG_LAST_APPL_IN_DAY**: Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice

- **RATE_DOWN_PAYMENT**: Down payment rate normalized on previous credit -- normalized
- **RATE_INTEREST_PRIMARY**: Interest rate normalized on previous credit -- normalized
- **RATE_INTEREST_PRIVILEGED**: Interest rate normalized on previous credit -- normalized
- **NAME_CASH_LOAN_PURPOSE**: Purpose of the cash loan

- **NAME_CONTRACT_STATUS**: Contract status (approved, cancelled, ...) of previous application

- **DAYS_DECISION**: Relative to current application when was the decision about previous application made time only relative to the application
- **NAME_PAYMENT_TYPE**: Payment method that client chose to pay for the previous application

- **CODE_REJECT_REASON**: Why was the previous application rejected
- **NAME_TYPE_SUITE**: Who accompanied client when applying for the previous application

- **NAME_CLIENT_TYPE**: Was the client old or new client when applying for the previous application
- **NAME_GOODS_CATEGORY**: What kind of goods did the client apply for in the previous application

- **NAME_PORTFOLIO**: Was the previous application for CASH, POS, CAR, Ö
- **NAME_PRODUCT_TYPE**: Was the previous application x-sell o walk-in

- **CHANNEL_TYPE**: Through which channel we acquired the client on the previous application

- **SELLERPLACE_AREA**: Selling area of seller place of the previous application

- **NAME_SELLER_INDUSTRY**: The industry of the seller

- **CNT_PAYMENT**: Term of previous credit at application of the previous application

- **NAME_YIELD_GROUP**: Grouped interest rate into small medium and high of the previous application -- grouped
- **PRODUCT_COMBINATION**: Detailed product combination of the previous application
- **DAYS_FIRST_DRAWING**: Relative to application date of current application when was the first disbursement of the previous application -- time only relative to the application
- **DAYS_FIRST_DUE**: Relative to application date of current application when was the first due supposed to be of the previous application -- time only relative to the application
- **DAYS_LAST_DUE_1ST_VERSION**: Relative to application date of current application when was the first due of the previous application -- time only relative to the application
- **DAYS_LAST_DUE**: Relative to application date of current application when was the last due date of the previous application -- time only relative to the application
- **DAYS_TERMINATION**: Relative to application date of current application when was the expected termination of the previous application -- time only relative to the application
- **NFLAG_INSURED_ON_APPROVAL**: Did the client requested insurance during the previous application

1.1.6 5. POS CASH Balance Data Table (POS_CASH_balance.csv)

```
In [20]: # Display the first five records
         display(POS_CASH_balance_data.head(n=5))
```

```
   SK_ID_PREV  SK_ID_CURR  MONTHS_BALANCE  CNT_INSTALMENT  \
0     1803195      182943             -31            48.0
1     1715348      367990             -33            36.0
2     1784872      397406             -32            12.0
3     1903291      269225             -35            48.0
4     2341044      334279             -35            36.0

   CNT_INSTALMENT_FUTURE NAME_CONTRACT_STATUS  SK_DPD  SK_DPD_DEF
0                   45.0              Active       0           0
1                   35.0              Active       0           0
2                    9.0              Active       0           0
3                   42.0              Active       0           0
4                   35.0              Active       0           0
```

POS CASH Balance Data Table Featureset Exploration

1. **SK_ID_PREV**: ID of previous credit in Home Credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit)

- **SK_ID_CURR**: ID of loan in our sample

- **MONTHS_BALANCE**: Month of balance relative to application date (-1 means the information to the freshest monthly snapshot, 0 means the information at application - often it will be the same as -1 as many banks are not updating the information to Credit Bureau regularly) -- time only relative to the application

- **CNT_INSTALMENT**: Term of previous credit (can change over time)
- **CNT_INSTALMENT_FUTURE**: Installments left to pay on the previous credit
- **NAME_CONTRACT_STATUS**: Contract status during the month
- **SK_DPD**: DPD (days past due) during the month of previous credit
- **SK_DPD_DEF**: DPD during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

1.1.7 6. Installments Payments Data Table (installments_payments.csv)

```
In [21]: # Display the first five records
         display(installments_payments_data.head(n=5))
```

	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	\
0	1054186	161674	1.0	6	
1	1330831	151639	0.0	34	
2	2085231	193053	2.0	1	
3	2452527	199697	1.0	3	
4	2714724	167756	1.0	2	

	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT	AMT_PAYMENT
0	-1180.0	-1187.0	6948.360	6948.360
1	-2156.0	-2156.0	1716.525	1716.525
2	-63.0	-63.0	25425.000	25425.000
3	-2418.0	-2426.0	24350.130	24350.130
4	-1383.0	-1366.0	2165.040	2160.585

Installments Payments Data Table Featureset Exploration

1. **SK_ID_PREV**: ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit) -- hashed

- **SK_ID_CURR**: ID of loan in our sample -- hashed
- **NUM_INSTALMENT_VERSION**: Version of installment calendar (0 is for credit card) of previous credit. Change of installment version from month to month signifies that some parameter of payment calendar has changed
- **NUM_INSTALMENT_NUMBER**: On which installment we observe payment
- **DAYS_INSTALMENT**: When the installment of previous credit was supposed to be paid (relative to application date of current loan) -- time only relative to the application
- **DAYS_ENTRY_PAYMENT**: When was the installments of previous credit paid actually (relative to application date of current loan) -- time only relative to the application
- **AMT_INSTALMENT**: What was the prescribed installment amount of previous credit on this installment
- **AMT_PAYMENT**: What the client actually paid on previous credit on this installment

1.1.8 7. Credit Card Balance Data Table (credit_card_balance.csv)

In [22]:
```
# Display the first five records
display(credit_card_balance_data.head(n=5))
```

```
   SK_ID_PREV  SK_ID_CURR  MONTHS_BALANCE  AMT_BALANCE  \
0     2562384      378907              -6       56.970
1     2582071      363914              -1    63975.555
2     1740877      371185              -7    31815.225
3     1389973      337855              -4   236572.110
4     1891521      126868              -1   453919.455

   AMT_CREDIT_LIMIT_ACTUAL  AMT_DRAWINGS_ATM_CURRENT  AMT_DRAWINGS_CURRENT  \
0                   135000                       0.0                 877.5
1                    45000                    2250.0                2250.0
2                   450000                       0.0                   0.0
3                   225000                    2250.0                2250.0
4                   450000                       0.0               11547.0

   AMT_DRAWINGS_OTHER_CURRENT  AMT_DRAWINGS_POS_CURRENT  \
0                         0.0                     877.5
1                         0.0                       0.0
2                         0.0                       0.0
3                         0.0                       0.0
4                         0.0                   11547.0

   AMT_INST_MIN_REGULARITY     ...      AMT_RECIVABLE  AMT_TOTAL_RECEIVABLE  \
0                 1700.325     ...              0.000                 0.000
1                 2250.000     ...          64875.555             64875.555
2                 2250.000     ...          31460.085             31460.085
3                11795.760     ...         233048.970            233048.970
4                22924.890     ...         453919.455            453919.455

   CNT_DRAWINGS_ATM_CURRENT  CNT_DRAWINGS_CURRENT  CNT_DRAWINGS_OTHER_CURRENT  \
0                       0.0                     1                         0.0
1                       1.0                     1                         0.0
2                       0.0                     0                         0.0
3                       1.0                     1                         0.0
4                       0.0                     1                         0.0

   CNT_DRAWINGS_POS_CURRENT  CNT_INSTALMENT_MATURE_CUM  NAME_CONTRACT_STATUS  \
0                       1.0                       35.0                Active
1                       0.0                       69.0                Active
2                       0.0                       30.0                Active
3                       0.0                       10.0                Active
4                       1.0                      101.0                Active

   SK_DPD  SK_DPD_DEF
0       0           0
```

```
1       0           0
2       0           0
3       0           0
4       0           0

[5 rows x 23 columns]
```

Credit Card Balance Data Table Featureset Exploration

1. **SK_ID_PREV**: ID of previous credit in Home credit related to loan in our sample. (One loan in our sample can have 0,1,2 or more previous loans in Home Credit) -- hashed

- **SK_ID_CURR**: ID of loan in our sample -- hashed
- **MONTHS_BALANCE**: Month of balance relative to application date (-1 means the freshest balance date) -- time only relative to the application
- **AMT_BALANCE**: Balance during the month of previous credit

- **AMT_CREDIT_LIMIT_ACTUAL**: Credit card limit during the month of the previous credit

- **AMT_DRAWINGS_ATM_CURRENT**: Amount drawing at ATM during the month of the previous credit

- **AMT_DRAWINGS_CURRENT**: Amount drawing during the month of the previous credit

- **AMT_DRAWINGS_OTHER_CURRENT**: Amount of other drawings during the month of the previous credit

- **AMT_DRAWINGS_POS_CURRENT**: Amount drawing or buying goods during the month of the previous credit

- **AMT_INST_MIN_REGULARITY**: Minimal installment for this month of the previous credit

- **AMT_PAYMENT_CURRENT**: How much did the client pay during the month on the previous credit

- **AMT_PAYMENT_TOTAL_CURRENT**: How much did the client pay during the month in total on the previous credit

- **AMT_RECEIVABLE_PRINCIPAL**: Amount receivable for principal on the previous credit

- **AMT_RECIVABLE**: Amount receivable on the previous credit

- **AMT_TOTAL_RECEIVABLE**: Total amount receivable on the previous credit
- **CNT_DRAWINGS_ATM_CURRENT**: Number of drawings at ATM during this month on the previous credit

- **CNT_DRAWINGS_CURRENT**: Number of drawings during this month on the previous credit
- **CNT_DRAWINGS_OTHER_CURRENT**: Number of other drawings during this month on the previous credit
- **CNT_DRAWINGS_POS_CURRENT**: Number of drawings for goods during this month on the previous credit

- **CNT_INSTALMENT_MATURE_CUM**: Number of paid installments on the previous credit
- **NAME_CONTRACT_STATUS**: Contract status (active signed,...) on the previous credit

- **SK_DPD**: DPD (Days past due) during the month on the previous credit
- **SK_DPD_DEF**: DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit