

# Kaggle Home Credit Default Risk Competition

<https://www.kaggle.com/c/home-credit-default-risk> (<https://www.kaggle.com/c/home-credit-default-risk>)

This competition is sponsored by [Home Credit](http://www.homecredit.net) (<http://www.homecredit.net>), 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 `SK_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.

## 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
```

## Data Description

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

### 2. **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.

### 3. **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.

### 4. **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.

### 5. **POS\_CASH\_balance.csv**

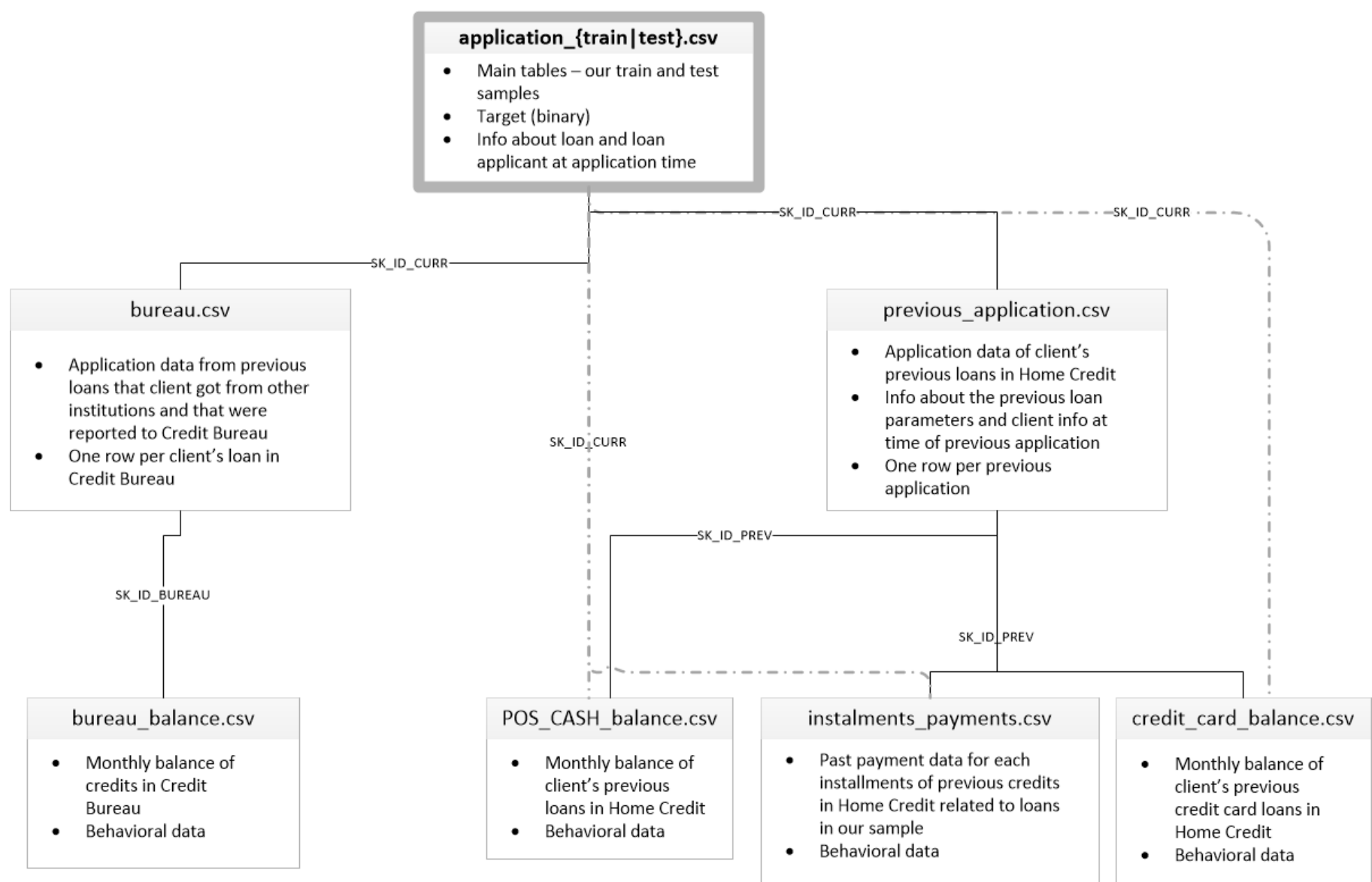
- Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample # of relative previous credits # of months in which we have some history observable for the previous credits) rows.

## 6. **instalments\_payments.csv**

- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
- 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.

## 7. **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. 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_CA
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

5 rows × 122 columns

In [15]:

```
# Total number of entries in training group
print("Total number of entries in training group: {}".format(application_train_data.shape[0]))
```

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
2. **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)
3. **NAME\_CONTRACT\_TYPE**: Identification if loan is cash or revolving
4. **CODE\_GENDER**: Gender of the client
5. **FLAG\_OWN\_CAR**: Flag if the client owns a car
6. **FLAG\_OWN\_REALTY**: Flag if client owns a house or flat
7. **CNT\_CHILDREN**: Number of children the client has
8. **AMT\_INCOME\_TOTAL**: Income of the client
9. **AMT\_CREDIT**: Credit amount of the loan
10. **AMT\_ANNUITY**: Loan annuity
11. **AMT\_GOODS\_PRICE**: For consumer loans it is the price of the goods for which the loan is given
12. **NAME\_TYPE\_SUITE**: Who was accompanying client when he was applying for the loan
13. **NAME\_INCOME\_TYPE**: Clients income type (businessman, working, maternity leave,Ö)
14. **NAME\_EDUCATION\_TYPE**: Level of highest education the client achieved
15. **NAME\_FAMILY\_STATUS**: Family status of the client
16. **NAME\_HOUSING\_TYPE**: What is the housing situation of the client (renting, living with parents, ...)
17. **REGION\_POPULATION\_RELATIVE**: Normalized population of region where client lives (higher number means the client lives in more populated region) -- normalized
18. **DAYS\_BIRTH**: Client's age in days at the time of application -- time only relative to the application
19. **DAYS\_EMPLOYED**: How many days before the application the person started current employment -- time only relative to the application
20. **DAYS\_REGISTRATION**: How many days before the application did client change his registration -- time only relative to the application
21. **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
22. **OWN\_CAR\_AGE**: Age of client's car
23. **FLAG\_MOBIL**: Did client provide mobile phone (1=YES, 0=NO)
24. **FLAG\_EMP\_PHONE**: Did client provide work phone (1=YES, 0=NO)
25. **FLAG\_WORK\_PHONE**: Did client provide home phone (1=YES, 0=NO)
26. **FLAG\_CONT\_MOBILE**: Was mobile phone reachable (1=YES, 0=NO)
27. **FLAG\_PHONE**: Did client provide home phone (1=YES, 0=NO)
28. **FLAG\_EMAIL**: Did client provide email (1=YES, 0=NO)
29. **OCCUPATION\_TYPE**: What kind of occupation does the client have
30. **CNT\_FAM\_MEMBERS**: How many family members does client have
31. **REGION\_RATING\_CLIENT**: Our rating of the region where client lives (1,2,3)
32. **REGION\_RATING\_CLIENT\_W\_CITY**: Our rating of the region where client lives with taking city into account (1,2,3)
33. **WEEKDAY\_APPR\_PROCESS\_START**: On which day of the week did the client apply for the loan
34. **HOURLY\_APPR\_PROCESS\_START**: Approximately at what hour did the client apply for the loan rounded
35. **REG\_REGION\_NOT\_LIVE\_REGION**: Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
36. **REG\_REGION\_NOT\_WORK\_REGION**: Flag if client's permanent address does not match work address (1=different, 0=same, at region level)

37. **LIVE\_REGION\_NOT\_WORK\_REGION:** Flag if client's contact address does not match work address (1=different, 0=same, at region level)
38. **REG\_CITY\_NOT\_LIVE\_CITY:** Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
39. **REG\_CITY\_NOT\_WORK\_CITY:** Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
40. **LIVE\_CITY\_NOT\_WORK\_CITY:** Flag if client's contact address does not match work address (1=different, 0=same, at city level)
41. **ORGANIZATION\_TYPE:** Type of organization where client works
42. **EXT\_SOURCE\_1:** Normalized score from external data source -- normalized
43. **EXT\_SOURCE\_2:** Normalized score from external data source -- normalized
44. **EXT\_SOURCE\_3:** Normalized score from external data source -- normalized
45. **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
46. **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
47. **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
48. **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
49. **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
50. **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
51. **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
52. **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
53. **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
54. **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
55. **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
56. **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
57. **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
58. **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
59. **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
60. **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
61. **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
62. **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
63. **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
64. **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

65. **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
66. **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
67. **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
68. **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
69. **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
70. **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
71. **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
72. **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
73. **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
74. **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
75. **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

76. **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
77. **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
78. **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
79. **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
80. **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
81. **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
82. **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
83. **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
84. **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
85. **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
86. **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

87. **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
88. **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
89. **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
90. **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
91. **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
92. **OBS\_30\_CNT\_SOCIAL\_CIRCLE**: How many observation of client's social surroundings with observable 30 DPD (days past due) default
93. **DEF\_30\_CNT\_SOCIAL\_CIRCLE**: How many observation of client's social surroundings defaulted on 30 DPD (days past due)
94. **OBS\_60\_CNT\_SOCIAL\_CIRCLE**: How many observation of client's social surroundings with observable 60 DPD (days past due) default
95. **DEF\_60\_CNT\_SOCIAL\_CIRCLE**: How many observation of client's social surroundings defaulted on 60 (days past due) DPD
96. **DAYS\_LAST\_PHONE\_CHANGE**: How many days before application did client change phone
97. **FLAG\_DOCUMENT\_2**: Did client provide document 2
98. **FLAG\_DOCUMENT\_3**: Did client provide document 3
99. **FLAG\_DOCUMENT\_4**: Did client provide document 4
100. **FLAG\_DOCUMENT\_5**: Did client provide document 5
101. **FLAG\_DOCUMENT\_6**: Did client provide document 6
102. **FLAG\_DOCUMENT\_7**: Did client provide document 7
103. **FLAG\_DOCUMENT\_8**: Did client provide document 8
104. **FLAG\_DOCUMENT\_9**: Did client provide document 9
105. **FLAG\_DOCUMENT\_10**: Did client provide document 10
106. **FLAG\_DOCUMENT\_11**: Did client provide document 11
107. **FLAG\_DOCUMENT\_12**: Did client provide document 12
108. **FLAG\_DOCUMENT\_13**: Did client provide document 13
109. **FLAG\_DOCUMENT\_14**: Did client provide document 14
110. **FLAG\_DOCUMENT\_15**: Did client provide document 15
111. **FLAG\_DOCUMENT\_16**: Did client provide document 16
112. **FLAG\_DOCUMENT\_17**: Did client provide document 17

- 113. **FLAG\_DOCUMENT\_18**: Did client provide document 18
- 114. **FLAG\_DOCUMENT\_19**: Did client provide document 19
- 115. **FLAG\_DOCUMENT\_20**: Did client provide document 20
- 116. **FLAG\_DOCUMENT\_21**: Did client provide document 21
- 117. **AMT\_REQ\_CREDIT\_BUREAU\_HOUR**: Number of enquiries to Credit Bureau about the client one hour before application
- 118. **AMT\_REQ\_CREDIT\_BUREAU\_DAY**: Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)
- 119. **AMT\_REQ\_CREDIT\_BUREAU\_WEEK**: Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
- 120. **AMT\_REQ\_CREDIT\_BUREAU\_MON**: Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
- 121. **AMT\_REQ\_CREDIT\_BUREAU\_QRT**: Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
- 122. **AMT\_REQ\_CREDIT\_BUREAU\_YEAR**: Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

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_CRED
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

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
- 2. **SK\_BUREAU\_ID**: Recoded ID of previous Credit Bureau credit related to our loan (unique coding for each loan application) -- hashed
- 3. **CREDIT\_ACTIVE**: Status of the Credit Bureau (CB) reported credits
- 4. **CREDIT\_CURRENCY**: Recoded currency of the Credit Bureau credit -- recoded
- 5. **DAYS\_CREDIT**: How many days before current application did client apply for Credit Bureau credit -- time only relative to the application
- 6. **CREDIT\_DAY\_OVERDUE**: Number of days past due on CB credit at the time of application for related loan in our sample
- 7. **DAYS\_CREDIT\_ENDDATE**: Remaining duration of CB credit (in days) at the time of application in Home Credit -- time only relative to the application
- 8. **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
- 9. **AMT\_CREDIT\_MAX\_OVERDUE**: Maximal amount overdue on the Credit Bureau credit so far (at application date of loan in our sample)
- 10. **CNT\_CREDIT\_PROLONG**: How many times was the Credit Bureau credit prolonged
- 11. **AMT\_CREDIT\_SUM**: Current credit amount for the Credit Bureau credit
- 12. **AMT\_CREDIT\_SUM\_DEBT**: Current debt on Credit Bureau credit
- 13. **AMT\_CREDIT\_SUM\_LIMIT**: Current credit limit of credit card reported in Credit Bureau
- 14. **AMT\_CREDIT\_SUM\_OVERDUE**: Current amount overdue on Credit Bureau credit
- 15. **CREDIT\_TYPE**: Type of Credit Bureau credit (Car, cash,...)
- 16. **DAYS\_CREDIT\_UPDATE**: How many days before loan application did last information about the Credit Bureau credit come -- time only relative to the application
- 17. **AMT\_ANNUITY**: Annuity of the Credit Bureau credit

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
- 2. **MONTHS\_BALANCE**: Month of balance relative to application date (-1 means the freshest balance date) -- time only relative to the application
- 3. **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 ])

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

5 rows × 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
- 2. **SK\_ID\_CURR**: ID of loan in our sample -- hashed
- 3. **NAME\_CONTRACT\_TYPE**: Contract product type (Cash loan, consumer loan [POS] ,...) of the previous application
- 4. **AMT\_ANNUITY**: Annuity of previous application
- 5. **AMT\_APPLICATION**: For how much credit did client ask on the previous application
- 6. **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
- 7. **AMT\_DOWN\_PAYMENT**: Down payment on the previous application
- 8. **AMT\_GOODS\_PRICE**: Goods price of good that client asked for (if applicable) on the previous

- application
9. **WEEKDAY\_APPR\_PROCESS\_START**: On which day of the week did the client apply for previous application
  10. **HOURLY\_APPR\_PROCESS\_START**: Approximately at what day hour did the client apply for the previous application -- rounded
  11. **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
  12. **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
  13. **RATE\_DOWN\_PAYMENT**: Down payment rate normalized on previous credit -- normalized
  14. **RATE\_INTEREST\_PRIMARY**: Interest rate normalized on previous credit -- normalized
  15. **RATE\_INTEREST\_PRIVILEGED**: Interest rate normalized on previous credit -- normalized
  16. **NAME\_CASH\_LOAN\_PURPOSE**: Purpose of the cash loan
  17. **NAME\_CONTRACT\_STATUS**: Contract status (approved, cancelled, ...) of previous application
  18. **DAYS\_DECISION**: Relative to current application when was the decision about previous application made time only relative to the application
  19. **NAME\_PAYMENT\_TYPE**: Payment method that client chose to pay for the previous application
  20. **CODE\_REJECT\_REASON**: Why was the previous application rejected
  21. **NAME\_TYPE\_SUITE**: Who accompanied client when applying for the previous application
  22. **NAME\_CLIENT\_TYPE**: Was the client old or new client when applying for the previous application
  23. **NAME\_GOODS\_CATEGORY**: What kind of goods did the client apply for in the previous application
  24. **NAME\_PORTFOLIO**: Was the previous application for CASH, POS, CAR, Ö
  25. **NAME\_PRODUCT\_TYPE**: Was the previous application x-sell o walk-in
  26. **CHANNEL\_TYPE**: Through which channel we acquired the client on the previous application
  27. **SELLERPLACE\_AREA**: Selling area of seller place of the previous application
  28. **NAME\_SELLER\_INDUSTRY**: The industry of the seller
  29. **CNT\_PAYMENT**: Term of previous credit at application of the previous application
  30. **NAME\_YIELD\_GROUP**: Grouped interest rate into small medium and high of the previous application -- grouped
  31. **PRODUCT\_COMBINATION**: Detailed product combination of the previous application
  32. **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
  33. **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
  34. **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
  35. **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
  36. **DAYS\_TERMINATION**: Relative to application date of current application when was the expected termination of the previous application -- time only relative to the application
  37. **NFLAG\_INSURED\_ON\_APPROVAL**: Did the client requested insurance during the previous application

## 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	CNT_INSTALM
0	1803195	182943	-31	48.0	45.0
1	1715348	367990	-33	36.0	35.0
2	1784872	397406	-32	12.0	9.0
3	1903291	269225	-35	48.0	42.0
4	2341044	334279	-35	36.0	35.0

### 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)
- 2. **SK\_ID\_CURR**: ID of loan in our sample
- 3. **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
- 4. **CNT\_INSTALMENT**: Term of previous credit (can change over time)
- 5. **CNT\_INSTALMENT\_FUTURE**: Installments left to pay on the previous credit
- 6. **NAME\_CONTRACT\_STATUS**: Contract status during the month
- 7. **SK\_DPD**: DPD (days past due) during the month of previous credit
- 8. **SK\_DPD\_DEF**: DPD during the month with tolerance (debts with low loan amounts are ignored) of the previous credit

## 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_INSTALLMENT_VERSION	NUM_INSTALLMENT_NUM
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

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
- 2. **SK\_ID\_CURR**: ID of loan in our sample -- hashed
- 3. **NUM\_INSTALLMENT\_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
- 4. **NUM\_INSTALLMENT\_NUMBER**: On which installment we observe payment
- 5. **DAYS\_INSTALLMENT**: When the installment of previous credit was supposed to be paid (relative to application date of current loan) -- time only relative to the application
- 6. **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
- 7. **AMT\_INSTALLMENT**: What was the prescribed installment amount of previous credit on this installment
- 8. **AMT\_PAYMENT**: What the client actually paid on previous credit on this installment

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	AMT_CREDIT_LIM
0	2562384	378907	-6	56.970	135000
1	2582071	363914	-1	63975.555	45000
2	1740877	371185	-7	31815.225	450000
3	1389973	337855	-4	236572.110	225000
4	1891521	126868	-1	453919.455	450000

5 rows × 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
2. **SK\_ID\_CURR**: ID of loan in our sample -- hashed
3. **MONTHS\_BALANCE**: Month of balance relative to application date (-1 means the freshest balance date) -- time only relative to the application
4. **AMT\_BALANCE**: Balance during the month of previous credit
5. **AMT\_CREDIT\_LIMIT\_ACTUAL**: Credit card limit during the month of the previous credit
6. **AMT\_DRAWINGS\_ATM\_CURRENT**: Amount drawing at ATM during the month of the previous credit
7. **AMT\_DRAWINGS\_CURRENT**: Amount drawing during the month of the previous credit
8. **AMT\_DRAWINGS\_OTHER\_CURRENT**: Amount of other drawings during the month of the previous credit
9. **AMT\_DRAWINGS\_POS\_CURRENT**: Amount drawing or buying goods during the month of the previous credit
10. **AMT\_INST\_MIN\_REGULARITY**: Minimal installment for this month of the previous credit
11. **AMT\_PAYMENT\_CURRENT**: How much did the client pay during the month on the previous credit
12. **AMT\_PAYMENT\_TOTAL\_CURRENT**: How much did the client pay during the month in total on the previous credit
13. **AMT\_RECEIVABLE\_PRINCIPAL**: Amount receivable for principal on the previous credit
14. **AMT\_RECIVABLE**: Amount receivable on the previous credit
15. **AMT\_TOTAL\_RECEIVABLE**: Total amount receivable on the previous credit
16. **CNT\_DRAWINGS\_ATM\_CURRENT**: Number of drawings at ATM during this month on the previous credit
17. **CNT\_DRAWINGS\_CURRENT**: Number of drawings during this month on the previous credit
18. **CNT\_DRAWINGS\_OTHER\_CURRENT**: Number of other drawings during this month on the previous credit
19. **CNT\_DRAWINGS\_POS\_CURRENT**: Number of drawings for goods during this month on the previous credit
20. **CNT\_INSTALLMENT\_MATURE\_CUM**: Number of paid installments on the previous credit
21. **NAME\_CONTRACT\_STATUS**: Contract status (active signed,...) on the previous credit
22. **SK\_DPD**: DPD (Days past due) during the month on the previous credit
23. **SK\_DPD\_DEF**: DPD (Days past due) during the month with tolerance (debts with low loan amounts are ignored) of the previous credit