Step by step description of homecredit_wrangling.R

P. A. Ortiz Otalvaro

3 September 2019

"Data Wrangling is the process of converting and mapping data from its raw form to another format with the purpose of making it more valuable and appropriate for advanced tasks such as Data Analytics and Machine Learning."

1. Load packages

```
library(dplyr)
library(tidyr)
library(ggplot2)
```

2. Read data

During the reading process, all blank and empty observations were replaced with NA.

```
homecredit_train <- readRDS("../LoanData_HomeCredit/application_train.RDS")
homecredit_test <- readRDS("../LoanData_HomeCredit/application_test.RDS")
bureau <- readRDS("../LoanData_HomeCredit/bureau.RDS")
bureaubalance <- readRDS("../LoanData_HomeCredit/bureau_balance.RDS")
previousapp <- readRDS("../LoanData_HomeCredit/previous_application.RDS")
poscashbalance <- readRDS("../LoanData_HomeCredit/POS_CASH_balance.RDS")
installments <- readRDS("../LoanData_HomeCredit/installments_payments.RDS")
creditcard <- readRDS("../LoanData_HomeCredit/credit_card_balance.RDS")
```

Clean up column names

In my opinion the columnn names do not need to be modified. They are already simple, short and descriptive.

3. Initial exploration of data sets

General description of data files. There are 7 csv files with information related to Home Credit customer's past financial data. All files are related directly or indirectly to application_{train|test}.csv. The relation between them (and the corresponding keys) are shown in Figure 1.

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

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- One row represents one loan in the data sample.
- For each loan there are 121 features describing the customer as well as the loan. Variables are in the categories: personal information, work status, family details, housing, properties and some information on customer's social circle related to Home Credit. Loan documentation and previous credit history according to Credit Bureau.

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. *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.

5. *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.

6. *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.

7. *installments_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.

8. *HomeCredit_columns_description.csv*

- This file contains descriptions for the columns in the various data files.

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

	Observations	Features	Character	Factor	Numeric	NAs
Train	307511	122	16	0	106	9152465
Test	48744	121	16	0	105	1404419
Bureau	1716428	17	3	0	14	3939947
Bureau balance	27299925	3	1	0	2	0
Previous Applications	1670214	37	16	0	21	11109336
POS cash balance	10001358	8	1	0	7	52158
Installment Payments	13605401	8	0	0	8	5810
Credit card	3840312	23	1	0	22	5877356

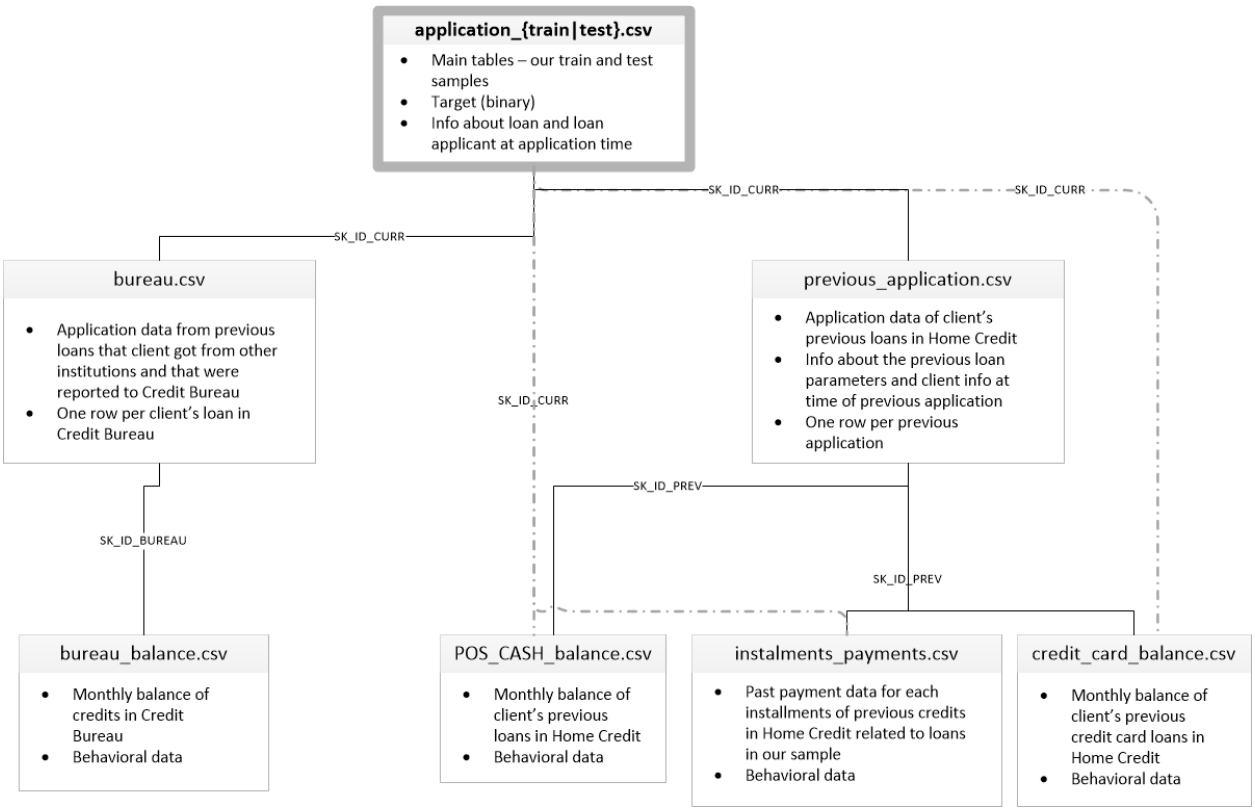

Figure 1: Figure 1: Connections between all data sets

4. Data types

Discrete and continuous

Separating the features into character and numeric does not totally give useful insights. For example: TARGET has only numeric values (1 and 0), however getting its mean or standard deviation does not make any sense and it would be more meaningful to analyze it as a factor and determine the number of appearances of each unique value. Therefore, the following step is to divide the variable set into categorical and non-categorical.

All following tables show the first 5 categorical and first 5 non-categorical columns of each data set. See appendix 1 for the complete tables.

train

Categorical columns:

	Col index	Unique total	Unique values
TARGET	2	2	1, 0
NAME_CONTRACT_TYPE	3	2	Cash loans, Revolving loans
CODE_GENDER	4	3	M, F, XNA
FLAG_OWN_CAR	5	2	N, Y
FLAG_OWN_REALTY	6	2	Y, N
CNT_CHILDREN	7	15	0, 1, 2, 3, 4, 7, 5, 6, 8, 9, 11, 12, 10, 19, 14

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
AMT_INCOME_TOTAL	307511	25650.0	117000000.00	237123.15	168797.92	135000
AMT_CREDIT	307511	45000.0	4050000.00	402490.78	599026.00	450000
AMT_ANNUITY	307499	1615.5	258025.50	14493.74	27108.57	9000
AMT_GOODS_PRICE	307233	40500.0	4050000.00	369446.46	538396.21	450000
REGION_POPULATION_RELATIVE	307511	0.0	0.07	0.01	0.02	0.035792

bureau

Categorical columns:

	Col index	Unique total	Unique values
CREDIT_ACTIVE	3	4	Closed, Active, Sold, Bad debt
CREDIT_CURRENCY	4	4	currency 1, currency 2, currency 4, currency 3
CREDIT_TYPE	15	15	Consumer credit, Credit card, Mortgage, Car loan, Microloan, Loan f

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
SK_ID_BUREAU	1716428	5000000	6843457	532265.73	5924434.49	5000000
DAYS_CREDIT	1716428	-2922	0	795.16	-1142.11	-364
CREDIT_DAY_OVERDUE	1716428	0	2792	36.54	0.82	0
DAYS_CREDIT_ENDDATE	1610875	-42060	31199	4994.22	510.52	0
DAYS_ENDDATE_FACT	1082775	-42023	0	714.01	-1017.44	-329

	Count	Min	Max	St Dev	Mean	Mode

bureau_balance

In *bureau_balance.csv* all columns are categorical except for the first one that contains the id of the applicants.

Categorical columns:

	Col index	Unique total	Unique values
MONTHS_BALANCE	2	97	0, -1, -2, -3, -4, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, -15, -16, -17, -18
STATUS	3	8	C, 0, X, 1, 2, 3, 5, 4

previous applications

Categorical columns:

	Col index	Unique total	Unique values
NAME_CONTRACT_TYPE	3	4	Consumer loans, Cash loans, Revolving loans, XN
WEEKDAY_APPR_PROCESS_START	9	7	SATURDAY, THURSDAY, TUESDAY, MONDA
HOUR_APPR_PROCESS_START	10	24	15, 11, 7, 9, 8, 10, 12, 13, 14, 16, 6, 4, 5, 19, 17, 18
FLAG_LAST_APPL_PER_CONTRACT	11	2	Y, N
NFLAG_LAST_APPL_IN_DAY	12	2	1, 0
NAME_CASH_LOAN_PURPOSE	16	25	XAP, XNA, Repairs, Everyday expenses, Car repa

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
AMT_ANNUITY	1297979	0.0	418058.2	14782.14	15955.12	2250
AMT_APPLICATION	1670214	0.0	6905160.0	292779.76	175233.86	0
AMT_CREDIT	1670213	0.0	6905160.0	318574.62	196114.02	0
AMT_DOWN_PAYMENT	774370	-0.9	3060045.0	20921.50	6697.40	0
AMT_GOODS_PRICE	1284699	0.0	6905160.0	315396.56	227847.28	45000

pos cash balance

Categorical columns:

	Col index	Unique total	Unique values
MONTHS_BALANCE	3	96	-31, -33, -32, -35, -38, -39, -34, -41, -37, -40, -43, -36, -42, -4
CNT_INSTALMENT	4	74	48, 36, 12, 24, 60, 18, 4, 42, 25, 14, 16, 13, 8, 10, 15, 11, 30,
CNT_INSTALMENT_FUTURE	5	80	45, 35, 9, 42, 12, 43, 36, 16, 24, 5, 15, 1, 28, 23, 56, 11, 7, 18
NAME_CONTRACT_STATUS	6	9	Active, Completed, Signed, Approved, Returned to the store

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
SK_DPD	10001358	0	4231	132.71	11.61	0
SK_DPD_DEF	10001358	0	3595	32.76	0.65	0

	Count	Min	Max	St Dev	Mean	Mode
NA	NA	NA	NA	NA	NA	NA
NA.1	NA	NA	NA	NA	NA	NA
NA.2	NA	NA	NA	NA	NA	NA

Installment payments

Categorical columns:

	Col index	Unique total	Unique values
NUM_INSTALMENT_VERSION	3	65	1, 0, 2, 4, 3, 5, 7, 8, 6, 13, 9, 21, 22, 12, 17, 18, 11, 14, 34,
NUM_INSTALMENT_NUMBER	4	277	6, 34, 1, 3, 2, 12, 11, 4, 14, 8, 20, 56, 7, 23, 38, 116, 5, 46,

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
DAYS_INSTALMENT	13605401	-2922	-1	800.95	-1042.27	-120
DAYS_ENTRY_PAYMENT	13602496	-4921	-1	800.59	-1051.11	-91
AMT_INSTALMENT	13605401	0	3771488	50570.25	17050.91	9000
AMT_PAYMENT	13602496	0	3771488	54735.78	17238.22	9000
NA	NA	NA	NA	NA	NA	NA

Credit card balance

Categorical columns:

	Col index	Unique total	Unique values
MONTHS_BALANCE	3	96	-6, -1, -7, -4, -5, -3, -2, -19, -13, -18, -15, -12, -22, -
CNT_DRAWINGS_ATM_CURRENT	16	45	0, 1, 3, 5, 2, NA, 6, 13, 4, 8, 9, 7, 10, 12, 11, 18, 19
CNT_DRAWINGS_CURRENT	17	129	1, 0, 8, 9, 5, 2, 30, 3, 4, 7, 14, 38, 12, 19, 26, 17, 33
CNT_DRAWINGS_OTHER_CURRENT	18	12	0, 1, NA, 3, 2, 4, 6, 5, 7, 8, 10, 12
CNT_DRAWINGS_POS_CURRENT	19	134	1, 0, 5, 6, 29, 4, 8, 2, NA, 14, 38, 12, 3, 17, 26, 7, 3
CNT_INSTALMENT_MATURE_CUM	20	122	35, 69, 30, 10, 101, 2, 6, 51, 3, 38, 0, 27, 59, 11, 42

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
AMT_BALANCE	3840312	-420250.18	1505902	106307.03	58300.16	0
AMT_CREDIT_LIMIT_ACTUAL	3840312	0.00	1350000	165145.70	153807.96	0
AMT_DRAWINGS_ATM_CURRENT	3090496	-6827.31	2115000	28225.69	5961.32	0
AMT_DRAWINGS_CURRENT	3840312	-6211.62	2287098	33846.08	7433.39	0
AMT_DRAWINGS_OTHER_CURRENT	3090496	0.00	1529847	8201.99	288.17	0

Subsetting features according to theme

train

5. NAs

The following tables give an idea of the missing values per column. Here only the top 5 columns with missing values are shown. Appendix 1 presents the complete list.

Missing values per column

IMPORTANT:

This section will be updated later on when I have decided how to replace these missing values in each column or if to remove the observations of with missing values. I need better understanding of the models to make this decision.

6. Anomalies in data set

This is the part of the data wrangling process to search for values in the columns that do not make sense as well as outliers or extreme measurements

COLUMNA 18 EN TRAIN, 17 EN TEST, DAYS_BIRTH: CAMBIARLA A UN TIEMPO M'AS DECENTE

change DAYS_BIRTH to postive value in years

```
homecredit_train <- homecredit_train %>% mutate(DAYS_BIRTH/-365)
```

HACER GRAFICA DE VALORES EN TARGET, USAR EJEMPLO DE DATACAMP https://campus.datacamp.com/courses/data-visualization-with-ggplot2-1/chapter-3-aesthetics?ex=10

Appendix 1: summary of categorical columns

Train

	Col index	Unique total	Unique values
TARGET	2	2	1, 0
NAME_CONTRACT_TYPE	3	2	Cash loans, Revolving loans
CODE_GENDER	4	3	M, F, XNA
FLAG_OWN_CAR	5	2	N, Y
FLAG_OWN_REALTY	6	2	Y, N
CNT_CHILDREN	7	15	0, 1, 2, 3, 4, 7, 5, 6, 8, 9, 11, 12, 10, 19, 14
NAME_TYPE_SUITE	12	8	Unaccompanied, Family, Spouse, partner, Childre
NAME_INCOME_TYPE	13	8	Working, State servant, Commercial associate, Pe
NAME_EDUCATION_TYPE	14	5	Secondary / secondary special, Higher education,
NAME_FAMILY_STATUS	15	6	Single / not married, Married, Civil marriage, Wi
NAME_HOUSING_TYPE	16	6	House / apartment, Rented apartment, With pare
FLAG_MOBIL	23	2	1, 0
FLAG_EMP_PHONE	24	2	1, 0
FLAG_WORK_PHONE	25	2	0, 1
FLAG_CONT_MOBILE	26	2	1, 0
FLAG_PHONE	27	2	1, 0
FLAG_EMAIL	28	2	0, 1
OCCUPATION_TYPE	29	19	Laborers, Core staff, Accountants, Managers, NA,
CNT_FAM_MEMBERS	30	18	1, 2, 3, 4, 5, 6, 9, 7, 8, 10, 13, NA, 14, 12, 20, 15,
REGION_RATING_CLIENT	31	3	2, 1, 3
REGION_RATING_CLIENT_W_CITY	32	3	2, 1, 3
WEEKDAY_APPR_PROCESS_START	33	7	WEDNESDAY, MONDAY, THURSDAY, SUNDA
HOUR_APPR_PROCESS_START	34	24	10, 11, 9, 17, 16, 14, 8, 15, 7, 13, 6, 12, 19, 3, 18, 1
REG_REGION_NOT_LIVE_REGION	35	2	0, 1
REG_REGION_NOT_WORK_REGION	36	2	0, 1
LIVE_REGION_NOT_WORK_REGION	37	2	0, 1
REG_CITY_NOT_LIVE_CITY	38	2	0, 1
REG_CITY_NOT_WORK_CITY	39	2	0, 1
LIVE_CITY_NOT_WORK_CITY	40	2	0, 1
ORGANIZATION_TYPE	41	58	Business Entity Type 3, School, Government, Reli
FONDKAPREMONT_MODE	87	5	reg oper account, NA, org spec account, reg oper
HOUSETYPE_MODE	88	4	block of flats, NA, terraced house, specific housing
WALLSMATERIAL_MODE	90	8	Stone, brick, Block, NA, Panel, Mixed, Wooden, C
EMERGENCYSTATE_MODE	91	3	No, NA, Yes
OBS_30_CNT_SOCIAL_CIRCLE	92	34	2, 1, 0, 4, 8, 10, NA, 7, 3, 6, 5, 12, 9, 13, 11, 14, 2
DEF_30_CNT_SOCIAL_CIRCLE	93	11	2, 0, 1, NA, 3, 4, 5, 6, 7, 34, 8
OBS_60_CNT_SOCIAL_CIRCLE	94	34	2, 1, 0, 4, 8, 10, NA, 7, 3, 6, 5, 12, 9, 13, 11, 14, 2
DEF_60_CNT_SOCIAL_CIRCLE	95	10	2, 0, 1, NA, 3, 5, 4, 7, 24, 6
FLAG_DOCUMENT_2	97	2	0, 1
FLAG_DOCUMENT_3	98	2	1, 0
FLAG_DOCUMENT_4	99	2	0, 1
FLAG_DOCUMENT_5	100	2	0, 1
FLAG_DOCUMENT_6	101	2	0, 1
FLAG_DOCUMENT_7	102	2	0, 1
FLAG_DOCUMENT_8	103	2	0, 1
FLAG_DOCUMENT_9	104	2	0, 1
FLAG_DOCUMENT_10	105	2	0, 1
FLAG_DOCUMENT_11	106	2	0, 1
FLAG_DOCUMENT_12	107	2	0, 1

	Col index	Unique total	Unique values
FLAG_DOCUMENT_13	108	2	0, 1
FLAG_DOCUMENT_14	109	2	0, 1
FLAG_DOCUMENT_15	110	2	0, 1
FLAG_DOCUMENT_16	111	2	0, 1
FLAG_DOCUMENT_17	112	2	0, 1
FLAG_DOCUMENT_18	113	2	0, 1
FLAG_DOCUMENT_19	114	2	0, 1
FLAG_DOCUMENT_20	115	2	0, 1
FLAG_DOCUMENT_21	116	2	0, 1
AMT_REQ_CREDIT_BUREAU_HOUR	117	6	0, NA, 1, 2, 3, 4
AMT_REQ_CREDIT_BUREAU_DAY	118	10	0, NA, 1, 3, 2, 4, 5, 6, 9, 8
AMT_REQ_CREDIT_BUREAU_WEEK	119	10	0, NA, 1, 3, 2, 4, 5, 6, 8, 7
AMT_REQ_CREDIT_BUREAU_MON	120	25	0, NA, 1, 2, 6, 5, 3, 7, 9, 4, 11, 8, 16, 12, 14, 10, 1
AMT_REQ_CREDIT_BUREAU_QRT	121	12	0, NA, 1, 2, 4, 3, 8, 5, 6, 7, 261, 19
AMT_REQ_CREDIT_BUREAU_YEAR	122	26	1, 0, NA, 2, 4, 5, 3, 8, 6, 9, 7, 10, 11, 13, 16, 12, 2

	Count	Min	Max	St Dev	Mean	Mode
AMT_INCOME_TOTAL	307511	25650.00	117000000.00	237123.15	168797.92	135000
AMT_CREDIT	307511	45000.00	4050000.00	402490.78	599026.00	450000
AMT_ANNUITY	307499	1615.50	258025.50	14493.74	27108.57	9000
AMT_GOODS_PRICE	307233	40500.00	4050000.00	369446.46	538396.21	450000
REGION_POPULATION_RELATIVE	307511	0.00	0.07	0.01	0.02	0.035792
DAYS_BIRTH	307511	-25229.00	-7489.00	4363.99	-16037.00	-13749
DAYS_EMPLOYED	307511	-17912.00	365243.00	141275.77	63815.05	365243
DAYS_REGISTRATION	307511	-24672.00	0.00	3522.89	-4986.12	-1
DAYS_ID_PUBLISH	307511	-7197.00	0.00	1509.45	-2994.20	-4053
OWN_CAR_AGE	104582	0.00	91.00	11.94	12.06	7
EXT_SOURCE_1	134133	0.01	0.96	0.21	0.50	0.3563266441
EXT_SOURCE_2	306851	0.00	0.85	0.19	0.51	0.2858978721
EXT_SOURCE_3	246546	0.00	0.90	0.19	0.51	0.7463002130
APARTMENTS_AVG	151450	0.00	1.00	0.11	0.12	0.0825
BASEMENTAREA_AVG	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_AVG	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_AVG	103023	0.00	1.00	0.11	0.75	0.8232
COMMONAREA_AVG	92646	0.00	1.00	0.08	0.04	0
ELEVATORS_AVG	143620	0.00	1.00	0.13	0.08	0
ENTRANCES_AVG	152683	0.00	1.00	0.10	0.15	0.1379
FLOORSMAX_AVG	154491	0.00	1.00	0.14	0.23	0.1667
FLOORSMIN_AVG	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_AVG	124921	0.00	1.00	0.08	0.07	0
LIVINGAPARTMENTS_AVG	97312	0.00	1.00	0.09	0.10	0.0504
LIVINGAREA_AVG	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_AVG	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_AVG	137829	0.00	1.00	0.07	0.03	0
APARTMENTS_MODE	151450	0.00	1.00	0.11	0.11	0.084
BASEMENTAREA_MODE	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_MODE	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_MODE	103023	0.00	1.00	0.11	0.76	0.8301
COMMONAREA_MODE	92646	0.00	1.00	0.07	0.04	0
ELEVATORS_MODE	143620	0.00	1.00	0.13	0.07	0
ENTRANCES_MODE	152683	0.00	1.00	0.10	0.15	0.1379

	Count	Min	Max	St Dev	Mean	Mode
FLOORSMAX_MODE	154491	0.00	1.00	0.14	0.22	0.1667
FLOORSMIN_MODE	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_MODE	124921	0.00	1.00	0.08	0.06	0
LIVINGAPARTMENTS_MODE	97312	0.00	1.00	0.10	0.11	0.0551
LIVINGAREA_MODE	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_MODE	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_MODE	137829	0.00	1.00	0.07	0.03	0
APARTMENTS_MEDI	151450	0.00	1.00	0.11	0.12	0.0833
BASEMENTAREA_MEDI	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_MEDI	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_MEDI	103023	0.00	1.00	0.11	0.76	0.8256
COMMONAREA_MEDI	92646	0.00	1.00	0.08	0.04	0
ELEVATORS_MEDI	143620	0.00	1.00	0.13	0.08	0
ENTRANCES_MEDI	152683	0.00	1.00	0.10	0.15	0.1379
FLOORSMAX_MEDI	154491	0.00	1.00	0.15	0.23	0.1667
FLOORSMIN_MEDI	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_MEDI	124921	0.00	1.00	0.08	0.07	0
LIVINGAPARTMENTS_MEDI	97312	0.00	1.00	0.09	0.10	0.0513
LIVINGAREA_MEDI	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_MEDI	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_MEDI	137829	0.00	1.00	0.07	0.03	0
TOTALAREA_MODE	159080	0.00	1.00	0.11	0.10	0
DAYS_LAST_PHONE_CHANGE	307510	-4292.00	0.00	826.81	-962.86	0

bureau

	Col index	Unique total	Unique values
CREDIT_ACTIVE	3	4	Closed, Active, Sold, Bad debt
CREDIT_CURRENCY	4	4	currency 1, currency 2, currency 4, currency 3
CREDIT_TYPE	15	15	Consumer credit, Credit card, Mortgage, Car loan, Microloan, Loan f

	Count	Min	Max	St Dev	Mean	Mode
SK_ID_BUREAU	1716428	5000000.0	6843457	532265.73	5924434.49	5000000
DAYS_CREDIT	1716428	-2922.0	0	795.16	-1142.11	-364
CREDIT_DAY_OVERDUE	1716428	0.0	2792	36.54	0.82	0
DAYS_CREDIT_ENDDATE	1610875	-42060.0	31199	4994.22	510.52	0
DAYS_ENDDATE_FACT	1082775	-42023.0	0	714.01	-1017.44	-329
AMT_CREDIT_MAX_OVERDUE	591940	0.0	115987185	206031.61	3825.42	0
CNT_CREDIT_PROLONG	1716428	0.0	9	0.10	0.01	0
AMT_CREDIT_SUM	1716415	0.0	585000000	1149811.34	354994.59	0
AMT_CREDIT_SUM_DEBT	1458759	-4705600.3	170100000	677401.13	137085.12	0
AMT_CREDIT_SUM_LIMIT	1124648	-586406.1	4705600	45032.03	6229.51	0
AMT_CREDIT_SUM_OVERDUE	1716428	0.0	3756681	5937.65	37.91	0
DAYS_CREDIT_UPDATE	1716428	-41947.0	372	720.75	-593.75	-7
AMT_ANNUITY	489637	0.0	118453424	325826.95	15712.76	0

Bureau balance

	Col index	Unique total	Unique values
MONTHS_BALANCE	2	97	0, -1, -2, -3, -4, -5, -6, -7, -8, -9, -10, -11, -12, -13, -14, -15, -16, -17, -18
STATUS	3	8	C, 0, X, 1, 2, 3, 5, 4

Previous applications

	Col index	Unique total	Unique values
NAME_CONTRACT_TYPE	3	4	Consumer loans, Cash loans, Revolving loans, XN
WEEKDAY_APPR_PROCESS_START	9	7	SATURDAY, THURSDAY, TUESDAY, MONDA
HOUR_APPR_PROCESS_START	10	24	15, 11, 7, 9, 8, 10, 12, 13, 14, 16, 6, 4, 5, 19, 17, 18
FLAG_LAST_APPL_PER_CONTRACT	11	2	Y, N
NFLAG_LAST_APPL_IN_DAY	12	2	1, 0
NAME_CASH_LOAN_PURPOSE	16	25	XAP, XNA, Repairs, Everyday expenses, Car repa
NAME_CONTRACT_STATUS	17	4	Approved, Refused, Canceled, Unused offer
NAME_PAYMENT_TYPE	19	4	Cash through the bank, XNA, Non-cash from you
CODE_REJECT_REASON	20	9	XAP, HC, LIMIT, CLIENT, SCOFR, SCO, XNA
NAME_TYPE_SUITE	21	8	NA, Unaccompanied, Spouse, partner, Family, Ch
NAME_CLIENT_TYPE	22	4	Repeater, New, Refreshed, XNA
NAME_GOODS_CATEGORY	23	28	Mobile, XNA, Consumer Electronics, Constructio
NAME_PORTFOLIO	24	5	POS, Cash, XNA, Cards, Cars
NAME_PRODUCT_TYPE	25	3	XNA, x-sell, walk-in
CHANNEL_TYPE	26	8	Country-wide, Contact center, Credit and cash of
NAME_SELLER_INDUSTRY	28	11	Connectivity, XNA, Consumer electronics, Industr
CNT_PAYMENT	29	50	12, 36, 24, 18, NA, 54, 30, 8, 3, 6, 0, 48, 10, 60, 42
NAME_YIELD_GROUP	30	5	middle, low_action, high, low_normal, XNA
PRODUCT_COMBINATION	31	18	POS mobile with interest, Cash X-Sell: low, Cash
NFLAG_INSURED_ON_APPROVAL	37	3	0, 1, NA

	Count	Min	Max	St Dev	Mean	Mode
AMT_ANNUITY	1297979	0.00	418058.2	14782.14	15955.12	2250
AMT_APPLICATION	1670214	0.00	6905160.0	292779.76	175233.86	0
AMT_CREDIT	1670213	0.00	6905160.0	318574.62	196114.02	0
AMT_DOWN_PAYMENT	774370	-0.90	3060045.0	20921.50	6697.40	0
AMT_GOODS_PRICE	1284699	0.00	6905160.0	315396.56	227847.28	45000
RATE_DOWN_PAYMENT	774370	0.00	1.0	0.11	0.08	0
RATE_INTEREST_PRIMARY	5951	0.03	1.0	0.09	0.19	0.189136348180891
RATE_INTEREST_PRIVILEGED	5951	0.37	1.0	0.10	0.77	0.835095137420719
DAYS_DECISION	1670214	-2922.00	-1.0	779.10	-880.68	-245
SELLERPLACE_AREA	1670214	-1.00	4000000.0	7127.44	313.95	-1
DAYS_FIRST_DRAWING	997149	-2922.00	365243.0	88916.12	342209.86	365243
DAYS_FIRST_DUE	997149	-2892.00	365243.0	72444.87	13826.27	365243
DAYS_LAST_DUE_1ST_VERSION	997149	-2801.00	365243.0	106857.03	33767.77	365243
DAYS_LAST_DUE	997149	-2889.00	365243.0	149647.42	76582.40	365243
DAYS_TERMINATION	997149	-2874.00	365243.0	153303.52	81992.34	365243

POS cash balance

	Col index	Unique total	Unique values
MONTHS_BALANCE	3	96	-31, -33, -32, -35, -38, -39, -34, -41, -37, -40, -43, -36, -42, -4

	Col index	Unique total	Unique values
CNT_INSTALMENT	4	74	48, 36, 12, 24, 60, 18, 4, 42, 25, 14, 16, 13, 8, 10, 15, 11, 30,
CNT_INSTALMENT_FUTURE	5	80	45, 35, 9, 42, 12, 43, 36, 16, 24, 5, 15, 1, 28, 23, 56, 11, 7, 18
NAME_CONTRACT_STATUS	6	9	Active, Completed, Signed, Approved, Returned to the store

	Count	Min	Max	St Dev	Mean	Mode
SK_DPD	10001358	0	4231	132.71	11.61	0
SK_DPD_DEF	10001358	0	3595	32.76	0.65	0

Installment payments

	Col index	Unique total	Unique values
NUM_INSTALMENT_VERSION	3	65	1, 0, 2, 4, 3, 5, 7, 8, 6, 13, 9, 21, 22, 12, 17, 18, 11, 14, 34,
NUM_INSTALMENT_NUMBER	4	277	6, 34, 1, 3, 2, 12, 11, 4, 14, 8, 20, 56, 7, 23, 38, 116, 5, 46,

	Count	Min	Max	St Dev	Mean	Mode
DAYS_INSTALMENT	13605401	-2922	-1	800.95	-1042.27	-120
DAYS_ENTRY_PAYMENT	13602496	-4921	-1	800.59	-1051.11	-91
AMT_INSTALMENT	13605401	0	3771488	50570.25	17050.91	9000
AMT_PAYMENT	13602496	0	3771488	54735.78	17238.22	9000

Credit card balance

	Col index	Unique total	Unique values
MONTHS_BALANCE	3	96	-6, -1, -7, -4, -5, -3, -2, -19, -13, -18, -15, -12, -22, -
CNT_DRAWINGS_ATM_CURRENT	16	45	0, 1, 3, 5, 2, NA, 6, 13, 4, 8, 9, 7, 10, 12, 11, 18, 19
CNT_DRAWINGS_CURRENT	17	129	1, 0, 8, 9, 5, 2, 30, 3, 4, 7, 14, 38, 12, 19, 26, 17, 33
CNT_DRAWINGS_OTHER_CURRENT	18	12	0, 1, NA, 3, 2, 4, 6, 5, 7, 8, 10, 12
CNT_DRAWINGS_POS_CURRENT	19	134	1, 0, 5, 6, 29, 4, 8, 2, NA, 14, 38, 12, 3, 17, 26, 7, 3
CNT_INSTALMENT_MATURE_CUM	20	122	35, 69, 30, 10, 101, 2, 6, 51, 3, 38, 0, 27, 59, 11, 42
NAME_CONTRACT_STATUS	21	7	Active, Completed, Demand, Signed, Sent proposa

	Count	Min	Max	St Dev	Mean	Mode
AMT_BALANCE	3840312	-420250.18	1505902	106307.03	58300.16	0
AMT_CREDIT_LIMIT_ACTUAL	3840312	0.00	1350000	165145.70	153807.96	0
AMT_DRAWINGS_ATM_CURRENT	3090496	-6827.31	2115000	28225.69	5961.32	0
AMT_DRAWINGS_CURRENT	3840312	-6211.62	2287098	33846.08	7433.39	0
AMT_DRAWINGS_OTHER_CURRENT	3090496	0.00	1529847	8201.99	288.17	0
AMT_DRAWINGS_POS_CURRENT	3090496	0.00	2239274	20796.89	2968.80	0
AMT_INST_MIN_REGULARITY	3535076	0.00	202882	5600.15	3540.20	0
AMT_PAYMENT_CURRENT	3072324	0.00	4289207	36078.08	10280.54	0
AMT_PAYMENT_TOTAL_CURRENT	3840312	0.00	4278316	32005.99	7588.86	0
AMT_RECEIVABLE_PRINCIPAL	3840312	-423305.82	1472317	102533.62	55965.88	0
AMT_RECIVABLE	3840312	-420250.18	1493338	105965.37	58088.81	0
AMT_TOTAL_RECEIVABLE	3840312	-420250.18	1493338	105971.80	58098.29	0

	Count	Min	Max	St Dev	Mean	Mode
SK_DPD	3840312	0.00	3260	97.52	9.28	0
SK_DPD_DEF	3840312	0.00	3260	21.48	0.33	0