Exploration of Home Credit application dataset

Author

November 22, 2018

1 Explore character variables

Practically the empty fields and the XNA values are the missing ones so we convert these values to missing.

```
        sk_id_curr target name_contract_type code_gender flag_own_car flag_own_realty cnt_
     1:     100002      1         Cash loans           M            N               Y
     2:     100003      0         Cash loans           F            N               N
     3:     100004      0    Revolving loans           M            Y               Y
     4:     100006      0         Cash loans           F            N               Y
     5:     100007      0         Cash loans           M            N               Y
    ---
307507:     456251      0         Cash loans           M            N               N
307508:     456252      0         Cash loans           F            N               Y
        amt_income_total amt_credit amt_annuity amt_goods_price name_type_suite     name_i
     1:           202500   406597.5     24700.5          351000   Unaccompanied
     2:           270000  1293502.5     35698.5         1129500          Family       Sta
     3:            67500   135000.0      6750.0          135000   Unaccompanied
     4:           135000   312682.5     29686.5          297000   Unaccompanied
     5:           121500   513000.0     21865.5          513000   Unaccompanied
    ---
307507:           157500   254700.0     27558.0          225000   Unaccompanied
307508:            72000   269550.0     12001.5          225000   Unaccompanied
                  name_education_type   name_family_status name_housing_type region_popula
     1: Secondary / secondary special Single / not married House / apartment
     2:             Higher education              Married House / apartment
     3: Secondary / secondary special Single / not married House / apartment
     4: Secondary / secondary special       Civil marriage House / apartment
     5: Secondary / secondary special Single / not married House / apartment
    ---
307507: Secondary / secondary special            Separated      With parents
307508: Secondary / secondary special                Widow House / apartment
        days_birth days_employed days_registration days_id_publish own_car_age flag_mobil
     1:      -9461          -637             -3648           -2120          NA          1
     2:     -16765         -1188             -1186            -291          NA          1
     3:     -19046          -225             -4260           -2531          26          1
     4:     -19005         -3039             -9833           -2437          NA          1
     5:     -19932         -3038             -4311           -3458          NA          1
    ---
```

```
307507:      -9327          -236             -8456           -1982          NA          1
307508:     -20775        365243             -4388           -4090          NA          1
       flag_work_phone flag_cont_mobile flag_phone flag_email occupation_type cnt_fam_mem
     1:               0                1          1          0        Laborers
     2:               0                1          1          0     Core staff
     3:               1                1          1          0        Laborers
     4:               0                1          0          0        Laborers
     5:               0                1          0          0     Core staff
    ---
307507:               0                1          0          0    Sales staff
307508:               0                1          1          0           <NA>
       region_rating_client region_rating_client_w_city weekday_appr_process_start hour_a
     1:                    2                          2                  WEDNESDAY
     2:                    1                          1                     MONDAY
     3:                    2                          2                     MONDAY
     4:                    2                          2                  WEDNESDAY
     5:                    2                          2                   THURSDAY
    ---
307507:                    1                          1                   THURSDAY
307508:                    2                          2                     MONDAY
       reg_region_not_live_region reg_region_not_work_region live_region_not_work_region
     1:                          0                          0                           0
     2:                          0                          0                           0
     3:                          0                          0                           0
     4:                          0                          0                           0
     5:                          0                          0                           0
    ---
307507:                          0                          0                           0
307508:                          0                          0                           0
       reg_city_not_live_city reg_city_not_work_city live_city_not_work_city      organiz
     1:                      0                      0                       0 Business Ent
     2:                      0                      0                       0
     3:                      0                      0                       0
     4:                      0                      0                       0 Business Ent
     5:                      0                      1                       1
    ---
307507:                      0                      0                       0
307508:                      0                      0                       0
       ext_source_1 ext_source_2 ext_source_3 apartments_avg basementarea_avg years_begin
     1:   0.08303697    0.2629486    0.1393758         0.0247           0.0369
     2:   0.31126731    0.6222458           NA         0.0959           0.0529
     3:           NA    0.5559121    0.7295667             NA               NA
     4:           NA    0.6504417           NA             NA               NA
     5:           NA    0.3227383           NA             NA               NA
    ---
307507:   0.14557045    0.6816324           NA         0.2021           0.0887
307508:           NA    0.1159921           NA         0.0247           0.0435
       years_build_avg commonarea_avg elevators_avg entrances_avg floorsmax_avg floorsmin
     1:          0.6192         0.0143          0.00        0.0690        0.0833        0.
     2:          0.7960         0.0605          0.08        0.0345        0.2917        0.
     3:              NA             NA            NA            NA            NA
     4:              NA             NA            NA            NA            NA
     5:              NA             NA            NA            NA            NA
    ---
307507:          0.8300         0.0202          0.22        0.1034        0.6042        0.
307508:          0.6260         0.0022          0.00        0.1034        0.0833        0.
        livingapartments_avg livingarea_avg nonlivingapartments_avg nonlivingarea_avg apar
     1:               0.0202         0.0190                  0.0000            0.0000
     2:               0.0773         0.0549                  0.0039            0.0098
     3:                   NA             NA                      NA                NA
     4:                   NA             NA                      NA                NA
     5:                   NA             NA                      NA                NA
    ---
307507:               0.1484         0.1965                  0.0753            0.1095
307508:               0.0202         0.0257                  0.0000            0.0000
        basementarea_mode years_beginexpluatation_mode years_build_mode commonarea_mode el
     1:            0.0383                       0.9722           0.6341          0.0144
     2:            0.0538                       0.9851           0.8040          0.0497
     3:                NA                           NA               NA              NA
     4:                NA                           NA               NA              NA
     5:                NA                           NA               NA              NA
    ---
307507:            0.0172                       0.9782           0.7125          0.0172
307508:            0.0451                       0.9727           0.6406          0.0022
        entrances_mode floorsmax_mode floorsmin_mode landarea_mode livingapartments_mode l
     1:         0.0690         0.0833         0.1250        0.0377                0.0220
     2:         0.0345         0.2917         0.3333        0.0128                0.0790
     3:             NA             NA             NA            NA                    NA
     4:             NA             NA             NA            NA                    NA
     5:             NA             NA             NA            NA                    NA
    ---
307507:         0.0345         0.4583         0.0417        0.0094                0.0882
307508:         0.1034         0.0833         0.1250        0.0592                0.0220
        nonlivingapartments_mode nonlivingarea_mode apartments_medi basementarea_medi
     1:                        0             0.0000          0.0250            0.0369
     2:                        0             0.0000          0.0968            0.0529
     3:                       NA                 NA              NA                NA
     4:                       NA                 NA              NA                NA
     5:                       NA                 NA              NA                NA
    ---
307507:                        0             0.0125          0.2040            0.0887
307508:                        0             0.0000          0.0250            0.0435
        years_beginexpluatation_medi years_build_medi commonarea_medi elevators_medi entra
     1:                       0.9722           0.6243          0.0144           0.00
     2:                       0.9851           0.7987          0.0608           0.08
     3:                           NA               NA              NA             NA
     4:                           NA               NA              NA             NA
     5:                           NA               NA              NA             NA
    ---
307507:                       0.9876           0.8323          0.0203           0.22
```

```
307508:                       0.9727           0.6310          0.0022           0.00
        floorsmax_medi floorsmin_medi landarea_medi livingapartments_medi livingarea_medi
     1:         0.0833         0.1250        0.0375                0.0205          0.0193
     2:         0.2917         0.3333        0.0132                0.0787          0.0558
     3:             NA             NA            NA                    NA              NA
     4:             NA             NA            NA                    NA              NA
     5:             NA             NA            NA                    NA              NA
    ---
307507:         0.6042         0.2708        0.0605                0.1509          0.2001
307508:         0.0833         0.1250        0.0589                0.0205          0.0261
        nonlivingapartments_medi nonlivingarea_medi fondkapremont_mode housetype_mode tota
     1:                   0.0000             0.0000   reg oper account block of flats
     2:                   0.0039             0.0100   reg oper account block of flats
     3:                       NA                 NA               <NA>           <NA>
     4:                       NA                 NA               <NA>           <NA>
     5:                       NA                 NA               <NA>           <NA>
    ---
307507:                   0.0757             0.1118   reg oper account block of flats
307508:                   0.0000             0.0000   reg oper account block of flats
        wallsmaterial_mode emergencystate_mode obs_30_cnt_social_circle def_30_cnt_social_
     1:       Stone, brick                  No                        2
     2:              Block                  No                        1
     3:               <NA>                <NA>                        0
     4:               <NA>                <NA>                        2
     5:               <NA>                <NA>                        0
    ---
307507:       Stone, brick                  No                        0
307508:       Stone, brick                  No                        0
        obs_60_cnt_social_circle def_60_cnt_social_circle days_last_phone_change flag_docu
     1:                        2                        2                  -1134
     2:                        1                        0                   -828
     3:                        0                        0                   -815
     4:                        2                        0                   -617
     5:                        0                        0                  -1106
    ---
307507:                        0                        0                   -273
307508:                        0                        0                      0
        flag_document_3 flag_document_4 flag_document_5 flag_document_6 flag_document_7 fl
     1:               1               0               0               0               0
     2:               1               0               0               0               0
     3:               0               0               0               0               0
     4:               1               0               0               0               0
     5:               0               0               0               0               0
    ---
307507:               0               0               0               0               0
307508:               1               0               0               0               0
        flag_document_9 flag_document_10 flag_document_11 flag_document_12 flag_document_1
     1:               0                0                0                0
     2:               0                0                0                0
     3:               0                0                0                0
     4:               0                0                0                0
     5:               0                0                0                0
    ---
307507:               0                0                0                0
307508:               0                0                0                0
        flag_document_15 flag_document_16 flag_document_17 flag_document_18 flag_document_
     1:                0                0                0                0
     2:                0                0                0                0
     3:                0                0                0                0
     4:                0                0                0                0
     5:                0                0                0                0
    ---
307507:                0                0                0                0
307508:                0                0                0                0
        flag_document_20 flag_document_21 amt_req_credit_bureau_hour amt_req_credit_bureau
     1:                0                0                          0
     2:                0                0                          0
     3:                0                0                          0
     4:                0                0                         NA
     5:                0                0                          0
    ---
307507:                0                0                         NA
307508:                0                0                         NA
        amt_req_credit_bureau_week amt_req_credit_bureau_mon amt_req_credit_bureau_qrt
     1:                          0                         0                         0
     2:                          0                         0                         0
     3:                          0                         0                         0
     4:                         NA                        NA                        NA
     5:                          0                         0                         0
    ---
307507:                         NA                        NA                        NA
307508:                         NA                        NA                        NA
        amt_req_credit_bureau_year
     1:                          1
     2:                          0
     3:                          0
     4:                         NA
     5:                          0
    ---
307507:                         NA
307508:                         NA
 [ reached getOption("max.print") -- omitted 3 rows ]

Skim summary statistics
 n obs: 307511
 n variables: 16

âŤĂâŤĂ Variable type:character âŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤ
                   variable missing complete      n min max empty n_unique
                code_gender       4   307507 307511   1   1     0        2
        emergencystate_mode  145755    161756 307511   2   3     0        2
               flag_own_car       0    307511 307511   1   1     0        2
            flag_own_realty       0    307511 307511   1   1     0        2
         fondkapremont_mode  210295     97216 307511  13  21     0        4
             housetype_mode  154297    153214 307511  14  16     0        3
         name_contract_type       0    307511 307511  10  15     0        2
        name_education_type       0    307511 307511  15  29     0        5
         name_family_status       0    307511 307511   5  20     0        6
          name_housing_type       0    307511 307511  12  19     0        6
           name_income_type       0    307511 307511   7  20     0        8
            name_type_suite    1292    306219 307511   6  15     0        7
            occupation_type   96391    211120 307511   7  21     0       18
          organization_type   55374    252137 307511   4  22     0       57
         wallsmaterial_mode  156341    151170 307511   5  12     0        7
 weekday_appr_process_start       0    307511 307511   6   9     0        7
```

2 Explore flags

We recode the flag_own_car and flag_own_reality from Yes:No to 1:0.

```
Skim summary statistics
 n obs: 307511
 n variables: 34

âŤĂâŤĂ Variable type:integer âŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâŤĂâ
                    variable missing complete      n        mean     sd p0 p25 p50 p75 p10
            flag_cont_mobile       0   307511 307511     1       0.043   0   1   1   1
            flag_document_10       0   307511 307511     2.3e-05 0.0048  0   0   0   0
            flag_document_11       0   307511 307511     0.0039  0.062   0   0   0   0
            flag_document_12       0   307511 307511     6.5e-06 0.0026  0   0   0   0
            flag_document_13       0   307511 307511     0.0035  0.059   0   0   0   0
            flag_document_14       0   307511 307511     0.0029  0.054   0   0   0   0
            flag_document_15       0   307511 307511     0.0012  0.035   0   0   0   0
            flag_document_16       0   307511 307511     0.0099  0.099   0   0   0   0
            flag_document_17       0   307511 307511     0.00027 0.016   0   0   0   0
            flag_document_18       0   307511 307511     0.0081  0.09    0   0   0   0
            flag_document_19       0   307511 307511 6e-04       0.024   0   0   0   0
             flag_document_2       0   307511 307511     4.2e-05 0.0065  0   0   0   0
            flag_document_20       0   307511 307511     0.00051 0.023   0   0   0   0
            flag_document_21       0   307511 307511     0.00033 0.018   0   0   0   0
             flag_document_3       0   307511 307511     0.71    0.45    0   0   1   1
             flag_document_4       0   307511 307511     8.1e-05 0.009   0   0   0   0
             flag_document_5       0   307511 307511     0.015   0.12    0   0   0   0
             flag_document_6       0   307511 307511     0.088   0.28    0   0   0   0
             flag_document_7       0   307511 307511     0.00019 0.014   0   0   0   0
             flag_document_8       0   307511 307511     0.081   0.27    0   0   0   0
             flag_document_9       0   307511 307511     0.0039  0.062   0   0   0   0
                  flag_email       0   307511 307511     0.057   0.23    0   0   0   0
              flag_emp_phone       0   307511 307511     0.82    0.38    0   1   1   1
                  flag_mobil       0   307511 307511     1       0.0018  0   1   1   1
                flag_own_car       0   307511 307511     0.34     0.47    0   0   0   1
             flag_own_realty       0   307511 307511     0.69     0.46    0   0   1   1
                  flag_phone       0   307511 307511     0.28     0.45    0   0   0   1
             flag_work_phone       0   307511 307511     0.2      0.4     0   0   0   0
     live_city_not_work_city       0   307511 307511     0.18     0.38    0   0   0   0
 live_region_not_work_region      0   307511 307511     0.041    0.2     0   0   0   0
      reg_city_not_live_city       0   307511 307511     0.078    0.27    0   0   0   0
      reg_city_not_work_city       0   307511 307511     0.23     0.42    0   0   0   0
  reg_region_not_live_region      0   307511 307511     0.015    0.12    0   0   0   0
  reg_region_not_work_region      0   307511 307511     0.051    0.22    0   0   0   0
```

3 Explore numeric variables

We drop all those numeric input factors where the missing ratio is above 30%.

```
> # Quickly explore numeric variables ----
> # Our general rule is that if the missing ratio is above 30%
> # then we drop the variable
> skim_dev_num <- skim_dev[type == 'numeric' & stat %in% c('missing','n'), .(variable, sta
> skim_dev_num <- dcast(skim_dev_num, variable ~ stat)
> skim_dev_num[, rat_missing := missing/n]
```

```
                     variable missing      n  rat_missing
 1:               amt_annuity      12 307511 3.902299e-05
 2:                amt_credit       0 307511 0.000000e+00
 3:           amt_goods_price     278 307511 9.040327e-04
 4:          amt_income_total       0 307511 0.000000e+00
 5:   amt_req_credit_bureau_day   41519 307511 1.350163e-01
 6:  amt_req_credit_bureau_hour   41519 307511 1.350163e-01
 7:   amt_req_credit_bureau_mon   41519 307511 1.350163e-01
 8:   amt_req_credit_bureau_qrt   41519 307511 1.350163e-01
 9:  amt_req_credit_bureau_week   41519 307511 1.350163e-01
10:  amt_req_credit_bureau_year   41519 307511 1.350163e-01
11:            apartments_avg  156061 307511 5.074973e-01
12:           apartments_medi  156061 307511 5.074973e-01
13:           apartments_mode  156061 307511 5.074973e-01
14:          basementarea_avg  179943 307511 5.851596e-01
15:         basementarea_medi  179943 307511 5.851596e-01
16:         basementarea_mode  179943 307511 5.851596e-01
17:           cnt_fam_members       2 307511 6.503832e-06
18:           commonarea_avg  214865 307511 6.987230e-01
19:          commonarea_medi  214865 307511 6.987230e-01
20:          commonarea_mode  214865 307511 6.987230e-01
21:     days_last_phone_change       1 307511 3.251916e-06
22:         days_registration       0 307511 0.000000e+00
23:    def_30_cnt_social_circle    1021 307511 3.320206e-03
24:    def_60_cnt_social_circle    1021 307511 3.320206e-03
25:             elevators_avg  163891 307511 5.329598e-01
26:            elevators_medi  163891 307511 5.329598e-01
27:            elevators_mode  163891 307511 5.329598e-01
```

```
28:                entrances_avg  154828 307511 5.034877e-01
29:               entrances_medi  154828 307511 5.034877e-01
30:               entrances_mode  154828 307511 5.034877e-01
31:                 ext_source_1  173378 307511 5.638107e-01
32:                 ext_source_2     660 307511 2.146265e-03
33:                 ext_source_3   60965 307511 1.982531e-01
34:                floorsmax_avg  153020 307511 4.976082e-01
35:               floorsmax_medi  153020 307511 4.976082e-01
36:               floorsmax_mode  153020 307511 4.976082e-01
37:                floorsmin_avg  208642 307511 6.784863e-01
38:               floorsmin_medi  208642 307511 6.784863e-01
39:               floorsmin_mode  208642 307511 6.784863e-01
40:                 landarea_avg  182590 307511 5.937674e-01
41:                landarea_medi  182590 307511 5.937674e-01
42:                landarea_mode  182590 307511 5.937674e-01
43:         livingapartments_avg  210199 307511 6.835495e-01
44:        livingapartments_medi  210199 307511 6.835495e-01
45:        livingapartments_mode  210199 307511 6.835495e-01
46:               livingarea_avg  154350 307511 5.019333e-01
47:              livingarea_medi  154350 307511 5.019333e-01
48:              livingarea_mode  154350 307511 5.019333e-01
49:      nonlivingapartments_avg  213514 307511 6.943296e-01
50:     nonlivingapartments_medi  213514 307511 6.943296e-01
51:     nonlivingapartments_mode  213514 307511 6.943296e-01
52:            nonlivingarea_avg  169682 307511 5.517916e-01
53:           nonlivingarea_medi  169682 307511 5.517916e-01
54:           nonlivingarea_mode  169682 307511 5.517916e-01
55:     obs_30_cnt_social_circle    1021 307511 3.320206e-03
56:     obs_60_cnt_social_circle    1021 307511 3.320206e-03
57:                  own_car_age  202929 307511 6.599081e-01
58:   region_population_relative       0 307511 0.000000e+00
59:               totalarea_mode  148431 307511 4.826852e-01
60:  years_beginexpluatation_avg  150007 307511 4.878102e-01
61: years_beginexpluatation_medi  150007 307511 4.878102e-01
62: years_beginexpluatation_mode  150007 307511 4.878102e-01
63:              years_build_avg  204488 307511 6.649778e-01
64:             years_build_medi  204488 307511 6.649778e-01
65:             years_build_mode  204488 307511 6.649778e-01
                        variable missing      n  rat_missing

>
```