homecredict_exploratory_analysis

August 30, 2023

1 The analysis of the Home Credit Group dataset (I part)

1.1 Introduction

The project uses the dataset of the Home Credit Group (https://www.homecredit.net/). The Home Credit is an international consumer finance provider operates in 9 countries and focuses on installment lending primarily to people with little or no credit history.

The dataset which is downloaded from the Kaggle repository (https://www.kaggle.com/competitions/home-credit-default-risk/data) includes these files:

1) two files with data for all loan applications, broken into two files for Train (with TARGET) and Test (without TARGET); one row represents one loan in our data sample (application_train.csv, application_test.csv);

2) a file containing data on 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 the sample, there are as many rows as number of credits the client had in Credit Bureau before the application date (bureau.csv);

3) a file with data on monthly balances of previous credits in Credit Bureau; the 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 (bureau_balance.csv);

4) a file with data on monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit; the 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 the sample – i.e. the table has (#loans in sample * # of relative previous credits * # of months) rows (POS_CASH_balance.csv);

5) a file containing data on 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 the sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months) rows (credit_card_balance.csv);

6) a file with data on 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 (previous_application.csv);

7) a file containing data on the repayment history for the previously disbursed credits in Home

Credit related to the loans in the 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 (installments_payments.csv).

The relationships between datasets can be seen in this image:

The purposes of the project: - to iteratively build and implement a plan for a large dataset based on business objectives; - to create a number of different models in order to develop a robust and diversified offering of a product of the risk evaluation as a service for retail banks.

Requirements: - Create a plan for your investigation, analysis, and POC building. This should include your assumptions, overall objectives, and objectives for each step in your plan. You are not expected to have a plan for the whole project but instead have a clear understanding of what you'll try to achieve in the next step and build the plan one step at a time. - Perform exploratory data analysis. This should include creating statistical summaries and charts, testing for anomalies, checking for correlations and other relations between variables, and other EDA elements. - Perform statistical inference. This should include defining the target population, forming multiple statistical hypotheses and constructing confidence intervals, setting the significance levels, conducting z or t-tests for these hypotheses. - Use machine learning models to predict the target variables based on your proposed plan. You should use hyperparameter tuning, model ensembling, the analysis of model selection, and other methods. The decision of where to use and not to use these techniques is up to you; however, they should be aligned with your team's objectives. - Deploy these machine learning models to Google Cloud Platform. You are free to choose any deployment option you wish as long as it can be called an HTTP request.

Objectives:

- Practice translating business requirements into data science tasks.
- Practice performing EDA.
- Practice applying statistical inference procedures.
- Practice using machine learning to solve business problems.
- Practice deploying multiple machine learning models.

1.1.1 Plan of the analysis

The analysis consists of two major parts: 1. the exploratory data analysis which includes: - importing and examining the datasets, - examining variables of the each dataset, - preprocessing variables (doing dimension reduction with principal component analysis, transforming variables, constructing new variables, etc.), - merging preprocessed variables from various dataframes into one dataframe, - examining relationships between variables (calculating correlation coeffficients, doing statistical tests), - checking for missing values, outliers, and duplicates, 2. the machine learning which includes: - building machine learning modelling pipelines and functions, - running functions on different combinations of features, parameters and classifiers, - recursivelly selecting features by shap values, - randomly selecting features, - doing hyperparameter tuning with the Bayesian optimization, - testing the best performing model on the test data, - creating and running a deep learning model with the tensorflow library, - creating an API which will be deployed to the Google Cloud Platform.

The exploratory analysis and machine learning parts are presented in two separate Jupyter notebooks.

1.1.2 Hypotheses

Those clients will likely experience loan payment difficulties:

- who have worse loan payment history;
- who take loans of larger ammount;
- who have less qualified jobs or are unemployed;
- who have lower levels of education;
- who live in rented houses; etc.

These and other hypotheses will be tested in the course of the exploratory data analysis.

Importing libraries The main libraries which will be used for the manipulation with data are pandas and numpy. Matplotlib, seaborn and yellowbrick will be used for data visualization. Scipy, Statsmodels, Researchpy, Math, Random will be used for conducting statistical tests, calculating confidence intervals. Sklearn modules will be used for spliting data into training and testing samples, building and testing machine learning models. Optuna will be used for the Bayesian optimization. Tensorflow will be used for the deep learning modelling.

1.2 Exploratory analysis

Importing the datasets The Home Credit datasets are imported and saved into pandas dataframes.

The general information on the dataframes and numerical variables of the dataframes was obtained by looping info(), head() functions on the elements of the list of dataframes.

```
DATASET: aptrain

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 307511 entries, 0 to 307510
Columns: 122 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(41), object(16)
memory usage: 286.2+ MB

DATASET: aptest

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48744 entries, 0 to 48743
Columns: 121 entries, SK_ID_CURR to AMT_REQ_CREDIT_BUREAU_YEAR
dtypes: float64(65), int64(40), object(16)
memory usage: 45.0+ MB

DATASET: bureau

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1716428 entries, 0 to 1716427
Data columns (total 17 columns):
 #   Column                  Dtype
---  ------                  -----
```

```
 0   SK_ID_CURR              int64
 1   SK_ID_BUREAU            int64
 2   CREDIT_ACTIVE           object
 3   CREDIT_CURRENCY         object
 4   DAYS_CREDIT             int64
 5   CREDIT_DAY_OVERDUE      int64
 6   DAYS_CREDIT_ENDDATE     float64
 7   DAYS_ENDDATE_FACT       float64
 8   AMT_CREDIT_MAX_OVERDUE  float64
 9   CNT_CREDIT_PROLONG      int64
 10  AMT_CREDIT_SUM          float64
 11  AMT_CREDIT_SUM_DEBT     float64
 12  AMT_CREDIT_SUM_LIMIT    float64
 13  AMT_CREDIT_SUM_OVERDUE  float64
 14  CREDIT_TYPE             object
 15  DAYS_CREDIT_UPDATE      int64
 16  AMT_ANNUITY             float64
dtypes: float64(8), int64(6), object(3)
memory usage: 222.6+ MB

DATASET: bbalance

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27299925 entries, 0 to 27299924
Data columns (total 3 columns):
 #   Column          Dtype
---  ------          -----
 0   SK_ID_BUREAU    int64
 1   MONTHS_BALANCE  int64
 2   STATUS          object
dtypes: int64(2), object(1)
memory usage: 624.8+ MB

DATASET: pcbalance

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10001358 entries, 0 to 10001357
Data columns (total 8 columns):
 #   Column                 Dtype
---  ------                 -----
 0   SK_ID_PREV             int64
 1   SK_ID_CURR             int64
 2   MONTHS_BALANCE         int64
 3   CNT_INSTALMENT         float64
 4   CNT_INSTALMENT_FUTURE  float64
 5   NAME_CONTRACT_STATUS   object
 6   SK_DPD                 int64
 7   SK_DPD_DEF             int64
```

```
dtypes: float64(2), int64(5), object(1)
memory usage: 610.4+ MB

DATASET: bbalance

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27299925 entries, 0 to 27299924
Data columns (total 3 columns):
 #   Column          Dtype
---  ------          -----
 0   SK_ID_BUREAU    int64
 1   MONTHS_BALANCE  int64
 2   STATUS          object
dtypes: int64(2), object(1)
memory usage: 624.8+ MB

DATASET: ccbalance

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3840312 entries, 0 to 3840311
Data columns (total 23 columns):
 #   Column                      Dtype
---  ------                      -----
 0   SK_ID_PREV                  int64
 1   SK_ID_CURR                  int64
 2   MONTHS_BALANCE              int64
 3   AMT_BALANCE                 float64
 4   AMT_CREDIT_LIMIT_ACTUAL     int64
 5   AMT_DRAWINGS_ATM_CURRENT    float64
 6   AMT_DRAWINGS_CURRENT        float64
 7   AMT_DRAWINGS_OTHER_CURRENT  float64
 8   AMT_DRAWINGS_POS_CURRENT    float64
 9   AMT_INST_MIN_REGULARITY     float64
 10  AMT_PAYMENT_CURRENT         float64
 11  AMT_PAYMENT_TOTAL_CURRENT   float64
 12  AMT_RECEIVABLE_PRINCIPAL    float64
 13  AMT_RECIVABLE               float64
 14  AMT_TOTAL_RECEIVABLE        float64
 15  CNT_DRAWINGS_ATM_CURRENT    float64
 16  CNT_DRAWINGS_CURRENT        int64
 17  CNT_DRAWINGS_OTHER_CURRENT  float64
 18  CNT_DRAWINGS_POS_CURRENT    float64
 19  CNT_INSTALMENT_MATURE_CUM   float64
 20  NAME_CONTRACT_STATUS        object
 21  SK_DPD                      int64
 22  SK_DPD_DEF                  int64
dtypes: float64(15), int64(7), object(1)
memory usage: 673.9+ MB
```

```
DATASET: instpayments

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13605401 entries, 0 to 13605400
Data columns (total 8 columns):
 #   Column                  Dtype
---  ------                  -----
 0   SK_ID_PREV              int64
 1   SK_ID_CURR              int64
 2   NUM_INSTALMENT_VERSION  float64
 3   NUM_INSTALMENT_NUMBER   int64
 4   DAYS_INSTALMENT         float64
 5   DAYS_ENTRY_PAYMENT      float64
 6   AMT_INSTALMENT          float64
 7   AMT_PAYMENT             float64
dtypes: float64(5), int64(3)
memory usage: 830.4 MB

DATASET: prevapplication

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1670214 entries, 0 to 1670213
Data columns (total 37 columns):
 #   Column                       Non-Null Count    Dtype
---  ------                       --------------    -----
 0   SK_ID_PREV                   1670214 non-null  int64
 1   SK_ID_CURR                   1670214 non-null  int64
 2   NAME_CONTRACT_TYPE           1670214 non-null  object
 3   AMT_ANNUITY                  1297979 non-null  float64
 4   AMT_APPLICATION              1670214 non-null  float64
 5   AMT_CREDIT                   1670213 non-null  float64
 6   AMT_DOWN_PAYMENT             774370 non-null   float64
 7   AMT_GOODS_PRICE              1284699 non-null  float64
 8   WEEKDAY_APPR_PROCESS_START   1670214 non-null  object
 9   HOUR_APPR_PROCESS_START      1670214 non-null  int64
 10  FLAG_LAST_APPL_PER_CONTRACT  1670214 non-null  object
 11  NFLAG_LAST_APPL_IN_DAY       1670214 non-null  int64
 12  RATE_DOWN_PAYMENT            774370 non-null   float64
 13  RATE_INTEREST_PRIMARY        5951 non-null     float64
 14  RATE_INTEREST_PRIVILEGED     5951 non-null     float64
 15  NAME_CASH_LOAN_PURPOSE       1670214 non-null  object
 16  NAME_CONTRACT_STATUS         1670214 non-null  object
 17  DAYS_DECISION                1670214 non-null  int64
 18  NAME_PAYMENT_TYPE            1670214 non-null  object
 19  CODE_REJECT_REASON           1670214 non-null  object
 20  NAME_TYPE_SUITE              849809 non-null   object
 21  NAME_CLIENT_TYPE             1670214 non-null  object
 22  NAME_GOODS_CATEGORY          1670214 non-null  object
 23  NAME_PORTFOLIO               1670214 non-null  object
 24  NAME_PRODUCT_TYPE            1670214 non-null  object
 25  CHANNEL_TYPE                 1670214 non-null  object
 26  SELLERPLACE_AREA             1670214 non-null  int64
 27  NAME_SELLER_INDUSTRY         1670214 non-null  object
 28  CNT_PAYMENT                  1297984 non-null  float64
 29  NAME_YIELD_GROUP             1670214 non-null  object
 30  PRODUCT_COMBINATION          1669868 non-null  object
 31  DAYS_FIRST_DRAWING           997149 non-null   float64
 32  DAYS_FIRST_DUE               997149 non-null   float64
 33  DAYS_LAST_DUE_1ST_VERSION    997149 non-null   float64
 34  DAYS_LAST_DUE                997149 non-null   float64
 35  DAYS_TERMINATION             997149 non-null   float64
 36  NFLAG_INSURED_ON_APPROVAL    997149 non-null   float64
dtypes: float64(15), int64(6), object(16)
memory usage: 471.5+ MB

DATASET: sampsubmission

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48744 entries, 0 to 48743
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   SK_ID_CURR  48744 non-null  int64
 1   TARGET      48744 non-null  float64
dtypes: float64(1), int64(1)
memory usage: 761.8 KB


DATASET: aptrain

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

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

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

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

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

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

[5 rows x 122 columns]

DATASET: aptest

   SK_ID_CURR NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY  \
0      100001         Cash loans           F            N               Y
1      100005         Cash loans           M            N               Y
2      100013         Cash loans           M            Y               Y
3      100028         Cash loans           F            N               Y
4      100038         Cash loans           M            Y               N

   CNT_CHILDREN  AMT_INCOME_TOTAL  AMT_CREDIT  AMT_ANNUITY  AMT_GOODS_PRICE  \
0             0          135000.0    568800.0      20560.5         450000.0
1             0           99000.0    222768.0      17370.0         180000.0
2             0          202500.0    663264.0      69777.0         630000.0
3             2          315000.0   1575000.0      49018.5        1575000.0
4             1          180000.0    625500.0      32067.0         625500.0

   … FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21  \
0  …                0                0                0                0
1  …                0                0                0                0
```

```
2  …                0                0                0                0
3  …                0                0                0                0
4  …                0                0                0                0

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

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

   AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
0                        0.0                         0.0
1                        0.0                         3.0
2                        1.0                         4.0
3                        0.0                         3.0
4                        NaN                         NaN

[5 rows x 121 columns]

DATASET: bureau

   SK_ID_CURR  SK_ID_BUREAU CREDIT_ACTIVE CREDIT_CURRENCY  DAYS_CREDIT  \
0      215354       5714462        Closed      currency 1         -497
1      215354       5714463        Active      currency 1         -208
2      215354       5714464        Active      currency 1         -203
3      215354       5714465        Active      currency 1         -203
4      215354       5714466        Active      currency 1         -629

   CREDIT_DAY_OVERDUE  DAYS_CREDIT_ENDDATE  DAYS_ENDDATE_FACT  \
0                   0               -153.0             -153.0
1                   0               1075.0                NaN
2                   0                528.0                NaN
3                   0                  NaN                NaN
4                   0               1197.0                NaN

   AMT_CREDIT_MAX_OVERDUE  CNT_CREDIT_PROLONG  AMT_CREDIT_SUM  \
0                     NaN                   0         91323.0
1                     NaN                   0        225000.0
2                     NaN                   0        464323.5
3                     NaN                   0         90000.0
```

```
4                 77674.5                   0       2700000.0

   AMT_CREDIT_SUM_DEBT  AMT_CREDIT_SUM_LIMIT  AMT_CREDIT_SUM_OVERDUE  \
0                  0.0                   NaN                     0.0
1             171342.0                   NaN                     0.0
2                  NaN                   NaN                     0.0
3                  NaN                   NaN                     0.0
4                  NaN                   NaN                     0.0

       CREDIT_TYPE  DAYS_CREDIT_UPDATE  AMT_ANNUITY
0  Consumer credit                -131          NaN
1      Credit card                 -20          NaN
2  Consumer credit                 -16          NaN
3      Credit card                 -16          NaN
4  Consumer credit                 -21          NaN

DATASET: bbalance

   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

DATASET: pcbalance

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

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

DATASET: bbalance

   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

DATASET: ccbalance

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

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

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

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

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

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

   SK_DPD  SK_DPD_DEF
0       0           0
```

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

[5 rows x 23 columns]

DATASET: instpayments

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

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

DATASET: prevapplication

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

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

   HOUR_APPR_PROCESS_START  … NAME_SELLER_INDUSTRY  CNT_PAYMENT  \
0                       15  …         Connectivity         12.0
1                       11  …                  XNA         36.0
2                       11  …                  XNA         12.0
3                        7  …                  XNA         12.0
4                        9  …                  XNA         24.0

   NAME_YIELD_GROUP       PRODUCT_COMBINATION  DAYS_FIRST_DRAWING  \
0            middle  POS mobile with interest            365243.0
```

```
1        low_action          Cash X-Sell: low            365243.0
2              high         Cash X-Sell: high            365243.0
3            middle       Cash X-Sell: middle            365243.0
4              high         Cash Street: high                 NaN

  DAYS_FIRST_DUE DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE DAYS_TERMINATION  \
0          -42.0                     300.0          -42.0             -37.0
1         -134.0                     916.0       365243.0          365243.0
2         -271.0                      59.0       365243.0          365243.0
3         -482.0                    -152.0         -182.0            -177.0
4            NaN                       NaN            NaN               NaN

  NFLAG_INSURED_ON_APPROVAL
0                       0.0
1                       1.0
2                       1.0
3                       1.0
4                       NaN

[5 rows x 37 columns]

DATASET: sampsubmission

   SK_ID_CURR  TARGET
0      100001     0.5
1      100005     0.5
2      100013     0.5
3      100028     0.5
4      100038     0.5
```

Numerical variables Statistics on numerical variables in all data sets will be presented by looping the describe() function on the elements of the list of dataframes. However, there is a need to separate pure numerical variables from binary categorical variables which have the numeric form (values 0 and 1). For the purpose of such separation the functions 'binary_numeric_or_zeros" and "get_binary_numeric_or_zeros" are created, which extract features which have 2 or less unique values (and avoids errors if there are no such features in a dataframe).

The function describe() is run on the datasets, and the outputs of numerical variables are printed with exception of the binary numeric variables.

```
DATASET: aptrain

          SK_ID_CURR   CNT_CHILDREN  AMT_INCOME_TOTAL    AMT_CREDIT  \
count  307511.000000  307511.000000      3.075110e+05  3.075110e+05
mean   278180.518577       0.417052      1.687979e+05  5.990260e+05
std    102790.175348       0.722121      2.371231e+05  4.024908e+05
min    100002.000000       0.000000      2.565000e+04  4.500000e+04
```

```
25%    189145.500000       0.000000      1.125000e+05  2.700000e+05
50%    278202.000000       0.000000      1.471500e+05  5.135310e+05
75%    367142.500000       1.000000      2.025000e+05  8.086500e+05
max    456255.000000      19.000000      1.170000e+08  4.050000e+06

         AMT_ANNUITY  AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE  \
count  307499.000000     3.072330e+05               307511.000000
mean    27108.573909     5.383962e+05                    0.020868
std     14493.737315     3.694465e+05                    0.013831
min      1615.500000     4.050000e+04                    0.000290
25%     16524.000000     2.385000e+05                    0.010006
50%     24903.000000     4.500000e+05                    0.018850
75%     34596.000000     6.795000e+05                    0.028663
max    258025.500000     4.050000e+06                    0.072508

          DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  …  \
count  307511.000000  307511.000000      307511.000000  …
mean   -16036.995067   63815.045904       -4986.120328  …
std      4363.988632  141275.766519        3522.886321  …
min    -25229.000000  -17912.000000      -24672.000000  …
25%    -19682.000000   -2760.000000       -7479.500000  …
50%    -15750.000000   -1213.000000       -4504.000000  …
75%    -12413.000000    -289.000000       -2010.000000  …
max     -7489.000000  365243.000000           0.000000  …

       DEF_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE  \
count             306490.000000             306490.000000
mean                   0.143421                  1.405292
std                    0.446698                  2.379803
min                    0.000000                  0.000000
25%                    0.000000                  0.000000
50%                    0.000000                  0.000000
75%                    0.000000                  2.000000
max                   34.000000                344.000000

       DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
count             306490.000000           307510.000000
mean                   0.100049             -962.858788
std                    0.362291              826.808487
min                    0.000000            -4292.000000
25%                    0.000000            -1570.000000
50%                    0.000000             -757.000000
75%                    0.000000             -274.000000
max                   24.000000                0.000000

       AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
count               265992.000000              265992.000000
mean                     0.006402                   0.007000
```

```
std                      0.083849                   0.110757
min                      0.000000                   0.000000
25%                      0.000000                   0.000000
50%                      0.000000                   0.000000
75%                      0.000000                   0.000000
max                      4.000000                   9.000000

       AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
count               265992.000000              265992.000000
mean                     0.034362                   0.267395
std                      0.204685                   0.916002
min                      0.000000                   0.000000
25%                      0.000000                   0.000000
50%                      0.000000                   0.000000
75%                      0.000000                   0.000000
max                      8.000000                  27.000000

       AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
count              265992.000000               265992.000000
mean                    0.265474                    1.899974
std                     0.794056                    1.869295
min                     0.000000                    0.000000
25%                     0.000000                    0.000000
50%                     0.000000                    1.000000
75%                     0.000000                    3.000000
max                   261.000000                   25.000000

[8 rows x 73 columns]

DATASET: aptest

          SK_ID_CURR  CNT_CHILDREN  AMT_INCOME_TOTAL    AMT_CREDIT  \
count   48744.000000  48744.000000      4.874400e+04  4.874400e+04
mean   277796.676350      0.397054      1.784318e+05  5.167404e+05
std    103169.547296      0.709047      1.015226e+05  3.653970e+05
min    100001.000000      0.000000      2.694150e+04  4.500000e+04
25%    188557.750000      0.000000      1.125000e+05  2.606400e+05
50%    277549.000000      0.000000      1.575000e+05  4.500000e+05
75%    367555.500000      1.000000      2.250000e+05  6.750000e+05
max    456250.000000     20.000000      4.410000e+06  2.245500e+06

         AMT_ANNUITY  AMT_GOODS_PRICE  REGION_POPULATION_RELATIVE  \
count   48720.000000     4.874400e+04                48744.000000
mean    29426.240209     4.626188e+05                    0.021226
std     16016.368315     3.367102e+05                    0.014428
min      2295.000000     4.500000e+04                    0.000253
25%     17973.000000     2.250000e+05                    0.010006
50%     26199.000000     3.960000e+05                    0.018850
```

```
75%     37390.500000     6.300000e+05                    0.028663
max    180576.000000     2.245500e+06                    0.072508

         DAYS_BIRTH  DAYS_EMPLOYED  DAYS_REGISTRATION  …  \
count  48744.000000   48744.000000       48744.000000  …   
mean  -16068.084605   67485.366322       -4967.652716  …   
std     4325.900393  144348.507136        3552.612035  …   
min   -25195.000000  -17463.000000      -23722.000000  …   
25%   -19637.000000   -2910.000000       -7459.250000  …   
50%   -15785.000000   -1293.000000       -4490.000000  …   
75%   -12496.000000    -296.000000       -1901.000000  …   
max    -7338.000000  365243.000000           0.000000  …   

       DEF_30_CNT_SOCIAL_CIRCLE  OBS_60_CNT_SOCIAL_CIRCLE  \
count              48715.000000              48715.000000   
mean                   0.143652                  1.435738   
std                    0.514413                  3.580125   
min                    0.000000                  0.000000   
25%                    0.000000                  0.000000   
50%                    0.000000                  0.000000   
75%                    0.000000                  2.000000   
max                   34.000000                351.000000   

       DEF_60_CNT_SOCIAL_CIRCLE  DAYS_LAST_PHONE_CHANGE  \
count              48715.000000            48744.000000   
mean                   0.101139            -1077.766228   
std                    0.403791              878.920740   
min                    0.000000            -4361.000000   
25%                    0.000000            -1766.250000   
50%                    0.000000             -863.000000   
75%                    0.000000             -363.000000   
max                   24.000000                0.000000   

       AMT_REQ_CREDIT_BUREAU_HOUR  AMT_REQ_CREDIT_BUREAU_DAY  \
count                42695.000000               42695.000000   
mean                     0.002108                   0.001803   
std                      0.046373                   0.046132   
min                      0.000000                   0.000000   
25%                      0.000000                   0.000000   
50%                      0.000000                   0.000000   
75%                      0.000000                   0.000000   
max                      2.000000                   2.000000   

       AMT_REQ_CREDIT_BUREAU_WEEK  AMT_REQ_CREDIT_BUREAU_MON  \
count                42695.000000               42695.000000   
mean                     0.002787                   0.009299   
std                      0.054037                   0.110924   
min                      0.000000                   0.000000   
```

```
25%                      0.000000                   0.000000
50%                      0.000000                   0.000000
75%                      0.000000                   0.000000
max                      2.000000                   6.000000

       AMT_REQ_CREDIT_BUREAU_QRT  AMT_REQ_CREDIT_BUREAU_YEAR
count               42695.000000                42695.000000
mean                    0.546902                    1.983769
std                     0.693305                    1.838873
min                     0.000000                    0.000000
25%                     0.000000                    0.000000
50%                     0.000000                    2.000000
75%                     1.000000                    3.000000
max                     7.000000                   17.000000

[8 rows x 73 columns]

DATASET: bureau

         SK_ID_CURR  SK_ID_BUREAU   DAYS_CREDIT  CREDIT_DAY_OVERDUE  \
count  1.716428e+06  1.716428e+06  1.716428e+06        1.716428e+06
mean   2.782149e+05  5.924434e+06 -1.142108e+03        8.181666e-01
std    1.029386e+05  5.322657e+05  7.951649e+02        3.654443e+01
min    1.000010e+05  5.000000e+06 -2.922000e+03        0.000000e+00
25%    1.888668e+05  5.463954e+06 -1.666000e+03        0.000000e+00
50%    2.780550e+05  5.926304e+06 -9.870000e+02        0.000000e+00
75%    3.674260e+05  6.385681e+06 -4.740000e+02        0.000000e+00
max    4.562550e+05  6.843457e+06  0.000000e+00        2.792000e+03

       DAYS_CREDIT_ENDDATE  DAYS_ENDDATE_FACT  AMT_CREDIT_MAX_OVERDUE  \
count         1.610875e+06       1.082775e+06            5.919400e+05
mean          5.105174e+02      -1.017437e+03            3.825418e+03
std           4.994220e+03       7.140106e+02            2.060316e+05
min          -4.206000e+04      -4.202300e+04            0.000000e+00
25%          -1.138000e+03      -1.489000e+03            0.000000e+00
50%          -3.300000e+02      -8.970000e+02            0.000000e+00
75%           4.740000e+02      -4.250000e+02            0.000000e+00
max           3.119900e+04       0.000000e+00            1.159872e+08

       CNT_CREDIT_PROLONG  AMT_CREDIT_SUM  AMT_CREDIT_SUM_DEBT  \
count        1.716428e+06    1.716415e+06         1.458759e+06
mean         6.410406e-03    3.549946e+05         1.370851e+05
std          9.622391e-02    1.149811e+06         6.774011e+05
min          0.000000e+00    0.000000e+00        -4.705600e+06
25%          0.000000e+00    5.130000e+04         0.000000e+00
50%          0.000000e+00    1.255185e+05         0.000000e+00
75%          0.000000e+00    3.150000e+05         4.015350e+04
max          9.000000e+00    5.850000e+08         1.701000e+08
```

```
       AMT_CREDIT_SUM_LIMIT  AMT_CREDIT_SUM_OVERDUE  DAYS_CREDIT_UPDATE  \
count          1.124648e+06            1.716428e+06        1.716428e+06
mean           6.229515e+03            3.791276e+01       -5.937483e+02
std            4.503203e+04            5.937650e+03        7.207473e+02
min           -5.864061e+05            0.000000e+00       -4.194700e+04
25%            0.000000e+00            0.000000e+00       -9.080000e+02
50%            0.000000e+00            0.000000e+00       -3.950000e+02
75%            0.000000e+00            0.000000e+00       -3.300000e+01
max            4.705600e+06            3.756681e+06        3.720000e+02

        AMT_ANNUITY
count  4.896370e+05
mean   1.571276e+04
std    3.258269e+05
min    0.000000e+00
25%    0.000000e+00
50%    0.000000e+00
75%    1.350000e+04
max    1.184534e+08

DATASET: bbalance

       SK_ID_BUREAU  MONTHS_BALANCE
count  2.729992e+07    2.729992e+07
mean   6.036297e+06   -3.074169e+01
std    4.923489e+05    2.386451e+01
min    5.001709e+06   -9.600000e+01
25%    5.730933e+06   -4.600000e+01
50%    6.070821e+06   -2.500000e+01
75%    6.431951e+06   -1.100000e+01
max    6.842888e+06    0.000000e+00

DATASET: pcbalance

         SK_ID_PREV    SK_ID_CURR  MONTHS_BALANCE  CNT_INSTALMENT  \
count  1.000136e+07  1.000136e+07    1.000136e+07    9.975287e+06
mean   1.903217e+06  2.784039e+05   -3.501259e+01    1.708965e+01
std    5.358465e+05  1.027637e+05    2.606657e+01    1.199506e+01
min    1.000001e+06  1.000010e+05   -9.600000e+01    1.000000e+00
25%    1.434405e+06  1.895500e+05   -5.400000e+01    1.000000e+01
50%    1.896565e+06  2.786540e+05   -2.800000e+01    1.200000e+01
75%    2.368963e+06  3.674290e+05   -1.300000e+01    2.400000e+01
max    2.843499e+06  4.562550e+05   -1.000000e+00    9.200000e+01

       CNT_INSTALMENT_FUTURE        SK_DPD    SK_DPD_DEF
count           9.975271e+06  1.000136e+07  1.000136e+07
mean            1.048384e+01  1.160693e+01  6.544684e-01
```

```
std             1.110906e+01  1.327140e+02  3.276249e+01
min             0.000000e+00  0.000000e+00  0.000000e+00
25%             3.000000e+00  0.000000e+00  0.000000e+00
50%             7.000000e+00  0.000000e+00  0.000000e+00
75%             1.400000e+01  0.000000e+00  0.000000e+00
max             8.500000e+01  4.231000e+03  3.595000e+03

DATASET: bbalance

       SK_ID_BUREAU  MONTHS_BALANCE
count  2.729992e+07    2.729992e+07
mean   6.036297e+06   -3.074169e+01
std    4.923489e+05    2.386451e+01
min    5.001709e+06   -9.600000e+01
25%    5.730933e+06   -4.600000e+01
50%    6.070821e+06   -2.500000e+01
75%    6.431951e+06   -1.100000e+01
max    6.842888e+06    0.000000e+00

DATASET: ccbalance

         SK_ID_PREV    SK_ID_CURR  MONTHS_BALANCE   AMT_BALANCE  \
count  3.840312e+06  3.840312e+06    3.840312e+06  3.840312e+06
mean   1.904504e+06  2.783242e+05   -3.452192e+01  5.830016e+04
std    5.364695e+05  1.027045e+05    2.666775e+01  1.063070e+05
min    1.000018e+06  1.000060e+05   -9.600000e+01 -4.202502e+05
25%    1.434385e+06  1.895170e+05   -5.500000e+01  0.000000e+00
50%    1.897122e+06  2.783960e+05   -2.800000e+01  0.000000e+00
75%    2.369328e+06  3.675800e+05   -1.100000e+01  8.904669e+04
max    2.843496e+06  4.562500e+05   -1.000000e+00  1.505902e+06

       AMT_CREDIT_LIMIT_ACTUAL  AMT_DRAWINGS_ATM_CURRENT  \
count             3.840312e+06              3.090496e+06
mean              1.538080e+05              5.961325e+03
std               1.651457e+05              2.822569e+04
min               0.000000e+00             -6.827310e+03
25%               4.500000e+04              0.000000e+00
50%               1.125000e+05              0.000000e+00
75%               1.800000e+05              0.000000e+00
max               1.350000e+06              2.115000e+06

       AMT_DRAWINGS_CURRENT  AMT_DRAWINGS_OTHER_CURRENT  \
count          3.840312e+06                3.090496e+06
mean           7.433388e+03                2.881696e+02
std            3.384608e+04                8.201989e+03
min           -6.211620e+03                0.000000e+00
25%            0.000000e+00                0.000000e+00
50%            0.000000e+00                0.000000e+00
```

```
75%            0.000000e+00                0.000000e+00
max            2.287098e+06                1.529847e+06

       AMT_DRAWINGS_POS_CURRENT  AMT_INST_MIN_REGULARITY  …  \
count              3.090496e+06             3.535076e+06  …
mean               2.968805e+03             3.540204e+03  …
std                2.079689e+04             5.600154e+03  …
min                0.000000e+00             0.000000e+00  …
25%                0.000000e+00             0.000000e+00  …
50%                0.000000e+00             0.000000e+00  …
75%                0.000000e+00             6.633911e+03  …
max                2.239274e+06             2.028820e+05  …

       AMT_RECEIVABLE_PRINCIPAL  AMT_RECIVABLE  AMT_TOTAL_RECEIVABLE  \
count              3.840312e+06   3.840312e+06          3.840312e+06
mean               5.596588e+04   5.808881e+04          5.809829e+04
std                1.025336e+05   1.059654e+05          1.059718e+05
min               -4.233058e+05  -4.202502e+05         -4.202502e+05
25%                0.000000e+00   0.000000e+00          0.000000e+00
50%                0.000000e+00   0.000000e+00          0.000000e+00
75%                8.535924e+04   8.889949e+04          8.891451e+04
max                1.472317e+06   1.493338e+06          1.493338e+06

       CNT_DRAWINGS_ATM_CURRENT  CNT_DRAWINGS_CURRENT  \
count              3.090496e+06          3.840312e+06
mean               3.094490e-01          7.031439e-01
std                1.100401e+00          3.190347e+00
min                0.000000e+00          0.000000e+00
25%                0.000000e+00          0.000000e+00
50%                0.000000e+00          0.000000e+00
75%                0.000000e+00          0.000000e+00
max                5.100000e+01          1.650000e+02

       CNT_DRAWINGS_OTHER_CURRENT  CNT_DRAWINGS_POS_CURRENT  \
count                3.090496e+06              3.090496e+06
mean                 4.812496e-03              5.594791e-01
std                  8.263861e-02              3.240649e+00
min                  0.000000e+00              0.000000e+00
25%                  0.000000e+00              0.000000e+00
50%                  0.000000e+00              0.000000e+00
75%                  0.000000e+00              0.000000e+00
max                  1.200000e+01              1.650000e+02

       CNT_INSTALMENT_MATURE_CUM        SK_DPD    SK_DPD_DEF
count               3.535076e+06  3.840312e+06  3.840312e+06
mean                2.082508e+01  9.283667e+00  3.316220e-01
std                 2.005149e+01  9.751570e+01  2.147923e+01
min                 0.000000e+00  0.000000e+00  0.000000e+00
```

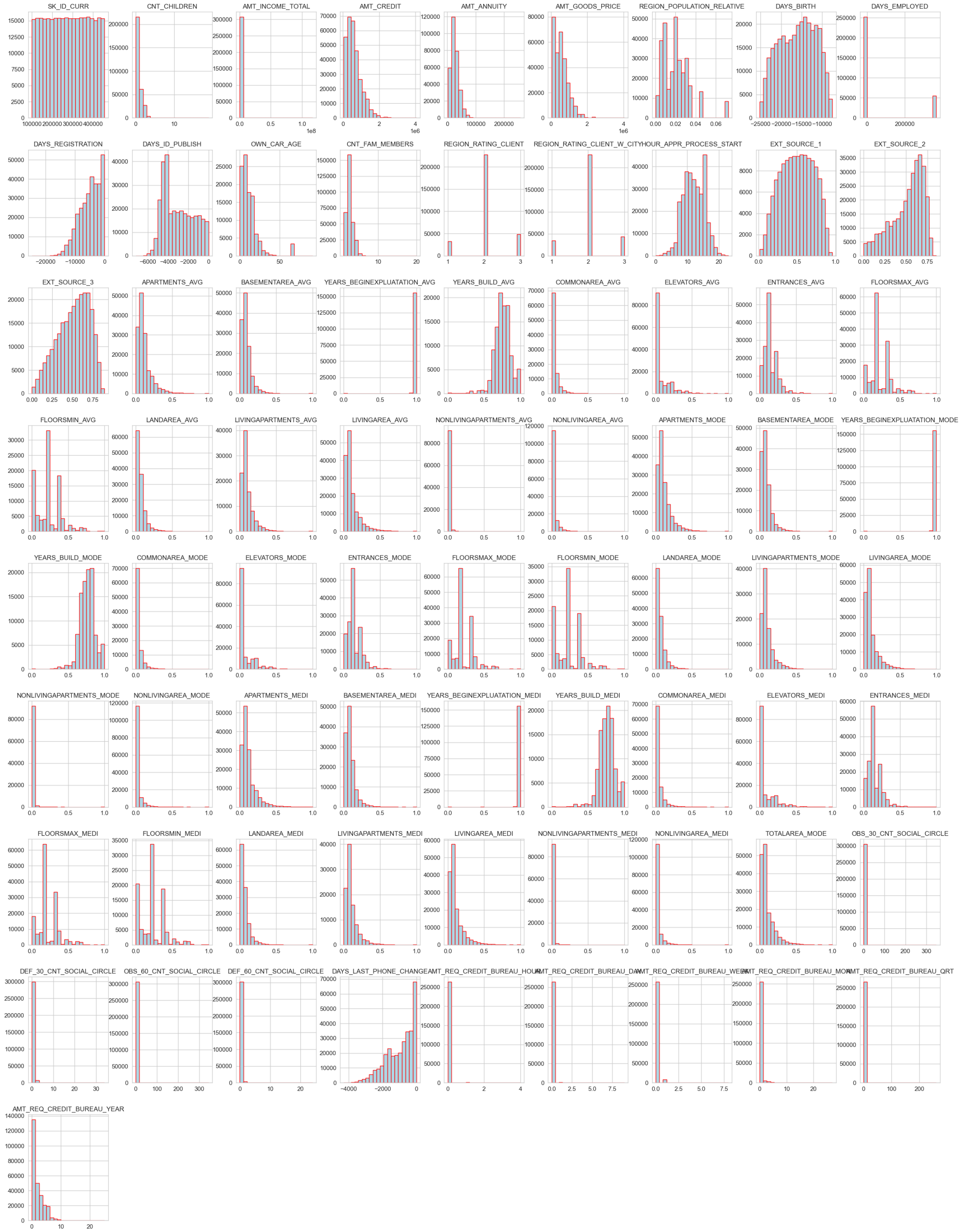
```
25%                 4.000000e+00  0.000000e+00  0.000000e+00
50%                 1.500000e+01  0.000000e+00  0.000000e+00
75%                 3.200000e+01  0.000000e+00  0.000000e+00
max                 1.200000e+02  3.260000e+03  3.260000e+03

[8 rows x 22 columns]

DATASET: instpayments

         SK_ID_PREV    SK_ID_CURR  NUM_INSTALMENT_VERSION  \
count  1.360540e+07  1.360540e+07            1.360540e+07
mean   1.903365e+06  2.784449e+05            8.566373e-01
std    5.362029e+05  1.027183e+05            1.035216e+00
min    1.000001e+06  1.000010e+05            0.000000e+00
25%    1.434191e+06  1.896390e+05            0.000000e+00
50%    1.896520e+06  2.786850e+05            1.000000e+00
75%    2.369094e+06  3.675300e+05            1.000000e+00
max    2.843499e+06  4.562550e+05            1.780000e+02

       NUM_INSTALMENT_NUMBER  DAYS_INSTALMENT  DAYS_ENTRY_PAYMENT  \
count           1.360540e+07     1.360540e+07        1.360250e+07
mean            1.887090e+01    -1.042270e+03       -1.051114e+03
std             2.666407e+01     8.009463e+02        8.005859e+02
min             1.000000e+00    -2.922000e+03       -4.921000e+03
25%             4.000000e+00    -1.654000e+03       -1.662000e+03
50%             8.000000e+00    -8.180000e+02       -8.270000e+02
75%             1.900000e+01    -3.610000e+02       -3.700000e+02
max             2.770000e+02    -1.000000e+00       -1.000000e+00

       AMT_INSTALMENT   AMT_PAYMENT
count    1.360540e+07  1.360250e+07
mean     1.705091e+04  1.723822e+04
std      5.057025e+04  5.473578e+04
min      0.000000e+00  0.000000e+00
25%      4.226085e+03  3.398265e+03
50%      8.884080e+03  8.125515e+03
75%      1.671021e+04  1.610842e+04
max      3.771488e+06  3.771488e+06

DATASET: prevapplication

         SK_ID_PREV    SK_ID_CURR   AMT_ANNUITY  AMT_APPLICATION  \
count  1.670214e+06  1.670214e+06  1.297979e+06     1.670214e+06
mean   1.923089e+06  2.783572e+05  1.595512e+04     1.752339e+05
std    5.325980e+05  1.028148e+05  1.478214e+04     2.927798e+05
min    1.000001e+06  1.000010e+05  0.000000e+00     0.000000e+00
25%    1.461857e+06  1.893290e+05  6.321780e+03     1.872000e+04
50%    1.923110e+06  2.787145e+05  1.125000e+04     7.104600e+04
```

```
75%    2.384280e+06  3.675140e+05  2.065842e+04     1.803600e+05
max    2.845382e+06  4.562550e+05  4.180581e+05     6.905160e+06

         AMT_CREDIT  AMT_DOWN_PAYMENT  AMT_GOODS_PRICE  \
count  1.670213e+06      7.743700e+05     1.284699e+06
mean   1.961140e+05      6.697402e+03     2.278473e+05
std    3.185746e+05      2.092150e+04     3.153966e+05
min    0.000000e+00     -9.000000e-01     0.000000e+00
25%    2.416050e+04      0.000000e+00     5.084100e+04
50%    8.054100e+04      1.638000e+03     1.123200e+05
75%    2.164185e+05      7.740000e+03     2.340000e+05
max    6.905160e+06      3.060045e+06     6.905160e+06

       HOUR_APPR_PROCESS_START  RATE_DOWN_PAYMENT  RATE_INTEREST_PRIMARY  \
count             1.670214e+06      774370.000000            5951.000000
mean              1.248418e+01           0.079637               0.188357
std               3.334028e+00           0.107823               0.087671
min               0.000000e+00          -0.000015               0.034781
25%               1.000000e+01           0.000000               0.160716
50%               1.200000e+01           0.051605               0.189122
75%               1.500000e+01           0.108909               0.193330
max               2.300000e+01           1.000000               1.000000

       RATE_INTEREST_PRIVILEGED  DAYS_DECISION  SELLERPLACE_AREA  \
count               5951.000000   1.670214e+06      1.670214e+06
mean                   0.773503  -8.806797e+02      3.139511e+02
std                    0.100879   7.790997e+02      7.127443e+03
min                    0.373150  -2.922000e+03     -1.000000e+00
25%                    0.715645  -1.300000e+03     -1.000000e+00
50%                    0.835095  -5.810000e+02      3.000000e+00
75%                    0.852537  -2.800000e+02      8.200000e+01
max                    1.000000  -1.000000e+00      4.000000e+06

        CNT_PAYMENT  DAYS_FIRST_DRAWING  DAYS_FIRST_DUE  \
count  1.297984e+06       997149.000000   997149.000000
mean   1.605408e+01       342209.855039    13826.269337
std    1.456729e+01        88916.115833    72444.869708
min    0.000000e+00        -2922.000000    -2892.000000
25%    6.000000e+00       365243.000000    -1628.000000
50%    1.200000e+01       365243.000000     -831.000000
75%    2.400000e+01       365243.000000     -411.000000
max    8.400000e+01       365243.000000   365243.000000

       DAYS_LAST_DUE_1ST_VERSION  DAYS_LAST_DUE  DAYS_TERMINATION  \
count              997149.000000  997149.000000     997149.000000
mean                33767.774054   76582.403064      81992.343838
std                106857.034789  149647.415123     153303.516729
min                 -2801.000000   -2889.000000      -2874.000000
```

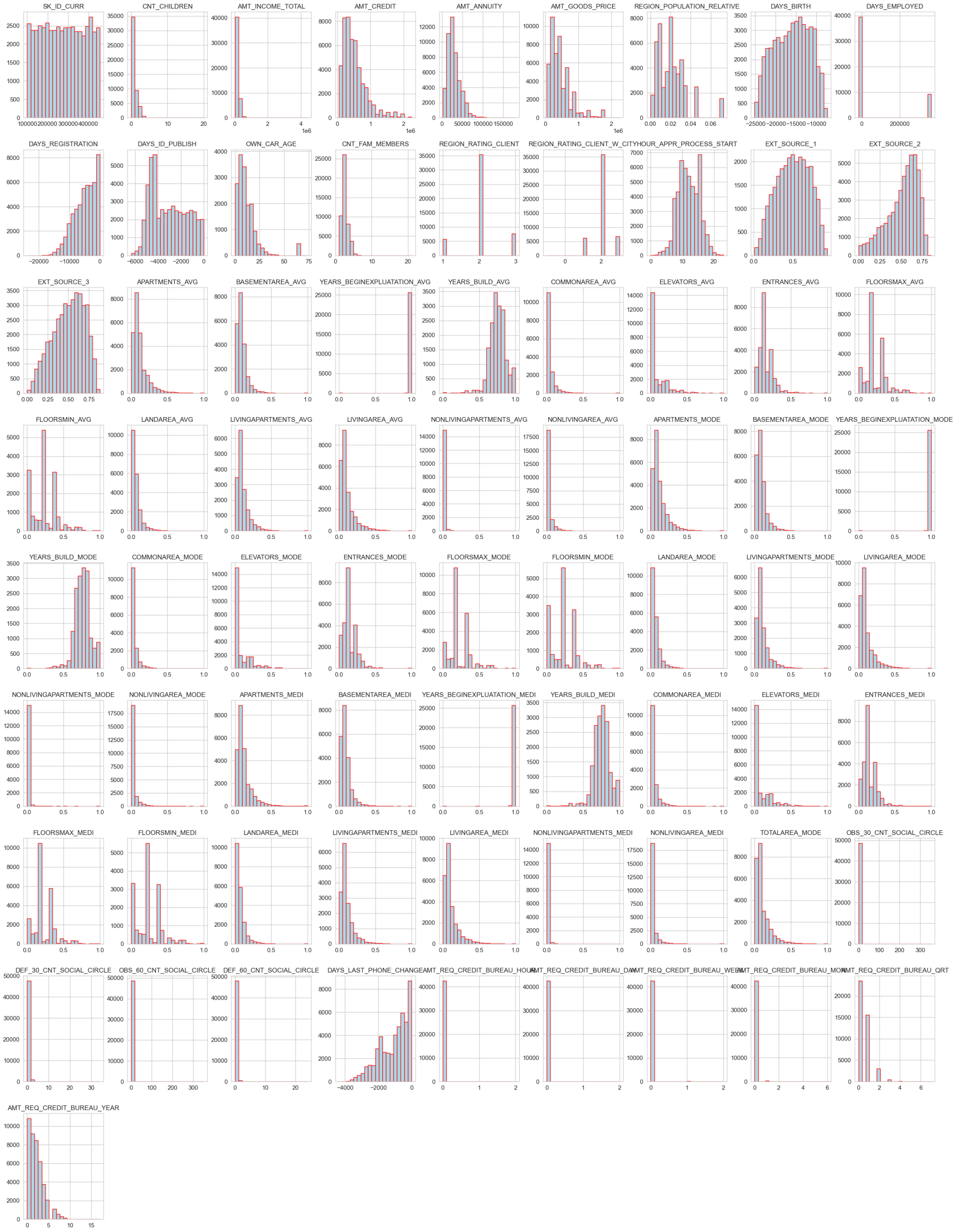
```
25%                -1242.000000   -1314.000000      -1270.000000
50%                 -361.000000    -537.000000       -499.000000
75%                  129.000000     -74.000000        -44.000000
max               365243.000000  365243.000000     365243.000000

       NFLAG_INSURED_ON_APPROVAL
count              997149.000000
mean                    0.332570
std                     0.471134
min                     0.000000
25%                     0.000000
50%                     0.000000
75%                     1.000000
max                     1.000000
```

Distributions of numerical variables in all dataframes are plotted with histograms (binary numeric variables are excluded; id variables were not excluded from the plots, though they should also not to be treated as numerical variables). Plots are saved to png files for the deeper examination.

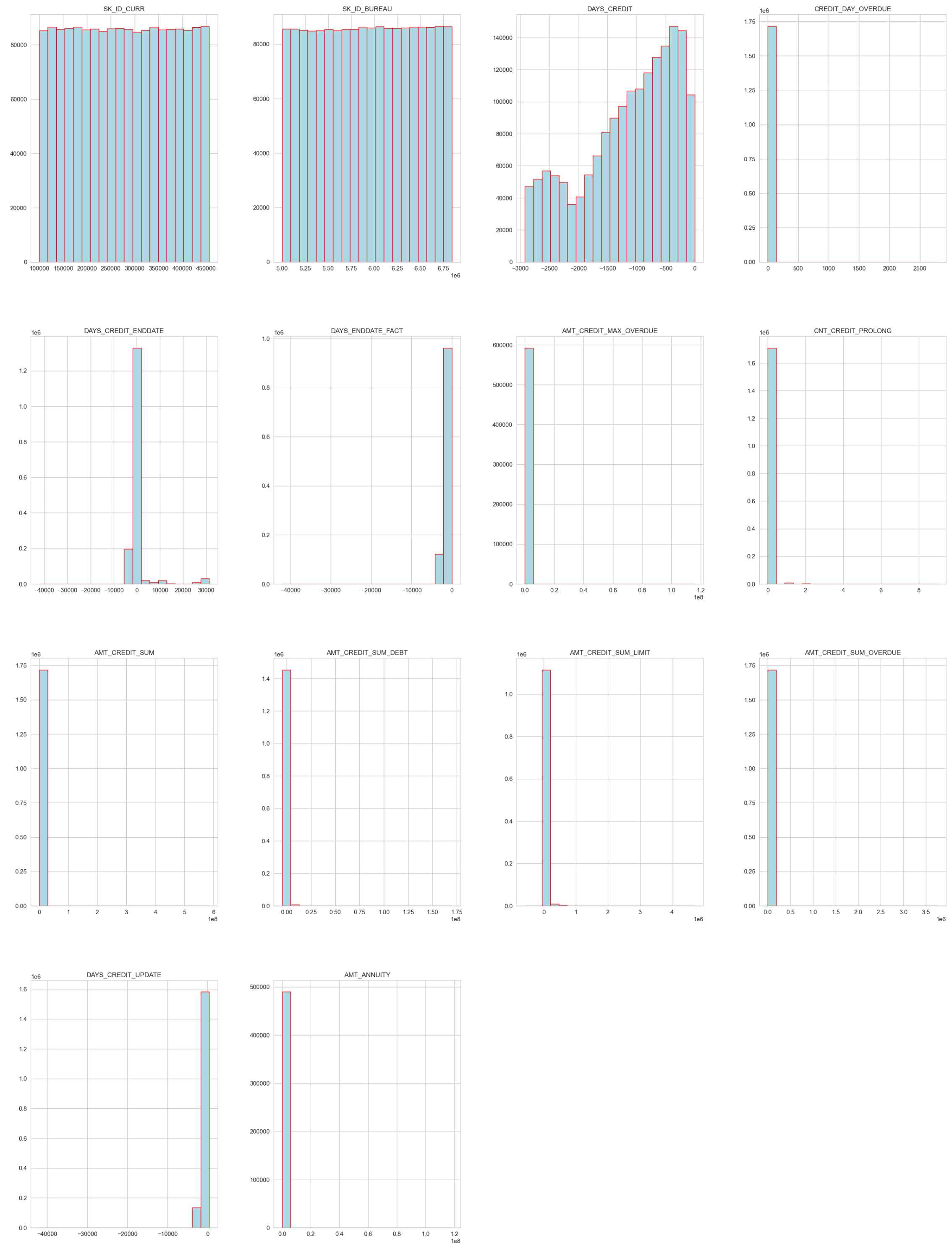
```
DATASET: aptrain
```



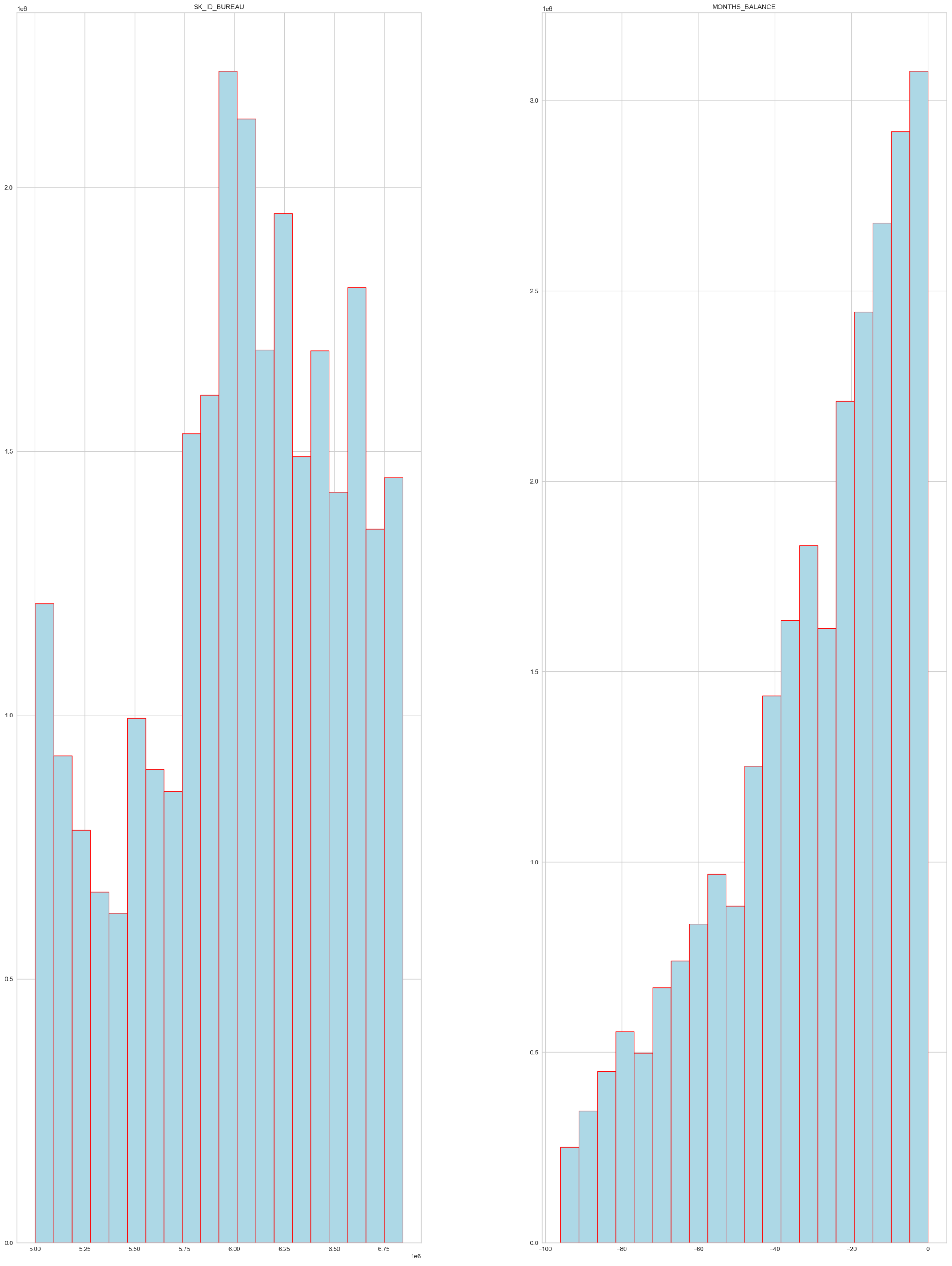
`DATASET: aptest`



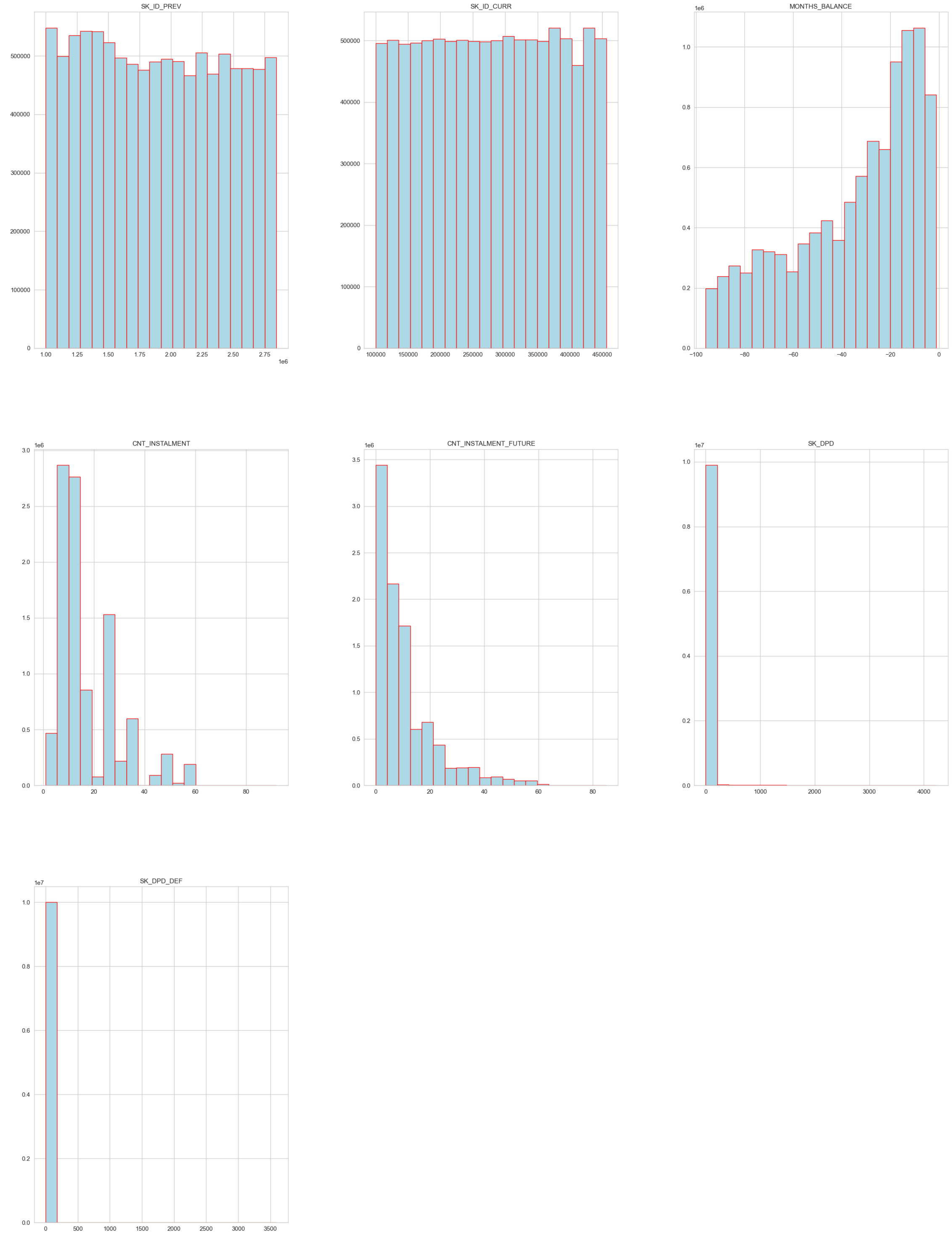
DATASET: bureau



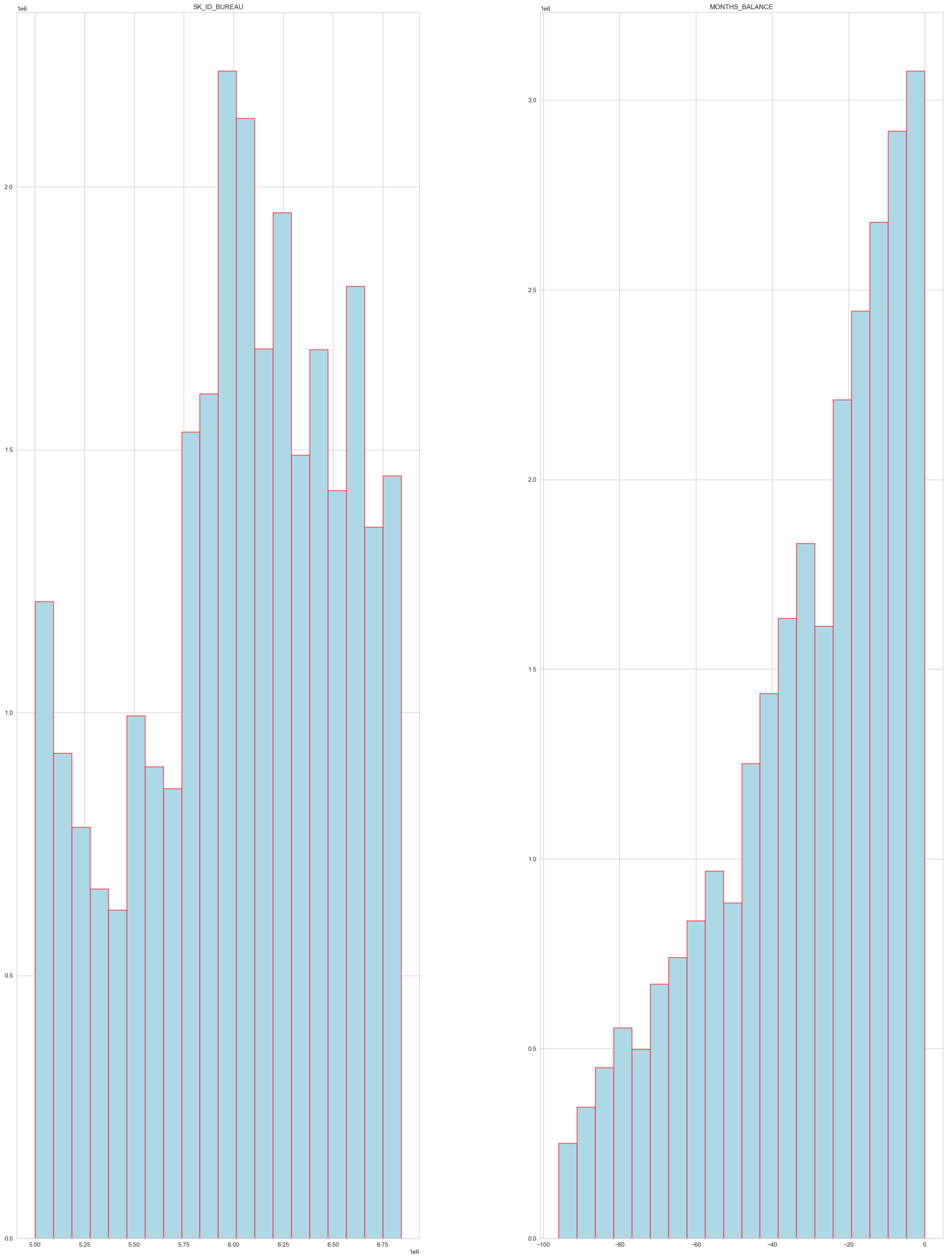
`DATASET: bbalance`



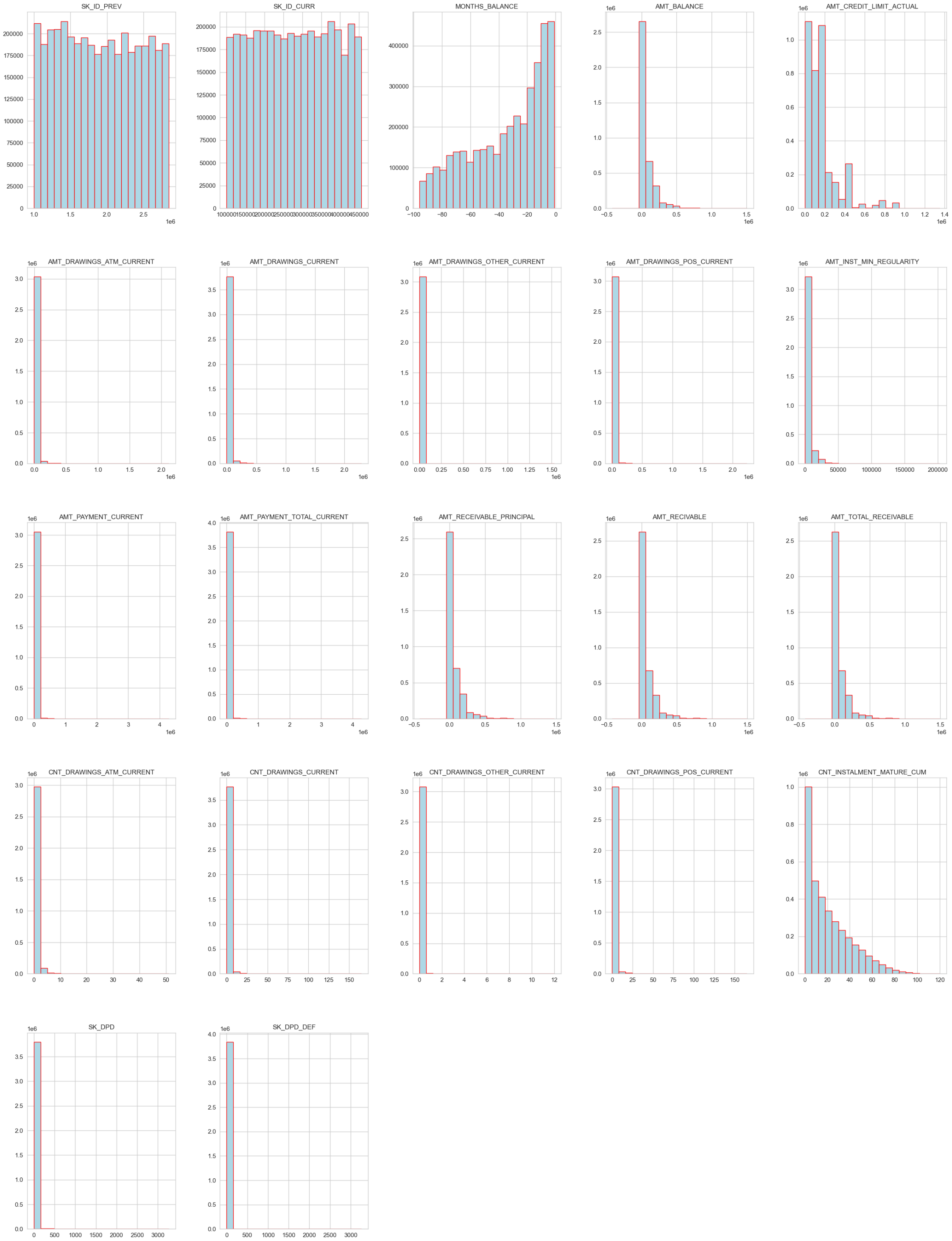
DATASET: pcbalance



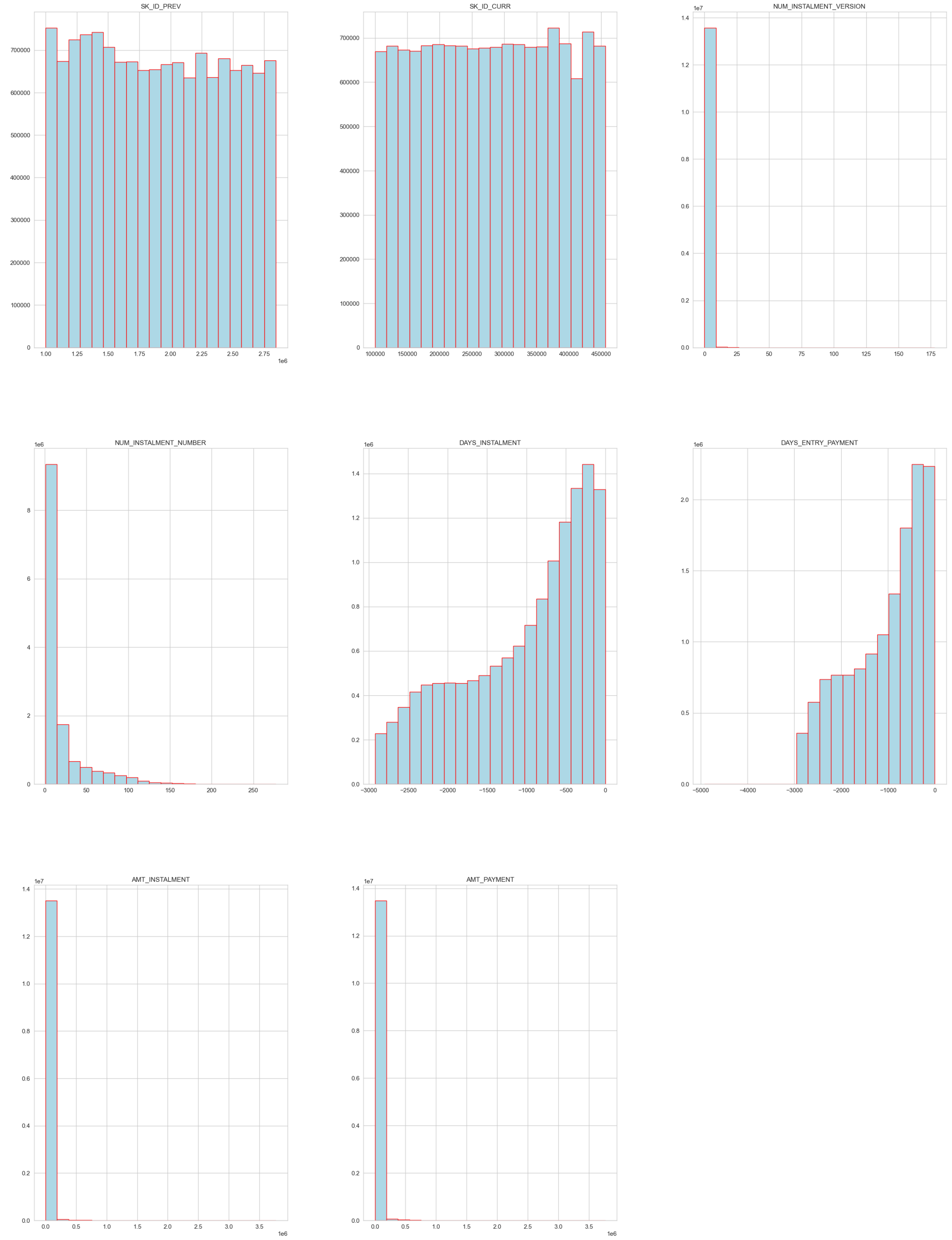
DATASET: bbalance



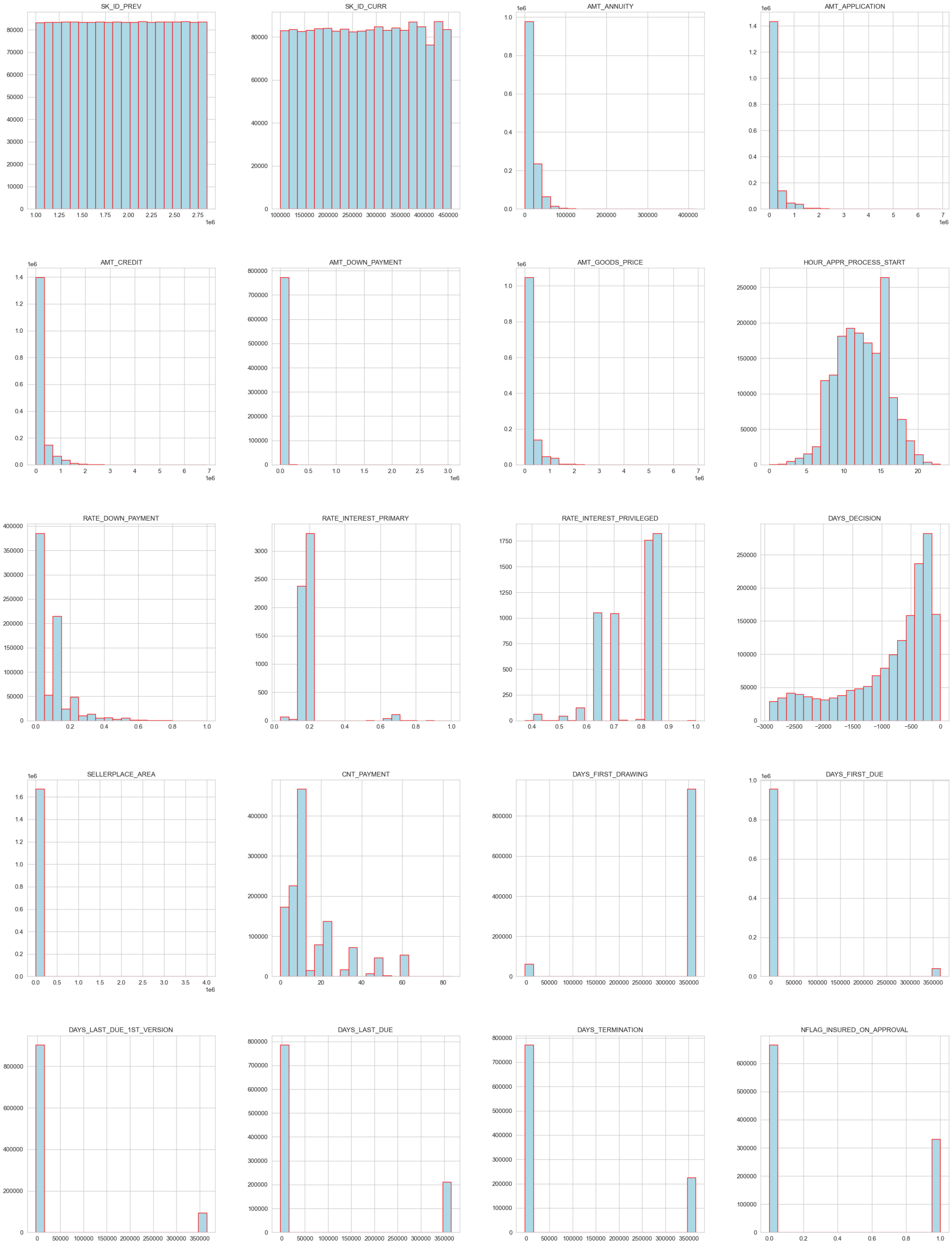
DATASET: ccbalance



DATASET: instpayments



`DATASET: prevapplication`



It can be observed that the majority of numerical variables are not distributed normally and there are high numbers of outliers in many variables.

Categorical and binary variables Functions to count values of all categorical and binary variables were created and looped over the elements of the list of dataframes.

```
DATASET: aptrain

CATEGORICAL:
Cash loans         278232
Revolving loans     29279
Name: NAME_CONTRACT_TYPE, dtype: int64

F      202448
M      105059
XNA         4
Name: CODE_GENDER, dtype: int64

N    202924
Y    104587
Name: FLAG_OWN_CAR, dtype: int64

Y    213312
N     94199
Name: FLAG_OWN_REALTY, dtype: int64

Unaccompanied      248526
Family              40149
Spouse, partner     11370
Children             3267
Other_B              1770
Other_A               866
Group of people       271
Name: NAME_TYPE_SUITE, dtype: int64

Working                 158774
Commercial associate     71617
Pensioner                55362
State servant            21703
Unemployed                  22
Student                     18
Businessman                 10
Maternity leave              5
Name: NAME_INCOME_TYPE, dtype: int64

Secondary / secondary special    218391
Higher education                  74863
Incomplete higher                 10277
Lower secondary                    3816
Academic degree                     164
Name: NAME_EDUCATION_TYPE, dtype: int64
```

```
Married                 196432
Single / not married     45444
Civil marriage           29775
Separated                19770
Widow                    16088
Unknown                      2
Name: NAME_FAMILY_STATUS, dtype: int64

House / apartment      272868
With parents            14840
Municipal apartment     11183
Rented apartment         4881
Office apartment         2617
Co-op apartment          1122
Name: NAME_HOUSING_TYPE, dtype: int64

Laborers                 55186
Sales staff              32102
Core staff               27570
Managers                 21371
Drivers                  18603
High skill tech staff    11380
Accountants               9813
Medicine staff            8537
Security staff            6721
Cooking staff             5946
Cleaning staff            4653
Private service staff     2652
Low-skill Laborers        2093
Waiters/barmen staff      1348
Secretaries               1305
Realty agents              751
HR staff                   563
IT staff                   526
Name: OCCUPATION_TYPE, dtype: int64

TUESDAY      53901
WEDNESDAY    51934
MONDAY       50714
THURSDAY     50591
FRIDAY       50338
SATURDAY     33852
SUNDAY       16181
Name: WEEKDAY_APPR_PROCESS_START, dtype: int64

Business Entity Type 3    67992
XNA                       55374
```

```
Self-employed             38412
Other                     16683
Medicine                  11193
Business Entity Type 2    10553
Government                10404
School                     8893
Trade: type 7              7831
Kindergarten               6880
Construction               6721
Business Entity Type 1     5984
Transport: type 4          5398
Trade: type 3              3492
Industry: type 9           3368
Industry: type 3           3278
Security                   3247
Housing                    2958
Industry: type 11          2704
Military                   2634
Bank                       2507
Agriculture                2454
Police                     2341
Transport: type 2          2204
Postal                     2157
Security Ministries        1974
Trade: type 2              1900
Restaurant                 1811
Services                   1575
University                 1327
Industry: type 7           1307
Transport: type 3          1187
Industry: type 1           1039
Hotel                       966
Electricity                 950
Industry: type 4            877
Trade: type 6               631
Industry: type 5            599
Insurance                   597
Telecom                     577
Emergency                   560
Industry: type 2            458
Advertising                 429
Realtor                     396
Culture                     379
Industry: type 12           369
Trade: type 1               348
Mobile                      317
Legal Services              305
Cleaning                    260
```

```
Transport: type 1           201
Industry: type 6            112
Industry: type 10           109
Religion                     85
Industry: type 13            67
Trade: type 4                64
Trade: type 5                49
Industry: type 8             24
Name: ORGANIZATION_TYPE, dtype: int64

reg oper account         73830
reg oper spec account    12080
not specified             5687
org spec account          5619
Name: FONDKAPREMONT_MODE, dtype: int64

block of flats      150503
specific housing      1499
terraced house        1212
Name: HOUSETYPE_MODE, dtype: int64

Panel           66040
Stone, brick    64815
Block            9253
Wooden           5362
Mixed            2296
Monolithic       1779
Others           1625
Name: WALLSMATERIAL_MODE, dtype: int64

No     159428
Yes      2328
Name: EMERGENCYSTATE_MODE, dtype: int64


BINARY NUMERIC:
0    282686
1     24825
Name: TARGET, dtype: int64

Cash loans         278232
Revolving loans     29279
Name: NAME_CONTRACT_TYPE, dtype: int64

N    202924
Y    104587
Name: FLAG_OWN_CAR, dtype: int64
```

```
Y    213312
N     94199
Name: FLAG_OWN_REALTY, dtype: int64

1    307510
0         1
Name: FLAG_MOBIL, dtype: int64

1    252125
0     55386
Name: FLAG_EMP_PHONE, dtype: int64

0    246203
1     61308
Name: FLAG_WORK_PHONE, dtype: int64

1    306937
0       574
Name: FLAG_CONT_MOBILE, dtype: int64

0    221080
1     86431
Name: FLAG_PHONE, dtype: int64

0    290069
1     17442
Name: FLAG_EMAIL, dtype: int64

0    302854
1      4657
Name: REG_REGION_NOT_LIVE_REGION, dtype: int64

0    291899
1     15612
Name: REG_REGION_NOT_WORK_REGION, dtype: int64

0    295008
1     12503
Name: LIVE_REGION_NOT_WORK_REGION, dtype: int64

0    283472
1     24039
Name: REG_CITY_NOT_LIVE_CITY, dtype: int64

0    236644
1     70867
Name: REG_CITY_NOT_WORK_CITY, dtype: int64
```

```
0    252296
1     55215
Name: LIVE_CITY_NOT_WORK_CITY, dtype: int64

0    307498
1        13
Name: FLAG_DOCUMENT_2, dtype: int64

1    218340
0     89171
Name: FLAG_DOCUMENT_3, dtype: int64

0    307486
1        25
Name: FLAG_DOCUMENT_4, dtype: int64

0    302863
1      4648
Name: FLAG_DOCUMENT_5, dtype: int64

0    280433
1     27078
Name: FLAG_DOCUMENT_6, dtype: int64

0    307452
1        59
Name: FLAG_DOCUMENT_7, dtype: int64

0    282487
1     25024
Name: FLAG_DOCUMENT_8, dtype: int64

0    306313
1      1198
Name: FLAG_DOCUMENT_9, dtype: int64

0    307504
1         7
Name: FLAG_DOCUMENT_10, dtype: int64

0    306308
1      1203
Name: FLAG_DOCUMENT_11, dtype: int64

0    307509
1         2
Name: FLAG_DOCUMENT_12, dtype: int64
```

```
0    306427
1      1084
Name: FLAG_DOCUMENT_13, dtype: int64

0    306608
1       903
Name: FLAG_DOCUMENT_14, dtype: int64

0    307139
1       372
Name: FLAG_DOCUMENT_15, dtype: int64

0    304458
1      3053
Name: FLAG_DOCUMENT_16, dtype: int64

0    307429
1        82
Name: FLAG_DOCUMENT_17, dtype: int64

0    305011
1      2500
Name: FLAG_DOCUMENT_18, dtype: int64

0    307328
1       183
Name: FLAG_DOCUMENT_19, dtype: int64

0    307355
1       156
Name: FLAG_DOCUMENT_20, dtype: int64

0    307408
1       103
Name: FLAG_DOCUMENT_21, dtype: int64


DATASET: aptest

CATEGORICAL:
Cash loans         48305
Revolving loans      439
Name: NAME_CONTRACT_TYPE, dtype: int64

F    32678
M    16066
Name: CODE_GENDER, dtype: int64
```

```
N    32311
Y    16433
Name: FLAG_OWN_CAR, dtype: int64

Y    33658
N    15086
Name: FLAG_OWN_REALTY, dtype: int64

Unaccompanied      39727
Family              5881
Spouse, partner     1448
Children             408
Other_B              211
Other_A              109
Group of people       49
Name: NAME_TYPE_SUITE, dtype: int64

Working                 24533
Commercial associate    11402
Pensioner                9273
State servant            3532
Student                     2
Businessman                 1
Unemployed                  1
Name: NAME_INCOME_TYPE, dtype: int64

Secondary / secondary special    33988
Higher education                 12516
Incomplete higher                 1724
Lower secondary                    475
Academic degree                     41
Name: NAME_EDUCATION_TYPE, dtype: int64

Married                 32283
Single / not married     7036
Civil marriage           4261
Separated                2955
Widow                    2209
Name: NAME_FAMILY_STATUS, dtype: int64

House / apartment      43645
With parents            2234
Municipal apartment     1617
Rented apartment         718
Office apartment         407
Co-op apartment          123
Name: NAME_HOUSING_TYPE, dtype: int64
```

```
Laborers                 8655
Sales staff              5072
Core staff               4361
Managers                 3574
Drivers                  2773
High skill tech staff    1854
Accountants              1628
Medicine staff           1316
Security staff            915
Cooking staff             894
Cleaning staff            656
Private service staff     455
Low-skill Laborers        272
Secretaries               213
Waiters/barmen staff      178
Realty agents             138
HR staff                  104
IT staff                   81
Name: OCCUPATION_TYPE, dtype: int64

TUESDAY      9751
WEDNESDAY    8457
THURSDAY     8418
MONDAY       8406
FRIDAY       7250
SATURDAY     4603
SUNDAY       1859
Name: WEEKDAY_APPR_PROCESS_START, dtype: int64

Business Entity Type 3    10840
XNA                        9274
Self-employed              5920
Other                      2707
Medicine                   1716
Government                 1508
Business Entity Type 2     1479
Trade: type 7              1303
School                     1287
Construction               1039
Kindergarten               1038
Business Entity Type 1      887
Transport: type 4           884
Trade: type 3               578
Military                    530
Industry: type 9            499
Industry: type 3            489
Security                    472
Transport: type 2           448
```

```
Police                      441
Housing                     435
Industry: type 11           416
Bank                        374
Security Ministries         341
Services                    302
Postal                      294
Agriculture                 292
Restaurant                  284
Trade: type 2               242
University                  221
Industry: type 7            217
Industry: type 1            178
Transport: type 3           174
Industry: type 4            167
Electricity                 156
Hotel                       134
Trade: type 6               122
Industry: type 5             97
Telecom                      95
Emergency                    91
Insurance                    80
Industry: type 12            77
Industry: type 2             77
Realtor                      72
Advertising                  71
Trade: type 1                64
Culture                      61
Legal Services               53
Mobile                       45
Cleaning                     43
Transport: type 1            35
Industry: type 6             27
Industry: type 10            24
Trade: type 4                14
Religion                     12
Trade: type 5                 9
Industry: type 13             6
Industry: type 8              3
Name: ORGANIZATION_TYPE, dtype: int64

reg oper account         12124
reg oper spec account     1990
org spec account           920
not specified              913
Name: FONDKAPREMONT_MODE, dtype: int64

block of flats      24659
```

```
specific housing      262
terraced house        204
Name: HOUSETYPE_MODE, dtype: int64

Panel           11269
Stone, brick    10434
Block            1428
Wooden            794
Mixed             353
Monolithic        289
Others            284
Name: WALLSMATERIAL_MODE, dtype: int64

No     26179
Yes      356
Name: EMERGENCYSTATE_MODE, dtype: int64


BINARY NUMERIC:
Cash loans         48305
Revolving loans      439
Name: NAME_CONTRACT_TYPE, dtype: int64

F    32678
M    16066
Name: CODE_GENDER, dtype: int64

N    32311
Y    16433
Name: FLAG_OWN_CAR, dtype: int64

Y    33658
N    15086
Name: FLAG_OWN_REALTY, dtype: int64

1    48743
0        1
Name: FLAG_MOBIL, dtype: int64

1    39469
0     9275
Name: FLAG_EMP_PHONE, dtype: int64

0    38766
1     9978
Name: FLAG_WORK_PHONE, dtype: int64

1    48666
```

```
0       78
Name: FLAG_CONT_MOBILE, dtype: int64

0    35918
1    12826
Name: FLAG_PHONE, dtype: int64

0    40816
1     7928
Name: FLAG_EMAIL, dtype: int64

0    47826
1      918
Name: REG_REGION_NOT_LIVE_REGION, dtype: int64

0    46055
1     2689
Name: REG_REGION_NOT_WORK_REGION, dtype: int64

0    46695
1     2049
Name: LIVE_REGION_NOT_WORK_REGION, dtype: int64

0    44968
1     3776
Name: REG_CITY_NOT_LIVE_CITY, dtype: int64

0    37793
1    10951
Name: REG_CITY_NOT_WORK_CITY, dtype: int64

0    40252
1     8492
Name: LIVE_CITY_NOT_WORK_CITY, dtype: int64

0    48744
Name: FLAG_DOCUMENT_2, dtype: int64

1    38343
0    10401
Name: FLAG_DOCUMENT_3, dtype: int64

0    48739
1        5
Name: FLAG_DOCUMENT_4, dtype: int64

0    48025
1      719
```

```
Name: FLAG_DOCUMENT_5, dtype: int64

0    44480
1     4264
Name: FLAG_DOCUMENT_6, dtype: int64

0    48742
1        2
Name: FLAG_DOCUMENT_7, dtype: int64

0    44432
1     4312
Name: FLAG_DOCUMENT_8, dtype: int64

0    48525
1      219
Name: FLAG_DOCUMENT_9, dtype: int64

0    48744
Name: FLAG_DOCUMENT_10, dtype: int64

0    48687
1       57
Name: FLAG_DOCUMENT_11, dtype: int64

0    48744
Name: FLAG_DOCUMENT_12, dtype: int64

0    48744
Name: FLAG_DOCUMENT_13, dtype: int64

0    48744
Name: FLAG_DOCUMENT_14, dtype: int64

0    48744
Name: FLAG_DOCUMENT_15, dtype: int64

0    48744
Name: FLAG_DOCUMENT_16, dtype: int64

0    48744
Name: FLAG_DOCUMENT_17, dtype: int64

0    48668
1       76
Name: FLAG_DOCUMENT_18, dtype: int64

0    48744
```

```
Name: FLAG_DOCUMENT_19, dtype: int64

0    48744
Name: FLAG_DOCUMENT_20, dtype: int64

0    48744
Name: FLAG_DOCUMENT_21, dtype: int64


DATASET: bureau

CATEGORICAL:
Closed      1079273
Active       630607
Sold           6527
Bad debt         21
Name: CREDIT_ACTIVE, dtype: int64

currency 1    1715020
currency 2       1224
currency 3        174
currency 4         10
Name: CREDIT_CURRENCY, dtype: int64

Consumer credit                                 1251615
Credit card                                      402195
Car loan                                          27690
Mortgage                                          18391
Microloan                                         12413
Loan for business development                      1975
Another type of loan                               1017
Unknown type of loan                                555
Loan for working capital replenishment              469
Cash loan (non-earmarked)                            56
Real estate loan                                     27
Loan for the purchase of equipment                   19
Loan for purchase of shares (margin lending)          4
Mobile operator loan                                  1
Interbank credit                                      1
Name: CREDIT_TYPE, dtype: int64


BINARY NUMERIC:
There are no binary numeric variables in the dataset.

DATASET: bbalance

CATEGORICAL:
```

```
C    13646993
0     7499507
X     5810482
1      242347
5       62406
2       23419
3        8924
4        5847
Name: STATUS, dtype: int64


BINARY NUMERIC:
There are no binary numeric variables in the dataset.

DATASET: pcbalance

CATEGORICAL:
Active                   9151119
Completed                 744883
Signed                     87260
Demand                      7065
Returned to the store       5461
Approved                    4917
Amortized debt               636
Canceled                      15
XNA                            2
Name: NAME_CONTRACT_STATUS, dtype: int64


BINARY NUMERIC:
There are no binary numeric variables in the dataset.

DATASET: bbalance

CATEGORICAL:
C    13646993
0     7499507
X     5810482
1      242347
5       62406
2       23419
3        8924
4        5847
Name: STATUS, dtype: int64


BINARY NUMERIC:
There are no binary numeric variables in the dataset.
```

```
DATASET: ccbalance

CATEGORICAL:
Active           3698436
Completed         128918
Signed             11058
Demand              1365
Sent proposal        513
Refused               17
Approved               5
Name: NAME_CONTRACT_STATUS, dtype: int64


BINARY NUMERIC:
There are no binary numeric variables in the dataset.

DATASET: instpayments

CATEGORICAL:

BINARY NUMERIC:
There are no binary numeric variables in the dataset.

DATASET: prevapplication

CATEGORICAL:
Cash loans         747553
Consumer loans     729151
Revolving loans    193164
XNA                   346
Name: NAME_CONTRACT_TYPE, dtype: int64

TUESDAY      255118
WEDNESDAY    255010
MONDAY       253557
FRIDAY       252048
THURSDAY     249099
SATURDAY     240631
SUNDAY       164751
Name: WEEKDAY_APPR_PROCESS_START, dtype: int64

Y    1661739
N       8475
Name: FLAG_LAST_APPL_PER_CONTRACT, dtype: int64

XAP                                 922661
XNA                                 677918
```

```
Repairs                              23765
Other                                15608
Urgent needs                          8412
Buying a used car                     2888
Building a house or an annex          2693
Everyday expenses                     2416
Medicine                              2174
Payments on other loans               1931
Education                             1573
Journey                               1239
Purchase of electronic equipment      1061
Buying a new car                      1012
Wedding / gift / holiday               962
Buying a home                          865
Car repairs                            797
Furniture                              749
Buying a holiday home / land           533
Business development                   426
Gasification / water supply            300
Buying a garage                        136
Hobby                                   55
Money for a third person                25
Refusal to name the goal                15
Name: NAME_CASH_LOAN_PURPOSE, dtype: int64

Approved        1036781
Canceled         316319
Refused          290678
Unused offer      26436
Name: NAME_CONTRACT_STATUS, dtype: int64

Cash through the bank                        1033552
XNA                                           627384
Non-cash from your account                      8193
Cashless from the account of the employer       1085
Name: NAME_PAYMENT_TYPE, dtype: int64

XAP       1353093
HC         175231
LIMIT       55680
SCO         37467
CLIENT      26436
SCOFR       12811
XNA          5244
VERIF        3535
SYSTEM        717
Name: CODE_REJECT_REASON, dtype: int64
```

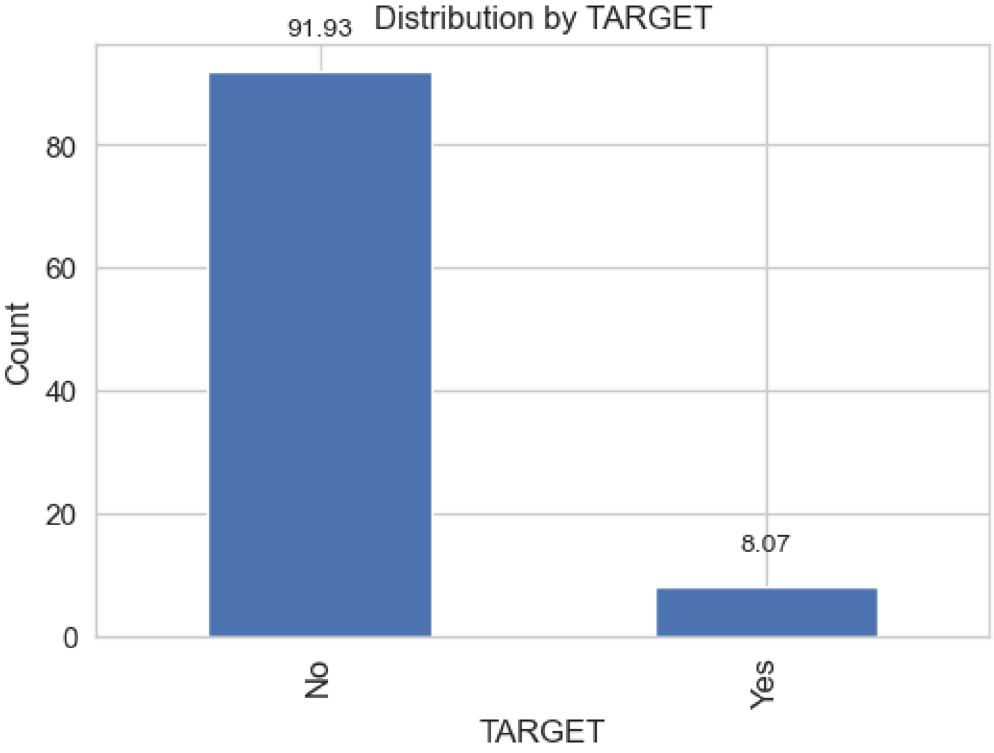
```
Unaccompanied      508970
Family             213263
Spouse, partner     67069
Children            31566
Other_B             17624
Other_A              9077
Group of people      2240
Name: NAME_TYPE_SUITE, dtype: int64

Repeater     1231261
New           301363
Refreshed     135649
XNA             1941
Name: NAME_CLIENT_TYPE, dtype: int64

XNA                         950809
Mobile                      224708
Consumer Electronics        121576
Computers                   105769
Audio/Video                  99441
Furniture                    53656
Photo / Cinema Equipment     25021
Construction Materials       24995
Clothing and Accessories     23554
Auto Accessories              7381
Jewelry                       6290
Homewares                     5023
Medical Supplies              3843
Vehicles                      3370
Sport and Leisure             2981
Gardening                     2668
Other                         2554
Office Appliances             2333
Tourism                       1659
Medicine                      1550
Direct Sales                   446
Fitness                        209
Additional Service             128
Education                      107
Weapon                          77
Insurance                       64
Animals                          1
House Construction               1
Name: NAME_GOODS_CATEGORY, dtype: int64

POS      691011
Cash     461563
XNA      372230
```

```
Cards    144985
Cars        425
Name: NAME_PORTFOLIO, dtype: int64

XNA        1063666
x-sell      456287
walk-in     150261
Name: NAME_PRODUCT_TYPE, dtype: int64

Credit and cash offices       719968
Country-wide                  494690
Stone                         212083
Regional / Local              108528
Contact center                 71297
AP+ (Cash loan)                57046
Channel of corporate sales      6150
Car dealer                       452
Name: CHANNEL_TYPE, dtype: int64

XNA                     855720
Consumer electronics    398265
Connectivity            276029
Furniture                57849
Construction             29781
Clothing                 23949
Industry                 19194
Auto technology           4990
Jewelry                   2709
MLM partners              1215
Tourism                    513
Name: NAME_SELLER_INDUSTRY, dtype: int64

XNA           517215
middle        385532
high          353331
low_normal    322095
low_action     92041
Name: NAME_YIELD_GROUP, dtype: int64

Cash                              285990
POS household with interest       263622
POS mobile with interest          220670
Cash X-Sell: middle               143883
Cash X-Sell: low                  130248
Card Street                       112582
POS industry with interest         98833
POS household without interest     82908
Card X-Sell                        80582
```

```
Cash Street: high                  59639
Cash X-Sell: high                  59301
Cash Street: middle                34658
Cash Street: low                   33834
POS mobile without interest        24082
POS other with interest            23879
POS industry without interest      12602
POS others without interest         2555
Name: PRODUCT_COMBINATION, dtype: int64


BINARY NUMERIC:
Y    1661739
N       8475
Name: FLAG_LAST_APPL_PER_CONTRACT, dtype: int64

1    1664314
0       5900
Name: NFLAG_LAST_APPL_IN_DAY, dtype: int64


DATASET: sampsubmission

CATEGORICAL:

BINARY NUMERIC:
0.5  48744
Name: TARGET, dtype: int64
```

It can be observed that quite high numbers of values in several different features are named XNA, XAP, XC, however, the meaning of these acronyms is unclear (explanations are not provided in the data dictionnary). Thus, these values could ve considered as missing.

'TARGET' as a target variable The 'TARGET' variable in the aptrain dataframe will be treated as the target variable for machine learning purposes, thus, the distribution of its values is presented here. It can be observed that this variable is heavily imbalanced - there are only 8 percent cases with value 1 (indicating that the person has payment difficulties).



Next, each of the dataframes and their variables will be examined separately.

APTRAIN AND APTEST: These dataframes contain 121 columns (aptest; the target variable 'TARGET' missing) and 122 columns (aptrain; the 'TARGET' is present). The dataframes include data on the HomeCredit club clients - their demographical characteristics (age, education, gender, family status), social conditions (income, living conditions, car, etc.), and the loans of the clients - contract type, annuity, credit amount, goods price amount, etc. Some features such as documents provided by clients or the day of the week and hour of the day when a client applied for a loan do not seem meaningful with regard to the effect on the clients' capability to pay the loan in time. These features will later be removed in one or another stage of feature engeneering.

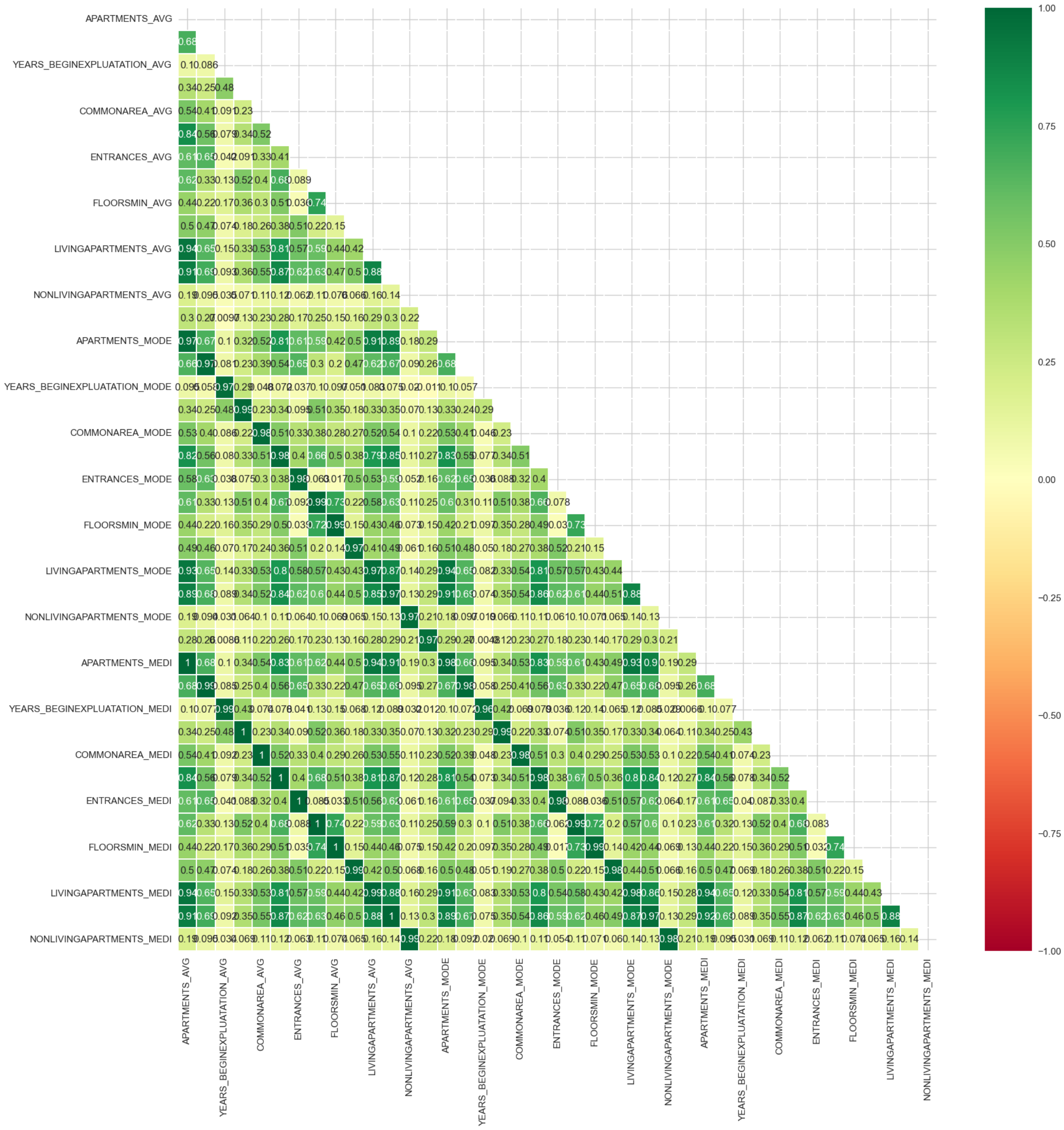
Bellow is the full list of features in two forms: with quation marks (for the use in pandas functions) and without quation marks (for the use in sql queries).

```
"SK_ID_CURR", "TARGET", "NAME_CONTRACT_TYPE", "CODE_GENDER", "FLAG_OWN_CAR",
"FLAG_OWN_REALTY", "CNT_CHILDREN", "AMT_INCOME_TOTAL", "AMT_CREDIT",
"AMT_ANNUITY", "AMT_GOODS_PRICE", "NAME_TYPE_SUITE", "NAME_INCOME_TYPE",
"NAME_EDUCATION_TYPE", "NAME_FAMILY_STATUS", "NAME_HOUSING_TYPE",
"REGION_POPULATION_RELATIVE", "DAYS_BIRTH", "DAYS_EMPLOYED",
"DAYS_REGISTRATION", "DAYS_ID_PUBLISH", "OWN_CAR_AGE", "FLAG_MOBIL",
"FLAG_EMP_PHONE", "FLAG_WORK_PHONE", "FLAG_CONT_MOBILE", "FLAG_PHONE",
"FLAG_EMAIL", "OCCUPATION_TYPE", "CNT_FAM_MEMBERS", "REGION_RATING_CLIENT",
"REGION_RATING_CLIENT_W_CITY", "WEEKDAY_APPR_PROCESS_START",
"HOUR_APPR_PROCESS_START", "REG_REGION_NOT_LIVE_REGION",
"REG_REGION_NOT_WORK_REGION", "LIVE_REGION_NOT_WORK_REGION",
"REG_CITY_NOT_LIVE_CITY", "REG_CITY_NOT_WORK_CITY", "LIVE_CITY_NOT_WORK_CITY",
"ORGANIZATION_TYPE", "EXT_SOURCE_1", "EXT_SOURCE_2", "EXT_SOURCE_3",
"APARTMENTS_AVG", "BASEMENTAREA_AVG", "YEARS_BEGINEXPLUATATION_AVG",
"YEARS_BUILD_AVG", "COMMONAREA_AVG", "ELEVATORS_AVG", "ENTRANCES_AVG",
"FLOORSMAX_AVG", "FLOORSMIN_AVG", "LANDAREA_AVG", "LIVINGAPARTMENTS_AVG",
"LIVINGAREA_AVG", "NONLIVINGAPARTMENTS_AVG", "NONLIVINGAREA_AVG",
"APARTMENTS_MODE", "BASEMENTAREA_MODE", "YEARS_BEGINEXPLUATATION_MODE",
"YEARS_BUILD_MODE", "COMMONAREA_MODE", "ELEVATORS_MODE", "ENTRANCES_MODE",
"FLOORSMAX_MODE", "FLOORSMIN_MODE", "LANDAREA_MODE", "LIVINGAPARTMENTS_MODE",
"LIVINGAREA_MODE", "NONLIVINGAPARTMENTS_MODE", "NONLIVINGAREA_MODE",
"APARTMENTS_MEDI", "BASEMENTAREA_MEDI", "YEARS_BEGINEXPLUATATION_MEDI",
"YEARS_BUILD_MEDI", "COMMONAREA_MEDI", "ELEVATORS_MEDI", "ENTRANCES_MEDI",
"FLOORSMAX_MEDI", "FLOORSMIN_MEDI", "LANDAREA_MEDI", "LIVINGAPARTMENTS_MEDI",
"LIVINGAREA_MEDI", "NONLIVINGAPARTMENTS_MEDI", "NONLIVINGAREA_MEDI",
"FONDKAPREMONT_MODE", "HOUSETYPE_MODE", "TOTALAREA_MODE", "WALLSMATERIAL_MODE",
"EMERGENCYSTATE_MODE", "OBS_30_CNT_SOCIAL_CIRCLE", "DEF_30_CNT_SOCIAL_CIRCLE",
"OBS_60_CNT_SOCIAL_CIRCLE", "DEF_60_CNT_SOCIAL_CIRCLE",
"DAYS_LAST_PHONE_CHANGE", "FLAG_DOCUMENT_2", "FLAG_DOCUMENT_3",
"FLAG_DOCUMENT_4", "FLAG_DOCUMENT_5", "FLAG_DOCUMENT_6", "FLAG_DOCUMENT_7",
"FLAG_DOCUMENT_8", "FLAG_DOCUMENT_9", "FLAG_DOCUMENT_10", "FLAG_DOCUMENT_11",
"FLAG_DOCUMENT_12", "FLAG_DOCUMENT_13", "FLAG_DOCUMENT_14", "FLAG_DOCUMENT_15",
"FLAG_DOCUMENT_16", "FLAG_DOCUMENT_17", "FLAG_DOCUMENT_18", "FLAG_DOCUMENT_19",
"FLAG_DOCUMENT_20", "FLAG_DOCUMENT_21", "AMT_REQ_CREDIT_BUREAU_HOUR",
"AMT_REQ_CREDIT_BUREAU_DAY", "AMT_REQ_CREDIT_BUREAU_WEEK",
"AMT_REQ_CREDIT_BUREAU_MON", "AMT_REQ_CREDIT_BUREAU_QRT",
"AMT_REQ_CREDIT_BUREAU_YEAR"
```

```
SK_ID_CURR, TARGET, NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR,
FLAG_OWN_REALTY, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY,
AMT_GOODS_PRICE, NAME_TYPE_SUITE, NAME_INCOME_TYPE, NAME_EDUCATION_TYPE,
NAME_FAMILY_STATUS, NAME_HOUSING_TYPE, REGION_POPULATION_RELATIVE, DAYS_BIRTH,
DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, OWN_CAR_AGE, FLAG_MOBIL,
FLAG_EMP_PHONE, FLAG_WORK_PHONE, FLAG_CONT_MOBILE, FLAG_PHONE, FLAG_EMAIL,
OCCUPATION_TYPE, CNT_FAM_MEMBERS, REGION_RATING_CLIENT,
REGION_RATING_CLIENT_W_CITY, WEEKDAY_APPR_PROCESS_START,
HOUR_APPR_PROCESS_START, REG_REGION_NOT_LIVE_REGION, REG_REGION_NOT_WORK_REGION,
LIVE_REGION_NOT_WORK_REGION, REG_CITY_NOT_LIVE_CITY, REG_CITY_NOT_WORK_CITY,
LIVE_CITY_NOT_WORK_CITY, ORGANIZATION_TYPE, EXT_SOURCE_1, EXT_SOURCE_2,
EXT_SOURCE_3, APARTMENTS_AVG, BASEMENTAREA_AVG, YEARS_BEGINEXPLUATATION_AVG,
YEARS_BUILD_AVG, COMMONAREA_AVG, ELEVATORS_AVG, ENTRANCES_AVG, FLOORSMAX_AVG,
FLOORSMIN_AVG, LANDAREA_AVG, LIVINGAPARTMENTS_AVG, LIVINGAREA_AVG,
NONLIVINGAPARTMENTS_AVG, NONLIVINGAREA_AVG, APARTMENTS_MODE, BASEMENTAREA_MODE,
YEARS_BEGINEXPLUATATION_MODE, YEARS_BUILD_MODE, COMMONAREA_MODE, ELEVATORS_MODE,
ENTRANCES_MODE, FLOORSMAX_MODE, FLOORSMIN_MODE, LANDAREA_MODE,
LIVINGAPARTMENTS_MODE, LIVINGAREA_MODE, NONLIVINGAPARTMENTS_MODE,
NONLIVINGAREA_MODE, APARTMENTS_MEDI, BASEMENTAREA_MEDI,
YEARS_BEGINEXPLUATATION_MEDI, YEARS_BUILD_MEDI, COMMONAREA_MEDI, ELEVATORS_MEDI,
```

ENTRANCES_MEDI, FLOORSMAX_MEDI, FLOORSMIN_MEDI, LANDAREA_MEDI, LIVINGAPARTMENTS_MEDI, LIVINGAREA_MEDI, NONLIVINGAPARTMENTS_MEDI, NONLIVINGAREA_MEDI, FONDKAPREMONT_MODE, HOUSETYPE_MODE, TOTALAREA_MODE, WALLSMATERIAL_MODE, EMERGENCYSTATE_MODE, OBS_30_CNT_SOCIAL_CIRCLE, DEF_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE, DEF_60_CNT_SOCIAL_CIRCLE, DAYS_LAST_PHONE_CHANGE, FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12, FLAG_DOCUMENT_13, FLAG_DOCUMENT_14, FLAG_DOCUMENT_15, FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21, AMT_REQ_CREDIT_BUREAU_HOUR, AMT_REQ_CREDIT_BUREAU_DAY, AMT_REQ_CREDIT_BUREAU_WEEK, AMT_REQ_CREDIT_BUREAU_MON, AMT_REQ_CREDIT_BUREAU_QRT, AMT_REQ_CREDIT_BUREAU_YEAR

living conditions features: Quite a high number of features in these datasets are related with living conditions of clients. Some variables present just different metrics of the same metrics (averages, medians, modes). From the correlation heatmap plot it can be observed that many of these features are highly correlated between each other. Thus, it is reasonable to reduce a number of features by applying a dimension reduction procedure.

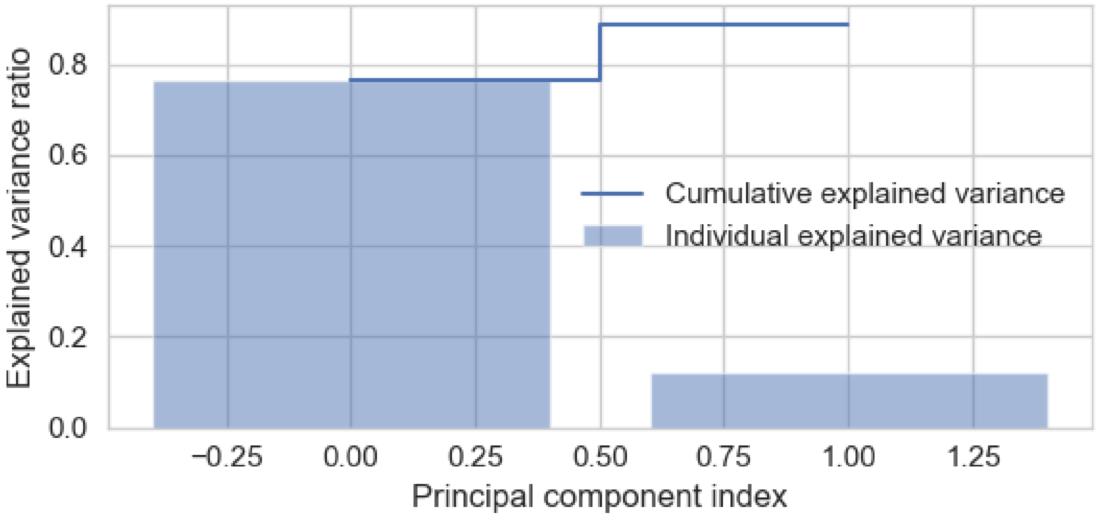


PCA for aptrain living conditions features: For the purpose of dimension reduction, the function for principle componend analysis is created and run on the features presenting averages of the apartment or house characteristics (median and mode were excluded).

Two principle components were identified. One of them explain 76 percent variance, another one - 12 percent. Data from 9 variables is transformed into 2 variables ‘LIVING_CONDITIONS_1’ and “LIVING_CONDITIONS_2’.

```
Explained variance ratio: [0.76303748 0.12254684]

Singular values: [332.75066999 133.35124361]
```

```
Components: [[ 0.10776881  0.06972266  0.78256887  0.5128919   0.03089267
0.07726968
   0.12553963  0.20204174  0.1616118   0.05052997  0.07139918  0.10111174
   0.00607464  0.02272648]
 [ 0.02156127  0.03084564 -0.57650737  0.75210819  0.05449739  0.05401282
  -0.03988908  0.01501277  0.27850645  0.01757961  0.11893494  0.02660957
   0.01033411 -0.00412828]]
```



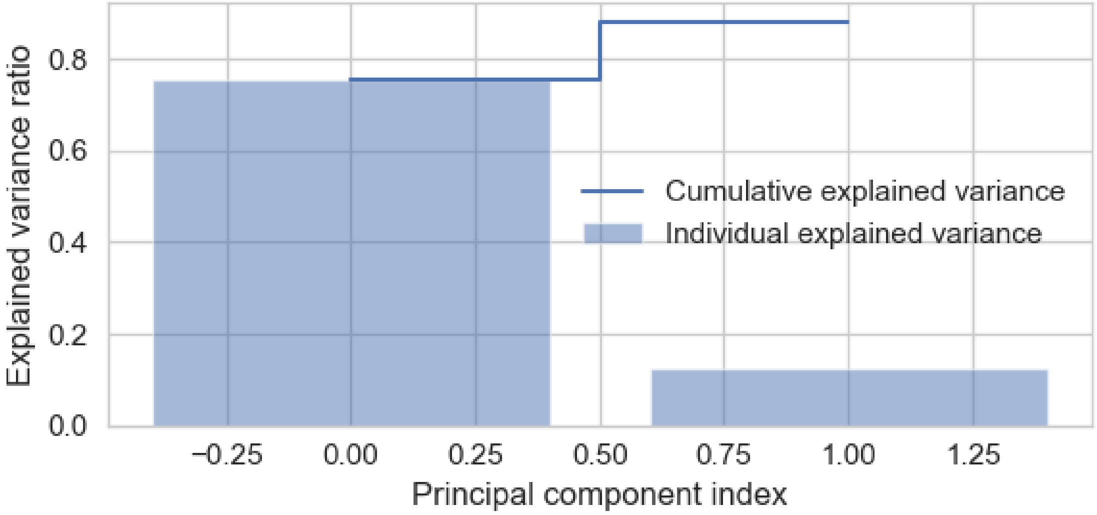
```
[400]:                                    PC1    PC2
        YEARS_BEGINEXPLUATATION_AVG  0.783 -0.577
        YEARS_BUILD_AVG              0.513  0.752
        FLOORSMAX_AVG                0.202  0.015
        FLOORSMIN_AVG                0.162  0.279
        ENTRANCES_AVG                0.126 -0.040
        APARTMENTS_AVG               0.108  0.022
        LIVINGAREA_AVG               0.101  0.027
        ELEVATORS_AVG                0.077  0.054
        LIVINGAPARTMENTS_AVG         0.071  0.119
        BASEMENTAREA_AVG             0.070  0.031
        LANDAREA_AVG                 0.051  0.018
        COMMONAREA_AVG               0.031  0.054
        NONLIVINGAREA_AVG            0.023 -0.004
        NONLIVINGAPARTMENTS_AVG      0.006  0.010
```

From the table above it can be observed that the first component is highly positively correlated with YEARS_BEGINEXPLUATATION_AVG, the second component - with 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

PCA for aptest living conditions features: The same transformations are done for the aptest data.

```
Explained variance ratio: [0.75345082 0.12475632]

Singular values: [133.01499148  54.12580998]

Components: [[ 0.11297998  0.0714217   0.77573999  0.51355282  0.03362968
0.08399889
   0.12703213  0.20931292  0.16781588  0.05114148  0.0758914   0.10651227
   0.00650168  0.02377694]
 [ 0.0254051   0.03065532 -0.58324207  0.74082172  0.05992641  0.06035285
  -0.04425825  0.01691045  0.28734686  0.01726803  0.12665822  0.03010373
   0.0112208  -0.00273198]]
```



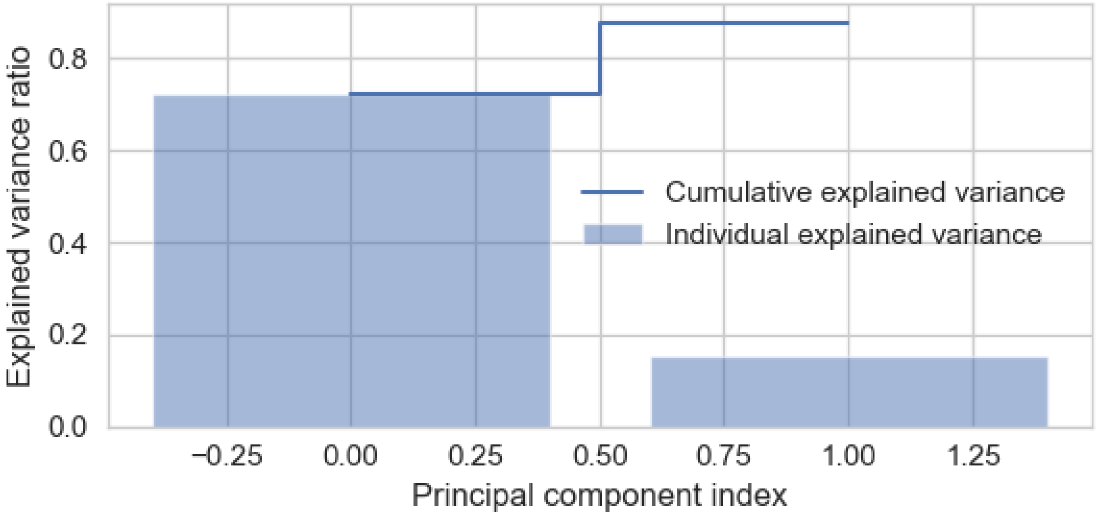
[402]:

```
                                PC1    PC2
YEARS_BEGINEXPLUATATION_AVG   0.776 -0.583
YEARS_BUILD_AVG               0.514  0.741
FLOORSMAX_AVG                 0.209  0.017
FLOORSMIN_AVG                 0.168  0.287
ENTRANCES_AVG                 0.127 -0.044
APARTMENTS_AVG                0.113  0.025
LIVINGAREA_AVG                0.107  0.030
ELEVATORS_AVG                 0.084  0.060
LIVINGAPARTMENTS_AVG          0.076  0.127
BASEMENTAREA_AVG              0.071  0.031
LANDAREA_AVG                  0.051  0.017
COMMONAREA_AVG                0.034  0.060
NONLIVINGAREA_AVG             0.024 -0.003
```

```
NONLIVINGAPARTMENTS_AVG      0.007  0.011
```

PCA for aptrain CB enquiries features: In the same way, the principal component analysis was applied to the variables on number of enquiries to the Credit Bureau.

```
Explained variance ratio: [0.72177081 0.1532178 ]

Singular values: [1030.54178595  474.8107654 ]

Components: [[ 2.32940681e-04  2.78130079e-04  4.03347266e-03  1.91425223e-02
   5.39843927e-02  9.98350063e-01]
 [ 2.51022169e-04 -3.73754815e-04 -2.13012090e-03  9.99809746e-01
   2.06496216e-03 -1.92735189e-02]]
```



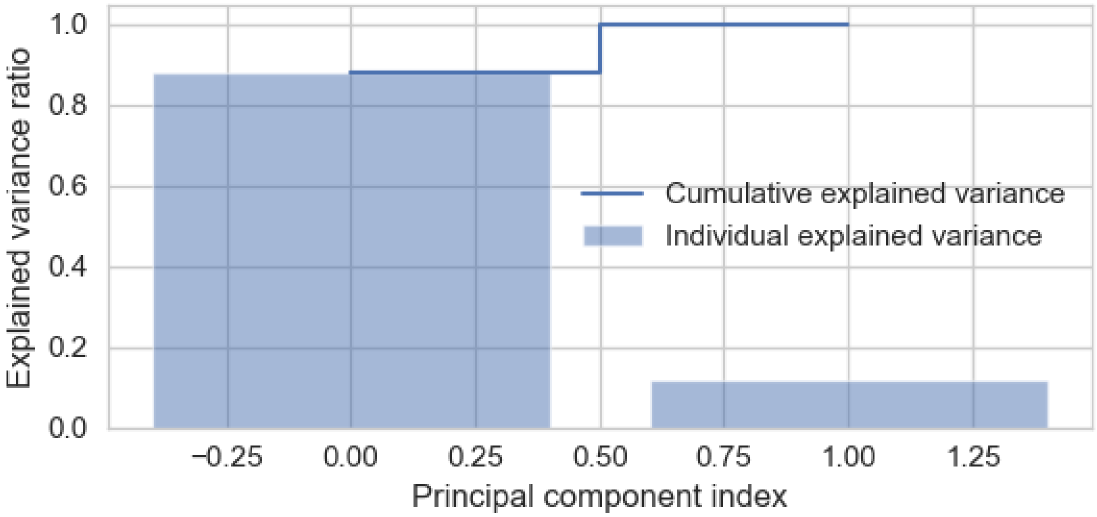
```
[405]:                              PC1    PC2
        AMT_REQ_CREDIT_BUREAU_YEAR  0.998 -0.019
        AMT_REQ_CREDIT_BUREAU_QRT   0.054  0.002
        AMT_REQ_CREDIT_BUREAU_MON   0.019  1.000
        AMT_REQ_CREDIT_BUREAU_WEEK  0.004 -0.002
        AMT_REQ_CREDIT_BUREAU_HOUR  0.000  0.000
        AMT_REQ_CREDIT_BUREAU_DAY   0.000 -0.000
```

Two principal components were identified. It can be observed that the first PC is highly positively correlated with the variable AMT_REQ_CREDIT_BUREAU_YEAR, the second component - with the AMT_REQ_CREDIT_BUREAU_MON.

PCA for aptest CB enquiries features:

```
Explained variance ratio: [0.87902704 0.11654932]
```

```
Singular values: [406.68930366 148.08685661]

Components: [[ 3.97178551e-04  1.61484153e-04  1.81058171e-04  1.98067966e-03
   3.51267400e-02  9.99380795e-01]
 [-1.84630986e-05  6.04206103e-04  2.15536790e-03  1.56061591e-03
   9.99379017e-01 -3.51302513e-02]]
```



```
[407]:                            PC1    PC2
       AMT_REQ_CREDIT_BUREAU_YEAR  0.999 -0.035
       AMT_REQ_CREDIT_BUREAU_QRT   0.035  0.999
       AMT_REQ_CREDIT_BUREAU_MON   0.002  0.002
       AMT_REQ_CREDIT_BUREAU_HOUR  0.000 -0.000
       AMT_REQ_CREDIT_BUREAU_DAY   0.000  0.001
       AMT_REQ_CREDIT_BUREAU_WEEK  0.000  0.002
```

for the aptest dataframe also two principal components were identified. It can be observed that the first PC is highly positively correlated with the variable AMT_REQ_CREDIT_BUREAU_YEAR, the second component - with the AMT_REQ_CREDIT_BUREAU_QRT.

gender and organization type: As it was noticed that gender and organization type variables contain value XNA which has to be removed, it was decided to encode the variables manually with pandas (rather than later in the machine learning pipeline). Also, categorical variable ORGANIZATION_TYPE has too many values, thus it was decided to do dimension reduction with PCA after one-hot encoding of this variable.

Gender variables were selected, removing variable gender_XNA.

days employed: It can be observed (see bellow) that there are some errors in the values of the variable 'DAYS_EMPLOYED'. These values were transformed to missing values.

```
[412]:  365243    55374
       -200         156
       -224         152
       -230         151
       -199         151
                  …
       -13961         1
       -11827         1
       -10176         1
       -9459          1
       -8694          1
       Name: DAYS_EMPLOYED, Length: 12574, dtype: int64
```

education: The categorical variable 'education' was transformed into an ordinal variable in the scale from 1 'lower secondary' to 5 'academic degree' which will be treated as a numerical variable (with an option to calculate the mean of education).

Recoded values of this variable: "1": Lower secondary "2": Secondary / secondary special "3": Incomplete higher "4": Higher education "5": Academic degree

car ownership: OWN_CAR_AGE and FLAG_OWN_CAR variables were transformed into a single variable CAR_OWN with values 0 indicating that a person does not own a car.

Getting modified aptrain and aptest datasets: Aptrain and aptest dataframes were modified by merging remaining variables of these dataframes with new modified variables and droping variables FLAG_DOCUMENT_2, FLAG_DOCUMENT_3, FLAG_DOCUMENT_4, FLAG_DOCUMENT_5, FLAG_DOCUMENT_6, FLAG_DOCUMENT_7, FLAG_DOCUMENT_8, FLAG_DOCUMENT_9, FLAG_DOCUMENT_10, FLAG_DOCUMENT_11, FLAG_DOCUMENT_12, FLAG_DOCUMENT_13, FLAG_DOCUMENT_14, FLAG_DOCUMENT_15, FLAG_DOCUMENT_16, FLAG_DOCUMENT_17, FLAG_DOCUMENT_18, FLAG_DOCUMENT_19, FLAG_DOCUMENT_20, FLAG_DOCUMENT_21, WEEKDAY_APPR_PROCESS_START, HOUR_APPR_PROCESS_START as they do not seem meaningful.

pcbalance: The dataframe pcbalance was examined. The dataframe contains data on monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit. The data has the time dimension (info about payment balance of previous loans of the client). It has 6 features as well as primary and foreign keys refering to a client and his or her previous loans.

```
[420]: Index(['SK_ID_PREV', 'SK_ID_CURR', 'MONTHS_BALANCE', 'CNT_INSTALMENT',
              'CNT_INSTALMENT_FUTURE', 'NAME_CONTRACT_STATUS', 'SK_DPD',
              'SK_DPD_DEF'],
             dtype='object')
```

A categorical variable NAME_CONTRACT_STATUS of this dataframe was transformed by on hot encoding this variable in pandas (creating dummy variables), grouping values of encoded variables by loan ids and getting means of these values. Numerical variables also were grouped by loan ids and their means calculated. New aggreraged variables indicate average monthly balance and contract status of of each loan.

```
[422]: Index(['SK_ID_PREV', 'SK_ID_CURR', 'MONTHS_BALANCE', 'CNT_INSTALMENT',
              'CNT_INSTALMENT_FUTURE', 'SK_DPD', 'SK_DPD_DEF',
              'NAME_CONTRACT_STATUS_Active', 'NAME_CONTRACT_STATUS_Amortized debt',
              'NAME_CONTRACT_STATUS_Approved', 'NAME_CONTRACT_STATUS_Canceled',
              'NAME_CONTRACT_STATUS_Completed', 'NAME_CONTRACT_STATUS_Demand',
              'NAME_CONTRACT_STATUS_Returned to the store',
              'NAME_CONTRACT_STATUS_Signed', 'NAME_CONTRACT_STATUS_XNA'],
             dtype='object')
```

```
SK_ID_PREV, SK_ID_CURR, MONTHS_BALANCE, CNT_INSTALMENT, CNT_INSTALMENT_FUTURE,
SK_DPD, SK_DPD_DEF, NAME_CONTRACT_STATUS_Active, NAME_CONTRACT_STATUS_Amortized
debt, NAME_CONTRACT_STATUS_Approved, NAME_CONTRACT_STATUS_Canceled,
NAME_CONTRACT_STATUS_Completed, NAME_CONTRACT_STATUS_Demand,
NAME_CONTRACT_STATUS_Returned to the store, NAME_CONTRACT_STATUS_Signed,
NAME_CONTRACT_STATUS_XNA
```

```
[426]:    SK_ID_PREV  SK_ID_CURR  WAVG_CNT_INSTALMENT  WAVG_CNT_INSTALMENT_FUTURE  \
       0     1002090      374073                 21.0                         1.5

          WAVG_SK_DPD  WAVG_SK_DPD_DEF
       0          0.0              0.0
```

Numerical and encoded categorical variables were concatenated.

```
[429]:             NAME_CONTRACT_STATUS_Active  NAME_CONTRACT_STATUS_Amortized_debt  \
       SK_ID_PREV
       1000001                        0.666667                                  0.0
       1000002                        0.800000                                  0.0
       1000003                        1.000000                                  0.0
       1000004                        0.875000                                  0.0
       1000005                        0.909091                                  0.0

                   NAME_CONTRACT_STATUS_Approved  NAME_CONTRACT_STATUS_Canceled  \
       SK_ID_PREV
       1000001                               0.0                            0.0
       1000002                               0.0                            0.0
       1000003                               0.0                            0.0
       1000004                               0.0                            0.0
       1000005                               0.0                            0.0

                   NAME_CONTRACT_STATUS_Completed  NAME_CONTRACT_STATUS_Demand  \
       SK_ID_PREV
       1000001                           0.333333                          0.0
       1000002                           0.200000                          0.0
       1000003                           0.000000                          0.0
       1000004                           0.125000                          0.0
```

```
1000005                           0.090909                          0.0

            NAME_CONTRACT_STATUS_Returned_to_the_store  \
SK_ID_PREV
1000001                                            0.0
1000002                                            0.0
1000003                                            0.0
1000004                                            0.0
1000005                                            0.0

            NAME_CONTRACT_STATUS_Signed  NAME_CONTRACT_STATUS_XNA  SK_ID_CURR  \
SK_ID_PREV
1000001                             0.0                       0.0      158271
1000002                             0.0                       0.0      101962
1000003                             0.0                       0.0      252457
1000004                             0.0                       0.0      260094
1000005                             0.0                       0.0      176456

            WAVG_CNT_INSTALMENT  WAVG_CNT_INSTALMENT_FUTURE  WAVG_SK_DPD  \
SK_ID_PREV
1000001                8.666667                    7.666667          0.0
1000002                     NaN                         NaN          NaN
1000003               12.000000                   10.500000          0.0
1000004                9.625000                    6.125000          0.0
1000005                     NaN                         NaN          NaN

            WAVG_SK_DPD_DEF
SK_ID_PREV
1000001                 0.0
1000002                 NaN
1000003                 0.0
1000004                 0.0
1000005                 NaN
```

Finally, averages of all loans for each client were calculated. A column with XNA was dropped as it was treated as a variable with only missing values.

[430]:

```
            AVG_NAME_CONTRACT_STATUS_Active  \
SK_ID_CURR
193774                                  1.0

            AVG_NAME_CONTRACT_STATUS_Amortized_debt  \
SK_ID_CURR
193774                                          0.0

            AVG_NAME_CONTRACT_STATUS_Approved  \
SK_ID_CURR
193774                                    0.0
```

```
            AVG_NAME_CONTRACT_STATUS_Canceled  \
SK_ID_CURR
193774                                    0.0

            AVG_NAME_CONTRACT_STATUS_Completed  \
SK_ID_CURR
193774                                     0.0

            AVG_NAME_CONTRACT_STATUS_Demand  \
SK_ID_CURR
193774                                  0.0

            AVG_NAME_CONTRACT_STATUS_Returned_to_the_store  \
SK_ID_CURR
193774                                                 0.0

            AVG_NAME_CONTRACT_STATUS_Signed  CNT_INSTALMENT_WAVG  \
SK_ID_CURR
193774                                  0.0                 10.0

            CNT_INSTALMENT_FUTURE_WAVG  SK_DPD_WAVG  SK_DPD_DEF_WAVG
SK_ID_CURR
193774                             5.5          0.0              0.0
```

ccbalance: Similar trasformations as for pcbalance dataframe were conducted for the ccbalance dataframe. This dataframe has data on credit card payment balance for previous periods.

[431]:

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

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

   AMT_DRAWINGS_OTHER_CURRENT  AMT_DRAWINGS_POS_CURRENT  \
0                         0.0                     877.5
1                         0.0                       0.0
2                         0.0                       0.0
3                         0.0                       0.0
```

```
4                         0.0                   11547.0

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

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

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

   SK_DPD  SK_DPD_DEF
0       0           0
1       0           0
2       0           0
3       0           0
4       0           0

[5 rows x 23 columns]
```

[432]:

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'MONTHS_BALANCE', 'AMT_BALANCE',
       'AMT_CREDIT_LIMIT_ACTUAL', 'AMT_DRAWINGS_ATM_CURRENT',
       'AMT_DRAWINGS_CURRENT', 'AMT_DRAWINGS_OTHER_CURRENT',
       'AMT_DRAWINGS_POS_CURRENT', 'AMT_INST_MIN_REGULARITY',
       'AMT_PAYMENT_CURRENT', 'AMT_PAYMENT_TOTAL_CURRENT',
       'AMT_RECEIVABLE_PRINCIPAL', 'AMT_RECIVABLE', 'AMT_TOTAL_RECEIVABLE',
       'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
       'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
       'CNT_INSTALMENT_MATURE_CUM', 'NAME_CONTRACT_STATUS', 'SK_DPD',
       'SK_DPD_DEF'],
      dtype='object')
```

[434]:

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'MONTHS_BALANCE', 'AMT_BALANCE',
       'AMT_CREDIT_LIMIT_ACTUAL', 'AMT_DRAWINGS_ATM_CURRENT',
       'AMT_DRAWINGS_CURRENT', 'AMT_DRAWINGS_OTHER_CURRENT',
       'AMT_DRAWINGS_POS_CURRENT', 'AMT_INST_MIN_REGULARITY',
       'AMT_PAYMENT_CURRENT', 'AMT_PAYMENT_TOTAL_CURRENT',
       'AMT_RECEIVABLE_PRINCIPAL', 'AMT_RECIVABLE', 'AMT_TOTAL_RECEIVABLE',
       'CNT_DRAWINGS_ATM_CURRENT', 'CNT_DRAWINGS_CURRENT',
       'CNT_DRAWINGS_OTHER_CURRENT', 'CNT_DRAWINGS_POS_CURRENT',
       'CNT_INSTALMENT_MATURE_CUM', 'SK_DPD', 'SK_DPD_DEF',
       'NAME_CONTRACT_STATUS_Active', 'NAME_CONTRACT_STATUS_Approved',
       'NAME_CONTRACT_STATUS_Completed', 'NAME_CONTRACT_STATUS_Demand',
       'NAME_CONTRACT_STATUS_Refused', 'NAME_CONTRACT_STATUS_Sent proposal',
       'NAME_CONTRACT_STATUS_Signed'],
      dtype='object')
```

```
SK_ID_PREV, SK_ID_CURR, MONTHS_BALANCE, AMT_BALANCE, AMT_CREDIT_LIMIT_ACTUAL,
AMT_DRAWINGS_ATM_CURRENT, AMT_DRAWINGS_CURRENT, AMT_DRAWINGS_OTHER_CURRENT,
AMT_DRAWINGS_POS_CURRENT, AMT_INST_MIN_REGULARITY, AMT_PAYMENT_CURRENT,
AMT_PAYMENT_TOTAL_CURRENT, AMT_RECEIVABLE_PRINCIPAL, AMT_RECIVABLE,
AMT_TOTAL_RECEIVABLE, CNT_DRAWINGS_ATM_CURRENT, CNT_DRAWINGS_CURRENT,
CNT_DRAWINGS_OTHER_CURRENT, CNT_DRAWINGS_POS_CURRENT, CNT_INSTALMENT_MATURE_CUM,
SK_DPD, SK_DPD_DEF, NAME_CONTRACT_STATUS_Active, NAME_CONTRACT_STATUS_Approved,
NAME_CONTRACT_STATUS_Completed, NAME_CONTRACT_STATUS_Demand,
NAME_CONTRACT_STATUS_Refused, NAME_CONTRACT_STATUS_Sent proposal,
NAME_CONTRACT_STATUS_Signed
```

[438]:

```
            SK_ID_CURR  WAVG_AMT_BALANCE  WAVG_AMT_CREDIT_LIMIT_ACTUAL  \
SK_ID_PREV
1830565         442368      227010.23426                 436568.877551

            WAVG_AMT_DRAWINGS_ATM_CURRENT  WAVG_AMT_DRAWINGS_CURRENT  \
SK_ID_PREV
1830565                       9734.693878               14236.058342

            WAVG_AMT_DRAWINGS_OTHER_CURRENT  WAVG_AMT_DRAWINGS_POS_CURRENT  \
SK_ID_PREV
1830565                          234.183673                    4267.180791

            WAVG_AMT_INST_MIN_REGULARITY  WAVG_AMT_PAYMENT_CURRENT  \
SK_ID_PREV
1830565                     11780.384694              18575.676046

            WAVG_AMT_PAYMENT_TOTAL_CURRENT  WAVG_AMT_RECEIVABLE_PRINCIPAL  \
SK_ID_PREV
1830565                       18575.676046                  220513.884337

            WAVG_AMT_RECIVABLE  WAVG_AMT_TOTAL_RECEIVABLE  \
SK_ID_PREV
1830565          227056.979158              227056.979158

            WAVG_CNT_DRAWINGS_ATM_CURRENT  WAVG_CNT_DRAWINGS_CURRENT  \
       SK_ID_PREV
       1830565                          0.397959                        0.5

                   WAVG_CNT_DRAWINGS_OTHER_CURRENT  WAVG_CNT_DRAWINGS_POS_CURRENT  \
       SK_ID_PREV
       1830565                            0.005102                       0.096939

                   WAVG_CNT_INSTALMENT_MATURE_CUM  WAVG_SK_DPD  WAVG_SK_DPD_DEF
       SK_ID_PREV
       1830565                          67.765306          0.0              0.0

[441]:             AVG_NAME_CONTRACT_STATUS_CC_Active  \
       SK_ID_CURR
       324478                                     1.0

                   AVG_NAME_CONTRACT_STATUS_CC_Approved  \
       SK_ID_CURR
       324478                                       0.0

                   AVG_NAME_CONTRACT_STATUS_CC_Completed  \
       SK_ID_CURR
       324478                                        0.0

                   AVG_NAME_CONTRACT_STATUS_CC_Demand  \
       SK_ID_CURR
       324478                                     0.0

                   AVG_NAME_CONTRACT_STATUS_CC_Refused  \
       SK_ID_CURR
       324478                                      0.0

                   AVG_NAME_CONTRACT_STATUS_CC_Sent_proposal  \
       SK_ID_CURR
       324478                                            0.0

                   AVG_NAME_CONTRACT_STATUS_CC_Signed  WWAVG_AMT_BALANCE  \
       SK_ID_CURR
       324478                                     0.0      542245.440938

                   WWAVG_AMT_CREDIT_LIMIT_ACTUAL  WWAVG_AMT_DRAWINGS_ATM_CURRENT  \
       SK_ID_CURR
       324478                           765000.0                        46968.75

                   …  WWAVG_AMT_RECEIVABLE_PRINCIPAL  WWAVG_AMT_RECIVABLE  \
       SK_ID_CURR  …
       324478      …                   527561.625937          541014.1875
```

```
            WWAVG_AMT_TOTAL_RECEIVABLE  WWAVG_CNT_DRAWINGS_ATM_CURRENT  \
SK_ID_CURR
324478                     541014.1875                          1.3125

            WWAVG_CNT_DRAWINGS_CURRENT  WWAVG_CNT_DRAWINGS_OTHER_CURRENT  \
SK_ID_CURR
324478                          4.0625                               0.0

            WWAVG_CNT_DRAWINGS_POS_CURRENT  WWAVG_CNT_INSTALMENT_MATURE_CUM  \
SK_ID_CURR
324478                                2.75                            4.875

            WWAVG_SK_DPD  WWAVG_SK_DPD_DEF
SK_ID_CURR
324478               0.0               0.0

[1 rows x 26 columns]
```

prevapplication: The prevapplication dataframe contains data on previous applications for Home Credit loans of clients who have loans. It has high numbers of numerical and categorical features. Categorical variables were one hot encoded and theirs means were calculated, grouped by client ids. Numerical variables were also grouped by client ids and their weighted means by the time variable DAYES_DECISION were calculated. Other time variables 'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION','DAYS_LAST_DUE', 'DAYS_TERMINATION' were removed.

```
Index(['SK_ID_PREV', 'SK_ID_CURR', 'NAME_CONTRACT_TYPE', 'AMT_ANNUITY',
       'AMT_APPLICATION', 'AMT_CREDIT', 'AMT_DOWN_PAYMENT', 'AMT_GOODS_PRICE',
       'WEEKDAY_APPR_PROCESS_START', 'HOUR_APPR_PROCESS_START',
       'FLAG_LAST_APPL_PER_CONTRACT', 'NFLAG_LAST_APPL_IN_DAY',
       'RATE_DOWN_PAYMENT', 'RATE_INTEREST_PRIMARY',
       'RATE_INTEREST_PRIVILEGED', 'NAME_CASH_LOAN_PURPOSE',
       'NAME_CONTRACT_STATUS', 'DAYS_DECISION', 'NAME_PAYMENT_TYPE',
       'CODE_REJECT_REASON', 'NAME_TYPE_SUITE', 'NAME_CLIENT_TYPE',
       'NAME_GOODS_CATEGORY', 'NAME_PORTFOLIO', 'NAME_PRODUCT_TYPE',
       'CHANNEL_TYPE', 'SELLERPLACE_AREA', 'NAME_SELLER_INDUSTRY',
       'CNT_PAYMENT', 'NAME_YIELD_GROUP', 'PRODUCT_COMBINATION',
       'DAYS_FIRST_DRAWING', 'DAYS_FIRST_DUE', 'DAYS_LAST_DUE_1ST_VERSION',
       'DAYS_LAST_DUE', 'DAYS_TERMINATION', 'NFLAG_INSURED_ON_APPROVAL'],
      dtype='object')
```

[443]:

```
            SK_ID_CURR  WAVG_AMT_ANNUITY  WAVG_AMT_APPLICATION  \
SK_ID_PREV
1049286         188536               NaN                   0.0

            WAVG_AMT_CREDIT  WAVG_AMT_DOWN_PAYMENT  WAVG_AMT_GOODS_PRICE  \
SK_ID_PREV
1049286                 0.0                    NaN                   NaN

            WAVG_RATE_DOWN_PAYMENT  WAVG_RATE_INTEREST_PRIMARY  \
SK_ID_PREV
1049286                        NaN                         NaN

            WAVG_RATE_INTEREST_PRIVILEGED
SK_ID_PREV
1049286                               NaN
```

[444]:

```
            WWAVG_AMT_ANNUITY  WWAVG_AMT_APPLICATION  WWAVG_AMT_CREDIT  \
SK_ID_CURR
313675               5871.501               148385.7          166054.5

            WWAVG_AMT_DOWN_PAYMENT  WWAVG_AMT_GOODS_PRICE  \
SK_ID_CURR
313675                     30328.2               296771.4

            WWAVG_RATE_DOWN_PAYMENT  WWAVG_RATE_INTEREST_PRIMARY  \
SK_ID_CURR
313675                     0.152227                          NaN

            WWAVG_RATE_INTEREST_PRIVILEGED
SK_ID_CURR
313675                                 NaN
```

```
"SK_ID_PREV", "SK_ID_CURR", "AMT_ANNUITY", "AMT_APPLICATION", "AMT_CREDIT",
"AMT_DOWN_PAYMENT", "AMT_GOODS_PRICE", "WEEKDAY_APPR_PROCESS_START",
"HOUR_APPR_PROCESS_START", "RATE_DOWN_PAYMENT", "RATE_INTEREST_PRIMARY",
"RATE_INTEREST_PRIVILEGED", "DAYS_DECISION", "SELLERPLACE_AREA",
"NAME_SELLER_INDUSTRY", "CNT_PAYMENT", "DAYS_FIRST_DRAWING", "DAYS_FIRST_DUE",
"DAYS_LAST_DUE_1ST_VERSION", "DAYS_LAST_DUE", "DAYS_TERMINATION",
"NAME_CONTRACT_TYPE_Cash loans", "NAME_CONTRACT_TYPE_Consumer loans",
"NAME_CONTRACT_TYPE_Revolving loans", "NAME_CONTRACT_TYPE_XNA",
"FLAG_LAST_APPL_PER_CONTRACT_N", "FLAG_LAST_APPL_PER_CONTRACT_Y",
"NFLAG_LAST_APPL_IN_DAY_0", "NFLAG_LAST_APPL_IN_DAY_1",
"NAME_CASH_LOAN_PURPOSE_Building a house or an annex",
"NAME_CASH_LOAN_PURPOSE_Business development", "NAME_CASH_LOAN_PURPOSE_Buying a
garage", "NAME_CASH_LOAN_PURPOSE_Buying a holiday home / land",
"NAME_CASH_LOAN_PURPOSE_Buying a home", "NAME_CASH_LOAN_PURPOSE_Buying a new
car", "NAME_CASH_LOAN_PURPOSE_Buying a used car", "NAME_CASH_LOAN_PURPOSE_Car
repairs", "NAME_CASH_LOAN_PURPOSE_Education", "NAME_CASH_LOAN_PURPOSE_Everyday
expenses", "NAME_CASH_LOAN_PURPOSE_Furniture",
"NAME_CASH_LOAN_PURPOSE_Gasification / water supply",
"NAME_CASH_LOAN_PURPOSE_Hobby", "NAME_CASH_LOAN_PURPOSE_Journey",
"NAME_CASH_LOAN_PURPOSE_Medicine", "NAME_CASH_LOAN_PURPOSE_Money for a third
person", "NAME_CASH_LOAN_PURPOSE_Other", "NAME_CASH_LOAN_PURPOSE_Payments on
```

```
other loans", "NAME_CASH_LOAN_PURPOSE_Purchase of electronic equipment",
"NAME_CASH_LOAN_PURPOSE_Refusal to name the goal",
"NAME_CASH_LOAN_PURPOSE_Repairs", "NAME_CASH_LOAN_PURPOSE_Urgent needs",
"NAME_CASH_LOAN_PURPOSE_Wedding / gift / holiday", "NAME_CASH_LOAN_PURPOSE_XAP",
"NAME_CASH_LOAN_PURPOSE_XNA", "NAME_CONTRACT_STATUS_Approved",
"NAME_CONTRACT_STATUS_Canceled", "NAME_CONTRACT_STATUS_Refused",
"NAME_CONTRACT_STATUS_Unused offer", "NAME_PAYMENT_TYPE_Cash through the bank",
"NAME_PAYMENT_TYPE_Cashless from the account of the employer",
"NAME_PAYMENT_TYPE_Non-cash from your account", "NAME_PAYMENT_TYPE_XNA",
"CODE_REJECT_REASON_CLIENT", "CODE_REJECT_REASON_HC",
"CODE_REJECT_REASON_LIMIT", "CODE_REJECT_REASON_SCO",
"CODE_REJECT_REASON_SCOFR", "CODE_REJECT_REASON_SYSTEM",
"CODE_REJECT_REASON_VERIF", "CODE_REJECT_REASON_XAP", "CODE_REJECT_REASON_XNA",
"NAME_TYPE_SUITE_Children", "NAME_TYPE_SUITE_Family", "NAME_TYPE_SUITE_Group of
people", "NAME_TYPE_SUITE_Other_A", "NAME_TYPE_SUITE_Other_B",
"NAME_TYPE_SUITE_Spouse, partner", "NAME_TYPE_SUITE_Unaccompanied",
"NAME_CLIENT_TYPE_New", "NAME_CLIENT_TYPE_Refreshed",
"NAME_CLIENT_TYPE_Repeater", "NAME_CLIENT_TYPE_XNA",
"NAME_GOODS_CATEGORY_Additional Service", "NAME_GOODS_CATEGORY_Animals",
"NAME_GOODS_CATEGORY_Audio/Video", "NAME_GOODS_CATEGORY_Auto Accessories",
"NAME_GOODS_CATEGORY_Clothing and Accessories", "NAME_GOODS_CATEGORY_Computers",
"NAME_GOODS_CATEGORY_Construction Materials", "NAME_GOODS_CATEGORY_Consumer
Electronics", "NAME_GOODS_CATEGORY_Direct Sales",
"NAME_GOODS_CATEGORY_Education", "NAME_GOODS_CATEGORY_Fitness",
"NAME_GOODS_CATEGORY_Furniture", "NAME_GOODS_CATEGORY_Gardening",
"NAME_GOODS_CATEGORY_Homewares", "NAME_GOODS_CATEGORY_House Construction",
"NAME_GOODS_CATEGORY_Insurance", "NAME_GOODS_CATEGORY_Jewelry",
"NAME_GOODS_CATEGORY_Medical Supplies", "NAME_GOODS_CATEGORY_Medicine",
"NAME_GOODS_CATEGORY_Mobile", "NAME_GOODS_CATEGORY_Office Appliances",
"NAME_GOODS_CATEGORY_Other", "NAME_GOODS_CATEGORY_Photo / Cinema Equipment",
"NAME_GOODS_CATEGORY_Sport and Leisure", "NAME_GOODS_CATEGORY_Tourism",
"NAME_GOODS_CATEGORY_Vehicles", "NAME_GOODS_CATEGORY_Weapon",
"NAME_GOODS_CATEGORY_XNA", "NAME_PORTFOLIO_Cards", "NAME_PORTFOLIO_Cars",
"NAME_PORTFOLIO_Cash", "NAME_PORTFOLIO_POS", "NAME_PORTFOLIO_XNA",
"NAME_PRODUCT_TYPE_XNA", "NAME_PRODUCT_TYPE_walk-in", "NAME_PRODUCT_TYPE_x-
sell", "CHANNEL_TYPE_AP+ (Cash loan)", "CHANNEL_TYPE_Car dealer",
"CHANNEL_TYPE_Channel of corporate sales", "CHANNEL_TYPE_Contact center",
"CHANNEL_TYPE_Country-wide", "CHANNEL_TYPE_Credit and cash offices",
"CHANNEL_TYPE_Regional / Local", "CHANNEL_TYPE_Stone", "NAME_YIELD_GROUP_XNA",
"NAME_YIELD_GROUP_high", "NAME_YIELD_GROUP_low_action",
"NAME_YIELD_GROUP_low_normal", "NAME_YIELD_GROUP_middle",
"PRODUCT_COMBINATION_Card Street", "PRODUCT_COMBINATION_Card X-Sell",
"PRODUCT_COMBINATION_Cash", "PRODUCT_COMBINATION_Cash Street: high",
"PRODUCT_COMBINATION_Cash Street: low", "PRODUCT_COMBINATION_Cash Street:
middle", "PRODUCT_COMBINATION_Cash X-Sell: high", "PRODUCT_COMBINATION_Cash
X-Sell: low", "PRODUCT_COMBINATION_Cash X-Sell: middle",
"PRODUCT_COMBINATION_POS household with interest", "PRODUCT_COMBINATION_POS
household without interest", "PRODUCT_COMBINATION_POS industry with interest",
"PRODUCT_COMBINATION_POS industry without interest", "PRODUCT_COMBINATION_POS
mobile with interest", "PRODUCT_COMBINATION_POS mobile without interest",
"PRODUCT_COMBINATION_POS other with interest", "PRODUCT_COMBINATION_POS others
without interest", "NFLAG_INSURED_ON_APPROVAL_0.0",
"NFLAG_INSURED_ON_APPROVAL_1.0"
```

[448]:

```
            NAME_CONTRACT_TYPE_Cash loans  NAME_CONTRACT_TYPE_Consumer loans  \
SK_ID_CURR
100001                                0.0                                1.0

            NAME_CONTRACT_TYPE_Revolving loans  NAME_CONTRACT_TYPE_XNA  \
SK_ID_CURR
100001                                     0.0                     0.0

            FLAG_LAST_APPL_PER_CONTRACT_N  FLAG_LAST_APPL_PER_CONTRACT_Y  \
SK_ID_CURR
100001                                0.0                            1.0

            NFLAG_LAST_APPL_IN_DAY_0  NFLAG_LAST_APPL_IN_DAY_1  \
SK_ID_CURR
100001                           0.0                       1.0

            NAME_CASH_LOAN_PURPOSE_Building a house or an annex  \
SK_ID_CURR
100001                                                      0.0

            NAME_CASH_LOAN_PURPOSE_Business development  …  \
SK_ID_CURR                                               …
100001                                              0.0  …

            PRODUCT_COMBINATION_POS household with interest  \
SK_ID_CURR
100001                                                  0.0

            PRODUCT_COMBINATION_POS household without interest  \
SK_ID_CURR
100001                                                     0.0

            PRODUCT_COMBINATION_POS industry with interest  \
SK_ID_CURR
100001                                                 0.0

            PRODUCT_COMBINATION_POS industry without interest  \
SK_ID_CURR
100001                                                    0.0

            PRODUCT_COMBINATION_POS mobile with interest  \
SK_ID_CURR
100001                                               1.0

            PRODUCT_COMBINATION_POS mobile without interest  \
SK_ID_CURR
100001                                                  0.0

            PRODUCT_COMBINATION_POS other with interest  \
SK_ID_CURR
100001                                              0.0

            PRODUCT_COMBINATION_POS others without interest  \
SK_ID_CURR
100001                                                  0.0

            NFLAG_INSURED_ON_APPROVAL_0.0  NFLAG_INSURED_ON_APPROVAL_1.0
SK_ID_CURR
100001                                1.0                            0.0

[1 rows x 129 columns]
```

All transformed categorical and numerical variables were concatenated.

instpayments: The dataframe instpayments contains data on the repayment history for the previously disbursed credits in Home Credit related to the loans in the sample. In order to use the information from the dataframe some data transformations were needed. Ducdb queries with subqueries were performed to obtain sums of differences between days and amounts of loans when payment was supposed to occur and when it actually occured in order to indentity how many days a client was late to proceed a payment and what amount.

```
SK_ID_PREV, SK_ID_CURR, NUM_INSTALMENT_VERSION, NUM_INSTALMENT_NUMBER,
DAYS_INSTALMENT, DAYS_ENTRY_PAYMENT, AMT_INSTALMENT, AMT_PAYMENT
```

```
FloatProgress(value=0.0, layout=Layout(width='100%'), style=ProgressStyle(bar_color='black'))
```

```
[452]:             sums_of_days_late  sums_of_days_in_time  sums_of_amounts_late  \
SK_ID_CURR
180748                   92.0                 225.0          1.455192e-11

            sums_of_amounts_in_time
SK_ID_CURR
180748                          0.0
```

bcbalance: The bcbalance dataframe contains data on monthly balances of previous credits in Credit Bureau. Transformations that were performed for this dataframe are similar to those which were performed on pcbalance and ccbalance dataframes. The variable "MONTH_BALANCE" was removed because it duplicates the variable "DAYS_DECISION" in the bureau dataset.

```
[453]: Index(['SK_ID_BUREAU', 'MONTHS_BALANCE', 'STATUS'], dtype='object')
```

```
[455]: Index(['SK_ID_BUREAU', 'MONTHS_BALANCE', 'STATUS_0', 'STATUS_1', 'STATUS_2',
              'STATUS_3', 'STATUS_4', 'STATUS_5', 'STATUS_C', 'STATUS_X'],
             dtype='object')
```

```
SK_ID_BUREAU, MONTHS_BALANCE, STATUS_0, STATUS_1, STATUS_2, STATUS_3, STATUS_4,
STATUS_5, STATUS_C, STATUS_X
```

```
[459]:    SK_ID_BUREAU  STATUS_0  STATUS_1  STATUS_2  STATUS_3  STATUS_4  STATUS_5  \
       0       5001709       0.0       0.0       0.0       0.0       0.0       0.0

          STATUS_C  STATUS_X
       0  0.886598  0.113402
```

```
[460]:    SK_ID_BUREAU  AVG_STATUS_0  AVG_STATUS_1  AVG_STATUS_2  AVG_STATUS_3  \
       0       5509124           1.0           0.0           0.0           0.0

          AVG_STATUS_4  AVG_STATUS_5  AVG_STATUS_C  AVG_STATUS_X
       0           0.0           0.0           0.0           0.0
```

bureau: The dataframe bureau contains data on all client's previous credits provided by other financial institutions that were reported to Credit Bureau.

First, the joined dataframe was created by combining burea data with averaged bcbalance data on the foreign key SK_ID_BUREAU.

```
Index(['SK_ID_CURR', 'SK_ID_BUREAU', 'CREDIT_ACTIVE', 'CREDIT_CURRENCY',
       'DAYS_CREDIT', 'CREDIT_DAY_OVERDUE', 'DAYS_CREDIT_ENDDATE',
       'DAYS_ENDDATE_FACT', 'AMT_CREDIT_MAX_OVERDUE', 'CNT_CREDIT_PROLONG',
       'AMT_CREDIT_SUM', 'AMT_CREDIT_SUM_DEBT', 'AMT_CREDIT_SUM_LIMIT',
       'AMT_CREDIT_SUM_OVERDUE', 'CREDIT_TYPE', 'DAYS_CREDIT_UPDATE',
       'AMT_ANNUITY'],
      dtype='object')
```

```
[462]:    SK_ID_CURR  SK_ID_BUREAU CREDIT_ACTIVE CREDIT_CURRENCY  DAYS_CREDIT  \
       0      296326       5857106        Active      currency 1         -256

          CREDIT_DAY_OVERDUE  DAYS_CREDIT_ENDDATE  DAYS_ENDDATE_FACT  \
       0                   0                 72.0                NaN

          AMT_CREDIT_MAX_OVERDUE  CNT_CREDIT_PROLONG  …  AMT_ANNUITY  \
       0                     NaN                   0  …          NaN

          SK_ID_BUREAU_2  AVG_STATUS_0  AVG_STATUS_1 AVG_STATUS_2  AVG_STATUS_3  \
       0         5857106      0.111111           0.0          0.0           0.0

          AVG_STATUS_4  AVG_STATUS_5  AVG_STATUS_C  AVG_STATUS_X
       0           0.0           0.0           0.0      0.888889
```

```
[1 rows x 26 columns]
```

Then the avarages of differences between days when a client was supposed to pay and actually paid were obtained, grouped by current client ids.

Averages of numerical variables of the dataframe were calculated, grouped by current client ids.

[463]:

```
            WAVG_CREDIT_END_LATE  WAVG_CREDIT_DAY_OVERDUE  \
SK_ID_CURR
411608                 37.666667                      0.0

            WAVG_AMT_CREDIT_MAX_OVERDUE  WAVG_CNT_CREDIT_PROLONG  \
SK_ID_CURR
411608                         1039.512                      0.0

            WAVG_AMT_CREDIT_SUM  WAVG_AMT_CREDIT_SUM_DEBT  \
SK_ID_CURR
411608               502424.058                       0.0

            WAVG_AMT_CREDIT_SUM_LIMIT  WAVG_AMT_CREDIT_SUM_OVERDUE  \
SK_ID_CURR
411608                            0.0                          0.0

            WAVG_DAYS_CREDIT_UPDATE  WAVG_AVG_STATUS_0  WAVG_AVG_STATUS_1  \
SK_ID_CURR
411608                  -774.866667           0.395211           0.018194

            WAVG_AVG_STATUS_2  WAVG_AVG_STATUS_3  WAVG_AVG_STATUS_4  \
SK_ID_CURR
411608               0.000794                0.0           0.000794

            WAVG_AVG_STATUS_5  WAVG_AVG_STATUS_C  WAVG_AVG_STATUS_X
SK_ID_CURR
411608               0.003175           0.500307           0.081525
```

Categorical variables were one hot encoded and grouped, averages calculated.

```
"SK_ID_CURR", "SK_ID_BUREAU", "DAYS_CREDIT", "CREDIT_DAY_OVERDUE",
"DAYS_CREDIT_ENDDATE", "DAYS_ENDDATE_FACT", "AMT_CREDIT_MAX_OVERDUE",
"CNT_CREDIT_PROLONG", "AMT_CREDIT_SUM", "AMT_CREDIT_SUM_DEBT",
"AMT_CREDIT_SUM_LIMIT", "AMT_CREDIT_SUM_OVERDUE", "DAYS_CREDIT_UPDATE",
"AMT_ANNUITY", "SK_ID_BUREAU_2", "AVG_STATUS_0", "AVG_STATUS_1", "AVG_STATUS_2",
"AVG_STATUS_3", "AVG_STATUS_4", "AVG_STATUS_5", "AVG_STATUS_C", "AVG_STATUS_X",
"CREDIT_ACTIVE_Active", "CREDIT_ACTIVE_Bad debt", "CREDIT_ACTIVE_Closed",
"CREDIT_ACTIVE_Sold", "CREDIT_CURRENCY_currency 1", "CREDIT_CURRENCY_currency
2", "CREDIT_CURRENCY_currency 3", "CREDIT_CURRENCY_currency 4",
"CREDIT_TYPE_Another type of loan", "CREDIT_TYPE_Car loan", "CREDIT_TYPE_Cash
loan (non-earmarked)", "CREDIT_TYPE_Consumer credit", "CREDIT_TYPE_Credit card",
"CREDIT_TYPE_Loan for business development", "CREDIT_TYPE_Loan for purchase of
shares (margin lending)", "CREDIT_TYPE_Loan for the purchase of equipment",
"CREDIT_TYPE_Loan for working capital replenishment", "CREDIT_TYPE_Microloan",
"CREDIT_TYPE_Mobile operator loan", "CREDIT_TYPE_Mortgage", "CREDIT_TYPE_Real
estate loan", "CREDIT_TYPE_Unknown type of loan"
```

Finally, modified variables were concatenated.

Merging transformed datasets In this step, after renaming columns in order to avoid repetitive names data from aptrain_mod dataframe were merged with other modified dataframes except aptest_mod). In the same way aptest_mod dataframe was merged with other dataframes (except aptrain_mod). Full datasets which could be used for machine learnining were obtained.

```
Index(['NAME_CONTRACT_TYPE_Cash loans', 'NAME_CONTRACT_TYPE_Consumer loans',
       'NAME_CONTRACT_TYPE_Revolving loans', 'NAME_CONTRACT_TYPE_XNA',
       'FLAG_LAST_APPL_PER_CONTRACT_N', 'FLAG_LAST_APPL_PER_CONTRACT_Y',
       'NFLAG_LAST_APPL_IN_DAY_0', 'NFLAG_LAST_APPL_IN_DAY_1',
       'NAME_CONTRACT_STATUS_Approved', 'NAME_CONTRACT_STATUS_Canceled',
       …
       'PRODUCT_COMBINATION_POS other with interest',
       'PRODUCT_COMBINATION_POS others without interest', 'WWAVG_AMT_ANNUITY',
       'WWAVG_AMT_APPLICATION', 'WWAVG_AMT_CREDIT', 'WWAVG_AMT_DOWN_PAYMENT',
       'WWAVG_AMT_GOODS_PRICE', 'WWAVG_RATE_DOWN_PAYMENT',
       'WWAVG_RATE_INTEREST_PRIMARY', 'WWAVG_RATE_INTEREST_PRIVILEGED'],
      dtype='object', length=131)
Index(['CREDIT_ACTIVE_Active', 'CREDIT_ACTIVE_Bad debt',
       'CREDIT_ACTIVE_Closed', 'CREDIT_ACTIVE_Sold',
       'CREDIT_CURRENCY_currency 1', 'CREDIT_CURRENCY_currency 2',
       'CREDIT_CURRENCY_currency 3', 'CREDIT_CURRENCY_currency 4',
       'CREDIT_TYPE_Another type of loan', 'CREDIT_TYPE_Car loan',
       'CREDIT_TYPE_Cash loan (non-earmarked)', 'CREDIT_TYPE_Consumer credit',
       'CREDIT_TYPE_Credit card', 'CREDIT_TYPE_Loan for business development',
       'CREDIT_TYPE_Loan for purchase of shares (margin lending)',
       'CREDIT_TYPE_Loan for the purchase of equipment',
       'CREDIT_TYPE_Loan for working capital replenishment',
       'CREDIT_TYPE_Microloan', 'CREDIT_TYPE_Mobile operator loan',
       'CREDIT_TYPE_Mortgage', 'CREDIT_TYPE_Real estate loan',
       'CREDIT_TYPE_Unknown type of loan', 'WAVG_CREDIT_END_LATE',
       'WAVG_CREDIT_DAY_OVERDUE', 'WAVG_AMT_CREDIT_MAX_OVERDUE',
       'WAVG_CNT_CREDIT_PROLONG', 'WAVG_AMT_CREDIT_SUM',
       'WAVG_AMT_CREDIT_SUM_DEBT', 'WAVG_AMT_CREDIT_SUM_LIMIT',
       'WAVG_AMT_CREDIT_SUM_OVERDUE', 'WAVG_DAYS_CREDIT_UPDATE',
       'WAVG_AVG_STATUS_0', 'WAVG_AVG_STATUS_1', 'WAVG_AVG_STATUS_2',
       'WAVG_AVG_STATUS_3', 'WAVG_AVG_STATUS_4', 'WAVG_AVG_STATUS_5',
       'WAVG_AVG_STATUS_C', 'WAVG_AVG_STATUS_X'],
      dtype='object')
Index(['sums_of_days_late', 'sums_of_days_in_time', 'sums_of_amounts_late',
       'sums_of_amounts_in_time'],
      dtype='object')
Index(['AVG_NAME_CONTRACT_STATUS_Active',
       'AVG_NAME_CONTRACT_STATUS_Amortized_debt',
       'AVG_NAME_CONTRACT_STATUS_Approved',
       'AVG_NAME_CONTRACT_STATUS_Canceled',
       'AVG_NAME_CONTRACT_STATUS_Completed', 'AVG_NAME_CONTRACT_STATUS_Demand',
       'AVG_NAME_CONTRACT_STATUS_Returned_to_the_store',
       'AVG_NAME_CONTRACT_STATUS_Signed', 'CNT_INSTALMENT_WAVG',
       'CNT_INSTALMENT_FUTURE_WAVG', 'SK_DPD_WAVG', 'SK_DPD_DEF_WAVG'],
      dtype='object')
Index(['AVG_NAME_CONTRACT_STATUS_CC_Active',
       'AVG_NAME_CONTRACT_STATUS_CC_Approved',
       'AVG_NAME_CONTRACT_STATUS_CC_Completed',
       'AVG_NAME_CONTRACT_STATUS_CC_Demand',
       'AVG_NAME_CONTRACT_STATUS_CC_Refused',
       'AVG_NAME_CONTRACT_STATUS_CC_Sent_proposal',
       'AVG_NAME_CONTRACT_STATUS_CC_Signed', 'WWAVG_AMT_BALANCE',
       'WWAVG_AMT_CREDIT_LIMIT_ACTUAL', 'WWAVG_AMT_DRAWINGS_ATM_CURRENT',
       'WWAVG_AMT_DRAWINGS_CURRENT', 'WWAVG_AMT_DRAWINGS_OTHER_CURRENT',
       'WWAVG_AMT_DRAWINGS_POS_CURRENT', 'WWAVG_AMT_INST_MIN_REGULARITY',
       'WWAVG_AMT_DRAWINGS_POS_CURRENT_2', 'WWAVG_AMT_PAYMENT_TOTAL_CURRENT',
       'WWAVG_AMT_RECEIVABLE_PRINCIPAL', 'WWAVG_AMT_RECIVABLE',
       'WWAVG_AMT_TOTAL_RECEIVABLE', 'WWAVG_CNT_DRAWINGS_ATM_CURRENT',
       'WWAVG_CNT_DRAWINGS_CURRENT', 'WWAVG_CNT_DRAWINGS_OTHER_CURRENT',
       'WWAVG_CNT_DRAWINGS_POS_CURRENT', 'WWAVG_CNT_INSTALMENT_MATURE_CUM',
       'WWAVG_SK_DPD', 'WWAVG_SK_DPD_DEF'],
      dtype='object')
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 307511 entries, 100002 to 456255
Columns: 320 entries, TARGET to WWAVG_SK_DPD_DEF
dtypes: float64(230), int32(1), int64(19), object(11), uint8(59)
memory usage: 638.9+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 48744 entries, 100001 to 456250
Columns: 319 entries, NAME_CONTRACT_TYPE to WWAVG_SK_DPD_DEF
dtypes: float64(230), int32(1), int64(18), object(11), uint8(59)
memory usage: 100.6+ MB
```

Checking for missing values

[478]:

```
TARGET                                  0
NAME_CONTRACT_TYPE                      0
FLAG_OWN_REALTY                         0
CNT_CHILDREN                            0
AMT_INCOME_TOTAL                        0
                                      …
WWAVG_CNT_DRAWINGS_OTHER_CURRENT   246371
```

```
       WWAVG_CNT_DRAWINGS_POS_CURRENT      246371
       WWAVG_CNT_INSTALMENT_MATURE_CUM     220606
       WWAVG_SK_DPD                        220606
       WWAVG_SK_DPD_DEF                    220606
       Length: 320, dtype: int64
```

```
[479]: NAME_CONTRACT_TYPE                      0
       FLAG_OWN_REALTY                         0
       CNT_CHILDREN                            0
       AMT_INCOME_TOTAL                        0
       AMT_CREDIT                              0
                                           …
       WWAVG_CNT_DRAWINGS_OTHER_CURRENT    37690
       WWAVG_CNT_DRAWINGS_POS_CURRENT      37690
       WWAVG_CNT_INSTALMENT_MATURE_CUM     32091
       WWAVG_SK_DPD                        32091
       WWAVG_SK_DPD_DEF                    32091
       Length: 319, dtype: int64
```

For the training of machine learning models the fulldata_train dataframe will be used. The aptest dataframe will be used solely for testing the models.

1.2.1 Cleaning up the data

In order to prepare the dataset which could be used for machine learning, data have to be cleaned. Bellow, several approaches to cleaning data were applied for different variables.

Dropping rows with certain values Rows with small numbers of values for some of the variables were dropped from the dataset in order to avoid the data not being present in either training or test datasets.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 307511 entries, 100002 to 456255
Columns: 315 entries, TARGET to WWAVG_SK_DPD_DEF
dtypes: float64(225), int32(1), int64(19), object(11), uint8(59)
memory usage: 627.1+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 48744 entries, 100001 to 456250
Columns: 314 entries, NAME_CONTRACT_TYPE to WWAVG_SK_DPD_DEF
dtypes: float64(225), int32(1), int64(18), object(11), uint8(59)
memory usage: 98.8+ MB
```

The final transformed datasets contain 307511 cases and 315 features (with the ’TARGET” variable) in the train dataset and 48744 cases and 314 features (without the “TARGET” variable) in the test dataset.

The transformed datasets were saved to a local computer in order to use them for the machine learning (see the “homecredit_machine_learning.ipynb” file).

1.2.2 Examining relationships between features and the target variable

The last part of the exploratory analysis contains examining relationships between different types of features (numerical, binary and categorical) and the target variable.

Examining relationships between numerical variables and the target variable - on-parametric test for statistical significance of mean rank differences (Mann Whitney U test) Relationships between the target variable and numerical variables were examined by testing statistical significance of differences between means of numerical variables in two groups - persons with payment difficulties and persons who do not have payment difficulties.

As only few numerical variables are normally distributed, it was decided to use the non-parametric Mann Whitney U test for testing statistical significance of differences between mean ranks in groups of clients who experience loan payment difficulties and those who do not. Results of the analysis for the confidence level 0.95 are presented at the end of the output bellow. The output also provides info for for which features mean ranks in one group are higher than in the other group.

First, numerical variables are distinguished from binary variables with values 0 and 1. These variables will be examined as categorical variables later.

```
[490]: Index(['TARGET', 'NAME_CONTRACT_TYPE', 'FLAG_OWN_REALTY', 'FLAG_MOBIL',
              'FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'FLAG_CONT_MOBILE', 'FLAG_PHONE',
              'FLAG_EMAIL', 'REG_REGION_NOT_LIVE_REGION',
              'REG_REGION_NOT_WORK_REGION', 'LIVE_REGION_NOT_WORK_REGION',
              'REG_CITY_NOT_LIVE_CITY', 'REG_CITY_NOT_WORK_CITY',
              'LIVE_CITY_NOT_WORK_CITY', 'ORGANIZATION_TYPE_Advertising',
              'ORGANIZATION_TYPE_Agriculture', 'ORGANIZATION_TYPE_Bank',
              'ORGANIZATION_TYPE_Business Entity Type 1',
              'ORGANIZATION_TYPE_Business Entity Type 2',
              'ORGANIZATION_TYPE_Business Entity Type 3',
              'ORGANIZATION_TYPE_Cleaning', 'ORGANIZATION_TYPE_Construction',
              'ORGANIZATION_TYPE_Culture', 'ORGANIZATION_TYPE_Electricity',
              'ORGANIZATION_TYPE_Emergency', 'ORGANIZATION_TYPE_Government',
              'ORGANIZATION_TYPE_Hotel', 'ORGANIZATION_TYPE_Housing',
              'ORGANIZATION_TYPE_Industry: type 1',
              'ORGANIZATION_TYPE_Industry: type 10',
              'ORGANIZATION_TYPE_Industry: type 11',
              'ORGANIZATION_TYPE_Industry: type 12',
              'ORGANIZATION_TYPE_Industry: type 13',
              'ORGANIZATION_TYPE_Industry: type 2',
              'ORGANIZATION_TYPE_Industry: type 3',
              'ORGANIZATION_TYPE_Industry: type 4',
              'ORGANIZATION_TYPE_Industry: type 5',
              'ORGANIZATION_TYPE_Industry: type 6',
              'ORGANIZATION_TYPE_Industry: type 7',
              'ORGANIZATION_TYPE_Industry: type 8',
              'ORGANIZATION_TYPE_Industry: type 9', 'ORGANIZATION_TYPE_Insurance',
              'ORGANIZATION_TYPE_Kindergarten', 'ORGANIZATION_TYPE_Legal Services',
              'ORGANIZATION_TYPE_Medicine', 'ORGANIZATION_TYPE_Military',
              'ORGANIZATION_TYPE_Mobile', 'ORGANIZATION_TYPE_Other',
              'ORGANIZATION_TYPE_Police', 'ORGANIZATION_TYPE_Postal',
              'ORGANIZATION_TYPE_Realtor', 'ORGANIZATION_TYPE_Religion',
              'ORGANIZATION_TYPE_Restaurant', 'ORGANIZATION_TYPE_School',
              'ORGANIZATION_TYPE_Security', 'ORGANIZATION_TYPE_Security Ministries',
              'ORGANIZATION_TYPE_Self-employed', 'ORGANIZATION_TYPE_Services',
              'ORGANIZATION_TYPE_Telecom', 'ORGANIZATION_TYPE_Trade: type 1',
              'ORGANIZATION_TYPE_Trade: type 2', 'ORGANIZATION_TYPE_Trade: type 3',
              'ORGANIZATION_TYPE_Trade: type 4', 'ORGANIZATION_TYPE_Trade: type 5',
              'ORGANIZATION_TYPE_Trade: type 6', 'ORGANIZATION_TYPE_Trade: type 7',
              'ORGANIZATION_TYPE_Transport: type 1',
              'ORGANIZATION_TYPE_Transport: type 2',
              'ORGANIZATION_TYPE_Transport: type 3',
              'ORGANIZATION_TYPE_Transport: type 4', 'ORGANIZATION_TYPE_University',
              'GENDER_F', 'GENDER_M', 'NAME_GOODS_CATEGORY_House Construction'],
             dtype='object')
```

```
[491]: Index(['TARGET', 'CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT',
              'AMT_ANNUITY', 'AMT_GOODS_PRICE', 'REGION_POPULATION_RELATIVE',
              'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION',
              …
              'WWAVG_AMT_RECEIVABLE_PRINCIPAL', 'WWAVG_AMT_RECIVABLE',
              'WWAVG_AMT_TOTAL_RECEIVABLE', 'WWAVG_CNT_DRAWINGS_ATM_CURRENT',
              'WWAVG_CNT_DRAWINGS_CURRENT', 'WWAVG_CNT_DRAWINGS_OTHER_CURRENT',
              'WWAVG_CNT_DRAWINGS_POS_CURRENT', 'WWAVG_CNT_INSTALMENT_MATURE_CUM',
              'WWAVG_SK_DPD', 'WWAVG_SK_DPD_DEF'],
             dtype='object', length=309)
```

```
For confidence level 0.95, there is the statistically significant difference
between means of CNT_CHILDREN in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CNT_CHILDREN is higher in the group of clients
with payment difficulties <Yes>.



For confidence level 0.95, there is the statistically significant difference
between means of AMT_INCOME_TOTAL in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AMT_INCOME_TOTAL is higher in the group of clients
who do not have payment difficulties <No>.



For confidence level 0.95, there is the statistically significant difference
between means of AMT_CREDIT in groups of clients with payment difficulties <Yes>
```

```
and those
who do not have payment difficulties <No>.

The mean of the feature AMT_CREDIT is higher in the group of clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is no statistically significant difference
between means of AMT_ANNUITY in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AMT_ANNUITY is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AMT_GOODS_PRICE in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AMT_GOODS_PRICE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of REGION_POPULATION_RELATIVE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature REGION_POPULATION_RELATIVE is higher in the group of
clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is the statistically significant difference
between means of DAYS_BIRTH in groups of clients with payment difficulties <Yes>
and those
who do not have payment difficulties <No>.

The mean of the feature DAYS_BIRTH is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of DAYS_EMPLOYED in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.
```

```
The mean of the feature DAYS_EMPLOYED is higher in the group of clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is the statistically significant difference
between means of DAYS_REGISTRATION in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature DAYS_REGISTRATION is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of DAYS_ID_PUBLISH in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature DAYS_ID_PUBLISH is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CNT_FAM_MEMBERS in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CNT_FAM_MEMBERS is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of REGION_RATING_CLIENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature REGION_RATING_CLIENT is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of REGION_RATING_CLIENT_W_CITY in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature REGION_RATING_CLIENT_W_CITY is higher in the group of
clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of EXT_SOURCE_1 in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature EXT_SOURCE_1 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of EXT_SOURCE_2 in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature EXT_SOURCE_2 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of EXT_SOURCE_3 in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature EXT_SOURCE_3 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of TOTALAREA_MODE in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature TOTALAREA_MODE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of OBS_60_CNT_SOCIAL_CIRCLE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature OBS_60_CNT_SOCIAL_CIRCLE is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of DAYS_LAST_PHONE_CHANGE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature DAYS_LAST_PHONE_CHANGE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of EDUCATION in groups of clients with payment difficulties <Yes>
and those
who do not have payment difficulties <No>.

The mean of the feature EDUCATION is higher in the group of clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is no statistically significant difference
between means of CAR_OWN in groups of clients with payment difficulties <Yes>
and those
who do not have payment difficulties <No>.

The mean of the feature CAR_OWN is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of LIVING_CONDITIONS_1 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature LIVING_CONDITIONS_1 is higher in the group of clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is the statistically significant difference
between means of LIVING_CONDITIONS_2 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature LIVING_CONDITIONS_2 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CB_enquiries_1 in groups of clients with payment difficulties
```

```
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CB_enquiries_1 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is the statistically significant difference
between means of CB_enquiries_2 in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CB_enquiries_2 is higher in the group of clients
who do not have payment difficulties <No>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_TYPE_Cash loans in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_TYPE_Cash loans is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_TYPE_Consumer loans in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_TYPE_Consumer loans is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_TYPE_Revolving loans in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_TYPE_Revolving loans is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_TYPE_XNA in groups of clients with payment
```

```
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_TYPE_XNA is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of FLAG_LAST_APPL_PER_CONTRACT_N in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature FLAG_LAST_APPL_PER_CONTRACT_N is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of FLAG_LAST_APPL_PER_CONTRACT_Y in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature FLAG_LAST_APPL_PER_CONTRACT_Y is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NFLAG_LAST_APPL_IN_DAY_0 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NFLAG_LAST_APPL_IN_DAY_0 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NFLAG_LAST_APPL_IN_DAY_1 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NFLAG_LAST_APPL_IN_DAY_1 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of NAME_CONTRACT_STATUS_Approved in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_STATUS_Approved is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_STATUS_Canceled in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_STATUS_Canceled is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_STATUS_Refused in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_STATUS_Refused is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CONTRACT_STATUS_Unused offer in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CONTRACT_STATUS_Unused offer is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PAYMENT_TYPE_Cash through the bank in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PAYMENT_TYPE_Cash through the bank is higher in the
group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of NAME_PAYMENT_TYPE_Cashless from the account of the employer in
groups of clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PAYMENT_TYPE_Cashless from the account of the
employer is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PAYMENT_TYPE_Non-cash from your account in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PAYMENT_TYPE_Non-cash from your account is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PAYMENT_TYPE_XNA in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PAYMENT_TYPE_XNA is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_CLIENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_CLIENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_HC in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_HC is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_LIMIT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_LIMIT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_SCO in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_SCO is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_SCOFR in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_SCOFR is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_SYSTEM in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_SYSTEM is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_VERIF in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_VERIF is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CODE_REJECT_REASON_XNA in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CODE_REJECT_REASON_XNA is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Children in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Children is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Family in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Family is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Group of people in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Group of people is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Other_A in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Other_A is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Other_B in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Other_B is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Spouse, partner in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Spouse, partner is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_TYPE_SUITE_Unaccompanied in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_TYPE_SUITE_Unaccompanied is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CLIENT_TYPE_New in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CLIENT_TYPE_New is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CLIENT_TYPE_Refreshed in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CLIENT_TYPE_Refreshed is higher in the group of
clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CLIENT_TYPE_Repeater in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CLIENT_TYPE_Repeater is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PORTFOLIO_Cards in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PORTFOLIO_Cards is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PORTFOLIO_Cars in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PORTFOLIO_Cars is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PORTFOLIO_Cash in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PORTFOLIO_Cash is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PORTFOLIO_POS in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PORTFOLIO_POS is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of NAME_PORTFOLIO_XNA in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PORTFOLIO_XNA is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PRODUCT_TYPE_XNA in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PRODUCT_TYPE_XNA is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PRODUCT_TYPE_walk-in in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PRODUCT_TYPE_walk-in is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_PRODUCT_TYPE_x-sell in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_PRODUCT_TYPE_x-sell is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_AP+ (Cash loan) in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_AP+ (Cash loan) is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Car dealer in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Car dealer is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Channel of corporate sales in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Channel of corporate sales is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Contact center in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Contact center is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Country-wide in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Country-wide is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Credit and cash offices in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Credit and cash offices is higher in the
group of clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Regional / Local in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Regional / Local is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CHANNEL_TYPE_Stone in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CHANNEL_TYPE_Stone is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_YIELD_GROUP_high in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_YIELD_GROUP_high is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_YIELD_GROUP_low_action in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_YIELD_GROUP_low_action is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_YIELD_GROUP_low_normal in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_YIELD_GROUP_low_normal is higher in the group of
clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_YIELD_GROUP_middle in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_YIELD_GROUP_middle is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NFLAG_INSURED_ON_APPROVAL_0.0 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NFLAG_INSURED_ON_APPROVAL_0.0 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NFLAG_INSURED_ON_APPROVAL_1.0 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NFLAG_INSURED_ON_APPROVAL_1.0 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Additional Service in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Additional Service is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Animals in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.
```

```
The mean of the feature NAME_GOODS_CATEGORY_Animals is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Audio/Video in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Audio/Video is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Auto Accessories in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Auto Accessories is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Clothing and Accessories in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Clothing and Accessories is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Computers in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Computers is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Construction Materials in groups of clients
with payment difficulties <Yes> and those
```

```
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Construction Materials is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Consumer Electronics in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Consumer Electronics is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Direct Sales in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Direct Sales is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Education in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Education is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Fitness in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Fitness is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of NAME_GOODS_CATEGORY_Furniture in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Furniture is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Gardening in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Gardening is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Homewares in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Homewares is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Insurance in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Insurance is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Jewelry in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Jewelry is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Medical Supplies in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Medical Supplies is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Medicine in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Medicine is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Mobile in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Mobile is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Office Appliances in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Office Appliances is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Other in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Other is higher in the group of
clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Photo / Cinema Equipment in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Photo / Cinema Equipment is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Sport and Leisure in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Sport and Leisure is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Tourism in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Tourism is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Vehicles in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_GOODS_CATEGORY_Vehicles is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_GOODS_CATEGORY_Weapon in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.
```

```
The mean of the feature NAME_GOODS_CATEGORY_Weapon is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Building a house or an annex in groups
of clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Building a house or an annex is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Business development in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Business development is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Buying a garage in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Buying a garage is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Buying a holiday home / land in groups
of clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Buying a holiday home / land is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Buying a home in groups of clients with
payment difficulties <Yes> and those
```

```
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Buying a home is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Buying a new car in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Buying a new car is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Buying a used car in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Buying a used car is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Car repairs in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Car repairs is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Education in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Education is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of NAME_CASH_LOAN_PURPOSE_Everyday expenses in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Everyday expenses is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Furniture in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Furniture is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Gasification / water supply in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Gasification / water supply is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Hobby in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Hobby is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Journey in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Journey is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Medicine in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Medicine is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Money for a third person in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Money for a third person is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Other in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Other is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Payments on other loans in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Payments on other loans is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Purchase of electronic equipment in
groups of clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Purchase of electronic equipment
is higher in the group of clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Refusal to name the goal in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Refusal to name the goal is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Repairs in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Repairs is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Urgent needs in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Urgent needs is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of NAME_CASH_LOAN_PURPOSE_Wedding / gift / holiday in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature NAME_CASH_LOAN_PURPOSE_Wedding / gift / holiday is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Card Street in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.
```

```
The mean of the feature PRODUCT_COMBINATION_Card Street is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Card X-Sell in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Card X-Sell is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash Street: high in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash Street: high is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash Street: low in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash Street: low is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash Street: middle in groups of clients
with payment difficulties <Yes> and those
```

```
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash Street: middle is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash X-Sell: high in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash X-Sell: high is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash X-Sell: low in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash X-Sell: low is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_Cash X-Sell: middle in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_Cash X-Sell: middle is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS household with interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS household with interest is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of PRODUCT_COMBINATION_POS household without interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS household without interest is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS industry with interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS industry with interest is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS industry without interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS industry without interest is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS mobile with interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS mobile with interest is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS mobile without interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS mobile without interest is
higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS other with interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS other with interest is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of PRODUCT_COMBINATION_POS others without interest in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature PRODUCT_COMBINATION_POS others without interest is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_ANNUITY in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_ANNUITY is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_APPLICATION in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_APPLICATION is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_CREDIT in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_CREDIT is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DOWN_PAYMENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DOWN_PAYMENT is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_GOODS_PRICE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_GOODS_PRICE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_RATE_DOWN_PAYMENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_RATE_DOWN_PAYMENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_RATE_INTEREST_PRIMARY in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_RATE_INTEREST_PRIMARY is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_RATE_INTEREST_PRIVILEGED in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_RATE_INTEREST_PRIVILEGED is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_ACTIVE_Active in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_ACTIVE_Active is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_ACTIVE_Bad debt in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_ACTIVE_Bad debt is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_ACTIVE_Closed in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_ACTIVE_Closed is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_ACTIVE_Sold in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_ACTIVE_Sold is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_CURRENCY_currency 1 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_CURRENCY_currency 1 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_CURRENCY_currency 2 in groups of clients with payment
```

```
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_CURRENCY_currency 2 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_CURRENCY_currency 3 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_CURRENCY_currency 3 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_CURRENCY_currency 4 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_CURRENCY_currency 4 is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Another type of loan in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Another type of loan is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Car loan in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Car loan is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of CREDIT_TYPE_Cash loan (non-earmarked) in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Cash loan (non-earmarked) is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Consumer credit in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Consumer credit is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Credit card in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Credit card is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Loan for business development in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Loan for business development is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Loan for purchase of shares (margin lending) in
groups of clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Loan for purchase of shares (margin lending)
is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Loan for the purchase of equipment in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Loan for the purchase of equipment is higher
in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Loan for working capital replenishment in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Loan for working capital replenishment is
higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Microloan in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Microloan is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Mobile operator loan in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Mobile operator loan is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Mortgage in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Mortgage is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Real estate loan in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Real estate loan is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CREDIT_TYPE_Unknown type of loan in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CREDIT_TYPE_Unknown type of loan is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_CREDIT_END_LATE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_CREDIT_END_LATE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_CREDIT_DAY_OVERDUE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_CREDIT_DAY_OVERDUE is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AMT_CREDIT_MAX_OVERDUE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AMT_CREDIT_MAX_OVERDUE is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of WAVG_CNT_CREDIT_PROLONG in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_CNT_CREDIT_PROLONG is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AMT_CREDIT_SUM in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AMT_CREDIT_SUM is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AMT_CREDIT_SUM_DEBT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AMT_CREDIT_SUM_DEBT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AMT_CREDIT_SUM_LIMIT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AMT_CREDIT_SUM_LIMIT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AMT_CREDIT_SUM_OVERDUE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AMT_CREDIT_SUM_OVERDUE is higher in the group of
clients
```

```
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_DAYS_CREDIT_UPDATE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_DAYS_CREDIT_UPDATE is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_0 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_0 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_1 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_1 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_2 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_2 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_3 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_3 is higher in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_4 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_4 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_5 in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_5 is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_C in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_C is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WAVG_AVG_STATUS_X in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WAVG_AVG_STATUS_X is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of sums_of_days_late in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature sums_of_days_late is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of sums_of_days_in_time in groups of clients with payment
```

```
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature sums_of_days_in_time is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of sums_of_amounts_late in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature sums_of_amounts_late is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of sums_of_amounts_in_time in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature sums_of_amounts_in_time is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Active in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Active is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Amortized_debt in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Amortized_debt is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Approved in groups of clients with
```

```
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Approved is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Canceled in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Canceled is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Completed in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Completed is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Demand in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Demand is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Returned_to_the_store in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Returned_to_the_store is higher
in the group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_Signed in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_Signed is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CNT_INSTALMENT_WAVG in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CNT_INSTALMENT_WAVG is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of CNT_INSTALMENT_FUTURE_WAVG in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature CNT_INSTALMENT_FUTURE_WAVG is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of SK_DPD_WAVG in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature SK_DPD_WAVG is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of SK_DPD_DEF_WAVG in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature SK_DPD_DEF_WAVG is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
```

```
between means of AVG_NAME_CONTRACT_STATUS_CC_Active in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Active is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Approved in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Approved is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Completed in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Completed is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Demand in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Demand is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Refused in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Refused is higher in the
group of clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Sent_proposal in groups of clients
with payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Sent_proposal is higher in
the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of AVG_NAME_CONTRACT_STATUS_CC_Signed in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature AVG_NAME_CONTRACT_STATUS_CC_Signed is higher in the
group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_BALANCE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_BALANCE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_CREDIT_LIMIT_ACTUAL in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_CREDIT_LIMIT_ACTUAL is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DRAWINGS_ATM_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DRAWINGS_ATM_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.
```

```
For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DRAWINGS_CURRENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DRAWINGS_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DRAWINGS_OTHER_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DRAWINGS_OTHER_CURRENT is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DRAWINGS_POS_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DRAWINGS_POS_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_INST_MIN_REGULARITY in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_INST_MIN_REGULARITY is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_DRAWINGS_POS_CURRENT_2 in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_DRAWINGS_POS_CURRENT_2 is higher in the group
```

```
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_PAYMENT_TOTAL_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_PAYMENT_TOTAL_CURRENT is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_RECEIVABLE_PRINCIPAL in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_RECEIVABLE_PRINCIPAL is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_RECIVABLE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_RECIVABLE is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_AMT_TOTAL_RECEIVABLE in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_AMT_TOTAL_RECEIVABLE is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_CNT_DRAWINGS_ATM_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.
```

```
The mean of the feature WWAVG_CNT_DRAWINGS_ATM_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_CNT_DRAWINGS_CURRENT in groups of clients with payment
difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_CNT_DRAWINGS_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_CNT_DRAWINGS_OTHER_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_CNT_DRAWINGS_OTHER_CURRENT is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_CNT_DRAWINGS_POS_CURRENT in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_CNT_DRAWINGS_POS_CURRENT is higher in the group of
clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_CNT_INSTALMENT_MATURE_CUM in groups of clients with
payment difficulties <Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_CNT_INSTALMENT_MATURE_CUM is higher in the group
of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_SK_DPD in groups of clients with payment difficulties
<Yes> and those
```

```
who do not have payment difficulties <No>.

The mean of the feature WWAVG_SK_DPD is higher in the group of clients
with payment difficulties <Yes>.


For confidence level 0.95, there is no statistically significant difference
between means of WWAVG_SK_DPD_DEF in groups of clients with payment difficulties
<Yes> and those
who do not have payment difficulties <No>.

The mean of the feature WWAVG_SK_DPD_DEF is higher in the group of clients
with payment difficulties <Yes>.




For confidence level 0.95 significant statistical differences in groups of
clients with payment difficulties <Yes> and those
who do not have payment difficulties <No> are for these features:
['CNT_CHILDREN', 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'REGION_POPULATION_RELATIVE',
'DAYS_BIRTH', 'DAYS_EMPLOYED', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH',
'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'EDUCATION',
'LIVING_CONDITIONS_1', 'LIVING_CONDITIONS_2', 'CB_enquiries_2'])


Significantly higher means in the group of clients with payment difficulties
<Yes> are for these features:
['CNT_CHILDREN', 'DAYS_BIRTH', 'DAYS_REGISTRATION', 'DAYS_ID_PUBLISH',
'REGION_RATING_CLIENT', 'REGION_RATING_CLIENT_W_CITY', 'LIVING_CONDITIONS_2']

Significantly higher means in the group of clients who do not have payment
difficulties <No>
are for these features:
['AMT_INCOME_TOTAL', 'AMT_CREDIT', 'REGION_POPULATION_RELATIVE',
'DAYS_EMPLOYED', 'EDUCATION', 'LIVING_CONDITIONS_1', 'CB_enquiries_2']
```

Examining relationships between categorical feature variables and a target variable (chi square tests) In order to examine which of the categorical feature variables has a effect on the target variable, chi square tests were performed.

First, dictionaries for appending contingency tables were created, contingency tables for binary and categorical variables were presented in the output.

```
                       No   Yes
NAME_CONTRACT_TYPE
Cash loans          91.65  8.35
```

```
Revolving loans     94.52  5.48

                    No   Yes
FLAG_OWN_REALTY
N                91.68  8.32
Y                92.04  7.96

                No   Yes
FLAG_MOBIL
0           100.00  0.00
1            91.93  8.07

                   No   Yes
FLAG_EMP_PHONE
0               94.60  5.40
1               91.34  8.66

                    No   Yes
FLAG_WORK_PHONE
0                92.31  7.69
1                90.37  9.63

                     No   Yes
FLAG_CONT_MOBILE
0                 92.16  7.84
1                 91.93  8.07

               No   Yes
FLAG_PHONE
0           91.52  8.48
1           92.96  7.04

               No   Yes
FLAG_EMAIL
0           91.92  8.08
1           92.12  7.88

                               No   Yes
REG_REGION_NOT_LIVE_REGION
0                           91.95  8.05
1                           90.70  9.30

                               No   Yes
REG_REGION_NOT_WORK_REGION
0                           91.97  8.03
1                           91.11  8.89

                                No   Yes
```

```
LIVE_REGION_NOT_WORK_REGION
0                            91.94  8.06
1                            91.55  8.45

                           No    Yes
REG_CITY_NOT_LIVE_CITY
0                       92.28   7.72
1                       87.77  12.23

                           No    Yes
REG_CITY_NOT_WORK_CITY
0                       92.69   7.31
1                       89.39  10.61

                            No   Yes
LIVE_CITY_NOT_WORK_CITY
0                        92.34  7.66
1                        90.03  9.97

                                  No   Yes
ORGANIZATION_TYPE_Advertising
0                              91.93  8.07
1                              91.84  8.16

                                  No    Yes
ORGANIZATION_TYPE_Agriculture
0                              91.95   8.05
1                              89.53  10.47

                           No   Yes
ORGANIZATION_TYPE_Bank
0                       91.90  8.10
1                       94.81  5.19

                                             No   Yes
ORGANIZATION_TYPE_Business Entity Type 1
0                                         91.93  8.07
1                                         91.86  8.14

                                             No   Yes
ORGANIZATION_TYPE_Business Entity Type 2
0                                         91.94  8.06
1                                         91.47  8.53

                                             No   Yes
ORGANIZATION_TYPE_Business Entity Type 3
0                                         92.28  7.72
1                                         90.70  9.30
```

```
                               No    Yes
ORGANIZATION_TYPE_Cleaning
0                           91.93   8.07
1                           88.85  11.15

                                   No    Yes
ORGANIZATION_TYPE_Construction
0                               92.01   7.99
1                               88.32  11.68

                              No   Yes
ORGANIZATION_TYPE_Culture
0                          91.92  8.08
1                          94.46  5.54

                                  No   Yes
ORGANIZATION_TYPE_Electricity
0                              91.92  8.08
1                              93.37  6.63

                                No   Yes
ORGANIZATION_TYPE_Emergency
0                            91.93  8.07
1                            92.86  7.14

                                 No   Yes
ORGANIZATION_TYPE_Government
0                             91.89  8.11
1                             93.02  6.98

                            No   Yes
ORGANIZATION_TYPE_Hotel
0                        91.92  8.08
1                        93.58  6.42

                              No   Yes
ORGANIZATION_TYPE_Housing
0                          91.93  8.07
1                          92.06  7.94

                                       No    Yes
ORGANIZATION_TYPE_Industry: type 1
0                                   91.94   8.06
1                                   88.93  11.07

                                        No   Yes
ORGANIZATION_TYPE_Industry: type 10
```

```
0                                    91.93  8.07
1                                    93.58  6.42

                                        No   Yes
ORGANIZATION_TYPE_Industry: type 11
0                                    91.93  8.07
1                                    91.35  8.65

                                        No   Yes
ORGANIZATION_TYPE_Industry: type 12
0                                    91.92  8.08
1                                    96.21  3.79

                                        No    Yes
ORGANIZATION_TYPE_Industry: type 13
0                                    91.93   8.07
1                                    86.57  13.43

                                       No   Yes
ORGANIZATION_TYPE_Industry: type 2
0                                   91.93  8.07
1                                   92.79  7.21

                                       No    Yes
ORGANIZATION_TYPE_Industry: type 3
0                                   91.95   8.05
1                                   89.38  10.62

                                       No    Yes
ORGANIZATION_TYPE_Industry: type 4
0                                   91.93   8.07
1                                   89.85  10.15

                                       No   Yes
ORGANIZATION_TYPE_Industry: type 5
0                                   91.92  8.08
1                                   93.16  6.84

                                       No   Yes
ORGANIZATION_TYPE_Industry: type 6
0                                   91.93  8.07
1                                   92.86  7.14

                                       No   Yes
ORGANIZATION_TYPE_Industry: type 7
0                                   91.93  8.07
1                                   91.97  8.03
```

```
                                       No    Yes
ORGANIZATION_TYPE_Industry: type 8
0                                   91.93   8.07
1                                   87.50  12.50

                                       No   Yes
ORGANIZATION_TYPE_Industry: type 9
0                                   91.91  8.09
1                                   93.32  6.68

                                No   Yes
ORGANIZATION_TYPE_Insurance
0                            91.92  8.08
1                            94.30  5.70

                                   No   Yes
ORGANIZATION_TYPE_Kindergarten
0                               91.90  8.10
1                               92.97  7.03

                                     No   Yes
ORGANIZATION_TYPE_Legal Services
0                                 91.93  8.07
1                                 92.13  7.87

                               No   Yes
ORGANIZATION_TYPE_Medicine
0                           91.87  8.13
1                           93.42  6.58

                               No   Yes
ORGANIZATION_TYPE_Military
0                           91.90  8.10
1                           94.87  5.13

                             No   Yes
ORGANIZATION_TYPE_Mobile
0                         91.93  8.07
1                         90.85  9.15

                            No   Yes
ORGANIZATION_TYPE_Other
0                        91.90  8.10
1                        92.36  7.64

                            No  Yes
ORGANIZATION_TYPE_Police
0                         91.9  8.1
```

```
1                         95.0  5.0

                             No   Yes
ORGANIZATION_TYPE_Postal
0                         91.93  8.07
1                         91.56  8.44

                              No    Yes
ORGANIZATION_TYPE_Realtor
0                          91.93   8.07
1                          89.39  10.61

                               No   Yes
ORGANIZATION_TYPE_Religion
0                           91.93  8.07
1                           94.12  5.88

                                 No    Yes
ORGANIZATION_TYPE_Restaurant
0                             91.95   8.05
1                             88.29  11.71

                             No   Yes
ORGANIZATION_TYPE_School
0                         91.86  8.14
1                         94.09  5.91

                               No   Yes
ORGANIZATION_TYPE_Security
0                           91.95  8.05
1                           90.02  9.98

                                          No   Yes
ORGANIZATION_TYPE_Security Ministries
0                                      91.91  8.09
1                                      95.14  4.86

                                    No    Yes
ORGANIZATION_TYPE_Self-employed
0                                92.23   7.77
1                                89.83  10.17

                               No   Yes
ORGANIZATION_TYPE_Services
0                           91.92  8.08
1                           93.40  6.60

                              No   Yes
```

```
ORGANIZATION_TYPE_Telecom
0                          91.93  8.07
1                          92.37  7.63

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 1
0                                91.93  8.07
1                                91.09  8.91

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 2
0                                91.92  8.08
1                                93.00  7.00

                                    No    Yes
ORGANIZATION_TYPE_Trade: type 3
0                                91.95   8.05
1                                89.66  10.34

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 4
0                                91.93  8.07
1                                96.88  3.12

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 5
0                                91.93  8.07
1                                93.88  6.12

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 6
0                                91.92  8.08
1                                95.40  4.60

                                    No   Yes
ORGANIZATION_TYPE_Trade: type 7
0                                91.96  8.04
1                                90.55  9.45

                                        No   Yes
ORGANIZATION_TYPE_Transport: type 1
0                                    91.92  8.08
1                                    95.52  4.48

                                        No   Yes
ORGANIZATION_TYPE_Transport: type 2
0                                    91.93  8.07
1                                    92.20  7.80
```

```
                                        No    Yes
ORGANIZATION_TYPE_Transport: type 3
0                                    91.96   8.04
1                                    84.25  15.75

                                        No   Yes
ORGANIZATION_TYPE_Transport: type 4
0                                    91.95  8.05
1                                    90.72  9.28

                                 No   Yes
ORGANIZATION_TYPE_University
0                             91.91  8.09
1                             95.10  4.90

             No    Yes
GENDER_F
0         89.86  10.14
1         93.00   7.00

             No    Yes
GENDER_M
0         93.00   7.00
1         89.86  10.14


                       No   Yes
NAME_CONTRACT_TYPE
Cash loans          91.65  8.35
Revolving loans     94.52  5.48

                    No   Yes
FLAG_OWN_REALTY
N                91.68  8.32
Y                92.04  7.96

                    No   Yes
NAME_TYPE_SUITE
Children         92.62  7.38
Family           92.51  7.49
Group of people  91.51  8.49
Other_A          91.22  8.78
Other_B          90.17  9.83
Spouse, partner  92.13  7.87
Unaccompanied    91.82  8.18

                          No    Yes
```

```
NAME_INCOME_TYPE
Businessman           100.00   0.00
Commercial associate   92.52   7.48
Maternity leave        60.00  40.00
Pensioner              94.61   5.39
State servant          94.25   5.75
Student               100.00   0.00
Unemployed             63.64  36.36
Working                90.41   9.59

                          No   Yes
NAME_FAMILY_STATUS
Civil marriage         90.06  9.94
Married                92.44  7.56
Separated              91.81  8.19
Single / not married   90.19  9.81
Unknown               100.00  0.00
Widow                  94.18  5.82

                        No    Yes
NAME_HOUSING_TYPE
Co-op apartment      92.07   7.93
House / apartment    92.20   7.80
Municipal apartment  91.46   8.54
Office apartment     93.43   6.57
Rented apartment     87.69  12.31
With parents         88.30  11.70

                          No    Yes
OCCUPATION_TYPE
Accountants            95.17   4.83
Cleaning staff         90.39   9.61
Cooking staff          89.56  10.44
Core staff             93.70   6.30
Drivers                88.67  11.33
HR staff               93.61   6.39
High skill tech staff  93.84   6.16
IT staff               93.54   6.46
Laborers               89.42  10.58
Low-skill Laborers     82.85  17.15
Managers               93.79   6.21
Medicine staff         93.30   6.70
Private service staff  93.40   6.60
Realty agents          92.14   7.86
Sales staff            90.37   9.63
Secretaries            92.95   7.05
Security staff         89.26  10.74
Waiters/barmen staff   88.72  11.28
```

```
                          No   Yes
FONDKAPREMONT_MODE
not specified          92.46  7.54
org spec account       94.18  5.82
reg oper account       93.02  6.98
reg oper spec account  93.44  6.56

                     No    Yes
HOUSETYPE_MODE
block of flats    93.06   6.94
specific housing  89.86  10.14
terraced house    91.50   8.50

                       No   Yes
WALLSMATERIAL_MODE
Block               92.98  7.02
Mixed               92.47  7.53
Monolithic          95.28  4.72
Others              91.69  8.31
Panel               93.65  6.35
Stone, brick        92.59  7.41
Wooden              90.30  9.70
```

The function for chi sqare test calculation was created and run for contingency tables in the dictionaries.

p-values indicating if the differences between proportions in the two groups are significant were calculated and assesed by the condition that values lower than 0.05 indicate significant differences. Names of variables were saved in lists of significant or insignificant and printed in the end.

```
Confidence level - 0.99:

Pearson chi square test:293.151
P_value: 0.0

With regard to the variable NAME_CONTRACT_TYPE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:11.576
P_value: 0.001

With regard to the variable FLAG_OWN_REALTY,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:0.0
P_value: 1.0

With regard to the variable FLAG_MOBIL,there are no statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:649.751
P_value: 0.0

With regard to the variable FLAG_EMP_PHONE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:249.94
P_value: 0.0

With regard to the variable FLAG_WORK_PHONE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.017
P_value: 0.898

With regard to the variable FLAG_CONT_MOBILE,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:174.084
P_value: 0.0

With regard to the variable FLAG_PHONE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.923
P_value: 0.337

With regard to the variable FLAG_EMAIL,there are no statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).
```

```
Pearson chi square test:9.394
P_value: 0.002

With regard to the variable REG_REGION_NOT_LIVE_REGION,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:14.703
P_value: 0.0

With regard to the variable REG_REGION_NOT_WORK_REGION,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:2.392
P_value: 0.122

With regard to the variable LIVE_REGION_NOT_WORK_REGION,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:605.482
P_value: 0.0

With regard to the variable REG_CITY_NOT_LIVE_CITY,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:799.218
P_value: 0.0

With regard to the variable REG_CITY_NOT_WORK_CITY,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:324.864
P_value: 0.0

With regard to the variable LIVE_CITY_NOT_WORK_CITY,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:0.0
P_value: 1.0

With regard to the variable ORGANIZATION_TYPE_Advertising,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:18.873
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Agriculture,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:28.005
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Bank,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.027
P_value: 0.87

With regard to the variable ORGANIZATION_TYPE_Business Entity Type 1,there are
no statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:2.992
P_value: 0.084

With regard to the variable ORGANIZATION_TYPE_Business Entity Type 2,there are
no statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:176.804
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Business Entity Type 3,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:2.926
P_value: 0.087

With regard to the variable ORGANIZATION_TYPE_Cleaning,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:119.961
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Construction,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:2.945
P_value: 0.086

With regard to the variable ORGANIZATION_TYPE_Culture,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:2.476
P_value: 0.116

With regard to the variable ORGANIZATION_TYPE_Electricity,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.534
P_value: 0.465

With regard to the variable ORGANIZATION_TYPE_Emergency,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:17.239
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Government,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:3.355
P_value: 0.067

With regard to the variable ORGANIZATION_TYPE_Hotel,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.05
P_value: 0.823

With regard to the variable ORGANIZATION_TYPE_Housing,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:12.203
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Industry: type 1,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.209
P_value: 0.648

With regard to the variable ORGANIZATION_TYPE_Industry: type 10,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:1.163
P_value: 0.281

With regard to the variable ORGANIZATION_TYPE_Industry: type 11,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:8.546
P_value: 0.003

With regard to the variable ORGANIZATION_TYPE_Industry: type 12,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:1.922
P_value: 0.166

With regard to the variable ORGANIZATION_TYPE_Industry: type 13,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.356
P_value: 0.551

With regard to the variable ORGANIZATION_TYPE_Industry: type 2,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:28.535
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Industry: type 3,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:4.828
P_value: 0.028

With regard to the variable ORGANIZATION_TYPE_Industry: type 4,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:1.06
P_value: 0.303

With regard to the variable ORGANIZATION_TYPE_Industry: type 5,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.035
P_value: 0.851

With regard to the variable ORGANIZATION_TYPE_Industry: type 6,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).
```

```
Pearson chi square test:0.0
P_value: 0.999

With regard to the variable ORGANIZATION_TYPE_Industry: type 7,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.178
P_value: 0.673

With regard to the variable ORGANIZATION_TYPE_Industry: type 8,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:8.707
P_value: 0.003

With regard to the variable ORGANIZATION_TYPE_Industry: type 9,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:4.242
P_value: 0.039

With regard to the variable ORGANIZATION_TYPE_Insurance,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:10.075
P_value: 0.002

With regard to the variable ORGANIZATION_TYPE_Kindergarten,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.001
P_value: 0.979

With regard to the variable ORGANIZATION_TYPE_Legal Services,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).
```

```
Pearson chi square test:34.468
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Medicine,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:30.705
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Military,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.36
P_value: 0.548

With regard to the variable ORGANIZATION_TYPE_Mobile,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:4.342
P_value: 0.037

With regard to the variable ORGANIZATION_TYPE_Other,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:29.641
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Police,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:0.342
P_value: 0.559

With regard to the variable ORGANIZATION_TYPE_Postal,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).
```

```
Pearson chi square test:3.095
P_value: 0.079

With regard to the variable ORGANIZATION_TYPE_Realtor,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.294
P_value: 0.588

With regard to the variable ORGANIZATION_TYPE_Religion,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:31.916
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Restaurant,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:57.175
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_School,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:15.799
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Security,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:27.146
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Security Ministries,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:260.775
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Self-employed,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:4.411
P_value: 0.036

With regard to the variable ORGANIZATION_TYPE_Services,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.101
P_value: 0.75

With regard to the variable ORGANIZATION_TYPE_Telecom,there are no statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.224
P_value: 0.636

With regard to the variable ORGANIZATION_TYPE_Trade: type 1,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:2.822
P_value: 0.093

With regard to the variable ORGANIZATION_TYPE_Trade: type 2,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:24.11
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Trade: type 3,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:1.498
P_value: 0.221

With regard to the variable ORGANIZATION_TYPE_Trade: type 4,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.057
P_value: 0.811

With regard to the variable ORGANIZATION_TYPE_Trade: type 5,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:9.836
P_value: 0.002

With regard to the variable ORGANIZATION_TYPE_Trade: type 6,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:20.334
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Trade: type 7,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:3.035
P_value: 0.081

With regard to the variable ORGANIZATION_TYPE_Transport: type 1,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).

Pearson chi square test:0.181
P_value: 0.67

With regard to the variable ORGANIZATION_TYPE_Transport: type 2,there are no
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is not rejected).
```

```
Pearson chi square test:93.698
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_Transport: type 3,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:10.645
P_value: 0.001

With regard to the variable ORGANIZATION_TYPE_Transport: type 4,there are
statistically signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:17.672
P_value: 0.0

With regard to the variable ORGANIZATION_TYPE_University,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:919.814
P_value: 0.0

With regard to the variable GENDER_F,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:920.104
P_value: 0.0

With regard to the variable GENDER_M,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).



Significant statistical differences in groups of persons with payment
difficulties and those who do not have payment difficulties are for these
variables:
    ['NAME_CONTRACT_TYPE', 'FLAG_OWN_REALTY', 'FLAG_EMP_PHONE',
'FLAG_WORK_PHONE', 'FLAG_PHONE', 'REG_REGION_NOT_LIVE_REGION',
'REG_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY',
'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY',
'ORGANIZATION_TYPE_Agriculture', 'ORGANIZATION_TYPE_Bank',
'ORGANIZATION_TYPE_Business Entity Type 3', 'ORGANIZATION_TYPE_Construction',
'ORGANIZATION_TYPE_Government', 'ORGANIZATION_TYPE_Industry: type 1',
'ORGANIZATION_TYPE_Industry: type 12', 'ORGANIZATION_TYPE_Industry: type 3',
'ORGANIZATION_TYPE_Industry: type 9', 'ORGANIZATION_TYPE_Kindergarten',
'ORGANIZATION_TYPE_Medicine', 'ORGANIZATION_TYPE_Military',
'ORGANIZATION_TYPE_Police', 'ORGANIZATION_TYPE_Restaurant',
'ORGANIZATION_TYPE_School', 'ORGANIZATION_TYPE_Security',
'ORGANIZATION_TYPE_Security Ministries', 'ORGANIZATION_TYPE_Self-employed',
'ORGANIZATION_TYPE_Trade: type 3', 'ORGANIZATION_TYPE_Trade: type 6',
'ORGANIZATION_TYPE_Trade: type 7', 'ORGANIZATION_TYPE_Transport: type 3',
'ORGANIZATION_TYPE_Transport: type 4', 'ORGANIZATION_TYPE_University',
'GENDER_F', 'GENDER_M'])
Unsignificant statistical differences in groups of of persons with payment
difficulties and those who do not have payment difficulties are for these
variables:
    ['FLAG_MOBIL', 'FLAG_CONT_MOBILE', 'FLAG_EMAIL',
'LIVE_REGION_NOT_WORK_REGION', 'ORGANIZATION_TYPE_Advertising',
'ORGANIZATION_TYPE_Business Entity Type 1', 'ORGANIZATION_TYPE_Business Entity
Type 2', 'ORGANIZATION_TYPE_Cleaning', 'ORGANIZATION_TYPE_Culture',
'ORGANIZATION_TYPE_Electricity', 'ORGANIZATION_TYPE_Emergency',
'ORGANIZATION_TYPE_Hotel', 'ORGANIZATION_TYPE_Housing',
'ORGANIZATION_TYPE_Industry: type 10', 'ORGANIZATION_TYPE_Industry: type 11',
'ORGANIZATION_TYPE_Industry: type 13', 'ORGANIZATION_TYPE_Industry: type 2',
'ORGANIZATION_TYPE_Industry: type 4', 'ORGANIZATION_TYPE_Industry: type 5',
'ORGANIZATION_TYPE_Industry: type 6', 'ORGANIZATION_TYPE_Industry: type 7',
'ORGANIZATION_TYPE_Industry: type 8', 'ORGANIZATION_TYPE_Insurance',
'ORGANIZATION_TYPE_Legal Services', 'ORGANIZATION_TYPE_Mobile',
'ORGANIZATION_TYPE_Other', 'ORGANIZATION_TYPE_Postal',
'ORGANIZATION_TYPE_Realtor', 'ORGANIZATION_TYPE_Religion',
'ORGANIZATION_TYPE_Services', 'ORGANIZATION_TYPE_Telecom',
'ORGANIZATION_TYPE_Trade: type 1', 'ORGANIZATION_TYPE_Trade: type 2',
'ORGANIZATION_TYPE_Trade: type 4', 'ORGANIZATION_TYPE_Trade: type 5',
'ORGANIZATION_TYPE_Transport: type 1', 'ORGANIZATION_TYPE_Transport: type 2'])



Confidence level - 0.99:

Pearson chi square test:293.151
P_value: 0.0

With regard to the variable NAME_CONTRACT_TYPE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:11.576
P_value: 0.001

With regard to the variable FLAG_OWN_REALTY,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:32.825
P_value: 0.0

With regard to the variable NAME_TYPE_SUITE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:1253.471
P_value: 0.0

With regard to the variable NAME_INCOME_TYPE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:504.694
P_value: 0.0

With regard to the variable NAME_FAMILY_STATUS,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:420.556
P_value: 0.0

With regard to the variable NAME_HOUSING_TYPE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:1402.847
P_value: 0.0

With regard to the variable OCCUPATION_TYPE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).
```

```
Pearson chi square test:16.81
P_value: 0.001

With regard to the variable FONDKAPREMONT_MODE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:27.633
P_value: 0.0

With regard to the variable HOUSETYPE_MODE,there are statistically signifficant
differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).

Pearson chi square test:139.235
P_value: 0.0

With regard to the variable WALLSMATERIAL_MODE,there are statistically
signifficant differences among
groups of persons with payment difficulties and those who do not have payment
difficulties (the H0 hypothesis is rejected).



Significant statistical differences in groups of persons with payment
difficulties and those who do not have payment difficulties are for these
variables:
    ['NAME_CONTRACT_TYPE', 'FLAG_OWN_REALTY', 'NAME_TYPE_SUITE',
'NAME_INCOME_TYPE', 'NAME_FAMILY_STATUS', 'NAME_HOUSING_TYPE',
'OCCUPATION_TYPE', 'FONDKAPREMONT_MODE', 'HOUSETYPE_MODE',
'WALLSMATERIAL_MODE'])
Unsignificant statistical differences in groups of of persons with payment
difficulties and those who do not have payment difficulties are for these
variables:
    [])
```

It was checked which proportion is higher for each variable (the one in the “yes” group or the one in the “no” group), and variable names were appended to separate lists “yes_list” and “no_list”. Those variable names were selected which were also present in the list “significant” in order to find out which variables have statistically significant higher proportions in “yes” and “no” groups (clients who have and do not have laon payment difficulties).

```
For these variables there are statistically significant
higher proportions in the <yes> group (persons having loan payment
difficulties):

 ['FLAG_EMP_PHONE', 'FLAG_WORK_PHONE', 'REG_REGION_NOT_LIVE_REGION',
'REG_REGION_NOT_WORK_REGION', 'REG_CITY_NOT_LIVE_CITY',
'REG_CITY_NOT_WORK_CITY', 'LIVE_CITY_NOT_WORK_CITY',
'ORGANIZATION_TYPE_Agriculture', 'ORGANIZATION_TYPE_Business Entity Type 3',
'ORGANIZATION_TYPE_Construction', 'ORGANIZATION_TYPE_Industry: type 1',
'ORGANIZATION_TYPE_Industry: type 3', 'ORGANIZATION_TYPE_Restaurant',
'ORGANIZATION_TYPE_Security', 'ORGANIZATION_TYPE_Self-employed',
'ORGANIZATION_TYPE_Trade: type 3', 'ORGANIZATION_TYPE_Trade: type 7',
'ORGANIZATION_TYPE_Transport: type 3', 'ORGANIZATION_TYPE_Transport: type 4',
'GENDER_M']

For these variables there are statistically significant
higher proportions in the <no> group (persons not having loan payment
difficulties):

 ['NAME_CONTRACT_TYPE', 'FLAG_OWN_REALTY', 'FLAG_PHONE',
'ORGANIZATION_TYPE_Bank', 'ORGANIZATION_TYPE_Government',
'ORGANIZATION_TYPE_Industry: type 12', 'ORGANIZATION_TYPE_Industry: type 9',
'ORGANIZATION_TYPE_Kindergarten', 'ORGANIZATION_TYPE_Medicine',
'ORGANIZATION_TYPE_Military', 'ORGANIZATION_TYPE_Police',
'ORGANIZATION_TYPE_School', 'ORGANIZATION_TYPE_Security Ministries',
'ORGANIZATION_TYPE_Trade: type 6', 'ORGANIZATION_TYPE_University', 'GENDER_F']
```

1.2.3 Conclusions for the exploratory analysis part

From this analysis of numerical variables (Mann Whitney U test) it can be concluded that clients that likely will experience loan payment difficulties are those who: - have higher numbers of children; - are of older age; - have been living longer in the same area; - have not changed their id document for a longer time; - live in a region with a rating of higher number (rather the region 3 than the region 1); - live in a region with a rating of higher number (rather the region 3 than the region 1) taking city into account; - the living conditions of the factor 2 of the clients have higher scores (e.g. have older houses);

It is more likely that clients will be paying loans in time it: - The have higher income; - they took credits of higher amount; - live in more populated regions; - have better education; - the living conditions of the factor 2 of the clients have higher scores (e.g have houses of with longer periods ofexploitation); - numbers of the Credit Bureau enquiries about the person of teh factor 2 are higher (i.e., more enquiries during the last quartier).

It can be observed from the analysis of binary and categorical variables (chi-square tests) that, for the confidence level 0.95, persons that will more likely have loan payment difficulties are those who: - take cash loans; - own real estate; - are on maternity leave or unemployed; - provided home and work phone numbers; - their permanent adress does not match contact or work adresses in region or city levels; - work in agriculture, business entity (type 3), industry (type 1,11, 13, 3, 4, 8), construction, cleaning, mobile, postal, realtor, restaurant, security, trade (type 1, 3, 7), transport (type 3, 4); - are self_employed; - are in civil mariage or single/ not married; - were unaccompanied or accompanied by a group of people when applying for a loan; - live in rented apartment or with parents; - work as low-skill laborers, laborers, drivers, security staff, waiters/ barmen staff, cooking staff (percentages higher than 10 percent in the “No” group); - live in specific housing, walls are wooden (percentages higher than 9 percent in the “Yes” group); - are men.

for the confidence level 0.95, persons that will more likely not have loan payment difficulties are those who: - take revolving loans; - own real estate; - provided their mobile phone number; - work for a bank, the government, industry (type 12, 9), kindergarden, medicine, military, police, school, security ministries, trade (type 6), university; - are married or widows; - work as core staff, accountants, medicine staff, managers, private service staff, high skill tech staff, hr staff (percentages less than 7 percent in the “Yes” group; - live in monolithic housing (percentages lower than 5 in the “Yes” group); - are women.