Wrangling Home Credit data set
Discussion

P. A. Ortiz Otalvaro

26 September 2019

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

1. Initial exploration of data sets

Goals of the initial exploration:

- Get a good idea of what the data is all about
- Define the criteria to delimite in the following steps in the data wrangling process.

1.1. General description of data files

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

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

- This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
- One row represents one loan in the data sample.
- For each loan there are 121 features describing the customer as well as the loan.

2. *bureau.csv*

- All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
- For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.

3. *bureau_balance.csv*

- Monthly balances of previous credits in Credit Bureau.
- This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample * # of relative previous credits * # of months where we have some history observable for the previous credits) rows.

4. *POS_CASH_balance.csv*

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

5. *credit_card_balance.csv*

- Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.

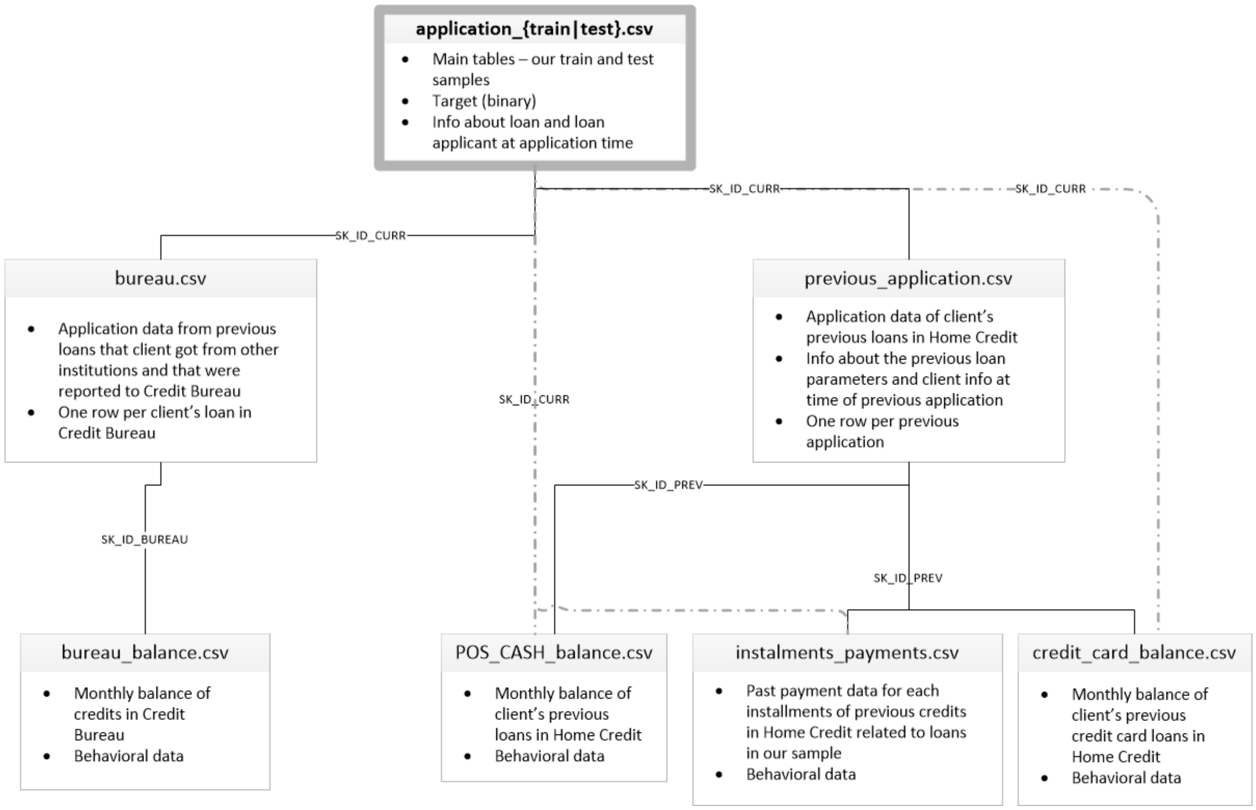


Figure 1: Connections between all data sets

- This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample * # of relative previous credit cards * # of months where we have some history observable for the previous credit card) rows.

6. *previous_application.csv*

- All previous applications for Home Credit loans of clients who have loans in our sample.
- There is one row for each previous application related to loans in our data sample.

7. *installments_payments.csv*

- Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
- There is
  a) one row for every payment that was made plus
  b) one row each for missed payment.
- One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.

8. *HomeCredit_columns_description.csv*

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

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

1.2. Exploration of data types

Variables types: general picture

Let's first get a general picture of all data sets by extracting: the number of variables and observations as well as the number of character, factor, and numeric variables in each file. This is all show in table 1 together with the total number of NAs in each data set.

Table 1: Table 1. General characteristics of data in all data sets.

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

Separating the features into character and numeric does not totally give useful insights on the values contained in each column. Therefore, to get even a better grasp on how the data set looks like, the following step will be to divide the variables according to topics

Initially I will perform this exploration on the main data set: application_train. Later on, I will explore the remaining data sets.

Variable types: Subsetting features according to topic

Application_train set: subsetting features according to topic

As mentioned before, each loan has 121 features describing customer and loan. Variables can be separated in the categories:

a) Loan data

Feature name	column number
Contract type	3
Amount of loan	9
Loan annuity	10
price of goods (loan purpose)	11
Client's companion during loan application	12
Weekday of loan application	33
Hour of loan application	34

b) Client's personal information

Feature name	column number
Gender	4
Level of education	14
Age	18
ID expedition time	21
Did client provide mobile phone? (flag)	23

Feature name	column number
Did client provide employer phone? (flag)	24
Did client provide work phone? (flag)	25
Was mobile phone reachable (flag)	26
Did client provide home phone? (flag)	27
Did client provide email?	28
How many days befor applicaiton client changed phone	96

c) Client's work information

Feature name	column number
Total income of client	8
Clients income type (businessman, working, maternity leave,...)	13
Number of days in current employment	19
Client's occupation	29
Type of organization where client works	41

d) Client's properties

Feature name	column number
Does client own a car? (flag)	5
Does client own a house or flat? (flag)	6
Age of client's car	22

e) Previous Credit history from Credit Bureau

Feature name	column number
Number of enqueries about the client at different times before application	117 : 122

f) Client's family details

Feature name	column number
Number of children	7
Family status	15
Number of family members	30

g) Client's social circle

Feature name	column number
Observations of client's social surroundings that defaulted	92 : 95

h) Housing

Feature name	column number
type of housing	16
population of housing region	17
rating of housing region	31, 32
Do contact, work and permanent addresses match?	35 : 40
apartment size	45, 59, 73
basement area	46, 60, 74
age of building	47, 48, 61, 62, 75, 76
common area	49, 63, 77
number of elevators	50, 64, 78
number of entrances	51, 65, 79
number of floors	52, 53, 66, 67, 80, 81
land area	54, 68, 82
living area/aparments	55, 56, 69, 70, 83, 84
nonliving area/apartm	57, 58, 71, 72, 85, 86
Type of housing	88
Total area	89
Walls material	90
Emergency state	91
?	87

i) Loan paperwork

Feature name	column number
Did client provide document 2	97
Did client provide document 3	98
Did client provide document 4	99
Did client provide document 5	100
Did client provide document 6	101
Did client provide document 7	102
Did client provide document 8	103
Did client provide document 9	104
Did client provide document 10	105
Did client provide document 11	106
Did client provide document 12	107
Did client provide document 13	108
Did client provide document 14	109
Did client provide document 15	110
Did client provide document 16	111
Did client provide document 17	112
Did client provide document 18	113
Did client provide document 19	114
Did client provide document 20	115

j) others

I do not know exactly what these features are about

Feature name	column number	Explanation in data set
DAYS_REGISTRATION	20	How many days before the application did client change his registration
EXT_SOURCE_1	42	Normalized score from external data source,normalized
EXT_SOURCE_2	43	Normalized score from external data source,normalized

2. Structuring

Restructure the data in a manner that better suits the following analysis

Column names

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

Order of columns (group)

For simplicity in future analysis the columns are reordered according to their corresponding category (topic as described earlier) as follows:

1. ID of applicant
2. Target
3. Loan
4. Paperwork
5. Personal (including contact details)
6. Work related
7. Properties (belongings) of client
8. Previous credit history (from Credit Bureau)
9. Housing
10. Columns which meaning is not fully clear

3. Cleaning

3.1. Subsetting features into categorical and non-categorical

To clean the data, an even more detailed exploration is needed. This can be done by dividing the variables into categorical and non-categorical and finding their distributions and patterns. *"In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property."*

Application_train set: subsetting features into categorical and non-categorical

Categorical columns:

	Col	Unique	Unique values
TARGET	2	2	1, 0
NAME_CONTRACT_TYPE	3	2	Cash loans, Revolving loans
CODE_GENDER	4	3	M, F, XNA
FLAG_OWN_CAR	5	2	N, Y
FLAG_OWN_REALTY	6	2	Y, N
CNT_CHILDREN	7	15	0, 1, 2, 3, 4, 7, 5, 6, 8, 9, 11, 12, 10, 19, 14
NAME_TYPE_SUITE	12	8	Unaccompanied, Family, Spouse, partner, Children, Other_A
NAME_INCOME_TYPE	13	8	Working, State servant, Commercial associate, Pensioner, Un
NAME_EDUCATION_TYPE	14	5	Secondary / secondary special, Higher education, Incomplete
NAME_FAMILY_STATUS	15	6	Single / not married, Married, Civil marriage, Widow, Separa
NAME_HOUSING_TYPE	16	6	House / apartment, Rented apartment, With parents, Munici
FLAG_MOBIL	23	2	1, 0
FLAG_EMP_PHONE	24	2	1, 0
FLAG_WORK_PHONE	25	2	0, 1
FLAG_CONT_MOBILE	26	2	1, 0
FLAG_PHONE	27	2	1, 0
FLAG_EMAIL	28	2	0, 1
OCCUPATION_TYPE	29	19	Laborers, Core staff, Accountants, Managers, NA, Drivers, Sa
CNT_FAM_MEMBERS	30	18	1, 2, 3, 4, 5, 6, 9, 7, 8, 10, 13, NA, 14, 12, 20, 15, 16, 11
REGION_RATING_CLIENT	31	3	2, 1, 3
REGION_RATING_CLIENT_W_CITY	32	3	2, 1, 3
WEEKDAY_APPR_PROCESS_START	33	7	WEDNESDAY, MONDAY, THURSDAY, SUNDAY, SATUR
HOUR_APPR_PROCESS_START	34	24	10, 11, 9, 17, 16, 14, 8, 15, 7, 13, 6, 12, 19, 3, 18, 21, 4, 5, 20,
REG_REGION_NOT_LIVE_REGION	35	2	0, 1
REG_REGION_NOT_WORK_REGION	36	2	0, 1
LIVE_REGION_NOT_WORK_REGION	37	2	0, 1
REG_CITY_NOT_LIVE_CITY	38	2	0, 1
REG_CITY_NOT_WORK_CITY	39	2	0, 1
LIVE_CITY_NOT_WORK_CITY	40	2	0, 1
ORGANIZATION_TYPE	41	58	Business Entity Type 3, School, Government, Religion, Other
FONDKAPREMONT_MODE	87	5	reg oper account, NA, org spec account, reg oper spec accoun
HOUSETYPE_MODE	88	4	block of flats, NA, terraced house, specific housing
WALLSMATERIAL_MODE	90	8	Stone, brick, Block, NA, Panel, Mixed, Wooden, Others, Mon
EMERGENCYSTATE_MODE	91	3	No, NA, Yes
OBS_30_CNT_SOCIAL_CIRCLE	92	34	2, 1, 0, 4, 8, 10, NA, 7, 3, 6, 5, 12, 9, 13, 11, 14, 22, 16, 15, 17
DEF_30_CNT_SOCIAL_CIRCLE	93	11	2, 0, 1, NA, 3, 4, 5, 6, 7, 34, 8
OBS_60_CNT_SOCIAL_CIRCLE	94	34	2, 1, 0, 4, 8, 10, NA, 7, 3, 6, 5, 12, 9, 13, 11, 14, 21, 15, 22, 16
DEF_60_CNT_SOCIAL_CIRCLE	95	10	2, 0, 1, NA, 3, 5, 4, 7, 24, 6
FLAG_DOCUMENT_2	97	2	0, 1

	Col	Unique	Unique values
FLAG_DOCUMENT_3	98	2	1, 0
FLAG_DOCUMENT_4	99	2	0, 1
FLAG_DOCUMENT_5	100	2	0, 1
FLAG_DOCUMENT_6	101	2	0, 1
FLAG_DOCUMENT_7	102	2	0, 1
FLAG_DOCUMENT_8	103	2	0, 1
FLAG_DOCUMENT_9	104	2	0, 1
FLAG_DOCUMENT_10	105	2	0, 1
FLAG_DOCUMENT_11	106	2	0, 1
FLAG_DOCUMENT_12	107	2	0, 1
FLAG_DOCUMENT_13	108	2	0, 1
FLAG_DOCUMENT_14	109	2	0, 1
FLAG_DOCUMENT_15	110	2	0, 1
FLAG_DOCUMENT_16	111	2	0, 1
FLAG_DOCUMENT_17	112	2	0, 1
FLAG_DOCUMENT_18	113	2	0, 1
FLAG_DOCUMENT_19	114	2	0, 1
FLAG_DOCUMENT_20	115	2	0, 1
FLAG_DOCUMENT_21	116	2	0, 1
AMT_REQ_CREDIT_BUREAU_HOUR	117	6	0, NA, 1, 2, 3, 4
AMT_REQ_CREDIT_BUREAU_DAY	118	10	0, NA, 1, 3, 2, 4, 5, 6, 9, 8
AMT_REQ_CREDIT_BUREAU_WEEK	119	10	0, NA, 1, 3, 2, 4, 5, 6, 8, 7
AMT_REQ_CREDIT_BUREAU_MON	120	25	0, NA, 1, 2, 6, 5, 3, 7, 9, 4, 11, 8, 16, 12, 14, 10, 13, 17, 24, 1
AMT_REQ_CREDIT_BUREAU_QRT	121	12	0, NA, 1, 2, 4, 3, 8, 5, 6, 7, 261, 19
AMT_REQ_CREDIT_BUREAU_YEAR	122	26	1, 0, NA, 2, 4, 5, 3, 8, 6, 9, 7, 10, 11, 13, 16, 12, 25, 23, 15, 1

Non-categorical columns:

General statistics of non-categorical features:

	Count	Min	Max	St Dev	Mean	Mode
AMT_INCOME_TOTAL	307511	25650.00	117000000.00	237123.15	168797.92	135000
AMT_CREDIT	307511	45000.00	4050000.00	402490.78	599026.00	450000
AMT_ANNUITY	307499	1615.50	258025.50	14493.74	27108.57	9000
AMT_GOODS_PRICE	307233	40500.00	4050000.00	369446.46	538396.21	450000
REGION_POPULATION_RELATIVE	307511	0.00	0.07	0.01	0.02	0.035792
YEARS_BIRTH	307511	20.52	69.12	11.96	43.94	36.79
YEARS_EMPLOYED	307511	-1000.67	49.07	387.06	-174.84	-1000.67
YEARS_REGISTRATION	307511	0.00	67.59	9.65	13.66	0.01
YEARS_ID_PUBLISH	307511	0.00	19.72	4.14	8.20	11.22
OWN_CAR_AGE	104582	0.00	91.00	11.94	12.06	7
EXT_SOURCE_1	134133	0.01	0.96	0.21	0.50	0.35632266441
EXT_SOURCE_2	306851	0.00	0.85	0.19	0.51	0.28589787214
EXT_SOURCE_3	246546	0.00	0.90	0.19	0.51	0.74630021305
APARTMENTS_AVG	151450	0.00	1.00	0.11	0.12	0.0825
BASEMENTAREA_AVG	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_AVG	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_AVG	103023	0.00	1.00	0.11	0.75	0.8232
COMMONAREA_AVG	92646	0.00	1.00	0.08	0.04	0
ELEVATORS_AVG	143620	0.00	1.00	0.13	0.08	0
ENTRANCES_AVG	152683	0.00	1.00	0.10	0.15	0.1379
FLOORSMAX_AVG	154491	0.00	1.00	0.14	0.23	0.1667

	Count	Min	Max	St Dev	Mean	Mode
FLOORSMIN_AVG	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_AVG	124921	0.00	1.00	0.08	0.07	0
LIVINGAPARTMENTS_AVG	97312	0.00	1.00	0.09	0.10	0.0504
LIVINGAREA_AVG	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_AVG	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_AVG	137829	0.00	1.00	0.07	0.03	0
APARTMENTS_MODE	151450	0.00	1.00	0.11	0.11	0.084
BASEMENTAREA_MODE	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_MODE	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_MODE	103023	0.00	1.00	0.11	0.76	0.8301
COMMONAREA_MODE	92646	0.00	1.00	0.07	0.04	0
ELEVATORS_MODE	143620	0.00	1.00	0.13	0.07	0
ENTRANCES_MODE	152683	0.00	1.00	0.10	0.15	0.1379
FLOORSMAX_MODE	154491	0.00	1.00	0.14	0.22	0.1667
FLOORSMIN_MODE	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_MODE	124921	0.00	1.00	0.08	0.06	0
LIVINGAPARTMENTS_MODE	97312	0.00	1.00	0.10	0.11	0.0551
LIVINGAREA_MODE	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_MODE	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_MODE	137829	0.00	1.00	0.07	0.03	0
APARTMENTS_MEDI	151450	0.00	1.00	0.11	0.12	0.0833
BASEMENTAREA_MEDI	127568	0.00	1.00	0.08	0.09	0
YEARS_BEGINEXPLUATATION_MEDI	157504	0.00	1.00	0.06	0.98	0.9871
YEARS_BUILD_MEDI	103023	0.00	1.00	0.11	0.76	0.8256
COMMONAREA_MEDI	92646	0.00	1.00	0.08	0.04	0
ELEVATORS_MEDI	143620	0.00	1.00	0.13	0.08	0
ENTRANCES_MEDI	152683	0.00	1.00	0.10	0.15	0.1379
FLOORSMAX_MEDI	154491	0.00	1.00	0.15	0.23	0.1667
FLOORSMIN_MEDI	98869	0.00	1.00	0.16	0.23	0.2083
LANDAREA_MEDI	124921	0.00	1.00	0.08	0.07	0
LIVINGAPARTMENTS_MEDI	97312	0.00	1.00	0.09	0.10	0.0513
LIVINGAREA_MEDI	153161	0.00	1.00	0.11	0.11	0
NONLIVINGAPARTMENTS_MEDI	93997	0.00	1.00	0.05	0.01	0
NONLIVINGAREA_MEDI	137829	0.00	1.00	0.07	0.03	0
TOTALAREA_MODE	159080	0.00	1.00	0.11	0.10	0
YEARS_LAST_PHONE_CHANGE	307510	0.00	11.76	2.27	2.64	0

After listing all the categorical variables, it is useful to plot them to see the proportions between the different categories of each variable. These plots are shown in Appendices 1 and 2.

From these summary tables and the plots it can be concluded that:

- There are far more loans that were repaid on time (TARGET=0) than loans that were not repaid (TARGET=1).
- Some of the features have a very considerable difference in the occurrence of its two criteria: one of the two criteria being observed 0.001% or less with respect to the total of observations (FLAG_MOBIL, FLAG_CONT_MOBIL, FLAG_DOCUMENT_2, FLAG_DOCUMENT_4, FLAG_DOCUMENT_10, FLAG_DOCUMENT_12).
- YEARS_EMPLOYED has bad measurements (bad observations): the minimum amount of years emplyed is negative and -1000. Unfortunately this is also the most ocurring value in this column (mode): 55374 times out of 307511 (total observations). This will be discussed further in section 4.1.

3.2. Formating

Columns with time

A few of the columns are given in days and to understand better what is in them (and if there are any outliers) it is handy to change it to years and as positive values. I changed: DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, and DAYS_LAST_PHONE_CHANGE. These variables in years are shown in tables of section 3.1.

Change variable types

I changed all categorical variables from integer or character type to factor type.

3.3. NAs

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

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

Missing values per column

Table 14: Missing values in Train data set

	MissingValues	percentage
COMMONAREA_AVG	214865	69.87
COMMONAREA_MODE	214865	69.87
COMMONAREA_MEDI	214865	69.87
NONLIVINGAPARTMENTS_AVG	213514	69.43
NONLIVINGAPARTMENTS_MODE	213514	69.43
NONLIVINGAPARTMENTS_MEDI	213514	69.43
FONDKAPREMONT_MODE	210295	68.39
LIVINGAPARTMENTS_AVG	210199	68.35
LIVINGAPARTMENTS_MODE	210199	68.35
LIVINGAPARTMENTS_MEDI	210199	68.35
FLOORSMIN_AVG	208642	67.85
FLOORSMIN_MODE	208642	67.85
FLOORSMIN_MEDI	208642	67.85
YEARS_BUILD_AVG	204488	66.50
YEARS_BUILD_MODE	204488	66.50
YEARS_BUILD_MEDI	204488	66.50
OWN_CAR_AGE	202929	65.99
LANDAREA_AVG	182590	59.38
LANDAREA_MODE	182590	59.38
LANDAREA_MEDI	182590	59.38
BASEMENTAREA_AVG	179943	58.52
BASEMENTAREA_MODE	179943	58.52
BASEMENTAREA_MEDI	179943	58.52
EXT_SOURCE_1	173378	56.38
NONLIVINGAREA_AVG	169682	55.18
NONLIVINGAREA_MODE	169682	55.18
NONLIVINGAREA_MEDI	169682	55.18
ELEVATORS_AVG	163891	53.30
ELEVATORS_MODE	163891	53.30
ELEVATORS_MEDI	163891	53.30
WALLSMATERIAL_MODE	156341	50.84

	MissingValues	percentage
APARTMENTS_AVG	156061	50.75
APARTMENTS_MODE	156061	50.75
APARTMENTS_MEDI	156061	50.75
ENTRANCES_AVG	154828	50.35
ENTRANCES_MODE	154828	50.35
ENTRANCES_MEDI	154828	50.35
LIVINGAREA_AVG	154350	50.19
LIVINGAREA_MODE	154350	50.19
LIVINGAREA_MEDI	154350	50.19
HOUSETYPE_MODE	154297	50.18
FLOORSMAX_AVG	153020	49.76
FLOORSMAX_MODE	153020	49.76
FLOORSMAX_MEDI	153020	49.76
YEARS_BEGINEXPLUATATION_AVG	150007	48.78
YEARS_BEGINEXPLUATATION_MODE	150007	48.78
YEARS_BEGINEXPLUATATION_MEDI	150007	48.78
TOTALAREA_MODE	148431	48.27
EMERGENCYSTATE_MODE	145755	47.40
OCCUPATION_TYPE	96391	31.35
EXT_SOURCE_3	60965	19.83
AMT_REQ_CREDIT_BUREAU_HOUR	41519	13.50
AMT_REQ_CREDIT_BUREAU_DAY	41519	13.50
AMT_REQ_CREDIT_BUREAU_WEEK	41519	13.50
AMT_REQ_CREDIT_BUREAU_MON	41519	13.50
AMT_REQ_CREDIT_BUREAU_QRT	41519	13.50
AMT_REQ_CREDIT_BUREAU_YEAR	41519	13.50
NAME_TYPE_SUITE	1292	0.42
OBS_30_CNT_SOCIAL_CIRCLE	1021	0.33
DEF_30_CNT_SOCIAL_CIRCLE	1021	0.33
OBS_60_CNT_SOCIAL_CIRCLE	1021	0.33
DEF_60_CNT_SOCIAL_CIRCLE	1021	0.33
EXT_SOURCE_2	660	0.21
AMT_GOODS_PRICE	278	0.09
AMT_ANNUITY	12	0.00
CNT_FAM_MEMBERS	2	0.00
YEARS_LAST_PHONE_CHANGE	1	0.00

Table 15: Missing values in Test data set

	MissingValues	percentage
COMMONAREA_AVG	33495	68.72
COMMONAREA_MODE	33495	68.72
COMMONAREA_MEDI	33495	68.72
NONLIVINGAPARTMENTS_AVG	33347	68.41
NONLIVINGAPARTMENTS_MODE	33347	68.41
NONLIVINGAPARTMENTS_MEDI	33347	68.41
FONDKAPREMONT_MODE	32797	67.28
LIVINGAPARTMENTS_AVG	32780	67.25
LIVINGAPARTMENTS_MODE	32780	67.25
LIVINGAPARTMENTS_MEDI	32780	67.25
FLOORSMIN_AVG	32466	66.61

	MissingValues	percentage
FLOORSMIN_MODE	32466	66.61
FLOORSMIN_MEDI	32466	66.61
OWN_CAR_AGE	32312	66.29
YEARS_BUILD_AVG	31818	65.28
YEARS_BUILD_MODE	31818	65.28
YEARS_BUILD_MEDI	31818	65.28
LANDAREA_AVG	28254	57.96
LANDAREA_MODE	28254	57.96
LANDAREA_MEDI	28254	57.96
BASEMENTAREA_AVG	27641	56.71
BASEMENTAREA_MODE	27641	56.71
BASEMENTAREA_MEDI	27641	56.71
NONLIVINGAREA_AVG	26084	53.51
NONLIVINGAREA_MODE	26084	53.51
NONLIVINGAREA_MEDI	26084	53.51
ELEVATORS_AVG	25189	51.68
ELEVATORS_MODE	25189	51.68
ELEVATORS_MEDI	25189	51.68
WALLSMATERIAL_MODE	23893	49.02
APARTMENTS_AVG	23887	49.01
APARTMENTS_MODE	23887	49.01
APARTMENTS_MEDI	23887	49.01
HOUSETYPE_MODE	23619	48.46
ENTRANCES_AVG	23579	48.37
ENTRANCES_MODE	23579	48.37
ENTRANCES_MEDI	23579	48.37
LIVINGAREA_AVG	23552	48.32
LIVINGAREA_MODE	23552	48.32
LIVINGAREA_MEDI	23552	48.32
FLOORSMAX_AVG	23321	47.84
FLOORSMAX_MODE	23321	47.84
FLOORSMAX_MEDI	23321	47.84
YEARS_BEGINEXPLUATATION_AVG	22856	46.89
YEARS_BEGINEXPLUATATION_MODE	22856	46.89
YEARS_BEGINEXPLUATATION_MEDI	22856	46.89
TOTALAREA_MODE	22624	46.41
EMERGENCYSTATE_MODE	22209	45.56
EXT_SOURCE_1	20532	42.12
OCCUPATION_TYPE	15605	32.01
EXT_SOURCE_3	8668	17.78
AMT_REQ_CREDIT_BUREAU_HOUR	6049	12.41
AMT_REQ_CREDIT_BUREAU_DAY	6049	12.41
AMT_REQ_CREDIT_BUREAU_WEEK	6049	12.41
AMT_REQ_CREDIT_BUREAU_MON	6049	12.41
AMT_REQ_CREDIT_BUREAU_QRT	6049	12.41
AMT_REQ_CREDIT_BUREAU_YEAR	6049	12.41
NAME_TYPE_SUITE	911	1.87
OBS_30_CNT_SOCIAL_CIRCLE	29	0.06
DEF_30_CNT_SOCIAL_CIRCLE	29	0.06
OBS_60_CNT_SOCIAL_CIRCLE	29	0.06
DEF_60_CNT_SOCIAL_CIRCLE	29	0.06
AMT_ANNUITY	24	0.05

	MissingValues	percentage
EXT_SOURCE_2	8	0.02

IMPORTANT:

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

4. Filter data

Select the features that are neeeded (remove non needed)

4.1. Unnecessary columns

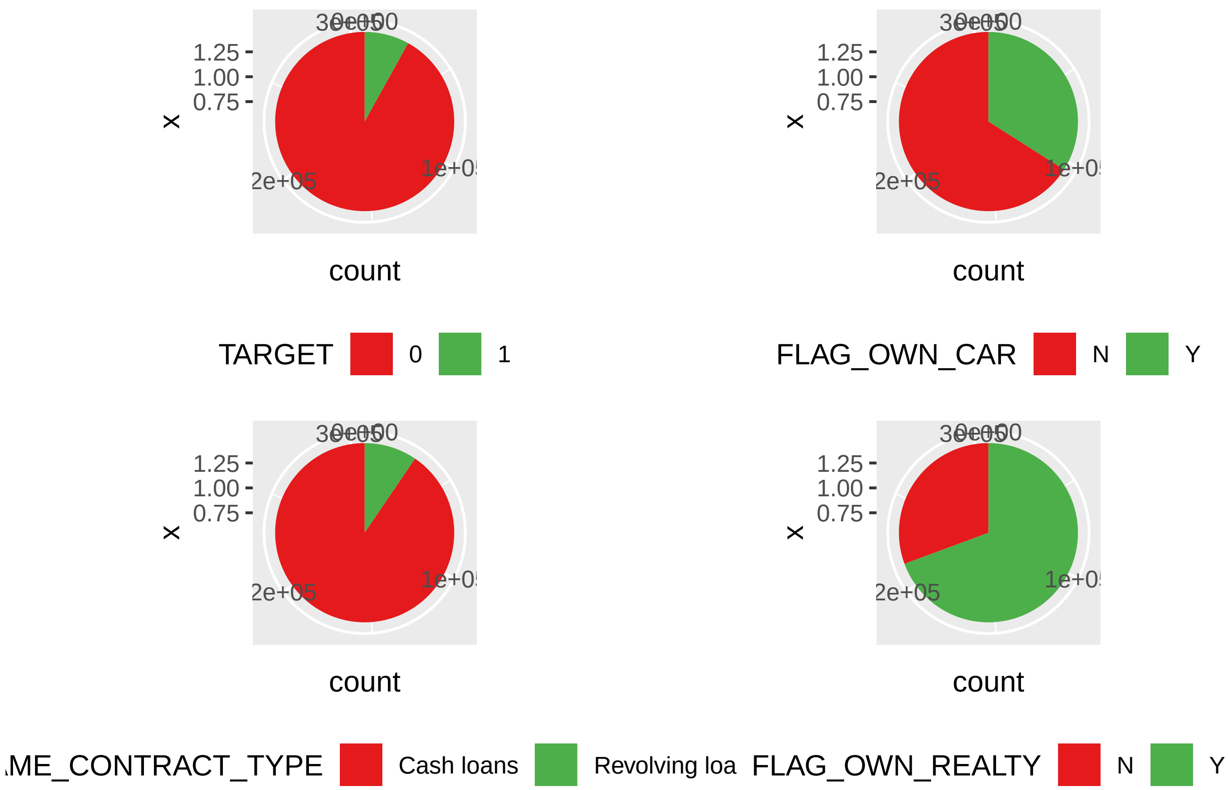
This may be required to be done after the exploration data analysis is done and after feature engeneer is done.

4.2. Bad data

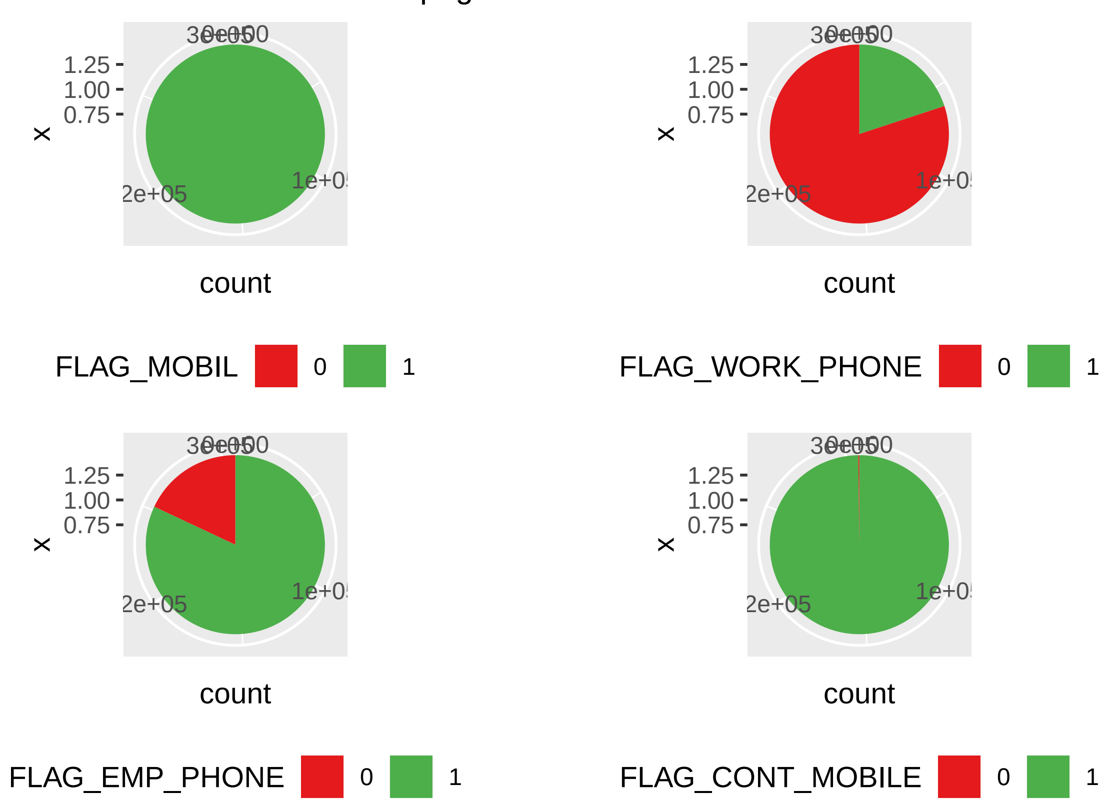
As discussed earlier in section 3.1. YEARS_EMPLOYED has a value that is not a sensitive time and so it needs to be replaced. For now I changed it into NaN.

Appendix 1: Pie charts of categorical variables (with 2 categories)

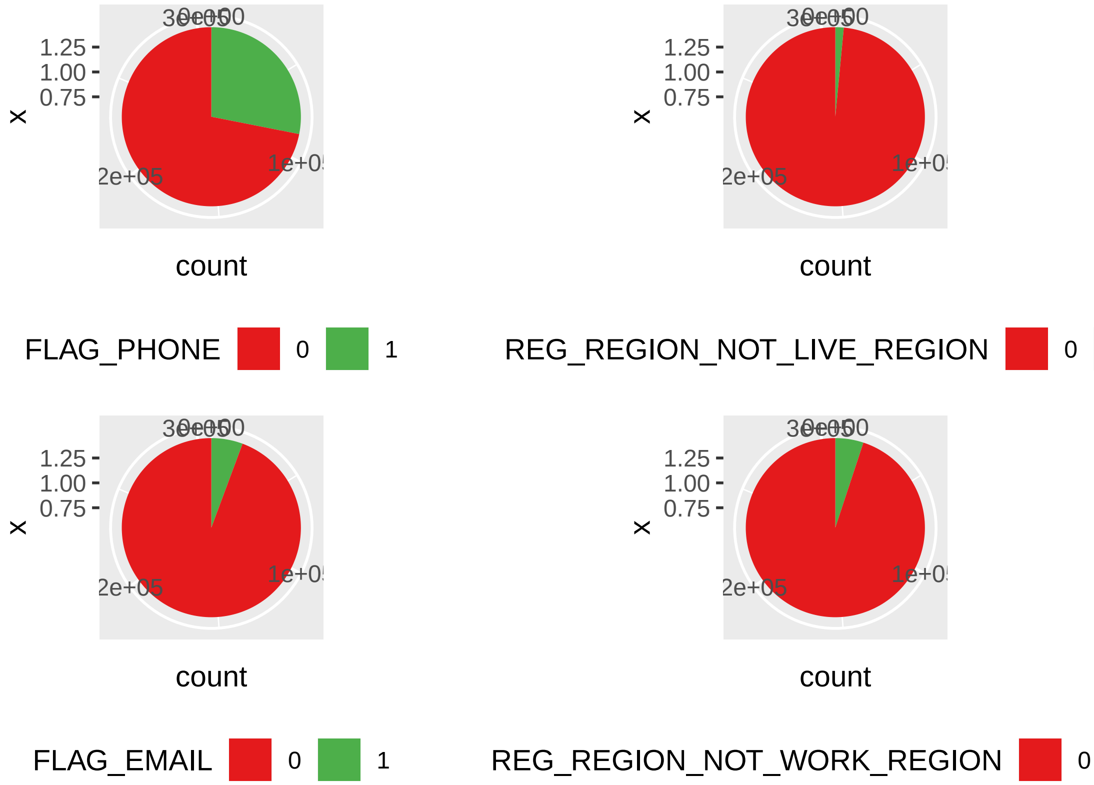
page 1 of 9



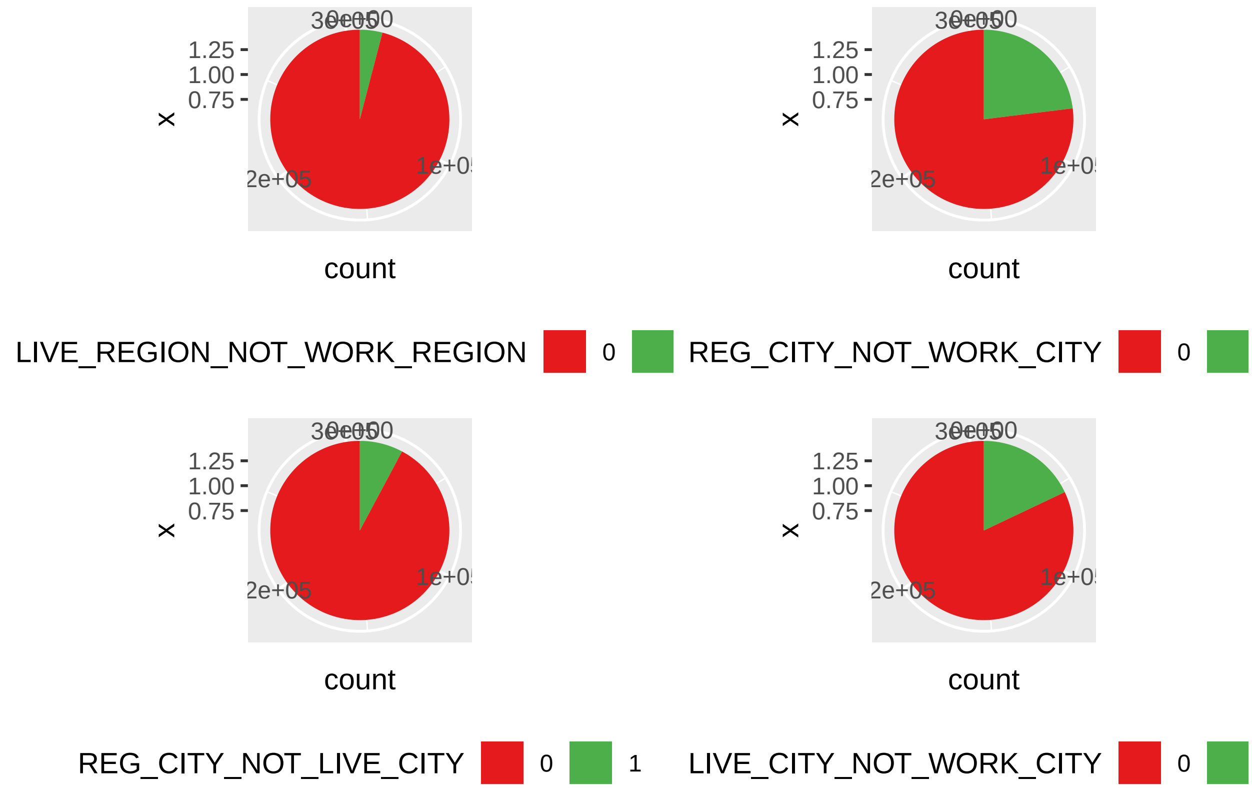
page 2 of 9



page 3 of 9

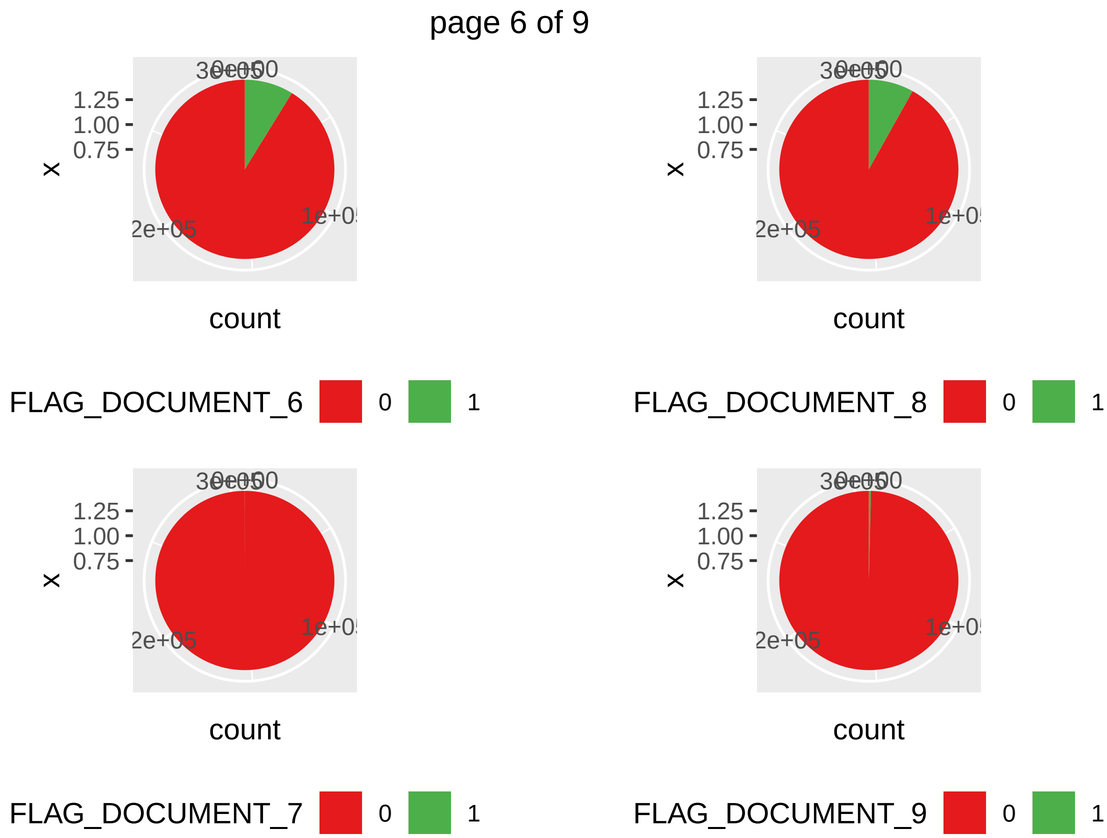

1.25
1.00
0.75
x
2e+05
count
FLAG_PHONE
0
1
REG_REGION_NOT_LIVE_REGION
0
FLAG_EMAIL
0
1
REG_REGION_NOT_WORK_REGION
0

page 4 of 9

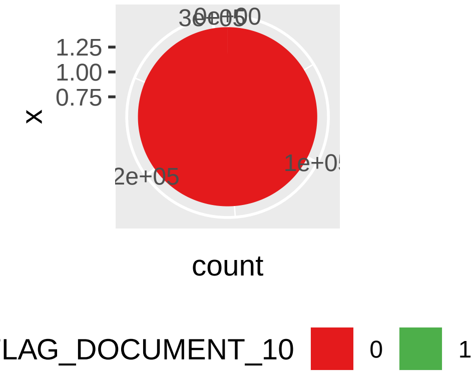

x
1.25
1.00
0.75
3e+05
0e+00
1e+05
2e+05
count
LIVE_REGION_NOT_WORK_REGION
0
REG_CITY_NOT_WORK_CITY
0
REG_CITY_NOT_LIVE_CITY
0
1
LIVE_CITY_NOT_WORK_CITY
0

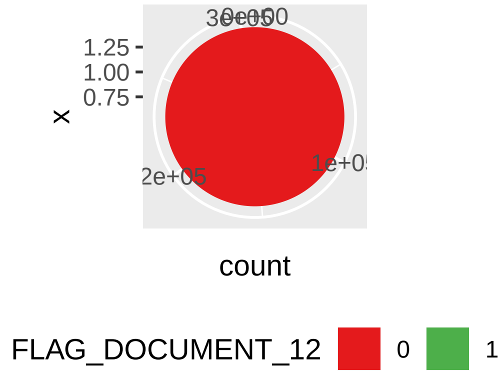
page 5 of 9

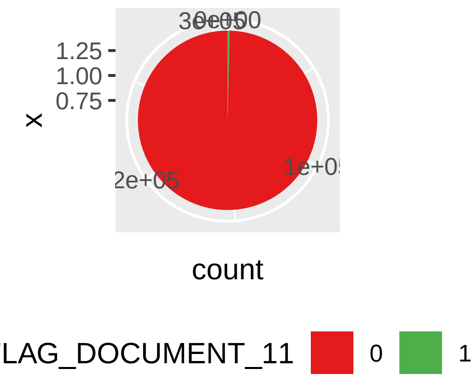
count
FLAG_DOCUMENT_2
0
1
count
FLAG_DOCUMENT_4
0
1
count
FLAG_DOCUMENT_3
0
1
count
FLAG_DOCUMENT_5
0
1

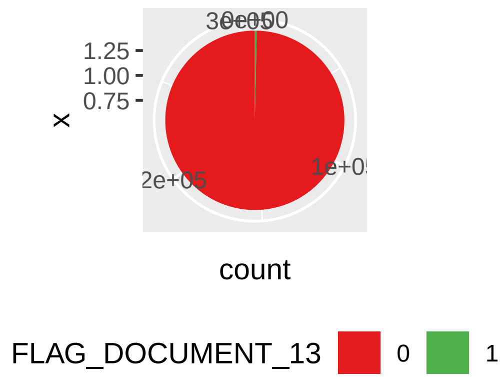

page 6 of 9
1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_6
0
1
FLAG_DOCUMENT_8
FLAG_DOCUMENT_7
FLAG_DOCUMENT_9

page 7 of 9

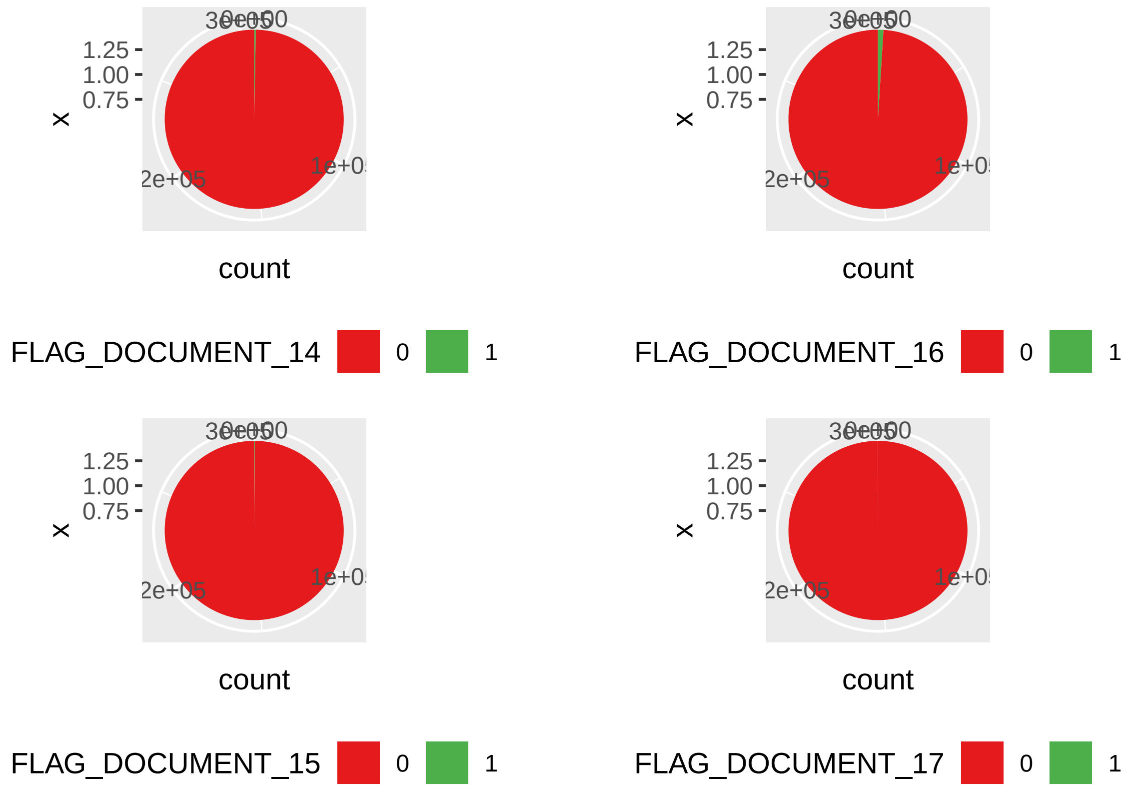

1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_10
0
1


1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_12
0
1

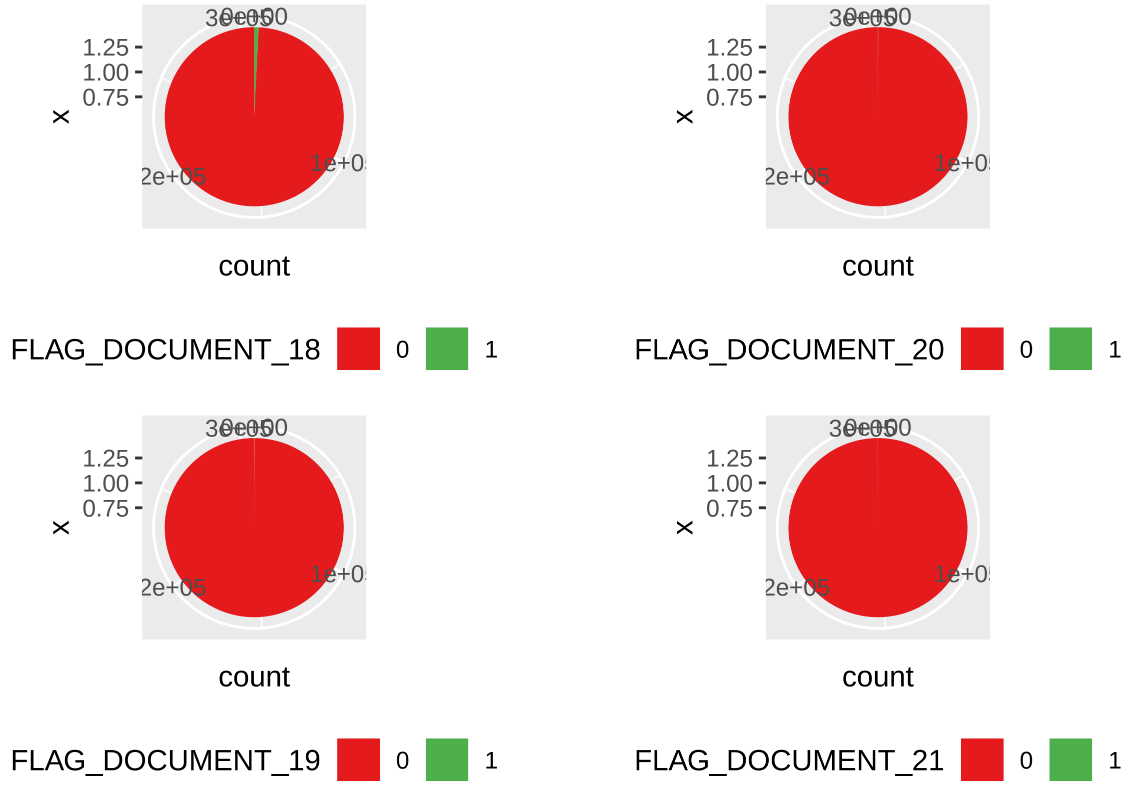

1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_11
0
1


1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_13
0
1

page 8 of 9

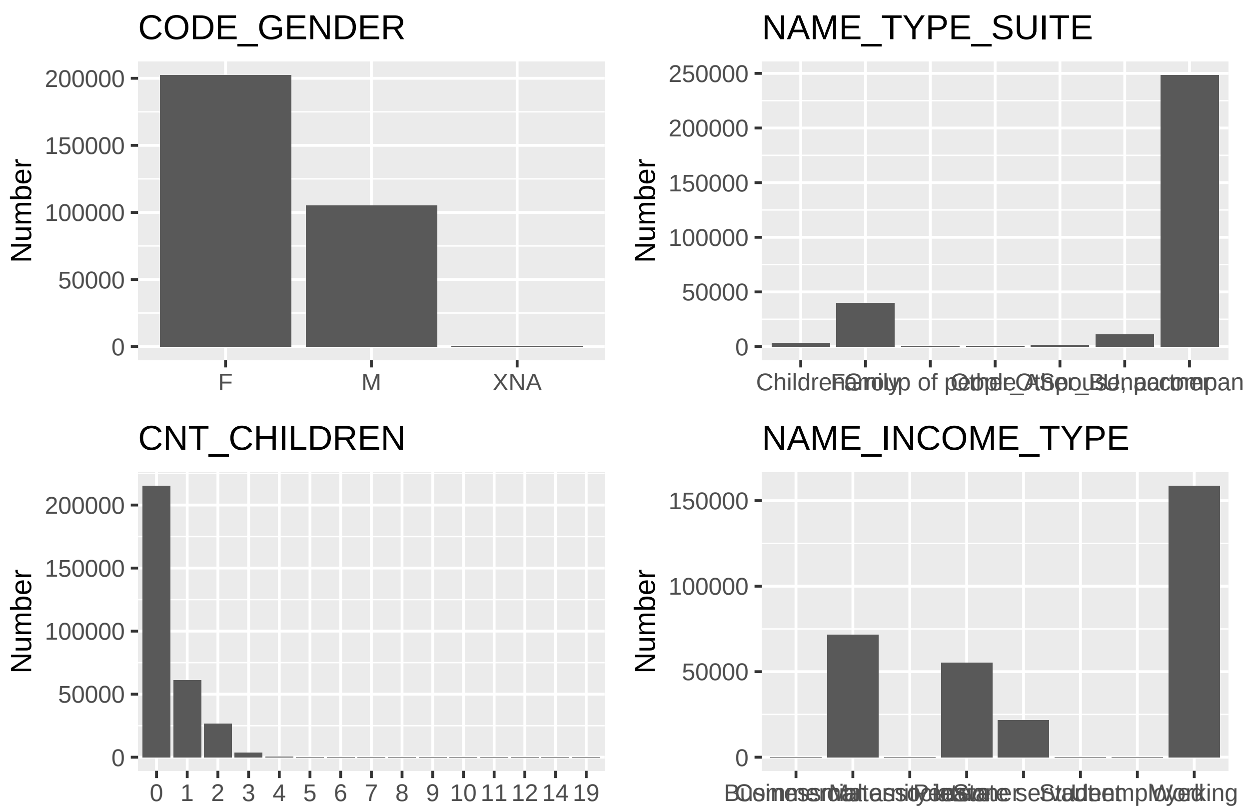

x
1.25
1.00
0.75
2e+05
count
FLAG_DOCUMENT_14
0
1
FLAG_DOCUMENT_16
0
1
FLAG_DOCUMENT_15
0
1
FLAG_DOCUMENT_17
0
1

page 9 of 9

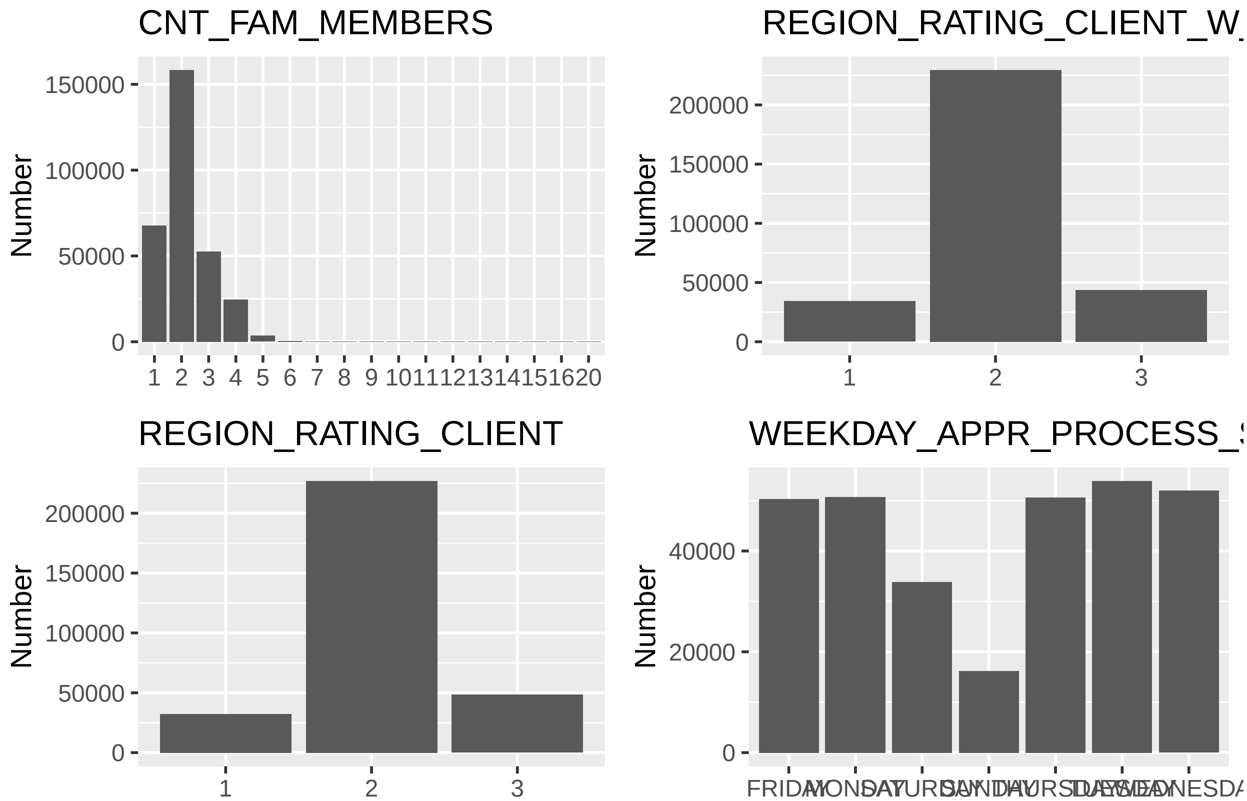

1.25
1.00
0.75
x
2e+05
count
FLAG_DOCUMENT_18
0
1
FLAG_DOCUMENT_20
FLAG_DOCUMENT_19
FLAG_DOCUMENT_21

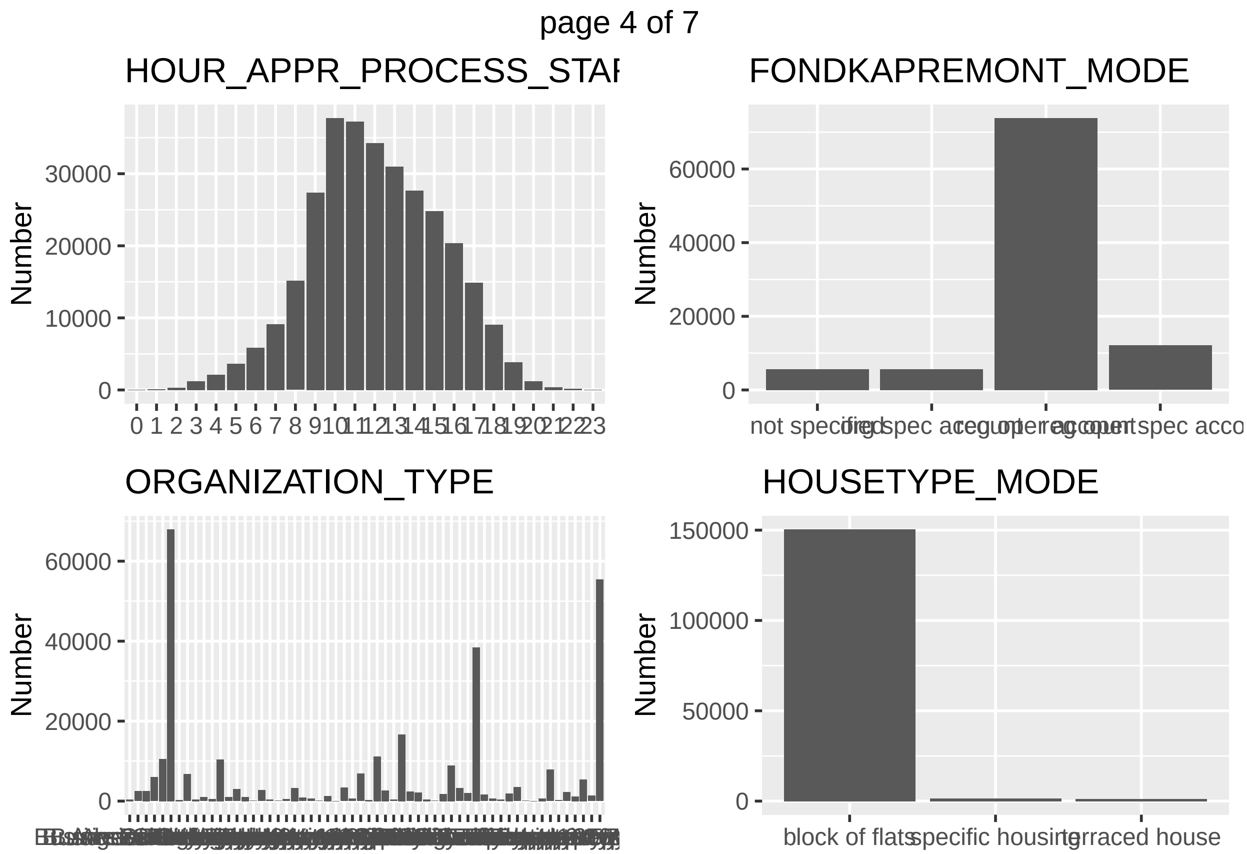
Appendix 2: Bar charts of categorical variables (with more than 2 categories)

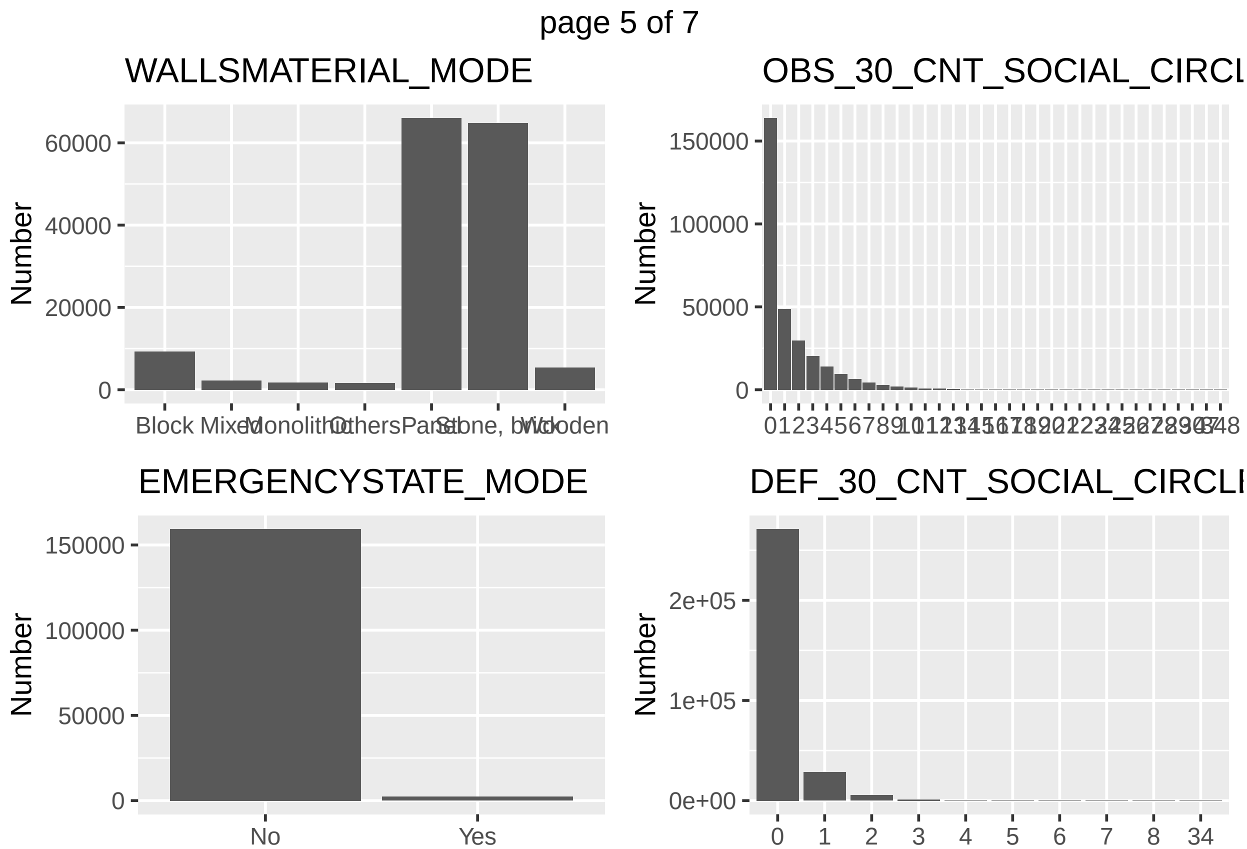
page 1 of 7

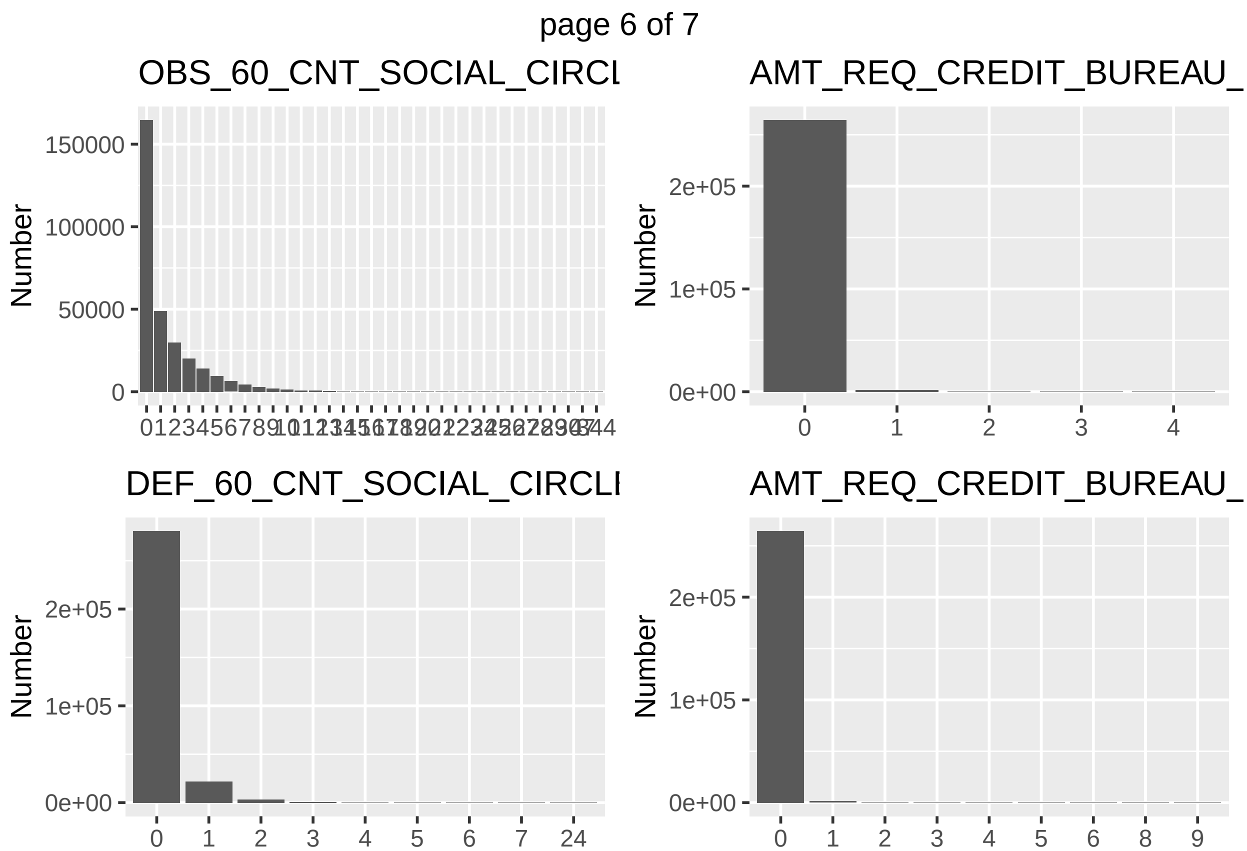


page 2 of 7

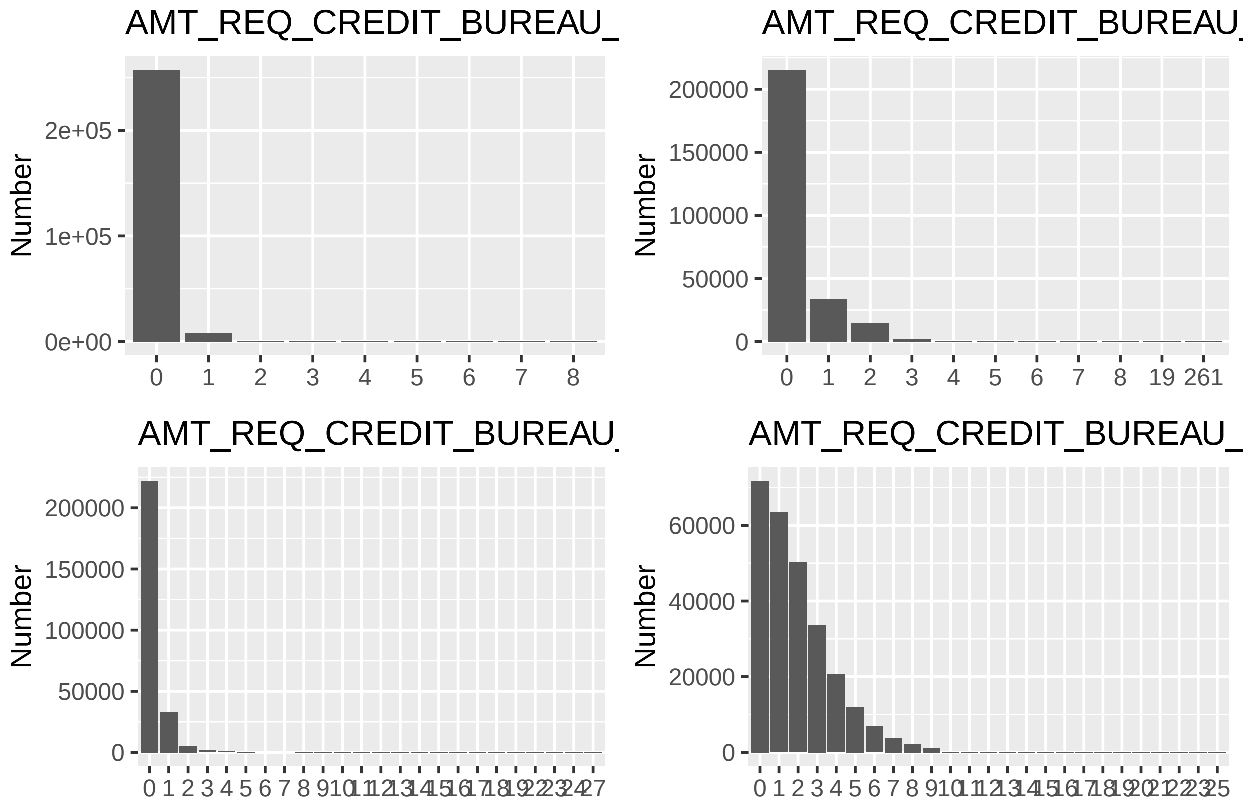

CNT_FAM_MEMBERS
Number
150000
100000
50000
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 20
REGION_RATING_CLIENT_W_
Number
200000
150000
100000
50000
0
1 2 3
REGION_RATING_CLIENT
Number
200000
150000
100000
50000
0
1 2 3
WEEKDAY_APPR_PROCESS_S
Number
40000
20000
0


page 4 of 7
HOUR_APPR_PROCESS_STAI
Number
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
FONDKAPREMONT_MODE
Number
ORGANIZATION_TYPE
Number
HOUSETYPE_MODE
Number
block of flats
terraced house

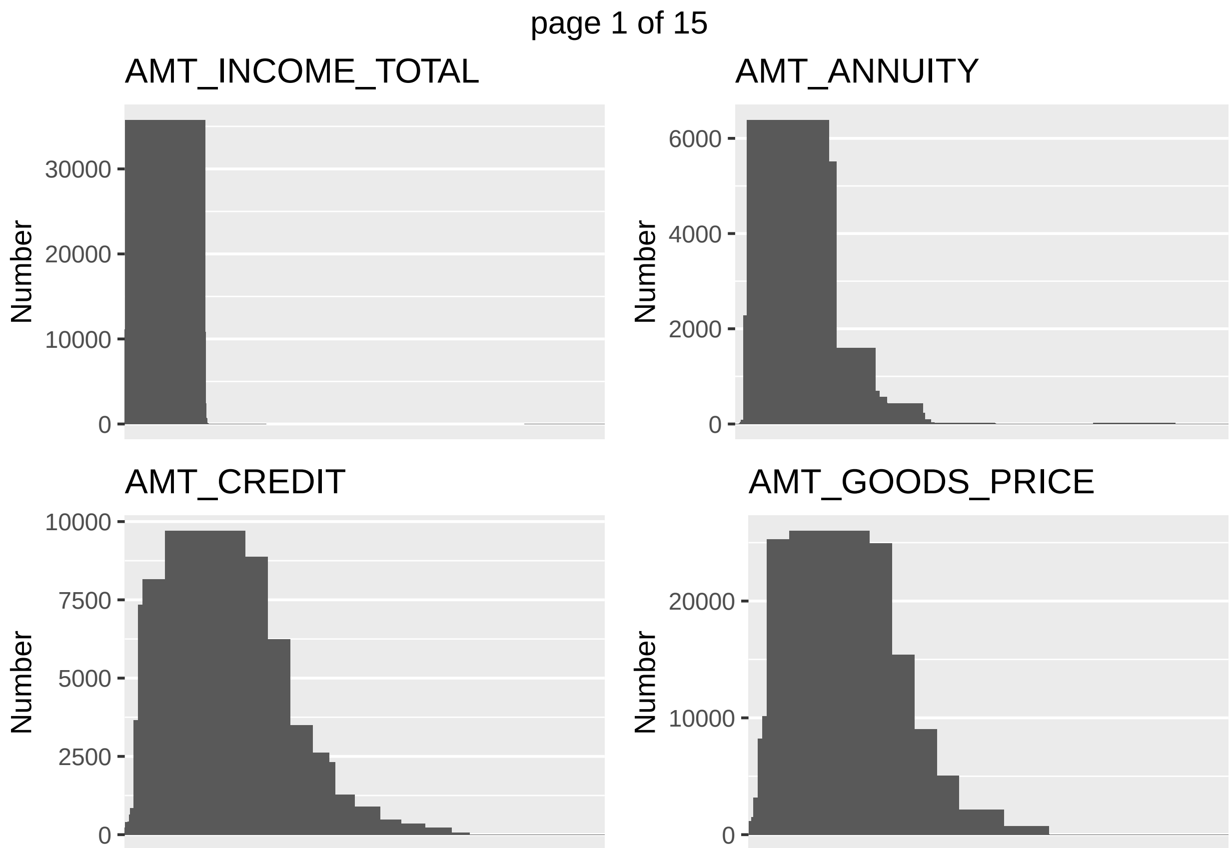

page 5 of 7
WALLSMATERIAL_MODE
Number
60000
40000
20000
0
Block
OBS_30_CNT_SOCIAL_CIRCL
150000
100000
50000
EMERGENCYSTATE_MODE
No
Yes
DEF_30_CNT_SOCIAL_CIRCLE
2e+05
1e+05
0e+00
0 1 2 3 4 5 6 7 8 34

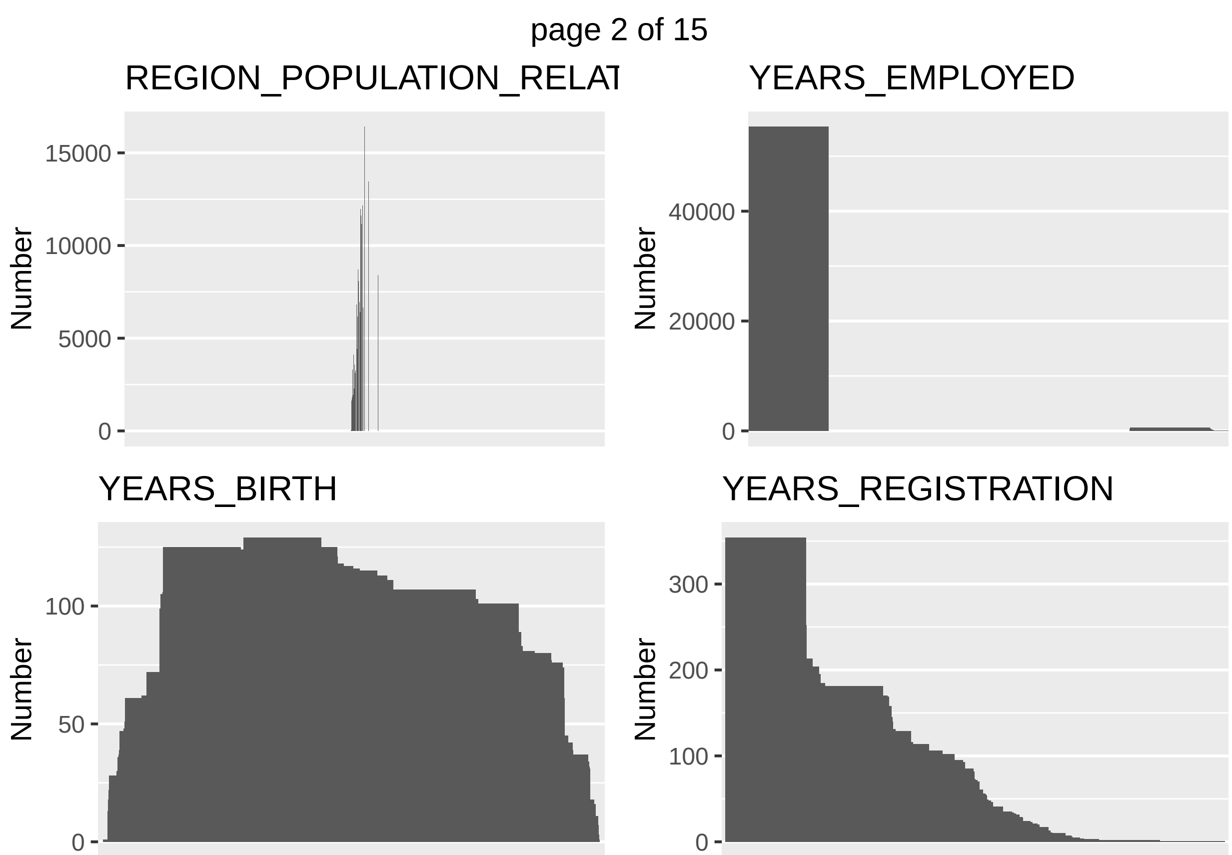

page 6 of 7
OBS_60_CNT_SOCIAL_CIRCL
Number
150000
100000
50000
0
AMT_REQ_CREDIT_BUREAU_
Number
2e+05
1e+05
0e+00
0
1
2
3
4
DEF_60_CNT_SOCIAL_CIRCLI
Number
2e+05
1e+05
0e+00
0
1
2
3
4
5
6
7
24
AMT_REQ_CREDIT_BUREAU_
Number
2e+05
1e+05
0e+00
0
1
2
3
4
5
6
8
9

page 7 of 7

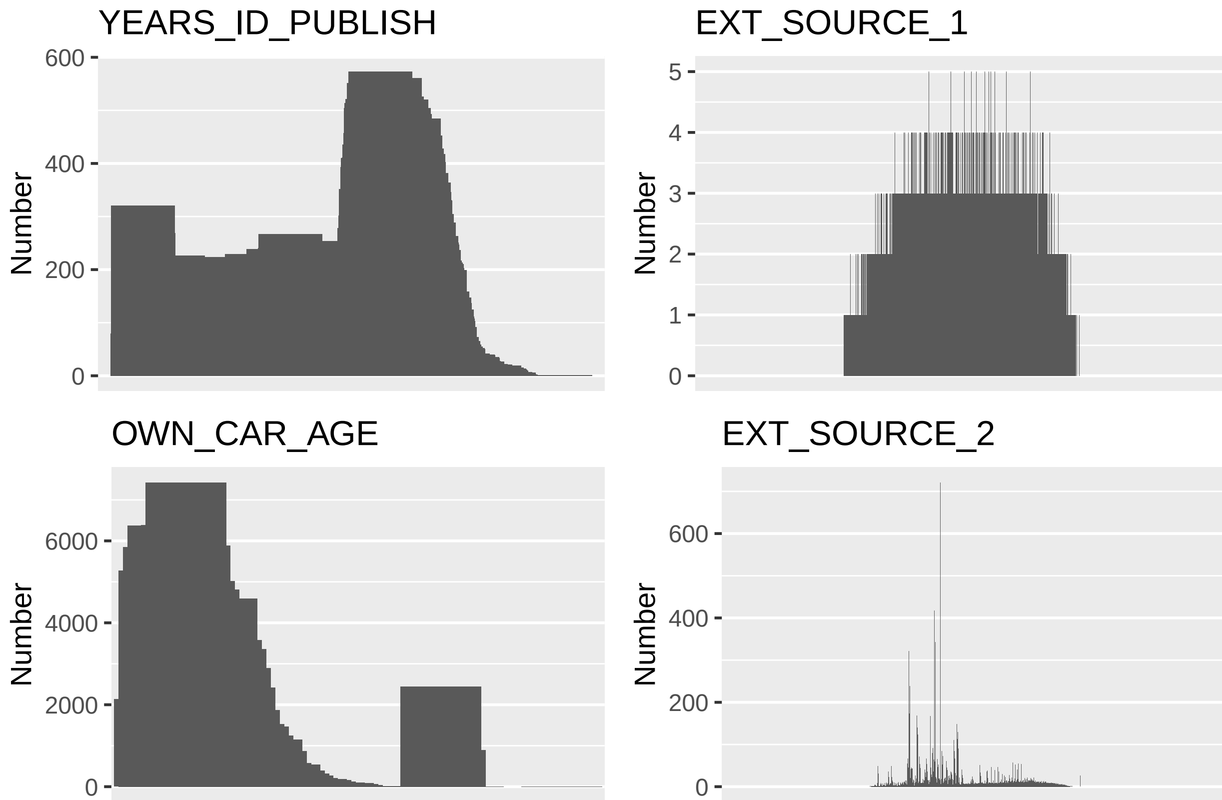

AMT_REQ_CREDIT_BUREAU_
Number
2e+05
1e+05
0e+00
0 1 2 3 4 5 6 7 8
AMT_REQ_CREDIT_BUREAU_
Number
200000
150000
100000
50000
0
0 1 2 3 4 5 6 7 8 19 261
AMT_REQ_CREDIT_BUREAU_
Number
200000
150000
100000
50000
0
AMT_REQ_CREDIT_BUREAU_
Number
60000
40000
20000
0

Appendix 3: Histograms of continuous variables

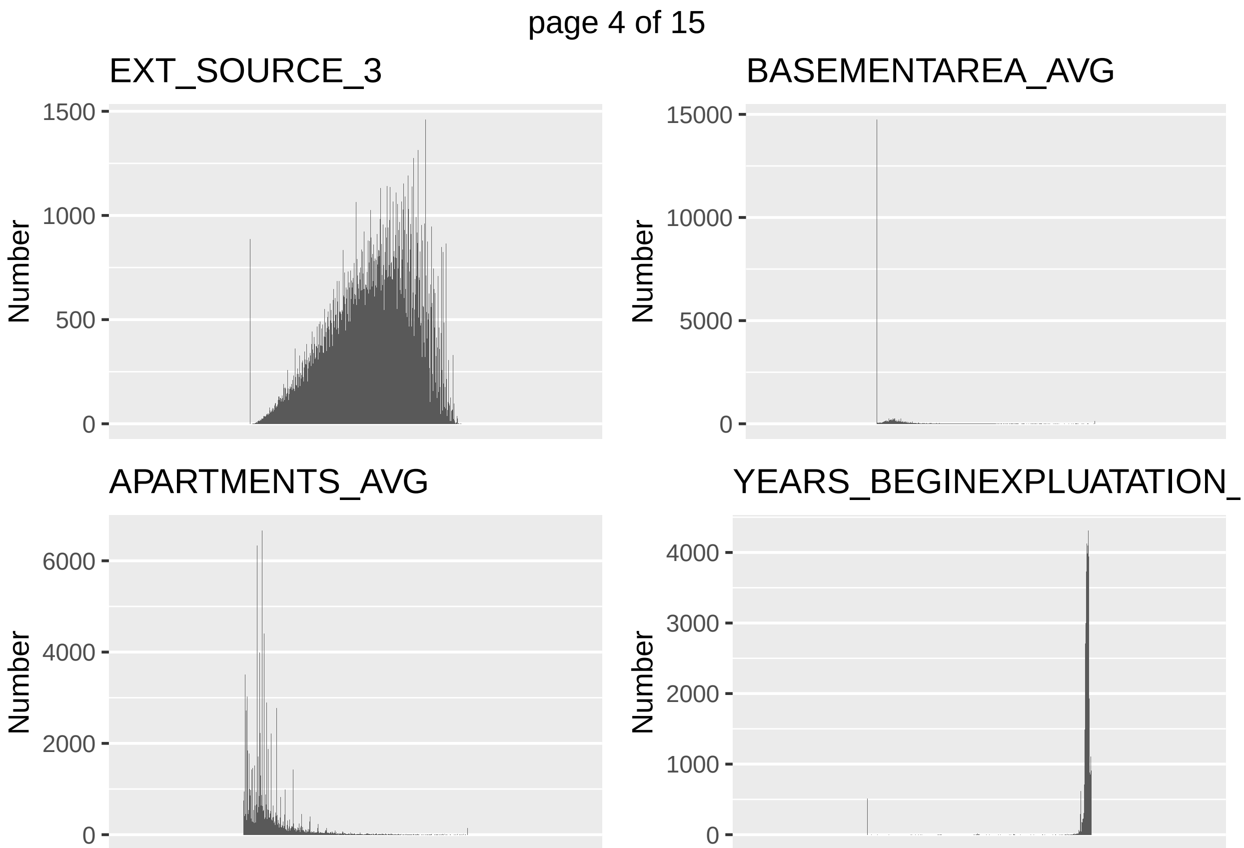




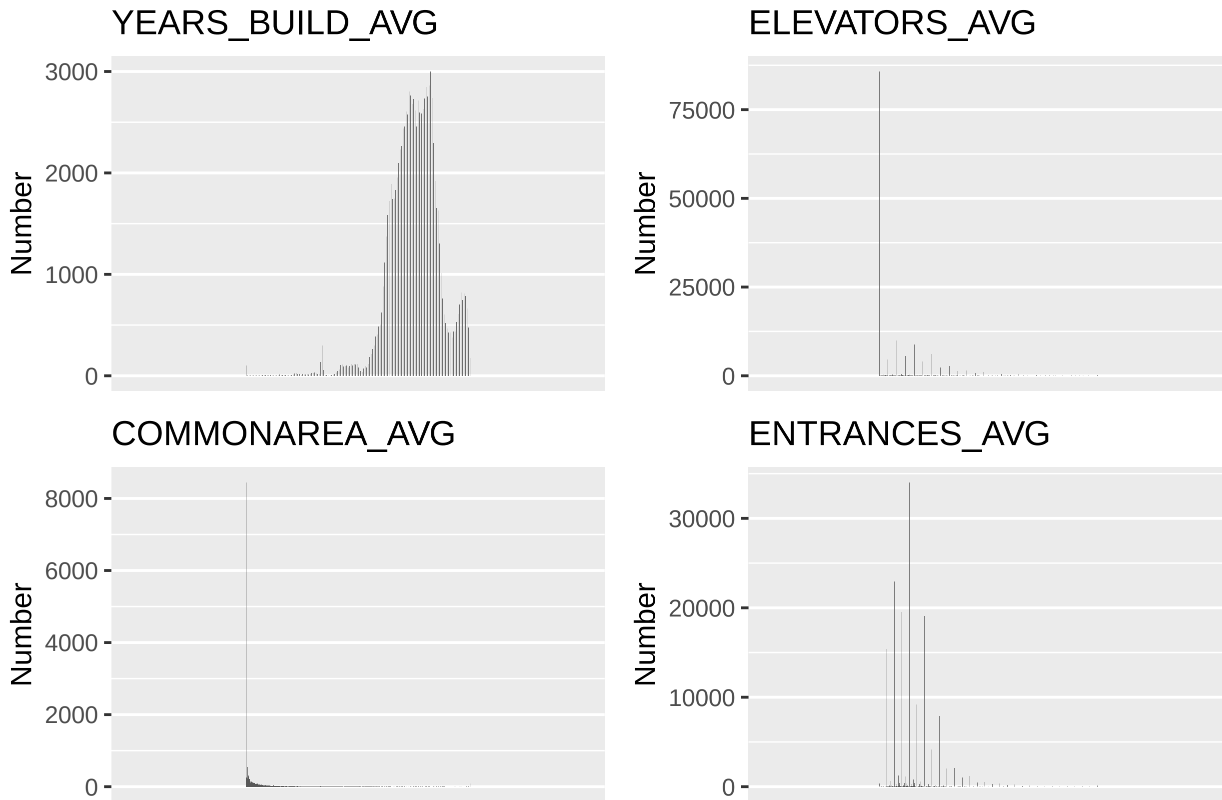
page 3 of 15

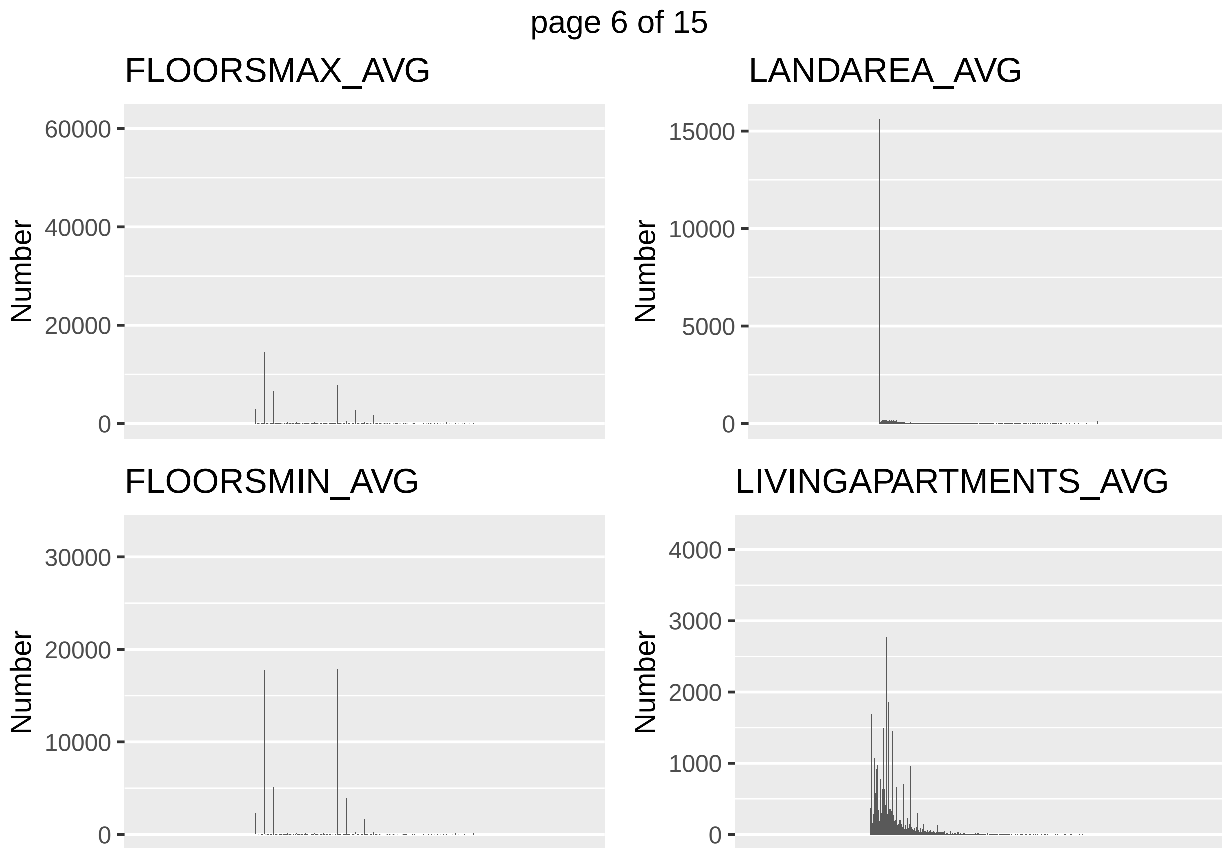


YEARS_ID_PUBLISH
EXT_SOURCE_1
OWN_CAR_AGE
EXT_SOURCE_2
Number
0
200
400
600
1
2
3
4
5
2000
4000
6000

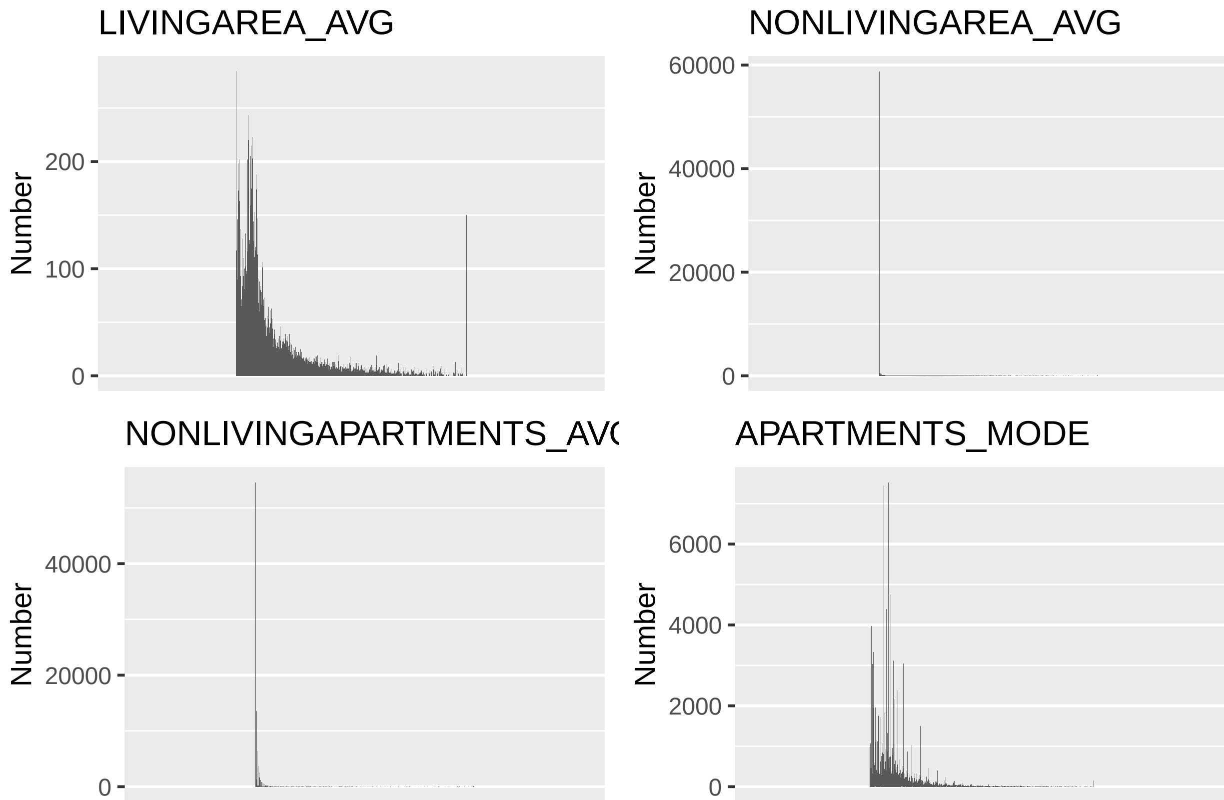

page 4 of 15
EXT_SOURCE_3
BASEMENTAREA_AVG
APARTMENTS_AVG
YEARS_BEGINEXPLUATATION_
Number
0
500
1000
1500
5000
10000
15000
2000
4000
6000
3000

page 5 of 15


YEARS_BUILD_AVG
Number
3000
2000
1000
0
ELEVATORS_AVG
Number
75000
50000
25000
0
COMMONAREA_AVG
Number
8000
6000
4000
2000
0
ENTRANCES_AVG
Number
30000
20000
10000
0

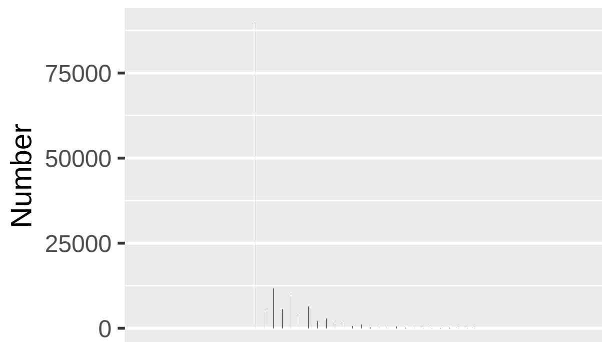

page 6 of 15
FLOORSMAX_AVG
Number
60000
40000
20000
0
LANDAREA_AVG
Number
15000
10000
5000
0
FLOORSMIN_AVG
Number
30000
20000
10000
0
LIVINGAPARTMENTS_AVG
Number
4000
3000
2000
1000
0

page 7 of 15

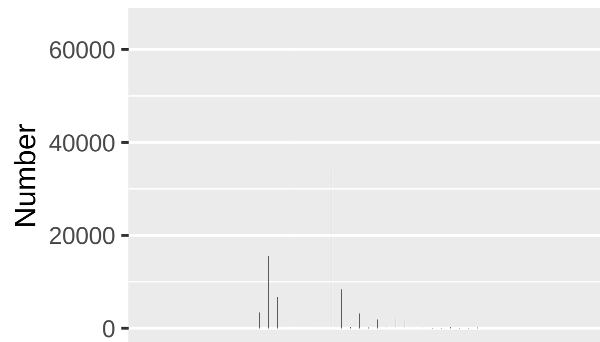

LIVINGAREA_AVG
Number
0
100
200
NONLIVINGAREA_AVG
Number
0
20000
40000
60000
NONLIVINGAPARTMENTS_AV(
Number
0
20000
40000
APARTMENTS_MODE
Number
0
2000
4000
6000


page 8 of 15
BASEMENTAREA_MODE
Number
15000
10000
5000
0
YEARS_BUILD_MODE
Number
3000
2000
1000
0
YEARS_BEGINEXPLUATATION_
Number
4000
3000
2000
1000
0
COMMONAREA_MODE
Number
10000
7500
5000
2500
0

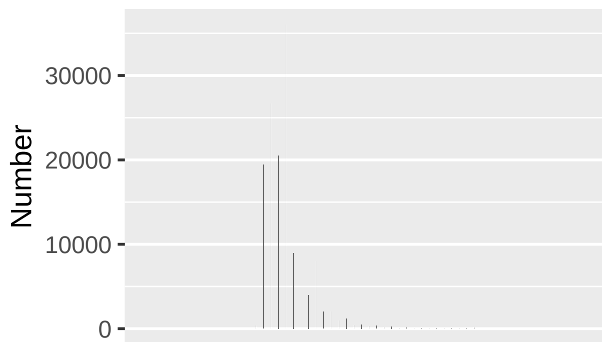
ELEVATORS_MODE



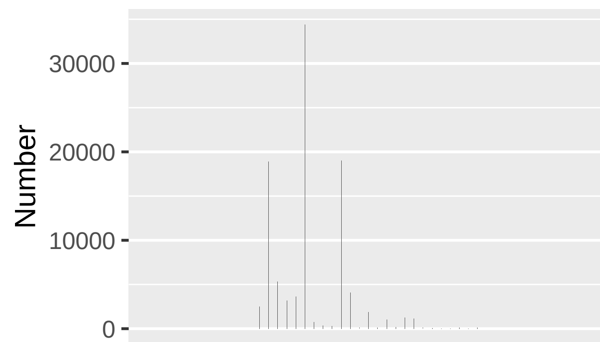
FLOORSMAX_MODE



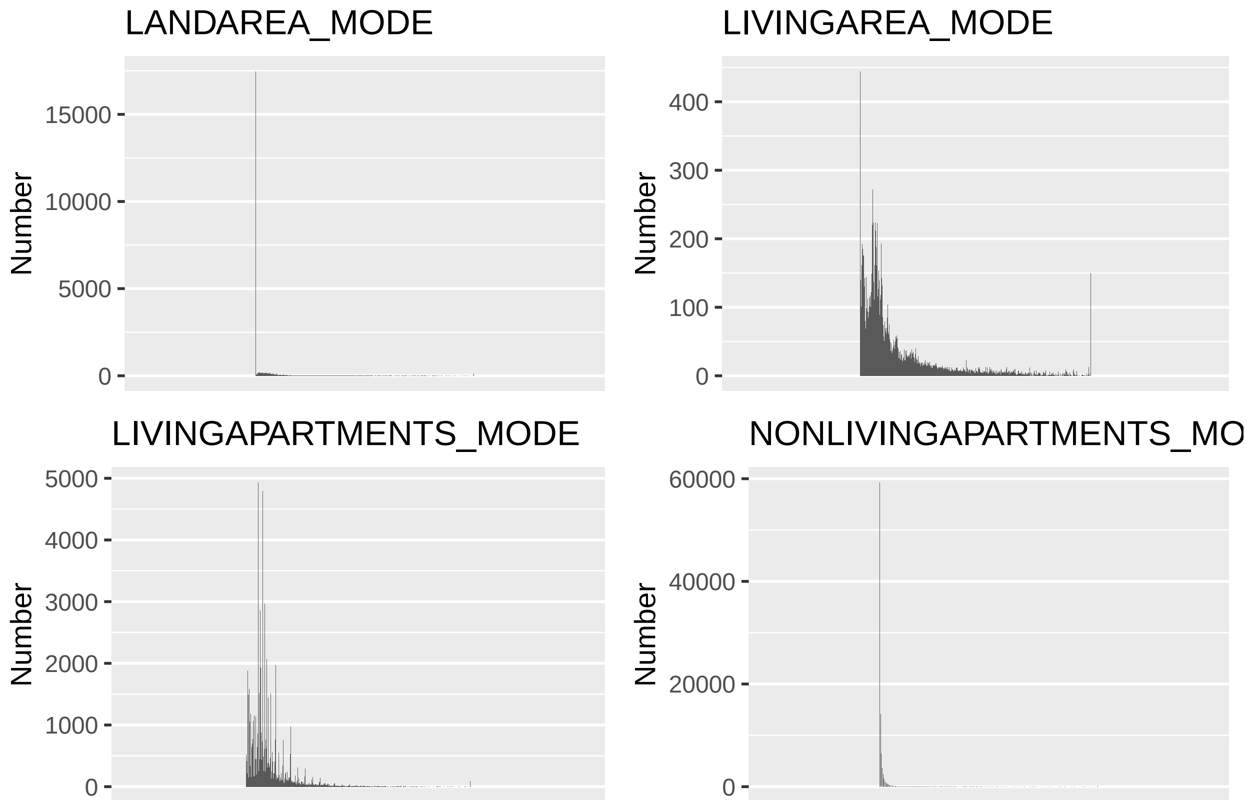
ENTRANCES_MODE



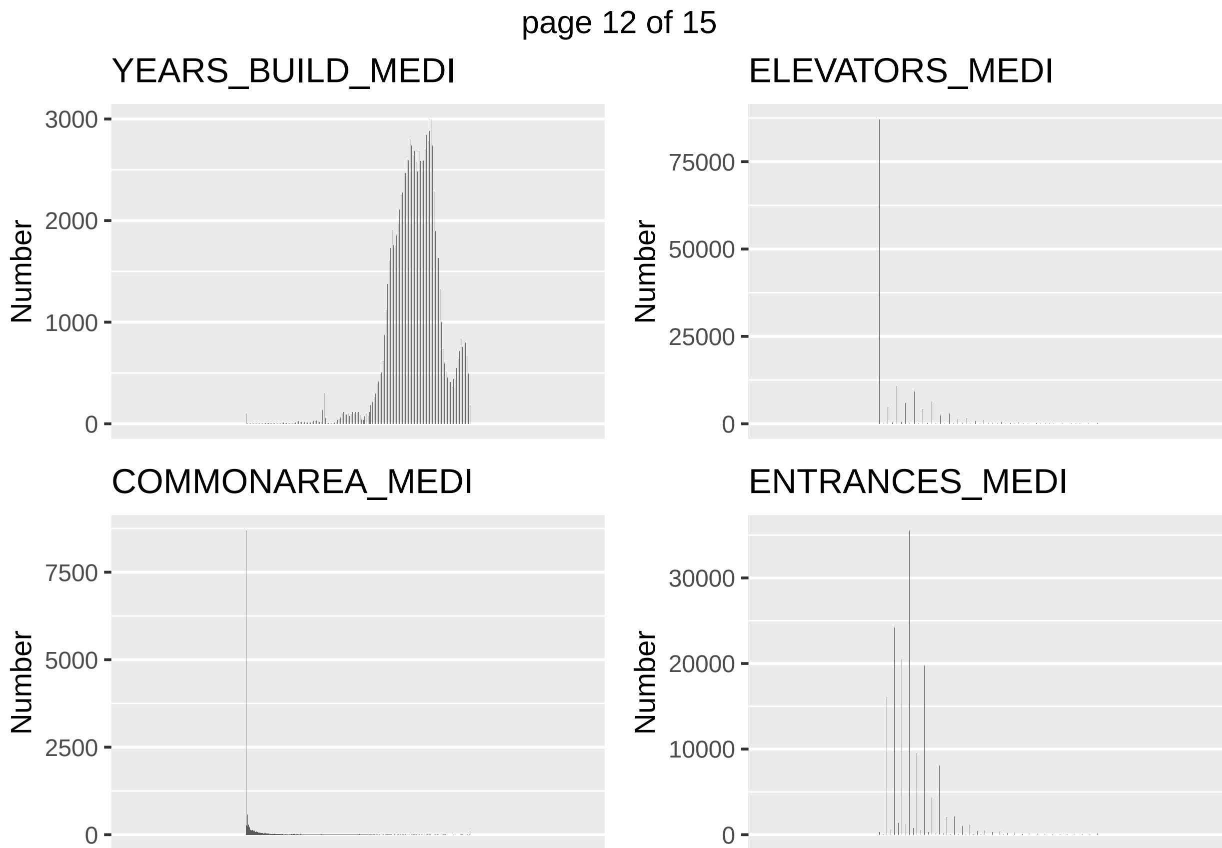
FLOORSMIN_MODE



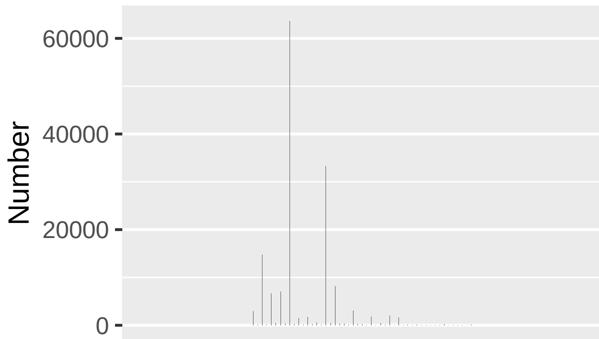
page 10 of 15


LANDAREA_MODE
Number
15000
10000
5000
0
LIVINGAREA_MODE
Number
400
300
200
100
0
LIVINGAPARTMENTS_MODE
Number
5000
4000
3000
2000
1000
0
NONLIVINGAPARTMENTS_MO
Number
60000
40000
20000
0

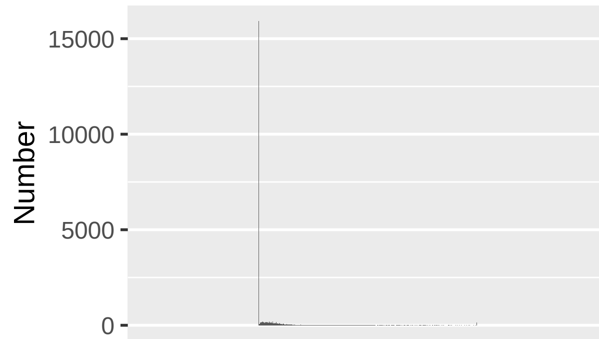

NONLIVINGAREA_MODE
Number
60000
40000
20000
0
BASEMENTAREA_MEDI
Number
15000
10000
5000
0
APARTMENTS_MEDI
Number
6000
4000
2000
0
YEARS_BEGINEXPLUATATION_
Number
4000
3000
2000
1000
0


page 12 of 15
YEARS_BUILD_MEDI
Number
3000
2000
1000
0
ELEVATORS_MEDI
Number
75000
50000
25000
0
COMMONAREA_MEDI
Number
7500
5000
2500
0
ENTRANCES_MEDI
Number
30000
20000
10000
0

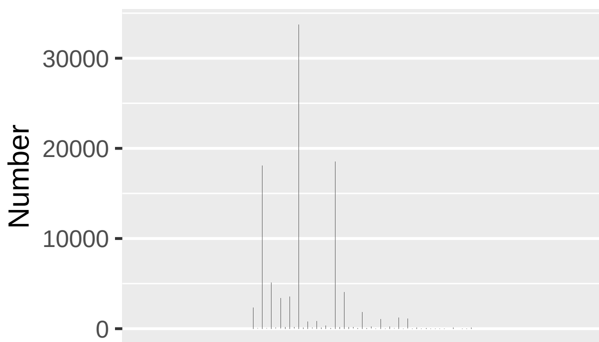
FLOORSMAX_MEDI



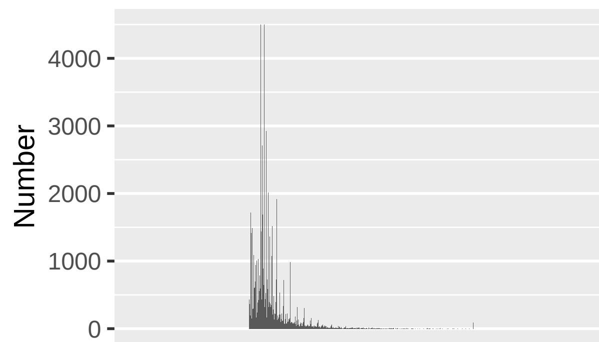
LANDAREA_MEDI

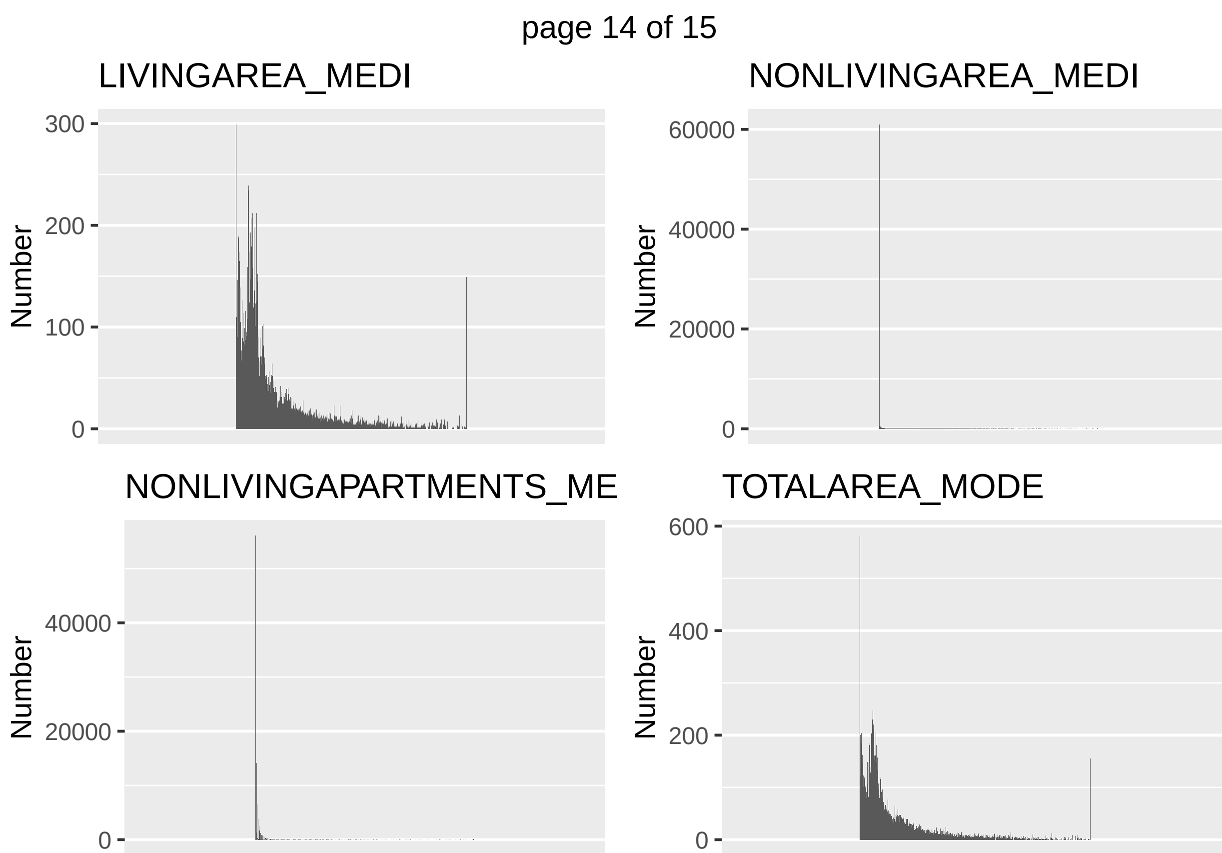


FLOORSMIN_MEDI



LIVINGAPARTMENTS_MEDI




page 14 of 15
LIVINGAREA_MEDI
Number
300
200
100
0
NONLIVINGAREA_MEDI
Number
60000
40000
20000
0
NONLIVINGAPARTMENTS_ME
Number
40000
20000
0
TOTALAREA_MODE
Number
600
400
200
0

page 15 of 15