

- Есть наны
 - ☐ выбросить строки с нанами
 - ☐ заполнить наны: street - ближайшими (соседями)

```
Requirement already satisfied: pygments in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (2.6.1)
Requirement already satisfied: simplegeneric>0.8 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.1.2)
Requirement already satisfied: pexpect in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.8.0)
Requirement already satisfied: setuptools>=18.5 in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (59.5.0)
Requirement already satisfied: decorator in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.7/dist-packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.5)
Requirement already satisfied: MarkupSafe>=0.23 in /usr/local/lib/python3.7/dist-packages (from Jinja2<3.0,>=2.11->dataprep) (2.0.1)
Collecting ply
  Downloading ply-3.11-py2.py3-none-any.whl (49 kB)
    |██████████| 49 kB 5.5 MB/s
Requirement already satisfied: six in /usr/local/lib/python3.7/dist-packages (from jsonpath-ng<2.0,>=1.5->dataprep) (1.15.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (3.2.0)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.7/dist-packages (from nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.7.0)
Requirement already satisfied: click in /usr/local/lib/python3.7/dist-packages (from nltk<4.0,>=3.5->dataprep) (7.1.2)
Requirement already satisfied: joblib in /usr/local/lib/python3.7/dist-packages (from nltk<4.0,>=3.5->dataprep) (1.0.1)
Requirement already satisfied: pyparsing>=2.0.2 in /usr/local/lib/python3.7/dist-packages (from packaging>=16.8->bokeh<3,>=2->dataprep) (2.4.7)
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/dist-packages (from pandas<2.0,>=1.1->dataprep) (2018.9)
Collecting locket
  Downloading locket-0.2.1-py2.py3-none-any.whl (4.1 kB)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.7/dist-packages (from prompt-toolkit<2.0.0,>=1.0.4->ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.1)
Collecting python-crfsuite>=0.7
  Downloading python_crfsuite-0.9.7-cp37-cp37m-manylinux1_x86_64.whl (743 kB)
    |██████████| 743 kB 58.6 MB/s
Requirement already satisfied: future>=0.14 in /usr/local/lib/python3.7/dist-packages (from usaddress<0.6.0,>=0.5.10->dataprep) (0.16.0)
Collecting probablenlp
  Downloading probablenlp-0.0.1-py2.py3-none-any.whl (3.1 kB)
Requirement already satisfied: notebook>=4.4.1 in /usr/local/lib/python3.7/dist-packages (from widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (4.4.1)
Requirement already satisfied: terminado>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.12.1)
Requirement already satisfied: Send2Trash in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.8.0)
Requirement already satisfied: nbconvert in /usr/local/lib/python3.7/dist-packages (from notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (5.6.1)
Requirement already satisfied: pyzmq>=13 in /usr/local/lib/python3.7/dist-packages (from jupyter-client->ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (19.0.0)
Requirement already satisfied: ptyprocess in /usr/local/lib/python3.7/dist-packages (from terminado>=0.8.1->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.0)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from wordcloud<2.0,>=1.8->dataprep) (3.2.2)
Requirement already satisfied: idna>=2.0 in /usr/local/lib/python3.7/dist-packages (from yarl<2.0,>=1.0->aiohttp<4.0,>=3.6->dataprep) (2.10)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (1.3.1)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->wordcloud<2.0,>=1.8->dataprep) (0.10.0)
Requirement already satisfied: entrypoints>=0.2.2 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.3)
Requirement already satisfied: testpath in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.6.0)
Requirement already satisfied: pandocfilters>=1.4.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.4.1)
Requirement already satisfied: mistune<2,>=0.8.1 in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.1)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.6.0)
Requirement already satisfied: bleach in /usr/local/lib/python3.7/dist-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (3.1.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.7/dist-packages (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~>3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.1)
Building wheels for collected packages: metaphone, python-Levenshtein
  Building wheel for metaphone (setup.py) ... done
  Created wheel for metaphone: filename=Metaphone-0.6-py3-none-any.whl size=13919 sha256=bd90eef3f0379dc261146a59af3e3a14791a41f1d2760dadced40cd7
  Stored in directory: /root/.cache/pip/wheels/1d/a8/cb/6f8902aa5457bd71344e00665c230e9c45255b3f57f2194a0f
  Building wheel for python-Levenshtein (setup.py) ... done
  Created wheel for python-Levenshtein: filename=python-Levenshtein-0.12.2-cp37-cp37m-linux_x86_64.whl size=149868 sha256=152abf514547a022b058d21
  Stored in directory: /root/.cache/pip/wheels/05/5f/ca/7c4367734892581bb5ff896f15027a932c551080b2abd3ce00d
Successfully built metaphone python-Levenshtein
Installing collected packages: multidict, locket, yarl, regex, python-crfsuite, probablenlp, ply, partd, fsspec, dask, asyncio-timeout, wordcloud
Attempting uninstall: regex
  Found existing installation: regex 2019.12.20
  Uninstalling regex-2019.12.20:
    Successfully uninstalled regex-2019.12.20
Attempting uninstall: dask
  Found existing installation: dask 2.12.0
  Uninstalling dask-2.12.0:
```

```
1 df = pd.read_csv(os.path.join(path, 'train.csv'))
```

◀ ▶

(279792, 77)

	city	floor	id	lat	lng	osm_amenity_points_in_0.001	osm_amenity_points_in_1
0	Пермь	NaN	COL_0	57.998207	56.292797		4
1	Шатура	NaN	COL_1	55.574284	39.543835		3
2	Ярославль	NaN	COL_2	57.619140	39.850525		1
3	Новокузнецк	NaN	COL_3	53.897083	87.108604		0
4	Москва	NaN	COL_4	55.802590	37.487110		1

```
Index(['city', 'floor', 'id', 'lat', 'lng', 'osm_amenity_points_in_0.001',
      'osm_amenity_points_in_0.005', 'osm_amenity_points_in_0.0075',
      'osm_amenity_points_in_0.01', 'osm_building_points_in_0.001',
      'osm_building_points_in_0.005', 'osm_building_points_in_0.0075',
      'osm_building_points_in_0.01', 'osm_catering_points_in_0.001',
      'osm_catering_points_in_0.005', 'osm_catering_points_in_0.0075',
      'osm_catering_points_in_0.01', 'osm_city_closest_dist',
      'osm_city_nearest_name', 'osm_city_nearest_population',
      'osm_crossing_closest_dist', 'osm_crossing_points_in_0.001',
      'osm_crossing_points_in_0.005', 'osm_crossing_points_in_0.0075',
      'osm_crossing_points_in_0.01', 'osm_culture_points_in_0.001',
      'osm_culture_points_in_0.005', 'osm_culture_points_in_0.0075',
      'osm_culture_points_in_0.01', 'osm_finance_points_in_0.001',
      'osm_finance_points_in_0.005', 'osm_finance_points_in_0.0075',
      'osm_finance_points_in_0.01', 'osm_healthcare_points_in_0.005',
      'osm_healthcare_points_in_0.0075', 'osm_healthcare_points_in_0.01',
      'osm_historic_points_in_0.005', 'osm_historic_points_in_0.0075',
      'osm_historic_points_in_0.01', 'osm_hotels_points_in_0.005',
      'osm_hotels_points_in_0.0075', 'osm_hotels_points_in_0.01',
      'osm_leisure_points_in_0.005', 'osm_leisure_points_in_0.0075',
      'osm_leisure_points_in_0.01', 'osm_offices_points_in_0.001',
      'osm_offices_points_in_0.005', 'osm_offices_points_in_0.0075',
      'osm_offices_points_in_0.01', 'osm_shops_points_in_0.001',
      'osm_shops_points_in_0.005', 'osm_shops_points_in_0.0075',
      'osm_shops_points_in_0.01', 'osm_subway_closest_dist',
      'osm_train_stop_closest_dist', 'osm_train_stop_points_in_0.005',
      'osm_train_stop_points_in_0.0075', 'osm_train_stop_points_in_0.01',
      'osm_transport_stop_closest_dist', 'osm_transport_stop_points_in_0.005',
      'osm_transport_stop_points_in_0.0075',
      'osm_transport_stop_points_in_0.01', 'per_square_meter_price',
      'reform_count_of_houses_1000', 'reform_count_of_houses_500',
      'reform_house_population_1000', 'reform_house_population_500',
      'reform_mean_floor_count_1000', 'reform_mean_floor_count_500',
      'reform_mean_year_building_1000', 'reform_mean_year_building_500',
      'region', 'total_square', 'street', 'date', 'realty_type',
      'price_type'],
      dtype='object')
```

16	osm_catering_points_in_0.01	279792	non-null	int64
17	osm_city_closest_dist	279792	non-null	float64
18	osm_city_nearest_name	279792	non-null	object
19	osm_city_nearest_population	279737	non-null	float64
20	osm_crossing_closest_dist	279792	non-null	float64
21	osm_crossing_points_in_0.001	279792	non-null	int64
22	osm_crossing_points_in_0.005	279792	non-null	int64
23	osm_crossing_points_in_0.0075	279792	non-null	int64
24	osm_crossing_points_in_0.01	279792	non-null	int64
25	osm_culture_points_in_0.001	279792	non-null	int64
26	osm_culture_points_in_0.005	279792	non-null	int64
27	osm_culture_points_in_0.0075	279792	non-null	int64
28	osm_culture_points_in_0.01	279792	non-null	int64
29	osm_finance_points_in_0.001	279792	non-null	int64
30	osm_finance_points_in_0.005	279792	non-null	int64
31	osm_finance_points_in_0.0075	279792	non-null	int64
32	osm_finance_points_in_0.01	279792	non-null	int64
33	osm_healthcare_points_in_0.005	279792	non-null	int64
34	osm_healthcare_points_in_0.0075	279792	non-null	int64
35	osm_healthcare_points_in_0.01	279792	non-null	int64
36	osm_historic_points_in_0.005	279792	non-null	int64
37	osm_historic_points_in_0.0075	279792	non-null	int64
38	osm_historic_points_in_0.01	279792	non-null	int64
39	osm_hotels_points_in_0.005	279792	non-null	int64
40	osm_hotels_points_in_0.0075	279792	non-null	int64
41	osm_hotels_points_in_0.01	279792	non-null	int64
42	osm_leisure_points_in_0.005	279792	non-null	int64
43	osm_leisure_points_in_0.0075	279792	non-null	int64

44	osm_leisure_points_in_0.01	279792	non-null	int64
45	osm_offices_points_in_0.001	279792	non-null	int64
46	osm_offices_points_in_0.005	279792	non-null	int64
47	osm_offices_points_in_0.0075	279792	non-null	int64
48	osm_offices_points_in_0.01	279792	non-null	int64
49	osm_shops_points_in_0.001	279792	non-null	int64
50	osm_shops_points_in_0.005	279792	non-null	int64
51	osm_shops_points_in_0.0075	279792	non-null	int64
52	osm_shops_points_in_0.01	279792	non-null	int64
53	osm_subway_closest_dist	279792	non-null	float64
54	osm_train_stop_closest_dist	279792	non-null	float64
55	osm_train_stop_points_in_0.005	279792	non-null	int64
56	osm_train_stop_points_in_0.0075	279792	non-null	int64
57	osm_train_stop_points_in_0.01	279792	non-null	int64
58	osm_transport_stop_closest_dist	279792	non-null	float64
59	osm_transport_stop_points_in_0.005	279792	non-null	int64
60	osm_transport_stop_points_in_0.0075	279792	non-null	int64
61	osm_transport_stop_points_in_0.01	279792	non-null	int64
62	per_square_meter_price	279792	non-null	float64
63	reform_count_of_houses_1000	279792	non-null	int64
64	reform_count_of_houses_500	279792	non-null	int64
65	reform_house_population_1000	265196	non-null	float64
66	reform_house_population_500	252558	non-null	float64
67	reform_mean_floor_count_1000	263084	non-null	float64
68	reform_mean_floor_count_500	249624	non-null	float64
69	reform_mean_year_building_1000	263553	non-null	float64
70	reform_mean_year_building_500	250155	non-null	float64
71	region	279792	non-null	object
72	total_square	279792	non-null	float64
73	street	278186	non-null	object
74	date	279792	non-null	object

▼ Пропущенные значения

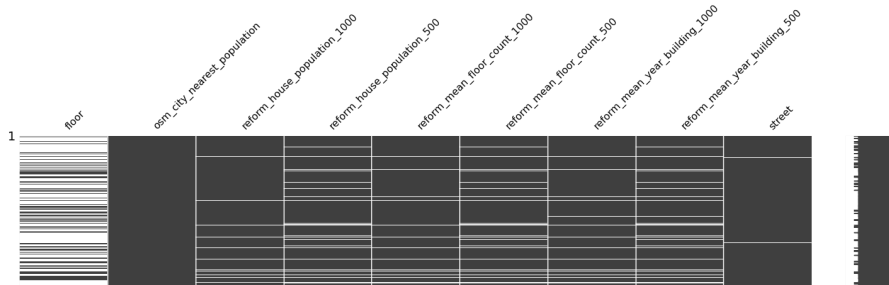
```

1 col_with_nan = []
2 ROW_NUM = df.shape[0]
3
4 for col in df.columns:
5     if df[col].isna().sum() != 0:
6         print(col, round(df[col].isna().sum() / ROW_NUM * 100, 2), '%')
7         col_with_nan.append(col)
8
9 col_with_nan

floor 62.99 %
osm_city_nearest_population 0.02 %
reform_house_population_1000 5.22 %
reform_house_population_500 9.73 %
reform_mean_floor_count_1000 5.97 %
reform_mean_floor_count_500 10.78 %
reform_mean_year_building_1000 5.8 %
reform_mean_year_building_500 10.59 %
street 0.57 %
['floor',
 'osm_city_nearest_population',
 'reform_house_population_1000',
 'reform_house_population_500',
 'reform_mean_floor_count_1000',
 'reform_mean_floor_count_500',
 'reform_mean_year_building_1000',
 'reform_mean_year_building_500',
 'street']

1 msno.matrix(df[col_with_nan]);

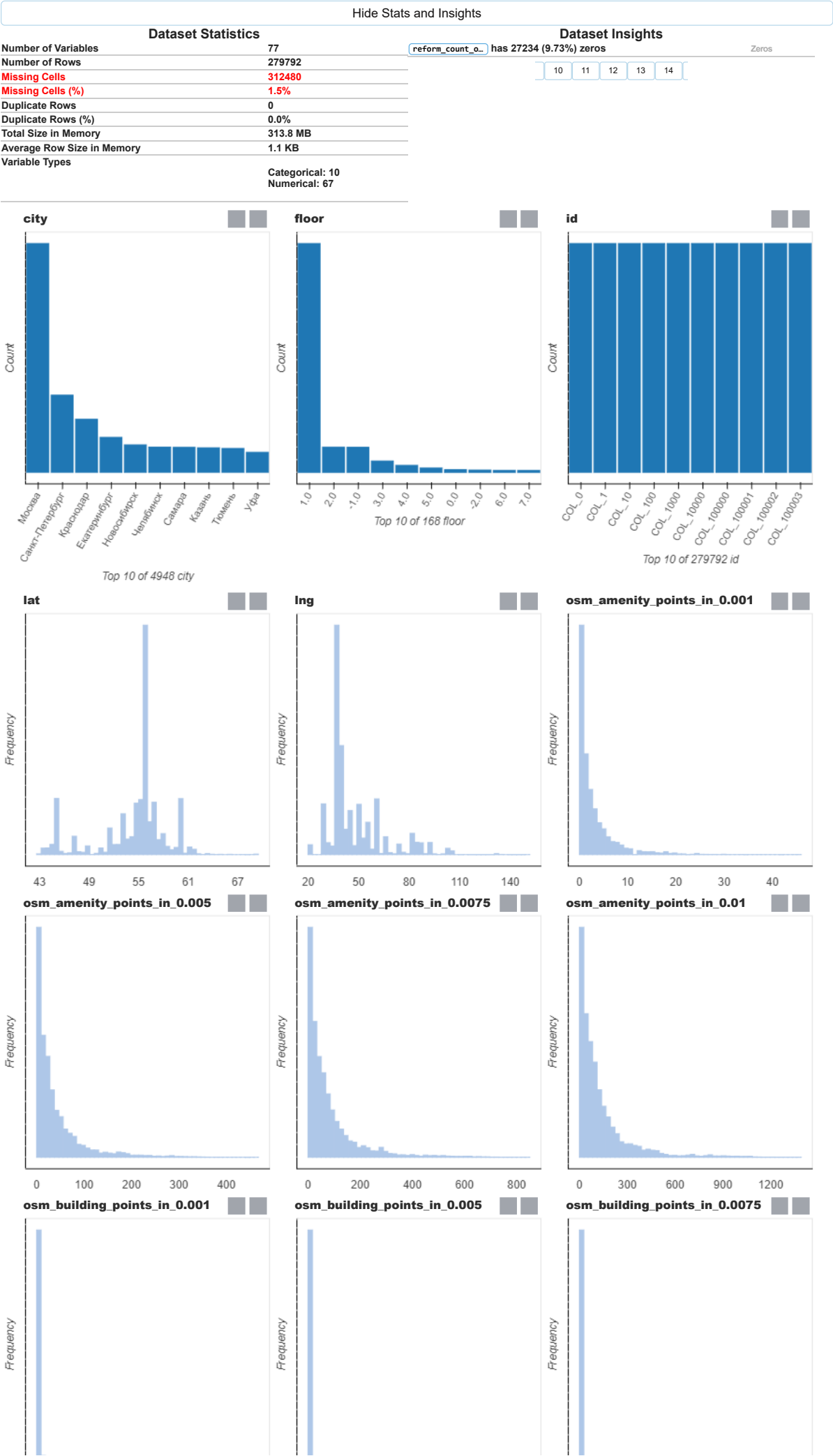
```

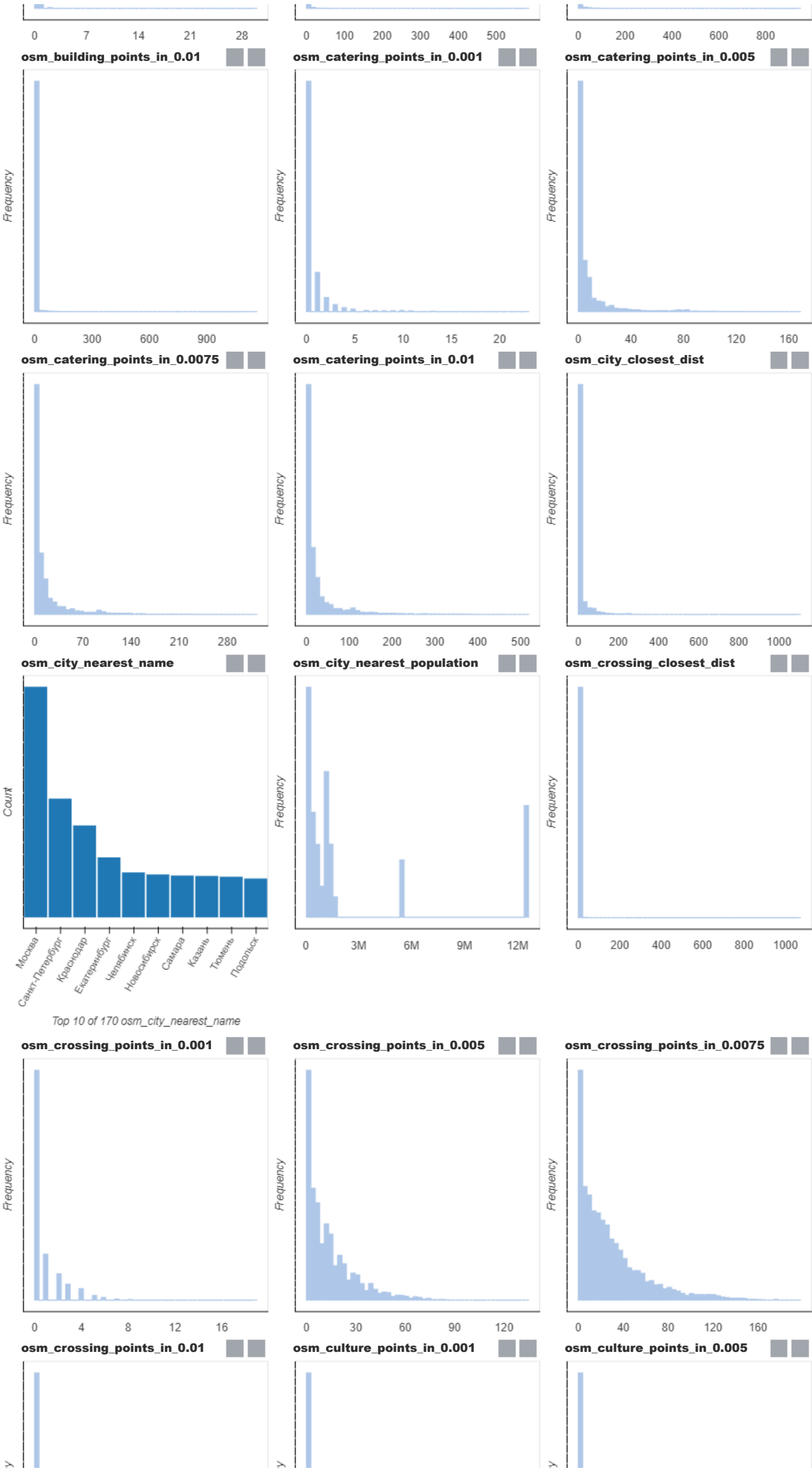


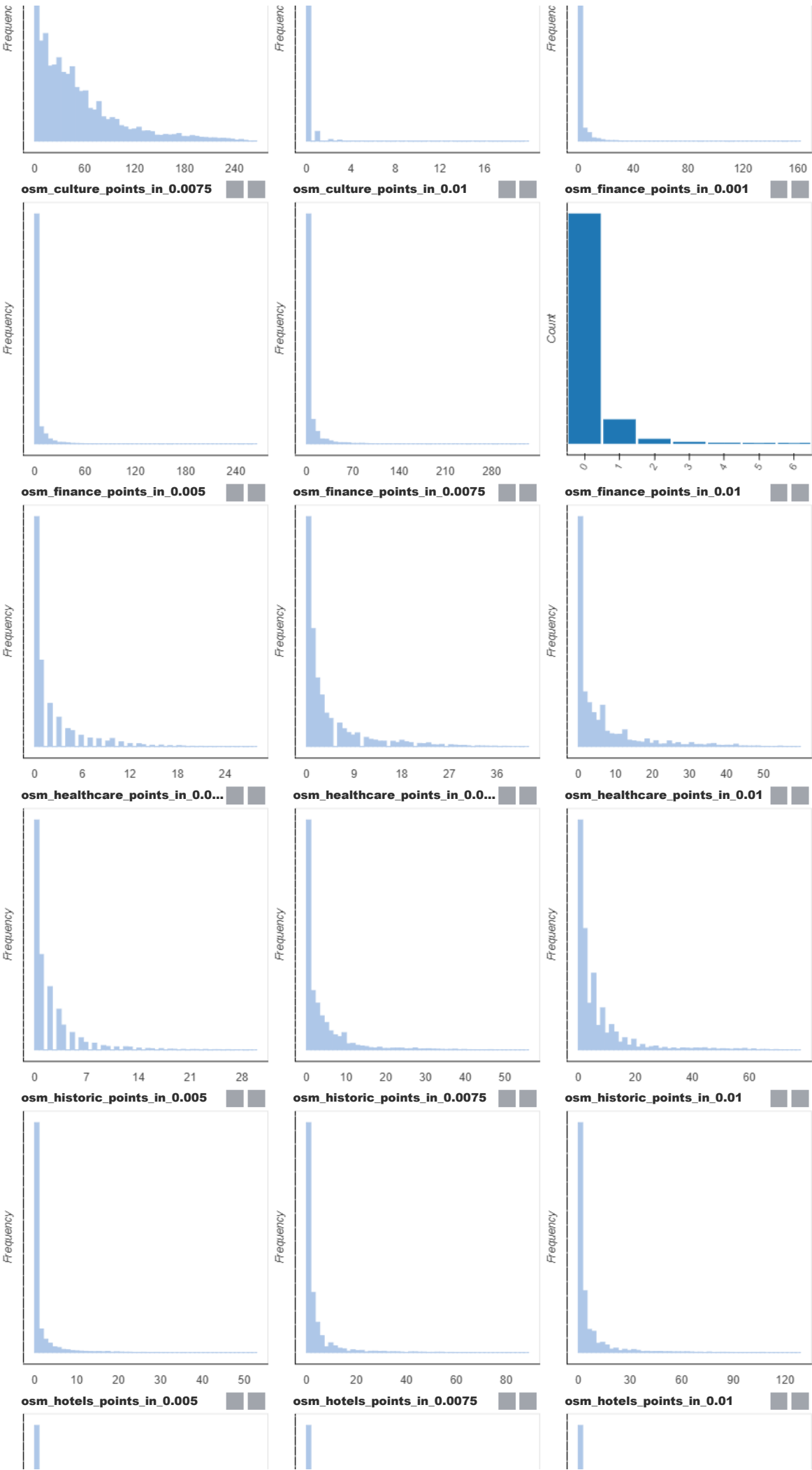
osm_amenity_points_in_0.001 and osm_offices_points_in_0.005 have similar distributions Similar Distribution osm_amenity_points_in_0.0075 and osm_shops_points_in_0.01 have similar distributions Similar Distribution osm_building_points_in_0.001 and osm_culture_points_in_0.001 have similar distributions Similar Distribution osm_building_points_in_0.001 and osm_train_stop_points_in_0.005 have similar distributions Similar Distribution osm_building_points_in_0.005 and osm_catering_points_in_0.001 have similar distributions Similar Distribution osm_building_points_in_0.0075 and osm_culture_points_in_0.005 have similar distributions Similar Distribution osm_building_points_in_0.01 and osm_historic_points_in_0.005 have similar distributions Similar Distribution osm_building_points_in_0.01 and osm_hotels_points_in_0.0075 have similar distributions Similar Distribution osm_culture_points_in_0.001 and osm_train_stop_points_in_0.005 have similar distributions Similar Distribution osm_finance_points_in_0.005 and osm_shops_points_in_0.001 have similar distributions Similar Distribution osm_finance_points_in_0.01 and osm_offices_points_in_0.0075 have similar distributions Similar Distribution osm_healthcare_points_in_0.005 and osm_leisure_points_in_0.005 have similar distributions Similar Distribution osm_healthcare_points_in_0.0075 and osm_leisure_points_in_0.0075 have similar distributions Similar Distribution osm_historic_points_in_0.005 and osm_hotels_points_in_0.0075 have similar distributions Similar Distribution osm_historic_points_in_0.0075 and osm_offices_points_in_0.005 have similar distributions Similar Distribution floor has 176237 (62.99%) missing values Missing reform_house_population_1000 has 14596 (5.22%) missing values Missing reform_house_population_500 has 27234 (9.73%) missing values Missing reform_mean_floor_count_1000 has 16708 (5.97%) missing values Missing reform_mean_floor_count_500 has 30168 (10.78%) missing values Missing reform_mean_year_building_1000 has 16239 (5.8%) missing values Missing reform_mean_year_building_500 has 29637 (10.59%) missing values Missing ... city has a high cardinality: 4948 distinct values High Cardinality floor has a high cardinality: 168 distinct values High Cardinality id has a high cardinality: 279792 distinct values High Cardinality osm_city_nearest_name has a high cardinality: 170 distinct values High Cardinality street has a high cardinality: 28841 distinct values High Cardinality osm_finance_points_in_0.001 has constant length 1 Constant Length date has constant length 10 Constant Length price_type has constant length 1 Constant Length id has all distinct values Unique osm_amenity_points_in_0.001 has 108604 (38.82%) zeros Zeros osm_amenity_points_in_0.005 has 17440 (6.23%) zeros Zeros osm_building_points_in_0.001 has 273479 (97.74%) zeros Zeros osm_building_points_in_0.005 has 220224 (78.71%) zeros Zeros osm_building_points_in_0.0075 has 188937 (67.53%) zeros Zeros osm_building_points_in_0.01 has 164807 (58.9%) zeros Zeros osm_catering_points_in_0.001 has 211258 (75.51%) zeros Zeros osm_catering_points_in_0.005 has 78425 (28.03%) zeros Zeros osm_catering_points_in_0.0075 has 51340 (18.35%) zeros Zeros osm_catering_points_in_0.01 has 37470 (13.39%) zeros Zeros osm_crossing_points_in_0.001 has 189497 (67.73%) zeros Zeros osm_crossing_points_in_0.005 has 39970 (14.29%) zeros Zeros osm_crossing_points_in_0.0075 has 24759 (8.85%) zeros Zeros osm_crossing_points_in_0.01 has 18199 (6.5%) zeros Zeros osm_culture_points_in_0.001 has 265293 (94.82%) zeros Zeros osm_culture_points_in_0.005 has 184687 (66.01%) zeros Zeros osm_culture_points_in_0.0075 has 146406 (52.33%) zeros Zeros osm_culture_points_in_0.01 has 119092 (42.56%) zeros Zeros osm_finance_points_in_0.005 has 133318 (47.65%) zeros Zeros osm_finance_points_in_0.0075 has 92889 (33.2%) zeros Zeros osm_finance_points_in_0.01 has 69505 (24.84%) zeros Zeros osm_healthcare_points_in_0.005 has 122621 (43.83%) zeros Zeros osm_healthcare_points_in_0.0075 has 83430 (29.82%) zeros Zeros osm_healthcare_points_in_0.01 has 62656 (22.39%) zeros Zeros osm_historic_points_in_0.005 has 162419 (58.05%) zeros Zeros osm_historic_points_in_0.0075 has 118207 (42.25%) zeros Zeros osm_historic_points_in_0.01 has 87518 (31.28%) zeros Zeros osm_hotels_points_in_0.005 has 191803 (68.55%) zeros Zeros osm_hotels_points_in_0.0075 has 154338 (55.16%) zeros Zeros osm_hotels_points_in_0.01 has 127295 (45.5%) zeros Zeros osm_leisure_points_in_0.005 has 122645 (43.83%) zeros Zeros osm_leisure_points_in_0.0075 has 82366 (29.44%) zeros Zeros osm_leisure_points_in_0.01 has 60814 (21.74%) zeros Zeros osm_offices_points_in_0.001 has 237130 (84.75%) zeros Zeros osm_offices_points_in_0.005 has 111467 (39.84%) zeros Zeros osm_offices_points_in_0.0075 has 76352 (27.29%) zeros Zeros osm_offices_points_in_0.01 has 56099 (20.05%) zeros Zeros osm_shops_points_in_0.001 has 138828 (49.62%) zeros Zeros osm_shops_points_in_0.005 has 26979 (9.64%) zeros Zeros osm_shops_points_in_0.0075 has 17337 (6.2%) zeros Zeros osm_train_stop_points_in_0.005 has 269647 (96.37%) zeros Zeros osm_train_stop_points_in_0.0075 has 258422 (92.36%) zeros Zeros osm_train_stop_points_in_0.01 has 243840 (87.15%) zeros Zeros osm_transport_stop_points_in_0.005 has 40729 (14.56%) zeros Zeros osm_transport_stop_points_in_0.0075 has 23541 (8.41%) zeros Zeros osm_transport_stop_points_in_0.01 has 17247 (6.16%) zeros Zeros reform_count_of_houses_1000 has 14596 (5.22%) zeros Zeros reform_count_of_houses_500 has 27234 (9.73%) zeros Zeros

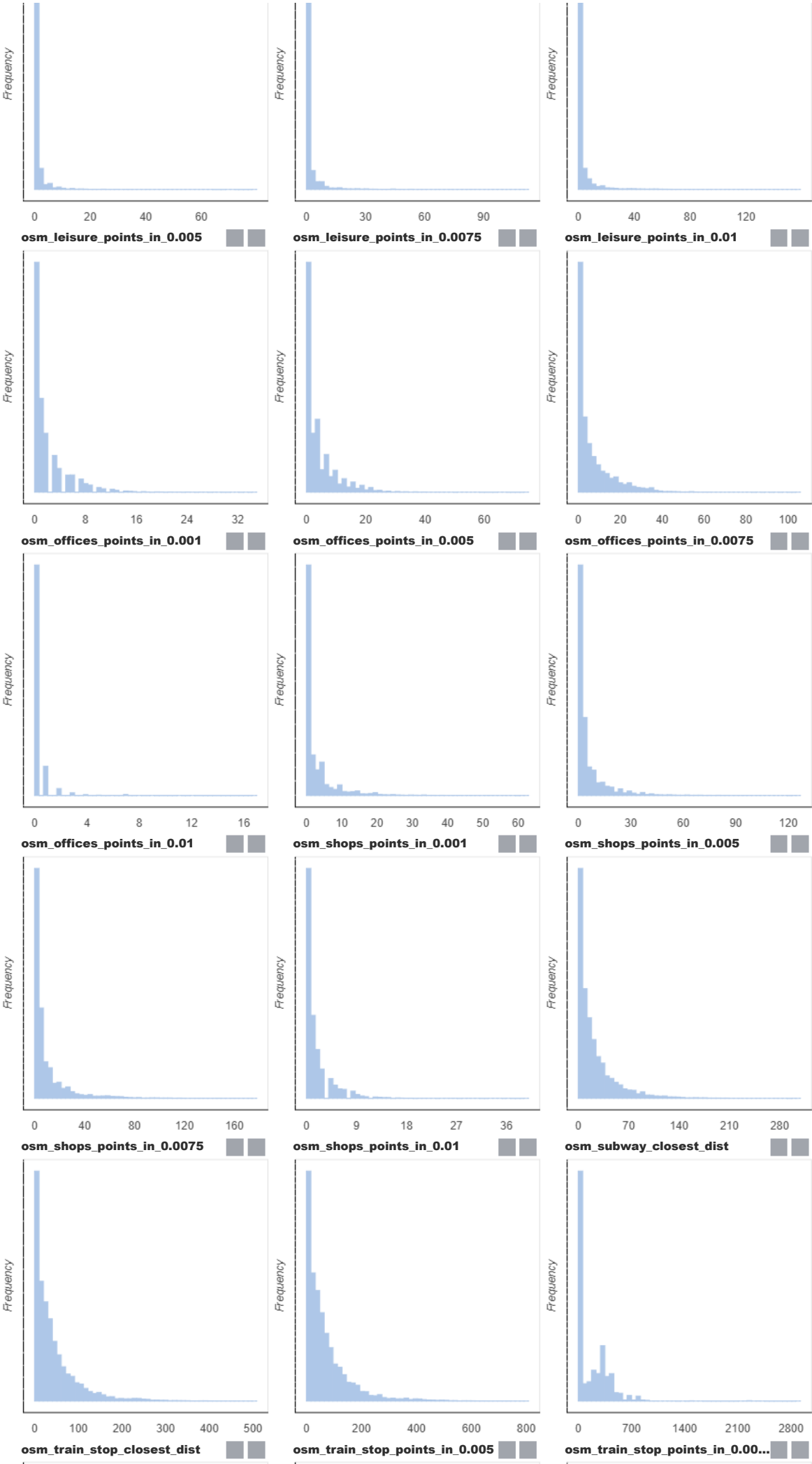
```
1 plot(df)
```

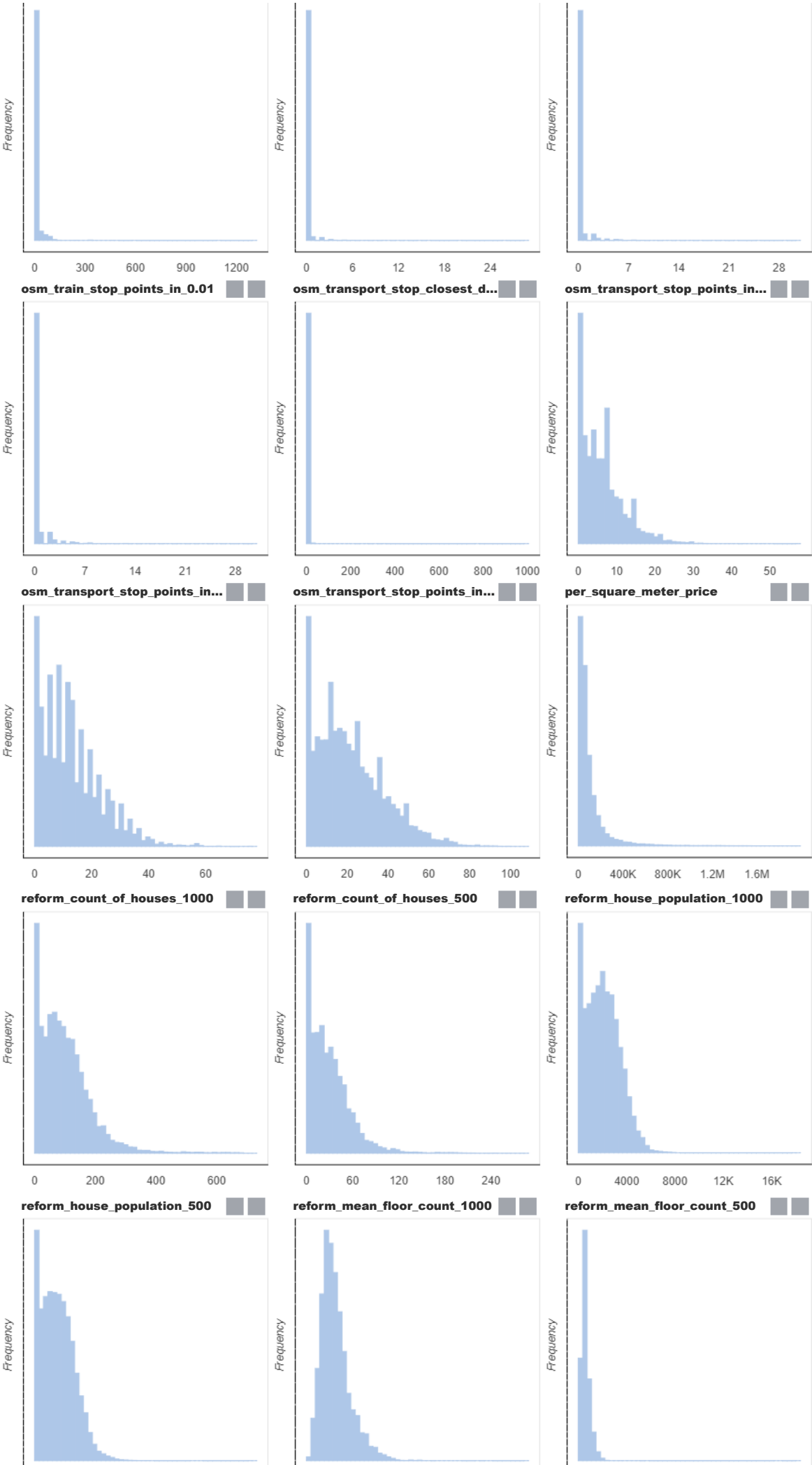


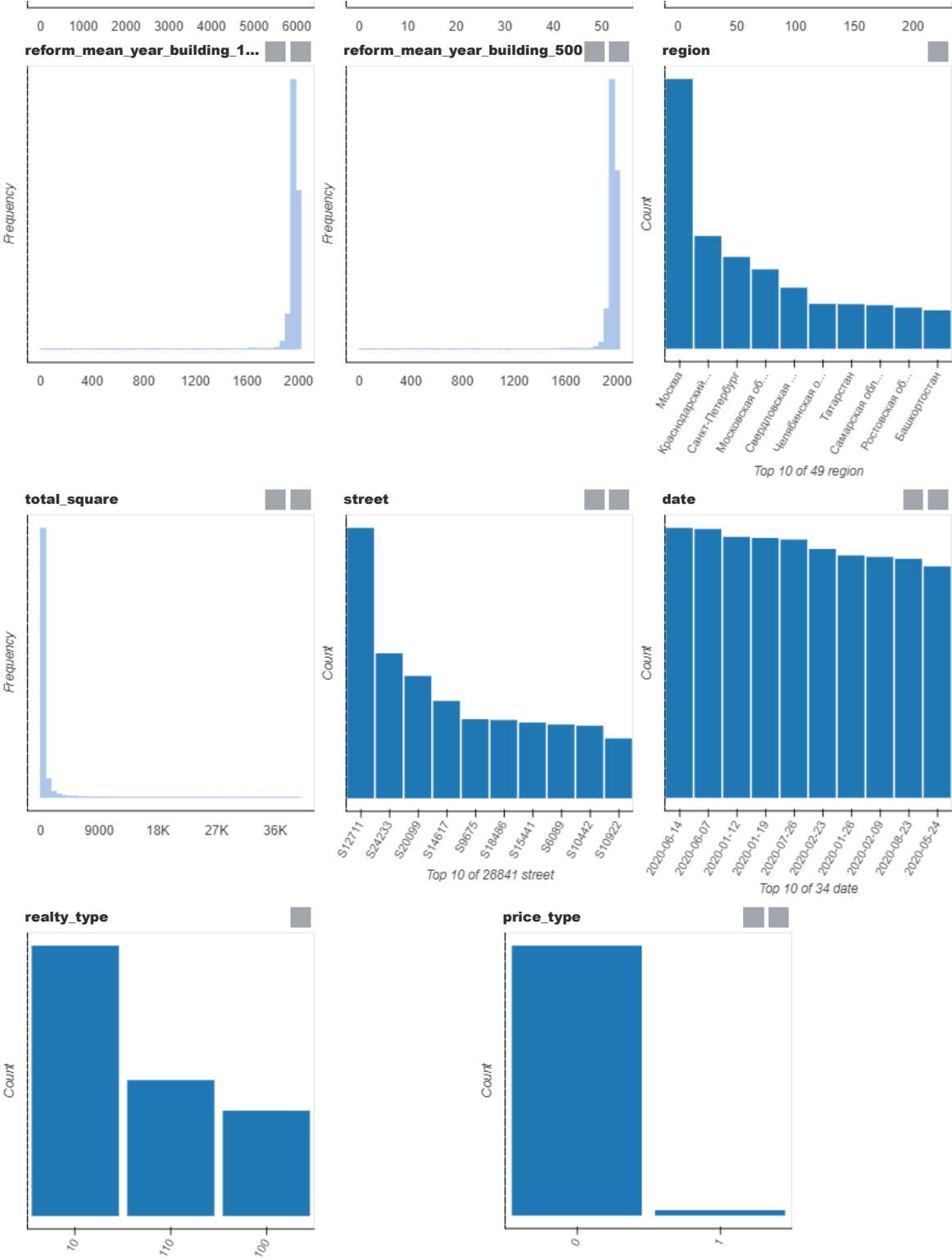












✓ 56 сек. выполнено в 01:22

● x