3-6 2017 ОКТЯБРЯ
ОТЕЛЬ «ПАРК ИНН ПРИБАЛТИЙСКАЯ»

Технологии Big Data в работе аналитика агентства недвижимости

Котриков Кирилл Юрьевич

Руководитель отдела аналитики,
Century 21 Россия с 25/12/2015

Образование:
СПбГУ, Экономическая социология
UCL, Comparative Business Economics
НИУ ВШЭ, Машинное обучение и анализ больших данных

- Big data, Machine learning, Data Mining
- Язык программирования Python
- Научные библиотеки Python
- Как сделать первый шаг
- Примеры практического применения
- Другие Open Source продукты

Что такое Big data и почему это важно (и не страшно)

Big data, термин который описывает большие массивы данных, как структурированные, так и неструктурированные, которые наводняют бизнес ежедневно. Но дело не в объеме. А в том, каким образом организации извлекают пользу из данных которая способствует принятию лучших решений и стратегических действий.

SAS

Big data, это новое, безразмерное и страшное, очень, очень страшное. Нет, подождите… Big data, это очередное название для старого доброго анализа данных, который используют все аналитики, и не такие они и большие, и это то, что мы должны изучить а не бояться. Подождите… И это не то… Что я хотел сказать, так это то, что Big data мощное как цунами, но это потоп, который можно контролировать и использовать для создания стоимости

Forbes

Большие данные — совокупность подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста (объемов)…

Wiki

Разработка и майнинг данных

KDD (knowledge discovery in databases)- Разработка данных/Генерирование знаний. Нетривиальный процесс идентификации актуальных, новых, потенциально полезных и очень понятных закономерностей/характеристик.

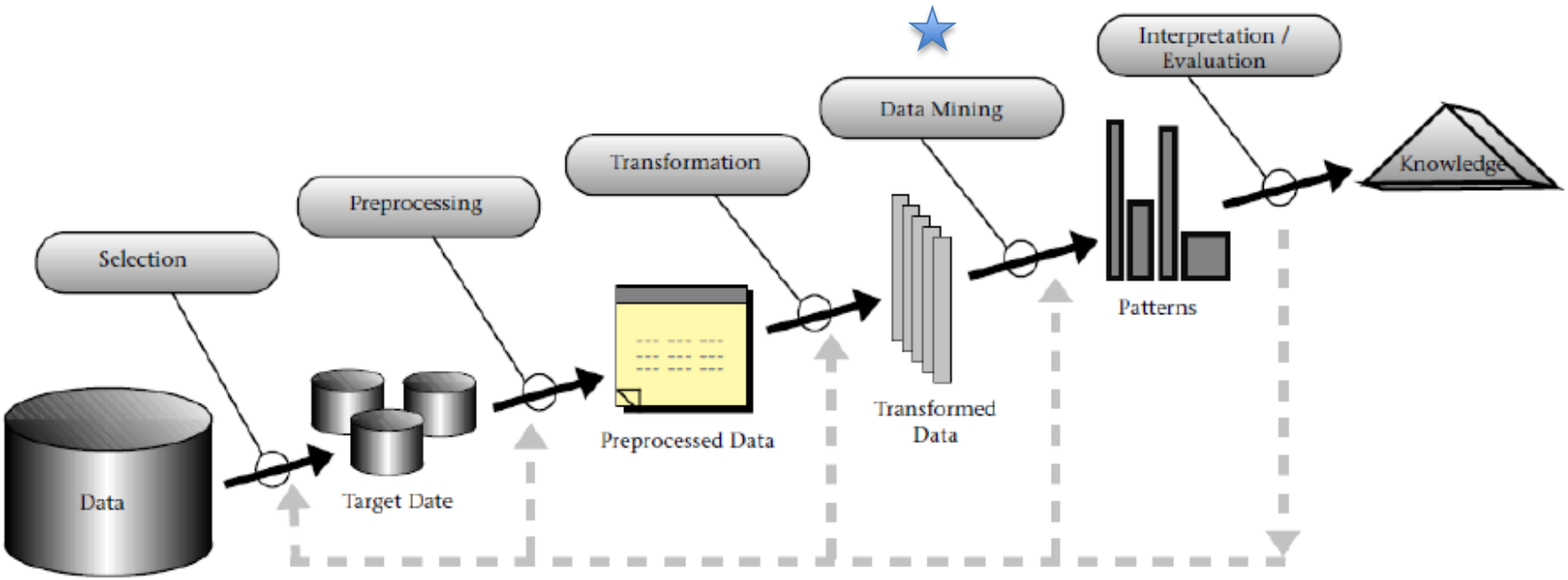
Файад, Пятецкий-Шапиро и Смит 1996

Data mining – Майнинг, это часть процесса разработки данных ,заключающаяся в использованием методов математического и статистического анализа с целью выявления зависимостей и создания моделей.

Файад, Пятецкий-Шапиро и Смит 1996

Machine Learning (с учителем) - обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.

Разработка и майнинг данных



Язык программирования Python (Питон) – инструмент безграничных возможностей и молниеносный анализ данных

Python - высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода. Синтаксис ядра Python минималистичен. В то же время наличие большого количества библиотек делает язык универсальным для решения различных задач.

В аналитике это :

- Анализ данных (больших данных), текста, изображений
- Манипуляции с массивами данных
- Визуализация данных
- Написание собственных алгоритмов
- Алгоритмы по решению рутинных задач
- Web scraping и Parsing
- и многое другое

Библиотеки для работы с данными

NumPy

Фундаментальные пакет
Для научных расчётов
С поддержкой многомерных массивов

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

- Манипуляции с массивами данных
- Работа с отсутствующими значениями
- Совмещение данных
- Сводные таблицы
- Срез, индексирование, выборки
- Работа с временными рядами
- И т.д.

ANACONDA

jupyter

scikit learn

- Классификация
- Регрессия
- Кластеризация
- Снижение размерности
- Перекрестная проверка
- Оценка модели

matplotlib

seaborn

Визуализация данных

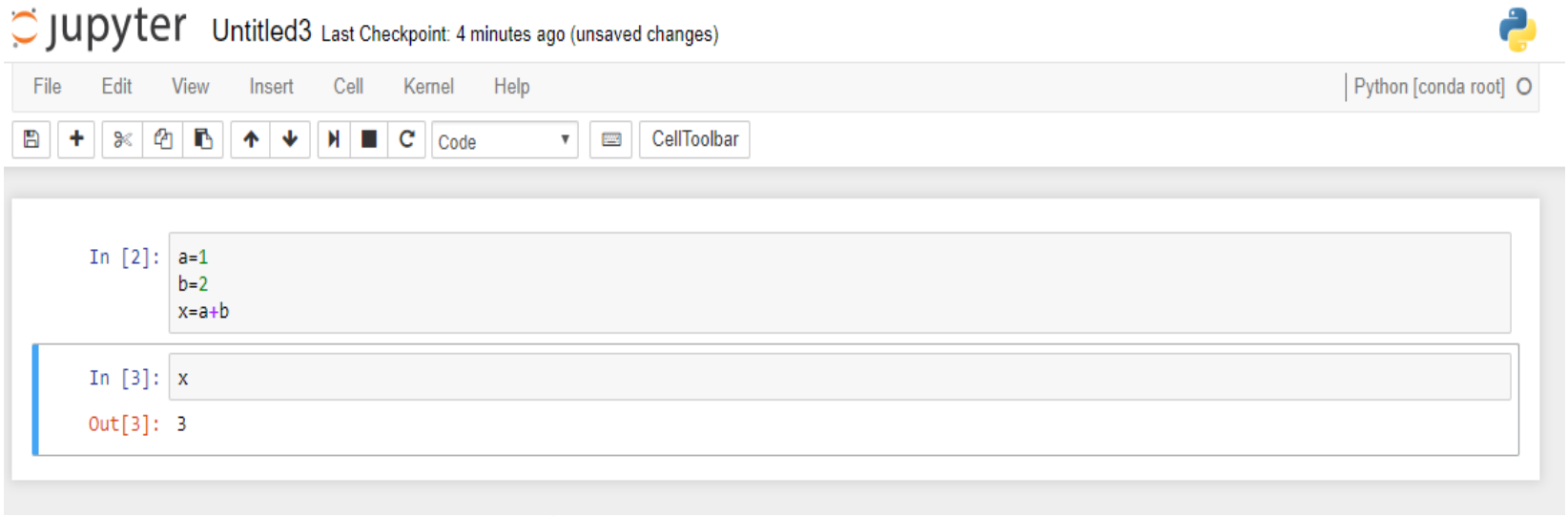
РЕЗУЛЬТАТ

С чего начать...

1 ANACONDA

OPEN SOURCE !

2 jupyter



3 Установка библиотек

In [ ]:
```
!pip install pandas
!pip install numpy
!pip install seaborn
# есть и другие варианты установки
```

4 Выбор библиотек для работы

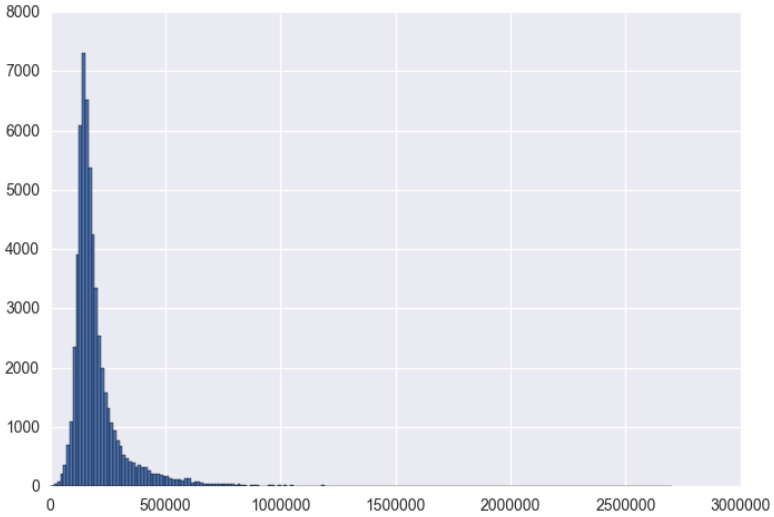
In [ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
# И так далее
```

Пример (логнормальное распределение Московского рынка)

```
Moscow=pd.read_excel('C:\Users\kirill
```

```
%time
Moscow['sqmprice'].hist(bins=200)
plt.show()
```

```
Wall time: 0 ns
```

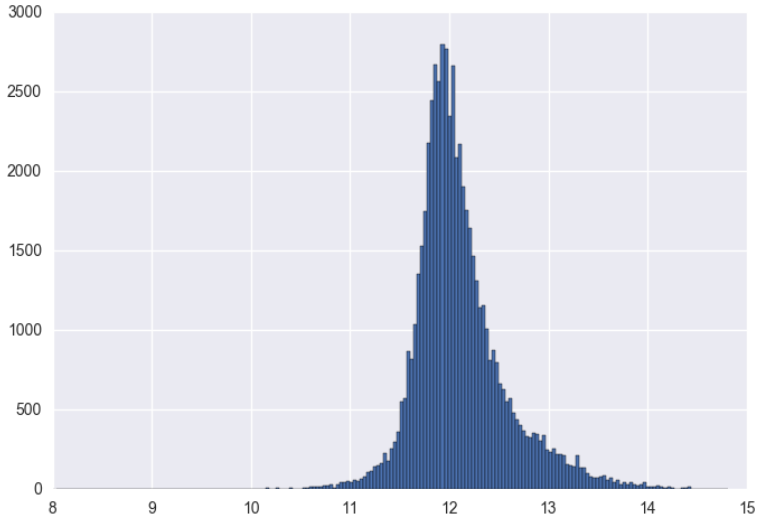



```
%time
Moscow['log_sqmprice']=np.log(Moscow.sqmprice)
```

```
Wall time: 0 ns
```

```
%time
Moscow['log_sqmprice'].hist(bins=200)
plt.show()
```

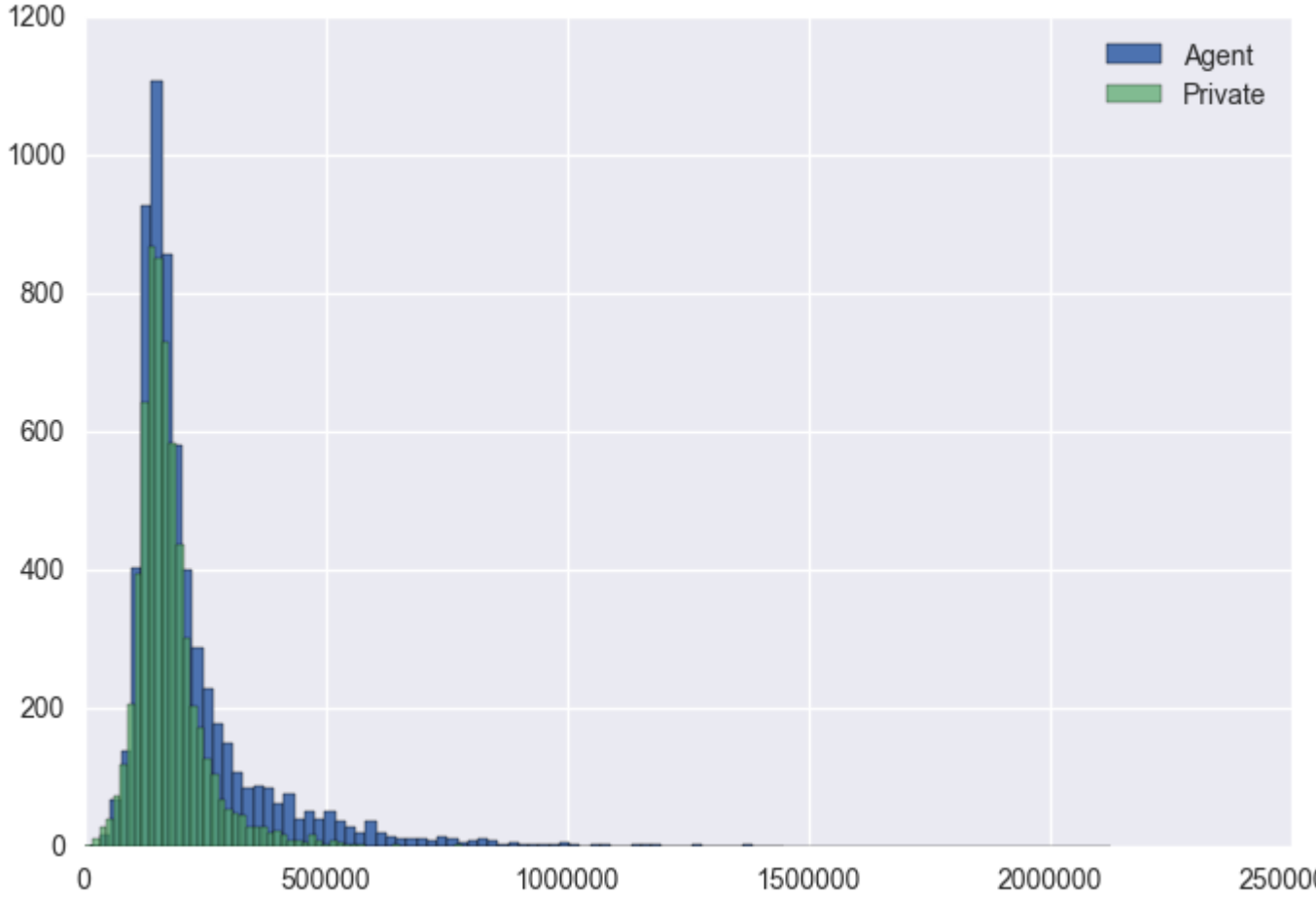
```
Wall time: 0 ns
```




```
%time
print 2.718281**(np.mean(Moscow.log_sqmprice)+((np.std(Moscow.log_sqmprice))**2)/2)
print 2.718282**(np.mean(Moscow.log_sqmprice))
```

```
Wall time: 0 ns
203678.61689
182094.422904
```

Пример (распределение объектов продаваемых агентом и собственником по цене)



80% объектов в Москве продается с участием посредников (в ЦАО- 85%)

Стоимость квадратного метра при продаже с агентом на 17% дороже

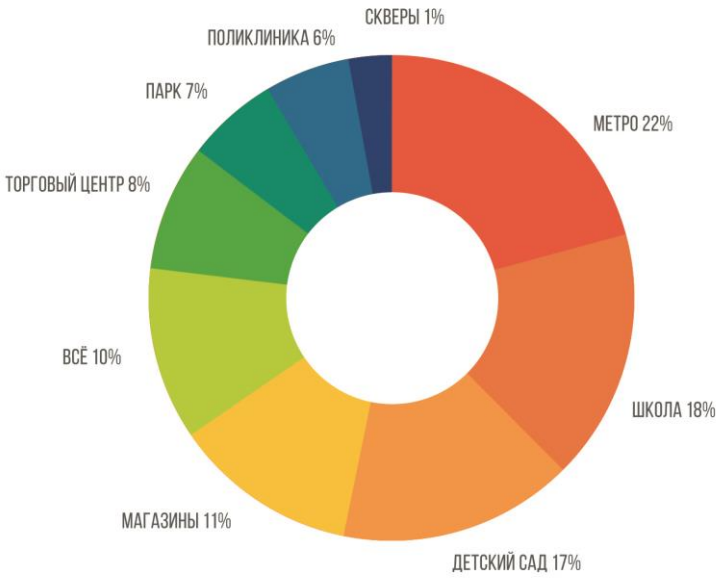
Время экспозиции 1 к 1,5

Примеры практического применения

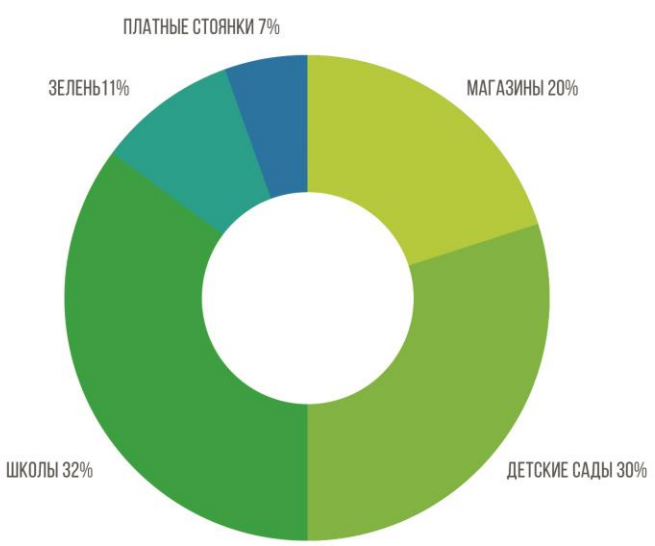
(анализ текста/text mining, как часть разработки самообучающейся модели для прогнозирования времени экспозиции объекта/определения стоимости/создания продающего объявления и т.д.)

Данные: тексты 10 000 объявлений о продаже квартир в Москве

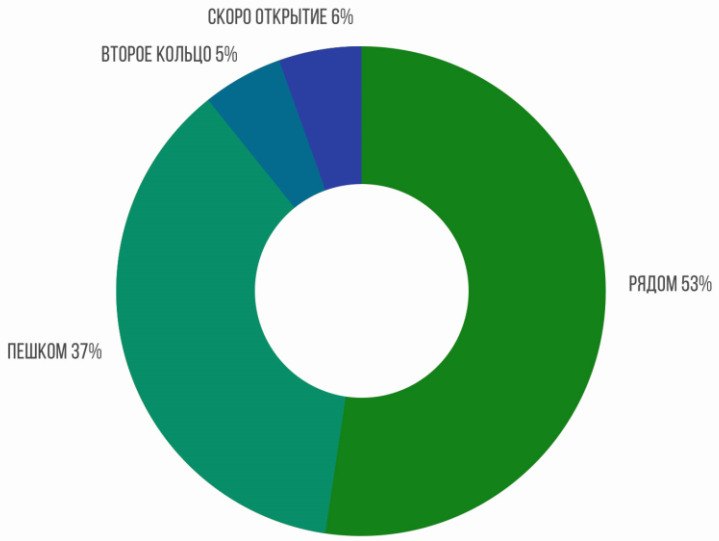
СЛОВА, ХАРАКТЕРИЗУЮЩИЕ ОБЪЕКТЫ РЯДОМ



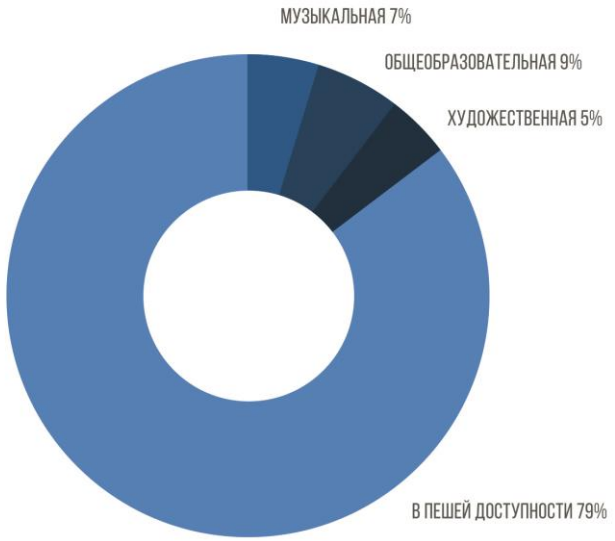
СЛОВА, ИДУЩИЕ В ПАРЕ СО СЛОВОМ «МНОГО»



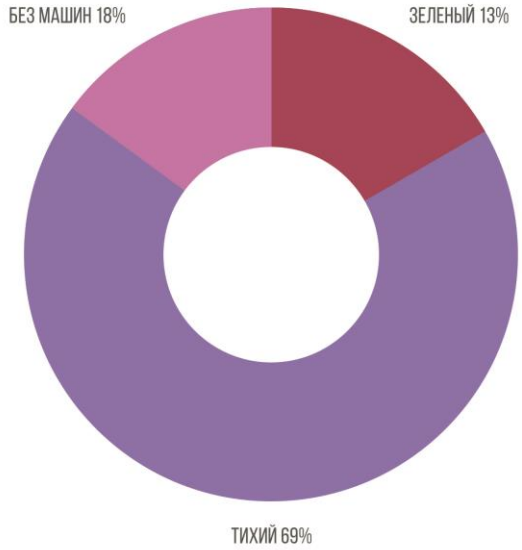
СЛОВА, ХАРАКТЕРИЗУЮЩИЕ МЕТРО



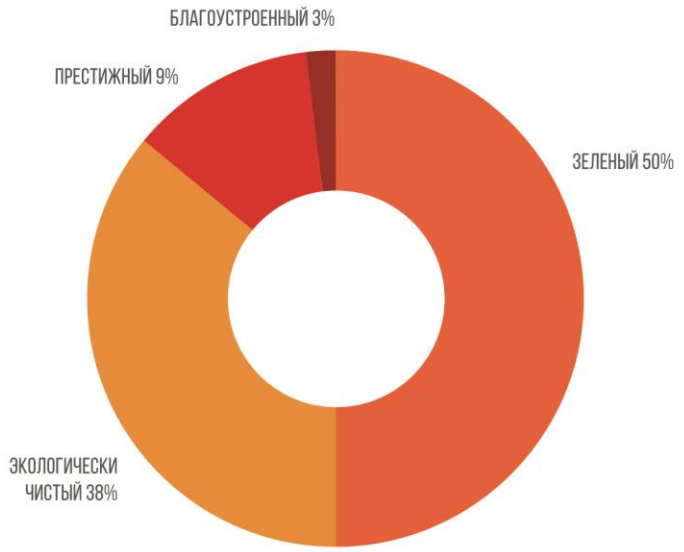
СЛОВА, ХАРАКТЕРИЗУЮЩИЕ ШКОЛУ



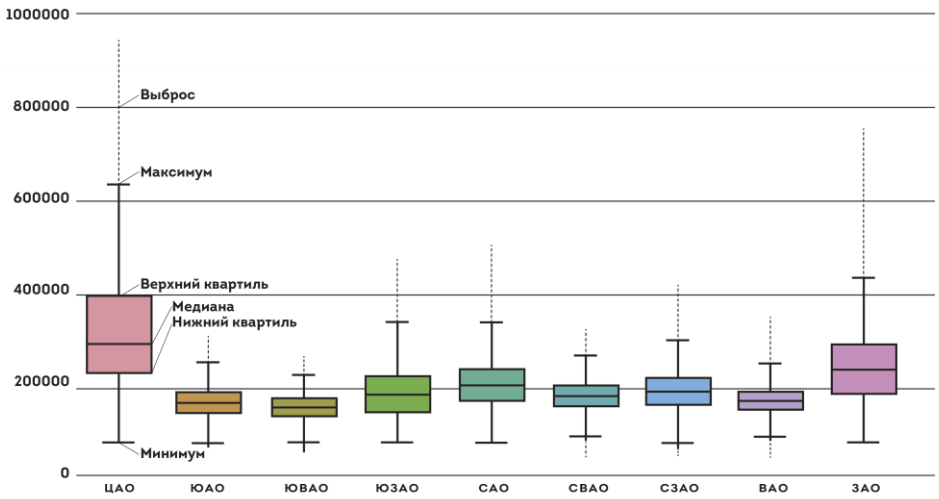
СЛОВА, ХАРАКТЕРИЗУЮЩИЕ ДВОР



СЛОВА, ХАРАКТЕРИЗУЮЩИЕ РАЙОН



Примеры практического применения
(разработка аналитических продуктов)



Средняя стоимость квадратного метра на вторичном рынке в зависимости от района и комнатности (на 19/06/2017)

Район	Студия	1-комнатные	2-комнатные	3-комнатные	4-комнатные
Ленинский	68 878	67 299	62 579	60 429	68056
Дзержинский	67 827	56 092	51 475	50 104	54 358
Свердловский	53 999	54 204	50 873	54 519	55 286
Мотовилихинский	61 660	55 971	49 108	46 403	43 348
Индустриальный	58 614	53 649	48 945	45 687	47 343
Кировский	54 047	45 149	41 963	40 493	38 205
Орджоникидзевский	47 602	41 674	37 294	34 906	35 684

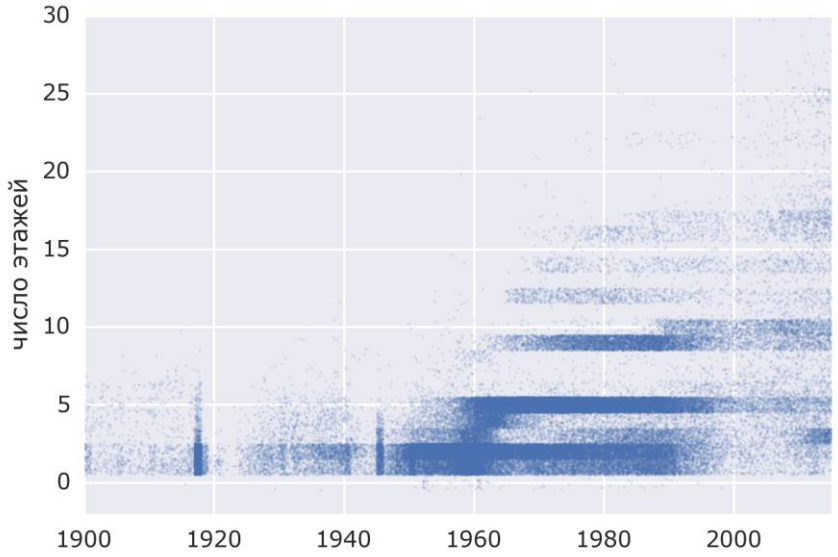
С использованием стандартных алгоритмов, создание аналитического отчета по конкретному региону
Занимает не более 20 минут

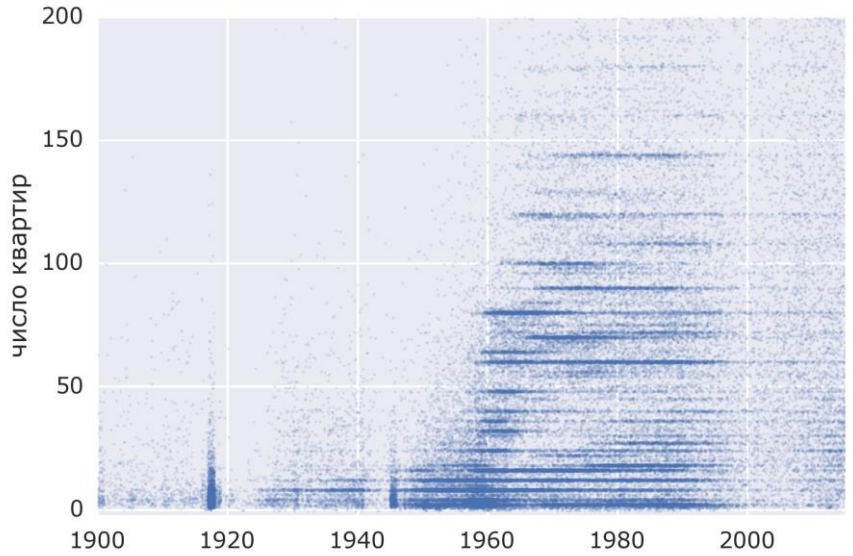
Примеры практического применения

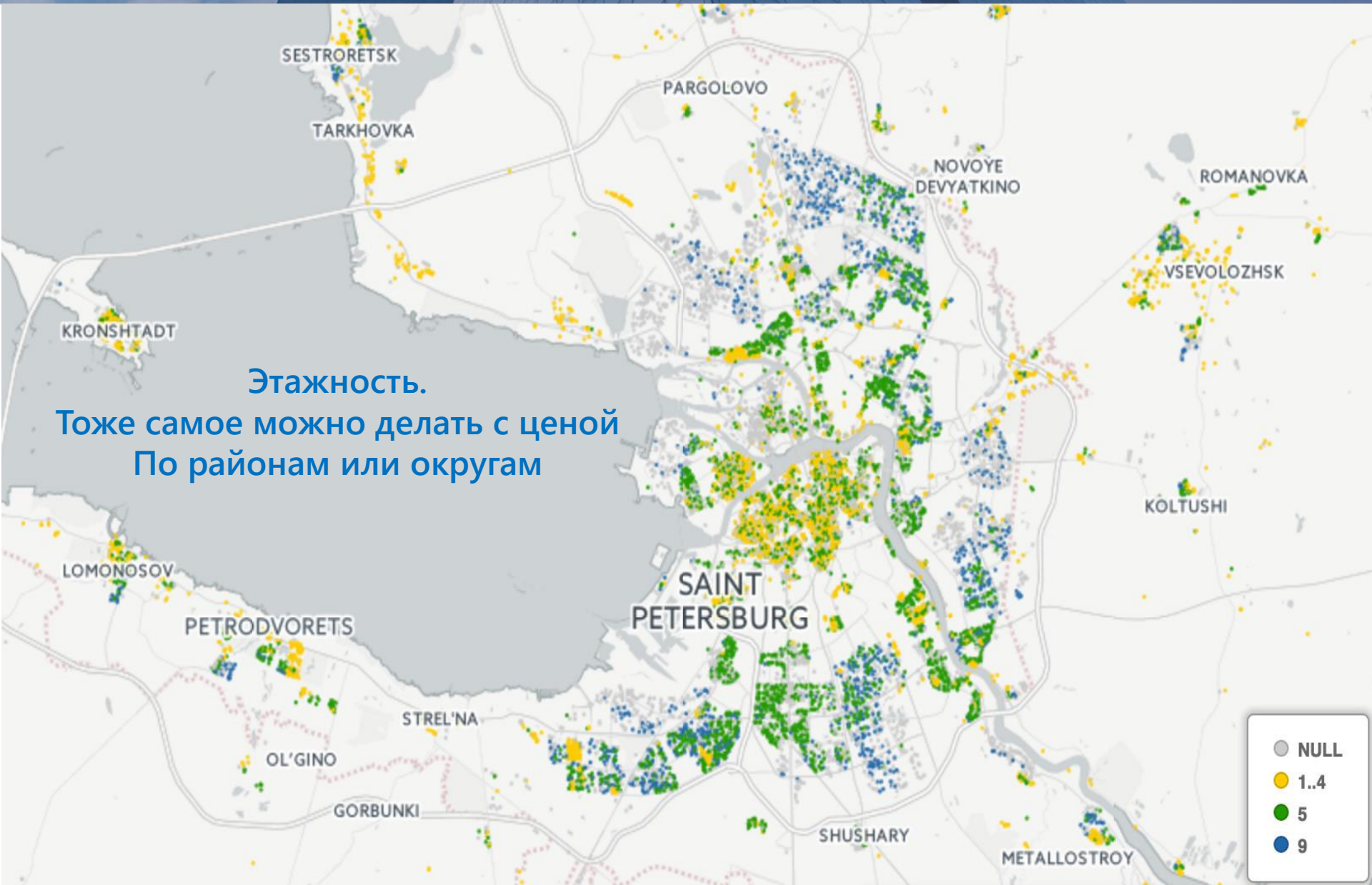
Работа с большими массивами

Результаты веб-скрэпинга и парсинга сайта www.reformagkh.ru

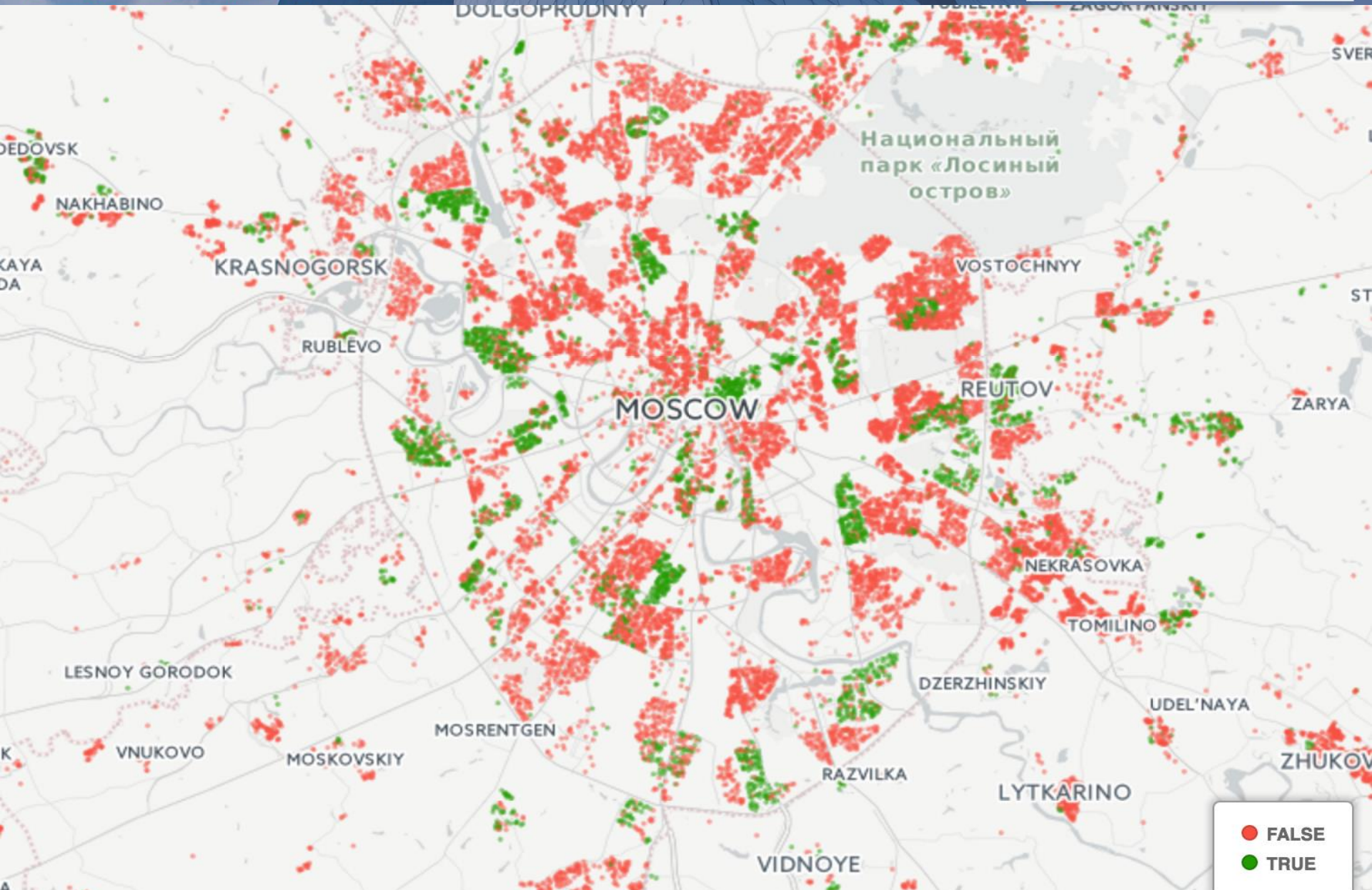
(данные по 1 млн. домов)

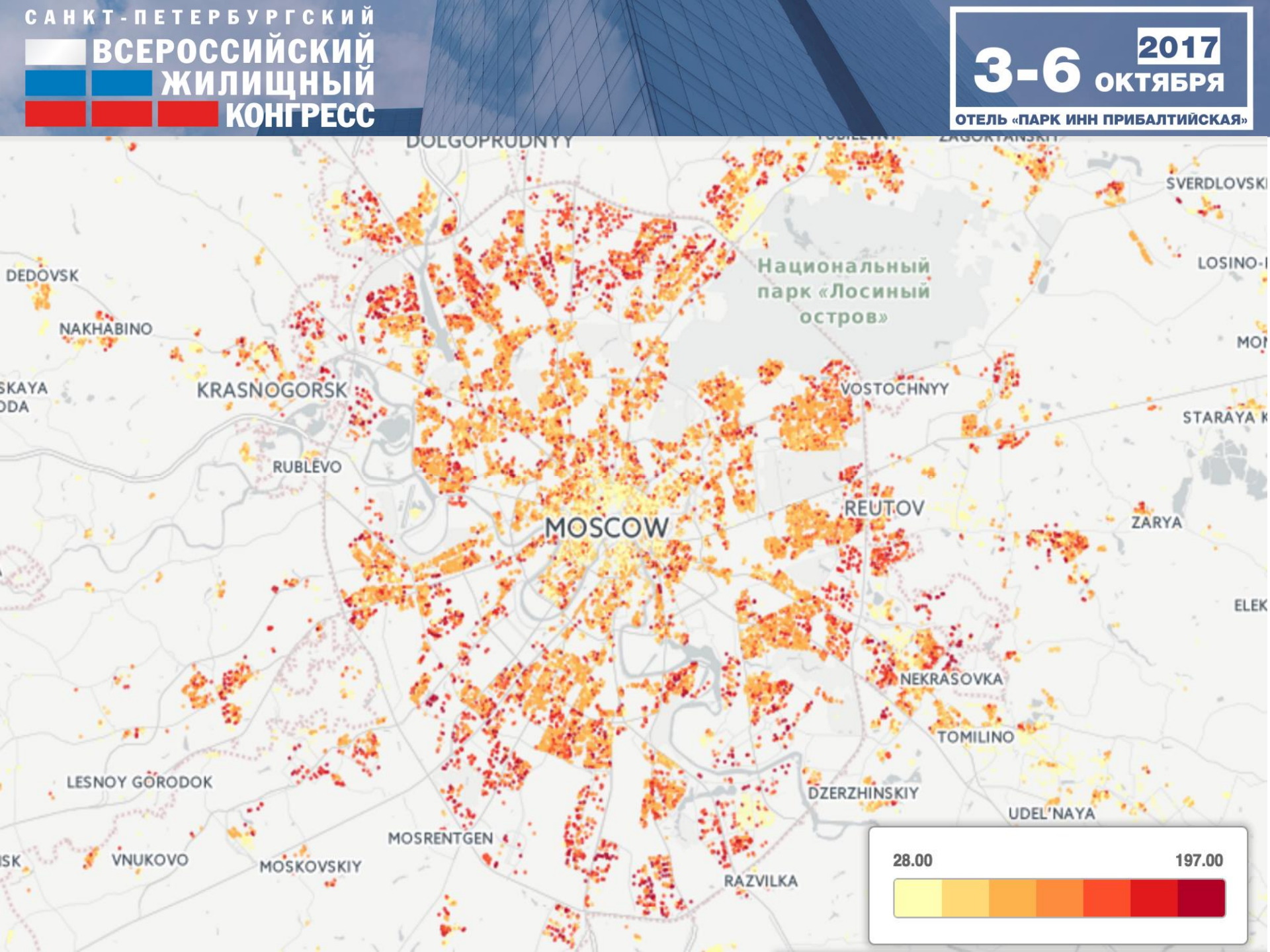






Этажность.
Тоже самое можно делать с ценой
По районам или округам


DOLGOPRUDNYY
Национальный парк «Лосиный остров»
NAKHABINO
KRASNOGORSK
RUBLEVO
MOSCOW
VOSTOCHNYY
REUTOV
ZARYA
NEKRASOVKA
TOMILINO
DZERZHINSKIY
UDEL'NAYA
LESNOY GORODOK
MOSRENTGEN
VNUKOVO
MOSKOVSKIY
RAZVILKA
LYTKARINO
VIDNOYE
FALSE
TRUE




```
train.shape
```

```
(30471, 292)
```

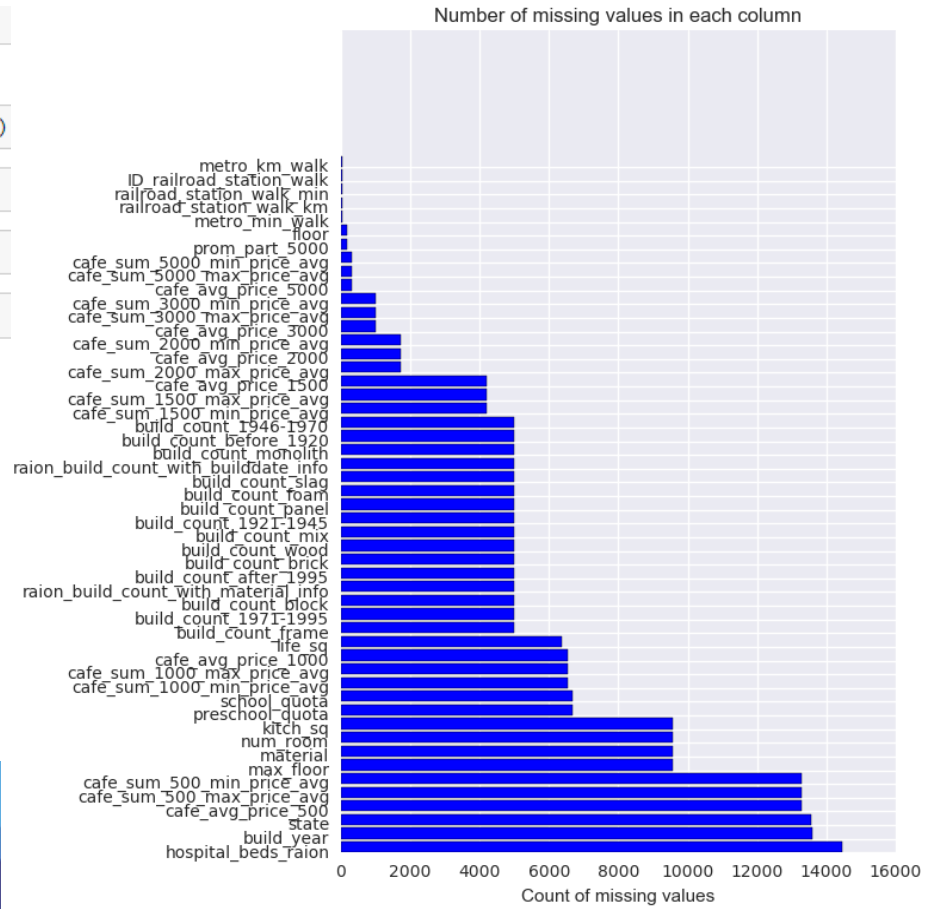
```
missing1=pd.DataFrame(train.isnull().sum().sort_values(ascending=False).reset_index())
```

```
missing1.columns=['var_name','missing_count']
```

```
missing1=missing1.loc[missing1.missing_count!=0,:]
```

```
missing1
```

	var_name	missing_count
0	hospital_beds_raion	14441
1	build_year	13605
2	state	13559
3	cafe_avg_price_500	13281
4	cafe_sum_500_max_price_avg	13281
5	cafe_sum_500_min_price_avg	13281
6	max_floor	9572
7	material	9572
8	num_room	9572



Другие Open Source продукты для работы с большими данными

https://www.r-project.org/

https://www.knime.com/

https://orange.biolab.si/

Не ждите, меняйтесь!

Спасибо!