

Государственное бюджетное профессиональное
образовательное учреждение Московской области
«Физико-технический колледж»

Аналитический отчет

**«Модель оценки цены квартиры на
вторичном рынке по Московскому региону:
Москва, Новая Москва, Московская область»**

Работу выполнила:

Студент группы ИСП-21
Замараева Виктоия

Долгопрудный 2024

Введение

В данном аналитическом отчете рассматривается модель оценки квартир по Московской области, Москве и Новой Москве. Московский регион является одним из крупнейших рынков недвижимости в России, привлекающий местных жителей и приезжих.

Цель:

Собрать данные и произвести анализ данных для построения модели, которая будет оценивать цену квадратного метра недвижимости в Московском регионе.

Задачи:

- Используя открытые источники (Циан) составить список параметров, значительно влияющих на цену квадратного метра жилой площади.
- С учетом выявленных факторов произвести парсинг данных по квартирам на продажу, используя различные парсеры.
- Произвести подготовку данных для анализа.
- Визуализировать взаимосвязь между ними; определить признаки, оказывающие наиболее сильное влияние.

Основная часть

Для выполнения задачи парсинга данных с сайта Циан, я сначала получила доступ к необходимой информации на сайте. Затем, мы с одноклассницей с помощью программы парсера `cianparser` извлекли все данные по квартирам. После этого выявила данные которые больше всех будут влиять на цену квартиры.

Для выполнения задачи подготовки данных для анализа я использовала несколько библиотек `pandas`, `numpy` и другие. В начале работы у меня возникли трудности с кодировкой данных, но с помощью онлайн редактора я решила эту проблему и начала работать с данными.

В начале я выведу какие данные у меня есть всего и какого они типа.

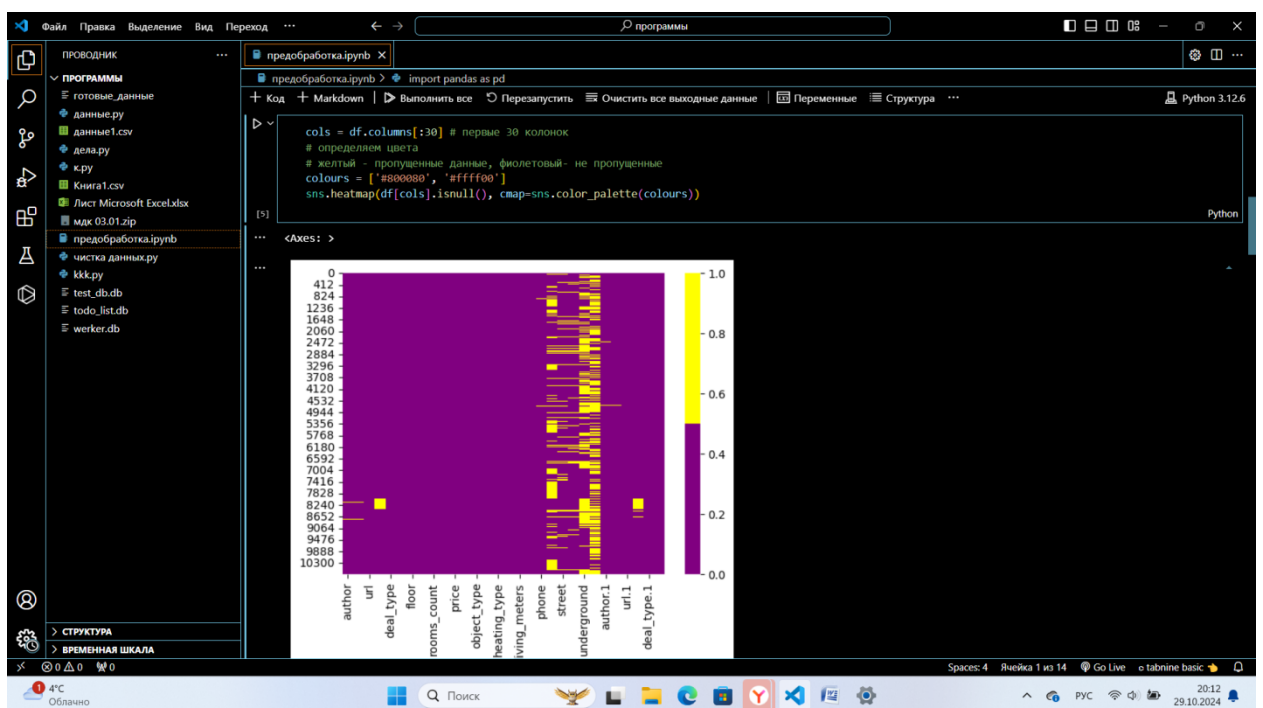
```
File Edit View Transition ...
предобработка.ipynb
import pandas as pd

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10696 entries, 0 to 10695
Data columns (total 48 columns):
 #   Column                Non-Null Count  Dtype
---  ---
 0   author                10695 non-null  object
 1   author_type           10695 non-null  object
 2   url                   10684 non-null  object
 3   location              10278 non-null  object
 4   deal_type             10686 non-null  object
 5   accommodation_type    10684 non-null  object
 6   floor                 10682 non-null  float64
 7   floors_count          10682 non-null  float64
 8   rooms_count           10682 non-null  float64
 9   total_meters          10682 non-null  object
10  price                 10662 non-null  float64
11  year_of_construction  10682 non-null  object
12  object_type           10682 non-null  object
13  house_material_type   10682 non-null  object
14  heating_type          10682 non-null  object
15  finish_type           10682 non-null  object
16  living_meters          10682 non-null  object
17  kitchen_meters         10682 non-null  object
18  phone                 10677 non-null  float64
19  district              7060 non-null   object
...
46  underground.1         6361 non-null   object
47  residential_complex.1 5125 non-null   object
dtypes: float64(10), object(38)
memory usage: 3.94 MB

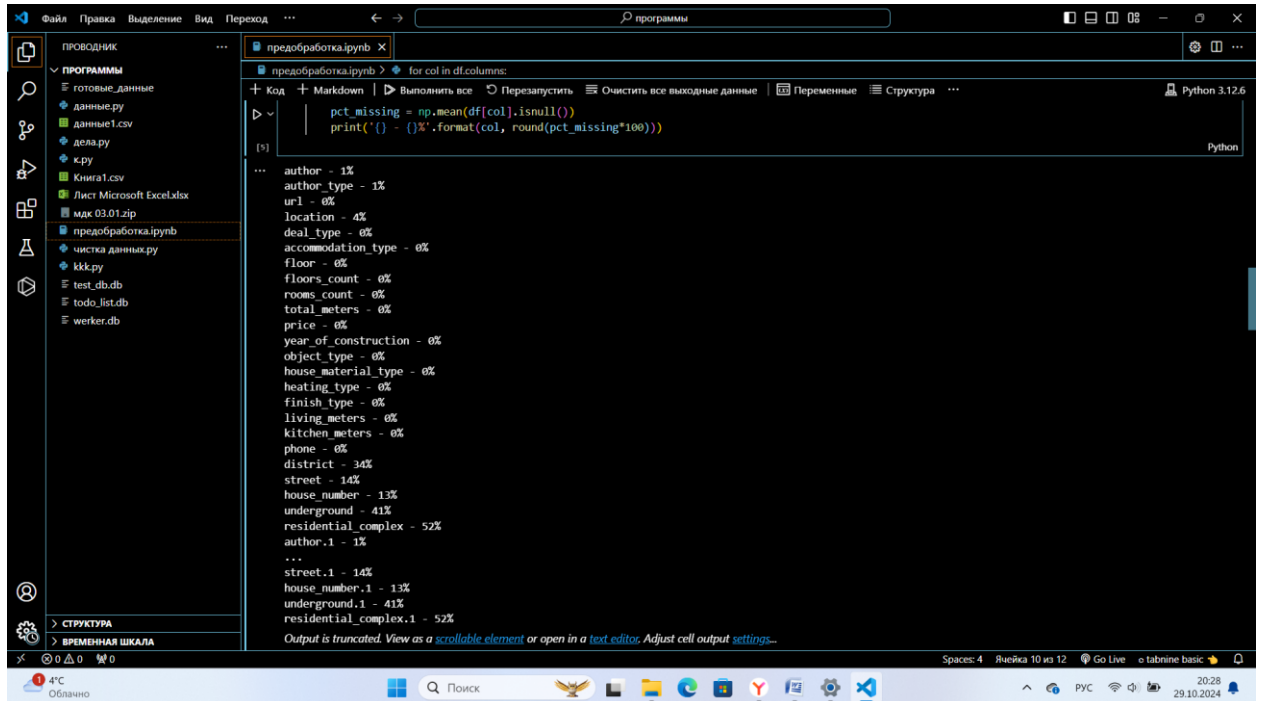
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Вывожу на тепловую карту пропущенных значений и смотрю какие столбцы или строки удалить.



Больше всего пропущенных данных в колонке phone, street и underground.

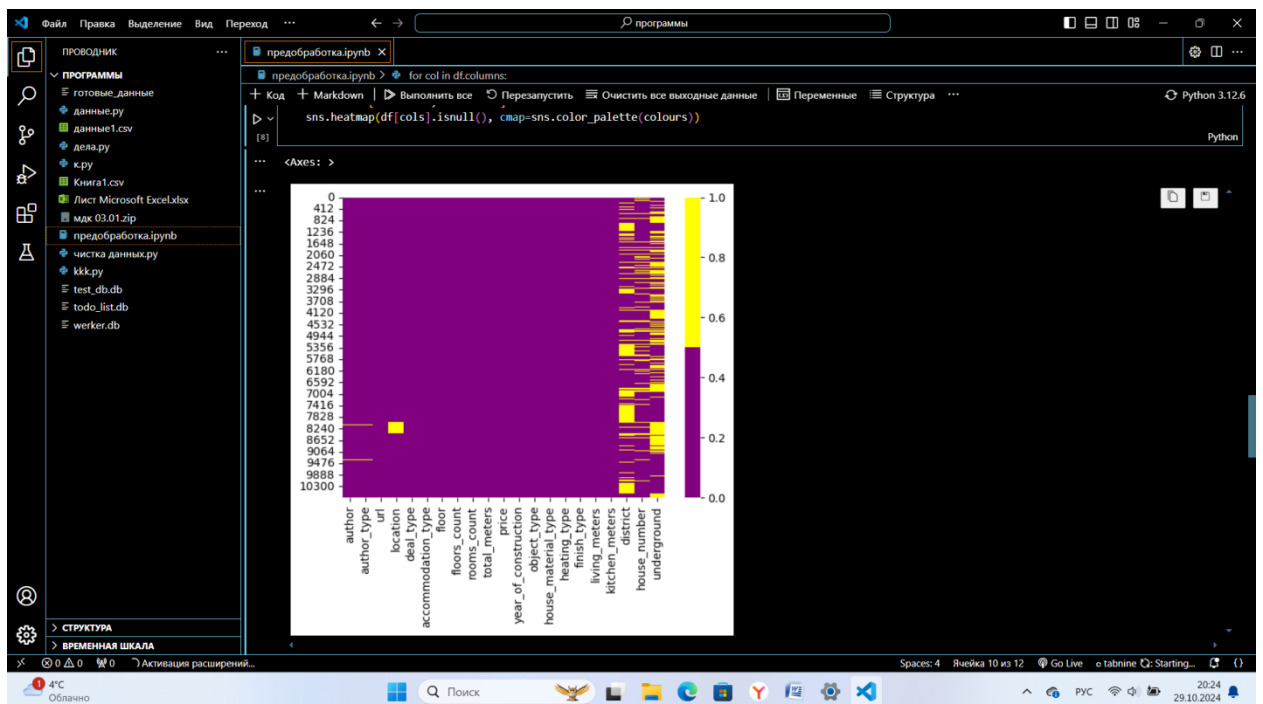
Далее вывожу процентное содержание пропусков.



```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}'.format(col, round(pct_missing*100)))
```

```
author - 1%
author_type - 1%
url - 0%
location - 4%
deal_type - 0%
accommodation_type - 0%
floor - 0%
floors_count - 0%
rooms_count - 0%
total_meters - 0%
price - 0%
year_of_construction - 0%
object_type - 0%
house_material_type - 0%
heating_type - 0%
finish_type - 0%
living_meters - 0%
kitchen_meters - 0%
phone - 0%
district - 34%
street - 14%
house_number - 13%
underground - 41%
residential_complex - 52%
author.1 - 1%
...
street.1 - 14%
house_number.1 - 13%
underground.1 - 41%
residential_complex.1 - 52%
```

Потом удаляю выбранные столбцы и дубликаты столбцов и опять смотрю в тепловую карту и удаляю дубликаты.



```
df_unique_subset = df.drop_duplicates() #удаление дубликатов
print(df_unique_subset)
```

	author	author_type	url	location	deal_type
0	P"PyCf-P"PyPIPyP"PyPSC,	developer			
1	Point Estate P"PsC"Pr	real_estate_agent			
2	PC,P"PrP" PyPScP"PrP	real_estate_agent			
3	Sminex	developer			
4	Whitewill	real_estate_agent			
...
10691	Наталья Бондаренко	realtor			
10692	Евгений Цопин	realtor			
10693	OPN PARTNER	real_estate_agent			
10694	OPN PARTNER	real_estate_agent			
10695	ПРТЬ ЗВЕЗД	real_estate_agent			

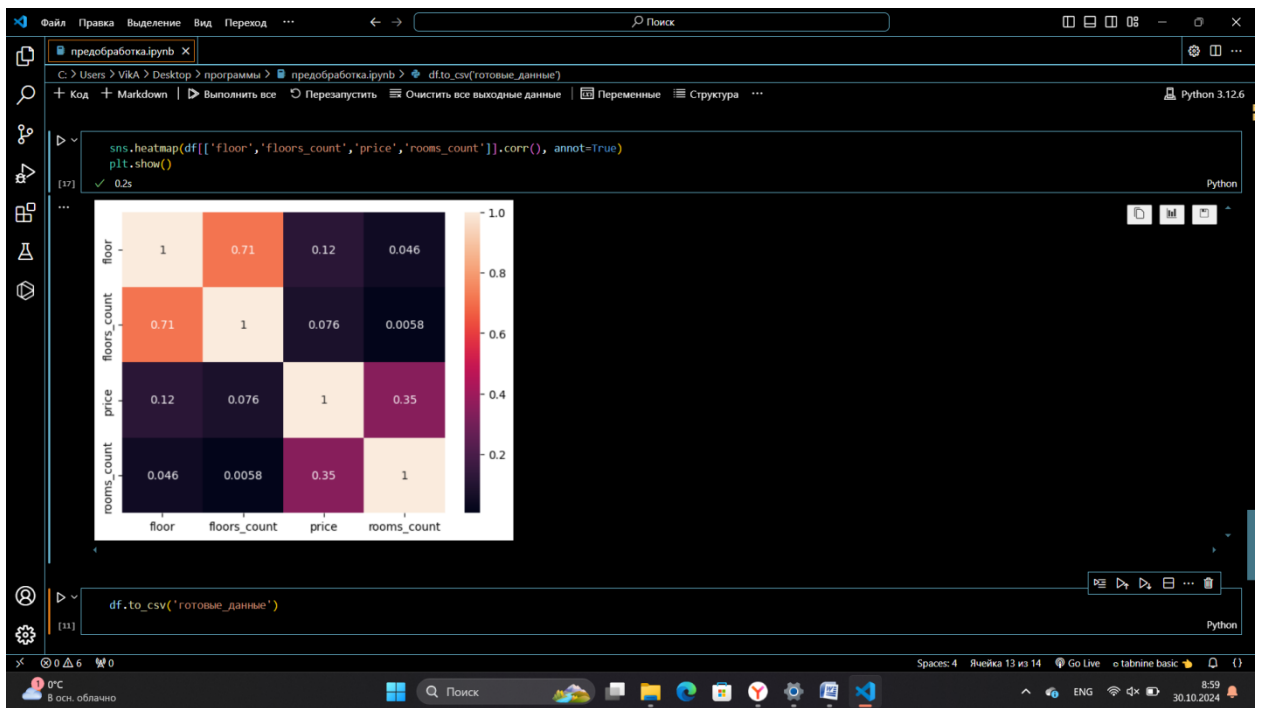
	url	location	deal_type
0	https://www.cian.ru/sale/flat/308080996/	PyPScP"PrP"	sale
1	https://www.cian.ru/sale/flat/307575540/	PyPScP"PrP"	sale
2	https://www.cian.ru/sale/flat/298845467/	PyPScP"PrP"	sale
3	https://www.cian.ru/sale/flat/293713880/	PyPScP"PrP"	sale
4	https://www.cian.ru/sale/flat/302177521/	PyPScP"PrP"	sale
...
10691	https://chekhov.cian.ru/sale/flat/30879862/	Чехов	sale
10692	https://chekhov.cian.ru/sale/flat/297269456/	Чехов	sale
10693	https://chekhov.cian.ru/sale/flat/205845891/	Чехов	sale
10694	https://chekhov.cian.ru/sale/flat/205818606/	Чехов	sale
10695	https://chekhov.cian.ru/sale/flat/307249555/	Чехов	sale
...
10694	Чехов	41	NaN
10695	Чехов	9	NaN

После удаление всех не нужных данных вывожу процентное содержание пропусков, их стало меньше.

```
for col in df.columns:
    pct_missing = np.mean(df[col].isnull())
    print('{} - {}'.format(col, round(pct_missing*100)))
```

author	- 1%
author_type	- 1%
url	- 0%
location	- 4%
deal_type	- 0%
accommodation_type	- 0%
floor	- 0%
floors_count	- 0%
rooms_count	- 0%
total_meters	- 0%
price	- 0%
year of construction	- 0%
object_type	- 0%
house_material_type	- 0%
heating_type	- 0%
finish_type	- 0%
living_meters	- 0%
kitchen_meters	- 0%
phone	- 0%
district	- 34%
street	- 14%
house_number	- 13%
underground	- 41%
residential_complex	- 52%
author.1	- 1%
...	...
street.1	- 14%
house_number.1	- 13%

Потом вывела корреляцию числовых значений. Больше всего влияют на цену квартиры параметры price и rooms_count.



Теперь можно идти визуализировать.

Я использовала в визуализации инструмент Power Bi. С его помощью я создала интерактивный дашборд для выявления факторов, которые больше влияют на цену квартиры.

В начале работы я закинула .csv файл в редактор Power Qwery

По моему мнению тип отделки квартиры очень важен, поэтому я заменила на среднее значение.

The screenshot shows the Power Qwery editor interface. The main table displays data for various properties, including columns for 'punt', 'total_meters', 'price', 'year_of_construction', 'finish_type', and 'kitchen_meters'. A dialog box titled 'Замена значений' (Replace values) is open, allowing the user to replace values in selected columns. The dialog box contains the following fields:

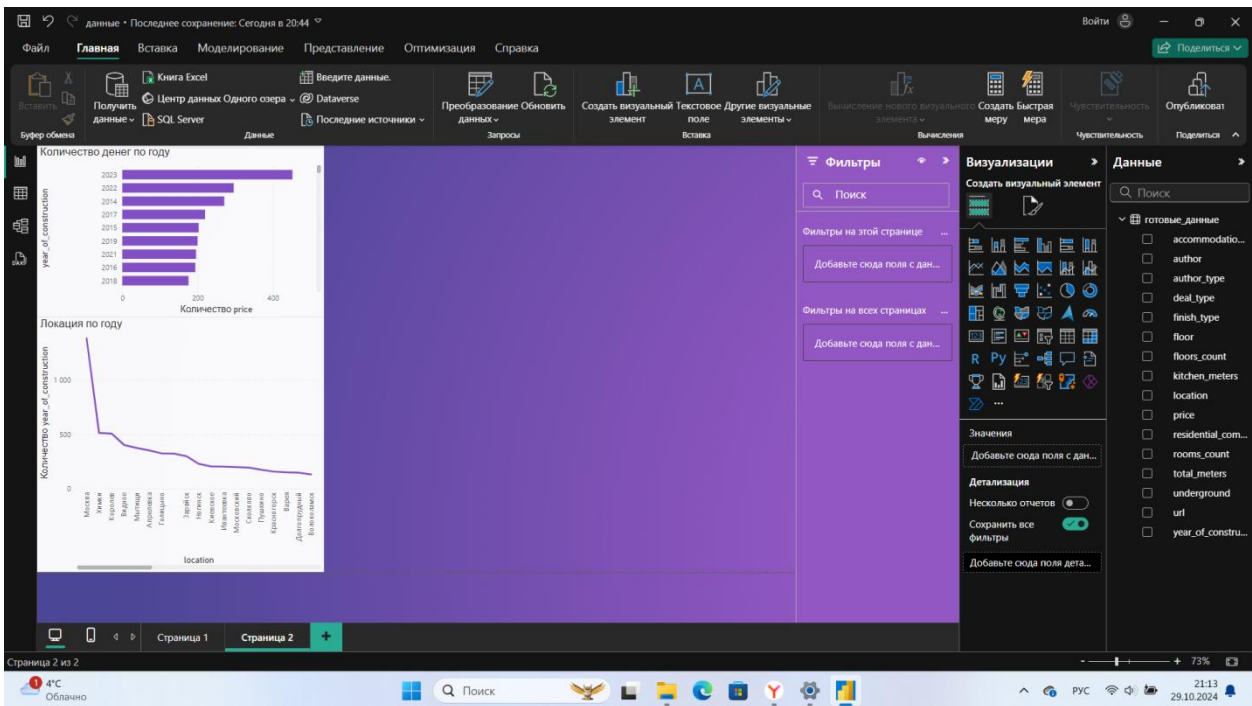
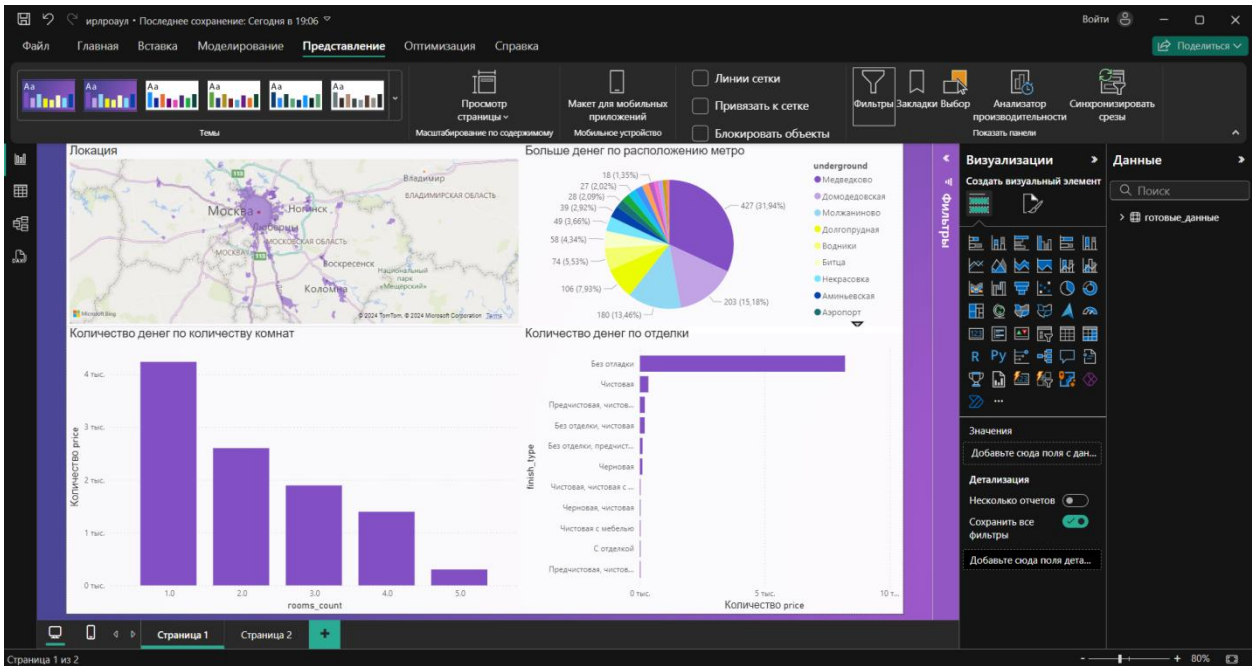
- Значение для поиска (Value to search for): -1
- Заменил на (Replace with): Без отделки
- Расширенные параметры (Advanced parameters):

The table data is as follows:

punt	total_meters	price	year_of_construction	finish_type	kitchen_meters
1	176.6	36000000.0	2006	-1	19.3 м²
2	98.8				13.8 м²
3	129.88				-1
4	244.1				-1
5	160.0				22 м²
6	125.08				24.2 м²
7	233.0				30 м²
8	167.0				20 м²
9	180.7				-1
10	157.0				-1
11	92.8				11 м²
12	126.55				-1
13	200.0				-1
14	105.0				-1
15	176.5				8 м²
16	200.0	110000000.0	2001	-1	70.7 м²
17	207.0	430000000.0	2012	-1	15 м²
18	153.6	97888600.0	2026	Без отделки, чистовая с мебелью	26.8 м²
19	168.3	249000000.0	2012	-1	20 м²
20	100.5	300000000.0	2019	-1	18 м²
21	102.9	44068057.0	2025	Без отделки	11.3 м²
22	145.0	299000000.0	2016	-1	20 м²
23	169.5	79570000.0	2000	-1	10 м²
24	112.75	86417000.0	2027	Без отделки	23 м²
25	93.0	79950000.0	-1	-1	20 м²
26	83.3	183181074.0	2026	Без отделки, чистовая с мебелью	13.3 м²

К сожалению из-за того что Power Vi иностранный инструмент для визуализации, price обозначается как текстовые тип данных. Но можно price можно сравнивать с другими данными только по количеству.

Вот такой дашборд получился.

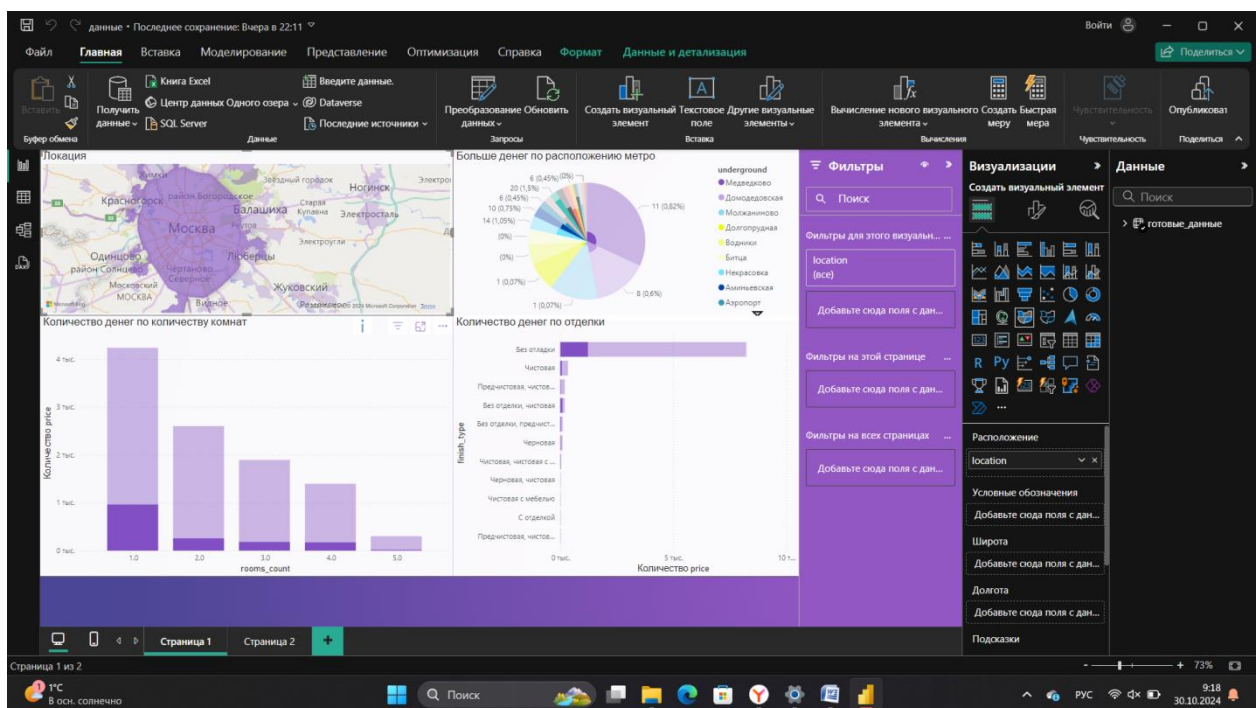


Анализ данных

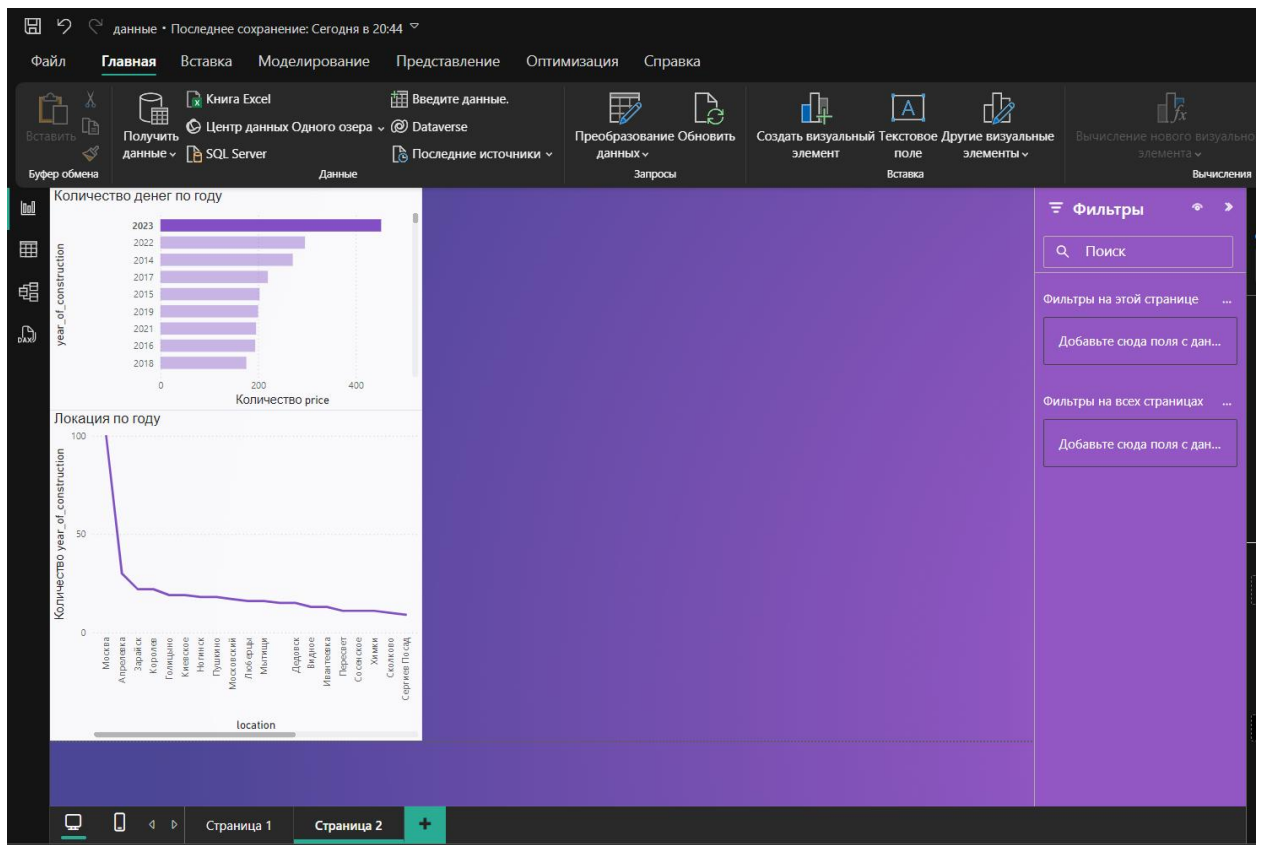
После полученного дашборда можно выделить важные критерии по цене квадратного метра квартиры.

Критерии:

1. Метро
2. Отделка
3. Нахождение в крупных городах
4. Количество комнат



Больше всего цена в центре Москвы, так как она соответствует всем важным критериям. Есть рядом метро, большинство квартир уже с отделкой и больше всего однокомнатных.



Так же квартиры в 2023 году выросли в цене и видно сколько объявление было сделано с информацией годом постройки.

Заключение

В результате анализа были собраны данные, построена модель и визуализация, в анализе которой выявлены важные критерии в оценивании цены квадратного метра недвижимости в Московском регионе. Основными факторами, оказывающими влияние на стоимость жилья, оказались расположение около метро, нахождение в крупных городах, наличие отладки в квартире. Полученные результаты могут быть использованы для дальнейшего прогнозирования цен на недвижимость.