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

Отчёт по кейсу «Анализ данных»

Работу выполнил:
Студент группы № ИСП-21
Благодарный Арсений

Долгопрудный, 2024

Введение

В данном отчёте рассматриваются выводы, полученные после анализа данных в области «Квартиры в Москве и Московской области».

Цель

Собрать данные и проанализировать их для будущего использования, например, обучение машины на основе выводов.

Задачи

1. Собрать данные с помощью данных инструментов.
2. работу над ними точнее удаление ненужных данных, дополнение необходимых и т.д.
3. Визуализация данных. Нахождение взаимосвязи между данными или её полное отсутствие для отчёта.

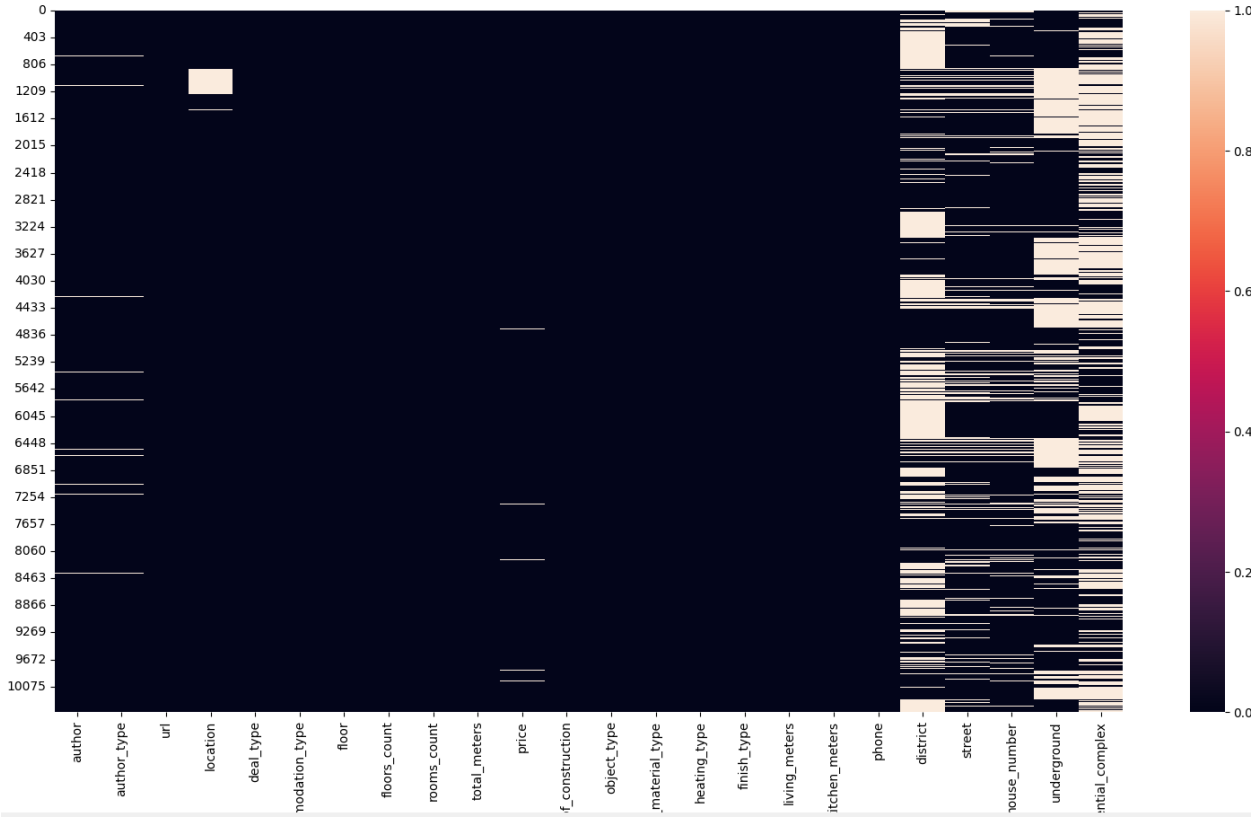
Основная часть

С помощью языка Python и библиотеки “CianParser” мы собрали чуть больше 10 тысяч данных по Москве и Московской Области.

Мы собрали данные, объединяем её и убираем дубликаты. Смотрим какого типа наши данные

```
RangeIndex: 10460 entries, 0 to 10459
Data columns (total 24 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   author                10289 non-null  object 
 1   author_type           10286 non-null  object 
 2   url                   10456 non-null  object 
 3   location              10057 non-null  object 
 4   deal_type             10459 non-null  object 
 5   accommodation_type    10459 non-null  object 
 6   floor                 10459 non-null  float64
 7   floors_count          10459 non-null  float64
 8   rooms_count           10459 non-null  float64
 9   total_meters          10459 non-null  float64
 10  price                 10426 non-null  float64
 11  year_of_construction  10459 non-null  float64
 12  object_type           10456 non-null  float64
 13  house_material_type   10456 non-null  object 
 14  heating_type          10456 non-null  float64
 15  finish_type           10456 non-null  object 
 16  living_meters         10458 non-null  object 
 17  kitchen_meters        10458 non-null  object 
 18  phone                 10456 non-null  float64
 19  district              5900 non-null   object 
 20  street                8981 non-null   object 
 21  house_number          9366 non-null   object 
 22  underground           6634 non-null   object 
 23  residential_complex   4733 non-null   object 
dtypes: float64(9), object(15)
```

Теперь мы визуализируем наши данные при помощи библиотеки seaborn и функции heatmap, и смотрим, какие данные мы не смогли собрать



Проанализировав нашу таблицу, мы видим, что у нас есть таблицы, которые будут мешать нашей работе. Как пример таблица residential_complex, которая имет слишком много пропусков, и нам следует избавиться от нее.

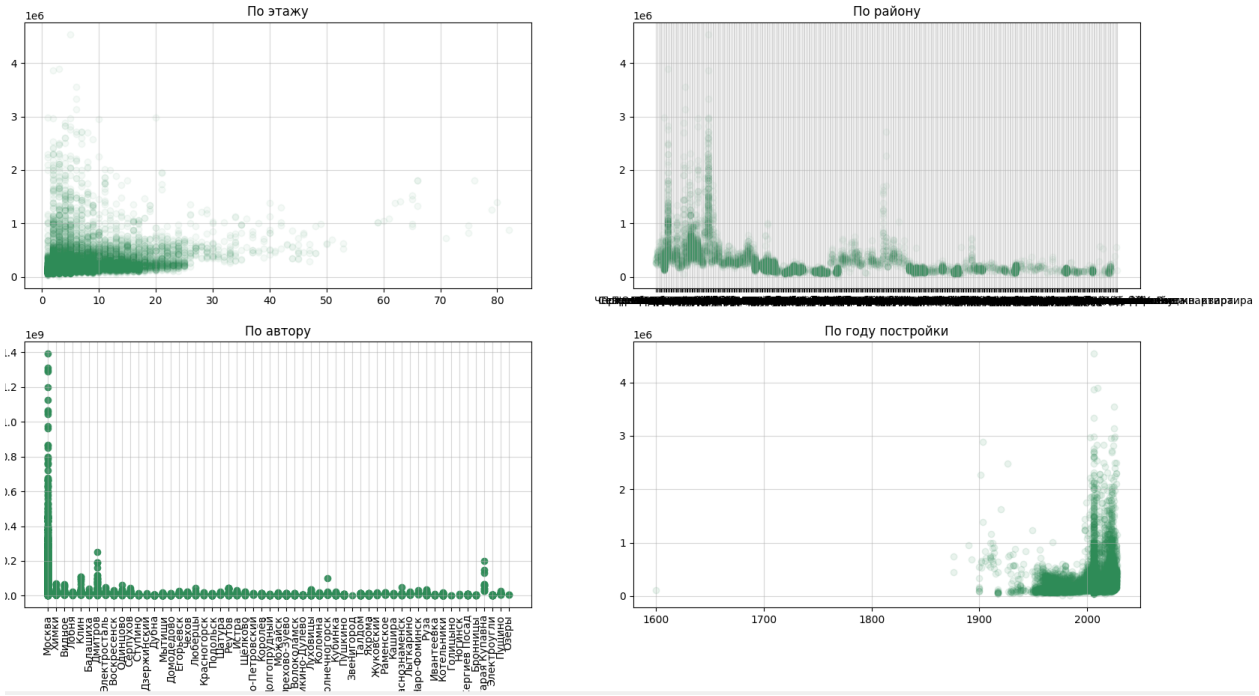
Далее фильтруем ненужные данные и форматируем столбцы, чтобы их было можно анализировать

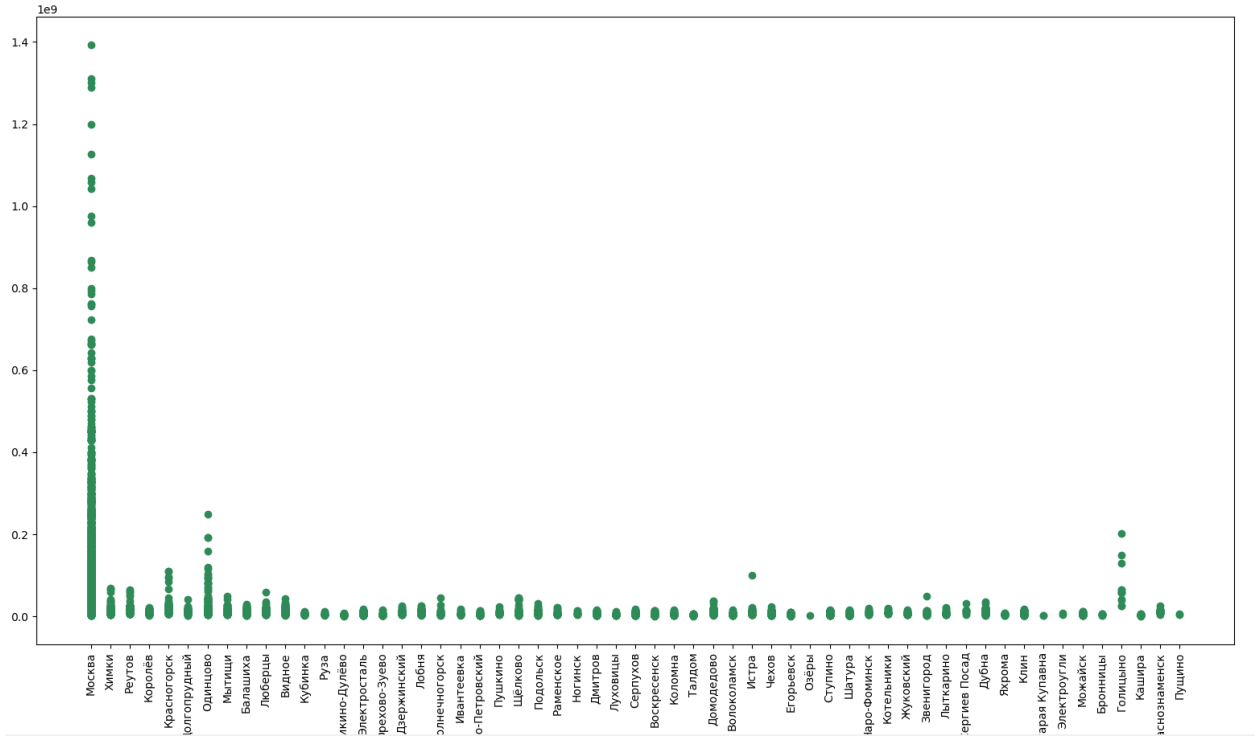
```
# Убираем колонки, не особо влияющие на цену, а так же дубликаты
useless_columns=["author","author_type","url","deal_type", "accommodation_type","object_type","house_material_type", "finish_type","phone",
"heating_type","street","house_number", "residential_complex"]
df=df.drop(columns=useless_columns).drop_duplicates()


# Заполняем/Удаляем пропуски
df.dropna(subset=["location","price"],inplace=True)
df.loc[df["district"].isna(),"district"]=df["location"]
df.loc[df["underground"].isna(),"underground"]=df["location"]
df.loc[df["living_meters"]=="-1","living_meters"]=df["total_meters"]
df.loc[df["living_meters"].isna(),"living_meters"]=df["total_meters"]
df.loc[df["kitchen_meters"]=="-1","kitchen_meters"]=0
df.loc[df["kitchen_meters"].isna(),"kitchen_meters"]=0
df.replace(["-1",-1,"-1.0",-1.0],df['year_of_construction'].median(),inplace=True)
df=df[df['rooms_count']!=2006.0]
```

После завершения процесса очистки данных можно сохранить очищенную базу, а затем перейти к созданию графиков и проведению аналитической работы с использованием библиотеки matplotlib для визуализации информации. В частности, планируется построить пять графиков:

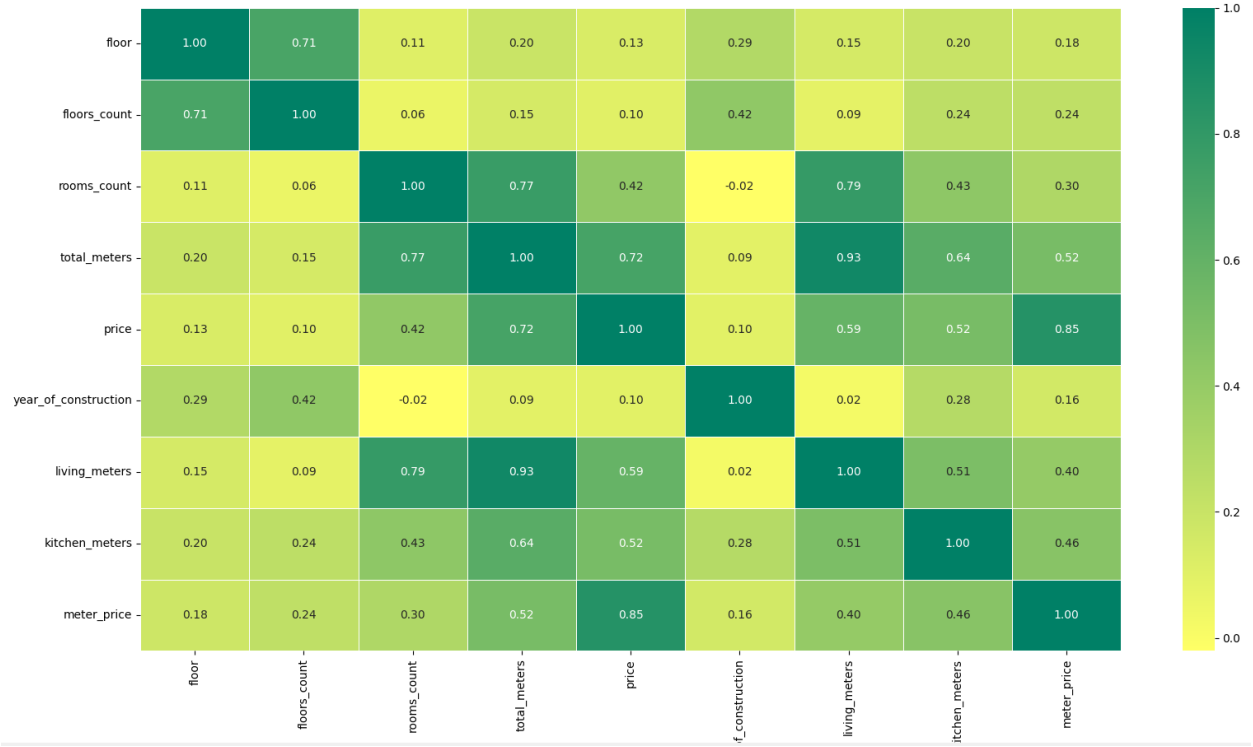
1. Цена за м² в зависимости от этажа, на котором расположена квартира.
2. Цена за м² по различным районам.
3. Цена за м² для всего города.

4. Цена за м² в зависимости от года постройки квартиры.
5. Количество объявлений в зависимости от числа комнат.





Так же мы выведем матрицу корреляций, которая очень хоро поможет нам понять зависимость значений друг от друга.



В итоге всех наших данных мы можем делатьв ыводы и анализировать дальнейшие задачи

Аналитика данных

Анализируя графики, можно заключить, что цена в основном определяется количеством комнат и площадью квартиры. Наименьшее влияние на цену оказывают год постройки и количество этажей.

Заключение

В ходе работы было собрано около одиннадцати тысяч квартир, данные были отсортированы и очищены, в результате чего осталось почти десять тысяч объектов. Были построены графики, которые упростили анализ и помогли выявить ключевые критерии для оценки стоимости недвижимости в Москве и Московской области. Основными факторами, влияющими на цену, оказались площадь и количество комнат. Мы также освоили работу с различными библиотеками и их функциями.