Аналитический отчет   
Зотеев А. А.  
  
  
1. Введение

Анализ рынка недвижимости представляет собой важный аспект в современном обществе, поскольку он влияет на множество факторов, включая экономическое развитие, инвестиционные решения и личные финансовые стратегии. В данном отчете рассматривается анализ данных о квартирах, с целью выявления ключевых факторов, влияющих на стоимость жилья, а также для понимания динамики рынка.

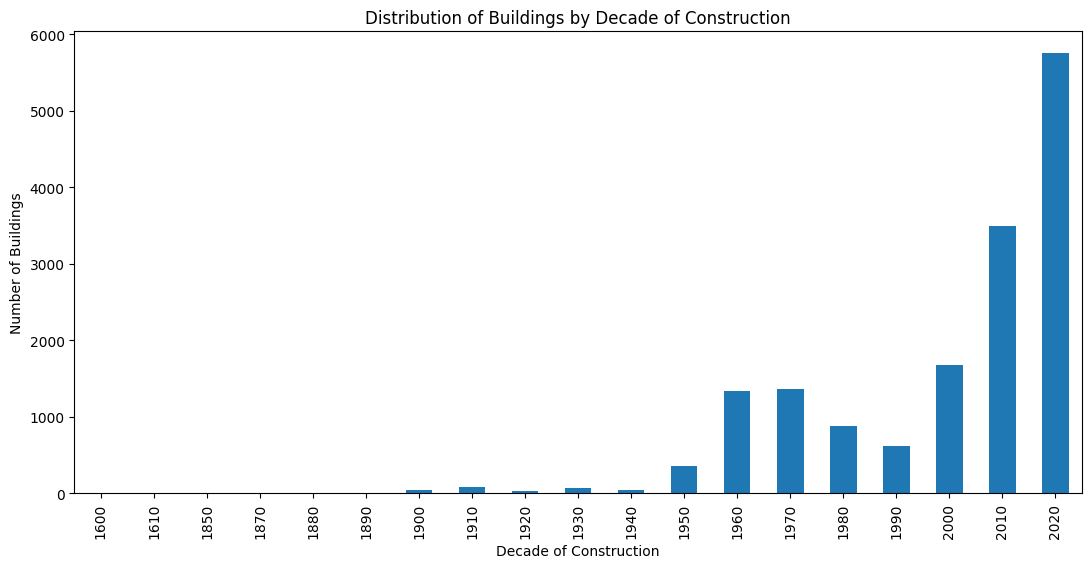
Данные, использованные в этом анализе, включают информацию о характеристиках квартир, таких как площадь, количество комнат, расположение и возраст зданий. Кроме того, учитываются такие параметры, как средняя цена за квадратный метр.

Цель данного отчета заключается в том, чтобы предоставить аналитическую информацию, которая поможет потенциальным покупателям, инвесторам и другим заинтересованным сторонам лучше понять текущую ситуацию на рынке квартир. В процессе анализа будут использованы различные методы, включая статистический анализ, визуализацию данных и машинное обучение, чтобы выявить закономерности и сделать обоснованные прогнозы.

2. Описание данных

Используя библиотеку cianparser, и объединившись со своими одногруппниками, я собрал набор данных состоящий из +-20тыс. данных. В нем были различные колонки, в частности:

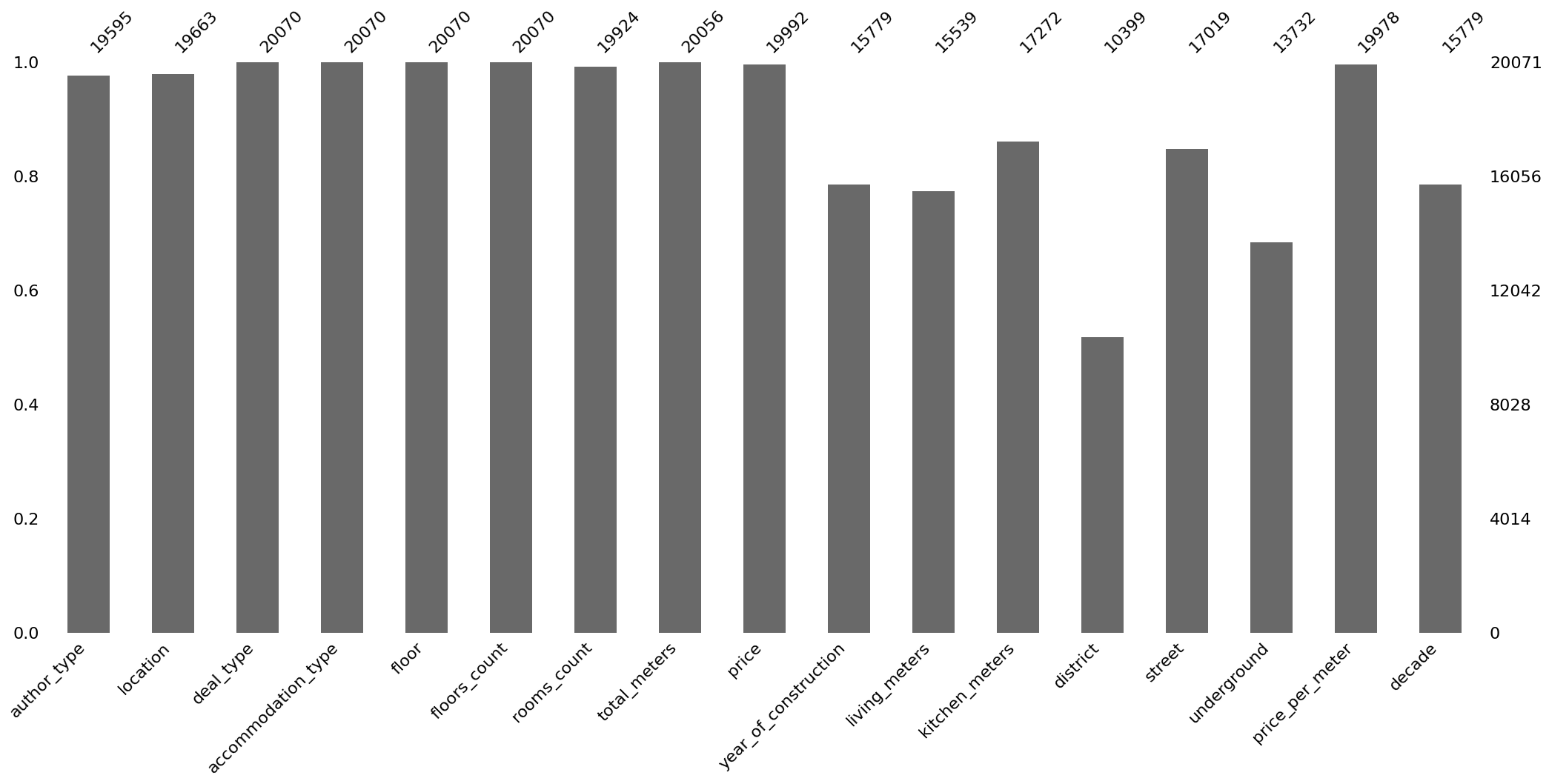
* Автор объявления
* Тип автора
* Ссылка на само объявление
* Город
* Тип сделки
* Тип строения
* Этаж
* Кол-во этажей в доме
* Кол-во комнат
* Метраж
* Цена
* Год постройки
* Тип объекта
* Тип материала
* Завершенность
* Жилой метраж
* Метраж кухни
* Телефон
* Район
* Улица
* Номер дома
* Метро
* Жилой комплекс



(График распределения кол-ва квартир по декадам)

До того, как посмотреть пропуски и начать их заполнять, я провел небольшую предобработку данных, а именно:

* Все значения -1, я заменил на NaN, тк в данном случае -1 означает отсутствие данных
* Везде, где значение location было неизвестно, и где ссылка имела формат <https://”city_name””.cian/..../>..., я достал город из ссылки, переделал с транслита на русский язык, и вставил полученное значение в location
* Также, все значение которые имели тип object, а должны были быть числовые я поменял на нужный тип данных (например колонки kitchen\_meters, floor, rooms\_count и так далее)



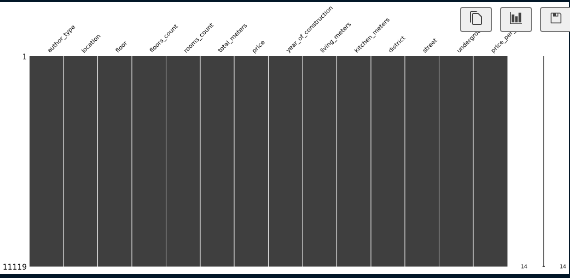
(График, показывающий пустые значения)

Для дальнейшей работы с датасетом, я убрал те колонки, которые посчитал необязательными - автор, номер дома, тип сделки, тип строения, ссылку и номер телефона, так как их значения никак не влияют на предсказываемую цену за квадратный метр, тип объекта и тип отопления из за того что там было 100% пропущенных значений, и тип материала, завершенность, жилой комплекс - из за того, что они не прошли трешхолд по пропущенным значениям - 50%

1. Пропуски в комнатах я заполнил медианой
2. Пропуски в типе автора я заполнил модой
3. Пропуски в локации я заполнил модой
4. Пропуски в этажах я заполнил модой, а пропуск в количестве этажей в доме я удалил, так как он был один
5. Чтобы заполнить пропуски цены, я нашел верхние и ниэние квартили цены, отфильтровал по ним датасет, и там где пропуски вставил среднее по городам



1. Пропуски в метраже квартиры я удалил, а также создал новую колонку - цена за квадратный метр, в которой вывел значение цены поделенной на квадратный метр
2. Пропуски в жилом метраже, а также в метраже кухни, я заполнил используя колонки без пропусков, и модель линейной регрессии
3. Пропуски в районе и метро, я заполнил используя значения из колонки location
4. Пропуски в колонке с улицей, я удалил.

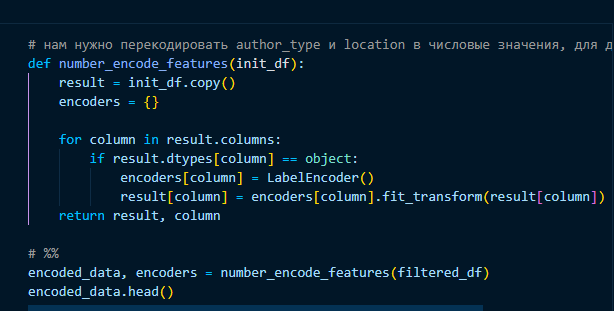


1. В конце, я отфильтровал датафрейм так, чтобы не было значений где жилой метраж и метраж кухни меньше 14 (если однушка, 16 если больше чем 1 комната) и 5 квадратных метров соответственно

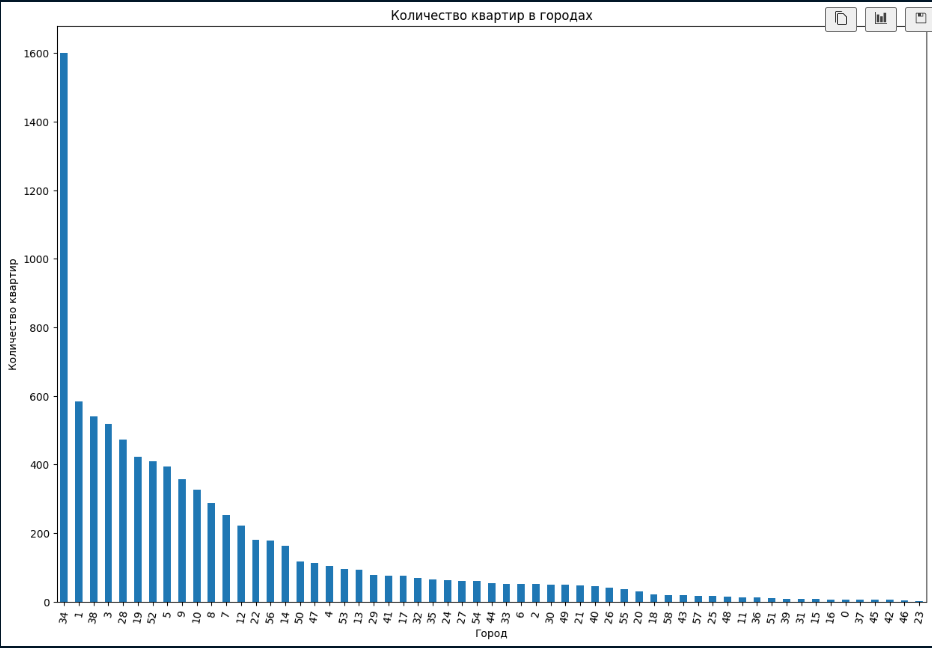
Можно заметить, что на данном этапе мы получили чистый датасет, который содержит в себе 10+- тысяч данных

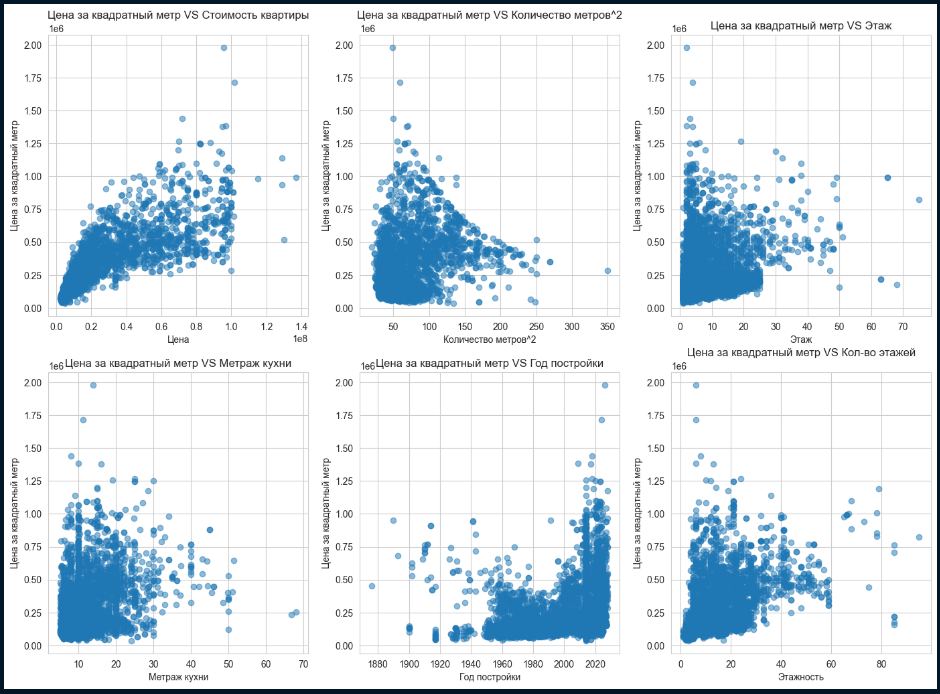
3. Анализ данных

Для того чтобы начать визуализировать и анализировать данные, нужно перекодировать все категориальные значения в числовые

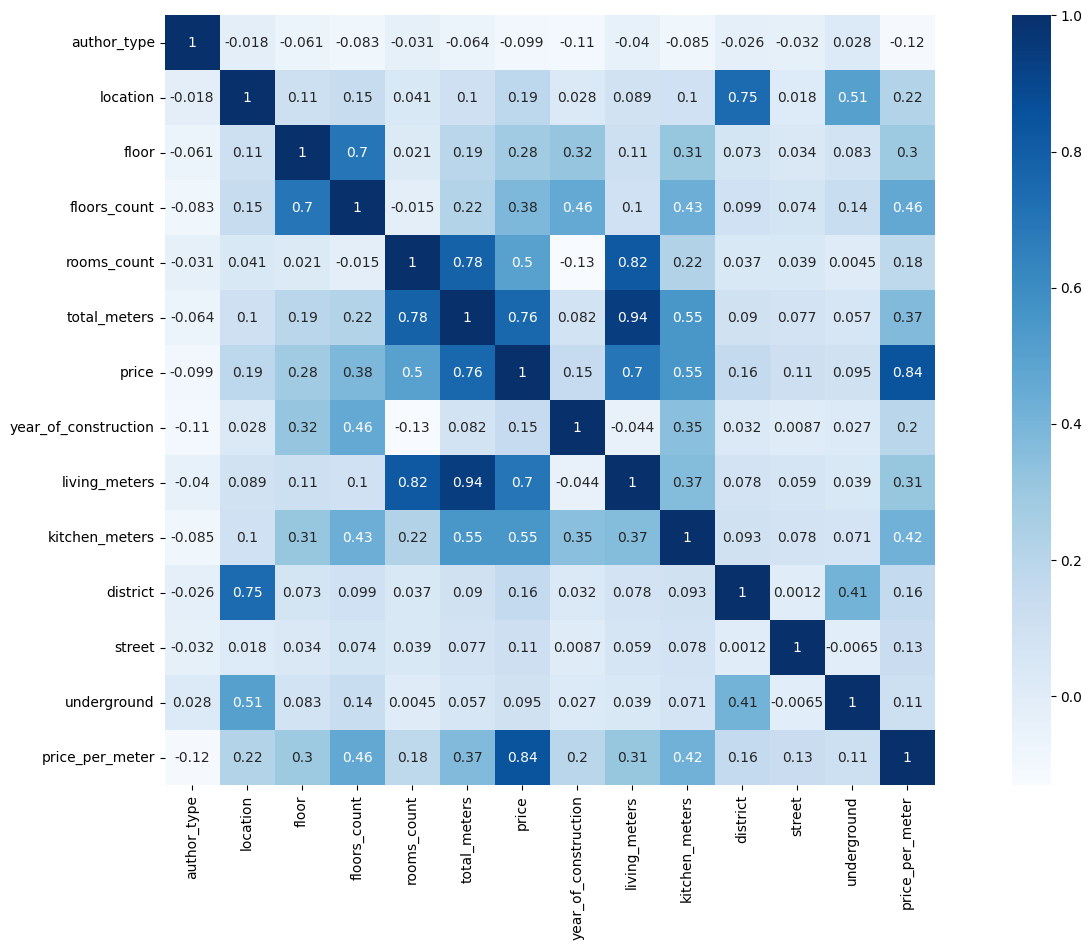


Выведем график распределения кол-ва квартир по городам - конкретно в этом случае я обрезал данные с Москвы до 1600, а также уже закодировал данные, так что вместо названий городов мы видим только их код





Но есть и хорошая новость - тк мы закодировали данные, теперь у нас есть возможность отобразить тепловую карту корреляций:



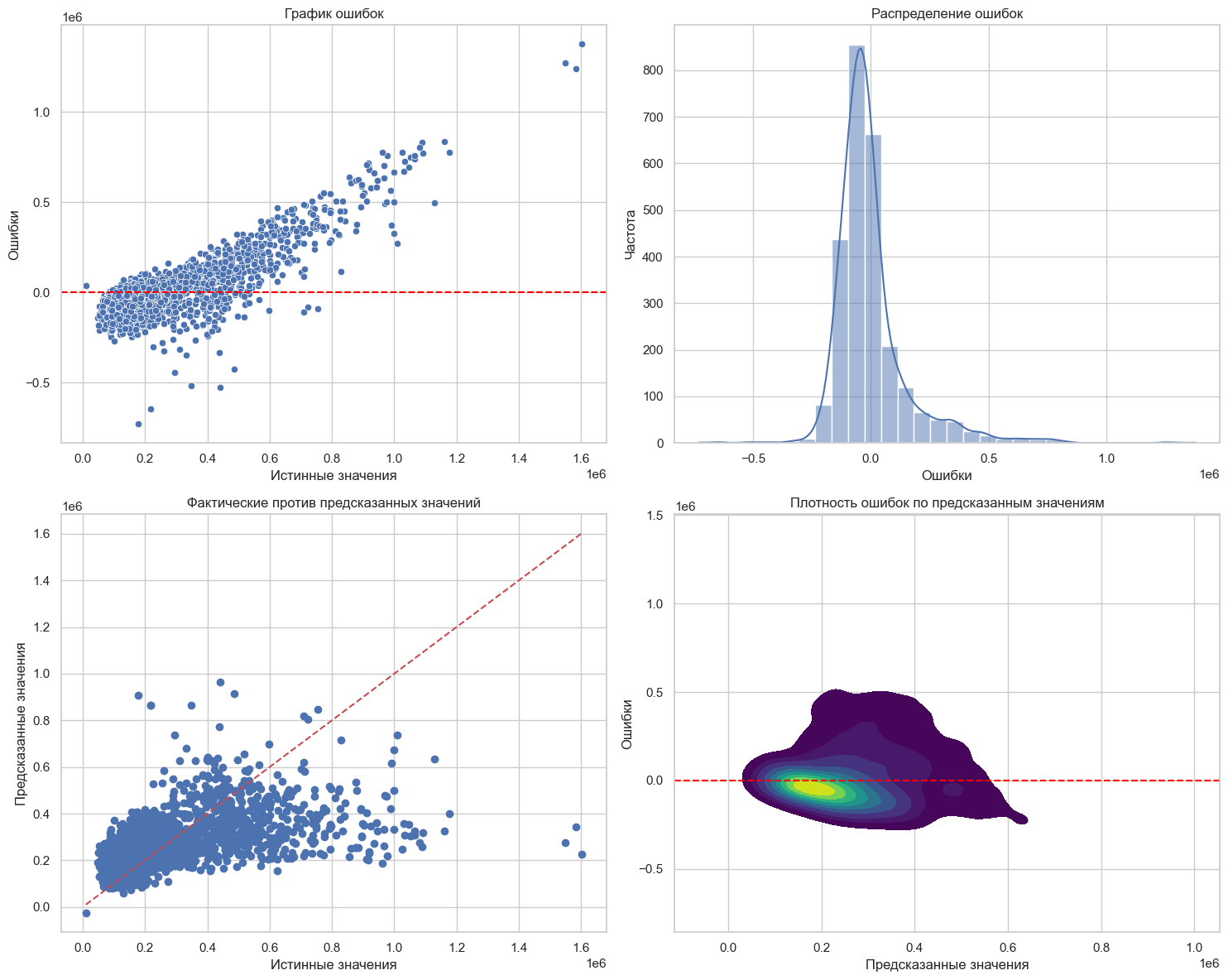
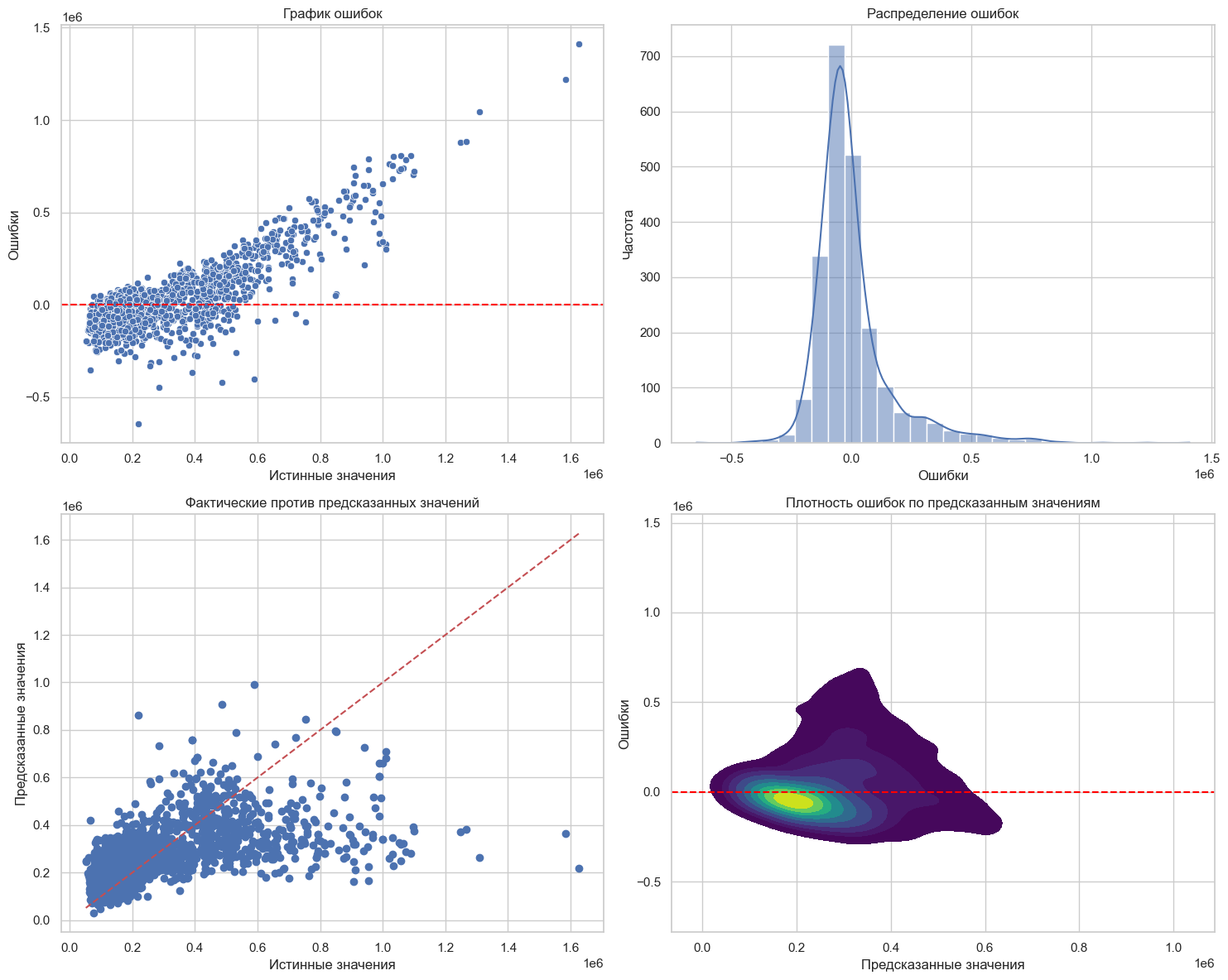
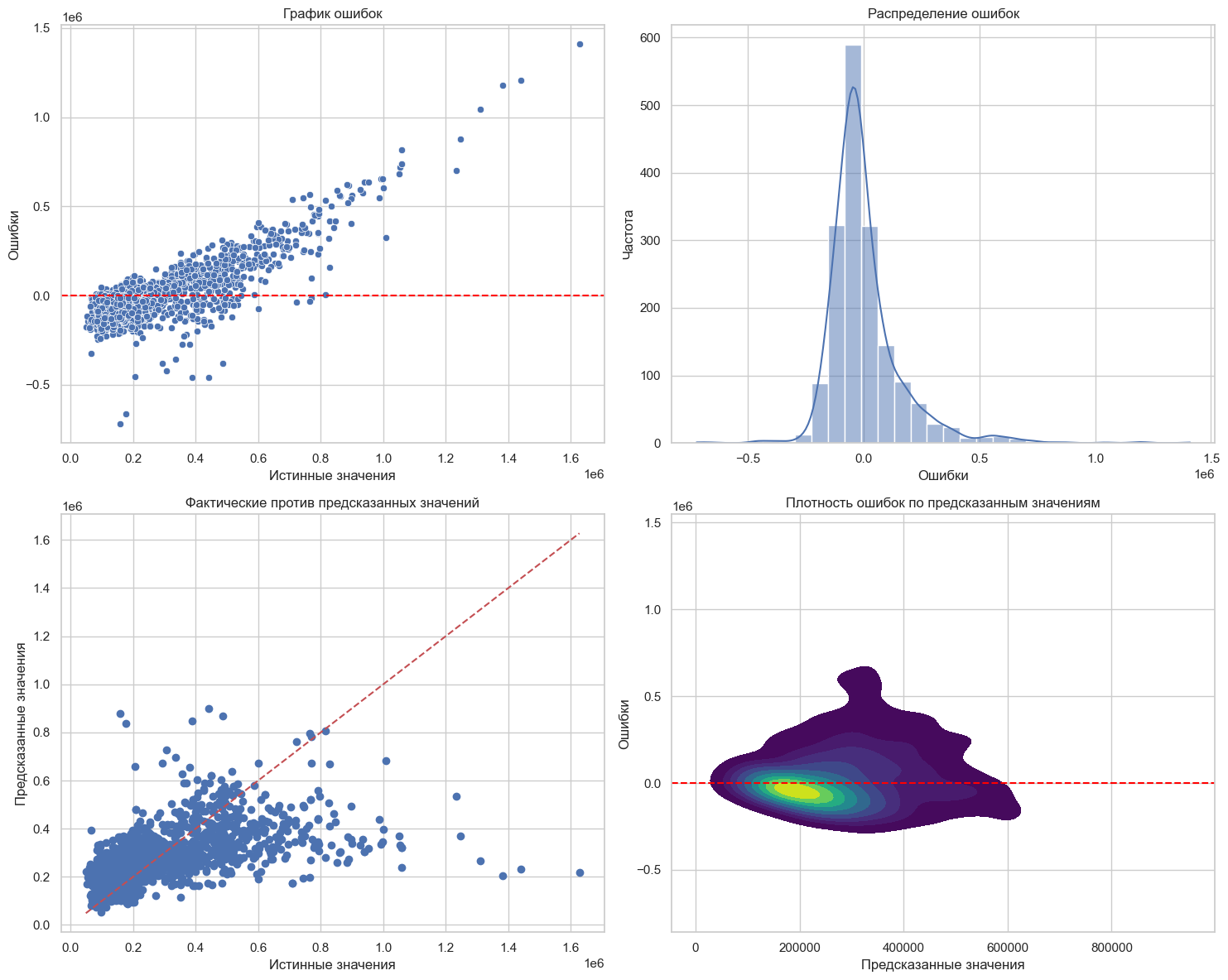
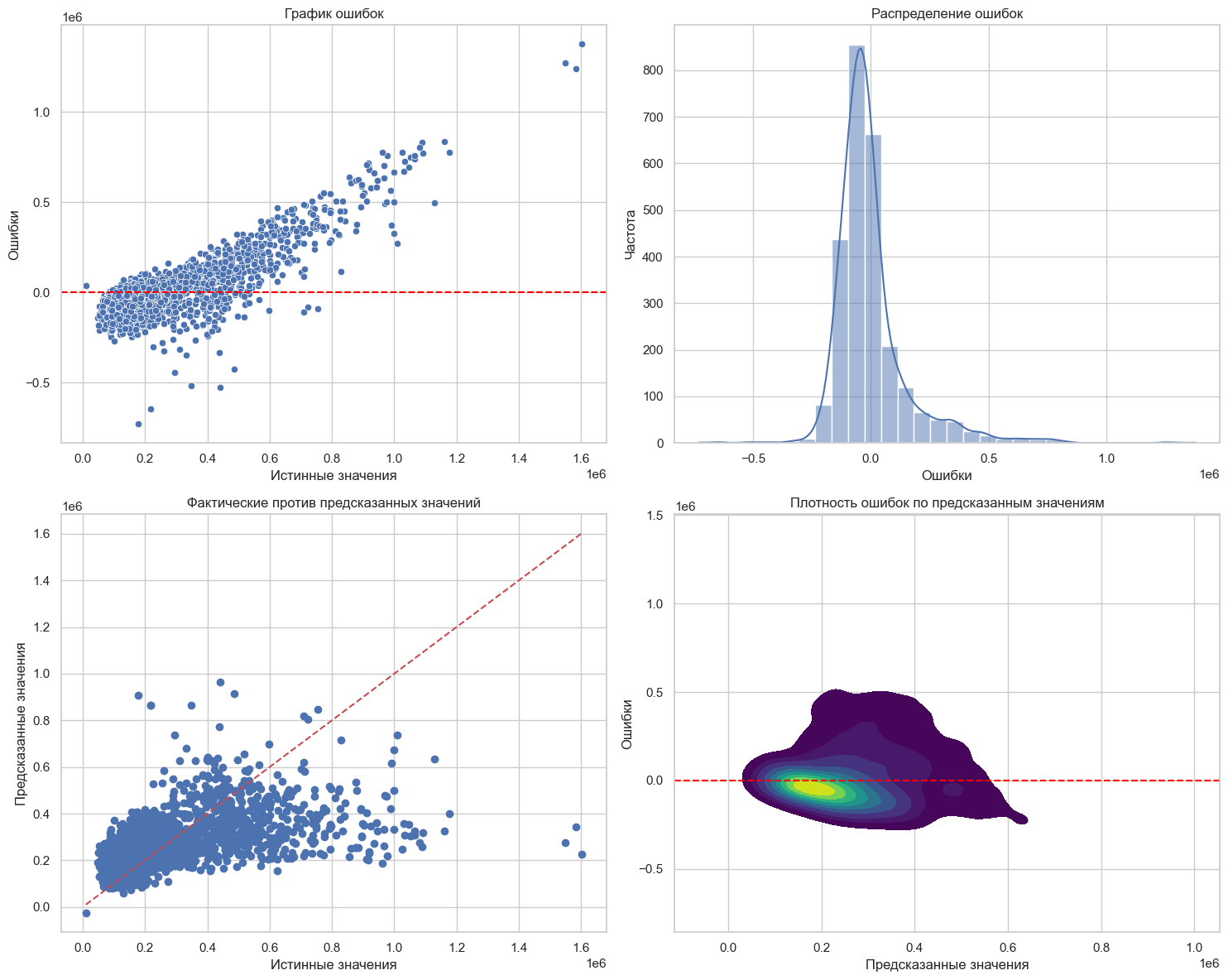
Можно заметить, что на нашу целевую переменную - цена за квадратный метр довольно сильно влияют такие переменные как : жилой и кухонный метражи, сам метраж, количество этажей в доме, сам этаж собственно и локация (город)

4. Моделирование

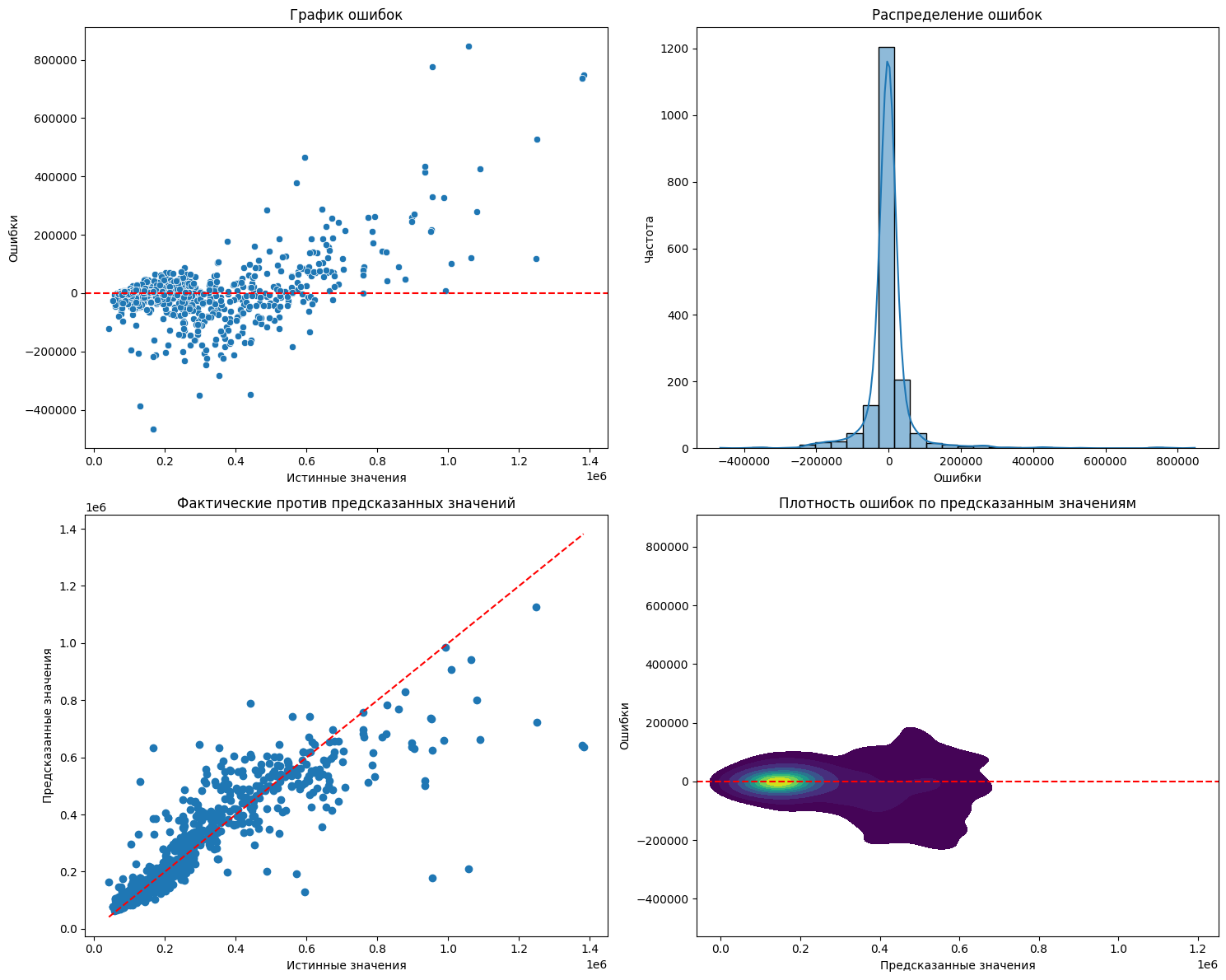
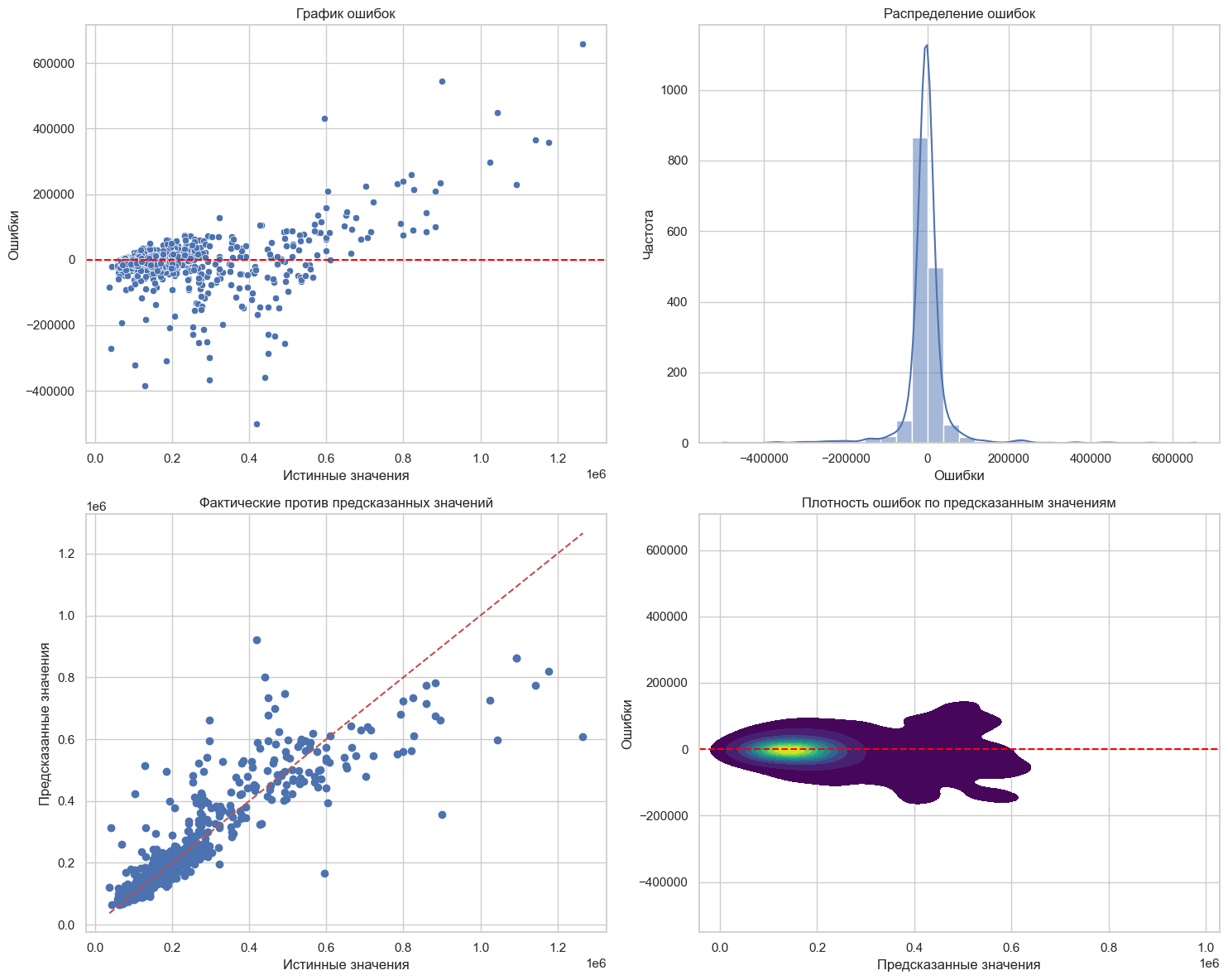
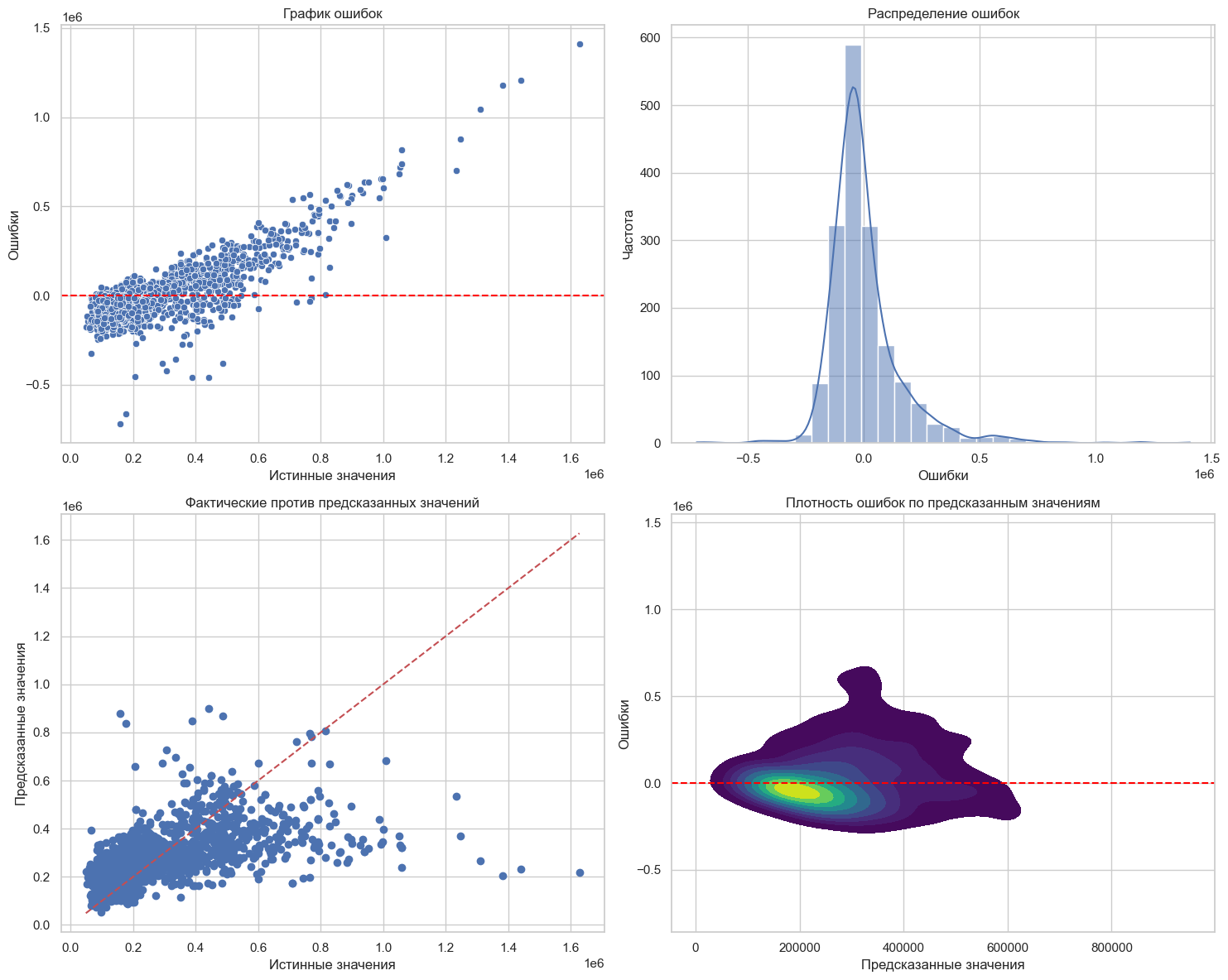
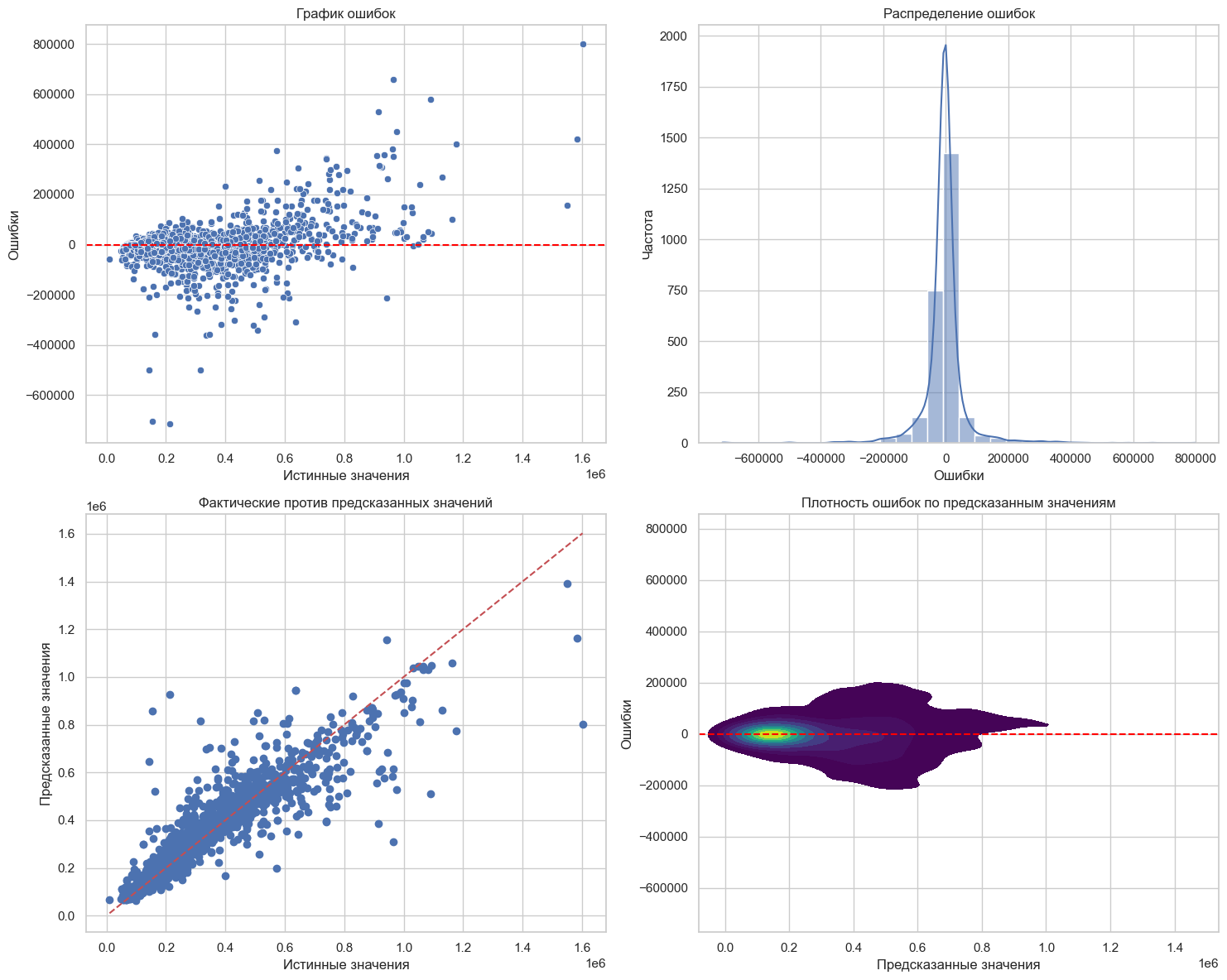
В данном случае мы решаем задачу регрессии, поэтому я буду использовать модели linear regression и random forest regression (также хотел попробовать catbost, но возникли проблемы на техническом уровне)  
  
Я создал отдельный ml\_df, который буду использовать при обучении и убрал из него колонку с ценой

Потом назначил Х, у, а также разделил датасет на тестовый и тренировочный

Основные метрики по которым я буду смотреть валидность модели - MAE и R^2  
  
Модель LinearRegression() показывала неутешительные результаты, и спустя 4 прогона с разными значениями, она в среднем показывала результат в 2-2.5 раза хуже чем модель RandomForestRegression()



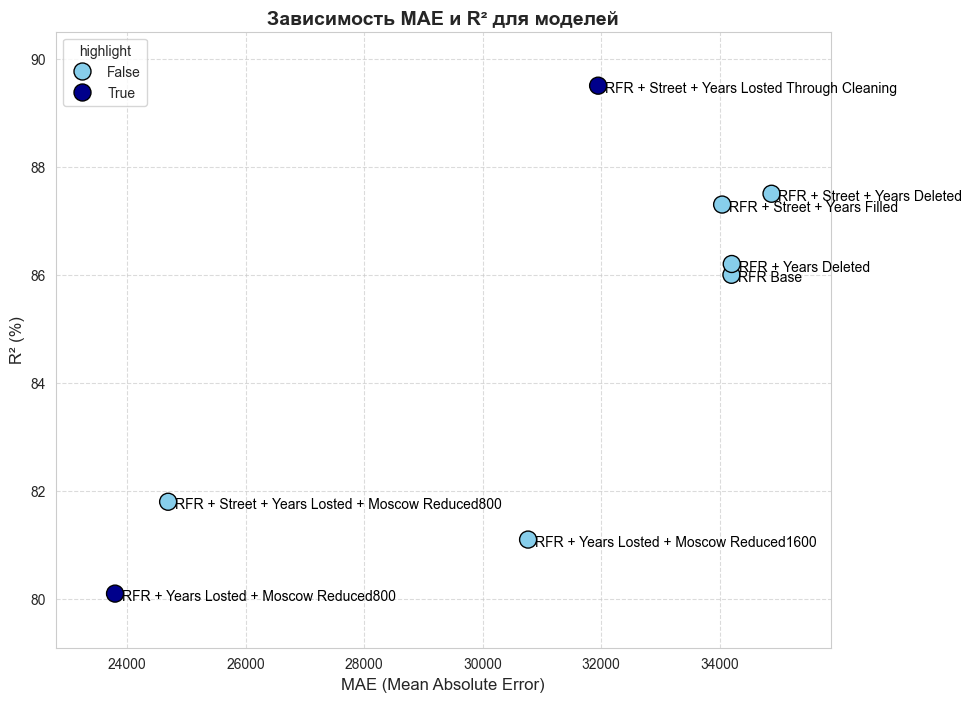
Это графики по линейной регрессии



Это графики по случайным регрессивным лесам

Можно заметить, что у RFR данные лучше разбросаны по графику, и имеет лучшую плотность

Поэкспериментировав со значениями и используемыми колонками, я получил следующий график:



Он собственно показывает, лучшую модель по метрике MAE и лучшую по метрике R^2. Если нам нужна точность - надо брать лучшую по MAE, если важнее объяснительная сила модели - стоит посмотреть на лучшую по метрике R^2

5. Выводы

Проведя сбор, чистку и анализ данных, мы выявили несколько основных параметров, от которых зависит значения цены за квадратный метр, и построили модель с хорошими показателями в метриках

Задания на будущее:

* Провести анализ отдельно по Москве и МО
* Провести анализ не только квартир, но и например домов, частных участков и так далее
* Создать удобный интерфейс для работы с моделями