Государственное бюджетное профессиональное

образовательное учреждение Московской области

«Физико-технический колледж»

**Отчёт по кейсу «Самолёт»:**

Работу выполнил:

Студент группы № ИСП-22

Доронин Андрей

**Введение:**

В данном отчёте представлены результаты аналитической работы с

данными по квартирам в Московской области, Новой Москве и

Москве.

**Цель:**

Целью работы является сбор и анализ данных для последующего

использования, например, в разработке модели прогнозирования.

**Задачи:**

Для достижения поставленной цели были определены следующие

задачи:

1. **Сбор данных**: Используя открытые источники, собрать

исчерпывающий набор данных о квартирах.

1. **Очистка и подготовка данных**: Провести предварительную

обработку данных, включая удаление ненужной информации,

дополнение недостающих данных, выявление и устранение

аномалий.

1. **Визуализация и анализ**: представить данные в наглядном

виде с помощью как минимум двух инструментов

визуализации. Выявить взаимосвязи между данными,

определить средние показатели и сделать выводы,

необходимые для составления отчёта.

**Основная часть:**

Для сбора данных использовалась библиотека Cianparser для

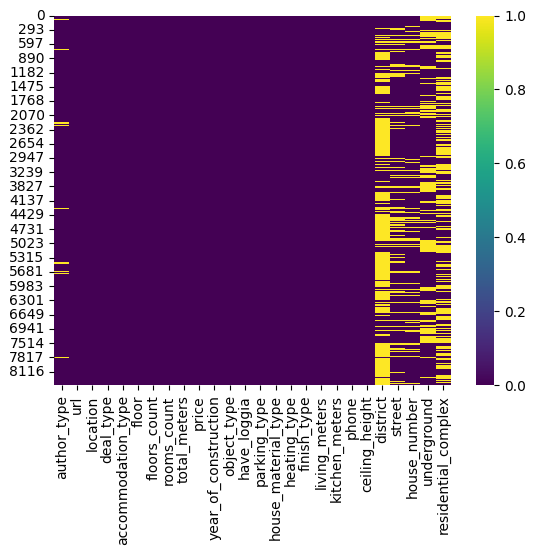
Python.С ее помощью была собрана информация о 20 тысячах

предложений о продаже квартир. Почти вся остальная работа подготовки данных к анализу была произведена с помощью библиотеки Pandas.

В полученном дата-сете было 16 видов данных:

1. **author** – имя автора, сразу удалил за ненадобностью
2. **author\_type** – тип автора (собственник,агент)
3. **url** – ссылка на обьявление
4. **location** - город
5. **deal\_type** – тип сделки(везде продажа
6. **accommodation\_type** – тип помещения(везде квартира)
7. **floor** - этаж
8. **floors\_count** – кол-во этажей в здании
9. **rooms\_count** – кол-во комнат
10. **total\_meters** - общая площадь
11. **price** - цена
12. **district** - район
13. **street** - улица
14. **house\_number** – номер дома(удален)
15. **underground** - ближайшая остановка метро/мцд
16. **residential\_complex** – жилой комплекс

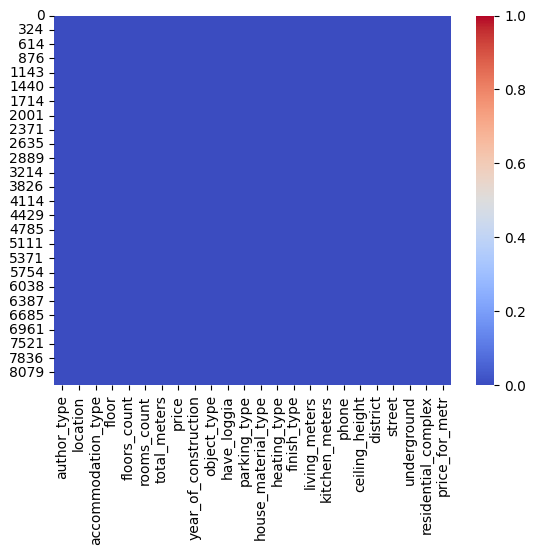
Затем я занялся удалением и заполнением пропусков.



Для этого был построен график, который показал, что больше всего пропусков содержится в колонках «district» и «residential\_complex». Чаще всего пропуски встречаются в строках, относящихся к квартирам в области. Чтобы заполнить эти пробелы, я использовал следующие подходы: значения для колонки «district» я взял из колонки «location», пропуски в колонке «underground» заполнил данными из «district», а для заполнения пустых значений в колонке «residential\_complex» использовал информацию из «underground».

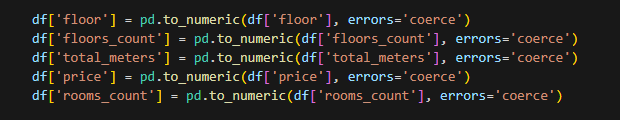
Так же просто удалил строки с любыми другими пропусками, так как их

число незначительно. После этого график стал выглядеть так:

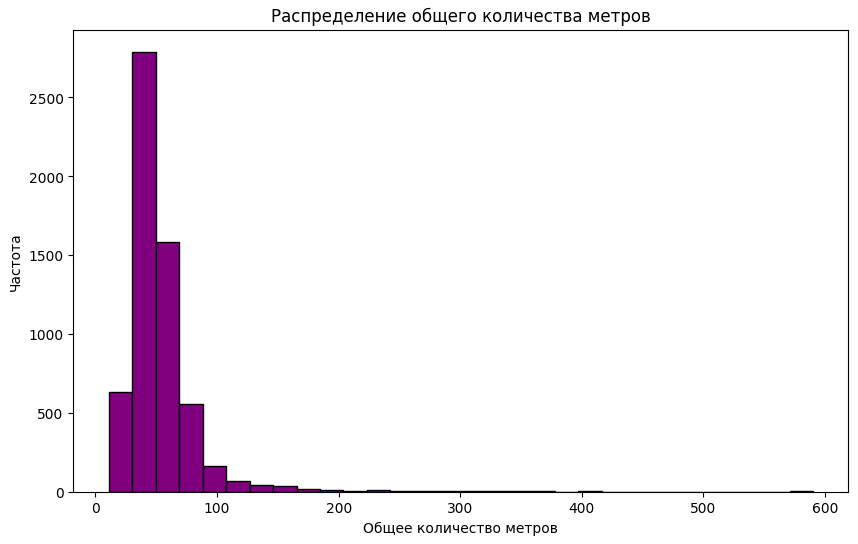


Данный график наглядно показывает, что пропусков больше нет.

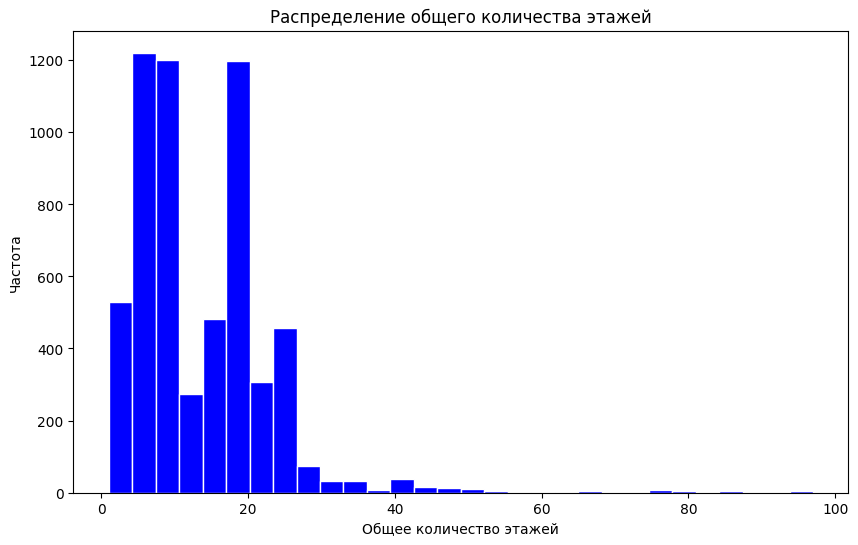
Дальше я занялся приведением типов данных, потому что данный всех колонках были типа “object”. Привел к числовым колонки которые должны быть числовыми:



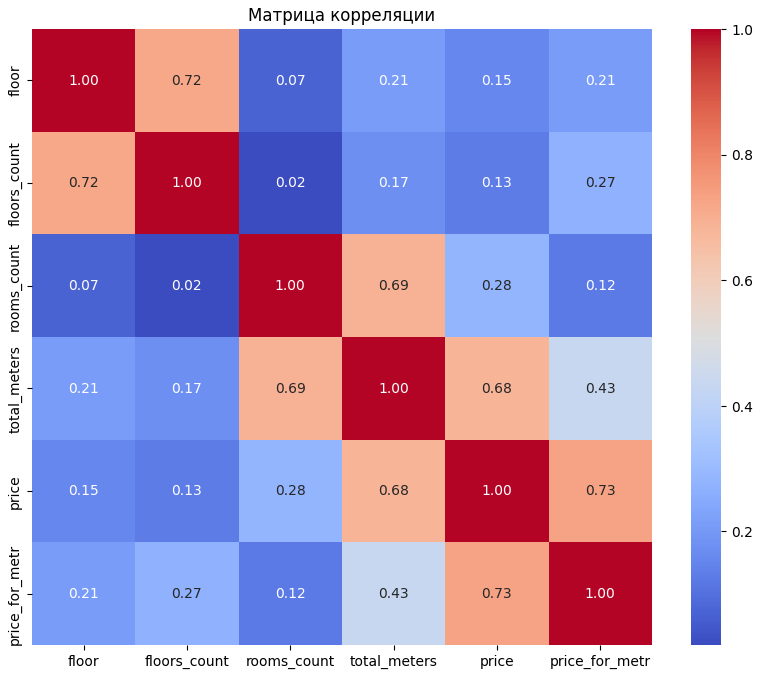
На основании полученных числовых значений я провёл анализ:



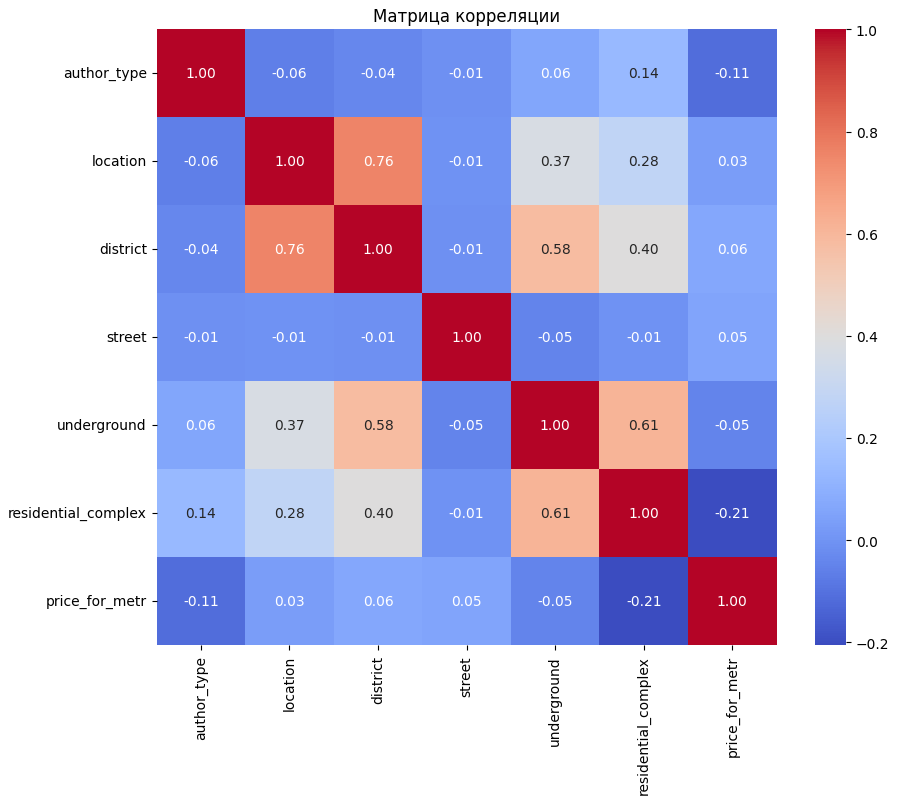
Чаще всего встречаются квартиры площадью 40-50 квадратных метров



Чаще всего встречаются 5-ти и 20-ти этажные здания.



Построив матрицу корреляции, мы видим, что целевая переменная значительно зависит от стоимости и площади квартиры. На стоимость влияют площадь и число комнат. Этажность связана с количеством этажей в доме. Площадь имеет корреляцию с числом комнат и стоимостью.



После преобразования категориальных данных в числа мы видим, что расположение оказывает очень слабое влияние на целевую переменную.

**Заключение:** анализ данных был проведён успешно. Цель проекта была выполнена.