# EDA: разведочный анализ данных по проекту

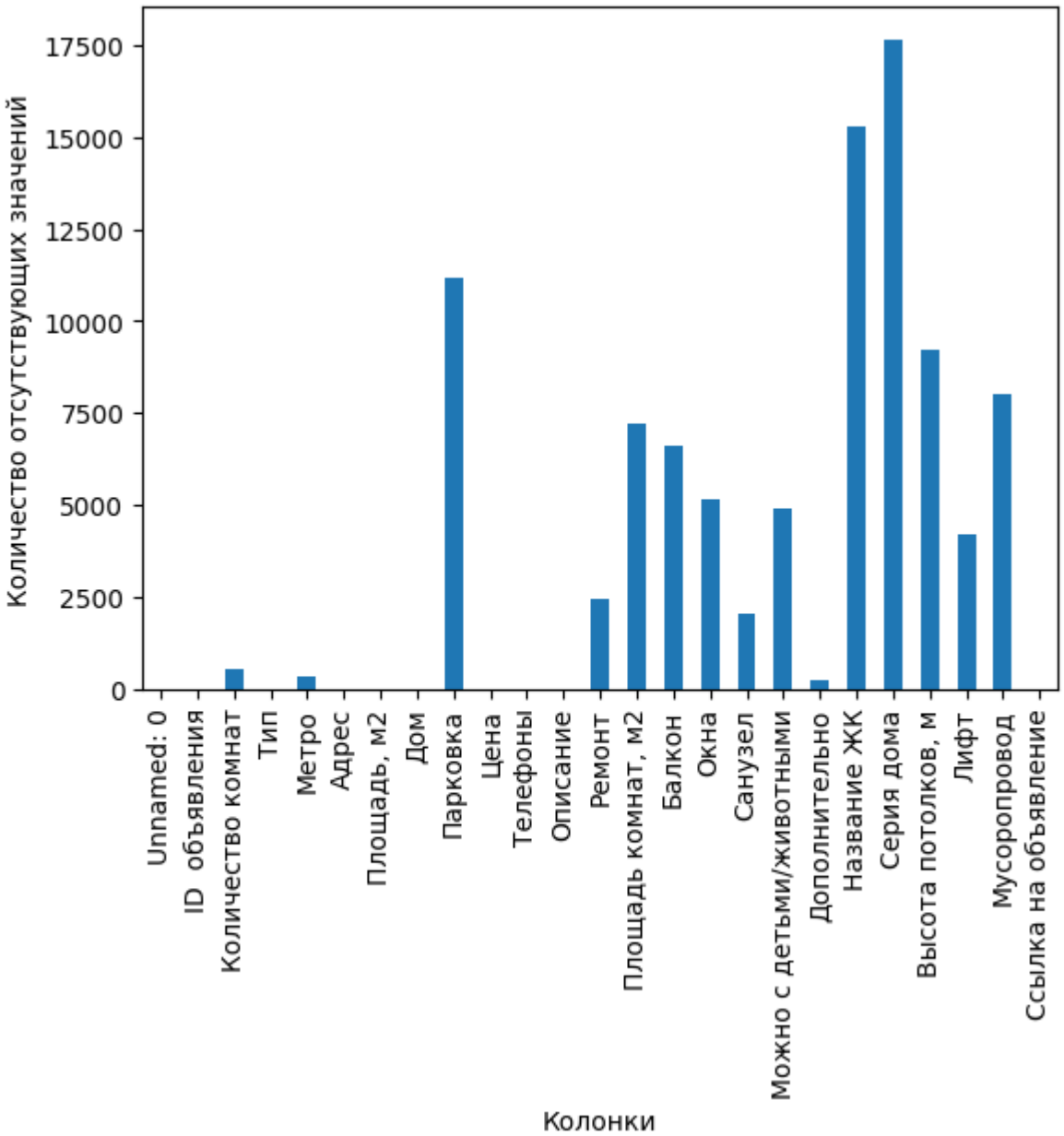
## Общие сведения о датасете:

```
Общее количество строк в датасете: 23368, количество колонок: 25.
Наименование колонок:
- Unnamed: 0
- ID  объявления
- Количество комнат
- Тип
- Метро
- Адрес
- Площадь, м2
- Дом
- Парковка
- Цена
- Телефоны
- Описание
- Ремонт
- Площадь комнат, м2
- Балкон
- Окна
- Санузел
- Можно с детьми/животными
- Дополнительно
- Название ЖК
- Серия дома
- Высота потолков, м
- Лифт
- Мусоропровод
- Ссылка на объявление
```

Учитывая, что пилотным регионом является Москва, из датасета были исключены строки с объектами недвижимости, расположенными не в Москве.

```
Количество строк, оставшихся после исключения объектов, расположенных не в Москве: 19737
```

Общие сведения об отсутствующих значениях по всему датасету:


Количество отсутствующих значений
17500
15000
12500
10000
7500
5000
2500
0
Unnamed: 0
ID объявления
Количество комнат
Тип
Метро
Адрес
Площадь, м2
Дом
Парковка
Цена
Телефоны
Описание
Ремонт
Площадь комнат, м2
Балкон
Окна
Санузел
Можно с детьми/животными
Дополнительно
Название ЖК
Серия дома
Высота потолков, м
Лифт
Мусоропровод
Ссылка на объявление
Колонки

В процессе реализации этапа EDA были исследованы данные по каждой колонке датасета, сделаны предварительные выводы о целесообразности либо нецелесообразности включения тех или иных колонок с данными в финальный датасет, о необходимости преобразования данных, о возможных стратегиях решения проблемы отсутствия значений.

При выполнении исследования и принятии решений об использовании либо отказе от использования тех или иных данных, содержащихся в датасете, принимались во внимание:

1. Предшествующий опыт и знания членов команды об исследуемом рынке.
2. Объективные ограничения команды по времени и доступным ресурсам.

## Результаты предварительного анализа по каждой колонке данных в датасете представлены ниже:

**Колонка** 0

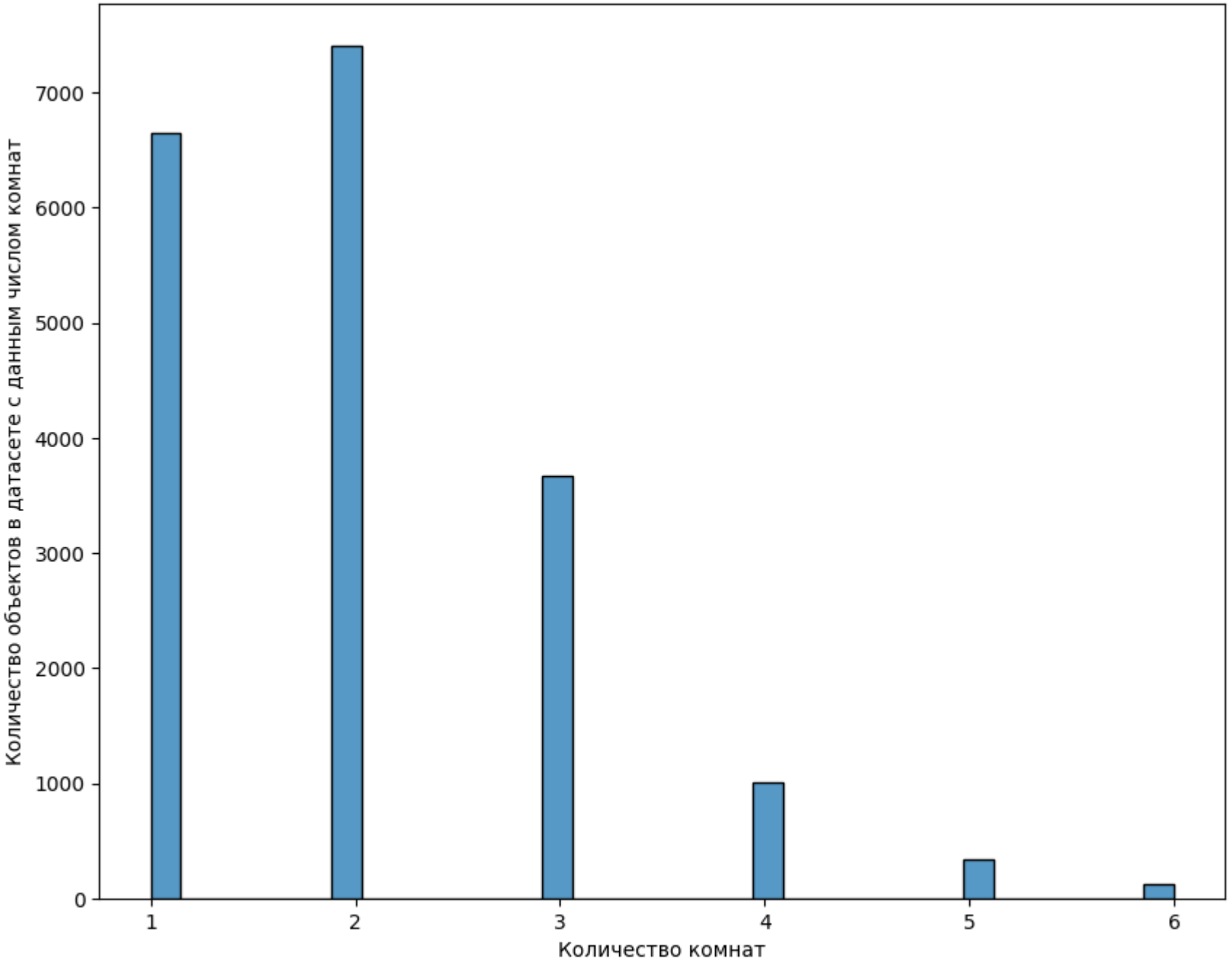
```
Название колонки: Unnamed: 0
Колонка представляет собой последовательный возрастающий ряд целых чисел, начиная от "0", дублирующий индекс первоначального датасета.
Уникальны ли значения в колонке:  да.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: int64
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.
```

**Колонка** 1

```
Название колонки: ID  объявления
Колонка содержит перечень идентификационных номеров предложений.
Уникальны ли значения в колонке:  да.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: int64
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.
```

**Колонка** 2

Название колонки: Количество комнат
Колонка содержит значение количества комнат в сдаваемом в аренду объекте недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 535. Доля отсутствующих значений в датасете: 2.71%.
Тип данных в колонке: object
Минимальное количество комнат в объекте недвижимости: 1, максимальное количество комнат в объекте недвижимости: 6


Количество объектов в датасете с данным числом комнат
0
1000
2000
3000
4000
5000
6000
7000
1
2
3
4
5
6
Количество комнат

Колонка подлежит использованию в финальном датасете, с переводом значений в формат int. Для разрешения проблемы частично отсутствующих значений
предполагается создание отдельной колонки с двумя возможными значениями: 0 или 1 (Indicator Method).

**Колонка** 3

Название колонки: Тип
Колонка содержит указание типа сдаваемого в аренду объекта недвижимости.
Уникальны ли значения в колонке: нет.
Перечень уникальных значений: ['Квартира']; таким образом, все значения в колонке идентичны друг другу.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Колонка подлежит исключению ввиду неинформативности (имеется единственное уникальное значение).

**Колонка** 4

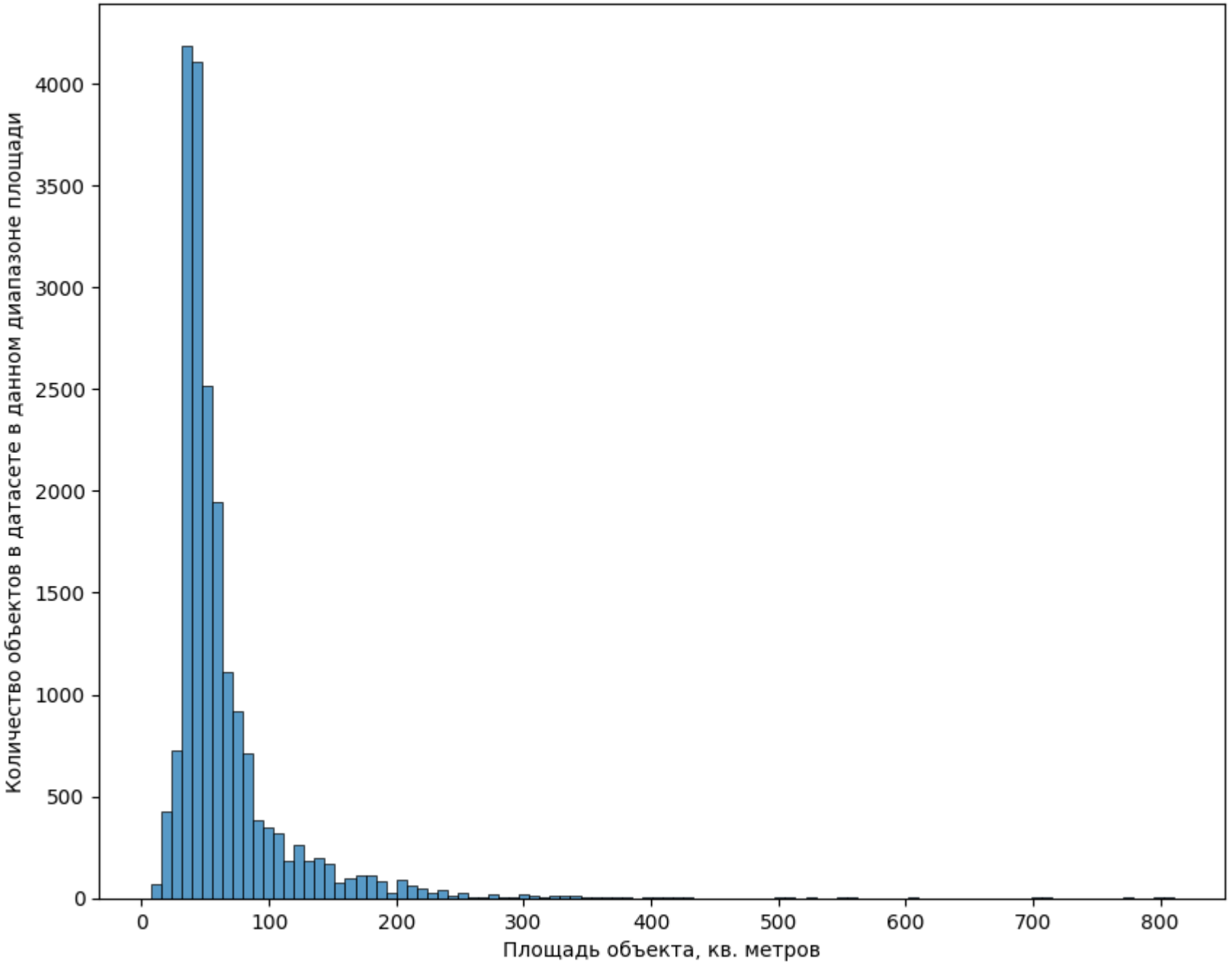
Название колонки: Метро
Колонка содержит информацию о станции метро, вблизи которой находится объект недвижимости, а также о времени, затрачиваемом на перемещение до станции метро.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 346. Доля отсутствующих значений в датасете: 1.7500000000000002%.
Тип данных в колонке: object
В финальный датасет будет включена (посредством метода One Hot Encoding) информация из колонки Метро (в части названия станций метро).
В случае сохранения информации о названиях станций метро, строки с отсутствующими значениями (ввиду их малочисленности) будут исключены из датасета (Complete Case Analysis).
Подлежит уточнению и дополнительной оценке целесообразность включения в финальный датасет информации о времени перемещения до станции метро.

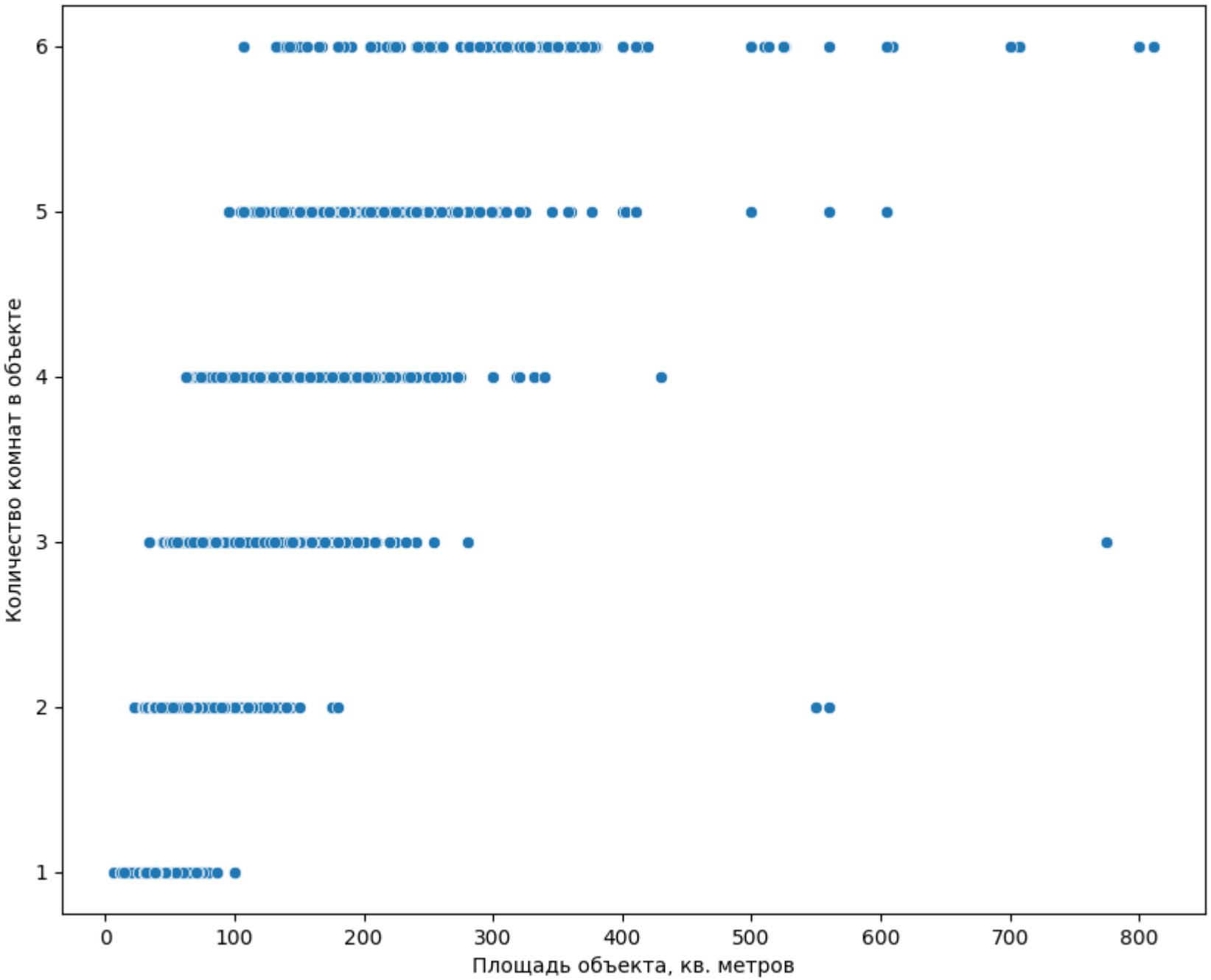
**Колонка** 5

Название колонки: Адрес
Колонка содержит информацию об адресе объекта.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Ввиду затруднительности применения метода One Hot Encoding к данным из колонки Адрес, целесообразно её исключение.

**Колонка** 6

```
Название колонки: Площадь, м2
Колонка содержит информацию о площади объекта недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Площадь объекта с минимальной площадью: 7.0, площадь объекта с максимальной площадью: 811.0
```


Количество объектов в датасете в данном диапазоне площади
4000
3500
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
Площадь объекта, кв. метров


Количество комнат в объекте
Площадь объекта, кв. метров
1
2
3
4
5
6
0
100
200
300
400
500
600
700
800

Колонка подлежит сохранению с переводом значений (первое значение в строке, соответствующее общей площади) в числовой формат.

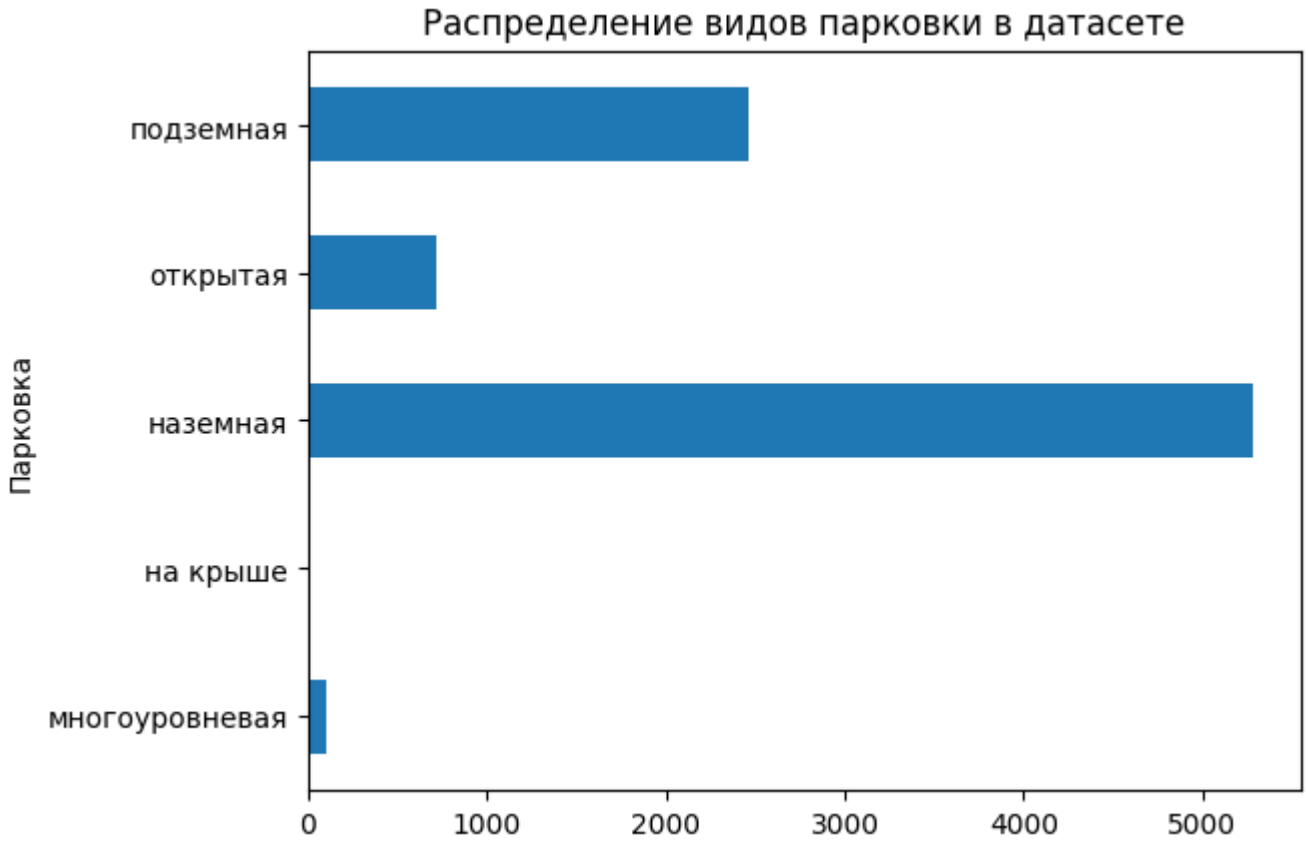
**Колонка** 7

Название колонки: Дом
Колонка содержит информацию об этажности домов, этажах расположения объектов недвижимости в домах, а также о типах домов.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Типы домов, представленных в датасете:
- Монолитный
- Монолитно-кирпичный
- Панельный
- Сталинский
- Кирпичный
- старый фонд
- Блочный
- Деревянный
- Щитовой

Данные из колонки подлежат использованию в финальном датасете. Для разрешения проблемы частично отсутствующих значений (в части типов домов) предполагается создание отдельной колонки с двумя возможными значениями: 0 или 1 (Indicator Method).

**Колонка** 8

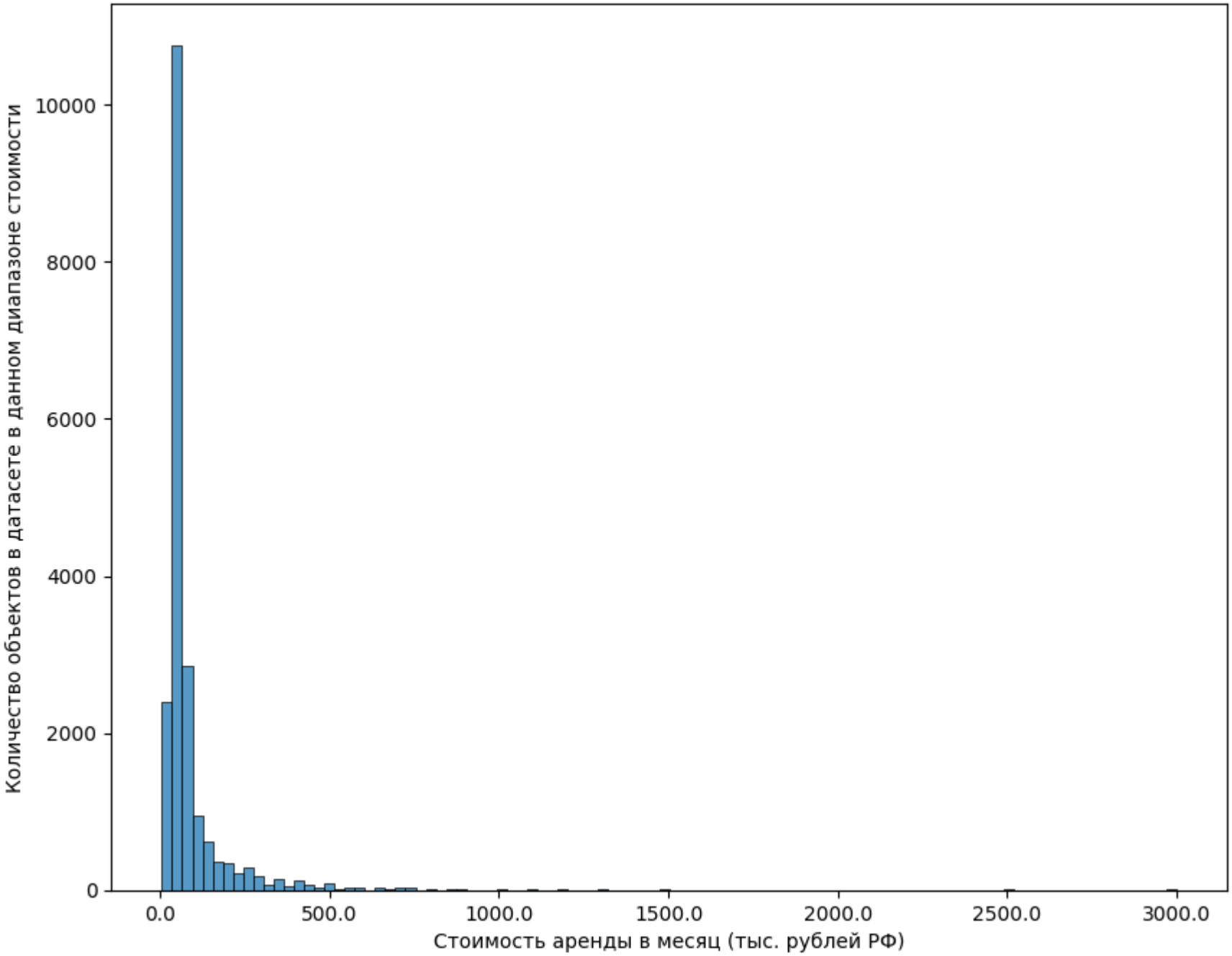
Название колонки: Парковка
Колонка содержит информацию о парковке.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 11174. Доля отсутствующих значений в датасете: 56.61000000000001%.
Тип данных в колонке: object

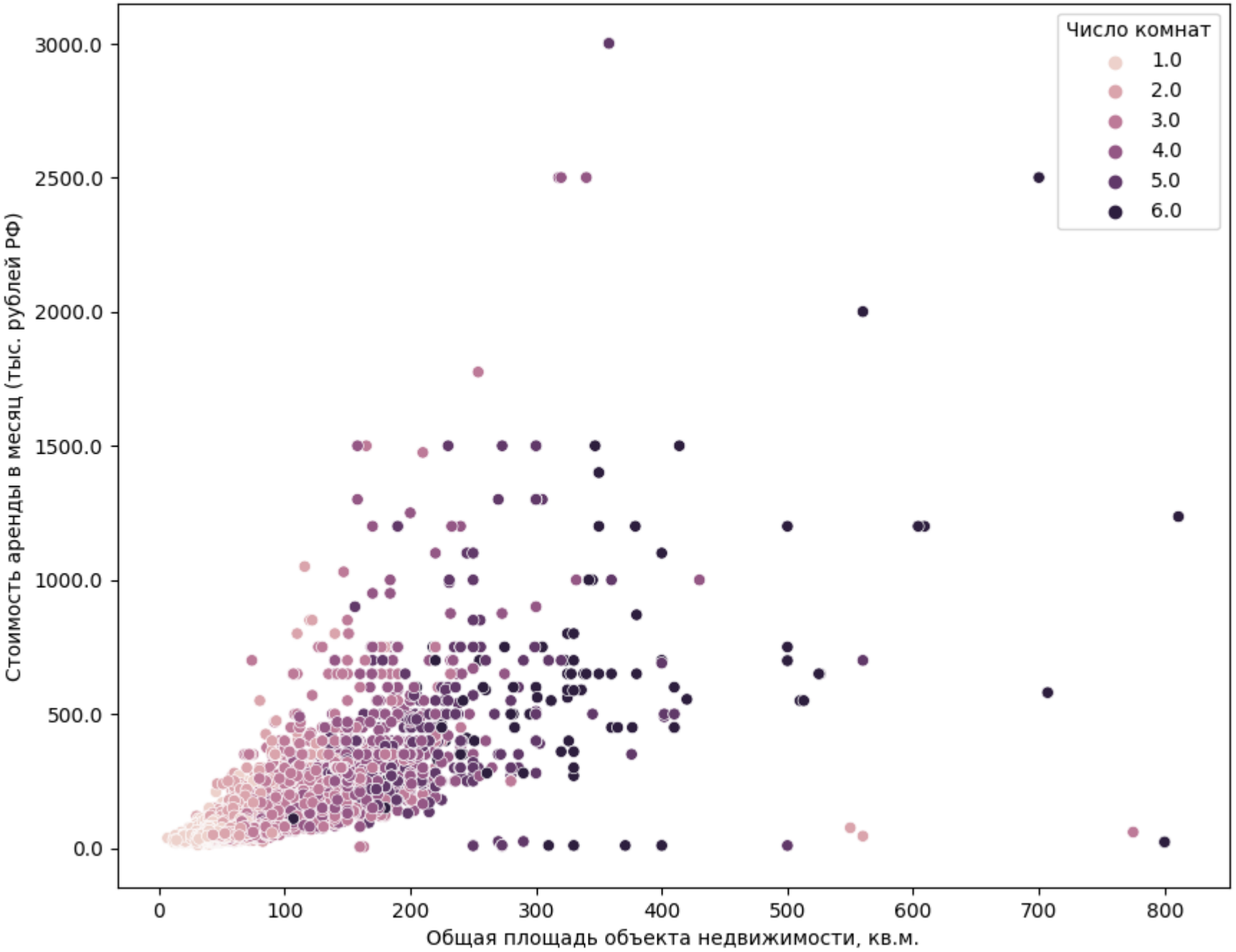



Ввиду большого числа пропусков данных сохранение колонки в финальном датасете нецелесообразно.

**Колонка** 9

Название колонки: Цена
Колонка содержит информацию о ежемесячной стоимости аренды (в рублях РФ), о дополнительных условиях, относящихся к цене (залог, включение коммунальных услуг в стоимость).
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Минимальная стоимость аренды: 5000.0 рублей РФ, максимальная стоимость аренды: 3000000.0 рублей РФ.


Количество объектов в датасете в данном диапазоне стоимости
10000
8000
6000
4000
2000
0
0.0
500.0
1000.0
1500.0
2000.0
2500.0
3000.0
Стоимость аренды в месяц (тыс. рублей РФ)


Число комнат
1.0
2.0
3.0
4.0
5.0
6.0
3000.0
2500.0
2000.0
1500.0
1000.0
500.0
0.0
Стоимость аренды в месяц (тыс. рублей РФ)
0
100
200
300
400
500
600
700
800
Общая площадь объекта недвижимости, кв.м.

Колонка подлежит сохранению в части ежемесячной стоимости аренды в требуемом числовом формате.

**Колонка** 1 0

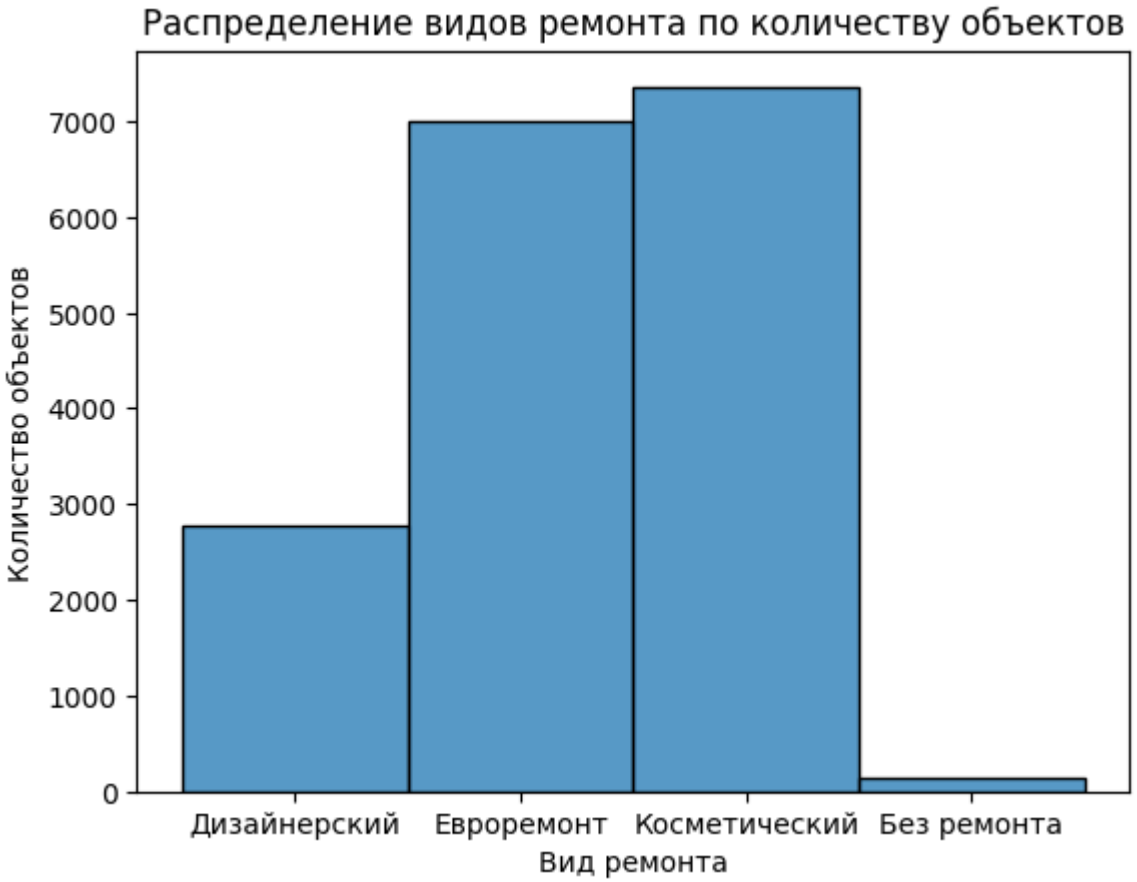
Название колонки: Телефоны
Колонка содержит информацию о контактных телефонах.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.

**Колонка** 1 1

Название колонки: Описание
Колонка содержит описания предложений по аренде в текстовом формате.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.

**Колонка** 1 2

Название колонки: Ремонт
Колонка содержит перечень идентификационных номеров предложений.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 2463. Доля отсутствующих значений в датасете: 12.479999999999999%.
Тип данных в колонке: object




Колонка подлежит сохранению с приведением в требуемый формат данных (с использованием метода OHE).

**Колонка** 1 3

Название колонки: Площадь комнат, м2
Колонка содержит указание площади комнат в объектах недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 7228. Доля отсутствующих значений в датасете: 36.620000000000005%.
Тип данных в колонке: object
Колонка подлежит исключению ввиду её производного характера (т.к. сохраняются колонки с указанием числа комнат и общей площади), а также большого числа пропусков данных.

**Колонка** 14

Название колонки: Балкон
Колонка содержит информацию о наличии и количестве балконов/лоджий в объектах недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 6630. Доля отсутствующих значений в датасете: 33.589999999999996%.
Тип данных в колонке: object
Колонка подлежит исключению ввиду большого числа пропусков данных.

**Колонка** 15

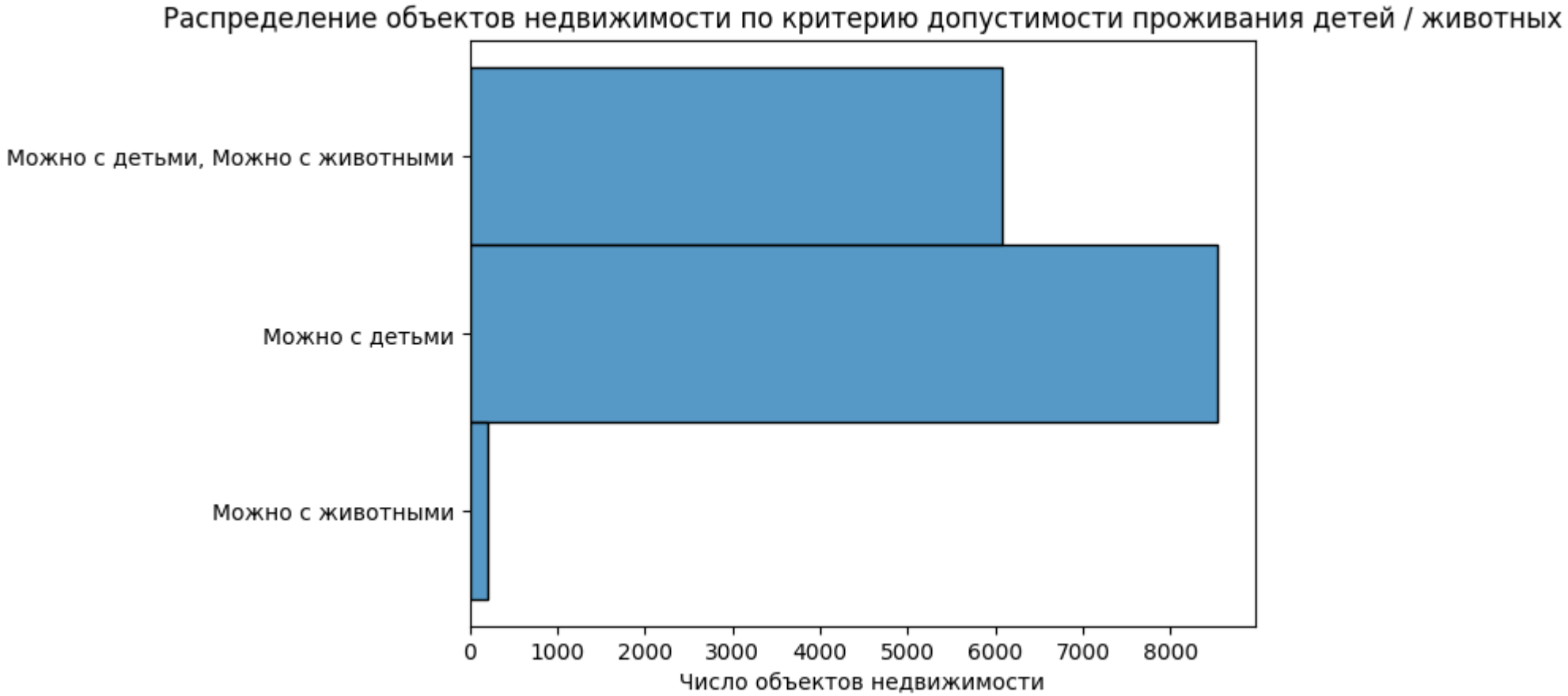
Название колонки: Окна
Колонка содержит информацию о направлении выхода окон объекта недвижимости (на улицу, во двор).
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 5150. Доля отсутствующих значений в датасете: 26.090000000000003%.
Тип данных в колонке: object
Колонка подлежит исключению ввиду большого числа пропусков данных.

**Колонка** 16

Название колонки: Санузел
Колонка содержит информацию о совмещённом / раздельном характере санузла на объекте недвижимости.
Уникальны ли значения значения в колонке: нет
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 2041. Доля отсутствующих значений в датасете: 10.34%
Колонка подлежит исключению ввиду большого числа пропусков данных и предварительно экспертно оценённой нерелевантности параметра для определения стоимости аренды.

**Колонка** 17

Название колонки: Можно с детьми/животными
Колонка содержит информацию о допустимости проживания детей / животных на объекте недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 4915. Доля отсутствующих значений в датасете: 24.9%.
Тип данных в колонке: object




Рассматривается возможность сохранения колонки для финального датасета в случае идентификации способа разрешения проблемы большого числа отсутствующих значений.

**Колонка** 1 8

Название колонки: Дополнительно
Колонка содержит о наличии мебели и иных предметов быта на объекте недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 272. Доля отсутствующих значений в датасете: 1.38%.
Тип данных в колонке: object
Колонка подлежит сохранению (в полном или неполном объёме) ввиду экспертной оценки существенного влияния данного параметра на стоимость аренды.

**Колонка** 19

Название колонки: Название ЖК
Колонка содержит перечень названий жилых комплексов (если это применимо), в которых находятся объекты недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 15281. Доля отсутствующих значений в датасете: 77.42%.
Тип данных в колонке: object
Колонка подлежит исключению ввиду нерелевантности данных и большого числа пропусков.

**Колонка** 20

Название колонки: Серия дома
Колонка содержит перечень серий домов, в которых находятся объекты недвижимости.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 17646. Доля отсутствующих значений в датасете: 89.41%.
Тип данных в колонке: object
Колонка подлежит исключению ввиду большого числа пропусков.

**Колонка** 21

Название колонки: Высота потолков, м
Колонка содержит информацию о высотке потолков.
Уникальны ли значения в колонке: нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 9202. Доля отсутствующих значений в датасете: 46.62%.
Тип данных в колонке: float64
Колонка подлежит исключению ввиду большого числа пропусков.

**Колонка** 22

Название колонки: Лифт
Колонка содержит информацию о наличии и типе имеющихся лифтов в доме.
Уникальны ли значения в колонке:  нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 4192. Доля отсутствующих значений в датасете: 21.240000000000002%
Тип данных в колонке: object
Рассматривается возможность сохранения колонки ввиду экспертно оценённого влияния параметра на стоимость аренды с учётом большого числа пропусков данных.

**Колонка** 2 3

Название колонки: Мусоропровод
Колонка содержит информацию о наличии / отсутствии мусоропровода.
Уникальны ли значения в колонке:  нет.
Содержит ли колонка отсутствующие значения: да. Количество отсутствующих значений: 8007. Доля отсутствующих значений в датасете: 40.57%.
Тип данных в колонке: object
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.

**Колонка** 2 4

Название колонки: Ссылка на объявление
Колонка содержит перечень ссылок на объявления о сдаче соответствующих объектов недвижимости в аренду.
Уникальны ли значения в колонке: да.
Содержит ли колонка отсутствующие значения: нет.
Тип данных в колонке: object
Колонка подлежит исключению, так как не содержит ценных данных для последующей работы.