

# Исследование статистики о недвижимости в Москве

Проект в рамках всероссийской научно-технологической программы по решению проектных задач в области искусственного интеллекта и смежных дисциплин «Сириус.ИИ»



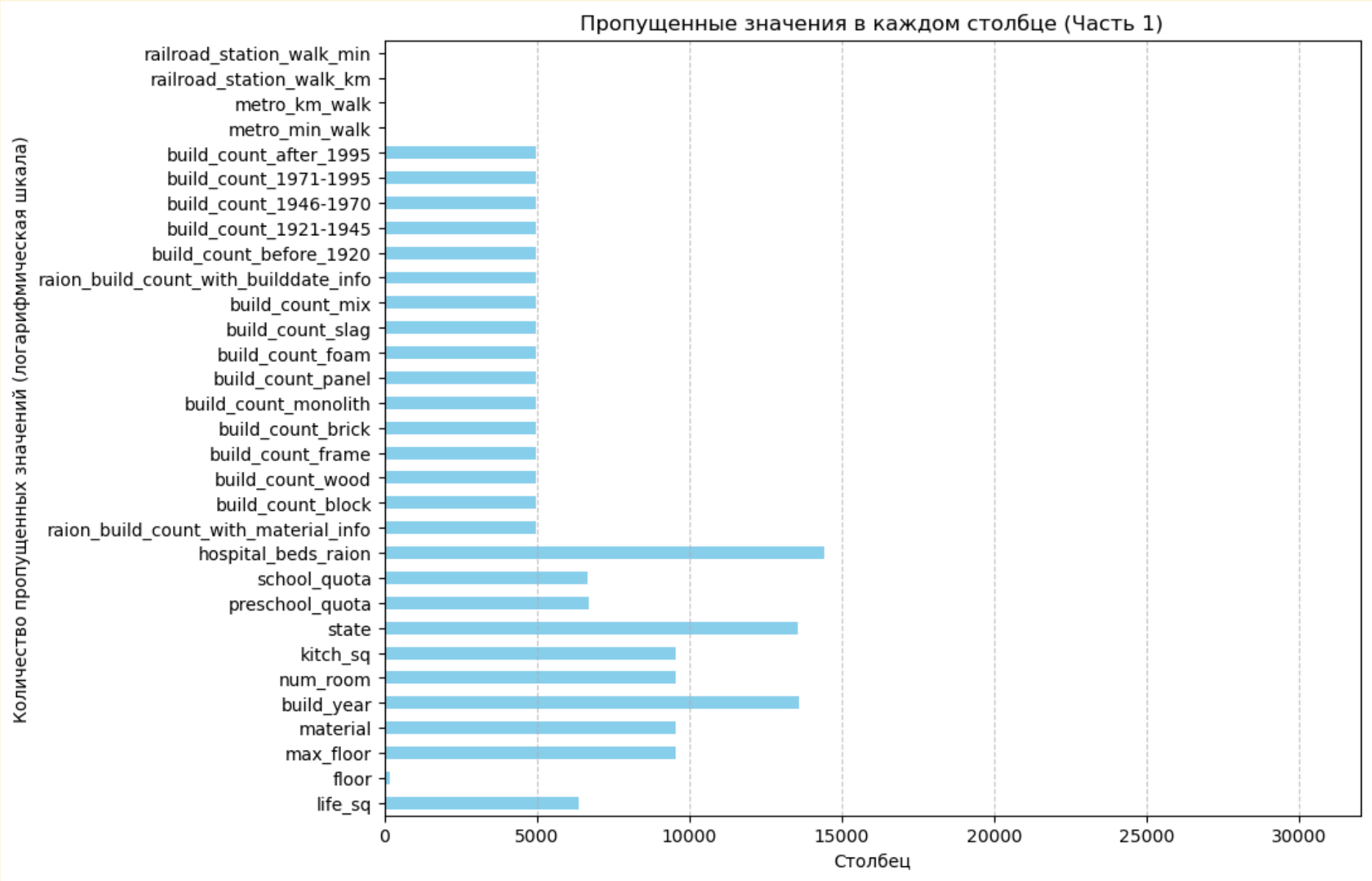
# Состав проектной команды

- Беляков Михаил Евгеньевич, Бийск, Алтайский край – Разработка кода исследования и машинное обучение.
- Мораст Марк Анатольевич, Бийск, Алтайский край – Анализ информации. Выгрузка на Github.

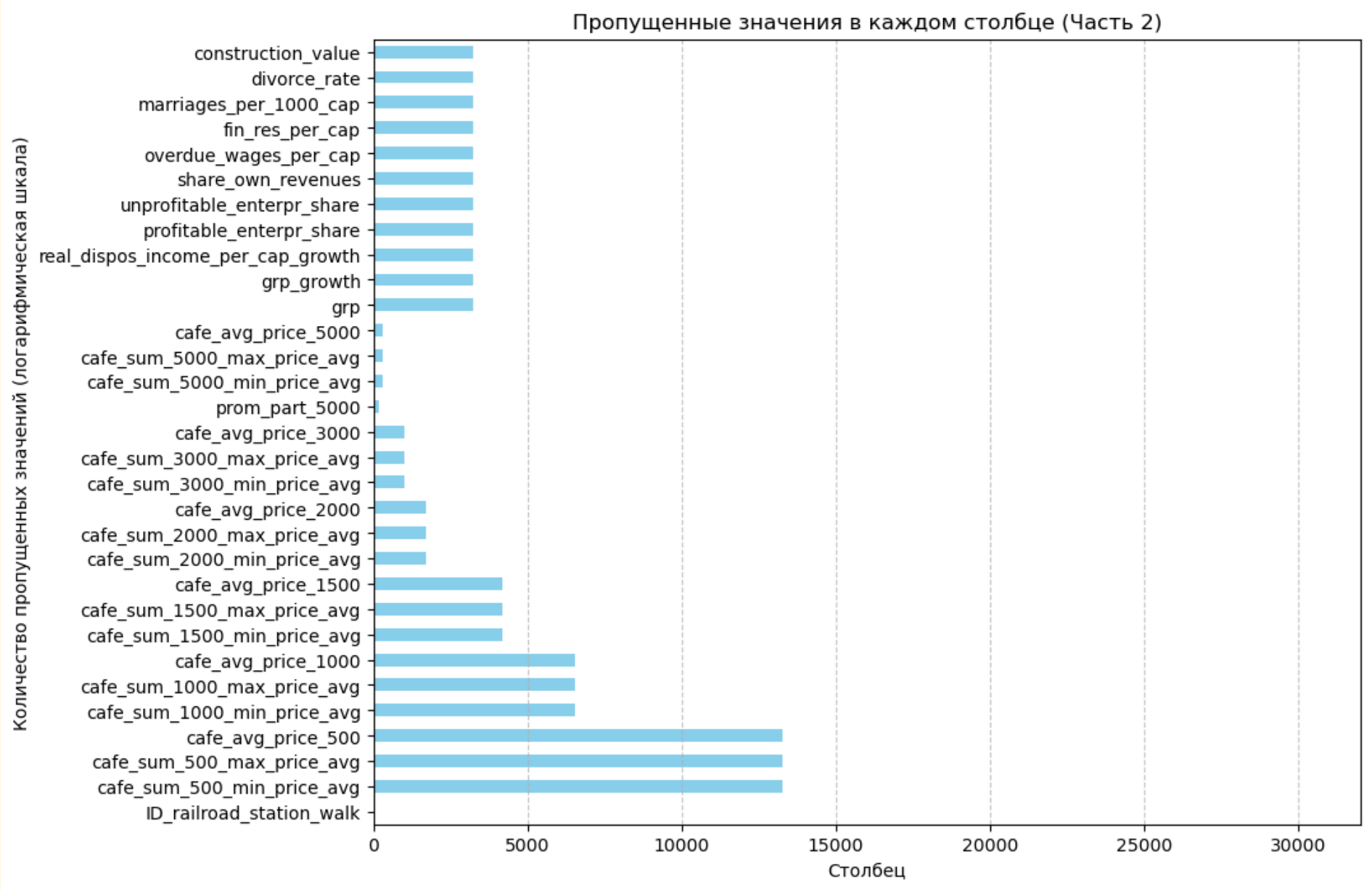
# Обработка отсутствующих значений

1. **Чтение и подготовка данных:** Импорт необходимых библиотек (pandas для работы с датасетами и matplotlib для визуализации). Загрузка двух наборов данных: train.csv и macro.csv, с указанием, что столбец "timestamp" следует интерпретировать как дату. Эти данные затем объединяются по столбцу "timestamp", создавая объединенный датасет для анализа.
2. **Проверка объединенных данных:** Вывод на экран первых нескольких строк объединенного датасета для визуальной проверки корректности объединения.
3. **Анализ пропущенных значений:** Определение количества пропущенных значений в каждом столбце. Этот шаг важен для понимания, насколько полны данные и какие столбцы содержат слишком много пропущенных значений, что может снизить качество анализа.
4. **Фильтрация столбцов по пропущенным значениям:** Отбор тех столбцов, где количество пропущенных значений превышает 10. Это предварительный шаг к решению об удалении или замене этих пропущенных значений.
5. **Визуализация пропущенных значений:** Деление отфильтрованных столбцов на части для визуализации количества пропущенных значений. Построение гистограмм позволяет наглядно увидеть, в каких столбцах проблема пропущенных значений наиболее остра. Используется разбиение на части, чтобы обеспечить читабельность графиков.
6. **Удаление столбцов с большим количеством пропущенных значений:** Исключение из анализа столбцов, в которых количество пропущенных значений превышает 13000. Это решение основано на предположении, что такое большое количество пропущенных данных может сделать эти переменные ненадежными для анализа.
7. **Оценка результатов обработки:** Вывод на экран количества столбцов до и после удаления, чтобы оценить, сколько данных было исключено из анализа.

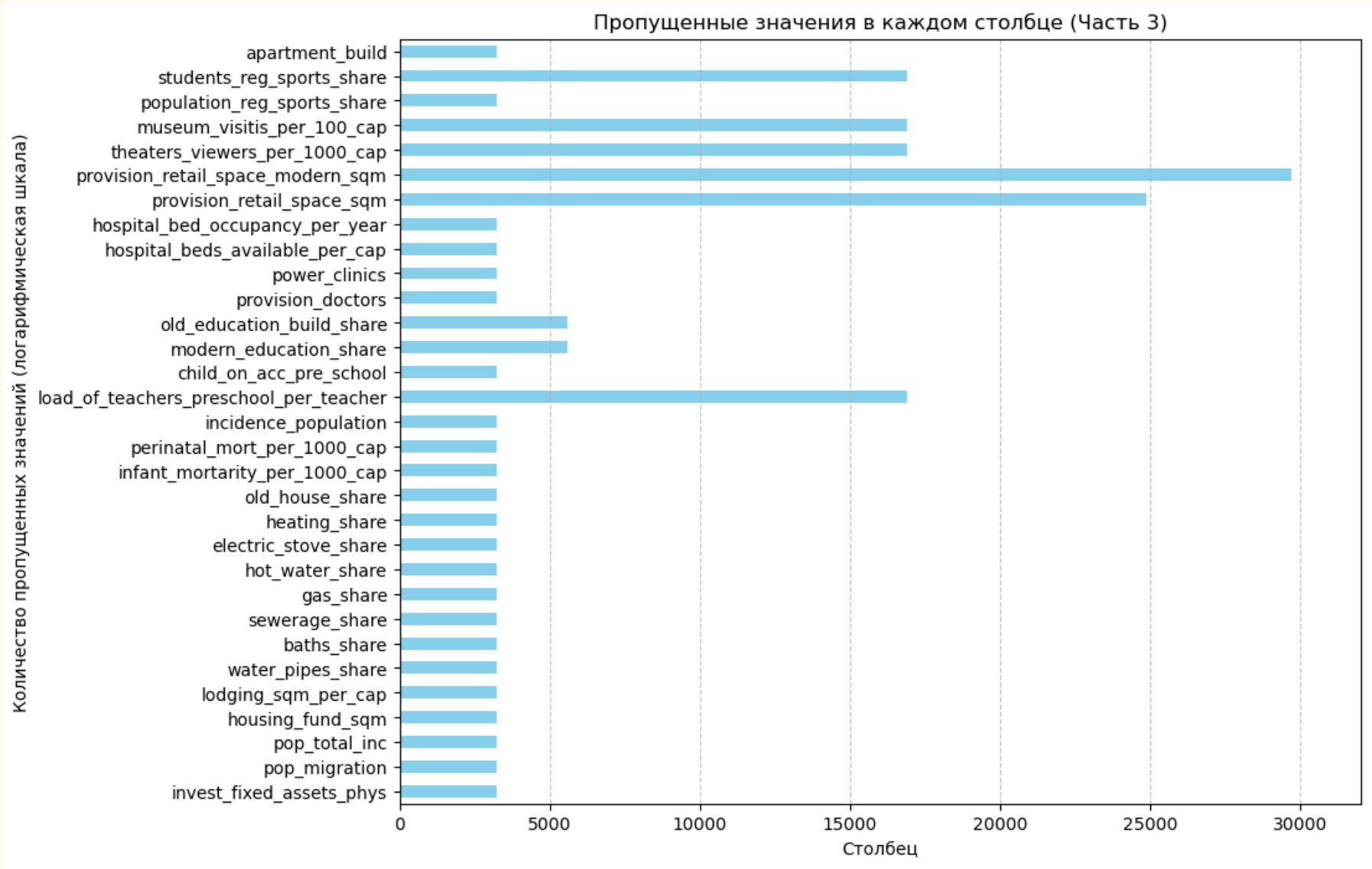
# Приложение к первому пункту (графический вывод)



# Приложение к первому пункту (графический вывод)



# Приложение к первому пункту (графический вывод)



# Обработка лишних значений

1. Сначала вычисляется матрица корреляции для всех признаков датасета. Используется метод `.corr()` `DataFrame`, абсолютные значения корреляций взяты методом `.abs()`, чтобы учитывать как положительную, так и отрицательную корреляцию.
2. Вывод количества столбцов до удаления: Перед применением каких-либо фильтров или удалений выводится текущее количество столбцов в `DataFrame`, чтобы можно было сравнить его с количеством после обработки.
3. Анализ корреляции с целевой переменной: Осуществляется расчет корреляции каждого признака с целевой переменной `price_doc`. Результаты сортируются по убыванию абсолютного значения корреляции, чтобы определить наиболее значимые признаки.
4. Определение и удаление столбцов с низкой корреляцией с целевой переменной: Задается порог корреляции (`threshold`), ниже которого считается, что признаки не имеют значимой связи с целевой переменной. Все столбцы, корреляция которых с `price_doc` ниже этого порога, подлежат удалению. Таким образом, уменьшается размерность данных, исключая менее информативные признаки.
5. Вывод информации о процессе удаления: Выводится информация об удаленных столбцах, что позволяет оценить, какие признаки были исключены из анализа.
6. Вывод информации о столбцах, оставшихся после удаления: После удаления выводится перечень оставшихся признаков, а также их количество, что позволяет оценить объем и структуру данных, с которыми предстоит работать далее. Он отсортирован по убыванию, что позволяет понять какие из столбцов самые важные и больше всех связаны с целевой переменной.

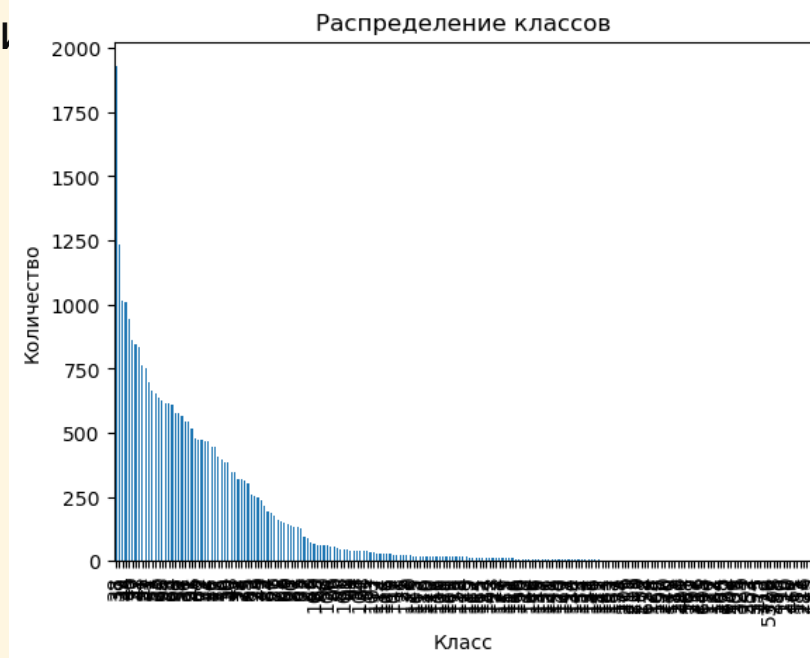
# Выявление аномалий

- В данном этапе мы строим гистограмму распределения значений для столбца 'full\_sq'. Гистограмма позволяет визуально оценить форму распределения данных и выявить возможные выбросы или аномалии.
- После построения гистограммы мы применяем метод межквартильного размаха для определения выбросов. Этот метод основан на интерквартильном размахе (IQR), который вычисляется как разница между третьим и первым квартилями. Затем выбросы определяются как значения, находящиеся за пределами верхней и нижней границы, определенных как  $Q1 - 1.5 * IQR$  и  $Q3 + 1.5 * IQR$  соответственно.



# Сбалансированность

- В этой части мы анализируем сбалансированность данных по каким-либо признакам, например, по значению 'full\_sq'. Мы сначала подсчитываем количество уникальных значений в столбце и вычисляем процентное соотношение каждого уникального значения от общего числа записей.
- После этого мы визуализируем распределение классов с использованием столбчатой диаграммы. Это позволяет наглядно представить баланс между различными категориями.



аковой имеется.

## 5. Базовый отбор признаков:

- В данной части мы анализируем влияние признаков на целевую переменную ('price\_doc'), которая представляет собой цену объекта недвижимости. Мы начинаем с расчета корреляционной матрицы, чтобы определить степень линейной зависимости между признаками и целевой переменной.
- После этого мы визуализируем корреляционную матрицу с помощью тепловой карты, чтобы наглядно представить взаимосвязи между признаками. Кроме того, мы строим диаграмму рассеяния, чтобы визуально оценить зависимость между переменными и целевой 'price\_doc'.

# Сравнение и вывод

|      | count     | mean         | std          | min      | 25%       | 50%       | \ |
|------|-----------|--------------|--------------|----------|-----------|-----------|---|
| year |           |              |              |          |           |           |   |
| 2011 | 753.0     | 5.929668e+06 | 3.829036e+06 | 340000.0 | 4300000.0 | 5500000.0 |   |
| 2012 | 4839.0    | 6.354435e+06 | 4.518082e+06 | 190000.0 | 4174267.0 | 5650000.0 |   |
| 2013 | 7978.0    | 6.712150e+06 | 4.313564e+06 | 260000.0 | 4462000.0 | 5900000.0 |   |
| 2014 | 13662.0   | 7.449468e+06 | 4.882734e+06 | 100000.0 | 5000000.0 | 6552000.0 |   |
| 2015 | 3239.0    | 8.183914e+06 | 5.610930e+06 | 500000.0 | 5350000.0 | 7100000.0 |   |
|      | 75%       | max          |              |          |           |           |   |
| year |           |              |              |          |           |           |   |
| 2011 | 7150000.0 | 37000000.0   |              |          |           |           |   |
| 2012 | 7300000.0 | 111111112.0  |              |          |           |           |   |
| 2013 | 7700000.0 | 91066096.0   |              |          |           |           |   |
| 2014 | 8500000.0 | 80777440.0   |              |          |           |           |   |
| 2015 | 9534027.5 | 95122496.0   |              |          |           |           |   |

---

|      | count      | mean         | std          | min | 25%       | 50%       | \ |
|------|------------|--------------|--------------|-----|-----------|-----------|---|
| year |            |              |              |     |           |           |   |
| 2021 | 11358150.0 | 6.787516e+06 | 1.977118e+08 | 0.0 | 2600000.0 | 3995000.0 |   |
|      | 75%        | max          |              |     |           |           |   |
| year |            |              |              |     |           |           |   |
| 2021 | 6500000.0  | 6.355524e+11 |              |     |           |           |   |

Преобразование столбца временной метки:  
В этой части мы преобразуем столбец временных меток (например, 'timestamp' или 'date') в формат даты и извлекаем из него год.

Затем мы агрегируем данные по годам и рассчитываем основные статистики (среднее, стандартное отклонение, минимум, максимум, квартили) для целевой переменной ('price\_doc') для каждого года.

Вывод:

После проведения всех этих шагов мы можем сделать более информативные выводы о данных. Аномалии и выбросы были обработаны, сбалансированность данных проанализирована, выявлены признаки, оказывающие влияние на целевую переменную, и произведено преобразование столбца временной метки для дальнейшего анализа.

Такой подробный анализ помогает лучше понять данные, выделить особенности и сделать более обоснованные выводы о закономерностях и зависимостях в данных.

# Результат исследования

Результатом исследования стали файлы на Github:

А так же вывод о том, что цена в основном зависит от площади квартиры, и то, что цена постепенно и прямо пропорционально растет во времени.