Анализ и прогнозирование цен на жилье в Москве

Процкая Вера, Серов Кирилл, Романовский Илья

332 группа

Введение

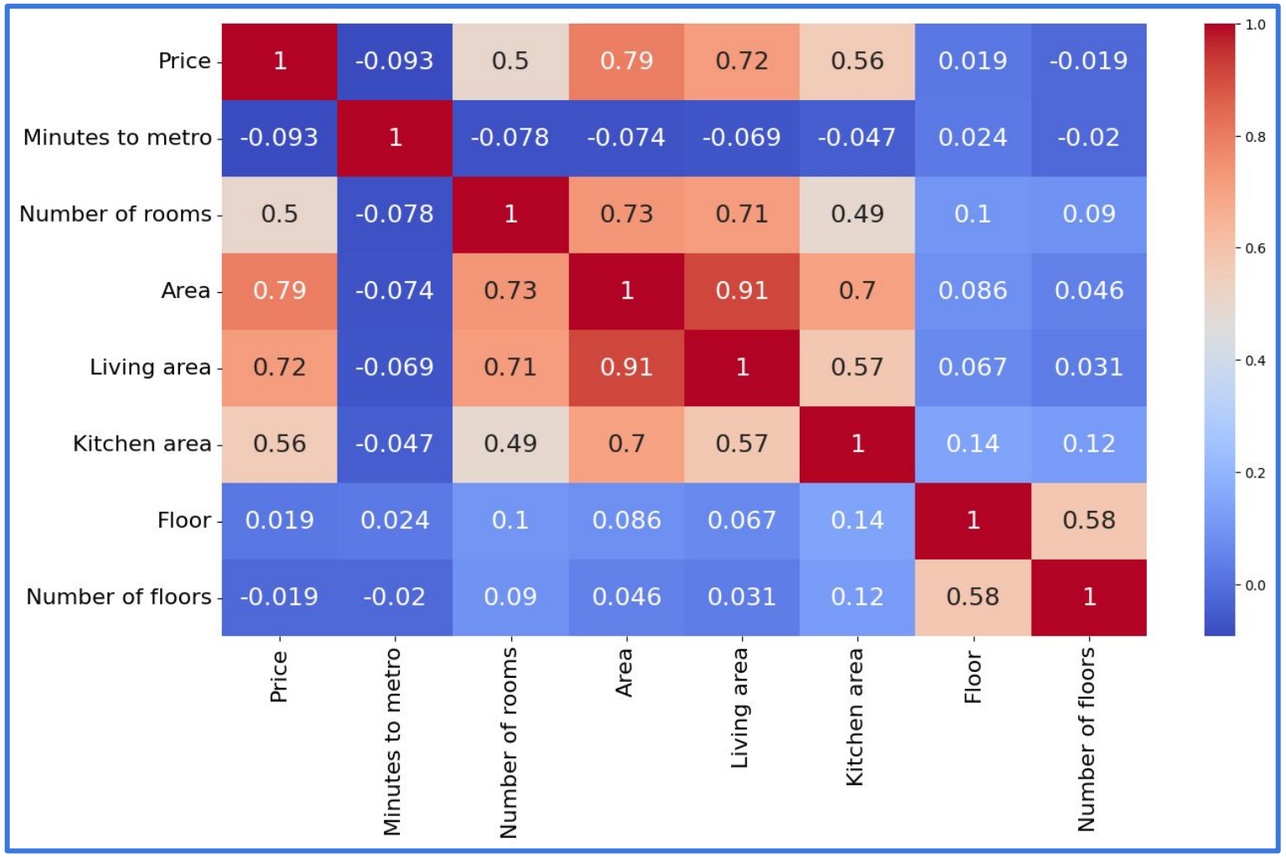
Проект посвящён исследованию и построению моделей машинного обучения для прогнозирования стоимости квартир в Москве на основе набора данных Moscow Housing Price Dataset. Мы изучаем влияние различных характеристик жилья, таких как район, площадь, количество комнат, этаж и другие, чтобы выявить ключевые факторы, определяющие цену. В презентации подробно рассматриваются этапы работы — от разведочного анализа данных до создания прогнозных моделей и кластеризации районов по ценам с помощью метода k-means.

Разведочный анализ данных (EDA)

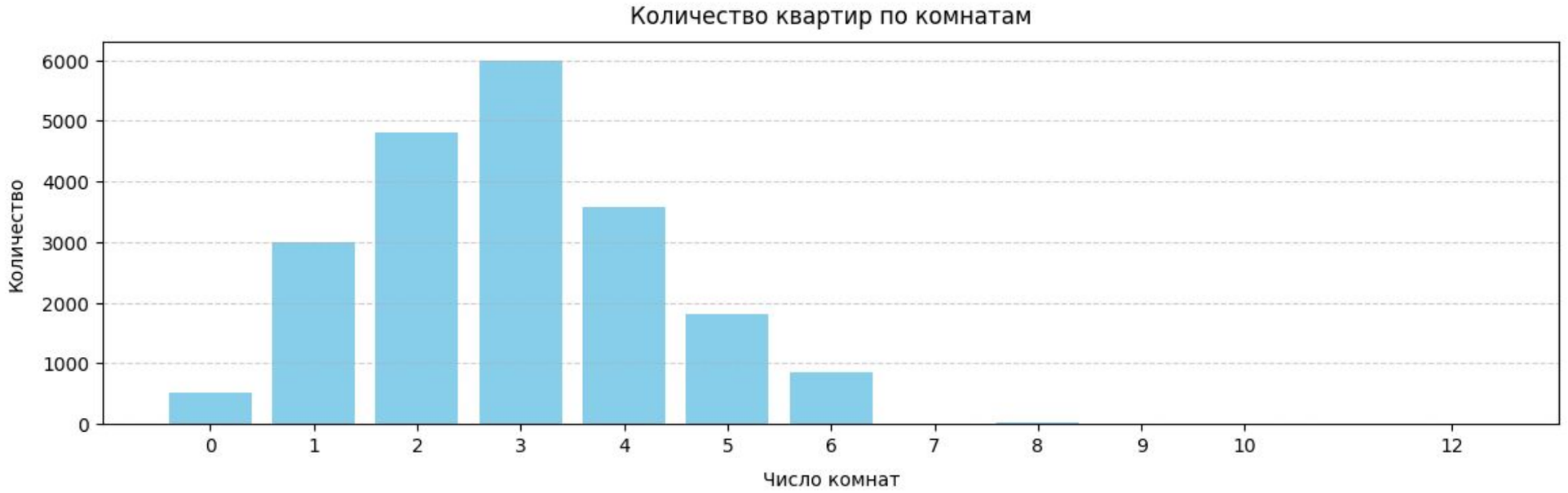
На начальном этапе проекта происходила загрузка и первичный осмотр датасета. Особое внимание уделялось обработке пропусков и выявлению аномальных значений. Так в процессе было убрано 1835 дубликатов, а также исключили 236 подозрительных объектов, которые имели 0 комнат и больше 50 кв. м. площади. Далее анализируется распределение целевой переменной — цена квартир, для удобства цена была переведена в миллионы рублей.

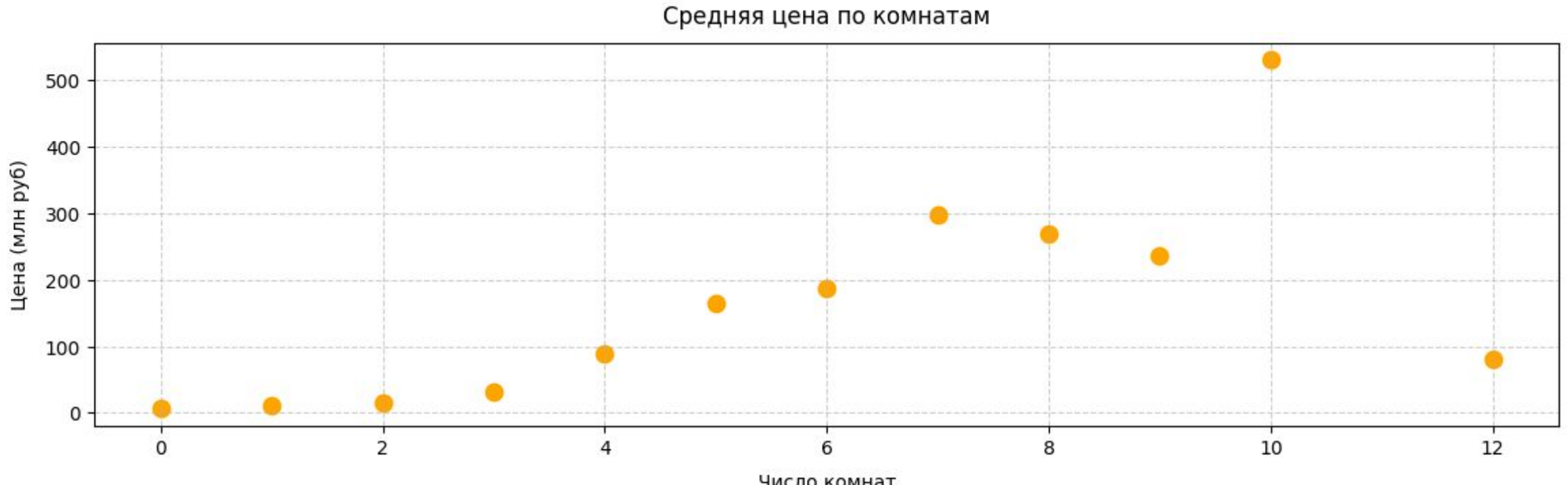
Помимо этого исследовали:

- Корреляция признаков
- Визуализация зависимостей
- Анализ влияния этажа и района

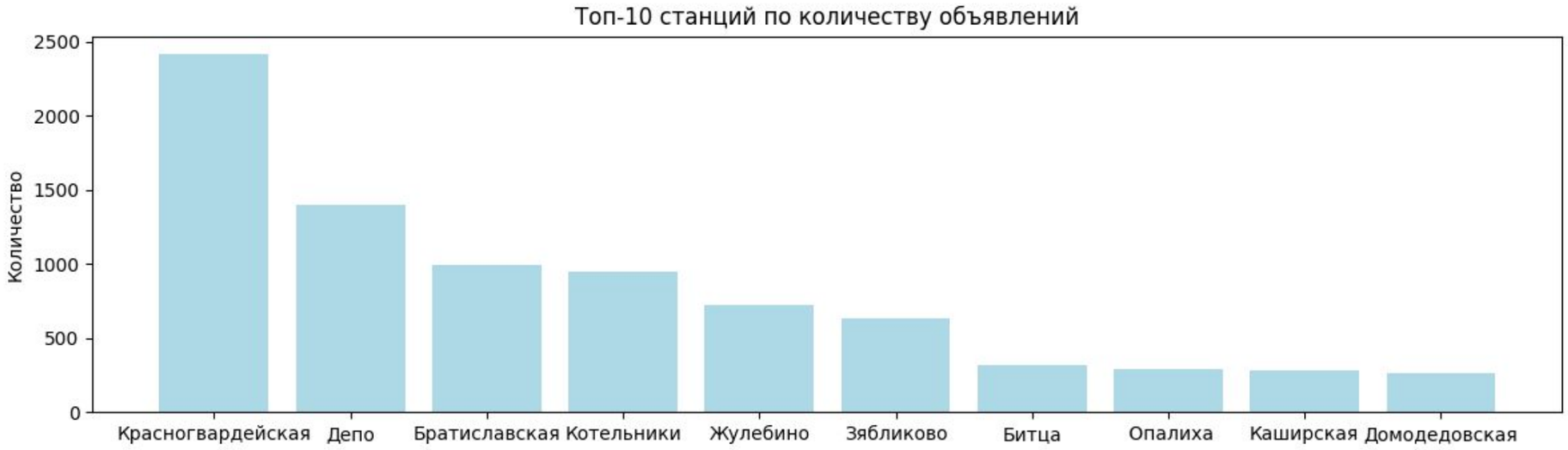


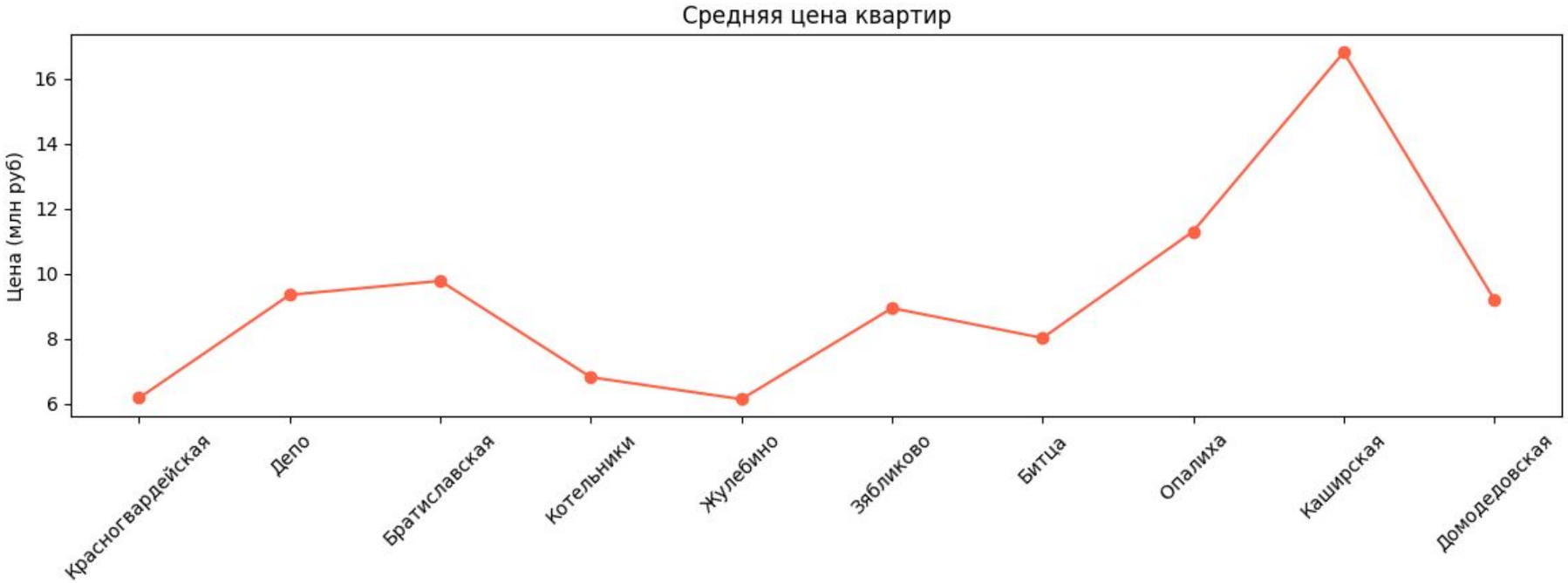
Визуализация ключевых зависимостей



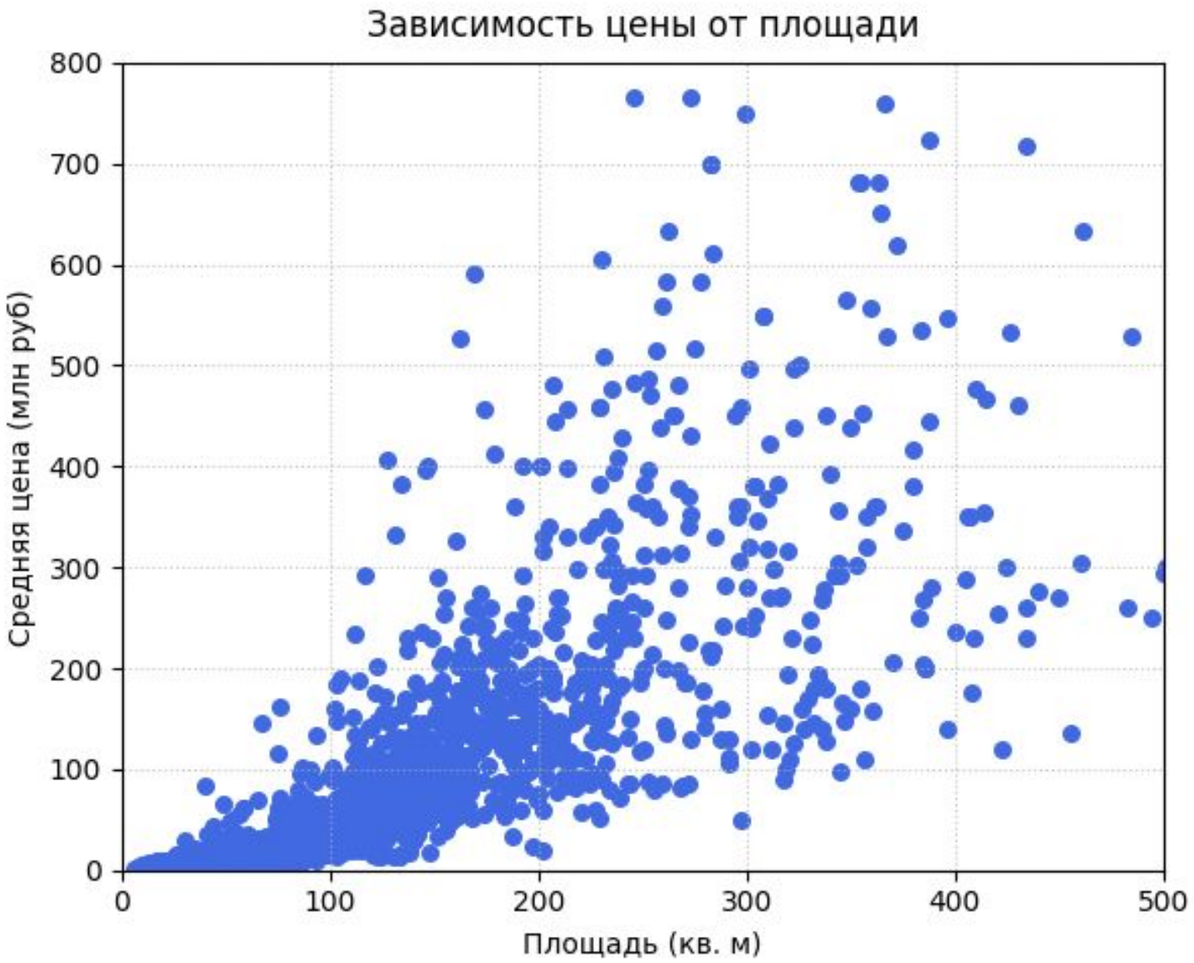


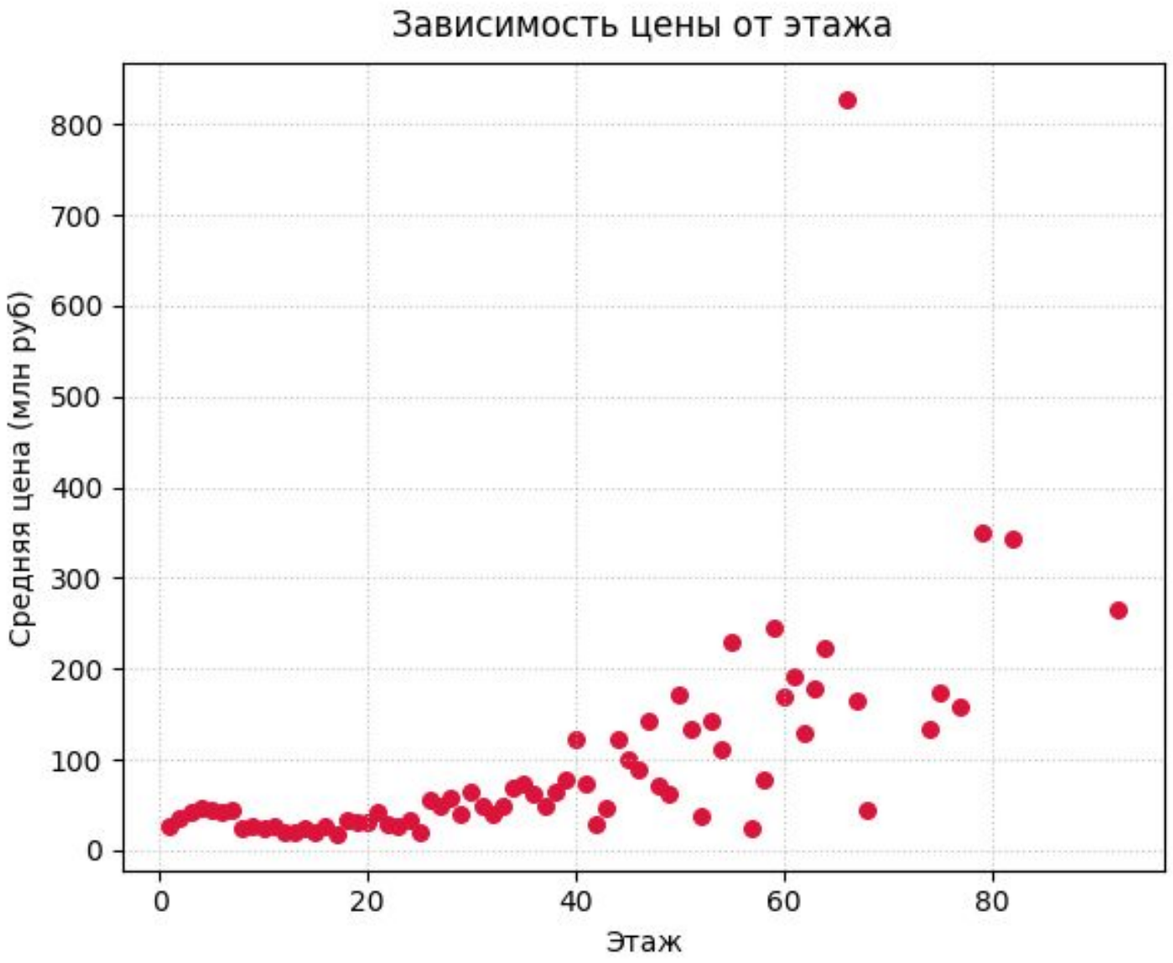
Визуализация ключевых зависимостей





Визуализация ключевых зависимостей





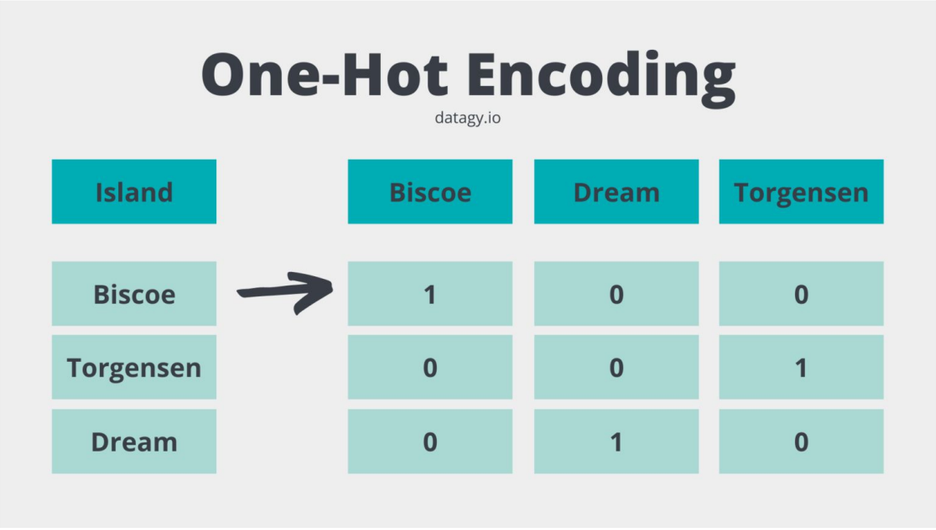
Предобработка данных для моделирования

На подготовительном этапе происходит нормализация числовых признаков для устранения разницы в масштабах, что повышает стабильность и сходимость моделей машинного обучения (особенно важно для линейной регрессии). Также в нашем датасете было 4 категориальных переменных: Apartment type, Region, Renovation и Metro Station.

Первые два являлись “бинарными” признаками. поэтому просто преобразовали в 0 и 1

Для Renovation был использован метод One-hot encoding

Для Metro Station – частотное кодирование



Нормализация признаков

Преобразование числовых характеристик к единому масштабу

Кодирование категорий

Преобразование районов и других категориальных данных в числовые значения

Разделение данных

Создание обучающих и тестовых выборок для оценки моделей

Модели машинного обучения: теория и выбор

Перед практической реализацией представлены теоретические основы используемых моделей.

Случайный лес (Random Forest) применяет ансамблевый подход, улучшая стабильность и точность за счёт усреднения решений множества деревьев. XGBoost, как градиентный бустинг, мощно справляется с нелинейными зависимостями и обладает эффективным механизмом регуляризации для борьбы с переобучением.

Построение и оценка моделей

Модели реализованы и обучены на подготовленном датасете. Для оценки качества используются метрики, такие как среднеквадратичная ошибка (RMSE), коэффициент детерминации (R^2) и средняя абсолютная ошибка (MAE).

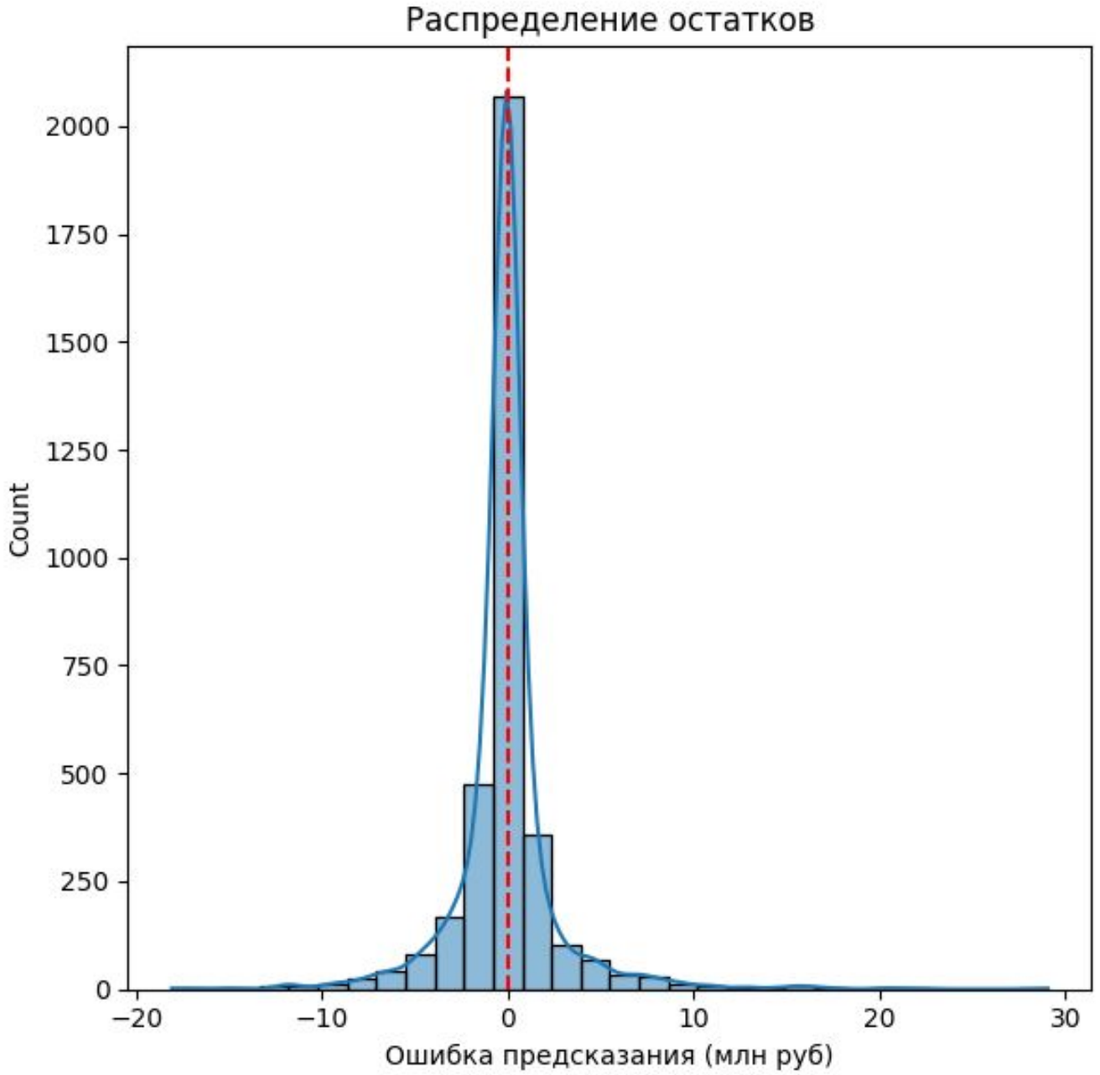
Сравнение результатов показывает, что модели ансамблей (Random Forest и XGBoost) достигают более высокой точности, особенно на тестовой выборке, демонстрируя устойчивость к шумам и способности улавливать сложные зависимости.

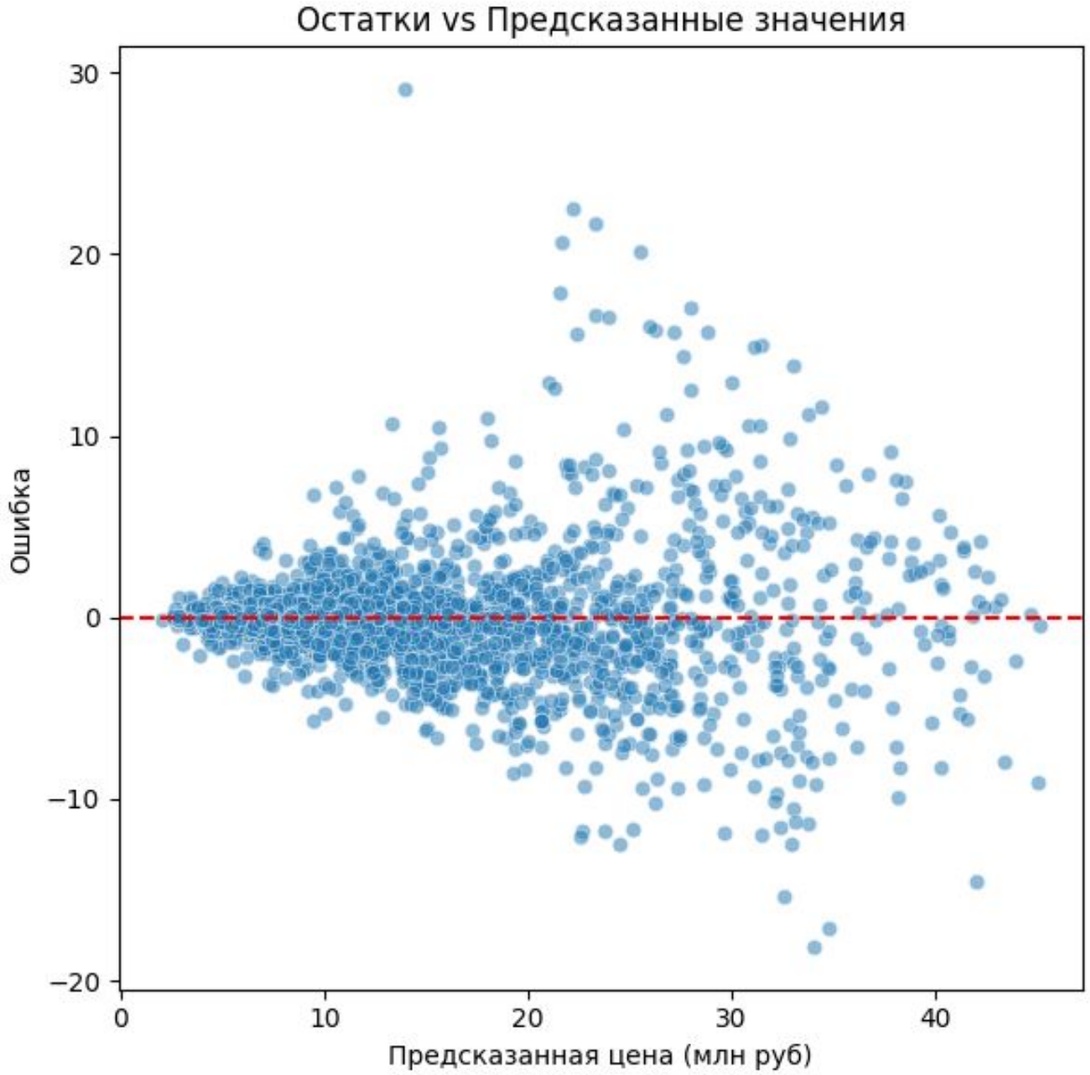
1. **Обучение модели**
   Использование обучающей выборки для подгонки параметров
2. **Тестирование**
   Оценка на отдельной тестовой выборке
3. **Сравнение метрик**
   Анализ RMSE, R^2, MAE для каждой модели
4. **Интерпретация результатов**
   Выявление сильных и слабых сторон моделей

Сравнение производительности:

Метод	R^2	MAE (млн руб)	RMSE (млн руб)
Linear	0.770	2.88	16.89
Random Forest	0.886	1.55	8.37
XGBoost	**0.897**	**1.45**	**7.58**

Анализ остатков

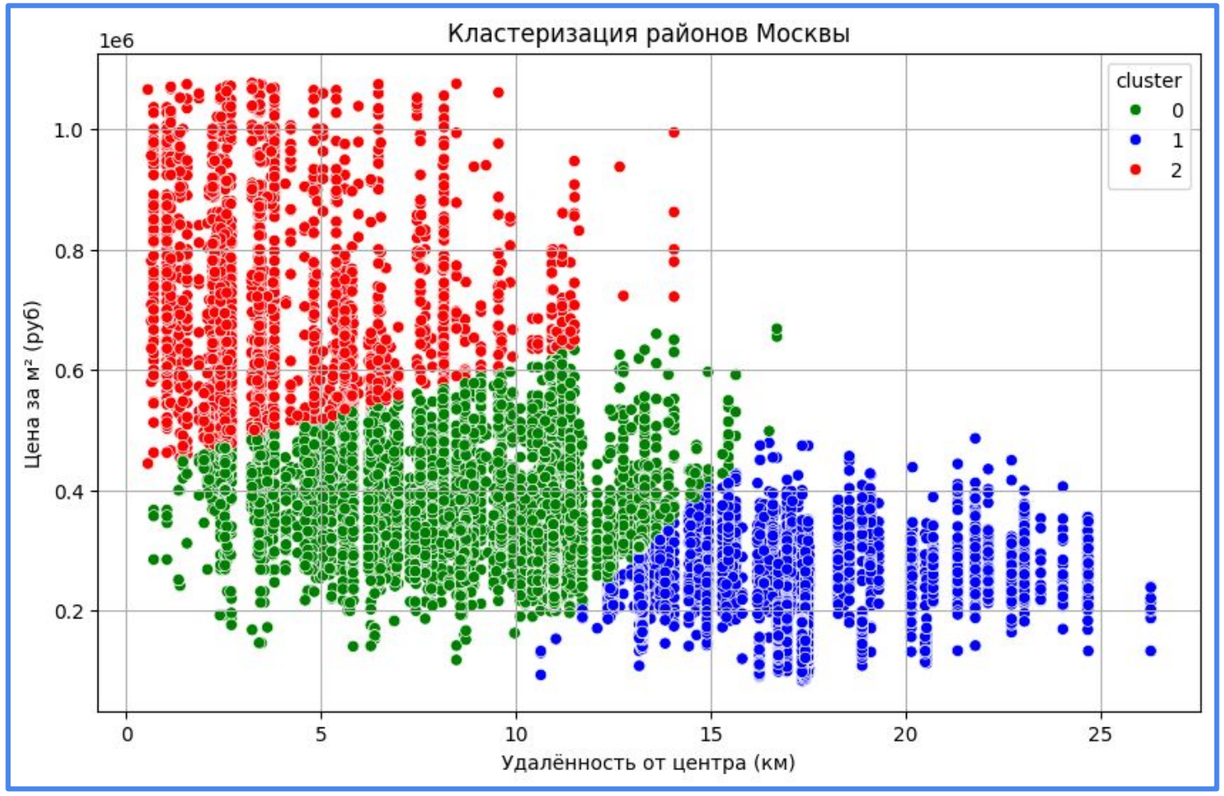




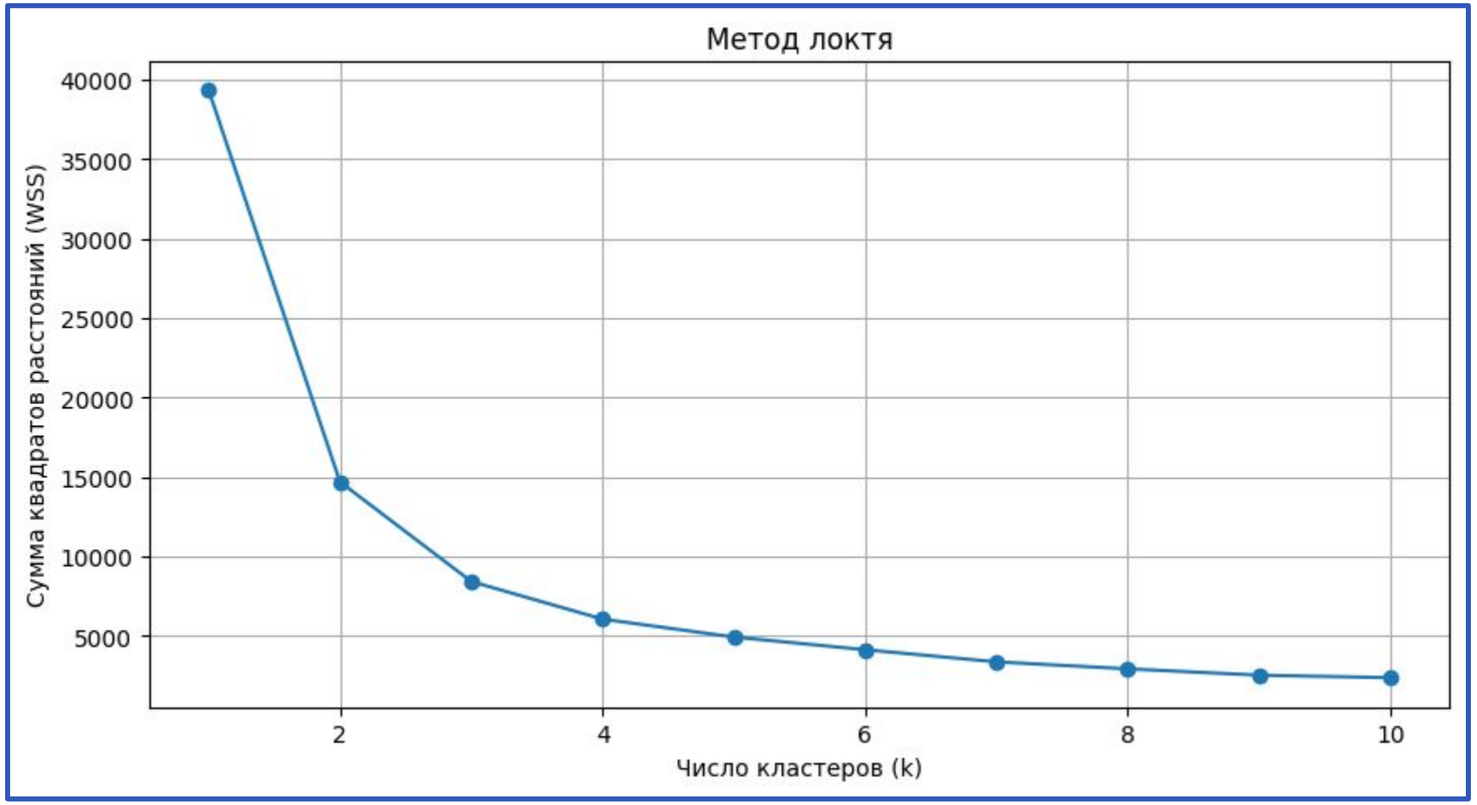
Кластеризация районов Москвы

Проанализируем стоимость квадратного метра жилья в Москве, используя кластеризацию методом K-means на основе признаков цены за м² и удалённости от центра. Прежде всего очистим наши данные и нормализуем их. Далее с помощью метода локтя определим оптимальное количество кластеров - 3. Визуализация результатов показывает чёткую зависимость цены за м² от расположения.

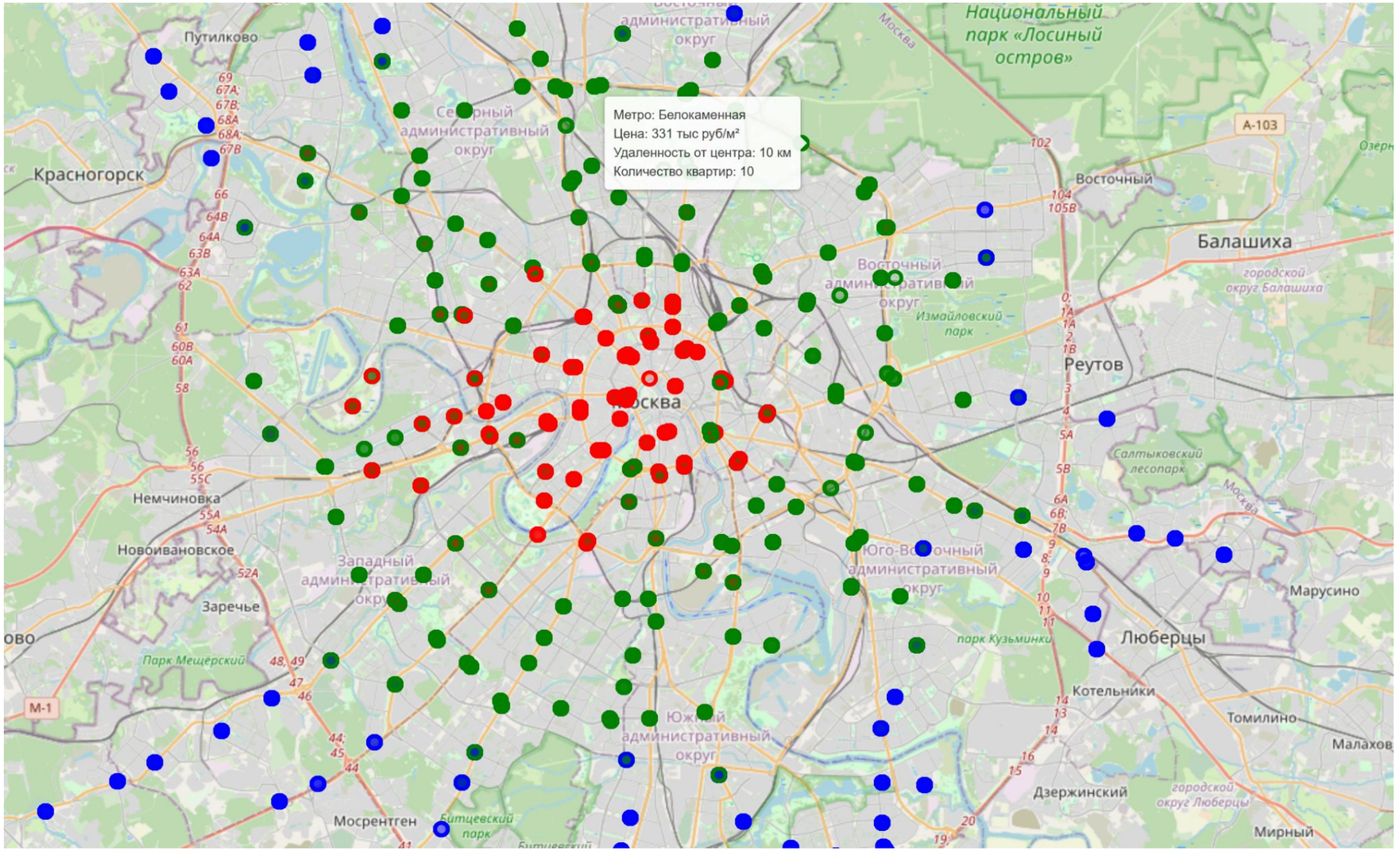
1. Нормализация данных
2. Выбор числа кластеров
3. Визуализация на карте
4. Анализ характеристик



Как мы определили оптимальное количество кластеров?



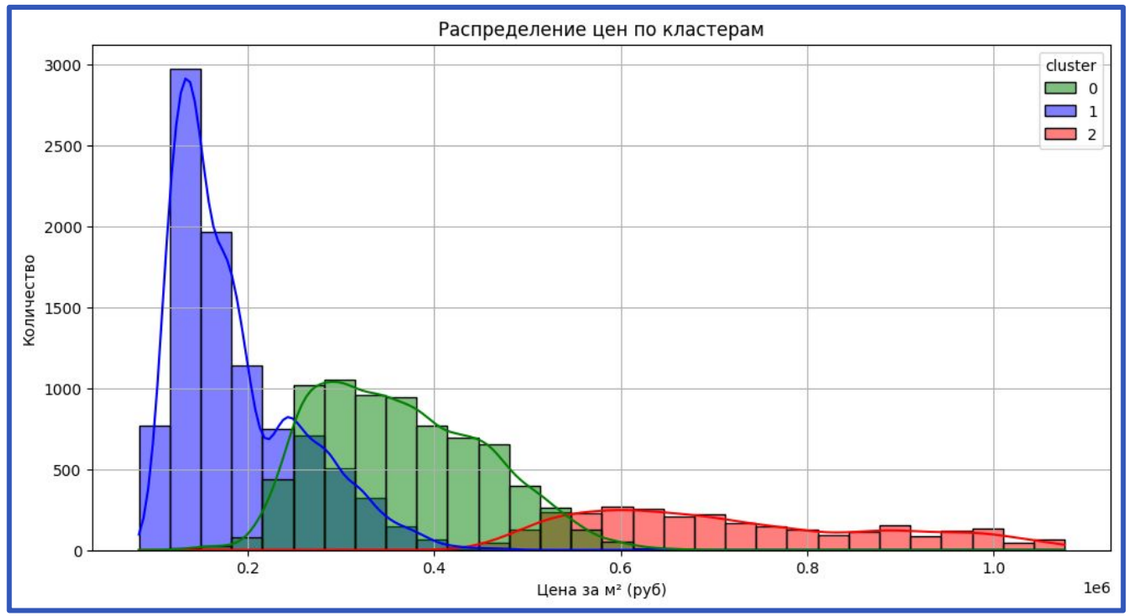
Интерактивная карта Москвы с визуализацией кластеров

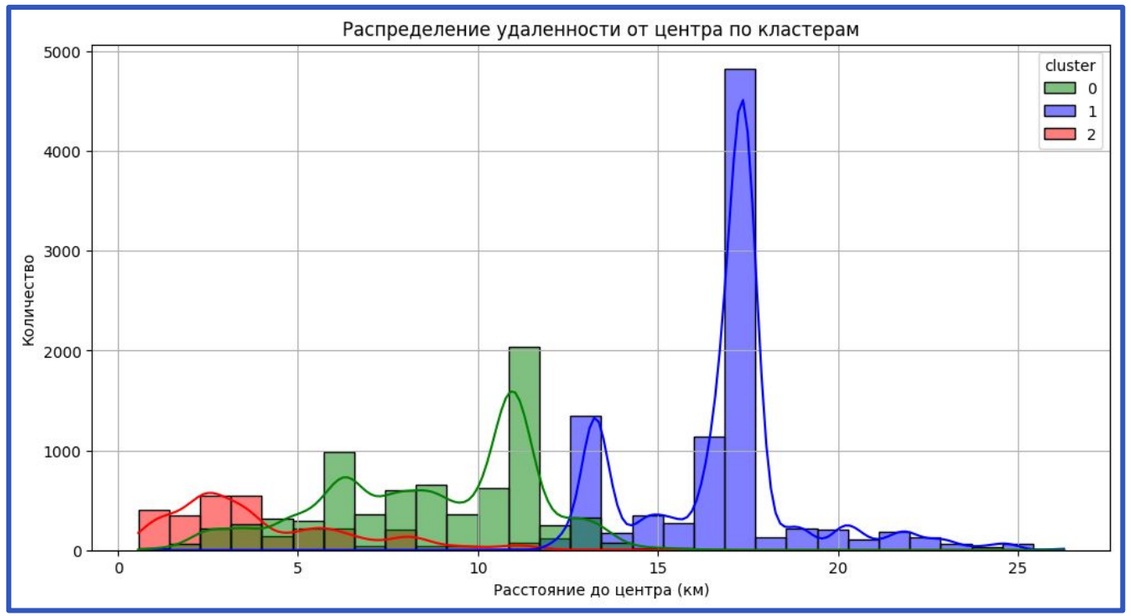


Характеристика кластеров и итоговые визуализации

На основе полученных характеристик кластеров мы можем описать каждый из них:

1. **0 кластер:** Средний ценовой сегмент, компромисс между удаленностью от центра и стоимостью квадратного метра, преобладающий тип жилья - вторичка
2. **1 кластер:** Дешевое жилье на окраинах в основном в новостройках, самый популярный кластер
3. **2 кластер:** Дорогие элитные квартиры в историческом центре города

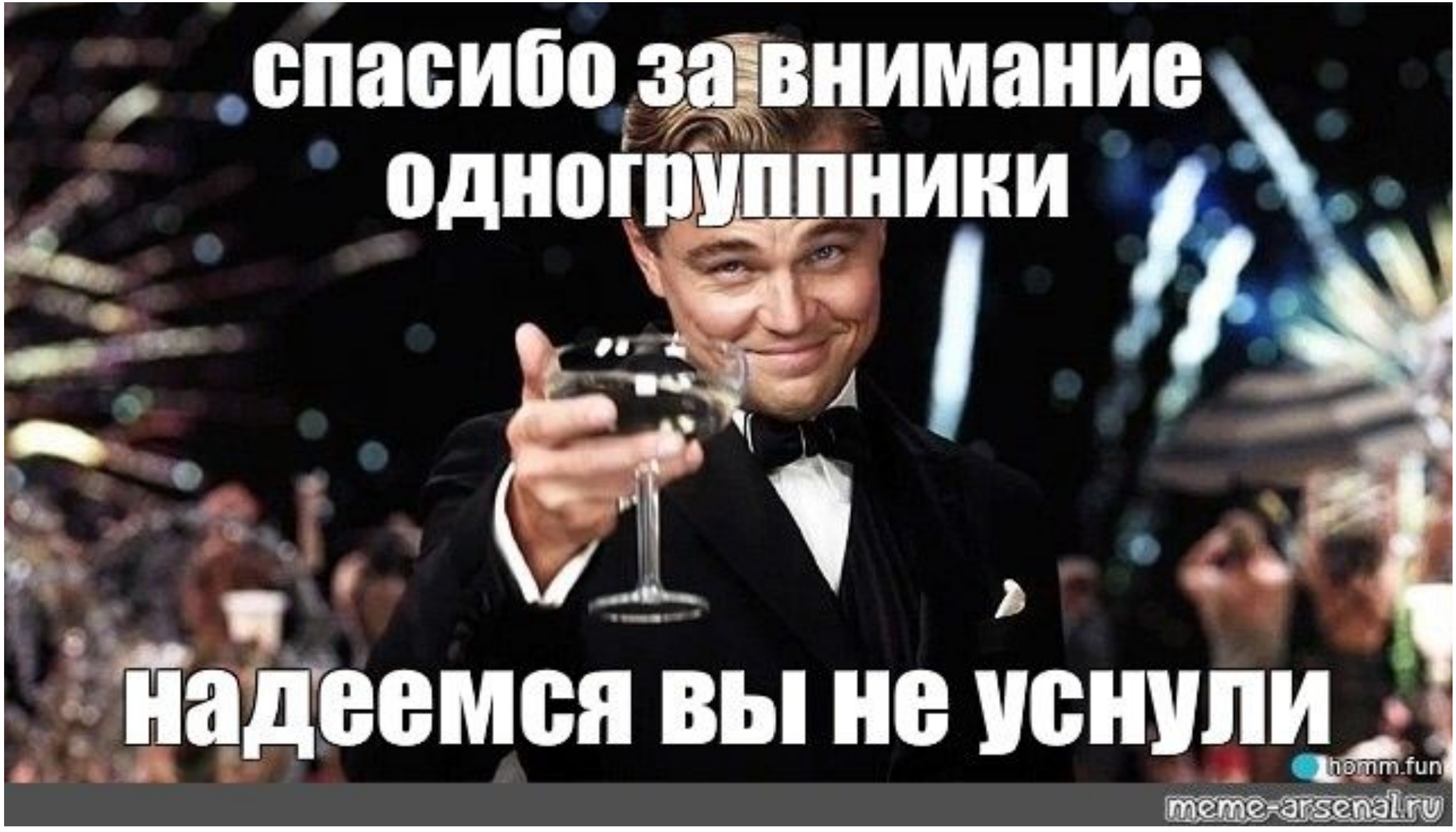




Выводы

Проект показал комплексный подход к анализу московского рынка жилья — от подробного разведочного анализа и предобработки до построения и сравнения моделей прогнозирования цен. Модель XGBoost показала лучший баланс между точностью и устойчивостью.

Кластеризация районов выявила четкие ценовые зоны с отличительными характеристиками, что позволяет применять результаты для более точного ценообразования и урбанистического планирования. Возможным направлением дальнейшей работы может стать интеграция дополнительных внешних данных и создание интерактивных дашбордов.


спасибо за внимание
одногруппники
надеемся вы не уснули
homm.fun
meme-arsenal.ru