Анализ и прогнозирование цен съёмного жилья Airbnb

Выполнили: Бавин Станислав, Васинков Никита, Сериков Александр.

332 группа

Введение

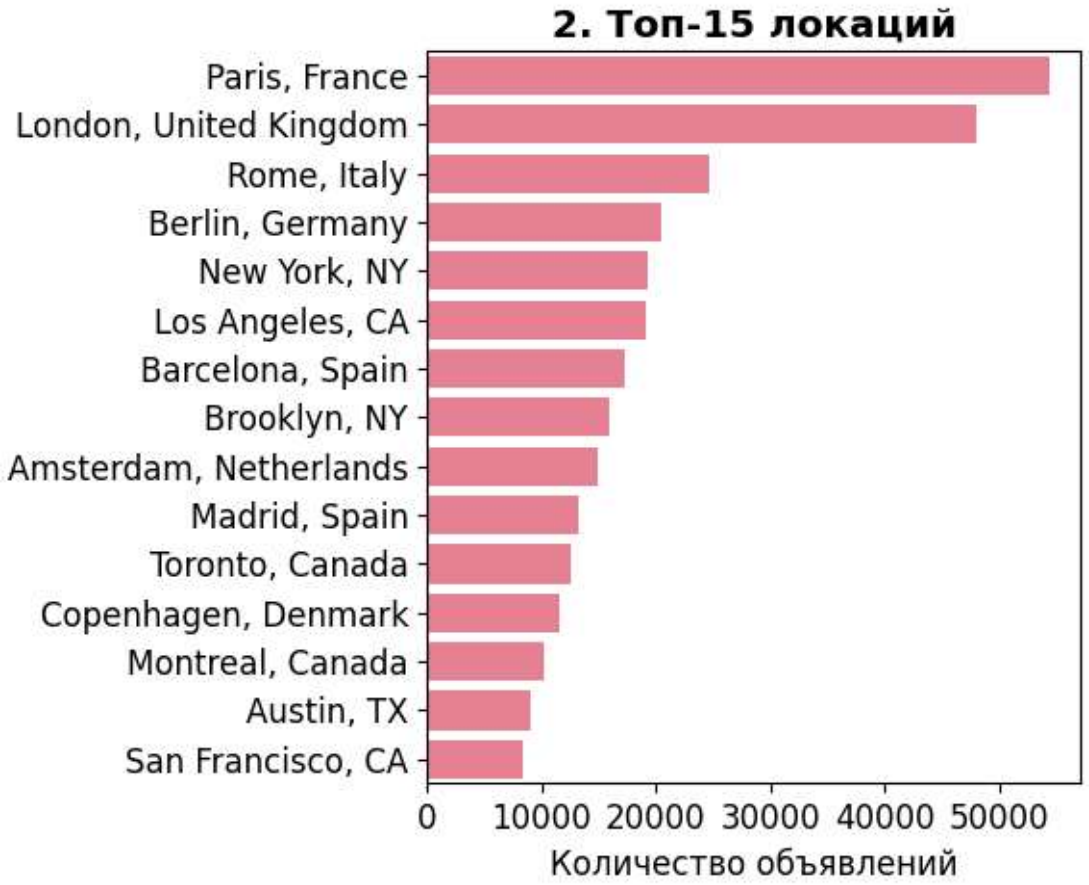
- В данном проекте проводится исследование датасета Airbnb с целью выявления факторов, влияющих на успешность размещений. Под успешностью понимается популярность объекта — например, высокая частота бронирований, большое количество отзывов или высокий средний рейтинг. Задача заключается в том, чтобы определить, какие характеристики объявления — такие как цена, расположение, тип жилья, наличие удобств или политика отмены — способствуют его привлекательности для гостей.

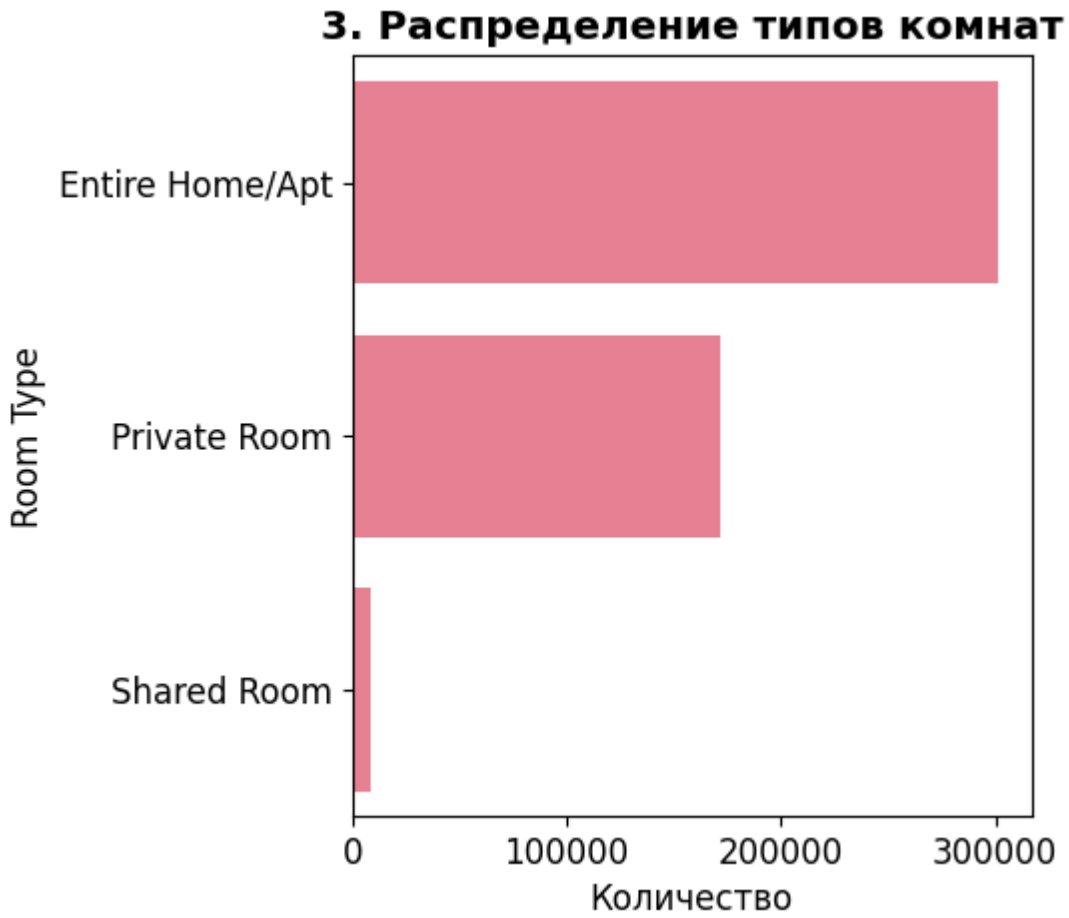
План исследования

- Исследование начинается с разведочного анализа данных (EDA), включающего оценку структуры датасета, выявление пропусков и выбросов, а также анализ распределения признаков и их взаимосвязей, чтобы сформулировать гипотезы о влиянии различных факторов на успешность жилья. Затем строятся модели машинного обучения (логистическая регрессия, kNN и случайный лес) для прогнозирования популярности объявлений на основе таких признаков, как цена, местоположение, отзывы и удобства. Ожидается, что исследование выявит ключевые факторы успеха, такие как разумная цена, положительные отзывы и удобное расположение, а также позволит построить эффективную модель прогнозирования, подтвержденную наглядными визуализациями зависимостей и важности признаков.

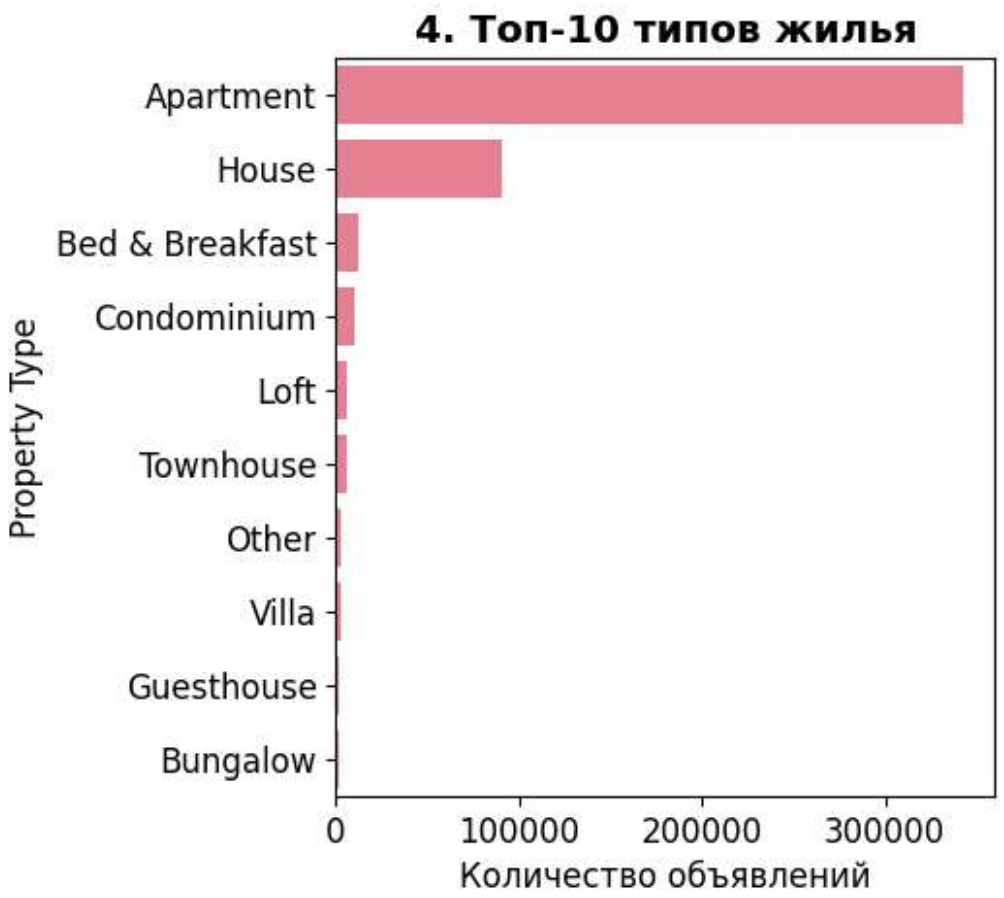
Визуализация ключевых признаков

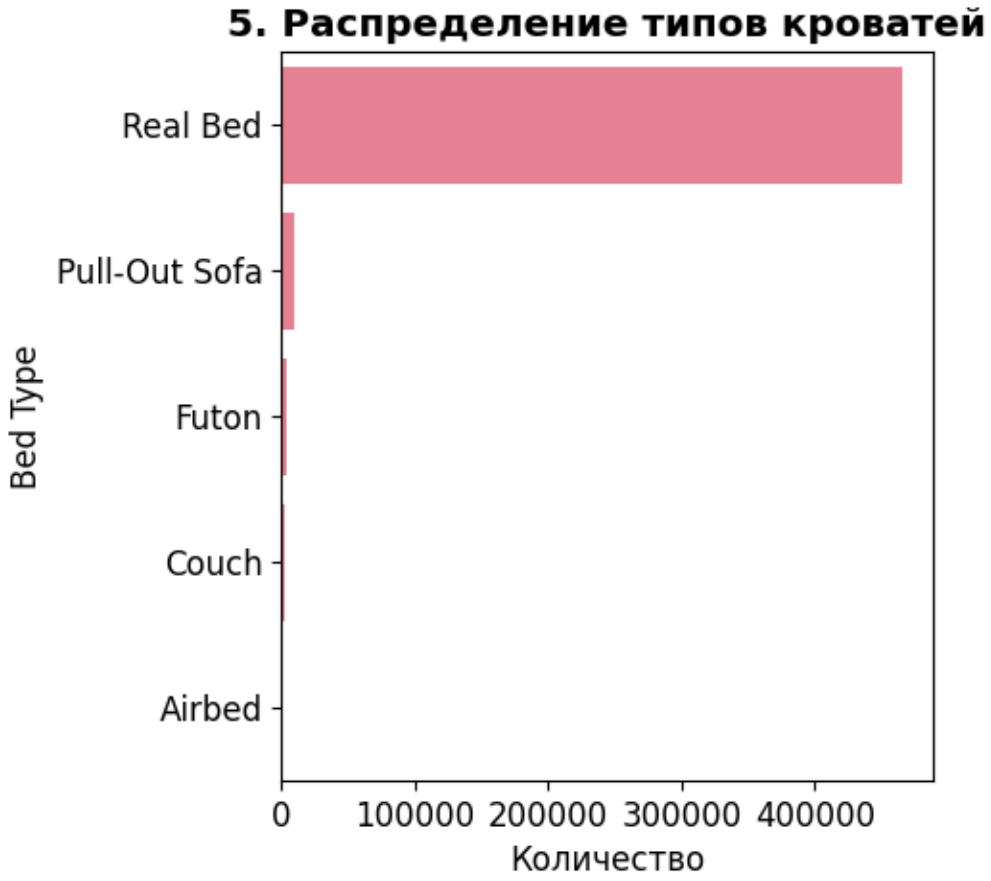
Визуализация ключевых признаков



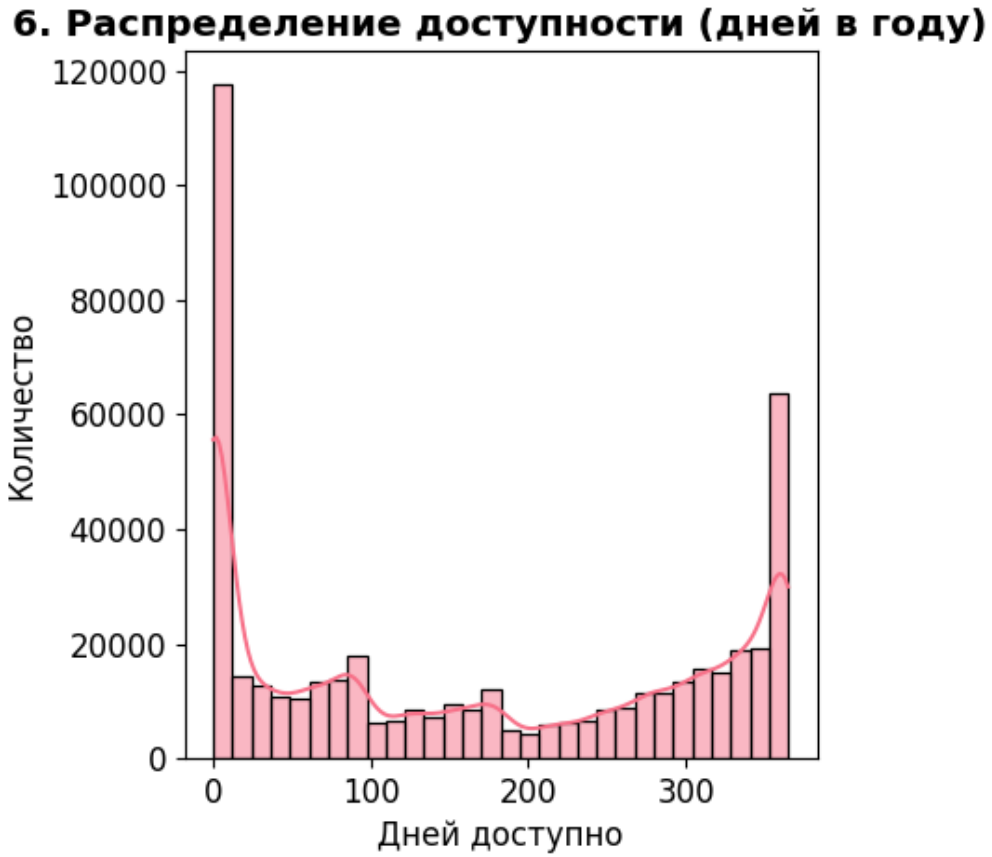


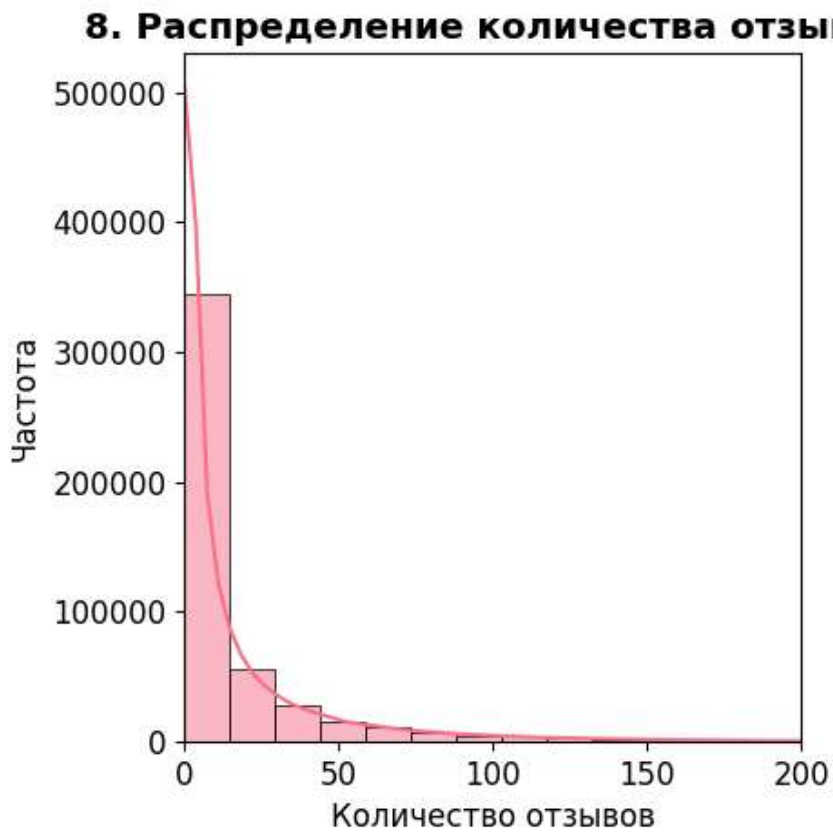
Визуализация ключевых признаков



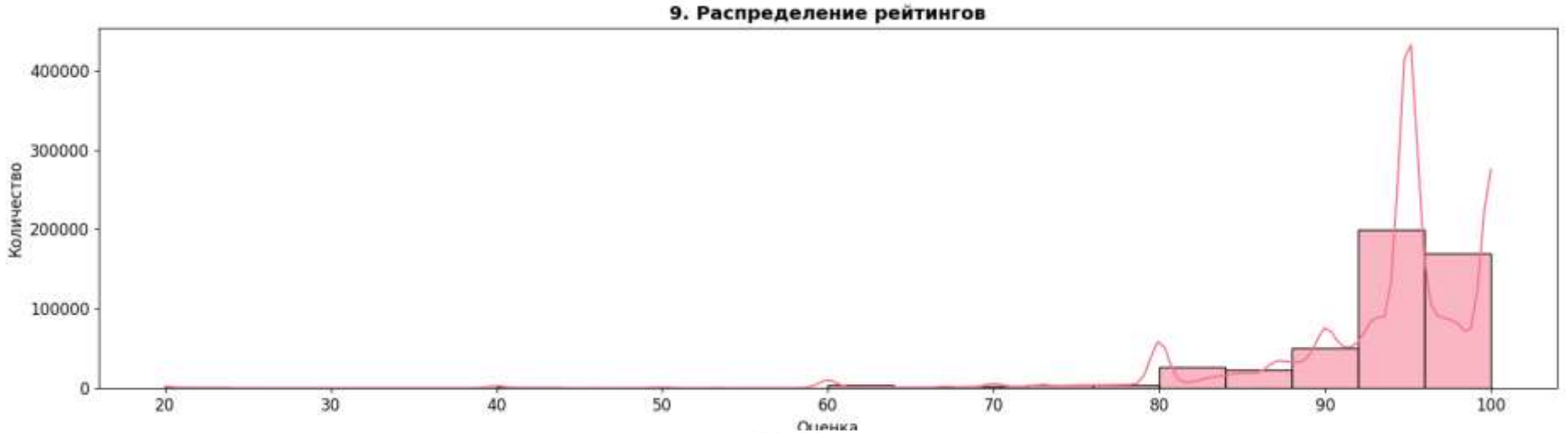


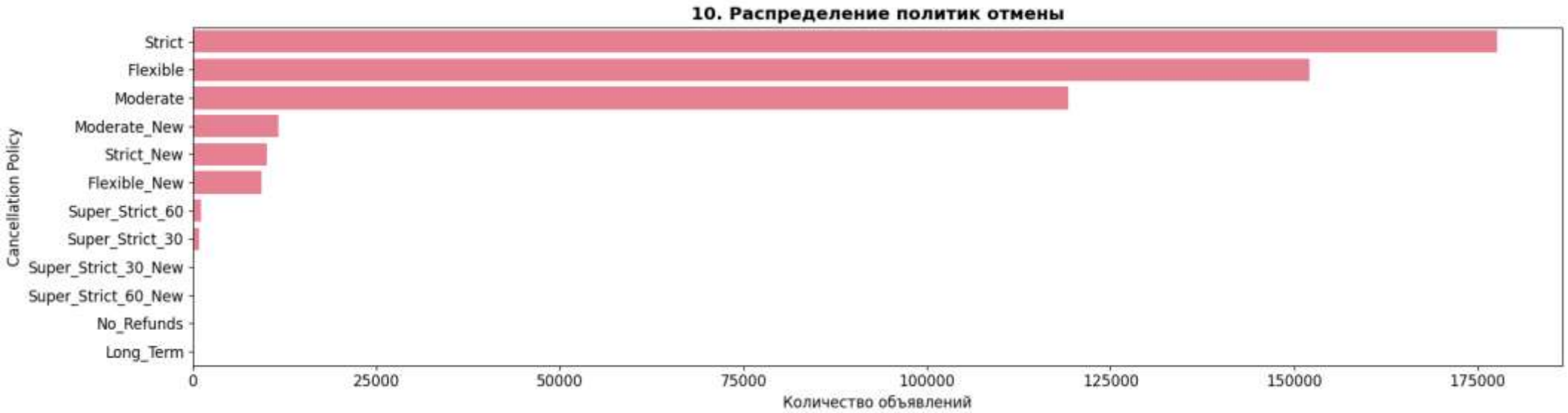
Визуализация ключевых признаков



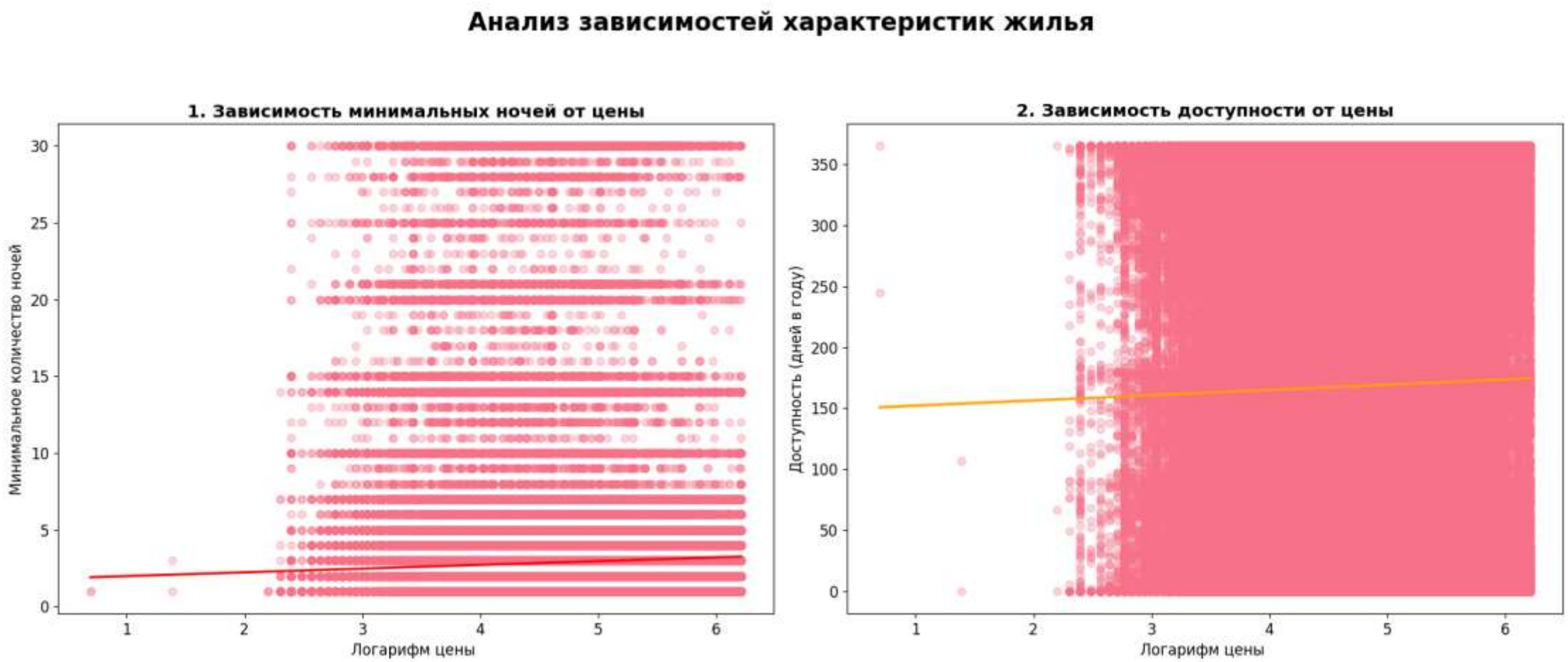


Визуализация ключевых

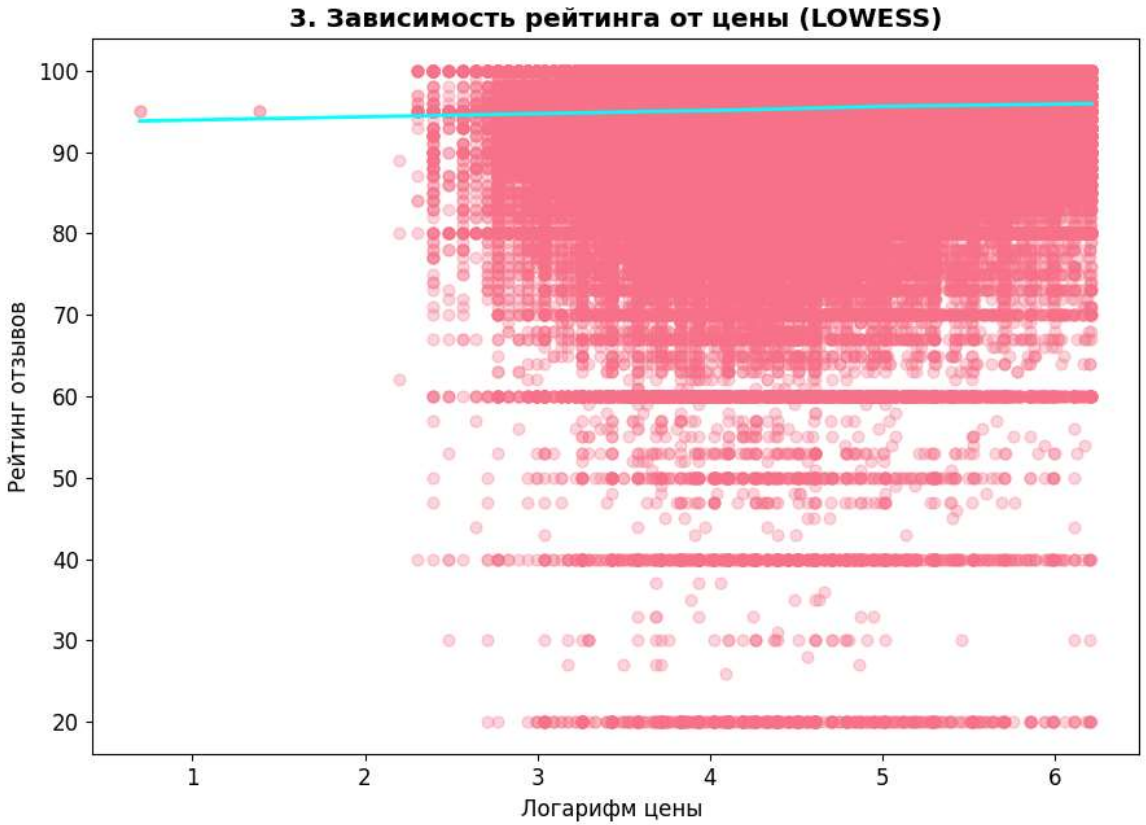


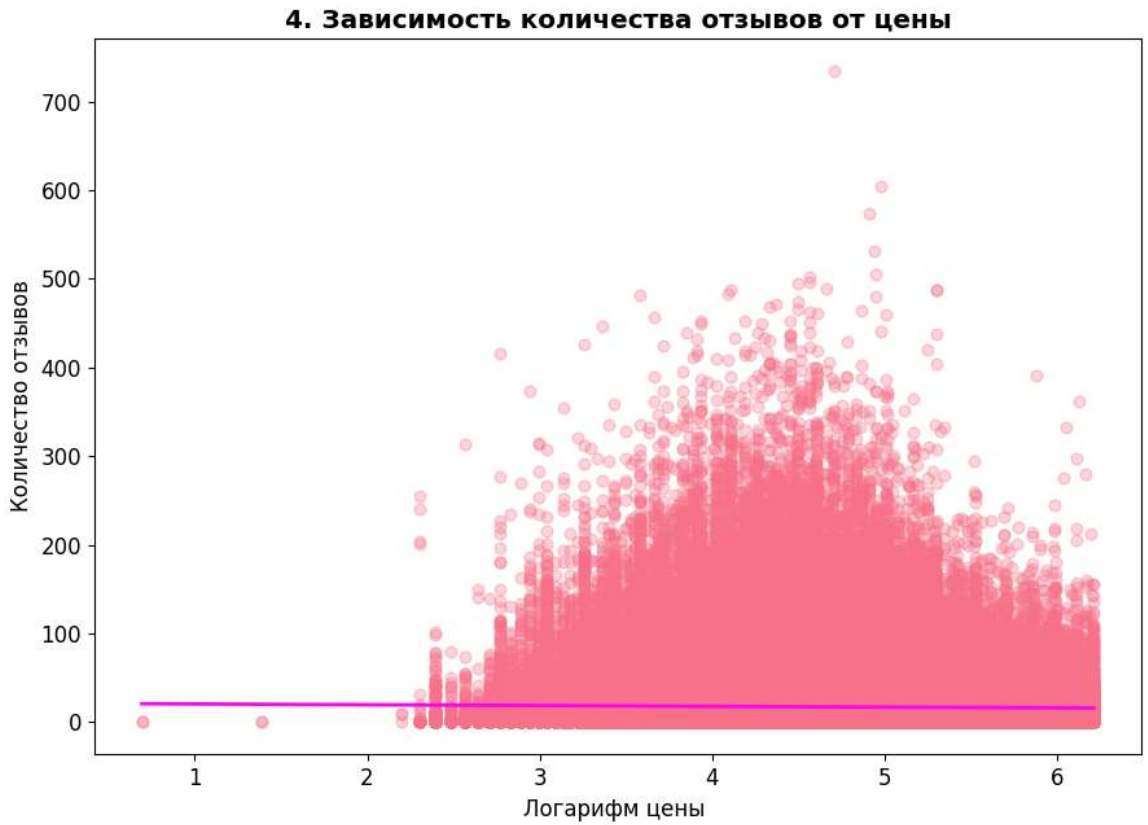


Анализ зависимостей характеристик жилья

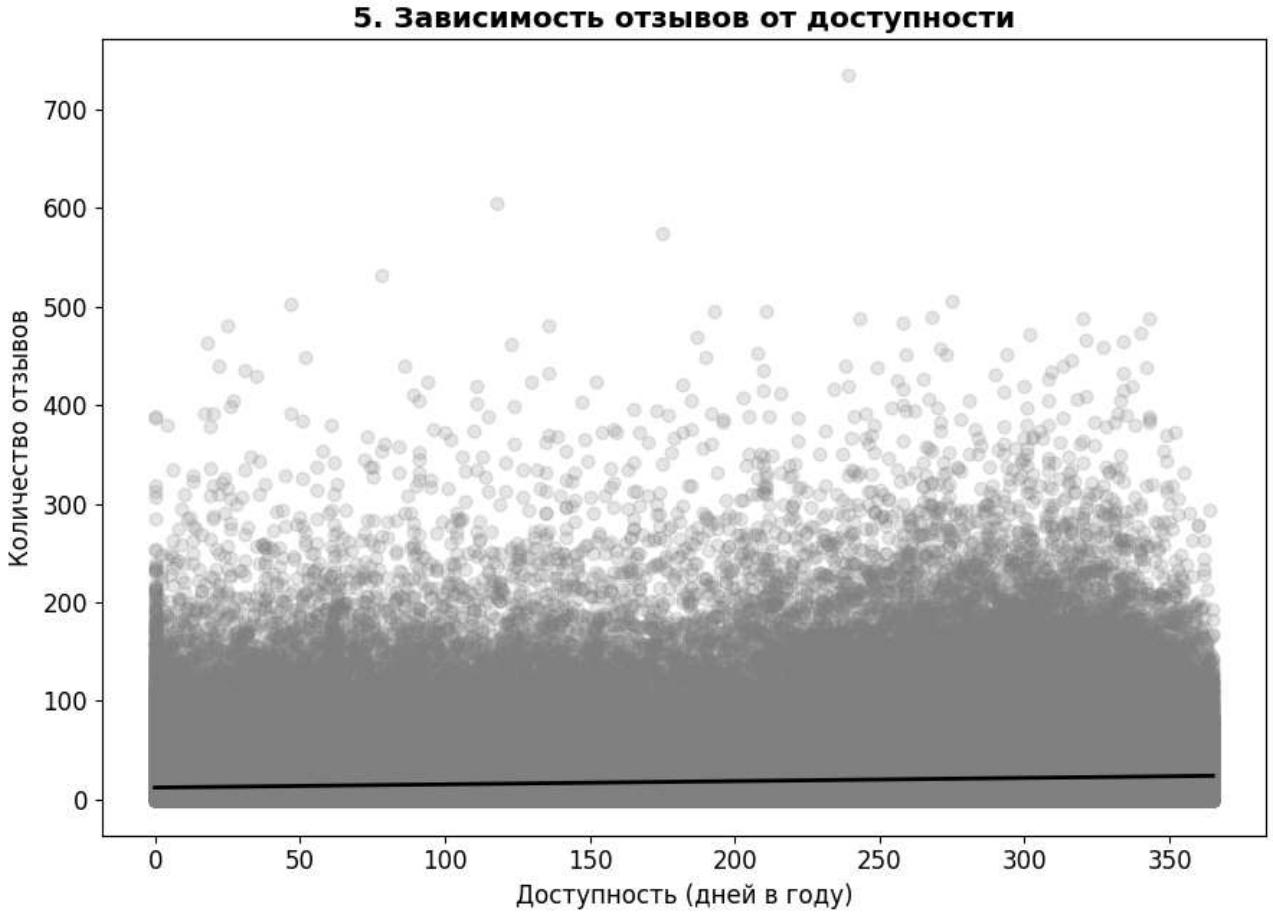


Анализ зависимостей характеристик жилья





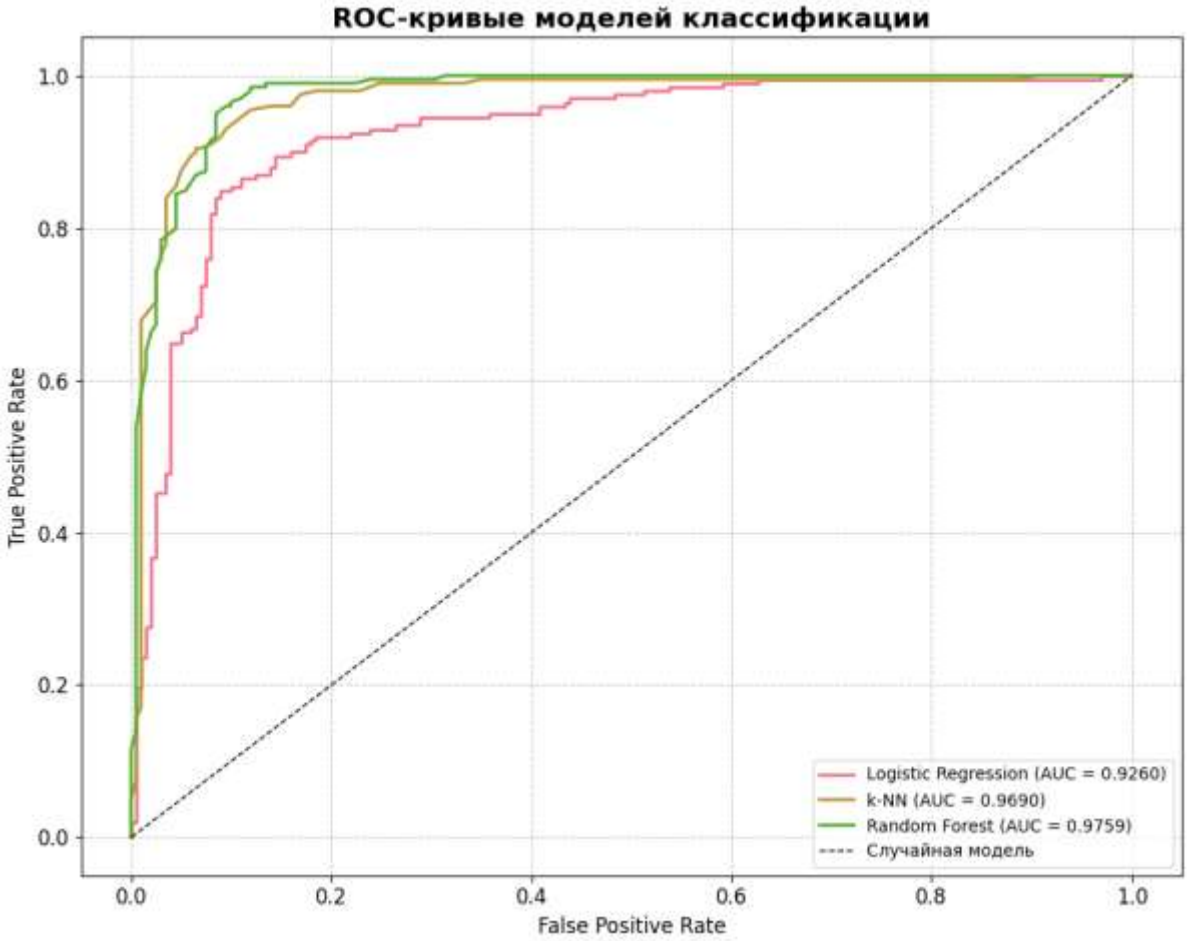
Анализ зависимостей характеристик жилья



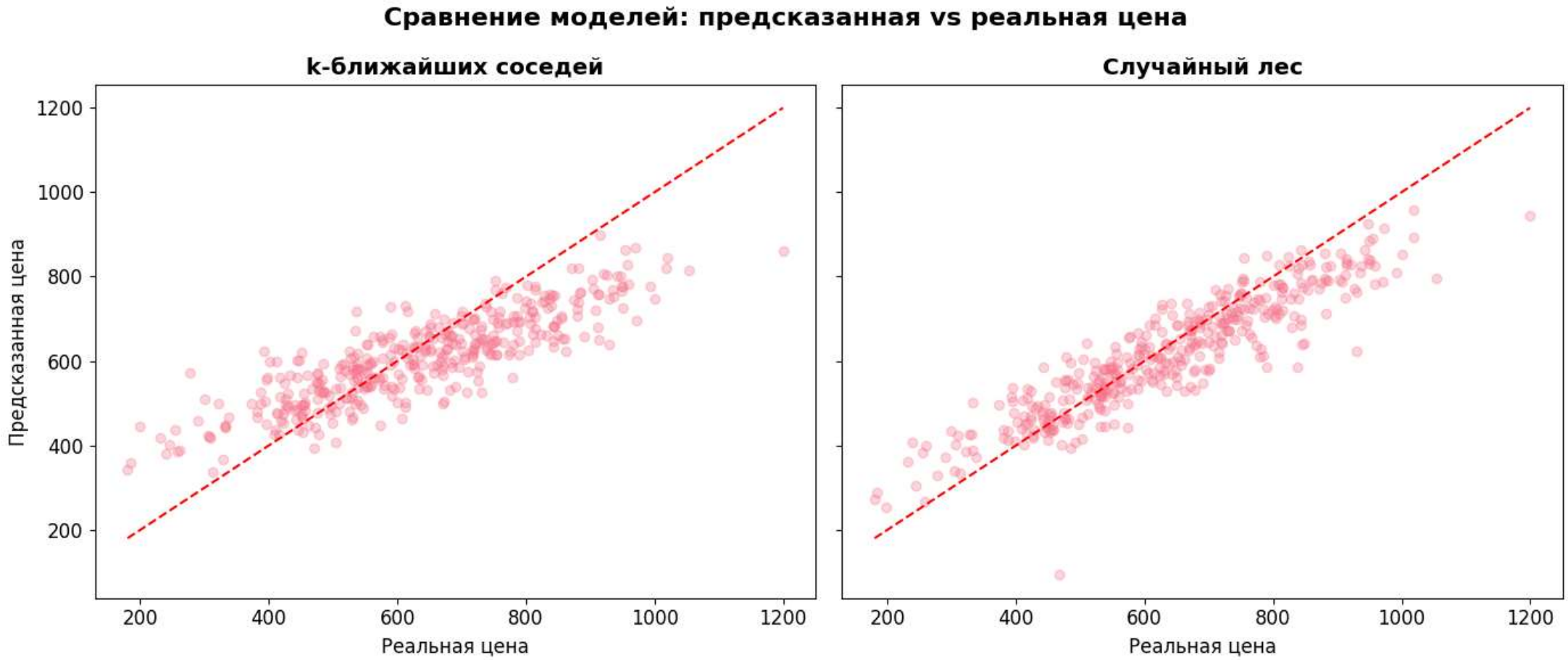
Модели машинного обучения: теория.

- Логистическая регрессия является линейной моделью классификации, которая оценивает вероятность принадлежности объекта к определённому классу с помощью сигмоидной функции, преобразующей линейную комбинацию признаков в значение от 0 до 1; её преимущества — интерпретируемость и простота, но она плохо справляется с нелинейными зависимостями.
- Метод kNN относится к непараметрическим алгоритмам и работает по принципу «подобное к подобному»: класс нового объекта определяется большинством голосов среди k ближайших к нему обучающих примеров, что делает модель гибкой, но чувствительной к шумам и требующей тщательного подбора расстояния и числа соседей.
- Случайный лес представляет собой ансамблевый метод, строящий множество решающих деревьев на случайных подвыборках данных и признаков, а итоговый прогноз формируется путём голосования всех деревьев.

Roc кривые выбранных моделей



Предсказанная и реальная цена в выбранных моделях.



Выводы

- Успешные объявления характеризуются доступной ценой, хорошими отзывами и высоким рейтингом.
- Расположение, тип жилья и политика отмены также влияют на привлекательность.
- Лучшие модели прогнозирования: случайный лес (AUC = 0.9759).