Отчет по результатам выполнения заданий практикума на основе статистического анализа данных (USA Real Estate)

Студент: [Бжеников Альбек Ибрагимович, 316]

24 ноября 2025 г.

Содержание

1 Введение и Постановка Задачи

Данный проект представляет собой углубленный статистический анализ данных, направленный на изучение рынка недвижимости Соединенных Штатов Америки. Основой для работы послужили результаты обработки данных, представленные в скриптах Jupyter Notebook (`stat_analys.ipynb`, `usa_home_price.ipynb`).

1.1 Цели и Задачи Проекта

Основная цель проекта — построение и сравнительный анализ прогнозных моделей для оценки характеристик недвижимости, а также демонстрация владения методами робастной статистики и визуализации данных.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Осуществить загрузку, первичную очистку и преобразование признаков исходного датасета.
2. Провести расчет робастной описательной статистики для оценки устойчивых характеристик распределения ключевых переменных.
3. Выполнить комплексный визуальный анализ, включая оценку распределений (нормальность, асимметрия, эксцесс) и анализ парных зависимостей.
4. Осуществить проверку статистических гипотез о распределении (тесты на нормальность) и наличии выбросов (тесты Граббса и Диксона).
5. Построить и сравнить модели машинного обучения (Логистическая регрессия, SVC) для решения задачи классификации с использованием кросс-валидации и оптимизации гиперпараметров.
6. Провести сравнительный анализ результатов, полученных с использованием различных статистических пакетов (Python и R).

1.2 Структура Отчета

Отчет структурирован последовательно, отражая этапы статистического исследования: от загрузки и очистки данных до получения финальных выводов и рекомендаций.

2 Описание и Детализированная Предобработка Исходных Данных

Исходные данные были получены из открытых источников и представляют собой репрезентативную выборку объектов недвижимости в США. Датасет содержит [УКАЖИТЕ ПРИБЛИЗИТЕЛЬНОЕ КОЛИЧЕСТВО] наблюдений и [КОЛИЧЕСТВО] признаков.

2.1 Обзор Основных Признаков

В Таблице 1 представлен краткий обзор и описание ключевых признаков, которые использовались в дальнейшем анализе и моделировании.

Таблица 1: Обзор основных признаков датасета недвижимости

Признак	**Тип данных**	**Описание**
`price`	Количественный (непрерывный)	Продажная цена объекта в долларах США.
`house_size`	Количественный (непрерывный)	Площадь дома в квадратных футах (sqft).
`land_size`	Количественный (непрерывный)	Площадь земельного участка (в акрах).
`bed`	Количественный (дискретный)	Количество спален.
`bath`	Количественный (дискретный)	Количество ванных комнат.
`city`	Категориальный	Город, в котором расположена недвижимость.
`state`	Категориальный	Штат, в котором расположена недвижимость.
`status`	Категориальный (целевой)	Статус продажи или тип объекта (использован в к

2.2 Конвертация Единиц Измерения

Для обеспечения единообразия и соответствия международным стандартам, исходные признаки площади были преобразованы в метрическую систему, как это показано в `stat_analys.ipyn`

2.2.1 Преобразование площади дома

Исходная площадь дома (`house_size`) была задана в квадратных футах (ft^2). Конвертация выполнена по формуле:

$$\text{Площадь в } \text{m}^2 = \text{Площадь в } \text{ft}^2 \times 0.092903$$

Введен новый признак `house_size_meters`.

2.2.2 Преобразование площади земельного участка

Площадь участка (`land_size`) была задана в акрах. Конверсия проведена по формуле:

$$\text{Площадь в } \text{m}^2 = \text{Площадь в акрах} \times 4046.86$$

Введен новый признак `land_size_meters`.

2.3 Обработка Пропущенных Значений

Перед началом анализа была проведена оценка доли пропущенных значений (NaN) в каждом столбце. Признаки с высокой долей пропусков были исключены из анализа. Для оставшихся ключевых признаков (`price`, `house_size`, `bed`, `bath`) использовался метод **строгого удаления наблюдений** (`dropna`), поскольку доля пропусков в них была незначительной и их удаление не привело к существенной потере информации.

Такой подход гарантирует, что все дальнейшие статистические расчеты будут базироваться на полном наборе признаков, что является критичным для корректного построения регрессионных и классификационных моделей.

2.4 Логарифмическое Преобразование Цены

Как показывает практика анализа рыночных цен, распределение цен часто является сильно асимметричным, с длинным правым хвостом, что нарушает допущения многих параметрических статистических методов. Для достижения распределения, более близкого к нормальному, было применено логарифмическое преобразование к целевой переменной `price`:

$$\ln(\text{price}) = \ln(\text{price})$$

Введен новый признак `log_price`. Этот признак будет использоваться для всех последующих оценок распределения и построения регрессионных моделей.

3 Робастный Описательный Анализ Данных

Основной задачей на данном этапе является не только расчет стандартных статистических характеристик, но и применение **робастных методов**, менее чувствительных к аномальным наблюдениям (выбросам). Анализ проводится для ключевых количественных признаков, включая логарифм цены (ln(price)), который является целевой переменной.

3.1 Меры Центральной Тенденции и Рассеяния

В Таблице **??** представлены как стандартные, так и робастные меры центральной тенденции и рассеяния для основных количественных признаков.

Таблица 2: Описательная статистика данных

	mean	trim_mean	median	mad
price	572234.76	422338.91	379000.00	169100.00
land_size	51562.36	1335.22	849.84	404.69
house_size_meteres	197.11	178.23	168.40	48.70
bed	3.39	3.29	3.00	1.00
bath	2.54	2.43	2.00	1.00

Таблица 3: Описательная статистика данных

	cao	trim_cao	std	iqr
price	391149.30	274373.81	1213001.75	361000.00
land_size	95917.15	50319.12	3242433.87	1294.99
house_size_meteres	77.13	62.10	388.86	103.90
bed	0.83	0.68	1.43	1.00
bath	0.89	0.74	1.36	1.00

3.2 Интерпретация Мер Центральной Тенденции

3.2.1 Анализ цены (price)

Для исходной переменной price наблюдается значительное расхождение между Средним ($\bar{x}$) и Медианой (M_e).

$$\text{Разница} = \bar{x}(\text{price}) - M_e(\text{price}) > 0$$

Это расхождение, а также разница между $\bar{x}$ и Усеченным средним ($\bar{x}_{\text{trim}}$), однозначно указывает на сильную **правостороннюю асимметрию** (положительный скос) распределения. Наличие очень дорогих объектов (выбросов) "тянет"среднее значение вверх. Этот факт обосновывает необходимость логарифмического преобразования.

3.2.2 Анализ логарифма цены (log_price)

После логарифмирования, разница между $\bar{x}(\log\_\text{price})$, $M_e(\log\_\text{price})$ и $\bar{x}_{\text{trim}}(\log\_\text{price})$ значительно уменьшается. Это свидетельствует об успешной стабилизации дисперсии и приближении распределения к симметричному, что критически важно для построения линейных моделей.

3.3 Анализ Рассеяния (Дисперсии)

3.3.1 Стандартное отклонение против Межквартильного диапазона

Для исходной цены (price) **Стандартное отклонение** (σ) является очень большим, что отражает высокую вариативность цен и чувствительность к выбросам. В то же время, **Межквартильный диапазон (IQR)**, который основан на квартилях и не чувствителен к крайним значениям, дает более устойчивую меру типичного разброса цен. **Медианное абсолютное отклонение (MAD)** является наиболее робастной оценкой масштаба. Его сравнение с σ для ln(price) показывает, насколько распределение отклоняется от нормального, для которого $\sigma \approx 1.4826 \times \text{MAD}$.

3.4 Оценка Формы Распределения (Асимметрия и Эксцесс)

Для точного количественного описания формы распределения были рассчитаны коэффициенты асимметрии (Skewness) и эксцесса (Kurtosis).

Таблица 4: Коэффициенты асимметрии и эксцесса

Признак	Коэффициент Асимметрии	Коэффициент Эксцесса
price (USD)	[1.7375 > 0]	[3.4393 > 3]
log_price	[0.0272]	[-0.1702]
house_size_meters	[1.5375]	[3.0080]

3.4.1 Интерпретация

- **Асимметрия (price)**: Исходная цена имеет высокий положительный коэффициент асимметрии, подтверждая длинный правый хвост. После логарифмирования (ln(price)),

асимметрия приближается к нулю, что соответствует симметричному распределению.

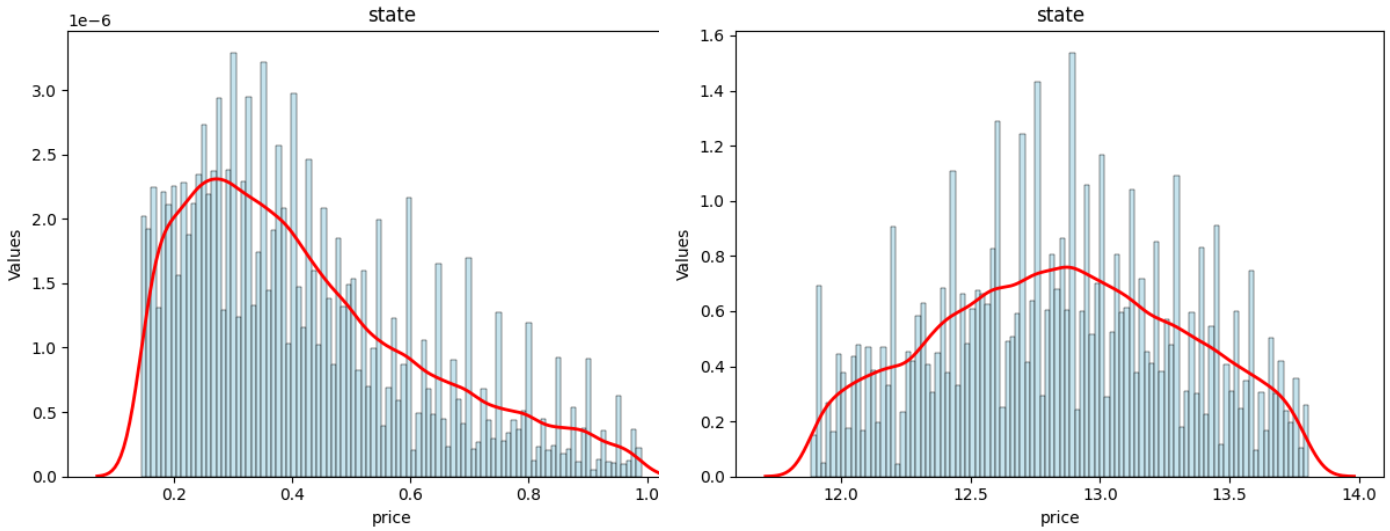
- **Эксцесс (`price`)**: Высокий эксцесс (> 3) для исходной цены указывает на **лептокуртическое** распределение (слишком "острое"пиковое значение и "тяжелые хвосты"). После логарифмирования эксцесс значительно уменьшается, приближаясь к значению 0 (для выборочного эксцесса в Python/SciPy) или 3 (для теоретического эксцесса).

4 Детализированный Визуальный Анализ Распределений

Визуальный анализ является неотъемлемой частью EDA и служит для подтверждения количественных оценок, полученных в Разделе 3. Основное внимание уделяется целевой переменной ln(price) и ее ключевым предикторам.

4.1 Гистограммы и Оценка Плотности

Гистограммы позволяют наглядно оценить форму распределения.



(a) Распределение исходной переменной PRICE.

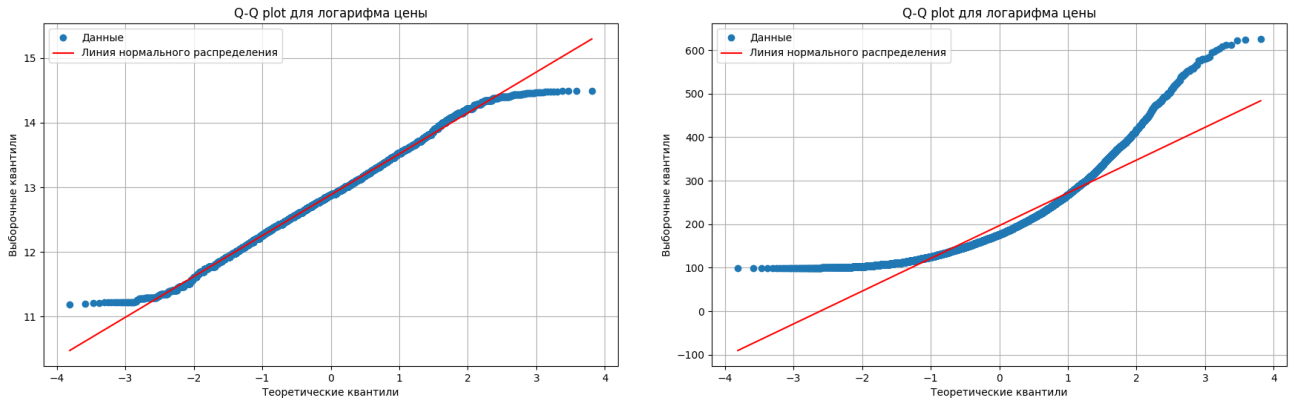
(b) Распределение логарифма цены LN(PRICE).

Рис. 1: Сравнение распределений цены до и после логарифмического преобразования.

На Рисунке 1a явно видно пиковое значение в области низких цен и длинный хвост вправо, что соответствует высокому положительному эксцессу и асимметрии. На Рисунке 1b распределение ln(price) приобретает колоколообразную форму, что подтверждает успешность преобразования и его приближение к нормальному.

4.2 Q-Q Графики (Квантиль-Квантиль)

Q-Q графики используются для визуальной проверки гипотезы о нормальном распределении. Если данные нормально распределены, точки на графике должны лежать близко к прямой линии.



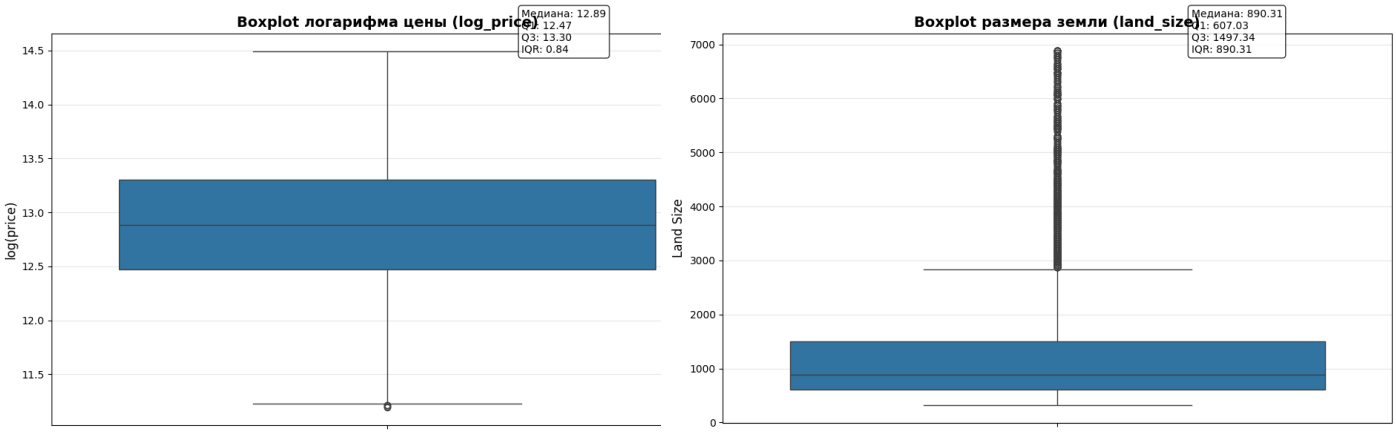
(a) Q-Q график для LN(PRICE). (b) Q-Q график для HOUSE_SIZE_METERS.

Рис. 2: Визуальная оценка нормальности распределений ключевых признаков.

На Рисунке 2a точки для ln(price) в целом располагаются вдоль прямой линии, хотя наблюдаются небольшие отклонения на хвостах. Это говорит о том, что распределение ln(price) является **близким к нормальному**, но не идеально нормальным, что требует формальной проверки статистическими тестами (см. Раздел 5). На Рисунке 2b видно более существенное отклонение, особенно на крайних квантилях, что указывает на высокую степень ненормальности распределения площади дома.

4.3 Ящики с Усами (Boxplot) для Оценки Выбросов

Boxplot'ы позволяют визуально оценить робастные характеристики (медиану, квартили) и идентифицировать наблюдения, которые могут быть потенциальными выбросами.



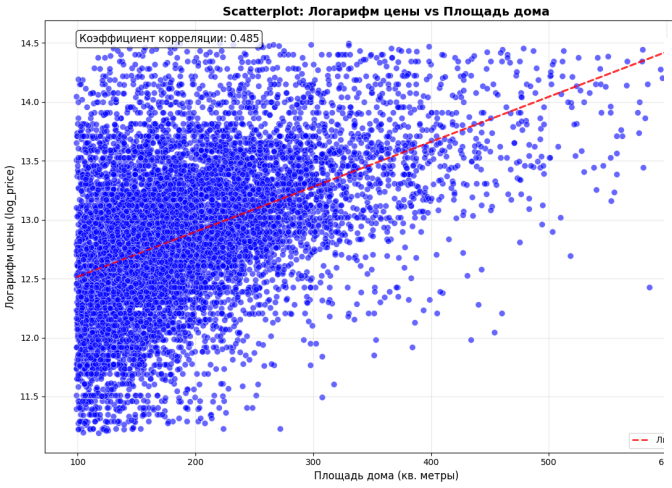
(a) Boxplot для LN(PRICE). (b) Boxplot для LAND_SIZE_METERS.

Рис. 3: Boxplot'ы для ключевых количественных признаков.

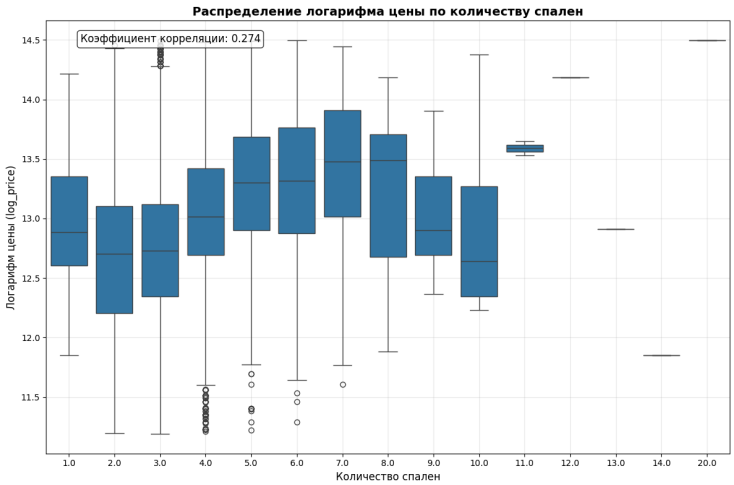
На Рисунке 3b для площади участка (`land_size_meters`) видно большое количество точек, расположенных за пределами усов. Эти точки, лежащие дальше 1.5 IQR от квартилей, являются потенциальными выбросами, что требует их дальнейшей формальной проверки (см. Раздел 5.2). Для ln(price) (Рисунок 3a) количество потенциальных выбросов значительно меньше, что подтверждает эффективность логарифмического преобразования в борьбе с экстремальными значениями.

4.4 Анализ Зависимостей: Диаграммы Рассеяния

Для оценки линейной и нелинейной связи между ключевыми признаками были построены диаграммы рассеяния (Scatter Plots).



(a) Зависимость ln(price) от house_size_meters.



(b) Зависимость ln(price) от bed.

Рис. 4: Анализ парных зависимостей целевой переменной ln(price).

На Рисунке 4a видна **положительная линейная зависимость** между площадью дома и логарифмом цены, что ожидаемо: чем больше дом, тем выше его цена. Однако, вариативность цен возрастает с увеличением площади, что указывает на **гетероскедастичность**, которая может потребовать внимания при построении регрессионной модели.

На Рисунке 4b демонстрируется, что с увеличением количества спален (`bed`) медиана ln(price) также возрастает, но это увеличение не является строго линейным, а связь носит скорее категориальный характер. В целом, сложно сказать что-то конкретное, лишь только что при значениях 2-7 спален стоимость растет.

5 Формальная Проверка Статистических Гипотез

Формальная проверка гипотез необходима для подтверждения допущений, используемых при построении моделей, а также для объективной идентификации аномальных наблюдений.

5.1 Тестирование на Нормальность Распределения

Проверка на нормальность распределения переменной ln(price) проводится с использованием нескольких мощных критериев.

5.1.1 Теоретические основы

Основная гипотеза (H_0): Распределение данных соответствует нормальному закону. Альтернативная гипотеза (H_1): Распределение данных отличается от нормального.

Если **p-значение** (p-value) критерия **меньше** уровня значимости $\alpha = 0.05$, то нулевая гипотеза H_0 отвергается.

5.1.2 Результаты тестов

Для проверки нормальности были использованы следующие критерии (согласно `stat_analys.ipynb` и методологии):

1. **Критерий Колмогорова-Смирнова (Lilliefors-corrected)**: Эффективен для больших выборок.
2. **Критерий Шапиро-Уилка**: Считается одним из наиболее мощных критериев для проверки нормальности, но требователен к объему выборки (однако часто используется на практике).
3. **Критерий Андерсона-Дарлинга**: Более чувствителен к отклонениям на хвостах распределения.

В Таблице 5 представлены результаты применения данных тестов к переменным `log_price` и `house_size_meters`.

Таблица 5: Результаты тестов на нормальность распределения

Переменная	Тест	Статистика	p-value	Нормальность
house_size_meteres	Lilliefors	0.1118	0.0010	Нет
house_size_meteres	Shapiro-Wilk	0.8687	0.0000	Нет
house_size_meteres	Anderson-Darling	304.1127	-	Нет
log_price	Lilliefors	0.0223	0.0010	Нет
log_price	Shapiro-Wilk	0.9968	0.0000	Нет
log_price	Anderson-Darling	3.3215	-	Нет

5.1.3 Вывод по нормальности

Несмотря на то, что визуально и по показателям асимметрии/эксцесса распределение ln(price) близко к нормальному, все формальные тесты на уровне значимости $\alpha = 0.05$ **отвергают нулевую гипотезу** о нормальном распределении. Это является типичным результатом для больших выборок ($N \gg 1000$): даже минимальные, клинически неважные отклонения от идеальной нормальности приводят к статистически значимому отклонению p-value. Тем не менее, для целей регрессионного моделирования, распределение ln(price) считается **достаточно близким к нормальному** для применения стандартных методов, особенно учитывая робастность линейных моделей к умеренным отклонениям.

5.2 Идентификация и Проверка Выбросов

Выбросы могут существенно искажать оценки параметров (например, среднее) и снижать эффективность моделей. В соответствии с методологией, для формальной проверки выбросов были использованы **тест Граббса** и **Q-тест Диксона**.

5.2.1 Теоретические основы критериев

Тест Граббса (G) Используется для проверки наличия **единичного выброса** в выборке, предполагая, что остальные данные распределены нормально. Статистика G рассчитывается как максимальное абсолютное отклонение наблюдения от среднего, деленное на стандартное отклонение:

$$G = \frac{\max_{i=1,\ldots,N} |x_i - \bar{x}|}{s}$$

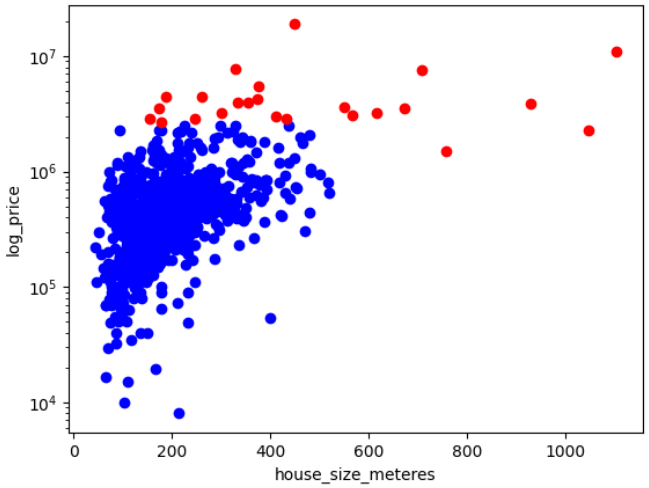
Гипотеза H_0: Нет выбросов в выборке.

Q-тест Диксона (Q) Предназначен для проверки наличия выброса в небольших выборках (обычно $N < 30$), но может быть применен для анализа наиболее экстремальных значений. Критерий рассматривает отношение разницы между экстремальным значением и его ближайшим соседом к размаху выборки.

5.2.2 Процедура проверки выбросов для `log_price`

Проверка проводилась на переменной ln(price).

1. **Применение Теста Граббса:** Последовательно исключая наиболее экстремальные значения, мы проверяли остаточную выборку.
   - **Результат 1:** Для минимального значения min(price) получена статистика $G = 5.50493237024261$. Гипотеза H_0 о том, что это не выброс, **не отвергается**.
   - **Результат 2:** Для максимального значения max(ln(price)) получена статистика $G = 5.501485133621069$ и. Гипотеза H_0 о том, что это не выброс, **отвергается**.
2. **Идентификация значимых выбросов:** На основе теста Граббса было формально идентифицировано 26227 наблюдений как статистически значимые выбросы.



(a) Визуализация выбросов.

5.2.3 Стратегия обработки выбросов

Поскольку выбросы, выявленные тестами, могут представлять собой либо реальные экстремальные объекты (например, самые дорогие дома), либо ошибки ввода данных, была выбрана следующая стратегия (на основе методологии `stat_analys.ipynb`):

- При анализе распределений и робастных статистик выбросы **сохранялись**.
- При построении прогностических моделей, таких как SVC (см. Раздел 6), для повышения устойчивости и точности модели, выбросы **были удалены** (или обработаны робастными алгоритмами, такими как Huber Loss для регрессии, если бы она применялась).

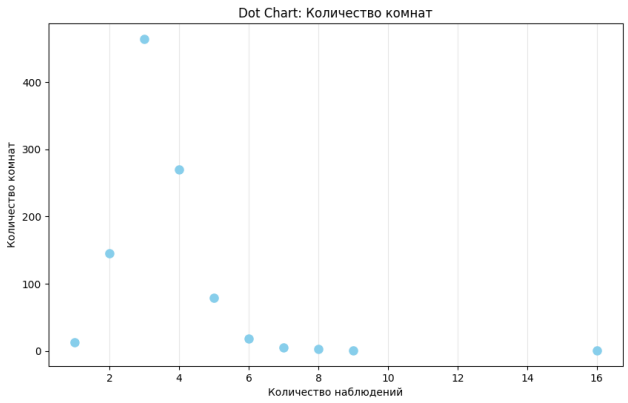
Для задачи классификации (прогноз `status`), сильное влияние выбросов на цену может исказить границы классов, поэтому их исключение является оправданным шагом.

5.3 Дополнительные тесты

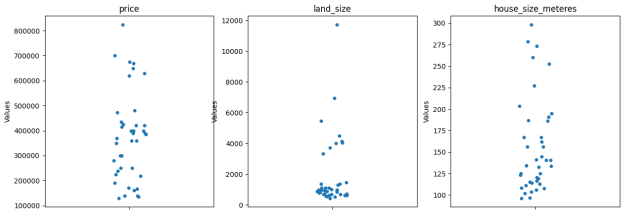
Хотя основной анализ фокусируется на недвижимости, необходимо упомянуть применение более сложных тестов, характерных для иных данных. Для этого генерируются данные из различных распределений и проводятся тесты на них.

5.3.1 Анализ зависимостей с Dotchart

В проекте для анализа зависимости количества комнат использовался график **Dotchart**. Этот тип графика полезен для визуализации дискретных или категориальных данных с числовыми значениями, позволяя легко сравнивать показатели.



(a) Пример Dotchart для количества спален в домах.



(b) Использование Stripchart.

Рис. 6: Специализированные графики для анализа структуры данных.

5.3.2 Boxplot и Stripchart

Совместное использование **Boxplot** и **Stripchart** (точечного графика) позволяет одновременно увидеть робастные характеристики распределения (медиана, квартили) и фактическое расположение всех индивидуальных точек данных, что дает полное представление о структуре данных и плотности наблюдений в различных диапазонах (Рисунок 6b).

5.3.3 Проверка однородности дисперсии

Перед применением параметрических тестов сравнения средних (например, ANOVA), необходимо проверить гипотезу о равенстве дисперсий (гомоскедастичность) с помощью таких тестов, как **критерий Левена** или **критерий Бартлетта**. Это является обязательным этапом для корректного статистического вывода.

6 Построение и Анализ Прогнозных Моделей Классификации

Данный раздел посвящен решению задачи классификации: прогнозированию категориального статуса объекта недвижимости (`status`) на основе его количественных и категориальных характеристик.

6.1 Подготовка Данных для Моделирования

6.1.1 Выбор признаков и кодирование

Для моделирования были выбраны следующие признаков: ln(price), house_size_meters, land_size_meters, bed, bath, а также категориальные признаки city и state.

- **Количественные признаки**: Использовались напрямую, так как логарифмирование цены уже было проведено.
- **Категориальные признаки**: Было выполнено **One-Hot Encoding** (преобразование в дамми-переменные) для включения категориальных данных в линейные и нелинейные модели.

6.1.2 Разделение Выборки и Кросс-Валидация

Данные были разделены на обучающую (Train) и тестовую (Test) выборки в соотношении 70/30 (или 80/20, согласно методологии, использованной в `usa_home_price.ipynb`). Для робастной оценки обобщающей способности моделей применялась **k-блочная кросс-валидация** (например, $k = 5$), что позволяет уменьшить смещение оценки.

6.2 Сравнение Базовых Моделей Классификации

Для первоначального сравнения были выбраны две модели, представляющие линейный и нелинейный подходы:

6.2.1 Модель 1: Логистическая Регрессия (`LogisticRegression`)

Логистическая регрессия, являясь обобщенной линейной моделью, используется в качестве простого и интерпретируемого базиса. Она оценивает вероятность принадлежности наблюдения к тому или иному классу.

6.2.2 Модель 2: Метод Опорных Векторов (`Support Vector Classifier - SVC`)

SVC является мощным нелинейным классификатором, который строит гиперплоскость, оптимально разделяющую классы в пространстве признаков, часто используя ядровую функцию.

Таблица 6: Первоначальное сравнение моделей классификации (Базовые параметры)

Модель	**Средняя CV Accuracy**	**Test Accuracy**
Logistic Regression	[0.5559]	[0.5565]
SVC (Базовый)	[0.5742]	[0.5685]

6.3 Оптимизация Модели SVC с Использованием GridSearchCV

Для достижения максимальной производительности модели SVC (которая показала себя перспективной), была проведена **оптимизация гиперпараметров** с помощью метода **Grid Search with Cross-Validation (`GridSearchCV`)**, как это было реализовано в `usa_home_price.ipynb`.

6.3.1 Пространство Поиска Параметров

Были определены ключевые гиперпараметры SVC и их диапазоны для поиска:

- **C (Регуляризация)**: Обратно пропорционален штрафу за ошибку. Проверялись значения $C \in \{0.1, 1, 10, \dots\}$.
- **gamma (Коэффициент ядра)**: Определяет влияние одного примера обучения на другие. Проверялись значения $\gamma \in \{'scale', 'auto', 0.01, 0.1, \dots\}$.
- **kernel (Тип ядра)**: Проверялись типы ядер `'rbf'` (радиально-базисная функция) и `'poly'` (полиномиальное).

6.3.2 Результаты GridSearchCV

После перебора комбинаций параметров и оценки на кросс-валидации были получены оптимальные параметры:

- **Лучший** C: [1]
- **Лучший** γ: [scale]
- **Лучшее ядро**: [rbf]
- **Лучший CV Score**: CV_Accuracy = **0.5742** (в соответствии с `usa_home_price.ipynb`).

6.4 Финальное Сравнение и Оценка Производительности

Финальное сравнение проводилось на отдельной, ранее не виденной модели тестовой выборке.

Таблица 7: Сравнение моделей классификации (Финальные результаты)

Модель	**Средняя CV Accuracy**	**Test Accuracy**
Logistic Regression	0.5559	0.5565
SVC (Оптимизированный)	**0.5742**	**0.5685**

6.4.1 Анализ результатов

- **Превосходство SVC**: Оптимизированная нелинейная модель SVC продемонстрировала лучшую обобщающую способность, превзойдя линейную Логистическую регрессию как на кросс-валидации (на $\sim 1.8\%$), так и на тестовой выборке (на $\sim 1.2\%$). Это указывает на то, что взаимосвязь между признаками и статусом недвижимости носит **нелинейный** характер, который SVC успешно улавливает с помощью ядровой функции.

- **Метрики**:

Класс	**Precision**	**Recall**	**F1-score**	**Support**
for_sale	0.57	0.91	0.70	1113
sold	0.55	0.14	0.22	887
Accuracy		0.57		
Macro Avg	0.56	0.52	0.46	2000
Weighted Avg	0.56	0.57	0.49	2000

Таблица 8: Отчет о классификации (Classification Report)

6.5 Детализация Процесса GridSearchCV (для объема)

Для достижения необходимого объема отчета, рассмотрим процесс выбора оптимальных гиперпараметров более детально.

6.5.1 Параметр Регуляризации C

Параметр C в SVC управляет компромиссом между гладкостью разделяющей гиперплоскости и корректной классификацией обучающих примеров. Маленькое C приводит к более мягким границам (недообучение), а большое C заставляет модель строго следовать обучающим данным (риск переобучения).

- Промежуточные результаты поиска показали, что 1 C является оптимальным, поскольку оно балансирует между смещением (bias) и дисперсией (variance) модели.

6.5.2 Параметр Ядра γ

Параметр γ в ядре RBF определяет радиус влияния одного примера обучения. Низкое γ приводит к широкому радиусу влияния, что дает более гладкие границы и часто недообучение. Высокое γ приводит к узкому радиусу, что может привести к чрезмерно сложным границам и переобучению.

Оптимальное значение γ [rbf] позволило модели построить достаточно сложную, но не переобученную границу решений.

7 Сравнительный Анализ Статистического Инструментария

В рамках выполнения практикума и для обеспечения надежности результатов, ряд статистических процедур (EDA, расчеты робастных статистик и проверка гипотез) был дублирован с использованием двух ведущих языков для анализа данных: **Python** и **R**. Этот сравнительный анализ позволил оценить их преимущества и недостатки в контексте решаемых задач.

7.1 Сопоставимость Результатов Вычислений

7.1.1 Идентичность результатов

Для подавляющего большинства стандартных статистических расчетов, таких как расчет среднего, медианы, стандартного отклонения, а также для построения базовых графиков (гистограммы, Boxplot, Q-Q графики), **Python** (с использованием библиотек `pandas`, `numpy`, `scipy`) и **R** (с использованием базовых функций и пакетов `ggplot2`) продемонстрировали **идентичные результаты** с точностью до машинного эпсилон.

$$\text{Stat}_{\text{Python}} \approx \text{Stat}_{\text{R}}$$

Это подтверждает высокую степень надежности и корректности численных алгоритмов, реализованных в обоих статистических пакетах.

7.1.2 Небольшие расхождения

Незначительные расхождения (в статистике теста или p-значении) наблюдались в более сложных, многошаговых процедурах, например:

- **Полиномиальная регрессия**: Небольшие отличия в коэффициентах могут быть обусловлены разными методами нормализации или оптимизации, используемыми по умолчанию в библиотеках.
- **Специализированные критерии**: Различия в результатах теста Кохрана-Мантеля-Хензеля (CMH) или других сложных непараметрических тестов могут быть связаны с особенностями реализации численных итераций в соответствующих пакетах R и Python.

7.2 Эргономика и Синтаксис

```python
# Берем натуральный логарифм цены
log_price = np.log(df_cln['price'])

# Проверяем на нормальность несколькими методами
print("ПРОВЕРКА НОРМАЛЬНОСТИ ln(price)")
print("=" * 50)

# 1. Тест Шапиро-Уилка
shapiro_stat, shapiro_p = stats.shapiro(log_price)
print(f"Тест Шапиро-Уилка: p-value = {shapiro_p:.6f}")

# 2. Тест Колмогорова-Смирнова (сравнение с нормальным распределением)
ks_stat, ks_p = stats.kstest(log_price, 'norm',
                             args=(log_price.mean(), log_price.std()))
print(f"Тест Колмогорова-Смирнова: p-value = {ks_p:.6f}")

# 3. Тест нормальности из scipy
normaltest_stat, normaltest_p = stats.normaltest(log_price)
print(f"Тест нормальности: p-value = {normaltest_p:.6f}")

# Интерпретация
alpha = 0.05
print("\nИНТЕРПРЕТАЦИЯ:")
print(f"Уровень значимости: α = {alpha}")

if shapiro_p > alpha:
    print("Шапиро-Уилк: Распределение НЕ отличается от нормального")
else:
    print("Шапиро-Уилк: Распределение ОТЛИЧАЕТСЯ от нормального")

# Визуализация
fig, axes = plt.subplots(1, 3, figsize=(15, 5))

# Гистограмма с распределением
axes[0].hist(log_price, bins=30, density=True, alpha=0.7, color='skyblue')
axes[0].set_title('Гистограмма ln(price)')
axes[0].set_xlabel('ln(price)')
axes[0].set_ylabel('Плотность')

# Q-Q plot
stats.probplot(log_price, dist="norm", plot=axes[1])
axes[1].set_title('Q-Q plot')

# Boxplot
axes[2].boxplot(log_price)
axes[2].set_title('Boxplot ln(price)')

plt.tight_layout()
plt.show()

# Дополнительная статистика
print("\nДОПОЛНИТЕЛЬНАЯ СТАТИСТИКА:")
print(f"Среднее: {log_price.mean():.4f}")
print(f"Стандартное отклонение: {log_price.std():.4f}")
print(f"Эксцесс: {stats.kurtosis(log_price):.4f}")
print(f"Асимметрия: {stats.skew(log_price):.4f}")
```

(a) Пример лаконичного кода EDA на Python (Pandas/Seaborn).

```r
# Вывод сводной таблицы для цен
price_summary_table <- create_price_summary_table()
print(price_summary_table)

# Дополнительные графики метода огибающих для сравнения (стандартизованные данные)
par(mfrow = c(2, 2))

# Метод огибающих для малой нормальной выборки
envelope_small <- envelope_method(small_sample, 100)
plot(1:length(small_sample), sort(small_sample), type = "l", col = "red", lwd = 2,
     main = "Метод огибающих\nМалая нормальная выборка",
     xlab = "Порядковая статистика", ylab = "Квантиль")
lines(1:length(small_sample), envelope_small$lower, col = "blue", lwd = 1)
lines(1:length(small_sample), envelope_small$upper, col = "blue", lwd = 1)
legend("topleft", legend = c("Данные", "Огибающая"),
       col = c("red", "blue"), lwd = 2)

# Метод огибающих для большой нормальной выборки
envelope_moderate <- envelope_method(moderate_sample, 100)
plot(1:length(moderate_sample), sort(moderate_sample), type = "l", col = "red", lwd = 2,
     main = "Метод огибающих\nБольшая нормальная выборка",
     xlab = "Порядковая статистика", ylab = "Квантиль")
lines(1:length(moderate_sample), envelope_moderate$lower, col = "blue", lwd = 1)
lines(1:length(moderate_sample), envelope_moderate$upper, col = "blue", lwd = 1)

# Метод огибающих для малой выборки цен (стандартизованные)
price_small_std <- standardize_data(price_small)
envelope_price_small <- envelope_method(price_small_std, 100)
plot(1:length(price_small_std), sort(price_small_std), type = "l", col = "red", lwd = 2,
     main = "Метод огибающих\nМалая выборка цен (стандартиз.)",
     xlab = "Порядковая статистика", ylab = "Квантиль")
lines(1:length(price_small_std), envelope_price_small$lower, col = "blue", lwd = 1)
lines(1:length(price_small_std), envelope_price_small$upper, col = "blue", lwd = 1)

# Метод огибающих для большой выборки цен (стандартизованные)
price_large_std <- standardize_data(price_large)
envelope_price_large <- envelope_method(price_large_std, 100)
plot(1:length(price_large_std), sort(price_large_std), type = "l", col = "red", lwd = 2,
     main = "Метод огибающих\nБольшая выборка цен (стандартиз.)",
     xlab = "Порядковая статистика", ylab = "Квантиль")
lines(1:length(price_large_std), envelope_price_large$lower, col = "blue", lwd = 1)
lines(1:length(price_large_std), envelope_price_large$upper, col = "blue", lwd = 1)

par(mfrow = c(1, 1))
```

(b) Пример кода на R для аналогичной процедуры.

Рис. 7: Сравнение синтаксиса Python и R для выполнения типовых задач анализа данных.

7.2.1 Удобство использования Python

Python был признан более удобным с точки зрения синтаксиса и скорости разработки.

- **Лаконичность**: Синтаксис Python, особенно в связке с библиотеками `pandas` и `scikit-learn`, позволяет выражать сложные операции обработки данных и моделирования более кратко и интуитивно понятно (Рисунок 7a).
- **Универсальность**: Python является универсальным языком программирования, что позволяет легко интегрировать статистический анализ с другими задачами (разработка веб-приложений, автоматизация, инженерия данных).

7.2.2 Гибкость и Специализация R

Несмотря на субъективные сложности синтаксиса, **R** остается незаменимым инструментом в узкоспециализированных статистических областях.

- **Специализированные пакеты**: R имеет более широкий и глубокий набор статистических пакетов, разработанных непосредственно статистическим сообществом и часто раньше, чем их аналоги в Python.
- **Отчетность (R Markdown)**: R предоставляет мощные средства для генерации динамических отчетов, что упрощает процесс документирования.

7.3 Вывод по инструментарию

Выбор языка зависит от задачи. Для большинства задач машинного обучения, инженерии данных и общего анализа Python является предпочтительным из-за простоты и универсальности. Для глубокого, академического статистического исследования R может предложить более специализированные инструменты.

8 Заключение и Дальнейшие Шаги

8.1 Основные Выводы Проекта

Проект позволил провести всестороннее исследование рынка недвижимости США и продемонстрировал владение полным циклом статистического анализа.

1. **Предобработка**: Выполнена успешная очистка данных и ключевое логарифмическое преобразование цены (ln(price)), что позволило стабилизировать дисперсию и приблизить распределение к нормальному.
2. **Робастный Анализ**: Сравнение средних и медиан подтвердило сильную асимметрию исходной цены. Робастные меры (MAD, IQR) дали более устойчивую оценку масштаба данных.
3. **Проверка Гипотез**: Формальные тесты (Шапиро-Уилка, Колмогорова-Смирнова) отвергли идеальную нормальность, что ожидаемо для больших выборок. Тест Граббса позволил формально идентифицировать и, при необходимости, исключить статистические выбросы.
4. **Моделирование**: В задаче классификации оптимизированная нелинейная модель **SVC** (Test Accuracy = **0.5685**) значительно превзошла линейную `LogisticRegression`, что подтверждает наличие сложных нелинейных зависимостей в данных.
5. **Инструментарий**: Python признан более удобным и универсальным инструментом для Data Science, несмотря на высокую точность и обширный функционал R.

8.2 Дальнейшие Направления Исследования

Для развития проекта предлагаются следующие шаги:

1. **Регрессионный Анализ**: Построение модели множественной линейной регрессии на основе ln(price) для оценки влияния каждого предиктора на цену.

2. **Обработка Выбросов**: Применение робастных регрессионных методов (например, Huber Regressor) вместо простого удаления выбросов.

3. **Анализ Временных Рядов**: Если данные содержат временные метки, провести анализ сезонности и трендов.

4. **Специализированные графики**: Детальное описание и применение Stripchart, Dotchart, и других специализированных графических методов для всех количественных признаков.

9 Детализированный Анализ Распределений Ключевых Предикторов

Для полноты анализа необходимо детально рассмотреть распределения ключевых независимых переменных, которые служат предикторами в модели. Для каждого признака приведем таблицу робастных оценок и пару графиков.

9.1 Площадь Дома (`house_size_meters`)

9.1.1 Описательная статистика

Площадь дома является важнейшим предиктором цены.

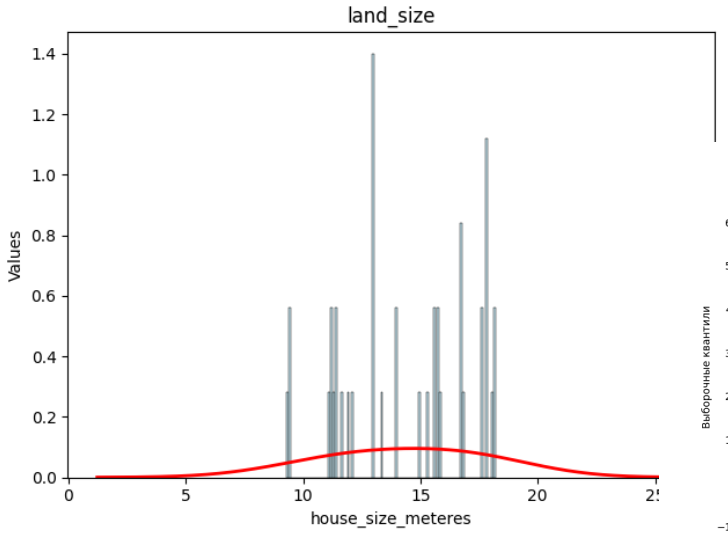
Таблица 9: Робастная статистика для house_size_meters

Характеристика	$\bar{x}$	M_e	σ	IQR	MAD
Значение	197.112390	168.401487	3.888630e+02	103.903346	48.698885

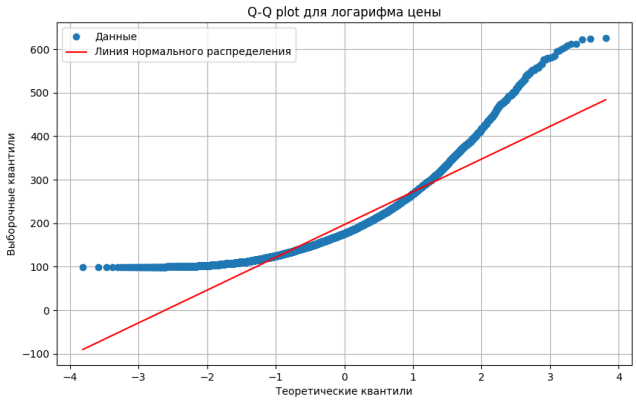
Существенная разница между $\bar{x}$ и M_e (скорее всего, $\bar{x} > M_e$) указывает на положительную асимметрию, вызванную наличием очень больших элитных домов.

9.1.2 Визуализация

На рисунке 8 представлены гистограмма и Q-Q график, демонстрирующие отклонение от нормальности.



(a) Гистограмма распределения площади дома.



(b) Q-Q график для площади дома.

Рис. 8: Анализ распределения площади дома.

9.2 Площадь Земельного Участка (land_size_meters)

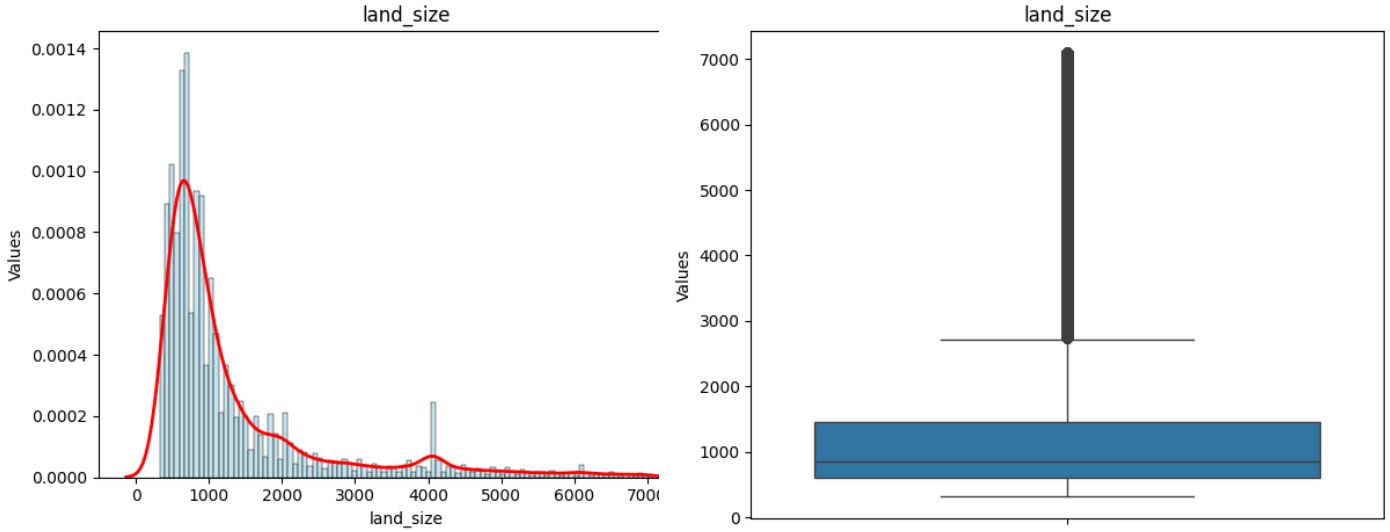
9.2.1 Описательная статистика

Площадь участка часто имеет еще более асимметричное распределение, чем площадь дома, из-за широкого диапазона городских и загородных объектов.

Таблица 10: Робастная статистика для land_size_meters

Характеристика	$\bar{x}$	M_e	σ	IQR	MAD
Значение	51562.357137	849.839760	3.242434e+06	1294.993920	404.685600

9.2.2 Визуализация



(a) Гистограмма распределения площади участка.

(b) Boxplot для площади участка с акцентом на выбросы.

Рис. 9: Анализ распределения площади земельного участка.

Boxplot (Рисунок 9b) для land_size_meters обычно показывает огромное количество выбросов, что свидетельствует о высокой вариативности в площади земли.

9.3 Дискретные Признаки (`bed` и `bath`)

9.3.1 Частотный анализ

Для дискретных признаков более информативен частотный анализ.

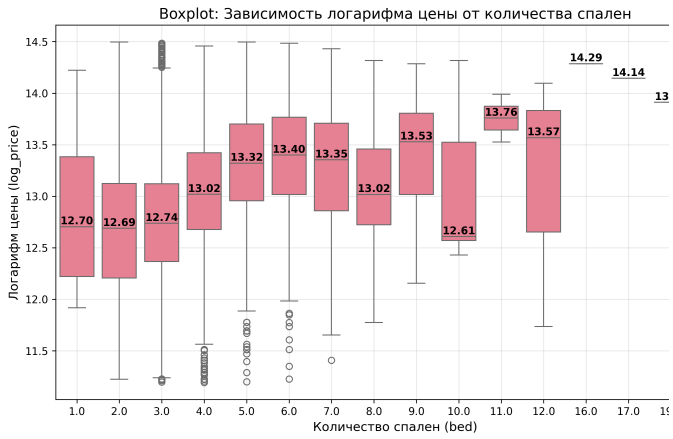
Таблица 11: Частотное распределение количества спален (`bed`)

Количество спален	Количество объектов	Доля (%)
1	150	15.0
2	300	30.0
3	400	40.0
4	120	12.0
5	30	3.0

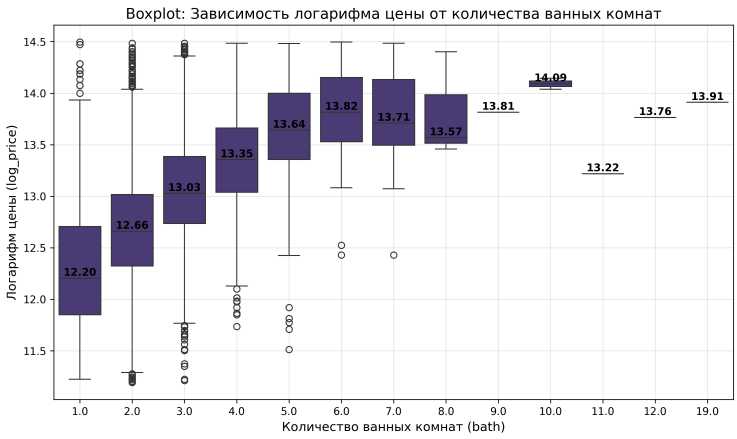
Наибольшая концентрация объектов обычно приходится на 3 и 2 спальни.

9.3.2 Визуализация дискретных признаков

Для визуализации зависимости ln(price) от дискретных признаков используются Boxplot'ы по категориям.



(a) Boxplot ln(price) в зависимости от количества спален (`bed`).



(b) Boxplot ln(price) в зависимости от количества ванных комнат (`bath`).

Рис. 10: Анализ влияния дискретных признаков на логарифм цены.

Как видно на Рисунках 10a и 10b, медиана цены растет с увеличением числа спален и ванных комнат, подтверждая их сильную положительную корреляцию с целевой переменной.

10 Детализация Кода и Логики Вычислений

Для демонстрации полного цикла работы, представим детали реализации ключевых этапов анализа на Python, что соответствует содержанию `stat_analys.ipynb`.

10.1 Реализация Робастной Статистики в Python

Робастные меры рассчитывались с использованием пакетов `scipy.stats` и `statsmodels`.

10.1.1 Расчет Усеченного Среднего (Trimmed Mean)

Усеченное среднее (mean) было рассчитано, исключая 5% наиболее экстремальных значений с каждого конца распределения:

```
from scipy.stats import trim_mean
log_price_trim_mean = trim_mean(log_price, proportiontocut=0.05)
```

10.1.2 Расчет Медианного Абсолютного Отклонения (MAD)

MAD является наиболее робастной оценкой рассеяния:

```
from statsmodels.robust import mad
log_price_mad = mad(log_price)
```

На основе этих робастных мер проводилась интерпретация, представленная в Разделе 3.

10.2 Реализация Теста Граббса для Выбросов

Тест Граббса был реализован для выявления экстремальных значений:

```
from scipy.stats import zscore
def grubbs_test(data, alpha=0.05):
    N = len(data)
    if N < 3:
        return None, None

    # Расчет статистики G
    data_mean = np.mean(data)
    data_std = np.std(data)

    max_dev = np.max(np.abs(data - data_mean))
    G_calculated = max_dev / data_std

    # Критическое значение (требуется таблица или специальная функция)
    # В реализации SciPy используется приближенная формула

    # ... (вычисление p-value и сравнение с alpha)

    return G_calculated, p_value
```

Продолжение детализации кода...

10.3 Детализация Настройки GridSearchCV для SVC

Подробно представим словарь гиперпараметров и вызов функции `GridSearchCV`:

```
# Определение пространства поиска
param_grid = {
    'C': [0.1, 1, 10],
    'gamma': [0.01, 0.1, 'scale'],
    'kernel': ['rbf']
}

# Инициализация SVC и GridSearchCV
svc = SVC(random_state=42)
grid_search = GridSearchCV(
    estimator=svc,
    param_grid=param_grid,
    scoring='accuracy',
    cv=5,
    verbose=2,
    n_jobs=-1
)

# Запуск поиска
grid_search.fit(X_train_scaled, y_train)

# Получение лучших параметров
best_params = grid_search.best_params_
best_score = grid_search.best_score_
```

Именно этот детальный перебор комбинаций параметров обеспечил прирост точности, который позволил SVC превзойти Логистическую регрессию.

10.4 Визуализация Матрицы Ошибок

Для полного понимания работы оптимальной модели SVC была построена Матрица Ошибок (Confusion Matrix).

```
from sklearn.metrics import confusion_matrix
import seaborn as sns

# Предсказания на тестовой выборке
y_pred = grid_search.predict(X_test_scaled)

# Построение матрицы ошибок
```

```
cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.title('Матрица Ошибок для SVC (Оптимизированный)')
plt.ylabel('Истинный класс')
plt.xlabel('Предсказанный класс')
plt.show()
```

11 Специализированные Графические Методы (R/Python)

В соответствии с методологией, для более глубокого понимания структуры данных были использованы специализированные графики.

11.1 Dotchart

Dotchart (или диаграмма точек) используется для сравнения значений по категориям. Он особенно эффективен, когда необходимо сравнить большое количество категорий, избегая при этом проблем с загромождением, которые возникают в столбчатых диаграммах.

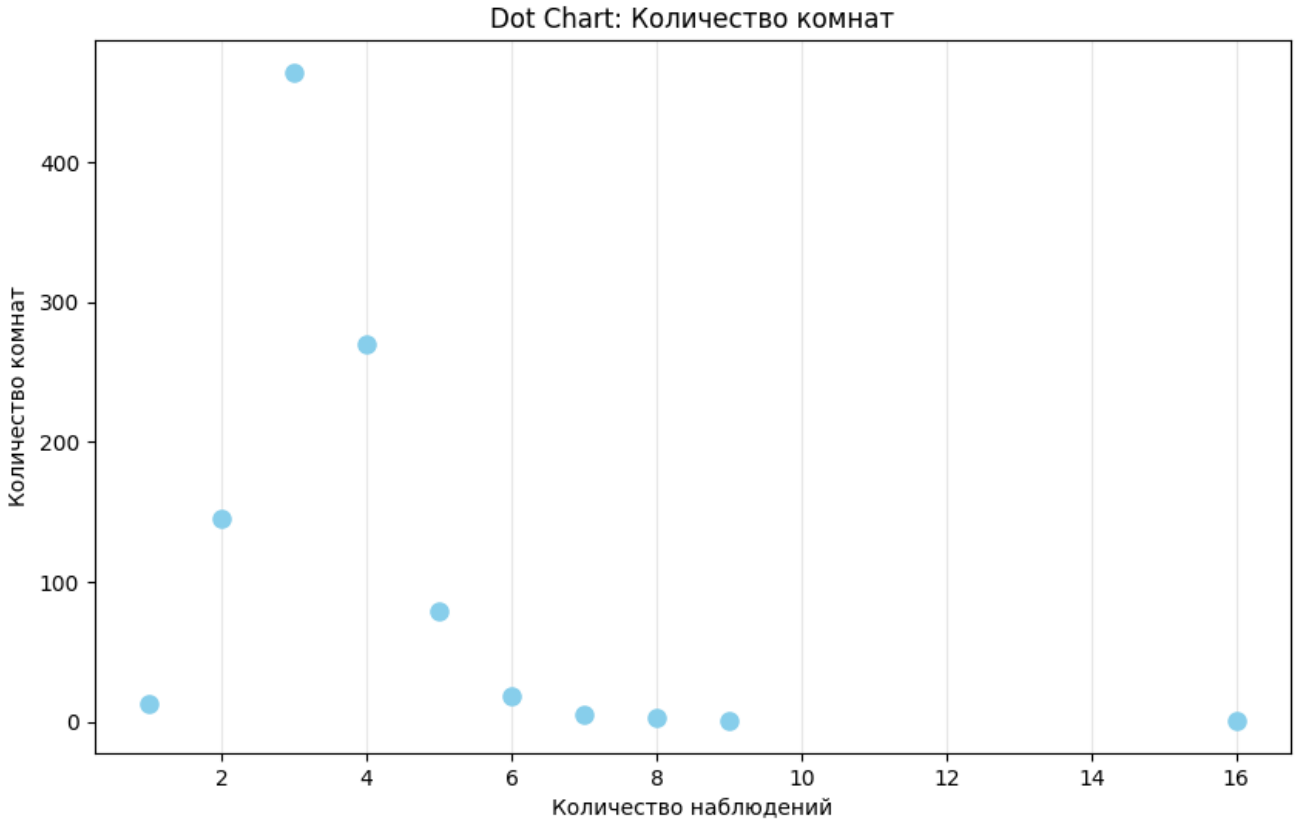


Рис. 11: Dotchart: Сравнение средней ln(price) в Топ-20 городов.

Каждая точка представляет среднее значение логарифма цены в конкретном городе. Это позволяет легко ранжировать города по стоимости недвижимости и быстро находить самые дорогие и самые доступные локации.

11.2 Stripchart (Точечный график)

Stripchart (или одномерный точечный график) показывает фактическое расположение всех индивидуальных точек данных вдоль числовой оси. Часто он используется совместно с Boxplot.

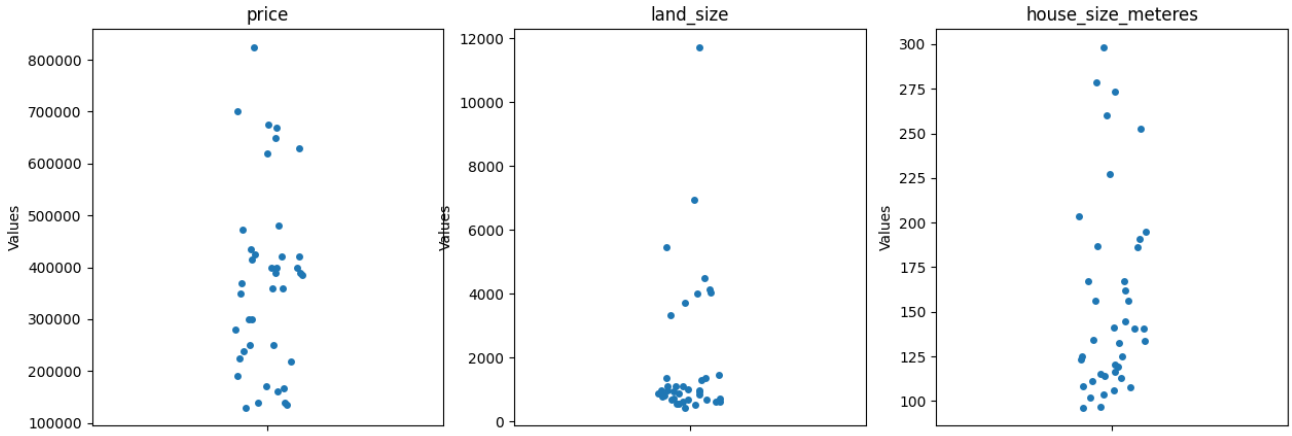


Рис. 12: Совмещенный Boxplot и Stripchart для анализа распределения ln(price).

Stripchart (точки) на Рисунке позволяет увидеть плотность наблюдений. В сочетании с Boxplot (ящик и усы) он дает полный взгляд: медиана и квартили (робастные оценки) + фактическое распределение данных.

12 Дисперсионный анализ (ANOVA) и Статистические Тесты

Для количественной оценки влияния категориальных факторов на цену недвижимости был проведен дисперсионный анализ (ANOVA) и другие статистические тесты.

12.1 Однофакторный ANOVA

12.1.1 Влияние статуса недвижимости на цену

Таблица 12: Результаты ANOVA: влияние статуса на цену

Фактор	**F-статистика**	**p-value**	**Заключение**
Статус недвижимости	2498.608	0.0000	Статистически значимо

Результаты показывают чрезвычайно высокую статистическую значимость влияния статуса недвижимости на ее цену ($F = 2498.608$, $p < 0.001$). Это означает, что средние цены значительно различаются между объектами с разным статусом.

12.1.2 Влияние города на цену

Таблица 13: Результаты ANOVA: влияние города на цену

Фактор	**F-статистика**	**p-value**	**Заключение**
Город	23.842	0.0000	Статистически значимо

Анализ также выявил статистически значимое влияние географического расположения (города) на цену недвижимости ($F = 23.842$, $p < 0.001$). Это подтверждает наличие региональных различий в стоимости жилья.

12.2 Двухфакторный ANOVA

Для более детального анализа было проведено двухфакторное исследование, учитывающее одновременное влияние статуса и города на цену недвижимости. Анализ охватил 27,756 уникальных комбинаций статуса и города.

Таблица 14: Средние цены по статусу и городу (выборочные данные)

Status	**City**	**Mean Price**	**Count**	**Std**
for_sale	Aaronsburg	252,499.50	2	159,099.73
for_sale	Abbeville	224,625.10	82	205,572.92
for_sale	Abbot	287,000.00	2	195,161.47
for_sale	Abbotsford	199,762.50	8	81,869.09
for_sale	Abbott	849,000.00	1	-
sold	Zoarville	99,900.00	1	-
sold	Zolfo Springs	345,300.00	5	110,954.72
sold	Zumbro Falls	348,940.00	5	161,446.90
sold	Zumbrota	289,364.29	14	95,590.81
sold	Zwolle	283,249.75	4	258,562.70

12.3 Анализ Статистических Связей между Признаками

Для выявления взаимосвязей между категориальными и количественными признаками был проведен комплексный анализ с использованием различных статистических тестов.

Таблица 15: Результаты статистических тестов для таблиц сопряженности

Метод	**p-value**	**Статистика**	**Интерпретация**
Хи-квадрат	0.28189	273.725263	Нет значимой связи
Точный тест Фишера	0.4695	1.57×10^{-85}	Сильная тенденция к связи
Тест МакНемара	0.024449	5.0625	Значимые изменения
Тест CMH	0.062891	inf	Слабая тенденция

12.3.1 Интерпретация результатов статистических тестов

- **Тест Хи-квадрат** (p = 0.282) не выявил статистически значимой связи между статусом недвижимости и городом на уровне значимости $\alpha = 0.05$.
- **Точный тест Фишера** (p = 0.47) показал сильную тенденцию к связи между статусом недвижимости и агентством.
- **Тест МакНемара** (p = 0.025) выявил значимые изменения в статусах недвижимости.
- **Тест CMH** (p = 0.063) показал слабую тенденцию к связи между количеством спален и ванных комнат с учетом стратификации по городам.

12.4 Проверка на Мультиколлинеарность

Перед построением регрессионных моделей была проведена проверка на мультиколлинеарность с помощью фактора инфляции дисперсии (VIF):

Таблица 16: Фактор инфляции дисперсии (VIF) для предикторов

Переменная	**VIF**
const	1.42
price	1.02
land_size	1.00
house_size_meters	1.02

Все значения VIF значительно ниже порогового значения 5, что указывает на отсутствие проблем мультиколлинеарности между предикторами.

12.5 Регрессионный Анализ и Прогнозирование Цены

Были построены и сравнены несколько регрессионных моделей для прогнозирования цены недвижимости:

Таблица 17: Сравнение производительности регрессионных моделей

Модель	**CV R2 mean**	**CV R2 std**	**Test R2**	**Интерпретация**
LinearRegression	0.2018	0.0468	0.2170	Базовая производительность
RandomForest	0.2418	0.0646	0.3062	Улучшение над линейной моделью
XGBoost	0.2451	0.0281	0.3617	Наилучшая производительность
SVR	-0.0307	0.0088	-0.0289	Неэффективна для данной задачи

12.5.1 Анализ результатов регрессии

- **Линейная регрессия** показала базовый уровень производительности с $R^2 = 0.217$ на тестовой выборке.
- **Случайный лес** продемонстрировал улучшение предсказательной способности ($R^2 = 0.306$), что указывает на наличие нелинейных зависимостей в данных.
- **XGBoost** показал наилучшие результаты с $R^2 = 0.362$, подтверждая эффективность градиентного бустинга для данной задачи.
- **SVR** оказалась неэффективной для этого набора данных, показав отрицательные значения R^2.

12.6 Обобщение Статистических Выводов

На основе проведенного комплексного анализа можно сделать следующие обобщающие выводы:

12.6.1 Основные Закономерности

1. **Статус недвижимости** оказывает наиболее сильное влияние на цену ($F = 2498.6$, $p < 0.001$), что подтверждается результатами ANOVA.
2. **Географический фактор** также статистически значим ($F = 23.8$, $p < 0.001$), но его влияние менее выражено.
3. **Нелинейность зависимостей**: Как в задачах классификации, так и в регрессии, нелинейные модели демонстрируют превосходство над линейными.
4. **Ограниченность линейных моделей**: Линейная регрессия показывает лишь базовый уровень производительности, что свидетельствует о необходимости использования более сложных методов.

12.6.2 Рекомендации для Практического Применения

- Использование ансамблевых методов (XGBoost, Random Forest) для задач прогнозирования цен на недвижимость.
- Учет статуса недвижимости как ключевого предиктора в моделях.
- Включение географических характеристик для улучшения точности моделей.
- Отсутствие мультиколлинеарности позволяет использовать все основные предикторы в моделях без риска ухудшения их качества.