Проект по анализу данных на Python: Богатый домовладелец

Лебедев Матвей[1, a] and Шаббир Кафи[1, b]

[1]Московский физико-технический институт, Долгопрудный, 141701
[a]lebedev.ma@phystech.edu M02-301я
[b]kafiulshabbir@phystech.edu M03-304б

5 июня 2024 г.

Аннотация

В последние годы рынок краткосрочной аренды жилья через платформы, такие как Airbnb, существенно вырос. Тем не менее, существует ограниченное понимание того, какие факторы наиболее сильно влияют на цены листингов. Это знание может помочь как владельцам жилья, так и арендаторам лучше ориентироваться на рынке.

Содержание

1 Введение

1.1 Проблема

В последние годы рынок краткосрочной аренды жилья через платформы, такие как Airbnb, существенно вырос [3]. Тем не менее, существует ограниченное понимание того, какие факторы наиболее сильно влияют на цены листингов. Это знание может помочь как владельцам жилья, так и арендаторам лучше ориентироваться на рынке [4].

1.2 Актуальность

Для владельцев жилья знание факторов, влияющих на цену, поможет оптимизировать свои предложения, а для арендаторов — выбрать наиболее выгодные варианты.

1.3 Исследовательский вопрос

Какие факторы объясняют вариации цен на листинги Airbnb в Амстердаме в будние дни?

1.4 Гипотезы

1. Листинги, расположенные ближе к центру города, имеют более высокие цены.
2. Листинги, предоставляемые суперхостами, имеют более высокие цены.
3. Листинги с большим количеством спален имеют более высокие цены.
4. Листинги с высоким рейтингом удовлетворенности гостей имеют более высокие цены.
5. Листинги, предлагающие целый дом или квартиру, имеют более высокие цены по сравнению с отдельными комнатами.
6. Листинги, расположенные ближе к популярным туристическим местам, имеют более высокие цены.
7. Листинги, находящиеся в районах с высоким уровнем бизнеса (biz), имеют более высокие цены.
8. Листинги с более высоким значением cleanliness_rating имеют более высокую цену.
9. Листинги, доступные для бронирования на несколько комнат (multi), имеют более высокие цены.

2 Методология

2.1 Методы сбора данных

Данные собраны и предоставлены в виде CSV-файла. Источник данных - платформа Airbnb

2.2 Описание данных

Данные содержат информацию о краткосрочной аренде жилья в Амстердаме в будние дни. Каждый объект недвижимости представлен множеством характеристик, включая тип комнаты, вместимость, оценки чистоты и удовлетворенности гостей, а также различные индексы привлекательности и доступности ресторанов. Данные могут использоваться для анализа факторов, влияющих на цены и популярность аренды, а также для оценки качества обслуживания.

Файл: amsterdam_weekends.csv [1].

2.3 Столбцы данных

1. **Unnamed:** 0: Идентификатор записи (тип: числовой)
2. **realSum:** Реальная сумма, стоимость проживания (тип: числовой)
3. **room_type:** Тип комнаты (категориальный)
4. **room_shared:** Комната общая (бинарный)
5. **room_private:** Комната частная (бинарный)
6. **person_capacity:** Вместимость комнаты (тип: числовой)
7. **host_is_superhost:** Хост является супер-хостом (бинарный)
8. **multi:** Количество объектов (тип: числовой)
9. **biz:** Бизнес-категория (тип: числовой)
10. **cleanliness_rating:** Оценка чистоты (тип: числовой)
11. **guest_satisfaction_overall:** Общая удовлетворенность гостей (тип: числовой)
12. **bedrooms:** Количество спален (тип: числовой)
13. **dist:** Расстояние до центра (тип: числовой)
14. **metro_dist:** Расстояние до метро (тип: числовой)
15. **attr_index:** Индекс привлекательности (тип: числовой)
16. **attr_index_norm:** Нормированный индекс близости к туристическим местам (тип: числовой)
17. **rest_index:** Индекс ресторанов (тип: числовой)
18. **rest_index_norm:** Нормированный индекс ресторанов (тип: числовой)
19. **lng:** Долгота (тип: числовой)
20. **lat:** Широта (тип: числовой)

2.4 Аналитические методы

1. Визуализация данных: Использование гистограмм, scatter plot, box plot и heatmap для анализа распределения данных и зависимостей между переменными.
2. Описательная статистика: Средние значения, медианы, стандартные отклонения и другие описательные статистики для ключевых переменных [6].

3. Корреляционный анализ: Определение корреляций между переменными с помощью коэффициента корреляции Пирсона [2].
4. Т-тест: Сравнение средних значений между двумя группами [5].

2.5 Настройка программы для считывания данных

Импорт библиотек

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind
%matplotlib inline
```

Считываем данные и предобрабатываем

```
file_path = 'amsterdam_weekdays.csv'
df = pd.read_csv(file_path)
df.head()
```

2]:

	Unnamed: 0	realSum	room_type	room_shared	room_private	person_capacity	host_is_superhost	multi	biz	cleanliness_rating	guest
0	0	194.033698	Private room	False	True	2.0	False	1	0	10.0	
1	1	344.245776	Private room	False	True	4.0	False	0	0	8.0	
2	2	264.101422	Private room	False	True	2.0	False	0	1	9.0	
3	3	433.529398	Private room	False	True	4.0	False	0	1	9.0	
4	4	485.552926	Private room	False	True	2.0	True	0	0	10.0	

Рис. 1: Как выглядят наши данные после импорта. А справа есть еще много столб.

3 Результаты

3.1 Корреляционная матрица

```
correlation_matrix = df[['realSum', 'dist', 'bedrooms', '
                                    guest_satisfaction_overall', '
                                    attr_index', 'multi', 'biz', '
                                    cleanliness_rating']].corr()

plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matri')
plt.show()
```

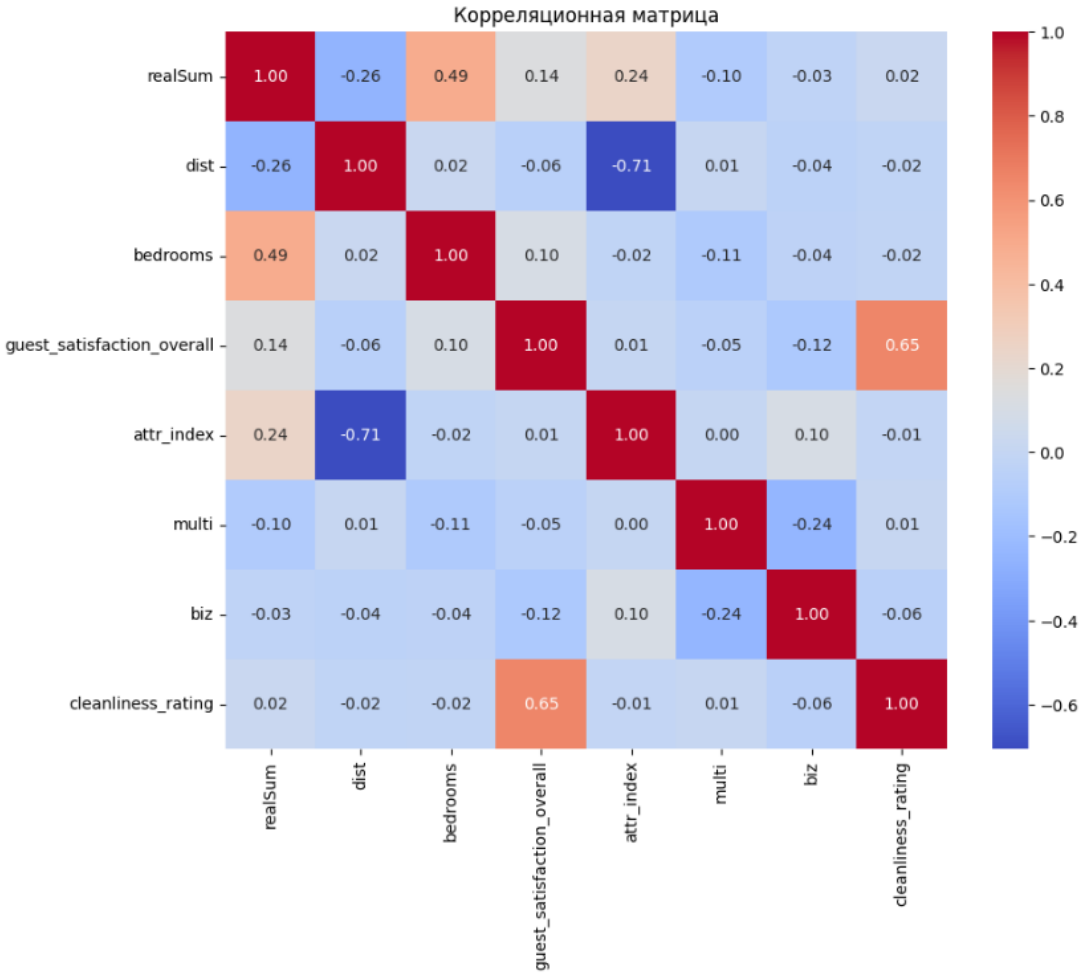

Рис. 2: тепловая карта корреляций.

3.2 Цены в зависимости от расстояния до центра города

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='dist', y='realSum', hue='dist', palette='
                                    coolwarm')
plt.title('Prices depending on the distance to the city center')
plt.xlabel('Distance to the city center (dist)')
plt.ylabel('Price (realSum)')
plt.show()
```

Рис. 3: для наших гипотез, чтобы подкрепить наши выводы из кореляционного анализа и описательной статистики. Листинги, расположенные ближе к центру города, имеют более высокие цены.

Значение корреляции между переменными -0.26, что говорит об небольшой обратной связи - чем меньше расстояние до центра, тем выше цена, это и подтверждается по графику, потому что видим сильное скопление точек на высоких значениях цен у малого значения dist и наоборот малое значение цены у точек с максимальной dist. Гипотеза подтверждается.

3.3 Цены в зависимости от того, является ли хозяин суперхостом

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='host_is_superhost', y='realSum')
plt.title('Prices depending on whether the host is a superhost')
plt.xlabel('Superhost')
plt.ylabel('Price (realSum)')
plt.show()

# t-test
superhost_prices = df[df['host_is_superhost'] == True]['realSum']
non_superhost_prices = df[df['host_is_superhost'] == False]['realSum']
t_stat_superhost, p_val_superhost = ttest_ind(superhost_prices,
                                              non_superhost_prices)
t_stat_superhost, p_val_superhost

# (-2.0985191376128354, 0.03608667618036517)
```

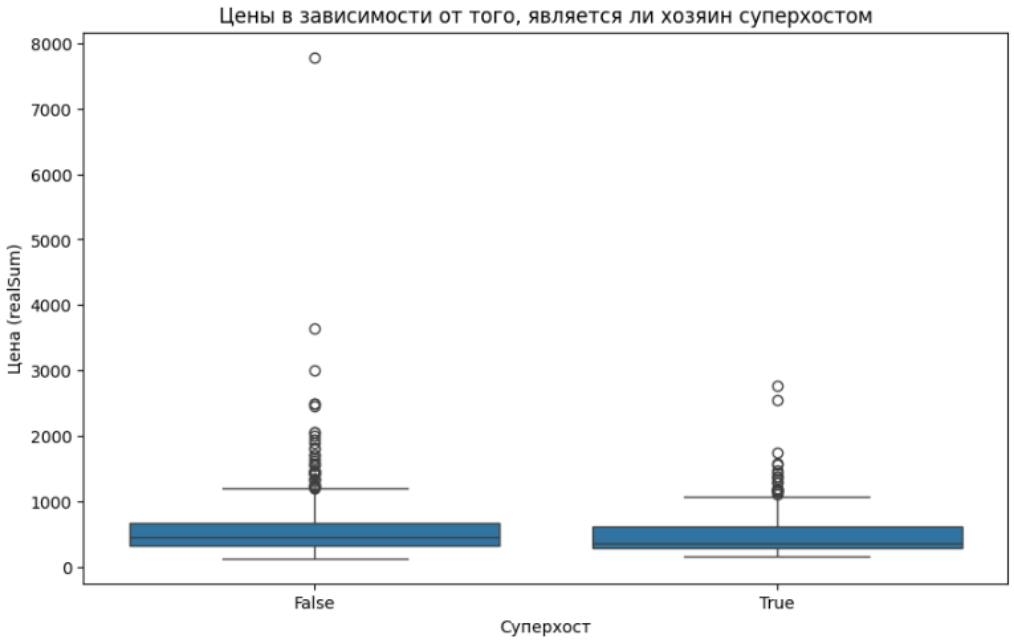

Рис. 4: Листинги, предоставляемые суперхостами, имеют более высокие цены.

Построили диарамму с усами и видим, что медианное значение ниже у True, но слабо заметно на графике, поэтому мы дополнительно сделали t-test и увидели значимую разницу в ценах причем цена у True Ниже. Гипотеза отвергается.

3.4 Цены в зависимости от количества спален

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='bedrooms', y='realSum')
plt.title('Prices depending on the number of bedrooms')
plt.xlabel('Number of bedrooms (bedrooms)')
plt.ylabel('Price (realSum)')
plt.show()
```

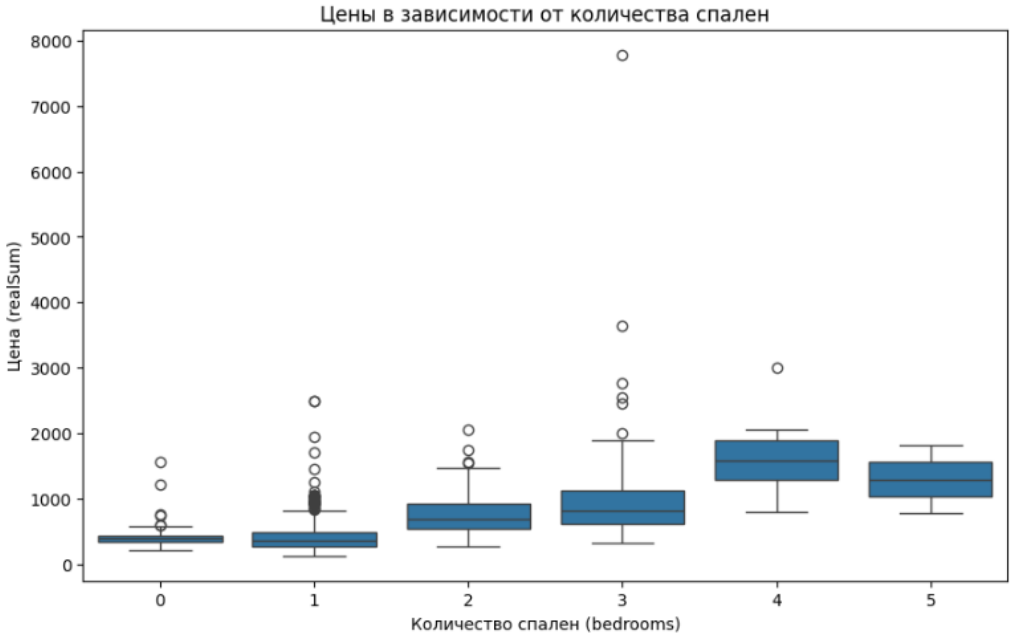

Рис. 5: Листинги с большим количеством спален имеют более высокие цены.

Построили диарамму с усами и видим, что медианное значение растет до 4 спален, но потом становится ниже, но это некритично, также наличие корреляции между переменными 0.49, что говорит о наличии положительной связи между переменными.

3.5 Цены в зависимости от рейтинга удовлетворенности гостей

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='guest_satisfaction_overall', y='realSum', hue='
                                    guest_satisfaction_overall', palette=
                                    'viridis')
plt.title('Prices depending on guest satisfaction rating')
plt.xlabel('Guest Satisfaction rating (guest_satisfaction_overall)')
plt.ylabel('Price (realSum)')
plt.show()
```

Рис. 6: Листинги с высоким рейтингом удовлетворенности гостей имеют более высокие цены.

По точечной диаграмме мы видим положительный тренд и можем сказать, что рейтинг удовлетворенности гостей влияет на цену. Корреляция между переменными 0.14, что говорит о наличии малой положительной линейной зависимости

3.6 Цены в зависимости от типа листинга

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='room_type', y='realSum')
plt.title('Prices depending on the type of listing')
plt.xlabel('Listing type (room_type)')
plt.ylabel('Price (realSum)')
plt.show()

# t-test
entire_home_prices = df[df['room_type'] == 'Shared room']['realSum']
private_room_prices = df[df['room_type'] == 'Private room']['realSum']
t_stat_room_type, p_val_room_type = ttest_ind(entire_home_prices,
                                              private_room_prices)
t_stat_room_type, p_val_room_type

# (-1.1628609255034663, 0.24537858346227584)
```

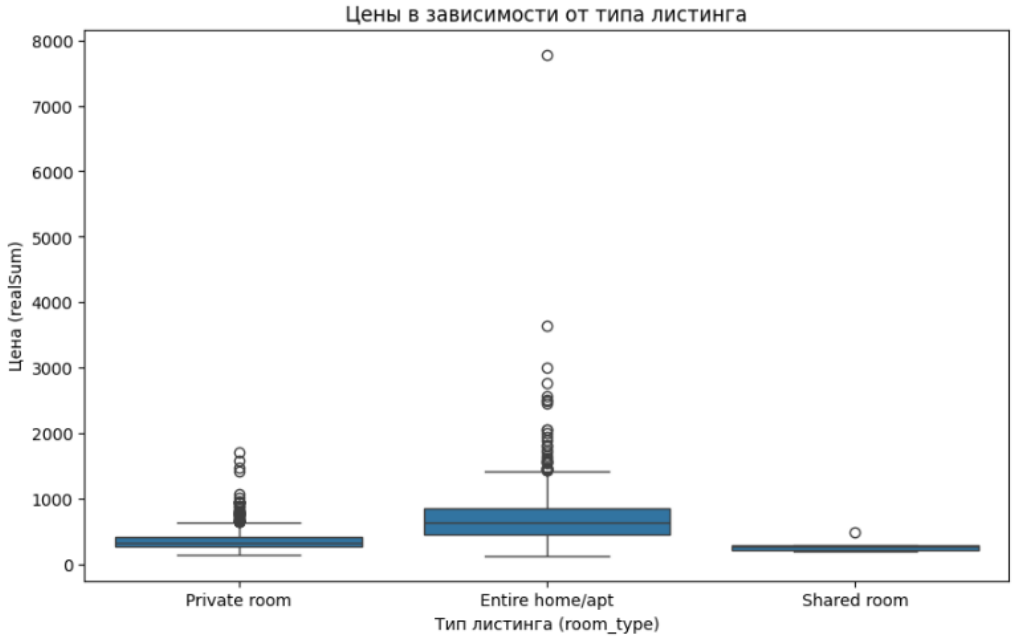

Рис. 7: Листинги, предлагающие целый дом или квартиру, имеют более высокие цены по сравнению с отдельными комнатами.

По диаграмме с усами мы видим, что дом дороже квартиры и отдельных комнат, но проведенный t-test показывает, что нет статистической разницы между квартиры и отдельной комнаты, поэтому можем принять гипотезу частично и сказать, что дом - самый дорой тип листинга

3.7 Цены в зависимости от близости к популярным туристическим местам

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='attr_index', y='realSum', hue='attr_index',
                                        palette='coolwarm')
plt.title('Prices depending on proximity to popular tourist spots')
plt.xlabel('Index of proximity to tourist places (attr_index)')
plt.ylabel('Price (realSum)')
plt.show()
```

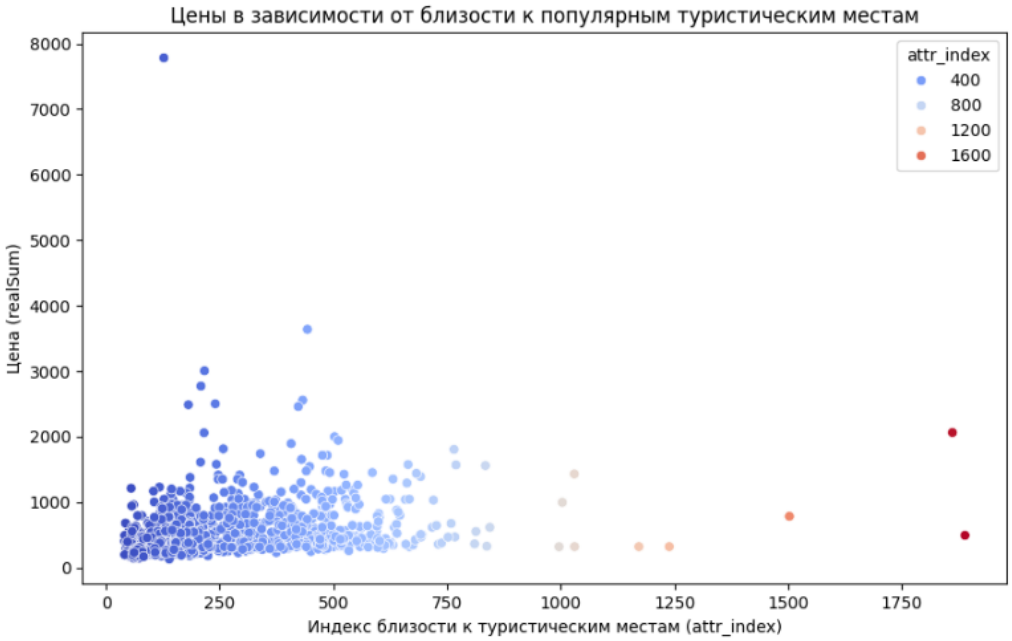

Рис. 8: Листинги, расположенные ближе к популярным туристическим местам, имеют более высокие цены.

По диаграмме видим, что все точки сконцентрированны в основном у маленьких значений индекса, также имеется положительная корелляция 0.24, что говорит наличии положительный связи

Рациональнее будет учесть, что чем меньше индекс близости тем больше цена, но если рассматривать большинство значений, котоый сконцентрированны от 0 до 750, то будем наблюдать положительный тренд(изходя из корреляционного анализа) и цена не будет зависеть от индекса.

3.8 Цены в зависимости от района с высоким уровнем бизнеса

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='biz', y='realSum')
plt.title('Prices depending on the area with a high level of business')
plt.xlabel('High level of business (biz)')
plt.ylabel('Price (realSum)')
plt.show()
```

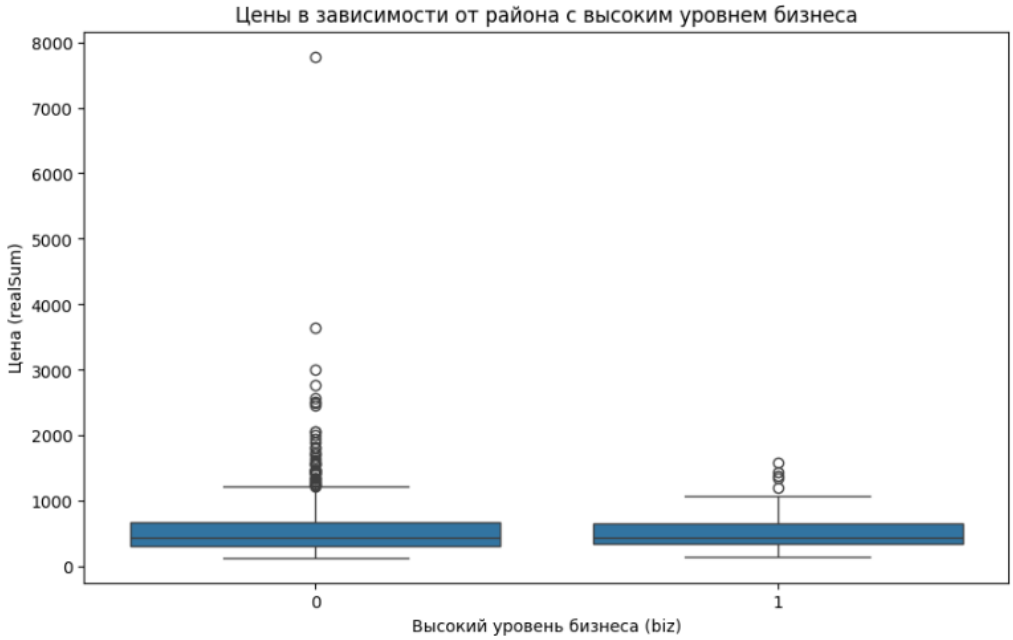

Рис. 9: Листинги, находящиеся в районах с высоким уровнем бизнеса (biz), имеют более высокие цены.

По диаграмме с усами тяжело что-то сказать, но мы проводим т-тест, который показывает p-value 0.37, что говорит о том, что нет статистической разницы между переменными.

3.9 Цены в зависимости от рейтинга чистоты

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='cleanliness_rating', y='realSum')
plt.title('Prices depending on purity rating')
plt.xlabel('Purity rating (cleanliness_rating)')
plt.ylabel('Price (realSum)')
plt.show()

# t-test
median_cleanliness = df['cleanliness_rating'].median()
low_cleanliness_prices = df[df['cleanliness_rating'] < median_cleanliness]['
                                    'realSum']
high_cleanliness_prices = df[df['cleanliness_rating'] >= median_cleanliness
                                    ]['realSum']
t_stat_cleanliness, p_val_cleanliness = ttest_ind(low_cleanliness_prices,
                                    high_cleanliness_prices)
t_stat_cleanliness, p_val_cleanliness
```

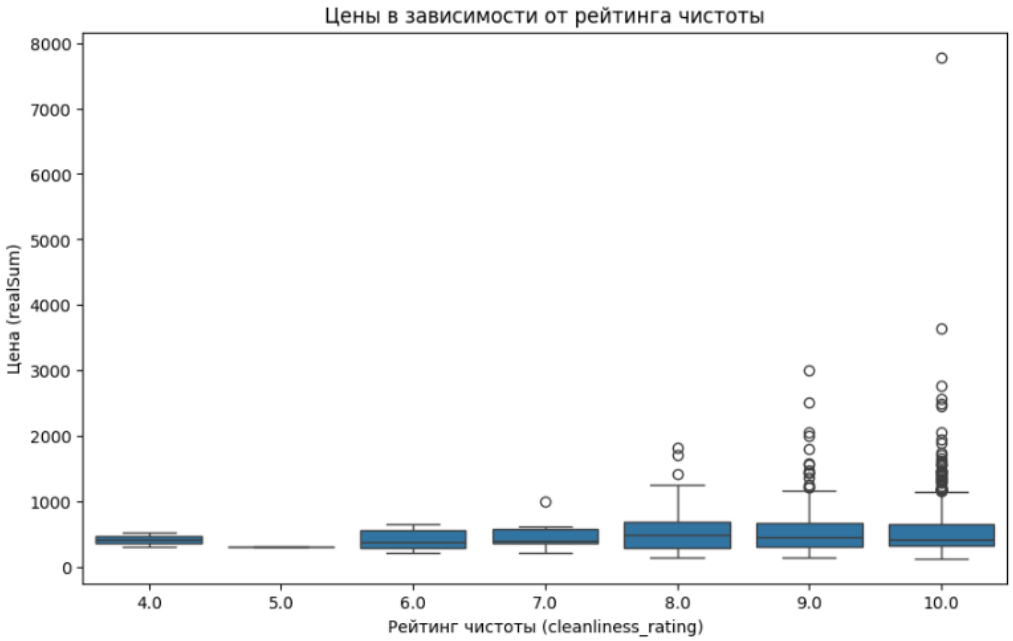

Рис. 10: По диаграмме с усами видим, что значения примерно одинаковые, корреляция близка к 0, проведенный t-test показывает значение 0.8, что говорит нам об отстутствии статистической разницы средних.

По диаграмме с усами видим, что значения примерно одинаковые, корреляция близка к 0, проведенный t-test показывает значение 0.8, что говорит нам об отстутствии статистической разницы средних.

3.10 Цены в зависимости от доступности для бронирования на несколько комнат

```
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='multi', y='realSum')
plt.title('Prices depending on availability for booking multiple rooms')
plt.xlabel('Availability for booking multiple rooms (multi)')
plt.ylabel('Price (realSum)')
plt.show()

# t-test
mult_1 = df[df['multi'] == 0]['realSum']
mult_0 = df[df['multi'] == 1]['realSum']
t_stat_mult, p_val_mult = ttest_ind(mult_1, mult_0)
t_stat_mult, p_val_mult

# (3.502504162873662, 0.0004793932043132782)
```

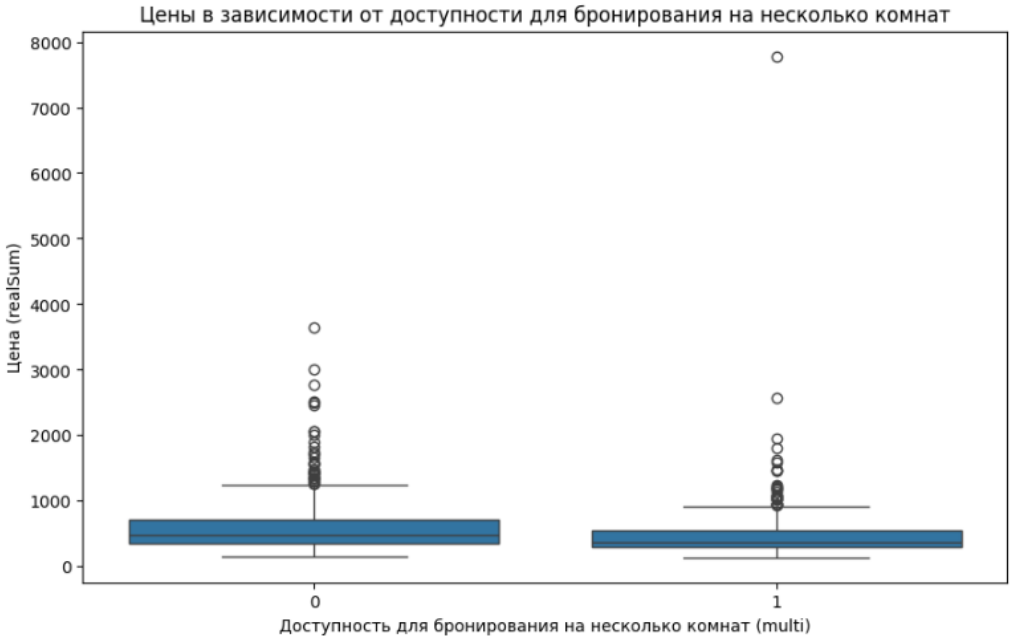

Рис. 11: Листинги, доступные для бронирования на несколько комнат (multi), имеют более высокие цене.

4 Выводы

На цену листинга в основном влияют следующие факторы:

1. Близость к центру города: Чем ближе листинг к центру, тем выше цена.
2. Количество спален: Листинги с большим количеством спален имеют более высокие цены.
3. Тип жилья: Целые дома или квартиры стоят дороже, чем отдельные комнаты.
4. Доступность для бронирования на несколько комнат: Листинги, доступные для бронирования на несколько комнат, имеют более высокие цены.
5. Рейтинг удовлетворенности гостей: чем больше рейтинг, тем дороже листинг.

Список литературы

[1] Airbnb Prices in European Cities — kaggle.com. `https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities?select=amsterdam_weekdays.csv`. [Accessed 05-06-2024].

[2] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

[3] Daniel Guttentag. Progress on airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, 10(4):814–844, 2019.

[4] Jonathan Halket, Lars Nesheim, and Florian Oswald. The housing stock, housing prices, and user costs: The roles of location, structure, and unobserved quality. *International Economic Review*, 61(4):1777–1814, 2020.

[5] Tae Kyun Kim. T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540, 2015.

[6] Prabhaker Mishra, Chandra M Pandey, Uttam Singh, Anshul Gupta, Chinmoy Sahu, and Amit Keshri. Descriptive statistics and normality tests for statistical data. *Annals of cardiac anaesthesia*, 22(1):67–72, 2019.