Exploratory Data Analysis (EDA) Assignment Report

Reemaa Sajad Hyder, Tharanitharan Muthuthirumaran, Tyler Poff

Khoury College of Computer Sciences, Roux Institute at Northeastern University

CS 4160: MACHINE LEARNING

September 25, 2024

HW Assignment 1 (A1): Exploratory Data Analysis (EDA)

Introduction

This assignment explores datasets from Airbnb and performs different types of analysis to identify trends, correlations, and patterns that will be conducive to good business decisions. As a data scientist at an analytics firm, the goal is to find trends in the data that could be used to drive good business decisions in tourism and real estate.

In order to accomplish this, we made use of the two different datasets, namely, Listings and Reviews. Both datasets were chosen for five different cities, resulting in 10 different datasets that needed to be analyzed. The datasets cover the period from the second quarter of 2024, providing information about the current Airbnb market in these tourist destinations. The focus on tourism and real estate formed the basis for our choice of cities. We chose the cities in the United States with the **highest footfall of international tourists all year round**, namely, **New York, Boston, Seattle, Hawaii and San Francisco**.

The main objective was to analyze the data and find the influence of various factors on the price of a property and the likelihood of a listing being popular. This analysis will provide valuable insights for various stakeholders, including Airbnb hosts, property investors, and tourism boards, enabling them to make data-driven decisions in their respective fields.

Analysis

1. Descriptive Statistics

The first task was focused on descriptive statistics. The goal was to understand the central tendency, dispersion and distribution of the various numerical variables in the Listings datasets for the various cities.

The relevant numerical attributes from the dataset were identified and the central tendency of the data was measured by means of mean, median and mode.
The selected attributes are ‘price', 'minimum_nights', 'maximum_nights', 'number_of_reviews', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value' and 'reviews_per_month'.

The distribution of the data can be better visualized by means of box plot distribution. The distribution varied from one dataset to another. Some of the datasets had a large number of entries in the interquartile range which could be easily visualized in the bar plot.

The individual variations in some of the analyzed metrics are as follows:

I. Price

The mean prices vary significantly across cities:

- Boston: $232.19

- Hawaii: $314.32
- New York: $211.62
- Seattle: $207.88
- San Francisco: $242.88.

Hawaii has the highest average price, while Seattle has the lowest.

However, there is a large difference between mean and median prices for all the cities, indicating skewed distributions with some high-priced outliers.

II. Minimum Nights

Mean minimum nights:

- Boston: 16.56
- Hawaii: 6.41
- New York: 25.19
- Seattle: 10.73
- San Francisco: 16.78.

New York has the highest average minimum stay requirement, while Hawaii has the lowest.

The median values are much lower (often 2-3 nights), suggesting that most listings have short minimum stays, but some have very long requirements, skewing the mean.

III. Maximum Nights

Mean maximum nights:

- Boston: 539.64
- Hawaii: 570.04
- New York: 475.87
- Seattle: 446.55
- San Francisco: 559.02

All cities have high maximum night averages, often around a year or more. This suggests that many listings allow long-term stays, thus catering to both short-term tourists and longer-term renters.

IV. Number of Reviews

Seattle has the highest average number of reviews, potentially indicating higher booking frequency or longer listing history.

V. Review Scores Rating

Review scores are consistently high across all cities, with San Francisco having the highest average rating.

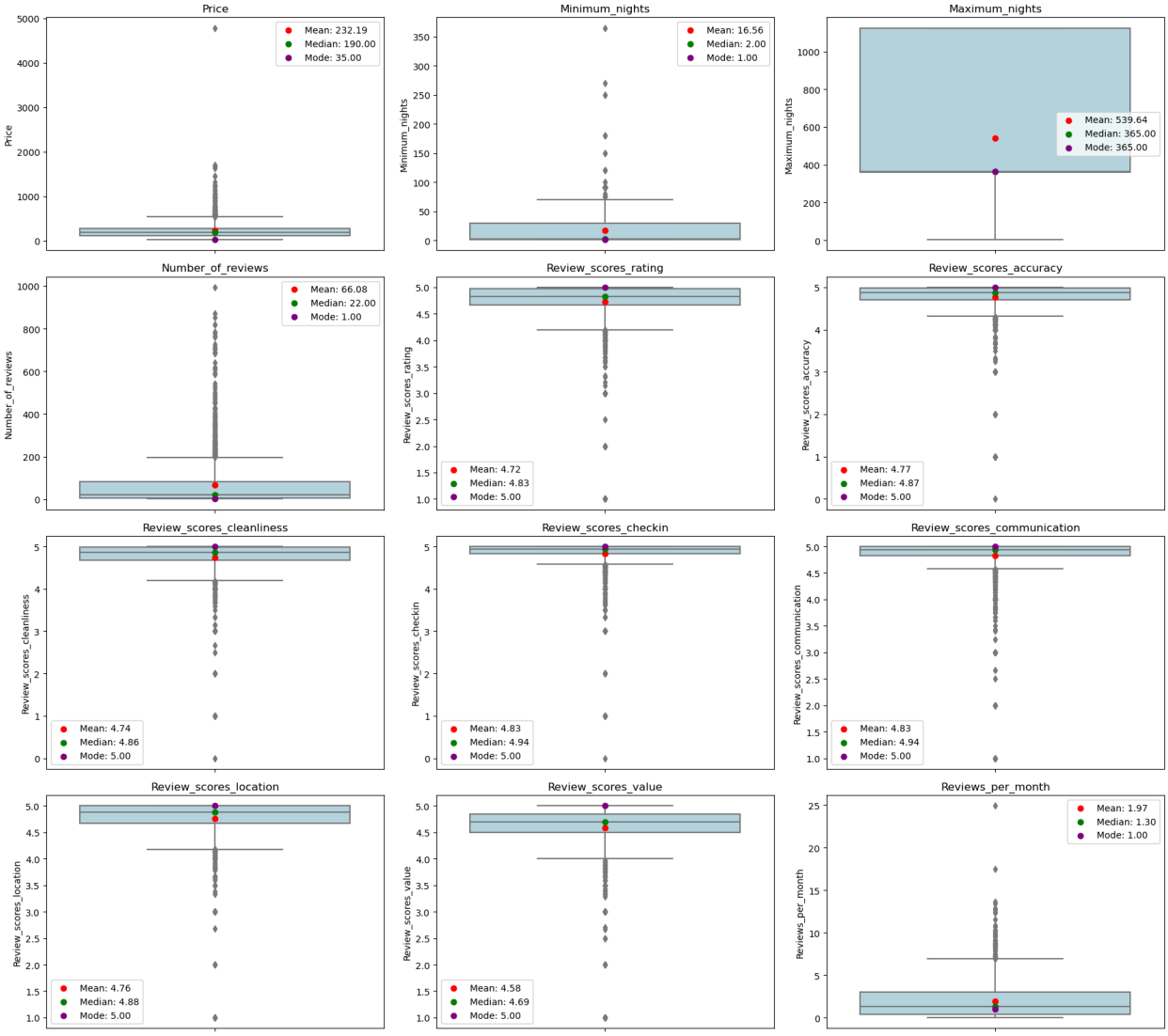

Fig 1: Box plot for Boston

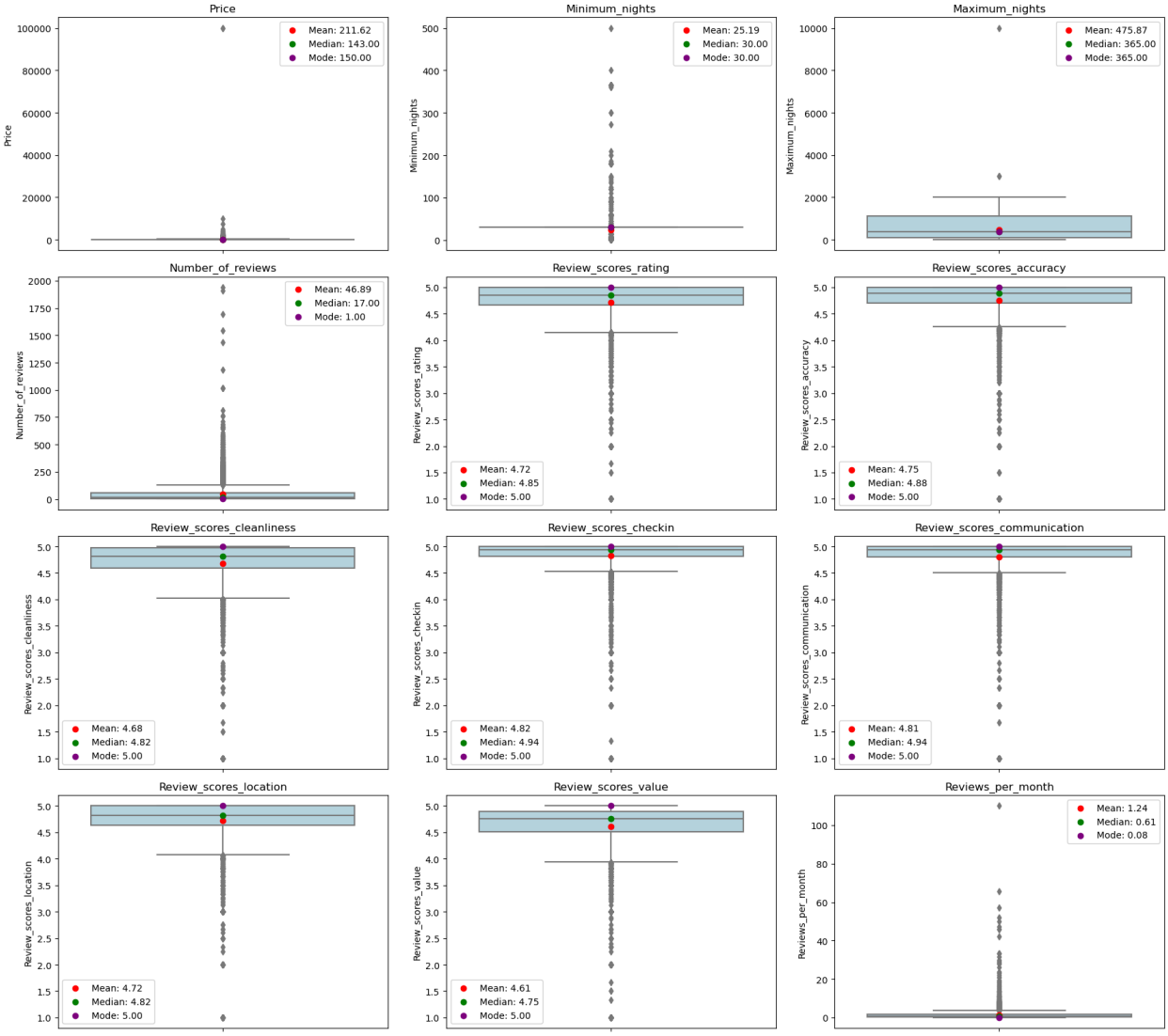

Fig 2: Box plot for New York

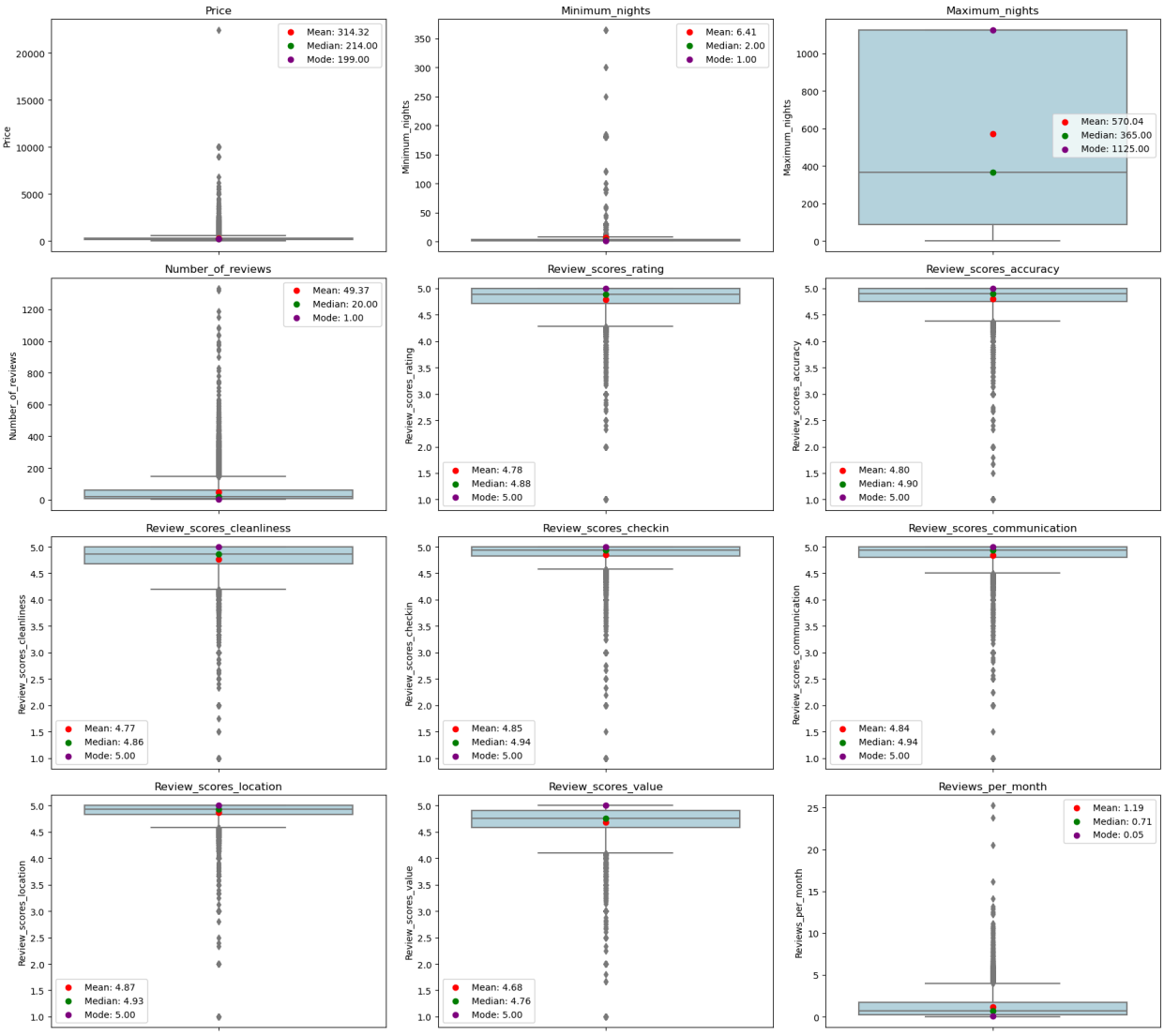

Fig 3: Box plot for Hawaii

The dispersion of the data was measured by means of variance, standard deviation, range and InterQuartileRange (IQR). This was visualized using a violin plot. Violin plots provide a comprehensive representation of the data distribution for the selected metrics. They combine box plots and kernel density plots, thus providing a nuanced view of the data shape, central tendency and variability.

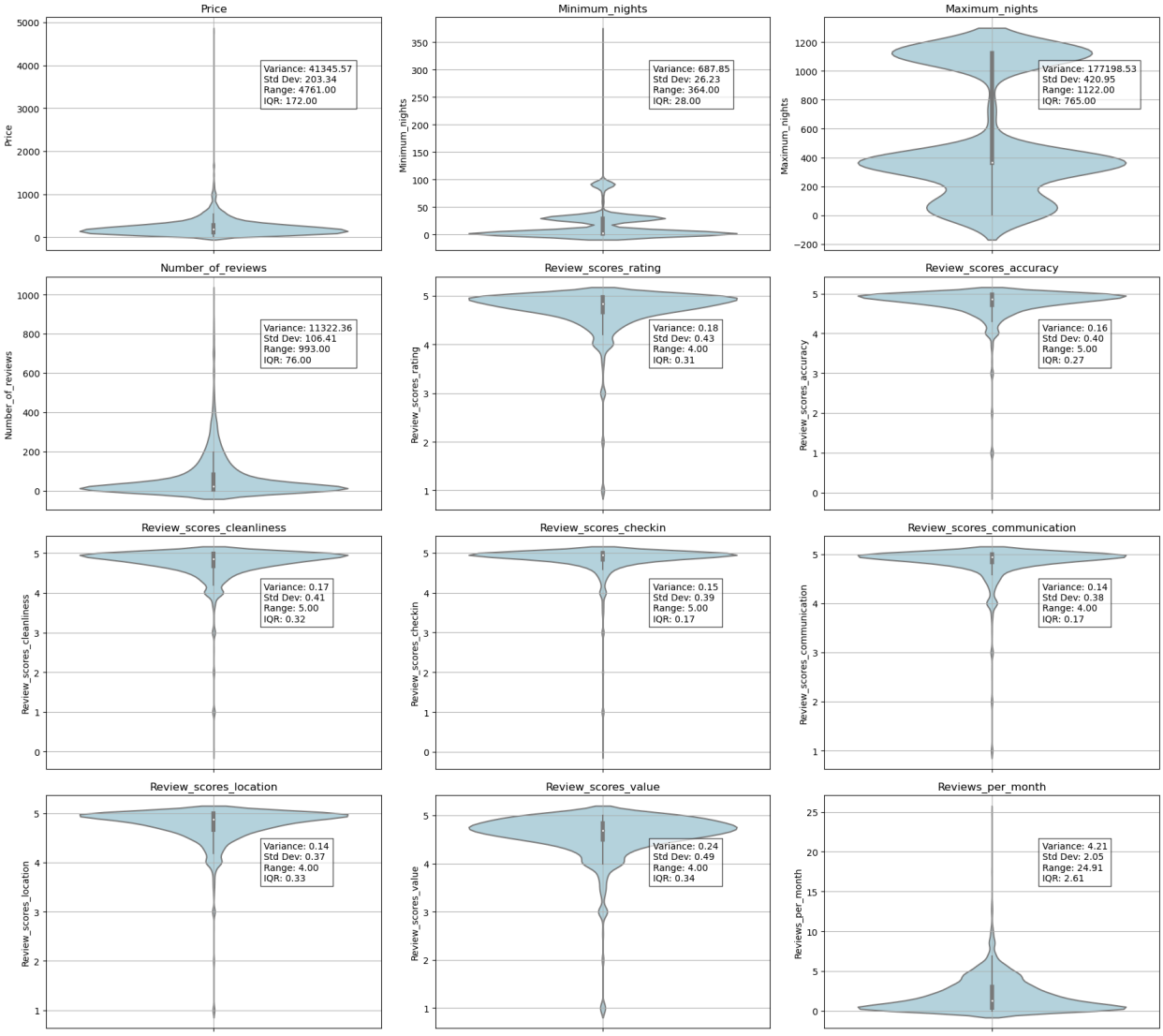

Figure 4: Violin Plot for Boston

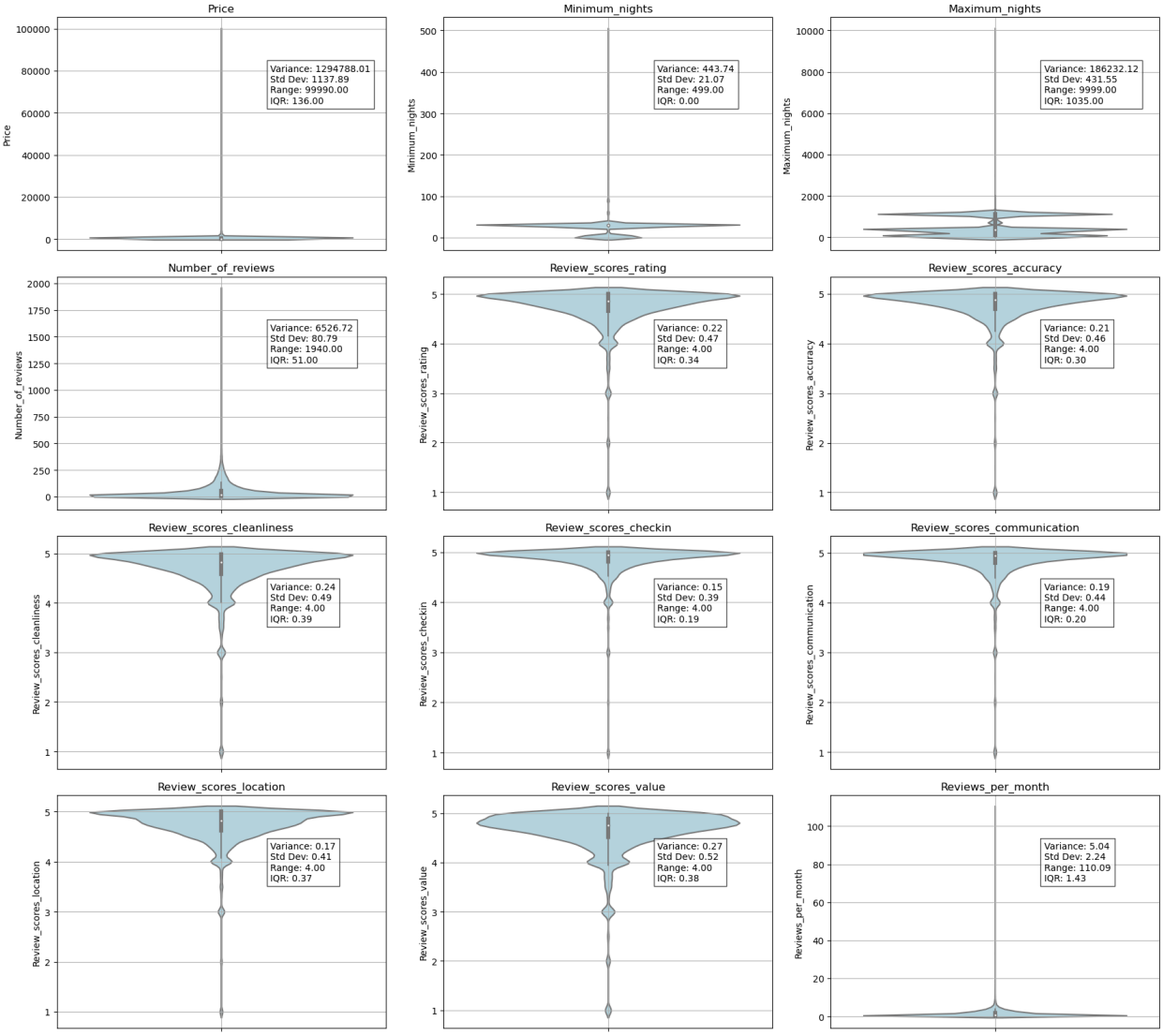

Fig 5: Violin Plot for New York

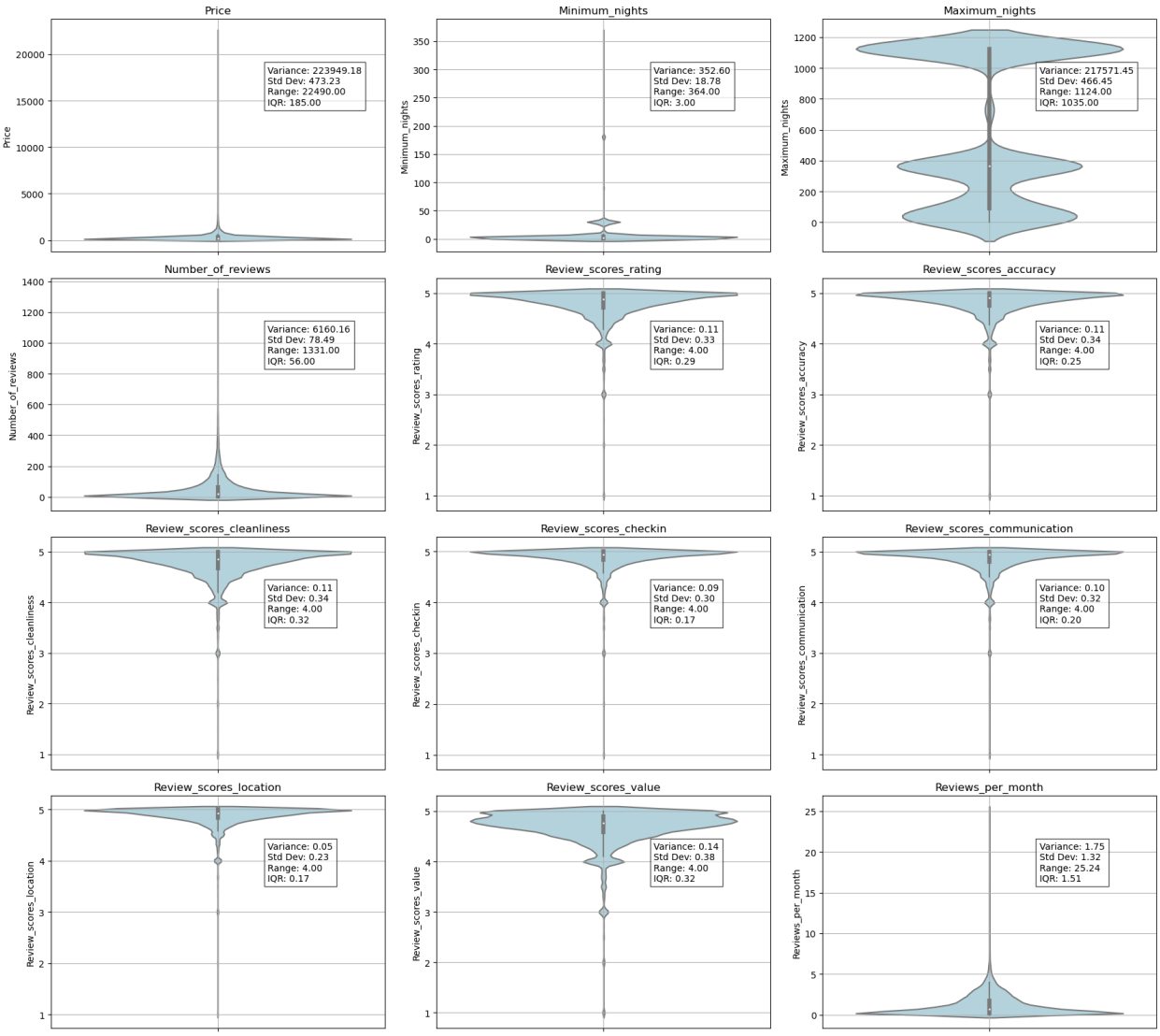

Fig 6: Violin Plot for Hawaii

Key analysis of results:

I. Price:

All cities show highly right-skewed price distributions with long upper tails. Boston, New York, and Seattle have similar price distributions, with New York showing slightly higher prices. San Francisco has the most extreme outliers, with some listings **priced up to $50,000.**

II. Minimum Nights:

Most cities have a strong concentration of listings with short minimum stays (1-3 nights). **New York shows a notably different distribution, with a higher median and more listings requiring longer minimum stays.** All cities have outliers with very high minimum night requirements.

III. Maximum Nights:

Distributions are multimodal across all cities, **with peaks around 30 days**, 365 days, and 1125 days. This suggests different categories of listings: short-term, month-long, and long-term stays.

IV. Number of Reviews:

All cities show right-skewed distributions. Seattle has the highest median number of reviews, potentially indicating higher booking frequency or longer listing history.

V. Review Scores:

All cities show left-skewed distributions for various review score categories (rating, accuracy, cleanliness, etc.). **Median scores are consistently high (above 4.5 out of 5)** across all categories and cities.

VI. Reviews per Month:

All cities show right-skewed distributions with long upper tails. New York has the highest variance in reviews per month, suggesting more diverse listing activity.

The violin plots reveal that while there are general similarities across cities (e.g., high review scores, right-skewed price distributions), there are also notable differences in pricing strategies, minimum stay requirements, and review patterns.

Overall, while there are similarities in distribution shapes across cities (e.g., right-skewed prices, left-skewed review scores), there are notable differences in central tendencies and extreme values, reflecting the unique characteristics of each city's Airbnb market.

2. Distribution Analysis

Histograms are plotted for the key numerical features for all 5 cities.

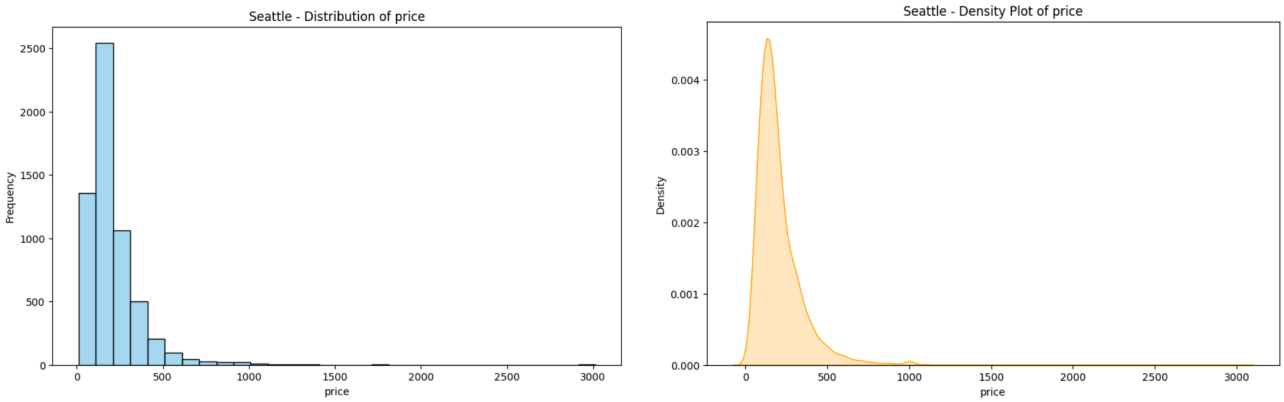

Fig 7: Histogram and density plot for price for Seattle

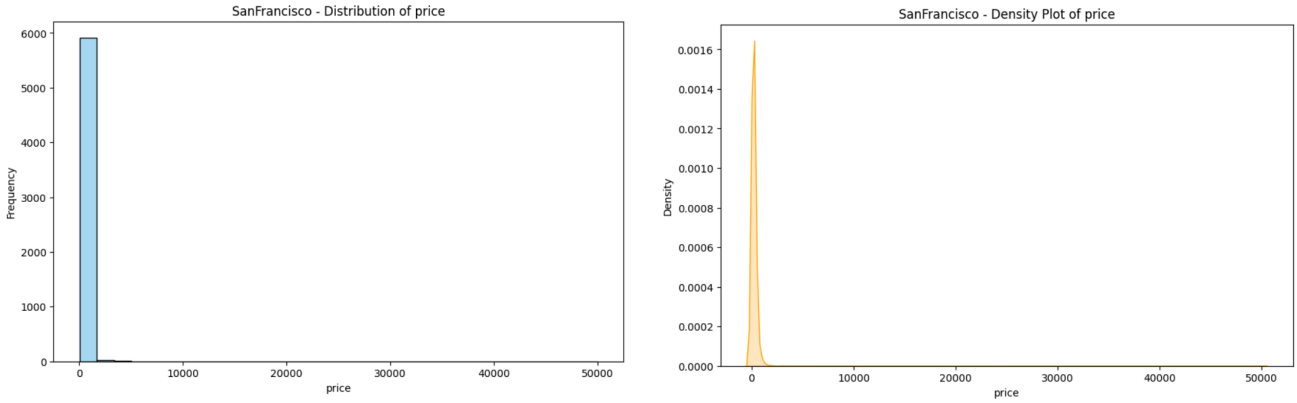

Fig 8: Histogram and Density Plot for Price for San Francisco

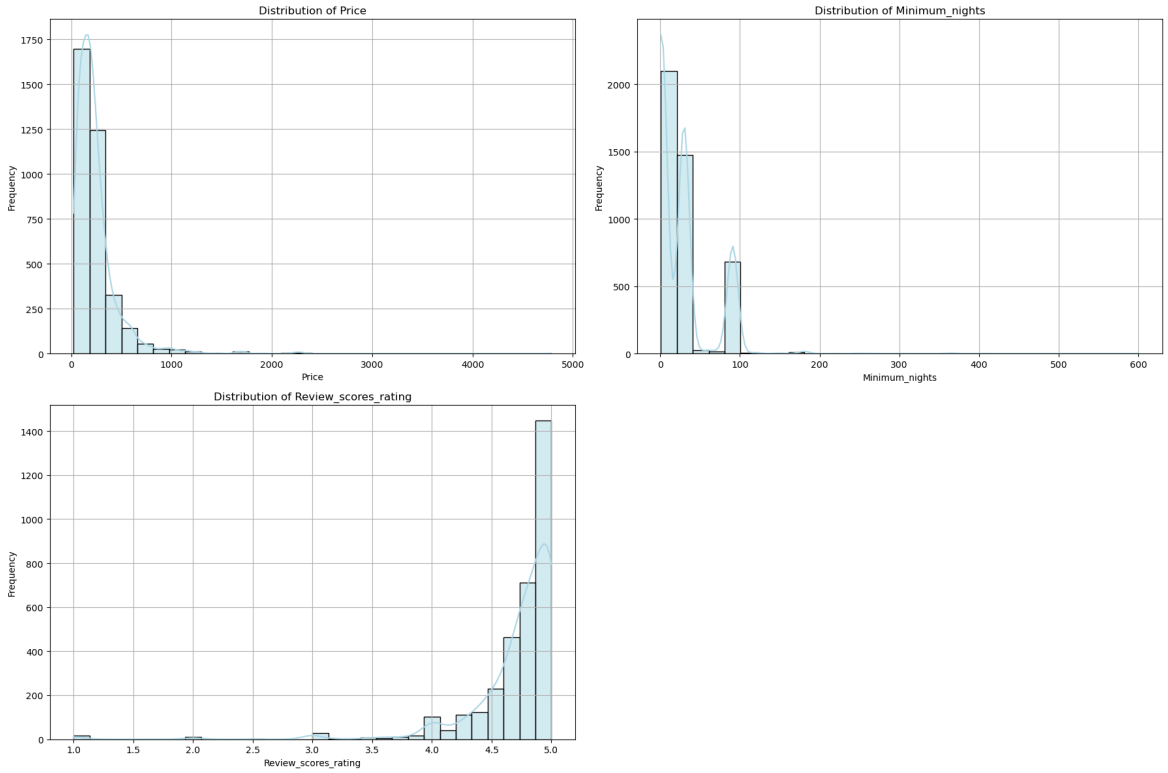

Fig 9: Histogram for Price, Minimum Nights and Review Scores Rating for Boston

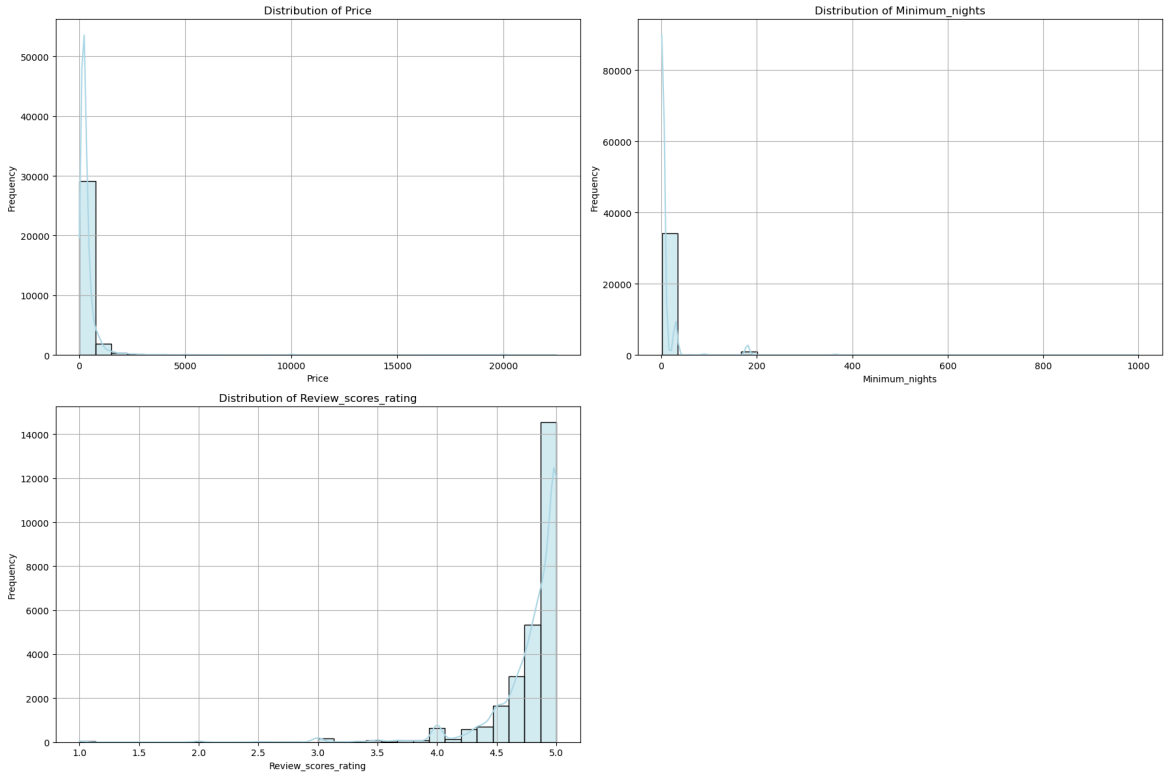

Fig 10: Histogram for Price, Minimum Nights and Review Scores Rating for Hawaii

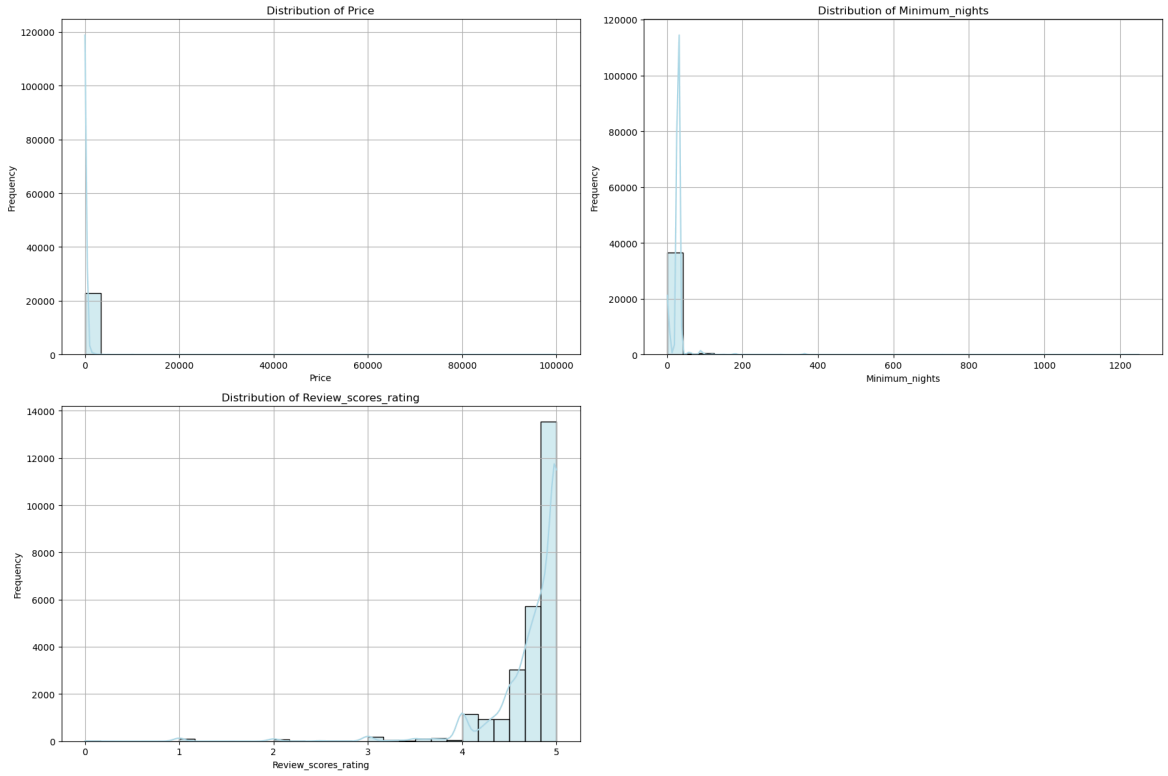

Fig 11: Histogram for Price, Minimum Nights and Review Scores Rating for New York

Key analysis of results:

I. Price

All cities show a right-skewed (positively skewed) distribution for prices, meaning that there are many listings at lower price points and fewer listings at higher price points. Likewise, all cities show the highest frequency of listings at the lower end of the price spectrum, **typically under $200-$300 per night.**

Each city's distribution has a long right tail, indicating the presence of some very high-priced listings. This is particularly pronounced in San Francisco.

II. Minimum Nights

The distribution of minimum nights is heavily right-skewed across all cities, with a very long tail. Most cities show a range from 0 to about 600 nights, with some extreme outliers.

All cities show a **long tail extending to higher minimum night values**, but the frequency drops off after the initial peak. There are visible outliers in all cities, with **some listings having very high minimum night requirements (up to 1000+ nights in some cases).**

There's a small segment of listings geared towards longer-term rentals, possibly monthly or yearly leases. The extreme outliers might represent errors in data entry or listings with special circumstances.

III. Review Scores Rating

The distribution of review scores is heavily left-skewed (negatively skewed) across all cities. The vast majority of ratings are clustered between 4.0 and 5.0, with a peak at the end of the scale, around 4.5 - 5.0. There is a very small number of listings with low ratings (1.0-3.0), visible as small bars on the left side of the histograms.

The small number of low-rated listings could represent new hosts, problematic properties, or listings that haven't been removed yet.

3. Correlation Analysis

Correlation analysis is performed by plotting the correlation matrix between the required variables. In addition to the existing numerical variables, we

perform encoding and convert some of the **categorical variables such as Neighbourhood Group, Room Type and Amenities to numerical values using One Hot Encoding** to check if there is a correlation between these factors and price. The correlation matrices for the various cities are as follows

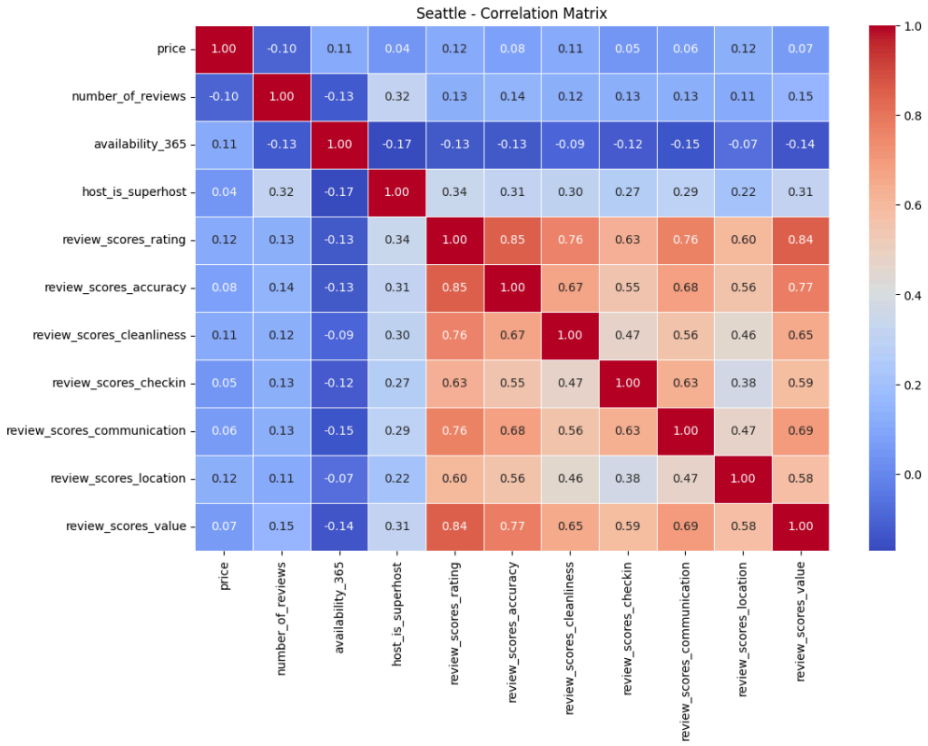

Fig 12: Heat map for correlation matrix for Seattle

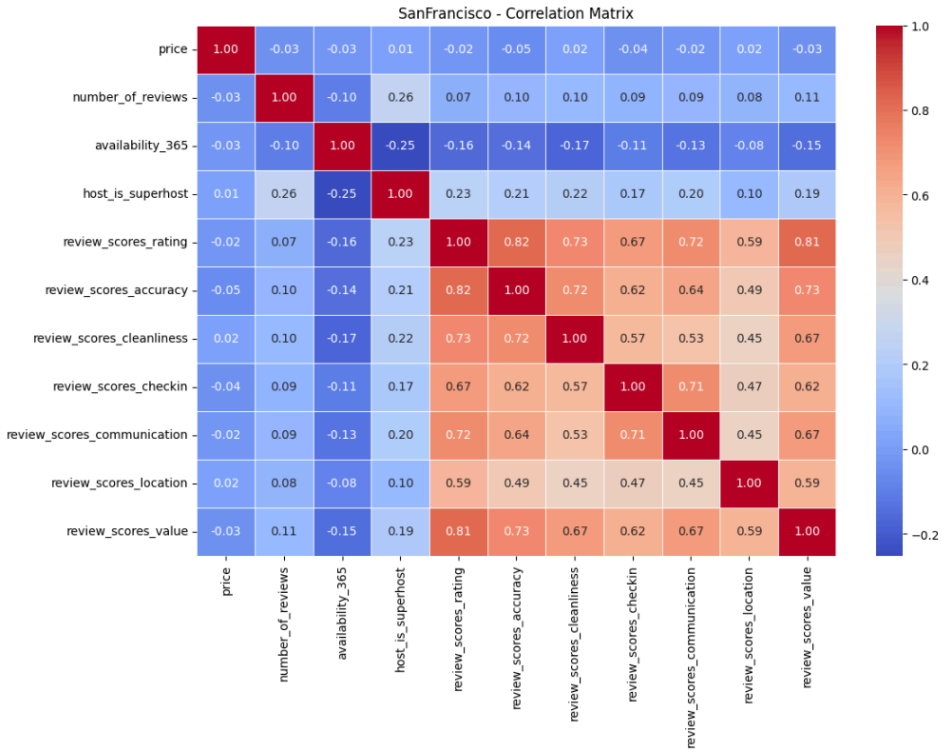

Fig 13: Heat map for correlation matrix for San Francisco

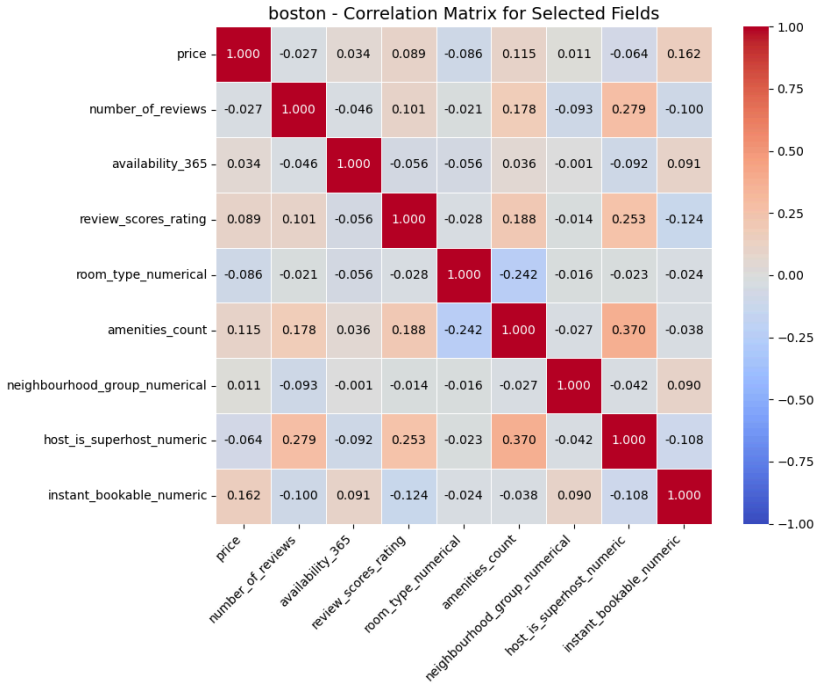

Fig 14: Heat map for correlation matrix for Boston

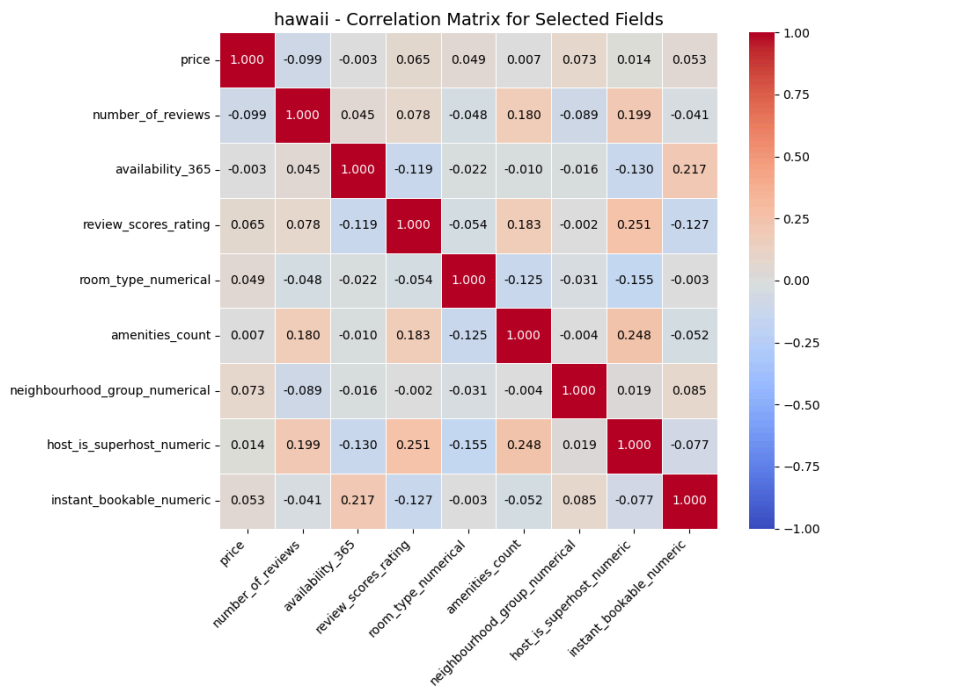

Fig 15: Heat map for correlation matrix for Hawaii

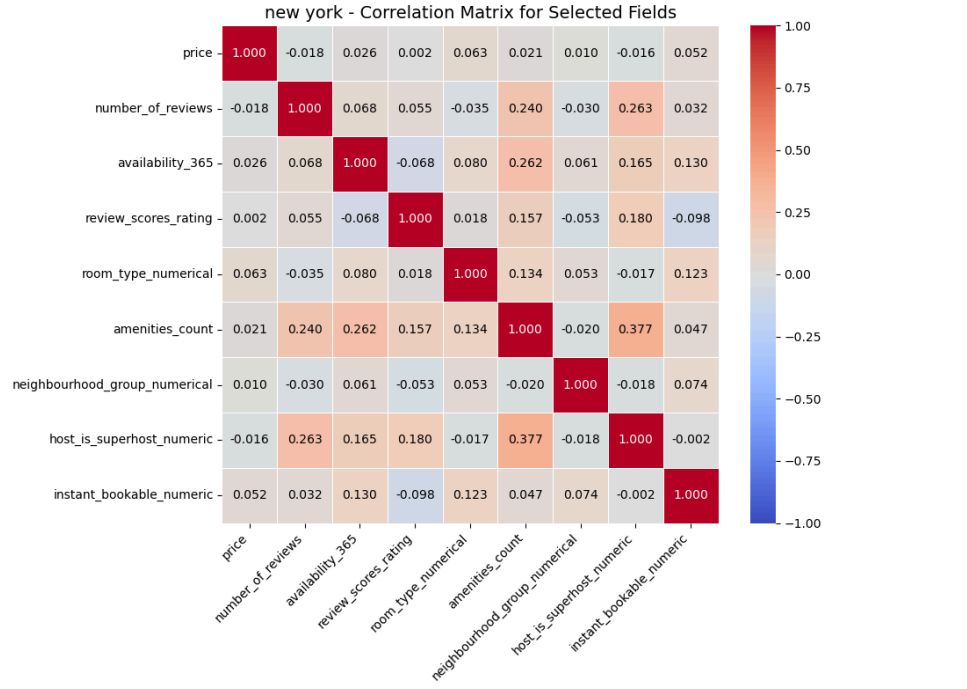

Fig 16: Heat map for correlation matrix for New York

Analysis of heat maps for some of the prominent attributes

I. **Price**

   Across all cities, the price attribute has weak correlations with most other variables.

II. **Review Score Related Attributes**

   Strong positive correlations exist between different review score categories (cleanliness, accuracy, communication, etc.) across all cities.

III. **Superhost Status**

   Moderately positive correlations exist between superhost status and review scores in all cities (ranging from 0.2 to 0.3)

IV. **Availability 365**

   Consistently shows weak negative correlations with review scores and superhost status across all cities.

V. **Number of reviews**

   Weak to moderate positive correlation with host_is_superhost in most cities (0.1 to 0.3). This suggests that superhosts tend to have more reviews.

Overall, the heat maps reveal that while there are some correlations between variables, most are weak to moderate.

4. **Price Analysis**

   We plot the bar graph for the mean price of listings grouped by neighborhood for each city.

   We make use of the `neighbourhood_groups_cleansed` attribute to group the different neighborhoods together. 'neighbourhood_groups_cleansed'

provides a more manageable level of geographic division, offering a balance between detail and generalization. This larger groupings, ensuring more listings per category and increasing the reliability of average price calculations.

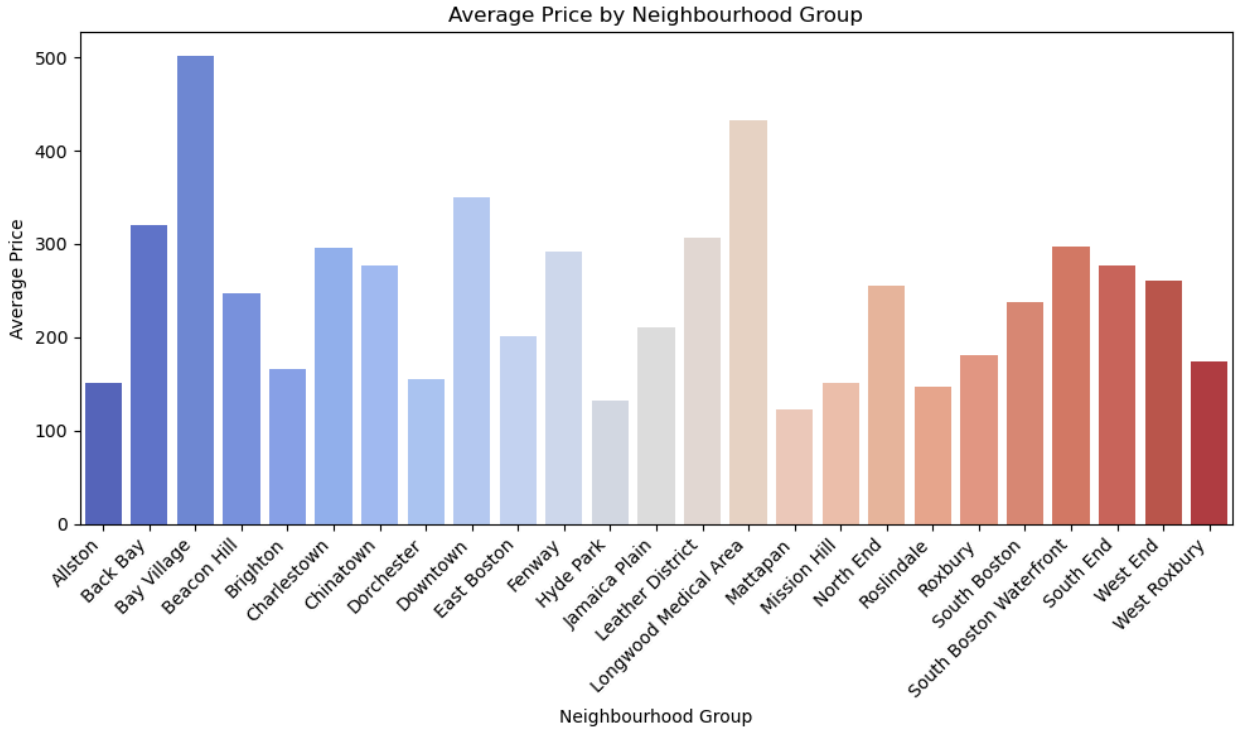

Fig 17: Price distribution by neighborhood - Boston

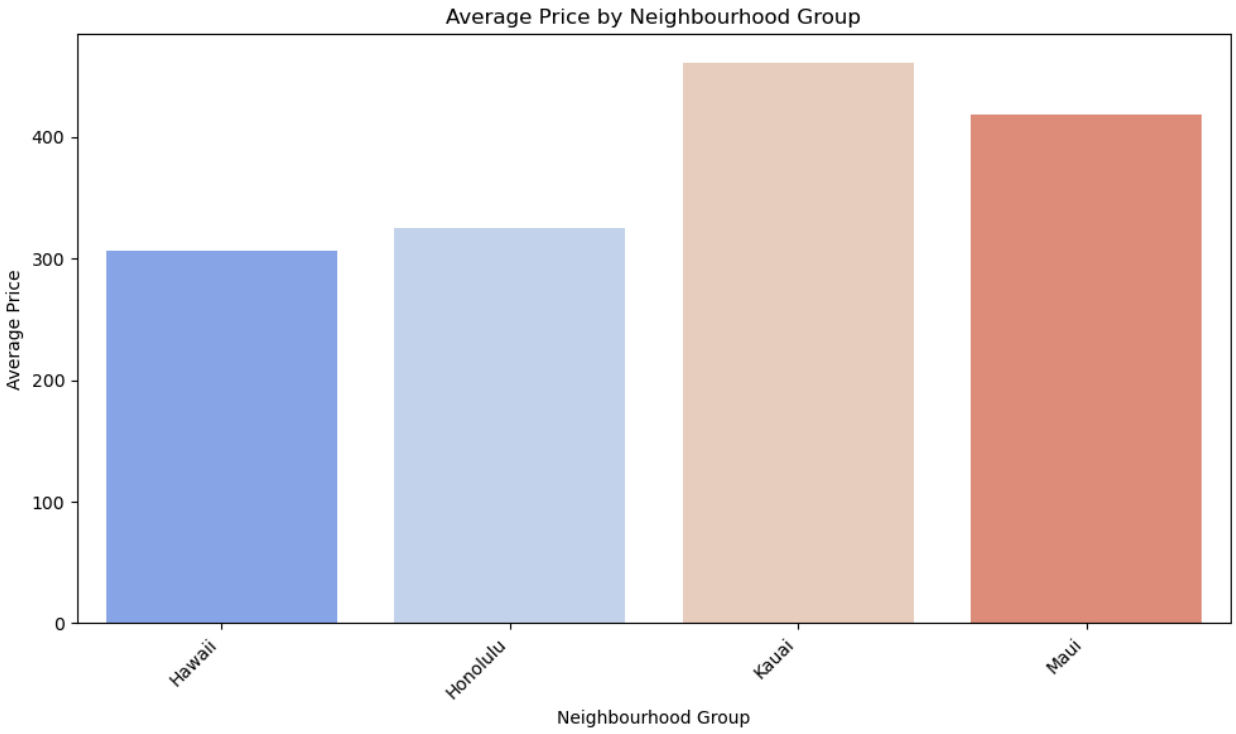

Fig 18: Price distribution by neighborhood - Hawaii

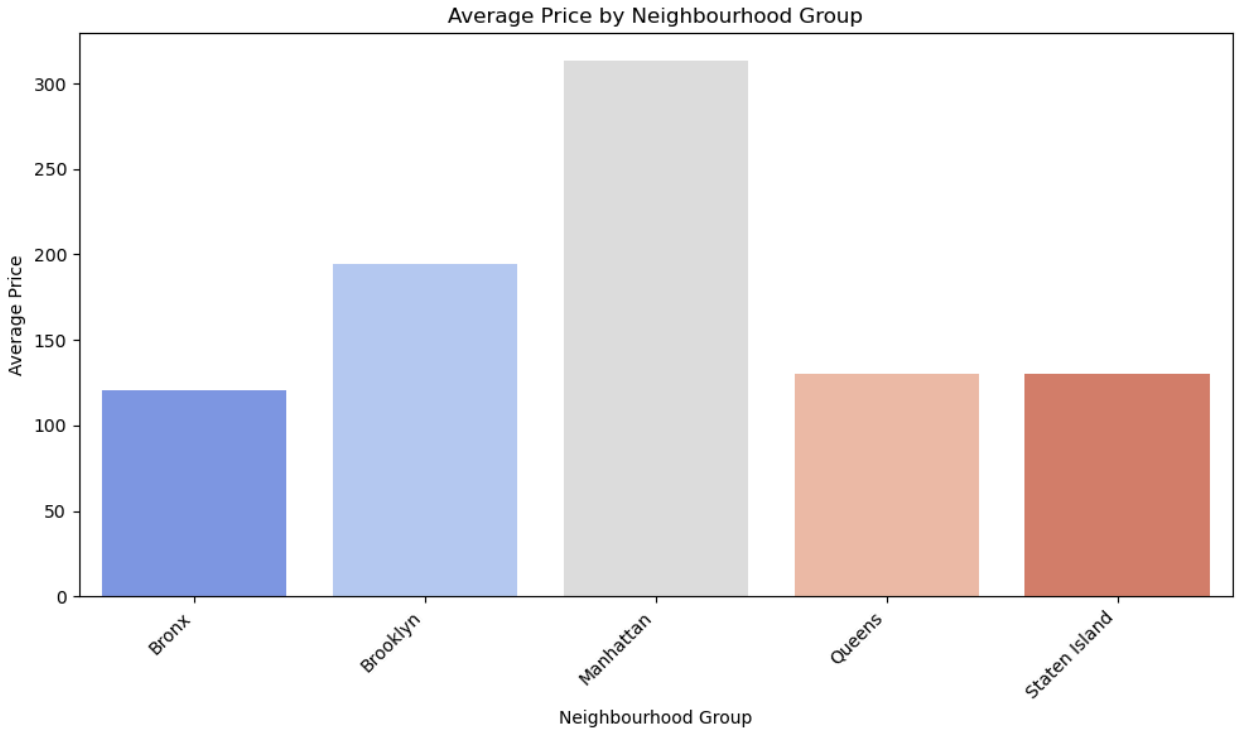

Fig 19: Price distribution by neighborhood - New York

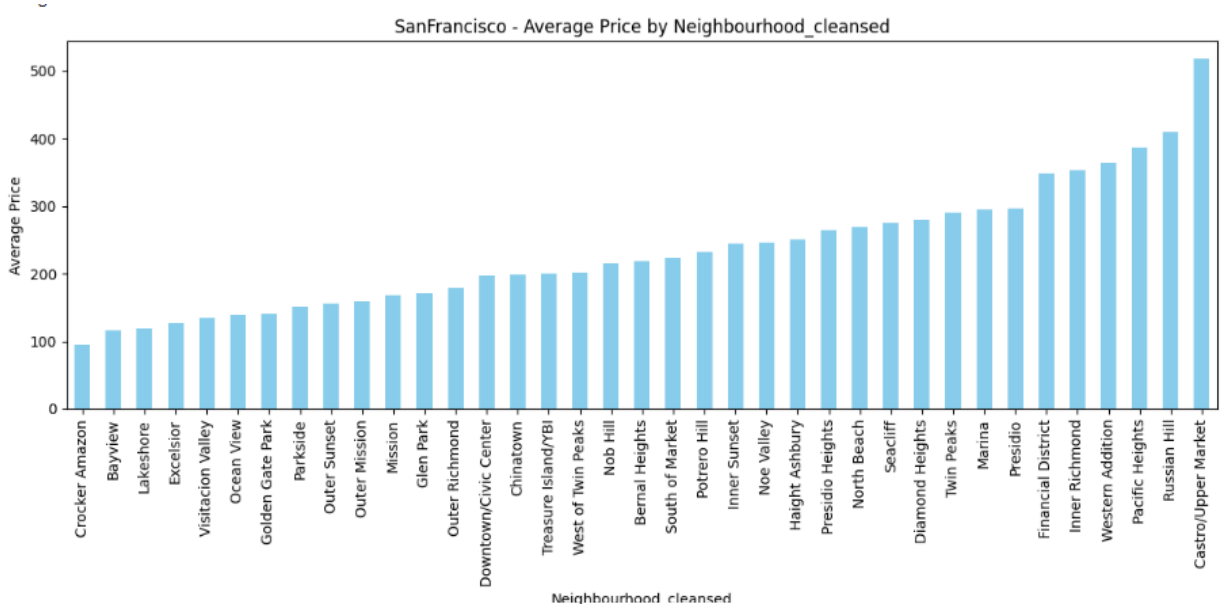

Fig 20: Price distribution by neighborhood - San Francisco

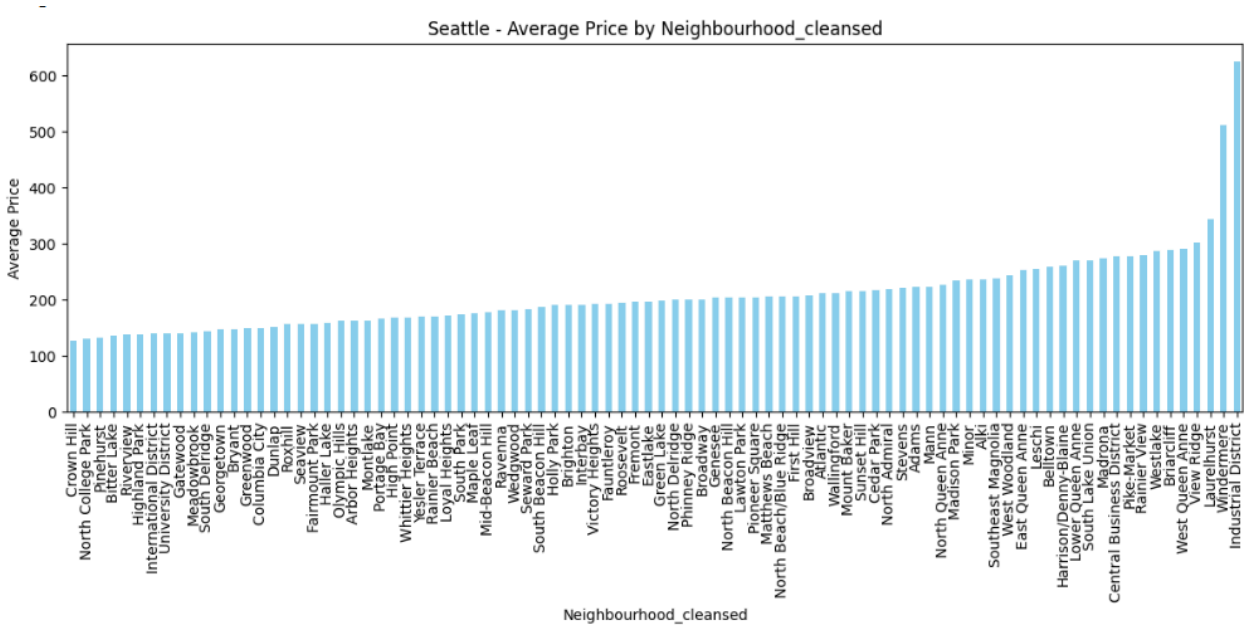

Fig 21: Price distribution by neighborhood - Seattle

Analysis of price distribution:

Among the cities analyzed, Manhattan in New York City consistently stands out as the most expensive neighborhood group. Similarly, Kauai in Hawaii has the highest mean price. Within each city, there's a wide range of prices between neighborhoods, indicating significant local variations.

Distribution of prices across room types

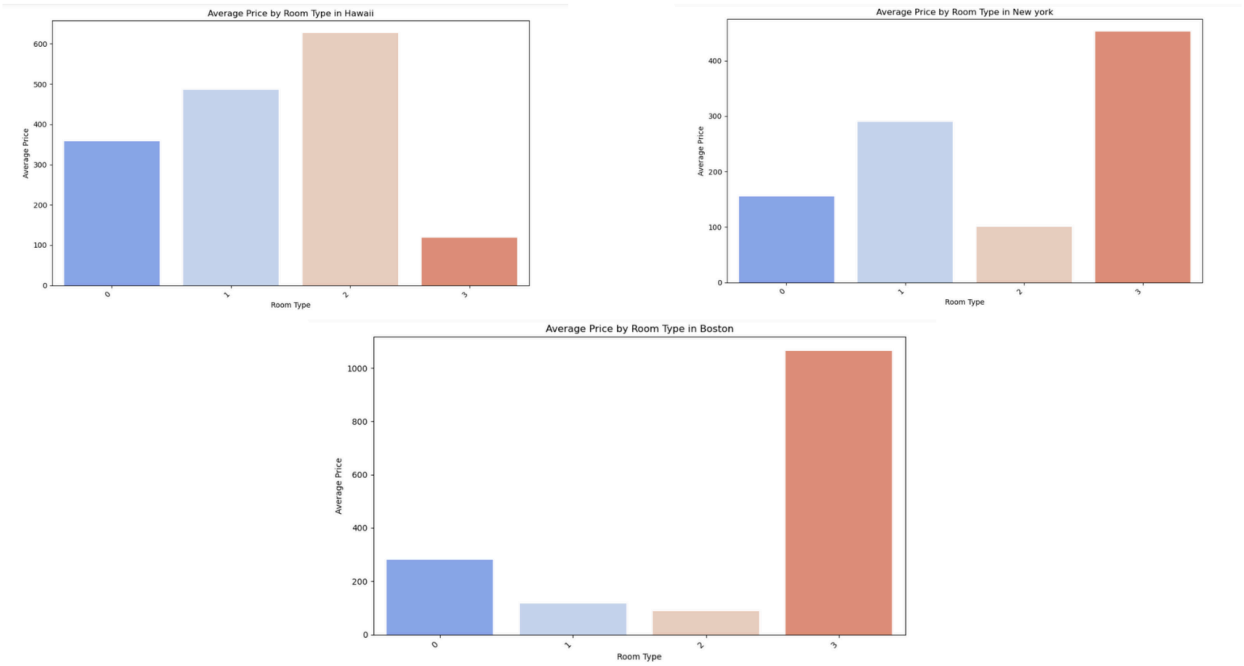

Fig 22: Price distribution by Number of Rooms

The magnitude of price increase varies between locations. New York and Boston exhibit more significant price jumps between room types compared to Hawaii. The distribution of price across different neighborhoods can be more effectively visualized as heat maps plotted with latitude and longitude as follows:

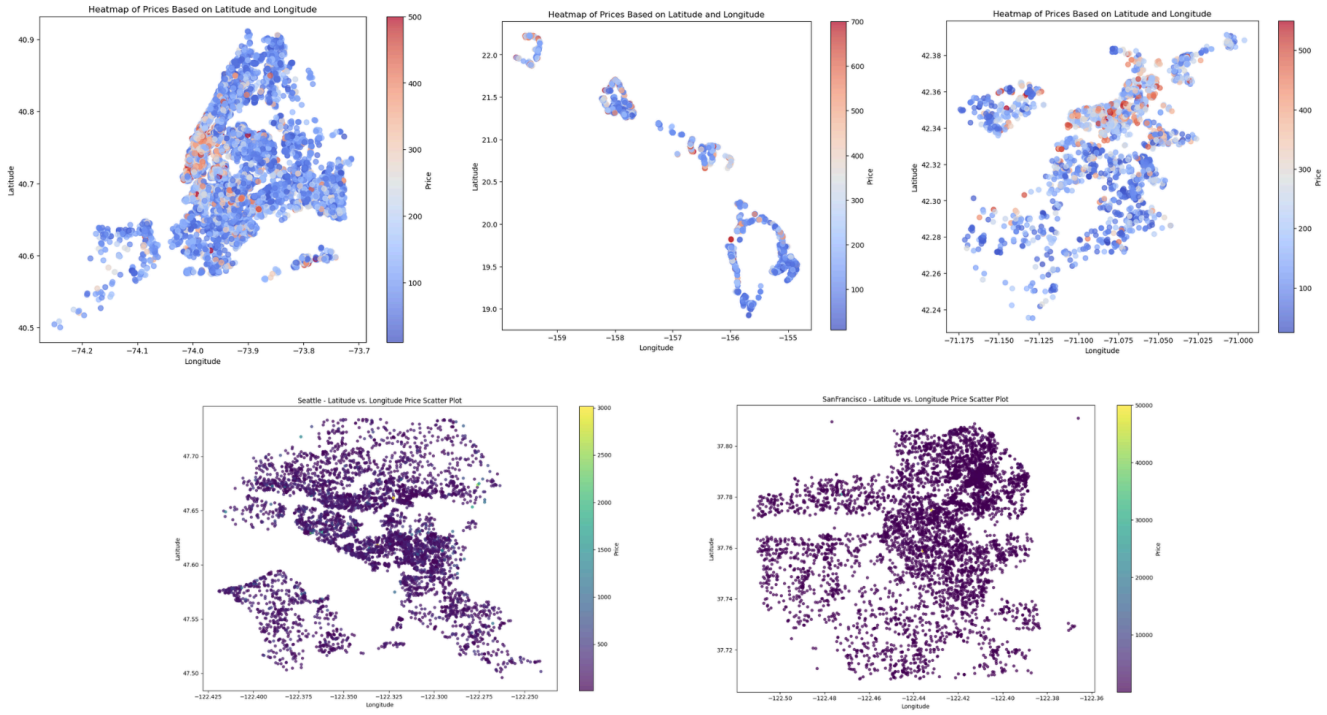

Fig 23: Price distribution across different neighborhoods in 5 cities

In order to find whether certain neighborhoods are more popular for short term or long term stays, we find the mean minimum_stay value associated with each neighborhood group for the 5 cities. Next, based on minimum_stay mean value, we classify each neighborhood based on the preferred stay term as long term or short term stay. For all 5 cities, the predominant stay type for the majority of the neighborhoods was Short Term.

stay_type	Long Term	Medium Term	Short Term	Predominant Stay Type
neighbourhood_group_cleansed				
Bronx	6	14	858	Short Term
Brooklyn	30	299	7648	Short Term
Manhattan	84	1006	8902	Short Term
Queens	14	92	3520	Short Term
Staten Island	3	4	310	Short Term

Table 1: Predominant Stay Type by neighborhood - New York

stay_type	Long Term	Medium Term	Short Term	Predominant Stay Type
neighbourhood_cleansed				
Allston	45.0	12.0	63.0	Short Term
Back Bay	12.0	57.0	266.0	Short Term
Bay Village	1.0	14.0	63.0	Short Term
Beacon Hill	6.0	19.0	148.0	Short Term
Brighton	36.0	51.0	148.0	Short Term
Charlestown	2.0	4.0	59.0	Short Term
Chinatown	4.0	17.0	13.0	Short Term
Dorchester	65.0	20.0	351.0	Short Term
Downtown	12.0	80.0	237.0	Short Term
East Boston	6.0	1.0	146.0	Short Term
Fenway	23.0	43.0	160.0	Short Term
Hyde Park	4.0	0.0	45.0	Short Term
Jamaica Plain	19.0	5.0	124.0	Short Term
Leather District	1.0	2.0	5.0	Short Term
Longwood Medical Area	0.0	0.0	7.0	Short Term
Mattapan	1.0	1.0	42.0	Short Term
Mission Hill	16.0	4.0	40.0	Short Term
North End	17.0	12.0	73.0	Short Term
Roslindale	2.0	1.0	68.0	Short Term
Roxbury	48.0	23.0	201.0	Short Term
South Boston	13.0	26.0	129.0	Short Term
South Boston Waterfront	1.0	23.0	20.0	Short Term
South End	9.0	26.0	245.0	Short Term
West End	0.0	24.0	15.0	Short Term
West Roxbury	8.0	1.0	58.0	Short Term

Table 2: Predominant Stay Type by neighborhood - Boston

stay_type	Long Term	Medium Term	Short Term	Predominant Stay Type
neighbourhood_group_cleansed				
Hawaii	4	53	7754	Short Term
Honolulu	33	109	8850	Short Term
Kauai	303	1	4935	Short Term
Maui	369	2	9350	Short Term

Table 3: Predominant Stay Type by neighborhood - Hawaii

5. Neighborhood Comparison

We plot a heat map of the review_scores_rating against latitude and longitude.

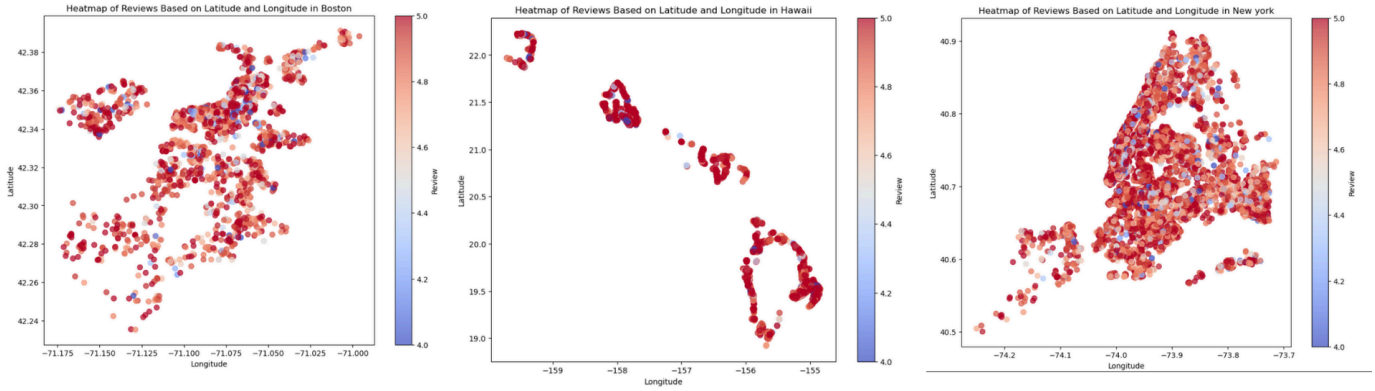

Fig 24: Review scores rating distribution across different neighborhoods in Boston, Hawaii and New York

From the heatmap, it is visible that the majority of review scores are high for all the cities. In order to quantify this, we classify the reviews as Top Review (Review Score > 4.75), Good Review and Low Review (Review Score < 3) and plot a bar graph with the percentage of reviews in each category. For all five cities, there are a negligible number of low reviews and a vast majority of the reviews are positive for each neighborhood.

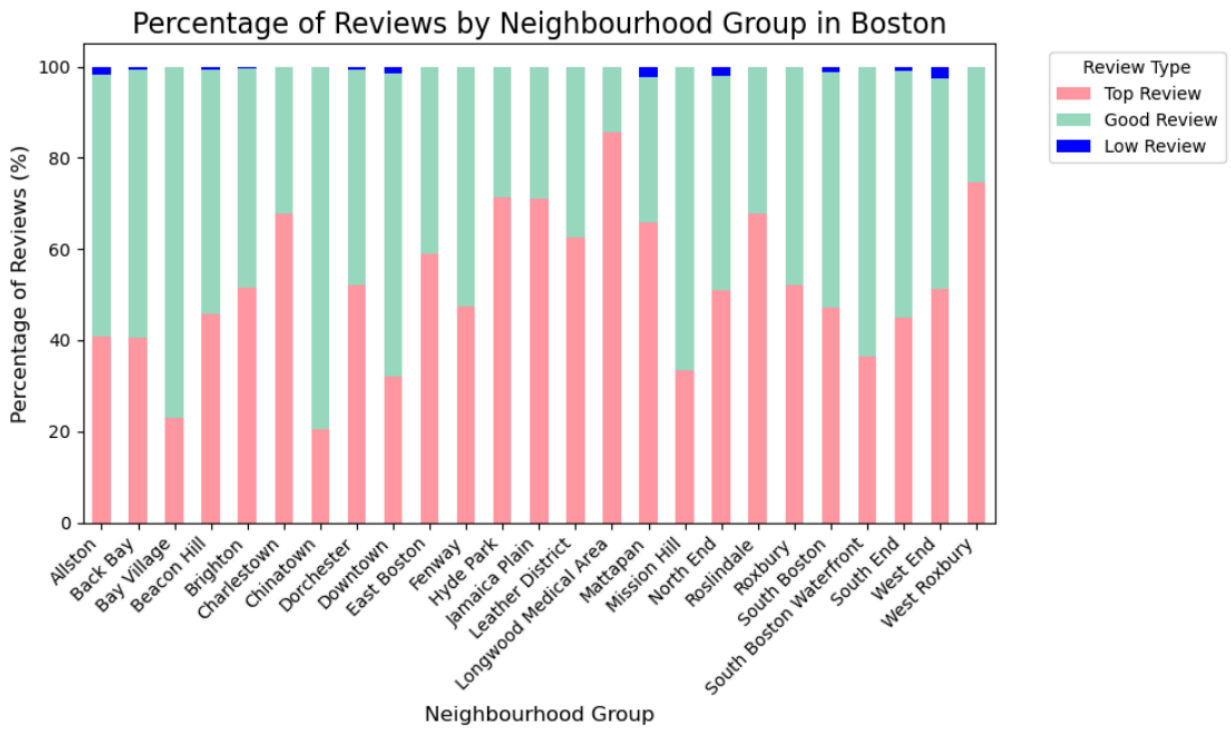

Fig 25: Percentage reviews by neighborhood groups - Boston

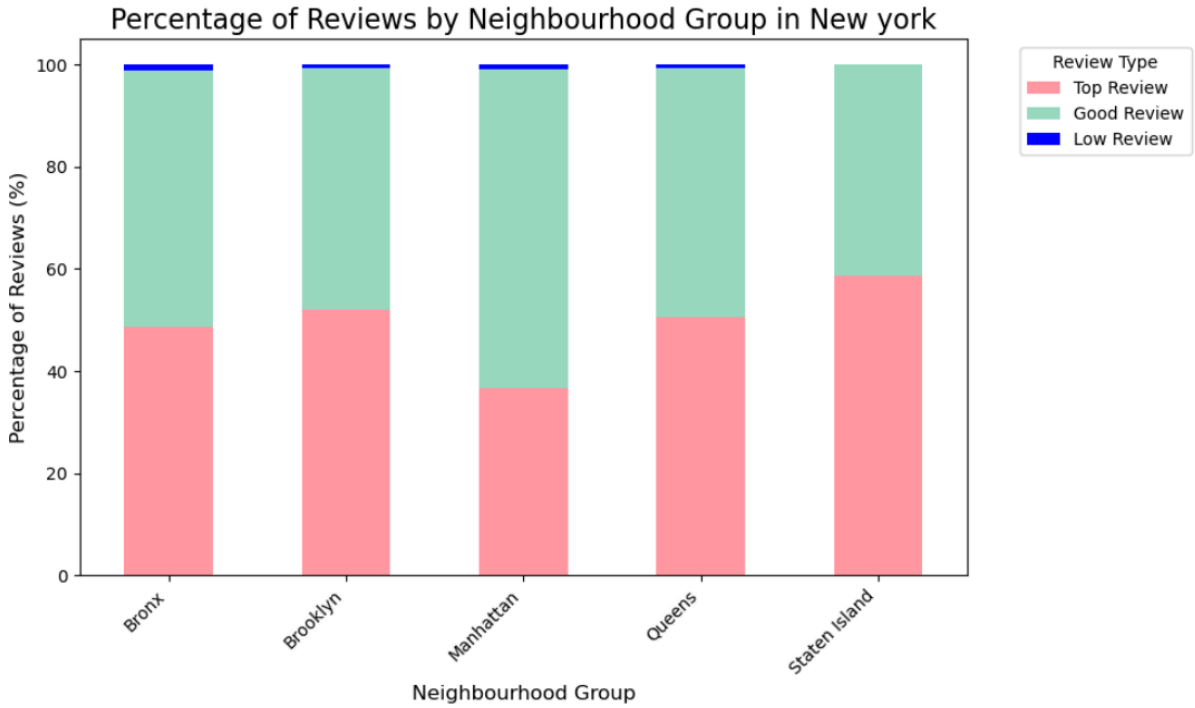

Fig 26: Percentage reviews by neighborhood - New York

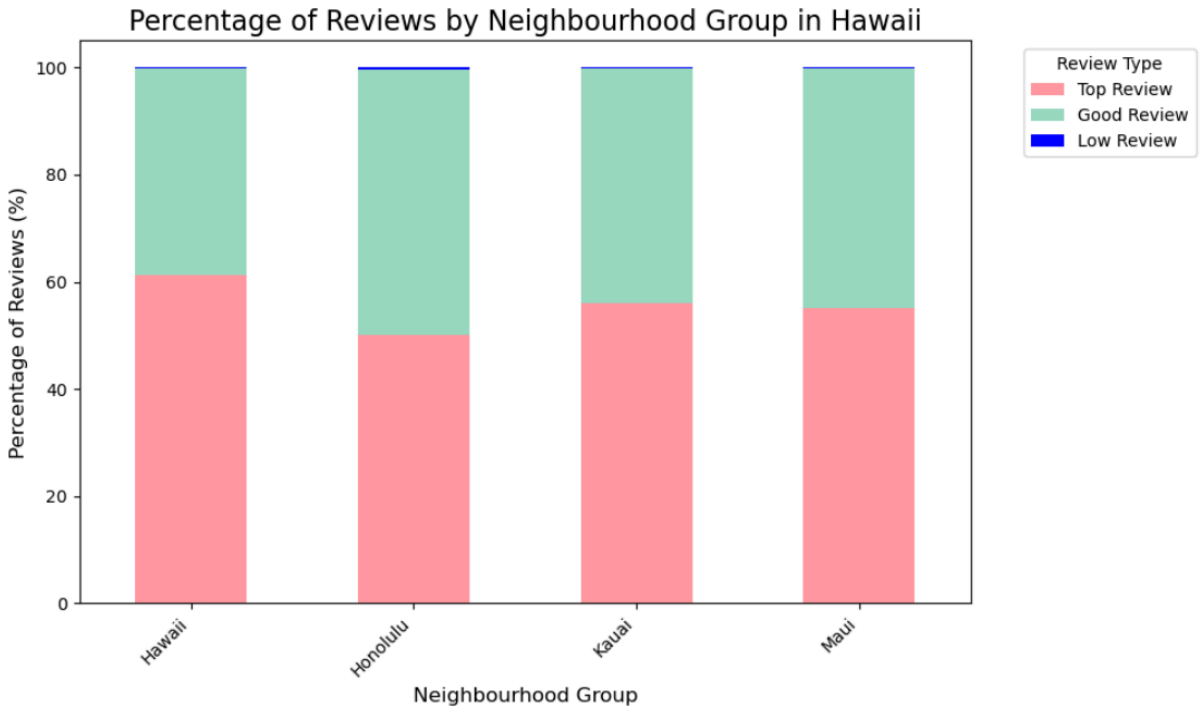

Fig 27: Percentage reviews by neighborhood - Hawaii

6. Outlier Detection

In order to find outliers, we use two methods - Calculating the IQR (InterQuartileRange) and Z- Score. We calculate the values for the following attributes - Price, Minimum nights, No of reviews, Review Score Rating as follows for the 5 cities:

Attribute	**Skewness**	**IQR**	**No. of outliers**
Price	16.46	(-174.5,701.5)	2904
Minimum nights	6.36	(-5,11)	4242
Review Score Rating	-4.39	(4.28,5.44)	1540

Table 4: Skewness and Outliers distribution - Hawaii

Attribute	Skewness	IQR	No. of outliers
Price	5.28	(-147.5,544.5)	241
Minimum nights	2.98	(-45.5,78.5)	716
Review Score Rating	-4.53	(4.15,5.48)	225

Table 5: Skewness and Outliers distribution - Boston

Attribute	Skewness	IQR	No. of outliers
Price	83.7	(-162.5,497.5)	1771
Minimum nights	17.04	(30,30)	7406
Review Score Rating	-4.39	(4.1,5.54)	1755

Table 5: Skewness and Outliers distribution - New York

First, we generate box plots for the attributes price, minimum_nights, review_scores_rating in order to visualize the outliers.

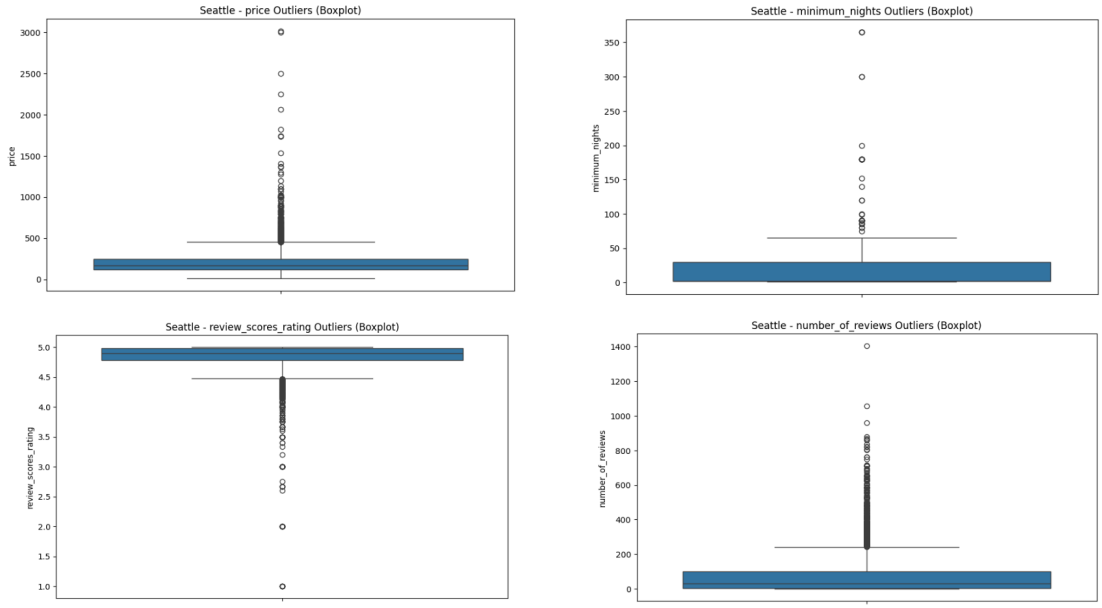

Fig 28: Box plot showing outliers - Seattle

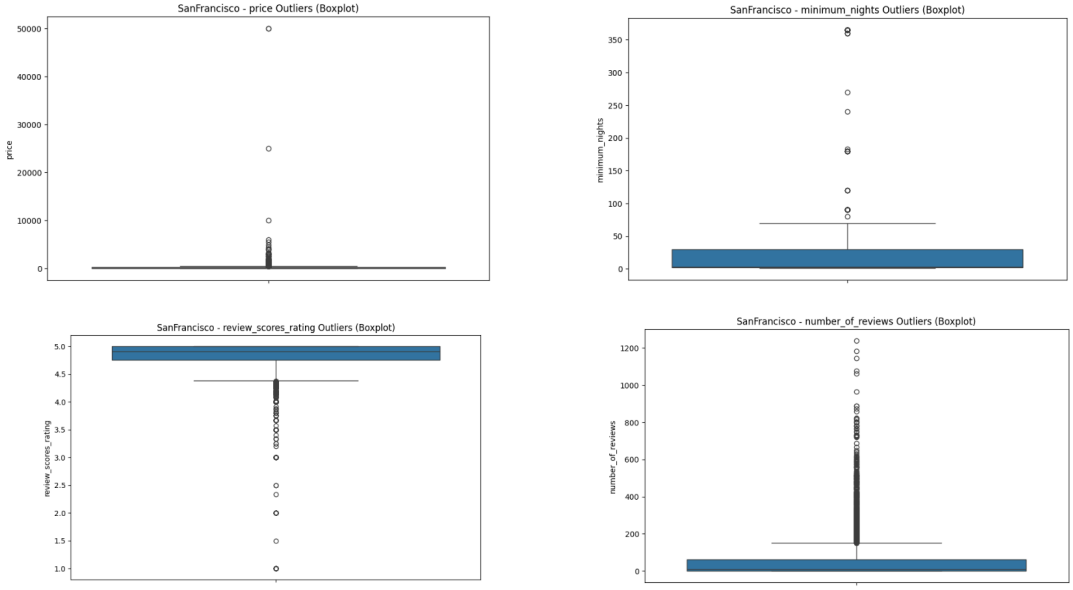

Fig 29: Box plot showing outliers - San Francisco

On removing the outliers, the data distribution changes significantly as evidenced by the following graphs:

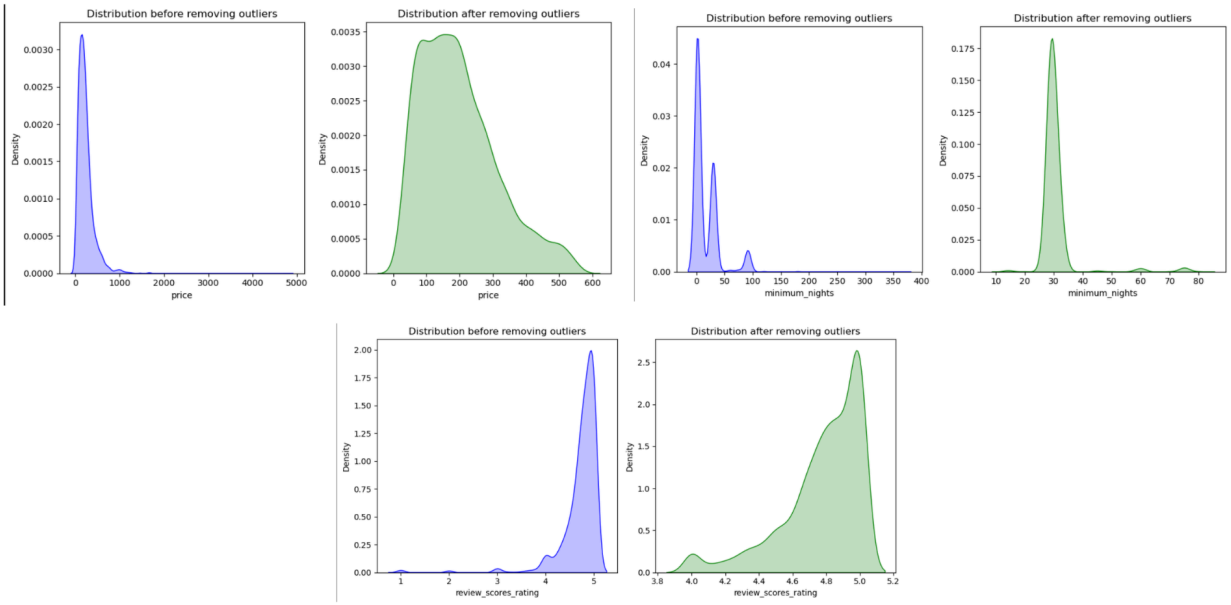

Fig 30: Distribution of the various attributes before and after removing outliers - Boston

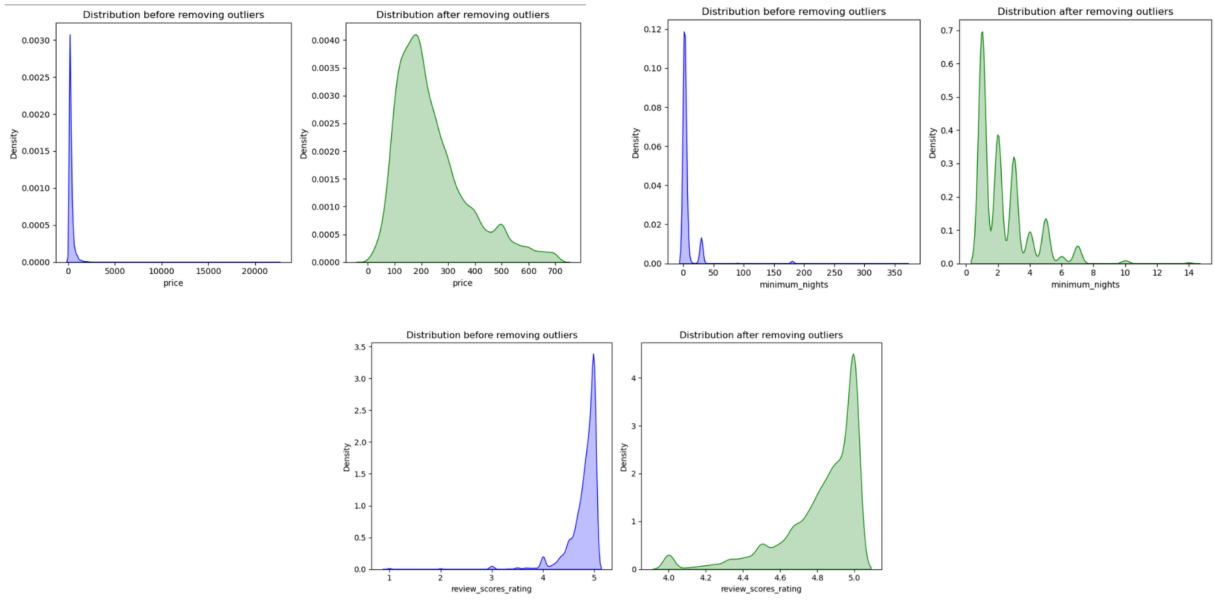

Fig 31: Distribution of the various attributes before and after removing outliers - Hawaii

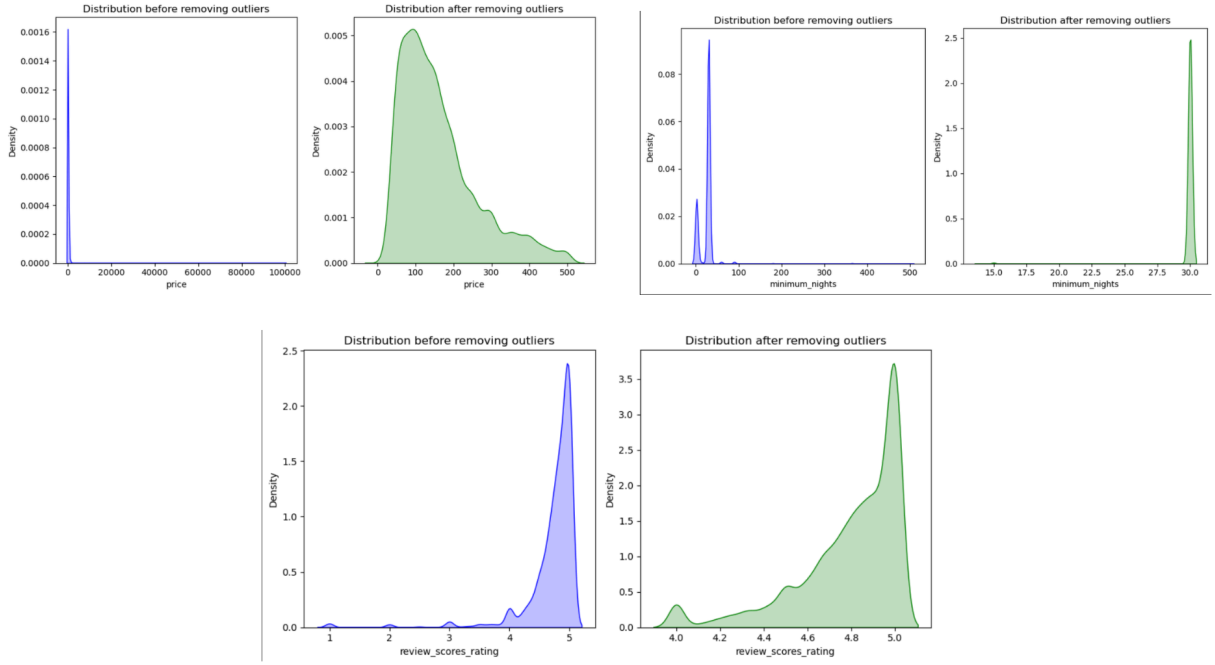

Fig 32: Distribution before and after removing outliers - New York

Price:

There is a significant narrowing of price range after outlier removal for all 5 cities.

Boston's price range narrowed from 0-5000 to 0-600.

Hawaii's range reduced from 0-20000+ to 0-700.

New York's range decreased from 0-100000 to 0-500.

Review Scores Rating:

The distribution remained right-skewed but became more concentrated since the lower ratings (below 4.0) were mostly removed as outliers.

Minimum nights:

The extreme values were removed.

Boston's range narrowed from 0-400 to 10-80 nights, with a clear peak around 30 nights.
Hawaii's range reduced from 0-350 to 0-15 nights, with multiple peaks visible.
New York's range decreased from 0-500 to 15-30 nights, with a sharp peak at 30 nights.

In conclusion, the effect of outlier removal was most pronounced for price and minimum nights for all 5 cities, with review scores not changing as much.

7. **Text length**

To find if there is any correlation between the number of characters and the sentiment of the review, we started with preprocessing the comments. This step involves cleaning the comments by removing the HTML tags, eliminating special characters, extra spaces and converting it into a lower-case character for consistency. Following this, we analyzed the data and inferred that **negative reviews tend to contain a higher word count compared to positive reviews.** This finding suggests that there is a possible correlation between review length and sentiment, with customers expressing negative experiences more extensively.

8. **Keyword Extraction**

We look at the most common words in both positive and negative reviews by visually representing them in the form of word clouds. This helped us identify key words that appeared frequently in each type of review. Following this, we count how often these important words showed up in each review.

We did some data analysis to find any connections between different aspects of the reviews and listings by looking at attributes like 'review_scores_rating', 'id', 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', 'review_scores_value', 'number_of_words', 'listing_id', 'sentiment'.

After examining all this information, we found that there **wasn't a strong correlation** between the number of positive or negative words we counted and the other details we looked at. This means that just counting certain words in a review doesn't tell us much about how people rated their stay or the specific aspects of their experience.

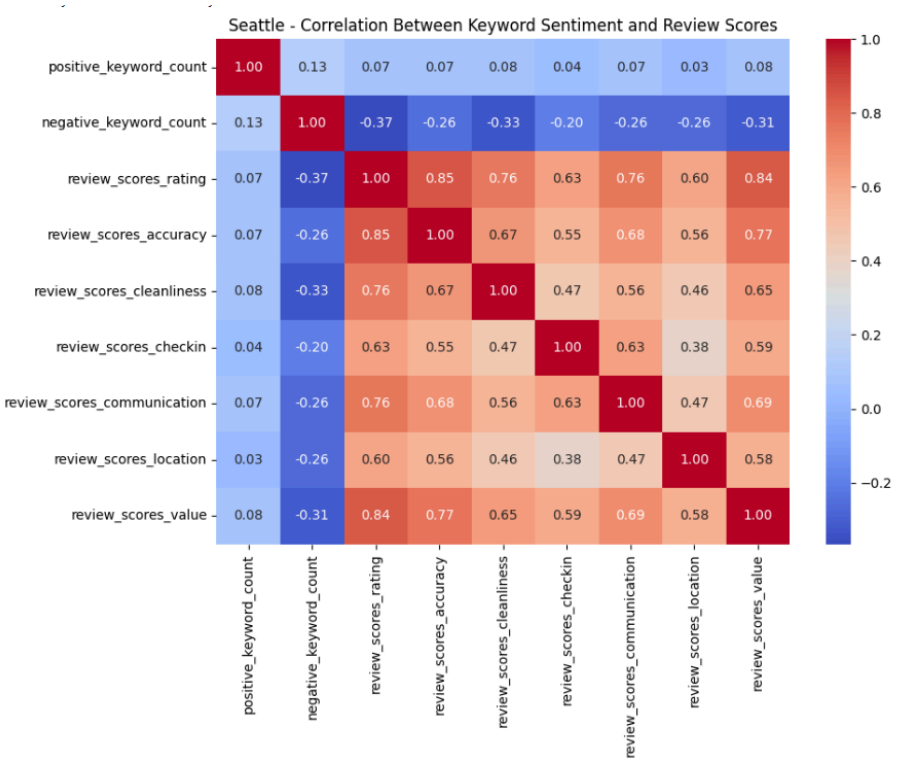

Fig 33: Keyword Sentiment Analysis for Seattle

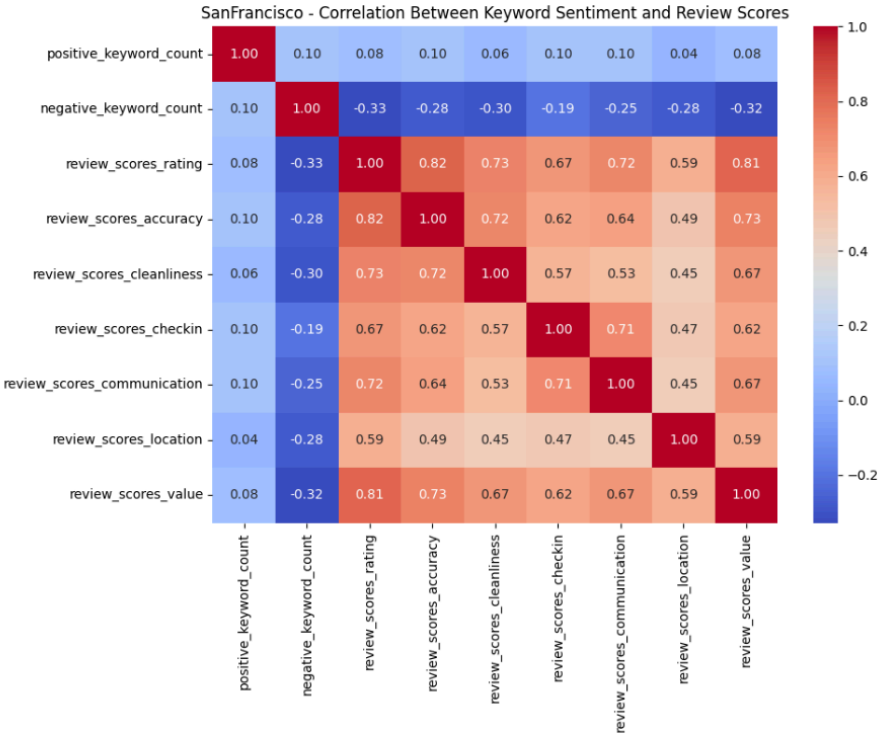

Fig 34: Keyword Sentiment Analysis for San Francisco

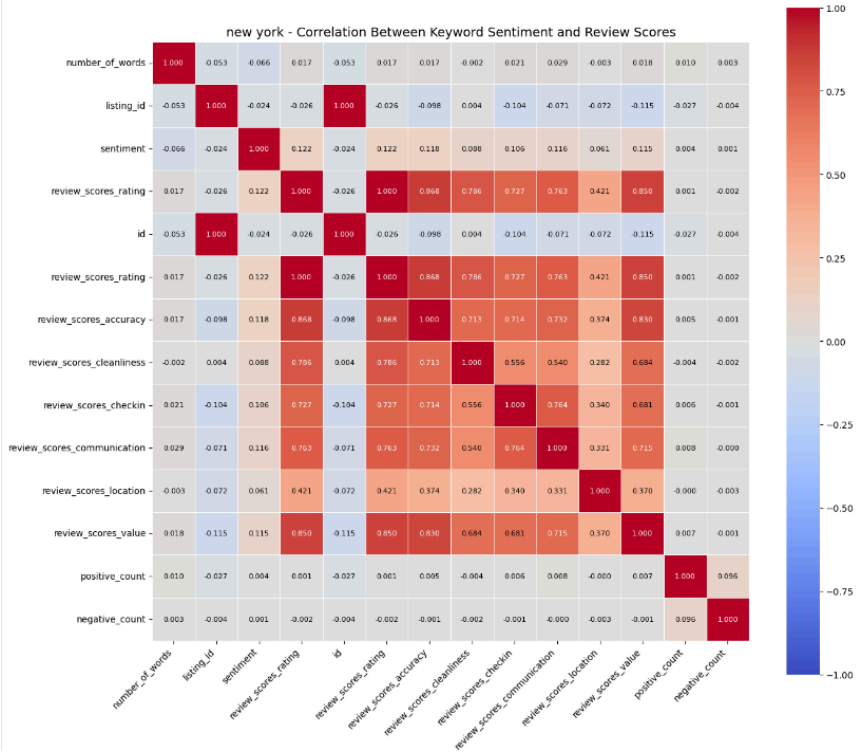

Fig 35: Keyword Sentiment Analysis for New York

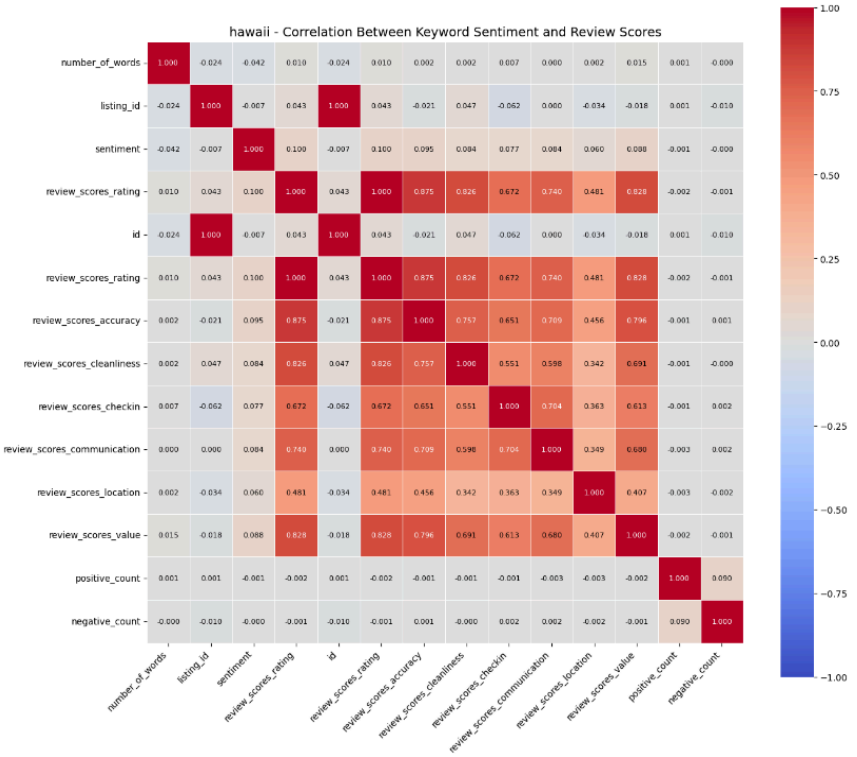

Fig 36: Keyword Sentiment Analysis for Hawaii

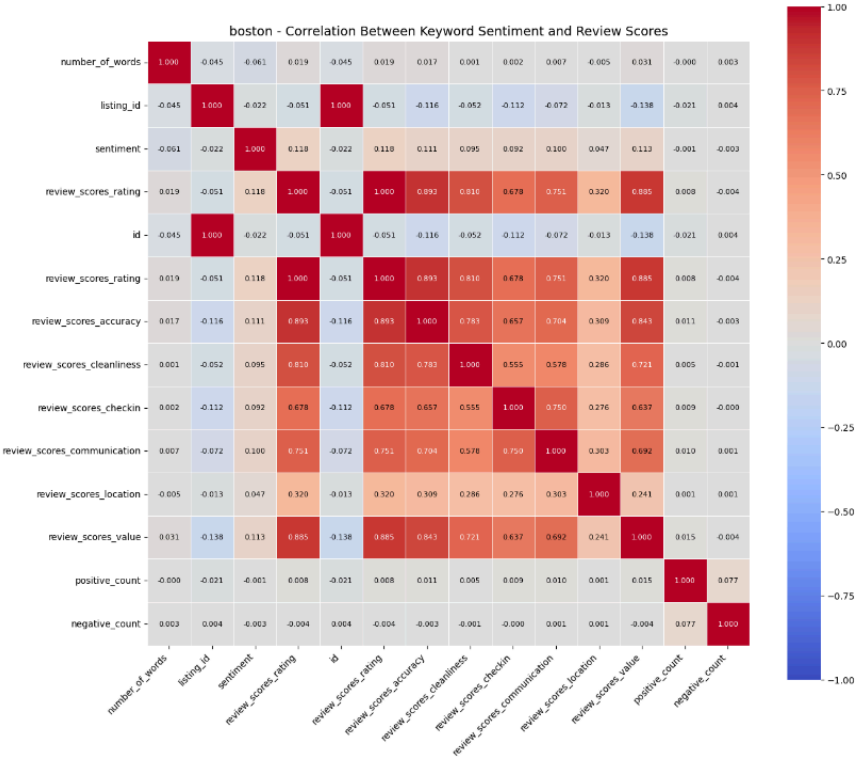

Fig 37: Keyword Sentiment Analysis for Boston

Limitations

Our analysis of the Airbnb data for New York, Boston, Seattle, Hawaii and San Francisco provided valuable insights, but it is very important to acknowledge the limitations of our analysis. Firstly, our data represents the second quarter of 2024, which is a limitation in itself, since it is a specific time period and the short-term rental market can change quickly. We also did not account for any seasonal variations in the pricing, which could be very significant for popular tourist destinations.

Additionally, we relied on the accuracy of the data provided by Airbnb and there might be some inaccuracies and inconsistencies which could affect our findings.

We also observed that there might be a bias in the review system as there are predominately more positive reviews when compared to the negative reviews.

Even though we removed some of the outliers, there might be some extremely high priced listings that might have skewed some of our average price calculations. Furthermore, while we considered various factors for our analysis, there could be other important elements affecting the prices like popularity or local attractions.

Our analysis doesn’t prove direct correlation between any of the parameters. More research needs to be done in order to establish such relationships conclusively.

Conclusion

Our exploratory data analysis of the Airbnb data (reviews and listings) across the following five cities - New York, Boston, Seattle, Hawaii, San Francisco has revealed some interesting insights. We found that the prices vary significantly with different cities and neighborhoods, with the prices being consistently high in certain neighborhoods. Manhattan in New York and the Kauai in Hawaii stood out as particularly expensive. This analysis would be helpful for property investors and budget constraint travelers.

We noticed that most of the listings across all cities have received higher ratings, with very few negative reviews. This suggests the overall experience of the Airbnb users, but we are also concerned that there is a potential bias in the review system.

One interesting insight we inferred from the data is that negative reviews tend to be longer than the negative reviews. However, we couldn't find a strong correlation between the use of specific words in reviews and the overall ratings or other fields of the listings.

Our analysis showed that most of the listings in the neighborhoods primarily cater to the short-term stays which aligns with the typical usage of tourists. However, there are also options for longer stays in all the cities.

Lastly, we found out that being a super host is usually associated with higher ratings and more reviews. This suggests that the superhost program is effective in identifying a good listing.

Overall, this analysis could be useful for the Airbnb host to improve their listings, planning for the tourists and Airbnb itself for refining their services. However, we need to point out that our analysis has some limitations and there is a need to do further research in order to provide more detailed insights about the rental market in these popular tourist destinations.

Individual Contributions

Name	Percentage
Tharanitharan Muthuthirumaran	34
Tyler Poff	33
Reemaa Sajad Hyder	33

References

U.S. cities with the most international tourists | Statista

Most visited states US 2022 | Statista

https://econometricstutors.com/exploratory-data-analysis-central-tendency/

https://mode.com/blog/violin-plot-examples

https://www.smartpls.com/documentation/functionalities/excess-kurtosis-and-skewness

https://www.datacamp.com/tutorial/understanding-skewness-and-kurtosis

https://insideairbnb.com/get-the-data/

seaborn: statistical data visualization — seaborn 0.13.2 documentation (pydata.org)