CS6140 Machine Learning Assignment 1

Exploratory Data Analysis Report

Problem Statement:

1. Explore Airbnb to gain hands-on experience with Exploratory Data Analysis and Feature Engineering.
2. Download listings & reviews for five cities and analyze these to identify trends, correlations and patterns that can drive business insights.
3. Analyze reviews dataset to engineer features aimed at predicting customer satisfaction and improving service offerings.

Introduction

Airbnb is a leading online marketplace that offers tailored accommodations and experiences for travelers seeking unique stays, connecting hosts and travelers. It was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb has revolutionized the travel industry by providing a platform for short-term and long-term stays in over 220 countries.

As data scientists at an analytics firm specializing in real estate and tourism, we are tasked with analyzing Airbnb data for five cities. The primary objective is to conduct exploratory data analysis and perform feature engineering to derive insights that will optimize property listings and improve guest satisfaction. The datasets under analysis are the listings and reviews datasets provided by Airbnb (available on: https://insideairbnb.com/get-the-data/). The listings dataset contains detailed information about each property, such as pricing, location, availability, and host details, while the reviews dataset contains guest feedback on various properties, explaining their experiences.

The tasks involve calculating descriptive statistics and performing distribution and correlation analysis on key features such as 'price', 'minimum_nights', 'maximum_nights', 'number_of_reviews', and 'review_scores_rating'. These statistics will help to understand the characteristics of these key features, and the analysis will explore relationships between them and help detect outliers. A deeper analysis of pricing in relation to neighborhoods, room types, and review score ratings will offer in-depth insights into the guest experience. Additionally, the tasks involve performing feature engineering focused on keyword frequency and review length to enhance the analysis and examine their correlation with guest satisfaction.

The cities considered for the analysis are five tech hubs in the United States, namely, Boston, Chicago, Dallas, Los Angeles, and San Francisco. The goal is to use the analysis for these five cities to provide insights that optimize property listings and enhance overall guest satisfaction.

Data Cleaning:

Removed all irrelevant attributes from the dataset. The 'price' column, originally a string with a '$' symbol, was cleaned by stripping the symbol and converting it to a float for easier analysis. Verified that no numerical attributes contained string or text data, converting any such values to null. Cleaned the 'room_type' attribute by removing extra white spaces and converting it to lowercase. Ensured that both 'room_type' and 'neighborhood_cleansed' attributes were free of null values. Additionally, processed the 'amenities' column by counting the number of amenities for each listing and adding this as a new feature in the data frame.

Analysis

1. Task 1: Descriptive Statistics

The goal of this task was to calculate summary statistics for some key numerical features and understand the central tendency, dispersion, and distribution of these variables.

This task was carried out with the help of the describe function, which provides a summary of the statistical characteristics of numerical columns in the DataFrame.

The central tendency of a variable can be determined with the help of mean, median or mode. The dispersion of a variable can be determined with the help of standard deviation. The distribution of the variable can be determined with range and visually plotting the data with figures like histograms or box plots. Here, we will look at the range.

The four main attributes that we analyzed are price, minimum_nights, maximum_nights and review_scores_rating.

Price:

These are the statistical values for the price attribute for all 5 cities.

City Name	Mean	Std	Min	Max	25%	50%	75%	Variance	Skewness	Kurtosis
Los Angeles	244.9	467.9	5.0	20343.0	96.0	151.0	250.0	219013.0	16.9	489.5
Boston	232.0	203.0	25.0	4786.0	112.0	190.0	284.0	41222.3	5.9	97.8
Chicago	216.9	218.0	19.0	2911.0	99.0	159.0	255.0	47532.3	4.4	31.9
Dallas	188.9	574.1	6.0	10000.0	79.0	112.0	178.0	329612.2	15.6	261.2
San Francisco	234.6	1050.4	25.0	50000.0	99.0	148.0	235.0	1103431	43.7	2061.8

Key Findings:

Central tendency, Dispersion and Distribution:

Los Angeles has the highest mean price, followed San Francisco, followed very closely by Boston and Chicago. Dallas has the least mean price. The standard deviation shows how much the prices deviate from the average. San Francisco has the highest price variability. Chicago and Boston have more stable price distributions, while Dallas and Los Angeles are more volatile than Boston and Chicago but less so than San Francisco.

Variance, Skewness and Kurtosis:

San Francisco has the highest variance, reflecting extremely wide price distribution, with outliers pulling the variance upward. Los Angeles and Dallas also have relatively high variances, suggesting significant price fluctuations. Boston and Chicago have lower variances, implying more consistent pricing compared to San Francisco.

Skewness shows the asymmetry in the price distribution. All cities have positive skewness, meaning they are right skewed with a longer tail on the higher price side. San Francisco has the highest skewness, showing a strong concentration of lower-priced listings with extreme high-price outliers. Los Angeles and Dallas also have high skewness values, while Boston and Chicago are relatively less skewed.

Kurtosis measures the peakedness of the distribution. San Francisco shows extremely high kurtosis, indicating sharp peaks and heavy tails, caused by rare but extreme unreasonable prices. Los Angeles and Dallas have high kurtosis too, meaning they have more frequent extreme price outliers compared to Boston and Chicago. Boston and Chicago exhibit much lower kurtosis, suggesting a relatively more normal distribution of prices without as many extreme outliers.

In summary, San Francisco stands out with extreme price variation and outliers, while Boston and Chicago show more moderate price distributions.

Minimum_nights:

City Name	Mean	Std	Min	Max	25%	50%	75%	Variance	Skewness	Kurtosis
Los Angeles	13.9	19.9	1.0	700.0	2.0	3.0	30.0	397.8	9.9	233.0
Boston	16.5	26.2	1.0	365.0	1.0	2.0	29.0	688.2	3.4	22.2
Chicago	9.8	21.1	1.0	365.0	1.0	2.0	5.0	448.5	9.3	140.1
Dallas	7.5	17.7	1.0	500.0	1.0	2.0	3.0	315.4	9.6	191.7
San Francisco	16.6	31.5	1.0	365.0	2.0	3.0	30.0	997.8	8.2	86.3

Key findings:

Central Tendency, Dispersion & Distribution:
San Francisco and Boston have the highest median minimum nights, suggesting longer stays on average compared to other cities. Los Angeles and Chicago have lower means, with Dallas having the shortest average stay.
San Francisco and Boston have the highest standard deviations, indicating a wide range of values. Los Angeles has a maximum value, indicating some extreme outliers. Meanwhile, Dallas has a smaller range.
Chicago and Dallas have smaller 75th percentiles, meaning most listings have relatively low minimum nights compared to other cities. Los Angeles and San Francisco have higher 75th percentiles, indicating longer minimum night stays for a sizable portion of listings.
San Francisco and Boston show the highest variances, highlighting significant variability in minimum stay requirements. Dallas and Los Angeles have lower variances, suggesting slightly less variation in minimum stay values.

Variance, Skewness and Kurtosis:
San Francisco and Boston show the highest variances, highlighting significant variability in minimum stay requirements. Dallas and Los Angeles have lower variances, suggesting slightly less variation in minimum stay values.

All cities have positive skewness, meaning they are right-skewed (longer tail on the higher end). Los Angeles and Dallas show the highest skewness, indicating that most properties require short stays, but a few listings have extremely long minimum nights. Boston has the lowest skewness, suggesting a relatively more balanced distribution of minimum stay requirements.
Los Angeles has the highest kurtosis, showing sharp peaks and more outliers (extremely long stays). San Francisco and Chicago also exhibit high kurtosis, suggesting a large presence of extreme values. Boston has the lowest kurtosis, indicating a more moderate distribution with fewer outliers.

In summary, San Francisco and Los Angeles show the most variability and extreme values for minimum nights, while Boston has a more balanced and consistent pattern. Dallas and Chicago demonstrate shorter stays on average with less variability in stay requirements.

Maximum_nights

City Name	Mean	Std	Min	Max	25%	50%	75%	Variance	Skewness	Kurtosis
Los Angeles	451.3	413.5	1.0	3650	90.0	365	730	171042.5	0.7	-0.7
Boston	539.8	420.8	3.0	1125.0	360.0	365.0	1125.0	177140.9	0.4	-1.3
Chicago	517.2	431.9	2.0	1125.0	95.0	365.0	1125.0	188798.4	0.5	-1.3
Dallas	515.9	390.5	1.0	1125	365	365	1125	152504.1	0.6	-1.0
San Francisco	583.3	14300.8	1.0	999999.9	29.0	182.0	365.0.	2045153300	69.8	4880.7

Key findings:

Central tendency, Dispersion and Distribution:

Most cities (Los Angeles, Boston, Chicago, Dallas) have a mean around 500 nights and a median of 365 nights, indicating that most properties allow stays of up to a year. San Francisco is an outlier, with an extremely high mean of 5833.3 nights due to an extreme maximum of 99,999 nights, though the median is much lower (182 nights)

San Francisco has the highest standard deviation (14,300.8), driven by the outlier of 99,999 nights, making it an anomaly compared to the other cities. The other four cities have standard deviations around 400-430, indicating moderate dispersion, with most properties offering stays within a range of 1 to 3650 nights (for Los Angeles) or 1 to 1125 nights (for the others).

All cities except San Francisco show a pattern where the majority of properties cluster around the 365-night mark. San Francisco's distribution is different, with the 25th percentile as low as 29 nights and the median at 182 nights, suggesting that most properties there allow shorter stays, but extreme outliers heavily influence the overall statistics.

Variance, Skewness and Kurtosis:

San Francisco has the most extreme values, with massive variance, high skewness, and extremely high kurtosis due to the outlier of 99,999 nights. The other cities have similar variance and skewness, showing moderate rightward skew and relatively flat distributions with some outliers, but nothing on the scale of San Francisco.

In conclusion, it is clear that San Francisco presents a unique market for long-term stays, where some properties allow for extraordinarily long maximum stays (nearly 100,000 nights), creating high variability. The other cities, however, offer more consistency with most properties clustered around a one-year maximum stay, making them more predictable for both short-term and long-term rental planning. In comparison, San Francisco's data needs to be analyzed separately due to its extreme outliers, which make it significantly different from the other cities.

Review_scores_rating

City Name	Mean	Std	Min	Max	25%	50%	75%	Variance	Skewness	Kurtosis
Los Angeles	4.7	0.4	1.0	5.0	4.7	4.9	5.0	0.16	-5.2	37.0
Boston	4.7	0.4	1.0	5.0	4.6	4.8	4.9	0.18	-4.6	30.7
Chicago	4.7	0.4	1.0	5.0	4.7	4.8	4.9	0.1	-5.3	40.1
Dallas	4.7	0.4	1.0	5.0	4.6	4.8	5.0	0.1	-4.4	28.3
San Francisco	4.7	0.3	1.0	5.0	4.7	4.9	5.0	0.1	-4.9	35.6

Key findings:

Central tendency, Dispersion and Distribution:

All cities have similar means, but Los Angeles and San Francisco have slightly higher medians, suggesting that ratings in these cities are clustered around 5 more frequently compared to others.

San Francisco has the lowest standard deviation and variance, indicating the tightest clustering of ratings around the mean. Boston has the highest variance, suggesting more variability in reviews than the other cities.

Los Angeles, Dallas, and San Francisco have a 75th percentile of 5.0, indicating that 25% of their reviews are perfect scores, reflecting extreme customer satisfaction. Boston and Chicago, with a 75th percentile of 4.9, show fewer perfect ratings but still remarkably high scores. While all cities tend toward high ratings, Los Angeles, Dallas, and San Francisco show a greater concentration of perfect reviews, whereas Boston and Chicago have a slightly more diverse range of high-end scores.

Variance, Skewness and Kurtosis:

Boston has the highest variance, indicating more variability in reviews, while Chicago, Dallas, and San Francisco have the lowest, showing that ratings are most consistent in these cities.
Chicago has the most negative skew, indicating that a significant majority of reviews are close to 5. Dallas has the least negative skew, meaning it has more evenly distributed ratings.
Chicago also has the highest kurtosis, indicating the sharpest peak and most concentrated distribution of high ratings, while Dallas has the lowest kurtosis, showing a more spread-out distribution.
In summary, this comparison shows that while all cities have a strong tendency towards high ratings, Los Angeles, Dallas, and San Francisco stand out with a greater concentration of perfect reviews, while Boston and Chicago have a slightly more diverse range of high-end scores.

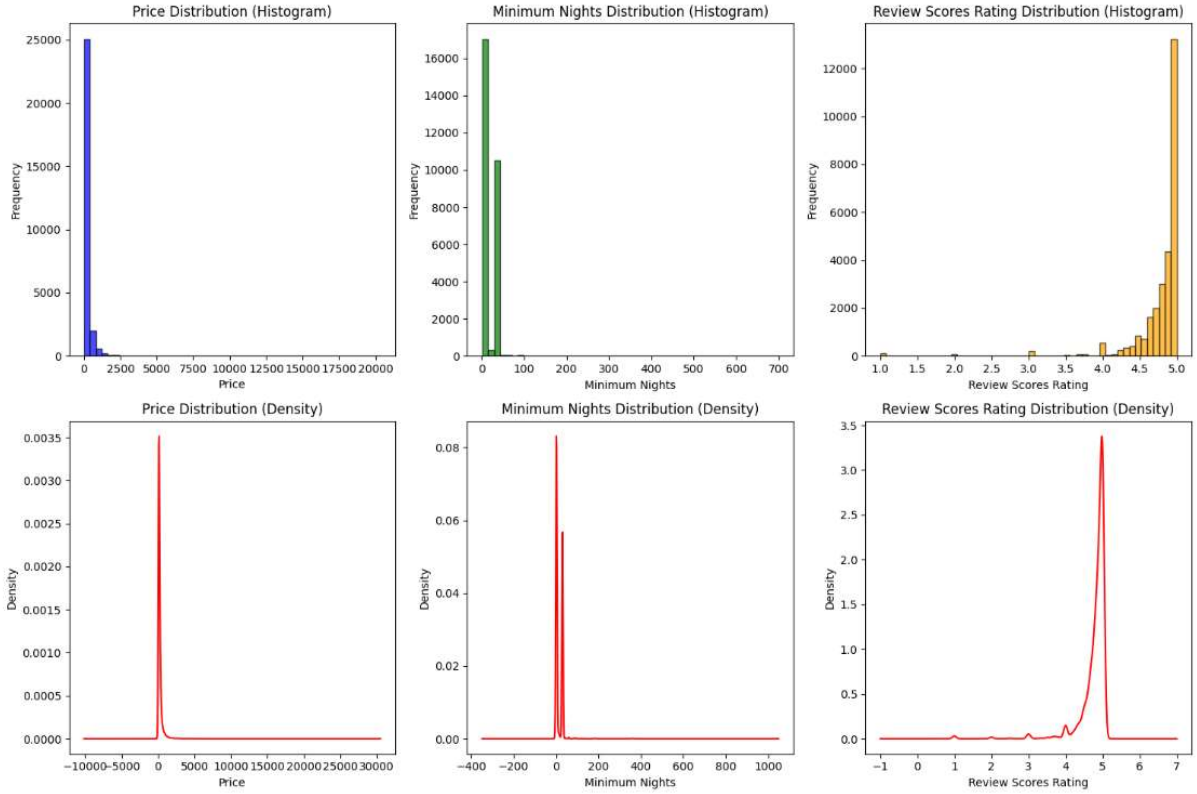
2. Task 2: Distribution Analysis

The goal of the task is to plot histograms or density plots for key numerical features and analyze the distribution of these features to identify any skewness or outliers.

We used pandas for data manipulation and matplotlib.pyplot for creating the plots. The function plot_numerical_distributions(dataframe) is defined to plot both histograms and

density plots for key numerical features in each DataFrame (dataframe). The function takes in a pandas DataFrame as input.

I. Los Angeles



Price Distribution:

- The price distribution is heavily right-skewed, with most data points concentrated around lower price values. A few extreme values are stretching out towards the higher end (up to 20,000), indicating outliers.
- Density Plot: The sharp peak near the lower prices with a long tail further confirms the presence of positive skewness and potential outliers in the higher price range.

Minimum Nights Distribution:

- Histogram: This distribution is also right skewed, with most listings requiring fewer than 100 minimum nights. There are fewer listings with higher minimum night requirements, but there are some extreme values, particularly around 700 nights, indicating outliers.

- Density Plot: The long tail in the density plot further supports the right skewness, with a remarkably high concentration of data near lower minimum night values.

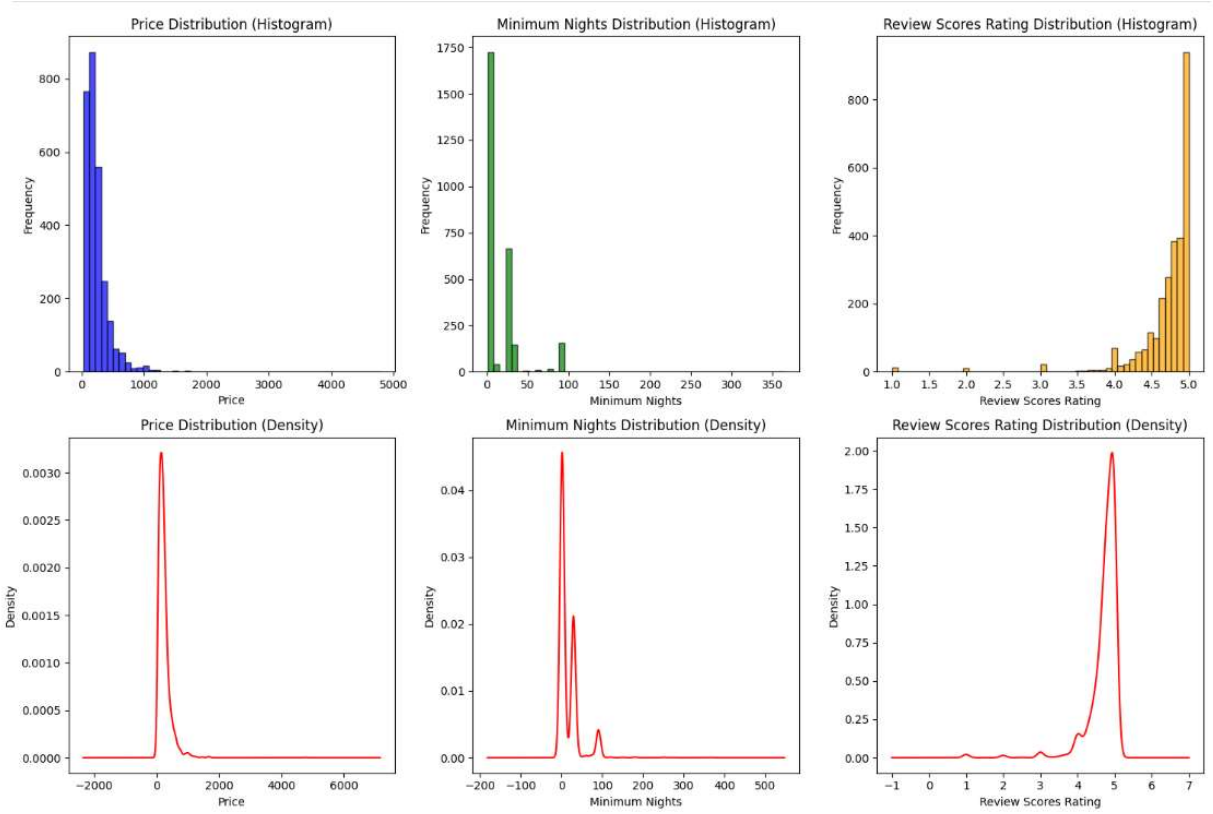
Review Scores Rating Distribution:

- Histogram: The distribution is left-skewed, with most reviews clustered around the highest scores (close to 5.0). This suggests that most properties have remarkably high review scores.
- Density Plot: The density plot shows a sharp peak near 4.5-5, confirming the left skewness, but no significant outliers are immediately evident in the review score.

Findings:

- Price and Minimum Nights show positive skewness and the presence of outliers at the higher ends.
- Review Scores Rating is negatively skewed with most values near the upper end (high ratings).
- Outliers should be handled carefully, especially in terms of the price and minimum night features, as they may distort the overall model performance. Skewed data may benefit from transformations (e.g., log transformations) to normalize distributions.

II. Boston:



Price Distribution (Histogram & Density)

- Histogram: The distribution of prices is right-skewed, meaning that most prices are concentrated at lower values, with a long tail extending towards higher prices. This indicates the presence of some high-priced listings.
- Density Plot: The density plot shows most prices are below $1000, with a sharp peak between $0 and $200. The long right tail further confirms the presence of outliers in the form of high-priced listings.

Minimum Nights Distribution (Histogram & Density)

- Histogram: The minimum nights feature also has a right-skewed distribution, with most listings having low minimum night requirements (fewer than 10 nights). However, there are some outliers with significantly higher minimum night requirements.
- Density Plot: Like the price, the density plot indicates that most listings require a minimum stay of less than 10 nights, with a long tail stretching to higher values.

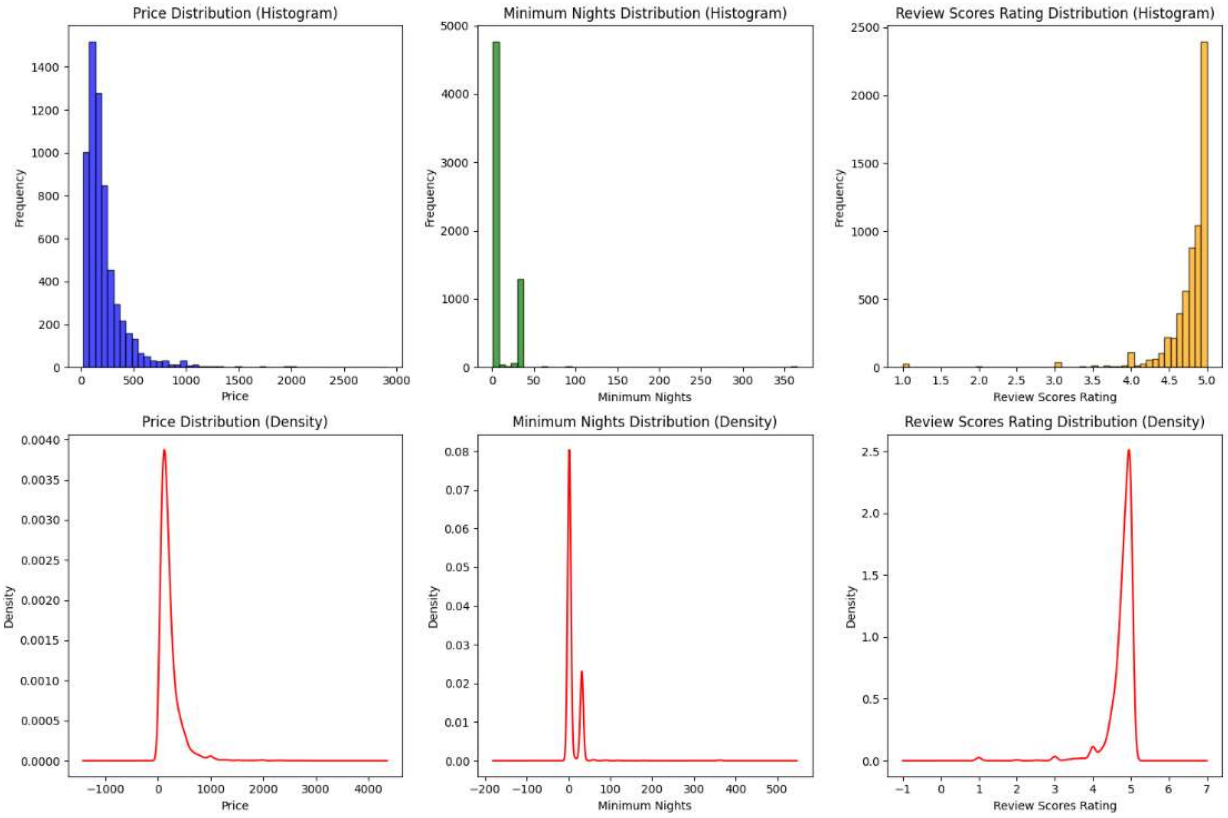
Review Scores Rating Distribution (Histogram & Density)

- Histogram: The review scores rating distribution is strongly left-skewed, meaning that most review scores are high, clustered around 4.5 and 5.0. There are very few low-rating reviews, which are outliers in this case.
- Density Plot: The density plot confirms the concentration of ratings between 4 and 5, with a steep peak at 5. This is common in review systems where high scores dominate, and exceptionally low ratings are rare.

Findings:

- Price and Minimum Nights: Both these features are right-skewed and contain outliers in the form of high values. These outliers could impact summary statistics like the mean, so measures like the median or robust statistical techniques should be considered when analyzing these features.
- Review Scores: The review scores are heavily left-skewed, with most ratings being high, which may indicate a positive bias in the reviews or customers being satisfied.

III. Chicago



Price Distribution (Histogram & Density)

- Histogram: The price distribution is strongly right-skewed, with most prices concentrated under $500. The tail extends towards $3000, indicating a few high-priced listings.
- Density Plot: The density plot confirms that the majority of the listings are priced below $500. The peak is sharp, followed by a long tail that includes higher prices.

Minimum Nights Distribution (Histogram & Density)

- Histogram: The minimum nights feature is highly right-skewed, with most listings requiring fewer than 20 nights. There are a few listings with significantly higher minimum night requirements.
- Density Plot: The density plot shows a sharp peak below 20 nights, and a long tail extends to higher values, indicating some outliers.

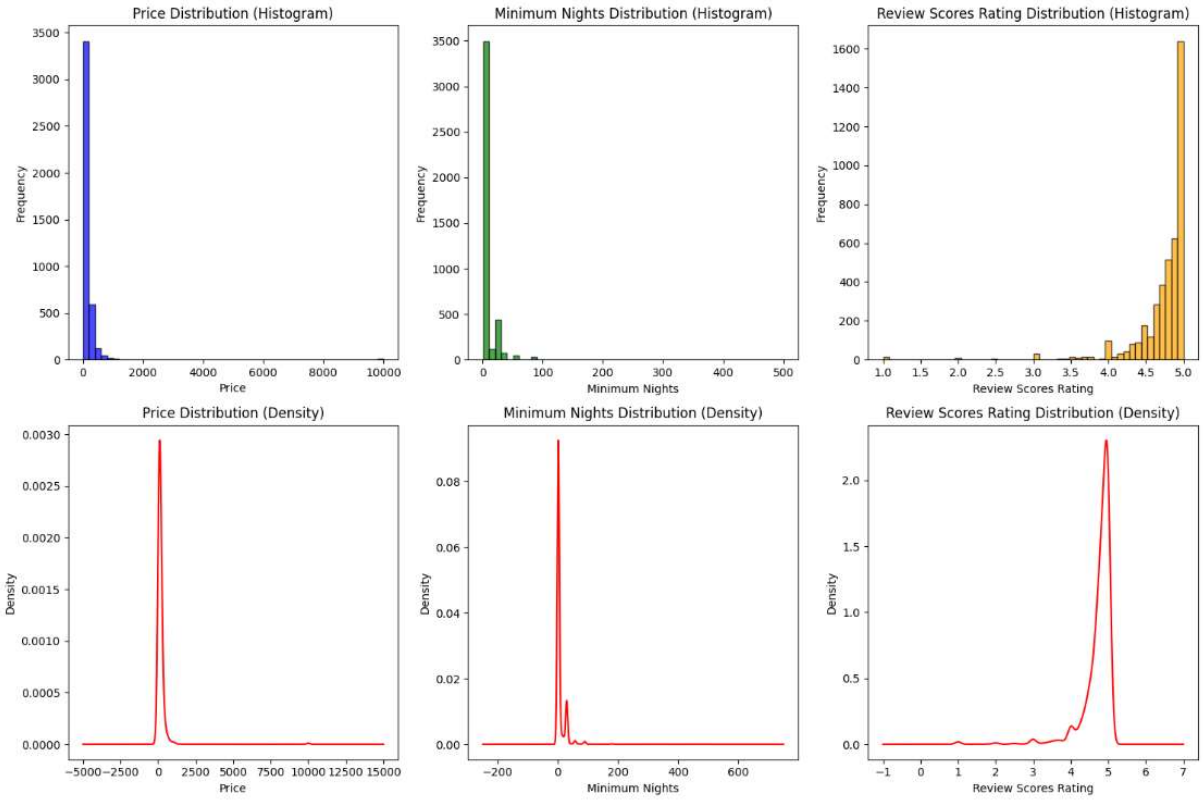
Review Scores Rating Distribution (Histogram & Density)

- Histogram: The review scores rating distribution is left-skewed, with most ratings clustered near the maximum of 5.0. There are only a few listings with lower review scores.
- Density Plot: The density plot reinforces this, showing a significant peak at 5.0, indicating a general bias toward high review scores.

Findings:

- Price and Minimum Nights: Both distributions are right-skewed with outliers in the higher ranges. A log transformation might help normalize these features, and outliers should be considered before analysis.
- Review Scores: The review scores are concentrated at the higher end, and the skewness suggests that most listings are rated highly. Low review scores are rare and can be considered outliers.

IV. Dallas



Price Distribution (Histogram & Density)

- Histogram: The price distribution is heavily right-skewed. The majority of listings are priced under $2000, but there are a few outliers that extend up to nearly $10,000, contributing to a long tail.
- Density Plot: The density plot shows a similar pattern, with a sharp peak around low prices and a long right tail extending towards higher prices.

Minimum Nights Distribution (Histogram & Density)

- Histogram: The minimum nights distribution is also right-skewed, with most listings requiring fewer than 50 nights. A small number of listings have very high minimum night requirements (up to 600).
- Density Plot: The density plot confirms the histogram, showing that the vast majority of listings require a low number of nights, with a sharp drop-off in frequency for higher values.

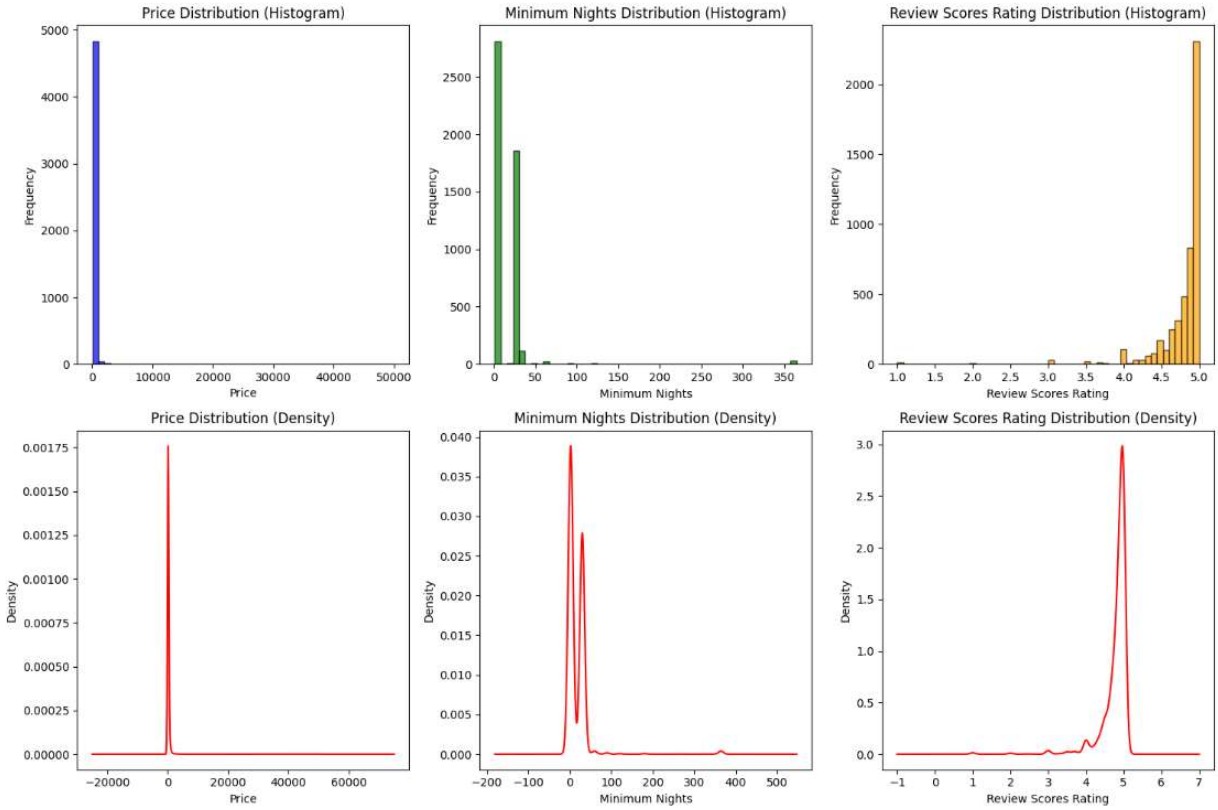
Review Scores Rating Distribution (Histogram & Density)

- Histogram: The review scores are heavily left-skewed, with most listings receiving high ratings, close to 5.0. There are very few listings with ratings below 3.0.
- Density Plot: The density plot shows a very high concentration of review scores near 5.0, confirming that most ratings are positive.

Findings:

- Price: The price distribution is right-skewed with extreme outliers that could distort averages. Most prices are clustered below $2000, with some extreme cases nearing $10,000.
- Minimum Nights: The distribution of minimum nights is also right-skewed, with most listings requiring fewer than 50 nights. However, some listings have very high minimum night requirements, extending the tail of the distribution.
- Review Scores: Review scores are highly left-skewed, with most listings receiving very high ratings. Low-rated listings are rare and can be considered outliers.

V. San Francisco



Price Distribution (Histogram & Density)

- Histogram: The price distribution appears highly right-skewed, with most listings priced under $1000. There are extreme outliers, with some prices reaching up to $50,000, which seems highly unusual for typical listings.
- Density Plot: The density plot further confirms the skewness, with a very sharp peak below $1000 and an extremely long tail stretching to high values, indicating outliers.

Minimum Nights Distribution (Histogram & Density)

- Histogram: The minimum nights distribution is also right-skewed, with the majority of listings requiring fewer than 30 nights. Some outliers show higher minimum night requirements, extending up to 350 nights.
- Density Plot: The density plot shows a sharp peak around 0 to 20 nights, followed by a long right tail. Some listings have abnormally high minimum night requirements, which are rare but should be considered outliers.

Review Scores Rating Distribution (Histogram & Density)

- Histogram: The review scores rating distribution is strongly left-skewed, with the majority of ratings clustered around the maximum value of 5.0. There are very few low-rated listings, with most falling between 4.0 and 5.0.
- Density Plot: The density plot confirms a steep concentration around 5.0, showing that high ratings are dominant in the data, and there are very few low ratings.

Findings:

- Price: Highly skewed with extreme outliers in the higher range. These outliers, potentially in the $10,000 to $50,000 range, may significantly impact the analysis, and measures such as the median or interquartile range could better represent the central tendency.
- Minimum Nights: Right-skewed with a majority of listings having low minimum night requirements but some notable outliers. These outliers can influence summary statistics and should be examined.
- Review Scores: Left-skewed towards high ratings, with most listings receiving very positive reviews. The few lower ratings are rare and can be treated as outliers in this context.

3. Task 3: Correlation Analysis

i. Introduction

This task aims to find correlations between numerical fields of the listing data such as 'price','minimum_nights','maximum_nights','number_of_reviews','bedrooms','bathrooms','beds','accommodates','availability_30','availability_365','number_of_reviews_ltm','amenities_count'. We need to identify the degree of correlation between these variables to use these in predictive modeling.

ii. Approach

Step 1: We first skim through the dataset to find all the columns which contain numerical data and select columns such as ‘price’, ‘maximum_nights’,’minimum_nights’, ‘bedrooms’, ‘bathrooms’ etc.

Step 2: We convert all the selected columns to numeric values; all the invalid values will be set as NaN.

Step 3: Drop all the rows that have NaN values with respect to the columns that were selected

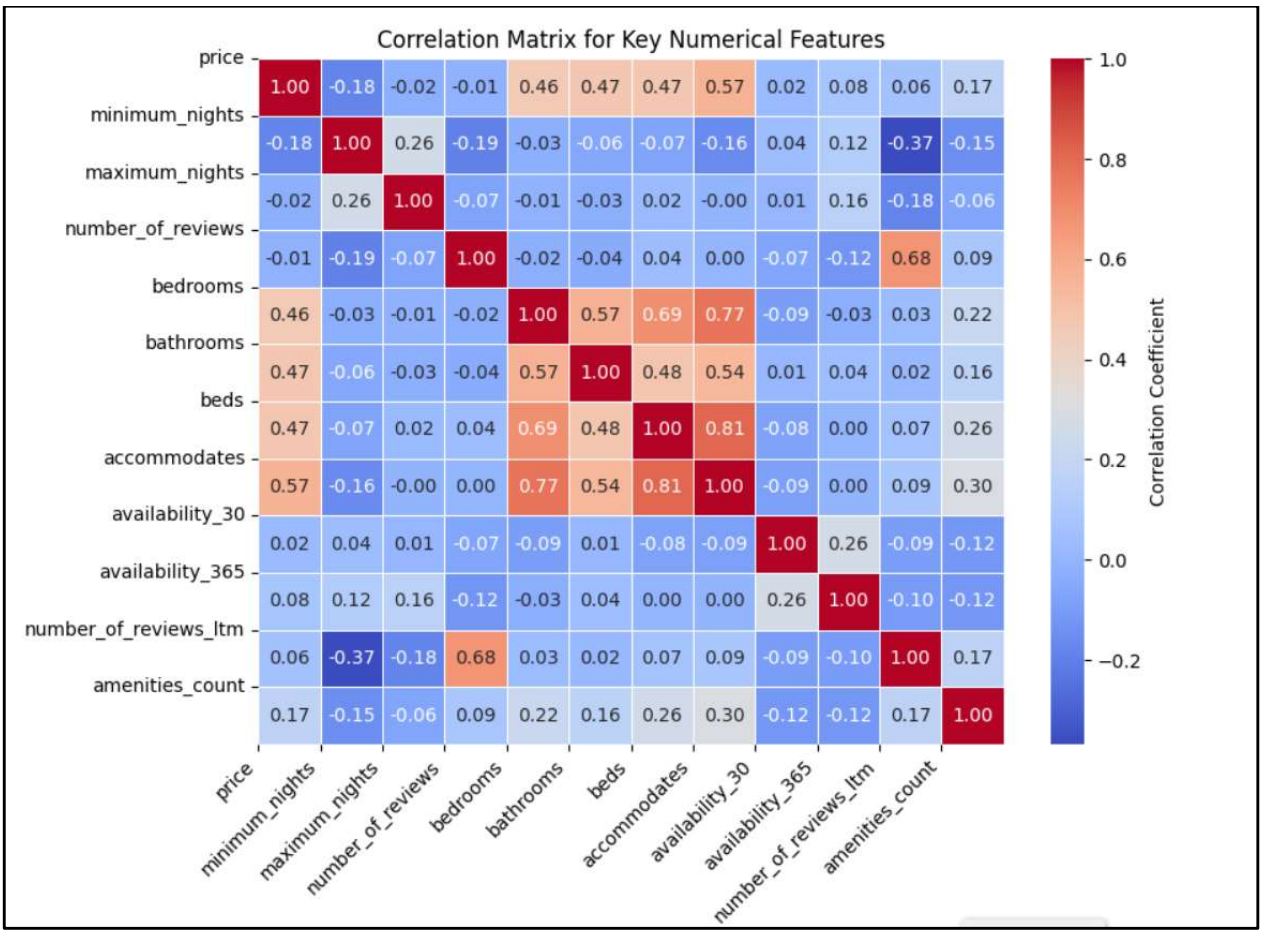
Step 4: Creating the correlation matrix, where the correlation will range from –1 to 1. The correlation matrix measures the linear relationship between two variables.

Step 5: We create a heatmap using the seaborn library to visualize the correlation matrix.

Step 6: Configure the heat map axis, using the column names as the names if both x-axis and y-axis. Finally display the heat map along with printing the correlation matrix as a text.

iii. Observations and Conclusion

a) Boston

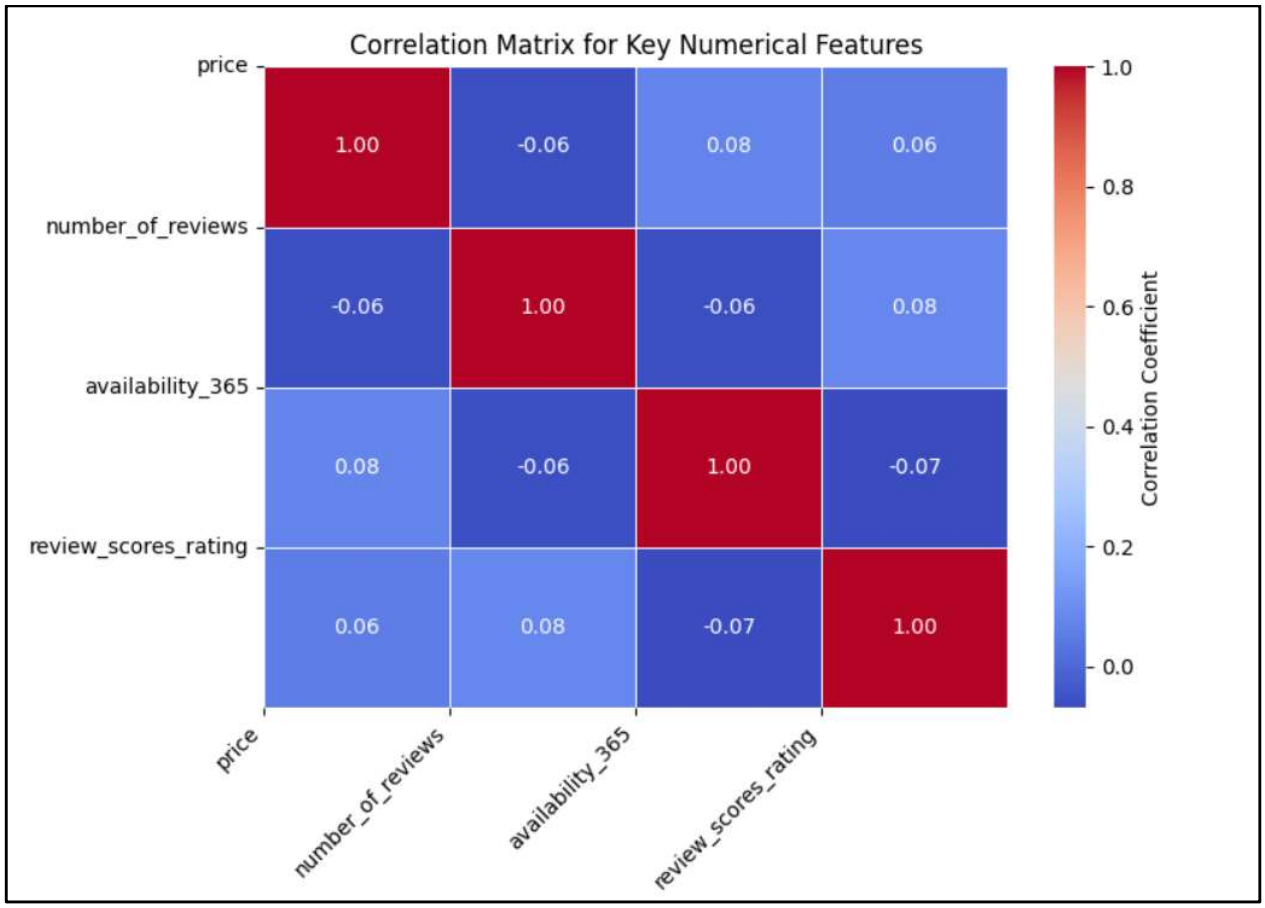


- Accomodates has strong correlation of (0.81) with bed, Bedroom, and beds (beds (0.69), accomodates and bedrooms (0.77) and finally Number of Reviews LTM (Last Twelve Months) with Number of Reviews (0.68).
- Price has a moderate correlation with accommodates (0.57) and accommodates also has a correlation with bathrooms (0.54)

Conclusion:

We can conclude that since the correlation is so strong almost close to 1.0, the pairs of variables are dependent on each other which indicates any change in one of those variables is likely to affect the other variable.

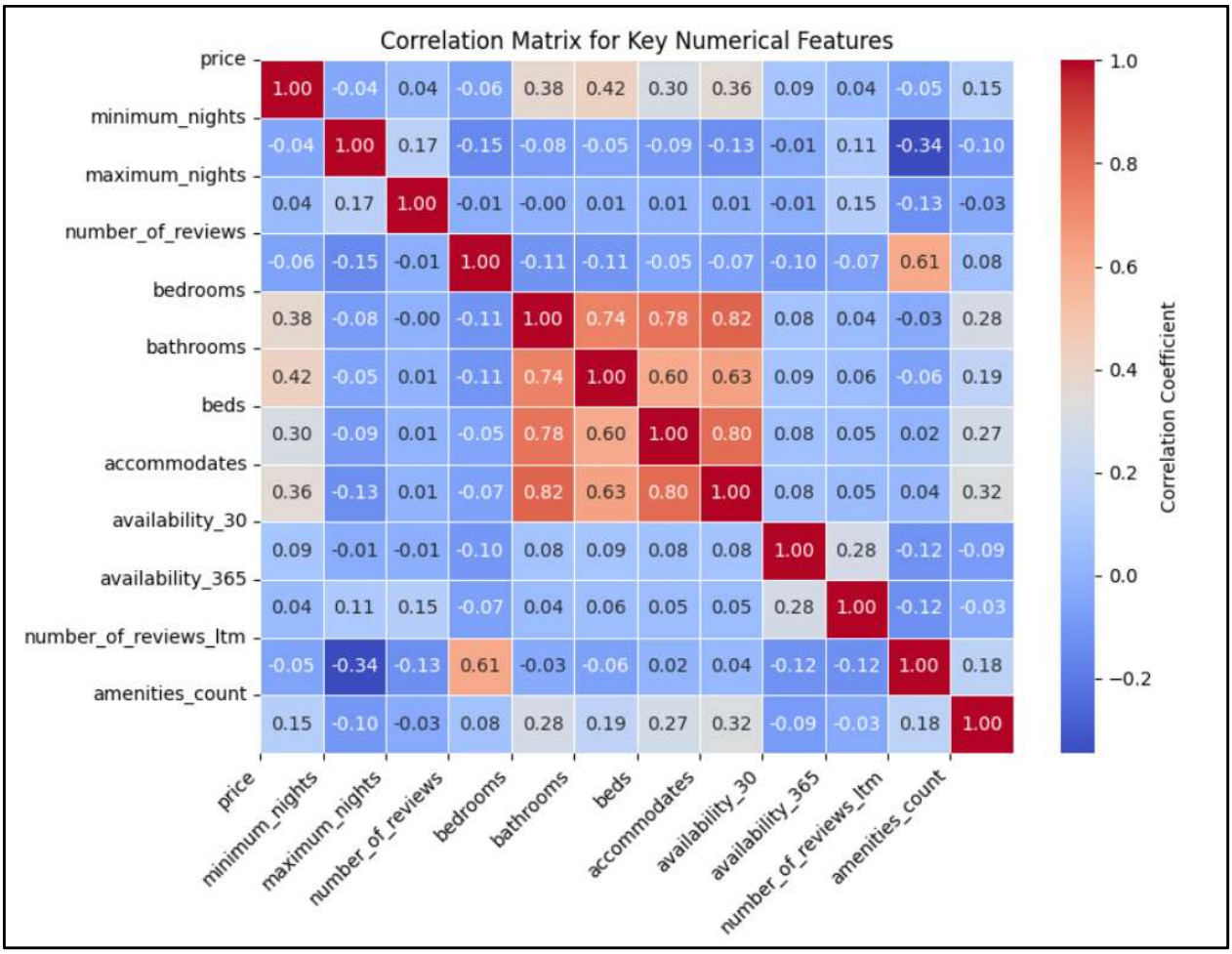
b) Chicago



Conclusion:

Here we can see that there is no strong or even moderate correlation between any pairs of variables. So, any change in one of these variables is not likely to affect the other variables.

c) LA

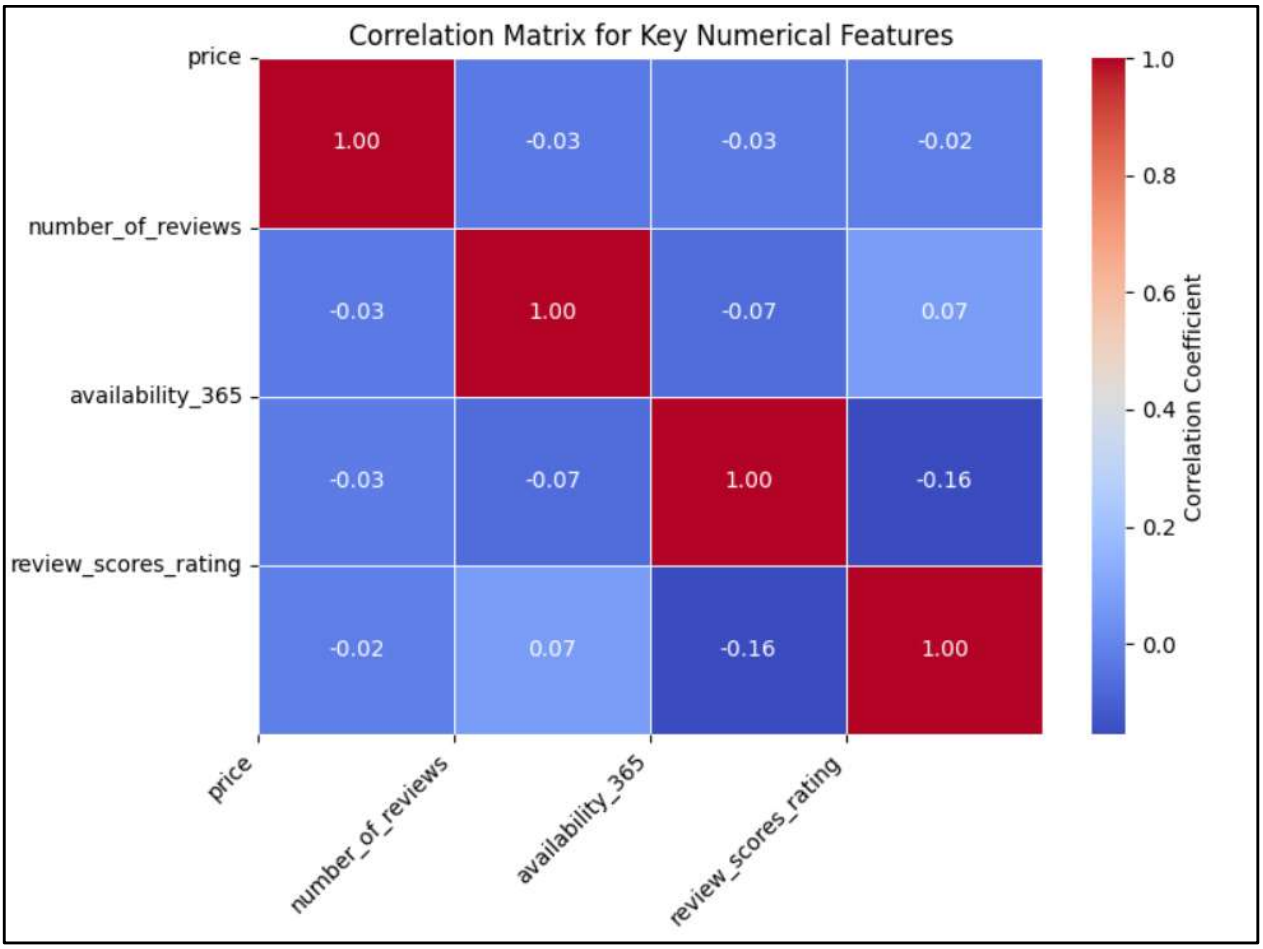


- Accomodates has strong correlation of (0.8) with bed, Bedroom, and beds (0.78), accomodates and bedrooms (0.82), Number of Reviews LTM (Last Twelve Months) with Number of Reviews (0.61) and finally accomodates has a strong correlation with bathrooms (0.74).

Conclusion:

We can conclude that since the correlation is so strong almost close to 1.0, the pairs of variables are dependent on each other which indicates any change in one of those variables is likely to affect the other variable.

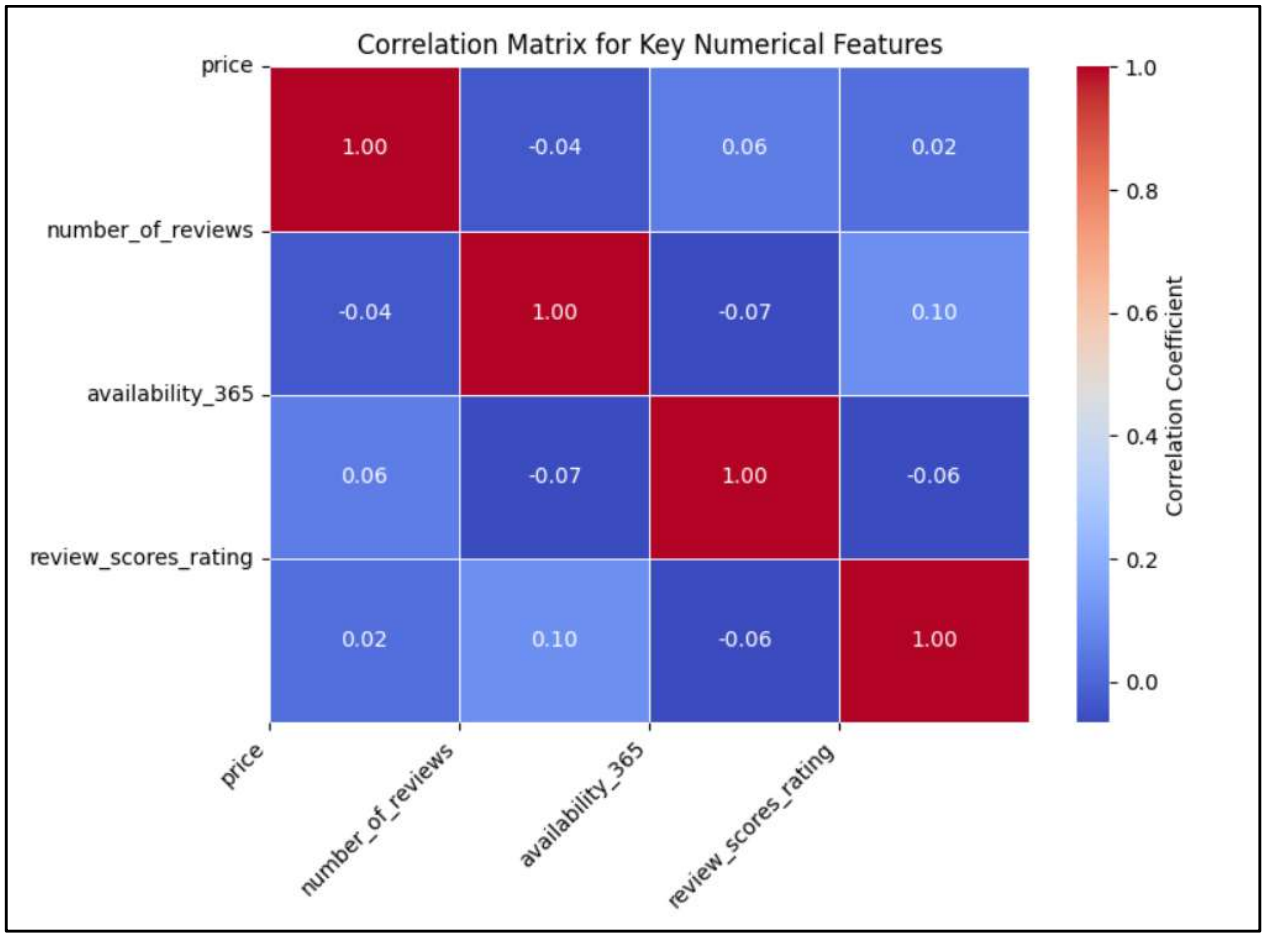
d) San Francisco



Conclusion:

Like Chicago, we can see that there is no strong or even moderate correlation between any pairs of variables. So, any change in one of these variables is not likely to affect the other variables.

e) Dallas



Conclusion:

Like Chicago & Dallas, we can see that there is no strong or even moderate correlation between any pairs of variables. So, any change in one of these variables is not likely to affect the other variables.

NOTE: Negative correlation between two variables indicates that if one variable increases the other variable will decrease.

4. Task 4: Price Analysis:

Introduction

This task investigates how prices vary between neighborhoods and room types and compares short-term vs. long-term stays. It helps figure out which neighborhoods are cheaper or more expensive and shows which areas are better for short or long stays.

Data Overview

The task checks out the listings dataset and looks at columns like price, neighbourhood_cleansed, room_type, and minimum_nights. The stay_type is decided based on the minimum_nights column—if it is over 30 nights, it is 'Long-term,' and anything less is 'Short-term.'

Approach

Step 1: Data Aggregation and Processing:

First, we added a new column, **stay_type**, based on the **minimum_nights** value to mark listings as either long-term or short-term stays. This is key to spotting price trends based on stay duration. After that, we pulled in the important columns (price, neighbourhood_cleansed, room_type, minimum_nights, maximum_nights, and stay_type) and saved them into a new DataFrame called **price_analysis_data**. Any missing values were removed to keep the analysis clean.

Step 2: Calculating Average Prices:

In this section, we looked at Price per Neighborhood and Average Price per Room Type.

- **Average Price per Neighborhood**: We figured out the average price for each neighborhood by grouping the data by neighbourhood_cleansed and calculating the average price. This gives an idea of which areas are pricier or cheaper for guests.
- **Average Price per Room Type**: We also calculated the average price for different room types (like Entire home or Private room) to see how the type of accommodation influences prices.

Step 3: Minimum and Maximum Nights Analysis:

We also figured out the average **minimum_nights** and **maximum_nights** for each neighborhood, in addition to price. This helps show whether a neighborhood is better for short-term or long-term stays based on its listing rules.

Step 4: Analyzing Short-Term and Long-Term Stays.

In this part, we looked at short-term and long-term stays.

- **Price by Neighborhood and Stay Type**: We worked out the average price for short-term and long-term stays in each neighborhood. This helps show which areas are more expensive or popular for distinct types of stays.
- **Price by Room Type and Stay Type**: We also figured out the average price for short-term and long-term stays by room type. This gives us a promising idea of how prices change depending on the room type and stay duration.

Step 5: Data Visualization

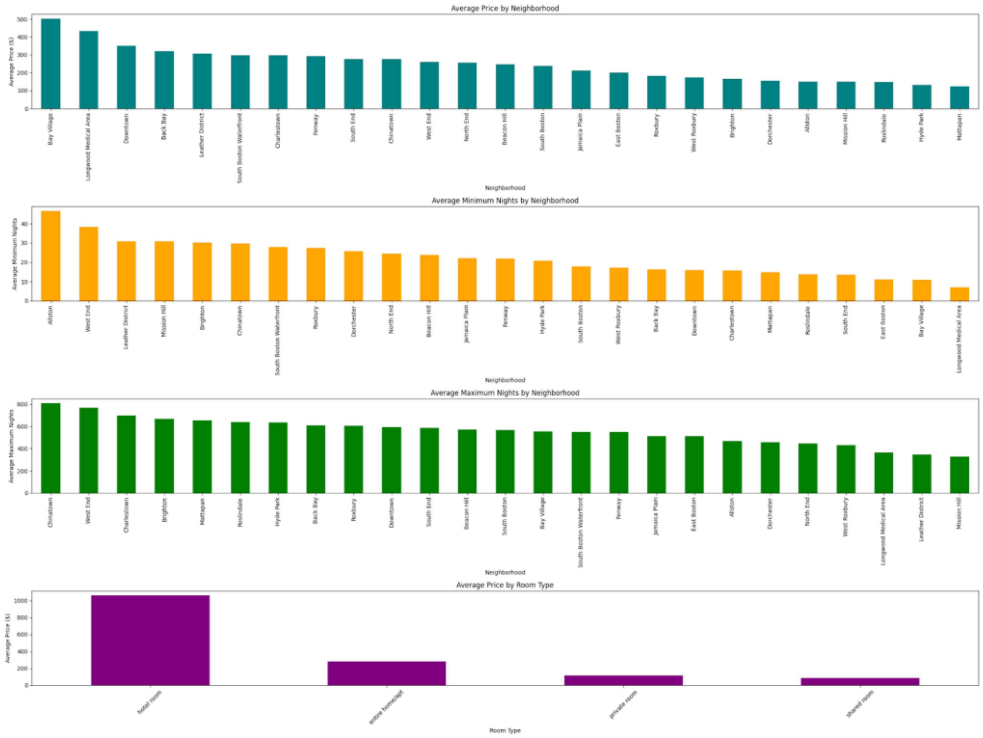
We generated multiple bar plots to visualize the results: Average price by neighborhood, Average minimum and maximum nights by neighborhood, Average price by room type, Average price by neighborhood for short-term vs. long-term stays, Average price by room types for short-term vs. long-term stays.

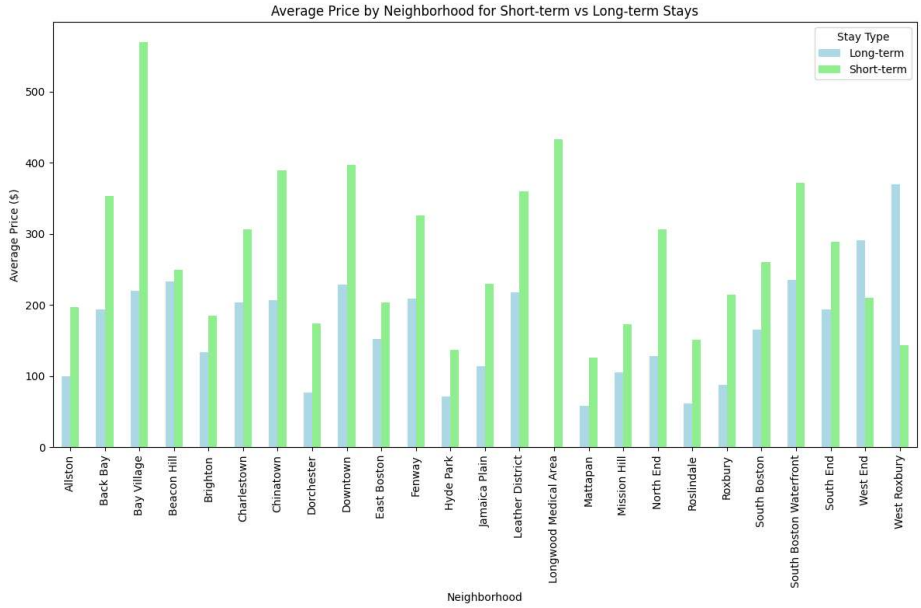
Observations

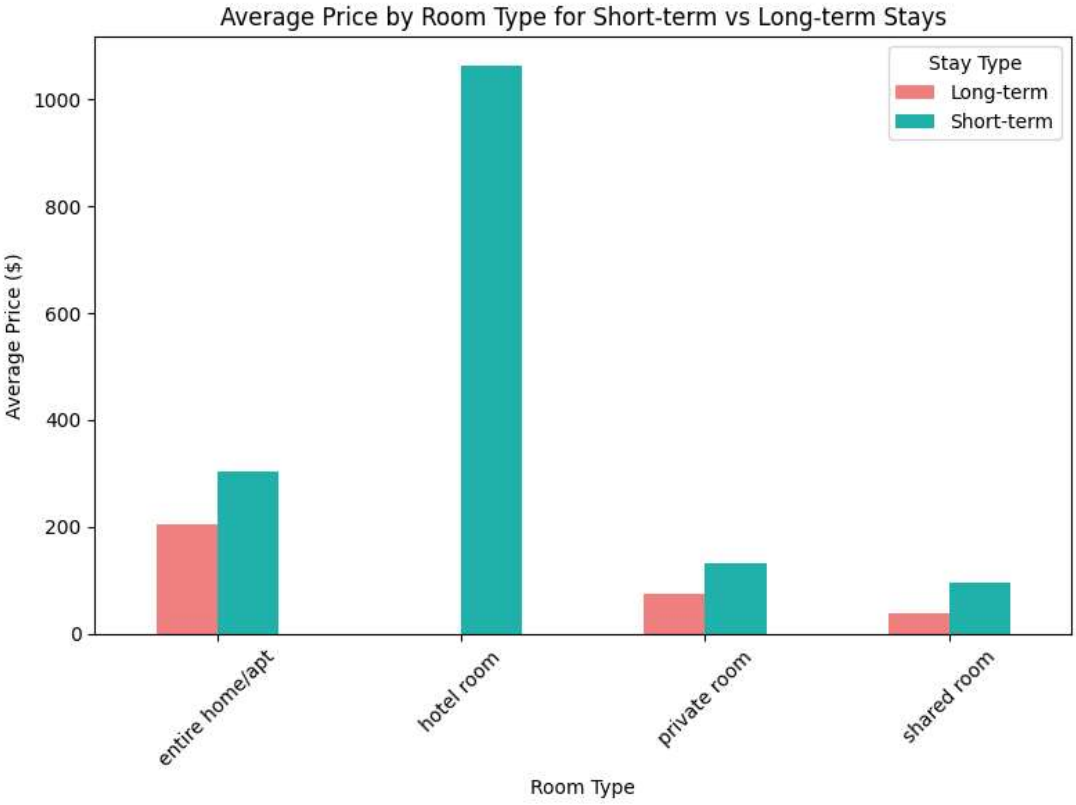
- **Neighborhoods with Higher Average Prices**: The plot for average price by neighborhood shows which areas are more expensive. This might suggest which neighborhoods are more popular with guests, possibly because of their location, amenities, or how close they are to tourist attractions. For example, in Boston, Bay Village is expensive, and Mattapan is the cheapest.
- **Neighborhoods Catering to Long-Term Stays**: The minimum_nights and maximum_nights analysis helps us see which neighborhoods are more suited for long-term stays. A higher average for these numbers means that hosts are offering longer stays. For example, in Boston, Allston has a higher average for minimum nights, so people tend to stay there longer. Meanwhile, Chinatown's average maximum nights suggest it is a popular spot for long-term stays.
- **Room Type Pricing**: The room type analysis shows how prices vary between different room types, like entire homes and private rooms. For instance, in Boston, hotel rooms are more expensive than entire homes, which is surprising.
- **Short-Term vs. Long-Term Price by neighborhood Trends**: Looking at prices for short-term and long-term stays by neighborhood and room type shows where there are big differences. Long-term stays might offer discounts, while some neighborhoods have higher prices for short-term stays because of demand.
- **Short-Term vs. Long-Term Price by room trends**: This shows that people usually prefer renting entire homes for long-term stays and hotel rooms for short-term stays.

All these observations are based on certain limitations which are provided in the 'Limitations' section under Task 4.

The following images show the analysis for all the five cities considered.
Boston:

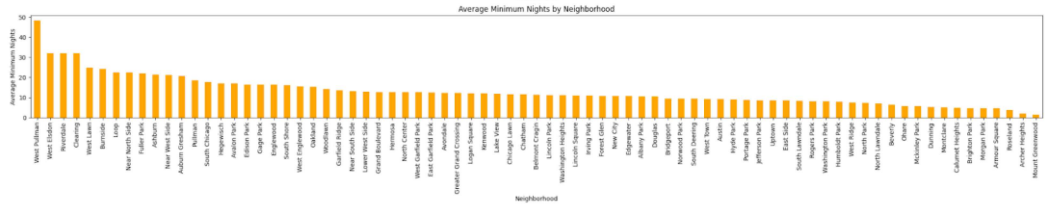


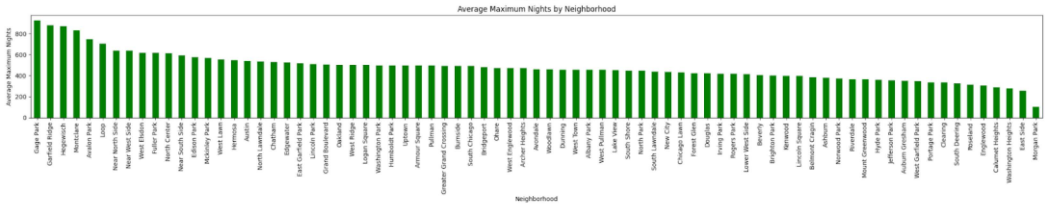



Average Price by Room Type for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
200
400
600
800
1000
entire home/apt
hotel room
private room
shared room
Room Type

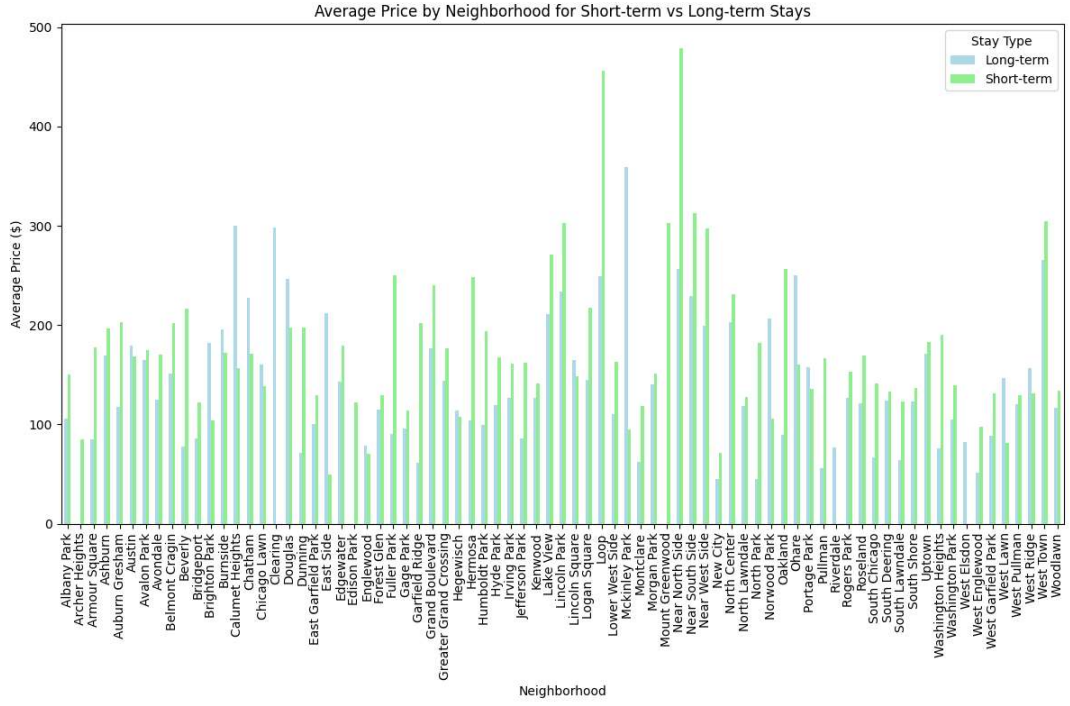
Chicago:

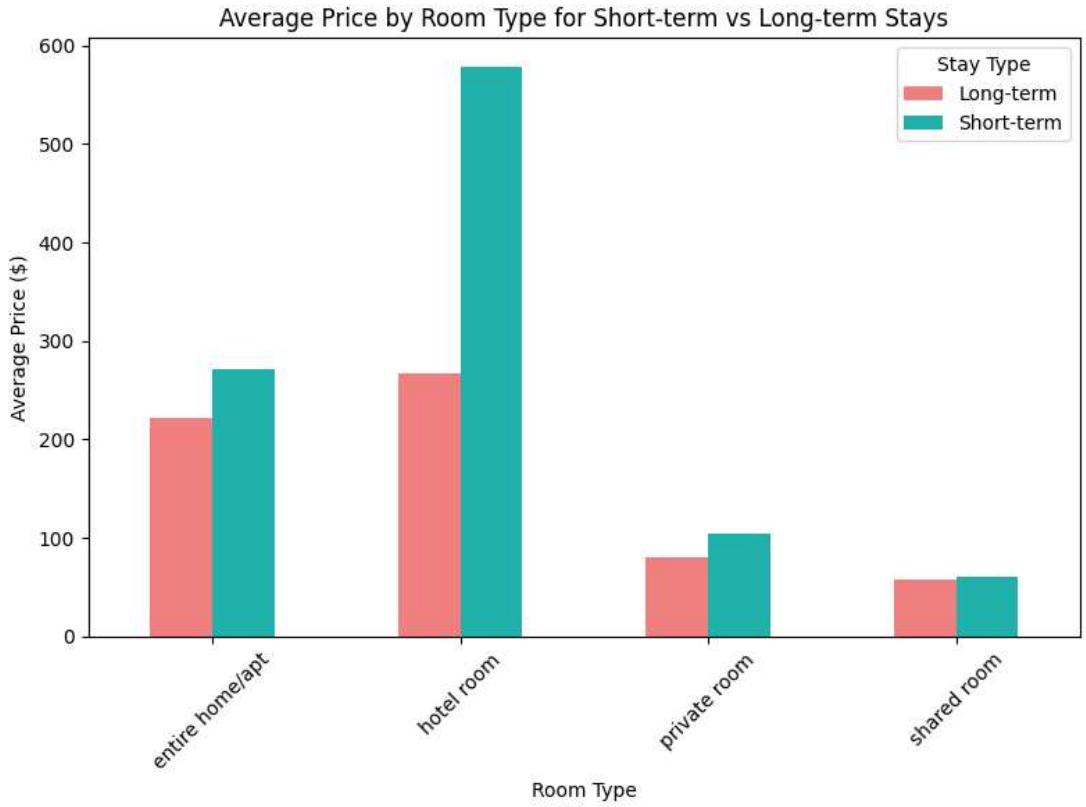

Average Price by Neighborhood
Average Price ($)
Neighborhood


Average Minimum Nights by Neighborhood
Average Minimum Nights
Neighborhood

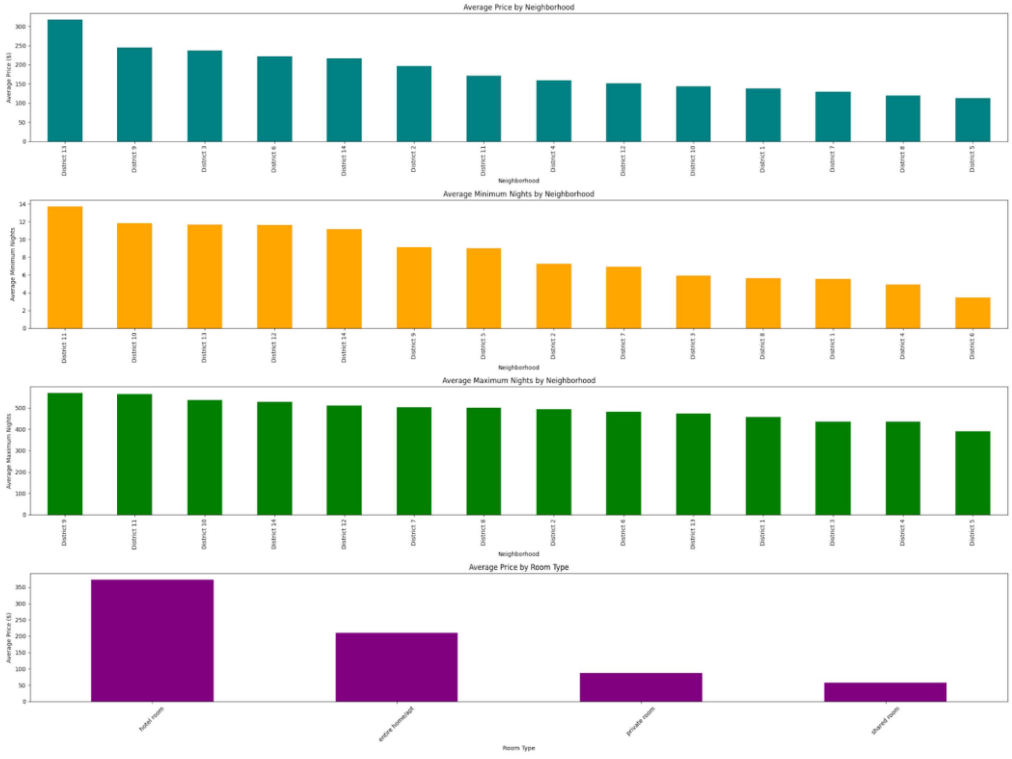

Average Maximum Nights by Neighborhood
Average Maximum Nights
Neighborhood

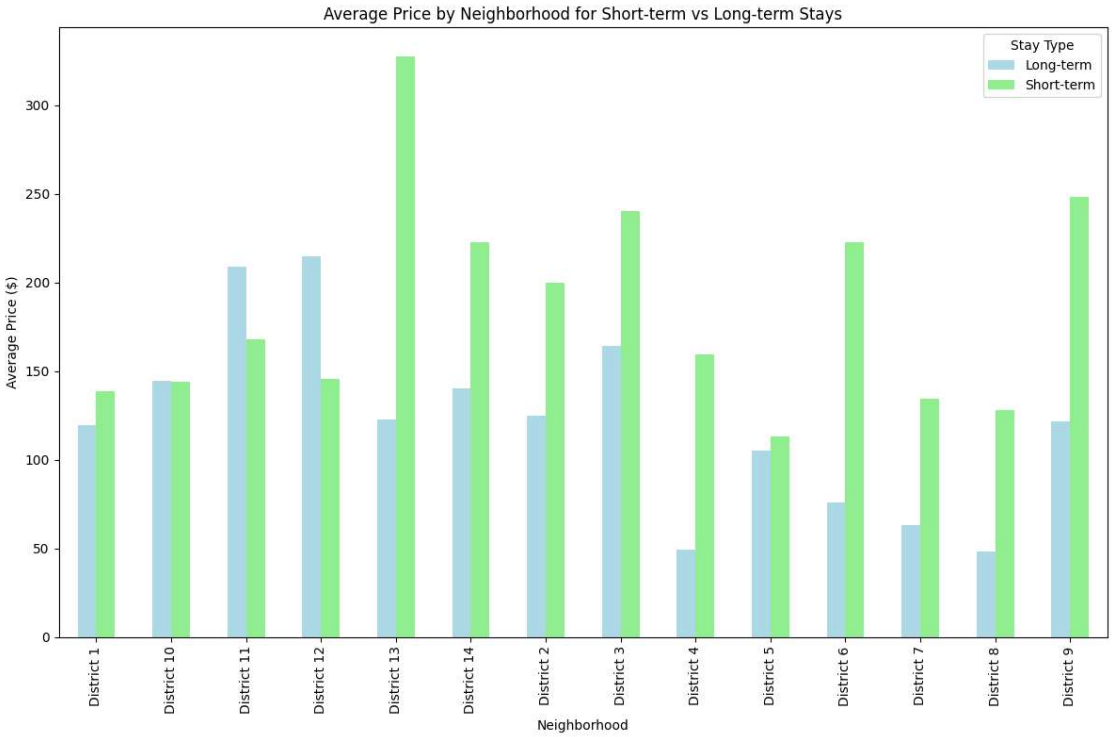

Average Price by Room Type
Average Price ($)
Hotel room
entire home/apt
private room
shared room
Room Type

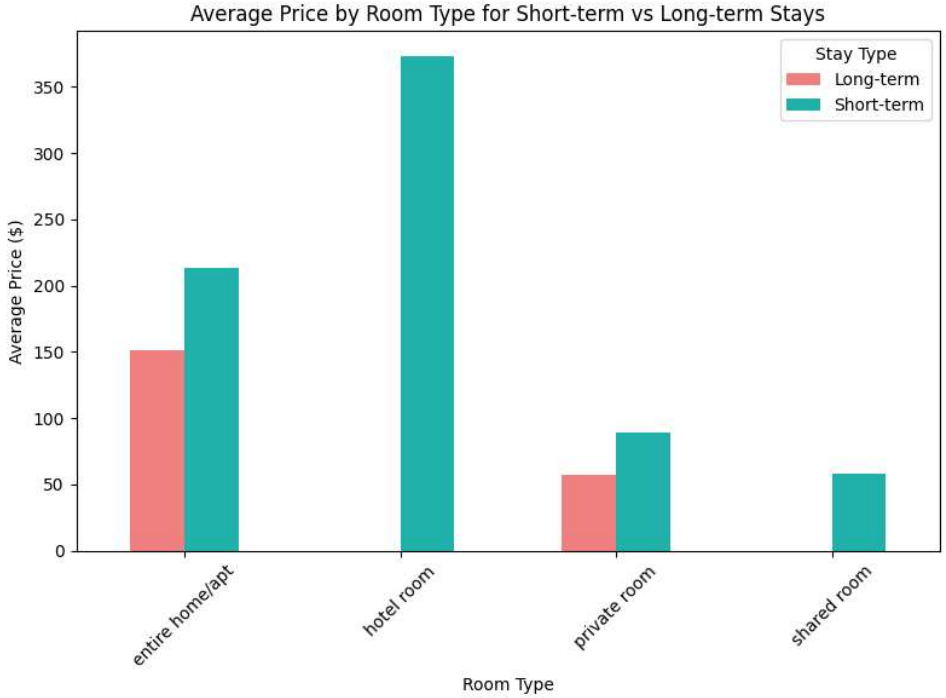

Average Price by Neighborhood for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
Albany Park
Archer Heights
Armour Square
Ashburn
Auburn Gresham
Austin
Avalon Park
Avondale
Belmont Cragin
Beverly
Bridgeport
Brighton Park
Burnside
Calumet Heights
Chatham
Chicago Lawn
Clearing
Douglas
Dunning
East Garfield Park
East Side
Edgewater
Edison Park
Englewood
Forest Glen
Fuller Park
Gage Park
Garfield Ridge
Grand Boulevard
Greater Grand Crossing
Hegewisch
Hermosa
Humboldt Park
Hyde Park
Irving Park
Jefferson Park
Kenwood
Lake View
Lincoln Park
Lincoln Square
Logan Square
Loop
Lower West Side
Mckinley Park
Montclare
Morgan Park
Mount Greenwood
Near North Side
Near South Side
Near West Side
New City
North Center
North Lawndale
North Park
Norwood Park
Oakland
Ohare
Portage Park
Pullman
Riverdale
Rogers Park
Roseland
South Chicago
South Deering
South Lawndale
South Shore
Uptown
Washington Heights
Washington Park
West Elsdon
West Englewood
West Garfield Park
West Lawn
West Pullman
West Ridge
West Town
Woodlawn
Neighborhood


Average Price by Room Type for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
100
200
300
400
500
600
entire home/apt
hotel room
private room
shared room
Room Type

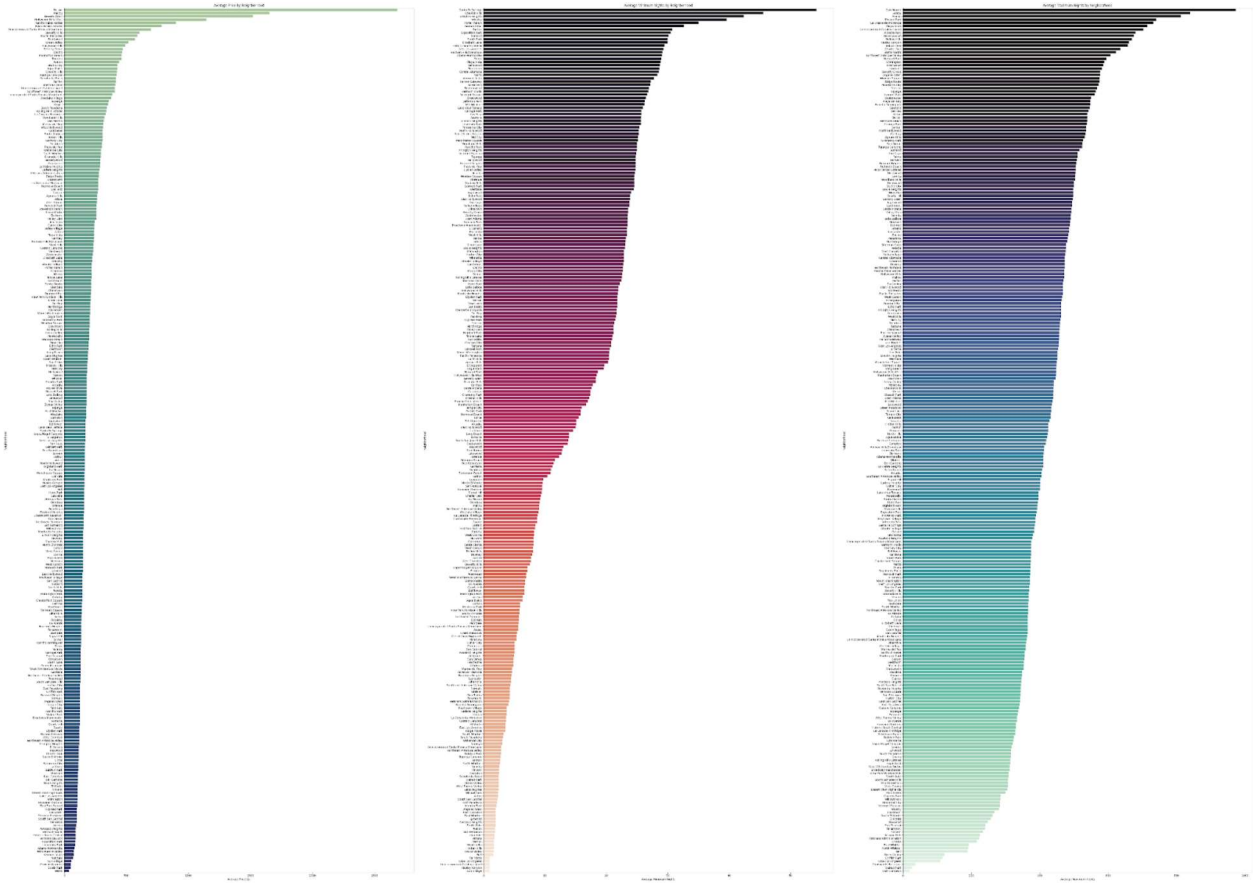
Dallas:

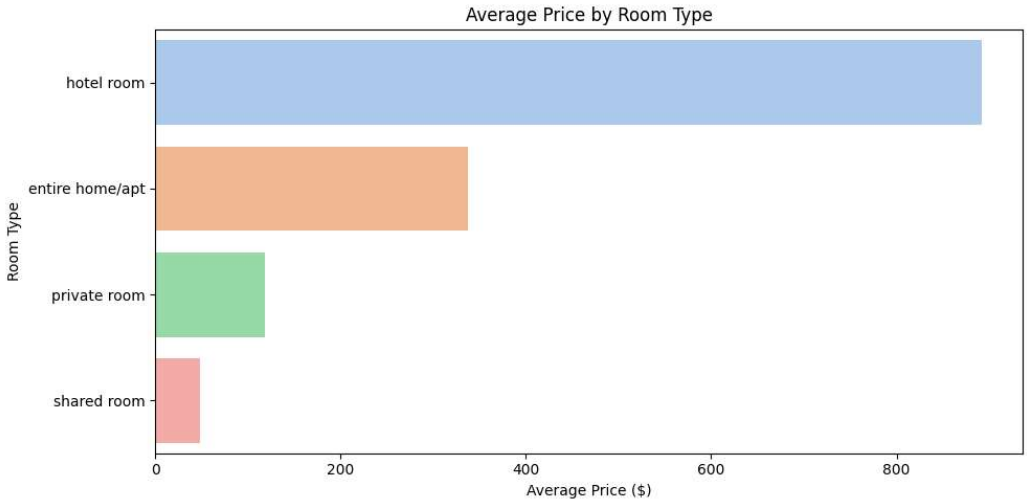

Average Price by Neighborhood
Average Minimum Nights by Neighborhood
Average Maximum Nights by Neighborhood
Average Price by Room Type
Neighborhood
Room Type

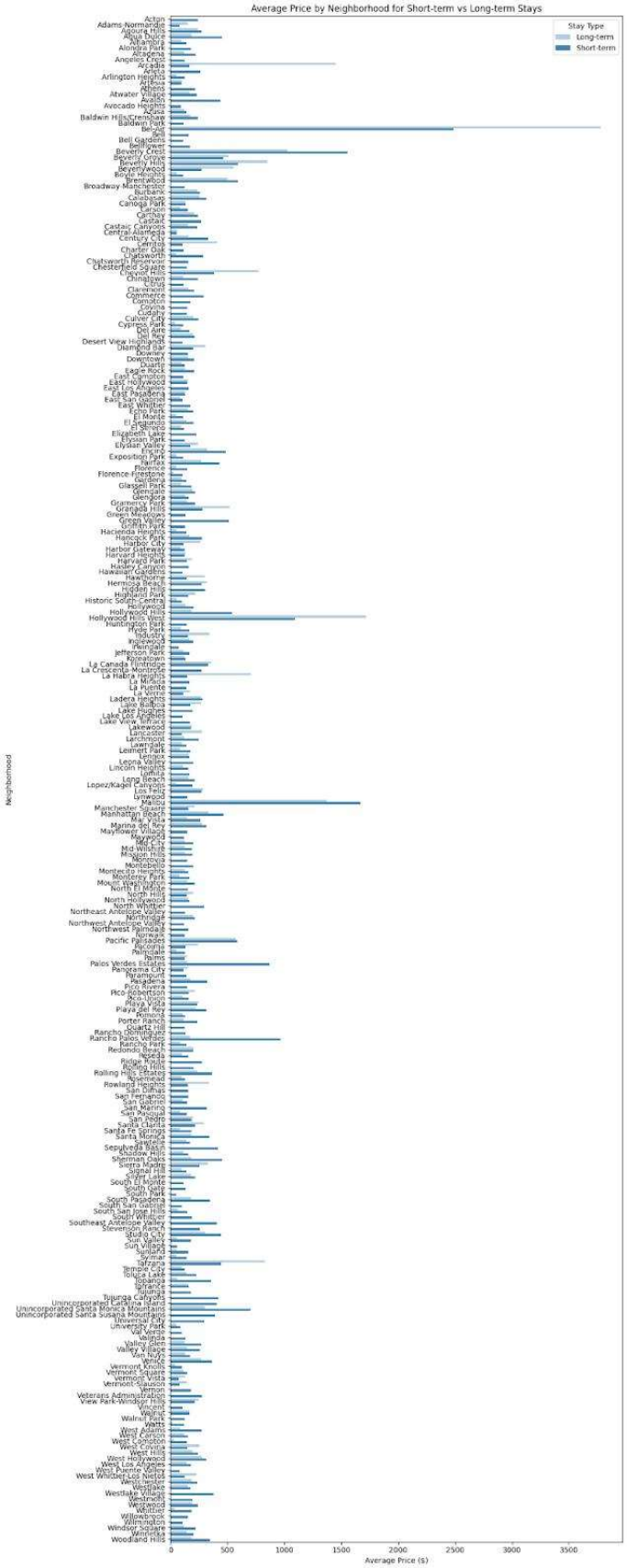

Average Price by Neighborhood for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
50
100
150
200
250
300
District 1
District 10
District 11
District 12
District 13
District 14
District 2
District 3
District 4
District 5
District 6
District 7
District 8
District 9
Neighborhood

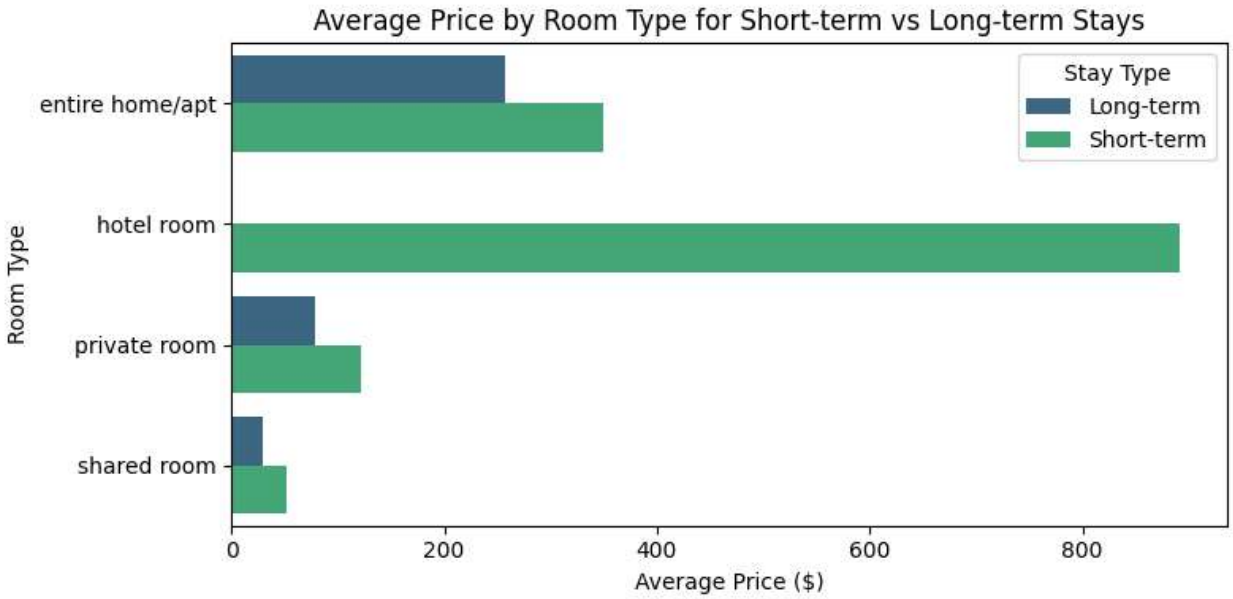

Average Price by Room Type for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
50
100
150
200
250
300
350
entire home/apt
hotel room
private room
shared room
Room Type

Los Angeles:

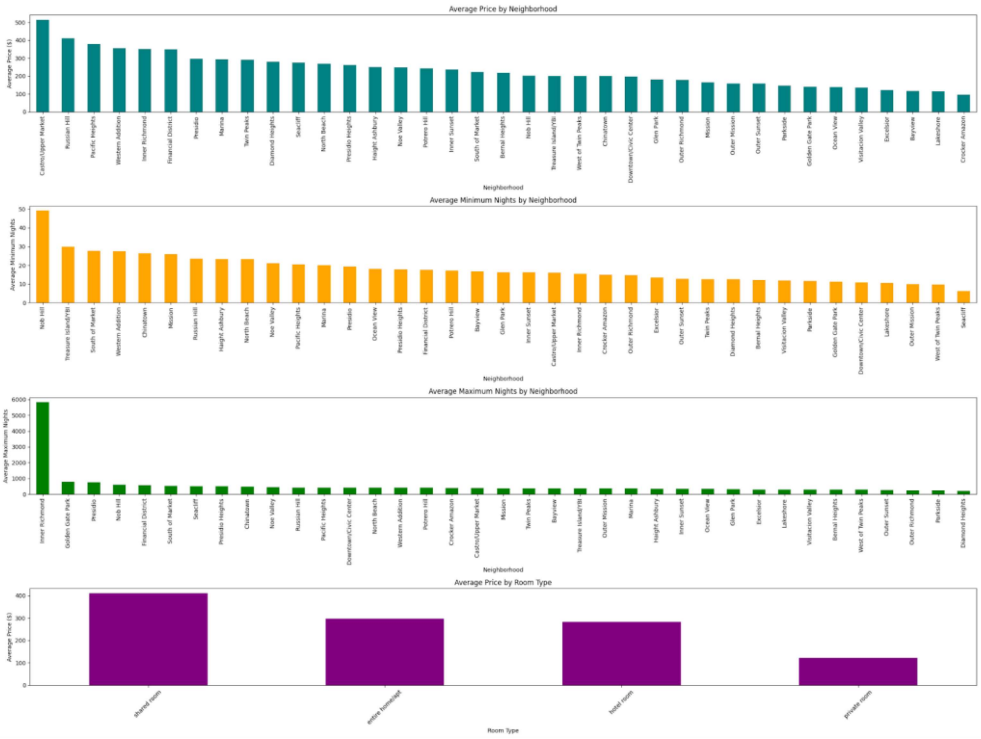


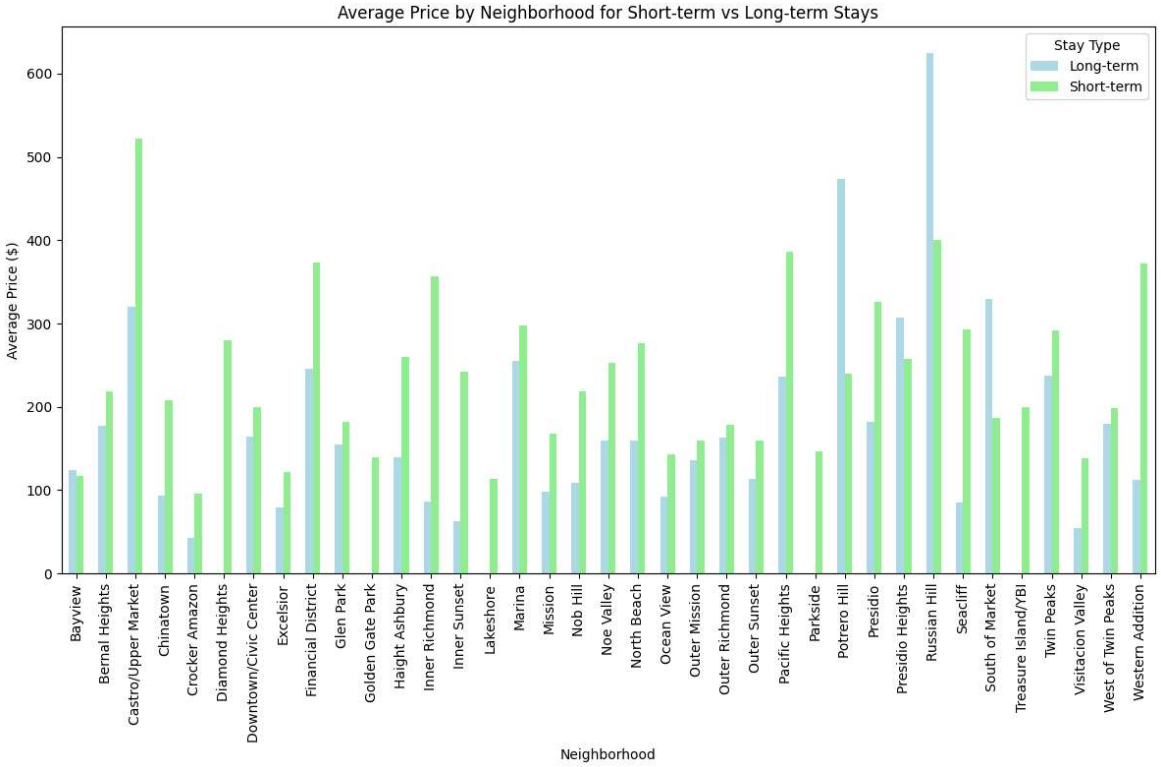

Average Price by Room Type
hotel room
entire home/apt
private room
shared room
Room Type
0
200
400
600
800
Average Price ($)

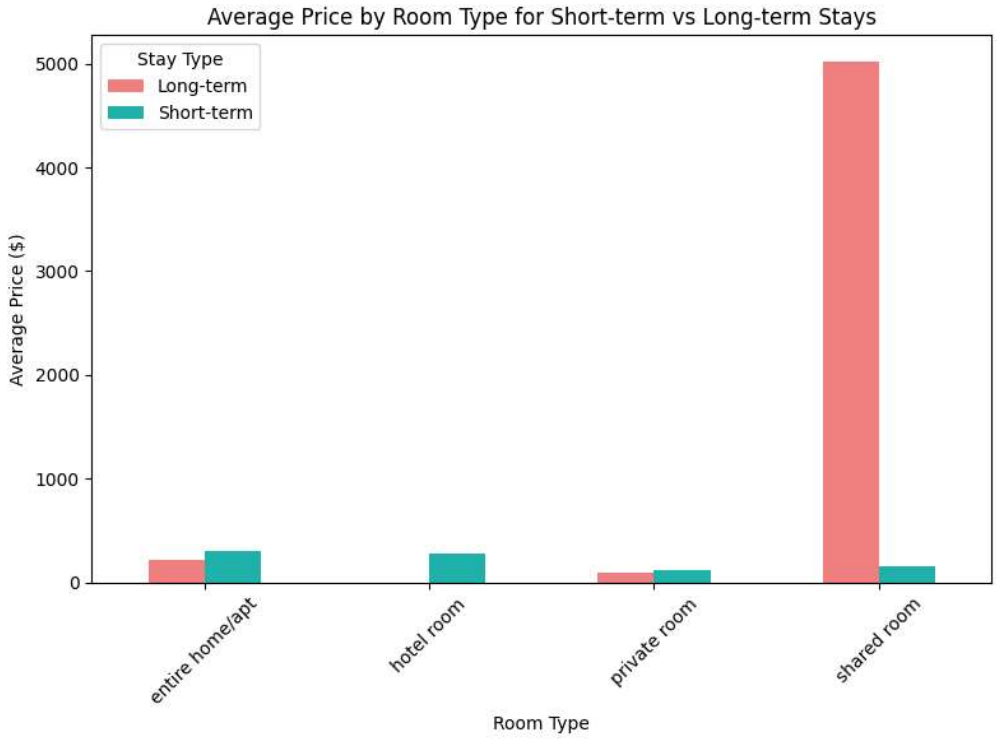

Average Price by Neighborhood for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Neighborhood
Average Price ($)
0
500
1000
1500
2000
2500
3000
3500



Sans Francisco:




Average Price by Neighborhood for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
100
200
300
400
500
600
Bayview
Bernal Heights
Castro/Upper Market
Chinatown
Crocker Amazon
Diamond Heights
Downtown/Civic Center
Excelsior
Financial District
Glen Park
Golden Gate Park
Haight Ashbury
Inner Richmond
Inner Sunset
Lakeshore
Marina
Mission
Nob Hill
Noe Valley
North Beach
Ocean View
Outer Mission
Outer Richmond
Outer Sunset
Pacific Heights
Parkside
Potrero Hill
Presidio
Presidio Heights
Russian Hill
Seacliff
South of Market
Treasure Island/YBI
Twin Peaks
Visitacion Valley
West of Twin Peaks
Western Addition
Neighborhood


Average Price by Room Type for Short-term vs Long-term Stays
Stay Type
Long-term
Short-term
Average Price ($)
0
1000
2000
3000
4000
5000
entire home/apt
hotel room
private room
shared room
Room Type

Conclusion

- **Average Price by Neighborhood:** Expensive neighborhoods are mostly found in city centers or fancy areas, which attract costly listings. These areas have things people want, so the prices are higher. Cheaper neighborhoods, on the other hand, are good for people who want to save money, offering affordable places to stay.
- **Average Minimum Nights by Neighborhood:** Certain neighborhoods are all about long-term stays, probably because they are near universities, or just because people like living there.
- **Average Maximum Nights by Neighborhood:** Neighborhoods where you can stay the longest tend to be more popular for extended visits, attracting business travelers or people moving for work or school. But areas with shorter stays are perfect for tourists, hospital visitors, or folks on quick trips.
- **Average Price by Room Type:** Hotel rooms are much pricier than other types of accommodations, likely because they offer more services or luxury features.
- **Average Price by Neighborhood for Short-term vs Long-term Stays:** Short-term stays are always more expensive, especially in neighborhoods with a lot of tourist or business traffic. But if you are staying long-term, premium neighborhoods usually offer way better rates, which is great for extended stays or if you are relocating.

5. Task 5: Neighborhood Comparison

i. Introduction

The task aims to compare the average 'review_score_rating' across various neighborhoods in the 'listings' dataset. The main objective is to determine which neighborhoods consistently receive higher or lower ratings from guests, offering insights to improve the guest experience.

ii. Data Overview

The 'listings' dataset contains 'review_score_rating' provided by guests for various listings. These ratings reflect their experiences of the properties and can be utilized to identify trends associated with neighborhoods. Since we are currently looking at only five cities, we used the data from Inside Airbnb for these cities (for reference, check the Appendix).

iii. Approach

The approach consisted of the following steps:

- **Step 1: Data Aggregation**

The first step was to group the listings data based on neighborhood to calculate the average review score ratings for each neighborhood. The missing values in review scores were removed to ensure that only relevant data was utilized for the analysis. The aggregated data was then sorted in descending order, highlighting the neighborhoods with the best average guest ratings.

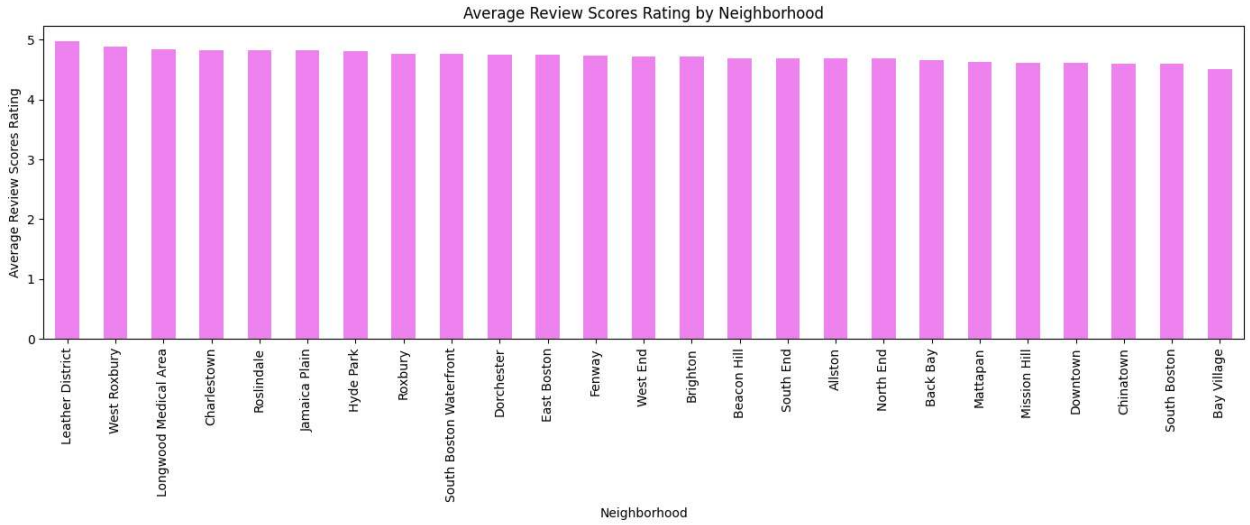
- **Step 2: Data Visualization**

The second step was to visualize the average review score ratings to provide a pictorial insight into the data, which can be used to determine neighborhood trends.

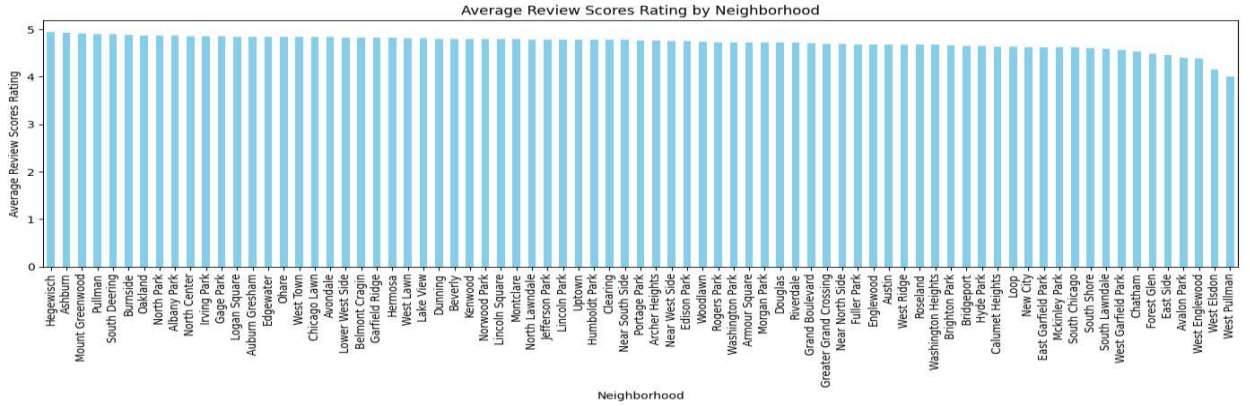
iv. Observations

- The visualization clearly distinguishes neighborhoods with higher guest satisfaction based on review score ratings.
- Neighborhoods with low average ratings could indicate areas where guests have experienced discomfort or issues with the listings. This is because guests are likely to rate a listing poorly if they encounter problems during their stay.
- All these observations are based on certain limitations which are provided in the 'Limitations' section under Task 5.
- The following images show the average review scores for different neighborhoods for all the five cities considered.

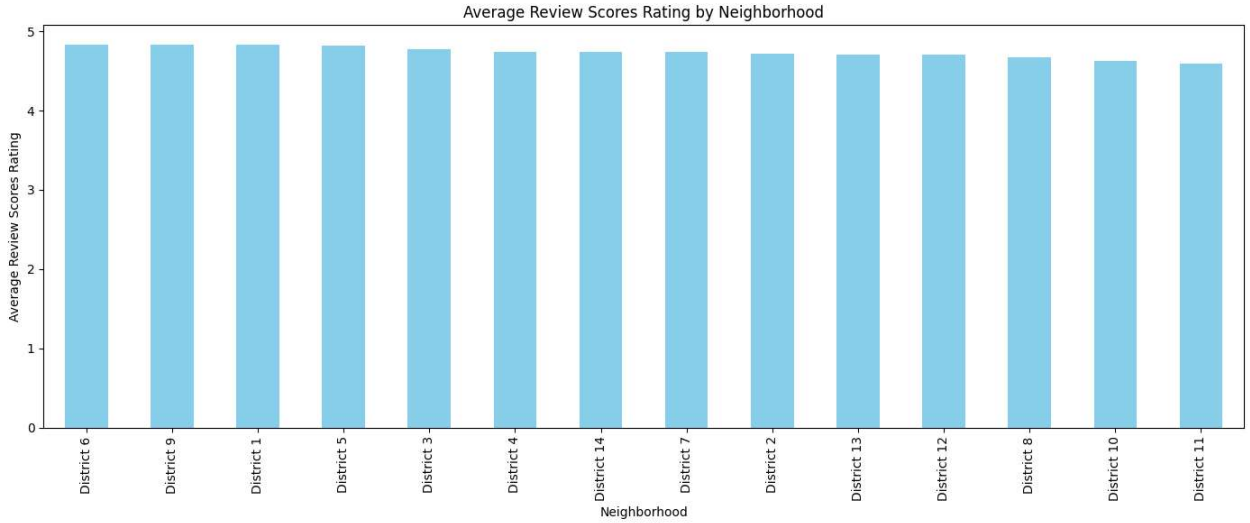
Boston -



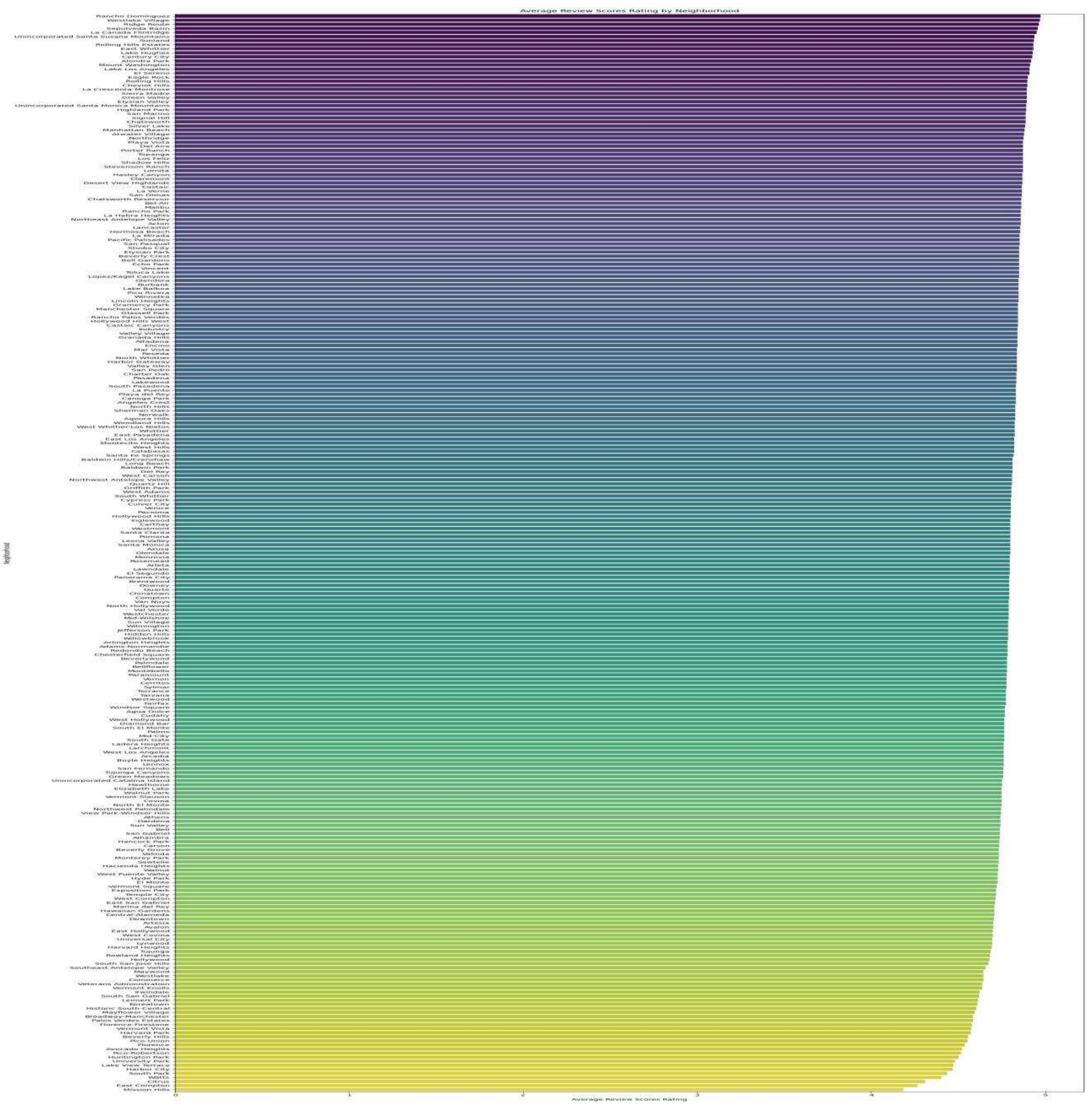
Chicago -



Dallas -

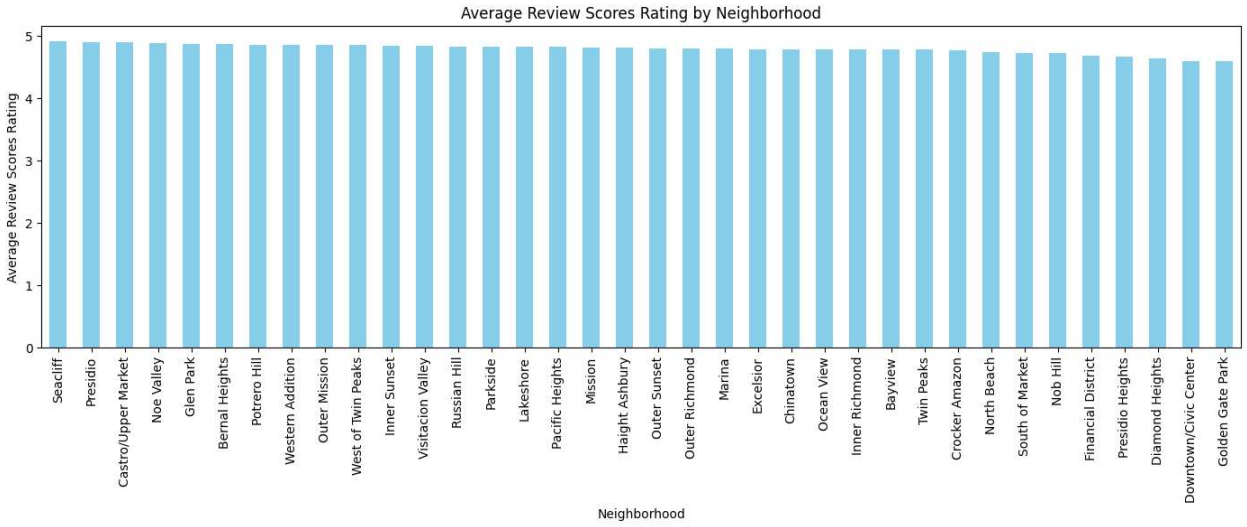


Los Angeles -



Average Review Scores Rating by Neighborhood
0
1
2
3
4
5
Average Review Scores Rating

San Francisco -



v. Conclusion

To conclude the observations, we identified neighborhoods that consistently receive higher ratings, reflecting better guest experiences, property conditions, and neighborhood advantages. Similarly, neighborhoods with low ratings indicate poor service conditions. These insights help prioritize areas requiring improvement, such as offering better amenities or enhancing the quality of listings and guest experience. A detailed analysis could be conducted in the future to explore more factors like amenities and proximity to tourist attractions to understand more about what drives guests towards a neighborhood.

6. Task 6: Outlier Detection

Introduction:

The goal of this task is to:

1. Find outliers in the 'price', 'minimum_nights', and 'review_scores_rating' columns using two common methods: the Interquartile Range (IQR) method and the Z-Score method.
2. Visualize the outliers to get a clearer picture of where the extreme values show up in the dataset.
3. Get rid of the outliers to clean up the dataset so extreme values do not mess up the analysis.

Data Overview:

The dataset used to detect outliers is a property listings dataset with info like price, minimum booking nights, and guest review scores.

Approach:

Step 1: Data Preparation:

- We are going to focus on these columns to detect outliers: price, minimum_nights, and review_scores_rating.
- Any rows with missing values in these key columns should be deleted to make sure our outlier detection is accurate.

Step 2: Detect Outliers Using the Interquartile Range (IQR) Method:

- Calculate Quartiles: For each column, calculate the first quartile (Q1) and the third quartile (Q3). These represent the 25th and 75th percentiles of the data.
- Compute the IQR: The IQR is calculated as the difference between Q3 and Q1 (IQR = Q3 - Q1). This represents the range of the middle 50% of the data.
- Define Outlier Boundaries:
    - The lower bound is Q1 - 1.5 * IQR.
    - The upper bound is Q3 + 1.5 * IQR.
    - Any value outside this range (below the lower bound or above the upper bound) is considered an outlier.
- Apply these boundaries to the price, minimum_nights, and review_scores_rating columns to identify extreme values.

Step 3: Detect Outliers Using the Z-Score Method

- Standardize Data: Calculate the Z-scores for each column. The Z-score represents how many standard deviations a data point is from the mean:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X is the data point,
- μ is the mean,
- σ is the standard deviation.
- Define Outlier Threshold: Set a Z-score threshold (commonly 3). Any data point with a Z-score greater than 3 (or less than -3) will be considered an outlier.
- For each column (price, minimum_nights, and review_scores_rating), identify the rows where the Z-score exceeds the threshold.

Step 4: Compare the Results

- **Summary Statistics:** After removing outliers using the IQR and Z-Score methods, calculate the summary statistics—mean, median, standard deviation, etc.—for the cleaned datasets. This will help assess the impact of outlier removal on the data distribution.
- **Compare Number of Outliers:** Check how many outliers each method catches to figure out which one picks up more extreme values.
- **Evaluate Methods:** Depending on the dataset's characteristics (whether it is skewed or normally distributed), figure out whether IQR or Z-Score is better for this analysis.

Step 5: Visualize the Cleaned Data

After taking out the outliers, use histograms or density plots to check the distribution of the cleaned features like price, minimum_nights, and review_scores_rating. This helps ensure there are not any more extreme outliers, and that the data looks good for the next steps.

Observations:

IQR Method Observations:

- **Price:** The IQR method detected many high-priced outliers in neighborhoods known for luxury properties. These outliers were well above the upper bound and contributed to the

price distribution's skewness. Meanwhile, some listings in less central areas had incredibly low prices, which were also flagged as outliers.

- **Minimum Nights:** The IQR method spotted some extreme values for minimum_nights, especially in listings that required super long stays (like over 365 days). These were marked as outliers since most places had much shorter minimum stays. On the other side, there were a few single-night stay outliers in areas where you usually expect longer-term stays.
- **Review Scores Rating:** Outliers in review_scores_rating were less common than in price and minimum nights. Some properties with incredibly low ratings (under 50%) were outliers. Properties with nearly perfect scores (close to 100%) were also flagged, but they may be suitable places, not just statistical outliers.

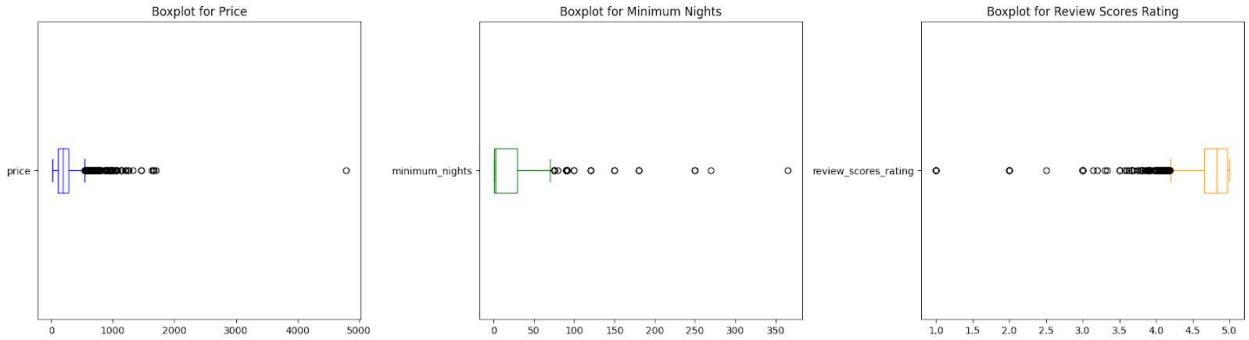
Z-Score Method Observations:

- **Price:** The Z-Score method caught more high-priced outliers than the IQR method since it is more sensitive to data with a lot of variation. Any listings with Z-Scores above 3 (more than three standard deviations from the mean) were considered outliers. There were fewer low-priced listings flagged by the Z-Score method compared to the IQR method, which suggests that cheaper listings are more tightly packed together.
- **Minimum Nights:** Like the IQR method, the Z-Score method also found long-stay listings as outliers, where the minimum nights were much higher than the average. The Z-Score method was more sensitive to short-stay listings and flagged more of them as outliers than the IQR method.
- **Review Scores Rating:** The Z-Score method flagged fewer review score outliers than the IQR method, suggesting that review scores are more consistent and do not show big differences from the mean.

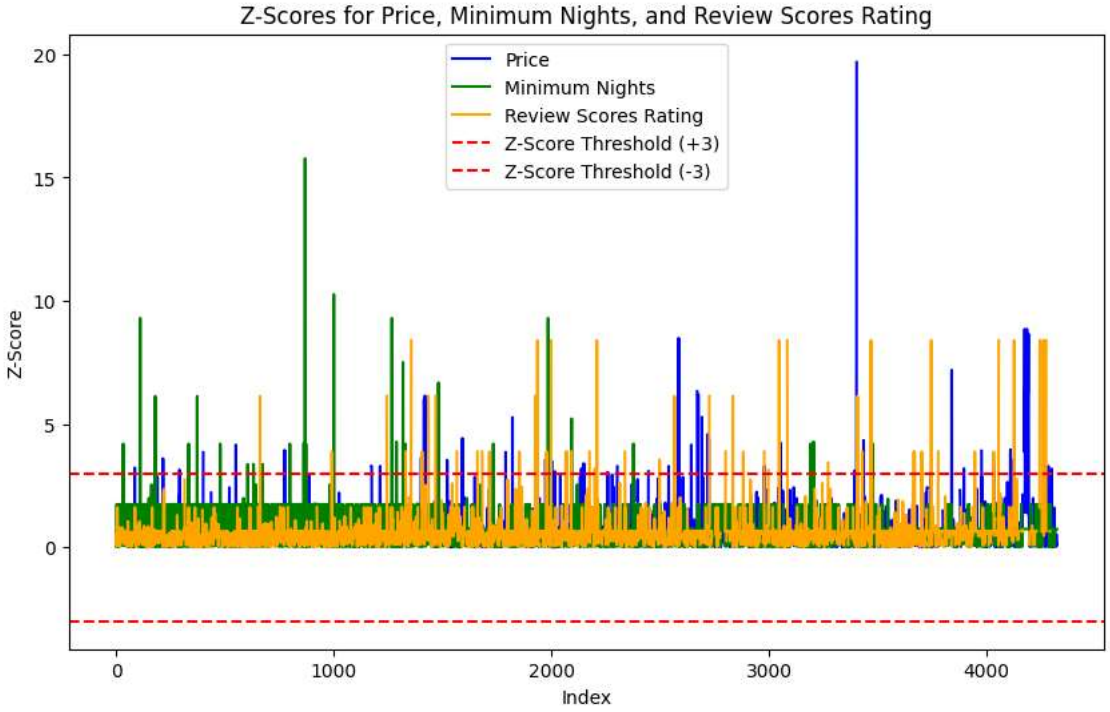
All these observations are based on certain limitations which are provided in the 'Limitations' section under Task 4.

The following images show the analysis for all the five cities considered.

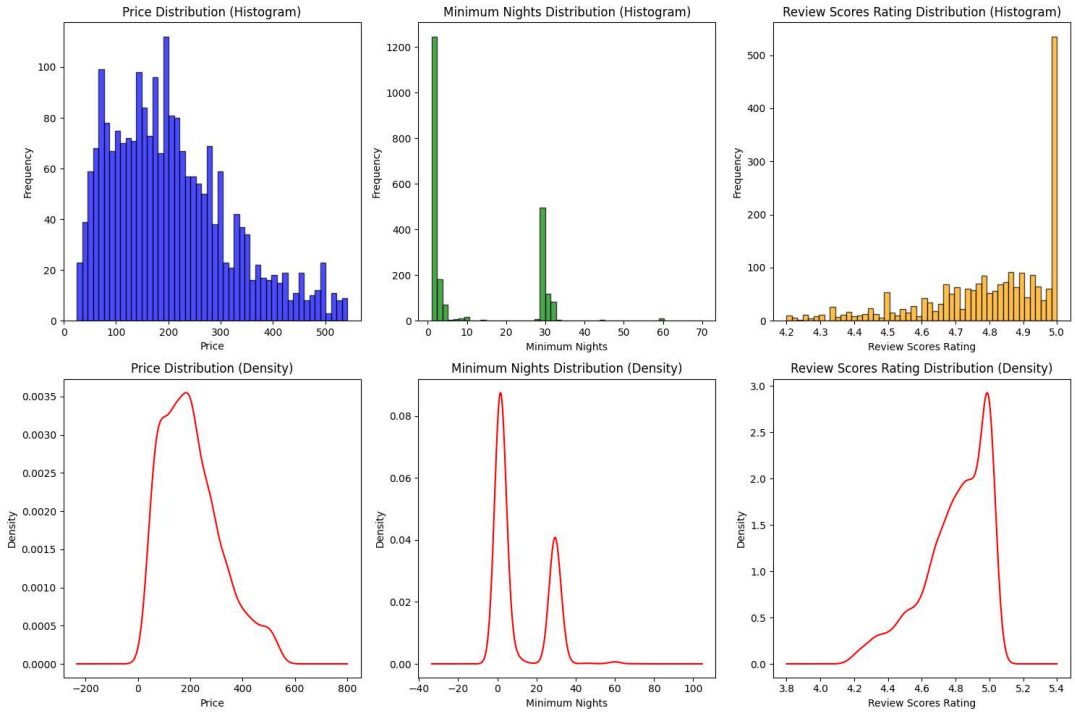
Boston:
Detecting outliers using IQR:



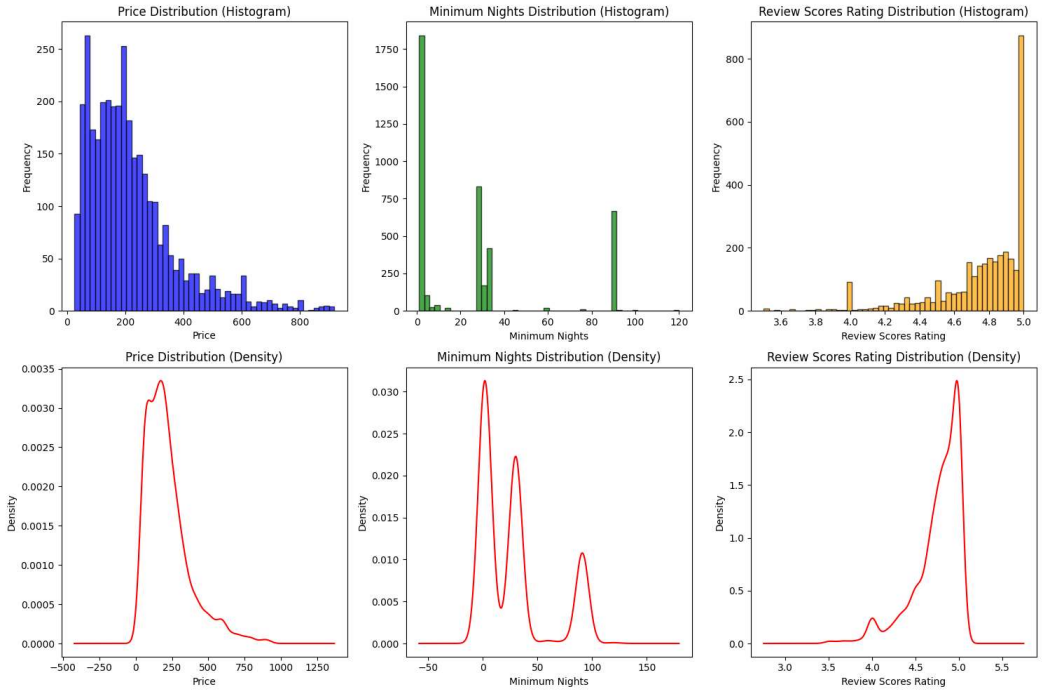
Detecting outliers using Z:



Visualizing after outlier removal using IQR:

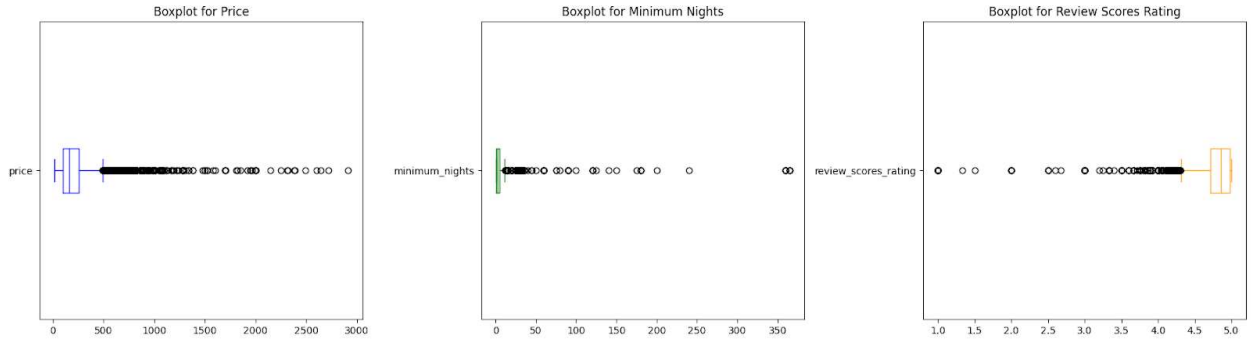


Visualizing after outlier removal using Z Score:

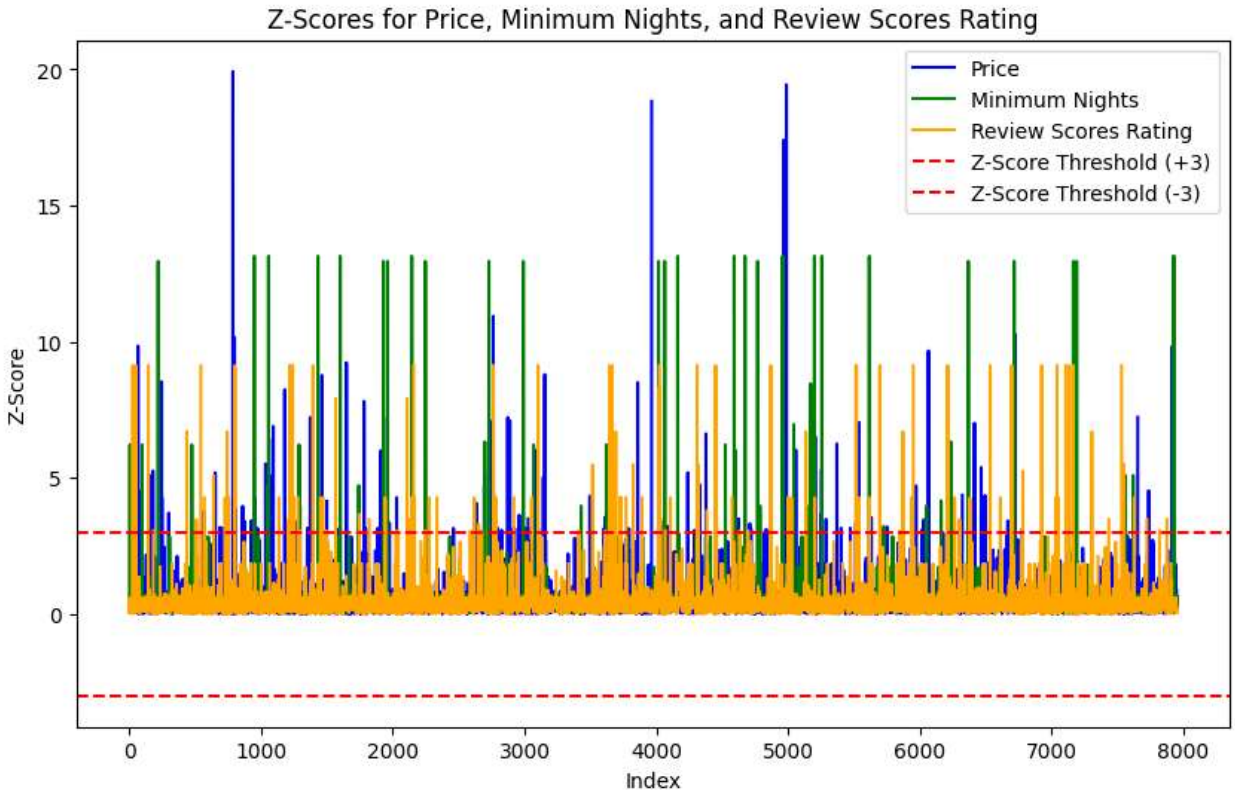


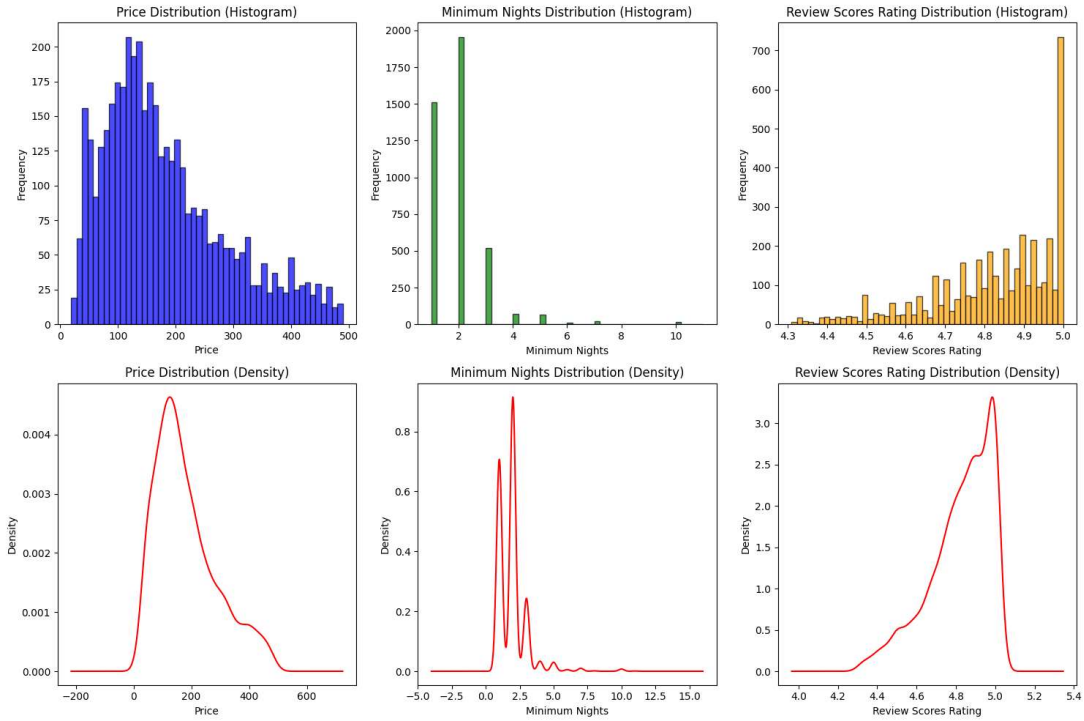
Chicago:

Detecting outliers using IQR:

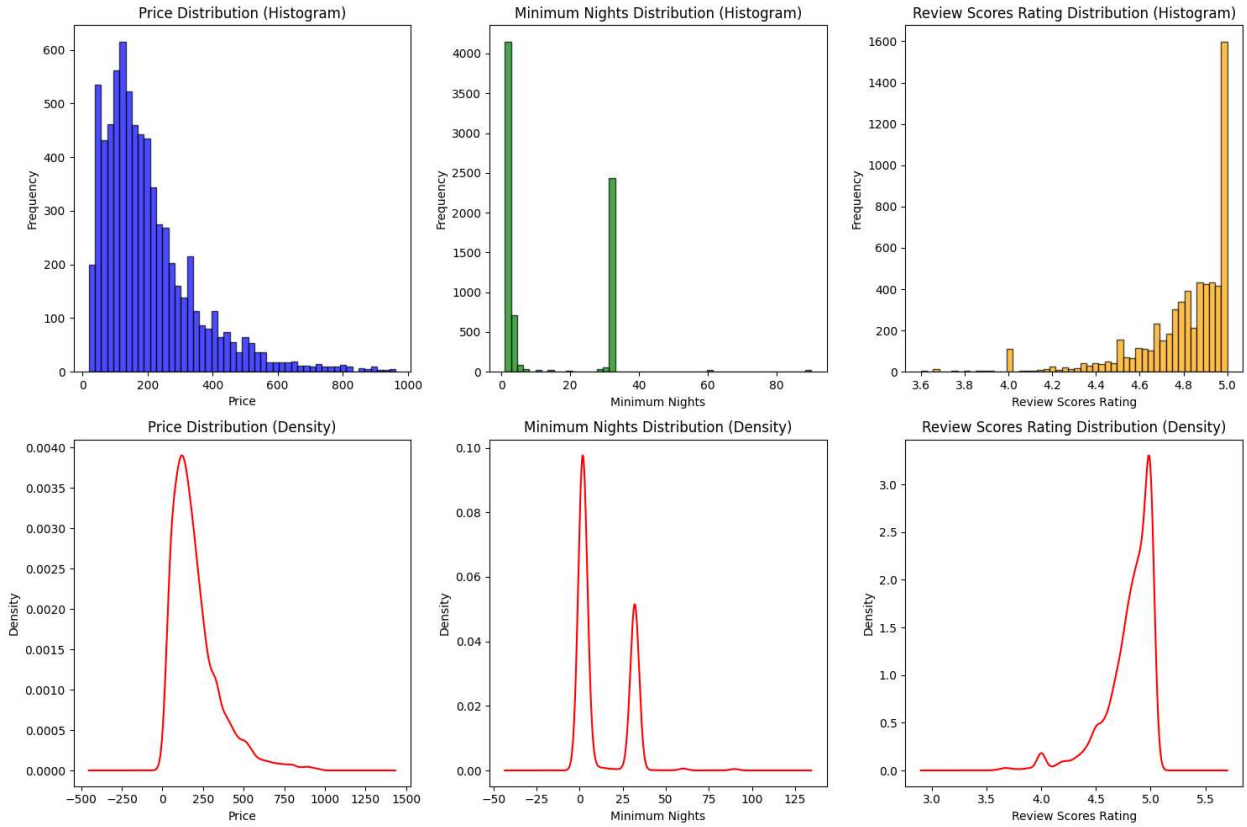


Outlier Detection using Z Score Technique:



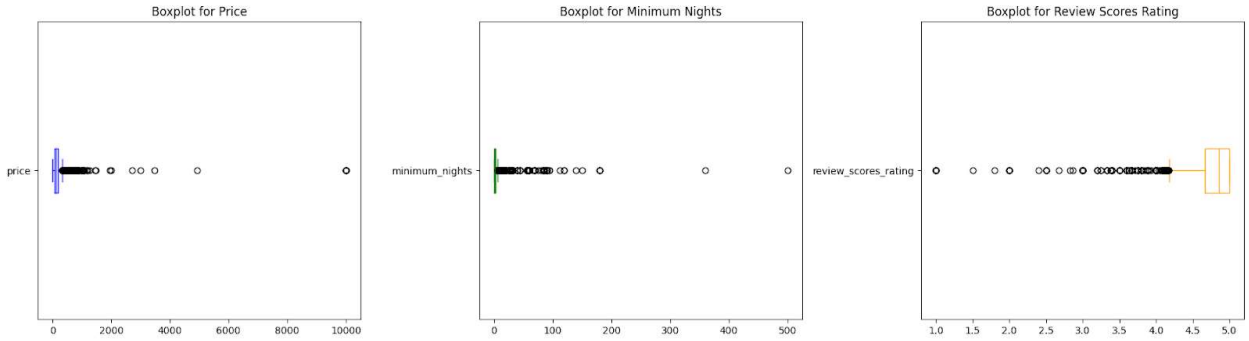


Visualizing after outlier removal using Z Score:

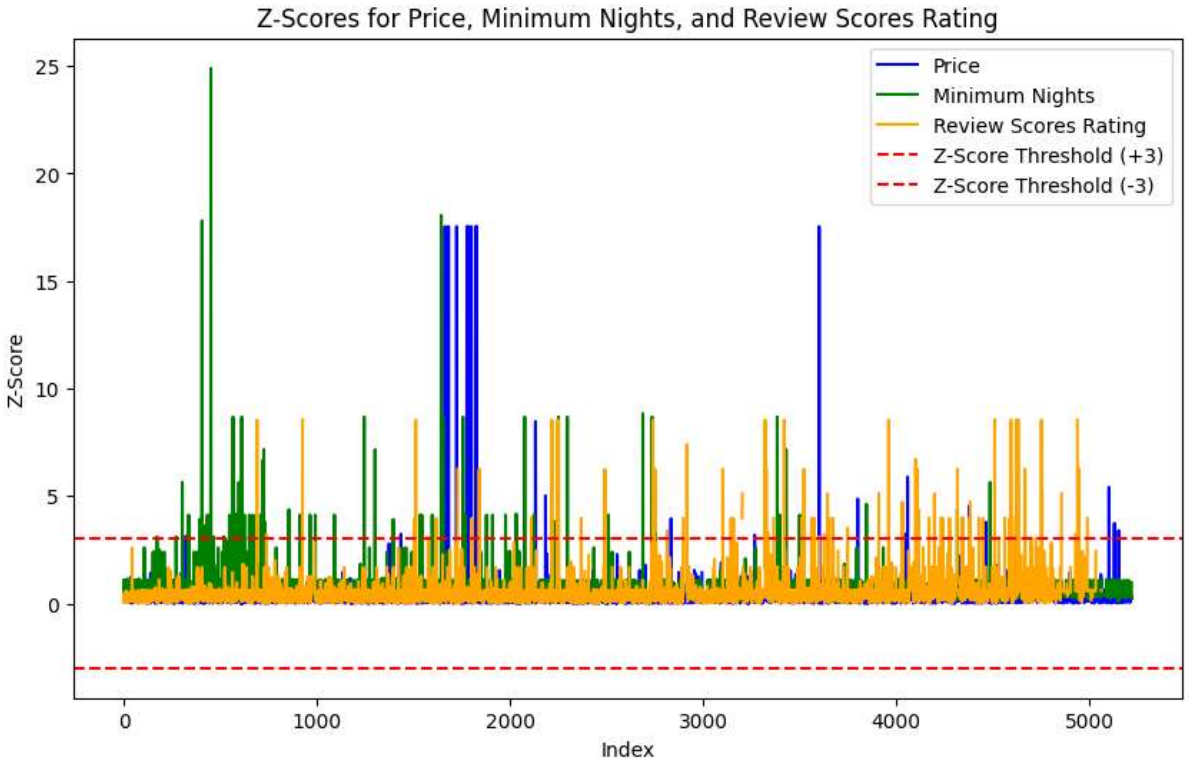


Dallas:

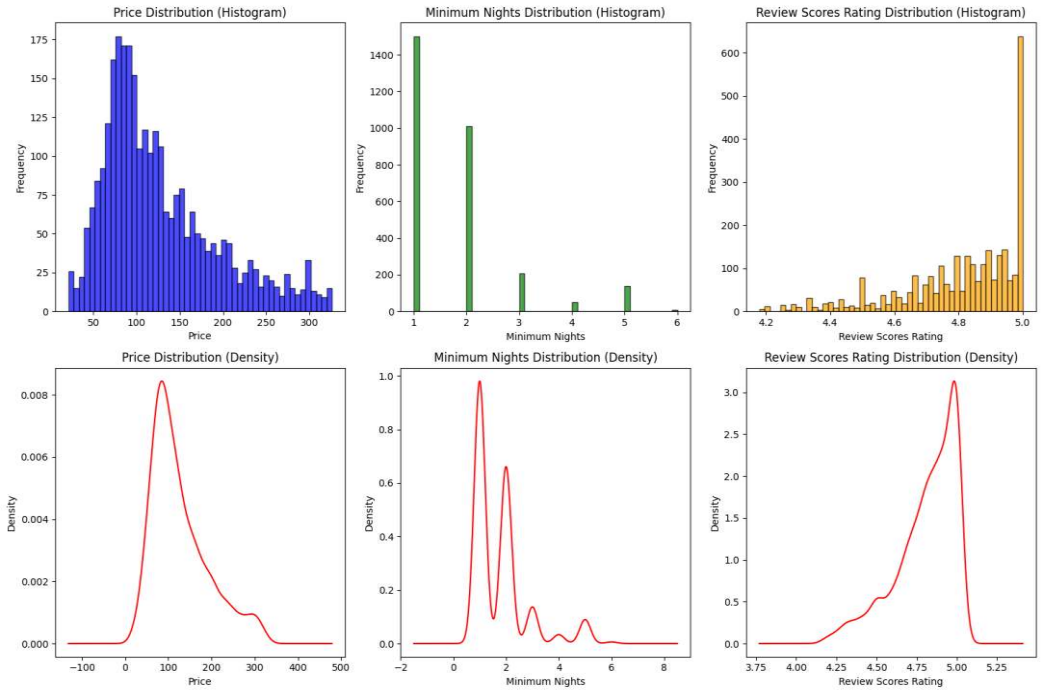
Detecting outliers using IQR:



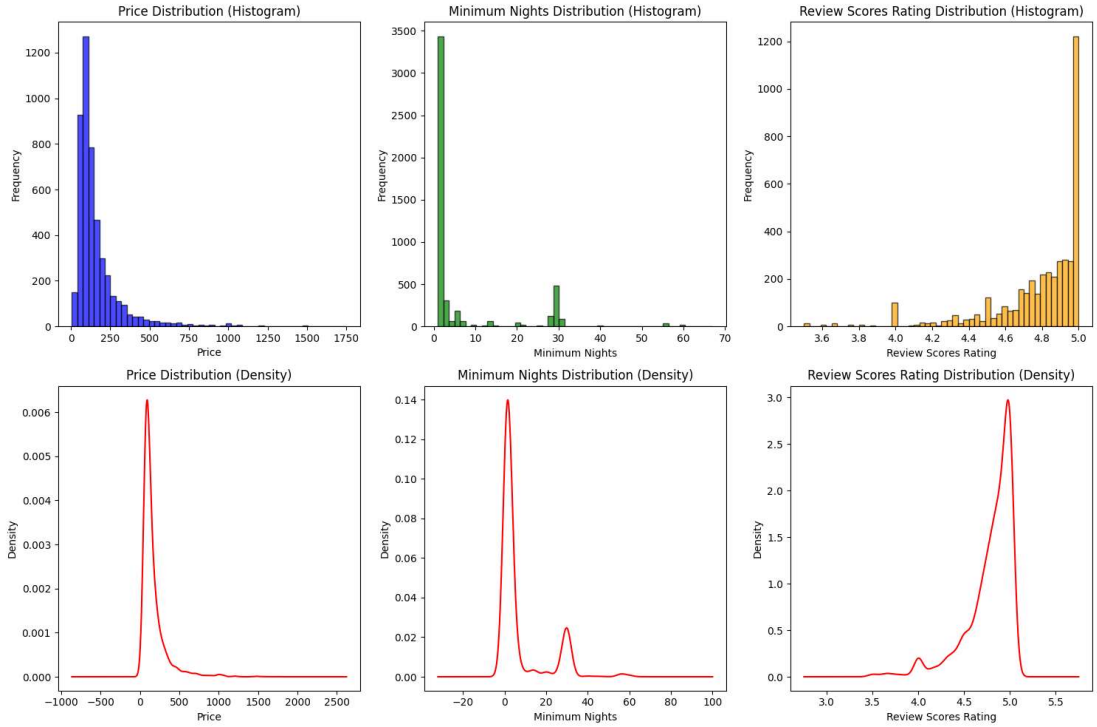
Outlier Detection using Z Score Technique:



Visualizing after outlier removal using IQR:

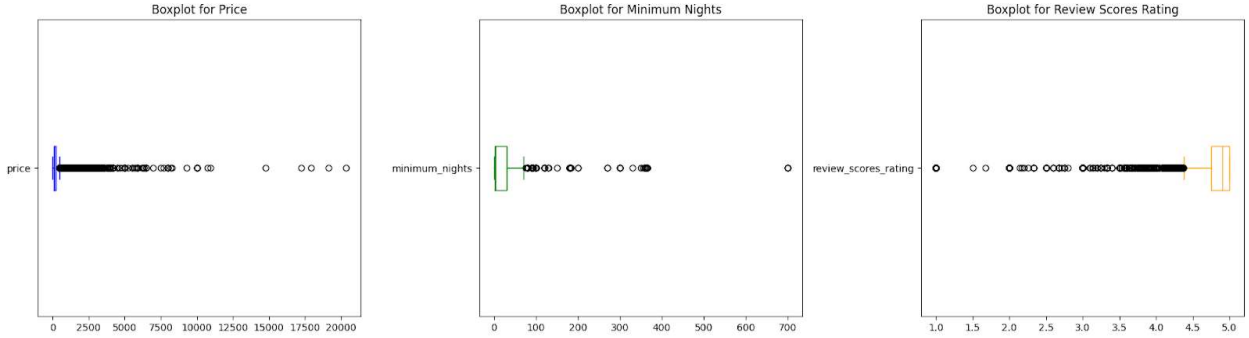


Visualizing after outlier removal using Z Score:

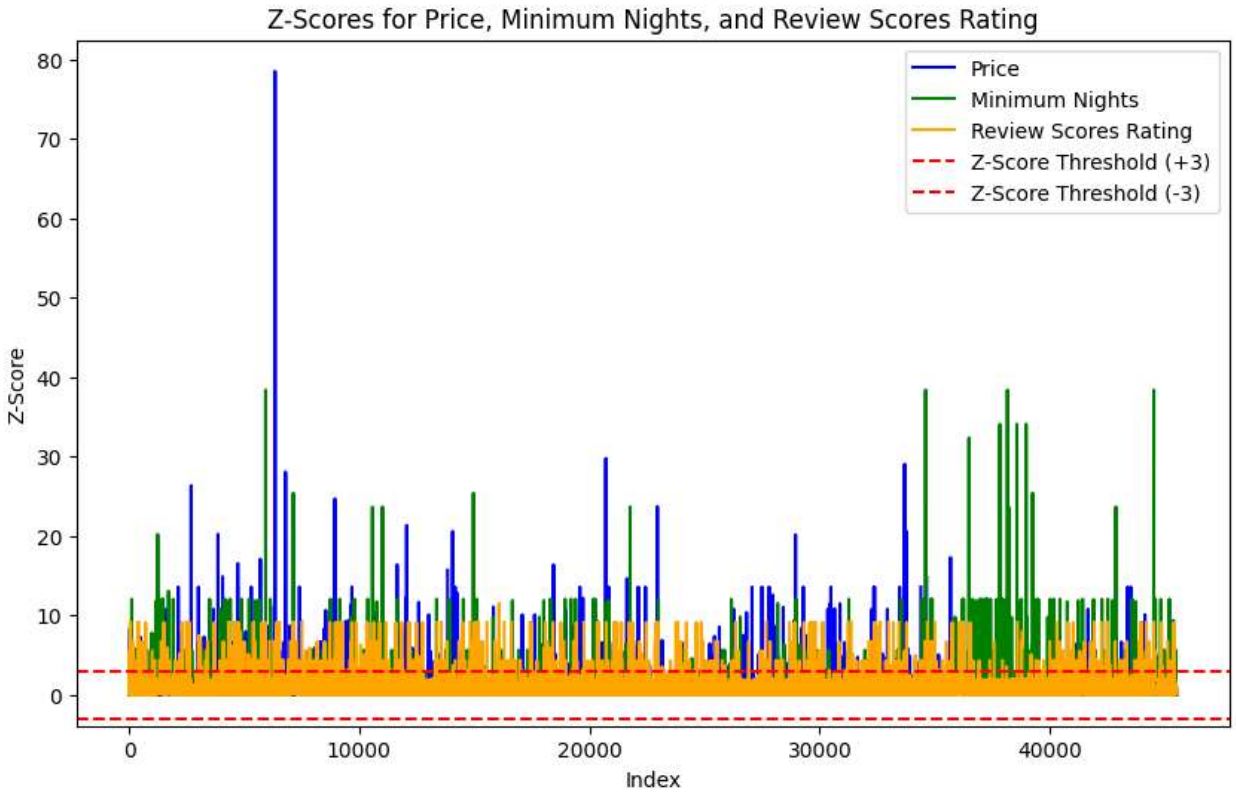


Los Angeles:

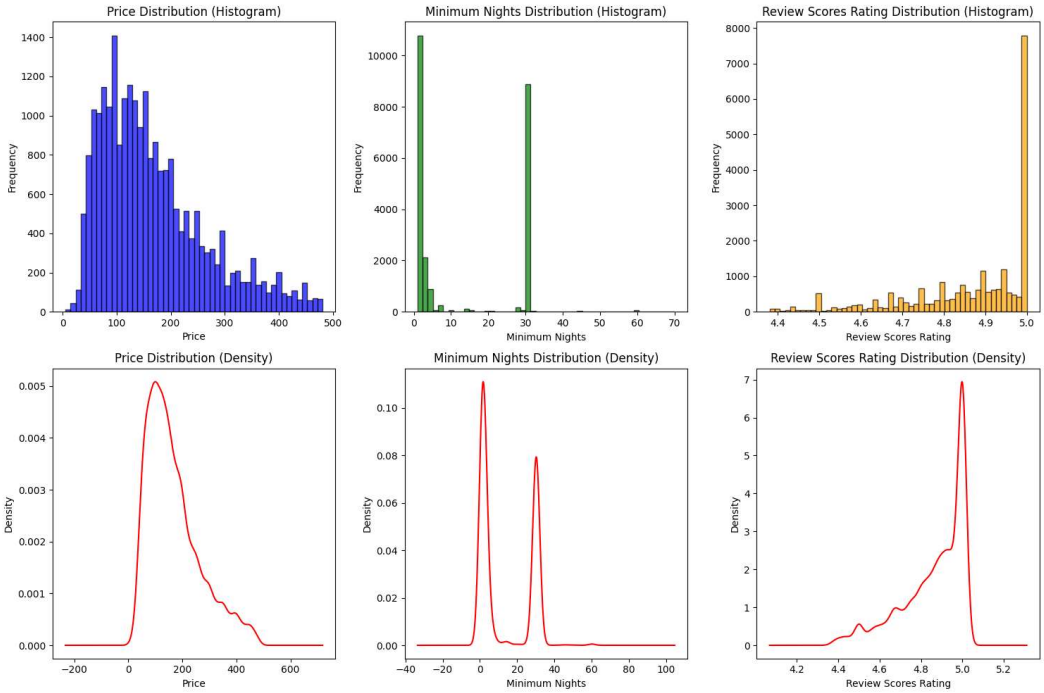
Detecting outliers using IQR:



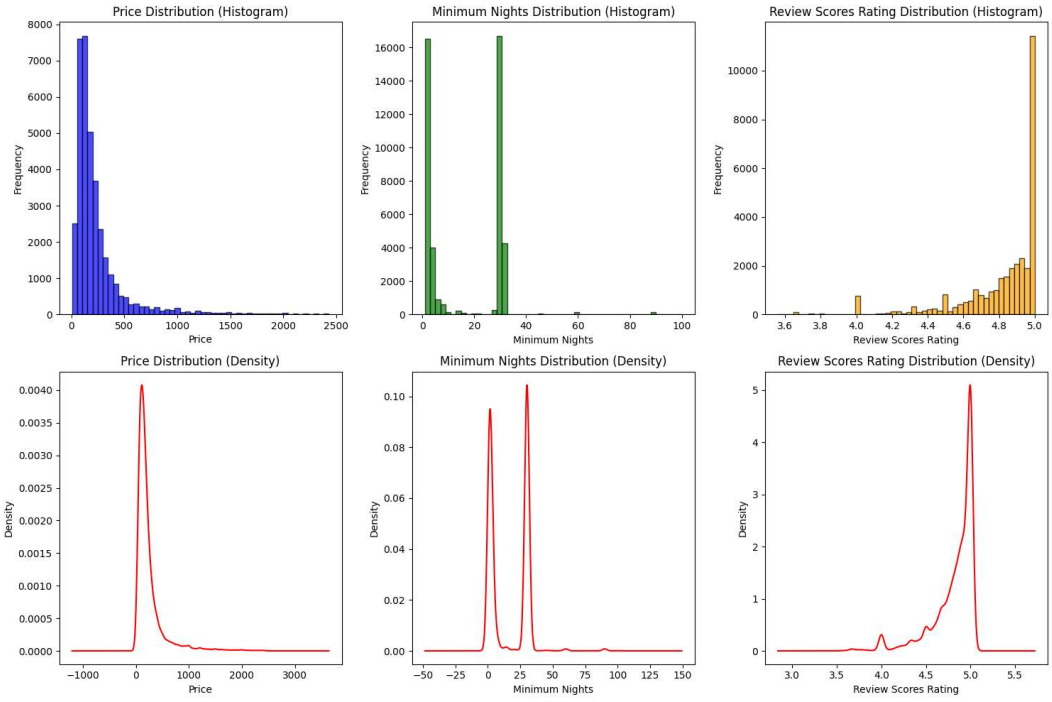
Outlier Detection using Z Score Technique:



Visualizing after outlier removal using IQR:

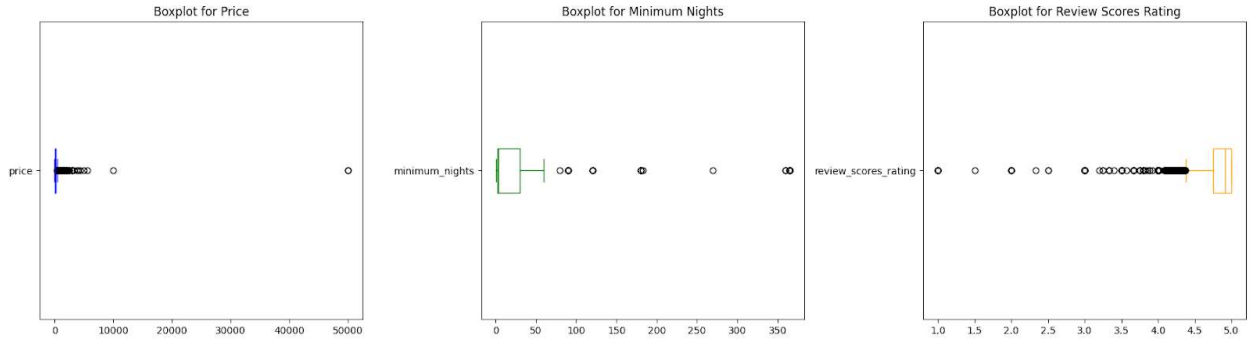


Visualizing after outlier removal using Z Score:

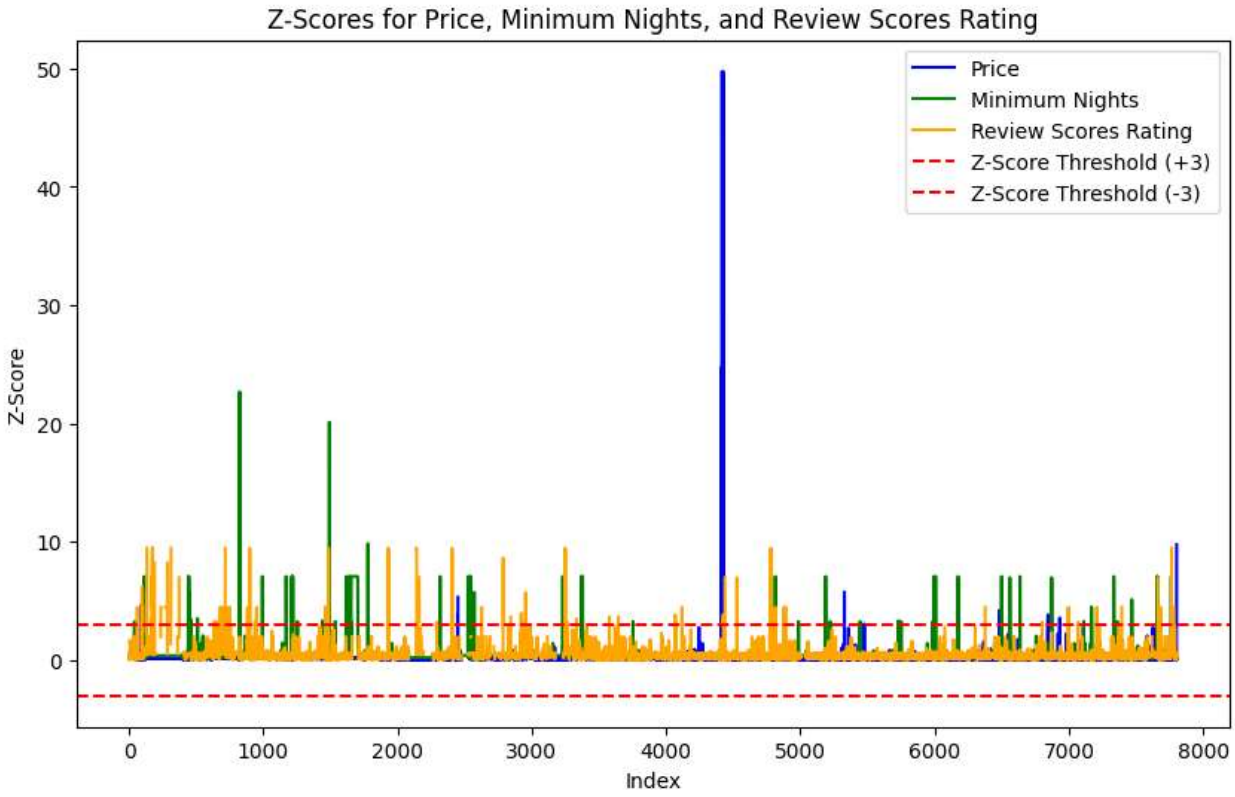


San Francisco:

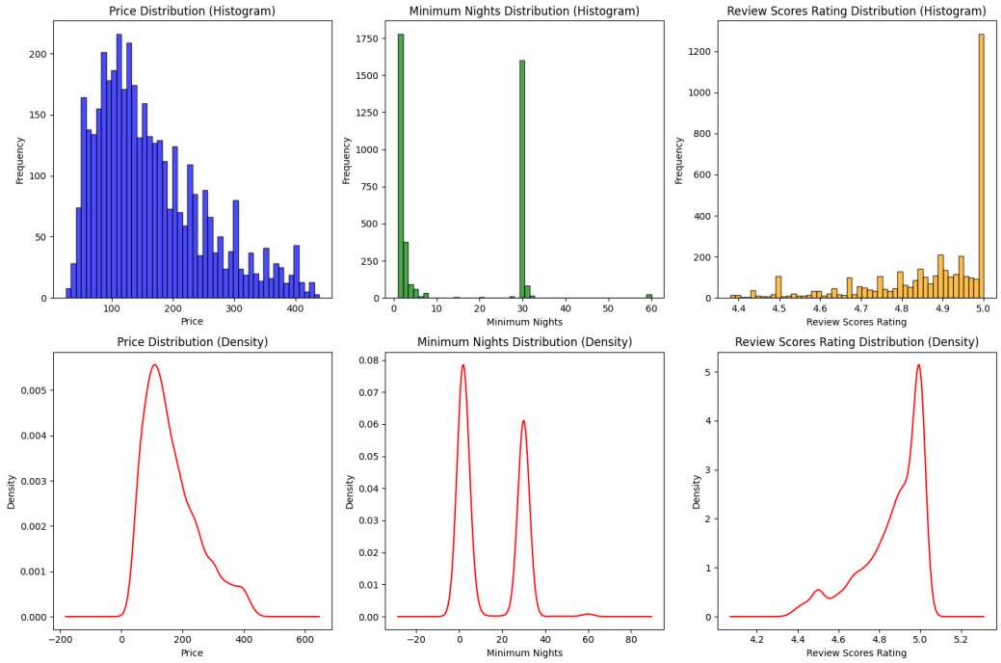
Detecting outliers using IQR:



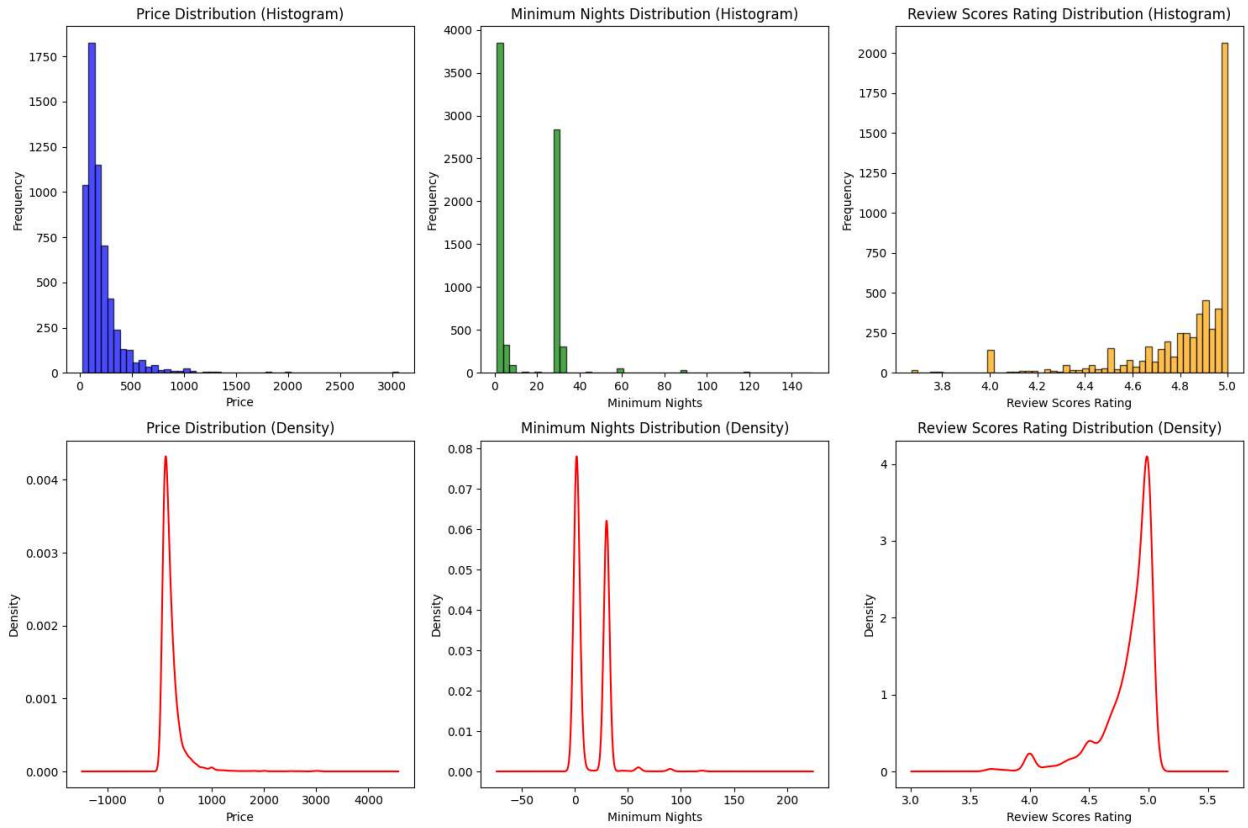
Outlier Detection using Z Score Technique:



Visualizing after outlier removal using IQR:



Visualizing after outlier removal using Z Score:



Conclusion:

IQR Method:

- **Boxplot for Price:** most of the data concentrated on the lower end but with some extremely high-priced outliers.
- **Histogram for Price:** Most listings are priced in the affordable to mid-range category, but there are a few luxury listings with significantly higher prices, which are reflected in the right tail of the distribution.
- **Boxplot for Minimum Nights:** There is a cluster of outliers requiring 100 nights or more, which suggests that while most listings cater to short-term stays, some listings are strictly for long-term rentals.
- **Histogram for Minimum Nights:** The dataset contains a mix of short-term and long-term rentals, with most listings requiring either 1 night or 30 nights.
- **Boxplot for Review Scores Rating:** While most properties receive high review scores, outliers with significantly low review scores suggest that a few listings have received poor feedback, likely due to negative guest experiences.
- **Histogram for Review Scores Rating:** Review scores are overwhelmingly positive, with most properties receiving high ratings. There are very few poorly rated properties in the dataset.

The Z-Score Plot:

- **Price (Blue Line): There are many price outliers, most of which are high-end luxury properties priced much higher than most listings.**
- **Histogram for Price:** Most listings are in the affordable price range, but some high-end or luxury properties exist, creating a skewed distribution with outliers on the high end.
- **Minimum Nights (Green Line): Many listings have high minimum night requirements, indicating long-term stay accommodations.**
- **Histogram for Minimum Nights:** The dataset contains both short-term and long-term rental options, with a sizable portion of listings allowing 1-night stays, while others require around 30 nights, likely for monthly rentals.
- **Review Scores Rating (Orange Line):** The dataset is consistent with review scores, with most listings having average to high scores.
- **Histogram for Review Scores Rating:** The review scores indicate strong customer satisfaction, with most properties receiving high ratings.

7. Task 7: Text Length

i. Introduction

This task aims to determine if there exists a positive or a negative correlation between the length of a review for a particular Airbnb listing with the review score. If a positive correlation (close to 1) is identified it indicates that if the length of the review is less the review score tends to increase, whereas a negative correlation would indicate as the length of review increases the review score is less likely to be good.

ii. Data Overview

We had to use both the listings and the reviews dataset for this problem as the two main columns to run our correlation lies in both these files. Additionally, we had to pre-process the data to get rid of empty values, emojis, special characters and HTML tags. Lastly, we also had to create a new data frame before running our correlation analysis as we noticed that each listing would have multiple reviews associated with it, so we need to calculate the average length of each comment for that listing before running our analysis.

iii. Approach

Step 1: For both the reviews and listing data after reading the file, firstly we created a duplicate column in listings dataset called listing_id since the reviews dataset and listing dataset were not consistent with the column name for listing_id, this duplicate column will help us later when we merge with the review's dataset.

Step 2: Secondly, we filled the empty comments data in reviews dataset with empty strings, since specifically for the comments data it had a lot of special characters and HTML tags, so we wrote a script using Python to clean the comments data by getting rid of all the special characters, emojis and HTML tags.

Step 3: Post getting rid of the special characters in comments column we count the length of each review, this gives us two new columns review_length_words and review_length_chars. Next, we run a sentiment analysis on the comments using TextBlob to assign each comment with a number ranging from -1 to 1. Now since the comments are grouped by listing_id, the mean review length is calculated for each listing. The results in a new dataframe avg_review_length which has two columns listing_id & avg_review_length_words. Any rows with missing values in avg_review_length_words or review_scores_rating are dropped to ensure valid data for analysis.

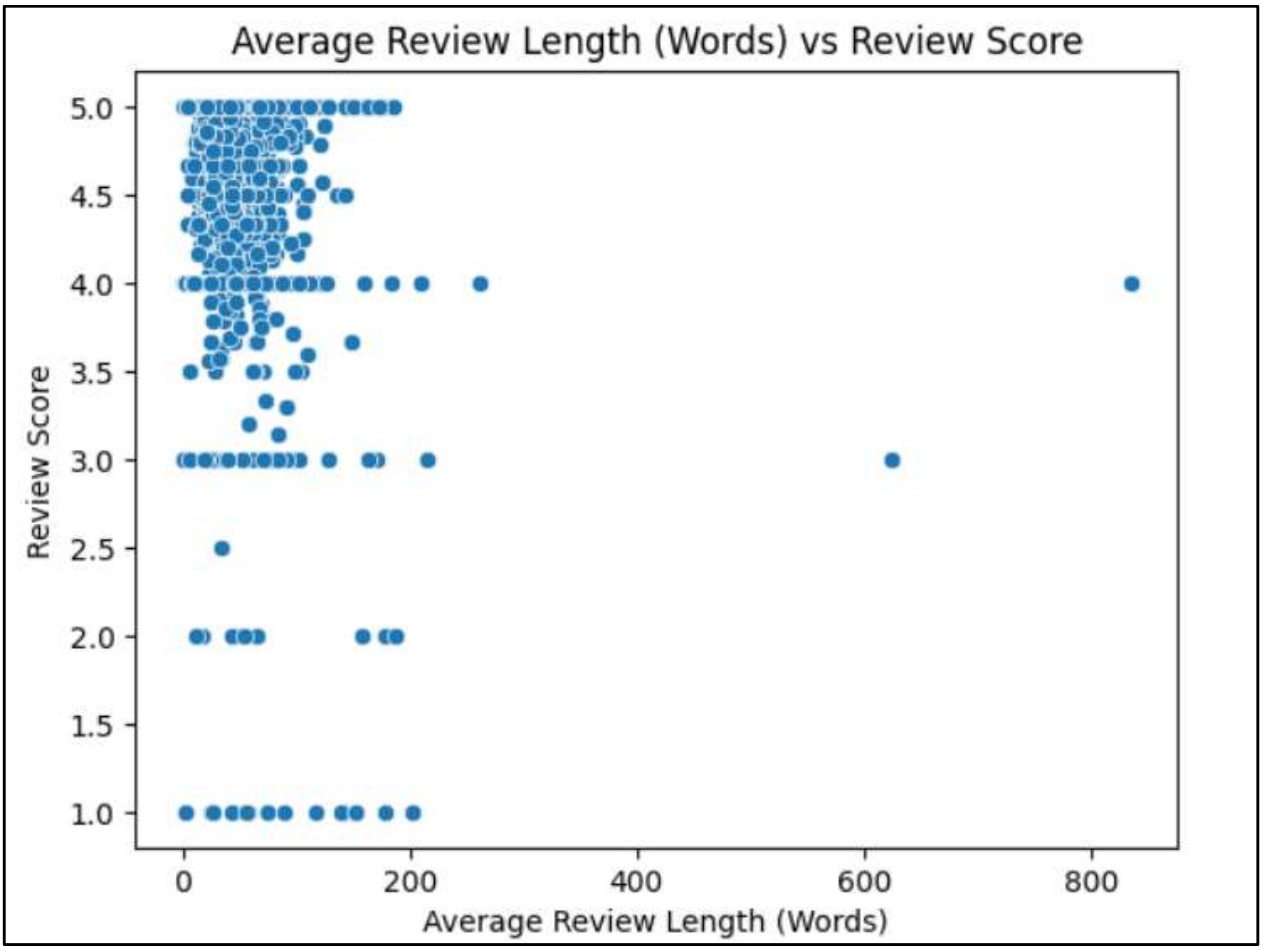
Step 4: We now move towards calculating the correlation between review length and review score using two columns namely 'avg_review_length_words', 'review_scores_rating', using the correlation function we calculate the correlation of these

two columns, the correlation value lies between -1 to 1. In our problem statement a positive correlation would indicate that as the review length increases (on average), the review score tends to increase slightly and a negative correlation would indicate as the review length increases, the review score tends to decrease.

Step 5: Finally, we visualize this correlation using a scatter plot where we pass the columns avg_review_length_words and review_scores_rating along with the merged_data variable, we use the X-axis as the avg_review_length_words and the Y-axis as the review_scores_rating.

iv. Observations

a) **Boston:**

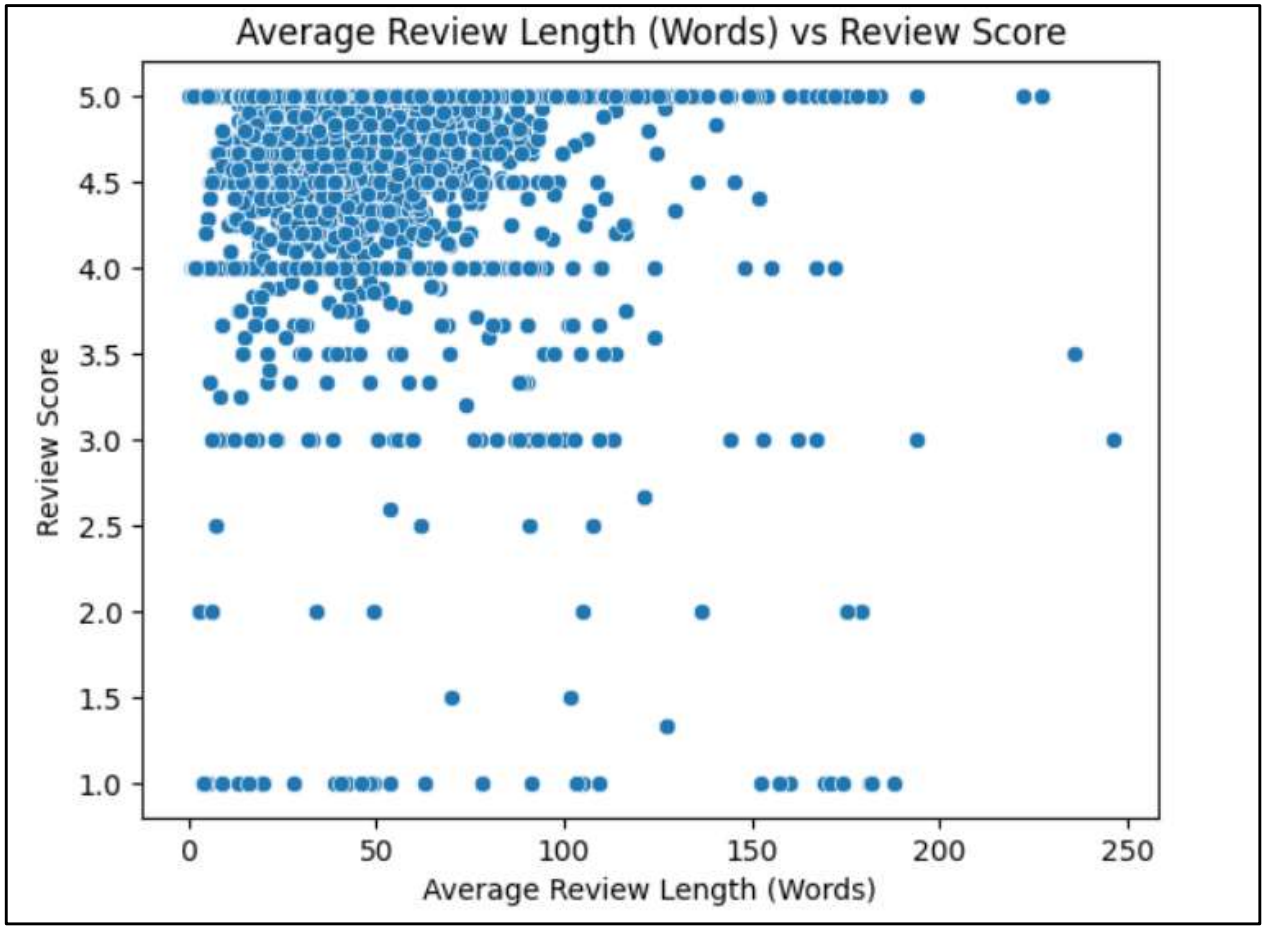


- High review scores are associated with shorter reviews (under 100 words), giving us a trend that satisfied customers often leave concise feedback.
- There is a significant concentration of reviews with a 5.0 score, all of which are short, with few going beyond 100 words.
- Reviews with lower scores (1.0 to 3.0) are more scattered, with some appearing at longer review lengths (200+ words) while others appearing at shorter review lengths as well. There are some outliers which have longer reviews with lower scores (over 600 words), but these are exceedingly rare.
- Outliers with exceptionally long reviews (500+ words) are few, and even though they have lower scores, they are not common enough or that many in numbers to form any clear pattern.

v. Conclusion:

There are not enough outliers or plot points to deny that the plot supports the conclusion that shorter review length tends to indicate that the review score is high (4.0 to 5.0) range thus indicating a positive correlation which indicate that shorter reviews are in fact associated with higher review score.

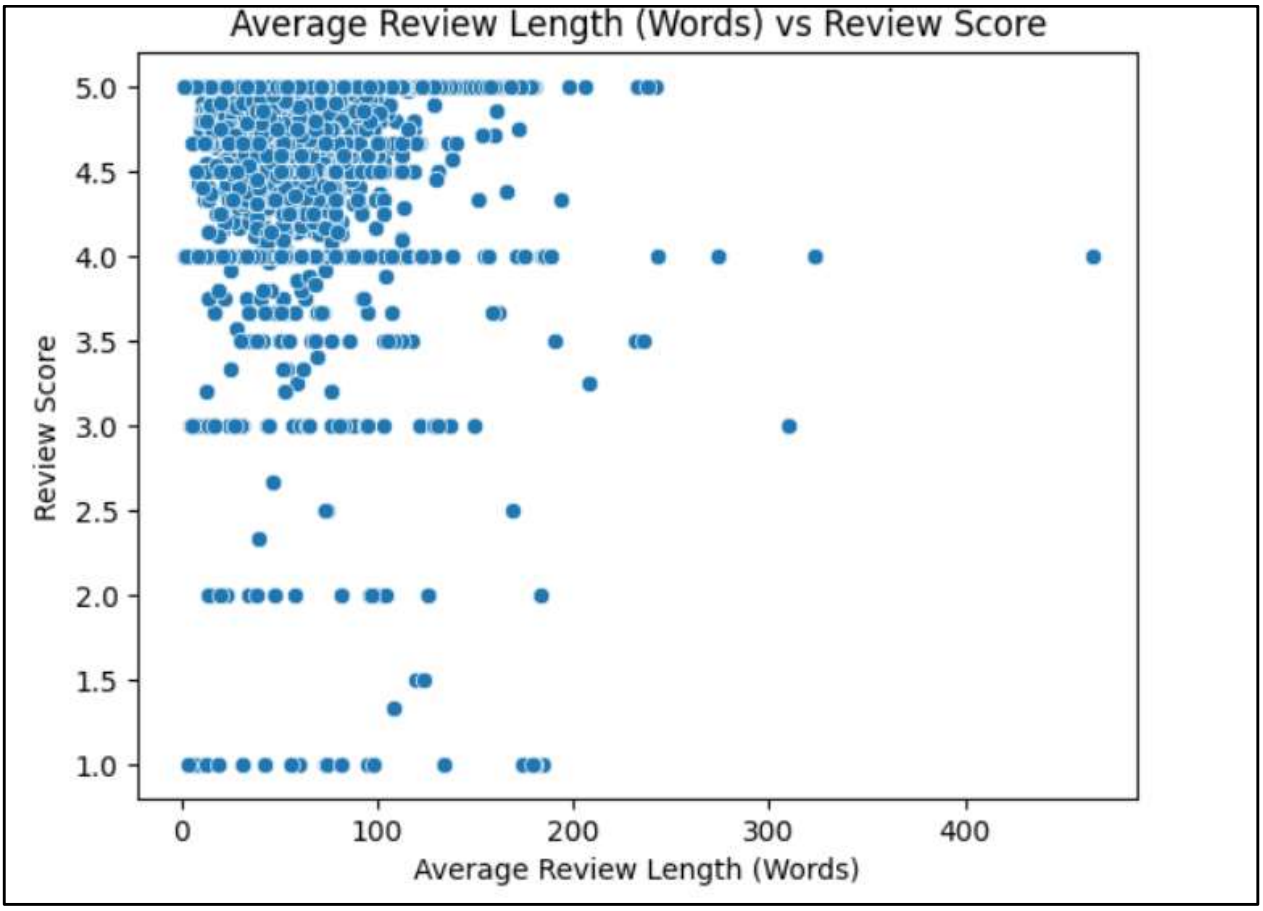
b) Chicago



- High review scores are once again associated with shorter reviews (under 100 words), repeating the trend that satisfied customers prefer to leave concise feedback.
- We see a clear cluster of high reviews with compact review lengths, thus supporting the conclusion of a positive correlation.
- There are a lot of outliers for different scenarios be it high reviews with longer review lengths or low reviews with shorter review lengths

Conclusion: Even with the distinct types of outliers we see in our graph, the density of the plot near the high reviews and short reviews area of the graph we can conclude that the correlation is in fact a positive one and short reviews are correlated to high reviews.

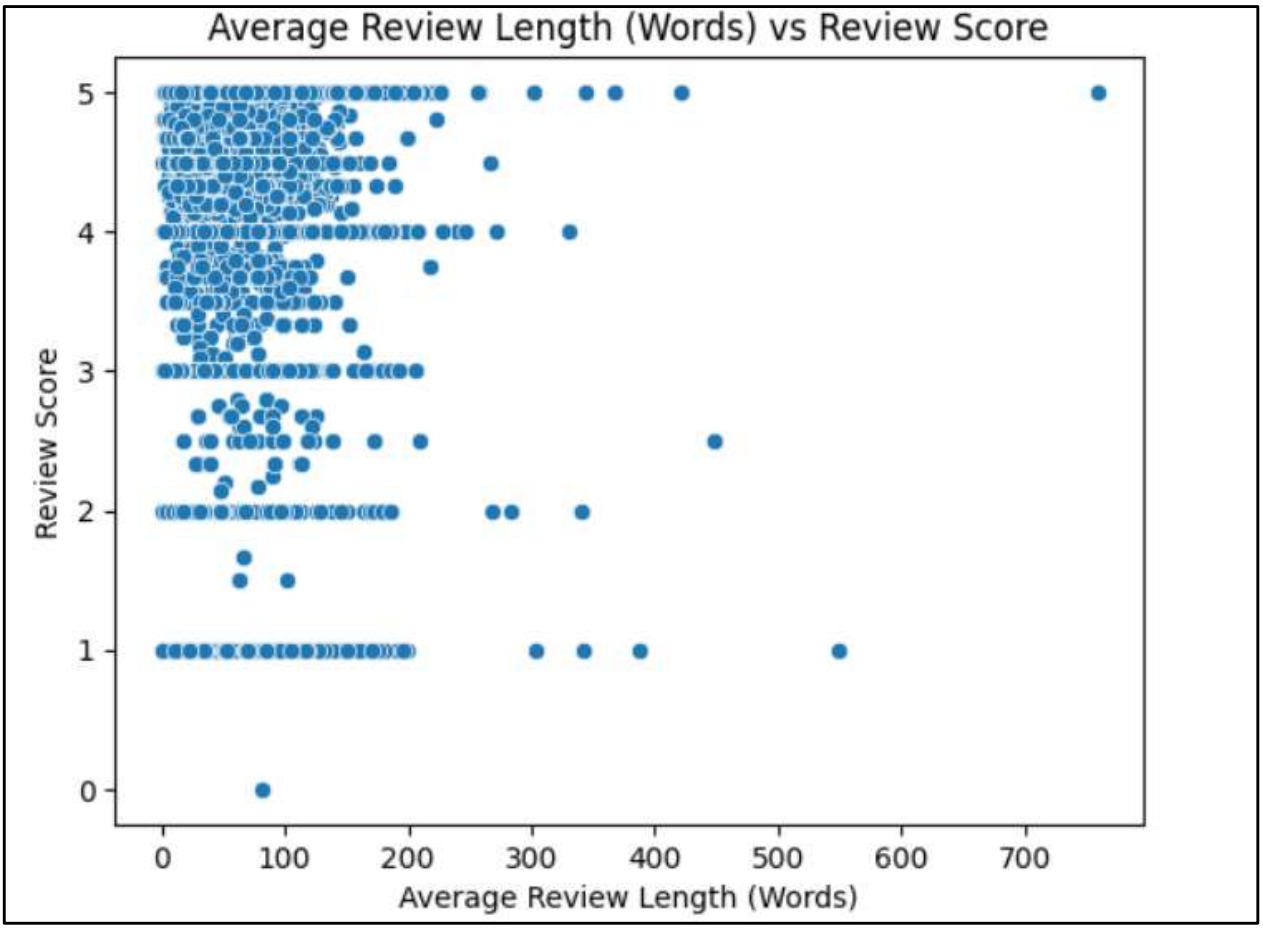
c) San Francisco



1. High reviews are once again associated with shorter reviews (under 100 words), repeating the previous trends.
2. There is a significant concentration of reviews with a score of 5.0 scores, all of which are short (around 100 words) with a few going beyond 100 words.
3. Reviews with low scores are more scattered with a good portion of them being under 100 words where there are some outliers which have low scores and longer reviews

Conclusion: The graph supports the previous conclusion making the correlation positive where we can say that short reviews are correlated to better review scores.

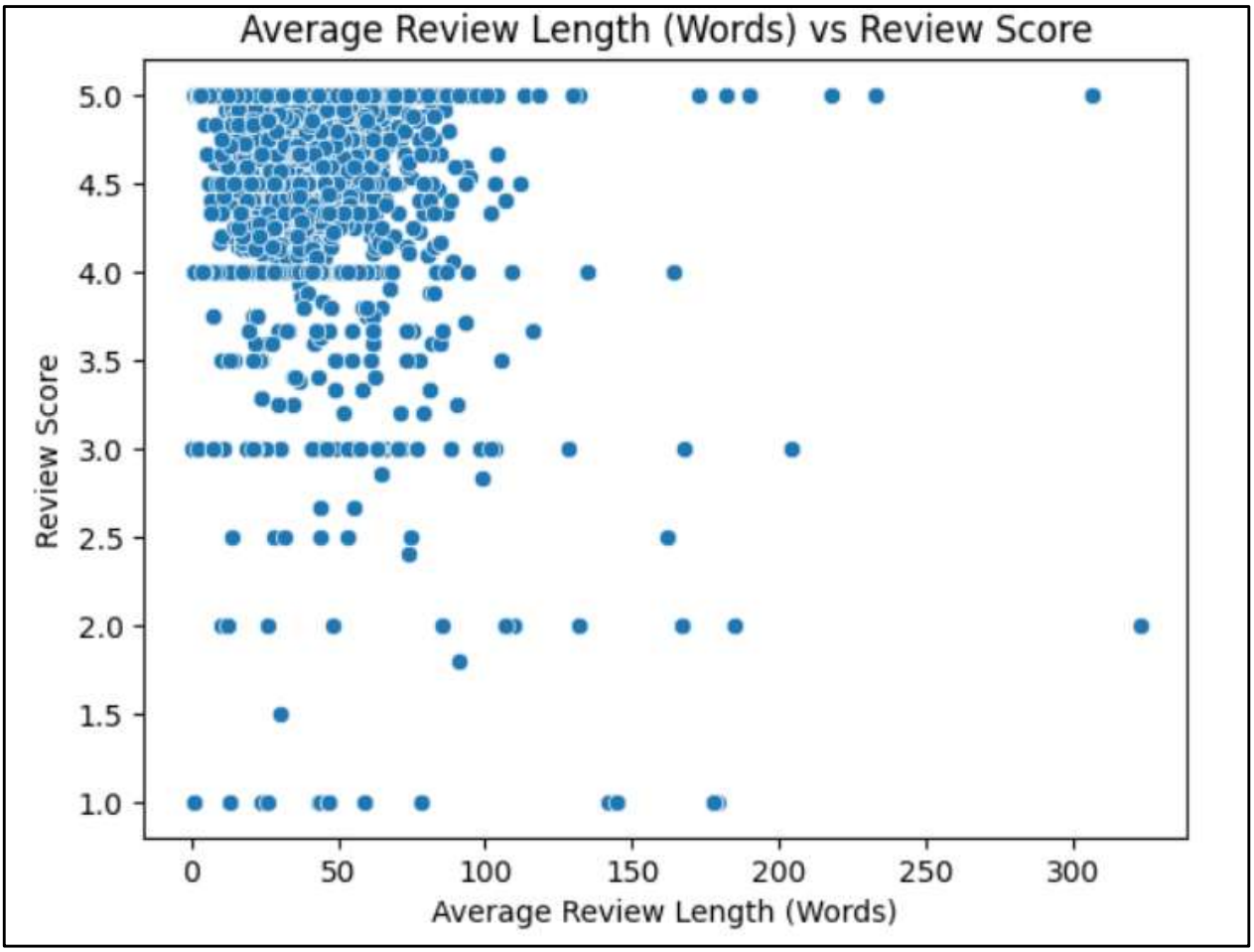
d) Los Angeles



- The scatter plots are dense in the range of 3 to 5 with shorter review lengths.
- Even though we can see that the range from 4 to 5 review scores have short reviews, we cannot ignore the plots in the range of 3 to 4
- Even though 3 to 4 review score is good it is not considered a high score, and, in this case, it also has shorter review lengths

Conclusion: There are far too many outliers with short reviews and average ratings that we cannot call this correlation to be a positive one.

e) Dallas



- Like the previous 3 cities we can see that the plots are clustered around the range of 4-5 review score with shorter reviews
- There are outliers with longer reviews, high review scores, shorter reviews but low review scores.

Conclusion: Just looking at the scatter plot we can conclude that the correlation is in fact a positive one as there are not enough outliers to prove the correlation either neutral or negative.

8. Task 8: Keyword Extraction

i. Introduction

The task aims to identify and count the occurrence of specific keywords that may influence guest satisfaction. Some common keywords associated with guest experiences include "clean," "comfortable," and "noisy," which can provide insights into listing quality. By generating new features based on these keywords, we can enhance model predictions related to guest satisfaction and offer recommendations for improving vacation rental services like Airbnb.

ii. Data Overview

The 'reviews' dataset contains comments provided by guests for various listings. These comments reflect their experiences of the properties and any feedback or suggestions, which could be utilized to identify trends associated with properties or listings. Since we are currently looking at only five cities, we used the data from Inside Airbnb for these cities (for reference, check the Appendix).

iii. Keyword Identification and Selection

To identify specific keywords, we selected a range of reviews for listings with ratings from 1 to 5. We chose five listings for each rating and identified the commonly used keywords in these reviews. We categorized the keywords into two sets: positive keywords expressing positive sentiments or experiences of guests, and negative keywords expressing negative sentiments. The keywords are:

- positive_keywords = ["great", "clean", "nice", "comfortable", "friendly", "spacious", "amazing", "fabulous"]
- negative_keywords = ["uncomfortable", "noisy", "bad", "dirty", "terrible", "unpleasant", "poor"]

These keywords are quite relevant to what a general person might experience during a stay at a property. There are libraries like 'WordCloud' that provide visual representations of text data, where the size of each word indicates its frequency or importance within the text. These libraries could be utilized to identify keywords based on the comments data. A sample figure of 'WordCloud' we generated using comments from the Boston reviews file is presented below.

iv. Approach

The approach consisted of the following steps:

- **Step 1: Keyword Counting**

We created a function to count the occurrence of each keyword in a particular review comment. The comments were converted to lowercase for uniformity, and the count of each keyword was saved.

- **Step 2: Dataframe Creation**

We generated two separate dataframes for positive and negative keyword counts, with columns prefixed with 'positive_' and 'negative_' to distinguish between positive and negative keywords.

- **Step 3: Data Aggregation**

The keyword counts for each review were then aggregated by 'listing_id' to generate an overall count for each keyword for each listing.

- **Step 4: Data Integration**

We then used the aggregated data and merged it with the 'listing' dataset using the 'listing_id' to generate additional features, such as the total number of positive and negative keyword occurrences for each listing.

- **Step 5: Handling missing data**

If a listing had no reviews or matching keywords, the missing data was filled with zero to ensure the dataset remained complete and could be utilized for further analysis.

v. Feature Generation

The new features generated are:

- **Positive keyword counts:** Number of occurrences for each positive keyword, e.g., positive_clean, positive_comfortable.
- **Negative keyword counts:** Number of occurrences for each negative keyword, e.g., negative_noisy, negative_dirty.

These new features are now part of the listing dataset, providing richer information about guest sentiment based on reviews. These features can help further optimize property listings. For example, these keywords can be used to filter listings based on certain parameters while a guest is searching. A guest looking for a clean and comfortable stay should receive suggestions or recommendations for listings with high positive keyword counts for "clean" and "comfortable." These keyword counts are directly proportional to the number of guests satisfied with the listing for that particular feature. This filtering mechanism can help users find listings that meet their preferences, enhancing their overall experience. However, we do have some limitations, which can be found in the 'Limitations' section under Task 8.

The results were saved in a CSV file, with each listing having the overall count of these specific keywords. Below are a few lines displaying the output for each city.

- Boston -

	A	B	C	D	E	L	M	N
1	id	name	review_scores_rating	positive_great	positive_clean	negative_uncomfortable	negative_noisy	negative_bad
2	4090224	Sanitized Modern 3BD private home-near BCEC	4.91	626	371	0	0	6
3	891661	GORGEOUS WATER FRT NORTH END STUDIO	4.7	476	180	1	9	6
4	13393418	*Airport/Dwtn/TD Garden/NorthEnd/Convention Ctr/T*	4.75	430	277	1	8	15
5	4568116	AKBrownstone: cozy private studio by T	4.9	428	345	2	3	4
6	11202183	Second Oldest Home in Beacon Hill	4.85	406	204	0	4	3
7	36237250	Staypineapple Boston, Back Bay King	4.67	373	219	3	2	10
8	18290558	â¯... Spacious & Modern â¯... Professionally Sanitized â¯...	4.85	358	443	4	6	3
9	257588	BOSTON LUXURY FOR LESS!! NEAR BCEC	4.85	351	170	2	2	4
10	12269155	Boston Apartment (North End)	4.85	349	132	0	6	3
11	197972	Serene Studio in a Perfect Location	4.88	346	148	0	0	2

- Chicago -

id	name	review_scores_rating	positive_great	positive_clean	negative_uncomfortable	negative_noisy	negative_bad
29819757	Hotel Perks - Private Bedroom | Private Bathroom	4.47	2039	664	25	23	54
29093384	Traveler's Dream - 1 bed in a shared bedroom	4.51	748	302	4	4	18
44126335	Kasa | 2BD, Walk to Lincoln Park Zoo | Chicago	4.83	726	291	3	3	4
38090971	Godfrey Hotel Lifestyle Rooftop 4.5*-King Deluxe	4.54	679	256	3	9	10
726376	Spacious Studio, Amazing Location!	4.7	669	267	1	12	18
350347	Urban Chicago Loft, 1 blk to train w/ 2 pkg spots	4.9	648	185	1	3	10
1171860	Lincoln Park Studio, Great Value!	4.76	603	261	4	1	4
44126327	Kasa | 1BD, Spectacular City Views | Chicago	4.7	581	296	3	8	6
10069247	Artist Loft, Private Room	4.9	548	77	4	7	10
464581	Large, Private Logan Square Studio	4.81	511	180	0	7	4

- Dallas -

id	name	review_scores_rating	positive_great	positive_clean	negative_uncomfortable	negative_noisy	negative_bad
16353509	Downtown/Deep Ellum Great Location Very Private	4.7	721	356	7	16	4
39844580	Splendorous Suite with Beautiful Luxury Shower	4.75	454	209	6	0	6
33198135	Private Guesthouse Close to Downtown, Deep Ellum.	4.84	441	283	3	8	13
1826550	Private Room in Canton Townhouse	4.94	435	219	0	4	5
24596033	Vintage Airstream Near Deep Ellum & Fair Park	4.95	428	121	3	0	1
15342315	Charming Cabin Near Deep Ellum & Fair Park	4.97	427	132	0	1	4
23681001	Artsy Eclectic Dallas Getaway	4.89	384	122	1	5	4
34000242	Your Own Private Tiny House In the Heart of Dallas	4.94	373	196	1	0	9
52126733	Sonder at Commerce | One-Bedroom Apartment	4.48	361	176	4	1	13
20300902	1920â€™s City Retreat with Dallas Arboretum Tickets	4.74	358	143	4	4	4

- Los Angeles -

id	name	review_scores_rating	positive_great	positive_clean	negative_uncomfortable	negative_noisy	negative_bad
42409434	The Burlington Hotel	4.67	1172	1275	8	39	68
1990543	Historic Bungalow	4.63	763	219	9	3	16
6527658	Private Loft In the Hollywood Hills (by Universal)	4.86	738	323	1	6	7
2034041	Venice Original Private Guest House	4.89	685	284	1	3	4
3259107	Super Venice Location!!	4.8	674	187	1	1	8
6438015	Venice Beach Guest Studio with Pool and Hot Tub	4.96	625	126	1	0	10
22215734	Guest Suite, Private Ent., 5 min LAX, Westchester	4.93	584	575	2	3	7
18278425	One bedroom Guest Home- Private Entrance, Near LAX	4.92	556	499	0	17	10
10454154	Private Bungalow, 3 Blks from Beach	4.56	539	174	4	1	14
21234077	5 min to LAX, Private Entrance, Studio Apartment	4.96	528	427	1	9	2

- San Francisco -

id	name	review_scores_rating	positive_great	positive_clean	negative_uncomfortable	negative_noisy	negative_bad
545685	Garden Suite by Golden Gate Park, Private Bathrm	4.82	647	558	10	9	10
6092596	Mission Dolores Suite	4.9	632	324	0	5	3
35642179	Grant Plaza Hotel, Standard Double	4.27	577	274	31	36	23
49634091	Quiet Marina District Garden Oasis W/Free Parking	4.72	564	228	2	12	12
14804950	Cozy, Comfortable, Private Studio in Bernal Hts	4.89	558	279	2	3	9
208831	Suite in Heart of North Beach. No Cleaning Fees.	4.81	555	130	1	1	1
585326	Cozy Suite by Golden Gate Park, Private Bathrm	4.81	538	470	1	7	12
1738415	Tranquil Updated Studio in Historic District	4.91	536	209	1	1	2
4464347	One Bed in Shared Dormitory at Social SF Hostel #1	4.76	535	207	3	13	11
6163821	Studio Potrero Hill-close to Chase and UCSF	4.9	518	234	2	1	1

vi. Conclusion

The keyword extraction task successfully identified specific keywords and their frequencies. The resulting features provided valuable insights into guest satisfaction and listings' optimization. Future improvements could include expanding the keyword set and employing more sophisticated techniques to capture context and sentiment more accurately.

Limitations

Task 3: Correlation Analysis

Handling of NaN values: We dropped the NaN values from the dataset assuming there is no significant difference in the data analysis.

Task 4: Price Analysis

1- Data Skewness: Pricey homes can totally mess up the average prices. The median's probably better for seeing what homes in a neighborhood actually go for.

2- Missing Key Variables: The analysis skips over stuff like the season, the quality of the property, and amenities—things that all affect prices. It also does not look at occupancy rates.

Task 5: Neighborhood Comparison

The limitations are as follows:

- **Missing Data:** Some neighborhoods do not have review scores listed and hence were excluded from the analysis, which might result in an incomplete comparison.
- **Unequal Distribution:** Some neighborhoods could have many listings while others may have fewer. The average rating thus may not accurately reflect guest experiences in neighborhoods with fewer listings.
- **Outliers:** Since we do not capture extreme values, a few listings with very extreme ratings could influence the overall average of a neighborhood.

Task 6: Outlier Detection

1- Sensitivity of Methods: IQR and Z-score methods could flag legit cases, like luxury homes or long-term rentals, as outliers, especially if the data's a bit skewed.

2- Over-simplification of Stay Durations: Places designed for longer stays could get wrongly flagged as outliers because of their minimum night stays.

3- Static Data: The analysis does not consider price changes over time, meaning the outliers it detects might only be temporary.

Task 7: Text Length

Blank Reviews: Some listing_id values did not have associated reviews. In the code, we handled this by excluding those rows, assuming their absence would not significantly impact the overall data analysis. Similarly for some listing_id there were no review score associated so we excluded those rows as well.

Task 8: Keyword Extraction

We have identified the following limitations while analyzing the keyword extraction to generate new features. The limitations are:

- **Keyword Limitations:** The selected keywords, although common, may not fully capture the entire scope of guest experiences. Additional keywords or specific vocabulary could enhance the analysis.

- **Keyword Context:** The approach counts keyword occurrences without considering the context. For example, the word “clean” may appear in both positive and negative scenarios (e.g., not clean). More advanced techniques, like NLP, could be used to address this.
- **Review Availability:** The availability of few or no reviews may not provide meaningful data for analysis, which can lead to incomplete insights for listings.

Conclusion

Descriptive statistics revealed variations in pricing and review scores across neighborhoods, with outliers suggesting areas for pricing adjustments. Correlation analysis showed strong relationships between review scores and factors like the number of reviews and availability, indicating that higher visibility may enhance guest experiences. Distribution analysis highlighted skewness in certain features, necessitating careful pricing strategies. Feature engineering, including text length measurement and keyword extraction, identified key attributes such as "clean" and "comfortable" that significantly influence guest satisfaction. These insights empower stakeholders to improve service offerings and tailor marketing strategies, enhancing Airbnb's competitive edge in the vacation rental market.

Appendix

1. Listings and Reviews data from Airbnb: https://insideairbnb.com/get-the-data/
2. City files -
    a. Boston:
        i. Listings: https://data.insideairbnb.com/united-states/ma/boston/2024-06-22/data/listings.csv.gz
        ii. Reviews: https://data.insideairbnb.com/united-states/ma/boston/2024-06-22/data/reviews.csv.gz
    b. Chicago
        i. Listings: https://data.insideairbnb.com/united-states/il/chicago/2024-06-21/data/listings.csv.gz
        ii. Reviews: https://data.insideairbnb.com/united-states/il/chicago/2024-06-21/data/reviews.csv.gz
    c. Dallas
        i. Listings: https://data.insideairbnb.com/united-states/tx/dallas/2024-08-17/data/listings.csv.gz

    ii. Reviews: https://data.insideairbnb.com/united-states/tx/dallas/2024-08-17/data/reviews.csv.gz
d. Los Angeles
    i. Listings: https://data.insideairbnb.com/united-states/ca/los-angeles/2024-09-04/data/listings.csv.gz
    ii. Reviews: https://data.insideairbnb.com/united-states/ca/los-angeles/2024-09-04/data/reviews.csv.gz
e. San Francisco
    i. Listings: https://data.insideairbnb.com/united-states/ca/san-francisco/2024-09-04/data/listings.csv.gz
    ii. Reviews: https://data.insideairbnb.com/united-states/ca/san-francisco/2024-09-04/data/reviews.csv.gz