SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طويق
TUWAIQ ACADEMY

COURSE END PROJECT:

Data Analysis Module

Exploratory Data Analysis

Prepared By:

Sara Alharbi

In this exploratory data analysis, I investigated a dataset containing information about [Airbnb Listings & Reviews dataset]. My analysis aimed to uncover insights, patterns, and trends within the data to inform decision-making and identify potential areas for further investigation.

Introduction and Dataset Overview:

The Airbnb Listings & Reviews Dataset offers a wealth of information about Airbnb listings, encompassing property details, host information, pricing, availability, and guest reviews. Also the dataset covers listings from major cities worldwide, including popular destinations such as New York City, Paris, and Sydney. This geographic diversity allows for analysis of accommodation trends across different regions and markets. Moreover it include property details such as property type (e.g., apartment, house, villa), room type (e.g., entire home, private room, shared room), amenities (e.g., Wi-Fi, parking), and pricing information (e.g., nightly rates, minimum stay requirements). Also guest reviews provide feedback on aspects such as cleanliness, location, and overall satisfaction, offering valuable insights into guest experiences.

The dataset comprises 5373143 row and 36 column after merged two dataset list and review. The data includes a variety of numerical columns (eg. minimum_nights , price, bedrooms accommodates and more) and categorical columns (eg. room_type, property_type, instant_bookable and more).

In preparation for analysis, thorough data cleaning was conducted to ensure the reliability and integrity of the dataset. Missing values were addressed through appropriate strategies such as deletion or imputation, while duplicate entries were identified and removed to prevent redundancy.

Data Cleaning:

During exploring the dataset, I noticed some null values in different columns which it can be a problem during the analysis so we can solve this during the cleaning process it involved several steps to address missing values, handle duplicate entries, and convert data types where necessary.

1. Identification of Missing Values and Handling Them:

The dataset was containing many null values, you can see some of them as shown in the below table:

Table.1 Number of null values

Name of column	Name	District	bedrooms	review_scores_rating
Number of null values	333	4525416	541413	6105

I dealt with missing values in several ways:

- I deleted the entire column ('district') because it contained 84% null values of the driven values.
- For columns with numerical data, I replaced the missing values with the average for each column (Mean).
- For columns with categorical data, I replaced missing values with the (Mode) for each column.

2. Handling Duplicates Rows (if any):

The dataset was not having any duplicates rows, so I do not have to do any preprocessing in this section.

3. Converting Categorical Variables:

In this dataset it has multiple categorical variables so we need to convert them to ‘category‘ type, I do this for these columns: ['host_is_superhost', 'host_has_profile_pic', 'host_identity_verified', 'neighbourhood', 'city', 'property_type', 'room_type', 'instant_bookable'].

4. Encoding:

I do label encoding to binary categorical columns containing 'T' and 'F' values. These columns, namely 'host_is_superhost', 'host_has_profile_pic', 'host_identity_verified', and 'instant_bookable', represent binary attributes indicating certain characteristics or features associated with Airbnb hosts and listings.

By applying label encoding to these binary categorical columns, we converted the 'T' and 'F' values into equivalent numerical representations, enabling the incorporation of these features into subsequent analysis and modeling tasks. This preprocessing step ensures compatibility with various machine learning algorithms that require numerical input data, facilitating the exploration of relationships and patterns within the dataset.

These are the main steps in data cleaning to help us in data analysis.

Exploratory Data Analysis (EDA) :

Exploratory data analysis (EDA) was undertaken to gain a deeper understanding of the dataset and uncover meaningful patterns and relationships. Univariate analysis helped explore the distribution of individual variables, while bivariate analysis delved into the relationships between variables. Visualizations such as histograms, box plots, and heatmaps were utilized to facilitate data exploration.

1. Univariate analysis:

I do a visualization for listings price distribution. After visualizing the distribution of listing prices using histograms as shown below (Figure.1), it was evident that the distribution exhibited positive skewness, also known as right skewness. This indicates that the majority of listing prices were clustered towards the lower end of the price range, with a long tail extending towards higher prices.

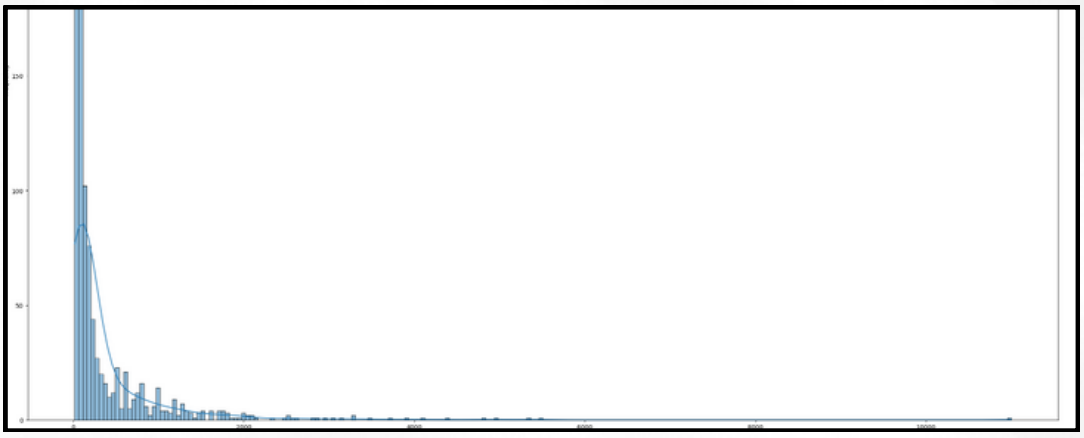

Figure.1 Listing Price Distribution

Moreover, after examining the distribution of review scores ratings through histograms as shown below (Figure.2), it became apparent that the distribution exhibited negative skewness, indicative of a left-skewed distribution. This implies that the majority of review scores ratings were concentrated towards the higher end of the rating scale, with fewer instances of lower ratings.

The presence of negative skewness in the review scores ratings distribution suggests a generally positive sentiment among guests who provided reviews for the Airbnb listings in the dataset. A left-skewed distribution indicates that a larger proportion of ratings fall within the higher range, reflecting overall satisfaction with the accommodations and experiences.

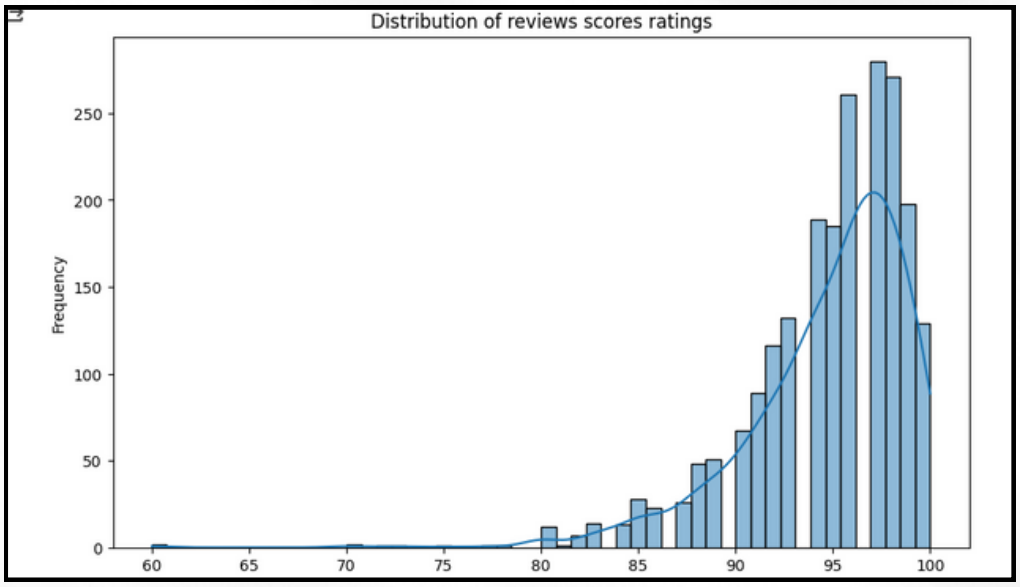

Figure.2 Review Score Rating Distribution

2. Bivariate analysis:

I do bivariate analysis between ‘property_type‘ and ‘price‘ as Average price by top 30 property types as shown below (Figure.3).

- 'Entire villa' emerged as the property type with the highest average price, exceeding 4000. This suggests that villas, often associated with luxury and exclusivity, command premium prices among accommodation options.
- The lowest average price was observed for 'shared room in apartment,' indicating that shared accommodations within apartment settings tend to be more budget-friendly options for travelers. This aligns with the notion that shared rooms provide cost-effective alternatives for budget-conscious travelers who prioritize affordability over privacy.

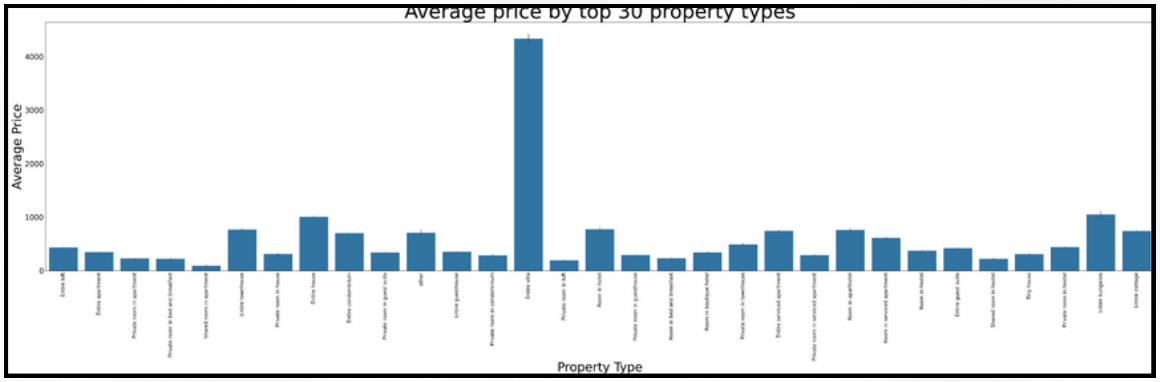

Figure.3 Average price of top 30 property types

Also I do a bivariate analysis between ‘neighbourhood‘ and ‘price‘ as average prices by the top 20 neighborhoods as shown below (Figure.4). As results:

- The neighborhood of 'Miguel Hidalgo' in Mexico emerged as the most expensive neighborhood, with an average price of approximately 1500. This suggests that properties located in Miguel Hidalgo command premium prices, likely due to factors such as upscale amenities, proximity to attractions, and desirable living environments.
- The neighborhood of 'VII San Giovanni/Cinecitta' in Rome exhibited the lowest average price among the analyzed neighborhoods. This finding indicates that properties in this neighborhood are more affordably priced compared to others in the dataset, potentially appealing to budget-conscious travelers or those seeking economical accommodation options.

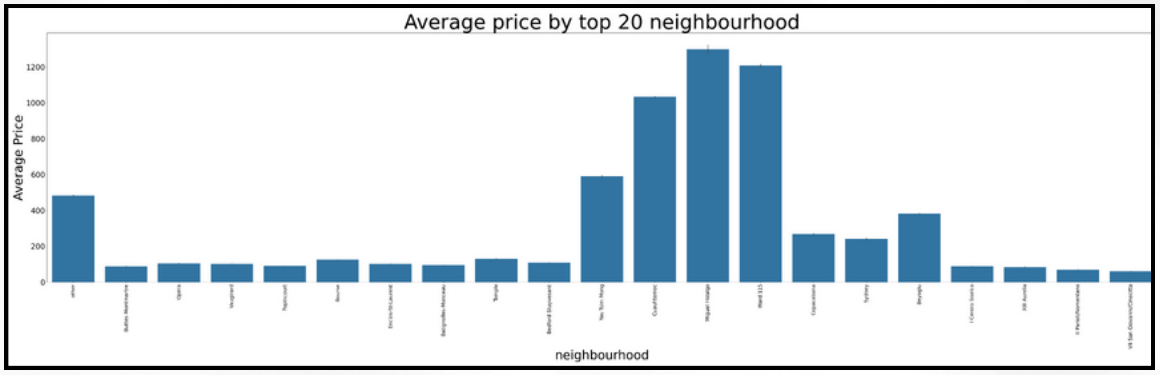

Figure.4 Average Prices Top 20 Neighborhoods

The visualization depicting the average accommodation prices in each city as shown below (Figure.5) revealed significant disparities in pricing levels among the analyzed cities.

- Bangkok emerged as the city with the highest average accommodation price, indicating that properties in Bangkok command premium prices within the Airbnb market. This finding underscores Bangkok's status as a popular tourist destination and suggests that factors such as high demand, limited supply, and the city's appeal to luxury travelers contribute to its elevated pricing levels.
- Rome exhibited the lowest average accommodation price among the analyzed cities. This suggests that properties in Rome are more affordably priced compared to other destinations in the dataset, potentially making it an attractive option for budget-conscious travelers or those seeking economical accommodation options. Rome's historical significance, cultural attractions, and diverse range of accommodations may contribute to its appeal as a cost-effective destination for travelers.

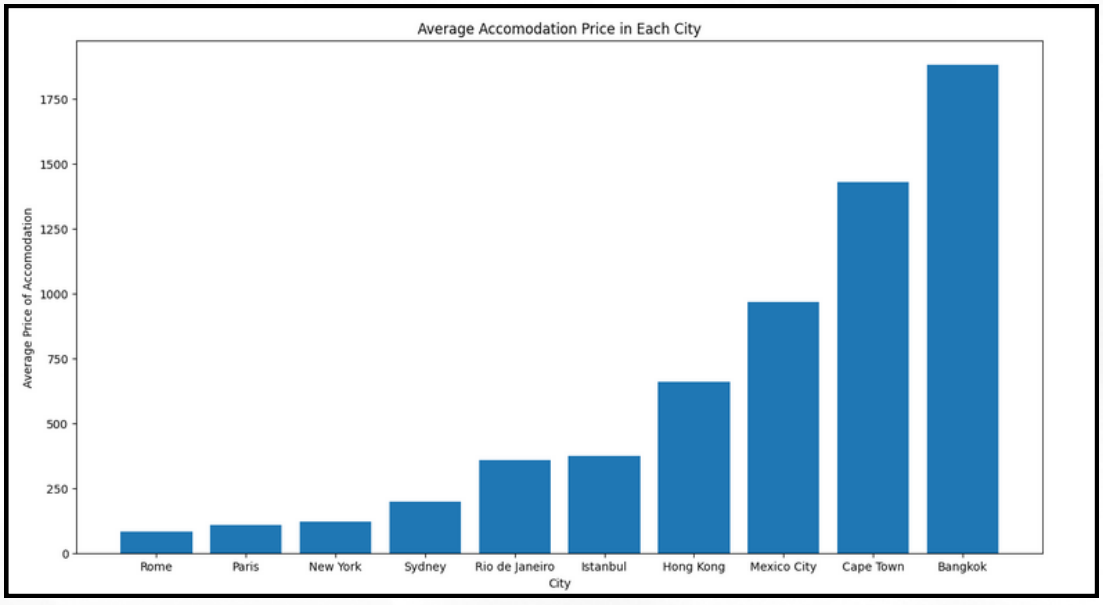

Figure.5 Average Accommodation Prices in each City

Also, based on the visualization showing the most popular property type in each city as shown below (Figure.6), with 'entire apartment' being the most prevalent type and Paris having the highest number of entire apartments followed by Rome, New York, and Hong Kong, I can draw several insights:

- The prevalence of 'entire apartment' as the most popular property type across multiple cities suggests a universal preference among travelers for this accommodation type. It may offer a combination of privacy, space, and amenities that appeal to a wide range of travelers, from families to adventurers.
- The high number of entire apartments in Paris reflects its reputation as a global cultural and economic hub, attracting a diverse array of visitors seeking long-term or short-term accommodations in the heart of the city.
- The lower prevalence of entire apartments in Hong Kong could be influenced by factors such as limited space, high property prices, and a preference for alternative accommodation options like hotels or serviced apartments.

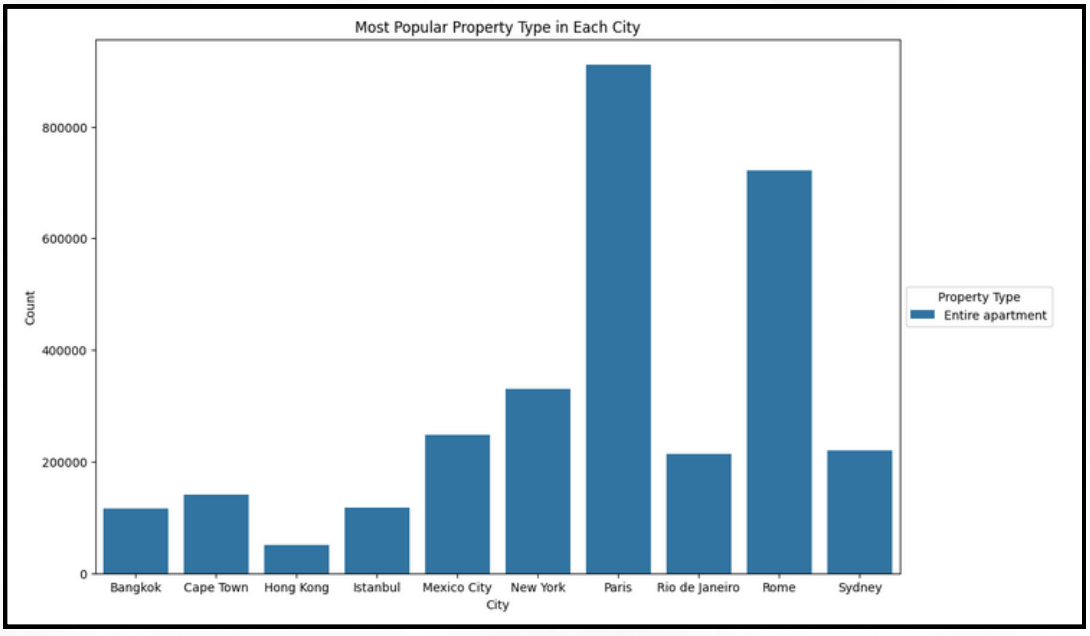

Figure.6 Most Popular Property type in each City

From heatmap as shown below (Figure.7) I observed a strong positive correlation between "review scores cleanliness" and "review scores rating", suggesting that properties with higher ratings for cleanliness tend to also receive higher overall ratings from guests. This insight underscores the importance of maintaining cleanliness standards in accommodations, as it directly impacts guest satisfaction and overall ratings.

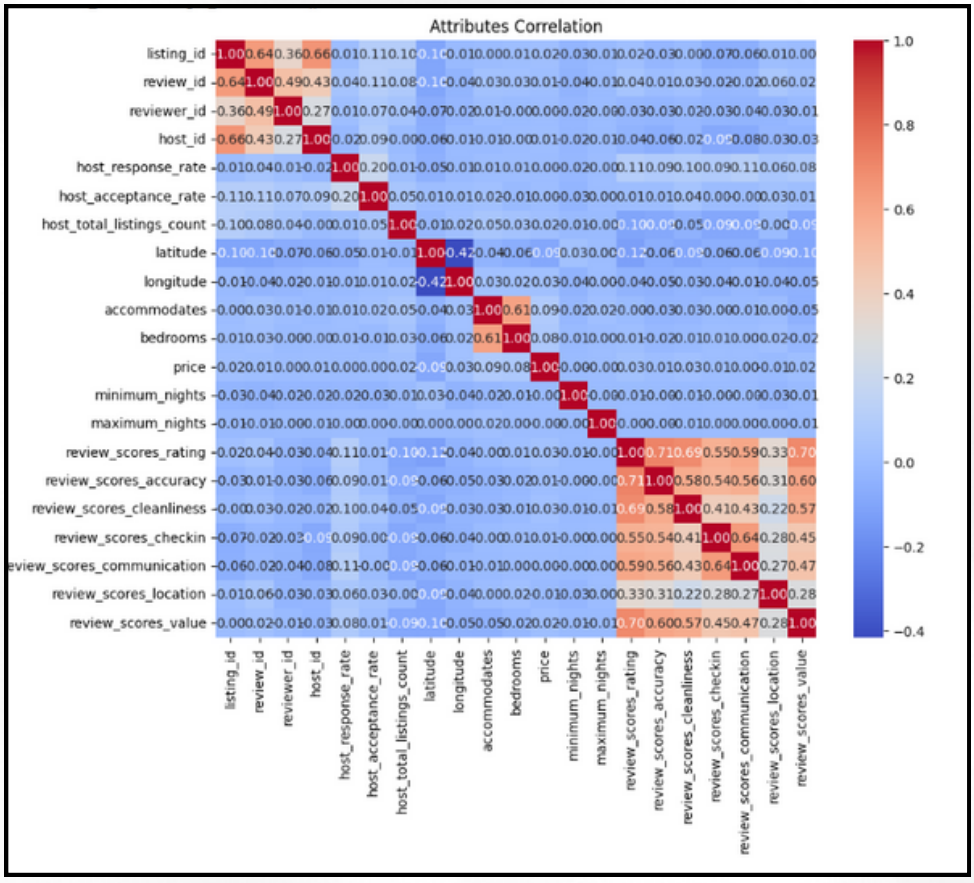

Figure.7 Heatmap (Relations between columns)

Moreover, based on the box plot visualization showing the distribution of the number of amenities by room type as shown below (Figure.8) , with 'entire place' having the highest number of amenities followed by private room and shared room, while hotel room has the lowest number of amenities, you can draw several insights:

- The higher number of amenities associated with 'entire place' accommodations suggests that hosts or property managers strive to provide a comprehensive range of facilities and services to enhance the guest experience.
- Private room and shared room accommodations typically offer a moderate number of amenities compared to 'entire place' listings. While they may not provide the same level of privacy and exclusivity as entire accommodations, they still offer essential amenities to ensure guests' comfort and convenience during their stay.
- The lower number of amenities associated with hotel rooms reflects that hotels may differentiate themselves based on service quality, brand reputation, and location rather than the sheer number of amenities offered. Guests may prioritize factors such as convenience, accessibility, and personalized service when choosing hotel accommodations.

Also, I noticed that all values have outliers that we should deal with it.

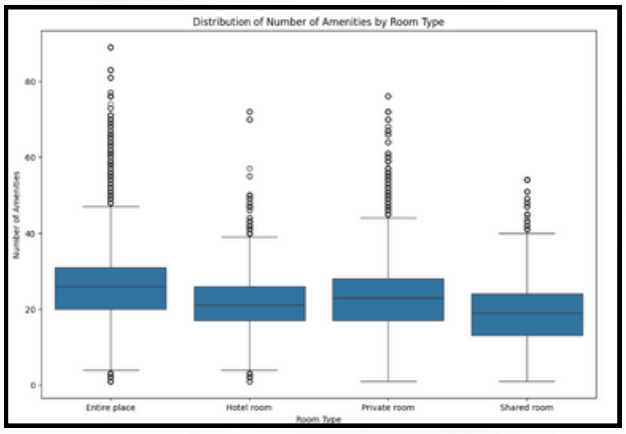

Figure.8 Number of amenities by room type

Feature Engineering:

In the Feature Engineering phase, I add a new feature named 'average_rating' was created to provide a consolidated metric representing the overall rating of Airbnb listings based on multiple review scores. This feature was derived by calculating the mean of seven individual review scores: 'review_scores_rating', 'review_scores_accuracy', 'review_scores_cleanliness', 'review_scores_checkin', 'review_scores_communication', 'review_scores_location', and 'review_scores_value'.

By aggregating these review scores into a single composite measure, the 'average_rating' feature offers a comprehensive assessment of the quality and satisfaction levels associated with each listing.

Upon calculating the 'average_rating' feature across the dataset, it was observed that the highest average rating recorded was approximately 22.86, indicating exceptional performance across all reviewed aspects, while the lowest average rating stood at approximately 4.57. The mode of the 'average_rating' feature was found to be approximately 22.57, suggesting that this value was the most frequently occurring average rating across the dataset.

In addition to the 'average_rating' feature, another valuable metric, 'occupancy_rate', was engineered to provide insights into the utilization and booking patterns of Airbnb properties. The 'occupancy_rate' feature was computed by dividing the number of minimum nights booked by the maximum nights available for booking, and then multiplying the result by 100 to obtain a percentage.

The resulting 'occupancy_rate' values represent the estimated occupancy level of each property, indicating the proportion of available nights that have been booked by guests. A higher 'occupancy_rate' suggests greater demand and utilization of the property, while a lower rate may indicate underutilization or availability for additional bookings.

Upon calculating the 'occupancy_rate' feature across the dataset, a range of values was observed, spanning from 10% to 0%. These values reflect the varying degrees of occupancy observed among the listed properties, with some experiencing higher booking rates and others having more availability for potential guests.

Recommendations:

Several suggestions for both visitors and hospitality industry stakeholders can be made based on the analysis that was done:

1. Prioritize Cleanliness Standards: It is imperative that lodging providers give priority to cleanliness standards because of the significant positive association that has been shown between review scores for cleanliness and overall ratings. Having regular maintenance and strict cleaning procedures in place can help increase visitor happiness and favorable feedback.
2. Competitive Pricing Strategies: In locations like Bangkok where the average cost of lodging is greater, lodging providers should make sure that their pricing strategies reflect the perceived value of their services. Affordability can also be used as a competitive advantage by places with lower average rates, like Rome, to draw tourists on a tight budget.
3. Diversify Property Offerings: Accommodation providers should consider diversifying their property offerings to meet the varied preferences of travelers. While 'entire apartment' listings are popular across multiple cities, there is still demand for other types of accommodations such as private rooms and shared rooms. Offering a diverse range of options can attract a broader customer base.

By implementing these recommendations, stakeholders can enhance the overall guest experience, drive customer satisfaction, and achieve business success in the competitive hospitality industry.