Airbnb Rental Trends in Los Angeles: A Big Data Analysis

Yuci Zhang, Lingyan Hu, Congxiao Wang

Ⅰ. Motivation

The sharing economy has significantly changed the way people travel and live, and platforms like Airbnb have played a significant role in this transformation. Understanding the impact of Airbnb on the market is essential for policymakers, industry leaders, and users alike. The Airbnb market in Los Angeles is a critical area for investigation due to its size, diversity, and economic significance. The dataset we will be exploring in this project includes several powerful pieces of information such as customer reviews and their ratings. By analyzing this data, we aim to investigate trends in pricing, customer behavior, and the interactions between different relevant factors.

Some studies have investigated these phenomena in the past, and we aim to build on this foundation to gain a more comprehensive understanding of the Los Angeles Airbnb market. "What Makes You Choose Airbnb Again?" by Wang et al. (2018), which explores the factors that influence Airbnb listing prices using regression analysis. Other relevant studies we will draw from include Heo and Blengini's "A macroeconomic perspective on Airbnb's global presence," which examines macroeconomic factors affecting Airbnb's global presence (Heo and Blengini 2019), and Leick's "Digital entrepreneurs in the sharing economy: A case study on Airbnb and regional economic development in Norway," which investigates the influence of traditional accommodation and unemployment on digital entrepreneurs associated with Airbnb (Leick 2020). All these studies suggest that there is ample room for us to analyze and research this market and its data.

We will explore certain aspects from our data, here are some examples. One aspect we aim to explore is the users' perspectives by examining the words behind their comments, which could be useful tools to understand the relationship between hosts and users on the platform. We will analyze users' reviews to identify factors that impact the quality of Airbnb listings, and we will use this information to build an evaluation criterion to quantify the Airbnb feedback provided by users to hosts. We also seek to identify what users care about the most. These insights will not only be beneficial to Airbnb users but also to companies operating in the sharing economy. Additionally, we will explore the market from hosts' perspectives. For instance, we will investigate factors that influence pricing, the relationships among the number of Airbnb properties each host owns, the response time, and the response rate for examining the efficiency of each host.

Having established the importance of studying Airbnb's data, we will then delve deeper into the specific topics of user perspectives, host factors, and market trends in the following paragraphs to further elucidate our understanding of the Los Angeles Airbnb market.

Ⅱ. Data Sources

The Airbnb dataset for Los Angeles is a valuable resource for companies seeking to gain insights into customer behavior and develop effective pricing strategies. The datasets contain a wealth of information, including customer reviews and listing attributes. The reviews dataset comprises 1,357,181 rows of data, spanning from May 26th, 2009 to December 23rd, 2022, and 6 features, such as the reviewer's information and comments on listings are contained. Meanwhile, the listing dataset contains information on 40,438 listings and 75 features, which is spanning from May 26th, 2009 to March 8th, 2023, including listings locations, room type, check-in information, and customer rating scores etc.

To access the Airbnb dataset, users can download it from the Inside Airbnb website in either CSV or JSON format[1]. What we choose is the CSV format. The website provides detailed information on the dataset's structure and variables, making it easy for users to navigate and analyze the data. By using this dataset, companies can gain valuable insights into customer behavior and develop effective pricing strategies tailored to specific listings.

Listing data

Name	Type	Description
price	text	daily price in local currency
Host_acceptance_rate	text	That rate at which a host accepts booking requests.
room_type	text	Three room types: Entire place, Private room, Shared room.
has_availability	bool	If the airbnb is available. [t=true; f=false]
review_scores_rating (Accuracy/ Cleanliness/ CheckIn/ Communication/ Location/Value)	numeric	Feedback from user.

Review data

Name	Type	Description
listing_id	numeric	The id from listing data, Airbnb's unique identifier for the listing.
date	date	Comment date.
reviewer_id	numeric	User id corresponding to the comment.
comments	text	The comment from the user.
id	numeric	Comment id.

III.Data Manipulation Methods

This enriched dataset, containing both listing and review information, facilitates a more in-depth analysis of user perspectives, host factors, and market trends in the Los Angeles Airbnb market.

3.1 Reviews data:

The review data is subject to a series of transformations to protect user privacy and enhance the analytical outcomes.

- **Step 1 - Remove data:** to begin with, we remove review names (using drop) to respect privacy concerns while keeping the reviewer ID to facilitate identification. Additionally, the review ID column is removed, as it doesn't contribute to the overall understanding of the data.
- **Step 2 - Convert Data:** we then modify the date format, converting it to a date_time format, which allows for more detailed categorizations such as year, month, and date (using to_datetime). This is beneficial for time-based analysis and trend identification.
- **Step 3 - Drop Missing Data:** the quality of the data is crucial for accurate analysis, and to ensure this, we implemented measures to remove 250 instances with missing comments. Given that our dataset contains 1,357,181 comments, removing these 250 instances will not significantly impact the overall analysis (using dropna).
- **Step 4 - Process the text value 'Comments':** Special symbols or signs in the comments variable are also deleted to ensure the text is standardized and easier to analyze (using str.replace) and merge all comments together(using join), select adjective words in all comments (using nltk - word_tokenize, pos_tag)
- **Step 5 - The variable used to join another dataset:** we retain the listing ID, which serves as a crucial link to combine review data with listing data. This step is essential, as performing an inner join with listing data enables us to create a more comprehensive dataset.

3.2 Listings data:

The listing data undergoes several manipulations to improve the quality and utility of the dataset for analysis:

- **Step 1 - Drop unuseful columns:** We remove several unnecessary columns (using drop), such as 'listing_url', 'description', 'host_thumbnail_url', 'host_picture_url', 'picture_url', 'host_url', and 'host_location' etc., to streamline the dataset.
- **Step 2 - Convert certain variables:**
    - *Response_time:* we replace null values with 'Unknown' and convert the variable to a categorical format (using pd.categorical).
    - *Response_rate:* we first convert 'host_response_rate' to a double format, then fill 'host_listings_count' null values with the mean value. We categorize the response rate as 'low' for values below 0.5, 'medium' for values between 0.5 and 0.9, and 'high' for values greater than 0.9 (using pd.cut).
    - *Price:* we process the price variable, which has no null values, by removing '$' and ',' symbols and converting it to a float format (using str.replace and astype).
    - *Acceptance_rate:* we process the acceptance rate variable (which has no null values) by removing '%' and converting it to a float format (using str.replace and astype).
- **Step 3 - Create a new variable:** count_amenitites, which counts the number of amenities each listing provides.(using for loop)

3.3 Combine data:

- **Step 1 - Rename columns:** we create a new column called 'list_id' so that it corresponds to the reviews data and drop the original ‘id’ column(using drop). This change enables us to perform an inner join between the two datasets.
- **Step 2 - Groupby:** For the reviews dataset, we calculated the mean of ‘sentimental_polarity’ for each property. To avoid data redundancy when merging two tables later, we first aggregate the df_reviews data frame by the id of each Airbnb listing, because each Airbnb listing has multiple reviews, through grouping by 'list_id' and ‘sentimental_polarity’ and for the following efficient and precise analysis while processing numerous dataset.(using groupby and mean)
- **Step 3 - Merge two datasets:** Using ‘pd.merge’ to combine Reviews and Listings datasets on the column of listing_id.(use pd.merge)

3.4 Challenge and Discussion:

- When using TextBlob for sentiment analysis, the tool may not accurately classify attitudes expressed in languages other than English, as non-English comments tend to be classified as neutral, disregarding positive or negative sentiments. Due to time constraints, we have decided to exclude non-English comments, which make up only a small portion of the dataset and are unlikely to significantly affect the overall analysis results.
- The listings dataset contains numerous features, and it is impractical to analyze all of them within a limited timeframe. Thus, it is essential for us to identify the most relevant features for our analysis and selectively preprocess and utilize those data points.

IV. Visualization and Analysis based on Listing Data

4.1 Host Perspectives

a. What is the relationship between the room type of Airbnb and their location?

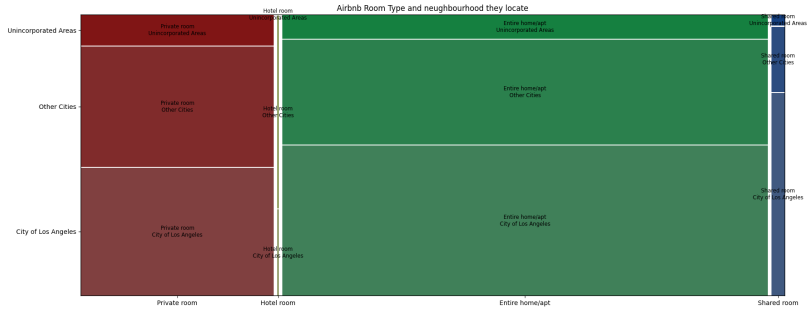

Figure 1. The ratio distribution of Airbnb room type and neighborhood located

From the visualization Figure 1, it is noticeable that Entire home/apt constitutes the largest proportion of room_type, followed by Private room. Shared room and Hotel room, on the other hand, make up a significantly smaller portion. This pattern suggests that most hosts prefer providing guests with a private environment, reflecting the importance of privacy in users' decision-making when choosing Airbnb accommodations in Los Angeles. Additionally, all types of rooms share a common characteristic regarding their location. The majority are situated in the city of Los Angeles, while a considerable number can also be

found in surrounding cities. A more detailed analysis of the specific cities where these listings are located is presented below.

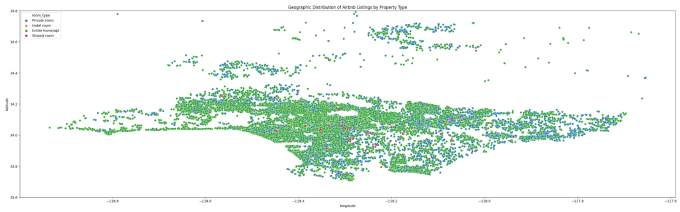

Figure 2. The map illustrates the distribution of room types of Airbnb in Los Angeles

The graph Figure 2 is created by using longitude and latitude data with room types, and shows that the most prevalent room type across the region is Entire Home/Apt, followed by a smaller number of Private and Shared rooms. Hotel rooms are relatively rare in the graph.

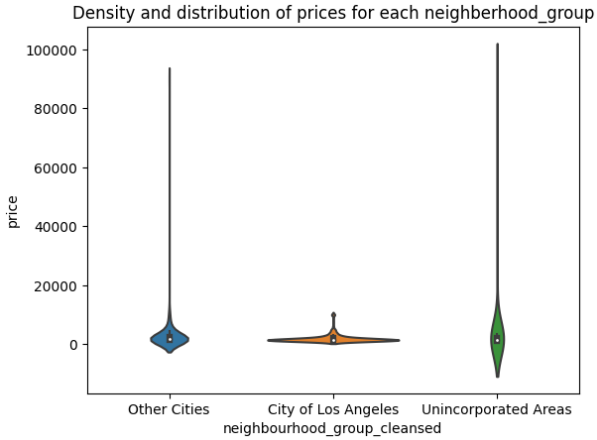

Figure 3. The distribution of prices over $1000 for each neighborhood in Los Angeles

While the city of Los Angeles hosts a significant number of properties, it's worth noting that other cities and unincorporated areas exhibit higher price ranges and extreme values, which is presented by Figure 3. These variations can stem from factors such as location, amenities, accessibility, and demand.

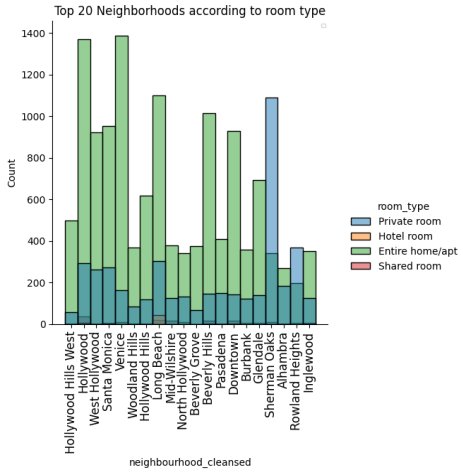

Figure 4. The distribution of listings with room types across the 20 most available neighborhoods

From the visualization Figure 4, we can distinguish that most Airbnb properties hosted are located in popular areas such as Hollywood Hills, Santa Monica, Long Beach, and Sherman Oaks, all of which are iconic neighborhoods in Los Angeles. Among these, only Sherman Oaks predominantly features Private rooms. This might be because Sherman Oaks is a relatively safe spot of the city of Los Angeles, characterized by high prices and low crime rate. Consequently, hosts may take users' financial considerations into account, and renting a single room offers a more affordable option compared to an entire Airbnb property.

In the top 20 neighborhoods, the majority of Airbnb properties are Entire homes or apartments. This suggests that as Los Angeles is a renowned tourist destination, visitors prioritize a comfortable and relaxing experience. By choosing entire Airbnb properties, they can ensure a higher quality stay while enjoying their travels accommodates their economic choices.

b. Host Service Quality: The host who owns more Airbnb may have a lower response rate and longer response time?

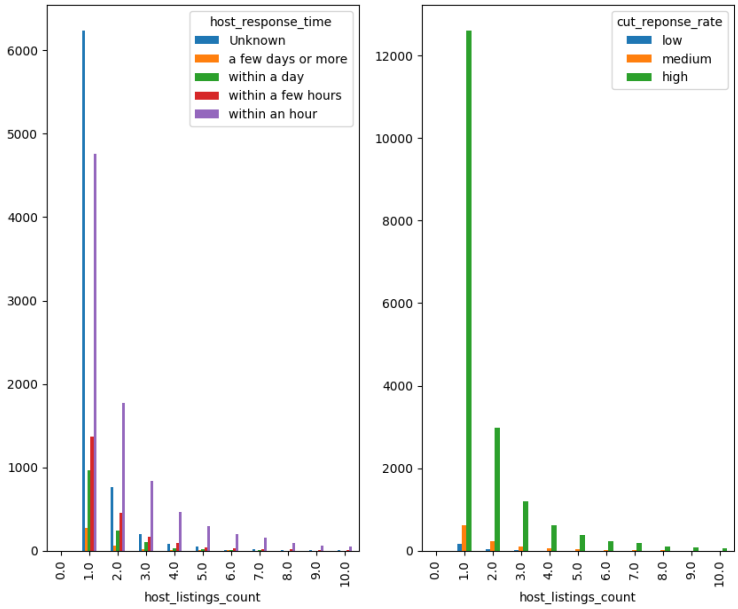

Figure 5. The relationship between Airbnb properties quantities and the response from hosts performances

In the listings data, we exclude extreme values, such as hosts with more than 10 listings. The mean is considerably high due to the influence of these extreme values. However, the 50th percentile indicates that most hosts manage just one Airbnb property, while the 75th percentile shows that hosts manage only two, demonstrating a discrepancy with the mean value, which is demonstrated by the left side chart on Figure 5.

The analysis presented by Figure 5, reveals that regardless of the number of listings a host manages, many can retain users over time and demonstrate significant response efficiency. The graph indicates that there is no clear relationship between the number of listings a host holds and their response time or response rate. For instance, hosts with a single listing may exhibit varying response times/rates. To gain a deeper understanding of the factors that could influence response speed, further investigation may be necessary.

4.2 User Perspectives

a. How could an user find the Airbnb they want?

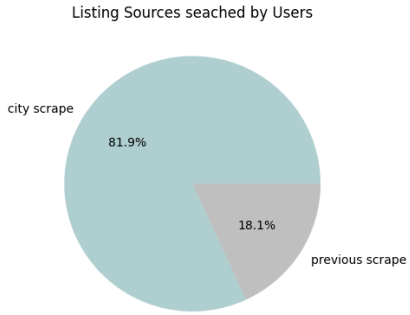

Figure 6. The pie chart illustrates the distribution of listings searched by users

"City scrape" refers to listings found by searching for the city, whereas "previous scrape" denotes listings that were seen in another scrape conducted within the past 65 days and confirmed to be still available on the Airbnb site. From the visualization Figure 6, we can see that the majority of users find an Airbnb by searching for the city in which it is located, and a smaller portion of listings is discovered through historical records.

b. Does an Airbnb with more amenities have a higher price? What about other factors?

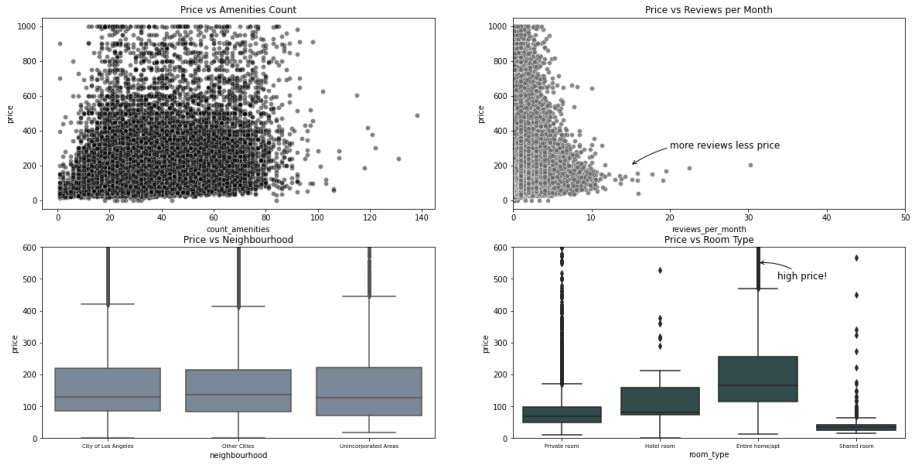

Figure 7. The overview analysis within price and four related factors

- The first visualization shows that the number of amenities provided by listings does not have a clear impact on its price.
- The second visualization reveals that the number of reviews per month has no significant influence on price when there are fewer than 10 reviews. However, when there are more than 15 reviews, the price tends to be relatively lower. This could be because budget-friendly Airbnb listings appeal to a wider audience.
- The third visualization indicates that the listing location does not significantly affect its price.
- In the fourth visualization, it is evident that Entire homes or apartments have the highest average price among room types. Following that, the prices from high to low are: Hotel rooms, Private rooms, and Shared rooms.

V. NLP and Analysis based on Reviews Data

After process Reviews and combine data, we make NLP related analysis:

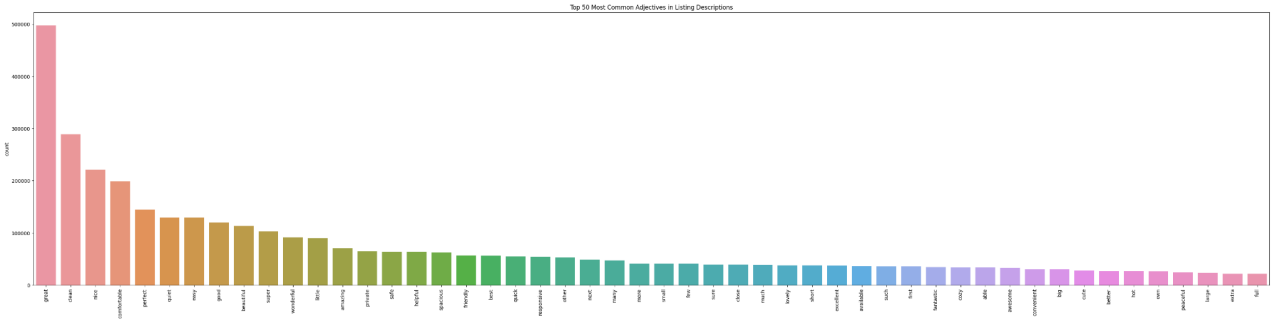

Figure 8. The top 50 most frequently used adjectives in reviews

The code generates a count plot, Figure 8, of the top 50 most common adjectives in user comments, sorted by the frequency of each adjective in descending order. This visualization highlights that many users employ multiple adjectives to describe similar qualities they experienced during their stay, such as "beautiful" and "wonderful" or "good" and "great". In essence, the plot sheds light on the language used in user comments and the characteristics guests found noteworthy during their Airbnb stays.

However, one limitation of the analysis is that it solely considers the frequency of each adjective in the comments, without accounting for the context in which they appear. For instance, the word "good" could be employed to positively describe a feature, or it might be used sarcastically to indicate a negative aspect. In order to address this concern, we have decided to exclude words like "great" and "good" and instead focus on words that reflect the factors that users value most on Airbnb, such as "safe" and "privacy."

To achieve this, we have created word clouds that visually illustrate the frequency of these words, with larger and bolder words indicating a higher frequency of use. Upon examining the word cloud, we have found that there are numerous high-frequency words that describe both the properties and the hosts. These words suggest that users are seeking accommodations that are clean, comfortable, and spacious, with modern amenities and a peaceful and comfortable atmosphere. Additionally, users place a high priority on safety and convenience when selecting their accommodations.

Figure 9. The top 10 positive most frequently used adjectives

The terms "friendly" and "responsive" are subjective and evaluate the behavior and gestures of hosts, while the other words used predominantly describe the ambiance of the places. Interestingly, the top 10 most

commonly used words in the review section are primarily related to the accommodations themselves rather than the hosts. In summary, the word cloud, Figure 9, offers valuable insights into the adjectives commonly used by users to describe their accommodation preferences, which can assist hosts and property managers in understanding guest expectations and adapting their offerings accordingly. This analysis also sheds light on the key attributes that are currently popular in the Airbnb market.

VI. Model and Analysis based on Combined data

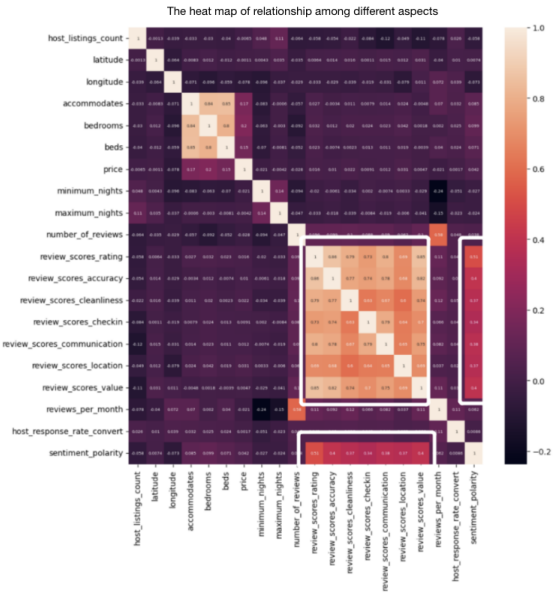

Figure 10. The heatmap that indicates the correlation among different valuable attributes

The graph, Figure 10, illustrates that review scores based on different aspects have a strong relationship with sentiment polarity. To delve further into this relationship, we will employ the Pearson correlation test to examine the connection between review scores and the sentiment expressed in users' comments. In addition, In airbnb feedback data, we worry about discrepancy between ratings and reviews on online marketplaces. For example Some users may receive messages from hosts and give high scores to Airbnb, but their actual experience may not be good, and they will express their true feelings in the comments. Therefore, we will use statistical tests to further examine the relationship between the two and evaluate whether the ratings given by Airbnb users are reliable.

In this study, we aim to explore the consistency between users' review scores and the content of their actual comments on Airbnb listings. By examining both quantitative and qualitative data, we assess the degree of alignment between the ratings given by guests and the sentiments expressed in their written feedback in the following. This analysis will provide a deeper understanding of guest satisfaction, allowing hosts and property managers to identify potential discrepancies and make more informed decisions when evaluating their property's performance based on user feedback.

```
import scipy.stats as stats
# perform Pearson correlation test between sentiment_polarity and each variable
for col in df_behavior.columns[:-1]:
    correlation, pvalue = stats.pearsonr(df_behavior[col], df_behavior['sentiment_polarity'])
    if pvalue < 0.05:
        print(f"There is a significant positive correlation between {col} and sentiment_polarity (p-value: {pvalue:.4f})")
    else:
        print(f"There is no significant correlation between {col} and sentiment_polarity (p-value: {pvalue:.4f})")
```

✓ 0.0s

```
There is a significant positive correlation between review_scores_rating and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_accuracy and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_cleanliness and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_checkin and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_communication and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_location and sentiment_polarity (p-value: 0.0000)
There is a significant positive correlation between review_scores_value and sentiment_polarity (p-value: 0.0000)
```

Figure 11. The screenshot of using statistical analysis for examining the correlation between reviews and sentiments

Upon applying the Pearson correlation test to investigate the hypothesis we described earlier, we discovered a strong positive correlation between users' review scores and the sentiment of their actual comments on Airbnb listings. This finding indicates that the review scores are highly consistent with the sentiments expressed in users' written feedback. As a result, hosts and property managers can confidently rely on guest ratings as a reliable indicator of their property's performance and guest satisfaction.

Ⅶ. Conclusion

In the listing dataset we visualize the data from host and user perspective and provide interesting views, such as hosts of different listing numbers having varying response times and rates, suggesting that the number of listings does not necessarily affect the quality of service.

In the reviews dataset, we analyze the most frequently used adjectives in user comments and highlights the qualities that guests find noteworthy during their Airbnb stays. The insights gained from the word cloud can assist hosts and property managers in understanding guest expectations and could be applied in line with adapting their offerings accordingly.

In the combined data, our analysis aimed to examine the consistency between guests' review scores and their actual comments on Airbnb listings, providing hosts and property managers with a more informed evaluation of their property's performance. The results showed a strong positive correlation between review scores and sentiment expressed in comments, indicating that guest ratings can be a reliable indicator of guest satisfaction. Therefore, hosts and property managers can use this information to improve their property's performance and enhance guest experience.

With the rise of the sharing economy, platforms like Airbnb have disrupted traditional industries and changed the way people travel and live. The sharing economy represents a shift towards more collaborative and community-based forms of consumption and production. By studying the dataset from Airbnb that consists of possible aspects impacting price, corresponding consumers’ reviews, and consumers’ attitudes and responses, researchers can gain insights into the values and preferences of consumers, as well as their willingness to participate in collaborative consumption models.

Ⅶ. Statement of Work

Our team of three worked collaboratively on this project, with each member contributing to various aspects of the tasks, including data preprocessing, model development, visualization, and reporting. The unique skills and expertise of each team member played a significant role in achieving our goals.

Lingyan Hu primarily focused on coding and implementation, ensuring that the datasets were properly preprocessed and the models developed correctly. She also contributed to the visualization and report sections. Congyang Wang and Yuci Zhang mainly worked on the report, interpreting the results and writing a clear and concise summary of our findings. Both also made contributions during the data preprocessing and model development stages, as well as the visualization.

Throughout the project, all team members were engaged in providing ideas and feedback for each portion of the work, fostering an environment of open communication and collaboration. In future projects, we aim to continue this collaborative approach, leveraging the diverse skills of each team member to optimize the quality of our work. By actively participating in all aspects of the project and maintaining open communication channels, we can ensure a successful outcome and continue to improve our collaboration.

References:

[1]: http://insideairbnb.com/get-the-data.html

Heo, Cindy Yoonjoung ; Blengini, Isabella. “A Macroeconomic Perspective on Airbnb’s Global Presence.” International Journal of Hospitality Management, vol. 78, Elsevier Ltd, pp. 47–49, doi:10.1016/j.ijhm.2018.11.013.

Leick, B., Kivedal, B. K., Eklund, M. A., & Vinogradov, E. (2022). Exploring the relationship between Airbnb and traditional accommodation for regional variations of tourism markets. Tourism Economics, 28(5), 1258–1279. https://doi.org/10.1177/1354816621990173

Wang, Chuhan (Renee) ; Jeong, Miyoung. “What Makes You Choose Airbnb Again? An Examination of Users’ Perceptions toward the Website and Their Stay.” International Journal of Hospitality Management, vol. 74, Elsevier Ltd, pp. 162–70, doi:10.1016/j.ijhm.2018.04.006.