Research Article

Use Python Data Analysis to Gain Insights from Airbnb Hosts

Zichun Tian [ID]

Olin Business School, Washington University in St. Louis, St. Louis 63130, USA

Correspondence should be addressed to Zichun Tian; t.zichun@wustl.edu

This paper uses Python and its external data processing package to conduct an in-depth analysis machine study of Airbnb review data. Increasingly, travelers are now using Airbnb instead of staying in traditional hotels. However, in such a growing and competitive Airbnb market, many hosts may find it difficult to make their listings attractive among the many. With the development of data science, the author can now analyse large amounts of data to obtain compelling evidence that helps Airbnb hosts find certain patterns in some popular properties. By learning and emulating these patterns, many hosts may be able to increase the popularity of their properties. By using Python to analyse all data from all aspects of Airbnb listings, the author proposes to test and find correlations between certain variables and popular listings. To ensure that the results are representative and general, the author used a database containing many multidimensional details and information about Airbnb listings to date. To obtain the desired results, the author uses the Pandas, NLTK, and matplotlib packages to better process and visualize the data. Finally, the author will make some recommendations to Airbnb hosts based on the evidence generated from the data in many ways. In the future, the author will build on this to further optimize the design.

1. Introduction

Even though Airbnb has gained a lot of popularity since its inception, there is increasingly emerging voice stating that they would rather choose hotels than Airbnb. According to a news report, some reasons include inaccurate descriptions of properties, prices being not cheaper than hotels, and long response time when communicating with hosts [1]. The annual booking times have a decreasing growth rate since 2018, the accumulated listing shows a decreasing growth rate since 2017, and the closest competitor Vrbo has started to occupy the market shares that originally belong to Airbnb [2].

When the time comes to 2020, the outbreak of COVID-19 has brought even more pressure on Airbnb hosts. Due to the uncertainties of the travel plans, Airbnb allowed a more flexible cancel policy when booking a place which will increase the opportunity costs of the hosts [3]. A report says that 70% of guests are fearful to stay at an Airbnb compared to hotels that have more strict and standardized cleaning protocols if they must travel during the pandemic. For the hosts, the cost of accepting guests also increases because of the stricter hygiene procedures [4]. As a result, the hosts

and Airbnb itself are now facing new challenges both from inside and outside. Since the diversified hosts and their properties that are located almost everywhere in the world are the core competitiveness of Airbnb, Airbnb should help them become more appealing through multiple ways to compete with hotels and other rivals.

The author tried to find some specific patterns that are most common among popular properties and apply these patterns to other properties to maximize advantages and attractiveness to guests [5].

To better help understand, the author wants to clarify some terms that are frequently used in the Airbnb system. The hosts mean the property owners who rent guests the places to stay. Guests are people that do not want to live in hotels and rent the hosts' properties to stay. The listings are created when hosts put their properties on the Airbnb system and are available for guests to stay. Listings include all the information that describes the properties like price, description, and location.

The database the author used is downloaded from a website called "Inside Airbnb" which provided all the relative data of Airbnb listings from all locations around the world.

The author used the data from the Los Angeles site which includes three CSV files. The first file is called calendar which includes all the listing details including the maximum and minimum price of every listing in Los Angeles daily [6]. Since each listing is highly likely to be listed more than one day, there will be many repeated data that need to be eliminated when the author processes the data [7]. The second file is called reviews, and it records all the reviews and everything related to that like dates, reviewer names, and unique IDs. These reviews are given by guests that are followed by every stay or transaction [8]. The reviews are very useful because the author can do sentiment analysis to find the most popular words that are valued by the guests to help Airbnb hosts improve. The last file is called listings which has much more data than the previous two files. It contains all the information of every Airbnb listing in Los Angeles. Some valuable information is room types, host response time, review scores of each section from guests, room descriptions, etc.

These three files can be connected through unique IDs, and then, the author can combine all these factors to find all the relations that existed in these data.

2. Current Status of Research

In many fields of scientific research, for the same problem, there are often different scholars using the same or different scientific methods to conduct research and experiments, and the conclusions are not the same. In this case, how can the author synthesize the results of different existing studies to get a more reliable conclusion? Meta-analysis is a statistical method that analyses and generalizes the data collected from multiple studies to provide a quantitative average effect to answer the research question [9]. Its advantage is that it increases the credibility of the conclusions by increasing the sample content and resolves inconsistencies in research findings. It is a quantitative review of the literature, which is a systematic, objective, and quantitative synthesis of the results of multiple independent studies on the same topic, based on a rigorous design, using appropriate statistical methods.

To conduct meta-analysis, the first step is to determine the effect values of the study results, i.e., the statistical quantities that can be used to measure the good or bad results of the study, and usually, correlation coefficients, relative ratios, and standardized relative differences can be used as effect values [10]. Consistent effect values are the basis of meta-analysis, and only when the effect values are unified is it possible and reliable to conduct a comprehensive analysis of the results [11]. In practical research, more than one effect value is often needed to evaluate the results of an experiment or study [12]. For example, in the field of education, to evaluate a student's good or bad performance, one cannot only look at the score of one subject but needs the combination of multiple subjects; in the medical field, to conduct a hypertension drug test, one needs to measure both the blood pressure of the heart during systole and the blood pressure of the heart during diastole; in the financial field, several indicators are reflecting the liquidity risk of an enterprise, such

as current ratio, quick ratio, and short-term cash service multiple [13].

To conduct simulation experiments and eventually analysis, you need to write programs through computer languages to achieve, at this stage, the most used computer language for analysing big data which is mainly two, one is Python and one is R. Among them, Python is more powerful; in addition to data processing, modelling analysis can also be website development, game development, etc., which can be said to be the hottest computer language. It is more suitable for technical people who have some computer theory foundation and aim at engineering development. R was originally developed to help users to do data analysis, statistical modelling, visualization models, etc., in a user-friendly and fast way, and it has a very powerful open-source library. Users can easily call the packages to build their models without programming their implementation, which is easier to implement for researchers who are not very good at computer languages and do not need to spend much time and effort here.

3. Python Methods

3.1. Experimental Method. Since the raw data the author collected is disorderly and chaotic, the author must choose the right method in Python to process and visualize these data [14]. There are many packages and preinstalled programs available to choose from in Python; the author must choose the right one to facilitate our works:

$$I_i = \alpha\theta_i^2 + \beta\theta_i^2 + \chi\theta_i^2. \tag{1}$$

The first method this paper used in Python is the Pandas package. Pandas is a package that has been frequently used when manipulating and analysing data. It allows this paper to import data from various formats and lets the author do some certain manipulation on data as desired [15]:

$$G_j = \frac{\sum_{i,j=1}^{m,n} R_i * U_{ij}^2}{\sum_{i,j=1}^{m,n} E_i * U_{ij}^2}. \tag{2}$$

The Python Pandas package has been frequently used in quantitative finance applications in recent years. In a report that mainly focuses on statistical computing in Python, the author used the Pandas package to reshape the primary data set which contains the stock prices of some certain companies and industries. With the function of easily changing DataFrame to series and removing dummy variables, the Pandas package can save much time used to prepare data:

$$CSV = (H(m_i) + H(n_i)) - \left(H\left(m_j\right) + H\left(n_j\right)\right), \tag{3}$$

$$M' = \left\{ M' = m - m_c, \ \forall m \in M \right\}, \tag{4}$$

$$M'' = \left\{ M'' \middle| m'' = \frac{m' \times d}{D}, \ \forall m' \in M' \right\}. \tag{5}$$
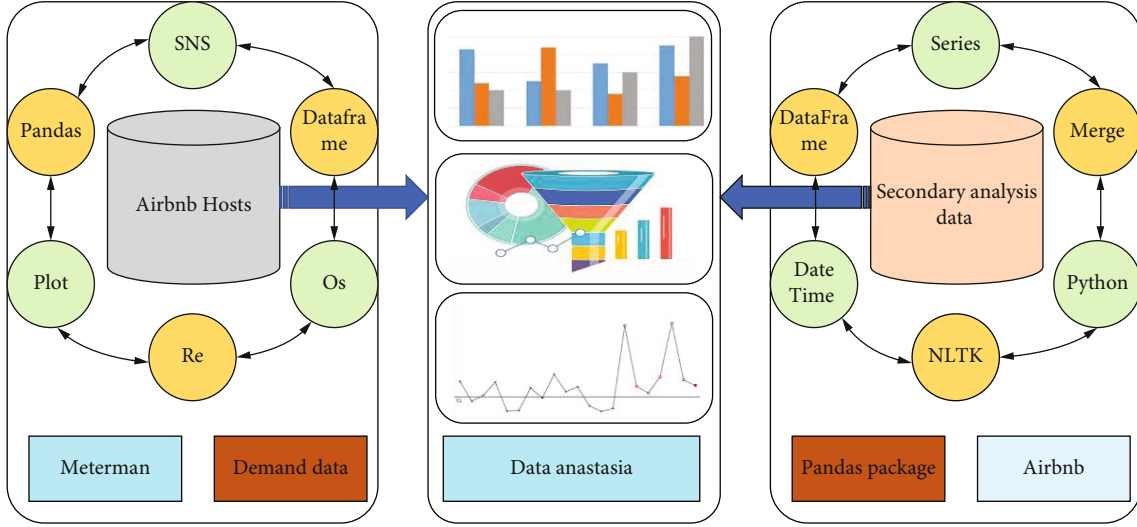
FIGURE 1: Python data analysis framework.

In this case, this paper use Pandas to read all the three CSV files and use series, DataFrame, merge, and DateTime code inside Pandas to help better understand and visualize the results, as shown in Figure 1.

Another method in Python the author used is NLTK which is also called Natural Language Toolkit. Its primary purpose is to let Python programming language work with human language data [16]. As this paper mentioned above, the author planned to do sentiment analysis to figure out the most popular words in the review from the database and thus making NLTK the most suitable tool in this scenario. According to research, sentiment analysis using Python NLTK has been applied to normal business operations to improve performance. They use sentiment analysis to study the customers' behaviour and let their business improve in that direction.

The last method this paper uses in Python is matplotlib which is a plotting library in Python. It is useful when visualizing the results. By drawing the plots, this paper can better understand the comparison from multiple aspects.

3.2. Sentiment Analysis. Sentiment analysis can study the customers' behaviour to some extent and can somewhat replace the time- and money-consuming traditional methods like surveys and focus groups. By just analysing the easily accessible sources of data, sentiment analysis can give the business owners relatively accurate feedbacks to measure the customers' tendencies:

$$P(y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}, \tag{6}$$

$$W_c = \int_R \frac{\varphi^2(w)}{w} \, dw. \tag{7}$$

However, sentiment analysis can be hard to finish and can be easily compromised. To get accurate results, the author must eliminate all the nonvariables and dummy variables to make the dataset clean [17].

Becoming a super host on Airbnb seems an official quality assurance that will bring the hosts more clicks and thus more stays and more revenue. Besides the indirect benefits (increase in the revenue), Airbnb will provide the super hosts more direct benefits like travel coupons, exclusive events, and priority support, as shown in Figure 2.

The criteria to become a super host can be fussy [18]. It asks hosts to be almost perfect in all aspects including but not limited to review scores, response rate, and cancellation rate. As a result, the Airbnb super hosts can be seen as examples for all hosts to learn from [19].

Some traditional positive words like recommend and good cannot fully reflect the guests' thoughts. Many people with ok feelings who are not willing to leave a long talk may use recommend and nice words [20]. As a result, these words can be misleading. So, filtering these reviews before processing is necessary when dealing with the review data [21].

4. Analysis of Results

4.1. Basic Attributes of the Datasets. Before diving deep into the dataset, the author did a basic statistical analysis to calculate the number of unique listings, hosts, and basic perimeter (mean, median, and standard deviation).

Since there are many repeated listings in all these three CSV files, the author used the drop_duplicates() function to make the dataset clean. After that, the author used the len() function to calculate; there are 38481 unique listings and 22274 unique hosts. The mean, median, standard deviation, minimum, and maximum number of listings per host are shown in Figure 3.

As is shown in Figure 3, most hosts only have one listing and some may have more than one listing. The author divided the hosts into super and nonsuper hosts and calculated the same perimeter as above to better understand the difference between each group. The numbers are shown in Figure 4.
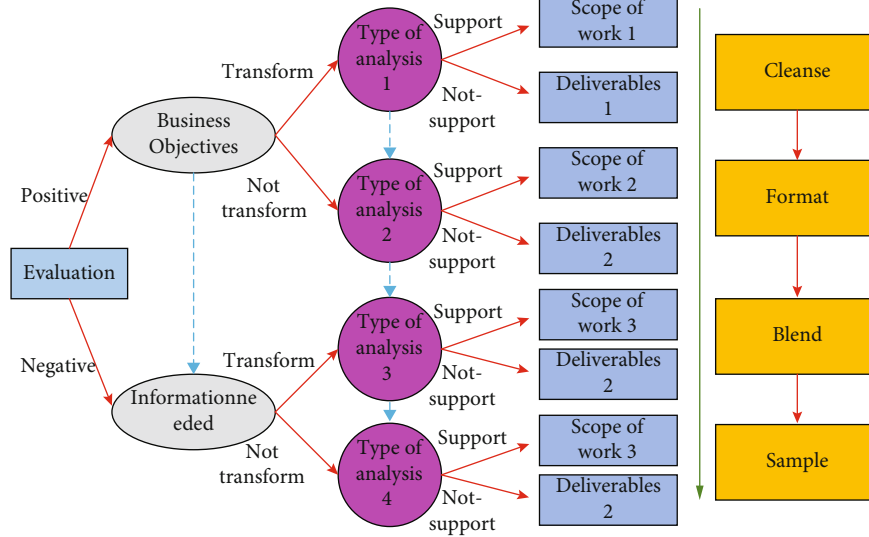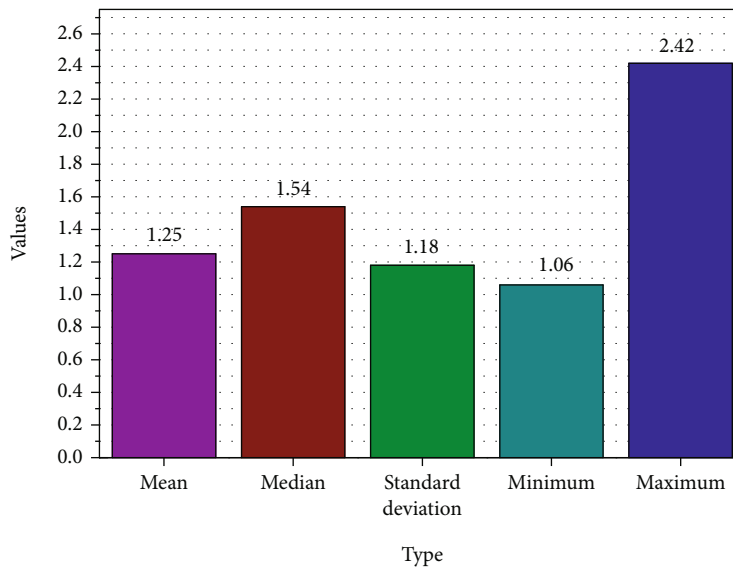
Figure 2: Data analysis steps.

Figure 3: Perimeter of number of listings per host.

Then, we want to analyse the datasets in two parts. The first part is to analyse some correlations among all the hosts, and the second part is to make some comparisons between super and nonsuper hosts.

First, the author wants to do a sentiment analysis of reviews among all the hosts. This paper intends to find the most 10 popular words in the reviews of the listings. After that, this paper can have a rough view of what the guests value when staying in an Airbnb, so this paper can let hosts improve towards that direction [22].

To generate the most accurate results, the author eliminate all the stopping words like punctuations and let the next popular words replace some words like "recommend" and "great" that cannot fully reflect the guests' opinion. This paper also made a histogram to visualize the results as shown in Figure 5.

Word "location" is ranked #3 in the most popular 10 words in the reviews. A good location is not just limited to close to attractions and landmarks or convenient transportation but depends on the purpose of the guests. Family travels may favour attractions nearby, but many other types of travelers like business or single travelers may favour quiet neighbourhoods, various selection of restaurants, safe community, etc. As a result, hosts should dig their properties' unique advantages as much as possible and add these words to the titles or descriptions of their listings to attract more guests.

Not surprisingly, "clean" is one of the most popular 10 words in the guests' reviews. Clean is the prerequisite for guests to perceive all other advantages that the hosts provided in properties. A clean property will not add scores but a slight lack of clean will reduce the scores largely.
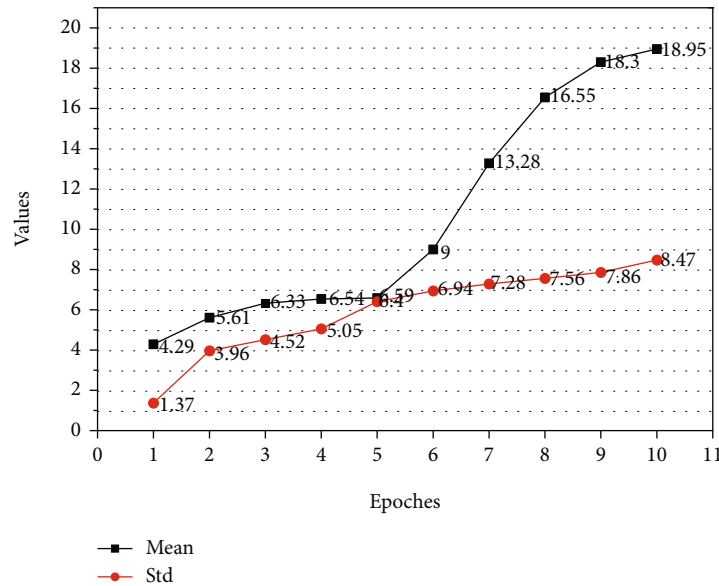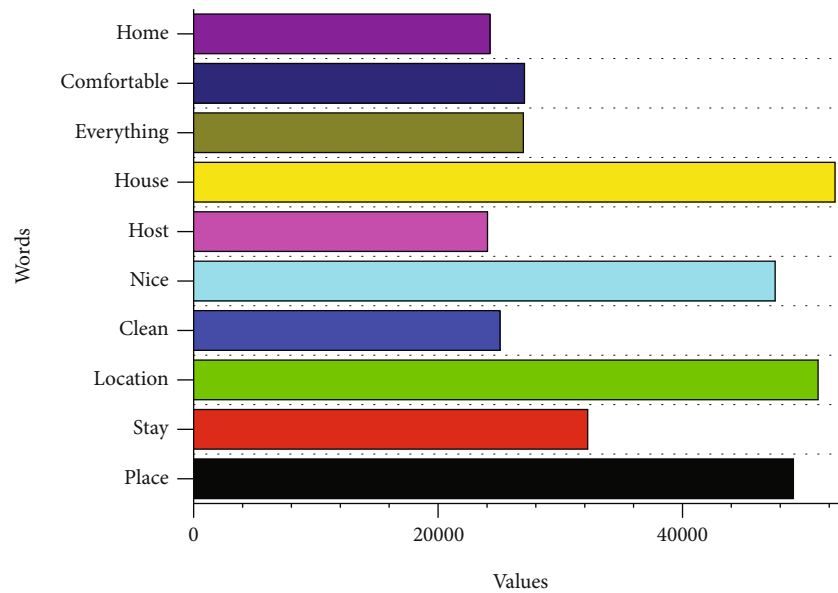
Figure 4: Number of listings per nonsuper hosts.

Figure 5: Most popular 10 words in reviews.

Another word "nice host" suggests hosts being polite, honest, and responsive when communicating with potential guests. Sometimes, using strategies like asking guests if they travel for special purposes, for example, anniversary, and preparing some small gifts for them would leave good impressions.

Since guests take "everything" as a criterion to measure properties, hosts should equip the properties with complete home appliances and living essentials. According to a report, some new Airbnb properties are now equipping gaming and cinema equipment like PlayStation, Xbox, and projectors to attract guests.

Second, the author proposed to find the correlation between the room popularity and the time (month of a year)

to suggest hosts and Airbnb system when to fully or part-time engage to operate their properties.

Since there is no such direct column that can imply room popularity, the author planned to count the room_type column in the listing. CSV is an indirect indicator of room popularity. There are four room types in this dataset which are the entire home/apt, private room, shared room, and hotel room. To get corresponding dates, the author merged listing and calendar DataFrame based on listing_id. Besides, an inner merge is appropriate in this case because it can eliminate listings that do not have corresponding dates. A line chart shown in Figure 6 has been made to help understand results.
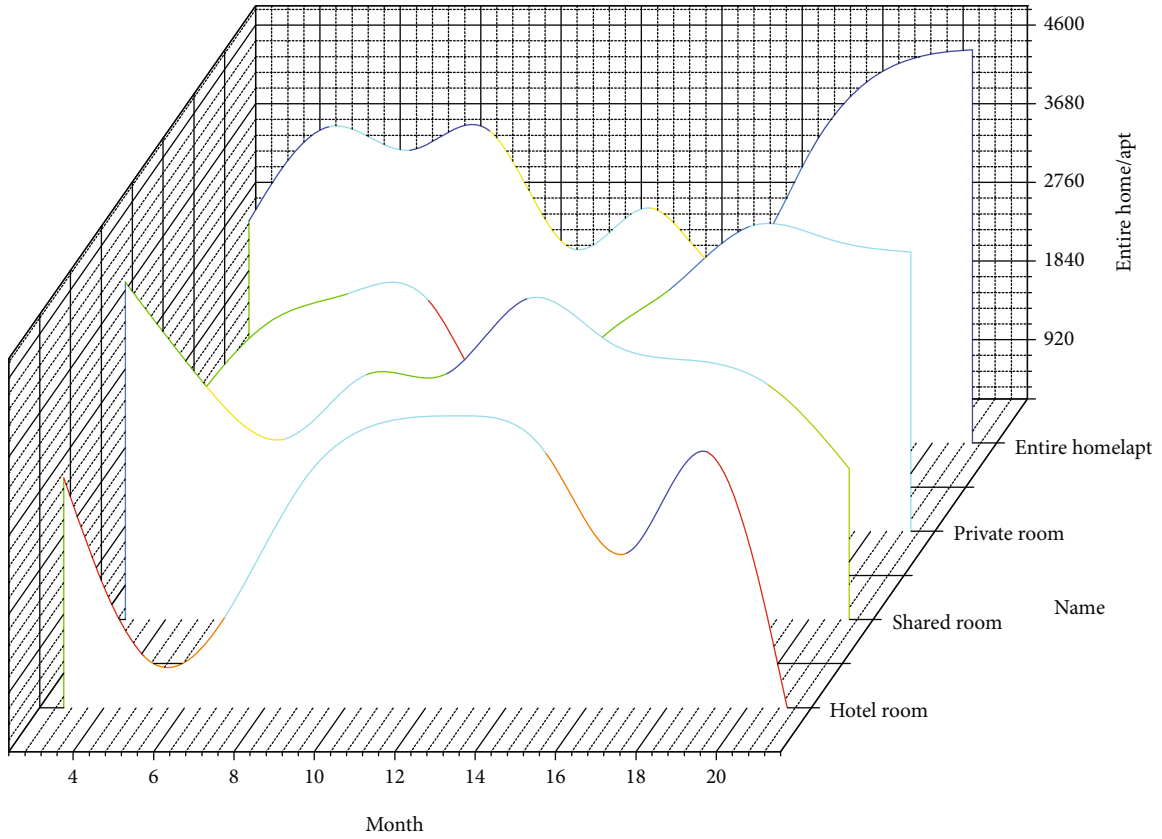
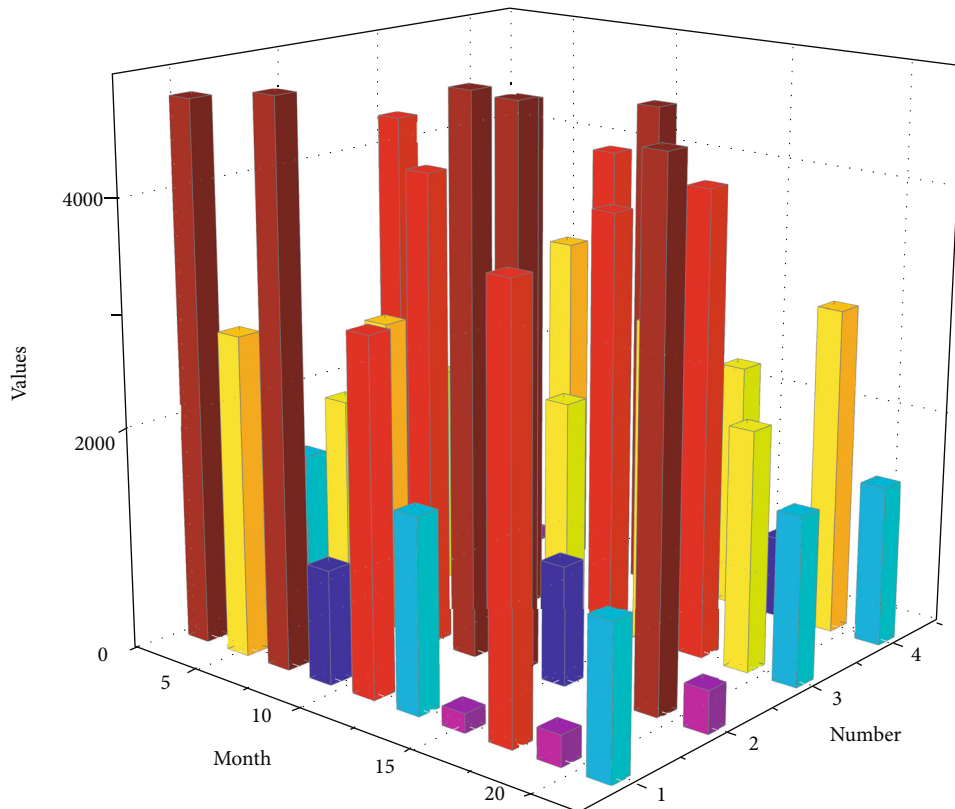Figure 6: Room popularity correlated with a month.

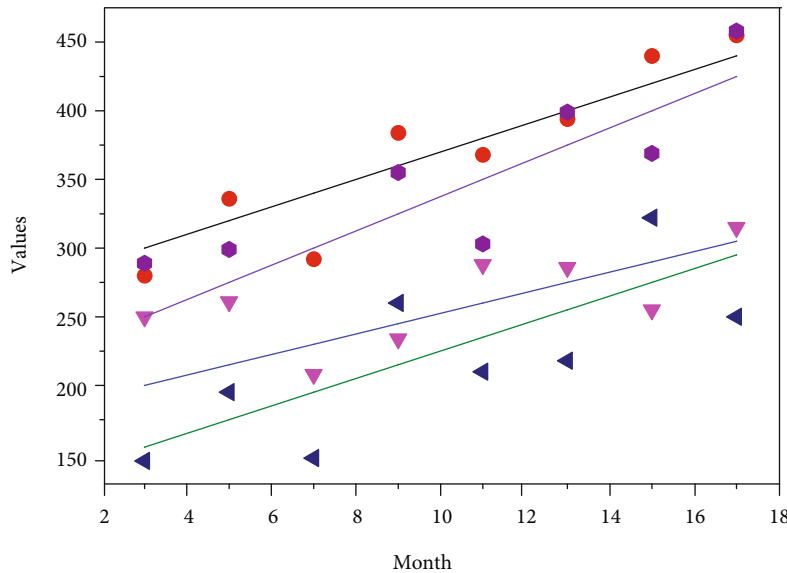Figure 7: Average monthly price correlated with the month.

Figure 8: Room price correlation with room type.

The difficulty of big data processing lies not only in the huge overall quantity of data but also in the fact that each data contains more features, i.e., the data has a high number of dimensions. This makes it difficult to analyse and understand the data intuitively, and then, the choice of data processing tools or algorithms is also blind, which affects the final data analysis results and reduces work efficiency. KL scatter is a measure of the difference between two different probability distributions p and g, also known as relative entropy. KL scatter calculates not only the spatial difference between two distributions but also the information loss of one distribution compared to the other. To data visualization, the loss of information after mapping data from a high-dimensional space to a low-dimensional space should be as small as possible.

SOM mimics the competitive learning mechanism in the biological brain's nervous system and is therefore also called a self-organizing competitive network. When the brain is stimulated by some input signals, a - number of neurons will start to excite. If these inputs are similar, then they stimulate the same neurons, while other neurons are inhibited by the excited neurons. This process is a competition between different neurons for the opportunity to be stimulated by the input signal. Corresponding to the clustering algorithm, the neurons can be considered clustering centres and the input signals as data.

When designing a new database, the researchers should not only carefully study the business requirements but also examine the existing systems. Most database projects are not built from scratch; usually, there will always be existing systems in the organization that meet specific needs (and may not have automatic calculations). Obviously, the existing system is not perfect; otherwise, the researchers do not have to build a new system. But research on the old system can reveal some subtle issues that you might overlook. It is good for researchers to look at existing systems.
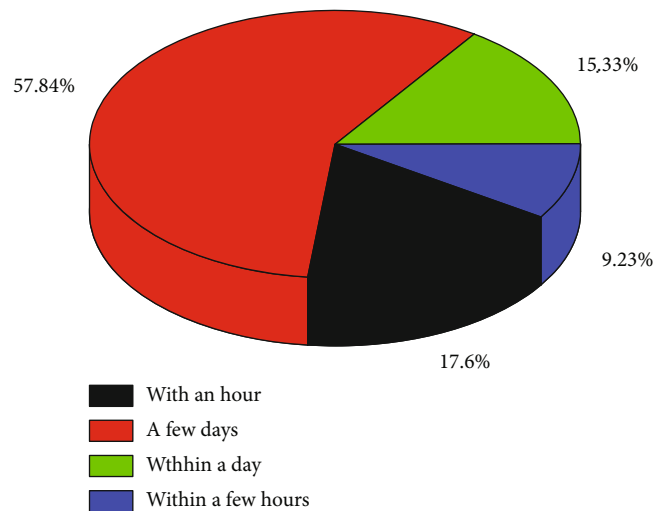

Figure 9: Average response time of super hosts.

From the graph above, the author can conclude that all the four types of rooms except hotel rooms (quantity is too small to generate statistical evidence) correlate with time (month of a year). The number of these three types of rooms being booked goes through a large boost around July and a slight increase around November and December. This meets the traditional traveling season which is June to August and before Christmas.

The trend shown in the graph suggests that the Airbnb hosts are well-prepared during the traveling season and Airbnb can advertise more to occupy market shares belonging to hotels.

Last, the author planned to test if there is a relationship between the popularity of each room type and their prices. This can imply whether the host should increase or decrease prices during season/off-season to increase property popularity.

Verified

UnVerified

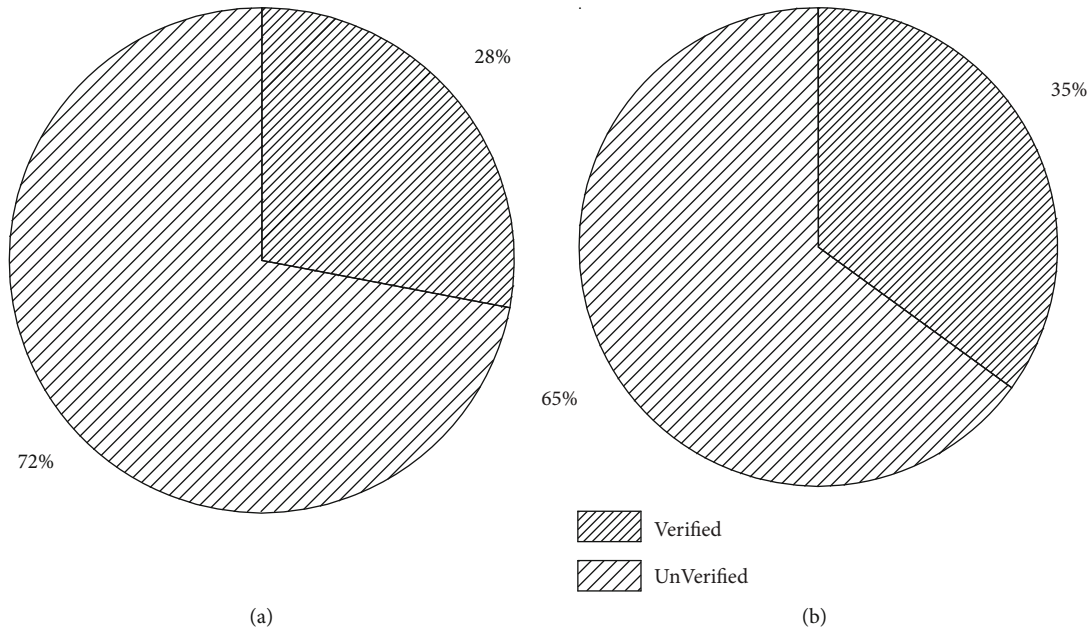(a)   (b)

Figure 10: Identity-verified distribution between super and nonsuper hosts.

To get results, this paper first calculated the average monthly price for each room type and visualized it with a line chart shown in Figure 7. Then, the author can use a linear regression model to test relations.

With each room type's average monthly price, the author can now generate regression analysis between price and popularity as shown in Figure 8.

From the graph above, the author can hardly conclude that the room popularity is correlated with its prices for all three types of room except hotel room (uncommon on Airbnb as quantity is too small to generate statistical evidence). As a result, an increase or decrease in price cannot generate more popularity, so Airbnb hosts should maintain their prices relatively stable during season/off-season to compete with the hotel industry.

4.2. Comparisons between Super and Nonsuper Host Patterns. As this paper mentioned above, becoming a super host can bring a lot of direct and indirect benefits. According to the Airbnb policies, a host needs to maintain a 90% or higher response rate, 1% or lower cancellation rate, and 4.8 or higher total review scores, as shown in Figure 9.

The author planned to divide all the hosts into nonsuper and super hosts and compare the same features between these two groups. After the author finds some specific patterns among super hosts, this paper will make suggestions for nonsuper hosts to learn from gaining advantages and super host to maintain its advantages.

To find the specific patterns of the super host, the author first examined the average response time of super and nonsuper hosts.

There are four types of response time in this dataset which are "within an hour," "within few hours," "within a day," and "within a few days." The author used SQL language in Python to first filter super hosts from nonsuper hosts then the author group by these four different response

times and finally count the number of each type of response time. A pie chart with the percentage shown in Figure 10 on each section has been made to visualize the results.

As is shown on the pie charts, super hosts are more likely to respond to the messages within an hour. Many potential guests may turn to other options when waiting for hosts' long-time responses which may lead to loss of potential customers. To conclude, nonsuper hosts should increase their communication efficiency to decrease the response time to gain some advantages.

Second, the author proposed to compare identity-verified situations between super and nonsuper hosts. Verified identity is an optional procedure for hosts. Hosts will get badges that say "identified verified" on their homepage if they upload their government-issued photo ID to the Airbnb system. Figure 10 shows the results of the comparison.

As shown from the graphs, super hosts are more likely to get their identity verified. According to research, guests say if all others stay the same; they prefer identity-verified owner's properties because that badge gives a sense of safety. So, hosts need to get their identity verified to gain advantages if Airbnb can stick to the user privacy policies. From the two evaluation criteria of deviation and standard deviation, each of the three models has its advantages and disadvantages. The metamodel without introducing time t performs best in terms of deviation, and the metamodel with introducing time t performs best in terms of standard deviation, but the differences between the three methods in terms of deviation and standard deviation are not large. In terms of martingale distance, the effect of the metamodel with the introduction of time t is much better than that of the metamodel without the introduction of time t.

The effect of the model with the introduction of time t and $t2$ is better than that of the metamodel with the introduction of time t, and the difference between the three models is larger. Therefore, the author believes that
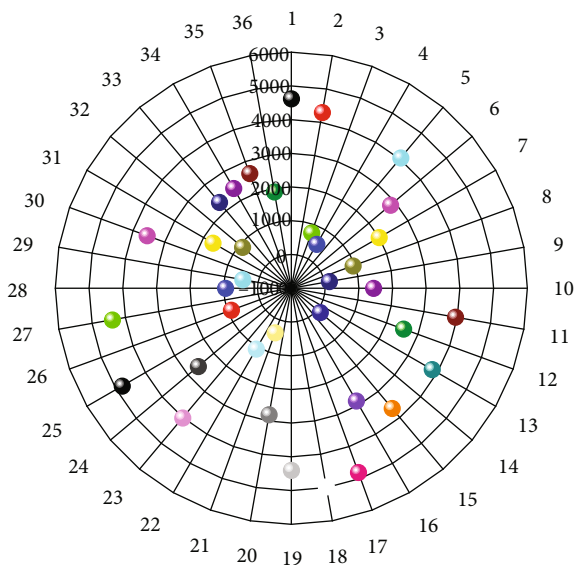
FIGURE 11: Comparison of mean scores of super/nonsuper hosts.

introducing time t and t to build a multivariate metaregression model is better than building a multivariate meta-analysis model in combining the parameter estimates. In summary, when it comes to the need to split the data according to features, and the features have an impact on the parameter estimates of the variables, the quantitative relationship between the features and each parameter estimate can be obtained through the multivariate metaregression model, obtaining a more accurate result of combining the parameter estimates, which is more reasonable and effective than the multivariate meta-analysis model. The basic attributes including the mean value were calculated before, and the median values were 115.2 and 25.1, respectively.

Last, the author intends to compare different review scores between super and nonsuper hosts. There are six different types of scores including accuracy, cleanliness, check-in, communication, location, and value. The guests will submit their evaluations after check-out from hosts, and a large part of an evaluation is to rate the properties from various aspects. These types of scores range from 1 to 10 where 10 is the best and 1 is awful.

To compare, the author calculated the mean of each type of score for super and nonsuper hosts and make all six scores in one histogram as shown in Figure 11 to better understand the differences.

The numbers on the horizontal axis are the mean review scores, and each paired column shown in the histogram is one of six types of scores.

As shown in the histogram, super hosts exceed nonsuper hosts on all types of scores, but super hosts only gain large advantages on three types of scores which are accuracy, cleanliness, and value. Accuracy measures whether the real condition of the properties is compatible with the hosts' description. Hosts can weaken the properties' drawbacks and emphasize the advantages in descriptions, but they need to achieve a balance to avoid being evaluated as false descriptions by guests. Cleanliness is another point that nonsuper

hosts need to improve. Value scores represent the degree of compatibility between what the hosts paid and what they have. The value is most difficult to interpret because it varies from person to people's different perceptions. It is also hard to be defined to a specific single point but goes through the entire procedure; from communicating, booking, staying, and leaving, the value can be defined in every moment in an Airbnb transaction. So, nonsuper hosts should try to make everything perfect in every link.

5. Conclusion

Based on the evidence found through data analysis, the author can now make some suggestions to Airbnb hosts. The first part is for all hosts (super and nonsuper hosts). First, by studying the guests' reviews, the author's recommendations are trying to dig their unique advantages, taking cleanliness as the top priority, and equipping the properties as much as possible. Second, by studying the room popularity changing with time, this paper suggests that Airbnb hosts and themselves should be fully engaged around July and December. Last but not the least, Airbnb hosts should remain their property prices stable to compete with hotels. The second part suggestions are for nonsuper hosts to improve and super hosts to maintain. First, nonsuper hosts should reduce response time when communicating with potential guests. Second, nonsuper hosts should get their identities verified. Last, accuracy, cleanliness, and perceived value are the top three problems that nonsuper hosts need to improve.

Data Availability

The simulation experiment data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] R. Bion, R. Chang, and J. Goodman, "How R helps Airbnb make the most of its data," *The American Statistician*, vol. 72, no. 1, pp. 46–52, 2018.

[2] C. Chang and S. Li, "Study of price determinants of sharing economy-based accommodation services: evidence from Airbnb. com," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 4, pp. 584–601, 2021.

[3] D. Coyle, "Do-it-yourself digital: the production boundary, the productivity puzzle and economic welfare," *Economica*, vol. 86, no. 344, pp. 750–774, 2019.

[4] L. Garay, S. Morales, and J. Wilson, "Tweeting the right to the city: digital protest and resistance surrounding the Airbnb effect," *Scandinavian Journal of Hospitality and Tourism*, vol. 20, no. 3, pp. 246–267, 2020.

[5] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, "" Did you suspect the post would be removed?" Understanding user

reactions to content removals on Reddit," in *Proceedings of the ACM on human-computer interaction*, vol. 3, pp. 1–33, 2019.

[6] J. Liu, J. Park, K. Xie, H. Song, and W. Chen, "Effect of commercial neighbors on the online popularity of peer-to-peer accommodation-sharing properties," *Journal of Hospitality & Tourism Research*, vol. 44, no. 5, pp. 780–805, 2020.

[7] L. Luo, "Intelligent led lighting centralized control system based on internet of things," *Shenyang Gongye Daxue Xuebao/Journal of Shenyang University of Technology*, vol. 40, no. 4, pp. 453–458, 2018.

[8] S. Ma, H. Gu, D. P. Hampson, and Y. Wang, "Enhancing customer civility in the peer-to-peer economy: empirical evidence from the hospitality sector," *Journal of Business Ethics*, vol. 167, no. 1, pp. 77–95, 2020.

[9] M. M. Mariani and M. Borghi, "Online review helpfulness and firms' financial performance: an empirical study in a service industry," *International Journal of Electronic Commerce*, vol. 24, no. 4, pp. 421–449, 2020.

[10] A. C. Mermet, "Who is benefiting from Airbnb? Assessing the redistributive power of peer-to-peer short-term rentals," *The Professional Geographer*, vol. 73, no. 3, pp. 553–566, 2021.

[11] M. Zhang, F. Conti, H. Le Sourne et al., "A method for the direct assessment of ship collision damage and flooding risk in real conditions," *Ocean Engineering*, vol. 237, p. 109605, 2021.

[12] R. Peng, M. Zhang, C. Guo, J. Cui, and J. Song, "Public bicycle operating system based on space-time security and the internet of things," *Wuhan University Journal of Natural Sciences*, vol. 23, no. 6, pp. 541–548, 2018.

[13] S. Ruiz-Correa, I. Ruiz-Correa, C. Olmos-Carrillo et al., "Mi Casa es su Casa? Examining Airbnb hospitality exchange practices in a developing economy," *ACM Transactions on Social Computing*, vol. 2, no. 1, pp. 1–24, 2019.

[14] L. Rupprecht, J. C. Davis, C. Arnold, Y. Gur, and D. Bhagwat, "Improving reproducibility of data science pipelines through transparent provenance capture," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3354–3368, 2020.

[15] S. Xiao, S.-H. Ma, L. Guo, and S. K. Mukhopadhyay, "European option pricing with a fast Fourier transform algorithm for big data analysis," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 3, pp. 1219–1231, 2016.

[16] A. A. Sinha, S. Rajendran, R. P. Nazareth, W. Lee, and S. Ullah, "Improving the service quality of telecommunication companies using online customer and employee review analysis," *Quality Management Journal*, vol. 27, no. 4, pp. 182–199, 2020.

[17] G. Williams, M. Tushev, F. Ebrahimi, and A. Mahmoud, "Modeling user concerns in sharing economy: the case of food delivery apps," *Automated Software Engineering*, vol. 27, no. 3-4, pp. 229–263, 2020.

[18] Y. Yao, J. R. Basdeo, O. R. Mcdonough, and Y. Wang, "Privacy perceptions and designs of bystanders in smart homes," in *Proceedings of the ACM on Human-Computer Interaction*, vol. 3(CSCW), pp. 1–24, New York, NY, USA, 2019.

[19] M. Yaşlıoğlu, "Dynamics of the network economy: a content analysis of the search engine trends and correlate results using word clusters," *International Journal of Data Analysis Techniques and Strategies*, vol. 12, no. 1, pp. 1–29, 2020.

[20] J. Zhang, "Listening to the consumer: exploring review topics on Airbnb and their impact on listing performance," *Journal of Marketing Theory and Practice*, vol. 27, no. 4, pp. 371–389, 2019.

[21] Z. Yong-zhang, P.-x. Lwe, W. A. N. G. Shu-gong, X. I. A. O. Fan, L. Jing-zhe, and G. A. O. Le, "Research progress on big data and intelligent modelling of mineral deposits," *Bulletin of Mineralogy, Petrology and Geochemistry*, vol. 36, no. 2, pp. 327–331, 2017.

[22] Z. Xie, J. Wang, and L. Miao, "Big data and emerging market firms' innovation in an open economy: the diversification strategy perspective," *Technological Forecasting and Social Change*, vol. 173, article 121091, 2021.