Airbnb Data Analysis Report

Lei Wang

04-03-2019

Introduction

Inside Airbnb provides tools and public data that allows anyone to explore how Airbnb is really being used in cities around the world. In my analysis, I mainly focus on the city of San Francisco. There are three main topics I want to analyze and discuss, which are 1. Price 2. Review 3. Availability

The three topics are closely related to if customers are willing to rent house from listings. Reasonable price, good review and certain availability rate would attract more customers to use Airbnb so that Airbnb will make more revenue. Here are several specific questions that I want to address through later analysis.

1. How does price change over the neighborhoods of San Francisco?
2. What factors are going to affect price?
3. What are reasons that host got lower review?
4. How is the availability rate over the neighborhoods of San Francisco?
5. How is the availability rate for the next 365 days?

Data selection and processing

Data is available through this link, http://insideairbnb.com/get-the-data.html .There are three main data sets available. For each data set, I primarily selected certain number of columns which are useful for my analysis.

1. Detailed Listings data for San Francisco, including. Columns include id, host response time, host response rate, host is super host, host identity verified, neighborhood cleansed, property type, room type, minimum nights, availability 365, number of reviews, last review, review scores rating, cancellation policy, reviews per month, calculated host listings count, price.
2. Detailed Calendar Data for listings in San Francisco, Columns include listing id, date, available.
3. Detailed Review Data for listings in San Francisco, Columns include listing id, comments.

Correlation matrix for continuous variables analysis.

The Matrix (Figure 1) gives me a preliminary idea about the possible correlation between each pair of variables. The target variables are price and review. It seems that there are no factors that have strong correlation with price or review. Therefore, in later analysis, I might need to explore it more with categorical variables.

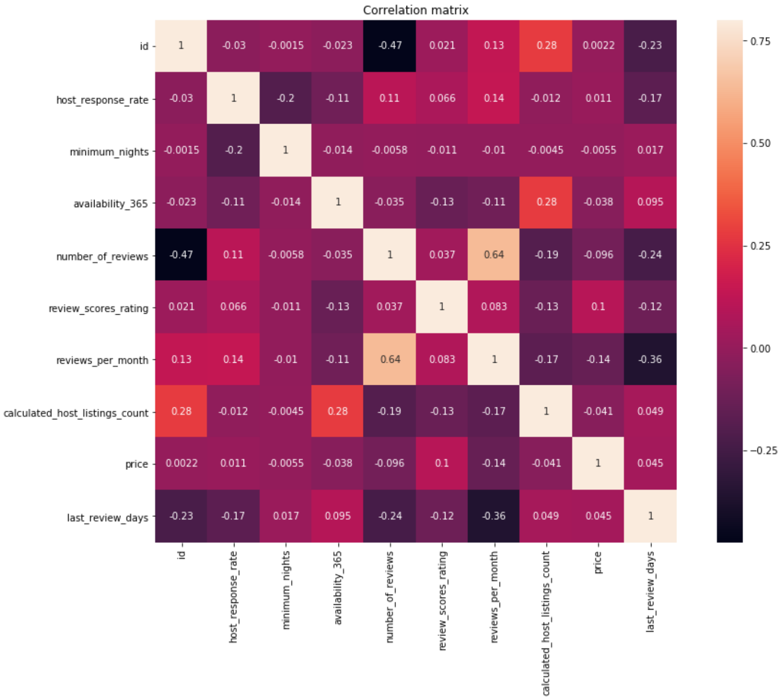


Figure 1. Correlation Matrix

Part I Price-related analysis

(1) Price Distribution over neighborhood

The graph of average price over neighborhood (Left panel of Figure 2) shows that the more expensive area is along the coast or near Golden Gate Park. The second expensive area is in the middle of downtown. The other areas have relatively lower average price. It proves that location plays an important role on listing price. I also created a graph for the number of listings over neighborhood (Right panel of Figure 2) as comparison. It shows that the middle of down has the most rent listings. However, the area along the coast and part have no much listings. The house/apartment along the coast or park might have good view, nice and spacious accommodation for customers. In addition, because of the rare listings available, the price should be higher. In downtown, more listings and crowded rooms do not allow the high price high even if the location is very convenient.

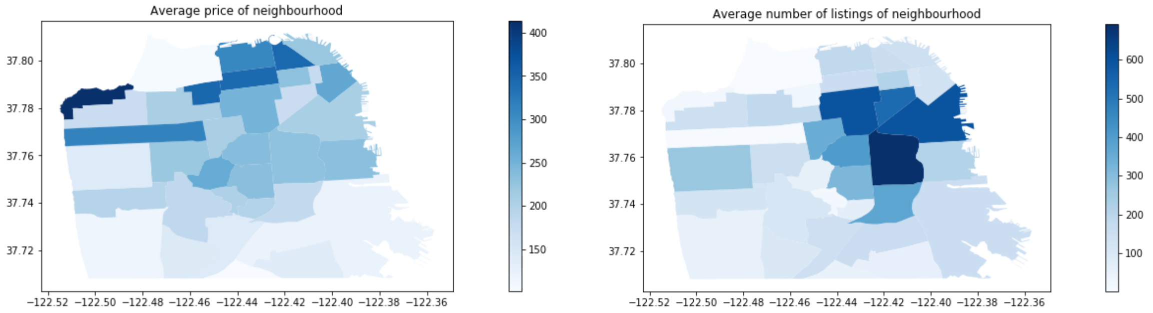


Figure 2. Average price and number of listings distribution over neighborhood

(2) Room type vs Price

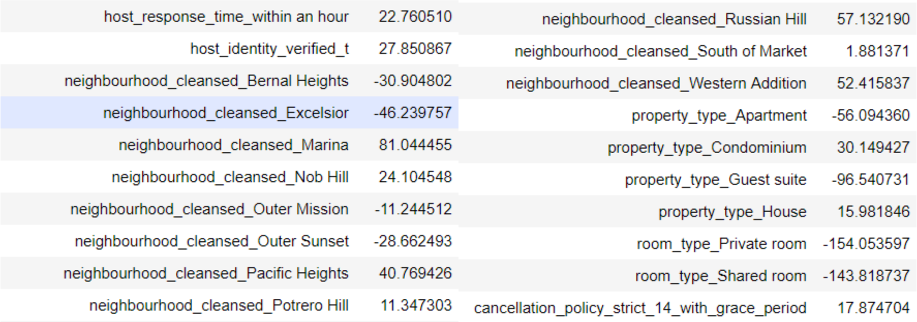
The second factor that affects price is the room type (Table 1). Obviously, the entire house/apartment has the highest price, private room is second high, shared room has the lowest price.

Table 1. Room Type vs number of listing and average price

Room Type	The number of listings	Average price
Entire home/apt	4366	267.96
Private room	2606	132.03
Shared room	179	77.49

(3) Other factors based on regression with Lasso

There are still other factors that might have influence on price. I build multiple linear regression with Lasso to select the influential variables. According to the coefficients (Figure 3), room type is the most influential factor to the price. Property type also has some influence, house or condo has higher price while apartment or guest suit has lower price. Another important factor is neighborhood, which is similar as what I discussed above. In addition, host response rate within an hour, identified host, cancellation policy have a certain degree of influence on price. They might be just correlated to price but not the reasons why price is high. In conclusion, hosts can set a more reasonable price based on these features.



host_response_time_within an hour	22.760510	neighbourhood_cleansed_Russian Hill	57.132190
host_identity_verified_t	27.850867	neighbourhood_cleansed_South of Market	1.881371
neighbourhood_cleansed_Bernal Heights	-30.904802	neighbourhood_cleansed_Western Addition	52.415837
neighbourhood_cleansed_Excelsior	-46.239757	property_type_Apartment	-56.094360
neighbourhood_cleansed_Marina	81.044455	property_type_Condominium	30.149427
neighbourhood_cleansed_Nob Hill	24.104548	property_type_Guest suite	-96.540731
neighbourhood_cleansed_Outer Mission	-11.244512	property_type_House	15.981846
neighbourhood_cleansed_Outer Sunset	-28.662493	room_type_Private room	-154.053597
neighbourhood_cleansed_Pacific Heights	40.769426	room_type_Shared room	-143.818737
neighbourhood_cleansed_Potrero Hill	11.347303	cancellation_policy_strict_14_with_grace_period	17.874704

Figure 3. The coefficients of variables

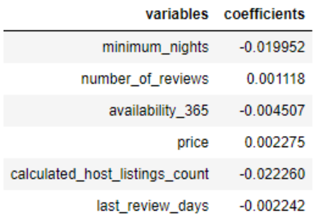
Part 2 Review analysis

(1) Review rating summary statistics

The summary statistics shows that the review rating performs very well. The mean value is 95.55, 70% of ratings are 100. In this case, it is hard to use other factors to differentiate the rating.

(2) Review rating regression result

The regression with lasso (Figure 4) proves the analysis above. Only 6 variables are selected by model. However, the coefficients are very small, which give no big impact to ratings.



variables	coefficients
minimum_nights	-0.019952
number_of_reviews	0.001118
availability_365	-0.004507
price	0.002275
calculated_host_listings_count	-0.022260
last_review_days	-0.002242

Figure 4. The coefficients of variables

(3) Word cloud

Since the review rating is too good. I decide to explore something from the comments. I select comments from review rating less than 90 and further select the negative words. The word cloud (Figure 5) gives me some insight about bad reviews. It shows that noise and dirty are the biggest problems.

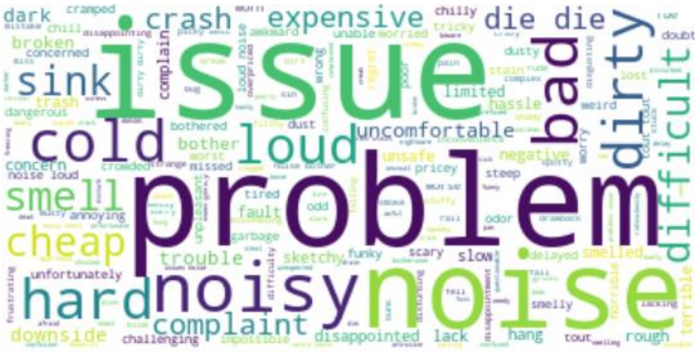


Figure 5. Word cloud of comments

Part 3 Availability analysis

(1) Average availability over neighborhood

The future 365 days of availability over neighborhood shows that the sources in middle town has lower availability even if it has more listings. The comparison with price over neighborhood and the number of listings over neighborhood would give us insight about which area has high availability, good price, which would be a useful method to give customers recommendation on house selection.

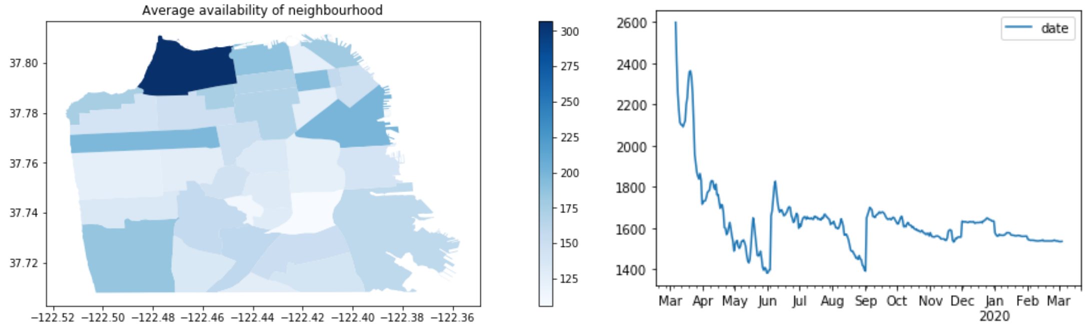


Figure 6. Availability distribution over neighborhood (left) and Unavailability for the next 365 days (right).

(2) Unavailability over time

It makes sense that the unavailability in March and April is very high because customers generally book place not very early. However, the other time gives us an idea of the availability trend. Start from Jun, order demand starts to increase and reach a peak order time quickly. It is a good time for hosts to have good business. Meanwhile, it is a good time that Airbnb to attract more hosts to join in because there is more demand during this time. It starts to reduce in August and get back on September. December is another peak time. This trend is also very useful to adjust price. If there is very low availability in a certain period, hosts can adjust price higher.

(3) Term vs Availability

Third, listings have minimum nights requirements. Some listings are only for long term rent while some are for short term rent. In table 3, it is obvious that long term rent has more days available. However, in peak time, it has better for Airbnb to motivate hosts to rent their room with short term contract. It will increase the total revenue of Airbnb.

Table 3. Rent term vs Availability

Minimum night level	Average availability
Long term (>=30 days)	214.08
Short term (<=30 days)	99.68

Conclusions

1. Hosts can refer to location, room type and availability to set price in general. But it is possible to build machine learning model to give host daily or weekly price suggestion with more factors.
2. The host in Airbnb performs so well on review part. Because hosts are always willing to provide good service to get good review so that more customers will come. What the customers most care about is cleanness and quiet. Airbnb should give hosts a standard to maintain cleanness. If some hosts can not reach the standard, Airbnb can take action to change it.
3. Availability is an effective real-time monitor. It affects the price, customers' choice, Airbnb's marketing strategy. It is important for Airbnb to attract more hosts, give reasonable recommendation to hosts and customers. More customers, more rooms rent out, more revenue for Airbnb.