Airbnb Price Per Night Prediction in Paris, France

Peter Mai
University of California, San Diego
December 2, 2017

ABSTRACT
With so many people connected online, it has never been easier for people to access crowd sharing resources online. Airbnb is one of those services, allowing everyday people to provide short-leases on their home to practically anyone in the world. As people tend to get a lot more space and amenities renting out an Airbnb than a traditional hotel, it's a no brainer these types of crowd sharing services are picking up in popularity faster than no other. However, with home owners in charge of deciding the prices of their lease, rather than a huge monopolistic company controlling the prices everyone pays, is there a reason to believe that there is a trend involved in how prices are determined or is it pure random? This article will attempt to explore this question by building a supervised machine learning predictive model for Airbnb listing prices through analyzing tens of thousands of Airbnb listing data gathered throughout Paris, France.

1. DATASET
1.1 Choosing the Dataset
The dataset we are using is provided by "Inside Airbnb", a non-commercial, independent site, whose pure purpose is to provide clean data on all the Airbnb listings/reviews from each of the main metropolitan cities in which Airbnb operates in. The neat thing about this dataset is that we get all the public information as provided on the official Airbnb website, regarding each and every single listing/review, all properly stored in CSV format. This is a huge benefit as it significantly reduced the time needed for us to crawl and clean the data ourselves.

Out of the 44 metropolitan cities provided by "Inside Airbnb", we chose to do our research on the city of Paris, France. With Paris being such a popular tourist destination, it became the city with the most Airbnb listing in the world. This was crucial as the more data we have, the easier it is for us to train our model without overfitting on a small subgroup of information. As a result, we ended up with a dataset of 56,535 unique Airbnb listings located in Paris, France, collected on April 4, 2017.

1.2 Dataset Description
As "Inside Airbnb" crawls for all the information provided on the Airbnb website, it was able to provide us with 95 unique features regarding each Airbnb listing. The feature list included various information ranging from descriptions/rules of the place provided by the host, location, amenities offered, average review score of the place, links to pictures of the house, all the way to the host's resident neighborhood. All of the 95 features are provided at the end of this report.

We can quickly see that this dataset included a mixture of qualitative, quantitative, and repeated information. The qualitative data included the summary and descriptions the host provided. Some of the text were given here in English, while others were given in French. This may pose a challenge later on when we decide to do text analysis to seek further insight, as the corpus we are analyzing has a mixture of two different languages. The data also has an abundant amount of quantitative information such as average review for the place in different categories, number of bedrooms/restrooms, owner's pricing for each night/week/month, number of guests allowed, longitude/latitude, etc. This information allowed for quick analysis on the data set without much preprocessing. Then we also had a lot of redundant information. For example, the dataset included street names on which the housing is located, but it also provided the city, zip code, and general neighborhood name. All of this information can be derived from just the street name. The description and the summary provided are also usually duplicates of each other. This meant that we can represent quite a few of these features by condensing them into one.

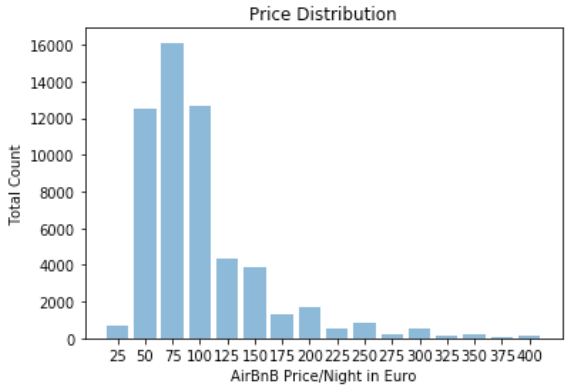
1.3 Exploratory Analysis
To get a sense of the type of data we are working with and maybe spot a few trends, we decided to go by intuition and graph a few features that may be correlated to one another.
We began by analyzing the price per night distribution across the entire Airbnb listings in Paris. The results are seen below:

Price Range	€ 0.00 to € 7,790.00
Mean	€ 96.12
Median	€ 75.00
Standard Deviation	€ 99.30

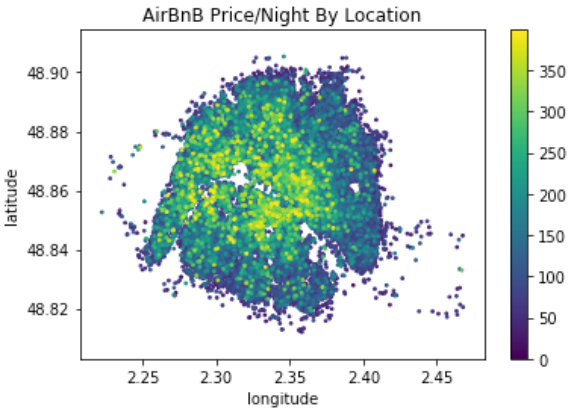
Based on these findings, we can see that the majority of the pricings were located around €75, however, our data was significantly skewed by a few large outliers. To try to make our data a bit nicer to analyze, we removed the 1% highest pricing from the dataset. This removed 629 listings from the total 56,535 total listings, giving us a remaining total of 55,906 listings. The new findings are shown below:

Price Range	€ 0.00 to € 399.00
Mean	€ 89.87
Median	€ 75.00
Standard Deviation	€ 56.23

Analyzing the statistical significant of these data, we can see that we now have a much nicer dataset to work with. A graph of the pricing distribution of Airbnb listings after removing the 1% highest pricing can be found on the next page.

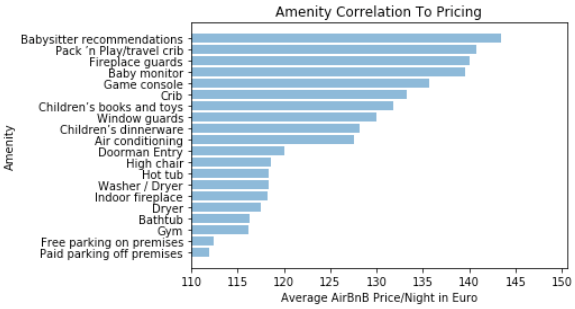


Locations often have a big impact on the pricing of the housing market. To get a better sense if this was indeed the case, we graph a map of the data along with their pricing as seen below.



From the graph above, Airbnb listings that are closer to the center of Paris will often have listings with higher prices. However, in those same place, there also exist listings with lower prices. This meant that though location do have some role in influencing price, using location alone cannot give us the full picture.

Let's take into account another feature that may influence the price: amenities. The dataset has 67 different unique amenities that an Airbnb host can provide for their clients. To do this, we calculated the average Airbnb price per night for each of the amenity listed. The graph is shown below.



From the data above, we can see that there is indeed a correlation between different amenities and the pricing of the Airbnb. Host that offer those more premium amenities such as babysitter recommendation, travel crib, fireplace guard, baby monitors, game consoles, etc. tend to charge higher for their Airbnb homes.

Though skeptical due to the multiple languages, we also graphed the correlation between different words used in the description of the Airbnb with its pricing. This was done by taking the top 2000 most used words and graphing the 20 words with the highest average pricing as seen below:

We see that words related to unique, beautiful, and luxury tend to have a higher average pricing. However, the word that have the highest average pricing was "restrooms". This could mean that a lot of hosts had this in their descriptions and Airbnb with higher pricings may skew this average up, resulting in the higher average score for "restrooms".

2. PREDICTIVE TASK

2.1 Identify Predictive Task

From our data exploratory analysis, we can see that there existed a correlation between different Airbnb listing features and its pricing. Therefore, our goal was to build a predictive model that would determine the estimated pricing of an Airbnb listing given various information on the listing.

To do this, we first shuffled our data as the original dataset was ordered by geographic locations. Then, we divided our data into three groups; training, validation, and testing. The first 80% of the dataset would be the training dataset, the next 10% would the validation dataset, and the remaining 10% would be the testing set. We would use the validation set to tune our parameters of our model, and finally perform predictions on the testing set with our tuned model.

2.2 Baseline and Validity Assessment

From the price distribution graph, we can see that the majority of the Airbnb listing prices tend to be around €75.00. A simple baseline could be a model that always predict the median for every Airbnb listing, or €75.00.

However, the price distribution graph is slightly skewed to the right. We can add a second baseline that could account for this right skew by always predict the mean Airbnb listing price, or roughly €89.87.

We will assess the performance of our model by calculating the coefficient of determination (R^2) and mean squared error (MSE) score of our model.

R^2 gives us a feel for how well our feature model correlated to the actual expected value. A score closer to 1 meant our model can replicate/explain the output fairly well, while a score closer to 0 meant that our model does a very poor job at explaining the variability of the output. We would be using sklearn's implementation of R^2, which may result in a negative number if the model is worse than a regression model that is a horizontal line.

On the other hand, MSE gave us an average score of the squared errors of all our results. We want an MSE that is closer to zero, indicating that we have a smaller error range. The equation for calculating MSE is

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(prediction - expected)^2 .$$

To assess the performance of our model, we will ensure that our model's R^2 and MSE scores yield a much more desirable number than those given by the two baseline models, always predicting median and mean.

Baseline 1: Always Predicting Median (€75.00) – Test Set

R^2	-0.0723373329977
MSE	3427.7703452

Baseline 2: Always Predicting Mean (€89.87) – Testing Set

R^2	-0.00003508451473
MSE	3196.65323716

2.3 Pre-Processing

The dataset does have some missing fields. However, most of these missing fields were related to the qualitative side such as missing notes, host about information, and tips provided by the owner. Some quantitative information that were missing were host average response time, and if they provided weekly/monthly rent prices. To ensure the consistency of our data, missing text were left as empty strings, and missing quantitative information were filled in with zeros. Furthermore, we tried to avoid using features that were not present in every single listing data, as these data fields were often optional and could be described by other features.

2.4 Features and Justifications

For the first half of the feature set, we included pure quantitative data as these don't require any preprocessing to attain. These included latitude, longitude, number of people housing accommodates, number of bathrooms/bedrooms/ beds, number of guests included, minimum nights required, if host's identity was verified, and each of the Airbnb's average review score for rating, accuracy, cleanliness, check-in, communication, and location. These features were chosen because these were the basic information every user took into consideration when they are booking an Airbnb to stay. These information are also the "featured" information that was highlighted on every Airbnb listing web page.

Then we added on other features a user always considered but are not exactly quantitative measurements. This included which of the 23 property types (e.g. Apartment, Loft, Boat, etc.) and which of the 3 room types (Entire home/apt, private room, or shared room) the Airbnb is. These information were also highlighted on the Airbnb websites. However, since they were not quantitative data, we had to pre-processed the data by representing each of the 23 property types as a one hot vector of length 23, and each of the room type as a one hot vector of length 3.

Since our data exploratory analysis indicated that location of the Airbnb may have an impact on the pricing of the rent, we went on to represent each of the 960 street names as a one hot vector as well. This gave us a more in depth understanding of where the Airbnb is located compared to using regular longitude and latitude, as neighboring streets may vary drastically in price points, while maintaining similar longitude and latitude readings. If the Airbnb location has an impact on pricing, our intuition gave us a sense that the host's resident neighborhood may have an impact as well. Therefore, we also included each of the 235 host's resident neighborhoods as a one hot vector feature as well.

Then we moved on to indicating which Airbnb listing provided which type of amenities. To do this, we saw that our dataset provided the amenities list of each Airbnb listing as a long string, with each amenity separated by a comma. To pre-process this information, we went through the entire dataset, parsing each of the string and keeping a list of all the unique different type of amenities offered. This yielded a list of 67 different amenities (e.g. Washer / Dryer, Gym, Shampoo, etc.). Each of the 67 amenities was then indexed into an array location. When generating the feature set for a listing, we would state either True or False for the given index if the Airbnb listing had the amenity correlated to that index. We also added on which of the five bed types (Real bed, pull-out sofa, couch, futon, or air bed) the Airbnb offered as a one hot vector

Furthermore, we noted earlier that we saw that some Airbnb descriptions that have certain words have a higher average Airbnb listing price. Therefore, we wanted to incorporate this data into our feature set as well. To do this, we went through the entire dataset to find the 2000 most popular words. Then we ranked each of the 2000 words by their average listing price. We then took the top 1000 words with the highest average listing price and represented it as an array of length 1000. For each listing, we represented the feature set as True or False for each array index based on if the listing's description had the word correlated to that array index in it.

3. MODEL

3.1 Type of Models and Justification

As we are trying to predict prices for certain Airbnb listings, the type of model we are looking for is a regression type of model. The most basic type of regression model we can choose from is linear regression. We will consider this type of model, but it might not perform ideally as the feature set

we have chosen have different variables that correlate to one another.

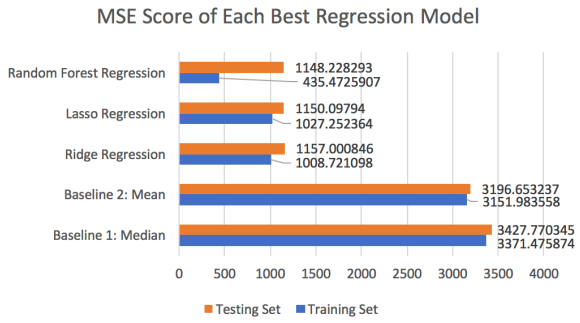
To tackle this issue of highly correlated independent variables, we would also use the ridge regression model to aid us in predicting the Airbnb listing price. We would like to also note that for each listing, we have a feature set of length 2,310. Since most of these features were 0 due to the nature of a one hot encoding, we ended up with a very sparse feature set. To tackle this second issue while holding onto the benefits provided by ridge regression, we would be adding on the lasso regression as our third regression model.

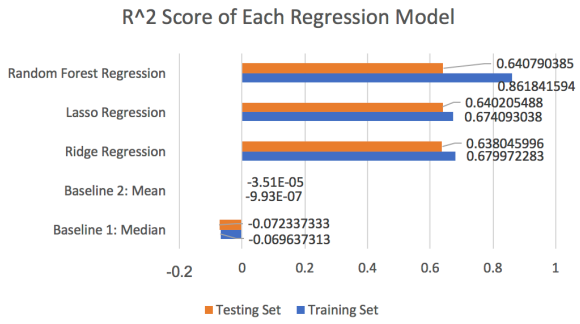
Finally, we will also try the random forest regression model as it used an ensemble technique, combining various models into one. This method usually reduced the variance and bias in our data resulting in a better prediction accuracy.

3.2 Optimization Techniques

Using the pre-processed scaled feature sets we came up with previously, we did our initial test using various combination of the feature set (e.g. excluding/including certain features) using the default parameter. We discovered that having all the features yielded the highest accuracy score. Then we began tuning the hyper-parameters of each model we trained, and finally choosing the model with the highest accuracy on the validation set.

For the ridge and lasso models, we tested with different alpha parameters (0.001, 0.01, 0.1, 1, 10, 100). For the random forest regression model, we tested a combination of max depth (1 to 20) and number of estimators (1 to 50) to find the best model on the validation set. The best model for each type of regression is described in the tables below:





Linear Regression

	Training Set	Testing Set
R^2	0.673598244	-1.06E+27
MSE	1028.811947	3.39508287814*10^30

Model Tuning

	Hyper-Parameters
Linear	Default
Ridge	Alpha = 100
Lasso	Alpha = 0.1
Random Forest	Max depth = 17, # Estimators = 50

3.3 Strength/Weaknesses of Each Model

We can see from the data that linear regression does a very poor job in predicting the expected price as our feature set have many correlated data. It over fitted on the training data, resulting in decent R^2 and MSE training score, but yielded a horrible R^2 and MSE score on the testing data.

Next up, we see that ridge, lasso, and random forest all beat the two baselines we have set up and have relatively similar testing scores. Lasso performed slightly better than ridge, and the model doesn't overfit too much on the training set. Random forest does slightly better than lasso on the testing set, however, we noticed that the random forest model tended to overfit a lot more as we increase the number of estimators and max depth. Therefore, the hyper-parameters chosen above attempted to strike a good balance between validation accuracy and overfitting for these models.

4. LITERATURE

Relating to other literature and works, Rady School of Management's academic paper on forecasting real estate prices [1] indicated that various economical features also impacted the price of the real estate. This included predictions on the average income of people in the area, economic/population growth, and monetary policies on the area. Since Paris was designed as a tourist destination, it was the prime market of economic boom in France. This correlated heavily with our findings on how location, such as those near the center of Paris, tended to have a higher Airbnb listing price.

On the other hand, Peng Liu Ph.D. from Cornell University School of Hotel Administration offered a bang for your buck economic approach to predicting housing prices [2]. Liu suggested that the right price is where the guest and the host strikes a balance value on what is offered. In other words, when people get what they value, they are willing to pay more for that purchase. This correlated with the idea that the different type of amenities offered by the Airbnb host carried a price tag with it. Certain amenities were considered luxuries and certain guests valuing those items were willing to pay more for it.

Visit Limsombunchai from Lincoln University performed a similar research in 2004, comparing housing price prediction accuracy between hedonic price model versus neural network [3]. His paper at the time aimed to prove the potential in neural networks capabilities as it outperformed

the hedonic price model, yielding an R^2 score of roughly 0.75. The feature set he included were land price, house's age/type, number of bathrooms/garage/bathrooms, and a list of different amenities. Though his paper was aimed at predicting housing prices, rather than crowd source home leasing prices, these two types of predictive tasks were very similar to each other. As a result, we can see that a lot of the features he used can transfer over to helping us predict the Airbnb listing prices as well. This backed up our finding that certain amenities offered have a high correlation with Airbnb listing prices.

5. RESULTS

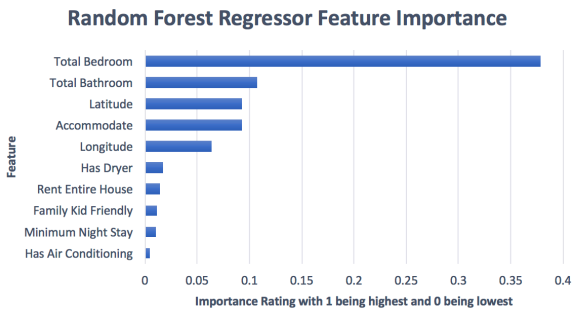
5.1 Model Predictive Abilities

After we trained our four models above, we saw that our models were able to predict Airbnb listing prices at a significantly higher accuracy as compared to using simpler statistical approaches such as predicting the prices as either the mean or the median of all the prices. We also learn that as our feature set included more correlated data, we should avoid using linear regression to train as this model tended to significantly overfit the training data and perform very poorly on unseen data. Ridge, lasso, and random forest regression tackled this issue of correlated feature variables and all resulted in roughly the same prediction accuracy on unseen data. Ridge and lasso tended to overfit less while random forest tended to overfit more as we increase the number of estimators and max depth.

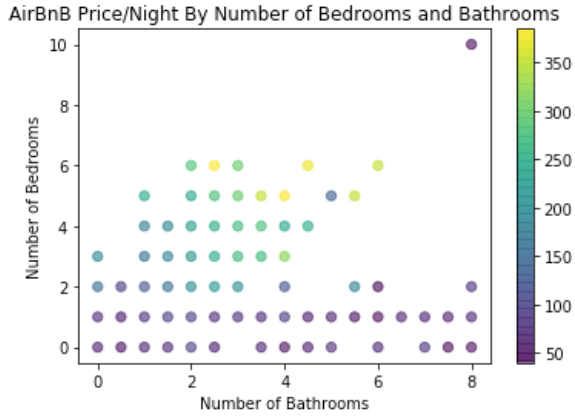
However, random forest does beat ridge and lasso slightly on predictive accuracy, yielding an R^2 score of 0.640790385272 and an MSE of 1148.22829271. Though these scores were lower than those found by Limsombunchai in predicting housing prices, we must keep in mind that Airbnb listing prices were determined by the host, rather than regulated by housing agencies. This gave more freedom for the host to pick varied pricing that fitted his/her needs and values. We can however conclude that with this ability to predict different Airbnb listing prices, there does exist different trends in how Airbnb hosts chose their listing prices.

5.2 Model Parameters Interpretation & Insight

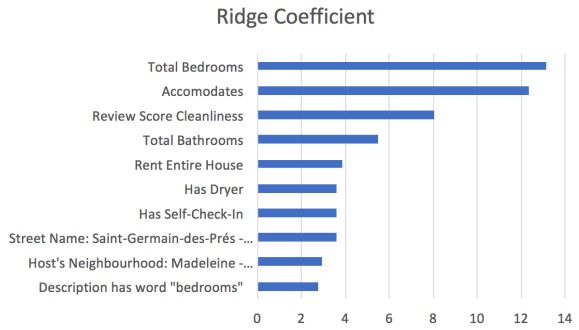
After training the random forest regression model, we were able to determine which feature had a higher importance weight when computing the listing prices. The results were taken from the train random forest sklearn model:

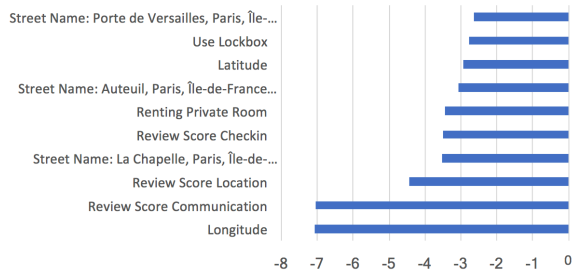


Drawing conclusion from the random forest model, we can see that the model picked up amenities and location as the main contributors to how it predicted the Airbnb listing prices. We see that the two most important features were the number of bedrooms and bathrooms offered by the place. The graph below illustrated how the number of bedrooms and bathrooms affected the listing prices.



These features resulted in a higher importance weight because they tended to be more linearly correlated. Those with higher bedroom and bathroom numbers tended to result in higher listing prices. Though feature importance highlighted different feature weight, it doesn't tell us how it affected the pricing model. To do this, we can look at feature coefficient from the ridge regression model as seen below:





The graph above shown how different features affected pricing. Those with a higher positive coefficient score would raise the pricing, whereas those with a lower negative coefficient score would lower the pricing. From the graph, we saw how certain amenities such as total bedrooms/ bathrooms, cleanliness score, or has dryer raised the prices. We also saw that some amenities were in charge of lowering

prices as well, such as if the host used a lock box and if the Airbnb is leasing private rooms rather than entire house. More interestingly, we could also see how different street names drove prices up or down. This aligned well with other research articles on how location of the housing played an important role on determining the pricing of the place.

We could therefore conclude pricing of Airbnb were highly correlated to the amenities they offered and locations they are located in, as the majority of the features driving prices were related to amenities/locations. Descriptions provided by hosts do play a role, but has a much lower significant as compared to the list of amenities, type of home, and location feature.

5.3 Conclusion

Training a predictive model on Airbnb listing prices using regression models is possible and will yield fairly decent results. It will also quickly provide insights on the large dataset we are analyzing by picking out certain features that are important, and forming correlation between how different features will drive prices up or down. Furthermore, it allowed us to explore our initial question regarding how given the fact that Airbnb listing prices were left in the hands of hosts, if there existed a correlation between how listing prices were determined or if it's pure random. The model not only allowed us to predict Airbnb listing prices, it also allowed us to dive deeper in the dataset and pull out trends we might not have thought about.

6. REFERENCES

[1] Eric Ghysels, Alberto Plazzi, Walter Torous, and Rossen Valkanov. 2012. Forecasting Real Estate Prices* *HandRE_GPTV. Rady School of Management* (July, 2012), 1-89. http://rady.ucsd.edu/faculty/directory/valkanov/pub/docs/HandRE_GPTV.pdf

[2] Peng Liu Ph.D. 2012. Optimizing Hotel Pricing: A New Approach to Hotel Reservations. *Cornell University School of Hotel Administration. The Scholarly Commons* (August. 2012) https://pdfs.semanticscholar.org/a295/45d40cac2c8161fe4befb7ae8259ac4e81af.pdf

[3] Visit Limsombunchai. 2004. House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *Commerce Division, Lincoln University* (June. 2004), 1-15. https://ageconsearch.umn.edu/bitstream/97781/2/2004-9-house%20price%20prediction.pdf

Airbnb Listing Features as Shown from "Inside Airbnb" Dataset

- id
- listing_url
- scrape_id
- last_scraped
- name
- summary
- space
- description
- experiences_offered
- neighborhood_overview
- notes
- transit
- access
- interaction
- house_rules
- thumbnail_url
- medium_url
- picture_url
- xl_picture_url
- host_id
- host_url
- host_name
- host_since
- host_location
- host_about
- host_response_time
- host_response_rate
- host_acceptance_rate
- host_is_superhost
- host_thumbnail_url
- host_picture_url
- host_neighbourhood
- host_listings_count
- host_total_listings_count
- host_verifications
- host_has_profile_pic
- host_identity_verified
- street
- neighbourhood
- neighbourhood_cleansed
- neighbourhood_group_cleansed
- city
- state
- zipcode
- market
- smart_location
- country_code
- country
- latitude
- longitude
- is_location_exact
- property_type
- room_type
- accommodates
- bathrooms
- bedrooms
- beds
- bed_type
- amenities
- square_feet
- price
- weekly_price
- monthly_price
- security_deposit
- cleaning_fee
- guests_included
- extra_people
- minimum_nights
- maximum_nights
- calendar_updated
- has_availability
- availability_30
- availability_60
- availability_90
- availability_365
- calendar_last_scraped
- number_of_reviews
- first_review
- last_review
- review_scores_rating
- review_scores_accuracy
- review_scores_cleanliness
- review_scores_checkin
- review_scores_communication
- review_scores_location
- review_scores_value
- requires_license
- license
- jurisdiction_names
- instant_bookable
- cancellation_policy
- require_guest_profile_picture
- require_guest_phone_verification
- calculated_host_listings_count
- reviews_per_month