Is the Price Reasonable?

Group 2

Xinyi Gong, Lanyu Shang, Yang Sun

Table of contents

Overview

Airbnb, founded in 2008, provides an online platform for local hosts to rent accommodations for outside travelers. Up to July 2016, there have been 100 million users and 2.3 million listings around the world.[1] In general, there are three categories of roles associated with the company: guests, hosts, and the company itself. In this project, we are going to build different models from the perspective of guests. Followed by each model with different predictors, we will conduct a comprehensive analysis to determine if guests should rent the hosts' accommodations based on model's results.

Problem Identification

Given various datasets in different cities, we will focus on San Francisco. Airbnb was founded and headquartered in San Francisco in 2008. It is convinced that the market in short-term housing is well developed in San Francisco. As guests, we can use regression models to predict a price range with given guests' attributes, such as known days of stay, numbers of companions etc., all of which are included in the dataset. With continuous dependent variables "price" and "log price", we will build linear and semi-log regression models to yield an estimated price, and decide whether the hosts' listing prices are above or under our budgets. Also, with nominal dependent variable "Target", we will build classification algorithms, such as logistic regression and bootstrap forest to classify whether the hosts' listing prices are overpriced, reasonable or underpriced. We will also discuss if applying text mining to extract users' reviews have any predictive power to make classifications.

[1] http://expandedramblings.com/index.php/airbnb-statistics/

Data-driven Analysis

2.1 In-sample Modeling

2.1.1 Data Overview and Preprocessing

We start the analysis with data preprocessing, which is considered as an important yet a very tedious step. We will mainly focus on "*listing2.csv*" that containing detailed listing records in this part of analysis. The dataset contains 8720 listing records in San Francisco and each record includes 95 variables. Intuitively, 33 variables are selected to be potential independent variables in predicting the target variable "*price*":

a. Nominal: "*id*", "*host_id*", "*host_response_time*", "*property_type*", "*room_type*", "*bed_type*", "*cancellation_policy*", "*host_is_superhost*", "*host_has_profile_pic*", "*host_identity_verified*", "*is_location_exact*", "*instant_bookable*".
b. Continuous: "*bathrooms*", "*bedrooms*", "*beds*", "*host_response_rate*", "*host_listings_count*", "*latitude*", "*longitude*", "*security_deposit*", "*guests_included*", "*extra_people*", "*minimum_nights*", "*availability_365*", "*review_scores_rating*", "*review_scores_accuracy*", "*review_scores_cleanliness*", "*review_scores_checkin*", "*review_scores_communication*", "*review_scores_location*", "*review_scores_value*", "*calculated_host_listings_count*", "*reviews_per_month*".

Next, 11 empty entries are removed and missing values are replaced with appropriate substitutions, e.g. mean, zero or a new category.

2.1.2 Bivariate Analysis

Besides intuition and common knowledge about price, data visualization also gives us a lot of

hints to seek for powerful predictors, especially when making bivariate analysis. First, we start understanding the variables with the bivariate analysis, that is, examine every variable with target variable "*price*". Our target is a continuous variable, so for the dependent variables for bivariate analysis, if it is nominal, we will apply t-test for two-level nominal (binary) variable and ANOVA for nominal variable with two more levels; if it is continuous, correlations and regression details will be applied to see it the variable can affect the dependent variable significantly. For example:

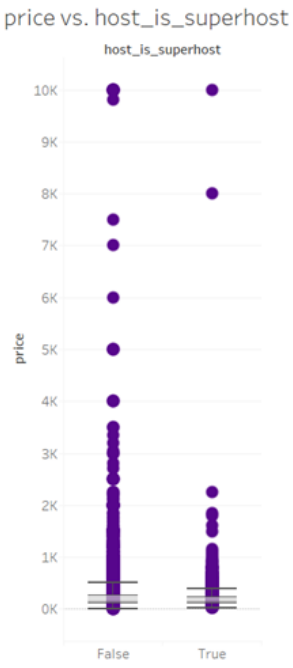


Figure 1

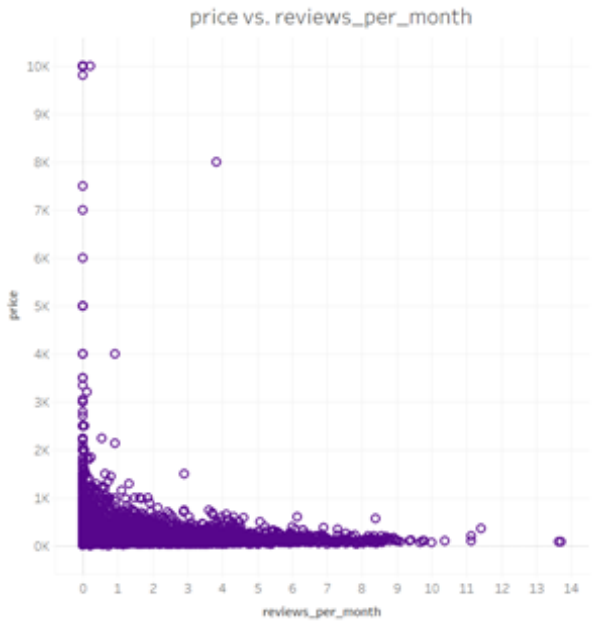


Figure 2

As shown in Figure 1, dependent variable "*price*" has a very different distribution with nominal variable "*host_is_superhost*", so we can assume that these two variables are strongly correlated. We continue our analysis using t-test, and the result shows a t-ratio of 3.91, which supports our assumption. Similarly, we plot the scatter plot first, and then analyze the correlation between two continuous "*price*" and "*reviews_per_month*", and the result shows they are highly correlated. In summary, we find 12 nominal predictors and 10 continuous variables that are strongly correlated

with "*price*", and use them as our predictors to build our regression models. Due to the limit length of report, the visualizations (scatterplots) of bivariate analysis have been attached in the appendix. To discuss them easily, according to the description, these predicting variables are divided into several groups. Here are the intuitions and statistic results of these 22 variables.

a) About Booking

- "*instant_bookable*": whether the listing can be booked immediately without the approval of hosts.
- "*cancellation_policy*": what will happen if guests want to cancel the booking.
- "*security_deposit*": deposit required to secure the stay (e.g., property damage).

For these three variables about booking process, listings with higher price may have higher quality and more regulations: not instantly bookable, strict cancellation policy or greater security deposit.

Variable Name	p-value	Correlation (if applicable)
"***instant_bookable***"	0.01253	---
"***cancellation_policy***"	6.94e-12	---
"***security_deposit***"	<2e-16	0.100689

b) Availability

- "*availability_365*": how many days in a year will the listing ready to be booked. Intuitively, listings with low availability (either booked or occupied) reflect a high demand and is expected to have higher price. The correlation to the target variable is -0.04 and the p-value by linear regression is 0.000197.

c) About Hosts

- "*host_response_time*": how long will the host reply the message.
- "*host_is_superhost*": whether the host is qualified as a superhost.
- "*host_identity_verified*": whether the identity of hosts is verified.
- "*calculated_host_listing_count*": the number of listings of a host.

For this group of variables, well-qualified hosts are expected to provide great service which may further affect the price, for example, a super-host is expected to know the market better and thus the price should be more reasonable.

Variable Name	p-value	Correlation (if applicable)
"***host_response_time***"	<2e-16	---
"***host_is_superhost***"	9.443e-05	---
"***host_identity_verified***"	0.001092	---
"***calculated_host_listing_count***"	2.51e-07	-0.05522791

d) Property Type

- "*property_type*": type of the listing. Intuitively, different types can affect the price, e.g., castle is more expensive than tent. By ANOVA test, the p-value is 8.67e-05.

e) Room Property

- "*room_type*": type of rooms.
- "*bathrooms*": number of bathrooms in the listing.
- "*bedrooms*": number of bedrooms in the listing.
- "*beds*": number of beds in the listing.

- "*bed_type*": type of beds in the listing.
- "*guests_included*": number of guests the listing supposes to hold.

All variables above reflect the listing size and condition, which can affect the price. The bigger and better-condition listings are tended to have higher prices than the smaller ones.

Variable Name	p-value	Correlation (if applicable)
"*room_type*"	<2e-16	---
"*bathrooms*"	<2e-16	0.2269761
"*bedrooms*"	<2e-16	0.2985625
"*beds*"	<2e-16	0.2389547
"*bed_type*"	0.00394	---
"*guests_included*"	2.89e-13	0.07811042

f) Reviews/Month

- "*reviews_per_month*": number of reviews received per month. A reasonably priced listing is expected to be more popular and thus to receive more reviews per month. The correlation to the target variable is -0.055 and the p-value by linear regression is <2e-16.

g) Review Scores

- "*review_scores_rating*": overall rating of stay.
- "*review_scores_accuracy*": the accuracy of listing description.
- "*review_scores_cleanliness*": the cleanliness of listing.
- "*review_scores_checkin*": the check-in process.
- "*review_scores_communication*": the communication with hosts.
- "*review_scores_location*": the location of listing.

Intuitively, a listing with higher review is expected to have higher price as it reflects good quality and service.

Variable Name	p-value	Correlation (if applicable)
"*review_scores_rating*"	1.797e-05	0.04593894
"*review_scores_accuracy*"	0.006464	0.02917963
"*review_scores_cleanliness*"	0.0005223	0.03716541
"*review_scores_checkin*"	0.04402	0.0215808
"*review_scores_communication*"	0.02993	0.02326415
"*review_scores_location*"	0.0008218	0.03584101

2.1.3 Linear and Semi-Log Regressions

According to above bivariate analyses, we can prove that 22 predictors are strongly correlated with price or log price. Comparatively, we found that as continuous dependent variable, semi-log price model performs much better than the linear model, as shown in Figure 3.

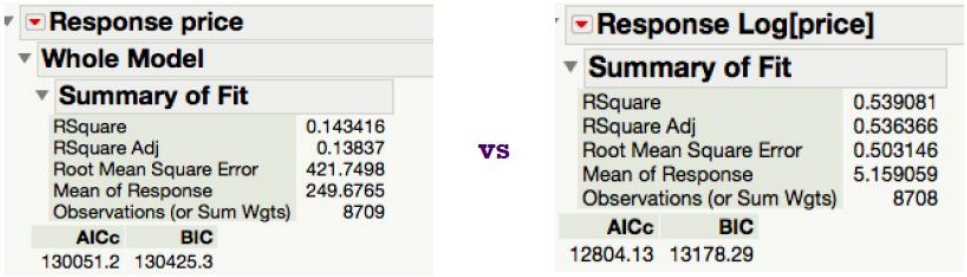
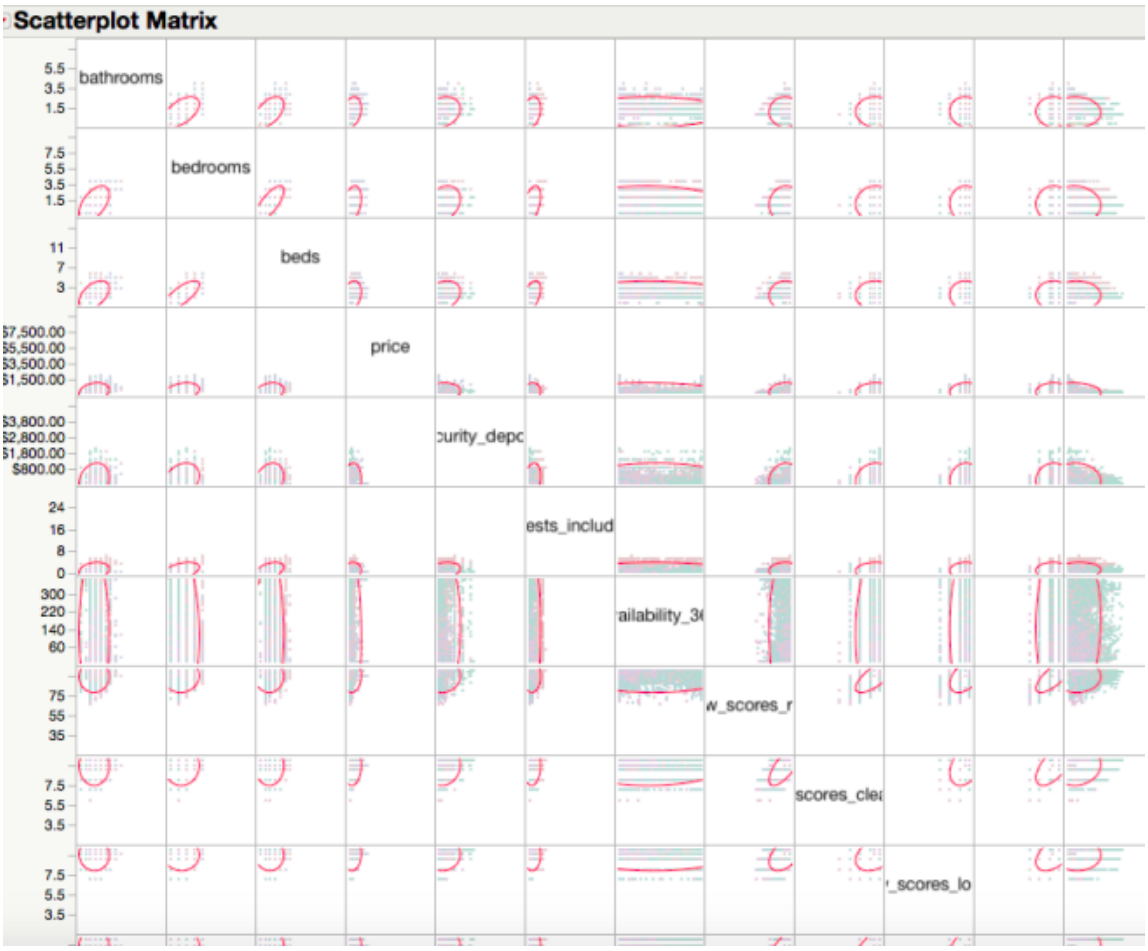


Figure 3

Among the 22 predictors, many of them have similar meanings, which are likely to contain similar information, so it is likely for them to be highly correlated. Therefore, we can apply multivariate analysis and factor analysis to see if we can simplify the model without reducing predictive power. Multivariate analysis and Principal Component Analysis result follows:



Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.2032	60.046		60.046
2	0.7630	10.900		70.946
3	0.6990	9.985		80.932
4	0.4033	5.762		86.693
5	0.3798	5.426		92.120
6	0.3444	4.921		97.040
7	0.2072	2.960		100.000

Figure 4

We can see that only one factor has an eigenvalue greater than 1, and grouping all 7 review scores and run a regression, we have following regression report:

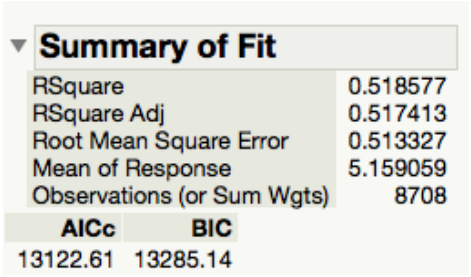


Figure 5

R-square decreases and AIC/BIC increases, so in this semi-log model, factor analysis is not applicable in our case. Even though column cannot be grouped by factor analysis, we can still combine rows using clustering analysis. It will be easier for us to search for outliers, and excluding them from the dataset will not only help us reduce dimensions, but also improve model accuracy. For example, there are outliers in "price" and "security deposit", and hiding 174 rows help us increase R-square by 1.6%.

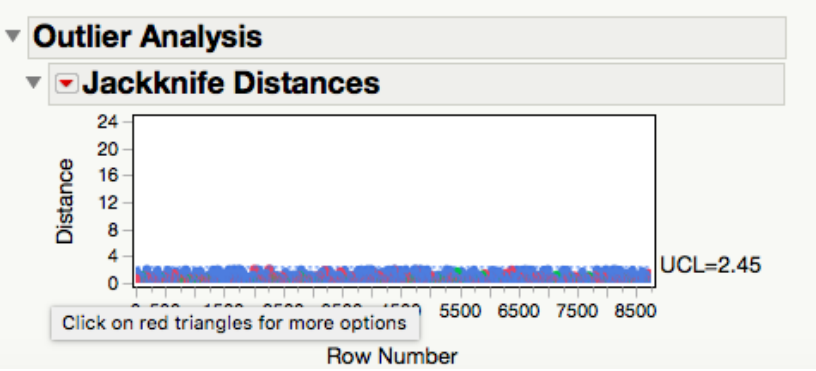


Figure 6

Our final regression model's results, features and coefficients are shown in Figure 7 and 8.

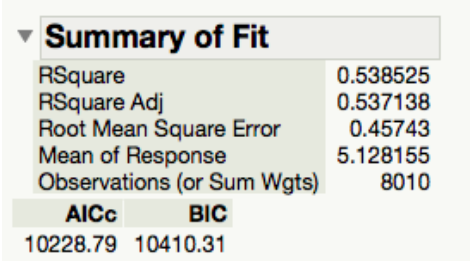


Figure 7

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.6076363	0.234367	11.13	<.0001*
host_response_time[a few days or more]	0.078334	0.041978	1.87	0.0621
host_response_time[N/A]	0.2049233	0.016141	12.70	<.0001*
host_response_time[within a day]	-0.015854	0.018256	-0.87	0.3852
host_response_time[within a few hours]	-0.038932	0.015729	-2.48	0.0133*
host_is_superhost[f]	-0.089616	0.015249	-5.88	<.0001*
host_identity_verified[f]	0.0237336	0.012213	1.94	0.0520
room_type[Entire home/apt]	1.141381	0.037536	30.41	<.0001*
room_type[Private room]	0.5698977	0.037204	15.32	<.0001*
bathrooms	0.155993	0.013081	11.93	<.0001*
bedrooms	0.1845283	0.010899	16.93	<.0001*
beds	0.0788507	0.009406	8.38	<.0001*
security_deposit	1.6821e-5	1.975e-5	0.85	0.3945
guests_included	0.0058403	0.005822	1.00	0.3158
availability_365	0.0005147	4.178e-5	12.32	<.0001*
instant_bookable[f]	-0.002534	0.014027	-0.18	0.8566
reviews_per_month	-0.049811	0.004042	-12.32	<.0001*
cancellation_policy[flexible]	-0.468438	0.188179	-2.49	0.0128*
cancellation_policy[moderate]	-0.552193	0.188131	-2.94	0.0033*
cancellation_policy[strict]	-0.535849	0.187956	-2.85	0.0044*
cancellation_policy[super_strict_30]	-1.202577	0.22172	-5.42	<.0001*
review_scores_rating	0.0050617	0.001842	2.75	0.0060*
review_scores_cleanliness	0.07296	0.011707	6.23	<.0001*
review_scores_location	0.1152075	0.011071	10.41	<.0001*
review_scores_value	-0.072432	0.013074	-5.54	<.0001*

Figure 8

2.1.4 Conclusion

As guests, we generally have an expected price for accommodation. Using the semi-log regression model, we can fill in our requirements for stay, such as room types and reviews for hosts etc, to get an estimated price range. If the hosts' listed prices are higher than the estimated price, then searching for another hosts with lower price might be better choice. During the process of model construction, we find some features such as "reviews" that are composed with English characters. They affect customers' attitude toward hosts' accommodations and are likely to be useful indicators to predict prices. We apply text mining to extract information from them, and help us make recommendation for guests.

2.2 Out-sample Modeling

2.2.1 Data Preprocessing

First, as machine learning algorithms are commonly used for classification tasks, we construct a new nominal target variable with 3 levels (*overpriced, reasonable,* and *underpriced*) from the original target "Price". K-means clustering is applied to group listings into 29 groups since the clustering algorithm suggests 29 clusters (Figure 9) will reach the optimal Cubic Clustering Criterion. Next, the mean price in each cluster is used as the benchmark of price ("*price_benchmark*") in each cluster and the percentage difference is calculated for each listing with the formula $\frac{price - price\_benchmark}{price\_benchmark}$. From the frequency distribution of percentage difference shown in Figure 10, we cut listings into 3 categories: listings whose percentage difference are in the first quartile (less than -0.4176) will be considered as *underpriced*, listings in the second and third quartiles will be considered as *reasonable*, and listings in the last quartile will be considered as *overpriced*. The distribution of the new target variable is shown in Figure 11.

Cluster Comparison

Method	NCluster	CCC	Best
K-Means Clustering	10	92.2042	
K-Means Clustering	11	97.1637	
K-Means Clustering	12	104.144	
K-Means Clustering	13	120.493	
K-Means Clustering	14	118.861	
K-Means Clustering	15	124.404	
K-Means Clustering	16	146.099	
K-Means Clustering	17	137.174	
K-Means Clustering	18	140.085	
K-Means Clustering	19	152.163	
K-Means Clustering	20	153.969	
K-Means Clustering	21	153.363	
K-Means Clustering	22	134.079	
K-Means Clustering	23	151.401	
K-Means Clustering	24	159.832	
K-Means Clustering	25	158.64	
K-Means Clustering	26	148.694	
K-Means Clustering	27	163.027	
K-Means Clustering	28	150.166	
K-Means Clustering	29	165.604	Optimal CCC
K-Means Clustering	30	146.585	

Figure 9

Quantiles

100.0%	maximum	13.142620232173
99.5%		3.9985802591282
97.5%		1.7020447906524
90.0%		0.6817951435285
75.0%	quartile	0.1779424956329
50.0%	median	-0.154223051797
25.0%	quartile	-0.416705552963
10.0%		-0.562529164722
2.5%		-0.697287697772
0.5%		-0.789129632558
0.0%	minimum	-0.94728240814

Figure 10

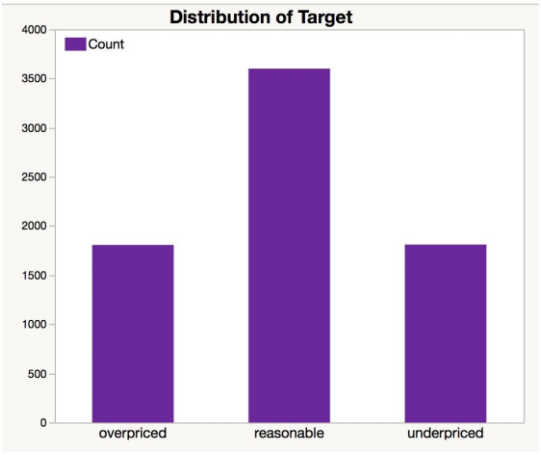


Figure 11

What is more, the review data ("*review2.csv*") has also been cleaned for text modeling. Reviews not in English or invalid, such as automated comments and comments with single character or punctuation, are removed. Reviews corresponding to the same "*listing_id*" are merged together, and joined with the listing dataset. In addition, texts in reviews are also preprocessed for text analysis with the following methods: lowering letters, stemming, removing punctuations, stop words and numbers.

2.2.2 Text Mining

As mentioned before, text mining can be used to extract information from variables that entirely consist of words. After creating three groups according to the difference between hosts' listing prices and our constructed "benchmark prices", we mine information from each group to see how different probability of words' appearances in each review are correlated groups' prices, as shown in Figure 12, 13 and 14. The larger word's font size, the more likely it appears in customers' reviews.

Figure 12. Reasonable

Figure 13. Overpriced

Figure 14. Underpriced

The three plots give us less information than expected, because the three groups all contain "*great*", "*place*", "*stay*", "clean" etc. The guests of Airbnb in San Francisco seem to be nice on the comments. Therefore, sentiment analysis for our target variable "*price*" may be unnecessary as the frequent terms in these 3 classes are not significantly different. However, we can still use text mining skills for exploring these three groups. Then, we explore the terms correlated to the word "*price*" and top topic to carry on data-driven analysis.

a. Group "Reasonable"

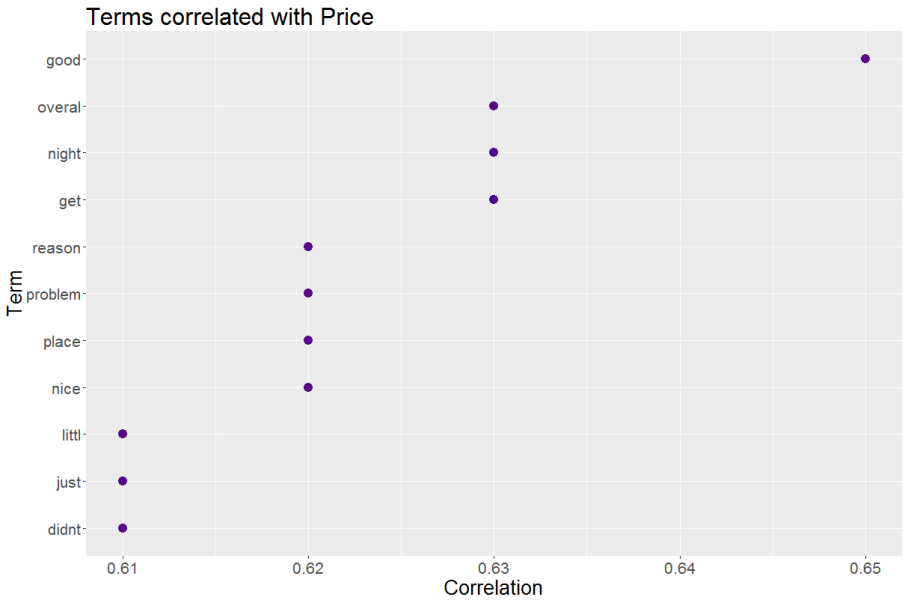


Figure 15. Terms correlated in "Reasonable"

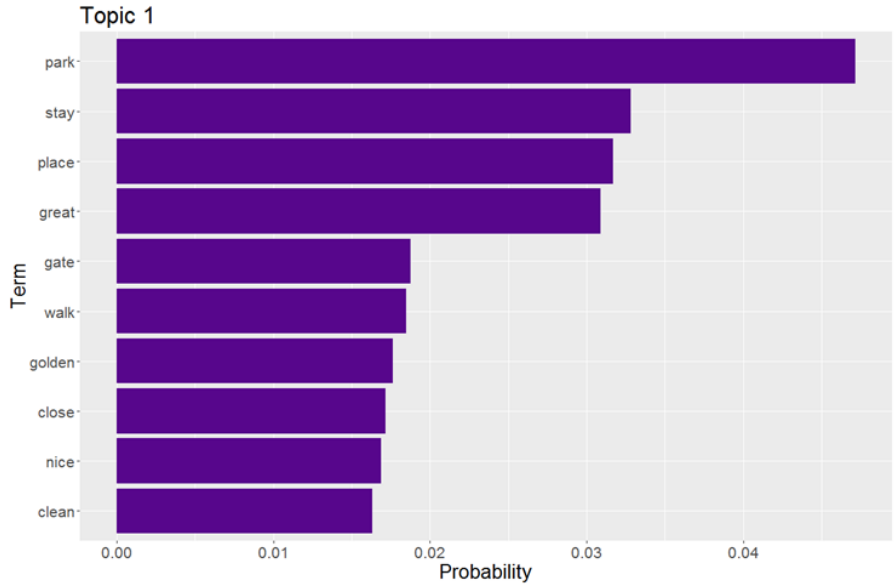


Figure 16. Top Topic in "Reasonable"

From Figure 15, "*good*" is the highest correlated word, which indicates that guests are fairly satisfied with the listing price. The topic with highest probability (Figure 16) is about the distance to the scenic spots, which also refers that the location and price are in the good trade-off.

b. Group "Overpriced"

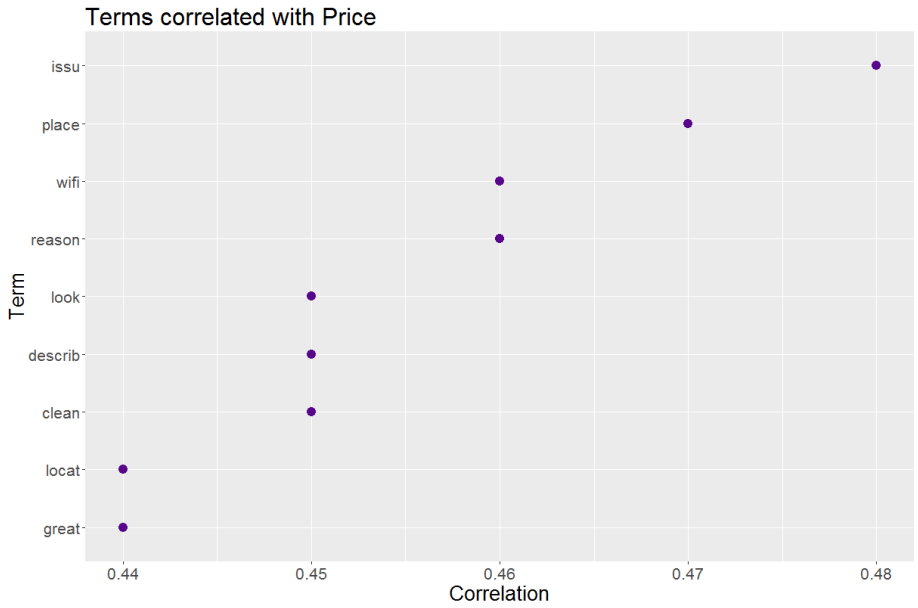


Figure 17. Terms correlated in "Overpriced"

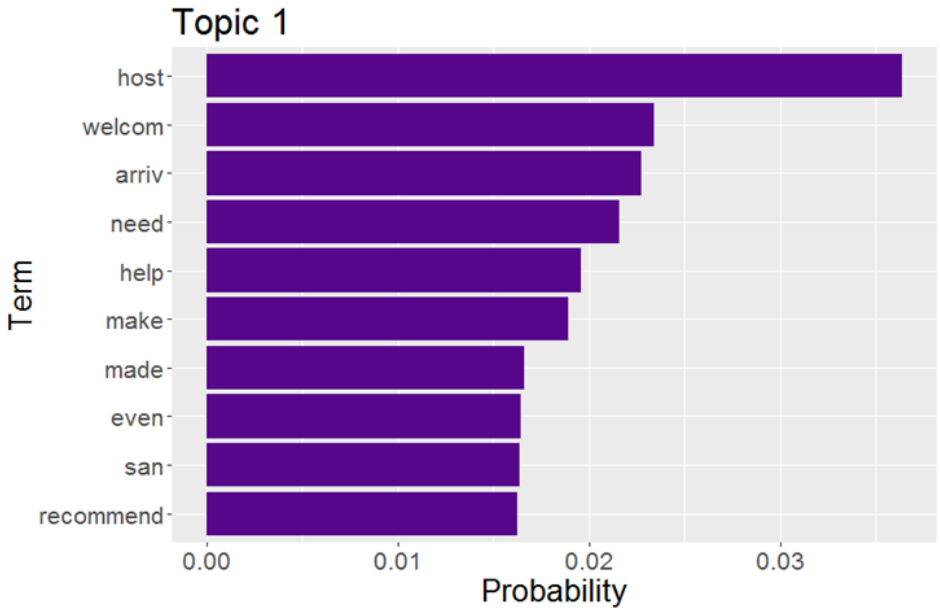


Figure 18. Top Topic in "Overpriced"

From Figure 17, "*issue*" is the highest correlated word with "*price*", which may reflect that there exist some disagreements on the listing price. For the top topic in "Overpriced" (Figure 18), it is about the customer services, which could be the reason for the extra fees that results in the higher

price.

c. Group "Underpriced"

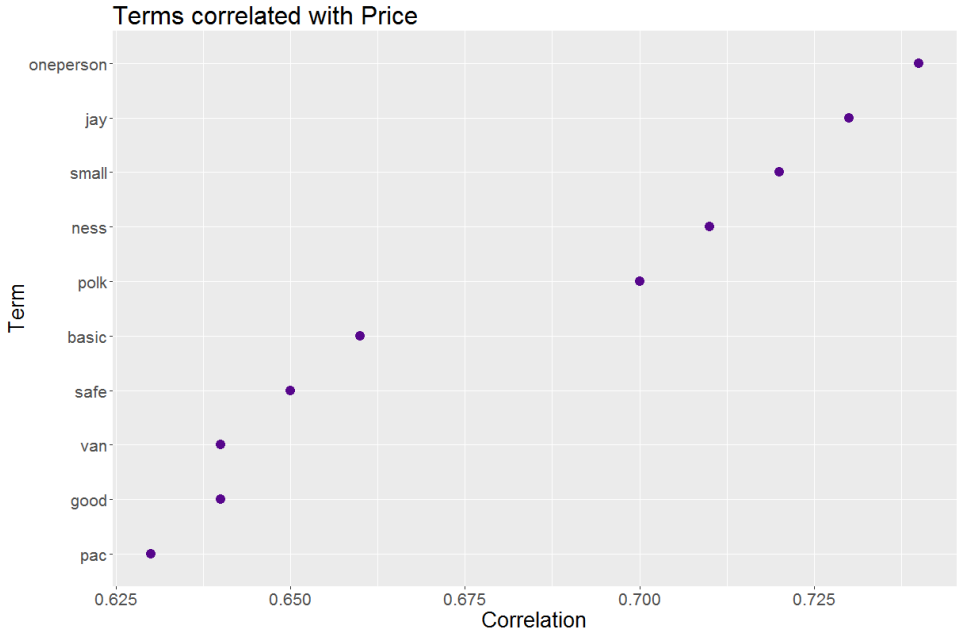


Figure 19. Terms correlated in "Underpriced"

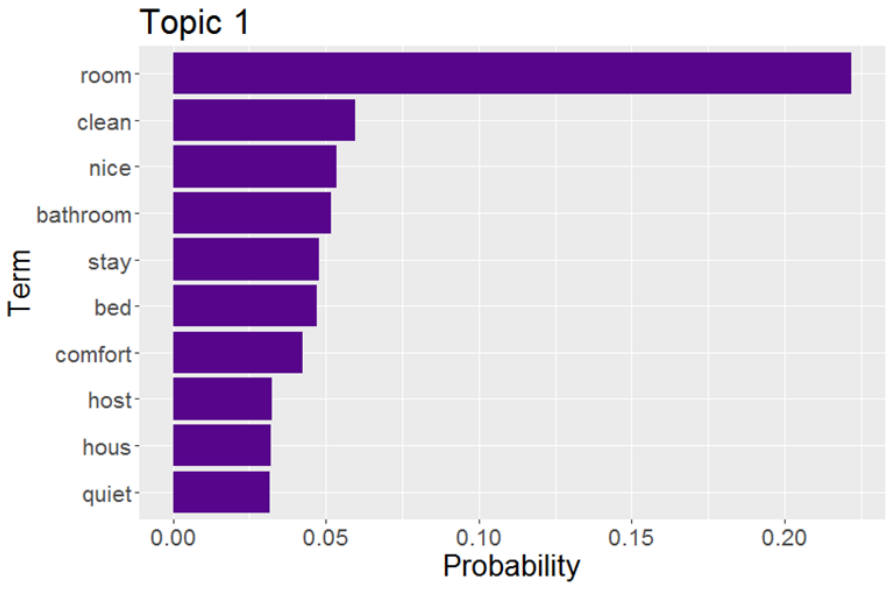


Figure 20. Top Topic in "Underpriced"

From Figure 19, the terms have higher correlations such as "*oneperson*", "*small*" and "*basic*" are likely to be related with a single room or shared room, which has small space and basic amenity.

It could be the reason for underpricing. For the top topic (Figure 20) in group "Underpriced" is talking about the quality of listing. Some positive words such as "*clean*", "*nice*" and "*quiet*" have higher probabilities, which shows that the quality of listing and experience of stay could exceed guests' expectations for the given listing price.

In short, the frequent terms in these three classes are not significantly different. Sentiment analysis could be used to predict rating, instead of our dependent variable --- "*price*". Therefore, use rating for the price prediction could be enough.

2.2.3 Machine Learning

Recalling that the new target created earlier is consisting of 3 levels, appropriate machine learning models are built to solve the multiclass classification problem. 80% of the data are used for training models and 20% of the data are used for validation.

a. Logistic Regression

Multinomial logistic regression is used to classify dependent variables with more than two levels. According to the fit details shown in Figure 21, the multinomial logistic regression could achieve a three-class misclassification rate of 0.3165 on the validation set. The root mean square error (RMSE) is about 0.4735.

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3729	0.2826	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6166	0.5086	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.6520	0.7484	∑ -Log(ρ[j])/n
RMSE	0.4628	0.4735	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.4031	0.4095	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2863	0.3165	∑ (ρ[j]≠ρMax)/n
N	5715	1504	n

Figure 21. Logistic Regression Fit Details

b. Bootstrap Forest

Bootstrap Forest, derived from decision tree method, could also be used to predict multiclass outcomes. Fitting the bootstrap forest with default parameters, the model achieves a multiclass misclassification rate of 0.3777 and RMSE of 0.5447. Tuning the parameters, default parameters appear to have the best result.

Number of Trees in the Forest	100
Number of Terms Sampled per Split:	7
Bootstrap Sample Rate	1
Minimum Splits per Tree:	10
Maximum Splits per Tree	2000
Minimum Size Split:	7

Measure	Training	Validation
Entropy RSquare	0.4469	0.2134
Generalized RSquare	0.6916	0.4103
Mean -Log p	0.5751	0.8206
RMSE	0.4403	0.5447
Mean Abs Dev	0.4155	0.5167
Misclassification Rate	0.1601	0.3777
N	5715	1504

Figure 22. Parameter Setting and Fit Details

c. Neural Network

As an emerging method in machine learning, neural network is also attempted to fit the data. We begin with training a neural network with one layer and the result is shown in Figure 23. It reaches a misclassification rate of 0.4309 and RMSE of 0.5604. In addition to the default model, we add a second layer to the model and achieve a relative lower misclassification rate of 0.3969 and RMSE of 0.5464.

Validation

target

Measures	Value
Generalized RSquare	0.299515
Entropy RSquare	0.1458293
RMSE	0.5604209
Mean Abs Dev	0.5277727
Misclassification Rate	0.4308511
-LogLikelihood	1340.2208
Sum Freq	1504

Confusion Matrix

Actual	Predicted Count		
target	overpriced	reasonable	underpriced
overpriced	118	226	28
reasonable	98	565	81
underpriced	28	187	173

Figure 23. Neural Network Fit Details

d. Model Comparison and Conclusion

According to three machine learning methods discussed above, it appears that multinomial logistic regression performs the best on misclassification rate and RMSE as seen in Figure 22. Therefore, logistic regression will be selected as the best model on predicting the target variable.

Measures of Fit for target

Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Training	Fit Nominal Logistic		0.3729	0.6166	0.652	0.4628	0.4031	0.2863	5715
Training	Bootstrap Forest		0.4469	0.6916	0.5751	0.4403	0.4155	0.1601	5715
Training	Neural		0.1996	0.3883	0.8322	0.5421	0.5108	0.4024	5715
Training	Neural		0.2269	0.4299	0.8039	0.5327	0.4948	0.3811	5715
Training	Bootstrap Forest		0.2570	0.4731	0.7726	0.5321	0.5146	0.3340	5715
Validation	Fit Nominal Logistic		0.2259	0.4291	0.8076	0.4735	0.4095	0.3165	1504
Validation	Bootstrap Forest		0.2134	0.4103	0.8206	0.5447	0.5167	0.3777	1504
Validation	Neural		0.1458	0.2995	0.8911	0.5604	0.5278	0.4309	1504
Validation	Neural		0.1802	0.3578	0.8552	0.5464	0.5064	0.3969	1504
Validation	Bootstrap Forest		0.1842	0.3643	0.8511	0.5621	0.5442	0.3983	1504

Figure 22. Model Comparison

2.3 Model Deployment

Finally, we will deploy our models on a real listing as shown in the table below and see how the predictions will be.

Variable Name	Value	Variable Name	Value
host_response_time	within a few hours	***instant_bookable***	False
property_type	House	***room_type***	Private room
cancellation_policy	Flexible	***bathrooms***	1
review_scores_accuracy	10	***bedrooms***	1
host_identity_verified	False	***bed_type***	Real Bed
calculated_host_listings_count	1	***beds***	1
review_scores_checkin	10	***guests_included***	1
review_scores_communication	10	***security_deposit***	NA

review_scores_location	10	***availability_365***	124
reviews_per_month	0.57	***review_scores_rating***	98
review_scores_cleanliness	10	***host_is_superhost***	False

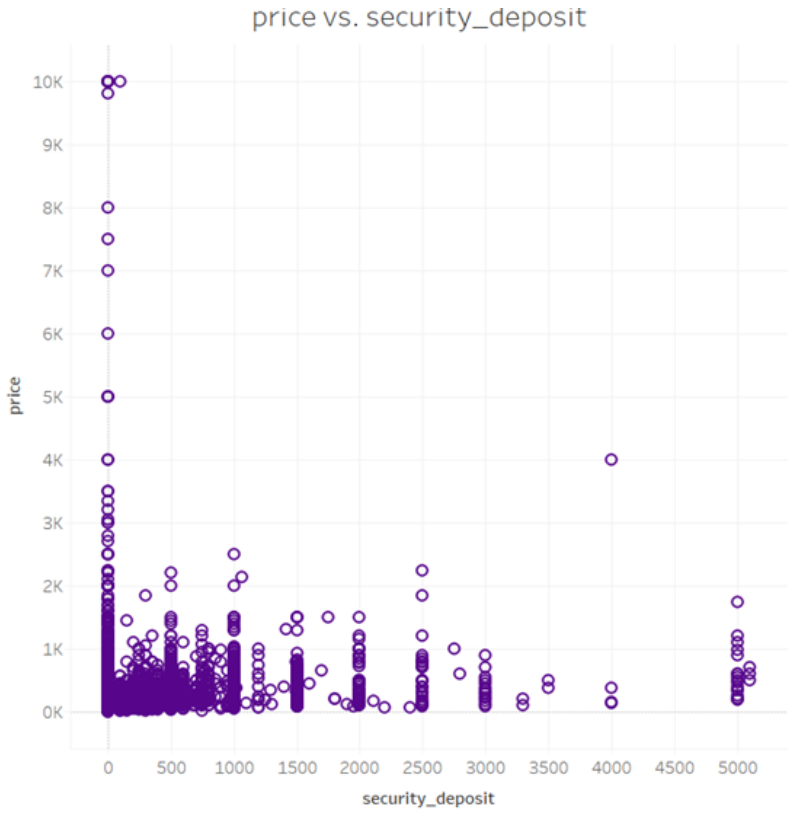
The actual price of the listing above is $80. The regression model we developed gives us a predicted value of $195.15 and a 95% confidence interval of (188.10, 202.35). The logistic regression also tells us that the actual price is underpriced. In addition, the price benchmark we derived from clustering is $171.44 which also confirms us that our prediction is accurate and this listing is a bargain on the market.

Conclusion

As what has been discussed above, using the data scraped from Airbnb, we finally derived both regression and classification oriented models. The semi-log regression model could provide an Airbnb guest with an expected range for a given listing, while the classification model could tell how the given listing is compared to similar listings on the marketplace. Both methods could better assist guests in choosing their desired vacation rentals in San Francisco in a smart and efficient way.

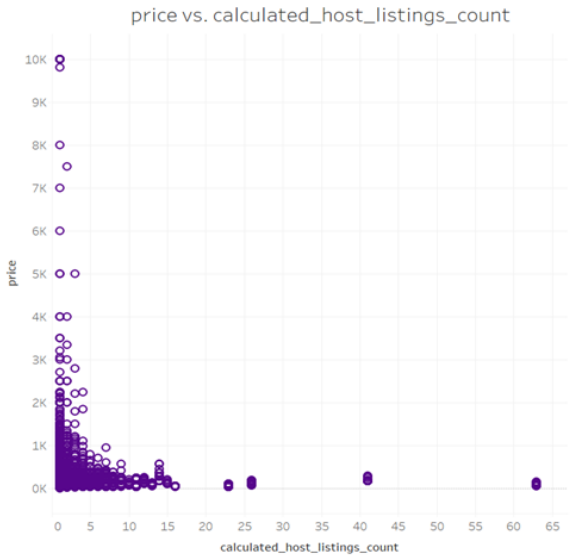
Appendix - Visualization for Bivariate Analysis

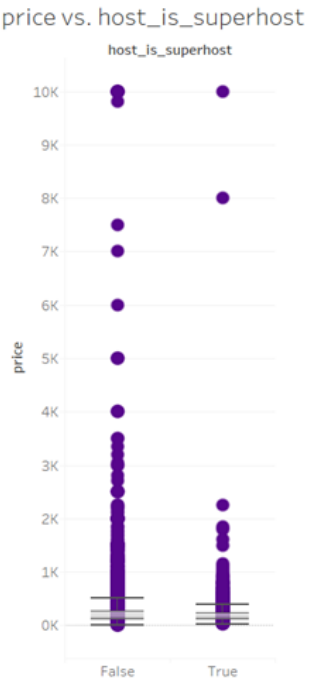
1. About booking

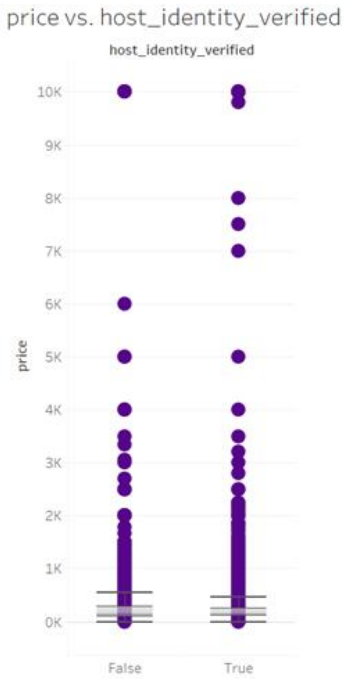


2. Availability

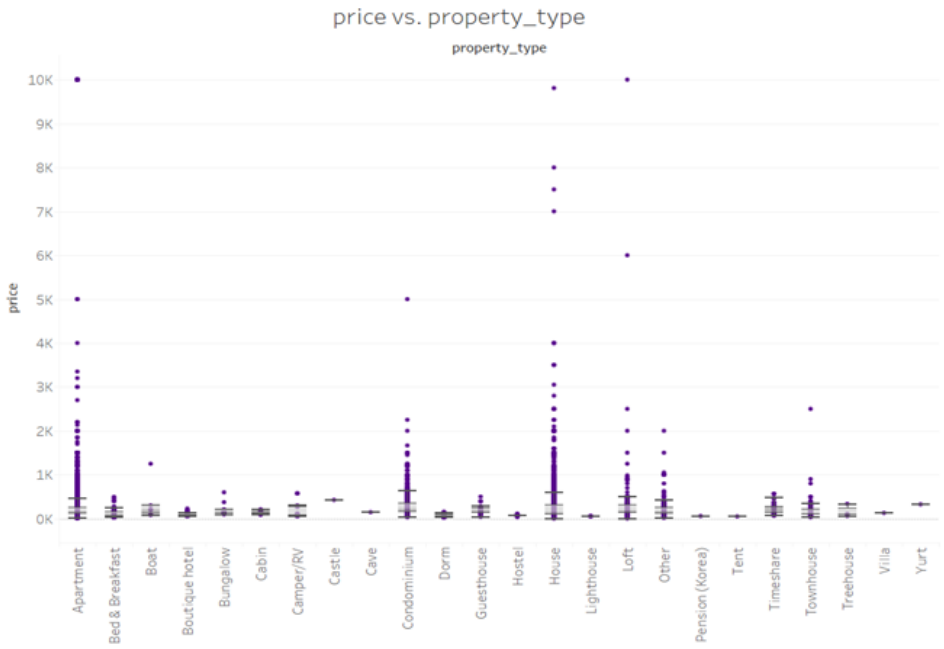
3. About hosts





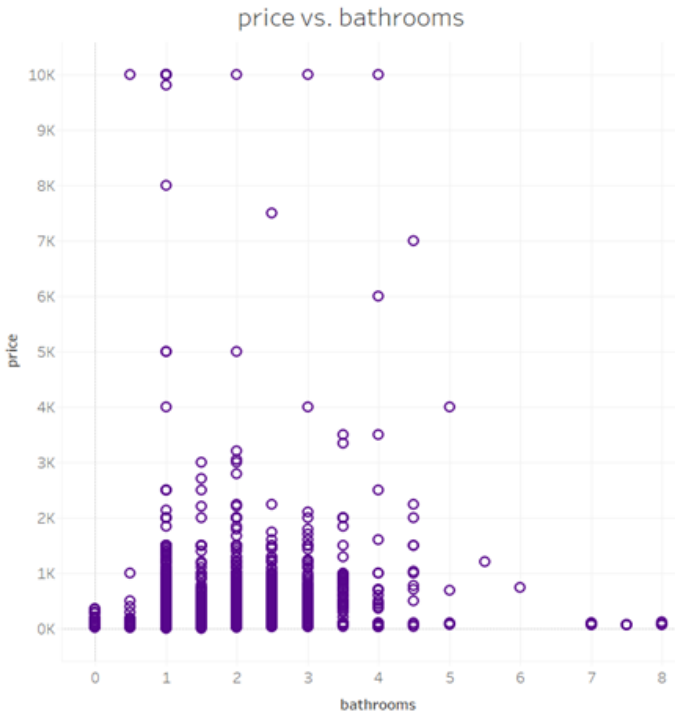


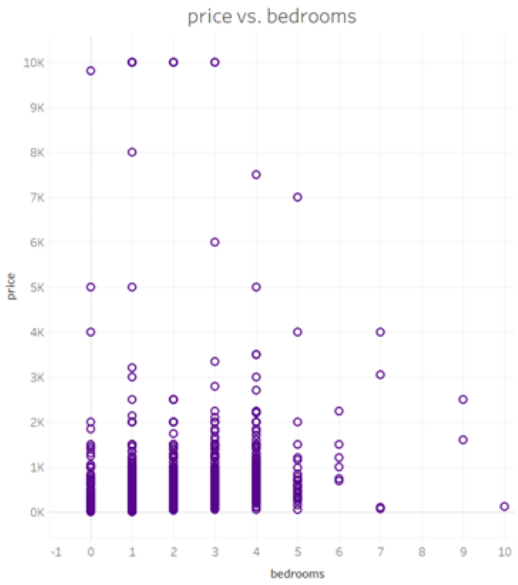
4. Property type

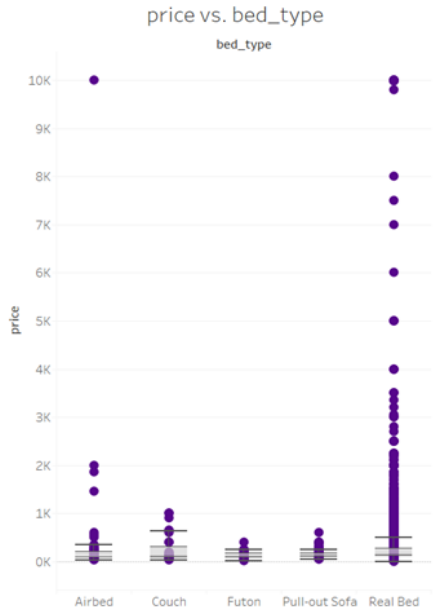


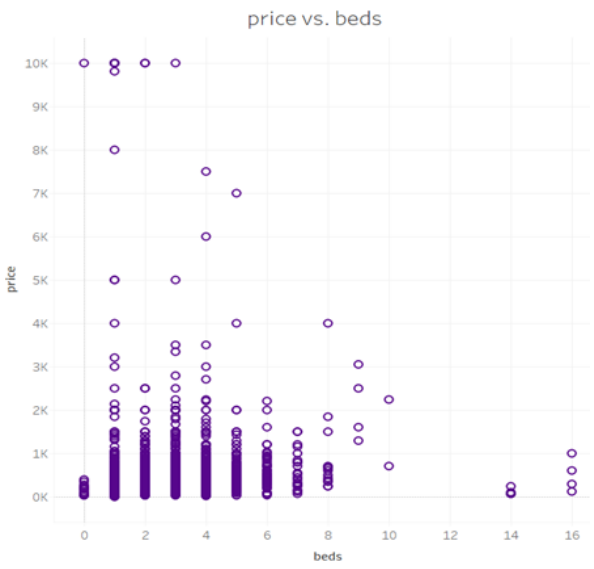
5. Room properties


price vs. room_type
room_type
price
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Entire home/apt
Private room
Shared room

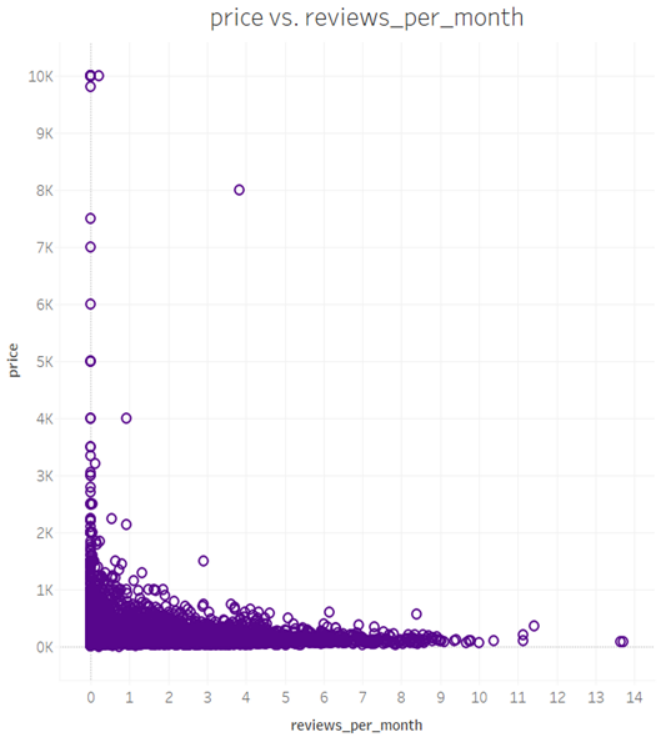

price vs. bathrooms
price
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
0
1
2
3
4
5
6
7
8
bathrooms


price vs. bedrooms
price
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
-1
0
1
2
3
4
5
6
7
8
9
10
bedrooms

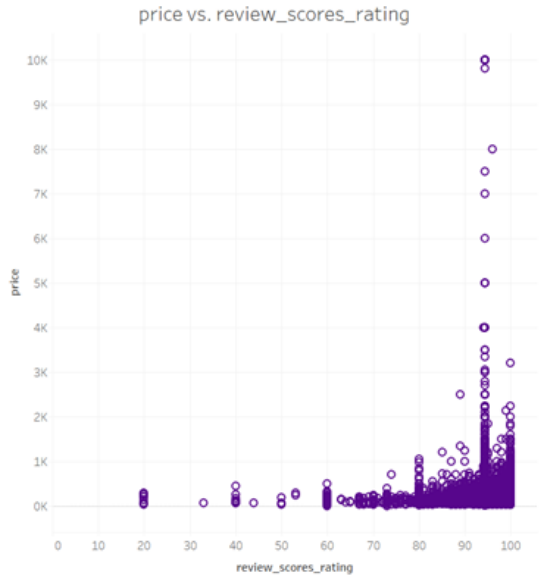

price vs. bed_type
bed_type
price
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
Airbed
Couch
Futon
Pull-out Sofa
Real Bed

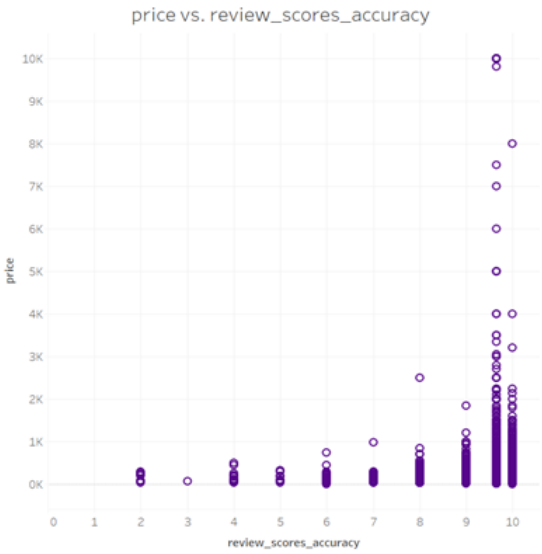


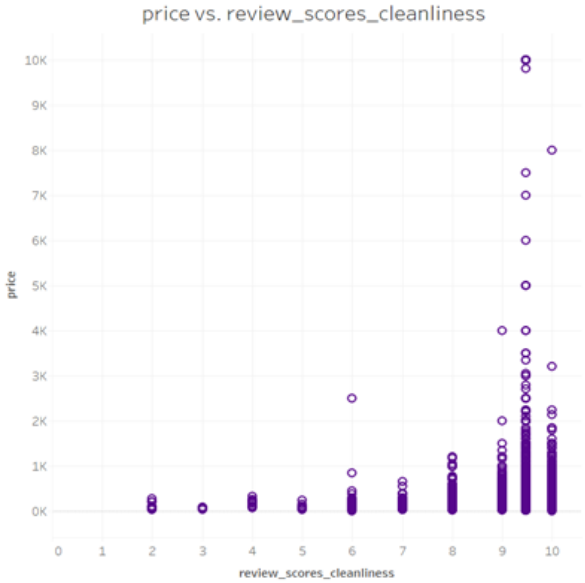
6. Reviews/month

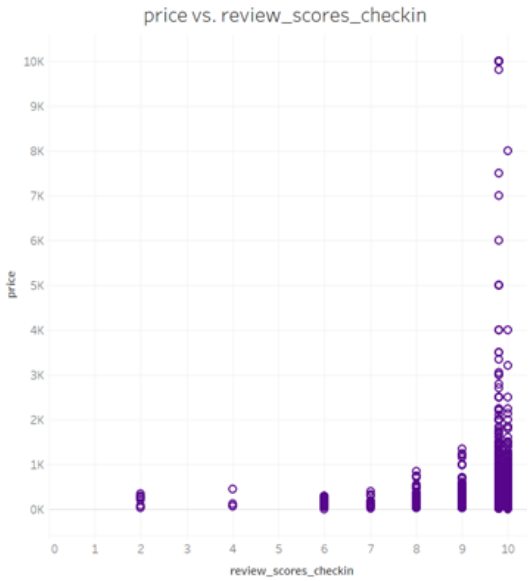


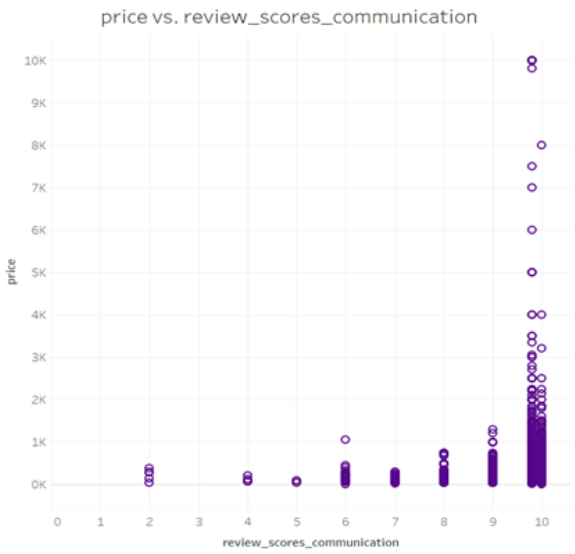
7. Review scores

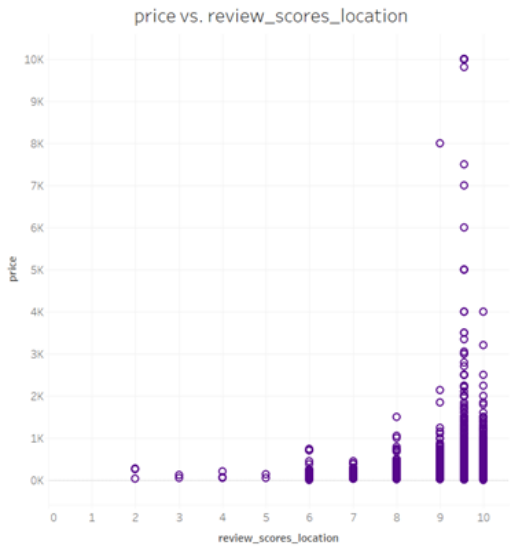

price vs. review_scores_rating
price
review_scores_rating


price vs. review_scores_accuracy
price
review_scores_accuracy


price vs. review_scores_cleanliness
price
review_scores_cleanliness


price vs. review_scores_checkin
price
review_scores_checkin


price vs. review_scores_communication
price
review_scores_communication
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
0 1 2 3 4 5 6 7 8 9 10


price vs. review_scores_location
price
review_scores_location
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
0 1 2 3 4 5 6 7 8 9 10