What Features Drive Demand High?

Airbnb in New York, Fitting a Demand Surface

Final Report for D3M @ Stern by Daniel Chin | Andrew Liu | Wendi Tang

Table of Contents

0. Research Question and Context

In this project, we use data analysis to inform Airbnb hosts in NY what are the best features that boost demand.

The demand curve describes the relation between P (price) and Q (quantity). In the context of Airbnb, Q is translated to “occupancy rate”, i.e. the percentage of available days that end up being booked by the guests. We will fit a model to predict occupancy rate from the price and the features of a listing. The model will reveal the features that influence demand.

1. Estimate Occupancy

The dataset is scraped from Airbnb public listings and hence does not contain the occupancy rate for the listings. Yet, we need the occupancy rate to calculate the sales per year for any given host. To estimate the occupancy rate, most analysts chose to use the number_of_reviews. For example, Inside Airbnb's "San Francisco Model" assumed a review rate of 50% [1]. In contrast, Airbnb's CEO Brian Chesky reported a review rate of 72% [2]. Their numbers diverge a lot, so we decide to calculate our own estimation from the dataset.

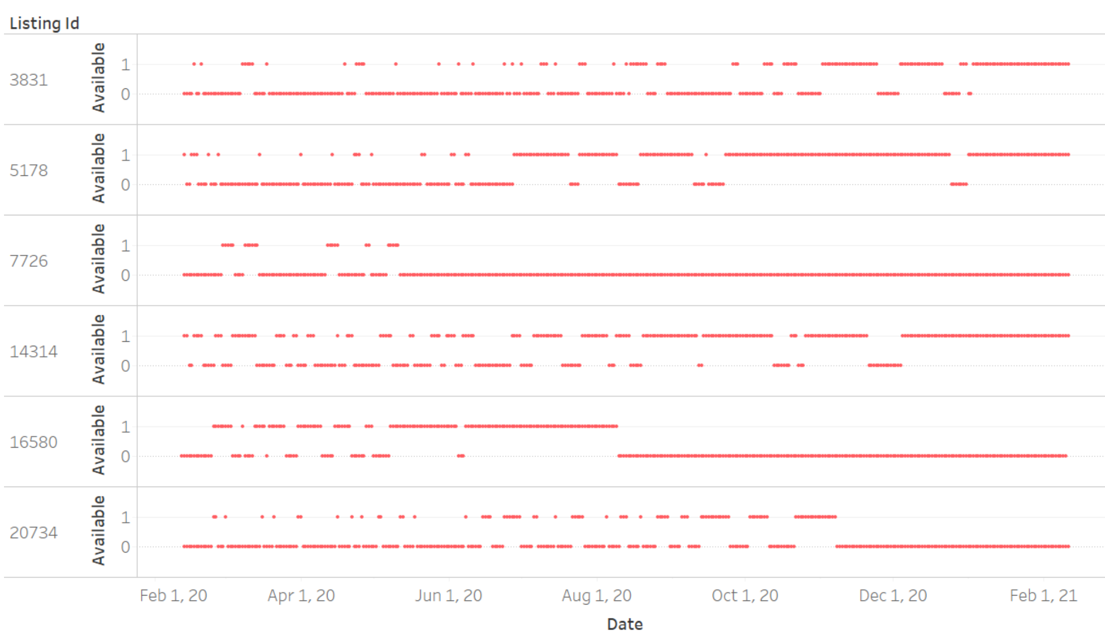


Figure 1

Figure 1 shows the availability calendar for six selected listings. It is a snapshot taken on scrape_day (Feb. 12-13, 2020). “Available” means the day is open for booking. “Unavailable” has two possible causes: a) the day is occupied by another guest, or b) the host closed it off. Usually, the dots trend upwards (Figure 1 listing 3831, 5178, 14314), hinting that the unavailable days on the left are sales. However, the trend is not observed in listing 7726, 16580, and 20734. That is due to large chunks of days being closed off by the hosts.

Thus, the process to estimate occupancy is clear. Denote the horizontal axis in figure 1 as “x”. First, estimate probability_unavailable for x=0 (i.e. the first day on the availability calendar). Second, estimate probability_closed (intrinsic property of the listing, independent of x). Third, subtract them, and we will get the estimation of the occupancy rate. We use the left half ($0<x<x_1$) to estimate probability_unavailable for x=0. We use the right half ($x_2<x<end$) to estimate probability_closed.[1]

We use machine learning to find the optimal x_1 and x_2. We download two instances of the dataset (whose scrape dates are different). For any given x_1, we can try to use *one dataset* to predict *the other dataset*’s availability at x=1. The goodness of x_1 is thus indicated by the cross-entropy loss. See figure 2. We find the optimal x_1=9 and x_2=93 (see Appendix 1).

Figure 2

[1] We also considered trying linear regression or sigmoid regression on the calendar, but we rejected those methods. The current model assumes: 1) the position of the closed-day chunks is random and independent of other listing features; 2) occupancy is 0 near x=end. (1) is reasonable since scrape_day randomizes the horizontal offset of the calendar. (2) makes our method a *conservative* estimation in extremely popular listings.

In this way, we calculate a rigid estimation of the occupancy rate for every listing. The next step is to bring number_of_reviews back into the picture. We run linear regression between number_of_reviews, length_of_stay and occupancy_rate (from calendar), filtering on the condition that last_review is recent enough[2]. The results show the probability of guests leaving a review is 38%. See Appendix 2. With this new information, we can now *reverse the direction* of this regression and predict occupancy_rate from the number_of_reviews.

Now we have two estimations for occupancy_rate, one from the calendar, and one from the reviews. Our final estimation is a weighted average of the two. Intuitively, at times when the calendar produces bad estimations, we should put a higher weight on the reviews estimation. Bad calendars are those with large chunks of closed days, and those with occupancy>0 near x=end.[3] A neural net is trained to look at the calendar[4] and predict the discrepancy between the calendar estimation and the review estimation. The predicted discrepancy is taken as the quality score for how bad the calendar is. See figure 3.

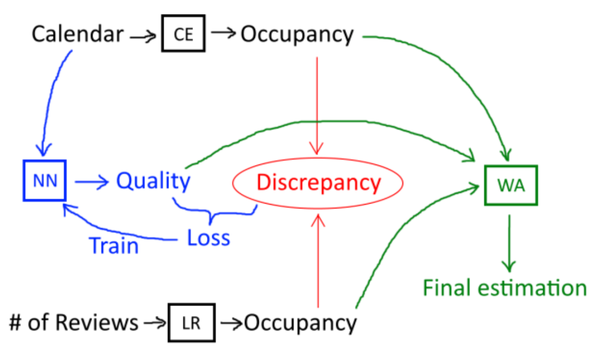


Figure 3.
CE: Calendar Estimation; NN: Neural Net; LR: Linear Model; WA: Weighted Average

[2] 30 / reviews_per_month gives the average review blank time. Multiply by 2 and we get the 95% confidence threshold that the listing is more inactive than usual.
[3] The bad features correspond nicely with the assumptions of the calendar model.
[4] Input includes: probability_unavailable, probability_closed, and the lengths of the longest three chunks of unavailable days.

The performance of our neural net[5] is evaluated using a Predicted vs Actual plot (figure 4). Alongside is a comparison with a linear regression model, whose inadequacy is visible to naked eyes. The usage of a neural net is thus justified because the linear model fails to capture the non-linear surface.

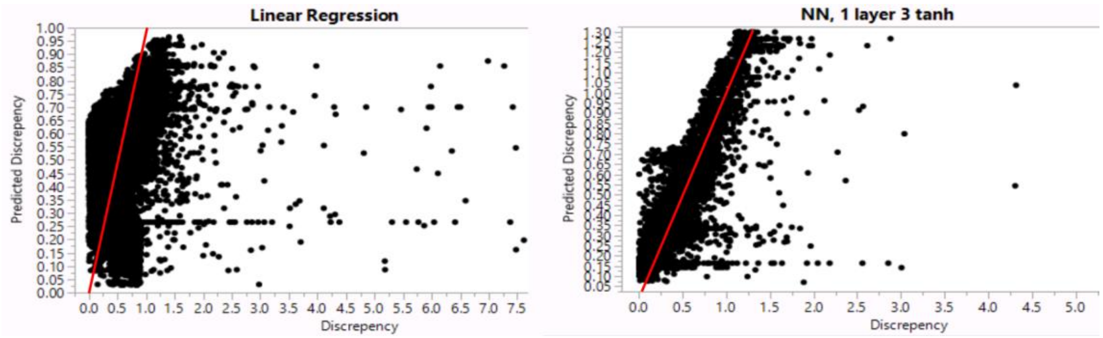


Figure 4. Predicted vs Actual, on validation set.

With the predicted discrepancy[6] as the quality score, we calculate the final estimation for occupancy rate:

```
occupancy_rate_final = (calendar_weight * occupancy_rate_calendar
                      + review_weight    * occupancy_rate_reviews
                       ) / (calendar_weight + review_weight)
```

Where calendar_weight = max(1 - predicted_discrepancy, 0),

and calendar_weight = constrain(1.5 - 0.5 * review_oldness_z_score, 0, 1)

The sum of the two weights is stored as a confidence indicator of our final occupancy estimation.

---

[5] We tried various NN structures. We found that giving the model more layers and parameters does not yield better results, so we settled with 1 layer of 3 tanh nodes to minimize room for overfitting. In the end, the validation set R-Square was higher than that of the training set, so we concluded there was no significant overfitting.

[6] Why not use actual discrepancy? If we used actual discrepancy, the final equation could be simplified (everything is linear!) to a weighted average, just as if we did not consider the quality score. It is also wrong because the calendar's quality score should only depend on the calendar, but not on both the calendar and the review counts. The NN-predicted discrepancy, on the other hand, does the right job. Without overfitting, the NN let no information from the review counts' side sneak into the predicted discrepancy.

2. Feature Cleaning

The dataset provides 105 different features for every listing. For us to fit a model to predict demand, we shall try to lower the number of x variables. There are three purposes of doing so: 1) lower the computation time, 2) avoid overfitting, and 3) avoid co-linearity.

The first thing we do is to exclude 34 obviously irrelevant features: listing_url, scrape_id, last_scraped, host_id, host_url, host_name, state, market, country_code, country, calendar_updated, availability_30, availability_60, availability_90, availability_365, calendar_last_scraped, number_of_reviews, number_of_reviews_ltm, first_review, last_review, review_scores_rating[7], review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, reviews_per_month, and 6 image urls. We discard notes, access because it has too many missings. This leaves us with 69 features.

Then, we remove duplicate information. We remove neighbourhood, latitude, longitude, smart_location, is_location_exact, and keep neighbourhood_cleansed and neighbourhood_group_cleansed. We remove minimum_minimum_nights, maximum_minimum_nights, minimum_maximum_nights, maximum_maximum_nights, minimum_nights_avg_ntm, maximum_nights_avg_ntm, and keep minimum_nights, maximum_nights. We remove requires_license and keep license. We remove calculated_host_listings_count_entire_homes, calculated_host_listings_count_private_rooms, calculated_host_listings_count_shared_rooms, and keep calculated_host_listings_count. This further reduces the number of columns to 54.

After inspecting the distribution of each feature, we find that some features have extremely skewed distributions. Therefore, we further discard license, experiences_offered, host_neighbourhood, square_feet, has_availability, jurisdiction_names, is_business_travel_ready. Also, Host_listings_count and host_total_listings_count are always equal, so we discard the latter. Amenities

[7] The reviews data are excluded because as a host, when we try to optimize our features, we do not want to hold the review scores as constant. The review scores can be a result of our features.

need to be expanded into 30 columns (See Appendix 3). This way, we have 75 features left to deal with.

Host_response_rate, host_acceptance_rate needs additional cleaning ("23%" becomes "0.23") to be parsed as numerical. Host_verifications is converted to num_of_host_verifications. Weekly_price and monthly_price are converted[8] into weekly_discount and monthly_discount. Missing values are filled with ones for the following variables: host_listings_count, host_total_listings_count, bedrooms, beds. Missings are filled with zeros for Security_deposit, cleaning_fee. Bathrooms missings are left as is. (Maybe leaving missing data could be a host strategy)

Lastly, we examine extreme outliers.[9] Apparently one host put the largest integer that can be stored in a 32-bit computer word (2,147,483,647) as maximum_nights:

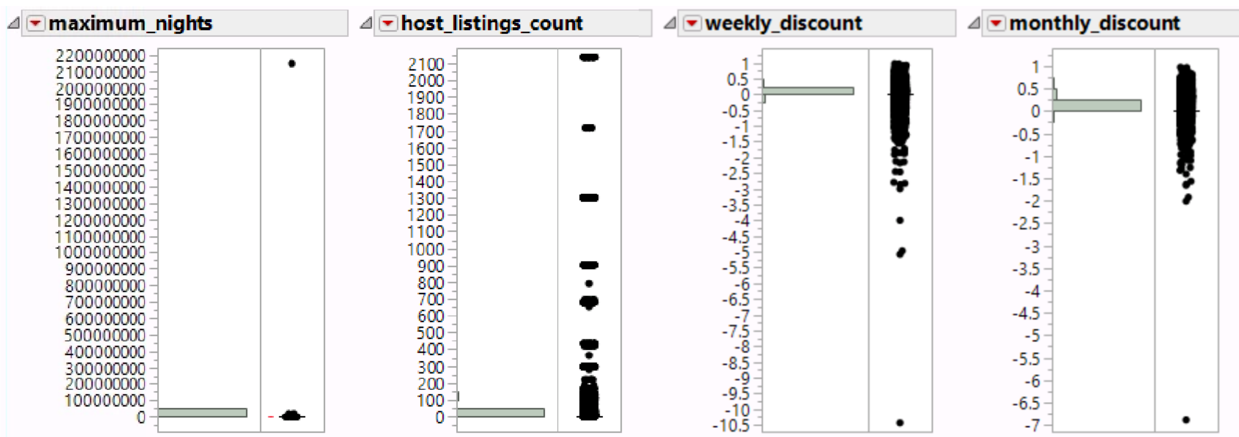


Figure 5. Strange Distributions

We set a cap of 2000 to maximum_nights. The distribution of host_listings_count also looks funny, but we decide it just needs a log transform, so we keep the outliers. Weekly_discount and monthly_discount, to our surprise, have negative values. Evidently, some hosts set the weekly

---

8 For example, $10 a day, $63 a week would yield weekly_discount = 10%. Missing values in weekly_price yields weekly_discount = 0. This transformation helps, because instead of tasking the machine to interpret *informative missings*, we interpret them correctly beforehand (using common sense) for the machine.

9 We tried PCA and PCA gave an unreasonably high loading to maximum_nights. Only then did we realize we forgot to clean outliers.

prices such that it is actually more expensive than the per-night prices. Albeit strange, we decide that it is a valid feature of a listing. We thus set a lower cap of -1 (you pay double!) for Weekly_discount and monthly_discount.

The location originally has 224 levels. We group the minority locations into "other" and end up having 45 levels. See figure 6.

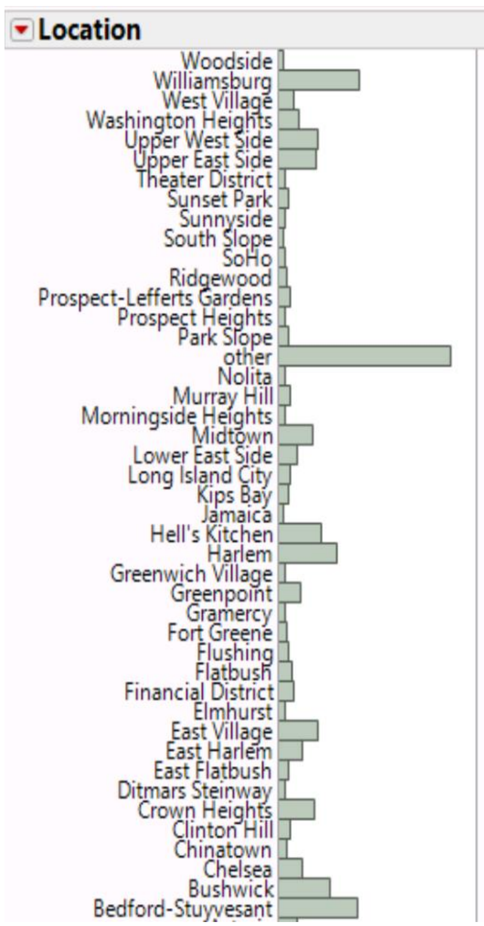


Figure 6. Grouping the Minority Locations into "other"

3. Text Mining

The data contain many text features. We must extract quantitative features from the texts. Here is how we handle each text feature.

Text Variable	Word count	Sentiment	Topics
name	✓		
summary	✓	✓	✓
space	✓	✓	✓
description	✓	✓	✓
interaction	✓	✓	✓
host_about	✓	✓	✓
neighborhood_overview	✓		
transit			✓
house_rules			✓

For name, we also generate is_all_cap and average_word_length. (Spoiler alert: According to section 6, uppercase letters do not attract guests at all…) Average_word_length is supposed to reflect how sophisticated the listing name is.

We first clean the text data. It is particularly annoying to see *multiple emojis not separated by space*, since every unique string of emojis are regarded as a unique word in the corpus.

Figure 7. A list of emojis displayed incorrectly in R terminal

We use library "textclean"[10] to resolve this issue. We replace emojis with word equivalences and remove non-ASCII characters. We replace word elongation ("soooo good" becomes "so good").

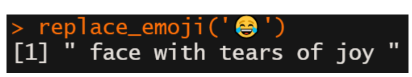


Figure 8

[10]github.com/trinker/textclean.

We then inspect for outliers (figure 9). The word count of summary is not normally distributed, so we add a column denoting whether the summary is empty, in case a linear model couldn't handle its nonlinearity. The word count of host_about has extreme outliers, so we cap the value at 300. Similarly, name_avg_word_len is capped at 14.

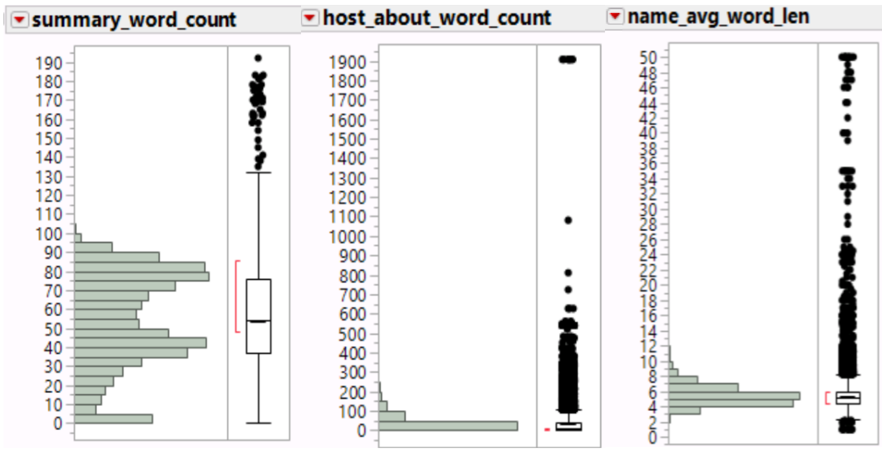


Figure 9

Most text fields are neutral in terms of sentiment; some are mostly positive.

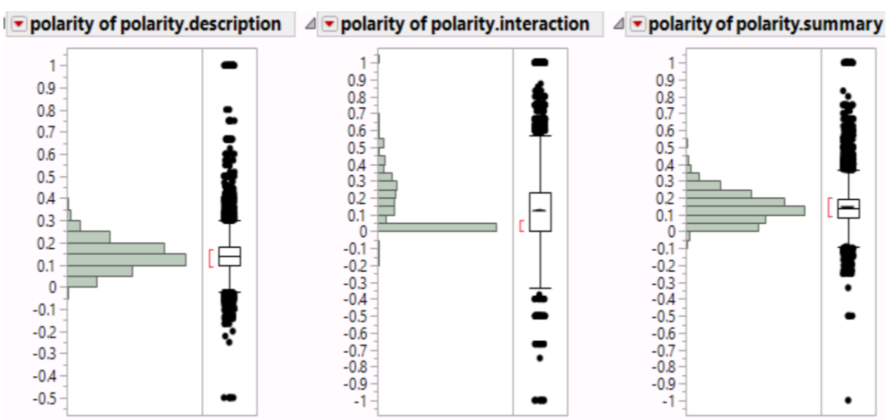


Figure 10

For topic modeling, we use top ~500 terms and the Gibbs method with 5000 iterations to find 20 topics for each text field. We identify the topics in long-appendix B.

Figure 11 shows the word cloud for listing description. See long-appendix A for more word clouds.

Figure 11

We end up having 155 columns of features related to text. Adding that to the $75 - 9 = 66$ non-text features, our model is finally ready to run with 221 columns!

4. Principal Component Analysis

The next step is Principal Component Analysis (PCA). We create a temporary copy of the data table. We exclude nominal variables that are unusable. We recode some nominal variables to continuous numerical values. The results show that PCA manages to summarize 57 columns to 15 factors while still preserving 59% of overall diversity. See Appendix 4.

Although the rotated factors are interesting to look at, we do not intend to use the rotated vectors as model input to fit a demand surface. Because the project goal is to provide guidance for the host to improve the listing's features, we do not want to add a layer of indirectness to the insights that the final model will give us.

Next, we add text mining results to the PCA. This time, 205 features can be summarized by 66 factors. The rotated factor loading matrix shows that some amenities are correlated with certain topics in a very sparse manner. We are amazed to see this result. See figure 12.

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
10 of topic.host_about	0.015779	-0.138263	-0.146081	0.000226	-0.095277	-0.027224
11 of topic.host_about	-0.050587	-0.160128	-0.175685	-0.000582	-0.138113	0.036848
12 of topic.host_about	-0.011948	0.946384	-0.061036	-0.004615	-0.080836	0.043737
13 of topic.host_about	0.004680	-0.160531	-0.184719	0.105075	-0.152102	-0.012345
14 of topic.host_about	0.031068	-0.041574	-0.050528	-0.011130	0.922993	0.030869
15 of topic.host_about	0.031106	-0.149644	-0.124791	0.016331	-0.123398	0.025435
16 of topic.host_about	-0.041408	-0.150718	-0.142307	-0.050432	-0.127637	-0.035766
17 of topic.host_about	0.077107	-0.089598	0.971836	0.000682	-0.086019	-0.001859
18 of topic.host_about	-0.023223	-0.161055	-0.138380	0.003673	-0.093842	0.023110
19 of topic.host_about	-0.014668	-0.103168	-0.117730	0.016259	-0.093658	0.031538
20 of topic.host_about	-0.003790	-0.123026	-0.106529	-0.009416	-0.117095	-0.008499
1 of topic.house_rules	-0.020735	-0.080033	-0.148126	-0.011159	-0.060433	0.040648
2 of topic.house_rules	0.002445	-0.122279	-0.129417	-0.025149	-0.059022	0.035671
3 of topic.house_rules	-0.008097	0.922312	-0.058880	0.003118	-0.053187	0.066842
4 of topic.house_rules	-0.039868	-0.164434	-0.141768	0.090420	-0.112880	0.044288
5 of topic.house_rules	0.036710	-0.121751	-0.148895	-0.035112	-0.038501	0.044221
6 of topic.house_rules	-0.010528	-0.015005	-0.133745	-0.037120	-0.061024	-0.169443
7 of topic.house_rules	0.006722	-0.076932	-0.089419	-0.034644	0.883227	0.038604
8 of topic.house_rules	-0.010105	-0.118424	-0.130161	-0.120096	-0.021164	0.017248
9 of topic.house_rules	0.016281	-0.049275	-0.084333	0.044271	-0.073296	-0.038669
10 of topic.house_rules	-0.061892	-0.165402	-0.168740	0.060844	-0.051588	-0.014755
11 of topic.house_rules	-0.049449	-0.142978	-0.149942	0.027847	-0.047918	0.090976

Figure 12

Another interesting observation is that the hosts whose listings have all-cap names are extraneous and usually have many verifications. Unfortunately, because many text fields are optional for the host, 80% of the columns have missing data; so this PCA result is unsuitable for cluster analysis. In the next section, we use the previous 15-factor PCA results to perform clustering.

5. Hierarchical Clustering

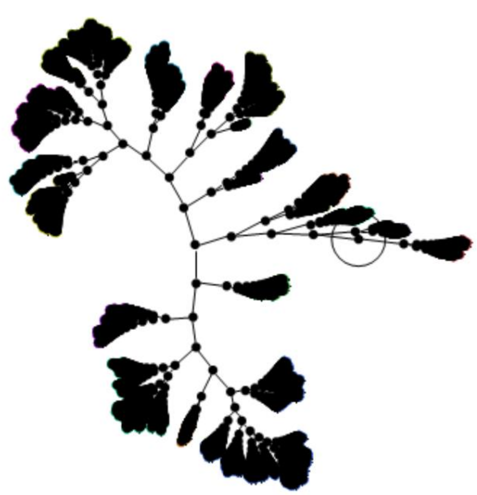


Figure 13. Constellation Plot.

The cluster analysis suggested 20 clusters. See figure 13 for the constellation plot and Appendix 5 for the dendrogram.

6. Demand Curve and Recommendations for the Host

To answer the research question, we need to find the demand curve.[11] The demand curve predicts occupancy rate when supplied with prices and features. The shape of the demand curve then informs the host 1) what features are important and 2) what price is optimal.

We first try a linear regression without regularization. 20% of the dataset is allocated to the validation set. Regression is weighted by the confidence attached to occupancy prediction.

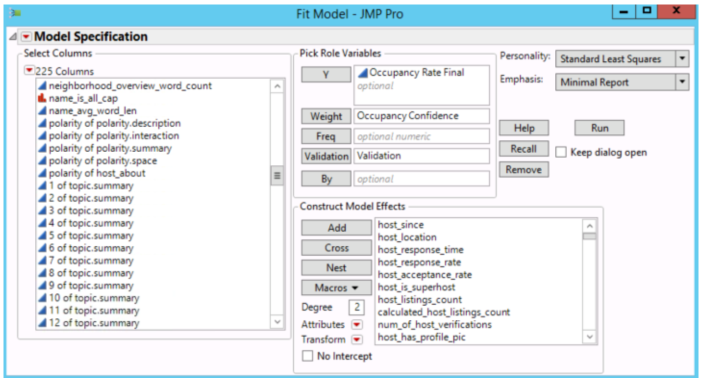


Figure 14

We encounter an "ill-conditioned regression" warning. That means there are perfectly correlated model effects. However, when we run a multivariate analysis of the 172 continuous model effects, we find no perfect correlations. We then look at the singularity report and realize there is co-linearity since the distribution of 20 topics must sum to 1. Therefore, we exclude topic 20. We also exclude host_location, zipcode, city and street because they somehow correlate with neighborhood_cleansed. We discard minimum_nights because we used it to estimate occupancy rate. We now have 209 model effects left.

We apply log transformation to price, accommodates, bathrooms, bedrooms, beds, security_deposit, cleaning_fee, guests_included, and extra_people. Informative missing is

---

[11] Due to its high dimensionality (200+), the demand curve we are solving for may as well be called "demand surface".

turned on, and indicator function parameterization is turned off. We multiply occupancy_rate by 365.25 so that the model parameters would be easier to comprehend.

We run the regression. It fits with adjust RSquare = 19%. Residuals look pretty normally distributed except 38 positive outliers. See figure 15 and long-appendix C.

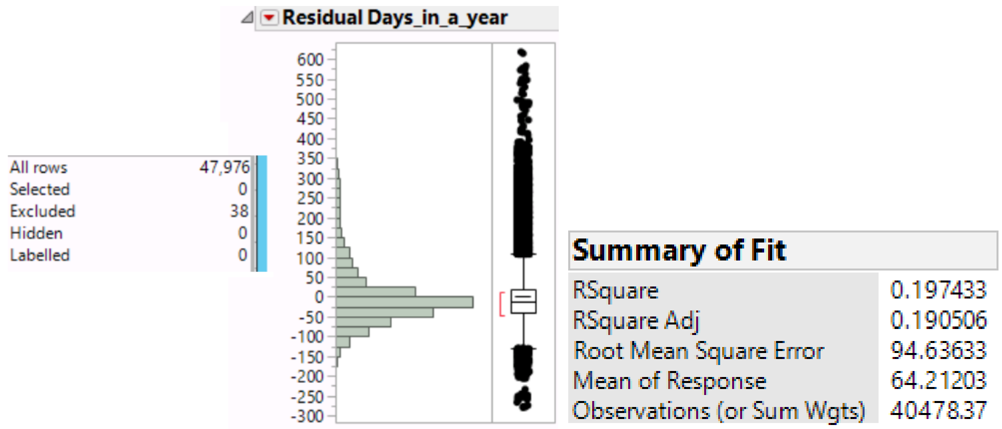


Figure 15

Next, we will look at the significant model effects one by one, and offer suggestions to a host looking to list a place on Airbnb. However, the host should ***beware of the confounder effect***. For example, it may appear that a positive sentiment in listing description helps; however in reality, a host only writes positive-sentiment description if the place is nice and they feel proud of it. Faking a positive sentiment in the description without having a nice place may be ineffective in attracting guests.

Log(price). -9.9 days/year. P-value: <.0001
The price sensitivity. For every 11% increase in price, the number of days booked per year decreases by 1. This sensitivity is lower than expected. There must be model effects not captured in the dataset that positively correlate with price and sales.
Host_since. 1 year newer → +0.9 days/year. P-value: 0.0011
HOVAC, if the host is 1 year newer, the number of days booked per year increases by 0.9 on average.
Host_acceptance_rate. Missing → -19.7 days/year. P-value: <.0001
This is a confounder effect. If the listing is bad, then nobody books it, so the host acceptance rate is missing.
Host_is_superhost. Yes → +11.5 days/year. P-value: <.0001

Being a super host may be effective on boosting sales. Or, maybe the causality is in reverse?
Log(calculated_host_listing_count). +9.9 days/year. P-value: <.0001
If you double the listing count, yearly sales +6.9 days. This may also be confounded.
Location[Flushing]. -12.7 days/year. P-value: 0.0297
Location[Greenwich Village]. +12.9 days/year. P-value: 0.0238
Location[Harlem]. -6.4 days/year. P-value: 0.0312
Location[Morningside Heights]. +34.5 days/year. P-value: <.0001
Location[Ridgewood]. -12.0 days/year. P-value: 0.0458
Location[Theater District]. +27.0 days/year. P-value: 0.0001
Location[Woodside]. -14.2 days/year. P-value: 0.0368
All locations are compared against "other".
Neighborhoob_group_cleaned[Brooklyn]. -5.8 days/year. P-value: 0.0098
HOVAC, Manhattan listings have higher demand compared to Brooklyn!
Log(Accommodates). -3.6 days/year. P-value: 0.0236
The more people you accommodate, the less demand there will be? This may be wrong.
Log(beds). +5.4 days/year. P-value: 0.0017
For every 10% more beds you have, you can increase your price by 5.5% and the demand will not change... The underlying confounder here, however, should be the size of the room. Increasing the size of the room and having more beds significantly raises the demand!
Log(security_deposit). +5.5 days/year. P-value: <.0001
This is, again, confounded. The better the place, the higher the demand, as well as the average security deposit.
Cleaning_fee. Missing → +3.0 days/year. P-value: 0.0227
If you don't collect a cleaning fee, there will be a higher demand.
Log(guests_included). +5.0 days/year. P-value: 0.0004
For every 22% more guests you include with the offer, the demand grows by 1 day/year.
Log(extra_people). -3.3 days/year. P-value: 0.0030
Extra_people is the fee you collect for every additional guest. If you collect a 35% higher fee, the demand drops by 1 day/year.
Room_type[Shared room]. -13.2 days/year. P-value: <.0001
Compared to a full house/apartment, offering a shared room lowers the demand.
Amen: Air conditioning. Yes → +2.3 days/year. P-value: 0.0018
Amen: Hangers. Yes → 3.3 days/year. P-value: <.0001
Amen: Carbon monoxide detector. Yes → +1.6 days/year. P-value: 0.0138
Amen: Shampoo. Yes → -2.4 days/year. P-value: <.0001
Whoever mentions shampoo in the listing must have a bad place? We suspect some sort of confounder effect here.
Amen: Hair dryer. Yes → +1.6 days/year. P-value: 0.0130
Amen: Iron. Yes → +1.3 days/year. P-value: 0.0451
Amen: Hot water. Yes → -2.0 days/year. P-value: 0.0050

Again, if you have to mention hot water, you probably don't have anything else.
Amen: Refrigerator. Yes → +3.4 days/year. P-value: 0.0055
Amen: Dishes and silverware. Yes → +2.1 days/year. P-value: 0.0383
Amen: Fire extinguisher. Yes → +2.4 days/year. P-value: <.0001
Amen: Lock on bedroom door. Yes → -1.4 days/year. P-value: 0.0171
This suggests it's a shared apartment, hence lower demand.
Amen: Free street parking. Yes → +3.7 days/year. P-value: <.0001
Amen: Internet. Yes → +2.7 days/year. P-value: <.0001
To sum up, it helps when you have more amenities.
Weekly_discount. -23.9 days/year. P-value: <.0001
For every 10% discount you offer if the guests book in bulks of weeks, you lose 2.4 days/year of sales? No, that is false. The confounder here is whether the listing is self-consistent or not.
Monthly_discount. +33.1 days/year. P-value: <.0001
For every 10% discount you offer if the guests book in bulks of weeks, you win 3.3 days of sales per year!
Maximum_nights. +.007 days/year. P-value: <.0001
It is not clear what this means.
Instant_bookable. Yes → +1.5 days/year. P-value: 0.0062
Making your listing instantly bookable creates value for your guests.
Cancellation_policy[flexible]. +18.6 days/year. P-value: 0.0015
Cancellation_policy[moderate]. +18.7 days/year. P-value: 0.0013
Cancellation_policy[strict_14_with_grace_period]. +22.5 days/year. P-value: <.0001
Compared to having a strict cancellation policy, it is better to be flexible.
Space_word_count. +.015 days/year. P-value: 0.0227
Consider talking about space! If you leave this field blank, you may be missing out a lot of guests.
Polarity.interaction. -7.8 days/year. P-value: 0.0381
In the "interaction" text field of the listing, positive sentiment leads to lower demand. This is consistent with Interaction topic 13. See below.
Summary topic 8 explore NYC. -90 days/year. P-value: 0.0157
Summary topic 15 kitchen. +107 days/year. P-value: 0.0033
Have a kitchen, and talk about it in summary!
Summary topic 17 high-rise. -83 days/year. P-value: 0.0302
Transit topic 2 generic transit. -102 days/year. P-value: 0.0150
Interaction topic 2 host family lives here. +200 days/year. P-value: 0.0023
Interaction topic 13 host is your friend. -176 days/year. P-value: 0.0058
Compare interaction topic 2 and 13. Be hospitable, but don't overshoot! This is consistent with "polarity.interaction". See above.
Description topic 5 generic description. +93 days/year. P-value: 0.0095
Description topic 10 kitchen towel microwave coffeemaker. +84 days/year. P-value: 0.0113
Host_about topic 1 lives in NYC. +242 days/year. P-value: <.0001

Host_about topic 2 welcome foreigners. +110 days/year. P-value: 0.0172
If you like to learn about foreign cultures.
Host_about topic 3 family here. +140 days/year. P-value: 0.0038
Host_about topic 5 Manhattan. +82 days/year. P-value: 0.0375
Host_about topic 10 will help. +187 days/year. P-value: <.0001
Host_about topic 12 tech. +207 days/year. P-value: <.0001
Host_about topic 13 Brooklyn couple (+dog). +95 days/year. P-value: 0.0446
Host_about topic 14 cares. +165 days/year. P-value: <.0001
Host_about topic 19 generic host about. +123 days/year. P-value: 0.0027
House_rules topic 2 host lives here, just ask. +112 days/year. P-value: 0.0032
House_rules topic 3 pet allowed. +173 days/year. P-value: <.0001
Allow pets! Compare this to cleaning fee (+3 days/year). You'd rather allow pets and collect cleaning fee.
House_rules topic 4 smoking. +148 days/year. P-value: 0.0002
House_rules topic 6 tenant. +168 days/year. P-value: <.0001
House_rules topic 8 share kitchen / bathroom. +88 days/year. P-value: 0.0241
House_rules topic 9 late fee. +231 days/year. P-value: <.0001
House_rules topic 16 plz clean dishes. +155 days/year. P-value: <.0001
House_rules topic 17 ???. +124 days/year. P-value: <.0001
House_rules topic 18 plz recycle and take trash out. +108 days/year. P-value: 0.0024
House_rules topic 19 fees and taxes. +100 days/year. P-value: 0.0023
Space topic 1 towel coffeemaker microwave dryer stove oven iron. +83 days/year. P-value: 0.0195
Space topic 10 fully equipped. +81 days/year. P-value: 0.0305
Space topic 11 clean home. +165 days/year. P-value: <.0001
Space topic 12 queen-sized bed. +118 days/year. P-value: 0.0039
Again, merely talking about a topic may not be the underlying causal effect; your place must actually have what the topic describes. Please see topic identification in long-appendix B. The coefficients are very big because the topic distributions are usually << 1. For example, "+100 days/year" means if you devote *all* the words in your summary to this one topic (which is impossible), you get +100 days/year.

The model also gives us the profiling ability. A host can input their specific features and plot a demand curve of sales vs price. For illustration purposes, here we solve for the optimal price assuming all other features are the population median. The optimal price that maximizes revenue turns out to be $160,000 per night. For comparison, the average price of NY Airbnb listings is $100 per night. We significantly underestimated the demand's sensitivity to prices! There must be model effects not captured in the dataset that positively correlate with price and sales. Although this result is disappointing, we know we did our best and we hope the other conclusions are still somewhat effective.

7. ML Models and Metrics

This section is unrelated to the research question. We use Random Forest, Boosted Tree, and NN to fit the demand curve. We then compare their prediction power.

We do not use these models to answer the research question for two reasons. 1) A lot of these models do not have human-interpretable parameters. 2) They do not offer p-value and are prone to overfitting. A simple linear model, on the other hand, gives conservative and interpretable insights. The linear regression presented in the last section performs very well on the validation set, so there is no need to run an elastic net model, as long as we only focus on those statistically significant coefficients.

The way we assess the performance of the ML models differ from what we learnt from the class (ROC, lift curve...). This is because instead of a classification problem, our ML models try to predict a continuous value.

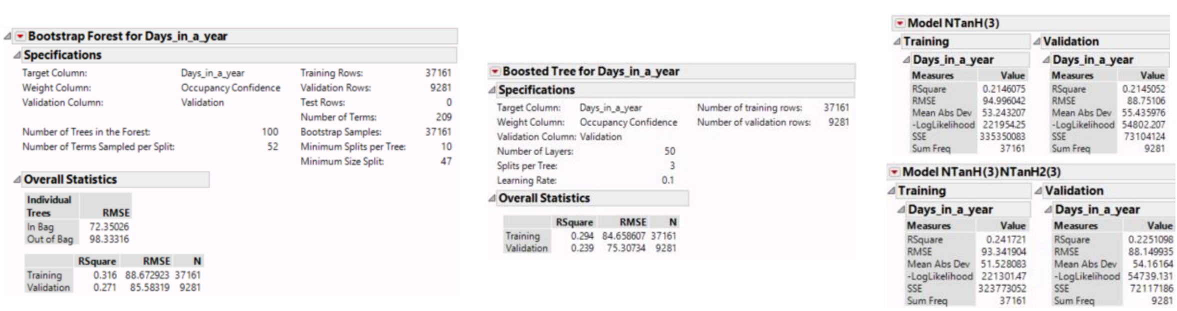


Figure 16. ML Model Results

Table 1. Assess ML Performance

	Linear	Random Forest	Boosted Tree	NN1	NN2
RSquare, Training Set	20%	32%	30%	21%	24%
RSquare, Validation Set	26%	27%	24%	21%	23%
Comment	Baseline.	100 trees.	50 layers.	3 tanh.	3 + 3 tanh.

In terms of prediction power, the linear regression and random forest do the best job.

8. Future Work

We are very pleased with our results. However, there are many improvements we can make. Here are a few.

Remove outliers. In this project, we remove outliers only when it occurs to us. We often forget to remove outliers. For the study to be more rigid, one should always remember to remove outliers.

Cleanse amenities. The amenities values are provided by the hosts and are not cleansed. For example, instead of {“Pets allowed”, “Wifi”} (2 items), one host listed {“Pets allowed, Wif”} (1 item, typo). We suspect that Airbnb provides some tags for the hosts to pick and also allows the hosts to create new tags, and hence there are typos and mistakes. If we had more time, we could use edit distance to smartly interpret the tags the hosts provide. We could also enlist typos as a model effect and see whether typos affect your sales.

Analyze images. The dataset contains urls to room photos and profile images. It would be helpful if we can extract quantitative data from the images.

Take out confounded variables.

9. Questions

This section lists out questions we need help with.

Adaptive Elastic Net with Validation Column

Model Summary

Response	Occupancy Rate Final
Distribution	Normal
Estimation Method	Adaptive Elastic Net
Validation Method	Validation Column with Early Stopping
Mean Model Link	Identity
Scale Model Link	Identity

Measure	Training	Validation
Number of rows	37161	9281
Sum of Frequencies	40478.365	10107.624
-LogLikelihood	478112283	80227715
Number of Parameters	45	45
BIC	956225044	160455845
AICc	956224656	160455521
RSquare	0.1744049	1

Estimation Details

Elastic Net Alpha	0.9
Number of Grid Points	150
Minimum Penalty Fraction	0.01
Grid Scale	Linear

Validation RSquare = 1? What's happening here?

10. References

[1] Inside Airbnb. San Francisco Model. insideairbnb.com/about.html, Section "The Occupancy Model".

[2] Brian Chesky. Review Rate. Quora. qr.ae/pNn4gn.

[3] Airbnb Economic Impact. blog.atairbnb.com/economic-impact-airbnb/#new-york.

11. Appendices

All source code can be found at github.com/Daniel-Chin/airbnb.

Appendix 1. Optimize x_1 and x_2

We tried two methods to optimize x_1 and x_2. The first method (legacy method) failed. The second method yields satisfying results.

Here goes the legacy method. For any given x_1, we can pretend we do not know the availability of x=1 and predict probability_unavailable for x=1. The goodness of x_1 is thus indicated by the cross-entropy loss on x=1. With the optimal x_1, we then make x=1 visible to the algorithm and predict probability_unavailable for x=0. See figure 17.

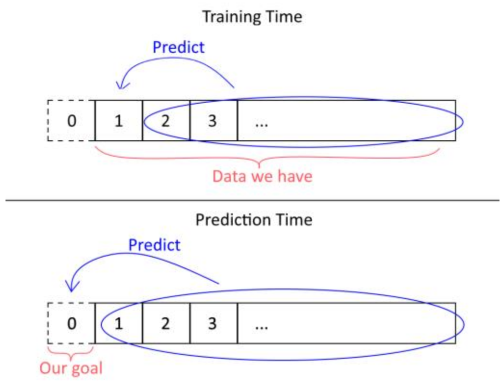
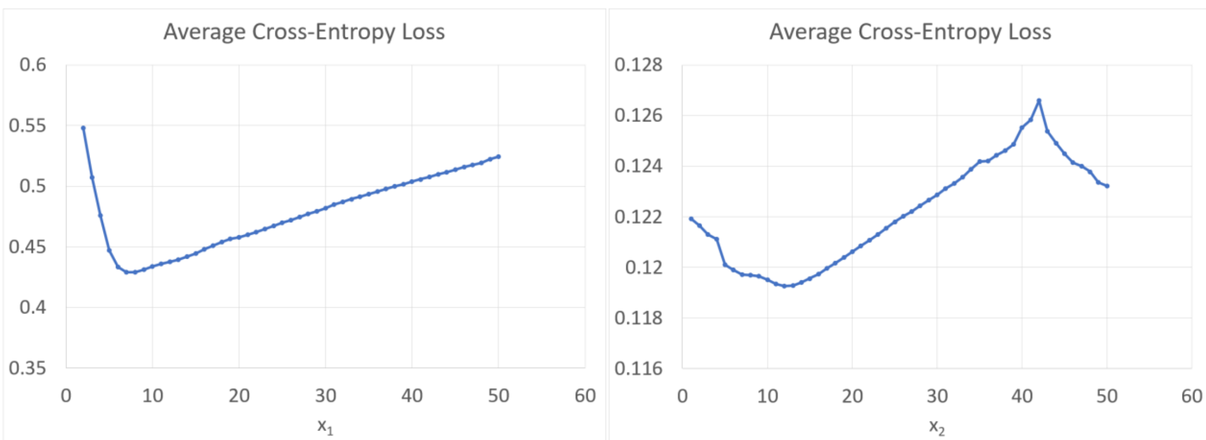


Figure 17

x_1=6 yields minimum average loss 0.429, x_2=12 yields minimum average loss 0.119

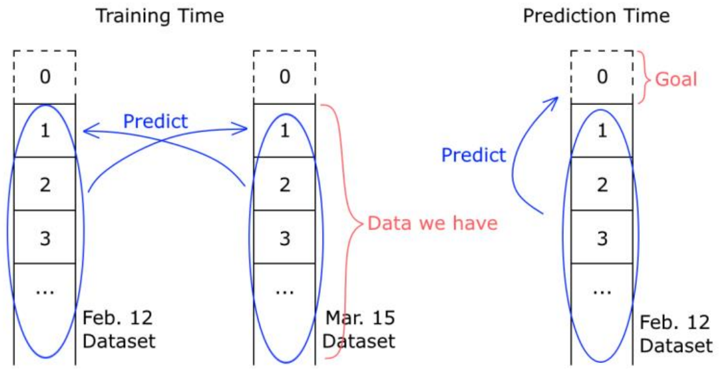
When |y_hat - y|=1, we adjust the loss from infinity to 5.

Source code is at ./demandAnalysis/x1x2.py

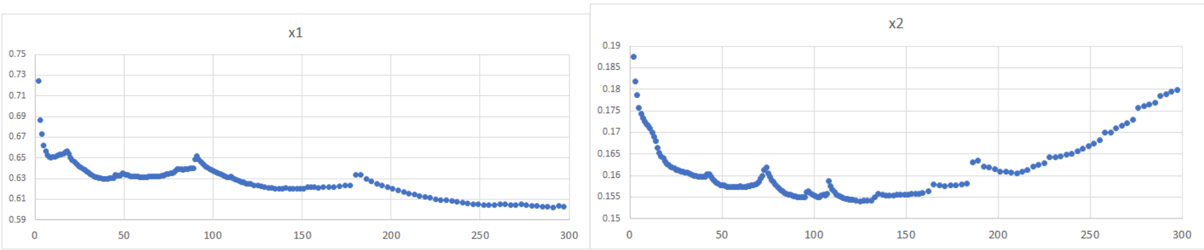


However, the occupancy estimated by the legacy method has a very low R-square with review scores. Unsatisfied, we look for reasons. Closer analysis reveals a serious mistake: this model assumes availability is independent across days. In reality, because days are booked in consecutive lumps, nearby days have positively correlated availability. The ML model can easily exploit that fact and give a very small x_1.

To fix that, we waited one month and downloaded a newer version of the calendar availability data. The ML method is described in the main text. Note that with the extra dataset, there can now be many better methods of estimating occupancy; however, we want to retain our project scope of using only the Feb. 12 dataset, so we do our best to leave the extra data untouched. Consider the extra data we use to obtain optimal x_1 as a "one-time local hack".



Plotting average loss against x_1 x_2:



We set x_2=93. For x_1, we take 9, 39, 135, 258, 291 as four candidates. To assess their validity, we check the correlation between number_of_reviews and occupancy_rate predicted by the calendar using x_1. Legacy method gives R-Square = .0051. The improved method gives the following results:

x_1	9	39	135	258	291
R-Square	.184	.139	.134	.123	.121
Estimated Review Rate	37%	38%	43%	49%	51%

We set x_1=9 to maximize R-Square.

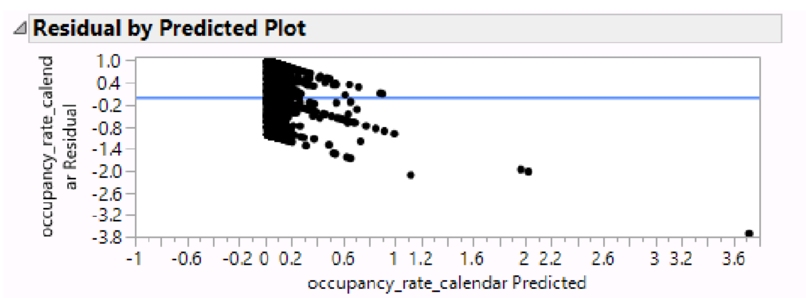
Source code is at ./demandAnalysis/m2_x1x2.py

Appendix 2. Linear Regression Model between number_of_reviews and occupancy_rate

We first calculate the average length_of_stay per listing. According to Airbnb, "visitors stay on average 6.4 nights" in New York [3], which we use as the baseline. Then, we observe that the acceptable minimum_nights is <7 for 41180 listings and >=7 for 9917 listings. We use the minimum_nights for those >=7. For those <7, we use 6.4. We initially wanted to adjust them so that the overall average remains 6.4, but the adjusted average turned out to be 1.031 if we assume equal weight to all the listings. This is obviously unrealistic, so we decided to use 6.4 instead. See discussion and source code at ./demandAnalysis/length_of_stay.py

Dividing occupancy_rate by length_of_stay, we get y. Remove outliers. Run a regression from review_per_day (calculated) to y. The probability of guests leaving reviews turns out to be 38.3% ± 1.2%. It is significantly lower than [1] and [2].

A linear regression between the two versions of occupancy_rate predicted leaves residuals:



This does not look too good; however, the relation between review counts and occupancy is, by nature, linear. Therefore, we are happy with the linear model.

Appendix 3. Most Common Amenities

Wifi	49830
Heating	48124
Essentials	47281
Kitchen	46473
"Smoke detector"	44595
"Air conditioning"	43493
Hangers	38471
"Carbon monoxide detector"	36249
TV	35086
Shampoo	34568
"Hair dryer"	34045
"Laptop friendly workspace"	32520
Iron	32353
"Hot water"	28744
Refrigerator	22867
"Dishes and silverware"	22132
Washer	20821
Dryer	20463
"Fire extinguisher"	20056
Microwave	19339
"Lock on bedroom door"	19046
Stove	18960
"Cooking basics"	18924
Oven	18553
"Free street parking"	18496
"Coffee maker"	17698
"First aid kit"	17583
"Bed linens"	16954
Internet	13857
Elevator	13631

For the sake of efficiency, we take the 30 most common amenities tags.

Appendix 4. PCA Results

Row	F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8	F 9	F 10	F 11	F 12	F 13	F 14	F 15
amen_"Hair dryer"	2	1	6	0	0	0	0	1	0	0	0	-1	-1	0	1
amen_Iron	2	1	6	0	0	1	0	1	1	0	0	0	-1	0	1
amen_Hangers	2	0	5	0	0	0	0	2	1	0	1	0	-1	0	1
amen_Shampoo	1	0	5	0	0	0	0	1	1	0	-1	-1	0	1	0
amen_"Laptop friendly workspace"	2	1	4	0	0	-1	0	0	1	0	1	-1	0	0	1
amen_Essentials	1	0	4	0	0	0	0	1	1	0	0	-1	0	0	0
amen_"Air conditioning"	1	1	3	1	0	0	0	-1	0	0	0	1	1	0	-1
amen_Heating	1	0	3	0	0	0	0	0	1	0	0	0	0	0	0
amen_"Hot water"	6	0	3	0	0	-1	0	1	1	0	1	0	-2	1	1
amen_TV	1	1	3	1	0	1	0	-1	1	0	0	2	1	1	-1
amen_"Smoke detector"	1	0	3	0	0	0	0	0	7	0	0	0	1	0	-1
amen_Wifi	0	0	2	0	0	0	0	-1	1	0	0	0	1	0	-1
amen_"Carbon monoxide detector"	2	1	2	0	0	0	0	1	7	0	1	0	0	0	0
amen_"First aid kit"	1	1	2	0	0	-1	0	0	3	0	-1	0	-1	1	1
amen_"Coffee maker"	7	1	2	0	0	0	0	0	1	0	0	0	0	0	-1
amen_"Bed linens"	6	0	2	0	0	1	0	1	1	0	1	0	-1	1	1
amen_"Fire extinguisher"	2	1	2	0	0	0	0	0	3	0	0	1	-1	1	1

host_is_superhost	3	0	2	0	1	0	0	0	1	0	0	-1	-1	1	0
amen_"Dishes and silverware"	9	0	2	0	0	-1	0	0	0	0	1	0	0	0	-1
amen_"Free street parking"	5	1	2	-1	0	-2	0	0	1	0	0	-1	-2	0	0
cancellation_policy	2	1	2	0	1	0	0	-1	0	0	3	0	0	1	0
cleaning_fee	2	4	2	1	0	3	0	0	0	0	5	0	1	1	-1
amen_"Cooking basics"	9	0	2	0	0	-1	0	0	0	0	1	0	2	0	0
amen_Refrigerator	9	0	1	0	1	1	0	1	1	0	0	0	0	0	0
amen_Microwave	8	1	1	0	1	1	0	1	1	0	0	0	-1	0	-1
accommodates	1	8	1	0	0	0	0	0	0	0	0	0	1	2	-1
host_acceptance_rate	2	0	1	-1	4	0	0	4	0	0	-1	-1	0	1	-1
amen_Stove	9	0	1	0	0	0	0	0	1	0	0	0	3	1	1
amen_Oven	9	0	1	0	0	1	0	0	1	0	0	0	3	0	2
amen_"Lock on bedroom door"	1	-1	1	-1	0	-2	0	2	2	0	-1	0	-2	0	2
host_response_rate	1	0	1	0	9	0	0	0	0	0	0	-1	0	0	0
beds	1	8	1	0	0	0	0	0	0	0	0	0	0	0	0
guests_included	1	5	1	0	0	0	0	0	0	0	1	0	0	4	-1
amen_Dryer	0	1	1	10	0	1	0	0	0	0	1	0	0	0	0
amen_Washer	0	1	1	10	0	1	0	0	0	0	1	-1	0	0	0
amen_Elevator	-1	-1	1	4	0	2	0	-1	0	0	0	1	1	0	0
amen_Kitchen	1	1	1	1	0	0	0	-1	0	0	0	-1	4	-1	0
instant_bookable	0	0	1	0	1	1	0	5	0	0	-1	0	0	0	-1
extra_people	1	1	1	0	0	-1	0	0	1	0	2	-1	-1	4	0
calculated_host_listings_count	1	0	1	1	0	8	0	1	0	0	0	0	0	0	0
host_since	0	0	1	0	0	1	-2	6	0	-1	-1	5	0	0	0
num_of_host_verifications	1	0	0	0	0	1	0	-1	0	0	0	-4	0	0	0
host_listings_count	1	0	0	1	0	8	0	1	0	0	1	-1	0	0	0
security_deposit	0	1	0	1	0	0	0	-1	0	0	5	0	0	1	0
price	0	2	0	0	0	0	0	0	0	0	1	1	0	0	0
require_guest_phone_verification	0	0	0	0	0	0	9	-1	0	0	0	0	0	0	0
require_guest_profile_picture	0	0	0	0	0	0	9	-1	0	0	0	0	0	0	0
host_identity_verified	-1	0	0	0	0	-1	0	-4	0	0	0	-4	0	0	-1
bedrooms	1	8	0	0	0	-1	0	0	0	0	1	-1	0	-1	1
bathrooms	0	5	0	1	0	0	0	1	1	0	1	0	0	-1	1

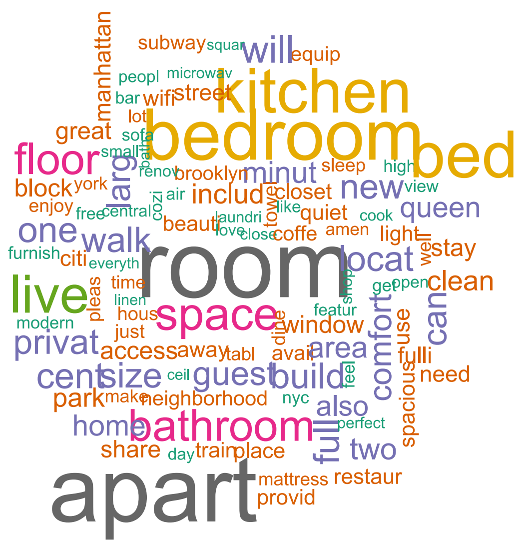
host_has_profile_pic	0	0	0	0	0	0	0	0	0	0	0	-1	0	0	0
amen_Internet	-1	0	0	0	0	0	1	-5	0	1	0	-1	1	0	-1
maximum_nights	-1	0	0	1	0	2	-1	0	0	0	0	0	1	-1	-1
minimum_nights	0	0	0	1	0	2	0	0	0	0	2	0	0	-1	1
weekly_discount	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
monthly_discount	0	0	0	0	0	0	1	-2	0	10	1	0	0	0	0
host_response_time	-1	0	-1	0	-9	0	0	-1	0	0	0	1	0	0	0

* Loading values multiplied by 10 and rounded to the nearest integer.

Factor 1: Appliances. Factor 2: Capacity. Factor 3: Good shower. Factor 4: Laundry. Factor 5: Host response. Factor 6: Host listings count. Factor 7: Require guest profile. Factor 8: No internet. Factor 9: Smoke detector. Factor 10: Bulk discount. Factor 11: Additional fees. Factor 12: Unverified. Factor 13: Cooking. Factor 14: Guest fee. Factor 15: Unknown.

Appendix 5. Dendrogram of Hierarchical Clustering

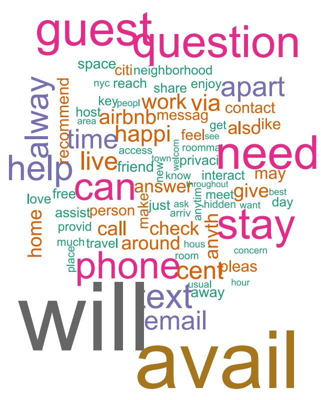
Long-Appendix A. Word Clouds



Summary

Description

Space



Interaction

Host “About Me”

House rules

Transit

Long-Appendix B. Identify the Topics

Here we look at the topics that are significantly correlated with demand in the linear regression results. Many of the topics are very interesting to look at!

Topic ID	Summary 8	Summary 15	Summary 17	Transit 2	Interaction 2	Interaction 13	Description 5	Description 10	Host about 1	Host about 2	Host about 3	Host about 5	Host about 10	Host about 12	Host about 13
	new	kitchen	build	just	live	will	new	wifi	time	peopl	home	manhattan	will	live	live
	citi	renov	view	block	apart	avail	citi	kitchen	work	meet	hous	locat	help	apart	brooklyn
	nyc	fulli	apart	right	hous	town	york	fulli	live	world	famili	park	make	peopl	love
	york	furnish	luxuri	away	also	friend	locat	includ	current	love	day	neighborhood	happi	citi	two
Terms	perfect	wifi	laundri	apart	build	issu	apart	equip	nyc	cultur	welcom	walk	know	home	year
	experi	includ	elev	locat	come	case	offer	towel	enjoy	new	book	restaur	feel	compani	husband
	explor	newli	amen	corner	floor	emerg	brand	provid	also	travel	one	room	can	real	famili
	offer	equip	rooftop	build	famili	travel	just	cabl	spend	around	come	east	need	tech	beauti
	best	unit	modern	also	upstair	also	nyc	microwav	student	differ	best	away	hope	estat	wife
	visit	applianc	beauti	front	manag	assist	everyth	coffe	school	learn	call	hotel	question	beauti	dog
Name	explore NYC	kitchen	high-rise	generic transit	host family lives here	host is your friend	generic description	kitchen towel micro-wave coffeemaker	lives in NYC	welcome foreigners	family here	Manhattan	will help	tech	Brooklyn couple (+dog)

Topic ID	Host about 14	Host about 19	House rules 2	House rules 3	House rules 4	House rules 6	House rules 8	House rules 9	House rules 16	House rules 17	House rules 18	House rules 19	Space 1	Space 10	Space 11	Space 12
	guest	stay	ask	pet	smoke	will	room	check	clean	use	pleas	fee	towel	fulli	home	bed
	cent	apart	just	may	pet	stay	kitchen	time	pleas	food	trash	charg	coffe	includ	apart	size
	room	local	make	check	parti	apart	use	will	dish	area	recycl	night	provid	equip	clean	queen
	communiti	comfort	friend	provid	allow	day	clean	late	use	clean	garbag	upon	microwav	kitchen	stay	sleep
Terms	care	provid	also	build	outsid	must	bathroom	fee	wash	remov	take	will	dryer	featur	day	sofa
	price	can	peopl	includ	insid	one	live	can	towel	etc	put	per	fresh	furnish	rent	full
	month	make	can	polici	absolut	month	cook	guest	bed	can	floor	tax	stove	applianc	place	comfort
	space	guest	feel	upon	drug	tenant	share	need	place	stay	paper	arriv	sheet	internet	long	mattress
	one	rent	need	prior	cat	reserv	eat	hour	keep	leav	toilet	room	oven	unit	everi	air
	properti	like	get	rent	cigarett	move	bedroom	provid	leav	turn	bin	airbnb	iron	cabl	year	twin
Name	cares	generic host about	host lives here, just ask	pet allowed	smoking	tenant	share kitchen / bathroom	late fee	plz clean dishes	???	plz recycle and take trash out	fees and taxes	towel coffeemaker micro-wave dryer stove oven iron	fully equipped	clean home	queen-sized bed

Long-Appendix C. Linear Regression Report

Response Days_in_a_year

Validation: Validation

Weight: Occupancy Confidence

Effect Summary

Source	LogWorth	PValue
host_acceptance_rate&MissingCoded	24.507	0.00000
host_is_superhost	15.525	0.00000
Log[price]&MissingCoded	14.859	0.00000
maximum_nights	13.649	0.00000
Log[calculated_host_listings_count]	12.882	0.00000
room_type	11.536	0.00000
monthly_discount	9.635	0.00000
Log[security_deposit]&MissingCoded	8.748	0.00000
9 of topic.house_rules&MissingCoded	8.602	0.00000
cancellation_policy	7.841	0.00000
amen_"Free street parking"	7.003	0.00000
6 of topic.house_rules&MissingCoded	6.990	0.00000
Location	6.633	0.00000
1 of topic.host_about&MissingCoded	6.338	0.00000
host_response_time	6.310	0.00000
amen_Hangers	5.917	0.00000
3 of topic.house_rules&MissingCoded	5.119	0.00001
weekly_discount	4.781	0.00002
amen_Internet	4.680	0.00002
10 of topic.host_about&MissingCoded	4.546	0.00003
Log[extra_people]&MissingCoded	4.441	0.00004
11 of topic.space&MissingCoded	4.418	0.00004
12 of topic.host_about&MissingCoded	4.408	0.00004
amen_Shampoo	4.403	0.00004
amen_"Fire extinguisher"	4.358	0.00004
17 of topic.house_rules&MissingCoded	4.321	0.00005
14 of topic.host_about&MissingCoded	4.111	0.00008
16 of topic.house_rules&MissingCoded	4.031	0.00009
4 of topic.house_rules&MissingCoded	3.606	0.00025
Log[guests_included]	3.448	0.00036
host_since&MissingCoded	2.969	0.00107
property_type	2.798	0.00159
amen_"Air conditioning"	2.754	0.00176
2 of topic.interaction&MissingCoded	2.641	0.00228
19 of topic.house_rules&MissingCoded	2.634	0.00232
18 of topic.house_rules&MissingCoded	2.620	0.00240
19 of topic.host_about&MissingCoded	2.576	0.00266
2 of topic.house_rules&MissingCoded	2.491	0.00323
15 of topic.summary&MissingCoded	2.484	0.00328
Log[beds]&MissingCoded	2.440	0.00363
3 of topic.host_about&MissingCoded	2.418	0.00382
12 of topic.space&MissingCoded	2.410	0.00389
amen_"Hot water"	2.302	0.00499
amen_Refrigerator	2.260	0.00549
13 of topic.interaction&MissingCoded	2.234	0.00583
instant_bookable	2.204	0.00625
Log[cleaning_fee]&MissingCoded	2.127	0.00747
5 of topic.description&MissingCoded	2.023	0.00948
10 of topic.description&MissingCoded	1.946	0.01133
amen_"Hair dryer"	1.887	0.01297
amen_"Carbon monoxide detector"	1.860	0.01381
2 of topic.transit&MissingCoded	1.822	0.01505
8 of topic.summary&MissingCoded	1.805	0.01567
amen_"Lock on bedroom door"	1.768	0.01708
2 of topic.host_about&MissingCoded	1.763	0.01725
1 of topic.space&MissingCoded	1.709	0.01953
space_word_count	1.643	0.02273
Log[accommodates]	1.626	0.02364
8 of topic.house_rules&MissingCoded	1.618	0.02412
neighbourhood_group_cleansed	1.570	0.02693
17 of topic.summary&MissingCoded	1.519	0.03024
10 of topic.space&MissingCoded	1.516	0.03047
5 of topic.host_about&MissingCoded	1.426	0.03750
polarity of polarity.interaction	1.419	0.03810
amen_"Dishes and silverware"	1.416	0.03835
13 of topic.host_about&MissingCoded	1.351	0.04457
amen_Iron	1.346	0.04512
amen_"Bed linens"	1.292	0.05103
12 of topic.house_rules&MissingCoded	1.253	0.05587
amen_Kitchen	1.239	0.05769
10 of topic.house_rules&MissingCoded	1.223	0.05983
14 of topic.interaction&MissingCoded	1.202	0.06276
3 of topic.transit&MissingCoded	1.191	0.06434
7 of topic.description&MissingCoded	1.186	0.06515
11 of topic.house_rules&MissingCoded	1.183	0.06555
polarity of polarity.summary	1.162	0.06894
16 of topic.host_about&MissingCoded	1.155	0.06991
amen_"First aid kit"	1.155	0.06994
4 of topic.space&MissingCoded	1.142	0.07209
15 of topic.house_rules&MissingCoded	1.140	0.07248
num_of_host_verifications	1.115	0.07672
6 of topic.host_about&MissingCoded	1.110	0.07755
amen_Oven	1.063	0.08657
4 of topic.description&MissingCoded	1.058	0.08756
1 of topic.house_rules&MissingCoded	1.033	0.09261
9 of topic.host_about&MissingCoded	1.011	0.09753
amen_TV	1.007	0.09833
3 of topic.space&MissingCoded	1.005	0.09894
polarity of host_about	1.003	0.09941
4 of topic.transit&MissingCoded	0.987	0.10299
5 of topic.transit&MissingCoded	0.986	0.10339
8 of topic.transit&MissingCoded	0.956	0.11066
9 of topic.summary&MissingCoded	0.948	0.11262
11 of topic.host_about&MissingCoded	0.909	0.12321
5 of topic.space&MissingCoded	0.901	0.12547
18 of topic.host_about&MissingCoded	0.898	0.12653
17 of topic.transit&MissingCoded	0.881	0.13167
6 of topic.transit&MissingCoded	0.878	0.13235
7 of topic.house_rules&MissingCoded	0.872	0.13425
4 of topic.host_about&MissingCoded	0.871	0.13467
9 of topic.interaction&MissingCoded	0.865	0.13641
7 of topic.space&MissingCoded	0.842	0.14395
13 of topic.house_rules&MissingCoded	0.836	0.14597
9 of topic.space&MissingCoded	0.832	0.14733
amen_"Laptop friendly workspace"	0.795	0.16025
log(host_listing_count)&MissingCoded	0.790	0.16205
16 of topic.description&MissingCoded	0.746	0.17968
15 of topic.description&MissingCoded	0.739	0.18234
2 of topic.space&MissingCoded	0.704	0.19751
amen_Essentials	0.702	0.19882
7 of topic.transit&MissingCoded	0.696	0.20134
13 of topic.transit&MissingCoded	0.674	0.21169
8 of topic.space&MissingCoded	0.664	0.21660
amen_"Coffee maker"	0.651	0.22316
interaction_word_count	0.635	0.23197
summary_word_count	0.632	0.23321
19 of topic.space&MissingCoded	0.623	0.23842
16 of topic.space&MissingCoded	0.618	0.24105
15 of topic.host_about&MissingCoded	0.613	0.24358
5 of topic.interaction&MissingCoded	0.609	0.24628
15 of topic.interaction&MissingCoded	0.593	0.25512
12 of topic.interaction&MissingCoded	0.578	0.26443
amen_Stove	0.563	0.27348
18 of topic.space&MissingCoded	0.560	0.27528
Log[bedrooms]&MissingCoded	0.515	0.30527
host_about_word_count	0.493	0.32124
host_has_profile_pic	0.478	0.33241
6 of topic.interaction&MissingCoded	0.466	0.34188
host_response_rate&MissingCoded	0.465	0.34296
has_summary	0.456	0.35001
17 of topic.interaction&MissingCoded	0.444	0.35982
1 of topic.interaction&MissingCoded	0.438	0.36441
4 of topic.interaction&MissingCoded	0.432	0.36975
10 of topic.interaction&MissingCoded	0.430	0.37133
5 of topic.house_rules&MissingCoded	0.423	0.37726
17 of topic.host_about&MissingCoded	0.414	0.38576
14 of topic.summary&MissingCoded	0.402	0.39605
16 of topic.interaction&MissingCoded	0.401	0.39746
6 of topic.summary&MissingCoded	0.374	0.42281
19 of topic.description&MissingCoded	0.364	0.43268
bed_type	0.342	0.45513
description_word_count	0.338	0.45963
17 of topic.description&MissingCoded	0.330	0.46769
10 of topic.summary&MissingCoded	0.303	0.49718
11 of topic.interaction&MissingCoded	0.302	0.49931
amen_Wifi	0.285	0.51892
19 of topic.transit&MissingCoded	0.283	0.52084
require_guest_phone_verification	0.277	0.52799
amen_Heating	0.277	0.52904
10 of topic.transit&MissingCoded	0.261	0.54846
3 of topic.description&MissingCoded	0.260	0.54921
16 of topic.summary&MissingCoded	0.253	0.55821
8 of topic.host_about&MissingCoded	0.252	0.55972
1 of topic.transit&MissingCoded	0.248	0.56525
16 of topic.transit&MissingCoded	0.236	0.58138
13 of topic.summary&MissingCoded	0.215	0.60892
19 of topic.interaction&MissingCoded	0.211	0.61537
name_is_all_cap	0.203	0.62677
19 of topic.summary&MissingCoded	0.200	0.63066
18 of topic.summary&MissingCoded	0.199	0.63207
6 of topic.description&MissingCoded	0.188	0.64801
amen_"Smoke detector"	0.184	0.65442
15 of topic.transit&MissingCoded	0.182	0.65762
amen_Microwave	0.182	0.65765
13 of topic.description&MissingCoded	0.179	0.66265
Log[bathrooms]&MissingCoded	0.174	0.66926
14 of topic.description&MissingCoded	0.168	0.67967
18 of topic.transit&MissingCoded	0.165	0.68458
9 of topic.description&MissingCoded	0.148	0.71175
amen_Dryer	0.142	0.72078
3 of topic.interaction&MissingCoded	0.131	0.73968
2 of topic.summary&MissingCoded	0.130	0.74183
13 of topic.space&MissingCoded	0.128	0.74449
host_identity_verified	0.123	0.75323
17 of topic.space&MissingCoded	0.122	0.75495
11 of topic.transit&MissingCoded	0.122	0.75500
polarity of polarity.space	0.121	0.75601
5 of topic.summary&MissingCoded	0.118	0.76194
14 of topic.transit&MissingCoded	0.107	0.78164
name_word_count	0.106	0.78301
neighborhood_overview_word_count	0.100	0.79406
14 of topic.space&MissingCoded	0.097	0.80071
name_avg_word_len	0.092	0.80921
15 of topic.space&MissingCoded	0.090	0.81195
14 of topic.house_rules&MissingCoded	0.084	0.82376
8 of topic.interaction&MissingCoded	0.080	0.83240
require_guest_profile_picture	0.076	0.83945
polarity of polarity.description	0.075	0.84209
3 of topic.summary&MissingCoded	0.073	0.84619
7 of topic.host_about&MissingCoded	0.071	0.84888
18 of topic.description&MissingCoded	0.070	0.85060
12 of topic.summary&MissingCoded	0.070	0.85176
12 of topic.description&MissingCoded	0.067	0.85723
11 of topic.summary&MissingCoded	0.063	0.86581
12 of topic.transit&MissingCoded	0.057	0.87716
7 of topic.interaction&MissingCoded	0.052	0.88724
6 of topic.space&MissingCoded	0.048	0.89591
11 of topic.description&MissingCoded	0.044	0.90305
8 of topic.description&MissingCoded	0.042	0.90696
9 of topic.transit&MissingCoded	0.038	0.91715
1 of topic.summary&MissingCoded	0.030	0.93360
4 of topic.summary&MissingCoded	0.016	0.96291
amen_Washer	0.015	0.96578
amen_Elevator	0.015	0.96660
7 of topic.summary&MissingCoded	0.015	0.96683
2 of topic.description&MissingCoded	0.014	0.96929
amen_"Cooking basics"	0.008	0.98188
18 of topic.interaction&MissingCoded	0.003	0.99224
1 of topic.description&MissingCoded	0.003	0.99283

Summary of Fit

RSquare	0.197433
RSquare Adj	0.190506
Root Mean Square Error	94.63633
Mean of Response	64.21203
Observations (or Sum Wgts)	40478.37

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	318	81170304	255253	28.5006
Error	36842	329958226	8956	Prob > F
C. Total	37160	411128531		<.0001*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	-794.7719	3020975	-0.00	0.9998
host_since Or Mean if Missing		2.8186e-8	8.618e-9	3.27	0.0011*
host_since Is Missing	Biased	3359.8279	18125848	0.00	0.9999
host_response_time[within an hour]	Biased	576.71266	3020975	0.00	0.9998
host_response_time[within a few hours]	Biased	568.15033	3020975	0.00	0.9998
host_response_time[within a day]	Biased	568.28805	3020975	0.00	0.9998
host_response_time[N/A]	Biased	544.40919	3020975	0.00	0.9999
host_response_time[a few days or more]	Biased	575.00369	3020975	0.00	0.9998
host_response_rate Or Mean if Missing		6.8706215	7.244895	0.95	0.3430
host_response_rate Is Missing	Zeroed	0	0	.	.
host_acceptance_rate Or Mean if Missing		4.9441656	2.815403	1.76	0.0791
host_acceptance_rate Is Missing		-19.74668	1.859336	-10.62	<.0001*
host_is_superhost[t]	Biased	11.500534	1.406353	8.18	<.0001*
host_is_superhost[f]	Zeroed	0	0	.	.
log(host_listing_count) Or Mean if Missing		1.1940286	1.225564	0.97	0.3299
log(host_listing_count) Is Missing		2.717722	1.981992	1.37	0.1703
Log[calculated_host_listings_count]		9.9238915	1.339644	7.41	<.0001*
num_of_host_verifications		-0.472983	0.267204	-1.77	0.0767
host_has_profile_pic[t]	Biased	-8.99371	9.278636	-0.97	0.3324
host_has_profile_pic[f]	Zeroed	0	0	.	.
host_identity_verified[t]	Biased	0.3508879	1.116124	0.31	0.7532
host_identity_verified[f]	Zeroed	0	0	.	.
Location[Astoria]		-3.178282	4.340364	-0.73	0.4640
Location[Bedford-Stuyvesant]		1.6185786	2.779958	0.58	0.5604
Location[Bushwick]		0.7829626	3.146389	0.25	0.8035
Location[Chelsea]		5.3339914	3.698491	1.44	0.1493
Location[Chinatown]		-0.837486	5.594044	-0.15	0.8810
Location[Clinton Hill]		3.904044	4.6916	0.83	0.4053
Location[Crown Heights]		-3.149256	3.33971	-0.94	0.3457
Location[Ditmars Steinway]		-6.690297	6.789233	-0.99	0.3244
Location[East Flatbush]		-4.151228	5.293419	-0.78	0.4329
Location[East Harlem]		-7.152428	3.786641	-1.89	0.0589
Location[East Village]		-0.800759	3.169077	-0.25	0.8005
Location[Elmhurst]		-10.34844	6.945561	-1.49	0.1362
Location[Financial District]		-2.482978	5.004987	-0.50	0.6198
Location[Flatbush]		-5.831585	4.774742	-1.22	0.2220
Location[Flushing]		-12.6528	5.820795	-2.17	0.0297*
Location[Fort Greene]		-1.310812	5.313706	-0.25	0.8052
Location[Gramercy]		-1.64918	5.84133	-0.28	0.7777
Location[Greenpoint]		1.7840453	3.861905	0.46	0.6441
Location[Greenwich Village]		12.856793	5.688828	2.26	0.0238*
Location[Harlem]		-6.4442	2.990599	-2.15	0.0312*
Location[Hell's Kitchen]		2.313366	3.319816	0.70	0.4859
Location[Jamaica]		-2.293913	7.508036	-0.31	0.7600
Location[Kips Bay]		-4.431171	5.288904	-0.84	0.4021
Location[Long Island City]		-3.861534	5.227953	-0.74	0.4601
Location[Lower East Side]		4.5071625	4.008655	1.12	0.2609
Location[Midtown]		-0.329645	3.625017	-0.09	0.9275
Location[Morningside Heights]		34.539681	5.954082	5.80	<.0001*
Location[Murray Hill]		6.5885011	5.119581	1.29	0.1981
Location[Nolita]		-6.551685	6.755978	-0.97	0.3322
Location[Park Slope]		9.3378447	4.994899	1.87	0.0616
Location[Prospect Heights]		2.8636292	5.963535	0.48	0.6311
Location[Prospect-Lefferts Gardens]		-3.364691	4.941602	-0.68	0.4959
Location[Ridgewood]		-11.98283	5.999271	-2.00	0.0458*
Location[SoHo]		-4.612317	5.840028	-0.79	0.4297
Location[South Slope]		-1.667455	6.470398	-0.26	0.7966
Location[Sunnyside]		-9.231779	5.992693	-1.54	0.1234
Location[Sunset Park]		5.5021832	5.298194	1.04	0.2990
Location[Theater District]		26.974787	7.055518	3.82	0.0001*
Location[Upper East Side]		1.3565087	3.238429	0.42	0.6753
Location[Upper West Side]		1.6404633	3.146487	0.52	0.6021
Location[Washington Heights]		4.2080299	4.075302	1.03	0.3018
Location[West Village]		0.5594685	4.381428	0.13	0.8984
Location[Williamsburg]		2.9325212	2.970183	0.99	0.3235
Location[Woodside]		-14.24485	6.82117	-2.09	0.0368*
neighbourhood_group_cleansed[Bronx]		0.5423548	2.990282	0.18	0.8561
neighbourhood_group_cleansed[Brooklyn]		-5.758839	2.228979	-2.58	0.0098*
neighbourhood_group_cleansed[Queens]		4.0091848	2.535105	1.58	0.1138
neighbourhood_group_cleansed[Staten Island]		-0.894447	4.817994	-0.19	0.8527
property_type[Aparthotel]	Biased	-51.21882	79.97686	-0.64	0.5219
property_type[Barn]	Biased	51.877436	98.68798	0.53	0.5991
property_type[Bed and breakfast]	Biased	12.53674	75.65696	0.17	0.8684
property_type[Boat]	Biased	-6.687846	90.27488	-0.07	0.9409
property_type[Boutique hotel]	Biased	11.466638	74.44911	0.15	0.8776
property_type[Bungalow]	Biased	13.32568	77.03051	0.17	0.8627
property_type[Bus]	Biased	-439.4913	2535.241	-0.17	0.8624
property_type[Cabin]	Biased	29.737414	107.2845	0.28	0.7816
property_type[Camper/RV]	Biased	-1.828976	78.86182	-0.02	0.9815
property_type[Casa particular (Cuba)]	Zeroed	0	0	.	.
property_type[Cave]	Biased	27.677467	96.68534	0.29	0.7747
property_type[Condominium]	Biased	27.556166	74.09056	0.37	0.7100
property_type[Cottage]	Biased	47.659344	98.55147	0.48	0.6287
property_type[Dome house]	Biased	51.895688	95.32718	0.54	0.5862
property_type[Dorm]	Biased	-83.39315	136.8622	-0.61	0.5423
property_type[Earth house]	Biased	-30.98558	83.06566	-0.37	0.7091
property_type[Farm stay]	Biased	16.322904	127.2273	0.13	0.8979
property_type[Guest suite]	Biased	23.942676	74.24825	0.32	0.7471
property_type[Guesthouse]	Biased	32.241019	75.13363	0.43	0.6678
property_type[Hostel]	Biased	83.46679	76.85959	1.09	0.2775
property_type[Hotel]	Biased	7.5833051	74.82262	0.10	0.9193
property_type[House]	Biased	18.656456	74.06642	0.25	0.8011
property_type[Houseboat]	Biased	-41.21868	94.6265	-0.44	0.6631
property_type[Island]	Biased	-34.08731	102.2843	-0.33	0.7389
property_type[Lighthouse]	Biased	-57.88492	105.6404	-0.55	0.5837
property_type[Loft]	Biased	23.589616	74.0957	0.32	0.7502
property_type[Other]	Biased	24.522798	74.81472	0.33	0.7431
property_type[Resort]	Biased	-29.10982	75.1728	-0.39	0.6986
property_type[Serviced apartment]	Biased	-1.073035	74.32292	-0.01	0.9885
property_type[Tent]	Biased	-8.773804	90.62992	-0.10	0.9229
property_type[Timeshare]	Biased	119.7113	124.4065	0.96	0.3359
property_type[Tiny house]	Biased	27.017769	76.8711	0.35	0.7252
property_type[Townhouse]	Biased	17.730549	74.09076	0.24	0.8109
property_type[Treehouse]	Biased	76.29967	110.9244	0.69	0.4915
property_type[Villa]	Biased	17.822069	76.99403	0.23	0.8169
property_type[Yurt]	Zeroed	0	0	.	.
Log[accommodates]		-3.561953	1.573992	-2.26	0.0236*
Log[bathrooms] Or Mean if Missing		-0.221867	2.269513	-0.10	0.9221
Log[bathrooms] Is Missing		8.1383814	9.125637	0.89	0.3725
Log[bedrooms] Or Mean if Missing		3.3837566	2.200737	1.54	0.1242
Log[bedrooms] Is Missing		-0.182353	1.82665	-0.10	0.9205
Log[beds] Or Mean if Missing		5.3611139	1.710146	3.13	0.0017*
Log[beds] Is Missing		-3.882206	2.774634	-1.40	0.1618
Log[security_deposit] Or Mean if Missing		5.4678912	0.976303	5.60	<.0001*
Log[security_deposit] Is Missing		-4.129884	1.126407	-3.67	0.0002*
Log[cleaning_fee] Or Mean if Missing		1.5585256	0.947031	1.65	0.0998
Log[cleaning_fee] Is Missing		3.0349284	1.332385	2.28	0.0227*
Log[guests_included]		4.9879699	1.396821	3.57	0.0004*
Log[extra_people] Or Mean if Missing		-3.326376	1.121152	-2.97	0.0030*
Log[extra_people] Is Missing		4.2930855	1.2319	3.48	0.0005*
room_type[Shared room]		-13.16312	3.360254	-3.92	<.0001*
room_type[Private room]		0.5686181	2.185972	0.26	0.7948
room_type[Hotel room]		1.3511537	5.7707	0.23	0.8149
bed_type[Airbed]		-2.997842	7.42937	-0.40	0.6866
bed_type[Couch]		-11.09772	10.43271	-1.06	0.2875
bed_type[Futon]		11.10755	5.930867	1.87	0.0611
bed_type[Pull-out Sofa]		0.3014427	6.427482	0.05	0.9626
amen_Wifi[t]		1.101153	1.707148	0.65	0.5189
amen_Heating[t]		0.6737418	1.070325	0.63	0.5290
amen_Essentials[t]		-1.314079	1.022686	-1.28	0.1988
amen_Kitchen[t]		-1.971083	1.038465	-1.90	0.0577
amen_"Smoke detector"[t]		-0.389762	0.870712	-0.45	0.6544
amen_"Air conditioning"[t]		2.2662468	0.724537	3.13	0.0018*
amen_Hangers[t]		3.2847618	0.676607	4.85	<.0001*
amen_"Carbon monoxide detector"[t]		1.6000525	0.649833	2.46	0.0138*
amen_TV[t]		-0.947155	0.572968	-1.65	0.0983
amen_Shampoo[t]		-2.410547	0.586449	-4.11	<.0001*
amen_"Hair dryer"[t]		1.5880765	0.639173	2.48	0.0130*
amen_"Laptop friendly workspace"[t]		-0.80892	0.576047	-1.40	0.1602
amen_Iron[t]		1.2530972	0.625425	2.00	0.0451*
amen_"Hot water"[t]		-1.99238	0.709538	-2.81	0.0050*
amen_Refrigerator[t]		3.4184587	1.231047	2.78	0.0055*
amen_"Dishes and silverware"[t]		2.1020564	1.014891	2.07	0.0383*
amen_Washer[t]		-0.077703	1.811366	-0.04	0.9658
amen_Dryer[t]		-0.647988	1.81295	-0.36	0.7208
amen_"Fire extinguisher"[t]		2.3543981	0.576079	4.09	<.0001*
amen_Microwave[t]		0.3839153	0.8663	0.44	0.6576
amen_"Lock on bedroom door"[t]		-1.368212	0.573644	-2.39	0.0171*
amen_Stove[t]		-1.391001	1.270203	-1.10	0.2735
amen_"Cooking basics"[t]		-0.025727	1.132784	-0.02	0.9819
amen_Oven[t]		-2.053361	1.198124	-1.71	0.0866
amen_"Free street parking"[t]		3.657719	0.686353	5.33	<.0001*
amen_"Coffee maker"[t]		0.9678941	0.794527	1.22	0.2232
amen_"First aid kit"[t]		-1.059087	0.58438	-1.81	0.0699
amen_"Bed linens"[t]		1.3163115	0.67459	1.95	0.0510
amen_Internet[t]		2.7489153	0.645907	4.26	<.0001*
amen_Elevator[t]		0.0288368	0.688606	0.04	0.9666
weekly_discount		-23.8996	5.54858	-4.31	<.0001*
monthly_discount		33.099696	5.220519	6.34	<.0001*
maximum_nights		0.007096	0.000929	7.64	<.0001*
instant_bookable[t]		1.5417837	0.563802	2.73	0.0062*
cancellation_policy[flexible]		18.579563	5.843088	3.18	0.0015*
cancellation_policy[moderate]		18.722556	5.817904	3.22	0.0013*
cancellation_policy[strict_14_with_grace_period]		22.536043	5.758734	3.91	<.0001*
cancellation_policy[strict]		16.601252	18.74609	0.89	0.3758
cancellation_policy[super_strict_30]		-30.31883	19.19968	-1.58	0.1143
require_guest_profile_picture[t]		-0.547261	2.701242	-0.20	0.8395
require_guest_phone_verification[t]		1.6532377	2.619671	0.63	0.5280
name_word_count		0.0863982	0.313717	0.28	0.7830
summary_word_count		-0.033655	0.028231	-1.19	0.2332
has_summary[1]		5.5320384	5.91932	0.93	0.3500
space_word_count		0.0352089	0.015456	2.28	0.0227*
description_word_count		0.0151112	0.020435	0.74	0.4596
interaction_word_count		-0.039854	0.033342	-1.20	0.2320
host_about_word_count		-0.015238	0.015362	-0.99	0.3212
neighborhood_overview_word_count		0.0038272	0.014661	0.26	0.7941
name_is_all_cap[1]		0.6178741	1.270626	0.49	0.6268
name_avg_word_len		0.10599	0.43898	0.24	0.8092
Log[price] Or Mean if Missing		-9.911031	1.204964	-8.23	<.0001*
Log[price] Is Missing		-31.87416	32.10847	-0.99	0.3209
polarity of polarity.description		-2.407568	12.08511	-0.20	0.8421
polarity of polarity.interaction		-7.800722	3.761441	-2.07	0.0381*
polarity of polarity.summary		-15.40645	8.470246	-1.82	0.0689
polarity of polarity.space		1.8029557	5.802431	0.31	0.7560
polarity of host_about		6.8207065	4.139357	1.65	0.0994
1 of topic.summary Or Mean if Missing		-3.066522	36.80792	-0.08	0.9336
1 of topic.summary Is Missing	Biased	6.7734232	11.98813	0.57	0.5721
2 of topic.summary Or Mean if Missing		-12.06333	36.61874	-0.33	0.7418
2 of topic.summary Is Missing	Zeroed	0	0	.	.
3 of topic.summary Or Mean if Missing		-7.059731	36.39415	-0.19	0.8462
3 of topic.summary Is Missing	Zeroed	0	0	.	.
4 of topic.summary Or Mean if Missing		1.7479153	37.58545	0.05	0.9629
4 of topic.summary Is Missing	Zeroed	0	0	.	.
5 of topic.summary Or Mean if Missing		-10.97794	36.23815	-0.30	0.7619
5 of topic.summary Is Missing	Zeroed	0	0	.	.
6 of topic.summary Or Mean if Missing		28.105947	35.0636	0.80	0.4228
6 of topic.summary Is Missing	Zeroed	0	0	.	.
7 of topic.summary Or Mean if Missing		1.5777808	37.93877	0.04	0.9668
7 of topic.summary Is Missing	Zeroed	0	0	.	.
8 of topic.summary Or Mean if Missing		-89.72871	37.12866	-2.42	0.0157*
8 of topic.summary Is Missing	Zeroed	0	0	.	.
9 of topic.summary Or Mean if Missing		-60.26096	37.98206	-1.59	0.1126
9 of topic.summary Is Missing	Zeroed	0	0	.	.
10 of topic.summary Or Mean if Missing		-28.29739	41.67868	-0.68	0.4972
10 of topic.summary Is Missing	Zeroed	0	0	.	.
11 of topic.summary Or Mean if Missing		5.8347835	34.5296	0.17	0.8658
11 of topic.summary Is Missing	Zeroed	0	0	.	.
12 of topic.summary Or Mean if Missing		-6.991809	37.41543	-0.19	0.8518
12 of topic.summary Is Missing	Zeroed	0	0	.	.
13 of topic.summary Or Mean if Missing		-18.05238	35.28451	-0.51	0.6089
13 of topic.summary Is Missing	Zeroed	0	0	.	.
14 of topic.summary Or Mean if Missing		-31.11481	36.66179	-0.85	0.3961
14 of topic.summary Is Missing	Zeroed	0	0	.	.
15 of topic.summary Or Mean if Missing		106.60785	36.25382	2.94	0.0033*
15 of topic.summary Is Missing	Zeroed	0	0	.	.
16 of topic.summary Or Mean if Missing		21.839453	37.30039	0.59	0.5582
16 of topic.summary Is Missing	Zeroed	0	0	.	.
17 of topic.summary Or Mean if Missing		-82.63847	38.13521	-2.17	0.0302*
17 of topic.summary Is Missing	Zeroed	0	0	.	.
18 of topic.summary Or Mean if Missing		-18.34716	38.31706	-0.48	0.6321
18 of topic.summary Is Missing	Zeroed	0	0	.	.
19 of topic.summary Or Mean if Missing		-19.82098	41.22457	-0.48	0.6307
19 of topic.summary Is Missing	Zeroed	0	0	.	.
1 of topic.transit Or Mean if Missing		-24.00469	41.74218	-0.58	0.5652
1 of topic.transit Is Missing	Biased	-0.641961	1.577847	-0.41	0.6841
2 of topic.transit Or Mean if Missing		-102.2594	42.05933	-2.43	0.0150*

2 of topic.transit Is Missing	Zeroed	0	0	.	.
3 of topic.transit Or Mean if Missing		-69.57174	37.60949	-1.85	0.0643
3 of topic.transit Is Missing	Zeroed	0	0	.	.
4 of topic.transit Or Mean if Missing		-69.84632	42.83506	-1.63	0.1030
4 of topic.transit Is Missing	Zeroed	0	0	.	.
5 of topic.transit Or Mean if Missing		-66.63929	40.91653	-1.63	0.1034
5 of topic.transit Is Missing	Zeroed	0	0	.	.
6 of topic.transit Or Mean if Missing		-61.78415	41.05431	-1.50	0.1323
6 of topic.transit Is Missing	Zeroed	0	0	.	.
7 of topic.transit Or Mean if Missing		-47.21622	36.95199	-1.28	0.2013
7 of topic.transit Is Missing	Zeroed	0	0	.	.
8 of topic.transit Or Mean if Missing		63.23535	39.63876	1.60	0.1107
8 of topic.transit Is Missing	Zeroed	0	0	.	.
9 of topic.transit Or Mean if Missing		-4.053271	38.96482	-0.10	0.9172
9 of topic.transit Is Missing	Zeroed	0	0	.	.
10 of topic.transit Or Mean if Missing		26.413607	44.01688	0.60	0.5485
10 of topic.transit Is Missing	Zeroed	0	0	.	.
11 of topic.transit Or Mean if Missing		-13.72456	43.98125	-0.31	0.7550
11 of topic.transit Is Missing	Zeroed	0	0	.	.
12 of topic.transit Or Mean if Missing		6.6572078	43.06913	0.15	0.8772
12 of topic.transit Is Missing	Zeroed	0	0	.	.
13 of topic.transit Or Mean if Missing		-53.26473	42.6472	-1.25	0.2117
13 of topic.transit Is Missing	Zeroed	0	0	.	.
14 of topic.transit Or Mean if Missing		12.392775	44.71038	0.28	0.7816
14 of topic.transit Is Missing	Zeroed	0	0	.	.
15 of topic.transit Or Mean if Missing		17.951327	40.50369	0.44	0.6576
15 of topic.transit Is Missing	Zeroed	0	0	.	.
16 of topic.transit Or Mean if Missing		-23.72429	43.0271	-0.55	0.5814
16 of topic.transit Is Missing	Zeroed	0	0	.	.
17 of topic.transit Or Mean if Missing		-61.18912	40.5877	-1.51	0.1317
17 of topic.transit Is Missing	Zeroed	0	0	.	.
18 of topic.transit Or Mean if Missing		-16.41594	40.41068	-0.41	0.6846
18 of topic.transit Is Missing	Zeroed	0	0	.	.
19 of topic.transit Or Mean if Missing		-24.82076	38.65821	-0.64	0.5208
19 of topic.transit Is Missing	Zeroed	0	0	.	.
1 of topic.interaction Or Mean if Missing		-55.12293	60.77411	-0.91	0.3644
1 of topic.interaction Is Missing	Biased	-0.552951	1.815402	-0.30	0.7607
2 of topic.interaction Or Mean if Missing		199.65277	65.44435	3.05	0.0023*
2 of topic.interaction Is Missing	Zeroed	0	0	.	.
3 of topic.interaction Or Mean if Missing		-19.71404	59.32852	-0.33	0.7397
3 of topic.interaction Is Missing	Zeroed	0	0	.	.
4 of topic.interaction Or Mean if Missing		60.989157	67.99567	0.90	0.3697
4 of topic.interaction Is Missing	Zeroed	0	0	.	.
5 of topic.interaction Or Mean if Missing		-69.52244	59.96175	-1.16	0.2463
5 of topic.interaction Is Missing	Zeroed	0	0	.	.
6 of topic.interaction Or Mean if Missing		-58.9975	62.07178	-0.95	0.3419
6 of topic.interaction Is Missing	Zeroed	0	0	.	.
7 of topic.interaction Or Mean if Missing		9.1563716	64.57527	0.14	0.8872
7 of topic.interaction Is Missing	Zeroed	0	0	.	.
8 of topic.interaction Or Mean if Missing		13.240442	62.56602	0.21	0.8324
8 of topic.interaction Is Missing	Zeroed	0	0	.	.
9 of topic.interaction Or Mean if Missing		-94.66689	63.56393	-1.49	0.1364
9 of topic.interaction Is Missing	Zeroed	0	0	.	.
10 of topic.interaction Or Mean if Missing		-60.15064	67.28239	-0.89	0.3713
10 of topic.interaction Is Missing	Zeroed	0	0	.	.
11 of topic.interaction Or Mean if Missing		-42.9489	63.57352	-0.68	0.4993
11 of topic.interaction Is Missing	Zeroed	0	0	.	.
12 of topic.interaction Or Mean if Missing		-69.65695	62.41621	-1.12	0.2644
12 of topic.interaction Is Missing	Zeroed	0	0	.	.
13 of topic.interaction Or Mean if Missing		-176.3675	63.96503	-2.76	0.0058*
13 of topic.interaction Is Missing	Zeroed	0	0	.	.
14 of topic.interaction Or Mean if Missing		-127.1598	68.32961	-1.86	0.0628
14 of topic.interaction Is Missing	Zeroed	0	0	.	.
15 of topic.interaction Or Mean if Missing		-68.48182	60.1761	-1.14	0.2551
15 of topic.interaction Is Missing	Zeroed	0	0	.	.
16 of topic.interaction Or Mean if Missing		-54.07533	63.9049	-0.85	0.3975
16 of topic.interaction Is Missing	Zeroed	0	0	.	.
17 of topic.interaction Or Mean if Missing		-51.19783	55.91029	-0.92	0.3598
17 of topic.interaction Is Missing	Zeroed	0	0	.	.
18 of topic.interaction Or Mean if Missing		-0.668078	68.65403	-0.01	0.9922
18 of topic.interaction Is Missing	Zeroed	0	0	.	.
19 of topic.interaction Or Mean if Missing		-31.41556	62.52818	-0.50	0.6154
19 of topic.interaction Is Missing	Zeroed	0	0	.	.
1 of topic.description Or Mean if Missing		0.3175209	35.33353	0.01	0.9928
1 of topic.description Is Missing	Biased	-8.888722	5.130179	-1.73	0.0832
2 of topic.description Or Mean if Missing		-1.48229	38.49768	-0.04	0.9693
2 of topic.description Is Missing	Zeroed	0	0	.	.
3 of topic.description Or Mean if Missing		-20.31428	33.91617	-0.60	0.5492
3 of topic.description Is Missing	Zeroed	0	0	.	.
4 of topic.description Or Mean if Missing		-67.52713	39.52564	-1.71	0.0876
4 of topic.description Is Missing	Zeroed	0	0	.	.
5 of topic.description Or Mean if Missing		92.818666	35.77478	2.59	0.0095*
5 of topic.description Is Missing	Zeroed	0	0	.	.
6 of topic.description Or Mean if Missing		16.033666	35.12028	0.46	0.6480
6 of topic.description Is Missing	Zeroed	0	0	.	.
7 of topic.description Or Mean if Missing		-67.33838	36.51182	-1.84	0.0651
7 of topic.description Is Missing	Zeroed	0	0	.	.
8 of topic.description Or Mean if Missing		-4.027002	34.45448	-0.12	0.9070
8 of topic.description Is Missing	Zeroed	0	0	.	.
9 of topic.description Or Mean if Missing		12.6804	34.31736	0.37	0.7118
9 of topic.description Is Missing	Zeroed	0	0	.	.
10 of topic.description Or Mean if Missing		83.704064	33.05176	2.53	0.0113*
10 of topic.description Is Missing	Zeroed	0	0	.	.
11 of topic.description Or Mean if Missing		4.2338449	34.75907	0.12	0.9031
11 of topic.description Is Missing	Zeroed	0	0	.	.
12 of topic.description Or Mean if Missing		-6.358368	35.34245	-0.18	0.8572
12 of topic.description Is Missing	Zeroed	0	0	.	.
13 of topic.description Or Mean if Missing		-14.74969	33.8096	-0.44	0.6627
13 of topic.description Is Missing	Zeroed	0	0	.	.
14 of topic.description Or Mean if Missing		-13.81533	33.45741	-0.41	0.6797
14 of topic.description Is Missing	Zeroed	0	0	.	.
15 of topic.description Or Mean if Missing		-48.24306	36.17453	-1.33	0.1823
15 of topic.description Is Missing	Zeroed	0	0	.	.
16 of topic.description Or Mean if Missing		47.393058	35.32133	1.34	0.1797
16 of topic.description Is Missing	Zeroed	0	0	.	.
17 of topic.description Or Mean if Missing		-23.87151	32.86965	-0.73	0.4677
17 of topic.description Is Missing	Zeroed	0	0	.	.
18 of topic.description Or Mean if Missing		6.5123275	34.57513	0.19	0.8506
18 of topic.description Is Missing	Zeroed	0	0	.	.
19 of topic.description Or Mean if Missing		-28.13909	35.86312	-0.78	0.4327
19 of topic.description Is Missing	Zeroed	0	0	.	.
1 of topic.host_about Or Mean if Missing		241.77867	47.93831	5.04	<.0001*
1 of topic.host_about Is Missing	Biased	1.6370804	1.596638	1.03	0.3052
2 of topic.host_about Or Mean if Missing		109.5838	46.01451	2.38	0.0172*
2 of topic.host_about Is Missing	Zeroed	0	0	.	.
3 of topic.host_about Or Mean if Missing		140.14269	48.4386	2.89	0.0038*
3 of topic.host_about Is Missing	Zeroed	0	0	.	.
4 of topic.host_about Or Mean if Missing		68.737752	45.94909	1.50	0.1347
4 of topic.host_about Is Missing	Zeroed	0	0	.	.
5 of topic.host_about Or Mean if Missing		81.854026	39.34591	2.08	0.0375*
5 of topic.host_about Is Missing	Zeroed	0	0	.	.
6 of topic.host_about Or Mean if Missing		87.210655	49.408	1.77	0.0776
6 of topic.host_about Is Missing	Zeroed	0	0	.	.
7 of topic.host_about Or Mean if Missing		9.5197154	49.95988	0.19	0.8489
7 of topic.host_about Is Missing	Zeroed	0	0	.	.
8 of topic.host_about Or Mean if Missing		27.891096	47.81952	0.58	0.5597
8 of topic.host_about Is Missing	Zeroed	0	0	.	.
9 of topic.host_about Or Mean if Missing		79.18654	47.78899	1.66	0.0975
9 of topic.host_about Is Missing	Zeroed	0	0	.	.
10 of topic.host_about Or Mean if Missing		187.30749	44.74601	4.19	<.0001*
10 of topic.host_about Is Missing	Zeroed	0	0	.	.
11 of topic.host_about Or Mean if Missing		72.013812	46.71811	1.54	0.1232
11 of topic.host_about Is Missing	Zeroed	0	0	.	.
12 of topic.host_about Or Mean if Missing		207.49882	50.44373	4.11	<.0001*
12 of topic.host_about Is Missing	Zeroed	0	0	.	.
13 of topic.host_about Or Mean if Missing		95.362933	47.47339	2.01	0.0446*
13 of topic.host_about Is Missing	Zeroed	0	0	.	.
14 of topic.host_about Or Mean if Missing		165.03695	41.75099	3.95	<.0001*
14 of topic.host_about Is Missing	Zeroed	0	0	.	.
15 of topic.host_about Or Mean if Missing		56.233175	48.22347	1.17	0.2436
15 of topic.host_about Is Missing	Zeroed	0	0	.	.
16 of topic.host_about Or Mean if Missing		85.870074	47.37463	1.81	0.0699
16 of topic.host_about Is Missing	Zeroed	0	0	.	.
17 of topic.host_about Or Mean if Missing		45.060916	51.95297	0.87	0.3858
17 of topic.host_about Is Missing	Zeroed	0	0	.	.
18 of topic.host_about Or Mean if Missing		72.942183	47.7382	1.53	0.1265
18 of topic.host_about Is Missing	Zeroed	0	0	.	.
19 of topic.host_about Or Mean if Missing		123.15981	40.98238	3.01	0.0027*
19 of topic.host_about Is Missing	Zeroed	0	0	.	.
1 of topic.house_rules Or Mean if Missing		65.258973	38.80192	1.68	0.0926
1 of topic.house_rules Is Missing	Biased	-2.53093	1.136855	-2.23	0.0260*
2 of topic.house_rules Or Mean if Missing		112.09046	38.05618	2.95	0.0032*
2 of topic.house_rules Is Missing	Zeroed	0	0	.	.
3 of topic.house_rules Or Mean if Missing		173.36682	38.72812	4.48	<.0001*
3 of topic.house_rules Is Missing	Zeroed	0	0	.	.
4 of topic.house_rules Or Mean if Missing		147.67274	40.29576	3.66	0.0002*
4 of topic.house_rules Is Missing	Zeroed	0	0	.	.
5 of topic.house_rules Or Mean if Missing		33.73529	38.20705	0.88	0.3773
5 of topic.house_rules Is Missing	Zeroed	0	0	.	.
6 of topic.house_rules Or Mean if Missing		167.79454	31.51968	5.32	<.0001*
6 of topic.house_rules Is Missing	Zeroed	0	0	.	.
7 of topic.house_rules Or Mean if Missing		64.497033	43.06708	1.50	0.1342
7 of topic.house_rules Is Missing	Zeroed	0	0	.	.
8 of topic.house_rules Or Mean if Missing		87.737799	38.90273	2.26	0.0241*
8 of topic.house_rules Is Missing	Zeroed	0	0	.	.
9 of topic.house_rules Or Mean if Missing		231.72023	38.86138	5.96	<.0001*
9 of topic.house_rules Is Missing	Zeroed	0	0	.	.
10 of topic.house_rules Or Mean if Missing		76.046004	40.40412	1.88	0.0598
10 of topic.house_rules Is Missing	Zeroed	0	0	.	.
11 of topic.house_rules Or Mean if Missing		72.837498	39.55278	1.84	0.0656
11 of topic.house_rules Is Missing	Zeroed	0	0	.	.
12 of topic.house_rules Or Mean if Missing		-81.60721	42.67926	-1.91	0.0559
12 of topic.house_rules Is Missing	Zeroed	0	0	.	.
13 of topic.house_rules Or Mean if Missing		63.222946	43.48332	1.45	0.1460
13 of topic.house_rules Is Missing	Zeroed	0	0	.	.
14 of topic.house_rules Or Mean if Missing		-8.299167	37.26344	-0.22	0.8238
14 of topic.house_rules Is Missing	Zeroed	0	0	.	.
15 of topic.house_rules Or Mean if Missing		71.344551	39.72053	1.80	0.0725
15 of topic.house_rules Is Missing	Zeroed	0	0	.	.
16 of topic.house_rules Or Mean if Missing		155.37422	39.75393	3.91	<.0001*
16 of topic.house_rules Is Missing	Zeroed	0	0	.	.
17 of topic.house_rules Or Mean if Missing		124.26647	30.55703	4.07	<.0001*
17 of topic.house_rules Is Missing	Zeroed	0	0	.	.
18 of topic.house_rules Or Mean if Missing		107.58837	35.43835	3.04	0.0024*
18 of topic.house_rules Is Missing	Zeroed	0	0	.	.
19 of topic.house_rules Or Mean if Missing		99.856771	32.78782	3.05	0.0023*
19 of topic.house_rules Is Missing	Zeroed	0	0	.	.
1 of topic.space Or Mean if Missing		82.814277	35.46188	2.34	0.0195*
1 of topic.space Is Missing	Biased	2.125961	1.988763	1.07	0.2851
2 of topic.space Or Mean if Missing		53.01534	41.13848	1.29	0.1975
2 of topic.space Is Missing	Zeroed	0	0	.	.
3 of topic.space Or Mean if Missing		58.456285	35.42718	1.65	0.0989
3 of topic.space Is Missing	Zeroed	0	0	.	.
4 of topic.space Or Mean if Missing		75.411263	41.92796	1.80	0.0721
4 of topic.space Is Missing	Zeroed	0	0	.	.
5 of topic.space Or Mean if Missing		59.990393	39.15167	1.53	0.1255
5 of topic.space Is Missing	Zeroed	0	0	.	.
6 of topic.space Or Mean if Missing		5.1965807	39.71865	0.13	0.8959
6 of topic.space Is Missing	Zeroed	0	0	.	.
7 of topic.space Or Mean if Missing		60.048023	41.09337	1.46	0.1440
7 of topic.space Is Missing	Zeroed	0	0	.	.
8 of topic.space Or Mean if Missing		50.16176	40.59583	1.24	0.2166
8 of topic.space Is Missing	Zeroed	0	0	.	.
9 of topic.space Or Mean if Missing		-56.19705	38.78209	-1.45	0.1473
9 of topic.space Is Missing	Zeroed	0	0	.	.
10 of topic.space Or Mean if Missing		80.778881	37.32747	2.16	0.0305*
10 of topic.space Is Missing	Zeroed	0	0	.	.
11 of topic.space Or Mean if Missing		165.25331	40.12093	4.12	<.0001*
11 of topic.space Is Missing	Zeroed	0	0	.	.
12 of topic.space Or Mean if Missing		117.84427	40.81408	2.89	0.0039*
12 of topic.space Is Missing	Zeroed	0	0	.	.
13 of topic.space Or Mean if Missing		-12.05985	37.00263	-0.33	0.7445
13 of topic.space Is Missing	Zeroed	0	0	.	.
14 of topic.space Or Mean if Missing		10.368363	41.07404	0.25	0.8007
14 of topic.space Is Missing	Zeroed	0	0	.	.
15 of topic.space Or Mean if Missing		-9.760877	41.0269	-0.24	0.8119
15 of topic.space Is Missing	Zeroed	0	0	.	.
16 of topic.space Or Mean if Missing		39.611958	33.78716	1.17	0.2410
16 of topic.space Is Missing	Zeroed	0	0	.	.
17 of topic.space Or Mean if Missing		12.597459	40.36013	0.31	0.7549
17 of topic.space Is Missing	Zeroed	0	0	.	.
18 of topic.space Or Mean if Missing		44.998238	41.24475	1.09	0.2753
18 of topic.space Is Missing	Zeroed	0	0	.	.
19 of topic.space Or Mean if Missing		46.458842	39.40647	1.18	0.2384
19 of topic.space Is Missing	Zeroed	0	0	.	.

Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.1974	90.285	40478.4
Validation Set	0.2596	74.270	10107.6