STA723 Case Study 2 - Group 5 Report

Youngsoo Baek, Phuc Nguyen, Irene Ji

Abstract

We modeled price and popularity of Airbnb listings in 2019 jointly using a bivariate-response regression with random effects. We found that boroughs and room types were significant predictors of listing price, while boroughs, listings' age, availability and the usage of foreign languages and whitespaces in listing names were predictors of popularity. Specifically, we can expect listings in Manhattan to be most expensive while listings in Queens to be most popular. Entire homes are on average most expensive, while younger listings are on average more popular.

1. Introduction

The Airbnb New York City Open Data, collected in 2019, contains information on 48,895 Airbnb listings in New York City. The dataset has 16 variables that contain different information about a listing. The analysis focuses on identifying interesting patterns among airbnb listings, especially influential factors for price and popularity.

2. Materials and Methods

We first removed 14 observations that have `minimum_nights` greater than 365 days, and imputed the `price` for listings with zero price as $5, which is half the minimum non-zero price available. We grouped `last_review` by year and used the difference between the last review year and 2019 as our predictor. Since it is unclear what quantity `availability_365` variable is measuring, and the variable measurement is extremely noisy, we created an indicator variable `available_spec` to indicate whether the listing is available (1) or not (0), based on the discussions on Kaggle forum [1].

For text analysis, we removed punctuations, numbers, and stop words and converted the listing names to lowercase. We then tokenized the names and calculated a frequency for each word over the whole corpus. Before fitting our main model (bivariate mixed-effect regression), we applied XGBoost to explore the importance of each predictor and conducted Pearson's chi-squared test to examine the heterogeneity across boroughs. Price and monthly reviews were grouped into 3 levels (below Q1, between Q1 and Q3, above Q3) for the chi-squared test. We also identified potentially influential characteristics and topics within the listing names using term-frequencies and included them in our main model below.

To understand which factors influence the price and popularity of a listing, we fit a bivariate response linear regression that has varying intercepts across different neighborhoods and boroughs. For the i-th listing in neighborhood j, within borough k, the responses of interest are price and monthly reviews, on log scale:

$$\left(\begin{array}{c} \text{Log Price}_{k[j[i]]} \\ \text{Log Monthly reviews}_{k[j[i]]} \end{array} \right) = \left(\begin{array}{c} \beta_1^T \mathbf{X}_i \\ \beta_2^T \mathbf{X}_i \end{array} \right) + \eta_{k[j]} + \theta_j + \epsilon_{k[j[i]]},$$

where $\eta_{k[j]}$ is the 2×1 vector of neighborhood-level random effect, θ_j is the borough-level random effect, and ϵ is an observation error that has $N(0, \sigma^2 \mathbf{I}_2)$ distribution. The random intercepts are assumed to have a bivariate normal distribution centered at zero, so we can estimate the between-response correlation between boroughs and between neighborhoods, within boroughs. The model is fit through a standard MLE procedure implemented in `lme4` package for `R`. The slopes remain fixed across different groups, and the predictors for the two responses remain identical for the two responses.

3. Results

3.1 Exploratary Data Analysis

As shown in Figures 1 and 2, variable importance plots of XGBoost suggest that room type, availability, monthly reviews, and boroughs are the most influential factors for the price of Airbnb. Last review year, availability, minimum nights and price are most important for monthly reviews. Figures 3 to 5 present boxplots and mosaic plots for price, monthly review and room type across boroughs. As shown in the plots, price, popularity and room type differ across boroughs. For example, Manhattan has the highest price and most listings there are entire homes or apartments. Queens, on the other hand, has the highest popularity and private rooms take up the most listings there. This agrees with the chi-squared test results, where the tests are all significant with p-values $< 2\text{e-}16$, suggesting there exists heterogeneity across boroughs. Figure 6 shows that the words "apartment", "bedroom", "private", "cozy", and "apt" appear most often in the corpus. We also removed the top five most frequent words and plotted words sized by co-occurrences with the boroughs' names in Figure 7. The frequent usage of "min" or "minutes" suggests proximity is a common theme in listings related to Manhattan. Listing names in other boroughs are more often related to famous landmarks, such as brownstone in Brooklyn, the Astoria neighborhood in Queens, or the zoo in the Bronx. Since we noticed substantial occurrences of foreign characters, misspellings and lack of whitespaces in listing names, we created indicators for these characteristics.

3.2 Main Results

Figures 9 and 10 each plot the interval estimates for the predictor coefficients for price and monthly reviews. Exact estimates, standard errors, and t-test statistics are shown in Table 1. Together, these results show that notwithstanding the "significance" of many of the predictors based on naive t-tests, many of the coefficients have too small a magnitude to have a meaningful impact on price or popularity adjusted for others. This is especially apparent when we compare them to the estimated standard deviations and correlations for the random intercepts (Tables 2 and 3). The only factor that has a greater influence on average on price than the heterogeneity across different neighborhood groups is a listing's room type, as apartments/entire rooms are the most expensive on average, followed by private and shared rooms. For monthly review rate, influential factors include a listing's age, and whether a listing is available at the moment data were collected (non-zero `availability_365` variable). The negative coefficient for `last_review` implies a listing must be younger on average to have more reviews per month. Listing names in foreign languages on average also have higher monthly reviews, while those with no whitespaces have lower monthly reviews.

3.3 Model Diagnostics

Our model does not explicitly account for possible spatial structure within neighborhoods. Assuming there exists such an autocorrelation structure, the semivariogram γ of the response Y, which is defined as

$$\gamma(||\mathbf{h}||) \equiv \frac{1}{2}\mathrm{E}[Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})],$$

will exhibit a clear pattern in increasing pairwise distance $d \equiv ||\mathbf{h}||$. Here, $\mathbf{s}$ indexes the listing's location, and $\mathbf{h}$ is the displacement vector. A customary, simple nonparametric estimator for semivariogram is [2]

$$\hat{\gamma}(d) = \frac{1}{2|N(d)|}\sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(d)}[Y(\mathbf{s}_i)-Y(\mathbf{s}_j)]^2,$$

where $N(d)$ consists of all pairs of locations that have a pairwise, geodesic distance d. The semivariogram estimators for each of the residuals are plotted on increasing pairwise distance in Figure 11. It is clear based on the plots that while for price, no evidence of significant spatial structure is seen conditional on the neighborhood effects, a negative spatial autocorrelation exists for monthly review rates: closer things have more different responses, rather than similar. While seemingly counterintuitive, the negative autocorrelation can be explained by strong competition between listings when in close proximity. Within the same neighborhoods, a potential customer is always being sapped away from one listing to another, so the latter has at least one more review. Such effect will be strongly visible when listings are closer, but as distance increases, the dependence will become weak to none, and the variability of the difference between two monthly review rates will stabilize. That we do not see a similar negative autocorrelation in price is informative, as it suggests the hosts' pricing policy remains indifferent to their neighboring hosts, conditional on the neighborhood they belong to.

4. Discussion

We found that boroughs and room types were significant predictors of listing price, while boroughs, listings' age, availability and the usage of foreign languages and whitespaces in listing names were predictors of popularity. Specifically, we can expect listings in Manhattan to be the most expensive while listings in Queens to be the most popular. Entire homes are on average most expensive, while younger listings are on average more popular. Listing names written without whitespaces were less popular, while those in foreign languages were more popular on average. Based on these results, potentially profitable listings could be "Luxury suit minutes from Midtown Manhattan" or "Entire home in Astoria, Queens".

What we discovered through EDA and model diagnostics has important implications for the application of more complex joint models. Most importantly, negative spatial autocorrelation for within-neighborhood monthly review rates raises concern for common spatial models including conditional autoregressive (CAR) models, which can only model positive autocorrelation. Models for repulsive point processes also exist in the literature, but our diagnostics show care needed in jointly modeling the two responses when taking into account possible spatial structures.

Finally, our model can be improved by letting the slopes also vary by neighborhoods and boroughs. While this model has not been implemented yet, mainly due to facilitating computational ease, it is expected that Bayesian estimation procedure with the choice of prior inducing strong shrinkage will improve the estimation of many slope parameters.

References

Sarthak, N., Post: "Availability_365=0?", Discussion thread: New York City Airbnb Open Data, Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/discussion/111835

Banrjee, S., Carlin, B. P., and Gelfand, A. E. (2011). *Hierarchical Modeling and Analysis for Spatial Data.*
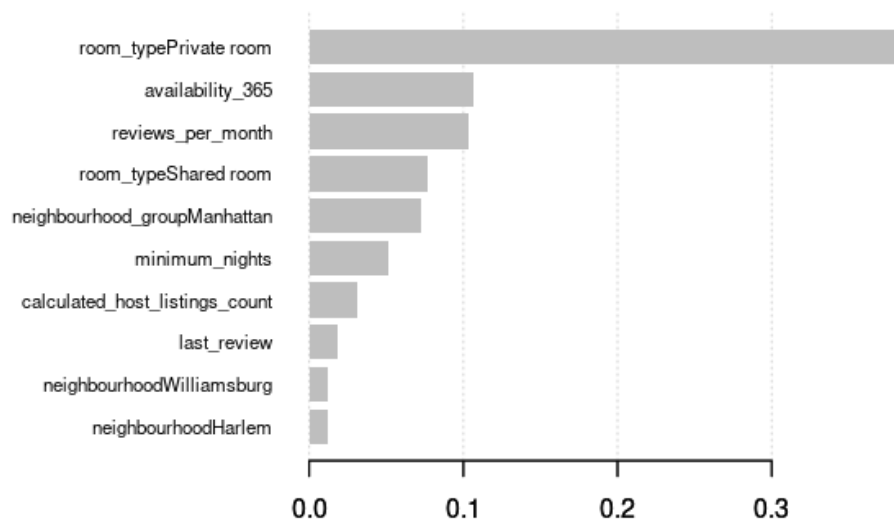
Figures and Tables

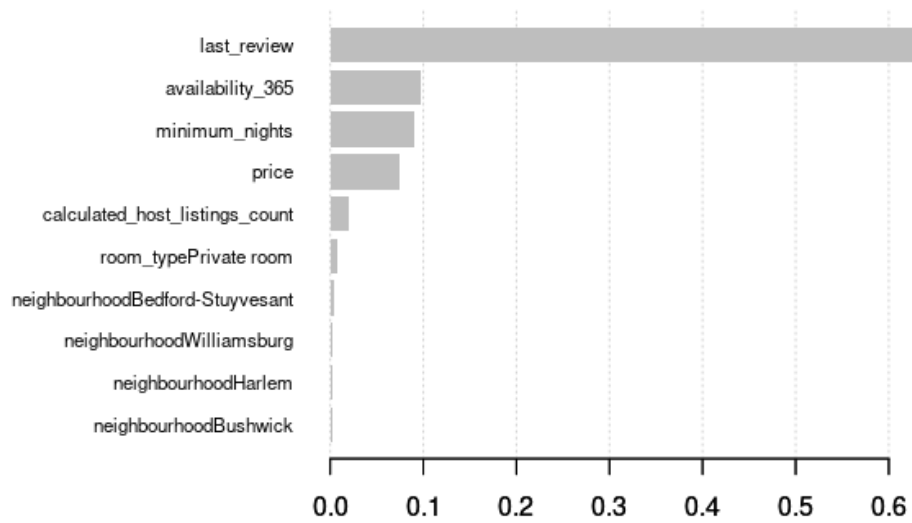Figure 1: XGBoost - Variable Importance Plots for Price

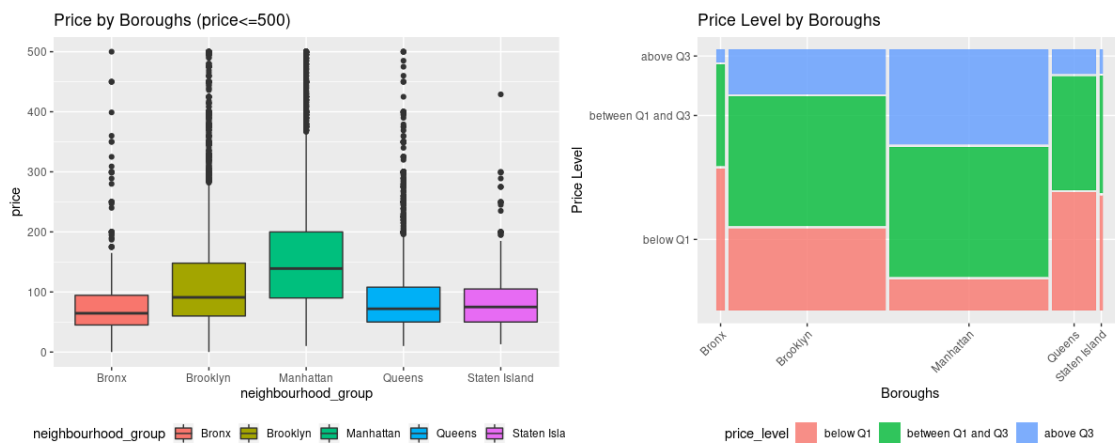Figure 2: XGBoost - Variable Importance Plots for Monthly Review

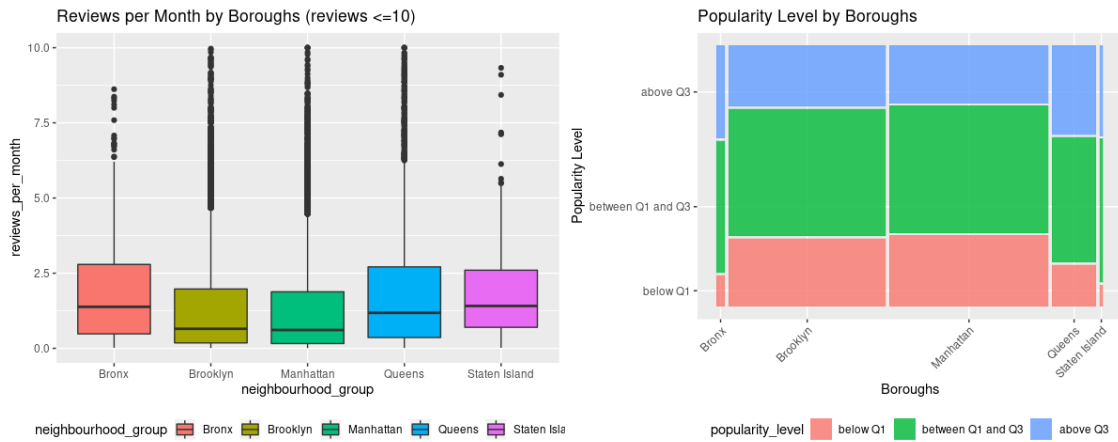Figure 3: Boxplot and Mosaic Plot for Price across Boroughs

Figure 4: Boxplot and Mosaic Plot for Monthly Review across Boroughs
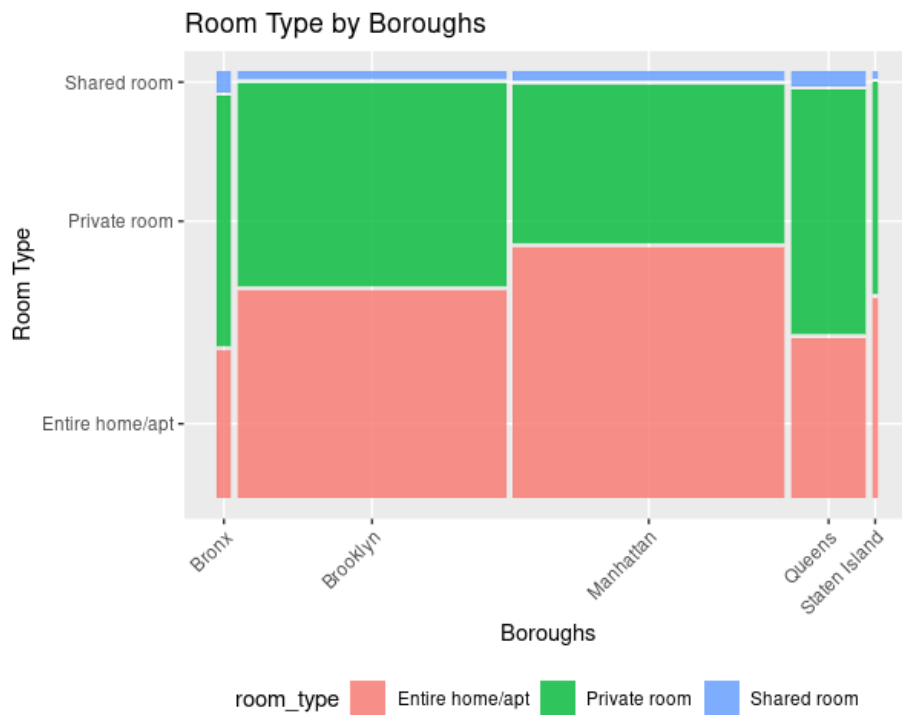

Figure 5: Mosaic Plot for Room Type across Boroughs

Figure 6: Wordcloud of words in listing names, sized relative to their term-frequencies in the corpus.

Figure 7: Wordcloud of words co-occurring with boroughs' names, sized by number of co-occurrences.
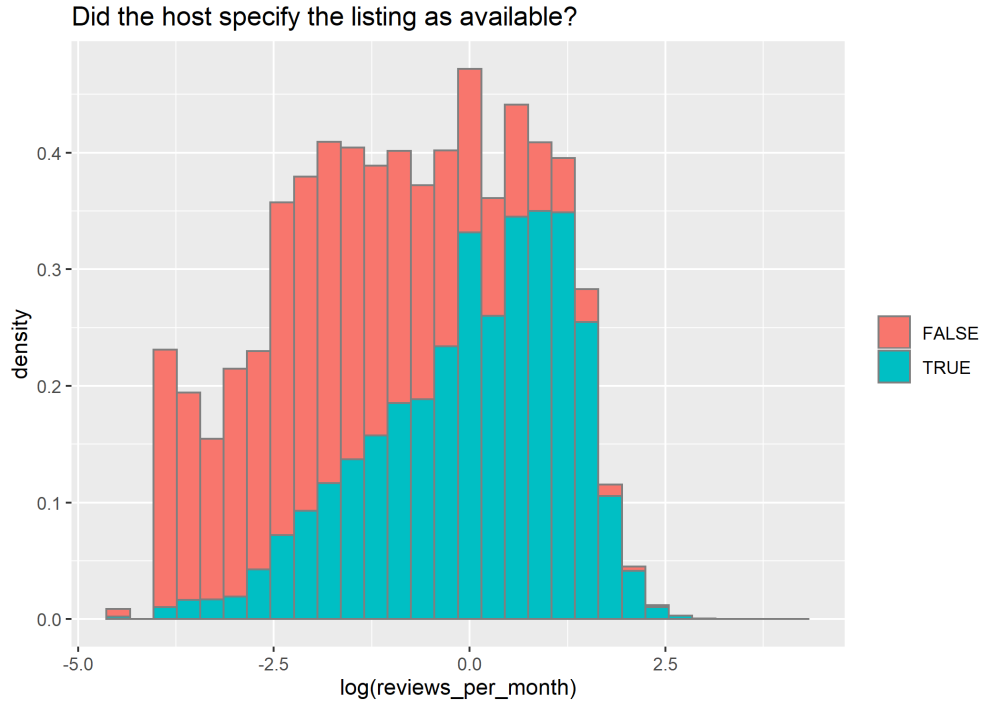
Figure 8: Distribution of log monthly review rates for listings with zero/non-zero availability feature.
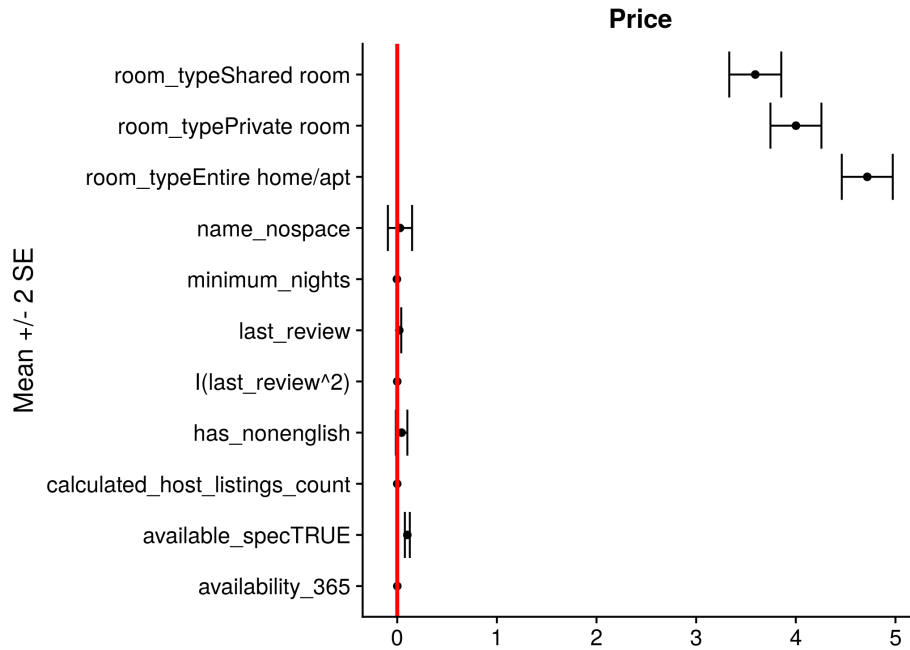

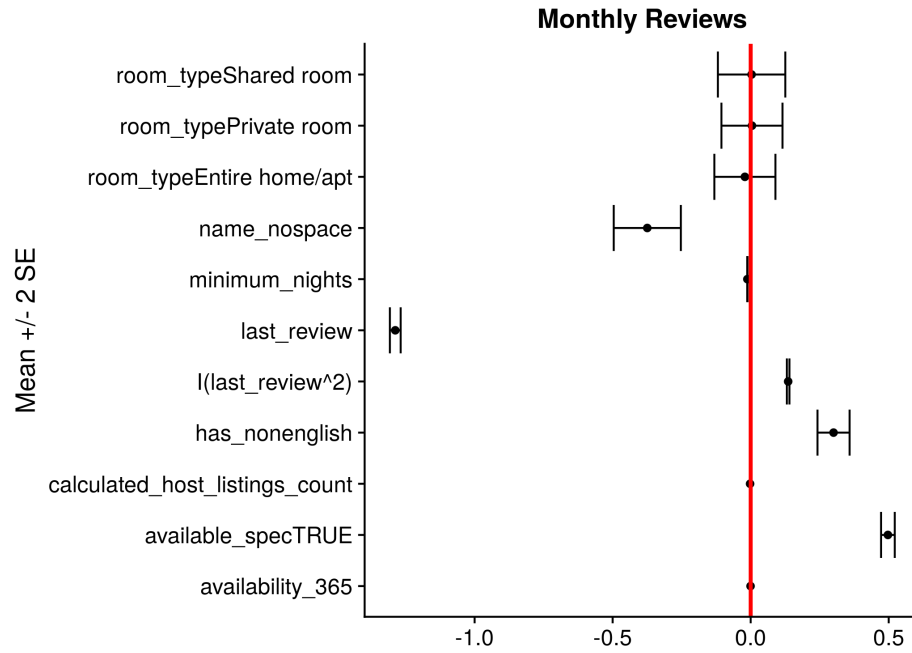Figure 9: Fixed effect estimates for log price.

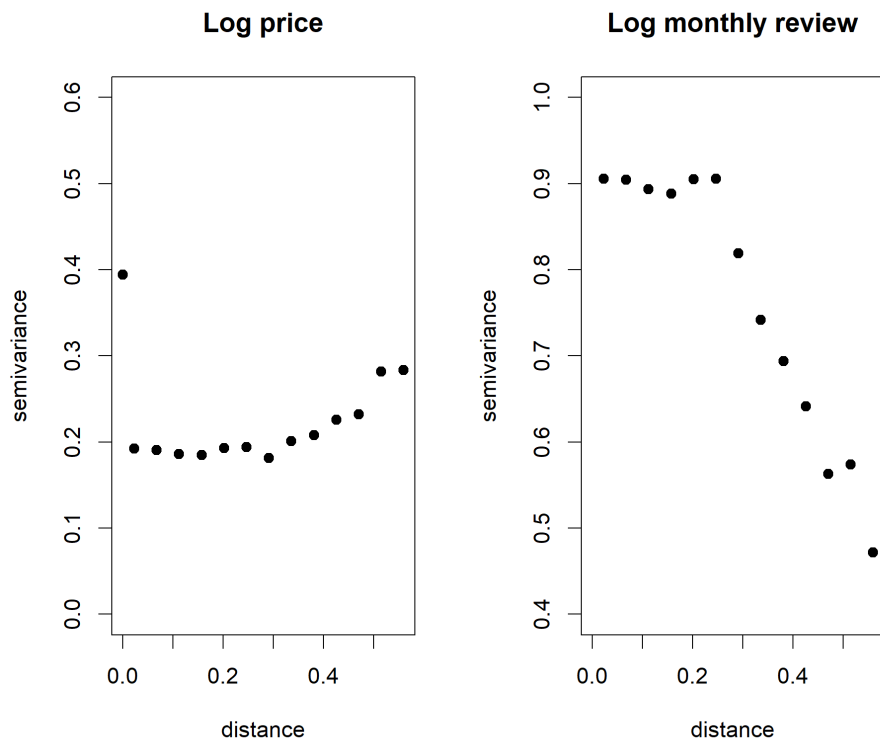Figure 10:   Fixed effect estimates for log monthly review rates.

Figure 11:   Semivariogram estimators calculated from model residuals for price and monthly review rates.

Table 1: Coefficient estimates for price

	Estimate	Std. Error	t value
room_typeEntire home/apt	4.72	0.13	36.96
room_typePrivate room	4.00	0.13	31.33
room_typeShared room	3.59	0.13	27.56
last_review	0.02	0.01	2.00
I(last_review^2)	0.00	0.00	-0.07
minimum_nights	0.00	0.00	-13.26
availability_365	0.00	0.00	15.04
calculated_host_listings_count	0.00	0.00	-1.96
available_specTRUE	0.10	0.01	8.42

Table 2: Coefficient estimates for monthly reviews

	Estimate	Std. Error	t value
room_typeEntire home/apt	-0.02	0.05	-0.43
room_typePrivate room	0.00	0.05	0.05
room_typeShared room	0.00	0.06	-0.03
last_review	-1.29	0.01	-133.01
I(last_review^2)	0.14	0.00	57.96
minimum_nights	-0.01	0.00	-43.03
availability_365	0.00	0.00	-11.75
calculated_host_listings_count	0.00	0.00	-11.99
available_specTRUE	0.49	0.01	39.94

Table 3: Random intercept fits

	Price	Reviews
Bronx	-0.21	0.05
Brooklyn	0.07	-0.12
Manhattan	0.45	-0.11
Queens	-0.10	0.11
Staten Island	-0.22	0.06

Table 4: Variance component estimates

Group	Variable	StdDev	Corr
Neighborhood	Price	0.17	NA
	Reviews	0.19	-0.23
Borough	Price	0.28	NA
	Reviews	0.11	-0.76
Residual		0.74	NA