Patterns of Airbnb Listings in NYC

Frances Hung, Yunran Chen, Keru Wu

Abstract

Airbnb home rental listings vary in price and popularity, and it is natural to explore reasons for this variance. We apply a multilevel conditional autoregressive Bayesian model to capture association between certain Airbnb rental characteristics (including neighbourhood location) and listing price/popularity in NYC. Room type and minimum nights required are the most influential factors for price and popularity respectively. Adjusting for influential factors, Midtown South in Manhattan and East Elmhurst (close to LaGuardia Airport) in Queens are the most expensive and the most popular neighbourhoods respectively. With respect to a balance between price and popularity, Yorkville in Manhattan is the most lucrative host neighborhood. Text analysis suggests including location, room type and positive adjectives in names of listings.

1. Introduction

Airbnb is a platform providing home rentals for travelers. Our observed data consists of 48,895 individual Airbnb listings in New York City. Each listing observation contains the following variables: host ID, neighbourhood group, neighbourhood, longitude/latitude, available days of the listing in a year, room type, price, minimum nights required, number of reviews, and reviews per month.

From the perspective of a host, we are interested in exploring the patterns in price and popularity. Specifically, we are interested in (1) quantifying the influential factors in the price/popularity and evaluating their influence (2) finding the most valuable neighborhoods adjusted for the influential factors (3) optimally choosing a location and a price for the listing (4) optimally naming the listing.

2. Materials and Methods

Since the price and popularity are strongly related to the location of listings (Fig. 1, 2), and neighborhoods provide a natural boundary for spatial characteristics of listings, we consider a multilevel conditional autoregressive Bayesian model (CARBayes)(Lee 2013) based on neighborhood units as follows:

$$Y_{kj}|\mu_{kj} \sim f(y_{kj}|\mu_{kj}, \nu^2), \quad \begin{array}{l} k = \text{neighbourhood} = 1, ..., K \\ j = \text{listings} = 1, ..., m_k \end{array}$$

$$g(\mu_{kj}) = x_{kj}^T \beta + \psi_{kj}$$

$$\psi_{kj} = \phi_k + \zeta_{kj}$$

, where β represents the potential effect of predictor x_{kj}, with a prior $\beta \sim N(\mu_\beta, \Sigma_\beta)$. ϕ_k and ζ_{kj} represents the neighbourhood effect and individual effect respectively. We consider an autoregressive prior for ϕ_k:

$$\phi_k|\phi_{-k} \sim N\Big(\frac{\rho \sum_{l=1}^K w_{kl}\phi_j}{\rho \sum_{j=1}^K w_{kl} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^K w_{kl} + 1 - \rho}\Big)$$

where $w_{kl} \in \{0, 1\}$ is known from data and $w_{kl} = 1$ denotes that neighbourhood k is adjacent to neighbourhood l. $\rho \sim U(0, 1)$ captures the relation between neighbourhood effects. In summary, this prior captures the spatial structure among neighbourhoods; each neighborhood's effect is centered at the weighted sum of effects from its neighbors.

We consider `log(price)` and `log(1+review_per_month)`(popularity) as response variables and model them separately. We include room type, price, minimum nights required, and popularity/price respectively as predictors based on EDA results. Additionally, we incorporate the logarithm distance from a

listing to the nearest metro station to account for the heterogeneity of individual spatial effects within the same neighborhood. We extracted features from names of listings by applying Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) model and introduced these features as predictors.

To carry out text analysis on names of listings, we first conducted a detailed text cleaning and applied Porter's stemmer algorithm to merge the words with the same root. Then we applied Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) to explore the latent topics. By assigning each word a weight of related topics (e.g. adjectives, locations), we extracted features from the listings' names and included them in our multilevel CARBayes model. In addition, we conducted word frequency analysis for different boroughs and different levels of price and used wordclouds to visualize the results.

3. Results

3.1 Exploratory Data Analysis

Initial data exploration suggests a clear spatial structure for price, popularity, and traffic (Fig. 1,2,3). High-priced listings are concentrated in midtown/downtown Manhattan with some spread into the part of Brooklyn closest to Manhattan; traffic follows a similar pattern. In contrast, most popular neighborhoods are located around the LGA airport. Room types appear to be strongly correlated with price, but not popularity (Fig. 4,5). They also seem to be heterogeneous across boroughs and neighborhoods (Fig. 6), and we corroborate this with a Pearson's Chi-squared test (p-value $<$ 2.2e-16). Our graphs suggest a non-linear effect of room type on price/popularity (Fig. 7,8).

3.2 Data Preprocessing

We remove 11 listings with price equal to 0 and impute 0 for listings with NA `reviews_per_month` values since they correspond to listings with zero-valued `number_of_reviews`. To improve scaling, we use a logarithm transformation for the response variables `price` and `reviews_per_month`. The choice of predictors are based on the results from EDA; we choose `reviews_per_month` as a proxy for popularity. Furthermore, we categorize `minimum_night` into 5 groups in order to account for it's nonlinear association with the response variables. To obtain the adjacency matrix of neighborhoods in NYC, we incorporate shape files for neighborhoods in New York [1] and reallocate the listings' neighborhoods based on latitude and longitude. To account for heterogeneity of spatial effects across listings within the same neighborhood, we introduce a new predictor: the logarithm of distance from a listing to the closest metro[2]. In order to carry out text analysis, we first preprocess the listings' names by transforming them to lower case and removing non-informative characters such as punctuations, stopwords, whitespace, and numbers. We then apply Porter's stemmer algorithm (Porter 2001) for word normalization, which extracts the common roots of informative words.

3.3 Main Results

From our model coefficient estimation (Fig. 9), our multilevel CAR model on price demonstrates the following patterns (numbers in parentheses are medians of corresponding coefficients). Entire rooms (0) are more expensive than private ones (-0.7), which in turn are more expensive than shared ones (-1.1). Manhattan (0.57) is the most expensive borough, and the Bronx (0) the cheapest. Availability (0.12) is positively correlated to price while reviews per month is negatively correlated. In addition, more strict requirements on minimum nights and longer distance to metro stations result in lower price. Room type is the most influential factor since compared to removing other predictors, our wAIC increases the most when it is removed from the full price model (Table 1).

[1] <https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq>

[2] Locations of metro stations: <https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49>

Our model on popularity (Fig. 10) yields mostly similar correlation signs but is different as follows. Compared to the other four boroughs, Queens (0.13) has the highest average review rate. Availability (0.15) still has a positive effect on popularity, while higher price (-0.12) corresponds to less popular listings. Moreover, metro distance is no longer significant for predicting popularity. Minimum nights is the most influential factor since our wAIC increases most when it is removed from the full popularity model (Table 2).

Heterogeneity across neighbourhoods is shown in Fig. 11 and 12. As shown in 11, neighbourhoods in Manhattan are more expensive on average, and their confidence intervals are narrower than in other boroughs. Fig. 13 and 14 present the posterior median of neighbourhoods' effects for price and popularity respectively. Among all neighbourhoods, Midtown South in Manhattan is the most expensive one, while New Drop-Midland Beach in Staten Island is the one with lowest prices. On the other hand, East Elmhurst (close to LaGuardia Airport) in Queens is the most popular neighbourhood, and Co-op City is the most unpopular one. If we consider the top 20 neighbourhoods for price and popularity seperately, one neighborhood appears in both: Yorkville in Manhattan (highlighted in both Fig. 13 and 14).

Our text analysis (Fig. 15, 16) indicates some critical words related to price: luxury, manhattan, beautiful. We also carry out LDA to find latent topics in listing names. Four discernable topics we found were adjectives, locations, Brooklyn-related and Manhattan-related words. Adding these 4 topics into our price model (as 4 indicators), we conclude that Brooklyn and Manhattan-related words have a positive significant coefficient, while the other two coefficients are significantly negative.

3.4 Sensitivity Analysis

The availability_365 variable has zero-valued observations which may correspond to hosts who temporarily take their listings off the market. Comparing the distribution of other variables for zero-valued vs. positive-valued availability_365 observations suggests that the data may be missing at random because we don't see an obvious pattern in missingness. Using MICE (Buuren and Groothuis-Oudshoorn 2010), we impute the data, treating the zero-valued observations as missing values.

Our model using the imputed data had indistinguishable AIC with our model without imputed data. As a result, we choose to use the original dataset and in future work, explore missingness of availability_365 further.

5. Discussion

Our multilevel CAR Bayesian model successfully discovers patterns of listings addressing both neighborhood level and individual level potential effects. We capture the spatial information at only the neighborhood levels, which facilitates interpretation as well as eases computation. However, the heterogeneity across individual spatial information within the same neighborhood may not be well caputured. To address the heterogeneity at the individual-level, a hierachical point-reference spatial model may be a better choice.

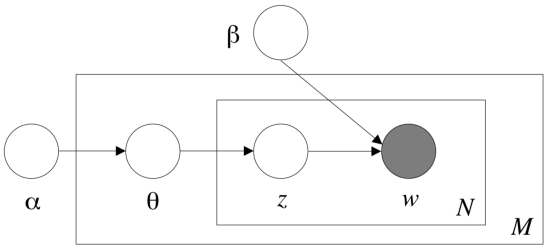
We assume linear relationships between response variables and predictors such as availability and distance to the closest metro station. We consider categorizing minimum night to account for the nonlinear effect we discovered in EDA. To better capture the nonlinear relationship and obtain a more flexible model, a nonlinear model using spline regression such as GAM would be more reasonable.

Another critical part of this analysis imputing missing data. Although MICE doesn't perform better than imputing with 0, exploring other imputation methods could be helpful. Moreover, since different hosts have different numbers of listings, we can further try approaches that account for their influence (e.g. random effects).

Appendix

Latent Dirichlet Allocation

- Terms:
  - Corpus $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}$
  - Doument $\boldsymbol{w} = \{w_1, w_2, ..., w_N\}$
  - Word $w_i \in \{1, ..., V\}$, V is total number of unique words.
- LDA Model:
  For all document $\boldsymbol{w}$ in D:
  1. $N \sim$ Poisson(ξ)
  2. $\theta \sim$ Dir(α)
  3. For word w_n $(n = 1, ..., N)$
  (a) choose a topic $z_n|\theta \sim$ Multinomial(θ)
  (b) choose a word $w_n|z_n, \beta \sim$ Multinomial(β_{z_n})



Tables

Model	All var	Room type	Availability	Reviews	Night	neighborhood
WAIC	63998	85372	66426	64501	66023	70860

Table 1: WAIC for model on price: without 1 variable

Model	All var	Room type	Availability	Price	Night	neighborhood
WAIC	74803	75370	78011	75297	80749	75881

Table 2: WAIC for model on popularity: without 1 variable

Figures

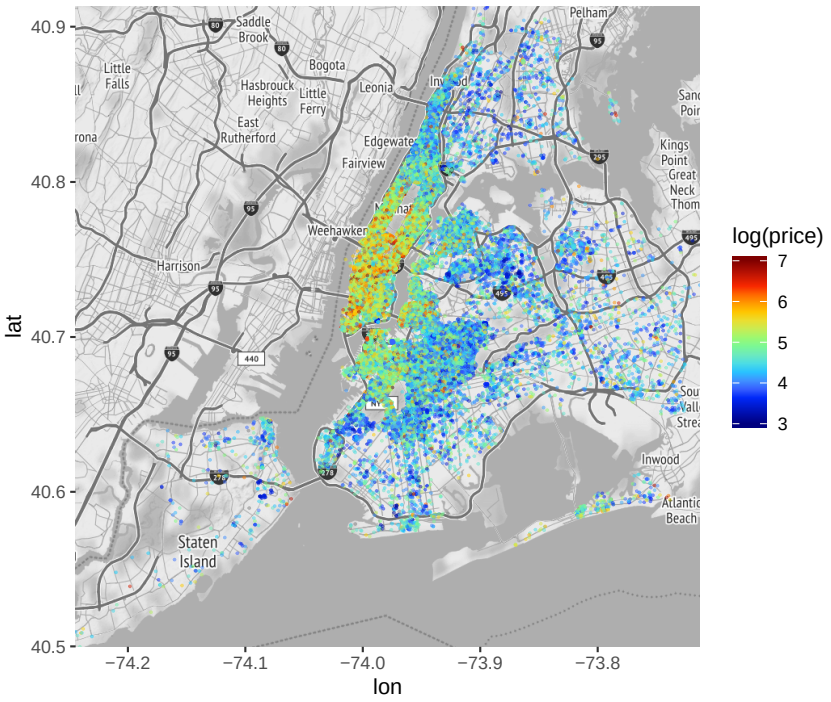


Figure 1: Distribution of log(price)

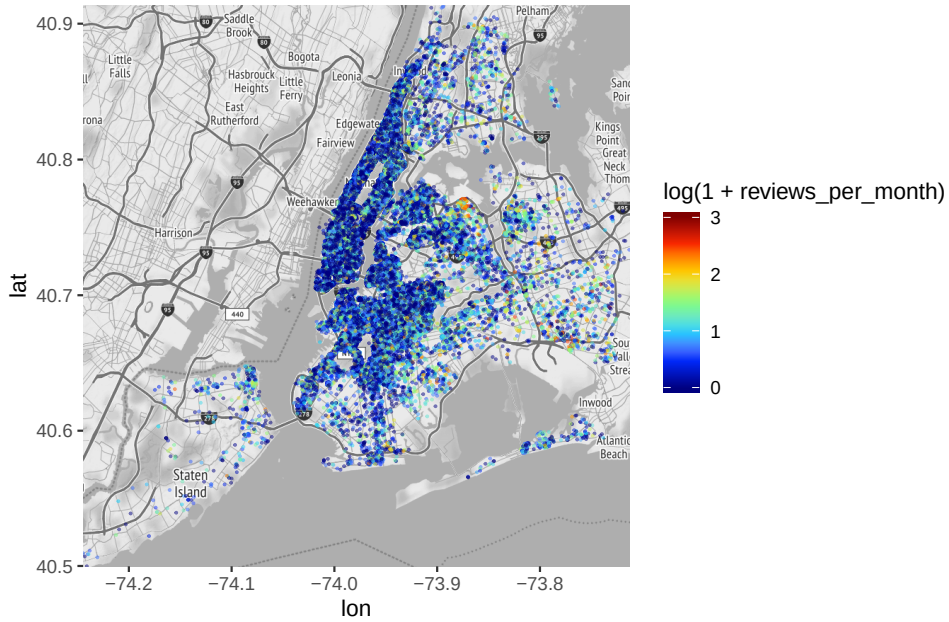


Figure 2: Distribution of log(1+reviews/mon)

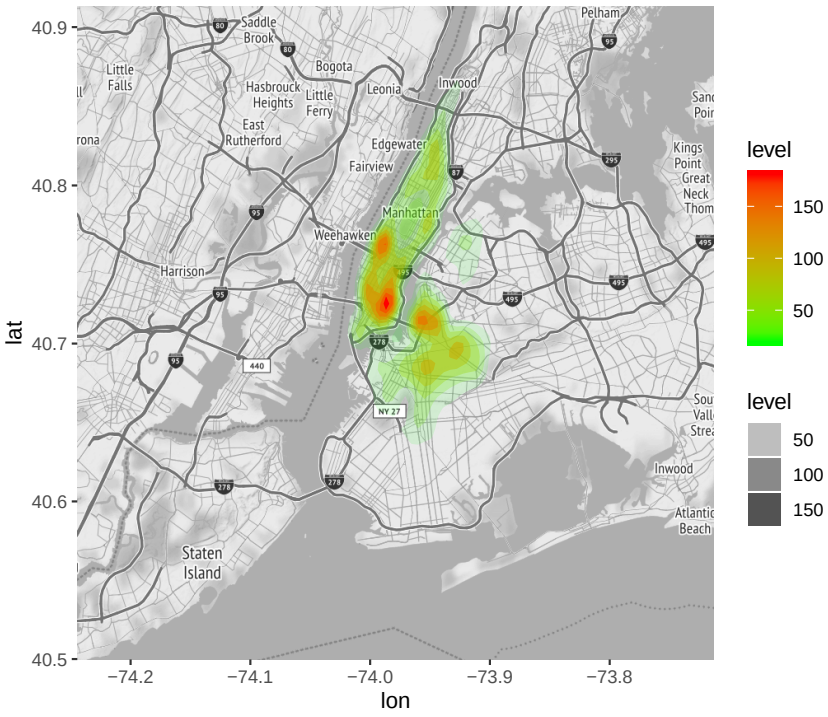


Figure 3: 2D-Density estimation

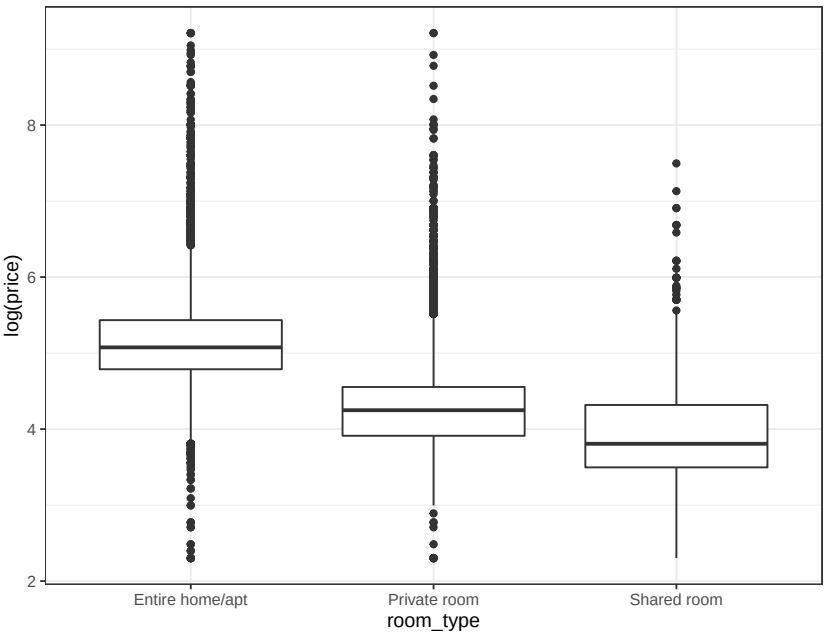


Figure 4: Association between price and room type

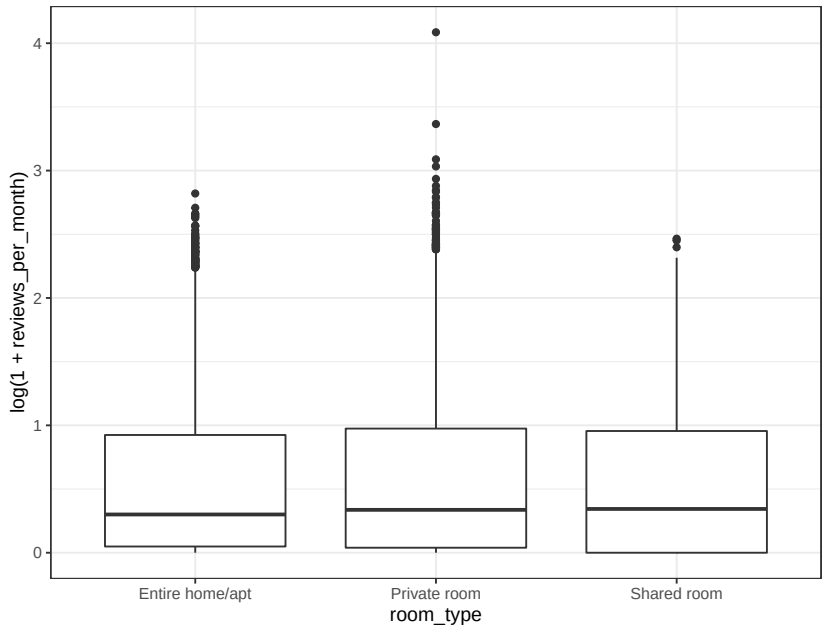


Figure 5: Association between review/mon and room type

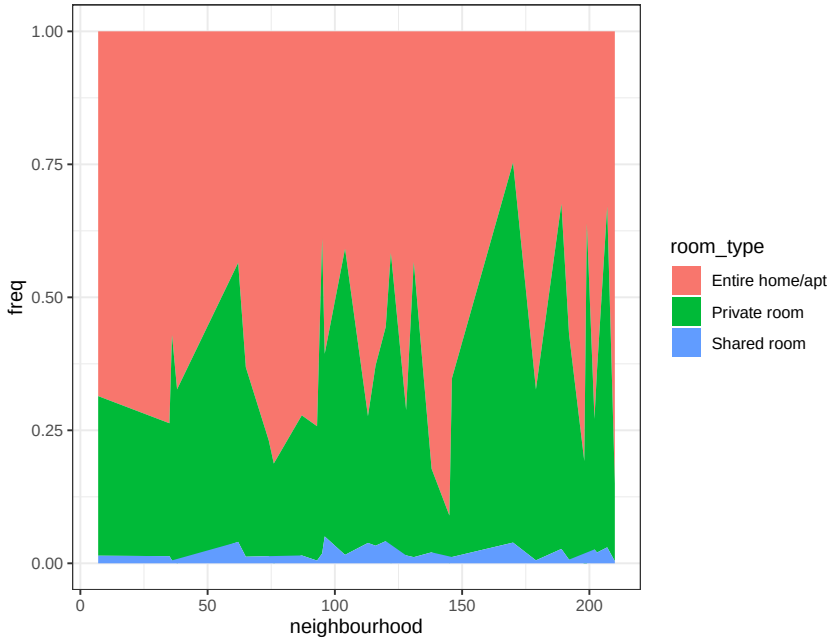


Figure 6: Heterogeneity of Room Type Across Neighborhoods (Manhattan)

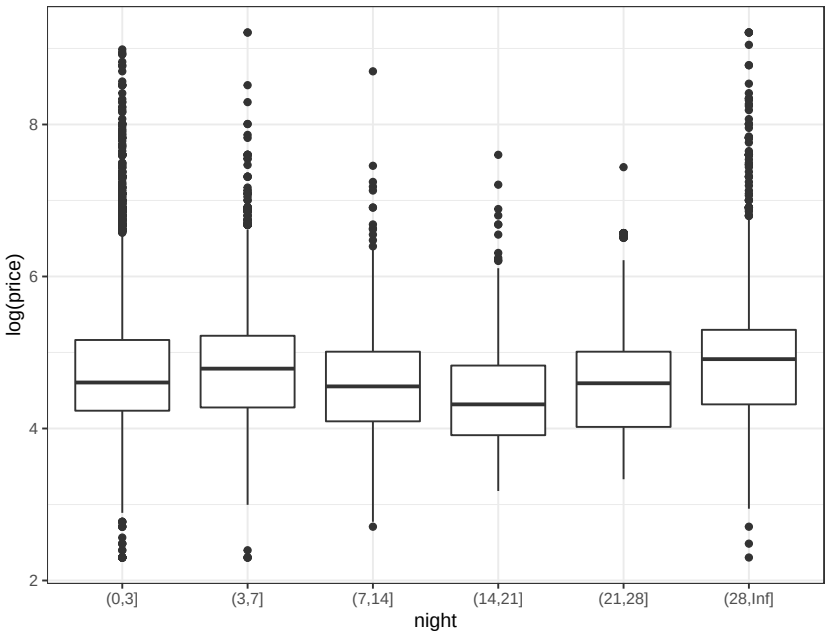


Figure 7: Association between price and minimum night

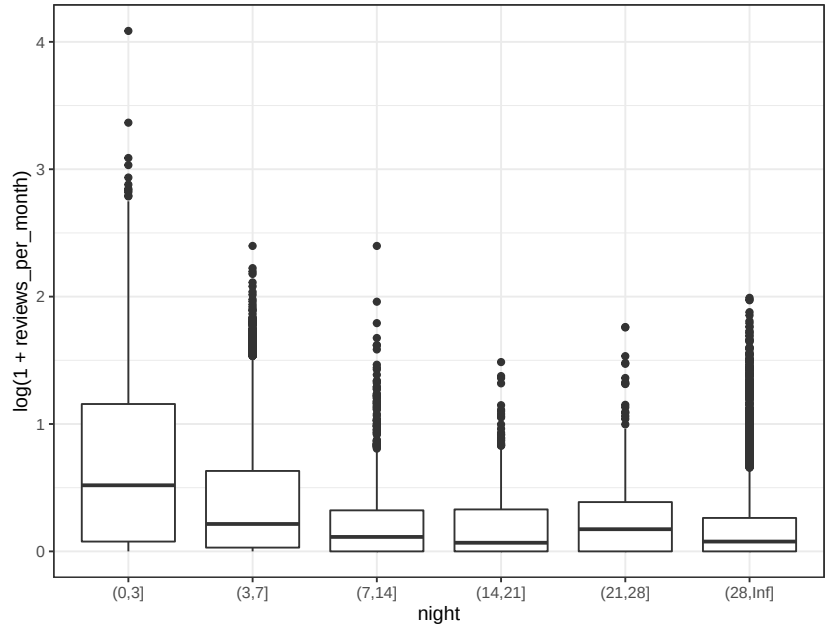


Figure 8: Association between review/month and minimum night

	Median	2.5%	97.5%
(Intercept)	4.8153	4.7443	4.8862
room_typePrivate room	-0.7238	-0.7322	-0.7142
room_typeShared room	-1.1091	-1.1379	-1.0836
neighbourhood_groupBrooklyn	0.1874	0.1089	0.2657
neighbourhood_groupManhattan	0.5775	0.4893	0.6526
neighbourhood_groupQueens	0.0964	0.0280	0.1787
neighbourhood_groupStaten Island	0.0404	-0.0698	0.1578
availability_365	0.1174	0.1129	0.1222
log(1 + reviews_per_month)	-0.0919	-0.1008	-0.0835
night(3,7]	-0.0758	-0.0871	-0.0646
night(7,14]	-0.2247	-0.2490	-0.2005
night(14,21]	-0.2865	-0.3193	-0.2503
night(21,28]	-0.2536	-0.3088	-0.2053
night(28,Inf]	-0.3288	-0.3452	-0.3141
metrodist	-0.0054	-0.0124	0.0017
topic1TRUE	-0.0655	-0.0767	-0.0532
topic2TRUE	0.0434	0.0270	0.0608
topic3TRUE	-0.0164	-0.0270	-0.0063
topic4TRUE	0.0283	0.0175	0.0391

Figure 9: CAR Model on price - Model Summary

	Median	2.5%	97.5%
(Intercept)	1.2715	1.1960	1.3567
room_typePrivate room	-0.1425	-0.1538	-0.1303
room_typeShared room	-0.2697	-0.3008	-0.2335
neighbourhood_groupBrooklyn	0.0065	-0.0621	0.0646
neighbourhood_groupManhattan	0.0026	-0.0746	0.0660
neighbourhood_groupQueens	0.1284	0.0507	0.1882
neighbourhood_groupStaten Island	0.0491	-0.0545	0.1513
availability_365	0.1508	0.1457	0.1553
log(price)	-0.1153	-0.1255	-0.1062
night(3,7]	-0.2439	-0.2560	-0.2314
night(7,14]	-0.4046	-0.4324	-0.3760
night(14,21]	-0.4521	-0.4925	-0.4141
night(21,28]	-0.4421	-0.5008	-0.3836
night(28,Inf]	-0.6003	-0.6175	-0.5833
metrodist	0.0007	-0.0071	0.0081
topic1TRUE	-0.0537	-0.0678	-0.0401
topic2TRUE	0.0099	-0.0112	0.0298
topic3TRUE	0.0006	-0.0115	0.0115
topic4TRUE	0.0318	0.0201	0.0447

Figure 10: CAR Model on popularity - Model Summary

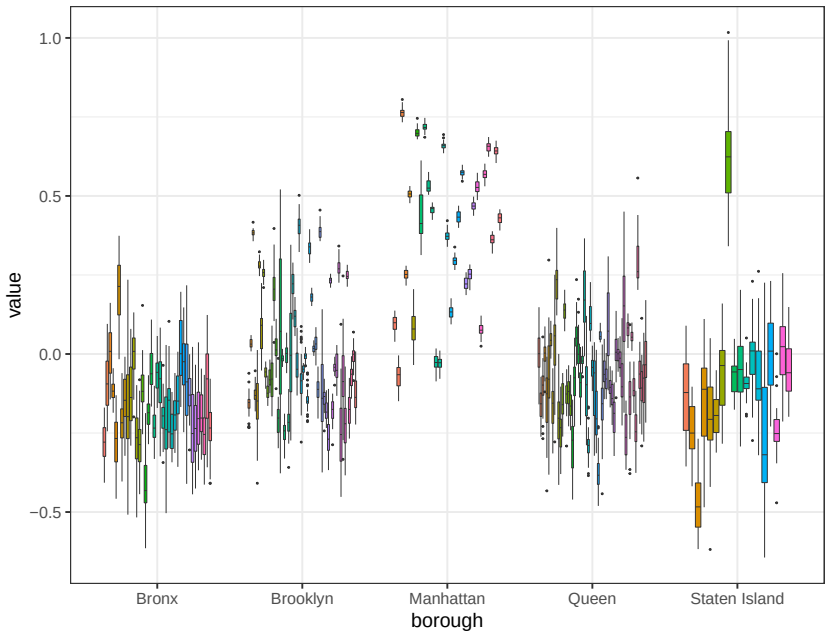


Figure 11: CAR Model on price - Neighbourhoods

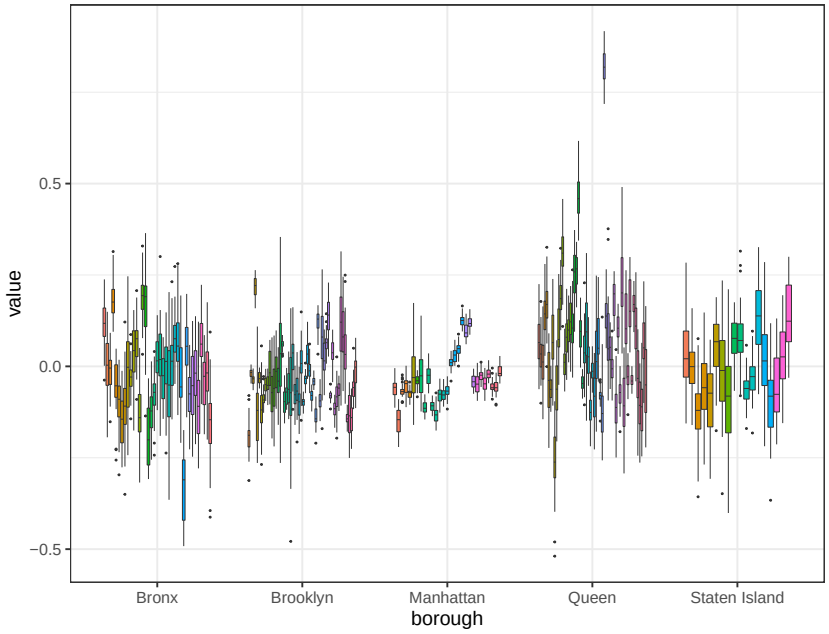


Figure 12: CAR Model on popularity - Neighbourhoods

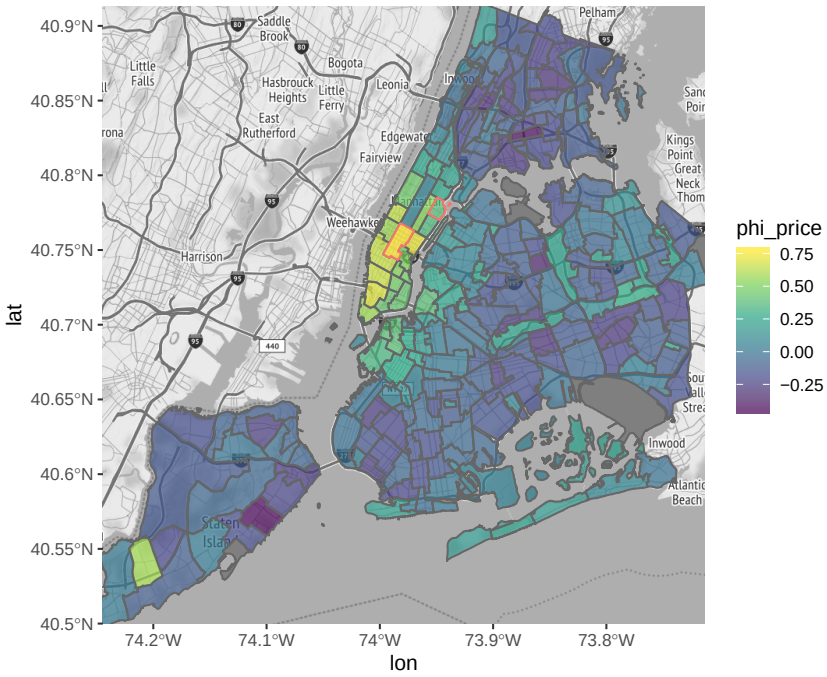


Figure 13: Neighborhoods' effects for price

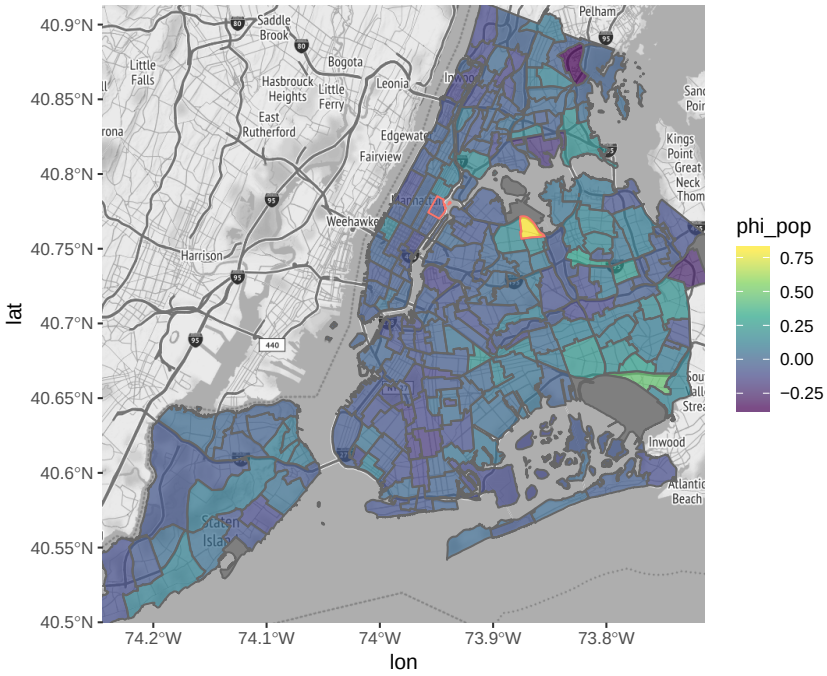


Figure 14: Neighborhoods' effects for popularity

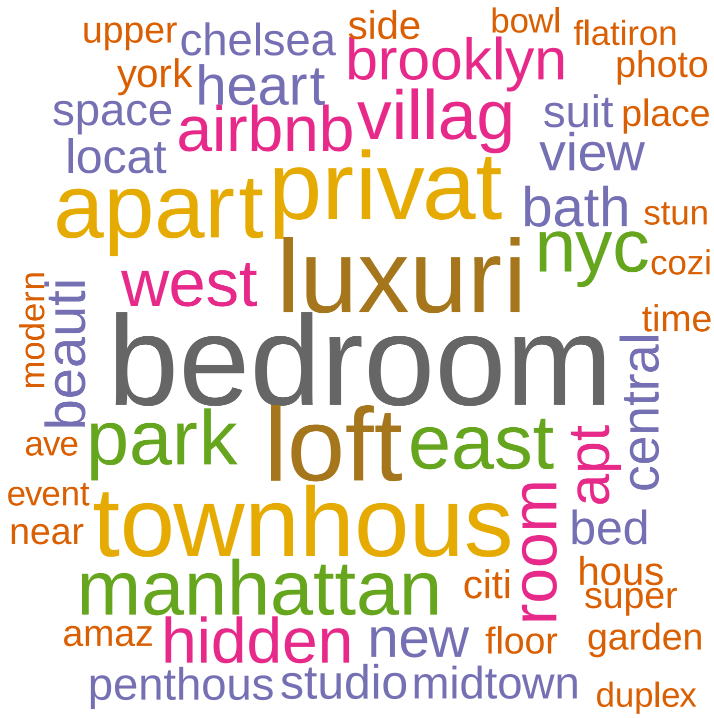


Figure 15: Wordcloud for listings with price > 2000

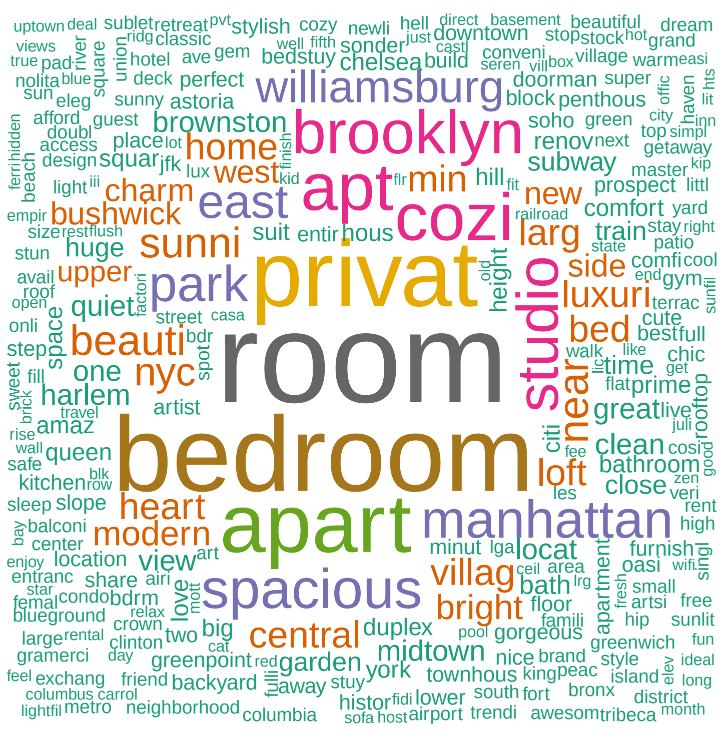


Figure 16: Wordcloud for listings

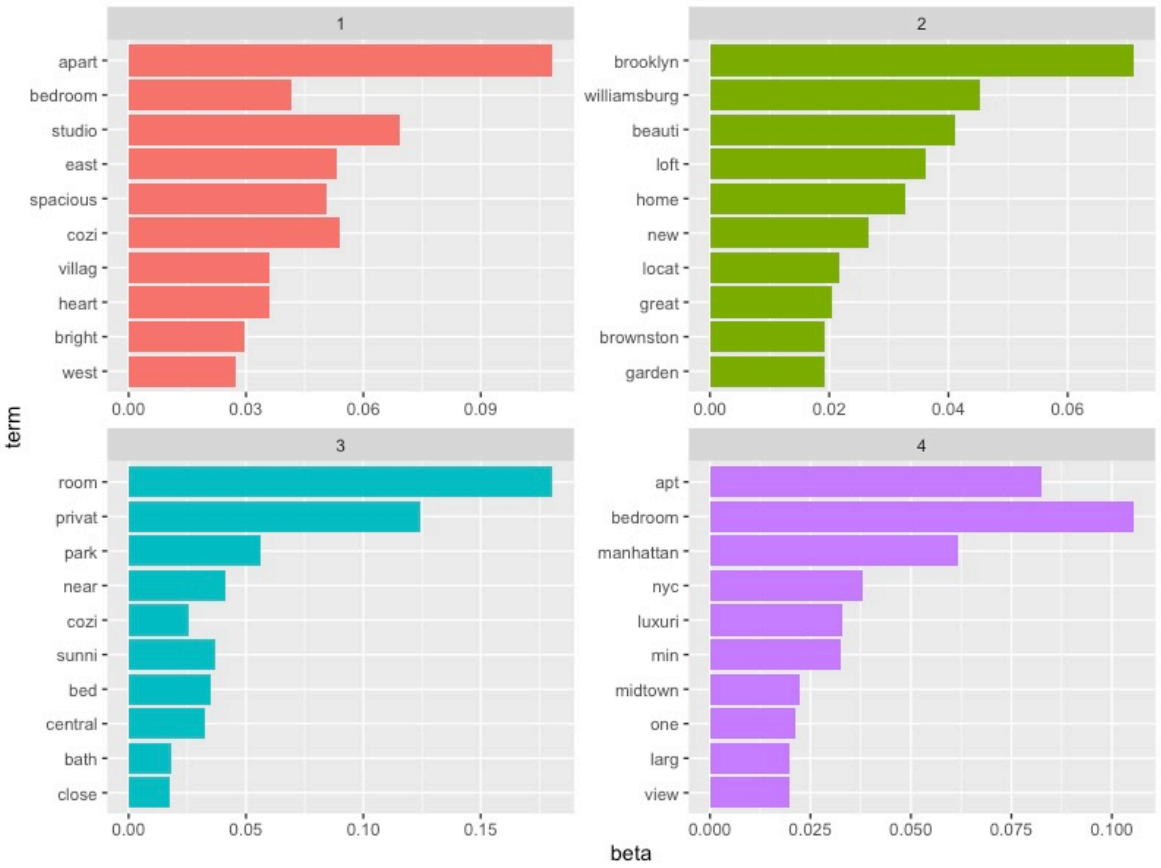


Figure 17: LDA: Top 10 words in each topic

References

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Buuren, S van, and Karin Groothuis-Oudshoorn. 2010. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*. University of California, Los Angeles, 1–68.

Lee, Duncan. 2013. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors." *Journal of Statistical Software* 55 (13). American Statistical Association: 1–24.

Porter, Martin F. 2001. "Snowball: A Language for Stemming Algorithms."