Patterns of Airbnb Listings in NYC

Frances Hung, Yunran Chen, Keru Wu

Abstract

Airbnb home rental listings vary in price and popularity; one question perninent to hosts is which listings are most successful. We explore the relationships between certain rental characteristics (including neighbourhood location) and listing price/popularity in NYC.

1. Introduction

Airbnb is a platform provides more personalized home rentals for travelers compared with hotels. Our data observations consist 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) exploring the most valuable neighborhoods with adjusting the influential factors (3) choose a location and set a price for the listing (4) name the listing.

(Most questions of interest we address in this case study involve finding patterns in price and popularity among neighbourhoods. We find the most influential factors in the price/popularity of a listing, and look for heterogeneity among neighbourhoods and boroughs in terms of traffic, price, and room type. From this and some text analysis, we build an Airbnb listing which would be predicted to be among the most profitable in NYC.)

2. Materials and Methods

2.1 Exploratory Data Analysis

2.2 Data Preprocessing and Missing data manipulation

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.

2.3 Main Model

Since the price and popularity are strongly related to the location of listings and neighboods provides a natural boundary for spatial characteristics of listings, we consider a multilevel conditional autoregressive Bayesian model (CARBayes)(Lee 2013) based on neighbood units as follows:

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

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

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

, 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 neighbourhoods' effect and individual effect. We consider a 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} is known from data with $w_{kl} = 1$ denotes neighbourhood k is adjacent to neighbourhood l, and 0 otherwise; $\rho \sim U(0,1)$ capture the relation between neighbourhood effects. This prior captures the spatial structure among neighbourhoods for each neighborhoods' effect is centered at the weighted sum of the effects from its neighbors.

We consider `log(price)` and `log(1+review_per_month)`(popularity) as response variable and model them separately. We include room type, price, minimum nights required, price/popularity as predictors. Additionally, we incorporate the logarithm distance from a listing to the nearest metro station (from extra data source) and extracted features from LDA model as predictors. To obtain the adjacency matrix for neighbourhoods, we incorporate shape file from NYC Opendata and re-allocate listings to new neighbourhoods.

To carry out text analysis on names of listings, we consider using text mining methods including Porter's stemmer algorithm (Porter 2001), wordcloud, Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), etc. We first preprocess these names by transforming them to lower case and removing non-informative characters (e.g. punctuations, stopwords, whitespace, numbers). Then we use Porter's stemmer algorithm for word normalization, which allows us to extract all the common roots of informative words. Based on the result, we further execute word frequency analysis for different boroughs, and use wordcloud to display frequent words. Moreover, we implement LDA to build up a Bayesian generative model, assigning each word a weight of related topics (e.g. adjectives, locations). Features obtained from LDA will be included in our multilevel CAR regression model.

3. Results

3.1 Exploratory Data Analysis

We obtain an intuitive understanding of the data based on plots of price, popularity and traffic (Figs **??**). Most high-priced listings are located in Manhattan, while some of them also lie in Brooklyn. Similar pattern is discovered for traffic. In addition, EDA plots for room type (Fig **??**) demonstrate that it matters for price but not for popularity. We can also see the heterogeneity of room type across boroughs/neighborhoods. Wordcloud (Fig **??**) implies some high-frequency words in high-priced listings: luxury, manhattan, apartment, etc.

3.2 Main Results

According to model coefficient estimation (Fig 1), our multilevel CAR model on price demonstrates the following patterns. Numbers in brackets are median of corresponding coefficients. For room type, entire room (0) is more expensive than private ones (-0.7) and shared ones (-1.1), with shared room being the cheapest. Manhattan (0.57) stands out to be the most luxurious borough, and Bronx (0) has the lowest price. Availability (0.12) is positively related to price while reviews per month (-0.0) is negatively related. In addition, more strict requirement on minimum nights results in lower price, which aligns with our common sense. And longer distance to metro stations also reduces the price (-0.005). Table 1 shows the WAIC of different models, which allow us to choose the most influential factor for price(room type).

Model on popularity (1) has some similarity but is different as follows. Compared to other four boroughs, Queens borough (0.13) has the highest average reviews. Availability still has a positive effect (0.15) while price (-0.12) leads to a negative influence. Moreover, metro distance is no longer significant for predicting popularity. Table 2 shows the WAIC of different models, which allow us to choose the most influential factor for popularity (availability & night).

Heterogeneity across neighbourhoods is shown in Fig 3 and 4. According to Fig 3, most neighbourhoods in Manhattan have higher average price, and their confidence interval is also narrower than others. Among all neighbourhoods, “Midtown South” in Manhattan turns out to be the most expensive one, while “New Drop-Midland Beach” in Staten Island becomes the one with lowest price. On the other hand, 4 indicates that “East Elmhurst” in Queen is the most popular neighbourhood, which makes sense since LaGuardia Airport is located here, and “Co-op City” is the most unpopular one. If we condsider top 20 neighbourhoods for price and popularity seperately, they have only one intersection at “Yorkville” in Manhattan.

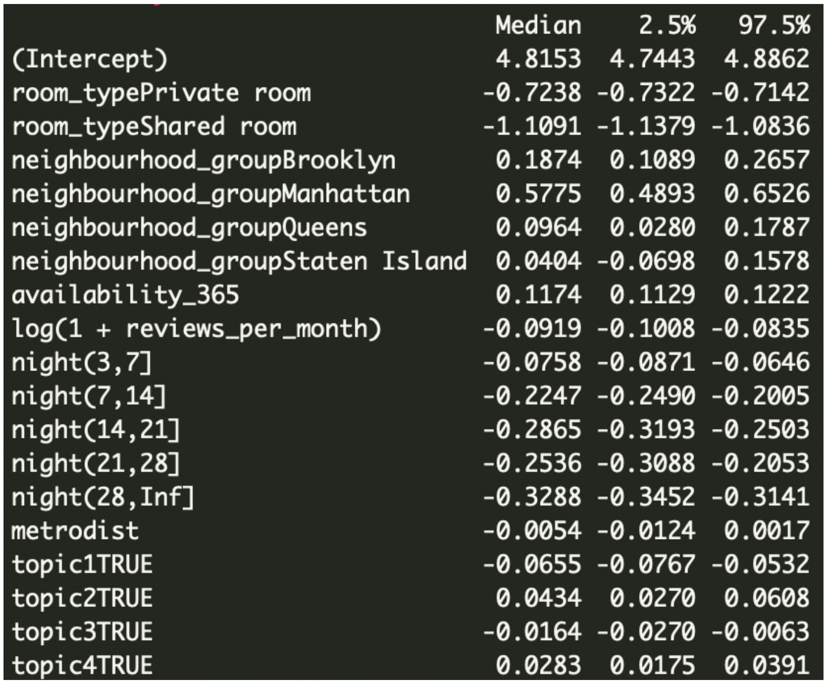
Our text analysis (Fig 5, 6) indicates some critical words: luxury, manhattan, beautiful (Note that we use stemming algorithm so we get stem of words rather than words themselves). We further carry out LDA to find latent topics in listing names. We choose 4 topics which is not too complicated and has a reasonable result (Fig 7). The 4 topics can be categorized as adjectives, locations, Brooklyn related and Manhattan related. If we further add these 4 topics into our model (4 indicators), we conclude that Brooklyn and Manhattan has a positive significant coefficient, while the other two is significantly negative.

3.3 Sensitivity Analysis

4. Answers to Questions

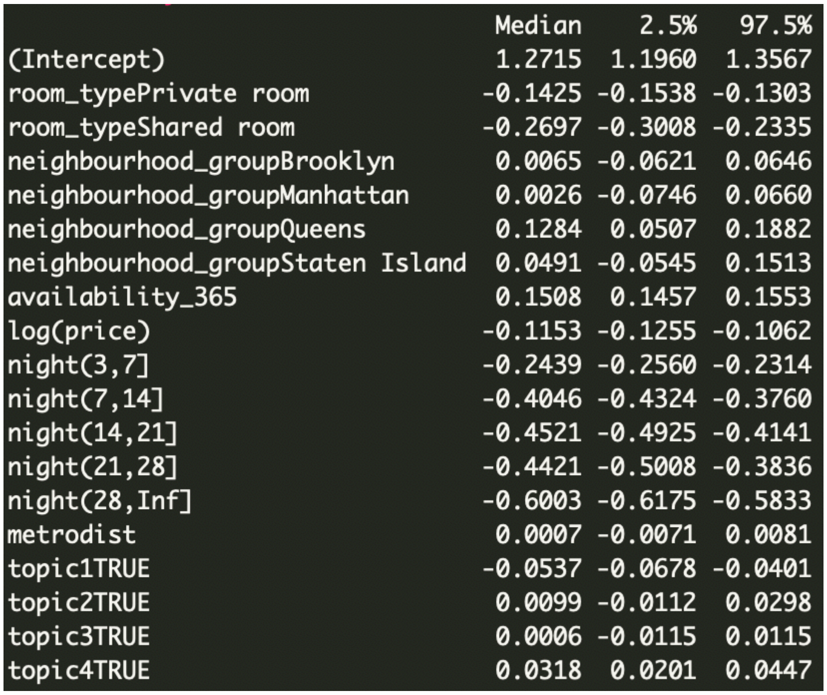
5. Discussion

Appendix



```
                                  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 1: 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 2: CAR Model on popularity - Model Summary

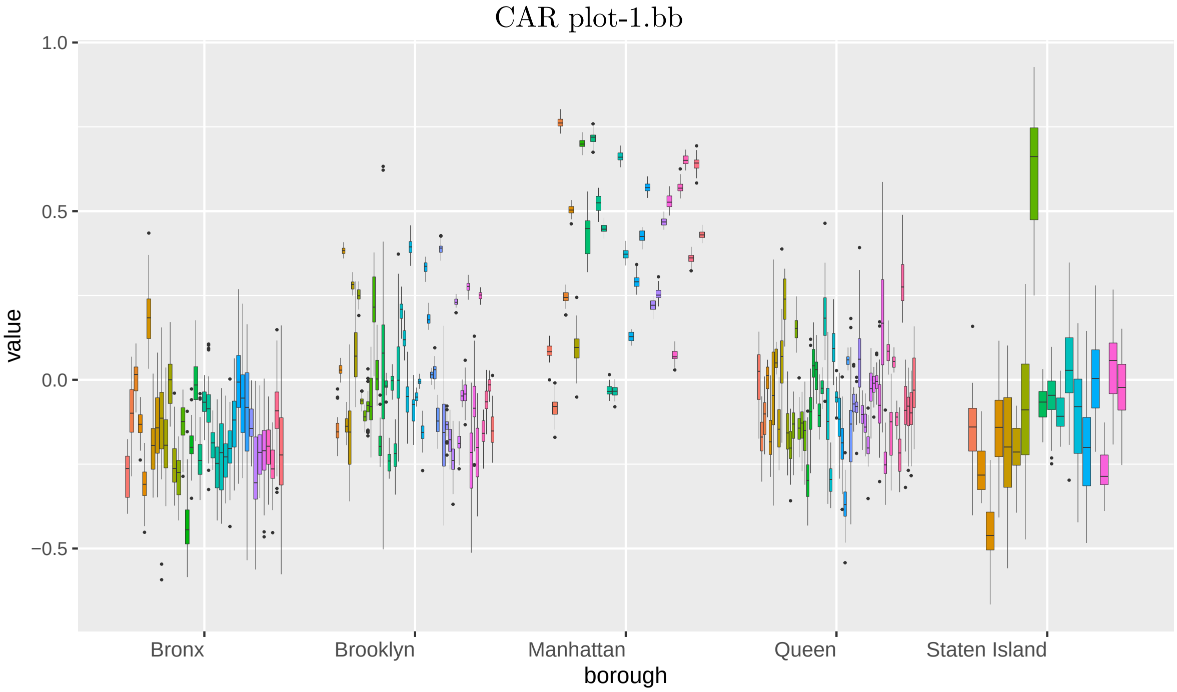


Figure 3: CAR Model on price - Neighbourhoods

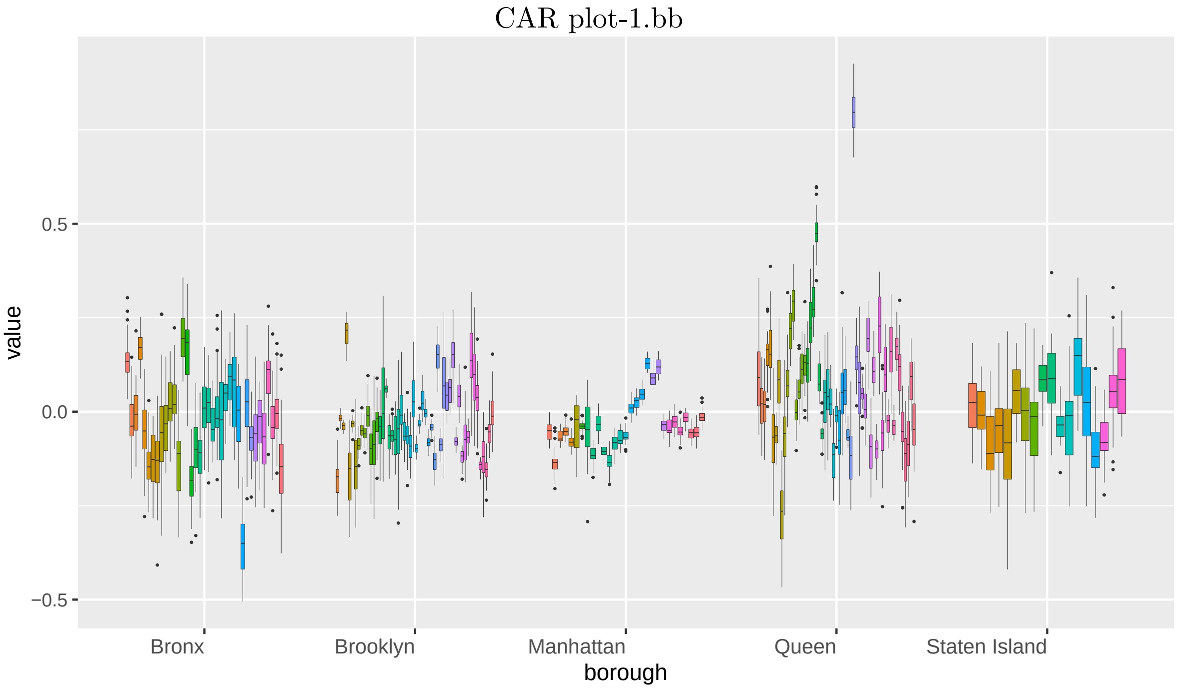


Figure 4: CAR Model on popularity - Neighbourhoods

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

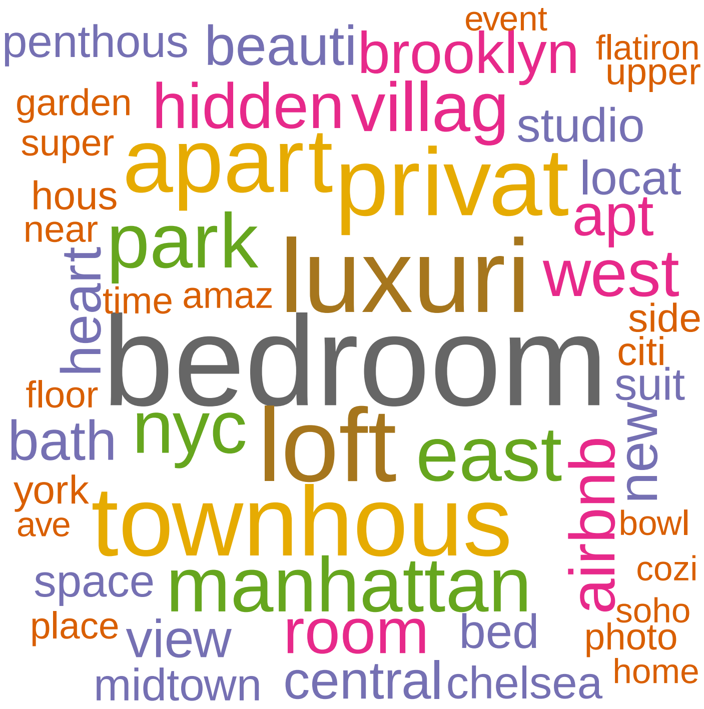


Figure 5: Wordcloud for listings with price > 2000

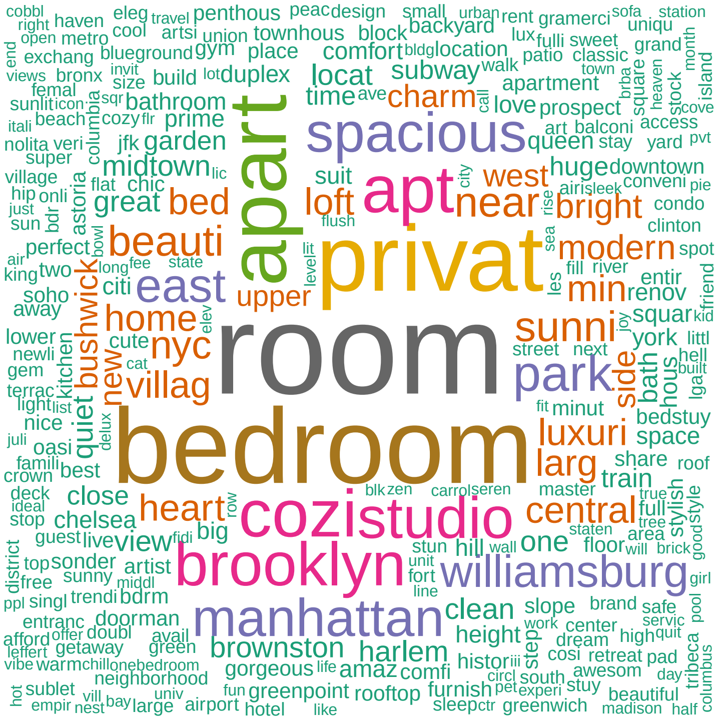


Figure 6: Wordcloud for listings

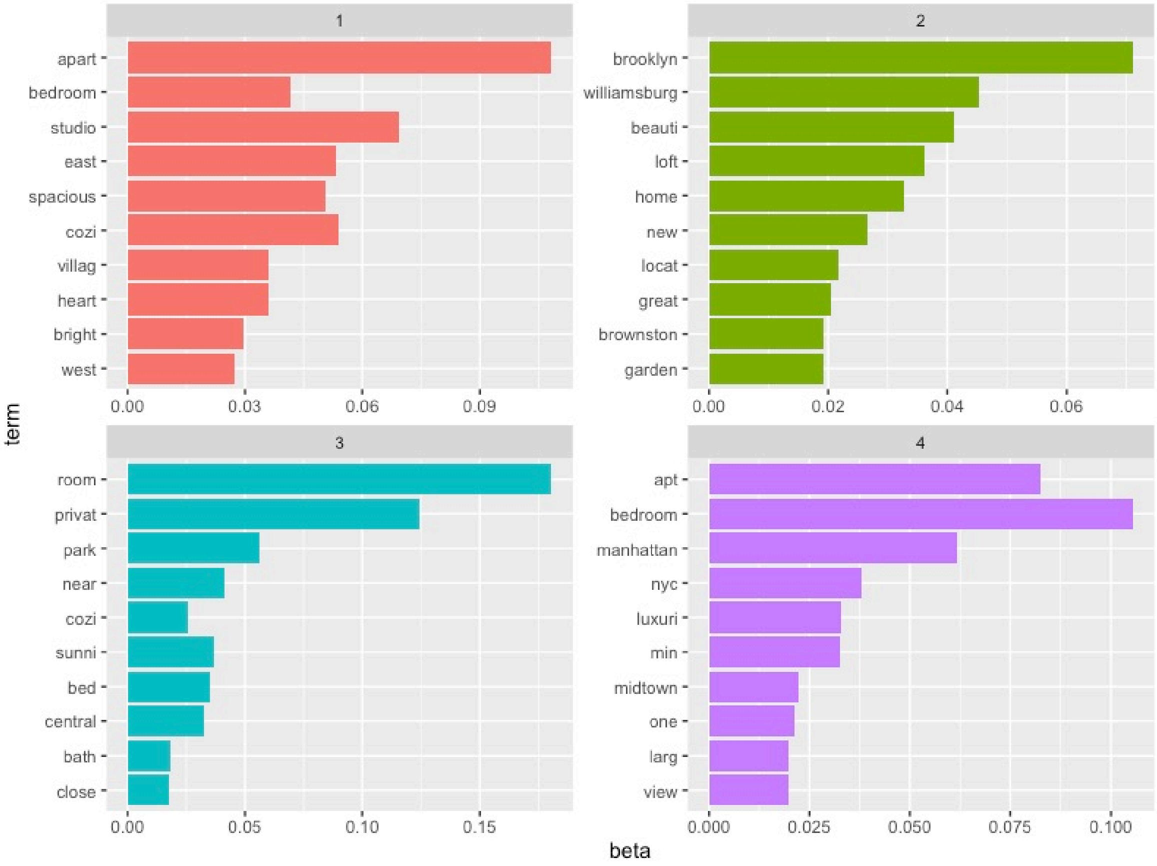


Figure 7: LDA: Top 10 words in each topic

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."