Patterns among Airbnb Listings in NYC

Frances Hung, Yunran Chen, Keru Wu

Abstract

1. Introduction

2. Materials and Methods

2.1 Exploratory Data Analysis

2.2 Missing data manipulation

Cite MICE: (Buuren and Groothuis-Oudshoorn 2010)

2.3 Multilevel Conditional Autoregressive (CAR) Model

Cite CARBayes: (Lee 2013)

2.4 Text Analysis

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. (Polularity?) 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. In addition, wordcloud (Fig **??**) suggests some high-frequency words in high-priced listings: luxury, manhattan, apartment, etc.

3.2 Main Results

According to model coefficient estimation (Fig **??**), 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).

Model on popularity (**??**) 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 negetive influence. Moreover, metro distance is no longer significant for predicting popularity.

Heterogeneity across neighbourhoods is shown in Fig 1 and 2. According to Fig 1, 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 luxurious one, while "New Dorp-Midland Beach" in Staten Island becomes the one with lowest price. On the other hand, 2 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 3, 4) 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 5). 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 Answers to Questions

3.4 Sensitivity Analysis

4. Discussion

```
##                                  Median    2.5%   97.5% n.sample % accept
## (Intercept)                      4.8168  4.7417  4.8823      333    100.0
## room_typePrivate room           -0.7232 -0.7316 -0.7142      333    100.0
## room_typeShared room            -1.1098 -1.1340 -1.0823      333    100.0
## neighbourhood_groupBrooklyn      0.1787  0.1119  0.2556      333    100.0
## neighbourhood_groupManhattan     0.5895  0.4886  0.6758      333    100.0
## neighbourhood_groupQueens        0.0978  0.0440  0.1749      333    100.0
## neighbourhood_groupStaten Island 0.0364 -0.0787  0.1611      333    100.0
## availability_365                 0.1175  0.1130  0.1222      333    100.0
## log(1 + reviews_per_month)      -0.0921 -0.0993 -0.0825      333    100.0
## night(3,7]                      -0.0749 -0.0856 -0.0647      333    100.0
## night(7,14]                     -0.2256 -0.2479 -0.1992      333    100.0
## night(14,21]                    -0.2837 -0.3169 -0.2529      333    100.0
## night(21,28]                    -0.2507 -0.2949 -0.2001      333    100.0
## night(28,Inf]                   -0.3277 -0.3449 -0.3139      333    100.0
## metrodist                       -0.0053 -0.0108  0.0003      333    100.0
```

```
## topic1TRUE                        -0.0657 -0.0763 -0.0520      333    100.0
## topic2TRUE                         0.0434  0.0256  0.0622      333    100.0
## topic3TRUE                        -0.0168 -0.0274 -0.0073      333    100.0
## topic4TRUE                         0.0280  0.0161  0.0383      333    100.0
## nu2                                0.2158  0.2132  0.2188      333    100.0
## tau2                               0.0402  0.0273  0.0637      333    100.0
## rho                                0.0971  0.0054  0.3133      333     63.5
##                                  n.effective Geweke.diag
## (Intercept)                            110.6         0.3
## room_typePrivate room                  417.2        -0.3
## room_typeShared room                   333.0         0.6
## neighbourhood_groupBrooklyn             66.6        -1.3
## neighbourhood_groupManhattan            28.0         2.2
## neighbourhood_groupQueens               88.2        -1.0
## neighbourhood_groupStaten Island       131.9        -0.7
## availability_365                       333.0         1.1
## log(1 + reviews_per_month)             333.0        -0.5
## night(3,7]                             333.0        -0.9
## night(7,14]                            333.0         1.4
## night(14,21]                           333.0         2.5
## night(21,28]                           298.8        -0.9
## night(28,Inf]                          925.3        -0.4
## metrodist                              333.0        -0.3
## topic1TRUE                             333.0        -0.2
## topic2TRUE                             333.0         1.7
## topic3TRUE                             333.0         0.3
## topic4TRUE                             333.0         1.0
## nu2                                    717.0        -0.1
## tau2                                    59.9         0.4
## rho                                     46.2         0.2

##                                   Median    2.5%   97.5% n.sample % accept
## (Intercept)                       1.2632  1.1911  1.3303       33    100.0
## room_typePrivate room            -0.1424 -0.1556 -0.1312       33    100.0
## room_typeShared room             -0.2631 -0.3011 -0.2448       33    100.0
## neighbourhood_groupBrooklyn       0.0334 -0.0383  0.0808       33    100.0
## neighbourhood_groupManhattan     -0.0184 -0.0652  0.0158       33    100.0
## neighbourhood_groupQueens         0.1353  0.0945  0.1941       33    100.0
## neighbourhood_groupStaten Island  0.0680 -0.0346  0.1383       33    100.0
## availability_365                  0.1510  0.1471  0.1552       33    100.0
## log(price)                       -0.1148 -0.1278 -0.1052       33    100.0
## night(3,7]                       -0.2423 -0.2550 -0.2315       33    100.0
## night(7,14]                      -0.4095 -0.4269 -0.3794       33    100.0
## night(14,21]                     -0.4465 -0.4909 -0.4177       33    100.0
## night(21,28]                     -0.4353 -0.4895 -0.4020       33    100.0
## night(28,Inf]                    -0.6021 -0.6202 -0.5884       33    100.0
## metrodist                         0.0011 -0.0059  0.0086       33    100.0
## topic1TRUE                       -0.0555 -0.0660 -0.0410       33    100.0
## topic2TRUE                        0.0103 -0.0081  0.0272       33    100.0
## topic3TRUE                        0.0003 -0.0096  0.0099       33    100.0
## topic4TRUE                        0.0317  0.0210  0.0452       33    100.0
## nu2                               0.2695  0.2671  0.2716       33    100.0
## tau2                              0.0193  0.0136  0.0234       33    100.0
## rho                               0.0330  0.0041  0.0888       33     79.5
##                                  n.effective Geweke.diag
## (Intercept)                             33.0        -1.8
```

```
## room_typePrivate room                33.0         1.0
## room_typeShared room                 33.0         0.8
## neighbourhood_groupBrooklyn           3.6        -2.2
## neighbourhood_groupManhattan         13.3        -0.6
## neighbourhood_groupQueens             4.1        -0.9
## neighbourhood_groupStaten Island     33.0         1.3
## availability_365                     33.0        -1.7
## log(price)                           33.0         3.3
## night(3,7]                           33.0         0.5
## night(7,14]                          33.0        -1.3
## night(14,21]                         77.8        -0.7
## night(21,28]                         33.0         2.3
## night(28,Inf]                        33.0         2.2
## metrodist                            39.2         0.8
## topic1TRUE                           33.0        -1.1
## topic2TRUE                           99.2         1.1
## topic3TRUE                           33.0         1.0
## topic4TRUE                           33.0         0.2
## nu2                                  17.7         3.6
## tau2                                 33.0         0.4
## rho                                   7.0         1.3
```

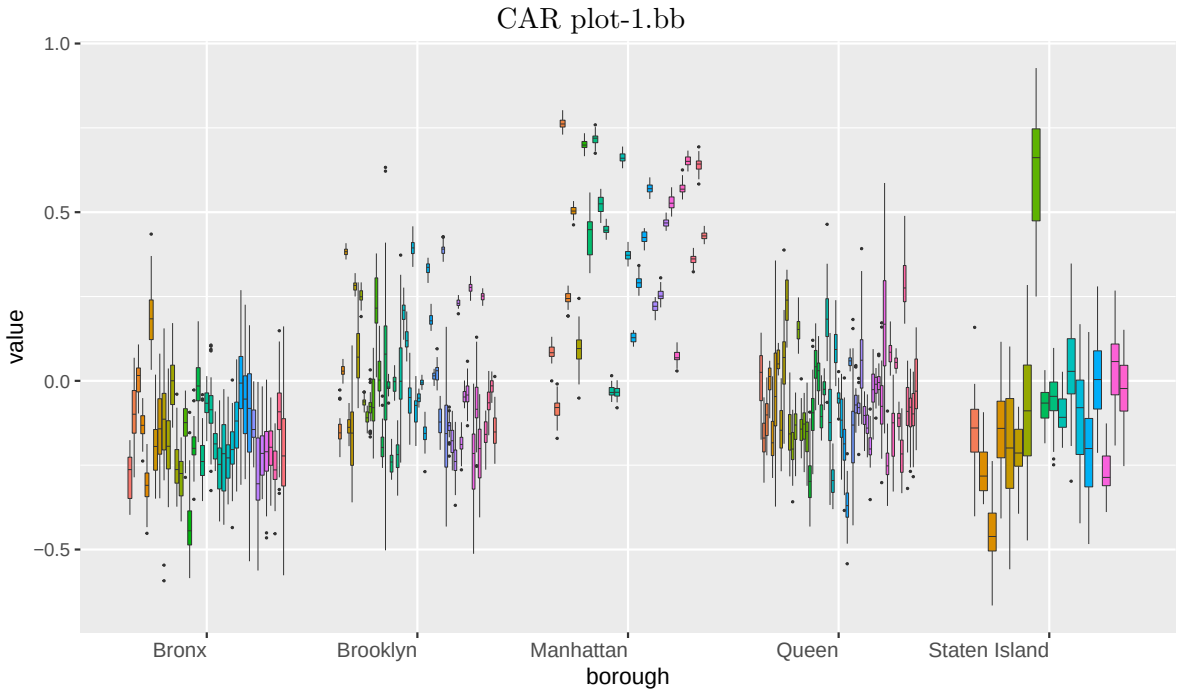

Figure 1: CAR Model on price - Neighbourhoods

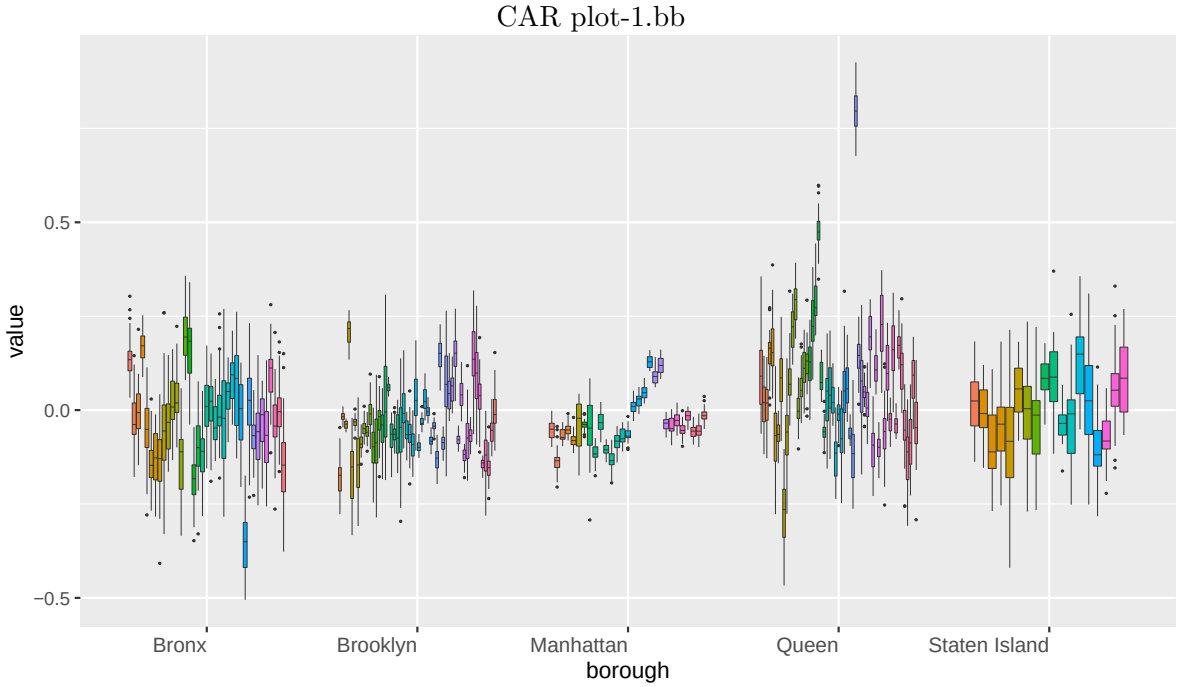

Figure 2: CAR Model on popularity - Neighbourhoods

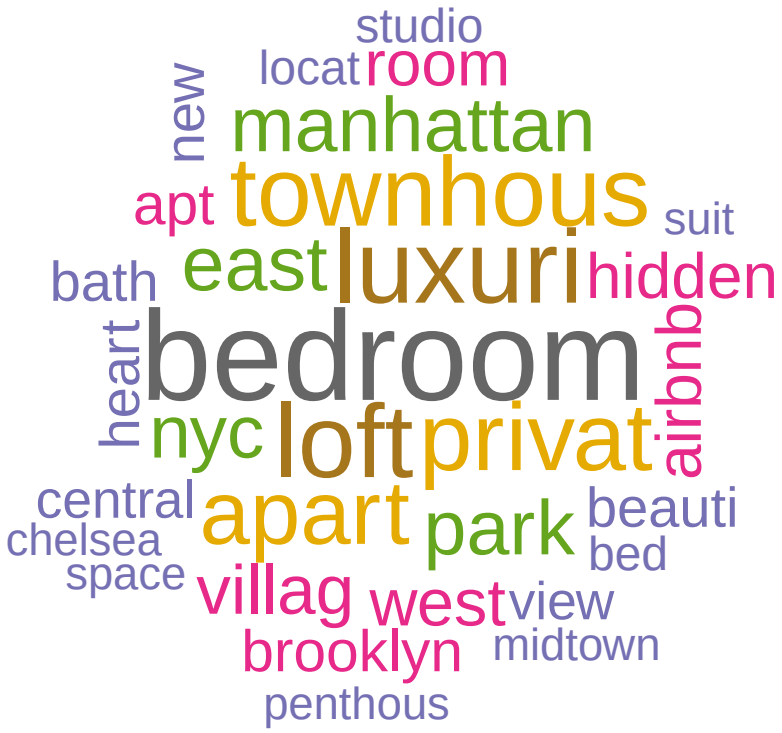

Figure 3: Wordcloud for listings with price > 2000

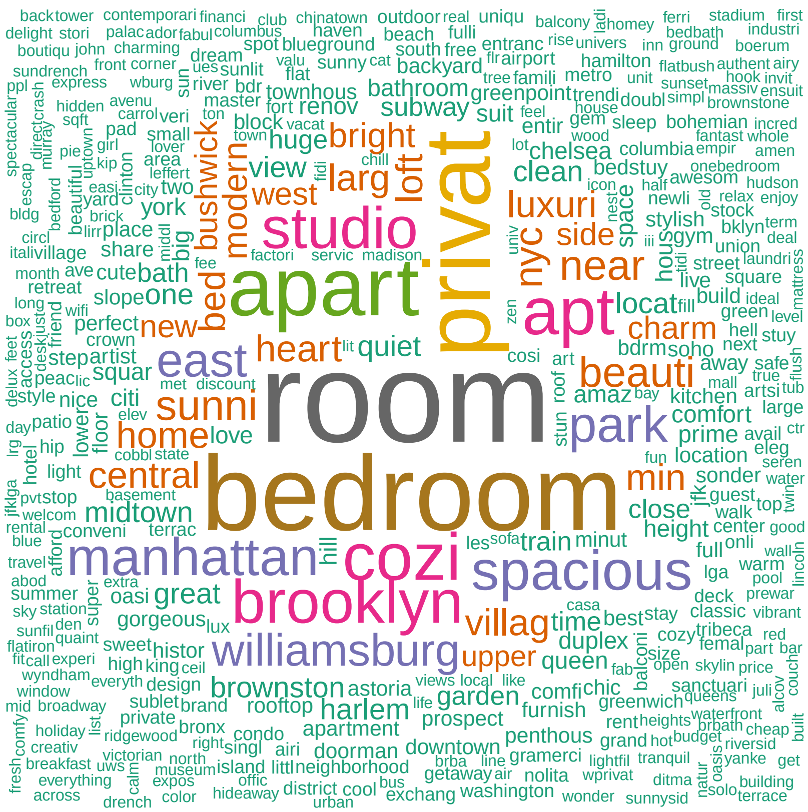

Figure 4: Wordcloud for listings

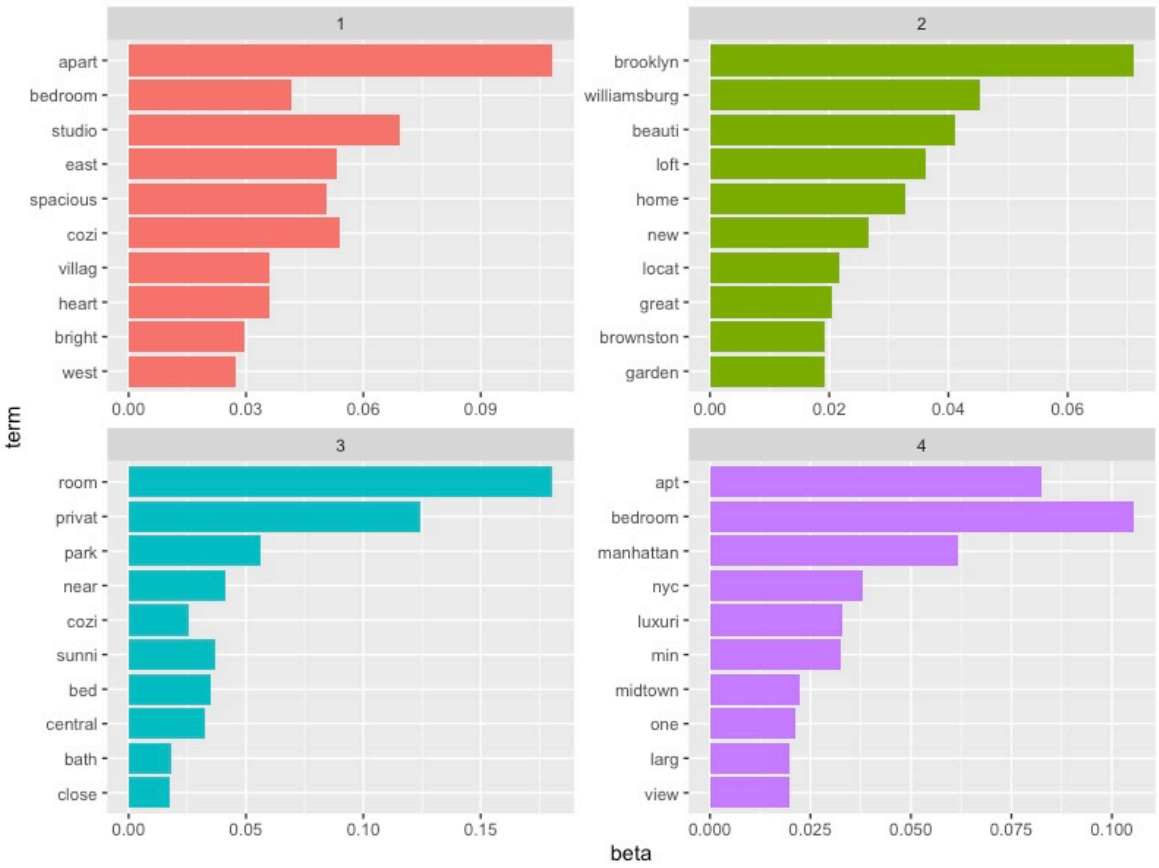

Figure 5: LDA: Top 10 words in each topic

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."