Case Study 2-Group 4

Frances Hung, Yunran Chen, Keru Wu

Department of Statistical Science, Duke University

01/30/2019

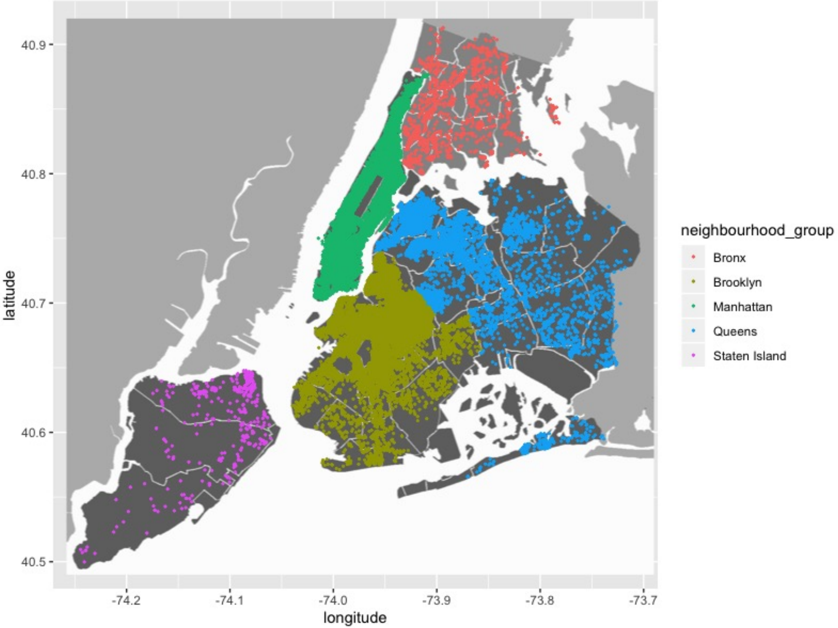
Introduction

- Data: 2019 Airbnb listings in NYC, 48895 observations.
- Goal: Identify discernible and interesting patterns among the listings in NYC.

Data Preprocessing

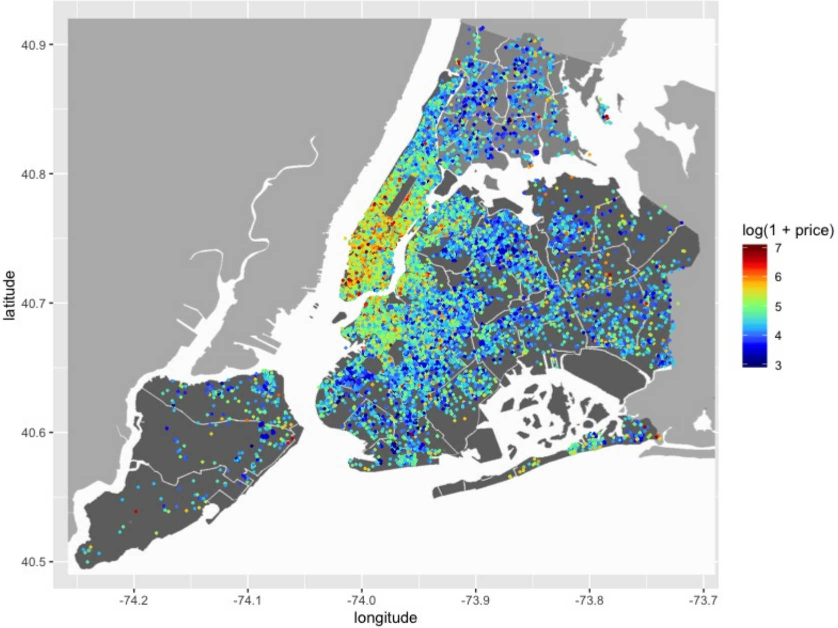
- Delete id, host_name and last_review.
- Delete 11 listings with price 0.
- Missing data: 10052 in reviews_per_month, impute with 0.
- Categorize minimum_nights to 5 groups: 1-2, 3-6, 7-13, 14-20, 21-27, 28+ days.
- Transformation: log(price), log(1+reviews_per_month).

EDA-borough



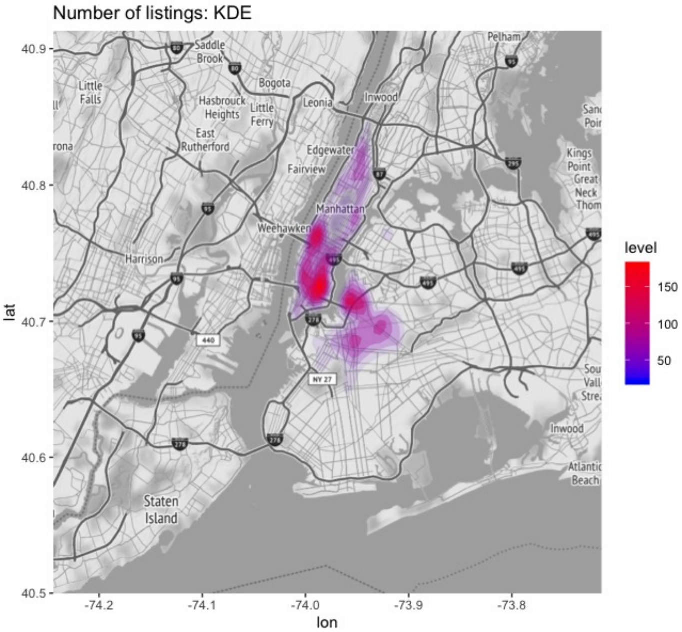
40.9
40.8
40.7
40.6
40.5
latitude
-74.2
-74.1
-74.0
-73.9
-73.8
-73.7
longitude
neighbourhood_group
Bronx
Brooklyn
Manhattan
Queens
Staten Island

EDA-price



40.9
40.8
40.7
40.6
40.5
latitude
-74.2
-74.1
-74.0
-73.9
-73.8
-73.7
longitude
log(1 + price)
7
6
5
4
3

EDA-Traffic



Number of listings: KDE
lat
lon
40.9
40.8
40.7
40.6
40.5
-74.2
-74.1
-74.0
-73.9
-73.8
level
150
100
50
Saddle Brook
Little Falls
Bogota
Hasbrouck Heights
Little Ferry
Leonia
Inwood
Pelham
East Rutherford
Edgewater
Fairview
Kings Point
Great Neck
Manhattan
Weehawken
Harrison
Staten Island
Atlantic Beach

Model

- Multilevel Conditional Autoregressive (CAR) Model

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

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

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

- Priors

$$\beta \sim N(\mu_\beta, \Sigma_\beta)$$

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

- w_{kl} denotes whether neighborhood k and l are adjacent.
- ρ denotes spatial dependence.

Model

- Priors (Cont'd)

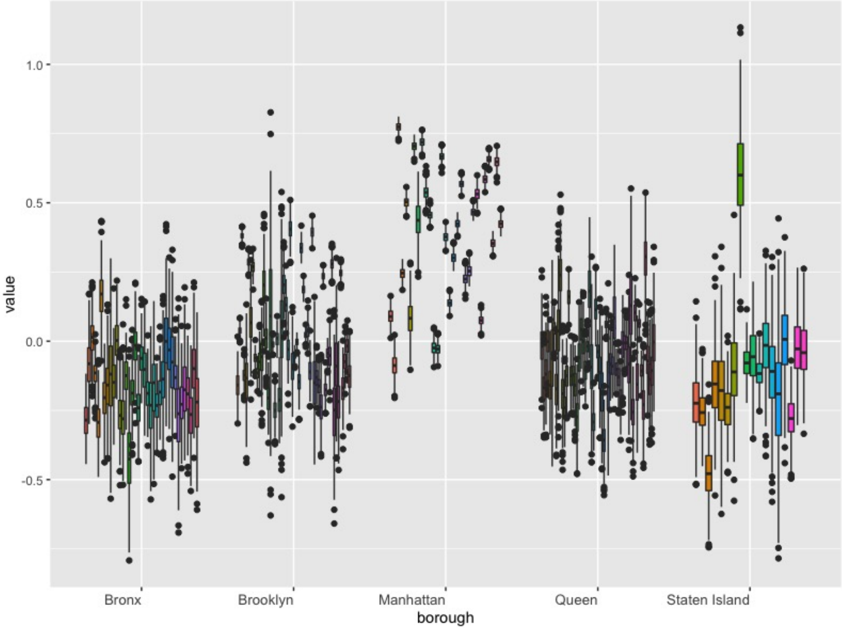
$$
\begin{aligned}
&\zeta_{kj} \sim N(0, \sigma^2) \\
&\tau^2, \sigma^2 \sim \text{Inv-Gamma}(a, b) \\
&\rho \sim \text{Uniform}(0,1)
\end{aligned}
$$

- x_{kj} include room_type, neighbourhood_group, availability_365, log(1+reviews_per_month), minimum_nights, etc.
- $\psi_{kj} = \phi_k + \zeta_{kj}$ includes both spatial information and individual random effect.

Further process the data

- No data for exactly 217 neighbourhoods.
- Relocate neighbourhoods according to formal NYC shapefile data (195 neighbourhoods).
- 191 neighbourhoods have airbnb listings.
- Obtain adjacency matrix $W = (w_{kl})$

Neighbourhood Effect on log(price)



Model Summary

	Median	2.5%	97.5%
(Intercept)	4.7689	4.7143	4.8368
room_typePrivate room	-0.7133	-0.7220	-0.7034
room_typeShared room	-1.0578	-1.0882	-1.0289
neighbourhood_groupBrooklyn	0.1521	0.0674	0.2138
neighbourhood_groupManhattan	0.5145	0.4261	0.5928
neighbourhood_groupQueens	0.0962	0.0072	0.1619
neighbourhood_groupStaten Island	0.0452	-0.0763	0.1555
log(1 + reviews_per_month)	-0.0382	-0.0462	-0.0299
minimum_nights2	-0.1510	-0.1707	-0.1334
minimum_nights3	-0.2481	-0.2784	-0.2179
minimum_nights4	-0.2884	-0.3405	-0.2270
minimum_nights5	-0.1791	-0.1957	-0.1634
nu2	0.2280	0.2255	0.2310
tau2	0.0387	0.0275	0.0541
rho	0.1104	0.0172	0.2226

Preliminary conclusion

- Manhattan has highest prices, Bronx the lowest.
- Midtown South in Manhattan has the highest price.
- Entire room > Private room > Shared room
- Higher minimum_nights leads to lower price.

Discussion

- Include last_review: spatial temporal model.
- Nonlinear model: spline regression for x_{kj}.
- More spatial information: longitude & latitude