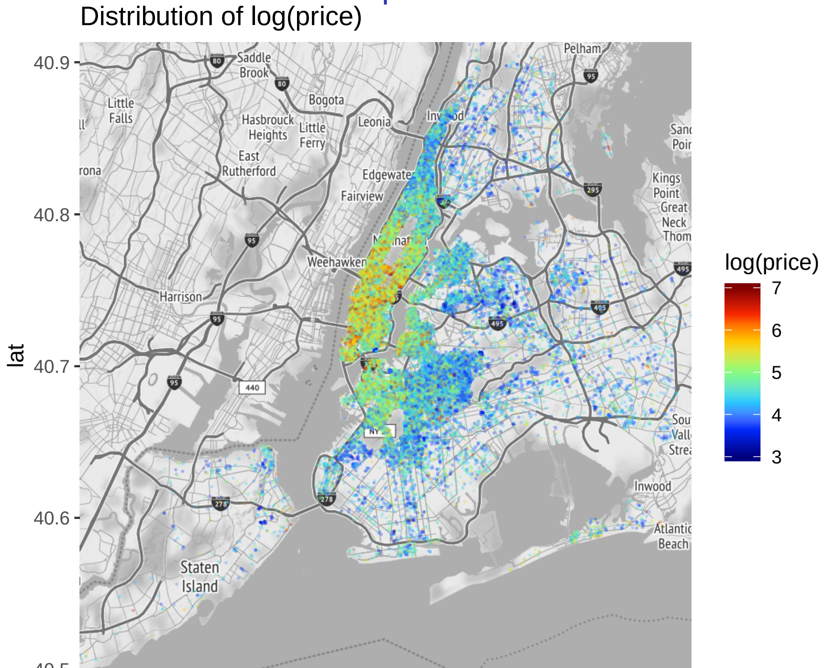
Patterns of Airbnb Listings in NYC

Frances Hung, Yunran Chen, Keru Wu

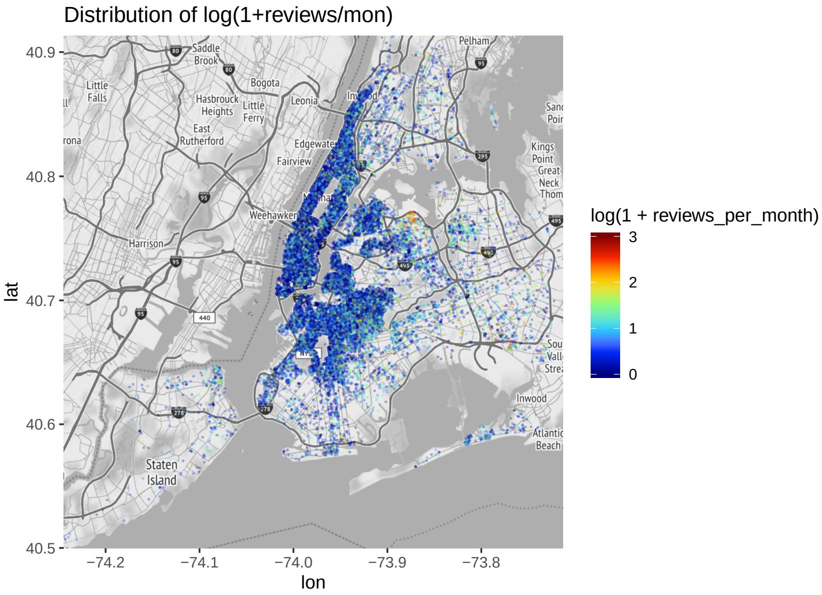
Introduction

- Data: 2019 Airbnb listings in NYC, 48895 observations.
- Goal: Explore Patterns of Listings
  - Care about price and popularity
  - What are the influential factors? quantify influence?
  - Find the most valuable neighborhoods based on price/popularity balance
  - Set the price of the listing
  - Name the listing
- Model:
  - CARBayes for `log(price)` and `log(1+reviews_per_month)` respectively.
  - LDA for text analysis

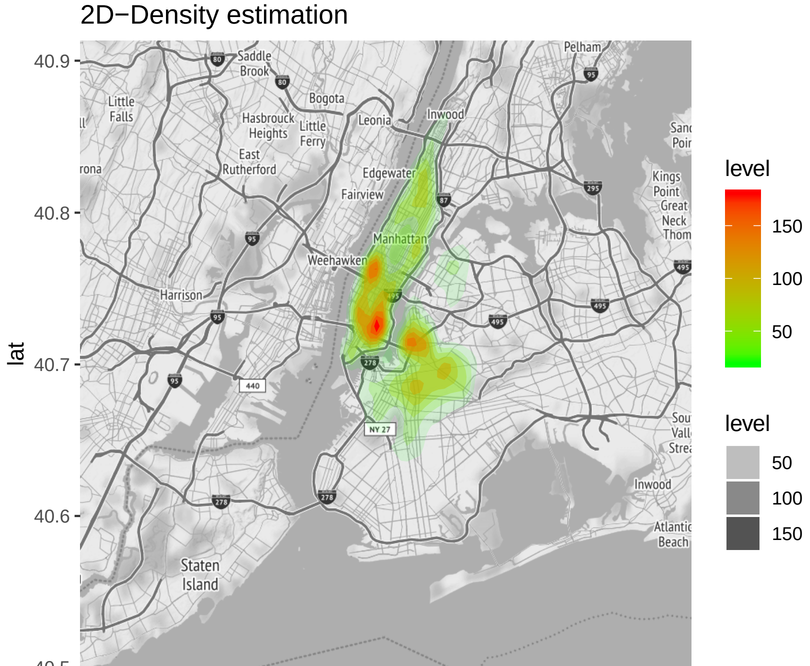
EDA: Location matters for price



EDA: Location matters for popularity

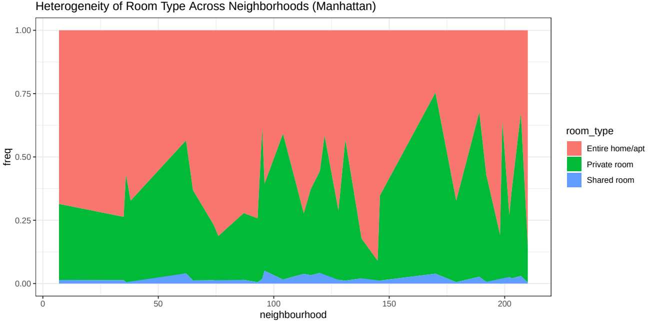


EDA: Location matters for traffic



EDA: Potential effects

- Neighborhoods/boroughs: spatial effect exist
- Room type
  - Room type matters for price but not for popularity
  - Heterogeneity of room type exists across boroughs/neighborhoods
    - Pearson's Chi-squared test (p-value:<2.2e-16)



- Minimum Night
  - nonlinear effect on price/popularity

Data Preprocessing

- Delete: `id`, `host_name` and `last_review`; 11 listings with price 0.
- Impute: impute 0's for `reviews_per_month` (10052 records).
- Categorize: `minimum_nights` to 5 groups by weeks.
- Transformation: `log(price)`, `log(1+reviews_per_month)`.
- Incorporate new dataset:
  - shape file for neighbourhoods (NYC Opendata)
  - locations for metro stations
- Text cleaning:
  - Remove punctuations, stopwords, etc.
  - Word nomalization (Porter's stemmer algorithm)

Model: CARBayes

- Interested in neighbourhood-based patterns
- Multilevel Conditional Autoregressive (CAR) Model

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

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

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

- 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: CARBayes

- Priors (Cont'd)

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

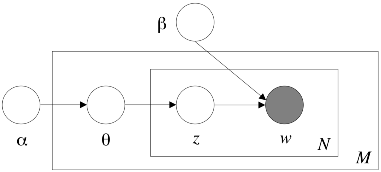
- x_{kj} include room_type, neighbourhood_group, availability_365, log(1+reviews_per_month), minimum_nights.
- $\psi_{kj} = \phi_k + \zeta_{kj}$ includes both spatial information and individual random effect.

Text Analysis: 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. $N \sim \text{Poisson}(\xi)$
  2. $\theta \sim \text{Dir}(\alpha)$
  3. For word w_n $(n = 1, ..., N)$
     (a) choose a topic $z_n|\theta \sim \text{Multinomial}(\theta)$
     (b) choose a word $w_n|z_n, \beta \sim \text{Multinomial}(\beta_{z_n})$



LDA results

- 4 topics: Adjectives, Locations, Brooklyn related, Manhattan related.

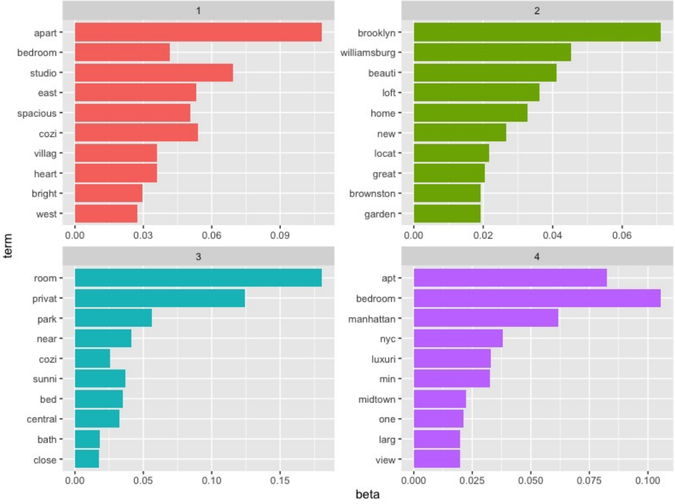


Figure 1: LDA results

Model Summary for log(price)

	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

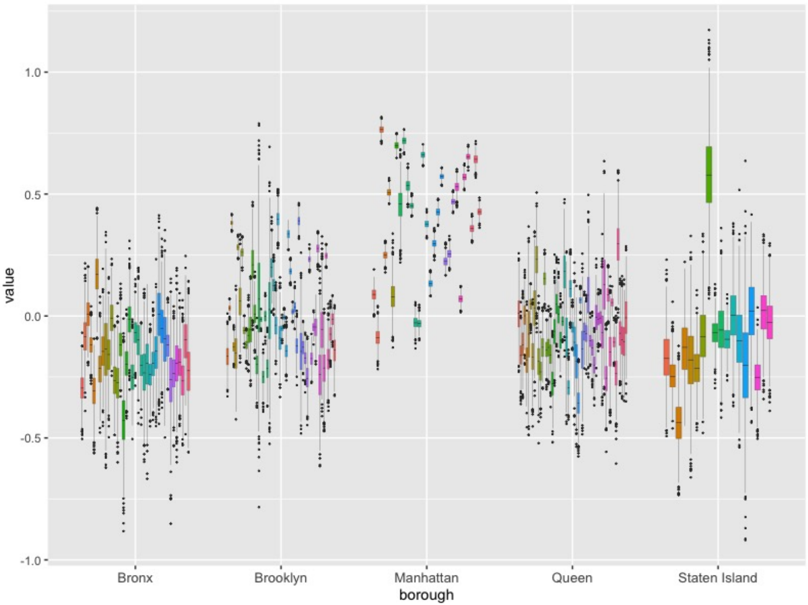
Figure 2: Summary for Model on price

Most influencial factors for `log(price)`

Model WAIC with all variables and without one variable:

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

Neighbourhood Effect on `log(price)`



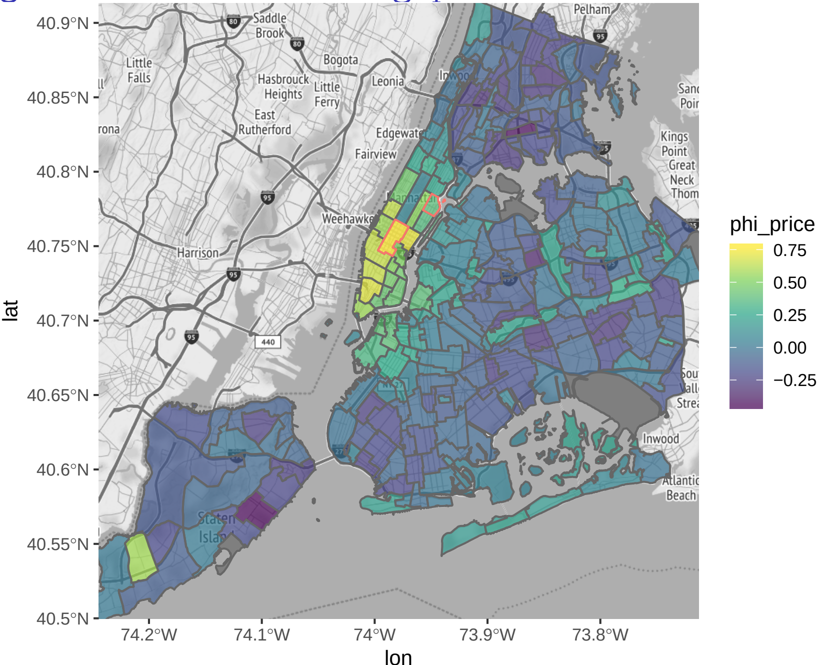
Text Analysis:

- Wordcloud for price < 1000 (left) and all listings(right)

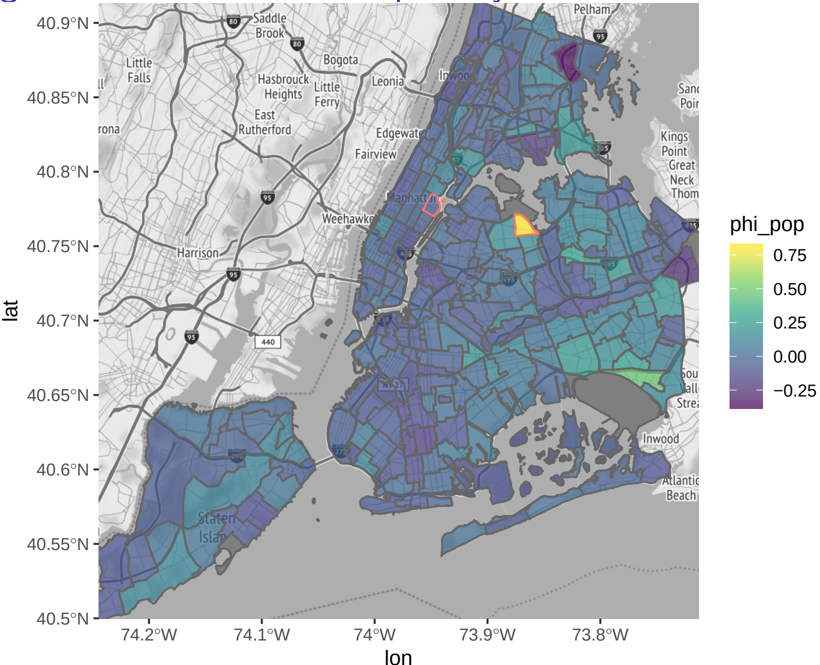
Neighburhood-Specific Conclusions

- Manhattan has the highest prices, Bronx the lowest.
- Midtown South (Manhattan) = most expensive, New Drop-Midland Beach (Staten Island) = cheapest.
- East Elmhurst (Queens) = most popular (LaGuardia Airport), Co-op City (Bronx) = least popular.
- East Village (Manhattan) = heavest traffic, park-cemetery-etc-Brooklyn (Queens) = lightest traffic.
- Yorkville = the most lucrative host neighborhood

Neighbourhood Effect on log(price)



Neighbourhood Effect on Popularity



Variable Importance Conclusions

- Price: Entire room $>$ Private room $>$ Shared room
- Higher minimum_nights $\sim$ lower price
- Shorter distance to metro stations, higher availability $\sim$ higher price
- Popularity: metro distance no longer significant

A Model Airbnb: One Example

- Spacious, Charming Loft in Upper East Side
- Location: Yorkville
- Entire Home, $130/night, between 200-300 available days
- Close to metro station

Some Interesting Inferences

- Sonder homes in Manhattan provide some competition!
- Vague descriptors (i.e. “great”) are associated with less popularity
- Missingness in `availability_365` is not MCAR

Future Directions

- availability_365 missing data
- Spatial-temporal model (last_review)
- Nonlinear model for minimum night stay
- Point reference spatial model (longitude and latitude)
- Random effect for host_id.