Best neighborhoods for Airbnb Listings in NYC

Frances Hung, Yunran Chen, Keru Wu

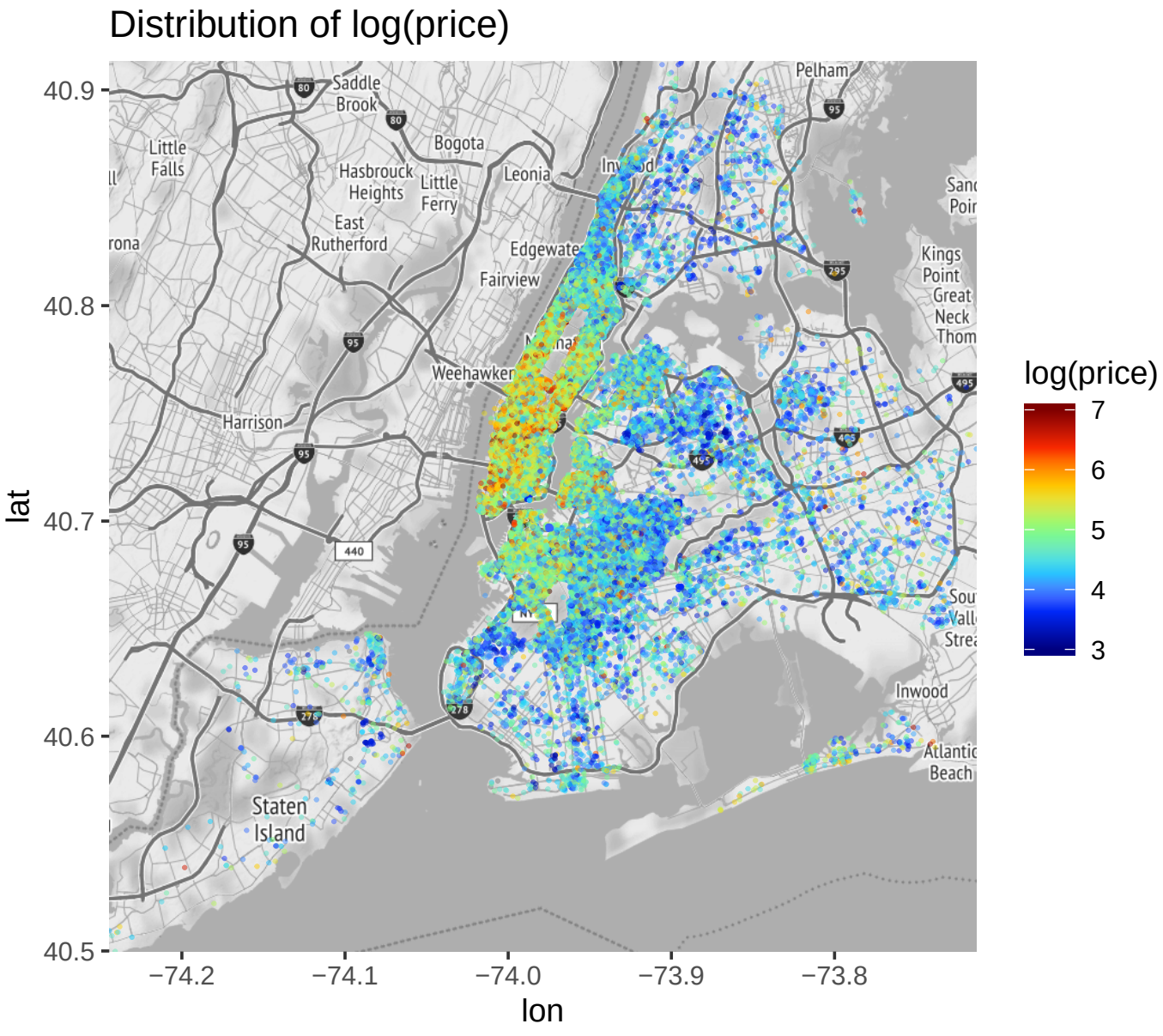
Introduction

- Data: 2019 Airbnb listings in NYC, 48895 observations.
- Goal: Identify patterns among the listings in NYC.
- Patterns for price/popularity: influential factors? quantify influence?
  - Patterns for locations: heterogeneity across boroughs/neighborhoods?
  - Patterns for the type of the listings: vary across neighborhoods?
  - Patterns for listings names: text analysis
- Rank the boroughs/neighborhoods based on price/popularity/traffic
- Post a listing: choice of location and name
- Model:
- CARBayes for `log(price)` and `log(1+reviews_per_month)` respectively.
- LDA for text analysis

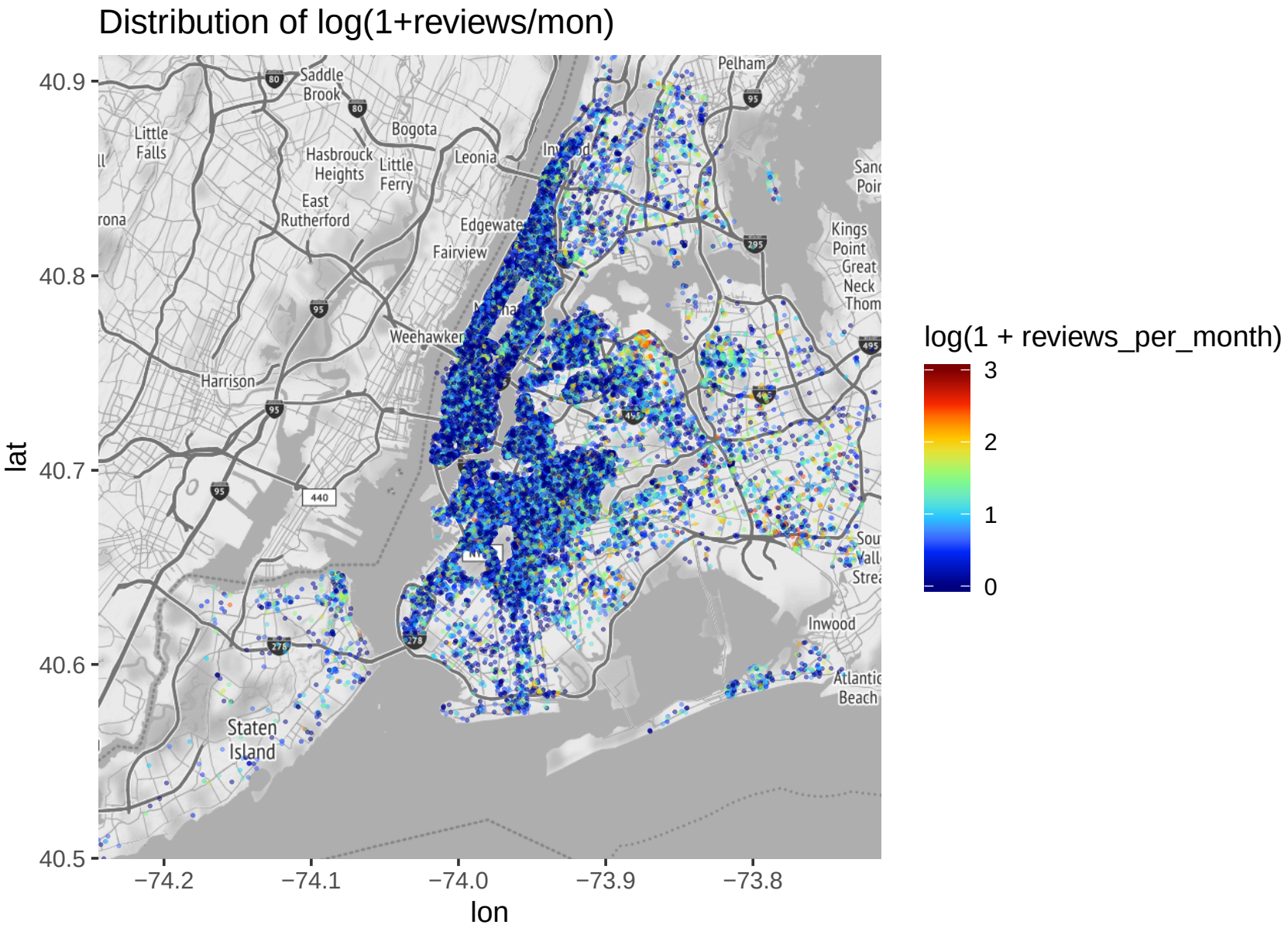
Data Preprocessing

- 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: (1,3], (3,7], (7,14], (14,21], (21,30], (30,∞) days.
- Transformation: log(price), log(1+reviews_per_month).
- Incorporate new dataset: shape file for neighbourhoods, locations for metro stations

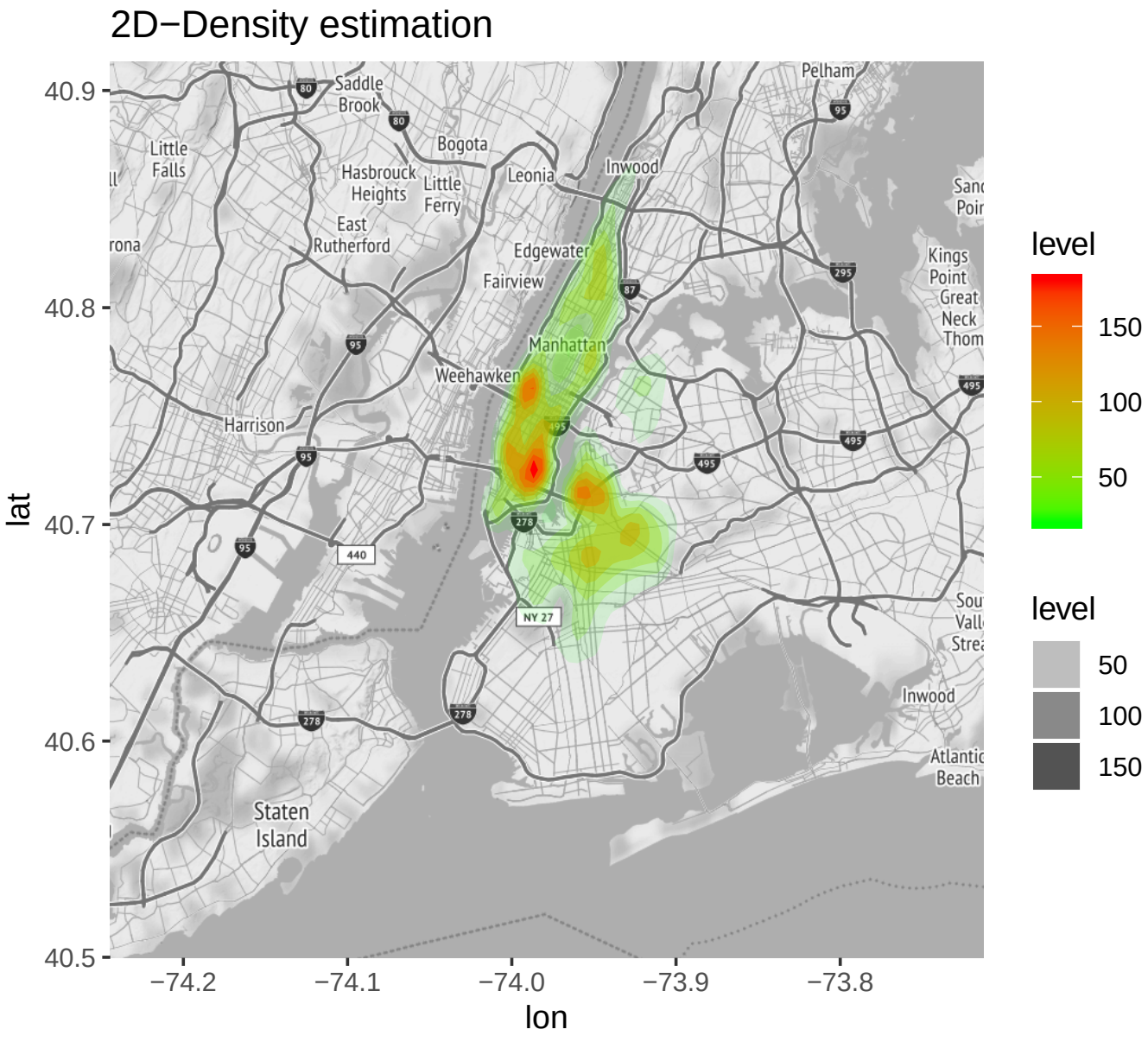
EDA: Location matters for price



EDA: Location matters for popularity

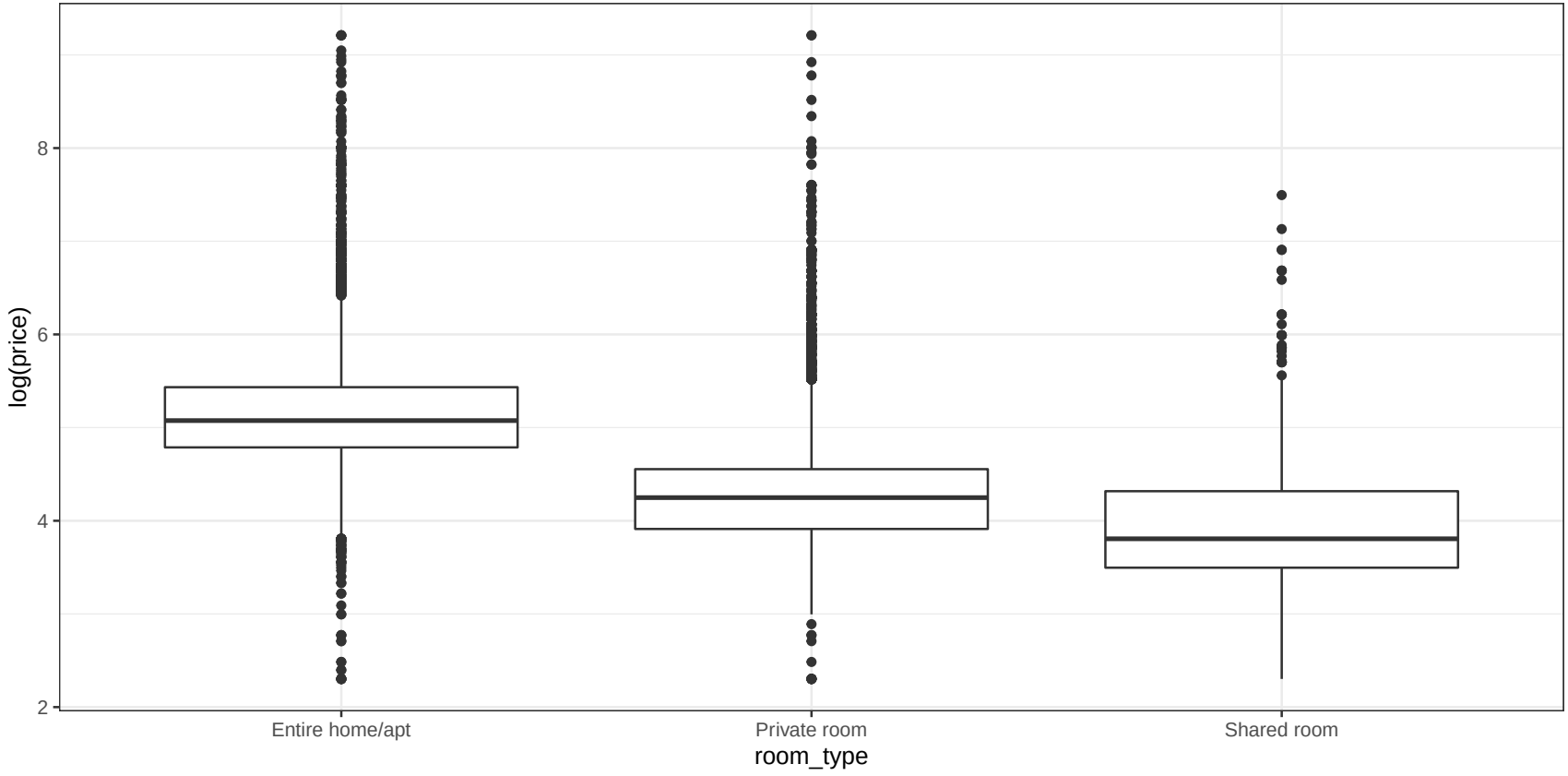


EDA: Location matters for traffic

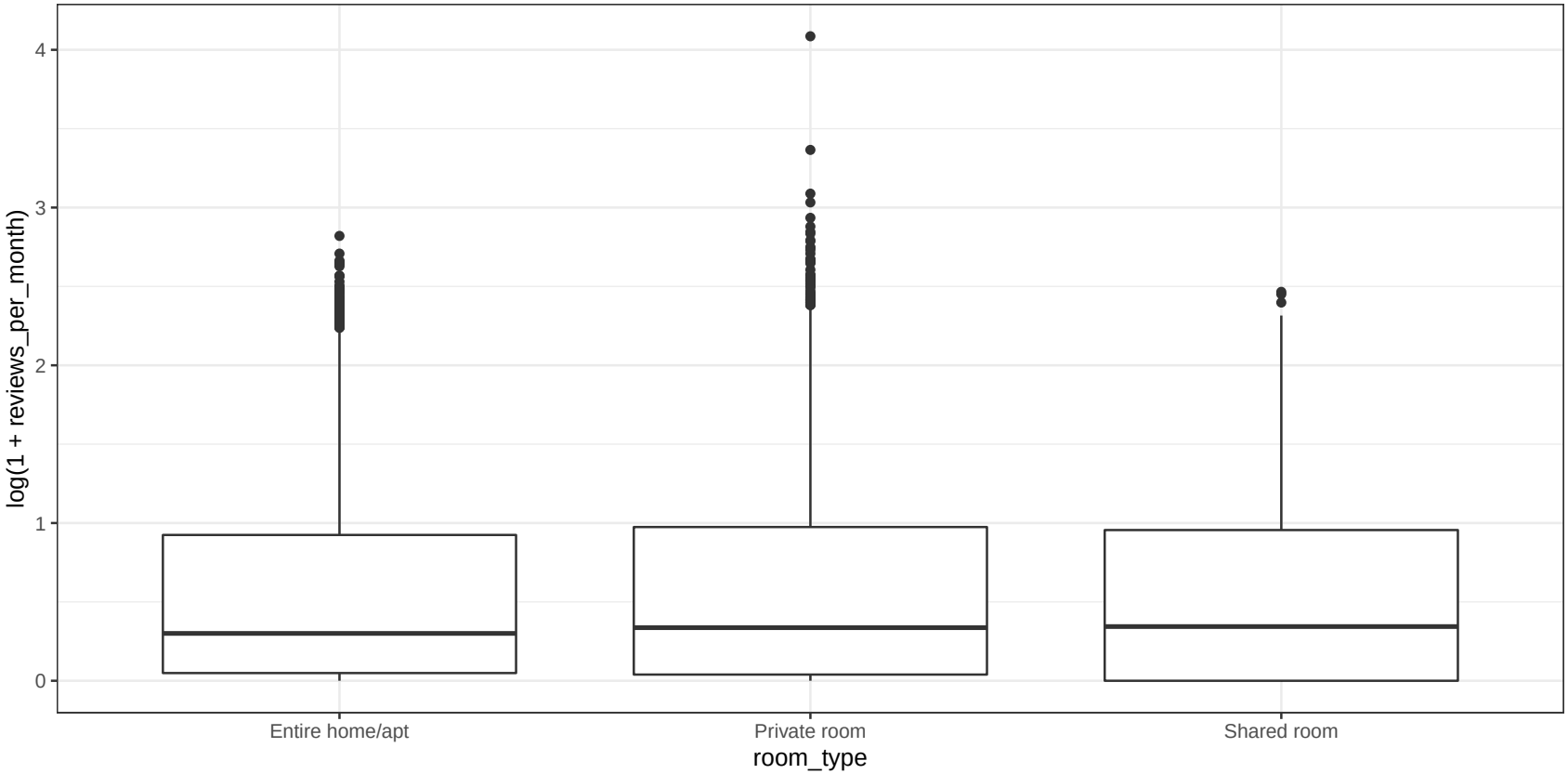


EDA: Room type

- Room type matters for price but not for popularity
- Heterogeneity of room type exists across boroughs/neighborhoods (testing needed)

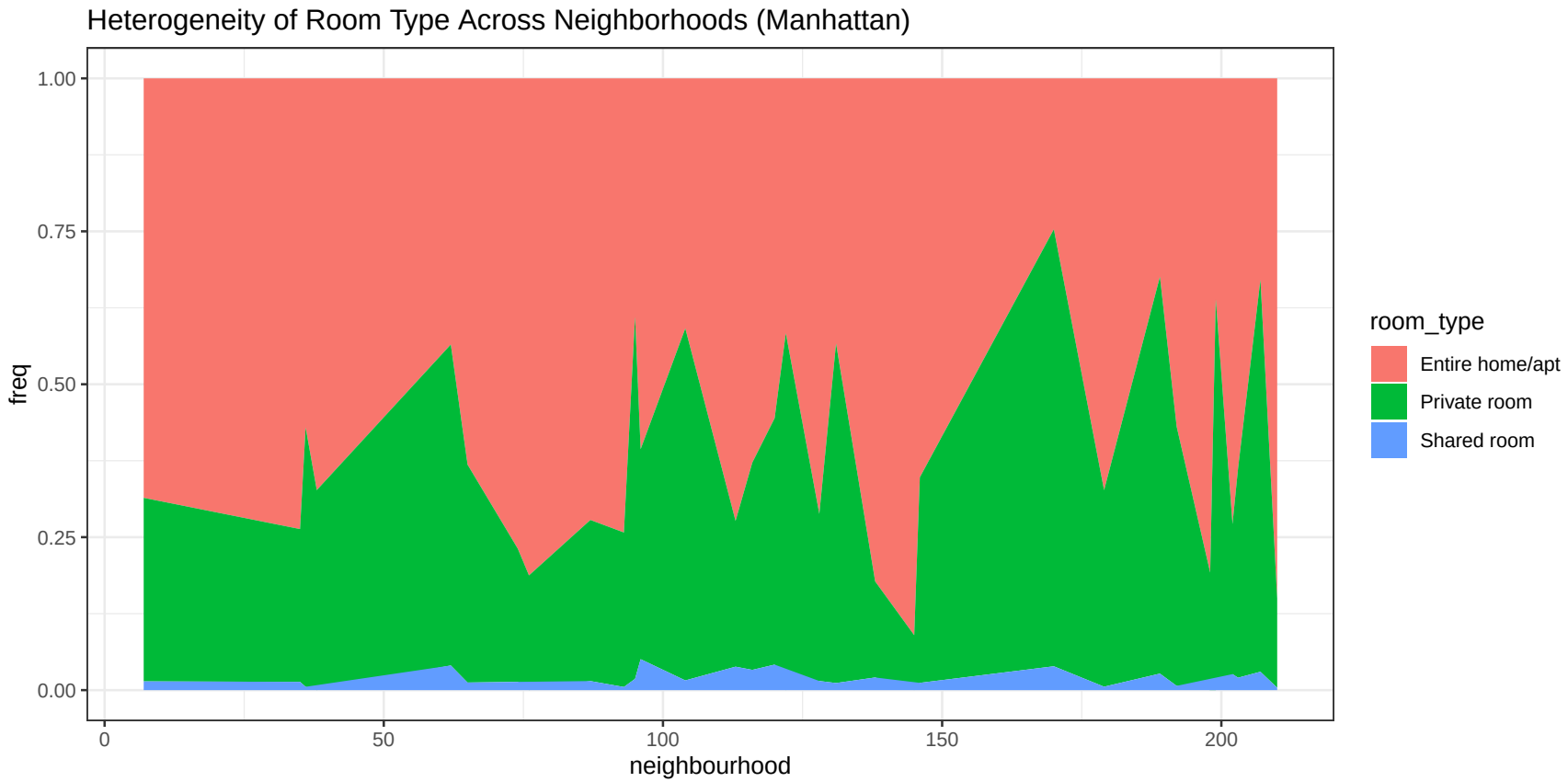


EDA: Room type



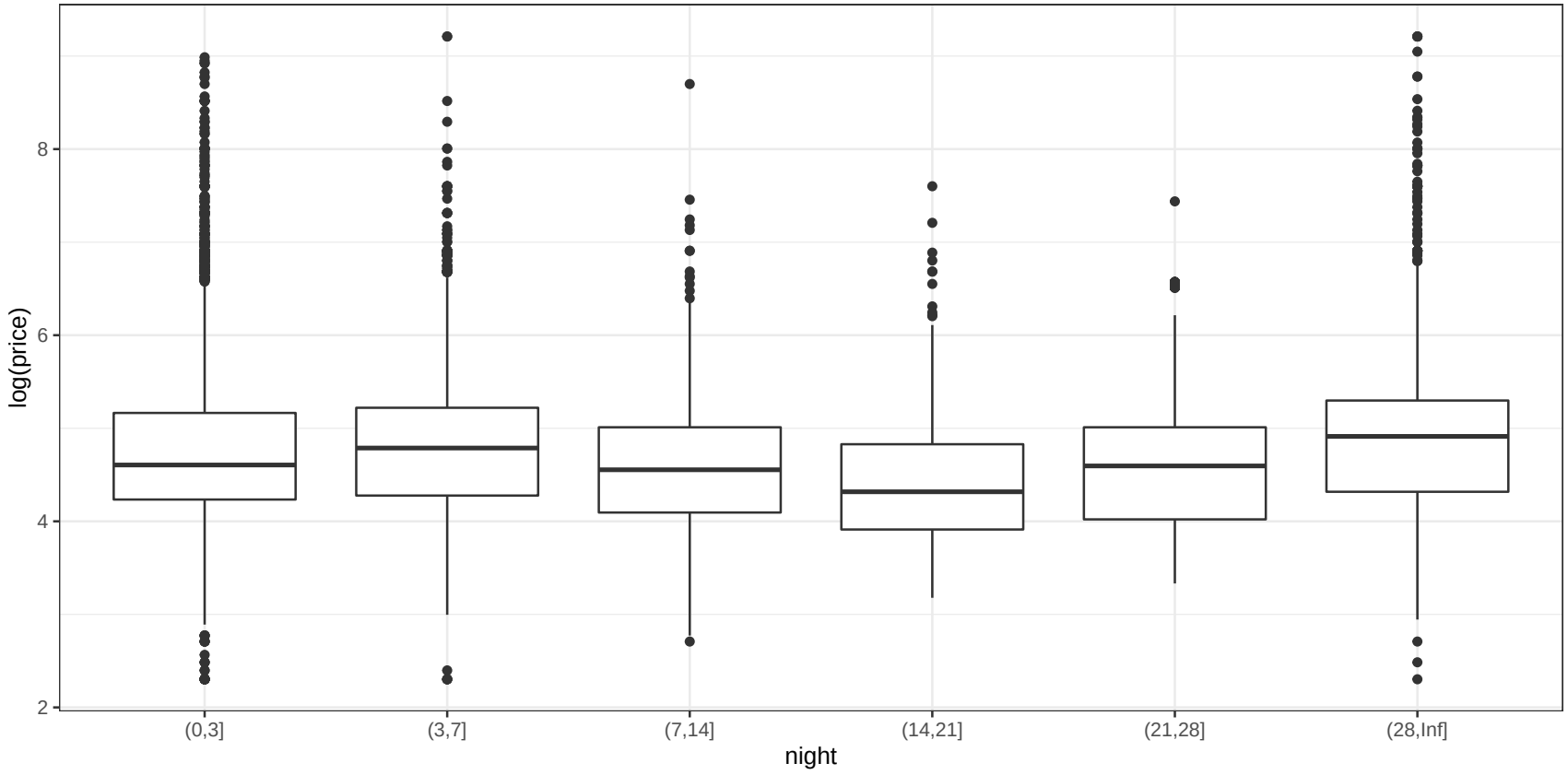
EDA: Room type

- Type of listing vary across neighborhoods.
- Pearson's Chi-squared test (stat:6363.5,df:440,p-value:$<$2.2e-16)

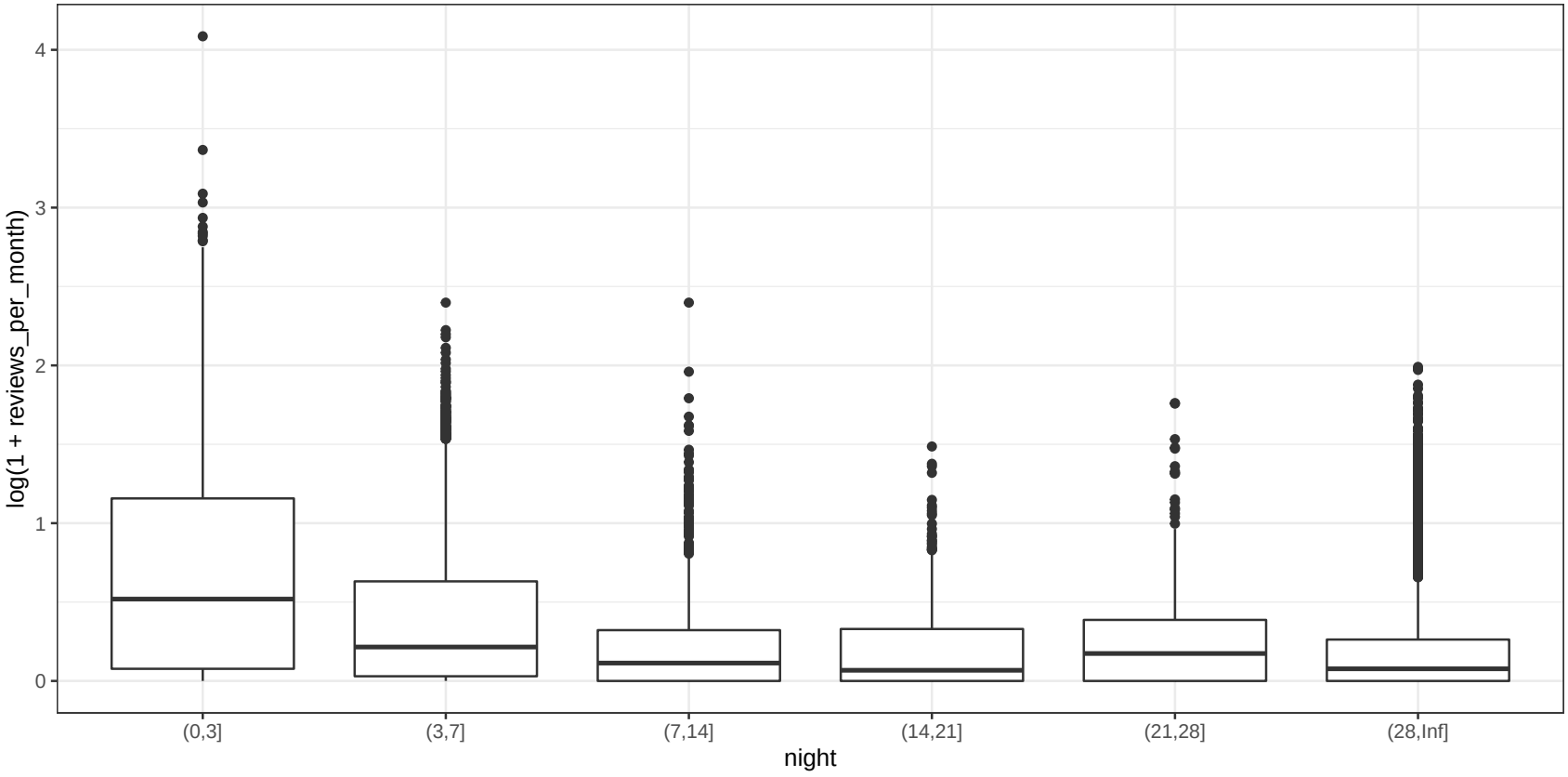


EDA: Minimum Night

- Potential effect on price/popularity exist



EDA: Minimum Night



Model: CARBayes

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

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

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

- 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: 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.

Extra predictors

- Incorporate shape file for nieghbourhoods from NYC Opendata.
- Relocate neighbourhoods (191 out of 195 neighbourhoods have airbnb listings)
- Obtain adjacency matrix $W = (w_{kl})$
- Incorporate location of metro stations from NYC Opendata
- Obtain a predictor represent the logarithm distance to the closest metro station for each listing
- Incorporate features from text analysis

Model Summary

	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: Summary for Model on price

Neighbourhood Effect on log(price)

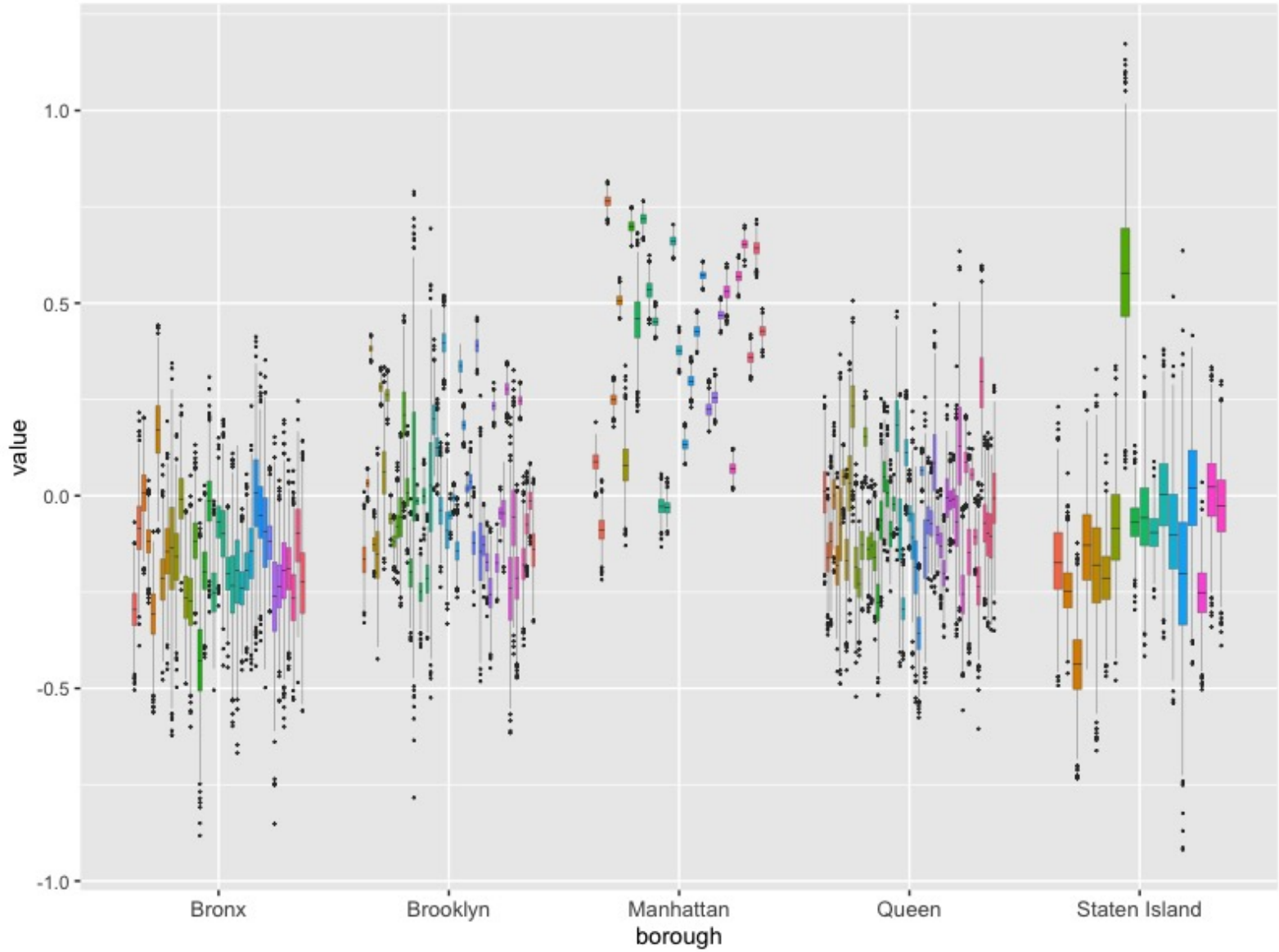


Figure 2: Neighbourhood Effect on log(price)

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 3: Summary for Model on popularity

Neighbourhood Effect on

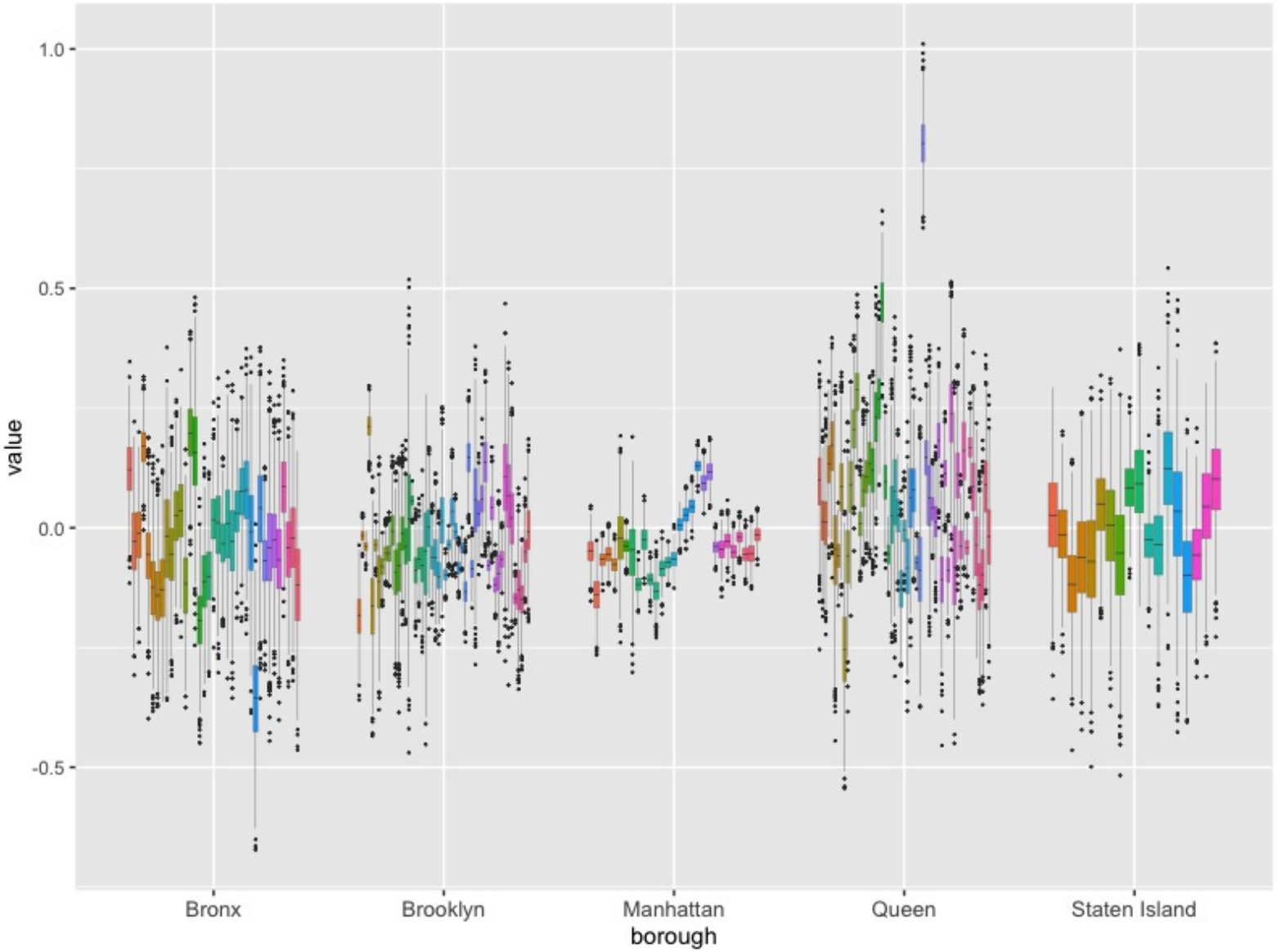


Figure 4: Neighbourhood Effect on log(1+reviews per month)

Conclusion

- Manhattan has the highest prices, Bronx the lowest.
- "Midtown South" in Manhattan is the most luxurious, "New Drop-Midland Beach" in Staten Island is the cheapest.
- "East Elmhurst" in Queen is the most popular (LaGuardia Airport), "Co-op City" in Bronx is the most unpopular.
- Entire room $>$ Private room $>$ Shared room.
- Higher minimum_nights leads to lower price.
- Longer distance to metro stations reduces price.

Text Analysis: Preprocessing

- Transform to lower case
- Remove non-informative characters
  - Punctuations, Stopwords, Whitespace, Numbers, etc.
- Word normalization
  - Porter's stemmer algorithm.
  - e.g. luxury, luxurious $\rightarrow$ luxuri

Text Analysis: Wordcloud of listings with price>2000

Figure 5: Wordcloud

Text Analysis: Latent Dirichlet Allocation

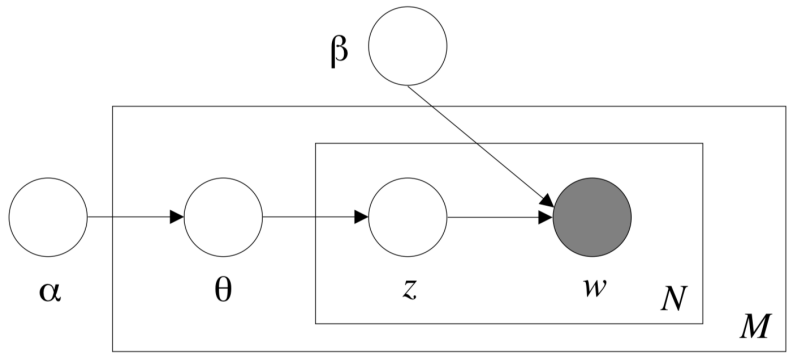
- Terms:
  - Corpus $D = \{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_M\}$
  - Doument $\boldsymbol{w} = \{w_1, w_2, ..., w_N\}$
  - Word $w_i \in \{1, ..., V\}$, V is total number of unique words.
- LDA Model:

  For all document $\boldsymbol{w}$ in D:

  1. $N \sim \text{Poisson}(\xi)$
  2. $\theta \sim \text{Dir}(\alpha)$
  3. For word w_n $(n = 1, ..., N)$

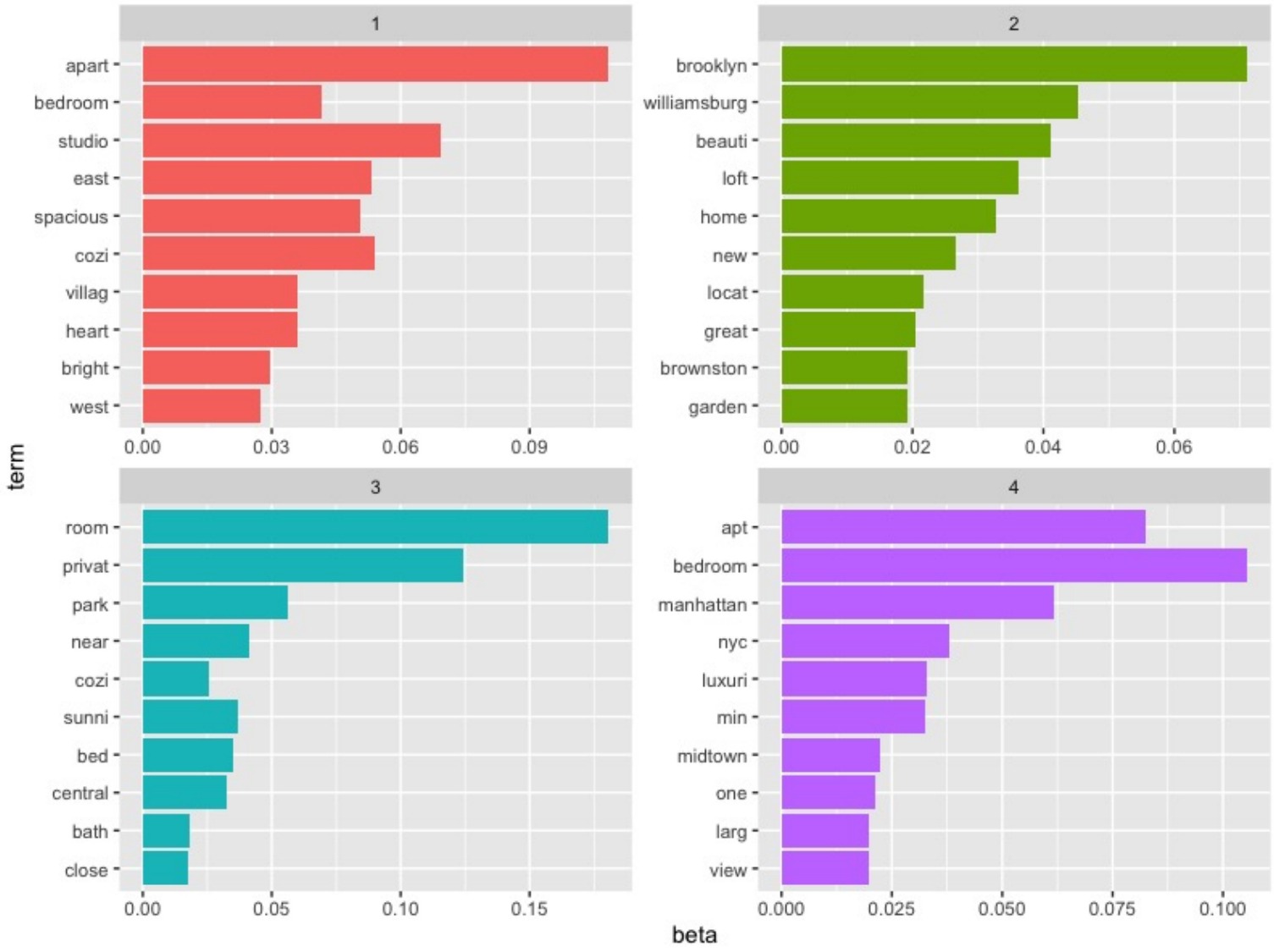
     (a) choose a topic $z_n|\theta \sim \text{Multinomial}(\theta)$

     (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.



Word frequency

Discussion

- MICE: missing data
- Include `last_review`: spatial temporal model.
- Nonlinear model: spline regression for x_{kj}.
- More spatial information (point-reference): longitude and latitude
- Add random effect for `host_id`.