NYC AirBNB Data

Shrey Gupta, Joseph Lawson, Joseph Mathews

Introduction/EDA

- New York City *Airbnb* data from 2019.
  - Variables include listing locations (neighbourhoods and NYC boroughs), price, and reviews per month.
  - Goal: Discover which variables contribute to price and popularity.
  - Presentation order: EDA/Text Analysis/Model.

Distribution of Price

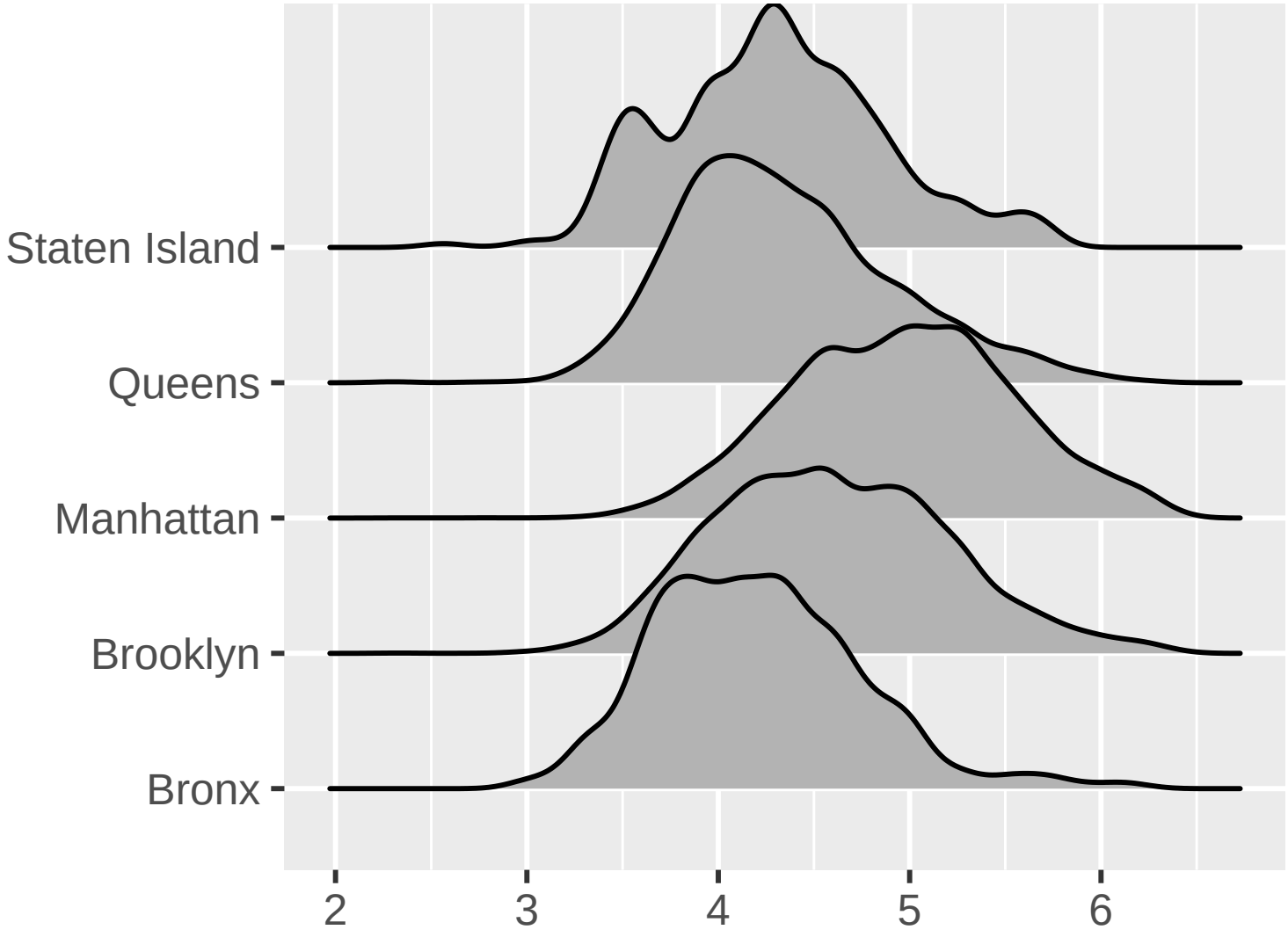


Figure 1: Log of Price

Distribution of Price

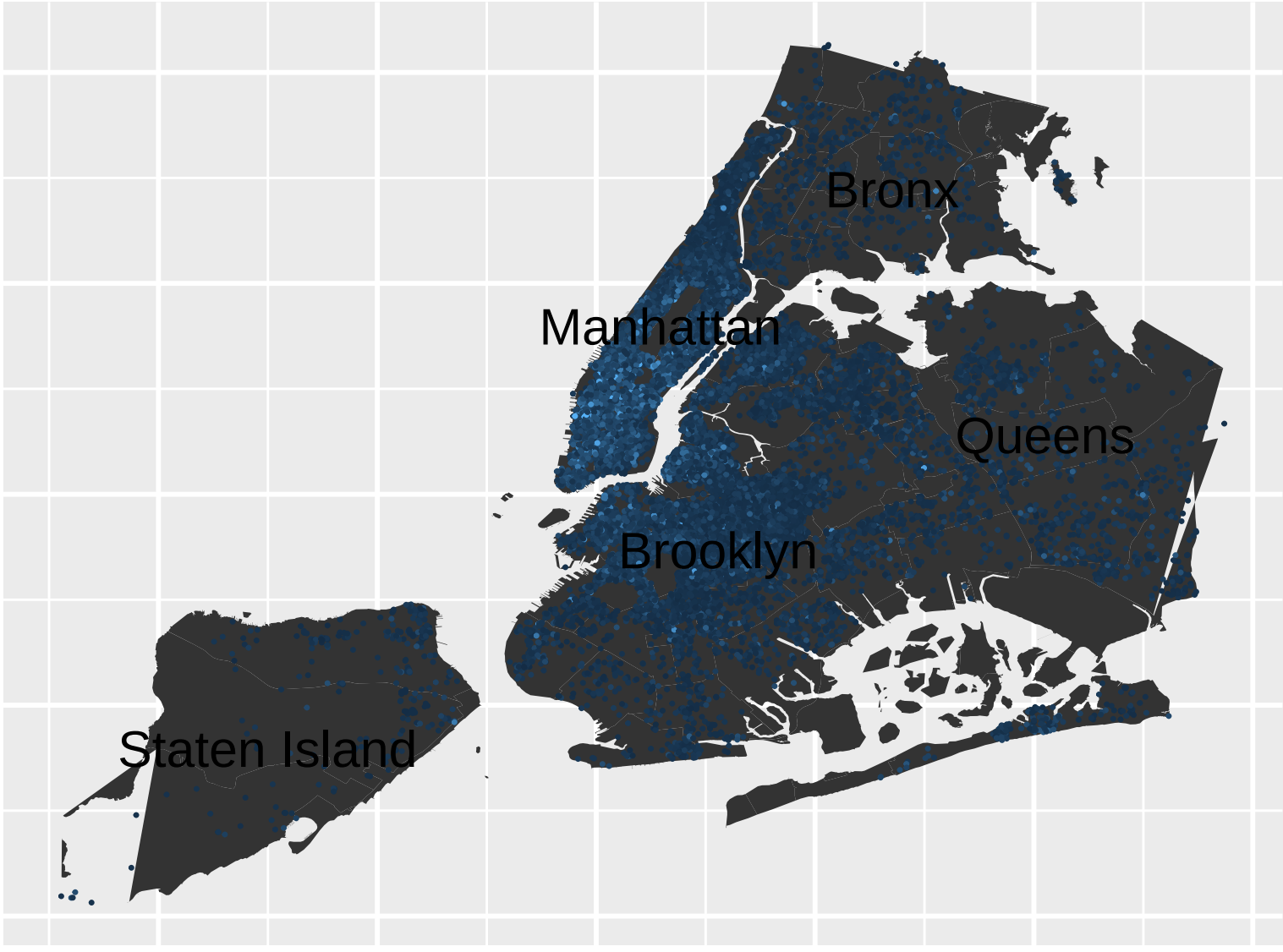


Figure 2: Distribution of Price

Distribution of Room Type

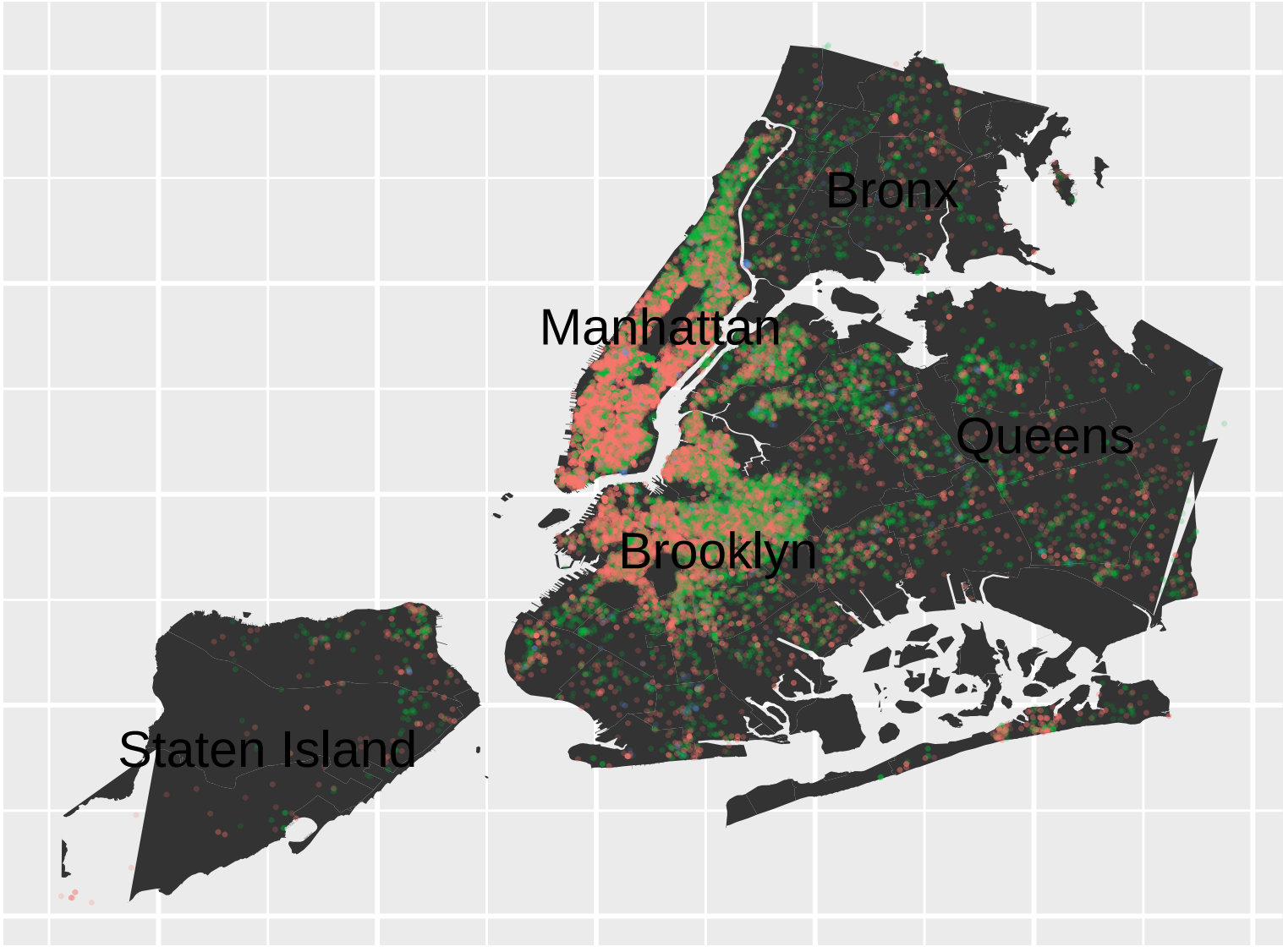


Figure 3: Distribution of Room Type Weighted by Price

Distribution of Reviews per Stay

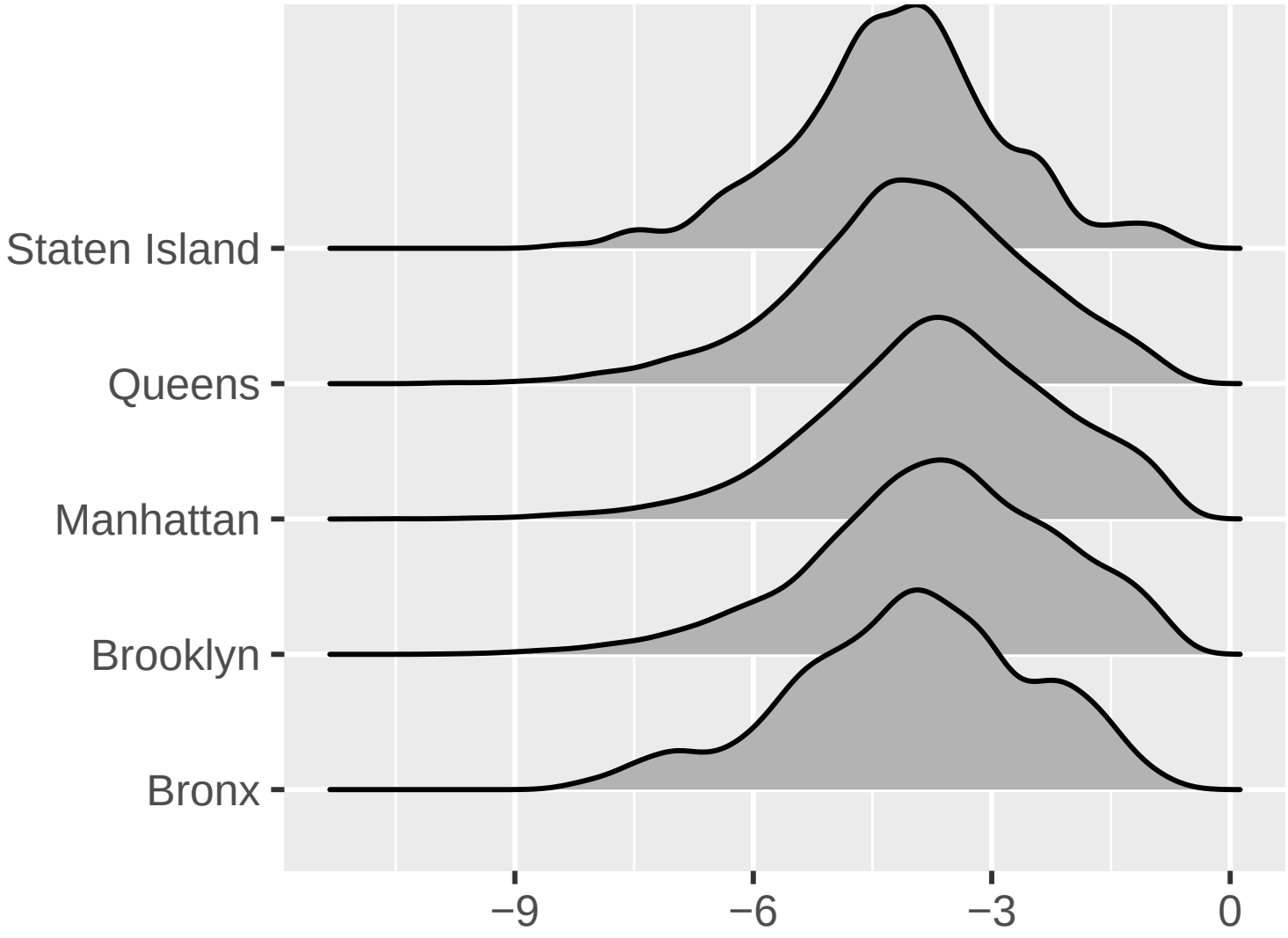


Figure 4: Log of Reviews Available Per Stay

Distribution of Reviews per Stay

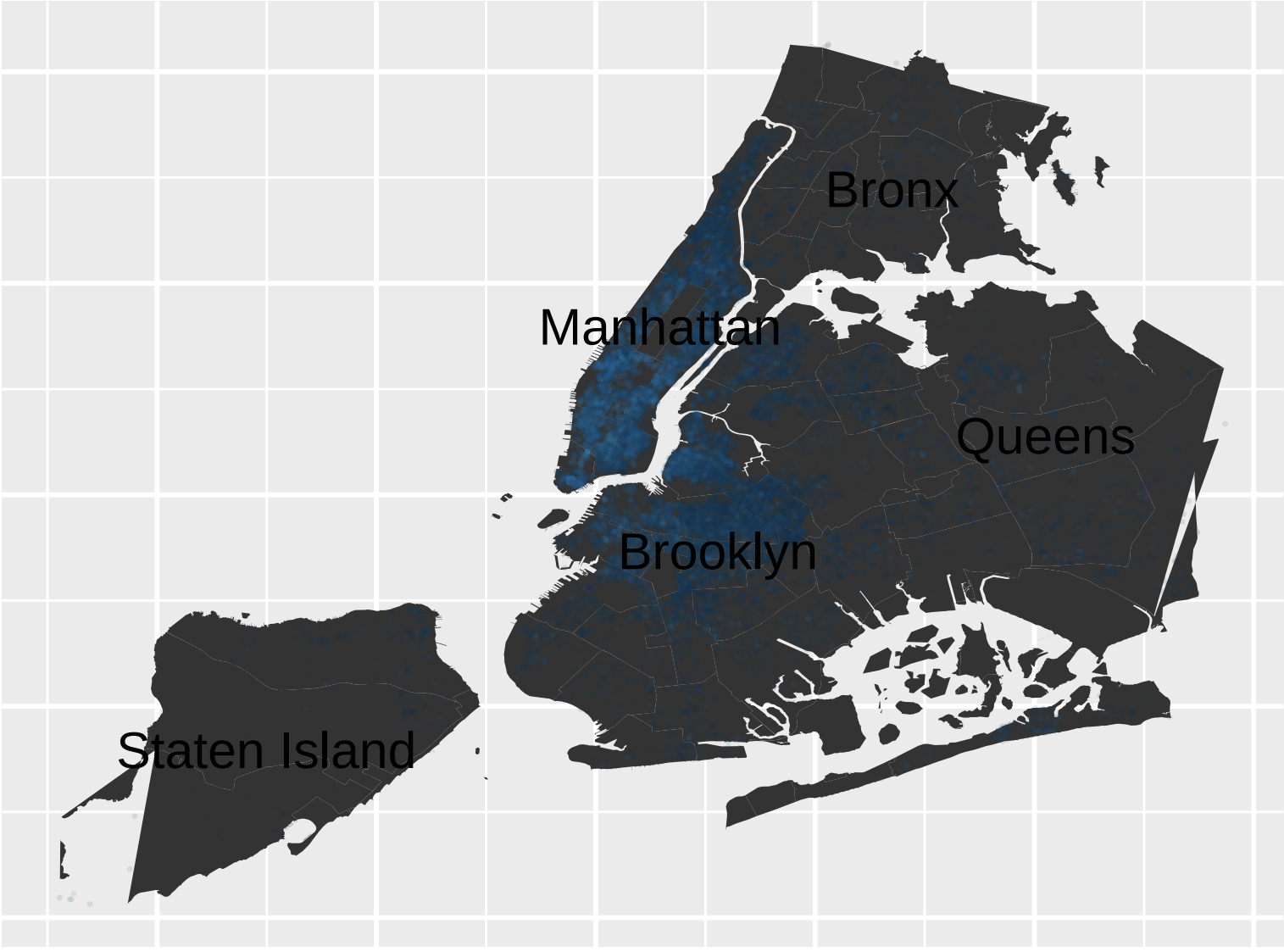
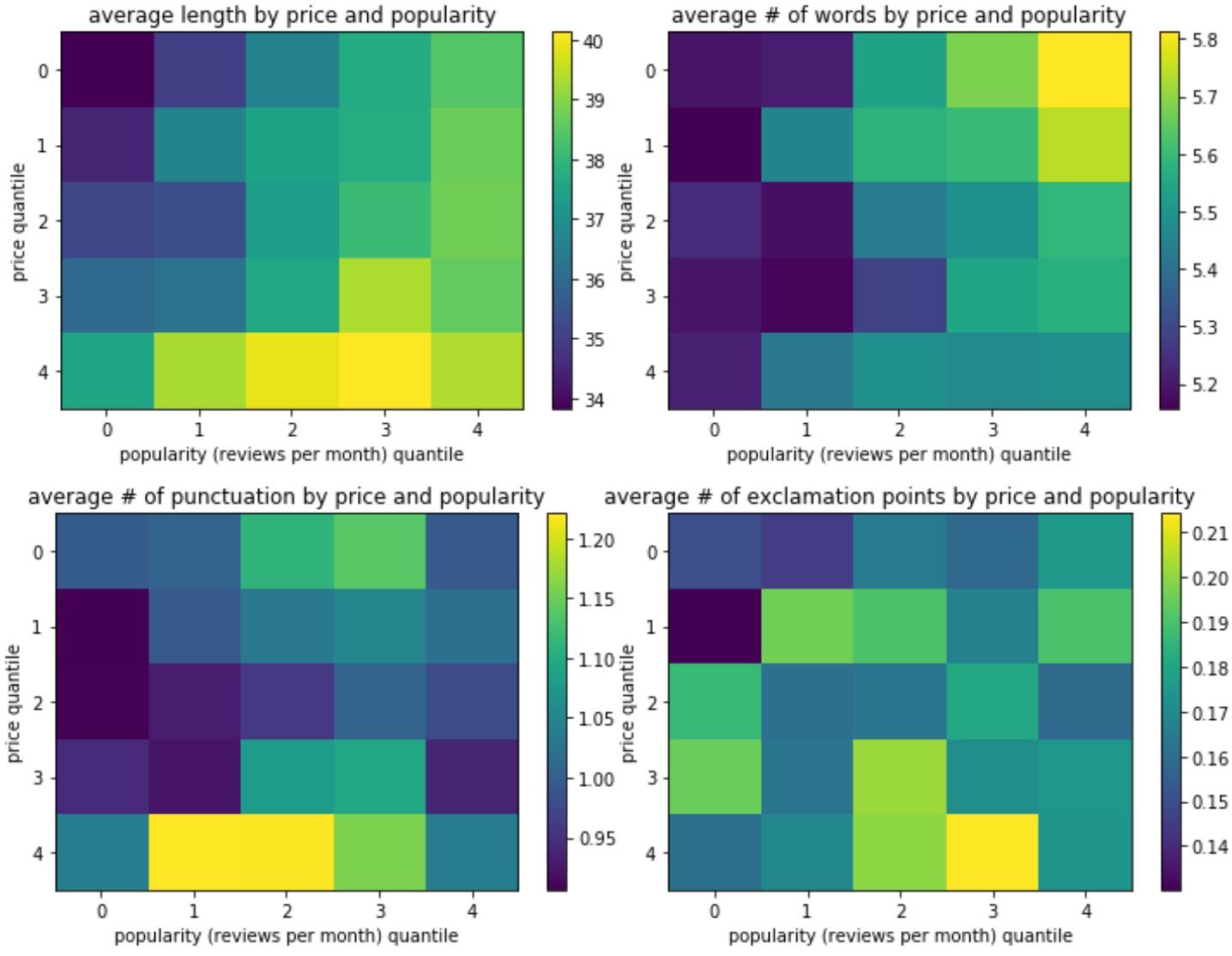


Figure 5: Distribution of Price Weighted by Reviews per Available Stay

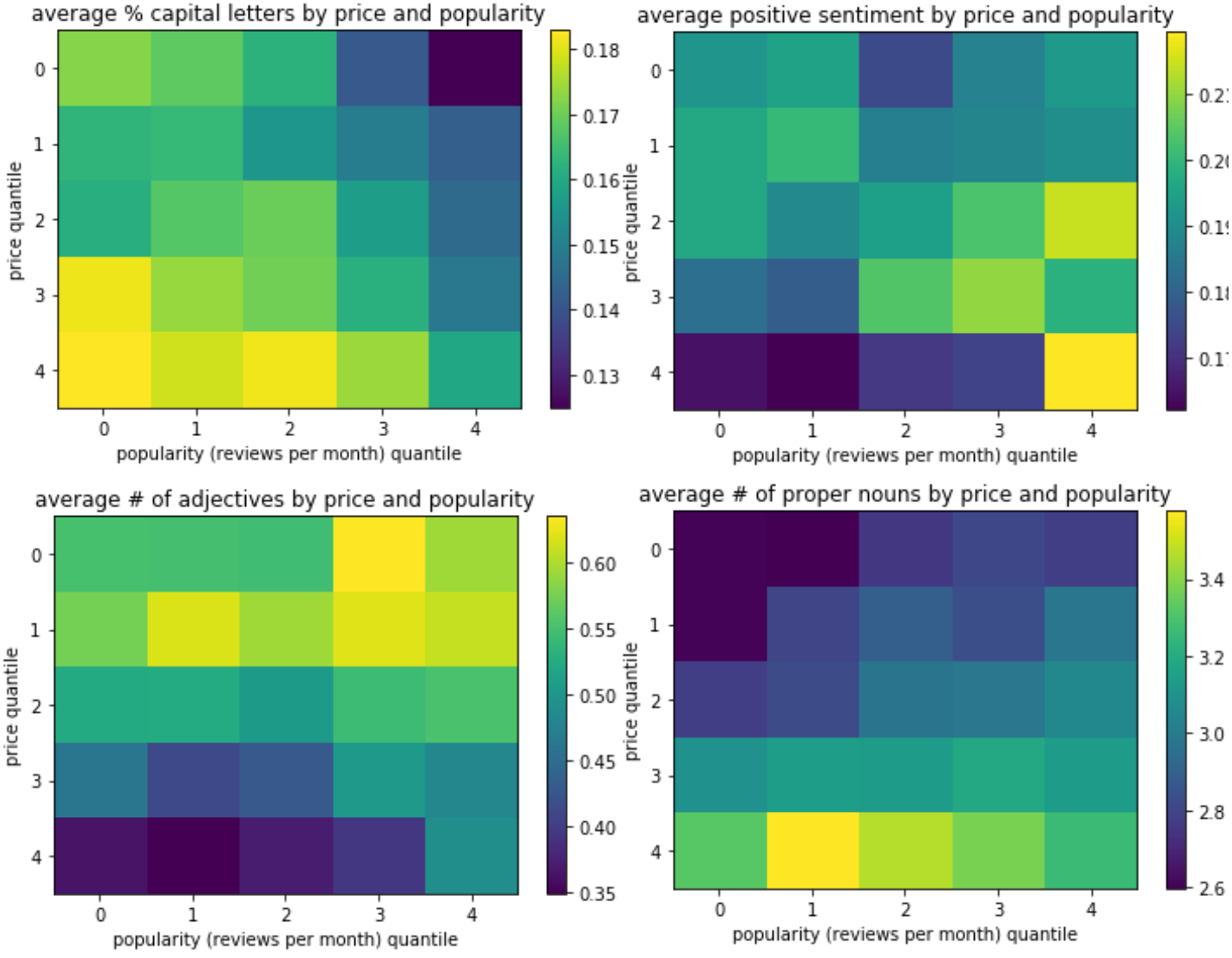
Text Analysis

- Analyzed various text features for listing names
- Length (in characters), # of words, % capitalization
- Number of punctuation, number of exclamations
- Sentiment analysis using Python's NLTK to determine positive sentiment
- Number of adjectives, number of proper nouns

Text Analysis



Text Analysis



Expensive & Popular Words

- Expensive: fee, beekman, wyndham, service, tower, spectacular, triplex
- Popular: convenience, long, easy, more, outdoor, jewel

Analysis

- Drop Zero Review/Zero Days Available
- Cap Price at 600/Cap Minimum Stay at 30
- Defined a popularity metric as:
  reviews per stay $= \frac{(\text{reviews per month})}{(\text{availability } 365)/(\text{minimum nights})}$
  - (availability 365)/(minimum nights) $\approx$ number of possible bookings
  - Serves as a "corrected" version of reviews per month.

CAR Model

Specify CAR structure for ϕ:

$$\phi_i|\phi_{-i}, \tau_i^2 \sim N\left(\rho \sum b_{ij}\phi_j, \tau_i^2\right)$$

- Set $b_{ij} = w_{ij}/w_{i+}$ (w_{i+} is $i's$ neighbor count)
  - Set $\tau_i^2 = \frac{\sigma^2}{w_{i+}}$

Then we have joint distribution:

$$\phi|W, \sigma^2 \sim N(0, \sigma^2(D - \rho W)^{-1})$$

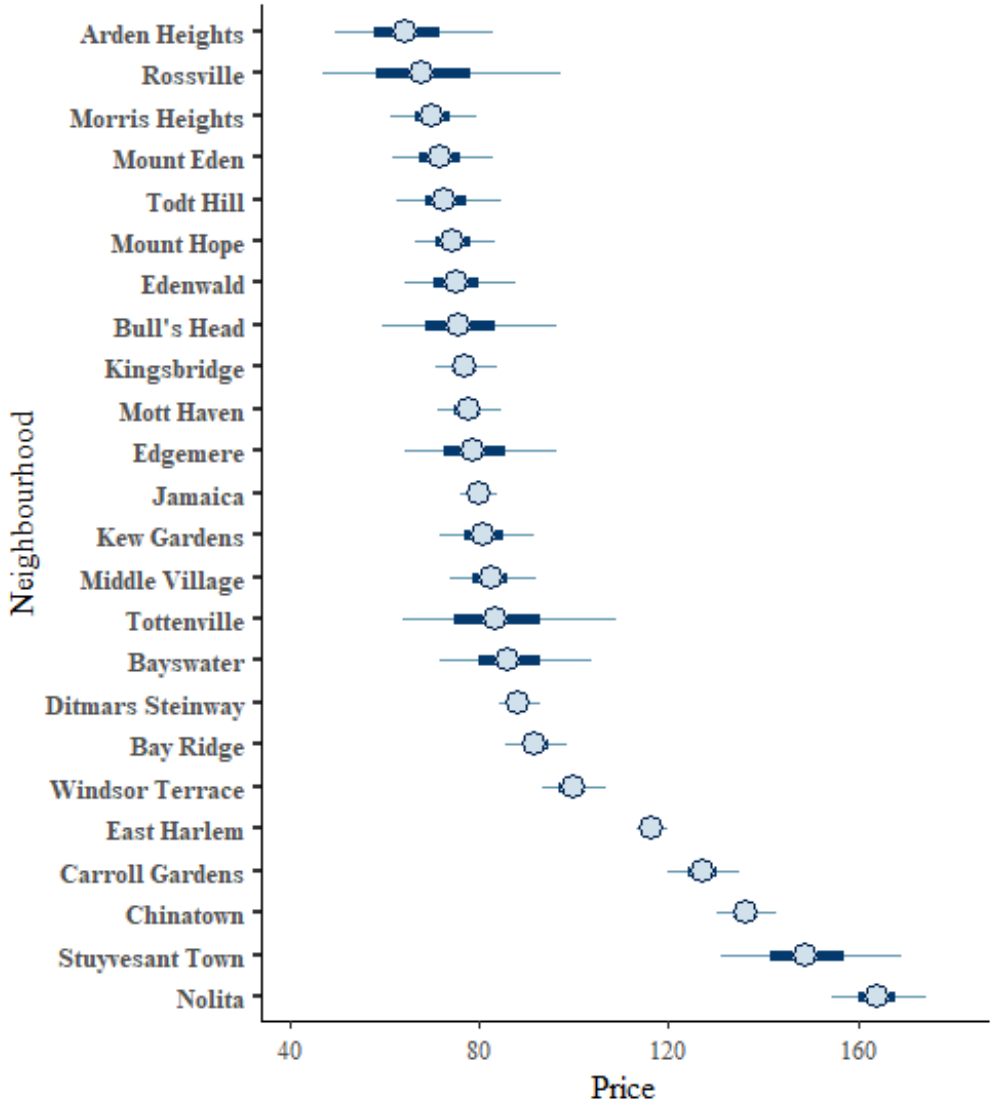
- W is adjacency matrix of neighborhoods
  - D is diagonal with d_{ii} the count of $i's$ neighbors
  - $\rho \in (0, 1)$ calibrates strength of relationship/controls pairwise covariance

(See (Gelfand, Vounatsou 2003.) for further details)

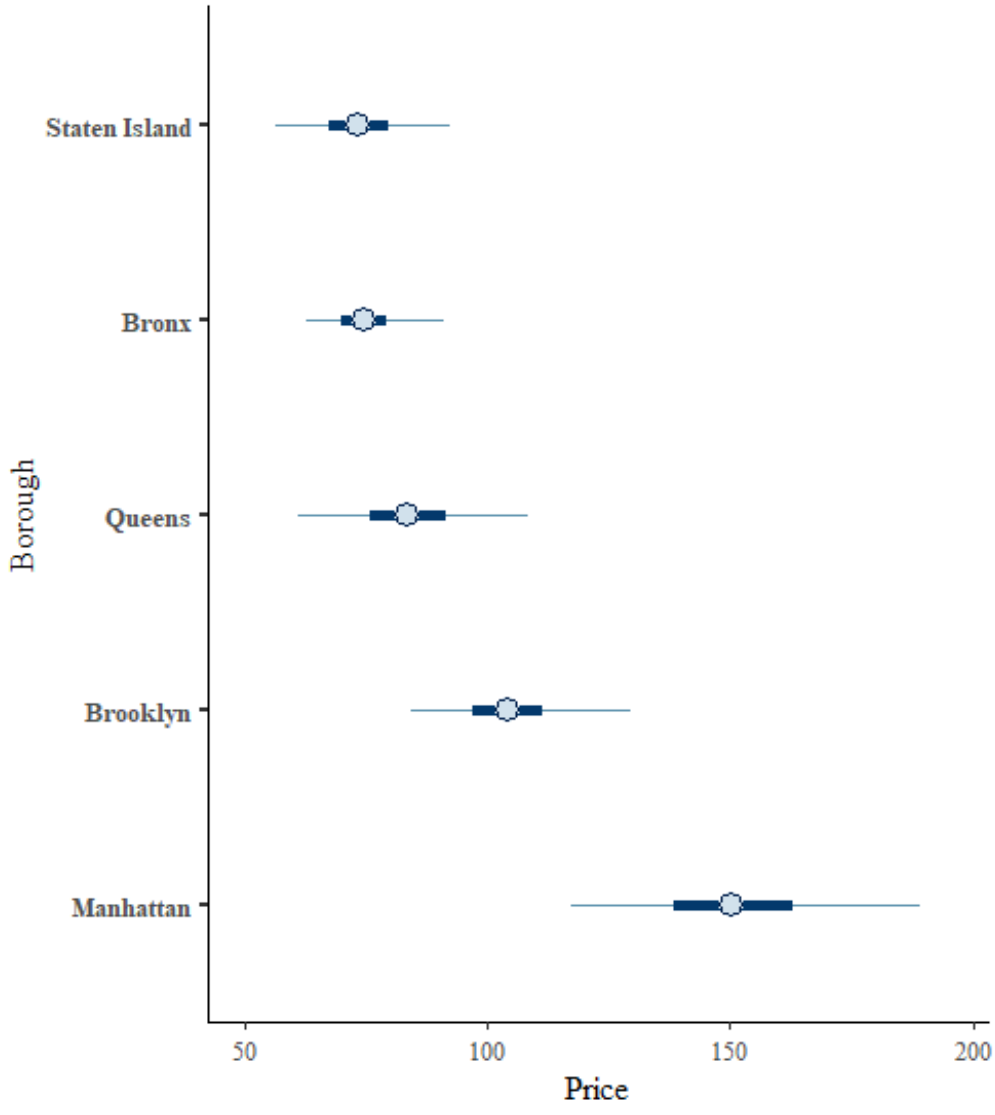
Adjacency Matrix

- Define closeness as mean pairwise distance between listings in pairs of neighborhoods
- Establish threshold (.05) for adjacency, and require same Borough
- Ends up similar to taking distance between mean coordinates
- Considered something like Earth Mover's Distance, but was computationally excessive
- More representative of geographic "closeness" than neighborhood boundaries

Price Model Output



Price Model Output

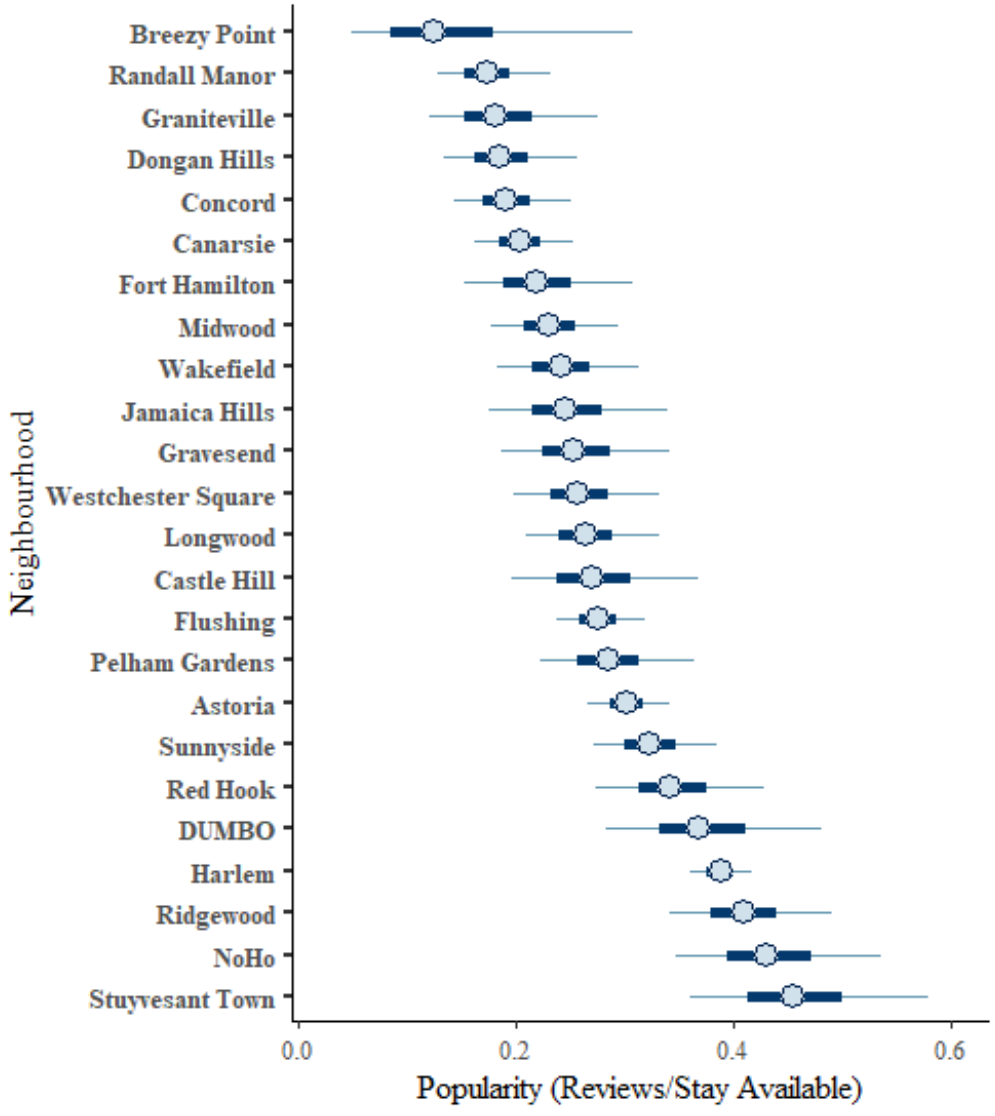


Price Model Output

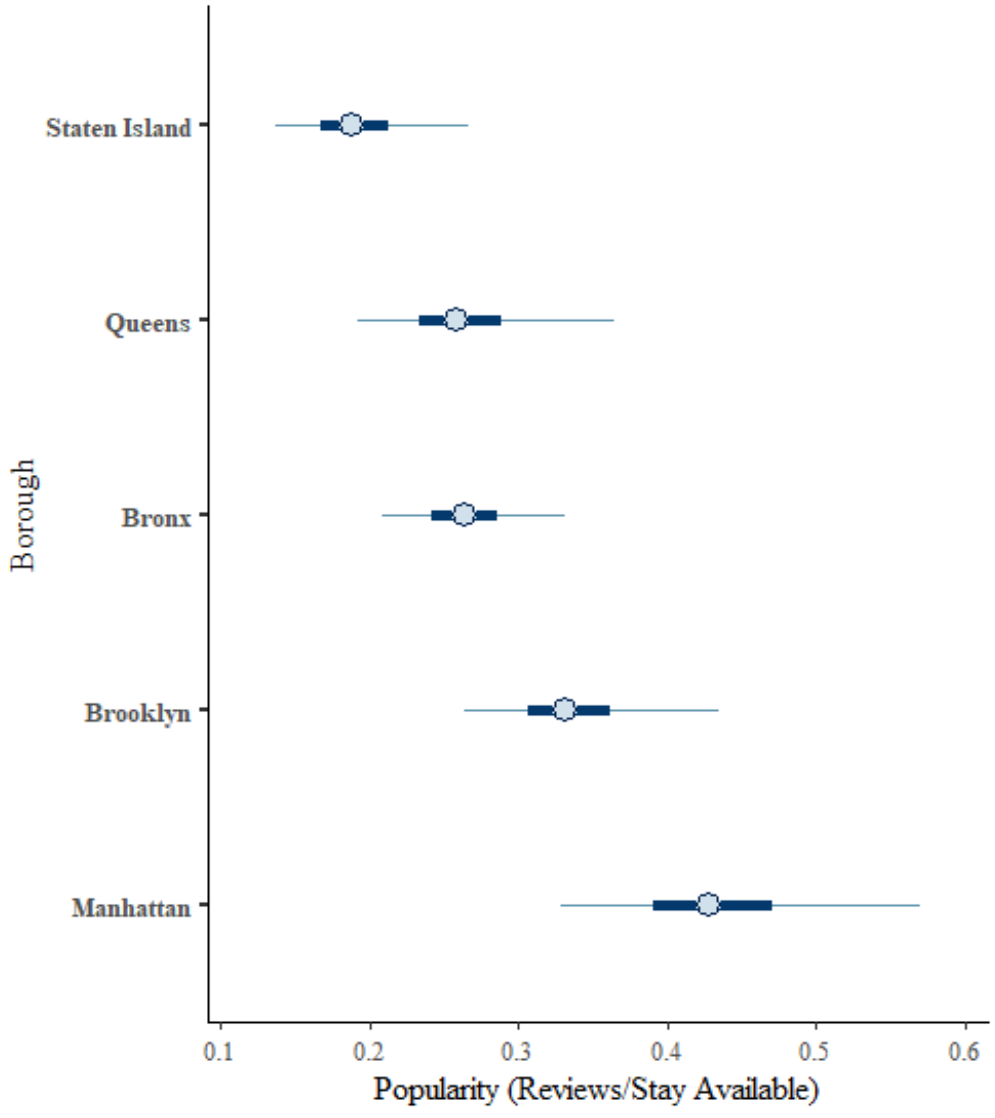
Table 1: Percentage Change in Price per Unit of Variable

	Lower	Mean	Upper	Significant
room_typePrivate room	-50.95	-50.43	-49.89	1
room_typeShared room	-69.19	-68.17	-67.12	1
calculated_host_listings_count	-0.08	-0.06	-0.04	1
availability_365	0.04	0.04	0.05	1
pos_sent	-6.77	-4.74	-2.69	1
exclamations	-1.33	-0.36	0.62	0
words	0.44	0.77	1.10	1
pct_capitals	-3.86	-0.48	2.97	0
avg_word_len	0.51	0.73	0.95	1
pct_adj	-8.89	-4.83	-0.60	1
pct_ppn	3.02	5.04	7.11	1
punc_no_excl	0.23	0.74	1.25	1

Popularity Model Output



Popularity Model Output



Popularity Model Output

Table 2: Percentage Change in Popularity per Unit of Variable

	Lower	Mean	Upper	Significant
room_typePrivate room	-41.32	-38.55	-35.73	1
room_typeShared room	-68.12	-63.36	-58.07	1
calculated_host_listings_count	-0.60	-0.52	-0.44	1
pos_sent	-1.66	7.70	17.92	0
exclamations	-7.18	-3.26	0.79	0
words	9.65	11.18	12.71	1
pct_capitals	-52.42	-44.75	-36.24	1
avg_word_len	1.27	2.23	3.20	1
pct_adj	-15.29	2.48	23.18	0
pct_ppn	-11.51	-3.65	4.66	0
punc_no_excl	1.13	3.35	5.59	1

Listing Type Hetergeneity Across Neighborhoods/Boroughs

Test	ChiSq.P.Val	Fisher.P.Val
Overall.Nbhd	0.000	0.0000
Brooklyn.Nbhd	0.000	0.0000
Manhattan.Nbhd	0.000	0.0000
Queens.Nbhd	0.000	0.0000
Staten Island.Nbhd	0.005	0.0024
Bronx.Nbhd	0.000	0.0000
Borough	0.000	0.0000

- Low cell count renders Chi-Square inexact
- Fisher Exact test as alternative
- Both indicate strong heterogeneity, as suggested by EDA
- Staten Island a bit less-so

Conclusion

- Based on our results, an example of an expensive listing is:
  - **entire apartment is available for rent conveniently located in the Tribeca neighborhood of Manhattan island. . . please contact for details. . . pricing subject to negotiation"**
- Replace location with “East Harlem” to maximize popularity.
- Replace location with “Financial District” to balance both.
- Posterior distribution of CAR parameter α suggests adjusting for spatial dependency is warranted.