Popularity of NYC Airbnb Listings

Olivier Binette, Justin Weltz, and Brian Kundinger

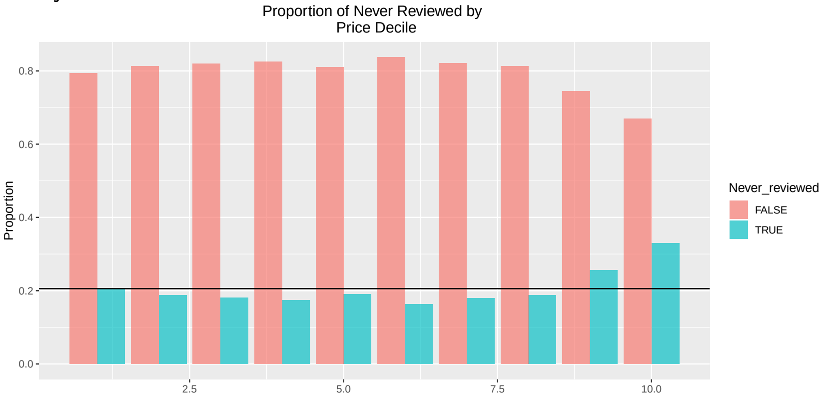
Goals

- Determine “neighborhood brand” effects.
- Examine the features of a popular listing.

Data cleaning and transformations

Listings with "minimum stay" greater than 30 often had extremely high prices, suggesting some were monthly prices. To avoid these innacuracies, we limit analysis to listings with "minimum stay" less than 7.

Listings with high prices have no reviews at disporportionately high rates. Therefore, we remove listings with no reviews from our analysis.



Unreliability of “Reviews per Month”

We can see that the study uses a different number of total months for each listing when calculating “average reviews per month.”

	id	No._reviews	last_review	Avg_reviews
1	5447434	20	2016-01-03	0.41
2	30232758	20	2019-07-02	2.84

However, we can recover the “months active” by using the “last review,” and construct a more reliable popularity metric.

Normalized Average Reviews

$$\frac{\text{Days between July 8, 2019 and Last Review}}{30.42} = \text{Months Inactive}$$

$$\frac{\text{No. of Reviews}}{\text{Months Active} + \text{Months Inactive}} = \text{Reviews per Month}$$

$$\text{Months Active} = \frac{\text{No. of Reviews - Avg Reviews*Months Inactive}}{\text{Reviews per Month}}$$

$$\text{Normalized Average Reviews} = \frac{\text{Total Reviews}}{\text{Months Active}}$$

Normalized Average Reviews

Through normalization, we recover the popularity of each listing during the time it was active.

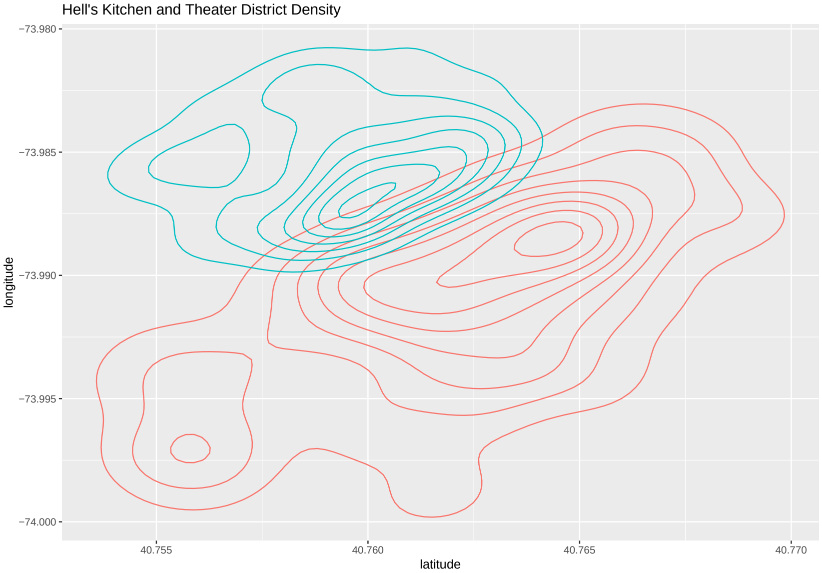
	id	No._reviews	last_review	Nrmlized_reviews
1	5447434	20	2016-01-03	2.86
2	30232758	20	2019-07-02	2.86

Text Analysis

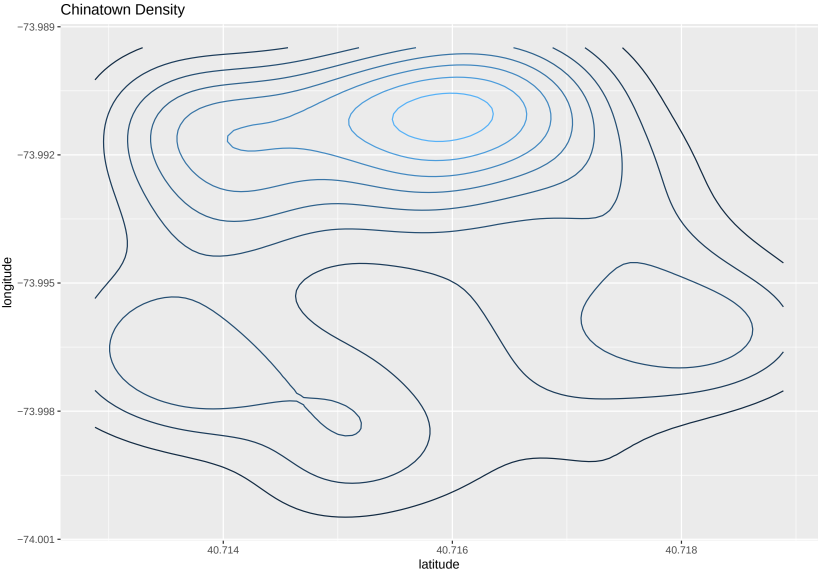
We identified all adjectives that occured in 10 or more listings (219 words). We then ran a linear regression of average reviews against these words, and identified all words with p-values below the Bonferonni adjustment of $\frac{0.05}{218}$. This left 18 words for further analysis.

private, square, minute, close, stock, apt, sunny, spacious, deluxe, newly, walking, central, fast, huge, west, green.

Spatial Auto-correlation: Theater District



Spatial Auto-correlation: Chinatown



Conditionally Autoregressive Model

The Leroux et. al. model:

Priors:

$$\phi_k|\phi_{-k}, W, \tau^2, \rho \sim N(\frac{\rho \sum_{i=1}^{k} w_{ki}\phi_i}{\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho})$$

$$\tau^2 \sim Inverse - Gamma(a, b)$$

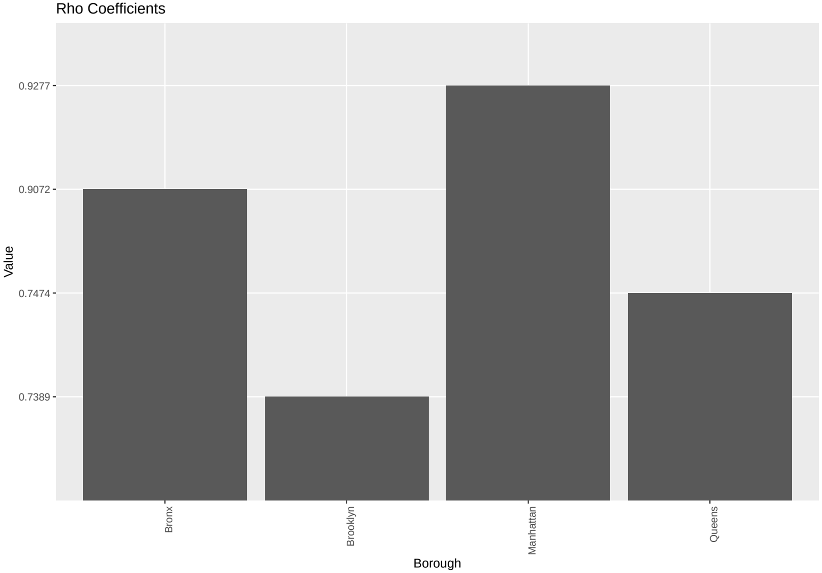
$$\rho \sim Uniform(0, 1)$$

This induces neighbor spatial correlation:

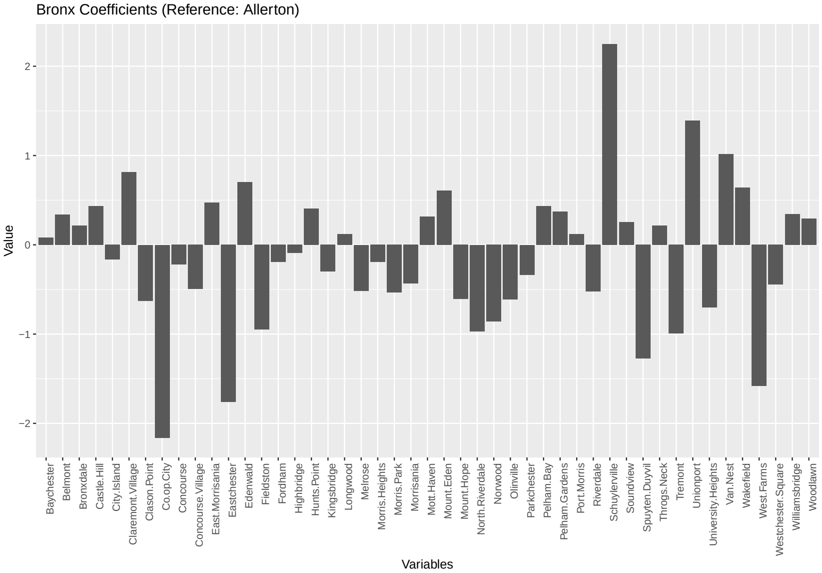
$$COR(\phi_k, \phi_j|\phi_{-kj}, W, \rho) =$$

$$\frac{\rho w_{kj}}{\sqrt{(\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho)(\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho}}$$

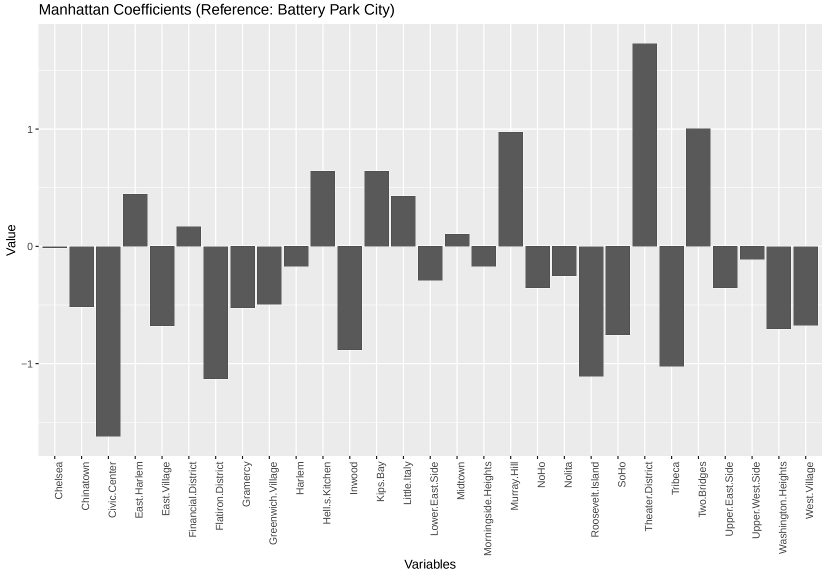
Evidence of Auto-Correlation



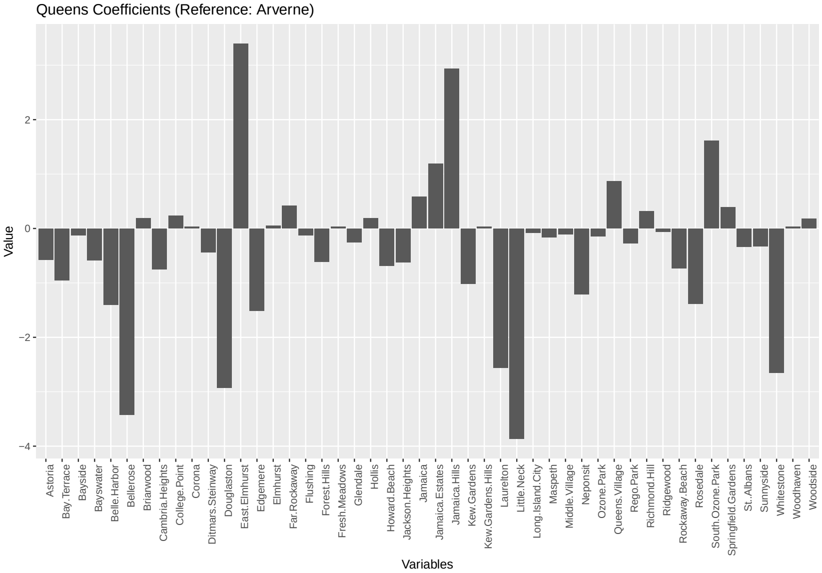
CAR Results Bronx



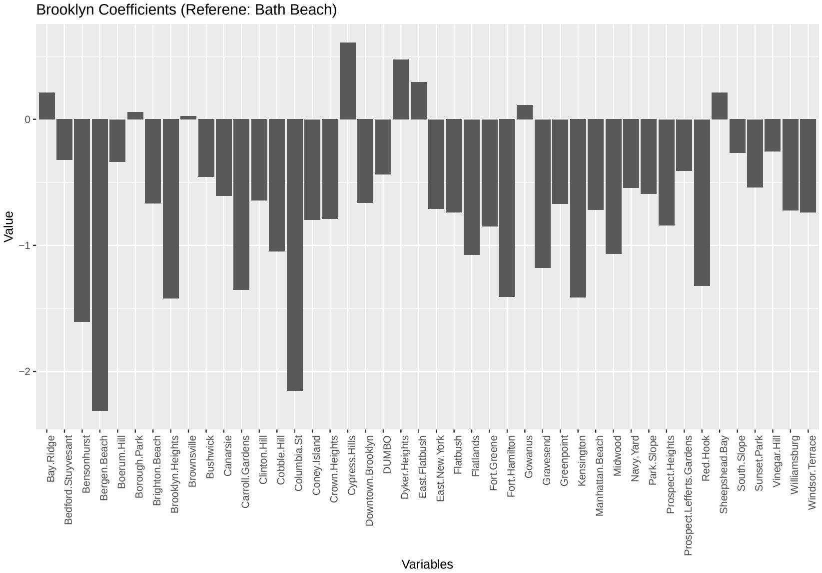
CAR Results Manhattan



CAR Results Queens



CAR Results Brooklyn



Global model: shape-constrained GAM

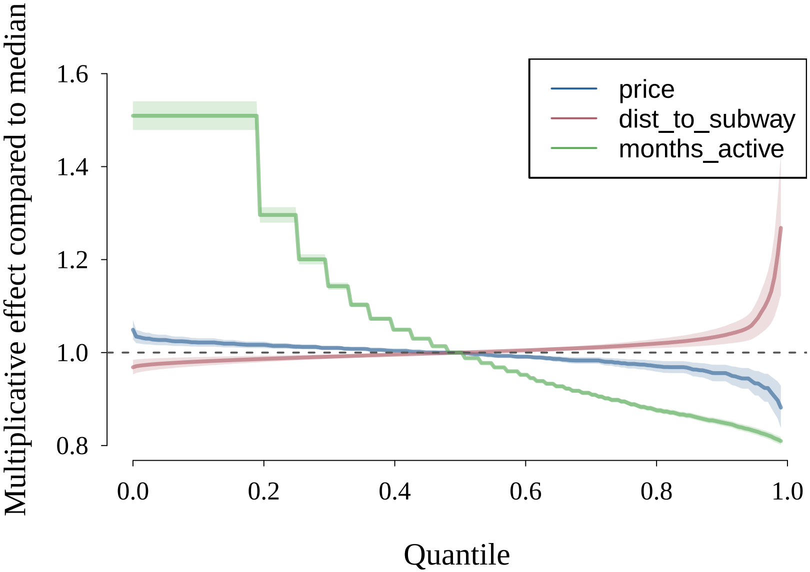
Log-linear model

$$\log(\texttt{monthly reviews}) \sim$$
$$f_1(\texttt{price}) + f_2(\texttt{dist to subway}) + f_3(\texttt{months active}) +$$
$$\texttt{entire home/apt} + \texttt{gender} + \texttt{private} + \cdots + \texttt{cozy} +$$
$$\texttt{neighbourhoods}$$

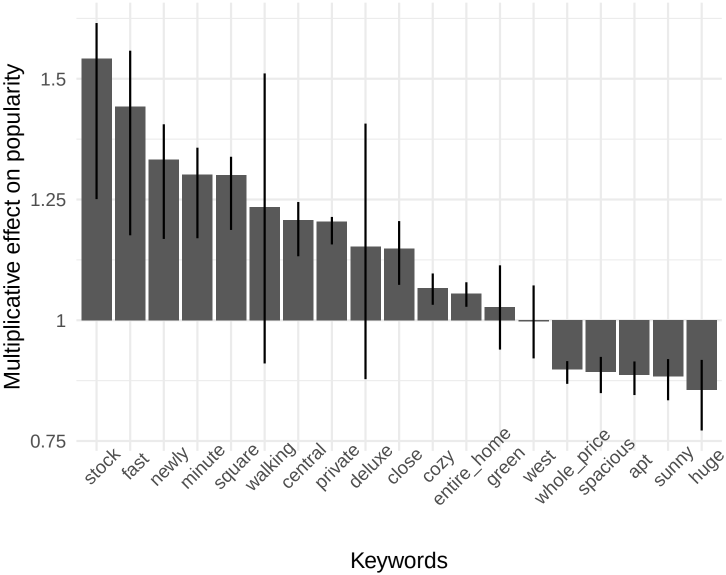
where f_1, f_2 and f_3 are increasing functions.

- Allows us to estimate trends in an interpretable way.
- Tried shape-constrained generalized additive model packages (cgam and scam), but this didn't run in finite time.
- Instead boxcox transforms of the predictors.

Results for the global model



Results for the global model



Improvements

- Integrating the analyses
  - Combining keyword selection with full model fitting (e.g. with a weighted lasso).
- Non-linear relationships