Case study 2: New York Airbnb listings

Olivier Binette, Brian Kundinger and Justin Weltz

1. Introduction

Airbnb is a online marketplace of rooms, houses and appartments to rent for short periods of time, as an alternative to hotels and other accomodation services. Having started just around 2008, the growth of Airbnb has been phenomenal. Now more than 6 million listings are available worldwide.

This case study focuses on New York City 2019 listings compiled by Murray Cox as part of the Inside Airbnb project, originally with the goal of supporting thoughtful debate surrounding the socio-economic impact of this growing business. We study this data more superficially, focusing on (1) the neighborhood brand effects on listing popularity; and (2) how different listing features are associated with popularity.

Our measure of popularity is defined as the monthly number of reviews per month, properly normalized to account for disrepancy in the data (see Section 2.2). A conditional auto-regressive spatial model is used to assess neighborhood brand effect (Section 2.4), while a global additive model allows us to extract conditional trends related to listing popularity (Section 2.5). The results of our analyses are presented in Section 3, supported with Figures listed in the Appendix. We reflect more broadly on the problem in Section 4.

2. Materials and methods

2.1 Data

The `Airbnb` dataset contains 48 895 listings from New York City which have been active before July 8, 2019. Information is given on the name of the listing, the name of of the host, the type of room available (entire home, private room or shared room), on the minimum number of nights for a rental, on the yearly availability, on the location of the listing (coordinates and neighborhood), on its price, on the number of reviews, and on the average number of reviews per months.

We note that 35% of the listings are recorded as having zero availability for booking. It is unclear in these cases if the listing is fully booked, or if it is simply no longer available for booking and has been temporarily disabled. Furthermore, the listed availability may have changed throughout time, and we are only provided with its value at the time when the data was collected. We do not use this variable in our analyses given these challenges in interpretability.

2.2 Data cleaning and transformations

Among the listings requiring stays of longer lengths of time, it was often unclear whether the price listed was a monthly, weekly, or daily price. We focus our analysis on listings which can be rented for less than 7 nights. Additionally, listings which had recieved no reviews had relatively high prices in comparison with the rest of the data set. Since listings that had recieved no reviews were not representative of the active listings, we removed these listings from our analysis. Lastly, we remove 11 listings that are priced at more than five thousands dollars a night. Many of these were event spaces or film and photography venues, and thus were not representative of typical Airbnb listing. We thus conducting our analysis with 32819 data points representing typical listings which can be rented for a few nights at a non-extravagant price.

As a measure of popularity, we consider a normalized average number of review per months. Indeed, it appears that the average number of review per months has been computed including the months a listing has been inactive. We have therefore renormalized the average by removing on the denominator the number of months since the last time a listing received a review.

2.3 Computed features

The Airbnb dataset was augmented with the following features:

1. Text analysis. We extracted all adjectives from the listing names and kept those with a frequency of at least 10. 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, and the data has been augmented to contain indicators of the presence or absence of each of these words.

Of important note, some of the words picked up by our algorithms are most likely not adjectives in the context of this dataset. For example, "apt" is most likely a shortening of "apartment," and not an adjective meaning "correct," and minute is most likely a measurement of time and proximity, and not an adjective meaning "small." We choose however to retain these words for the purpose of this analysis.

2. Distances to subway stations. We computed distance between listings and the nearest subway entrance using data from New York's Open Data portal.

3. Gender imputation. We imputed hosts gender using data in the R package `genderdata` through the interface package `gender`. Host names were compared with records from the U.S. Social Security Administration, and imputation was made when there was more than 90% match to a given gender. Otherwise, such as when the host name consisted of a pair of person, an "Unknown" gender was recorded.

2.4 Conditional autoregressive model to assess neighborhood effect on popularity

In order to advise a new Airbnb owner on the best neighborhood to set up a property, we have to be careful to separate the neighborhood effect on popularity from auto-correlation over the spatial dimension. For example, figure 1 and 2 demonstate two density patterns that would make an analysis of neighborhood popularity without controlling for distance problematic. The Theater District in New York City includes Times Square and many other major tourist attractions. In figure 1, we can see that most of the Airbnb locations in Hell's Kitchen are concentrated next to the Theater Disctrict, meaning that the measured effect on popularity of being in the latter nieghborhood may be inflated because it is highly correlated with the desirability of the theater district. Figure 2, depicts another interesting density pattern. We can see that although there are many Airbnb locations on the outskirts of Chinatown, there are very few in the heart of the neighborhood. This could cause Chinatown's popularity coefficient to be closely correlated with the neighborhoods that surround it. If we observe high popularity in this neighborhood and then advise a new Airbnb owner with this information, they could make a large strategic blunder by placing their property in middle of the neighborhood. This seems problematic!

Therefore, we will account for the density of Airbnb locations so that we can measure the "name brand effect" of neighborhoods on popularity. In order to control for the distance between properties, we will use the conditionally auto-regressive model proposed by Lareaux et al. in "Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence." This Bayesian approach models spatial auto-correlation as a smooth function of distance using an adjacency matrix and a series of priors on distance effects specified in the Appendix.

We determine the distance between neighbours so that every property has a neighbor and the mean number of neighbours is around 30. The correlation between neighbor's spatial effects is determined by the equation below, where ρ is a global parameter that determines the smoothing of the auto-correlation function (aka the magnitude and decay of correlation between neighboring properties).

$$CAR(\phi_k, \phi_j | \phi_{-kj}, W, \rho) = \frac{\rho w_{kj}}{\sqrt{(\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho)(\rho \sum_{i=1}^{k} w_{ki} + 1 - \rho}}$$

Unfortunately, since the CAR model take a very long time to run on large datasets, we weren't able to model this spatial relationship for every observation and ended up running seperate analyses on each of the boroughs with 1000 randomly selected observations (in each borough). Lastly, we also controlled for price and distance to subway when measuring neighborhood effects in this model.

2.5 Global additive model for feature effects

We compare the effect of the different given and computed features of the dataset using a log-linear additive model. Using standard *R* formula representation, our model is therefore

$$\begin{aligned}\log(\texttt{monthly reviews}) \sim & \overbrace{f_1(\texttt{price}) + f_2(\texttt{dist to subway}) + f_3(\texttt{months active})}^{\text{continuous variables}} \\ & + \underbrace{\texttt{entire home/apt} + \texttt{female} + \texttt{round\_price}}_{\text{binary features}} + \underbrace{\texttt{private} + \cdots + \texttt{cozy}}_{\textit{keywords}} \\ & + \texttt{neighbourhoods}\end{aligned}$$

where f_1, f_2 and f_3 are increasing functions, and we assume a normal error distribution. The variables `entire home/apt`, `female` and `round_price` are indicators that the listing is an entire home/apt, that the host is thought to be female, and that the price of the listing is a multiple of 50. The "keyword" variables are indicators that the title contains the words identified by our text analysis, and finally the baseline effect of the neighborhood factor is controlled.

The functions f_1, f_2 and f_3 are Box-Cox transformations of the form $f(x) = (x^\lambda - 1)/\lambda$ and are estimated by maximum likelihood using the `boxTidwell` function of the `car` package. We also tried using monotonicity-constrained splines as implpemented in the shape constrained generalized additive model packages `cgam` and `scam`, but these took too long to run on our machines.

This log-linear additive model, with monotonically transformed continuous variables, provides us with easily interpretable estimates of linear trends.

3. Results

3.1 Neighborhood effects

Figure 3, which depicts the ρ parameter for each borough, clearly indicates that spatial auto-correlation is present in the dataset even after accounting for neighborhood effects. Even though figures 4-7 indicated a lot of diversity in effects and effect sizes based on neighborhoods, there are only a few coefficients whose 95% posterior credible intervals do not include zero. Only Schuresville and Eastchester in the Bronx (figure 4) and Jamaica Hills and East Elmhurst in Queens (figure 6) have a "significant" effect. Since Queens has the highest popularity intercept and East Elmhurt has a significant positive effect on top of this baseline, it seems like this is the best location for a new Airbnb owner to locate his property! However, overwhelming, the effect of neighbourhoods controlling for distance seems to be negligible. This finding may be the result of NYC's highly interconnected and effecient transit system, which enables visitors to travel to major destinations with ease regardless of their starting point.

3.2 Feature effects

The global log-linear additive model provides estimates of trends associating features to listing popularity, after accounting for neighborhood baseline effects and other factors. In the left panel of Figure 8, we see the estimated multiplicative effect of the continuous variables on popularity compared to the median point of the dataset, as a function of the quantiles of this variable. For instance, looking at the price in blue, we see that lower prices correspond to slighly higher popularity (multiplicative effect relative to median > 1), while higher prices are slightly less popular. Here, price and distance to subway have very small effects, while the number of months a listing has been active has a much higher effect on popularity.

On the right of Figure 8, we see that certain keywords (e.g. "stock", "fast", "minute", "walking", "central") are associated with more popular listings, while others ("spacious", "sunny", "apt", "huge") are associated with less popular ones. Notable, round priced listings (price a multiple of 50) are less popular, and female hosts' listings are also associated with a slightly smaller average number of reviews per month.

4. Discussion

Using the CAR model, we were able to show high spatial correlation between the different Airbnb. We also conclusively pointed to the fact that unhears has the highest mean popularity after controlling for spatial correlation. Looking at features of the listings, it is interesting to see that keywords had the most important effect on popularity. This hints at the possibility that the most important features of the listings were in fact what they specifically contained such as the comfort of the space and particular amenities. The dataset only provides limited grasp on this content through the keywords, and any more in-depth analysis might require a more complete dataset.

Due to computational limitations, we did not jointly model the spatial effects, variable selection of keywords, and non-parametric specification of the transformation of the other continuous variables. Furthermore, we focused on estimating individual trends and our results do not provide insights into the interactions between different variables. The first point could be addressed using the specification of a full Bayesian model with Bayesian model selection/averaging. Second, we could have accounted for non-linear relationships in the variables using more complex non-parametric models such as random forests.

Appendix

Proportion of Never Reviewed Listings by Price Decile

CAR Priors

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

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

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

Spatial Auto-correlation and CAR Model

CAR Model Results

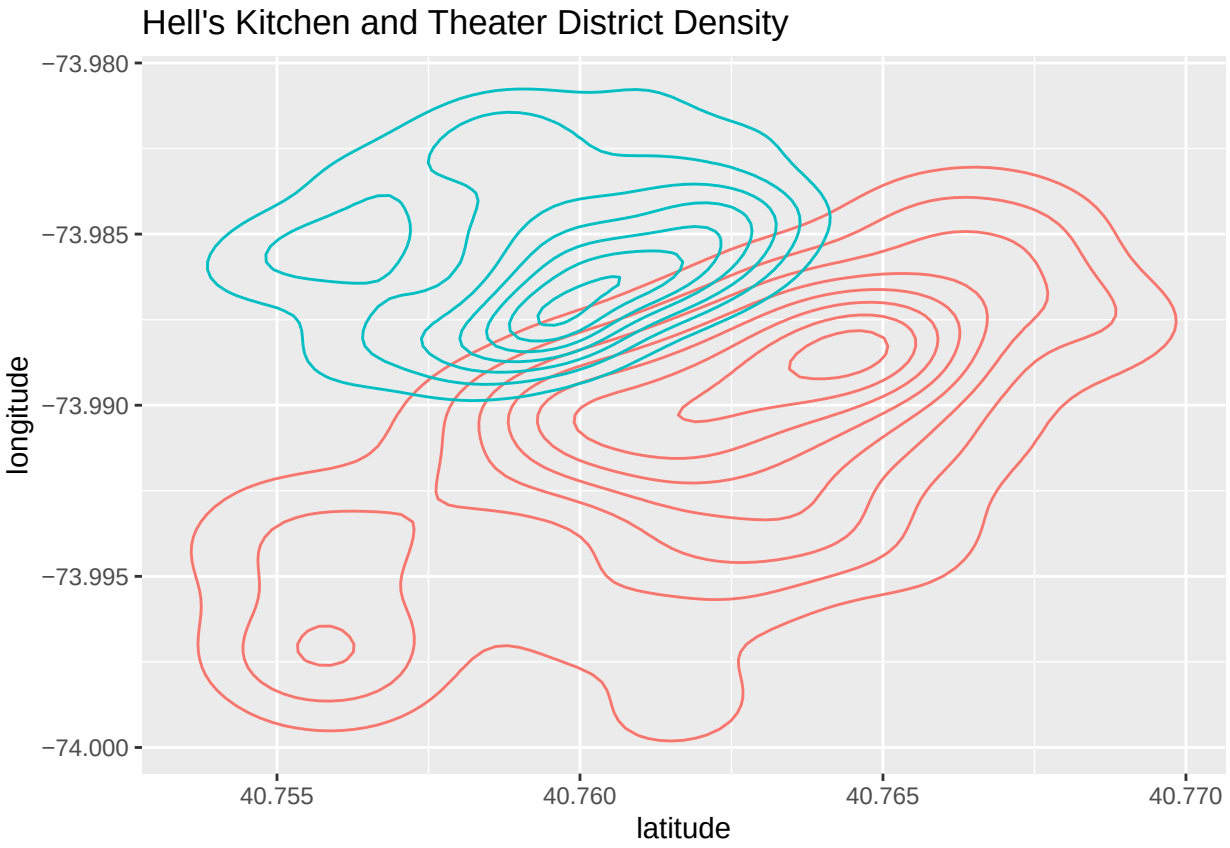

Figure 1: Theater District

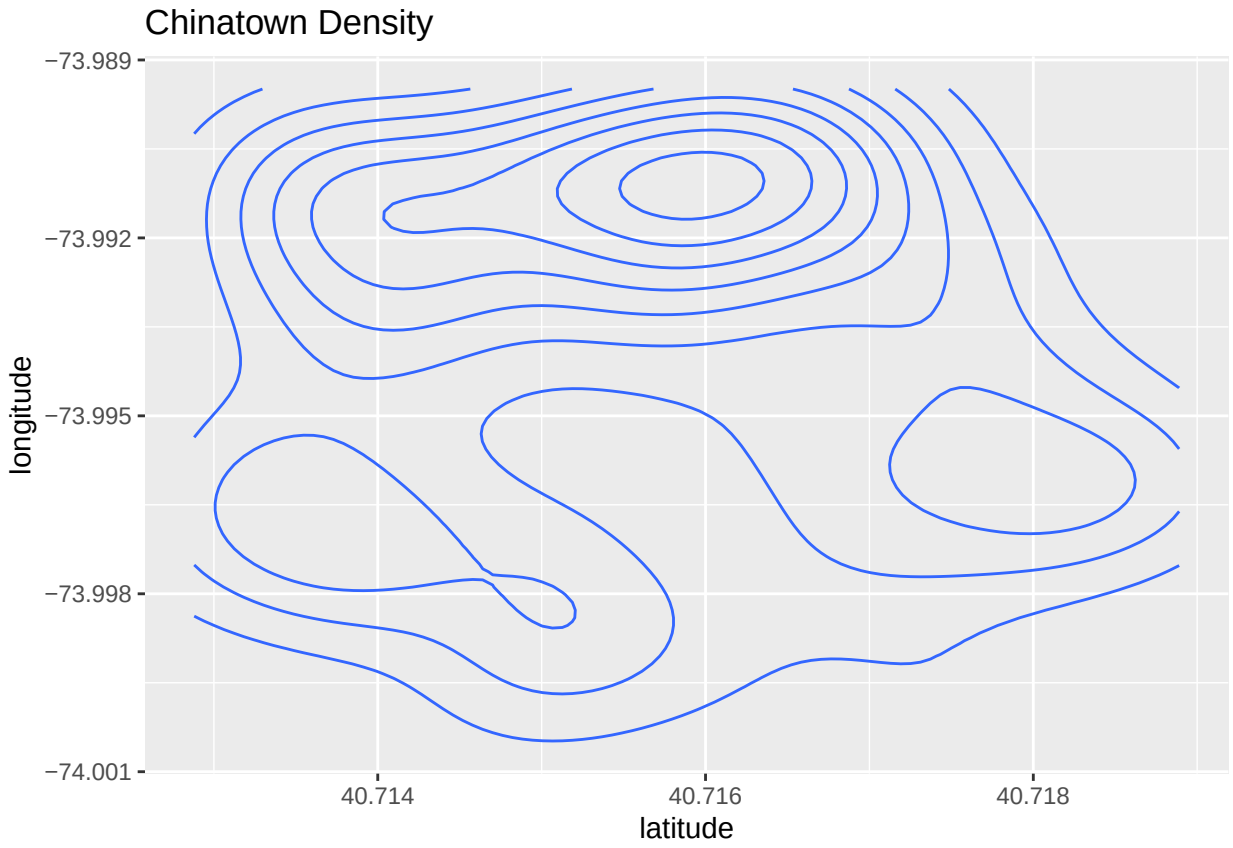

Figure 2: Chinatown

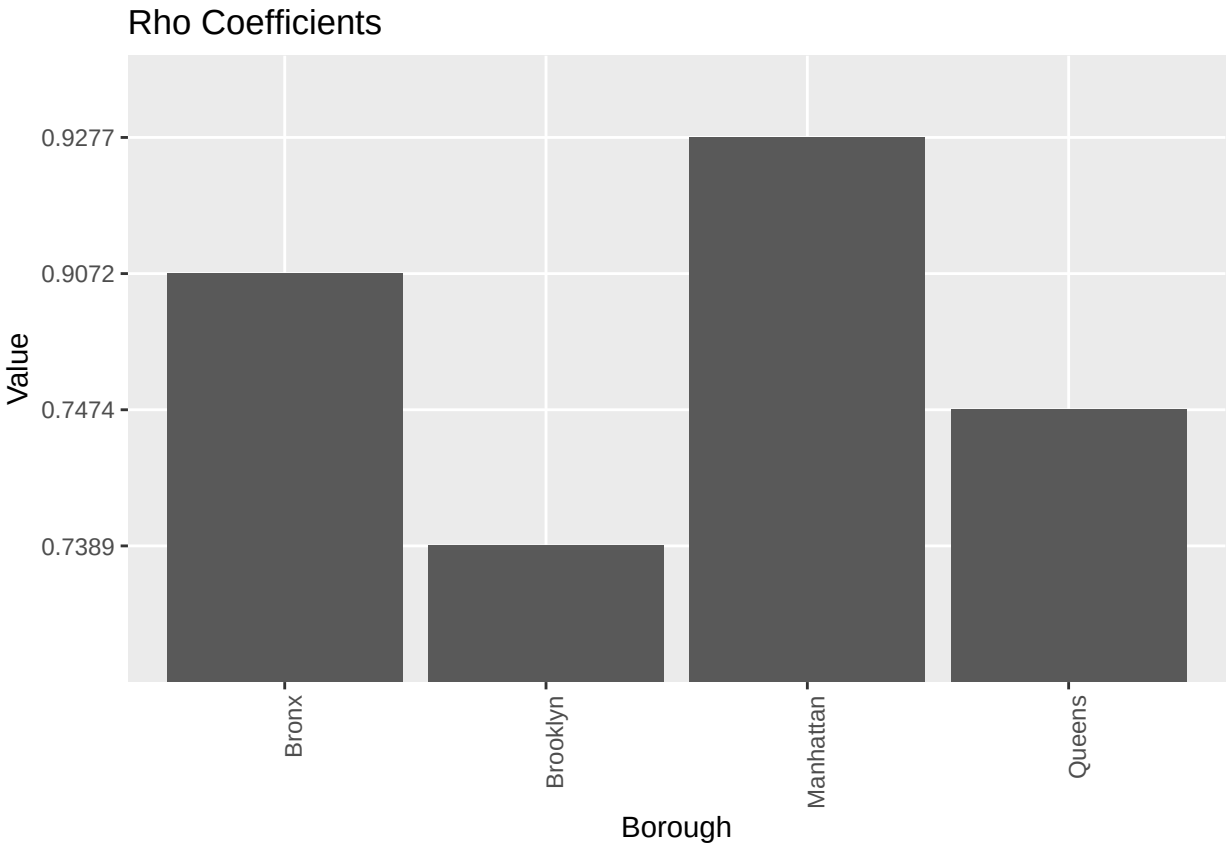

Figure 3: Rho Coefficients

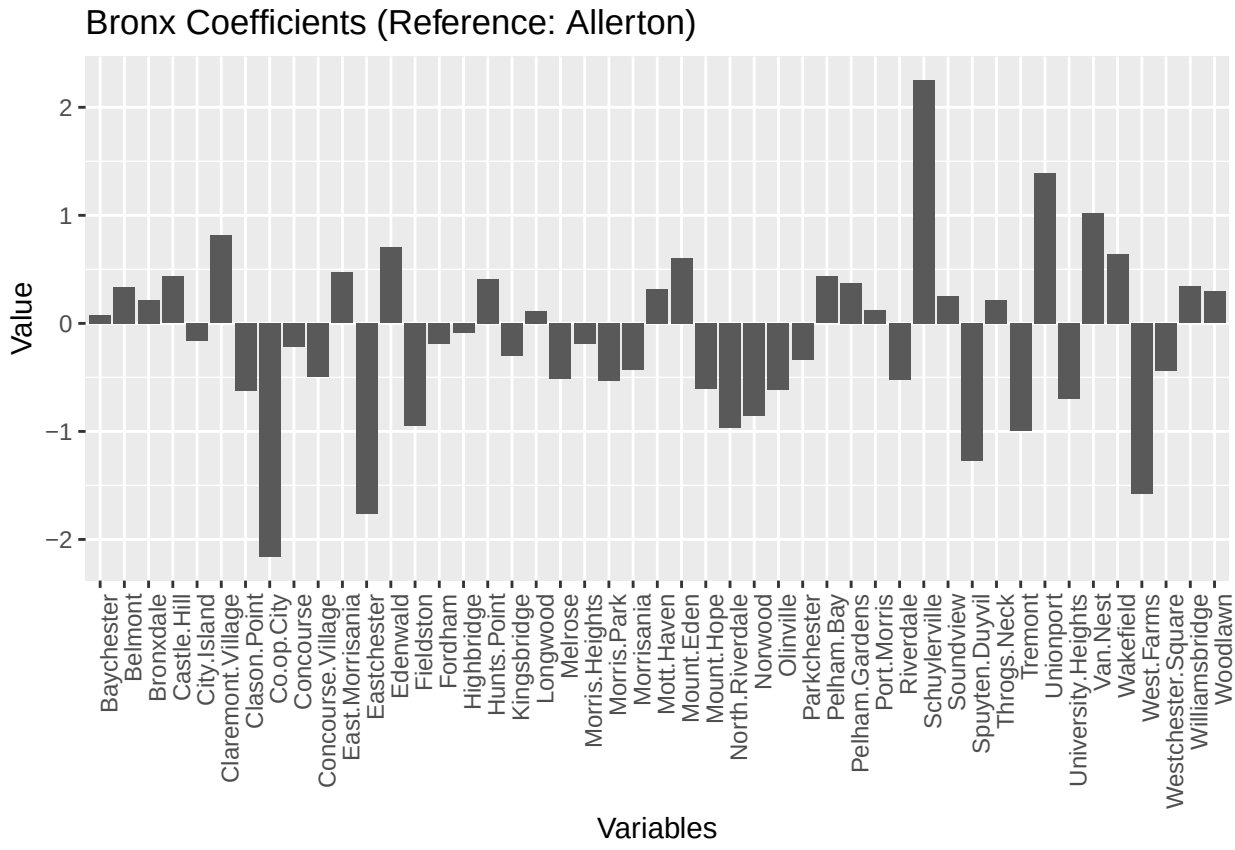

Figure 4: Bronx

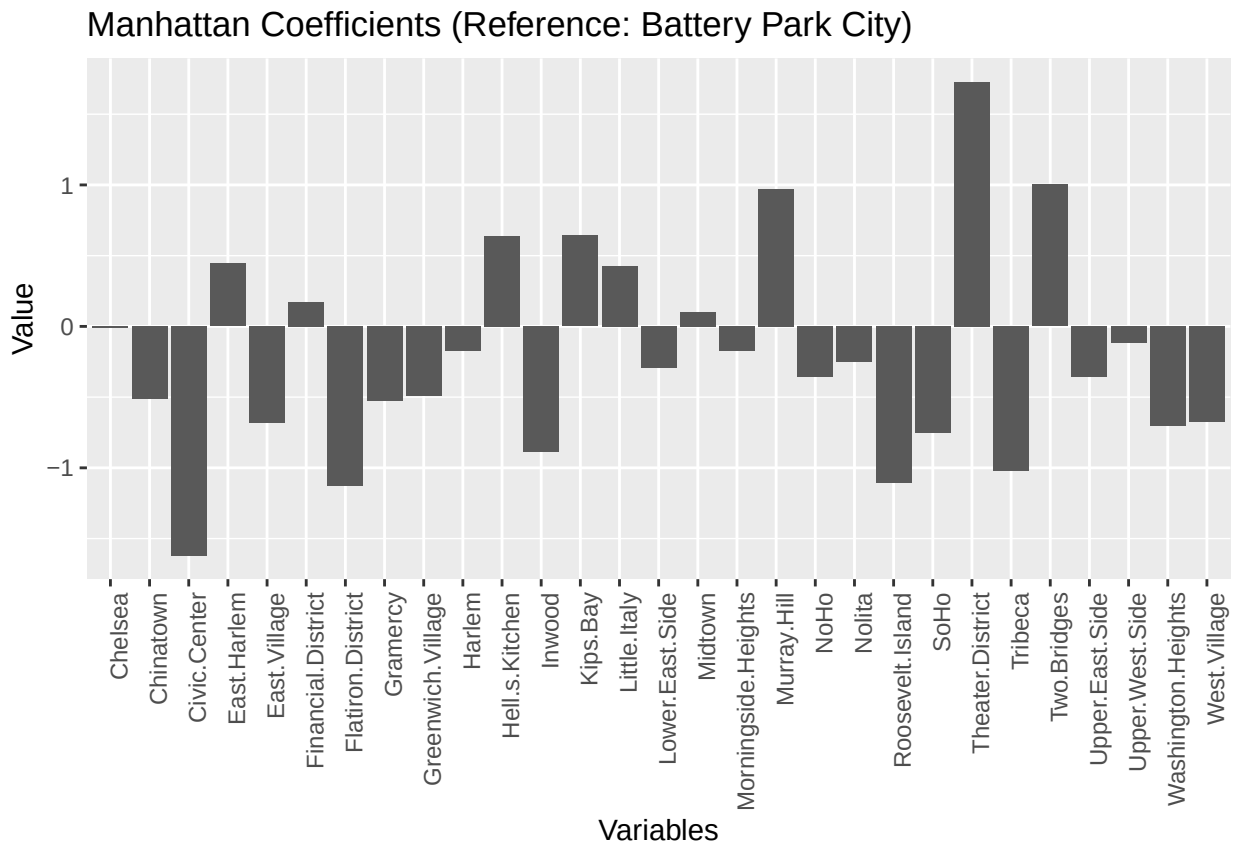

Figure 5: Manhattan

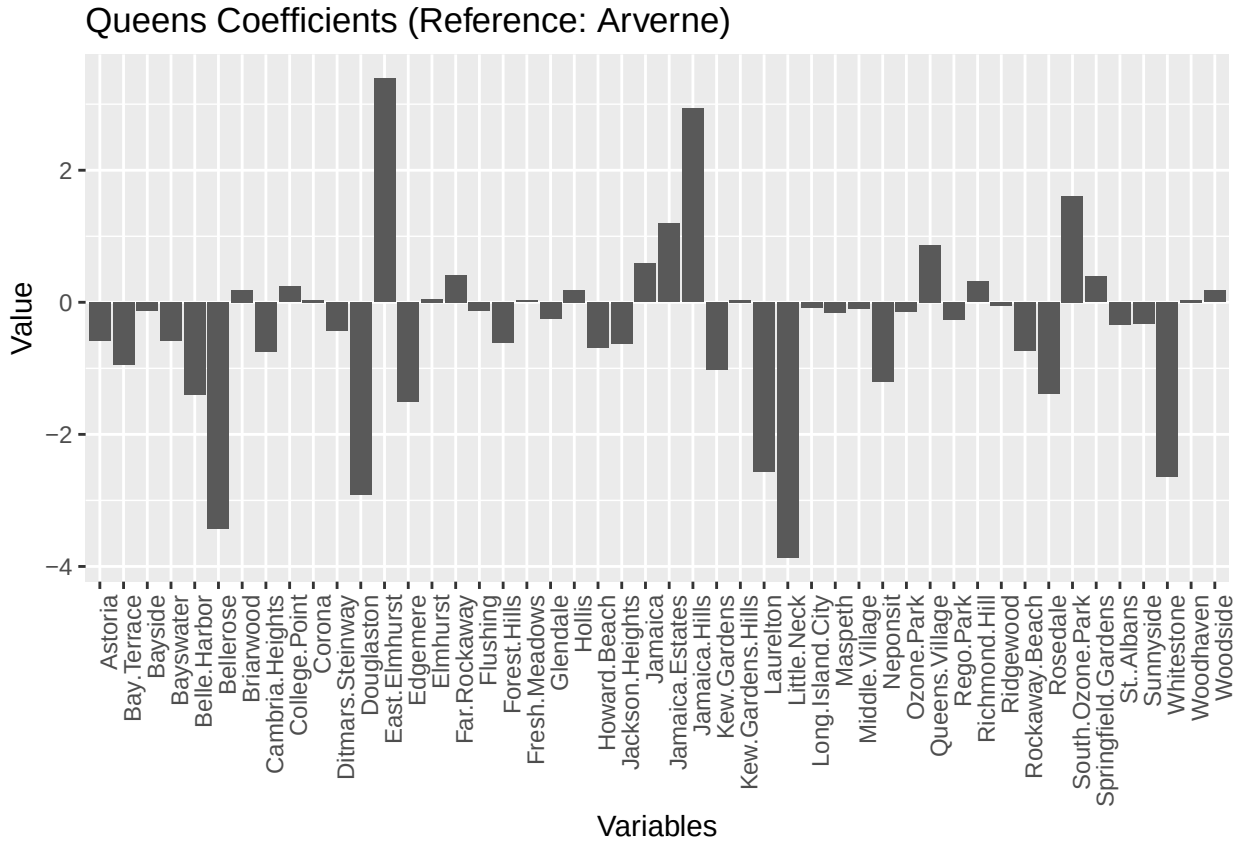

Figure 6: Queens

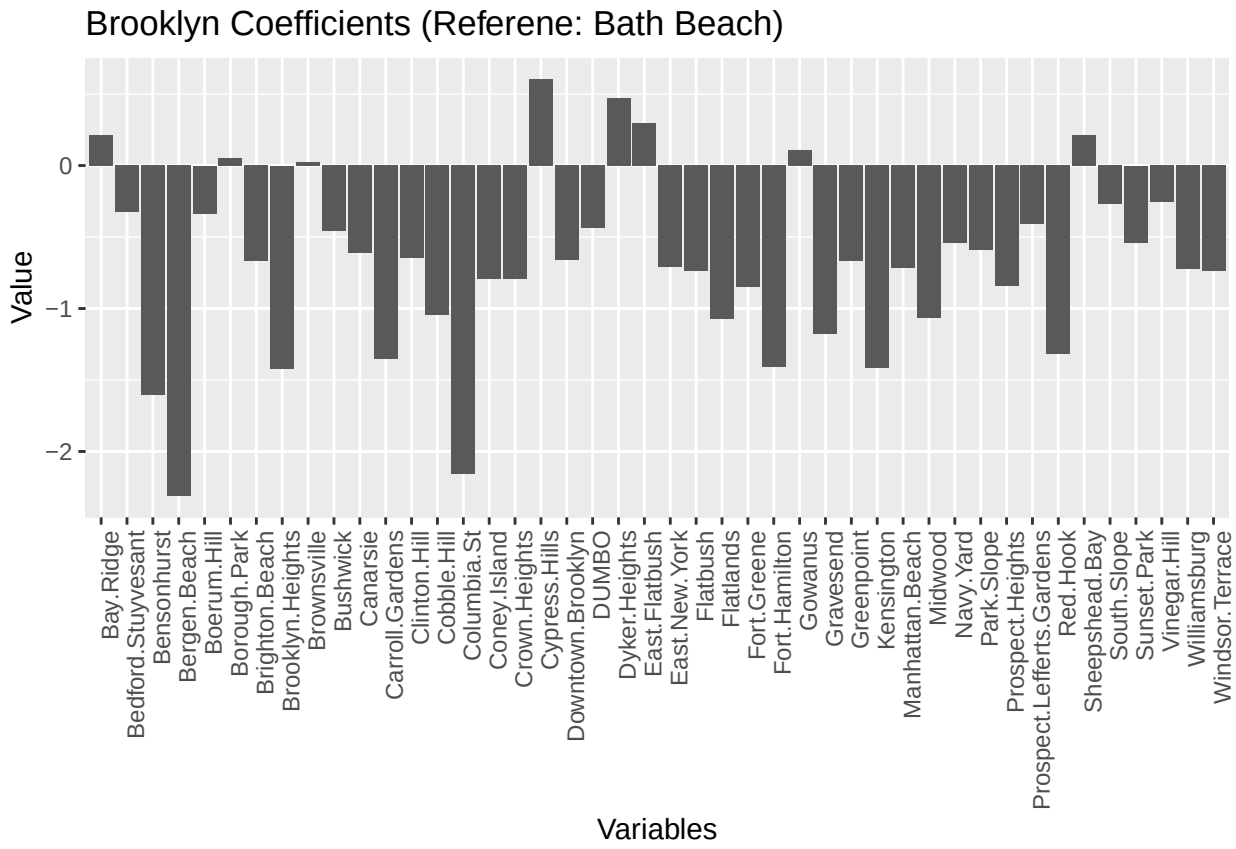

Figure 7: Brooklyn

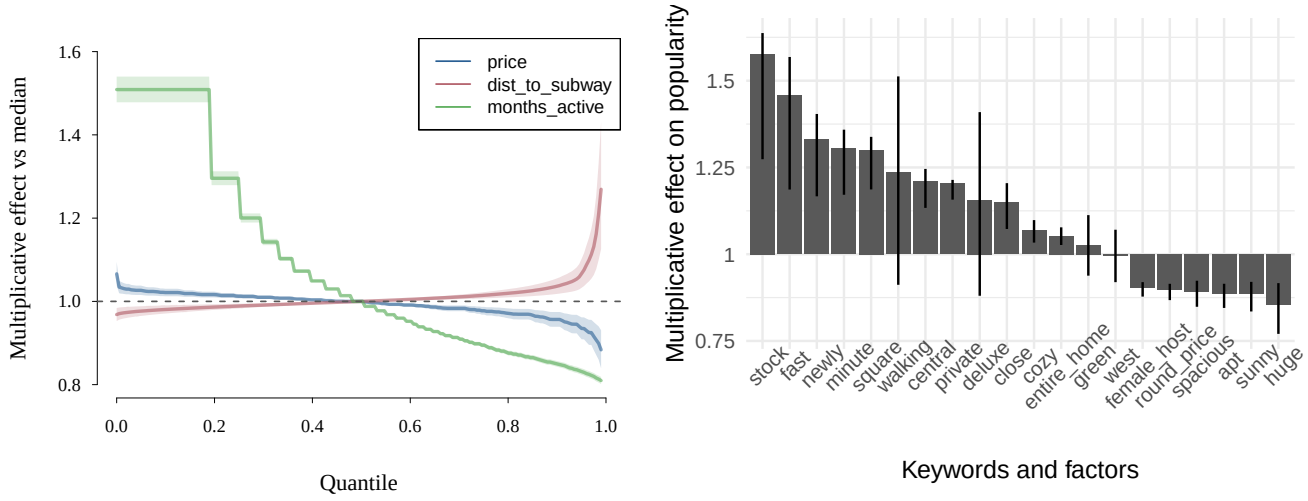

Figure 8: