STA723 Case Study 2 - Group 5 Report

Youngsoo Baek, Phuc Nguyen, Irene Ji

Executive Summary

1. Introduction

Airbnb New York City Open Data collected in 2019 contains information of 48,895 airbnb listings in New York. The dataset has 16 variables, including listing name, location, price, number of reviews, etc. The analysis focuses on identifying most influential factors for price and popularity, examining heterogeneity across boroughs and neighbourhoods, as well as recommending best location and name for airbnb.

2. Materials and Methods

Before fitting our main model (bivariate mixed effect regression), we applied XGBoost to explore the importance of each predictor and conducted Pearson's chi-squared test to examine the heterogeneity across boroughs. Price and monthly reviews were grouped into 3 levels (below Q1, between Q1 and Q3, above Q3) for chi-squared test.

3. Results

3.1 Exploratary Data Analysis and Preprocessing

For data preprocessing, we removed 14 observations with `minimum_nights` greater than a year and imputed missing data by setting missing `price` to 5 and missing `reviews_per_month` to 0. We grouped `last_review` by year and used the difference between last review year and 2019 as our predictor. Since it is unclear what quantity `availability_365` variable is measuring, or how precise a measure it can serve as for whatever quantity inherent to a listing, we created an indicator variable `available_spec` (availability specification) to indicate whether the listing is available (1) or not (0). We took natural logarithm of price and monthly review and use them as metrics of price and popularity respectively.

As shown in Figures 1 and 2, variable importance plots of XGBoost suggest that room type, availability, monthly reviews and boroughs are the most influential factors for price of airbnb. Last review year, availability, minimum nights and price are most important for monthly reviews.

Figures 3 to 5 present boxplots and mosaic plots for price, monthly review and room type across boroughs. As shown in the plots, price, popularity and room type differ across boroughs. For example, Manhanttan has the highest price and most listings there are entire homes or apartments. Queens, on the other hand, has the highest popularity and private rooms take up the most listings

there. This agrees with the chi-squared test results, where the tests are all significant with p-values $<$ 2e-16, suggesting there exists heterogeneity across boroughs.

3.2 Main Results

...Assuming log price and log number of monthly reviews are both isotropic and intrinsically stationary processes, we can define its semivariogram as

$$\gamma(||\mathbf{h}||) \equiv \frac{1}{2}\mathrm{E}[Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})].$$

The customary, simple nonparametric estimator for semivariogram is (BCG, 2011)

$$\hat{\gamma}(d) = \frac{1}{2|N(d)|} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(d)} [Y(\mathbf{s}_i) - Y(\mathbf{s}_j)]^2,$$

where $N(d)$ consists of all pairs of locations that have a pairwise Euclidean distance d.

...It is clear that for monthly review rates, we observe a negative spatial autocorrelation: closer things have more different response rather than similar. This must raise alarming sign to possible efforts to apply traditional spatial models for this response, as naive conditional autogressive models assumes positive spatial autocorrelation.

4. Discussion

References

Sarthak, N., Post: "Availability_365=0?", Discussion thread: New York City Airbnb Open Data, Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/discussion/111835

Banrjee, S., Carlin, B. P., and Gelfand, A. E. (2011). *Hierarchical Modeling and Analysis for Spatial Data.*

Appendix: Figures and Tables

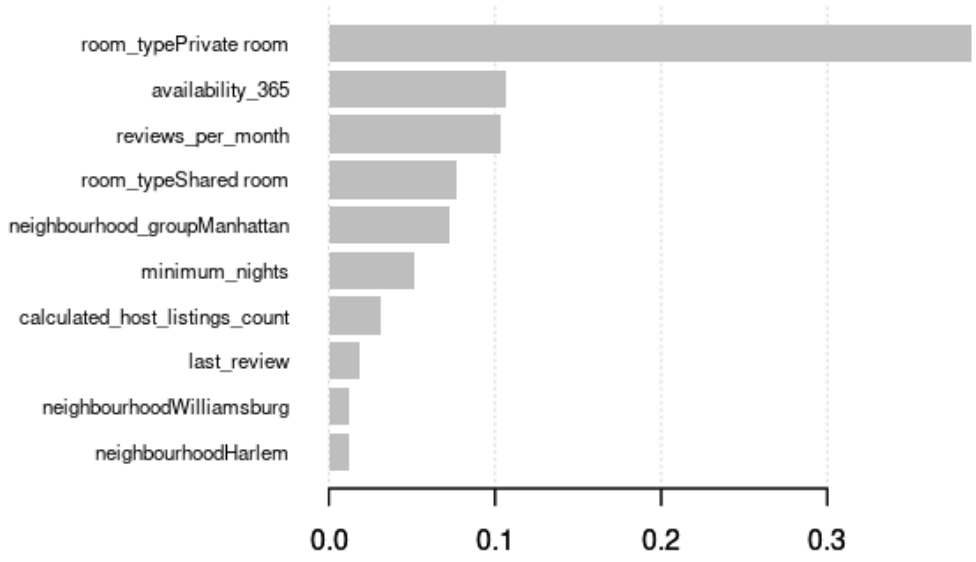

Figure 1. XGBoost − Variable Importance Plots for Price

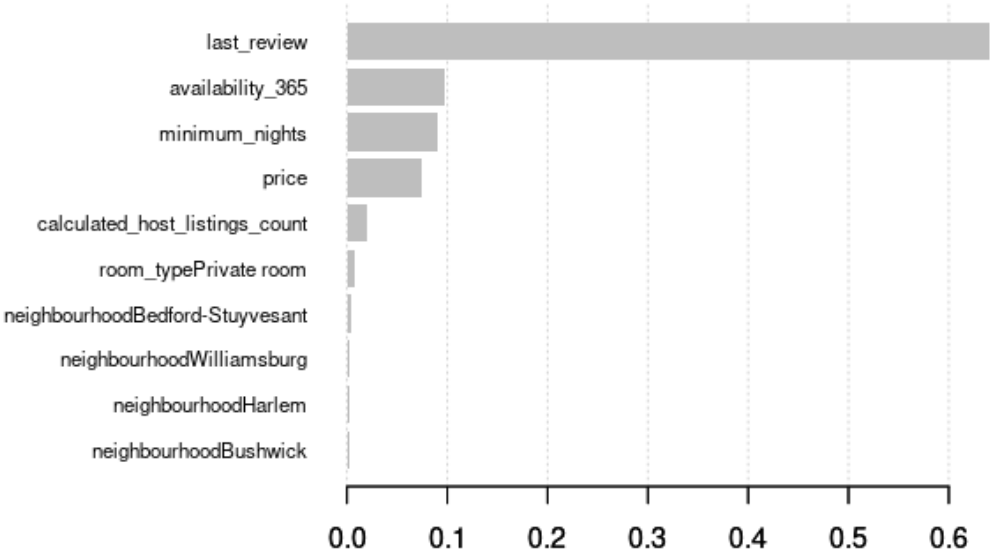

Figure 2. XGBoost − Variable Importance Plots for Monthly Review

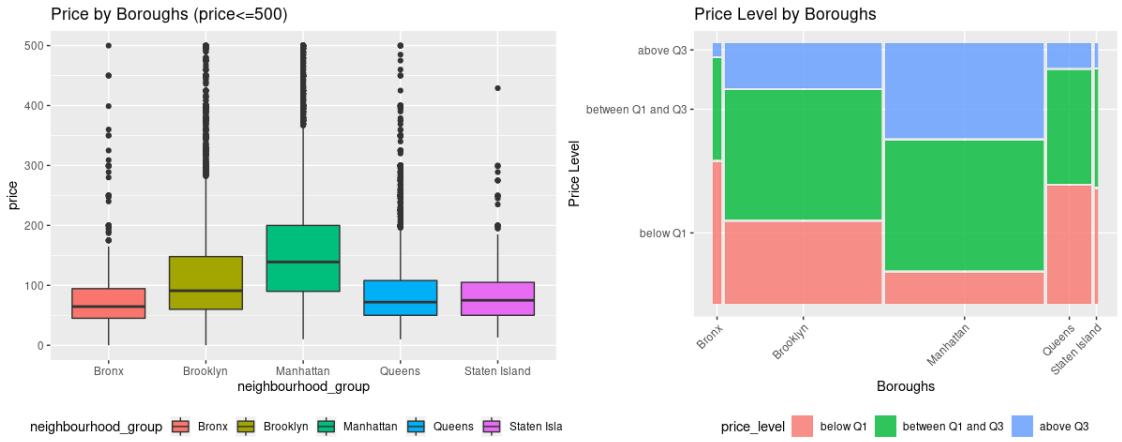

Figure 3. Boxplot and Mosaic Plot for Price across Boroughs

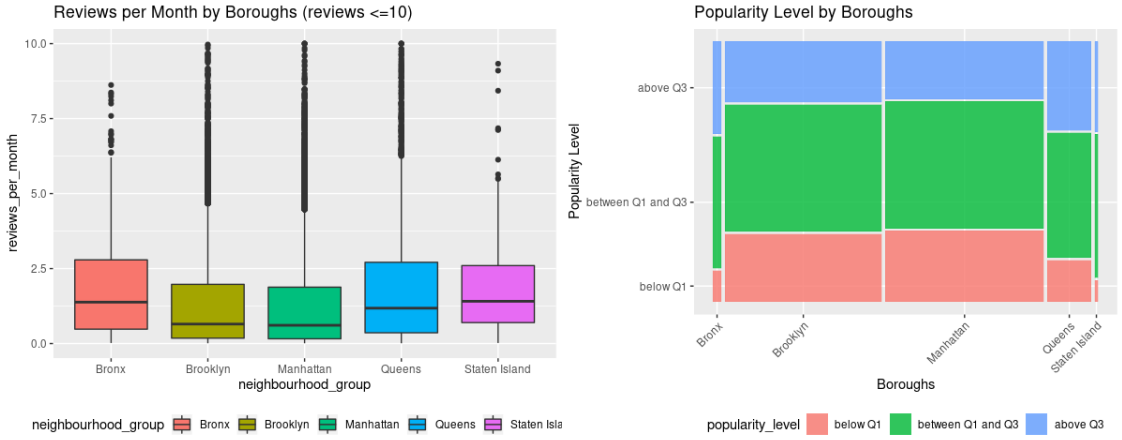

Figure 4. Boxplot and Mosaic Plot for Monthly Review across Boroughs

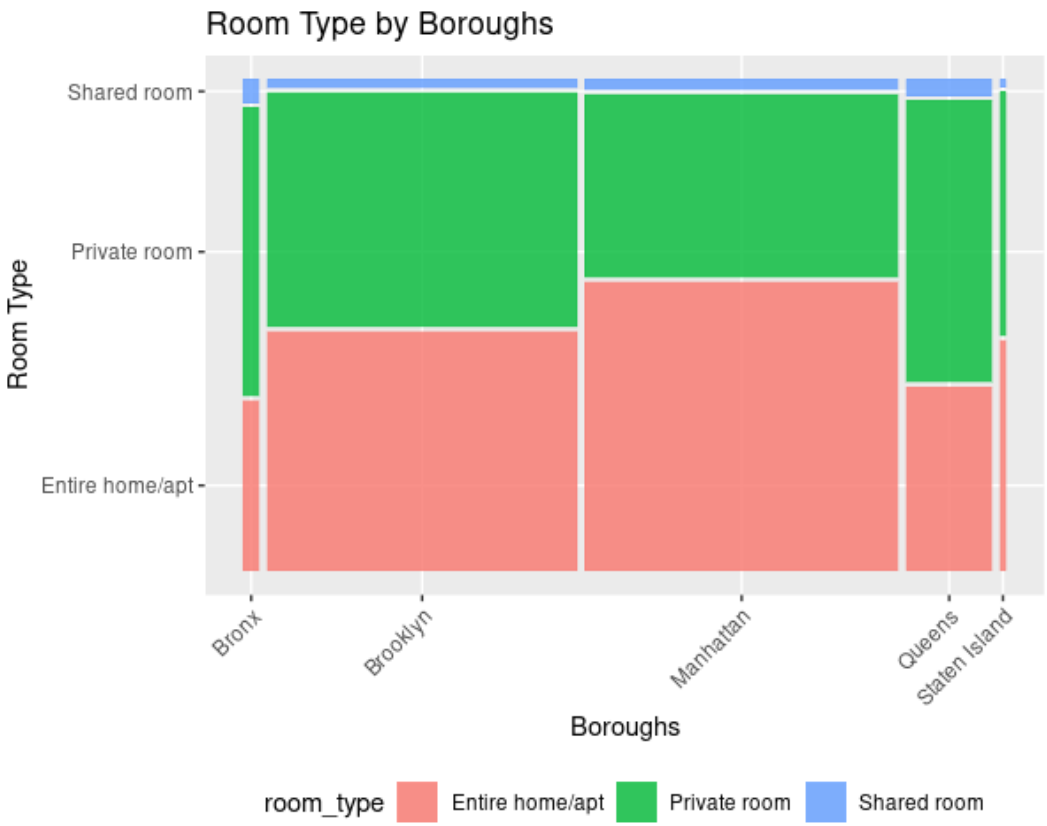

Figure 5. Mosaic Plot for Room Type across Boroughs