Exploratory Analysis of Data for Airbnb Listings in NYC

Youngsoo Baek, Irene Yi Ji, Phuc Nguyen

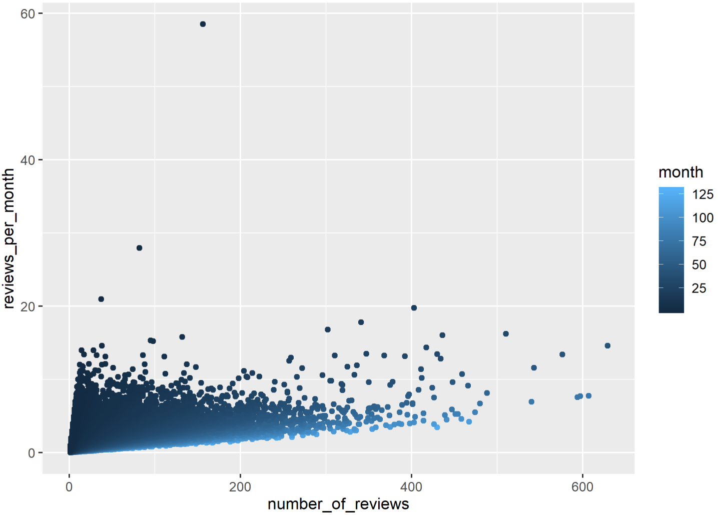
Introduction

- Data: Airbnb New York City open data collected from 2019, with 48,895 listings and 16 variables.
- Goals:
  - Identify most influential factors for price/popularity
  - Examine heterogeneity across boroughs and neighbourhoods
  - Recommend best location and name for airbnb

Data Processing

- Remove 14 observations with *minimum_nights* > 365
- *Price*: the lowest non-zero value is 10, added 5 to 0's
- *Reviews per Month*: missing values are set to 0 (last review dates are missing and total number of reviews are 0)
- *Last Review*: group by years from 2019 (e.g. 2019 -$>$ 0; 2018 -$>$ 1, etc.)
- *availability_365*: create a new variable *available_spec* to indicate whether the value is 0

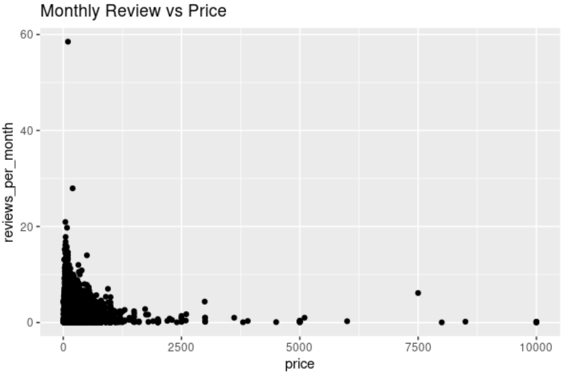
What is a Valid Metric for Popularity?



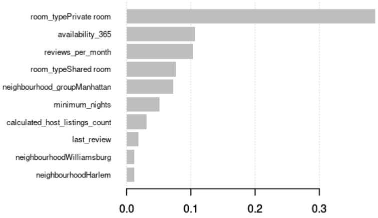
- **Monthly reviews** adjusts for the history of a listing (albeit not perfectly)

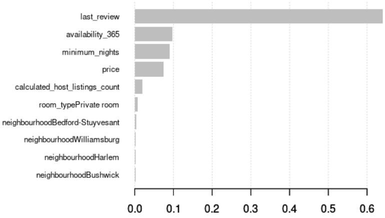
EDA - Price and Popularity

- Price and popularity seem to be negatively correlated (with extreme values):



XGBoost for Important Variables



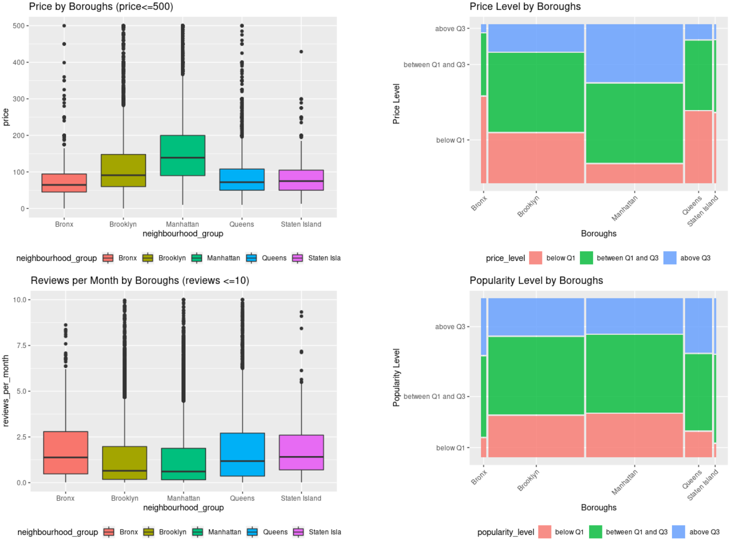


- The most influential factors for price include: room type, availability, monthly reviews, boroughs, etc.
- The most influential factors for popularity include: last review, availability, price, etc.
- Price and popularity are closely related, both being an important variable of the other. We may consider model them as bivariate reponse.

Heterogeneity of Price / Popularity across Boroughs

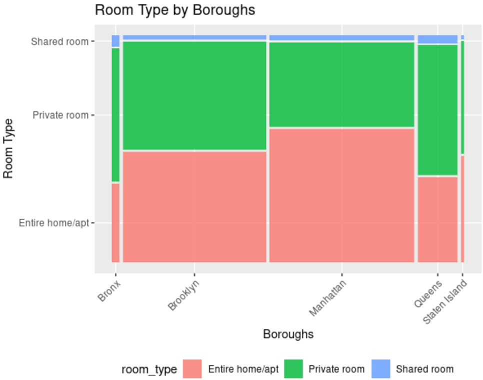
- Create new variables “Price Level” and “Popularity Level”:
  - “Low” for values $<$ 25th Percentile
  - “Medium” for values between 25th and 75th Percentile
  - “High” for values $>$ 75th Percentile
- Create contingency table and conduct Chi-squared Test for Homogeneity

Heterogeneity of Price / Popularity across Boroughs



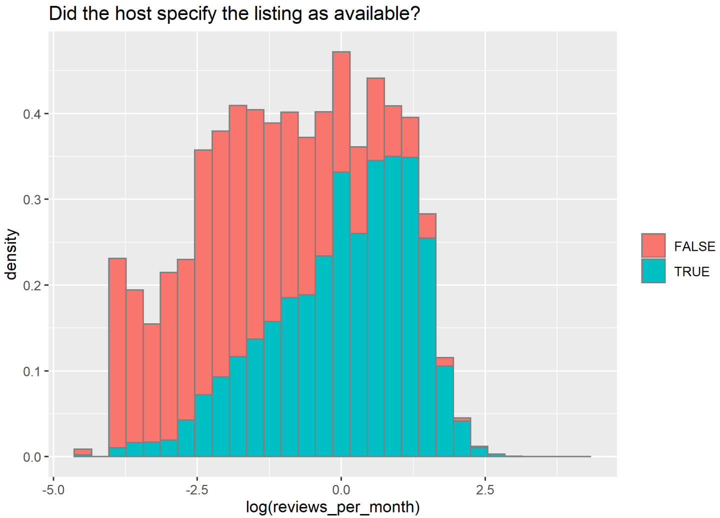
- Small p-value suggests heterogeneity across boroughs.

Heterogeneity of Room type across Boroughs



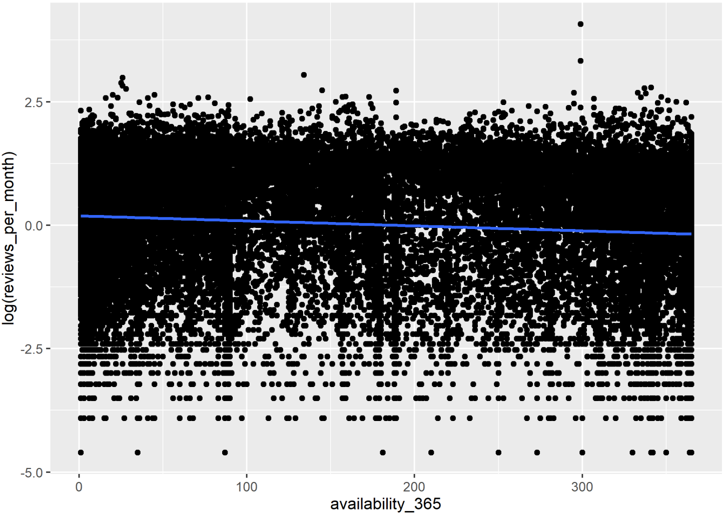
- Small p-value suggests heterogeneity across boroughs.

Unreliability of Availability Feature



On average, it seems the listings that are "temporarily unavailable" (zero availability) have lower monthly review rate. . .

Unreliability of Availability Feature



. . . but *conditioned on* non-zero availability, the association is less obvious (can be negative?).

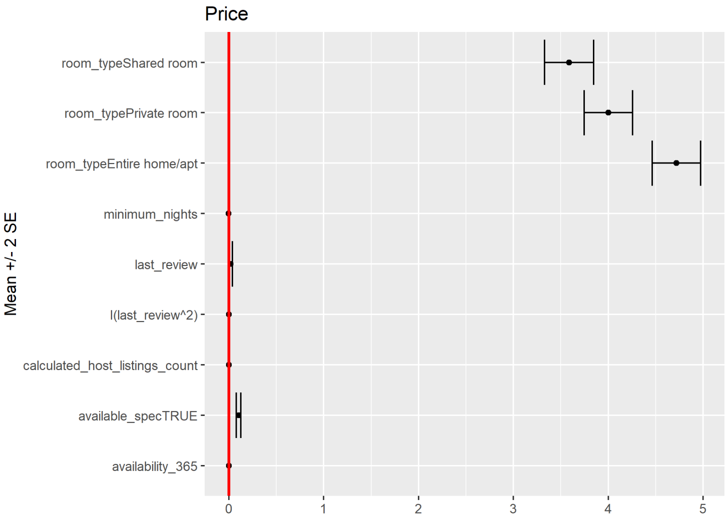
Modeling: Bivariate Mixed Effects Regression

- Varying intercept model: For the i-th listing in neighborhood j, within borough k,

$$\begin{pmatrix} \text{Price}_{k[j[i]]} \\ \text{Monthly review}_{k[j[i]]} \end{pmatrix} = \begin{pmatrix} \beta_1^T \mathbf{X}_i \\ \beta_2^T \mathbf{X}_i \end{pmatrix} + \boldsymbol{\eta}_{k[j]} + \boldsymbol{\theta}_j + \boldsymbol{\epsilon}_{k[j[i]]}.$$

- Both "availability specification" and raw availability count are included as predictors
- Quadratic term of the listing's age is included
- Observations with no reviews excluded (21% of the data)

What Are the Important Predictors for Price?



- Many predictors are significant, but **room type** only seems to be associated to large enough increase in price

What Are the Important Predictors for Popularity?

- The younger the listing is, the more it is popular on average (in spite of significance of the quadratic term)

Estimates for Group Heterogeneities

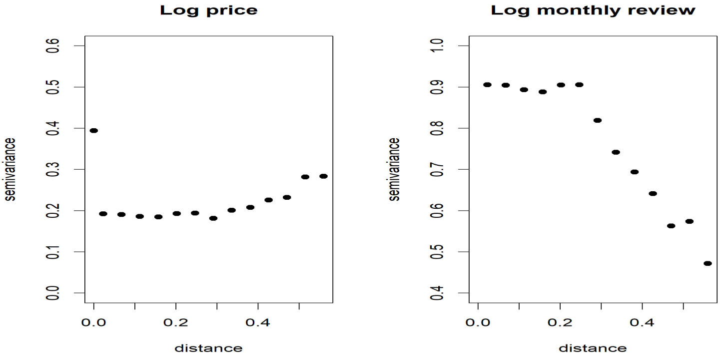
variableprice	variablereviews_per_month
0.03	-0.01
-0.01	0.04

variableprice	variablereviews_per_month
0.08	-0.02
-0.02	0.01

- Many significant coefficients can be swamped by the variability within/between different neighborhoods and boroughs
- Strong negative correlation between two random intercepts between boroughs (-0.76)

Examining Spatial Correlation of the Residuals

- Semivariograms: For location $\mathbf{s}_i$, estimate $\mathrm{Var}(Y(\mathbf{s}_i + d) - Y(\mathbf{s}_i))$ in increasing distance d.



- We observe large semivariogram for price when listings are extremely close, and **negative spatial correlation** for monthly review rates

Possible Insights

- When two listings are very close (identical coordinates), the market effect takes sway over all others. One potential customer is being sapped away from one listing to another.
- As a result, closer things have more dissimilar popularity measures. As distance increases, however, the effect becomes less severe and association between a listing's features and sales becomes noticeable.
- However, price is relatively "inelastic"; unless two listings are extremely close to each other, the hosts' pricing policy remains relatively indifferent to their neighbors, adjusted for other features of a listing.
- Hence, we observe no evidence of spatial correlation, conditional on what neighborhood a listing belongs to, except in extreme proximity (high semivariogram).

Text Analysis for Listing Names

(. . . Phuc's analysis. . .)

Limitations and Further Work

- Including varying slopes calls for strong shrinkage
- Care is needed for spatial covariance models: "soft" adjacency matrix for neighborhoods/boroughs, negative autocorrelation, etc.
- Missing data/latent space model for `availability_365`
- Nonparametric approach for bivariate model