Exploratory Analysis of Data for Airbnb Listings in NYC

Youngsoo Baek, Irene Yi Ji, Phuc Nguyen

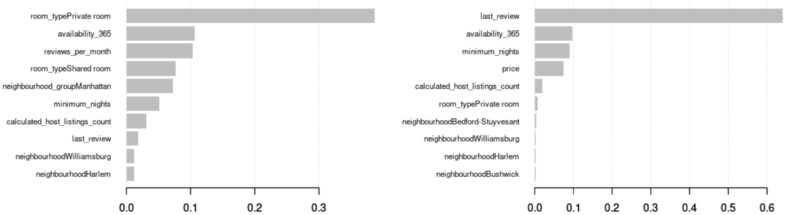
Data Processing

- Remove 14 observations with *minimum_nights* > 365
- *Price*: the lowest non-zero value is 10, change 0 to 5
- *Reviews per Month*: missing values are set to 0 (there is no review for these listings)
- *Last Review*: group by years from 2019 (e.g. 2019 -$>$ 0; 2018 -$>$ 1, etc.)

Response Variables - Price and Popularity

- Metric for price: **price**
- Metric for popularity: **monthly reviews** adjusts for the history of a listing (albeit not perfectly)
- Price and popularity seem to be negatively correlated (in log scale)

XGBoost for Important Variables

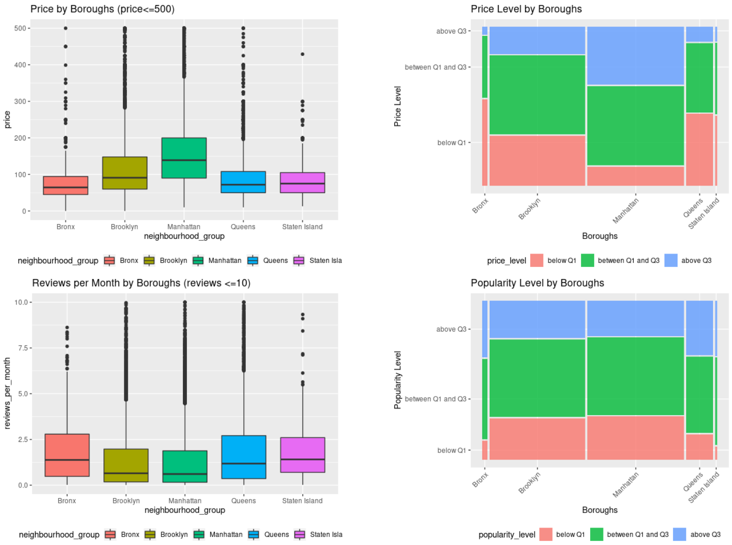


- Most important variable for price: *Room Type*.
- Most important variable for popularity: *Last Review* (age of the listings).
- Price and popularity are closely related, both being an important variable of the other. We may consider model them as bivariate reponse.

Heterogeneity of Price / Popularity across Boroughs

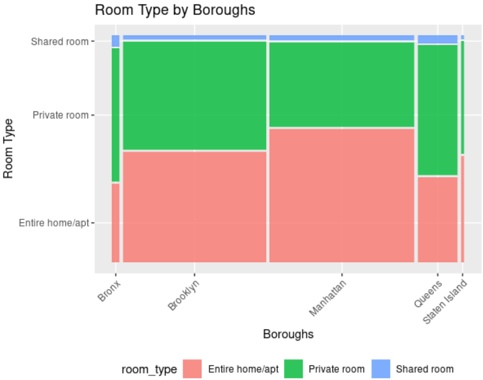
- Create new variables “Price Level” and “Popularity Level”:
  - “Low” for values $<$ 25th Percentile
  - “Medium” for values between 25th and 75th Percentile
  - “High” for values $>$ 75th Percentile
- Create contingency table and conduct Chi-squared Test for Homogeneity

Heterogeneity of Price / Popularity across Boroughs



- Small p-value suggests heterogeneity across boroughs.

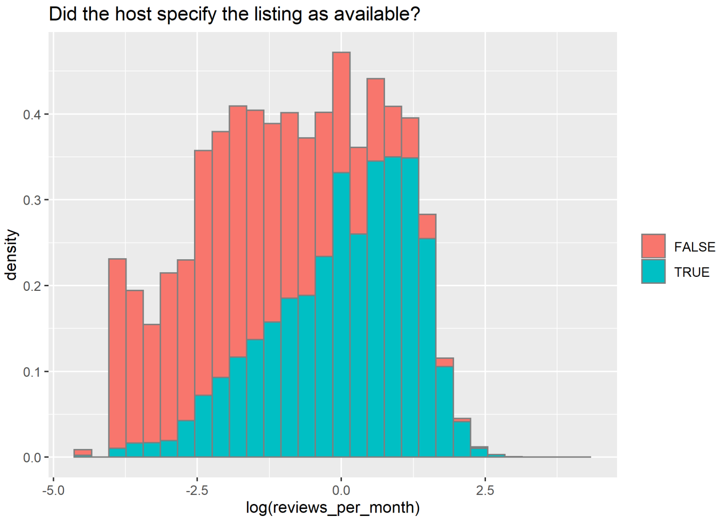
Heterogeneity of Room type across Boroughs



- Small p-value suggests heterogeneity across boroughs.

Unreliability of Availability Feature

- `availability_365`: Only important info. seems to come from whether it is zero or not



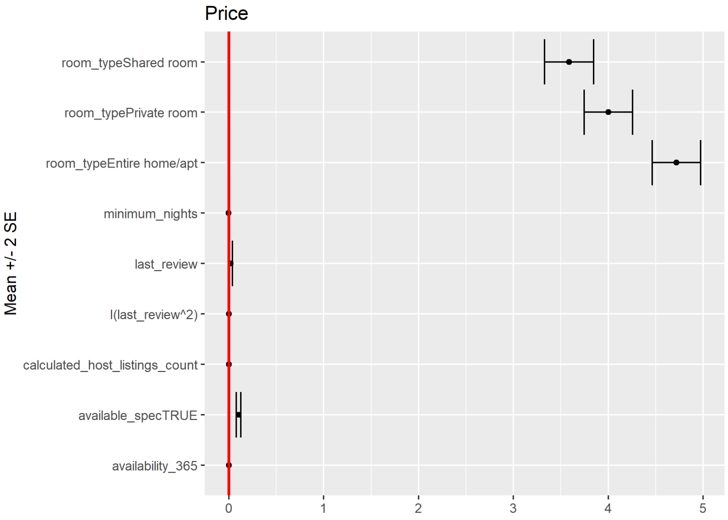
Modeling: Bivariate Mixed Effects Regression

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

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

- Availability is included both as an non-zero indicator and numeric variables
- 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?



- Apartment > Pvt room > Shared room in price

What Are the Important Predictors for Popularity?

- To have reviews, a listing on average should be young and actually on business

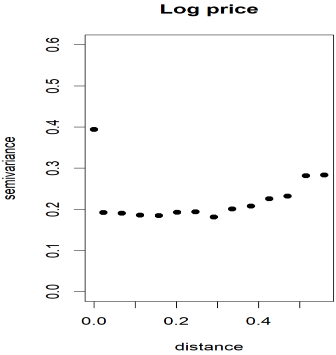
Random Intercepts for Groups

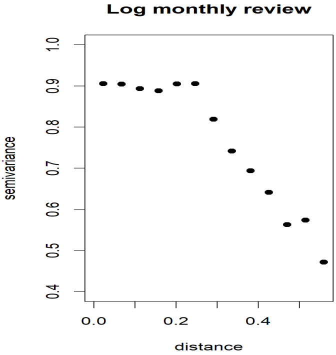
	variableprice	variablereviews_per_month
Bronx	-0.21	0.05
Brooklyn	0.07	-0.12
Manhattan	0.45	-0.11
Queens	-0.10	0.11
Staten Island	-0.22	0.06

- Manhattan most expensive, Queens most popular
- Strong negative correlation between two random intercepts between boroughs (-0.76)
- Neighborhood-level variation is relatively minute

Examining Spatial Correlation of the Residuals

- Plotting semivariogram as evidence of within-neighborhood spatial structure



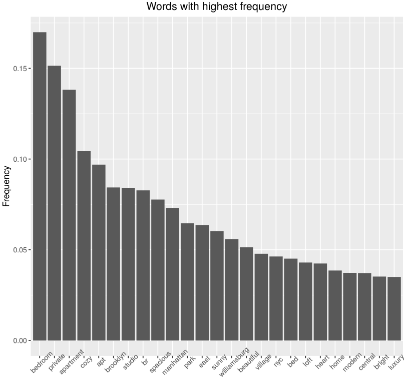


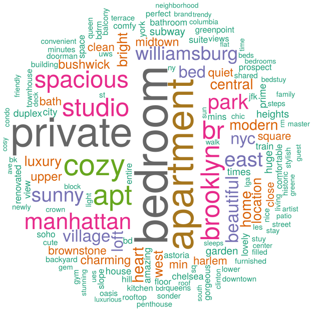
- We observe large semivariogram for price when listings are extremely close, and **negative spatial correlation** for monthly review rates

Possible Insights

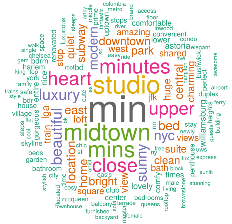
- When two listings are very close (identical coordinates), the market effect takes sway over all others. One potential customer is being sapped away from one listing to another.
- As a result, closer things have more dissimilar popularity measures. As distance increases, however, the effect becomes 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).

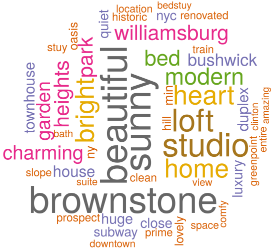
Text Analysis for Listing Names

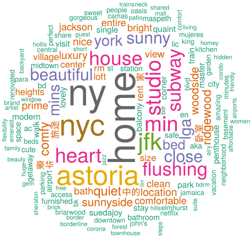




Text Analysis for Listing Names







Foreign language, Special Characters, and Misspelling

“WilliamsburgBrooklynPrivateBedroom” *“NiceRoom-NiceNeighborhoodCloseMaimonidesHospital”*

Conclusions

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