Exploratory Analysis of Data for Airbnb Listings in NYC

Youngsoo Baek, Irene Yi Ji, Phuc Nguyen

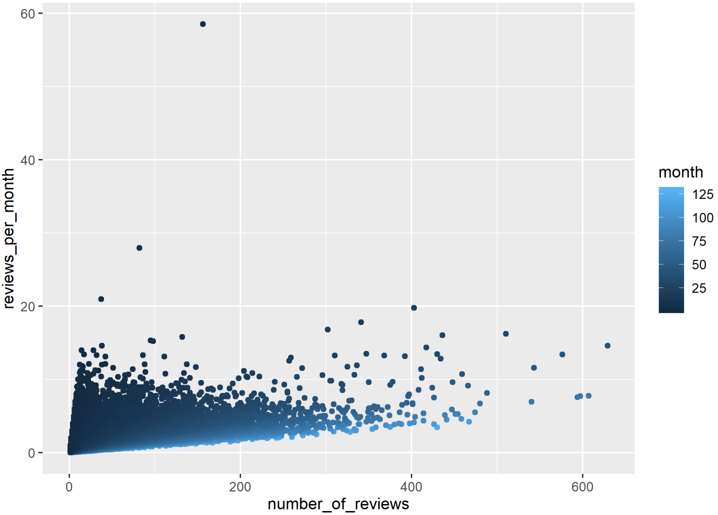
Introduction

- 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

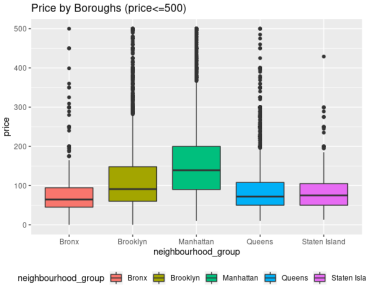
- Remove 14 observations with *minimum_nights* > 365
- *Price*: the lowest non-zero value is 10, added 5 to 0's
- *Reviews per Month*: missing values are set to 0 (last review dates are missing and total number of reviews are 0)
- *Last Review*: group by years from 2019 (e.g. 2019 -$>$ 0; 2018 -$>$ 1,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

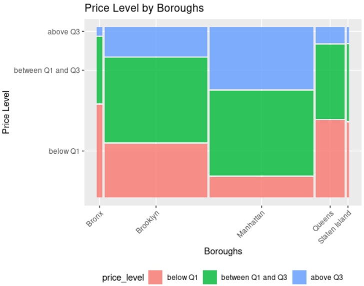
What is a Valid Metric for Popularity?



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

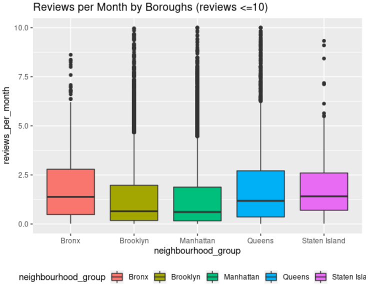
Heterogeneity of Price across Boroughs

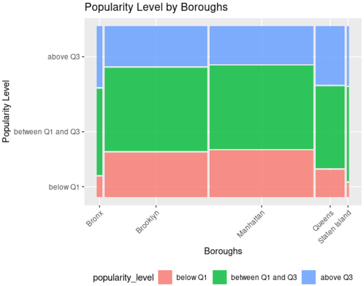




- Generate 3 price levels:
  "below Q1", "between Q1 and Q3", "above Q3"
- Pearson's Chi-squared test: p-value $< 2.2\text{e-}16$

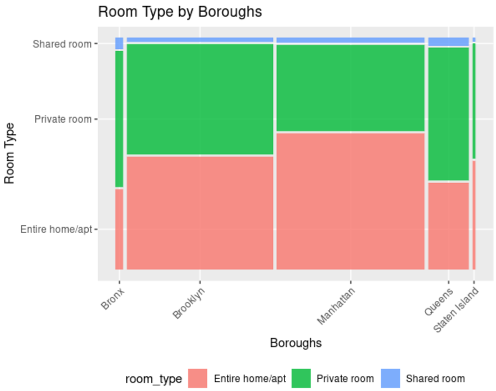
Heterogeneity of Popularity across Boroughs





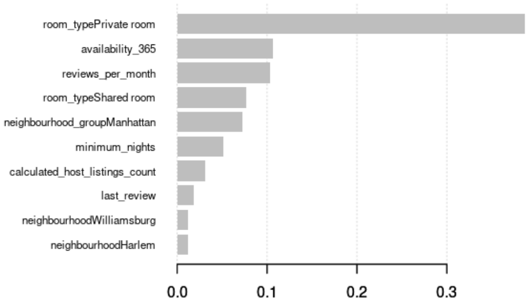
- Generate 3 popularity levels:
  "below Q1", "between Q1 and Q3", "above Q3"
- Pearson's Chi-squared test: p-value $< 2.2e\text{-}16$

Heterogeneity of Room type across Boroughs



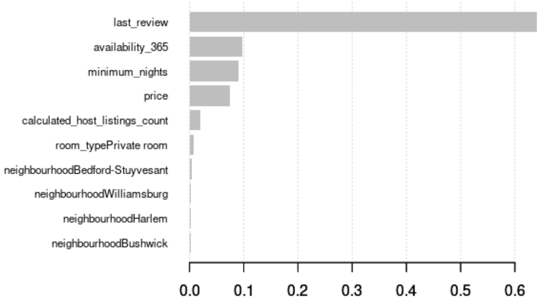
- Pearson's Chi-squared test: p-value $< 2.2e\text{-}16$

Price: XGBoost for Important Variables



- The most influential factors for price of airbnb include: room type (private room), availability, monthly reviews, boroughs (Manhattan), etc.

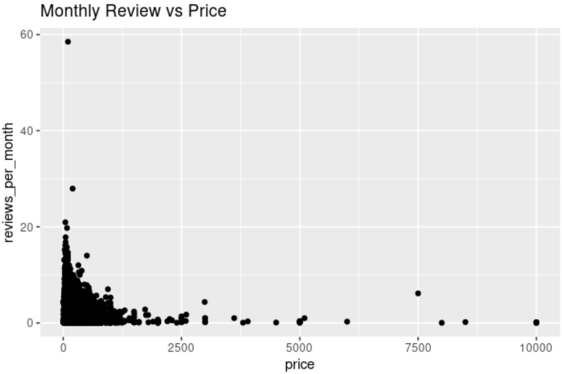
Popularity: XGBoost for Important Variables



- The most influential factors for popularity of airbnb include: last review (in years from 2019), availability, minimum nights, price, etc.

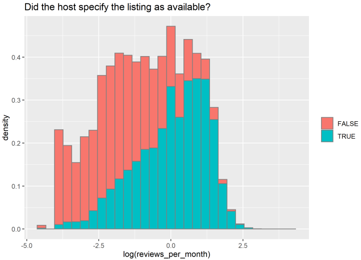
EDA - Price and Popularity

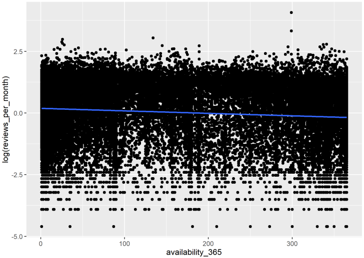
- From XGBoost outputs, price and popularity are closely related, both being an important variable of the other.
- The plot below shows a negative correlation between them on *log-scale*:



- We may consider model them as bivariate reponse.

Possibly Unreliable Predictors





Modeling: Bivariate Mixed Effects Regression

- Mixed effect (to better understand heterogeneities) + Joint model for price and populairty (to better understand their negative correlation)
- Group-varying intercept model

$$\begin{pmatrix} \text{Log price} \\ \text{Log monthly review} \end{pmatrix} \sim$$

$$1 + 1|\text{Borough:Neighborhood} + 1|\text{Borough}+$$

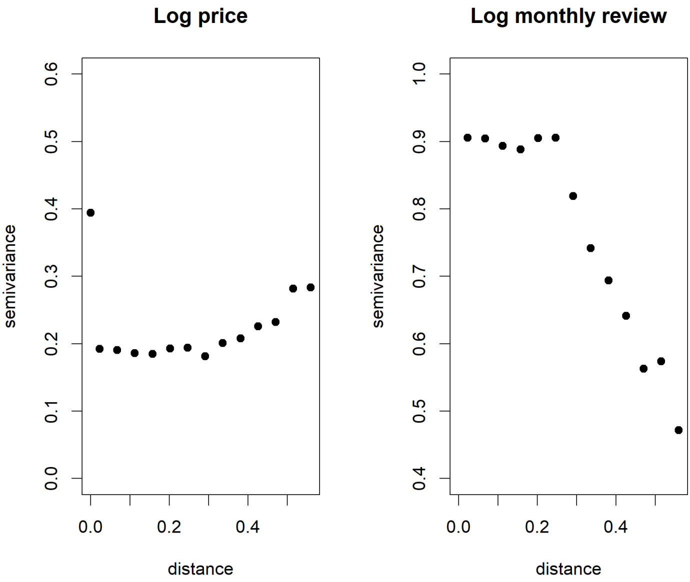
$$\text{Room type} + \text{Minimum nights} + \text{Last review}+$$

$$\text{Host listings count} + \text{Non-zero avail.} + \text{Available days}+$$

- Observations with no reviews are excluded (21% of the data)

Model Estimates

Did We Miss Spatial Correlation Within Neighbourhoods?



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 for price).

Text Analysis for Listing Names

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`