Exploratory Analysis of Data for Airbnb Listings in NYC

Youngsoo Baek, Irene Yi Ji, Phuc Nguyen

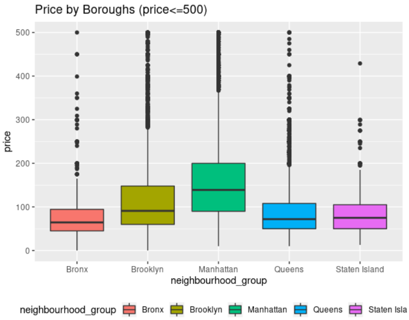
Introduction

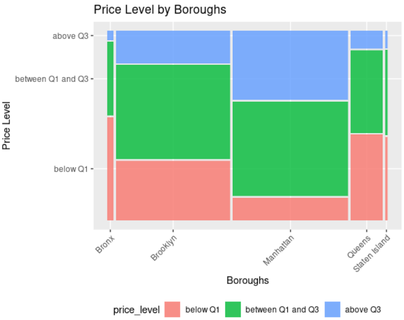
- Data: Airbnb New York City open data collected from 2019, with 48,895 listings and 16 variables.
- Goals:
  - Identify most influential factors for price/popularity
  - Examine heterogeneity across boroughs and neighbourhoods
  - Recommend best location and name for airbnb

Data Processing

- Remove 14 observations with *minimum_nights* > 365
- *Price*: the lowest non-zero value is 10, added 5 to 0's and take natural logarithm
- *Reviews per Month*: missing values are set to 0, since last review dates are missing and total number of reviews are 0 and take natural logarithm
- *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

EDA - Heterogeneity across Boroughs (Price)

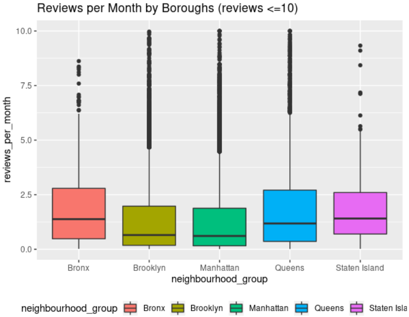




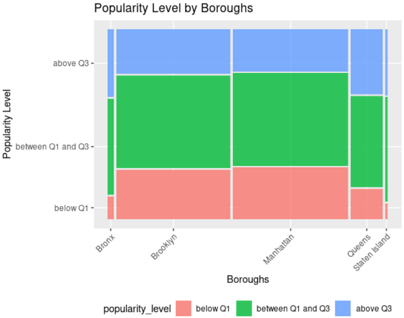
- Generate 3 price levels:
  "below Q1", "between Q1 and Q3", "above Q3"

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

EDA - Heterogeneity across Boroughs (Popularity)



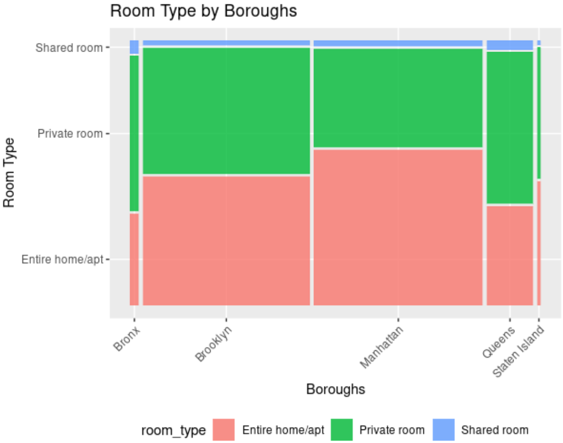
Reviews per Month by Boroughs (reviews <=10)



Popularity Level by Boroughs

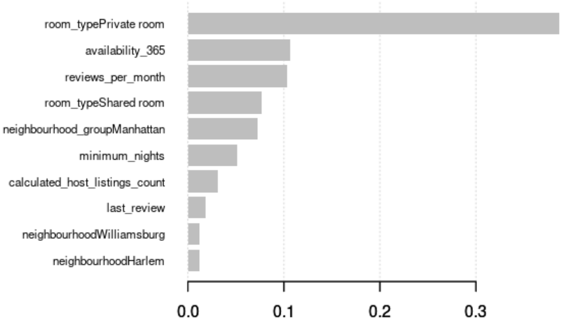
- Generate 3 popularity levels:
  “below Q1”, “between Q1 and Q3”, “above Q3”
- Pearson’s Chi-squared test: p-value $< 2.2\text{e-}16$

EDA - Heterogeneity across Boroughs (Room Type)



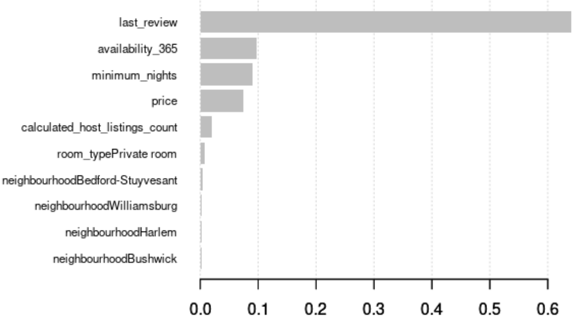
- Pearson’s Chi-squared test: p-value $< 2.2e\text{-}16$

EDA - XGBoost for Important Variables (Price)



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

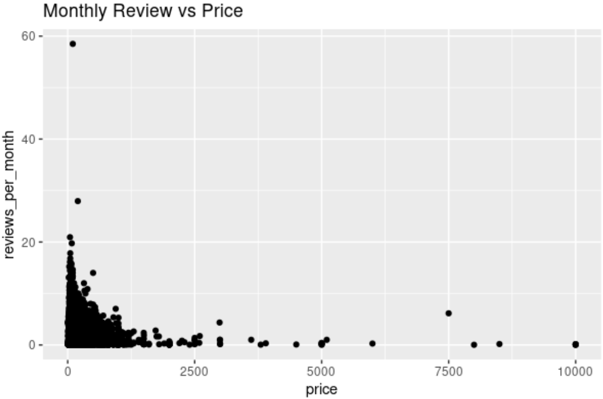
EDA - XGBoost for Important Variables (Popularity)



- 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:



- We may consider model them as a bivariate reponse.

Response of Interest: Price and Popularity

Choosing a Meaningful Measure of Popularity

Heterogeneity across Neighbourhoods/Boroughs

Spatial Correlation

Predictors of Interest

Possibly Unreliable Predictors

Modeling

Price and Popularity: Bivariate Mixed Effects Regression

Did We Miss Spatial Correlation Within Neighbourhoods?

Text Analysis for Listing Names

Further Work