# Modeling Price and Popularity of AirBnB listings in New-York

Melody Jiang, Raphael Morsomme, Ezinne Nwankwo

Department of Statistical Science, Duke University

02/03/2020

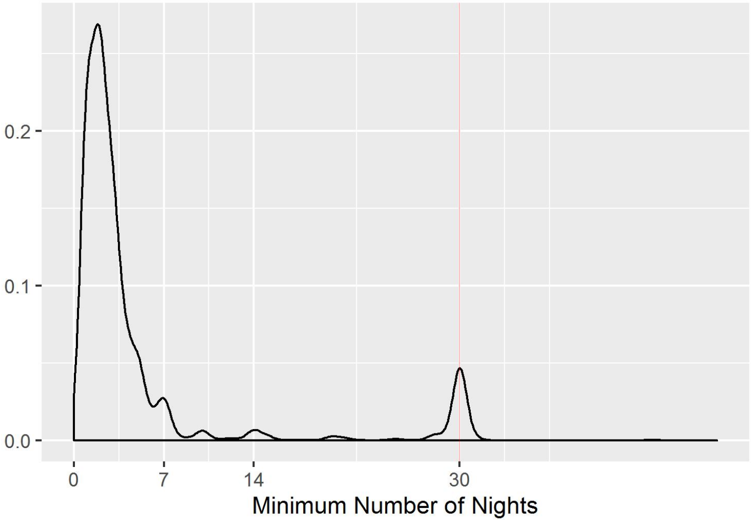
# Overview

- Data: Airbnblistings in NYC from 2019, 48,895 listings, 16 variables
- Questions
  - Influential factors on popularity / price
  - Heterogeneity among boroughs
  - If the type of listing vary across neighbourhoods
  - Where to locate listing and how to name listing to make listing most expensive and popular

# EDA - Issues with data

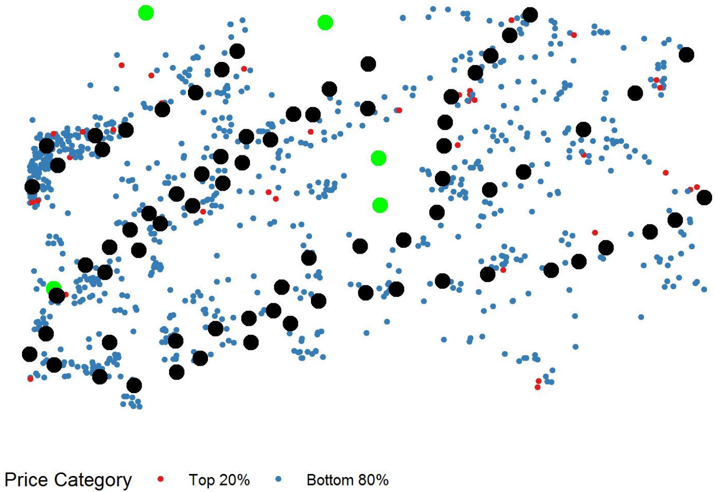
- Constructing a measure of popularity from limiting variables (time of last review)
- Improbable values
- ideally , tidy data (Wickham, 2009) with one row per booking
- Focus on data cleanin g and feature en g ineering over modeling.
- EDA will motivate the creation of new variables and the cleaning of the data.

# EDA - A city of two tales

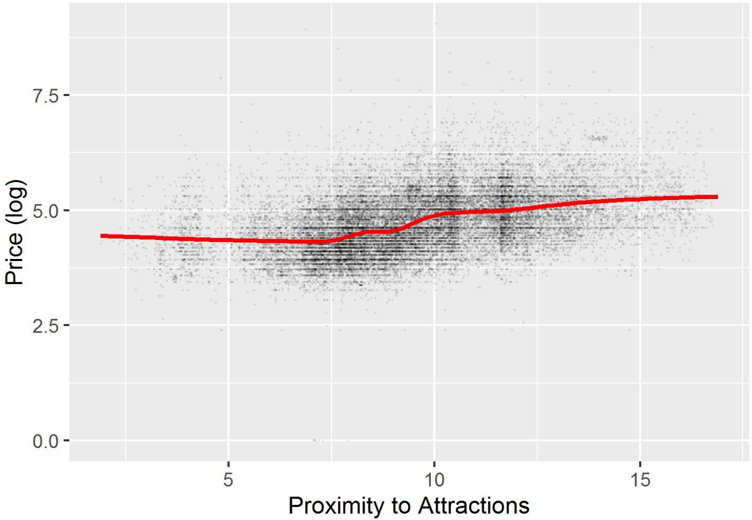


# EDA - Are you available?

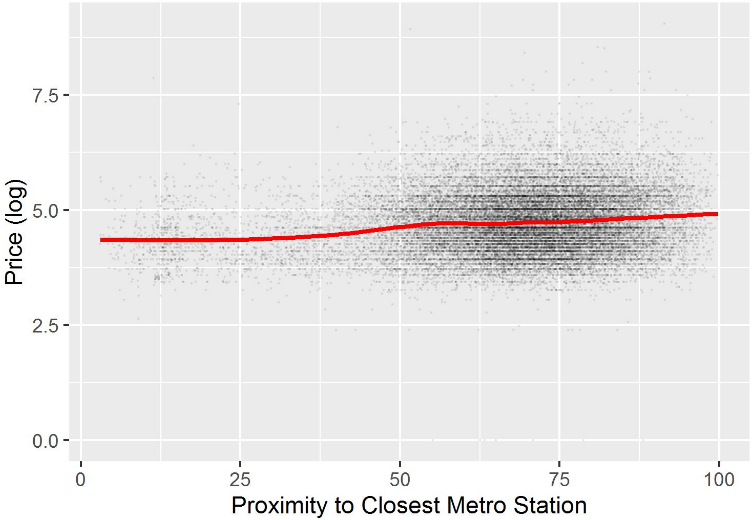
# EDA - Attractions



# EDA - Attractions



# EDA - Metro


Price (log)
7.5
5.0
2.5
0.0
0
25
50
75
100
Proximity to Closest Metro Station

# Data Cleaning

Drawing on the EDA, focus on active listings for short stay:
Keep listings with
(i) last review ¡ 1 year old [lose $15,000$]
(ii) minimum number days $< 30$ (short type of stay) [lose $XXX$]

# Data Cleaning

- Days since last review
  - not indicative of price nor popularity
  - A rough indicator of activeness of listing
- Calculated host listings
  - Exclude listings whose calculated host listings $> 5$
  - Different type of business
- Number of available days in a year
  - Excluded this variable, as we used it to calculate popularity

# Feature Engineering - Proximity

EDA shows impact of attraction on price. This suggests the creation of a variable measuring the proximity of a listing to attractions. The proximity variable is defined as the average proximity of the listing to the attractions

$$proximity(X) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{dist(X, attraction_i)}$$

where

$$dist(x, y) = \mid latitude_x - latitutde_y \mid + \mid longitude_x - longitude_y \mid .$$

is the Manhattan distance.
Similarly, we compute the proximity to the closest metro station.

# Feature Engineering - Textual Data

Sentiment analysis of listing name
- "documents" too short for topic modeling - Afinn dictionary (gradual rating)

$$Sentiment(X) = \frac{1}{n}\sum_{i=1}^{n} dictionary(x_i)$$

where $Afinn(x) \in \{-5, -4, \dots, 5\}$.
Origin of host name
- use name frequency as a proxy

# Models

Linear regression model $Y = X\beta$ where $X$ consists of:
proximity metro, proximity attraction, host name frequency, listing name sentiment, [newly created variables]
X1, X2, X3 [regular variable]
Random forest (n $= 1,500$, m $= 2/3$)
BMA (setting)

# Influential Factors

*Variable Importance* metric from the random forest (n $= 1,500$, m $= 2/3$)

*Posterior Inclusion Probability* from the BMA

# Sensitivity Analysis

Vary the setting of the RF: different levels of pruning, different values for $m$.
Vary the priors in the BMA: prior1, prior2, prior3

# Results - Influential Factors

¡Table of variable importance¿
¡Table of PIP¿

# Results - Q3

¡Figure for Q3¿

# Discussions

- Incusion of boroughs does not
  work well. Spatial modeling that addresses relationship between houses.(http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5
- Submarkets influence pricing. Finite mixture model. (https://pages.jh.edu/jrer/papers/pdf/forth/accepted/using

# References

Whickam, H.
Tidy Data
*Journal*, month year