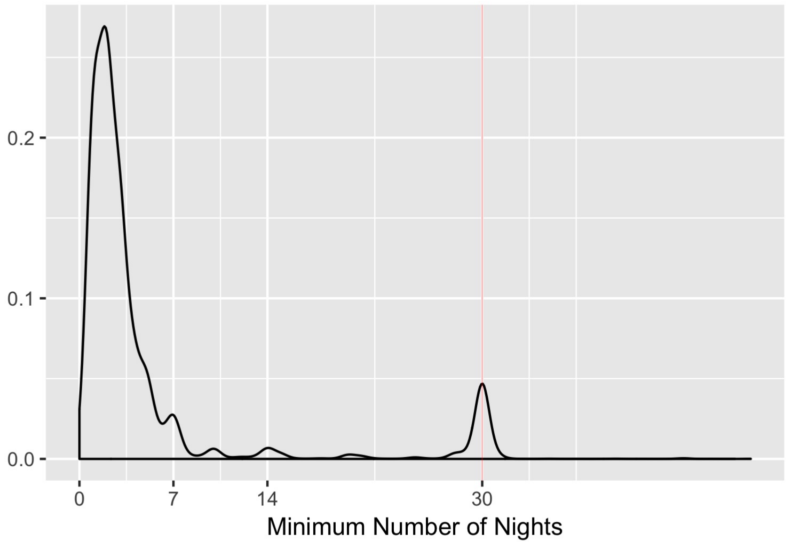
Modeling Price and Popularity of AirBnB listings in New-York

Melody Jiang, Raphael Morsomme, Ezinne Nwankwo

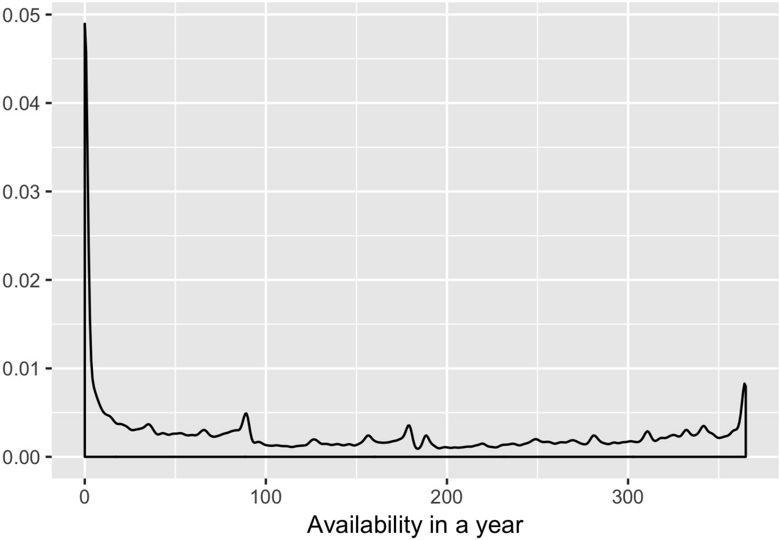
Department of Statistical Science, Duke University

02/03/2020

EDA - A city of two tales

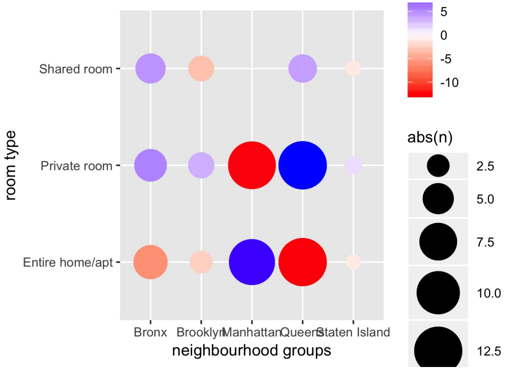


EDA - Are you available?

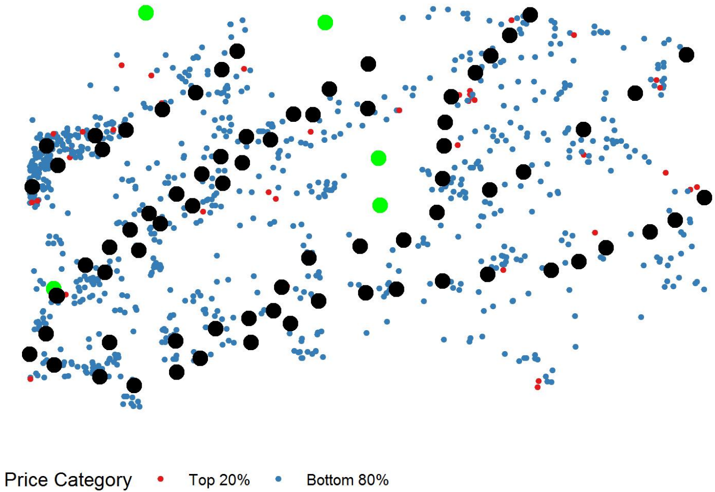


EDA - Room type

- $p < 1e-16$



EDA - Attractions



Data Cleaning

Since the data quality is questionable, we narrow down our focus to *active* listing for *short term* stays that are *private*:

- last review is less than 12 months old
- min number of days inferior to to 29
- less than 5 listings per owner

We also incorporate external data:

- Location of metro stations
- Location of main attractions

Feature Engineering - Proximity to Metro and Attractions

EDA suggests spatial modeling:

- Proximity to closest metro stations

$$dist_{Manhattan}(x, y) = |lat_x - lat_y| + |long_x - long_y|$$

- Average proximity to (36) attractions

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

Feature Engineering - Textual Data

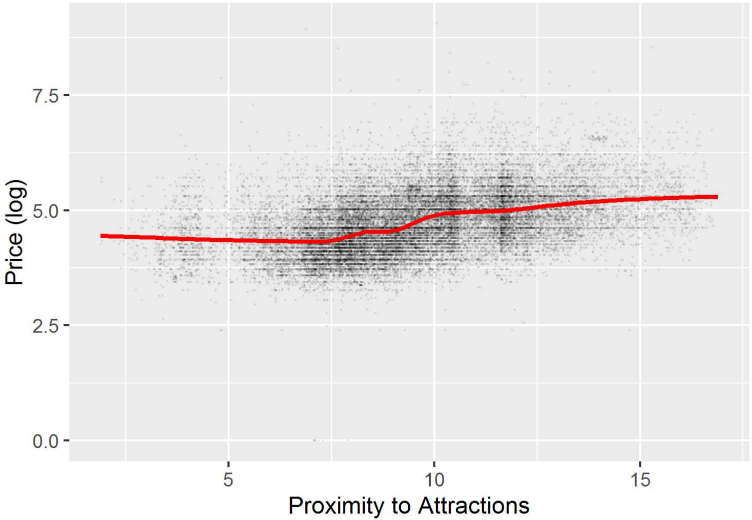
Textual data always invites creativity:

- Sentiment analysis of listing name

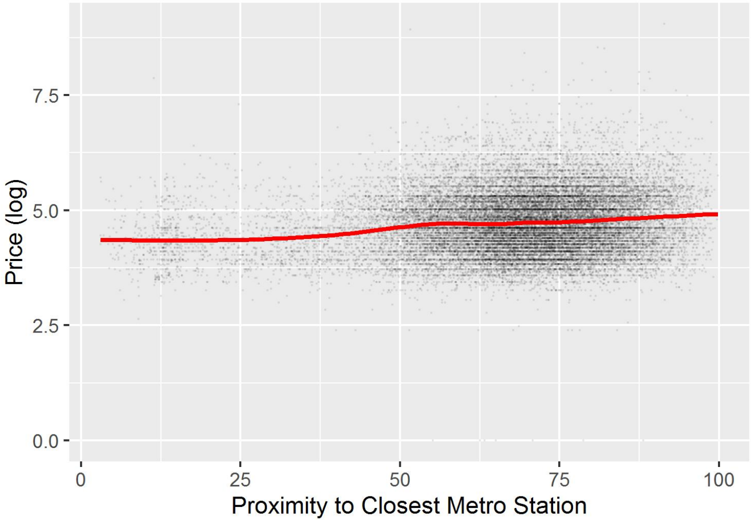
$$Sentiment(W) = \frac{1}{n}\sum_{i=1}^{n} dictionary(w_i)$$

  where $dictionary_{Affin}(w) \in \{-5, -4, \ldots, 5\}$
- Origin of host $\Rightarrow$ frequency of name

Feature Engineering - Attractions



Feature Engineering - Metro



Models

Two regression models

- Regression Model
  - Outcome: price, popularity
    - popularity $= \frac{\text{reviews per month}}{\text{availability}}$
  - Predictors: listing sentiment, name frequency, proximity metro, proximity attraction, room type, etc.
  - $n = 25,000$, $p = 14$
- Random Forest
  - $1,900$ trees, subsamples $19,000$
  - $m = \frac{p}{3} = 4$ variables at each split, $n_{leaf} \geq 5$.
- BMA
  - Linear combination of predictors
  - Prior: Cauchy
  - N.Models: 2^{15}
  - MCMC.iterations: 10^{16}
  - initprobs = "marg-eplogp"

Models - continued

- Important Predictors
  - RF: variable importance
  - BMA: PIP
- Sensitivity Analysis
  - RF: vary m and minimum n_{leaf}.
  - BMA: consider different priors (Cauchy prior, g prior for $g = 1, 5, 8, 100, 500, 1000$)

Results - Random Forest Price

	Predictors	Variable.Importance
1	longitude	978.10
2	proximity_metro	231.64
3	reviews_per_month	107.88
4	listing_sentiment	101.15
5	number_of_reviews	100.73
6	neighbourhood_group	95.16
7	latitude	94.24
8	room_type	90.71
9	name_listing_length	89.26
10	minimum_nights	88.53
11	popularity_log	61.00
12	name_host_special	53.69
13	proximity_attraction	47.78
14	name_host_freq	39.16
15	listing_count	16.42

Results - Random Forest Popularity

	Predictors	Variable.Importance
1	latitude	62.77
2	listing_sentiment	59.50
3	minimum_nights	57.85
4	neighbourhood_group	54.93
5	room_type	50.03
6	reviews_per_month	49.69
7	longitude	45.87
8	name_listing_length	41.50
9	name_host_special	35.34
10	popularity_log	33.77
11	proximity_attraction	20.73
12	listing_count	8.76
13	name_host_freq	8.72
14	proximity_metro	0.42

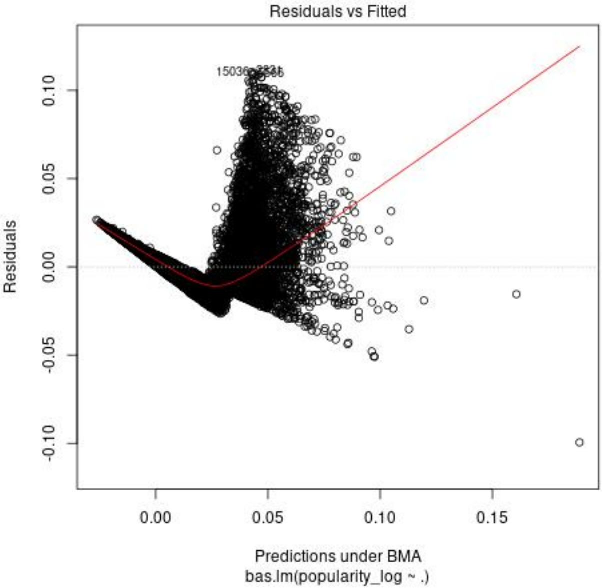
Results - BMA Price

	Predictors	PIP	Estimate	Lower.Confint	Upper.Confint	Significance
1	Intercept	1.00	4.70	4.70	4.71	*
2	name_host_freq	1.00	0.07	0.07	0.07	*
3	Neighbourhood_group:Manhattan	1.00	0.17	0.15	0.18	*
4	number of reviews	1.00	-0.01	-0.02	-0.01	*
5	Room_type:Entire home/apt	1.00	0.74	0.73	0.75	*
6	availability_365	1.00	0.00	0.00	0.00	*
7	name_listing_sentiment	1.00	0.00	0.00	0.00	*
8	last_review	1.00	-0.00	-0.00	-0.00	*
9	Neighbourhood_group:Queens	1.00	-0.11	-0.13	-0.10	*
10	name_host_special:True	1.00	10.42	7.38	13.55	*
11	Room_type:Shared room	1.00	-0.51	-0.54	-0.47	*
12	calculated_host_listings_count	1.00	-0.01	-0.02	-0.01	*
13	type_stay:Long	1.00	0.00	0.00	0.00	*
14	Neighbourhood_group:Bronx	1.00	-0.17	-0.20	-0.14	*
15	name_listing_length	1.00	0.06	0.04	0.08	*
16	proximity_metro	0.89	-0.00	-0.00	0.00	
17	proximity_attraction	0.50	-0.01	-0.04	0.00	
18	reviews_per_month	0.07	-0.00	-0.00	0.00	
19	Neighbourhood_group:Staten Island	0.02	-0.00	0.00	0.00	

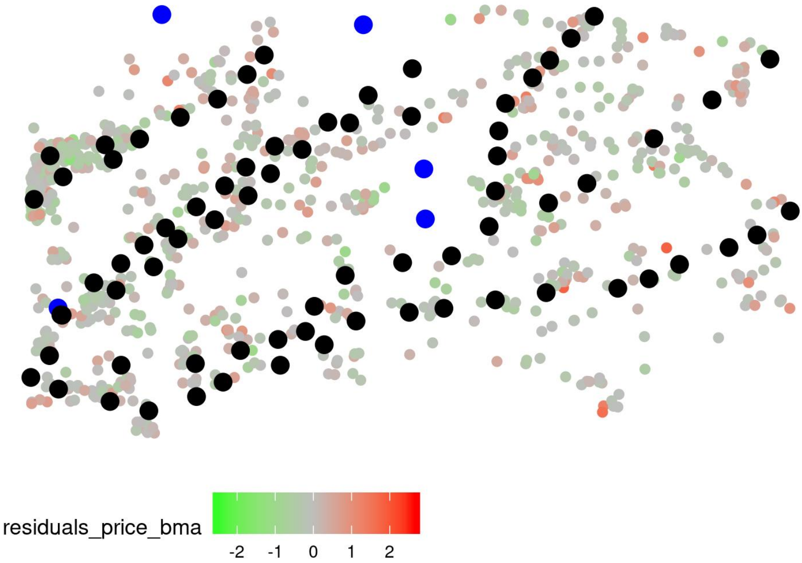
Results - BMA Popularity

	Predictors	PIP	Estimate	Lower.Confint	Upper.Confint	Significance
1	Intercept	1.00	0.02	0.02	0.02	*
2	calculated_host_listings_count	1.00	0.01	0.01	0.01	*
3	reviews_per_month	1.00	0.00	0.00	0.00	*
4	name_listing_sentiment	1.00	-0.00	-0.00	-0.00	*
5	Room_type:Entire home/apt	1.00	-0.00	-0.00	-0.00	*
6	number of reviews	0.95	-0.00	-0.00	0.00	
7	last_review	0.74	-0.00	-0.00	0.00	
8	proximity_metro	0.62	0.00	0.00	0.00	
9	Room_type:Shared room	0.23	0.00	0.00	0.00	
10	Neighbourhood_group:Queens	0.14	-0.00	-0.00	0.00	
11	name_listing_length	0.05	-0.00	0.00	0.00	
12	name_host_freq	0.03	0.00	0.00	0.00	
13	Neighbourhood_group:Bronx	0.02	0.00	0.00	0.00	
14	Neighbourhood_group:Staten Island	0.01	0.00	0.00	0.00	
15	name_host_special:True	0.01	-0.00	0.00	0.00	
16	availability_365	0.01	0.00	0.00	0.00	
17	Neighbourhood_group:Manhattan	0.01	-0.00	0.00	0.00	
18	type_stay:Long	0.01	-0.00	0.00	0.00	
19	proximity_attraction	0.01	-0.00	0.00	0.00	

Diagnostic Plots - BMA - Popularity



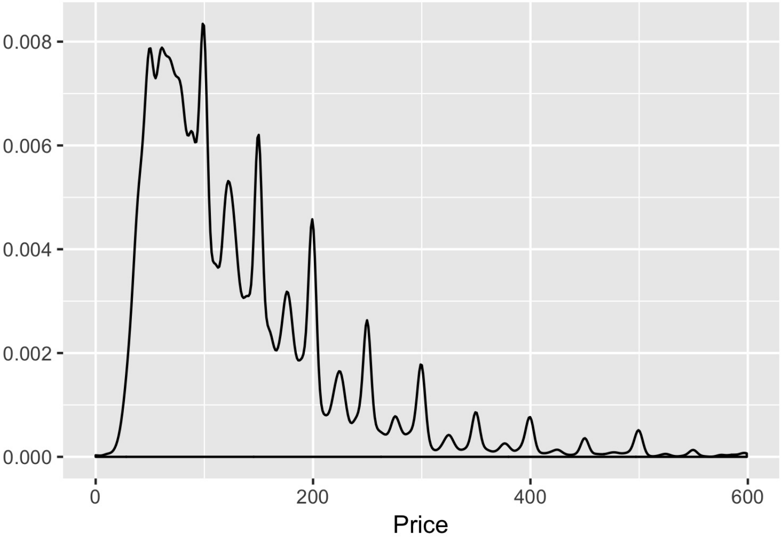
Residuals - BMA



Discussions - Future Directions

- Data collection
  - One row per booking
- Spatial modeling
  - Incusion of boroughs does not work well. Spatial modeling that addresses relationship between houses (Valente, 2005).
- Domain-specific knowledge
  - Submarkets influence pricing. Finite mixture model by Belasco1, 2012.
  - Knowledge in marketing. E.g. consider form of pricing as an influencer of popularity (99 vs 100).

Discussions



References

Whickam, H.
Tidy Data
Journal, month year

Valente, j.
Apartment Rent Prediction Using Spatial Modeling
Journal, month year

Belasco, E.
Using a Finite Mixture Model of Heterogeneous Households to Delineate Housing Submarkets
Journal, month year