Understanding Price and Popularity in New York City Airbnb Sites

Rihui Ou, David Buch

Duke University

February 4, 2020

Problem Statement

- AirBnB is an online marketplace allowing property owners to advertise housing available for short-term rent and connect with guests seeking accommodation
- We want to understand the role of location and room type in determining the price and popularity of AirBnB listings in NYC
- Data describe nearly 50k sites listed on AirBnB website in 2019
- Specifically we want to know:
  1. What are Influential Factors in Price/Popularity?
  2. Does Neighborhood/Borough Affect Price/Popularity? (Rank Ordering)
  3. Does Listing Type differ among Neighborhoods/Boroughs
  4. How (including site name) might a host Maximize a Listing's Price/Popularity?

Data Formatting/EDA

Problems with the Data:

- `price = 10k` too often
- Some sites never available, some sites have no reviews. Both present potential problems.
- **No good measure of popularity**

Defining a Proxy for Popularity:

- Define: "Popularity[1]" is proportional to unique customers per day available
- Assume: Review probability independent of experience. Popularity constant over time. Availability constant over time.
- $\text{Popularity} = \frac{12 * \textit{reviews\_per\_month}}{\textit{availability\_365}}$

[1] Note this would bias popularity towards sites with shorter length of stay (pointed out by Joseph Lawson)

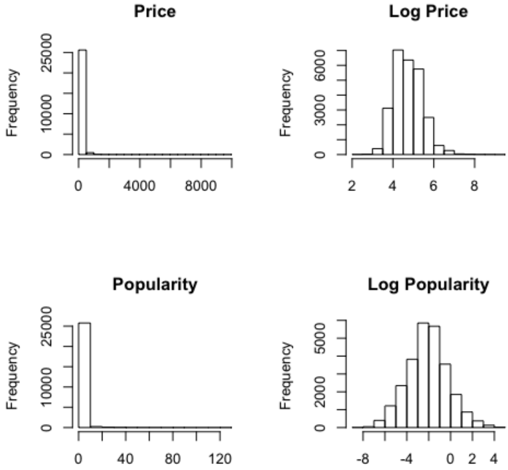
Name Features

- Want to understand how site name influences price and popularity.
- We specifically look at choice of words describing site: adjectives
- Check for most common words
- Remove location and room type words
- Subjectively group adjectives into major categories

comfort	space	beauty	upkeep	luxury	superlative
cozy	spacious	beautiful	new	luxury	great
comfy	large	lovely	bright		amazing
comfortable	huge	gorgeous	clean		prime
charm	big	view	sunny		best
charming	space		modern		perfect
quiet					

Logging Responses

Note: Price of goods and popularity are both commonly modeled with lognormal or Pareto distributions



Model

Gaussian spatial regression model

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})^\top \boldsymbol{\beta} + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \quad \epsilon(\boldsymbol{s}) \overset{iid}{\sim} N\left(0, \tau^2\right), \quad w(\boldsymbol{s}) \sim GP(\mathbf{0}, C(\cdot,\cdot))$$

where $\boldsymbol{s} = (\text{longtitude}, \text{latitude}) \in \Re^2$ is a location on the 2-D plane.

- $y(\boldsymbol{s})$ responses: log(Price) or log(Popularity)
- $\boldsymbol{x}(\boldsymbol{s})$ covariates: room_type, minimum_nights, availability_365, calculated_host_listings_count
- $w(\boldsymbol{s})$ a Gaussian spatial process with covariance function $C(\cdot,\cdot)$: these are random intercepts with spatial dependence which capture the effect that cannot be explained by covariates.
- $\epsilon(\boldsymbol{s})$: a white noise process that stands for measurement errors.

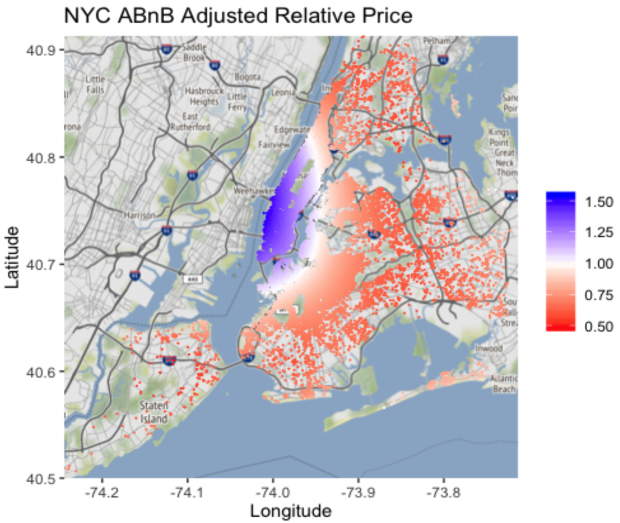


Figure: Relative price map of NYC. Thanks to **Alessandro Zito** for guidance and support in creating this plot.

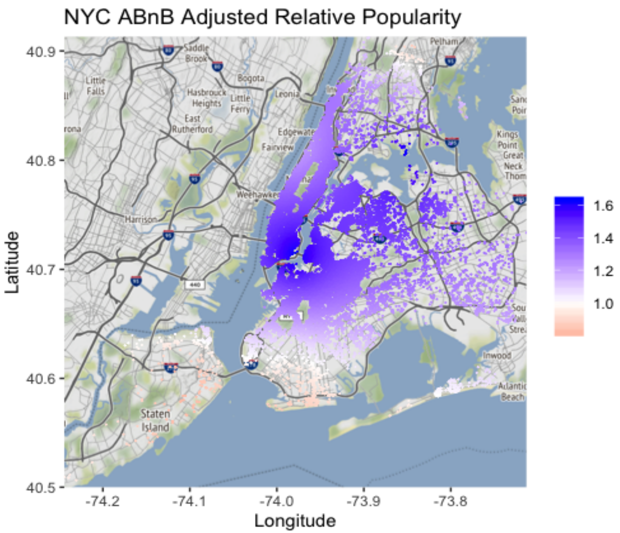


Figure: Relative popularity map of NYC. Thanks to **Alessandro Zito** for guidance and support in creating this plot.

(i) - Influential Factors

Price				Popularity		
Effect Size						
	mean	2.5%	97.5%	mean	2.5%	97.5%
room_typePrivate room	-51.3540	-51.9198	-50.7618	-12.8378	-16.8896	-8.6252
room_typeShared room	-68.5131	-69.6083	-67.3524	-38.0226	-46.2593	-28.4590
minimum_nights	-0.2397	-0.2597	-0.2098	-1.9801	-2.0977	-1.8723
comfort	-11.0059	-12.3747	-9.6428	25.7216	18.3292	33.4691
space	0.7125	-0.9752	2.4700	-3.5360	-9.7873	2.8087
beauty	1.4301	-0.6678	3.5516	-8.8442	-15.5238	-1.4987
upkeep	-2.2347	-3.8249	-0.6578	7.0365	0.7125	13.8145
luxury	31.9034	27.4431	36.3016	-19.0226	-28.6305	-8.0293
superlative	-3.1493	-5.1999	-1.0940	-1.5184	-9.3442	7.1758
calculated_host_listings_count	0.0100	-0.0100	0.0300	-0.7373	-0.8067	-0.6678

Figure: 95% CIs and means of the fixed effects on the percentage scale

Influential factors for

- Price: room type, minimum nights, some words...
- Popularity: room type, minimum nights, listing per host, some words...

(ii) - Heterogeneity among neighbourhoods and boroughs

Price — Borough Ranking

	2.5%	97.5%	mean
Staten Island	-57.8005	3.9696	-34.7004
Bronx	-57.8917	3.3409	-33.4774
Queens	-53.5366	14.4812	-26.0713
Brooklyn	-46.1235	32.7740	-13.2858
Manhattan	-25.6824	82.6514	20.5608

Popularity — Borough Ranking

	2.5%	97.5%	mean
Staten Island	-39.7221	56.5907	-6.1193
Bronx	-16.9913	106.6722	25.6879
Brooklyn	-9.4662	120.4741	35.4045
Queens	-8.1576	127.3634	38.3053
Manhattan	-6.1284	128.6615	40.2268

Price — Neighbourhood (Worst, Best)

	2.5%	97.5%	mean
Todt Hill	-68.1350	-21.2969	-51.9842
Emerson Hill	-67.5731	-19.9242	-51.0236
Willowbrook	-66.5642	-18.0779	-49.3892

	2.5%	97.5%	mean
Greenwich Village	-11.2131	117.5139	43.9197
Chelsea	-9.0925	122.8541	47.1832
West Village	-7.8282	125.3125	49.1954

Popularity — Neighbourhood (Worst, Best)

	2.5%	97.5%	mean
Todt Hill	-50.76306	42.37842	-18.13709
Grant City	-47.39362	36.40533	-17.95273
New Dorp Beach	-47.24878	35.50817	-17.47429

	2.5%	97.5%	mean
Vinegar Hill	4.561689	153.4457	57.71984
Clason Point	-1.083876	158.2423	57.78168
Lower East Side	5.248765	154.3549	58.13260

Figure: Expected values and 95% CIs inferred from resampling sites within boroughs/neighbourhoods on the percentage scale

Most

- Expensive borough: **Manhattan**, Expensive neighbourhood: **West Village**
- Popular borough: **Manhattan**, Popular neighbourhood: **Lower East**

(iii) - Independence of room type among neighborhoods

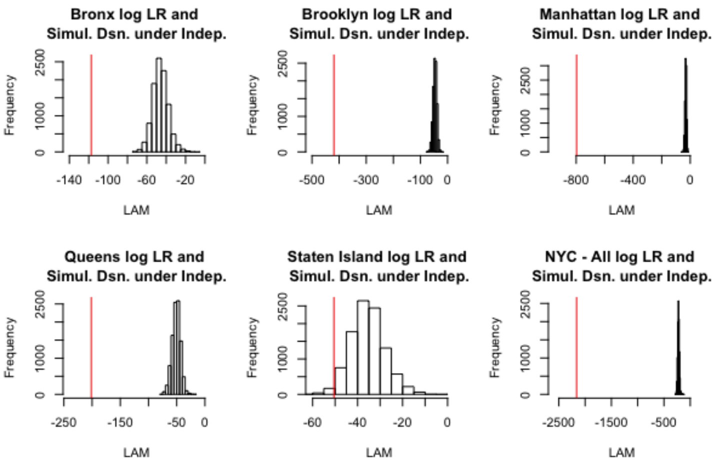


Figure: Due to low cell counts, χ^2 test asymptotic assumptions were not valid. Here we see log likelihood test statistics of independence plotted against simulated distributions of the test statistic under independence. (Staten Island $p \approx 0.02$)

Summary

- **Maximize Price :** "Luxury Entire Home in West Village" $(-15\%, 388\%)$ increase compared to an average location and no selected adjective used
- **Maximize Popularity :** "New cozy entire home in Lower East Side Manhattan" $(-14\%, 470\%)$ increase
- Future Work:
  - Improved proxy for popularity; better data
  - Joint modeling of price and popularity; inference about potential revenue