

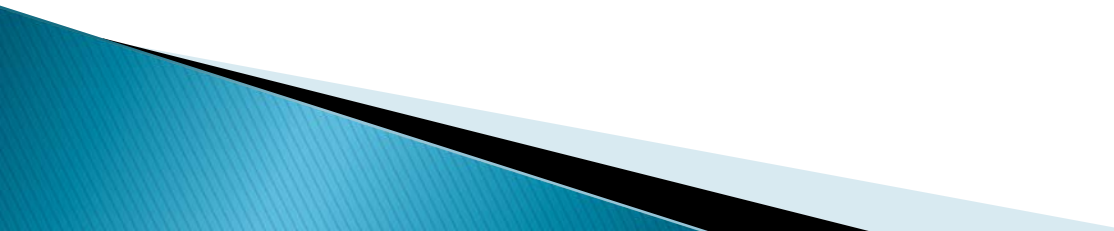
# New York City Neighborhood Search Tool

Jim Cole  
October, 2019

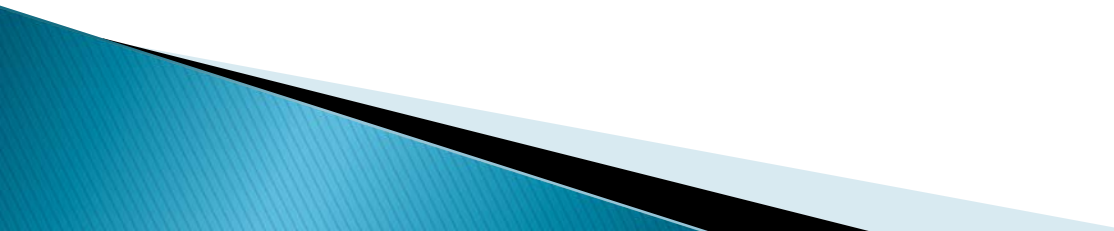
# Business Need

- ▶ Approximately 500,000 people move to or within Manhattan and Brooklyn every year
- ▶ Many of them wonder:
  - Which neighborhoods in NYC have the amenities I want?
  - Which of those neighborhoods can I afford?
- ▶ It's difficult-to-impossible to find answers for an arbitrary set of interests
- ▶ Let's see how this tool can help us figure out which are the best and most affordable neighborhoods with many **jazz clubs** and also **burrito shops**

# Data Required

- ▶ A list of FourSquare categories and sub-categories
  - ▶ The latitude and longitude of rectangular sections of Manhattan and Brooklyn, used to request data specific to those two boroughs from FourSquare
  - ▶ Lists of venues from various categories from FS
  - ▶ Reverse geocoding data from Bing to retrieve missing zip codes
  - ▶ The zip code(s) in each neighborhood in the two boroughs
  - ▶ The population of each zip code in the two boroughs
  - ▶ A geojson file with the boundaries of each zip code
  - ▶ The median cost of a 1-bedroom apartment rental in each neighborhood
- 

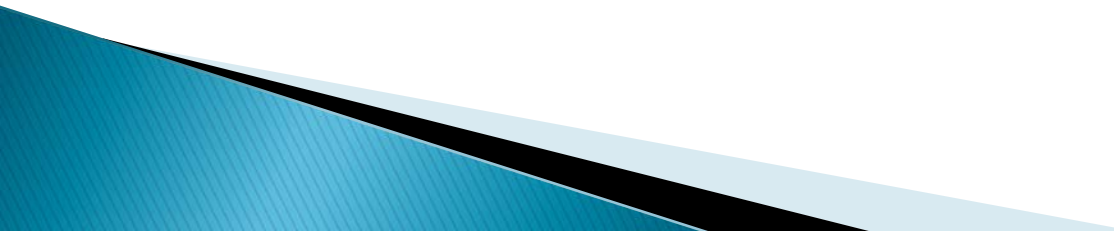
# Data Sources and Cleaning

- ▶ FourSquare venue categories are retrieved as hierarchical JSON data and flattened
  - ▶ A list of venues matching chosen categories is retrieved via numerous calls to FourSquare
    - Each call specifies a small geographic area because FS returns at most 50 venues/call
    - Duplicate entries are deleted
    - Missing venue zip codes are retrieved via Bing's reverse geocoding API
    - Venues outside the two boroughs are discarded
- 

# Data Sources and Cleaning

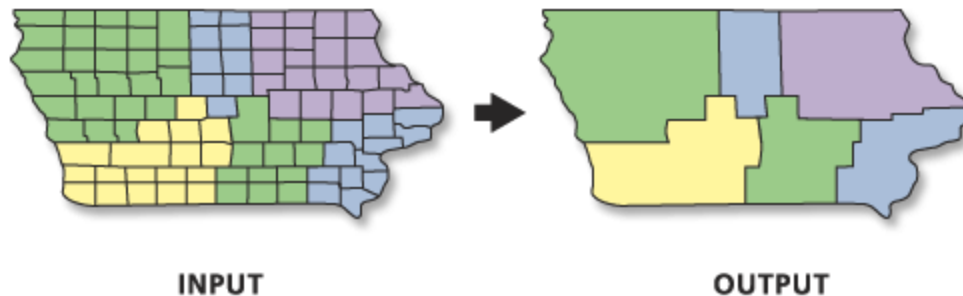
- ▶ Biggest issue: **Loosely matching FS venues**
  - FS returns venues with any remote connection to the queried category
  - E.g., Madison Square Garden is returned for almost any category queried because so many different types of events happen there
  - Thus, 1000's of venues are returned for almost any query, but most are irrelevant
  - The first category listed for a venue identifies its primary usage; this allows the tool to discern between strong matches and weak matches:
    - Village Vanguard, a famous jazz club – keep
    - Joe's Italian Restaurant (that has music sometimes) – discard

# Data Sources and Cleaning

- ▶ Recent data on median rent for 1BR apartment from rental site [zumper.com](https://www.zumper.com)
  - ▶ Neighborhood populations from [NYC Open Data Project](https://data.cityofnewyork.us/Neighborhoods/NYC-Open-Data-Project)
  - ▶ A list of zip codes in each neighborhood from [NY State Department of Health](https://www.health.ny.gov/data/zip_codes/), but the data is incomplete
  - ▶ Missing zip codes looked up on [PropertyShark.com](https://www.propertyshark.com)
- 

# Data Sources and Cleaning

- ▶ An open source [geojson file](#) containing boundaries of every zip code in NYC
  - To obtain neighborhood boundaries:
    - Zip code boundaries are geospatially “dissolved”
    - Inner zip code boundaries are removed, outer neighborhood boundaries are retained
    - 97 zip codes → 30 neighborhoods



[Image from ESRI](#)

# Methodology

- ▶ Venues matched to neighborhoods by zip
- ▶ Venue density: per-capita number of venues
- ▶ Rent factor: Geometric ratio of a neighborhood's rent compared to the least expensive neighborhood
- ▶ Attractiveness: Venue density that takes relative rent into account, normalized

$$\text{venue density} = \frac{\text{Number of venues in the neighborhood}}{10,000 \text{ (residents)}}$$

$$\text{rent factor} = (\text{median monthly rent for a 1 - bedroom apt.} / \text{cheapest rent})^3$$

$$\text{attractiveness} = \frac{\text{venue density}}{\text{rent factor}}$$



# Methodology

- ▶ Machine Learning
- ▶ k-means clustering used to find clusters of venues, regardless of neighborhood boundaries

# Results

- ▶ Neighborhoods with highest per-capita density of jazz clubs and burrito shops are identified
- ▶ It'd be great to live in Chelsea if you can afford it

	Borough	SubBoroNumber	Name	Population	Rent	VenueCount	VenueDensity
21	Manhattan	4	Chelsea, Clinton	103245	3700	22.0	2.130854
22	Manhattan	5	Midtown Business District	51673	3125	11.0	2.128771
19	Manhattan	2	Greenwich Village, Soho	90016	3375	18.0	1.999645
18	Manhattan	1	Battery Park City, Tribeca	60978	3877	10.0	1.639936
1	Brooklyn	2	Brooklyn Heights, Fort Greene	99617	2975	9.0	0.903460
3	Brooklyn	4	Bushwick	112634	2200	7.0	0.621482
27	Manhattan	10	Central Harlem	115723	2010	7.0	0.604893
5	Brooklyn	6	Park Slope, Carroll Gardens	104709	2555	6.0	0.573017
24	Manhattan	7	West Side, Upper West Side	209084	3150	11.0	0.526104
23	Manhattan	6	Stuyvesant Town, Turtle Bay	142745	3300	7.0	0.490385

# Results

- ▶ But Chelsea is the 2<sup>nd</sup> most expensive neighborhood, many people cannot afford it

Here are the 30 neighborhoods listed from most expensive rent to least expensive

Borough	Name	Rent			
Manhattan	Battery Park City, Tribeca	3877	Brooklyn	Bedford Stuyvesant	2200
Manhattan	Chelsea, Clinton	3700	Manhattan	East Harlem	2145
Manhattan	Greenwich Village, Soho	3375	Manhattan	Central Harlem	2010
Manhattan	Stuyvesant Town, Turtle Bay	3300	Manhattan	Manhattanville, Hamilton Heights	1925
Manhattan	West Side, Upper West Side	3150	Brooklyn	Flatbush, Midwood	1800
Manhattan	Midtown Business District	3125	Manhattan	Washington Heights, Inwood	1775
Brooklyn	Brooklyn Heights, Fort Greene	2975	Brooklyn	Bay Ridge, Dyker Heights	1700
Manhattan	Lower East Side, Chinatown	2950	Brooklyn	Borough Park, Ocean Parkway	1690
Manhattan	Upper East Side	2800	Brooklyn	Sheepshead Bay, Gerritsen Beach	1690
Brooklyn	Williamsburg, Greenpoint	2720	Brooklyn	East Flatbush, Rugby, Farragut	1650
Brooklyn	Park Slope, Carroll Gardens	2555	Brooklyn	Canarsie, Flatlands	1542
Brooklyn	Bushwick	2200			

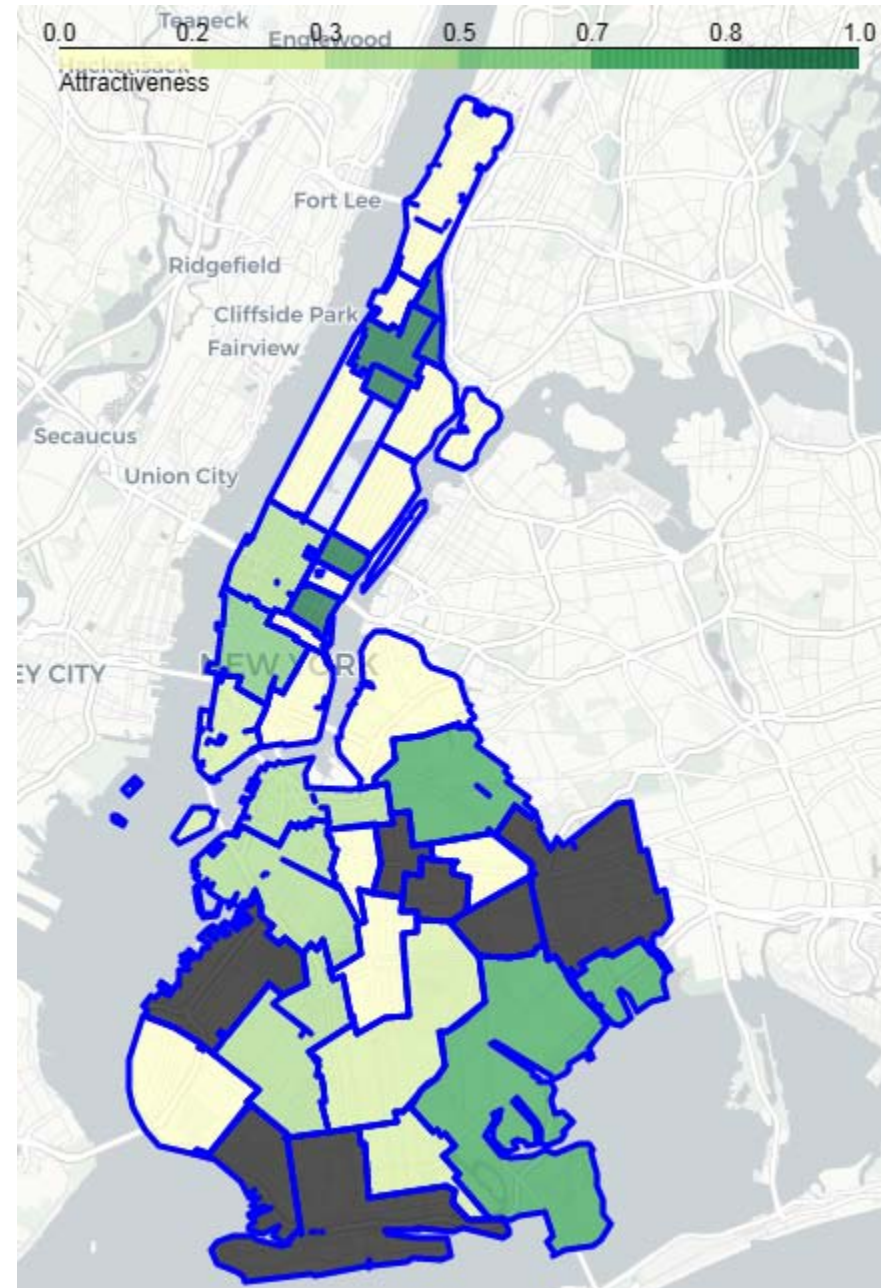
# Results

- ▶ “Attractiveness” factors in the cost of rent
- ▶ Neighborhoods from less expensive Brooklyn have moved up the list; Chelsea down

	Borough	SubBoroNumber	Name	Population	Rent	VenueCount	VenueDensity	Attractiveness	RentFactor
27	Manhattan	10	Central Harlem	115723	2010	7.0	0.604893	1.000000	3.123777
22	Manhattan	5	Midtown Business District	51673	3125	11.0	2.128771	0.926197	11.739294
3	Brooklyn	4	Bushwick	112634	2200	7.0	0.621482	0.748602	4.096000
17	Brooklyn	18	Canarsie, Flatlands	193543	1542	4.0	0.206672	0.717440	1.410409
19	Manhattan	2	Greenwich Village, Soho	90016	3375	18.0	1.999645	0.649576	14.788129
21	Manhattan	4	Chelsea, Clinton	103245	3700	22.0	2.130854	0.494463	19.484850
5	Brooklyn	6	Park Slope, Carroll Gardens	104709	2555	6.0	0.573017	0.374205	6.416004
1	Brooklyn	2	Brooklyn Heights, Fort Greene	99617	2975	9.0	0.903460	0.373532	10.128679
11	Brooklyn	12	Borough Park, Ocean Parkway	191382	1690	3.0	0.156755	0.344898	1.856744
16	Brooklyn	17	East Flatbush, Rugby, Farragut	155252	1650	2.0	0.128823	0.285671	1.728000

# Results

- ▶ The “attractiveness” of each neighborhood is easy to compare on a choropleth map

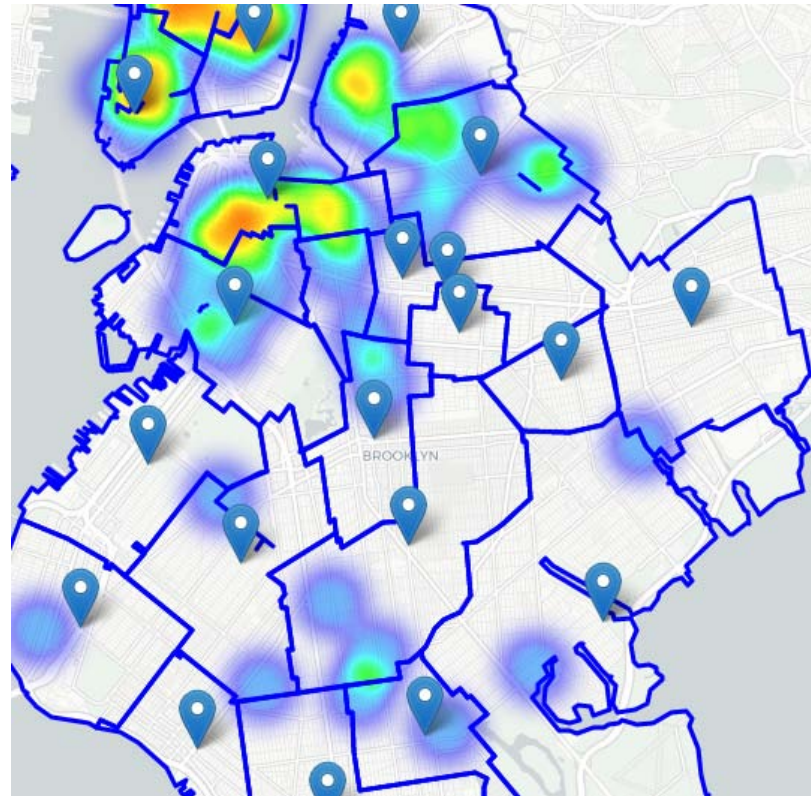




# Results

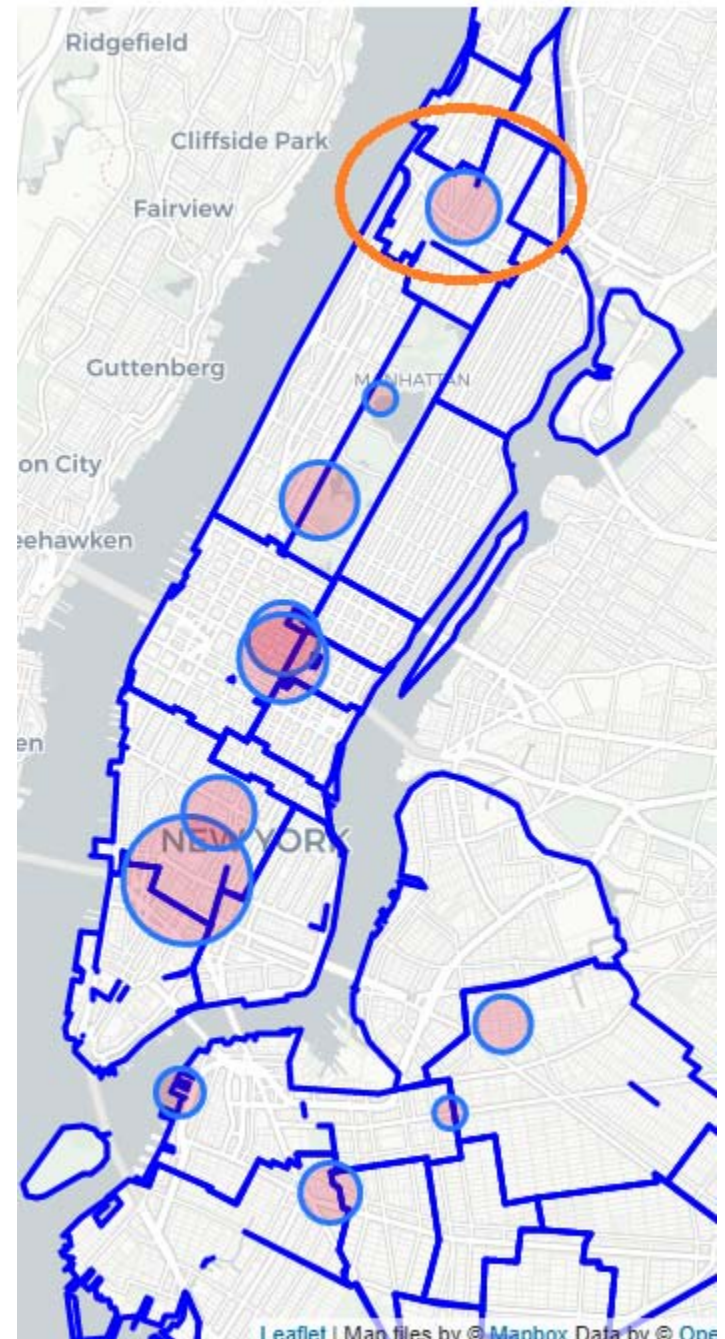


- ▶ A heatmap drills down to show where the venues are in the neighborhoods



# Results

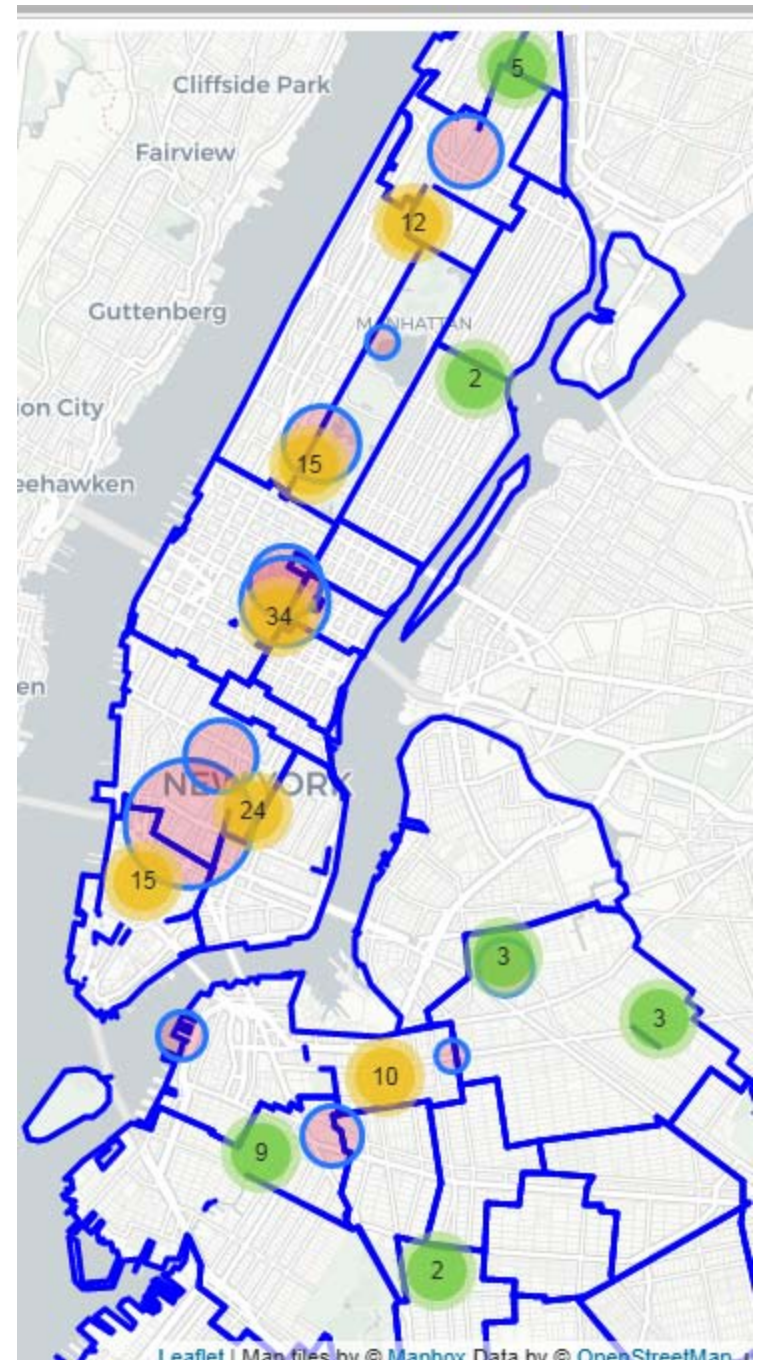
- ▶ k-means clustering shows groups of venues without regard to neighborhood
- ▶ Not surprisingly, there's a cluster of jazz clubs in Harlem, a center of jazz for over 100 years





# Results

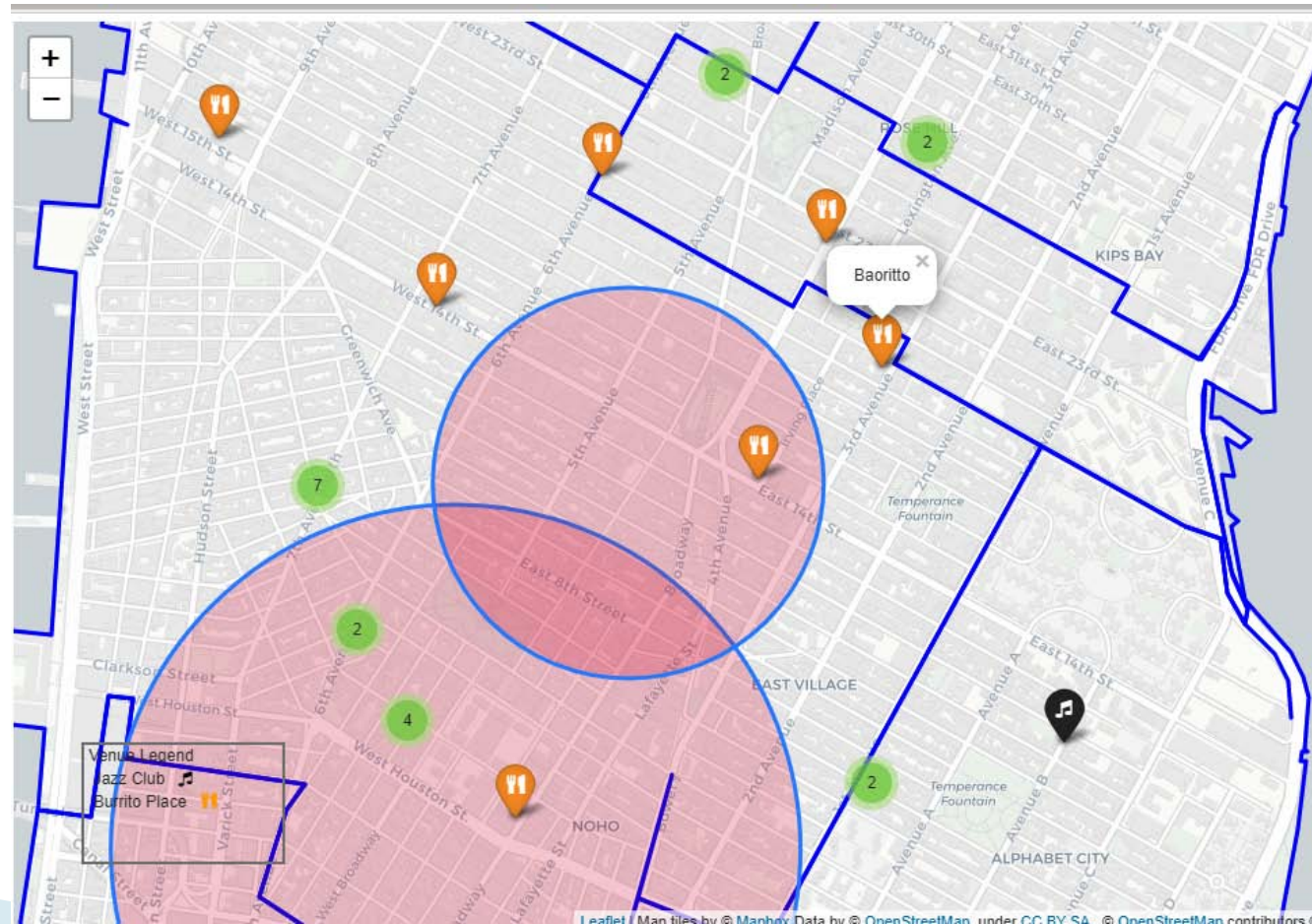
- ▶ The cluster map becomes an interactive tool when all of the venues are added to it
- ▶ It's easy to see how many venues are in any area
- ▶ And...





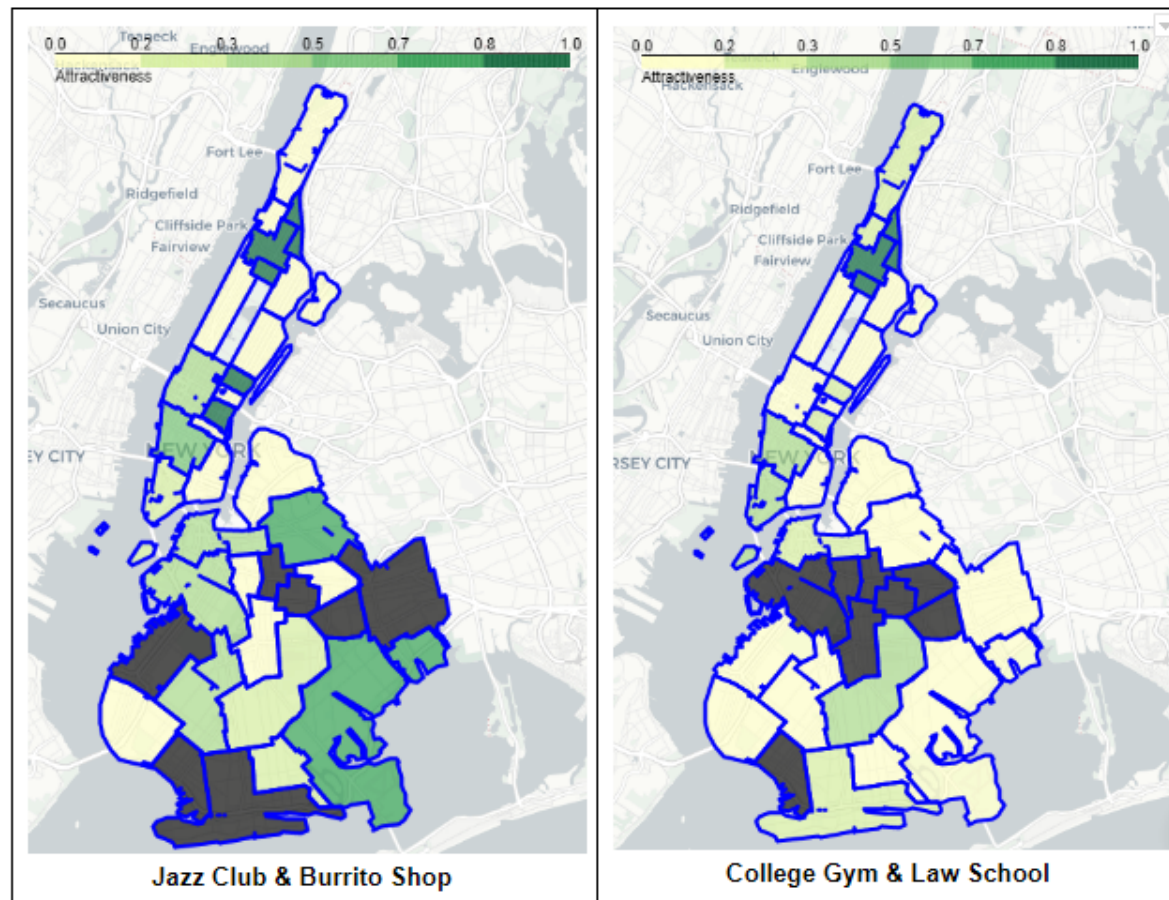
# Results

- Users can zoom in to see individual venues



# Results

- Does the tool really highlight different neighborhoods for different venue choices?
- Yes



# Conclusions

- ▶ Data science and machine learning techniques can provide insights for geographic questions that are very difficult to obtain with other techniques
  - ▶ Developing this type of tool requires finding and cleaning up data from many sources
  - ▶ Some neighborhoods have many more venues than others, but once you take the cost of living into account, neighborhoods with fewer venues can become more attractive
- 