New York City Neighborhood Search Tool

Jim Cole

October, 2019

1. Introduction

1.1 Background

According to USA Today, approximately 14% of the U.S. population moves every year. The U.S. Census Bureau estimates that about 4.2 million people live in the two New York City boroughs of Manhattan and Brooklyn, which means that it's likely that over 500,000 people move to or within those two boroughs each year.

1.2 Problem

It's often said that New York City is the "greatest city in the world," offering virtually anything a person could want, but the questions for many people are, "where in New York City are the things I'm looking for?, and "what are the most affordable neighborhoods with those amenities?" There are, of course, many web sites that list the "best neighborhoods for Mexican food" or "where to find the best jazz clubs in New York City," or "New York City's most affordable neighborhoods," but finding a web site that has researched a particular type of venue can be very difficult, and finding a web site that has researched an arbitrary combination of venue types (e.g., "lots of jazz clubs and burrito shops") is almost impossible, as is finding one that has researched multiple venue types and also factored in affordability.

This tool is designed to help answer those two key questions for any combination of venue types:

- Which neighborhoods have the largest number of venues of the types someone is looking for
- Which of those neighborhoods are most affordable.

1.3 Interest

With 500,000 people moving into or within the two boroughs, the number of potential stakeholders for such a tool is very large. This tool is personally interesting to me because I have two family members who are considering moving to New York City, and they would find this tool useful.

2. Data Acquisition and Cleansing

This project requires the following data:

- The 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 FourSquare.
- 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.

2.1 Data Sources

List of Categories

The tool first retrieves a hierarchical JSON list of all the venue categories FourSquare supports by calling FourSquare's venue categories API call. It flattens the list FourSquare returns to make it easy to look up FourSquare's category ID for any category. For example, the "Arts & Entertainment" category has a "Movie Theater" sub-category, and that in turn has three sub-categories of its own: "Drive-in Theater," "Indie Movie Theater," and "Multiplex." The tool recursively traverses the data and produces a simple list of category name/ID dictionary pairs. That is, it will turn this data:

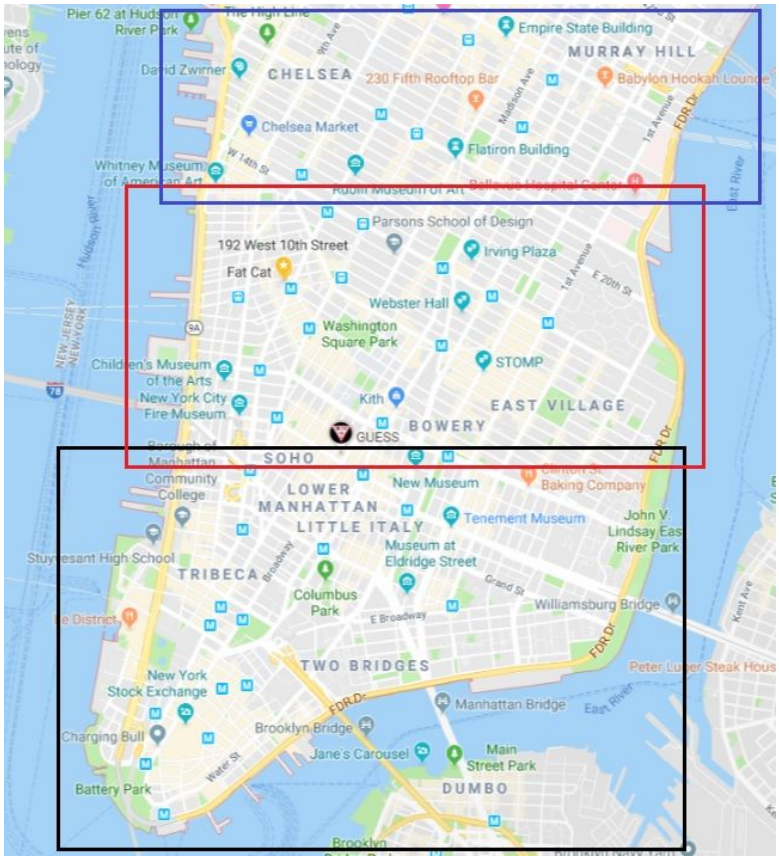
- Arts & Entertainment: 87654
    - Movie Theater: 12345
        - Drive-in Theater: 56781
        - Indie Movie Theater: 90210
        - Multiplex: 44455

Into this data:

- Arts & Entertainment: 87654
- Movie Theater: 12345
- Drive-in Theater: 56781
- Indie Movie Theater: 90210
- Multiplex: 44455

List of Venues

To retrieve venue data for the two boroughs, I used Google Maps to find the latitude and longitude of overlapping rectangles, as shown in the picture below, that encompass Manhattan and Brooklyn, and I created a CSV file containing these SW/NE pairs.



The tool uses FourSquare's venue search API call to retrieve a list of venues that match the user-specified venue category; e.g., "Jazz Club." The tool uses the Southwest (SW) and Northeast (NE) corners of a rectangular geographic area to limit the search. These rectangles are defined in the CSV file I created with data from Google maps, but because FourSquare will return a maximum of 50 venues per query, the tool has to slice those SW/NE rectangles in both directions (latitude and longitude) so that each query represents a very small geographic area.

Median Rent and Population

I found recent data on the median rent of a one bedroom apartment in different NYC neighborhoods on zumper.com, and I entered that data into a CSV file. I used data from the NYC Open Data project to find the population of each neighborhood; this data was added to the CSV file:

Borough	SubBoroNumber	Name	Population	Rent
Brooklyn	1	Williamsburg, Greenpoint	173083	2720
Brooklyn	2	Brooklyn Heights, Fort Greene	99617	2975
Brooklyn	3	Bedford Stuyvesant	152985	2200
Brooklyn	4	Bushwick	112634	2200
Brooklyn	5	East New York, Starrett City	182896	1375
Brooklyn	6	Park Slope, Carroll Gardens	104709	2555
Brooklyn	7	Sunset Park, Windsor Terrace	126230	1950

Neighborhood Zip Codes, Boundaries

The tool uses a geojson file from github containing geographic boundaries of NYC zip codes. But the tool analyzes data and presents it to the user based not on zip codes, but on neighborhoods, so zip codes have to be combined into neighborhoods. The tool uses data from the New York State Department of Health to match zip codes to NYC neighborhoods; that data is contained in a CSV file. That data was incomplete, so I used data from Property Shark to fill in missing zip codes in the CSV file:

Manhattan	1	Battery Park City, Tribeca	10282
Manhattan	1	Battery Park City, Tribeca	10285
Manhattan	2	Greenwich Village, Soho	10003
Manhattan	2	Greenwich Village, Soho	10011
Manhattan	2	Greenwich Village, Soho	10012
Manhattan	2	Greenwich Village, Soho	10014
Manhattan	3	Lower East Side, Chinatown	10002
Manhattan	3	Lower East Side, Chinatown	10009

2.2 Data Cleansing

There was a lot of data cleansing required for this project. Some data was incomplete, some data was superfluous or irrelevant, and some was duplicated.

Removing Duplicate Values

Because the SW/NE search rectangles are slightly overlapping, FourSquare returns duplicate entries in the results of different queries. The pandas method drop_duplicates easily eliminates these duplicate entries. For a query looking for “Jazz Club” and “Burrito Place,” the tool reports that it:

Dropped 3822 duplicate venues, 2665 venues remain

Missing Venue Zip Codes

Some of the venues returned by FourSquare include a street address, but not a zip code. The tool uses Bing’s reverse geocoding capability to fill in any missing zip codes in the venue data.

Eliminating Venues Outside the Two Boroughs

FourSquare returns some venues that are not in either Brooklyn or Manhattan. These irrelevant venues are returned because both boroughs are irregularly shaped, so some of the rectangles used to specify the SW and NE corners of a search area included parts of other boroughs. For instance, some of the Brooklyn rectangles in the above picture also include part of another borough, Queens, and some Upper Manhattan rectangles include part of the Bronx. The tool uses a CSV file that lists all zip codes in each neighborhood to check the zip code of each venue returned by FourSquare or Bing to eliminate venues that are not in any of the two borough’s neighborhoods.

Loosely Matching Venues

FourSquare returns venues that match a specified category, even if that category doesn't represent the main service the venue provides. For instance, FourSquare returns Madison Square Garden for almost any venue category query, because that facility is used for so many different things, from basketball games to monster truck rallies to political conventions.

This means that FourSquare will return a very large number of results for most queries, and that's undesirable for two reasons. First, Folium cannot easily produce cluster maps with thousands of points; it either takes a very long time, or fails. Second, and more importantly, many of the venues in the results won't actually match what the user is looking for. If, for instance, you query FourSquare for jazz clubs, it returns thousands of results; pretty much every music club of any type, plus restaurants that have music, plus any venue somehow related to music. But there aren't thousands of jazz clubs in Manhattan and Brooklyn, so even if the tool could show all of those results to the user, that would be misleading and not useful.

I attempted to filter out these superfluous results by looking to see if a returned venue's categoryID matched the desired venue type's categoryID. This didn't work, because FourSquare substitutes the categoryID being queried for in the query for a venue's main categoryID in the results it returns. For instance, if you queried for "Jazz Club," FourSquare returns Madison Square Garden as one of the query results, and its categoryID will match the categoryID of "Jazz Club." If you search for "Basketball Stadium," FourSquare will again return Madison Square Garden as one of the results, but this time, its categoryID will be the one for "Basketball Stadium."

So to filter out these superfluous results, the tool instead examines the name of the first FourSquare category listed for each venue, and considers that category to be the primary use of, or service provided by that venue. Thus, Madison Square Garden will show up in the final results if you are looking for venues in the category "Basketball Stadium," but not if you're looking for "Jazz Club."

The tool does one iteration through the venue data to:

- Look up missing venue zip codes
- Eliminate venues outside of the two boroughs
- Eliminate venues whose main category does not match any of the search categories

For a query for "Jazz Club" and "Burrito Place," the tool reports that most of the venues FourSquare returned did not actually match one of those categories or were outside the true search area:

> **Before processing, there are 2665 venues in our list**
>
> **151 venues kept, 2514 venues discarded, 20 zip codes fixed, 115 unique zip code**

3. Methodology

3.1 Exploratory Data Analysis

Original Plan: Compare All Counties in New York State

Originally, I planned to compare venue density between every county in New York State (NYS). The required tasks would have been:

- Divide NYS up into slightly-overlapping rectangular regions
- Call FourSquare repeatedly using the SW/NE coordinates of each rectangle
- Assign each venue to its county by looking up which county the venue's zip code was in

But I could not find a data file that listed all zip codes in all counties. Also, a quick manual analysis of FourSquare venue data for less-populated areas of NYS showed that many had very few venues of any kind, so it became apparent that comparing Manhattan (New York County) to upstate Broome county would be, as the old saying goes, like comparing apples to oranges.

Based on the lack of data for some rural areas of NYS, and the large difference in venue density between urban and suburban/rural areas, I changed the project to compare neighborhoods in two boroughs of New York City: Manhattan and Brooklyn.

Incorporating Venue Ratings

I had originally intended to incorporate venue ratings as part of the calculation of venues' "attractiveness," but the limited number of venue detail API calls allowed by FourSquare for free accounts precluded including this feature.

Technical Issue: Maximum Results per Query vs. Limited Number of FourSquare API calls per Hour and per Day

FourSquare returns at most 50 venues for any venue query. I found that depending on what type of venue you were searching for, and which area of Manhattan or Brooklyn you were searching in, FourSquare returned exactly 50 results. This happened for enough rectangular areas that it was obvious that the tool was running into the 50-venue limit, and that it needed to query smaller rectangular areas. But the smaller the area queried each time, the more API calls would be required to FourSquare. I quickly ran into the daily limit of 950 API calls, so I upgraded my FourSquare account to allow 99,500 calls per day. But if I sliced the geographic regions very small to avoid running into the 50-venue/call limit, the tool then ran afoul of the 500 call/hour limit. After trying variously-sized rectangles, I finally found a size that does not run into either limit.

Finding the Map Center

In order to locate the latitude and longitude to use as the center of the maps, I researched the geographic centers of Manhattan and Brooklyn. I determined that the geographic center of Manhattan is approximately where the Guggenheim Museum is, and that many sources consider Brooklyn College, at 2900 Bedford Avenue, to be the geographic center of Brooklyn.

The tool uses Bing's reverse geocoding API to retrieve the latitude and longitude of those two locations, then finds the point in the middle of those two points and uses that as the map center.

```
Manhattan center: [40.78292465209961, -73.95895385742188]
Brooklyn center: [40.63181, -73.953412]
Map center: [40.707367326049805, -73.95618292871094]
```

3.2 Geographic Locating

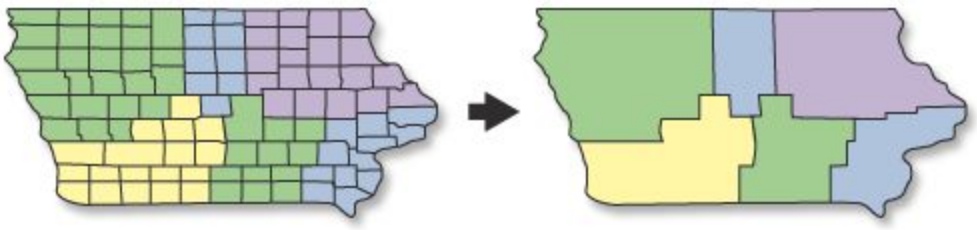
Matching Venues to Neighborhoods

Manhattan has 12 neighborhoods, and Brooklyn has 18. Using the CSV file that lists all zip codes in each neighborhood, the tool uses the zip code of each venue returned by FourSquare or found by Bing to determine which neighborhood a venue is in. For a query of "Jazz Club" and "Burrito Place," the tool reports:

- **151 venues processed**
- **5 removed due to zip codes outside the target areas**
- **51 remaining unique zip codes**

Combining Zip Codes into Neighborhoods

The tool has a geojson file giving the boundaries of every zip code in NYC, but the tool matches venue data from FourSquare to a neighborhood, not a zip code within that neighborhood. To create a geojson file so that neighborhood boundaries can be drawn on the maps, the tool uses a geospatial technique called "dissolving" to combine the polygons of multiple zip codes, eliminate the inner boundaries, and leave just the outer boundary of the combined areas, i.e., the outer boundary of each neighborhood. This picture from GIS vendor ESRI shows graphically how counties could be dissolved into regions using this technique:



To do this, the tool first reads in the geojson file, which contains boundaries for every zip code in all five boroughs of NYC. The tool discards data for any zip code outside of Manhattan or Brooklyn, which leaves 97 polygonal regions:

Before processing, 262 rows in the geopandas DataFrame

After processing, 97 rows in the geopandas DataFrame

It then does the geospatial dissolve, resulting in a geopandas DataFrame containing the boundaries of 30 neighborhoods:

Merged the geographic boundaries of 97 zip codes into 30 neighborhoods

3.3 Metric Calculation

Now that the tool has determined:

- Which venues really match the categories the user is looking for
- Which neighborhood each venue is in
- The population of each neighborhood
- The median rent of each neighborhood

it can calculate each neighborhood's *venue density* and *attractiveness.*

Venue Density

A neighborhood's *venue density* is the number of venues in the categories the user has specified per 10,000 residents of each neighborhood

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

Neighborhood Attractiveness

The "attractiveness" of a neighborhood is also based on venue density, but factors in the relative cost of living in that neighborhood versus living in other Manhattan or Brooklyn neighborhoods. A *rent factor* is calculated for each neighborhood; it is the cube of the multiple of a neighborhood's rent versus the rent in the cheapest neighborhood that has venues in the specified categories.

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

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

The *attractiveness* value is then normalized to fall between 0.0 and 1.0:

$$normalizedAttractiveness\ =\ \frac{attractiveness-min(attractiveness)}{max(attractiveness)-min(attractiveness)}$$

3.4 Machine Learning

So far, the tool has focused on identifying neighborhoods with many venues and affordable rent. But groupings of venues don't necessarily respect neighborhood boundaries. There could be a cluster of jazz clubs that are near one another, but are in two different neighborhoods. E.g., a club on the west side of a major street could be in neighborhood A, while a club right across the street is in neighborhood B. The tool uses machine learning, specifically k-means clustering, to identify clusters of venues regardless of neighborhood boundaries.

Since there are 30 neighborhoods in Manhattan and Brooklyn, and since this cluster analysis is supposed to identify groups of venues that cross neighborhood boundaries, I chose 15 as the number of clusters to group the venues in.

3.5 Mapping

Since the tool is answering geographic questions, it also shows the computed information on three types of maps:

- A choropleth map of the neighborhoods, shaded based on their *attractiveness* value.
- An overlay of neighborhood boundaries on a heatmap of venues, making it possible to visually compare the density of the selected venue categories among the neighborhoods.
- A cluster map of venues, showing:
    - The location of each venue, color-coded by venue category
    - A circle from the centroid of each cluster, with each circle's radius proportional to the number of venues in that cluster.

  This map enables visual exploration of clusters of venues, regardless of neighborhood boundaries.

To display the neighborhood boundaries on the maps, the geojson boundary information had to be converted to the World Geographic System 1984 coordinate system. To enable the user to see which neighborhood is which, a popup marker with the neighborhood name is added at the centroid of each neighborhood.

4. Results & Discussion

Rent

The median cost of a 1-bedroom apartment varies widely between neighborhoods:

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

Borough	Name	Rent	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			

It costs more than twice as much to live in several Manhattan neighborhoods as it does to live in some Brooklyn neighborhoods. If cost was the only factor for someone moving to NYC, they might concentrate on finding housing in Canarsie or East Flatbush. But most people want to live in an area with certain amenities, and that's where the data this tool analyzes and presents is helpful.

Venue Density

Venue density is the first piece of information this tool provides to people moving to NYC. It goes beyond a simple count of the number of venues in a neighborhood by taking the neighborhood's population into account to calculate a per-capita number of venues, the venue density. This approach was chosen because neighborhood 'A' might have twice as many jazz clubs and burrito shops as neighborhood 'B', but 'A' also has twice as many people as 'B', so on a per-capita basis, the two neighborhoods are relatively equivalent.

For "Jazz Club" and "Burrito Place," the tool finds these neighborhoods have the highest venue density:

The 10 neighborhoods with the highest density of venues in the categories ['Jazz Club', 'Burrito Place'] are:

	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

You can see that the Chelsea, Clinton neighborhood has twice as many venues as the Midtown Business District, but also twice as many people; thus, they are fairly equivalent for Jazz Clubs and Burrito Places on a per-capita basis.

Neighborhood Attractiveness

High venue density of the types of venues a person is seeking is great, but that metric doesn't take the cost of living into account. The *attractiveness* value does take rent into account. For "Jazz Club" and "Burrito Place," the tool finds these neighborhoods have the highest attractiveness value::

The 10 neighborhoods with the highest attractiveness ratings in the categories ['Jazz Club', 'Burrito Place'] are:

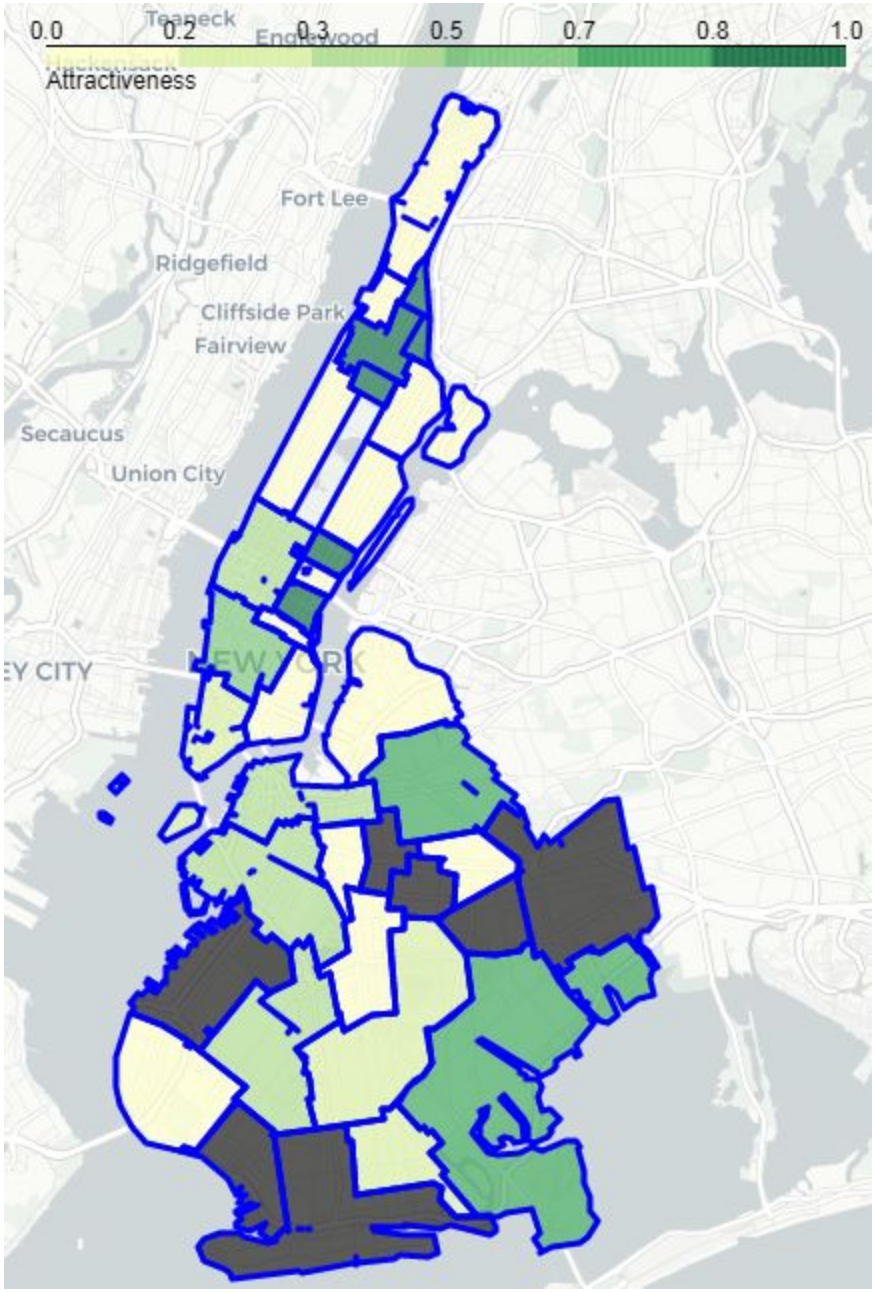
	Borough	SubBoroNumber	Name	Population	Rent	VenueCount	VenueDensity	Attractiveness	RentFactor
27	Manhattan	10	Central Harlem	115723	2010	7.0	0.604893	1.000000	3.123777
22	Manhattan	5	Midtown Business District	51673	3125	11.0	2.128771	0.926197	11.739294
3	Brooklyn	4	Bushwick	112634	2200	7.0	0.621482	0.748602	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

So taking into account the cost of housing, Central Harlem has moved from 7th place on the venue density list to 1st place on the attractiveness list. Bushwick, in Brooklyn, where the rent is generally less expensive than in Manhattan, has moved from 6th up to 3rd, and Canarsie/Flatlands in Brooklyn, which wasn't even in the top 10 for venue density, is 4th on the

list for attractiveness. If a jazz and burrito lover can afford to live in Chelsea, great, but if not, they would do well to consider Harlem and Bushwick.

Choropleth Map

The choropleth map shades each neighborhood according to its attractiveness, on a normalized scale of 0.0 to 1.0. You can see that some neighborhoods in Brooklyn have no jazz clubs or burrito shops, and thus have an attractiveness rating of zero, and are darkly shaded.

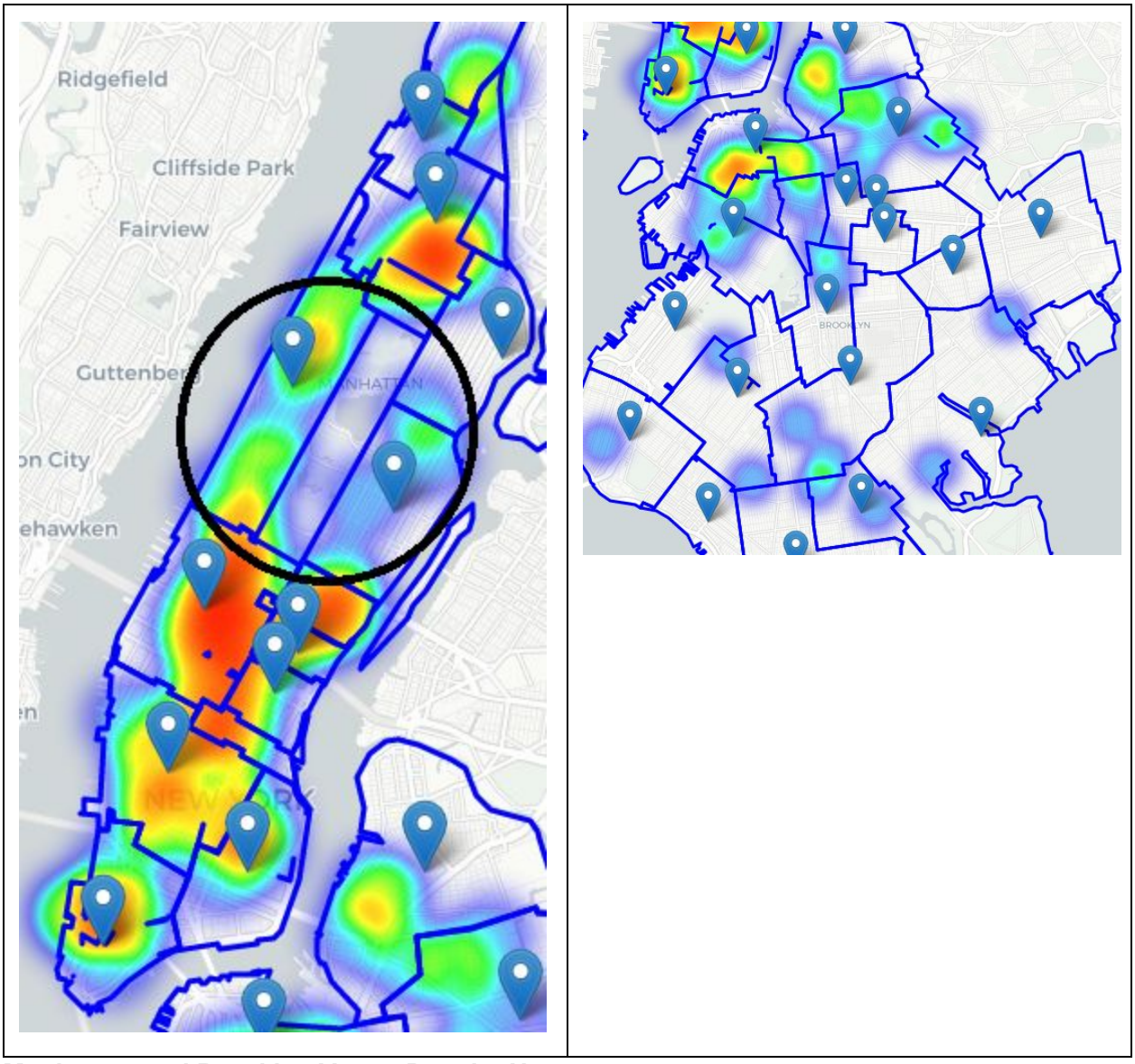


Choropleth Map of Neighborhood Attractiveness

Looking at the choropleth map, it appears that the neighborhoods just west and east of Central Park (the long, thin, empty rectangle in the middle of Manhattan), are not the most affordable places for a jazz and burrito lover to live. But that map doesn't answer the question of whether those neighborhoods have a lot of jazz clubs and burrito shops. The next map will help answer that question.

Heatmap

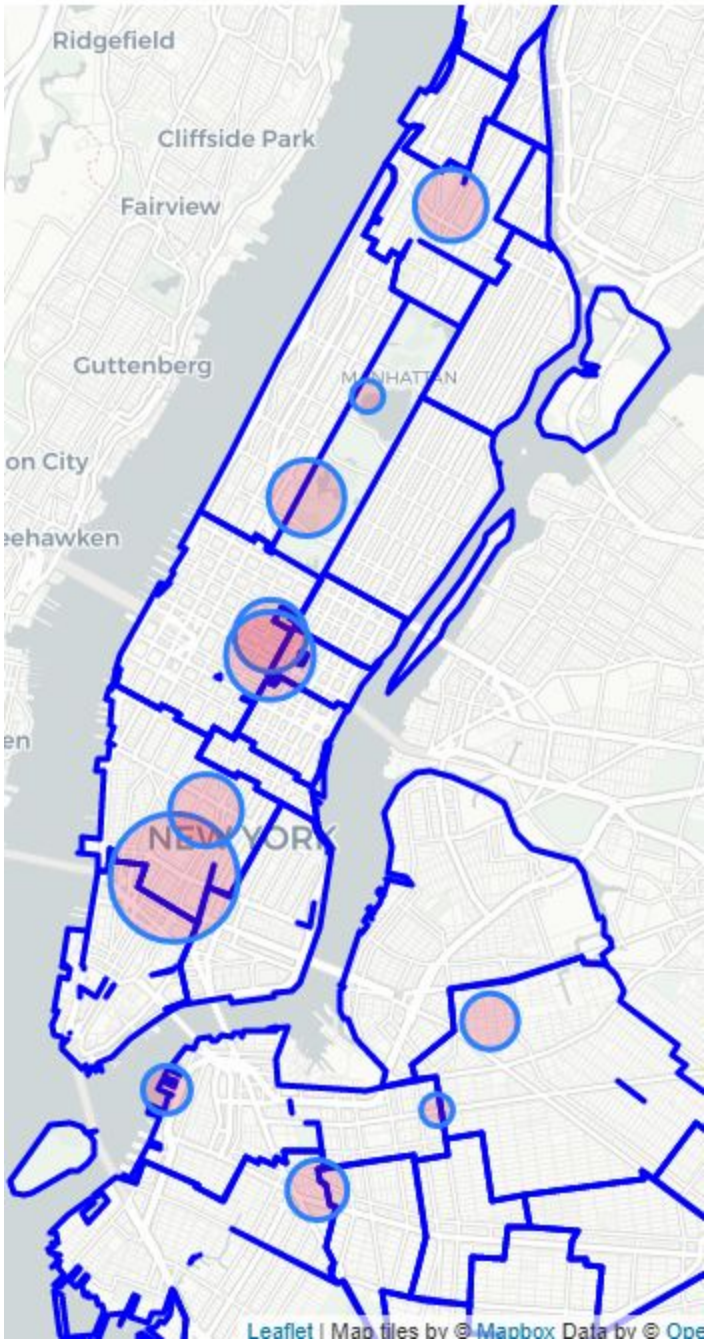
The second map is a heatmap that shows venue density within neighborhoods. You can see several areas with high concentrations of jazz clubs and burrito shops, including Central Harlem, just north of Central Park. This makes sense because Harlem has been known as a center of jazz music for over 100 years. On this map, you can also see that the neighborhoods next to Central Park (inside the black circle), which were lightly shaded on the choropleth map because rents are expensive, do have jazz and burrito venues.



Manhattan and Brooklyn Venue Density Heatmaps

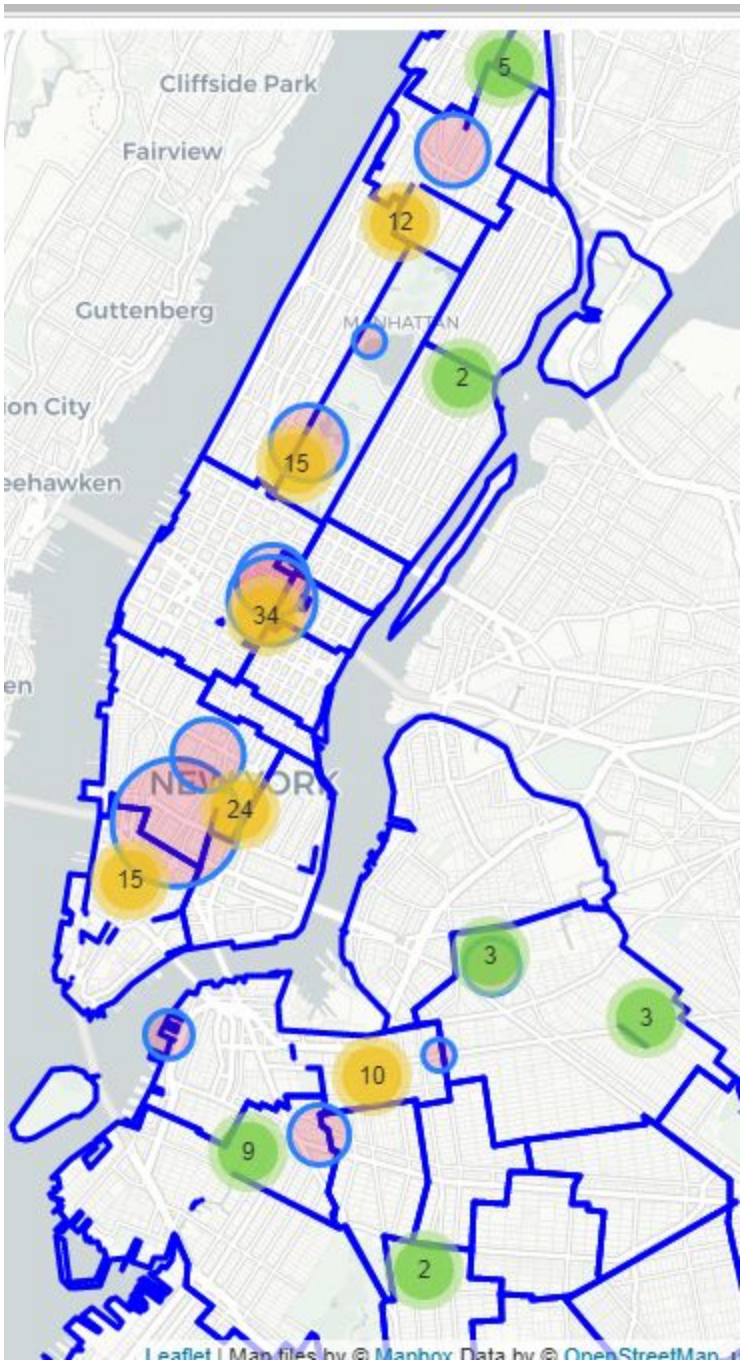
Cluster Map

A user might say, “Forget about neighborhood boundaries, and forget about per-capita venue density - just show me where there are clusters of the types of venues I’m looking for.” Using k-means clustering, the tool does exactly that, producing a cluster map with the size of each circle proportional to the number of venues in that cluster. It’s easy to see that the largest cluster of jazz clubs and burrito shops is in Greenwich Village, and that there are also sizable clusters on the West Side of Central Park, and in Harlem, and smaller clusters in the parts of Brooklyn closest to Manhattan.

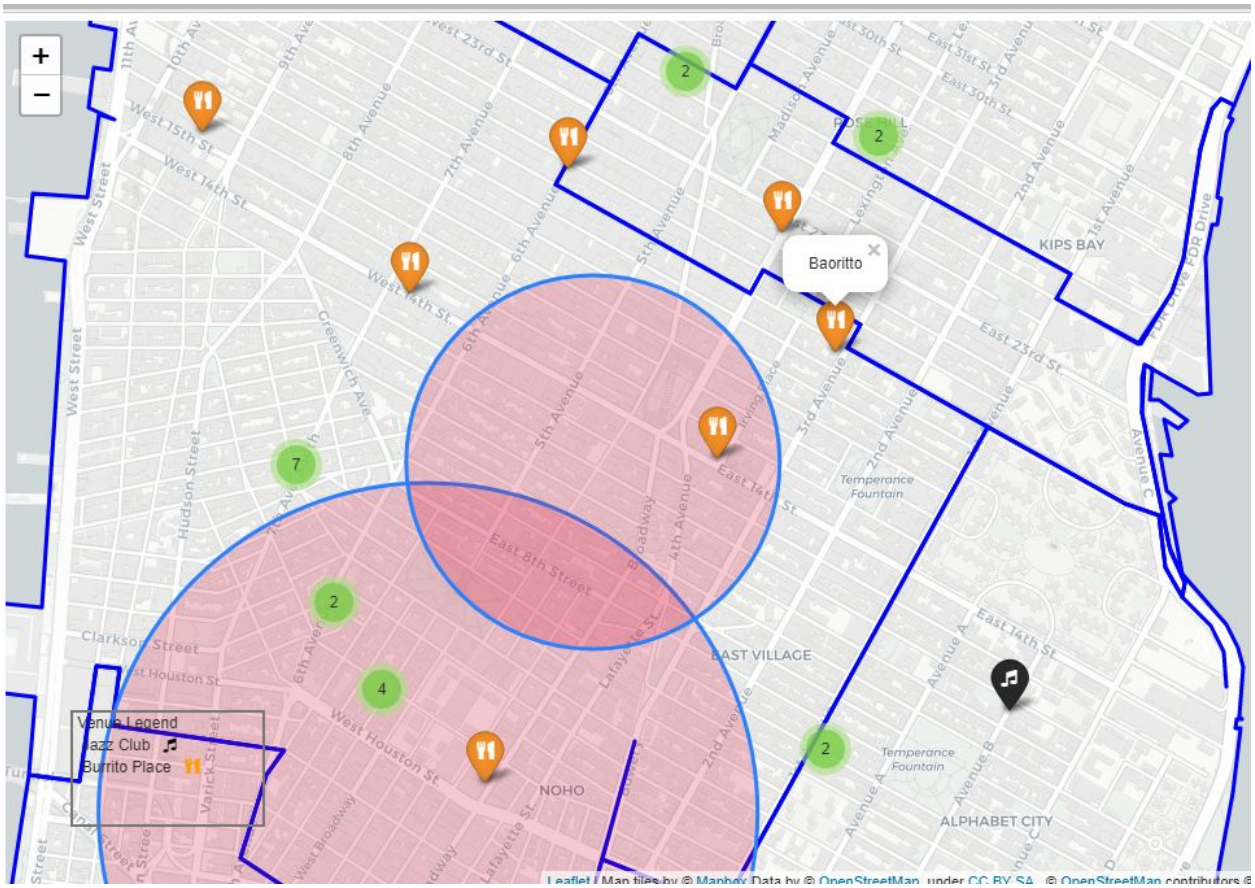


Venue cluster map

The cluster map becomes much more useful once all of the venues are added to it. This final map allows the user to see how many venues are clustered together.

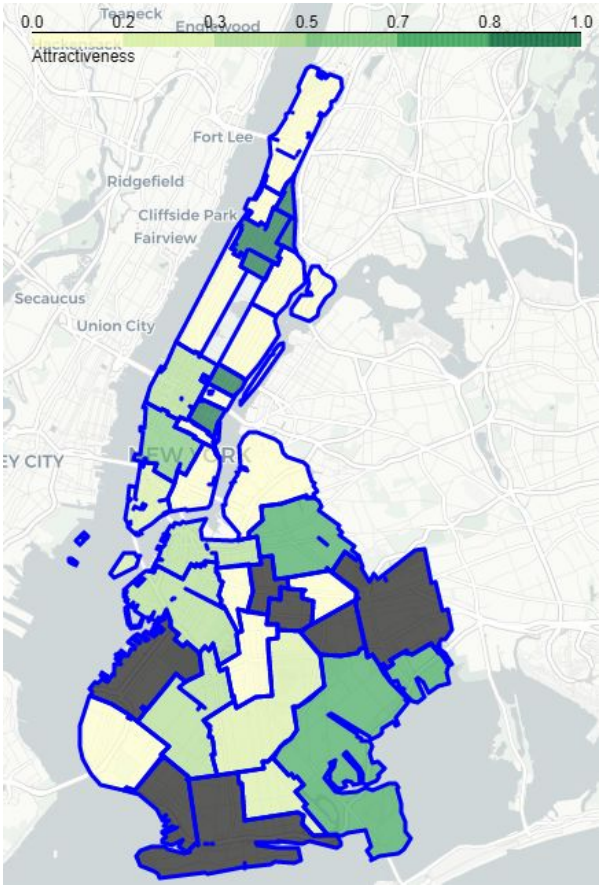


It also enables the user to zoom in and explore the individual venues. The different types of venues have different color pins (orange for burrito places, black for jazz clubs), and the pins' icons also identify the type of venue.

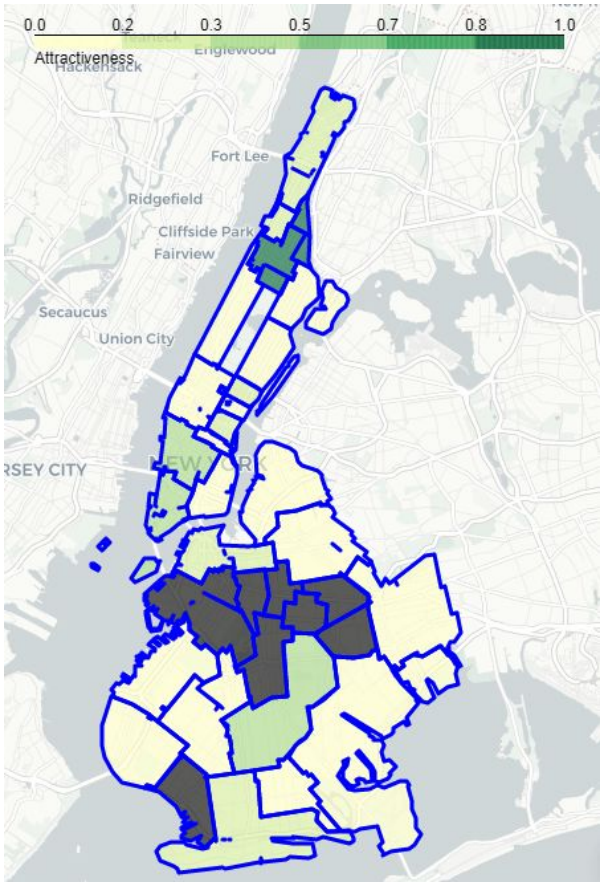


Examining Other Venue Choices

Since the tool is designed to enable users to research the density and attractiveness (in terms of rent) of any venue categories that FourSquare supports, it's interesting to compare the venue combination we've been examining, Jazz Club and Burrito Place, with a different combination. The figure below shows the choropleth "attractiveness" neighborhood map for two different venue combinations. The Manhattan neighborhoods are fairly similar, but the Brooklyn neighborhoods are quite different.



Jazz Club & Burrito Shop



College Gym & Law School

6. Conclusions

Creating a tool to answer geographic questions requires finding, merging, cleaning up, and analyzing data from disparate sources. This is true even if the metrics a tool calculates are fairly simple, as they are for this tool. But it can be worth it, because data science and machine learning techniques can provide insights for geographic questions that are very difficult to obtain via other techniques.

Developing and using this tool has shown me that certain neighborhoods have far more venues than other neighborhoods. I kept having to slice the SW/NE latitude/longitude rectangles smaller and smaller for parts of Manhattan and Brooklyn, because the geographic region passed to FourSquare had to be very small to avoid running into the limit of 50 venues returned per query, while other parts of the two boroughs returned very few venues for any categories I tested.

Is the old adage “you get what you pay for” true? Yes and no. For the categories tested, the tool determined that some neighborhoods have many more venues than other neighborhoods on an absolute basis and even on a per-capita basis, but when you factor in the cost of rental housing,

some of the less-expensive neighborhoods score better. As you would expect, exactly which neighborhoods are most “attractive” depends in part on what venue categories you’re seeking.

6.1 Going Forward

There are several ways this tool could be enhanced to make it even more useful:

- Incorporate FourSquare ratings into the attractiveness calculation; this was not possible due to the limited number of venue detail API calls allowed for a free FourSquare account.
- Include other NYC boroughs, such as Queens.
- Make it easier to add categories to the comparison “on the fly,” rather than having to run the script again from the beginning, as it the case now.