How “pizza” is New York?

Nasser Santiago Boan

November 30, 2019

1. Introduction

1.1 Background

New York is a huge multicultural city where millions of people commute, eat and work everyday. In the city we can find an almost infinite assortment of restaurants and behind each one a story. Different groups of immigrants came to America trying to improve their lives and introducing us to their cousine was the way they found, not only to rewrite their history, but also to enrich ours.

Italians are a big part of this immigrant chapter of NYC, but by 1860 the city had only 1400 new yorkers with some Italian descent, they were working as dockworkers, fruit vendors, organ grinders or rag pickers. In the late 1860's a wave of immigration from Italy began and culminated in a scenario where from 1900 to 1914 almost two million Italians immigrated to all of America, including NY. By 1930 NYC was home to one million Italian Americans1.

Pizza is an Italian dish largely appreciated in the whole world and it was brought to America by Italian immigrants. The first pizzeria in America was Lombardi's, by Genaro Lombardi in Manhattan's Little Italy in 1905. In 1924 Tottono's opened up in Coney Island and in 1933 Patsy's opened in Harlem, and all the three restaurants are still open today2.

Living a huge city like NY can be stressful, mood and anxiety disorders are the most prevalent mental health problems among city dwellers3. Dealing with stress, work, bills is a daunting task and most of us try to find a way to cope with everything, some turn to meditation, some turn to addiction, some turn to games and some turn to eating a good delicious meal. Pizza became a part of every new yorker day to day routine, it can be a fast meal for a busy day or something you eat to relax with friends, either way it has an weight in peoples lives.

1.2 Problem

Eating a good slice of pizza is a stress release to some people and finding the right place is crucial to that. I will define a standardized way to evaluate how ‘pizza’ a place can be using data collected from FourSquare about pizzerias. My analysis will be bound to NTA (Neighborhood tabulation areas) that have at least one pizzeria within it. Finally I’ll compare each NTA using a clustering algorithm to find distinguishable groups that can give us valuable information about the popularity and grading of a NTA based on their pizzerias.

Business problem: How different are the NYC pizzerias? Where can I find the most popular pizzerias in NYC? Where can I find the most rated pizza in NYC?

1.3 Interest

Entrepreneurs can take the findings in this project into considerations when defining the best place to open their pizza places. People can search and define the best course for a pizzeria crawl over NYC, tasting different pizzas and experience the NYC pizza culture. Finally, users can search for neighborhoods with great pizza places near and far from home.

2. Data acquisition

2.1 Data Sources

This project gathered data from two sources:

a. FourSquare API (https://developer.foursquare.com/places);

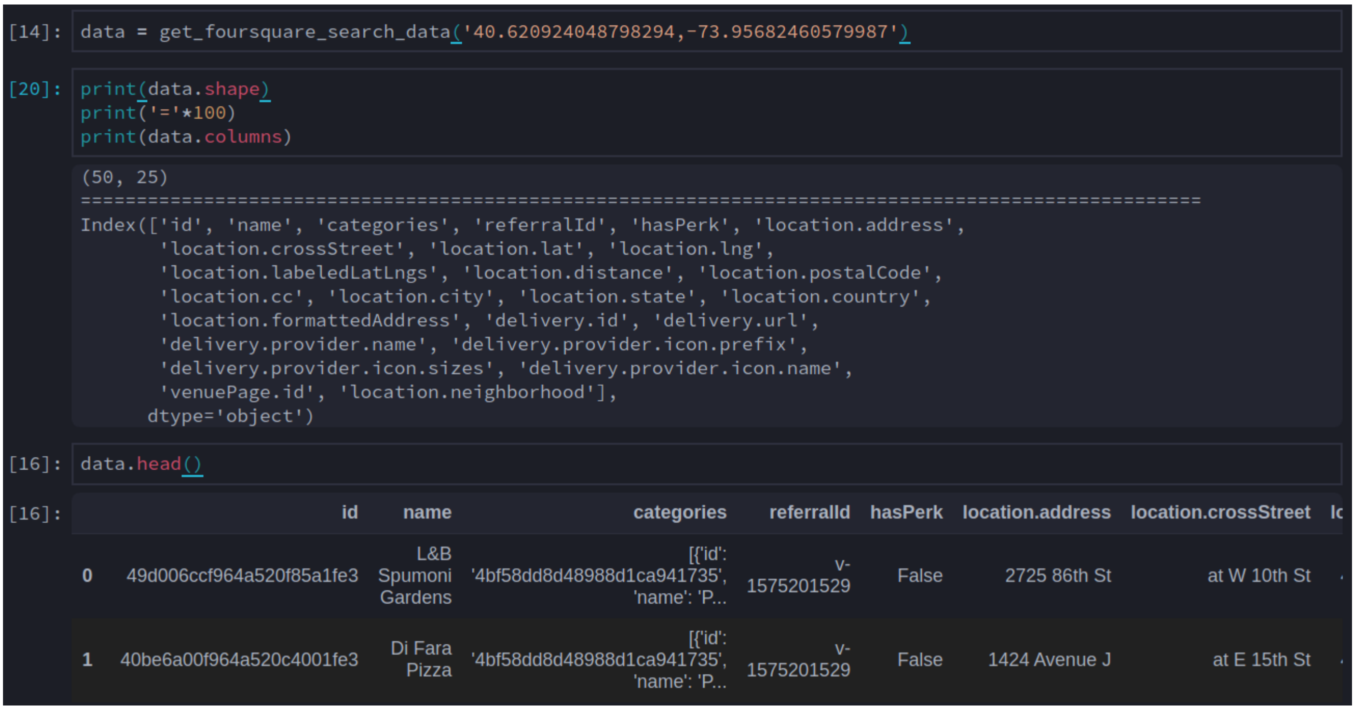
b. NYC OpenData (https://data.cityofnewyork.us/City-Government/NTA-map/d3qk-pfyz)

Foursquare defines itself as “a location technology platform dedicated to improving how people move through the real world”. In 2009 Foursquare invented the check-in, a way for people to rate and review and share places where they’ve been, as of today the Foursquare platform has almost 13+ billion check-ins. This geolocated data is a gold mine of information and is being used by Uber, Tencent, Apple, Samsung and Twitter.

> *“If it tells you where,*
> *it's probably built on Foursquare.”*
> - Foursquare website.

This project made use of two endpoints from Foursquare API: the “search” (regular call) endpoint and the “details” (premium call) endpoint. The first endpoint gives you venues information (up to 50 venues on each call) around a specified location using a latitude, a longitude, a radius and some other information, here is the full documentation. The second gives you details about an specific venue using it’s ID, the information returned from this endpoint can be found in the venue page from Foursquare, here is the endpoint full documentation.

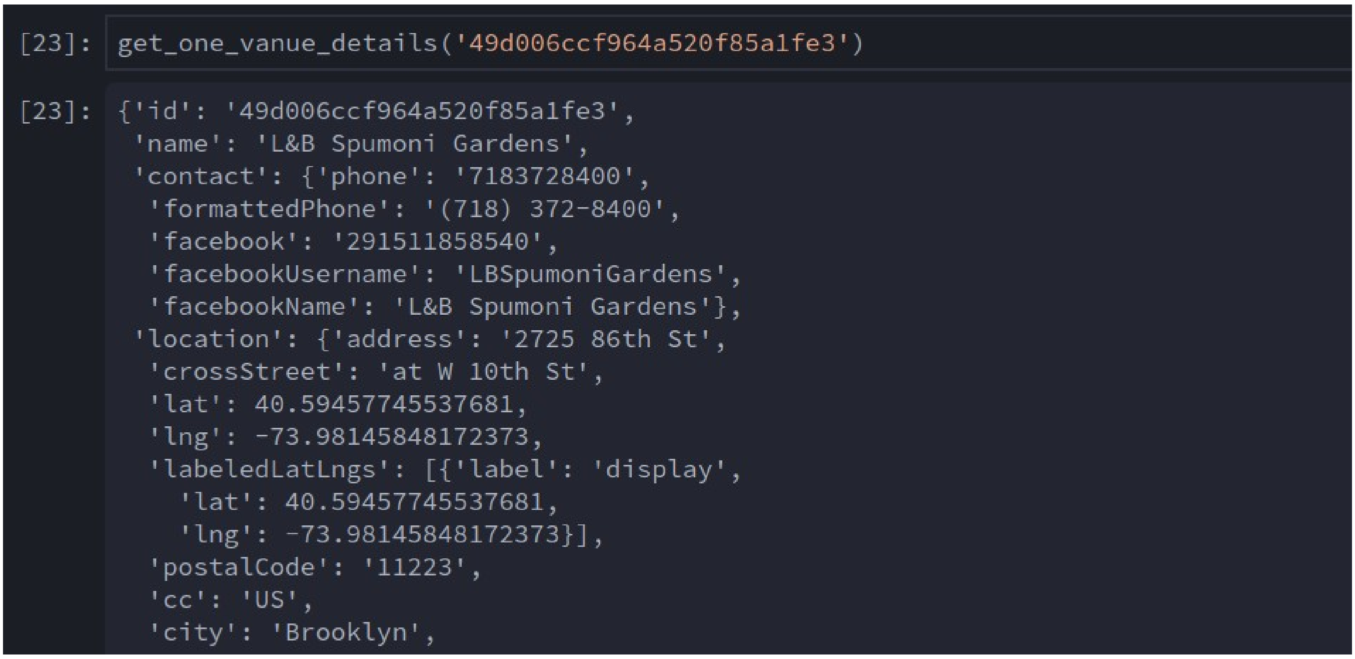
Here’s is an example using the search endpoint:



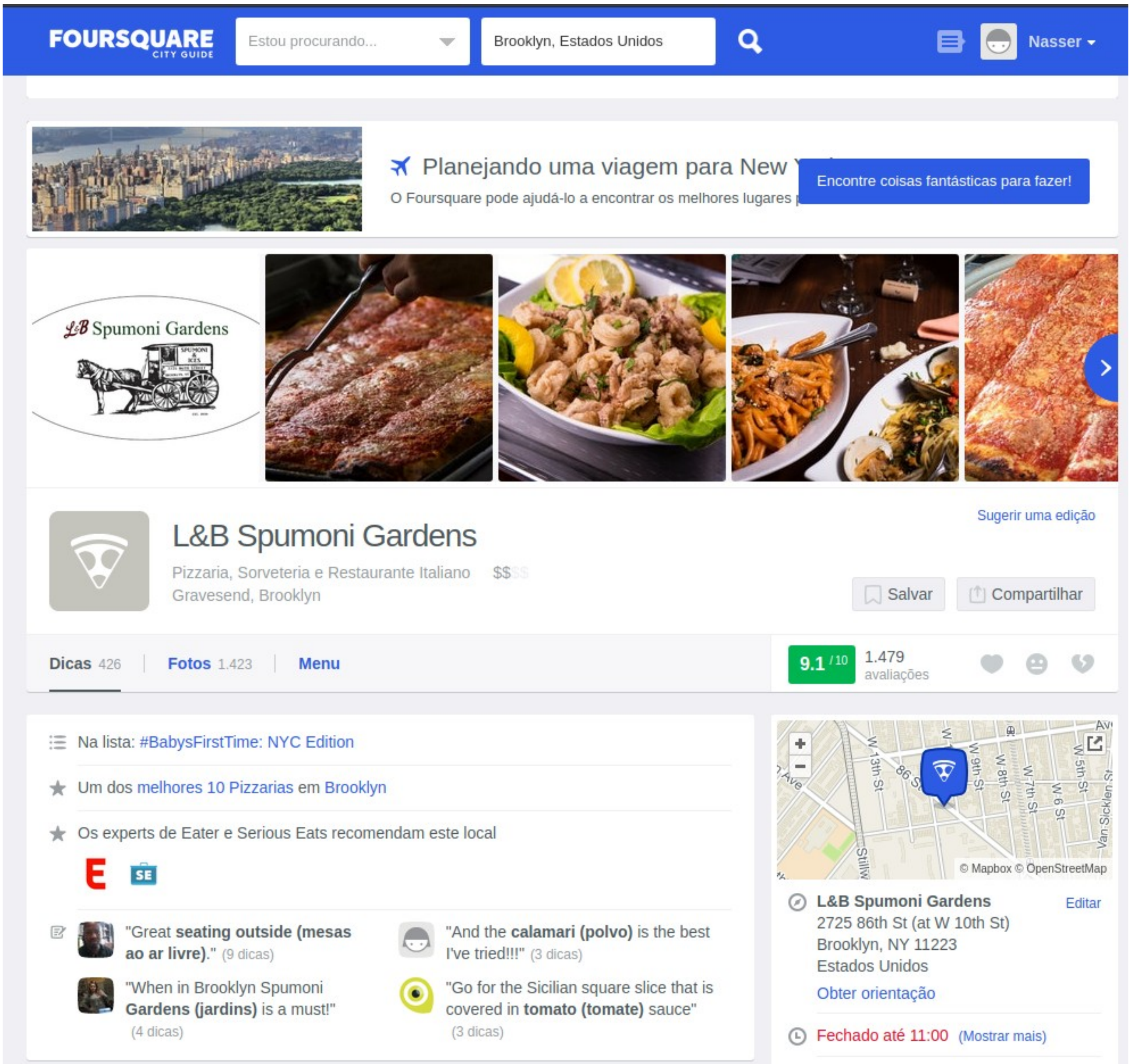
```
[14]: data = get_foursquare_search_data('40.620924048798294,-73.95682460579987')

[20]: print(data.shape)
      print('='*100)
      print(data.columns)

      (50, 25)
      ====================================================================================================
      Index(['id', 'name', 'categories', 'referralId', 'hasPerk', 'location.address',
             'location.crossStreet', 'location.lat', 'location.lng',
             'location.labeledLatLngs', 'location.distance', 'location.postalCode',
             'location.cc', 'location.city', 'location.state', 'location.country',
             'location.formattedAddress', 'delivery.id', 'delivery.url',
             'delivery.provider.name', 'delivery.provider.icon.prefix',
             'delivery.provider.icon.sizes', 'delivery.provider.icon.name',
             'venuePage.id', 'location.neighborhood'],
            dtype='object')

[16]: data.head()

[16]:
```

	id	name	categories	referralId	hasPerk	location.address	location.crossStreet	lc
0	49d006ccf964a520f85a1fe3	L&B Spumoni Gardens	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	v-1575201529	False	2725 86th St	at W 10th St	
1	40be6a00f964a520c4001fe3	Di Fara Pizza	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	v-1575201529	False	1424 Avenue J	at E 15th St	

Here’s an example using the details endpoint:



```
[23]: get_one_vanue_details('49d006ccf964a520f85a1fe3')

[23]: {'id': '49d006ccf964a520f85a1fe3',
       'name': 'L&B Spumoni Gardens',
       'contact': {'phone': '7183728400',
        'formattedPhone': '(718) 372-8400',
        'facebook': '291511858540',
        'facebookUsername': 'LBSpumoniGardens',
        'facebookName': 'L&B Spumoni Gardens'},
       'location': {'address': '2725 86th St',
        'crossStreet': 'at W 10th St',
        'lat': 40.59457745537681,
        'lng': -73.98145848172373,
        'labeledLatLngs': [{'label': 'display',
          'lat': 40.59457745537681,
          'lng': -73.98145848172373}],
        'postalCode': '11223',
        'cc': 'US',
        'city': 'Brooklyn',
```



FOURSQUARE CITY GUIDE | Estou procurando... | Brooklyn, Estados Unidos | Nasser

Planejando uma viagem para New Y

O Foursquare pode ajudá-lo a encontrar os melhores lugares p

Encontre coisas fantásticas para fazer!

L&B Spumoni Gardens

Sugerir uma edição

L&B Spumoni Gardens

Pizzaria, Sorveteria e Restaurante Italiano $$$$

Gravesend, Brooklyn

Salvar | Compartilhar

Dicas 426 | Fotos 1.423 | Menu

9.1 /10 | 1.479 avaliações

Na lista: #BabysFirstTime: NYC Edition

Um dos melhores 10 Pizzarias em Brooklyn

Os experts de Eater e Serious Eats recomendam este local

"Great **seating outside (mesas ao ar livre)**." (9 dicas)

"And the **calamari (polvo)** is the best I've tried!!!" (3 dicas)

"When in Brooklyn Spumoni **Gardens (jardins)** is a must!" (4 dicas)

"Go for the Sicilian square slice that is covered in **tomato (tomate)** sauce" (3 dicas)

L&B Spumoni Gardens | Editar

2725 86th St (at W 10th St)
Brooklyn, NY 11223
Estados Unidos

Obter orientação

Fechado até 11:00 (Mostrar mais)

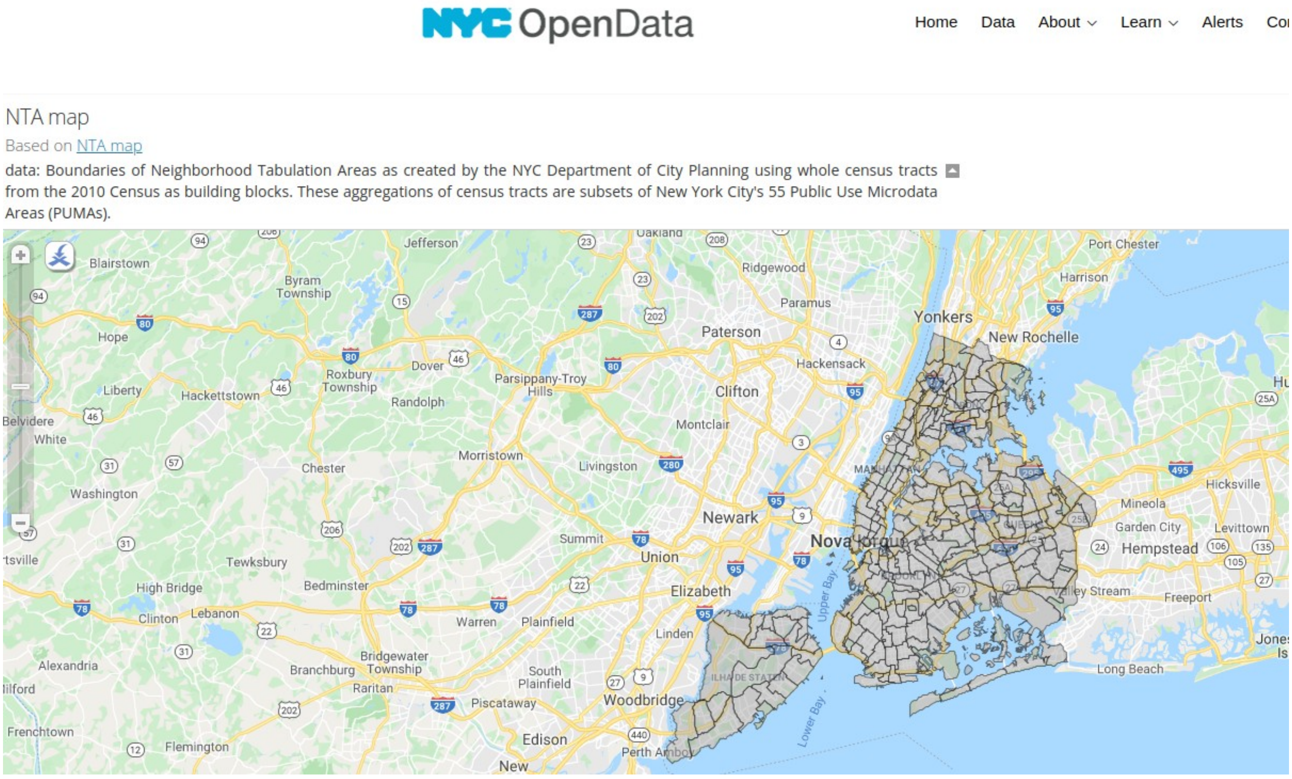
Complete JSON response here: https://pastebin.com/WxtnKVB8

Note that the ‘search’ endpoint returns information about 50 venues per call using a specific latitude and longitude and I’ve configured the request to only consider pizzerias. If I keep calling this endpoint with the same latitude and longitude it will return me same set of 50 venues. The ‘details’ returns me a lot of information, but only from one venue at a time.

The Foursquare API was used in the ‘PERSONAL’ tier, meaning that I could make 99,500 regular (‘search’ endpoint) calls and up to 500 premium (‘details’ endpoint) calls per day. The fact that I could only make 500 calls per day to get the venues’ details and that I had to use different latitude and longitude to get data from different pizzerias were the two main challenges from this project.

The NYC OpenData is a free repository of data published by NYC agencies and other partners. It has datasets from almost 50 different agencies, everything free and public. I used the NTA map for this project.

According to the Department of City Planning of NYC NTAs “*were created to project populations at a small area level, from 2000 to 2030 for PlaNYC, the long-term sustainability plan* *for New York City. Since population size affects the error associated with population projections, these geographic units needed to have a minimum population, which we determined to be 15,000".*



NYC OpenData

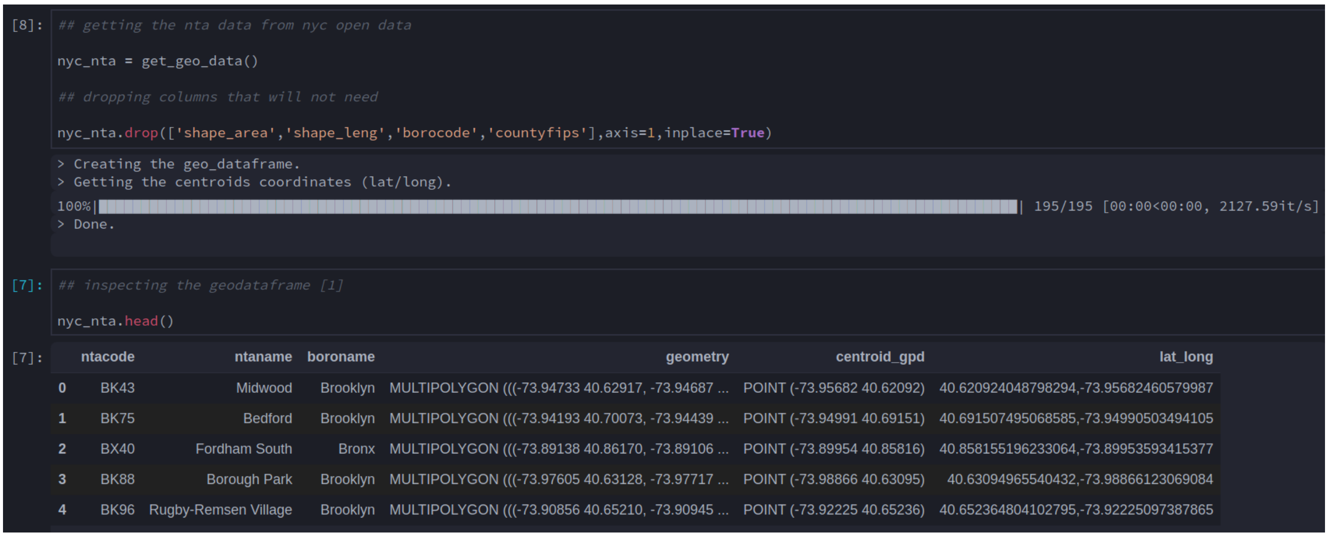
Home Data About ⌵ Learn ⌵ Alerts Co

NTA map

Based on NTA map

data: Boundaries of Neighborhood Tabulation Areas as created by the NYC Department of City Planning using whole census tracts from the 2010 Census as building blocks. These aggregations of census tracts are subsets of New York City's 55 Public Use Microdata Areas (PUMAs).

The NTA geographic multipolygons are being used to determine the unit of analysis for this project meaning that in the end I'll have information about each NTA as rows of my final dataset. I'll will also use the each NTA centroid's latitude and longitude to gather information. Here's how the NTA data is being gathered (geopandas' geodataframe):



[8]:

```
## getting the nta data from nyc open data

nyc_nta = get_geo_data()

## dropping columns that will not need

nyc_nta.drop(['shape_area','shape_leng','borocode','countyfips'],axis=1,inplace=True)
```

```
> Creating the geo_dataframe.
> Getting the centroids coordinates (lat/long).
100%|████████████████████████████████████████████████████████████████████████████| 195/195 [00:00<00:00, 2127.59it/s]
> Done.
```

[7]:

```
## inspecting the geodataframe [1]

nyc_nta.head()
```

[7]:

	ntacode	ntaname	boroname	geometry	centroid_gpd	lat_long
0	BK43	Midwood	Brooklyn	MULTIPOLYGON (((-73.94733 40.62917, -73.94687 ...	POINT (-73.95682 40.62092)	40.620924048798294,-73.95682460579987
1	BK75	Bedford	Brooklyn	MULTIPOLYGON (((-73.94193 40.70073, -73.94439 ...	POINT (-73.94991 40.69151)	40.691507495068585,-73.94990503494105
2	BX40	Fordham South	Bronx	MULTIPOLYGON (((-73.89138 40.86170, -73.89106 ...	POINT (-73.89954 40.85816)	40.858155196233064,-73.89953593415377
3	BK88	Borough Park	Brooklyn	MULTIPOLYGON (((-73.97605 40.63128, -73.97717 ...	POINT (-73.98866 40.63095)	40.63094965540432,-73.98866123069084
4	BK96	Rugby-Remsen Village	Brooklyn	MULTIPOLYGON (((-73.90856 40.65210, -73.90945 ...	POINT (-73.92225 40.65236)	40.652364804102795,-73.92225097387865

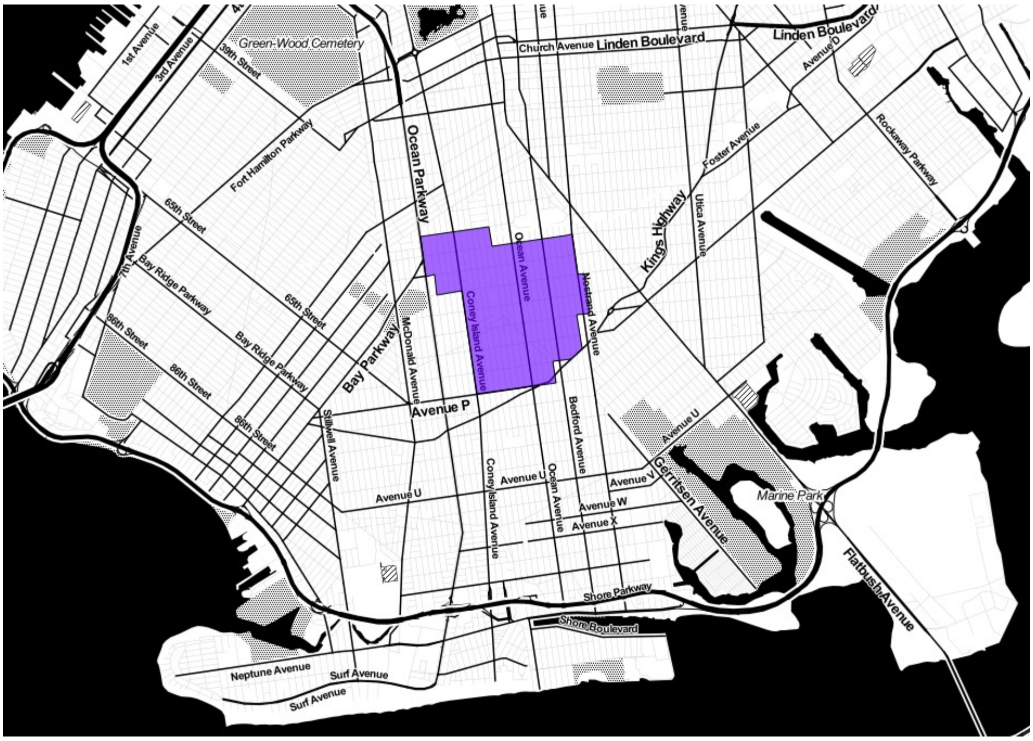
2.2 Acquisition and Cleaning

2.1.1 Getting Data – Stage 1

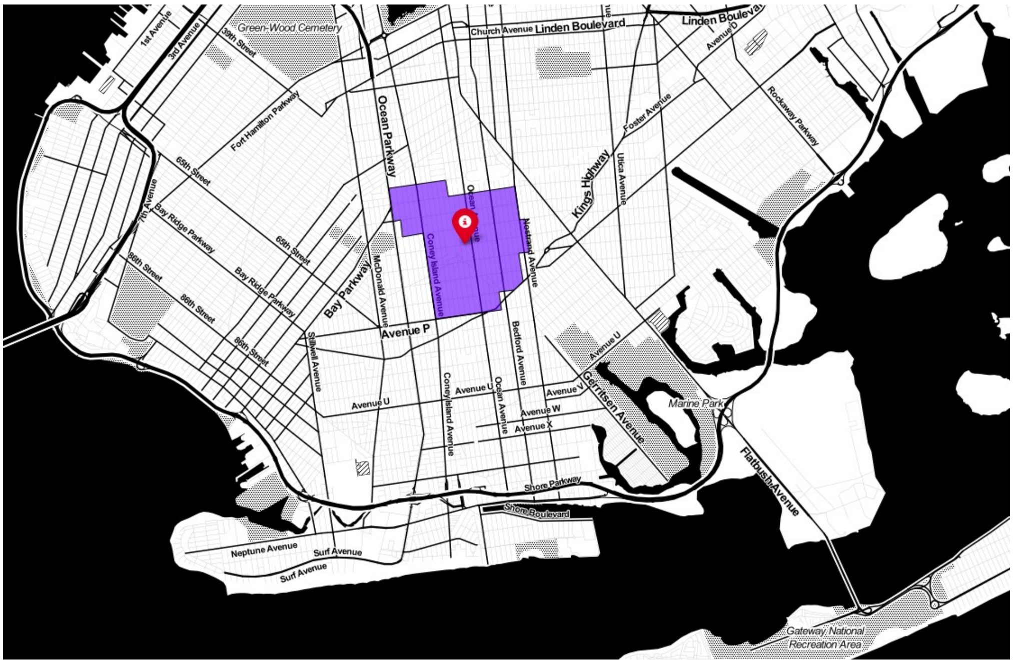
This stage's main objective is to get the most amount of data possible about venues from NYC. To do that my approach was to use each NTA centroid's latitude and longitude and pass both to Foursquare 'search' endpoint.

Here's an example from a NTA called Midwood in Brooklyn:

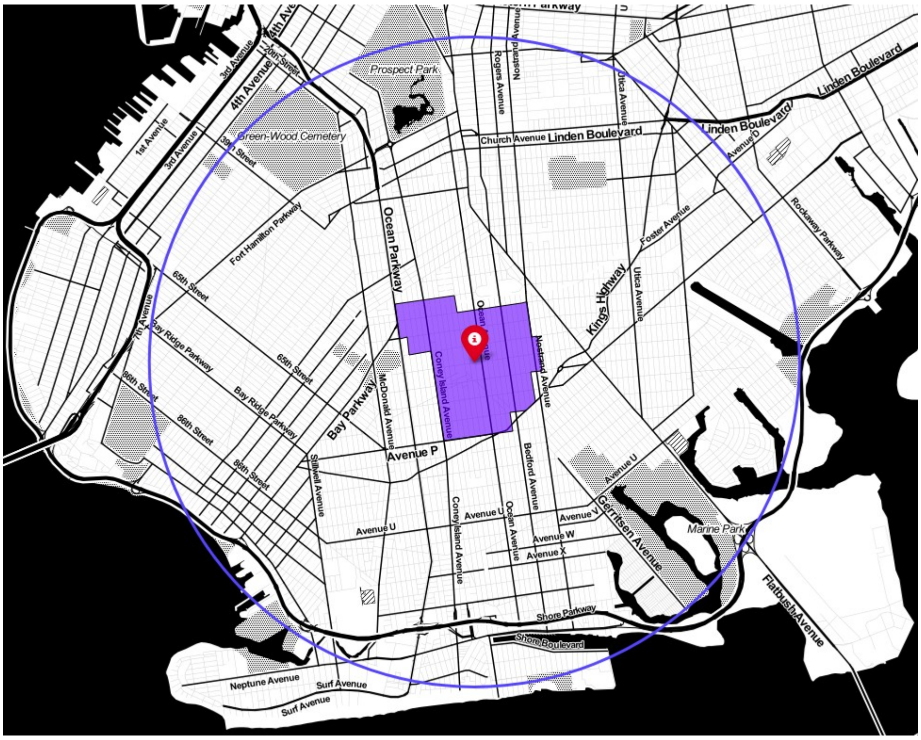
a. Choropleth



b. Choropleth + Centroid Marker

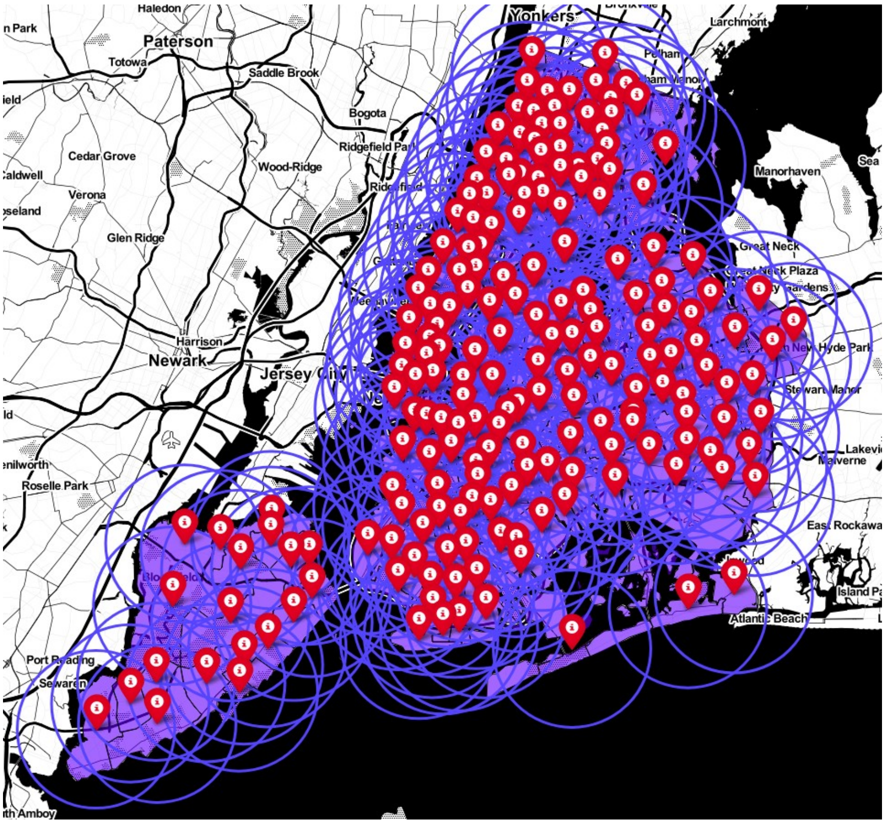


c. Choropleth + Centroid Marker + 5 km radius



Doing that for all of the 195 NTA:

d. Choropleth + Centroid Marker + 5 km radius (all NTAs)



This last map is quite unreadable, I know. But it shows that using the 5km radius from the NTA’s centroid helps me get all the possible pizzerias in the NYC area.

Using this radius I was able to use the ‘search’ endpoint and get 7882 pizzerias, most of them duplicated due to overlapping radius. However I achieved this first stage objective and got a lot of pizzerias.

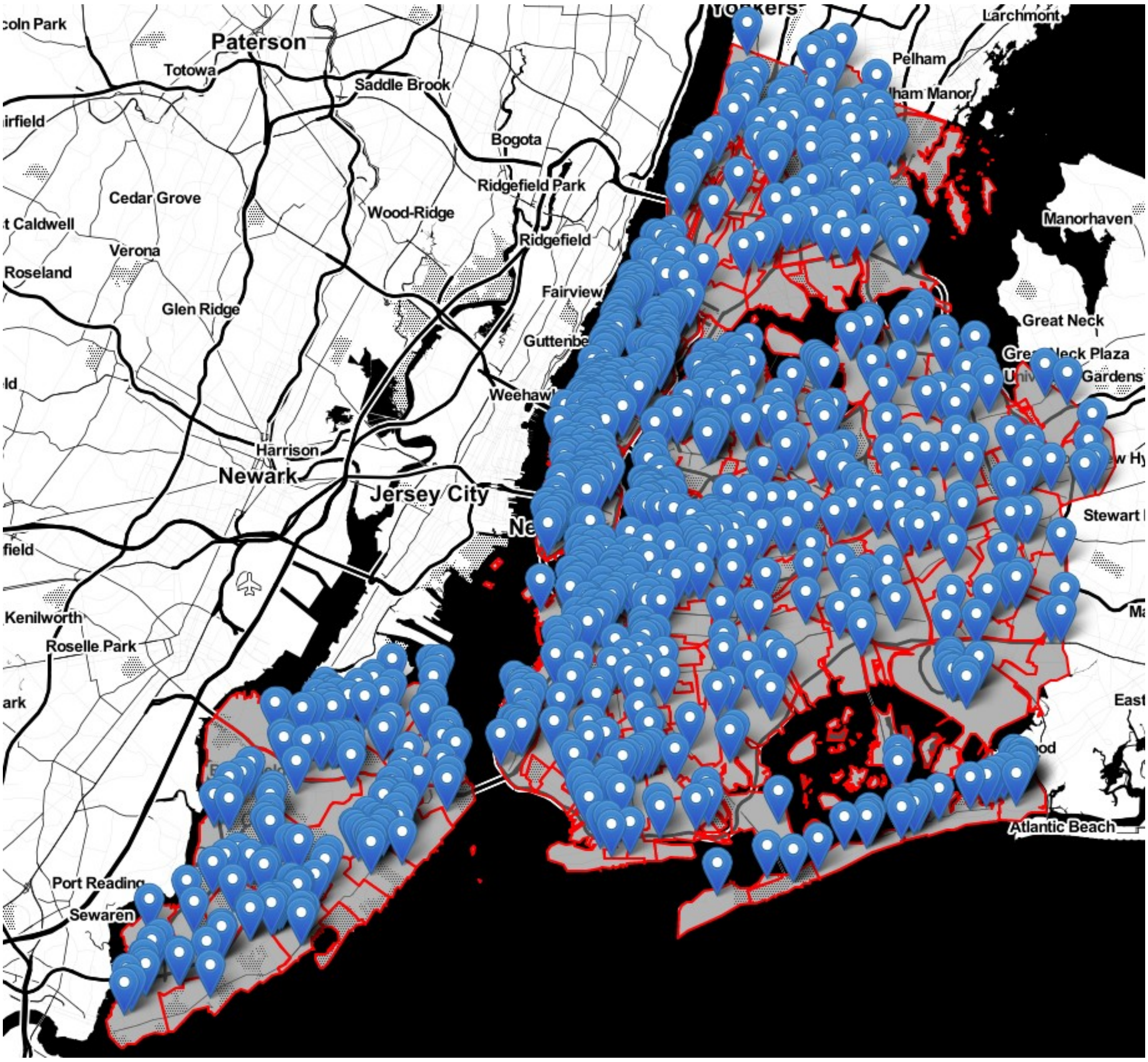
This stage created two outputs:

> raw_venues_data.csv

> raw_nta_data.csv

2.1.2 Data Prep – Stage 1

The main objective of this stage was to treat the venues data, that means dropping duplicate entries and columns that won’t be necessary for the final analysis. Finally I created a map with each pizzeria inside it’s corresponding NTA and dropped the pizzerias with no NTA associated. Result (each marker is a pizzeria):



This stage created two outputs:

> map_nta_pizzerias.html

> treated_venues_data.csv

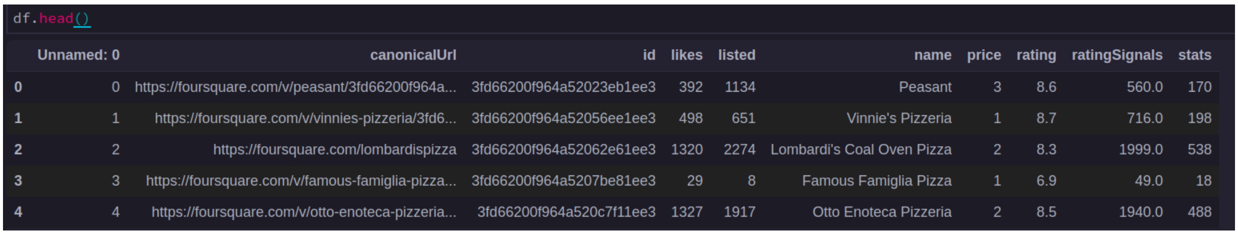
2.1.3 Getting Data – Stage 2

This stage objective was to get the details of each pizzeria using the ‘details’ endpoint of the Fourquare API. This endpoint is considered a premium call, meaning that I could only make 500 calls a day, each call returning a single pizzeria details. By this stage I had to process 903 different pizzerias, after calling the endpoint 500 times I switched to a friend’s credentials and used it to get the rest of the details needed.

This stage created one output:
> raw_venues_details.json

2.1.4 Data Prep – Stage 2

This stage objective was to treat the venues details gathered from the previous stage, this means dropping all the columns that are not needed to my analysis and treating the rest that are actually needed. The final result:



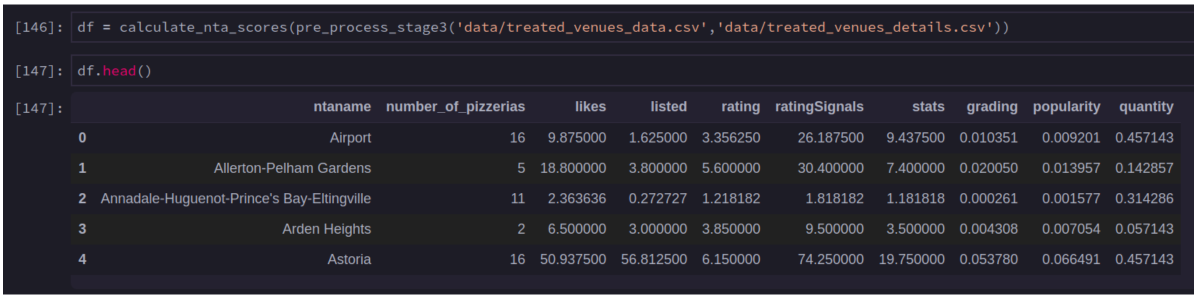
```
df.head()
```

	Unnamed: 0	canonicalUrl	id	likes	listed	name	price	rating	ratingSignals	stats
0	0	https://foursquare.com/v/peasant/3fd66200f964a...	3fd66200f964a52023eb1ee3	392	1134	Peasant	3	8.6	560.0	170
1	1	https://foursquare.com/v/vinnies-pizzeria/3fd6...	3fd66200f964a52056ee1ee3	498	651	Vinnie's Pizzeria	1	8.7	716.0	198
2	2	https://foursquare.com/lombardispizza	3fd66200f964a52062e61ee3	1320	2274	Lombardi's Coal Oven Pizza	2	8.3	1999.0	538
3	3	https://foursquare.com/v/famous-famiglia-pizza...	3fd66200f964a5207be81ee3	29	8	Famous Famiglia Pizza	1	6.9	49.0	18
4	4	https://foursquare.com/v/otto-enoteca-pizzeria...	3fd66200f964a520c7f11ee3	1327	1917	Otto Enoteca Pizzeria	2	8.5	1940.0	488

This stage created one output:
> treated_venues_details.csv

2.1.5 Data Prep – Stage 3

This stage objective was to merge the treated venues data that came from the ‘search’ endpoint to it’s correspondent treated details that came from the previous stage. Finally this stage create the three features that will be used to cluster the different NTAs. Final result:



```
[146]: df = calculate_nta_scores(pre_process_stage3('data/treated_venues_data.csv','data/treated_venues_details.csv'))

[147]: df.head()

[147]:
```

	ntaname	number_of_pizzerias	likes	listed	rating	ratingSignals	stats	grading	popularity	quantity
0	Airport	16	9.875000	1.625000	3.356250	26.187500	9.437500	0.010351	0.009201	0.457143
1	Allerton-Pelham Gardens	5	18.800000	3.800000	5.600000	30.400000	7.400000	0.020050	0.013957	0.142857
2	Annadale-Huguenot-Prince's Bay-Eltingville	11	2.363636	0.272727	1.218182	1.818182	1.181818	0.000261	0.001577	0.314286
3	Arden Heights	2	6.500000	3.000000	3.850000	9.500000	3.500000	0.004308	0.007054	0.057143
4	Astoria	16	50.937500	56.812500	6.150000	74.250000	19.750000	0.053780	0.066491	0.457143

This stage created did not created any document as a output. The output of this stage is used directly inside the jupyter for the modeling process.

3. Feature Engineering

For this project I used six features calculated for every NTA, the following:

1. number_of_pizzerias : absolute number of pizzerias inside that NTA;
2. likes : mean of likes of every pizzeria inside that NTA;
3. listed: mean of how many times each pizzeria of that NTA was listed by someone;

4. rating: mean of the ratings of every pizzeria of that NTA;
5. ratingSignals: mean of how many times every pizzeria of that NTA was rated;
6. stats: mean of how many written reviews every pizzeria of that NTA has.

These features were calculated in the ‘Data Prep – Stage 3’. I’ve used them all to come up with 3 main indicators for every NTA they’re : ‘grading’, ‘popularity’ and ‘quantity’.

3.1 Grading

I’ve defined ‘grading’ as:

$$grading=\frac{rating\times ratingSignals}{max(rating\times ratingSignals)}$$

Giving me an indicator that ranges from 0 to 1 which penalizes NTAs where the pizzerias have a low number of ‘ratingSignals’ and rewarding NTAs where the pizzerias have a high number of ‘ratingSignals’.

3.2 Popularity

I’ve defined ‘popularity’ as:

$$popularity=\frac{\left(\prod likes, listed, stats\right)^{\frac{1}{3}}}{max\left(\prod likes, listed, stats\right)^{\frac{1}{3}}}$$

Giving me an indicator that ranges from 0 to 1 which penalizes NTAs where the pizzerias got 0 likes, never been listed or never been reviewed by anyone.

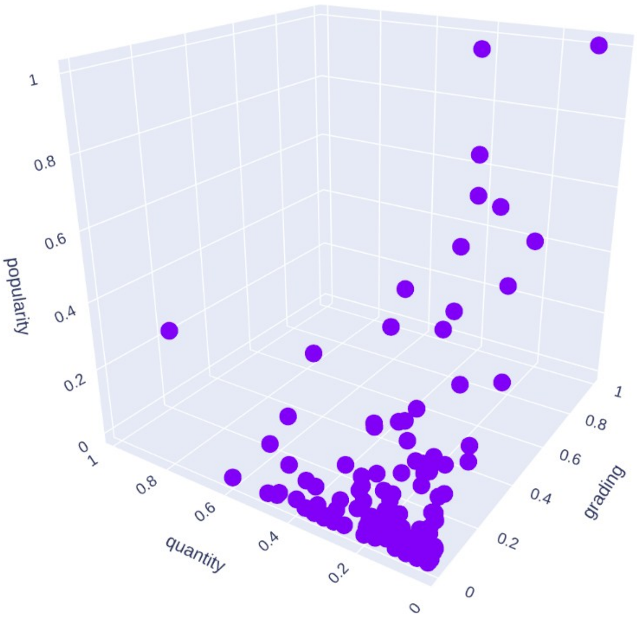
3.3 Quantity

I’ve define ‘quantity’ as:

$$quantity=\frac{number\ of\ pizzerias}{max(numer\ of\ pizzerias)}$$

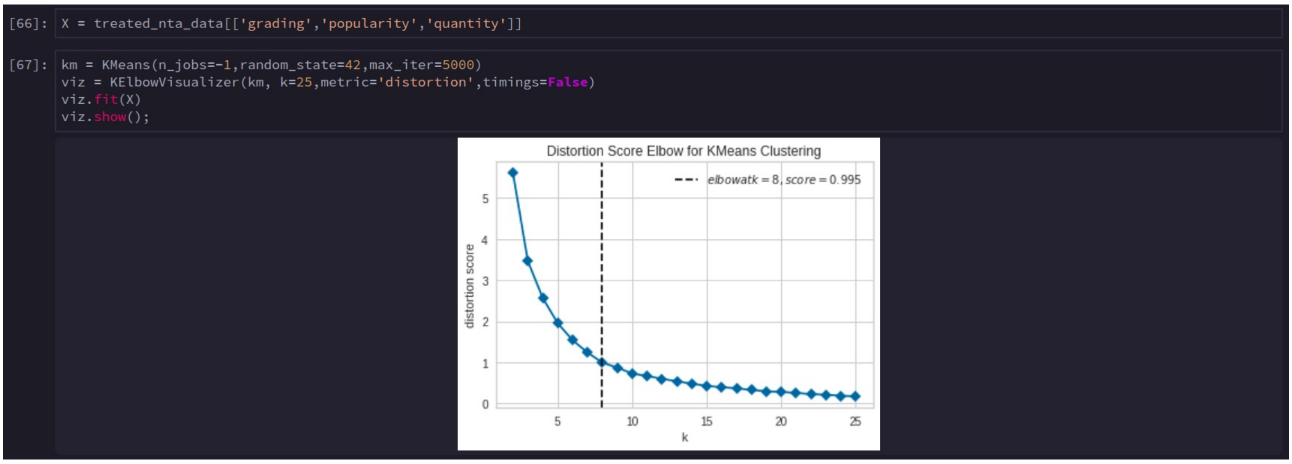
Giving me an indicator that ranges from 0 to 1 which penalizes NTAs with no pizzerias or a number too small of pizzerias.

Plotting all three features in a 3D scatter plot gives us the following (each NTA is a dot) :



4. Clustering

I used a clustering algorithm called KMeans from scikit-learn to find the optimal clusters and the Yellowbrick library to find the optimal number of clusters.

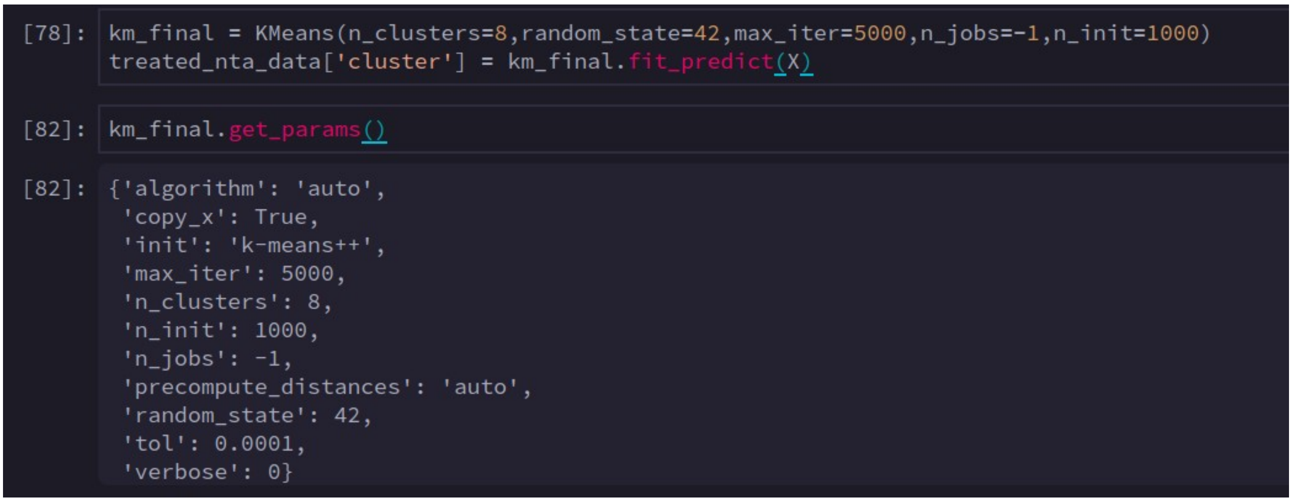


```
[66]: X = treated_nta_data[['grading','popularity','quantity']]

[67]: km = KMeans(n_jobs=-1,random_state=42,max_iter=5000)
      viz = KElbowVisualizer(km, k=25,metric='distortion',timings=False)
      viz.fit(X)
      viz.show();
```

Distortion Score Elbow for KMeans Clustering
elbowatk = 8, score = 0.995
distortion score
0
1
2
3
4
5
5
10
15
20
25
k

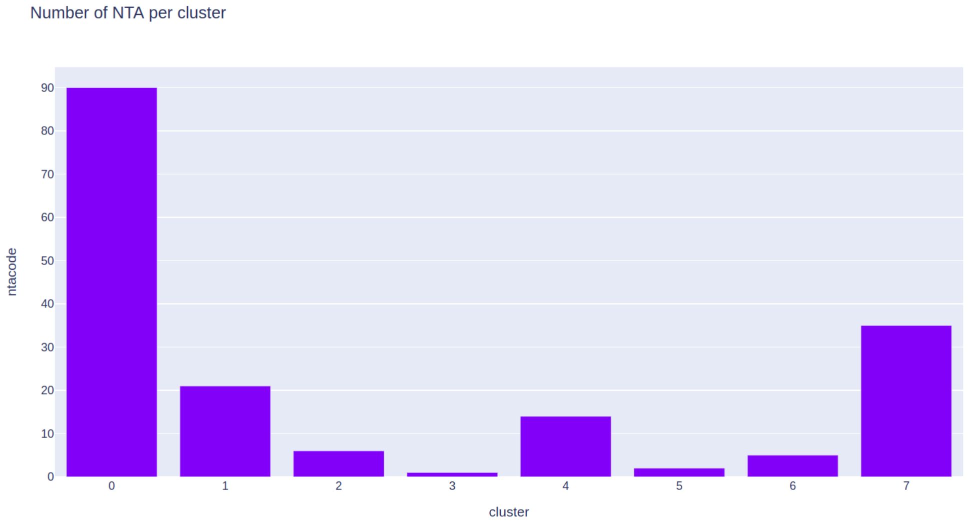
The Yellowbrick library uses the elbow method to calculate the optimal number of cluster, the image above shows a graph defining that the optimal k is 8.

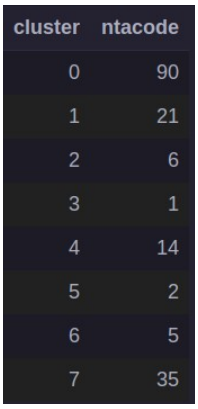
The final clustering algorithm:



```
[78]: km_final = KMeans(n_clusters=8,random_state=42,max_iter=5000,n_jobs=-1,n_init=1000)
      treated_nta_data['cluster'] = km_final.fit_predict(X)

[82]: km_final.get_params()

[82]: {'algorithm': 'auto',
       'copy_x': True,
       'init': 'k-means++',
       'max_iter': 5000,
       'n_clusters': 8,
       'n_init': 1000,
       'n_jobs': -1,
       'precompute_distances': 'auto',
       'random_state': 42,
       'tol': 0.0001,
       'verbose': 0}
```

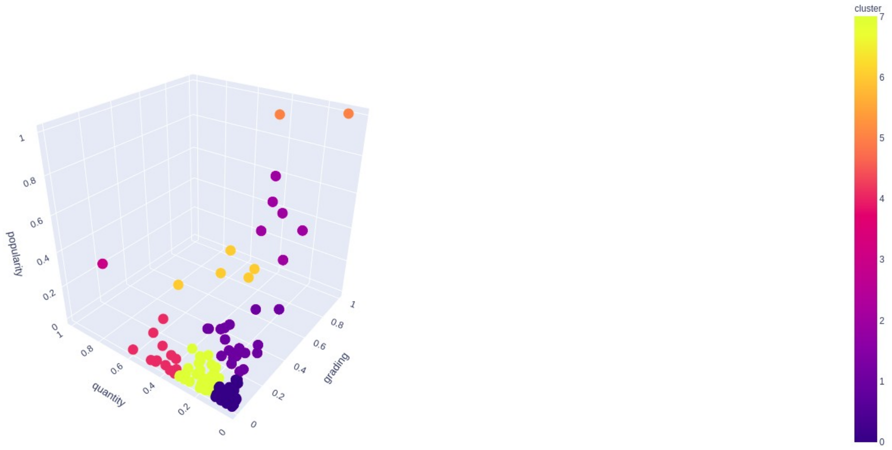
The distribution of NTA per cluster:





cluster	ntacode
0	90
1	21
2	6
3	1
4	14
5	2
6	5
7	35

The distribution of the three features (grading, popularity, quantity) plotted with cluster as a color dimension.

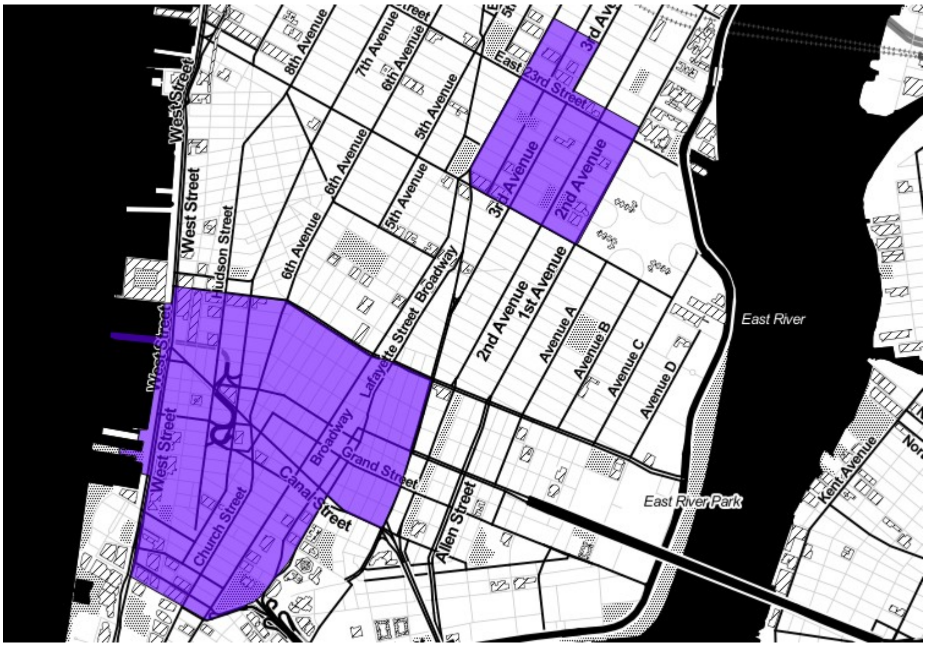


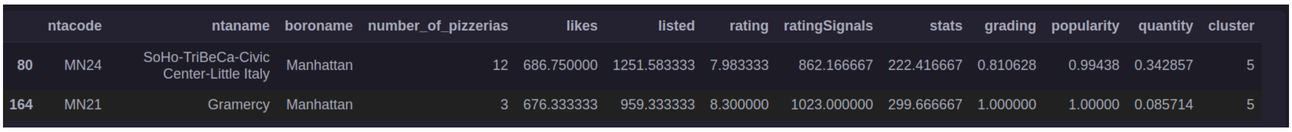
5. Results

Investigating the cluster centers:

Cluster	Quantity	Grading	Popularity
Cluster 0	Low	Low	Low
Cluster 1	Low	Low	Medium
Cluster 2	Low	Medium	Medium
Cluster 3	High	Medium	Medium
Cluster 4	Medium	Low	Low
Cluster 5	Low	High	High
Cluster 6	Medium	Medium	Medium
Cluster 7	Medium	Low	Low

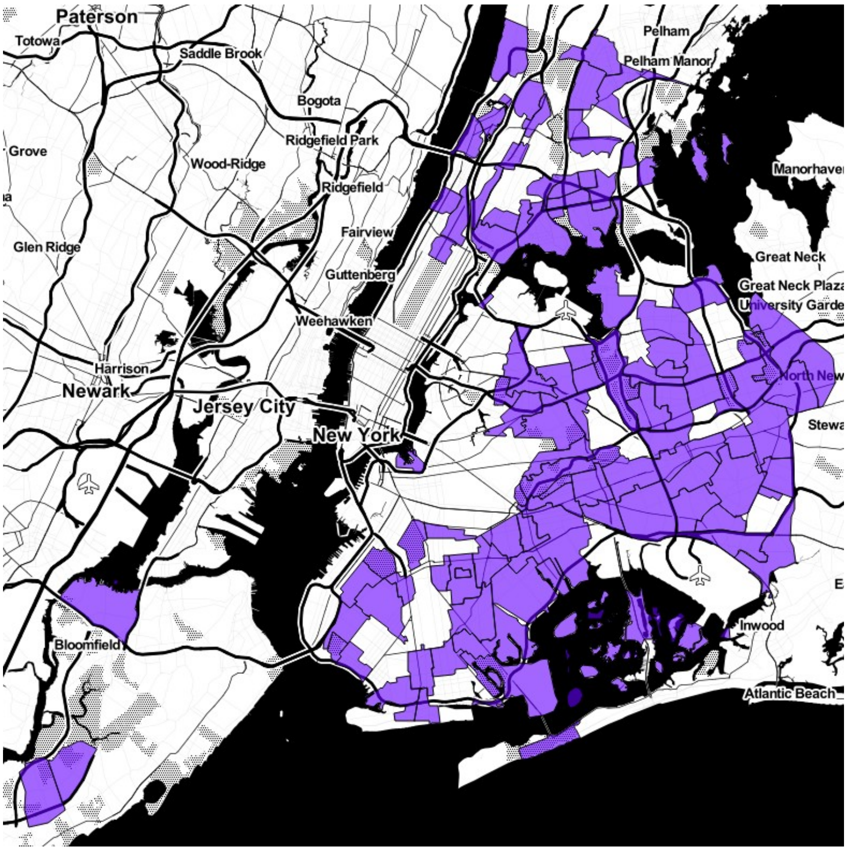
The best NTAs to try the pizzerias in terms of Grading and Popularity (cluster 5):





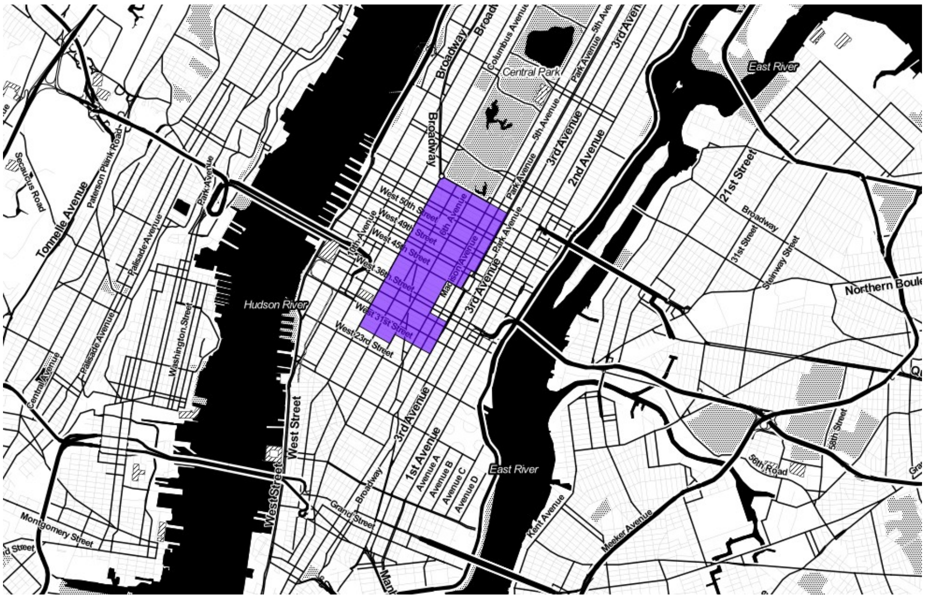
	ntacode	ntaname	boroname	number_of_pizzerias	likes	listed	rating	ratingSignals	stats	grading	popularity	quantity	cluster
80	MN24	SoHo-TriBeCa-Civic Center-Little Italy	Manhattan	12	686.750000	1251.583333	7.983333	862.166667	222.416667	0.810628	0.99438	0.342857	5
164	MN21	Gramercy	Manhattan	3	676.333333	959.333333	8.300000	1023.000000	299.666667	1.000000	1.00000	0.085714	5

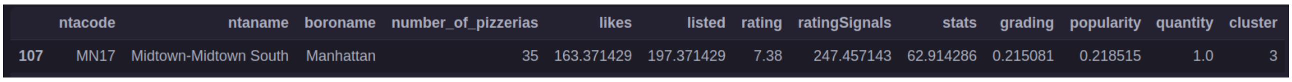
The ‘unknown’ pizzerias, almost no grading, low popularity and few pizzerias (cluster 5):



That does not mean that the pizzerias in these NTA are not good. It means that using data from Foursquare we can’t certainly give a recommendation because people that went to these pizzerias probably didn’t bother to check-in or leave a review.

The NTA with the most amount of pizzeria (cluster 3 , quantity = 1):





	ntacode	ntaname	boroname	number_of_pizzerias	likes	listed	rating	ratingSignals	stats	grading	popularity	quantity	cluster
107	MN17	Midtown-Midtown South	Manhattan	35	163.371429	197.371429	7.38	247.457143	62.914286	0.215081	0.218515	1.0	3

7. References

1. https://www.walksofnewyork.com/blog/italians-new-york-city
2. https://www.thespruceeats.com/what-is-new-york-style-pizza-2708764
3. https://www.nature.com/articles/nature10190