A tailored Neighborhood in NYC

Alex Vecchiettini

05 May 2020

Table of contents

Introduction

Background

New York is a dynamic and ever-changing city, moving into the city can be daunting and disorienting as prices are high, offer is varied and commuting not always easy or fast. People with different needs and interest all need to go through the complex process of finding the place that best suits their needs.

Problem

To better serve its clients, mostly working in Manhattan, a real estate agent needs to get a better insight on how New York city neighborhoods stack up against each other in order to focus the research of apartments in those with most value.

Analysis

Neighborhoods are complex systems and many different variables are at play. For our analysis we will focus on median rent price, distance from Manhattan and the kind of venues present around the area.

We will start by segmenting neighborhoods on the base of their venues profile.

We want to create many different profiles so that they could more closely relate to the many different interests people might have.

The distance from Manhattan will be calculated as 0 for the neighborhoods in the Borough of Manhattan and as the distance from the closest neighborhood in Manhattan Borough for all the others.

We will also create a weighting system to balance the variables of distance from Manhattan (giving a better score to those in Manhattan or closer to it) and median cost of rent to better suit customers needs as different people might value differently price and distance. After these steps we should be able to create a scoring system so that we can rank each neighborhood in their respective category.

Data acquisition and cleaning

Data sources

- Data for the geography and median rent cost (we will use the data for 2BR apartments as a reference) by neighborhood will be gathered from the website renthop.com (https://www.renthop.com/study/assets/new-york-city-cost-of-living-2017/nyc_col_geojson.js) which is a GeoJson that also contains our rent info.

- The venues data will be retrieved from FourSquare.com using the API provided

Data cleaning and feature selection

In the GeoJson many information are available, for this analysis makes sense to persist Neighborhoods:

Name through **‘features.properties.Neighborhood’**,

Borough through **‘features.properties.Borough’**,

Median rent through **‘features.properties.median2’**,

All the information to build the neighborhoods map that are available at **‘features.geometry’** which will be mostly used by Folium but we will need them to define the Neighborhood center as we will see in more details later,

All other information will be discarded.

Borough	Neighborhood	Median rent	geometry.coordinates	geometry.type
Bronx	Bedford Park-Fordham North	1895.0	[[[-73.883625, 40.867258], [-73.886833, 40.865...	Polygon
	East Concourse-Concourse Village	2225.0	[[[-73.909587, 40.842756], [-73.909625, 40.842...	Polygon
	Highbridge	2008.0	[[[-73.917287, 40.845104], [-73.917507, 40.844...	Polygon
	Hunts Point	1937.5	[[[-73.88439, 40.822967], [-73.883788, 40.8219...	Polygon
	Melrose South-Mott Haven North	1875.0	[[[-73.901293, 40.820475], [-73.90301, 40.8163...	Polygon

Figure 1 a snapshot of how the data extracted from the GeoJson looks like.

FourSquare API will return a verbose Json from which only the top 7 most common Venues categories (and their count) will be kept. Distance will be briefly used to filter out the venues categories over the 7th in case two or more categories are equally present (see Figure 2) in this situation the categories closer to the neighborhood center on average will be consider over the more distant ones.

categories	name	distance
Bar	3	653
Pizza Place	3	678
Mexican Restaurant	3	1085
Park	3	1422
Café	2	834
Scenic Lookout	2	888
Discount Store	1	196
Latin American Restaurant	1	207
Indian Restaurant	1	288
Grocery Store	1	299
Tapas Restaurant	1	312
Wine Shop	1	341
American Restaurant	1	353
Chinese Restaurant	1	353
Coffee Shop	1	381
Spa	1	426
Cocktail Bar	1	443
Garden	1	464
Market	1	476
Seafood Restaurant	1	565

Figure 2 an aggregate of all venues categories surrounding a point, ordered by descending frequency and increasing distance, in this example the first 6 most popular venue categories are clear but in regards to the 7th multiple categories have the same frequency and the one with the smallest distance will be selected.

Calculation of Neighborhood center

FourSquare API are able to provide venues from a pair of GPS coordinates, that means a point that is not present in the geometry of the Neighborhood so the need of calculating one arise in order to get the Venues in the area.

Three different methods come to mind:

1- Calculating the average of all points that define the border of a Neighborhood;
2- Calculating the center of the rectangle inscribing the polygonal Neighborhood;
3- Calculating the centroid of a quadrilateral which vertices coincides with the North-East-South-West most points in the Neighborhood.

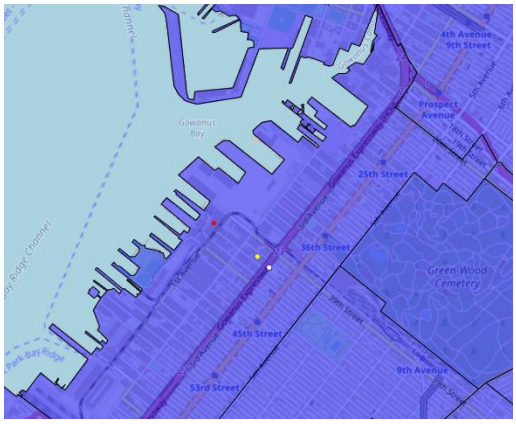


Figure 3 the different results of the 3 different method to evaluate the neighborhood center. Red first method, yellow second method and white third method.

While the first one might seem the most accurate it actually get biased towards zone of the neighborhood that have a complex shape and many points to define them (see Figure 3).

The second way is a good approximation but the third one on average returns the best results and has been chosen as preferred method to describe the neighborhoods center.

Calculation of Neighborhoods distances

Being Manhattan a Borough composed by many different neighborhoods calculating the distance from it requires making an assumption:

- calculate the distance from a central point in Manhattan;

- getting the minimum distance towards all the neighborhoods in Manhattan.

These are the 2 possibilities analyzed but not the only possible.

Considering no information is given on which area of Manhattan to prioritize and different clients might need to reach different areas, along with the consideration that the borough is well served by transportation the option of calculating the distance from the closest Neighborhood in Manhattan, hence the minimum distance to reach it.

Dimension reduction

A vast majority of the venues are food related and especially restaurants. To lessen the effect all the different restaurants and food venues will have on the clustering an aggregation between similar categories is a possible solution, the aggregation was done as per the table below. The frequency has been maintained as sum of all the cooperating sub categories.

Categories to be aggregated	Final category	Dimension reduction
'Spanish Restaurant', 'Greek Restaurant', 'Eastern European Restaurant', 'Russian Restaurant', 'Italian Restaurant', 'Mediterranean Restaurant'	'European Restaurant'	5
'Japanese Restaurant', 'Chinese Restaurant', 'Shanghai Restaurant', 'Cantonese Restaurant', 'Thai Restaurant', 'Malay Restaurant', 'Sushi Restaurant', 'Vietnamese Restaurant', 'Szechuan Restaurant', 'Korean Restaurant', 'Taiwanese Restaurant', 'Indonesian Restaurant', 'Hotpot Restaurant', 'Dumpling Restaurant'	'Asian Restaurant'	13
'Empanada Restaurant', 'Latin American Restaurant', 'Caribbean Restaurant', 'Cuban Restaurant', 'Mexican Restaurant', 'Peruvian Restaurant'	'South American Restaurant'	6
'Turkish Restaurant', 'Falafel Restaurant'	'Middle Eastern Restaurant'	2
'American Restaurant', 'New American Restaurant', 'Southern / Soul Food Restaurant', 'BBQ Joint', 'Diner', 'Steakhouse', 'Breakfast Spot'	'North American Restaurant'	7
'Juice Bar', 'Bar', 'Cocktail Bar', 'Wine Bar', 'Sake Bar', 'Karaoke Bar', 'Sports Bar'	'Bar'	6

Clustering algorithm

To lessen the effects of central area against rural data of each neighborhood category has been normalized so that each category does not contain the amount of venues in a neighborhood but the frequency inside that neighborhood of a particular category.

K-Means

K-Means algorithm ensure a division of our data in k cluster, it is a randomized Heuristic algorithm so it might find different solutions every time is run, consequently the resulting clusters can differ by Neighborhood composition and consequently the type of venues each cluster represent. Is unfortunately very difficult to visualize data in a high dimension sparse matrix such as the one composed of more than 100 neighborhoods and 150 venue categories so to find the best k the elbow method has been attempted. In order to have a more precise idea a silhouette analysis is performed (see Figure 5), and the results seems to confirm 7 as the best k for our data as below it clusters are to wide and above it they tend to have many elements not really pertaining to the cluster they are assigned.

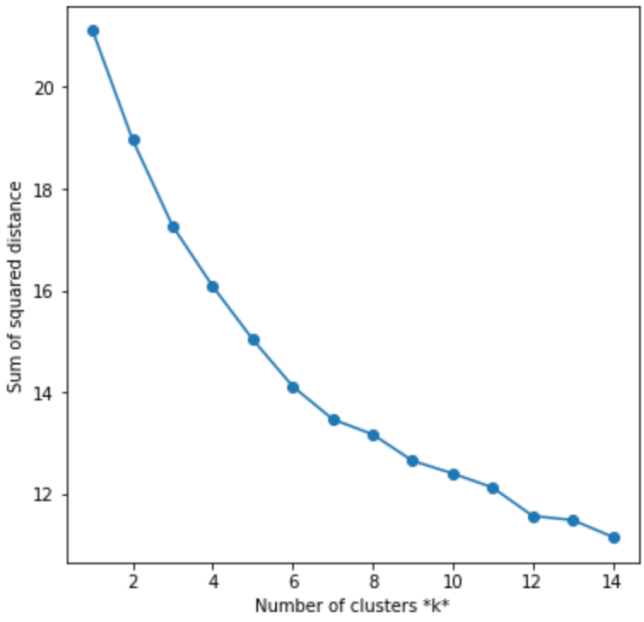


Figure 4 Elbow method graph. No clear elbow point but 7 cluster seem a promising compromise.

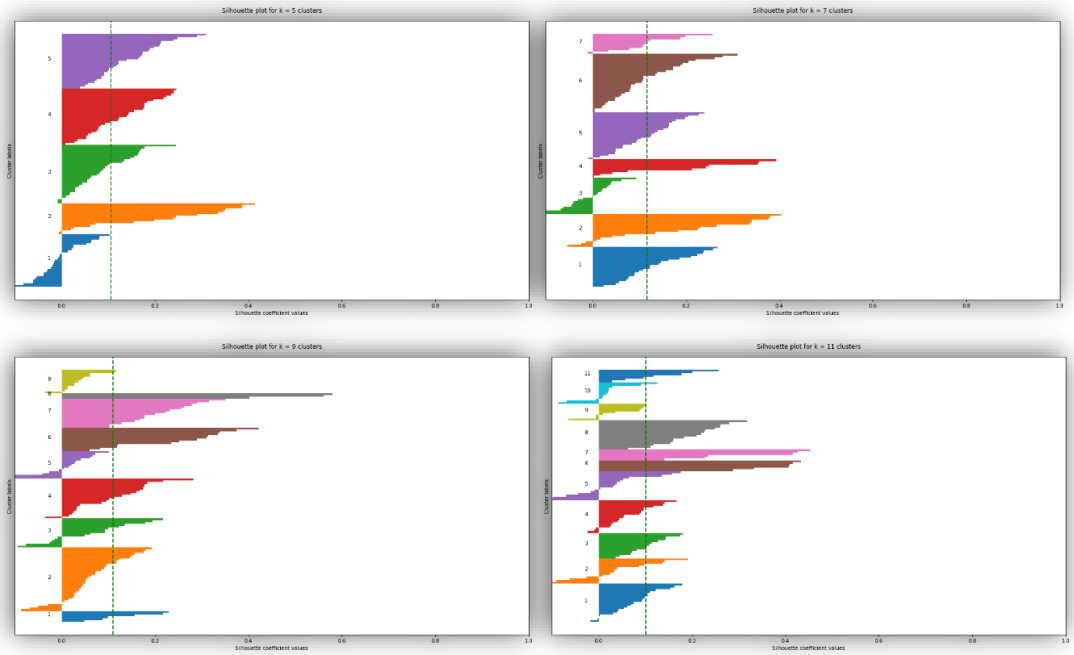


Figure 5 Silhouette analysis

Hierarchical

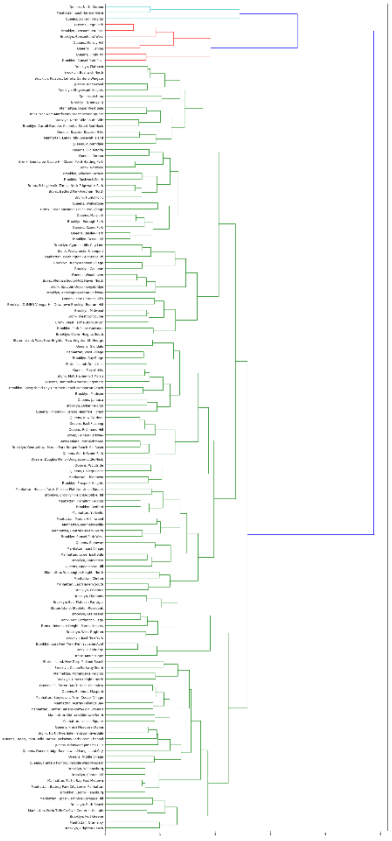


Figure 6 Hierarchical clustering dendogram

An attempt to Hierarchical clustering has been made as well, this would have two substantial advantages over the K-Means, it's not randomized so clusters would be defined once and for all and we would be able to tweak the distance between clusters and have the number of cluster accordingly distributed. The resulting dendrogram (see Figure 6), however, already gave insights on the results and why it's not a viable option.

The tree is not balanced so some cluster would end up with very few Neighborhoods if not a single one whilst few others would cover the vast majority of neighborhoods. The implications of this on our end goal would mean that few cluster would be very specific on few categories and would represent a small amount of neighborhoods while others would be very wide and generic.

Results comparison

Seeing the neighborhoods categories distributions (see Figure 7) is clear that the more balanced solution is the best one for the final purpose of clusters like these as they should be balanced as much as possible to give equal opportunities to different preference.

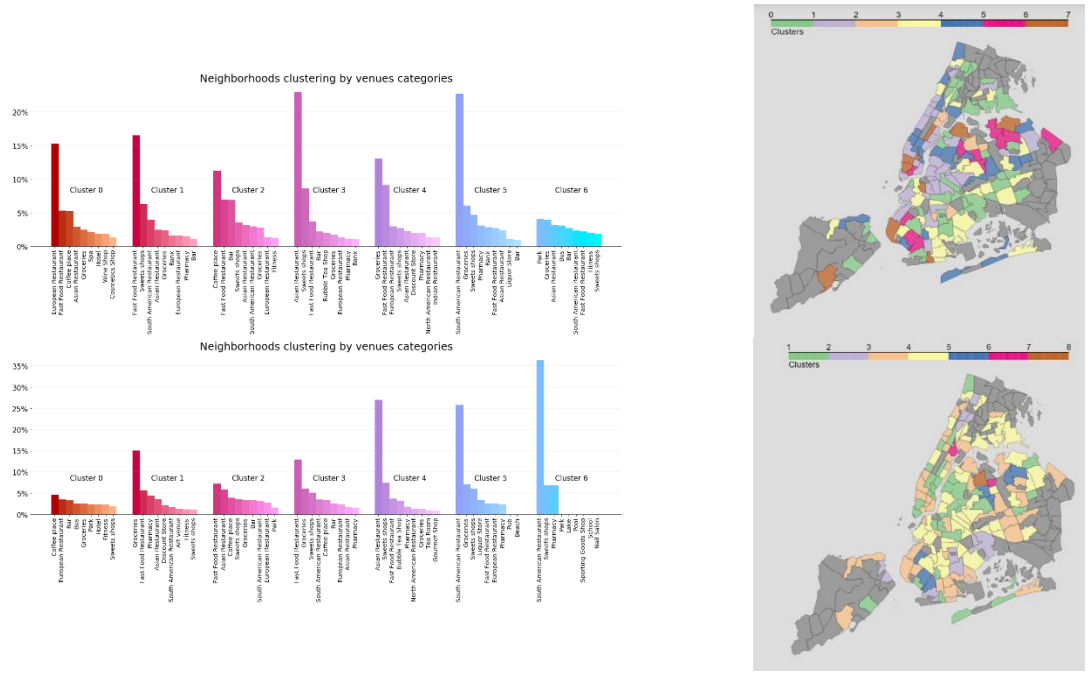


Figure 7 K-Means vs Hierarchical - Clusters composition and distribution

These results highlight how K-Means is a better algorithm for the goal and will be used for clustering the Dataset.

Exploratory Data Analysis

Relationship between Distance and Median Rent

To better understand how to create the scoring system and how the variables influence each other it can be helpful to find the relationship between median rent and distance.

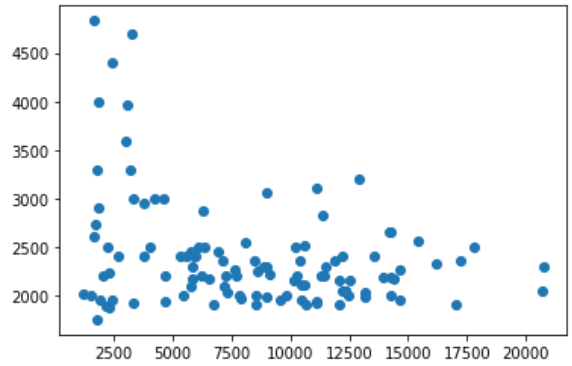


Figure 8 A scatter plot of Distance vs Median rent price (Neighborhoods from Manhattan have been removed since their distance is 0)

As visible in figure 8 a reciprocal function seems to be in place between the two variables, it's also unlikely we will be able to fit a reliable curve through our point as many other factors influence the rent price other then distance, such as reputation of the neighborhood, crime rate, services and so on.

Also the distance itself has been defined in a particular way to suit the problem and neighborhood close to a part of Manhattan that is not in the real heart of the city are bound to have a lower price even if their distance in our metric might be small, this is particularly evident for neighborhood in the borough of Bronx which constitute the datapoint on the bottom left of the screen.

Relationship between Cluster and Distance

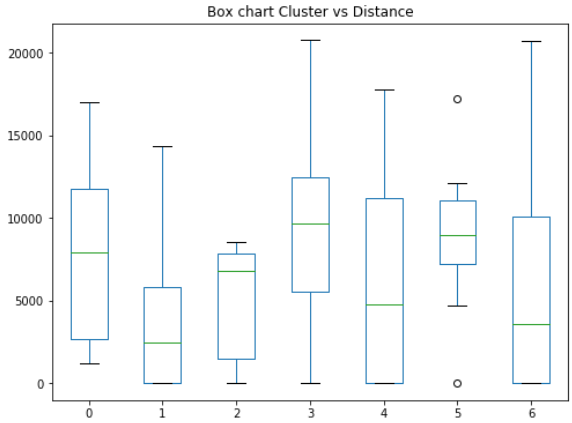


Figure 9 Box chart of Cluster vs Distance

Clusters are randomly defined and is observable (see Figure 9) that most of them are present in a wide range of distances, that means a good job was made avoiding the clusters to be focused on the density of venues in an area and direct the "attention" of the K-Means algorithm towards the quality and frequency of venues.

Clearly, the scoring will be higher for all the neighborhoods in the bottom part of this graph as it will be for the variable of Median rent.

Relationship between Cluster and Median Rent

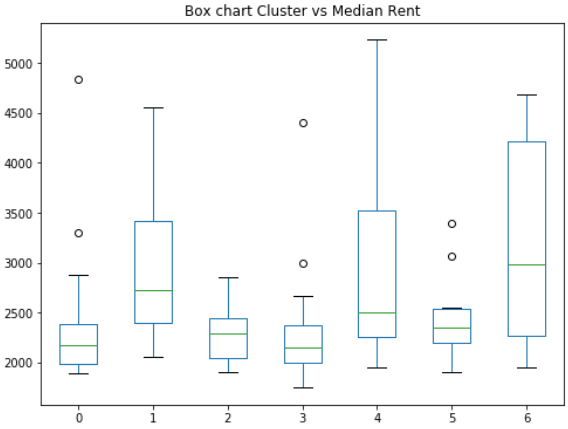


Figure 10 Box chart of Clusters vs Median Rent price

As seen already for the variable of Distance price is quite well distributed even if in this case is more evident how some neighborhoods have a higher concentration of low rent values and some outliers in the high-end price.

Customers categories

Results apparent the need of adding a customer categorization as the different variables (in this case only distance and median rent) may be have more or less importance to different individuals, and this would affect the interest they might have for a neighborhood, hence the scoring system should also take this into account. As a simplified example 3 dummy categories of customers are taken into account:

1. Young and wealthy professionals with high salaries working in Manhattan that are more interested in having a short commute then the rent bills, for these people we will assume that distance is 3 times more important then rent price;

2. Manhattan labor workers that have to equally balance rent and distance;
3. Families that are not as interested in living in the city center and they rather save money on the rent; these people will be biased towards rent by a factor of 3.

Conclusions

With all the consideration taken into account so far define a scoring system based on each cluster individually and a division of the cluster in percentile for both the Distance and Median rent variables. The score for each Neighborhood is then calculated as the difference between 100 and the percentile to which that Neighborhood belongs to.

Borough	Neighborhood	Median rent	Distance	Distance score	Rent score	Cluster
Bronx	Highbridge	2008.0	1195.246079	100.0	66.0	0
	University Heights-Morris Heights	1995.0	1519.418663	96.0	72.0	0
Brooklyn	Brooklyn Heights-Cobble Hill	4834.0	1691.836317	93.0	1.0	0
Queens	Old Astoria	2725.0	1759.349931	90.0	12.0	0
	Queensbridge-Ravenswood-Long Island City	3300.0	1826.999881	87.0	3.0	0
Bronx	Van Cortlandt Village	1950.0	1942.315424	84.0	81.0	0
	Bedford Park-Fordham North	1895.0	2208.230802	81.0	100.0	0
	East Concourse-Concourse Village	2225.0	2279.359329	78.0	39.0	0
	Mount Hope	1950.0	2432.869944	75.0	81.0	0
	Norwood	1925.0	3332.561539	72.0	90.0	0
	Hunts Point	1937.5	4690.441039	69.0	84.0	0
Brooklyn	Bushwick South	2400.0	5563.189030	66.0	24.0	0
Bronx	Pelham Parkway	2169.0	5876.415985	63.0	51.0	0
Brooklyn	Crown Heights North	2400.0	5941.622167	60.0	24.0	0
	Windsor Terrace	2875.0	6313.912394	57.0	6.0	0
	Ocean Hill	2099.0	7169.130660	54.0	54.0	0
Bronx	Soundview-Castle Hill-Clason Point-Harding Park	2025.0	7305.927004	51.0	63.0	0
Queens	Corona	2355.0	8452.406055	48.0	27.0	0
Brooklyn	East New York (Pennsylvania Ave)	2300.0	8888.956950	45.0	36.0	0
	Borough Park	1975.0	9011.315616	42.0	75.0	0
	East Flatbush-Farragut	2212.0	9096.323289	39.0	42.0	0

Borough	Neighborhood	Median rent	Distance	Distance score	Rent score	Cluster
Manhattan	Stuyvesant Town-Cooper Village	4256.5	0.000000	100.0	5.0	4
	Washington Heights North	2275.0	0.000000	100.0	73.0	4
	Murray Hill-Kips Bay	3498.0	0.000000	100.0	26.0	4
	Midtown-Midtown South	3850.0	0.000000	100.0	15.0	4
	Central Harlem North-Polo Grounds	2300.0	0.000000	100.0	68.0	4
	Clinton	3600.0	0.000000	100.0	21.0	4
	Lincoln Square	5232.5	0.000000	100.0	1.0	4
Bronx	Mott Haven-Port Morris	2200.0	2029.299712	63.0	84.0	4
Queens	Hunters Point-Sunnyside-West Maspeth	3969.0	3060.077683	57.0	10.0	4
Bronx	North Riverdale-Fieldston-Riverdale	3295.0	3197.289448	52.0	31.0	4
Queens	Elmhurst-Maspeth	2495.0	6360.585761	47.0	57.0	4
	Glendale	2000.0	8542.203516	42.0	89.0	4
	Forest Hills	2500.0	10230.874333	36.0	52.0	4
Brooklyn	Ocean Parkway South	2350.0	10399.029817	31.0	63.0	4
Staten Island	West New Brighton-New Brighton-St. George	3100.0	11121.409376	26.0	42.0	4
Brooklyn	Bath Beach	2200.0	11456.252873	21.0	84.0	4
Queens	Ft. Totten-Bay Terrace-Clearview	3200.0	12915.788568	15.0	36.0	4
	Pomonok-Flushing Heights-Hillcrest	1975.0	13161.969336	10.0	94.0	4
	Fresh Meadows-Utopia	1950.0	14643.139739	5.0	100.0	4
	Breezy Point-Belle Harbor-Rockaway Park-Broad Channel	2500.0	17779.454996	1.0	52.0	4

Figure 11 Cluster 0 and 4 sorted by "Distance"

All the neighborhood in Manhattan will have a score of 100 for the distance, across all the clusters and in each cluster the cheapest Neighborhood will get a 100 score as well.

If a cluster do not have neighborhoods in Manhattan the closest one will have a perfect score and this will be diminished for every percentile after that.

Borough	Neighborhood	Median rent	Distance	Distance score	Rent score	Cluster
Bronx	West Concourse	1750.0	1786.057916	92.0	100.0	3
	Melrose South-Mott Haven North	1875.0	2307.358459	88.0	96.0	3
	Westchester-Unionport	1900.0	6721.589002	64.0	92.0	3
Brooklyn	Canarsie	1937.5	11124.712784	36.0	88.0	3
	Dyker Heights	1950.0	9541.204290	52.0	84.0	3
Queens	South Ozone Park	2000.0	14276.442642	8.0	80.0	3
Bronx	West Farms-Bronx River	2000.0	5432.214347	76.0	80.0	3
Queens	Middle Village	2000.0	7819.353132	56.0	80.0	3
Brooklyn	Cypress Hills-City Line	2000.0	9840.687245	48.0	80.0	3
Queens	Kew Gardens Hills	2050.0	12333.694862	28.0	64.0	3
	Maspeth	2099.0	5754.640311	72.0	60.0	3
	Woodhaven	2100.0	10601.119074	40.0	56.0	3
Brooklyn	Midwood	2150.0	10119.183687	44.0	52.0	3
	Madison	2150.0	12109.042458	32.0	52.0	3
	Georgetown-Marine Park-Bergen Beach-Mill Basin	2150.0	12514.396610	24.0	52.0	3
	Sheepshead Bay-Gerritsen Beach-Manhattan Beach	2187.5	14265.767094	16.0	40.0	3
	Kensington-Ocean Parkway	2199.5	7725.310746	60.0	36.0	3
Manhattan	Washington Heights South	2295.0	0.000000	100.0	32.0	3
Queens	Hammels-Arverne-Edgemere	2300.0	20758.757298	1.0	28.0	3
	Auburndale	2400.0	13585.444069	20.0	24.0	3
Brooklyn	Crown Heights South	2497.0	6088.128127	68.0	20.0	3

Borough	Neighborhood	Median rent	Distance	Distance score	Rent score	Cluster
Queens	Fresh Meadows-Utopia	1950.0	14643.139739	5.0	100.0	4
	Pomonok-Flushing Heights-Hillcrest	1975.0	13161.969336	10.0	94.0	4
	Glendale	2000.0	8542.203516	42.0	89.0	4
Bronx	Mott Haven-Port Morris	2200.0	2029.299712	63.0	84.0	4
Brooklyn	Bath Beach	2200.0	11456.252873	21.0	84.0	4
Manhattan	Washington Heights North	2275.0	0.000000	100.0	73.0	4
	Central Harlem North-Polo Grounds	2300.0	0.000000	100.0	68.0	4
Brooklyn	Ocean Parkway South	2350.0	10399.029817	31.0	63.0	4
Queens	Elmhurst-Maspeth	2495.0	6360.585761	47.0	57.0	4
	Forest Hills	2500.0	10230.874333	36.0	52.0	4
	Breezy Point-Belle Harbor-Rockaway Park-Broad Channel	2500.0	17779.454996	1.0	52.0	4
Staten Island	West New Brighton-New Brighton-St. George	3100.0	11121.409376	26.0	42.0	4
Queens	Ft. Totten-Bay Terrace-Clearview	3200.0	12915.788568	15.0	36.0	4
Bronx	North Riverdale-Fieldston-Riverdale	3295.0	3197.289448	52.0	31.0	4
Manhattan	Murray Hill-Kips Bay	3498.0	0.000000	100.0	26.0	4
	Clinton	3600.0	0.000000	100.0	21.0	4
	Midtown-Midtown South	3850.0	0.000000	100.0	15.0	4
Queens	Hunters Point-Sunnyside-West Maspeth	3969.0	3060.077683	57.0	10.0	4
Manhattan	Stuyvesant Town-Cooper Village	4256.5	0.000000	100.0	5.0	4
	Lincoln Square	5232.5	0.000000	100.0	1.0	4

Figure 12 Cluster 3 and 4 sorted by "Median rent"

After each cluster has been scored by Distance and Median rent the final score is calculated for each customer category weighting the two scores accordingly to each category.

Borough	Neighborhood	Professionals score	Cluster	Median rent	Distance
Bronx	Highbridge	91.50	0	2008.0	1195.246079
	University Heights-Morris Heights	90.00	0	1995.0	1519.418663
	Bedford Park-Fordham North	85.75	0	1895.0	2208.230802
	Van Cortlandt Village	83.25	0	1950.0	1942.315424
	Mount Hope	76.50	0	1950.0	2432.869944
Manhattan	Marble Hill-Inwood	97.50	1	2200.0	0.000000
	Hamilton Heights	90.25	1	2495.0	0.000000
	Manhattanville	87.75	1	2650.0	0.000000
	Central Harlem South	85.25	1	3000.0	0.000000
	Morningside Heights	83.00	1	3200.0	0.000000
	East Harlem North	80.00	2	2495.0	0.000000
	East Harlem South	75.25	2	2850.0	0.000000
Queens	Jackson Heights	95.00	2	2300.0	5858.042084
	North Corona	45.00	2	2270.0	7660.850681
Brooklyn	Erasmus	35.00	2	1962.5	7886.885414
Bronx	West Concourse	94.00	3	1750.0	1786.057916
	Melrose South-Mott Haven North	90.00	3	1875.0	2307.358459
Manhattan	Washington Heights South	83.00	3	2295.0	0.000000
Bronx	West Farms-Bronx River	77.00	3	2000.0	5432.214347
	Spuyten Duyvil-Kingsbridge	75.00	3	2600.0	1679.299997
Manhattan	Washington Heights North	93.25	4	2275.0	0.000000

Borough	Neighborhood	Labour score	Cluster	Median rent	Distance
Bronx	Bedford Park-Fordham North	90.5	0	1895.0	2208.230802
	University Heights-Morris Heights	84.0	0	1995.0	1519.418663
	Highbridge	83.0	0	2008.0	1195.246079
	Van Cortlandt Village	82.5	0	1950.0	1942.315424
	Norwood	81.0	0	1925.0	3332.561539
Manhattan	Marble Hill-Inwood	95.0	1	2200.0	0.000000
	Hamilton Heights	80.5	1	2495.0	0.000000
	Manhattanville	75.5	1	2550.0	0.000000
	Central Harlem South	70.5	1	3000.0	0.000000
	Morningside Heights	68.0	1	3200.0	0.000000
	East Harlem North	60.0	2	2495.0	0.000000
Brooklyn	Rugby-Remsen Village	50.5	2	1900.0	8515.104734
Manhattan	East Harlem South	50.5	2	2850.0	0.000000
Brooklyn	Erasmus	50.0	2	1962.5	7886.885414
Queens	Jackson Heights	50.0	2	2300.0	5858.042084
Bronx	West Concourse	96.0	3	1750.0	1786.057916
	Melrose South-Mott Haven North	92.0	3	1875.0	2307.358459
	West Farms-Bronx River	78.0	3	2000.0	5432.214347
	Westchester-Unionport	78.0	3	1900.0	6721.589002
Brooklyn	Dyker Heights	68.0	3	1950.0	9541.204290
Manhattan	Washington Heights North	86.5	4	2275.0	0.000000

Borough	Neighborhood	Families score	Cluster	Median rent	Distance
Bronx	Bedford Park-Fordham North	95.25	0	1895.0	2208.230802
	Norwood	89.00	0	1925.0	3332.561539
	Van Cortlandt Village	81.75	0	1950.0	1942.315424
	Hunts Point	80.25	0	1957.5	4690.441039
Brooklyn	East New York	80.25	0	1900.0	10609.692674
Manhattan	Marble Hill-Inwood	92.50	1	2200.0	0.000000
Bronx	Van Nest-Morris Park-Westchester Square	76.00	1	2168.0	6523.541717
Brooklyn	Homecrest	79.75	1	2650.0	12315.800919
Queens	Ridgewood	72.25	1	2200.0	6216.763356
Manhattan	Hamilton Heights	70.75	1	2495.0	0.000000
Brooklyn	Rugby-Remsen Village	76.25	2	1900.0	8515.104734
	Erasmus	65.00	2	1962.5	7886.885414
Queens	North Corona	55.00	2	2270.0	7660.850681
	Jackson Heights	45.00	2	2300.0	5858.042084
Manhattan	East Harlem North	40.00	2	2495.0	0.000000
Bronx	West Concourse	98.00	3	1750.0	1786.057916
	Melrose South-Mott Haven North	94.00	3	1875.0	2307.358459
	Westchester-Unionport	85.00	3	1900.0	6721.589002
	West Farms-Bronx River	79.00	3	2000.0	5432.214347
Brooklyn	Dyker Heights	76.00	3	1950.0	9541.204290
Manhattan	Washington Heights North	79.75	4	2275.0	0.000000

Figure 13 Top 5 Neighborhoods for Professionals, Labor and Families respectively.

Scores will clearly vary according with the customer category and (see Figure 13) what can be the best neighborhood for a “Professional” might not be in the top 5 for a customer belonging to the “Families” category. After this all neighborhoods are trivially ranked in their cluster and for customer categories. All Neighborhoods with the same score in a cluster and in a customer, category are ranked at the same (minimum) position so there might be more than one Neighborhood ranked as 1st.

	Professionals	Labour	Families
0	Bronx, Bedford Park-Fordham North /-----/ Bronx, Highbridge /-----/ Bronx, University Heights-Morris Heights	Bronx, Bedford Park-Fordham North /-----/ Bronx, Highbridge /-----/ Bronx, University Heights-Morris Heights	Bronx, Bedford Park-Fordham North /-----/ Bronx, Norwood /-----/ Bronx, Van Cortlandt Village
1	Manhattan, Hamilton Heights /-----/ Manhattan, Manhattanville /-----/ Manhattan, Marble Hill-Inwood	Manhattan, Hamilton Heights /-----/ Manhattan, Manhattanville /-----/ Manhattan, Marble Hill-Inwood	Bronx, Van Nest-Morris Park-Westchester Square /-----/ Brooklyn, Homecrest /-----/ Manhattan, Marble Hill-Inwood
2	Manhattan, East Harlem North /-----/ Manhattan, East Harlem South /-----/ Queens, Jackson Heights	Brooklyn, Rugby-Remsen Village /-----/ Manhattan, East Harlem North /-----/ Manhattan, East Harlem South	Brooklyn, Erasmus /-----/ Brooklyn, Rugby-Remsen Village /-----/ Queens, North Corona
3	Bronx, Melrose South-Mott Haven North /-----/ Bronx, West Concourse /-----/ Manhattan, Washington Heights South	Bronx, Melrose South-Mott Haven North /-----/ Bronx, West Concourse /-----/ Bronx, West Farms-Bronx River /-----/ Bronx, Westchester-Unionport	Bronx, Melrose South-Mott Haven North /-----/ Bronx, West Concourse /-----/ Bronx, Westchester-Unionport
4	Manhattan, Central Harlem North-Polo Grounds /-----/ Manhattan, Murray Hill-Kips Bay /-----/ Manhattan, Washington Heights North	Bronx, Mott Haven-Port Morris /-----/ Manhattan, Central Harlem North-Polo Grounds /-----/ Manhattan, Washington Heights North	Bronx, Mott Haven-Port Morris /-----/ Manhattan, Washington Heights North /-----/ Queens, Glendale
5	Manhattan, Chinatown /-----/ Queens, Elmhurst /-----/ Queens, Woodside	Brooklyn, Bensonhurst West /-----/ Brooklyn, Sunset Park East /-----/ Queens, Elmhurst /-----/ Queens, Woodside	Brooklyn, Bensonhurst East /-----/ Brooklyn, Bensonhurst West /-----/ Queens, Woodside
6	Manhattan, Gramercy /-----/ Manhattan, Upper East Side-Carnegie Hill /-----/ Manhattan, West Village	Manhattan, Gramercy /-----/ Manhattan, Upper East Side-Carnegie Hill /-----/ Manhattan, West Village /-----/ Queens, Whitestone	Brooklyn, Brighton Beach /-----/ Queens, Whitestone /-----/ Staten Island, Great Kills

Figure 14 List of Neighborhoods ranked in the top 3 for each cluster and customer category (not sorted by rank)

Visualization

A sizable amount of information needs to be conveyed and this requires a proper way to visualize them.

Cluster distribution

A consideration could be made of visualizing all the data together (see Figure 15).

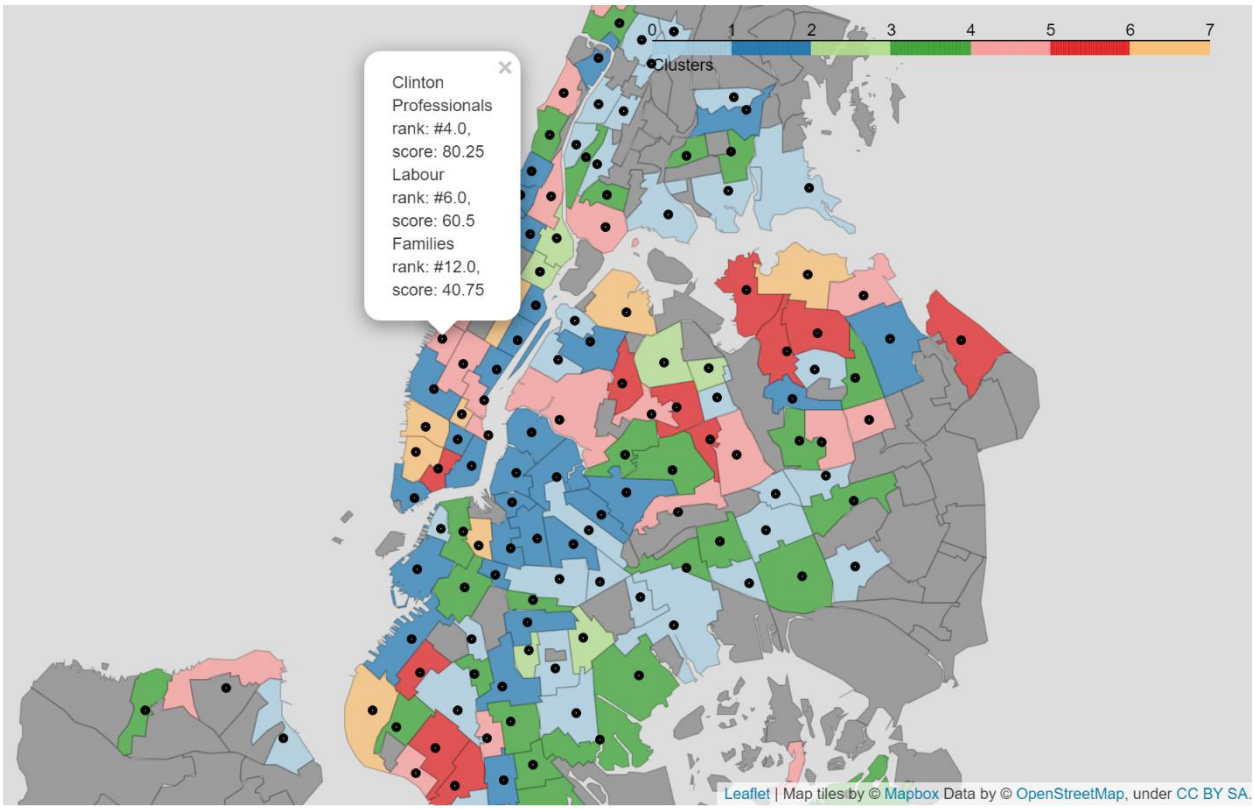


Figure 15 New York Neighborhoods color coded by cluster, each mark on the map contains all the information for each customer category.

This is a very informative map but definitely not the friendliest to human eyes and intuition so this might be presented to a real estate agent to keep all the information at a glance but for a single customer might be more relevant to see only information pertaining its customer category and his preferred cluster (see Figure 16).

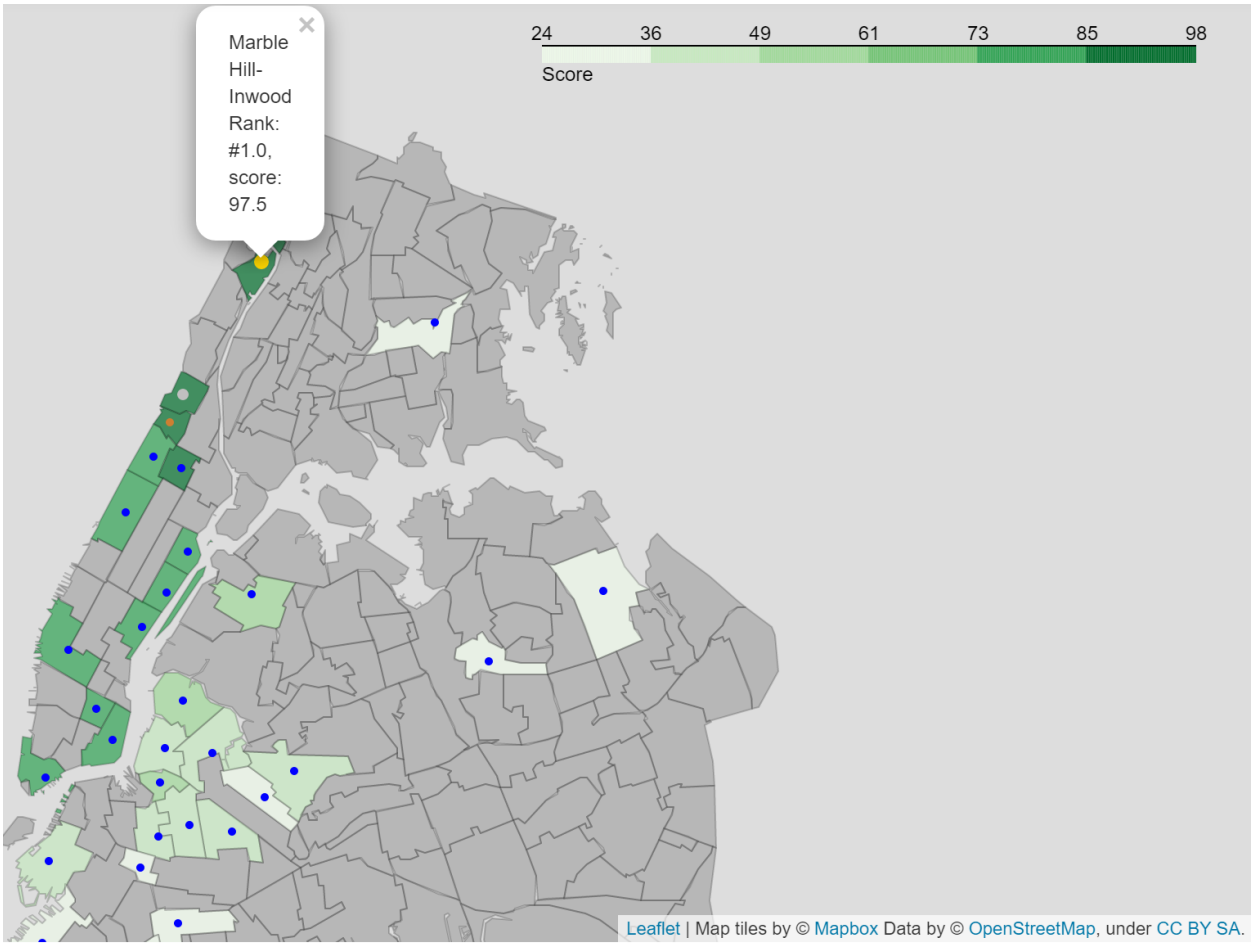


Figure 16 New York Neighborhoods color coded by score, each point on the map contains the information about the pertaining cluster and customer category only (cluster 1 – customer category "Professionals"). Different dimension and color are given to the marker of the top 3 Neighborhoods to get an immediate indication of their location.

Cluster categories

Once again information about the cluster can be given as an aggregate that is very informative but not straight forward (such as the one in Figure 17).

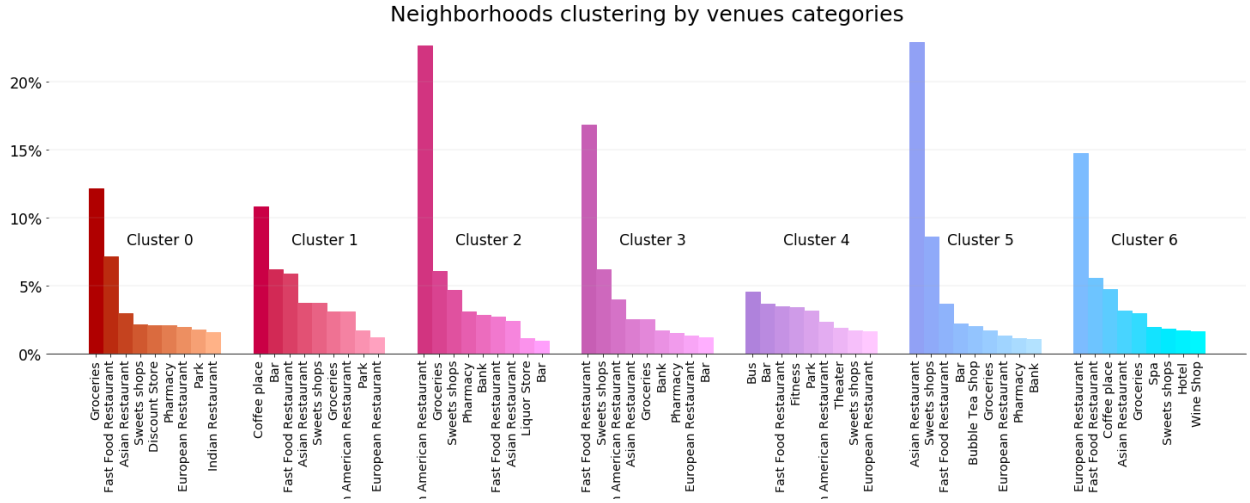
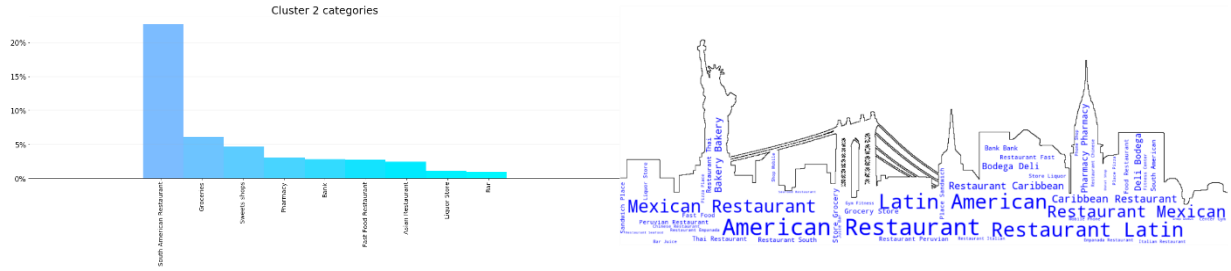


Figure 17 Cluster venues category composition

A more straight forward way could be presenting only the cluster of interest and maybe a more eye catching way of doing it could be utiliizing a word cloud.



Future development

Improvements can be made in different areas:

1- More variables can be considered for the scoring system, as we have seen distance is not a great predictor of price and many other factors can come into place such as criminality rate, public transport are just the firsts that comes to mind;
2- A better dimensions reduction can be made which would improve the results of the clustering algorithm;
3- Other clustering algorithms could be taken into consideration;
4- A better metric for defining distance could be implemented;
5- With more variables included a different scoring could be created;
6- More customer categories can be evaluated or better yet each customer could be interrogated to gauge how important each variable is for the subject.