A TAILORED NEIGHBORHOOD IN NYC

ANALYSIS OF NEIGHBORHOODS PROFILE FOR REAL ESTATE CONSIDERATION

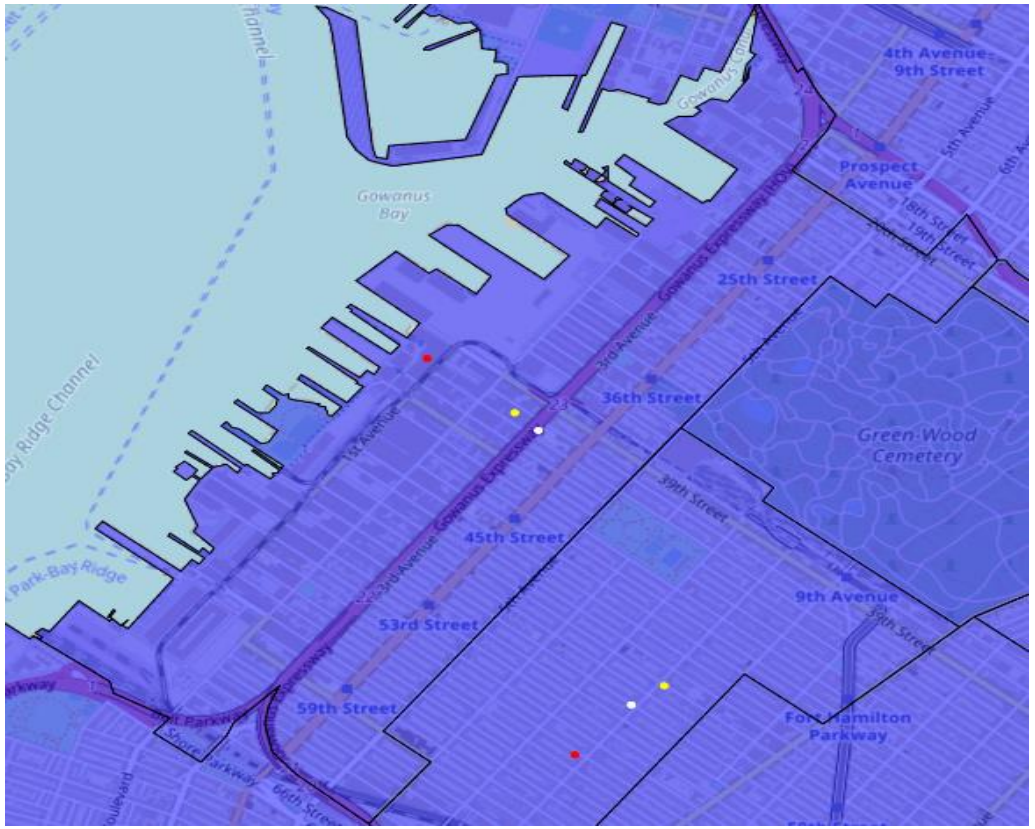
HOW NEIGHBORHOODS CLUSTERING CAN HELP REAL ESTATE

- Different profiles can be created and the city can be divided in those cluster
- Real estate agent can gauge customers taste and needs with few simple question
- Consequently the research of appropriate real estate can be focused in specific areas of interest
- A better fit between customers and neighborhoods can be found
- Customer satisfaction can be improved
- Customers and Agents time and efforts can be saved

DATA ACQUISITION AND CLEANING

- Geographical data and Median rent are collected from https://www.renthop.com/study/assets/new-york-city-cost-of-living-2017/nyc_col_geojson.js
- Venues information are provided by FourSquare
- Only Median rent and geography data are kept form the Renthop GeoJson
- A central point is calculated for each Neighborhood
- Venues information are filtered to keep only categories and their frequency in the proximity of the Neighborhoods
- A Distance metric is introduce and defined as 0 for all Neighborhoods in Manhattan and the distance to the closest Manhattan's neighborhood for all other Neighborhoods

CALCULATING NEIGHBORHOODS CENTER AND DIMENSION REDUCTION

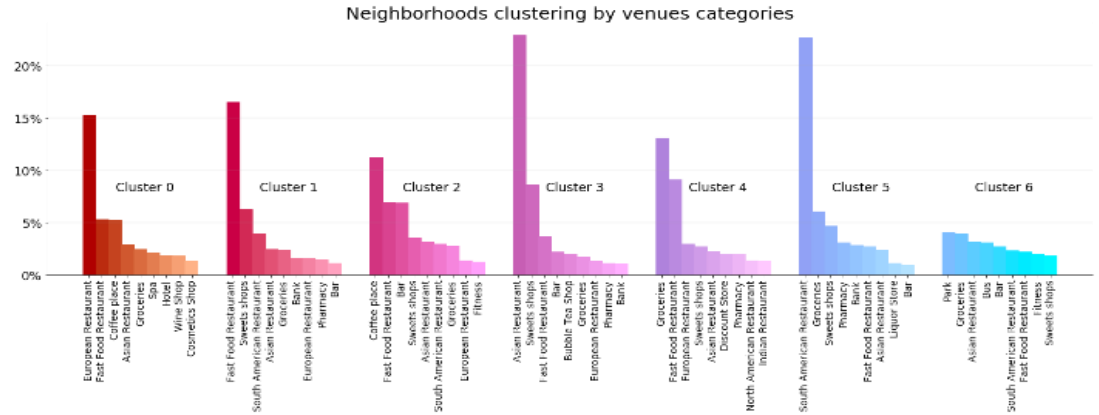


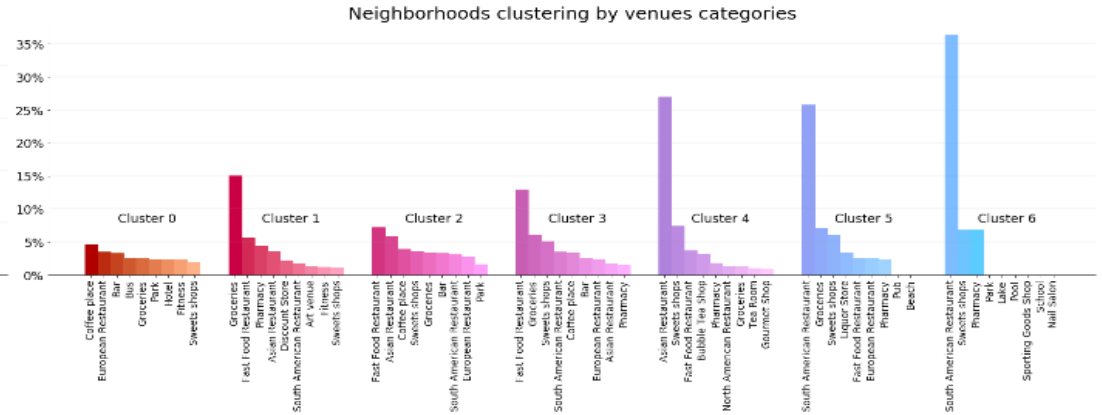
3 methods of calculating the neighborhoods center are taken into exam and the chosen one is the centroid of the quadrilateral formed by the North-East-South-West extremities

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

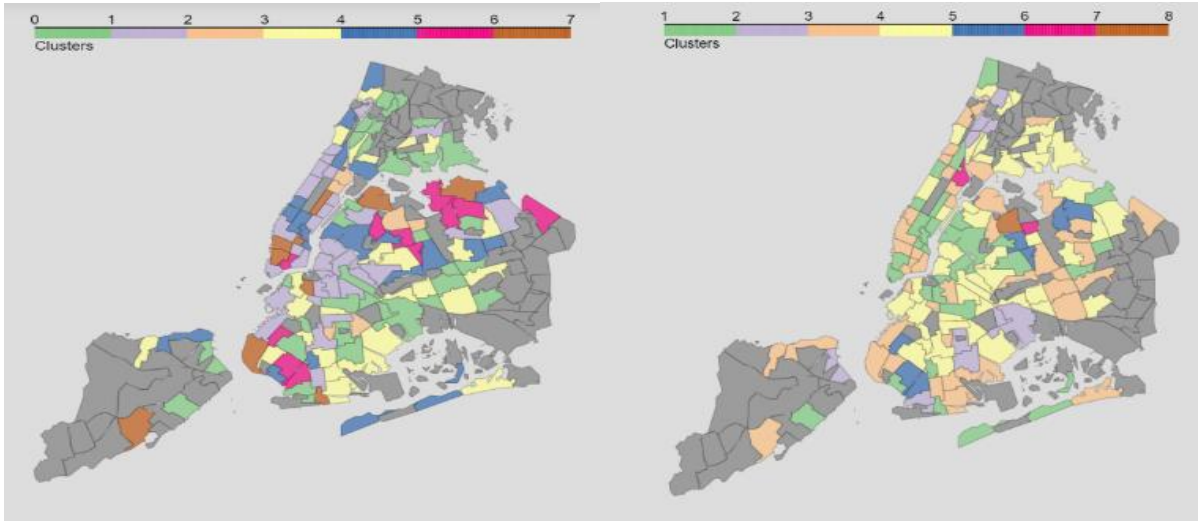
A predominance of food related categories is detected, an aggregation of these detailed options is made in order to have better clustering result not only based on restaurant categories

CLUSTERING:
K-MEANS VS HIERARCHICAL



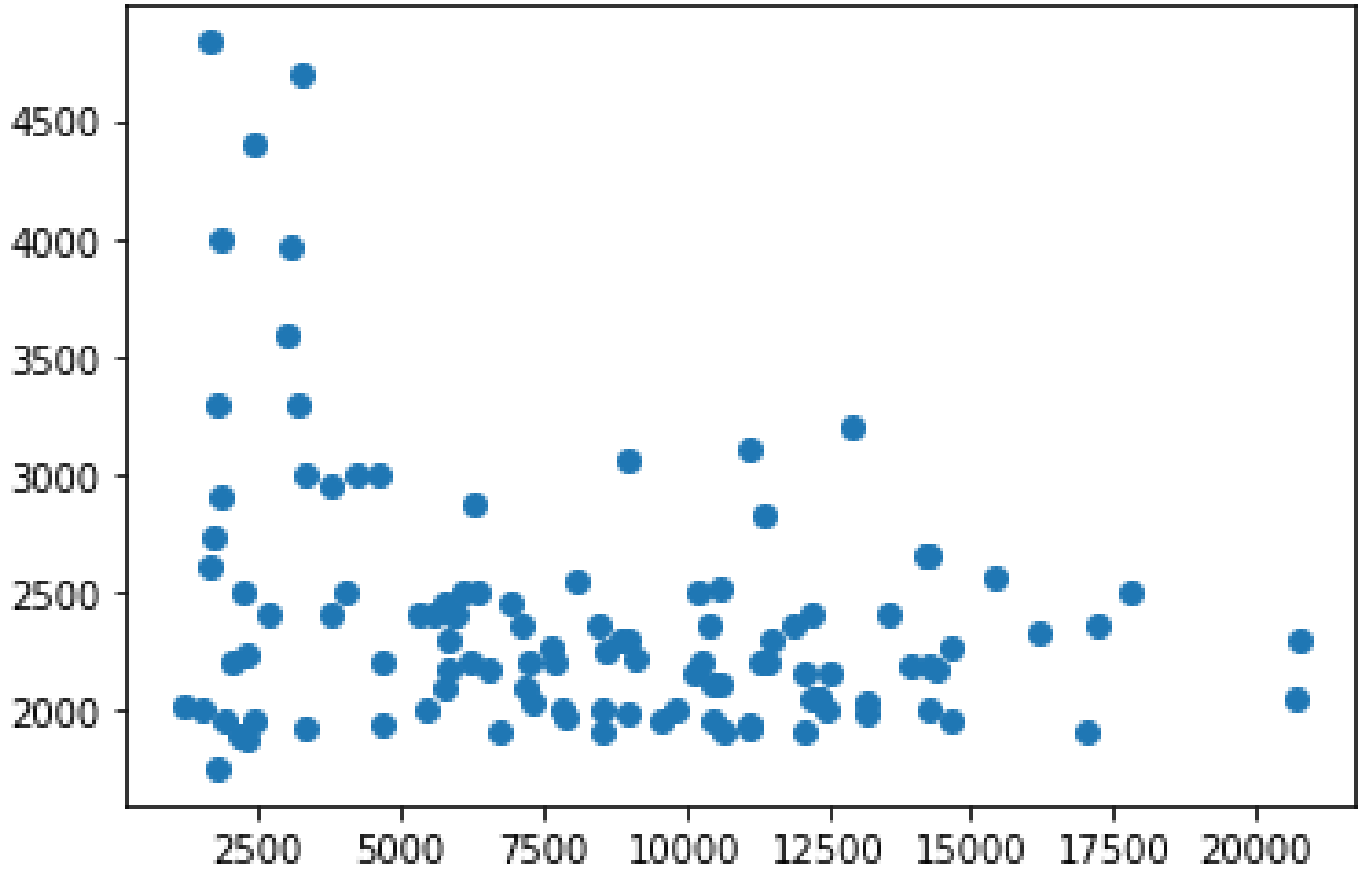


K-Means algorithm result in more balanced cluster's size and category distribution when compared to Hierarchical, this is partly due to the sparse nature of our datapoints and their distribution.



Hierarchical clustering have the advantage of not being a randomized algorithm, hence the clusters would be established always in the same way but the results are not promising as some clusters are composed by a single Neighborhood whilst others are very generic.

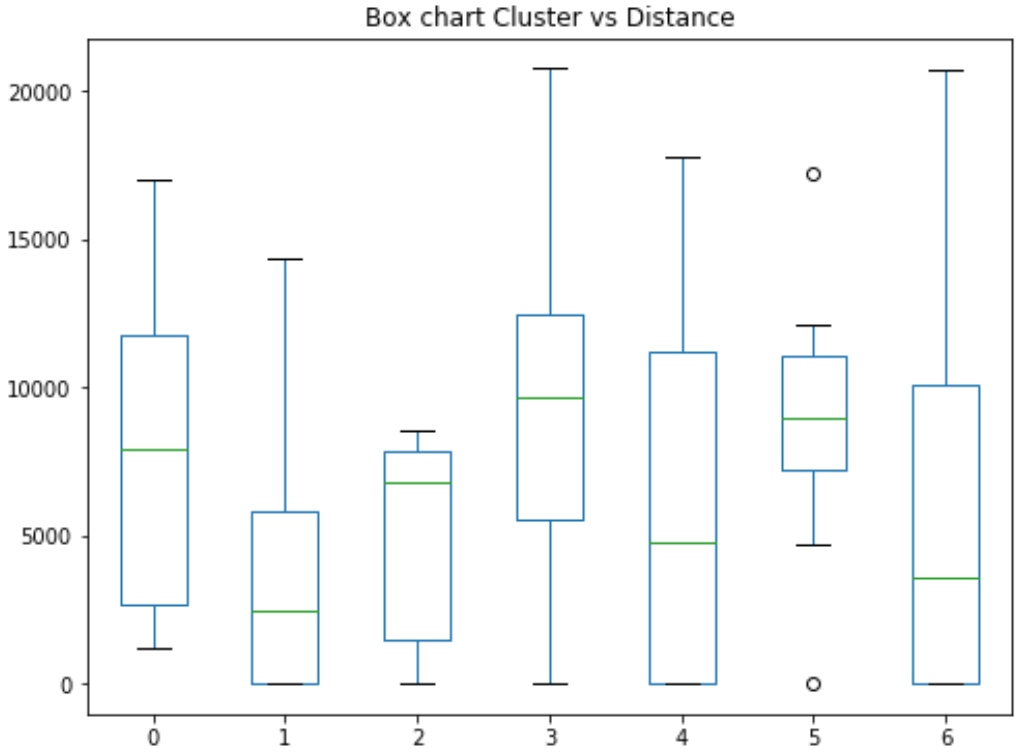
DATA ANALYSIS:
MEDIAN RENT VS DISTANCE

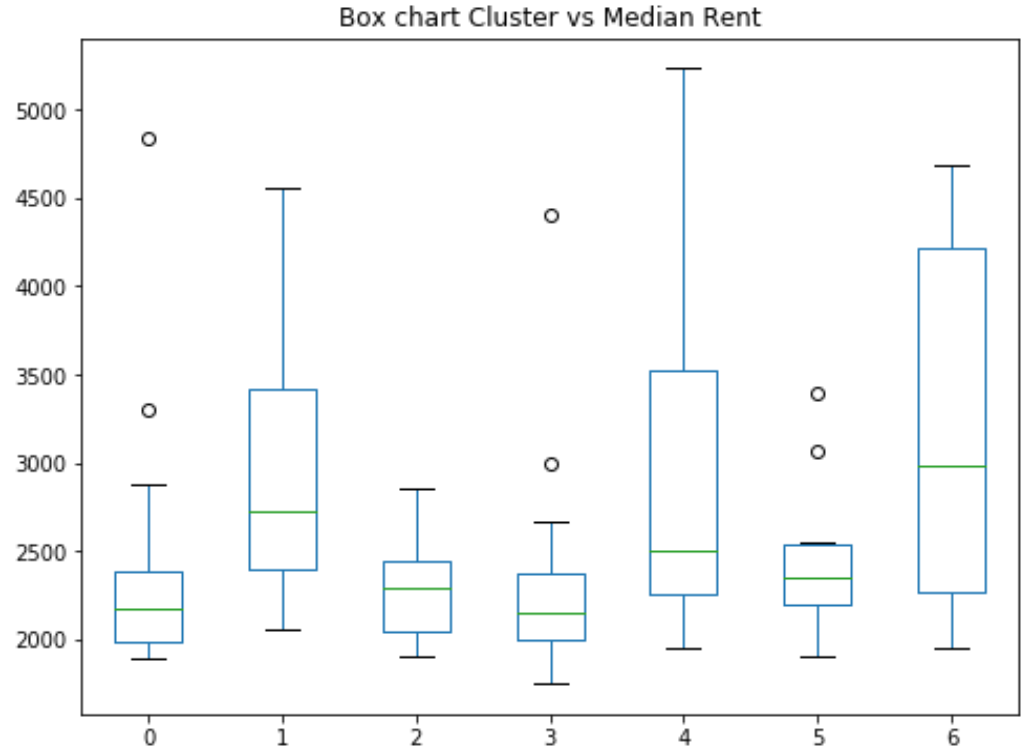


Here a Cluster plot of Median rent vs Distance, a reciprocal distribution seems to be at play but is not possible to fit a curve through this points accurately because of some datapoints, mainly in the bottom left side and central right area, we conclude the correlation is weak and there are 2 probable concurring cause for this:

- Our custom metric for the Distance variable
- Other factors influencing the price values like neighborhood reputation, crime rate, buildings quality, public transportation and so on

DATA ANALYSIS:
DISTANCE AND MEDIAN RENT VS CLUSTERS





Both Distance and Median rent are quite well distributed over the clusters which is desirable as we don't want cluster to be specific to a central location or price point so that any potential customer interested in a cluster would not be turned away from it because of the price tag or the distance. In our final score Neighborhoods will be evaluated only in comparison with the other neighborhoods in their cluster by associating its feature to a percentile in the cluster. All the scores will span from 0 to 100.

DATA ANALYSIS: CUSTOMER CATEGORIZATION

Not all people would value Distance and Price in the same way, some customers might prefer having a higher rent in favor of a closer location while the opposite is very much possible as well, in this analysis 3 mock up categories have been established and with them different weighting systems:

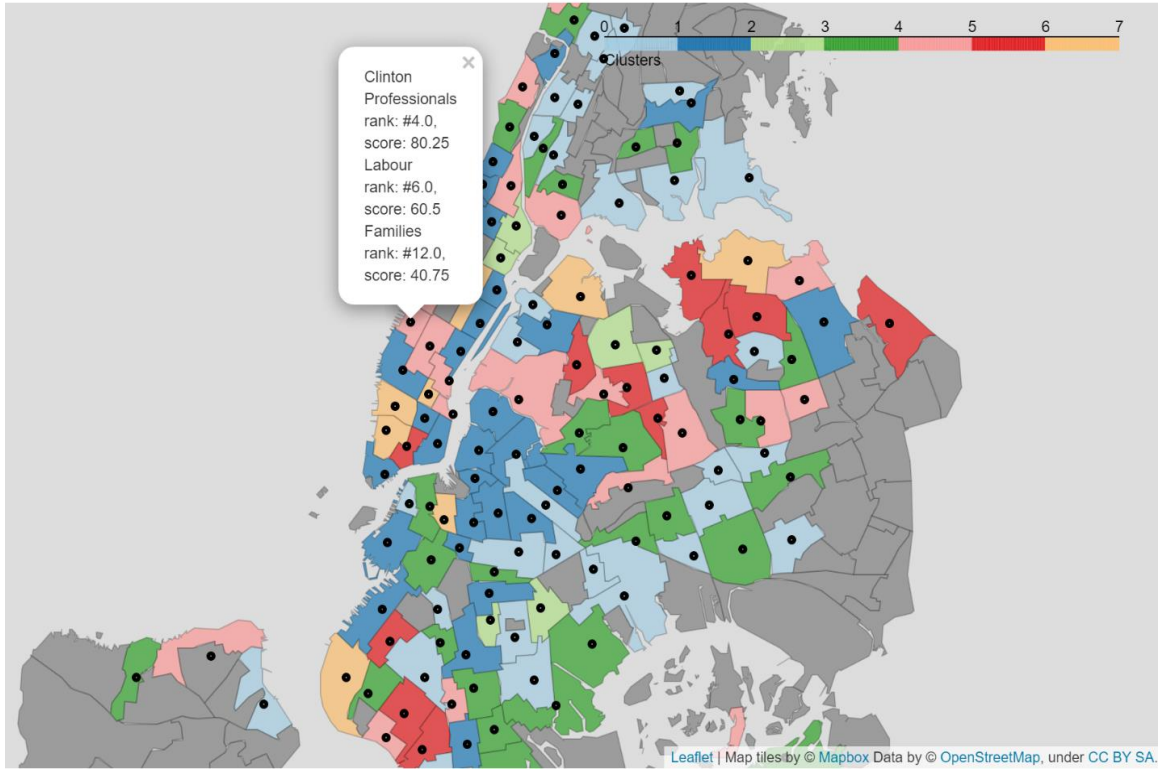
- Young and wealthy professionals with high incomes, working in Manhattan, that are more interested in having a short commute then having a high rent, for these people we will assume that distance is 3 times more important then rent price;
- Labor working in Manhattan that have to equally balance rent and distance;
- 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.

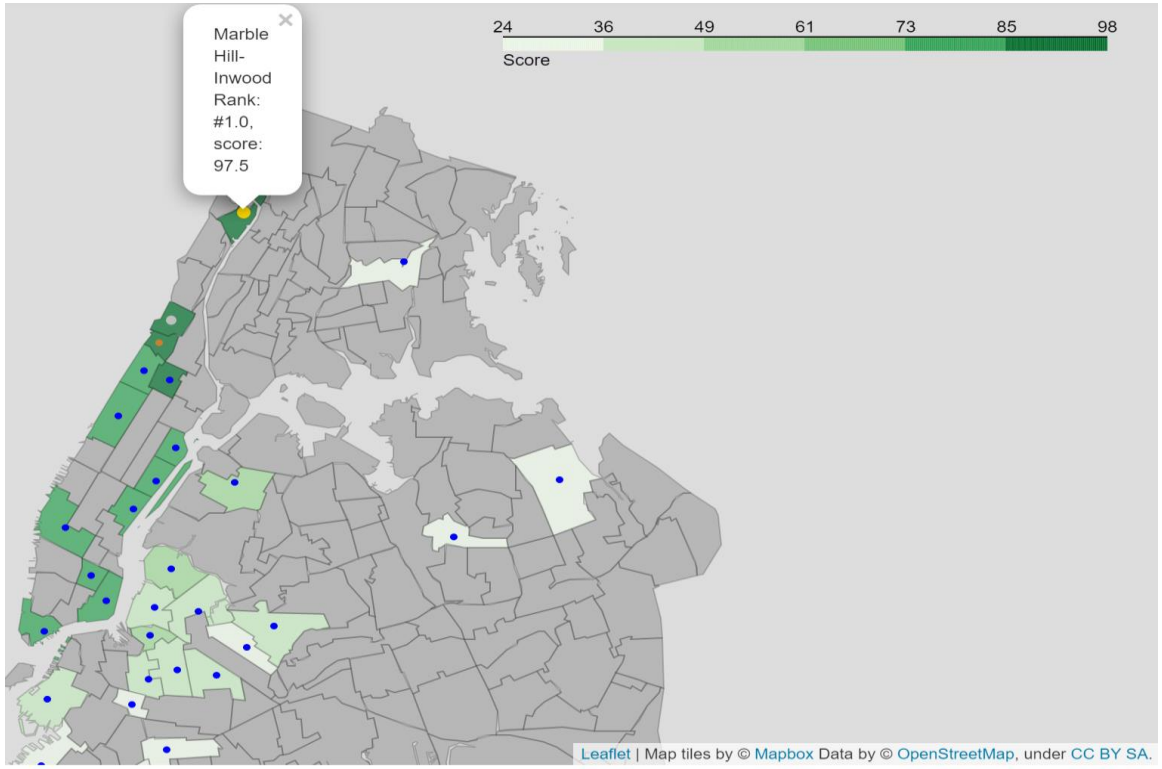
More categories may be evaluated or better yet is possible to gauge a client interest in the variables at play and weight the system accordingly.

RESULTS:
TOP NEIGHBORHOODS IN EACH CLUSTER AND CUSTOMER CATEGORY

	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

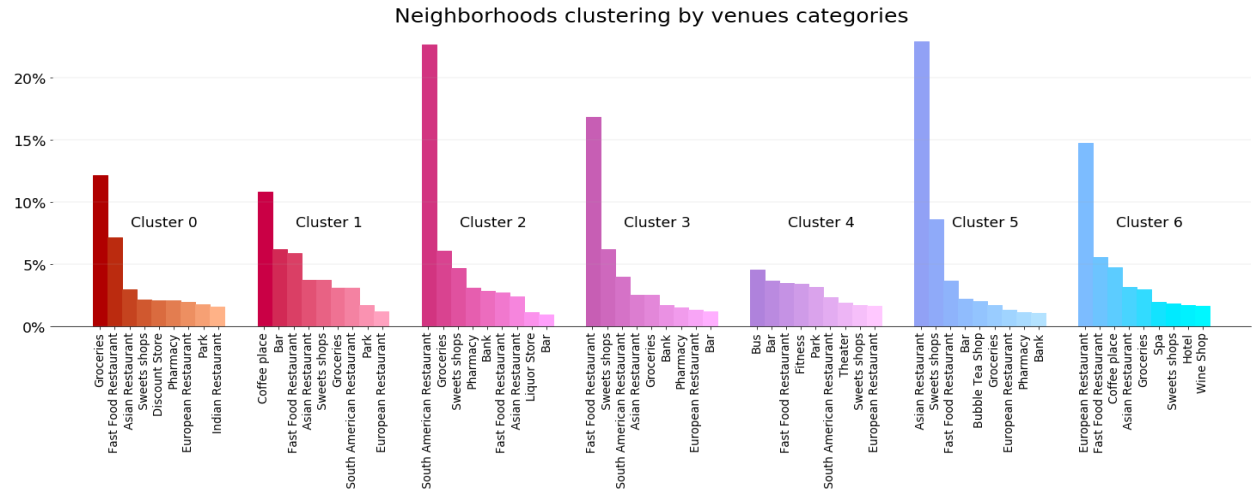
RESULTS:
VISUALIZATION OF CLUSTER DISTRIBUTION

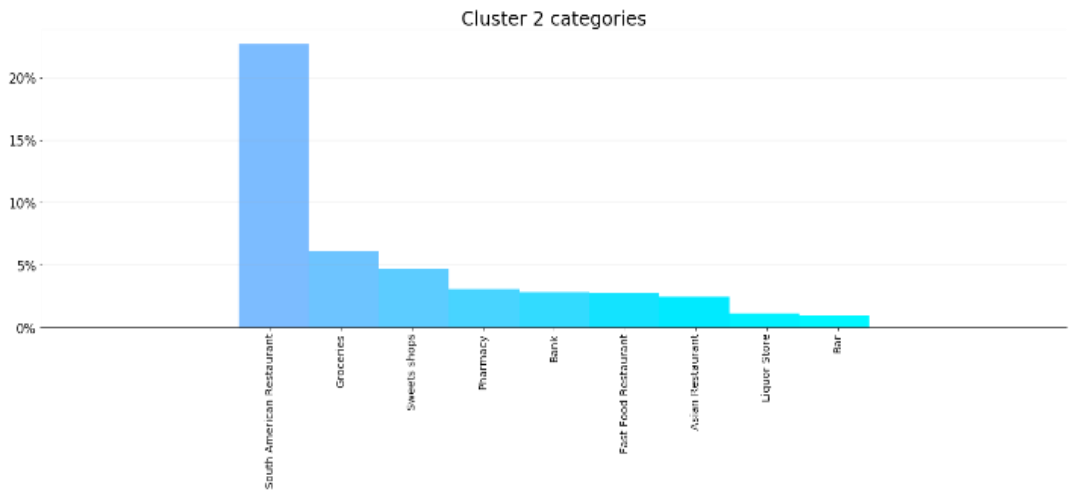




While the first map is full of information, color coded by cluster and contains all the relevant information about a Neighborhood might not be straightforward to understand, the second image is more specific and clear, so while the first one could be used by a professional to keep track of the big picture, the second one is suggested for clients that might be interested only on their preference.

RESULTS:
VISUALIZATION OF CLUSTER CATEGORIES





Once again while a general view contains more information a more specific one can be more clear to understand and, since not everyone is keen to read graphs a more stimulating image might be a better option for clients.
A word Cloud seems a good option to deliver a message clearly.

FUTURE DEVELOPMENTS

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.