Property Price in Venue Based Neighborhood Clustering

Anqi Lin

March 2019

1 Introduction

1.1 Background

Venue based neighborhood clustering groups geographically defined neighborhoods based on the number of similar venues found in each neighborhood. This method defines "similarity" based on the number of similar venues neighborhoods share. It is generally assumed that "similar" neighborhoods should foster similar success of the same business model. In other words, if a coffee shop business model in neighborhood A, then it should also be successful in neighborhood B that is "similar" as suggested by the venue based neighborhood system.

Similar venues uggest concentrated demand for similar types of business and is certainly a critical factor in growing a successful business. With this in mind, the venue based neighborhood clustering method generally produced moderately accurate results when compared against empirical evidence.

1.2 Problem

While there is truth to this assumption, it seems oversimplifying upon further examination. In a simple example, both the Whole Foods Market and C-Town Supermarkets fall under the "Supermarket" category for similarities in their product offering. However, while Whole Foods Market offers organic and preservatives-free products, C-Town offers lower price-points. Whole Foods Market are generally located near affluent neighborhoods, whereas C-Town are generally located near thrift ones.

1.3 Interest

It is clear that for operators from either of the two supermarket chains to make an informed decision, more factors need to be included in the clustering algorithm, chief among which: income.

2 Data Acquisition and Cleaning

2.1 Data Source

2.1.1 Geo Data

Geo data is conveniently provided by the NYC Open Data Project and made available through the Socrata Open Data API. This data set tabulates New York City neighborhoods by borough and provides latitude and longitude attributes.

2.1.2 Income / Property Price Data

For easy of data sourcing, property sale data is used to approximate income. Property sales data is also provided by NYC Open Data Project and the Socrata Open Data API. The data set encompasses 46 different transaction categories from condo/co-op sales to recreational facilities sales.

For the purpose of this study, the following slice of the data set is taken (Table 1):

#	**Category**
1	07 RENTALS - WALKUP APARTMENTS
2	08 RENTALS - ELEVATOR APARTMENTS
3	09 COOPS - WALKUP APARTMENTS
4	10 COOPS - ELEVATOR APARTMENTS
5	12 CONDOS - WALKUP APARTMENTS
6	13 CONDOS - ELEVATOR APARTMENTS

Table 1: Condo Group

Higher property sales price indicates a generally more affluent neighborhood, and therefore higher powering power. A price per square feet metric is calculated to eliminate the effect of square footage on sale price.

2.1.3 Venue Data

Like in previous modules, venue data is obtained through the FourSquare API and generated based on the neighborhoods contained in the geo data set and their corresponding latitude and longitude attributes.

2.2 Data Cleaning

The most challenging part of the data cleaning process involves lining up neighborhood names from the geo data set and the property sales data set. To start things off, the geo data set uses actual borough names (Manhattan: Alphabet City) while the property sales uses a numbering system to denote borough name (1: Alphabet City). A translation table was applied to the property sales data set to convert numbers back to names.

Other common causes of misalignments are as follows (Table 2):

#	**Description**	**Data Set 0**	**Data Set 1**
1	Granularity	Harlem	West Harlem, East Harlem
2	Concatenation	Midtown West	Midtown West (Hells Kitchen)
3	Abbreviation	South Jamaica	SO. Jamaica
4	Capitalization	Manhattan	MANHATTAN

Table 2: Causes for Misalignment

A few steps are taken including removing keywords like "East" and "West" appearing at the end of string; Removing texts after special characters like "-" and "(".

Other notable steps taken to further cleaning up the data sets are:

2.2.1 Obtaining Unique Neighborhood Names

Creative as the New York City, some boroughs contains neighborhoods of the same name. To overcome this issue, borough names and neighborhood names are concatenated to form unique strings identifying each neighborhood.

2.2.2 High Number of Missing Data in Property Sales Data

Some entries in the property sales data set has "0" for gross square foot and sale price columns, resulting in 0 per square foot and inf per square foot entries. Further analysis after correcting this issue indicates that entries still exist where price per square foot is unreasonable high or low. To correct for this, the 5 to 95 percentile was used.

3 Exploratory Data Analysis

3.1 Property Sales Price Data

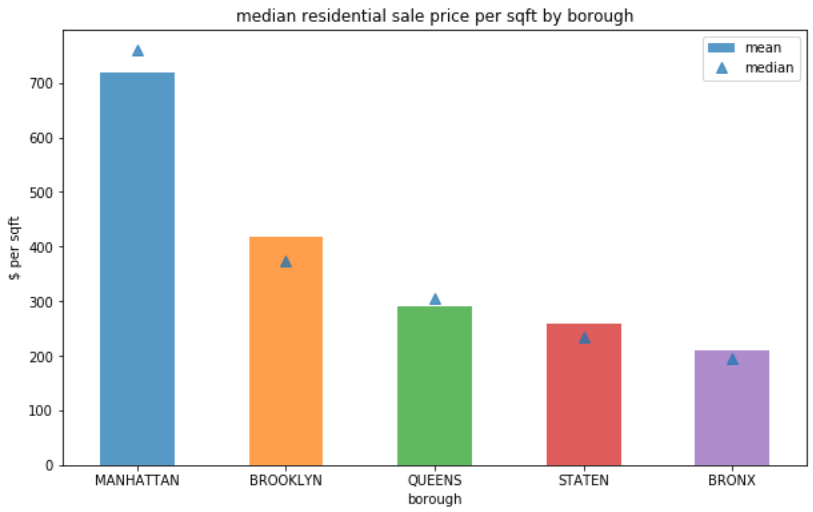

Figure 1: Mean Median Price

A quick look at the mean and median of the property sales (Figure 1) data set reveals that, not surprisingly, Manhattan has the highest mean residential sale price per square foot, leading by a considerable margin as compared to the other boroughs. Brooklyn comes in second, then Queens, Staten Island, and Bronx. It is worth noting that the median for Manhattan is significantly higher than the mean, indicating outliers on the lower end of the price spectrum or, in other words, a left-skewed population distribution. Brooklyn, on the other hand, exhibits the exact opposite.

This is reflective of the empirical observations that most of Manhattan has higher sale price with a handful of neighborhoods (eg. Harlem) as outliers on the lower end. On the other hand, most of Brooklyn has lower sale price (compared to Manhattan) with a handful of neighborhoods (eg. Brooklyn Heights) as outliers on the higher end.

Distribution analysis of the sale prices further supports this observation.

In addition, we observe that although data points from Manhattan, Queens and the Bronx generally fall into a single bell curve, data points from Brooklyn and Staten Island are broken into two groups with majority of the population falling into a bigger group on the lower end of the

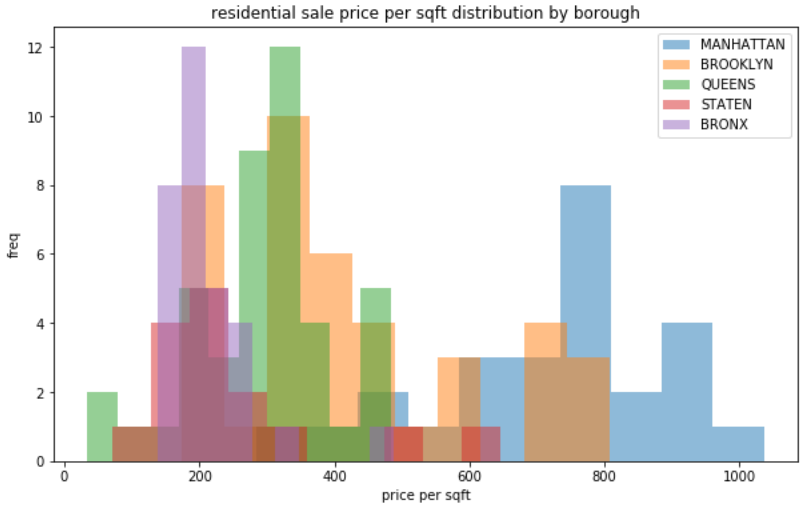

Figure 2: Distribution of Price

price spectrum and a few data points falling into a smaller group on the higher end, indicating the existence of high-end neighborhoods within these boroughs.

3.2 Venue Data

3.2.1 Venue Distribution Across Boroughs

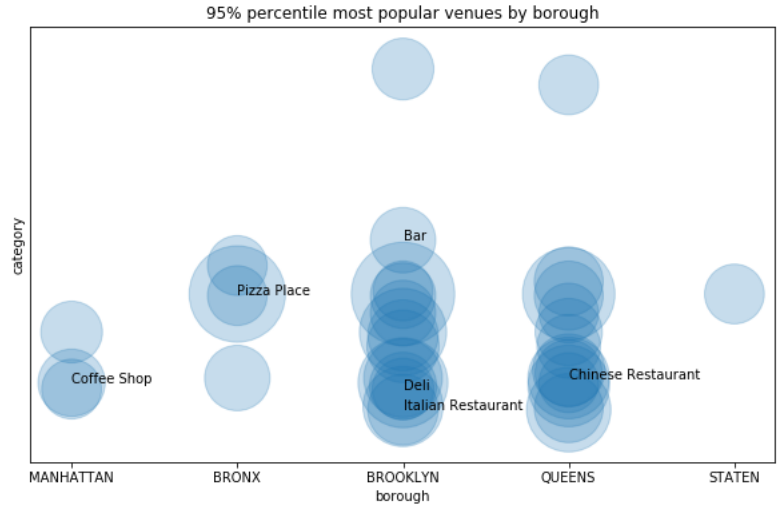

Figure 3: Popular Venues

It is easy to imagine that each borough has its unique most popular venues. For example, it would not be difficult to imagine that Coffee Shops must be really popular on Manhattan. However, a cursory read into Figure 3 reveals that boroughs also differ on the number and variety of venues within their respective top 5%.

Brooklyn and Queens, as two of the largest boroughs have a good variety of venue within their top 5%. Moreover, the respective 5% are highly similar. For example, both borough as Pizza

Place, Italian Restaurant, and Deli in their top 5% most popular venues. This observation seem to indicate that the two borough are highly similar.

On the other hand, Manhattan has only 3 different types of venues in their top 5% most popular venues: Coffee Shop, Italian Restaurant, and Cafe, seemingly indicate a higher kurtosis in the distribution.

3.2.2 Venues Types of "Lushness" of Neighborhoods

The following tables summarize the type of venues associated with the highest and lowest surrounding residential property price.

#	**High End Venue Type**
1	Yoga Studio
2	French Restaurant
3	Clothing Store
4	Women's Store
5	Cocktail Bar
6	Wine Bar
7	Furniture, Home Store
8	Hotel
9	Japanese Restaurant
10	Wine Shop

#	**Low End Venue Type**
1	Beach
2	Discount Store
3	Bus Stop
4	Fried Chicken Joint
5	Donut Shop
6	Mobile Phone Shop
7	Breakfast Spot
8	Supermarket
9	Bank
10	Caribbean Restaurant

When composition these table, the total city-wide number of venues per venue type is required to be higher than 10 to remove outliers. Overall the two tables are reasonably self-explanatory. High-end neighborhoods in general host venues that caters to demands higher up on Maslow's hierarchy like life-style (yoga) and entertainment (bars). On the other hand, lower-end neighborhoods are more packed with venues that fulfill more basic needs like sustenance (food joint) or finance (bank).

4 Predictive Modeling

Predictive models employed is K-Means Clustering. In addition, two distinct feature sets were created: one includes residential property sales series and one does not. In the end, the prediction results are compared against empirical evidence to evaluate the usefulness of K-Means clustering for identifying similar neighborhoods.

4.1 Selecting the Best K

A range of 2 to 12 clusters were calculated and the result plotted (Figure 4):

A best value of 8 is selected for K.

5 Result

Results are depicted in Figure 5 and 6.

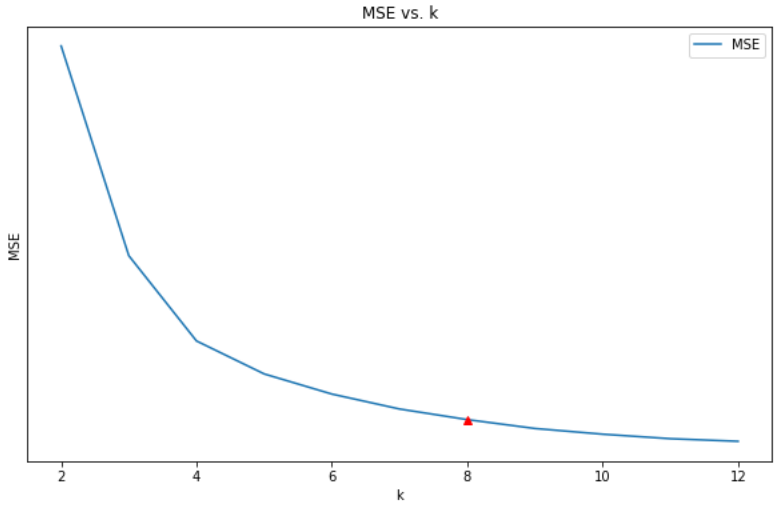

Figure 4: 8 Group Clustering without Property Sale Price

5.1 Without Property Sales Price (Figure 5)

8 Group K-Means Clustering (Clustering) was able to distinguish between higher and lower end neighborhood. Most of Manhattan, Downtown Brooklyn in Brooklyn, and Long Island City area in Queens have been classified as one group (orange). Most of the Bronx, Central Brooklyn, and part of Queens near Jamaica have been clustered as another (purple). Further, Lower Bronx, part of Queens near J.F.K Airport and Nassau county of Long Island are grouped together.

The result of the clusters seemed scattered and lack similarities from intra-cluster data points. The red cluster consists points in Chinatown, Inwoods Manhattan, and Pelham Bay Park Bronx. The blue cluster consists parts of Queens along the Long Island Railroad track and Manhattan Beach in Brooklyn. The rest of the clusters lack decent-sized population to be taken into consideration.

5.2 With Property Sale Price (Figure 6)

With property sale price data, the clusters seem more geographically connected. Specifically (Table 3):

6 Discussion

Evidently, by including property sale price data, the Clustering performed much better than with venue data only.

6.0.1 Looking into the Clusters from Empirical Evidence Perspective

From number 1 to 8 in Table 3, average property sale price decreases. Hind bias aside, the group seem to capture some crucial difference between neighborhoods, even within the same high end borough like Manhattan. Overall, property sale price seems to be a good complement to a otherwise limited data set of purely venue count.

#	Color	Includes:
1	Orange	Upper East and West, Midtown West, and Financial District on Manhattan; Downtown Brooklyn, Brooklyn Heights, and Park Slope in Brooklyn
2	Blue	Midtown East, East Village, Soho, and Tribeca on Manhattan
3	Red	Chelsea, Gramercy, Chinatown on Manhattan; Clinton Hill, Red Hook, and Gowanus in Brooklyn
4	Light Green	Manhattan Valley on Manhattan; Williamsburg, Prospect Heights, Madison in Brooklyn; Long Island City, Astoria, Rego Park in Queens
5	Teal	Ridge Bay and Ocean Parkway in Brooklyn; Elmhurst and Woodhaven in Queens
6	Purple	Inwood on Manhattan; Fordham in the Bronx; Brighton Beach in Brooklyn; Ozone Park and Richmond Hill in Queens; Stapleton and Willowbrookn on Staten Island
7	Bright Green	Most of Bronx; Part of Queens neighboring Nassau County of Long Island
8	Dark Orange	Misc. & Errors

Table 3: Clusters

6.1 Further Directions

A number of other factors could have been included in the predictive modeling. A few examples could be: crime data, average household level of education, or average age of residents.

Overall the decision to open a new branch for an existing business goes way beyond the analysis of demand. Further research into the cost structure of the business in question could help better finding the most suitable expansion. An example could be looking into the property sale / rental price of commercial real estate.

As always, other clustering methods could be used to cluster New York City neighborhoods.

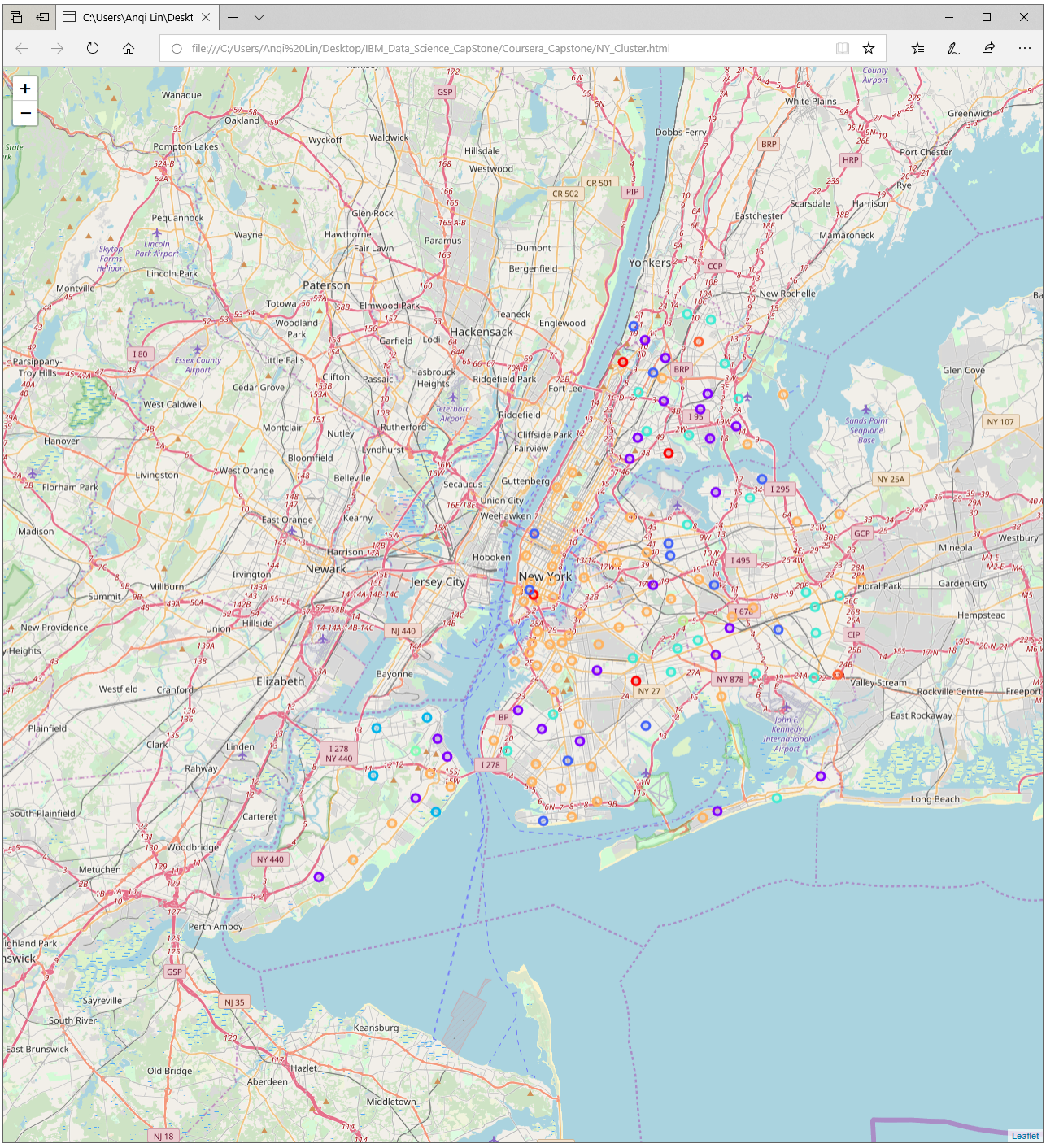

Figure 5: 8 Group Clustering without Property Sale Price

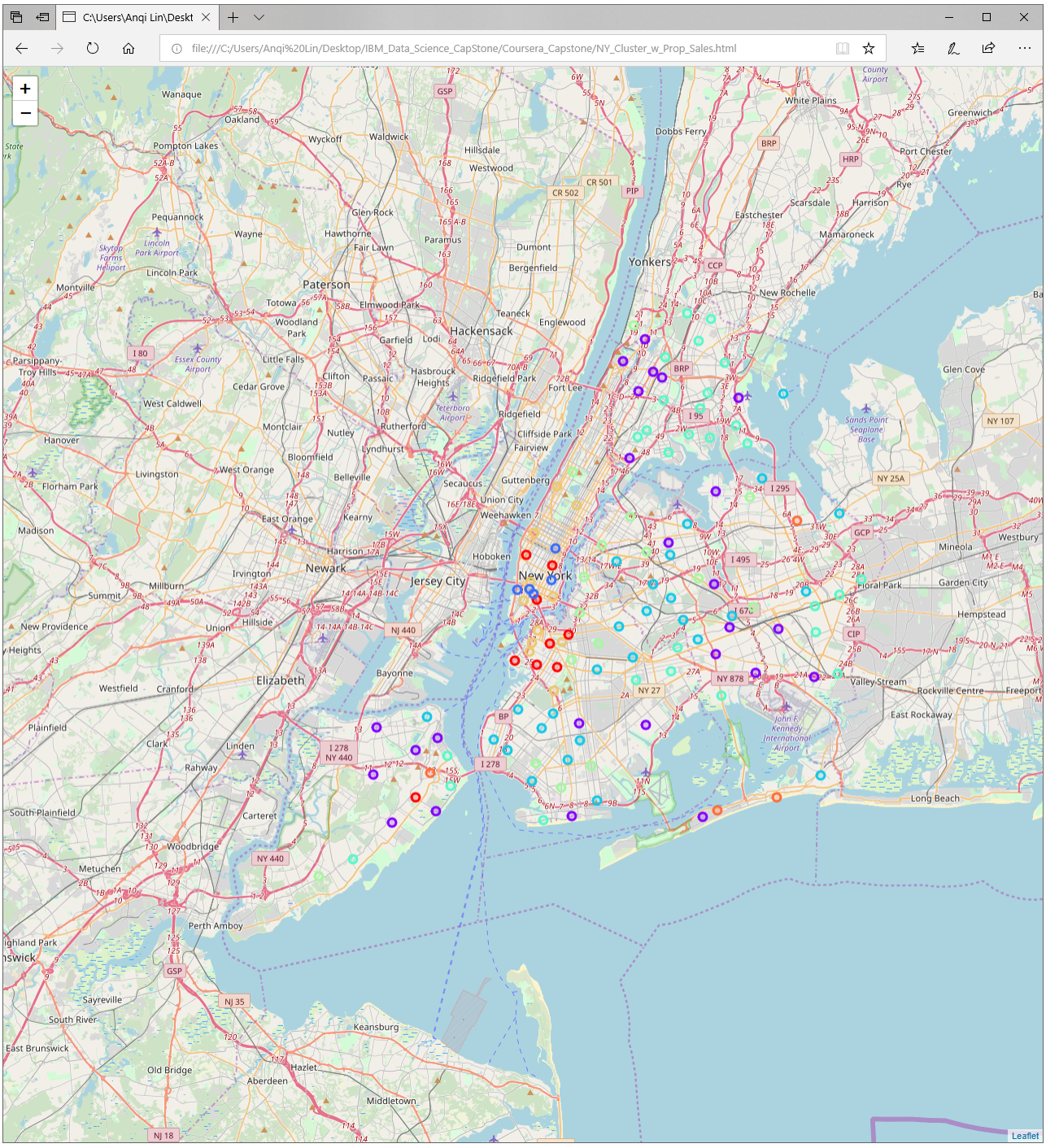

Figure 6: 8 Group Clustering with Property Sale Price