The Battle of Neighborhoods

Minshen Li

September 26, 2019

1. Introduction

When deciding on where to start their business or open a new branch, firms want to find a location where the new store has the highest potential to thrive. While factors such as real estate costs, local economy, and demographics profiles are important bulletin points for consideration, it is critical to understand the basics: local preferences for the business. While human preferences are subjective and cannot be directly inferred, examining how local businesses perform is a reasonable way to make the inference. For example, if there are many Italian restaurants in a certain neighborhood, that means in this area, people like Italian food. Nonetheless, this does not mean one should rush into the market and open an Italian restaurant in this specific area. The market has been filled with competitors in Italian food. In such case, a neighborhood that is similar to this "successful area" that has not yet many Italian restaurants available should be considered. Similarly, when a firm has succeeded in a local area and is contemplating opening a new branch, the firm should consider opening a new store somewhere else that is similar to the current area rather than insert one more store in the original place. In a word, it is important to understand the similarity among different neighborhoods in terms of preferences for different types of business and. I get a snapshot of local preferences for certain types of businesses by looking at the distribution of types of businesses that thrive in the local area to

In this case study, I analyze the similarity among the neighborhoods of the three cities, Philadelphia (Center City), New York City, and Toronto through cluster analysis and computation of Euclidean distance matrix. This analysis helps firms formulate steps in pinning down areas that potentially favor the business they do.

2. Approach

The distribution of different categories of venues is an accurate snapshot of what the neighborhood looks like and what kinds of venues thrive in the local area. Therefore, I collect information on the distribution of venue categories within a neighborhood using Foursquare API data. I run cluster analysis to capture a sense of what neighborhoods appear relatively similar. Using Word Cloud, I capture what the top venue categories are within each cluster. I measure how dissimilar the clusters are to each other using Euclidean distance across clusters. Finally, I obtain the distribution of clusters for each city, and compute the Euclidean distance among the cities based on the distribution to quantify their dissimilarity.

3. Data

I scrape neighborhood information of Philadelphia from Wikipedia. I use Foursquare data for the cluster analysis. Given the coordinate of a selected location, Foursquare API generates the most popular 100 venues within the radius of 2 kilometers of the location. I input the CenterPoint of each neighborhood and use the generated venues to approximate the most popular venues within that neighborhood. I define similarity between two neighborhoods as the similarity in the distribution of venue categories between the two neighborhoods. Foursquare data has category information for each venue. Using such information, I obtain the distribution of venue categories in each neighborhood by computing the frequency of occurrence of venue categories within the neighborhood. Table 1 gives a snapshot for what a data entry, also a neighborhood looks like (not from real data, just for the purpose of illustration). I calculate Euclidean distance among neighborhoods based on such constructed venue category distributions. Two neighborhoods are more similar if the Euclidean distance is smaller.

Table 1 A snapshot of an entry of the constructed dataset used for analysis

City	Borough	Neighborhood	Frequency of Category X	…	Frequency of Category Z
Philadelphia	Center City	Chinatown	0.05		0.22

I focus on New York City, Philadelphia, and Toronto. Table 1 provides an overview of the neighborhoods of the three cities. Philadelphia has 13 boroughs and 161 neighborhoods, New York City has 5 boroughs with 306 neighborhoods, and Toronto has 11 boroughs with 101 neighborhoods. In total, there are 570 neighborhoods. Table 2 provides an overview of the neighborhoods of each city across boroughs.

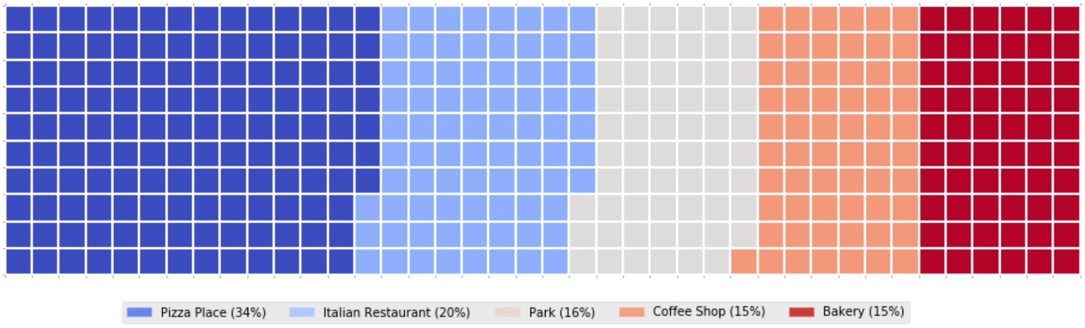
Table 2 Overview of Neighborhoods and Boroughs of the Three Cities.

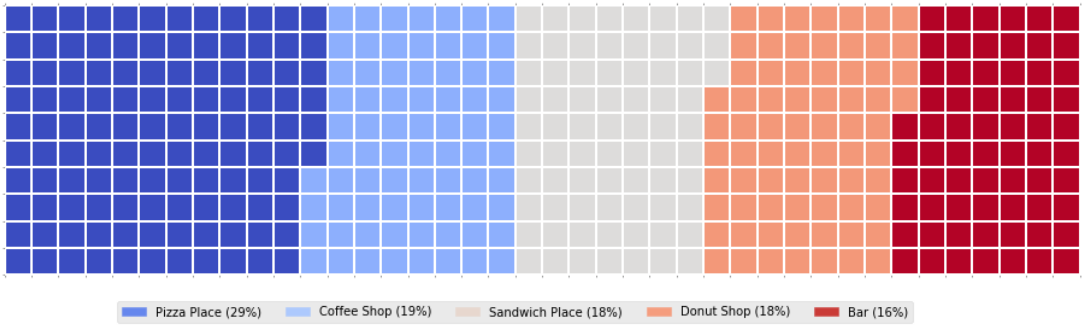
Philadelphia	
Boroughs	**Num. of Neighborhoods**
South Philadelphia	29
West Philadelphia	26
Center City	18
Near Northeast Philadelphia	18
Far Northeast Philadelphia	16
Southwest Philadelphia	13
Lower North Philadelphia	12
Olney-Oak Lane	8
Germantown-Chestnut Hill	6
Roxborough-Manayunk	5
Upper North Philadelphia	5
Bridesburg-Kensington-Port Richmond	5
Total	**161**
New York City	
Boroughs	**Num. of Neighborhoods**
Queens	81
Brooklyn	70
Staten Island	63
Bronx	52
Manhattan	40

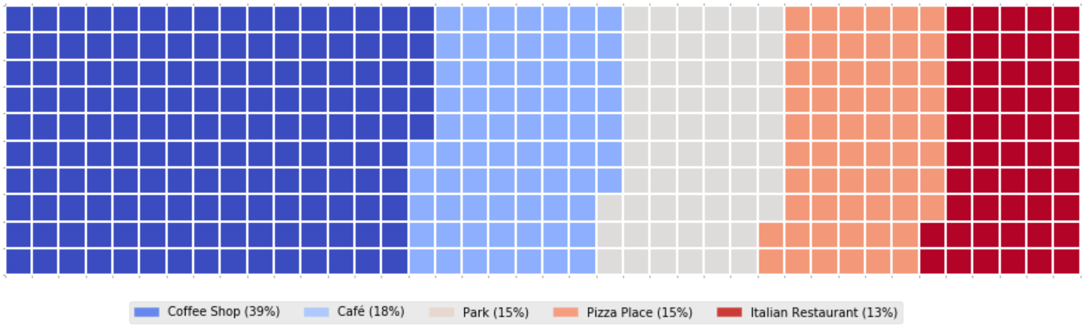
Total	**306**
Toronto	
Boroughs	**Num. of Neighborhoods**
North York	24
Downtown Toronto	18
Scarborough	17
Etobicoke	12
Central Toronto	9
West Toronto	6
East Toronto	5
East York	5
York	5
Queen's Park	1
Mississauga	1
Total	**103**
Total: 570	

Overall, 50,894 Top-100 venues from 479 categories are generated from Foursquare API. New York City has 28,304 (55.6%) popular venues, Philadelphia has 14,144 (27.8%) popular venues, and Toronto has 8,446 (16.6%) popular venues. The average numbers of venues per neighborhood are respectively 92.5 in Philadelphia, 87.9, and 83.6. New York City is the financial capital of U.S., so not surprisingly, the per-neighborhood number of venues is the highest among three.

Figure 1 shows Top 5 categories (based on occurrences of venues) for each city as well as overall. The Top 5 categories in New York City include pizza places, Italian restaurants, parks, coffee shops, and bakery shops. In Philadelphia, the Top 5 categories include pizza places, coffee shops, sandwich places, donut shops, and bars. In Toronto, the Top 5 categories include coffee shops, cafes, parks, pizza places, and Italian restaurants. There is a large overlapping in the popular categories across the three cities, except that the ranking of the popularity among the categories looks a bit different. Overall, pizza places, coffee shops, and Italian restaurants are prevalent “everywhere”. Figure 2 shows the joint distribution of frequency of sandwich places and parks. From the graph it appears that the 570 neighborhoods are quite different and could be divided into clusters. For example, a large square-shaped chunk of neighborhoods has both low to medium value of frequency of parks and of pizza places; the other two clusters involve high frequency of parks but low frequency of pizza places, and high frequency of pizza places but low frequency of parks.


New York City


Philadelphia


Toronto

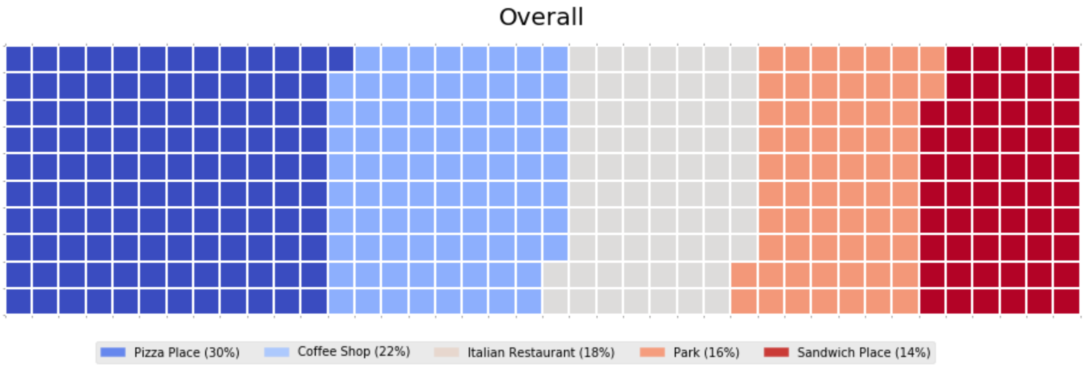


Figure 1. Top 5 Categories of New York City, Philadelphia, Toronto, and Overall

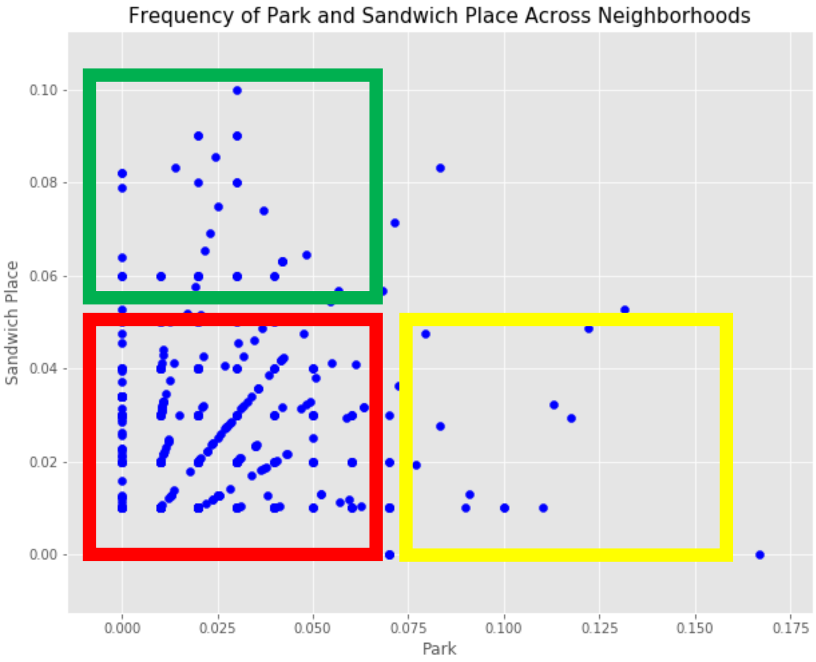


Figure 2. Joint distribution of frequency of sandwich places and parks

4. Methodology and Analyses

4.1 K-Means Cluster Analysis on Neighborhoods

I adopt the conventional method for cluster analysis - K-Means clustering method. For model selection, I use the elbow method: The optimal k value is where the decrease in distortion (within-cluster sum of square) starts slowing down significantly. Figure 3 shows the results.

While the distortion keeps dropping significantly as the number of clusters increases, I choose k=4 as the decrease in distortion for all k values starting from 4 are smaller than the decrease in distortion when k is increased from 3 to 4. Therefore, I go with four clusters.

Table 3 shows the results. The largest cluster (Cluster 1) accounts for 36% of the neighborhoods, while the smallest cluster (Cluster 3) only has 57 neighborhoods (10%). Cluster 2 and 4 have similar sizes, respectively accounting for 32% and 22% of the total neighborhoods.

Table 3. Distribution of Cluster Labels

Cluster Label	**# of Neighborhoods**	**Proportion**
1	205	0.36
4	182	0.32
2	126	0.22
3	57	0.10
Total	560	1.00

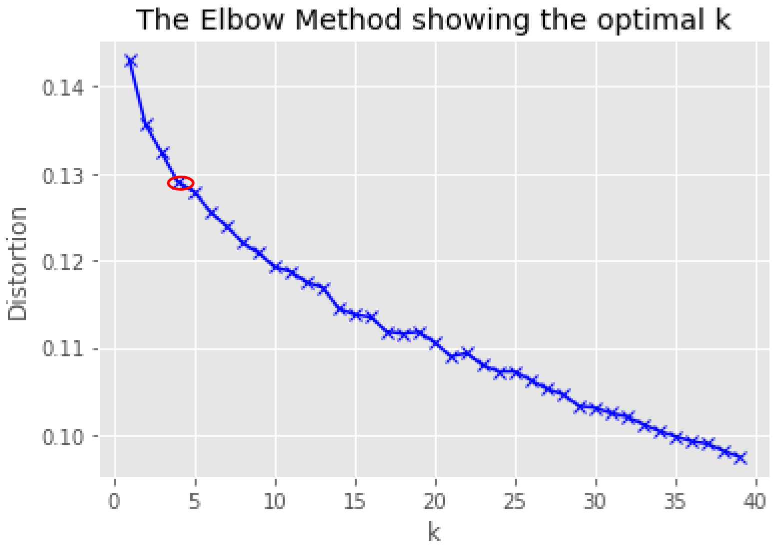


Figure 3. Distortion Across k Values

4.2 Cluster Keywords

I use word clouds to summarize the key categories within each cluster. In word clouds, the size of the word is proportional to the frequency of the word. Therefore, if a category is more popular, the size of the category is bigger. Figure 4 shows the results. In Cluster 1, the top categories include coffee shops, pizza places, gyms, bars, cafés, bakery shops, and Italian restaurants. In Cluster 2, the top categories include pizza places, donut shops, and pharmacies. In Cluster 3, the top category is mainly pizza places, sandwich places, and fast food restaurants. Note that Cluster 3 has only 57 neighborhoods, thus few categories, so the word cloud looks sparse. In Cluster 4, the top categories include Italian restaurants, pizza places, and sandwich places. In Cluster 1, the size of "Coffee Shop" is extremely large, suggesting that coffee shops are very popular in those local areas. In Cluster 2, pizza places, donut shops, and pharmacies seem equally very popular.

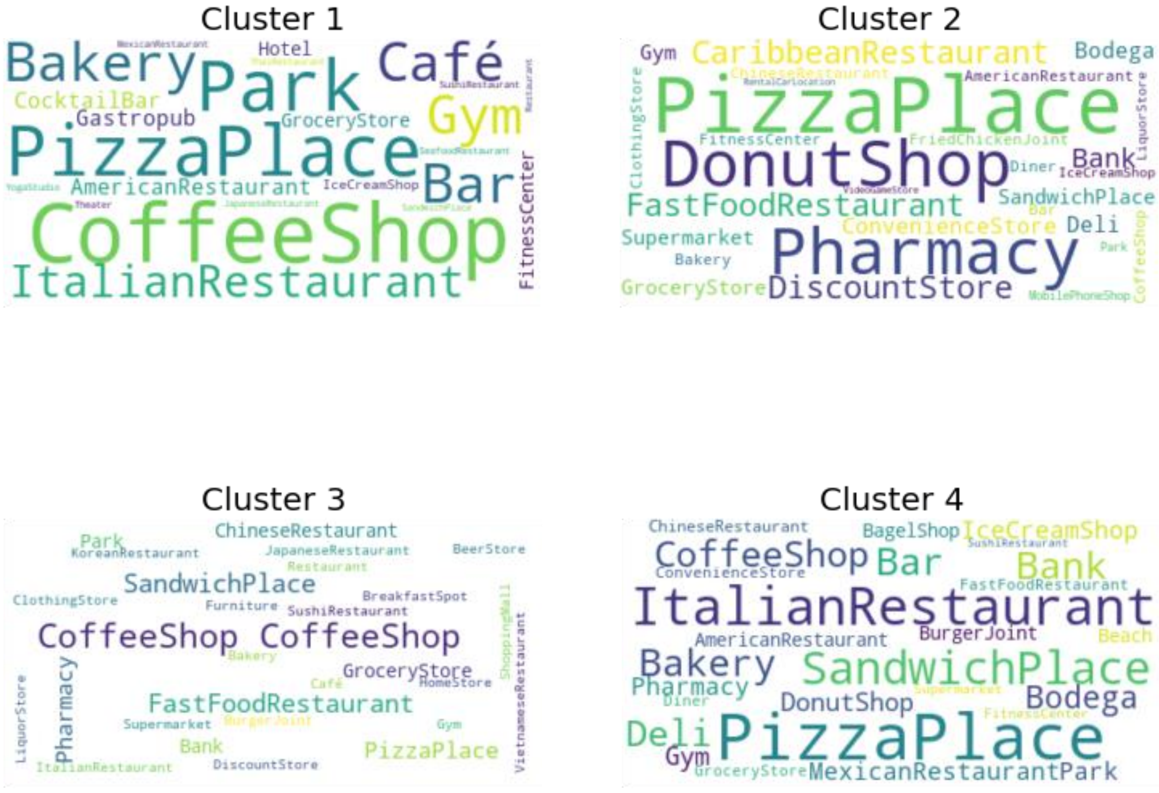


Figure 4. Top Categories in Different Clusters

It is highly possible that neighborhoods within the same cluster share common preferences for merchants, foods, and activities. The key words thus help firms identify potential areas for business. For example, if people want to open a donut shop, then they should first consider the neighborhoods in Cluster 2 that have non-saturated market of donuts.

4.3 Similarity of Clusters

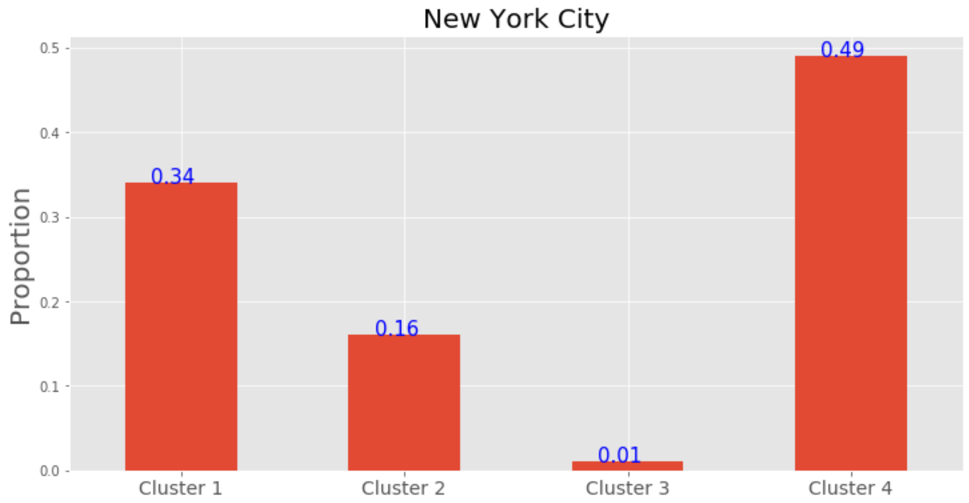
If all markets within Cluster 2 are saturated, which cluster should the donut shop owner consider? It is useful to know how similar or dissimilar different clusters are to each other. The distance between two clusters is computed as the Euclidean distance between their centroids. Table 4 shows the calculation results. The closest pairs of clusters include (Cluster 1, Cluster 4) and (Cluster 2, Cluster 4). Continuing with the previous example, if neighborhoods in Cluster 2 are all crowded with donut shops, then neighborhoods in Cluster 4 can be considered next.

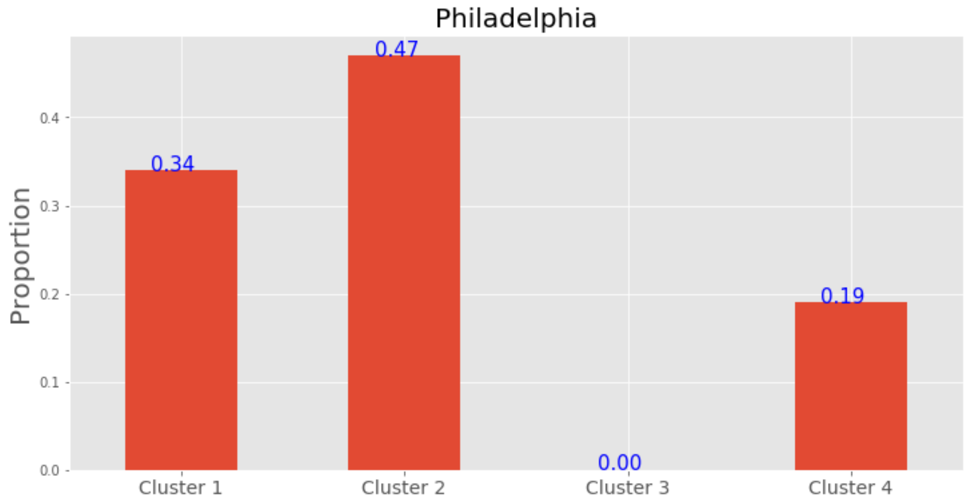
Table 4. Euclidean Distance Among Clusters

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1	0	0.126	0.113	0.083
Cluster 2		0	0.127	0.084
Cluster 3			0	0.118
Cluster 4				0

4.4 Distribution of Clusters Among Cities

To get a sense of the similarity across cities, I use the cluster labels of neighborhoods and obtain the distribution of clusters within each city. Figure 5 shows the results. New York City, not surprising, has very rich cluster structure as it has all four clusters. Philadelphia lacks Cluster 3, and Toronto lacks Cluster 2. New York City uniquely has a large size of Cluster 4, which is famous for the combination of Italian restaurants, pizza places, and sandwich places. Philadelphia uniquely has a large size of Cluster 2, where donut shops and pharmacies are top categories, therefore, firms should consider opening these stores in the neighborhoods of Cluster 2 in Philly. Finally, Toronto uniquely has a large size of Cluster 3, which has top categories like fast food restaurants. Therefore, if firms want to open a fast food restaurant, consider the neighborhoods in this cluster in Toronto.





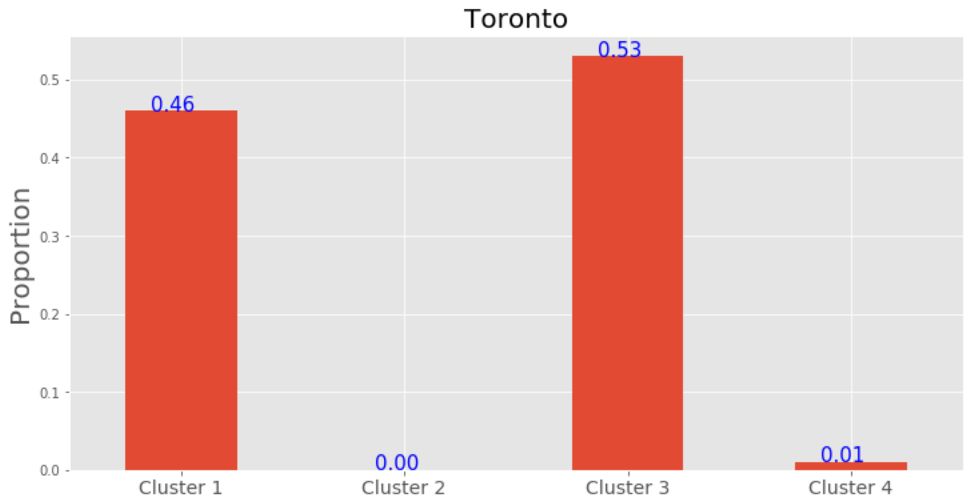


Figure 5. Distribution of Clusters for the Three Cities

4.5 Similarity Among Cities

To get a sense of how similar the cities are, I compute the Euclidean distance using the distribution of clusters of the cities. Table 5 shows the result. New York City is more similar to Philadelphia than Toronto. Ideally, when a large number of cities are involved, further cluster analysis can be conducted, which is out of the scope this project, but left for future exploration.

Table 5. Euclidean Distance Among Cities

	New York City	Philadelphia	Toronto
New York City	0	0.432	0.735
Philadelphia		0	0.741
Toronto			0

5. Concluding Remarks

In this project I use Foursquare API location data and cluster analysis to find clusters of neighborhoods in New York City, Philadelphia, and Toronto. Neighborhoods are regarded as similar if they share similar distributions of popular local venues. Similar neighborhoods are regarded as sharing similar preferences for types of local business. The cluster results are Threefold: First, the neighborhoods can be divided into four clusters. Therefore, for a certain category, to start up a new business, firms should start with clusters where the category is hot and search for a neighborhood that has non-saturated market. To extend the business by opening a new branch, neighborhoods within the same cluster but with non-saturated markets are the ideal places for consideration. Second, two clusters are more similar if the Euclidean distance between them is smaller. Based on the Euclidean distance, for each cluster, a ranking of similarity could be generated, so firms could go down the ranking list to search for the next best places if their best choice of market has already been saturated. Third, based on the distribution of clusters, we can cluster cities and rank cities based on similarity. In this project, I find that New York City is more similar to Philadelphia than to Toronto. Therefore, firms should first consider switching between New York City and Philadelphia for extending their business outreach.

There are three directions for pursuit in the future: First, with a larger database involving more cities, we can further conduct cluster analysis on the cities and summarize the key categories for city clusters. Second, future exploration could also take advantage of more information, such as local employment rates, housing prices, etc., to generate more insightful cluster results. Third, algorithm improvement. The dataset seems not perfect for k-means cluster analysis, though it is the fastest, easiest way to cluster data. Using k-means cluster method results in elbow method not generating clearly optimal k value. Future work should find better solutions to that issue.