Comparing Neighborhoods of Manhattan with old Toronto and Paris

Meredith Cao

April 2020

1. Introduction

1.1. Background

Is New York City more like Toronto or Paris?

It would be interesting to compare the neighborhoods of these multicultural cities as the financial capital of different countries and determine how similar or dissimilar they are. However, While the area of Paris is 40.7 mi² with its 20 Arrondissements, in our data New York City is 302.6 mi² with 302 neighborhoods and Toronto 243.3 mi² with 98 neighborhoods. Old Toronto(https://en.wikipedia.org/wiki/Toronto) and Manhattan (22.82 mi²) as the most densely populated part of the respective city, should be more comparable.

1.2. Problem

Therefore, our **problem** is refined as: Is Manhattan more like core Toronto or Paris? To be specific, which neighborhoods in Toronto and Paris would have a similar setting or vibe to Manhattan?

1.3. Interest

If a NYC-based brand would like to expand its business to the overseas, especially the brick-and-mortar, whether they should expect similarity or dissimilarity in its options of cities? Further, if Toronto and Paris are taken into consideration, people are likely to be also interested in which neighborhoods are better choices.

2. Data Acquisition and Cleaning

2.1. Data sources

The sources of city neighborhoods, their respective latitude and longitude are as follows.
Toronto: http://cocl.us/Geospatial_data,
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
New York City: https://cocl.us/new_york_dataset

Paris: http://download.geonames.org/export/zip/FR.zip, processed and stored in https://docs.google.com/spreadsheets/d/e/2PACX-1vQXNcKhCQi7jy0AoRKKAafKRqZsLJhwC8jfgBrdHLOqFiPo09A8IHb6C-GpNWI21g/pub?output=xlsx

Based on the geo coordinates of neighborhoods in the three cities, I obtain the data of venues around each neighborhood via Foursquare API.

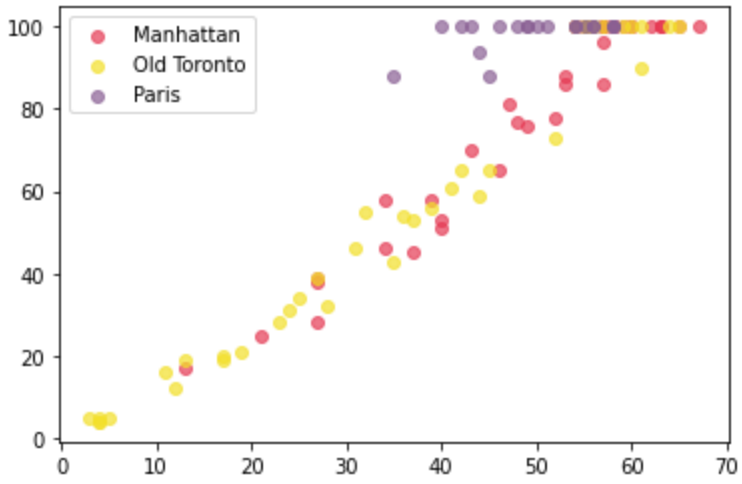
The old Toronto data (the borough name of a neighborhood contains “Toronto”) is filtered out from Toronto data; Manhattan data (the borough name of a neighborhood is “Manhattan”) out of New York City data; The Paris data is cleaned out of the geonames dataset of France.

2.2. Data cleaning

- To estimate the radius in Foursquare, apply: (City Area/ Number of Neighborhoods /)^(½) * 1609.34 meters/mile - the result would be 500 for Manhattan, 600 for Old Toronto, 800 for Paris.
- There are overlaps of radii, i.e. some venues are included in more than one neighborhood, so two types of dummy datasets are derived.
  - When analyzing based down on neighborhoods, directly get the venue category dummies from venues data around the neighborhoods.
  - When analyzing based on city level, remove duplicates: depends on subsets ‘Venue + Venue Latitude + Venue Longitude’. (‘Venue Latitude + Venue Longitude’ insufficient. There are possibilities that different venues are at the same coordinates so add ‘Venue’; In case one venue is recorded for more than one category in the database, 'Venue Category' is added to retain the complete category info.)
- Going further, the venue category frequencies (as percentages of group sum) - venues are grouped by city, then neighborhood.

3. Methodology

Figure 1. Scatter Plot - Venue Quantities against Venue Categories



3.1. Kolmogorov-Smirnov Test

Import the Kolmogorov-Smirnov test to compare 1-D arrays of the *venue category frequencies in the three cities (venues grouped by Toronto, Manhattan, and Paris*).

The Kolmogorov-Smirnov test (KS-test) tries to determine if two datasets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data. (Technically speaking it is non-parametric and distribution free.)

The KS-test result between venues frequencies in Manhattan and that in Old Toronto is 0.18045112781954886, with p-value of 1.1211445674119701e-07 indicating the lack of fit is significant between in statistical sense (the two datasets come from different distribution) the two places in terms of venues located in their neighborhoods;

In the meantime, the KS-test result between venues frequencies in Manhattan and that in Paris is 0.1748768472906404, with p-value of 1.1211445674119701e-07 indicating the datasets of the two places' venues are lacking statistical fit as well.

```
ks_2samp(freq_manhattan, freq_old_toronto)
```

```
Ks_2sampResult(statistic=0.18045112781954886, pvalue=4.3114250019683e-06)
```

```
ks_2samp(freq_manhattan, freq_paris)
```

```
Ks_2sampResult(statistic=0.2832080200501253, pvalue=1.6929118257415623e-14)
```

Figure 2. KS-test: Statistical Fit between Manhattan and Old Toronto/Paris

3.2. K-Means Clustering

As the neighborhoods are unlabeled data, to deal with the unlabelled data without features vector, I need to allow the model to work on its own to discover information and unknown patterns - it is why unsupervised learning is used in which algorithms are used against data that is not labeled, instead of supervised learning like classification.

K-Means algorithm is one of the most common cluster methods of unsupervised learning - aggregate the venues by neighborhood in the 3 cities, K-Means cluster the neighborhoods into 3 clusters. See how the neighborhoods are grouped in 3 clusters merely according to the *venue category frequencies (grouped by neighborhood)*, regardless of the geography i.e. the city where they are actually located.

While running K-Means to cluster the boroughs into 3 clusters, I analyze the K-Means with the elbow method to ensure that the 3-degree for optimum k of the K-Means. Both Distortion and Inertia are reviewed. Shown in the line plots below - the elbow is less obvious at k = 3 or 5.

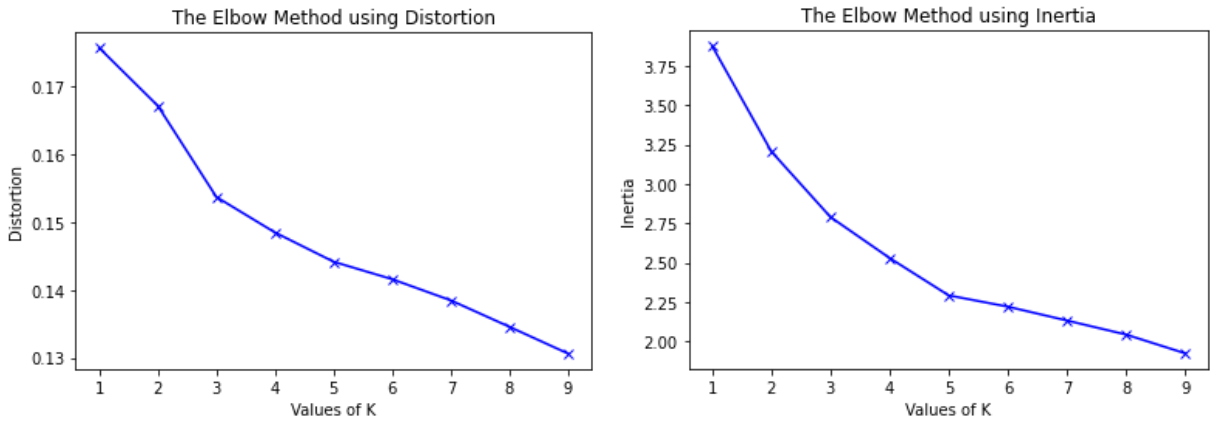


Figure 3. Line Plot - Elbow Method

The Silhouette Coefficient is bounded between -1 for incorrect clustering and +1 for highly dense clustering, according to scikit learn documentation:
https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

For clustering (k=3) in this case, the Silhouette score of 0.1559 does not suggest dense clustering.

```
from sklearn import metrics
metrics.silhouette_score(venues_grouped_clustering, kmeans.labels_, metric='euclidean')
```

```
0.1559057370552637
```

3.3. Content-Based Filtering (Recommendation System)

To complement K-Means Clustering as well as to explore further after clustering all the neighborhoods, we may consider which neighborhoods in the core Toronto area or Paris have the most similar setting or vibe as in Manhattan. We can refer to Content-Based Filtering in Recommendation Systems, and learn the venued-based “score” of each neighborhood in Toronto and Paris.

To calculate the scores, I turn the venue category frequencies of Manhattan into weights. We can do this by using the category frequencies in Manhattan and multiplying them into the venues dummies table and then summing up the resulting table by column. This operation is actually a dot product between a matrix and a vector, so simply accomplished by calling Pandas's "dot" function.

When applying this calculation on the neighborhoods in Manhattan, the mean score would be 1.024975. The more closely the score of the Toronto/Paris neighborhood approximates this figure, i.e. the lower the Absolute Difference, the higher the similarity.

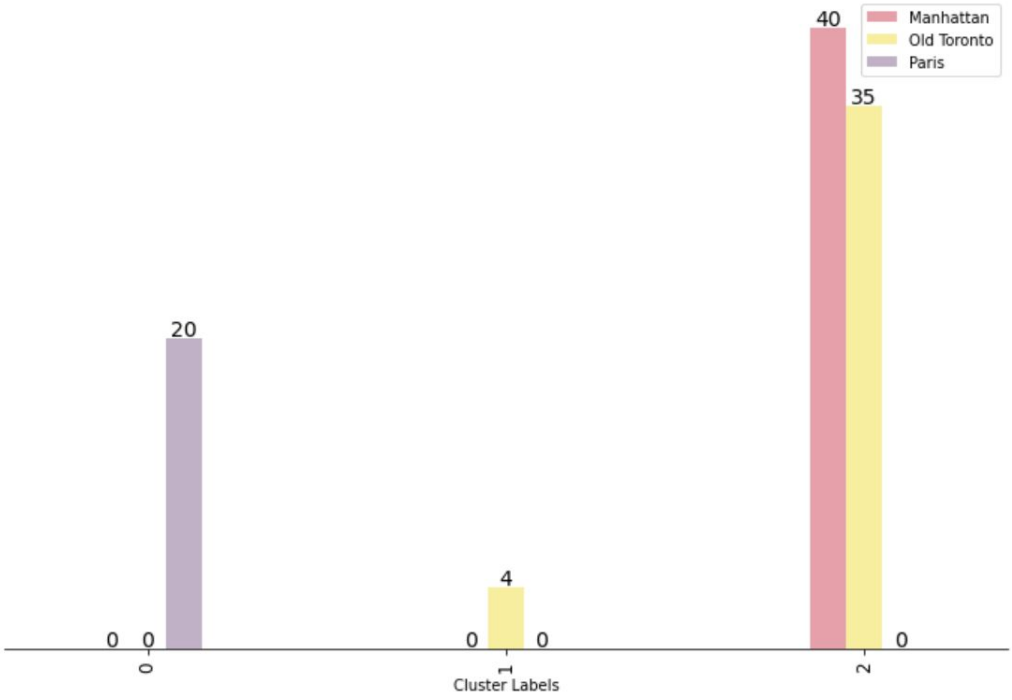
Since there is no knowledge of ground truth classes, no classification model should be built.

4. Results

K-Means Clustering

Among the 3 clusters by K-means clustering, however, 19 out of the 20 (95%) Paris neighborhoods are in Cluster 0, while 35 out of the 39 (90%) old Toronto neighborhoods fall together with all the 40 (100%) Manhattan neighborhoods in Cluster 2. The remaining 4 Toronto neighborhoods are in Cluster 1.

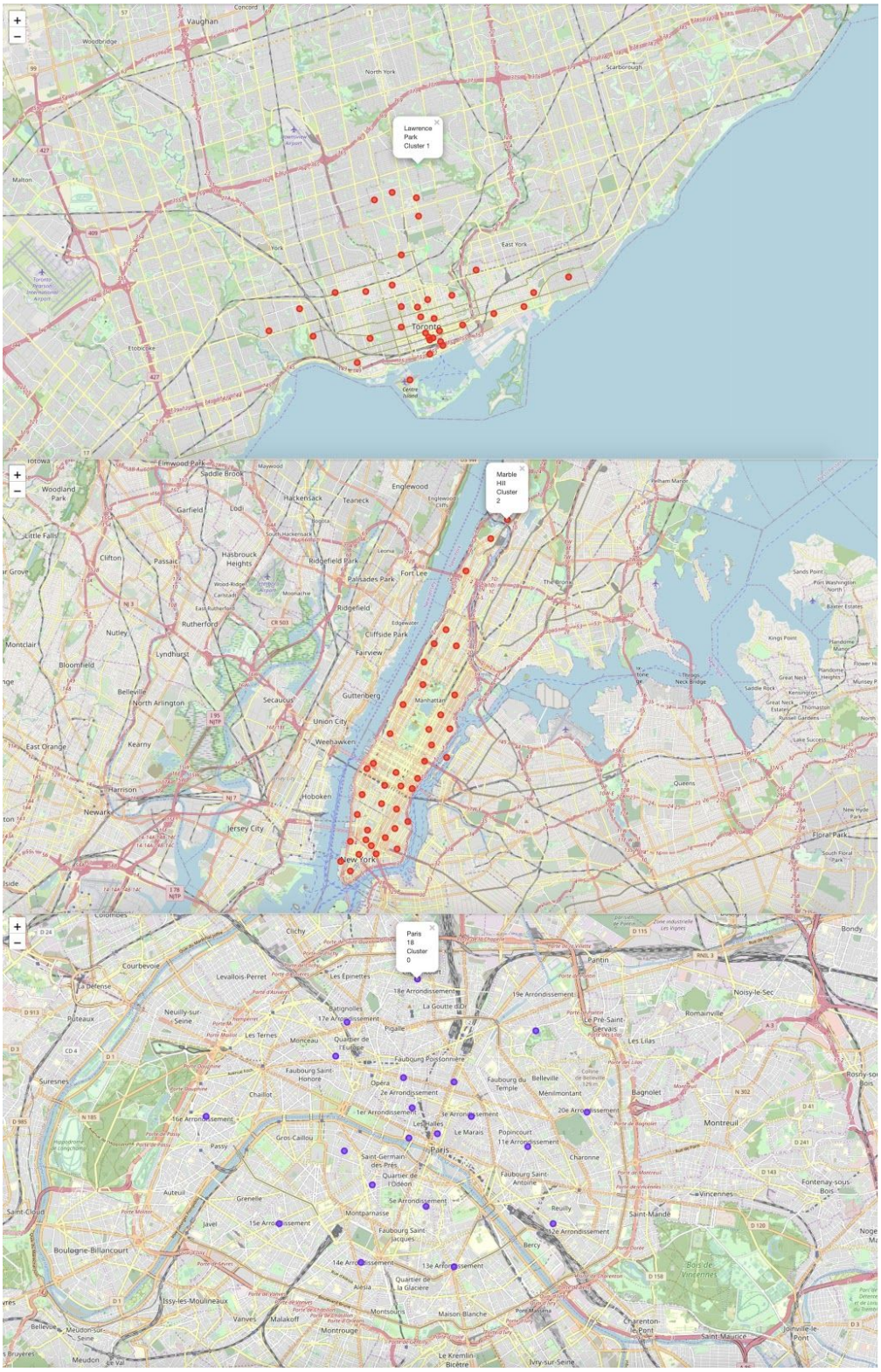
Figure 4. Bar Plot - Number of Neighborhoods in 3 Cities by Cluster



Cluster Labels	0	1	2
1st Most Common Venue	Park	Coffee Shop	French Restaurant
2nd Most Common Venue	Trail	Café	Hotel
3rd Most Common Venue	Playground	Italian Restaurant	Bar
4th Most Common Venue	Swim School	Park	Italian Restaurant
5th Most Common Venue	Bus Line	Pizza Place	Bakery
6th Most Common Venue	Jewelry Store	Bakery	Plaza
7th Most Common Venue	Sushi Restaurant	Bar	Wine Bar
8th Most Common Venue	Gym	Gym	Japanese Restaurant
9th Most Common Venue	Eastern European Restaurant	Sandwich Place	Café
10th Most Common Venue	Dog Run	American Restaurant	Bistro

The clustering shows Paris can be clearly separated, while all neighborhoods of Manhattan are also clustered together but with the majority of old Toronto neighborhoods.

Figure 5. Folium Map - Manhattan, Old Toronto, Paris



Content-Based Filtering (Recommendation System)

Sorted by the Absolute Difference with the Manhattan score, the 10 most similar neighborhoods in old Toronto and Paris to Manhattan neighborhoods are listed, also accompanied by the cluster number as well as the most popular venue categories in each neighborhood.

	City	Neighborhood	AD_with_Manhattan	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Paris	Paris 07	0.000834	0	French Restaurant	Hotel	Plaza	Garden	Café	History Museum	Historic Site	Restaurant	Italian Restaurant	Tailor Shop
2	Old Toronto	Queen's Park, Ontario Provincial Government	0.002231	2	Coffee Shop	Burger Joint	Sandwich Place	Italian Restaurant	Burrito Place	Gastropub	Falafel Restaurant	Café	Diner	Park
3	Paris	Paris 06	0.008702	0	French Restaurant	Hotel	Italian Restaurant	Plaza	Wine Bar	Ice Cream Shop	Garden	Bookstore	Chocolate Shop	Pastry Shop
4	Paris	Paris 12	0.018239	0	French Restaurant	Hotel	Bistro	Supermarket	Bar	Bakery	Italian Restaurant	Plaza	Chinese Restaurant	Japanese Restaurant
5	Paris	Paris 19	0.037313	0	French Restaurant	Bar	Café	Italian Restaurant	Pool	Restaurant	Diner	Japanese Restaurant	Supermarket	Park
6	Old Toronto	Kensington Market, Chinatown, Grange Park	0.038334	2	Bar	Café	Vegetarian / Vegan Restaurant	Dessert Shop	Chinese Restaurant	Mexican Restaurant	Coffee Shop	Yoga Studio	Record Shop	Pizza Place
7	Paris	Paris 01	0.045453	0	French Restaurant	Plaza	Hotel	Historic Site	Art Museum	Ice Cream Shop	Bakery	Restaurant	Art Gallery	Cocktail Bar
8	Paris	Paris 11	0.077163	0	French Restaurant	Bar	Italian Restaurant	Cocktail Bar	Bistro	Café	Wine Bar	Pizza Place	Restaurant	Vegetarian / Vegan Restaurant
9	Paris	Paris 04	0.078525	0	French Restaurant	Bakery	Ice Cream Shop	Art Gallery	Café	Plaza	Hotel	Gourmet Shop	Coffee Shop	Cocktail Bar
10	Paris	Paris 02	0.086018	0	French Restaurant	Wine Bar	Cocktail Bar	Hotel	Pedestrian Plaza	Plaza	Coffee Shop	Japanese Restaurant	Bakery	Cheese Shop

Figure 6. Top 10 Similar Neighborhoods to Manhattan

5. Discussion

5.1. Limitation

As the venue categories obtained from Foursquare are not exactly the same, 0 is actually filled in to complete the concat/union set. However, as the venue categorization can be different in the database through the countries, some errors or loss of similarity might be caused here, i.e. similar venues may be classified under a different “venue category” name.

The limit of “get nearby venues” queries is set at 100, seen from Figure 1. Scatter Plot - Venue Quantities against Venue Categories, most of Paris neighborhoods as well as a part of Manhattan neighborhoods have reached the upper limit, which may not adequately disclose actual venue quantities.

5.2. Scale

Based on the data of venues in the neighborhoods, at the city level, Manhattan as the whole seems similar to neither core Toronto nor Paris according to the non-parametric test - Kolmogorov-Smirnov test results. However, the Paris dataset has the higher fit.

While it is the frequencies of venue categories that are utilized in K-Means clustering where Manhattan venues are clustered with Most Toronto venues but Paris is separated, the quantities are taken into account in Content-Based Filtering, which should be the reason for more Paris neighborhoods being ranked in the front of the similarity list to Manhattan (referring to Figure 1).

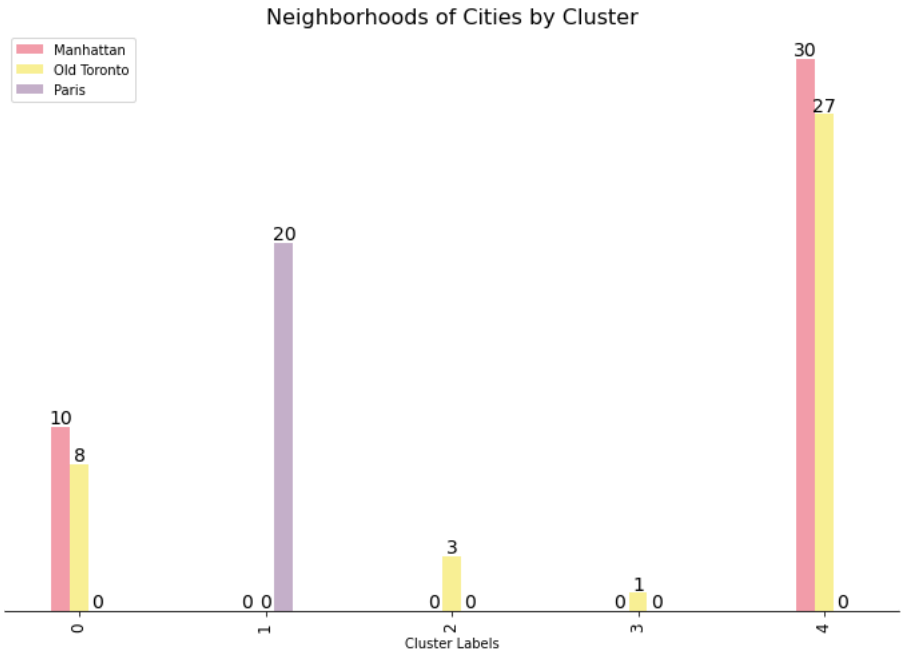
5.3. Clustering

For clustering (k=5), the Silhouette score of 0.2838 indicates denser clustering.

```
metrics.silhouette_score(venues_grouped_clustering_1, kmeans_1.labels_, metric='euclidean')
```

```
0.2837947753020613
```

Clustering (k=5, the other "elbow") further splits Cluster 1, 2 from clustering (k=3) in the section above. If discussing a certain Manhattan neighborhood rather than Manhattan as a whole, this clustering degree can be more helpful.



Cluster Labels	City	
0	Manhattan	10
	Old Toronto	9
1	Paris	20
2	Old Toronto	3
3	Old Toronto	1
4	Manhattan	30
	Old Toronto	26

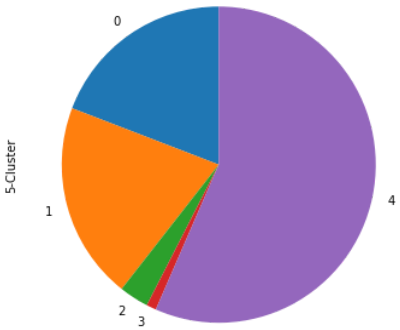


Figure 7. Pie Plot - Number of Neighborhoods by Cluster (5-Cluster)

6. Conclusion

In this study, the similarity/dissimilarity between the neighborhoods in Manhattan and in core Toronto, Paris is analyzed based on the presence of venues across hundreds of venue categories. I used KS-test to compare the distribution of neighborhoods by city; followed by K-Means Clustering all the neighborhoods aggregated together regardless of their geographic location into 3 clusters, then matching them with their respective city.

Taking Content-Based Filtering in Recommendation Systems for reference, I built a variable 'score' to compare each neighborhood to the average of Manhattan neighborhoods, sorted by the absolute difference as the similarity rank. Combined with the clustering results, even without a classification model, it can be very useful helping international brick-and-mortar businesses pick the location, from the scale of city to neighborhood. It can also be helpful for urban planning and investment selection of venue categories.