The Mover's Guide: Toronto to New York

1. Introduction

1.1 Background

In the age of globalization, people migrate between countries for purposes varying from business and education to adventure and recreation.
Such movement of people is most prominent between metropolitan cities with a large population.
This is evident in the number of expats living in major global cities worldwide.

Two of the largest cities in Northern America are New York and Toronto, both of which are the most populous of their respective countries.
New York has a population of greater than 8 million residents and Toronto has a population of close to 3 million residents.
It is natural to assume that there is plenty of movement and relocation of citizens from both cities, especially given their geographic proximity and economic influence.
It would therefore be helpful to have a guide for people planning to relocate from Toronto to New York or vice versa.

1.2 Problem

New York city is split into five boroughs, each of which houses several neighborhoods, the most populous being Brooklyn, Manhattan and Queens.
Toronto is also split into several neighborhoods.
For someone moving from one of these cities to the other, it would be useful to have an idea about which neighborhood to move into, based on their current neighborhood of residence.
This project aims to cluster the neighborhoods in the two cities in order to predict which neighborhood in the new city would be similar to the neighborhood of residence (or any other, depending on the user's choice) in the old city.

1.3 Interest

This guide is primarily targeted at anyone planning to move between the two cities and used to a certain lifestyle in either city.

It would be convenient if they could be able to reproduce a similar lifestyle in the other city - this could be quantified using several indicators such as proximity to parks, public transportation, distance from the downtown, types of restaurants and educational facilities to name a few.

2. Data

2.1 Data Source

Information about neighborhoods and their geographical coordinates are obtained using two different methods.

For the New York boroughs of Brooklyn, Manhattan and Queens, the data is available online and we use the same file as in the Week 2 assignment from the course.

For Toronto, since there is no widely available database, we make use of the wikipedia page -
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M - to extract information about the geographical coordinates of the neighborhoods.
We use the BeautifulSoup package to import the wiki page as a table into the Jupyter notebook.

The data required for comparing the neighborhoods based on its attributes is achieved using data from Foursquare about these locations.

2.2 Data Cleaning

The data extracted for Toronto does not have the same format as the data extracted the New York.
For example, there are certain neighborhoods which are not assigned a name.
For these neighborhoods, we replace the neighborhood name with the name of the borough it is part of.
Additionally, while importing the table, the line breaker '\n' appears as part of the string, which is cumbersome to work with.
Hence, we manually remove '\n's from the dataframe.

Through these methods, we are able to generate a dataframe for Toronto containing the neighborhood names and geographical coordinates, consistent with the format for New York.
This is essential, because during the clustering step, these two datasets will be combined.

2.3 Feature Selection

The primary features required for comparison include attributes of the neighborhood such as restaurants, educational and sports facilities, access to various modes of transportation, recreational areas, etc.
These attributes are imported using freely available data from Foursquare.
Importing Foursquare data is the crucial step because it lists the features as a function of the geographical location, which is key in order to compare neighborhoods.

The features are chosen in a 500 metre radius from the center of the particular neighborhood as given by its latitude and longitude.

3. Methodology

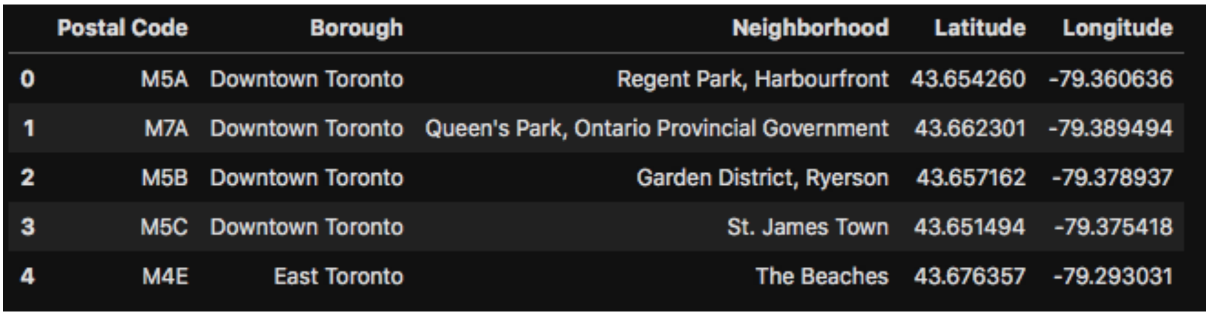
3.1 Displaying Maps

Our guide is based on the assumption the following assumptions:

1) Neighborhood coordinates extracted from the Geocoder can be used along with Foursquare API data.
2) The data from Foursquare contains relevant information about the attributes of neighborhoods described in Sec. 2.3.

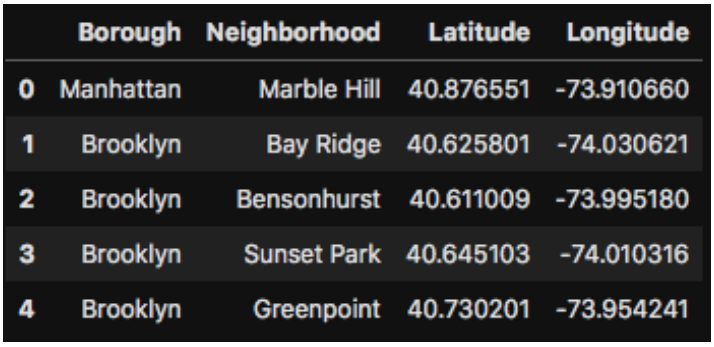
In order to verify the above two claims, we perform an exploratory analysis one a single neighbourhood and try to use its geographical coordinates to extract meaningful attributes.

After cleaning the data and obtaining the neighbourhood coordinates for Toronto and New York, the dataframes looks as follows:



	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
1	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
2	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
3	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
4	M4E	East Toronto	The Beaches	43.676357	-79.293031

Table 1: Neighbourhoods in Toronto listed along with their geographical coordinates.



	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Brooklyn	Bay Ridge	40.625801	-74.030621
2	Brooklyn	Bensonhurst	40.611009	-73.995180
3	Brooklyn	Sunset Park	40.645103	-74.010316
4	Brooklyn	Greenpoint	40.730201	-73.954241

Table 2: Neighbourhoods in the New York boroughs of Brooklyn, Manhattan and Queens listed along with their geographical coordinates.

Although only five neighbourhoods are displayed for each of these cities, the dataframes contain an extensive list of neighbourhoods in Toronto and the three most populous boroughs of New York - Brooklyn, Manhattan and Queens.

They can be displayed as geographical maps using the Folium package. The position of the neighbourhoods displayed corresponds to the latitude and longitude values from the respective dataframes of the cities.

The maps are shown in Figure 1 (Toronto) and Figure 2 (New York boroughs of Brooklyn, Manhattan and Queens). Each blue dot corresponds to the center of a neighbourhood listed in the dataframes for wither city.

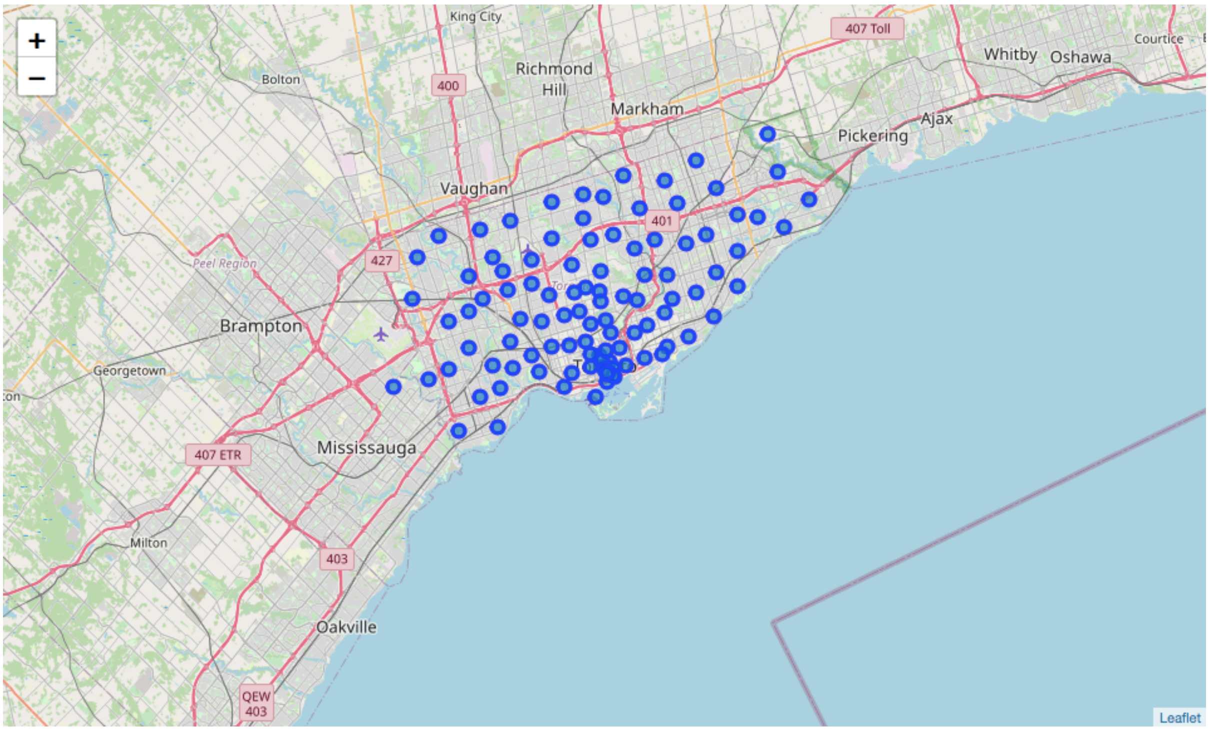


Figure 1: Geographical map displaying the position of the center of the neighborhoods in Toronto.

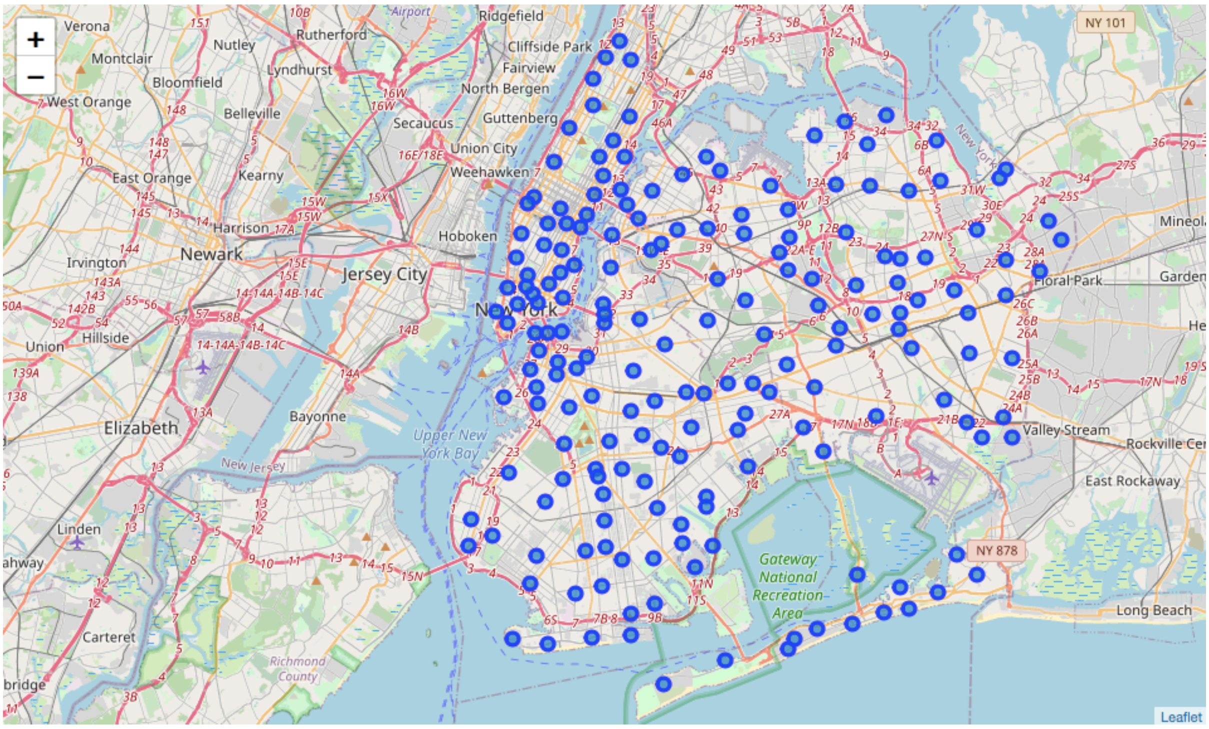
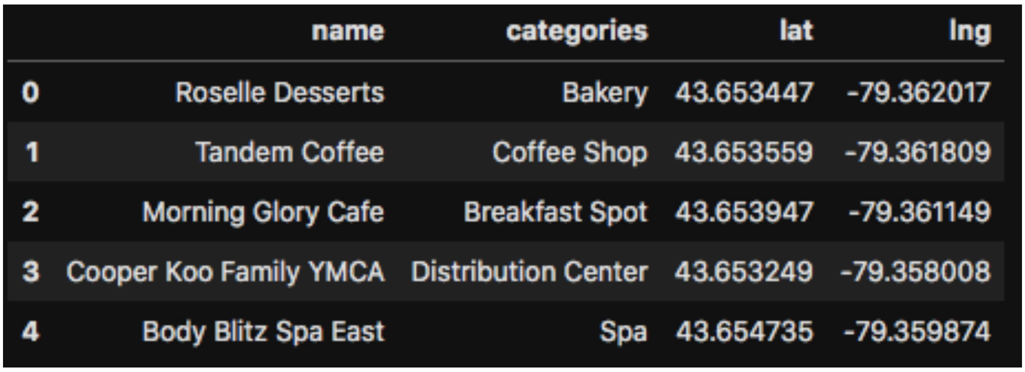


Figure 2: Geographical map displaying the position of the center of the neighborhoods in the three most populous New York boroughs – Brooklyn, Queens and Manhattan.

3.2 Exploratory Analysis

This is a great starting point to we perform an exploratory analysis one a single neighbourhood and try to use its geographical coordinates to extract meaningful attributes.
As an example, we choose St. James Town in Toronto (see Table 1), with the Foursquare API. Using the geographical coordinates of St. James Town, we explore the features and attributes available in a region of 500m from the center of the neighbourhood.

Using methods discussed previously in this course, we extract the various establishments in St. James Town and display the resulting dataframe:



	name	categories	lat	lng
0	Roselle Desserts	Bakery	43.653447	-79.362017
1	Tandem Coffee	Coffee Shop	43.653559	-79.361809
2	Morning Glory Cafe	Breakfast Spot	43.653947	-79.361149
3	Cooper Koo Family YMCA	Distribution Center	43.653249	-79.358008
4	Body Blitz Spa East	Spa	43.654735	-79.359874

Table 3: List of establishments in the Toronto neighbourhood of St. James Town. Data from the Foursquare API is used to generate this table.

The dataframe contains information about the establishments in the neighbourhood, their names, their category and their geographical position, all of which are crucial information which describe the neighbourhood.
Later, we will quantify these establishments to measure the similarity between neighborhoods.

The above analysis confirms that the neighbourhood coordinates extracted from Geocoder can be used along with the Foursquare API and that the resulting data from contains detailed information about the attributes of the neighbourhood, including name, type and geographical position.

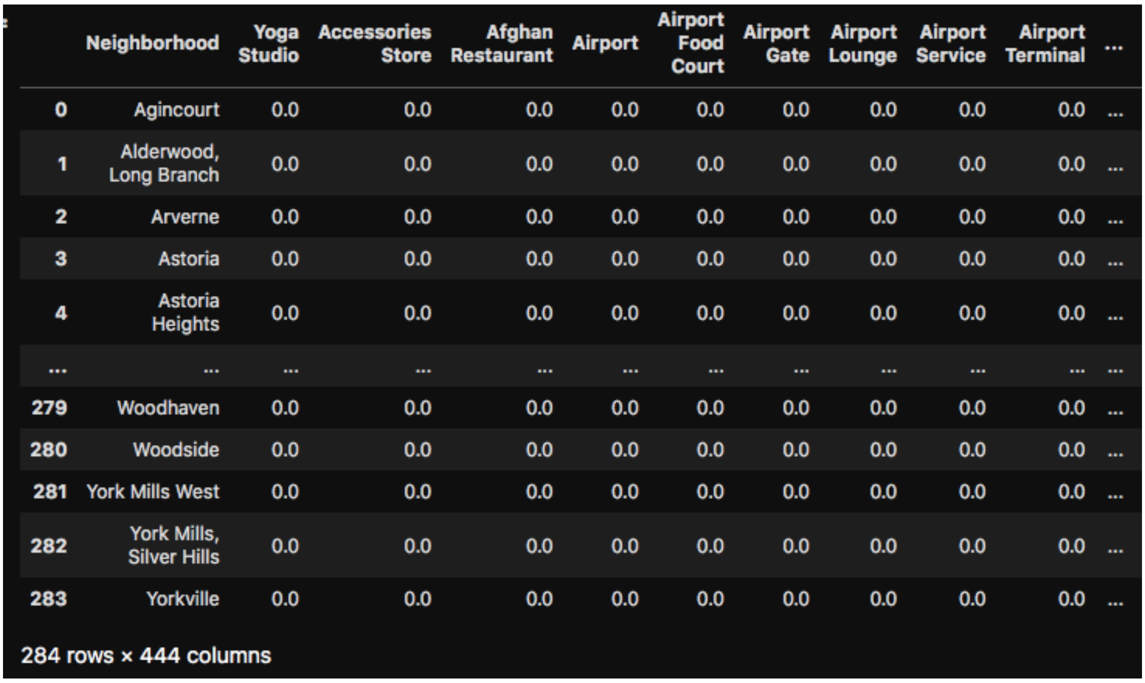
With this working method to extract attributes, it is a simple extension to extract similar data for all neighbourhoods in Toronto and in New York.

3.3 Predictive Modeling – Clustering

The objective of this guide is to cluster neighborhoods in Toronto and New York in order to assist and inform people move between the two cities.
To achieve this goal, we make use of the k-means clustering machine learning algorithm. It quantifies the selected features and attributes of a neighbourhood and uses this information to cluster them into groups that are similar, regardless of whether the neighbourhood is in Toronto or in New York.

The methodology used to prepare the data for clustering is described in this section:

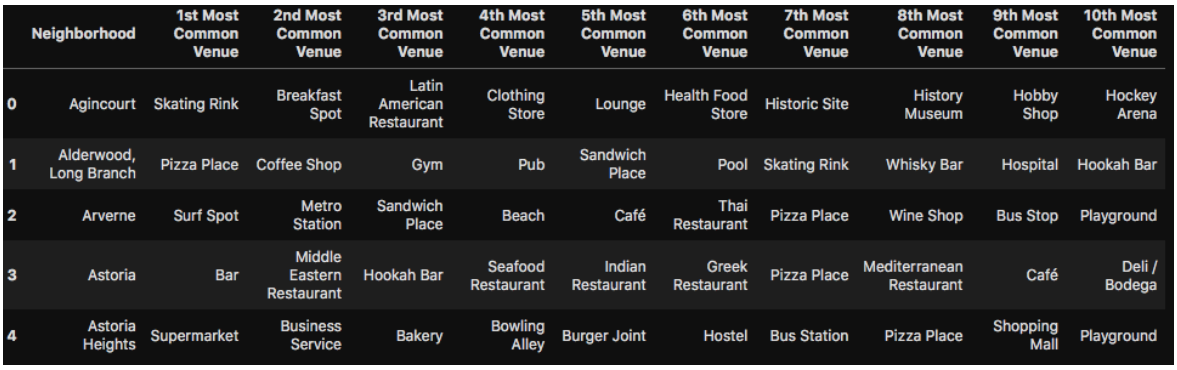
1) Data from Foursquare API for each neighbourhood in either city is collected in the form of dataframes similar to the one displayed in Table 3 for St. James Town in Toronto.
2) The lists of the neighbourhoods and their features of Toronto and New York are combined into one dataframe.
3) Each desired attribute (such as a specific type of restaurant, recreational facilities, airports, etc) in a 500m radius around the centre of the neighbourhood is quantified numerically with 0s (feature does not exist) and 1s (feature exists). The corresponding table with one-hot encoding is shown in Table 4 for the neighbourhoods in New York and Toronto. It lists 443 different features for each of the 285 neighbourhoods considered.
4) The occurrences of features are summed up and the most common features of a neighbourhood are listed for each neighbourhood, as shown in Table 5.
5) Clustering based on the k-means algorithm is performed. It divides the neighbourhoods into distinct groups based on their attributes. In order to provide a comprehensive guide to the user, it is important to choose a reasonably high number of *k* so that smaller differences between neighbourhoods can be distinguished, thereby giving the user a more accurate prediction. However, it should be too large either, because that would severely restrict the choices available. For the current notebook, the value of *k* is kept at 50 clusters, but it is important to note that the user is free to change this value depending on their requirements.



	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	...
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
2	Arverne	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
3	Astoria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
4	Astoria Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...
279	Woodhaven	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
280	Woodside	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
281	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
282	York Mills, Silver Hills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
283	Yorkville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...

284 rows × 444 columns

Table 4: One-hot encoding of features of each neighbourhood in both New York and Toronto. It lists 443 different features ranging from proximity to local bars and restaurants to train stations and airports.



	Neighborhood	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
0	Agincourt	Skating Rink	Breakfast Spot	Latin American Restaurant	Clothing Store	Lounge	Health Food Store	Historic Site	History Museum	Hobby Shop	Hockey Arena
1	Alderwood, Long Branch	Pizza Place	Coffee Shop	Gym	Pub	Sandwich Place	Pool	Skating Rink	Whisky Bar	Hospital	Hookah Bar
2	Arverne	Surf Spot	Metro Station	Sandwich Place	Beach	Café	Thai Restaurant	Pizza Place	Wine Shop	Bus Stop	Playground
3	Astoria	Bar	Middle Eastern Restaurant	Hookah Bar	Seafood Restaurant	Indian Restaurant	Greek Restaurant	Pizza Place	Mediterranean Restaurant	Café	Deli / Bodega
4	Astoria Heights	Supermarket	Business Service	Bakery	Bowling Alley	Burger Joint	Hostel	Bus Station	Pizza Place	Shopping Mall	Playground

Table 5: Occurrences of values from Table 4 are summed up to obtain the most common features of each neighbourhood.

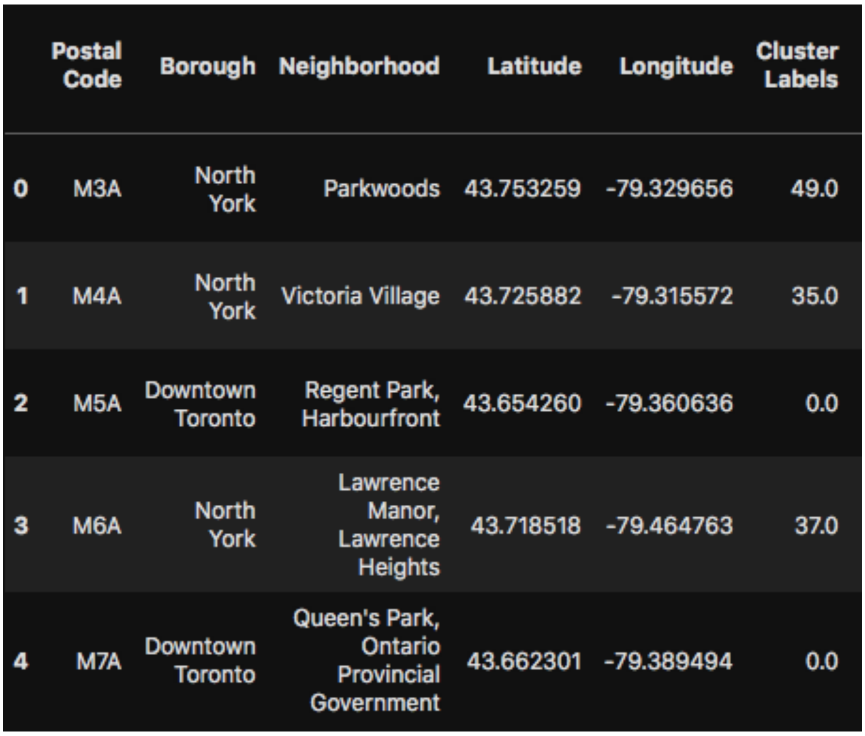
Using the methodology discussed in this section, clusters of neighbourhoods are created with similar attributes, regardless of their actual geographical coordinates, enabling the comparison of neighbourhoods in far away cities. In fact, this methodology can be readily adapted to any pair of cities, as long as the geographical coordinates of their neighbourhoods can be extracted.

4. Results

4.1 Cluster Generation

Upon applying the k-means algorithm, we obtain a dataframe listing each neighbourhood along with which cluster it has been grouped into.
For the chosen value of *k*=50, fiftly clusters of neighbourhoods are generated and the labels of these clusters range from 0 to 49.

Table 6 shows the neighbourhoods along with the cluster they've been grouped into.



	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	M3A	North York	Parkwoods	43.753259	-79.329656	49.0
1	M4A	North York	Victoria Village	43.725882	-79.315572	35.0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	0.0
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	37.0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	0.0

Table 6: List of neighbourhoods in Toronto and New York (only the first five entries shown here) along with the cluster they belong to, as indicated by the cluster label.

4.2 Geographical Visualisation

This clustering is the result of the k-means algorithm based on the information from Foursquare API and since the geographical coordinates are available, one can easily use the Folium package to display the neighbourhoods on a real map and indicate similar neighbourhoods with the same colour. Such maps are shown in the next page in Figure 3 and Figure 4.

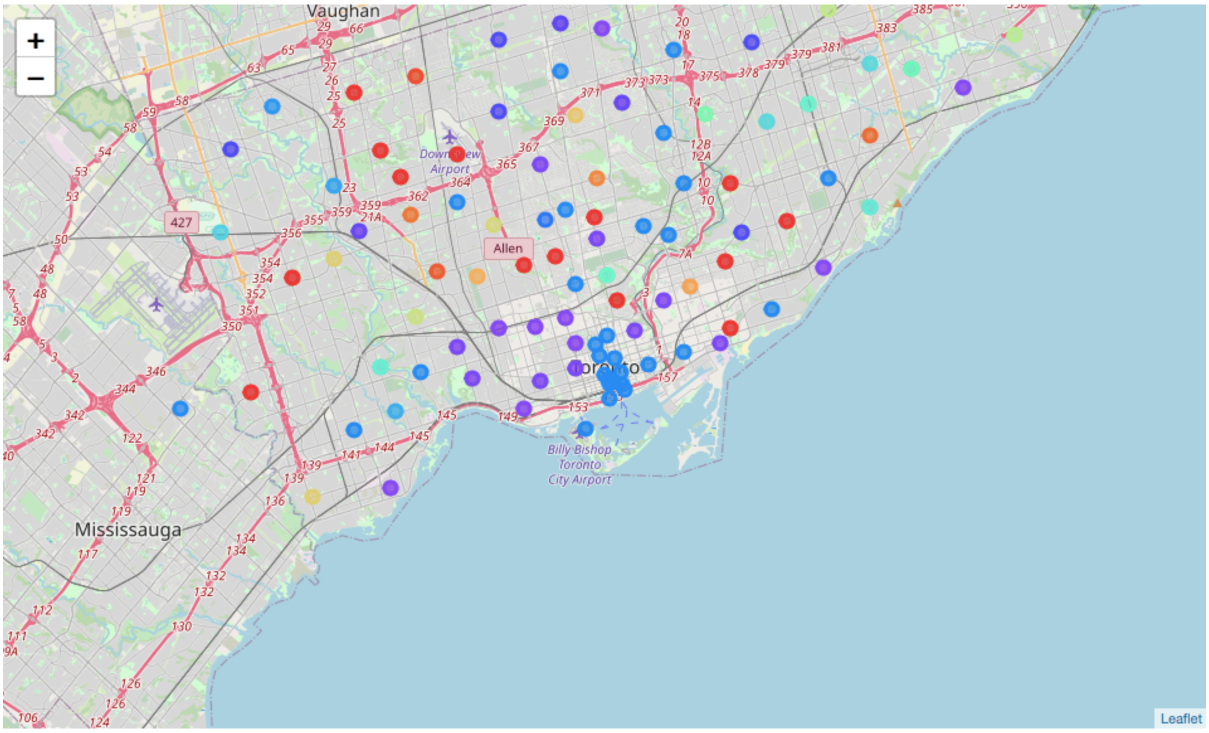


Figure 3: Map of neighbourhoods in Toronto clustered into groups with the colours indicating the cluster label.

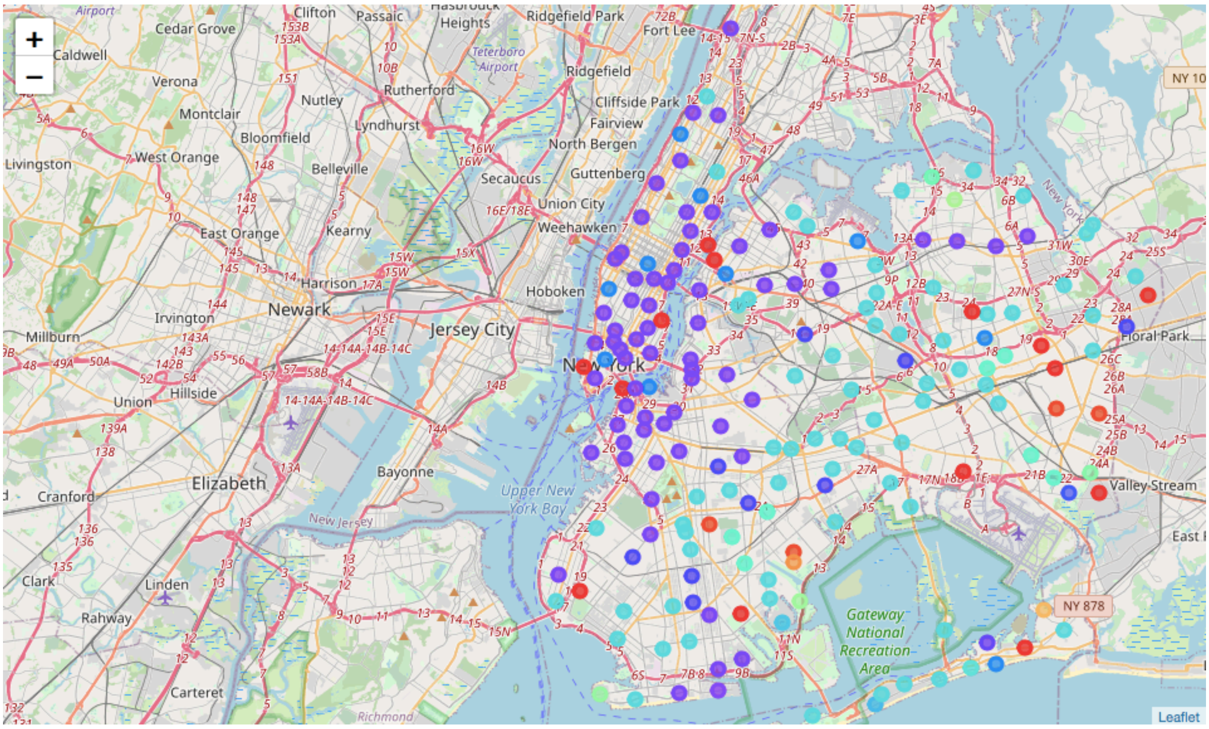


Figure 4: Map of neighbourhoods in New York clustered into groups with the colours indicating the cluster label.

Such a direct visual representation is an extremely powerful tool for any user planning to move between the two cities. The cluster labels are independent of the city the neighbourhood belongs to, meaning that a purple-coloured neighbourhood in Toronto has similar attributes as a purple-coloured neighbourhood in New York.
Thus, by simply looking at the maps, a decision can be made as to which neighbourhoods are worth looking into.

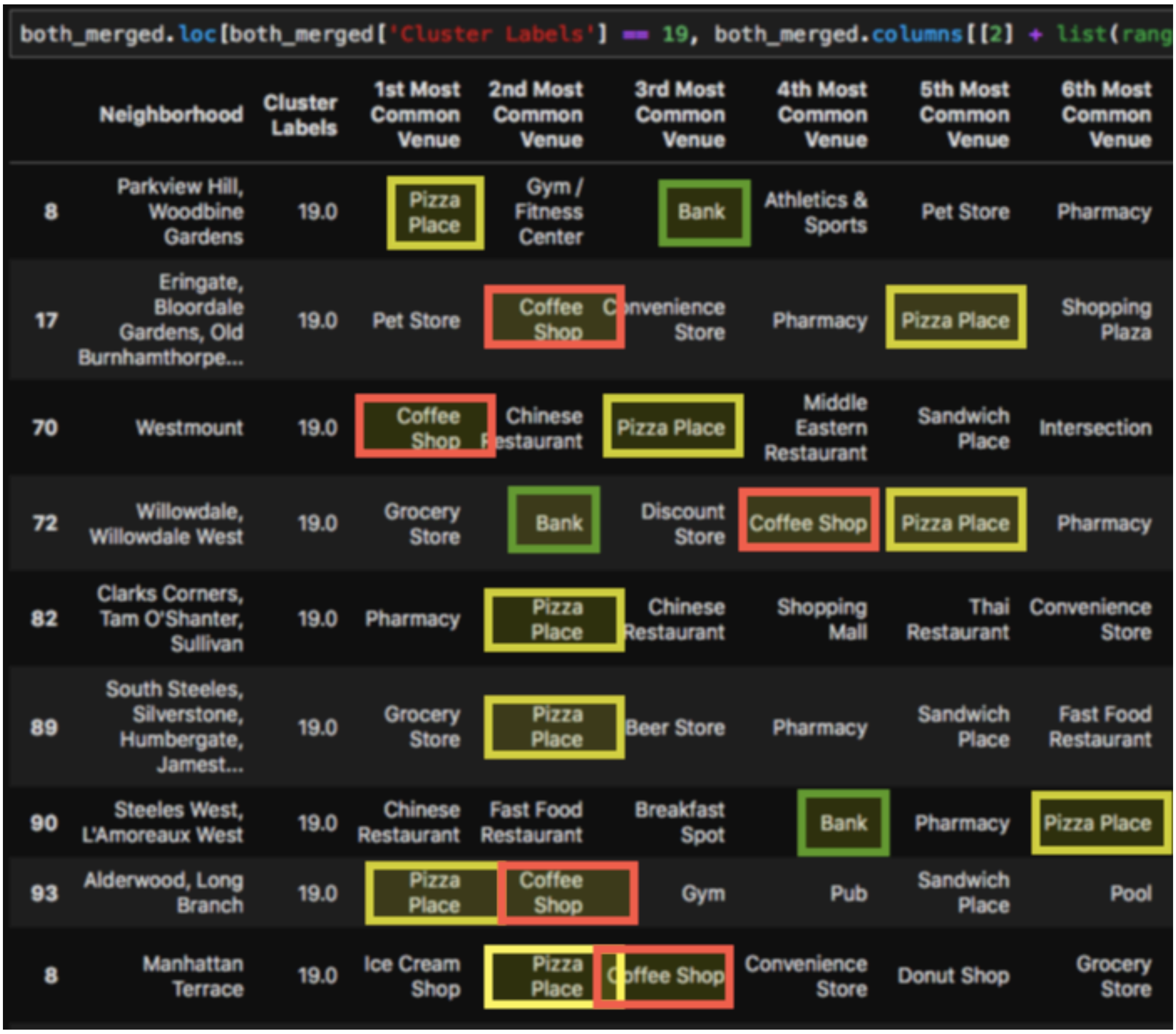
5. Discussion

The maps in Figure 3 and Figure 4 are generated without any personal biases and purely based on data from the Foursquare API database.
Certain obvious clusters can be distinguished, such as downtown areas in either city (or boroughs in the case of New York), which have similar attributes.
For example, clusters corresponding to dark-blue and purple-coloured neighbourhoods seem to be located closer to the downtown part of the cities.
Similarly, red and orange coloured neighbourhoods seem to be located towards the outskirts.

However, several clusters only be understood with a closer to look at their attributes. As an example of recommendations which could be made, here are some examples of similar neighbourhoods:

1) Table 7 shows a list of neighbourhoods which seem to be located in a densely populated downtown region, with several pizza places and coffee shops. They also seem to be neighbourhoods with proximity to banks, which would be an interesting parameter in case the user works at a bank.
   A bar plot listing the most common attributes shows gives an indication of the make-up of the cluster (Figure 5).
2) Tables 8 and 9 show clusters based on a common liking of Caribbean or Indian food. In case the user is a fan of either, this would be an option to consider.
   A bar plot for the cluster relating to Table 8 indicates the make-up of that cluster (Figure 6).
3) Clusters can be based on more than just taste in food, such as proximity to a beach or beach resorts, as in Table 10. It could also indicate whether a neighbourhood is close to natural trails, as in Table 11.

A user would be able to extract several such correlations by looking at clusters for different values of *k* and come up with a cluster which is ideally suited to their requirements.



```
both_merged.loc[both_merged['Cluster Labels'] == 19, both_merged.columns[[2] + list(rang
```

	Neighborhood	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
8	Parkview Hill, Woodbine Gardens	19.0	Pizza Place	Gym / Fitness Center	Bank	Athletics & Sports	Pet Store	Pharmacy
17	Eringate, Bloordale Gardens, Old Burnhamthorpe...	19.0	Pet Store	Coffee Shop	Convenience Store	Pharmacy	Pizza Place	Shopping Plaza
70	Westmount	19.0	Coffee Shop	Chinese Restaurant	Pizza Place	Middle Eastern Restaurant	Sandwich Place	Intersection
72	Willowdale, Willowdale West	19.0	Grocery Store	Bank	Discount Store	Coffee Shop	Pizza Place	Pharmacy
82	Clarks Corners, Tam O'Shanter, Sullivan	19.0	Pharmacy	Pizza Place	Chinese Restaurant	Shopping Mall	Thai Restaurant	Convenience Store
89	South Steeles, Silverstone, Humbergate, Jamest...	19.0	Grocery Store	Pizza Place	Beer Store	Pharmacy	Sandwich Place	Fast Food Restaurant
90	Steeles West, L'Amoreaux West	19.0	Chinese Restaurant	Fast Food Restaurant	Breakfast Spot	Bank	Pharmacy	Pizza Place
93	Alderwood, Long Branch	19.0	Pizza Place	Coffee Shop	Gym	Pub	Sandwich Place	Pool
8	Manhattan Terrace	19.0	Ice Cream Shop	Pizza Place	Coffee Shop	Convenience Store	Donut Shop	Grocery Store

Table 7: A cluster corresponding to neighbourhoods which seem to have a high frequency of pizza places and coffee shops.

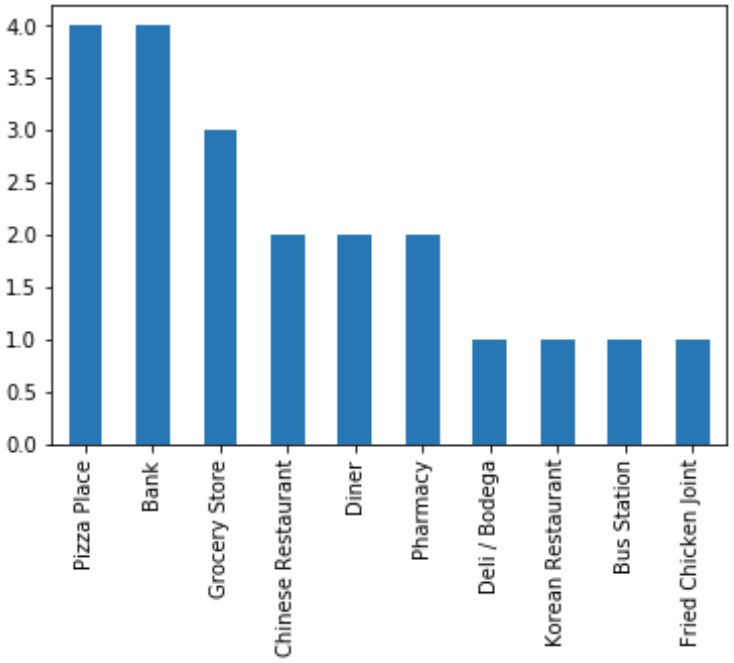
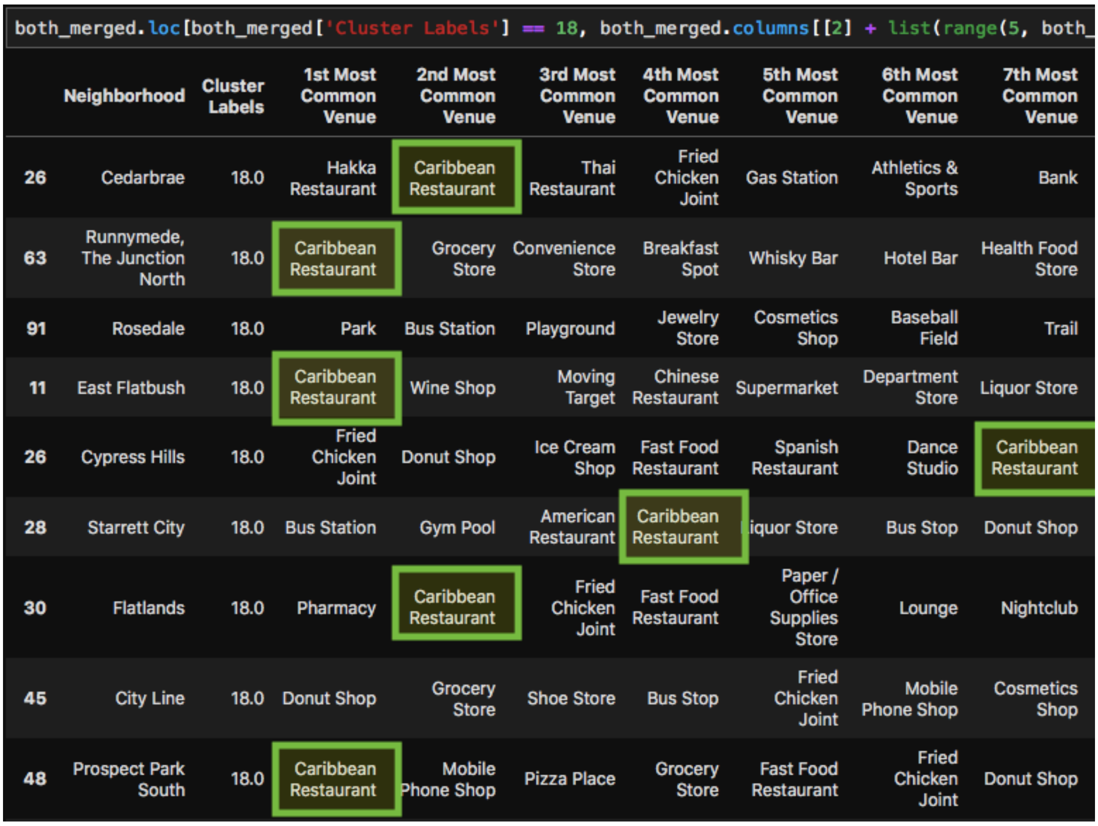


Figure 5: Bar plot showing the most common attributes of the neighbourhoods of the cluster corresponding to Table 7.



```
both_merged.loc[both_merged['Cluster Labels'] == 18, both_merged.columns[[2] + list(range(5, both_
```

	Neighborhood	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
26	Cedarbrae	18.0	Hakka Restaurant	Caribbean Restaurant	Thai Restaurant	Fried Chicken Joint	Gas Station	Athletics & Sports	Bank
63	Runnymede, The Junction North	18.0	Caribbean Restaurant	Grocery Store	Convenience Store	Breakfast Spot	Whisky Bar	Hotel Bar	Health Food Store
91	Rosedale	18.0	Park	Bus Station	Playground	Jewelry Store	Cosmetics Shop	Baseball Field	Trail
11	East Flatbush	18.0	Caribbean Restaurant	Wine Shop	Moving Target	Chinese Restaurant	Supermarket	Department Store	Liquor Store
26	Cypress Hills	18.0	Fried Chicken Joint	Donut Shop	Ice Cream Shop	Fast Food Restaurant	Spanish Restaurant	Dance Studio	Caribbean Restaurant
28	Starrett City	18.0	Bus Station	Gym Pool	American Restaurant	Caribbean Restaurant	iquor Store	Bus Stop	Donut Shop
30	Flatlands	18.0	Pharmacy	Caribbean Restaurant	Fried Chicken Joint	Fast Food Restaurant	Paper / Office Supplies Store	Lounge	Nightclub
45	City Line	18.0	Donut Shop	Grocery Store	Shoe Store	Bus Stop	Fried Chicken Joint	Mobile Phone Shop	Cosmetics Shop
48	Prospect Park South	18.0	Caribbean Restaurant	Mobile Phone Shop	Pizza Place	Grocery Store	Fast Food Restaurant	Fried Chicken Joint	Donut Shop

Table 8: A cluster corresponding to neighbourhoods which seem to have a high frequency of Caribbean restaurants.

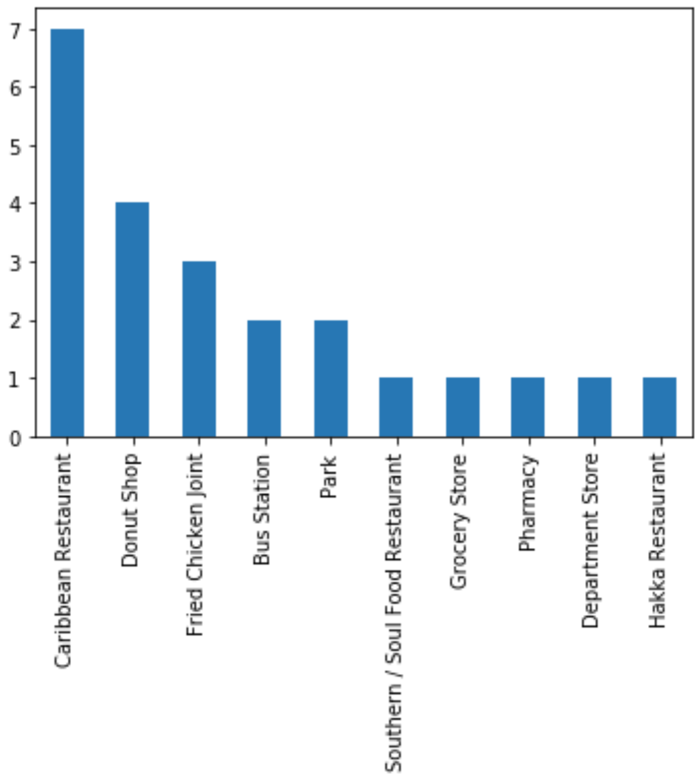
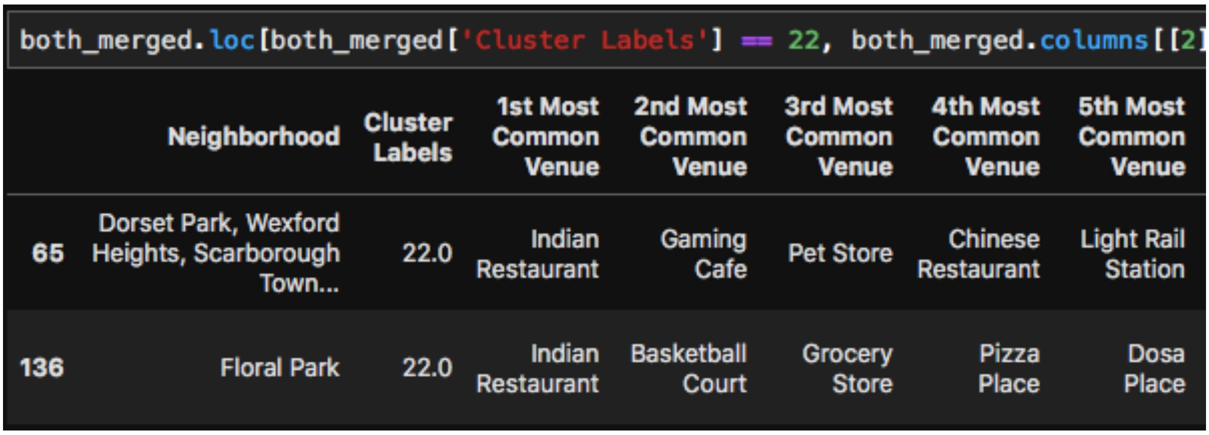


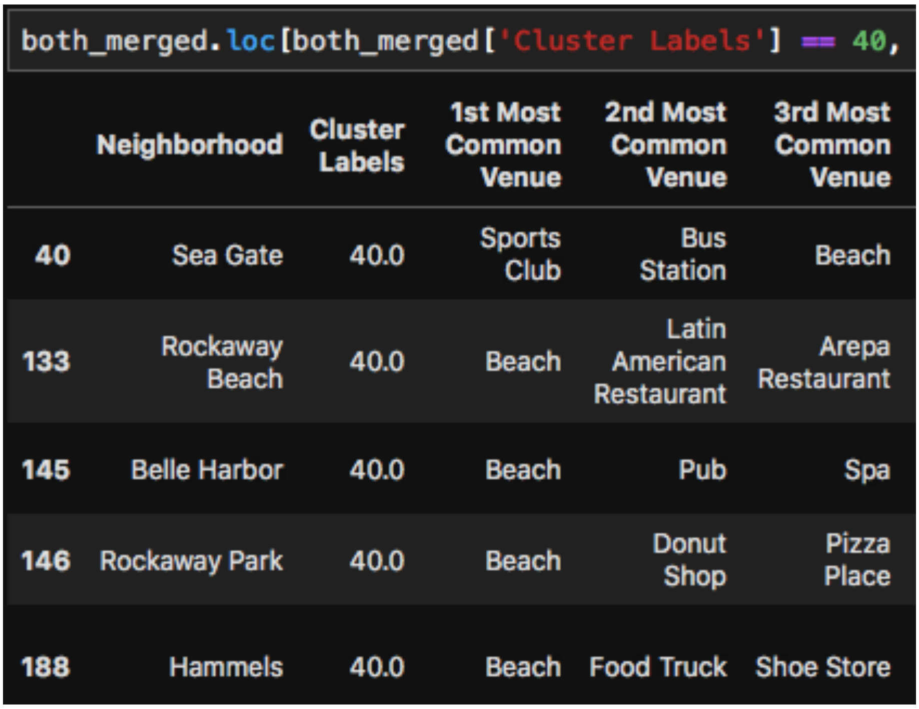
Figure 6: Bar plot showing the most common attributes of the neighbourhoods of the cluster corresponding to Table 8.



```
both_merged.loc[both_merged['Cluster Labels'] == 22, both_merged.columns[[2
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
65	Dorset Park, Wexford Heights, Scarborough Town...	22.0	Indian Restaurant	Gaming Cafe	Pet Store	Chinese Restaurant	Light Rail Station
136	Floral Park	22.0	Indian Restaurant	Basketball Court	Grocery Store	Pizza Place	Dosa Place

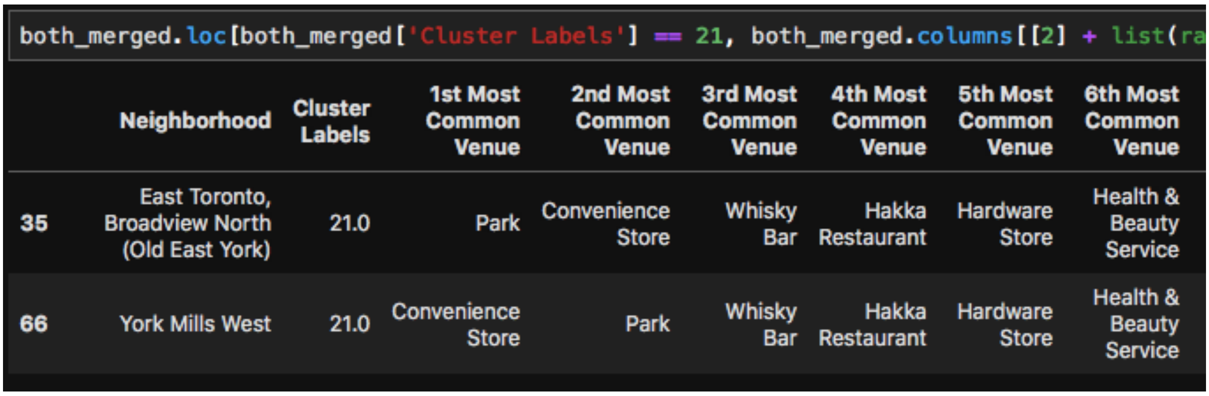
Table 9: A cluster corresponding to neighbourhoods which seem to have a high frequency of Indian restaurants.



```
both_merged.loc[both_merged['Cluster Labels'] == 40,
```

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
40	Sea Gate	40.0	Sports Club	Bus Station	Beach
133	Rockaway Beach	40.0	Beach	Latin American Restaurant	Arepa Restaurant
145	Belle Harbor	40.0	Beach	Pub	Spa
146	Rockaway Park	40.0	Beach	Donut Shop	Pizza Place
188	Hammels	40.0	Beach	Food Truck	Shoe Store

Table 10: A cluster corresponding to neighbourhoods located close to a beach.



```
both_merged.loc[both_merged['Cluster Labels'] == 21, both_merged.columns[[2] + list(ra
```

	Neighborhood	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
35	East Toronto, Broadview North (Old East York)	21.0	Park	Convenience Store	Whisky Bar	Hakka Restaurant	Hardware Store	Health & Beauty Service
66	York Mills West	21.0	Convenience Store	Park	Whisky Bar	Hakka Restaurant	Hardware Store	Health & Beauty Service

Table 11: A cluster corresponding to suburban neighbourhoods located close to parks.

6. Conclusion

In view of the requirement for people to make informed decisions about switching cities or neighbourhoods, it is useful to have a guide which makes suggestions based on preferences of the user.

To this end, this guide provides a detailed clustering mechanism along with visualization tools to help people moving between New York and Toronto or vice versa to choose which neighbourhood to move into.
The detailed analysis enables users to see precisely what sets different clusters apart and what the underlying attributes are.

It is also important to note that such a guide, although shown here for people moving between the two cities, is equally valid for people moving within either city, making this a highly versatile guide.

A relatively simple extension of this guide is to compare and cluster any arbitrary pair of cities, providing universal support for users moving out of their city. This can be done easily by having the geographical coordinates of the two cities to be compared.