Applied Data Science Capstone - Course Project

The Battle of Neighborhoods: NY Vs Toronto

Leonardo Bertolucci Coelho

1. Introduction

The background of this project if the importance of geographical data in the solution of Business problems in the open market.

In particular, stakeholders would more likely benefit from location data analysis to evaluate the risk of investing in new businesses, thus providing support to their decisions. The justification of the proposed Business Problem is here addressed in detail.

The main idea behind this project is the following: different big cities (or Capitals) around the world could be compared from diverse perspectives. For example: is there a group of Neighborhoods in NY city that could somehow be considered as similar to another group of Neighborhoods in Toronto? Further developing the idea: would there be a way of clustering different cities' Neighborhoods based on their Venues?

Considering an affirmative response to the raised up questions, it would be very interesting to compare the Venues of clustered Neighborhoods from different cities. This approach could lead to identifying business trends and market gaps in a given area of interest.

Indeed, for a group of stakeholders doubting which business to implement in Woodhaven (NY), the access to such Data-oriented information from a similar Neighborhood (of a comparable city) might undoubtedly help in the decision making process.

Similarly, if a stakeholder wants to open the first Ice Cream Shop in a Toronto Neighborhood but cannot decide (for example) between Woburn or York Mills West, the fact that a given NY Neighborhood well known for Ice Cream Shops is similar to Woburn might be tiebreaking.

Furthermore, considering a Neighborhood in Toronto as similar to one in NY, but with apparent dissimilarities in terms of frequency of African Restaurants, this observation potentially points out to a market growth direction.

Therefore, based on a geolocation approach, the purpose of this project would be to contribute to choosing locations for open businesses.

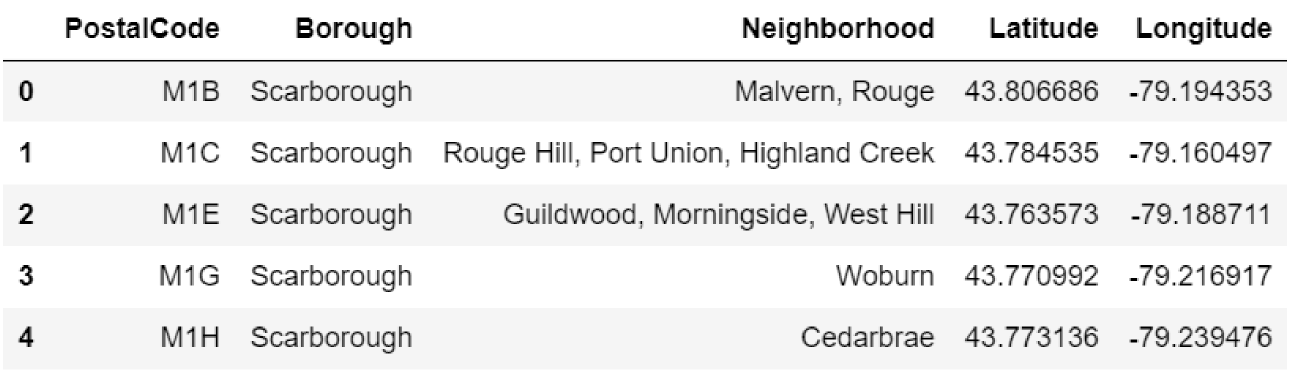
It is believed that this methodology could unravel unseen market opportunities at given locations and contribute to the growth of markets already established in specific Neighborhoods.

2. Data Description

The Toronto neighborhood Data was imported by web scraping using the BeautifulSoup package. This data table was read in a Pandas DataFrame.

Latitudes/longitudes values were updated to the toronto_data, by using the join() method. The Toronto postal-code data was imported by pd.read_csv() method using a link provided within the Applied Data Science Capstone course.

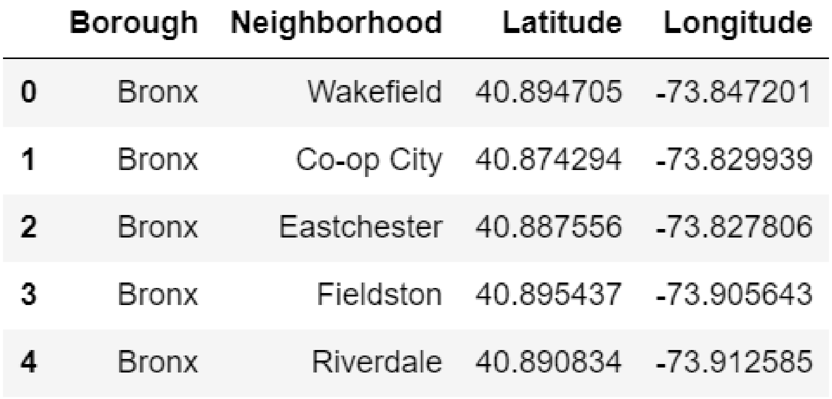
Upon data cleaning, the following DataFrame (toronto_data) was obtained:



	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

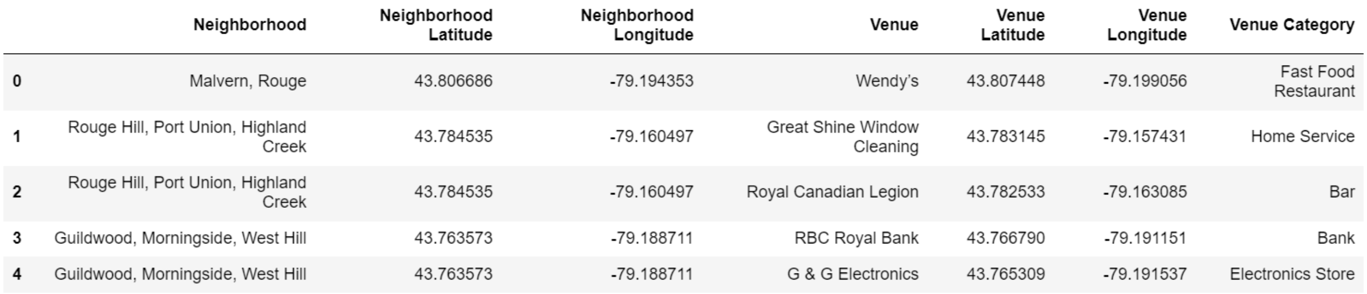
The NY city neighborhood Data was download using the wget() method. The json file was loaded to newyork_data.

A Pandas DataFrame was created, and Neighborhood/latitudes/longitudes from the newyork_data were appended. The following DataFrame (ny_data) was obtained:

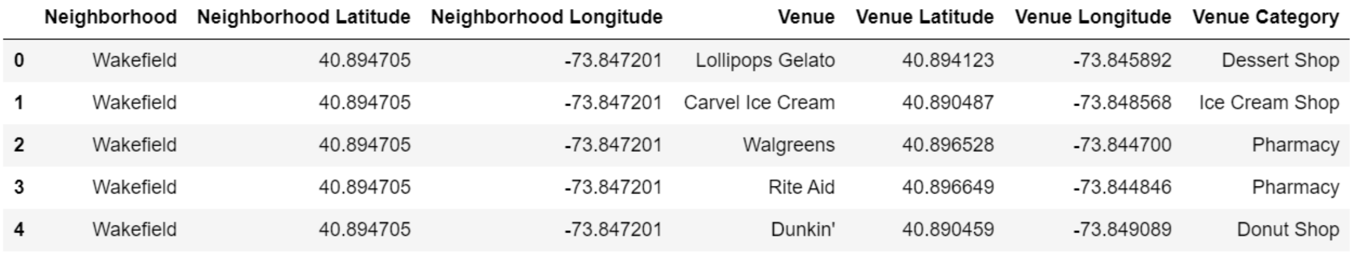


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Next, the Foursquare API was used to explore the neighborhoods and segment them, creating new data frames. The tor_venues and ny_venues data frames comprising location data of both Neighborhoods and Venues are respectively presented below:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Great Shine Window Cleaning	43.783145	-79.157431	Home Service
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

The Foursquare returned 2135 Venues and 99 Neighborhoods from the Toronto data (272 Venue categories); and 10131 Venues and 301 Neighborhoods from the NY data (439 Venue categories)

Upon completing the data collection and preparation steps, the Neighborhood Dataframes were used to explore the Toronto and NY Venues.

3. Methodology

The Foursquare API was used to obtain geolocation data of Venues from multiple Categories.

A function including an API request URL was created to get the latitudes/longitudes of all Venues returned by Foursquare (radius=500). The function mentioned above ran on both Toronto and NY city neighborhoods.

Considering that common Venue categories were determined for both NY and Toronto neighborhoods, it was decided to use this attribute to cluster together both cities data.

The clustering algorithm chosen was the K-Means, which is one of the most common methods of unsupervised learning.

For that purpose, the Kmeans tool was imported from sklearn.cluster library (random_state=0).

For applying the Kmeans, the DataFrames should contain only the numerical entries related to the Venue category (in 0 or 1 in the present case). Therefore, all other columns ("neighborhood", for exp) need to be removed.

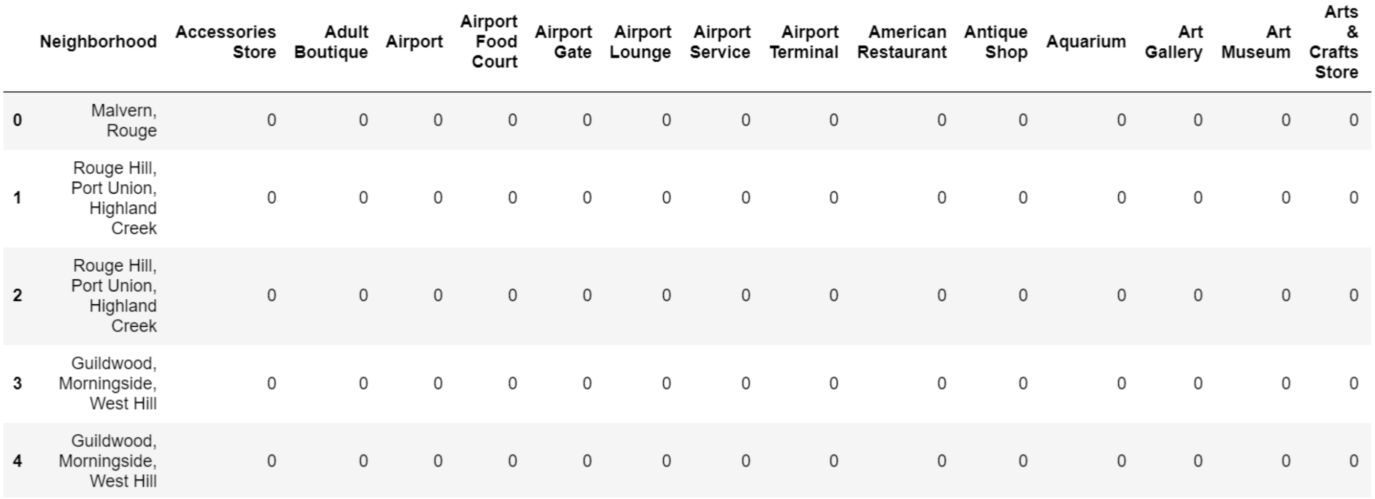
Next, The elbow method was applied to determine the optimal k value for the K-means algorithm. For that, plot libraries (matplotlib) were imported. Reference: https://predictivehacks.com/k-means-elbow-method-code-for-python/#:~:text=K%2DMeans%20is%20an%20unsupervised,optimal%20for%20the%20specific%20case.

Concerning the addition of cluster labels to the Dataframes comprising latitudes/longitudes, the following plotting libraries: matplotlib.cm as cm, matplotlib.colors as colors

The geopy library was used to get the latitudes/longitudes, and the folium library was imported for plotting of clusters superimposed to the cities' maps,

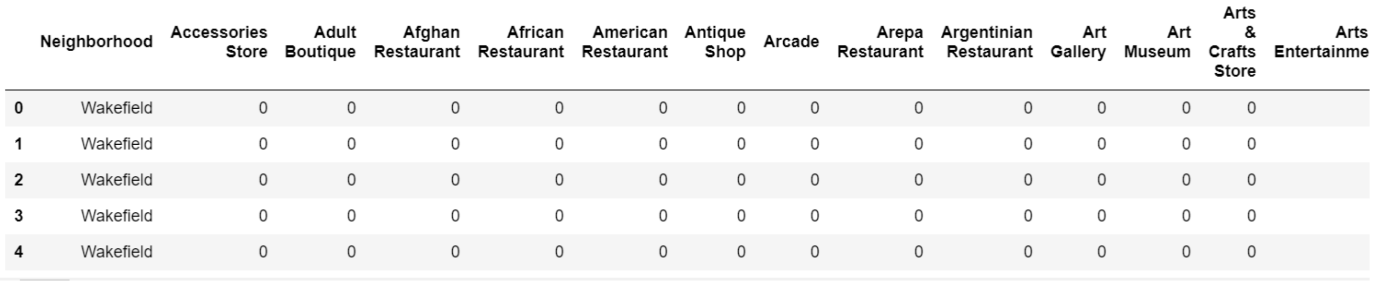
3.1. Exploratory Data Analysis

In the first step of data analysis, the Venue categories were analyzed I -detail by using the pd.get_dummies() method. New Dataframes comprising The Venue category (columns) for each Venue (rows) were obtained. For each Venue (represented by its Neighborhood), a cell value equal to 1 is attributed to the respective Venue category (otherwise, a cell value equal to 0 is given). The obtained Toronto (tor_onehot) and NY (ny_onehot) Dataframes are respectively presented below:



	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store
0	Malvern, Rouge	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Rouge Hill, Port Union, Highland Creek	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	0	0	0	0	0	0	0	0

:



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Arts Entertainme
0	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	
1	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	
2	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	
3	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	
4	Wakefield	0	0	0	0	0	0	0	0	0	0	0	0	

In the second step of the exploraroty analysis, the Venues were grouped by Neighborhood, and the mean of the frequency of occurrence of each Venue category was achieved.

The NY Dataframe still had 439 Venue categories distributed across 301 Neighborhoods, while the Toronto Dataframe still had 272 Venue categories for 99 Neighborhoods.

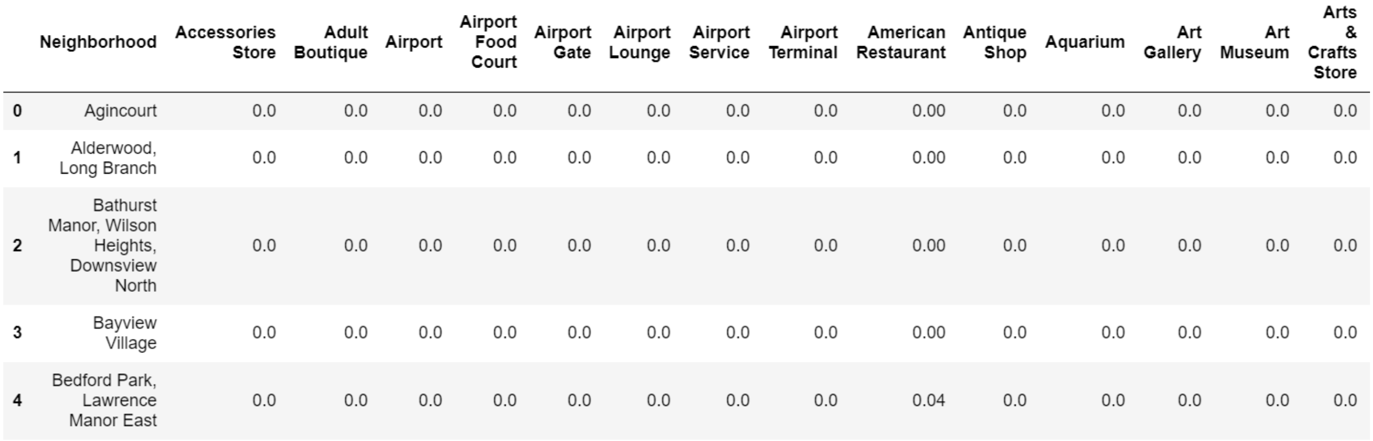
Next, both Toronto and NY DataFrames displaying the frequency of occurrence of Venue categories were merged.

The idea is that from the Merged Toronto-NY DataFrame, the Toronto/NY Neighborhoods will be clustered based on the Venue Category of all Venues returned by Foursquare. These clustering outcomes will be addressed in the Results section.

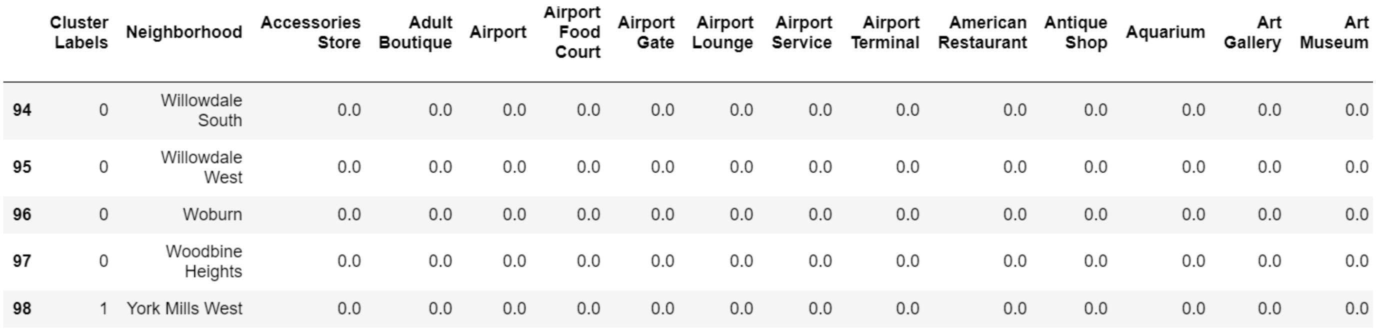
The merged Dataframe had 471 Venue categories in total for both 400 Toronto/NY Neighborhoods. The resulting NaN values were substituted by 0 for clustering purposes(NaN values corresponded to Venue categories belonging to the NY or the Toronto Dataframes only).

The exactness of the Toronto and NY subsets from the Merged Toronto-NY DataFrame (torny_grouped) was checked. The sub-Dataframes below display the frequency of occurrence of Venue categories per Neighborhood for:

Toronto (head() and tail()):

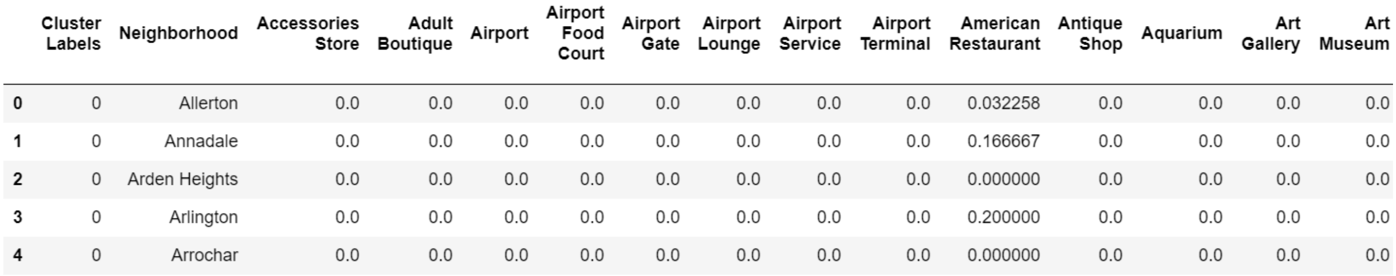


	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.0	0.0	0.0

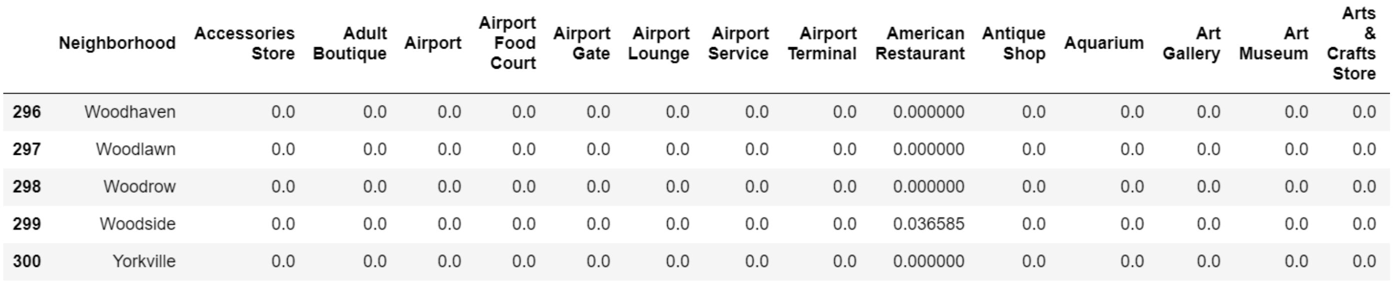


	Cluster Labels	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
94	0	Willowdale South	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
95	0	Willowdale West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
96	0	Woburn	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
97	0	Woodbine Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
98	1	York Mills West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

New York (head() and tail()):



	Cluster Labels	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
0	0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.032258	0.0	0.0	0.0	0.0
1	0	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.166667	0.0	0.0	0.0	0.0
2	0	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	0	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.200000	0.0	0.0	0.0	0.0
4	0	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0



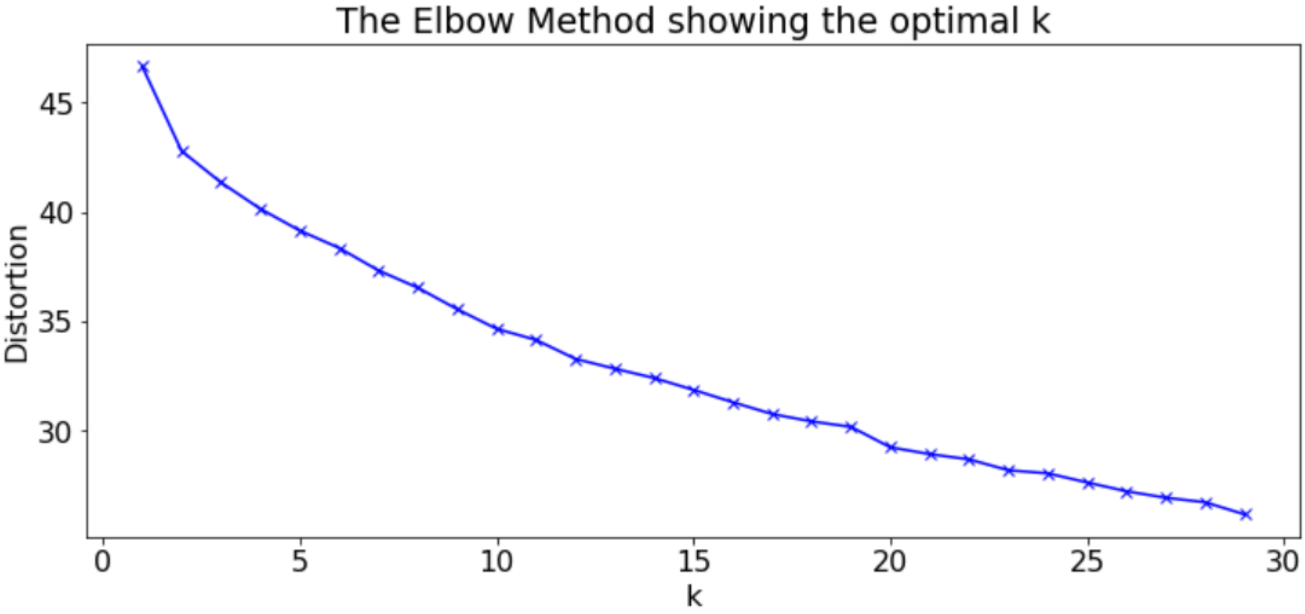
	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store
296	Woodhaven	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
297	Woodlawn	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
298	Woodrow	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0
299	Woodside	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.036585	0.0	0.0	0.0	0.0	0.0
300	Yorkville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0

4. Results and discussion

4.1. Clustering

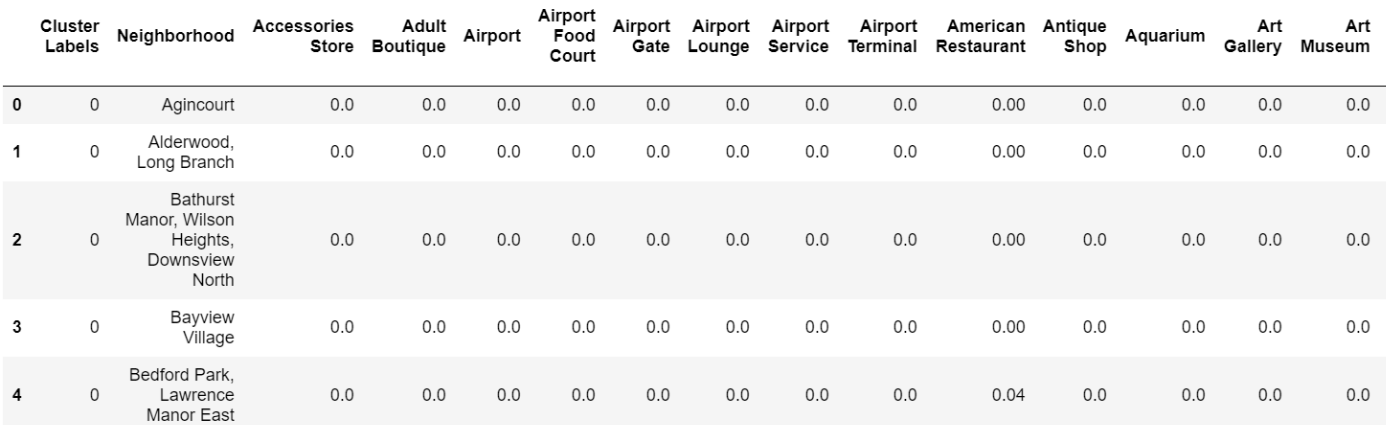
The KMeans algorithm was applied to Cluster Neighborhoods from the Merged Toronto-NY geolocation data.

The K-Means Elbow Method was applied for defining of the optimal number of clusters. The resulting Elbow plot was obtained (kclusters=range(1,30)):



Upon analysis of the K-means elbow outcomes, it was determined that the optimal number of clusters was equal to 2 (kclusters = 2)

Then, the K-means algorithm was able to cluster the Merged Toronto-NY data into 2 cluster of neighborhoods. Next, the cluster labels (0 or 1) were inserted into the Merged DataFrame, as shown hereafter:



	Cluster Labels	Neighborhood	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum
0	0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
1	0	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
2	0	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
3	0	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.0	0.0	0.0
4	0	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.0	0.0	0.0	0.0

Upon analysis, 297 and 4 Neighborhoods were respectively found for NY clusters 0 and 1; while 84 and 15 Neighborhoods were respectively found for Toronto clusters 0 and 1.

In terms of Venue categories, Cluster 0 seemed to be quite diverse for both cities, comprising over 1d hundred neighborhoods.

On the contrary, the dimension of cluster 1 was quite reduced: 15 and 4 neighborhoods for Toronto and NY, respectively.

Therefore, the business analysis will be focused on Cluster 1 Neighborhoods.

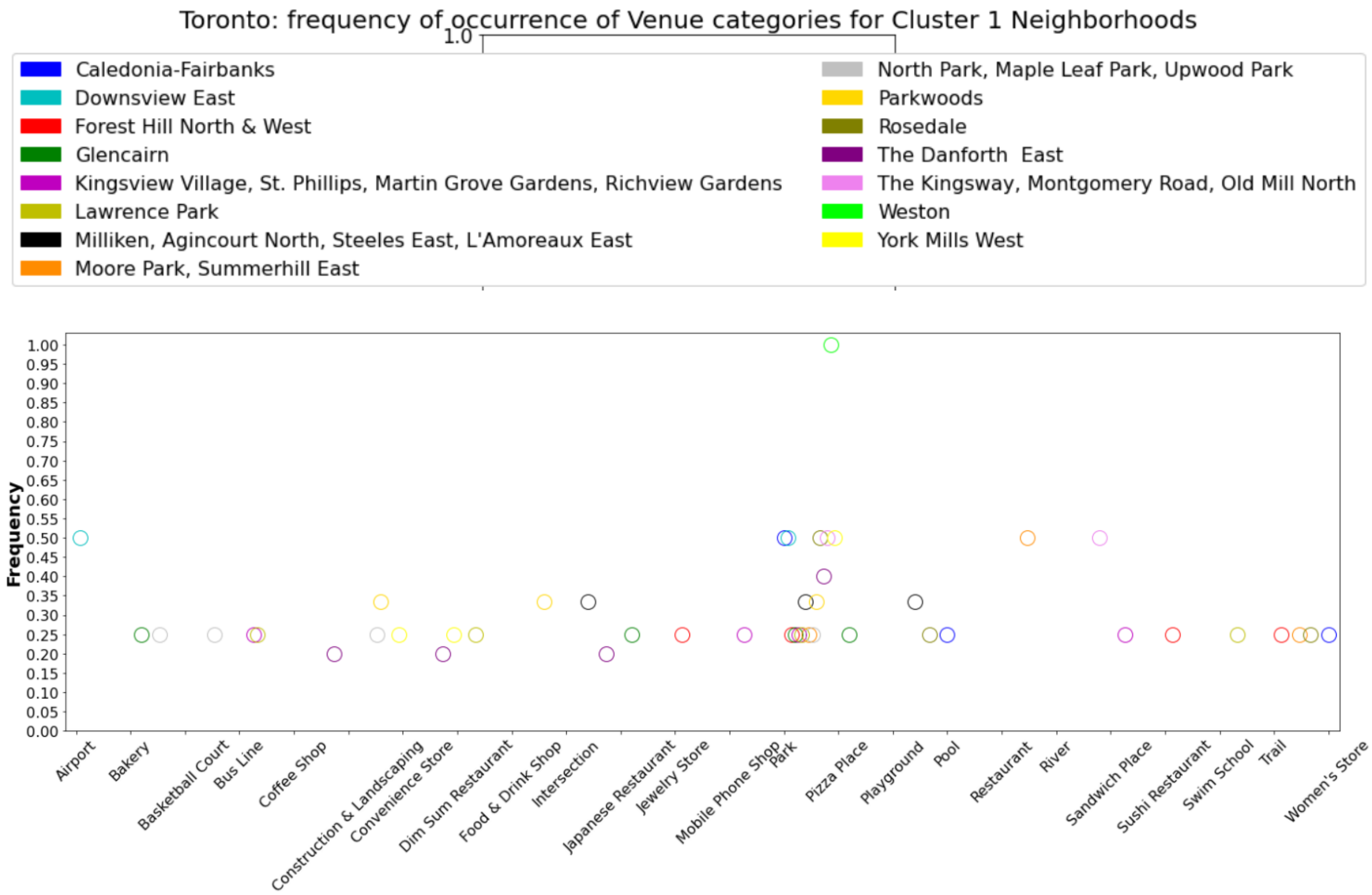
Considering only cluster 1 neighbothoods, only Toronto present the following 19 Venue categories: 'Airport', 'Bakery', 'Basketball Court', 'Bus Line', 'Coffee Shop', 'Construction & Landscaping', 'Dim Sum Restaurant', 'Food & Drink Shop', 'Intersection', 'Japanese Restaurant', 'Jewelry Store', 'Mobile Phone Shop', 'Restaurant', 'River', 'Sandwich Place', 'Sushi Restaurant', 'Swim School', 'Trail', "Women's Store".

Also considering only cluster 1 neighborhoods, only NY neighborhoods present the following 7 Venue categories: ‘Bagel Shop’, ‘Boat or Ferry’, ‘Bus Stop’, ‘Deli / Bodega’, ‘Grocery Store’, ‘Home Service’, ‘South American Restaurant’. As these Venue Categories are not yet present in Toronto, they might suggest new businesses opportunities in Toronto cluster 1 neighborhoods.

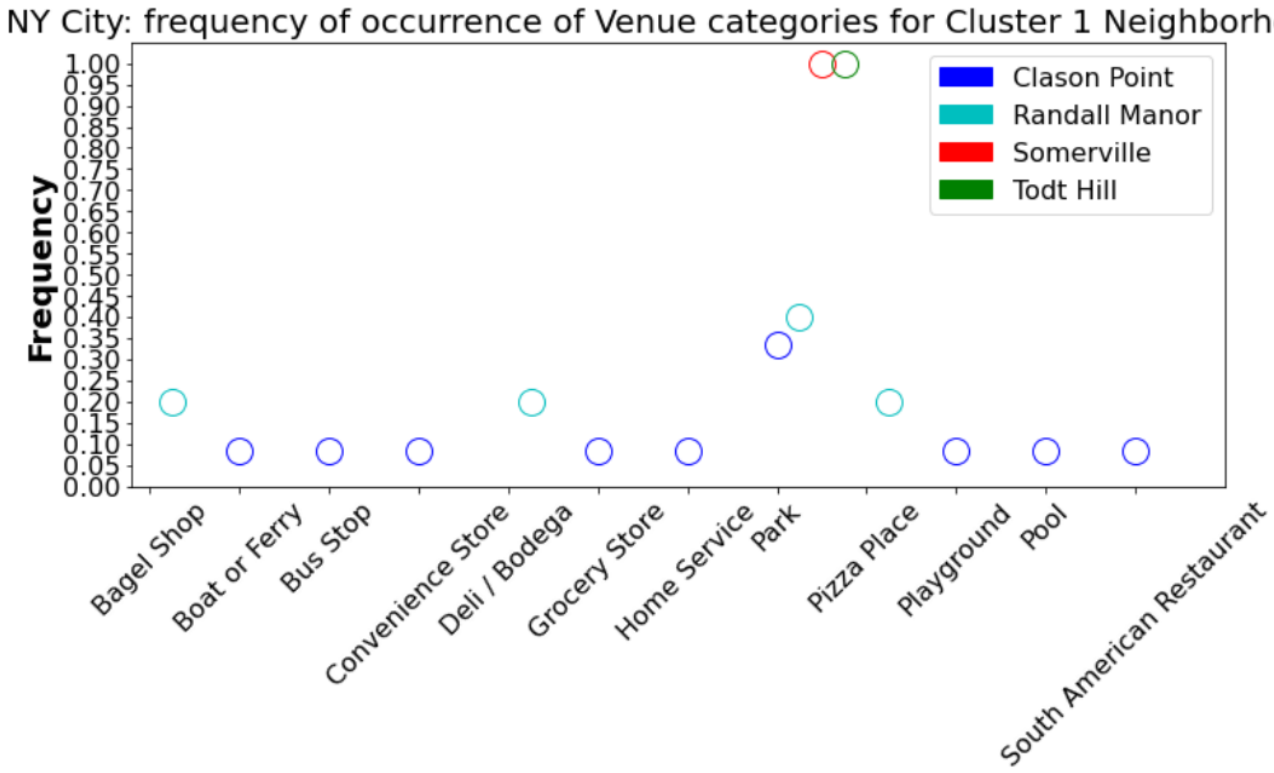
4.2. Plot analysis from Cluster 1 Neighborhoods

For plotting purposes, the cluster 1 datasets from both Toronto or NY were extracted from the merged Dataframe.

The plot below displays the frequency of occurrence of Venue categories for Cluster 1 Neighborhoods in Toronto:



The plot below displays the frequency of occurrence of Venue categories for Cluster 1 Neighborhoods in NY:



The plots shown above are helpful for illustrating the trends of Cluster 1 in terms of Venue category diversity. In other words, they refer to the Venue categories that are commonly found in the Cluster 1 neighborhoods from both Toronto and New York.

There, it can be seen that all neighborhods presented a Park as a Venue. Thus, it seems that this Venue category might be the main element defining cluster 1.

It is possible to infer that these Park-containing neighborhoods have various options of Food-related Venues, which is typical of locations where Parks are the main centre of activities. Moreover, a few sport-related Venues could be noted, such as Basketball court, River, Swim school, Pools, Trail, Landscaping, Boat or Ferry, Playgrounds; which reinforces the notion of Park areas frequented by a sportive public.

Furthermore, Bus Stop or Bus line are also common Venues among Cluster 1.

The “Bakery”, “Construction & Landscaping”, “Playground”, “Trail” Venue Categories are only present in Toronto, each one appearing in two neighborhoods. As these Venue categories are not present in NY neighborhoods, they might indicate an unfilled gap for new businesses in NY cluster 1 locations.

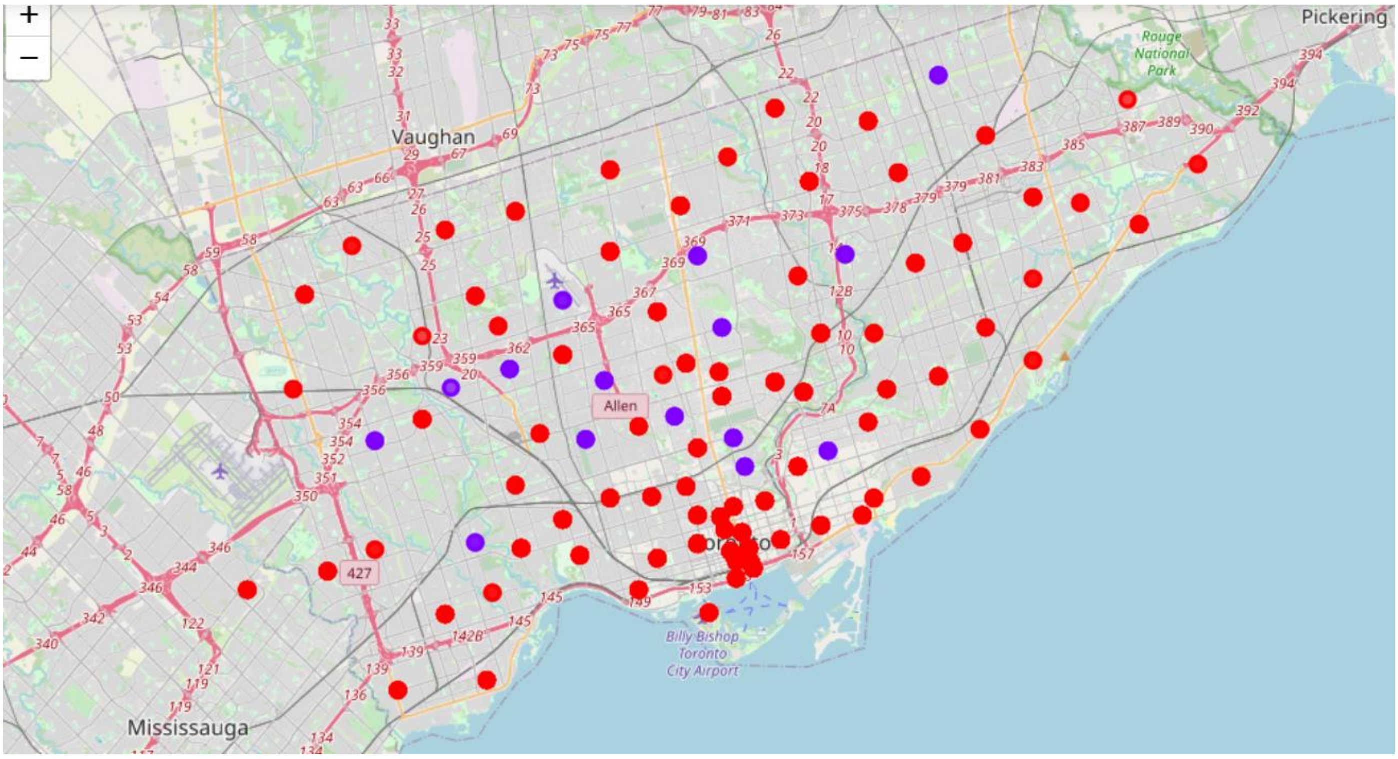
Concerning the Airport Venue observed from the Toronto results, it seemed to be an outlier at first sight. But closer inspection showed that the Airport is located at Downsview East, which is in fact a located in the Toronto district of North York.

Finally, the New York plot revealed two classes of Venues that might have a similarity: Jewelery store and Women’s store. This fact could indicate that women might preferentially frequent these park areas in NY city.

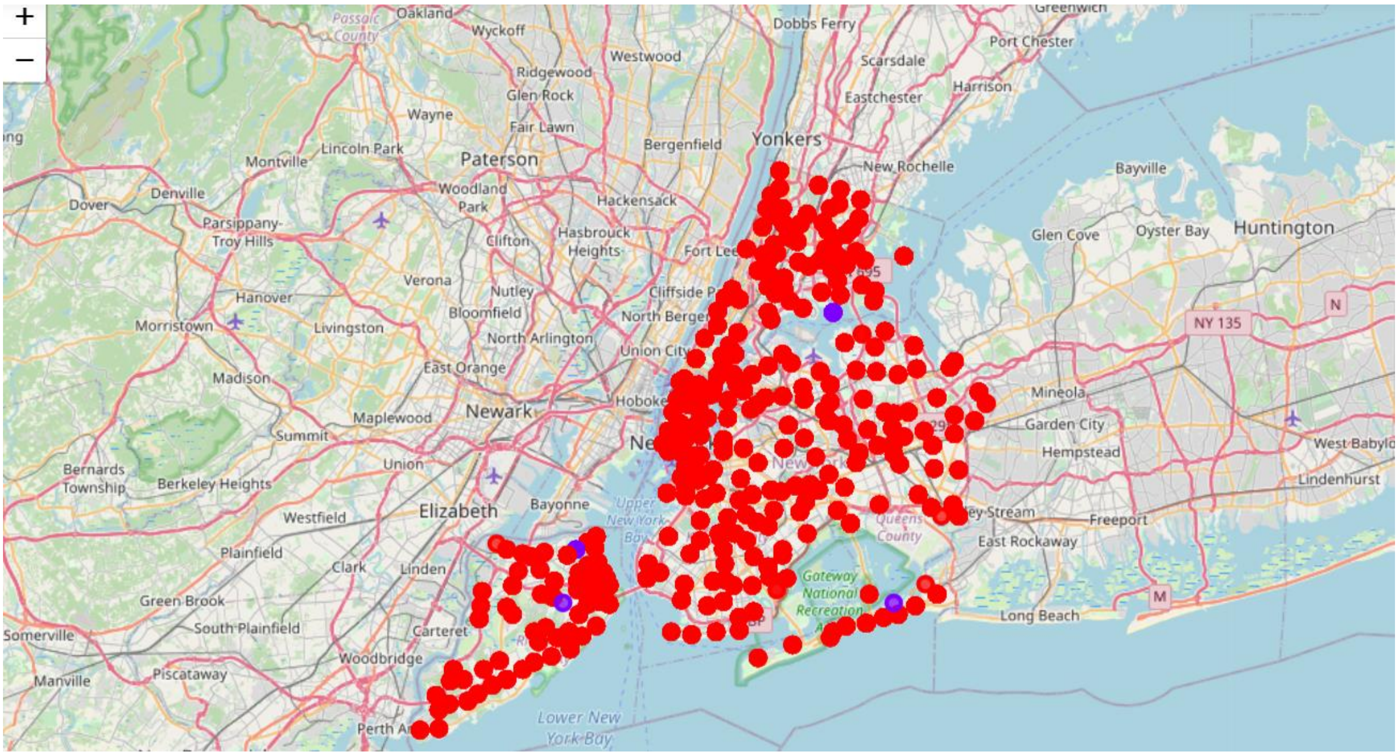
4.3. Clustered Toronto and NY maps using latitude and longitude values

In the cities’ maps with neighborhoods superimposed on top, clusters 0 and 1 appear in red and purple, respectively.

Toronto map with clustered neighborhoods:



NY City map with clustered neighborhoods:



From these cluster maps, it could be observed that Cluster 1 neighborhoods are not present in Central locations. They are rather present in peripherical areas, which corresponds well to the hypothesis that the primary defining characteristic of this cluster is the Park.

5. Conclusions

This project provided a data-oriented tool for promising neighborhoods for starting new businesses in NY city or Toronto.

The approach was based on the clustering of both cities neighborhoods according to their distributions in Venue categories. For that purpose, the K-means algorithm was employed, resulting in neighborhoods falling into two categories (cluster 0 and 1). Cluster 0 was shown to encompass the most of the neighborhoods, indicating a high diversity. On the contrary, cluster 1 presented a limited diversity in terms of the existing types of venues.

From further analysis of cluster 1, it was encountered that related neighborhoods were defined mainly by the presence of a Park. In addition, it was shown that the Venues categories in these locations are related to activities gravitating around the Park. In particular, it seemed that Food-related and Sport-related activities attract interest in these areas.

Therefore, stakeholders seeking for investing in Food/Sport-related business in Toronto or NY City, might benefit from the Neighborhoods indicated in Cluster 1. Alternatively, suppose potential stakeholders would be interested in new Venues in Cluster 1 neighborhoods, but doubt about the type of business. In that case, they might also consider the Women’s market, such as Jewelery stores.

The predictive task of determining optimal locations for new Venues was based on the distribution of Venues category solely.

The approach here discussed was based on the assumption that neighborhoods from different cities might belong to the same cluster defined by their Venue categories. Although able to provide some data-oriented guidance, it should be emphasized that the Venue category is certainly not the only parameter influencing the process of defining optimal locations. The final decision of a stakeholder should also consider other neighborhood characteristics, potentially presenting a clear correlation with the rate of the business rate of success.