FINE TUNING OF PARAMETERS FOR IMPROVED CLUSTERING OF NEIGHBORHOODS USING KMEANS

INTRODUCTION/BUSINESS PROBLEM

In week 3, a project about clustering the Toronto neighborhoods was assigned. As a result of the project, Foursquare was used to obtain venue data, and Kmeans was used to cluster the Toronto neighborhoods. Figure 1 shows the results of clustering the Toronto neighborhoods in 5 groups using a radius of 500 m to explore the neighborhoods.

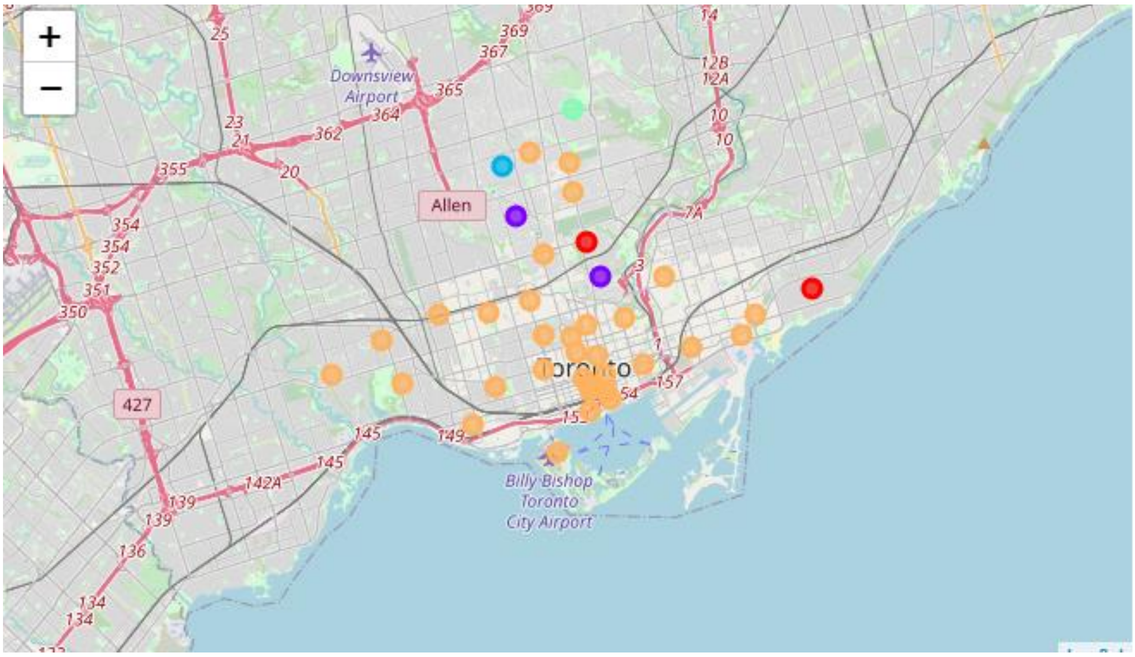


Figure 1. Clustering of 39 Toronto neighborhoods using a radius of 500 m for exploration.

Here, we observe several problems:

1- The number of neighborhoods is not evenly distributed, this is, the number of neighborhoods in each cluster is the following:
   a. Cluster 0: 2.
   b. Cluster 1: 2.
   c. Cluster 2: 1.
   d. Cluster 3: 1.
   e. Cluster 4: 34.
2- Clusters 0 (red) and 1 (violet) with 2 neighborhoods, occupy different areas in the map.
3- Cluster 4 (orange) covers almost the entire area of analysis. Only cluster 3 (green) in the northern part of Toronto, appears outside the area covered by cluster 4.

In order to solve this problem, the radius of the exploration data was changed from 500 m to 1000 m, yielding results shown in figure 2, which is a much better result.

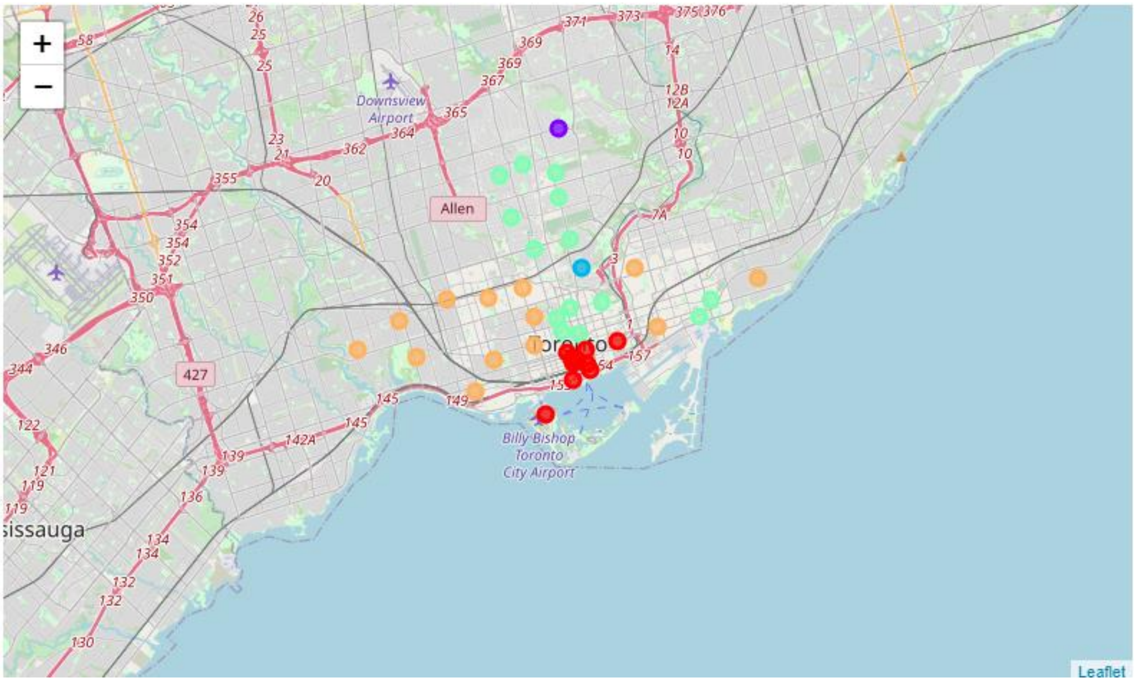


Figure 2. Clustering of 39 Toronto neighborhoods using a radius of 1,000 m for exploration.

From figure 2 we can make several observations:

1- The distribution of the clusters is more even. The new distribution is as follows:
   a. Cluster 1: 10 neighborhoods.
   b. Cluster 2: 1 neighborhood.
   c. Cluster 3: 1 neighborhood.
   d. Cluster 4: 14 neighborhoods.
   e. Cluster 5: 13 neighborhoods.
2- Clusters 0 (red) and 3 (green) appear to be “clustered” in single regions, while cluster 4 (orange) appears to be split in two regions.

Even though the result in figure 2 is much better than the one in figure 1, one can ask why we still have issues. The main reason is that we did not provide the location information to the Kmeans clustering algorithm, therefore, the algorithm does not know where the neighborhoods are located, and makes the clustering based on the similarities among venues alone. Hence, it should not surprise us that the clusters are split in several regions as we are trying to display them in a map without providing the neighborhood location. This situation might bewilder the user as it happened to me. Therefore, one of the objectives of this project is to improve the results of the kmeans algorithm by including the location data among the parameters to be passed to the algorithm, in order to make the clusters appear more grouped in the map.

In order to test the previous hypothesis, an experiment was designed. In this experiment, only the location of the 39 Toronto neighborhoods was passed to the Kmeans algorithm to perform the analysis. The results are shown in figure 3.

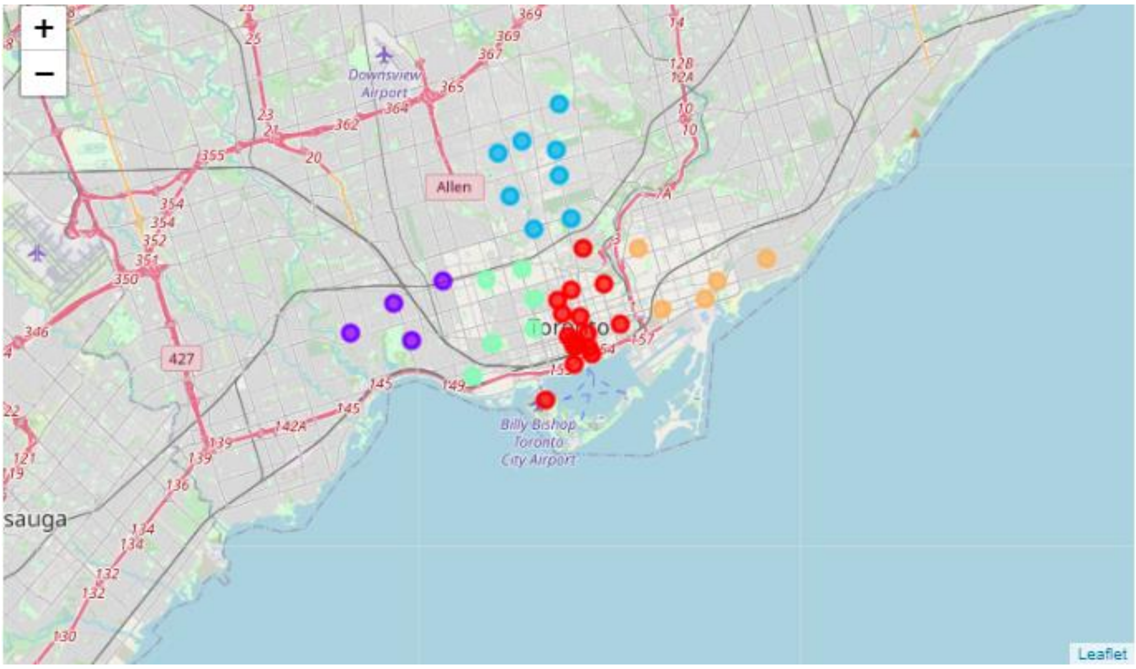


Figure 3. Clustering of 39 Toronto neighborhoods using location data only.

By analyzing figure 3, several observations can be made:

1- The cluster areas are more evenly distributed on the map, this is, the clusters appear to occupy similar size areas.
2- The number of clusters is more evenly distributed, having the following number of neighborhoods:
    a. Cluster 1: 16.
    b. Cluster 2: 4.
    c. Cluster 3: 8.
    d. Cluster 4: 6.
    e. Cluster 5: 5.

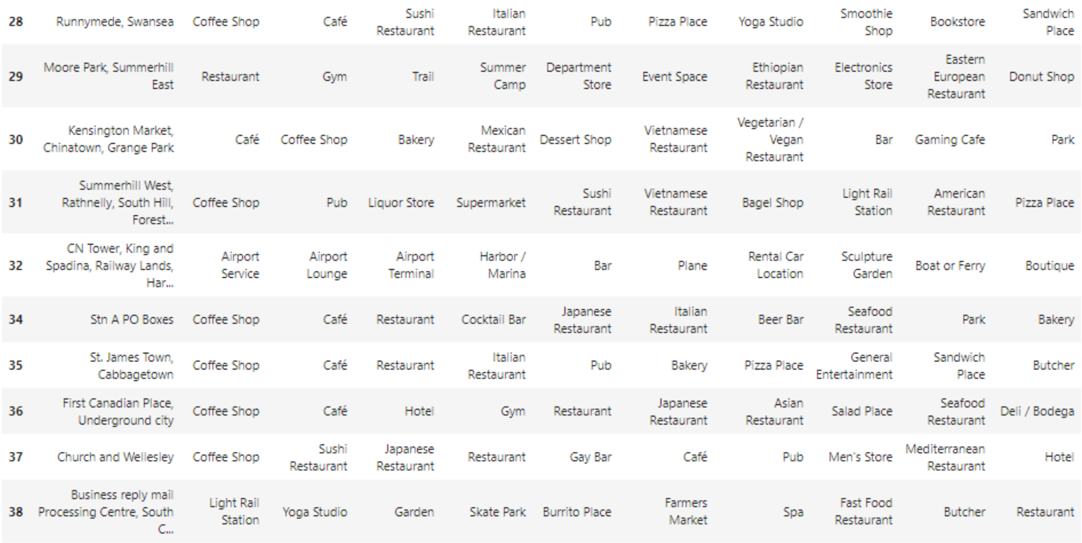
    Now, we do not have clusters with one neighborhood, and the values among them do vary as much as in the previous two cases.
3- Each cluster occupies a single area, this is, the clusters are not split in two or more regions, which is what we would expect from a clustering algorithm.

The results of this experiment show that the location information can be used in combination with the venue data to improve the results of the kmeans algorithm. The issue/problem is that the results of the Kmeans algorithm as shown in figure 2, using only the venue data, do not provide a good clustering of the neighborhoods, even though a good amount of time and effort was spent adjusting the parameters to provide the best possible clustering of the neighborhoods.

Another issue with the values passed to the Kmeans algorithm, is the density of the venues in each neighborhood. The venue information passed to the Kmeans algorithm is in the Toronto_grouped dataframe. An analysis of this dataframe reveals that the sum of all the venues in one neighborhood equals 1. The problem is that each neighborhood has a different number of venues and the sum of all the venues for each neighborhood is equal to one. Therefore, a neighborhood with two restaurants and two parks and another with one restaurant and one park will appear with the same values in the Toronto_grouped dataframe and therefore will appear equal to the Kmeans algorithm, while the reality is that one of them has double the amount of venues, this is what I call the density issue. I think this problem might be solved by not normalizing the sum of all venues in one row to one, which will allow us to make a better clustering of the neighborhoods.

Another issue that was observed while developing the assignment for week 3, is that the one-hot-coding technique assigns a number 1 to a venue, if present in that neighborhood, no matter what the category of the venue is. For example, if a restaurant exists, a number one will be assigned. The problem is that the restaurant category has been subdivided in many different subcategories. As a result, we may have 20 types of restaurants, and 1 type of park, this causes the results of the Kmeans algorithm to be biased towards restaurants. For example, a neighborhood with 10 different types of restaurants and one bank will appear very similar to a neighborhood having 10 different types of restaurant and no banks, while it may be fairer to say that neighborhood one has restaurants and banks and neighborhood two has restaurants and no banks.

Let us look at table 1 to illustrate this point.



28	Runnymede, Swansea	Coffee Shop	Café	Sushi Restaurant	Italian Restaurant	Pub	Pizza Place	Yoga Studio	Smoothie Shop	Bookstore	Sandwich Place
29	Moore Park, Summerhill East	Restaurant	Gym	Trail	Summer Camp	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop
30	Kensington Market, Chinatown, Grange Park	Café	Coffee Shop	Bakery	Mexican Restaurant	Dessert Shop	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Bar	Gaming Cafe	Park
31	Summerhill West, Rathnelly, South Hill, Forest...	Coffee Shop	Pub	Liquor Store	Supermarket	Sushi Restaurant	Vietnamese Restaurant	Bagel Shop	Light Rail Station	American Restaurant	Pizza Place
32	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Harbor / Marina	Bar	Plane	Rental Car Location	Sculpture Garden	Boat or Ferry	Boutique
34	Stn A PO Boxes	Coffee Shop	Café	Restaurant	Cocktail Bar	Japanese Restaurant	Italian Restaurant	Beer Bar	Seafood Restaurant	Park	Bakery
35	St. James Town, Cabbagetown	Coffee Shop	Café	Restaurant	Italian Restaurant	Pub	Bakery	Pizza Place	General Entertainment	Sandwich Place	Butcher
36	First Canadian Place, Underground city	Coffee Shop	Café	Hotel	Gym	Restaurant	Japanese Restaurant	Asian Restaurant	Salad Place	Seafood Restaurant	Deli / Bodega
37	Church and Wellesley	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Restaurant	Gay Bar	Café	Pub	Men's Store	Mediterranean Restaurant	Hotel
38	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Garden	Skate Park	Burrito Place	Farmers Market	Spa	Fast Food Restaurant	Butcher	Restaurant

Table 1. Top 10 most common venues for a cluster of neighborhoods in the Toronto area (radius = 500 m).

If we look at neighborhoods 32 (CN Tower King and Spadina…) and 34 (Stn A PO Boxes), we notice that their most common venues are completely different. For neighborhood 32, the most common venues are related to transportation services, while for neighborhood 34, the most common venues are restaurants and a park. This is, neighborhood 32 seems to be a commercial area, and neighborhood 32 is very likely a residential area. The question is, why are they in the same cluster? Why do the other clusters have only 2, 1, 1, and 1 neighborhood and one cluster have 34 neighborhoods? The answer might be in the large number of subcategories in some categories (like restaurants), that seems to bias the results of the Kmeans algorithm. While I am not 100% positive that this is the reason why neighborhoods 32 and 34 are in the same cluster, I am confident that we are losing information when we split a category into many different subcategories, and distort the results of the Kmeans algorithm.

Finally, another aspect to be considered is the radius of the neighborhoods for venue data collection. This parameter had a large impact in the results of the Kmeans algorithm. I believe that some study needs to be done to provide guidance on how to choose this parameter. For example, at first, a radius of 500 m was chosen and Kmeans returned clusters with 34, 2, 1, 1 and 1 neighborhood. When using a radius of 1,000 m the numbers were 10, 1, 1, 14 and 13 neighborhoods, a much better result. The radius should not be chosen too small because we would not get enough venue information, and should not be too large, because we might have some overlap with other neighborhoods. From this reasoning, we conclude that the neighborhood location and separation between neighborhoods is of importance and should be taken into account in order to make a good selection of the radius.

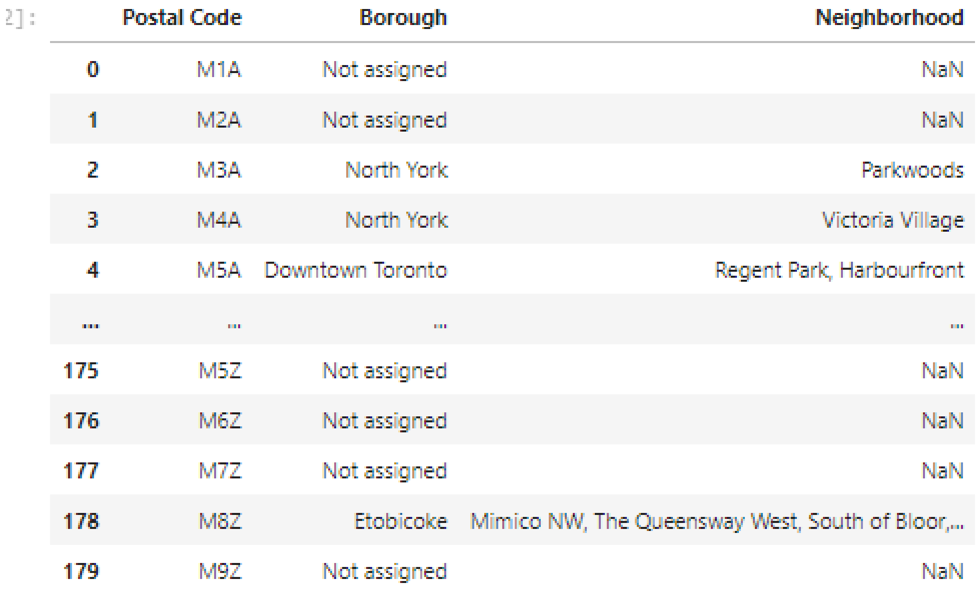
All the previously mentioned issues/problems affect the performance of the Kmeans algorithm, since they negatively influence the clustering results no matter what the city is. Therefore, the solution of this problem is of importance to anyone interested in exploring any city/place in any part of the world using the Kmeans algorithm.

DATA

The data that will be used to solve the problem are the following:

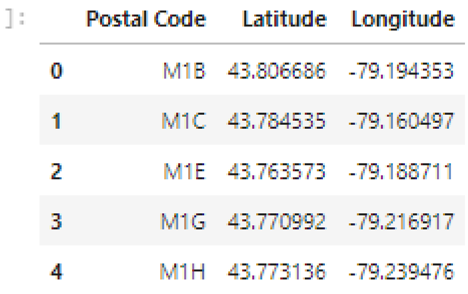
1. Toronto neighborhood information. This information was downloaded from the url: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Table 2 shows top and bottom rows of the table.
2. CSV file with location coordinates for the Toronto neighborhoods. Downloaded from the url: http://cocl.us/Geospatial_data. Table 3 shows the 5 top rows of the csv file.
3. Map of Toronto provided by Folium. Shown in figure 4.
4. Toronto venue information provided by Foursquare (see table 4).

First, the data will be cleaned following the directions given in the assignment for week 3, this is, the Postal Codes rows with no borough assigned will be removed, then the neighborhood data will be merged with the location data to create a new dataframe, and the neighborhoods without the word Toronto in their names will be dropped, to end up with a dataframe containing 39 neighborhoods with their location as shown in table 5.



	Postal Code	Borough	Neighborhood
0	M1A	Not assigned	NaN
1	M2A	Not assigned	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
...	...	...	...
175	M5Z	Not assigned	NaN
176	M6Z	Not assigned	NaN
177	M7Z	Not assigned	NaN
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	NaN

Table 2. Toronto neighborhood information.



	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Table 3. Toronto neighborhood coordinates.

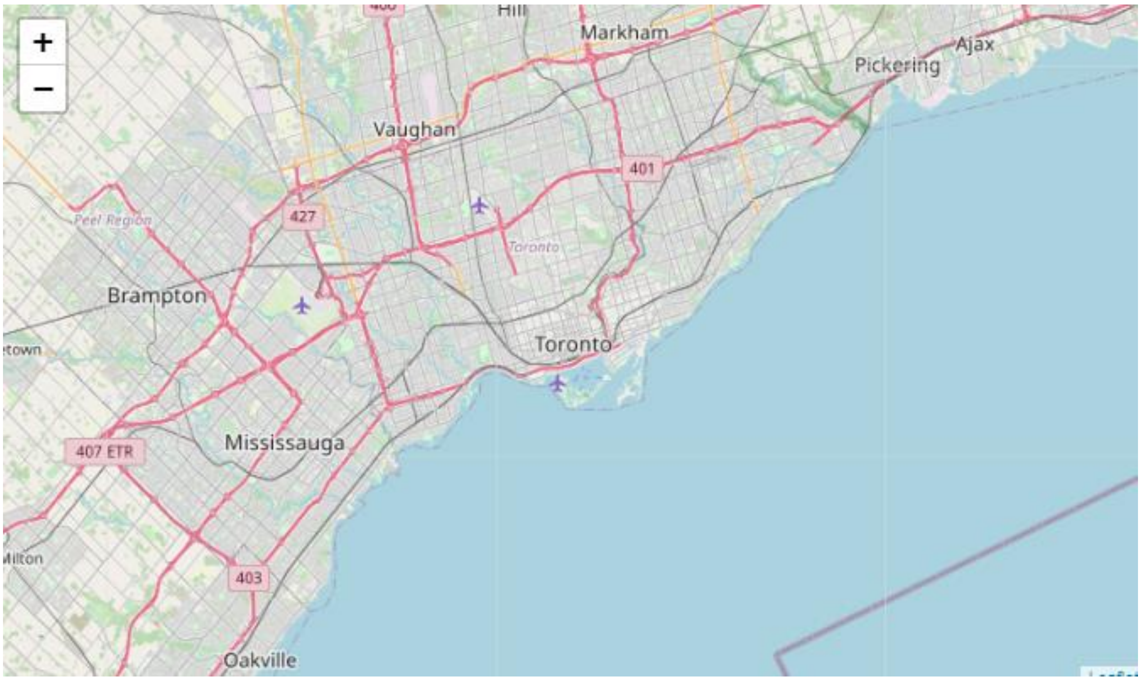
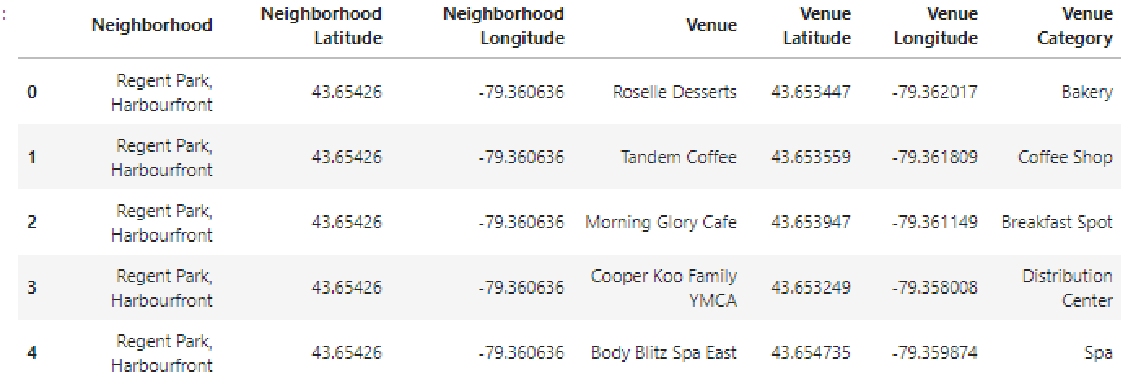
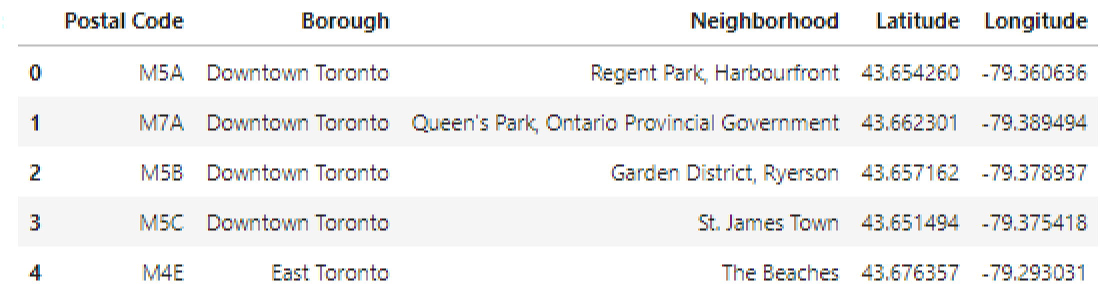


Figure 4. Map of Toronto provided by Folium.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
3	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
4	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa

Table 4. Toronto venue information.



	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 5. Toronto neighborhood information.

The data will be used in the following way to try to find the solution to the problem:

1. The neighborhood location data will be passed to the Kmeans algorithm to help cluster neighborhoods within the same area.
2. The venue information provided by Foursquare will be processed in a different way so that normalization does not make neighborhoods appear as if they have the same number of venues.
3. The venue information will have to be analyzed to take into account that different venues may belong to the same category. In order to solve this problem the category of the venues provided by foursquare will be used.
4. The location data will be used to calculate the distance between neighborhoods, to help make a conscious selection of the radius.

METHODOLOGY

Addition of the location information to the toronto_grouped dataframe

The first step was to add the location information to the toronto_grouped dataframe. In order to do this, the location data was normalized with zero mean and standard deviation equal to 1, then the location data was added to the toronto_grouped dataframe, and the clustering done. Figure 5 shows the results of the Kmeans algorithm, and table 6 shows the distribution of the neighborborhoods in each cluster.

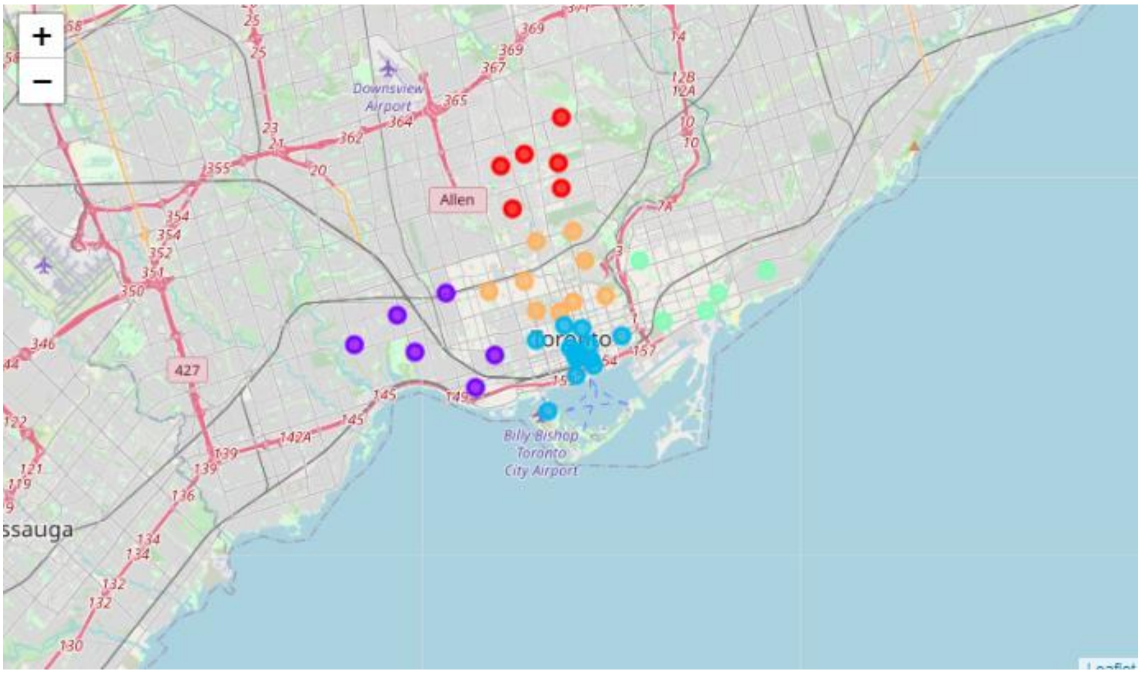
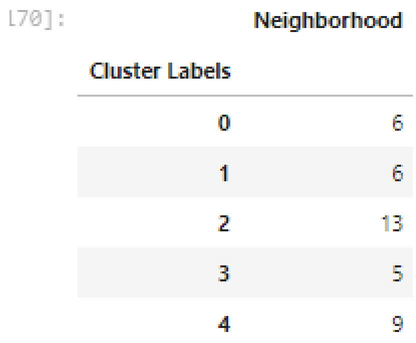


Figure 5. Effect of adding the location information to the Toronto_grouped dataframe.



[70]:

Cluster Labels	Neighborhood
0	6
1	6
2	13
3	5
4	9

Table 6. Distribution of the neighborhoods in figure 5.

Comparing figures 1 and 5 we notice that the addition of the location information have had a profound effect on the results. In figure 1, no location information was passed to the Kmeans algorithm, in figure 5, the location and venue information was passed to the Kmeans algorithm, resulting in a much better clustering as seen in table 6. Now, let's compare figures 3 and 5. In figure 3, only location data was passed to the Kmeans algorithm. The distribution of the neighborhoods in each cluster for figure 3 is 16, 4, 8, 6 and 5 neighborhoods, and for figure 5 is 6, 6, 13, 5 and 9 neighborhoods. We also notice that the distribution of the clusters on the maps is different. However, it seems that the location information has influenced the results more than what we would like to, this is, I would like to see more effect of the venue information in the distribution of the clusters.

In order to let the venue information have more effect on the clustering results, a scale factor of 1/5 was applied to the location data. Figure 6 shows the neighborhood distribution on the map, and table 7 shows the numerical distribution. Both, figure 6 and table 7 show how the venue information tends to have a larger effect on the results (compare with figure 5).

Processing of Results of the One Hot Coding Technique

The results provided by the One-Hot-Coding technique were processed in a different way. Instead of grouping and taking the mean of the toronto_onehot dataframe, the results were grouped and added, so that for each neighborhood, the total number of venues can be calculated by adding the number of individual venues, while in the previous case (week 3 assignment), the summation of all the venues for each neighborhood was equal to 1, which is obviously incorrect. This processing technique preserves the number of venues per neighborhood (the density) which allows us to obtain better results.

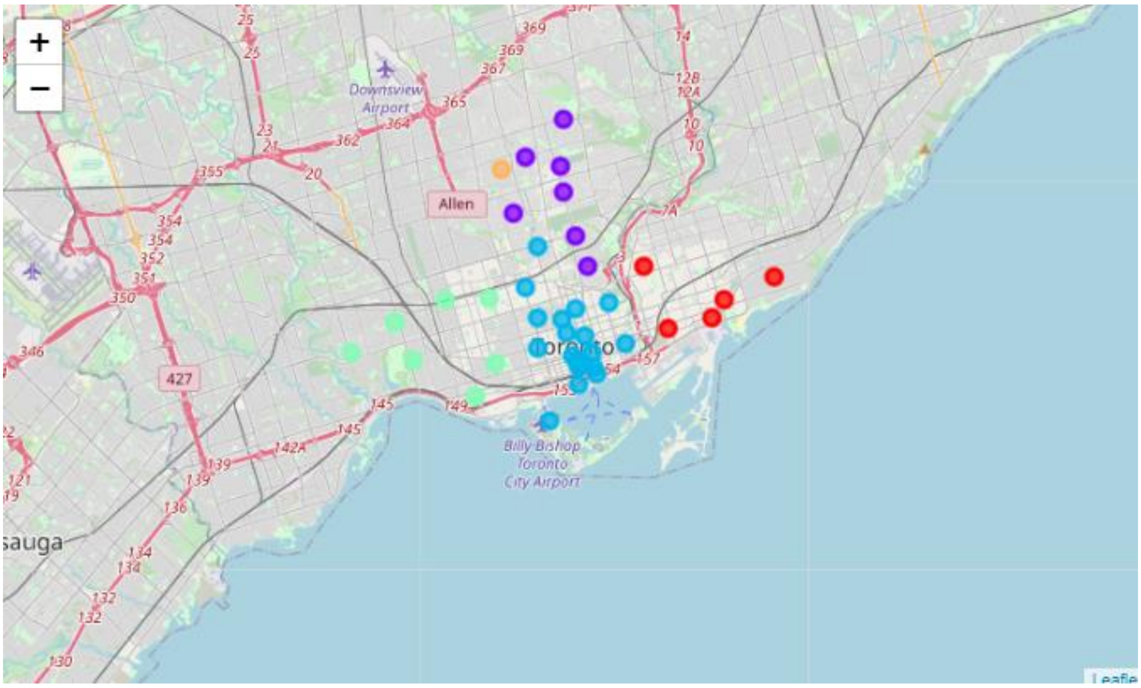
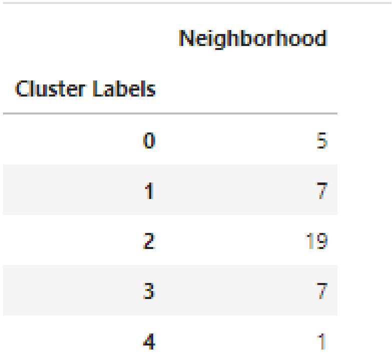


Figure 6. Toronto neighborhood clustering using scaled location data.



	Neighborhood
Cluster Labels	
0	5
1	7
2	19
3	7
4	1

Table 7. Distribution of neighborhoods for figure 6.

Figure 7 shows the new distribution of the neighborhoods. The effect of the new processing technique can be seen as the clusters tend to be distributed around the downtown area. Cluster 3 in green is in the middle, surrounded by clusters 0 (red) and 4 (orange), and then clusters 1 (violet) and 2 (blue). This distribution makes sense as the venues of the downtown area should be denser than in the suburbs of the big city of Toronto. Table 8 displays the number of neighborhoods in each cluster.

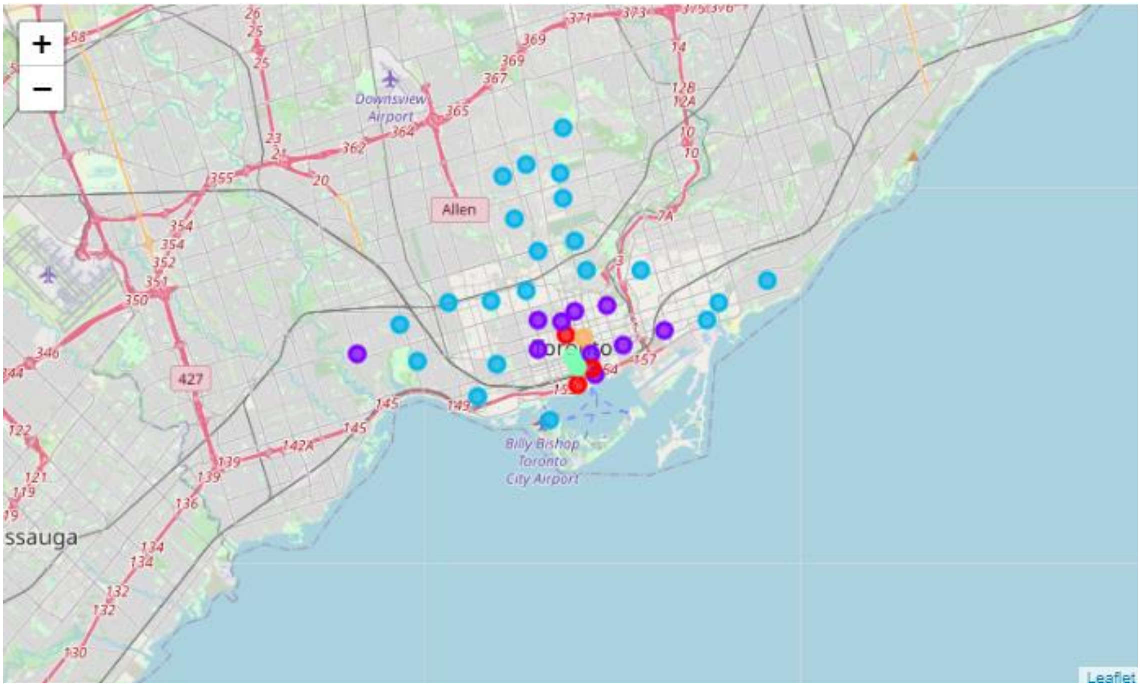
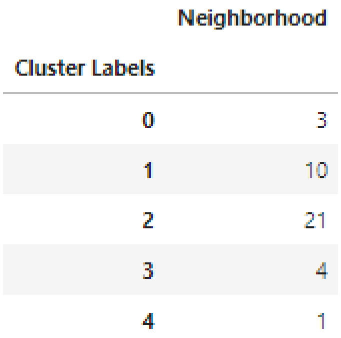


Figure 7. Distribution of the Toronto neighborhoods taking into account the number of venues per neighborhood.



	Neighborhood
Cluster Labels	
0	3
1	10
2	21
3	4
4	1

Table 8. Distribution of the neighborhoods in figure 7.

Regrouping the venue categories

For the 39 Toronto neighborhoods under analysis, Foursquare returned 237 venue categories. As explained above, a better result might be obtained by regrouping the venue categories into main groups. While reading the Foursquare website, it was found the following main venue categories:

1. Arts & enterteinment
2. College & university
3. Event
4. Food

5. Nightlife spot
6. Outdoors & recreation
7. Professional & other places
8. Residence
9. Shop & service
10. Travel & transport

The 237 venue categories of each neighborhood were regrouped in a new dataframe called Toronto_regrouped, where each category was mapped into one of the 10 main categories. The resulting dataframe (including location data) was input to the Kmeans algorithm to cluster the neighborhoods. Also, the location data was normalized to the range between 0 and 1 for easier scaling and control.

Figure 8 shows the new distribution of the neighborhoods. It can be noticed that the clusters are centered on the downtown area, and that clusters 0 (red), 2 (blue) and 3 (green) are split in different regions due to the low range of the location data (from 0 to 1) compared to the values of venue data, which was done on purpose to let the algorithm do the clustering based on venue information mostly. Table 9 shows the number of neighborhoods in each cluster.

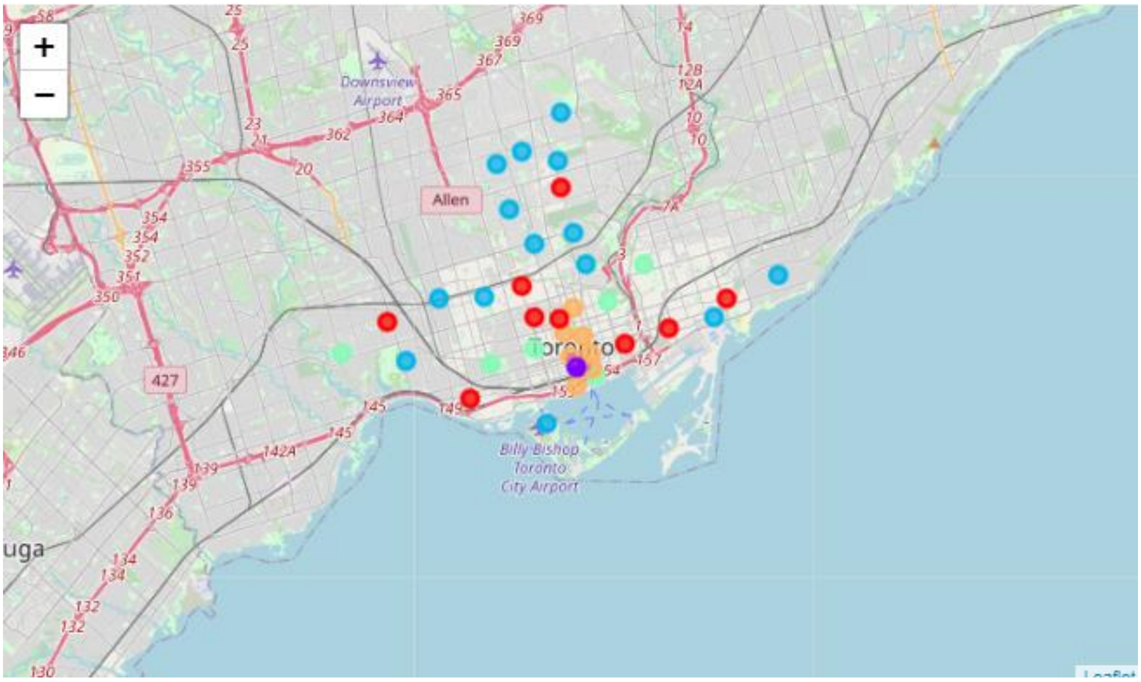
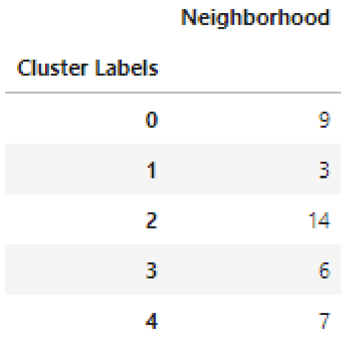


Figure 8. Distribution of the neighborhoods using 10 main venue categories.



	Neighborhood
Cluster Labels	
0	9
1	3
2	14
3	6
4	7

Table 9. Number of neighborhoods per cluster in figure 8.

Calculation of the distance between neighborhoods

The distance between two locations based on their coordinates can be calculated using the Harvesine formula. A function was implemented in the notebook to calculate the distance from each of the neighborhoods to the remaining 38 neighborhoods. From these results, it was found that the minimum distance between two neighborhoods is 150 m, this is well below the 1,000 m distance used in the assignment of week 3, and also below the 500 m that is being used in this final project.

Further analysis of the distance data showed that the average distance among all neighborhoods is 4.94 Km, and only 5 neighborhoods are separated from other neighborhoods by less than 500 m. Due to the large variation in the distance among neighborhoods and since the number of neighborhoods with a separation of less than 500 m is only 5, It was decided to keep using 500 m as the radius for exploration of the Toronto neighborhoods.

RESULTS

In order to regroup clusters 0 (red), 2 (blue), and 3 (green) from figure 8, the location data was scaled by a factor of 50, giving the results in figure 9, where we notice that 4 neighborhoods have be regrouped in single areas, only cluster 0 in red appears to be surrounding cluster 1 in violet. Table 10 shows the distribution of the neighborhoods in each cluster.

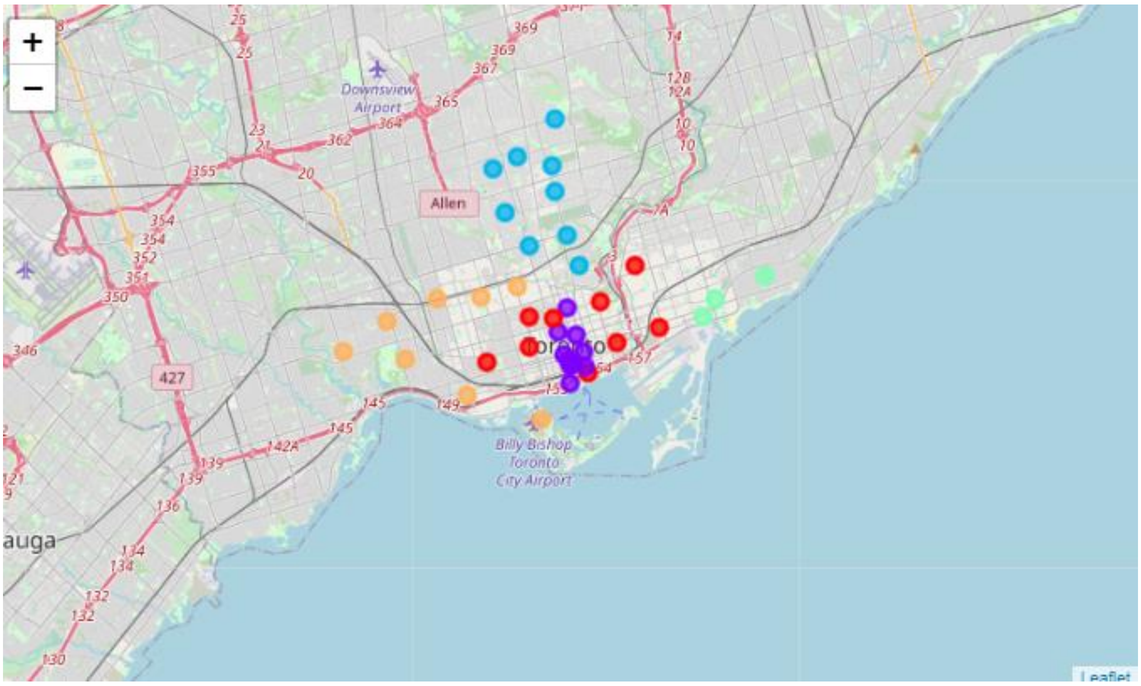
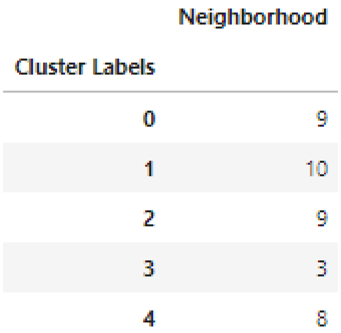


Figure 9. Clustering of the Toronto neighborhoods using a scale factor of 50.

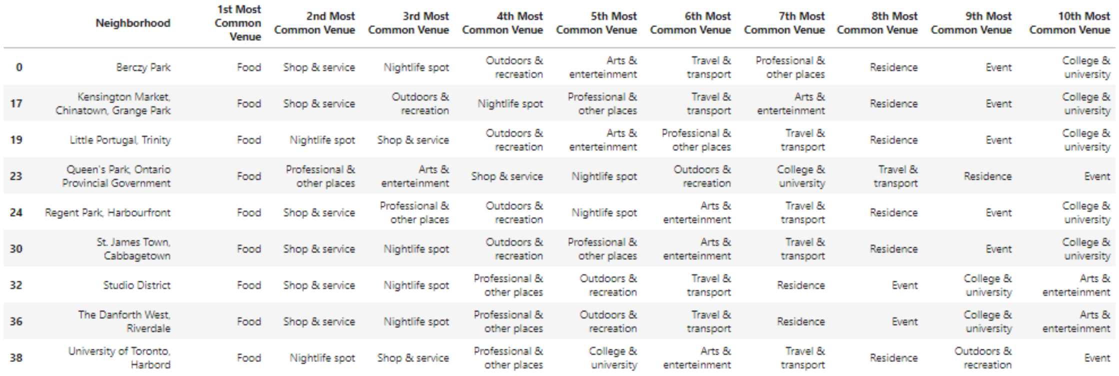


	Neighborhood
Cluster Labels	
0	9
1	10
2	9
3	3
4	8

Table 10. Distribution of the neighborhoods in figure 9.

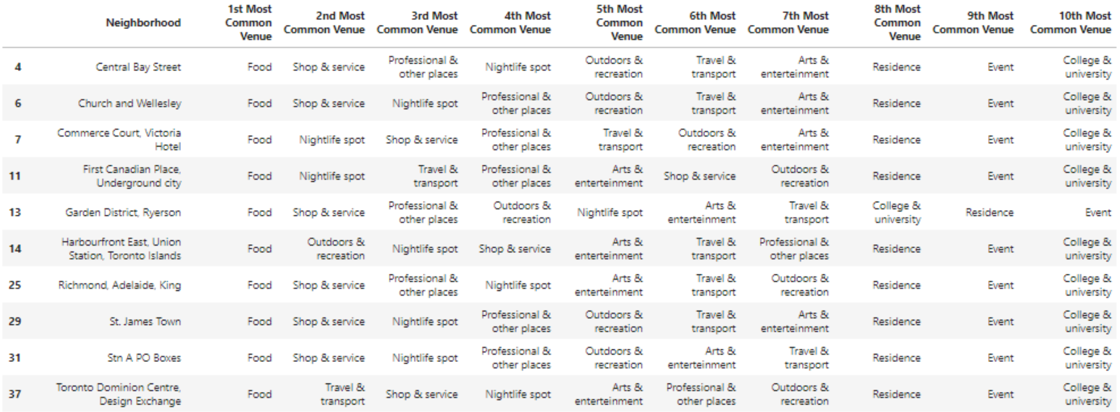
Tables 11, 12, 13, 14 and 15 show the most common venues for all the neighborhoods in each cluster. From the location of the clusters in figure 9 and the most common venues of each neighborhood, the main features of each neighborhood can be extracted in order to label each cluster. The labels assigned are the following:

1. Cluster 0 (red): Central zone, shopping-recreational area.
2. Cluster 1 (violet): Downtown, business-recreational area.
3. Cluster 2 (blue): Northern zone, residence-recreational area.
4. Cluster 3 (green): Eastern zone, recreational area.
5. Cluster 4 (orange): Western zone, residential area.



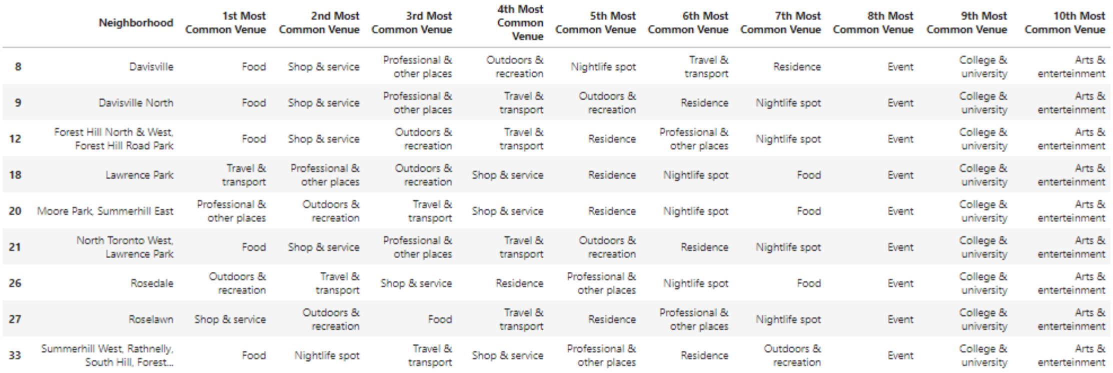
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Food	Shop & service	Nightlife spot	Outdoors & recreation	Arts & enterteinment	Travel & transport	Professional & other places	Residence	Event	College & university
17	Kensington Market, Chinatown, Grange Park	Food	Shop & service	Outdoors & recreation	Nightlife spot	Professional & other places	Travel & transport	Arts & enterteinment	Residence	Event	College & university
19	Little Portugal, Trinity	Food	Nightlife spot	Shop & service	Outdoors & recreation	Arts & enterteinment	Professional & other places	Travel & transport	Residence	Event	College & university
23	Queen's Park, Ontario Provincial Government	Food	Professional & other places	Arts & enterteinment	Shop & service	Nightlife spot	Outdoors & recreation	College & university	Travel & transport	Residence	Event
24	Regent Park, Harbourfront	Food	Shop & service	Professional & other places	Outdoors & recreation	Nightlife spot	Arts & enterteinment	Travel & transport	Residence	Event	College & university
30	St. James Town, Cabbagetown	Food	Shop & service	Nightlife spot	Outdoors & recreation	Professional & other places	Arts & enterteinment	Travel & transport	Residence	Event	College & university
32	Studio District	Food	Shop & service	Nightlife spot	Professional & other places	Outdoors & recreation	Travel & transport	Residence	Event	College & university	Arts & enterteinment
36	The Danforth West, Riverdale	Food	Shop & service	Nightlife spot	Professional & other places	Outdoors & recreation	Travel & transport	Residence	Event	College & university	Arts & enterteinment
38	University of Toronto, Harbord	Food	Nightlife spot	Shop & service	Professional & other places	College & university	Arts & enterteinment	Travel & transport	Residence	Outdoors & recreation	Event

Table 11. Cluster 0 (red): Central zone, shopping-recreational area.



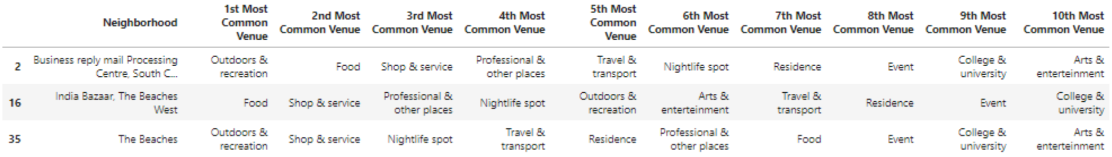
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Central Bay Street	Food	Shop & service	Professional & other places	Nightlife spot	Outdoors & recreation	Travel & transport	Arts & enterteinment	Residence	Event	College & university
6	Church and Wellesley	Food	Shop & service	Nightlife spot	Professional & other places	Outdoors & recreation	Travel & transport	Arts & enterteinment	Residence	Event	College & university
7	Commerce Court, Victoria Hotel	Food	Nightlife spot	Shop & service	Professional & other places	Travel & transport	Outdoors & recreation	Arts & enterteinment	Residence	Event	College & university
11	First Canadian Place, Underground city	Food	Nightlife spot	Travel & transport	Professional & other places	Arts & enterteinment	Shop & service	Outdoors & recreation	Residence	Event	College & university
13	Garden District, Ryerson	Food	Shop & service	Professional & other places	Outdoors & recreation	Nightlife spot	Arts & enterteinment	Travel & transport	College & university	Residence	Event
14	Harbourfront East, Union Station, Toronto Islands	Food	Outdoors & recreation	Nightlife spot	Shop & service	Arts & enterteinment	Travel & transport	Professional & other places	Residence	Event	College & university
25	Richmond, Adelaide, King	Food	Shop & service	Professional & other places	Nightlife spot	Arts & enterteinment	Travel & transport	Outdoors & recreation	Residence	Event	College & university
29	St. James Town	Food	Shop & service	Nightlife spot	Professional & other places	Outdoors & recreation	Travel & transport	Arts & enterteinment	Residence	Event	College & university
31	Stn A PO Boxes	Food	Shop & service	Nightlife spot	Professional & other places	Outdoors & recreation	Arts & enterteinment	Travel & transport	Residence	Event	College & university
37	Toronto Dominion Centre, Design Exchange	Food	Travel & transport	Shop & service	Nightlife spot	Arts & enterteinment	Professional & other places	Outdoors & recreation	Residence	Event	College & university

Table 12. Cluster 1 (violet): Downtown, business-recreational area.



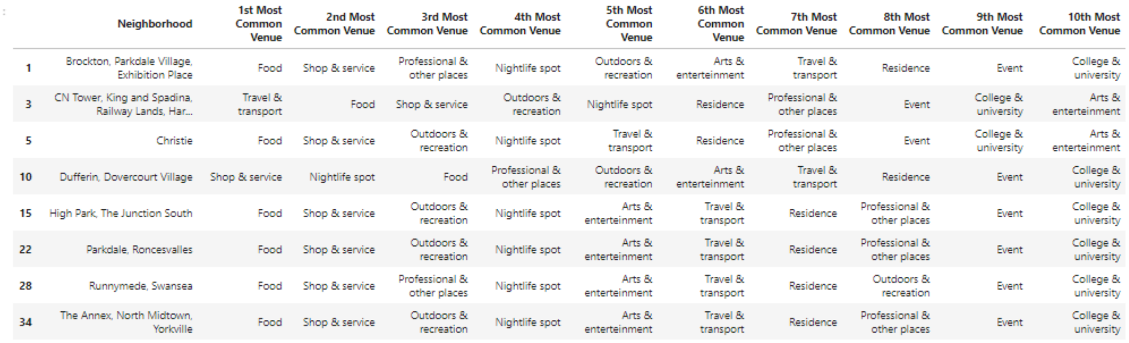
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Davisville	Food	Shop & service	Professional & other places	Outdoors & recreation	Nightlife spot	Travel & transport	Residence	Event	College & university	Arts & enterteinment
9	Davisville North	Food	Shop & service	Professional & other places	Travel & transport	Outdoors & recreation	Residence	Nightlife spot	Event	College & university	Arts & enterteinment
12	Forest Hill North & West, Forest Hill Road Park	Food	Shop & service	Outdoors & recreation	Travel & transport	Residence	Professional & other places	Nightlife spot	Event	College & university	Arts & enterteinment
18	Lawrence Park	Travel & transport	Professional & other places	Outdoors & recreation	Shop & service	Residence	Nightlife spot	Food	Event	College & university	Arts & enterteinment
20	Moore Park, Summerhill East	Professional & other places	Outdoors & recreation	Travel & transport	Shop & service	Residence	Nightlife spot	Food	Event	College & university	Arts & enterteinment
21	North Toronto West, Lawrence Park	Food	Shop & service	Professional & other places	Travel & transport	Outdoors & recreation	Residence	Nightlife spot	Event	College & university	Arts & enterteinment
26	Rosedale	Outdoors & recreation	Travel & transport	Shop & service	Residence	Professional & other places	Nightlife spot	Food	Event	College & university	Arts & enterteinment
27	Roselawn	Shop & service	Outdoors & recreation	Food	Travel & transport	Residence	Professional & other places	Nightlife spot	Event	College & university	Arts & enterteinment
33	Summerhill West, Rathnelly, South Hill, Forest...	Food	Nightlife spot	Travel & transport	Shop & service	Professional & other places	Residence	Outdoors & recreation	Event	College & university	Arts & enterteinment

Table 13. Cluster 2 (blue): Northern zone, residence-recreational area.



	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
2	Business reply mail Processing Centre, South C...	Outdoors & recreation	Food	Shop & service	Professional & other places	Travel & transport	Nightlife spot	Residence	Event	College & university	Arts & enterteinment
16	India Bazaar, The Beaches West	Food	Shop & service	Professional & other places	Nightlife spot	Outdoors & recreation	Arts & enterteinment	Travel & transport	Residence	Event	College & university
35	The Beaches	Outdoors & recreation	Shop & service	Nightlife spot	Travel & transport	Residence	Professional & other places	Food	Event	College & university	Arts & enterteinment

Table 14. Cluster 3 (green): Eastern zone, recreational area.



	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
1	Brockton, Parkdale Village, Exhibition Place	Food	Shop & service	Professional & other places	Nightlife spot	Outdoors & recreation	Arts & enterteinment	Travel & transport	Residence	Event	College & university
3	CN Tower, King and Spadina, Railway Lands, Har...	Travel & transport	Food	Shop & service	Outdoors & recreation	Nightlife spot	Residence	Professional & other places	Event	College & university	Arts & enterteinment
5	Christie	Food	Shop & service	Outdoors & recreation	Nightlife spot	Travel & transport	Residence	Professional & other places	Event	College & university	Arts & enterteinment
10	Dufferin, Dovercourt Village	Shop & service	Nightlife spot	Food	Professional & other places	Outdoors & recreation	Arts & enterteinment	Travel & transport	Residence	Event	College & university
15	High Park, The Junction South	Food	Shop & service	Outdoors & recreation	Nightlife spot	Arts & enterteinment	Travel & transport	Residence	Professional & other places	Event	College & university
22	Parkdale, Roncesvalles	Food	Shop & service	Outdoors & recreation	Nightlife spot	Arts & enterteinment	Travel & transport	Residence	Professional & other places	Event	College & university
28	Runnymede, Swansea	Food	Shop & service	Professional & other places	Nightlife spot	Arts & enterteinment	Travel & transport	Residence	Outdoors & recreation	Event	College & university
34	The Annex, North Midtown, Yorkville	Food	Shop & service	Outdoors & recreation	Nightlife spot	Arts & enterteinment	Travel & transport	Residence	Professional & other places	Event	College & university

Table 15. Cluster 4 (orange): Western zone, residential area.

DISCUSSION

The addition of the location data to the Toronto_grouped dataframe resulted in a much better clustering, as the Kmeans algorithm received both the location data and the venue information, however, the effect of the location data needs to be observed and controlled by a scale factor, so that it does not bias the results of the Kmeans algorithm.

By processing the results of the one-hot-coding technique using addition, instead of averaging, it was possible to preserve the number of venues in each neighborhood, and pass the information to the Kmeans algorithm, for better discrimination of the neighborhoods.

The gathering of the venue information into 10 main categories also resulted in more meaningful results of the algorithm, as the Kmeans algorithm does not have the capacity of associating all the different venues that belong to the same main category.

The calculation of the distance between neighborhoods revealed a large range of distances. This measure reveals that the radius parameter should be selected carefully.

CONCLUSION

The location data affected in a positive way the results provided by the Kmeans algorithm. Since the results of the Kmeans algorithm are plotted in the Toronto map, one would expect that the clusters appear in single areas in the map, which may or may not happen if the location data is not passed to the Kmeans algorithm. By passing the location data to the Kmeans algorithm, one can control the grouping of the neighborhoods in single regions (see figure 9).

The one-hot-coding technique proved to be a powerful tool for processing venue information, however, the processing of the results need to be done carefully to preserve the venue information. As a result, a modification was made to take into account the number of venues in each neighborhood (which I called the density), which proved to positively influence the results of the Kmeans algorithm, as seen in figure 7, where the clusters appear centered in the downtown area.

A new way to preserve venue information was found by regrouping the 237 venue categories into 10 main venue categories, this way, similar venues share the same category, and Kmeans can make a more meaningful clustering as seen in figure 9. This last statement can be confirmed by looking at tables 11 to 15, which show the most common venues of each cluster. This is most clearly seen in cluster 3 (in green in figure 9), where Kmeans was able to detect a cluster of neighborhoods that make a recreational area. In summary, the neighborhoods could be classified as follows (see figure 9):

1. Cluster 0 (red): Central zone, shopping-recreational area.
2. Cluster 1 (violet): Downtown, business-recreational area.
3. Cluster 2 (blue): Northern zone, residence-recreational area.
4. Cluster 3 (green): Eastern zone, recreational area.
5. Cluster 4 (orange): Western zone, residential area.

The distance between neighborhoods is an important measure that needs to be taken into account before deciding the radius for venue exploration, as there is a great variability in this parameter. The minimum value found was 150 m, and the average value was 4,94 Km.

As a result of this capstone project, several modifications were made to the week 3 assignment. These modifications resulted in a new, more meaningful clustering of the Toronto neighborhoods. These ideas implemented in this project are general, and can be applied to the analysis of any other cities or places.