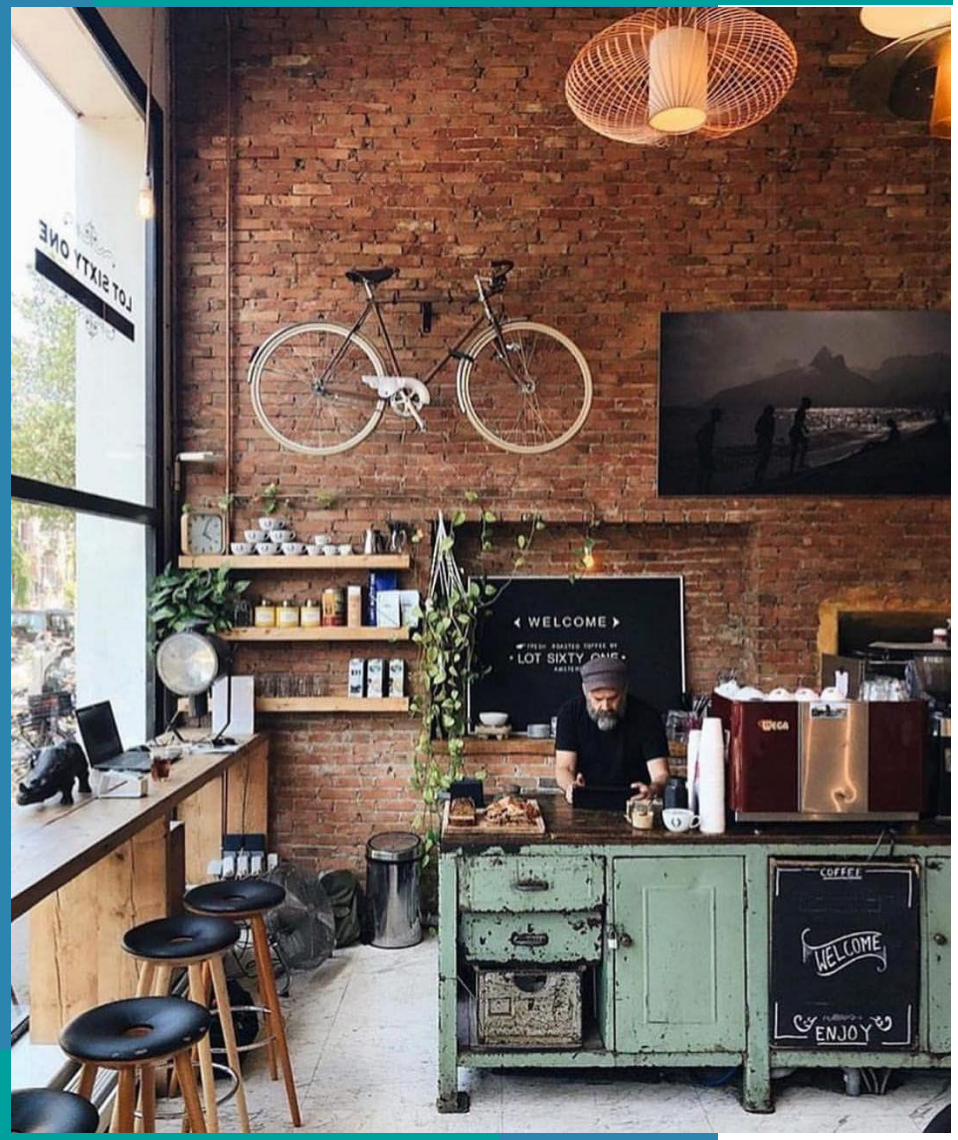
The Battle of Neighborhoods

Capstone Project:

Coffee Shop in Toronto

Applied Data Science Capstone

IBM Data Science Professional Certificate on Coursera



Background and Problem

- Toronto is the provincial capital of Ontario. With a recorded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America
- Toronto is a prominent center for music, theatre, motion picture production, and television production, and is home to the headquarters of Canada's major national broadcast networks and media outlets. Its varied cultural institutions, which include numerous museums and galleries, festivals and public events, entertainment districts, national historic sites, and sports activities, attract over 43 million tourists each year.
- The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a Coffee Shop. This city is home to many coffee shops where coffee addicts find their favorite drink.

Data acquisition and cleaning

- For the Toronto neighborhood data, a Wikipedia page exists that has all the information you need to explore and cluster the neighborhoods in Toronto.

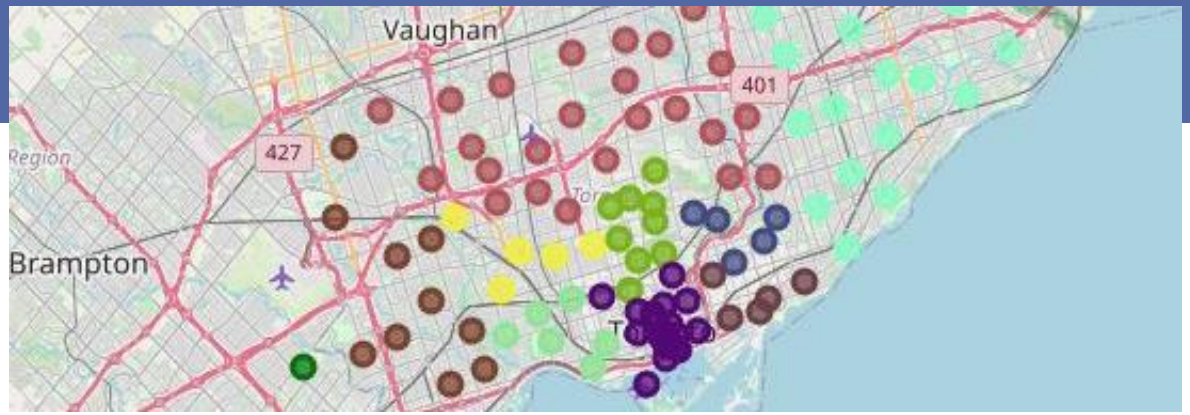
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Geocoder Python package.

https://geocoder.readthedocs.io/index.html

- Toronto city geographical coordinates data will be utilized as input for the Foursquare API that will be leveraged to provision venues information for each neighborhood.

Analytic Approach



- First, it is created a map using Folium color coded each Neighborhood depending on borough it is located.
- Next, we used the Foursquare API to get a list of all the Venues in Toronto which included Spa, Bus Line, Coffee Shops, Italian Restaurants, etc.
- Then to analyze the data you perform a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms.

There is a total of 177 Coffee Shops in Toronto.

This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.

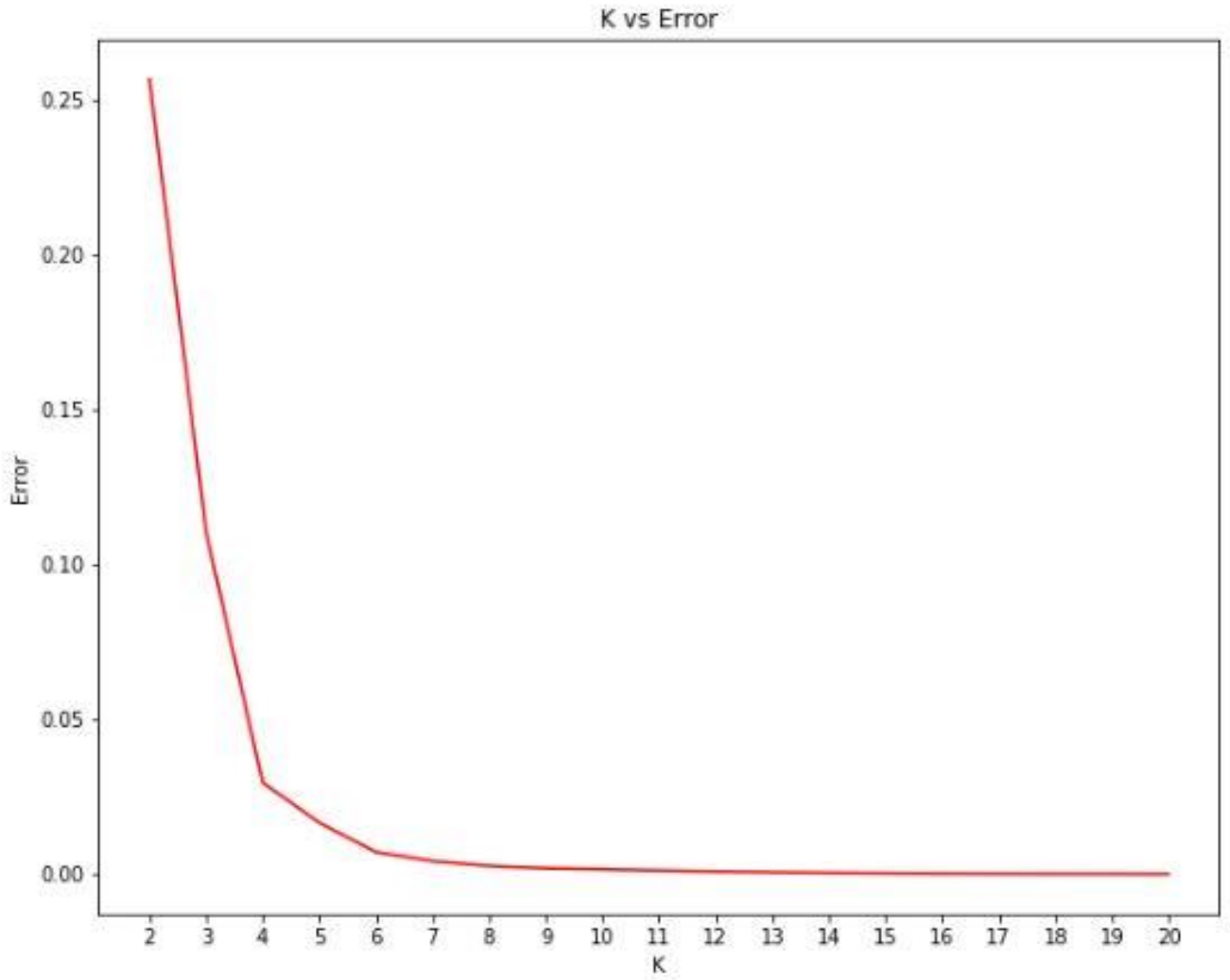
	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wine Shop	Wings Joint	Wome St
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

Then you group those rows by Neighborhood and by taking the Average of the frequency of occurrence of each Venue Category.

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Vegetarian / Vegan Restaurant	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Wine Bar	Wi Sh
0	Agincourt	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
1	Alderwood, Long Branch	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	...	0.000000	0.000000	0.000	0.000000	0.000000	0.000000	0.00

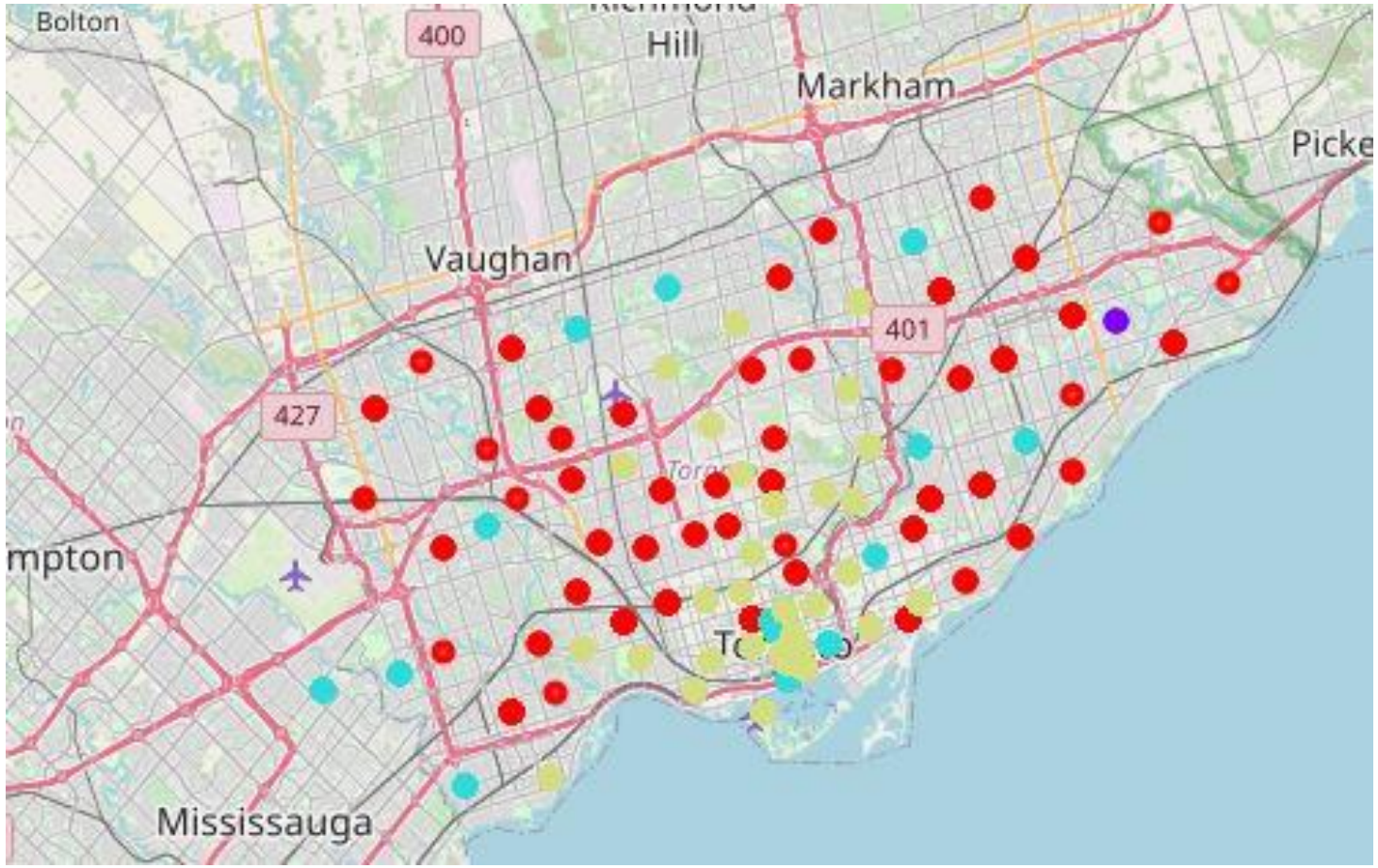
K-means Clustering

- To get our optimum K value that was neither overfitting or underfitting the model, we used the Elbow Point Technique. The best K value is chosen at the point in which the line has a sharpest turn. In our case we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.



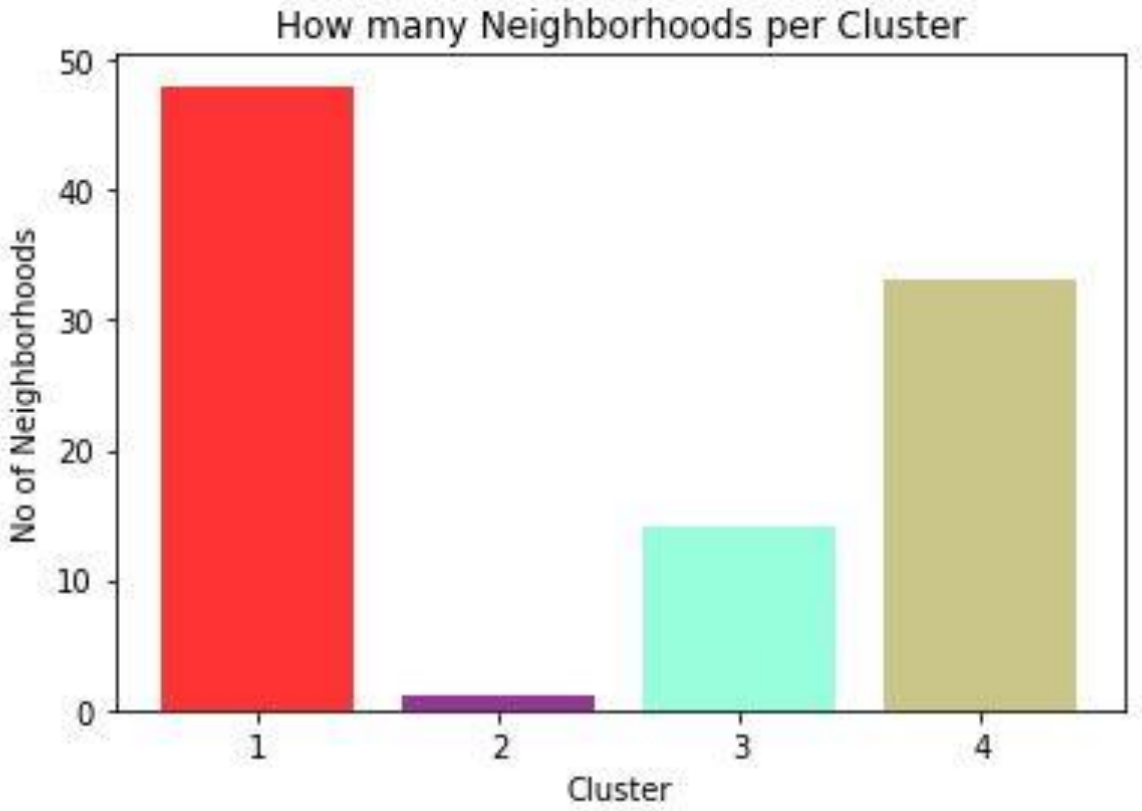
Results

- After, it is merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a new Coffee Shop in Toronto. Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.



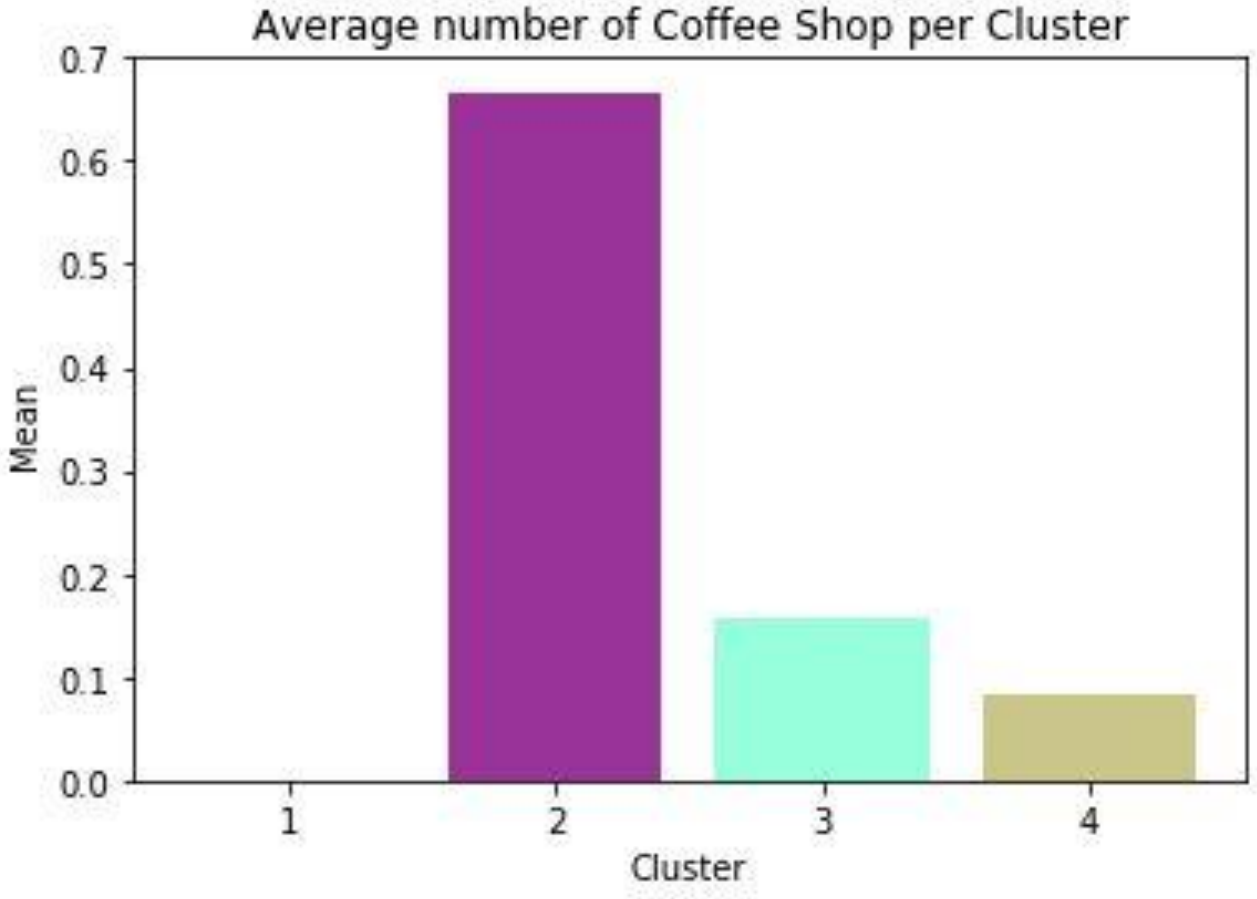
Neighborhoods per Cluster

- We have a total of 4 clusters (0,1,2,3). Before it has analyzed them one by one lets check the total amount of neighborhoods in each cluster and the average Coffee Shops in that cluster.
- We see that Cluster 1 has the most neighborhoods (48) while cluster 2 has the least (1). Cluster 4 has 33 neighborhoods and cluster 3 has only 14.



Average number of Coffee Shop per Cluster

- This information is crucial as we can see that even through there is only 1 neighborhood in Cluster 2, it has the highest number of Coffee Shops (0.66) while Cluster 1 has the most neighborhoods but has the least average of Coffee Shops (0.0).



Discussion and Conclusion

- Most of the Coffee Shops are in cluster 4 represented by the turquoise clusters. Even though there is a huge number of Neighborhoods in cluster 1, there is little to no Coffee Shops. We see that in the cluster 3 has the second last average of Coffee Shops but it is not one of the most populate cluster.
- Looking at the nearby venues, the optimum place to put a new Coffee Shop is in East York as there are many Neighborhoods in the area but little to no Coffee Shops therefore, eliminating any competition.