Establishing a new Italian restaurant in the city of Toronto

1. Introduction

Let's assume that our problem is to find a suitable area to open a new successful Italian restaurant in Toronto.

To succeed in our mission, there are several factors to considered it, such as, population base, accessibility, local employment, and one of the most important is to choose a suitable location where the concurrence is not too intense.

For our business, one of the key factors to gain more customers is to establish in the city centre of Toronto, but as I previously mentioned, the competition has to be low or non-existent.

If we cannot find a suitable area, that meets our requirement, the research has to propose an alternative location that will be unique.

In this work, we are designing an unsupervised recommendation system based on a number of clusters. It aims to partition a number of observations into k clusters in which each observation belongs to the cluster with the nearest mean of frequencies of occurrence in each restaurant category (Cuisine).

2. Data

The Initial data that we are using is provided by Wikipedia. This data consists of a set of information about Toronto's city: Postal Code, Borough, Neighbourhood.

To read our HTML file we are using a module called BeautifulSoup, that allows us to scrape files from HTML (our case) or XML.

When we look up the data, the first things that we notice is that some data is missing, so we are grouping each Postal code with all neighbourhoods.

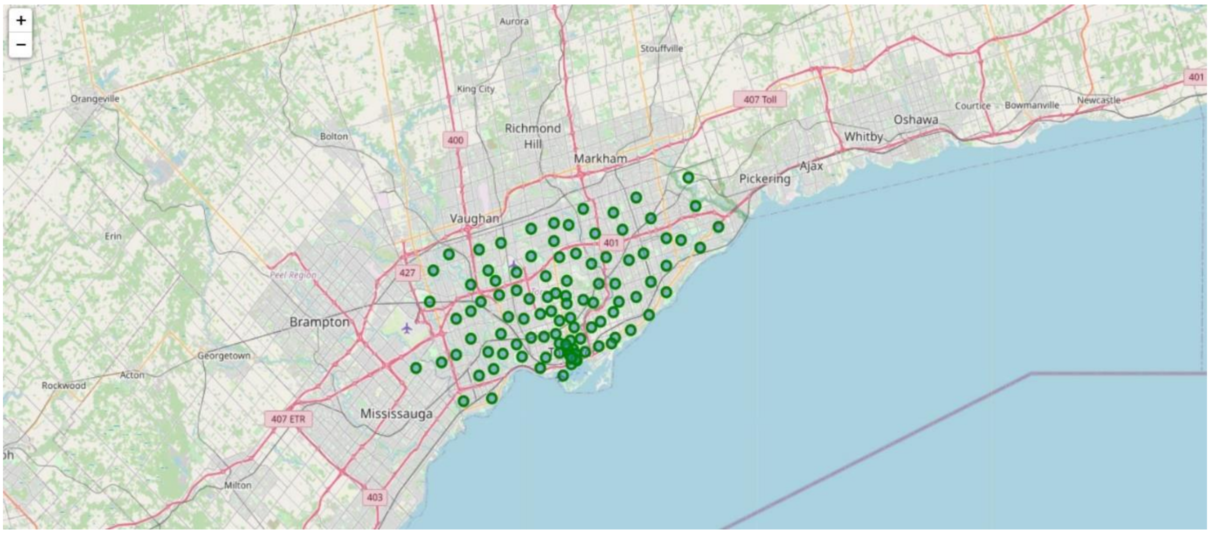
	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Now, that our database is ready, we load a CVS file into our project that contains Geo-spatial coordinates (Latitude, Longitude, Postal Code) and we stored our result into a Pandas module data frame.

To get a new data frame with each postal code and neighbourhood and the latitude and longitude coordinates we going to perform a left joint into our pre-existing database.

	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
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	Kennedy Park, Ionview, East Birchmount Park	43.727929	-79.262029
7	M1L	Scarborough	Golden Mile, Clairlea, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848

Then, using Folium library we can plot our results onto a map:



Now that we have plotted our data frame onto our map we can move forward and use Foursquare API to explore the neighbourhoods and segment them.

We can get the top 100 venues within a radius of 500 meters for each postal code area, the result of our call request will be to retrieve a JSON file object, that we will insert into a Pandas data frame and merge it into one data frame, that will include neighbourhoods and their top rated venues.

As we can see, the top 100 venues are not just food businesses

	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
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
5	Guildwood, Morningside, West Hill	43.763573	-79.188711	Sail Sushi	43.765951	-79.191275	Restaurant
6	Guildwood, Morningside, West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant
7	Guildwood, Morningside, West Hill	43.763573	-79.188711	Enterprise Rent-A-Car	43.764076	-79.193406	Rental Car Location
8	Guildwood, Morningside, West Hill	43.763573	-79.188711	Woburn Medical Centre	43.766631	-79.192286	Medical Center
9	Guildwood, Morningside, West Hill	43.763573	-79.188711	Lawrence Ave E & Kingston Rd	43.767704	-79.189490	Intersection
10	Guildwood, Morningside, West Hill	43.763573	-79.188711	Eggsmart	43.767800	-79.190466	Breakfast Spot
11	Woburn	43.770992	-79.216917	Starbucks	43.770037	-79.221156	Coffee Shop
12	Woburn	43.770992	-79.216917	Tim Hortons	43.770827	-79.223078	Coffee Shop
13	Woburn	43.770992	-79.216917	Korean Grill House	43.770812	-79.214502	Korean BBQ Restaurant
14	Woburn	43.770992	-79.216917	El rey del cabrito, monterrey city mexico	43.768800	-79.219800	Mexican Restaurant
15	Cedarbrae	43.773136	-79.239476	Drupati's Roti & Doubles	43.775222	-79.241678	Caribbean Restaurant
16	Cedarbrae	43.773136	-79.239476	Federick Restaurant	43.774697	-79.241142	Hakka Restaurant
17	Cedarbrae	43.773136	-79.239476	Thai One On	43.774468	-79.241268	Thai Restaurant
18	Cedarbrae	43.773136	-79.239476	Centennial Recreation Centre	43.774593	-79.236500	Athletics & Sports
19	Cedarbrae	43.773136	-79.239476	TD Canada Trust	43.774830	-79.241251	Bank
20	Cedarbrae	43.773136	-79.239476	Petro-Canada	43.774106	-79.243097	Gas Station

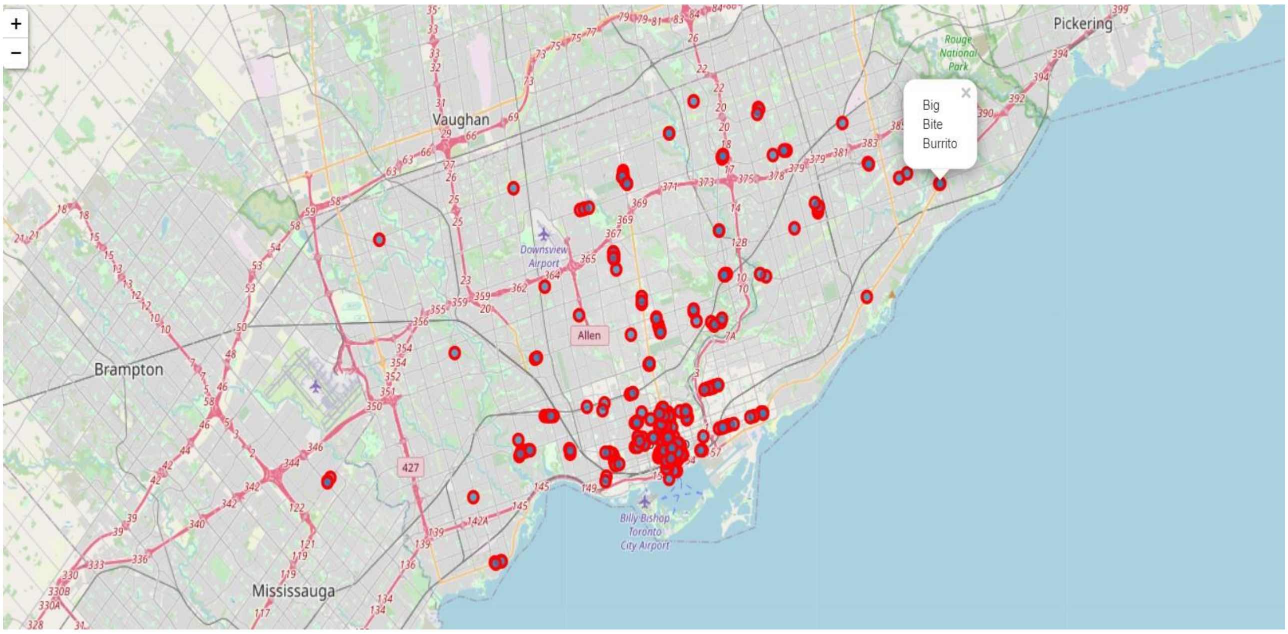
Because our main goal is to recommend a neighbourhood for a new Italian restaurant, we need to filter our data, so we will have only venues categories that contain the word 'Restaurant'.

	index	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	5	Guildwood, Morningside, West Hill	43.763573	-79.188711	Sail Sushi	43.765951	-79.191275	Restaurant
2	6	Guildwood, Morningside, West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant
3	13	Woburn	43.770992	-79.216917	Korean Grill House	43.770812	-79.214502	Korean BBQ Restaurant
4	14	Woburn	43.770992	-79.216917	El rey del cabrito, monterrey city mexico	43.768800	-79.219800	Mexican Restaurant
5	15	Cedarbrae	43.773136	-79.239476	Drupati's Roti & Doubles	43.775222	-79.241678	Caribbean Restaurant
6	16	Cedarbrae	43.773136	-79.239476	Federick Restaurant	43.774697	-79.241142	Hakka Restaurant
7	17	Cedarbrae	43.773136	-79.239476	Thai One On	43.774468	-79.241268	Thai Restaurant
8	41	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476	Vincent's Spot	43.717002	-79.242353	American Restaurant
9	46	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304	Kim Kim restaurant	43.753833	-79.276611	Chinese Restaurant

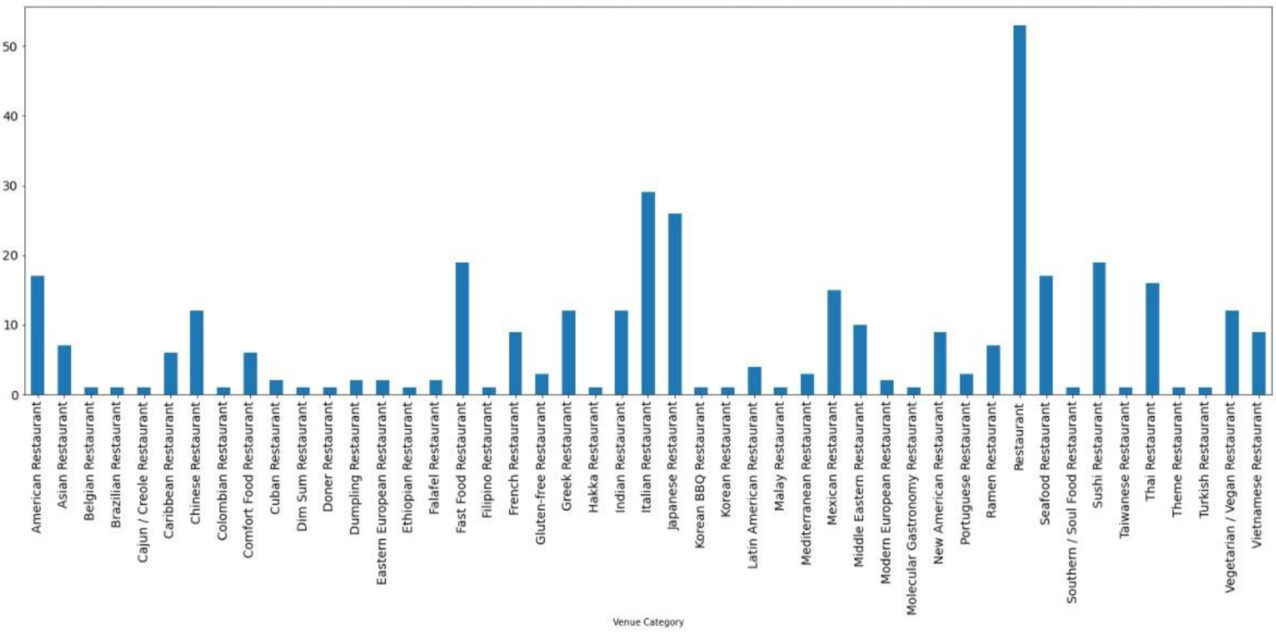
3. Methodology

From the dataset, we have filtered out the neighbourhoods that contain venues that include the word 'Restaurant'. We can see, after filtering out the data, that our dataset shape is (514, 8) which means that the data contains 514 rows that are 'Restaurant' venues and 8 columns, as we can see above.

Here, there is a map of all the top-rated restaurants in Toronto:



And here, we can find the distribution of all cuisines:



4. Results

We can see from our histogram bar chart, that the most frequent restaurant cuisine is represented by the generic category 'restaurant', followed by 'Italian restaurant'.

For each neighbourhood we can extract the top 5 frequencies of restaurant cuisines, as shown in the image below:

```
----Bathurst Manor, Wilson Heights, Downsview North----
                      venue  freq
0        Chinese Restaurant  0.25
1          Sushi Restaurant  0.25
2  Middle Eastern Restaurant  0.25
3                Restaurant  0.25
4       American Restaurant  0.00


----Bayview Village----
                    venue  freq
0     Chinese Restaurant   0.5
1    Japanese Restaurant   0.5
2    American Restaurant   0.0
3  Portuguese Restaurant   0.0
4      Korean Restaurant   0.0


----Bedford Park, Lawrence Manor East----
                     venue  freq
0       Italian Restaurant  0.22
1  Comfort Food Restaurant  0.11
2          Thai Restaurant  0.11
3         Sushi Restaurant  0.11
4               Restaurant  0.11
```

As it shows in our results, in Bedford Park the top cuisine is represented by "Italian restaurant" with a mean frequency of 0.22, followed by "Comfort food restaurant" with a 0.11.

Let's continue to use Pandas library to refine our findings:

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Agincourt	Latin American Restaurant	Vietnamese Restaurant	Indian Restaurant	Greek Restaurant	Gluten-free Restaurant
1	Bathurst Manor, Wilson Heights, Downsview North	Sushi Restaurant	Chinese Restaurant	Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant
2	Bayview Village	Japanese Restaurant	Chinese Restaurant	Vietnamese Restaurant	Dumpling Restaurant	Greek Restaurant
3	Bedford Park, Lawrence Manor East	Italian Restaurant	Sushi Restaurant	Comfort Food Restaurant	Greek Restaurant	Indian Restaurant
4	Berczy Park	Seafood Restaurant	Restaurant	Italian Restaurant	French Restaurant	Vegetarian / Vegan Restaurant
5	Brockton, Parkdale Village, Exhibition Place	Italian Restaurant	Restaurant	Doner Restaurant	Greek Restaurant	Gluten-free Restaurant
6	Business reply mail Processing Centre, South C...	Fast Food Restaurant	Restaurant	Vietnamese Restaurant	Doner Restaurant	Greek Restaurant
7	Canada Post Gateway Processing Centre	American Restaurant	Mediterranean Restaurant	Middle Eastern Restaurant	Hakka Restaurant	Gluten-free Restaurant
8	Cedarbrae	Hakka Restaurant	Thai Restaurant	Caribbean Restaurant	Doner Restaurant	Greek Restaurant
9	Central Bay Street	Italian Restaurant	Ramen Restaurant	Falafel Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant

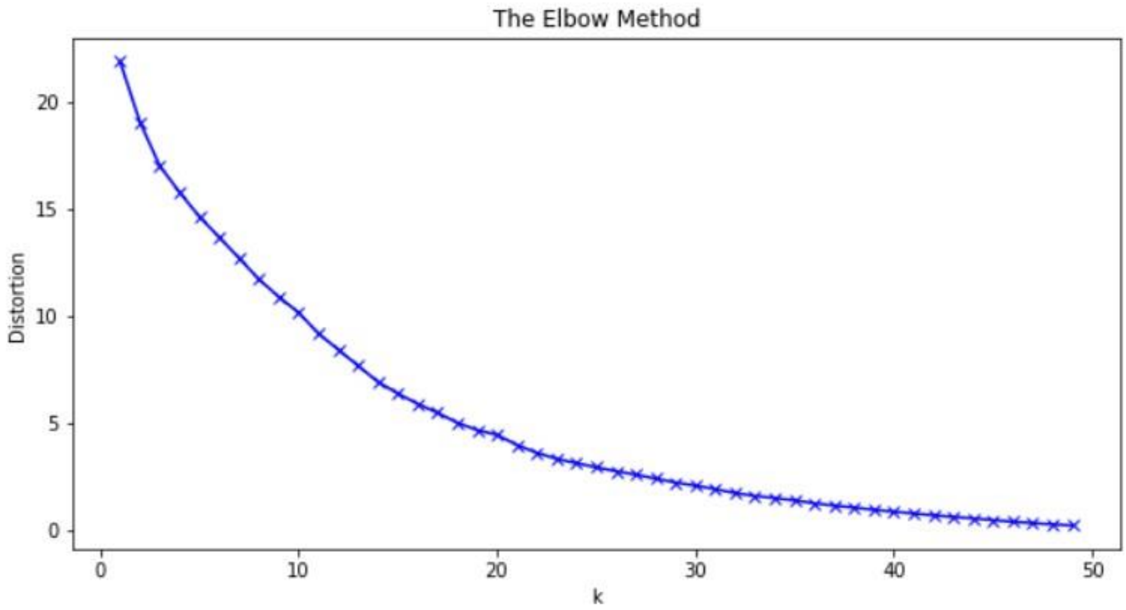
4.1 Clustering

Clustering is a technique for finding subgroups of observations within a data set. When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn't a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the n observations without being trained by a response variable.

Clustering allows us to identify which observations are alike, and potentially categorize them.

K-means clustering is the most commonly used unsupervised machine learning algorithm for partitioning a given data set into a set of k groups, where k represents the number of groups pre-specified by the analyst.

Before assigning the number of clusters let's perform the elbow method to see which is an ideal number to work with:



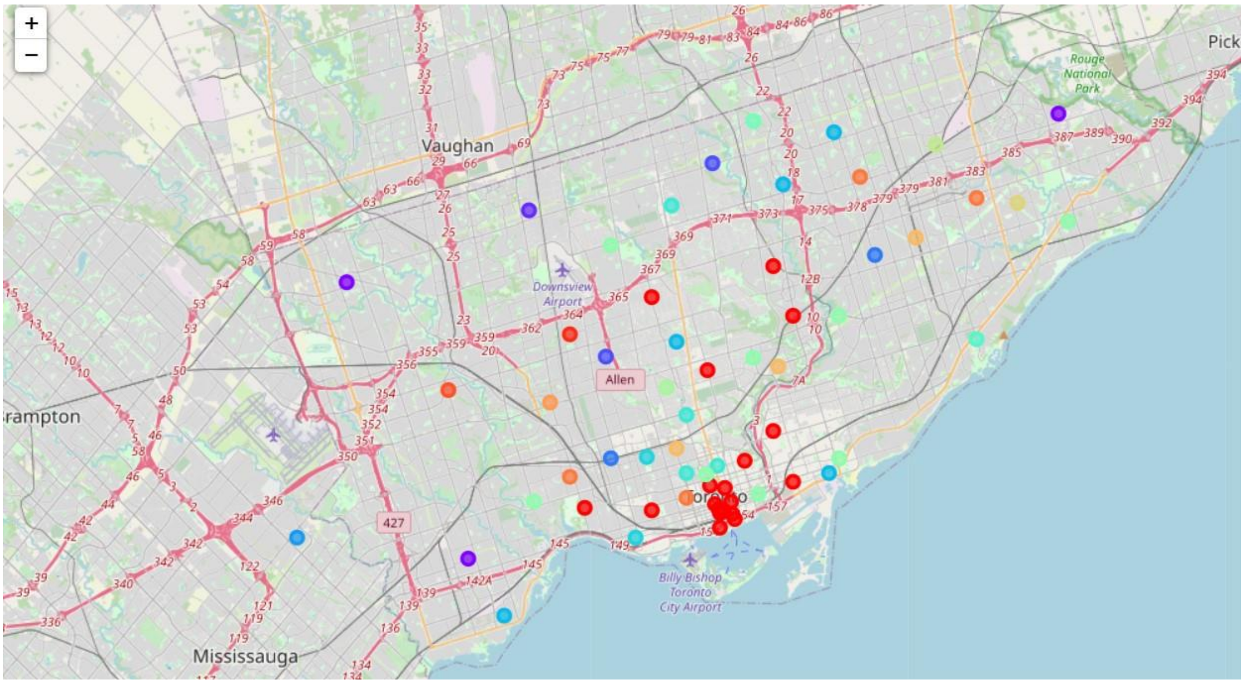
As shown in the graphic above, the perfect number of clusters is situated where there is the picking of the curve, which in our case is situated at 20. This will be the number of clusters that we going to use.

After applying our clustering, we merged the result with the previous data frame, and we can see that our neighbourhood falls in a specific cluster:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster_Labels
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	Fast Food Restaurant	Vietnamese Restaurant	Doner Restaurant	Greek Restaurant	Gluten-free Restaurant	1
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	Mexican Restaurant	Restaurant	Vietnamese Restaurant	Doner Restaurant	Gluten-free Restaurant	11
3	M1G	Scarborough	Woburn	43.770992	-79.216917	Korean BBQ Restaurant	Mexican Restaurant	Vietnamese Restaurant	Hakka Restaurant	Gluten-free Restaurant	14
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	Hakka Restaurant	Thai Restaurant	Caribbean Restaurant	Doner Restaurant	Greek Restaurant	17
8	M1M	Scarborough	Cliffside, Cliffcrest, Scarborough Village West	43.716316	-79.239476	American Restaurant	Doner Restaurant	Greek Restaurant	Gluten-free Restaurant	French Restaurant	9
10	M1P	Scarborough	Dorset Park, Wexford Heights, Scarborough Town...	43.757410	-79.273304	Indian Restaurant	Vietnamese Restaurant	Chinese Restaurant	Dumpling Restaurant	Greek Restaurant	15
11	M1R	Scarborough	Wexford, Maryvale	43.750072	-79.295849	Middle Eastern Restaurant	Vietnamese Restaurant	Indian Restaurant	Greek Restaurant	Gluten-free Restaurant	4

5. Discussion

we can plot our result onto a map, to see the distribution of the clusters:

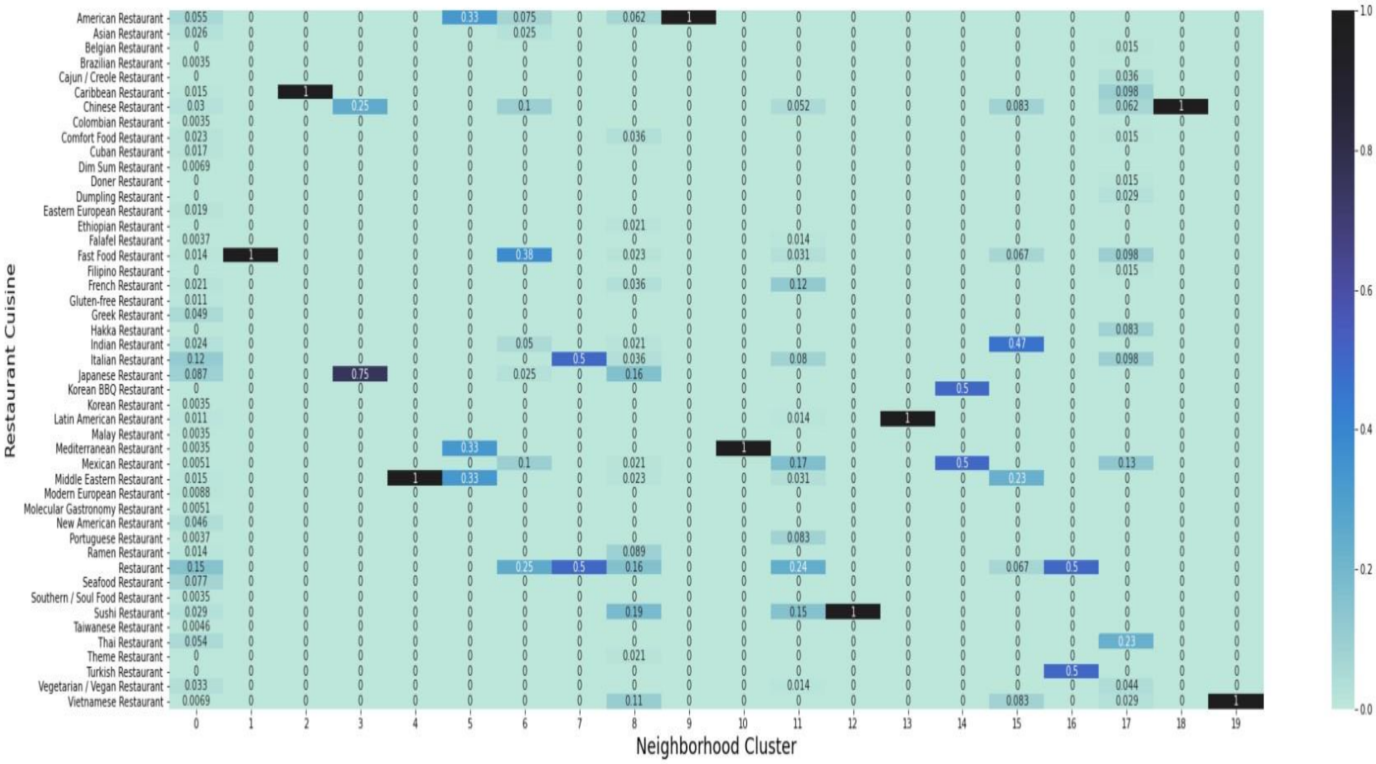


Easily we can see that cluster 0 (red) is the crowded one, that means that all neighbourhoods in cluster 0 are similar with top rated restaurants cuisines.

Now, after I performed one hot encoding on the data to get a binary representation of data set and then grouped rows by clusters and by taking the mean of the frequency of each cuisine:

Cluster_Labels	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
American Restaurant	0.054898	0.0	0.0	0.00	0.0	0.333333	0.075	0.0	0.062500	1.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Asian Restaurant	0.025603	0.0	0.0	0.00	0.0	0.000000	0.025	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Belgian Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.014706	0.0	0.0
Brazilian Restaurant	0.003472	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Cajun / Creole Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.035714	0.0	0.0
Caribbean Restaurant	0.015046	0.0	1.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.098039	0.0	0.0
Chinese Restaurant	0.030093	0.0	0.0	0.25	0.0	0.000000	0.100	0.0	0.000000	0.0	0.0	0.052083	0.0	0.0	0.0	0.083333	0.0	0.062500	1.0	0.0
Colombian Restaurant	0.003472	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Comfort Food Restaurant	0.022797	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.035714	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.014706	0.0	0.0
Cuban Restaurant	0.017361	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Dim Sum Restaurant	0.006944	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Doner Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.014706	0.0	0.0
Dumpling Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.029412	0.0	0.0
Eastern European Restaurant	0.018519	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Ethiopian Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.020833	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Falafel Restaurant	0.003704	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.013889	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Fast Food Restaurant	0.013573	1.0	0.0	0.00	0.0	0.000000	0.375	0.0	0.022727	0.0	0.0	0.031250	0.0	0.0	0.0	0.066667	0.0	0.098214	0.0	0.0
Filipino Restaurant	0.000000	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.014706	0.0	0.0
French Restaurant	0.020824	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.035714	0.0	0.0	0.118056	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Gluten-free Restaurant	0.011409	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0
Greek Restaurant	0.048721	0.0	0.0	0.00	0.0	0.000000	0.000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0

This allowed us to examine our findings through an heatmap:



6. Conclusion section

The conclusion of this report can be taken from the previous heatmap.

The darker the color, the higher concentration of a single restaurant cuisine in the corresponding neighbourhood cluster.

Now, we can deduce from our findings, that if we want to open an Italian restaurant in Toronto, it is not advisable to open in cluster number 7, because of the high concentration of Italian Restaurants.