Capstone Project - Report

"The Battle of Neighborhoods"

in Old Toronto

for the Applied Data Science course (IBM)

by Coursera

Introduction, The business problem

The current study will try to provide an optimum location for opening a new business in a specific city, based on:

- A neighborhoods property, e.g. the second most common language spoken (after English) in the neighborhood,
- The number of competitors in the neighborhood,
- The population density in the neighborhood.
- The average income of the neighborhood.

Let's investigate in the ***old City of Toronto*** (Canada), and propose the best possible place for opening ***a new restaurant with ethnic cuisine***. Assume that preferably we would like the new restaurant to be located in a neighborhood with a high degree of the same ethnic characteristics, i.e. assume the languages spoken in that neighborhood, so to make advantage of the cultural element of the area. In order to sustain the new business, there should be a lot of population, the less number of competitors possible. The restaurant should be of middle class and above.

The results could be highly usable ***for people having ethnic cooking skills or restaurant-businessmen, who want to open an ethnic restaurant in a neighborhood having some degree of the same ethnic culture in Old Toronto.***

2. Data description

The datasets that will be used are for retrieving information about city's neighborhoods and their characteristics are taken from Wikipedia website for the Toronto demographic information:

- https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

The geographic coordinates of the neighborhoods of Old Toronto are taken from:

- File 'oldToronto.csv' which the latitude and longitude per neighborhood in Old Toronto area.

Combining the above data sets, we get ***demographic information, focused in Old Toronto's neighborhoods*** and the exact coordinates per neighborhood. Then by using the Foursquare API, we can retrieve further information for venues, venue categories and venue coordinates for every area. ***The Foursquare data set combined with the neighborhood's data set with demographic information will be the main data set*** that we will be used for the analysis. Visualization of the results via maps and graphs, where possible, will help to explain the data.

Based on the language spoken (second language spoken after 'English'), the neighborhood's population, the level of wealth, and the number of ethnic restaurants (restaurants with ethnicity common with the language spoken) *the best possible set of candidate neighborhoods can be retrieved*. Then by using ***k-means algorithm*** the candidate neighborhoods will be further analyzed. The final results, via tables and maps will conclude on finding ***the best neighborhood to start an ethnic restaurant in an ethnic-cultural neighborhood,*** show any existing patterns and similarities between ethnic restaurants and ethnic populated neighborhoods in Old Toronto area.

3. Methodology

1. Data wrangling

For first step, the information in the Wikipedia link has to be transformed in a suitable form that enables further dataframe analysis. The link that provides the demographic data is the following:

'https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods'

By using 'BeautifulSoup' we retrieve the json data and fetch the wanted tags. We then clear the data via regular expressions for unwanted characters (e.g. remove '\n', empty spaces, etc.), remove non-meaningful data, rename index and columns and the dataframe with the demographic data is as shown below (df : demographic.head(10)):

	Neighborhood	Population	Density	Average income	Second language after English, %	Second language after English, name	Second language population
0	Agincourt	44577	3580	25750	19.3	Cantonese	8603
1	Alderwood	11656	2360	35239	6.2	Polish	722
2	Alexandra Park	4355	13609	19687	17.9	Cantonese	779
3	Allenby	2513	4333	245592	1.4	Russian	35
4	Amesbury	17318	4934	27546	6.1	Spanish	1056
5	Armour Heights	4384	1914	116651	9.4	Russian	412
6	Banbury	6641	2442	92319	5.1	Chinese	338
7	Bathurst Manor	14945	3187	34169	9.5	Russian	1419
8	Bay Street Corridor	4787	43518	40598	9.6	Mandarin	459
9	Bayview Village	12280	2966	46752	8.4	Cantonese	1031

Each neighborhood is depicted via its population, density, average income, the most common language after English spoken in the area (assume it as named as 'language' from now on), the name of the language, the population speaking that language (assume it as 'ethnic population')

2. Add coordinates in the demographic data

It is possible to add latitude and longitude data, by merging the demographic dataframe and the neighborhoods of Old Toronto only. File 'oldToronto.csv' contains the coordinates of Old Toronto. The updated dataframe is as shown below (df : Toronto.head(10)):

	Neighborhood	Population	Density	Average income	Percentage	Language	Second language population	Latitude	Longitude
0	Alexandra Park	4355	13609	19687	17.9	Cantonese	779	43.71627	-79.40555
1	Allenby	2513	4333	245592	1.4	Russian	35	43.71275	-79.54746
2	Bay Street Corridor	4787	43518	40598	9.6	Mandarin	459	43.65777	-79.38619
3	Bedford Park	13749	6057	80827	0.7	Greek	96	43.73138	-79.42116
4	Bloor West Village	5175	6993	55578	3.6	Ukrainian	186	43.65936	-79.48543
5	Bracondale Hill	5343	8618	41605	4.8	Greek	256	43.67600	-79.42803
6	Brockton	9039	8217	27260	19.9	Portuguese	1798	43.66055	-79.40531
7	Cabbagetown	11120	7943	50398	1.6	Chinese	177	43.66763	-79.36606
8	Carleton Village	6544	8843	23301	17.0	Portuguese	1112	43.67200	-79.45700
9	Casa Loma	3597	5369	82203	1.8	Korean	64	43.67000	-79.41000

We can visualize the map of Toronto, with the neighborhoods, the language used (after English) – which depicts the ethnic group, the percentage of the ethnic group. Different color is used for each language (e.g. 'pink' is used for 'Portuguese'). (map: map_Toronto_neighborhoods)

3. Foursquare API

Now that we have the demographic information per neighborhood in Old Toronto, lets collect all venues within 1km radius from the center of each neighborhood (limit to 100 venues) and store the results in a dataframe (df: Toronto_venues.head(10)):

```
0 Alexandra Park
1 Allenby
2 Bay Street Corridor
```

```
3 Bedford Park
4 Bloor West Village
5 Bracondale Hill
6 Brockton
7 Cabbagetown
8 Carleton Village
9 Casa Loma
10 Chaplin Estates
11 Christie Pits
```

…

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alexandra Park	43.71627	-79.40555	Sheridan Nurseries	43.719005	-79.400500	Flower Shop
1	Alexandra Park	43.71627	-79.40555	Himalayan Java	43.713486	-79.399811	Café
2	Alexandra Park	43.71627	-79.40555	De Mello Palheta Coffee Roasters	43.711791	-79.399403	Coffee Shop
3	Alexandra Park	43.71627	-79.40555	Barreworks	43.714070	-79.400109	Yoga Studio
4	Alexandra Park	43.71627	-79.40555	Starbucks	43.711200	-79.399182	Coffee Shop
5	Alexandra Park	43.71627	-79.40555	Uncle Betty's Diner	43.714452	-79.400091	Diner
6	Alexandra Park	43.71627	-79.40555	Douce France	43.711534	-79.399255	Bakery
7	Alexandra Park	43.71627	-79.40555	Alexander Muir Memorial Gardens	43.721315	-79.400822	Garden
8	Alexandra Park	43.71627	-79.40555	Cibo Wine Bar	43.711464	-79.399570	Italian Restaurant
9	Alexandra Park	43.71627	-79.40555	Sign of the Skier	43.719395	-79.401234	Sporting Goods Shop

We only interested in 'Restaurants' so we filter the venue category by this type (df: Toronto_restaurants_coord.head(10)):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alexandra Park	43.71627	-79.40555	Cibo Wine Bar	43.711464	-79.399570	Italian Restaurant
1	Alexandra Park	43.71627	-79.40555	La Vecchia Ristorante	43.710167	-79.399086	Italian Restaurant
2	Alexandra Park	43.71627	-79.40555	Sushi Shop	43.713609	-79.399844	Sushi Restaurant
3	Alexandra Park	43.71627	-79.40555	Grazie Ristorante	43.709329	-79.398823	Italian Restaurant
4	Alexandra Park	43.71627	-79.40555	Tio's Urban Mexican	43.714630	-79.400000	Mexican Restaurant
5	Alexandra Park	43.71627	-79.40555	C'est Bon	43.716785	-79.400406	Chinese Restaurant
6	Alexandra Park	43.71627	-79.40555	Sorn Thai Restaurant	43.713425	-79.399799	Thai Restaurant
7	Alexandra Park	43.71627	-79.40555	Banh Mi Boys	43.709217	-79.398777	Fast Food Restaurant
8	Alexandra Park	43.71627	-79.40555	Touhenboku Ramen 唐変木	43.711425	-79.399278	Ramen Restaurant
9	Alexandra Park	43.71627	-79.40555	Sushi Rock Café	43.709089	-79.398641	Sushi Restaurant

Then we calculate the sum of ethnic restaurants per neighborhood (note ethnic assumed the ethnic group speaking the second most common language after English in the area, (Toronto_restaurants.head(10))).

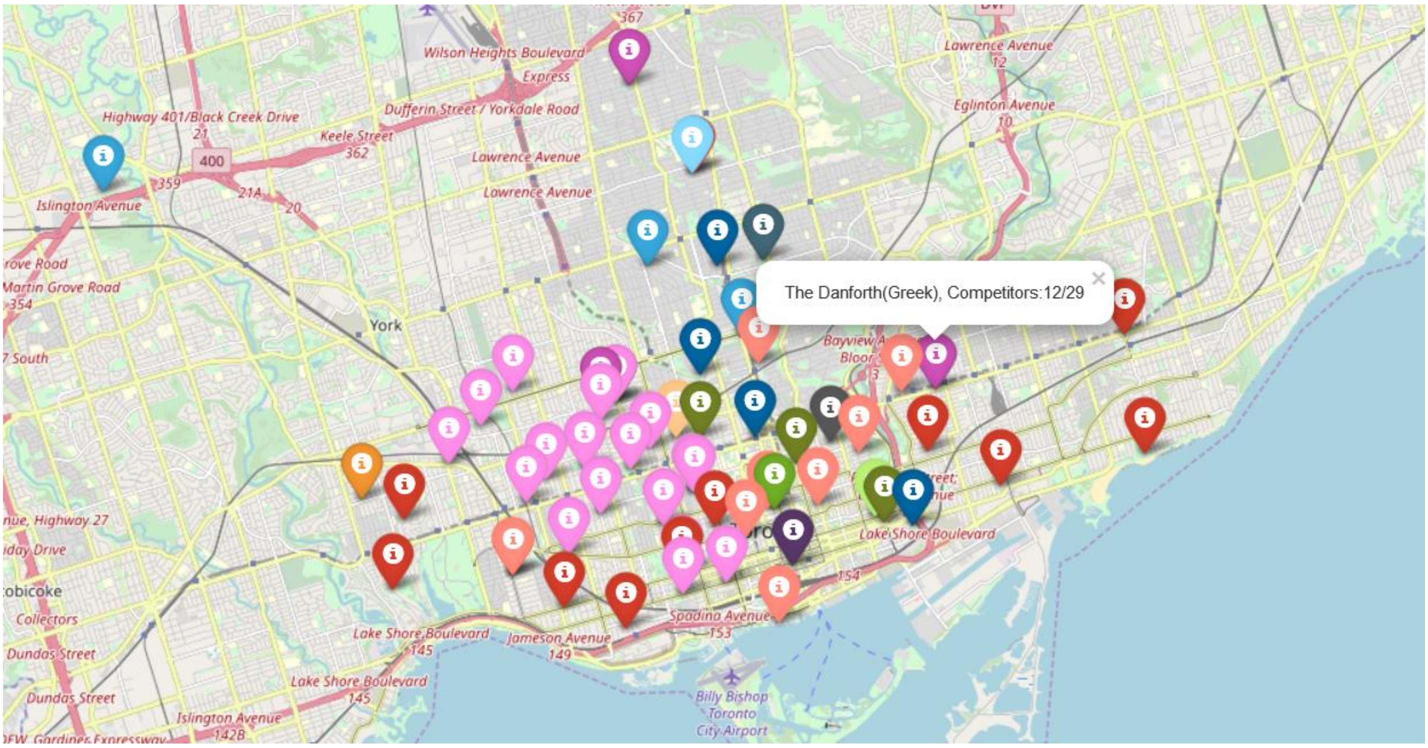
	Neighborhood	Population	Density	Average income	Percentage	Language	Second language population	Latitude	Longitude	Total Restaurants
0	Alexandra Park	4355	13609	19687	17.9	Cantonese	779	43.71627	-79.40555	11
1	Allenby	2513	4333	245592	1.4	Russian	35	43.71275	-79.54746	4
2	Bay Street Corridor	4787	43518	40598	9.6	Mandarin	459	43.65777	-79.38619	23
3	Bedford Park	13749	6057	80827	0.7	Greek	96	43.73138	-79.42116	16
4	Bloor West Village	5175	6993	55578	3.6	Ukrainian	186	43.65936	-79.48543	5
5	Bracondale Hill	5343	8618	41605	4.8	Greek	256	43.67600	-79.42803	27
6	Brockton	9039	8217	27260	19.9	Portuguese	1798	43.66055	-79.40531	32
7	Cabbagetown	11120	7943	50398	1.6	Chinese	177	43.66763	-79.36606	13
8	Carleton Village	6544	8843	23301	17.0	Portuguese	1112	43.67200	-79.45700	20
9	Casa Loma	3597	5369	82203	1.8	Korean	64	43.67000	-79.41000	37

Since a language can be spoken by more than one country (and represent more than one cuisines), the following speaking groups are formed:

For Portuguese assume common ethnic group for Brazilian and Portuguese Restaurants. For Japanese assume common ethnic group for Sushi and Japanese Restaurants. For Cantonese assumed common ethnic group for Thai, Taiwanese, Vietnamese, Cantonese, Indonesian Restaurants. For Mandarin as Chinese for Chinese Restaurant. This is important as it enables to differentiate the ethnic restaurants per neighborhood. These restaurants will be the competitors if we want to open a new ethnic restaurant (df: Toronto_restaurants_conclude.head(10)). We remove for now the ‘Average income’ and ‘Density’ columns:

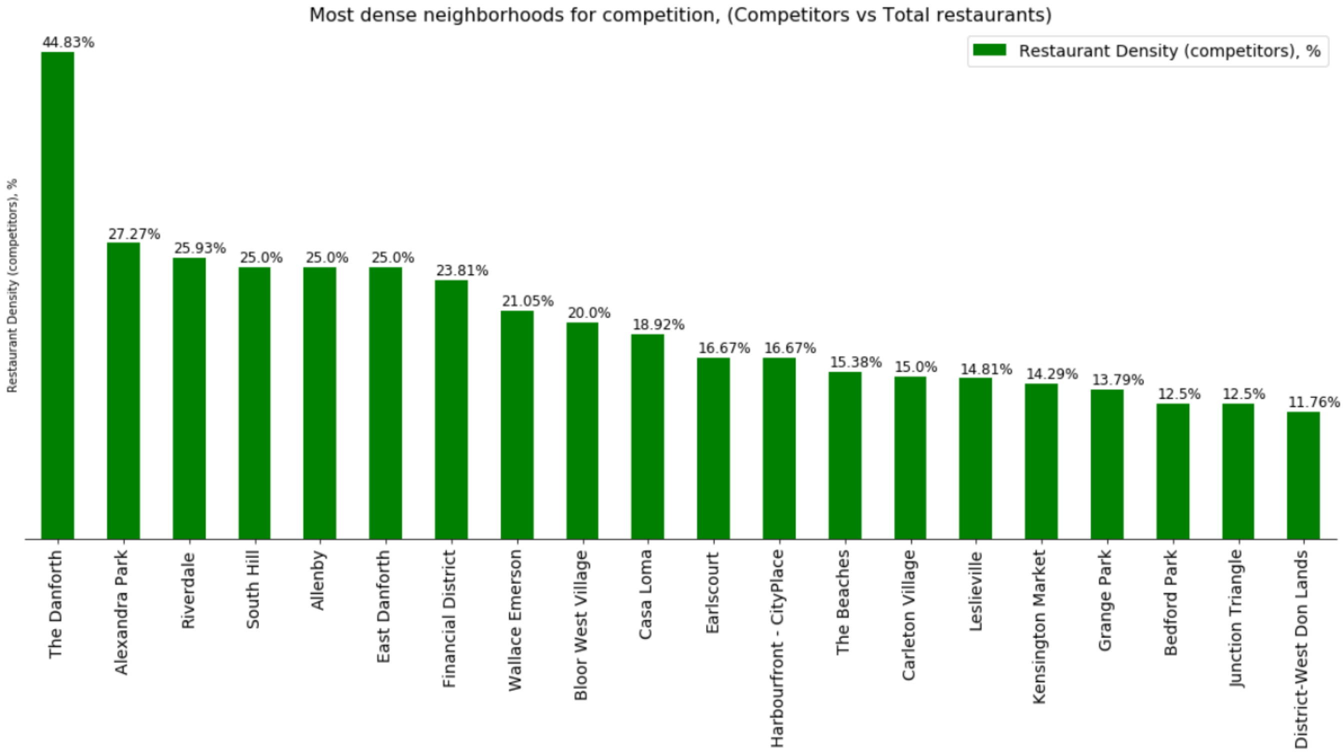
	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude
0	Alexandra Park	Cantonese	4355	779	11	2	43.71627	-79.40555
1	Allenby	Russian	2513	35	4	0	43.71275	-79.54746
2	Bay Street Corridor	Mandarin	4787	459	23	1	43.65777	-79.38619
3	Bedford Park	Greek	13749	96	16	1	43.73138	-79.42116
4	Bloor West Village	Ukrainian	5175	186	5	0	43.65936	-79.48543
5	Bracondale Hill	Greek	5343	256	27	0	43.67600	-79.42803
6	Brockton	Portuguese	9039	1798	32	0	43.66055	-79.40531
7	Cabbagetown	Chinese	11120	177	13	0	43.66763	-79.36606
8	Carleton Village	Portuguese	6544	1112	20	2	43.67200	-79.45700
9	Casa Loma	Korean	3597	64	37	6	43.67000	-79.41000

We can then visualize in a map the number of competitors for each neighborhood:



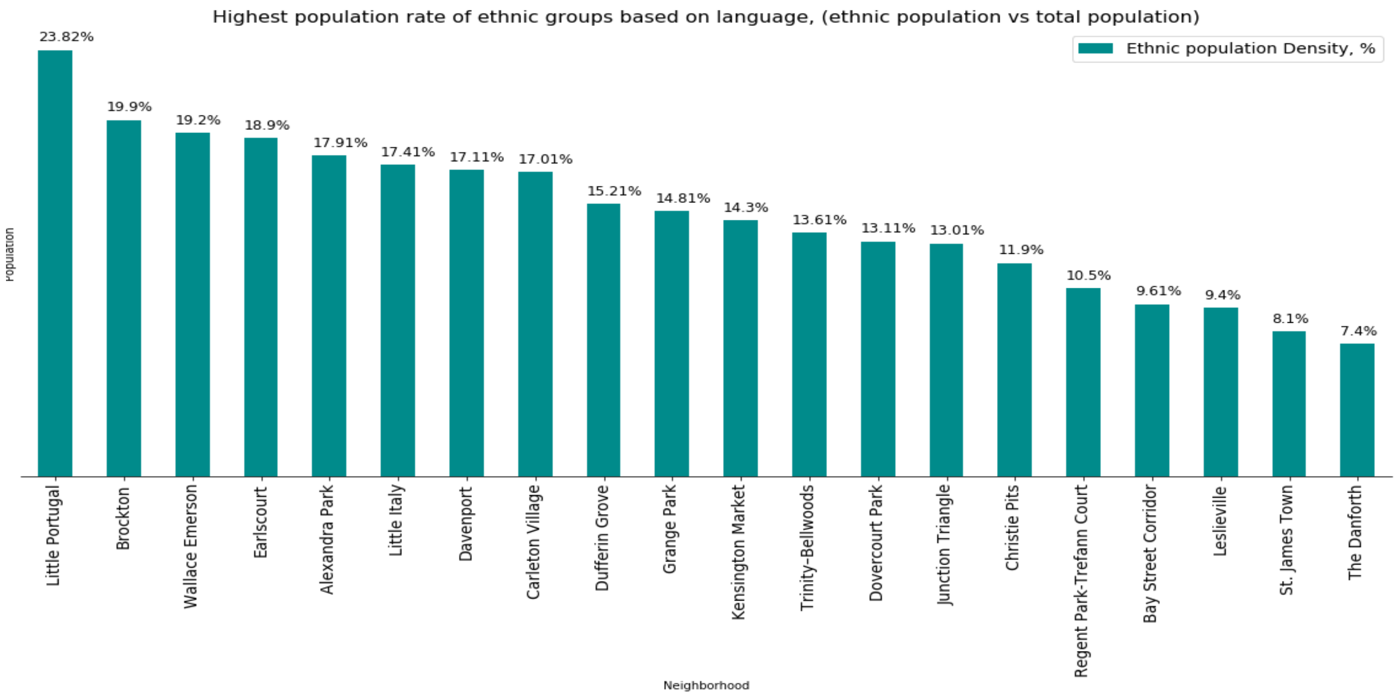
4. Introduce weight factors

Based on the initial requirements, to have the less competition possible, it is needed to introduce a new factor to depict the "Density of ethnic restaurants (same as language) out of total restaurants" (less is best). According to this factor the neighborhoods that should be avoided due to high number of competitors (ethnic restaurants same as the language spoken), is as follows:



We can see that "The Danforth", "Alexandra Park" and "Riverdale" are highly competitive for these kind of business and better to be avoided.

Similarly for the requirement to have strong ethnic presence in a neighborhood, the density of ethnic population over the total population is introduced (large is best). Neighborhoods such "Little Portugal", 'Brockton", "Wallace Emerson" have a high degree of the dominant language-ethnic groups. They should be considered in relation to average income and population later on if they are good candidates.



We combine the 2 density factors with the main dataframe to a final dataframe, as shown below (df: Toronto_restaurants_final.head(10)):

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %
0	Earlscourt	Portuguese	17240	3258	18	2	43.678000	-79.449000	16.67	18.90
1	Leslieville	Cantonese	23567	2215	27	3	43.661927	-79.332039	14.81	9.40
2	Riverdale	Cantonese	31007	2077	27	6	43.667750	-79.349610	25.93	6.70
3	Wallace Emerson	Portuguese	10338	1984	19	3	43.663000	-79.441000	21.05	19.20
4	Brockton	Portuguese	9039	1798	32	0	43.660550	-79.405310	3.12	19.90
5	Davenport	Portuguese	8781	1501	19	1	43.673000	-79.428000	10.53	17.11
6	Dufferin Grove	Portuguese	9875	1501	28	1	43.657000	-79.428000	7.14	15.21
7	Little Italy	Portuguese	7917	1377	33	1	43.655000	-79.413000	6.06	17.41
8	Grange Park	Chinese	9007	1333	29	3	43.653000	-79.393000	13.79	14.81
9	Little Portugal	Portuguese	5013	1193	25	1	43.650000	-79.435556	8.00	23.82

4. Results

We could manually try to search for the optimum location base on the following criteria:

1) Assume middle-class and above neighborhoods only, i.e. merge the dataframe with the 'average-income' available from demographic information. Find the

average income and filter neighborhoods above the man value, i.e. middle and above class.

2) Large population, so to attract as many people as possible. Further filter the above dataframe for neighborhoods with population above the mean population value

3) Large ethnic community, so to have significant cultural characteristics. Consider for ethnic group significant high, i.e. above average number of all ethnic groups.

4) Less number of competitors, so to avoid competition as much as possible. Based on restaurant density we keep only the neighborhoods where the competition is below the average number of competitors.

The result is (df: T_manual):

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income
35	The Annex	Spanish	15602	202	36	0	43.670000	-79.404000	2.78	1.3	63636
27	Davisville	Persian	23727	355	35	0	43.701000	-79.389000	2.86	1.5	55735
38	Deer Park	Russian	15165	166	17	0	43.688056	-79.394028	5.88	1.1	80704
28	Swansea	Polish	11133	333	14	0	43.643889	-79.477778	7.14	3.0	58681
22	Forest Hill	Russian	24056	577	12	0	43.700000	-79.416667	8.33	2.4	101631

Conclusion by observation:

“The Annex” is the first candidate, although “Davisville” has larger population and similar restaurant and population density, i.e. “Davisville” is better. “Deer Park” and Swansea have less ethnic population and larger restaurant densities, i.e. no good. “Forest Hill” is similar to “Davisville” but with mush worse density factors.

The winner seems to be “Davisville” for opening a Persian, medium-upper class restaurant.

Analysis of results by machine learning (k-means)

Let us add to the main dataframe “Toronto_restaurants_final” the information about the average income, since this is relative with the type of restaurant that will open, i.e. lower, middle, upper, high class and then try to apply the k-means algorithm and see the results.

For the clustering algorithm, a cluster of 5 groups (k=5) will be sufficient for the analysis. The following clusters are formed:

Cluster0:

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	The Annex	Spanish	15602	202	36	0	43.670000	-79.404000	2.78	1.30	63636	0
1	Fashion District	Portuguese	4642	51	32	0	43.645000	-79.398000	3.12	1.12	63282	0
2	Summerhill	Chinese	5100	56	30	0	43.683000	-79.390000	3.33	1.12	88937	0
3	Deer Park	Russian	15165	166	17	0	43.688056	-79.394028	5.88	1.10	80704	0
4	Chaplin Estates	French	4906	58	33	1	43.700000	-79.400000	6.06	1.20	81288	0
5	Bedford Park	Greek	13749	96	16	1	43.731380	-79.421160	12.50	0.71	80827	0
6	The Beaches	Cantonese	20416	142	13	1	43.667266	-79.297128	15.38	0.70	67536	0
7	Casa Loma	Korean	3597	64	37	6	43.670000	-79.410000	18.92	1.81	82203	0
8	Financial District	Japanese	548	9	21	4	43.647935	-79.381752	23.81	1.82	63952	0

Cluster1:

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	Rosedale	Chinese	7672	76	24	0	43.646231	-79.449048	4.17	1.00	213941	1
1	Allenby	Russian	2513	35	4	0	43.712750	-79.547460	25.00	1.43	245592	1

Cluster2:

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	Seaton Village	Portuguese	5259	262	38	0	43.668000	-79.416000	2.63	5.00	41506	2
1	Davisville	Persian	23727	355	35	0	43.701000	-79.389000	2.86	1.50	55735	2
2	Harbord Village	Portuguese	5906	242	33	0	43.661000	-79.406000	3.03	4.11	45792	2
3	Church and Wellesley	Spanish	13397	241	29	0	43.665694	-79.380956	3.45	1.81	37653	2
4	Playter Estates	Chinese	3968	71	29	0	43.678056	-79.355556	3.45	1.81	44557	2
5	Bracondale Hill	Greek	5343	256	27	0	43.676000	-79.428030	3.70	4.81	41605	2
6	Upper Beaches	Cantonese	19830	138	25	0	43.646667	-79.408333	4.00	0.70	44346	2
7	Roncesvalles	Polish	15996	703	24	0	43.646231	-79.449048	4.17	4.40	46820	2
8	Garden District	Chinese	8240	247	24	0	43.658500	-79.375800	4.17	3.01	37614	2
9	Niagara	Portuguese	6524	260	23	0	43.643000	-79.408000	4.35	4.00	44611	2
10	Corktown	Spanish	4484	94	22	0	43.655518	-79.359712	4.55	2.12	54681	2
11	High Park North	Polish	22746	682	18	0	43.656000	-79.475000	5.56	3.00	46437	2
12	Wychwood	Portuguese	4182	112	29	1	43.676200	-79.424400	6.90	2.70	53613	2
13	Swansea	Polish	11133	333	14	0	43.643889	-79.477778	7.14	3.00	58681	2
14	Cabbagetown	Chinese	11120	177	13	0	43.667630	-79.366060	7.69	1.60	50398	2
15	Bay Street Corridor	Mandarin	4787	459	23	1	43.657770	-79.386190	8.70	9.61	40598	2
16	Discovery District	Chinese	7262	472	26	2	43.658000	-79.388000	11.54	6.51	41998	2
17	Bloor West Village	Ukrainian	5175	186	5	0	43.659360	-79.485430	20.00	3.61	55578	2
18	Riverdale	Cantonese	31007	2077	27	6	43.667750	-79.349610	25.93	6.70	40139	2
19	The Danforth	Greek	7849	580	29	12	43.678472	-79.347222	44.83	7.40	44979	2

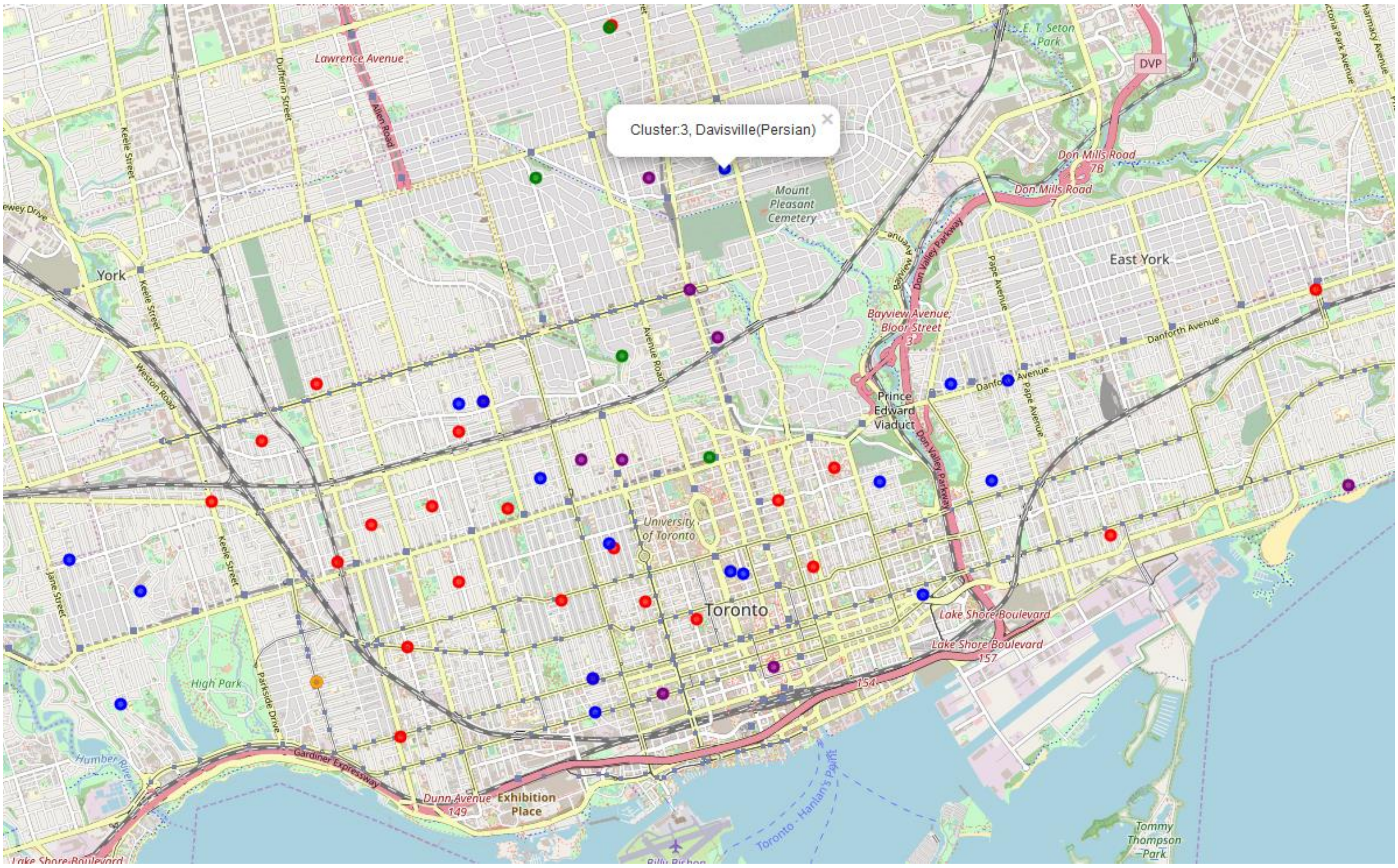
Cluster3:

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	Forest Hill	Russian	24056	577	12	0	43.700000	-79.416667	8.33	2.40	101631	3
1	Lytton Park	Serbian	6494	58	11	0	43.716000	-79.406000	9.09	0.91	127356	3
2	Yorkville	French	6045	114	38	3	43.670278	-79.391111	10.53	1.90	105239	3
3	South Hill	French	6218	62	16	3	43.681000	-79.404000	25.00	1.01	120453	3

Cluster4:

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	Christie Pits	Portuguese	5124	609	37	0	43.664722	-79.420833	2.70	11.90	30556	4
1	Brockton	Portuguese	9039	1798	32	0	43.660550	-79.405310	3.12	19.90	27260	4
2	Parkdale	Polish	28367	822	29	0	43.640454	-79.436731	3.45	2.90	26314	4
3	Regal Heights	Spanish	2719	149	29	0	43.676200	-79.424400	3.45	5.52	36652	4
4	Trinity–Bellwoods	Portuguese	8687	1181	25	0	43.646667	-79.408333	4.00	13.61	31106	4
5	The Junction	Portuguese	11391	467	23	0	43.665556	-79.464444	4.35	4.11	34906	4
6	Dovercourt Park	Portuguese	8497	1113	21	0	43.665000	-79.432000	4.76	13.11	28311	4
7	Little Italy	Portuguese	7917	1377	33	1	43.655000	-79.413000	6.06	17.41	31231	4
8	Dufferin Grove	Portuguese	9875	1501	28	1	43.657000	-79.428000	7.14	15.21	27961	4
9	St. James Town	Filipino	14666	1187	27	1	43.669167	-79.372778	7.41	8.10	22341	4
10	Little Portugal	Portuguese	5013	1193	25	1	43.650000	-79.435556	8.00	23.82	29224	4
11	Davenport	Portuguese	8781	1501	19	1	43.673000	-79.428000	10.53	17.11	28335	4
12	Junction Triangle	Portuguese	6666	866	24	2	43.659000	-79.446000	12.50	13.01	28067	4
13	Grange Park	Chinese	9007	1333	29	3	43.653000	-79.393000	13.79	14.81	35277	4
14	Kensington Market	Cantonese	3740	534	35	4	43.654772	-79.400678	14.29	14.30	23335	4
15	Leslieville	Cantonese	23567	2215	27	3	43.661927	-79.332039	14.81	9.40	30886	4
16	Carleton Village	Portuguese	6544	1112	20	2	43.672000	-79.457000	15.00	17.01	23301	4
17	Earlscourt	Portuguese	17240	3258	18	2	43.678000	-79.449000	16.67	18.90	26672	4
18	Wallace Emerson	Portuguese	10338	1984	19	3	43.663000	-79.441000	21.05	19.20	25029	4
19	East Danforth	Cantonese	21440	900	12	2	43.688056	-79.301944	25.00	4.20	33847	4
20	Alexandra Park	Cantonese	4355	779	11	2	43.716270	-79.405550	27.27	17.91	19687	4

The above results can be visualized in a map as follows:



5. Discussion

We can see from the results, that the main factor that the machine learning algorithm has used to divide the neighborhoods is the **'Average income'** data. This property proved to be the more decisive from all other properties of the neighborhoods.

In more details per cluster, we can see the following:

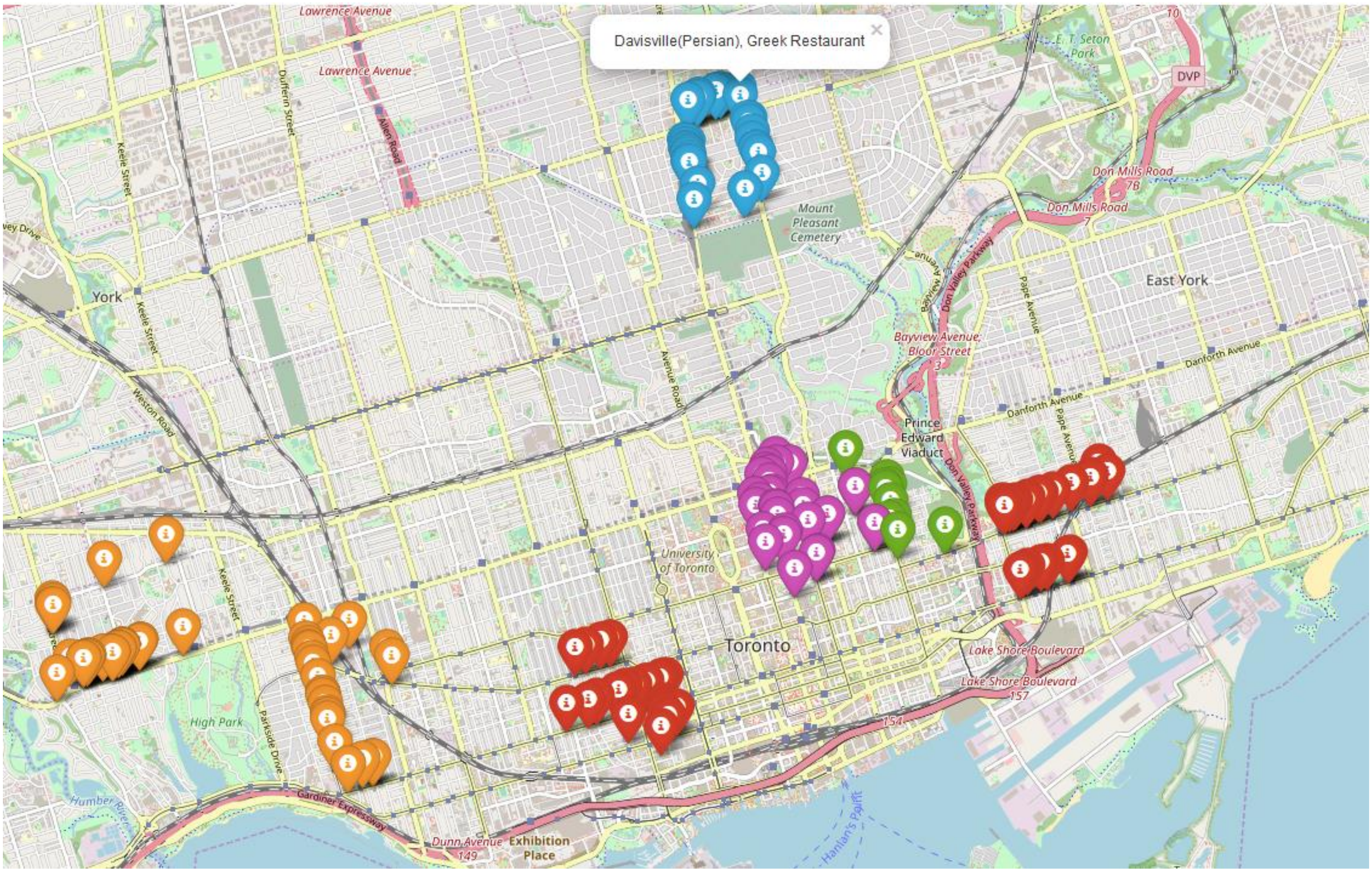
Cluster 0, has the upper-class population (60k – 90k). All neighborhoods either have small ethnic group, or small population relative to neighborhoods of other clusters. "The Annex" (Spanish) seems to be the best option for this group. Not efficient enough.

Cluster 1, has the most expensive areas (>200k), although the population at these areas is small and the ethnic group small, i.e. The areas do not represent a high a cultural neighborhood. Not efficient enough.

Cluster 2, has middle class areas (40k – 50k). We could say that being at the average class, both low-level and high-level income citizens can be attracted, i.e. this is the most representative group of neighborhoods. Let us further filter for population more than the average of the cluster (df: Cluster2_final):

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
6	Upper Beaches	Cantonese	19830	138	25	0	43.646667	-79.408333	4.00	0.70	44346	2
1	Davisville	Persian	23727	355	35	0	43.701000	-79.389000	2.86	1.50	55735	2
14	Cabbagetown	Chinese	11120	177	13	0	43.667630	-79.366060	7.69	1.60	50398	2
3	Church and Wellesley	Spanish	13397	241	29	0	43.665694	-79.380956	3.45	1.81	37653	2
11	High Park North	Polish	22746	682	18	0	43.656000	-79.475000	5.56	3.00	46437	2
13	Swansea	Polish	11133	333	14	0	43.643889	-79.477778	7.14	3.00	58681	2
7	Roncesvalles	Polish	15996	703	24	0	43.646231	-79.449048	4.17	4.40	46820	2
18	Riverdale	Cantonese	31007	2077	27	6	43.667750	-79.349610	25.93	6.70	40139	2

The strongest ethnic groups are at "Davisville", "High Park North", "Roncesvalles" and "Riverdale" (max). From all the above "Davisville" (Persian) has the lowest competition (1.5%) and the second largest population after "Riverdale". **So for this cluster and overall clusters, "Davisville" (Persian) is the best option for opening a new Persian restaurant (middle-class).**



Cluster3, has high-class areas (>100k). Small ethnic groups relatively to population and not many restaurants in the area. “Forest Hill” (Russian) seems the exception and for this cluster is the best option. For high-class restaurant “Forest Hill” (Russian) is the best option.

Cluster4, has the low-class areas (<35k). At these areas there is very high competition for almost the half of the neighborhoods. Best of all seems to be “Parkdale” (Polish) with very high population, very strong Polish group representative, no competition for other ethnic restaurants and relatively low competition from other types of restaurants. For low-level class “Parkdale” (Polish) is the clear winner.

Conclusion

We have analyzed the neighborhoods of Old Toronto with respect to

- Population
- Competition

- Ethnic group presence (language oriented)
- Average income

The results from observation are the same as the ones from applying the k-means algorithm. The 'Average income' was the most distinctive property for the neighborhoods, more important than other significant properties such as the population. Below the best candidates, based on "Average income":

	Neighborhood	Language	Population	Second language population	Total Restaurants	Number of Competitors	Latitude	Longitude	Restaurant Density (competitors), %	Ethnic population Density, %	Average income	Cluster Labels
0	Parkdale	Polish	28367	822	29	0	43.640454	-79.436731	3.45	2.9	26314	1
1	Davisville	Persian	23727	355	35	0	43.701000	-79.389000	2.86	1.5	55735	3
2	Forest Hill	Russian	24056	577	12	0	43.700000	-79.416667	8.33	2.4	101631	4

The best option to open an ethnic Persian restaurant is "Davisville". This area has less competition, and second best population from other candidates. It is in area with average income, i.e. can attract better all classes equally, in contrast to "Forest Hill" which seems rather expensive.

Below a map of competition in Davisville (restaurant types are represented via different colors):