Segmentation for a Restaurant Location Planning

Michael Dubem

May 23, 2021

1. Introduction

1.1 Background

There's a significant rise in entrepreneurs, online vendors, digital/affiliate marketing and the likes in recent times. In addition, a lot of new small scale businesses are surfacing, coupled with the expansion of already existing medium scale businesses and large scale businesses. In summary, whether a business is online or offline, small, medium or big, starting anew, afresh or expanding, the importance of market analysis can't be overemphasized. Now, if one was to be asked, 'Would you rather drive with your eyes closed or open?', the obvious answer would be 'open', sadly many individuals go into business and carry out operations with their eyes closed, either because they don't know better or they feel that market analysis or reliance on historical or current data is for bigger firms with physical presence.

1.2 Problem

When opening a new business, there are a couple of analysis one is expected to carry out, e.g. SWOT, but for the sake of simplicity, this report will follow a more fundamental approach to the concept of market analysis. Now looking at the case study (a restaurant), a new business would want to check the area that would have the optimal demand and supply, with moderate or little competition (if possible). As good as that sound, it is very hard to come by in real case scenarios, because most areas with high demand and good proximity to supply, also have high competition, which is where the analysis comes in to help the business see the optimal options based on the categories being considered and make a data driven decision.

1.3 Interest

This is already a widely adopted approach to business, by larger firms. It just isn't as popular as it should amongst smaller businesses. Any business would want to allocate its resources efficiently to avoid waste and get the most results, so factors of business operations like ads targeting, customer hunts, customer retention promos and others are pushed to the right audience to get the best results.

2. Data Acquisition and Cleaning

2.1 Data Sources

The research was focused on the boroughs and neighbourhoods of Toronto, Canada, majorly because this is a developed region and a lot of data about this region is readily available online. For some less popular location, on ground/field survey might be essential for an applicable analysis. A lot of the data needed for the market analysis can be gotten through location APIs. The Foursquare Location API was used in this analysis to get crucial details about the geographical locations under scrutiny. Other APIs and platforms like the Geocoder was used to get the geographical coordinates of the areas under scrutiny and google maps was used to get the approximate land area covered by the respective boroughs of Toronto. The research also covered web scraping Wikipedia: (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&direction=prev&oldid=926287641) to get data like the list of neighbourhoods in each borough and their respective postal codes.

2.2 Data Extraction and Cleaning

The first data gotten was scraped from the Wikipedia page for list of postal codes in Toronto, and was loaded in to a data frame. Unassigned neighbourhoods and boroughs were cleaned off the table, neighbourhoods with the same postal codes were merged into the same rows and some neighbourhood names were changed.

The next step was getting the geographical coordinates of each postal code and loading it in to a data frame and merging it to the previous list of boroughs and neighbourhoods, so each neighbourhood was matched to its respective coordinate. The coordinate for each borough was also gotten and loaded into a data frame, because the boroughs were to be analysed first before analysing the neighbourhoods of the optimal borough. The land area of the respective boroughs was also manually measured and merged into the existing data frame, so that any location call made on each borough would reflect the true size of the borough in relation to others. Instead of having to go through the stress of analysing and visualizing all the neighbourhoods, it seemed like a more efficient approach to analyse the optimal borough, then focus on only the neighbourhoods of that borough. The final part of the data extraction was using the Foursquare API to call up the number of venues in some certain categories that could give us insight into the nature of the demand and supply, as well as the competition in the areas under scrutiny. The categories covered were split into three groups, which was supply, demand and competition. For supply, the count on the number of supermarkets and shopping malls in each borough was extracted. The concept behind these venues is that, proximity to shopping malls and supermarkets means readily available raw materials for operations. There are other sources of supply, but these two were the focus in this research. For demand, the count on the number of colleges and universities, outdoors and recreation centres, hotels, professional complexes, and inhabitant's residences was gotten, due to the fact that these venues give an insight into the level of population and the type of population in the particular borough. Some age grades are more welcoming of the concept of eating out. The final category was competition, and the count for the number of restaurants in each borough was also extracted. After the optimal borough was selected, the same extraction process was repeated for the neighbourhoods in the optimal borough.

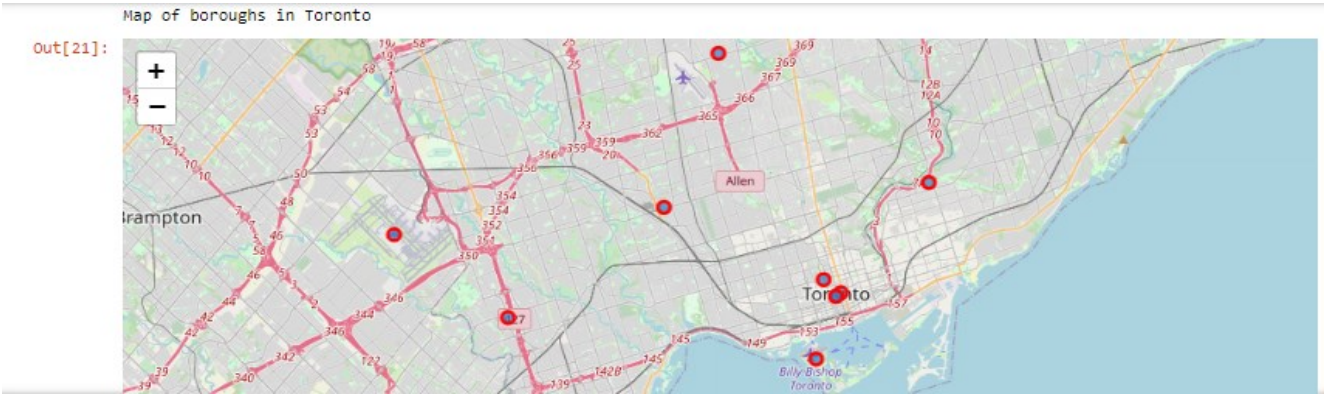
Out[56]:

	PostalCode	Borough	Borough Lat.	Borough Long.	Borough Rad.	Neighborhood	Latitude	Longitude
0	M3A	North York	43.754326	-79.449117	6558	Parkwoods	43.753259	-79.329656
1	M4A	North York	43.754326	-79.449117	6558	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	43.654174	-79.380812	1609	Harbourfront, Regent Park	43.654260	-79.360636
3	M6A	North York	43.754326	-79.449117	6558	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	43.659659	-79.390340	1609	Queen's Park	43.662301	-79.389494

Out[15]:

	Borough	Borough Lat.	Borough Long.
0	North York	43.754326	-79.449117
1	Downtown Toronto	43.654174	-79.380812
2	Scarborough	43.772974	-79.257648
3	Etobicoke	43.643556	-79.565633
4	Central Toronto	43.652384	-79.383568
5	West Toronto	43.652384	-79.383568
6	York	43.689619	-79.479188
7	East Toronto	43.626122	-79.395035
8	East York	43.699971	-79.332520
9	Queen's Park	43.659659	-79.390340
10	Mississauga	43.678524	-79.629129

Out[21]:



2.3 Feature Selection

K-means clustering was used to detect similarities amongst boroughs and neighbourhoods, and the major focus for checking for similarities was the features that fell under the three outlined categories (supply,

demand and competition). Hence, the features under each category were merged using a simple average. The clustering was then carried out on only the numeric columns in the data frame (except the coordinates)

Out[24]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Supermarkets	Shopping Malls	College & University	Outdoors & Recreation	Hotels	Professional & Other Places	Residence	Food
0	North York	43.754326	-79.449117	6558	30	27	63	12	13	23	0	152
1	Downtown Toronto	43.654174	-79.380812	1609	16	14	99	41	86	34	6	166
2	Scarborough	43.772974	-79.257648	6437	36	35	51	2	10	7	0	227
3	Etobicoke	43.643556	-79.565633	6598	18	21	22	7	52	8	3	176
4	Central Toronto	43.652384	-79.383568	1609	17	15	86	51	87	38	7	191
5	West Toronto	43.652384	-79.383568	1609	17	15	86	51	87	38	7	191
6	York	43.689619	-79.479188	3017	6	4	3	3	2	3	0	105
7	East Toronto	43.626122	-79.395035	1609	2	0	3	5	7	0	0	21
8	East York	43.699971	-79.332520	2293	4	1	8	2	3	3	0	83
9	Queen's Park	43.659659	-79.390340	1609	17	19	102	40	81	36	6	200
10	Mississauga	43.678524	-79.629129	12271	59	66	43	11	95	21	4	240

Out[25]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Competition	Supply	Demand
0	North York	43.754326	-79.449117	6558	152	28.5	22.2
1	Downtown Toronto	43.654174	-79.380812	1609	166	15.0	53.2
2	Scarborough	43.772974	-79.257648	6437	227	35.5	14.0
3	Etobicoke	43.643556	-79.565633	6598	176	19.5	18.4
4	Central Toronto	43.652384	-79.383568	1609	191	16.0	53.8
5	West Toronto	43.652384	-79.383568	1609	191	16.0	53.8
6	York	43.689619	-79.479188	3017	105	5.0	2.2
7	East Toronto	43.626122	-79.395035	1609	21	1.0	3.0
8	East York	43.699971	-79.332520	2293	83	2.5	3.2
9	Queen's Park	43.659659	-79.390340	1609	200	18.0	53.0
10	Mississauga	43.678524	-79.629129	12271	240	62.5	34.8

3. Methodology

In this section of the project, an unsupervised machine learning algorithm (k-means clustering) was applied on the selected features from the dataset in order to find similarities between neighbourhood so as to narrow down the selection. Two clusters labels were produced, the first was the supply and demand cluster, as our target was to get neighbourhood with high or moderate demand and supply. The second cluster was on approx. borough land radius and level of competition in the respective boroughs. Our focus in the second cluster was to find neighbourhoods with moderate land area and low or moderate competition.

3.1 Selection of the optimal borough

The optimal boroughs with respect to each cluster will be selected, and the recurring boroughs in each selection will be shortlisted. The most likely optimal borough will then be picked from comparing the boroughs within the shortlisted boroughs. The concept behind this analysis is that, for someone looking to open a restaurant, you will want to open in an environment with high demand and supply of the products you need for operation, and you won't want somewhere with excessive competition.

Another not so obvious observation is that, you could see a borough with adequate demand and supply, and relatively moderate competition but with a relatively very large area coverage, which will still impact negatively on the demand (as someone on one side of town might feel discouraged to drive all the way to another side of town, just to eat out). You could also see a borough with relatively moderate competition, and relatively very small area coverage, which would in turn amp up the struggle for placement and relative competition.

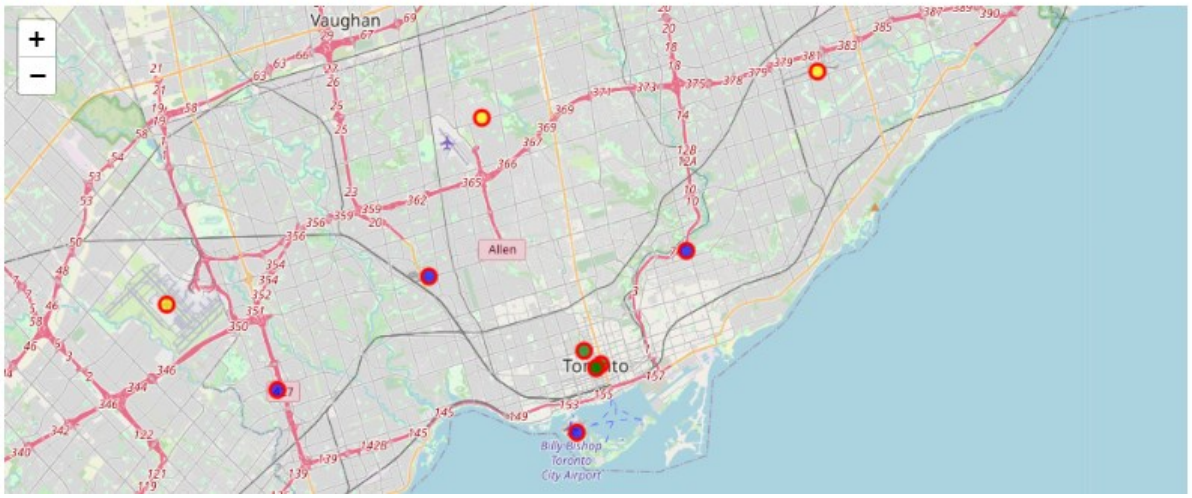
So our focus here is to find a borough with moderate area coverage and competition, with adequate demand and supply.

Out[33]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster
0	North York	43.754326	-79.449117	6558	152	28.5	22.2	0	0
1	Downtown Toronto	43.654174	-79.380812	1609	166	15.0	53.2	1	2
2	Scarborough	43.772974	-79.257648	6437	227	35.5	14.0	0	0
3	Etobicoke	43.643556	-79.565633	6598	176	19.5	18.4	2	0
4	Central Toronto	43.652384	-79.383568	1609	191	16.0	53.8	1	2
5	West Toronto	43.652384	-79.383568	1609	191	16.0	53.8	1	2
6	York	43.689619	-79.479188	3017	105	5.0	2.2	2	1
7	East Toronto	43.626122	-79.395035	1609	21	1.0	3.0	2	1
8	East York	43.699971	-79.332520	2293	83	2.5	3.2	2	1
9	Queen's Park	43.659659	-79.390340	1609	200	18.0	53.0	1	2
10	Mississauga	43.678524	-79.629129	12271	240	62.5	34.8	0	0

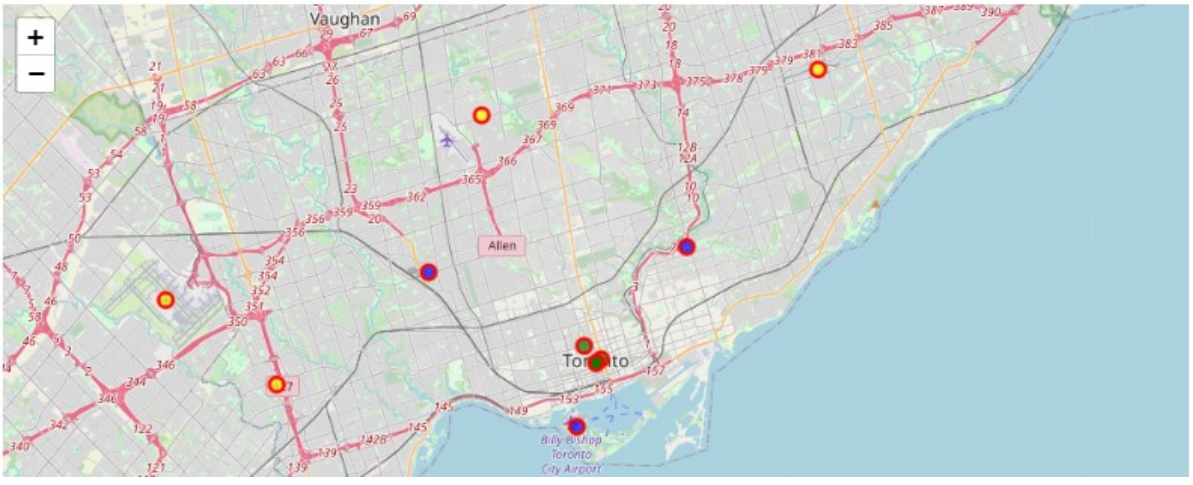
```
Geospatial Map showing the different Supply and Demand Clusters

Color: Blue
Cluster: 2
Characteristic: Low Average Supply and Demand

Color: Green
Cluster: 1
Characteristic: Moderate Average Supply and Demand

Color: Yellow
Cluster: 0
Characteristic: High Average Supply and Demand
```

Out[36]:




```
Geospatial Map showing the different Area Coverage and Competition Level Clusters

Color: Blue
Cluster: 1
Characteristic: Low Competition Level and Moderate Area Coverage

Color: Green
Cluster: 2
Characteristic: Moderate Competition Level and Low Area Coverage

Color: Yellow
Cluster: 0
Characteristic: High Competition Level and High Area Coverage
```

Out[38]:



3.2 Further selection of the optimal neighbourhood in the selected borough

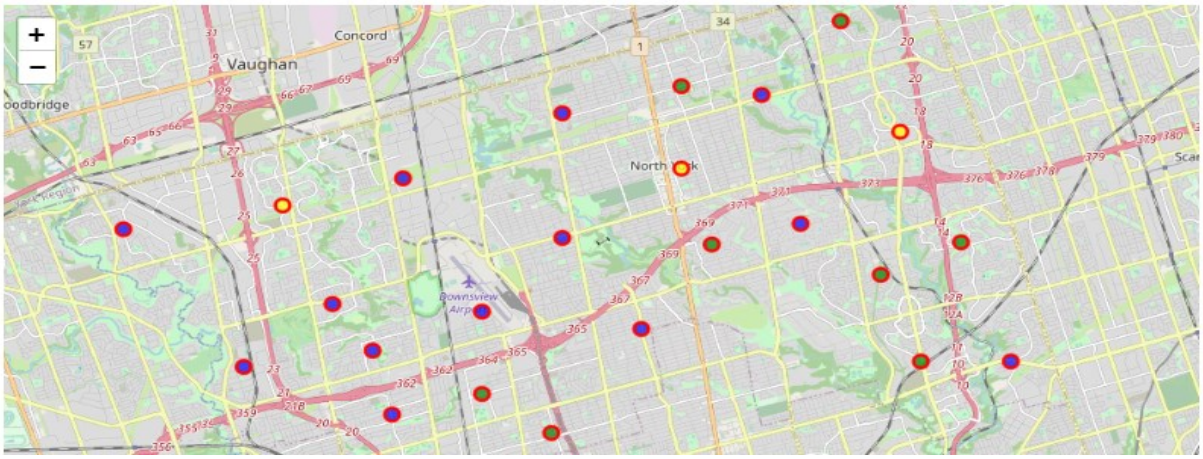
After the optimal borough is picked, the same process is repeated on the neighbourhoods within the selected optimal borough, and the optimal neighbourhood is then picked in a similar manner.

```
Map of neighbourhoods in North York, Toronto
```

Out[48]:

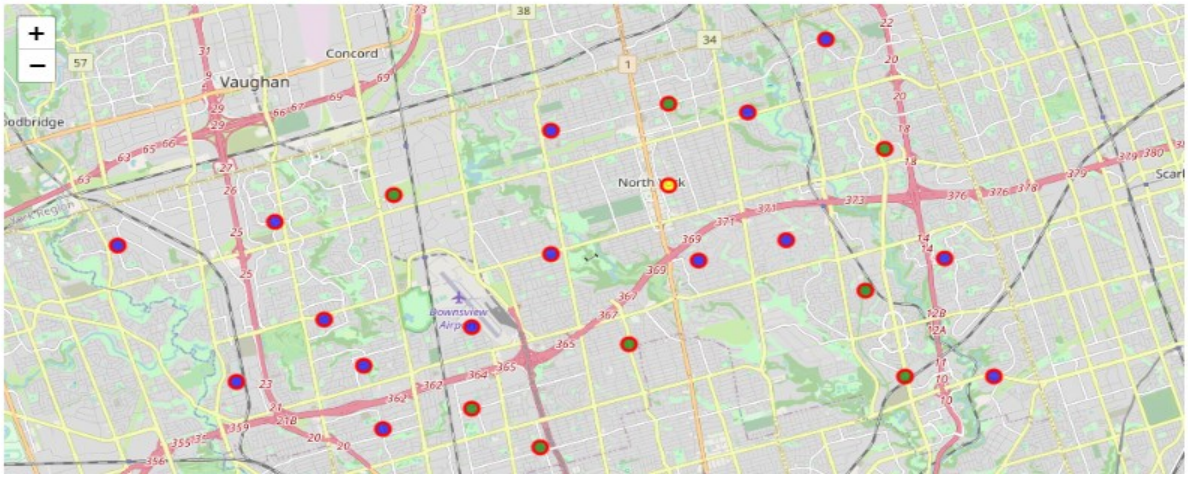

```
Geospatial Map showing the different Supply and Demand Clusters

Color: Blue
Cluster: 2
Characteristic: Low Average Supply and Demand

Color: Green
Cluster: 0
Characteristic: Moderate Average Supply and Demand

Color: Yellow
Cluster: 1
Characteristic: High Average Supply and Demand
```

Out[59]:




```
Geospatial Map showing the different Competition Level Clusters

Color: Blue
Cluster: 0
Characteristic: Low Competition Level

Color: Green
Cluster: 2
Characteristic: Moderate Competition Level

Color: Yellow
Cluster: 1
Characteristic: High Competition Level
```

Out[61]:



4. Results and Discussion

4.1 Results

After the first selection process for boroughs, there were only two remaining boroughs; North York and Downtown Toronto. Downtown Toronto had considerably higher supply and demand, but the competition was significantly higher due to its small land area in comparison to North York. Downtown Toronto had higher competition in a much smaller land area which meant more restaurants in a tighter cluster competing for its higher demand and supply, so North York was the natural pick.

In [42]:
```
#Let's get the boroughs with the least competition
opt_loc = borough_coord.sort_values(['Competition'], ascending=True)
opt_loc = opt_loc.head(6)
opt_loc
```

Out[42]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster	Avg. S&D
7	East Toronto	43.626122	-79.395035	1609	21	1.0	3.0	2	1	2.00
8	East York	43.699971	-79.332520	2293	83	2.5	3.2	2	1	2.85
6	York	43.689619	-79.479188	3017	105	5.0	2.2	2	1	3.60
0	North York	43.754326	-79.449117	6558	152	28.5	22.2	0	0	25.35
1	Downtown Toronto	43.654174	-79.380812	1609	166	15.0	53.2	1	2	34.10
3	Etobicoke	43.643556	-79.565633	6598	176	19.5	18.4	2	0	18.95

In [43]:
```
#Let's get the boroughs with the highest Supply and Demands
opt_loc1 = borough_coord.sort_values(['Avg. S&D'], ascending=False)
opt_loc1 = opt_loc1.head(6)
opt_loc1
```

Out[43]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster	Avg. S&D
10	Mississauga	43.678524	-79.629129	12271	240	62.5	34.8	0	0	48.65
9	Queen's Park	43.659659	-79.390340	1609	200	18.0	53.0	1	2	35.50
4	Central Toronto	43.652384	-79.383568	1609	191	16.0	53.8	1	2	34.90
5	West Toronto	43.652384	-79.383568	1609	191	16.0	53.8	1	2	34.90
1	Downtown Toronto	43.654174	-79.380812	1609	166	15.0	53.2	1	2	34.10
0	North York	43.754326	-79.449117	6558	152	28.5	22.2	0	0	25.35

In [45]:
```
#Now that we have narrowed our consideration to two boroughs, let's compare their parameters
opt_bor = borough_coord.Borough.isin(opt_list)
opt_bor = borough_coord[opt_bor]
opt_bor
```

Out[45]:

	Borough	Borough Lat.	Borough Long.	Borough Rad.	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster	Avg. S&D
0	North York	43.754326	-79.449117	6558	152	28.5	22.2	0	0	25.35
1	Downtown Toronto	43.654174	-79.380812	1609	166	15.0	53.2	1	2	34.10

From the table above, Downtown Toronto could have been considered but it has a higher competition value that North York, and an even much smaller area coverage which mean the competition would be amplified further. Hence our optimal choice is North York.

The same selection process was carried out on the neighbourhoods of North York. Due to the small marginal difference in the area of each neighbourhood, all neighbourhoods were assumed the same size, hence land area was not considered in this part of the analysis. In a similar manner, when the neighbourhoods were narrowed down to highest supply and demand, and lowest competition, Downsview Northwest was the optimal choice as it had the highest supply and demand, with a moderately low competition level.

In [66]:
```
#Let us first eliminat areas of low demand and supply
opt_neigh = df_1['S&D Cluster'].isin([0,1])
opt_neigh = df_1[opt_neigh]
opt_neigh
```

Out[66]:

	PostalCode	Borough	Borough Lat.	Borough Long.	Borough Rad.	Neighborhood	Latitude	Longitude	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster
0	M3A	North York	43.754326	-79.449117	6558	Parkwoods	43.753259	-79.329656	5	1.5	1.0	0	0
3	M6A	North York	43.754326	-79.449117	6558	Lawrence Heights, Lawrence Manor	43.718518	-79.464763	20	0.5	1.2	0	2
7	M3B	North York	43.754326	-79.449117	6558	Don Mills North	43.745906	-79.352188	19	0.5	1.4	0	2
10	M6B	North York	43.754326	-79.449117	6558	Glencairn	43.709577	-79.445073	17	2.0	0.6	0	2
13	M3C	North York	43.754326	-79.449117	6558	Flemingdon Park, Don Mills South	43.725900	-79.340923	21	1.5	1.0	0	2
27	M2H	North York	43.754326	-79.449117	6558	Hillcrest Village	43.803762	-79.363452	7	1.0	1.0	0	0
33	M2J	North York	43.754326	-79.449117	6558	Fairview, Henry Farm, Oriole	43.778517	-79.346556	31	3.0	1.2	1	2
52	M2M	North York	43.754326	-79.449117	6558	Newtonbrook, Willowdale	43.789053	-79.408493	23	2.0	1.2	0	2
59	M2N	North York	43.754326	-79.449117	6558	Willowdale South	43.770120	-79.408493	90	4.0	3.0	1	1
60	M3N	North York	43.754326	-79.449117	6558	Downsview Northwest	43.761631	-79.520999	12	3.0	1.6	1	0
66	M2P	North York	43.754326	-79.449117	6558	York Mills West	43.752758	-79.400049	4	0.0	1.6	0	0

In [67]:
```
#Next we'll eliminate the area of high competition
opt_neigh = opt_neigh[opt_neigh['Competition']<20]
opt_neigh
```

Out[67]:

	PostalCode	Borough	Borough Lat.	Borough Long.	Borough Rad.	Neighborhood	Latitude	Longitude	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster
0	M3A	North York	43.754326	-79.449117	6558	Parkwoods	43.753259	-79.329656	5	1.5	1.0	0	0
7	M3B	North York	43.754326	-79.449117	6558	Don Mills North	43.745906	-79.352188	19	0.5	1.4	0	2
10	M6B	North York	43.754326	-79.449117	6558	Glencairn	43.709577	-79.445073	17	2.0	0.6	0	2
27	M2H	North York	43.754326	-79.449117	6558	Hillcrest Village	43.803762	-79.363452	7	1.0	1.0	0	0
60	M3N	North York	43.754326	-79.449117	6558	Downsview Northwest	43.761631	-79.520999	12	3.0	1.6	1	0
66	M2P	North York	43.754326	-79.449117	6558	York Mills West	43.752758	-79.400049	4	0.0	1.6	0	0

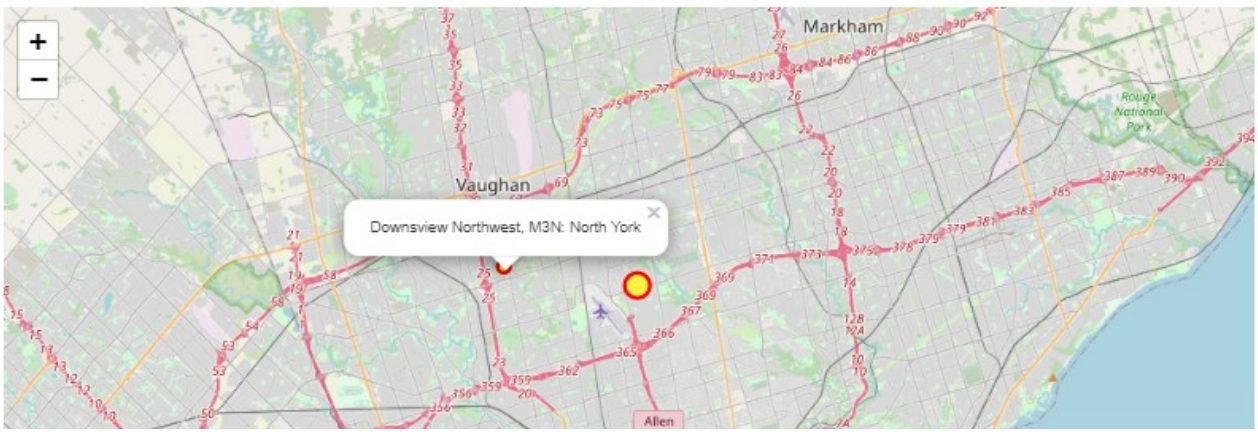
In [68]:
```
# Since we know that the S&D Cluster with label 1 have the highest supply and demand, and the Competiton Clusters with label 0
# have the least competition,
# Let us see if there are any neighbourhood that fits both categories

opt_neigh = opt_neigh[(opt_neigh['S&D Cluster'] == 1) & (opt_neigh['Area&Compt Cluster'] == 0)]
opt_neigh
```

Out[68]:

	PostalCode	Borough	Borough Lat.	Borough Long.	Borough Rad.	Neighborhood	Latitude	Longitude	Competition	Supply	Demand	S&D Cluster	Area&Compt Cluster
60	M3N	North York	43.754326	-79.449117	6558	Downsview Northwest	43.761631	-79.520999	12	3.0	1.6	1	0

From this selection process, it is clear that the neighbourhood with the best supply and demand, while having moderate competition is Downsview Northwest, North York, Toronto.

Map showing the optimal location: Downsview Northwest, North York, Toronto



4.2 Further Discussion

There are a lot more factors that can be considered when opening a business in a real case scenario, and these factor can also be considered in further studies. Factors like the age distribution of the inhabiting population also serves a significant purpose. Even though knowing the kind of recreation centres and professional complexes in the areas gives a bit of an insight to age distribution, having the actual values goes a longer way. Some age grades are more receptive of the idea of eating out than others. Other factors like safety index are also beneficial to a business owner, as this can help you decide on business hours without fear of harm to workers or damage to properties. All these are factors that can be considered in a further study.

The analysis done in this project was a top-down approach. The boroughs were first narrowed down before focusing on a select number of neighbourhoods, and this left out the neighbourhoods of every other borough, even though there's a small chance some neighbourhood in other boroughs could fit our target better than that which we've selected. Hence, in a further study, one could follow a bottom-up approach, and bear in mind that this approach could get considerably more time consuming and complicated with a significant increase in data.

It can also be noticed that some boroughs are significantly closer to each other than other boroughs, so despite these boroughs not meeting up the expectation separately, they could meet up when combined, which is also another area that can be covered in further studies.

5. Conclusion

The purpose of this project was to help a hypothetical stakeholder select an optimal location in Toronto to open a restaurant business. This was done by gathering data on the existing boroughs and neighbourhoods in each borough and using the said data to model the different locations. The various locations were then clustered and the optimal neighbourhood in the optimal borough was selected. From all the outlined neighbourhoods in all the boroughs in Toronto, Canada, after carrying out a two level selection process that involves the use of k-means clustering in picking out the optimal borough, and then selecting the optimal neighbourhood in that borough, it was found that 'Downsview Northwest, North York, Toronto, Canada' was the most likely optimal neighbourhood to open a restaurant.

In a similar manner, this same process could be carried out by a variety of other businesses for a variety of other purposes besides opening a new business spot or expanding. This is the overall aim of this project.