The Battle of the Neighbourhoods

Applied Data Science Capstone Project

Sean Large

1 The Problem

1.1 Introduction

Toronto is the most populous city in Canada with a recorded population of nearly 3 million people. It is the capital city of the province of Ontario and is widely recognised as one of the most multicultural and cosmopolitan cities in the world. As an international centre of business, finance, arts and culture, Toronto is extremely popular with both tourists and residents alike.

1.2 Business Problem

My client is the chief executive officer (CEO) of a large retail clothing business. As a budding data scientist, my client has asked me to decide on the most suitable neighbourhood in Toronto to open a new store. My client has stressed that the key to retail success on the high street boils down to four factors; great products, attentive customer service, consistently high foot-fall, and convenient parking.

While the products themselves and customer service are not my responsibility, I can leverage the Foursquare API to ensure that the recommended neighbourhoods have busy streets and nearby parking locations. Additionally, I will be clustering Toronto's neighbourhoods by popular venues to determine which of them can be considered hot-spots for retail outlets and eateries.

2 The Data

2.1 Required data

The following data will be required to provide an accurate recommendation to my client:

- A list of Toronto's neighbourhoods, with latitude and longitude coordinates, calculated by geopy's Nominatim
- A list of the most popular venues for each postal code region retrieved via the Foursquare API
- A list of suitable car parks for each postal code region retrieved via the Foursquare API
- A list of total populations for each postal code region retrieved via Statistics Canada [1]

2.2 Assumptions

There will be some assumptions made to keep this project relatively simple. Firstly, I am making the assumption that all postal-codes cover the same land area and hence population density will have a direct linear relationship with total population. Additionally, I am making the assumption that while

total populations may have changed since this data was curated (2016), postal-code population sizes will be largely similar in relation to each other. Furthermore, it will be assumed all car parks retrieved from the Foursquare API are deemed suitable and that only one car park is required to meet the parking criteria outline in the first section.

3 THE METHOD

3.1 GATHERING NEIGHBOURHOOD DATA

By scraping Wikipedia for postal code data and reading in a csv file of longitude/latitude coordinates, I was able to construct a Pandas dataframe of Neighbourhood information. The first 5 rows looks like so:

	Postal Code	Borough	Neighbourhood	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

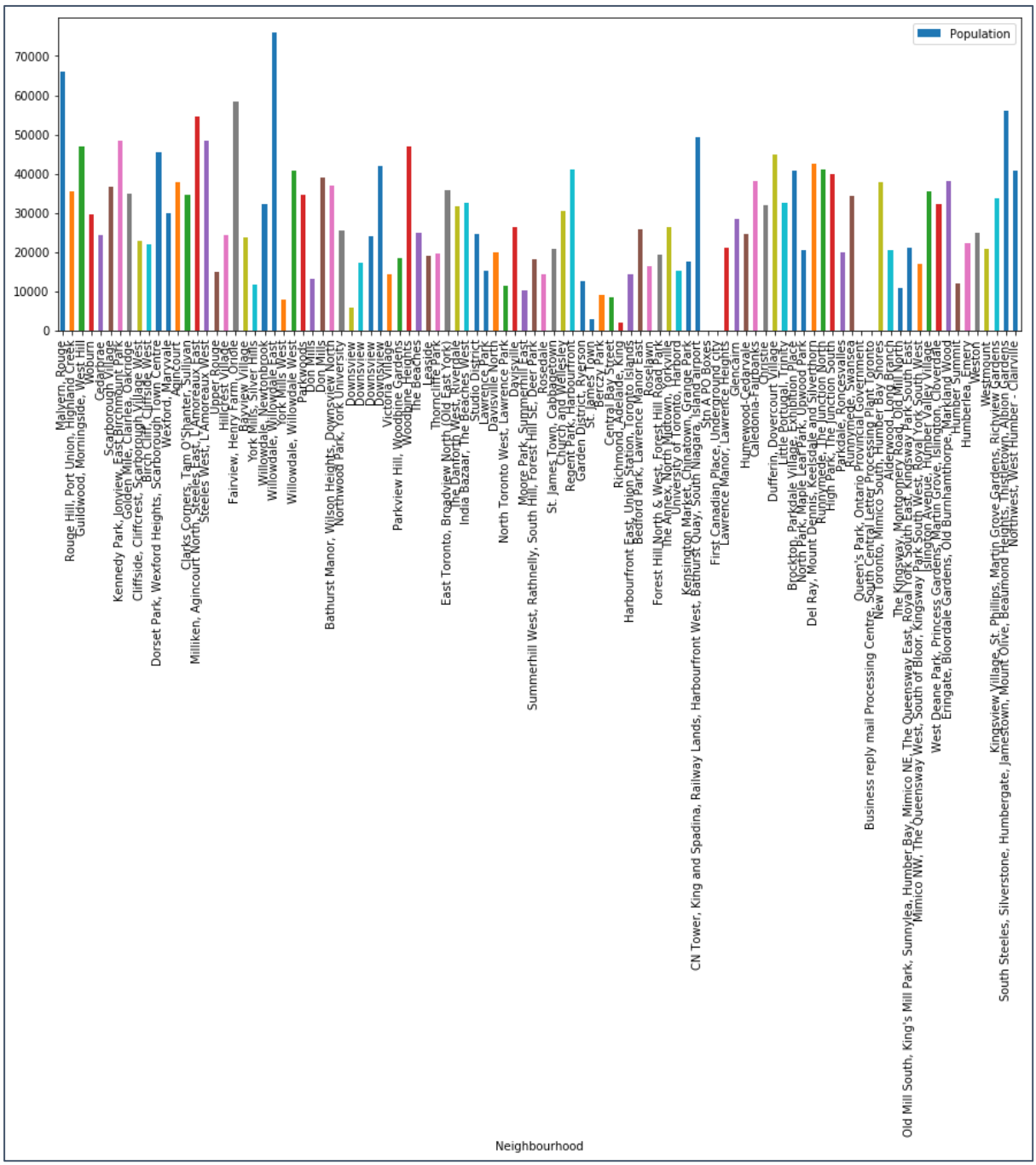
Additionally, I was able to obtain population statistics for individual postcode regions from the 2016 Toronto Census. Of course, I am making the assumption that population data has not changed since then, however I believe that this assumption is justified as I would not expect population proportions to have changed drastically.

I then appended the population to the data to produce the following dataframe:

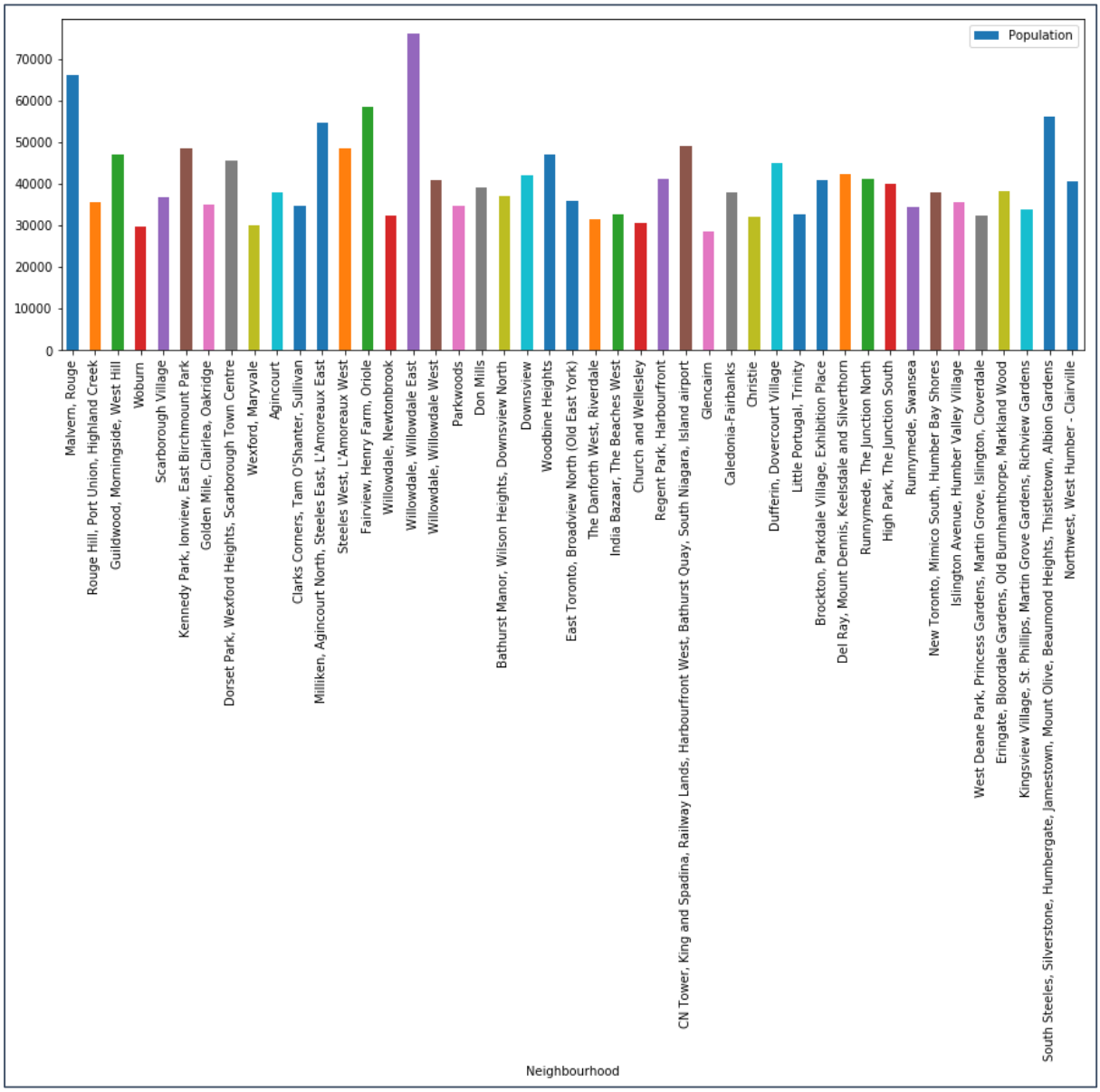
	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Population
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	66108
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	35626
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	46943
3	M1G	Scarborough	Woburn	43.770992	-79.216917	29690
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	24383

3.2 POPULATION AND PARKING CRITERIA

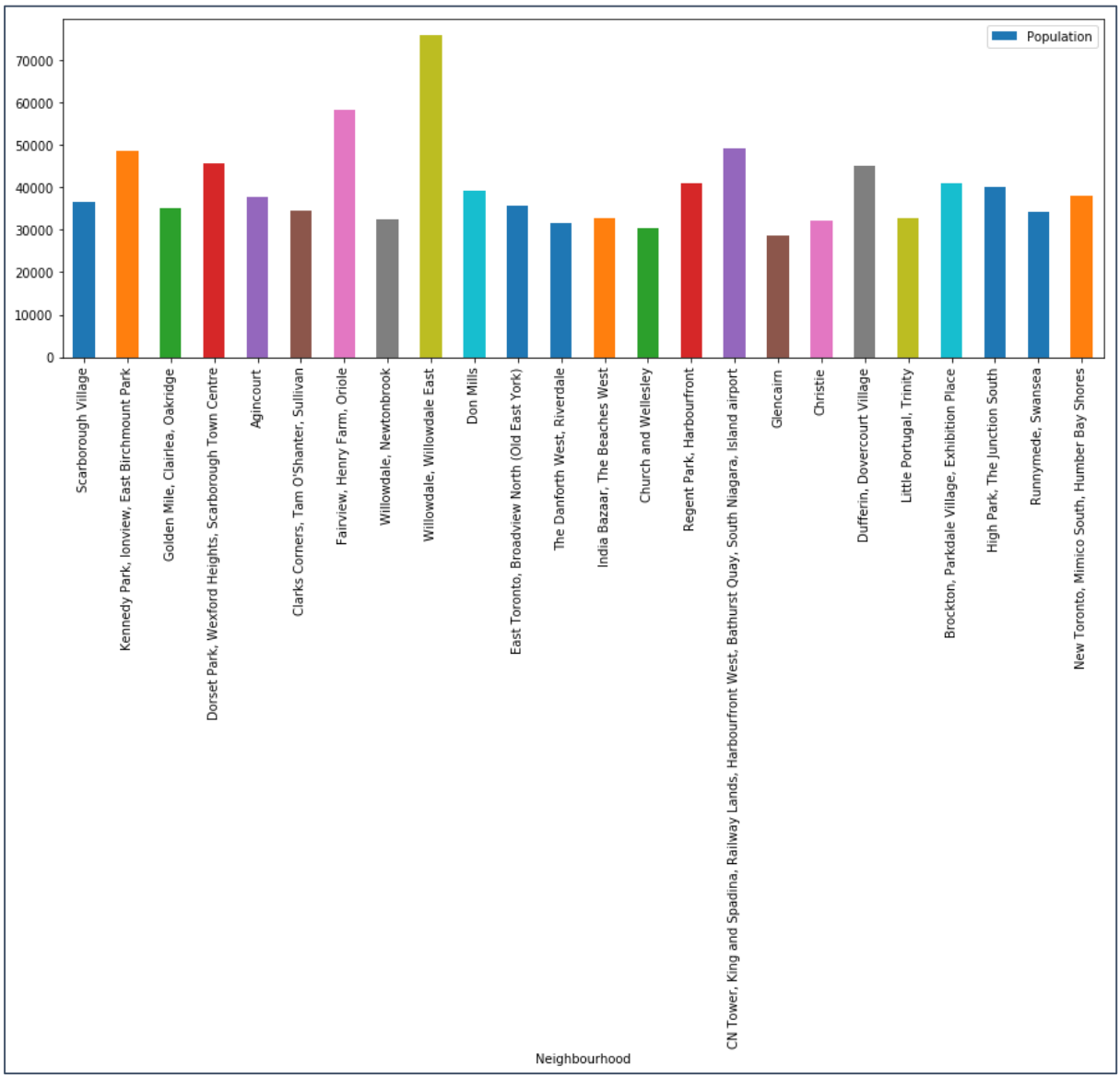
By plotting the population data in a simple bar chart, we can see that population varies drastically between postcodes:



Our client's domain expertise in the retail industry has told us that only areas with a high population should be considered when opening a new store. Consequently, we can remove all neighbourhoods with a population below the mean. We are then left with the following 45 regions:



In addition to the population criteria, our client has also expressed the importance of parking proximity. Customers will be more inclined to visit shops and carry more bags if there car is close by. To ensure that our final recommendation includes parking, we can filter these 45 regions by leveraging the Foursquare API. As a result, we are left with the remaining 24 regions.



3.3 Popular Venues

Now we have narrowed down our potential neighbourhoods to just 24 candidates, we can use the Foursquare API once more to obtain the 100 most popular venues for each. This results in the following dataframe:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Scarborough Village	43.744734	-79.239476	Diamond Pizza	43.743699	-79.245922	Pizza Place
1	Scarborough Village	43.744734	-79.239476	Tim Hortons	43.738992	-79.238961	Coffee Shop
2	Scarborough Village	43.744734	-79.239476	Dairy Queen	43.739506	-79.236894	Ice Cream Shop
3	Scarborough Village	43.744734	-79.239476	Dairy Queen	43.739580	-79.236991	Ice Cream Shop
4	Scarborough Village	43.744734	-79.239476	Subway	43.738284	-79.236792	Sandwich Place

By performing one-hot encoding, grouping rows by neighbourhood and taking the mean frequency of each venue, we obtain the following dataframe:

	Neighbourhood	Accessories Store	Afghan Restaurant	Airport	Airport Lounge	American Restaurant	Animal Shelter	Antique Shop	Art Gallery	Arts & Crafts Store	...	Trail	Train Station	Turkish Restaurant	Vegetarian / Vegan Restaurant
0	Agincourt	0.00	0.00	0.000000	0.000000	0.023256	0.0	0.0	0.00	0.00	...	0.00	0.0	0.0	0.00
1	Brockton, Parkdale Village, Exhibition Place	0.01	0.00	0.000000	0.000000	0.010000	0.0	0.0	0.01	0.02	...	0.01	0.0	0.0	0.02
2	CN Tower, King and Spadina, Railway Lands, Har...	0.00	0.00	0.066667	0.066667	0.000000	0.0	0.0	0.00	0.00	...	0.00	0.0	0.0	0.00
3	Christie	0.00	0.00	0.000000	0.000000	0.010000	0.0	0.0	0.01	0.00	...	0.00	0.0	0.0	0.02
4	Church and Wellesley	0.00	0.01	0.000000	0.000000	0.010000	0.0	0.0	0.01	0.01	...	0.00	0.0	0.0	0.01

4 RESULTS

With the above dataframe, we are now in a position to perform k-means clustering to separate the neighbourhoods into distinct clusters based on similarity of their venues. By examining these clusters and determining their categories of venues, we will be in a strong position to recommend several potential regions to our client.

Upon performing k-means clustering using 7 distinct clusters, we obtain the following:

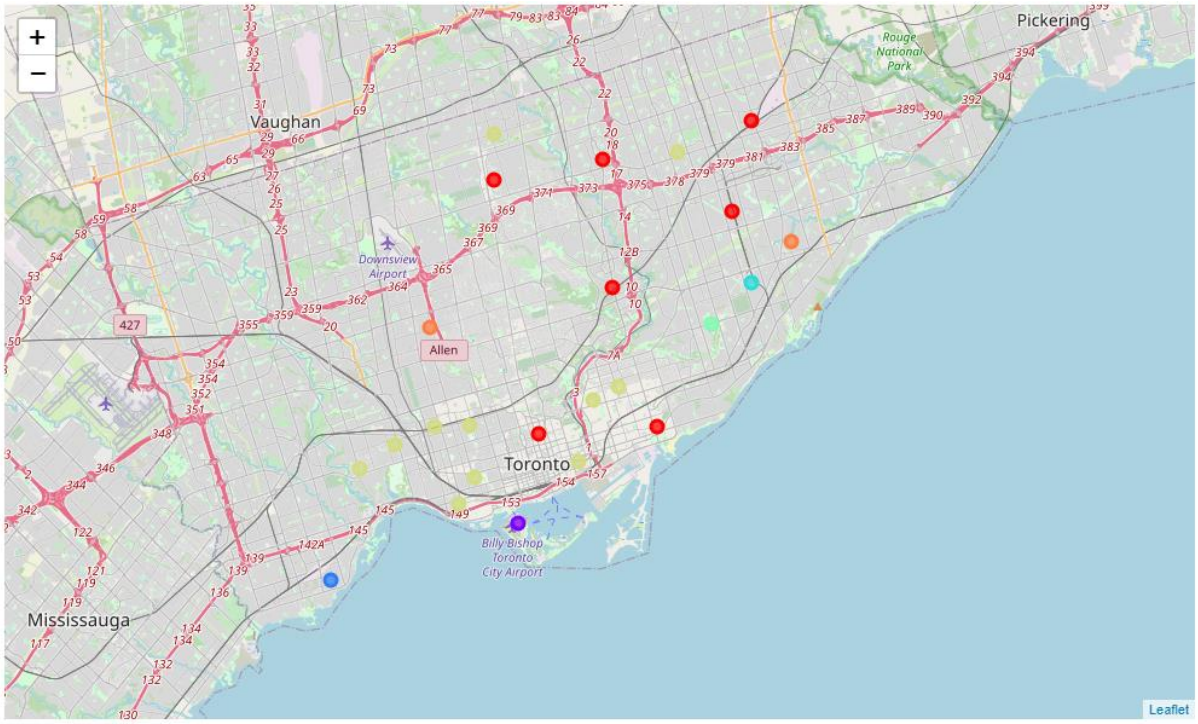


Figure 1: Cluster 0 (red), 1 (purple), 2 (blue), 3 (cyan), 4 (green), 5 (yellow), 6 (orange)

At a cursory glance, the map above suggests that in general, neighbourhoods in cluster 5 are in close proximity to Toronto city centre which may be convenient. Additionally, cluster 1 is separated from the mainland and hence may not be convenient for some.

By examining the individual clusters, we find that in general, cluster 5's most popular venues tend to be restaurants, cafés and gourmet shops. In general, this is good for footfall but there is little

indication that retail outlets are popular venues in these regions. On the other hand, many retail venues in cluster 0 are shown to be popular. This includes shopping malls and clothing stores.

5 DISCUSSION & CONCLUSION

Given the results above, I would make the following recommendation to my client. Neighbourhoods in cluster 0 appear to show the strongest evidence of having popular retail outlets. These neighbourhoods include Dorset Park, Wexford Heights, Scarborough Town, Agincourt, Fairview, Henry Farm, Oriole, Willowdale, Willowdale East, Don Mills, India Bazaar, The Beaches West and Church and Wellesley.

Of these neighbourhoods, those in the borough's of Scarborough and North York appear most popular. In particular, for the neighbourhoods Fairview, Henry Farm, Oriole, their second most popular venues are clothing stores. Additionally, those neighbourhoods have the second highest population of the regions examined. As their location is quite distant from Toronto city centre, we might also expect overhead costs to be cheaper than otherwise.

Of course, there are other factors that may be required to make a final decision such as costs, availability of retail space, target audience etc but I believe the analysis undertaken in this report should provide a robust indication of regions that will host a successful retail outlet.