Srishti Arora

Selecting location for a new health cafe in North York, Canada

1. Introduction:

When it comes to deciding a neighbourhood to position a new cafe, the important aspect to consider is who is the target audience and which neighbourhood would provide the maximum accessibility to that audience.

Let's say that the new cafe is a health cafe serving high protein, low calorie drinks and snacks. The best neighbourhood for such a cafe would be one with maximum number of health and training centres, gyms and other workout places and minimum number of other cafes.

It is not feasible to suggest a single neighbourhood to the entrepreneur since we are not aware of her budget and the property prices in that specific neighbourhood. So, instead we will suggest the neighbourhoods to be preferred based on number of fitness centres and neighbourhoods to be avoided based on number of cafes/restaurants. We will also create clusters of neighbourhoods based on their similarity in terms of number of cafes, types of restaurants, number of workout studios, etc. This will give entrepreneur the flexibility to select a neighbourhood based on her requirements. In the end we will visualize these clusters on the map.

2. Data Description and Collection:

Data for this project is collected and merged from three different sources – Wikipedia, Geospatial Coordinates csv and Foursquare API. We perform the following steps in order and generate the required dataset for analysis

a. Scrape list of postal codes of Canada from Wikipedia to get Postal Code, Borough, and Neighbourhood details
b. Read the Geospatial Coordinates csv to get latitude and longitude details corresponding to each postal code
c. Merge the datasets obtained in a) and b)
d. Fetch top 100 venues that are within a radius of 500 meters to all the neighbourhoods in North York using Foursquare API based on latitude and longitude details collected in c)

The data collected from Foursquare API contains Neighbourhood, Neighbourhood Latitude, Neighbourhood Longitude, Venue, Venue Latitude, Venue Longitude, and Venue Category details. Once we have all this information, we will fetch the neighbourhoods which have fitness related venues, and which have the largest number of eateries. This will give a good estimation of preferred and undesirable locations for a new health café in North York. To provide flexibility to the entrepreneur in terms of budget and the property prices, we performed clustering on the neighbourhoods to provide a list of locations similar to the preferred (undesirable) ones.

3. Methodology:

Our aim in this project is to pick out the most and least favourable locations for setting up a new health café in North York, Canada. The first step in this process is to collect all neighbourhoods in the North York borough and their latitude and longitude details. In order to do so, we fetch the postal codes, boroughs and neighbourhoods in Canada by web scrapping Wikipedia and merge it with the Geospatial Coordinates csv which contains latitude and longitude details of each postal code. We subset the North York Borough data from this merged dataset and create a data frame called *neighYork.* The first 5 rows of *neighYork* are:

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
3	M3B	North York	Don Mills	43.745906	-79.352188
4	M6B	North York	Glencairn	43.709577	-79.44507

Next task is to fetch the venue details in each neighbourhood so as to understand the market demand. Neighbourhoods which already have a lot of restaurants, cafes and other eateries are unlikely to welcome a new and unconventional café, whereas neighbourhoods which are more inclined towards fitness and house a number of fitness centres and sports related venues are expected to provide the target audience for the new café.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop

We exploit the Foursquare API to explore the neighbourhoods and fetch top 100 venues that are within a radius of 500 meters to all the neighbourhoods in North York. This will give us ample data to do a preliminary location analysis. This analysis will involve two different approaches. First is the regex search for eateries and fitness related venues and second involves clustering of neighbourhoods based on venue types and density. Based on these two methods, we will try to answer three questions listed below. Before jumping into the analysis, it would be better to understand what data is returned from Foursquare API and how best it can be used. This data is placed in a data frame named *york_venues* and we display the first five rows above for reference.

a. Which neighbourhoods to prefer?

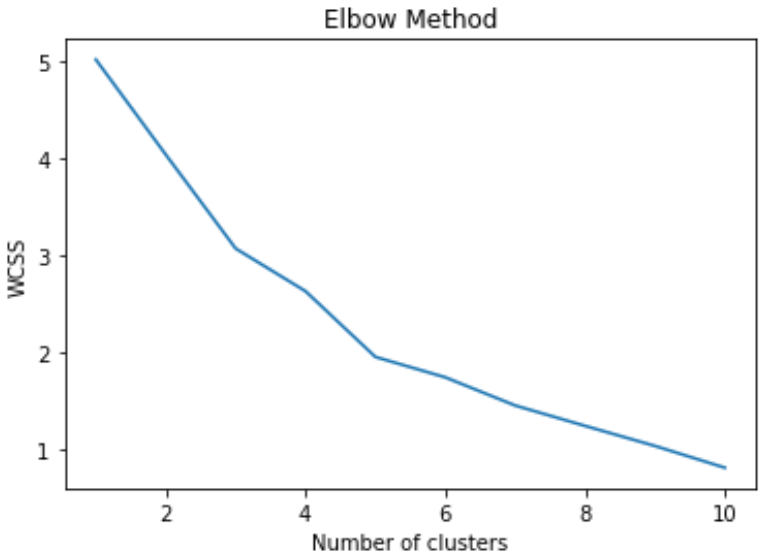
Thinking from business perspective, an entrepreneur who is venturing into a health café business, would like to figure out where can he find customers who will prefer health over money or taste. There is a high probability that someone who is visiting a gym, fitness centre or sports related venues is health conscious. Although we understand that this is not the sole criterion to classify someone as health vigilant, but this definitely increase the chances. Hence, as an owner I would like to know which neighbourhoods have more such venues. So, we run a regex to subset such neighbourhoods and make more informed recommendations.

b. Which neighbourhoods to avoid?

Not just market demand but also competition is important to consider otherwise the sales might get cannibalised and business might not flourish. Even though we are talking about opening a health café, availability of multiple alternatives to eat at will reduce the footfall at the café irrespective of type of eateries in the vicinity. In order to avoid unnecessary competition, we find out the neighbourhoods which already have a lot of eating joints and recommend the entrepreneur against selecting those neighbourhoods. Similar to a), we apply a regex and list the neighbourhoods that should be avoided while considering a plausible location.

c. Alternate neighbourhoods

As also mentioned earlier, we want to give the owner more flexibility and alternate options other than the recommendations made in a) and b), reason being we are ignoring the property prices and operational costs in any particular neighbourhood and our recommendations might not fit into the entrepreneur's budget. This involves clustering the neighbourhoods based on their similarity with respect to presence of particular venue category and frequency of occurrence of each category. We used the standard elbow method to determine the number of clusters for k-means clustering and then perform k-means clustering with the optimal number of clusters which is 7 in this case as can be seen from the plot below.



4. Results:

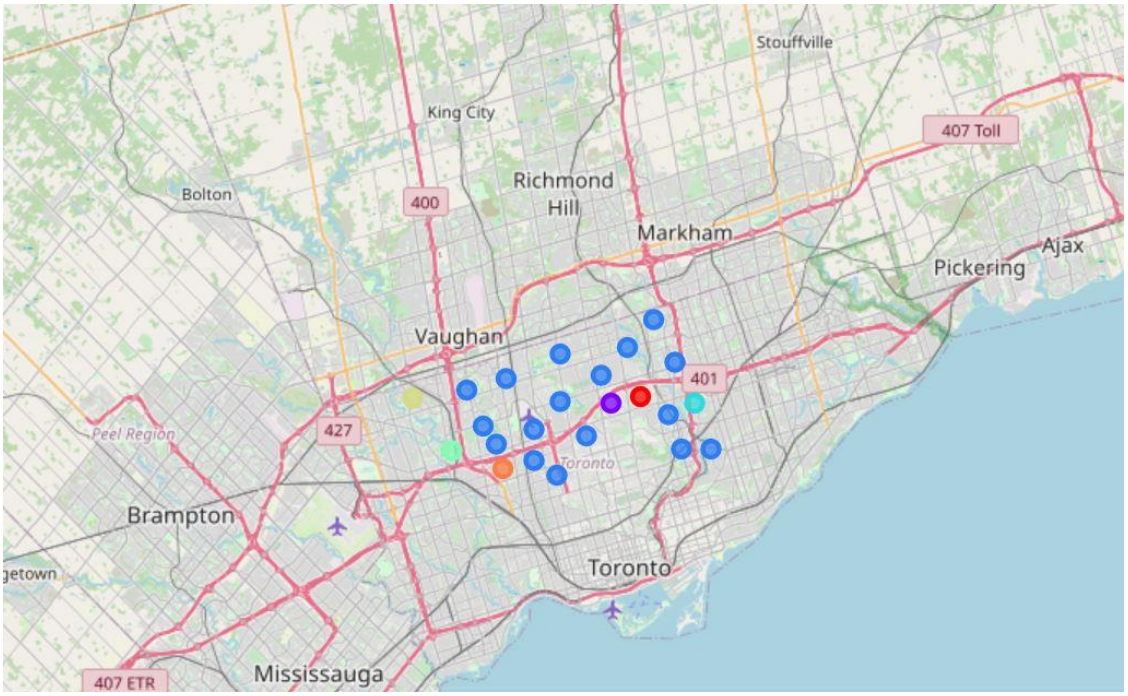
a. Top five neighbourhoods to be preferred based on number of fitness related venues
   - Don Mills
   - Downsview
   - Fairview, Henry Farm, Oriole
   - Hillcrest Village
   - Humberlea, Emery

b. Five neighbourhoods to be avoided based on number of cafes/restaurants
   - Fairview, Henry Farm, Oriole
   - Willowdale, Willowdale East
   - Bedford Park, Lawrence Manor East
   - Don Mills
   - Bathurst Manor, Wilson Heights, Downsview North

Based on the above results **Downsview**, **Hillcrest**, **Humberlea** and **Emery** (set a. minus set b.) are the top four neighbourhood recommendations for opening health café based on our analysis

c. In case the above four are not suitable because of some factors that are not considered here, we do provide alternate options. Refer the Jupyter notebook that is shared to get the full results. A snapshot of our clustering result is presented below. Based on the most preferred neighbourhood from the ones recommended above, the owner can look for other options in the same cluster and make a more informed decision also based on frequency of occurrence of each venue category.

	Postal Code	Borough	Neighborhoods	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	3	Food & Drink Shop	Park	Women's Store	Diner	Coffee Shop
1	M4A	North York	Victoria Village	43.725882	-79.315572	2	Coffee Shop	Pizza Place	Hockey Arena	Portuguese Restaurant	Intersection
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	2	Clothing Store	Women's Store	Vietnamese Restaurant	Boutique	Coffee Shop
3	M3B	North York	Don Mills	43.745906	-79.352188	2	Gym	Beer Store	Coffee Shop	Japanese Restaurant	Caribbean Restaurant
4	M3C	North York	Don Mills	43.725900	-79.340923	2	Gym	Beer Store	Coffee Shop	Japanese Restaurant	Caribbean Restaurant



5. Discussion:

There are certain limitations to this study, and we would like to remind the readers that we could not consider all the important factors while making recommendations due to lack of data availability. Other three most important features that should be considered while selecting a location are:

- Property prices
- Population density
- Operational costs

An important aspect of any good analysis is to be aware of its limitations and scope of improvement. Even though a data scientist is not able to include all components in his work, due to any kind constraints, it is critical that he is aware of the shortcomings and is able to highlight the same to the end users so that the recommendations made are at least harmless, even if not useful.

6. Conclusion:

The main aim of this analysis was to present the capabilities of data science. If we understand the problem statement well and know how to play with data, merge different data sources, clean and format data as per the requirements then 80% of the work is done. Data speaks a lot and beautiful insights can be generated with properly formatted data. Lastly, it is not only important to do great technical analysis, ability to write comprehensible reports that can be understood by non-technical people is also an important component to make your work worthwhile.