Applied Datascience Capstone Project

Final Report - The Battle of Neighbourhoods

By How Chih Lee

March 12, 2021

Criteria: This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

In this week, you will continue working on your capstone project. Please remember by the end of this week, you will need to submit the following:

1. *A full report consisting of all of the following components (**15 marks**):*
   - *Introduction where you discuss the business problem and who would be interested in this project.*
   - *Data where you describe the data that will be used to solve the problem and the source of the data.*
   - *Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.*
   - *Results section where you discuss the results.*
   - *Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.*
   - *Conclusion section where you conclude the report.*
2. *A link to your Notebook on your Github repository pushed showing your code. (**15 marks**)*
3. *A presentation or blogpost on your Github presenting your results. (**10 marks**)*

---

1. Introduction

1.1 Background

We are a company that provides Software as a Service (SaaS) to small and medium sized enterprises (SMEs) in the Greater Toronto Area (GTA). One of our key client groups are restaurant and cafe owners (RCO) in the Food and Beverage (F&B) sector. They come to us for business consultancy services that leverage off our Big Data and Machine Learning capabilities.

They want our advice on (i) where to locate their restaurant or cafe and (ii) how to promote it to ensure success.

In this case, we have a client (Client V) who was a sous-chef working at a famous restaurant in Manhattan, New York. She is returning to Toronto where she grew up and wants to open up her own restaurant or cafe in the GTA. We interviewed her and she has certain requirements she wants us to fulfil:

(1) Client V wants to open a vegan restaurant that serves organic food for breakfast and lunch, but not dinner.
(2) her menu price point is CAD15 to 30 per customer per meal.

(3) she wants to appeal to health conscious customers who go to fitness centres, yoga studios or pools nearby and are choosing to try vegan for health reasons.
(4) she wants to appeal to environmentally conscious customers who care about excessive carbon emissions generated by the meat industry and are choosing to try vegan for environmental reasons.
(5) she believes that repeat clients from the neighbourhood are stickier customers than clients who have to travel a long distance to her restaurant.

For each neighbourhood in the GTA, we provide basic data on the different types of venues, including restaurants and cafes, that operate there, including but not limited to, addresses, contact details, category of venue, type of F&B etc. which are provided to our RCO clients for free. Using this data, we assess:

(1) what competitive businesses there are in the area for the same type of cuisine
(2) what symbiotic businesses there are in the area from a different type of venue

In addition to the basic data, we provide a paid version for our services tailored to each client where we also:

(1) analyse the average spending power of local residents in the neighbourhood, which are often the first customers to try out a new restaurant or cafe in their locale
(2) identify surrounding venues of interest, like fitness centres, yoga studios or pools where digital and traditional forms of advertising for our clients, and which can be an important source of foot traffic from the neighbourhood
(3) monitor the different ratings and reviews from Key Opinion Leaders (KOL) active in that neighbourhood with large followings, and invite them to promotional events for our clients, like tastings and cocktails, to promote the restaurant or cafe

1.2 The Problem

Location: Traditionally, RCO clients relied on local knowledge and word of mouth to decide where to open a new restaurant or cafe. Many RCOs open up a new restaurant in a neighbourhood they are familiar with or where they know enough customers who will jump with them to their new restaurant. They do not explore outside familiar neighbourhoods and they do not have the resources to do market research, customer segmentation and wallet analysis. For example, if a bubble tea cafe opens at a specific locale and does well, many copy-cat bubble tea cafes quickly follow, driving up rents and driving down profits. The failure rate for starting a new restaurant or cafe business is high and profit margins are low.

With the power of crowd-sourced data on platforms like Foursquare and combining that with other compiled data like average spending power of residents in the neighbourhood, we can apply data analytics to help our clients make better business decisions where to locate their restaurant. For example, if an upscale supermarket opens in a neighbourhood with above average spending power, opening a boutique cafe next door may better capture foot traffic from customers of the supermarket than opening a Tim Horton's fast-food restaurant.

Promotion: Traditional advertising was centred on a local columnist visiting the restaurant or cafe and writing a good review in the local newspaper or magazine. Some restaurants would even pay popular actors/local celebrities to dine at their restaurant and take photographs with them for publicity. With the crowd-sourced platforms like Foursquare

and social media, good reviews by KOLs which have built up large followings of “foodies” can matter more than traditional advertising. Passing out fliers at the corner of the block can be replaced by email advertising sent straight to the Inbox of customers or by location-based messaging sent straight to smartphones of passer-byes. Working on joint promotions with surrounding venues of interest which attract customers of a similar profile can be more effective than billboard advertising to the general public.

1.3 The Solution

The preparation and serving of good food hasn’t changed much, but choosing a Location and Promotion of a good restaurant or cafe has changed greatly. Our RCOs clients have an enormous need for data analytics but few have the resources to dedicate to these functions. Client V has very specific requirements and our company can fulfil her needs with SaaS. Part 1 identifies several Locations for her vegan restaurant/cafe in the GTA. Part 2 details a Promotion plan. We hope to overcome the Problem and improve her chances of launching a successful F&B business through data analytics/machine learning.

2. Data

2.1 Data Sources

Data is sourced from multiple channels. Foursquare provides location based data on different sorts of venues obtained through crowd-sourcing. Foursquare indexes each location through the longitude and latitude of that location, and provides the data based on the endpoints selected. Regular endpoints are provided free and include basic venue firmographic data, category of venue, and venue ID. Premium endpoints require payment and include rich content such as user ratings, URLs, photos, tips, menus, hours of operation, etc. Endpoints classifications and links can be found here https://developer.foursquare.com/docs/places-api/endpoints/

To solve our Location Problem, we use “search?” and “explore?” functions to call up data on the category of venue, type of cuisine, geographical coordinates, URL links to websites, menus and pricing of competing F&B outlets in the neighbourhood. In Toronto, neighbourhoods are classified by Postal Code, but Foursquare does not map Toronto locations to their postal codes. We separately obtain Postal Codes for each neighbourhood from the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, After that, we use pgeocode API to get the geographical coordinates of each Postal Code before adding the Postal Codes to the dataframe downloaded from Foursquare. See Fig 1.

```
103 103 103
```

Out[14]:

	Postal Code	Borough	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude
2	M3A	North York	Parkwoods	43.7545	-79.3300
3	M4A	North York	Victoria Village	43.7276	-79.3148
4	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626
5	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6641	-79.3889
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.6662	-79.5282
9	M1B	Scarborough	Malvern, Rouge	43.8113	-79.1930
11	M3B	North York	Don Mills	43.7450	-79.3590
12	M4B	East York	Parkview Hill, Woodbine Gardens	43.7063	-79.3094
13	M5B	Downtown Toronto	Garden District, Ryerson	43.6572	-79.3783

Fig 1 - GTA Postal Codes with Geographical Coordinates

For the average spending power of GTA residents, this is difficult to obtain as income tax returns are private and confidential in Canada. We use average housing prices in the different neighbourhoods as a proxy for average spending power as economic studies have shown that higher housing prices have a strong correlation to higher spending power in any given neighbourhood [See *House Price and Income Inequality in Canada by Canada Mortgage and Housing Corporation, Nov 2020*][1]. Historical average housing prices are indexed by Postal Codes and can be obtained by scraping data from https://housepricehub.com.

To solve our Promotion Problem, we use the "explore?" function to call up Other Venues under the category "Gyms" that attract customers who are health conscious, but we found it hard to identify a venue category for customers who care about the environment. These Other Venues will be approached to become Promotion Partners for joint advertising to target customers for Client V.

We also use the "tips?" function to call up data from Foursquare endpoints that contain tips and reviews from customers and KOLs, User ID of KOLs, their names and contacts, and also the number of agreeCount or disagreeCount from their followers.

2.2 Data Cleaning

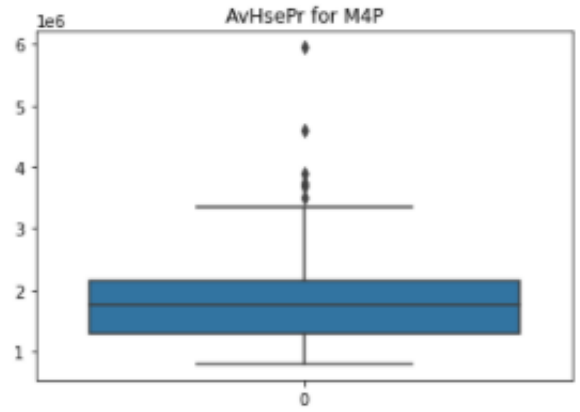
Postal Code data from Wikipedia shows 103 postal codes starting with M, which is reserved for the GTA, whereas data on historical average housing prices from housepricehub.com shows only 99 postal codes starting with M. Further inspection revealed that certain postal codes have not been assigned, whilst others are assigned by Canada Post to high-traffic areas like a Business Processing Centre in East Toronto and a Gateway sorting facility, etc. These postal codes have no residential homes and therefore no average housing prices. See Fig 2.

168	M7Y	East Toronto	Business reply mail Processing Centre, South C...
169	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
170	M9Y	Not assigned	Not assigned
171	M1Z	Not assigned	Not assigned
172	M2Z	Not assigned	Not assigned
173	M3Z	Not assigned	Not assigned
174	M4Z	Not assigned	Not assigned
175	M5Z	Not assigned	Not assigned
176	M6Z	Not assigned	Not assigned
177	M7Z	Not assigned	Not assigned
178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...
179	M9Z	Not assigned	Not assigned

Fig 2. Postal Codes Not Assigned

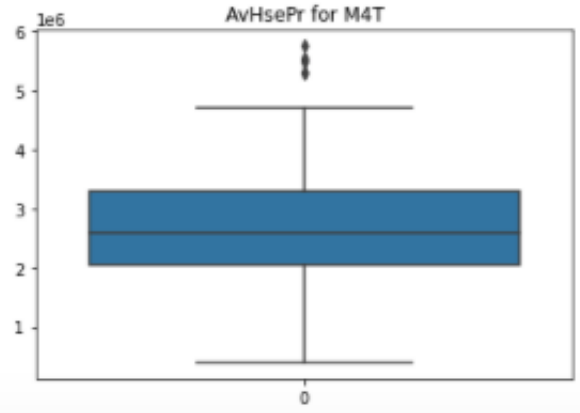
[1] https://www.cmhc-schl.gc.ca/en/data-and-research/publications-and-reports/research-insight-house-price-income-inequality-canada

```
Text(0.5, 1.0, 'AvHsePr for M4P')
```




```
sns.boxplot(data=AHP7nprice)
print(AHP7nprice.shape)
plt.title('AvHsePr for M4T')
```

```
(143,)
Text(0.5, 1.0, 'AvHsePr for M4T')
```



Further inspection of housepricehub.com reveals that different Postal Codes have different numbers of entries due to difference in popularity of neighbourhoods. The historical average price needs to be inferred from tables with enough entries (>100) such that sample sizes are large enough to return meaningful average housing prices and there is no skew.

Tables are embedded in different webpages, one for each Postal Code, so scraping data from each webpage requires cycling through several url links. Average housing prices for different Postal Codes are then stored in different dataframes as string objects. We convert these to numerical objects to return the means and normalise the data across different Postal Codes. We create box plots to visualise the distribution of Average House Prices, disregarding outliers at the high-end and low-end, which can lead to misleading results. See Fig 3.

Fig 3. BoxPlot of Locations

For Part 2 - Promotion, certain venues have identical names, which could mean they belong to the same chain store, like Goodlife Fitness. These are kept in the database and distinguished using their geographical coordinates because different outlets of the same chain may have different tips and reviews from their customers.

Tips data can often produce a null return, resulting in “No Ratings" for the Venue. This could be because Venues in Toronto were not popular enough with users of Foursquare to generate enough real-time reviews and tips to allow for the scraping of reliable data. We cycle through venueID and userID of several KOLs to find the KOL with the largest following and best agreeCounts for the Location.

3. Methodology

3.1 Feature Selection and Analysis for Part 1 - Location

After Data Cleaning, there are 99 neighbourhoods in the GTA with more than 2,150 venues and more than twenty features of which about eight were selected for their relevance to solving Part 1 - Location of our Problem.

```
(95, 253)
----Agincourt----
                       venue  freq
0  Latin American Restaurant   0.2
1                  Newsagent   0.2
2            Breakfast Spot   0.2
3           Badminton Court   0.2
4              Skating Rink   0.2
5       Monument / Landmark   0.0
6                    Museum   0.0
7             Moving Target   0.0
8             Movie Theater   0.0
9       Moroccan Restaurant   0.0


----Alderwood, Long Branch----
                        venue  freq
0                 Pizza Place  0.17
1          Convenience Store   0.17
2                        Gym   0.17
3             Sandwich Place   0.17
```

The features relevant to Part 1- Location are: Name, Neighbourhood, Postal Code, Geographical Coordinates, Average Housing Prices, Venue Category, Type of F&B. With the specifications of Client V in mind, we use OneHot encoding to analyse the Venue Category and do a venue count for the 10 most common Venue Category per Neighbourhood. We rank the Venues by the most common Venue Category and return a dataframe with the 5 Most Common Venues for each Postal Code. See Fig 4 and Fig 5.

Fig 4. Top 10 Venues by Neighbourhood

```
(99, 11)
```

click to unscroll output; double click to hide

	Postal Code	Borough	Neighbourhood	Neighbourhood Latitude	Neighbourhood 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	43.7545	-79.3300	0	Park	Food & Drink Shop	Yoga Studio	Eastern European Restaurant	Flea Market
1	M4A	North York	Victoria Village	43.7276	-79.3148	1	French Restaurant	Hockey Arena	Park	Pizza Place	Financial or Legal Service
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626	1	Coffee Shop	Breakfast Spot	Restaurant	Yoga Studio	Distribution Center
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504	1	Clothing Store	Coffee Shop	Women's Store	Restaurant	Cosmetics Shop
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6641	-79.3889	1	Sushi Restaurant	Gym	Italian Restaurant	Ramen Restaurant	Burrito Place

Fig 5. OneHot Venue Category

We use the unsupervised machine learning method of K-means clustering (with number of clusters = 10) to return Cluster Labels for each Neighbourhood. Clustering has the effect of grouping Neighbourhoods with the greatest overlap in Most Common Venues and puts these into the same Cluster Label.

Using the 'Folium' visualisation library, we create a map of GTA, and plot the different Neighbourhoods on the map, colour-coding them according to their Cluster Labels. We cycle through different Clusters to identify the Cluster that is most beneficial to solving our Location problem. See Fig 6.

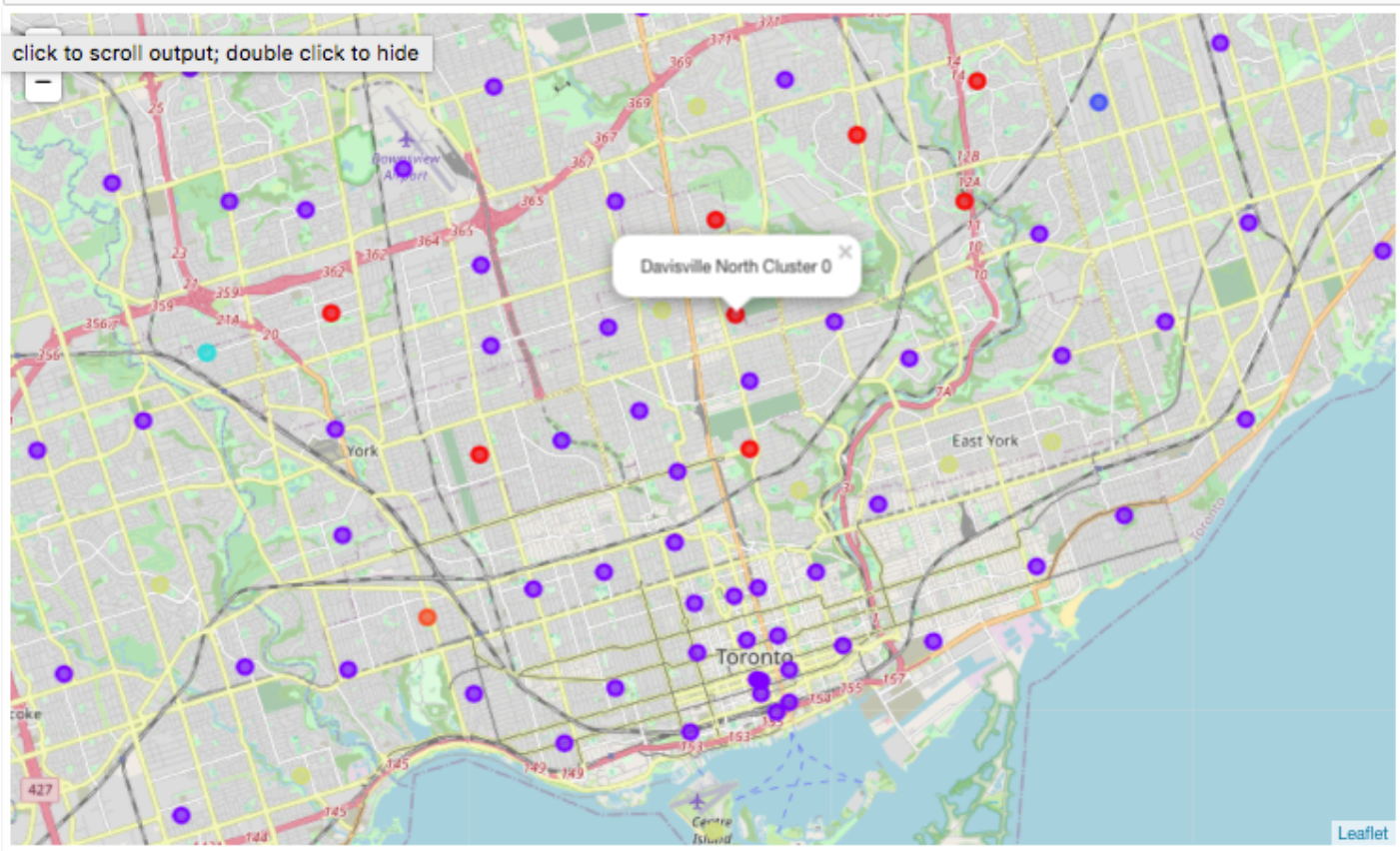


Fig 6. GTA Cluster Map

We arrive at a target Cluster (Cluster 0, colour = red) which appears to have the most symbiotic businesses and no competing businesses. The target Cluster has several Postal Codes distributed across several Neighbourhoods in the GTA. We return the Average Housing Prices for these Postal Codes and disregard the Neighbourhoods with the three highest Average Housing Prices and the three lowest Average Housing Prices. In certain exclusive areas, multi-million dollar mansions dominate the neighbourhood and skew Average Housing Prices to the high side. However, this is not a good indicator for Location as there are no commercial spaces for rent in such exclusive neighbourhoods due to zoning restrictions. On the low side, the customers may not have the spending power that Client V is targeting. We narrow down to two Postal Codes and recommend these as the solutions to Part 1 - Location of our Problem. See Fig 7.

	AvHsePr	Postal Code	Borough	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
66	1.840344e+06	M4P	Central Toronto	Davisville North	43.7135	-79.3887	0	Food & Drink Shop	Breakfast Spot	Park	Department Store	Dog Run
81	2.763354e+06	M4T	Central Toronto	Moore Park, Summerhill East	43.6899	-79.3853	0	Gym	Park	Playground	Thai Restaurant	Grocery Store

Fig 7. Two Recommended Locations

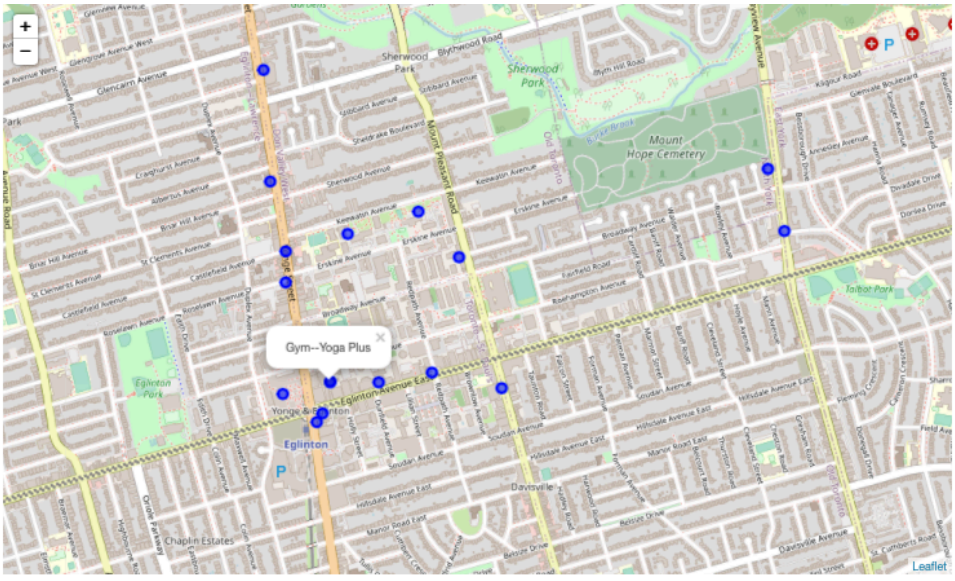
3.2 Feature Selection and Analysis for Part 2 - Promotion

The features relevant to Part 2 - Promotion are: Tips and Reviews, Agree and Disagree Counts, User ID and User Name of Reviewers, Venue ID and Contacts of Other Venues identified through the search query = “Gym”, which encompasses gyms, fitness centres, yoga studios and swimming pools.

We notice from the GTA Clusters Map that the two recommended Locations happen to be physically adjacent to each other. We decide to use one Location and expand the search to a bigger radius instead of searching two Locations. We use the function “explore?” within a 1200 meter radius to identify 17 Other Venues with symbiotic businesses. See Fig 8.

	name	categories	address	crossStreet	lat	lng	labeledLatLngs	distance	formattedAddress	postalCode	
0	Crossfit Metric	Gym	756 Mt Pleasant Rd	Eglinton	43.707480	-79.389857	[{'label': 'display', 'lat': 43.70747956557104...	676	[756 Mt Pleasant Rd (Eglinton), Toronto ON, Ca...	NaN	524ac4b1498e8498a174a
1	GoodLife Fitness Toronto Dunfield and Eglinton	Gym	110 Eglinton Ave E	at Dunfield Ave.	43.707645	-79.395303	[{'label': 'display', 'lat': 43.707645, 'lng':...	840	[110 Eglinton Ave E (at Dunfield Ave.), Toront...	M4P 1A6	4bd6e35d637ba5933ad3f
2	Barreworks	Yoga Studio	2576 Yonge St	NaN	43.714070	-79.400109	[{'label': 'display', 'lat': 43.71407030751952...	920	[2576 Yonge St, Toronto ON, Canada]	NaN	51606062e4b0878cf540f
3	GoodLife Fitness Toronto Yonge Eglinton Centre	Gym	2300 Yonge St	at Eglinton Ave. W	43.707276	-79.399562	[{'label': 'display', 'lat': 43.707276, 'lng':...	1115	[2300 Yonge St (at Eglinton Ave. W), Toronto O...	M4P 1E4	4e80de0930f8844fb2bd7
4	Gym	Gym	140 Erskine	NaN	43.713126	-79.393537	[{'label': 'display', 'lat': 43.71312601210131...	391	[140 Erskine, Toronto ON, Canada]	NaN	4da99b34a86e771ea70e8
5	900 Mount Pleasant - Residents Gym	Gym / Fitness Center	900 Mount Pleasant Road	NaN	43.711671	-79.391767	[{'label': 'display', 'lat': 43.71167058860572...	319	[900 Mount Pleasant Road, Toronto ON M4P 3J9, ...	M4P 3J9	4c3f2724db3b1b8d635e6
6	Yoga Tree Midtown	Yoga Studio	40 Eglinton Ave. E	at Yonge St.	43.707642	-79.397472	[{'label': 'display', 'lat': 43.70764167668336...	960	[40 Eglinton Ave. E (at Yonge St.), Toronto ON...	M4P 3A2	4efa4c786c25c411edcaf
7	Womens Fitness Clubs of Canada	Gym	1820 Bayview Ave Unit 1	NaN	43.712484	-79.377341	[{'label': 'display', 'lat': 43.71248383171391...	920	[1820 Bayview Ave Unit 1, Toronto ON M4G 4G7, ...	M4G 4G7	4bc6fb8b92b376b091d54
8	Yoga Plus	Gym	40 Eglinton Ave E	at Yonge	43.707674	-79.397478	[{'label': 'display', 'lat': 43.707674, 'lng':...	958	[40 Eglinton Ave E (at Yonge), Toronto ON, Can...	NaN	4b156d91f964a520b9ac2
9	CYCLEBAR	Gym	1866 Bayview Avenue, Suite 103	NaN	43.714468	-79.378066	[{'label': 'display', 'lat': 43.71446846114807...	862	[1866 Bayview Avenue, Suite 103, Toronto ON M4...	M4G 0C3	58cd67d92f91cb28ec3d6
10	Anytime	Gym	2739 Yonge	NaN	43.717654	-79.400434	[{'label': 'display', 'lat':	1051	[2739 Yonge St, Toronto ON M4N	M4N 2H9	53bdcf7a498ed37c87ab9

Fig 8. List of Other Venues



Using the ‘Folium' visualisation library, we create a map of the target Location and plot the 17 Other Venues on the map (colour=blue), attaching labels with their Venue Category and Names. We notice from the Map of Other Venues that there’s a high concentration around major crossroads in within 1200m radius. See Fig 9.

Fig 9. Map of Other Venues

With the VenueID, we cycle through different Venues to identify the Venue with the highest User Rating because good ratings tend to return supportive customers who are more likely to support joint promotions by Client V. Conversely, a Venue with poor ratings will tend to return fewer repeat customers who are not likely to support joint promotions by that Venue. Using the function “tips?” we return KOLs with large followings who are active in that neighbourhood.

We identify the User ID of the KOLs and analyse the agreeCount and disagreeCount, as well as the text of their tips on Other Venues to identify KOLs who are well regarded by their “followers”. See Fig 10.

	text	agreeCount	disagreeCount	id
0	Beef Bolognaise, white chocolate score cheesecake = SUPREME. Try the grapefruit lager. Best this summer #foodie #saucy	1	0	5399350211d2b5f1381bdd3f
1	So much fun! Gotta come back with more friends #dodgeball #foampitshouldhavemorelanes	0	0	530041ad11d24ed4822a569e
2	Came out of the parking garage & found a better parking area that's a lot closer to the path. #morningglory	1	0	4f915933e4b048b2e456c9cd
3	Try the 30 day unlimited classes for $40...great new studio.	1	0	4f302be9e4b07ca31811dc13
4	Get here early to avoid the long lines.	1	0	4f1eb59ce4b0d881fd7d6a10
5	Everything I need in one place...good place to get all your vegetables	0	0	4f0fa53fe4b06c5888c1b0f6
6	Must try the Truffle Brie and the 12yr old chedder...WOW!	0	0	4ef3cbd649010be35db4d44d
7	Technique is king! Great instructors.	0	0	4ee03ffcb8f741383026f77d
8	Good whiskey, good vodka and good vibes are welcome here.	0	0	4eb71bd5f5b94bd85c88fbd4
9	it's not just a donut shop, it's Canada's donut shop	1	0	4e71ea07aeb79ea0952497bd
10	Get your ThomasCook Visa TravelMoney Card here at TC Financial Services	0	0	4e6f7cd7d1647b1137f87094
11	Pleased to announce the launch of www.thomascookagent.ca, if you're an agent, come visit	0	0	4e6f7be4fa765666a2ec553d
12	Great place for a #TIFF11 party	1	0	4e6ef6082271c30fff3183fd
13	Sit at the bar. Great bourbon sour, it's true.	0	0	4e6ef2ad1838a9b627711262
14	Great view...but where are the wait staff?	1	0	4e6ef1b245dd49e0f19327a0
15	Where to get vegetables	0	0	4e547bb388770c1e522b7d4e
16	Great Monthly Unlimited promo $70. Clean beds & standup and people working there are really nice.	0	0	4dbc1a6243a1d8504b817770
17	If you come after 5pm there's no line up. Open till 8pm M-F	0	0	4db9de46fa8c2e303f1674f6
18	It is practice that guests visiting Jude's Joint bring Whiskey and/or Guinness although most alcohol will suffice.	0	0	4cca0603e47f5481eb53c5f7
19	Must get the french onion soup, bacon wrapped scallops & prime rib medium rare.	2	0	4c1d60f6b9f876b027927d46
20	I got $1 off my however-you-want-it frappuccino cause foursquare said I was the Mayor here. YAY! Good till June 28/10...get yours	0	0	4bf5945370c603bbee489cb4
21	I must admit, the machanic's are quick and let me hang out to show me what needed to be done and why. Not every CT let's you so this though	0	0	4bf3117c70c603bbd2069cb4
22	Greatest service, makes you want to come back again & again. Every dish is fantastic but do leave room for an avocado or mango shake. Party in your mouth!	0	0	4be6424070c603bb6ac99ab4
23	Get all your travel needs here. Also ask about Thomas Cook travel Insurance & our exclusive four currency ATM's	0	0	4be1811770c603bb08539ab4

Fig 10. KOL Tips with Highest AgreeCount

Not all of the Other Venues have Ratings and not all Users have good followings. It takes data analysis to filter and choose the best Promotion Partners from Other Venues and top KOLs active in the neighbourhood. We then approach these Promotion Partners (using their names and contact details scraped from the json files) to maximise the impact of every advertising dollar for Client V.

4. Results

From 99 Postal Codes and more than 2,150 Venues in the GTA, we managed to identify one Postal Code (M4P) to solve Part 1 - Location and the best Other Venue (Gym - Yoga Plus) and the top KOL (Judes C) to solve Part 2 - Promotion of our Problem.

(1) *Client V wants to open a vegan restaurant that serves organic food for breakfast and lunch, but not dinner.* ==>There are no restaurants/cafes with competing cuisine in this Location, and this was provided to Client V for free.
(2) *her menu price point is CAD15 to 30 per customer per meal.* ==> The Average Housing Price for the Neighbourhoods in Postal Code M4P is above the mean for Toronto, which should offer customers with above average spending power in this Location, and this was provided to Client V for free.
(3) *she wants to appeal to health conscious customers who go to fitness centres, yoga studios or pools nearby and are choosing to try vegan for health reasons*. ==> There are 17 Other Venues within 1200m radius of this Location that attract health conscious customers and we identified Gym Yoga Plus as the most effective one with the highest user ratings. We charged for these recommendations and contacts.
(4) *she wants to appeal to environmentally conscious customers who care about excessive carbon emissions generated by the meat industry and are choosing to try vegan for environmental reasons.*==> We did not do so well to identify Other Venues that meet this criteria.
(5) *she believes that repeat clients from the neighbourhood are stickier customers than clients who have to travel a long distance to her restaurant.*==> We identified Judes C as a KOL active in the Location with a good following that can help draw customers to Client V. We charged for this recommendation and contact.

5. Discussion

Notwithstanding our significant results set out above, we notice certain limitations when using data analysis and machine learning to solve the Problem. Most of these are associated with the Data Source and significant effort needs to be put into identifying the right source. For example, we could not identify environmental conscious customers and may need to expand our search beyond location based databases like Foursquare, to include social media like Facebook or Reddit, which maintain databases according to interest groups and blogger opinions.

We also noticed the bias inherent in clustering techniques like K-means Clustering. We understand the coding in the library uses least square of Euclidean distance techniques to group samples around centroids. This function finds a local minimum but may miss the global minimum. This is evident when we set the number of clusters differently, and we get different results and finding the right “k” is more an art than a science. There may also be limitations to machine learning in other circumstances where we want to group samples according to their differences and not similarities. We understand many so called “black swan” events are missed due to this limitation of machine learning that looks for statistical similarities and miss the differences.

6. Conclusion

We have met 4 out of the 5 criteria set out by Client V and provided her with some advice that is free and other good advice that she paid for, but we didn't charge for advice we couldn't give. We hope we've added significant value and wish her every success of her restaurant.

Good Luck Client V!