Applied Data Science Capstone: The Battle of the Neighbourhoods

Kishan Narotam
https://knarotam.github.io/
15/01/2021

I Introduction

A venture capitalist is known as a private equity investor who seeks opportunity in high-growth potential companies such as small businesses and startups in exchange for a stake in the respective company. The risk-reward factor of investing in these companies can be drastic, and if invested correctly can yield a substantial return for the investor. Venture capital firms began in the United States in the early to mid-1900s and has continued to grow exponentially as the world evolved through the Dot-Com burst and into the Fourth Industrial Revolution [1].

The popular television show Shark Tank and its respective spinoff such as Dragon's Den in the UK has brought to life the way investors, specifically venture capitalists invest their money in small businesses and startups. In order to mitigate the risks, investors must know about the business and what the plan would be to succeed.

This project aims to provide information to venture capitalists on what businesses are popular based on data about Toronto, Canada. The idea is to be able to make valid assumptions based on the popularity of certain venues in the various boroughs. The information gathered will allow venture capitalists to know what type of businesses are in high-demand as well as potential opportunities for less popular businesses.

II Data

Based on what we aim to achieve with this project, the data required includes:

- List of postal codes, corresponding boroughs and neighbourhoods for the City of Toronto.
- The demographics of the Toronto neighbourhoods.
- The various venues such as restaurants, bars, coffee shops, malls, etc. around each of the neighbourhoods.
- The longitude and latitude of each neighbourhood and venues.

A. *Data Sources*

The list of postal codes, boroughs and neighbourhoods are retrieved from a Wikipedia table listing all of the postal codes in Canada that begin with the letter M. This was chosen as the postal codes that begin with the letter M are the boroughs and neighbourhoods that are found within the city of Toronto. The original table is found at the link below:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This postal code data along with the corresponding boroughs and neighbourhoods will correlate directly with the geospatial data file. This data file provides the longitude and latitude of each postal code that is stored in a *csv* file. The link to the file can be found at the link below:

http://cocl.us/Geospatial_data

The demographics data is taken from the Toronto Open Data Catalogue, relating to neighbourhood profiles. The csv file consists of the neighbourhood profiles from a census done in 2016, which includes population distribution across various races and religions, languages spoken, immigration and citizenship, education and finances. This file is provided as a *csv* file and can be found at the link below:

https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv

These above links alongside the Foursquare API will be used to map the neighbourhoods and retrieve the data relating to the various venues. For this project, due to the limitations of the free account on Foursquare, the search limit of the venues is set to 100 with a radius of 500 metres of each neighbourhood.

III METHODOLOGY

A. Importing the Libraries

Various libraries will be used through the implementation of this project, with the main ones being `pandas` and `numpy` to handle the data itself. The `geopy`, `folium` and `requests` libraries will handle longitude and latitude conversion, JSON handling and map rendering respectively. The `Sci-kit learn` library gives us access to the k-means clustering model for our project execution and analysis. Lastly, the `BeautifulSoup` library will extract data from the respective HTML pages and allow us to use that data in a dataframe for analysis and modelling.

B. Importing the Data Sources

1) Toronto Neighbourhoods Data

In order to obtain this data from the Wikipedia page, the `get` function is used to request the page and convert it to raw HTML text. Using the `BeautifulSoup` library the table can be identified and converted into raw HTML text. Following this, the table is then read as an HTML file and converted into a dataframe for processing. Figure 1a shows the original Wikipedia table whereas Figure 1b shows the same data after being converted to a dataframe.

Postal Code	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront

(a) Wikipedia table view

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

(b) Extracted dataframe of wikipedia table

Fig. 1: Toronto Neighbourhoods Data tables

2) Preprocessing the Neighbourhoods Data

With the data having been imported correctly, it now must be preprocessed before any modelling and analysis can be done. The first step is to drop all rows from the table where the boroughs are *Not assigned*. The next step is to assign all *Not assigned* neighbourhoods the value of their respective boroughs. Figure 2 shows the table after being preprocessed and ready for assigning the respective longitude and latitude values to each location.

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
5	M9A	Etobicoke	Islington Avenue, Humber Valley Village
6	M1B	Scarborough	Malvern, Rouge
7	M3B	North York	Don Mills
8	M4B	East York	Parkview Hill, Woodbine Gardens
9	M5B	Downtown Toronto	Garden District, Ryerson
10	M6B	North York	Glencairn
11	M9B	Etobicoke	West Deane Park, Princess Gardens, Martin Grov...
12	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
13	M3C	North York	Don Mills
14	M4C	East York	Woodbine Heights
15	M5C	Downtown Toronto	St. James Town
16	M6C	York	Humewood-Cedarvale
17	M9C	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...
18	M1E	Scarborough	Guildwood, Morningside, West Hill
19	M4E	East Toronto	The Beaches

Fig. 2: Neighbourhoods dataframe after being preprocessed

3) Geospatial Data

The Geospatial data is the data that will provide the geographical coordinates to the neighbourhoods, specifically the centre point of each neighbourhood. The data is stores in a *csv* file, and is read in as such and converted into a dataframe. Subsequently, this data must be merged with the neighbourhoods dataframe for the data to be modelled and analysed. A left join is done on the neighbourhoods table with the geospatial data table and the resulting dataframe can be seen in Figure 3.

	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	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Fig. 3: Merged dataframe containing Neighbourhoods data

4) Demographics data (Used for in-depth analysis)

The demographics data is a *csv* file that was extracted from the Toronto open Data Catalogue and contained data from a 2016 census. With the amount of information in this file only two rows were extracted, specifically the population of the neighbourhoods and the average income of the neighbourhoods. This data was read in, preprocessed and merged with the Neighbourhoods dataframe. The resulting table can be seen in Figure 4.

	index	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Population	Average Income
0	1	M4A	North York	Victoria Village	43.725882	-79.315572	17,510	35,786
1	16	M6C	York	Humewood-Cedarvale	43.693781	-79.428191	14,365	65,274
2	19	M4E	East Toronto	The Beaches	43.676357	-79.293031	21,567	92,580
3	22	M1G	Scarborough	Woburn	43.770992	-79.216917	53,485	30,878
4	27	M2H	North York	Hillcrest Village	43.803762	-79.363452	16,934	40,442
5	29	M4H	East York	Thorncliffe Park	43.705369	-79.349372	21,108	28,875
6	32	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	16,724	32,913
7	39	M2K	North York	Bayview Village	43.786947	-79.385975	21,396	52,035
8	50	M9L	North York	Humber Summit	43.756303	-79.565963	12,416	30,731
9	64	M9N	York	Weston	43.706876	-79.518188	17,992	32,997

Fig. 4: Dataframe used for the in-depth analysis

C. *Creating the Map*

In order to generate an interactive map with points over each neighbourhood, the `folium` and `geopy` libraries are used. The `geopy` retrieved the centre coordinates of the city of Toronto, and the `folium` library is able to take the data from the dataframe as well as the location of Toronto and map it out accordingly as seen in Figure 5.

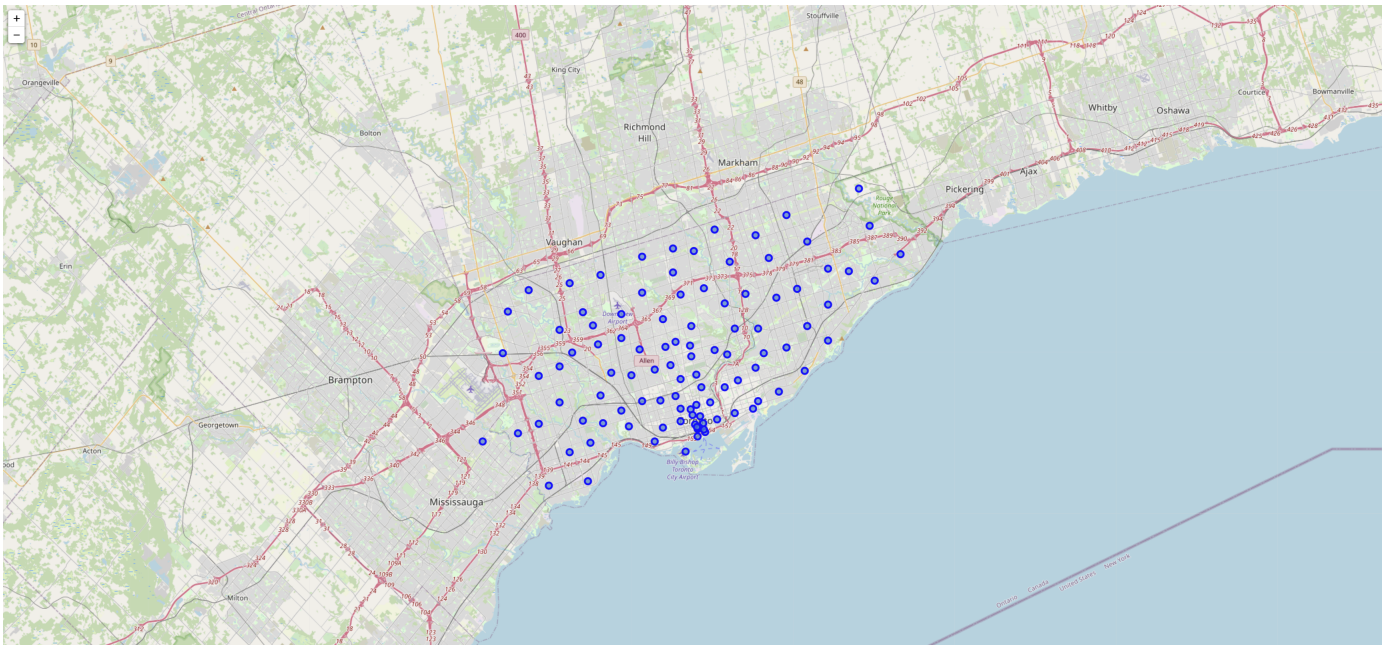

Fig. 5: Generated map of Toronto and respective neighbourhoods

D. *Retrieving the venues*

With half of the data ready for modelling and analysis, we can now extract the data required from the Foursquare API. The initial step is to set up your Foursquare credentials in order to retrieve the requested information. Once authenticated, 100 venues within a 500 metre radius of each neighbourhood. This data is requested as a JSON file and once received is converted into a dataframe which can be seen in Figure 6.

	Postal Code	Borough	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	M3A	North York	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	M3A	North York	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	M4A	North York	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	M4A	North York	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant
4	M4A	North York	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
5	M4A	North York	Victoria Village	43.725882	-79.315572	The Frig	43.727051	-79.317418	French Restaurant
6	M4A	North York	Victoria Village	43.725882	-79.315572	Eglinton Ave E & Sloane Ave/Bermondsey Rd	43.726086	-79.313620	Intersection
7	M4A	North York	Victoria Village	43.725882	-79.315572	Pizza Nova	43.725824	-79.312860	Pizza Place
8	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
9	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
10	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	Morning Glory Cafe	43.653947	-79.361149	Breakfast Spot
11	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center

Fig. 6: Dataframe of the neighbourhoods and requested venues

1) Grouping the Venues

All the venues within the requested radius has been presented, however in order to gain a full understanding of the data, the venues are grouped and counted. This is to observe what are some of the popular venues within the city of Toronto. The results can be seen in Figure 7 which amounts to a total of 273 unique categories of venues. Since no dictionary or classification is done on the venue categories, a caf and a coffee shop and a breakfast place are all categorised as unique venues, which for this project was an accepted trade off.

Neighbourhood	Postal Code	Borough	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Agincourt	5	5	5	5	5	5	5	5
Alderwood, Long Branch	8	8	8	8	8	8	8	8
Bathurst Manor, Wilson Heights, Downsview North	23	23	23	23	23	23	23	23
Bayview Village	4	4	4	4	4	4	4	4
Bedford Park, Lawrence Manor East	24	24	24	24	24	24	24	24
Berczy Park	58	58	58	58	58	58	58	58
Birch Cliff, Cliffside West	4	4	4	4	4	4	4	4
Brockton, Parkdale Village, Exhibition Place	23	23	23	23	23	23	23	23
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	17	17	17	17	17	17	17	17
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	14	14	14	14	14	14	14	14
Caledonia-Fairbanks	4	4	4	4	4	4	4	4

Fig. 7: Dataframe of the venues and the number of times it has appeared in the request

E. One Hot Encoding

One Hot Encoding is a process where categorical variables can be converted in order for a machine learning algorithm to process the data [2]. This converts the variables into a binary format, where a 0 indicates no occurrence and a 1 indicates an occurrence of that respective variable. Figure 8 shows the venues dataframe after being passed through the `onehot` function.

	Neighbourhood	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Baseball Stadium	Basketball Court	Basketball Stadium	B
0	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Parkwoods	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7	Victoria Village	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
9	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Fig. 8: One Hot Encoded dataframe

1) Calculating the average occurrence of each venue

Once processed and done, One Hot Encoding simply states if that categorical variable, in this case a venue, is present in that neighbourhoods radius. It does not give an indication of how many times that specific venue occurs in the radius. In order to achieve this, the neighbourhoods are grouped and the means of each venue is calculated as seen in Figure 9.

	Neighbourhood	Accessories Store	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage	Auto Workshop	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Baseball Stadium	Basketball Court	Basketball Stadium	Beach
0	Agincourt	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000
1	Alderwood, Long Branch	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.000000	0.000000	0.086957	0.000000	0.000000	0.00	0.00	0.000000	0.000000
3	Bayview Village	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.000000	0.000000	0.250000	0.000000	0.000000	0.00	0.00	0.000000	0.000000
4	Bedford Park, Lawrence Manor East	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.041667	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.00	0.000000	0.000000
5	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.017241	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0000	0.017241	0.034483	0.000000	0.000000	0.000000	0.00	0.00	0.017241	0.017241

Fig. 9: Grouped Neighbourhoods and means of respective venues

2) Presenting the top 10 venues of each neighbourhood

Now that the means are calculated for each venue for each neighbourhood,the goal is to retrieve the top 10 venues in each neighbourhood to allow investors to see what is the most popular venue categories and what would pose the most competition. Figure 10 displays an example of the top 10 venues of three different neighbourhoods. This raw text data can now be converted into a dataframe for clustering, which can be seen in Figure 11.

```
----Alderwood, Long Branch----
                      Venue  Frequency
0               Pizza Place      0.250
1                       Pub      0.125
2            Sandwich Place      0.125
3               Coffee Shop      0.125
4                       Gym      0.125
5                  Pharmacy      0.125
6                      Pool      0.125
7                      Park      0.000
8             Movie Theater      0.000
9  Mediterranean Restaurant      0.000
**********************************************


----Bathurst Manor, Wilson Heights, Downsview North----
                       Venue  Frequency
0                Coffee Shop    0.08696
1                       Bank    0.08696
2          Health Food Store    0.04348
3  Middle Eastern Restaurant    0.04348
4            Sushi Restaurant    0.04348
5                Supermarket    0.04348
6         Frozen Yogurt Shop    0.04348
7        Fried Chicken Joint    0.04348
8                Bridal Shop    0.04348
9             Sandwich Place    0.04348
**********************************************


----Bayview Village----
                  Venue  Frequency
0                  Bank       0.25
1   Chinese Restaurant       0.25
2  Japanese Restaurant       0.25
3                 Café       0.25
4  Monument / Landmark       0.00
5               Museum       0.00
6        Movie Theater       0.00
7                Motel       0.00
8  Moroccan Restaurant       0.00
9    Accessories Store       0.00
**********************************************
```

Fig. 10: Top 10 venues in various neighbourhoods

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Skating Rink	Lounge	Breakfast Spot	Latin American Restaurant	Clothing Store	Escape Room	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
1	Alderwood, Long Branch	Pizza Place	Pool	Pub	Coffee Shop	Gym	Pharmacy	Sandwich Place	Escape Room	Electronics Store	Eastern European Restaurant
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Health Food Store	Supermarket	Bridal Shop	Shopping Mall	Sandwich Place	Restaurant	Pizza Place	Pharmacy
3	Bayview Village	Japanese Restaurant	Bank	Chinese Restaurant	Café	Yoga Studio	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
4	Bedford Park, Lawrence Manor East	Coffee Shop	Sandwich Place	Thai Restaurant	Italian Restaurant	Café	Japanese Restaurant	Sushi Restaurant	Restaurant	Indian Restaurant	Juice Bar
5	Berczy Park	Coffee Shop	Cocktail Bar	Restaurant	Farmers Market	Beer Bar	Bakery	Cheese Shop	Seafood Restaurant	Sporting Goods Shop	Fish Market
6	Birch Cliff, Cliffside West	College Stadium	Café	General Entertainment	Skating Rink	Dumpling Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant
7	Brockton, Parkdale Village, Exhibition Place	Café	Bakery	Coffee Shop	Breakfast Spot	Convenience Store	Furniture / Home Store	Climbing Gym	Stadium	Bar	Italian Restaurant
8	Business reply mail Processing Centre, South C...	Yoga Studio	Auto Workshop	Garden Center	Garden	Light Rail Station	Fast Food Restaurant	Farmers Market	Comic Shop	Park	Pizza Place
9	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Harbor / Marina	Coffee Shop	Airport	Airport Food Court	Rental Car Location	Boat or Ferry	Sculpture Garden
10	Caledonia-Fairbanks	Park	Women's Store	Bar	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
11	Canada Post Gateway Processing Centre	Coffee Shop	Hotel	Intersection	American Restaurant	Fried Chicken Joint	Gas Station	Sandwich Place	Middle Eastern Restaurant	Burrito Place	Mediterranean Restaurant

Fig. 11: Dataframe of the top 10 venues in each neighbourhood

F. Clustering

The clustering used in this project is the *k-means* clustering. Five clusters were chosen for the categorisation in order to prevent underfitting and overfitting of the data given the concentration of the various neighbourhoods in the city of Toronto. Figure 12 shows the final processed dataframe after being passed through the machine learning algorithm.

	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	0	Park	Food & Drink Shop	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Escape Room
1	M4A	North York	Victoria Village	43.725882	-79.315572	4	Pizza Place	Hockey Arena	Intersection	French Restaurant	Portuguese Restaurant	Coffee Shop	Electronics Store	Escape Room	Eastern European Restaurant	Dumpling Restaurant
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	4	Coffee Shop	Park	Bakery	Café	Pub	Breakfast Spot	Restaurant	Theater	Performing Arts Venue	Brewery
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	4	Clothing Store	Furniture / Home Store	Accessories Store	Coffee Shop	Sporting Goods Shop	Event Space	Boutique	Vietnamese Restaurant	Coworking Space	Dog Run
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494	4	Coffee Shop	Sushi Restaurant	Yoga Studio	Mexican Restaurant	Italian Restaurant	Japanese Restaurant	Beer Bar	Smoothie Shop	Burrito Place	Sandwich Place

Fig. 12: Dataframe of all neighbourhoods, cluster groups and top 10 venues

IV RESULTS

A. Cluster Map

The neighbourhoods are all clustered, and in order to gain a better understanding on how they were clustered, a map is generated with each cluster being presented in a different colour. Figure 13 displays the map, where the following clusters correspond to the following clusters: gray is cluster 1; black is cluster 2; red is cluster 3; blue is cluster 4; purple is cluster 5.

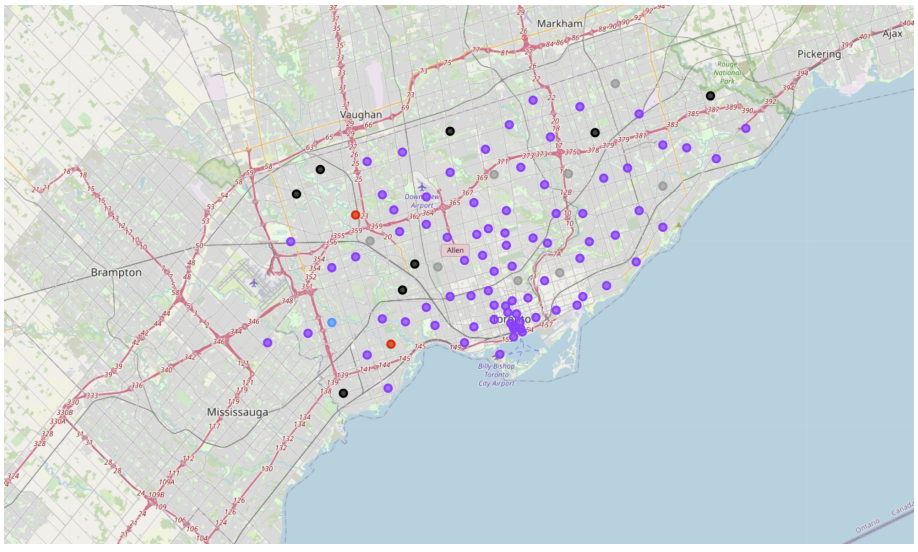

Fig. 13: Map showing the clusters

B. Cluster 1

If we analyse cluster 1, we can see that based on the most popular venue across all the neighbourhoods are parks and playgrounds. Across the top 10, it seems that all of these neighbourhoods have similarities as the rank of the venues decrease, for example the fifth to the 8th most common venues for the first four neighbourhoods have *Dog run*, *Doner Restaurant*, *Donut shop* and *Drugstore* in the same order.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	North York	0	Park	Food & Drink Shop	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Escape Room
21	York	0	Park	Women's Store	Bar	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
32	Scarborough	0	Playground	Jewelry Store	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Field
35	East York	0	Park	Metro Station	Convenience Store	Electronics Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
64	York	0	Park	Jewelry Store	Convenience Store	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
66	North York	0	Park	Convenience Store	Yoga Studio	Electronics Store	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
85	Scarborough	0	Playground	Park	Intersection	Concert Hall	Comic Shop	Farmers Market	Falafel Restaurant	Event Space	Ethiopian Restaurant	Escape Room
91	Downtown Toronto	0	Park	Playground	Trail	Yoga Studio	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop

Fig. 14: Table of cluster 1

C. Cluster 2

Cluster 2 has a *Pizza place* as it's most common venue in the various neighbourhoods. The various types of restaurants are spread across the rank of each neighbourhood, however almost all exist in each other, albeit at a lower or higher rank. An example of this would be the *Eastern European restaurant* being the third most common venue for the first two neighbourhoods, but the fifth most common venue for the fourth neighbourhood.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	Scarborough	1	Fast Food Restaurant	Yoga Studio	Eastern European Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
50	North York	1	Pizza Place	Yoga Studio	Eastern European Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
56	York	1	Fast Food Restaurant	Sandwich Place	Discount Store	Fried Chicken Joint	Restaurant	Donut Shop	Distribution Center	Dog Run	Doner Restaurant	Yoga Studio
63	York	1	Pizza Place	Brewery	Convenience Store	Grocery Store	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant
72	North York	1	Pizza Place	Grocery Store	Butcher	Coffee Shop	Pharmacy	Concert Hall	Discount Store	Falafel Restaurant	Event Space	Ethiopian Restaurant
82	Scarborough	1	Pizza Place	Gas Station	Noodle House	Chinese Restaurant	Fast Food Restaurant	Fried Chicken Joint	Bank	Italian Restaurant	Intersection	Thai Restaurant
89	Etobicoke	1	Pizza Place	Grocery Store	Video Store	Sandwich Place	Fast Food Restaurant	Beer Store	Fried Chicken Joint	Pharmacy	Golf Course	Department Store
93	Etobicoke	1	Pizza Place	Pool	Pub	Coffee Shop	Gym	Pharmacy	Sandwich Place	Escape Room	Electronics Store	Eastern European Restaurant

Fig. 15: Table of cluster 2

D. Cluster 3

Cluster 3 may not seem to have any similarities with regards to their venues, however these two neighbourhoods are clustered together due to the similarities of their location. Although not regarded as a venue, these two neighbourhoods are situated next to middle schools.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
57	North York	2	Baseball Field	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store	Filipino Restaurant
101	Etobicoke	2	Breakfast Spot	Business Service	Baseball Field	Yoga Studio	Electronics Store	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant	Escape Room

Fig. 16: Table of cluster 3

E. Cluster 4

Cluster 4 is somewhat of an outlier as no other neighbourhoods are similar to this, as it's most common venue is a *Filipino restaurant*. As no other cluster has this in common, this neighbourhood is clustered by itself.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
11	Etobicoke	3	Filipino Restaurant	Eastern European Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store	Health & Beauty Service

Fig. 17: Table of cluster 4

F. Cluster 5

Cluster 5 has the most neighbourhoods in it and shows blatant similarities in the most common venues. These neighbourhoods are clustered together as they are all high-density residential areas.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	North York	4	Pizza Place	Hockey Arena	Intersection	French Restaurant	Portuguese Restaurant	Coffee Shop	Electronics Store	Escape Room	Eastern European Restaurant	Dumpling Restaurant
2	Downtown Toronto	4	Coffee Shop	Park	Bakery	Café	Pub	Breakfast Spot	Restaurant	Theater	Performing Arts Venue	Brewery
3	North York	4	Clothing Store	Furniture / Home Store	Accessories Store	Coffee Shop	Sporting Goods Shop	Event Space	Boutique	Vietnamese Restaurant	Coworking Space	Dog Run
4	Downtown Toronto	4	Coffee Shop	Sushi Restaurant	Yoga Studio	Mexican Restaurant	Italian Restaurant	Japanese Restaurant	Beer Bar	Smoothie Shop	Burrito Place	Sandwich Place
7	North York	4	Gym	Coffee Shop	Restaurant	Japanese Restaurant	Beer Store	Asian Restaurant	Caribbean Restaurant	Italian Restaurant	Chinese Restaurant	Discount Store
8	East York	4	Pizza Place	Athletics & Sports	Café	Breakfast Spot	Flea Market	Bank	Intersection	Gastropub	Pharmacy	Gym / Fitness Center
9	Downtown Toronto	4	Coffee Shop	Clothing Store	Café	Bubble Tea Shop	Middle Eastern Restaurant	Cosmetics Shop	Hotel	Bookstore	Theater	Diner
10	North York	4	Pizza Place	Pub	Japanese Restaurant	Asian Restaurant	Bakery	Metro Station	Park	Escape Room	Electronics Store	Eastern European Restaurant
12	Scarborough	4	Bar	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store	Filipino Restaurant
13	North York	4	Gym	Coffee Shop	Restaurant	Japanese Restaurant	Beer Store	Asian Restaurant	Caribbean Restaurant	Italian Restaurant	Chinese Restaurant	Discount Store
14	East York	4	Skating Rink	Curling Ice	Park	Spa	Beer Store	Dance Studio	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
15	Downtown Toronto	4	Coffee Shop	Café	Cocktail Bar	American Restaurant	Gastropub	Restaurant	Lingerie Store	Farmers Market	Clothing Store	Seafood Restaurant

Fig. 18: Table of cluster 5

V DISCUSSION & IN-DEPTH ANALYSIS

For the in-depth analysis, 10 neighbourhoods were chosen and their respective populations and average income was extracted and combined into a single dataframe. These 10 cities will stand as an example to what can be done with this type of clustering and analysis. Figure 19 shows the dataframe of the 10 neighbourhoods while Figure 20 presents the previously clustered neighbourhoods and corresponding clusters.

	Postal Code_x	Borough_x	Neighbourhood	Latitude_x	Longitude_x	Population	Average Income	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M4A	North York	Victoria Village	43.725882	-79.315572	17,510	35,786	4	Pizza Place	Hockey Arena	Intersection	French Restaurant	Portuguese Restaurant	Coffee Shop	Electronics Store	Escape Room	Eastern European Restaurant	Dumpling Restaurant
1	M6C	York	Humewood-Cedarvale	43.693781	-79.428191	14,365	65,274	4	Field	Dog Run	Hockey Arena	Trail	Yoga Studio	Dumpling Restaurant	Distribution Center	Doner Restaurant	Donut Shop	Drugstore
2	M4E	East Toronto	The Beaches	43.676357	-79.293031	21,567	92,580	4	Health Food Store	Neighborhood	Trail	Pub	Drugstore	Discount Store	Distribution Center	Dog Run	Doner Restaurant	Donut Shop
3	M1G	Scarborough	Woburn	43.770992	-79.216917	53,485	30,878	4	Coffee Shop	Korean BBQ Restaurant	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
4	M2H	North York	Hillcrest Village	43.803762	-79.363452	16,934	40,442	4	Fast Food Restaurant	Dog Run	Pool	Mediterranean Restaurant	Athletics & Sports	Golf Course	Yoga Studio	Dumpling Restaurant	Doner Restaurant	Donut Shop
5	M4H	East York	Thorncliffe Park	43.705369	-79.349372	21,108	28,875	4	Indian Restaurant	Sandwich Place	Yoga Studio	Bank	Grocery Store	Gas Station	Fast Food Restaurant	Middle Eastern Restaurant	Discount Store	Park
6	M1J	Scarborough	Scarborough Village	43.744734	-79.239476	16,724	32,913	0	Playground	Jewelry Store	Dumpling Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Eastern European Restaurant	Field
7	M2K	North York	Bayview Village	43.786947	-79.385975	21,396	52,035	4	Japanese Restaurant	Bank	Chinese Restaurant	Café	Yoga Studio	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Eastern European Restaurant
8	M9L	North York	Humber Summit	43.756303	-79.565963	12,416	30,731	1	Pizza Place	Yoga Studio	Eastern European Restaurant	Distribution Center	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant	Electronics Store
9	M9N	York	Weston	43.706876	-79.518188	17,992	32,997	0	Park	Jewelry Store	Convenience Store	Yoga Studio	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	Drugstore	Dumpling Restaurant

Fig. 19: Dataframe of 10 neighbourhoods chosen for in-depth analysis

The main analysis would rather happen from the dataframe as the pertinent data is presented. If we take the second neighbourhood, *Humewood-Cedervale*, for example we can analyse the following:

- The neighbourhood is in the fifth cluster, meaning it is in a residential area.
- The top five most common venues are all healthy or sport related venues.
- There is no healthy restaurant or healthy food store in the top 10 most common venues.
- Given the smaller population size, but with the second highest average income in this list, it could stand that a venture capitalist may see an opportunity to invest in a healthy food store that already exists in the area to begin a small franchise.
- Increasing the number of same-branded healthy food stores can result in more people having access to the stores and make it potentially the sixth most common, if not one of the top five most common venues in the neighbourhood.

If we take the *Woburn* neighbourhood, with the largest population size in the list, and one of the lowest average income, we can deduce that this neighbourhood could be a potential location to open a fast food restaurant. However, in the borough of *Scarborough*, the population is potentially ethnically diverse given the various restaurants seen in the top 10 most common venues.

This implementation of this project can be drastically improved by utilising more data from the census file such as the ethnicity of people that live in the various neighbourhoods. Furthermore, the overall model

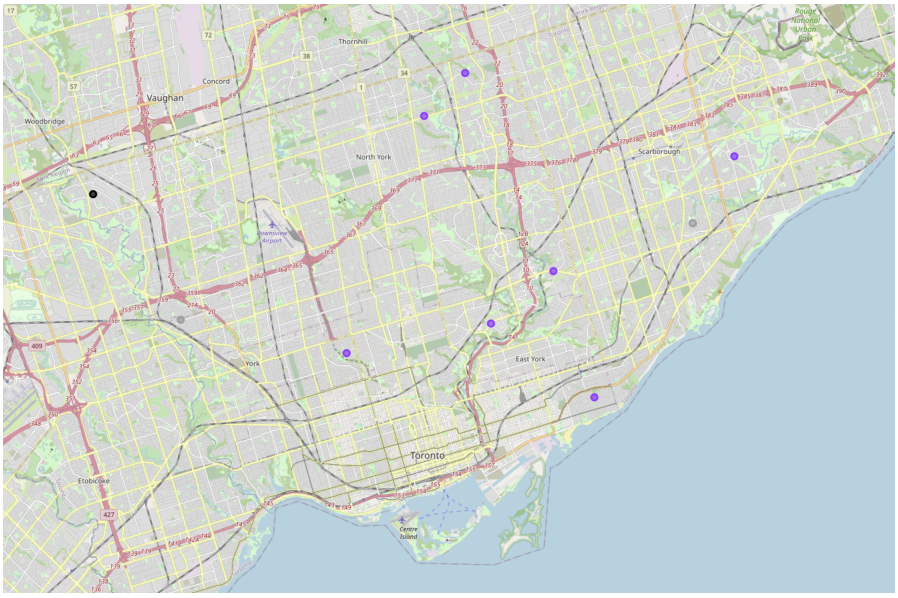

Fig. 20: Map of 10 neighbourhoods chosen for in-depth analysis

can be improved by finding the optimal number of clusters and with increased usage of the Foursquare API, a user can map each neighbourhood's population as a chloropleth, while adding markers to the various venues that are present.

VI CONCLUSION

A project was implemented on analysing data of the city of Toronto relating to its neighbourhoods. The data was sourced from various locations, preprocessed and presented accordingly. The data was then modelled and clustered using the k-means clustering machine learning algorithm, and an analysis was done on the results. An in-depth analysis was done given more information such as population and average income of the neighbourhood to answer the problem posed in this project. Overall, the project was successful but can stand to be improved with recommendations being given.

REFERENCES

[1] Ganti, A; *Venture Capitalist (VC) Definition*; https://www.investopedia.com/terms/v/venturecapitalist.asp#:~:text=A%20venture%20capitalist%20(VC)%20is,have%20access%20to%20equities%20markets; Last Accessed: 15/01/2021

[2] Vasudev, R; *What is One Hot Encoding? Why and When Do You Have to Use it?*; https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f; Last Accessed: 15/01/2021