CAPSTONE PROJECT

“BATTLE OF THE NEIGHBORHOODS”

Introduction

CITY OF TORONTO

The city of Toronto is the capital of Ontario and is the most popular city in Canada with a population that exceeds 2.5 million people as reported in 2016. The city is a bustling vibrant place that is a melting pot of diverse cultures and an international centre for trade, finance and business.

In this report, we will draw upon a database of location-based information from the foursquare database, namely venues, to develop a better understanding of the nuisances between different postcodes. Using this information we’ll develop a model to group postcodes into categories based on characteristical similarities and draw comparisons of likeness using a machine learning segmentation technique called k-means

This report will provide an understanding of which venues appear most frequently and whether there are correlations between frequently occurring venues and others. This would be particularly important for those looking to understand if certain venue category types are dependant on others.

City of Toronto “https://nugget.travel/podcast/forkids/toronto-canada-for-kids/”

Data

FOURSQUARE

For this analysis, we will be leveraging the location-based database from FourSquare. The FourSquare database is a comprehensive dataset containing over 60 million commercial venue points of interest world-wide and venue has over 900 categories with over 30 attributes where applicable.

Postal Code	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
M1C	Scarborough	43.784535	-79.160497	Fratelli Village Pizzeria	43.784008	-79.169787	Italian Restaurant
M1H	Scarborough	43.773136	-79.239476	Scaddabush Italian Kitchen & Bar	43.777460	-79.254358	Italian Restaurant
M1P	Scarborough	43.757410	-79.273304	Nova Ristorante	43.749434	-79.277630	Italian Restaurant
M1R	Scarborough	43.750072	-79.295849	Nova Ristorante	43.749434	-79.277630	Italian Restaurant
M1T	Scarborough	43.781638	-79.304302	Remezzo Italian Bistro	43.778649	-79.308264	Italian Restaurant
...	...	...	...	...	...	...	...
M8Z	Etobicoke	43.628841	-79.520999	Rocco's Plum Tomato	43.634898	-79.519951	Italian Restaurant
M8Z	Etobicoke	43.628841	-79.520999	Pazzia	43.624854	-79.509473	Italian Restaurant
M9L	North York	43.756303	-79.565963	Albertos Trattoria	43.748744	-79.560714	Italian Restaurant
M9M	North York	43.724766	-79.532242	Jolly II Italian Restaurant	43.711946	-79.531510	Italian Restaurant
M9N	York	43.706876	-79.518188	Jolly II Italian Restaurant	43.711946	-79.531510	Italian Restaurant

Venue Categories

The FourSquare database was used to obtain information with respect to the venue's location in order to locate its position spatially along with relative position to determine which postcode to associate with. In addition, further modelling will be done using the venue categories, this is important as it will provide the right baseline for grouping venues of similar characteristics.

WIKIPEDIA

Postcode, Borough and Neighborhood information will be scraped from the Wikipedia website. This will form the basis for dissecting the neighbourhood into parts for analysis.

	Postal Code	Borough	Neighborhood	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
...	...	...	...	...	...
98	M9N	York	Weston	43.706876	-79.518188
99	M9P	Etobicoke	Westmount	43.696319	-79.532242
100	M9R	Etobicoke	Kingsview Village, St. Phillips, Martin Grove ...	43.688905	-79.554724
101	M9V	Etobicoke	South Steeles, Silverstone, Humbergate, Jamest...	43.739416	-79.588437
102	M9W	Etobicoke	Northwest	43.706748	-79.594054

103 rows × 5 columns

Postal Code Dataframe

The first step in the process was to arrange and populate the information from Wikipedia into a Dataframe. The data contains 103 rows which represent the number of postcodes in Toronto and 5 columns, Postal Code, Borough, Neighborhood and location information which will be used to call the FourSquare API.

Methodology

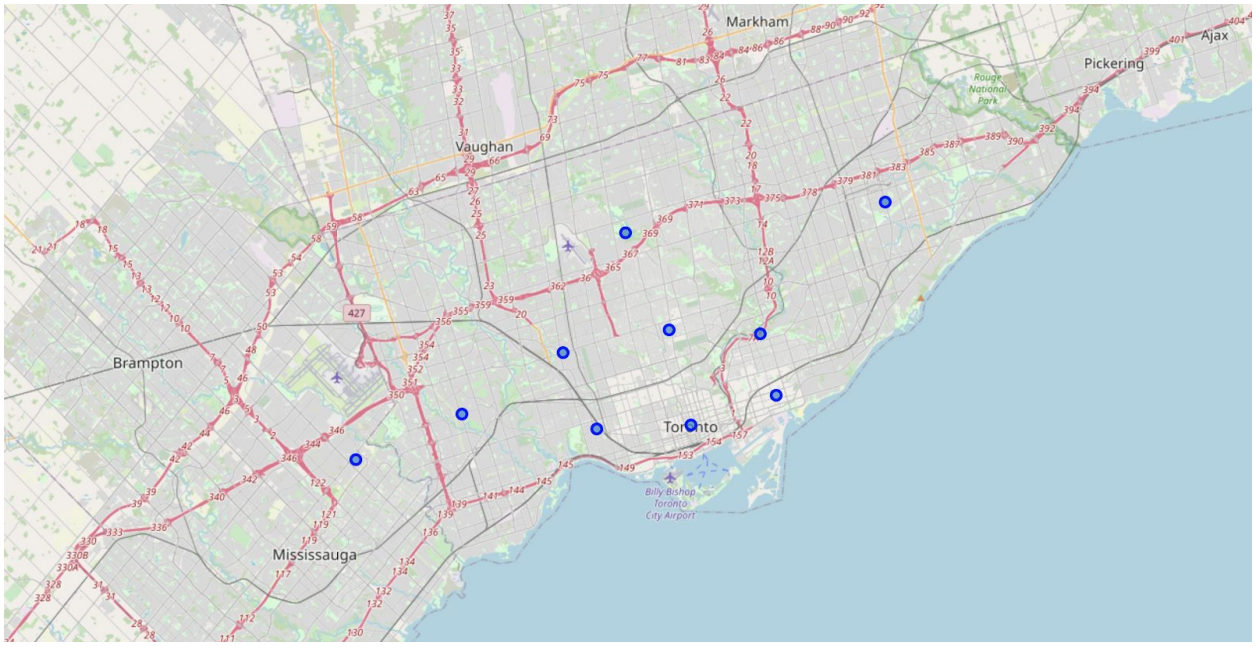
The first step in the analysis is to establish a base dataframe that determines the centres from which the FourSquare request can return venues. In this analysis there were two methods that were tested, calculate borough locations from the mean of postal code locations that fall within the borough or calculation of location based on postal code alone.

Grouping of the Data

An initial model was performed by clustering Boroughs. The location information of all postal codes within a Borough are grouped and a mean of these locations generated to determine centroids for each Borough. Using the Foursquare API we requested all venues within a 1500m radius from the centroids of the Borough's this returned 600+ venues.

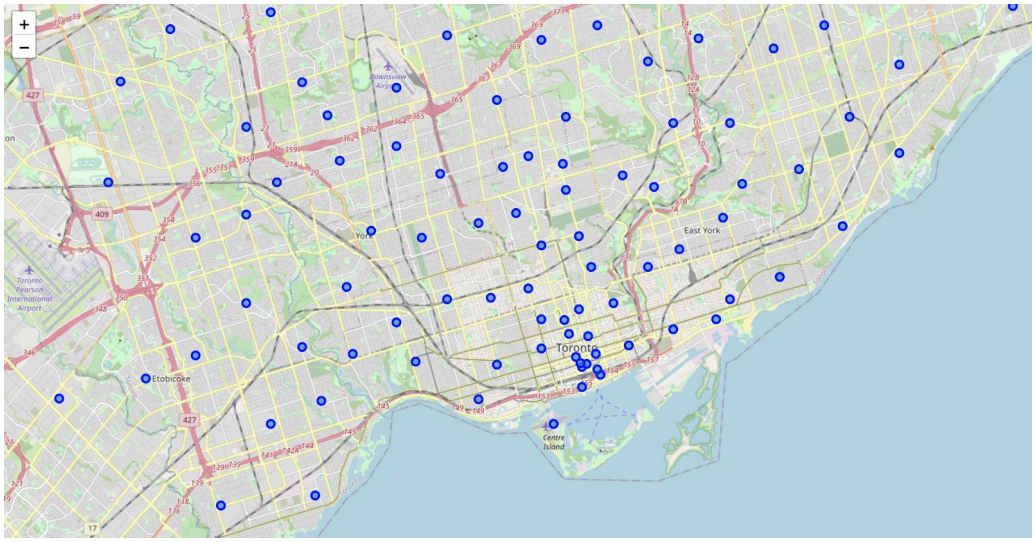
The image below shows the location distribution of Boroughs. As you can see using the mean locations is not an effective way to analyse the data for the following reasons;

- The model assumes that each postal code is uniformly distributed and of the same size which may not necessarily be the case
- The borough centroids are too sparsely distributed therefore venues that may be within the borough but outside of the 1500m radius net are not captured for analysis.



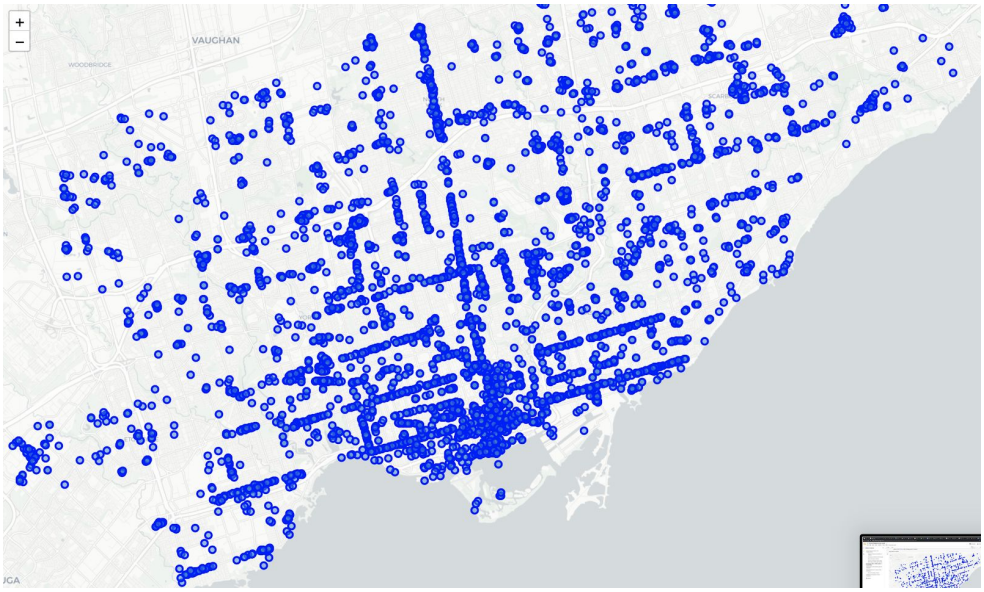
Mean Borough centroids

Using a different approach that uses the distinctively use the postal code would return a better result as is takes into account the uneven spatial distribution of postcode centres. This process requires a greater number of API calls from FourSquare but will return significantly more venues that are accurately associated with the postcodes.



Postal Code Locations

Using the postal code locations alone, the FourSquare database returned significantly more venues to the order of **6922** venues within the postal code of which **346** unique venue categories. Unique categories provide insight into how diverse the dataset is the smaller the value the more each venue is alike the larger the value the more each venue is distinct. This a more superior model for the analysis in comparison to groups by boroughs.



6922 Venue Locations.

Determining Likeness

Likeness would be determined by the characteristics of each neighbourhood. We approach this by determining what establishments existed within a specific distance of each neighbourhood. To do this we sorted the 6922 venues by Postal Code and performed a count of how many venues of a specific type appeared. Once the frequency had been established a mean was taken and each venue category was ranked in order from most to least frequently occurring in each postal code.

	Postal Code	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports
0	M1B	0.0	0.0	0.028571	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.028571	0.000000	0.000000
1	M1C	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
2	M1E	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.031250	0.000000
3	M1G	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
4	M1H	0.0	0.0	0.000000	0.0	0.015873	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.015873
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	M9N	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.019231	0.000000
99	M9P	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
100	M9R	0.0	0.0	0.000000	0.0	0.047619	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
101	M9V	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
102	M9W	0.0	0.0	0.000000	0.0	0.050000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000

1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Zoo Exhibit	Restaurant	Fast Food Restaurant	Pizza Place	Women's Store
Park	Gym / Fitness Center	Italian Restaurant	Burger Joint	Neighborhood
Pizza Place	Breakfast Spot	Bank	Coffee Shop	Restaurant
Coffee Shop	Pizza Place	Pharmacy	Fast Food Restaurant	Indian Restaurant
Coffee Shop	Restaurant	Sandwich Place	Clothing Store	Gas Station

Sample of Frequency Dataframe

The Dataframe was cleaned up and the top 5 venue categories were selected and ranked from most to least frequent. The reasoning behind using the top 5 as opposed to any other number such as 10, 15 or even 20 is that fact that we want to provide a higher weighting towards the top 5 venues as we feel this would provide a more accurate result of grouping using a machine learning segmentation technique called **k-means.**

K-means Clustering Technique

K-means was the best choice for this analysis because we working with an unlabeled dataset. K-means is a machine learning technique that works with an unlabeled dataset to segment data into groups based on certain characteristics that make groups a-like. The algorithm functions by calculating distances between centroids and venue categories within a cluster, after which a new cluster centre is calculated based on the mean of all the distances within that cluster to the centroid of that cluster. This process is repeated until a minimum distance all the venue categories within a venue cluster is achieved relative to other cluster centroids in the calculation. What's important to note here is the that is difficult to determine an appropriate starting k-value for the number of clusters required to properly label the dataset. For this analysis, the approach that was taken was to run the model several times at intervals of 5, what we discovered an optimum k-value of 8 as no matter what value was used beyond this the algorithm would always return 8 clusters. We also want to get as close to the global as possible in order to achieve this the algorithm was set to repeat the k-means algorithm and re-calculation of new centres 30 times, values in excess of 30 had little to no impact on the results, therefore, it was deduced that 30 would be an optimum value.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353	7
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	0
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711	5
3	M1G	Scarborough	Woburn	43.770992	-79.216917	5
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476	1

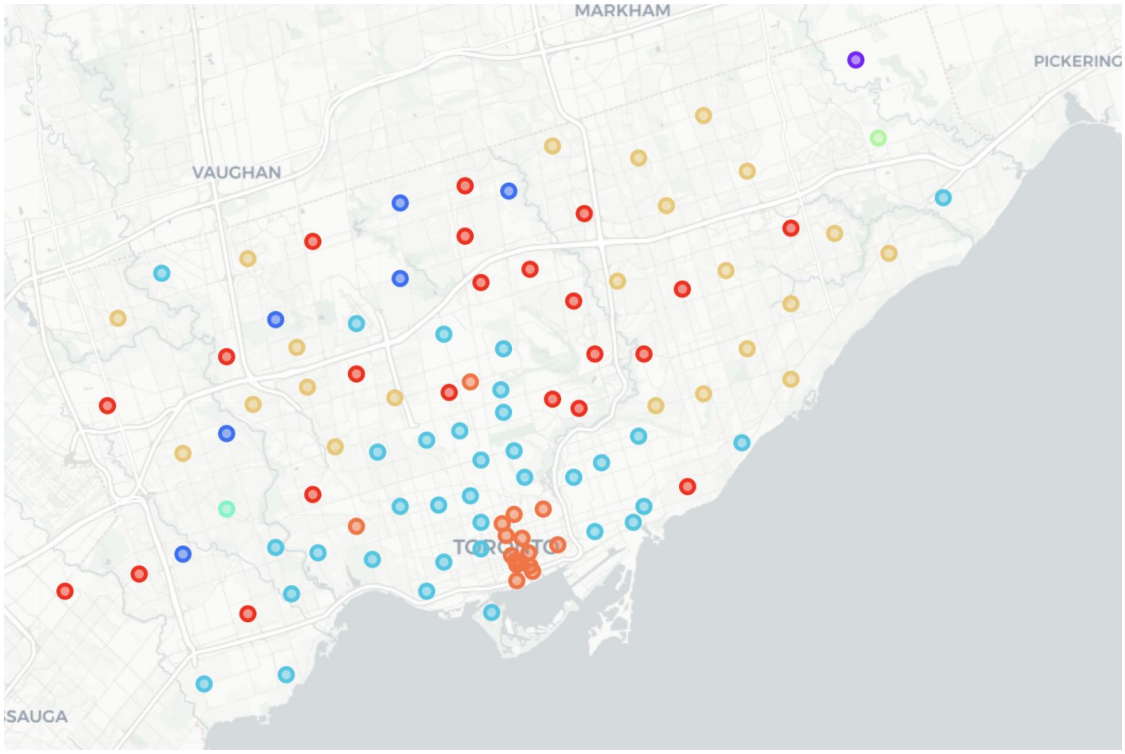
Cluster labels for each Post Code

Venue Analysis

Lastly, to determine whether or not an Italian restaurant should be opened and if so in which postal code, we looked at the venue category frequency once again. We needed to understand first how many Italian restaurants we're already in Toronto, where were they all located and are they dependant on other venue categories. In order to do this, a count was performed across the venues categories to understand the frequency and the results of the k-means segmentation was used to highlight postal codes with high occurrences of Italian Restaurants. The next step was to analyse of the postal codes that ranked Italian Restaurant as most frequently occurring which how many had similar venue categories that ranked 2nd or 3rd

Results

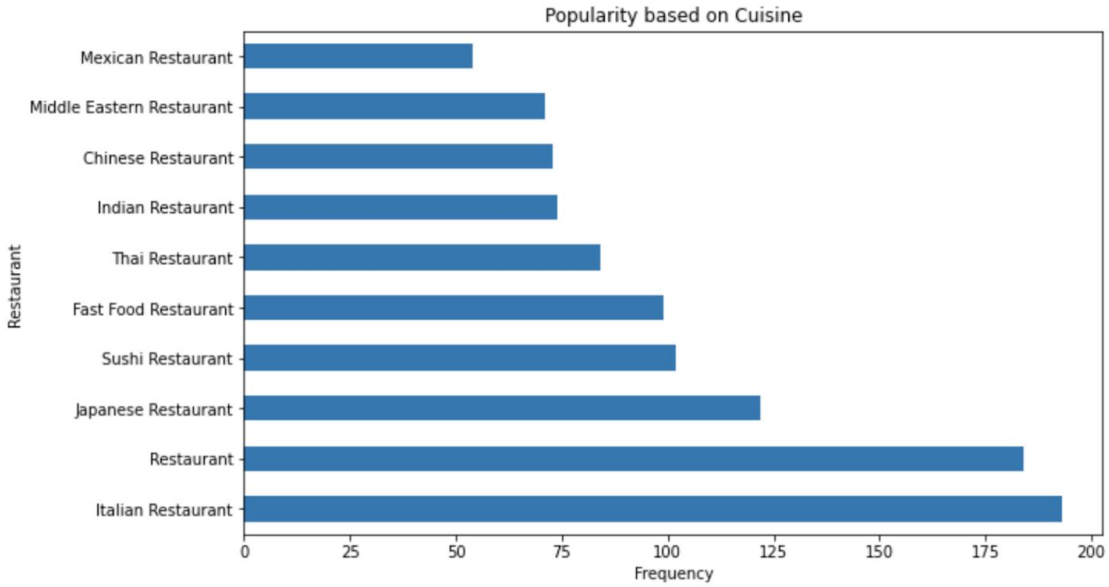
Applying the k-means segmentation algorithm using a k-value of 8 resulted in the segmentation of venue categories of Toronto as shown in the image below.



Segmentation results of Toronto

Cluster	Cluster Type	Legend
1	A predominantly high concentration of Coffee Shops	
2	Farms, National Parks	
3	Banks	
4	Coffee, Cafes and Italian Cuisine	
5	Pharmacy and Shopping Mall	
6	Zoo Exhibits	
7	Asian Cuisine and Pharmacies	
8	Coffee, Cafe, Park and Japanese Restaurants	

The area coloured circles represent the location of the postal code and the colours represent which category it belongs in.

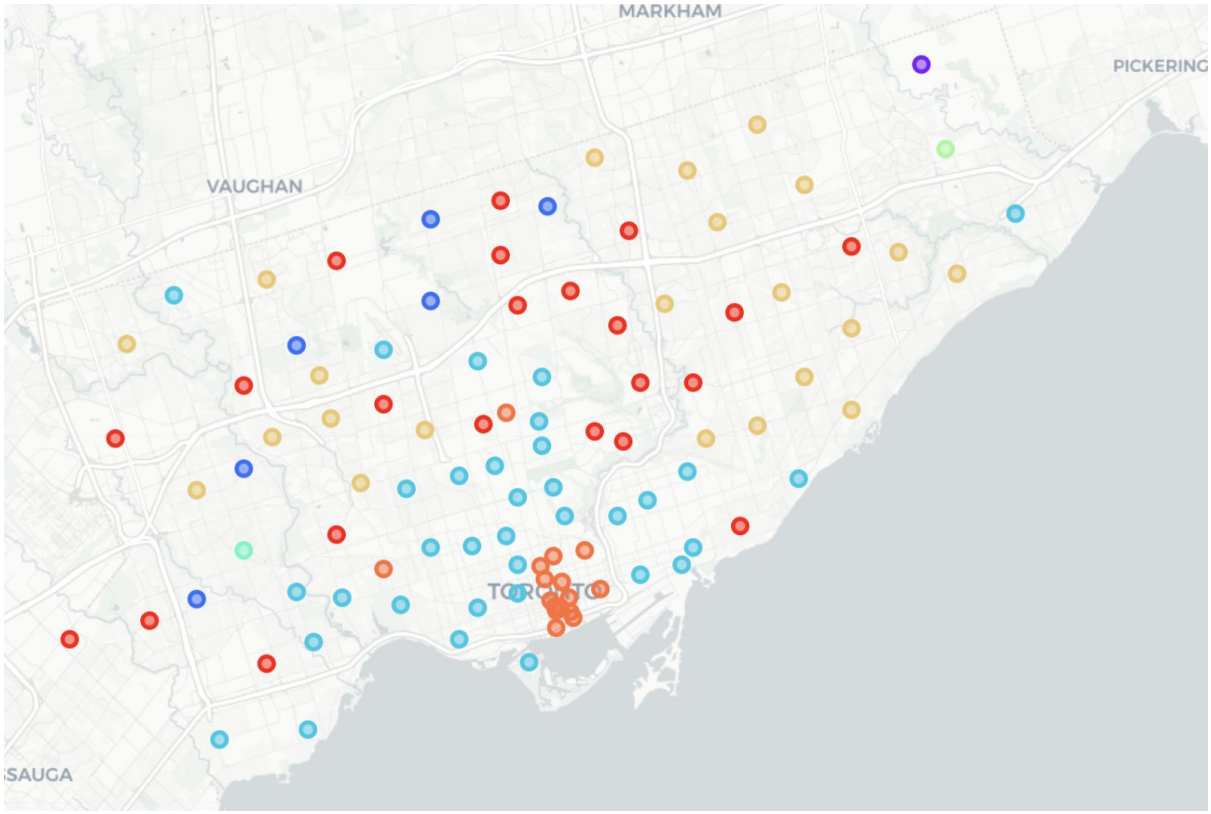


There are 193 Italian restaurants in Toronto, which makes it the most popular restaurant in Toronto, we also analysed to understand which postal codes had the highest concentration of Italian restaurants. What we discovered is shown in the image below.

	Postal Code	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
45	M4P	Central Toronto	Davisville North	43.712751	-79.390197	3	Italian Restaurant	Coffee Shop	Park	Restaurant	Indian Restaurant
47	M4S	Central Toronto	Davisville	43.704324	-79.388790	3	Italian Restaurant	Pizza Place	Coffee Shop	Bakery	Park
48	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	3	Italian Restaurant	Park	Sushi Restaurant	Coffee Shop	Grocery Store
49	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	3	Italian Restaurant	Café	Park	Sushi Restaurant	Coffee Shop
62	M5M	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750	3	Italian Restaurant	Coffee Shop	Bakery	Sushi Restaurant	Bagel Shop
65	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	3	Italian Restaurant	Coffee Shop	Park	Grocery Store	Café
74	M6E	York	Caledonia-Fairbanks	43.689026	-79.453512	3	Italian Restaurant	Coffee Shop	Furniture / Home Store	Pizza Place	Bakery
91	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.636258	-79.498509	3	Italian Restaurant	Coffee Shop	Restaurant	Pizza Place	Park

The results show that Central Toronto has the highest concentration of Italian Restaurants in Toronto. The results also show that North York, York, Etobicoke are similar in characteristics.

Discussion



The results of the K-means segmentation shows that Down Town Toronto is unique. There are no other postal codes areas aside from M6P in West Toronto that share the same characteristics of an alfresco cafe/coffee culture with a dash of sushi. The high concentration of coffee/cafes and Japanese cuisine also only spans across a very small geographic footprint. Given the similarities, M6P lacks a Japanese Restaurant in its top 5 list of frequently occurring venues by category. Looking at the similarities, a Japan restaurant would be the best option for this postal area.

Italian Restaurants are also the most popular cuisine in by frequency followed by Japanese and then fast food. Coffee/cafes and Italian cuisine form a belt around Down Town Toronto, however, more spread out geographically. The Italian restaurants are correlated to coffee shops and cafes. Therefore East York or West Toronto would be ideal choices for an Italian Restaurant because coffee shops and cafes are within the top 5 list of frequently occurring venues by categories and lacks an Italian Restaurant.

Conclusion

In Conclusion, the objective of this report was to apply data science techniques to gather insights into how similar postal codes were from one another. The importance of this was to better understand whether the frequency of certain venue types was correlated or dependant on other venues.

This would form the determining factor in deciding what cuisine the restaurant would cater to and the ideal location of the restaurant. Leveraging the FourSquare, we requested all the venues within a 500m radius of the postal code location and applied a machine learning segmentation technique called k-means to categories each neighbourhood into clusters of likeness based on frequently occurring venues by category.

The clusters highlighted that Japanese Cuisine was correlated with Cafes and Coffee shops within the Down Town Toronto area and Italian Cuisine was also correlated with Cafes and Coffee shops outside of Down Town Toronto.

Therefore, it would be safe to presume that Restaurants catering to Japanese cuisine should be located in Down Town Toronto and restaurants catering to Italian Cuisine should be located outside of Down Town Toronto and it would be prudent to find postal code locations with have both cafes and coffee shops appearing in high frequency.

We did note that there is also a frequency of un-categorised restaurants, without further analysis, there is no means of utilising this information. Therefore, un-categorised restaurant information was omitted for the purposes of this report.