Capstone project report

Problem :

As a long time toronto resident , one could be wondering , is that it with toronto ? Could there possibly be a part of Toronto in which I have never seen before ?
The idea that I am presenting right now originated from a match making app where one could be wondering is anyone out there that could actually gain your attention in life no matter what the differences or similarities are .

Aim :

Therefore the question that I am trying to tackle now is to provide a local solution to the residents in toronto to identify similar or different suburbs surrounding them .
With this solution , the residents or a new visitor in toronto could easily locate the difference around every corner of Toronto and could possibly promote every suburb out there.

Target application / audience :

With the ability to cluster the neighborhood in Toronto ,the result generated would be useful to the local council or others (eg : tourist or travel author) who would like to promote downtown Toronto to the world) . This would enable them to easily set up useful selections and recommendations in their publications , flyer or event a local guide app , therefore highlighting the power of clustering different suburbs .

Method :

This could be solved with the foursquare API , bundled with K-means clustering and other python tools as part of the process to effectively cluster the neighborhood and then effectively provide recommendations to the user that enquire it .

Location Data

Location in examination

Toronto . The neighborhood data of Toronto will be obtained via the scraping method (with the beautifulsoup library) that was introduced in week 3's assignment . For my case , I will be narrowing down my data only to suburbs that has the word 'Toronto' in it's borough. To obtain data on Toronto's neighborhood , I will be obtaining data from Toronto's neighborhood **.** With beautifulsoup , the data extracted from wikipedia contains the desired content and in the right format, therefore no further data cleaning and arrangement is required with the exception of converting the array of results over to panda's dataframe format .

Methodology

Explanation & discussion

In order to effectively mark the location of all neighborhood , I will also be appending the scrapped dataset with their respective latitude and longitude , example as below :

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

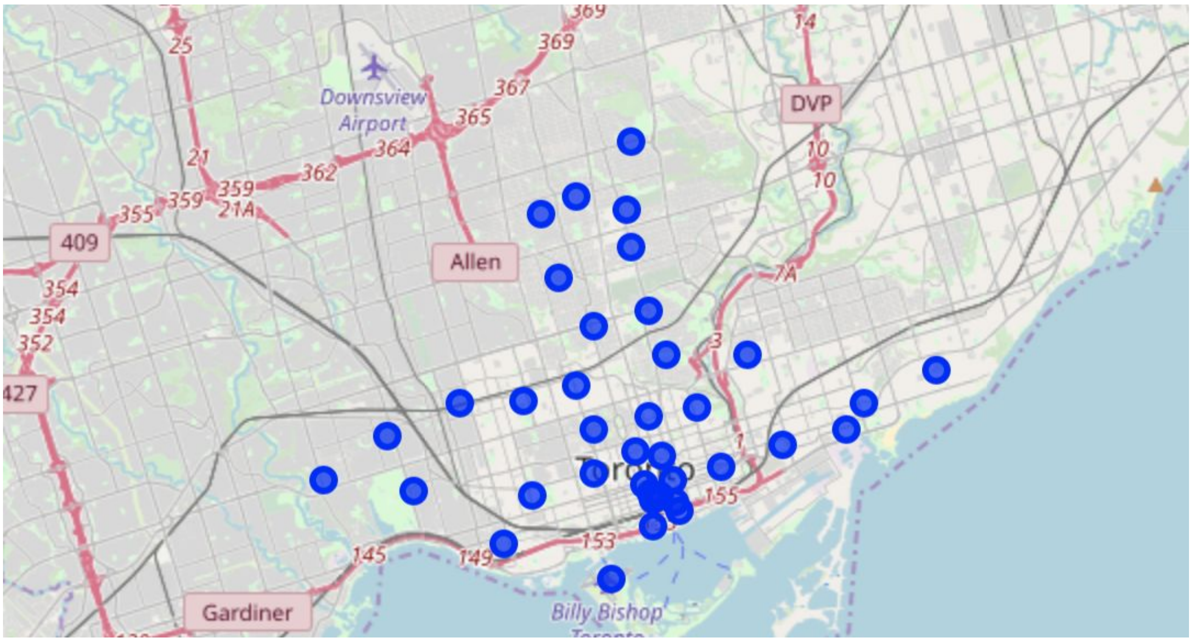
K-means clustering will then be applied onto the data sets where each suburbs will be marked with specific clusters on the map of Toronto.

The neighborhood will be clustered according to the venues that are in them. To obtain extra details regarding the venues , I will be using the EXPLORE endpoint as part of foursquare location API , where the result would look like this :

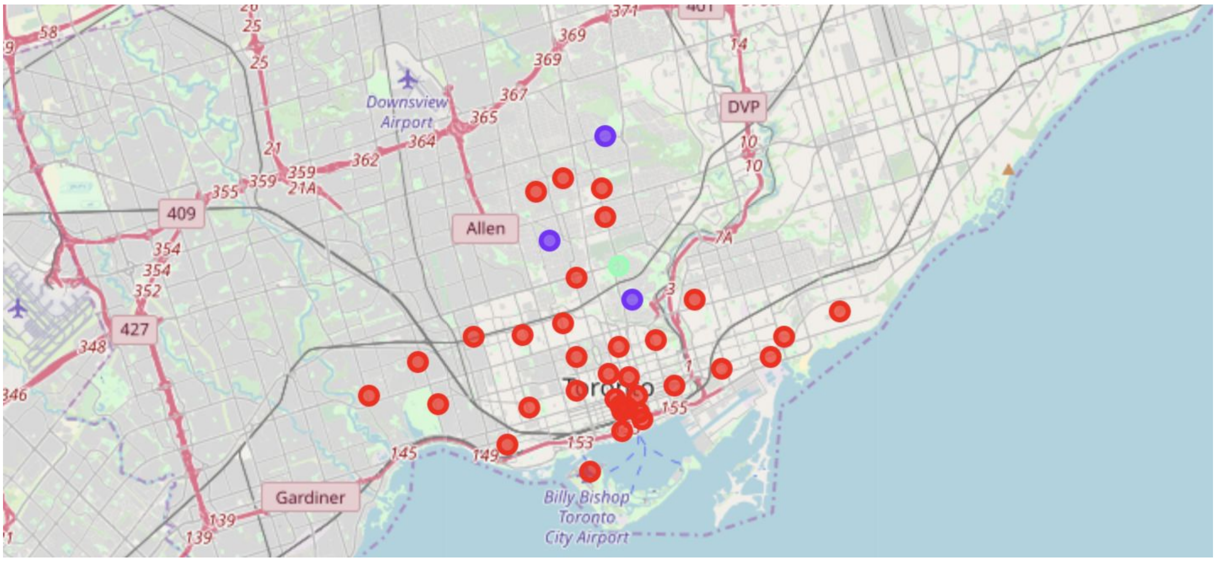
	Neighborhod	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aq
0	The Beaches	0	0	0	0	0	0	0	0	0	0	0
1	The Beaches	0	0	0	0	0	0	0	0	0	0	0
2	The Beaches	0	0	0	0	0	0	0	0	0	0	0
3	The Beaches	0	0	0	0	0	0	0	0	0	0	0
4	The Danforth West, Riverdale	0	0	0	0	0	0	0	0	0	0	0

Results

Before k-means clustering was applied onto the data set , each neighborhood was marked onto the Toronto map , having applied the same colour blue .



After k-means clustering was applied and Machine Learning had differentiate the neighborhoods into 3 different clusters , this map was produced with different colours applied onto each neighborhood .



The clusters labelled red , green and purple are detailed below :

cluster 1 (labelled RED)

```
In [37]: tor_merged.loc[tor_merged['Cluster Labels'] == 0, tor_merged.columns[[1] + list(range(5, tor_merged.shape[1]))]]
```

Out[37]:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	East Toronto	0	Health Food Store	Pub	Trail	Neighborhood	Dog Run
1	East Toronto	0	Greek Restaurant	Coffee Shop	Italian Restaurant	Ice Cream Shop	Furniture / Home Store
2	East Toronto	0	Park	Gym	Italian Restaurant	Pizza Place	Pub
3	East Toronto	0	Café	Coffee Shop	Gastropub	Italian Restaurant	Bakery
5	Central Toronto	0	Park	Clothing Store	Breakfast Spot	Gym	Hotel
6	Central Toronto	0	Coffee Shop	Yoga Studio	Bagel Shop	Park	Clothing Store
7	Central Toronto	0	Pizza Place	Dessert Shop	Sandwich Place	Italian Restaurant	Café
9	Central Toronto	0	Coffee Shop	Pub	Liquor Store	Sushi Restaurant	Supermarket
11	Downtown Toronto	0	Coffee Shop	Restaurant	Pizza Place	Café	Bakery
12	Downtown Toronto	0	Japanese Restaurant	Coffee Shop	Sushi Restaurant	Restaurant	Gay Bar
13	Downtown Toronto	0	Coffee Shop	Park	Bakery	Pub	Theater

cluster 2 (labelled PURPLE)

```
In [38]: tor_merged.loc[tor_merged['Cluster Labels'] == 1, tor_merged.columns[[1] + list(range(5, tor_merged.shape[1]))]]
```

Out[38]:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Central Toronto	1	Park	Swim School	Bus Line	Yoga Studio	Doner Restaurant
10	Downtown Toronto	1	Park	Playground	Trail	Dog Run	Fish & Chips Shop
23	Central Toronto	1	Sushi Restaurant	Park	Jewelry Store	Trail	Yoga Studio

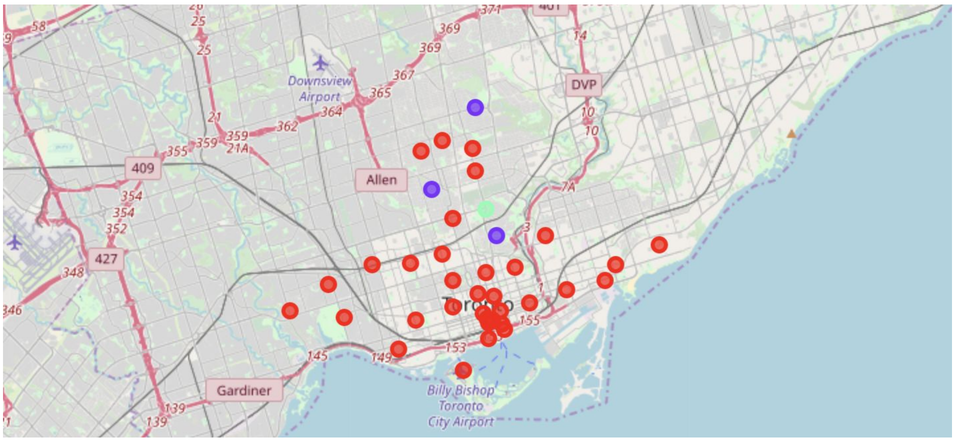
cluster 3 (labelled GREEN)

```
In [39]: tor_merged.loc[tor_merged['Cluster Labels'] == 2, tor_merged.columns[[1] + list(range(5, tor_merged.shape[1]))]]
```

Out[39]:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
8	Central Toronto	2	Playground	Tennis Court	Concert Hall	Convenience Store	Filipino Restaurant

Discussion



First , lets look at the result map .
We can see that most neighborhoods has been labelled with cluster 1 . Where the hotspot for neighborhood of cluster 1 are at the center of Toronto .
Getting further away ,you can see members of cluster 2 and 3 appearing , heading up north of the center business district .

Looking at the member breakdown of each clusters in the result section , the number 1 most common venue for cluster 1 is eatery such as restaurants and coffee shop .
For cluster 2 ,the most common venue is the park and for cluster 3 , the most common spot is the recreational spot such as playground .
Cluster 1 contains the most number of members followed by cluster 2 and 3 . This could be seen on the result map above where most of Toronto have been marked red due to the presence of cluster 1 .

Conclusion

With most of Toronto being marked red as the result of k-means clustering , it tells us that Toronto is a city that would be a joy to visit for food lovers . Therefore the local council should promote the hidden gems (eateries) in Toronto as the main highlight of the city .

As for a visitor , if you would like to explore the different looks of Toronto , head up north , where you will be greeted with a different demographic .