Data Science Capstone Project

Introduction to the Business Problem:

We are going to use k-means algorithms to cluster Downtown Toronto , Toronto. Downtown Toronto is pretty expensive place which has great amenities such as restaurants,bars,cafes,gyms,playgrounds,schools,saloons all of which are some of the most quite sophisticated in the world.The people moving to this area spend a lot of money to move into such an expensive neighborhood because of their world class architecture/amenities and breathtaking landscapes.They want to leave no stones unturned before moving into this neighborhood so that they can have an amazing life which they won't regret.Since Downtown Toronto is a large area it can have its differences which caters to different individuals.As a data scientist i have to provide the best bang for my client's buck using various machine algorithms (i.e k-means in this case) and most importantly to leverage the power of data to make better choices. This problem focuses to subdivide the Downtown Toronto area into clusters so that clients can make better decisions based on their preferences.

Source:Unsplash ,courtesy of Krishna C Koganti

Data

As a data scientist the most important and time consuming part of this process is to obtain,clean and analyse the data.For this project i have obtained from 2 sources mainly :

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

2.Foursquare api

Wikipedia was used to obtain details of various Boroughs (since Downtown Toronto happens to be one such Borough) along with their respective postal code and neighborhood.

	Postalcode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

We then add the longitude and latitude corresponding to the postal code to the above table with the help of python libraries i.e pandas,geocoder

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.81139	-79.19662
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.78574	-79.15875
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.76575	-79.17470
3	M1G	Scarborough	Woburn	43.76812	-79.21761
4	M1H	Scarborough	Cedarbrae	43.76944	-79.23892

Now the above table only contains a small section,in reality the table is large consisting of data of all the neighborhoods (i.e their postal code,borough,neighborhood,lat,long).

However we are only concerned with the data relating to Downtown Toronto,so we create a new table which is a subset of the above table that contains the data only about Downtown Toronto. We use the pandas library of python to select data related to Downton Toronto.

	Postalcode	Borough	Neighborhood	Latitude	Longitude
0	M4W	Downtown Toronto	Rosedale	43.68190	-79.37829
1	M4X	Downtown Toronto	St. James Town, Cabbagetown	43.66788	-79.36649
2	M4Y	Downtown Toronto	Church and Wellesley	43.66659	-79.38133
3	M5A	Downtown Toronto	Regent Park, Harbourfront	43.65512	-79.36264
4	M5B	Downtown Toronto	Garden District, Ryerson	43.65739	-79.37804

Next to visualize this data in a map we use the folium library

Methodology

We need to get more data about various venues and their respective categories.We use the foursquare library for this.We communicate with the api and obtain the data.The data obtained which is in json format is then converted to pandas dataframe which can be then used for analysing and processing it.

	name	categories	lat	lng
0	Rosedale Tennis Club	Tennis Court	43.683226	-79.378984
1	Rosedale Park	Playground	43.682328	-79.378934
2	Betline Trail at Roxborough dr.	Bike Trail	43.680530	-79.381490
3	Whitney Park	Park	43.682036	-79.373788
4	Petite & Sweet	Shop & Service	43.685982	-79.376072

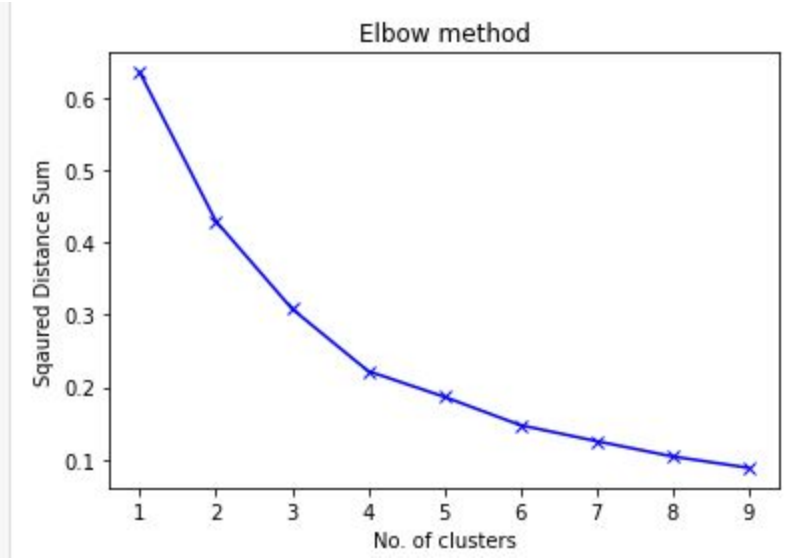
We then explore various neighborhoods and see how many venues they have,this is done using the count function in pandas, which splits each neighborhood and then counts the venues across them,the result obtained is then merged.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Berczy Park	60	60	60	60	60	60
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	76	76	76	76	76	76
Central Bay Street	76	76	76	76	76	76
Christie	11	11	11	11	11	11
Church and Wellesley	79	79	79	79	79	79
Commerce Court, Victoria Hotel	100	100	100	100	100	100
First Canadian Place, Underground city	100	100	100	100	100	100
Garden District, Ryerson	100	100	100	100	100	100
Harbourfront East, Union Station, Toronto Islands	58	58	58	58	58	58

Now we wish to see the top 10 venues,for that we have to first convert our venue categories which is in string format(eg: coffee shops,gyms,studios) to dummy variables.We then sort these values and put them into a table.The final result contains top 10 venues for each of its corresponding neighborhoods.

	Neighborhood	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	Berczy Park	Coffee Shop	Bakery	Breakfast Spot	Seafood Restaurant	Cheese Shop	Farmers Market	Beer Bar	Cocktail Bar	Restaurant	Pub
1	CN Tower, King and Spadina, Railway Lands, Har...	Italian Restaurant	Coffee Shop	Café	French Restaurant	Park	Bar	Sandwich Place	Speakeasy	Lounge	Bakery
2	Central Bay Street	Coffee Shop	Clothing Store	Bookstore	Hotel	Cosmetics Shop	Restaurant	Bubble Tea Shop	Plaza	Sushi Restaurant	Café
3	Christie	Café	Grocery Store	Candy Store	Coffee Shop	Playground	Baby Store	Italian Restaurant	Athletics & Sports	Women's Store	Electronics Store
4	Church and Wellesley	Coffee Shop	Japanese Restaurant	Restaurant	Gay Bar	Sushi Restaurant	Café	Pub	Bubble Tea Shop	Dance Studio	Hotel

Now we need to cluster the above data into various categories,to obtain how many categories i.e clusters we need we plot the clusters into a graph and find the elbow point(the point where the graph becomes less steep.)

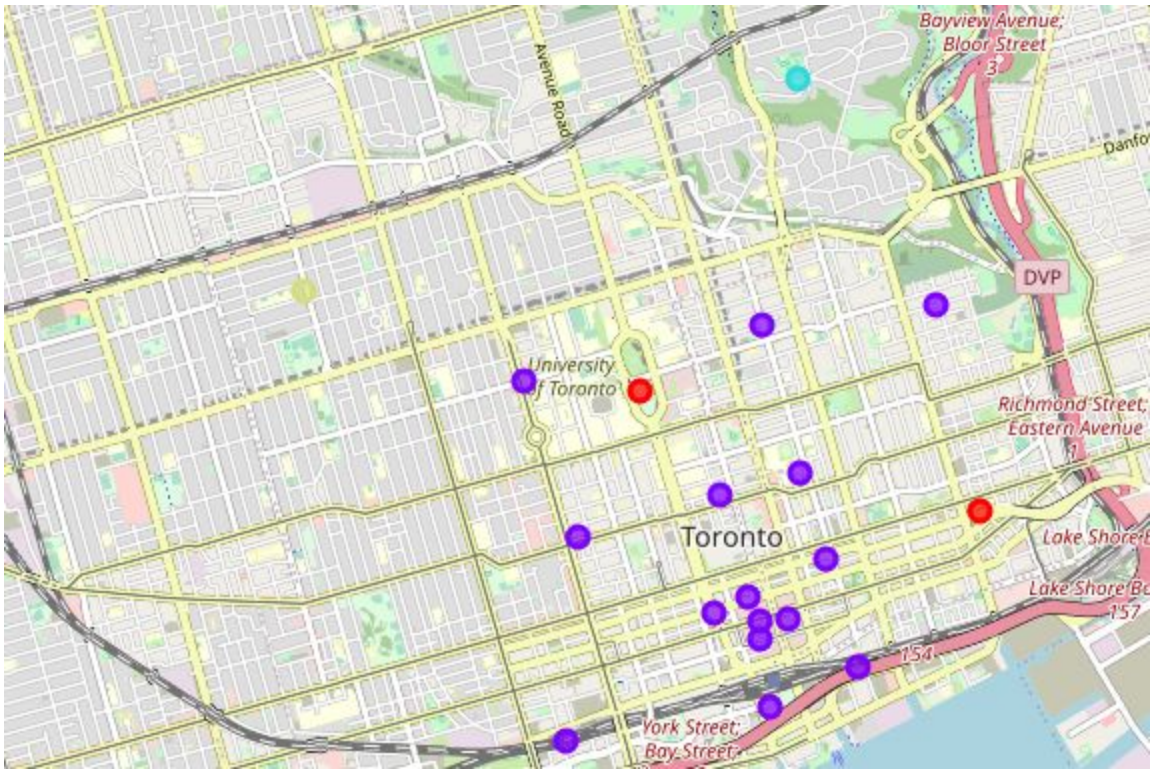


In the above graph ,if we notice the x-axis after 4(No. Of clusters) the graph becomes less steep and resembles a linearly decreasing graph.So we choose 4 as our number of clusters.After applying k-means we get 4 clusters labeled from 0 to 3

```
Out[30]: array([1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 0, 0, 1, 2, 1, 1, 1, 1, 1],
                dtype=int32)
```

Results

After mapping the labels using folium library ,our clustered map looks as follows



Here the 4 clusters are represented using the colors purple,red,yellow(upper-middle left) and light blue.

Cluster 1:

	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
3	Downtown Toronto	0	Coffee Shop	Breakfast Spot	Thai Restaurant	Event Space	Spa	Electronics Store	Food Truck	Restaurant	Bakery	Pub
18	Downtown Toronto	0	Coffee Shop	Sandwich Place	Park	Theater	Bank	Falafel Restaurant	Gastropub	Italian Restaurant	Portuguese Restaurant	Café

Cluster 2:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	
1	Downtown Toronto	1	Coffee Shop	Café	Restaurant	Pizza Place	Park	
2	Downtown Toronto	1	Coffee Shop	Japanese Restaurant	Restaurant	Gay Bar	Sushi Restaurant	
4	Downtown Toronto	1	Coffee Shop	Clothing Store	Café	Cosmetics Shop	Japanese Restaurant	
5	Downtown Toronto	1	Coffee Shop	Cocktail Bar	Hotel	Restaurant	Gastropub	
6	Downtown Toronto	1	Coffee Shop	Bakery	Breakfast Spot	Seafood Restaurant	Cheese Shop	Fa
7	Downtown Toronto	1	Coffee Shop	Clothing Store	Bookstore	Hotel	Cosmetics Shop	
8	Downtown Toronto	1	Coffee Shop	Hotel	Café	Japanese Restaurant	Restaurant	
9	Downtown Toronto	1	Coffee Shop	Hotel	Japanese Restaurant	Restaurant	Plaza	
10	Downtown Toronto	1	Coffee Shop	Hotel	Restaurant	Café	Seafood Restaurant	
	Downtown						Japanese	

Cluster 3:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Cor V
0	Downtown Toronto	2	Park	Shop & Service	Bike Trail	Tennis

Cluster 4:

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Downtown Toronto	3	Café	Grocery Store	Candy Store	Coffee Shop	Playground

Discussions

Since Downtown Toronto is a small areas which has total area of 17km sq,we could not expect large clusters which are of similar shape and size.If we notice the above map the purple cluster is the largest whereas other clusters are comparatively small having less than 2 members.One point in the **red cluster** which is near to the University of Toronto,which comprises mostly of students hence their lifestyle is completely different from other groups of people like the working class and families.Most of what a student needs are provided by the university such as dorms,food,gyms,ground,labs.Whereas the **purple clusters** which consist of the most points comprise of working class,where most of the popular spots are coffee shops,restaurants, gyms and other services. **Other 2 clusters** and suitable for newly wed couples or those who are have recently got a job.

Conclusion

As a conclusion note ,I would like to recommend any future clients or stakeholders to use other sources to get more holistic opinions such as talking to people who live there to get more accurate information.This analysis should only be used as an initial guiding tool,one must not be dependent on it. This analysis was performed during the covid 19 pandemic so most of the amenities and venues are partially or under complete lockdown.So as time changes,the recommendations will also change.This is also one of the biggest challenges as data scientist must obtain new data when times change to create a more accurate model.However this model provides a general idea to start with.