Selection of a city based on Individual's interest for relocation purpose

Scope

This document is a detailed report on a business case statement. It mentions the different aspects of a business case.

Target Audience

For this problem statement, the prospective target audience are individuals who based on their preferences are relocating to a new city, a city where they are unaware of which location would be best suited according to their choices and current lifestyle. Though this case study deals with the city of Toronto, it can be used for any city based on the postcodes and foursquare data.

Problem Statement

Ravi Iyer, an Indian is on his first ever trip overseas to the city of Toronto for a year. He belongs to an orthodox family and his family had been skeptical if he will be able to keep up with his cultural and food habits in Toronto.

Ravi has been exploring different neighbourhoods of Toronto. His preferences are

- He would not like to spend too much on commuting. Hence he would prefer a location with local transport.
- He is a vegetarian and he would prefer Asian food as much as possible.
- He is an avid reader and fitness conscious and would prefer having a book store and a fitness center close by
- Keen on not spending too much either on food, he would prefer to cook on his own and would hence need a grocery store near his accommodation
- For recreation and relaxation he prefers living close to parks and movie theatres.

With such specific needs, Ravi would like to explore neighbourhoods of Toronto and the possibility of living a lifestyle he expects.

Data Section

The data used for this analysis will be from FourSquare.com and Wikipedia. For all the zip codes available on Wikipedia for Canada, since our problem statement deals with the city of Toronto, we extract data for Toronto, clean the data and analyze it.

After cleaning of data for all neighbourhoods of Toronto, use foursquare.com to fetch venue information. Based on the data extracted and analyzed come up with the most favourable neighbourhood. We also apply statistical methods to cluster the data based on neighbourhoods.

To explain the data processing in detail, we first extract all zip code information from Wikipedia from the link:

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

The city of Toronto is mainly divided into four major areas viz. East, West, Central and DownTown Toronto, we get a list of venues from foursquare.com, the venues which meet our problem statement criteria. Based on the grouped data figures and the availability of services in each area, we decide on DownTown Toronto as the most favourable place. The selection is based purely on the counts of desired venues in each area.

Once the selection of the area is done, we then go on to get more information about the venues in DownTown Toronto. The venues are categorized, and are ranked based on their popularity. Using statistical method of kmeans clustering we find out the most optimum number of clusters that can be applied and then perform k means clustering on the data based on the neighbourhoods in DownTown Toronto.

Methodology

Once the data is identified, the right methodology is used to get inferences out of it. Since our problem statement is for the city of Toronto, from all the zip codes of Canada, we filter out only the relevant information we need for the city of Toronto. We also clean up the data for it is possible for some of the zip codes, we get incomplete or missing data.

Once this is done, we are left with details of only the city of Toronto. The city of Toronto is divided into four main areas Central, West, DownTown and East Toronto.

In our problem statement, we are looking at a locality which caters to very specific type of venues. To categorize in detail, the preferred categories in our problem statement drill down to:

'Asian Restaurant', 'Vegetarian / Vegan Restaurant', 'Health Food Store', 'Yoga Studio', 'Bookstore', 'Park', 'Grocery Store', 'Movie Theater', 'Café', 'Gym / Fitness Center', 'Light Rail Station'

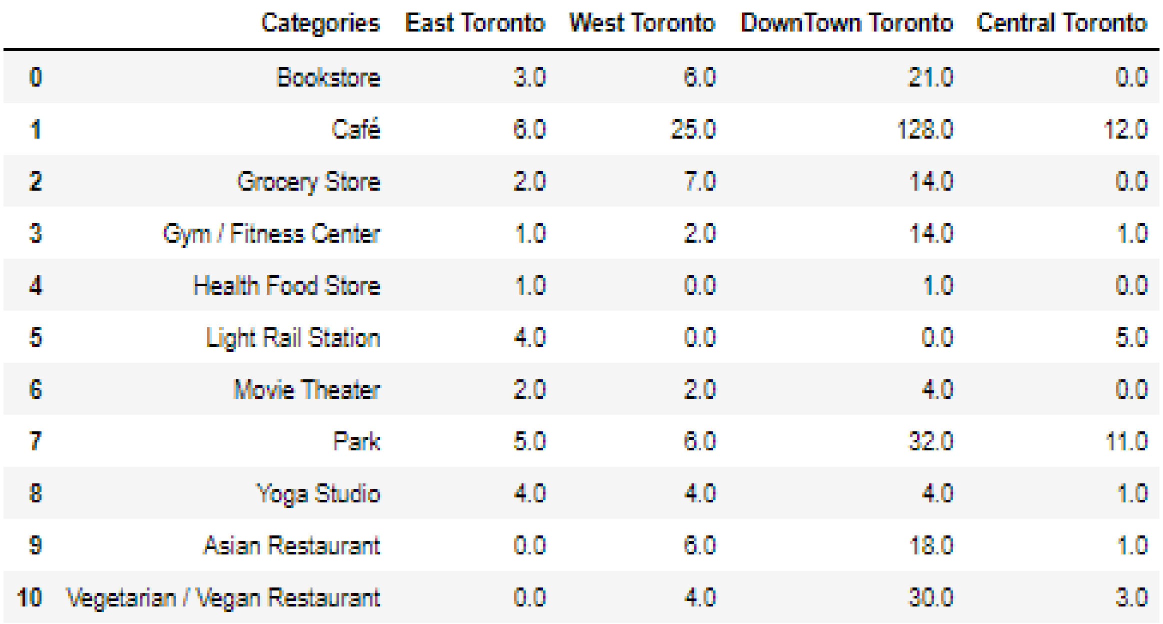
Of all four areas in Toronto, after categorizing the venues and finding out the count of venues specific to our problem statement, we conclude DownTown Toronto as the most viable venue to meet our requirements.

Once we have narrowed down on DownTown Toronto as our area of interest, we perform further analysis on the venues in DownTown Toronto for our preferred list of categories.

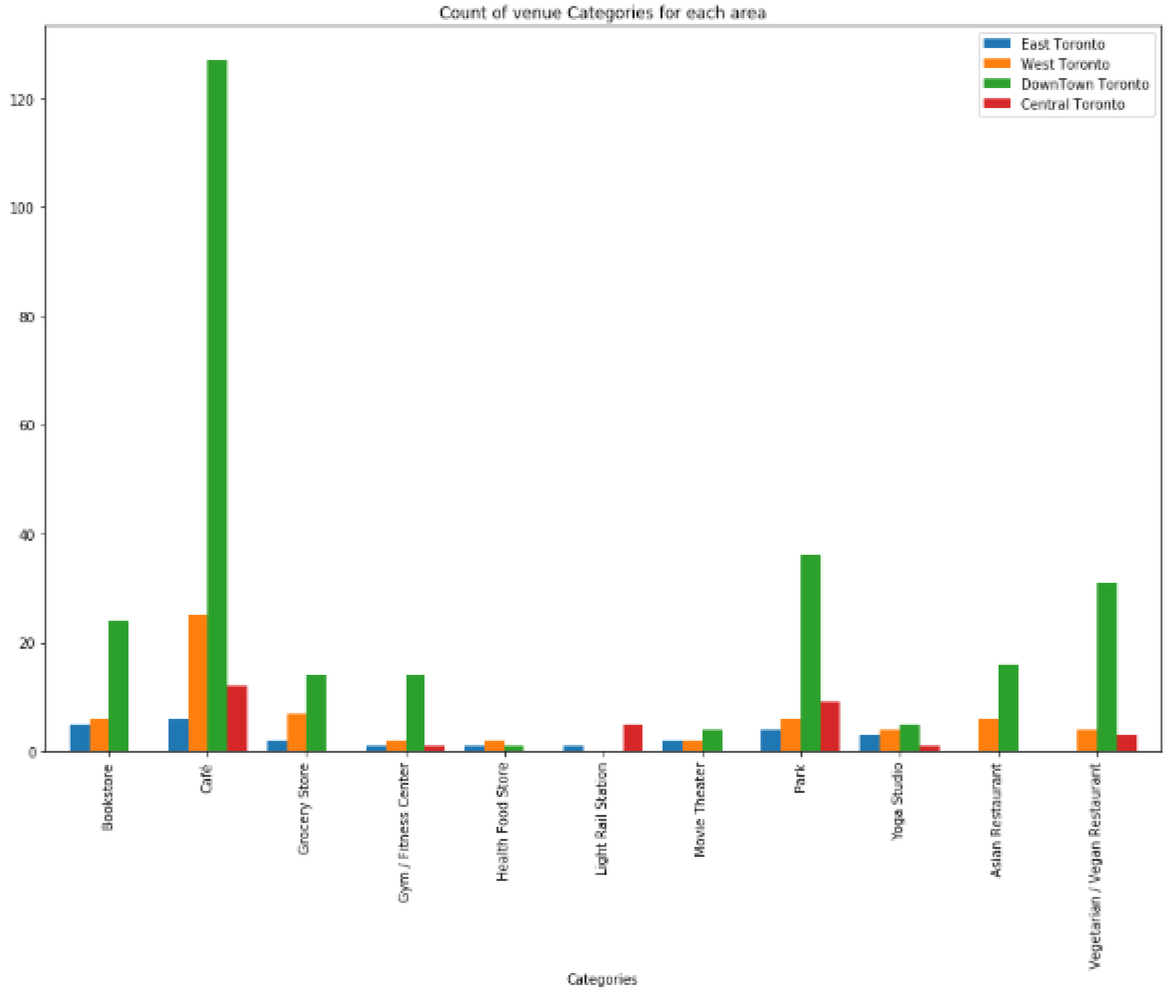
We list the most 10 popular venues for each Neighbourhood of DownTown Toronto. This is followed by using the elbow method to find the best value of k for k means clustering on Neighbourhood of DownTown Toronto.

Results

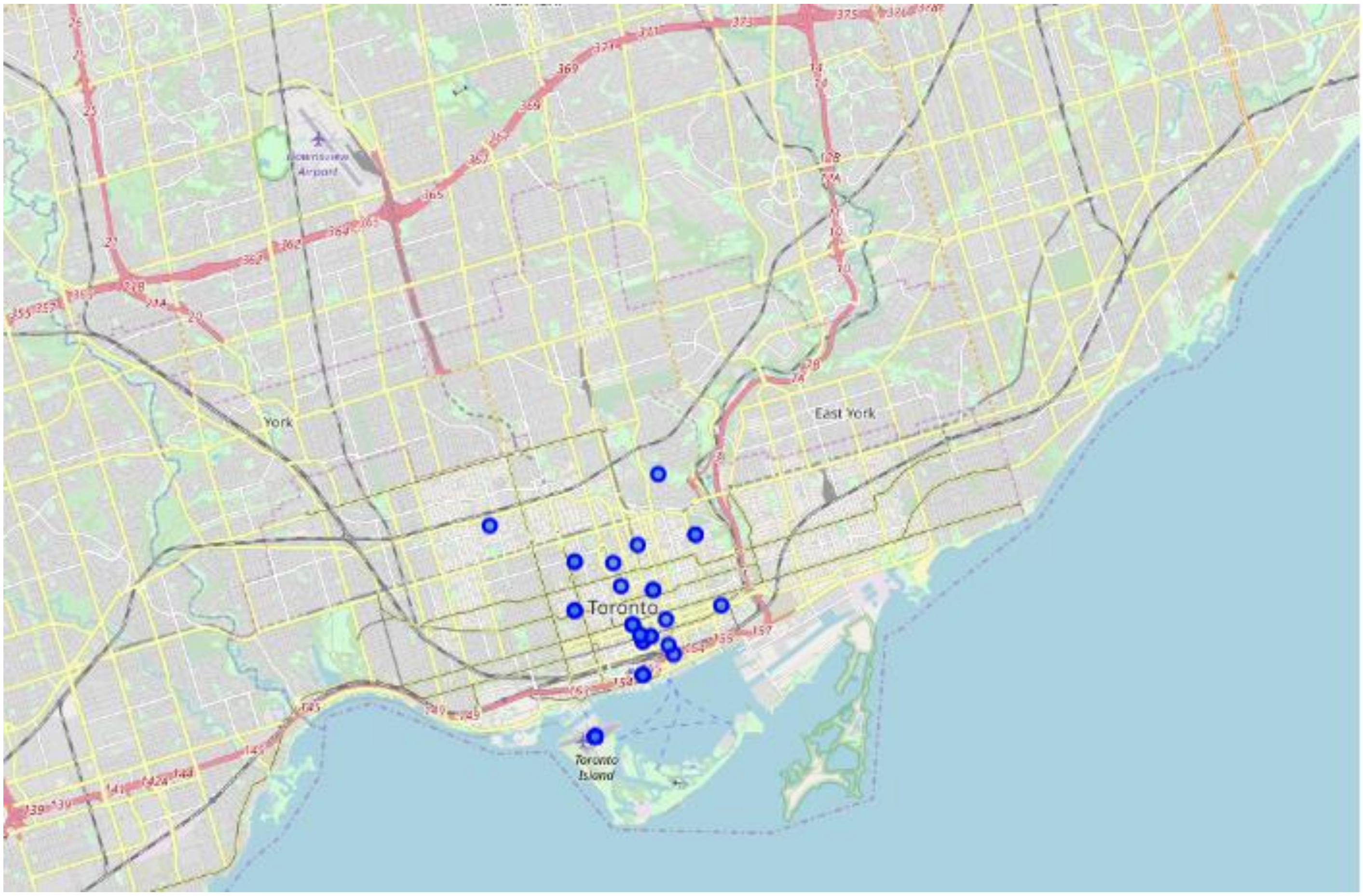
With the initial analysis of all areas of Toronto, our results are the cumulative values of preferred venue categories for each area. The following table and the subsequent graph shows us the count of venue categories of interest in each area.



	Categories	East Toronto	West Toronto	DownTown Toronto	Central Toronto
0	Bookstore	3.0	6.0	21.0	0.0
1	Café	6.0	25.0	128.0	12.0
2	Grocery Store	2.0	7.0	14.0	0.0
3	Gym / Fitness Center	1.0	2.0	14.0	1.0
4	Health Food Store	1.0	0.0	1.0	0.0
5	Light Rail Station	4.0	0.0	0.0	5.0
6	Movie Theater	2.0	2.0	4.0	0.0
7	Park	5.0	6.0	32.0	11.0
8	Yoga Studio	4.0	4.0	4.0	1.0
9	Asian Restaurant	0.0	6.0	18.0	1.0
10	Vegetarian / Vegan Restaurant	0.0	4.0	30.0	3.0

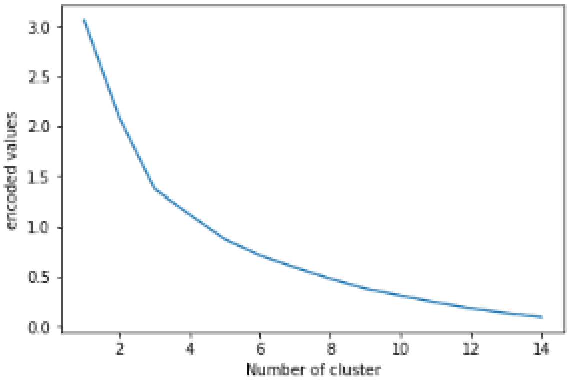


It is evident from the graph that DownTown Toronto has a very good distribution of venue categories of our choice. With DownTown Toronto as our preferred locality, we plot a map to show the distribution of Neighbourhoods in DownTown Toronto

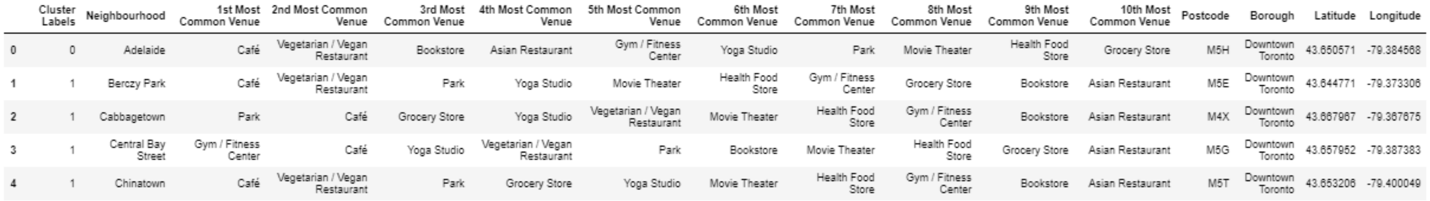


With this defined set of information, we then find the 10 most common venues for each Neighbourhood in DownTown Toronto. With the data scaled, we take the mean of the findings, grouping them by Neighbourhood.

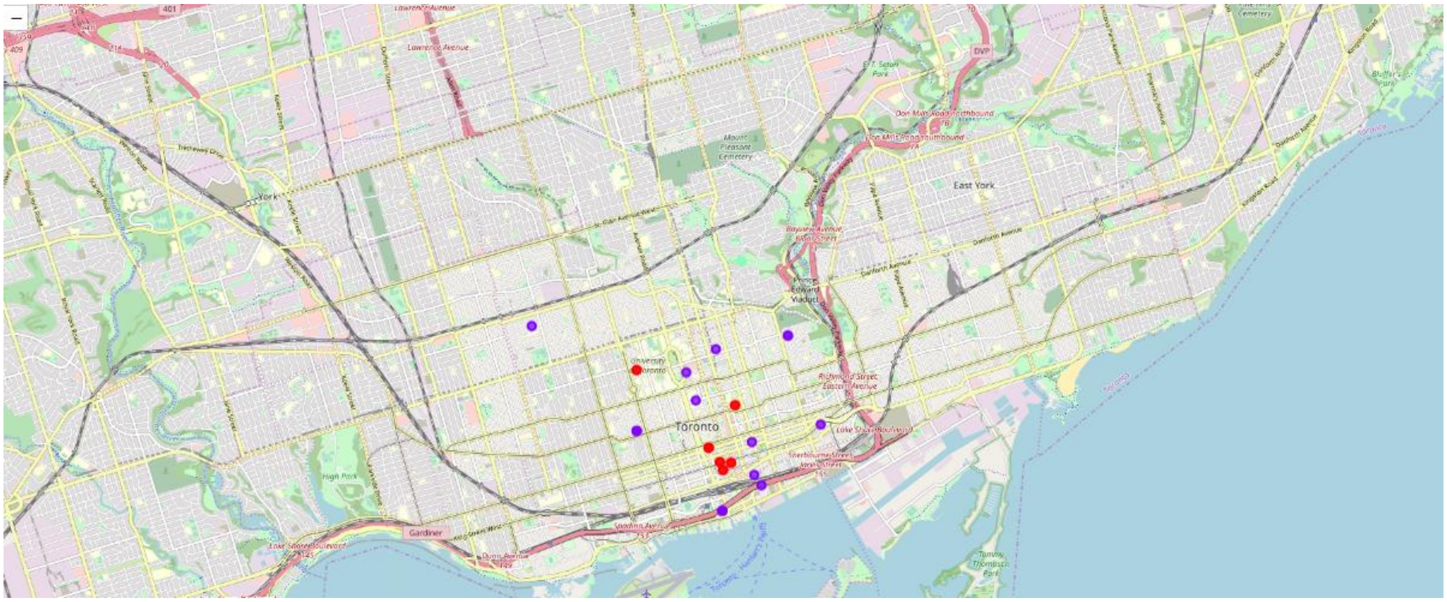
This is then followed by applying k means clustering on the data. The value of k is determined using the elbow method.



The elbow method shows the optimum value of k is 3. With this value of k, we perform k means clustering and plot a map with clusters marked.



	Cluster Labels	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	Postcode	Borough	Latitude	Longitude
0	0	Adelaide	Café	Vegetarian / Vegan Restaurant	Bookstore	Asian Restaurant	Gym / Fitness Center	Yoga Studio	Park	Movie Theater	Health Food Store	Grocery Store	M5H	Downtown Toronto	43.650571	-79.384568
1	1	Berczy Park	Café	Vegetarian / Vegan Restaurant	Park	Yoga Studio	Movie Theater	Health Food Store	Gym / Fitness Center	Grocery Store	Bookstore	Asian Restaurant	M5E	Downtown Toronto	43.644771	-79.373306
2	1	Cabbagetown	Park	Café	Grocery Store	Yoga Studio	Vegetarian / Vegan Restaurant	Movie Theater	Health Food Store	Gym / Fitness Center	Bookstore	Asian Restaurant	M4X	Downtown Toronto	43.667967	-79.367675
3	1	Central Bay Street	Gym / Fitness Center	Café	Yoga Studio	Vegetarian / Vegan Restaurant	Park	Bookstore	Movie Theater	Health Food Store	Grocery Store	Asian Restaurant	M5G	Downtown Toronto	43.657952	-79.387383
4	1	Chinatown	Café	Vegetarian / Vegan Restaurant	Park	Grocery Store	Yoga Studio	Movie Theater	Health Food Store	Gym / Fitness Center	Bookstore	Asian Restaurant	M5T	Downtown Toronto	43.653206	-79.400049



Discussion

From the analysis done on different types of venues in the city of Toronto, it is evident that the most viable locality as per our customized requirement is Down Town Toronto. Though this case study is limited to the city of Toronto, with a wider scope of area, it is possible, we may come up with a better areas or localities meeting the business requirements.

To achieve the goal, the data from Toronto was cleaned , analyzed, grouped, with each venue categorized, it gave us a better understanding of what information we have and how it can be leveraged to meet our requirement. Further in the analysis, by finding the optimum value of k for k means clustering, we have clustered the data into three major clusters.

Conclusion

By using the methodology used here to find out which locality best suits a person based on his or her needs, it can in future help people decide on a locality even before they relocate. It becomes easier for a person to gather such information and decide on a locality to relocate rather than move to a place first and then relocate later.