Coursera Applied Data Science Capstone

Introduction/Business Problem:

This project is aimed at visitors to the Toronto area. There are a lot of places to see, restaurants to dine in, and parks to walk or bike in spread throughout the city. I will explore the Toronto area by first clustering neighborhoods, and then by using Foursquare to visualize different venues and businesses located around the city. This information can be used by visitors to the Toronto area to find businesses of their choice within each neighborhood, i.e., downtown Toronto. Clustering the data will aid visitors in knowing where in the city they would enjoy staying.

Data:

The data I will be using is the borough and neighborhood names of the Toronto area combined with areas of general interest. The areas of interest can include parks, top attractions, and types of restaurants. The Toronto borough information will come from the postal code data found on Wikipedia, and it will be merged with data culled from the Foursquare API to return locations of interest in each area.

The data flow will be:

- Use Pandas and Beautiful Soup to import and scrape the postal code information
- Determine the latitude and longitude information for the boroughs
- Cluster Toronto based on borough names
- Use the Foursquare API to find the number of businesses (grouped by neighborhood)
- Sort the businesses based on their type
- Find the most popular businesses in a neighborhood
- Use k-means clusters to exhibit the businesses in each area

	PostalCode	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	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	1	Coffee Shop	Bakery	Park	Café	Pub	Hotel	Restaurant	Spa	French Restaurant	Mexican Restaurant
1	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	1	Coffee Shop	Café	Clothing Store	Hotel	Japanese Restaurant	Sandwich Place	Cosmetics Shop	Pizza Place	Middle Eastern Restaurant	Bookstore
2	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	1	Coffee Shop	Italian Restaurant	Café	Cocktail Bar	Clothing Store	Restaurant	Gym	Theater	Cosmetics Shop	Department Store
3	M4E	East Toronto	The Beaches	43.676357	-79.293031	1	Coffee Shop	Health Food Store	Pub	Monument / Landmark	Malay Restaurant	Market	Martial Arts School	Mediterranean Restaurant	Men's Store	Mexican Restaurant
4	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	1	Coffee Shop	Cocktail Bar	Sandwich Place	Bakery	Seafood Restaurant	Vegetarian / Vegan Restaurant	Beer Bar	Farmers Market	Sporting Goods Shop	Beach

Above is an example of the output from the data flow.

Methodology:

The data flow is driven by a step-down process where data is further refined in each step. The initial steps include gathering information and parsing it for select data. Next, the Foursquare API is utilized to gain geographical information about the select areas. Finally, the area is separated into five different clusters (see below), and each cluster is analyzed for its most common venue types.

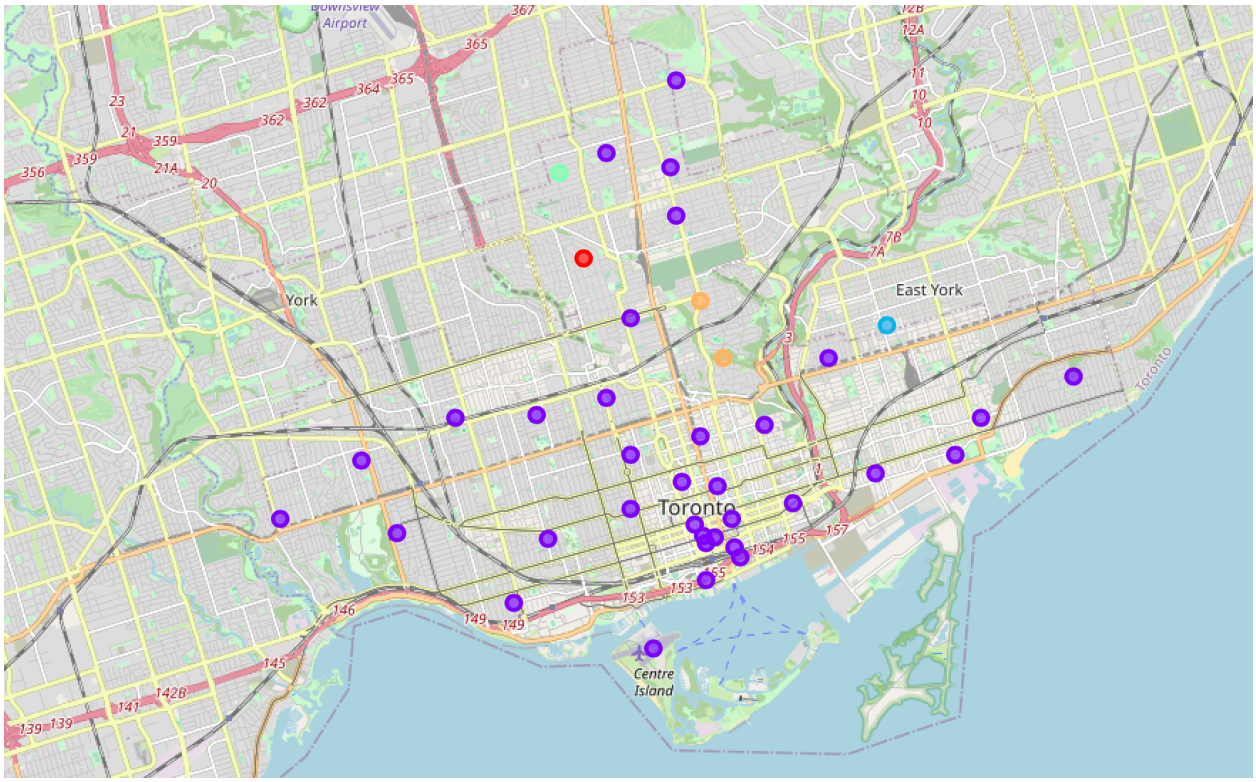
```
In [22]: kclusters = 5

         toronto_denc_grouped_clustering = toronto_denc_grouped.drop('Neighborhood', 1)

         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(toronto_denc_grouped_clustering)
         kmeans.labels_[0:10]
```

```
Out[22]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int32)
```

Results:



The map above is an example of the various clusters that were created using the k-means method. Each color represents a different cluster with red indicating cluster zero, purple representing cluster one, blue being cluster two, aqua as cluster three, and orange indicating cluster four.

Cluster 0

The table below shows that cluster zero occurs in Central Toronto and its most common attraction is a hiking trail.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
21	Central Toronto	0	Trail	Jewelry Store	Mexican Restaurant	Sushi Restaurant	Yoga Studio

Cluster 1

Cluster one occurs mainly in downtown Toronto and its most common venue is a coffee shop, with different types of food-centric businesses being the second most frequent.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Downtown Toronto	1	Coffee Shop	Bakery	Park	Café	Pub
1	Downtown Toronto	1	Coffee Shop	Café	Clothing Store	Hotel	Japanese Restaurant
2	Downtown Toronto	1	Coffee Shop	Italian Restaurant	Café	Cocktail Bar	Clothing Store
3	East Toronto	1	Coffee Shop	Health Food Store	Pub	Monument / Landmark	Malay Restaurant
4	Downtown Toronto	1	Coffee Shop	Cocktail Bar	Sandwich Place	Bakery	Seafood Restaurant
5	Downtown Toronto	1	Coffee Shop	Sandwich Place	Sushi Restaurant	Italian Restaurant	Japanese Restaurant

Cluster 2

The table below shows cluster two's most common attraction is a park located in the East York/ East Toronto neighborhood.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	East York/East Toronto	2	Park	Convenience Store	Yoga Studio	Molecular Gastronomy Restaurant	Mac & Cheese Joint

Cluster 3

Cluster three is in central Toronto. A home service is the top venue, while a garden and yoga studio are second and third.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
19	Central Toronto	3	Home Service	Garden	Yoga Studio	Monument / Landmark	Mac & Cheese Joint

Cluster 4

The final cluster's most common venue is a park. This cluster serves both central and downtown Toronto.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
29	Central Toronto	4	Park	Restaurant	Playground	Mac & Cheese Joint	Malay Restaurant
33	Downtown Toronto	4	Park	Playground	Trail	Molecular Gastronomy Restaurant	Mac & Cheese Joint

Conclusion:

The ability to cluster data relative to the types of venues that are in proximity to each other allows a visitor to the Toronto area the ability to visualize their surroundings. For example, an individual that seeks nature and peacefulness may choose to stay in the cluster three area to have access to a garden, yoga studio, and landmark. Someone with kids may choose the cluster four area for its access to a park and playground. In conclusion, incorporating the Foursquare geographical information in a mapped form based on most popular venue arms visitors with the knowledge they need to make an educated choice on the area in which they stay.