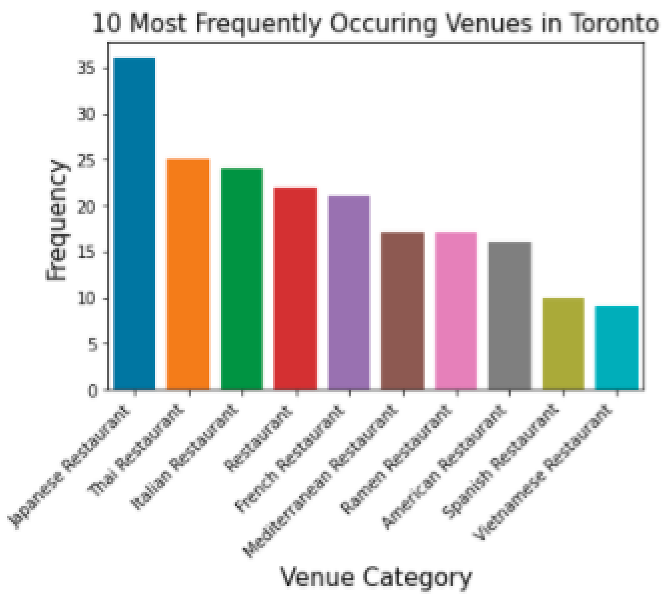
Applied Data Science Capstone Project – The Restaurant Battle of Neighborhoods

Introduction/Business Problem:

Japanese cuisine is one of the most famous cuisine in the world. This is because of they use only the best ingredients for their food. A client of mine is planning to supply some ingredients to Japanese restaurant that are located at Downtown Toronto. With this, our goal is to generate clusters of these restaurant for better service and deliveries.

Based on initial data gathered, there are a total of 36 Japanese Restaurant around Downtown Toronto, as we can see on the figure below:



For our client to have a better and more efficient way delivering the ingredients, the goal is to provide a clusters of Japanese restaurant.

Description of Data

As instructed by the assigned task, the data that will be used will come mainly from foursquare. Foursquare is an American Technological Company from New York focusing on location data. We will use their product, foursquare data, to get the location data about the Japanese restaurants in Toronto.

The data is not only limited to Foursquare data, we can also use other data. For this project we will also consider data coming from Wikipedia. Using beautifulsoup, we can scrape the data Neighborhood and

Boroughs in Toronto. We will then merge the Wikipedia data with the COCL data which has the coordinates of the Neighborhood in Toronto.

Methodology

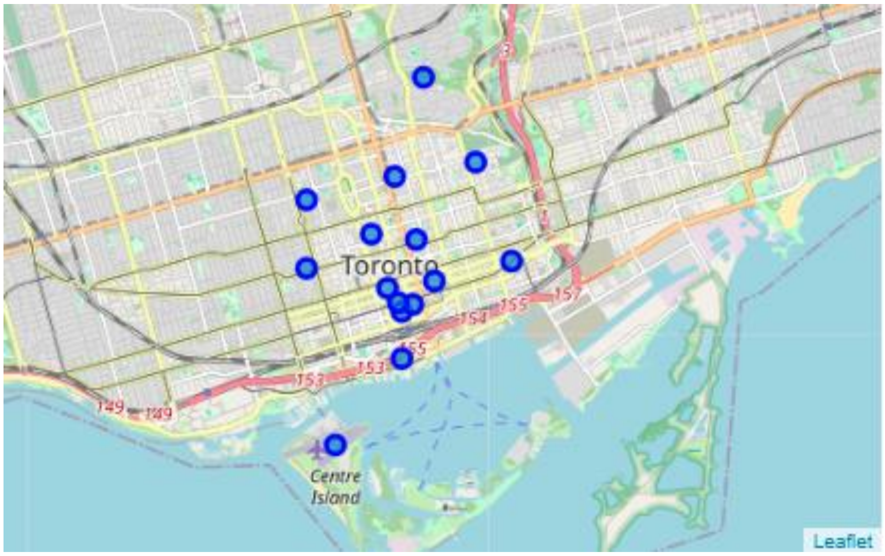
First create a data frame through webscraping of Wikipedia page. Using *beautifulsoup* function, the Wikipedia page: *'https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M'* will be scraped, cleaned, and preprocessed. It will generate the below data frame:

	Postcode	Borough	Neighborhood
0	M1B	Scarborough	Malvern
1	M1C	Scarborough	Rouge Hill
2	M1E	Scarborough	Guildwood
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Since the data frame above is still lacking the necessary data like coordinates, another data frame will be generated from *'https://cocl.us/Geospatial_data'*. This data frame will contain the coordinates data and will be merged with the other data frame. For this Capstone Project, only the borough of Downtown Toronto will be considered. After all that manipulation, the final data frame will look like this:

	Postcode	Borough	Neighborhood	Latitude	Longitude
44	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529
45	M4X	Downtown Toronto	St. James Town	43.667967	-79.367675
46	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160
47	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
48	M5B	Downtown Toronto	Garden District	43.657162	-79.378937
49	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
50	M5G	Downtown Toronto	Bay Street	43.657952	-79.387383
51	M5H	Downtown Toronto	Richmond	43.650571	-79.384568
52	M5J	Downtown Toronto	Harbourfront	43.640816	-79.381752
53	M5K	Downtown Toronto	Toronto Dominion Centre	43.647177	-79.381576
54	M5L	Downtown Toronto	Commerce Court	43.648198	-79.379817
57	M5S	Downtown Toronto	University of Toronto	43.662696	-79.400049
58	M5T	Downtown Toronto	Kensington Market	43.653206	-79.400049
59	M5V	Downtown Toronto	CN Tower	43.628947	-79.394420
60	M5X	Downtown Toronto	First Canadian Place	43.648429	-79.382280

Now, we can print the map of Toronto containing the selected neighborhood:



For us to collect the other necessary data like restaurant location data, a Foursquare API Credentials must be established first. This will allow the user to extracted and generate data frames from Foursquare. After creation of the Foursquare API, data can now be extracted. The venue name and category were extracted and resulted to a data frame below:

	name	categories	lat	lng
0	BATLgrounds	Athletics & Sports	43.647088	-79.351306
1	Pan Am Path, Don Landing	Trail	43.653752	-79.350744
2	McCleary Park	Baseball Field	43.652116	-79.340865
3	BATL Grounds	Recreation Center	43.647160	-79.351525

Now that the necessary data can be extracted, the restaurant category and frequency will be extracted and made into a data frame. Using "One Hot coding", the restaurant data will be expanded to its unique data. The data frame will look something like this:

	Neighborhood	American Restaurant	Caribbean Restaurant	Doner Restaurant	French Restaurant	Greek Restaurant	Indian Restaurant	Italian Restaurant
1	Rosedale	0	0	0	0	0	0	0
2	Rosedale	0	0	0	0	0	0	0
3	Rosedale	0	0	0	0	0	0	1
4	Rosedale	0	0	0	0	0	0	0
5	Rosedale	0	0	0	0	0	1	0

5 rows × 23 columns

This will be the resulting data frame after grouping the data with respect to their neighborhood:

	Neighborhood	American Restaurant	Caribbean Restaurant	Doner Restaurant	French Restaurant	Greek Restaurant	Indian Restaurant	Italian Restaurant
0	Bay Street	1	0	1	1	0	0	0
1	CN Tower	0	1	0	2	0	0	3
2	Church and Wellesley	2	0	1	2	0	0	1
3	Commerce Court	1	0	0	2	0	0	1
4	First Canadian Place	1	0	0	1	0	0	1
5	Garden District	2	0	0	1	0	0	0
6	Harbourfront	1	1	0	2	0	0	2
7	Kensington Market	1	1	1	1	0	0	1
8	Regent Park	1	0	0	1	0	0	1
9	Richmond	1	0	0	1	0	0	2
10	Rosedale	0	2	0	2	0	3	7
11	St. James Town	3	1	0	2	1	2	2
12	Toronto Dominion Centre	1	0	0	2	0	0	2
13	University of Toronto	1	0	1	1	0	0	1

14 rows × 23 columns

For this capstone project we will only consider the top 10 neighborhood and only the Japanese restaurants. We will arrive with this data frame:

	Neighborhood	Total Restaurants	Japanese Restaurants
0	Bay Street	17	3
1	CN Tower	15	1
2	Church and Wellesley	17	2
3	Commerce Court	14	2
4	First Canadian Place	12	2
5	Garden District	14	3
6	Harbourfront	20	3
7	Kensington Market	15	1
8	Regent Park	12	3
9	Richmond	18	4
10	Rosedale	27	3
11	St. James Town	32	5
12	Toronto Dominion Centre	15	2
13	University of Toronto	17	2

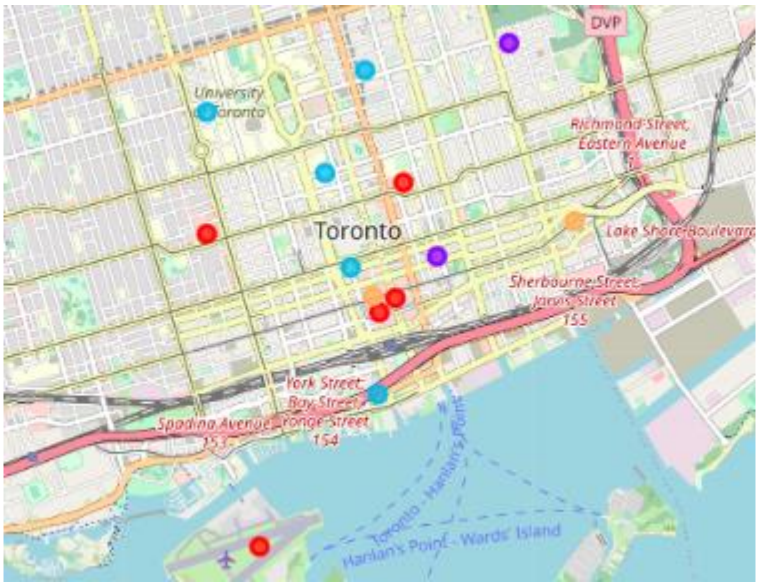
The neighborhood column will be dropped since it is not needed in the kmeans clustering. For k-means clustering, we will be using '*k = 5*'. After applying k-means, our final data frame will look like this:

```
# set number of clusters
kclusters = 5

grouped_clustering = df_restaurants.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
]: array([2, 0, 2, 0, 4, 0, 2, 0, 4, 2], dtype=int32)
```

```
# add clustering labels
df_restaurants.insert(0, 'Cluster Labels', kmeans.labels_)
```

```
# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborh
data_merged = df_final.join(df_restaurants.set_index('Neighborhood'), on='Neighborhood'

data_merged.head()
```

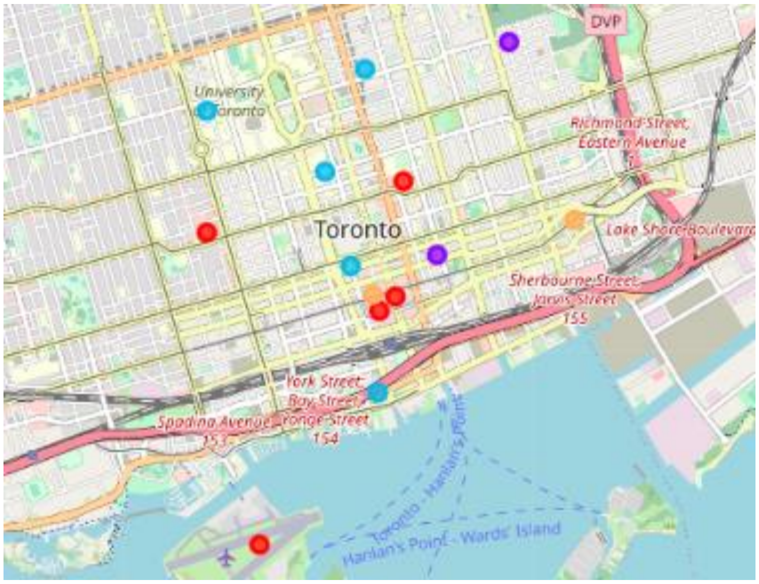


Results

And here already comes the result:

	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	Bay Street	Japanese Restaurant	Restaurant	Ramen Restaurant	Thai Restaurant	Spanish Restaurant	Middle Eastern Restaurant	Mediterranean Restaurant	American Restaurant	French Restaurant	Doner Restaurant
1	CN Tower	Italian Restaurant	French Restaurant	Seafood Restaurant	Ramen Restaurant	Spanish Restaurant	Thai Restaurant	Japanese Restaurant	Mediterranean Restaurant	Mexican Restaurant	Caribbean Restaurant
2	Church and Wellesley	Thai Restaurant	American Restaurant	Japanese Restaurant	French Restaurant	Restaurant	Tapas Restaurant	Mediterranean Restaurant	Ramen Restaurant	Italian Restaurant	Doner Restaurant
3	Commerce Court	French Restaurant	Japanese Restaurant	American Restaurant	Mexican Restaurant	Thai Restaurant	Spanish Restaurant	Seafood Restaurant	Restaurant	Ramen Restaurant	Mediterranean Restaurant
4	First Canadian Place	Japanese Restaurant	Thai Restaurant	American Restaurant	Spanish Restaurant	Restaurant	Mediterranean Restaurant	Ramen Restaurant	Italian Restaurant	French Restaurant	Vietnamese Restaurant

The data frame above are the most common venue and we assigned five different cluster labels 1 to 4. With this we can create a cluster specific color on map, using folium function. Below is the clustering of map:



We can see here the most common venue of Japanese restaurant per neighborhood.

Discussion

Here are the five clusters that we can follow to provide better service and deliveries:

Cluster 5 (Delivery Route 1):

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total Restaurants	Japanese Restaurants
47	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636	4	12	3
60	M5X	Downtown Toronto	First Canadian Place	43.648429	-79.382280	4	12	2

Cluster 4 (Delivery Route 2):

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total Restaurants	Japanese Restaurants
44	M4W	Downtown Toronto	Rosedale	43.679563	-79.377529	3	27	3

Cluster 3 (Delivery Route 3):

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total Restaurants	Japanese Restaurants
46	M4Y	Downtown Toronto	Church and Wellesley	43.665860	-79.383160	2	17	2
50	M5G	Downtown Toronto	Bay Street	43.657952	-79.387383	2	17	3
51	M5H	Downtown Toronto	Richmond	43.650571	-79.384568	2	18	4
52	M5J	Downtown Toronto	Harbourfront	43.640816	-79.381752	2	20	3
57	M5S	Downtown Toronto	University of Toronto	43.662696	-79.400049	2	17	2

Cluster 2 (Delivery Route 4):

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total Restaurants	Japanese Restaurants
45	M4X	Downtown Toronto	St. James Town	43.667967	-79.367675	1	32	5
49	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	1	32	5

Cluster 1 (Delivery Route 5):

	Postcode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	Total Restaurants	Japanese Restaurants
48	M5B	Downtown Toronto	Garden District	43.657162	-79.378937	0	14	3
53	M5K	Downtown Toronto	Toronto Dominion Centre	43.647177	-79.381576	0	15	2
54	M5L	Downtown Toronto	Commerce Court	43.648198	-79.379817	0	14	2
58	M5T	Downtown Toronto	Kensington Market	43.653206	-79.400049	0	15	1
59	M5V	Downtown Toronto	CN Tower	43.628947	-79.394420	0	15	1

Conclusion

Overall, we achieved the main goal of this capstone project. That is to provide delivery route through clustering for a client.