Capstone Project – Paris to Toronto

Samuel O

1. Introduction

1.1 Problem Background

Toronto is by far one of the greatest cities in the Canada and the world at large. With its multicultural diversity and capitalistic nature, it possesses a booming atmosphere for any business to successfully run. The heart of Canada's finance and banking industries lies in Toronto. It is no surprise that it is the Financial capital of Canada. From Bay Street's high-rises, which house corporations, law firms and the Toronto Stock Exchange to IBM, PepsiCo, etc., The city has shown its financial power and range of opportunities from fashion to technology, legal services, sports, real estate, tourism and finance amongst others. It really is a wonder to see and to live in!

For any young graduate looking to progress in his/her career, Toronto provides the perfect atmosphere of job opportunities, culture, housing, places, dining options ranging from cozy cafes to chic power lunch restaurants, etc. Moving from a different city to Toronto can be quite a change thus it is nice to switch easily. As such, it is very important for any newcomer interested in the business space of Toronto to analyze specifically the similarities between the current residency location and Toronto to ease the change of environment. These analyses would enhance decision making and eventually provide an edge to any person because he/she feels comfortable and is ready to withstand competition and survive in the Toronto scene. These analyses would improve understanding and prove to be valuable.

1.2 Problem Description

As a young graduate looking and pursuing a career in Data Science. I just received an offer of a lifetime to work with great engineers at IBM!

The sad part? I must leave Paris for Toronto, Canada. The IBM Toronto Software Lab is the largest software development laboratory in Canada, and IBM's third largest software lab. It is located in Markham, a community in Toronto, Canada.

Paris is known for its culture, sights, venues, rich diversity and in all a calming environment. Toronto also boasts of the same. For me, moving to Toronto would be quite the challenge and the environment I will have to go to is very important to me. There lies my problem, I must find a location that reminds me of home and has the potential to become my new home!

Several venues remind me the most of home but majorly the variety of food to select from after a very long day. Also, the several amenities ranging from the gyms to the clubs to transportation, etc.

The challenge would be to locate a similar community that benefits my transition from Paris to Toronto.

1.3 Interest

This research will be very interesting for any youth or career person moving from one city to another. It would provide a methodology and approach on how to locate the similarities between the two cities and aid decision making. Others who could be interested in this data are researchers on socio-geographical studies of cities by looking at the social factors as well as the geographical factors.

2. **Data Acquisition and Cleaning**

2.1 **Data Sources**

To solve my problem, I will need to source for data that would show me the top venues and locations at the neighbourhood I reside at in Paris. As such I would require:

1.Data on the Latitude and Longitude of my location (Using Nominatim or Google Map Locator)

2.Venues located in the neighbourhood at a specific radius and at a time version of 7th July, 2020. (Applying the Foursquare API)

With this dataset, I have a basis to begin my research.

I would then need to gather data for Toronto (latitude, longitude, venues, ratings):

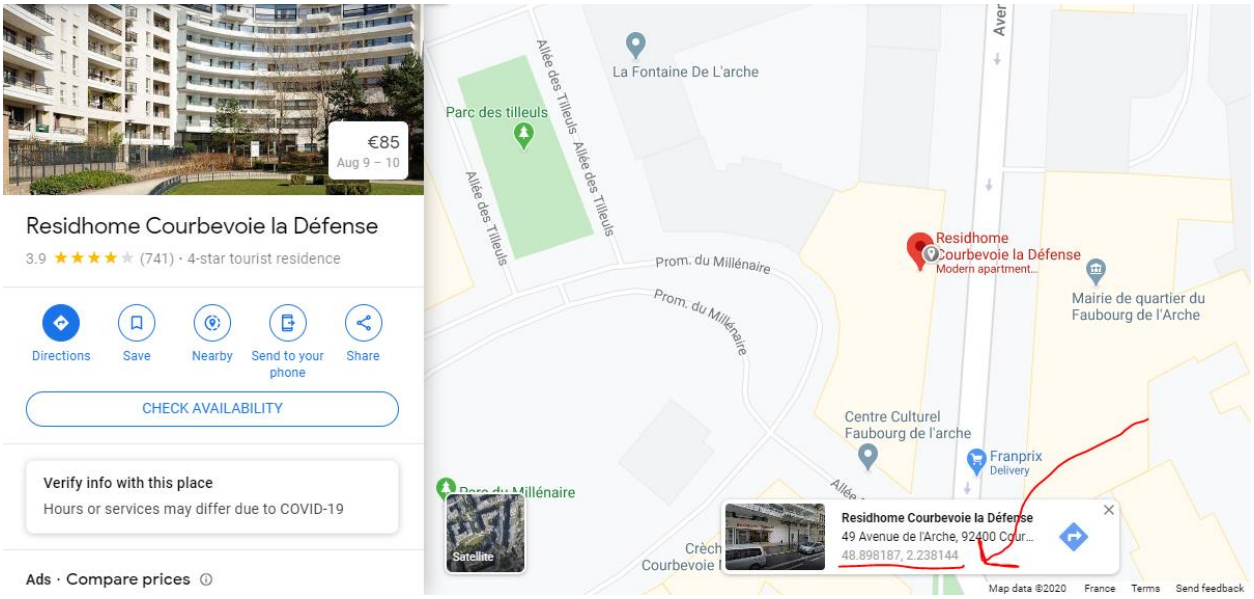
These data would be sourced from a Wikipedia page containing the Postal Codes of Ontario and a CSV File containing Geospatial Data of the several Ontario Postal Codes:

1.List of Postal Codes for Ontario: (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

2. Geospatial CSV (http://cocl.us/Geospatial_data)

2.2 **Data Cleaning**

The data on my location in Paris was sourced using Google Locator Maps and the Latitude and Longitude were found thus



With this and using the Foursquare API, the venues located at a radius of 650m to this residential area in Paris were located and then read into a DataFrame for further analysis later.

	name	categories	lat	lng
0	Parc du Millénaire	Park	48.898194	2.236577
1	Thaïoria	Thai Restaurant	48.900409	2.239551
2	Hôtel Pullman Paris La Défense	Hotel	48.895096	2.239006
3	Starbucks	Coffee Shop	48.896689	2.238164
4	Monoprix	Supermarket	48.896720	2.236800
5	La French Touch	Burger Joint	48.900660	2.232121
6	Grande Arche de la Défense	Monument / Landmark	48.892565	2.235882
7	Place Carpeaux	Plaza	48.893531	2.238529
8	Sushi Fukunoya	Japanese Restaurant	48.900200	2.238700
9	Fnac	Department Store	48.892625	2.239585
10	Villa Min	Korean Restaurant	48.899282	2.245689
11	Decathlon	Sporting Goods Shop	48.892972	2.240244
12	Allée de l'Arche	Plaza	48.894852	2.239296
13	So Thaï	Thai Restaurant	48.896766	2.245842

The data from the Wikipedia page which shows the Postal Codes of Ontario, Canada and the Geospatial data were combined (Will be shown further in Methodology)

This information would be classified by Boroughs and Neighbourhoods and then rated based on the top locations in such areas

In turn the communities would be clustered to aid me by visual and content analysis of the similarities between those clusters and my current location.

All these will in turn aid my Decision Making in picking a similar location and will ease my transition from Paris to Toronto

3. Methodology

3.1 Data Scraping and Wrangling

3.1.1 Libraries

To proceed with the require data scraping for the information we need onto this project, the following Libraries were installed and imported

1. **Numpy** library to handle data in a vectorized manner

2. **Pandas** library for data analysis

3. **Json** library to handle JSON files

4. **Requests** library to handle requests

5. **Json_normalize** to transform JSON file into a pandas dataframe

6. **Nominatim** to convert an address into latitude and longitude values

7. Lxml to be able to read html text file

8. **Folium** map rendering library and its **plugins**

9. **Matplotlib** and associated plotting modules for Data Visualization

10. **Seaborn** Library for Data Visualization

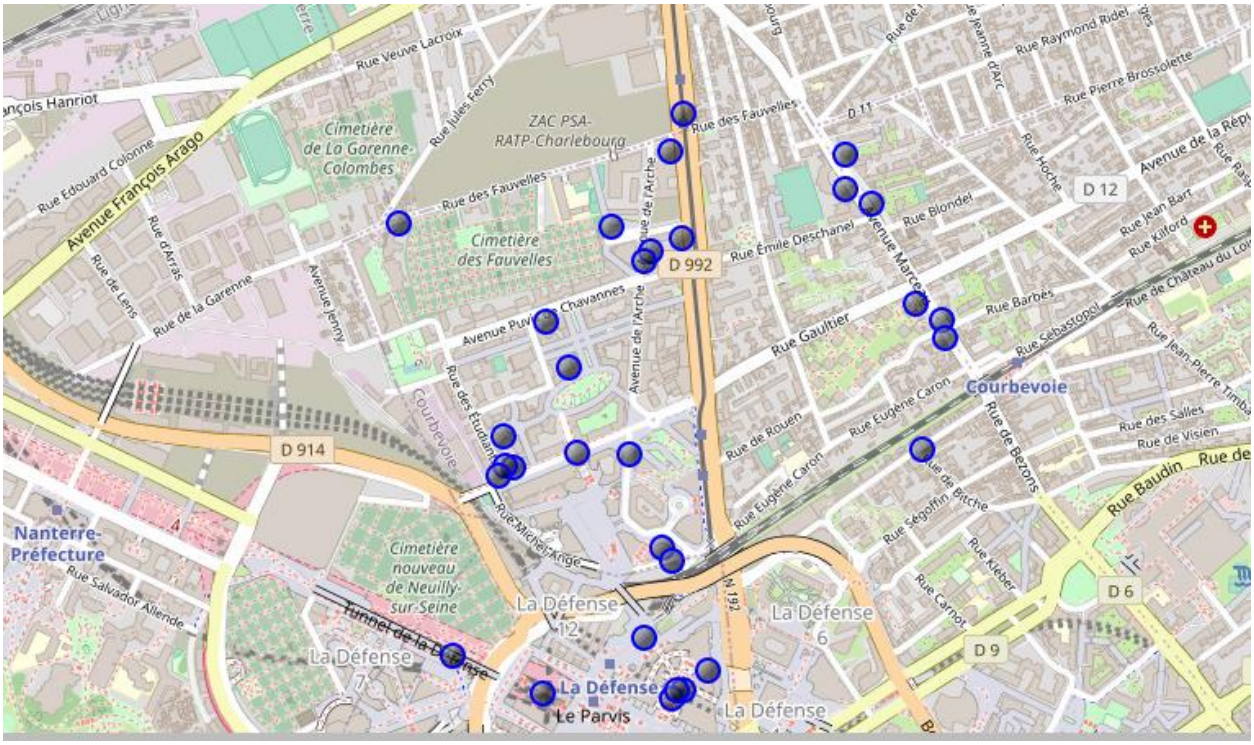
11. K-means from Sklearn.cluster for the Machine Learning and Clustering Analysis

3.1.2. Paris Location and Data

The geographical coordinates of Current Location in Paris were identified using Google Maps Locator as 48.898163, 2.238128 (as shown in the picture above)

The next approach was to identify the venues around the Current Location in Paris and in turn read that information into a Pandas DataFrame for further analysis later. This was done by utilizing the Foursquare API (Table shown above in Chapter 2)

To aid understanding of the venues and their locations in Paris, the Folium Library was used to create a map of the Neighborhood as shown below:



3.1.3 Toronto Location and Data

Lxml and Json were utilized to read the table of postal codes and locations in Ontario from the Wikipedia file. In turn the resulting data had data reflecting Postal Code, Borough and Neighbourhood. This data in turn had to scraped to remove Rows that had ‘Not Assigned’ for their Boroughs as such data is vague and not required for our analysis. The resulting Data Frame was thus:

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

The Geospatial data was read from the link as stated in Chapter 2 using the Pandas Library and this data was presented in a DataFrame thus:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

The next step was to merge both DataFrames by corresponding Postal Codes to eventually have one single DataFrame with the Postal Codes, Borough, Neighbourhood and Geospatial Data (Latitude and Longitude):

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Since my target location are neighbourhoods in Toronto, the next step was to scrape off other Boroughs and focus on the Boroughs of Toronto (Central Toronto, Downtown Toronto, East and West Toronto):

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049
1	M4S	Central Toronto	Davisville	43.704324	-79.388790
2	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160
3	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307
4	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678

Folium Map was then utilized to have a visualization of the several communities in Toronto

To compare the current residential neighborhood in Paris and the target neighbourhoods in Toronto, the aim was to locate data that possessed information about the venues in these neighbourhoods.

This data in turn would be the focus of our clustering and machine learning analysis.

Using the Foursquare API and a radius of 650m, the venues in each neighbourhood was requested in this format:


```
{'meta': {'code': 200, 'requestId': '5f2fbc10e48ac773fa54994c'},
 'response': {'headerLocation': 'Deer Park',
  'headerFullLocation': 'Deer Park, Toronto',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 55,
  'suggestedBounds': {'ne': {'lat': 43.69226230585, 'lng': -79.39197457567654},
   'sw': {'lat': 43.680562294149986, 'lng': -79.40812402432347}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '55c78cef498ec4095e9fba41',
       'name': 'LCBO',
       'location': {'address': '111 St. Clair West',
        'lat': 43.686990631074885,
        'lng': -79.39923810519545,
        'labeledLatLngs': [{'label': 'display',
          'lat': 43.686990631074885,
          'lng': -79.39923810519545}],
        'distance': 91,
        'cc': 'CA',
        'city': 'Toronto',
        'state': 'ON',
        'country': 'Canada',
```


This data had to be normalized by Json and then read into a Pandas DataFrame. In turn, the corresponding location of each Venue and their neighbourhood was scraped for further analysis:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	LCBO	43.686991	-79.399238	Liquor Store
1	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	The Market By Longo's	43.686711	-79.399536	Supermarket
2	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	The Bagel House	43.687374	-79.393696	Bagel Shop
3	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	Capocaccia Café	43.685915	-79.393305	Italian Restaurant
4	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	Scaramouche	43.681293	-79.399492	French Restaurant

The eventual data frame was **Dummied** and then grouped by Neighbourhoods by taking the mean of the frequency of occurrence of each venue category. The dataframe is thus:

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Aquarium	A Resta
0	Berczy Park	0.000000	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	0.0	0.0	0.01	0.0	
1	Brockton, Parkdale Village, Exhibition Place	0.017544	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	0.0	0.0	0.00	0.0	
2	Business reply mail Processing Centre, South C...	0.000000	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	0.0	0.0	0.00	0.0	
3	CN Tower, King and Spadina, Railway Lands,	0.000000	0.0	0.0	0.05	0.05	0.1	0.1	0.15	0.0	0.0	0.0	0.00	0.0	

With this data (which will be further used for Clustering), we have information of the most common venues of each neighborhood. This was used (as shown in the Notebook) to create a dataframe of the Top 10 Venues in each Neighborhood:

	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
0	Berczy Park	Coffee Shop	Hotel	Café	Bakery	Restaurant	Pub	Beer Bar	Japanese Restaurant	Seafood Restaurant	Cocktail Bar
1	Brockton, Parkdale Village, Exhibition Place	Coffee Shop	Café	Breakfast Spot	Supermarket	Nightclub	Sandwich Place	Bakery	Restaurant	Gift Shop	Pharmacy
2	Business reply mail Processing Centre, South C...	Light Rail Station	Brewery	Park	Fast Food Restaurant	Burrito Place	Coffee Shop	Gym / Fitness Center	Recording Studio	Restaurant	Sandwich Place
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Terminal	Coffee Shop	Airport Lounge	Airport Service	Boutique	Tunnel	Sculpture Garden	Rental Car Location	Boat or Ferry	Pier
4	Central Bay Street	Coffee Shop	Café	Sandwich Place	Art Gallery	Bubble Tea Shop	Sushi Restaurant	Italian Restaurant	Burger Joint	Salad Place	Breakfast Spot
5	Christie	Grocery Store	Café	Park	Diner	Coffee Shop	Baby Store	Candy Store	Italian Restaurant	Nightclub	Restaurant
6	Church and Wellesley	Coffee Shop	Japanese Restaurant	Restaurant	Sushi Restaurant	Gay Bar	Diner	Café	Yoga Studio	Men's Store	Hotel

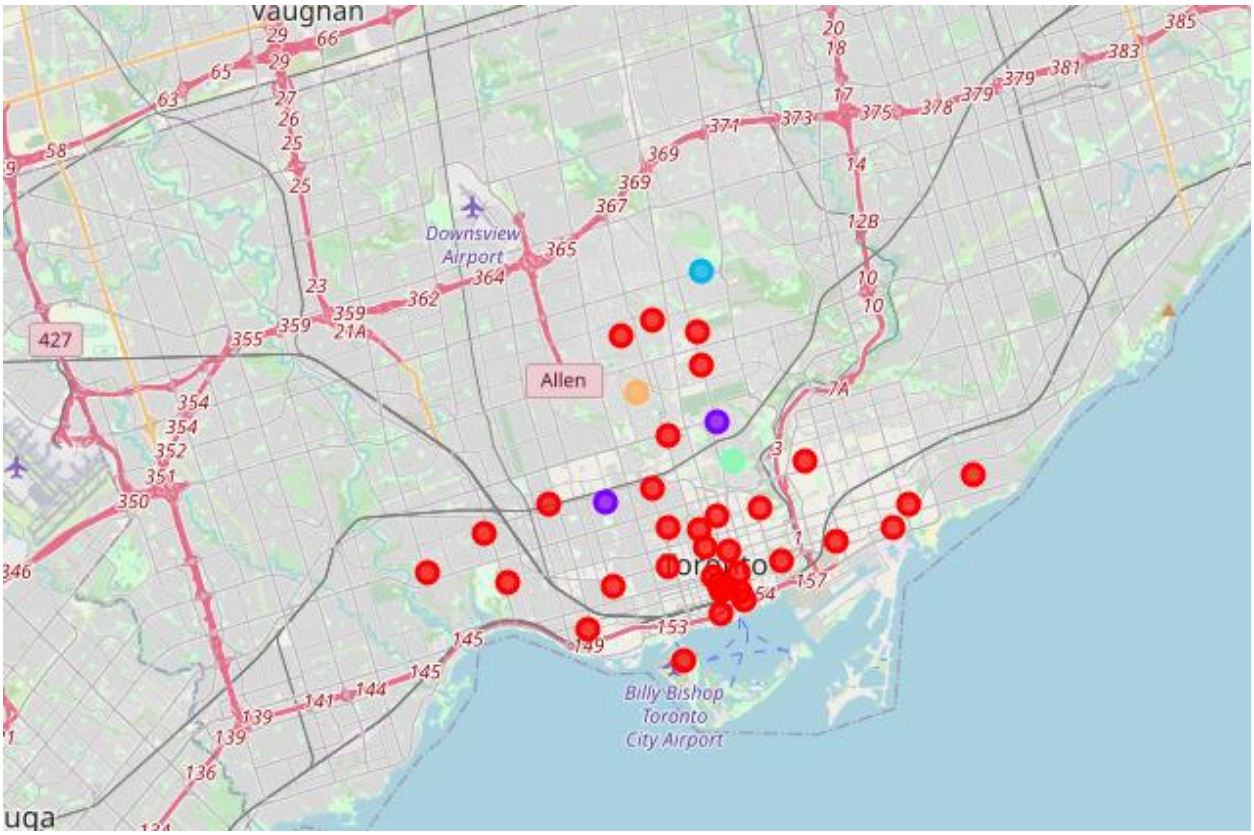
3.2 Data Science Methods and Machine Learning for Data Analysis

K-Means is an unsupervised algorithm that divides data into non-developing subsets (clusters) without any cluster internal structures. The aim is to minimize intra-cluster distances and maximize the inter-cluster distances.

To apply K-Means Clustering to our data frame, we initialized a number of Clusters (5 in this project) and the algorithm runs unsupervised to cluster our neighbourhoods based on similarities of venues. Thus we have 5 Clusters as shown below and CLUSTER LABELS assigning each neighbourhood to a Cluster:

	Postal Code	Borough	Neighbourhood	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
0	M4V	Central Toronto	Summerhill West, Rathnelly, South Hill, Forest...	43.686412	-79.400049	0	Coffee Shop	Italian Restaurant	Sushi Restaurant	Thai Restaurant	Gym	Light Rail Station	Pizza Place	Pub
1	M4S	Central Toronto	Davisville	43.704324	-79.388790	0	Pizza Place	Coffee Shop	Café	Italian Restaurant	Dessert Shop	Sandwich Place	Gym	Fast Food Restaurant
2	M4T	Central Toronto	Moore Park, Summerhill East	43.689574	-79.383160	1	Park	Tennis Court	Gym	Grocery Store	Thai Restaurant	Dumpling Restaurant	Distribution Center	Dive Bar
3	M5P	Central Toronto	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307	4	Bus Line	Jewelry Store	Trail	Sushi Restaurant	Yoga Studio	Doner Restaurant	Donut Shop	Dumpling Restaurant
4	M5R	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	0	Café	Coffee Shop	Park	Sandwich Place	Pub	Historic Site	Burger Joint	Indian Restaurant

To aid better understanding of the Clusters, the Folium Library is used to make a Geographical representation of each of the Clusters based on the K-Means Algorithm data:



We can then proceed to analyse these results and make observations.

4. Results and Discussions

Looking at each of the Clusters, there are striking similarities between Cluster 4 in Borough Central Toronto and the current Parisian Neighbourhood.

```
T_merge.loc[T_merge['Cluster Labels'] == 4, T_merge.columns[[1] + list(range(5, T_merge.shape[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	Central Toronto	4	Bus Line	Jewelry Store	Trail	Sushi Restaurant	Yoga Studio	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Electronics Store

The next step would be to locate the exact neighbourhood in Borough Central Toronto that bears this similarities to the Parisian Neighbourhood

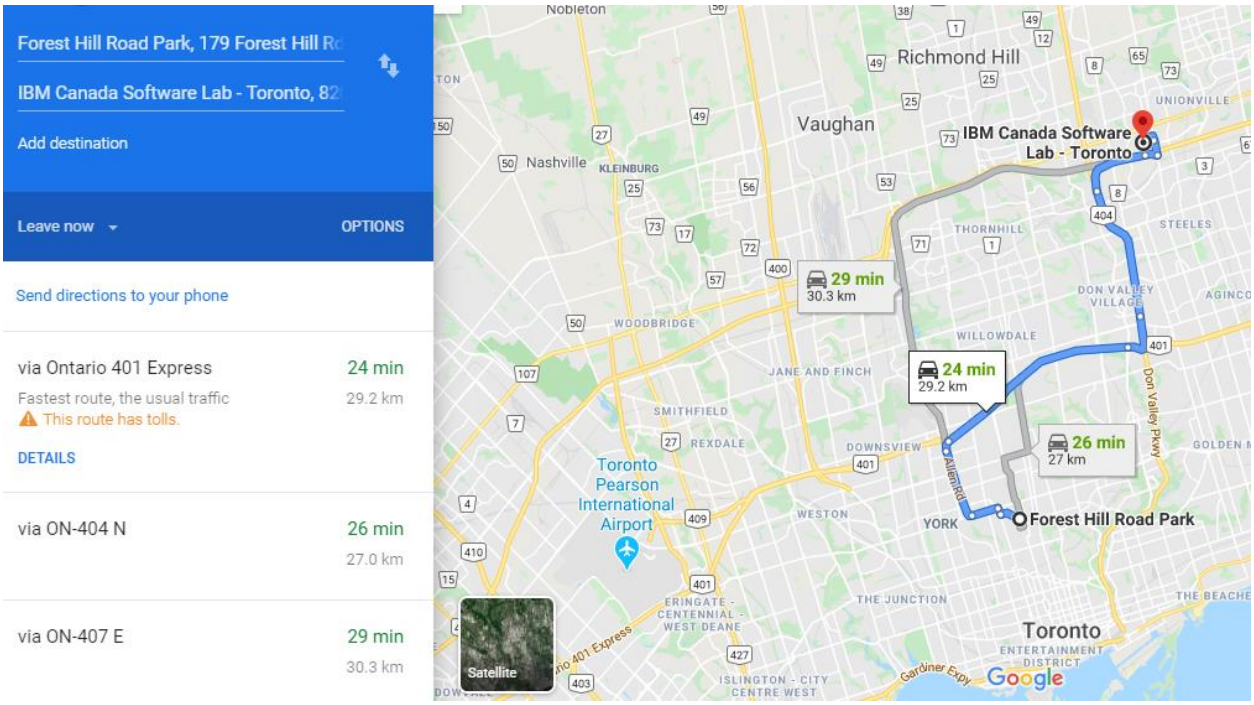
```
Postal Code                                                  M5P
Borough                                           Central Toronto
Neighbourhood           Forest Hill North & West, Forest Hill Road Park
Latitude                                                  43.6969
Longitude                                                -79.4113
Cluster Labels                                                  4
1st Most Common Venue                                    Bus Line
2nd Most Common Venue                                Jewelry Store
3rd Most Common Venue                                       Trail
4th Most Common Venue                             Sushi Restaurant
5th Most Common Venue                                  Yoga Studio
6th Most Common Venue                             Doner Restaurant
7th Most Common Venue                                    Donut Shop
8th Most Common Venue                           Dumpling Restaurant
9th Most Common Venue                 Eastern European Restaurant
10th Most Common Venue                            Electronics Store
Name: 3, dtype: object
```

This reflects that the Neighbourhoods with the closest similarities are **Forest Hill North & West and Forest Hill Road Park.**

5. **Observations of Discussions and Recommendations**

5.1 Observations

Looking at the locations of both Forest Hill Road Park and Forest Hill North & West, they are similar neighbourhoods to the Parisian Neighbourhood and by observation on Google Maps, they are located about 20 minutes by driving to the IBM Software Lab in Markham Toronto.



This Capstone project provided me insights on how to use data science analysis and methodology to tackle a presented problem and solve it with support of Data Visualization.

I believe this will help me in my Engineering Career as we progress into Industry 4.0 and the Digitalization of the Energy Industry.

5.2 Recommendations

Further Recommendations to this project would be to finetune the number of Clusters by possibly selecting 4 Clusters as against 5.

Other Recommendations would be better explanations on how to Push Python Files to Github locally without using the Atom Platform. This proved difficult and in turn resulted in Direct Uploading of Files

6. Conclusions

In Conclusion this project aided my decision in locating a similar neighbourhood in Toronto like my current one in Paris. This in turn will help me ease my transition to Toronto and work with IBM!

Thanks to IBM for this course and project.