DATA ACQUISITION

1.1 DATA SOURCES

For the purpose of this data analysis report, we obtain geographical data from the cities of Toronto, New York, and Paris. This includes information regarding the neighborhoods of each city and their coordinates (longitude and latitude) which were scraped and transformed into pandas data frames from the following websites.

1. Toronto: The Wikipedia page for the list of postal codes in Toronto and a csv file from Coursera for the geospatial coordinates.

Links:

1) Toronto Neighborhood Data :
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2) https://cf-courses-data.s3.us.cloud-objectstorage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701ENSkillsNetwork/labs_v1/Geospatial_Coordinates.csv

3) Venue data nearby: Foursquarer API

2. New York: I used the New York json file for the list of New York postal codes and coordinates, which were linked to their respective neighborhoods.

Links:

1) New York Neighborhood Data:

   New York City (JSON) : https://cocl.us/new_york_dataset

2) Venue data nearby: Foursquare API

3. Paris: This Wikipedia page for the list of areas of Paris wherein the coordinates were easily obtained from the table.

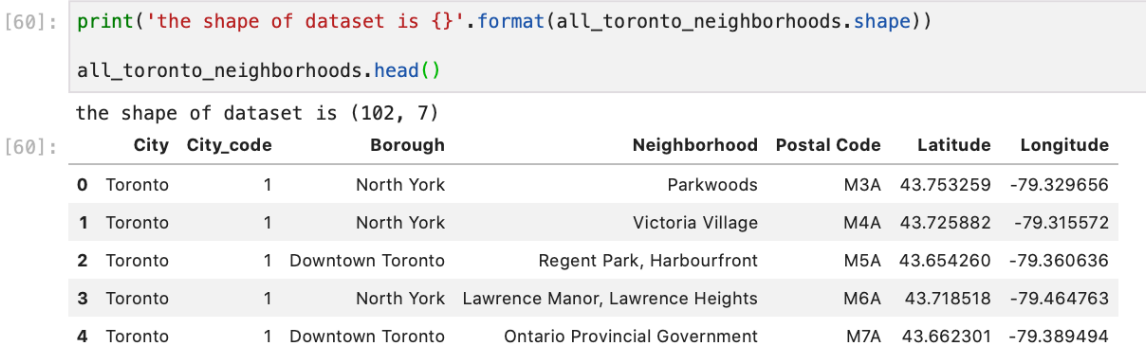
Links:

- https://en.wikipedia.org/wiki/Arrondissements_of_Paris

I imported dataframe from file of Paris_geo, which was obtained from https://en.wikipedia.org/wiki/Arrondissements_of_Paris.

This resulted in a dataset of 103 neighborhoods in Toronto, 306 neighborhoods in New York, and 29 neighborhoods in Paris.

Toronto neighborhoods has 102 neighborhoods and 9 boroughs.



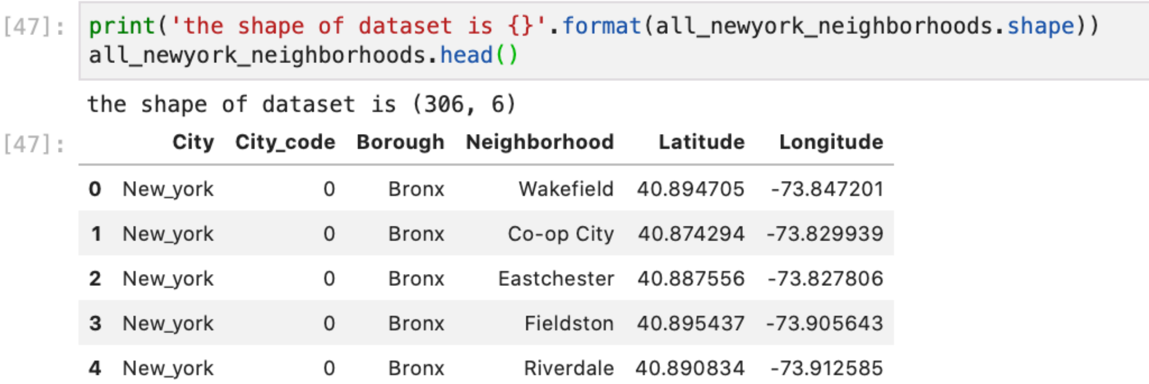
```
[60]: print('the shape of dataset is {}'.format(all_toronto_neighborhoods.shape))

      all_toronto_neighborhoods.head()

      the shape of dataset is (102, 7)
[60]:
```

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

New York neighborhoods has 306 neighborhoods and 5 boroughs.

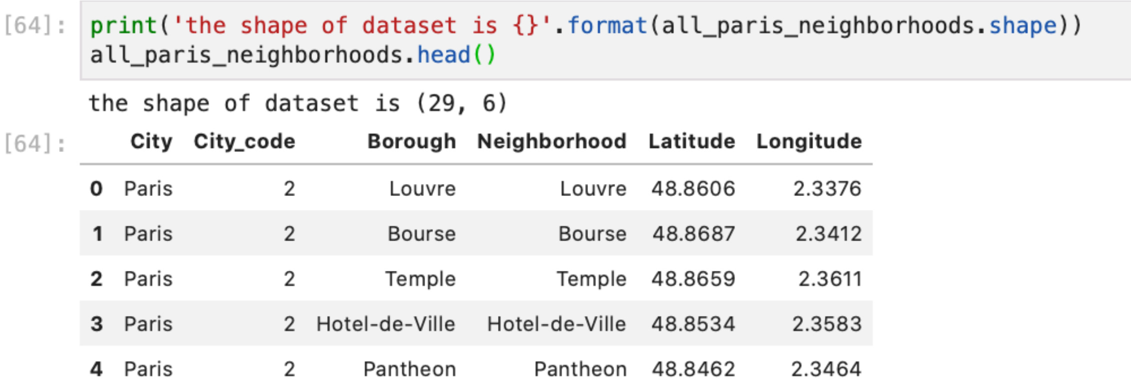


```
[47]: print('the shape of dataset is {}'.format(all_newyork_neighborhoods.shape))
      all_newyork_neighborhoods.head()

      the shape of dataset is (306, 6)
[47]:
```

	City	City_code	Borough	Neighborhood	Latitude	Longitude
0	New_york	0	Bronx	Wakefield	40.894705	-73.847201
1	New_york	0	Bronx	Co-op City	40.874294	-73.829939
2	New_york	0	Bronx	Eastchester	40.887556	-73.827806
3	New_york	0	Bronx	Fieldston	40.895437	-73.905643
4	New_york	0	Bronx	Riverdale	40.890834	-73.912585

Paris neighborhoods has 29 neighborhoods.



```
[64]: print('the shape of dataset is {}'.format(all_paris_neighborhoods.shape))
      all_paris_neighborhoods.head()

      the shape of dataset is (29, 6)
[64]:
```

	City	City_code	Borough	Neighborhood	Latitude	Longitude
0	Paris	2	Louvre	Louvre	48.8606	2.3376
1	Paris	2	Bourse	Bourse	48.8687	2.3412
2	Paris	2	Temple	Temple	48.8659	2.3611
3	Paris	2	Hotel-de-Ville	Hotel-de-Ville	48.8534	2.3583
4	Paris	2	Pantheon	Pantheon	48.8462	2.3464

The coordinates are used to decide the venues within a 500-meters radius of the neighborhoods. Therefore, the records on the venue's names and site groups around each coordinate is collected from the Foursquare API1 , and are limited to 100 venues for each neighborhood.

I use the following data analysis technique to recommend the business type to start up for a businessperson or entrepreneur.

The first technique is city analysis. In this analysis, we found and prepared data to discover how many place of venues are there, based on this result we can know the venue measure of each city. Additionally, we can examine the diversity of venues in each city, which may tell us how varied the urban venues can be and how interrelated between cities.

In order to get the objectives, we merged all city datasets together to do the clustering analysis, we need to find how the machine learning allocate all cities venues, and exclusively in each cluster how Toronto, New York and Paris venues are scattered? The essential hypothesis is that if two cities are similar, their distribution among clusters should be comparable.

The second way to analyze is borough / neighborhood analysis. we must practice Foursquare figures to investigate the top 10 venues in each district, so that we can catch what kinds of businesses are trendy there. The fundamental viewpoint here is the neighborhood market may have adequate capacity and keenness to opportune a new restaurant if the previously current restaurants here are located.