Coursera Capstone Project

Clustering neighborhood of Toronto

1. Introduction

- Toronto is a global city filled with vast opportunity and is home to an array of distinctive and dynamic neighborhoods that reflect the diversity of its population. The city is known for its vibrant arts and entertainment scene, incredible cultural festivals, delicious food, thriving sports culture, excellent shopping, beautiful parks and beaches, and much more.

- But, in order to open the business in Toronto and become success the owner need the understand the market and their competitor properly.

2. Problem

Our problem is simple question.

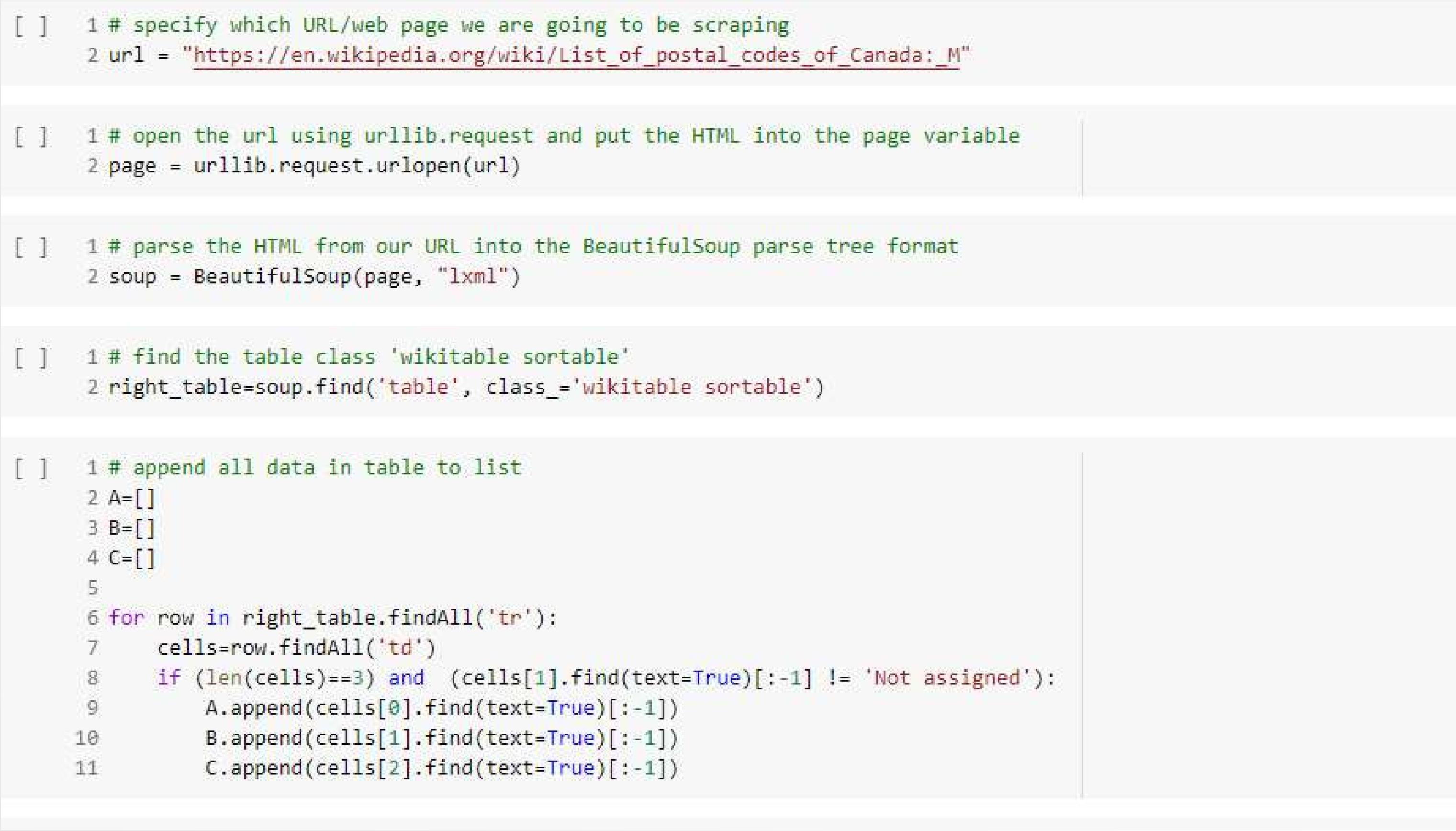
- "if someone is looking to open a restaurant, where would we recommend that they open it?

3. Data

- The data for this project has been collected from multiple sources.

3.1. Neighborhood

- The data of neighborhood in city of Toronto is obtained by web scraping using BeatifulSoup library in Python. The data is scaped from Wikipedia website.



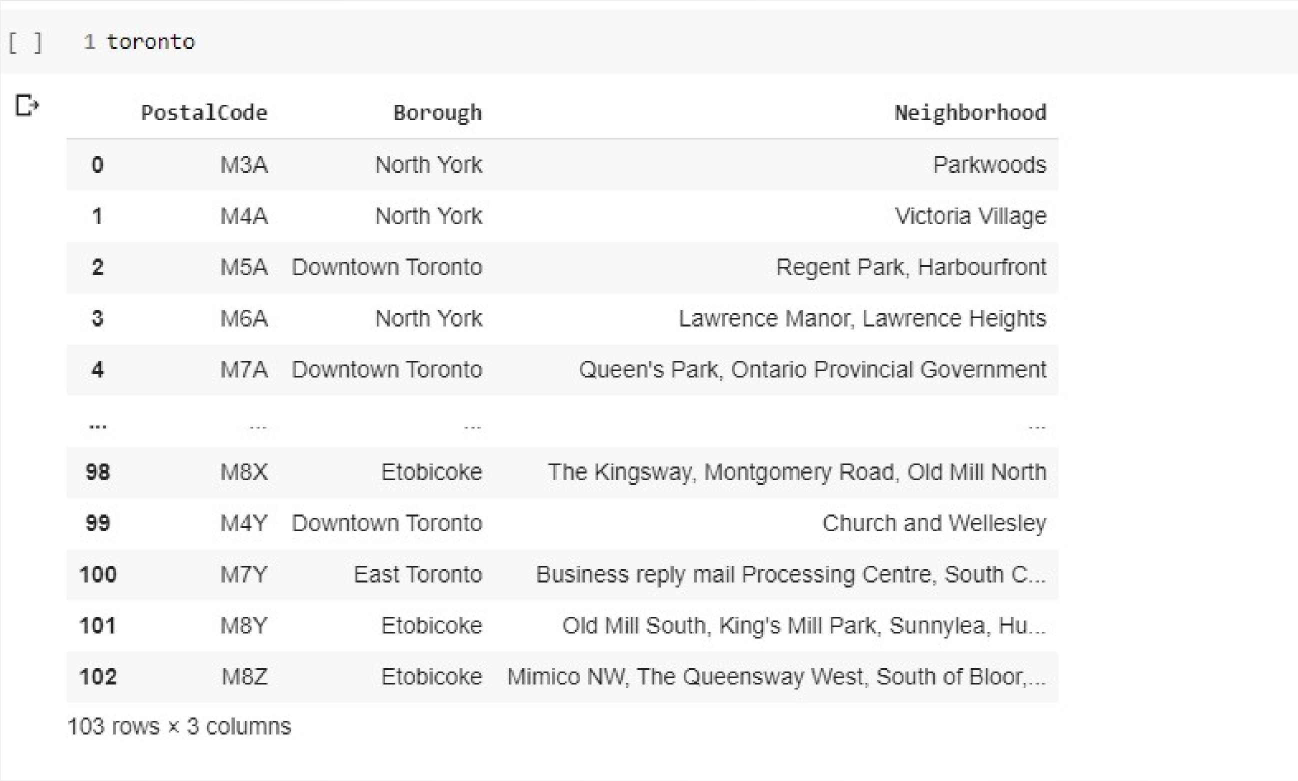
```
# specify which URL/web page we are going to be scraping
url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
```

```
# open the url using urllib.request and put the HTML into the page variable
page = urllib.request.urlopen(url)
```

```
# parse the HTML from our URL into the BeautifulSoup parse tree format
soup = BeautifulSoup(page, "lxml")
```

```
# find the table class 'wikitable sortable'
right_table=soup.find('table', class_='wikitable sortable')
```

```
# append all data in table to list
A=[]
B=[]
C=[]

for row in right_table.findAll('tr'):
    cells=row.findAll('td')
    if (len(cells)==3) and  (cells[1].find(text=True)[:-1] != 'Not assigned'):
        A.append(cells[0].find(text=True)[:-1])
        B.append(cells[1].find(text=True)[:-1])
        C.append(cells[2].find(text=True)[:-1])
```

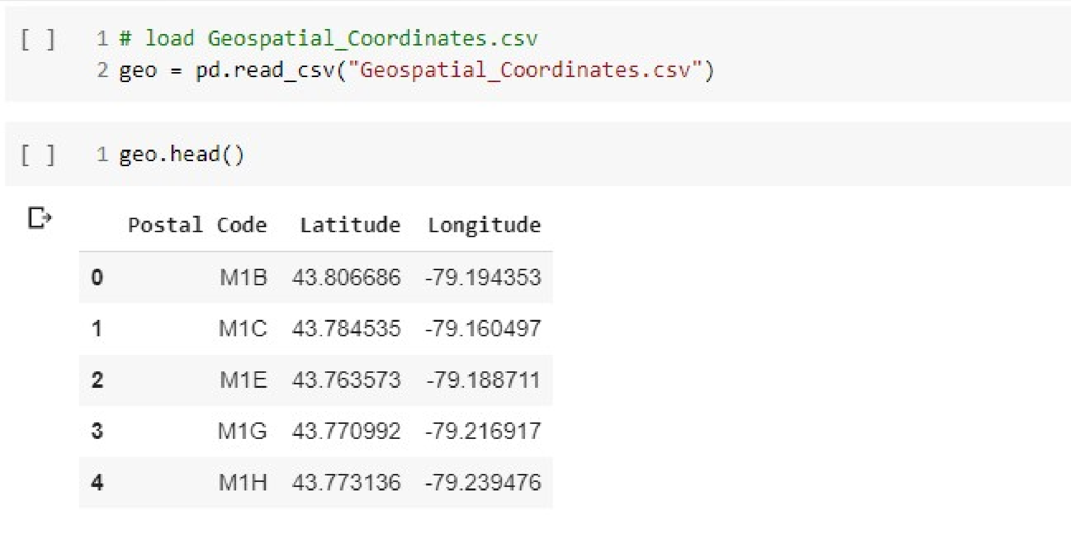


```
toronto
```

```
     PostalCode           Borough                                  Neighborhood
  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
...         ...               ...                                           ...
 98         M8X         Etobicoke  The Kingsway, Montgomery Road, Old Mill North
 99         M4Y  Downtown Toronto                        Church and Wellesley
100         M7Y      East Toronto Business reply mail Processing Centre, South C...
101         M8Y         Etobicoke     Old Mill South, King's Mill Park, Sunnylea, Hu...
102         M8Z         Etobicoke  Mimico NW, The Queensway West, South of Bloor,...

103 rows × 3 columns
```

3.2. Geospatial data

- The geospatial data of Toronto is provided by Coursera in form of csv file which contained latitude and longitude data in Toronto.



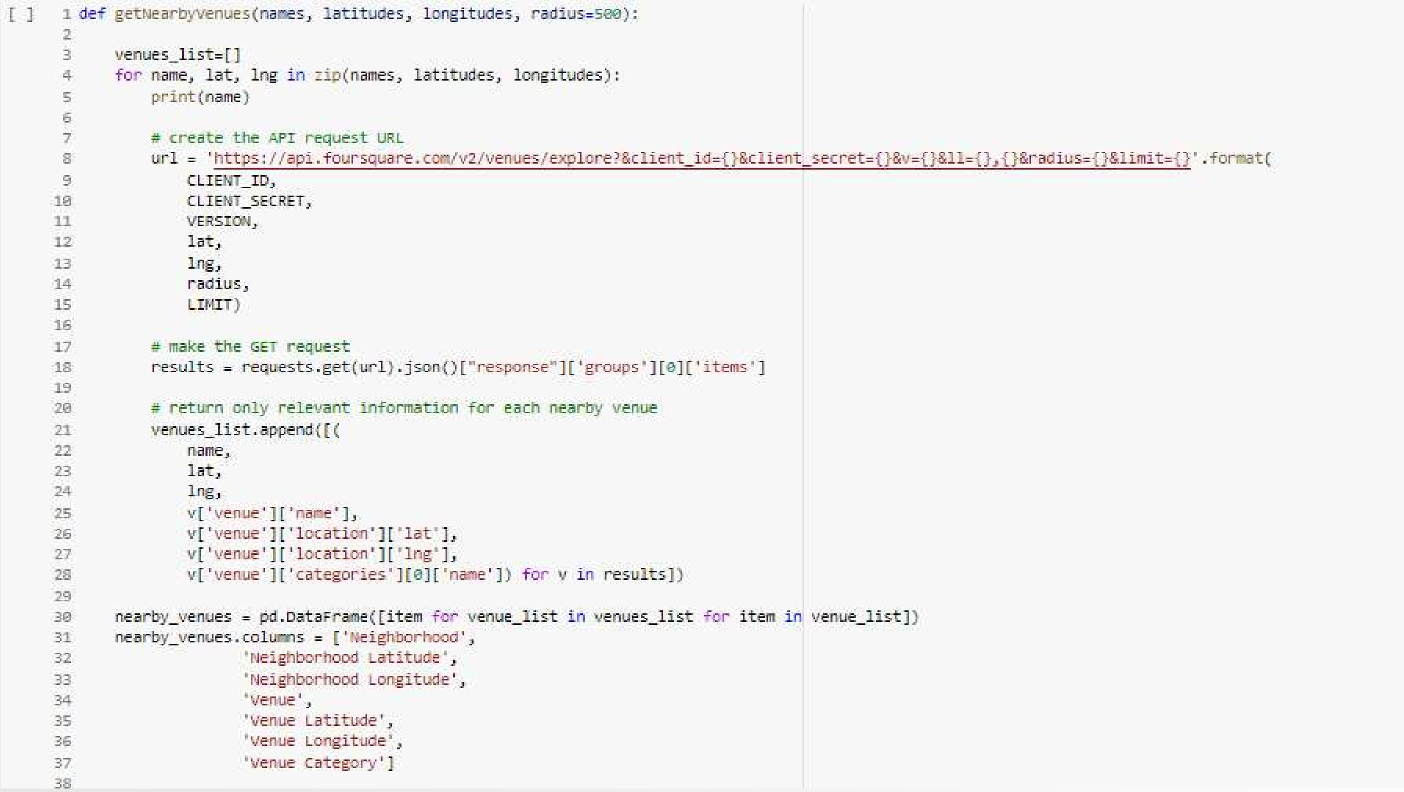
```
# load Geospatial_Coordinates.csv
geo = pd.read_csv("Geospatial_Coordinates.csv")
```

```
geo.head()
```

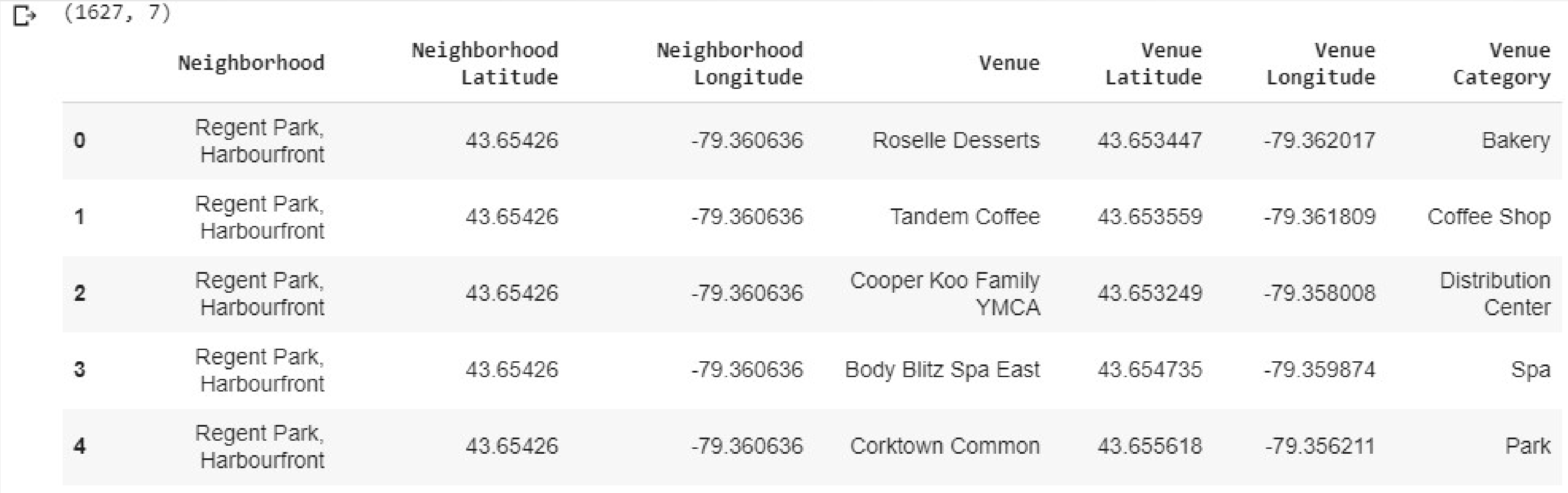
```
  Postal Code   Latitude   Longitude
0         M1B  43.806686  -79.194353
1         M1C  43.784535  -79.160497
2         M1E  43.763573  -79.188711
3         M1G  43.770992  -79.216917
4         M1H  43.773136  -79.239476
```

3.3. Venue category data

- The venue category data is collected by using FourSquare API and new data frame is created with the respective neighborhood.



```
[ ] def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

```



```
(1627, 7)
```

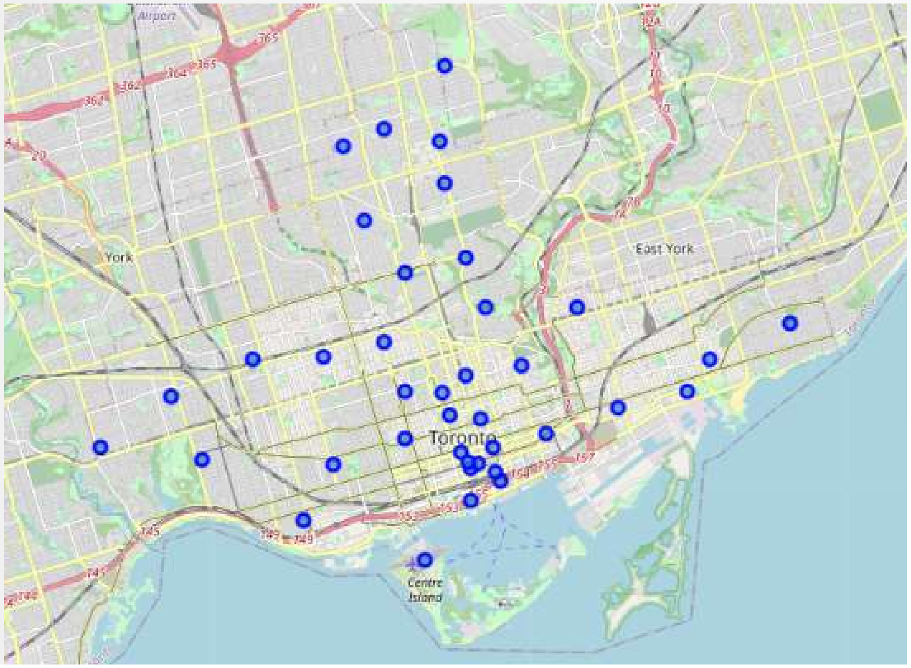
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Regent Park, Harbourfront	43.65426	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Regent Park, Harbourfront	43.65426	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
2	Regent Park, Harbourfront	43.65426	-79.360636	Cooper Koo Family YMCA	43.653249	-79.358008	Distribution Center
3	Regent Park, Harbourfront	43.65426	-79.360636	Body Blitz Spa East	43.654735	-79.359874	Spa
4	Regent Park, Harbourfront	43.65426	-79.360636	Corktown Common	43.655618	-79.356211	Park

4. Methodology

4.1. Folium

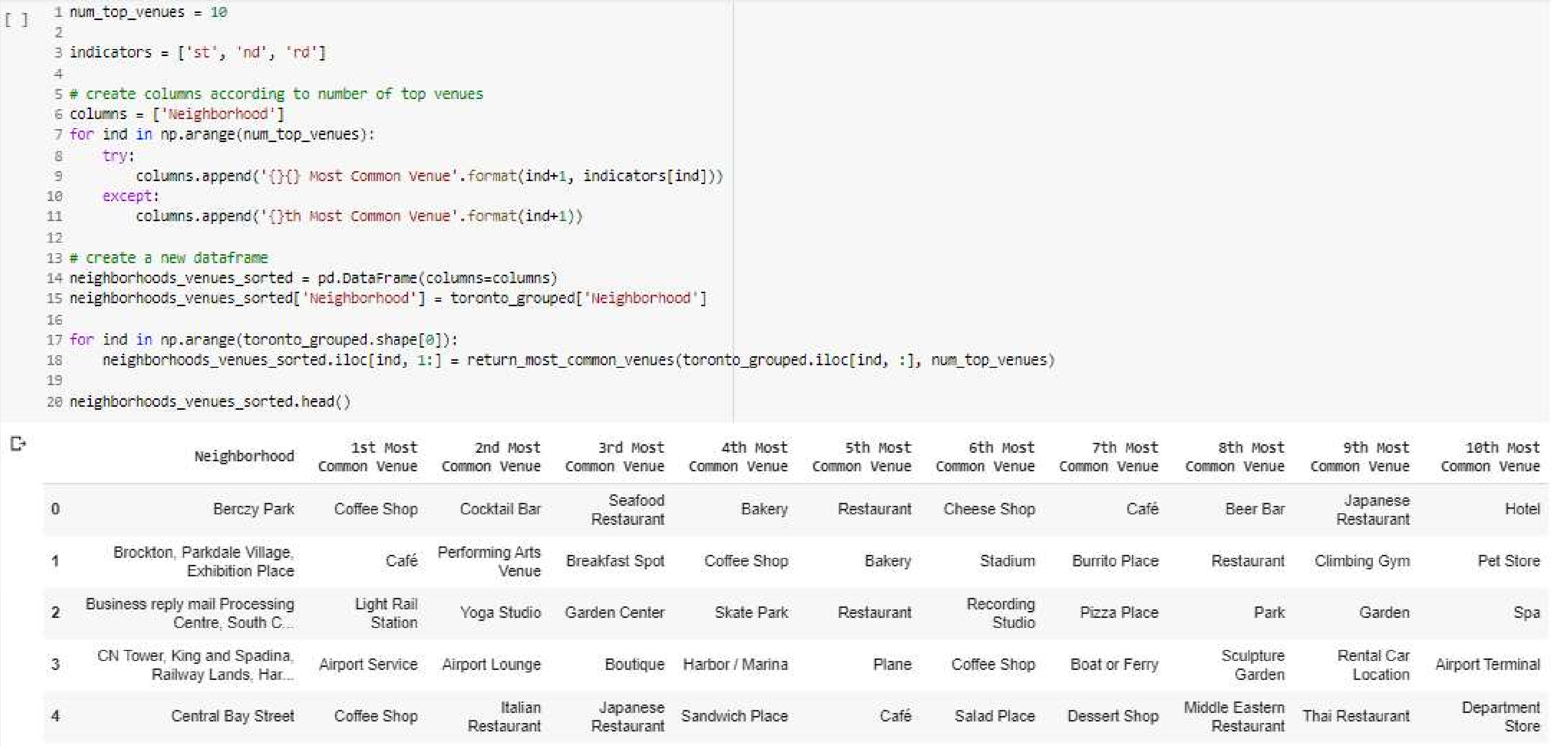
- folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. Manipulate your data in Python, then visualize it in on a Leaflet map via folium.

- The cluster visualizations are created by Folium to generates the leaflet map with marker on the map.



4.2. Top 10 most common venue

- Because of high variety in venues. We selected only top 10 common category in each neighborhood to train K-means Clustering Algorithm.

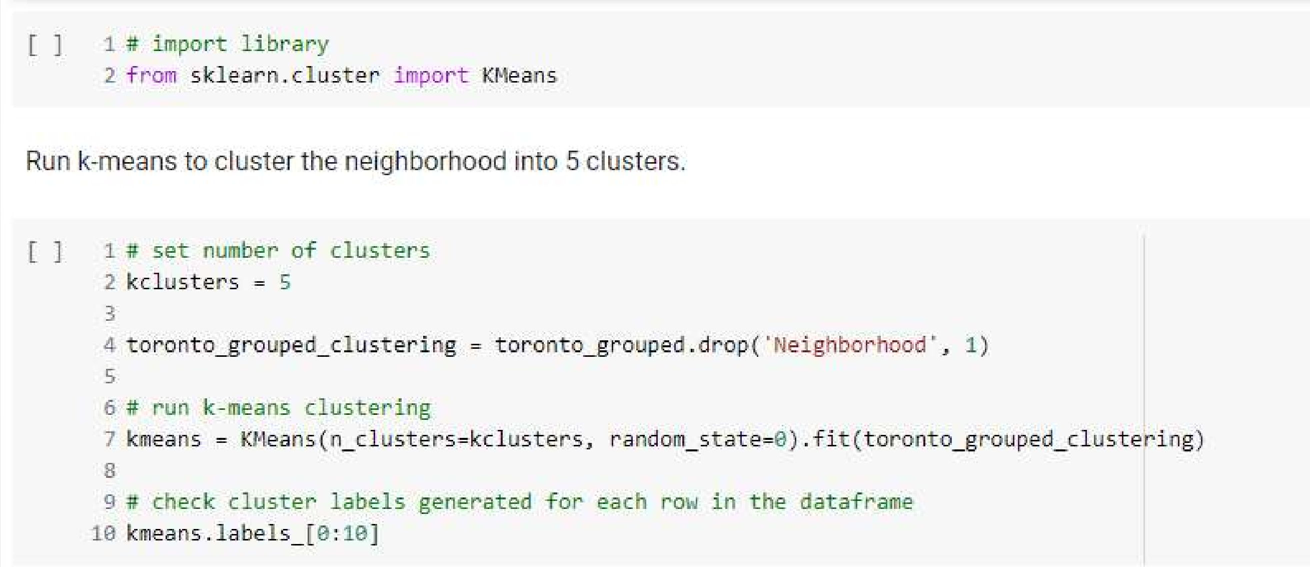


```
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = toronto_grouped['Neighborhood']

for ind in np.arange(toronto_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(toronto_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

	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	Berczy Park	Coffee Shop	Cocktail Bar	Seafood Restaurant	Bakery	Restaurant	Cheese Shop	Café	Beer Bar	Japanese Restaurant	Hotel
1	Brockton, Parkdale Village, Exhibition Place	Café	Performing Arts Venue	Breakfast Spot	Coffee Shop	Bakery	Stadium	Burrito Place	Restaurant	Climbing Gym	Pet Store
2	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Garden Center	Skate Park	Restaurant	Recording Studio	Pizza Place	Park	Garden	Spa
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Boutique	Harbor / Marina	Plane	Coffee Shop	Boat or Ferry	Sculpture Garden	Rental Car Location	Airport Terminal
4	Central Bay Street	Coffee Shop	Italian Restaurant	Japanese Restaurant	Sandwich Place	Café	Salad Place	Dessert Shop	Middle Eastern Restaurant	Thai Restaurant	Department Store

4.3. K-means clustering

- We used sklearn library to generates K-means Clustering with hyperparameter n_clusters = 5



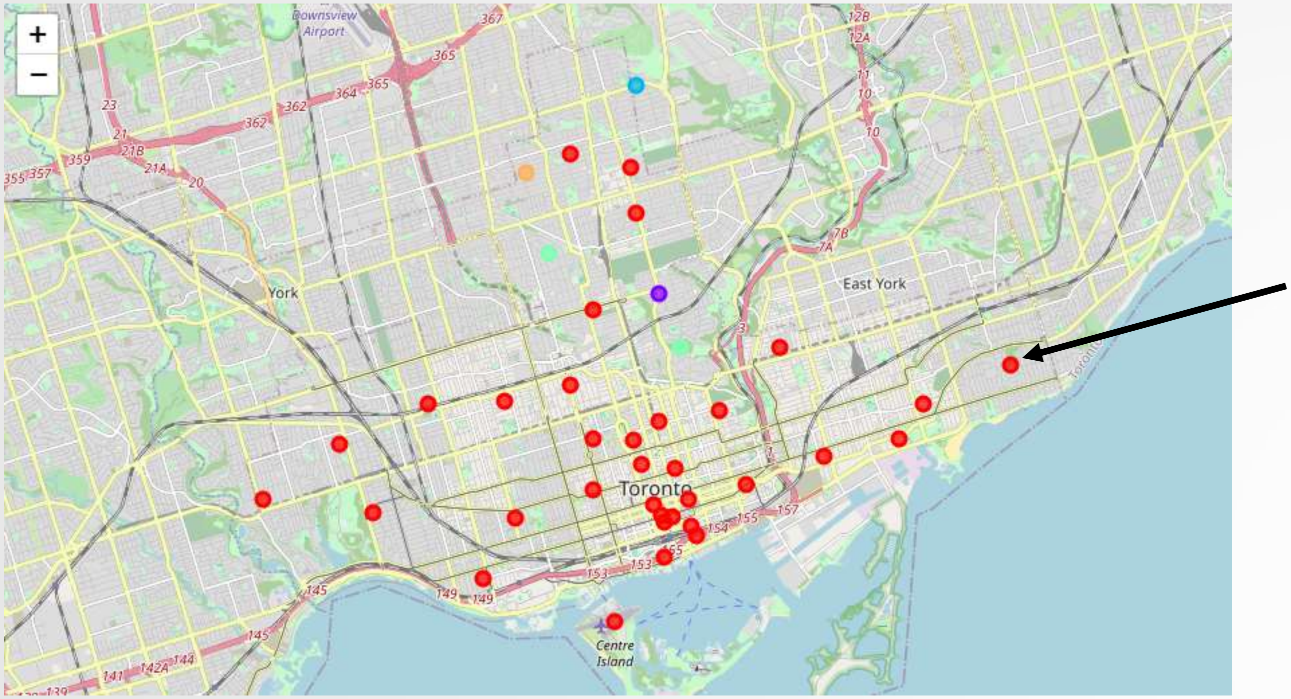
```
# import library
from sklearn.cluster import KMeans
```

Run k-means to cluster the neighborhood into 5 clusters.

```
# set number of clusters
kclusters = 5

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)

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

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

5. Result

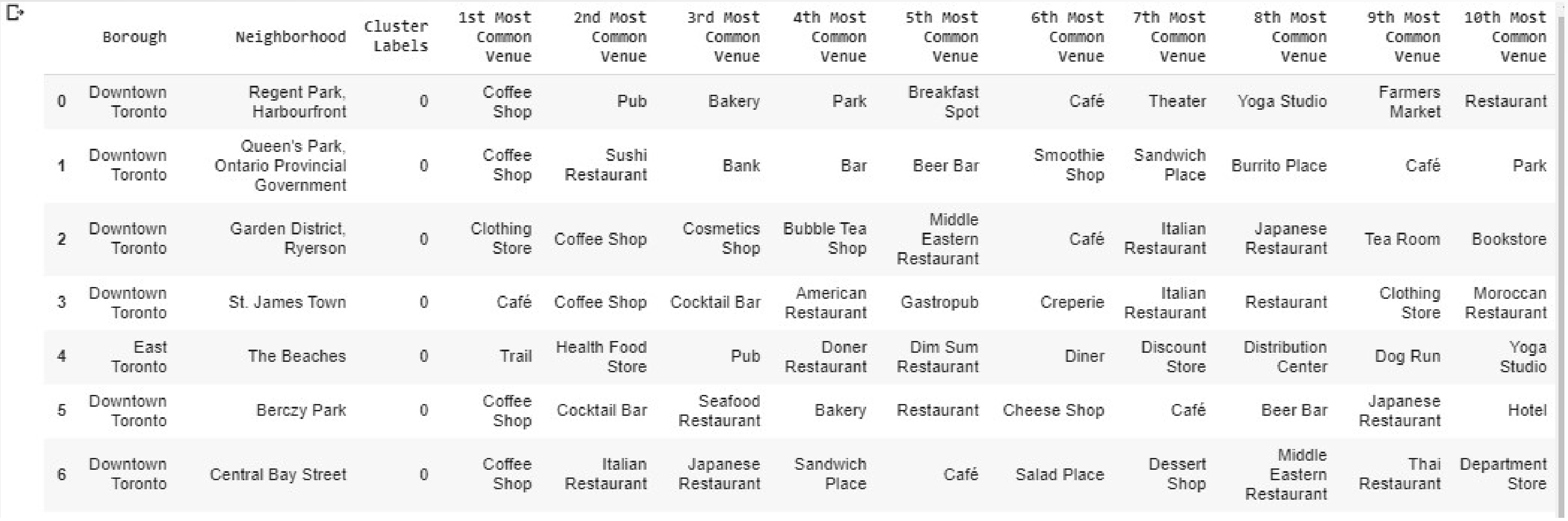
- The neighborhoods are divided in 5 clusters and visualized on the leaflet map. Each clusters are shown in different color on the map.



The most common

5. Result

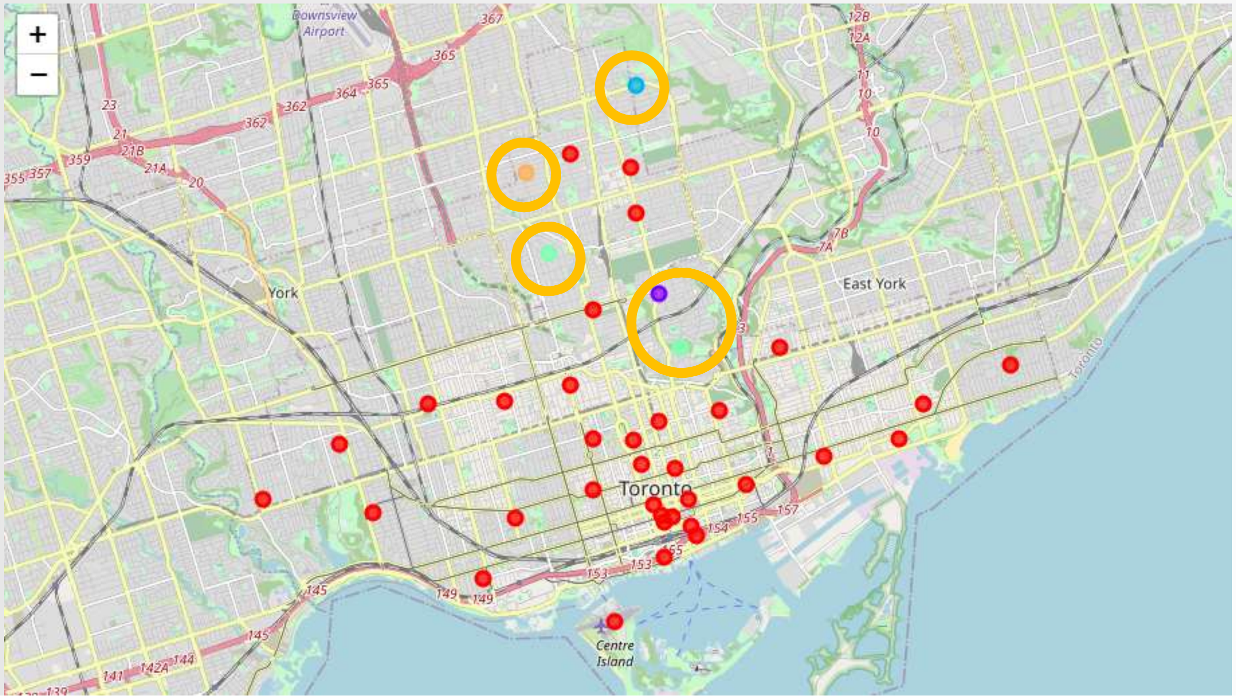
- The result from clustering shows that the most common venue category in city of Toronto is cluster 0 as shown in the map as red marker. The result in cluster 0 shows that the most common venues in the cluster are coffee shop and café. Inaddition, the next categories of common venue in the cluster 0 are many types of restaurant and food store as shown below.



	Borough	Neighborhood	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	Downtown Toronto	Regent Park, Harbourfront	0	Coffee Shop	Pub	Bakery	Park	Breakfast Spot	Café	Theater	Yoga Studio	Farmers Market	Restaurant
1	Downtown Toronto	Queen's Park, Ontario Provincial Government	0	Coffee Shop	Sushi Restaurant	Bank	Bar	Beer Bar	Smoothie Shop	Sandwich Place	Burrito Place	Café	Park
2	Downtown Toronto	Garden District, Ryerson	0	Clothing Store	Coffee Shop	Cosmetics Shop	Bubble Tea Shop	Middle Eastern Restaurant	Café	Italian Restaurant	Japanese Restaurant	Tea Room	Bookstore
3	Downtown Toronto	St. James Town	0	Café	Coffee Shop	Cocktail Bar	American Restaurant	Gastropub	Creperie	Italian Restaurant	Restaurant	Clothing Store	Moroccan Restaurant
4	East Toronto	The Beaches	0	Trail	Health Food Store	Pub	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribution Center	Dog Run	Yoga Studio
5	Downtown Toronto	Berczy Park	0	Coffee Shop	Cocktail Bar	Seafood Restaurant	Bakery	Restaurant	Cheese Shop	Café	Beer Bar	Japanese Restaurant	Hotel
6	Downtown Toronto	Central Bay Street	0	Coffee Shop	Italian Restaurant	Japanese Restaurant	Sandwich Place	Café	Salad Place	Dessert Shop	Middle Eastern Restaurant	Thai Restaurant	Department Store

5. Conclusion

- From the result, we can suggest that if someone is looking to open a restaurant in city of Toronto, they should open their restaurant in cluster 1,2,3 or 4 neighborhoods because in cluster 0 there are too many restaurants in neighborhood.


+
−
Downsview Airport
York
East York
Toronto
Centre Island