The Battle of the Neighborhoods.

Where to place a new yoga studio in Toronto?

1. Introduction.

1.1 Background.

Opening a Yoga studio is different from opening of ordinary fitness center. Indeed, fitness center hold many activities, most of which are dynamic and are accompanying by energetic music. That is why it can be situated anywhere, even on a busy street or in a commercial center. People that practice yoga are calm, concentrated and nature-oriented. Some yoga classes include meditation.
Taking all this into account, it is better to place a yoga studio in a quite district.

1.2 Business problem.

The client wants to open a yoga studio in Toronto. What is the best Toronto neighborhood to open it?
It is known that location is the most important characteristic for any business that imply direct contact with customers. That is why the neighborhoods should be examined carefully.
There are some requirements that should be fulfilled:

1) Not a lot of similar studios around. If yoga studios are common in the area, we can imagine that populations needs are saturated.

2) Relatively calm environment (minimum of restaurants and night life). Practicing yoga requires certain state of mind. We can presume that busy area with vibrant nightlife can be disturbing and repulse future customers from visiting the venue.

Therefore, the ideal neighborhood to open a new yoga studio should response to next criteria:

- Absence of yoga studio;
- Small number of restaurants.

2. Data description.

Our analysis should reveal the list of venue categories for each neighborhood and their frequency. Then, we will choose neighborhoods without yoga studio and cluster them according to restaurant frequency.

To perform this analysis, I will retrieve the list of Toronto neighborhoods from Wikipedia's List of postal codes of Canada.

I will transform it to the data frame, clean it and add a geospatial coordinates from http://cocl.us/Geospatial_data.

To retrieve type and number of venues (i.e. Yoga Studios, Restaurants) I will use Foursquare social network. Foursquare is a location-based social network that collect data

about venues. It also includes venues coordinates, so its data can be easily superposed on the map of Toronto.

2.1. Data collection.

- The most convenient way to segment the map of Toronto is to use city division on neighborhood based on its postal code. I will retrieve postal codes from Wikipedia's List of postal codes of Canada.
- To place neighborhood on a map I will download its geospatial coordinates available at http://cocl.us/Geospatial_data
- To retrieve type and number of venues (i.e. Yoga Studios, Restaurants) I will use Foursquare social network.

2.2 Data wrangling.

Using pandas.read_html() function I retrieve the data frame containing Toronto's neighborhoods Postal Codes from wiki. It looks like this:

Out[21]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Borough	Neighborhood	Postal code
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Not assigned	NaN	M1A
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Not assigned	NaN	M2A
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	North York	Parkwoods	M3A

I remove unnecessary columns as well as places where Borough name is not assigned:

Out[30]:

	Postal code	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill

Next, using pandas.read_csv() function I extract the list of Neighborhoods with its Postal Codes from http://cocl.us/Geospatial_data and get the following table:

Out[33]:

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711

In the next step, I concatenate two mentioned above data frames.
Now we have a data frame that contains Toronto's neighborhoods with their geospatial coordinates (first 5 rows):

	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

This data frame contains 103 rows. It means, there are 103 neighborhoods to analyze.

Now we can easily place neighborhoods on the map of Toronto.
First, I get geographical coordinates of Toronto and get a map of it using folium map builder:

```
address = 'Toronto, Canada'
geolocator = Nominatim(user_agent="toronto_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitude, longitude))
```

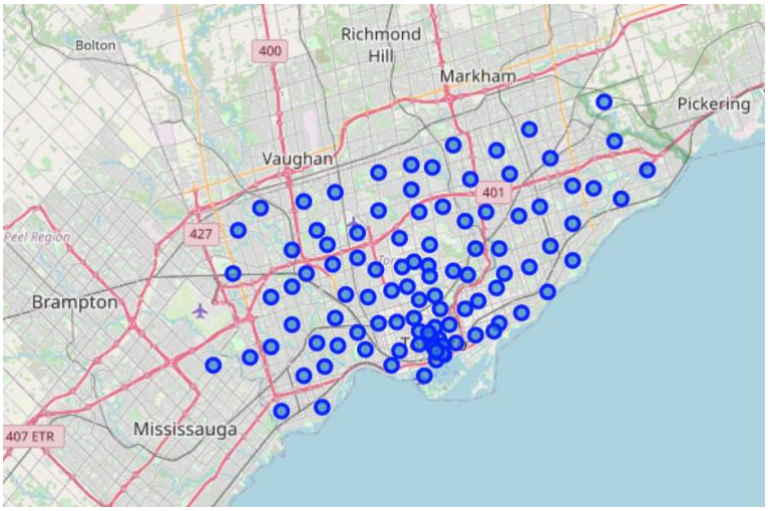
```
The geograpical coordinate of Toronto are 43.6534817, -79.3839347.
```

```
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=12)
map_toronto
```

Using folium python package we can draw a map centered around any location (latitude and longitude are needed).

Here is the map of Toronto we will work with:

I add neighborhoods to the map according to their coordinates (latitude and longitude):



2.3 Aquiring data from Foursquare API

Foursquare API collect the information about all venues in a certain location. For our purpose, we need to extract venues location (Neighborhood, latitude and longitude) and venue category. We will apply a special function that extract necessary information from Foursquare API. This function takes neighborhood's name and coordinates as arguments and returns venues and venue categories as well as their geolocation in a form of a data frame.

Here is a data frame extracted from Foursquare API. It contains all necessary information about venues of Toronto neighborhoods (first 5 rows of total 2128):

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant

3. Methodology section (Exploratory Data Analysis).

Now, when we have all data that we need, we can start filter it to get the information that we want.

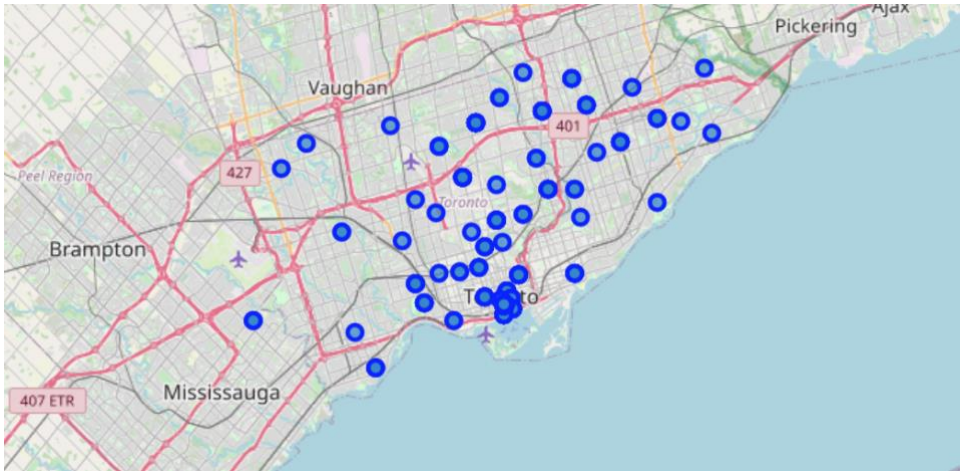
Firstly, lets count Yoga Studios in the data frame:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	Big Bite Burrito	43.766299	-79.190720	Mexican Restaurant

```
yoga_count = toronto_venues['Venue Category'].str.count('Yoga Studio').sum()
print('There are', yoga_count, 'Yoga Studios in Toronto!')
```

```
There are 14 Yoga Studios in Toronto!
```

We can see that there are only 14 Yoga Studios in Toronto. Now let's find neighborhoods with yoga studios and remove them from database. After filtering, we get a new data frame with 82 neighborhoods without yoga studios:



We want to choose a calm area without active nightlife. So, we will sort neighborhoods according to the restaurant frequency. Then we will choose less crowded areas.

Here are the necessary steps to perform:

1) Select only neighborhoods with restaurants (we do not consider neighborhoods without restaurants at all as it can be an industrial area);

2) Get a data frame containing list of restaurants with 1 (yes) or 0 (no) value for each type of restaurant for each neighborhood like this (53 neighborhoods, 45 venues):

	Neighborhood	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	R
0	Malvern, Rouge	0	0	0	0	0	0	
4	Guildwood, Morningside, West Hill	0	0	0	0	0	0	
11	Woburn	0	0	0	0	0	0	
13	Cedarbrae	0	0	0	0	0	0	
14	Cedarbrae	0	0	0	0	0	1	

3) Group neighborhoods by frequency of each venue;

	Neighborhood	American Restaurant	Asian Restaurant	Belgian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Chinese Restaurant
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.5

4) Apply K-means clustering to cluster neighborhoods according to restaurant frequency.

K-means clustering is widely used unsupervised machine learning algorithms that forms clusters of data based on the similarity between data instances.

First, to start clustering, we should arbitrary define number of clusters.

Then, The K-means algorithm starts by randomly choosing a centroid value for each cluster. After that the algorithm iteratively performs three steps: (i) Find the Euclidean distance between each data instance and centroids of all the clusters; (ii) Assign the data instances to the cluster of the centroid with nearest distance; (iii) Calculate new centroid values based on the mean values of the coordinates of all the data instances from the corresponding cluster.

As a result, we get an array of cluster labels:

```
array([0, 0, 8, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 4, 0, 0, 5, 3, 0, 0, 0, 8,
       0, 0, 0, 1, 0, 3, 0, 6, 9, 0, 7, 7, 0, 3, 0, 0, 7, 0, 7, 0, 0, 3,
       0, 0, 0, 0, 5, 5, 0, 2], dtype=int32)
```

4. Results.

K-means clustering with 10 clusters reveals the following distribution of neighborhoods:

```
Cluster Labels
0     33
1      1
2      1
3      5
4      1
5      3
6      1
7      4
8      2
9      1
Name: Neighborhood, dtype: int64
```

The most abandon is Cluster # 0.

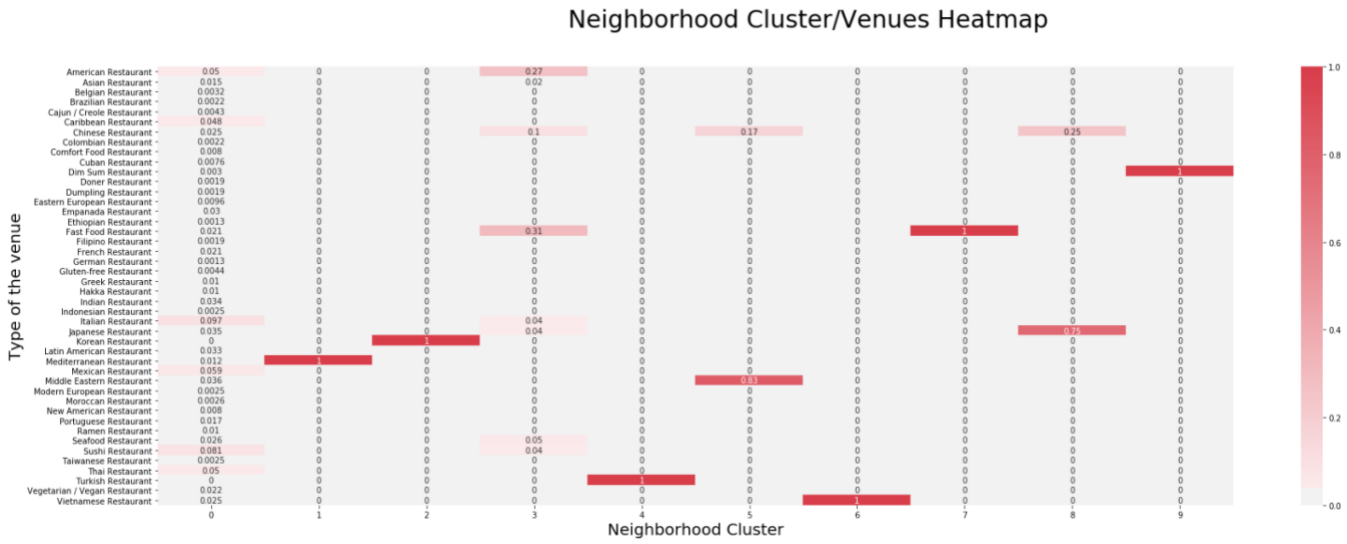
Let’s add cluster labels to each neighborhood and look more closely on our data frame again:

	Cluster Labels	Neighborhood	Longitude	Latitude
0	0	Agincourt	-79.262029	43.794200
1	0	Bathurst Manor, Wilson Heights, Downsview North	-79.442259	43.754328
2	8	Bayview Village	-79.385975	43.786947
3	0	Bedford Park, Lawrence Manor East	-79.419750	43.733283
4	0	Berczy Park	-79.373306	43.644771
5	0	Brockton, Parkdale Village, Exhibition Place	-79.428191	43.636847
6	0	Canada Post Gateway Processing Centre	-79.615819	43.636966
7	0	Cedarbrae	-79.239476	43.773136
8	0	Christie	-79.422564	43.669542
9	0	Clarks Corners, Tam O'Shanter, Sullivan	-79.304302	43.781638
10	3	Cliffside, Cliffcrest, Scarborough Village West	-79.239476	43.716316

To have a better idea what is difference between clusters, I will make a heatmap showing frequency distribution of different restaurants for each cluster. Firstly, I group datasets by frequency occurrence of each cluster. Them I transpose the database. As a result, I get a following database that can be easily transformed to the heatmap:

Cluster Labels	0	1	2	3	4	5	6	7	8	9
American Restaurant	0.049877	0.0	0.0	0.27	0.0	0.0	0.0	0.0	0.0	0.0
Asian Restaurant	0.014875	0.0	0.0	0.02	0.0	0.0	0.0	0.0	0.0	0.0
Belgian Restaurant	0.003211	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
Brazilian Restaurant	0.002222	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0
Cajun / Creole Restaurant	0.004329	0.0	0.0	0.00	0.0	0.0	0.0	0.0	0.0	0.0

Here is a heatmap originated from above data frame:



5. Discussion.

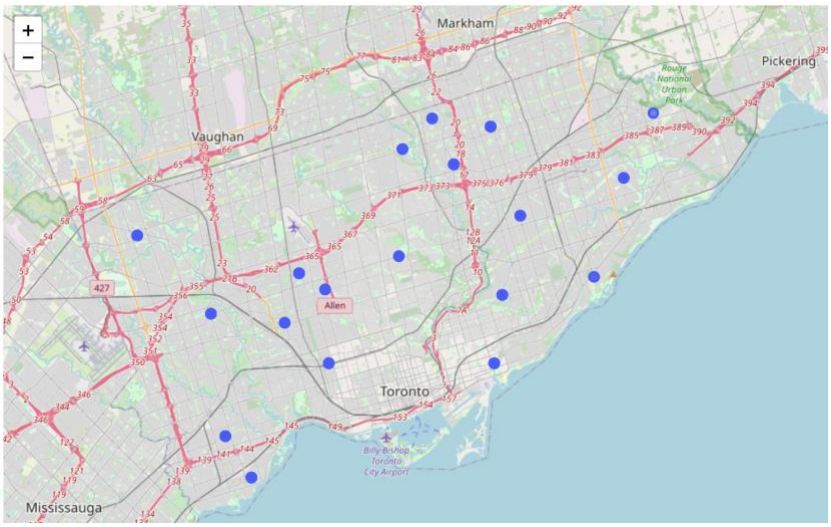
We can see that Cluster 0 is too much crowded with restaurants. That is why 33 neighborhoods that belong to cluster 0 are not the best place for a yoga studio:

0	Agincourt
1	Bathurst Manor, Wilson Heights, Downsview North
2	Bedford Park, Lawrence Manor East
3	Berczy Park
4	Brockton, Parkdale Village, Exhibition Place
5	Canada Post Gateway Processing Centre
6	Cedarbrae
7	Christie
8	Clarks Corners, Tam O'Shanter, Sullivan
9	Commerce Court, Victoria Hotel
10	Davisville
11	Don Mills
12	Dorset Park, Wexford Heights, Scarborough Town...
13	First Canadian Place, Underground city
14	Forest Hill North & West
15	Garden District, Ryerson
16	Guildwood, Morningside, West Hill
17	Harbourfront East, Union Station, Toronto Islands
18	High Park, The Junction South
19	Humber Summit
20	Kensington Market, Chinatown, Grange Park
21	Leaside
22	Moore Park, Summerhill East
23	Northwood Park, York University
24	Parkdale, Roncesvalles
25	Richmond, Adelaide, King
26	St. James Town
27	St. James Town, Cabbagetown
28	Summerhill West, Rathnelly, South Hill, Forest...
29	The Annex, North Midtown, Yorkville
30	Toronto Dominion Centre, Design Exchange
31	Victoria Village
32	Willowdale

Therefore, other 19 neighborhoods can be considered as a calm place for yoga center. Here is the list:

0	Bayview Village
1	Cliffside, Cliffcrest, Scarborough Village West
2	Del Ray, Mount Dennis, Keelsdale and Silverthorn
3	Dufferin, Dovercourt Village
4	Fairview, Henry Farm, Oriole
5	Glencairn
6	Hillcrest Village
7	India Bazaar, The Beaches West
8	Lawrence Manor, Lawrence Heights
9	Lawrence Park
10	Malvern, Rouge
11	Mimico NW, The Queensway West, South of Bloor,...
12	New Toronto, Mimico South, Humber Bay Shores
13	Parkview Hill, Woodbine Gardens
14	South Steeles, Silverstone, Humbergate, Jamest...
15	Steeles West, L'Amoreaux West
16	Westmount
17	Wexford, Maryvale
18	Woburn

Let’s place them on the map for better visualization:



6. Conclusion.

In this project, we identified 19 Toronto’s neighborhoods that are most suitable for an opening of a new yoga studio according to two criteria: i) absence of other yoga centers and ii) low number of restaurants. 14 neighborhoods of Toronto already have yoga studios.

Further analysis can include crime situation in above mentioned neighborhoods, traffic analysis and so on.