Peer-graded Assignment: Capstone Project - The Battle of Neighbourhoods

1. Introduction/ Business Problem

1.1. Background

Yoga is becoming an increasingly popular activity with statistics showing that the number of Americans practicing Yoga grew by 50% between 2012 and 2016[1]. The statistics also show that 1 in 3 Americans have tried Yoga in the last 6 months and spend a total of $16 billion on classes and equipment each year. Yoga has a number of health benefits including improving flexibility, building muscle strength, perfecting posture and preventing cartilage and joint breakdown[2].

1.2. Problem

My client has come to me for advice on a suitable location to open a new Yoga studio in Toronto. They would like to locate the studio where there are currently no other Yoga studios and close to other amenities that may be used by the members. They are interested in analysis another city to see where Yoga studios are located there.
The city I will use for segmentation and clustering comparison will be New York City.

1.3. Interest

Yoga is enjoyed by all ages. The chosen location could be in the CBD area to attract office workers to keep them active during their working day. Or alternatively the location could be in the suburbs where they may attract retirees or stay at home parents during the day. My client is flexible on location and would like to locate the studio in the location where he can expect to attract the most clients.

2. Data acquisition and cleaning

2.1. Data Sources

The following sources will be used for the investigation:

1. Neighbourhoods of New York data: https://geo.nyu.edu/catalog/nyu_2451_34572
2. Neighbourhoods of Toronto data: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
3. Geospatial Coordinates of Toronto: http://cocl.us/Geospatial_data
4. Foursquare - a technology company with a massive dataset of location data. A Foursquare API developer account is required and queries can be run on chosen locations using your Foursquare API Client ID and Client Secret.

2.2. Data Cleaning and methodology

2.2.1. New York dataset (data source 1)

A dataset that contains the New York boroughs and neighbourhoods as well as the latitude and longitude coordinates is already in existence and can be used for this project. A wget command is

[1] https://www.thegoodbody.com/yoga-statistics/

[2] https://www.yogajournal.com/lifestyle/count-yoga-38-ways-yoga-keeps-fit

used to download the dataset into a json file and then transfom it into a *pandas* dataframe for analysis– an empty dataframe is set up and then the data looped through.

2.2.2. Toronto dataset (data sources 2 and 3)

There is no existing dataset showing both neighbourhoods and coordinates so this is created using 2 datasets. The Neighbourhoods of Toronto data set is imported as a *pandas* dataframe (or by using the BeautifulSoup package) and the data is cleaned and the duplicates removed. Where data is missing for a Post Code, these boroughs are removed and in the case of missing data for a neighbourhood, then the borough is used.

The Geospatial csv data file is then also imported as a *pandas* dataframe and the 2 tables merged together using Post Code which is included in both datasets.

We then have a complete table including borough and neighbourhood and the longitude and latitude coordinates.

2.2.3. Foursquare

Foursquare API will be used to locate existing Yoga studios in each location and to explore the area around the venue. The longitude and latitude coordinates will be used on both New York City and Toronto, the search query will be set to 'Yoga' and a limit and radius chosen. A URL is generated using this data along with Client ID and Client Secret. A Get request is sent as a JSON file and then transformed into a *pandas* dataframe.

Folium is then used to visualise the Yoga studios on a map of the Cities.

We can then select a Yoga studio from the results to explore the area and see what else tends to be located around a Yoga studio and see if there are any patterns. If the Yoga studios have been rated we can also obtain to help determine the most popular ones.

2.2.4. K-means Clustering

We can use K-means clustering to cluster the neighbourhoods based on similarities and identify the top 10 most common venues in a neighbourhood. We can also use this to determine which neighbourhoods, if any, have a Yoga studio within the top 10.

3. Methodology

3.1. Import Dependencies and Foursquare

3.1.1. Dependencies

The first step is to import all the dependencies we need for the analysis:

- requests – library to handle requests
- numpy – to handle data in a vectorized manner
- pandas – library for data analysis
- random – for random number generation
- json – to handle JSON files
- Nominatim – to convert an address into Latitude and Longitude values
- Image and HTML – from IPython for displaying images
- json_normalize – to transform JSON file into a pandas dataframe
- matplotlib – for map plotting
- sklearn – for k-means clustering
- folium – map rendering library

3.1.2. Foursquare Credentials and Version

The Foursquare API Credentials are set using Client ID and Client Secret. The version is set 20191001 for up to date data and the limit initially set to 100 (but amended to depending on the queries executed). Radius and coordinates are added into the url each time a query is run.

3.2. New York data

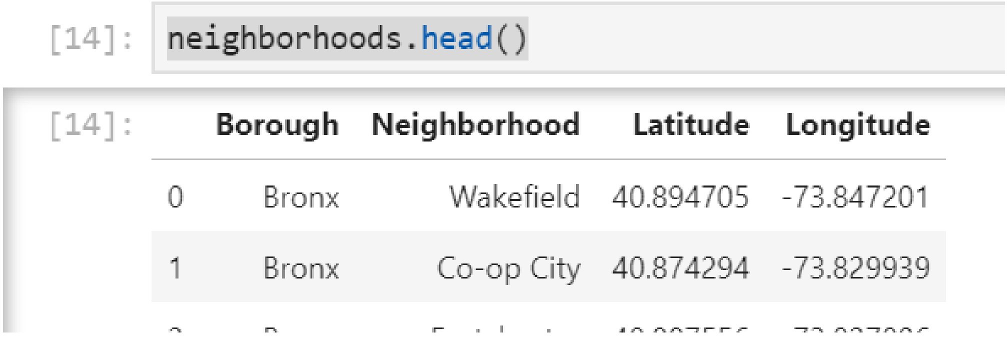
3.2.1. Download and Explore Dataset

The required data exists for free on the web (data source 1). As described in 2.1.1, a wget command is used to download the dataset into a json file.

We identify by running the data that 'features' key contains the most relevant data so a new variable is defined that includes this data.

We then transfom the data into a *pandas* dataframe for analysis, using the column names Borough, Neighbourhood, Latitude and Longitude.

The dataframe is named 'neighborhoods' and the import is checked by running neighborhoods.head() to showing the dataframe:



```
[14]: neighborhoods.head()
```

[14]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939

The number of boroughs and neighbourhoods is shown as 5 and 306 by running the neighbourhoods.shape [0] formula.

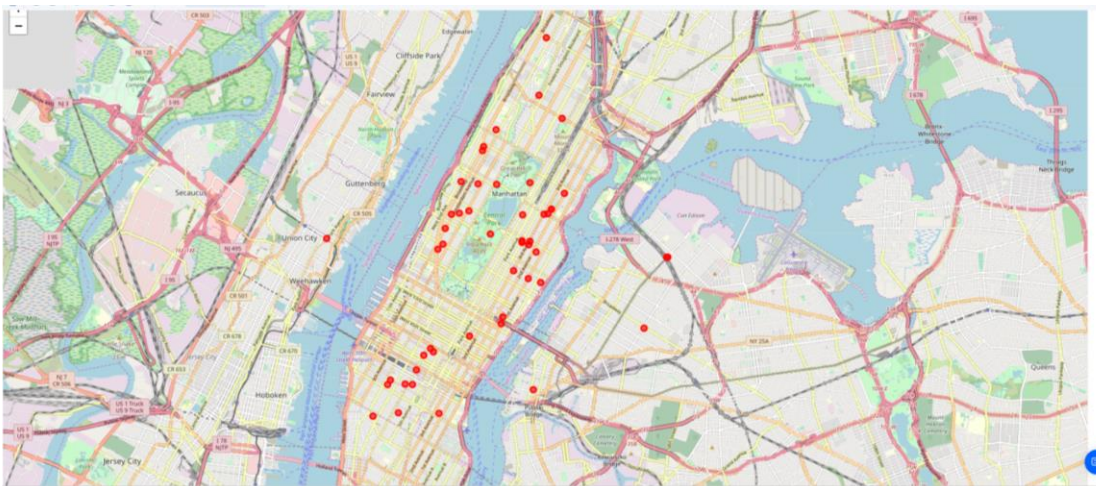
3.2.2. Use geopy library to get the coordinates of New York City and run Foursquare API Request

Nominatim is used to work out the coordinates of Manhattan to be 40.7896239 and -73.9598939.

The coordinates are used to run a Foursquare API request to identify where Yoga Studios are located in New York City.

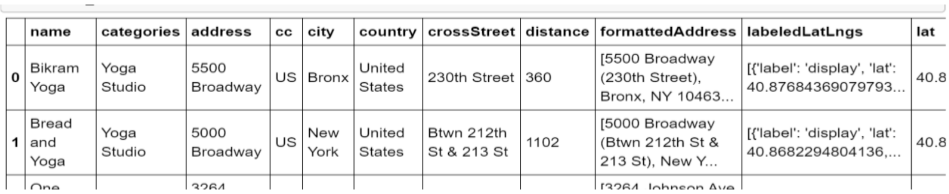
A JSON request is then run and the data transformed into a pandas dataframe.

The Yoga Studios are then marked on a New York City map by running Folium.

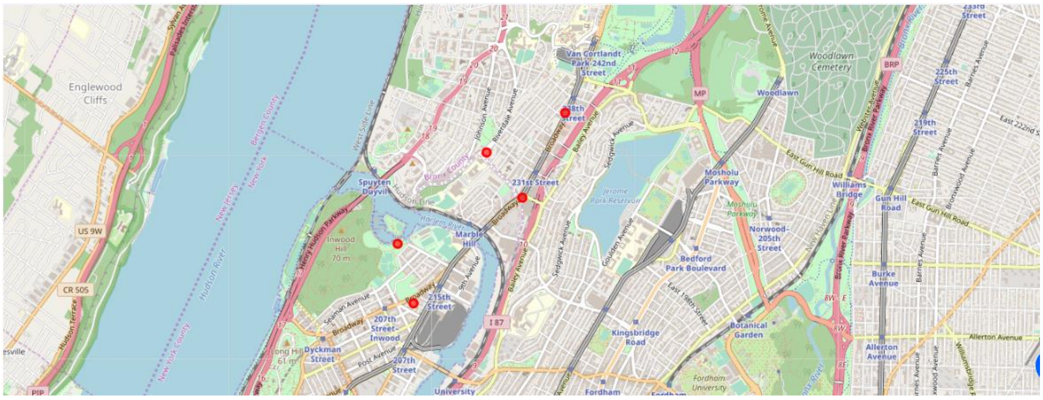


3.2.3. Narrow down search by running query for 1 individual Neighbourhood

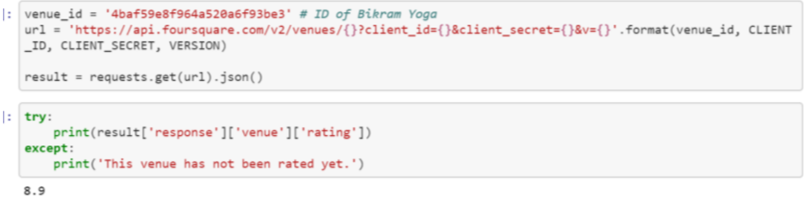
As there are many Yoga Studios in New York City, the request is reduced by running the query for 1 neighbourhood within New York City.
The geopathy library is used to find the coordinates for neighbourhood location 0 which is Marble Hill and then the Foursquare API run again followed by the Folium map.



	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat
0	Bikram Yoga	Yoga Studio	5500 Broadway	US	Bronx	United States	230th Street	360	[5500 Broadway (230th Street), Bronx, NY 10463...	[{'label': 'display', 'lat': 40.87684369079793...	40.8
1	Bread and Yoga	Yoga Studio	5000 Broadway	US	New York	United States	Btwn 212th St & 213 St	1102	[5000 Broadway (Btwn 212th St & 213 St), New Y...	[{'label': 'display', 'lat': 40.8682294804136,...	40.8
	One		3264						[3264 Johnson Ave		



A Foursquare API query is then run to discover the user rating of the first Yoga studio location identified in Marble Hill.



```
venue_id = '4baf59e8f964a520a6f93be3' # ID of Bikram Yoga
url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(venue_id, CLIENT
_ID, CLIENT_SECRET, VERSION)

result = requests.get(url).json()
```

```
try:
    print(result['response']['venue']['rating'])
except:
    print('This venue has not been rated yet.')
```

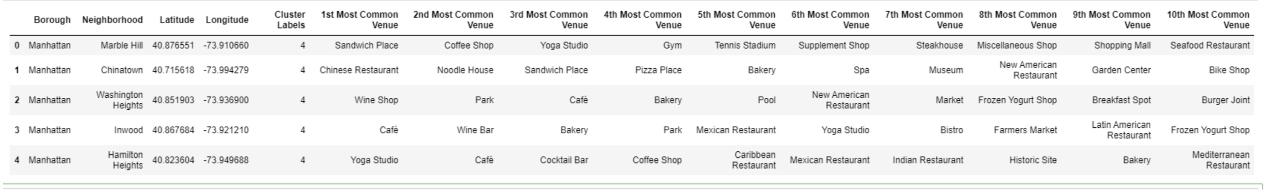
```
8.9
```

The Yoga Studio is popular with a rating of 8.9 so it was decided to continue using the Marble Hill area for the analysis.

3.2.4. Investigating nearby areas – nearby_venues and k-means

Further API requests are run using the coordinates of Marble Hill and nearby venues are identified using the get_category_type function and converting to pandas dataframe. This combined with manhattan_venues data can be used for k-means clustering.

One-hot coding is used and the data grouped by neighbourhood and k-means clustering run which then shows the top 10 most common venues for each neighbourhood.



	Borough	Neighborhood	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	9th Most Common Venue	10th Most Common Venue
0	Manhattan	Marble Hill	40.876551	-73.910660	4	Sandwich Place	Coffee Shop	Yoga Studio	Gym	Tennis Stadium	Supplement Shop	Steakhouse	Miscellaneous Shop	Shopping Mall	Seafood Restaurant
1	Manhattan	Chinatown	40.715618	-73.994279	4	Chinese Restaurant	Noodle House	Sandwich Place	Pizza Place	Bakery	Spa	Museum	New American Restaurant	Garden Center	Bike Shop
2	Manhattan	Washington Heights	40.851903	-73.936900	4	Wine Shop	Park	Café	Bakery	Pool	New American Restaurant	Market	Frozen Yogurt Shop	Breakfast Spot	Burger Joint
3	Manhattan	Inwood	40.867684	-73.921210	4	Café	Wine Bar	Bakery	Park	Mexican Restaurant	Yoga Studio	Bistro	Farmers Market	Latin American Restaurant	Frozen Yogurt Shop
4	Manhattan	Hamilton Heights	40.823604	-73.949688	4	Yoga Studio	Café	Cocktail Bar	Coffee Shop	Caribbean Restaurant	Mexican Restaurant	Indian Restaurant	Historic Site	Bakery	Mediterranean Restaurant

Marble Hill has Yoga Studio as number 3 and Tennis Court as number 5.

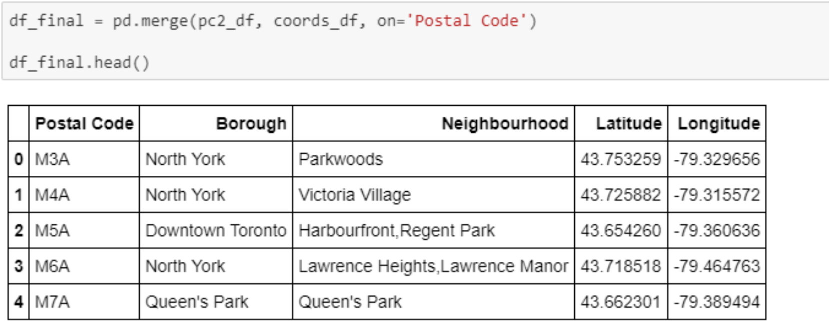
It is likely that a user of the Yoga studio may also be interested in other leisure activities so we will consider if there is a location within Toronto where there are tennis courts but no Yoga studio.

3.3. Toronto Data

3.3.1. Import and clean data

As noted in Section 2, there is no existing dataset showing both Toronto neighbourhoods and coordinates and so we need to combine 2 sources – see Section 2 for detail on the importing and data cleaning.

The final dataset is the 2 merged together:



```
df_final = pd.merge(pc2_df, coords_df, on='Postal Code')

df_final.head()
```

	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	Harbourfront,Regent Park	43.654260	-79.360636
3	M6A	North York	Lawrence Heights,Lawrence Manor	43.718518	-79.464763
4	M7A	Queen's Park	Queen's Park	43.662301	-79.389494

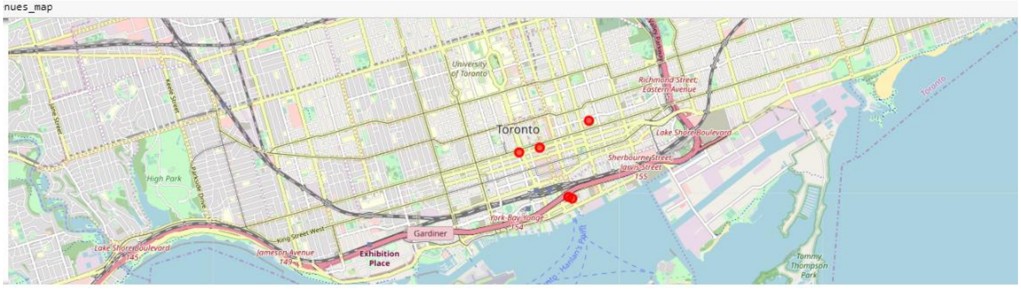
We will then use Foursquare API and Folium to run a map to identify the tennis courts in Toronto and a map to identify the Yoga studios.

If we find a location with tennis courts and no Yoga studios, then this is where we will recommend the open the new yoga studio.

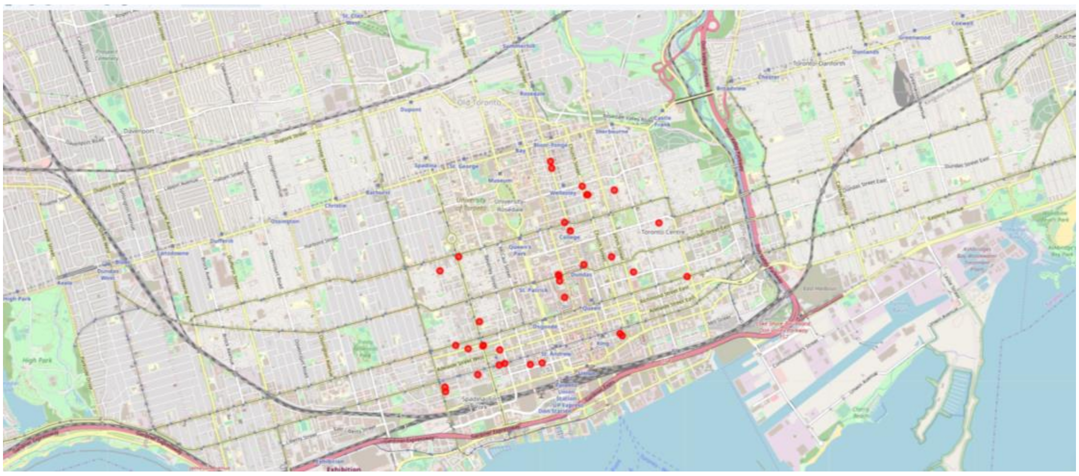
4. Results, discussion and recommendation

As per the analysis undertaken with the New York data, we identified that Marble Hill had a Yoga studio as the 3rd most common venue and also tennis courts as 4th most common venue. The Yoga studio we checked the rating for also had a high rating of 8.9.

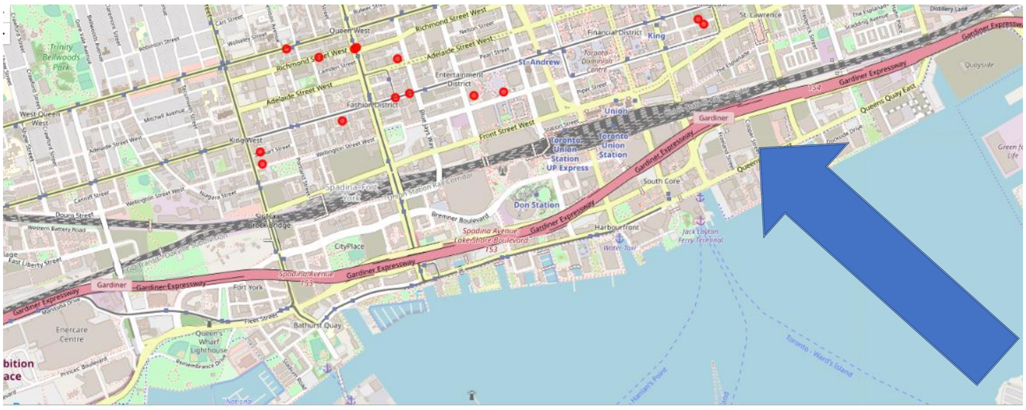
The resulting map using Foursquare API on Toronto and then Folium shows the tennis courts distributed around Toronto as follows:



The same run for Yoga studios in Toronto is:



We can identify that there are 2 tennis courts in the bottom centre where there are no Yoga studios:



This is where it will be recommended to open the Yoga studio. The location is also close to transport links by a main road and railway station.