IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

The Battle of Neighborhoods – A new gym in Toronto

Prepared by Maciej Skoczny

WEEK 4 – INTRODUCTION

Background:

Our client, Jane Doe, is looking for a suitable location in Toronto to open her new gym. She has experience in the fitness industry and runs a handful of fitness studios in Canada's other regions. She would like to expand her Canadian business to Toronto; however, she doesn't know the city well.

Business problem:

What Toronto neighborhood/borough should Jane open her new gym in? It is crucial to solving this problem by providing appropriate analysis and conclusions so that Jane can make profitable business decisions.

Target Audience:

Jane Doe is our target audience, as she's our client. The report could be useful for all the entrepreneurs who want to open a new fitness venue in Toronto.

Data:

- List of Toronto neighborhoods and boroughs
- Geographical coordinates of Toronto, including neighborhoods and boroughs
- Gym venue data

Proposed solution:

To solve the business problem and help Jane find a suitable sport for her new gym in Toronto, we'll use the data science methods to determine which neighborhood of Toronto has the smallest number of gyms. We'll combine the data extracted by Foursquare API with the geographical coordinates from Wikipedia and find an area with the smallest number of gyms. We'll use the Jupyter Notebook on the Anaconda platform and Python programming language. We'll use the following data science techniques:

- Data extraction
- Data scrapping
- Data segmentation
- Data clustering
- Data visualization

Data sources:

- List of Toronto neighborhoods and boroughs (including postal codes): https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Geographical coordinates of Toronto, including neighborhoods and boroughs (latitude, longitude): GeoCoder pack
- Gym venue data, including addresses and types - https://foursquare.com/

WEEK 5 – FINAL REPORT

Methodology:

As mentioned above, in order to help Jane find a suitable sport for her new gym in Toronto:

1. We imported all the necessary libraries

```
#Importing relevant libraries
import requests
import pandas as pd
import numpy as np
import folium
import lxml
import geocoder
from sklearn.cluster import KMeans
print("Libraries imported :)")
```

2. We extracted the list of Toronto neighborhoods and associated postal codes from Wikipedia.

```
#Extracting data from Wikipedia
url  = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
url_data = requests.get(url)
```

3. We scrapped the web data with pandas and transformed it into a data frame. Cleaning of data was needed to remove NaN and Not Assigned values. We also saved the results to the CSV file.

	index	Postal Code	Borough	Neighbourhood
0	2	M3A	North York	Parkwoods
1	3	M4A	North York	Victoria Village
2	4	M5A	Downtown Toronto	Regent Park, Harbourfront
3	5	M6A	North York	Lawrence Manor, Lawrence Heights
4	6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
...	...	...	...	...
98	160	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	165	M4Y	Downtown Toronto	Church and Wellesley
100	168	M7Y	East Toronto	Business reply mail Processing Centre, South C...
101	169	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...
102	178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...

103 rows × 4 columns

4. We extracted the geographical coordinates with Geocoder and created a new data frame by merging the coordinates and postal codes data. We also saved the results to the CSV file.

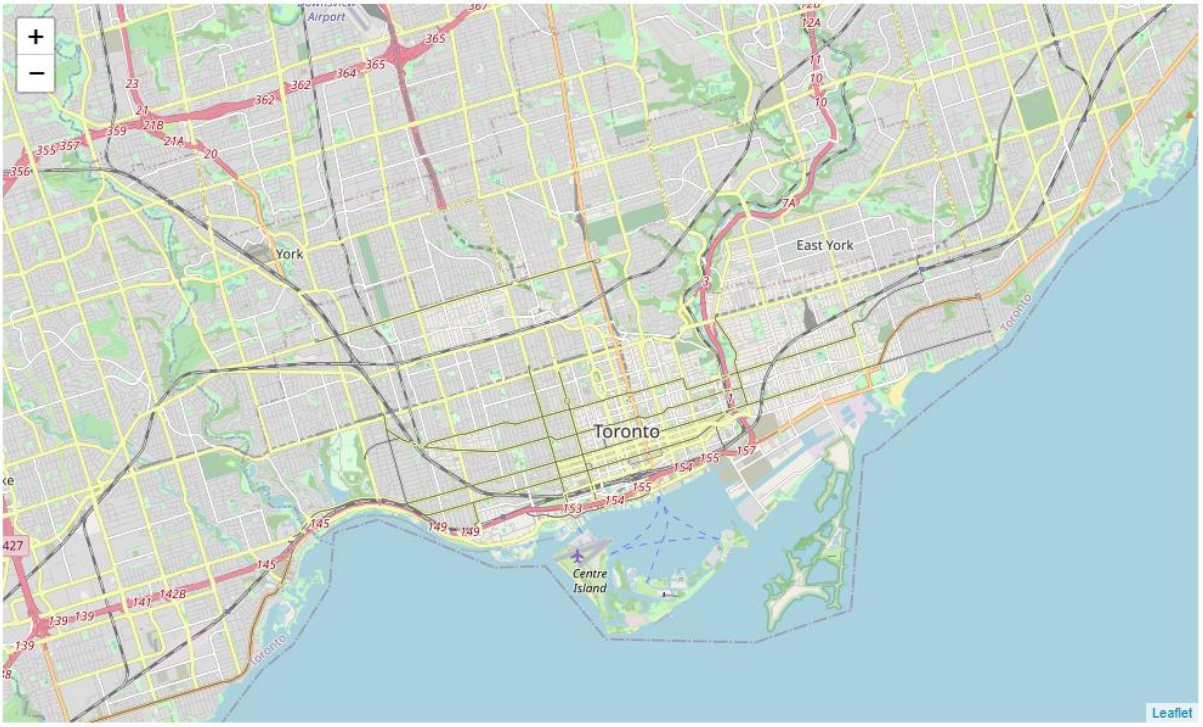
	index	Postalcode	Borough	Neighbourhood	Latitude	Longitude
0	2	M3A	North York	Parkwoods	43.75245	-79.32991
1	3	M4A	North York	Victoria Village	43.73057	-79.31306
2	4	M5A	Downtown Toronto	Regent Park, Harbourfront	43.65512	-79.36264
3	5	M6A	North York	Lawrence Manor, Lawrence Heights	43.72327	-79.45042
4	6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.66253	-79.39188
...	...	...	...	...	...	...
98	160	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North	43.65319	-79.51113
99	165	M4Y	Downtown Toronto	Church and Wellesley	43.66659	-79.38133
100	168	M7Y	East Toronto	Business reply mail Processing Centre, South C...	43.64869	-79.38544
101	169	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu...	43.63278	-79.48945
102	178	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,...	43.62513	-79.52681

103 rows × 6 columns

5. We then checked the neighborhoods' list and removed values that did not include "Toronto" in the borough's name.

```
Borough
Central Toronto      9
Downtown Toronto    19
East Toronto         5
West Toronto         6
```

6. Another step was to visualize the map of Toronto. We used the folium package and mean of Toronto's latitude and longitude, extracted from the previous data frame.



7. To extract the venue data, we used the Foursquare API (top 500 venues within a 500m radius). We created another data frame containing the neighborhood's information, latitude, longitude, venue, latitude, longitude, and category. We received 1712 records.

8. We then extracted a list of venue categories to pick the exact type for gym/fitness venues. We then converted categorical variables into numerical variables by using the one-hot method. It allows us to use data science/machine learning for clustering.

9. Another step was to sort and group the list by neighborhoods and means per each venue category, check the numbers of gym/fitness venues within the list (12 venues), and create a list containing neighborhoods and a mean of gym/fitness venues (37 records).

10. We clustered the neighborhoods into 3 clusters (0, 1, 2) and checked if there are only three types of clusters in the array. We then added cluster information to the table with venues.

	Neighborhood	Gym / Fitness Center	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Berczy Park	0.0	2	43.64536	-79.37306	The Keg Steakhouse + Bar - Esplanade	43.646712	-79.374768	Restaurant
0	Berczy Park	0.0	2	43.64536	-79.37306	Fresh On Front	43.647815	-79.374453	Vegetarian / Vegan Restaurant
0	Berczy Park	0.0	2	43.64536	-79.37306	LCBO	43.642944	-79.372440	Liquor Store
0	Berczy Park	0.0	2	43.64536	-79.37306	Meridian Hall	43.646292	-79.376022	Concert Hall
0	Berczy Park	0.0	2	43.64536	-79.37306	St. Lawrence Market (South Building)	43.648743	-79.371597	Farmers Market
...	...	...	...	...	...	...	...	...	...
38	University of Toronto, Harbord	0.0	2	43.66311	-79.40180	Second Cup (Miles Nadal JCC Fitness)	43.666527	-79.403872	Coffee Shop
38	University of Toronto, Harbord	0.0	2	43.66311	-79.40180	Shoppers Drug Mart	43.666562	-79.405007	Pharmacy
38	University of Toronto, Harbord	0.0	2	43.66311	-79.40180	Al Green Theatre	43.666547	-79.404053	Concert Hall
38	University of Toronto, Harbord	0.0	2	43.66311	-79.40180	Kenzo Japanese Noodle House	43.666295	-79.406002	Ramen Restaurant
38	University of Toronto, Harbord	0.0	2	43.66311	-79.40180	Swiss Chalet	43.666577	-79.404338	Restaurant

Results:

After filtering, we received the following list of gym/fitness venues for clusters 0, 1, and 2:

- Cluster 0

	Neighborhood	Gym / Fitness Center	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
3	CN Tower, King and Spadina, Railway Lands, Har...	0.025641	0	43.64082	-79.39818	FlirtyGirl Fitness	43.644005	-79.397590	Gym / Fitness Center
3	CN Tower, King and Spadina, Railway Lands, Har...	0.025641	0	43.64082	-79.39818	Cykl	43.642778	-79.402361	Gym / Fitness Center
4	Central Bay Street	0.016129	0	43.65609	-79.38493	Hard Candy Fitness	43.659556	-79.382440	Gym / Fitness Center
24	Regent Park, Harbourfront	0.041667	0	43.65512	-79.36264	The Extension Room	43.653313	-79.359725	Gym / Fitness Center

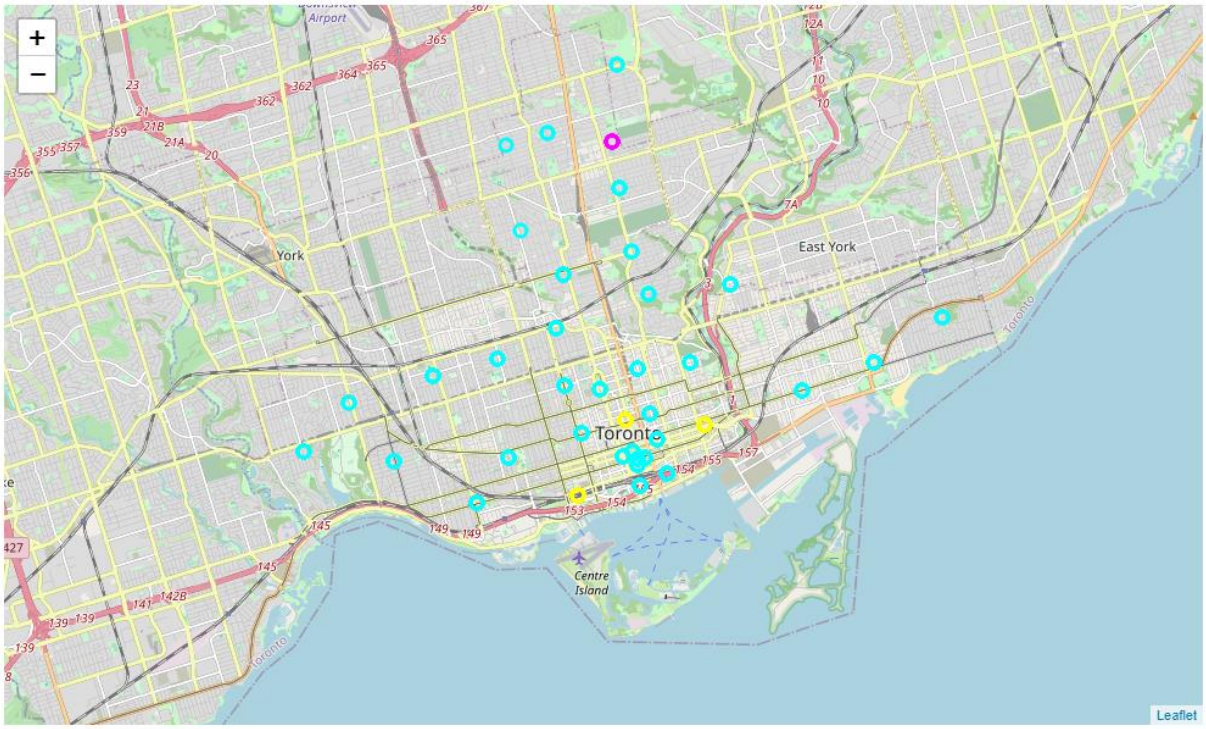
- Cluster 1

	Neighborhood	Gym / Fitness Center	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
9	Davisville North	0.125	1	43.71276	-79.38851	900 Mount Pleasant - Residents Gym	43.711671	-79.391767	Gym / Fitness Center

- Cluster 2

	Neighborhood	Gym / Fitness Center	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Business reply mail Processing Centre, South C...	0.010000	2	43.64869	-79.38544	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center
6	Church and Wellesley	0.012658	2	43.66659	-79.38133	Verve Gym	43.666702	-79.376539	Gym / Fitness Center
7	Commerce Court, Victoria Hotel	0.010000	2	43.64840	-79.37914	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center
11	First Canadian Place, Underground city	0.010000	2	43.64828	-79.38146	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center
13	Garden District, Ryerson	0.010000	2	43.65739	-79.37804	Hard Candy Fitness	43.659556	-79.382440	Gym / Fitness Center
25	Richmond, Adelaide, King	0.010000	2	43.64970	-79.38258	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center
31	Stn A PO Boxes	0.010000	2	43.64869	-79.38544	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center
37	Toronto Dominion Centre, Design Exchange	0.010000	2	43.64710	-79.38153	Adelaide Club Toronto	43.649279	-79.381921	Gym / Fitness Center

- Map of Toronto including clusters:



Discussion and conclusion:

Our client, Jane Doe, is looking for a suitable location in Toronto to open her new gym.
Following our analysis, three clusters have been identified:

Cluster 0 – 4 gym/fitness venues
Cluster 1 – 1 gym/fitness venue
Cluster 2 – 8 gym/fitness venues

Based on the criteria of gym/fitness venues clustering and competitors numbers, our recommendation for Jane Doe is to find a place for her new business venue within cluster 1, Dashville North, consisting of only one gym/fitness venue. Hence, Jane Doe would face a smaller number of competitors than in the other two clusters.