Coursera

Capstone Project:

IBM DATA SCIENCE

The Battle of Neighborhoods

Find a suitable location for *EAT-TO-GET-FIT* first outlet in Toronto

-Vishal Arora

Jan'2020

Introduction: Business Problem

The fitness industry in Canada is growing at a rapid pace of more than 5% year on year. Being a fitness enthusiast myself, I would like to work on a problem faced by majority of people involved in fitness. We all know that exercise plays a vital in becoming fit. However, the *DIET* plays an even more important role in the overall process.

For exercise, people join GYM, FITNESS CENTERS or simply go to park. But when it comes to the diet, we are unable to manage it on our own. Hence, I have come up with an idea of launching a chain of *EAT-TO-GET-FIT* outlets in Canada. It will take care of the complete meal for the person enrolling for the service. The unique-selling-point of these outlets will be that they will be strategically located very near to the GYMs/Fitness Centers etc. so that the enrolled member can eat at the outlet/take home the food after their work-out.

In today's world, people are crazy about fitness and to sustain with the habit of exercising daily, they join a gym/fitness center. Meanwhile, *EAT-TO-GET-FIT* outlets will take care of their diet. I believe that this idea has a great potential and to test it in Canada, I would like to start with one outlet in Toronto. Now, to find the most optimal location for this outlet, I am taking up this project.

In this project, we will put our data science mind to work and find out the most promising location in Toronto. We plan to have the outlet near the Gym or Fitness Centers because post the workout, people need food to re-energize themselves.

The **target audience** for this project will be the entrepreneur who wants to find the location to open an eat to get fit outlet(restaurant)

Data Objective

Data Requirements:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue Category data to understand the type of venues in each neighborhood

Data Fetching:

- Using the technique of web-scrapping to gather the list of neighborhoods in Toronto
- Installing the Geocoder package to fetch the Latitudes and Longitudes the neighborhoods
- Calling Foursquare API to get the details of various types of venues in these neighborhoods

Out of all the features considered, we will consider Venue Category equal to Gym to segment the data. This data will be used for data analysis (using techniques like clustering) to come up with the most suitable location for opening the first outlet of *EAT-TO-GET-FIT*.

Methodology

In this project, we will direct our efforts on detecting areas of Toronto that have high Gym/Fitness Center density. Listing down the steps:

- Install xmld, folium and geopy libraries on the Jupyter platform. After which we will import them and start working on them

- Use the following URL to scrap the borough and neighborhood data in Toronto, Canada

    "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"

Merging both datasets

Borough	Count (Neighborhood)
Central Toronto	9
Downtown Toronto	19
East Toronto	5
East York	5
Etobicoke	11
Mississauga	1
North York	24
Queen's Park	1
Scarborough	17
West Toronto	6
York	5

Figure 1

- Work on data wrangling to clean it and make it suitable for analysis

- Fetch the co-ordinates of these neighborhoods using the postal code

- Merge the newly fetched coordinates with the neighborhood data and name it df_toronto

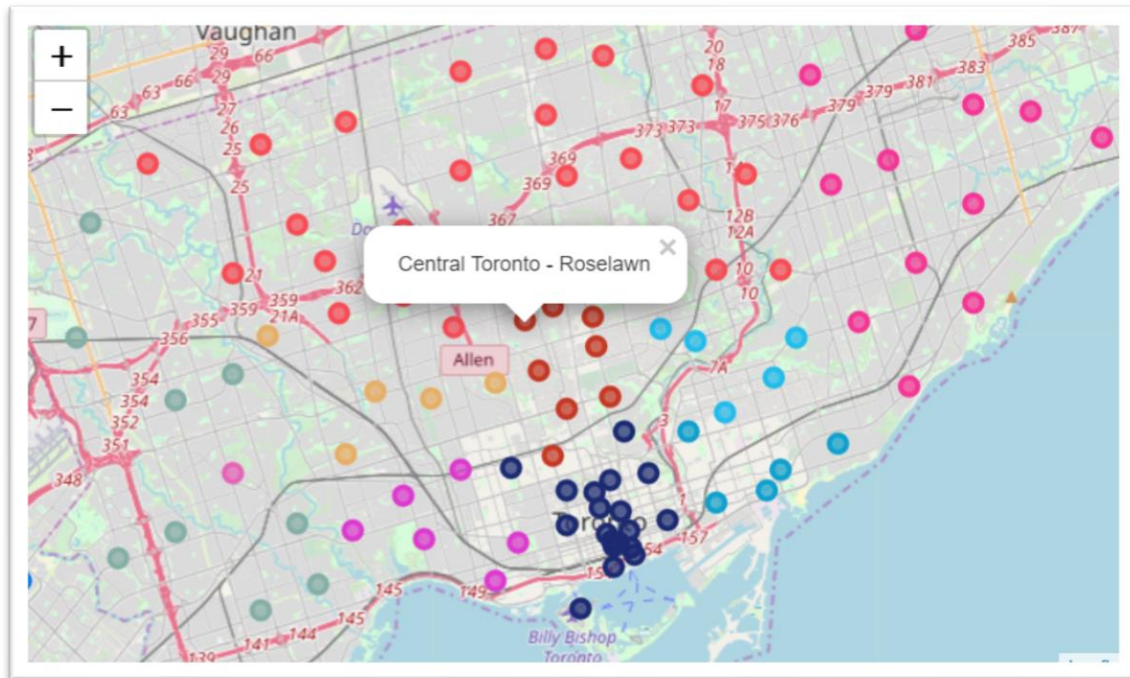- Visualize this data on the Toronto Map to understand the location of different neighborhoods

Figure 2

- Call the Foursquare API to get data on venues in every neighborhood. This API would return Venue Name, Venue Longitude, Venue Latitude and Venue Category (we explicitly ask for these details from the API in the get request)

- After receiving the data, we will clean it and make it suitable by:

  - Changing the venue category of all the different types of Gym to Gym in order to achieve consistency in the data as we would like to have our outlet near a Gym / Fitness Centers / Other Fitness related area

- Before analyzing the data, use the one-hot encoding technique to convert the categorical variable named "Venue Category" to a numerical variable.

- Calculate the mean occurrence of "Gym" in every neighborhood and save it in a data frame

- Perform Cluster analysis on this data frame and apply the k_means values to the df_toronto variable saved earlier

- Visualize the data frame df_toronto to understand the clusters formed and identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location

Results

We have used K-means clustering to form 3 kinds of clusters:
Cluster 2: Neighborhoods with a lot of Gyms in the area,
Cluster 1: Neighborhoods with no Gym in the area,
Cluster 0: Neighborhood with less or no Gyms in the area

Based on the clusters formed, there are 2 Neighborhoods with high number of Gyms/Fitness Centers. These are **Don Mills North** and **Downsview Northwest**. These locations belong to the **Cluster 2** and have **Turquoise** colored markers in the MapView. (Please find the cluster 0,1 and 2 in the Appendix)

Discussion

Based on the results drawn from the analysis, I would like to recommend *Don Mills North* and *Downsview Northwest* for opening the first outlet of *EAT-TO-GET-FIT*. Now, the next step will be to incorporate factors like size of the available space, distance from gym/fitness center, price of the available space, demographic factors of the neighborhood, etc. to recommend the final specific location is outside the scope of this project due to the timeline attached. However, the key take-away would be that these 2 areas are great to start with the on-ground street exploration for the ultimate outlet location.

Conclusion

In this study, I have analyzed the neighborhoods and venue categories in Toronto, Canada. As I wanted to open the first outlet near the fitness centers/gyms, I built clustering model to identify the most potential areas. Based on the data, Don Mills North and Downsview are the 2 most suitable areas for opening the outlet. This model will be very useful when we will expand the chain of EAT-TO-GET-FIT to other parts of Toronto and later, Canada.

References

- https://foursquare.com/developers/apps

- http://cocl.us/Geospatial_data

- https://www.ibisworld.com/canada/market-research-reports/gym-health-fitness-clubs-industry/

Appendix A: Visualization of the clusters on the map of Toronto:

Turquoise markers are the areas where we have high occurrence of Gyms/Fitness Centers
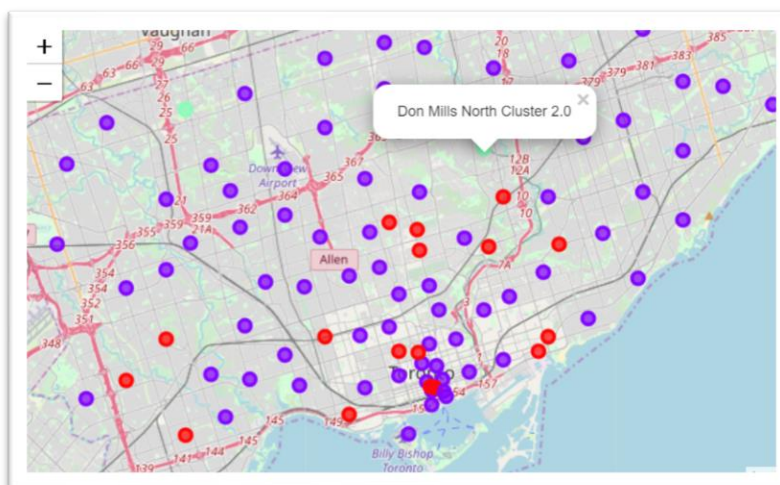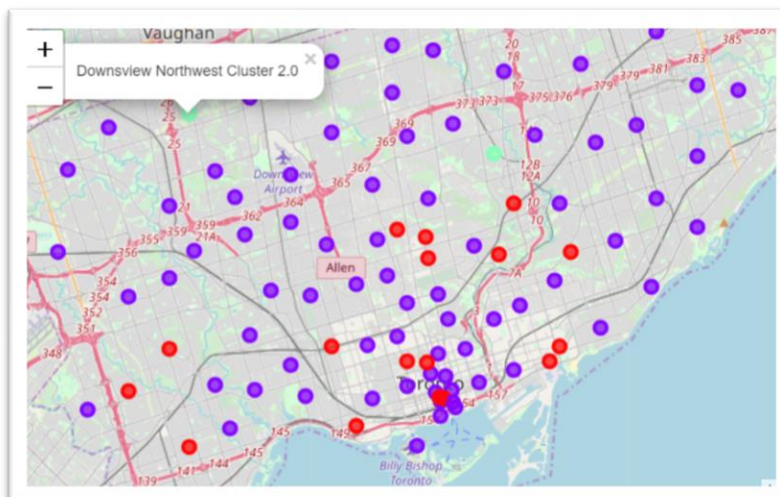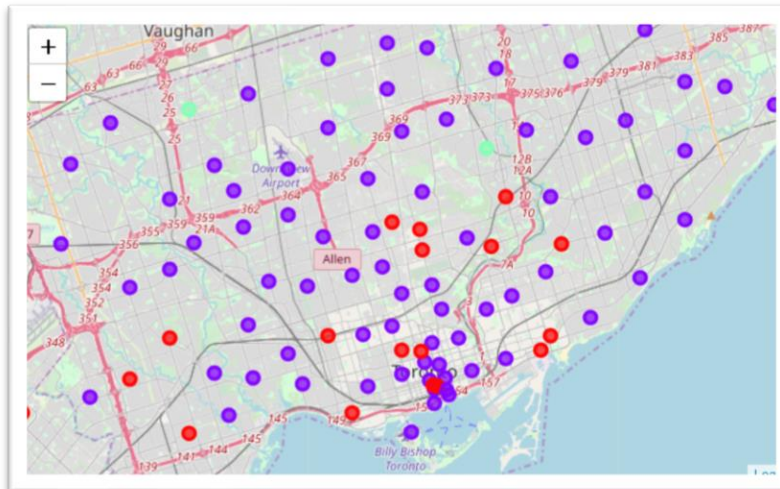
Appendix B: The table of 3 clusters formed after performing K-Means Clustering

Cluster 0

	Neighborhood	Gym	Cluster Labels
4	Alderwood, Long Branch	0.111111	0
11	Brockton, Exhibition Place, Parkdale Village	0.090909	0
12	Business Reply Mail Processing Centre 969 Eastern	0.066667	0
17	Canada Post Gateway Processing Centre	0.090909	0
26	Commerce Court, Victoria Hotel	0.04	0
27	Davisville	0.055556	0
28	Davisville North	0.125	0
34	Dovercourt Village, Dufferin	0.066667	0
43	First Canadian Place, Underground city	0.04	0
44	Flemingdon Park, Don Mills South	0.095238	0
48	Harbord, University of Toronto	0.054054	0
54	Humber Bay Shores, Mimico South, New Toronto	0.083333	0
59	Kingsway Park South West, Mimico NW, The Queen...	0.076923	0
68	North Toronto West	0.05	0
73	Queen's Park	0.052632	0
86	The Beaches West, India Bazaar	0.05	0
90	Thorncliffe Park	0.052632	0
97	Woodbine Gardens, Parkview Hill	0.076923	0

Cluster 1

	Neighborhood	Gym	Cluster Labels
0	Adelaide, King, Richmond	0.03	1
1	Agincourt	0	1
2	Agincourt North, L'Amoreaux East, Milliken, St...	0	1
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0	1
5	Bathurst Manor, Downsview North, Wilson Heights	0	1
...	...	...	...
94	Willowdale South	0	1
95	Willowdale West	0	1

	Neighborhood		
96	Woburn	0	1
98	Woodbine Heights	0	1
99	York Mills West	0	1

Cluster 2

	Neighborhood	Gym	Cluster Labels
32	Don Mills North	0.2	2
36	Downsview Northwest	0.2	2