Wellness in Toronto: IBM Applied Data Science Capstone

By Joshua Kehler

Introduction:

On March 11th, 2020, the World Health Organization determined that COVID-19 had reached the level of pandemic at 118,000 cases and 4,291 deaths across 114 countries (1). As of August 12th, 2020 the number of Global cases surges to 20,439,000 with 744,941 deaths to date (2). With a reproduction number estimated at 3.28, the number of infected will continue to increase (3) with proportional loss of life. To anticipate need and safeguard those most vulnerable, risk factors and the elucidation of exacerbating comorbidities are of great interest. Research has shown that those with heart disease may be up to 7X more likely to die from COVID-19 (4). As doctors and scientist work towards treatment and vaccines, the general public can strive towards solvency as well.

Studies indicate that males who run for more than an hour per week at moderate intensity may realize a 42% reduction in risk of heart disease than those who do not (5). Resistance training has shown to decrease glycosylated hemoglobin (HbA1c) levels, a condition associated with diabetes and cardiovascular disease (6). Diets such as the Mediterranean diet, high in mixed nuts, fish, vegetables and low in red meats, sugars may reduce risk of heart disease by 31% (7). Taken together, cardiovascular training, strength training, and nutrition guidance coalesce into a treatment plan readily deployed and exercised by the general public. In observance of social distancing and CDC best practice, the local gym may be the best place to program the initiative.

Business Problem:

To identify optimal locations for future gym sites based upon current density of related venues in Toronto, Canada. Leverage an array of data science techniques; such as web scraping,

Foursquare API, machine learning algorithms and visualization methods to develop a tool useful to real estate developers, sole proprietors, and fitness enthusiasts looking to add gyms or wellness centers to their community.

Target Audience:

The tools and data herein will prove useful for real estate developers, sole proprietors, and fitness enthusiasts looking to add gyms and wellness centers to their community. Moreover, it is valuable to those looking to encourage greater health and wellness in the face of crisis by establishing more venues designed to facilitate healthier living. As gyms tighten hygiene and social distancing standards, these venues may very well establish a first line of defense against future pandemics.

Data:

- List of postal codes from Canada to establish the scope of data to be refined to the city, neighborhood, and venue level.
- Geographical coordinates for neighborhoods to plot the map and return venue data.
- Gym Venue data used to explore and cluster neighborhoods.

Sources of data:

Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) will be scraped to retrieve the data contained in the table of postal codes and transformed into a pandas dataframe for further analysis. Python requests and beautfulsoup packages will be employed towards this end. Retrieve geographical coordinates such as latitude and longitude for neighborhoods via Python

geocoder package. Use the Foursquare API to extract venue data for all neighborhoods and focus our attention on the gym category.

Methodology:

We begin by extracting a list of postal codes to limit our geographical scope to Toronto, Canada. This data is contained within a table format on the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). This task will be completed through web scraping, which will depend on the Python packages; requests and beautiful soup. To acquire the geographical coordinates, we leverage the Geocoder package to convert an address into latitude and longitude. With data in hand, we generate our pandas dataframe for further analysis.

We proceed with the Foursquare API to retrieve the top 100 venues within 500 meters. This step requires registration with Foursquare as a Developer to attain the Foursquare ID and Foursquare secret key to perform API calls by passing in the neighborhood latitude and longitude values from above via loop. Foursquare returns the data in the prescribed form for convenient extraction of venue name, category, and geographical coordinates. Next, determine the number of venues returned to establish a sense of venue density, and determine the number of unique categories curated from each returned venue. Group rows by neighborhood and by mean frequency of occurrence per venue category and begin analysis of each neighborhood. This process is required to prepare data for our machine learning analysis, which is further refined to our venue data of choice, the “Gym” category.

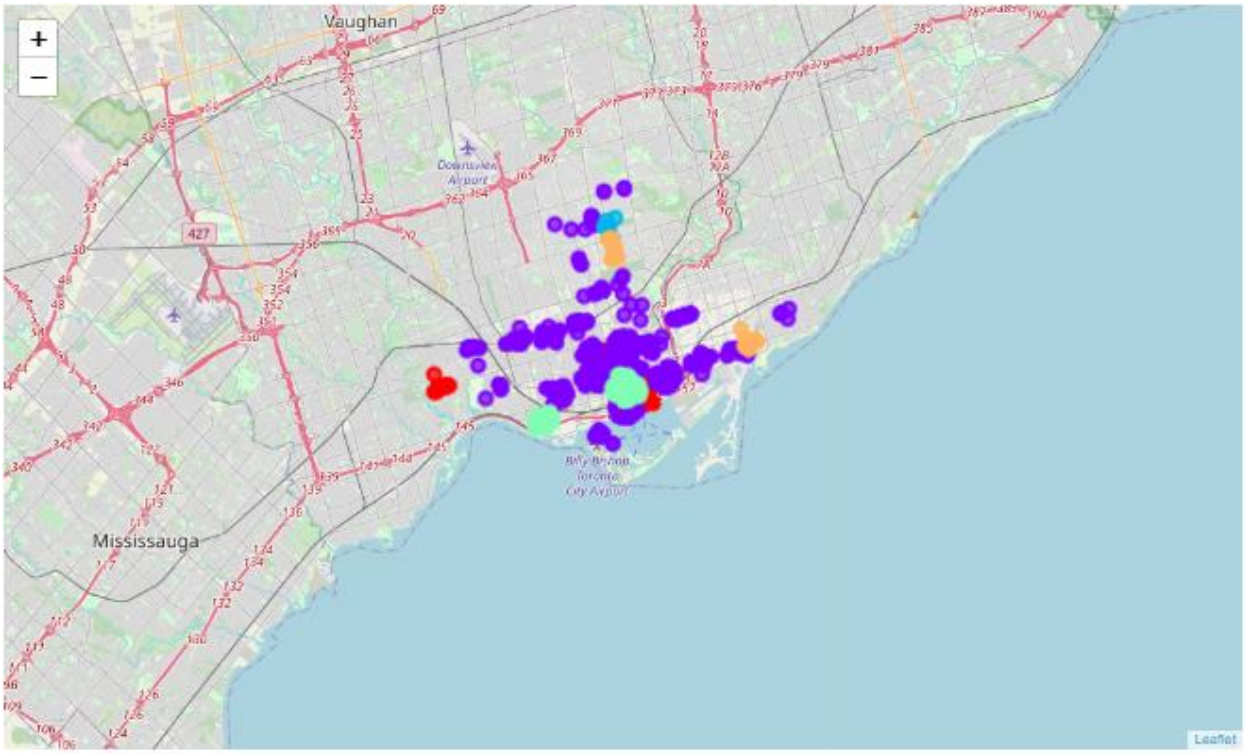
We then deploy K-means clustering, a simple yet popular unsupervised machine learning algorithm used in vector quantization. The process may be conceptualized as the partitioning of n data points into k clusters determined by distance to a cluster centroid, the nearest mean. We used k=5 to organize data points and elucidate five clusters of varied density from which we may determine which neighborhoods have the fewest gyms and which the most. To visualize the data, we generate a map of

each cluster using the folium package. From here, we may begin to make inferences about which neighborhoods are optimally suited to hosting another gym venue.

Results:

By K-means clustering we have elucidated five clusters, distinguished from another by density of "Gyms".

- Cluster 0: Low Gym Density
- Cluster 1: No Gym Density
- Cluster 2: High Gym Density
- Cluster 3: Low- Moderate Gym Density
- Cluster 4: Moderate Gym Density



Discussion:

We see that cluster 0 and cluster 1 have the lowest density of gyms, thus these clusters may represent a population of peoples currently underserved in the domain of fitness & health. In contrast, cluster 2 has the highest density of gyms and may be saturated by over-supply. Further analysis is required to identify those contributing factors which may inform the low occurrence of gyms in cluster 0 and cluster 1 in contrast to the oversaturation of gyms in cluster 2. Limitations include the use of one factor, i.e. the density of gyms in the area. Factors such as median household income, demography information such as age, characterization of the cluster as residential, business, or other may provide further guidance to real estate developers and those interested in developing gyms and wellness centers. Based on this study, cluster 0 and cluster 1 would benefit greatly from the introduction of a gym or wellness center and should be the focus of future development.

Conclusion:

In summation, we present a methodology to determine an optimal location for future gym or wellness center development by identifying data sources, extracting and wrangling the required data, applying machine learning techniques, and offering recommendations to potential developers. Cluster 0 and cluster 1 represent two areas of underserved populations which would benefit from the development of gyms and wellness centers. Thus, improving the community as a whole by providing greater access to health and wellness resources, a vital component in the initiative against chronic disease.

References:

1. WHO. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://web.archive.org/web/20200502133342/https://www.who.int/dg/

speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefingon-covid-19–11-march-2020 (2020).

2. Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/map.html. Accessed 8/12/20.
3. Liu Y, Gayle AA, Wilder-Smith A, Rocklov J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med. 2020.
4. Deng, G., Yin, M., Chen, X. & Zeng, F. Clinical determinants for fatality of 44,672 patients with COVID-19. Crit. Care 24, (2020).
5. Tanasescu M, Leitzmann MF, Rimm EB, Willett WC, Stampfer MJ, Hu FB. Exercise type and intensity in relation to coronary heart disease in men. Journal of the American Medical Association 2002; 288:1994–2000.
6. Dunstan DW, Daly RM, Owen N, Jolley D, De Courten M, Shaw J, Zimmet P. High-intensity resistance training improves glycemic control in older patients with type 2 diabetes. *Diabetes Care.* 2002; *25*: 1729–1736
7. Estruch R, Ros E, Salas-Salvado J, Covas MI, Corella D, Ar os F, G omez-Gracia E, Ruiz-Gutie´rrez V, Fiol M, Lapetra J, Lamuela-Raventos RM, Serra-Majem L, Pinto X, Basora J, Mu ~noz MA, Sorlı JV, Martınez JA, Fito M, Gea A, Hern an MA, Martınez-Gonzalez MA; PREDIMED Study Investigators. Primary prevention of cardiovascular disease with a Mediterranean diet supplemented with extravirgin olive oil or nuts. N Engl J Med 2018;378:e34.