Battle of the Neighborhoods

By Omar Ragab

Introduction

With obesity becoming a major problem around the world health and fitness are more important than ever. Worldwide obesity rates nearly tripled from 1975 to 2016 and heart disease is the leading cause of death everywhere in the world end obesity is a major link to heart disease. A gym owner is looking to expand his portfolio and open up a gym in the North York area of Canada in order to promote the importance of physical active lifestyles. This gym owner wants to target areas that have few options when it comes to gyms and other outdoor activities. The gym owner would like to know which neighborhood would be the best neighborhood to open up his new gym and have the least amount of competition in the area with the most amount of demand.

Data

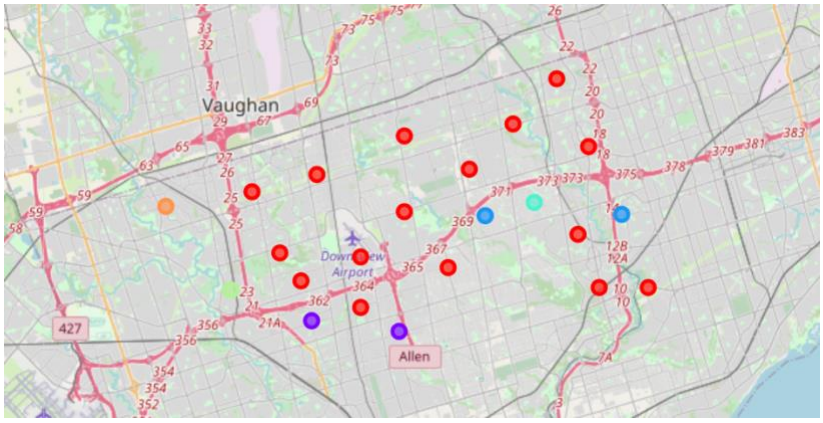
The data used for this capstone project will be obtained by web scrapping Wikipedia in order to first acquire information about the different neighborhoods in Toronto, Canada such as the neighborhood name, the neighborhood's postal code and its respective borough. The CSV file provided by IBM contains the data for the longitude and latitude of each postal code in Toronto, Canada and will be merged with the data from Wikipedia. The foursquare API will be used to gather information about each neighborhood's venues. K-means clustering will be utilized to cluster the neighborhoods based on similarities which the unsupervised machine leaning model will interpret. The folium python package will be used to visualize the clusters layered on top of a map of North York, Canada.

Methodology

The python package called pandas was utilized to scrap data from Wikipedia that contained information on the neighborhoods, boroughs and postal codes of Toronto, Canada. The data obtained from Wikipedia was then merged with the CSV file provided by Coursera on the postal codes. Using the panda's module, the data was then cleaned, and the North York borough set singled out to apple this project on. The foursquare API was used to make a request about additional information about the venues in each neighborhood in North York. Once the venue data was obtained from the foursquare API, one hot encoding was applied to the data frame of venues in their neighborhood. From this, the top 10 venues for each neighborhood were calculated. The sklearn module in python was used for its unsupervised machine learning model code k means clustering. The k means clustering algorithm was applied to the data set and formed 5 clusters as a result. The folium python package was utilized to apply a graphical visualization of the clusters over top of a map of North York, Canada.

Results

Using the unsupervised machine learning technique, k-means clustering. Of the 5 clusters created, the majority of neighborhoods fell into cluster 4 (seen in red). In cluster 4, of the top ten venues within the neighborhoods, gyms, parks, athletics, pools and other physically active venues were all a common them. The neighborhood in cluster 3 did not have any venues that had to do with physical activity while the neighborhoods in clusters 0,1 and 2 had one venue in their top ten venues that was related to physical activity.



Discussion

North York, Canada has 24 neighborhoods and based on the clusters seen in the results section is would be appropriate to say that the location with the least amount of physical activity venues would be cluster 3 which is indicated in orange. It is different from the rest of the clusters which all have at least one venue related to physical activity within their top 10 most common venues. Cluster 4 (seen in red) would have the most competition when it comes to opening a gym because within this cluster in their top 10 most common venues there are multiple places that people can be physically active. This decreases the chances that someone who is looking to be more physically active with the sign up for the new gym. Humber Summit (cluster 3) looks like the appropriate choice for somewhere that would have the least amount of competition and also is a place where there aren't many other locations in the area to be physically active.

Conclusion

With obesity being a major problem around the world more physical activity and better nutrition is needed. From the data obtained from the machine learning model it is recommended that the gym owner open up his/her gym in the Humber Summit neighborhood of North York, Canada (cluster 3, orange). This would be an appropriate location to open a gym because of all of the venues, in the top ten most common, none of them have anything to do with physical activity. Therefore, this neighborhood with have little competition and not many other options when it comes to a place to be physically active.