Data Science Capstone Report

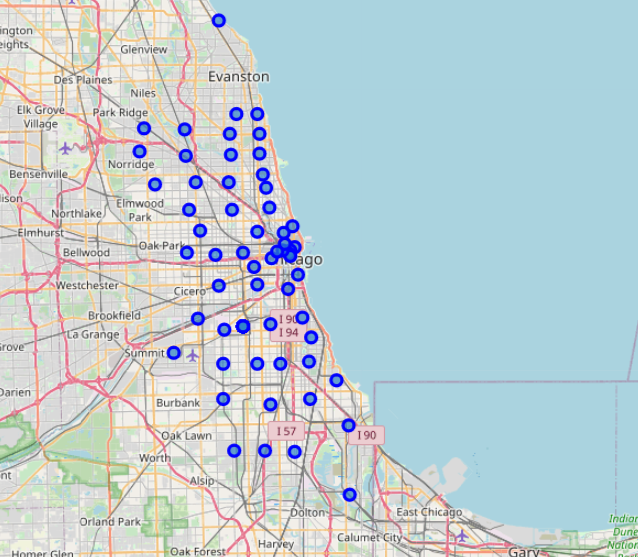
**Introduction**

Looking at the clustered neighbourhoods in Toronto, I focused on one of my clusters in which every neighbourhood had a park as its most or second most frequent venue. These neighbourhoods also had many playgrounds and dog parks, and often had schools. This cluster seems to be mainly suburban areas. I also noticed that every neighbourhood in this Toronto had a yoga studio in the top four most common venues, making yoga the most popular business.

My idea is to apply this clustering to other cities and find the suburban areas that don’t have many yoga studios. Any neighbourhood in the suburban cluster that doesn’t have a yoga studio in its top 5 most common venues will be recommended as a good area to open a yoga studio. This will be a neighbourhood with a likely unmet demand for a yoga studio.

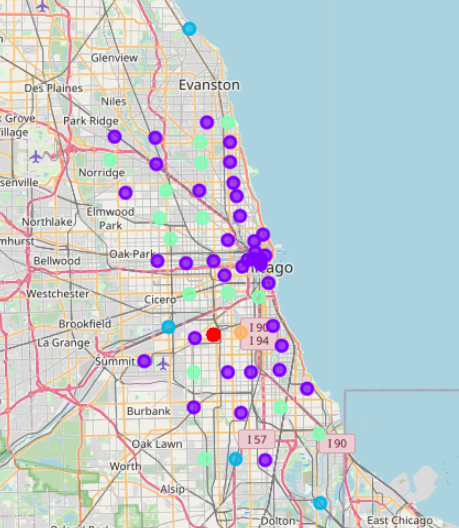
**Data**

This problem was solved using foursquare location data to find the common venues in each of a city’s neighbourhoods. The neighbourhoods were clustered based on common venues. In this example, Chicago was selected as the target city, because it has similar size, population, and location to Toronto. Zip codes were used to identify the different neighbourhoods, and the zip code data was taken from the city of Chicago website. The data was cleaned to include only the areas within Chicago city limits, and a map was generated to visualize the different areas.



**Methodology**

To cluster the neighbourhoods, a k-means clustering algorithm was used, with 5 different clusters. Before clustering, the data was arranged to show the top 10 most common venues in each zip code. After clustering, another map was generated, this time using different colours to indicate zip codes in different clusters.



Next, the process of determining which cluster or clusters were suburban type neighbourhoods by examining the frequency of parks, gyms, and dog parks. In this example it was determined that clusters 2 and 3 (blue and green) were suburban neighbourhoods, making them potential candidates for opening a new yoga studio. Finally, each of the zip codes in the suburban cluster was examined to see if it already had many yoga studios. Any area that did not have a yoga studio in its top eight most common venues was recommended.

**Results**

16 zip codes in Chicago were recommended for opening a new yoga studio. With this narrowed down list, a businessperson could further explore each of these areas to find a suitable candidate for their new business.

**Conclusion**

Based off the success of yoga studios in suburban neighbourhoods in Toronto, a clustering algorithm was used to identify the suburban areas in other cities and recommend possible areas that might have an unmet demand for a yoga studio. This tool could potentially help future business owners find the location to start a business that will provide optimal success.