1. **Introduction**

A friend of mine is planning to immigrate to Toronto, Canada from Hong Kong with her family. She is with her husband and has two son and a dog. I'm trying to identify the characteristic of each neighbourhood, their most common venues nearby and their clustering similarity, and try to leverage the data science technique to evaluate the neighbourhood of Toronto, and recommend the best option for her to stay and start her new live.

1. **Data**

I will be using the Toronto neighborhood data in Wikipedia, which offer the postal code, and map to the coordinate of each neighborhood. Then by leveraging foursquare location data, I can locate the venue or attraction nearby. After extracting all the venue of each neighborhood, I will evaluate the most common (top 10 revenue) features of each district, then make use of KNN technique to cluster each neighborhood with similarity to make my recommendation to my friend.

1. **Methodology**

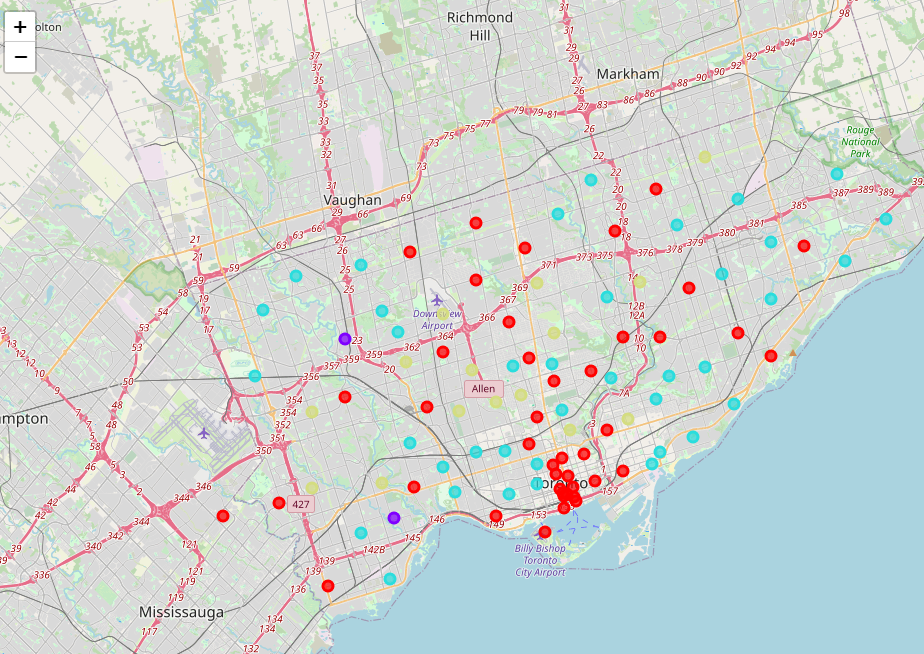
The Toronto neighborhood data will first be extracted with web scrapping. Python library beautiful soup is being used to extract the full list of postal code, neighborhood and borough. It will then be map into a csv file containing the latitude and longitude information of the neighborhood.

Then we will explore each neighborhood by making use of the foursquare api. We extract all the nearby revenue within 500 meters of each neighborhood, and group the revenue by frequency of occurrence. The top 5 most common venues of each neighborhood will be used for next steps of machine learning analysis.

Lastly we make use of the KNN machine learning library to cluster the neighborhood into 4 categories. These categories will be examined and proposed to my friends for the most optimal place for her to live.

1. **Results**

The map below shows the clustering result of Toronto by making use of KNN=4.



Each of these clusters of neighborhood has its unique characteristic of revenue:

Cluster 0 is common with coffee shop, with large variety of restaurants of different cuisine.

Cluster 1 is dominated with baseball field, drug store and diner. There are few venues and attraction in these two neighorhood.

Cluster 2: Offer most variety of venues and restaurants, some unique services such as business service, candy storem, golf course of hakka restaurants can also be found here.

Cluster 3 is most common with park, with playground and swimming school. This may be optimal for children to growth and play.

1. **Discussion & Conclusion**

Choosing the right K in KNN clustering is the key to success. When using large or small value of K, the resulting clustering will be dominated in particular clusters, resulting in over fit in some cluster with 1 or 2 neighborhood. By choosing the value of 4, each cluster contains sufficient amount of neighorhood with certain similarity. From the result of the 4 cluster, I would recommend my friend to live in neighorhoods in cluster 3, in which there are lots of park and playground for her children and dogs. At the same time it also offer variety of restaurant and living venue there may be convenience for her live.