**Capstone Project: Building a Public Park in Toronto**

**Introduction / Business Problem**

Major cities with dense population such as Toronto often lack public areas designated for outdoor leisure. This project will use Foursquare API to understand at a high level the infrastructures of Toronto and determine an ideal location to build a public park.

**Data**

The Wikipedia page used in week 3 of this course contains a list of neighborhoods and establishments in Toronto. Web scraping techniques were used to extract the data from the web page using various packages from Python. Then geographic coordinates of the different locations will be acquired using Python Geocoder package. Lastly, Foursquare API will be used to get the detailed venue data to determine an ideal location for a new park.

Data Required:

1. List of neighborhoods in Toronto
2. Geographic coordinates and map of Toronto
3. Venue data from Foursquare

**Methodology**

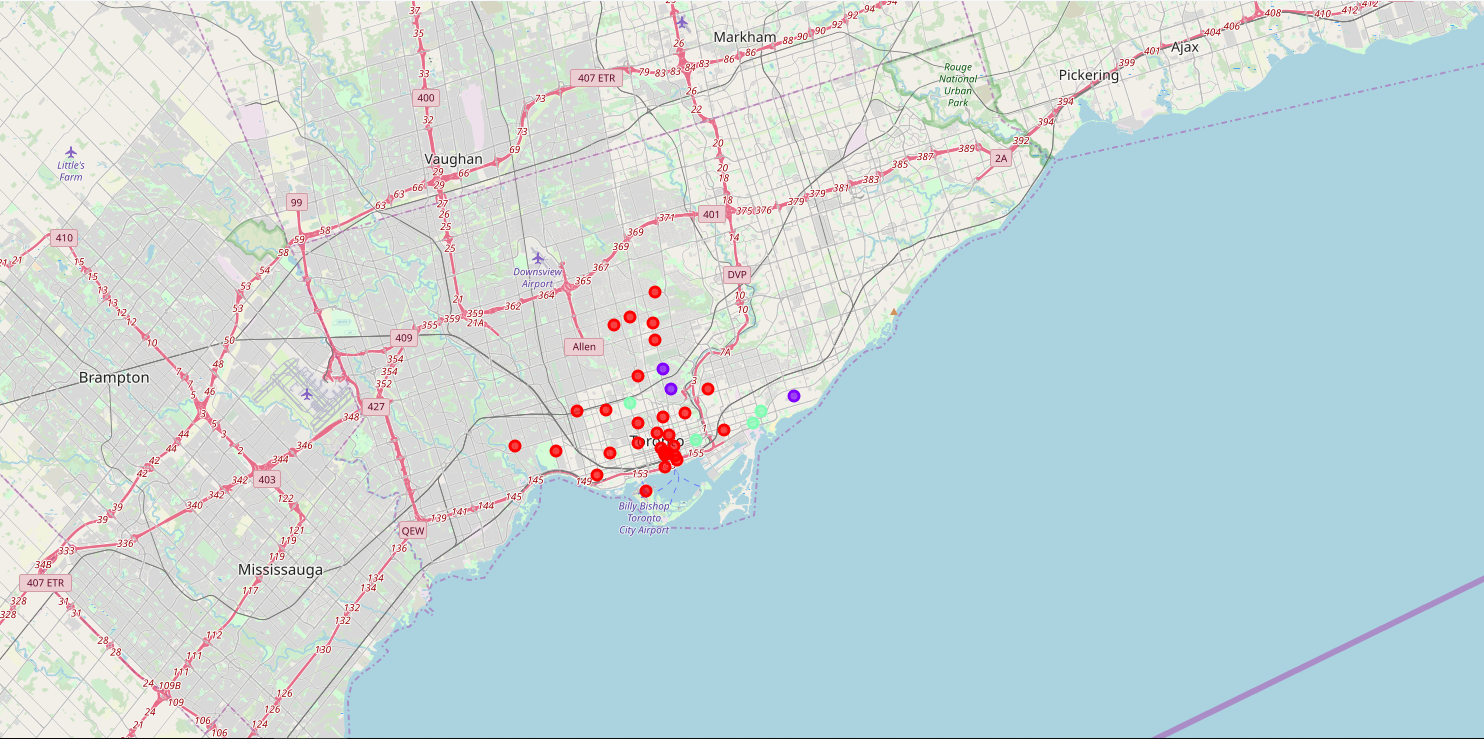
First, web scraping via Python beautifulsoup package was used to gather a list of neighborhoods in Toronto. Subsequently, Geocoder package was used to acquire latitude and longitude of the locations from web scrapping. Once the required information was gathered, data wrangling was performed to clean and manipulate the data, and then to populate them into Pandas dataframe. Folium package was used to map the locations and thus verify that the correct dataset was gathered for the analysis.

After the first session of data gathering, Foursquare API was called to retrieve the top 100 venues within a radius of 300 meters. Making API calls to Foursquare by passing the geographical coordinates of the neighborhoods will return venue data in JSON format. From there venue category, venue latitude and longitude will be extracted and mapped to each neighborhood. Then, mean of the frequency of occurrence of each venue category will be calculated by neighborhood and then filtered to “Parks”.

Final step of the analysis is to perform k-means clustering. K-means clustering algorithm identifies k number of centroids (3 in this analysis), and then allocates every neighborhood data point to the nearest cluster centroid, while keeping the radius as small as possible. The results will be used to identify different concentrations of parks throughout Toronto.

**Results**

Three clusters were created. Red being Cluster 0, blue being Cluster 1, and green being Cluster 3. Cluster 1 has many (relatively) while cluster 2 has moderate number of parks. Cluster 0 on the other hand has little to no parks. Cluster 0 also has the most neighborhoods.



**Discussion**

Cluster 1 has many (relatively) while cluster 2 has moderate number of parks. Cluster 0 on the other hand has little to no parks. Parks should be built in neighborhoods in cluster 0. Cluster 0 has the most neighborhoods, providing several options for potential park locations. This represents a great opportunity and high potential areas to open new parks. Therefore, this project recommends stakeholders to start building parks within neighborhoods in Cluster 0, raising parks density similar to that of Cluster 1.

**Conclusion**

In this Capstone project, business problem was identified, data requirement was specified, data wrangling was performed, and machine learning by clustering was performed. The conclusion to the business question raised in the introduction section is the following: The neighborhoods in cluster 0 have little to no public parks and thus the most preferred locations to open one. The conclusion can be used by stakeholders to improve public infrastructure in Toronto.