**The Battle of Neighborhoods**

* **Introduction/Business Problem**

A Group of Medical Practitioners (Medical Laboratory scientists) decides to move their abode from an African Country to the city of Toronto. They intend to arrive in Toronto and establish various medical diagnostic centers (Medical laboratories) across various Neighborhoods in the city of Toronto. Faced with the difficulty of selecting the best and most favorable neighborhood out of 140 neighborhoods in the city of Toronto, they employed the help of a Data scientist.

The Top factors to be considered in choosing the most favorable neighborhoods are: neighborhood with low amounts of already established medical diagnostic centers/Medical laboratories, neighborhood with a high number of Hospitals around, neighborhood with a high population and neighborhood with a low crime rate.

* **Data**

1. **Datasets from Foursquare:**

**The main data required to solve this problem was collected from Foursquare using the foursquare api. The data consist of the result of a venue search for the number of all available Medical laboratories within a neighborhood at a range of 500 meters with a limit of 100.**

**Another set of data also collected from the foursquare api was the number of all available Hospitals within a neighborhood at a radius of 500 meters with a limit of 100.**

1. **Datasets from** <https://edu.hub.arcgis.com/datasets>:

The other datasets which was needed to complete this project was collected from the website <https://edu.hub.arcgis.com/datasets>. These datasets includes:

1. CityofToronto\_WFL1: A spreadsheet (csv-file) that contains the name of all the 140 neighborhoods in Toronto alongside their respective latitude and longitude
2. CrimebyNeighbhToronto2019: A spreadsheet (csv-file) that contains all the 140 neighborhoods in Toronto alongside the number of crimes that has been committed in each respective neighborhood.
3. Toronto\_Neighbourhoods: A spreadsheet (csv-file) that contains the name of all the 140 neighborhoods in Toronto alongside their respective Population.

* **Methodology**

**Data Pre-processing**

After collecting the needed data as described above in the data description section, I imported all data sets into a pandas data frame. The first thing done after importing data sets into pandas was to write a forward loop program to iterate through the latitude and longitude of the respective neighborhoods on the cityofToronto\_WFL1 Data frame, passed this latitude and longitude into foursquare api and made a search query for available Medical Laboratories and Hospitals per 500 meter radius with 100 limit of each neighborhood. The Json file returned from the api calls was also iterated through, to get a number of the respective hospitals and Medical laboratories in each neighborhood and the numbers was appended into a python list and later updated into the cityofToronto\_WFL1 Data frame. Also from CrimebyNeighbhToronto2019, I summed up the total crime per neighborhood to get the total crime rate, then I concatenated the value of the total crime to its respective neighborhoods in the cityofToronto\_WFL1 Data frame.

Lastly from the Toronto\_Neighbourhoods I extracted the Population values for the respective neighborhoods and joined also to the cityofToronto\_WFL1 Data frame. The data is now ready to be used in solving the problem

**Finding the best Neighborhood**

***The best Neighborhood is to be chosen based on:***

* Neighborhoods with low amounts of already established medical diagnostic centers/Medical laboratories ; the lower the available established medical diagnostic centers per neighborhood, the lower the competition and the higher the customer.
* Neighborhoods with a high number of Hospitals around; the higher the number of hospitals around the better the referrals and the higher the customer.
* Neighborhoods with a high population; Higher populations also imply higher customers
* Neighborhoods with a low crime rate for safety purposes.

**Machine Learning Algorithm**

A clustering algorithm **(K-Means**) was used in selecting the best neighborhoods.

**WHY KMEANS ?**

**Since the aim of the project is to select the best neighborhood that meets the criteria listed above, a clustering is done to group the neighborhoods. With kmeans, neighborhoods with similar traits fall into same group and the group that best suites our best neighborhood description is chosen**

**Results**

**On doing kmeans clustering with a set n\_cluster of 5, the best clusters that suites our best neighborhood description was cluster 0, which was then exported to another data frame for final analysis. Of the 140 neighborhoods, 115 fall into cluster 0.**

**From the Cluster 0 data, the average population and average crime rate was calculated and neighborhoods with a crime rate higher than average and neighborhoods with population lower than average were dropped. This reduces the remaining available neighborhoods to 18.**

**Finally, of the 18 remaining neighborhoods, those with an already established laboratory were also dropped, leaving our final best neighborhoods to 11 neighborhoods.**

**Visualization**

**A folium map of Toronto with markers for the 11 selected final best neighborhoods was generated for visualization purposes.**

**Conclusion**

**Of the 11 finally selected best neighborhoods, 2 of the neighborhoods are further recommended that better suites the best neighborhood description.**