**Capstone Final Project**

**Introduction/Business Problem**

Life is precious. Any threats to life needs to be dealt with swiftly, responsibly and in line with the laws of the land. The current COVID-19 pandemic has taught us the importance of swift action in the face of an invisible enemy. Scrutinising this pandemic further, there are several lessons learned which should be implemented at the soonest to tackle future pandemics. One such problem is the identification of appropriate screening locations. This project aims to present weighted (where the weights are the populations of the individual boroughs) k means clustering approach to identify appropriate locations within the city of Toronto, Canada for setting up of future ad-hoc virus detection centres. This project also aims to find if any parks are available vicinity of the chosen location as open air spaces is perhaps better for this purpose.

**Data**

The data for this study was scrapped from the following websites:

1. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>
2. <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0>
3. <http://cocl.us/Geospatial_data>

Here are the first 12 rows of the data.



Basically, this dataset gives a selected neighbourhoods, geographical locations, population and borough based on the different postal codes of the city of Toronto, Canada.

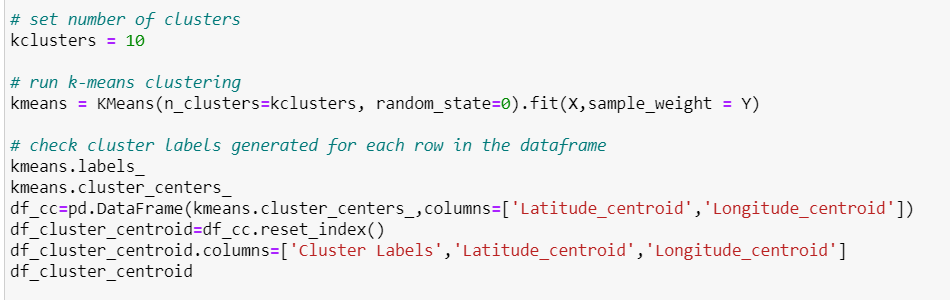
Foursquare API will be used at the end of the study (and presented in a separate section of this report) to identify the closest parks to the kmeans centroid cluster

**Methodology**

Firstly, the total population in Toronto was determined in order to estimate the number of detection centres that would be required.



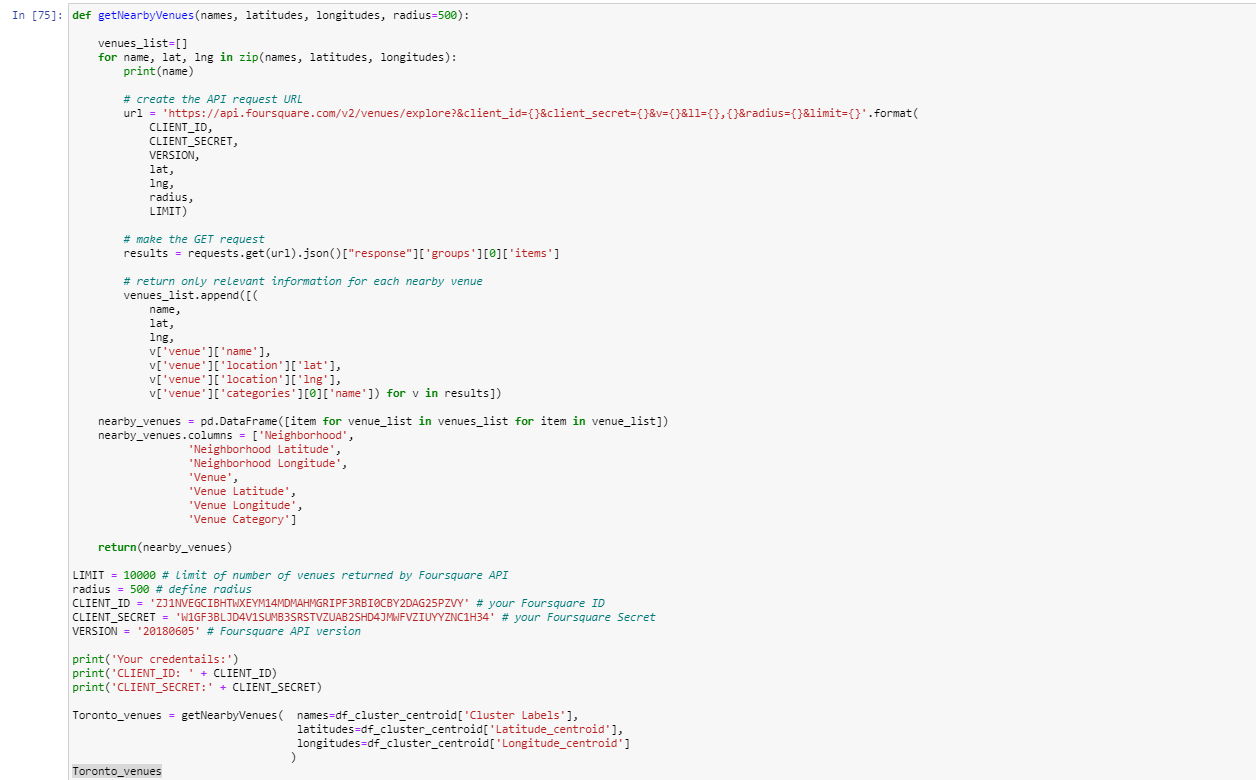
On a rough estimate, 10 centres suffices for a 2million plus population. As such weighted k-means clustering with k=10 was used to identify the 10 detection centres. The image below captures the utilised code.



The resulting data were then plot using folium.map as shown in the code below.



A total of 10000 venues within a 500m radius of the detection centres were then obtained via Foursquares API.

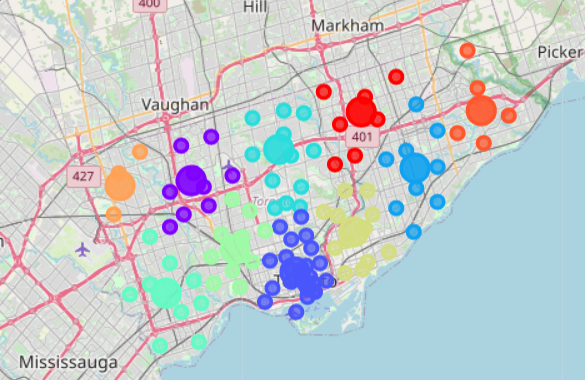


Finally, the following code produce the closest park to the detection centres:

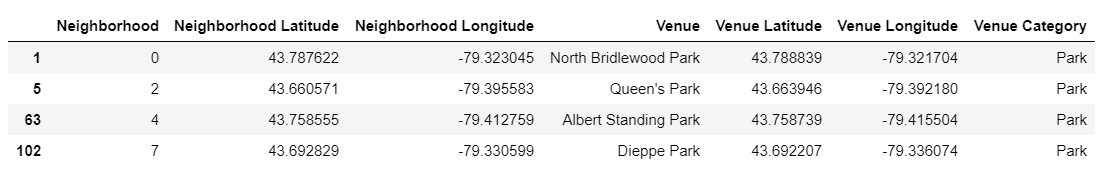


**Results**

The map below displays the sectored postal codes in the city of Toronto. The smaller circles represent the individual postal code locations while the bigger circles are the weighted k-means centroids.

****

The table below captures parks that are closest to four detection centres.

****

**Discussions**

The map in the previous section nicely captures the possible locations of the detection centres in preparation for a future pandemic. Nevertheless, these locations can be slightly optimised to the take various factors into consideration. One consideration could be the requirement of an open space for which parks may be a good candidate. Looking at the table in the previous section, only 4 of the 10 locations had parks and as such other venues such as government clinics or nearby hospitals could also be considered. Nevertheless, this is out of the scope of the current project.

**Conclusions**

This project successfully presented how k-means clustering could be used to identify detection centres for future pandemics. Nevertheless, the centroids does not necessarily give the optimal location of the detection centre. The vicinity of the centroids should be scouted for optimal placement of the detection centres.