**FINAL CAPSTONE PROJECT FOR IBM PROFESSIONAL DATA SCIENCE CERTIFICATION ON COURSERA**

**Prepared by:** Nachiket Karajagi

**Introduction:**

As part of the IBM Data science certification on Coursera, we were asked to perform analysis and clustering based on location data from FourSquare. In the earlier classes, we had done segmentation of New York and Toronto cities. I decided to extend the Toronto analysis further for the project. As we are in the midst of a COVID-19 pandemic, I planned to the Toronto location data along with the COVID-19 reported cases in Toronto, to identify clusters of locations which might be risky for people to go to. This could help people avoid venues with higher concentration of active cases. This can further be segmented by Age groups of people who have been infected and the kind of venues which are located close by.

**Business problem:**

As we move into a Post-COVID world, and the cities relaxing the restrictions, it will be important for people to know the high risk areas with active cases and take necessary precaution. This may also help the health care workers to focus on specific age groups that may be at higher risk, and plan for medical facilities accordingly.

**Data:**

I will be leveraging 4 data sources for this project

1. The neighborhoods of Toronto as per the Toronto Open data portal. This data set has the list of 140 neighborhoods in Toronto along with latitude & longitude, and also the geometry of the neighborhood.
2. Foursquare data for the most popular public places in certain areas. This data set will be based on the output of the API from Foursquare to fetch the location and venue details.
3. COVID-19 cases reported in each neighborhood from the Toronto Open data portal. This dataset is refreshed daily by the Toronto portal. This data set has details of each case reported, the age group of the person, neighborhood in Toronto, gender, and status pf the case (Active, Resolved, Fatal). For the purpose of this project, I will use only active cases for analysis.
4. Pedestrian foot traffic across Toronto for prior 8 hours.

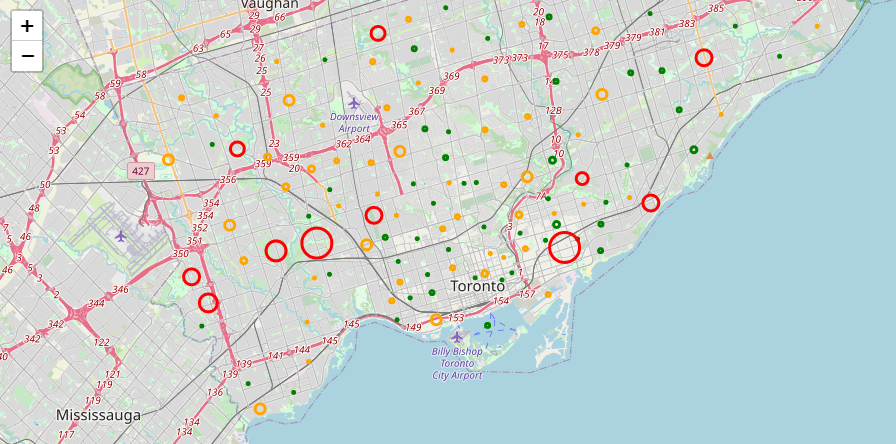
By analyzing these 4 sources of data, we can cluster the venues based on the density of COVID cases by location.

**Methodology:**

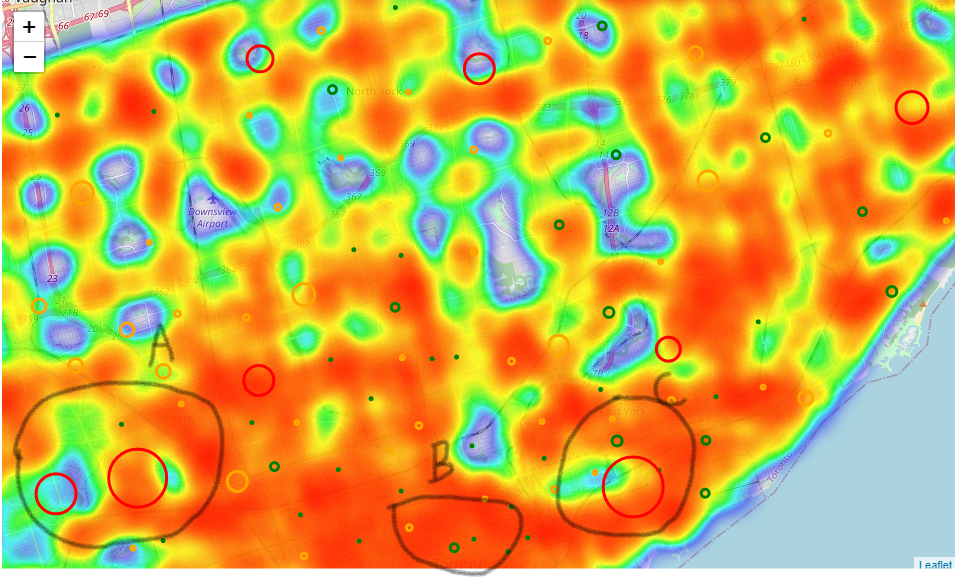
1. **Fetch data for Toronto neighborhoods:** I fetched the data for the Toronto neighborhoods from the Toronto Open data porta API into a data frame. Once I cleaned the dataset and removed any spaces in the area name, I created a clean data frame to analyze further.
2. **Fetch data for COVID cases reported in Toronto :** The Toronto open data portal also publishes and updates daily the number of reported COVID cases (anonymized for personal data), which provided the details of age group, gender, and neighborhood in which the case was reported along with the outcome of the case. The status could be Active, Resolved or Fatal. For this analysis, I just used the active cases as that would be more beneficial to people visiting the neighborhoods. Once this data was downloaded and cleaned, I was able to create a one hot encoding for the data to calculate the total cases for each neighborhood and also by age group.
3. **Merge Toronto neighborhood data with COVID cases:** The data obtained in step **a** and **b** were merged to generate one dataset that could be visualized. I created a map of Toronto to indicate the neighborhoods with cases indicated in Red, Green or Yellow markers based on the following scale – (Greater than 10 cases = Red; Cases between 5 and 10 = Orange ; cases less than 5 = Green)
4. **Fetch pedestrian foot traffic for recent 8-hour period:** The Toronto open data portal also publishes the vehicular and pedestrian traffic for the most recent 8 hours by street locations. I used only the pedestrian foot traffic for this analysis. With this data, I was able to create a heat map for Toronto highlighting the pedestrian foot traffic overlaid with the neighborhoods with active cases. This would give a quick overview of the areas where there is considerable number of people and the number of active cases.
5. **Fetch popular venues for Toronto:**  I used the Foursquare API from the previous assignment to fetch the most popular venue locations for each neighborhood. Once the data was downloaded, and processed, it was one hot encoded to identify the various categories of venues and their occurrences in each neighborhood. Top 3 venues were fetched for each neighborhood to create a dataset with 1st, 2nd and 3rd most popular venues in each neighborhood.
6. **Merge COVID cases counts with venues:** I merged the COVID cases by neighborhood along with the venues dataset by neighborhood. This provided the final dataset that was used for clustering.
7. **Clustering the venues and assign to neighborhoods:** By using the k-means clustering technique, I clustered the data set from previous step into 5 clusters. Once the cluster numbers were derived, they were mapped back into a dataset which had the neighborhood, cluster label and the top 3 venue categories for the neighborhood.
8. **Analyze the clusters:** The dataset from the previous step was then analyzed to provide an idea on the popular venue categories in each cluster identified. I then merged this dataset along with the COVID cases to analyze the relationship between cluster label and the number of COVID cases.

**Results & recommendation:**

1. The first visualization was for the view of Toronto and the number of COVID cases sorted into three categories as described in step **c** of the methodology section.

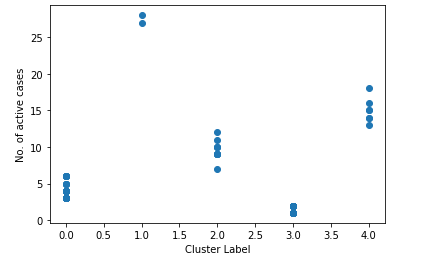


1. The second visualization was an overlay of the pedestrian traffic for prior 8 hours along with the number of COVID cases. I have marked 2 areas of interest.
   1. Section A is where there is moderate pedestrian traffic, but higher COVID cases. Its good that people are avoiding these areas.
   2. Section B is where the pedestrian traffic is high but number of reported COVID cases there are low. This could be because they are commercial areas and no residential locations resulting in fewer cases.
   3. Section C is an area where the cases reported are high, and with high pedestrian traffic. These areas should ideally be avoided.

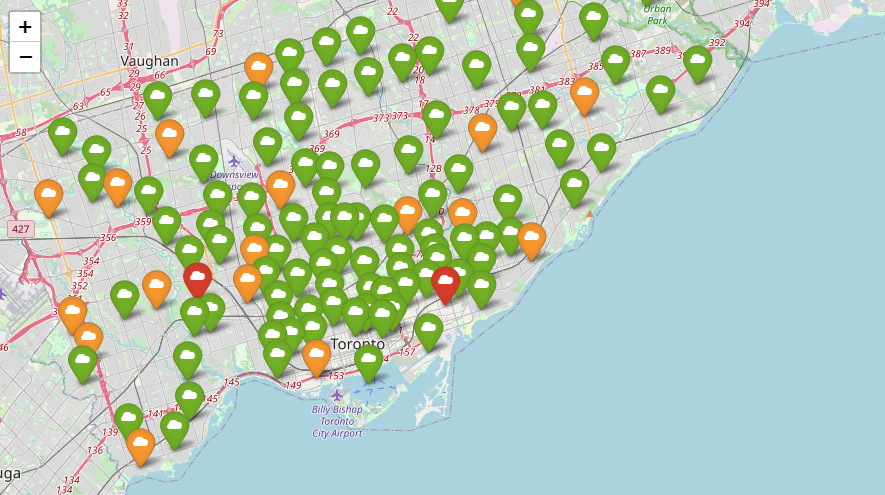


1. Cluster label vs number of cases was plotted based on the dataset from step **g.** Here it was evident that:
   1. Neighborhoods in cluster 1 were the highest risk. Venues in this location should be avoided
   2. Neighborhoods in Cluster 2 and 4 were moderate risk. These areas could be visited along with precautions such as wearing masks.
   3. Neighborhoods in cluster 0 and 3 were relatively low risk. These areas may be safer to visit.

DISCLAIMER: Since this data set is refreshed everyday, the cluster label may vary when the model is run as per the latest dataset. For the purpose of this analysis, I have used a static categorization to visualize these neighborhoods.



1. The last visualization was to visualize the high risk neighborhoods based on the clustering performed. They were color coded for easy identification.



**Limitations & future opportunities:**

Due to the paucity of time and my limited skills, I was able to analyze the data based on only the total number of cases. A few opportunities to further extend this idea are:

1. Clustering the dataset based on the age group of the COVID cases reported and the demographic data for Toronto. This can help the medical teams identify the high-risk age group neighborhoods and prioritize the care.
2. Cluster the dataset based on gender of the COVID case as this may help target the medical supplies.
3. Visualize the pedestrian traffic data along with the neighborhood clusters. I was unable to do this due to time constraints.

**Conclusion:**

This project gave me an opportunity to put the data analysis and Python skills acquired along the course and explore my limitations. I believe this is just the beginning of my learning journey. This course is definitely recommended for anyone wanting to start their data science learning journey.