Clustering Restaurants based on nearby fire-stations in Toronto

Vidit Agrawal

April 23, 2019

Introduction

Background

Big Metropolis are the financial and cultural centers of modern day civilization. City planning is an essential aspect for any metropolis and residents. Since, these cities are ever-growing and changing, the people responsible to take care of the city need to keep up with the infrastructure needs of the city. One of the infra-structures that a city municipal corporation has duty to provide is fire-safety. In this regard, fire-stations play the key role. The locations and the numbers of fire-stations should be such that in fire calamities, quick and efficient deployment can be made. This would ensure minimization of life and property damages. One major business establishment that is at risk of fire-accidents are the restaurants. Restaurants work with fire and heavy duty electrical equipment in day-to-day business. The city planners need to keep track of the how restaurants are growing or decreasing in neighborhoods. Accordingly, they need to keep track of fire-stations that are in that area.

Problem

The main problem that we are addressing is how the restaurants in different neighborhoods in the city cluster with respect to the fire-stations that are in their proximity. The city we are looking at is Toronto which is a major cultural and financial hub of Canada.

Interest

First and foremost, this problem would interest the city-planners of Toronto. This would enable them to take preemptive actions necessary to avoid potential weak points in fire-accidents responses. For example, if in certain area there are too many restaurants, then the city might hold off on opening of new restaurants till they either establish more fire-stations or upscale the existing ones.

Data acquisition and cleaning

Data sources

To address this problem, we first need data from different neighborhoods of the city of Toronto. We need the latest data on the restaurants and their geo-locations. We also need the fire-stations and their geo-locations as well. To get the data on restaurants first we obtain the data on the neighborhoods and their geo-locations for the city of Toronto. We used the Wikipedia page for the city of Toronto. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

The geo-coordinates (latitude, longitude) data is been taken from source made available by the course instructors.

http://cocl.us/Geospatial_data

Next, by using the neighborhood data we obtain the latest data on restaurants and their geo-coordinates from Foursquare API. This task requires one to create an account as foursquare developer via the link:

https://foursquare.com/developers/apps

Getting the latest data is important for restaurants as they open and close daily.

Finally, the fire-stations and their geo-coordinates data is made available by the city of Toronto at their Open Data portal that can be accessed via weblink:

https://portal0.cf.opendata.inter.sandbox-toronto.ca/dataset/fire-station-locations/

Data cleaning

I started with the neighborhood data so to get the different neighborhoods in the city of Toronto we looked at the postal codes and grouped neighborhoods in terms of postal codes. This is done by scrapping through the webpage on Wikipedia and saving the data-table where information is provided for the neighborhoods, the postal code and the boroughs they belong to.

For details, please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_neighborhoods_Toronto.ipynb

To this data-table we also add the information about the latitude and longitude geo-coordinates for each postal code.

Next step is to use these geo-coordinates of different neighborhoods to access the foursquare API and perform a search for ‘query = restaurants’ in the geo-location of each postal code. This search returns a json file containing details about the restaurants such as address, name, geo-coordinates etc. We extract name and geo-coordinates only as per the requirement of current project. One must be careful while getting data from Foursquare as there is a daily limit on the number of calls that can be made.

For details, please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_Restaurants_Toronto.ipynb

Finally, we get the data for fire-stations as their geo-locations from the Open data Portal for the city of Toronto. The geo-locations are provided in the UTM 6-degree coordinates (x,y) format. We convert these to latitude and longitude so that the clustering and comparison with restaurant data can be performed.

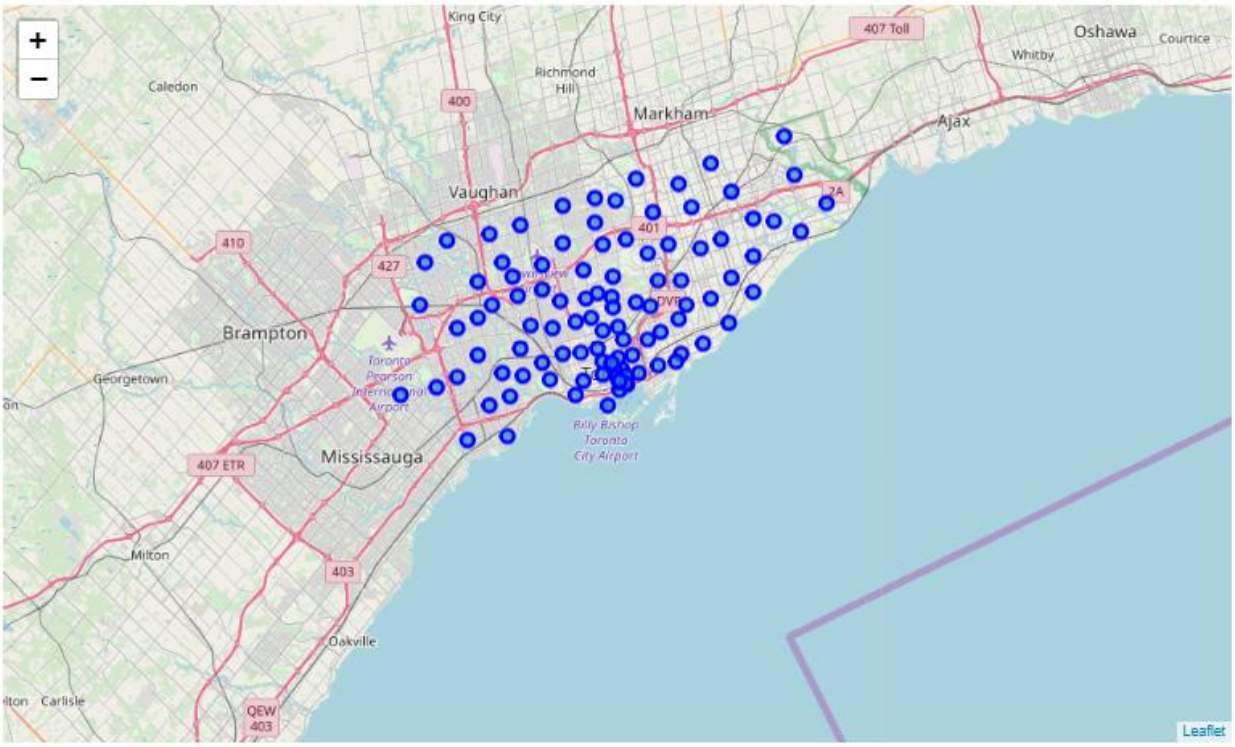
Please refer to following link to the jupyter notebook:

https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Get_Firestations_Toronto.ipynb

All the 3 datasets are saved in csv format and then finally loaded into data-frame to perform final analysis.

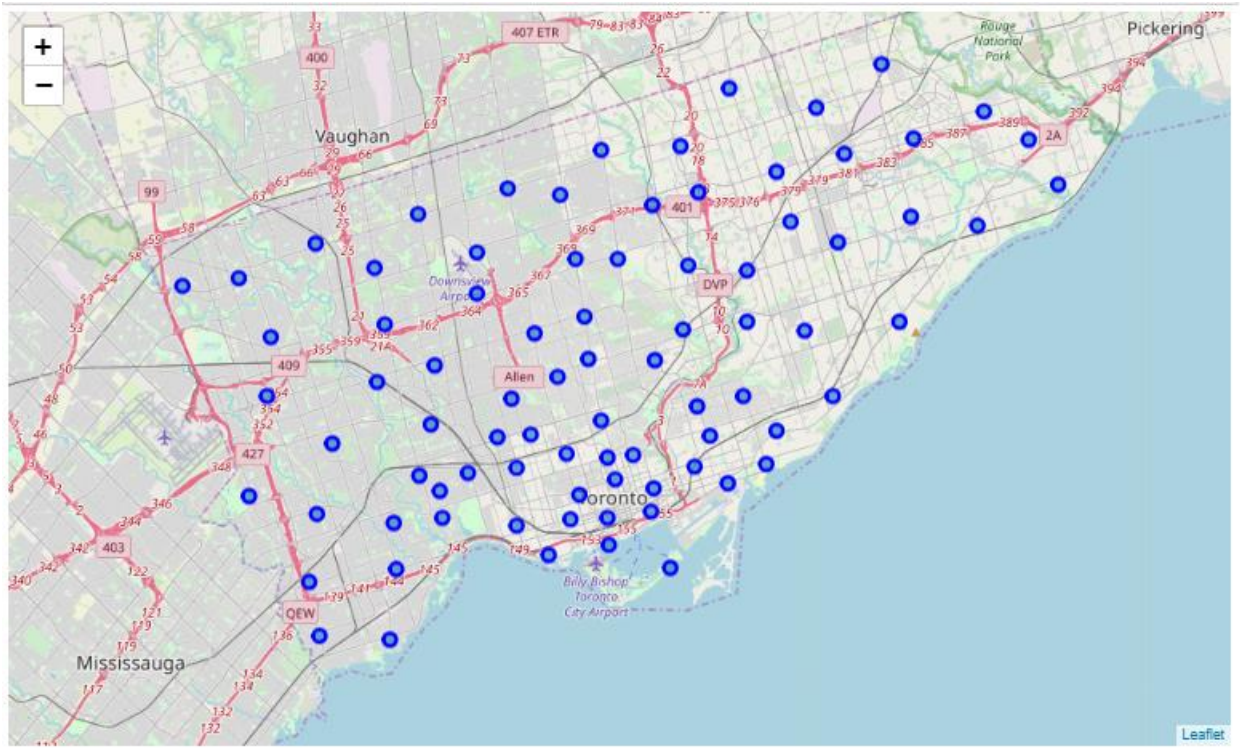
Methodology

Exploratory Analysis



First, we visualize different neighborhoods in Toronto on the map of Toronto. Here we have grouped the neighborhoods on the basis of postal codes. So, each point on this map represents the geo-location belonging to a postal code. For our analysis we had 103 rows all total. From simple visual inspection we can see that there are parts of the city which are denser meaning more postal codes and parts of city that is sparse as well. We save the neighborhood data into a data frame and first few rows are shown below:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Port Union, Rouge Hill	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
5	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
6	M1K	Scarborough	East Birchmount Park, Ionview, Kennedy Park	43.727929	-79.262029
7	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577
8	M1M	Scarborough	Cliffcrest, Cliffside, Scarborough Village West	43.716316	-79.239476
9	M1N	Scarborough	Birch Cliff, Cliffside West	43.692657	-79.264848



Next, we visualize the fire-stations on the map of Toronto. All together there are 83 fire-stations. We save the data into dataframe and first few rows are shown below:

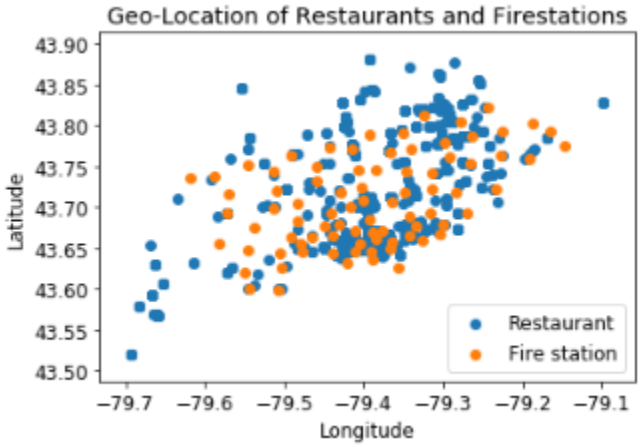
	Fire_station_#	x_6degree_UTM_coordinate	y_6degree_UTM_coordinate	longitude	latitude
0	111	629254.562	4849606.500	-79.393539	43.788230
1	112	625241.895	4847922.113	-79.443786	43.773759
2	113	632675.000	4849871.000	-79.350979	43.790005
3	114	627526.500	4847651.500	-79.415473	43.770934
4	115	633496.245	4847881.079	-79.341271	43.771948
5	116	631535.936	4847282.749	-79.365766	43.766914
6	121	628261.812	4844955.500	-79.406984	43.746542
7	122	630091.688	4844954.000	-79.384266	43.746209
8	123	633104.622	4844755.216	-79.346910	43.743885
9	125	632958.450	4841971.252	-79.349413	43.718855

A noteworthy thing here is that each fire-station is given a numerical code, so we refer to the fire-stations by their numbers.

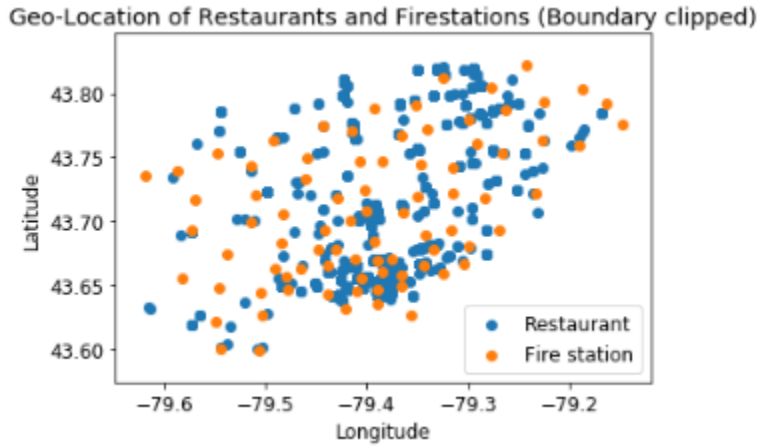
We combined the restaurant data with the neighborhood data and saved it in a single dataframe. This final dataframe has restaurants name and geo-coordinates, neighborhood it is located in, neighborhoods geo-coordinates and the Borough it is located in. Altogether we have data for 3090 restaurants that could be obtained via search call on Foursquare API under the query “restaurants”. Below we show the first few rows of this dataframe.

	Neighborhood	Neighborhood_lat	Neighborhood_long	Restaurant	Restaurant_lat	Restaurant_long	PostalCode	Borough
0	Malvern, Rouge	43.806686	-79.194353	Ted's Restaurant	43.784468	-79.169200	M1B	Scarborough
1	Malvern, Rouge	43.806686	-79.194353	Perfect Chinese Restaurant 雅璞海鲜酒家	43.787774	-79.270294	M1B	Scarborough
2	Malvern, Rouge	43.806686	-79.194353	Alton Restaurant 益衡坊	43.825582	-79.276038	M1B	Scarborough
3	Malvern, Rouge	43.806686	-79.194353	東海漁村 Tasty BBQ Seafood Restaurant (Tasty BBQ S...	43.794425	-79.353300	M1B	Scarborough
4	Malvern, Rouge	43.806686	-79.194353	Federick Restaurant	43.851124	-79.253210	M1B	Scarborough

Before we perform the clustering, we looked at how the fire-station locations and restaurants locations lie on the map. The folium maps that was used to display maps did not work properly due to large number of points, so we plotted a scatter plot with longitude and latitude as x and y axis and plotted the geo-coordinated of fire-stations and the restaurants. The plot is shown below:



We observed that some restaurants are well outside the range of fire-stations, but the foursquare API gave that information as well. Since, this would bias the clustering we removed those restaurants data that have geo-coordinates beyond the fire-stations range.



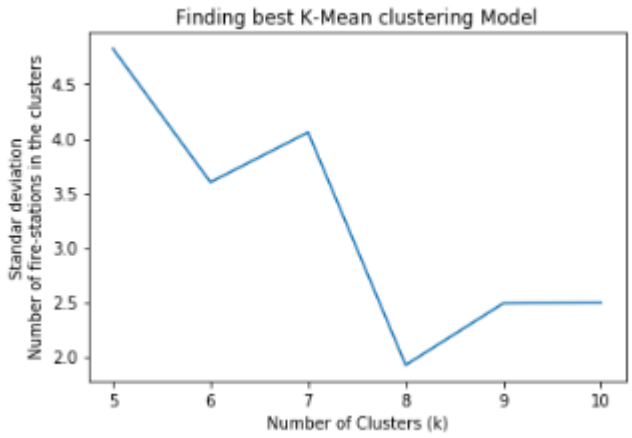
This the final data we used for clustering analysis.

Clustering Algorithm

First, we grouped the fire-stations into clusters based on there latitude and longitude data. So, this would cluster the fire-stations based on their proximity. For this we employ K-Means clustering algorithm. The main idea here is to divide the datapoints into k-clusters based on the similarity with each other. The datapoints in same clusters are more similar than the datapoints belonging to different cluster. The criteria for similarity are the distances between the data-points as we use geo-coordinates to perform K-Means clustering. Here is a link if you would like more details.

https://en.wikipedia.org/wiki/K-means_clustering

We want have clusters that have most possible uniformity. We want each cluster of fire-stations to have similar number of fire-stations under it. For this we check a range of values for the parameter 'Number of Clusters (k)'. For each k value we perform K-Mean clustering on the fire-stations data. After clustering we count the number of fire-stations in each cluster. To find the best k value we choose the one that has smallest standard deviation in the number of fire-stations in each cluster.



We find that k=8 gives the smallest standard deviation for size of fire-station clusters so we train the K-Means Model using k=8 parameter value.

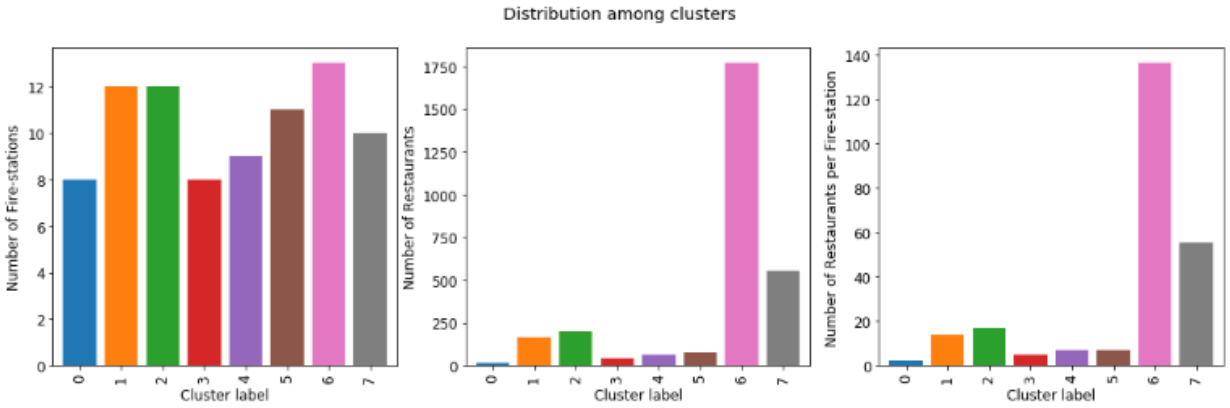
After fitting the Model on the fire-station data we finally predicted the clusters label for the restaurants data using their geo-locations. The results of this analysis are shown in next section.

For details, refer to following link to the jupyter notebook:

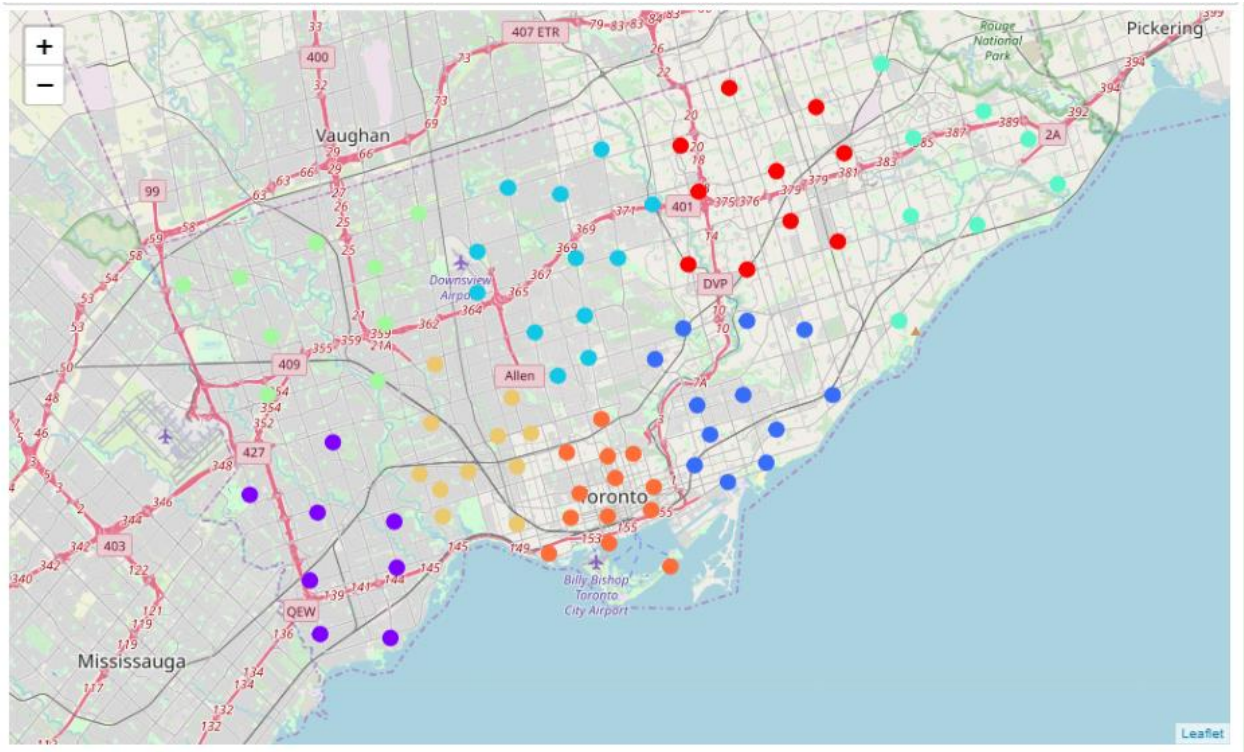
https://nbviewer.jupyter.org/github/ViditAg/Restaurant_fire_stations_clustering/blob/master/Capstone_Project_Restaurants_Analysis.ipynb

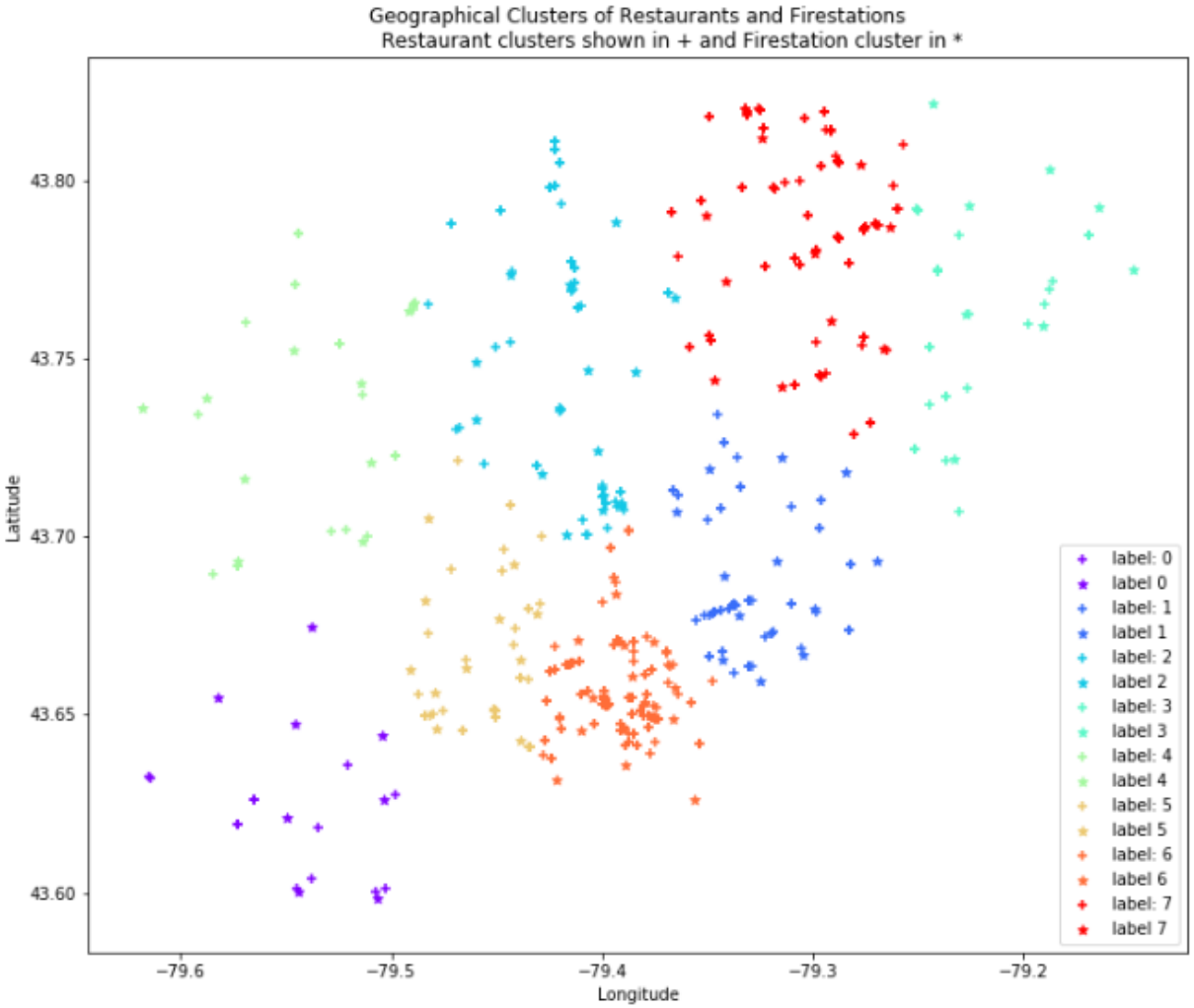
Results

We fitted the K-Means clustering model with 8 clusters on the fire-stations locations. The result that we got is shown below in the first panel. The histogram shows that in all clusters the number of fire-stations is relatively similar with highest of **13 in cluster label # 6** and **lowest of 8 in cluster label # 0**. Next, we predicted the cluster labels for the restaurants using the Kmeans model fitted on fire-station data. Here we saw huge disparity in number of restaurants in each cluster. Specially, **cluster label # 6 and label # 7 have significantly higher number of restaurants.** We can see that the average number of restaurants that will be covered by each fire-station in those clusters is much higher than the others.



Next, we visualized the different clusters on the map of Toronto and in a scatter plot. On the map we have shown the fire-station clusters. On the scatter-plot we have shown both the restaurants clusters and fire-stations clusters along with their respective clusters.





We Identified the Boroughs that are covered by clusters that have the highest Restaurants per fire-station ratios. Starting with Cluster label # 6. A total of 1768 restaurants are covered by 13 fire-stations. These restaurants are in the following boroughs.

```
Downtown Toronto    533
North York          256
Etobicoke           233
Central Toronto     210
West Toronto        146
York                116
East Toronto         92
East York            72
Scarborough          66
Queen's Park         30
Mississauga          14
```

Next, highest Restaurants per fire-station ratios. Starting with Cluster label # 7. A total of 555 restaurants are covered by 10 fire-stations. These restaurants are in the following boroughs.

```
Scarborough        300
North York         192
East York           22
Central Toronto     17
Etobicoke           10
York                 7
East Toronto         7
```

Discussion

From our analysis of clustering the fire-stations and restaurants we found one aspect of city planning that the city of Toronto keep in mind to prepare for fire accidents that might occur in restaurants. The restaurants handle fire in day-to-day business, so the fire-stations that are closest to them should be in proportion to restaurants in the area. We identified clusters where the ratio of restaurants to fire-stations was highly skewed. So, the city of Toronto needs to pay attention to this fact, either by increasing fire-stations in those areas or upscaling the ones already there. The most critical Boroughs under which these clusters lie are Downtown Toronto, North York and Scarborough. Similar clustering analysis can be done for other types of venues with regard to fire-stations or any other state amenities as well. In completion, I would mention the advantage simple clustering analysis and in turn data-science can provide to planning of big Metropolis.

Conclusion

This project is done as part of Data science professional certificate specialization with IBM on coursera.org. The data used in this project is obtained from 3 sources Wikipedia, Foursquare API and Open Data portal for city of Toronto. All these sources are open for public access and can be used for data-science projects. In this project we performed clustering analysis (K-Means algorithm) of the locations of fire-stations in the city of Toronto. We further predicted which restaurant falls in which cluster based on its location. These clusters have uniform number of fire-stations but high skewed number of restaurants. This helped us in uncovering areas where there is a much higher per fire-station load in terms of restaurants.