Coursera Capstone Project

By: Alissa Stuart

Introduction/Business Problem

Our client, a bottled water distributor, has an established role in the market of Toronto. He is the top supplier of bottled water with more than 1500 clients in the Toronto neighborhood. His main clients are hotels, coffee shops, restaurants, and bars. Currently, the products are stored in a big central warehouse outside of Toronto and distributed to different venues daily. The main problem with this is that the distribution of the product becomes increasingly time-consuming and costly. Our client wants to increase efficiency and reduce the cost by building 5 smaller warehouses in Toronto to serve his clients locally. This approach will reduce the time to spend on roads, fuel cost and become more environmentally friendly.

To do so our client asked us to find the best 5 locations in Toronto at which he can build warehouses to create smaller distribution clusters. After this, he will try to build his warehouses at the center of those clusters to minimize the distance to each venue.

Data

The data required are the locations of hotels, coffee shops, bars and restaurants in Toronto. To gather the data, we will use the locations of all neighborhoods in Toronto gathered from Wikipedia. Based on these locations we will gather the locations of all venues in these neighborhoods from Foursquare. We will filter the data to acquire the locations of the targeted venues. To inspect the data, we will use the folium library to extract the map of Toronto and visualize the locations of the venues on the map.

A k-means algorithm will be applied to the locations features to define the 5 clusters of venue. The locations of the warehouse will be defined as the centroids of the clusters. Again, to visualize the map of Toronto, the 5 clusters and the locations of the warehouses we will use the folium library.

Methodology

To define the location of each warehouse, we chose to use the k-means algorithm. By default, this algorithm minimizes the distance of each point from the centroid to the cluster. As a result, the output of the k-mean algorithm is a set of clusters whose points are lying at the minimum distance from the determined centroids.

As an input, we used the set of locations (lat, long) of venues of interest. In this example, the category of each venue is not required in the algorithm. All venues belonging to hotels, bars, restaurants and coffee shops are included in the list. The locations of hotels, bars, restaurants and coffee shops are shown as they were gathered from the Foursquare database in figures 1, 2, 3 and 4. A total of 249 coffee shops, 961 restaurants, 81 hotels, and 172 bars were found and used in the k-means algorithm.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
2	The Beaches	43.676357	-79.293031	Tori's Bakeshop	43.672114	-79.290331	Vegetarian / Vegan Restaurant
14	The Beaches	43.676357	-79.293031	Delina Restaurant	43.668867	-79.305404	Middle Eastern Restaurant
18	The Beaches	43.676357	-79.293031	Veloute Bistro	43.672267	-79.289584	French Restaurant
24	The Beaches	43.676357	-79.293031	Xola	43.672603	-79.288080	Mexican Restaurant
28	The Beaches	43.676357	-79.293031	Budapest Restaurant	43.680946	-79.310110	Hungarian Restaurant
30	The Beaches	43.676357	-79.293031	Isabella's Boutique Restaurant	43.673767	-79.282703	Japanese Restaurant
34	The Beaches	43.676357	-79.293031	Lake Inez	43.672520	-79.320712	Asian Restaurant
38	The Beaches	43.676357	-79.293031	Green Eggplant	43.670517	-79.298660	Mediterranean Restaurant
39	The Beaches	43.676357	-79.293031	Udupi Palace	43.672480	-79.321275	Indian Restaurant
47	The Beaches	43.676357	-79.293031	Jatujak	43.688421	-79.270073	Thai Restaurant

Table 1: Sample of Restaurants Locations in Toronto

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
53	The Beaches	43.676357	-79.293031	Relish Bar & Grill	43.686280	-79.310980	Bar
77	The Beaches	43.676357	-79.293031	The Shore Leave	43.684200	-79.319413	Cocktail Bar
83	The Beaches	43.676357	-79.293031	Hitch Bar	43.663250	-79.330649	Bar
85	The Beaches	43.676357	-79.293031	Pinkerton Snack Bar	43.668900	-79.337309	Cocktail Bar
120	The Danforth West, Riverdale	43.679557	-79.352188	The Only Cafe	43.680409	-79.337898	Beer Bar
159	The Danforth West, Riverdale	43.679557	-79.352188	The Comrade	43.659346	-79.347932	Bar
161	The Danforth West, Riverdale	43.679557	-79.352188	Rooftop At Broadview Hotel	43.659109	-79.350074	Hotel Bar
180	The Danforth West, Riverdale	43.679557	-79.352188	Hitch Bar	43.663250	-79.330649	Bar
183	The Danforth West, Riverdale	43.679557	-79.352188	FUEL+	43.664399	-79.380427	Juice Bar
188	The Danforth West, Riverdale	43.679557	-79.352188	Greenhouse Juice Co	43.679101	-79.390686	Juice Bar

Table 2: Sample of Bars Locations in Toronto

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
138	The Danforth West, Riverdale	43.679557	-79.352188	The Broadview Hotel	43.659060	-79.350030	Hotel
161	The Danforth West, Riverdale	43.679557	-79.352188	Rooftop At Broadview Hotel	43.659109	-79.350074	Hotel Bar
292	The Beaches West, India Bazaar	43.668999	-79.315572	The Broadview Hotel	43.659060	-79.350030	Hotel
306	Studio District	43.659526	-79.340923	The Broadview Hotel	43.659060	-79.350030	Hotel
320	Studio District	43.659526	-79.340923	Rooftop At Broadview Hotel	43.659109	-79.350074	Hotel Bar
395	Studio District	43.659526	-79.340923	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel
826	Moore Park, Summerhill East	43.689574	-79.383160	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
921	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
1011	Rosedale	43.679563	-79.377529	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
1063	Rosedale	43.679563	-79.377529	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel

Table 3: Sample of Hotels Locations in Toronto

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
12	The Beaches	43.676357	-79.293031	The Remarkable Bean	43.672801	-79.287038	Coffee Shop
13	The Beaches	43.676357	-79.293031	Buds Coffee Bar	43.669375	-79.303218	Coffee Shop
49	The Beaches	43.676357	-79.293031	Press Books Coffee Vinyl	43.687672	-79.304457	Coffee Shop
76	The Beaches	43.676357	-79.293031	Pomarosa Coffee & Kitchen	43.683201	-79.325849	Coffee Shop
79	The Beaches	43.676357	-79.293031	Starbucks	43.682379	-79.327249	Coffee Shop
123	The Danforth West, Riverdale	43.679557	-79.352188	Hailed Coffee	43.666900	-79.345432	Coffee Shop
132	The Danforth West, Riverdale	43.679557	-79.352188	Merchants of Green Coffee	43.659916	-79.353963	Coffee Shop
147	The Danforth West, Riverdale	43.679557	-79.352188	Te Aro	43.661373	-79.338577	Coffee Shop
149	The Danforth West, Riverdale	43.679557	-79.352188	Dark Horse Espresso Bar	43.658498	-79.352356	Coffee Shop
172	The Danforth West, Riverdale	43.679557	-79.352188	Rooster Coffee House	43.669654	-79.379871	Coffee Shop

Table 4: Sample of Coffee Shops Locations in Toronto

Having extracted clusters, descriptive statistics were used to determine the radius of each cluster and therefore the area I which the venues of interest will be served by the specific warehouse. Moreover, to examine the result, the density of venues in each area was calculated. Mean values were used to determine the exact location of each warehouse.

Results

In figure 1 we present the map of Toronto. The locations of interest are marked on the map with different colors according to their category:
Green – Hotels
Blue – Coffee Shops
Red – Restaurants
Black – Bars

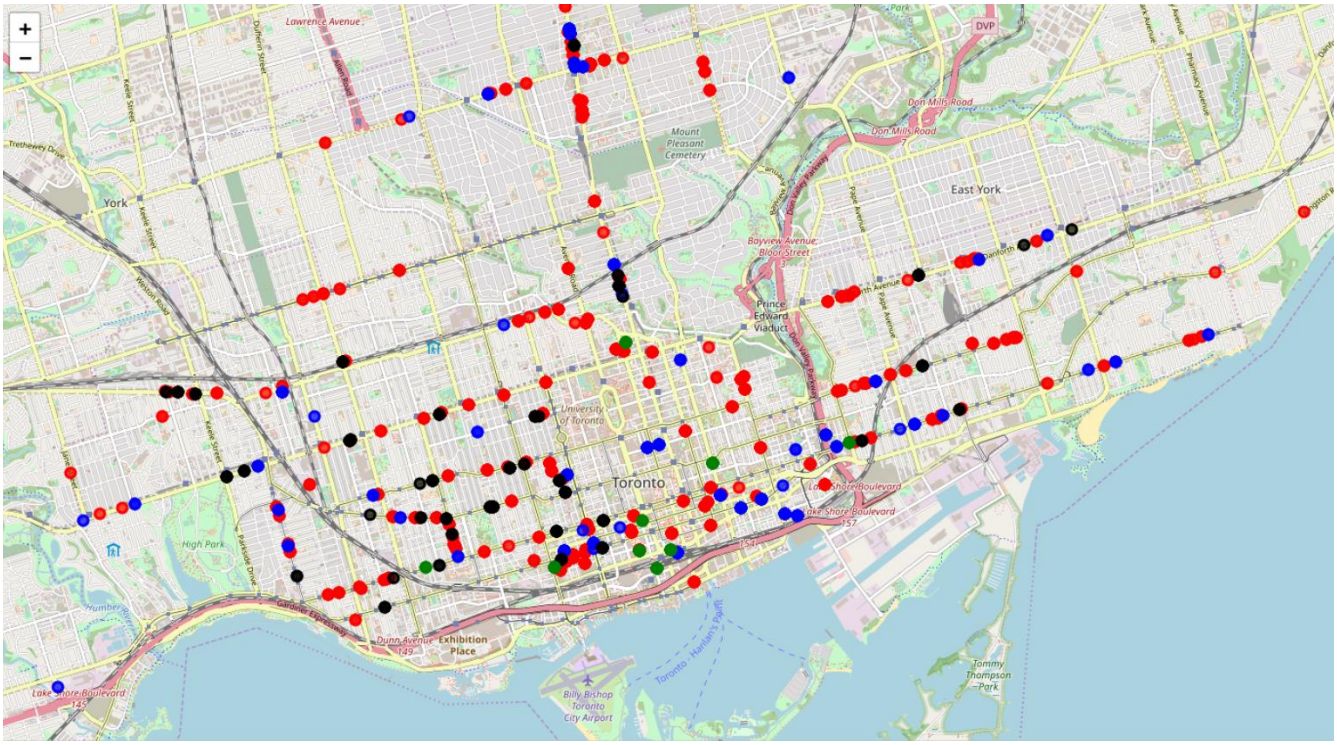

Figure 1: Locations to which Company Delivers its Product

All of the above points were used in the k-mean algorithm. The resulted clusters are shown in Figure 2.

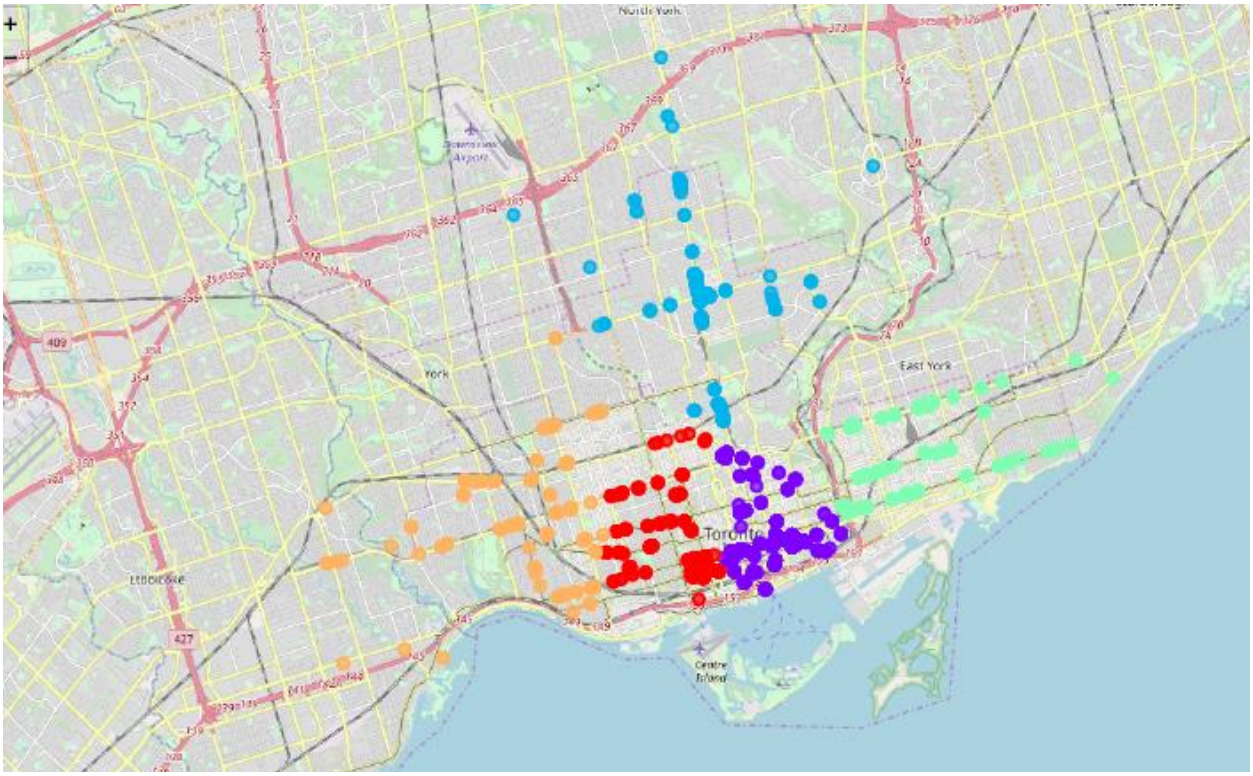

Figure 2: The 5 Clusters of the Venues Indicated by Different Colors.

The mean radius was calculated for each cluster. The location of the warehouse was defined by the cluster centroids in the k-mean algorithm. Moreover, the number of venues in each cluster was extracted and used to define the density in each area. The results are shown in the table below.

Cluster	Location Center (latitude)	Location Center (longitude)	Mean Radius (km)	Area (km^2)	Number of Venues	Density
Red	43.651	-79.406	3.773	44.729	367	8.204
Magenta	43.654	-79.378	3.575	40.155	502	12.501
Green	43.670	-79.334	5.430	92.624	191	2.062
Blue	43.707	-79.398	4.687	57.246	227	3.965
Orange	43.658	-79.448	6.188	120.314	176	1.463

In figure 3 we present the area of each cluster overlaid in the map of Toronto.

Figure 3: Clusters Area

Discussion

From the results, we can see the variety in the distribution of venues in Toronto city. As seen the two central clusters in the middle of the city contain at least a double number of hotels, restaurants, bars and coffee shops than three other clusters. For this reason, their areas are smaller than those of the other three clusters. Based on this observation we suggest:

- The size of the warehouse at the red and purple areas be at least double than the size of the other warehouses
- The number of employees at the warehouse at the red and purple areas should be at least double than the in the other warehouses
- The number of tracks in each warehouse should correlate to the number of venues in each cluster

Conclusion

The main problem in this work is the definition of the locations where 5 warehouses can be built to improve the process of bottled water distribution in the city of Toronto. Our approach was to devise Toronto in 5 different subregions. The criterion used for this division was the density of venues of customer venues in the entire city. The center of each subregion was estimated using the k-means algorithm. The standard deviation of each cluster was used to define the area containing the venues to which the warehouse must provide its products. Based on the results, we are able to define the location of the warehouses, their size, number of employees in each warehouse and the number of trucks required.