Coursera Capstone Project
Week 5

Introduction/Business Problem

Our client, a bottled water brand, has an established role in th market of Toronto. He is the first supplier of bottled water with more than 1500 clients in the neighborhoods of Toronto. His main clients are hotels, coffee shops, restaurants and bars. Currently, the products are stored in a big central warehouse outside Toronto and distributed in the different venues in a daily basis. The main problem of this business plan is that the distribution of the product becomes time consuming, inefficient, and costly. Our client wants to reduces the costs of distribution by building 5 smaller warehouse in Toronto to serve locally his clients. This approach will reduce the time which is required to serve his clients, the fuel costs, and will also transform his business to a green chain.

To do so, our client asked from us to find the 5 best locations in Toronto at which the warehouses must be build in order to create smaller distribution clusters. The warehouses locations must be at the centers of these clusters in order to minimize the relative distance from each of the venues.

Data

The data required for this task are the locations of the hotels, coffee shops, bars and restaurants in Toronto. To gather the data we ‘ll use the locations of all neighborhoods in Toronto gathered from Wikipedia. Based on these location we ll gather the locations of all venues in these neighborhoods from Foursquare. We will filter the data to acquire the locations of the targeted venues. In order to inspect the data, we ll 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 on the location features to define the 5 clusters of venues. 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 ll use the folium library.

Methodology

In order to define the location of each warehouse, we choose to use the kmeans algorithm. By default, this algorithm minimize the distance of each point in a cluster from the centroid of 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 the 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 the were gathered from the Foursquared database in Figure 1, 2, 3, 4. A total of 258 coffee shops, 909 restaurants, 178 bars and 73 hotels were found and used in the k-means algorithm.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
46	The Beaches	43.676357	-79.293031	Relish Bar & Grill	43.686280	-79.310980	Bar
82	The Beaches	43.676357	-79.293031	Pinkerton Snack Bar	43.668900	-79.337309	Cocktail Bar
88	The Beaches	43.676357	-79.293031	The Only Cafe	43.680409	-79.337898	Beer Bar
89	The Beaches	43.676357	-79.293031	Hitch Bar	43.663250	-79.330649	Bar
92	The Beaches	43.676357	-79.293031	The Shore Leave	43.684259	-79.319474	Cocktail Bar
118	The Danforth West, Riverdale	43.679557	-79.352188	The Only Cafe	43.680409	-79.337898	Beer Bar
128	The Danforth West, Riverdale	43.679557	-79.352188	Pinkerton Snack Bar	43.668900	-79.337309	Cocktail Bar
173	The Danforth West, Riverdale	43.679557	-79.352188	The Comrade	43.659346	-79.347932	Bar
193	The Danforth West, Riverdale	43.679557	-79.352188	Greenhouse Juice Co	43.679101	-79.390686	Juice Bar
236	The Beaches West, India Bazaar	43.668999	-79.315572	Pinkerton Snack Bar	43.668900	-79.337309	Cocktail Bar

Figure 1: Locations of Bars in Toronto (sample).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
8	The Beaches	43.676357	-79.293031	The Remarkable Bean	43.672801	-79.287038	Coffee Shop
11	The Beaches	43.676357	-79.293031	Buds Coffee Bar	43.669375	-79.303218	Coffee Shop
54	The Beaches	43.676357	-79.293031	Circus Coffee House	43.685483	-79.315364	Coffee Shop
68	The Beaches	43.676357	-79.293031	Starbucks	43.668370	-79.308015	Coffee Shop
75	The Beaches	43.676357	-79.293031	Starbucks	43.682446	-79.327232	Coffee Shop
120	The Danforth West, Riverdale	43.679557	-79.352188	Hailed Coffee	43.666900	-79.345432	Coffee Shop
140	The Danforth West, Riverdale	43.679557	-79.352188	Merchants of Green Coffee	43.659986	-79.354299	Coffee Shop
150	The Danforth West, Riverdale	43.679557	-79.352188	Te Aro	43.661373	-79.338577	Coffee Shop
154	The Danforth West, Riverdale	43.679557	-79.352188	Dark Horse Espresso Bar	43.658498	-79.352356	Coffee Shop
169	The Danforth West, Riverdale	43.679557	-79.352188	Sumach Espresso	43.658135	-79.359515	Coffee Shop

Figure 2: Locations of Coffee Shops in Toronto

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
0	The Beaches	43.676357	-79.293031	Tori's Bakeshop	43.672114	-79.290331	Vegetarian / Vegan Restaurant
15	The Beaches	43.676357	-79.293031	Delina Restaurant	43.668867	-79.305404	Middle Eastern Restaurant
19	The Beaches	43.676357	-79.293031	Veloute Bistro	43.672267	-79.289584	French Restaurant
22	The Beaches	43.676357	-79.293031	Budapest Restaurant	43.680946	-79.310110	Hungarian Restaurant
25	The Beaches	43.676357	-79.293031	Jatujak	43.688421	-79.270073	Thai Restaurant
26	The Beaches	43.676357	-79.293031	Udupi Palace	43.672480	-79.321275	Indian Restaurant
30	The Beaches	43.676357	-79.293031	Lake Inez	43.672520	-79.320712	Asian Restaurant
40	The Beaches	43.676357	-79.293031	The Wren	43.682467	-79.328079	American Restaurant
44	The Beaches	43.676357	-79.293031	Melanie's Bistro	43.684800	-79.317167	French Restaurant
47	The Beaches	43.676357	-79.293031	Maha's Fine Egyptian Cuisine	43.671758	-79.328444	African Restaurant

Figure 3: Location of Restaurant in Toronto (sample).

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Latitude	Longitude	Category
145	The Danforth West, Riverdale	43.679557	-79.352188	The Broadview Hotel	43.659072	-79.350074	Hotel
293	The Beaches West, India Bazaar	43.668999	-79.315572	The Broadview Hotel	43.659072	-79.350074	Hotel
307	Studio District	43.659526	-79.340923	The Broadview Hotel	43.659072	-79.350074	Hotel
399	Studio District	43.659526	-79.340923	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel
830	Moore Park, Summerhill East	43.689574	-79.383160	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
929	Deer Park, Forest Hill SE, Rathnelly, South Hi...	43.686412	-79.400049	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
1012	Rosedale	43.679563	-79.377529	Four Seasons Hotel Toronto	43.671796	-79.389457	Hotel
1066	Rosedale	43.679563	-79.377529	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel
1091	Rosedale	43.679563	-79.377529	The Broadview Hotel	43.659072	-79.350074	Hotel
1118	Cabbagetown, St. James Town	43.667967	-79.367675	The Grand Hotel & Suites Toronto	43.656449	-79.374110	Hotel

Figure 4: Locations of Hotels in Toronto

Having extracted the clusters, descriptive statistics were used to determine the radius of each cluster and therefore the area in which the venues of interest will be served by the specific warehouse. Morover, in order 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 5 we present the map of Toronto. The locations of the venues of interest are marked on the map with different colors according to their category.

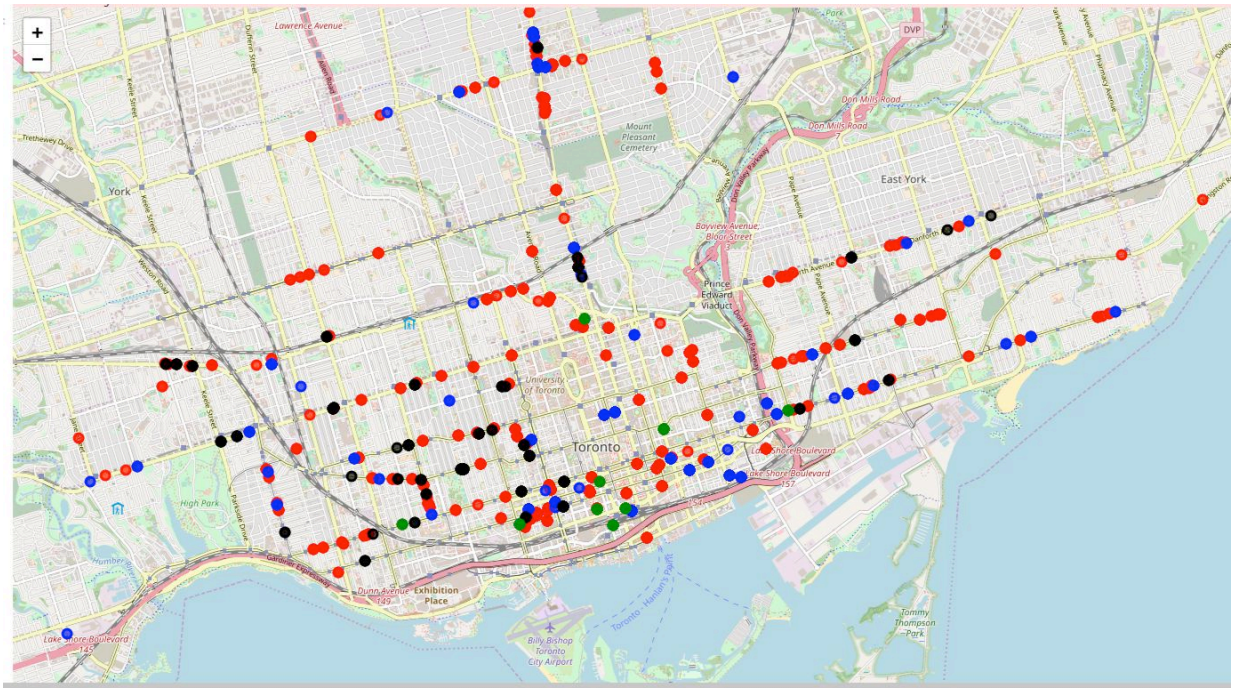

Figure 5: Hotels (green points), Coffee shops (blue points), Restaurants (red points) and Bars (black points) in Toronto

All of the above points were used in the k-mean algorithm. The resulted clusters are shown in Figure 6.

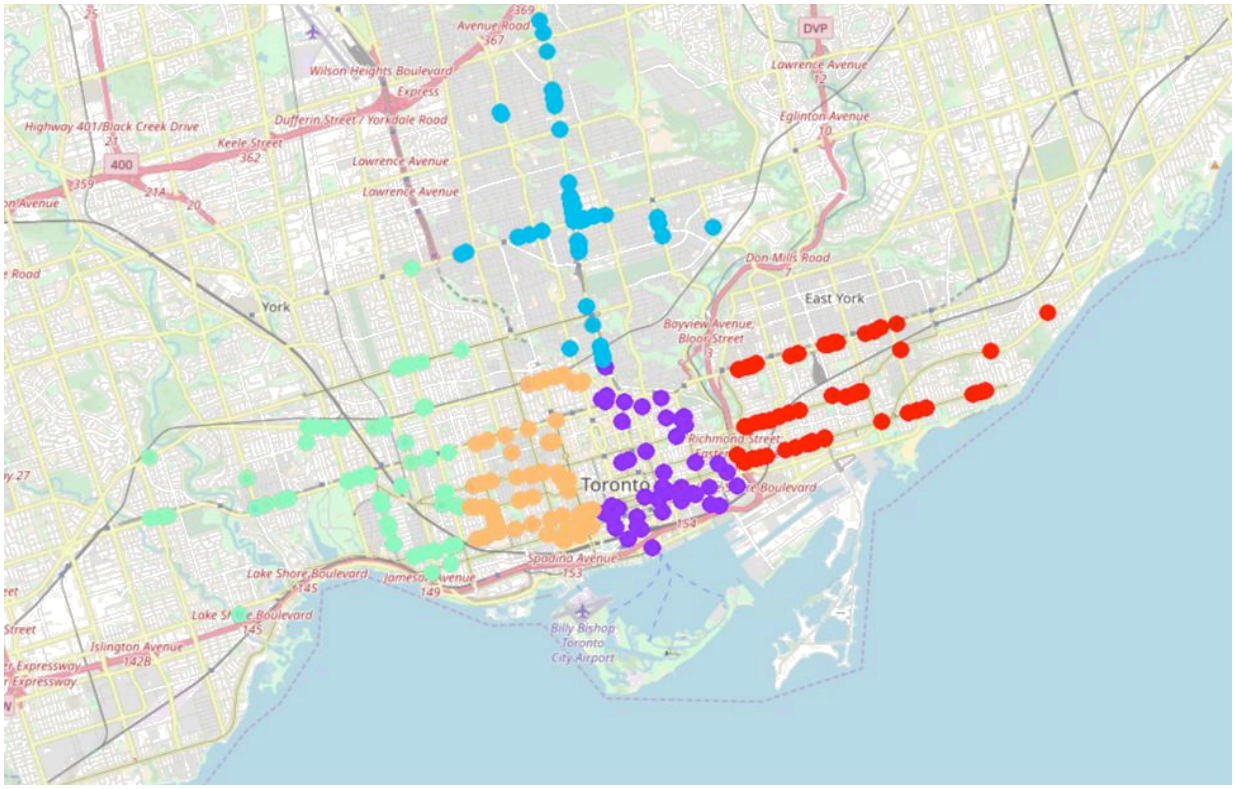

Figure 6: The 5 clusters of the venues indicated by the different colors. The clusters were extracted with the k-means algorithm.

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

Cluster	Location-Centre (latitude)	Location-Centre (longitude)	Mean Radius (km)	Area (km^2)	Number of Venues	Density (#/km^2)
Red	43.66962	-79.33436	5.24	86.27	182	2.1
Magenta	43.65410	-79.37749	3.49	38.42	425	11.1
Light blue	43.70438	-79.39764	4.18	54.99	231	4.2
Light green	43.65665	-79.44908	5.87	108.33	166	1.5
Orange	43.65278	-79.40520	3.81	45.7	414	9.1

In Figure 7 were present the area of each cluster overlayed on the map of Toronto

Figure 7: Clusters Area. Each warehouse will be responsible for the venues lying at the area indicated by the corresponding circle.

Discussion

From the results we can see the variability in the distribution of venues in the 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 the three other cluster. 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 placed at the orange and purple areas be at least double than the size of the other warehouse.
- Similarly, th employees employed in the central warehouses be at leas double than those employed in the other three warehouse,
- the number of trucks in each warehouse to chosen regarding the number of venues in each cluster.

Conclusions

The main problem solved in this work is the definition of the locations where 5 warehouses can be build to improve the process of bottled water distribution in the Toronto city. Our approach was to divide the Toronto city in 5 different subregions. The criterion used for this division was the distribution (density) of the venues of interest 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 the products. Based on the results, we are able to define the location of the warehouses, their size, the number of employees in each warehouse and the number of trucks required.