Introduction/Business

A bottling company has an established role in the market of Toronto. They are the premier supplier of bottled water with over 1000 clients in Toronto’s neighborhoods. His best clients consist of the following types of businesses: restaurants and bars, hotels and coffee shops. The bottled water is currently stored in a central warehouse outside of the city and is distributed to the clientele daily. The main problem with this business plan is that the distribution becomes is inefficient and becomes time consuming and costly.

Our client seeks to reduce the distribution costs by building 5 smaller warehouses in Toronto to serve as smaller hubs and distribute to his clients from those points since it’d be closer, eliminating the time and resource costs caused by the distance of the original warehouse from his clients. In order to accomplish this, our client has asked us to find the 5 best locations in the city of Toronto where the warehouses could be constructed in order to create smaller distribution clusters. The warehouses must form the centers of the clusters in order to minimize their relative distances from each client.

Data

The data required for this task are the locations of the hotels, coffee shops, bars and restaurants in Toronto. In order to gather the required data, we will gather the locations of Toronto neighborhoods from Wikipedia. With this data, we will pinpoint the locations of these neighborhoods on Foursquare. The data will then be filtered in order to acquire the locations of the targeted venues needed for our study. We will utilize Folium to build a map of Toronto and visualize the delivery venues. Following this, the k-means algorithm will be used to define the centroids of the five clusters which will serve as the best possible points at which distribution warehouses can be built.

Methodology

In order to define the location of each warehouse, we choose to use the k-means 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-means 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 they were gathered from the Foursquare 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.



*Fig 1. Sample of bar locations in Toronto*

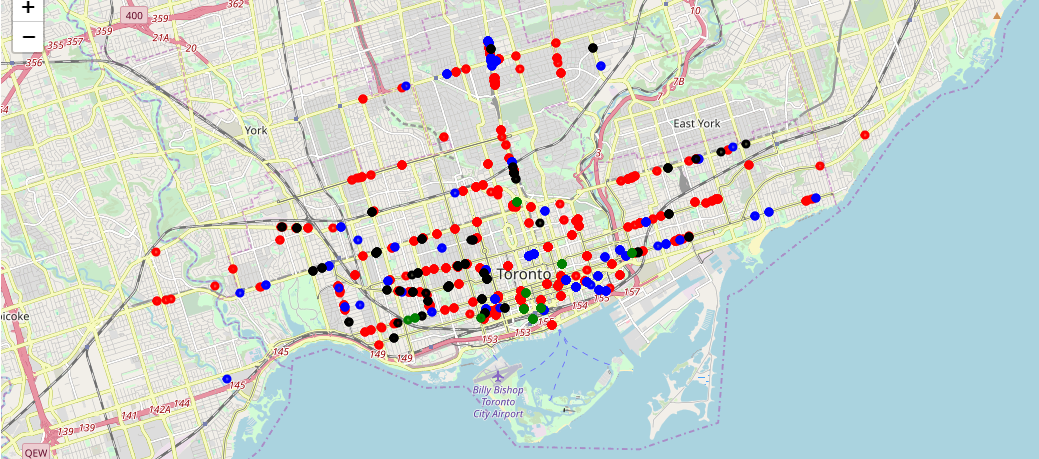


*Figure 2. Sample of hotel locations in Toronto*

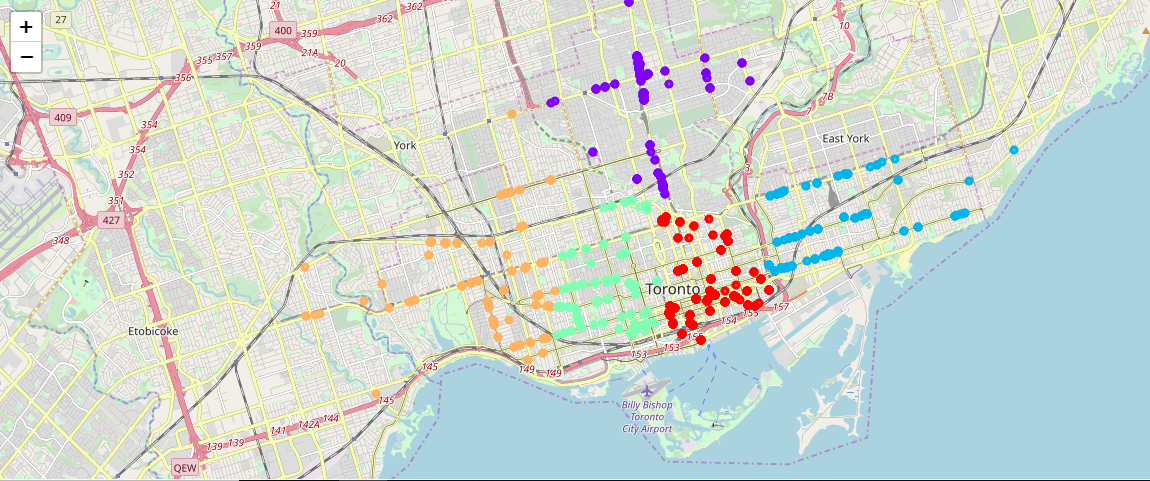
  
*Figure 3. Sample of coffee shop locations in Toronto*

  
*Figure 4. Sample of restaurant locations in Toronto*

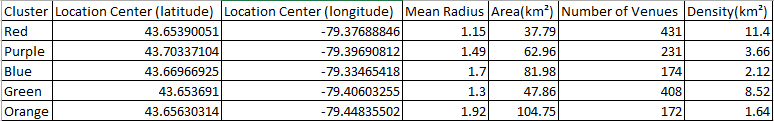
After I extracted the clusters, the radius of each cluster was determined using descriptive statistics and consequently, the approximate areas by where venues will be served by each warehouse. Also, the density of the venues in each cluster was calculated. Mean values were used to determine the best precise locations for the warehouses to be built.

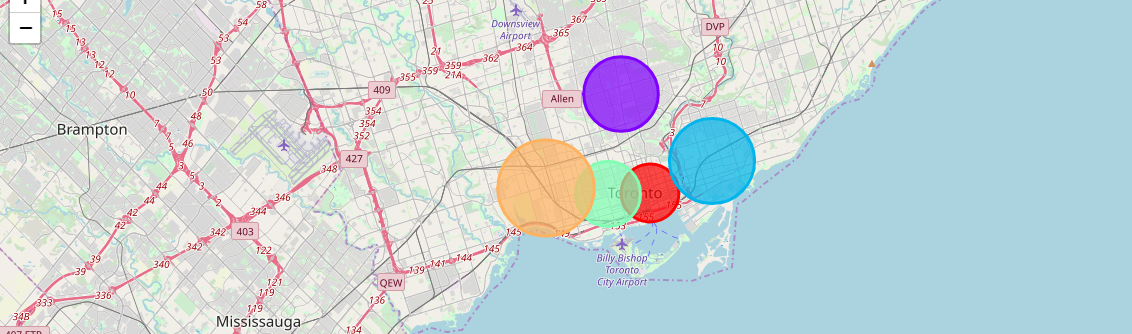
 *Figure 5. Bars(black), Hotels(green), Restaurants(red), Coffee Shops(blue)*

The above points were all used in the k-Means algorithm with the resultant clusters shown in Fig 6.

  
*Figure 6: The venue clusters are indicated by different colors, assigned with colormap. The clusters were created using the k-Means algorithm.*

The mean radius was calculated for each cluster and the locations of the warehouses defined by the clusters’ centroids which resulted from the k-means algorithm. The number of venues in each cluster has been extracted and used in order to define each area’s density. The results are shown in the following table, and in Figure 7, the area of each cluster is presented.



  
*Fig. 7: Area of the Clusters. Each warehouse will be responsible for distribution to the venues within the areas highlighted by the circles.*

Discussion

From our results we can see the variability in the distribution of different venues in Toronto. As can be seen, the two central clusters(green and red) in the city center contain at least twice as many venues as the remaining three clusters. Based on this observation, I suggest that the warehouses to be built in the green and red areas be much larger than the remaining three warehouses, as they will need the extra storage space to serve the bulk of venues. On that same thread, there should likewise be far more people employed at the warehouses serving red and green areas to help meet demand in a productive fashion. It also follows that the warehouses in those two areas should also have more delivery trucks than those from the remaining three clusters.

Conclusion

We have solved the problem of defining the locations for the five distribution warehouses to be built for the purpose of improving bottled water distribution in Toronto’s most populated city area. The approach was to divide the city into five subregions via Clustering. We used the distribution, ie, the density of the venues in the city to accomplish this. The centers of these subregion clusters were generated by using the k-means algorithm. The standard deviation of the clusters were used to generate and define the areas which contained the venues of interest to which the warehouses must distribute its products. Based on our results, we have successfully defined the warehouse locations, their size, numbers to be employed and number of distribution trucks required.