**Capstone Project Report**

**Introduction/Business Problem**

**DDC- A US based International Coffee Shop/Cafe chain wants to expand in Canada and has decided that Toronto would be a good city to start. But they were not sure as to which location would be most advantageous for them. Hence they called a Data Scientist and wanted a region-based report regarding the inclinations of the people living in different regions of Toronto. They want to leverage Foursquare location data to predict how like people living in a given area would be inclined to spend money in a cafe. This will help them pick a target region to open their first branch in Canada.**

**Data**

First set of Data is been scrapped from the **Wikipedia page** of Toronto. We have gathered information such as Postal code, Boroughs’ Data and corresponding Neighbourhood data. These regions will act as a data set upon which the models will be based. Furthermore, longitude and latitude data was also gathered for these regions in order to make visualisation clearer.

Second set of data was taken from Foursquare.com. It provided the data on the preference of the people living in the region or near the region (that were input from the previous set).

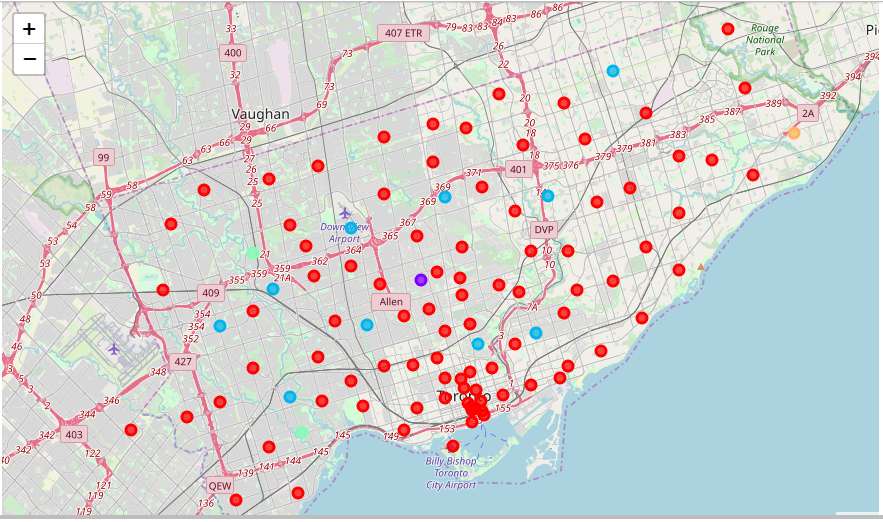
Since it is evident that segmentation among the people is needed (based on their preference) and since the nature of this segmentation is unsupervised; K mean clustering is the model selected to solve the problem

**Model**

K mean clustering is an unsupervised data clustering model used to segment data points into clusters/segments where data points in the same cluster are similar to each other in a certain way and different to those of other clusters.

Here, data collected from Foursquare.com acts as a decider upon which different regions get certain scores based on the behaviour of its residents. Based on these scores, the regions are divided into clusters where each cluster has a significantly close score.

Here, the regions are divided into 5 clusters and top 10 recreational preferences of people in that region of each region were listed.



The clusters were represented on a map for better representation and data visualisation.

**Results**

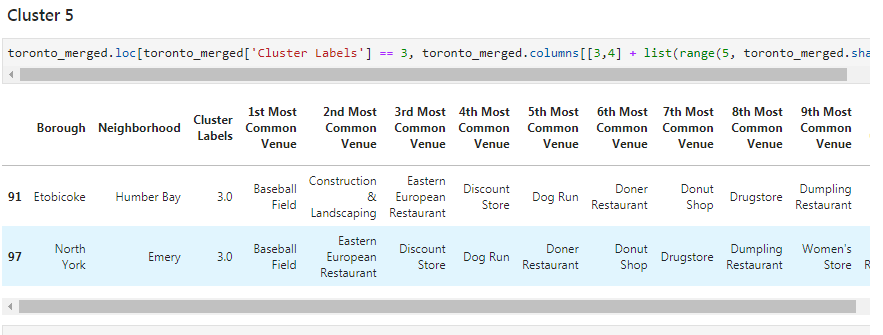
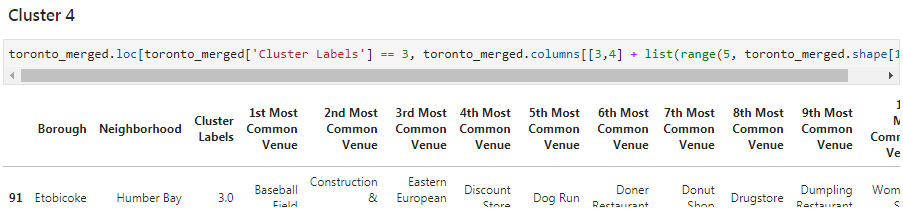
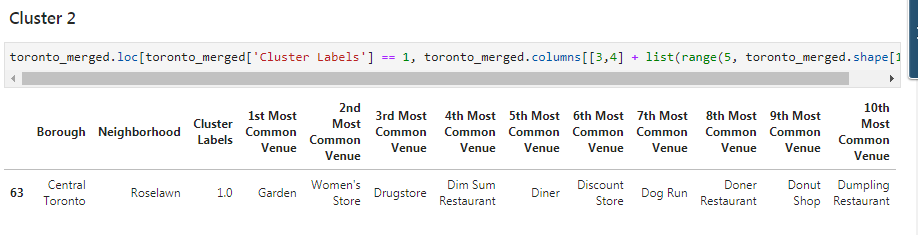
The tables of each cluster showed specific boroughs, neighbourhoods in those boroughs and top 10 recreational preferences of residents of those areas. These preferences include Park, Cafe, Coffee Shop, Bank etc.

**Discussion**

According to the data, the region whose people have Cafe/ Coffee Shop as their 1st or 2nd Preference would be most favourable for the company.

As evident from the data, neighbourhoods in Downtown Toronto borough consistently have Coffee shop/Cafe as either first or second preference. Hence Downtown Toronto will be an ideal region to build a branch.

Additionally the cluster that Downtown Toronto borough is the part of (Cluster 0) has the maximum number of neighbourhoods within it. This shows that many other neighbourhoods are very similar to Downtown Toronto. This shows a great possibility of expansion in those regions in near future.



**Conclusion**

As per the K mean clustering model “Downtown Toronto” would be the ideal region to open the first branch of DDC brand.