Capstone Project Notebook

This notebook contains the final capstone project for the IBM Data Science Professionals course on Coursera

1.1 Introduction/ Business Problem

One of the greatest challenges to someone who wants to start their own cafe is to decide its location. A lot of factors play into this decision, but one of the most important elements to consider is the placement of its competition. Ideally a proprietor wouldn't find it in his/her best interests to open cafe in an area which already has a several cafes because that will make it harder for them to compete for customers. However, on the other hand it would be futile for the cafe to open in a area completely devoid of other cafes as it is an indication of either low population of that area or low public interest in cafes.

Through analyzing data regarding number of cafes in a particular area, I hope to find an ideal location where it would be most profitable for the stakeholder to start a cafe restaurant. The data will be sourced from Foursquare.

1.2 Data

The requirements are as follows:

1. Cafes/coffee shops in the city of interest (Toronto)
2. Location of the cafe in the form of its Latitude and Longitude coordinates.

The data wil be sourced from Foursquare by making appropriate API calls. The data will be cleaned, preprocessed and made to go through KMeans clustering algorithm.

1.3 Methodology

The primary data that was required for my data analysis was details regarding coffee shops in Toronto. The details were gathered by making an API call to Foursquare services, specifically the 'search' call. In the call, information like latitude and longitude of the target area, i.e Toronto, and the query, which was 'coffee', was passed. The get request containing the api call returned a json file with all the relevant data. The data was then pre-processed and cleaned to create a pandas dataframe out of it. The dataframe contained the following feature sets:

1. Name of the coffee shop
2. Latitude co-ordinate
3. Longitude co-ordinate

The latitude and longitude subset of the dataframe were fitted using KNN clustering algorithm. The number of clusters was set to 5. This created 5 labelled clusters and the labels were added to the dataframe.

Using matplotlib rainbow libraries, different colors were assigned to each cluster to help in visualization.

Using Folium, a map of the Toronto area was generated and markers for each coffee shop was placed on it. The marker was colored with the specific color associated with their cluster to help visualize the clustering.

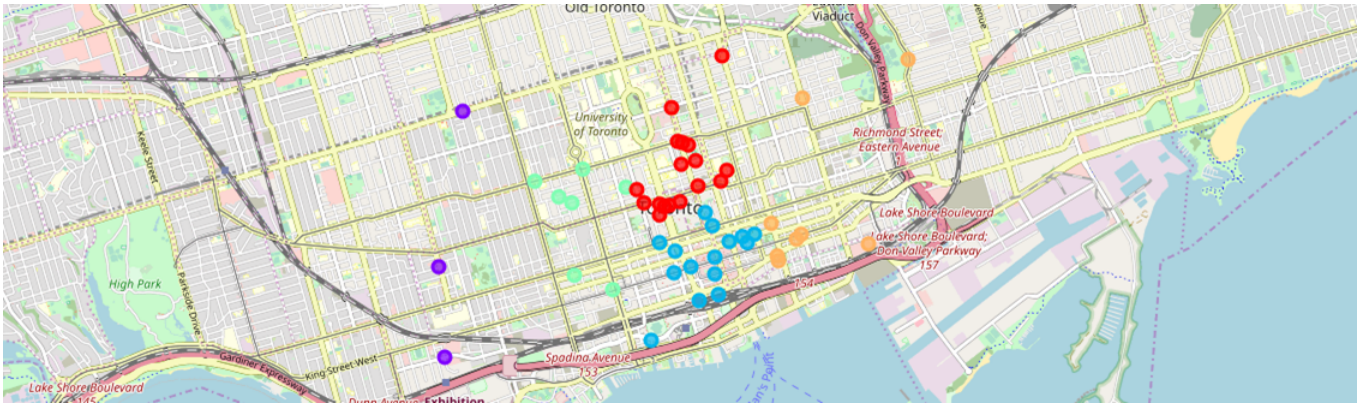
1.4 Results

As can be clearly seen on the map, the coffee shops are clearly divided into five clusters. This can give key insights to the stakeholder as to where they can best place their coffee shop.

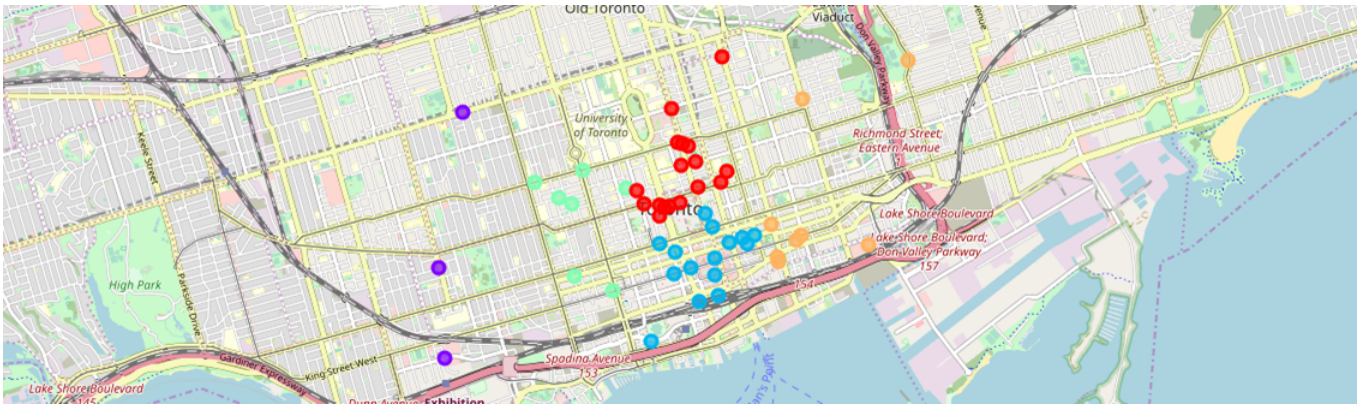
Some areas are packed densely with coffee shops while some are relatively dense. Now it is up to the stakeholder to study the map and decide what is in his/her best interest, based on the business model.

They may choose to build their shop in an area with not many shops around to eliminate competition. On the other hand the shop may be built in an area with a large number of competing shops as it is generally an indication of a large target customer base existing there.

1.4.1 Clustering of coffee shops in Toronto:



1.4.2 Clustering of neighborhoods in Toronto



Using these two maps we can make observations about which neighborhoods have what concentration of coffee shops.
**For example we find out that the areas in and around Kingsway have very few cafes so that might be a good location for the stakeholder to build his/her own shop. **

1.5 Discussion

Further research should be done regarding the observations that were reported from the data analysis. The visual representation of the map of Toronto gives us key insights as to the placement of coffee shops. Diving deeper, we need to find out the context behind these insights such as:

1. **If an area is densely packed with coffee shops, what is the reason. Is it because of a large population there who is willing to pay for coffee,or is it something else?**
2. **Conversely, if an area doesn't have many cafes what is the reason? Is it because of a low population there or is it something else?**
3. **What all factors influence the density of a partiular cluster of coffee shops? For example, its proximity to a college or office campus?**

1.6 Conclusion

In conclusion, the location data of coffee shops across Toronto was taken and plotted on a map where they were clustered using KNN algorithm to show different clusters of shops. The data analysis should help stakeholders make decisions about placement of a new coffee shop to maximize profits.