**BATTLE OF NEIGHBOURHOODS**

Akshat Singhal

June 26, 2019

1. **Introduction**
   1. **Background**

A chain of restaurant owners in Ontario, Canada want to expand their business. Currently they have their restaurants open in cities like Ottawa, Brampton and Hamilton. They figured out that they would make more profit by opening up a restaurant in Toronto as Toronto is the largest city of Canada. So they want to open up a new restaurant some place nice with good neighbourhood in Toronto.

* 1. **Problem**

They are having trouble figuring out which place to choose within Toronto to open their new restaurant. We have to help them figure out which place to choose where their business will be good, they have less competition and nice people live around. They want to know about 3-4 such places so that they can decide for themselves which one is the best according to the type of their restaurant.

* 1. **Interest**

Obviously, people in the business of restaurant chains, hotels, etc. who are willing to expand their business in new cities would be very interested for competitive advantage and business values. Others who are new to this business might also be interested.

1. **Data Acquisition and cleaning**
   1. **Data Sources**

There were two main datasets that were used for this project.

**First Dataset: List of neighbourhoods in Toronto**

Firstly, data from a Wikipedia page was used which provides information about list of neighbourhoods in Toronto, Canada. Then a web scrapping tool named BeautifulSoup was used for extracting the data in the form of a table from this Wikipedia page. This table contains 3 columns: Postal Code, Borough and Neighbourhood. The link for this Wikipedia page: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> . After pre-processing the table and adding two more columns of Latitude and Longitude of each Neighbourhood, this dataset was ready for use. Final Data Frame will have 5 columns: Postal Code, Borough, Neighbourhood, Latitude, Longitude. And it will contain 103 rows having 103 unique neighbourhoods of Toronto and 5 unique Boroughs. For example, below photo depicts first 5 rows of the dataset:



**Second Dataset: List of different venues in the neighbourhoods of Toronto:**

This dataset will be formed using the Foursquare API. I will use the Foursquare location data to explore different venues in each neighbourhood of Toronto.

These venues can be any place. For example: Parks, Coffee Shops, Hotels, Gyms, etc.

Using the Foursquare location data, information about these venues can be taken and the neighbourhoods of Toronto can be easily analysed based on this information.

We will use the geographical coordinates from above dataset to generate this Location dataset. This dataset is named **toronto\_venues.**



For example, the neighbourhood named Parkwoods contains 3 nearby venues depicted by first 3 rows of above dataset. Information about these venues is also provided in this dataset.

* 1. **Data Pre-processing**

After the 2 datasets were obtained, pre-processing of these two datasets was needed so that they can be used for clustering algorithms easily. A data frame named **venues\_sorted** was created which list of neighbourhoods of Toronto along with their respective 5 most common venues. This dataset would eventually help in visualising the solution.



1. **Methodology**

We have a dataset named **toronto\_onehot** that is pre-processed and through one-hot encoding, it is ready to be used for clustering technique. But this dataset contains information about all the nearby venues like Park, Gym, Shops, etc.

As we are interested in venues in 'food' category, but only those that are proper restaurants. Therefore, coffee shops, pizza places, bakeries etc. are not direct competitors, so we don't care about those. Hence we will include in our list only venues that have 'restaurant' in category name, and we'll make sure to detect and include all the subcategories of different restaurants in the neighbourhood. For example, Afghan restaurant, Italian restaurant, etc. For this, we locate venues from **toronto\_onehot** data frame that are restaurants only and store this in a new data frame named **toronto\_restaurants**.

I also added a column containing total number of restaurants in that neighbourhood in **toronto\_restaurants** data frame. This will help us in making clusters using K-Means clustering algorithm.

Now we use K-Means clustering algorithm to make clusters of dataset so that our analysis of the neighbourhoods is easy. For this we set number of clusters to be 5. The input for this clustering algorithm was **toronto\_restaurants** data frame.

After the clusters were made, we merged the first dataset and the **venues\_sorted** data frame and inserted cluster labels also. The result data frame was named **toronto\_merged** which looked like this:



Next part was Analysis of each cluster to get the correct neighbourhood. I calculated Total number of neighbourhoods and Total number of restaurants for each cluster. Then I calculated Restaurant/Neighbourhood ratio and found that this ratio was lowest for cluster with cluster label=4. Hence this cluster was chosen for further analysis.

Cluster 4 contained total 8 neighbourhoods. Out of these, 2 had very high total number of restaurants, therefore these 2 neighbourhoods were discarded. Out of remaining 6, 2 more were discarded because they had Restaurant as their most common venue more than once and hence these neighbourhoods were not suitable for Restaurant business and hence discarded.

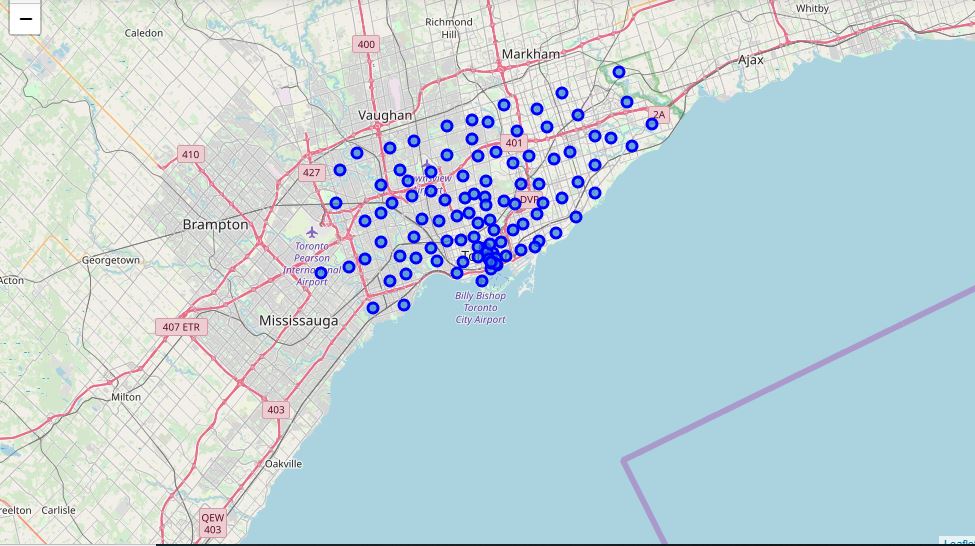
The final dataset contains all the information about these remaining 4 neighbourhoods:



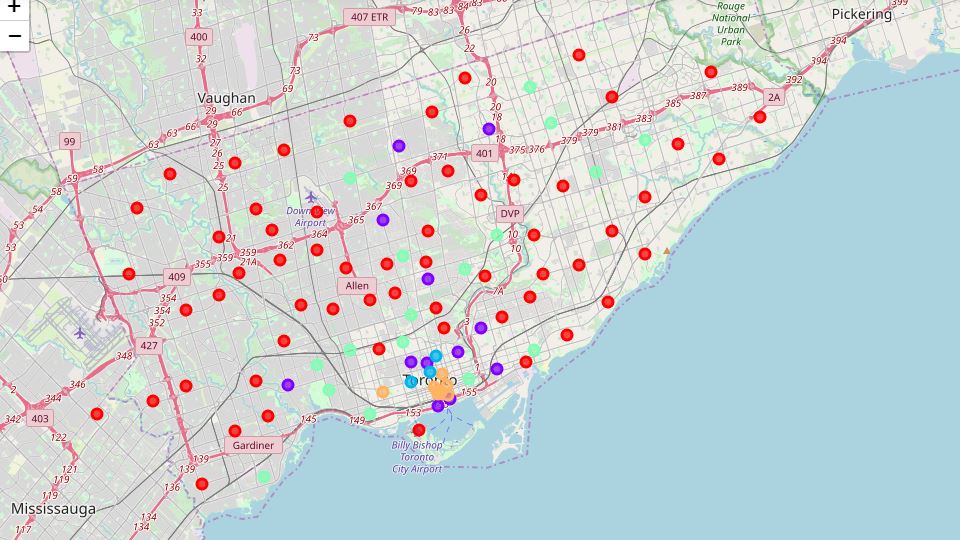
The owners can further choose from these 4 locations which will be the best according to the type of restaurant they are trying to open.

1. **Data Visualisation**

A map of Toronto city was generated using a great visualisation library named **Folium**. All the 103 neighbourhoods of Toronto were also marked with blue circles on the map with help of first dataset. The map looked like this:

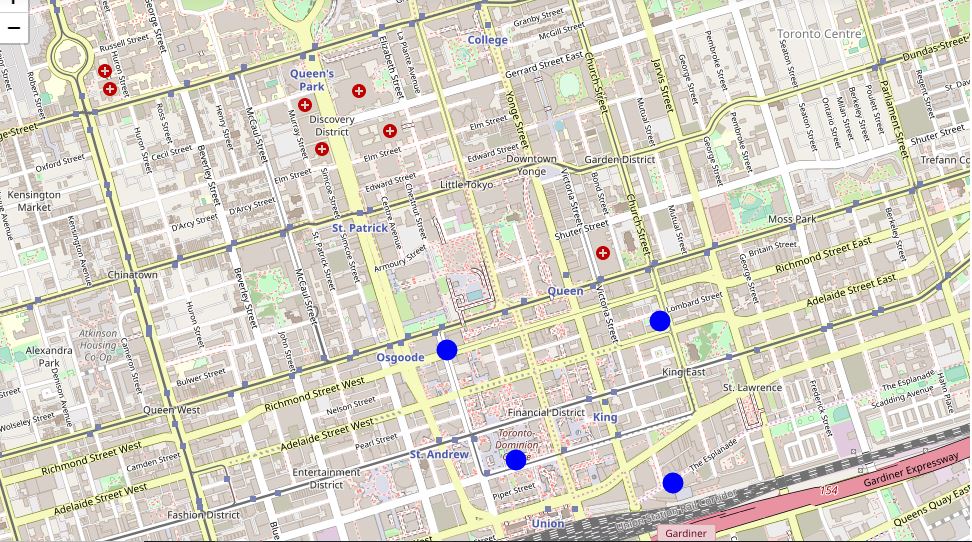


After using the clustering algorithm and creating 5 different clusters where each neighbourhood belong to one of these clusters, the new map of Toronto looked like this:



In the above map, 5 different colours, one for each cluster are used for representing each neighbourhood in Toronto.

The final 4 neighbourhoods were also presented on a map:



The 4 neighbourhoods are depicted by 4 blue dots in the above map.

1. **Result and Discussion**

Our analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density fairly close to city centre. To identify these pockets, we used clustering algorithm and segmented our neighbourhood dataset accordingly.

We used K-means clustering algorithm for making 5 clusters each containing some neighbourhoods based on number of restaurants they have in their vicinity. Then we analysed each cluster by calculating Restaurant/Neighbourhood ratio of each cluster. We saw that cluster 4 had lowest ratio, which means very few restaurants are present within vicinity of each neighbourhood. There were total 6 neighbourhoods belonging to cluster 4. Then upon further analysis, we found that 2 among those were not good for opening up a new restaurant. Hence, only 4 neighbourhoods left.

According to our analysis, we got a total of 4 neighbourhoods where restaurant business will be good. There are two reasons for that. First reason is, we saw that these neighbourhoods does not contain much restaurants around their vicinity which will lower the competition in the restaurant business. Second reason is that, as we can see in the above map that these 4 neighbourhoods lie in the centre of Toronto which means these neighbourhoods have high population density which means more customers and hence more profit.

The final 4 neighbourhoods that are perfect for opening a new restaurant are stored in a data frame named final which contains information about latitude, longitude and borough of these neighbourhoods.

The owners can further choose from these 4 locations which will be the best according to the type of restaurant they are trying to open.

1. **Conclusion**

Purpose of this project was to identify neighbourhoods in Toronto low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for a new restaurant. By calculating restaurant density distribution from Foursquare data we have first identified the most common nearby venues of each neighbourhood. Then with the help of clustering techniques and further analysis we were able to narrow down to 4 neighbourhoods which were good for opening up a new restaurant. This concludes this project of Battle of neighbourhoods.