BATTLE OF NEIGHBOURHOODS

Surendra kumar kumawat
August 28, 2021

1. Introduction

1.1 Background

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

1.2 Problem

As Toronto is a very large city, they are having trouble figuring out which place to choose within Toronto for 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 for them according to the type of their restaurant.

1.3 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 in my project for competitive advantage and business values. Others who are new to this business and want to set up their business in a new city might also be interested.

2. Data Acquisition and cleaning

2.1 Data Sources

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

First Dataset: List of all the neighborhoods in Toronto

Firstly, I used data from a Wikipedia page which provides information about all the neighborhoods of Toronto, Canada. Then I used a web scrapping tool named

BeautifulSoup for extracting the data in the form of a CSV table from this Wikipedia page. This table consisted of 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 importing this table into a data frame, pre-processing this data frame, and adding two more columns of Latitude and Longitude of each Neighbourhood, this data frame was ready for use. The final data frame will have 5 columns: Postal Code, Borough, Neighbourhood, Latitude, Longitude. And it will contain 103 rows having 103 unique neighborhoods of Toronto and 10 unique Boroughs.

Second Dataset: List of different venues in the neighborhoods of Toronto:

This dataset will be formed using the Foursquare API. Foursquare is a website that provides any information about a particular venue. I used the Foursquare location data to explore different venues in each neighborhood 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 neighborhoods of Toronto can be easily analyzed based on this information.

I will use the geographical coordinates from the above dataset to generate this Location dataset. This dataset is named **toronto_venues**.

2.2 Data Pre-processing

After the 2 datasets were obtained, pre-processing of the second dataset was needed so that it can be used for clustering algorithms easily. I pre-processed the **toronto_venues** data frame using **one-hot encoding** tool. The pre-processed data was stored in a data frame named **toronto_onehot**.

Now, 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. which is not necessary. As we are only interested in venues in 'food' category, therefore venues like Park, Gym, playgrounds are discarded from the **toronto_onehot** data frame.

Also, we are looking for only those venues that are proper restaurants. Hence venues such as coffee shops, pizza places, bakeries, etc. are not direct competitors of the restaurant business, 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 neighborhood. 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**. This new data frame will now be used for clustering algorithm.

Also, a data frame named **venues_sorted** was also created which listed all the neighborhoods of Toronto along with their respective 12 most common venues. This dataset would eventually help in visualizing the solution.

3. Methodology and Analysis

In the **toronto_restaurants** data frame, I also added a column containing total number of restaurants in that neighborhood. This will help us in making good clusters using the K-Means clustering algorithm.

Now I use K-Means clustering algorithm to make clusters of dataset so that our analysis of the neighborhoods is easy. For this, I set the number of clusters to be 5. The input for this clustering algorithm was **toronto_restaurants** data frame.

After the clusters were made, I merged the first dataset and the **venues_sorted** data frame and inserted cluster labels also.

The next part was Analysis of each cluster to get the correct neighborhood. I calculated total number of neighborhoods 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=0. Hence this cluster was chosen for further analysis.

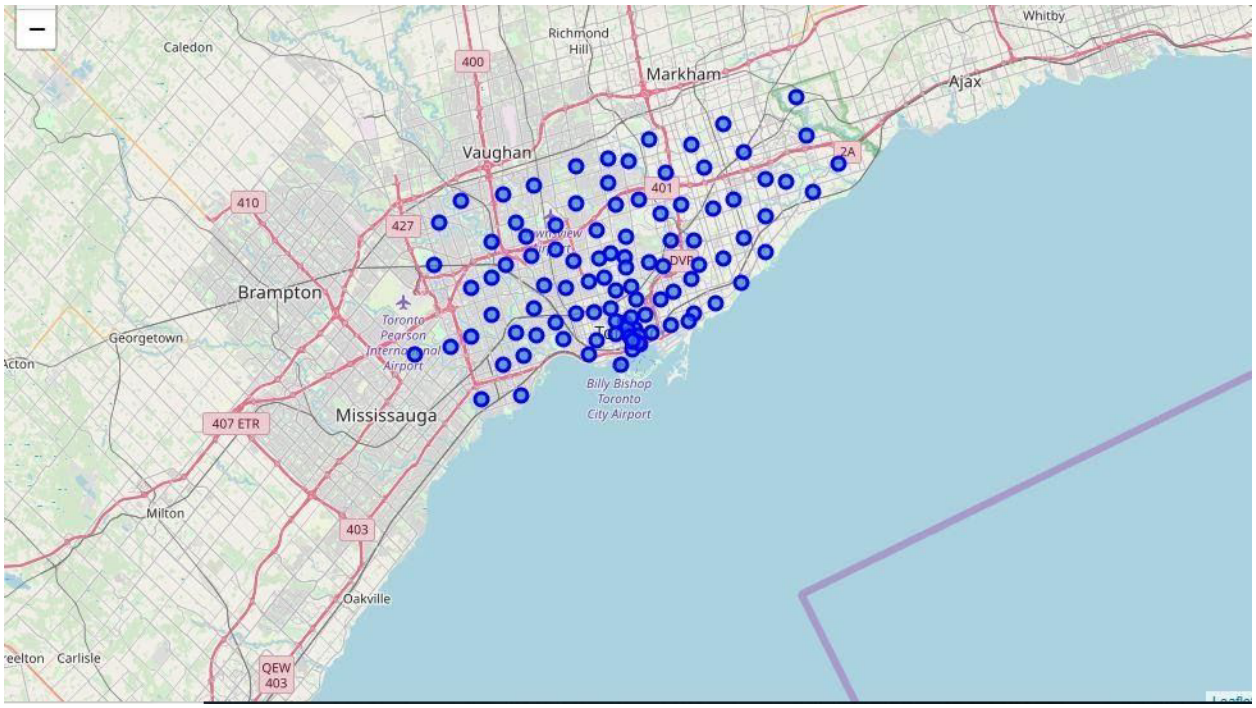
Cluster 4 consisted of total 5 neighborhoods. Out of these, 1 had very high total number of restaurants, therefore their 1 neighborhood was discarded. Out of the remaining 4 neighborhoods, 1 more was discarded because it had Restaurant as their most common venue more than once in the **toronto_merged** data frame and hence these neighborhoods were not suitable for Restaurant business and hence discarded.

The final dataset contains all the information about these remaining 3 neighborhoods.

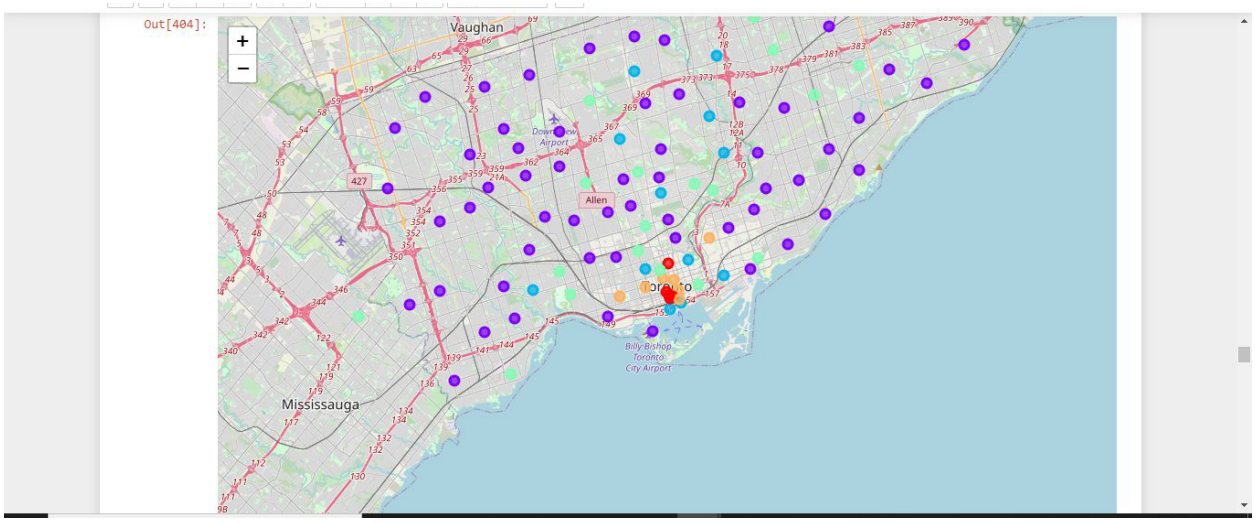
The owners can further choose from these 3 locations which will be the best according to the type of restaurant they are trying to open.

4. Data Visualisation

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

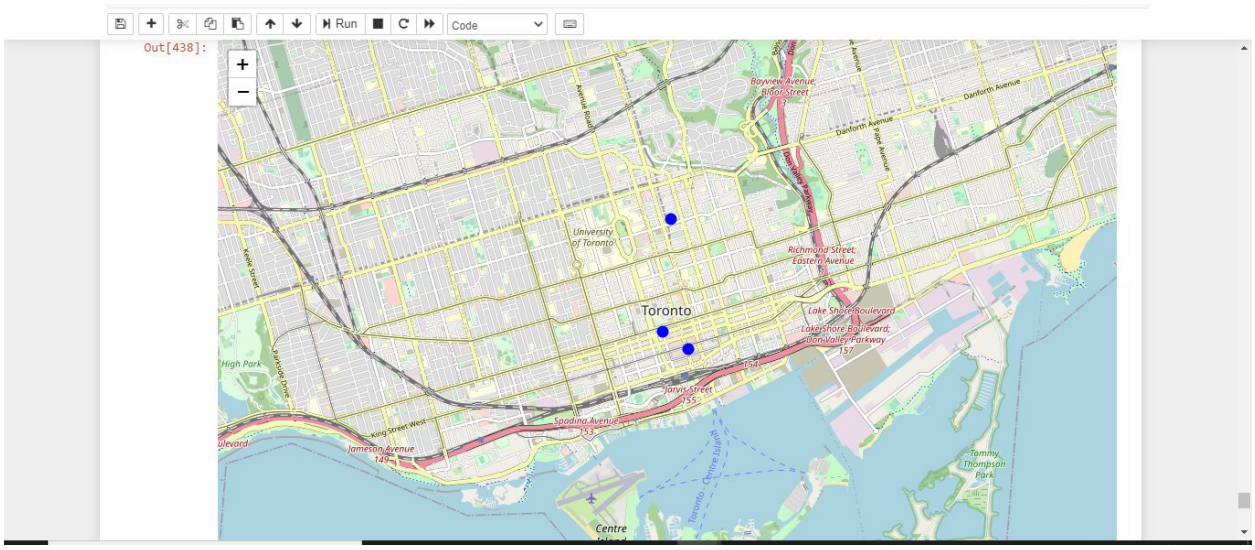


After using the clustering algorithm and creating 5 different clusters where each neighborhood belong to one of these clusters, the new map of Toronto looked like this:



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

The final 3 neighborhoods were also presented on a map:



The 3 neighborhoods are depicted by 3 blue dots in the above map.

5. Result and Discussion

Our analysis shows that although there is a great number of restaurants in Toronto, there are pockets of low restaurant density spread across Toronto city. To identify these pockets, I used clustering algorithm and segmented our neighborhood dataset accordingly.

I used K-means clustering algorithm for making 5 clusters each containing some neighborhoods based on number of restaurants they have in their vicinity. Then I analyzed each cluster by calculating the Restaurant/Neighbourhood ratio of each cluster. I saw that cluster 0 had the lowest ratio, which means very few restaurants are present within the vicinity of each neighborhood that belonged to that cluster. There were a total of 5 neighborhoods belonging to cluster 0. Then upon further analysis, I found that 2 among those were not good for opening up a new restaurant. Hence, only 3 neighborhoods were left.

According to my analysis, I got a total of 3 neighborhoods where the restaurant business will be good. There are two reasons for that. The first reason is that we saw that these neighborhoods do not contain many restaurants around their vicinity which will lower the competition in the restaurant business and give them a competitive advantage. The second reason is that, as we can see in the above map that these 3 neighborhoods lie nearly in the center of Toronto city which means these neighborhoods must have high population density which means more customers and hence more profit.

The final 3 neighborhoods 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 neighborhoods.

The owners can further choose from these 3 locations which will be the best according to the type of restaurant they are trying to open.

6. Conclusion

The purpose of this project was to identify neighborhoods in Toronto which have low number of restaurants in order to aid stakeholders in narrowing down the search for the 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 neighborhood. Then with the help of clustering techniques and further analysis we were able to narrow down our analysis to 3 neighborhoods which were good for opening up a new restaurant. This concludes this project of **Battle of Neighbourhoods.**