**BATTLE OF NEIGHBOURHOODS**

Kushal Deswal

July 12, 2020

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

A chain of restaurant owners in Ontario, Canada want 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 of Canada and has large population density. So they want to open up a new restaurant some place nice with good neighbourhood in Toronto.

* 1. **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. **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.

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

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

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

Firstly, I used data from a Wikipedia page which provides information about all the neighbourhoods 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. 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 11 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. Foursquare is a website that provides any information about a particular venue. I used 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.

I 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 in Toronto 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 the second dataset was needed so that it can be used for clustering algorithm easily. I pre-processed **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, Playground 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 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**. 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 neighbourhoods of Toronto along with their respective 5 most common venues. This dataset would eventually help in visualising the solution. First 10 rows of this data frame is depicted in figure below:

