**1 Introduction**

Recent years, drinking bubble teas has slowly becoming a trend for Toronto citizens due to the cultural diversity of the city. Therefore opening up a bubble tea shop has been a huge market for Chinese investors not only because of the trend but also the low cost of making bubble teas. Hence opening a new bubble tea shop is a difficult task due to its competiveness. If the shop is opened in the area with high competition, the business will not be feasible. If the shop is opened in the area where bubble tea is not so common for the people who lived in that area, the business will not be successful. Therefore, choosing a right location is a huge first step for a successful bubble shop.

**1.1 Business Problem**

The objective of this capstone project is to find a right location for opening a new bubble tea shop in the city of Toronto, Canada. We will use data from Foursquare API and use Data Science methodologies to analysis the data. We will use clustering in machine learning to come up with a model that will predict that most appropriate location to open a new bubble tea shop.

**2 Data acquisition and cleaning**

**2.1 Data Source**

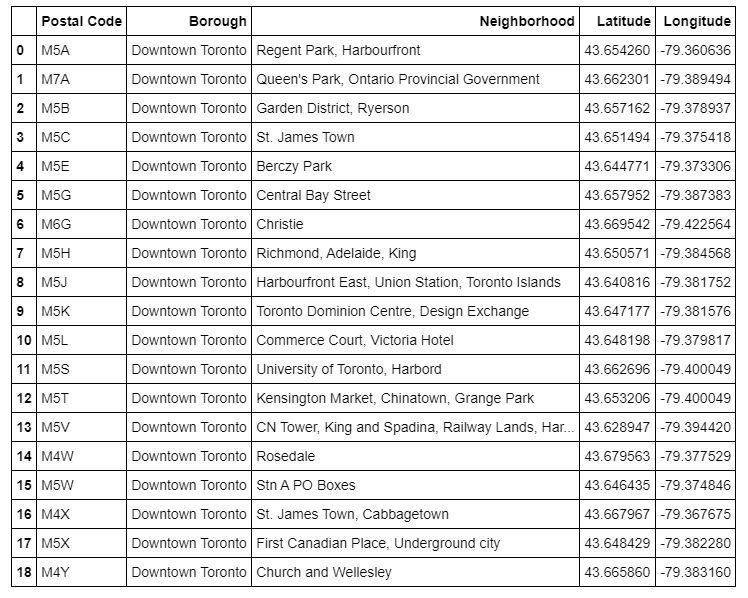
a. list of postal code of Canada from the wiki page: (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>).

b. geographical coordinates of the neighborhoods from <https://cocl.us/Geospatial_data>

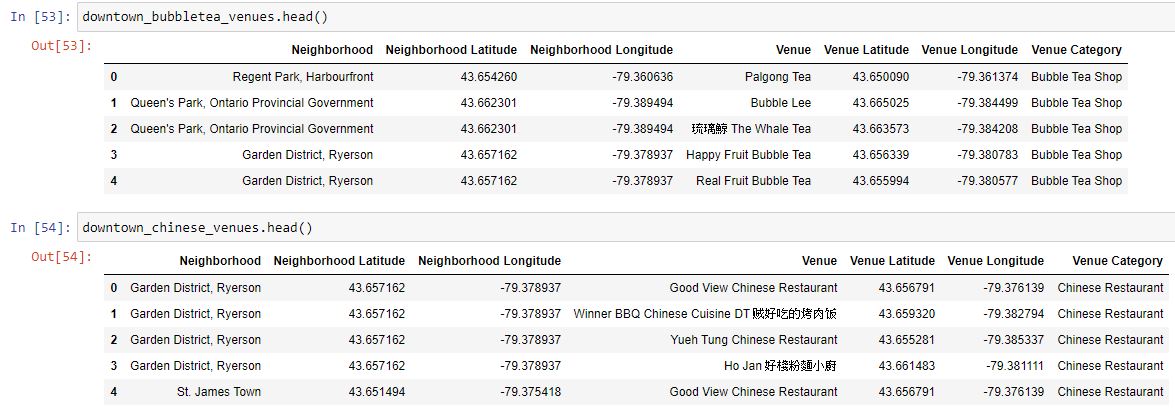
c. explore various venues using Foursquare’s explore API

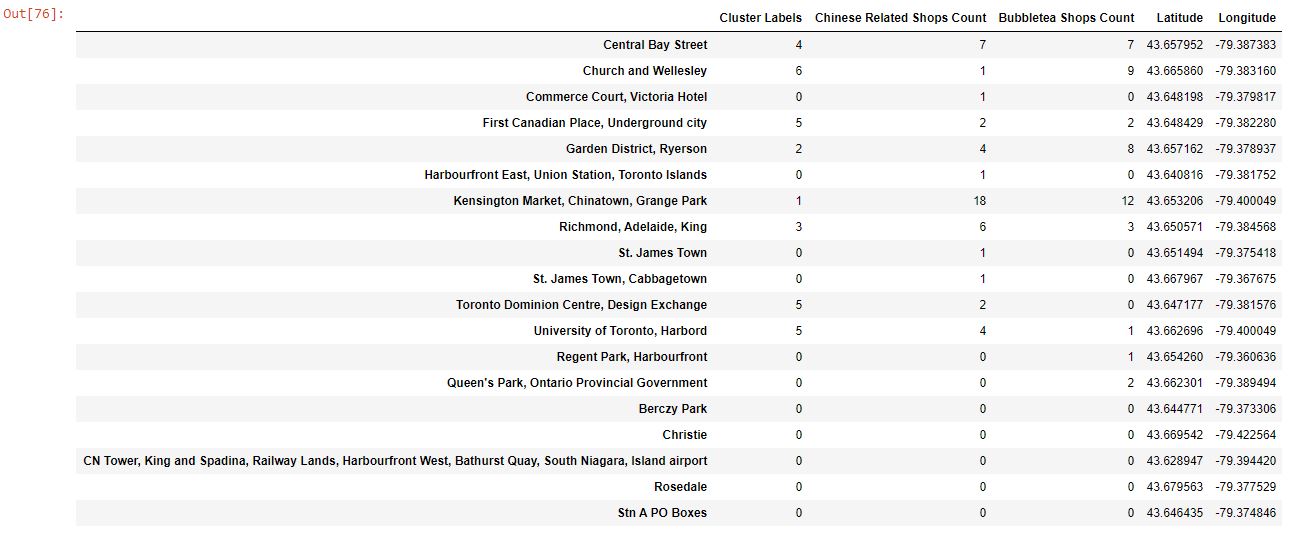
**2.2 Data Cleaning**

Since the project only focus on downtown region of Toronto, I would need to come out with a dataframe that only consists information strictly from downtown region. Then I will combine the data from geographical coordinates of the neighborhoods to get the coordinates of each neighbourhood.



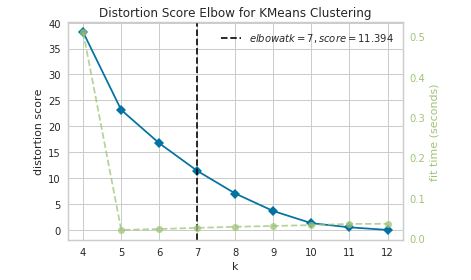
The next step is use the Foursquare API to explore various venues from each neighbourhood. There are two dataframes, one contains all the available bubble tea in each neighbourhood, the other one contains all Chinese related shops. Once I got all the venues from all the neighbourhoods, I combined all the venues to single dataframe and only contain the information I need for analysis.





**3 Predictive Modelling**

K-mean clustering was used to cluster the above datatframe. First the elbow method was used to determine the appropriate k value.

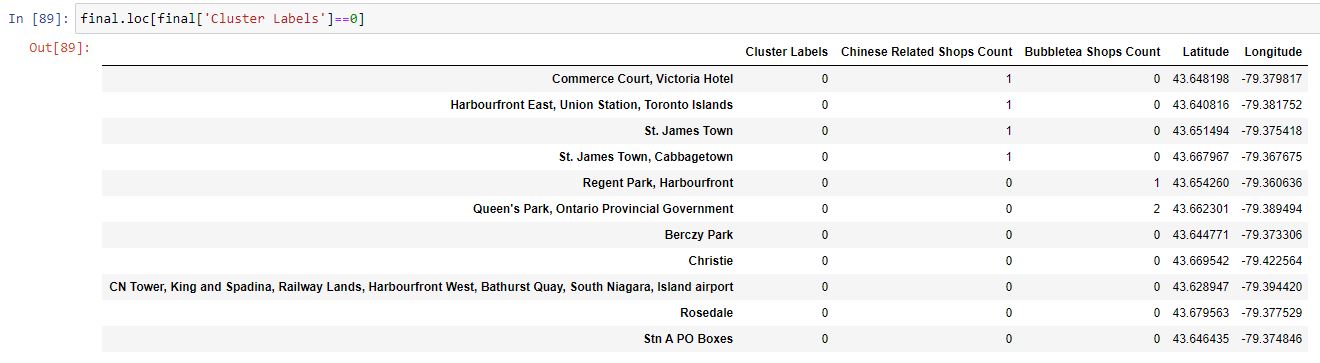


Then k = 7 was used to do the modeling.



**4.2 Examination and Conclusion:**

Let’s examining cluster 0.



This cluster contains the least amount of competition on bubble tea shops. There aren’t many Chinese related shops opening in the areas, that means the neighbourhoods are not populated with Chinese people. So opening a bubble tea shop in these areas is not really a good ideal.

Let’s examining Cluster 6



Cluster 6 contains neighbourhoods that have too much competition on bubble tea shops and also there aren’t many Chinese related shops opening in the areas either. Therefore opening a bubble tea shop in this cluster is not ideal.

Cluster 1, 2, 3, 4 and 5 are good candidates for opening a bubble tea shop.



Out of these clusters, cluster 3 would be the best candidate to open a bubble tea shop. The cluster contains good amount of Chinese related shops which means the neighbourhoods involve a lot of Chinese related activities. At the same time they are not packed with bubble tea shops which mean less competition. Since the cluster only contains a neighbourhood which is “Richmond, Adelaide, King”, this would be the neighbourhood to open a bubble tea shop!