IBM Applied Data Science Capstone Project

1. Introduction

This project will attempt to explore the hypothetical business problem of opening up a Chinese restaurant in Toronto. Opening up a restaurant that provides Chinese cuisine in Toronto can prove to be a lucrative idea considering that South Asians and Chinese immigrants are the leading minority categories in Toronto; both of which enjoy/are familiar with Chinese cuisines. Not only will opening up such a restaurant attract that demographic, it will also give them a sense of belonging when living in a foreign country. The **business problem** to be solved here is then where exactly should an entrepreneur open up the Chinese restaurant in Toronto, in the sense that a strategic location would bring about competitive advantage amongst the other restaurants. By that logic, it is to this research's best interest that the clustering method can be used in order to discover the ideal location in Toronto with less competitors – so that the entrepreneur can obtain a large part of the area's citizens as his/her uncontested market. Ergo, the **target audience** of this deliverable is the entrepreneurs who want to open up a Chinese restaurant in Toronto with a strategic location.

2. Data

In order to solve such a business problem, the following data will be required:

- List of neighbourhoods in Toronto
  - Description: Contains a list of all neighbourhoods in Toronto with their associated postal codes and boroughs
  - Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - Collection method: Web scraping using BeautifulSoup package
- Values of latitude and longitude of the associated neighbourhoods
  - Description: Needed to specify the neighbourhoods' locations in order to further interact with Foursquare API
  - Source: GeoSpace data of Toronto
  - Collection method: Using GeoCoder package in python or CSV file containing the GeoSpace data
- Location data regarding venues that are present within those neighbourhoods
  - Description: Consists of all venues in the neighbourhoods of Toronto
  - Source: Foursquare database
  - Collection method: Using the latitude and longitude values mentioned above, we can communicate and request information from Foursquare using the Foursquare API

3. Methodology

3.1 Data collection

First and foremost in order to start the project, the data mentioned above have to be collected accordingly. By using the BeautifulSoup package, it allows us to web scrape the necessary table that contains the data of Toronto's neighbourhoods from Wikipedia in the form of HTML data. Combining it with Pandas function to read HTML data, we can then turn it into a dataframe to be further prepared in Python.

To obtain the values of latitude and longitude of all the neighbourhoods in Toronto (which is needed to interact with the Foursquare API), the GeoSpace data of Toronto is required. While the GeoCoder package in Python can be used to do so, it proved to be unreliable at the time of this writing. As such, a CSV file that consisted of the GeoSpace data of Toronto neighbourhoods was imported instead.

Last but not least, location data regarding venues present within the Toronto neighbourhoods are to be collected using the Foursquare API. By defining our credentials (client ID and client secret), we can interact with the Foursquare API to gain access to their location data – in which we can select the location data of Toronto's neighbourhoods by including the previously mentioned latitudes and longitudes as parameters of the call request to the API.

3.2 Data wrangling/preparation

The dataframe of Toronto's neighbourhoods had missing values, whereby some postal codes had boroughs and neighbourhoods that were not assigned to them. As a result, rows with missing borough values are dropped, and neighbourhoods with missing values are replaced with the same values as the borough in their respective rows. The latitude and longitude values of all the neighbourhoods were also initially imported as a separate dataframe, so both dataframes were also merged with postal codes being the column to join based on. Afterwards, the location data obtained from using the Foursquare API were also merged by grouping them via the neighbourhoods and by taking the mean of the frequency of occurrence of each venue category. To prepare for further data analysis, the data set was also further filtered to only contain the neighbourhoods and their associated mean frequencies of Chinese restaurants.

3.3 Method of analysis

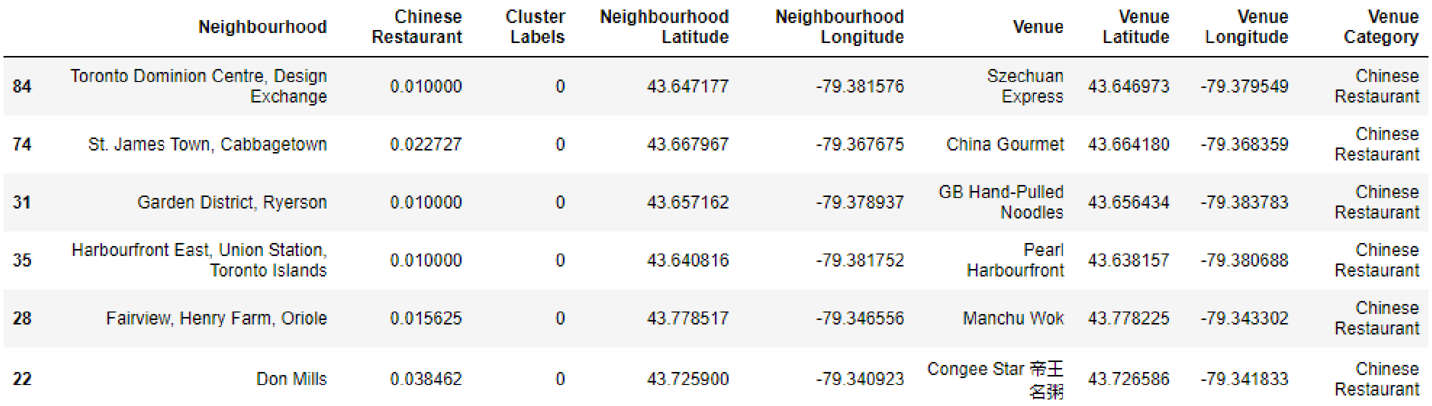
It is important to first explain why clustering is an appropriate machine learning model to be used in formulating a solution for the aforementioned business problem. Clustering is an unsupervised machine learning model that groups a set of data points in such a way that data points within the same groups (also called as clusters) are more similar to each other compared to objects present in other clusters. This way, with the context of the business problem at hand, clustering can be used to group Toronto neighbourhoods which have similar attributes – in which the attribute of interest here is the amount of Chinese restaurants present within the neighbourhood. By knowing which clusters of neighbourhoods are lacking and abundant in Chinese restaurants, we will be able to formulate strategies on where exactly should an entrepreneur attempt to strategically locate their new Chinese restaurant. For the case of this project, k-means clustering will be used. K-means clustering divides the data points into k number of centroids, in which then every data point will be allocated to the nearest centroid – eventually forming a cluster. This simple and straightforward unsupervised machine learning model is appropriate for the scope of this research.

After obtaining a dataframe that contains of the neighbourhoods and their associated mean frequencies of Chinese restaurants, the model can then be created using those data. By setting a k of 3, it will allow us to divide the data points into 3 clusters – which will signify low, medium, and high concentration of Chinese restaurants. Based on the concentrations, the entrepreneur can then strategically locate the Chinese restaurant in the sense that they can select areas with uncontested market.

4. Results

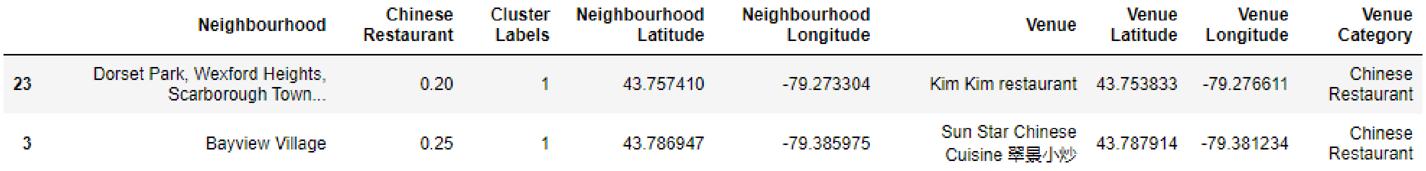
Using the aforementioned parameters, the results are as follows:

Cluster 1:



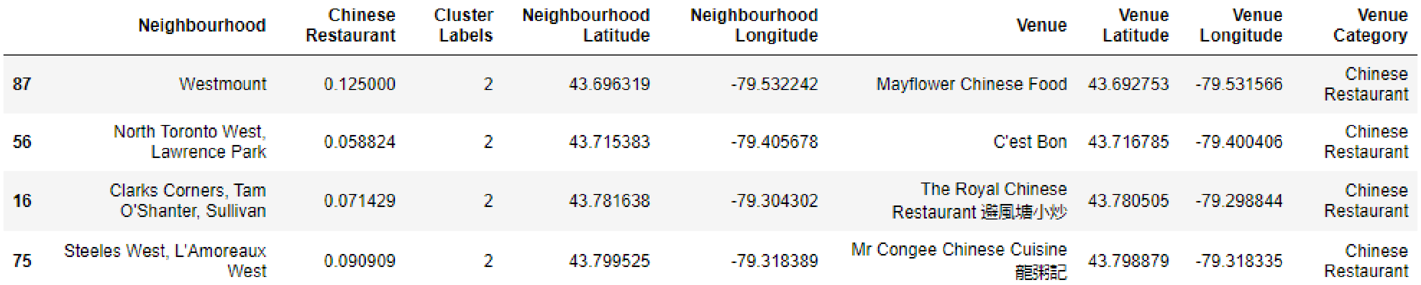
	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
84	Toronto Dominion Centre, Design Exchange	0.010000	0	43.647177	-79.381576	Szechuan Express	43.646973	-79.379549	Chinese Restaurant
74	St. James Town, Cabbagetown	0.022727	0	43.667967	-79.367675	China Gourmet	43.664180	-79.368359	Chinese Restaurant
31	Garden District, Ryerson	0.010000	0	43.657162	-79.378937	GB Hand-Pulled Noodles	43.656434	-79.383783	Chinese Restaurant
35	Harbourfront East, Union Station, Toronto Islands	0.010000	0	43.640816	-79.381752	Pearl Harbourfront	43.638157	-79.380688	Chinese Restaurant
28	Fairview, Henry Farm, Oriole	0.015625	0	43.778517	-79.346556	Manchu Wok	43.778225	-79.343302	Chinese Restaurant
22	Don Mills	0.038462	0	43.725900	-79.340923	Congee Star 帝王名粥	43.726586	-79.341833	Chinese Restaurant

Cluster 2:



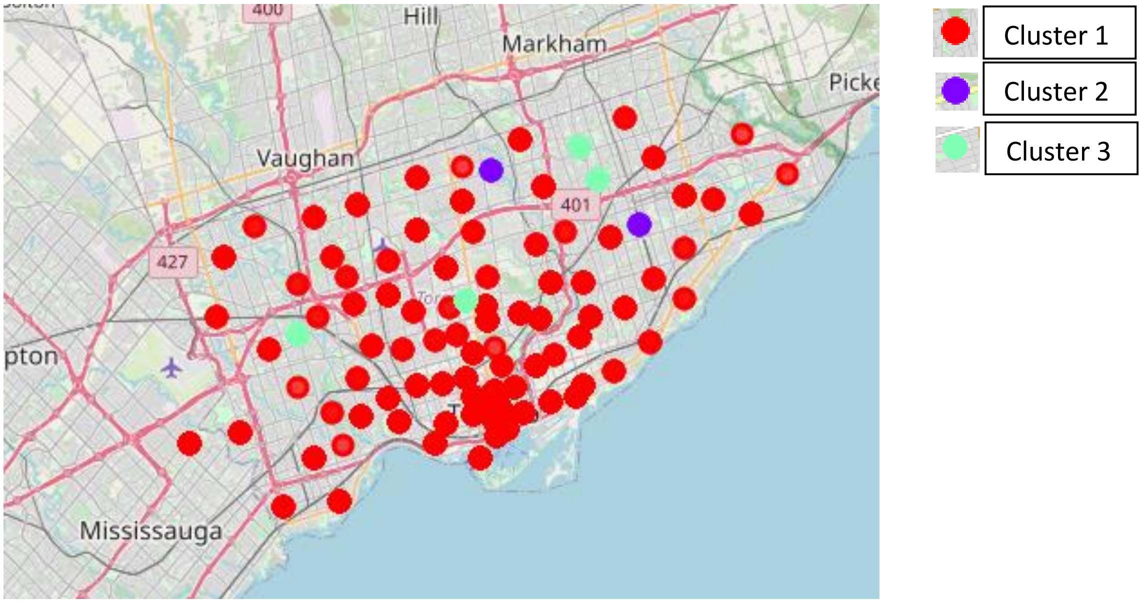
	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
23	Dorset Park, Wexford Heights, Scarborough Town...	0.20	1	43.757410	-79.273304	Kim Kim restaurant	43.753833	-79.276611	Chinese Restaurant
3	Bayview Village	0.25	1	43.786947	-79.385975	Sun Star Chinese Cuisine 翠景小炒	43.787914	-79.381234	Chinese Restaurant

Cluster 3:



	Neighbourhood	Chinese Restaurant	Cluster Labels	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
87	Westmount	0.125000	2	43.696319	-79.532242	Mayflower Chinese Food	43.692753	-79.531566	Chinese Restaurant
56	North Toronto West, Lawrence Park	0.058824	2	43.715383	-79.405678	C'est Bon	43.716785	-79.400406	Chinese Restaurant
16	Clarks Corners, Tam O'Shanter, Sullivan	0.071429	2	43.781638	-79.304302	The Royal Chinese Restaurant 避風塘小炒	43.780505	-79.298844	Chinese Restaurant
75	Steeles West, L'Amoreaux West	0.090909	2	43.799525	-79.318389	Mr Congee Chinese Cuisine 龍粥記	43.798879	-79.318335	Chinese Restaurant

Using the values of mean frequencies of Chinese restaurants found in the second column of the above tables, it can be seen that neighbourhoods in Cluster 1 has low concentration of Chinese restaurants, neighbourhoods in Cluster 3 has medium concentration of Chinese restaurants, and neighbourhoods in Cluster 2 has high concentration of Chinese restaurants. In more detail, these clusters can also be seen in the visualization displayed below; which is the map of Toronto with the 3 clusters superimposed on top.



5. Discussion

For an entrepreneur to open up a Chinese restaurant in a strategic location, it can be assumed that a location with less competitors is an ideal one – seeing that it poses the potential of an untapped and uncontested market. This follows the commonly referred blue ocean strategy as opposed to a red ocean strategy, in which entrepreneurs are to strategically create and capture new demand as opposed to competing for existing demand. Following this notion, it might be beneficial for the entrepreneur to open up their Chinese restaurant in neighbourhoods that are in Cluster 1 considering that it has the lowest concentration of Chinese restaurants. As such, based on the k-means clustering machine learning model, it is **recommended** that entrepreneurs open a Chinese restaurant in the following neighbourhoods if they want to target an uncontested market:

- Toronto Dominion Centre
- Design Exchange
- St. James Town
- Cabbagetown
- Garden District
- Ryerson
- Harbourfront East
- Union Station
- Toronto Islands
- Fairview
- Henry Farm
- Oriole
- Don Mills

6. Conclusion

The aim of this report is to solve the business problem of opening up a Chinese restaurant in Toronto with a strategic location, for the main target audience of entrepreneurs who want to do so. Using the unsupervised machine learning model of k-means clustering, this report then formulated the solution to such a business problem by identifying the areas of Toronto which are less concentrated in Chinese cuisine competition; in which it is discussed previously to be neighbourhoods in Cluster 1. By providing a recommendation that entrepreneurs should open a Chinese restaurant in those neighbourhoods, it is to this project's best interest that the results of this research can be used productively to cater to the untapped market – increasing the welfare of all economic parties involved.