**Location for a Chinese restaurant in Toronto**

1. **Introduction**

The purpose of this Project is to find a better location for a Chinese restaurant to open in Toronto. Toronto is the capital city of the Canadian province of Ontario. It is the most populous city in Canada and the fourth most populous city in North America. The diverse population of Toronto, which means it's going to be a wide variety of food there，everyone loves good food.

So there are multiple opportunities for investors in Toronto. First, a lot of Chinese people living in Toronto. Second, Chinese food is a very popular Asian cuisine. Last but not least, there may not be enough good Chinese restaurants in certain area.

1. **Business Problem**

The aim is suggest a better location for entrepreneur to open a Chinese restaurant. There are many problems we need to consider.

* Which neighborhood has the most Chinese people?
* Is there already a similar Chinese restaurant around?
* Who are the competitors in that area? ......

By using data science methods and machine learning techniques, I

think the answer will become clearer as the project progresses.

Based on the above, our target customer is clear： Entrepreneurs who want to open a Chinese restaurant in Toronto.

1. **Data Description**

To solve this project, we need to use data from:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>  
This wikipedia page has information about all the neighborhoods that we need.

* Postal Code: Postal codes for Toronto
* Borough: Name of Neighborhood
* Neighborhood

Maybe we can also take population distribution into account:

<https://en.wikipedia.org/wiki/Demographics_of_Toronto>

Finally, the data we got from last week: Geospatial\_Coordinates.csv  
This csv file will give us the latitude and longitude of these neighborhoods.

* 1. **Foursquare API Data**

In addition to the above data， we will need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

In conclusion, we need to explore, segment them, and group them into clusters to find some similarity. To be able to do that, we need to cluster data which is a form of unsupervised machine learning, so we choose k-means clustering here.

* 1. **Libraries**
* matplotlib: for generate maps and data visualization
* numpy: handle data in a vectorized manner
* pandas: for data analysis
* sklearn: for machine learning
* json: handle JSON files ...

1. **Data Analysis**

Use the data from wikipedia about the population distribution of each Borough. With the following table we can see Scarborough has the largest population of Chinese people in Toronto's Borough. So we can focus on Scarborough. This significantly reduced data scope.

**Table

Description automatically generated**

This is the map of all theneighborhoods in Scarborough, you can see that there are not too many neighborhoods, which makes it less difficult.

**Map

Description automatically generated**

It’s time for Foursquare! We found 82 restaurants are already opened in Scarborough. Here are some of the restaurants:

Table

Description automatically generated

Table

Description automatically generated

And we can also look at the distribution of cuisines in these restaurants：

Chart

Description automatically generated

As we suspected, there are many Chinese restaurants in Scarborough. It means that the Chinese restaurants are very popular and hot in Scarborough.

And we still need to figure out the number and distribution of restaurants in each neighborhood. So as before, we use matplotlib for

data visualization.

The following table shows the wide variation in the number of restaurants in each neighborhood. Neighborhoods with too many restaurants may not be suitable for another Chinese restaurant, and neighborhoods with too few restaurants may not have a good market for one, so further analysis is still needed.

Chart

Description automatically generated

Next, we will use one hot encoding to analyze different neighborhoods,

the distribution of the top ten restaurants in each neighborhood.

Here we use K-means clustering. Because K-means can group data only unsupervised based on the similarity to each other. It’s very simple and suitable for our project.

Map

Description automatically generated

1. **Conclusion**

K-means clustering separated the neighborhoods into 3 different clusters. And cluster 4 preform best, it looks like a great location for a Chinese restaurant.