Capstone Project - The Battle of Neighborhoods

Clustering strategy for finding potential optimal location of a Chinese restaurant in Toronto

Yilun Zhang

2021.05.21

Introduction - Background

Canadians who identify themselves as being of Chinese ethnic origin make up about 4.6% of the Canadian population, or about 1.57 million people according to the 2016 census. In addition, 11.1% of total population in Toronto are Chinese.

Introduction – Business Problem

- Under this background, opening a Chinese restaurant in Toronto seems to be a great business opportunity for our clients/stakeholders. However, according to my personal experience, not all Chinese restaurant can have a big success or live long in Toronto.

- Although there are many factors accounting for this situation, I think a cautious decision on restaurant location is the first thing needs to be considered for our clients/stakeholders.

- In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening a Chinese restaurant in Toronto. By intuition, population, the number and types of existing Chinese restaurants should be taken into consideration when deciding a good location candidate.

Data

- Based on the population of a neighborhood, the number of Chinese restaurants in a neighborhood and the types of Chinese restaurants in a neighborhood, we can use clustering strategy to segment these neighborhoods in Toronto and decide which neighborhoods our clients/stakeholders should choose when they want to start up a new Chinese restaurant in Toronto.

- We will use web-crawling technique to get all neighborhoods in Toronto with the corresponding postal codes and boroughs.

- Then we will use the GeoSpatial Dataset given by Coursera to get the latitude and longitude coordinates of each neighborhood.

- We will also use the census data to get the population information of each neighborhood downloaded from Canada open data portal.

- Lastly, we will utilize Foursquare API to get the venue information of each neighborhood, or in other words, Chinese restaurant related information.

Methodology - Exploratory data analysis
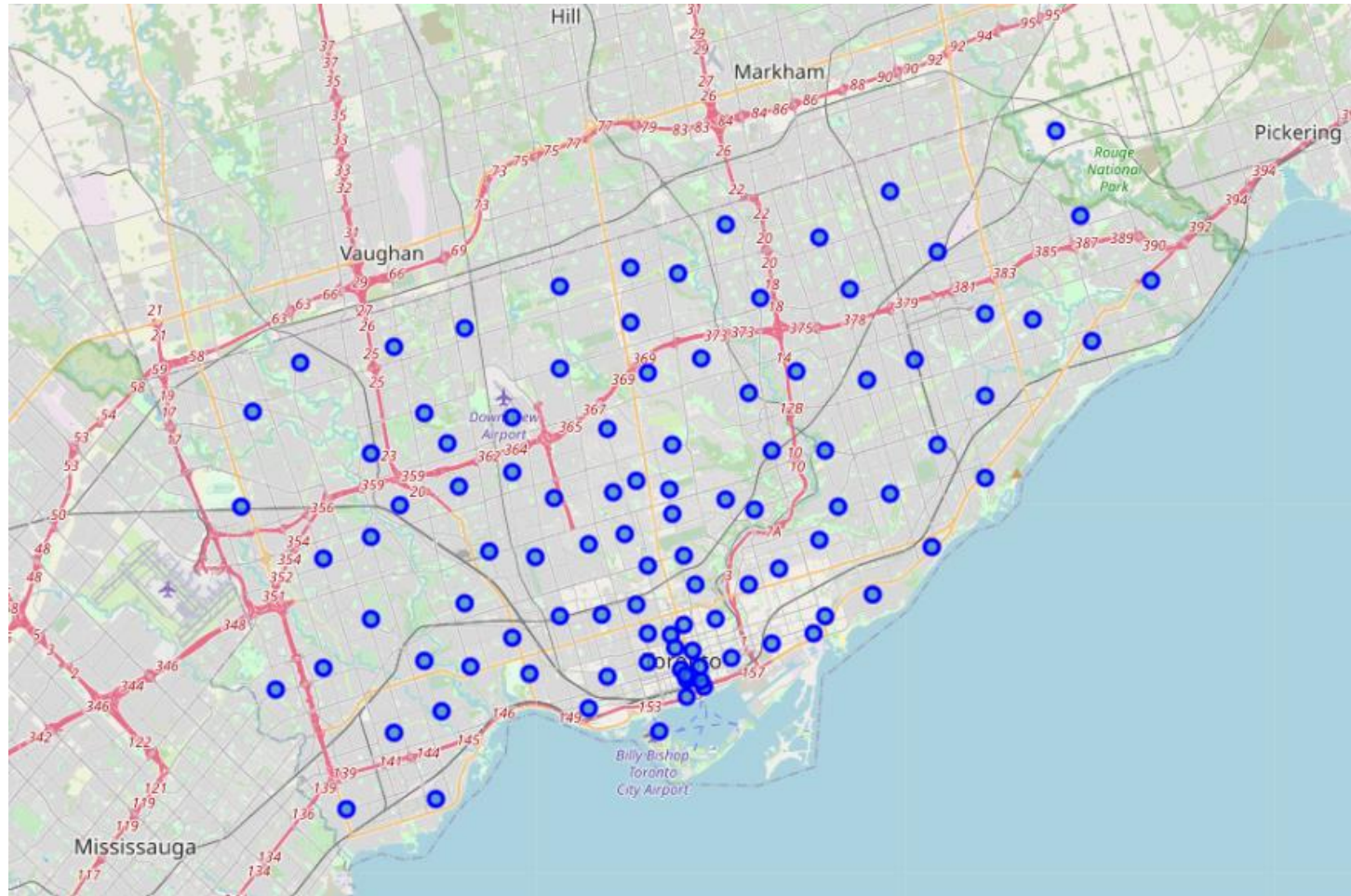
Toronto Map with our target neighborhoods

Methodology - Model

We will use K-Means clustering algorithm to divide neighborhoods into clusters in order to find potential location for a new restaurant. Since K-Means clustering algorithm needs us to predefine the number of clusters K, we need to use the elbow method to find the optimal K first.
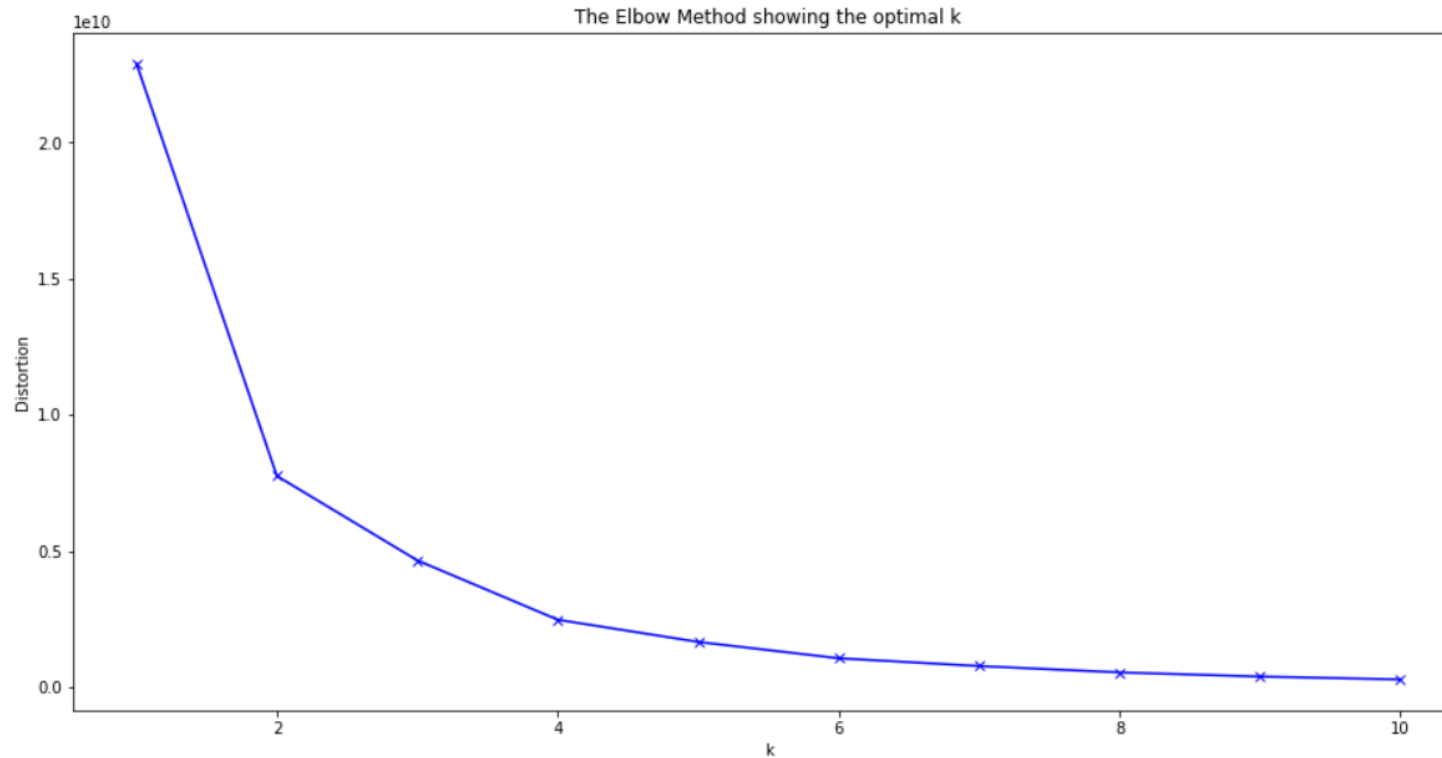
From the plot above we can see that, 4 probably is a good choice for our optimal K.

Results – Visualization of clustering results
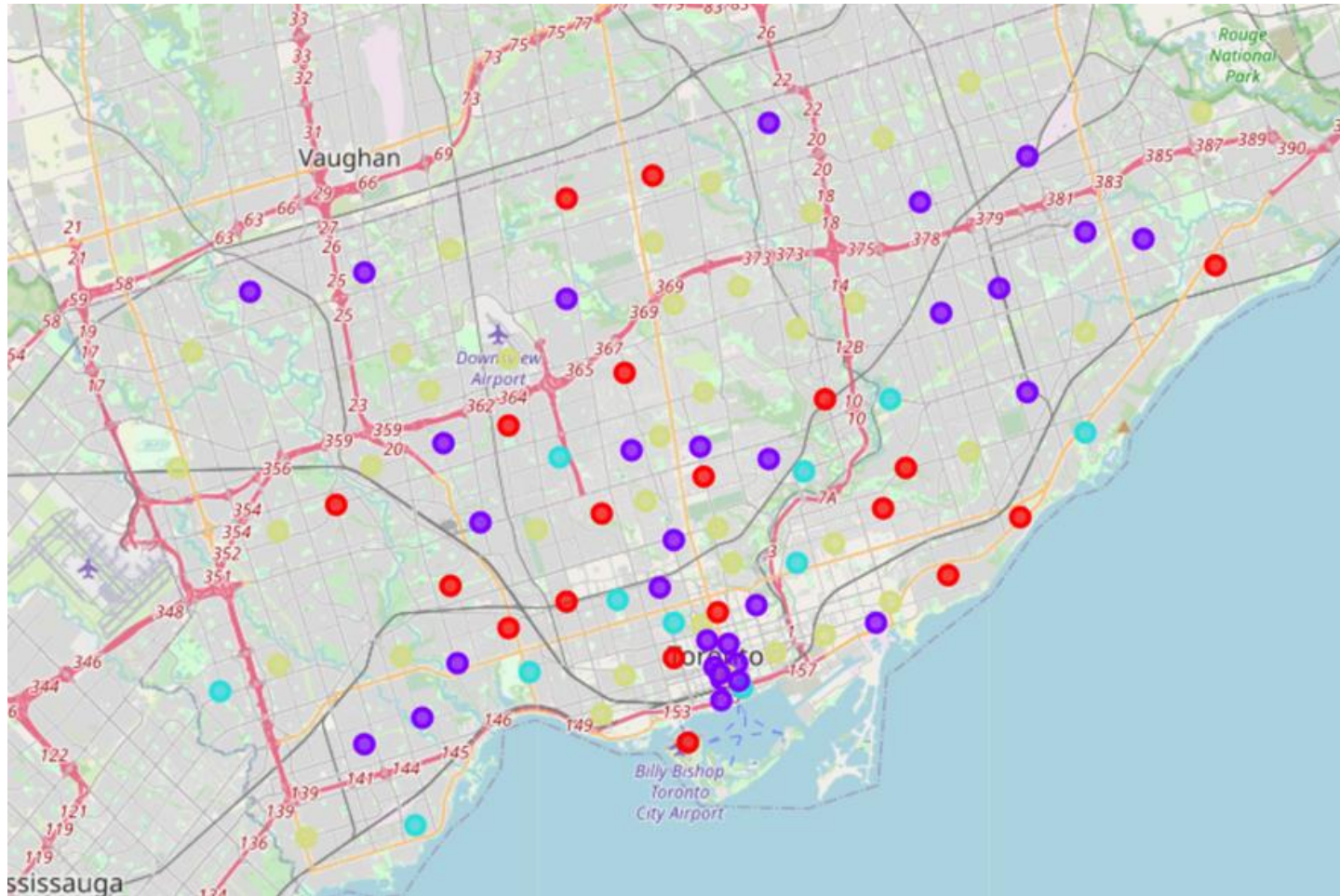
Red dots: Cluster 0
Purple dots: Cluster 1
Blue dots: Cluster 2
Yellow dots: Cluster 3

Results – Summary of clustering results

	Number of neighborhoods	Mean population	Mean number of Chinese restaurants
Cluster 0	19	32343.37	14.47
Cluster 1	31	21400.74	31.03
Cluster 2	11	24531.64	17.36
Cluster 3	37	29074.70	14.38

Discussion

The most interesting result should be Cluster 1 which has the highest mean number of Chinese restaurants and the lowest mean population. From the distribution of purple dots in the map we can see that, neighborhoods in Cluster 1 mainly gather in 4 areas, which are Downtown, North York, Etobicoke and Scarborough. This result makes sense since there are lots of universities and colleges in these 4 areas where many Chinese students are attending. For example, University of Toronto (St. George Campus and Scarborough Campus), York University and Humber College. Thus, even though the mean population is low in these areas, the number of Chinese restaurants are very high. This kind of popular regions should be a good choice for starting up a Chinese restaurants.

For more discussion, please refer to the capstone report.

Conclusion and future improvement

In this report, we find some potential optimal locations for starting up a new Chinese restaurant in Toronto using K-Means clustering algorithm. Popular regions near universities where many Chinese students are attending and regions near popular regions are some good choices. However, these results are still quite rough since location selection for a restaurant involves many more factors than what we used in this report. For example, rent and surrounding infrastructure. So this report is more like a demo. In order to make the results more promising, we need to collect more data to cover more factors and do further analysis.

Thank you for reviewing my assignment!