Capstone Project – The Battle of Neighborhood

Finding a new location for opening a Japanese Restaurant in Toronto

Alan Ng

1. Introduction

The aim of this project is to find an optimal place for a Japanese Restaurant in Toronto. As there are many restaurants in Toronto, we hope to find a location not packed with restaurants. We also hope to find a place not containing many Japanese restaurants to reduce competition.

1. Data

Here are the data the project is used.

1. Toronto Neighborhoods from Wikipedia: it contains almost all the information about the neighborhoods, e.g. postal code, borough, and the name of the neighborhoods.

https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

1. Venue Data from Foursquare API: it is for the number of restaurants and their type and location in every neighborhood
2. Location data from Geocoder Package: it is for getting Geographical coordinates of the neighborhoods. To simplify our analysis, we use the csv file containing the coordinates instead.
3. Methodology
4. Data Cleansing

First, the data of Toronto Neighborhood is scrapped from the Wikipedia. The data frame contains Postal Code, Borough, and Neighborhood. The cells with a borough “Not assigned” will not be processed. If there is more than one neighborhood exist in one postal code area, then it will be merged with 1 row separated with a comma, i.e. the Postal Code is unique and can contain multiple neighborhoods.

Second, the Geographical coordinates are retrieved from a csv file and the coordinates are combined with the data frame above.

Third, the venue data would be retrieved from Foursquare API for each neighborhood. Specifically, the venue category would be retrieved for each neighborhood. Since the problem of interest is Japanese Restaurant, the average Japanese Restaurant of each neighborhood would be retrieved. Totally, there are 39 neighborhoods. And this data would be processed for training the model.

1. Machine Learning

After getting the data, it is then proceeded to training the model. Here the methodology used is k-mean clustering. The neighborhoods are clustered based on similar Japanese Restaurants in that neighborhood. Here the neighborhood is clustered into 5 clusters.

1. Results
2. Map

After clustering, here is the map for different clusters.

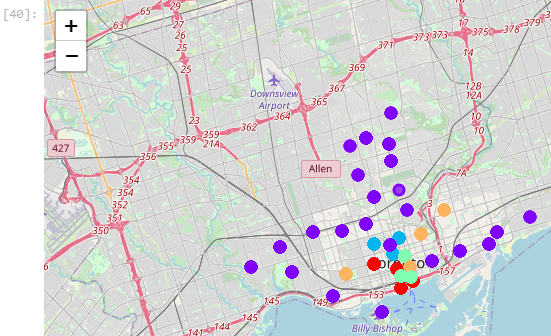
Cluster 1: Purple

Cluster 2: Blue

Cluster 3: Green

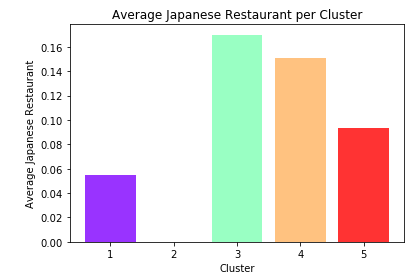
Cluster 4: Orange

Cluster 5: Red



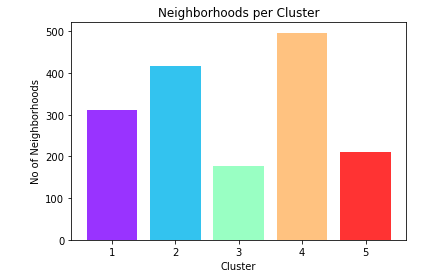
1. Average Japanese Restaurant per Cluster

In Cluster 2, there is no Japanese Restaurant, while Cluster 3 and Cluster 4 contains the most and second most Japanese Restaurant on average



1. Neighborhoods per Cluster

Cluster 4 and Cluster 2 contains the most and second most neighborhoods respectively while cluster 3 contains the least neighborhood.



1. Discussion

Since the problem of interest is to open a new Japanese Restaurant in Toronto, as cluster 2 appears no Japanese Restaurant. So, it might be quite viable to open a new one in cluster 2 as there is fewer competition. Plus, cluster 2 contains the second most neighborhoods, so there should be sufficient choices of locations in cluster 2 so that the population should be sufficient to support the business. Therefore, it is worthwhile to explore cluster 2.

As from the map in cluster 2, basically it is near college area, like University of Toronto, University-Rosedale, etc. It seems it’s a good choice to open a new Japanese Restaurant near there as the customer could be the exchange students from different countries, college students which are more likely to try different cuisine, and young adults which are more willing to pay for food.

1. Conclusion

In this project, the problem of interest is to open a new Japanese Restaurant in Toronto. First, the data is retrieved from different sources and it is used for clustering into different clusters for easier examination using k-means algorithm. After getting the result, based on the requirement, cluster 2 in the result is used for further investigation, such as potential location, and potential customers. At the end, an area near the college area is recommended since there is fewer competition, and good potential customers that can sustain the business.

Throughout this project, the programming language used is Python, and different modules are used for data retrieving, data visualization, machine learning algorithm, and data analysis. This complex problem can be decomposed into different pieces to be tackled easier.

One potential improvement is to include more data like the composite of different population, so as to understand the potential customers better and develop a better strategy to maintain the business, such as how to cater the need of them.