Coursera Capstone Project – “The Battle of Neighborhoods”

Opening a new Chinese Restaurant in Toronto

By Zufar Haseeb

July 2019

Introduction

Chinese cuisine (Chinese dishes) originated from different regions of China and has become widespread in many other parts of the world. As Toronto is a multicultural city and being the provincial capital Ottawa (Capital of Canada), a new Chinese restaurant may generate consistent revenue for any investor while the location of the restaurant being the key determinant of consistent revenue through the success of the Hookah Bar.

Business Problem

Although many neighborhoods in Toronto will have restaurants serving Chinese cusine, many Neighborhoods may not boast of having an authentic Chinese restaurant providing the typical Chinese food experience. Objective of this project is to analyze and select the best location in Toronto, Ottawa to open a new Chinese Restaurant. Using data science methodologies such as clustering, this project aims to provide suggestions for locations in Toronto for a new Chinese Restaurant to be opened.

Target Audience

- Investors
- Tourists
- Franchises, etc.

Data

We will require following data to solve this problem

- List of neighborhoods in Toronto. This defines the scope of the project which will be confined to the city of Toronto, the provincial capital of Ottawa, which is the Capital of Canada.
- Latitude and Longitude coordinates of above neighborhoods in order to plot the maps and obtain venue data
- Venue data, particularly data of Chinese restaurants. This can be used to perform clustering of neighborhoods.

Sources of Data

- The Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) contains a list of postal codes in Toronto with a total of 103 postal codes. We will be using web scraping techniques to extract data from the Wikipedia page with the help of Python and beautifulsoup packages. Then we will get the geographical coordinates of the above neighborhoods using Python's Geocoder package.
- Foursquare API to get venue data for above neighborhoods. Foursquare has one of the largest databases of over 100 million places and is used by over 150,000 developers across the world. Foursquare API will provide many categories of venue data and we will be focusing on Chinese Restaurant category in order for us to solve the problem that has been put forward.

Methodology

Firstly, we need to obtain the list of neighborhoods in Toronto. Fortunately, the list is available on the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We will be performing web scraping using Python requests and beautiful soup packages to extract the list of neighborhood data. However, this will be just a list of names. We will require the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will be using the Geocoder package that will help us in converting addresses in to geographical coordinates in the form of latitude and longitude. Upon gathering the data, we will populate the data into a Pandas DataFrame and then visualize the neighborhoods in a map with the help of Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by the Geocoder are correctly plotted in the city of Toronto.

Next, we will use Foursquare API to get the top 100 venues within the radius of 500 meters. We need to register a Foursquare developer account in order to obtain the Foursquare ID and the secret. We then make API calls to the Foursquare passing in the geographical coordinates of the neighborhoods in a python loop. Foursquare will return the data in JSON format and we will extract the venue name, venue category, geographical coordinates of the venue (latitude and longitude. With the data we can also check how many venues were returned for each neighborhood and examine how many unique categories can be curated from the returned venues. Then we will analyze each neighborhood by grouping the rows by postcode and taking the means of frequency of occurrence for each venue category. By doing so we are also preparing the data for clustering. Since we are analyzing the Chinese Restaurants data, we will filter the “Chinese Restaurants” as venue category for neighborhoods.

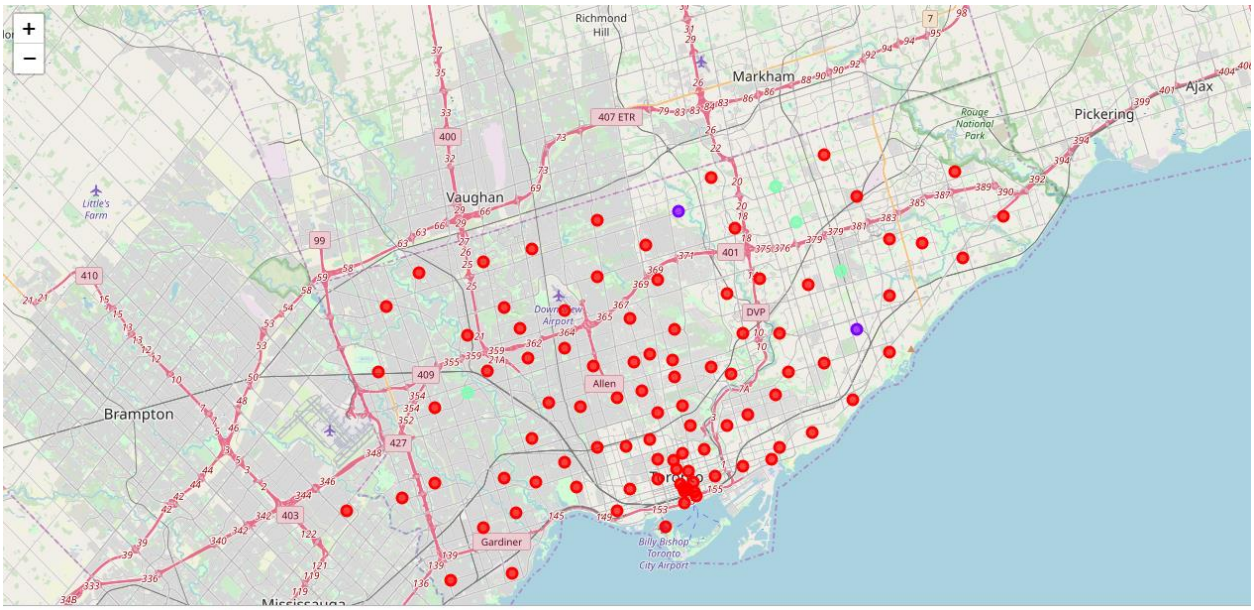
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Chinese Restaurant”. The results will allow us to identify which neighborhoods have higher concentration of Chinese restaurants while which neighborhoods have fewer number of Chinese restaurants. Based on the occurrence of Chinese restaurants in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Chinese restaurants.

Results

Results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Chinese Restaurant”.

- Cluster 0: Neighborhoods with low number to no existence of Chinese restaurants
- Cluster 1: Neighborhoods with moderate number of Chinese restaurants
- Cluster 2: Neighborhoods with high concentration of Chinese restaurants

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color and cluster 2 in mint green color.



Discussion

As observations noted from the map in the results section, most of the Chinese restaurants are concentrated in the southern eastern part of Toronto, with the with the highest number in cluster 2 and moderate number in cluster 1. On the other hand, cluster 0 has very low number to totally no Chinese restaurant in its neighborhoods. This represents a great opportunity and high potential areas to open new Chinese restaurants in cluster 0 as there is very little to no competition from existing Chinese restaurants. Meanwhile, Chinese restaurants in cluster 2 are likely suffering from intense competition due to oversupply and high concentration of Chinese restaurants. From another perspective, this also shows that the oversupply of Chinese restaurants mostly happened in the southern part of Toronto, with the suburb area still have very few Chinese restaurants. Therefore, this project recommends franchises and investors to capitalize on these findings to open new Chinese restaurants in neighborhoods in cluster 0 with little to no competition. Investors/Franchises with unique selling propositions to stand out from the competition can also open new Chinese restaurants in neighborhoods in cluster 1 with moderate competition. Lastly, investors/franchises are advised to avoid neighborhoods in cluster 2 which already have high concentration of Chinese restaurants and suffering from intense competition.

Limitations and suggestions for future research

In this we consider on one factor. i.e. the frequency of occurrence of Chinese restaurants, there are other factors such as population mix and income of residents that could influence the location decision for a new Chinese restaurant. However, to the best knowledge of me, such data is not available to the neighborhood level required by this project. Future research could develop a methodology to estimate such data to be used in the clustering algorithm to determine preferred locations for new Chinese restaurants. In addition, this project made use of the free sandbox tier account of Foursquare API that came with its own limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more accurate results.

Conclusion

In this project we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to relevant stakeholders i.e. investors and franchises regarding the best locations to open a new Chinese restaurant. The findings of this project will help the relevant stakeholders to capitalize on opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Chinese restaurant.

References

Category: Neighborhoods of Toronto. *Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Foursquare Developer Documentation. *Foursquare.* Retrieved from https://developer.foursquare.com/docs

Appendix

Cluster 0

Neighbourhood
Rouge, Malvern
Highland Creek, Rouge Hill, Port Union
Guildwood, Morningside, West Hill
Woburn
Cedarbrae
Scarborough Village
Clairlea, Golden Mile, Oakridge
Cliffcrest, Cliffside, Scarborough Village West
Birch Cliff, Cliffside West
Maryvale, Wexford
Agincourt
Agincourt North, L'Amoreaux East, Milliken, St...
Hillcrest Village
Fairview, Henry Farm, Oriole
Willowdale South
York Mills West
Willowdale West
Parkwoods
Don Mills North
Flemingdon Park, Don Mills South
Bathurst Manor, Downsview North, Wilson Heights
Northwood Park, York University
CFB Toronto, Downsview East
Downsview West
Downsview Central
Downsview Northwest
Victoria Village
Woodbine Gardens, Parkview Hill
Woodbine Heights
The Beaches
Ryerson, Garden District
St. James Town
Berczy Park
Central Bay Street
Adelaide, King, Richmond
Harbourfront East, Toronto Islands, Union Station
Design Exchange, Toronto Dominion Centre
Commerce Court, Victoria Hotel
Bedford Park, Lawrence Manor East
Roselawn
Forest Hill North, Forest Hill West
The Annex, North Midtown, Yorkville
Harbord, University of Toronto
Chinatown, Grange Park, Kensington Market
CN Tower, Bathurst Quay, Island airport, Harbo...
Stn A PO Boxes 25 The Esplanade
First Canadian Place, Underground city
Lawrence Heights, Lawrence Manor
Glencairn
Humewood-Cedarvale
Caledonia-Fairbanks
Christie
Dovercourt Village, Dufferin
Little Portugal, Trinity
Brockton, Exhibition Place, Parkdale Village
Downsview, North Park, Upwood Park
Del Ray, Keelesdale, Mount Dennis, Silverthorn
The Junction North, Runnymede
High Park, The Junction South
Parkdale, Roncesvalles
Runnymede, Swansea
Leaside
Thorncliffe Park
East Toronto
The Danforth West, Riverdale
The Beaches West, India Bazaar
Studio District
Lawrence Park
Davisville North
North Toronto West
Davisville
Moore Park, Summerhill East
Deer Park, Forest Hill SE, Rathnelly, South Hi...
Rosedale
Cabbagetown, St. James Town
Church and Wellesley
Harbourfront, Regent Park
Queen's Park
Canada Post Gateway Processing Centre
Business Reply Mail Processing Centre 969 Eastern
Humber Bay Shores, Mimico South, New Toronto
Alderwood, Long Branch
The Kingsway, Montgomery Road, Old Mill North
Humber Bay, King's Mill Park, Kingsway Park So...
Kingsway Park South West, Mimico NW, The Queen...
Cloverdale, Islington, Martin Grove, Princess ...
Bloordale Gardens, Eringate, Markland Wood, Ol...
Humber Summit
Emery, Humberlea
Weston
Kingsview Village, Martin Grove Gardens, Richv...
Albion Gardens, Beaumond Heights, Humbergate, ...
Northwest

Cluster 1

Neighbourhood
East Birchmount Park, Ionview, Kennedy Park
Bayview Village

Cluster 2

Neighbourhood
Dorset Park, Scarborough Town Centre, Wexford ...
Clarks Corners, Sullivan, Tam O'Shanter
L'Amoreaux West
Westmount