Applied Data Science Capstone

Gunay Ismayilova

Optimal Location for Opening a Chinese Restaurant in Toronto

1. Introduction:

1.1 Background

While opening a restaurant is one of the profitable businesses, wrong location may lead to failure within a year. Location should be the first consideration before opening a restaurant because it directly determines demand for the restaurant.

1.2 Business Problem:

The objective of this Capstone project is to find an optimal location for opening a Chinese Restaurant in Toronto. With its high income level, being home to diverse nationalities and many successful businesses, Toronto is where many entrepreneurs are willing to open a restaurant. With its unique cuisine and tastes, Chinese restaurant is in high demand all over the world, including in Toronto. So, for this project, we will assume that an entrepreneur plans to open a Chinese restaurant in Toronto. Our focus is to determine neighborhoods where the demand for Chinese restaurant is the highest.

1.3 Target Audience:

The target audience of the project are listed below:

1. Local entrepreneurs planning to open a Chinese restaurant in Toronto
2. Chinese businessmen planning to open a Chinese restaurant in Toronto
3. Chinese restaurant owners willing to open their branch in Toronto

All the stakeholders listed above are interested to find out an optimal neighborhood in Toronto to open their Chinese restaurant. Choosing a wrong neighborhood means failure of a business for them. This analysis will help them to better understand Toronto neighborhoods and find the best location for their restaurant.

2. Data acquisition and cleaning:

2.1 Data sources

a) I'm using "List of Postal code of Canada: M" (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) wiki page to get all the information about the neighborhoods present in Toronto. This page has the postal code, borough & the name of all the neighborhoods present in Toronto.

b) Then I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighborhoods.

c) To get information about population size, population density and average income in each neighborhood of Toronto, I'm using "Demographics of Toronto neighborhoods" (https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods) wiki page.

d) To extract number of Chinese restaurants in each neighborhood, I'm using Foursquare's explore API.

2.2. Data cleaning

a) Scraping Toronto Neighborhoods Table from Wikipedia

I scraped the following Wikipedia page, "List of Postal code of Canada: M" in order to obtain the data about the Toronto & the Neighborhoods in it. Then I dropped all the rows with missing data. Below is the cleaned dataframe of Toronto Neighborhoods:

	Postcode	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Harbourfront
5	M5A	Downtown Toronto	Regent Park
6	M6A	North York	Lawrence Heights

b) Adding geographical coordinates to the neighborhoods

Next step is adding the geographical coordinates to these neighborhoods. To do so I'm extracting geographical data from Geospatial Data csv file.

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Then I'm combining it with the existing neighborhood dataframe by merging them both based on the postal code.

```
In [41]: toronto_postal_coordinates = pd.merge(toronto_postal, coordinates, on='PostalCode', how='inner')
         toronto_postal_coordinates.head(10)
```

Out[41]:

	PostalCode	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M5A	Downtown Toronto	Regent Park	43.654260	-79.360636
4	M6A	North York	Lawrence Heights	43.718518	-79.464763
5	M6A	North York	Lawrence Manor	43.718518	-79.464763
6	M7A	Queen's Park	Not assigned	43.662301	-79.389494
7	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
8	M1B	Scarborough	Rouge	43.806686	-79.194353
9	M1B	Scarborough	Malvern	43.806686	-79.194353

c) Scraping population size, population density and average income for each neighborhood of Toronto

Demand for Chinese restaurants in Toronto is likely to be more where population size is large, population density and average income are high. To acquire the indicators above I import "Demographics of Toronto neighborhoods" table from Wikipedia.

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4	NaN	NaN	NaN
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	NaN
2	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	06.2% Polish	NaN
3	Alexandra Park	OCoT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	NaN
4	Allenby	OCoT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4	Russian (1.4%)	01.4% Russian	NaN

Then I extract population size, population density and average income from that table.

```
In [26]: toronto_demographics=toronto_demographics[['Name','Population', 'Density (people/km2)','Average Income']]
         toronto_demographics.head()
```

Out[26]:

	Name	Population	Density (people/km2)	Average Income
0	Toronto CMA Average	5113149	866	40704
1	Agincourt	44577	3580	25750
2	Alderwood	11656	2360	35239
3	Alexandra Park	4355	13609	19687
4	Allenby	2513	4333	245592

d) Using Foursquare's *explore API, acquire number of Chinese restaurants for each neighborhood.*

A new Chinese restaurant is likely to more less profitable when there are already many Chinese restaurants in its neighborhood and competition is high. Therefore, a new Chinese restaurant should be opened in a neighborhood with the minimum number of Chinese restaurants. For that, I will acquire number of Chinese restaurants in each neighborhood using Foursquare's API.

Foursquare data is very comprehensive, and it powers location data for Apple, Uber etc. For this business problem I have used it's explore API to get venues in Toronto with their locations and categories. With some data manipulation, I got neighborhood of each venue acquired from Foursquare.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Then by using one hot encoding I found number of venues in each category. After that I grouped venue categories by Neighborhood. At the end I found number of Chinese restaurants in each neighborhood by filtering only Chinese restaurants.

	Neighborhood	Chinese Restaurant
0	Adelaide	0
1	Agincourt	1
2	Agincourt North	0
3	Albion Gardens	0
4	Alderwood	0
5	Bathurst Manor	0
6	Bathurst Quay	0
7	Bayview Village	1
8	Beaumond Heights	0
9	Bedford Park	0