Capstone Project

Finding the best location in Toronto to open a Chinese food market

Qian Li

July 29, 2020

Agenda

Part 1: Intro/Business problem

Part 2: Data and Methodology

Part 3: Finding areas most populated by Chinese immigrants

Part 4: Decide on area that is least competitive

Part 5: Decide on area with lightest leasing cost

Part 6: Modelling and Testing

Part 7: Result

Part 8: Discussion

Part 9: Conclusion

Part 1: Intro/Business problem

Mr. Chen is a new immigrant who owned a chain supermarket company back in China.

He would like to apply his business talents to the Canadian market and open a supermarket targeting Chinese immigrants.

With everything else prepared, he now needs to know which community is the best choice for him to start his business.

His requirements for the ideal place are:

1. large Chinese population;

2. few competitors;

3. reasonable rent.

Part 2: Data and Methodology

To meet Mr. Chen's requirements, we need both location data and demographic data for the Toronto area. Hence the 3 main data sources will be **Toronto public data portal,** **Foursquare** and **CBRE**.

The whole research will be carried out in 4 main steps.

First, we use Toronto data to target the main candidate regions. We collect data from Toronto public data portal to find out the most populated regions. We need pandas and matplotlib. Second, we use data from Foursquare and decide on the number of competitors and population coverage per competitor. We need folium for visualization. Third, we got data from CBRE to see the leasing costs in shortlisted regions. Fourth, we crunch relevant data into one data frame, and import reference data, which is used to do supervised machine learning in deciding which region has the best chance for success.

Part 3: Finding areas most populated by Chinese immigrants

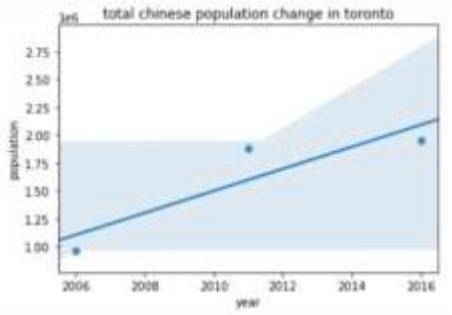
This will be solved by leveraging the demographic data from public data portal for Toronto(https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/) in the most recent years. First is to locate the promising areas where lots of Chinese people reside. We will look at two numbers; one is the total number of Chinese population in each region. Second, we will need these numbers for the most recent 3 years, as a reference to the demographic trend in these areas.

There are 48 wards in Toronto. Data for 3 years are found: 2006, 2011, and 2016. By adding up all related Chinese language we got a total population in each ward, and then putting the total population of each ward of those 3 years into one data frame. The result is as follow:

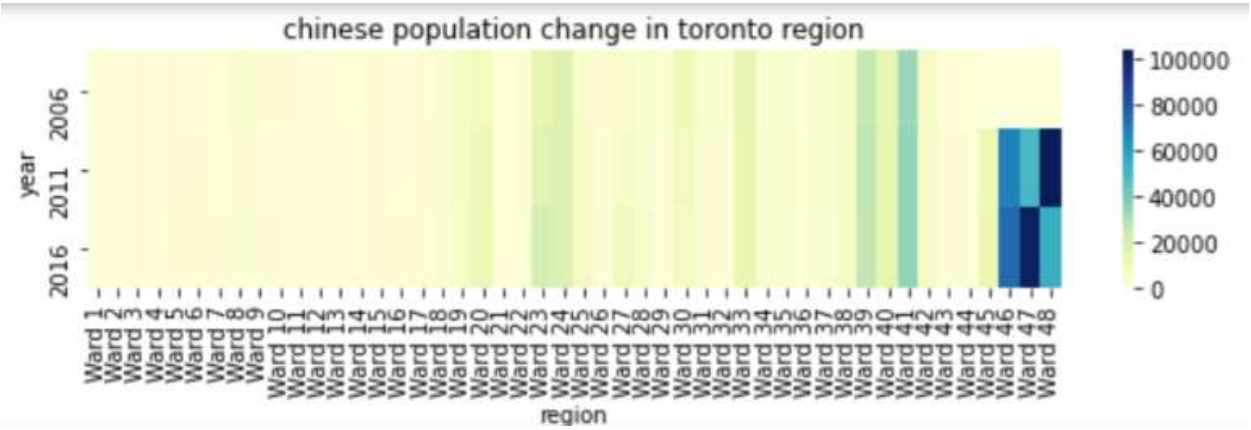
year	Ward 1	Ward 2	Ward 3	Ward 4	Ward 5	Ward 6	Ward 7	Ward 8	Ward 9	Ward 10	...	Ward 39	Ward 40	Ward 41	Ward 42	Ward 43	Ward 44	Ward 45	Ward 46	Ward 47	Ward 48
2006	1125.0	665.0	1530.0	900.0	1250.0	1070.0	1145.0	3805.0	1645.0	1480.0	...	26315.0	16000.0	33980.0	6935.0	1105.0	2085.0	0.0	0.0	0.0	0.0
2011	665.0	610.0	1410.0	885.0	1360.0	1230.0	1110.0	2545.0	1470.0	1170.0	...	27305.0	15390.0	35685.0	6180.0	1260.0	1950.0	11900.0	68780.0	50450.0	103840.0
2016	800.0	565.0	1365.0	935.0	1745.0	1715.0	1345.0	2100.0	1360.0	1620.0	...	26995.0	15185.0	35505.0	5955.0	1395.0	2000.0	12780.0	74625.0	102130.0	54425.0

After calculation, we see that the total Chinese population in Toronto is in an increasing trend in recent years, which is a good sign for launching new business.

year	total
2006	962140
2011	1879760
2016	1951620

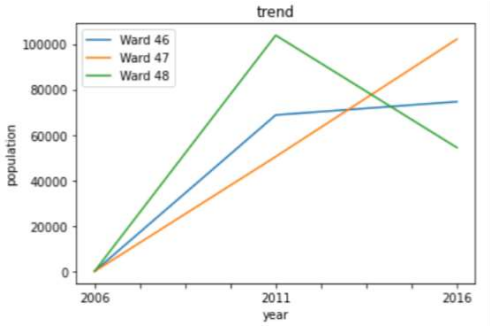


To make these data more readable, I plot them in a heat map, which clearly shows which region has the largest Chinese population.



Obviously the last 3 wards are more populated than other places. We will select these 3 wards as candidates. Let's explore the data further by putting the data of these wards into another plot, in order to see the trend of population change.

year	Ward 46	Ward 47	Ward 48
2006	0.0	0.0	0.0
2011	68780.0	50450.0	103840.0
2016	74625.0	102130.0	54425.0



See in the data frame that 2006 data is missing, but it doesn't matter a lot. When referring back to the original data from Toronto portal, we know that Ward 46 is North York, Ward 47 is East York and Ward 48 is Scarborough. The plot above shows that in both North York and East York, the Chinese population is growing, and East York is having a rapid growth. While in Scarborough the Chinese population is decreasing sharply. Nevertheless, Scarborough still has a large Chinese population, so it is still shortlisted.

Part 4: Decide on area that is least competitive

Foursquare will be the main source of data for this step. In the previous step we get a list of wards with large and growing Chinese population. Using the geographical info of those wards, we put it into Foursquare and explore those areas.

We use geolocator to decide the ward centers' latitude and longitude. According to the acreages of the shortlisted wards, the searching radius of North York and Scarborough is set to be 5 km while that for East York 2 km.

All Asian markets are considered as competitors. So when doing research on Foursquare, the searching keyword is set to be "Asian Market". Below are the 3 dataframes and corresponding maps showing the competitors' information in each ward.

North York:

	name	categories	lat	lng	labeledLatLngs	distance	cc	country	formattedAddress	address	postalCode	city	state	crossStreet	neighborhood	id
1	Whole Foods Market	Grocery Store	43.760662	-79.411044	[{'label': 'display', 'lat': 43.760662201105053...	3141	CA	Canada	[4771 Yonge Street, North York ON M2N 5M5, Can...	4771 Yonge Street	M2N 5M5	North York	ON	NaN	NaN	57f97b2d67c8ac5e8b34dd9
2	Asian Gourmet	Asian Restaurant	43.795940	-79.422880	[{'label': 'display', 'lat': 43.79594, 'lng':...	3089	CA	Canada	[Centrepoint, Toronto ON, Canada]	Centrepoint	NaN	Toronto	ON	NaN	NaN	4c4743d576d71d7fad8c0c4d
7	Asian Course	Asian Restaurant	43.811346	-79.454070	[{'label': 'display', 'lat': 43.81134645979900...	6359	CA	Canada	[Vaughan ON, Canada]	NaN	NaN	Vaughan	ON	NaN	NaN	4c8a5f1817972d6a0afc6866
11	Coppa's Fresh Market	Supermarket	43.779956	-79.469547	[{'label': 'display', 'lat': 43.77985629281025...	3307	CA	Canada	[4750 Dufferin St (at Martin Ross Ave), North ...	4750 Dufferin St	M3H 5S7	North York	ON	at Martin Ross Ave	NaN	4b4a5a41f964a520ed8426e3
18	Sensory Fresh Market	Market	43.718137	-79.420909	[{'label': 'display', 'lat': 43.718137, 'lng':...	4344	CA	Canada	[Canada]	NaN	NaN	NaN	NaN	NaN	NaN	5766c802496e92581f3ad72af
19	DPY Mama's Food Market	Grocery Store	43.737232	-79.436057	[{'label': 'display', 'lat': 43.737232000504916...	2161	CA	Canada	[322 Wilson Ave (Collinson Blvd), North York O...	322 Wilson Ave	M3H 1S8	North York	ON	Collinson Blvd	NaN	5d50879e02acfd0009bee9e6
20	Bathurst Village Market	Supermarket	43.764063	-79.445984	[{'label': 'display', 'lat': 43.764063, 'lng':...	3319	CA	Canada	[Toronto ON M2R 1Z1, Canada]	NaN	M2R 1Z1	Toronto	ON	NaN	NaN	5c29f5b4ab42a9002cd3dc63
23	Head Research (market research)	None	43.767594	-79.412041	[{'label': 'display', 'lat': 43.767594189096105...	3326	CA	Canada	[5075 Yonge St, Toronto ON, Canada]	5075 Yonge St	NaN	Toronto	ON	NaN	NaN	4d000b1121aa6aaf80cca3f9f

Scarborough (dataframe is abridged):

	name	categories	address	crossStreet	lat	lng	labeledLatLngs	distance	cc	city	state	country	formattedAddress	postalCode	neighborhood	id
4	Asian Specialty	Vietnamese Restaurant	2301 Brimley Rd	at Huntingwood Dr.	43.797819	-79.270737	[{'label': 'display', 'lat': 43.797819, 'lng'...	2944	CA	Toronto	ON	Canada	[2301 Brimley Rd. (at Huntingwood Dr.), Toront...	M1S 3B8	NaN	4f79f9c0e4b0f60afe4f9f1d
5	Asian Gourmet	Chinese Restaurant	Lawrence	Markham	43.758263	-79.227862	[{'label': 'display', 'lat': 43.75826275564416...	2918	CA	Toronto	ON	Canada	[Lawrence (Markham), Toronto ON, Canada]	NaN	NaN	4c14fee382a3c0b60009fef8
7	Jocelyn's Asian Cuisine	None	NaN	NaN	43.765086	-79.260463	[{'label': 'display', 'lat': 43.765086, 'lng'...	1354	CA	NaN	NaN	Canada	[Canada]	NaN	NaN	4f021269a60e45461fab5ea7
12	Edgecombes Marketing And Promotions (warehouse)	Office	2250 Midland Ave, Toronto, ON M1P 3B6	NaN	43.775696	-79.274705	[{'label': 'display', 'lat': 43.77569613234338...	1391	CA	Toronto	ON	Canada	[2250 Midland Ave, Toronto, ON M1P 3B6, Toront...	M1P 3B6	NaN	4f819561e4b02b9583ed08a7
20	Kostas Meat Market	Greek Restaurant	259 Ellesmere Rd	Warden Ave	43.760605	-79.301830	[{'label': 'display', 'lat': 43.76060511487278...	3804	CA	Toronto	ON	Canada	[259 Ellesmere Rd (Warden Ave), Toronto ON M1R...	M1R 4L4	NaN	4cd5fd3c194845c6a2b1aebb1
21	Adlytic	Advertising Agency	2190 Ellesmere Road	NaN	43.777470	-79.230201	[{'label': 'display', 'lat': 43.77746995819856...	2269	CA	Toronto	ON	Canada	[2190 Ellesmere Road, Toronto ON M1G 3M5, Canada]	M1G 3M5	Scarborough	5e4ccbb7e1617b6f0085ca33d
23	Marketplace Ministries	Church	1775 Midland Ave	Midland And Ellesmere	43.762053	-79.270091	[{'label': 'display', 'lat': 43.76205334249128...	1576	CA	Scarborough	ON	Canada	[1775 Midland Ave (Midland And Ellesmere), Sca...	M1P 3P2	NaN	4f3fc0b2e46b0ae06556b2e9f

East York (dataframe is abridged):

	name	categories	address	crossStreet	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress	id
1	raise the root organic market	Fruit & Vegetable Store	1164 Queen Street east	Jones Avenue	43.668100	-79.309304	[{'label': 'display', 'lat': 43.66618055866422...	315	NaN	CA	Toronto	ON	Canada	[1164 Queen Street east (Jones Avenue), Toront...	54b1a7bad98a8d9a417cc8bd
2	Beaches Wholesome Market	Health Food Store	2234 Queen St E	Beech Ave	43.672661	-79.287567	[{'label': 'display', 'lat': 43.67296063441251...	1648	M4E 1G2	CA	Toronto	ON	Canada	[2234 Queen St E (Beech Ave), Toronto ON M4E 1...	4c547fd8f95d13a5a2c5afe
3	Toby's Food Market	Convenience Store	86 Kingston Rd	NaN	43.669536	-79.311346	[{'label': 'display', 'lat': 43.6695.56, 'lng':...	348	M4L 1G6	CA	Toronto	ON	Canada	[86 Kingston Rd, Toronto ON M4L 1G6, Canada]	4da1dec12f09b1f7a2fefd58
4	Leslieville Farmers Market	Farmers Market	NaN	NaN	43.664801	-79.319784	[{'label': 'display', 'lat': 43.664901396008562...	1088	NaN	CA	Toronto	ON	Canada	[Toronto ON, Canada]	4deb0ba68874880e3387c0c
5	The Big Carrot Natural Food Market	Health Food Store	125 Southwood Dr	NaN	43.678879	-79.297734	[{'label': 'display', 'lat': 43.678876, 'lng':...	1380	M4E 0B8	CA	Toronto	ON	Canada	[125 Southwood Dr, Toronto ON M4E 0B8, Canada]	4ad4c062f964e52011f620e3
7	Farmer's Fruit Market	Farmers Market	NaN	NaN	43.673050	-79.310514	[{'label': 'display', 'lat': 43.67304887074939...	1107	NaN	CA	NaN	NaN	Canada	[Canada]	4f0f9f06a4b00d8da7d60d11
8	The Dog Market	Pet Store	2116 Queen St E	Hammersmith Ave	43.671163	-79.294773	[{'label': 'display', 'lat': 43.67116322361954...	1044	M4E 1E2	CA	Toronto	ON	Canada	[2116 Queen St E (Hammersmith Ave), Toronto ON...	4ea34f63f93ad512d8d0f511c
11	The Big Carrot Natural Food Market	Grocery Store	125 Southwood Dr	NaN	43.678730	-79.297478	[{'label': 'display', 'lat': 43.67873, 'lng':...	1378	M4E 0B8	CA	Toronto	ON	Canada	[125 Southwood Dr, Toronto ON M4E 0B8, Canada]	5b52eb3501bc5a002ce53ebb
12	Rowan Homespun Market	Market	2196 Queen St E	Balsam Ave	43.672218	-79.289780	[{'label': 'display', 'lat': 43.67221832275390...	1463	M4E 1E6	CA	Toronto	ON	Canada	[2196 Queen St E (Balsam Ave), Toronto ON M4E ...	51ac0cb1496e8f910f1d79aa

By using dataframe.shape we can conclude that North York has 8 competitors, Scarborough has 8, and East York has 16. The number of competitors is not representing the full picture, because different ward has different population. The number will be meaningful only if we factor in the population. That is, divide the population by competitor number to see how large is the population each competitor cover, which we will name it as population coverage in the following part. The larger is the population coverage, the larger potential is the market, because that means each potential customer gets less choice.

Now let's see what happens if we do the population coverage calculation for the 3 shortlisted wards:

	ward	population	competitors number	population coverage per competitor
0	North York	74625	8	9328.125
1	Scarborough	54425	8	6803.125
2	East York	102130	16	6383.125

The above dataframe lists 3 wards in descending order for their population coverage, with North York having the biggest population coverage.

Part 5: Decide on area with lightest leasing cost

Apart from the population coverage, Mr. Chen also wants to know the leasing cost, and uses both factors to determine the location choice.

CBRE will be the data source for getting this information. CBRE offers clear data report about business rental information for the Toronto area. Here is a snapshot showing the leasing costs of the 3 wards:

(https://www.cbre.ca/en/research-and-reports/Toronto-Office-MarketView-Q1-2020)

The corresponding leasing price for the 3 wards are included in the following chart with the population coverage information:

	ward	population coverage per competitor	leasing cost
0	North York	9328.125	21.56
1	Scarborough	6803.125	12.20
2	East York	6383.125	12.52

Part 6: Modelling and Testing

Now we have the information that Mr. Chen requires at hand. We can start our analysis.

The final answer we want to get is whether the business will be successful or not, by analyzing the information we got. This is very similar to a prediction model in machine learning. Hence classification will be the method we use. We will try to use data features to predict the business in each ward will be successful (labelled as Y) or not (labelled as N).

We found some historical data for supermarkets running in Toronto areas, which can be used to train the machine learning model. We use train-split-test to get the best model and then apply the model to real data for prediction.

Here is the data for modelling:

	Name	population coverage per competitor	leasing cost	Success or not
0	Market 1	10500	27	N
1	Market 2	7500	20	N
2	Market 3	6800	10	Y
3	Market 4	6600	9	Y
4	Market 5	7700	15	Y
5	Market 6	8000	13	Y
6	Market 7	9900	28	N
7	Market 8	11000	24	Y
8	Market 9	9800	19	Y
9	Market 10	8800	23	N

For the convenience of modelling, we change the Y/N values into 1/0 as follow:

	Name	population coverage per competitor	leasing cost	Success or not
0	Market 1	10500	27	0
1	Market 2	7500	20	0
2	Market 3	6800	10	1
3	Market 4	6600	9	1
4	Market 5	7700	15	1
5	Market 6	8000	13	1
6	Market 7	9900	28	0
7	Market 8	11000	24	1
8	Market 9	9800	19	1
9	Market 10	8800	23	0

We use 80% of the data to train the model and 20% to test it. We process the “competitor” and “leasing cost” into numpy array as X, “success or not” as Y. KNN will be the method for use. We then use accuracy metrics to test how accurate it is. Setting K=2, we got the following result:

```
[51]: from sklearn.neighbors import KNeighborsClassifier
      #Start testing by using k as 2.
      k = 2
      #Train Model and Predict
      neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,Y_train)
      neigh

      /home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/ipykernel_laun
      the shape of y to (n_samples, ), for example using ravel().
        """

[51]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                           metric_params=None, n_jobs=None, n_neighbors=2, p=2,
                           weights='uniform')

[52]: #We use the model to predict the test result.
      Yhat = neigh.predict(X_test)
      Yhat

[52]: array([1, 0])
```

Yhat is 1 and 0, by referring back to original data, the prediction is correct. Accuracy for train set is 0.875 and test set 0.5.

The result looks nice, but we want to find the best model, so we try to test every K value possible, from 1 to 9. The calculation tells us that the best K is 1, when train set accuracy is 1 and test set 0.5. Then we redo the model with k=1 and apply the model to actual data of the 3 shortlisted wards.

```
[57]: #so the best k is 1. Redo the model.
      k = 1
      neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,Y_train)
      neigh

      /home/jupyterlab/conda/envs/python/lib/python3.6/site-packages/ipykernel_launcher.py:3: Da
      the shape of y to (n_samples, ), for example using ravel().
        This is separate from the ipykernel package so we can avoid doing imports until

[57]: KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                           metric_params=None, n_jobs=None, n_neighbors=1, p=2,
                           weights='uniform')

[58]: #imput the data that we need to predict for the 3 wards.
      X1=df[['population coverage per competitor','leasing cost']].values
      X1=preprocessing.StandardScaler().fit(X1).transform(X1.astype(float))

[59]: #now let's see using the KNN model we have how's the result for each ward.
      Y1hat = neigh.predict(X1)
      Y1hat

[59]: array([0, 1, 1])
```

It gives a final result as an array [0,1,1] for the 3 wards.

Part 7: Result

We transform that Boolean values back into Y/N values and integrate them into the data frame:

	ward	population coverage per competitor	leasing cost	Success or not
0	North York	9328.125	21.56	N
1	Scarborough	6803.125	12.20	Y
2	East York	6383.125	12.52	Y

The analysis for the 3 wards has a preliminary conclusion:

To open a Chinese market in the 3 wards, North York will be not successful, Scarborough and East York will be successful.

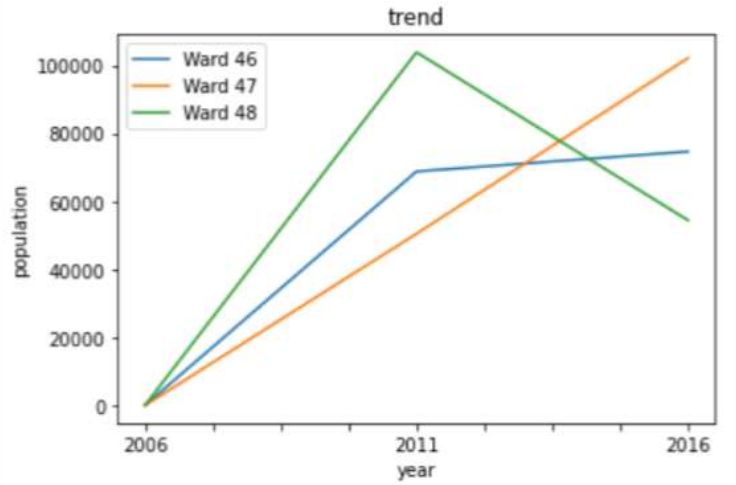
Part 8: Discussion

Unfortunately, North York fails the test. The leasing cost is too high that the population coverage cannot justify it. For the other 2 wards, does it mean that both Scarborough and East York have the same advantages for opening a Chinese market?

Of course there is difference. First we take a look at the straightforward information showing by charts and plots.

In the result chart in Part 7, we can see that in population coverage, Scarborough slightly wins over East York (6803 over 6383, the difference is about 5%). For the leasing costs, too, Scarborough wins by 0.32 CAD per square foot (12.20 compared with 12.52 in East York, the difference is about 2%).

This is the current situation. However, opening a business needs to look into the future. As per the result we got from the population analysis part, we can see that the total population in Toronto is in an increasing trend, hence we suppose the population coverage and leasing cost will increase with it. However, something special about Toronto is that it is a diverse city with people of different ethnic origins tend to live in the same region, so we have China town, little Portugal, little Italy etc. There is a trend for the region population change, i.e., if there are people starting leaving a certain region to live in another region, the rest population may follow the trend. Now recall the following plot:



Just to remind that Ward 47 is East York and Ward 48 is Scarborough. We can see that the Chinese population in Scarborough is decreasing sharply by 47.5%! While that in East York is growing by 102%! We can conclude that Chinese immigrants are leaving Scarborough and more like to settle down in East York.

Population is the most important feature we need to consider for potential market. Besides, when comparing with the population trend change, the difference of leasing costs and population coverage look small. Therefore, even though East York is slightly less advantageous to Scarborough in leasing cost and population coverage, the former's population trend benefit is much better than the latter. East York is a better choice.

Part 8: Conclusion

Weighing in the population trend, population coverage and leasing cost, the final winner is:

East York!!!