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**.

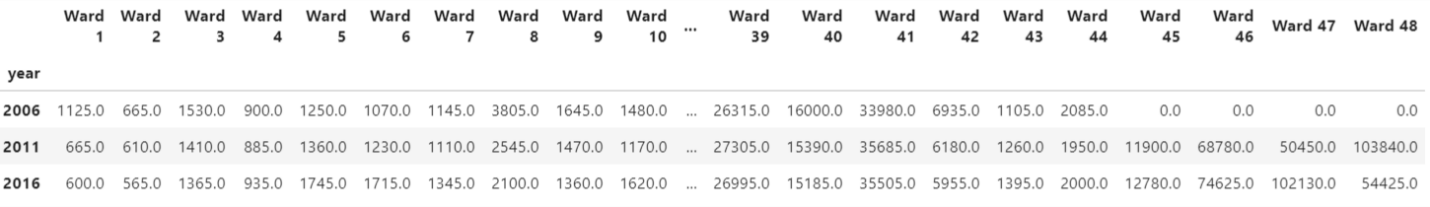
The whole research will be carried out in 4 main steps.

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.

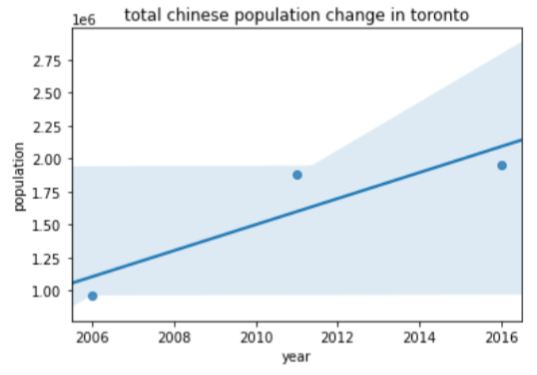
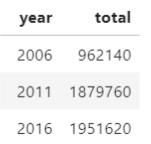
* **Part 3: Finding areas most populated by Chinese immigrants**

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.

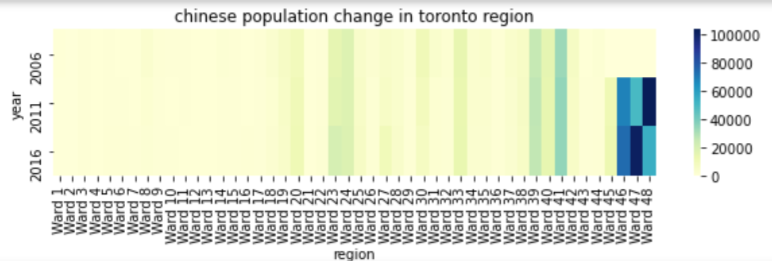
There are 48 wards in Toronto. Data for 3 years are found: 2006, 2011, and 2016. By adding up all related Chinese language we got a total population in each ward, and then putting the total population of each ward of those 3 years into one data frame. The result is as follow:



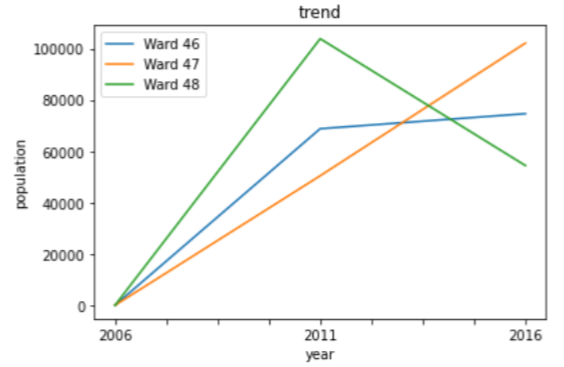
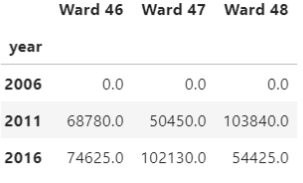
After calculation, we see that the total Chinese population in Toronto is in an increasing trend in recent years, which is a good sign for launching new business.



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.



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**

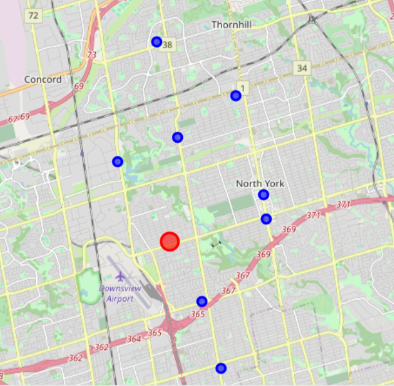
Foursquare will be the main source of data for this step. In the previous step we get a list of wards with large and growing Chinese population. Using the geographical info of those wards, we put it into Foursquare and explore those areas.

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.

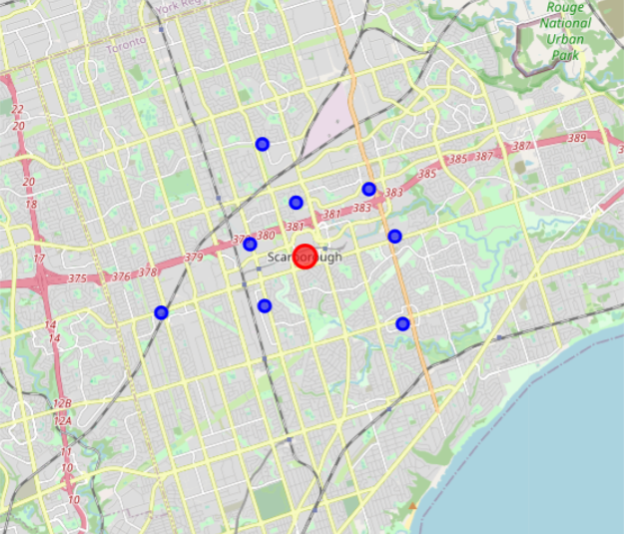
All Asian markets are considered as competitors. So when doing research on Foursquare, the searching keyword is set to be “Asian Market”. Below are the 3 dataframes and corresponding maps showing the competitors’ information in each ward.

North York:

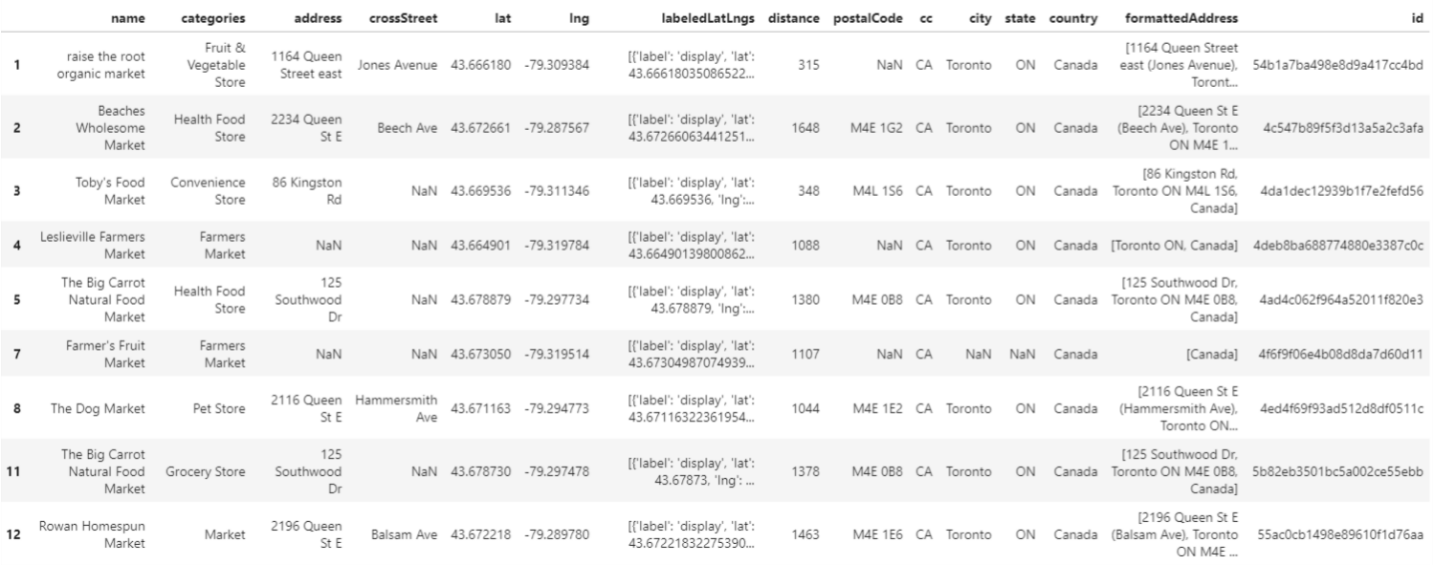




Scarborough (dataframe is abridged):

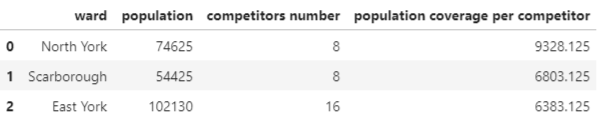


East York (dataframe is abridged):



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:

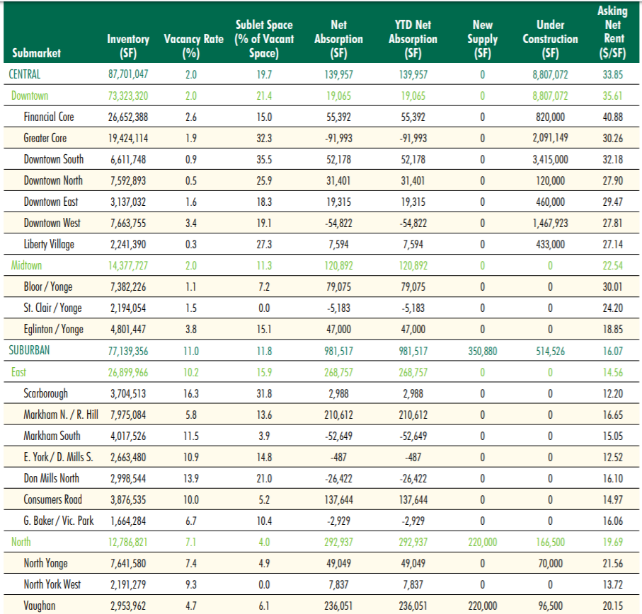


The above dataframe lists 3 wards in descending order for their population coverage, with North York having the biggest population coverage.

* **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:



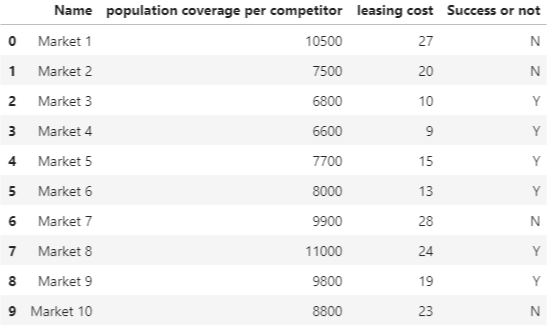
* **Part 6: Modelling and Testing**

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

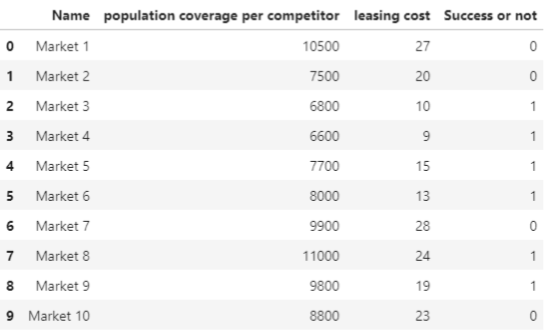
The final answer we want to get is whether the business will be successful or not, by analyzing the information we got. This is very similar to a prediction model in machine learning. Hence classification will be the method we use. We will try to use data features to predict the business in each ward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



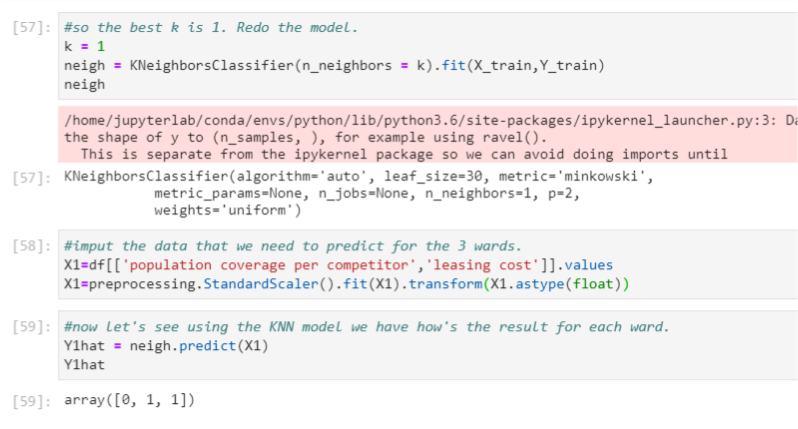
For the convenience of modelling, we change the Y/N values into 1/0 as follow:



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:  


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.

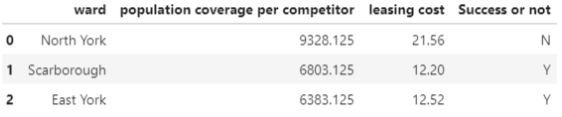
The result looks nice, but we want to find the best model, so we try to test every K value possible, from 1 to 9. The calculation tells us that the best K is 1, when train set accuracy is 1 and test set 0.5. Then we redo the model with k=1 and apply the model to actual data of the 3 shortlisted wards.



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:



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.

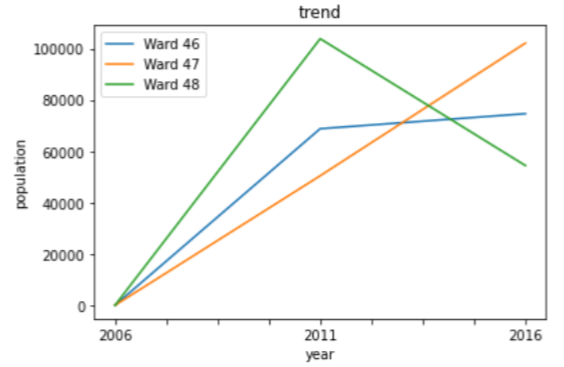
* **Part 8: Discussion**

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!!!