# Data Science Project Report

**Finding the best location for new business**

By Sam Zhang

2019-07-31

## Introduction

Toronto is an interesting city with diversity. I’ve been to it twice and enjoyed both times. The Chinese food is fascinating. As I almost chose to live there, I wonder if I were given the chance to start a Chinese restaurant, how do I choose the best location?

This actually is a very practical question, and potentially useful to anyone who are looking to invest.

## Data

My solution will be based on the following data (for now):

1. Defined location areas in Toronto. Will use neighborhood list scraped off a web Site (https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M).
2. Geo location data of the neighborhoods. For this I will use geolocator to find center of each neighborhood
3. Chinese restaurant data, with location data. For this I will use Foursquare Venue service to list Chinese restaurant in that area, then use Foursquare Venue detail service to get details of it. For version one will just get the createdAt property and use that as the business start date.

Combine the above data, we get a list of Chinese restaurants, each with its neighborhood, location, and date business started.

I collected 559 Venue records of Chinese Restaurants in 209 neighborhoods, added geo location and “createdAt” data for each.

## Methodology

Overall I will take the experimental methodology. Start with something simple, and try to improve it with better models; version and document each model; compare and evaluate.

### Model Version 1

With version one, my idea is to find the hottest area for Chinese restaurants, and recommend it. For version one I will use density as the measure of hotness. In plain words, simply count the number of Chinese restaurants in each neighborhood.

The advantage of this approach is its simplicity but is likely practically effective.

The disadvantage is that it may be too simple and not being logically 100% sound. Imagine, if everyone choose next location based on this, then all new future business will be opened in the same area. We all know that cannot be right.

Version turned out not successful, but it is a starting point to build up on. Details follow.

### Model Version 1 Validation

The validation is done in the following way:

1. Step 1: start with the full list of “Venue” records for all the Chinese Restaurants
2. Step 2: in this dataset, find out the maximum “createdAt” value, thus the “Venue” record with the maximum “createdAt” value. The “neighborhood” name of that Venue will be the result of the training sample. Let’s call it ***venue\_Y*** and ***neighborhood\_Y.***
3. Step 3: re-filter the dataset to be rid of the Venue with the maximum “createdAt” value. This would be the existing restaurants just before the ***venue\_Y*** was opened in ***neighborhood\_Y***. Let’s call this reduced dataset ***dataset\_X***.
4. Step 4: follow the version 1 methodology, find a recommended ***neighborhood\_y*** based on ***dataset\_X.***
5. Step 5: use this modified dataset (***dataset\_X***), repeat steps 2-4.
6. Complete validation on 100 or 200 recommendations, compare each pair of ***neighborhood\_y*** of ***neighborhood\_Y***, and calculate accuracy score.

## Results section

The result of validation: accuracy being 12% on 200 samples. Have to admit it was far from sufficient.

What I learned from this is that I will need to add more feature(s), involve more data, improve the model, etc. – which will be improvements in future versions.

## Discussion section

Lack of data is probably a major reason in the building of the unsuccessful version 1 model. Such decision-making is a complex brain process. It cannot be simply based on one factor - like finding the most crowded area.

In version 1, I was only able to use one feature – the established hotness of the area - out of the many factors that affect human decision making on where to open the business. It would be helpful to have demographic data, neighborhood polygon data, etc.

The more features the better. With all features available, AI model should be able to find what human cannot see directly.

## Conclusion section

So with this project I experimented with AI in real-world decision making. Although the result is not satisfying, the experience is good. And I have ideas of improvements.

So keep working, adding data, adding features, improve processes, revise ideas, etc.