

Capstone Project – Instructions

Now that you have been equipped with the skills and the tools to use location data to explore a geographical location, over the course of two weeks, you will have the opportunity to be as creative as you want and come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve. If you cannot think of an idea or a problem, here are some ideas to get you started:

1. In Module 3, we explored New York City and the city of Toronto and segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. One interesting idea would be to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto or Paris or some other multicultural city? I will leave it to you to refine this idea.
2. In a city of your choice, if someone is looking to open a restaurant, where would you recommend that they open it?

For this week, you will be required to submit the following:

1. A description of the problem and a discussion of the background. **(15 marks)**
2. A description of the data and how it will be used to solve the problem. **(15 marks)**

Part 1: Description of the problem and background discussion

Introduction

As Asians from the far east (e.g. Hong Kong, Singapore) with exposure to both Western and Asian culture, migrates to the U.S., a transferable skill would be in the food hospitality industry. It is quite safe to say that Asian cuisine is almost everywhere in the U.S. of A and a method to help provide some statistical data can prove useful to understand the existing market in a particular location.

Problem Description

For this project, we will analyse the greater Dallas area (or Dallas county) to derive the statistics for various cuisine for the purpose of understanding opportunities of starting up an Asian cuisine. The term Asian cuisine is used loosely here to cover the wide range of cuisines found in the greater Asia and can include fusion for variety and to increase the offering.

The definition of Asian cuisine encompasses a wide range of cooking practises and traditions and there is no enforcement on how the term is being used. Just in the Asia region alone, there is Chinese cuisine which varies greatly in taste and flavor at different locations, to Japan, countries in South Asia, and to western parts including India. As such, the definition of Asian cuisine will depend strictly on how it is defined by the source of the data.

For this exercise, Dallas was pick for the study for the following reasons:

1. It has a large white racial makeup, about 50% and while Asian constitutes about 6%.
2. It has mild weather with summer in the mid 30s and its dry, while winter is in the teens with at worse 1 inch of snow or so.
3. While it has issues with tornado, the inland is less vulnerable to
4. The current unemployment rate is about 10% below the nationwide numbers and it is at its lowest since 2002.

In this exercise, we will attempt to understand the statistics around the frequency of various categories of venues and the ratio to the Asian venues for the cities that falls within the Dallas greater area.

Part 2: Data description & how it will be used to solve the problem

Data Analytics

Data sources and description of the data

There are two sources identified

1. It is logical to assume that neighborhoods/towns/cities are not homogeneous. They will have different demographics profile, different distribution of wealth, infrastructure, etc ... Hence, a logical method would be to analyse by postal codes or zip codes as they are called in the U.S. Hence, the first data set will be Dallas Zip codes, along with their population. The name Dallas can refer to a county, in which there are cities and towns as listed below:

1. Addison
2. Balch Springs
3. Carrollton
4. Cedar Hill
5. Coppell
6. Dallas
7. Desoto
8. Duncanville
9. Garland
10. Grand Prairie
11. Hutchins
12. Irving
13. Lancaster
14. Mesquite
15. Richardson
16. Rowlett
17. Sachse
18. Seagoville
19. Sunnyvale
20. Wilmer

Each zip code will have a corresponding latitude and longitude value assigned.

Unfortunately, a city, by virtue of its size, can have multiple zip codes, and the results have to be grouped by city/town. For the purpose of aligning the nomenclature, cities and towns will be called '**neighborhood**', and Dallas county will be referred to '**city**' per the the data from Foursquare.

2. The second data set will be the venue data. Details on the venues will be derived from Foursquare.com website via an API to the application. Foursquares provides a rough guide on the types of cuisine according to a predefined set of categories as documented on its website <https://developer.foursquare.com/docs/resources>. While it also returns the venues' frequency by neighborhoods which is defined by their zip codes and their respective latitude and longitude. This information can only be used as a rough guide as Foursquare returns the findings based on a specified radius from a given latitude and longitude. This already assumes that all neighborhoods are circular and of a fixed size with its latitude and longitude in the centre of the circle and it is not capable of limiting its search within the boundaries of a given city or town or neighbourhood. In this exercise, we will not attempt to

'scrub' the information coming back from Foursquare for duplicates, or venues returned that don't match the search criteria.

Analytical Methods

The basic "Demand and Supply" approach will be used. To achieve this:

1. Statistical analysis has to be performed on the types of venues and its frequency by neighborhoods.
2. Understanding of how the Asian cuisine market segments itself such as a generic Chinese restaurant vs a Sushi or Peking Duck restaurant. As Chinese forms the greatest land mass in Asia, and the highest population count, understanding will have to be derived from how Foursquare defines this.
3. Any correlation between Asian and non-Asian cuisine to derive acceptance rates and opportunities for growth.

Interpretation of results

1. If an area has a high volume of Asian cuisine with respect to the overall venues count, it could mean that the market is moving towards saturation
2. If an area has a very low count of Asian cuisine, with respect to the overall venues count, it could mean that the penetration rate is poor.
3. There could be potential correlation between population density, the various types of cuisine and their frequencies.
4. Again, the data cannot be interpreted strictly and it should only be used as guide