Exploring San Francisco Bay Area

-Date Science in Real Life

1. **Introduction**

This project is the final capstone project for the course Applied Data Science Capstone which is the last course of the IBM Data Science Professional Certificate program. In this project, I will use the data science skills and tools learned from the series of courses to use location data to explore a geographical location. The goal is to prove that I am capable of defining an idea or problem, looking for data in the web, using the Foursquare location data to explore different areas, effectively analyzing the data, and presenting the results.

**1.1 Description of the problem and a discussion of the background**

I live in a suburban city in the grand Houston area. My neighborhood is very diverse, including a large amount resident originally from China. However, when I am tired of cooking and want to dine in a Chinese restaurant, I have very limited choices in the nearby neighborhood. I have to drive 15-20 minutes (without traffic) away to get one. Based on the information from my neighbors and friends, there is actually a demand to the traditional Chinese food. Since I am equipped with skills and tools from this series courses, an idea to my mind is to explore the relation between the number of Chinese restaurants and the population distribution in an area. This idea can be used to explore the similar relation between other venues and the population distribution.

At the beginning I used Foursquare to search the venues information in my city but found out the search didn’t work well since some restaurants I know well are not searched out. I choose to explore the San Francisco Bay Area because it is a metropolitan region with 7 million inhabitants and more than 100 cities/town in the 7000 square miles of land. The high-density population supports a sufficient number of venues to use in this project. In addition, I don’t have a chance to visit the Bay Area yet and it is in my visit list in near future, so I’d like to get some impression of it in advance.

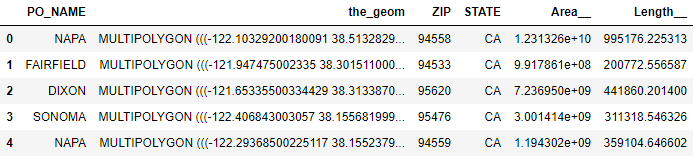
**1.2 Target audience**

Some data scientists might be interested in this project since it includes some real-life data and exploratory data analysis methods, and the data is able to tell a story out of it. Business personnel who want to invest a venue such as a Chinese restaurant, might be interested in this project too since its analysis results provide a comprehensive understanding on the relation between the number of a venue in an area and the population distribution. Some analysis is kind of easy to understand, such as that the number of Chinese restaurants is positively correlated to the Chinese population. Some others might be surprise.

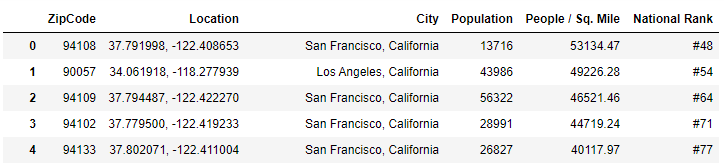
1. **Description of the data and how it will be used to solve the problem**

Three set of data are used in this project.

* 1. Bay Area ZIP Codes from <https://catalog.data.gov/dataset/bay-area-zip-codes>. The data was downloaded as a csv file. It includes zip codes, city names, geometry locations, area information. This data is used to limit the zip codes within the Bay Area.



* 1. Scraping the population distribution of California from <http://zipatlas.com/us/ca/zip-code-comparison/population-density.htm>. This data includes zip codes, geographic location (latitude, longitude), population of all zip codes in California.



* 1. Foursquare location date. Foursquare is a location technology platform offering business solutions and consumer products through a deep understanding of location. It gives the venues information in Bay Area including venue category, geographic location, etc.

