The Battle of the Neighborhoods

IBM Applied Data Science Capstone Project

Jack Perng

1. Introduction

Business Problem: If someone just obtained their California acupuncture license and resided in San Jose, where would be a good location to setup their practice?

The idea is to explore the San Jose neighborhoods, according to ZIP code, based on the follow factors that may be related to business growth:

(1) Population
(2) Per capita income
(3) Existing competition
(4) Crime rate
(5) Unemployment rate
(6) Bachelor degree percentage
(7) Median home price

The neighborhoods would be clustered via **k-means clustering**, analyzed and compared against one another. The goal is to identify, via the clustered results, potential neighborhood candidates to start an acupuncture clinic in.

2. Data

Data Sources: The data for the project would come from the various sources:

(1) **Kaggle dataset** "US Wages via Zipcode", containing the following information:
    a. U.S. ZIP codes
    b. Geographic coordinates
    c. Estimated Population
    d. Total Wages

    From this dataset, the relevant data subset for San Jose, CA can be extracted. In addition, the per capita income can be calculated from. (https://www.kaggle.com/pavansanagapati/us-wages-via-zipcode)

(2) **Foursquare location data**, which will be used to find and locate the acupuncturists in the San Jose area. The results are used as the metric for existing competition in the same ZIP code.

(3) **ADT Security Services**. The website contains an ADT Crime Map which provides the crime rate of each ZIP code. The "Total Crime" rate metric is selected.

(4) **City-Data**. The website provides the following data of interest. For population 25 and over:
    a. Unemployment rate
    b. Bachelor degree or higher percentage
(5) **Zillow** contains data on typical home prices based on ZIP code. The latest home prices at the time of completing this project, published on 05/31/2020, are used.

The combined information of all the data described above provide the input for the clustering algorithm model.

3. Methodology