Week 4 Homework

# Capstone Project:

# Identification of Best Location for a Medical Practice Using Public Health and Geographic Data

Dated: June 30, 2020

Class: Applied Data Science Capstone

William Sanborn

July 2020

## Introduction

The purpose of my project is to provide prioritized recommendations based on data analysis to a hypothetical doctor as to where he or she might want to set up his or her practice in the United States. These recommendations will consider both the market need for the doctor's services as well as that doctor's personal preferences. For purposes of this project, I am making the following assumptions:

- The hypothetical doctor is a cardiologist and, in order to ensure a successful practice, he/she would like to set up the practice in an area that has a significant need for heart related medical services.
- The cardiologist would also set like to set up his/her practice in an area that features as many areas of personal interest as possible. For purposes of this analysis, I am further assuming that:
    - The doctor enjoys sushi, wine, museums, and live music.
    - The doctor is an avid golfer.
    - The doctor has a dog that requires regular exercises.

In order to identify the best possible solution, I will use publicly available data sources and tools from Coursera to design and build a model that provides optimal recommendations for the hypothetical cardiologist.

## Data

To answer the problem posed in the introduction, I intend to design and build metrics that measure the following:

- The level of need in a city for heart related medical services.
- The attractiveness of the identified city based on the doctor's personal interests

To gauge the need for heart related services, I will use information from the Center for Disease Control's "500 City" database (https://chronicdata.cdc.gov/500-Cities/500-Cities-Coronary-heart-disease-among-adults-age/cqcq-r6f8/data). I will use this information to identify and/or design metrics that can be used to measure the level of heart health in each of the 500 cities. I will then identify the 25 US cities with the worst levels of heart health and assume that these are the cities with the greatest need for cardiologist services. The final product in this stage will be a dataframe incorporating the health information for these cities.

To measure how well these cities align to my doctor's personal interests, I will use the FourSquare API to build a database of all the venues in each of the 25 cities that I target. After building this dataframe, I will then extract the venues that align to the doctor's areas of interest. More specifically, I will pull the information/fields related for these types of venues:

- ‘Music Venue'
- 'Sushi Restaurant'
- 'Golf Course'
- 'Dog Run'
- 'Wine Bar'
- 'Museum'

I will use this subset of information to build an additional dataframe from which I can calculate metrics that reflect the frequency/concentration of each of these interesting venues in each of the targeted cities. After evaluating this information, I will determine a final metric that measures the potential level of interest that my hypothetical doctor might have to live in each city.

It should be noted the FourSquare API relies on GeoLocation coordinates (Latitude, Longitude) to identify nearby venues. As such, I will need to obtain these coordinates for the 25 cities that I intend to explore. I will use data from the following database for this purpose: https://public.opendatasoft.com/explore/dataset/1000-largest-us-cities-by-population-with-geographic-coordinates/table/?sort=-rank

Finally, all of the aforementioned information will be consolidated into a single dataframe that will be used to build the final data tables, Scatter charts and/or clustering maps that visually portray the requisite data and metrics. From these outputs, I intend to provide a recommendation of the optimum location(s) for the cardiologist's prospective medical practice based on market need and personal preferences.

## Data Sources

1) **Information on Health Issues By City**: https://chronicdata.cdc.gov/500-Cities/500-Cities-Coronary-heart-disease-among-adults-age/cqcq-r6f8/data

   Note: Measure: % Respondents aged ≥18 years who report ever having been told by a doctor, nurse, or other health professional that they had angina or coronary heart disease. https://www.cdc.gov/500cities/definitions/health-outcomes.htm

2) **Database of Geographic (Latitude/Longitude) Coordinates By City:** https://public.opendatasoft.com/explore/dataset/1000-largest-us-cities-by-population-with-geographic-coordinates/table/?sort=-rank

3) **Database(s) of City Venue Detail:** *FourSquare API:* https://developer.foursquare.com/