Where to open a Chinese restaurant

2. Data Description

a. Data sources

We could find Chinese-American population data at different cities in this Wikipedia page (https://en.wikipedia.org/wiki/List_of_U.S._cities_with_significant_Chinese-American_populations)
It has listed all the large-sized, medium-sized and small-sized cities with significant Chinese-American populations. I only used the large-sized and medium-sized cities, since for the small-sized cities (cities with a population fewer than 100,000) although the percentage of Chinese-American population might be relatively high, the absolute number would not be big enough.

After getting a list of all the large-sized and medium-sized cities with significant Chinese-American populations, I used geolocator to get the geographical coordinates of each city, and then used Foursquare to get the total number of Chinese restaurants in each city. Besides the number of available Chinese restaurants, I also considered livability in each city, and I got the AreaVibes Livability Score from https://www.areavibes.com .

After I decide that Cary in North Carolina is the best city to open a Chinese restaurant, I went on to find Cary's neighborhoods information on its city-data webpage http://www.city-data.com/city/Cary-North-Carolina.html and did more analysis about the neighborhoods using Foursquare data, one hot encoding and k-means clustering.

b. Data cleaning

As mentioned in the "Data Source" section, I used the population tables for large-sized and medium-sized cities in the Wikipedia page and combined them into one table. There was one city "Irvine" duplicated in both tables, and I removed before combining the two tables.

When dealing with the neighborhood information of Cary, I could only find a pure text containing all the neighborhood names in Cary on its city-data webpage http://www.city-data.com/city/Cary-North-Carolina.html. I replaced each 'neighborhood' word with symbols '","' (double quotes and comma), so that the pure text could be transformed to a list and then converted into a Pandas Dataframe.