2. Data Description

a. Data sources

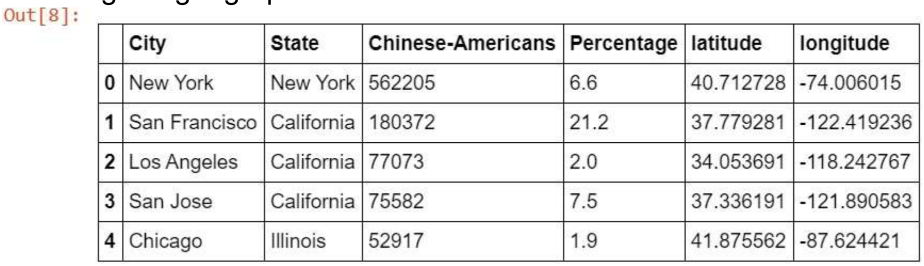
We could find Chinese-American population data at different cities in this Wikipedia page (https://en.wikipedia.org/wiki/List_of_U.S._cities_with_significant_Chinese-American_populations)
It has listed all the large-sized, medium-sized and small-sized cities with significant Chinese-American populations. 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

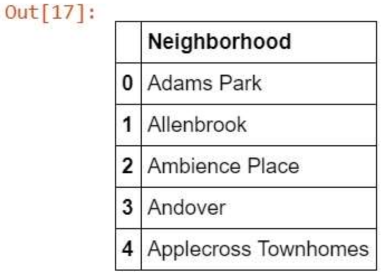
As mentioned in the "Data Source" section, I used the population tables for large-sized and medium-sized cities in the Wikipedia page, combined them into one table and sorted with population descendingly. There was one city "Irvine" duplicated in both tables, so I removed it before combining the two tables. The first five lines of the city data table including the geographical coordinates looks like this:



Out[8]:

	City	State	Chinese-Americans	Percentage	latitude	longitude
0	New York	New York	562205	6.6	40.712728	-74.006015
1	San Francisco	California	180372	21.2	37.779281	-122.419236
2	Los Angeles	California	77073	2.0	34.053691	-118.242767
3	San Jose	California	75582	7.5	37.336191	-121.890583
4	Chicago	Illinois	52917	1.9	41.875562	-87.624421

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, which looks like this:

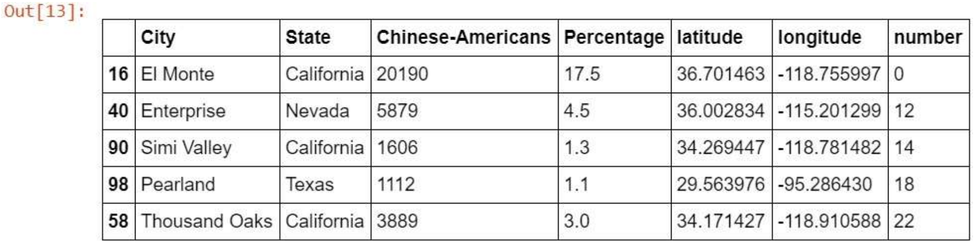


Out[17]:

	Neighborhood
0	Adams Park
1	Allenbrook
2	Ambience Place
3	Andover
4	Applecross Townhomes

c. Feature selection

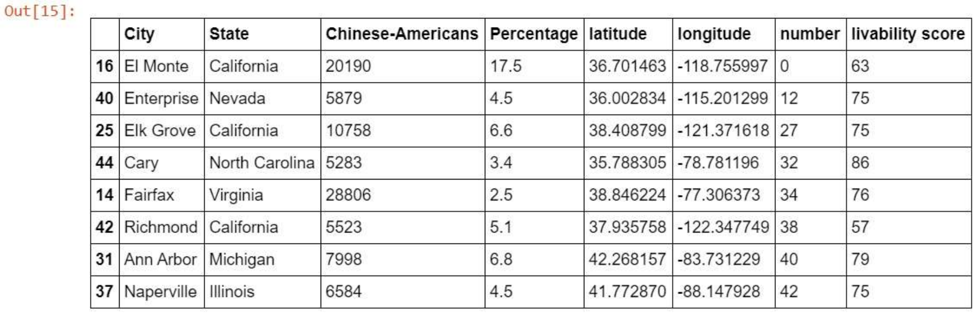
After getting the final city data table with geographical coordinates, I used Foursquare to get the total number of Chinese restaurants in each city. Here are the first five lines after sorting the cities using the available restaurants number ascendingly:



Out[13]:

	City	State	Chinese-Americans	Percentage	latitude	longitude	number
16	El Monte	California	20190	17.5	36.701463	-118.755997	0
40	Enterprise	Nevada	5879	4.5	36.002834	-115.201299	12
90	Simi Valley	California	1606	1.3	34.269447	-118.781482	14
98	Pearland	Texas	1112	1.1	29.563976	-95.286430	18
58	Thousand Oaks	California	3889	3.0	34.171427	-118.910588	22

I chose the cities that have relatively small number of Chinese restaurants (‘number’ < 50) but still a large Chinese-American population (index < 50, that is, the first fifty cities that have the largest number of Chinese-American population), then added the AreaVibes Livability Score for each city, and the final table looks like this:



Out[15]:

	City	State	Chinese-Americans	Percentage	latitude	longitude	number	livability score
16	El Monte	California	20190	17.5	36.701463	-118.755997	0	63
40	Enterprise	Nevada	5879	4.5	36.002834	-115.201299	12	75
25	Elk Grove	California	10758	6.6	38.408799	-121.371618	27	75
44	Cary	North Carolina	5283	3.4	35.788305	-78.781196	32	86
14	Fairfax	Virginia	28806	2.5	38.846224	-77.306373	34	76
42	Richmond	California	5523	5.1	37.935758	-122.347749	38	57
31	Ann Arbor	Michigan	7998	6.8	42.268157	-83.731229	40	79
37	Naperville	Illinois	6584	4.5	41.772870	-88.147928	42	75

Since “Cary, North Carolina” has the highest livability score among these cities, I chose Cary as our target city and continued to explore its neighborhoods (the data table of its neighborhoods has been shown in the section “b. data cleaning”).