The Battle of the Neighborhoods

IBM Applied Data Science Capstone Project

Jack Perng

07/2020

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.

Target Audience: The audience of the report will be recent licensed California acupuncturists in San Jose who are interested in opening up a clinic there. This report will provide information and insight on which neighborhoods might be good potential candidates.

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.
    (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 the data described above provides the input to the clustering algorithm model.

3. Methodology

3.1. Kaggle Dataset Preparation and Exploration

The raw imported dataframe contains the following columns of information:

	Zipcode	ZipCodeType	City	State	LocationType	Lat	Long	Location	Decommisioned	TaxReturnsFiled	EstimatedPopulation	TotalWages
0	705	STANDARD	AIBONITO	PR	PRIMARY	18.14	-66.26	NA-US-PR-AIBONITO	False	NaN	NaN	NaN
1	610	STANDARD	ANASCO	PR	PRIMARY	18.28	-67.14	NA-US-PR-ANASCO	False	NaN	NaN	NaN
2	611	PO BOX	ANGELES	PR	PRIMARY	18.28	-66.79	NA-US-PR-ANGELES	False	NaN	NaN	NaN
3	612	STANDARD	ARECIBO	PR	PRIMARY	18.45	-66.73	NA-US-PR-ARECIBO	False	NaN	NaN	NaN
4	601	STANDARD	ADJUNTAS	PR	PRIMARY	18.16	-66.72	NA-US-PR-ADJUNTAS	False	NaN	NaN	NaN

The cleaned-up data includes only entries from San Jose, CA.

ZIP codes that are not standard (P.O. Box and others) are also removed, in addition to those missing Estimated Population and Total Wages.

"Per capita wages" is computed by dividing TotalWages by EstimatedPopulation, and the additional column is added.

The final dataset looks like the following:

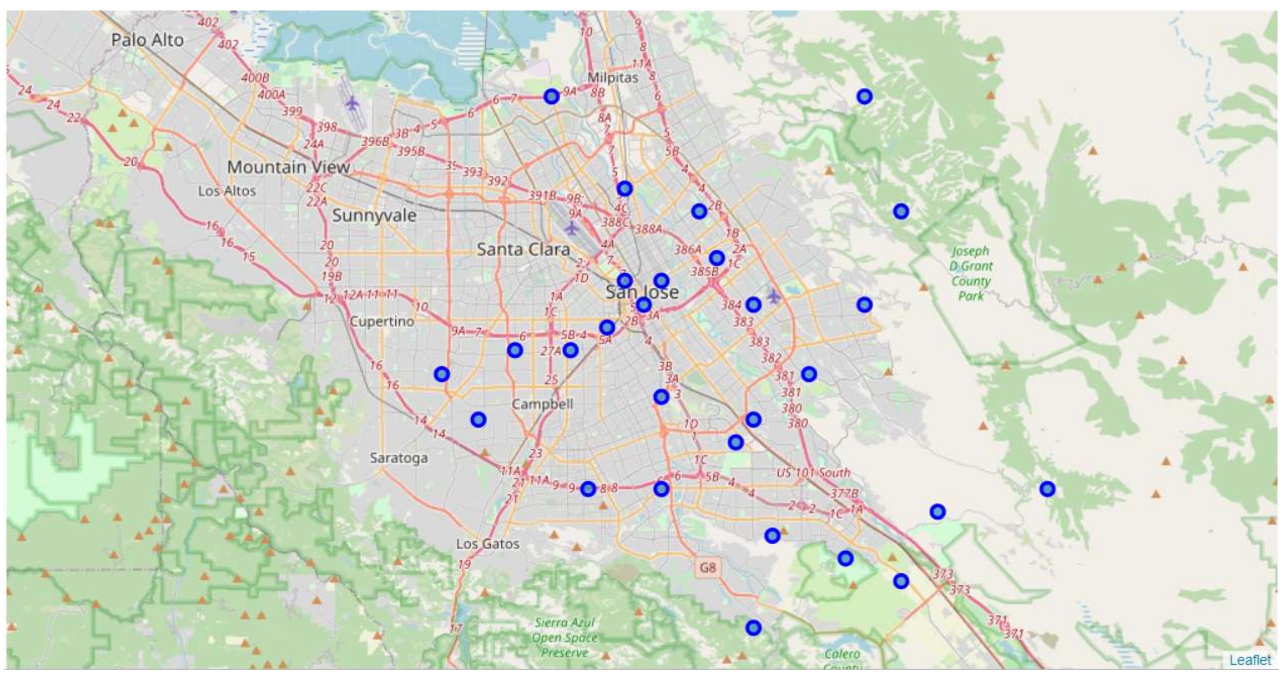
	Zipcode	Lat	Long	EstimatedPopulation	TotalWages	PerCapitaWages
0	95110	37.34	-121.90	12621.0	366468568.0	29036.412963
1	95111	37.28	-121.83	43578.0	866020686.0	19872.887374
2	95112	37.34	-121.88	34111.0	891795651.0	26143.931606
3	95113	37.33	-121.89	1049.0	37924110.0	36152.631077
4	95116	37.35	-121.85	35357.0	623888214.0	17645.394519

Inspecting the number of ZIP codes (neighborhoods), there are 28 of them. This ZIP code set will be used as the basis for subsequent collected data.

```
The are 28 ZIP codes

[95110, 95111, 95112, 95113, 95116, 95117, 95118, 95119, 95120, 95121,
95122, 95123, 95124, 95125, 95126, 95127, 95128, 95129, 95130, 95131,
95132, 95133, 95134, 95135, 95136, 95138, 95139, 95148]
```

Initial map plot of San Jose and the neighborhoods using Folium.



3.2. Foursquare Location Data Preparation and Exploration

Foursquare location data was utilized by an API call of venue search "**acupuncture**" near San Jose.

An estimated radius of 16000 m, roughly 10 mi, of San Jose was used.

A limit of 500 was passed as a parameter, but it became apparent afterwards that a Foursquare venue search limits its number of results to 50. This was a constraint that could not be circumvented.

The results were transformed into a dataframe and cleaned up by extracting the venue category name and keeping the relevant columns. A snapshot of the dataframe is shown below:

	name	categories	address	lat	lng	labeledLatLngs	distance	postalCode	cc	city	state	country	formattedAddress	crossStreet
0	Nurture Acupuncture	Acupuncturist	1520 The Alameda #130	37.335472	-121.915164	[{'label': 'display', 'lat': 37.33547199999999...	2177	95126	US	San Jose	CA	United States	[1520 The Alameda #130, San Jose, CA 95126, Un...	NaN
1	Charles Lin Acupuncture Clinic	Acupuncturist	475 N 1st St Ste 200	37.343643	-121.896553	[{'label': 'display', 'lat': 37.34364318847656...	983	95112	US	San Jose	CA	United States	[475 N 1st St Ste 200, San Jose, CA 95112, Uni...	NaN
2	Acupuncture Orthopedics & Natural Healing Center	Acupuncturist	259 Meridian Ave Ste 8	37.324388	-121.914624	[{'label': 'display', 'lat': 37.324388, 'lng':...	2500	95126	US	San Jose	CA	United States	[259 Meridian Ave Ste 8, San Jose, CA 95126, U...	NaN
3	Numo Acupuncture	Acupuncturist	1630 Oakland Rd Ste A110	37.381672	-121.894552	[{'label': 'display', 'lat': 37.3816716, 'lng'...	5075	95131	US	San Jose	CA	United States	[1630 Oakland Rd Ste A110, San Jose, CA 95131,...	NaN
4	1-2-3 Acupuncture Clinic (Santa Clara)	Acupuncturist	3700 Thomas Rd Ste 215	37.386284	-121.960762	[{'label': 'display', 'lat': 37.386284, 'lng':...	8345	95054	US	Santa Clara	CA	United States	[3700 Thomas Rd Ste 215 (San Thomas EXP), Sant...	San Thomas EXP

The data was processed further by keeping only the venues that are located in San Jose.

Furthermore, there are a few venues missing ZIP codes. This information was found through an internet search. The ones that are currently still practicing at the given location have this information filled in. Likewise, the ones who have moved or closed down their business are removed. In particular, one clinic has already moved to a different city.

Finally, the venues in each neighborhood (ZIP code) can be counted. The results are as follows:

postalCode	Zipcode	numbers
95110	95110	2
95112	95112	1
95117	95117	2
95120	95120	1
95121	95121	1
95122	95122	2
95123	95123	1
95125	95125	5
95126	95126	4
95128	95128	9
95131	95131	6

Note that out of the 28 ZIP codes in section 3.1, there are only 11 ZIP codes that contain non-zero entries. This is most likely due to the limit of 50 venue search results.

3.3. Other Data and Aggregation

The remaining data of crime rate, unemployment rate, bachelor degree percentage, and typical home prices, were tabulated manually in an Excel spreadsheet. A snippet of the imported data is shown below. The crime and unemployment rates are in percentages, while the Zillow median home price is in units of thousands of dollars.

	Zipcode	Crime_rate	Unemployment_Rate_25	Bachelor_Degree_Percentage	Median_Home_Zillow
0	95110	91	4.9	32.4	820
1	95111	32	6.1	19.6	774
2	95112	85	6.2	35.0	849
3	95113	86	4.5	72.8	763
4	95116	37	6.9	16.6	712

The final complete aggregated dataframe, with columns renamed, looks like the following:

	Zipcode	Lat	Long	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	37.34	-121.90	12621.0	29036.412963	2	91	4.9	32.4	820
1	95111	37.28	-121.83	43578.0	19872.887374	0	32	6.1	19.6	774
2	95112	37.34	-121.88	34111.0	26143.931606	1	85	6.2	35.0	849
3	95113	37.33	-121.89	1049.0	36152.631077	0	86	4.5	72.8	763
4	95116	37.35	-121.85	35357.0	17645.394519	0	37	6.9	16.6	712

3.4. Data Transformation and Modeling

Before feeding the data into model, the different columns have to be normalized because of their different scales. The columns of interest are EstPop, PerCapitaWages, Clinics, CrimeRate, UnemployRate, BSPercent, and HomePrice. The **Standard Scaler** was used to rescale the columns.

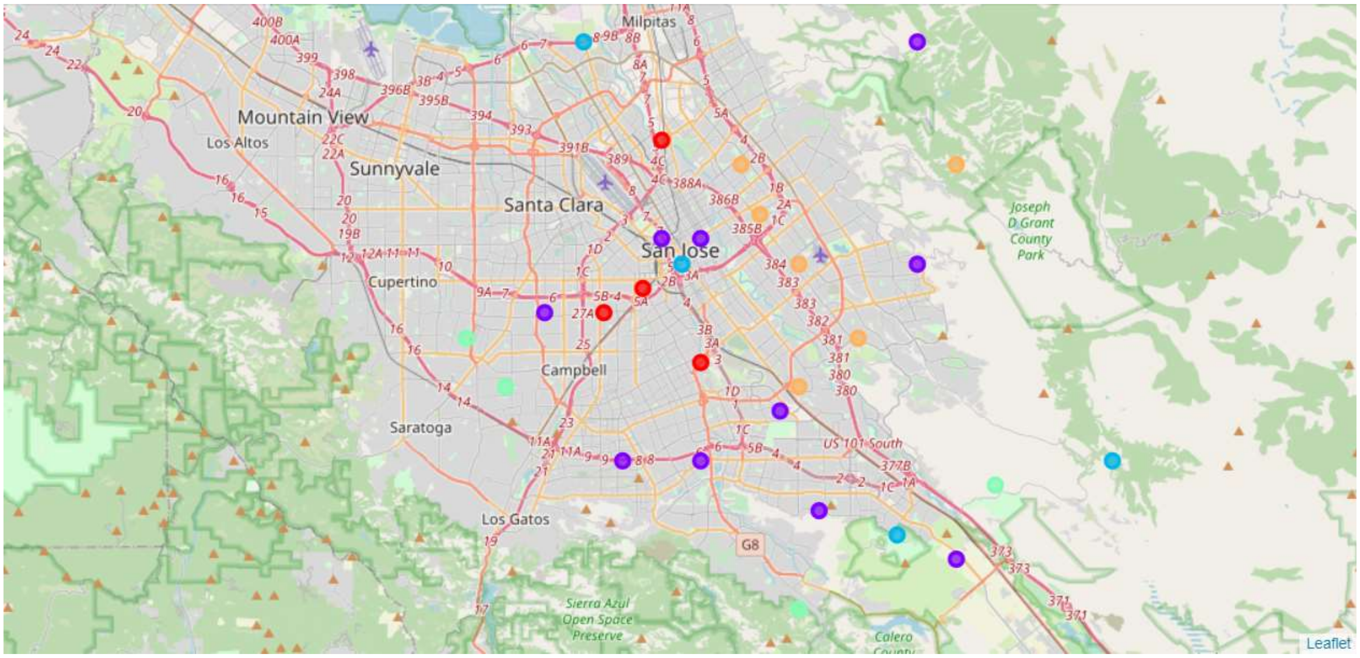
The transformed data is finally run through the k-means clustering algorithm with the number of clusters k set to 5.

4. Results

The clustered labels are added to the merged dataframe in Section 3.3, the beginning which is shown below.

	Zipcode	Cluster Labels	Lat	Long	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	1	37.34	-121.90	12621.0	29036.412963	2	91	4.9	32.4	820
1	95111	4	37.28	-121.83	43578.0	19872.887374	0	32	6.1	19.6	774
2	95112	1	37.34	-121.88	34111.0	26143.931606	1	85	6.2	35.0	849
3	95113	2	37.33	-121.89	1049.0	36152.631077	0	86	4.5	72.8	763
4	95116	4	37.35	-121.85	35357.0	17645.394519	0	37	6.9	16.6	712

The neighborhood map is re-drawn with the color-coded labels.



We begin by examining Cluster 1, which has the most neighborhoods.

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
0	95110	12621.0	29036.412963	2	91	4.9	32.4	820
2	95112	34111.0	26143.931606	1	85	6.2	35.0	849
5	95117	22030.0	30240.189696	1	21	5.5	45.5	1275
6	95118	26249.0	33438.090632	0	65	4.3	46.6	1128
11	95123	50481.0	32382.124086	1	67	4.0	41.1	946
12	95124	39234.0	38621.629938	0	69	4.4	53.0	1275
20	95132	34344.0	30705.994118	0	65	5.9	45.7	1151
24	95136	35078.0	33568.082074	0	68	4.7	46.7	961
26	95139	5634.0	36503.131523	0	40	6.2	47.7	904
27	95148	37541.0	30403.984470	0	61	5.7	39.0	1026

In this cluster, all the parameters such as per capita wages, unemployment rate, bachelor degree percentage, and home prices, are average compared to the other clusters, neither too high nor too low. The number of clinics here are very sparse.

Let's continue to examine Cluster 4, which has the 2nd most neighborhoods.

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
1	95111	43578.0	19872.887374	0	32	6.1	19.6	774
4	95116	35357.0	17645.394519	0	37	6.9	16.6	712
9	95121	30427.0	25115.081638	1	59	6.0	29.6	852
10	95122	41936.0	16719.609953	2	29	6.4	15.1	726
15	95127	46641.0	22820.174160	0	45	5.5	23.4	792
21	95133	20337.0	26582.617544	0	24	7.3	35.3	870

Cluster 4 does not appear to be wealthy, with lower per capita wages and home prices. The unemployment rate is higher, and bachelor degree percentage is low. The crime rate is low, a result that may run a bit counterintuitive. There is not a lot of existing competition, and the number of clinics is sparse as well.

Let's examine the remaining clusters, which have 4 neighborhoods each. We will start with Cluster 0

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
13	95125	41048.0	43425.502022	5	92	4.8	54.0	1331
14	95126	23076.0	36464.311362	4	93	4.8	52.0	1025
16	95128	25327.0	33020.557231	9	80	3.6	43.8	1159
19	95131	24403.0	37677.994919	6	55	4.4	55.9	1077

This is an affluent cluster, with high per capita wages and home prices. The unemployment rate is also low, as expected. The crime rate is rather high for this area. However, the biggest observation is the stiff competition, with a large number of existing clinics.

Next, we will examine Cluster 2.

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
3	95113	1049.0	36152.631077	0	86	4.5	72.8	763
7	95119	8171.0	35339.883368	0	128	4.2	41.5	953
22	95134	12670.0	51631.955722	0	92	3.0	75.8	951
23	95135	17221.0	42079.977063	0	121	4.7	61.3	1169

This cluster is wealthy and highly educated, with low unemployment rates. The home prices are not as high. There is also no competition in this area either. However, the recorded crime rate is high.

Lastly, let's look at Cluster 3.

	Zipcode	EstPop	PerCapitaWages	Clinics	CrimeRate	UnemployRate	BSPercent	HomePrice
8	95120	33486.0	50890.606821	1	52	3.1	71.3	1508
17	95129	32839.0	41750.644569	0	53	3.7	72.9	1751
18	95130	10841.0	36396.430311	0	64	3.7	55.7	1419
25	95138	15421.0	57789.241554	0	42	5.3	59.2	1213

This cluster appears to be the wealthiest, both in terms of wages and home prices. This cluster is also highly educated with high bachelor degree percentages, along with low unemployment rates and low crime rates. The existing competition is also very weak, with only 1 clinic in the entire cluster.

5. Discussion

From the results of the previous section, it is reasonable to avoid setting up a clinic in Cluster 0. The biggest drawback is the existing stiff competition, and the high saturation can be a challenge getting the business off the ground, in addition to future business growth.

Another cluster that I would advise against would be Cluster 4. Despite weak competition and low crime rate, the neighborhoods are not wealthy, with higher unemployment rates. This is not an ideal location to place your business in.

As for Cluster 1, whose attributes are average, these neighborhoods probably would be considered a safe bet and recommended over Clusters 0 and 4. However, it is also possible to do better by examining the results of Clusters 2 and 3.

Both are affluent and highly educated, with weak competition. The downside to Cluster 2 is the high recorded crime rate. The lower home prices may or may not be a reflection of that. It can also indicate that it is a neighborhood currently under growth.

Based on the findings of the data, we can conclude that Cluster 3 contains the neighborhoods with conditions most favorable to start an acupuncture clinic in.

6. Conclusion and Future Directions

This project provided me an opportunity to compile, analyze, and process demographic data of real neighborhoods. The work was done under the context of finding a suitable business location for an acupuncture clinic. The parameters I inspected were population, per capita wages, acupuncture clinics in the neighborhood, crime rate, unemployment rate (over 25), bachelor degree percentage (over 25), and median home prices.

This data was further analyzed using the k-means clustering algorithm, whose results helped provide insight as to which neighborhoods would be most ideal to setup an acupuncture practice.

There are a few things worth pointing out about the model that could also serve as future directions.

(1) The average commercial rent prices were not included in the model, which may be an important deciding factor. For example, a wealthier neighborhood may also charge higher commercial rent, which can be a deterrent for setting up a business there.
(2) As mentioned earlier, Foursquare location data only returns 50 results for a venue search. Based on personal experience residing in this area, the number of acupuncture clinics reported is lower than expected. This factor also affects the accuracy of the model.
(3) Using the total crime rate may not give the entire picture either. Violent crimes tend to discourage starting a business in the area, while the impact of lesser crimes can be much lower.

That said, the overall findings of this project prove to be valuable and provide a good starting point for further analysis.