The Battle of the Neighborhoods

Introduction

The client is interested in opening a yoga studio in the Sacramento County, California. According to 2010 census, the population of the county is 1,418,788 and are distributed in seven cities whose total area is 2570 km^2. The population of the county is diverse, consisting of neighborhoods of varying mean incomes, housing prices, infrastructures and so forth. To be profitable, the studio should be opened in an area with few other yoga studios and consisting of large population with relatively high income. Choosing a location to open a studio in such a large area consisting of varied demographic is a daunting task if done manually. To this end, various data analysis techniques will be employed to choose the best neighborhood for the studio to thrive.

The analysis used in this project will be of interest to individuals or institutions looking to find an ideal location to open a business.

Data

Various kinds of information is needed to suggest the best neighborhood to open the studio. The information about the latitude and the longitude of different zip codes in the Sacramento County will be downloaded from:

https://geo.nyu.edu/catalog/stanford-cm480wn0393

The information is available in the form of GeoJson file. The data can be imported into Python using **json** library.

Data of median incomes, housing prices, age and population of all the zip codes in the Sacramento County can be found in:

https://www.bestplaces.net/find/zip.aspx?st=CA&county=06067

The website will be scrapped using **BeautifulSoup** library to obtain the socio-economic data of all the zip codes in the Sacramento County. The name of the venues in different areas will be obtained from the **Foursquare** database.

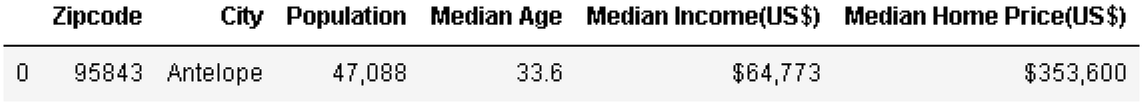
Methodology

The ideal place to open a yoga studio will be a location with relatively high population and median income but not so high median age. There are 54 zip codes in the Sacramento County and it will be quite a laborious task to scan the database for the proper zip code. The best way to obtain the ideal locations with these characteristics is to run the K-Mean clustering on standardized socio-economic database with the information on population, median age, median income and median home price. This will yield a group of zip codes with right properties. After this, we can get total number of yoga centers, gym and fitness centers in each zip code from Foursquare. The best zip code from the best group will be chosen such that it will not have yoga centers and gyms close by. Here, an assumption is made that the gym/fitness centers also have yoga classes. This assumption is based on general observation.

Results

The data with the information about the population, median age, median income and home price was scraped from the website https://www.bestplaces.net/find/zip.aspx?st=CA&county=06067

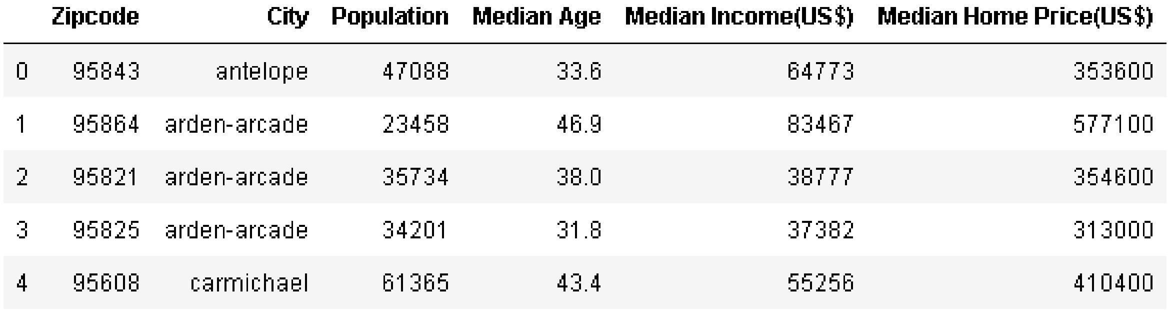
The scrapping was done using BeautifulSoup library and the relevant information was stored in pandas data frame. The detailed method of the information extraction can be seen in the Jupyter note book. The first row of the data frame is shown in figure 1.



	Zipcode	City	Population	Median Age	Median Income(US$)	Median Home Price(US$)
0	95843	Antelope	47,088	33.6	$64,773	$353,600

Figure 1. The first row of the data frame made using BeautifulSoup and Pandas library.

The data needs to be cleaned so that K-Means clustering can be run. All the values in the data frame shown in figure 1 are strings. They need to be changed into integers. The cleaned data frame is shown in figure 2.



	Zipcode	City	Population	Median Age	Median Income(US$)	Median Home Price(US$)
0	95843	antelope	47088	33.6	64773	353600
1	95864	arden-arcade	23458	46.9	83467	577100
2	95821	arden-arcade	35734	38.0	38777	354600
3	95825	arden-arcade	34201	31.8	37382	313000
4	95608	carmichael	61365	43.4	55256	410400

Figure 2. Cleaned data frame with the socio-economic data.

The non-relevant columns ‘Zipcode’ and ’City’ was dropped and the new data frame thus obtained was standardized, i.e. each column now had mean of zero and standard deviation of 1. This is necessitated by the fact that the values in one column are orders of magnitude different than in the other and need to be brought to the same scale for proper clustering.

The optimum value of number of clusters, as determined by ‘elbow method’, was found to be 5. This number was also verified by Calinski-Harabasz score. This method was used after installing the ‘yellowbrick’ library. The distortion of the cluster points and the Calinski-Harabasz score plotted against the number of clusters are shown in figures 3(a) and 3(b) respectively.

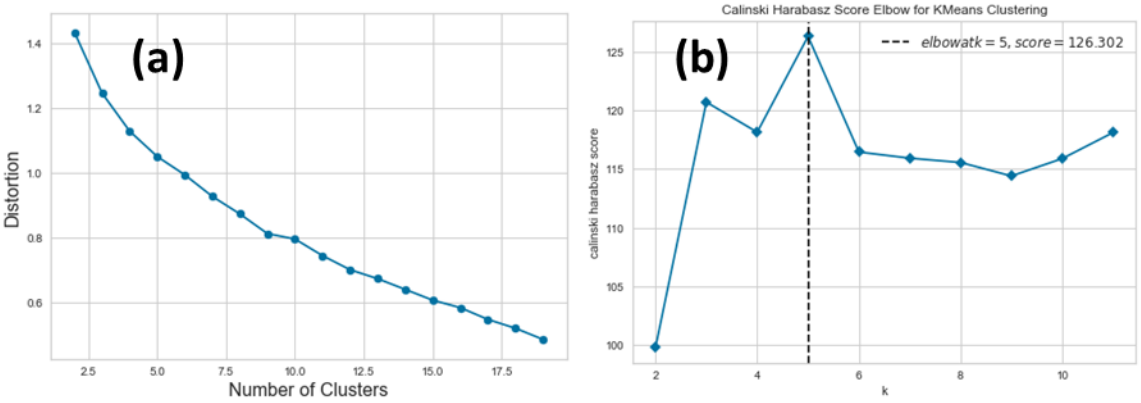
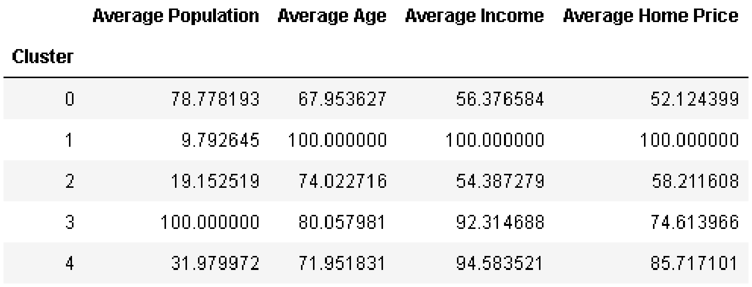


Figure 3(a). A plot of distortion vs the number of clusters. The elbow point seems to occur at number of clusters=5 but not apparent. 3(b). The Calinski-Harabasz score for different cluster numbers revels that the elbow occurs at the cluster number of 5.

The K-Means was run at the optimum cluster number of 5 on the standardized data. The average value of each column of each cluster was calculated to get the better picture of the cluster. Then, each column was normalized to 100 so that all the columns can be visualized in one plot.



Cluster	Average Population	Average Age	Average Income	Average Home Price
0	78.778193	67.953627	56.376584	52.124399
1	9.792645	100.000000	100.000000	100.000000
2	19.152519	74.022716	54.387279	58.211608
3	100.000000	80.057981	92.314688	74.613966
4	31.979972	71.951831	94.583521	85.717101

Figure 4. Data frame with normalized average values.

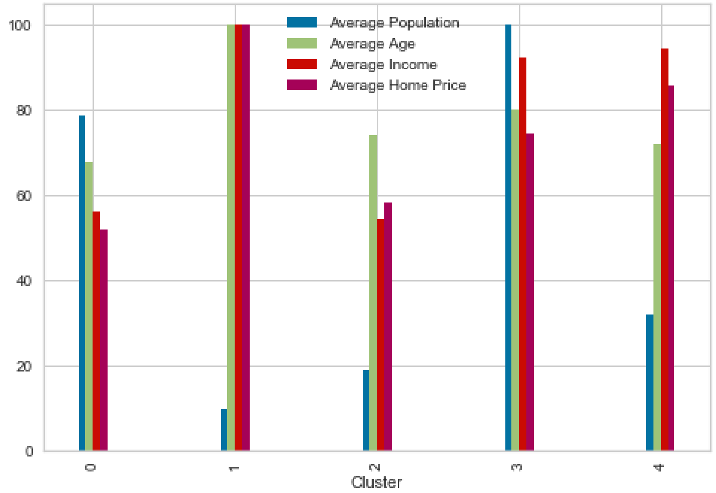


Figure 5. A bar graph of average values of population, age, income and home price of each cluster. The values are normalized to 100 so that all of them can be seen together. It reveals that the cluster 3 has the best zip codes for our need.

The bar graph of figure 5 revels that the cluster 3 consists of zip codes with highest population and the second highest median income. The median age is also neither too low nor too high. Thus, these locations need to be probed further. The locations of zip codes belonging to each cluster is shown in figure 6.

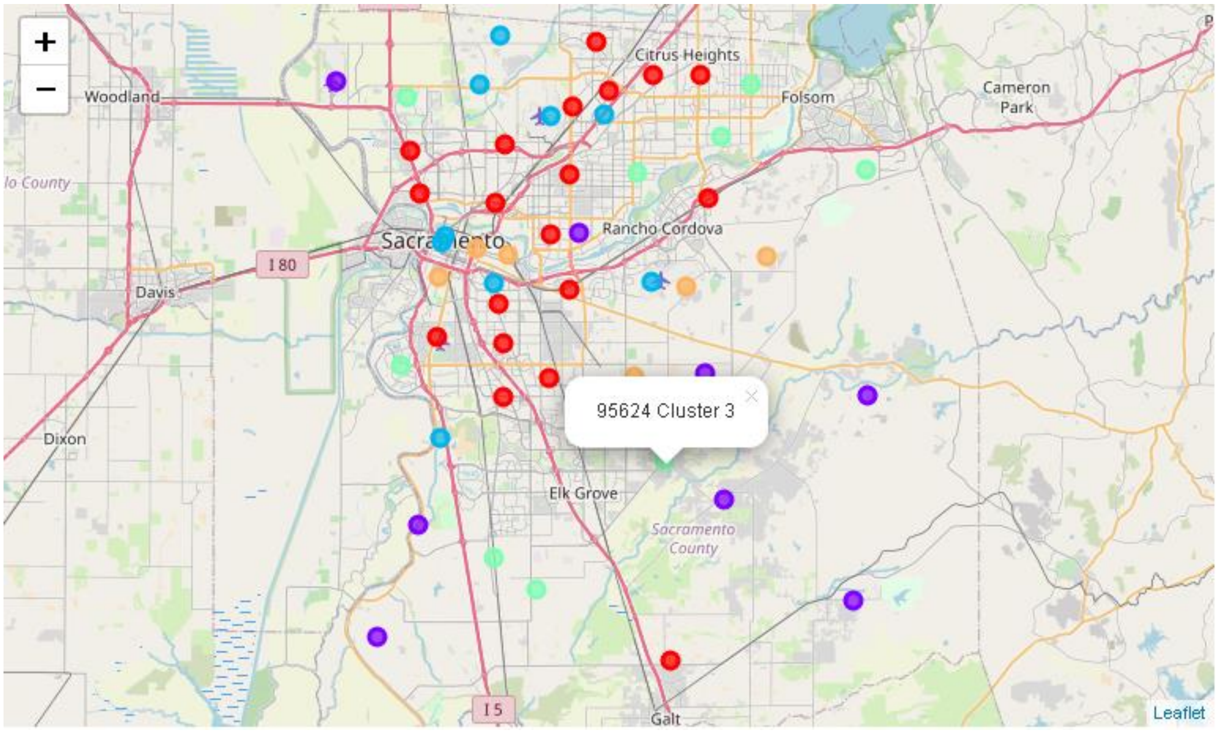
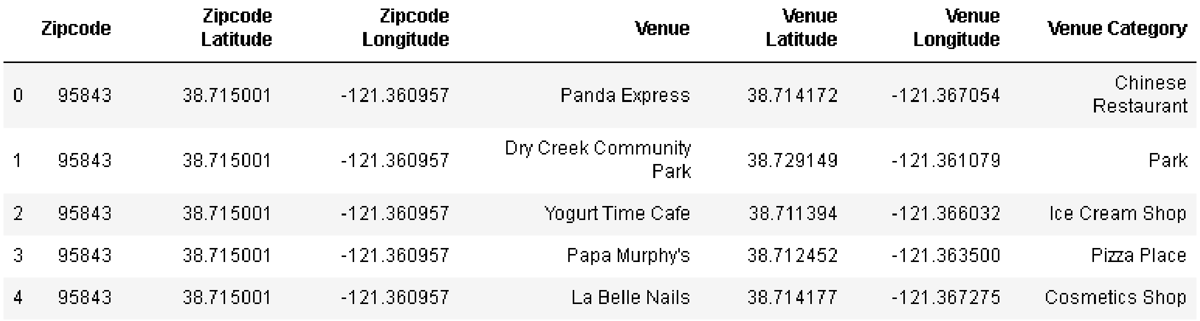


Figure 6. Map showing the zip codes belonging to different cluster. The map was produced by Folium.

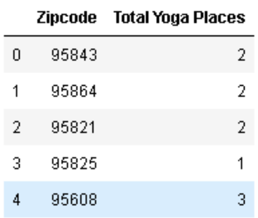
The next job is to find the total number of yoga centers and gym in each zip codes of cluster3. To this end, Foursquare was used to extract the venue data. The data frame containing this information is shown in figure 7.



	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	95843	38.715001	-121.360957	Panda Express	38.714172	-121.367054	Chinese Restaurant
1	95843	38.715001	-121.360957	Dry Creek Community Park	38.729149	-121.361079	Park
2	95843	38.715001	-121.360957	Yogurt Time Cafe	38.711394	-121.366032	Ice Cream Shop
3	95843	38.715001	-121.360957	Papa Murphy's	38.712452	-121.363500	Pizza Place
4	95843	38.715001	-121.360957	La Belle Nails	38.714177	-121.367275	Cosmetics Shop

Figure 7. Venue information extracted from Foursquare for each zip codes of the Sacramento County.

Another data frame was derived from the above data frame to yield the total number of yoga centers and gym in each zip code. This new data frame is shown in figure 8.



	Zipcode	Total Yoga Places
0	95843	2
1	95864	2
2	95821	2
3	95825	1
4	95608	3

Figure 8. A data frame showing total number of yoga centers and gym in each zip code.

The yoga centers and gyms where added to the map of figure 6 to get an idea about their distribution. The new map is shown in figure 9.

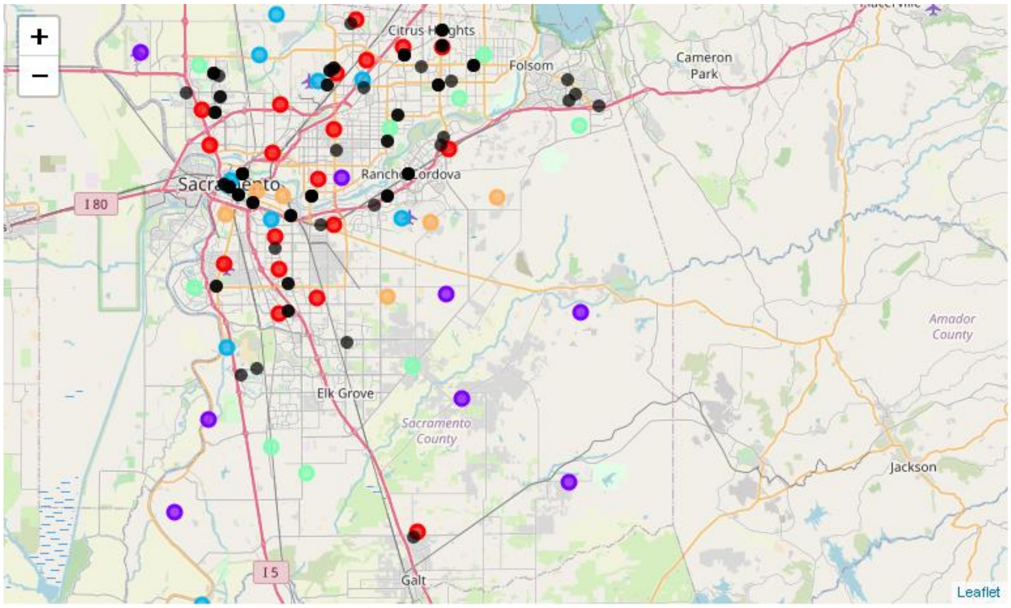
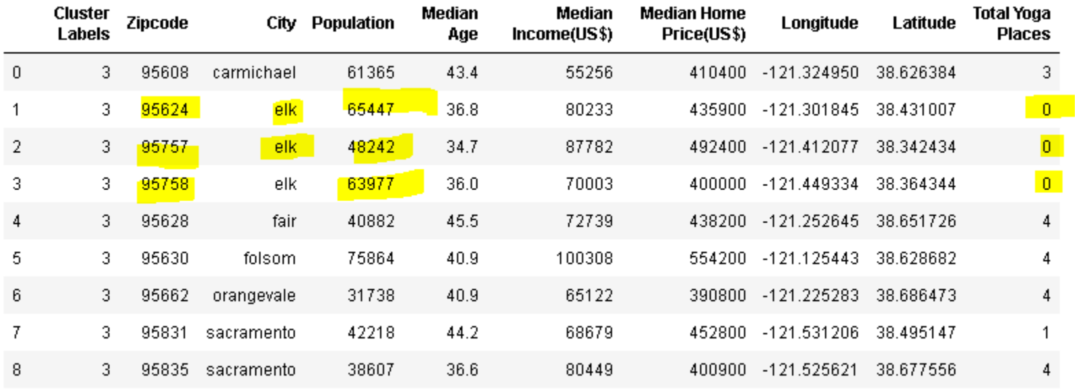


Figure 9. The black dots are gyms and yoga centers.

The information in figure 8 was merged with the zip codes of cluster 3. The new data frame is shown in figure 10.



	Cluster Labels	Zipcode	City	Population	Median Age	Median Income(US$)	Median Home Price(US$)	Longitude	Latitude	Total Yoga Places
0	3	95608	carmichael	61365	43.4	55256	410400	-121.324950	38.626384	3
1	3	95624	elk	65447	36.8	80233	435900	-121.301845	38.431007	0
2	3	95757	elk	48242	34.7	87782	492400	-121.412077	38.342434	0
3	3	95758	elk	63977	36.0	70003	400000	-121.449334	38.364344	0
4	3	95628	fair	40882	45.5	72739	438200	-121.252645	38.651726	4
5	3	95630	folsom	75864	40.9	100308	554200	-121.125443	38.628682	4
6	3	95662	orangevale	31738	40.9	65122	390800	-121.225283	38.686473	4
7	3	95831	sacramento	42218	44.2	68679	452800	-121.531206	38.495147	1
8	3	95835	sacramento	38607	36.6	80449	400900	-121.525621	38.677556	4

Figure 10. Data frame showing the total number of yoga centers and gyms in each zip codes of cluster 3. The data shows that the city of Elk has no gym or yoga centers, has high population and median income.

Discussion

The K-Means clustering applied to the socio-economic data was able to identify the ideal zip codes for opening a yoga studio. The zip codes, on average, has highest population, second highest median income and optimum median age. In addition to the information about the population, age, income and home price, we also need to know the total number of yoga studios in each zip codes of cluster 3. These information, obtained from Foursquare, was concatenated to the socio-economic data. The result was very insightful, as it revealed that, of the all zip codes in the cluster 3, the best ones are in the city of Elk. **There are three zip codes in Elk Grove and yoga center and gym are not present in any one of them. Out of the three zip codes, the best one is 95624 because it has the largest population and second highest median income.**

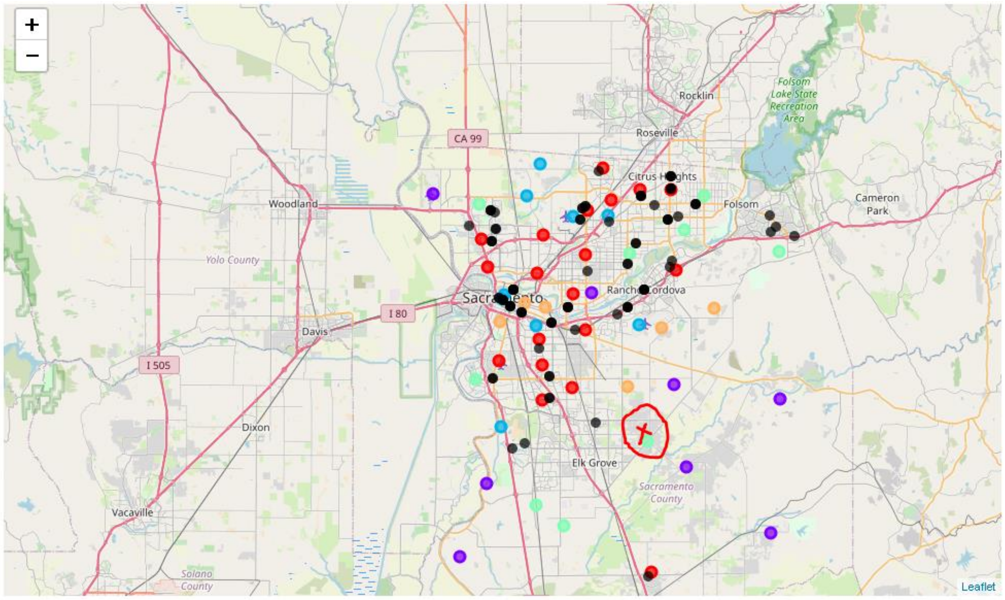


Figure 11. The best zip code to open a yoga studio is marked by a red cross inside a red circle.

Conclusion

The aim of this project was to suggest the best area to open a yoga studio. It is logical to think that the people who will join yoga studios will have relatively high income and live in a well to do neighborhoods, as the cost to join yoga studios are pretty high in the US. First, data with population, median age, income, and house price was scrapped from the website:

https://www.bestplaces.net/find/zip.aspx?st=CA&county=06067

After cleaning the data frame, dropping irrelevant information and standardization, the data frame was fed into K-Means clustering algorithm. The result of the analysis made it possible to narrow down on certain zip codes that fulfill our criteria. The group of neighborhoods with highest population and second highest median income was chosen. The venue information was obtained from Foursquare which enabled to find total number of yoga studios and gyms in each zip code, as well as their locations. **The result of the clustering of the zip codes and the venue information lead to the conclusion that Elk Grove city is the best place to open a yoga studio. Of the three zip codes in the city, 95624 looks more promising because of the high population and median income.**