Using Census Data to Plan where to open a new Gym

John Lunn 04/23/2020

Introduction

Background

In 2018, there were 62.5 million gym members in the United States. From 2000 to 2017, gym membership rates have experienced study growth over the past several years. In 2017, there were 38,477 gyms in the U.S. The types of gyms that people attend evolve based on fashion as well as research around the benefits of different exercises as well as the detriments. The most popular recent Gym types are Crossfit, Orange Theory, Stationary Cycles, Pilates, and Yoga has got increasingly more popular. The Bay Area and California are considered some of the fittest areas in the US.

Problem

Owning a Gym franchise is a good way for people who are passionate about fitness to combine their hobby with their employment. However, the Bay Area in California has many competitive gyms and which of the current fitness trends is suitable for which area. Can we use publicly available census data to predict where and what type of gym to open?

Interest

A prospective Gym franchise owner would be interested in this research in order to help make the difficult decision on where to open a gym. Franchise brands like “Orange Theory” or “SoulCycle” could also be interested in the outcomes of this study.

Data acquisition and cleaning

Data Sources

A variety of data sources are required for this project and rather than use Kaggle as a shortcut and to improve my Data Science skill the data was sought from a variety of places to build a rich dataset. Each data source had to solve a specific need.

What is the Bay Area?

In order to define the Bay area we used Beautiful Soup to scrape data from Wikipedia https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_the_San_Francisco_Bay_Area This scraped data lists all the cities and counties in the Bay Area. After cleaning the data we end with a data frame.

	City	City_Name	County	Land Area Km	Pop
0	Alameda	Alameda	Alameda	27.5	73812.0
1	Albany	Albany	Alameda	4.6	18539.0
2	American Canyon	American Canyon	Napa	12.5	19454.0
3	Antioch	Antioch	Contra Costa	73.4	102372.0
4	Atherton	Atherton	San Mateo	13.0	6914.0

What are the Zip Code, Long, and Lat for our cities?

A google search revealed a useful repository of data at https://public.opendatasoft.com/explore/?sort=modified from which I was able to obtain a .csv file of every Zip code and Geodata in the USA “us-zip-code-latitude-and-longitude.csv”. I cleaned the data of whitespace and removed non-relevant columns. Then I was able to merge the bay area data above with the zip code list to generate a list of all cities in the bay area, their zip codes (important as many cities have multiple Zip Codes), and their geodata.

Zip	City	County	Land Area Km	Pop	Latitude	Longitude
94501	Alameda	Alameda	27.5	73812	37.769528	-122.25937
94502	Alameda	Alameda	27.5	73812	37.734379	-122.23952
94706	Albany	Alameda	4.6	18539	37.889125	-122.29371
94503	American Canyon	Napa	12.5	19454	38.170372	-122.25605
94509	Antioch	Contra Costa	73.4	102372	37.991571	-121.80207

Census Data?

After an extensive search and attempts to extract the data directly from the Census board as well as other sites eventually I was able to find https://censusreporter.org/

This website allowed me to build reports for specific for every Zip code in California with data from the most recent census whether its 2016 or 2018/19 depending on what was surveyed. Unfortunately, each report could only list certain data before it could be downloaded in .csv format. Some manipulation was required using a spreadsheet as it was more efficient than using panda's alone. Each data set was picked based the assumptions as below before being cleaned and merged with the bay are zip data above to provide a dataframe for each. These dataframes have a format like so :

Zip	Total	Owned	Rent	City	County	Land Area Km	Pop	Latitude	Longitude
94002	10380	6308	4072	Belmont	San Mateo	12.0	25835	37.516687	-122.29026
94005	1836	1354	482	Brisbane	San Mateo	8.0	4282	37.682882	-122.40422
94010	16037	9460	6577	Burlingame	San Mateo	11.4	28806	37.574936	-122.36332
94014	14028	8169	5859	Daly City	San Mateo	19.8	101123	37.699182	-122.45035
94015	20052	11199	8853	Daly City	San Mateo	19.8	101123	37.682583	-122.48209

Commute Data

It was assumed that data around the length of a commute would be relevant to the location of Gyms. e.g. would people with long commutes work out near their work location rather than near home?

Age Data

The age distribution of a population would surely have and impact on the Gyms in their area. The highest number of Gym members fall between 25 and 60. Most commercial gyms also do not allow under 18's and over 70's are unlikely to engage in HIT workouts like Orange Theory. Also does a high amount of under 4 children in an area change gym membership as parents stay at home.

Income Data

It has been found that the lower the income the less likely an individual is to work out. This is caused by many factors but it cannot be ignored that Gyms and fitness are expensive and especially the trendy fitness gyms in this study with some offering classes at over $60 a session. Also for the very high income, it is increasingly more likely that they will have a home gym or personal trainer reducing the likelihood of attending a commercial gym.

Own/Rent Data

Do homeowners attend the gym more frequently and does renting mean you are less likely to commit to a long term membership or mean you have less disposable income for membership. Does rental accommodation have more access to complex condo gyms?

The Gym Data

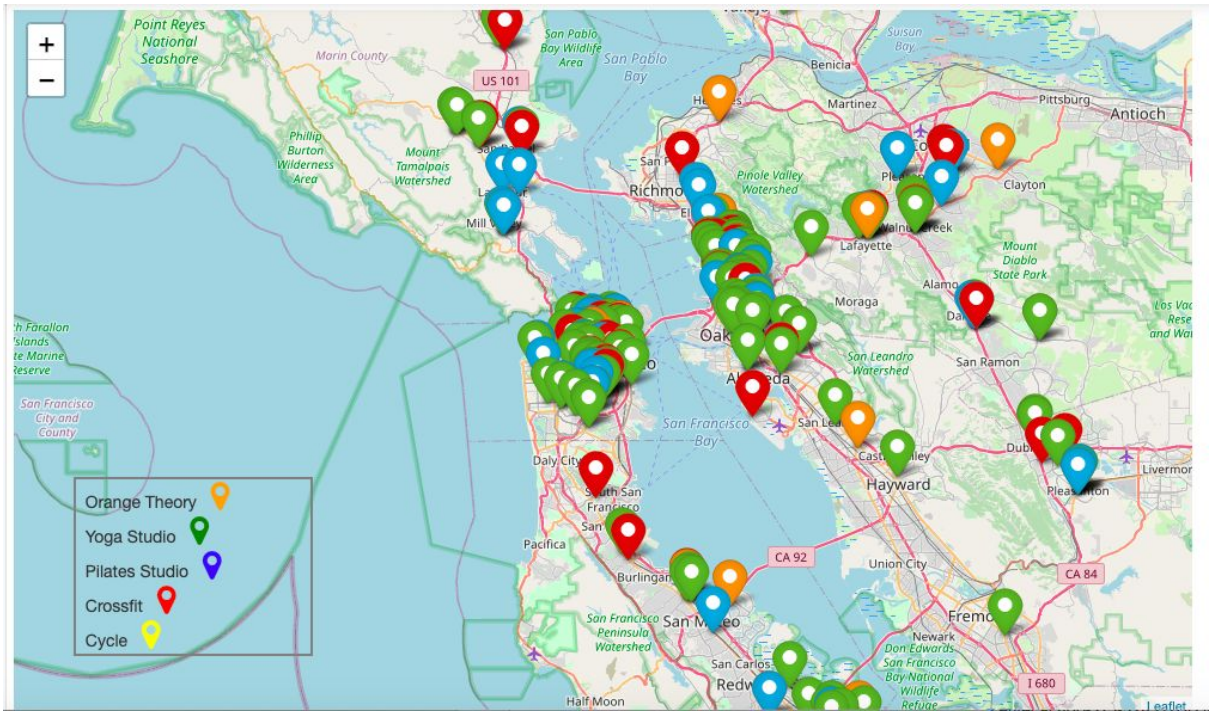
As part of the assignment, we were required to use Foursquare. Therefore using the Foursquare explore API and the Zip Code/City data frame I built above, I was able to build a data frame of all of the Gyms in the Bay Area.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Category1	Venue Category
0	Belmont	37.516687	-122.29026	Yvonne and Schuyler Moore Athletic Pavilion	37.515504	-122.285230	4bf58dd8d48988d176941735	Gym
1	Belmont	37.516687	-122.29026	United Studios of Self Defense	37.511324	-122.292948	4bf58dd8d48988d101941735	Martial Arts Dojo
2	Burlingame	37.574936	-122.36332	Ludus Fitness LLC	37.567474	-122.367654	4bf58dd8d48988d175941735	Gym / Fitness Center
3	Daly City	37.699182	-122.45035	Pointe Pacific Gym	37.701279	-122.456457	4bf58dd8d48988d176941735	Gym
4	Daly City	37.682583	-122.48209	Westmoor High School Track	37.680401	-122.479619	4bf58dd8d48988d106941735	Track

Then I used search and filtering to build separate data frame for each of our trending gym types “Orange Theory’, Crossfit, Yoga, Cycle, Pilates.

Then I was able to display these locations on a map of the Bay Area.

Location of Trending Gym Types in the Bay Area.



Unfortunately, the availability of Zip Code in the foursquare data is not reliable and therefore I had to merge with data from our Zip Code data frame based on Longitude and Latitude in order to populate the zip codes into our data.
I merged all individual Gym Types into 1 data frame and removed all data apart from the Gym types and zip codes.

Zip	Orange	yoga	crossfit	pilates	Cycle
94022	0	1	0	1	0
94022	0	1	0	1	0
94025	0	1	0	0	0
94030	0	1	0	0	1
94040	0	1	1	0	0

Then I wrote my own classification algorithm to classify the data into groups based on the location and Gyms in the zip code and deduped the data.

Zip	Orange	yoga	crossfit	pilates	Cycle	thesum	Category
94102	1	1	1	1	0	4	OrYoCrPi
94549	1	1	1	1	0	4	OrYoCrPi
94549	1	1	1	1	0	4	OrYoCrPi
94703	1	1	1	1	0	4	OrYoCrPi
94703	1	1	1	1	0	4	OrYoCrPi

The Features Data Frames

Each of our Census data frames now was merged with the Gym Matrix data frame above and then numbers we converted to %'s so that the data can be used more efficiently to predict which features determine whether a zip code will have each type of trending Gym.

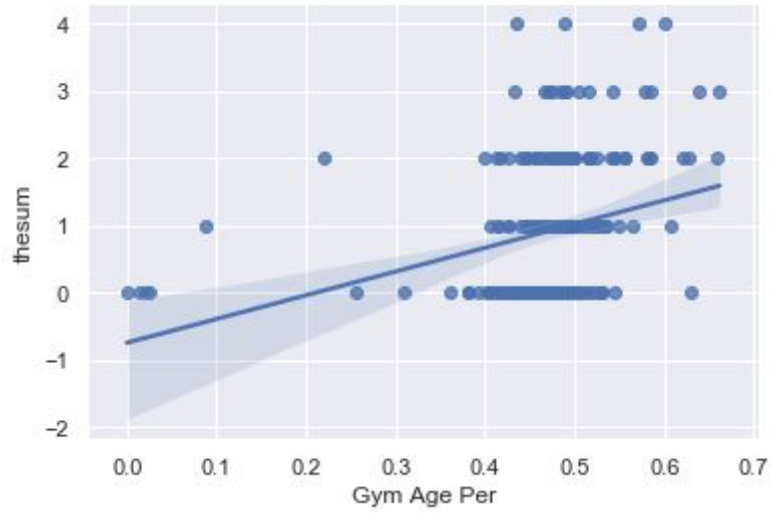
Each Dataframe had to be treated differently based on the data typed they contained, for example, the age dataset looked like this:

Zip	Male Per	Total	Gym Age Per	Young Kids	City	Density	Land Area Km	Orange	yoga	crossfit	pilates	Cycle	thesum	Category
94002	0.499228	27202	0.485405	0.061613	Belmont	2266	12.0	0.0	0.0	0.0	0.0	0.0	0.0	0
94005	0.471867	4692	0.518968	0.046675	Brisbane	586	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0
94010	0.476457	42730	0.462368	0.056611	Burlingame	3748	11.4	0.0	0.0	0.0	0.0	0.0	0.0	0
94014	0.502878	49515	0.483550	0.047157	Daly City	2500	19.8	0.0	0.0	0.0	0.0	0.0	0.0	0
94015	0.488310	64887	0.482994	0.040825	Daly City	3277	19.8	0.0	0.0	0.0	0.0	0.0	0.0	0
94019	0.475386	20314	0.460274	0.049621	Half Moon Bay	1223	16.6	0.0	0.0	0.0	0.0	0.0	0.0	0
94022	0.493033	19378	0.417381	0.027299	Los Altos	1153	16.8	0.0	1.0	0.0	1.0	0.0	2.0	YoPi
94024	0.471808	23961	0.429949	0.044405	Los Altos	1426	16.8	0.0	0.0	0.0	0.0	0.0	0.0	0

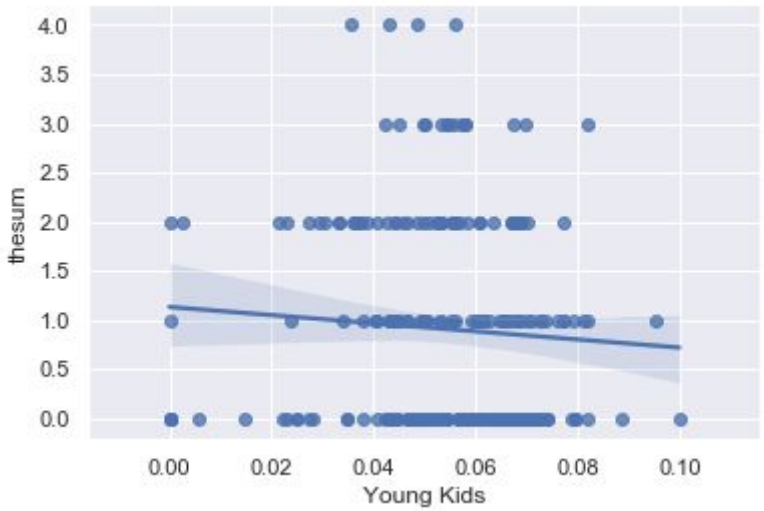
Feature Selection

For census data frame I looked at which features could be relevant and then ran a regression plot to see if that feature could have an influence on whether an area had a gym or not and how many.

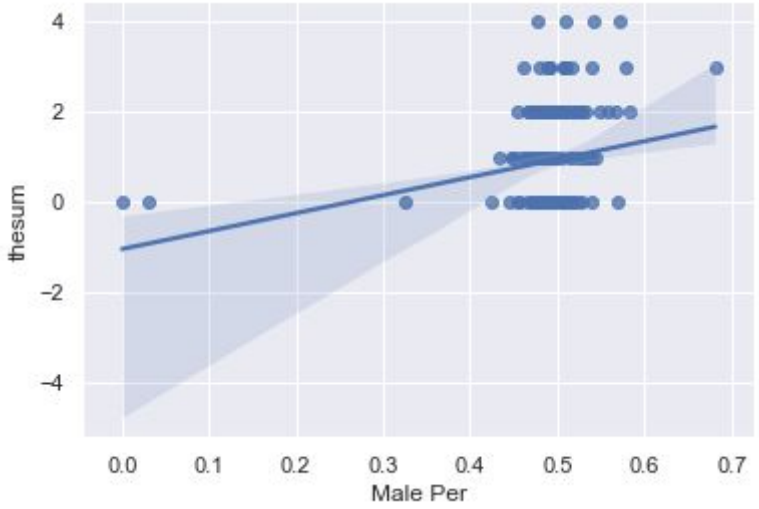
Age Features



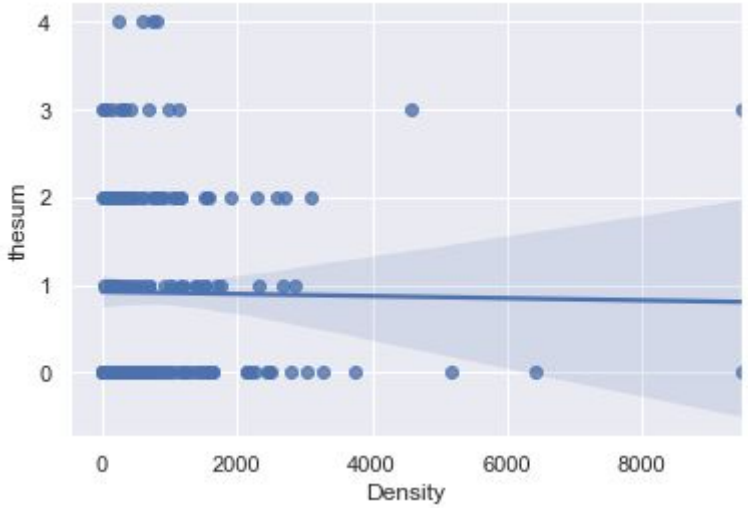
The data suggests that the percentage of the population in the age range 20-60 has an influence on the number of Gym's in a particular area.



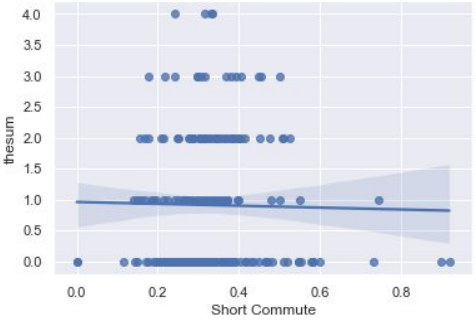
The data suggests that there is a log type relationship between the percentage of population under 4 and the number of gyms. It is interesting that when there are few children there are fewer gyms and when there is a high percentage there are also fewer gyms and there seems to be an optimal % between 2 - 8% of the population.

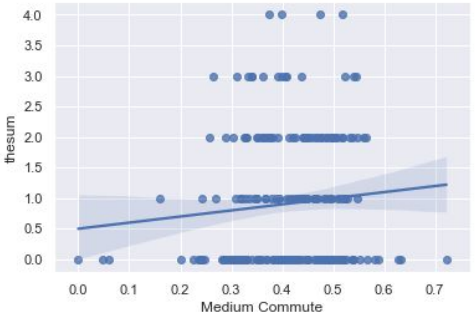


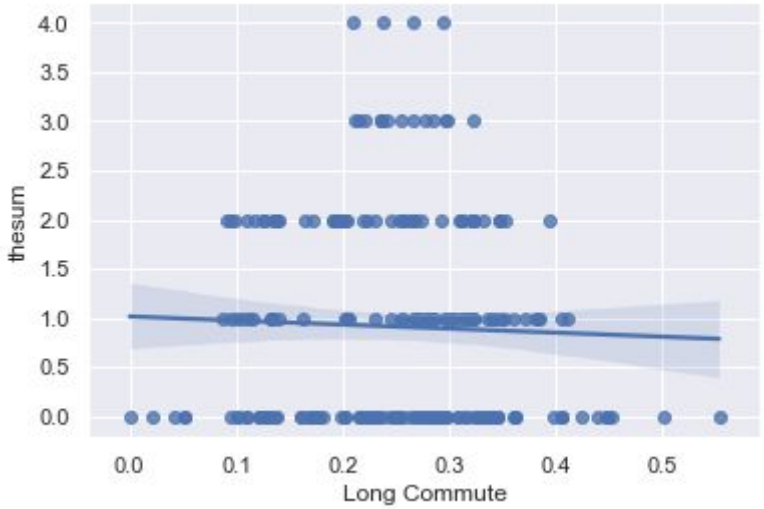
The data suggests that that the ratio of Male to female skews toward more Gyms where the data leans towards a greater number of Males. This makes sense as external data shows men are less likely to quit the gym than women.



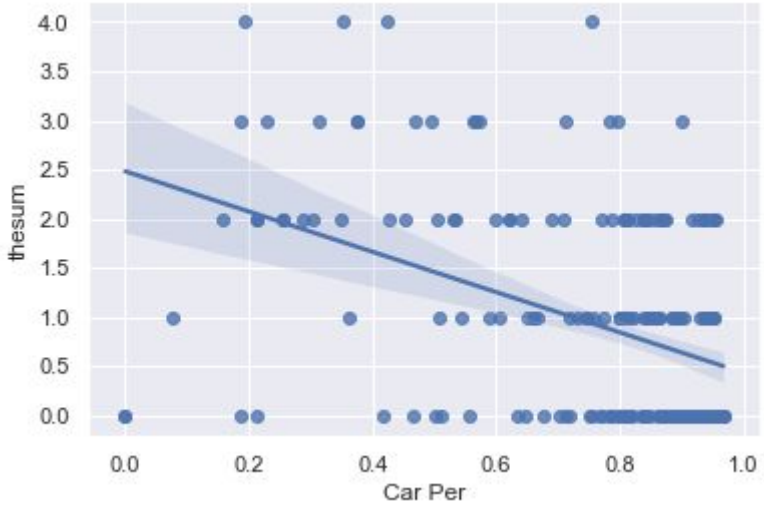
This data shows that population density has low correlation with these gyms which is surprising but we will not use the feature.





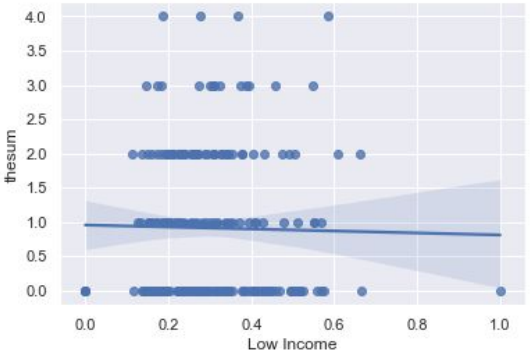


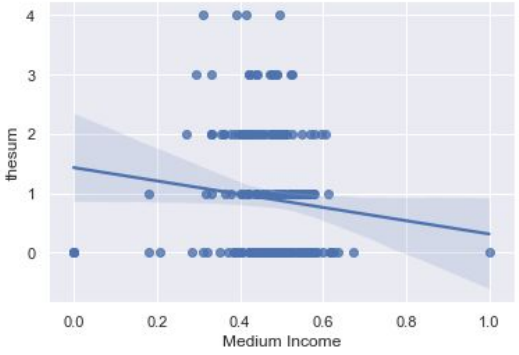
The best indicator of gym number out of commute length seems to be % of the population who have long commutes so this is our selected feature.

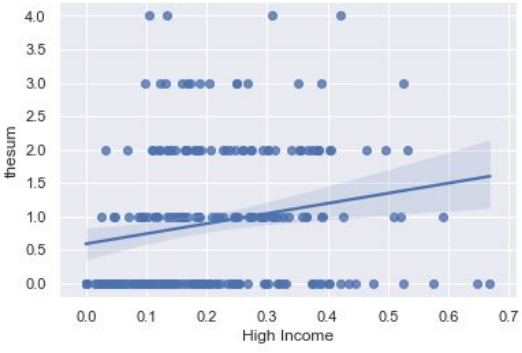


The Cars per population seems to be pretty strongly inversely corelated to the number of gyms. We will use this feature

Income Features

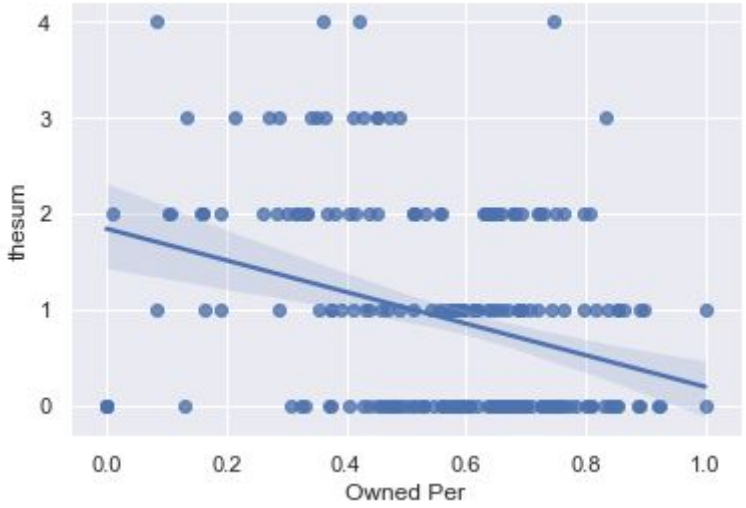






High Income and Medium Income seem to be good features to use whereas Low Income does not seem significant.

Own vs Rent Data



Increased ownership has a stong inversed correlation with gym numbers so this is a good feature to use.

Building the Feature DataSet

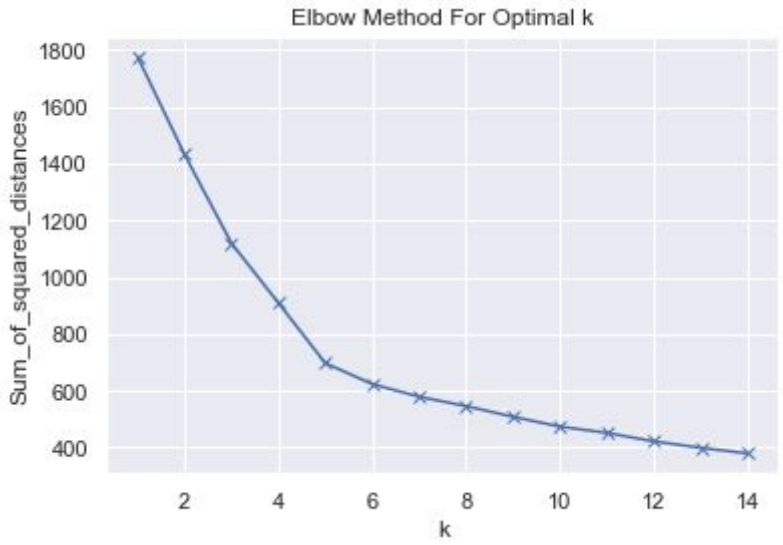
All of the above selected features were merged into one data set that i could use to build our clustering model and predictions.

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category
94002	Belmont	0.499228	0.485405	0.061613	0.885982	0.245061	0.433911	0.326108	0.607707	0
94005	Brisbane	0.471867	0.518968	0.046675	0.801282	0.329915	0.453704	0.253813	0.737473	0
94010	Burlingame	0.476457	0.462368	0.056611	0.801361	0.255832	0.421401	0.374571	0.589886	0
94014	Daly City	0.502878	0.483550	0.047157	0.711263	0.254946	0.549401	0.110493	0.582335	0
94015	Daly City	0.488310	0.482994	0.040825	0.785084	0.220862	0.563585	0.122033	0.558498	0
94019	Half Moon Bay	0.475386	0.460274	0.049621	0.939937	0.234712	0.489219	0.242222	0.699236	0
94022	Los Altos	0.493033	0.417381	0.027299	0.917826	0.117475	0.331752	0.532868	0.806141	YoPi
94024	Los Altos	0.471808	0.429949	0.044405	0.939652	0.102104	0.310006	0.574761	0.889611	0
94025	Menlo Park	0.476886	0.460246	0.077335	0.799008	0.133657	0.377974	0.390003	0.612938	Yo
94027	Atherton	0.499060	0.380978	0.053869	0.890902	0.095411	0.208387	0.648508	0.921747	0

Data Analysis

K Means Clustering of our Data

Firstly the aim was to see if our features dataset can be used to create clusters in the Bay Area. I used K Means clustering and firstly ran it so that to find the elbow so we could select optimal K.



Based on this I decided to use a k value of 6 and ran the K means clustering merged the new labels back into the dataset.

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94002	Belmont	0.499228	0.485405	0.061613	0.885982	0.245061	0.433911	0.326108	0.607707	0	5
94005	Brisbane	0.471867	0.518968	0.046675	0.801282	0.329915	0.453704	0.253813	0.737473	0	0
94010	Burlingame	0.476457	0.462368	0.056611	0.801361	0.255832	0.421401	0.374571	0.589886	0	5
94014	Daly City	0.502878	0.483550	0.047157	0.711263	0.254946	0.549401	0.110493	0.582335	0	0
94015	Daly City	0.488310	0.482994	0.040825	0.785084	0.220862	0.563585	0.122033	0.558498	0	0
94019	Half Moon Bay	0.475386	0.460274	0.049621	0.939937	0.234712	0.489219	0.242222	0.699236	0	0
94022	Los Altos	0.493033	0.417381	0.027299	0.917826	0.117475	0.331752	0.532868	0.806141	YoPi	5
94024	Los Altos	0.471808	0.429949	0.044405	0.939652	0.102104	0.310006	0.574761	0.889611	0	5
94025	Menlo Park	0.476886	0.460246	0.077335	0.799008	0.133657	0.377974	0.390003	0.612938	Yo	2
94027	Atherton	0.499060	0.380978	0.053869	0.890902	0.095411	0.208387	0.648508	0.921747	0	5
94028	Portola Valley	0.477286	0.408707	0.024898	0.935805	0.176763	0.283523	0.526150	0.808494	0	5
94030	Millbrae	0.490401	0.476750	0.045795	0.787282	0.253980	0.481297	0.258481	0.641357	YoCv	0

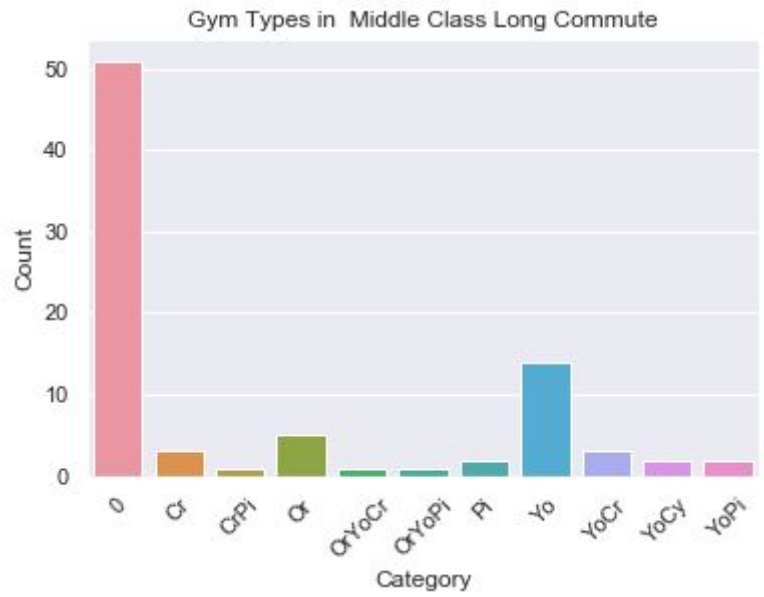
Looking at our 5 clusters

Cluster 0

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94005	Brisbane	0.471867	0.518968	0.046675	0.801282	0.329915	0.453704	0.253813	0.737473	0	0
94014	Daly City	0.502878	0.483550	0.047157	0.711263	0.254946	0.549401	0.110493	0.582335	0	0
94015	Daly City	0.488310	0.482994	0.040825	0.785084	0.220862	0.563585	0.122033	0.558498	0	0
94019	Half Moon Bay	0.475386	0.460274	0.049621	0.939937	0.234712	0.489219	0.242222	0.699236	0	0
94030	Millbrae	0.490401	0.476750	0.045795	0.787282	0.253980	0.481297	0.258481	0.641357	YoCy	0
94044	Pacifica	0.494381	0.485831	0.059644	0.864524	0.269096	0.578544	0.189551	0.687330	0	0
94112	San Francisco	0.499683	0.503699	0.042846	0.603633	0.295750	0.541606	0.145921	0.643041	Yo	0
94116	San Francisco	0.481463	0.480313	0.050248	0.588920	0.339522	0.491168	0.246322	0.687758	Yo	0
94124	San Francisco	0.481775	0.464469	0.061731	0.633951	0.268855	0.385472	0.098995	0.515714	0	0

Looking at the Data we can classify this group as middle class with a long commute i.e suburbia.

The distribution of Gym types for this cluster is as so:



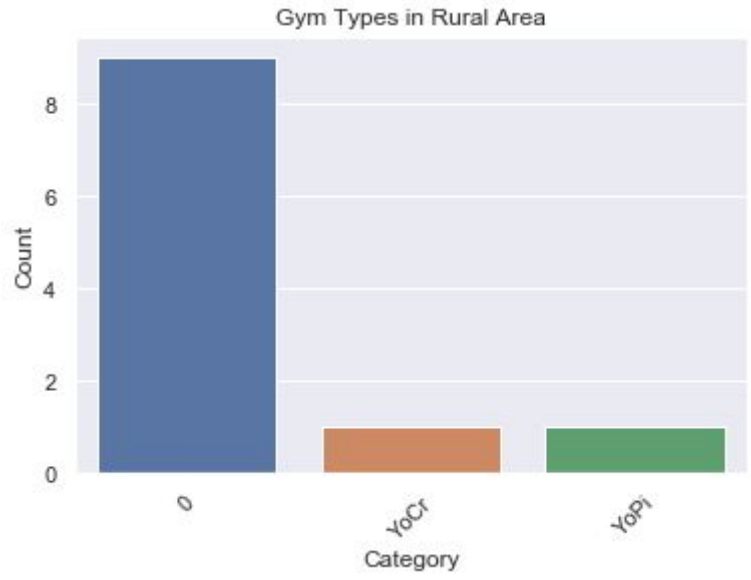
With Yoga and Crossfit being the strongest showing with a large number of zips with no Gym's at all.

Cluster 1

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94515	Calistoga	0.507800	0.474045	0.043706	0.812462	0.110874	0.438384	0.176094	0.618855	0	1
94558	Napa	0.498930	0.436479	0.048596	0.935119	0.135504	0.522262	0.138919	0.672743	0	1
94595	Walnut Creek	0.424964	0.254934	0.028041	0.835375	0.291203	0.427607	0.148317	0.841296	0	1
95053	Santa Clara	0.464655	0.014682	0.000000	0.211392	0.021519	1.000000	0.000000	1.000000	0	1
95404	Santa Rosa	0.501260	0.454366	0.044498	0.908792	0.120331	0.482263	0.106849	0.561863	0	1
95409	Santa Rosa	0.455001	0.360757	0.038079	0.959387	0.130873	0.480946	0.115062	0.643233	0	1
95425	Cloverdale	0.475912	0.413093	0.053089	0.933580	0.160061	0.482084	0.085642	0.668615	0	1
95448	Healdsburg	0.487816	0.413168	0.036952	0.874593	0.138052	0.491078	0.166401	0.635717	YoPi	1
95472	Sebastopol	0.487777	0.434465	0.034913	0.938318	0.160749	0.518185	0.128213	0.677596	0	1

These are Rural areas

Gym Distribution is as follows:



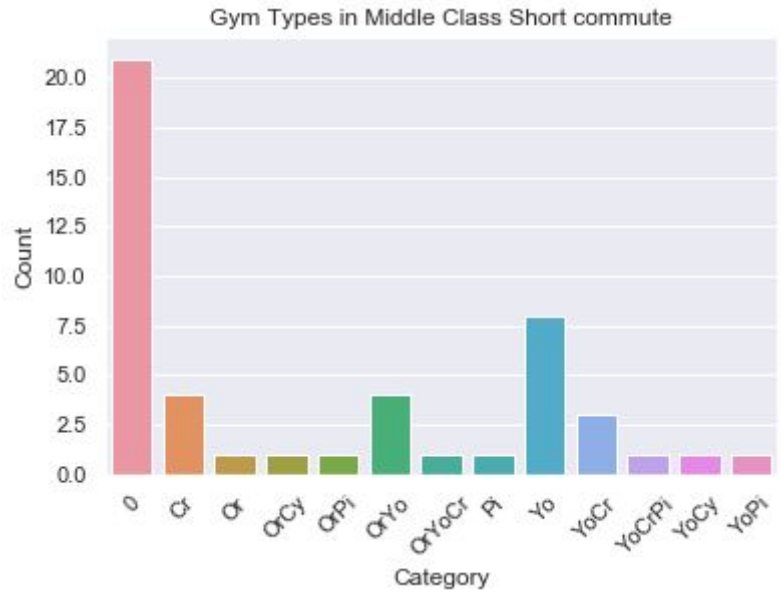
A very small cluster but again Yoga, Pilates, and Crossfit only.

Cluster 2

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94025	Menlo Park	0.476886	0.460246	0.077335	0.799008	0.133657	0.377974	0.390003	0.612938	Yo	2
94040	Mountain View	0.527661	0.514521	0.068629	0.838405	0.099376	0.429313	0.300540	0.416505	YoCr	2
94041	Mountain View	0.531680	0.556690	0.053425	0.769920	0.092032	0.496702	0.272613	0.315327	YoCy	2
94043	Mountain View	0.532044	0.532393	0.073647	0.807185	0.106003	0.492706	0.287316	0.458001	Yo	2
94061	Redwood City	0.489602	0.495987	0.070109	0.846037	0.133657	0.481359	0.232115	0.511948	OrPi	2
94063	Redwood City	0.540475	0.495667	0.070950	0.859930	0.132877	0.453825	0.118510	0.356322	Yo	2
94066	San Bruno	0.490655	0.497635	0.054309	0.818926	0.201402	0.565003	0.175390	0.577159	0	2
94080	South San Francisco	0.490140	0.474503	0.050091	0.811770	0.204697	0.546689	0.168553	0.611866	Cr	2
94085	Sunnyvale	0.534601	0.549128	0.095333	0.857437	0.116140	0.526492	0.244349	0.373640	Cr	2
94086	Sunnyvale	0.518275	0.544812	0.077252	0.838748	0.127429	0.493192	0.275114	0.327681	OrCy	2

These are Middle Class areas with low commutes .

Gym Distribution is as follows:

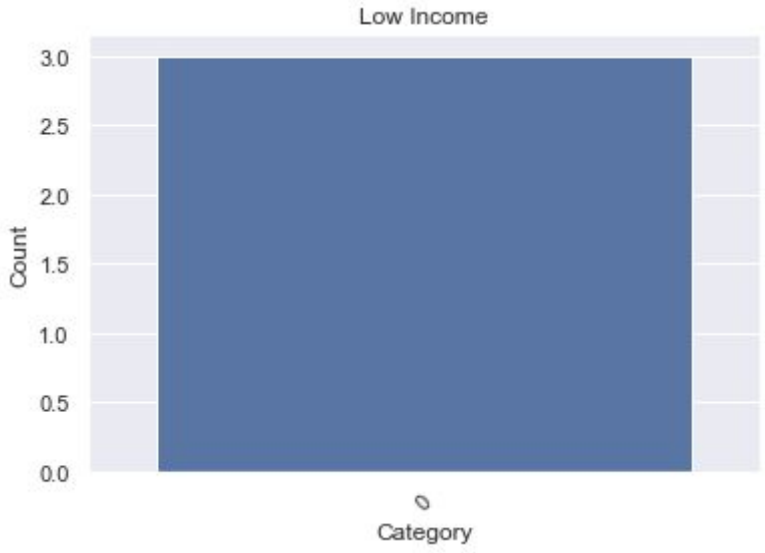


Half of the zip's have gyms and again Yoga and Crossfit show the strongest but there is a nice spread of Gym types.

Cluster 3

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94575	Moraga	0.324786	0.020513	0.000000	0.501144	0.041190	0.000000	0.000000	0.000000	0	3
94613	Oakland	0.030197	0.024390	0.005807	0.185185	0.051852	0.000000	0.000000	0.000000	0	3
94704	Berkeley	0.482528	0.220486	0.002501	0.255236	0.204495	0.270068	0.068392	0.106652	OrYo	3
94720	Berkeley	0.471895	0.087176	0.000000	0.076056	0.095775	0.181818	0.590909	1.000000	Yo	3
94850	Richmond	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0	3

These are the very working-class areas and the 2 Berkley zips are outliers as they are both zip codes for the University Campus.

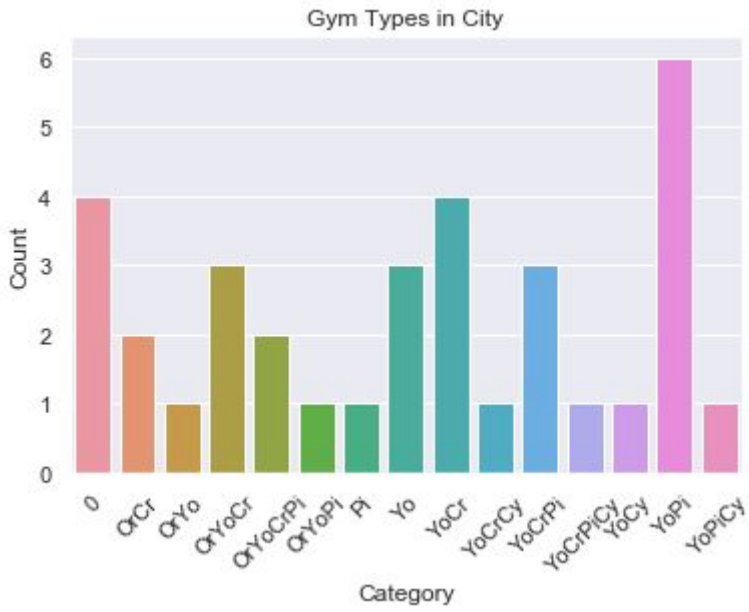


There are no trending gyms in these areas once you remove the outliers.

Cluster 4

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94102	San Francisco	0.571668	0.571861	0.035536	0.192003	0.210695	0.309960	0.105151	0.082843	OrYoCrPi	4
94103	San Francisco	0.584235	0.628293	0.022934	0.211000	0.193509	0.331379	0.164005	0.155650	OrCr	4
94104	San Francisco	0.681319	0.476190	0.053114	0.185185	0.236111	0.294118	0.158088	0.132353	OrYoCr	4
94105	San Francisco	0.538831	0.661824	0.058110	0.230317	0.236741	0.329666	0.524654	0.427058	OrYoCr	4
94107	San Francisco	0.526053	0.621644	0.044259	0.348836	0.265453	0.377623	0.384500	0.403409	YoCr	4

These are large City dwellers in San Francisco and Oakland

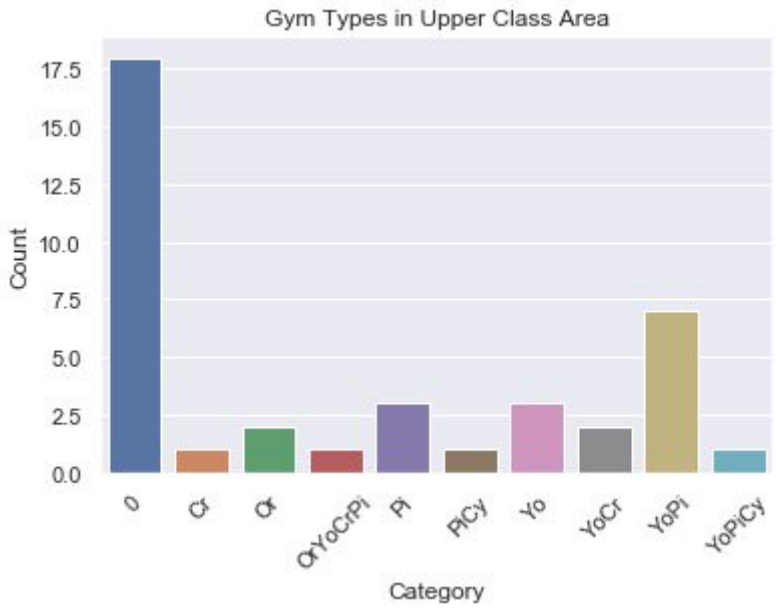


This is the nicest spread with large number of every type of gym. Only 4 zips have no Gyms and on analysis shows that these zips are the airport, a park, and one in an upcoming area in Oakland.

Cluster 5

Zip	City	Male Per	Gym Age Per	Young Kids	Car Per	Long Commute	Medium Income	High Income	Owned Per	Category	Cluster
94002	Belmont	0.499228	0.485405	0.061613	0.885982	0.245061	0.433911	0.326108	0.607707	0	5
94010	Burlingame	0.476457	0.462368	0.056611	0.801361	0.255832	0.421401	0.374571	0.589886	0	5
94022	Los Altos	0.493033	0.417381	0.027299	0.917826	0.117475	0.331752	0.532868	0.806141	YoPi	5
94024	Los Altos	0.471808	0.429949	0.044405	0.939652	0.102104	0.310006	0.574761	0.889611	0	5
94027	Atherton	0.499060	0.380978	0.053869	0.890902	0.095411	0.208387	0.648508	0.921747	0	5
94028	Portola Valley	0.477286	0.408707	0.024898	0.935805	0.176763	0.283523	0.526150	0.808494	0	5
94062	Redwood City	0.486220	0.471521	0.050150	0.910534	0.171424	0.403201	0.401347	0.735461	0	5

These are the Upper Class areas.



The upper class areas have mainly Yoga and Pilates Gyms with a large % with no Gyms at all which as discussed earlier is probably due to personal trainers and home gyms.

Conclusions from Clustering

For a person deciding to set up a Gym in the bay area then the order of preference of where they should set up is as follows.

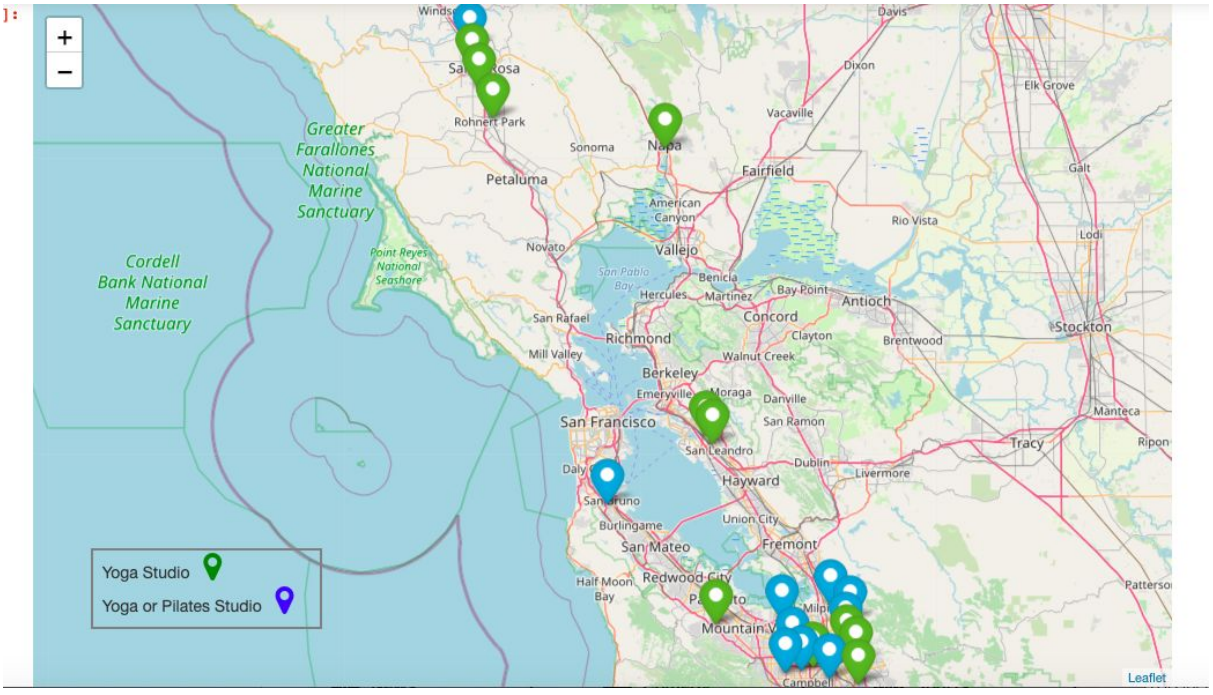
1) **The Middle-Class low commute:** the area has the most opportunity with the lowest risk, Over half already have gyms and this leave half of the area as potential venues for gyms with the right demographics
2) **Middle-Class long commute:** There are plenty of no gym zips but the large % of empty zips suggests demand is not there possibly caused by workers working out at work or at gyms near work or not at all due to exhausting commute times.
3) **Upper-Class areas:** If you were to set up a Gym in these areas you would be best off choosing to build a Yoga or Pilates Studio.
4) **City:** There is over-saturation in the city of Gyms and smaller Zip codes mean it's not hard to travel. The one exception is the upcoming area of Oakland by the Bay Bridge which has the right demographics but no Gym yet.
5) **Low Income:** The cost of trending Gyms is too high for these areas which is backed by the data.

Building a Model to Help A Gym owner select the Gym type they should build

As part of the study, I wanted to examine whether a model could be used to help a Gym owner pick the Gym type to build in a particular Zip code. The easiest way to provide this to a Gym owner would be in the format of a Decision Tree which they could use on publicly available census data. Therefore the first model I tried was a Decision Tree. We are using Cluster 2 as the area to predict. Unfortunately cluster 2 is small so this was not successful with it having an accuracy of 0.
Expanding the training dataset to the entirety of the Bay Area returned an accuracy of 0.14 Clearly a decision tree will not serve us here.

The second method i tried was an SVM model which returned and accuracy of 0.31428 and so did a Logistical Regression Model.

Therefore although is a low level of confidence I decided to use the SVM Model and predict the type of Gym's we should put in our Cluster 2 Cities with no current gyms.



The model tells us we should put Yoga or studios is all of them and Pilates in some. Although the model is not that accurate the conclusion is not wrong in that this is the most common gym type in our dataset.

Conclusion

The results of this study can show is that using census data, some areas are more likely to succeed as others as locations for Gyms and that you are better off choosing middle-class non-Large City areas as the place to locate such a Gym. In so much as using the census data to build a model to help you chose a gym type, we have a very low accuracy model. However, the model has determined the most likely to succeed would be a Yoga or Pilates Gym. These gym types have the lowest starting capital investment in that very little equipment is required to start this type of gyms and all that really is needed is floor space. Therefore as a conclusion as a prospective gym owner in the Bay area you should pick a middle-class low commute area and set up a Yoga/Pilates studio. Namaste!

Further Studies

The same dataset could be used to look at increased exercise types like martial arts dojo's and more traditional weights gyms as well as Boxing to see if these would help improve accuracy. However to truly predict success then attendance and membership data would be needed which could be hard to obtain but could definitely help improve accuracy and predictions. 2020 Census data could also be interesting.