IBM Applied Data Science Capstone

Coursera

Week 5

Giampaolo Tumminello

Gym market in San Francisco

Introduction

The rise in awareness among millennials about health issues has resulted in an increase in expenditure on healthy lifestyle and fitness activities, which is encouraging them to join fitness clubs. Revenue of the fitness industry is expected to show an annual growth rate (CAGR 2020-2024) of 0.6%, resulting in a projected market volume of US$23,127m by 2024.

As a result, new fitness clubs are rapidly emerging in the streets of every city worldwide.

Problem

Given that, for entrepreneurs who wants to enter this market, the knowledge of where to open a new gym is crucial for the success of the investment.

In this project, the geographical distribution of gyms in the City of San Francisco will be analysed, to see **where the "good spots" are hided** in the city.

The definition of a "good spot" is fundamental for the analysis, thus it's important to find the right characteristics:

- Great potential client base.
- No strong direct competition.
- Good geographical position overall.

The first condition means to have a big amount of people living in the neighbourhoods surrounding the spot; the second means that no other gyms (or very few) are surrounding the spot and the third refers to all other geographical characteristics which can result in a competitive advantage, for instance a good building, a good view from the window, nice entrance and many more.

The analysis is intended to be useful for entrepreneurs who wants to enter in the fitness industry in San Francisco.

Data

The Data needed for the analysis regards the neighborhoods of San Francisco, which are easily retrievable from the web and from government websites in particular. Since this is a geographical analysis, the coordinates of each Neighbourhood are required and would be acquired using the Geocoder library in python programming language.

The second category of data regards the venues for each neighbourhood, divided into categories, so as to identify the fitness centre and gyms present in each area. These data can be retrieved using Foursquare API. The advantage of this method is that venues will be already divided for each neiborhood of belonging.

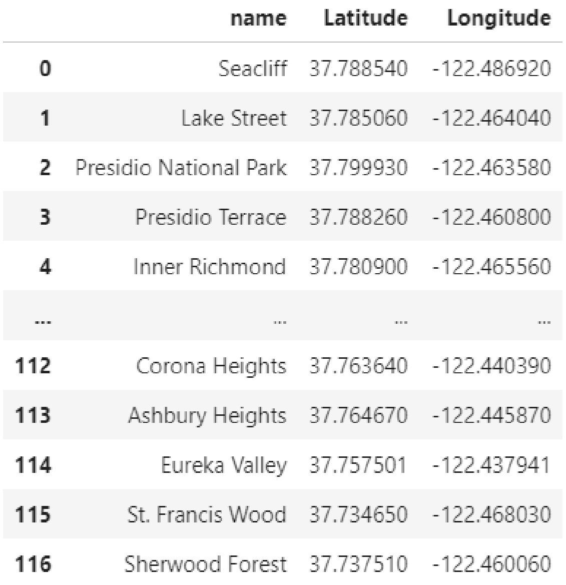
Methodology

1. Creating the data frame and visualize it

Initially, the Neighbourhoods of San Francisco are imported from "dataSF" government website1. The coordinates (latitude and longitude) of each neighbourhood are imported via Geocoder and a while loop to avoid some connection problems.

The resulting dataset consists in 117 rows containing the name of each neighbourhood and the respective coordinates.

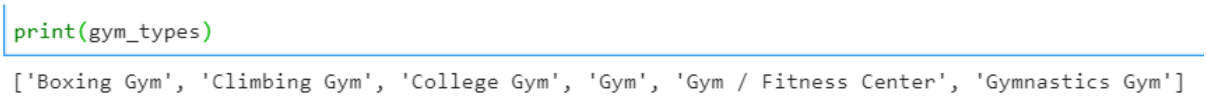
[1] 'https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4



	name	Latitude	Longitude
0	Seacliff	37.788540	-122.486920
1	Lake Street	37.785060	-122.464040
2	Presidio National Park	37.799930	-122.463580
3	Presidio Terrace	37.788260	-122.460800
4	Inner Richmond	37.780900	-122.465560
...	...	...	...
112	Corona Heights	37.763640	-122.440390
113	Ashbury Heights	37.764670	-122.445870
114	Eureka Valley	37.757501	-122.437941
115	St. Francis Wood	37.734650	-122.468030
116	Sherwood Forest	37.737510	-122.460060

2. Getting the venues data and merge with the data frame

After, using Foursquare API, the venues for each neighbourhood are imported (max 100 per neighbourhood). It results in a dataset containing nearly all the venues in San Francisco, divided and grouped by neighbourhood. The following passage is to encode the venue types and creating a dummy variable for each category (380 to be precise) found. In this way, querying the dataframe for "gym" will return all different kind of gyms in San Francisco.



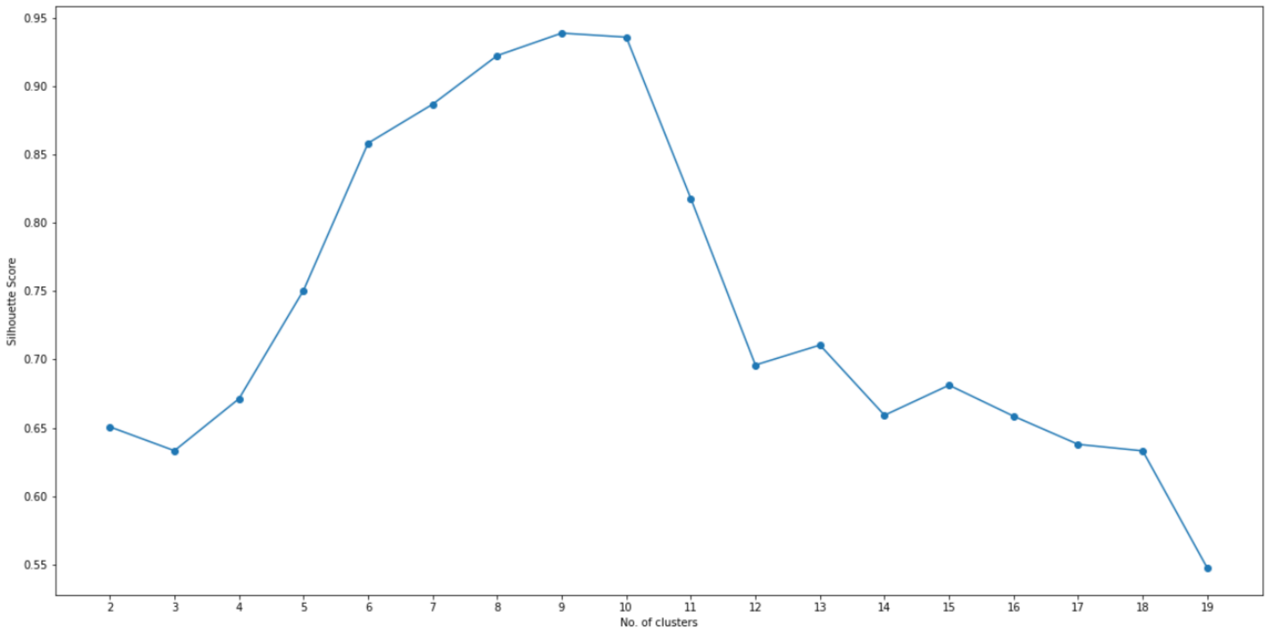
```
print(gym_types)
```

```
['Boxing Gym', 'Climbing Gym', 'College Gym', 'Gym', 'Gym / Fitness Center', 'Gymnastics Gym']
```

These gym_types are merged and summed up to get a category "Gyms" which includes all of them. Now data are ready to be clustered.

3. Clustering algorithm

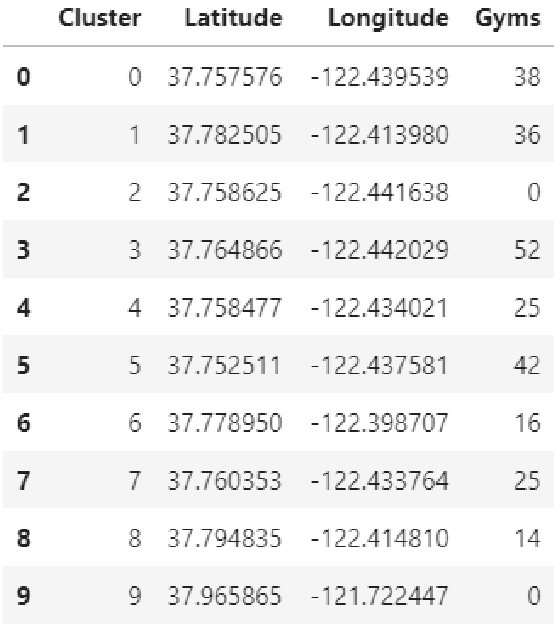
After data preparation, a k-means clustering algorithm is chosen to divide the neighbourhoods in clusters and see whether the "good spot" is present in the city od San Francisco. The first analysis regards the number of clusters, which is chosen to be 10, via silhouette score.



After, the k-means algorithm is applied and the clusters are aggregated. For the latitude and longitude, the mean of the neighbourhoods for each cluster is used, and, for the “Gyms” the sum of the values of the neighbours for each cluster is used.

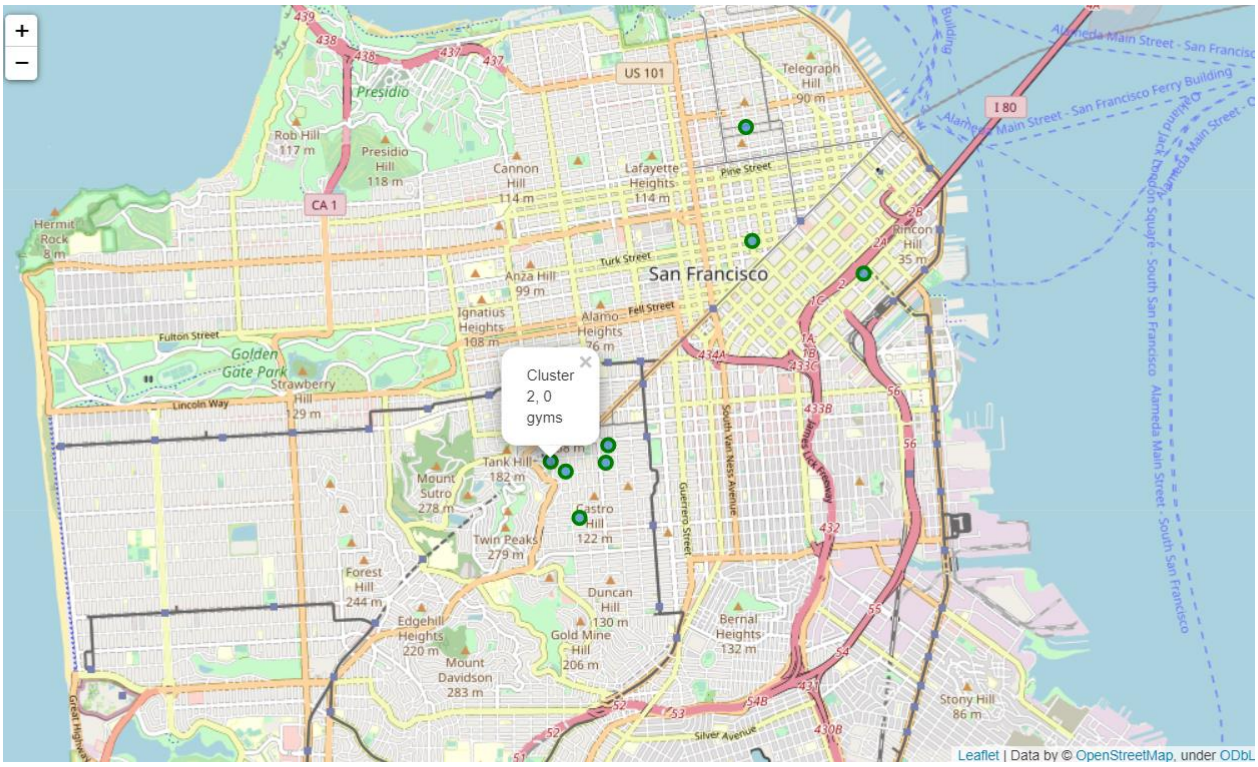
Results

After applying the algorithm, the dataset of the Cluster shows some interesting insights about the city.



	Cluster	Latitude	Longitude	Gyms
0	0	37.757576	-122.439539	38
1	1	37.782505	-122.413980	36
2	2	37.758625	-122.441638	0
3	3	37.764866	-122.442029	52
4	4	37.758477	-122.434021	25
5	5	37.752511	-122.437581	42
6	6	37.778950	-122.398707	16
7	7	37.760353	-122.433764	25
8	8	37.794835	-122.414810	14
9	9	37.965865	-121.722447	0

It’s important to specify that the cluster division is made not only using Euclidian distance between neighbourhoods, but also the number of gyms of each neighbourhood as a parameter. This was done to identify areas which, even if they are near other clusters, they have less gyms. In fact we can notice that Cluster number 2 has zero gyms in it, even if it is surrounded by other clusters containing respectively 38 gyms (cluster 0) and 52 gyms (cluster 3). It’s easier to get sense of this finding by looking at the geographical distribution of clusters.



Conclusion

The cluster number 2 represents a great opportunity for entrepreneurs who wants to enter in the fitness industry in San Francisco. In fact, it's reasonable that a gyms in cluster 2 would have a good amount of clients, since other surrounding areas are crowded with gyms and no one is present there. The criteria specified in the problem section are met in the cluster 2 area.

As stated before, this represents a good spot to enter in the fitness market and take advantage of the situation.