Data

To solve the problem, we will need the following data:

- List of neighborhoods in the Bay Area. This defines the scope of this project.
- Latitude and longitude coordinate of those neighborhoods. This will be needed to get venue data and plot on the map
- Venue data, particularly regarding fitness centers. This will be used to perform clustering.

Sources of data and methods that will be applied:

This wikipedia page (https://en.wikipedia.org/wiki/Category:Counties_in_the_San_Francisco_Bay_Area) contains the list of neighborhoods in the San Francisco Bay Area. We will use web scrapping technique to extract data from this page. We will use this in combination with pythons BeautifulSoup package and the geocoder package to give us the latitude and longitude of the venues.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare API has one of the largest databases of over 105+ million places, used by over 125,000 developers. We are particularly interested in the fitness center category(gyms,spas,health centers) that will help solve our problem.

This project will make use of many data science skills from web scraping , working with Foursquare API, data cleaning, data wrangling, to machine learning (k-means clustering) and map visualisation using Folium.