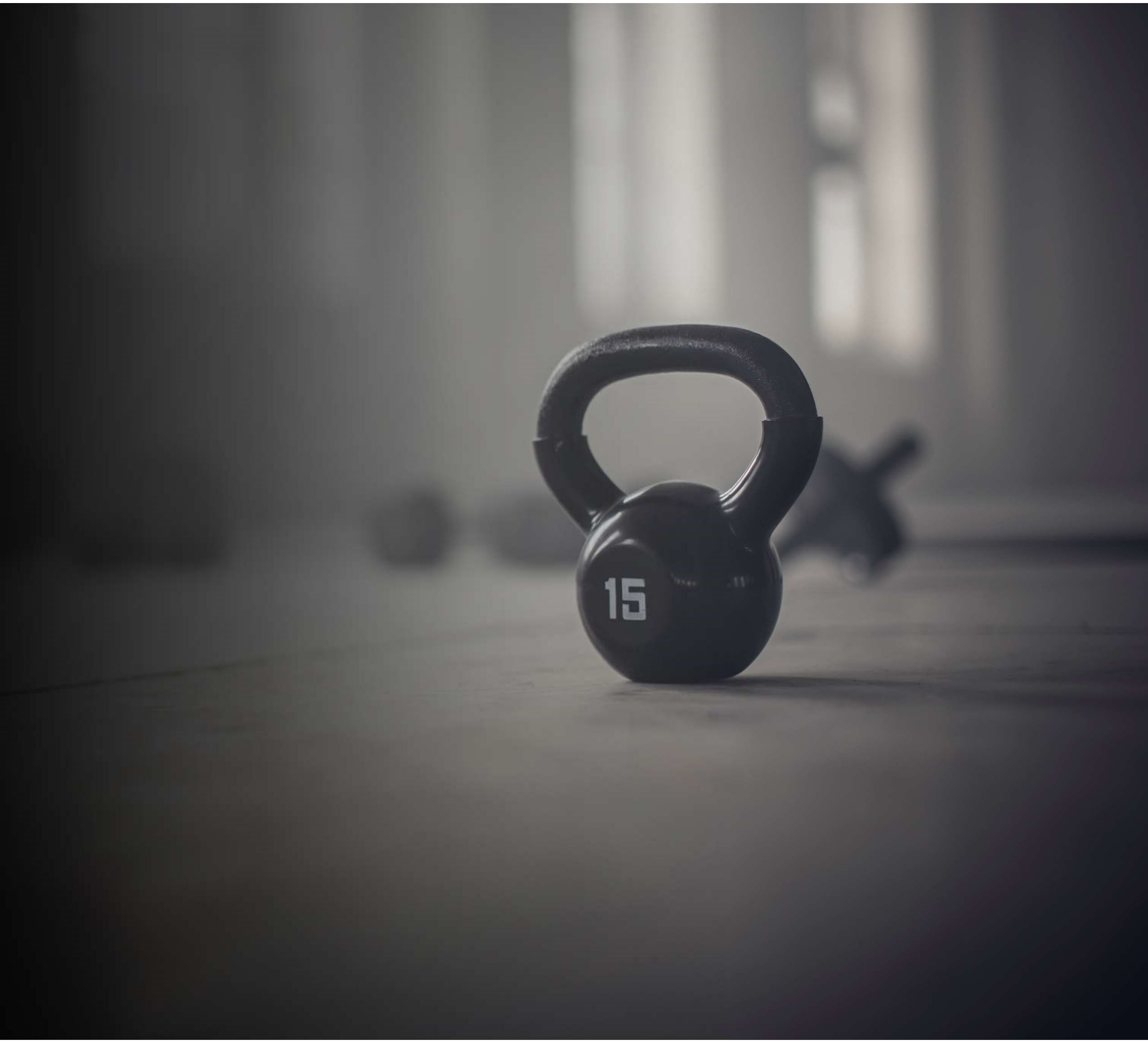
Battle of the Neighborhoods – An Analysis of Fitness in San Francisco

An analysis of the most ideal San Francisco Neighborhoods to open a brand-new fitness center.



Introduction

- According to US News in 2019, San Francisco was considered the 4th fittest city in the United States.
- With the trend of fitness and healthy living, now would be the best time to open a new fitness center. With the intention of opening many more!!
- This will be an analysis of each the San Francisco neighborhoods to isolate key locations.

Business Problem

- With finding the most optimal locations to open a new fitness center, these are the following questions that will need to be answered.
  - Using FourSquare API, can we get a visual map of different locations with the nearest venues
  - From those venues, how many gyms and fitness centers are the the most common in those neighborhoods
  - Can we conduct an analysis that can isolate neighborhoods?

Data

- To conduct this analysis we will be using data from the following locations.
  1. For San Francisco zip code and neighborhood data, we will be using the following url:
     a. http://www.healthysf.org/bdi/outcomes/zipmap.htm
  2. Will use pgeocode package to get latitude and longitude data
  3. Using the Foursquare API, we will gather the following venue data.
     a. Name of Venue
     b. Venue Latitude
     c. Venue Longitude
     d. Venue Category

Methodology(Part I)

- Data Collection and Cleaning
  - Using the request package, we begin by scraping the data from the [SF ZIP Map (healthysf.org)]() URL.
  - From there we read the zip code and neighborhood data into a pandas data frame.
  - We then clean the data in the data frame to remove and unnecessary data points.

In [90]:
```
# scrap the data from the url
url = "http://www.healthysf.org/bdi/outcomes/zipmap.htm"
san_fran_url = requests.get(url).text
soup = BeautifulSoup(san_fran_url,'lxml')
table = soup.find_all("table")

# move the data into a dataframe
san_fran_df = pd.read_html(str(table))

# clean the dataframe to fit
san_fran_df = pd.DataFrame(san_fran_df[4])
san_fran_df.columns = san_fran_df.iloc[0]
san_fran_df = san_fran_df.iloc[1:]
san_fran_df.drop(index = san_fran_df.index[21],axis = 0, inplace = True)
san_fran_df = san_fran_df.iloc[:,0:2]
san_fran_df.head()
```

Out[90]:

	Zip Code	Neighborhood
1	94102	Hayes Valley/Tenderloin/North of Market
2	94103	South of Market
3	94107	Potrero Hill
4	94108	Chinatown
5	94109	Polk/Russian Hill (Nob Hill)

Methodology(Part II)

- Pgeocode package
  - After the data frame has been collected and clean, we use the pgeocode library to start gathering latitude and longitude data.
  - Once that data is collected, it is added back into the San Francisco data frame.

In [91]:

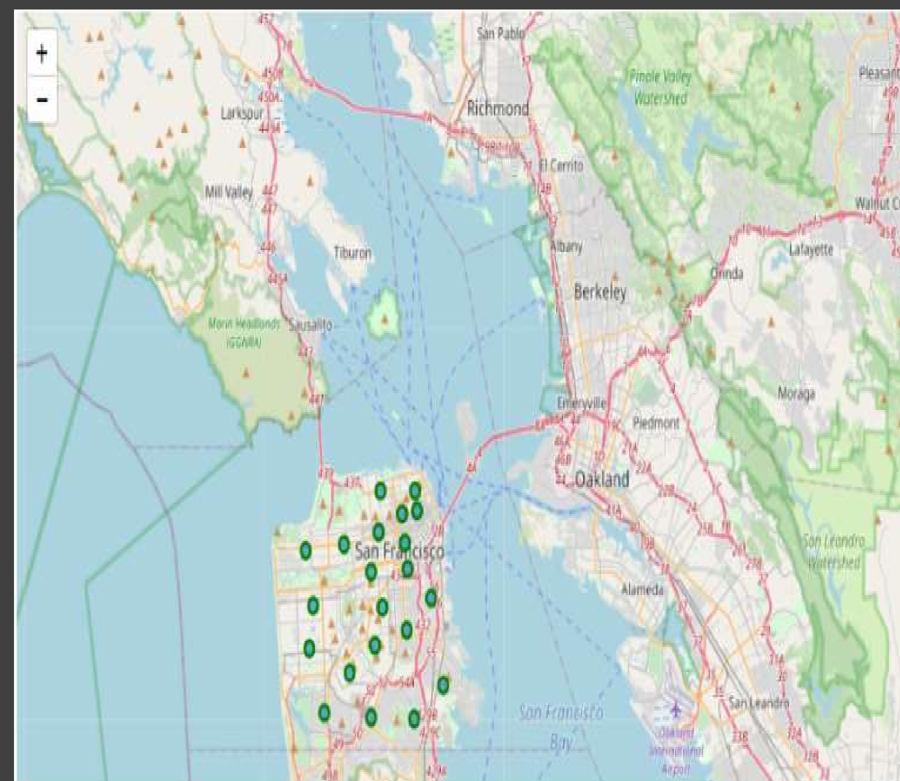
```
# Will use the pgeocode library to get the latitude and longitude coordinates for each neighborhood
nomi_object = pgeocode.Nominatim('us')
latitude = []
longitude = []

for index,row in san_fran_df.iterrows():
    zipcode = nomi_object.query_postal_code(row["Zip Code"])
    latitude.append(zipcode.latitude)
    longitude.append(zipcode.longitude)

san_fran_df["Latitude"] = latitude
san_fran_df["Longitude"] = longitude

san_fran_df.head()
```

Out[91]:

	Zip Code	Neighborhood	Latitude	Longitude
1	94102	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167
2	94103	South of Market	37.7725	-122.4147
3	94107	Potrero Hill	37.7621	-122.3971
4	94108	Chinatown	37.7929	-122.4079
5	94109	Polk/Russian Hill (Nob Hill)	37.7917	-122.4186

Methodology(Part III)

- <u>Visualizing the Data with Folium</u>
  - With the updated data frame, it can now be mapped.
  - To do this, we use the Folium package to build a visual map of the data using the latitude and longitude data.



Methodology(Part IV)

- <u>Foursquare API</u>
  - In order to get nearest venue data, we need to use the Foursquare API. This API gathers data regarding the nearest venue to a particular location.
  - We can use the API to insert a new data frame that includes venue name, venue category, venue latitude and longitude.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Latitude	Venue Longitude	Venue Name	Venue Category
0	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	37.780178	-122.416505	Asian Art Museum	Art Museum
1	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	37.782751	-122.415656	Ales Unlimited: Beer Basement	Beer Bar
2	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	37.783084	-122.417650	Saigon Sandwich	Sandwich Place
3	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	37.781266	-122.416901	Philz Coffee	Coffee Shop
4	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	37.782896	-122.418897	Brenda's French Soul Food	Southern / Soul Food Restaurant

Methodology(Part V)

- EDA
  - After the data frame has been updated, we can begin the explore this data.
  - First, we take a look at the number of venues in each neighborhood.
  - Since we have some categorical data, we need to convert it into a numeric value. Using one hot encoding, we can convert each category into a number. This needs to be done to create our model.
  - Since we need to look at the most common venues, we need to convert the data frame into the frequency of each category and then calculate the most common venues in a particular neighborhood. This data frame will be used to create our model.

Out[99]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bayview-Hunters Point	Park	Southern / Soul Food Restaurant	Light Rail Station	Chinese Restaurant	Pharmacy	Theater	Grocery Store	BBQ Joint	Mexican Restaurant	Gym
1	Castro/Noe Valley	Gay Bar	Thai Restaurant	Coffee Shop	Yoga Studio	Pharmacy	Flower Shop	Mediterranean Restaurant	New American Restaurant	Clothing Store	Convenience Store
2	Chinatown	Chinese Restaurant	Bakery	Hotel	Coffee Shop	Vietnamese Restaurant	Dim Sum Restaurant	Tea Room	Cocktail Bar	Szechuan Restaurant	Bank
3	Haight-Ashbury	Coffee Shop	Boutique	Park	Ice Cream Shop	Bakery	Bus Stop	Breakfast Spot	Bubble Tea Shop	Mexican Restaurant	Burrito Place
4	Hayes Valley/Tenderloin/North of Market	Vietnamese Restaurant	Sandwich Place	Hotel	Thai Restaurant	Theater	Hotel Bar	Coffee Shop	Beer Bar	Concert Hall	Bakery

Methodology(Part VI)

- <u>Kmean Cluster</u>
  - <u>Using the calculate most common values data frame, we begin building our data frame.</u>
  - <u>For the number of clusters we chose 2 clusters as the number of neighborhoods in San Francisco is quite small.</u>
  - <u>The Cluster labels are then added to the data frame</u>

	Zip Code	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Co
0	94102	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	1	Vietnamese Restaurant	Sandwich Place	Hotel	Thai Restaurant	Theater	Hotel Bar	Coffee Shop	Be
1	94103	South of Market	37.7725	-122.4147	1	Nightclub	Coffee Shop	Food Truck	Gay Bar	Cocktail Bar	Sushi Restaurant	Pizza Place	Rent Lo
2	94107	Potrero Hill	37.7621	-122.3971	1	Breakfast Spot	Coffee Shop	Café	Brewery	Cosmetics Shop	Office	Grocery Store	Rest
3	94108	Chinatown	37.7929	-122.4079	1	Chinese Restaurant	Bakery	Hotel	Coffee Shop	Vietnamese Restaurant	Dim Sum Restaurant	Tea Room	C
4	94109	Polk/Russian Hill (Nob Hill)	37.7917	-122.4186	1	Grocery Store	Thai Restaurant	Italian Restaurant	Massage Studio	Vietnamese Restaurant	Bakery	French Restaurant	

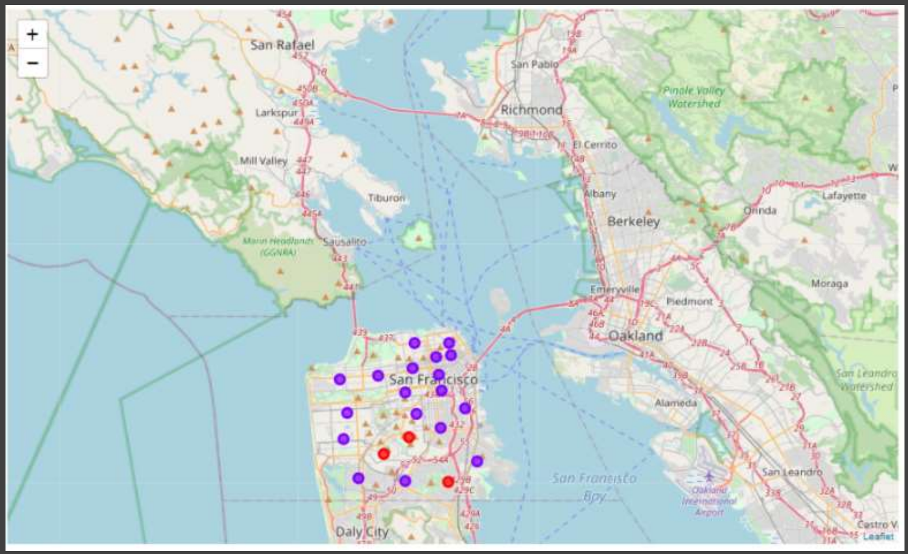
Methodology(Part VI)

- Kmean Cluster
  - Using the calculate most common values data frame, we begin building our data frame.
  - For the number of clusters we chose 2 clusters as the number of neighborhoods in San Francisco is quite small.
  - The Cluster labels are then added to the data frame

	Zip Code	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Co
0	94102	Hayes Valley/Tenderloin/North of Market	37.7813	-122.4167	1	Vietnamese Restaurant	Sandwich Place	Hotel	Thai Restaurant	Theater	Hotel Bar	Coffee Shop	Be
1	94103	South of Market	37.7725	-122.4147	1	Nightclub	Coffee Shop	Food Truck	Gay Bar	Cocktail Bar	Sushi Restaurant	Pizza Place	Rent Lo
2	94107	Potrero Hill	37.7621	-122.3971	1	Breakfast Spot	Coffee Shop	Café	Brewery	Cosmetics Shop	Office	Grocery Store	Rest
3	94108	Chinatown	37.7929	-122.4079	1	Chinese Restaurant	Bakery	Hotel	Coffee Shop	Vietnamese Restaurant	Dim Sum Restaurant	Tea Room	C
4	94109	Polk/Russian Hill (Nob Hill)	37.7917	-122.4186	1	Grocery Store	Thai Restaurant	Italian Restaurant	Massage Studio	Vietnamese Restaurant	Bakery	French Restaurant	

Results

- Now that the model is built we can take a visual look at the clusters. We separated each cluster by color.
  - Cluster 1 is Red
  - Cluster 2 is Purple



Discussion

- By analyzing each of the clusters, we can see each of the characteristics for these particular neighborhoods
  - For Cluster #1, it is clear that this cluster is mainly for outdoor activities with the higher number of trails and parks.
  - Cluster #2 seems to be where the higher populations seem to be with more indoor venues such as restaurants, coffee shops and even gyms.

Conclusion

- After analyzing the data of the San Francisco neighborhoods, it would appear the cluster #2 would be the ideal locations to open a new fitness center.
- Cluster #2 has more accessibility to public transportation and would ideally have more foot traffic that would be a perfect location for a new fitness center!