Applied Data Science - Capstone Project

Finding suitable locations to open a Gym in Singapore,

- ## Introduction / Business Problem

  The aim of this project is to find suitable locations to open a gym in the Singapore greater metropolitan area.

  The first requirement is that the new gym should be easily accessible by its prospective customers and more specifically it should be located near a metro station. The number of gyms already existing in an area should also be considered so that fierce competition be avoided if possible.

  Apart from the obvious intended stakeholders, entrepreneur looking to start a gym business, similar methodology could be used for other specific types of businesses. It can serve as an initial starting point of locations to consider starting their business.

  For the project objectives to be achieved, python geolocation libraries will used, along with the Foursquare API. Also, in order to create clusters of similar candidate locations, the KMeans machine learning clustering algorithm will be used.

- ## Data

The necessary data for this project, based on the above stated requirements, are:

- The metro stations in the Singapore greater metropolitan area
- Number of existing gyms near each station
- In addition, the distance to the nearest gym for every metro station will be used

In order to obtain the data, a combination of the geopy Python library and the Foursquare API will be used:

1. 'Central Area' will be considered as the center of Singapore. It is indeed one of the most central location in the city. I will obtain its geospatial coordinates using the geopy library
2. Having the coordinates of the 'center' of Singapore, the Foursquare API will be used to retrieve data for all the metro stations in Singapore greater area in a radius of 15 km
3. To find the existing gyms near the metro stations, the Foursquare API will again be used for every station. I will obtain data for all the gyms located in a radius of 750 meters of every metro station

Using the collected data, I will calculate the number of existing gyms near each station. I will also be able to determine the minimum distance to a gym for every metro station from the 3rd step of the above process. This minimum distance to every metro station from a gym, along with the number of already existing gyms near the station will be used as input to KMeans clustering algorithm to obtain the clusters of areas (metro stations).

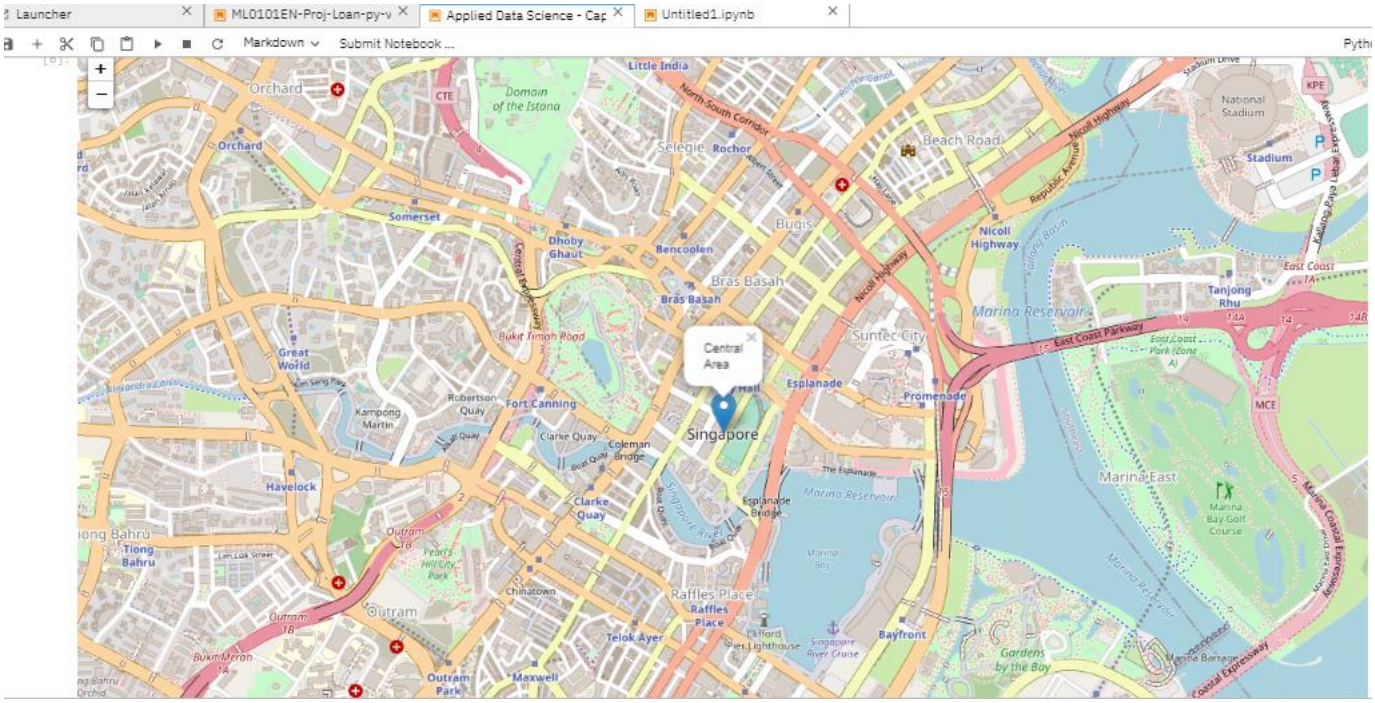
- ## Methodology

The objective of this project is to obtain information about metro stations in the greater metropolitan Singapore area with potential for opening a gym, and having as criteria:

- Low number of already existing gyms
- Minimum distance of each station to its nearby gyms

The steps I followed to identify potential areas (metro stations) were:

1. Considered Central Area as the 'center' of Singapore (indeed probably the most central location of the city) and acquired its latitude and longitude geospatial coordinates.



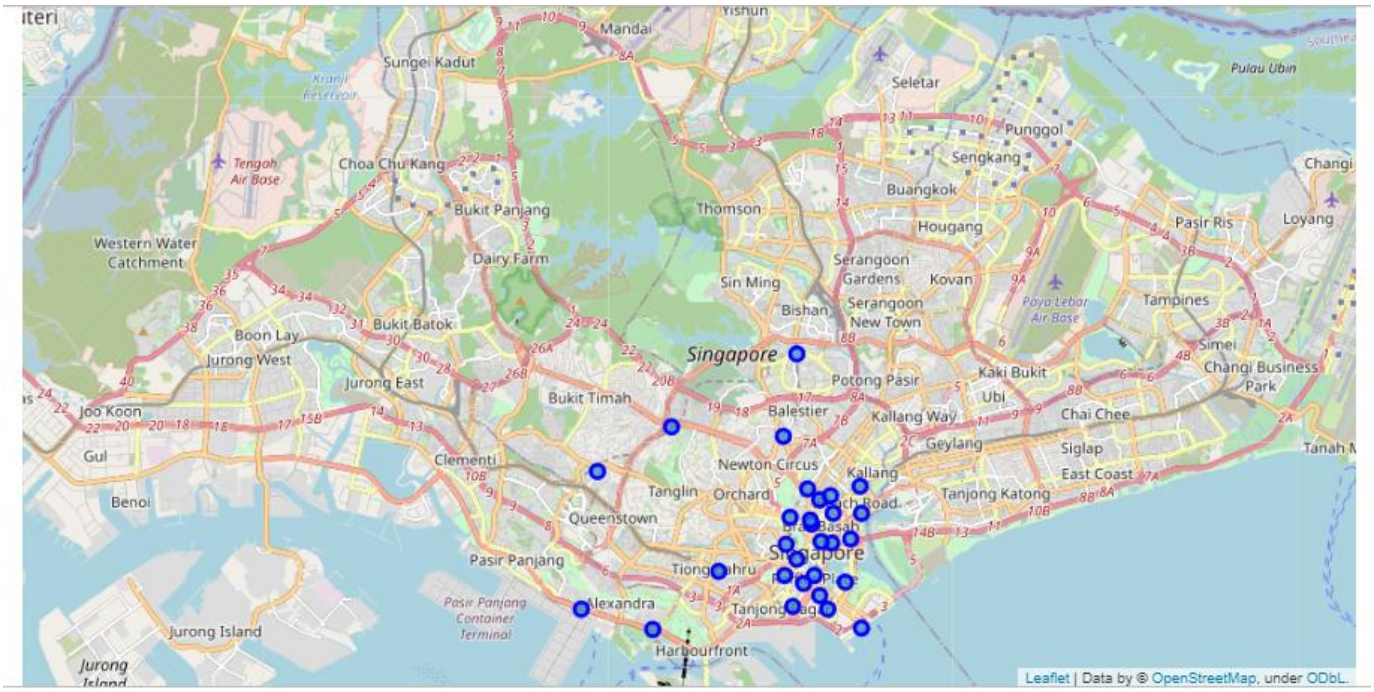
2. Based on the coordinates of Central Area, I obtained information about metro stations in a radius of 15 km using the Foursquare API. At this stage I removed from the above dataset 10 rows of data that although they are identified as 'Metro stations' by the Foursquare API, they are only used as depots or maintenance gathering for the metro carriages.

[14]:
```
stations_df[stations_df['name'].str.find('Station') == -1]
```

[14]:

	name	lat	lng	distance	postalCode	venue_type
0	Bugis MRT Interchange (EW12/DT14)	1.300476	103.856094	1201	188022	Metro Station
1	Raffles Place MRT Interchange (EW14/NS26)	1.284516	103.851446	666	048618	Metro Station
4	Bayfront MRT Interchange (CE1/DT16)	1.283062	103.859167	1144	018970	Metro Station
5	Botanic Gardens MRT Interchange (CC19/DT9)	1.322324	103.814880	5446	257494	Metro Station
11	Promenade MRT Interchange (CC4/DT15)	1.293731	103.860465	1005	039193	Metro Station
15	City Hall MRT Interchange (EW13/NS25)	1.293146	103.853053	318	179100	Metro Station
17	Dhoby Ghaut MRT Interchange (CC1/NE6/NS24)	1.299225	103.845343	1226	238826	Metro Station
19	Marina Bay MRT Interchange (NS27/CE2)	1.276144	103.854788	1624	018990	Metro Station
24	Chinatown MRT Interchange (NE4/DT19)	1.284482	103.843842	1129	059443	Metro Station
29	Little India MRT Interchange (NE7/DT12)	1.306457	103.849606	1799	229900	Metro Station

A visualization of the remaining metro stations on an Singapore city map:



3. After the collection of metro stations information, I again utilized the Foursquare API to locate all the existing gyms in a radius of 750 meters from each station. The resulting subcategories of businesses found were:

```
gyms_venues_df.groupby(['Venue Category']).count()
```

Venue Category	station	lat	lng	Venue	Venue Latitude	Venue Longitude	Distance from Station
Building	1	1	1	1	1	1	1
Climbing Gym	3	3	3	3	3	3	3
College Gym	3	3	3	3	3	3	3
Gym	156	156	156	156	156	156	156
Gym / Fitness Center	317	317	317	317	317	317	317
Gym Pool	9	9	9	9	9	9	9
Gymnastics Gym	1	1	1	1	1	1	1
Hotel Pool	3	3	3	3	3	3	3
Martial Arts Dojo	31	31	31	31	31	31	31
Pizza Place	1	1	1	1	1	1	1
Residential Building (Apartment / Condo)	1	1	1	1	1	1	1
Spa	5	5	5	5	5	5	5
Sporting Goods Shop	4	4	4	4	4	4	4
Track	1	1	1	1	1	1	1
Track Stadium	1	1	1	1	1	1	1
Wine Bar	2	2	2	2	2	2	2
Yoga Studio	70	70	70	70	70	70	70

- I kept as my data set the results that correspond only to ‘Gym / Fitness Center’ and ‘Gym’ subcategories. I removed the rest of the subcategories such as ‘Dance Studio’, ‘Yoga Studio’, ‘Martial Arts Dojo’ etc.
- I ignored for the purposes of clustering below metro stations that based on the results of the Foursquare API don’t have any existing gyms in their vicinity. There can either exist no data in the Foursquare database, or indeed there are no existing gyms near the corresponding stations.

```
missing_stations_df = pd.merge(stations_df[['station', 'lat', 'lng']], gyms_df,
                               on='station', how='left')
missing_stations_df[missing_stations_df['Gym Count'].isnull()]
```

	station	lat_x	lng_x	lat_y	lng_y	Min Distance from Station	Gym Count
14	Marina South Pier MRT Station	1.271067	103.863216	NaN	NaN	NaN	NaN

The resulting data set will also contain the distance of each gym to the corresponding station.

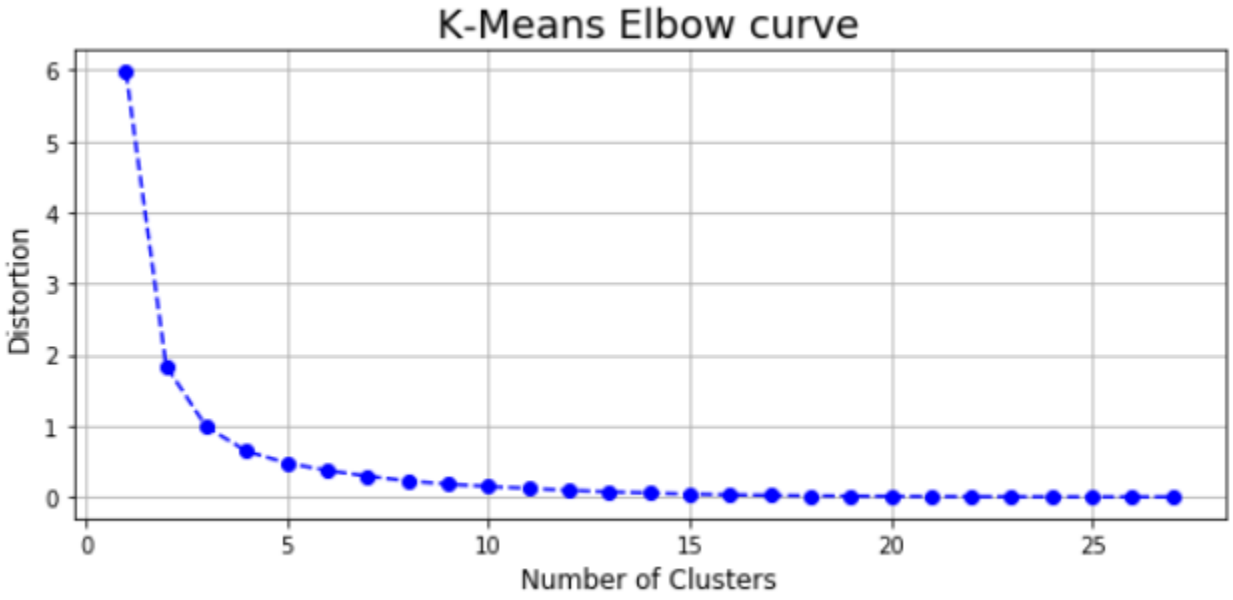
4. Having the information about gyms around metro stations, I calculated the number of existing gyms near each station as well as the minimum distance from each station to a gym using available python statistical functions. Part of the data set containing the minimum distance and number of existing gyms for each station:

```
stations_gyms_grouped_df.head()
```

	station	lat	lng	Min Distance from Station	Gym Count
0	Bugis MRT Interchange	1.300476	103.856094	187	24
24	Raffles Place MRT Interchange	1.284516	103.851446	11	27
51	Esplanade MRT Station	1.292907	103.855946	106	28
79	Telok Blangah MRT Station	1.270729	103.809998	365	4
83	Bayfront MRT Interchange	1.283062	103.859167	119	6

5. The data will be normalized so that both factors (minimum distance, number of existing gyms) will have equal weight when they will be used by a machine learning method.

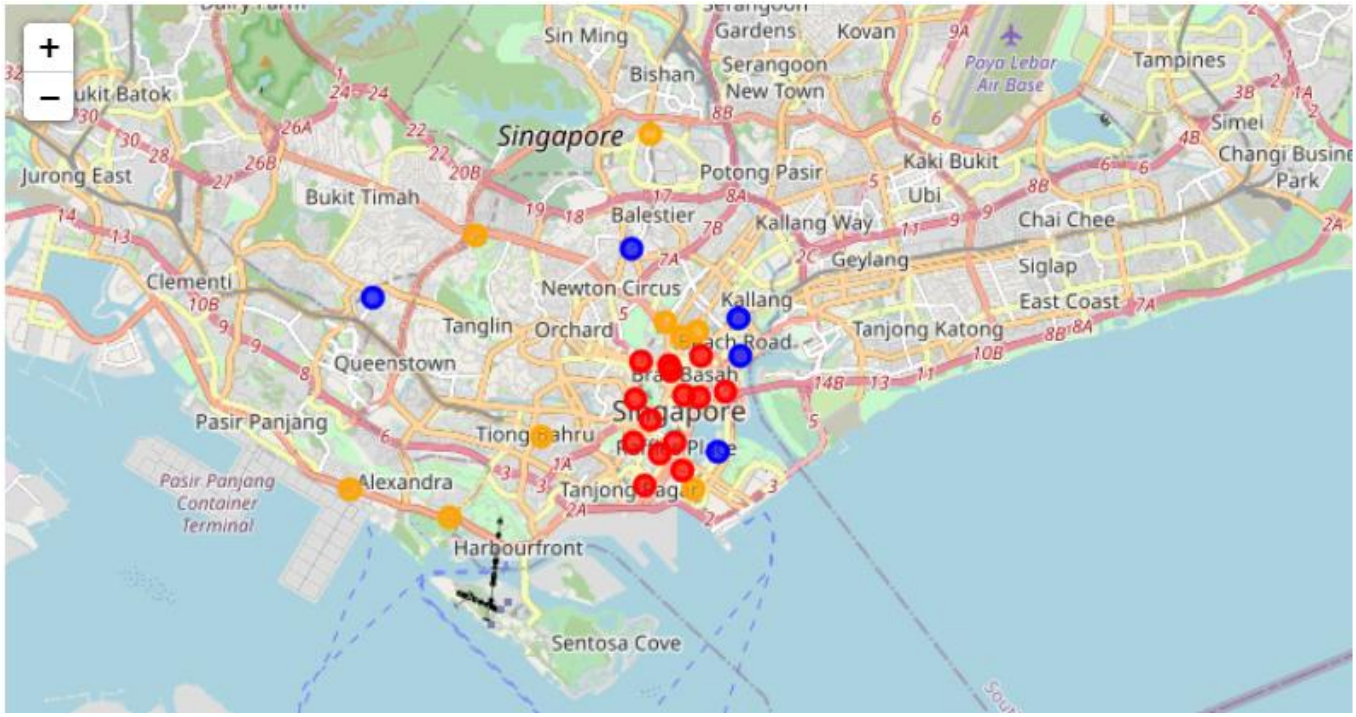
6. The K-Means Machine Learning clustering algorithm will be used to divide the stations and gyms data set into clusters of similar locations. The elbow method will be used to find the most suitable number of clusters



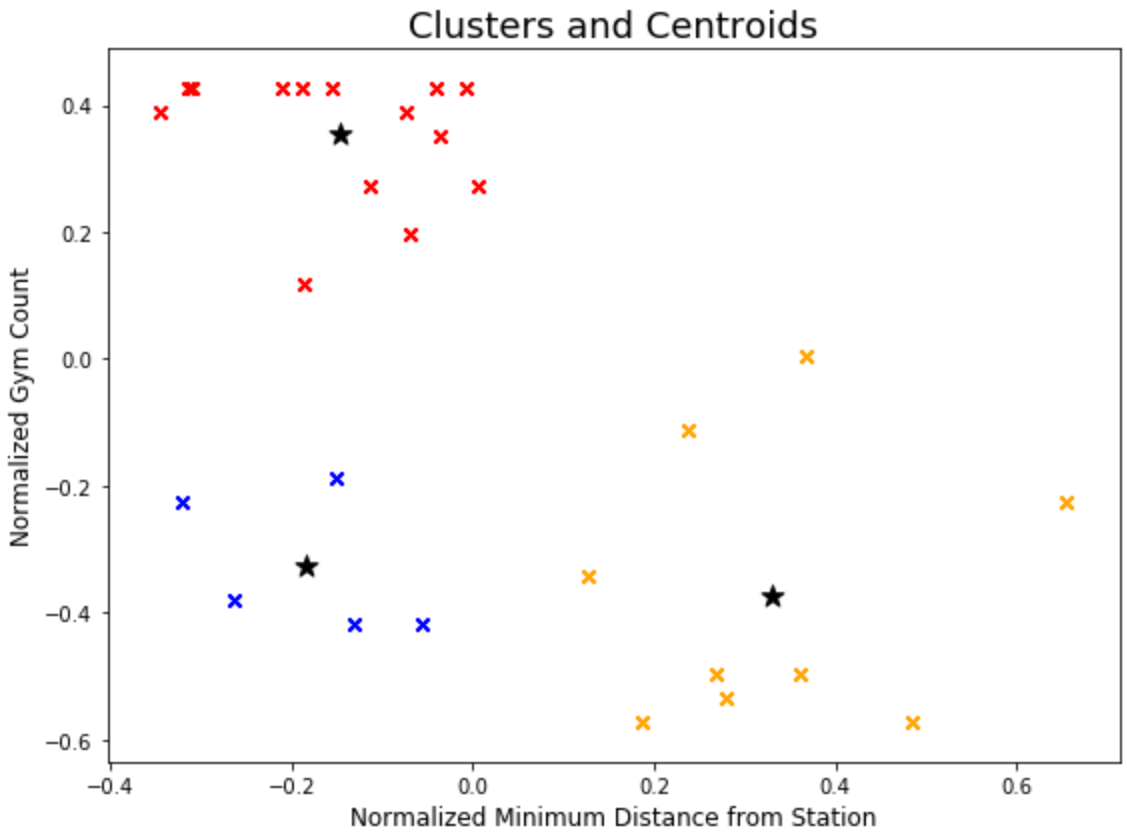
Although the elbow curve is not very steep, an elbow point of 3 clusters is clear, so this is the number of clusters that will be used for the K-Means clustering algorithm

- Results

After executing the K-Means clustering algorithm three clusters of Metro stations were created, identified by their respective colors on the following map



And a graph of the distribution of clusters and their final centroids (center points) in black, based on normalized values of minimum distance and number of existing gyms.



The three Metro stations clusters can be described as follows:

Cluster 1 [Cluster Label 0] – AVERAGE potential (color Orange on the map)

Although not a prohibitive metro station to open a gym in its vicinity, there is already a fair number of gyms in the area and the nearest one is not far from the metro station.

Examples:

	station	Cluster Label	lat	lng	Min Distance from Station	Gym Count	Norm Min Distance from Station	Norm Gym Count
3	Telok Blangah MRT Station	0	1.270729	103.809998	365	4	0.360202	-0.495879
5	Botanic Gardens MRT Interchange	0	1.322324	103.814880	278	2	0.186895	-0.572802
12	Tiong Bahru MRT Station	0	1.285758	103.826982	325	3	0.280521	-0.534341
15	Pasir Panjang MRT Station	0	1.276075	103.791973	427	2	0.483708	-0.572802
18	Marina Bay MRT Interchange	0	1.276144	103.854788	303	14	0.236696	-0.111264
19	Braddell MRT Station	0	1.340742	103.847020	319	4	0.268569	-0.495879
24	Rochor MRT Station	0	1.303877	103.852635	369	17	0.368170	0.004121
25	Jalan Besar MRT Station	0	1.305004	103.855339	513	11	0.655023	-0.226648
27	Little India MRT Interchange	0	1.306457	103.849606	248	8	0.127134	-0.342033

Cluster 2 [Cluster Label 1] – LOW potential (color Red on the map)

There are already many existing gyms in the area and the nearest gym is in most cases in a relatively short distance from the station.

Examples:

	station	Cluster Label	lat	lng	Min Distance from Station	Gym Count	Norm Min Distance from Station	Norm Gym Count
0	Bugis MRT Interchange	1	1.300476	103.856094	187	24	0.005620	0.273352
1	Raffles Place MRT Interchange	1	1.284516	103.851446	11	27	-0.344977	0.388736
2	Esplanade MRT Station	1	1.292907	103.855946	106	28	-0.155734	0.427198
6	Telok Ayer MRT Station	1	1.282401	103.848756	26	28	-0.315097	0.427198
7	Clarke Quay MRT Station	1	1.288804	103.846968	29	28	-0.309121	0.427198
11	Promenade MRT Interchange	1	1.293731	103.860465	91	20	-0.185615	0.119505
13	Fort Canning MRT Station	1	1.292371	103.844272	164	28	-0.040196	0.427198
14	City Hall MRT Interchange	1	1.293146	103.853053	90	28	-0.187607	0.427198
16	Dhoby Ghaut MRT Interchange	1	1.299225	103.845343	149	22	-0.070077	0.196429
17	Bras Basah MRT Station	1	1.297506	103.850506	147	27	-0.074061	0.388736
20	Bencoolen MRT Station	1	1.298700	103.850176	127	24	-0.113902	0.273352
22	Chinatown MRT Interchange	1	1.284482	103.843842	166	26	-0.036212	0.350275
23	Downtown MRT Station	1	1.279544	103.852828	180	28	-0.008324	0.427198
26	Tanjong Pagar MRT Station	1	1.276543	103.845943	79	28	-0.209519	0.427198

Cluster 3 [Cluster Label 2] – HIGH potential (color Blue on the map)

There are not many already existing gyms in the area and the nearest gym is in most cases relatively not in a short distance to the metro station

Examples:

	station	Cluster Label	lat	lng	Min Distance from Station	Gym Count	Norm Min Distance from Station	Norm Gym Count
4	Bayfront MRT Interchange	2	1.283062	103.859167	119	6	-0.129838	-0.418956
8	Nicoll Highway MRT Station	2	1.300246	103.863451	156	6	-0.056133	-0.418956
9	Novena MRT Station	2	1.320086	103.843592	23	11	-0.321073	-0.226648
10	Lavender MRT Station	2	1.307418	103.862860	108	12	-0.151750	-0.188187
21	Holland Village MRT Station	2	1.311008	103.795959	52	7	-0.263304	-0.380495

- ## Discussion

  Clusters of areas (in our case Metro stations) were identified as groups of similar in their potential locations for opening a gym.

  Possible areas that were not in the Foursquare database should also be examined so that it can be determined if it is just lack of data about these stations or indeed there are no gyms in the vicinity of the stations.

  A lot more factors can be considered when choosing an appropriate location. Some examples of extra factors can be:

  - Population density in the area
  - Number of businesses operating in the area (people may want to go to a gym close to work)
  - Average age and household income in the area
  - Property prices in the area

- ## 6. Conclusion

  The above results can be a good starting point for a prospective businessman that is interested in opening a gym. Similar methodology can be used for other types of businesses probably with customized criteria.

  With the availability of a number of different tools and Machine Learning algorithms, it is possible to find solutions (or possible solutions) to an ever-increasing number of problems and queries.

  And it is getting better and better!