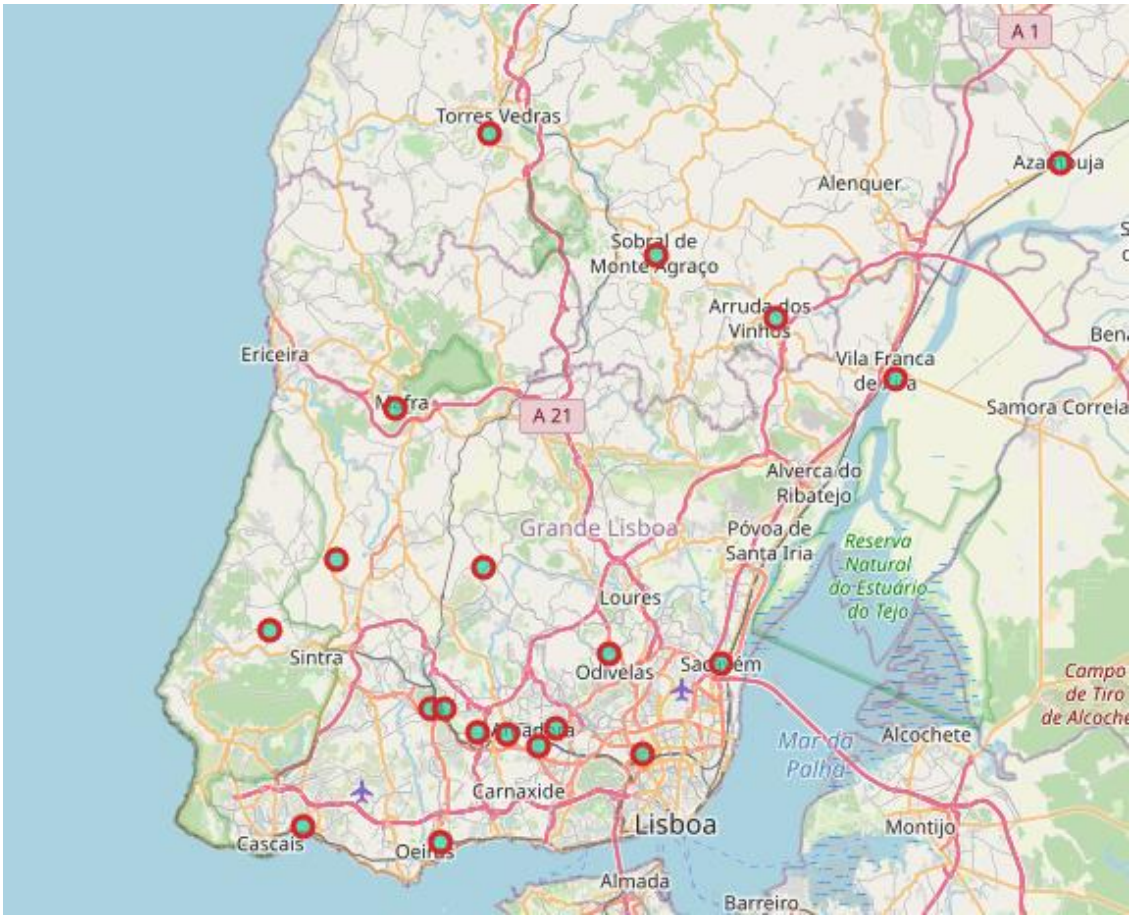
Study about the best location for the opening of a new gym in Lisbon



João Lopes for Coursera Capstone Project

1. Introduction 3
1.1 Business Problem 3
2.0 Data 4
3.0 Methodology 6
4.0 Results 9
5.0 Discussion 11
6.0 Conclusion 12

1. Introduction

Lisbon is the capital district of Portugal, making it the most important financially and the most populous, creating a lot of business opportunities. Nevertheless, Lisbon is the area with the higher price per square meter in all of the country, creating the necessity of a good prior analysis to choose the best location to open a new venue.

1.1. Business Problem

Being expected a higher percentage of the population to have obesity, after the lockdown imposed by the current pandemic disease (covid-19), I believe that there will be a greater search for gyms as it happened after the first lockdown in the country.

Having this special occasion, it will be a great opportunity to open a new facility in this city, that fulfill the necessities of the population.

My objective with this project is to localize the best spot in Lisbon to open a new facility, that ideally will be located in an area with lower offer of this kind of facilities, an area that have a low ratio of gym per inhabitant and also an area that have a low rating for the current opportunities.

2. Data

The data used was obtained in two websites:

- The first file that I used was a CSV obtained in https://simplemaps.com/data/world-cities, with all the geographic and demographic data necessary about the cities in the Lisbon District, creating a database with the following columns:

1)City Name

2)City Latitude

3)City Longitude

4)City Population

Table 1 – Database of cities of the world (sample)

	city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
0	Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN	Tōkyō	primary	37977000.0	1392685764
1	Jakarta	Jakarta	-6.2146	106.8451	Indonesia	ID	IDN	Jakarta	primary	34540000.0	1360771077
2	Delhi	Delhi	28.6600	77.2300	India	IN	IND	Delhi	admin	29617000.0	1356872604
3	Mumbai	Mumbai	18.9667	72.8333	India	IN	IND	Mahārāshtra	admin	23355000.0	1356226629
4	Manila	Manila	14.5958	120.9772	Philippines	PH	PHL	Manila	primary	23088000.0	1608618140

- To get the data about the gyms I used the Foursquare API, serching for gyms in a radius of 500m from the city coords and obtained a database with the folowing columns:

1)Venue Name

2)Venue Latitude

3)Venue Longitude

Table 2 – Database of gyms in Lisbon District (sample)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lisbon	38.7452	-9.1604	Malo Clinic	38.746271	-9.163602	Medical Center
1	Lisbon	38.7452	-9.1604	Fitness Center	38.745469	-9.159747	Gym / Fitness Center
2	Lisbon	38.7452	-9.1604	Marriott Gym	38.746945	-9.164175	Gym / Fitness Center
3	Lisbon	38.7452	-9.1604	Malo Clinic Sports	38.745799	-9.164073	Gym
4	Amadora	38.7500	-9.2333	Soul Lifting	38.747696	-9.236370	Gym
5	Queluz	38.7566	-9.2545	Life Gymnasium	38.756580	-9.252048	Gym
6	Queluz	38.7566	-9.2545	Workit24hours	38.757268	-9.251929	Gym
7	Cacém	38.7704	-9.3081	Olival Gym	38.770185	-9.305836	Gym / Fitness Center
8	Cacém	38.7704	-9.3081	+ Leve	38.769888	-9.303864	Gym / Fitness Center

3. Methodology

To start with the data of this project, it was necessary to clean the data from the CSV file, with the world cities information. The first steps of this cleaning were choosing only the cities of Portugal and then only the Lisbon district. These assured that I didn't have data of another Lisbon district in another country. Although the dataset was reduced substantially in the number of rows, I still had a lot of useless information in the 11 columns. So, I created the final dataset only selecting the columns with the name of the cities (city), latitude (lat), longitude (lng) and the number of residents (population).

These two cleaning steps transformed the original dataset with 26569 rows and 11 columns to a reduced dataset of 21 rows and 4 columns.

Out[2]:

	city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
0	Tokyo	Tokyo	35.6897	139.6922	Japan	JP	JPN	Tōkyō	primary	37977000.0	1392685764
1	Jakarta	Jakarta	-6.2146	106.8451	Indonesia	ID	IDN	Jakarta	primary	34540000.0	1360771077
2	Delhi	Delhi	28.6600	77.2300	India	IN	IND	Delhi	admin	29617000.0	1356872604
3	Mumbai	Mumbai	18.9667	72.8333	India	IN	IND	Mahārāshtra	admin	23355000.0	1356226629
4	Manila	Manila	14.5958	120.9772	Philippines	PH	PHL	Manila	primary	23088000.0	1608618140

```
In [3]: df.shape
Out[3]: (26569, 11)

In [4]: portugal=df.loc[df['country'] == 'Portugal']

In [5]: lisbon=df.loc[df['admin_name'] == 'Lisboa']

In [6]: lisbon.head()
```

Out[6]:

	city	city_ascii	lat	lng	country	iso2	iso3	admin_name	capital	population	id
748	Lisbon	Lisbon	38.7452	-9.1604	Portugal	PT	PRT	Lisboa	primary	506654.0	1620619017
2909	Amadora	Amadora	38.7500	-9.2333	Portugal	PT	PRT	Lisboa	minor	175136.0	1620896557
2993	Oeiras	Oeiras	38.6970	-9.3017	Portugal	PT	PRT	Lisboa	minor	172120.0	1620375757
3400	Odivelas	Odivelas	38.8000	-9.1833	Portugal	PT	PRT	Lisboa	minor	144549.0	1620010482
3638	Vila Franca de Xira	Vila Franca de Xira	38.9500	-8.9833	Portugal	PT	PRT	Lisboa	minor	136886.0	1620859041

```
In [7]: lisbon.shape
Out[7]: (21, 11)

In [8]: imp_lisbon=lisbon[['city','lat','lng','population']].reset_index(drop=True)

In [9]: imp_lisbon.shape
Out[9]: (21, 4)
```

Figure 1 – Representation of the cleaning phase for the cities' dataset

To better understand the data, I used the folium library to create a visualization map which enabled me to see the geographical distribution of the cities. It was easy to understand that some cities were very close to each other, so the presence of a gym in one of them could surpress the needs of the other (ex: Agualva and Cacém). On the other hand, there were cities too distant from each other, creating the necessity that both cities must have a gym (ex: Estoril and Azambuja). Having these information, I decided to create proximity clusters for the cities, creating 4 groups, using KMeans Cluster from sklearn.cluster library.

```
In [14]: kmeans = KMeans(n_clusters=4, random_state=0).fit(imp_lisbon[['lat','lng']])
         kmeans.labels_

Out[14]: array([0, 0, 2, 0, 1, 0, 2, 2, 2, 2, 0, 0, 3, 2, 1, 3, 1, 3, 2, 2, 2],
               dtype=int32)

In [15]: imp_lisbon.insert(0, 'Cluster Labels', kmeans.labels_)

In [16]: imp_lisbon
```

Out[16]:

	Cluster Labels	city	lat	lng	population
0	0	Lisbon	38.7452	-9.1604	506654
1	0	Amadora	38.7500	-9.2333	175136
2	2	Oeiras	38.6970	-9.3017	172120
3	0	Odivelas	38.8000	-9.1833	144549
4	1	Vila Franca de Xira	38.9500	-8.9833	136886
5	0	Queluz	38.7566	-9.2545	78273
6	2	Cacém	38.7704	-9.3081	21289
7	2	Agualva	38.7700	-9.2988	35824
8	2	Massamá	38.7568	-9.2748	28112
9	2	Estoril	38.7057	-9.3977	26399
10	0	Falagueira	38.7590	-9.2199	14530
11	0	Sacavém	38.7944	-9.1053	18469
12	3	Torres Vedras	39.0833	-9.2667	79465
13	2	Mafra	38.9333	-9.3333	76685
14	1	Azambuja	39.0667	-8.8667	21814
15	3	Cadaval	39.2500	-9.1000	14525
16	1	Arruda dos Vinhos	38.9833	-9.0667	13391
17	3	Sobral de Monte Agraço	39.0167	-9.1500	10156
18	2	Almargem	38.8475	-9.2714	8983
19	2	São Martinho	38.8125	-9.4208	6226
20	2	Terrujem	38.8511	-9.3747	5113

Figure 2 – Creation of the final dataset for Lisbon district cities

Being satisfied with the dataset for the Lisbon district cities, I started the collection of data for the venues in those cities using the Foursquare API.

In [19]:
```
Category='4bf58dd8d48988d175941735'
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?categoryId=4bf58dd8d48988d175941735&client_id={}&clien
t_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

In [20]:
```
lisbon_gyms = getNearbyVenues(names=imp_lisbon['city'],latitudes=imp_lisbon['lat'],longitudes=imp_lisbon['lng'])
```

```
Lisbon
Amadora
Oeiras
Odivelas
Vila Franca de Xira
Queluz
Cacém
Agualva
Massamá
Estoril
Falagueira
Sacavém
Torres Vedras
```

Figure 3 – Collection of venue data

4. Results

Tables 3 and 4 – Frequency of gyms in each city and cities cluster

	Cluster Labels	city	lat	lng	population	frequency
20	2	Terrujem	38.8511	-9.3747	5113	0
2	2	Oeiras	38.6970	-9.3017	172120	0
3	0	Odivelas	38.8000	-9.1833	144549	0
4	1	Vila Franca de Xira	38.9500	-8.9833	136886	0
18	2	Almargem	38.8475	-9.2714	8983	0
17	3	Sobral de Monte Agraço	39.0167	-9.1500	10156	0
16	1	Arruda dos Vinhos	38.9833	-9.0667	13391	0
15	3	Cadaval	39.2500	-9.1000	14525	0
14	1	Azambuja	39.0667	-8.8667	21814	0
19	2	São Martinho	38.8125	-9.4208	6226	0
13	2	Mafra	38.9333	-9.3333	76685	0
12	3	Torres Vedras	39.0833	-9.2667	79465	0
1	0	Amadora	38.7500	-9.2333	175136	1
10	0	Falagueira	38.7590	-9.2199	14530	2
6	2	Cacém	38.7704	-9.3081	21289	2
5	0	Queluz	38.7566	-9.2545	78273	2
11	0	Sacavém	38.7944	-9.1053	18469	2
9	2	Estoril	38.7057	-9.3977	26399	3
8	2	Massamá	38.7568	-9.2748	28112	4
7	2	Agualva	38.7700	-9.2988	35824	4
0	0	Lisbon	38.7452	-9.1604	506654	4

In [33]:
```
decision=complete.groupby('Cluster Labels').sum()
decision
```

Out[33]:

	lat	lng	population	frequency
Cluster Labels				
0	232.6052	-55.1567	937611	11
1	117.0000	-26.9167	172091	0
2	349.1443	-83.9813	380751	13
3	117.3500	-27.5167	104146	0

Looking at these tables, it is easy to see that in the Lisbon district we have 12 cities that doesn't have any gym, 1 city with 1 gym, 4 cities with 2 gyms, 1 city with 3 gyms and 3 cities with 4 gyms.

With a broader perspective, looking for the clusters created, 2 of them (1 and 3) do not have any gym, the cluster 0 have 11 gyms which means 1 gym for 85237 habitants, and the cluster 2 have 13 gyms that represents 1 gym for 29289 habitants.

5. Discussion

During the cleaning phase of the data, I quickly realized a major limitation of Foursquare in Portugal, that is the unfamiliarity with this app causing deficits in the data. This could be observed with the absence of any data about Holmes Place gyms that is the biggest franchise of premium gyms in the country.

Assuming the data extracted from Foursquare as being accurate and updated, the results obtained show a vast opportunity for someone to open this kind of business in the Lisbon district, since the majority of the cities do not offer any facility of this kind and the ones that have, offer a really poor ratio of venues per habitant.

Having to take a final decision, I would see which cluster would have a higher ratio of habitant per gym to choose the area that would create the most opportunities for my business. Looking to the results the first logical conclusion is to choose the cluster 1. Analyzing these cluster and maintaining the same evaluation criteria, the city with the highest population is Vila France de Xira, and that looks like the best option.

Given these results, I don't believe that it is necessary a more profound study, where it would be compared the ratings of the gyms in each city.

6. Conclusion

Knowing the limitations explained previously, about the unfamiliarity with Forsquare App, my suggestion based on the available data, for this study, is that the best city to open a new gym is Vila Franca de Xira, being the most populous and inserted in a cluster without any gym.

Nevertheless, I strongly suggest a future study to construct a more complete database about the location of the existing gyms in Portugal, particularly in Lisbon, to give a more robust opinion about this matter.