Hospitals Investment

Gerardo Navarro Guerrero

1. INTRODUCTION

1.1 Background

The healthcare industry is an integration of sectors that provides goods and services to treat patients with curative, preventive, rehabilitative, and palliative care. The modern healthcare industry includes three essential branches, which are services, products, and finance, and may be divided into many sectors and categories and depends on the interdisciplinary teams of trained professionals to meet the health needs of individuals and populations. This industry is a strategic sector of a country, whereby this is one of the most important areas to do business.

1.2 Problem description

A Corporate group in the Health industry is studying the possibility of invest in the construction and management of five Hospitals. The Corporate wants to know where to allocate that investment, namely, in which countries, and in which cities.

1.3 Interest

This idea comes from personal curiosity to know more about the health industry a make a comparison of the healthcare system between countries.

2. DATA ACQUISITION AND CLEANING

2.1 Data sources

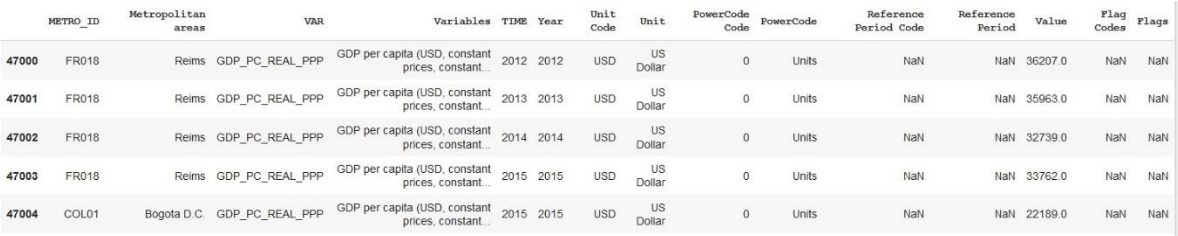
We need general information about the healthcare system of the countries, for example, population, GDP, number of hospitals, and any relevant information that helps us to make an informed decision. It is not easy to find all the data needed in one place. Our primary sources of information will be the OCDE [1] and the World Health Organization (WHO) [2]. Any additional data we will retrieved from various sources, given that those datasets are inconsistent and incomplete.
There are dozens of even hundreds of variables to study the health system of a country. We are going to focus only on some relevant features to assess the suitability of a city to allocate the investment using only with data publicly available online.

We will build two datasets with information about the Countries, and the other with details of the Cities. The primary feature will be the population to decide which cities we will study. With that, we will choose the 20 most populated cities of countries members of the OCDE.
One feature that is difficult to obtain is the number of hospitals in each city, the OCDE and the WHO datasets does not provide this information. In this case, we are going to use the Foursquare API to obtain the number of hospitals given the geolocalization coordinates and a radius of search.

Once we collected and prepared the data, we are going to use clustering to determine if there is a cluster of cities more suitable for the allocation of the investment.

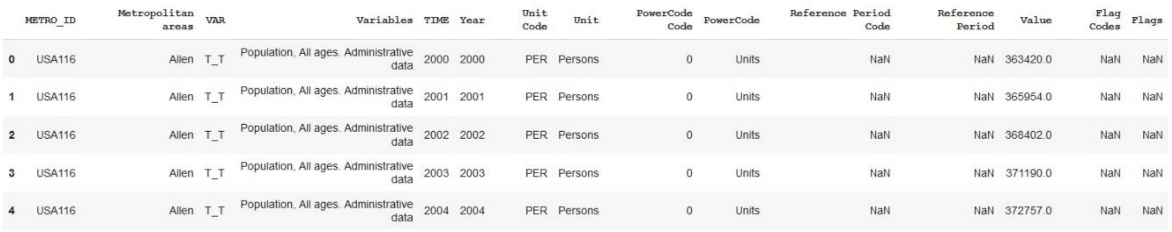
2.2 Data Cleaning

The raw datasets from the OCDE and the WHO (samples shown in Fig. 1 and Fig. 2) have 28473 rows and 15 columns, with information on each city from the years 2000 – 2016 and several variables measured.



	METRO_ID	Metropolitan areas	VAR	Variables	TIME	Year	Unit Code	Unit	PowerCode Code	PowerCode	Reference Period Code	Reference Period	Value	Flag Codes	Flags
47000	FR018	Reims	GDP_PC_REAL_PPP	GDP per capita (USD, constant prices, constant...	2012	2012	USD	US Dollar	0	Units	NaN	NaN	36207.0	NaN	NaN
47001	FR018	Reims	GDP_PC_REAL_PPP	GDP per capita (USD, constant prices, constant...	2013	2013	USD	US Dollar	0	Units	NaN	NaN	35963.0	NaN	NaN
47002	FR018	Reims	GDP_PC_REAL_PPP	GDP per capita (USD, constant prices, constant...	2014	2014	USD	US Dollar	0	Units	NaN	NaN	32739.0	NaN	NaN
47003	FR018	Reims	GDP_PC_REAL_PPP	GDP per capita (USD, constant prices, constant...	2015	2015	USD	US Dollar	0	Units	NaN	NaN	33762.0	NaN	NaN
47004	COL01	Bogota D.C.	GDP_PC_REAL_PPP	GDP per capita (USD, constant prices, constant...	2015	2015	USD	US Dollar	0	Units	NaN	NaN	22189.0	NaN	NaN

Fig. 1 First Five rows of the raw GDP dataset



	METRO_ID	Metropolitan areas	VAR	Variables	TIME	Year	Unit Code	Unit	PowerCode Code	PowerCode	Reference Period Code	Reference Period	Value	Flag Codes	Flags
0	USA116	Allen	T_T	Population, All ages. Administrative data	2000	2000	PER	Persons	0	Units	NaN	NaN	363420.0	NaN	NaN
1	USA116	Allen	T_T	Population, All ages. Administrative data	2001	2001	PER	Persons	0	Units	NaN	NaN	365954.0	NaN	NaN
2	USA116	Allen	T_T	Population, All ages. Administrative data	2002	2002	PER	Persons	0	Units	NaN	NaN	368402.0	NaN	NaN
3	USA116	Allen	T_T	Population, All ages. Administrative data	2003	2003	PER	Persons	0	Units	NaN	NaN	371190.0	NaN	NaN
4	USA116	Allen	T_T	Population, All ages. Administrative data	2004	2004	PER	Persons	0	Units	NaN	NaN	372757.0	NaN	NaN

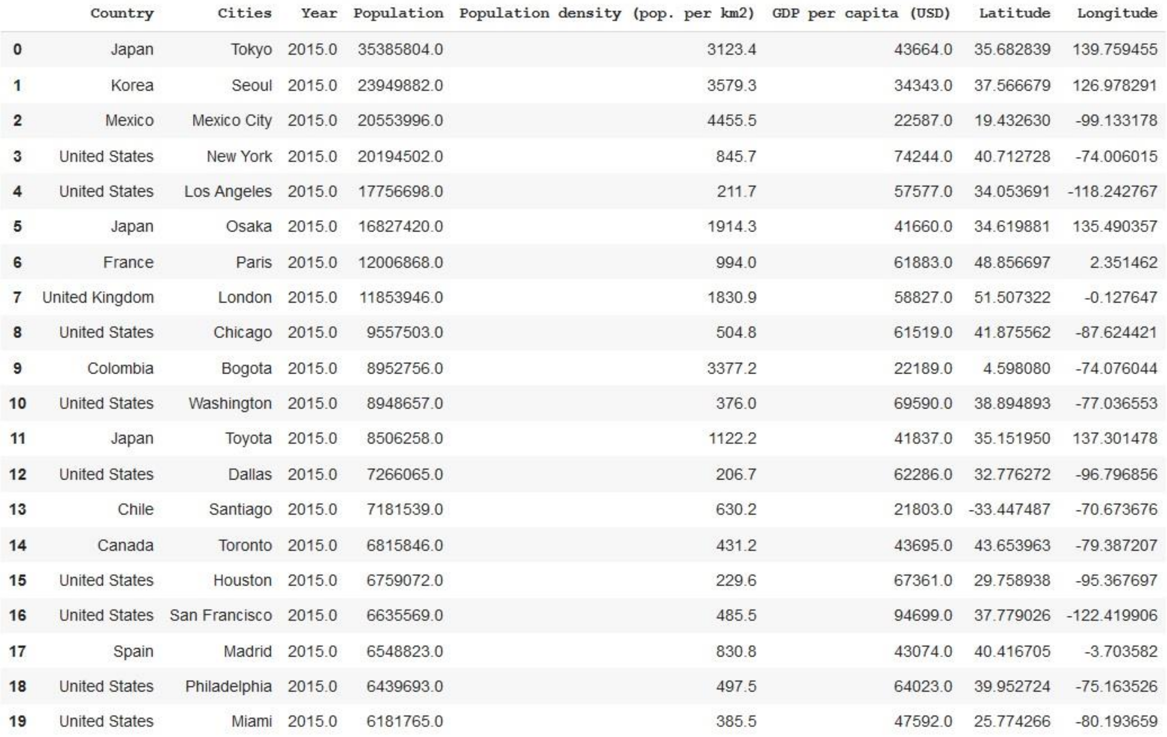
Fig. 2 Last five rows of the Population dataset

The variables we are going to use for the analysis are:

- Cities population dataset: Population, Population density
- Cities GDP: GDP per capita
- Country Hospitals: Hospitals (total number of hospitals)
- Country Healthcare access: Percentage of the population to access to healthcare services

The dataset has information from several years, so we are going to use data from 2016 because it is the most up to date information in all datasets.

We need to filter de dataset, first, by variable and year, then sorting the dataframe by population. For some cities, for example, Tokyo, there is no information regarding 2016; this is also true for various cities. Then we are going to restring the information to data from 2015.

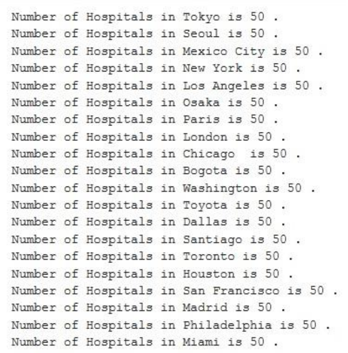


	Country	Cities	Year	Population	Population density (pop. per km2)	GDP per capita (USD)	Latitude	Longitude
0	Japan	Tokyo	2015.0	35385804.0	3123.4	43664.0	35.682839	139.759455
1	Korea	Seoul	2015.0	23949882.0	3579.3	34343.0	37.566679	126.978291
2	Mexico	Mexico City	2015.0	20553996.0	4455.5	22587.0	19.432630	-99.133178
3	United States	New York	2015.0	20194502.0	845.7	74244.0	40.712728	-74.006015
4	United States	Los Angeles	2015.0	17756698.0	211.7	57577.0	34.053691	-118.242767
5	Japan	Osaka	2015.0	16827420.0	1914.3	41660.0	34.619881	135.490357
6	France	Paris	2015.0	12006868.0	994.0	61883.0	48.856697	2.351462
7	United Kingdom	London	2015.0	11853946.0	1830.9	58827.0	51.507322	-0.127647
8	United States	Chicago	2015.0	9557503.0	504.8	61519.0	41.875562	-87.624421
9	Colombia	Bogota	2015.0	8952756.0	3377.2	22189.0	4.598080	-74.076044
10	United States	Washington	2015.0	8948657.0	376.0	69590.0	38.894893	-77.036553
11	Japan	Toyota	2015.0	8506258.0	1122.2	41837.0	35.151950	137.301478
12	United States	Dallas	2015.0	7266065.0	206.7	62286.0	32.776272	-96.796856
13	Chile	Santiago	2015.0	7181539.0	630.2	21803.0	-33.447487	-70.673676
14	Canada	Toronto	2015.0	6815846.0	431.2	43695.0	43.653963	-79.387207
15	United States	Houston	2015.0	6759072.0	229.6	67361.0	29.758938	-95.367697
16	United States	San Francisco	2015.0	6635569.0	485.5	94699.0	37.779026	-122.419906
17	Spain	Madrid	2015.0	6548823.0	830.8	43074.0	40.416705	-3.703582
18	United States	Philadelphia	2015.0	6439693.0	497.5	64023.0	39.952724	-75.163526
19	United States	Miami	2015.0	6181765.0	385.5	47592.0	25.774266	-80.193659

Fig. 3 Cities dataset filtered

Now we have the 20 most populated cities in our datasets. One feature missing is the number of hospitals in each city. Using the **Foursquare API**, we are going to retrieve this feature.

The maximum radius of search in the Foursquare API is 100km, but it appears that it only gives us 50 venues per search. Even if we change the radius up to 20Km, the number of venues in the response in the same, and for less of 20km, we received less than 50 venues. So, we need to get this information from other sources. We are going to use the data from [3].

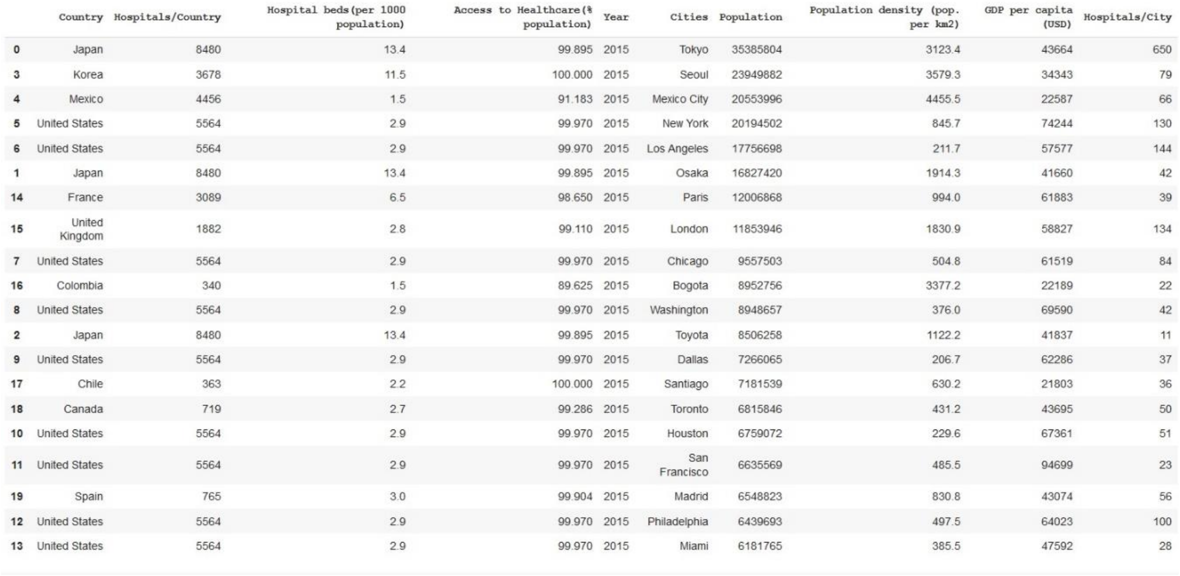


```
Number of Hospitals in Tokyo is 50 .
Number of Hospitals in Seoul is 50 .
Number of Hospitals in Mexico City is 50 .
Number of Hospitals in New York is 50 .
Number of Hospitals in Los Angeles is 50 .
Number of Hospitals in Osaka is 50 .
Number of Hospitals in Paris is 50 .
Number of Hospitals in London is 50 .
Number of Hospitals in Chicago  is 50 .
Number of Hospitals in Bogota is 50 .
Number of Hospitals in Washington is 50 .
Number of Hospitals in Toyota is 50 .
Number of Hospitals in Dallas is 50 .
Number of Hospitals in Santiago is 50 .
Number of Hospitals in Toronto is 50 .
Number of Hospitals in Houston is 50 .
Number of Hospitals in San Francisco is 50 .
Number of Hospitals in Madrid is 50 .
Number of Hospitals in Philadelphia is 50 .
Number of Hospitals in Miami is 50 .
```

Fig. 4 Number of venues given by the Foursquare API

Now from the WHO datasets, we need to retrieve the number of hospitals per country, the number of beds density, and the percentage of the population that have access to healthcare services.

The previous datasets are incomplete (see the notebook), do not have information from the year 2015 nor of the countries needed. The missing information we are going to retrieved from [4].
The Figure 5 show the final dataset that contain information of each country and each city, and Figure 6 show the localization of each City.



	Country	Hospitals/Country	Hospital beds(per 1000 population)	Access to Healthcare(% population)	Year	Cities	Population	Population density (pop. per km2)	GDP per capita (USD)	Hospitals/City
0	Japan	8480	13.4	99.895	2015	Tokyo	35385804	3123.4	43664	650
3	Korea	3678	11.5	100.000	2015	Seoul	23949882	3579.3	34343	79
4	Mexico	4456	1.5	91.183	2015	Mexico City	20553996	4455.5	22587	66
5	United States	5564	2.9	99.970	2015	New York	20194502	845.7	74244	130
6	United States	5564	2.9	99.970	2015	Los Angeles	17756698	211.7	57577	144
1	Japan	8480	13.4	99.895	2015	Osaka	16827420	1914.3	41660	42
14	France	3089	6.5	98.650	2015	Paris	12006868	994.0	61883	39
15	United Kingdom	1882	2.8	99.110	2015	London	11853946	1830.9	58827	134
7	United States	5564	2.9	99.970	2015	Chicago	9557503	504.8	61519	84
16	Colombia	340	1.5	89.625	2015	Bogota	8952756	3377.2	22189	22
8	United States	5564	2.9	99.970	2015	Washington	8948657	376.0	69590	42
2	Japan	8480	13.4	99.895	2015	Toyota	8506258	1122.2	41837	11
9	United States	5564	2.9	99.970	2015	Dallas	7266065	206.7	62286	37
17	Chile	363	2.2	100.000	2015	Santiago	7181539	630.2	21803	36
18	Canada	719	2.7	99.286	2015	Toronto	6815846	431.2	43695	50
10	United States	5564	2.9	99.970	2015	Houston	6759072	229.6	67361	51
11	United States	5564	2.9	99.970	2015	San Francisco	6635569	485.5	94699	23
19	Spain	765	3.0	99.904	2015	Madrid	6548823	830.8	43074	56
12	United States	5564	2.9	99.970	2015	Philadelphia	6439693	497.5	64023	100
13	United States	5564	2.9	99.970	2015	Miami	6181765	385.5	47592	28

Fig. 5 Final Dataset

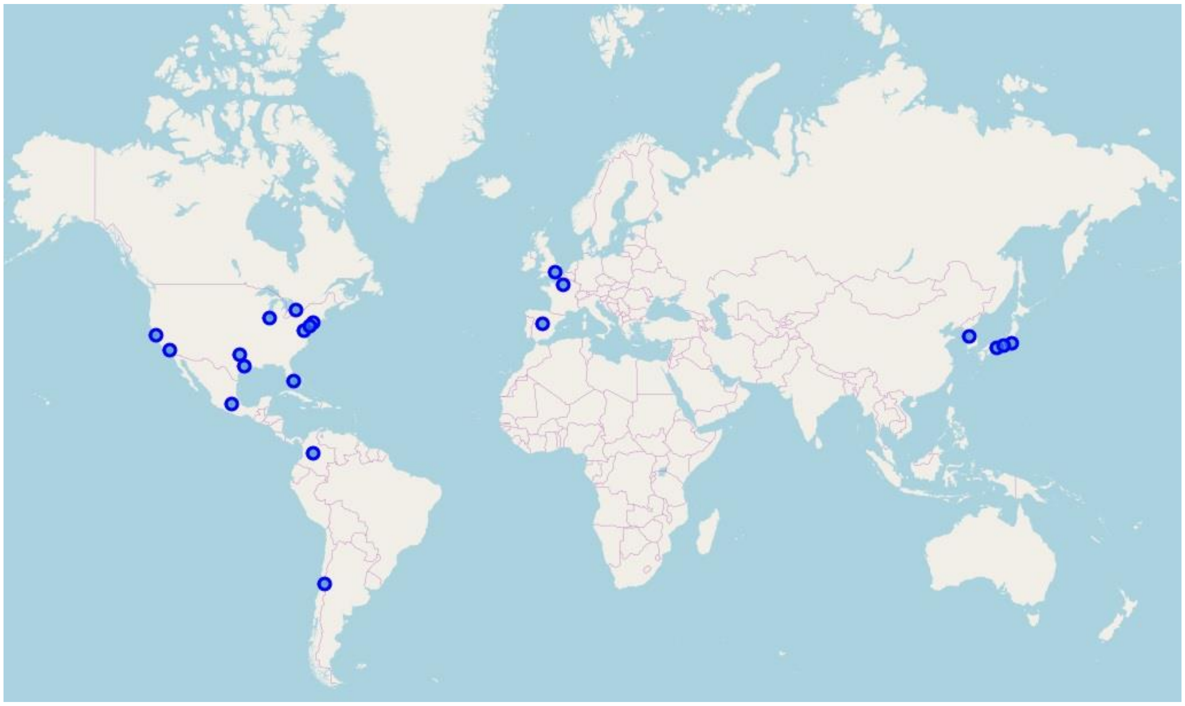


Fig. 6 Cities

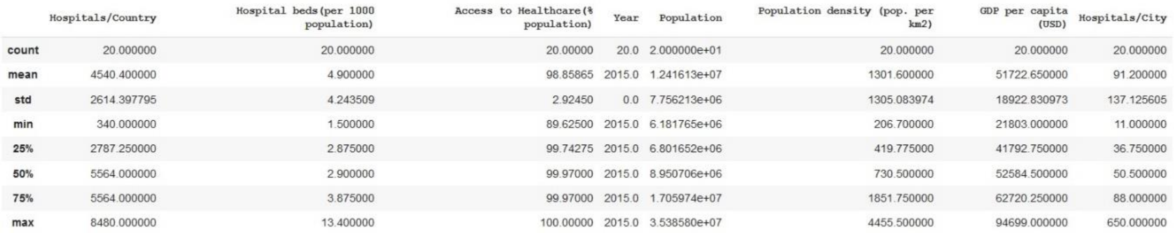
The final dataset is finished, the features are more general regarding the information of each country and city, specific data of the healthcare system of each city is more difficult to obtain, you have to search this data in local institutions.

We are going to treated analysis as a first approximation for solving the problem; determine which five cities are the best to build a hospital according to the selected features. Once we selected those cities, we can pass to the next stage analyzing more specific and local data of the healthcare system of each city and iterate the methodology from there.

3. EXPLORATORY ANALYSIS

Figure 7 shows the basic statistical details of the data frame, the standard deviation is wide, except for the "Access to Healthcare" attribute. Figure 8 and Figure 9 show the cities and countries features, respectively.

This first analysis does not give us any preliminary insight. We can say, as a first approximation, that in order to decide on where to build a hospital, first, we need to know in which cities are deficits of hospitals and if the population that is able to pay for private healthcare services.



	Hospitals/Country	Hospital beds(per 1000 population)	Access to Healthcare(% population)	Year	Population	Population density (pop. per km2)	GDP per capita (USD)	Hospitals/City
count	20.000000	20.000000	20.00000	20.0	2.000000e+01	20.000000	20.000000	20.000000
mean	4540.400000	4.900000	98.85865	2015.0	1.241613e+07	1301.600000	51722.650000	91.200000
std	2614.397795	4.243509	2.92450	0.0	7.756213e+06	1305.083974	18922.830973	137.125605
min	340.000000	1.500000	89.62500	2015.0	6.181765e+06	206.700000	21803.000000	11.000000
25%	2787.250000	2.875000	99.74275	2015.0	6.801652e+06	419.775000	41792.750000	36.750000
50%	5564.000000	2.900000	99.97000	2015.0	8.950706e+06	730.500000	52584.500000	50.500000
75%	5564.000000	3.875000	99.97000	2015.0	1.705974e+07	1851.750000	62720.250000	88.000000
max	8480.000000	13.400000	100.00000	2015.0	3.538580e+07	4455.500000	94699.000000	650.000000

Fig. 7 Basic statistical details

It appears that these features tell us more about the quality of life in a city, but do not tell us much about the local healthcare system.

We are going to do a segmentation analysis using K-Means clustering to see the similarity between cities.

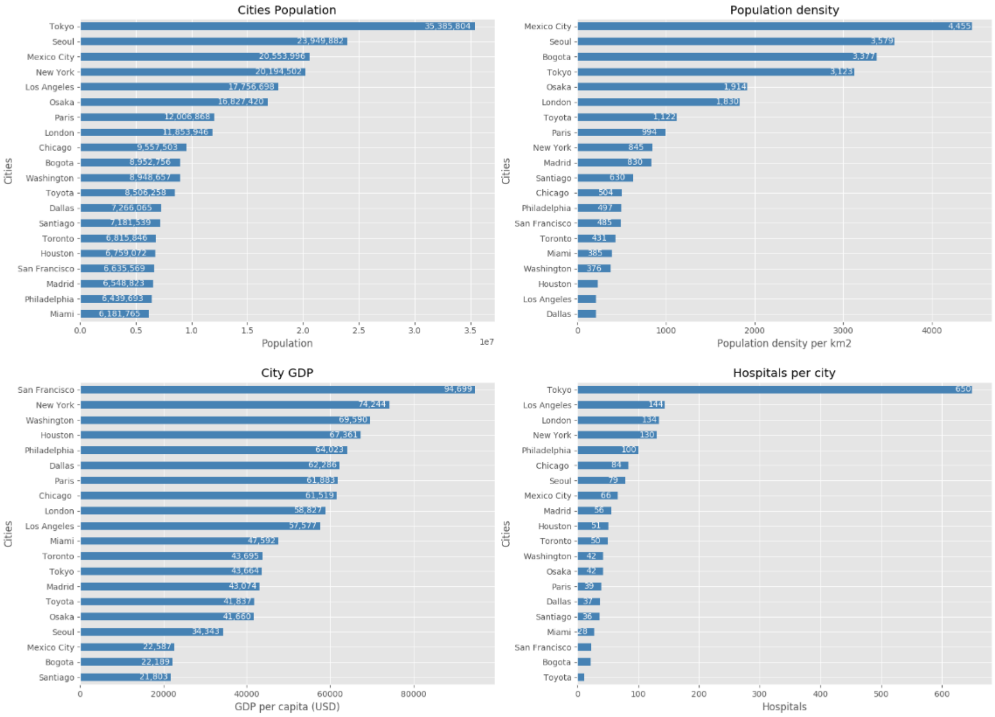


Fig. 8 Cities features

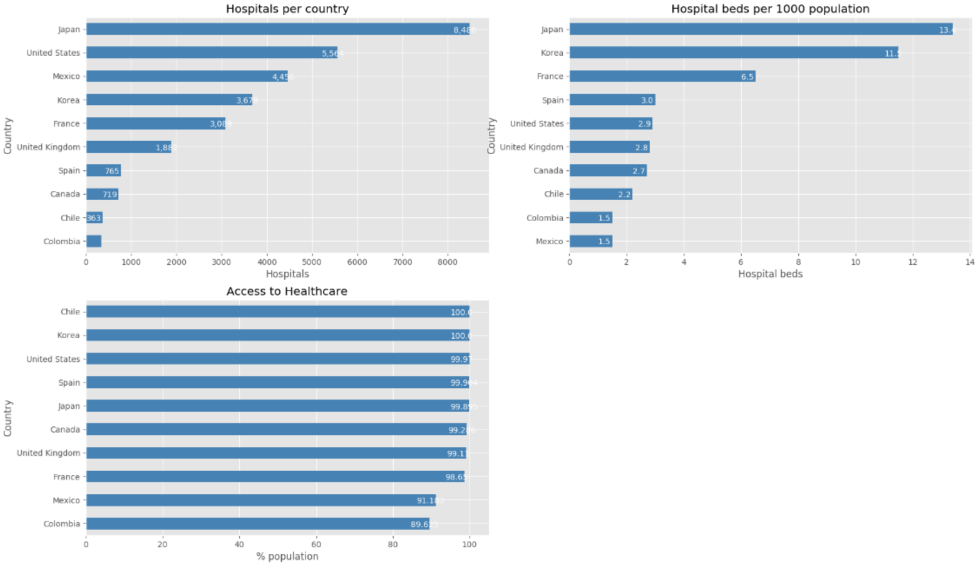
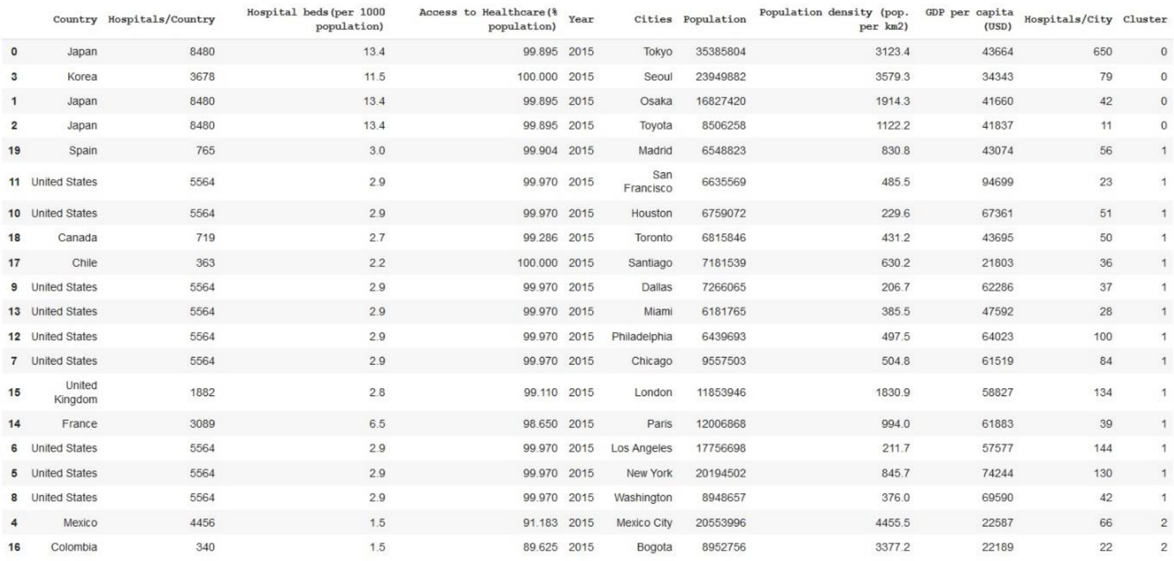


Fig. 9 Country features

4. SEGMENTATION

We are going to use K-means from the scikit-learn library using three clusters (see the notebook for more details). Figure 10 shows the cluster in the data frame and Figure 10 shows the localization of the clusters.



	Country	Hospitals/Country	Hospital beds (per 1000 population)	Access to Healthcare (% population)	Year	Cities	Population	Population density (pop. per km2)	GDP per capita (USD)	Hospitals/City	Cluster
0	Japan	8480	13.4	99.895	2015	Tokyo	35385804	3123.4	43664	650	0
3	Korea	3678	11.5	100.000	2015	Seoul	23949882	3579.3	34343	79	0
1	Japan	8480	13.4	99.895	2015	Osaka	16827420	1914.3	41660	42	0
2	Japan	8480	13.4	99.895	2015	Toyota	8506258	1122.2	41837	11	0
19	Spain	765	3.0	99.904	2015	Madrid	6548823	830.8	43074	56	1
11	United States	5564	2.9	99.970	2015	San Francisco	6635569	485.5	94699	23	1
10	United States	5564	2.9	99.970	2015	Houston	6759072	229.6	67361	51	1
18	Canada	719	2.7	99.286	2015	Toronto	6815846	431.2	43695	50	1
17	Chile	363	2.2	100.000	2015	Santiago	7181539	630.2	21803	36	1
9	United States	5564	2.9	99.970	2015	Dallas	7266065	206.7	62286	37	1
13	United States	5564	2.9	99.970	2015	Miami	6181765	385.5	47592	28	1
12	United States	5564	2.9	99.970	2015	Philadelphia	6439693	497.5	64023	100	1
7	United States	5564	2.9	99.970	2015	Chicago	9557503	504.8	61519	84	1
15	United Kingdom	1882	2.8	99.110	2015	London	11853946	1830.9	58827	134	1
14	France	3089	6.5	98.650	2015	Paris	12006868	994.0	61883	39	1
6	United States	5564	2.9	99.970	2015	Los Angeles	17756698	211.7	57577	144	1
5	United States	5564	2.9	99.970	2015	New York	20194502	845.7	74244	130	1
8	United States	5564	2.9	99.970	2015	Washington	8948657	376.0	69590	42	1
4	Mexico	4456	1.5	91.183	2015	Mexico City	20553996	4455.5	22587	66	2
16	Colombia	340	1.5	89.625	2015	Bogota	8952756	3377.2	22189	22	2

Fig. 10 Clusters

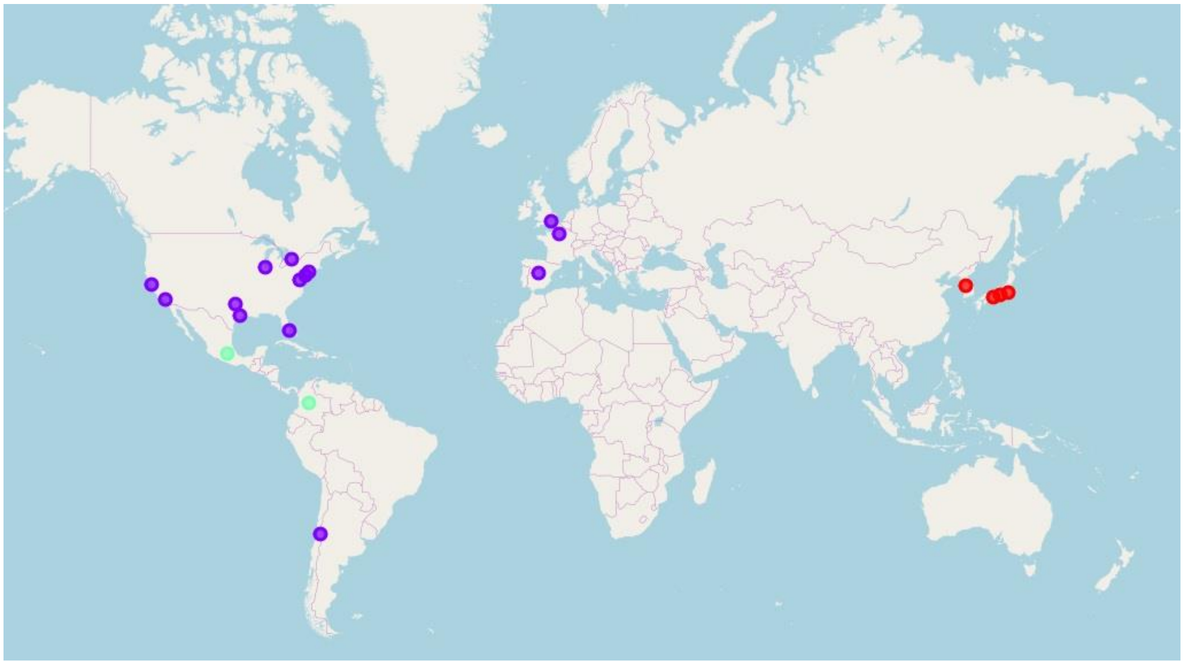


Fig. 11 Localization of each cluster

5. DISCUSION

After this analysis, I think there is no surprise. The features represent the quality of life in each city; we can observe a clear segmentation into regions. There is one cluster that represents the so-called "west" that includes west Europe and North America, another cluster that represents Latin America except for Santiago de Chile, and finally, a third cluster that represents the cities in East Asia.

6. CONCLUSION

In conclusion, we can say we need more data; the data used is not enough to solve the problem. What we can say according to the results obtained is that it is a regional problem, so we need to study the features of each region and see what benefits and disadvantages have each one. Then, do a local analysis of each city to explore the local healthcare system.

References

[1] https://stats.oecd.org/Index.aspx?DataSetCode=CITIES

[2] https://www.who.int/data/gho

[3] http://www.city-data.com/world-cities/index.html

[4] https://www.indexmundi.com/