Comparison of Europe's Top Football Stadiums

INDERPREET SINGH TOOR

28th August, 2019

1. Introduction

1.1 Background

Football (soccer) is the number 1 sport in the world in terms of fan base and viewership. Europe is the hotbed of football since its inception. In Europe's top 5 leagues a season usually lasts for about 9 months i.e. from mid-August to mid-May. During this period thousands of fans travel to watch their teams play in the stadiums. Where there are people there are chances of things happening, be it business related or a tragedy waiting to happen. Football fans might want to explore the area around the stadium if there are venues of entertainment or food/coffee shops nearby. Also there is a possibility of a medical emergency. So people might need the information if there are adequate number of hospitals nearby in case of medical emergency.

1.2 Problem

Fans travelling for games away from home might want to know if there are good places of their interest nearby the stadium they are going to visit and might seek a comparison between different stadiums in respect to the nearby venues such as bars, restaurants or hotels. They come to a decision that which stadium is best according to their taste.

Also with a large gathering of crowd on regular basis there is always a chance of medical emergency on a personal or a mass scale for e.g. a terrorist attack or riots leading to injuries to thousands. So, Stadium authorities, Public Administration and general public may need crucial information of Hospitals nearby so that planning can be done in advance if there is any medical emergency and necessary measures can be taken to avoid major mishappenings.

This project will aim to compare the stadium surroundings in Europe's Top 5 football leagues for the people to choose and make planning ahead of time.

2. Data

2.1 Data Requirements

We need detailed information about the stadiums such as home club, geo-coordinates, capacity, city where the stadium is situated. Apart from this we will require the information of entertainment venues, fast food joints, restaurants and hotels nearby the stadium. We will need the names, geo-coordinates and categories of these places.

We will also require the geo-coordinates and name of hospitals nearby.

2.2 Data Sources

The stadium data we need can be accessed through Wikipedia pages of respective leagues for the season 2019-20. I am taking the latest data of the teams playing in Europe's top 5 leagues for up to date information. You can find the links to each of the league's Wikipedia page below:

- English Premier League
- La Liga
- Bundesliga
- Serie A
- Ligue 1

There are tables on these pages containing the required information. I scrapped the data using pandas library in Python.

For the geo-coordinates of stadiums I used the Geocoder library which takes an address of a location as an input string and throw out its latitude and longitude values.

In order to get information about the nearby venues of entertainment, bars, hotels and restaurants I used foursquare API and its Explore URL type.

In order to get information about the nearby hospitals I used foursquare API and its Search URL type. The radius was set to 2.5 kilometres with centre as the coordinates of the stadium.

2.3 Data Cleaning

Data was stored in individual league table as there were some dissimilarities in terms of formatting, unnecessary data, faulty information and language conversion.

There were problems with how the scrapped data presented itself in the dataframe. I had to make all the column names uniform and in exact same order in all the league tables. For e.g. one table had teams under 'Team'

column and another table had teams under ‘Club’ column. So I uniformly put all teams under ‘Club’ column. Some tables had additional information per column which I had to remove and some tables had additional columns so I dropped those columns. Now some of the information like names of the stadium had old names so I had to manually change those cells with the new names. Some stadium names were giving wrong coordinates when run through Geocoder. I realized that their names were not complete so I had to correct those names as well.

When the above information was corrected I sorted the tables according to the stadium capacity, reset the index and stored them as csv files for later use.

3. Methodology

As this project is divided into following two parts:

a. Comparison of Stadiums based on nearby Restaurants, Hotels, Pubs and Bars.
b. Comparison of Stadiums based on nearby Hospitals.

For Hotels, Restaurants, Pubs, Bars:

Then with the help of Foursquare API I retrieved the data of nearby venues and stored them in the table after taking only the relevant parts of the json file. The radius was of 2 Kilometres. We select only those required venue categories and then one hot encoding was performed on frequency of venue category to make the data ready for K-Means clustering.

For Hospitals:

We use Foursquare API and retrieve data using Search URL feature. The radius was of 2.5 Kilometres. We store the data and group all clubs to get the frequency of hospitals. We use this data to visualize and find out which clubs and countries are the best in terms of hospital frequency.

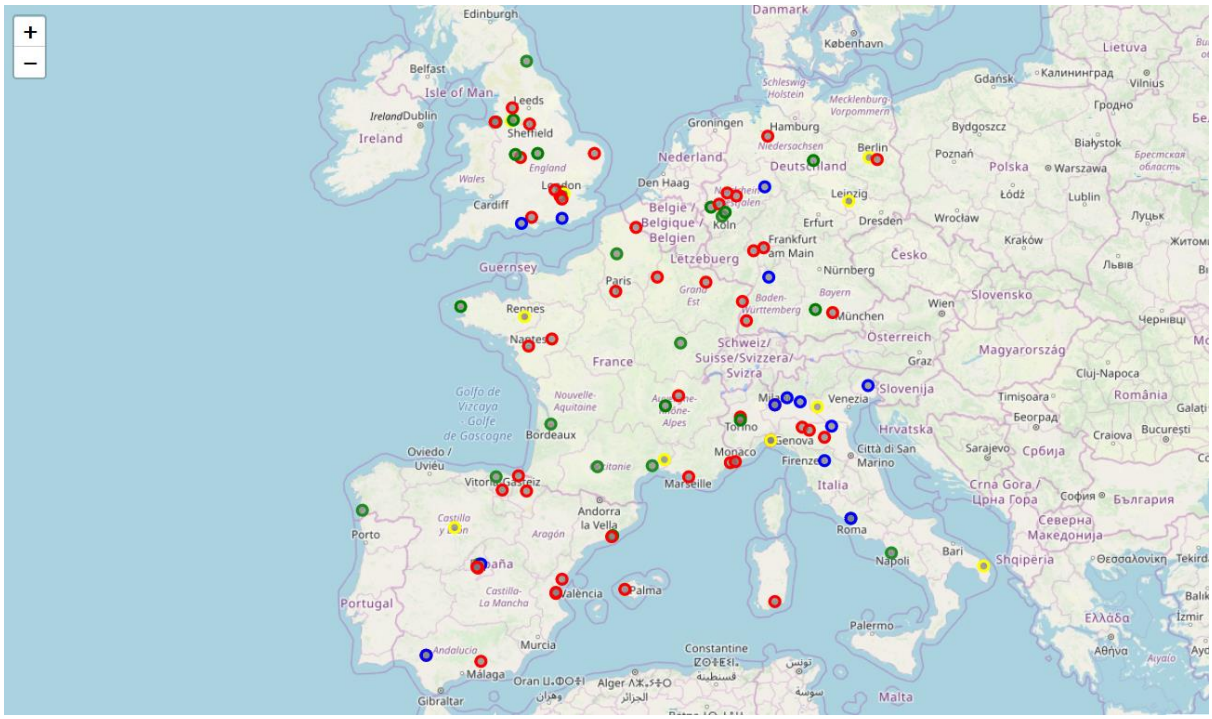
I will discuss about the first part next.

3.1 Exploratory Data Analysis

As our data set consists of data of 97 clubs from Europe’s top 5 football leagues which is huge.

	Club	Stadium Latitude	Stadium Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
5	Manchester United	53.455681	-2.285531	Pi	53.446778	-2.277464	Bar
14	Manchester United	53.455681	-2.285531	The Hillary Step	53.451818	-2.269583	Bar
22	Manchester United	53.455681	-2.285531	Manchester United Red Café	53.464080	-2.292145	Bar
25	Manchester United	53.455681	-2.285531	Proof	53.443948	-2.278135	Bar
26	Manchester United	53.455681	-2.285531	Electrik	53.442469	-2.276222	Bar

After one hot encoding is performed on the selected venues categories data we go for Clustering Analysis and use K-Means Clustering technique with 5 centroids. We add the clusters back to the combined data of all leagues and use Folium maps to visualize our clustered clubs on the map.



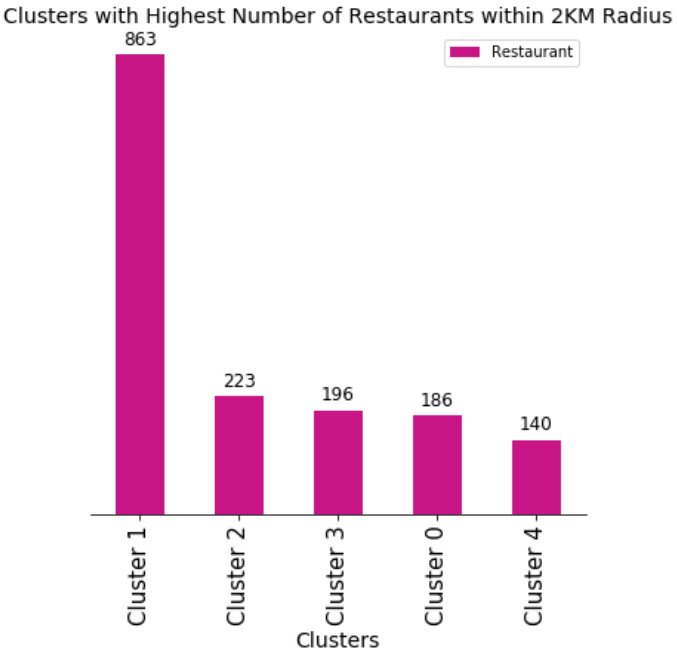
We then combine the clustered data with each venue category we need and make a table like the one shown below which is of Cluster 0.

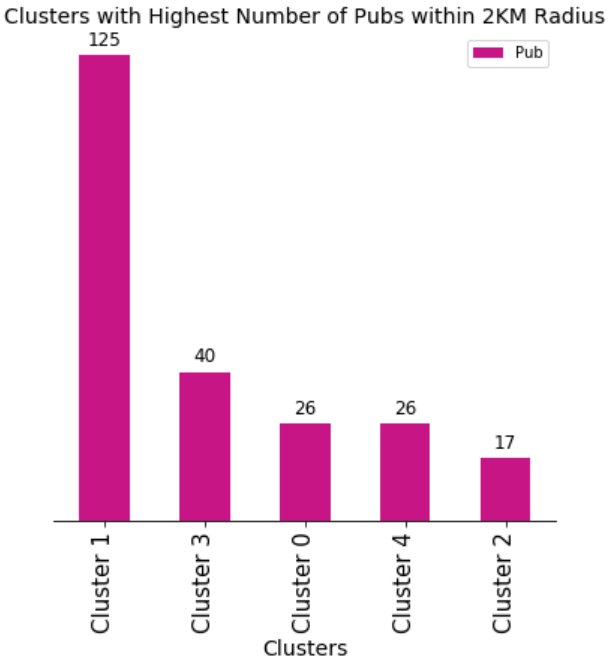
	Cluster Labels	Club	Stadium	Location	Capacity	Latitude	Longitude	Restaurant	Bar	Pub	Hotel
5	0	Liverpool	Anfield	Liverpool	54074	53.430836	-2.960910	5	1	7	1
6	0	Newcastle United	St James' Park	Newcastle upon Tyne	52354	54.975469	-1.621874	22	14	9	1
21	0	Real Madrid	Santiago Bernabéu	Madrid	81044	40.452367	-3.690725	49	2	2	4
25	0	Athletic Bilbao	San Mames Barria	Bilbao	53000	43.264204	-2.949391	29	11	2	6
29	0	Celta Vigo	Abanca-Balaídos	Vigo	29000	42.211780	-8.739715	30	11	0	2
46	0	1. FC Köln	RheinEnergieStadion	Cologne	49698	50.933510	6.876026	18	0	0	2
61	0	Napoli	Stadio San Paolo	Naples	60240	40.827977	14.192889	10	0	4	9
81	0	Bordeaux	Matmut Atlantique	Bordeaux	42115	44.897555	-0.560679	7	1	0	8
85	0	Toulouse	Stadium Municipal	Toulouse	33150	43.577670	1.468602	4	0	2	0
86	0	Montpellier	Stade de la Mosson	Montpellier	32939	43.622022	3.812090	1	0	0	2
93	0	Dijon	Stade Gaston Gérard	Dijon	18376	47.324308	5.068339	10	2	0	6
96	0	Amiens	Stade de la Licorne	Amiens	12097	49.893992	2.263474	1	0	0	0

Now we are ready to explore the cluster data and see what we can find out about the stadiums in regards to venues of our interest.

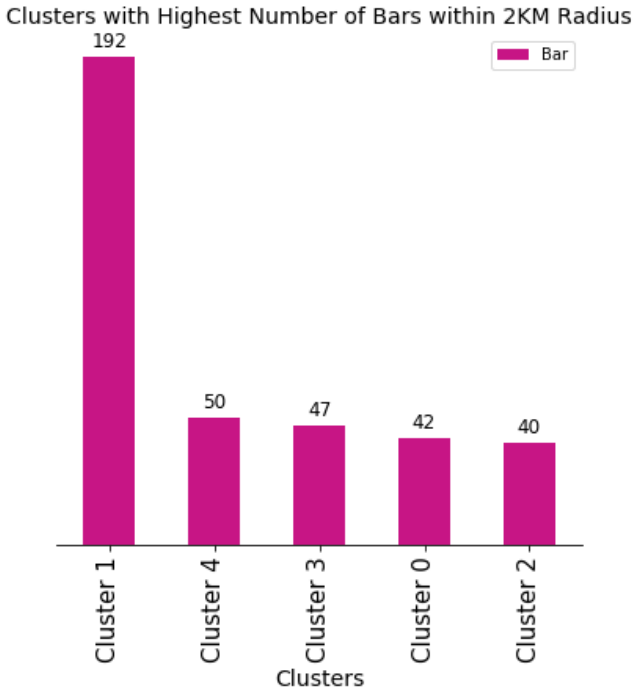
So I now visualized the data with the help of bar graph using Matplotlib library. We draw separate graphs for Restaurants, Bars, Hotels and Pubs.

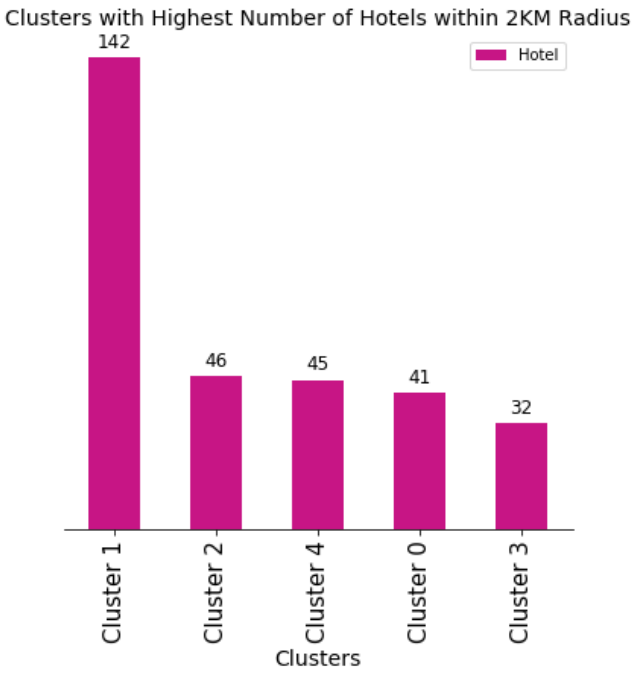
Cluster 1 by far dominates in all categories which indicates that stadiums in





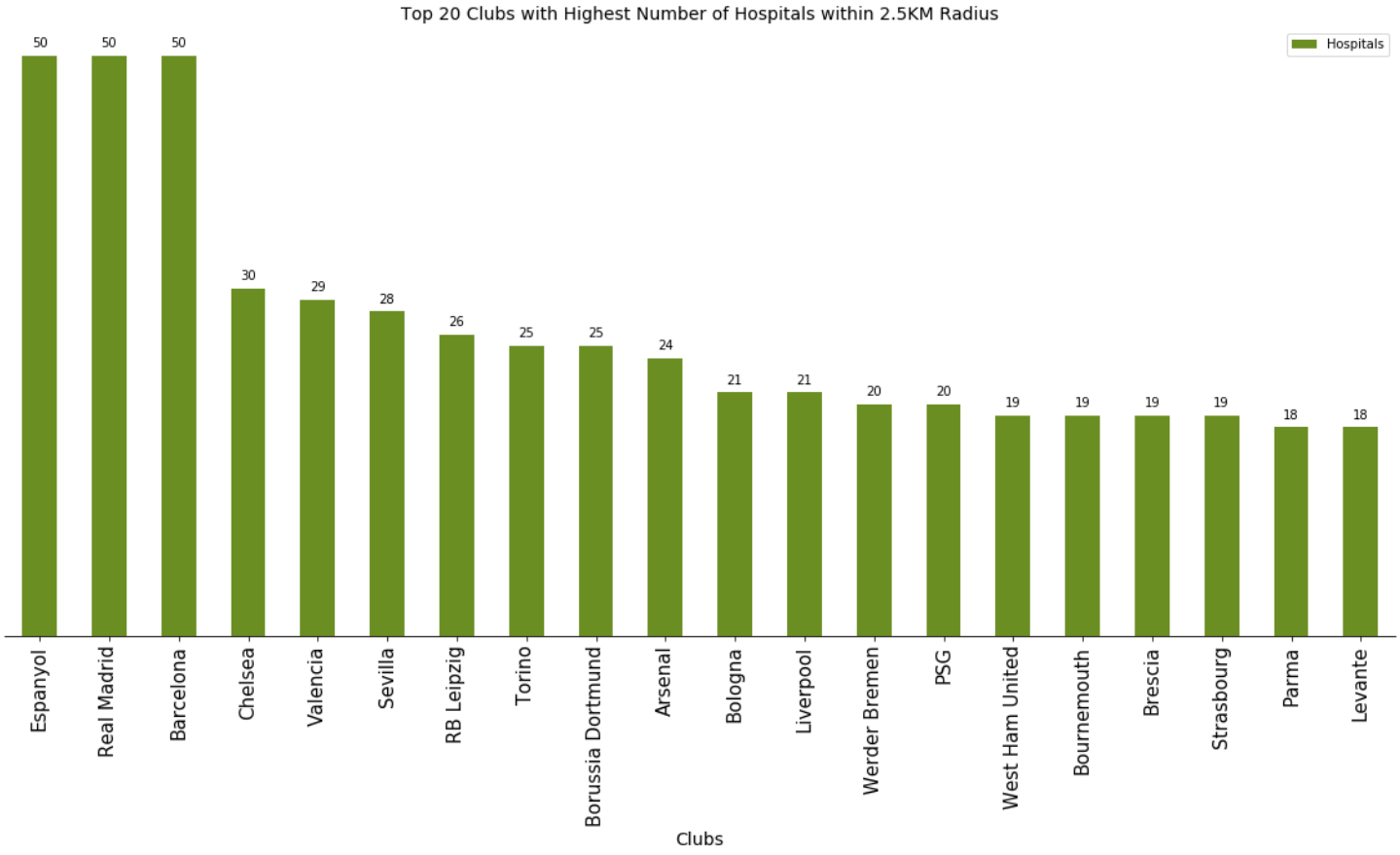
Cluster 1 are the best of the best in terms of services we are looking for. Stadiums in Cluster 3 has good number of Pubs while stadiums in Cluster 4 and Cluster 2 have sufficient numbers of Bars and Hotels respectively.





3.2 Visualization

The second part was Stadium Comparison based on number of nearby Hospitals. We set a parameter using Foursquare API and used its Search URL feature to search for all nearby Hospital. Data was collected individually and merged as dataset of all stadiums from the 5 leagues. Then we grouped all the clubs to calculate the number of Hospitals for each club. Here I went straight to visualization as there weren't enough variables to perform machine learning algorithm to help us with an answer.



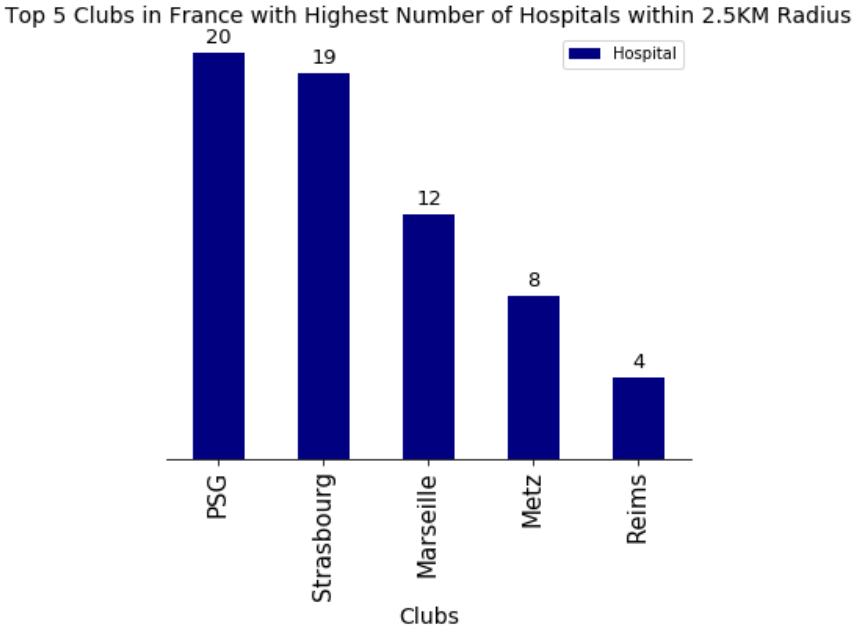
From what we can see from above i.e. Top 20 clubs with highest number of hospital bar chart that Spanish clubs Espanyol, Barcelona and Real Madrid are the top clubs in terms nearby hospitals with 50 hospitals each. Spain is known for its cheap medical services which combined with frequency of hospitals turns out to be the best in this department.

Chelsea is the number one club from England with 29 hospitals followed by another two Spanish clubs Valencia and Sevilla. Parma and Levante with 18 hospitals each round off this list.

Now let's check out each country and their top 5 Clubs in terms of Hospital frequency.

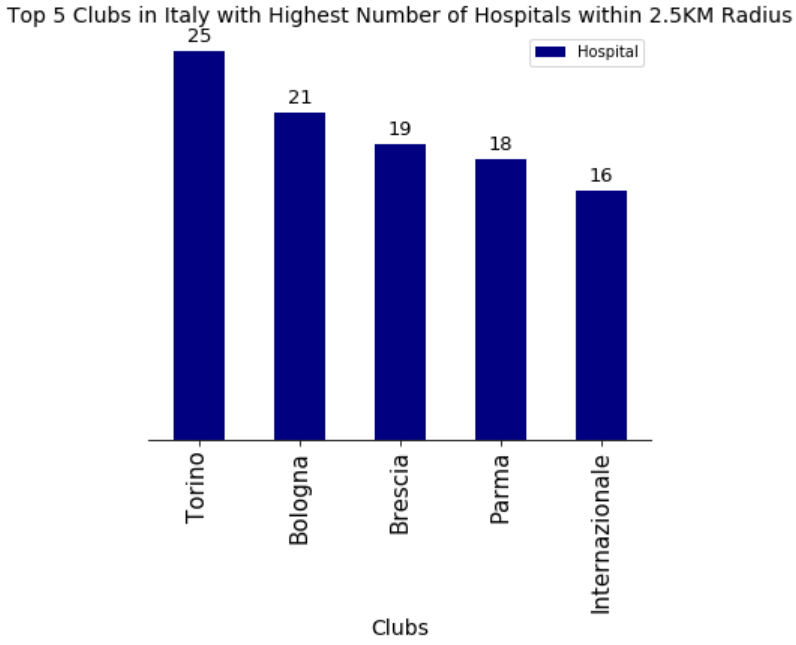
Let's Start with France first.

We notice that PSG being the Capital club rules in French Division in terms of emergency healthcare. Also Strasbourg is close enough to PSG while rest others are lagging far behind.



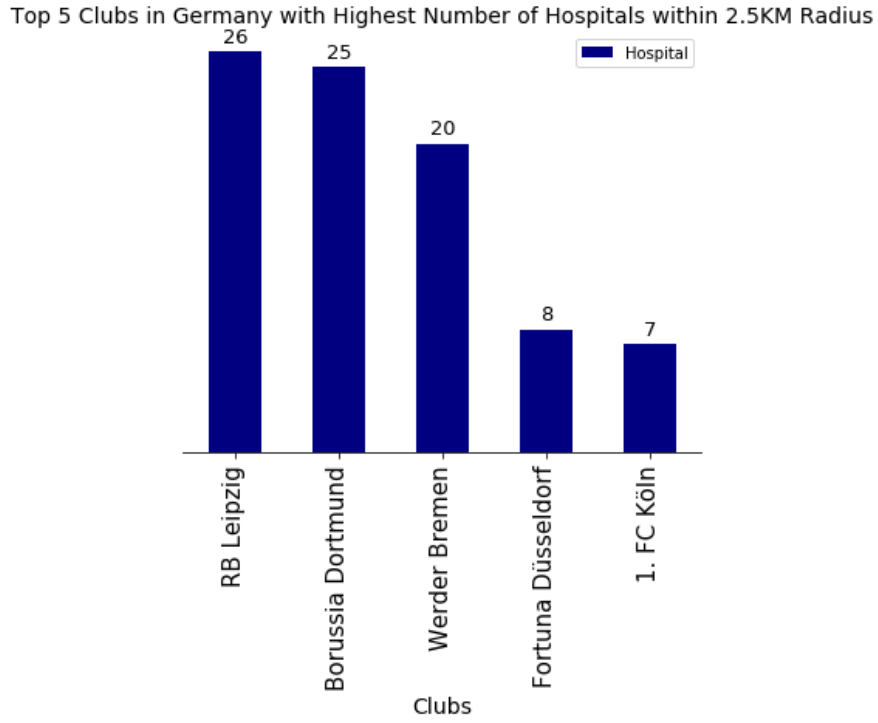
Second we look at Italy.

Torino based in the centre of Turin leads the way with 25 hospitals which rest of the top 5 give a healthy competition stating that Hospitals are a more uniform feature in Italy.

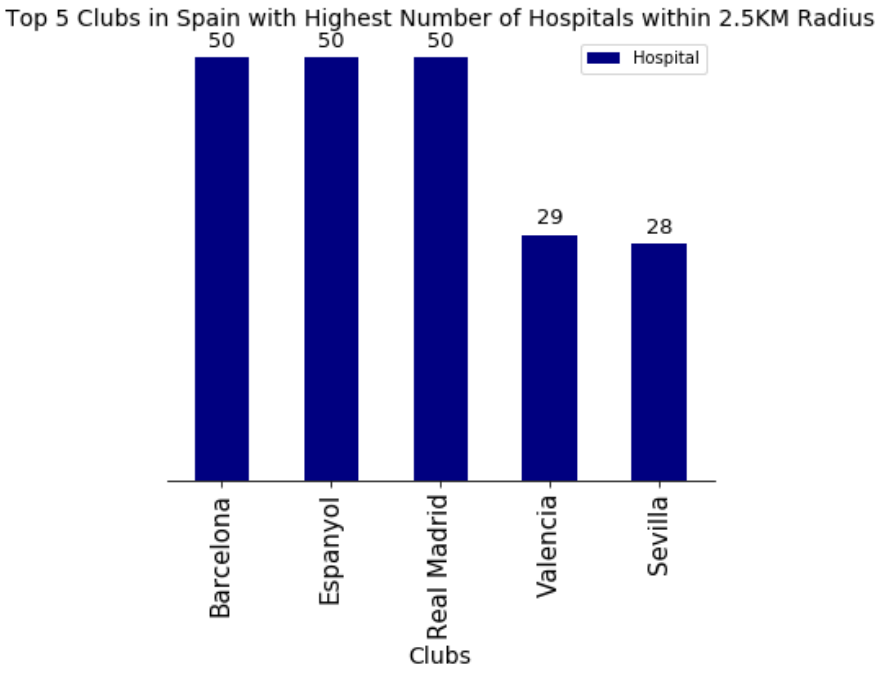


Now let's look at Germany's Data.

RB Leipzig and Borussia Dortmund lead with 26 and 27 hospitals respectively. Bayern Munich is missing from this list as it's stadium is outside the main city.

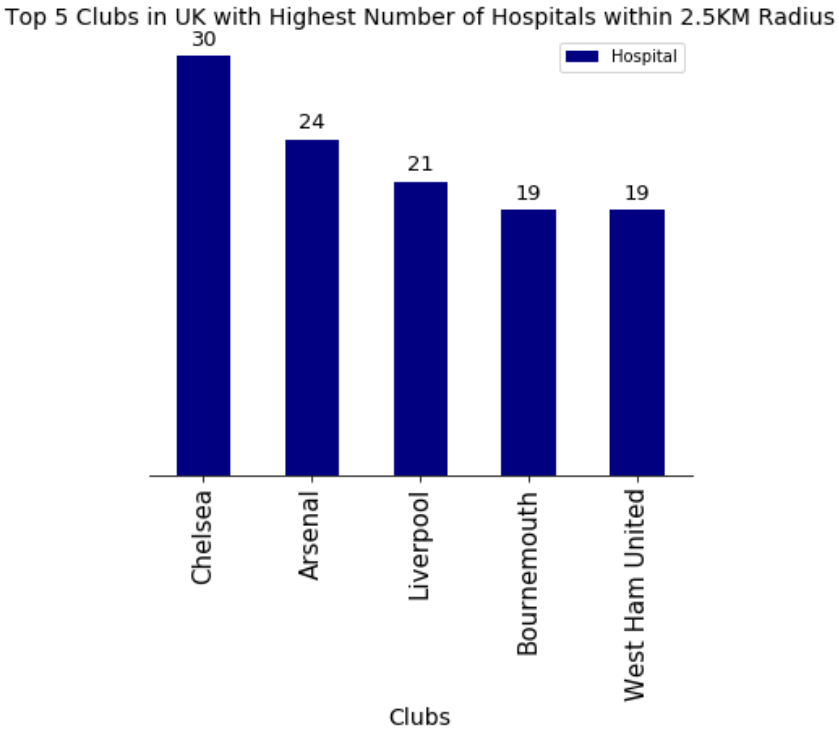


Looking at the Spanish graph we can adjudge them to be the best in class in term of hospital frequency. Even the lesser teams have more hospitals than other league leaders.



Let's check the data from United Kingdom.

EPL stadiums are average in terms of hospital frequency in Europe's top 5 football leagues.



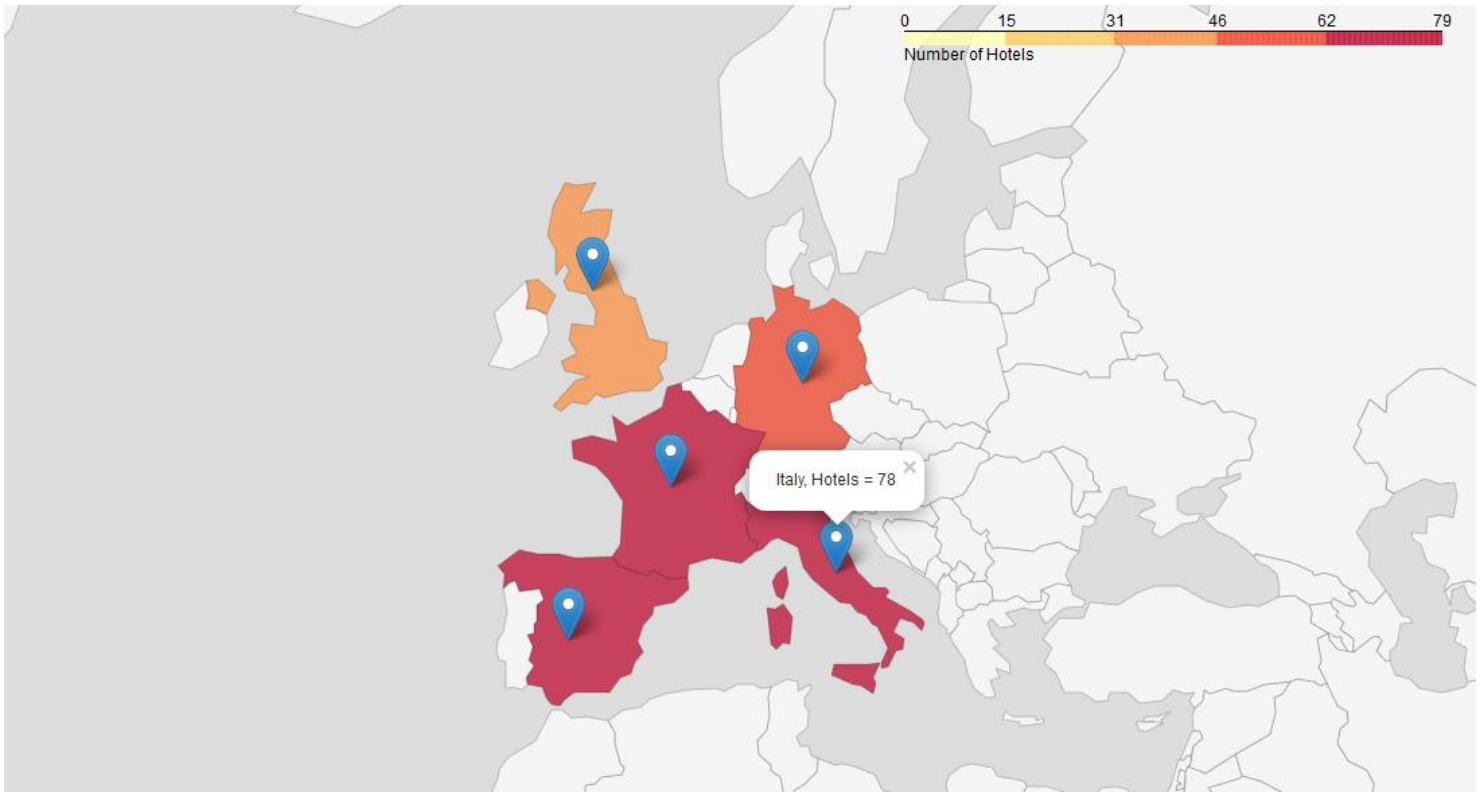
4. Results

We calculate the frequency of each venue of our choice for each league and combine the dataframe by country as shown in table below:

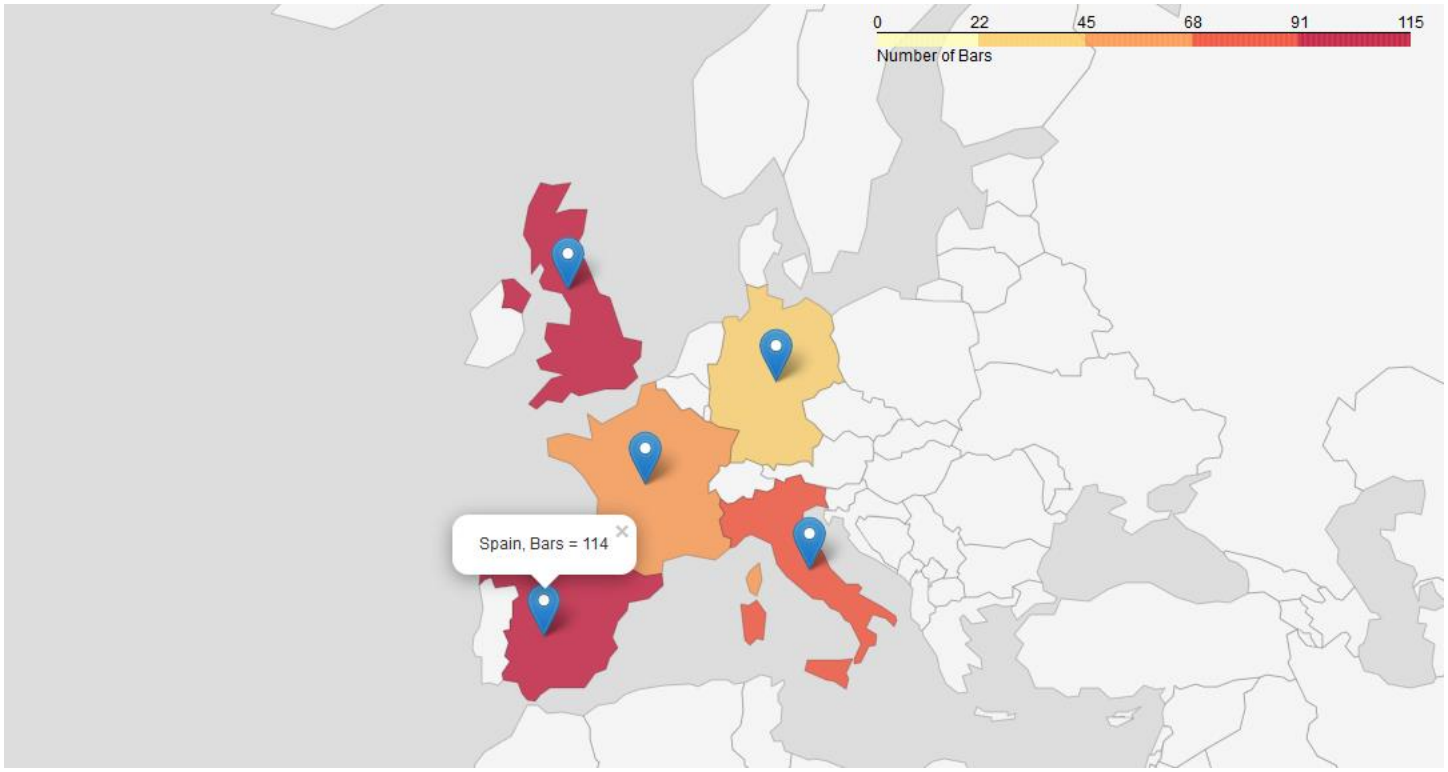
	Country	Restaurant	Bar	Pub	Hotel	Hospital
0	United Kingdom	288	103	155	39	212
1	Spain	504	114	20	66	334
2	Germany	201	32	10	46	118
3	Italy	385	76	33	78	187
4	France	230	46	16	77	79

Now we will use Choropleth Maps to visualize the frequency of Venues by country and judge which country is best in which department.

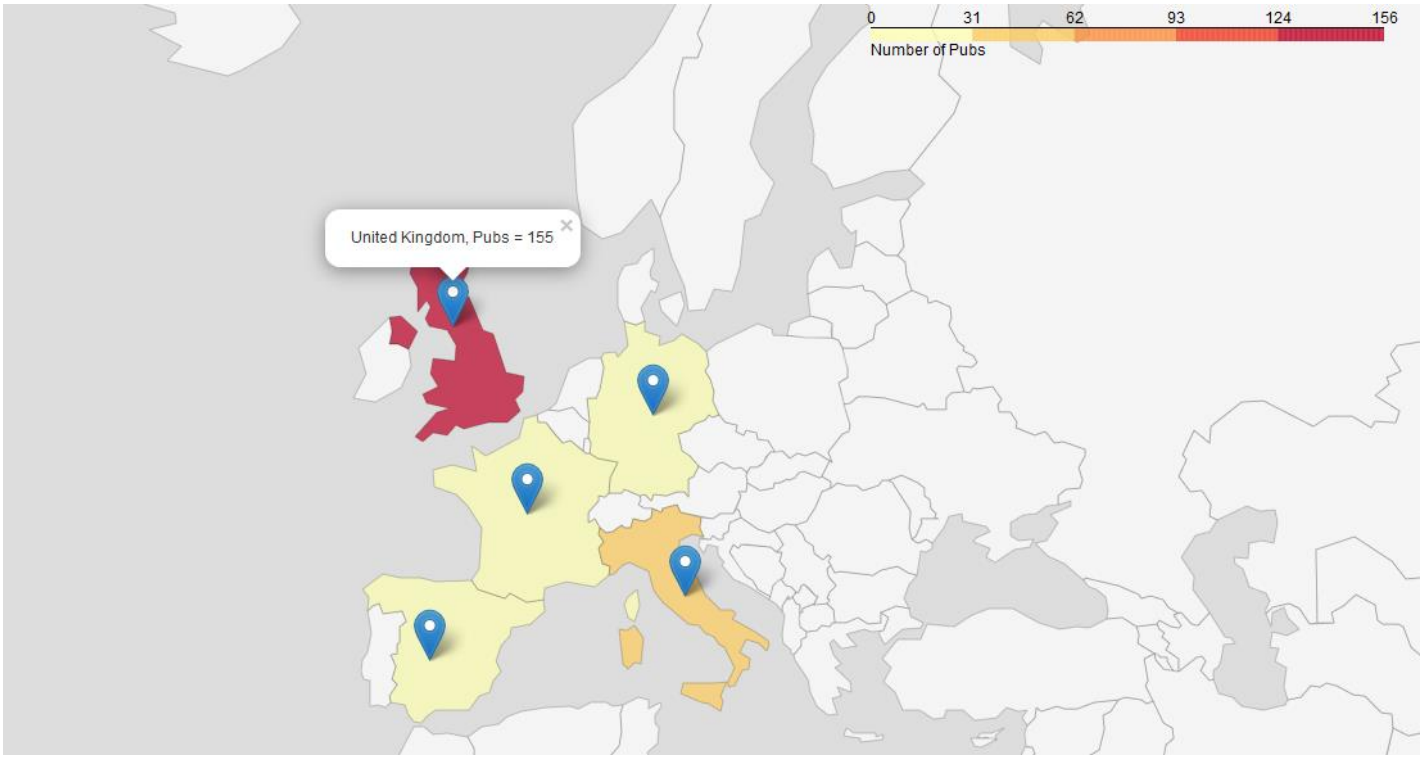
1. **Number of Hotels: Winner - ITALY**



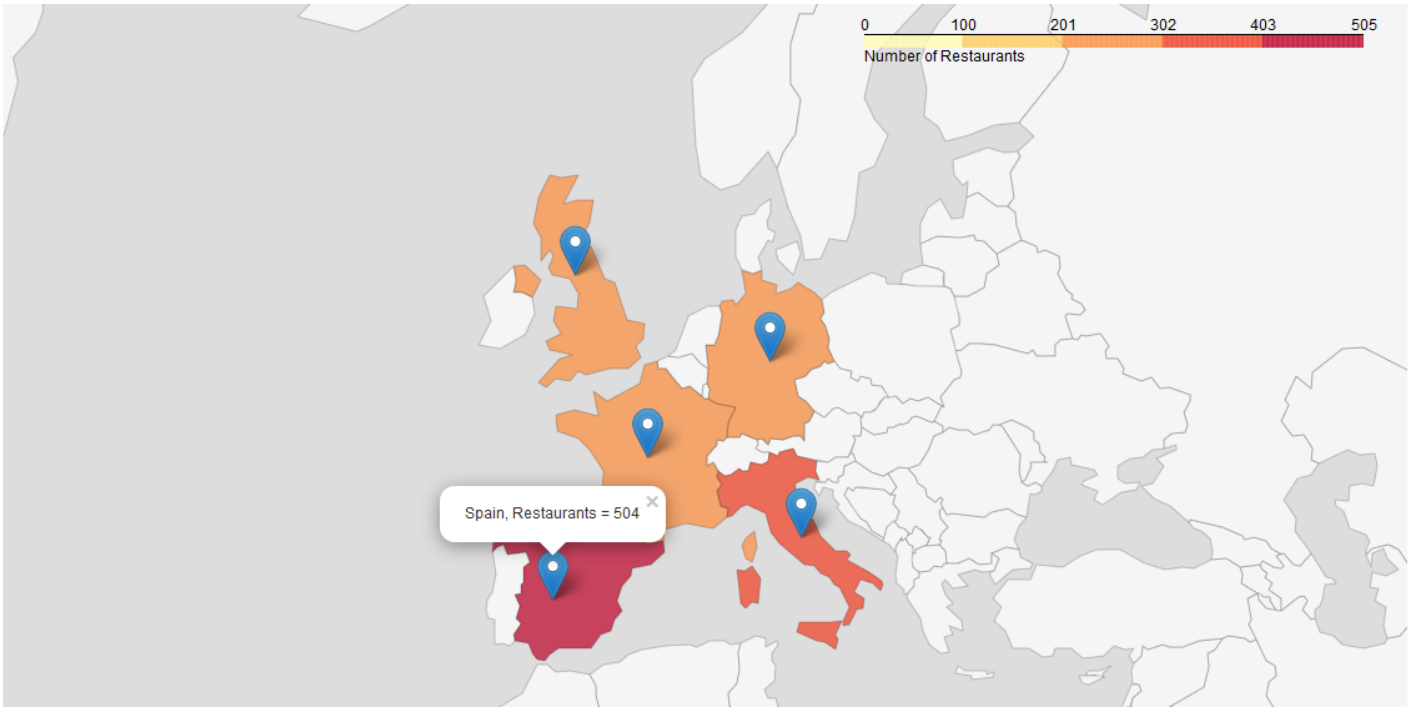
2. Number of Bars: Winner – SPAIN



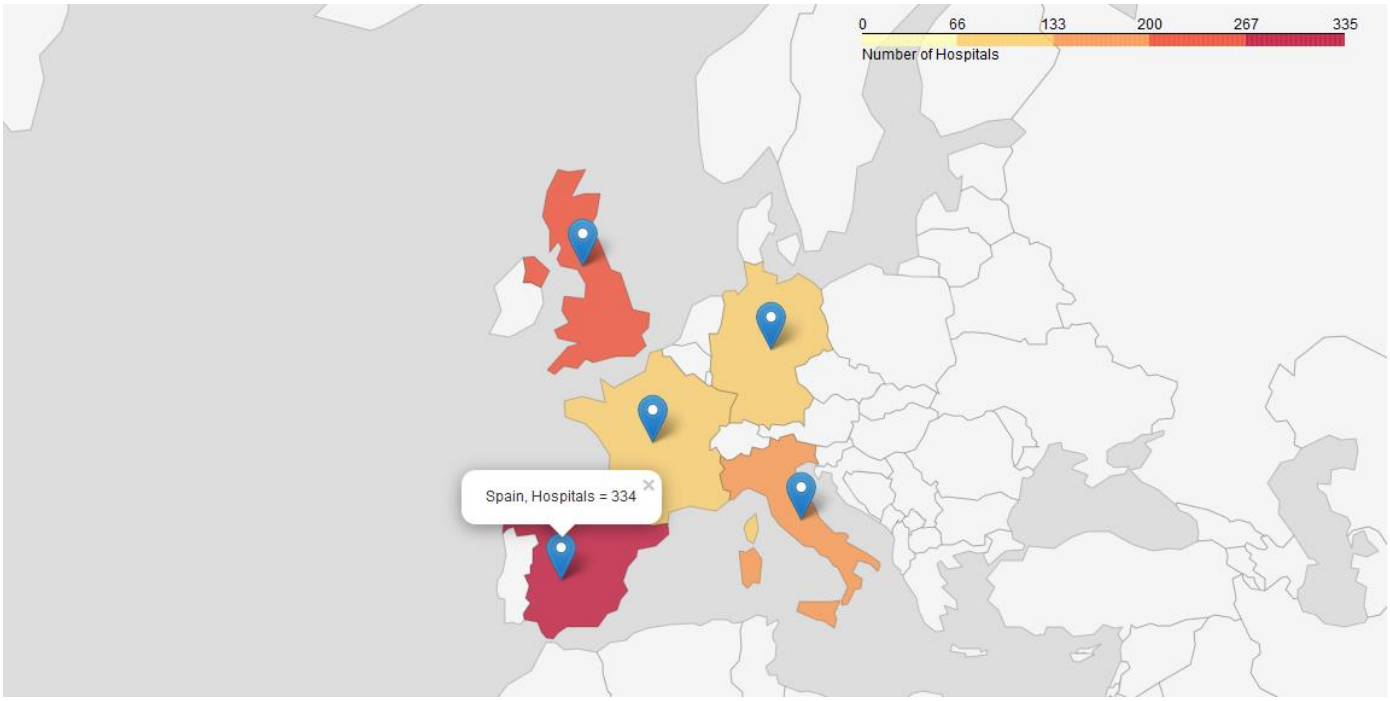
3. Number of Pubs: Winner – ENGLAND



4. Number of Restaurants: Winner – SPAIN



5. Number of Hospitals: Winner - SPAIN



5. Discussion

Let's discuss the results separately for module 1 and module 2.

Module 1: Hospitals

We observe that stadiums Spain have the best frequency of Hospitals in whole of our European Dataset. If you are travelling to Spain for football matches and encounter an accident or there is a medical emergency at the stadium then Spain is in perfect order to handle those scenarios. This also implies that most of the Spanish Stadiums are near city centres which results in greater number of hospitals.

Stadiums in England and Italy, although behind Spain, are still suitable for any medical emergency while attending the games. The graph is evenly distributed which means there is not much difference between in frequency of health care facilities if needed one in times of emergency. We can imply that although Stadiums are a little off from city centres yet there are reasonably close to many hospitals.

While on the other side Stadiums in Germany and France lag far behind those in Spain in number of hospitals. Also the difference between the highest and lowest number of hospital takes a sharp dip in the graph. We can imply that most of the stadiums in these countries are far from city centres and hospitals are therefore, less nearby. There is potential for hospital grown here.

Module 2: Restaurants, Bars, Pubs and Hotels

Italy leads the chart in number of Hotels from our results meaning there is an active tourism industry which is doing well. France almost matches Italy in this department. So if you are planning to go to Italy or France to watch matches then you are sure to have a lot of choice if you want to stay for a couple of days or more and prices will be reasonable as well given the number of choices. We can also imply that Italy and France are a better tourist destination than other countries in our list.

England is the home pubs and it is evident from the data. It's a place people go to chill and England has plenty of those. If you are travelling to England and want to have hangout then you have plethora of pubs nearby where you can enjoy after the games.

Spain is clear winner in Restaurants and Bars. There will be no shortage of these if someone is travelling to Spain for those Champions League games. You might even hang out in bar to enjoy your teams victory or sob after your teams loss.

Germany wins in no department which indicates that there may be a scope of improvement and potential business opportunities in hospitality business if Germany can attract more football tourists.

6. Conclusion

Football Stadiums all around the world are hotspots of events happening on weekly bases and Europe is the hub of Football in the world. There are millions of fans travelling on daily bases to see their stars play in whichever stadium they are. In terms of facilities these fan have in Europe, we can conclude that there is a lot of competition all around Europe in terms of which country is best in providing treatment to the travelling fans. There are a surprising fact that although English Premier League is there number one league in the world in terms of revenue yet Spanish La Liga ruled the charts. We can conclude from this finding that there is a shortage of real estate in prime location in England as England is a small country in terms of Area. This gives new businesses an edge in Spain due to its size in comparison to England. If Spain can increase the flow of more tourists on its land and improve in Hotel business then it will completely dominate the hospitality business for football fans in whole of Europe and evidently will become the best country to visit in the whole world for a football fan.

Other countries like France and Germany have the potential as there is enough real estate for new businesses to emerge and give fans plenty of choice. Italy and England are still doing great in these categories albeit the lack of usable land for setting up hospitality or emergency healthcare business.