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.

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. Then one hot encoding was performed on frequency of venue category to make the data ready for K-Means clustering.