Title:

ELT Project (Football Statistics)

Team: Sam Struthers, Bassam Aziz, Bryan Alcazar

Background:

In this project we sought to create a database on current professional soccer players and their expected goals. To achieve this we used, Pandas, MongoDB, BeautifulSoup and requests. We then created loops which go into the database and list the players by years active, position, age and nationality. We can then see each player’s performance stats like Expected Goals on FBref. Expected Goals is a metric in soccer which predicts how many goals a player is expected to score within 90 minutes. This is a very useful tool for people to plan for and predict what can happen in a soccer game, advancing our understanding and potentially the strategy within soccer.

We will be using web scraping tools to retrieve data from a football statistics webpage (FBRef.com) to extract various player’s information from six countries. Our dependencies utilized for extracting from the source page include BeautifulSoup, requests, pymongo, and pandas. We noticed that the URLs for Player information all were a base URL + that players id in the websites database. Once establishing the connection to the URL using requests and Beautifulsoup, our first step was to then go through player indexes and pull the player id from the href link inside of the html. We then inspected the page to view which class contains player information and find that it was under the <div class =”section-content”> part of html.

Once determining the class in which our players data is stored it now needs to be parsed based on the location of the players name, id, and position. To accomplish this task separate dictionaries are created to store every individual statistic. For extracting data, a loop is created to run through the section content data using the body text p and a to split the href for each dictionary. We then ran a for loop to run through the entire page and pull each player name, id, active years and position, placing them into the appropriate list. Once that was done we pushed the data to a MongoDB named Soccer, in the collection Players. When we had a final Database made up of these 6 countries, we exported it out of MongoDB compass to CSV so we could clean it up efficiently. We deleted non active players for this exercise (years active not ending in 2020). We ended with a final csv that was clean, containing the player information we needed to then create our ultimate database.

Our idea was that using this data (Player IDs) we could create a loop to run through that players information page and pull all the stats that were available in the FBref database. All the information was stored in tables so in order to clean it, we created a for loop that looped through each table under the class “section-content” that printed out the body of that table. We then mapped column headers to the appropriate row in the data pulled (Age is always the first row of the column, Country the second, etc.) so that we could then append those lists with the data. The idea would be that we could then create a mongo database under the player name with all of their statistics (such as expected goals). From there we would be able to take that data and create player profiles based on historical and current performance. (Which older players are starting to decline, younger players starting to peak)