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.

Extract (Data sources and how the data was formatted):

We will be using mongo scrapping to retrieve data from a football statistics webpage to extract various player’s information from six countries. Our dependencies utilized for extracting from the source page include BeautifulSoup, requests, pymongo, and pandas. Once establishing the connection to the url through mongo and Beautifulsoup, we then inspected the page to view which class contains player information. For our specific needs, the data of interest is stored within the div class named section content, which was further parsed based on the location of the players information.

Transform (What data cleaning or transformation was required):

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. From here, the data was further parsed and cleaned to extract as separate csv files for further manipulation.

Load (Final database, tables collected, and why):