Web Scraping and Social Media Scraping

Final Project

May 2022

Yelyzaveta Nemchynova 444467

Project Objective
Web Scraping of Premier League fixture results.

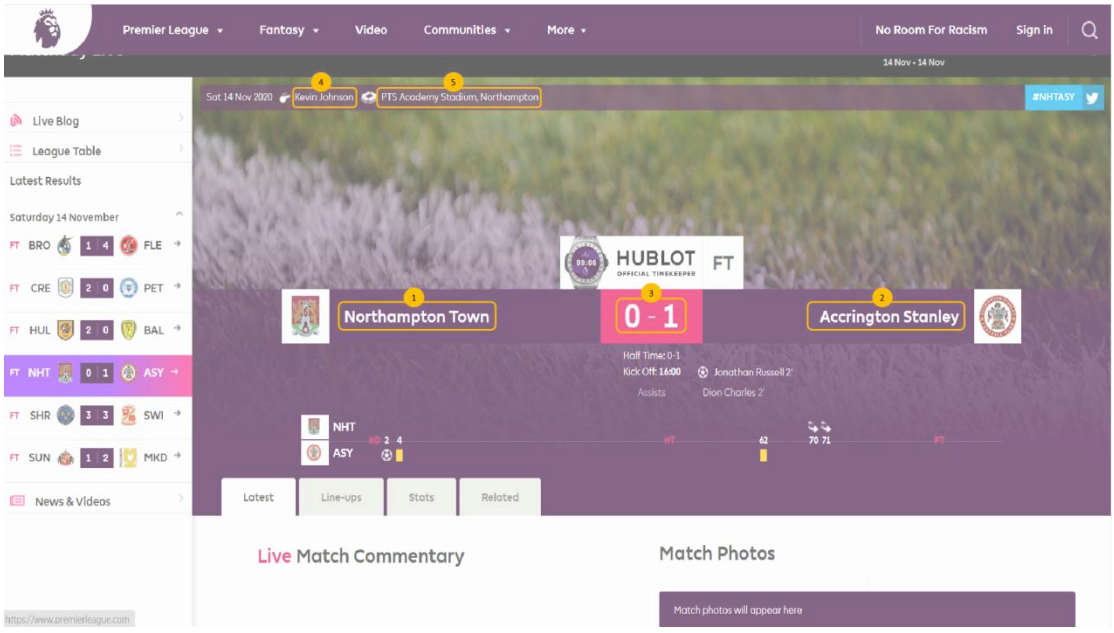
Business Objective
To investigate Premier League matches results in order to understand what are the most influential factors on a win of a match. (Surely, to answer this question, we would need quite a lot of variables and perform a regression analysis, but this business objective is provided as a potential use case if the scraped data is extended).
The questions which we can investigate with the scraped data:
1) Who scored the most at home? 2) Who scored the most away? 3) What is team's home/away performance?

Target data points:
- Home team (1)
- Away team (2)
- Home score (3)
- Away score (3)
- Referee (4)
- Stadium (5)

The source domain is: https://www.premierleague.com/

Some data points (such as match date, match stats) are dynamic parts of the page and are generated using JavaScript, therefore, omitted in order to be able to use Beautiful Soup in this project.

Understanding the Website
The URL for a match info is comprised of "https://www.premierleague.com/match/" followed by a distinctive match ID. Each ID is a 5-digit number for all matches, and IDs are sequenced.
For example, first match in 2021/2022 season has ID 58898 and, since the requirement is to scrape 100 pages, we would loop through this interval and collect the data from each match.



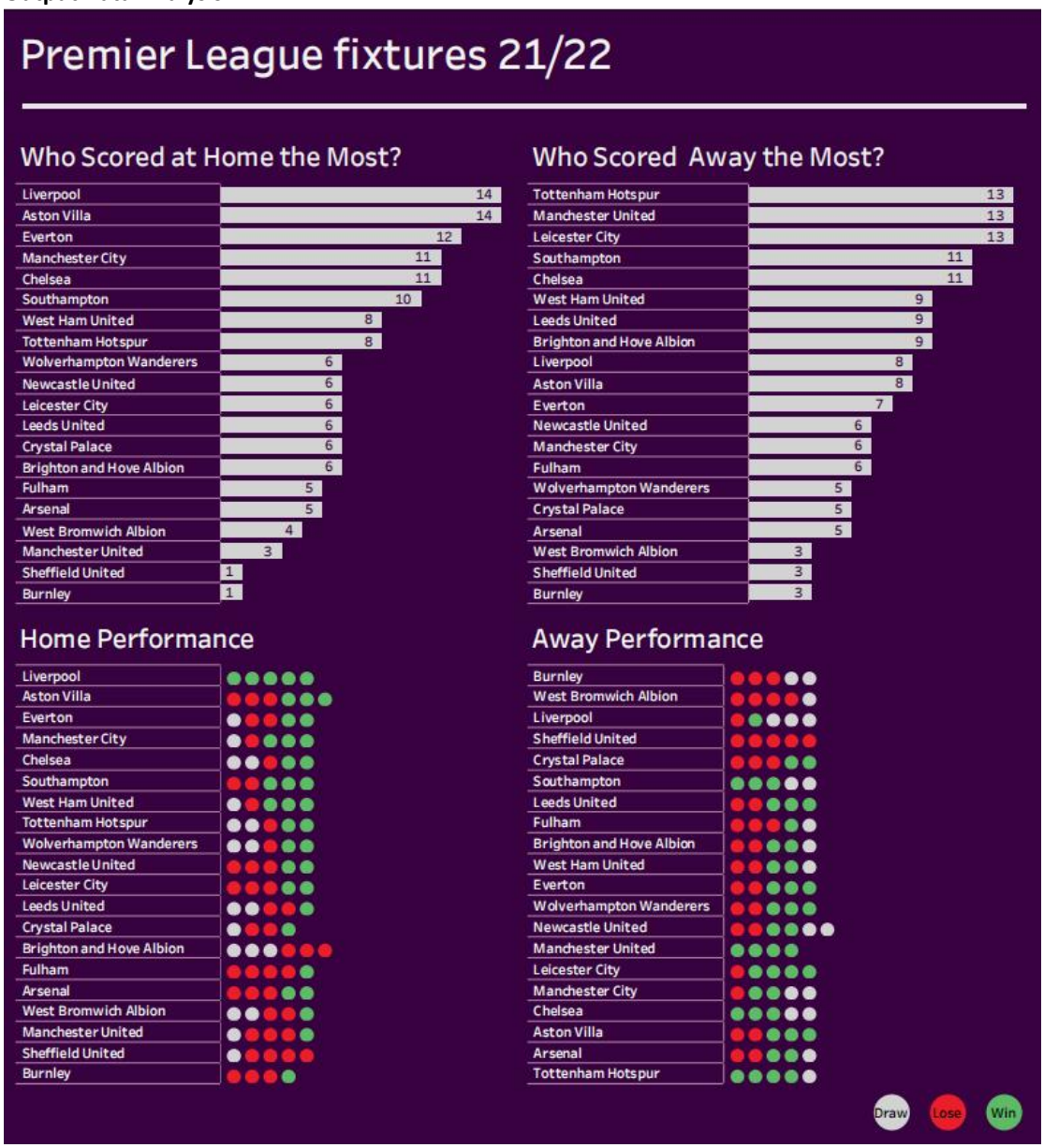
Scraping Data
The required data is scraped with a use of 3 different libraries: BeautifulSoup, Selenium and Scrapy. Conceptually, all three scrapers are doing the same thing:
1) Establishing a connection with the URL
2) Looping through a required interval of match IDs (limited to 100 pages)

3) Scraping the required data (Selenium and Scrapy: each data point is found through its Xpath, BeautifulSoup: each element is found by its class, tag)
4) Arranging data to the data frame and exporting the dataset to .csv
5) Calculating a runtime in order to compare a performance of 3 scrapers

Output Dataset
100 rows, 6 columns

column name	*data type*
home_team	object
away_team	object
home_score	object
away_score	object
referee	object
stadium	object

Output Data Analysis



Performance Comparison

Program	*Scraping Library*	*Runtime (seconds)*	*Comment*
PL_soup	BeautifulSoup	168.11734294891357	Performance of the scraper built using BeatuifulSoup has quite good performance in comparison with the scraper built using Selenium.
PL_selenium	Selenium	826.943377494812	Scraper which is built using Selenium framework ended up to have the highest runtime.
PL_scrapy	Scrapy	5.01	The most time efficient scraper. Scrapy framework is knows for very fast performance.