Web Scraping and Social Media Scraping

Project

Group members

Beata Michaluk, 412461

Przemysław Kępiński, 393785

Topic & web page

Our topic is Polish national football players. We choose the site http://www.hppn.pl/reprezentacja/pilkarze which contains all of the Polish national football team players. Through this page we can access the sub-pages with the basic data of each player – like name and surname, his matches or matches started on the bench.

What is our program doing?

First of all, our program downloads links from the homepage (http://www.hppn.pl/reprezentacja/pilkarze) – the links to pages about each player. Program collects 100 unique links. Next, the program goes to each site in turn (to each of the 100 links) and extracts specific data about the players. The program extracts following data: name and surname (‘name’), the number of played matches (‘matches’), the matches started on the bench (‘bench) and minutes played in all of matches (‘minutes’).

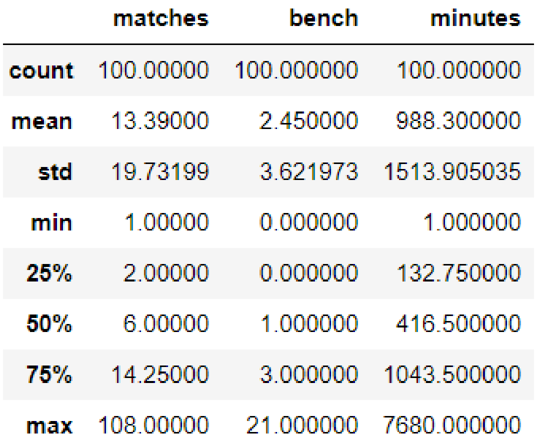
What do we get from the page?

From the homepage we get 100 unique links for pages about Polish national football players. Next, from each page, we extract data about players - name and surname (‘name’), the number of played matches (‘matches’), the matches started on the bench (‘bench) and minutes played in all of matches (‘minutes’).

Data analysis

The database contains 100 unique observations - 100 football players. There are 4 characteristics assigned to each player, i.e. name, number of matches played in the national team, number of the matches started on the bench and number of minutes played on the football field. We conducted an analysis of descriptive statistics of continuous data. The results are shown in the following table. Based on the extracted data and these from the table, it can be concluded that among the collected group of football players, they played an average of 13 matches. The highest number of matches played by a single player is 108. Each of the 100 players started his match on the bench about 2.5 times. Among the

100 players there were some who did not start any game on the bench and others who started on the bench 21 matches. Jakub Błaszczykowski spent the most number of minutes on the field with a total of 7680 minutes.



	matches	bench	minutes
count	100.00000	100.000000	100.000000
mean	13.39000	2.450000	988.300000
std	19.73199	3.621973	1513.905035
min	1.00000	0.000000	1.000000
25%	2.00000	0.000000	132.750000
50%	6.00000	1.000000	416.500000
75%	14.25000	3.000000	1043.500000
max	108.00000	21.000000	7680.000000

Of course our analysis is not exhaustive - with our data we can conduct further analyses. First of all, a simple econometric model can be created from the data collected. This would allow one to measure, for example, the effect of goals scored on the number of matches played. By extracting more data and combining them with our database, it would be possible to analyse even more precisely how the variables affect each other .

Distribution of tasks:

Beata Michaluk – Beautiful Soup, Scrapy, GitHub, Description

Przemysław Kępiński – Scrapy, Selenium, Description, Github