Names:

Michał Kunstler 412032, Mateusz Cedro 409632

Short description of the topic and the web page

Web site which we scrapped, The Basketball Reference web page (https://www.basketball-reference.com/leagues/NBA_2022_per_game.html), is a massive platform where any data which is related to National Basketball Association (NBA) can be found. There are statistics of the performance of all the teams, all the matches, and all the players from each season from the very beginning of the NBA competitions. Our project was to scrap the data about all the NBA players form current season – their name, number of games played, points scored, total rebounds, number of assists, and their physical information. In total we scrapped the data about the 616 NBA players.

Short description of the mechanism of our scrappers

Selenium – Our scrapper with the use of Selenium scrapped iteratively all the *hrefs* of all the NBA's players from the main page of the web site, and stored all of the in a list. Next, our scrapper went through all of the links to players and gathered the detail information of the full name of the player, number of games played, points scored, total rebounds, number of assists, and their physical information like weight and height. It was done with the use of specific *spans*.

Beautiful Soup – Our scrapper with the use of Beautiful Soup firstly found all hrefs which contained the '/players/' key world. Then with the use of extracted tags (hrefs) we created a list of set of the combinations of the main link to our website extended by the particular href to a particular player. We created the list of sets because there are some of players overlapped on the main website (and we do not know why). Next, our scrapper went iteratively through the entire list of links and scrapped the same information as it was done with the use of Selenium – looking for the exact *spans* which contained the data which we were looking for.

Scrapy –

Short technical description of the output

As a final output we gathered the data of all of the 616 NBA players who actively participate in current season competition. The data was collected to the pandas data frame for the ease of further data analysis.

Extremely elementary data analysis

Collection of such a data is very useful for calculating the descriptive statistics, and can be crucial for developing econometrics models which e.g. would predict the matches' results basing on

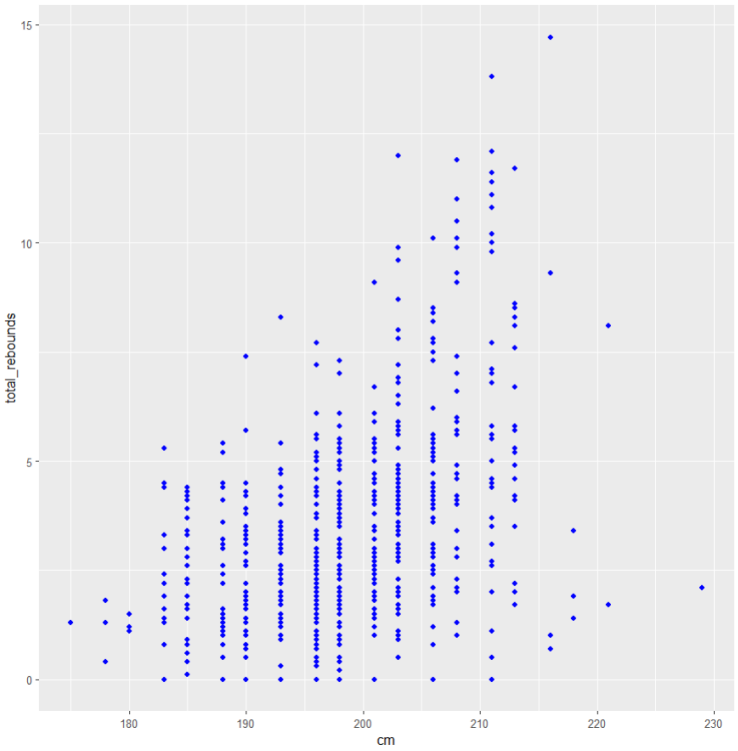
the set of players which play in both teams. Moreover, such data analysis might be helpful to find the optimal player for particular position on the pitch.

Since we get (scrap) the great amount of data, we can perform variety of analysis which might be the breakeven for the teams' performance and quality.

Therefore, when it comes to descriptive statistics:

	minimum	average	maximum
Games played	1	43.0	82
Assists	0	1.868	10.8
Points	0	8.24	30.6
Total rebounds	0	3.44	14.7
Height in cm	175	198.71	229
Weight in kg	72	96.99	141

For instance we can plot the data and see that the total rebounds is rather positively correlated with the height of the player:



Or we can develop models:

Call:

lm(formula = points ~ games + total_rebounds + assists + cm + kg, data = y)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.180457	4.615773	-0.256	0.798	
games	0.044092	0.006471	6.814	2.33e-11	***
total_rebo.	0.964120	0.091812	10.501	< 2e-16	***
assists	1.747893	0.104583	16.713	< 2e-16	***
cm	0.015311	0.027686	0.553	0.580	
kg	-0.021744	0.019705	-1.103	0.270	

--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.357 on 598 degrees of freedom

Multiple R-squared: 0.7211, Adjusted R-squared: 0.7187

F-statistic: 309.2 on 5 and 598 DF, p-value: < 2.2e-16

Participants' role:

Scrapy – Michał Kunstler

Beautiful Soup – Michał Kunstler

Selenium – Michał Kunstler & Mateusz Cedro

Data Analysis – Mateusz Cedro

Description – Mateusz Cedro

Comments in script – Mateusz Cedro