Web Scraping and Social Media Scraping

Final project

www.transfermarkt.com

Daniel Śliwiński
405848

Jarosław Leski
411174

Description of the project

The website that was selected for the project contains data on soccer clubs and players at all levels around the world. The aim of the project was to obtain comprehensive data on the Barclays Premier League (the UK's top football league) for recent seasons using three scrapers: BeautifulSoup, Scrapy and Selenium.

The design of transfermarkt.com allows for intuitive movement between teams and seasons. This has made it possible to scrape together information for the clubs such as position at the end of a given season, number of players, average age of players, average value of players and the total value of players playing for a given club. We have not limited ourselves to scraping data at club level. The next step of the project was to scrape the data for each player for each team in each season from 2015/2016 (the selection of both the starting and ending season is possible for the user by modifying the parameters that call the functions).

More technical information about the project and the resulting output can be found in the paragraphs below

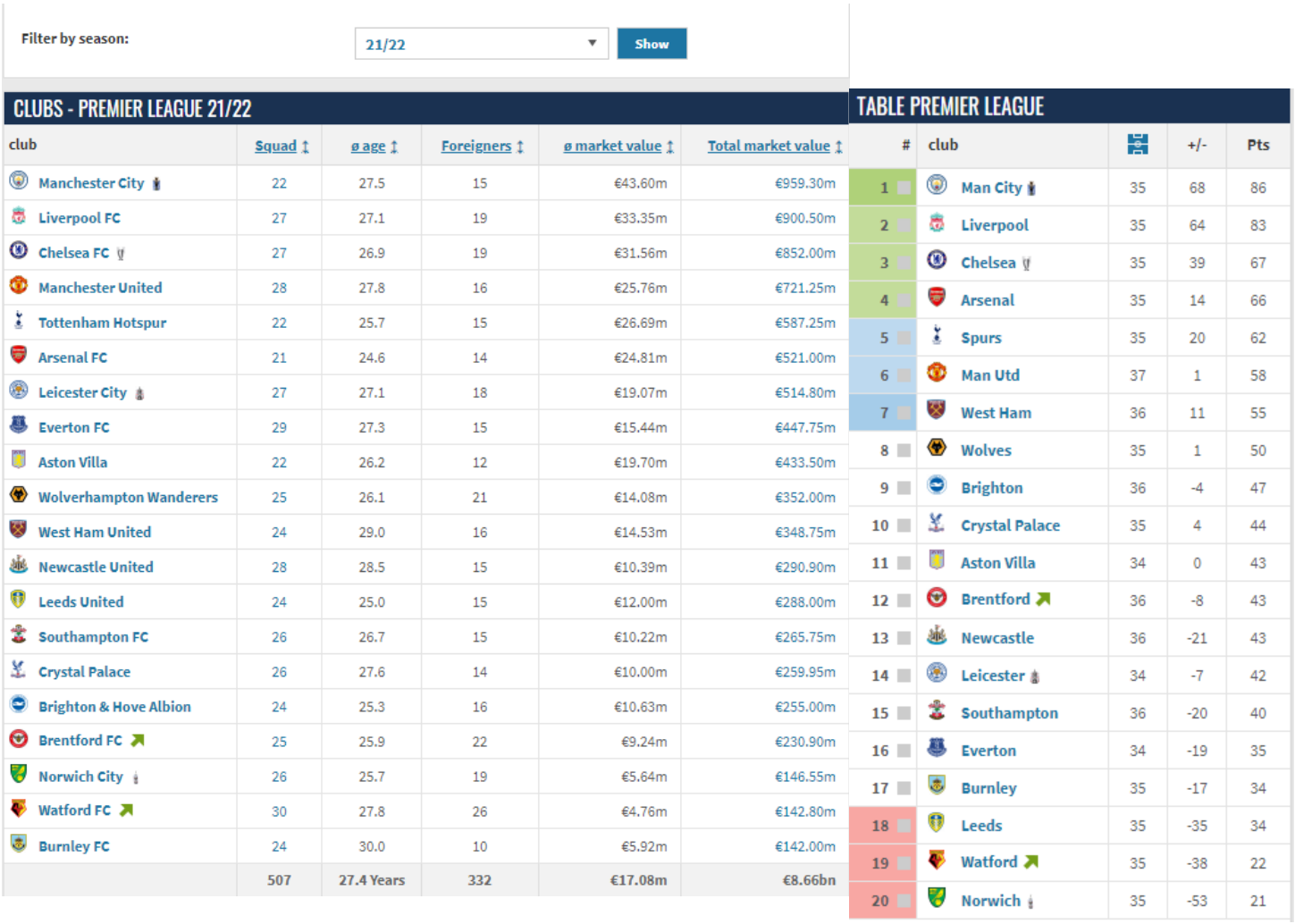
Description of scrapers mechanics

As mentioned above, we used three types of scrapers in this project, which ultimately produce the same output. The mechanics of the three scrapers are very similar, except of course for the obvious technical differences between them. The main conclusion to be drawn from this analysis is that scrapy is by far the fastest scraper. BeautifulSoup also performed well. Selenium, due to the need for time sleeps, takes the longest time and is inefficient when scraping many pages.

The program starts on the main Premier League page for the first season selected by the user, from which it downloads two tables:

- the first contains team statistics such as age and value of players
- the second one contains information about the performance of the teams during the season such as the number of points scored, matches played and goal difference

Figure 1: Two tables on the first page of the program activity.



Filter by season: 21/22 Show

CLUBS - PREMIER LEAGUE 21/22

club	Squad	ø age	Foreigners	ø market value	Total market value
Manchester City	22	27.5	15	€43.60m	€959.30m
Liverpool FC	27	27.1	19	€33.35m	€900.50m
Chelsea FC	27	26.9	19	€31.56m	€852.00m
Manchester United	28	27.8	16	€25.76m	€721.25m
Tottenham Hotspur	22	25.7	15	€26.69m	€587.25m
Arsenal FC	21	24.6	14	€24.81m	€521.00m
Leicester City	27	27.1	18	€19.07m	€514.80m
Everton FC	29	27.3	15	€15.44m	€447.75m
Aston Villa	22	26.2	12	€19.70m	€433.50m
Wolverhampton Wanderers	25	26.1	21	€14.08m	€352.00m
West Ham United	24	29.0	16	€14.53m	€348.75m
Newcastle United	28	28.5	15	€10.39m	€290.90m
Leeds United	24	25.0	15	€12.00m	€288.00m
Southampton FC	26	26.7	15	€10.22m	€265.75m
Crystal Palace	26	27.6	14	€10.00m	€259.95m
Brighton & Hove Albion	24	25.3	16	€10.63m	€255.00m
Brentford FC	25	25.9	22	€9.24m	€230.90m
Norwich City	26	25.7	19	€5.64m	€146.55m
Watford FC	30	27.8	26	€4.76m	€142.80m
Burnley FC	24	30.0	10	€5.92m	€142.00m
	507	27.4 Years	332	€17.08m	€8.66bn

TABLE PREMIER LEAGUE

#	club		+/-	Pts
1	Man City	35	68	86
2	Liverpool	35	64	83
3	Chelsea	35	39	67
4	Arsenal	35	14	66
5	Spurs	35	20	62
6	Man Utd	37	1	58
7	West Ham	36	11	55
8	Wolves	35	1	50
9	Brighton	36	-4	47
10	Crystal Palace	35	4	44
11	Aston Villa	34	0	43
12	Brentford	36	-8	43
13	Newcastle	36	-21	43
14	Leicester	34	-7	42
15	Southampton	36	-20	40
16	Everton	34	-19	35
17	Burnley	35	-17	34
18	Leeds	35	-35	34
19	Watford	35	-38	22
20	Norwich	35	-53	21

Source: https://www.transfermarkt.com/premier-league/startseite/wettbewerb/GB1

The program then iterates through the rows of the table by clicking on the hyperlinks of all the teams in the season, which lead to a view of the roster the team had in that season.

Below is a screenshot for a sample team. The operation of the program at this stage is to iterate through the rows and columns to retrieve data on identity, position, age, current team, historical market value and nationality. The iteration is dynamic, i.e. the program catches how many players have played for the given team (this is not a fixed parameter as these values vary from season to season)

Figure 2. Sample page view from the team level

#	player	Date of birth / Age	Nat.	Current club	Market value
31	Ederson Goalkeeper	Aug 17, 1993 (26)			€50.00m
13	Zack Steffen Goalkeeper	Apr 2, 1995 (25)			€6.00m
-	Scott Carson Goalkeeper	Sep 3, 1985 (34)			€300Th.
-	James Trafford Goalkeeper	Oct 10, 2002 (17)			
3	Rúben Dias Centre-Back	May 14, 1997 (23)			€75.00m
14	Aymeric Laporte Centre-Back	May 27, 1994 (26)			€45.00m
6	Nathan Aké Centre-Back	Feb 18, 1995 (25)			€32.00m
5	John Stones Centre-Back	May 28, 1994 (26)			€30.00m
50	Eric García Centre-Back	Jan 9, 2001 (19)			€20.00m

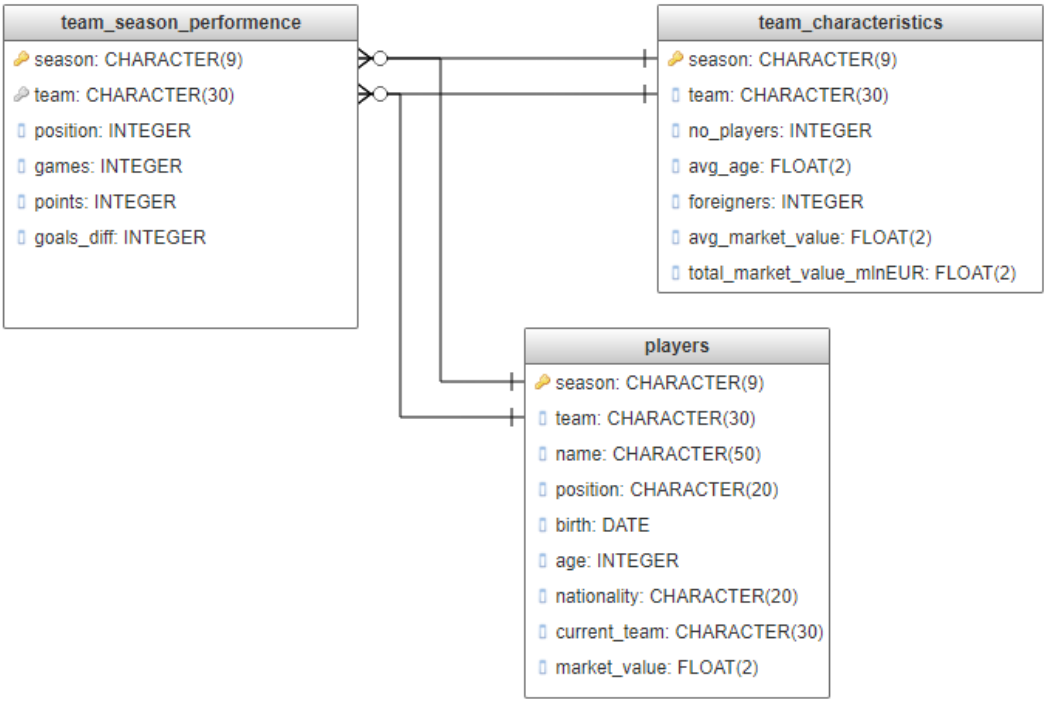
Source: https://www.transfermarkt.com/manchester-city/startseite/verein/281?saison_id=2020

The above steps are repeated for all seasons passed in the function call parameter.

Description of the output

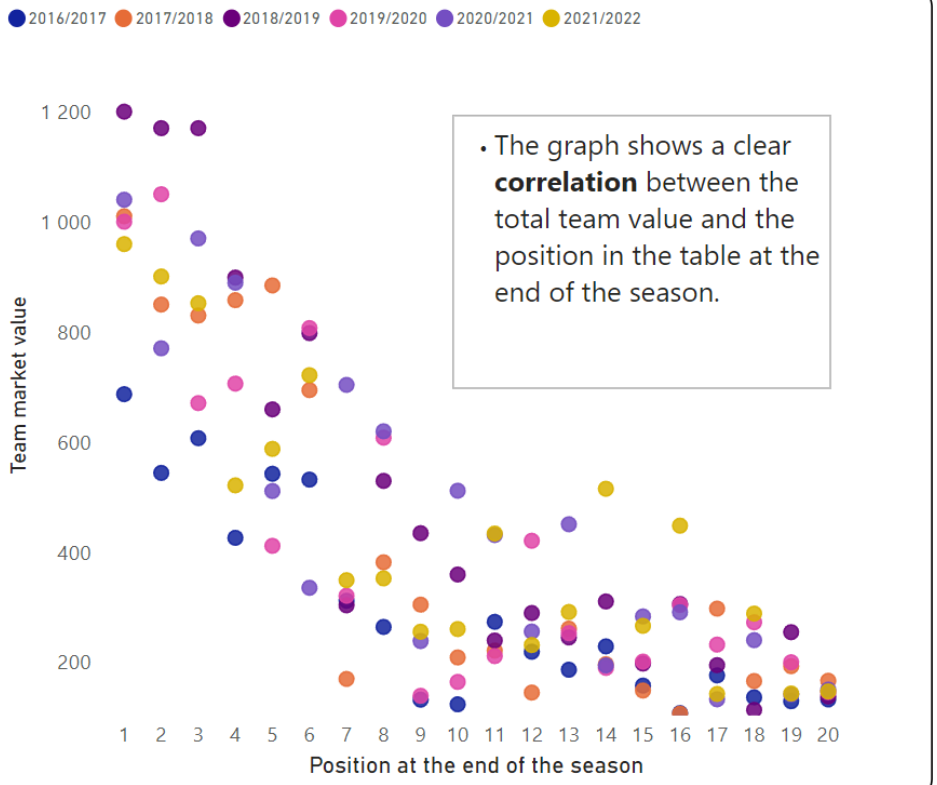
Three tables (which sources are described and visualized above) were obtained as the output of all three scrapers. Below is a graph of the resulting 'database'. Each of the scraped observations has a season and team combination which allows us to combine the data into a relational model that can be used to analyze the data.

Figure 3. Database diagram



Source: Own study

Elementary data analysis



- The graph shows a clear **correlation** between the total team value and the position in the table at the end of the season.

Season

2021/2022

- The youngest player currently playing is Evan Ferguson, while the oldest is Willy Caballero.
- The most non-British players are from France, Brazil and Spain.
- The most expensive players in the league are currently Harry Kane and Mohamed Salah (100 million euros).
- On average, the most expensive players in the league are strikers at 21.69 million euros.

Youngest Player

Evan Ferguson

17
age

2004
birth_year

Oldest Player

Willy Caballero

40
age

1981
birth_year

Nationalities

France **28**

Brazil **25**

Spain **25**

The most expensive players

Harry Kane

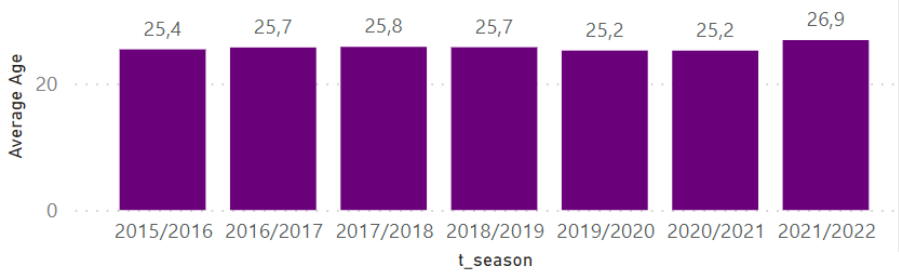
100,00
market_value[mlnE]

Mohamed Salah

100,00
market_value[mlnE]

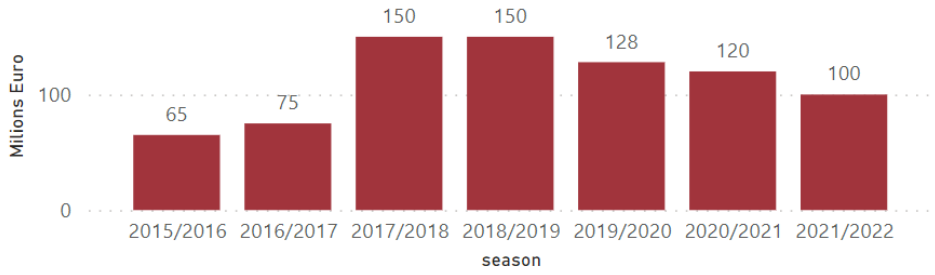
Average market value [mln EUR] per position

Average age over the seasons



- The average age in the league has clearly increased in the current season.
- The most expensive player in league history played in the 2017/2018 and 2018/2019 seasons.

The most expensive players over the seasons



Distribution of the work among team members

Project part		Jarosław Leski	Daniel Śliwiński
Beautiful Soup:			
	Code		**X**
	Paths		**X**
	Documentation		**X**
Scrapy:			
	Code		**X**
	Paths	**X**	
	Documentation		**X**
Selenium:			
	Code	**X**	
	Paths	**X**	
	Documentation	**X**	
Description file		**X**	