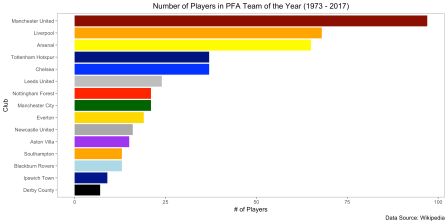
The Professional Footballers’ Association (PFA) [Team of the Year](https://en.wikipedia.org/wiki/PFA_Team_of_the_Year) is released in England at the end of each season, picking the 11 most influential players in each of Britain’s leagues.

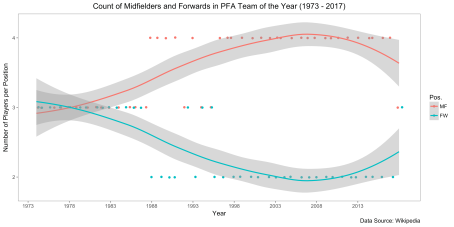
The Team of the Year award was launched in the 1973-1974 season, meaning there was 44 years worth of data to web scrape. Using Wikipedia’s [PFA Team of the Year pages](https://en.wikipedia.org/wiki/PFA_Team_of_the_Year) (filtered by decade) and the rvest package, I was left with a dataframe of 484 soccer players (44 years \* 11 players per/year).

Here are some visualizations I thought were cool:

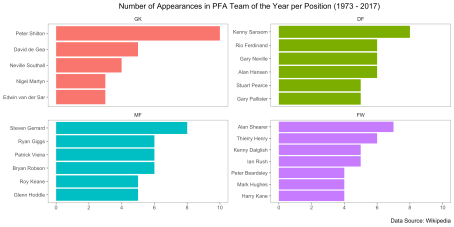
* Liverpool, Arsenal, and (most significantly) Manchester United are represented more than other clubs. United need just three more players to hit the 100 mark.

[](https://worldsocceranalytics.files.wordpress.com/2018/10/static_plot_players_per_club1.png)

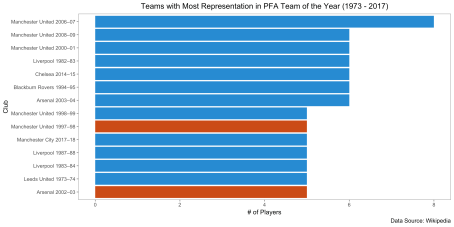
* The PFA Team of the Year formation has changed over time from 4-3-3 to 4-4-2. The graph below shows the number of forwards and midfielders who featured in the ranking each year. Up until the late 1980s, the formation included three midfielders and three forwards (the Team of the Year has always had four defenders and one goalkeeper). In the 1990s, both formations were used but since the 2000s, the use of two strikers has been the status quo.



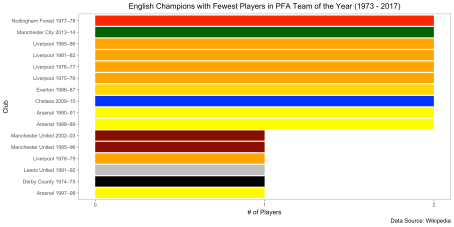
* Additionally, I thought it was interesting to see the top individual players in each position. Peter Shilton remains the only player to hit 10 First Division PFA Team of the Year Awards.

[](https://worldsocceranalytics.files.wordpress.com/2018/10/static_plot_players_appearances_number1.png)

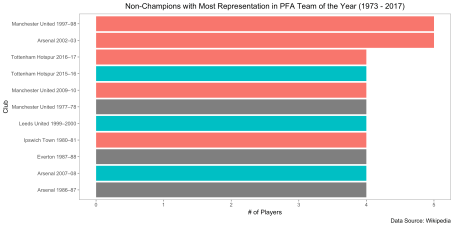
* The more interesting analysis involved merging Team of the Year data with [historical league position.](https://en.wikipedia.org/wiki/List_of_English_football_champions) We can now ask questions along the lines of: did the majority of players in the Team of the Year come from the champions that year or other teams? Looking at the specific seasons with the most representation in the award, only 14 times did a team have 5+ players on the list; the majority of these teams came from the champions (12 out of 14); only Manchester United in 1998 and Arsenal in 2003 (both runner-ups) feature in this list.

[](https://worldsocceranalytics.files.wordpress.com/2018/10/teams_most_representation_per_year2.png)

* And here are the champions of England with the *fewest* players on the Team of the Year. There are many from the 1970s and 1980s, which hints to a change in the way the Team of the Year was designed over the years (it may have been purposefully less top-heavy in the past, whereas today the English champions often features more significantly on the list). We can see that more recently Manchester City in 2014 and Chelsea in 2010 were on the lower end in terms of representation.

[](https://worldsocceranalytics.files.wordpress.com/2018/10/teams_most_representation_per_year_lowest_champions1.png)

* While we’re at it, which teams that didn’t finish 1st had the most players represented? It’s clear that there were some pretty high quality teams that didn’t win the league these years (runners-up are in red, third place teams in green, and any from 4th to 20th in grey).



I think that’s enough visualizations for today, but there’s definitely a lot more we can analyze with this data. Let me know if you have any questions or feedback.

**R Code Snapshot (full code can be found on** [**Github**](https://github.com/stefangouyet/epl_team_of_year)**):**

Step 1 – Web Scraping:

#initialize data.frame

df <- data.frame(`Pos.`= as.character(),

Player = as.character(),

Club = as.character(),

`App.` = as.double(),

year = as.character())

div\_table\_numbers <- c(1,5,9,13,17,21,25,29,33,37)

urls <- c("https://en.wikipedia.org/wiki/PFA\_Team\_of\_the\_Year\_(2000s)",

"https://en.wikipedia.org/wiki/PFA\_Team\_of\_the\_Year\_(1990s)",

"https://en.wikipedia.org/wiki/PFA\_Team\_of\_the\_Year\_(1980s)")

for (i in 1:length(urls)){

for (j in 1:length(div\_table\_numbers)) {

  xpath\_base <- '//\*[@id="mw-content-text"]/div/table['

new\_data <- urls[i] %>%

html() %>%

html\_nodes(xpath = paste0(xpath\_base,

div\_table\_numbers[j],

"]")) %>%

html\_table()

new\_data <- new\_data[[1]]

year\_xpath\_base <- '//\*[@id="mw-content-text"]/div/h3['

year <- urls[i] %>%

html() %>%

html\_nodes(xpath = paste0(year\_xpath\_base,

j,

"]")) %>%

html\_text()

year <- year %>% str\_remove\_all(fixed("[edit]"))

new\_data$year <- year

df <- bind\_rows(df,

new\_data)

}

}

Step 2: Exploratory data analysis and visualizations

club\_count <- df %>% count(Club, sort = TRUE)

club\_count %>%

top\_n(n = 15) %>%

ggplot(aes(x = reorder(Club,n),

y = n,

fill = Club)) +

geom\_col() +

coord\_flip()+

theme\_few()+

guides(fill=FALSE)+

labs(y = "# of Players",

x = "Club",

title = "Number of Players in PFA Team of the Year (1973 - 2017)",

caption = "Data Source: Wikipedia")+

theme(plot.title = element\_text(hjust = 0.5)) +

scale\_fill\_manual(values = c("Manchester United" = "darkred",

"Liverpool" = "orange",

"Arsenal" = "yellow",

"Chelsea" = "blue",

"Blackburn Rovers" = "lightblue",

"Leeds United" = "grey",

"Manchester City" = "dark green",

"Derby County" = "black",

"Everton" = "gold",

"Nottingham Forest" = "red",

"Ipswich Town" = "darkblue",

"Southampton" = "orange",

"Newcastle United" = "darkgrey",

"Aston Villa" = "purple",

"Tottenham Hotspur" = "navy"

))

df <- df %>%

mutate(short\_year = str\_sub(year,1,4) %>%

as.numeric() + 1)

order\_positions <- c("GK","DF","MF","FW")

df <- df %>% mutate(Pos. = fct\_relevel(Pos., order\_positions))

count\_position <- df %>%

filter(Pos. %in% c("MF","FW")) %>%

count(Pos.,short\_year, sort = TRUE)

count\_position %>%

ggplot(aes(x=short\_year,y=n,color=Pos.,group=Pos.)) +

geom\_point(position=position\_jitter(h=0.005))+

geom\_smooth(method = "loess")+

scale\_x\_continuous(breaks = seq(1973, 2017, 5))+

scale\_y\_continuous(breaks=seq(2,4,1))+

labs(x = "Year",

y = "Number of Players per Position",

title = "Count of Midfielders and Forwards in English PFA Team of the Year (1973 - 2017)",

caption = "Data Source: Wikipedia")+

theme\_few()+

theme(plot.title = element\_text(hjust = 0.5))

Step 3: More web scraping and merging datasets

first\_div\_top\_three\_url <- "https://en.wikipedia.org/wiki/List\_of\_English\_football\_champions"

first\_div\_top\_three\_xpath <- '//\*[@id="mw-content-text"]/div/table[2]'

first\_div\_top\_three <- first\_div\_top\_three\_url %>%

html() %>%

html\_nodes(xpath = first\_div\_top\_three\_xpath) %>%

html\_table()

first\_div\_top\_three <- first\_div\_top\_three[[1]]

first\_div\_top\_three <- first\_div\_top\_three %>% filter(!(Year %in% c("1915/16–1918/19",

"1939/40–1945/46")))

first\_div\_top\_three$Goals <-first\_div\_top\_three$Goals %>% as.numeric()

first\_div\_top\_three <- first\_div\_top\_three %>%

rename(`Champions` = `Champions(number of titles)`,

`Top goalscorer` = `Leading goalscorer`)

epl\_top\_three\_url <- "https://en.wikipedia.org/wiki/List\_of\_English\_football\_champions"

epl\_top\_three\_xpath <- '//\*[@id="mw-content-text"]/div/table[3]'

epl\_top\_three <- epl\_top\_three\_url %>%

html() %>%

html\_nodes(xpath = epl\_top\_three\_xpath) %>%

html\_table()

epl\_top\_three <- epl\_top\_three[[1]]

epl\_top\_three <- epl\_top\_three %>%

rename(`Champions` = `Champions (number of titles)`)

english\_top\_three\_total <- bind\_rows(first\_div\_top\_three,

epl\_top\_three)

english\_top\_three\_total$Champions <- english\_top\_three\_total$Champions %>% str\_remove\_all(regex("\\([^)]\*\\)"))

english\_top\_three\_total$Champions <- english\_top\_three\_total$Champions %>% str\_remove\_all(regex("\\[.\*?\\]"))

english\_top\_three\_total <- english\_top\_three\_total %>%

mutate(short\_year = str\_sub(Year,1,4) %>% as.numeric() + 1)

english\_top\_three\_total <- english\_top\_three\_total %>%

filter(short\_year > 1973) %>%

select(-c(`Top goalscorer`,Goals))

english\_top\_three\_total\_melted <- english\_top\_three\_total %>%

melt(id.vars=c("Year","short\_year"),

value.name = "Club",

variable.name = "Team\_Ranking")

english\_top\_three\_total\_melted$Club <- english\_top\_three\_total\_melted$Club %>% str\_trim(side = c( "right"))

df\_merged <- df %>% left\_join(english\_top\_three\_total\_melted,

by = c("Club","short\_year"))

Step 4: More data visualizations with merged dataset

club\_count\_year <- df\_merged %>%

count(Club, year, Team\_Ranking, sort = TRUE) %>%

mutate(club\_year = paste(Club, year))

club\_count\_year %>%

top\_n(n = 10, wt = n) %>%

ggplot(aes(x = reorder(club\_year,n),

y = n))+

geom\_col(aes(fill = factor(ifelse(Team\_Ranking == "Champions",

1,

2)))) +

coord\_flip()+

theme\_few()+

guides(fill=FALSE)+

labs(y = "# of Players",

x = "Club",

title = "Teams with Most Representation in PFA Team of the Year (1973 - 2017)",

caption = "Data Source: Wikipedia")+

theme(plot.title = element\_text(hjust = 0.5)) +

scale\_fill\_solarized()

club\_count\_year\_champions <- df\_merged %>%

count(Club, year, Team\_Ranking, sort = TRUE) %>%

mutate(club\_year = paste(Club, year)) %>%

filter(Team\_Ranking == "Champions")

club\_count\_year\_champions %>%

top\_n(n = -10, wt = n) %>%

ggplot(aes(x = reorder(club\_year,n),

y = n))+

geom\_col(aes(fill = Club)) +

coord\_flip()+

theme\_few()+

guides(fill=FALSE)+

labs(y = "# of Players",

x = "Club",

title = "English Champions with Fewest Players in PFA Team of the Year (1973 - 2017)",

caption = "Data Source: Wikipedia")+

theme(plot.title = element\_text(hjust = 0.5)) +

scale\_y\_continuous(breaks = seq(0,2,1)) +

scale\_fill\_manual(values = c("Manchester United" = "darkred",

"Liverpool" = "orange",

"Arsenal" = "yellow",

"Chelsea" = "blue",

"Blackburn Rovers" = "lightblue",

"Leeds United" = "grey",

"Manchester City" = "dark green",

"Derby County" = "black",

"Everton" = "gold",

"Nottingham Forest" = "red"))