Revisiting the 2022-2023 Soccer Teams of the Season

Final Technical Report

Ebubechukwu U. Nwoke

Oral Roberts University

CSC 201: Introduction to Data Science

Professor Jonathan Weed

December 13th, 2024

Abstract

Like in every other sport, statistics and data have made their way into the worldwide game of soccer, and year by year, using numbers and using them accurately can be the difference between winning games and winning trophies. However, from a spectator's point of view, using numbers to figure out how a team can improve doesn't mean much. What's more important for fans is being able to use numbers to back up claims and agendas they may or may not make while watching games. These different viewpoints all build towards the end of the season, where fans, the media, and even the players get into extensive debates about who the best and worst performers of the season are. While stats play a part, almost all final decisions for awards and nominations come down to human decisions, which can be heavily biased. For this project, I wanted to see what it would take for an algorithm to decide who the season's best performers were.

Revisiting the 2022-2023 Soccer Teams of the Season

I downloaded a data set that tracked the stats of every player in Europe's Top 5 domestic leagues for the 2022-23 season. The first thing I plan on doing is cleaning the data in Excel. From briefly scanning the data set, I knew there were a lot of issues with spelling, so I knew I would need to go through it closely to ensure all the non-numeric data was accurate. After cleaning the data, I plan on doing some exploratory data analysis in Excel, finding common traits between the better players and visualizing the stylistic differences between positions, teams, and even leagues. After doing this, I plan on going into RStudio and using some of the more important stats to rank all the players. Ultimately, I aim to have a decently accurate ranking system weighted correctly for every position. This process should leave me with a ranking of the players, which I'd be able to use their real-life positions to fit into my Team of the Season, one for each of the domestic leagues and one for all of Europe.

There will be a lot of challenges when it comes to sorting through the data, as well as deciding what statistics are more important. However, as someone who follows the sport closely and is used to playing soccer competitively, my knowledge of the game should help me make decisions. I also plan to verify every decision I make, ensuring that the players that appear at the top of my rankings are consistent with the players I remember were elite that year, as that season, I was able to follow many of the teams very closely. Lastly, once I have my six sets of teams, I can compare who made my list and who didn't with the teams compiled that season, and I can try to figure out why certain players may have gotten snubbed or players my program missed made the cut.

Cleaning the Data

I first found a data set that contained players from the top 5 leagues in Kaggle. There were three that I saw, but I eventually settled with one with 126 total columns, which gave me a ton of detailed data about each player. One caveat I noticed after downloading the data was that it was taken midway through the season. I figured this out because one of the players from the team I support, Erling Haaland, had a goal tally of 25 listed, and in that season, he ended the campaign with 36 goals, breaking the record for most goals in a season (in the English League). Regardless, I decided to continue, as the other data sets I looked at either didn't have all the players from the leagues or had less detailed data. With my analysis for the rest of the project, we have to consider that these statistics were taken from mid-February 2023, about 60% of the way through the season. Anything any player did from late February to the end of the season was voided.

Data Cleaning

Step 1: Adjusting Player Names. Once I downloaded the data, I began cleaning, which was time-consuming and tedious. I chose to do this in Excel because it was easier to see the live changes and undo any mistakes, as many of the changes I had to do were by hand. The first change I made was adjusting the names of the players. I originally planned on giving every player a first and last name column, but after experimenting, I decided to leave their full name in one cell. Many players' names, like Rodrygo or Charles De Ketelaere, weren't two names. To prevent me from having to deal with empty name cells or having to decide how to categorize first and last names, leaving all names as a singular string and then adjusting as needed for visualizations is the safer option. From there, I also had to account for fixing players with Slavic names. Kaggle was not able to register certain letters like 'č' or 'Ł,' so a lot of players from

Croatia, Poland, Serbia, and other Slavic countries had question marks in their names, pictured here:

Player	Nation	Pos	Squad	Comp	Age
Toma Baši?	CRO	MF	Lazio	Serie A	26
Kristijan Bistrovi?	CRO	MF	Lecce	Serie A	24
Domagoj Bradari?	CRO	DF	Salernitana	Serie A	23
Josip Brekalo	CRO	FW	Fiorentina	Serie A	24
Josip Brekalo	CRO	MFFW	Wolfsburg	Bundesliga	24
Marcelo Brozovi?	CRO	MF	Inter	Serie A	30
Ante Budimir	CRO	FW	Osasuna	La Liga	31
Duje ?aleta-Car	CRO	DF	Southampton	Premier League	26
Duje ?aleta-Car	CRO	DF	Marseille	Ligue 1	26
David ?olina	CRO	DFMF	Augsburg	Bundesliga	22
Dion Drena Beljo	CRO	FW	Augsburg	Bundesliga	20
Martin Erlic	CRO	DF	Sassuolo	Serie A	25
Bartol Franji?	CRO	MF	Wolfsburg	Bundesliga	23
Ivo Grbi?	CRO	GK	Atlético Madrid	La Liga	27
Joško Gvardiol	CRO	DF	RB Leipzig	Bundesliga	21
Kristijan Jaki?	CRO	DFMF	Eint Frankfurt	Bundesliga	25
Marin Jakoliš	CRO	FW	Angers	Ligue 1	26
Josip Juranovi?	CRO	DF	Union Berlin	Bundesliga	27
Mateo Kova?i?	CRO	MF	Chelsea	Premier League	28

To remedy this, I had to copy and paste the player's name into Google manually, figure out what letters were missing, and paste the names back into the dataset. Fortunately, Excel and R allow special characters like this, so I didn't have to worry about conversions. This was the most prolonged process in the data cleaning, and it took about three hours over multiple days.

Step 2: Removing Duplicate Players. The next part of the cleaning was to remove duplicates. In August and January, players and teams are given the option to initiate transfers, and players can switch their clubs halfway through. This data took the players who experienced this change and split their stats, one being their club at the start of the season and the other being the club they transferred to. Initially, I planned to merge the duplicates. After reading into a lot of the stats, however, recalculating a lot of them would be too difficult to track, as all the stats besides goals and assists are on a per90 basis, which means stats get inflated/deflated depending

on playing time. I removed the row with fewer total minutes, meaning the spell when the player contributed the most is still in the set.

Step 3: Trimming the Player Pool. After working with the names, I could finally start cleaning the numbers in the data. I made two significant changes to this part. The first change I made was establishing a minimum requirement for the data. As I previously mentioned, many of the advanced statistics provided are on a per90 basis. If a substitute player who doesn't get much playing time gets a goal or creates a key chance in the little time they get, their stats would get inflated, meaning better players who play more often might not beat them in the rankings. If this dataset had used counting stats, I would have been able to do the calculations myself to establish a fairer metric, but this is what I was given. To decide this cutoff, I filtered out the team I follow: Manchester City. From that pool of about 20 players, I was able to sort them based on their minutes played, and it looked like this:

Player	Nation	Position	Team	League	Age	Appearances	Starts	Minutes
Kalvin Phillips	ENG	MF	Manchester City	Premier League	27	3	0	21
Sergio Gómez	ESP	DFMF	Manchester City	Premier League	22	5	0	111
Cole Palmer	ENG	FWMF	Manchester City	Premier League	20	9	0	119
Aymeric Laporte	ESP	DF	Manchester City	Premier League	28	5	4	376
Rico Lewis	ENG	DFMF	Manchester City	Premier League	18	9	5	453
Julián Álvarez	ARG	FWMF	Manchester City	Premier League	23	16	5	578
Kyle Walker	ENG	DF	Manchester City	Premier League	32	11	10	793
Rúben Dias	POR	DF	Manchester City	Premier League	25	13	9	874
Riyad Mahrez	ALG	FWMF	Manchester City	Premier League	31	16	10	898
Jack Grealish	ENG	FW	Manchester City	Premier League	27	15	11	1040
Phil Foden	ENG	FWDF	Manchester City	Premier League	22	18	13	1045
Nathan Aké	NED	DF	Manchester City	Premier League	27	15	13	1156
John Stones	ENG	DF	Manchester City	Premier League	28	14	13	1181
İlkay Gündoğan	GER	MF	Manchester City	Premier League	32	18	14	1252
João Cancelo	POR	DF	Manchester City	Premier League	28	17	16	1274
Manuel Akanji	SUI	DF	Manchester City	Premier League	27	15	14	1286
Bernardo Silva	POR	MFFW	Manchester City	Premier League	28	20	16	1410
Kevin De Bruyne	BEL	MF	Manchester City	Premier League	31	20	18	1599
Erling Haaland	NOR	FW	Manchester City	Premier League	22	20	19	1636
Rodri	ESP	MF	Manchester City	Premier League	26	20	20	1733
Ederson	BRA	GK	Manchester City	Premier League	29	21	21	1890

I know my team well, so I picked my cutoff based on who I considered important that season. I settled on between Julián Álvarez, our backup striker, and Rico Lewis, one of our younger players who didn't get many valuable minutes. Since Julián played a significant enough role for

us during this period in the season, I used his minutes to calculate the cutoff for every player in the dataset. At the time the data was collected, Julián had amassed 573 minutes. When we divide that by the length of a football match, 90 minutes, we get a match equivalent of about 6.37 full games played. Using that, I set my cutoff at a match equivalent of 6, meaning only players who had played a minimum of 540 minutes would be kept in the set, and everyone else would be taken out. This took my player pool from 2530 to 1511, which is about 16 players per team, right around the ratio that should be eligible for the awards calculation I plan on doing.

Step 4: Removing Unnecessary Data. Lastly, I had to not only rename the columns of the stats but also remove duplicate stat columns and columns of data I thought were optional. I first renamed everything using the Kaggle site where I got the dataset. Thankfully, the user who compiled the data gave a key explaining what each column tracked. From this, I was able to rename all the columns, and the final adjustments looked like this:

Age	Appearances	Starts	Minutes	Total 90s	Goals	Shots	ShotsOnTarget	ShotsOnTargetPct	GoalsPerShot	GoalsPerShotOnTarget
30	17	16	1244	13.8	0	0.22	0.14	66.7	0	0
33	17	14	1219	13.5	2	1.26	0.44	35.3	0.12	0.33
26	20	20	1733	19.3	1	1.61	0.36	22.6	0.03	0.14
25	13	9	874	9.7	0	0.52	0.1	20	0	0
32	11	10	793	8.8	0	0.34	0.11	33.3	0	0
27	15	13	1156	12.8	0	0.31	0	0	0	0
26	10	9	723	8	0	0.38	0.25	66.7	0	0
33	14	9	931	10.3	1	0.87	0.19	22.2	0.11	0.5
36	20	18	1534	17	0	0.59	0.24	40	0	0
24	19	19	1700	18.9	0	0.26	0.05	20	0	0
26	13	7	566	6.3	0	0.95	0.32	33.3	0	0
21	18	13	1216	13.5	0	0.44	0.15	33.3	0	0
23	20	20	1700	18.9	1	0.42	0.11	25	0.13	0.5
27	15	14	1286	14.3	0	0.77	0.14	18.2	0	0
28	16	12	927	10.3	0	0.97	0.29	30	0	0
26	17	15	1377	15.3	1	1.05	0.39	37.5	0.06	0.17
28	14	13	1181	13.1	0	0.53	0.08	14.3	0	0
28	19	18	1599	17.8	3	1.18	0.51	42.9	0.14	0.33
26	14	11	941	10.5	1	1.24	0.19	15.4	0.08	0.5
32	12	6	620	6.9	1	0.43	0.14	33.3	0.33	1

I then went through each column, deciding if I deemed it necessary to take into account when determining what would go into my calculations. After all this, I was left with a dataset with 1511 players and 74 columns of data, which is still more than enough to do my analysis. I also noticed the lack of specific stats for goalkeeping, such as saves, goals prevented, etc. I

considered taking them out of the dataset, but later in the analysis, I will try and use some of the passing and distribution stats to find a fair metric for that position, as being good on the ball has become increasingly more critical for keepers as the game evolves.

Exploratory Data Analysis

With the data finally cleaned, I wanted to look at different statistics and metrics within the data to see if anything stood out. This was separate from the teams of the season, as different averages, rankings, and everything in between wouldn't be necessary to use for my final team of the season calculations. In this section, I will discuss the differences in the five leagues, the players with the most goal contributions, comparisons between the goal scorers for two league winners, and an in-depth look at the most dangerous dribblers across Europe. Lastly, I will conclude this EDA portion with a deep dive into Erling Haaland's unprecedented start to that season. I will see if I can use linear regression to predict his final goal contribution tally.

League Comparisons

The first thing on my agenda was to compare the differences between the top five leagues. To do this, I made a different data frame and filled it with the number of players, average age, average and standard deviation of minutes, goals, and average fouls. I made sure to have five different rows for the five leagues, then took the averages for all the players in each, filtering it out with R's "which" function. The data frame ended up looking like this:

League	NumPlayers	AVGAge	AVGMinutes	STDVMinutes	Goals	SoloGoals	AVGFouls
Premier League	299	26.77926	1290.378	426.5979	518	172	0.977291
La Liga	317	27.61199	1124.735	373.0547	421	118	1.237697
Bundesliga	258	26.55814	1174.938	360.9604	497	150	1.070039
Ligue 1	316	26.44304	1246.953	416.7045	558	176	1.144873
Serie A	320	26.81563	1169.459	383.3795	487	140	1.158469

There are a couple of observations to note, but I will talk about the most significant three that stood out to me and reasonings for why.

I first noticed that the number of players in the Bundesliga was significantly lower than in the other five leagues, at 258. This does make sense because the German top flight only has 18 teams compared to 20 in the other four leagues. However, this led me to see that they had the third-highest total goals despite this. This could mean two things: either the teams in the Bundesliga are very aggressive, leading to more goals, or the gap between the league's top and bottom teams is massive. I believe it is the latter from watching the league occasionally, as the same 2 or 3 teams consistently blow their competition out in high-scoring affairs.

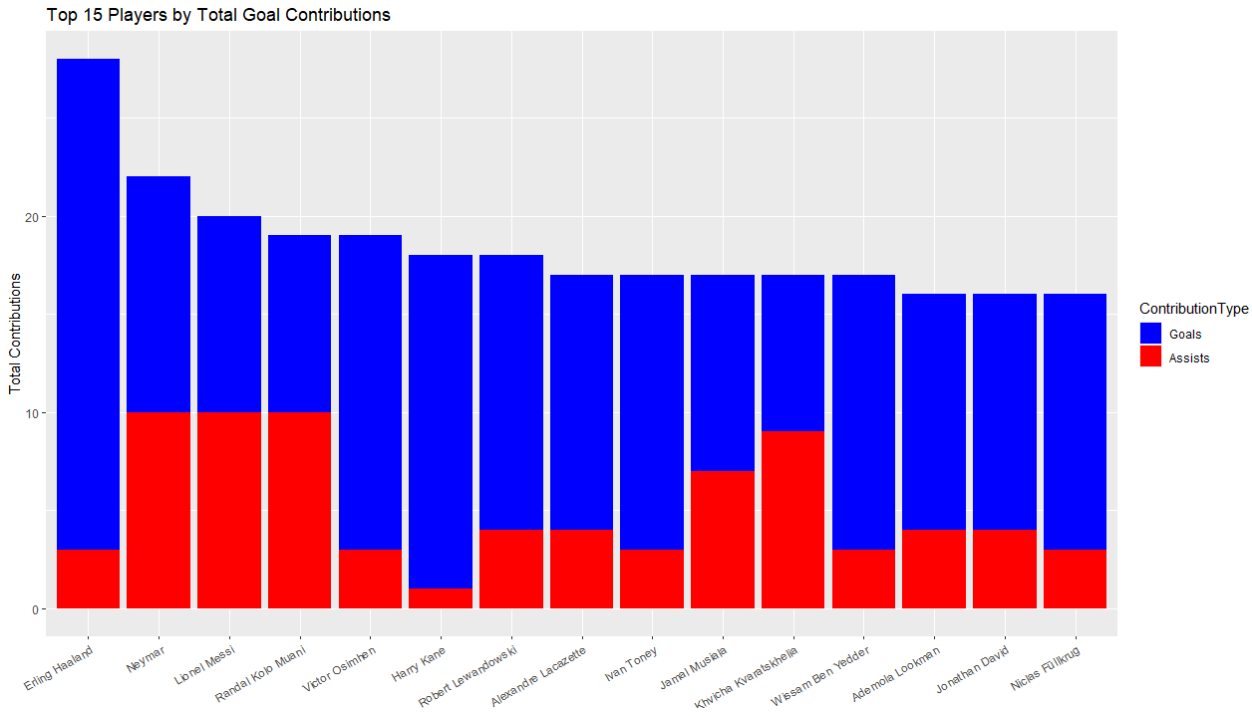
The second thing I noticed was the lack of goals in La Liga. The Spanish league only had 421 goals, over 60 behind the next lowest tally. While La Liga is in a similar situation to the Bundesliga, where only a few top teams are miles ahead of the competition, it tends to be much more competitive and tighter. This is because Spanish sides are extremely good at defending and are often very hard to break down. While this doesn't mean much against the better teams, it does mean that amongst two mid-table sides, there will be little to nothing between them, meaning very low-scoring games.

The last thing that jumped out to me was the average fouls. While looking at the data frame, it might seem like La Liga was the most physical league, but this is not necessarily the case. In sports, as games get more physical and contentious, fewer fouls get called, allowing the players to play more often. This also translates to soccer, where more physical games will frequently have fewer fouls called. This led me to conclude that despite the Spanish league being way more defensive than the others, the Premier League experienced the most physicality.

Top Goal Contributors

In soccer, the main objective is to score goals. However, assists, the passes leading to a goal, are just as important. In soccer, they combine these two counting stats into goal

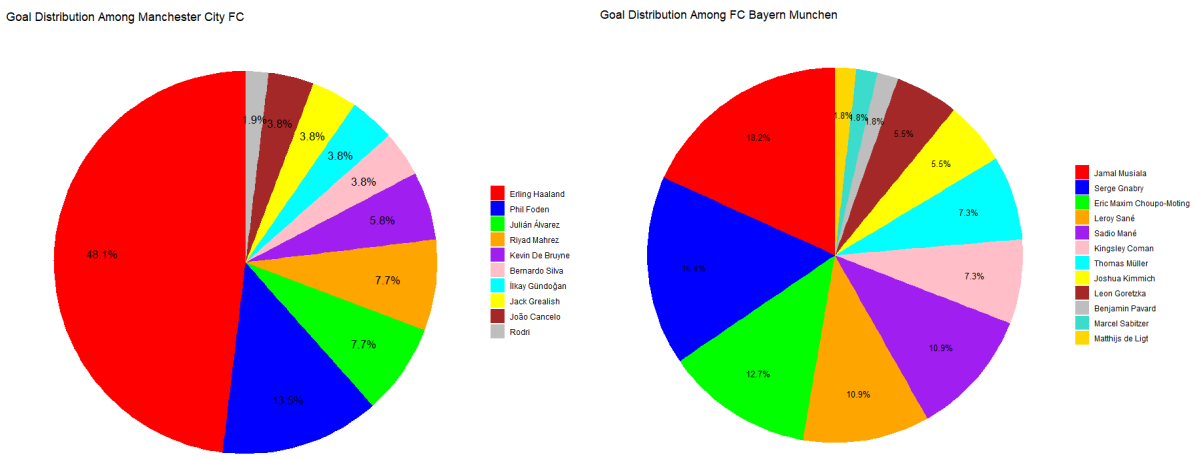
contributions, or G/A, which keeps track of how much a player is involved in their team's output. When I add every player's goals and assists together, these are the top 15 in terms of total.



Erling Haaland is far ahead of the pack, with 28 goal contributions, six ahead of Neymar in second. He is a constant outlier in goal involvements, and this base stat is no different.

Goal Distribution Among League Winners

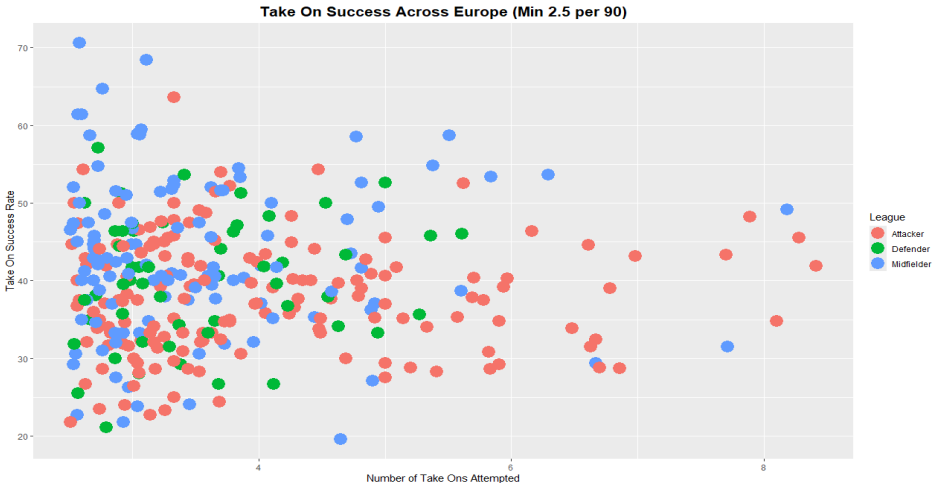
Next, I wanted to see how goals are shared between two teams who won their league. For this, I took England's and Germany's champions for that year: Manchester City and Bayern Munich. For each of these two clubs, I made a pie chart showing the percentage of goals each player scored for the two clubs, which is pictured here.



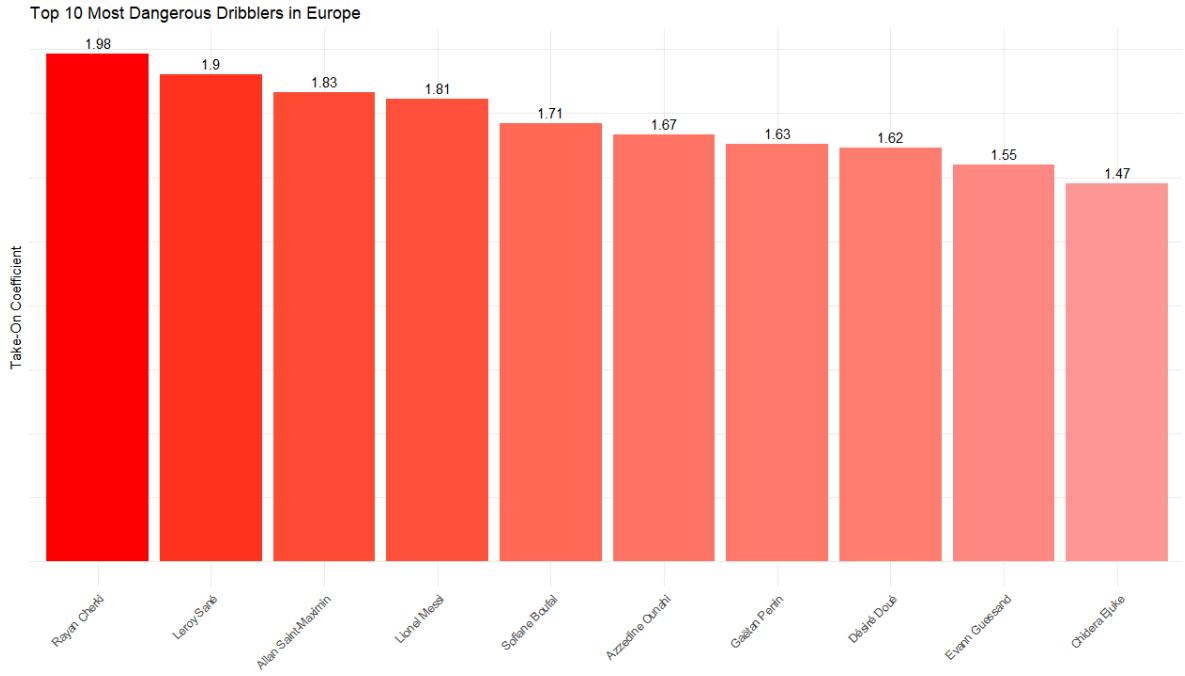
It's very easy to see the massive difference in the scoring, as with Man City, Erling Haaland takes up 48% of all goals scored by the club. In contrast, the scoring load is way more spread out with Bayern Munich, as their top scorer, Jamal Musiala, only has 18% of the goals.

Most Dangerous Dribblers in Europe

Ball carrying and beating your man is a massive part of the game of soccer and, honestly, one of the most entertaining. Seeing good dribblers slice their way through defenses is extremely pleasing to the eye, and some of the best goals ever scored stem from plays like this. I wanted to use the dataset to see the best dribblers statistically in the 2022/23 season. I first graphed a scatter plot showing every player's number of take-on attempted and their take-on success rate. It looked like this.

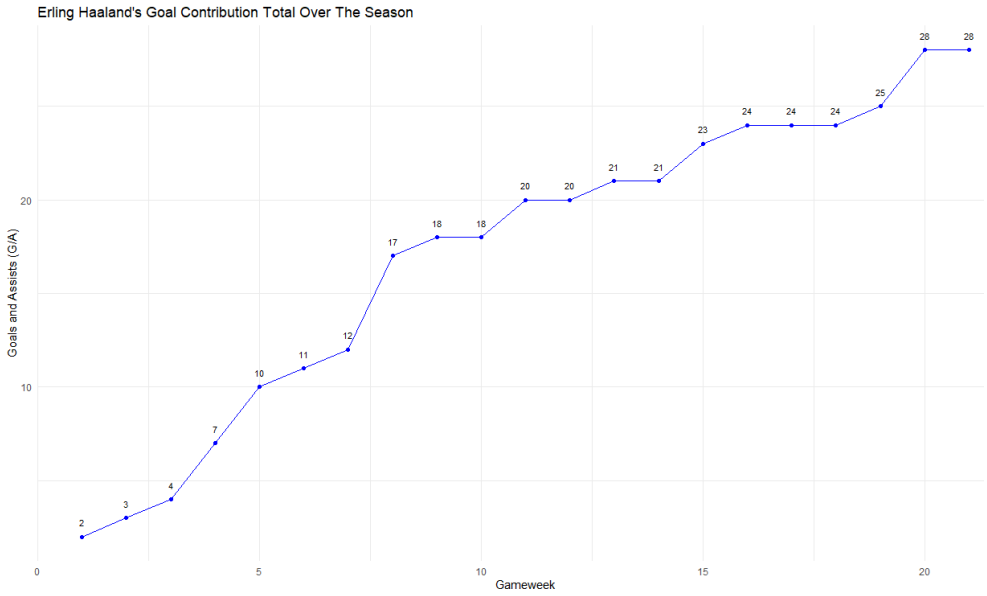


After adjusting the minimum requirements to prevent players who don't dribble often from appearing, we are still left with about 300 players, most of whom are midfielders and attackers. This makes sense because defenders and goalkeepers rarely, if ever, dribble the ball, and when I looked into the defenders that did make this cutoff, the few that did were wide defenders, meaning they go up the field more often, leading to more dribbles. The scatter plot can be interpreted in two different ways. I could have looked at the highest dribble percentages and called it a day, but everyone with high percentages has lower attempt rates. On the other hand, if I went by attempt rates, players who dribble a lot but aren't successful get rewarded. So, I gave all 331 players their own Takeon Coefficient. I multiplied their total number of completions by their success rate; that way, they not only get rewarded for completing dribbles often but for doing so at a higher rate. After doing this, I could graph the top 10 players on this coefficient to get the ten most dangerous dribblers in Europe, pictured here.



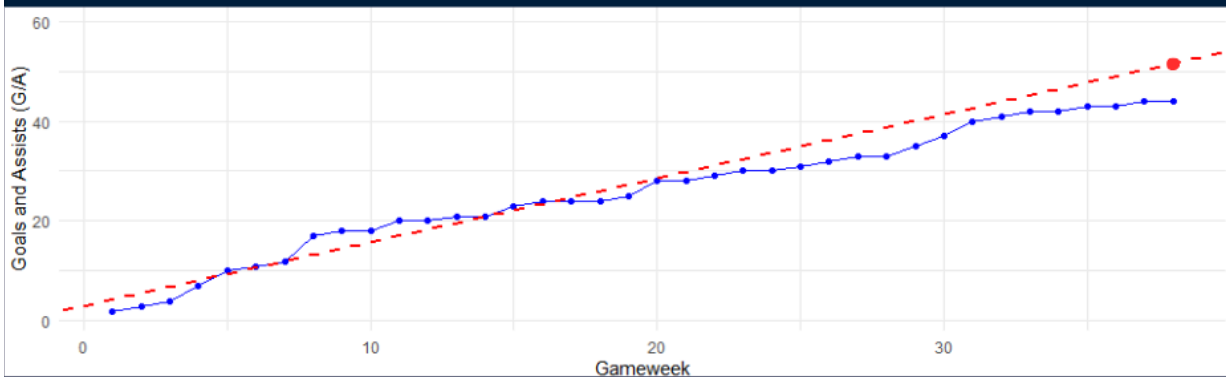
Erling Haaland Linear Regression

As mentioned earlier, this year was very fruitful for Erling Haaland. Because these stats covered only 60% of the season, I wanted to use linear regression to predict his final goal contributions. I first went to his player profile page on Transfermarkt to set it up. Since the dataset didn't track precisely when he scored, I used this site to review his game logs. I then made a data frame keeping track of the week and headlands g/a tally at each point. With this, I could plot a line graph of his start to the season.



From this, I used R's built-in linear modeling tools to fit a line to the points on this data frame. The only independent variable was the game week, so all I had to do for my predictor was plug in 38, the final number of weeks played in a domestic season. When I did, I was given back 52 goal contributions. Erling Haaland was eight off of this, finishing on 44 G/A, but that can be attributed to not only a few knocks and injuries during the back half of the season but also other competitions ramping up, meaning he has to play more than just his domestic games. Here is what the final tally looked like compared to the linear regression line we made.

```r
## predict haaland's end of season tally and add it to df for plotting
eosPredict <- predict(haalandmodel, newdata = data.frame(gameweeks = 38))
haaland[22, ] <- c(38, eosPredict)
eosPredict <- round(eosPredict)
eosPredict
```



Revised Teams of The Season

Now that I had a better sense of the data I was dealing with, it was time to recalculate the teams of the season. It was a long process, so I broke it down into a few steps. The first step was to separate all the players into separate data frames depending on their position. This gave me four datasets, and I could compare players to similar player profiles more fairly. Next, I took all the relevant numeric statistics and re-adjusted them, making them Z numbers relative to the stat's mean value rather than just the statistic itself. This would give the better players higher Z number values, and the more 'outlying' players would end up higher in the overall rankings, which is what we wanted. I then picked specific metrics depending on the position to prioritize more important things for each position class. This would prevent things like goalkeepers from being compared by their goals. Then, I gave these metrics varying coefficients based on importance and added the final scores. Lastly, I multiplied this score by a player's total of 90s played (90 minutes = one game) to favor players who played more throughout the season. This left me with a ranking that looked something like this.

	Player	Nation	Position	Team	League	Age	MFOVRRanks
761	Kevin De Bruyne	BEL	MF	Manchester City	Premier League	31	420.82287
715	Joshua Kimmich	GER	MF	Bayern Munich	Bundesliga	28	390.81003
1241	Rodri	ESP	MF	Manchester City	Premier League	26	382.38203
1071	Neymar	BRA	MFFW	Paris S-G	Ligue 1	31	381.07096
832	Lionel Messi	ARG	MFFW	Paris S-G	Ligue 1	35	380.72368
1413	Toni Kroos	GER	MF	Real Madrid	La Liga	33	350.35994
197	Branco van den Boomen	NED	MF	Toulouse	Ligue 1	27	338.52504
1177	Pierre Højbjerg	DEN	MF	Tottenham	Premier League	27	272.04693
42	Aleix García	ESP	MF	Girona	La Liga	25	253.54315
1159	Pedri	ESP	MFFW	Barcelona	La Liga	20	250.71467
207	Bruno Fernandes	POR	MFFW	Manchester Utd	Premier League	28	238.35439

The higher the score, the more of a statistical outlier the player is, meaning they are, in comparison, better than their competition. So, with this, I took the top 5 outliers in each position

class and the top goalkeeper and made a starting XI to compare to the official teams of the season. In this section, I will first explain my metrics for the four position classes, then go through the five leagues and compare and contrast the differences in team selection. Lastly, I will provide my calculated overall team of the season, where I used special weighting to adjust for the difficulty of the league the player was in.

Calculation Metrics

Attackers and midfielders, while having different jobs, benefitted a lot by having most of their stats tracked in this set. The same can not be said about defenders and goalkeepers, as much of their data was not tracked in this set. This meant that I had a lot to go off for the more attacking roles, and in turn, this would make my predictions for those players more accurate. Here are the metrics and weights I used for the four position classes.

```
######### Forward Scale #########
## 1.00 -- Goals, Goals Per Shot
## 0.85 -- Shots, Shots On Target
## 0.70 -- Assists
## 0.60 -- Shot Assists
## 0.55 -- Touches In Penalty Area, Goal Creating Actions
## 0.50 -- Shot Creating Actions
## 0.45 -- Offsides(penalize), Successful Take Ons, Take On Success Rate
## 0.40 -- Passes Into The Penalty Area, Touches In Attacking Third
## 0.35 -- Penalty Box Entries
## 0.30 -- Average Shot Distance
## 0.20 -- Fouls Drawn, Fouls Committed(penalize)
## Multiply final score by total 90s to give credit to players who play more
```

For forwards, the goal of their game is simple: who can create the most goals for their team? Making the metrics for this was very straightforward, as anything involving goal generation was included. What I prioritized more than anything was being able to finish chances, which is why the most important stats are around just pure goalscoring.

```
######### Midfielder Scale #########
## 1.00 -- Assists, Passes Completed
## 0.85 -- Shot Assists, Pass Completion PCt
## 0.70 -- Total prog Pass Distance, Carries, Passes Leading to shots
## 0.60 -- Goals, Goals per Shot, crosses
## 0.55 -- Recoveries, progressive Passes
## 0.50 -- Final Third Passes, Passes into Penalty AREA
## 0.45 -- Through Balls, Switches, shot creating Actions, Carry Dist
## 0.40 -- Touches, Aerial Duels won, Aerial duel win Rate
## 0.35 -- Fouls Drawn, Fouls Committed (penalize)
## 0.30 -- Take ons Completed, Tackles, Amt Dispossessed (penalize)
## 0.20 -- Interceptions, Passes Blocked
## Multiply final score by total 90s to give credit to players who play more
```

For midfielders, it is not only about goal creation but goal prevention on the opposite end. This was the most complete set of rules, as the midfield is typically where every player is involved. As I said prior, goal generation was taken into account a lot, but as you can see from some of the other metrics, defensive stats like interceptions and blocked passes made their way into the scales. Lastly, I considered ball progression, as midfielders take most of the load from building up play.

```
######### Defender Scale #########
## 1.00 -- Tackle Success Rate, Tackles Won
## 0.85 -- Interceptions, Clearances, Errors(penalize)
## 0.70 -- Shots Blocked, Pass Completion Pct
## 0.60 -- Recoveries, Dribbled Past (Penalize)
## 0.55 -- Fouls Committed (penalize), Passes Blocked
## 0.50 -- Touches, Long Passes Completed, Long Pass Pct
## 0.45 -- Aerial Duels Won, Aerial Duel Win Rate
## 0.40 -- Progressive Carries, Prog Carry Dist
## 0.35 -- Passes Completed, Progressive Passes
## 0.30 -- Final Third Entries
## 0.20 -- Goals, Assists, Shot Assists
## Multiply final score by total 90s to give credit to players who play more
```

Defenders are all about goal prevention, so I took any stat that could be taken as defensive. This wasn't a lot, and I still factored in some attacking metrics, but in the future, I would want to find more stats such as maybe clean sheets, defensive actions per 90, or expected goal against, as

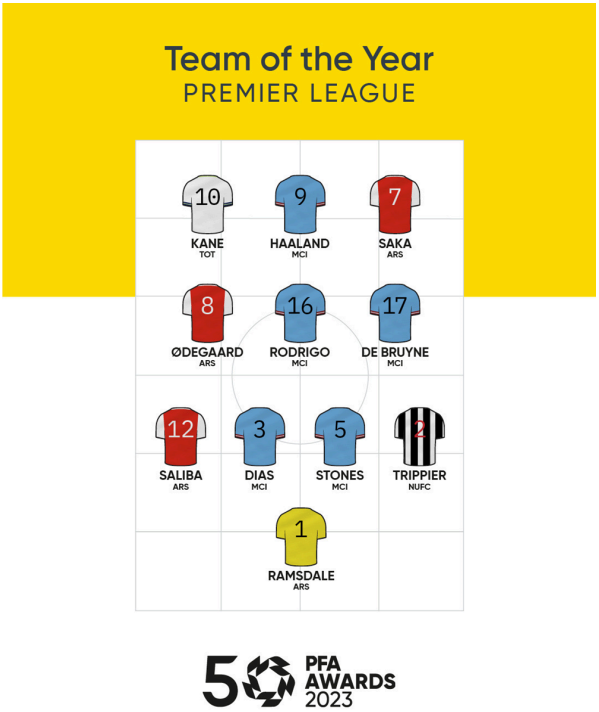
these statistics would better show how much of an impact a defender has on the game. Like I said, however, the ones I used have some defending aspects, but many of the stats available were based on how good they are at ball distribution, like Long Passes.

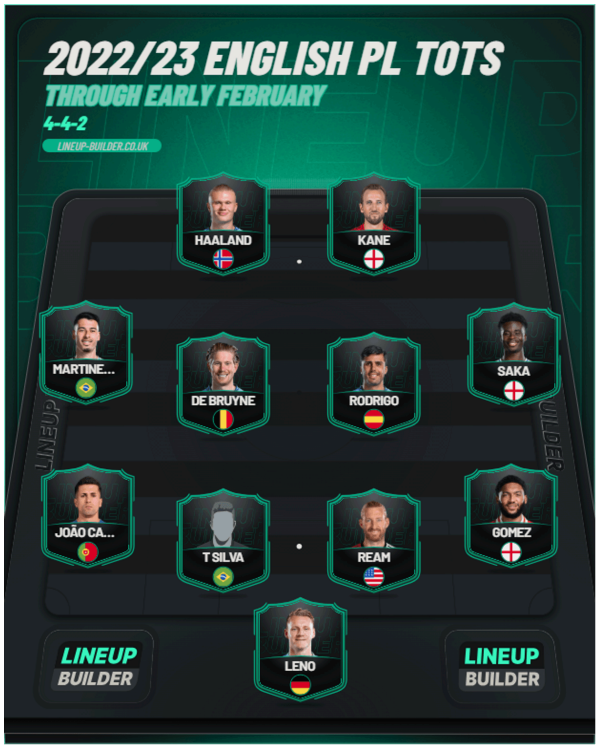
```
######### Goalkeepeers Scale #########
## 1.00 -- Errors (penalize)
## 0.85 -- Clearances, Recoveries
## 0.70 -- Aerial Duels Won, Aerial Duel Win Rate
## 0.60 -- Fouls (penalize)
## 0.50 -- Pass Completion Pct, Touches
## 0.40 -- Passes Completed, Long Pass Completion Pct
## Multiply final score by total 90s to give credit to players who play more
## No save/goal prevention data so not a lot I can use
```

Goalkeeper stats were even more sparse, as there was no save data. So, instead, I had to see which keepers are the least error-prone, penalizing errors and fouls more than every other position. The rest of the metrics were more muted defending stats, but in the future, I want to track things like expected goals prevented, number of shots faced, and standard counting stats like saves and ball claims.

The next part will be the actual teams I calculated. In each header, I listed their country and their UEFA League Coefficient, which ranks leagues based on their difficulty. I decided to use this as a multiplier for the overall team of the season, which we will get to at the end. My calculated team will be on the right, with the official team on the left.

Premier League Team of the Season (England: 1.04785)





I went 5/11 in England, with Harry Kane, Erling Haaland, Bukayo Saka, Kevin De Bruyne, and Rodrigo all featuring on both teams. One interesting thing I noticed was that the left-back, Joao Cancelo, didn't end the season in the league, as he transferred to Bayern Munich in Germany halfway through the season. This meant his start to the season was still better than every other defender, bar 3.

La Liga Team of the Season (Spain: 0.96499)

I have four players matching in La Liga: Vinicius Junior, Robert Lewandowski, Antoine Griezmann, and Marc Andre Ter Stegen. It was interesting to note that despite all the metric issues I had with goalkeepers, I still picked correctly compared to the general consensus for the year.

Bundesliga Team of the Season (Germany: 0.91241)

The Bundesliga was another 4/11 with me, as I correctly got Randall Kolo Muani, Jamal Musiala, Jude Bellingham, and Nico Schlotterbeck. This year, the teams in first and second place, Bayern Munich and Borussia Dortmund, finished with the same amount of points (71). My calculation reflected shared dominance, as 7 of the 11 featured players were on one of the two teams.

Serie A Team of the Season (Italy: 0.90963)

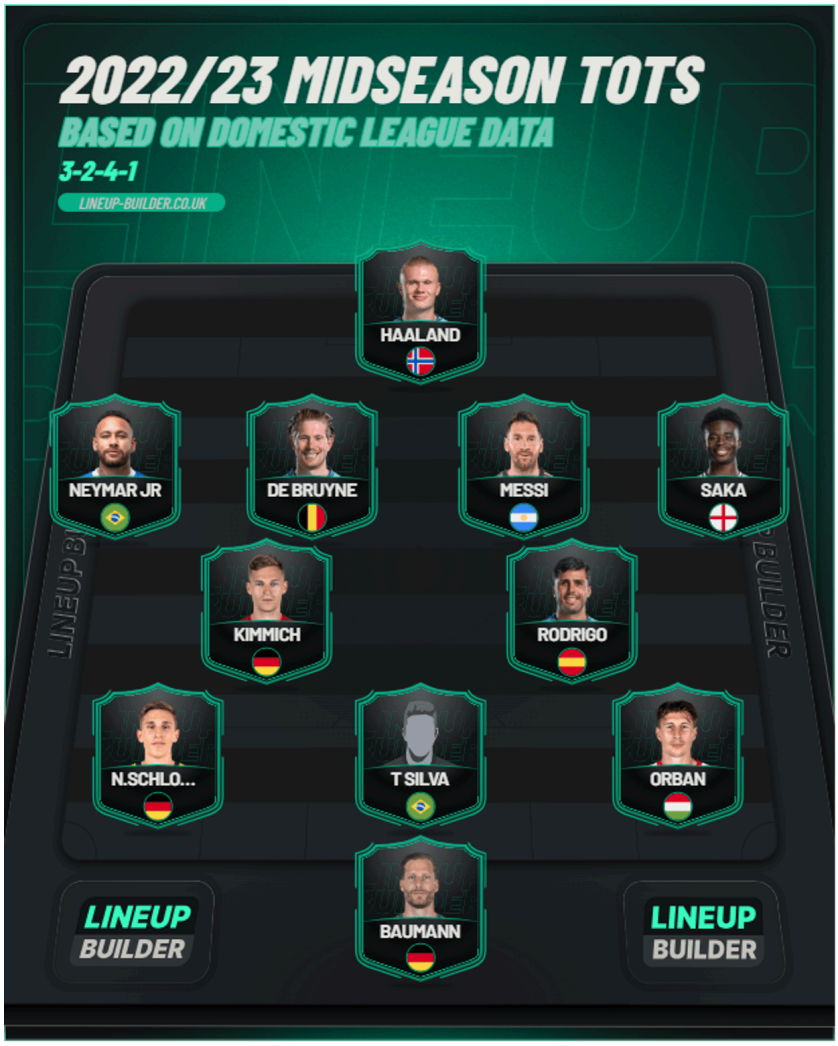
I scored another 4/11 in Serie A, correctly selecting Rafael Leao, Kvicha Kvaratskhelia, Nico Barella, and Kim Min Jae. One cool observation I noted was that the player I selected as center forward, Ademola Lookman, was of the same nationality as the center forward for the official team, Victor Osimhen. There were only 5 Nigerians in the league that year, so it was nice to see.

Ligue 1 Team of the Season (France: 0.80582)

Lastly, I also got four correct in the French League. They were Alexandre Lacazette, Kylian Mbappe, Lionel Messi, and Branco Van Den Boomen. Van Den Boomen's club only managed 13th in the league that season, so it's cool to see that both my calculations and the official teams were able to point him out as a highlight despite the unremarkable results from his team.

Overall Team of the Season

Once again, I used the UEFA club coefficients to fairly rate the players based on the difficulty of their league, and the above result was that. The highest-rated players by position class were Lionel Messi (FW, 305), Kevin De Bruyne (MF, 421), Willi Orban (DF, 181), and Oliver Baumann (GK, 105).



There were three things I noticed from this overall grouping. The first was that Neymar and Messi were rated exceptionally highly in both forward and midfield rankings. They placed 3rd and 1st among all forwards and 4th and 5th among all midfielders. They were in both datasets because their position was listed as 'MFFW' (both midfield and forward), but it is cool to see they still dominated both sets of testing metrics. The second thing I noticed was that there were zero La Liga players in the final eleven, with the closest being Toni Kroos, ranking 6th among all midfielders. The last thing that stuck out to me was regarding Lionel Messi, as he won the Ballon d'Or that year, meaning not only was he voted the best player in the world, but according to this data, he was statistically the best player in the world.

Conclusion

I would do a couple of things differently if I were given the chance and more time. Like I said multiple times throughout this report, there were not enough statistics to fairly rate the defenders and goalkeepers, and it showed, as throughout all five leagues, I only got one defender and one goalkeeper right. I also only had 60% of the data, so the back 40%, the most crucial part of the season, was not a part of the data. Another issue was that there was not a lot of flexibility with positions. If given more time, I would manually put in every player's position, and then the metrics for each position would be more catered to the actual position rather than the class. Lastly, I was interested in seeing if the data could predict league winners. This would take a lot more time as I would have to combine team data, player data, and matchups, but it could be an exciting extension to this project with the right tools.