Random forest EPL predictions using FM24 attributes

November 24, 2025

Contents

Abstract

Descriptive part

For the 24/25 English Premier League (EPL) season, player attribute data (e.g. speed, passing, shooting,etc) gathered from the simulation game "Football Manager 24" are descriptively analyzed with standard methods (PCA).

Predictive part

Firstly, after a brief revision of the theory of linear regression, we present the "ANOVA" method through the lens of this general theory. We then apply ANOVA to the simplest case possible (two categories) and answer the question "*are the top teams physically stronger than the bottom teams ?*"

Secondly, with the PCA method of the descriptive part, we reduce the 35 attribute variables to about 7 principal components (PC), keeping about 70% total inertia. For each team, for each PC, we compute the effective PC as the average of all the player PC in that team. That defines our predictors $X_1, \ldots, X_7$.

From the site "FBref" we gather data on the outcomes (win-draw-lose ; home xG, away xG) of the past 325 EPL games played this season. Using then the PCA reduced variables as predictors (14 of them, the X_i for home+away teams) we train/tune a *random forest* model using the built-in o-o-b error: (*i*) first as a classifier giving the win-draw-lose probabilities, (*ii*) second as a regressor giving "total expected goals" as the target.

We "test" the two models on the remaining EPL games (50 or so) ; then we examine the predictions - they are deemed very reasonable. In particular, variable importance (using o-o-b error) is mentioned and the "strongest/weakest" variables are identified and discussed.

Gambling part

We treat our model's win-draw-lose probabilities are *true* probabilities p and identify *favorable* sport casino bets say $\tilde{X}_j$, favorable in the sense that the expected value $Pp > 1$ with P the odds. We gamble our starting wealth S_0 multiplicatively and sequentially at each $t = 1, \ldots, m$ on $Y_t(\boldsymbol{b}) = \boldsymbol{b}'\tilde{\boldsymbol{X}}_{t,j}$ - weighted sums of the favorable casino bets $\tilde{\boldsymbol{X}}_j$. The optimal betting fractions $\boldsymbol{b}_t^*$ are provided by Algoet's theorem , they are the fractions that maximize $E[\log Y_t(\boldsymbol{b})]$.

Can we make a profit in m rounds ?

Appendix - SAS studio

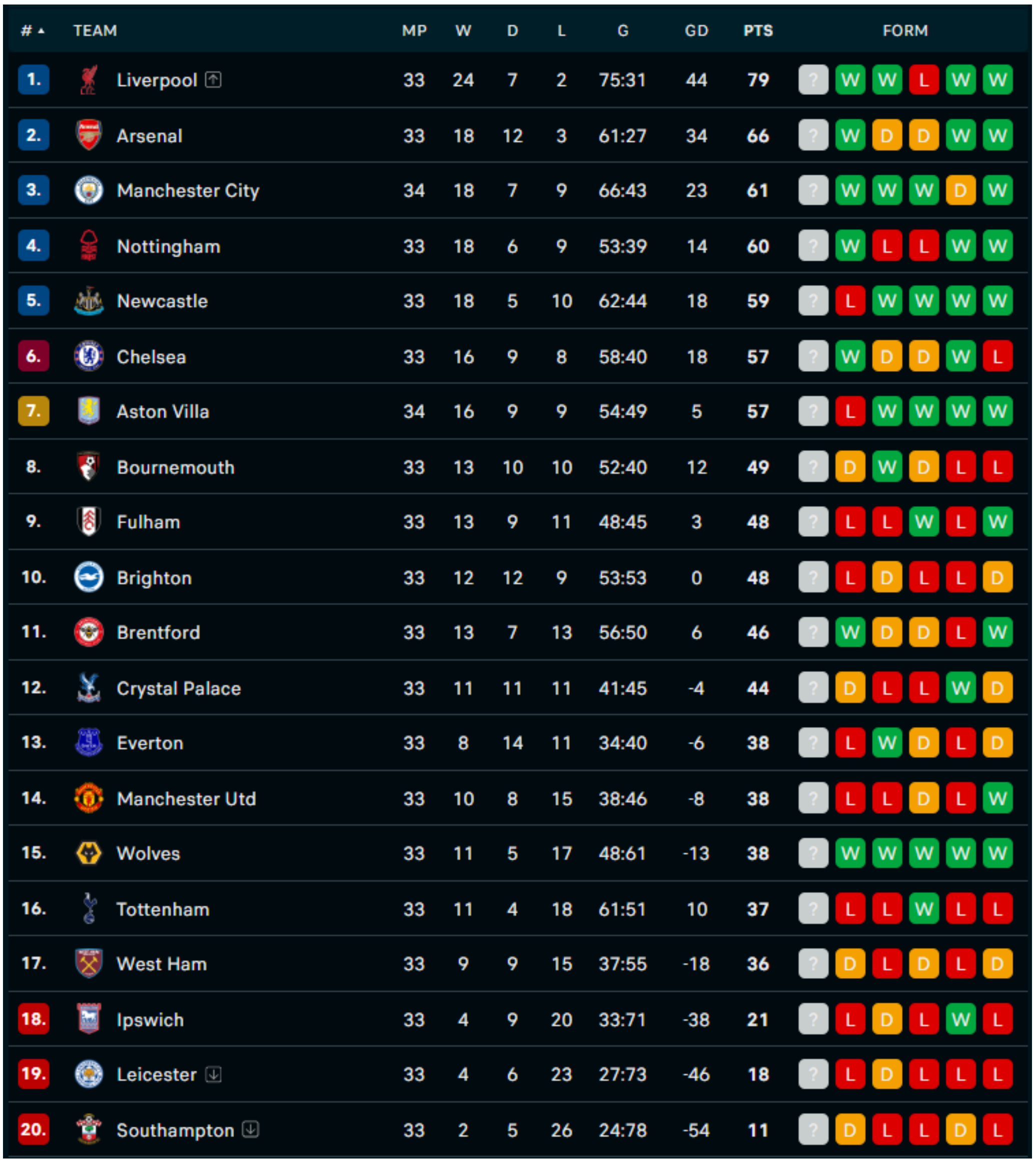
We present some very nice "bubble graphs" made with the UI in SAS studio (using the free student license).The goal is to present the maximum amount of data while keeping a reasonable readability. We also present a hypothetical process flow that would trivialize all the "manual" work done in R studio. The process is hypothetical in the sense that it would require a full SAS license to do. Data editing and random forest "tasks" are locked behind a paywall...

Introduction

The English Premier League (EPL)

The EPL is a double round robin tournament played in between 20 different teams. That gives a total of $M = 380$ matches played. So far (as of 21st April) 325 matches have been played , so 55 matches are remaining. Each team has about $K = 25$ players.

The above introduction is sufficient for our purposes, more info can be found anywhere else online.

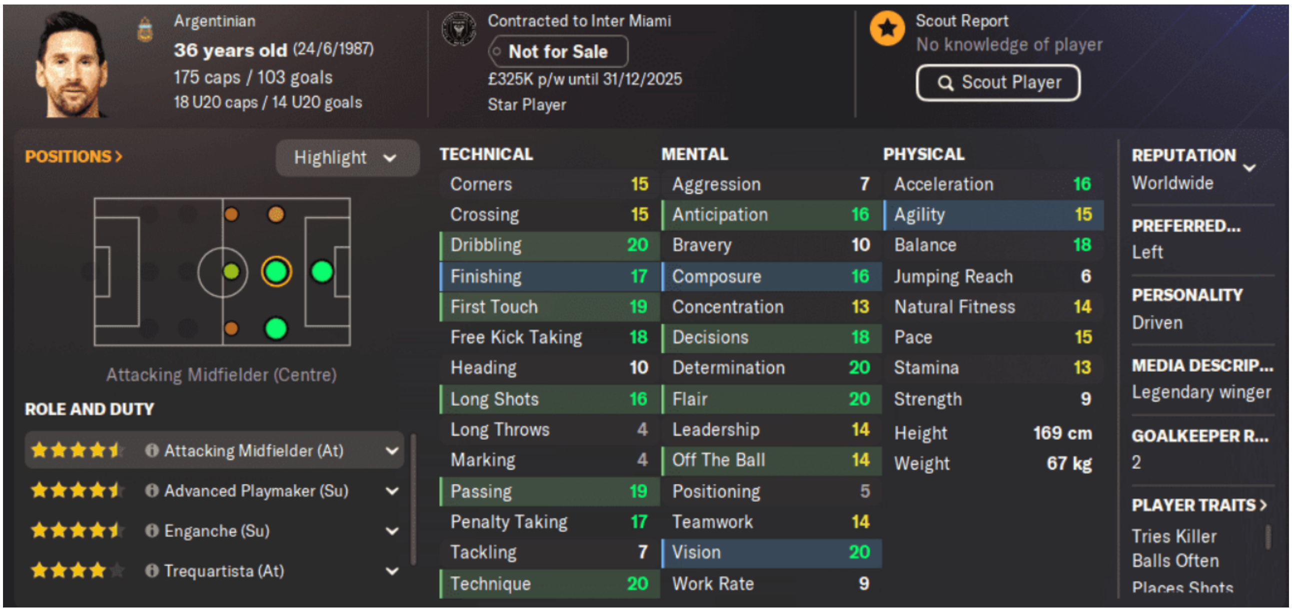


# ▴	TEAM	MP	W	D	L	G	GD	PTS	FORM					
1.	Liverpool ⇧	33	24	7	2	75:31	44	79	?	W	W	L	W	W
2.	Arsenal	33	18	12	3	61:27	34	66	?	W	D	D	W	W
3.	Manchester City	34	18	7	9	66:43	23	61	?	W	W	W	D	W
4.	Nottingham	33	18	6	9	53:39	14	60	?	W	L	L	W	W
5.	Newcastle	33	18	5	10	62:44	18	59	?	L	W	W	W	W
6.	Chelsea	33	16	9	8	58:40	18	57	?	W	D	D	W	L
7.	Aston Villa	34	16	9	9	54:49	5	57	?	L	W	W	W	W
8.	Bournemouth	33	13	10	10	52:40	12	49	?	D	W	D	L	L
9.	Fulham	33	13	9	11	48:45	3	48	?	L	L	W	L	W
10.	Brighton	33	12	12	9	53:53	0	48	?	L	D	L	L	D
11.	Brentford	33	13	7	13	56:50	6	46	?	W	D	D	L	W
12.	Crystal Palace	33	11	11	11	41:45	-4	44	?	D	L	L	W	D
13.	Everton	33	8	14	11	34:40	-6	38	?	L	W	D	L	D
14.	Manchester Utd	33	10	8	15	38:46	-8	38	?	L	L	D	L	W
15.	Wolves	33	11	5	17	48:61	-13	38	?	W	W	W	W	W
16.	Tottenham	33	11	4	18	61:51	10	37	?	L	L	W	L	L
17.	West Ham	33	9	9	15	37:55	-18	36	?	D	L	D	L	D
18.	Ipswich	33	4	9	20	33:71	-38	21	?	L	D	L	W	L
19.	Leicester ⇩	33	4	6	23	27:73	-46	18	?	L	D	L	L	L
20.	Southampton ⇩	33	2	5	26	24:78	-54	11	?	D	L	L	D	L

Figure 1: The league table (April 23rd). The names of the 20 clubs can be seen. Screenshot taken from Premier League Standings – Flashscore.

Football manager 24

Football manager 24 (FM24) is a management simulation game where the player takes on the role of the manager of a club (among other things). Important is to note that all the major leagues in the world are included in the game, in particular the EPL. To every player are associated attributes which reflect their individual strengths/weaknesses (cf. fig below).



Argentinian
36 years old (24/6/1987)
175 caps / 103 goals
18 U20 caps / 14 U20 goals

Contracted to Inter Miami
Not for Sale
£325K p/w until 31/12/2025
Star Player

Scout Report
No knowledge of player

Attacking Midfielder (Centre)

ROLE AND DUTY

- Attacking Midfielder (At)
- Advanced Playmaker (Su)
- Enganche (Su)
- Trequartista (At)

TECHNICAL		MENTAL		PHYSICAL	
Corners	15	Aggression	7	Acceleration	16
Crossing	15	Anticipation	16	Agility	15
Dribbling	20	Bravery	10	Balance	18
Finishing	17	Composure	16	Jumping Reach	6
First Touch	19	Concentration	13	Natural Fitness	14
Free Kick Taking	18	Decisions	18	Pace	15
Heading	10	Determination	20	Stamina	13
Long Shots	16	Flair	20	Strength	9
Long Throws	4	Leadership	14	Height	169 cm
Marking	4	Off The Ball	14	Weight	67 kg
Passing	19	Positioning	5		
Penalty Taking	17	Teamwork	14		
Tackling	7	Vision	20		
Technique	20	Work Rate	9		

REPUTATION
Worldwide

PREFERRED...
Left

PERSONALITY
Driven

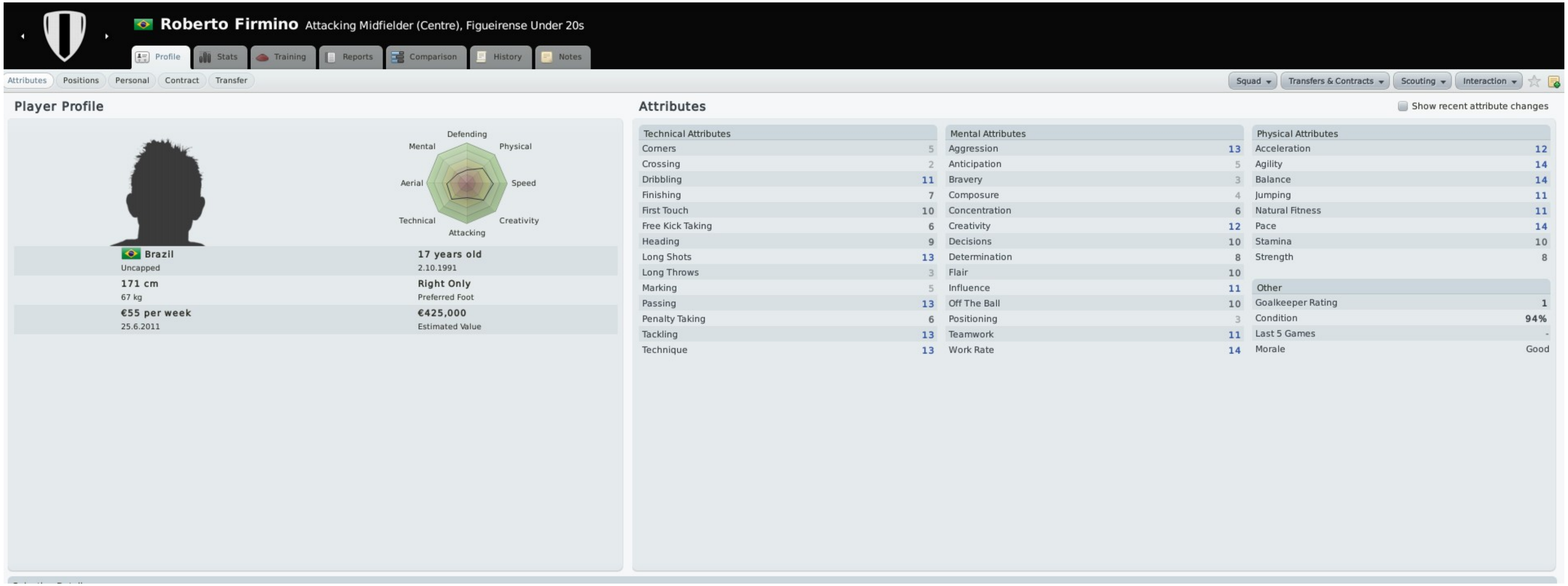
MEDIA DESCRIP...
Legendary winger

GOALKEEPER R...
2

PLAYER TRAITS >
Tries Killer Balls Often
Places Shots

Figure 2: Different player attributes ranging from $1-20$. They are categorized as "Technical, Mental , Physical". The other stats (height, weight, age, etc.) we did not consider. Taken from an in-game screenshot (FM24).

These attributes are not randomly chosen. Sports Interactive (SI), the company that develops Football Manager 2024 (FM24), follows a strict procedure to determine them. This procedure can be found here. In short, certain individuals, after passing an "admission" test, are allowed to provide subjective assessments of player attributes to SI. SI then uses these assessments to calculate the final player attributes.



Roberto Firmino Attacking Midfielder (Centre), Figueirense Under 20s

Player Profile

Brazil Uncapped	**17 years old** 2.10.1991
171 cm 67 kg	**Right Only** Preferred Foot
€55 per week 25.6.2011	**€425,000** Estimated Value

Attributes

□ Show recent attribute changes

Technical Attributes		Mental Attributes		Physical Attributes	
Corners	5	Aggression	13	Acceleration	12
Crossing	2	Anticipation	5	Agility	14
Dribbling	11	Bravery	3	Balance	14
Finishing	7	Composure	4	Jumping	11
First Touch	10	Concentration	6	Natural Fitness	11
Free Kick Taking	6	Creativity	12	Pace	14
Heading	9	Decisions	10	Stamina	10
Long Shots	13	Determination	8	Strength	8
Long Throws	3	Flair	10		
Marking	5	Influence	11	**Other**	
Passing	13	Off The Ball	10	Goalkeeper Rating	1
Penalty Taking	6	Positioning	3	Condition	94%
Tackling	13	Teamwork	11	Last 5 Games	-
Technique	13	Work Rate	14	Morale	Good

Figure 3: It's worth mentioning a popular myth often cited by fans: the case of Roberto Firmino. According to the story, Hoffenheim's scouts discovered Firmino, then an unknown player in Brazil's second division, using data from FM10. The narrative claims that "testers" correctly identified his potential and relayed the information to Sports Interactive. This myth has been debunked by Lutz Pfannenstiel the actual scout who brought Firmino to Hoffenheim, who stated he had never played or even heard of Football Manager at the time. Firmino's talent was eventually recognized by Liverpool, who signed him, and he went on to achieve remarkable success with the club, including winning the UEFA Champions League. The figure is the FM10 in-game screenshot of Roberto Firmino's attributes. We remark that the game got his height wrong - he is actually 181cm tall !

The data (a) - FM24 player attributes

The data involves $A = 36$ variables, attributes ranging from 1 to 20. The only way to get the data is to directly export it from the game FM24, albeit an exportation en masse, i.e. for every single team in the game at once, is impossible. One needs to load the game as the manager of each different team, and then export data; so a procedure of $L = 20$ steps each taking about 10min. This bottleneck effectively prevents us from having a bigger dataset, a data set of the top 5 european leagues say, as was originally intended.

Note 1: After completing the procedure L times , we realized that the attribute "Marking" was missing, so our data actually has $B = 35$ attributes instead of $A = 36$; we are not too worried about that.

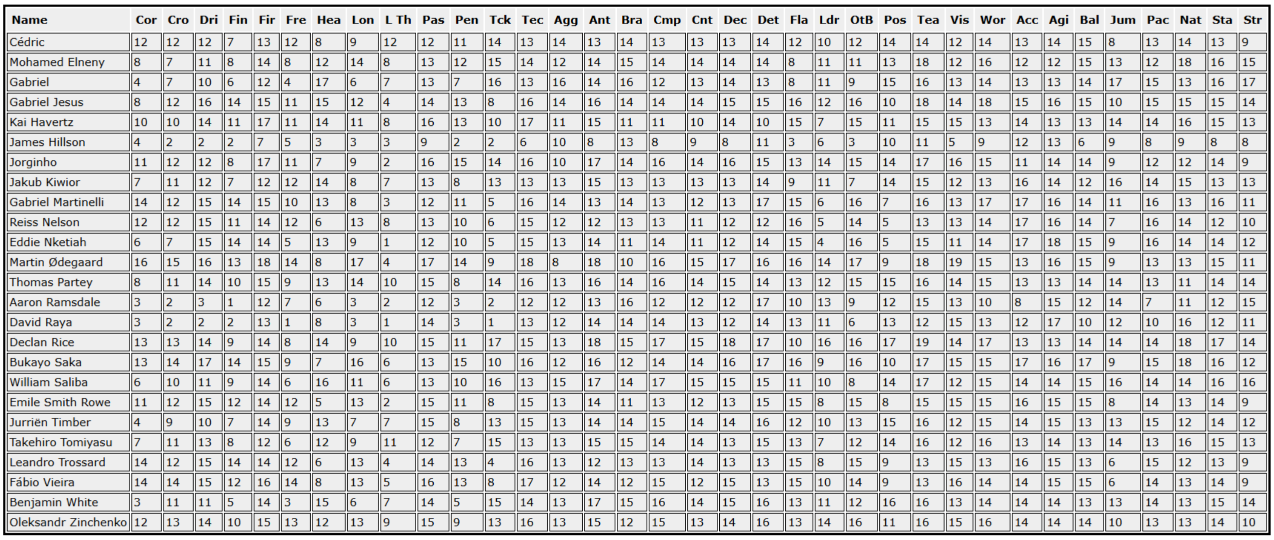
The 35 attributes are split into three categories:

Technical (T)	Mental (M)	Physical (P)
Corners	Aggression	Acceleration
Crossing	Anticipation	Agility
Dribbling	Bravery	Balance
Finishing	Composure	Jumping Reach
First Touch	Concentration	Natural Fitness
Free Kick Taking	Decisions	Pace
Heading	Determination	Stamina
Long Shots	Flair	Strength
Long Throws	Leadership	
Passing	Off The Ball	
Penalty Taking	Positioning	
Tackling	Teamwork	
Technique	Vision	
	Work Rate	

(see Figure 2 for example)

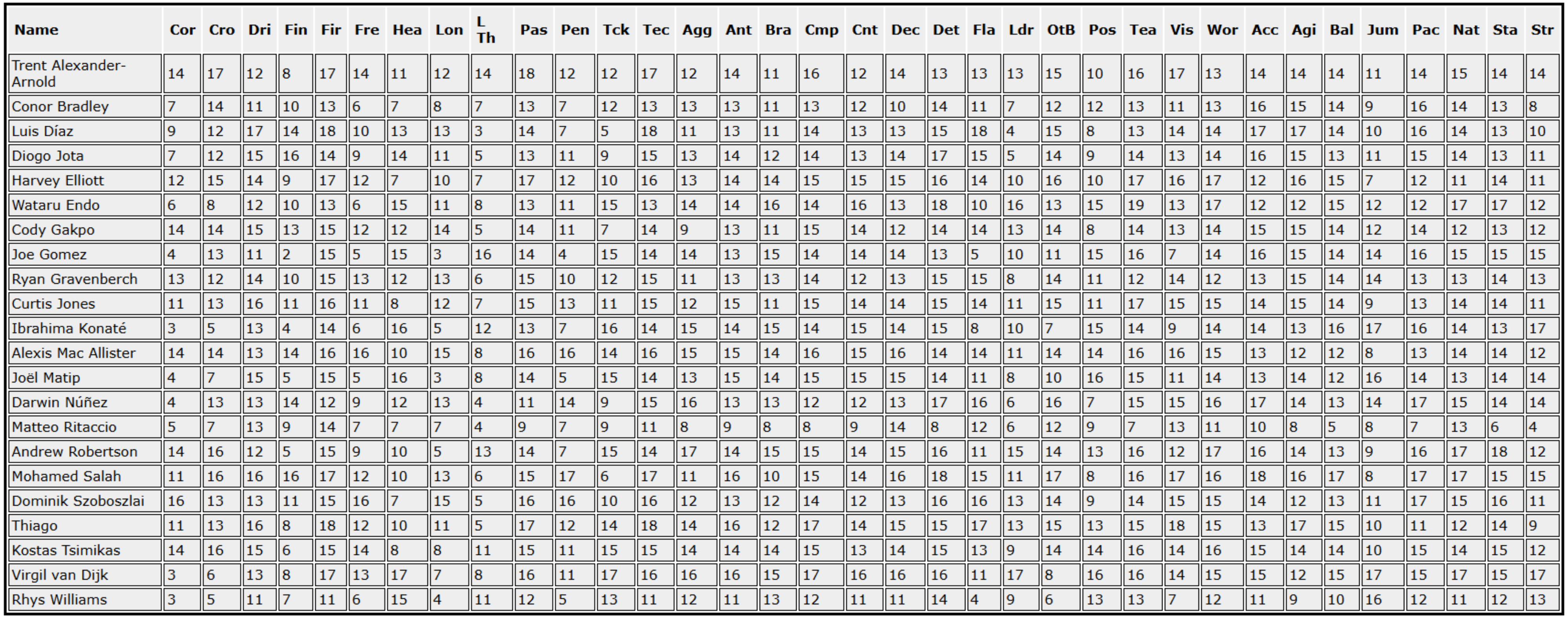
Note 2: The exact meaning of each attribute is fairly self explanatory, we will not bother to define them precisely, for the definitions see here.

Our final data set (a) (ATTRIBUTES.csv) has 436 players (goal keepers excluded!), their team name and 35 of their attributes as columns.



Name	Cor	Cro	Dri	Fin	Fir	Fre	Hea	Lon	L Th	Pas	Pen	Tck	Tec	Agg	Ant	Bra	Cmp	Cnt	Dec	Det	Fla	Ldr	OtB	Pos	Tea	Vis	Wor	Acc	Agi	Bal	Jum	Pac	Nat	Sta	Str
Cédric	12	12	12	7	13	12	8	9	12	12	11	14	13	14	13	14	13	13	13	14	12	10	12	14	14	12	14	13	14	15	8	13	14	13	9
Mohamed Elneny	8	7	11	8	14	8	12	14	8	13	12	15	14	12	14	15	14	14	14	14	8	11	11	13	18	12	16	12	12	15	13	12	18	16	15
Gabriel	4	7	10	6	12	4	17	6	7	13	7	16	13	16	14	16	12	13	14	13	8	11	9	15	16	13	14	13	13	14	17	15	13	16	17
Gabriel Jesus	8	12	16	14	15	11	15	12	4	14	13	8	16	14	16	14	14	14	15	15	16	12	16	10	18	14	18	15	16	15	10	15	15	15	14
Kai Havertz	10	10	14	11	17	11	14	11	8	16	13	10	17	11	15	11	11	10	14	10	15	7	15	11	15	15	13	14	13	13	14	14	16	15	13
James Hillson	4	2	2	2	7	5	3	3	3	9	2	2	6	10	8	13	8	9	8	11	3	6	3	10	11	5	9	12	13	6	9	8	9	8	8
Jorginho	11	12	12	8	17	11	7	9	2	16	15	14	16	10	17	14	16	14	16	15	13	14	15	14	17	16	15	11	14	14	9	12	12	14	9
Jakub Kiwior	7	11	12	7	12	12	14	8	7	13	8	13	13	13	15	13	13	13	13	14	9	11	7	14	15	12	13	16	14	12	16	14	15	13	13
Gabriel Martinelli	14	12	15	14	15	10	13	8	3	12	11	5	16	14	13	14	13	12	13	17	15	6	16	7	16	13	17	17	16	14	11	16	13	16	11
Reiss Nelson	12	12	15	11	14	12	6	13	8	13	10	6	15	12	12	13	13	11	12	12	16	5	14	5	13	13	14	17	16	14	7	16	14	12	10
Eddie Nketiah	6	7	15	14	14	5	13	9	1	12	10	5	15	13	14	11	14	11	12	14	15	4	16	5	15	11	14	17	18	15	9	16	14	14	12
Martin Ødegaard	16	15	16	13	18	14	8	17	4	17	14	9	18	8	18	10	16	15	17	16	16	14	17	9	18	19	15	13	16	15	9	13	13	15	11
Thomas Partey	8	11	14	10	15	9	13	14	10	15	8	14	16	13	16	14	16	14	15	14	13	12	15	15	16	14	15	13	13	14	14	13	11	14	14
Aaron Ramsdale	3	2	3	1	12	7	6	3	2	12	3	2	12	12	13	16	12	12	12	17	10	13	9	12	15	13	10	8	15	12	14	7	11	12	15
David Raya	3	2	2	2	13	1	8	3	1	14	3	1	13	12	14	14	14	13	12	14	13	11	6	13	12	15	13	12	17	10	12	10	16	12	11
Declan Rice	13	13	14	9	14	8	14	9	10	15	11	17	15	13	18	15	17	15	18	17	10	16	16	17	19	14	17	13	13	14	14	14	18	17	14
Bukayo Saka	13	14	17	14	15	9	7	16	6	13	15	10	16	12	16	12	14	14	16	17	16	9	16	10	17	15	15	17	16	17	9	15	18	16	12
William Saliba	6	10	11	9	14	6	16	11	6	13	10	16	13	15	17	14	17	15	15	15	11	10	8	14	17	12	15	14	14	15	16	14	14	16	16
Emile Smith Rowe	11	12	15	12	14	12	5	13	2	15	11	8	15	13	14	11	13	12	13	15	15	8	15	8	15	15	15	16	15	15	8	14	13	14	9
Jurriën Timber	4	9	10	7	14	9	13	7	7	15	8	13	15	13	14	14	15	14	14	16	12	10	13	15	16	12	15	14	15	13	13	15	12	14	12
Takehiro Tomiyasu	7	11	13	8	12	6	12	9	11	12	7	15	13	13	15	15	14	14	13	15	13	7	12	14	16	12	16	13	14	13	14	13	16	15	13
Leandro Trossard	14	12	15	14	14	12	6	13	4	14	13	4	16	13	12	13	13	14	13	13	15	8	15	9	13	15	13	16	15	13	6	15	12	13	9
Fábio Vieira	14	14	15	12	16	14	8	13	5	16	13	8	17	12	14	12	15	12	15	13	15	10	14	9	13	16	14	14	15	15	6	14	13	14	9
Benjamin White	3	11	11	5	14	3	15	6	7	14	5	15	14	13	17	15	16	14	15	16	13	11	12	16	16	13	14	14	14	13	13	14	13	15	14
Oleksandr Zinchenko	12	13	14	10	15	13	12	13	9	15	9	13	16	13	15	12	15	13	14	16	13	14	16	11	16	15	16	14	14	14	10	13	13	14	10

Figure 4: In game we define "attribute filters" and export data. This table is the result of the procedure done on the team "Arsenal". We remark that indeed the attribute "Marking" is missing. Also important is that we do not include goal keepers in this data, since they have their separate attributes.



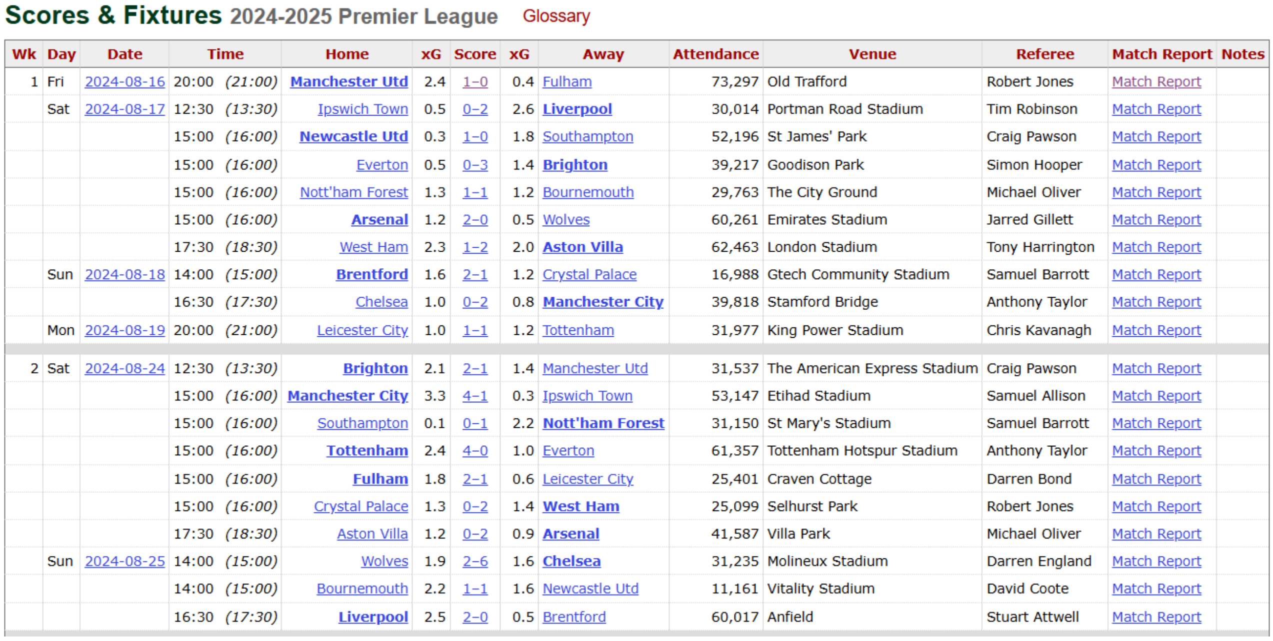
Name	Cor	Cro	Dri	Fin	Fir	Fre	Hea	Lon	L Th	Pas	Pen	Tck	Tec	Agg	Ant	Bra	Cmp	Cnt	Dec	Det	Fla	Ldr	OtB	Pos	Tea	Vis	Wor	Acc	Agi	Bal	Jum	Pac	Nat	Sta	Str
Trent Alexander-Arnold	14	17	12	8	17	14	11	12	14	18	12	12	17	12	14	11	16	12	14	13	13	13	15	10	16	17	13	14	14	14	11	14	15	14	14
Conor Bradley	7	14	11	10	13	6	7	8	7	13	7	12	13	13	13	11	13	12	10	14	11	7	12	12	13	11	13	16	15	14	9	16	14	13	8
Luis Díaz	9	12	17	14	18	10	13	13	3	14	7	5	18	11	13	11	14	13	13	15	18	4	15	8	13	14	14	17	17	14	10	16	14	13	10
Diogo Jota	7	12	15	16	14	9	14	11	5	13	11	9	15	13	14	12	14	13	14	17	15	5	14	9	14	13	14	16	15	13	11	15	14	13	11
Harvey Elliott	12	15	14	9	17	12	7	10	7	17	12	10	16	13	14	14	15	15	15	16	14	10	16	10	17	16	17	12	16	15	7	12	11	14	11
Wataru Endo	6	8	12	10	13	6	15	11	8	13	11	15	13	14	14	16	14	16	13	18	10	16	13	15	19	13	17	12	12	15	12	12	17	17	12
Cody Gakpo	14	14	15	13	15	12	12	14	5	14	11	7	14	9	13	11	15	14	12	14	14	13	14	8	14	13	14	15	15	14	12	14	12	13	12
Joe Gomez	4	13	11	2	15	5	15	3	16	14	4	15	14	14	13	15	14	14	14	13	5	10	11	15	16	7	14	16	15	14	14	16	15	15	15
Ryan Gravenberch	13	12	14	10	15	13	12	13	6	15	10	12	15	11	13	13	14	12	13	15	15	8	14	11	12	14	12	13	15	14	14	13	13	14	13
Curtis Jones	11	13	16	11	16	11	8	12	7	15	13	11	15	12	15	11	15	14	14	15	14	11	15	11	17	15	15	14	15	14	9	13	14	14	11
Ibrahima Konaté	3	5	13	4	14	6	16	5	12	13	7	16	14	15	14	15	14	15	14	15	8	10	7	15	14	9	14	14	13	16	17	16	14	13	17
Alexis Mac Allister	14	14	13	14	16	16	10	15	8	16	16	14	16	15	15	14	16	15	16	14	14	11	14	14	16	16	15	13	12	12	8	13	14	14	12
Joël Matip	4	7	15	5	15	5	16	3	8	14	5	15	14	13	15	14	15	15	15	14	11	8	10	16	15	11	14	13	14	12	16	14	13	14	14
Darwin Núñez	4	13	13	14	12	9	12	13	4	11	14	9	15	16	13	13	12	12	13	17	16	6	16	7	15	15	16	17	14	13	14	17	15	14	14
Matteo Ritaccio	5	7	13	9	14	7	7	7	4	9	7	9	11	8	9	8	8	9	14	8	12	6	12	9	7	13	11	10	8	5	8	7	13	6	4
Andrew Robertson	14	16	12	5	15	9	10	5	13	14	7	15	14	17	14	15	15	14	15	16	11	15	14	13	16	12	17	16	14	13	9	16	17	18	12
Mohamed Salah	11	16	16	16	17	12	10	13	6	15	17	6	17	11	16	10	15	14	16	18	15	11	17	8	16	17	16	18	16	17	8	17	17	15	15
Dominik Szoboszlai	16	13	13	11	15	16	7	15	5	16	16	10	16	12	13	12	14	12	13	16	16	13	14	9	14	15	15	14	12	13	11	17	15	16	11
Thiago	11	13	16	8	18	12	10	11	5	17	12	14	18	14	16	12	17	14	15	15	17	13	15	13	15	18	15	13	17	15	10	11	12	14	9
Kostas Tsimikas	14	16	15	6	15	14	8	8	11	15	11	15	15	14	14	14	15	13	14	15	13	9	14	14	16	14	16	15	14	14	10	15	14	15	12
Virgil van Dijk	3	6	13	8	17	13	17	7	8	16	11	17	16	16	16	15	17	16	16	16	11	17	8	16	16	14	15	15	12	15	17	15	17	15	17
Rhys Williams	3	5	11	7	11	6	15	4	11	12	5	13	11	12	11	13	12	11	11	14	4	9	6	13	13	7	12	11	9	10	16	12	11	12	13

Figure 5: Liverpool player attributes, table exported from FM24.

The data (b) - FBref match results

Data (b) (epl.csv) contains the results of 325 matches played in the EPL so far alongside the expected goals for each match.

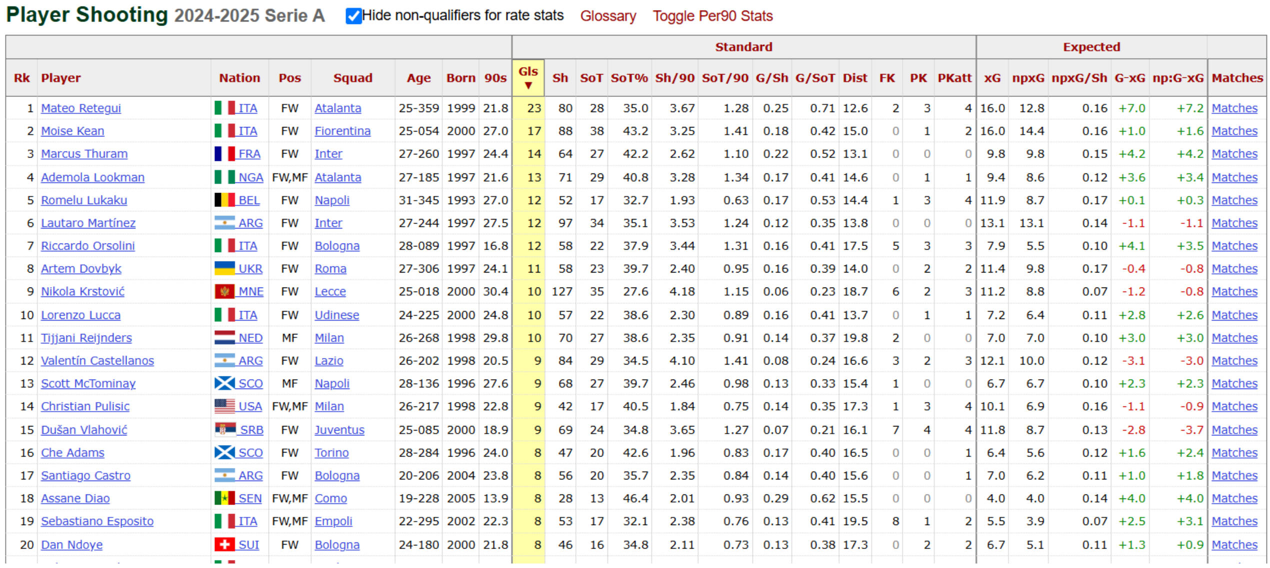
It was taken from the free website "FBref", see here Premier League Scores and Fixtures.



Scores & Fixtures 2024-2025 Premier League Glossary

Wk	Day	Date	Time	Home	xG	Score	xG	Away	Attendance	Venue	Referee	Match Report	Notes
1	Fri	2024-08-16	20:00 *(21:00)*	**Manchester Utd**	2.4	1–0	0.4	Fulham	73,297	Old Trafford	Robert Jones	Match Report	
	Sat	2024-08-17	12:30 *(13:30)*	Ipswich Town	0.5	0–2	2.6	**Liverpool**	30,014	Portman Road Stadium	Tim Robinson	Match Report	
			15:00 *(16:00)*	**Newcastle Utd**	0.3	1–0	1.8	Southampton	52,196	St James' Park	Craig Pawson	Match Report	
			15:00 *(16:00)*	Everton	0.5	0–3	1.4	**Brighton**	39,217	Goodison Park	Simon Hooper	Match Report	
			15:00 *(16:00)*	Nott'ham Forest	1.3	1–1	1.2	Bournemouth	29,763	The City Ground	Michael Oliver	Match Report	
			15:00 *(16:00)*	**Arsenal**	1.2	2–0	0.5	Wolves	60,261	Emirates Stadium	Jarred Gillett	Match Report	
			17:30 *(18:30)*	West Ham	2.3	1–2	2.0	**Aston Villa**	62,463	London Stadium	Tony Harrington	Match Report	
	Sun	2024-08-18	14:00 *(15:00)*	**Brentford**	1.6	2–1	1.2	Crystal Palace	16,988	Gtech Community Stadium	Samuel Barrott	Match Report	
			16:30 *(17:30)*	Chelsea	1.0	0–2	0.8	**Manchester City**	39,818	Stamford Bridge	Anthony Taylor	Match Report	
	Mon	2024-08-19	20:00 *(21:00)*	Leicester City	1.0	1–1	1.2	Tottenham	31,977	King Power Stadium	Chris Kavanagh	Match Report	
2	Sat	2024-08-24	12:30 *(13:30)*	**Brighton**	2.1	2–1	1.4	Manchester Utd	31,537	The American Express Stadium	Craig Pawson	Match Report	
			15:00 *(16:00)*	**Manchester City**	3.3	4–1	0.3	Ipswich Town	53,147	Etihad Stadium	Samuel Allison	Match Report	
			15:00 *(16:00)*	Southampton	0.1	0–1	2.2	**Nott'ham Forest**	31,150	St Mary's Stadium	Samuel Barrott	Match Report	
			15:00 *(16:00)*	**Tottenham**	2.4	4–0	1.0	Everton	61,357	Tottenham Hotspur Stadium	Anthony Taylor	Match Report	
			15:00 *(16:00)*	**Fulham**	1.8	2–1	0.6	Leicester City	25,401	Craven Cottage	Darren Bond	Match Report	
			15:00 *(16:00)*	Crystal Palace	1.3	0–2	1.4	**West Ham**	25,099	Selhurst Park	Robert Jones	Match Report	
			17:30 *(18:30)*	Aston Villa	1.2	0–2	0.9	**Arsenal**	41,587	Villa Park	Michael Oliver	Match Report	
	Sun	2024-08-25	14:00 *(15:00)*	Wolves	1.9	2–6	1.6	**Chelsea**	31,235	Molineux Stadium	Darren England	Match Report	
			14:00 *(15:00)*	Bournemouth	2.2	1–1	1.6	Newcastle Utd	11,161	Vitality Stadium	David Coote	Match Report	
			16:30 *(17:30)*	**Liverpool**	2.5	2–0	0.5	Brentford	60,017	Anfield	Stuart Attwell	Match Report	

Figure 6: The results of the previous EPL matches along side with the expected goals (xG). Screenshot taken from the website. Data collection is done by copy-pasting the entire table. Originally we intended to include results from all of the teams from the top 5 leagues: EPL (England), Serie A (Italy), La Liga (Spain), Bundesliga (Germany), Ligue 1 (France), but the data extraction process of the FM24 attributes was too cumbersome. (It is also "illegal" to share FM24 data online...)



Player Shooting 2024-2025 Serie A ☑Hide non-qualifiers for rate stats Glossary Toggle Per90 Stats

								Standard											Expected					
Rk	Player	Nation	Pos	Squad	Age	Born	90s	Gls ▼	Sh	SoT	SoT%	Sh/90	SoT/90	G/Sh	G/SoT	Dist	FK	PK	PKatt	xG	npxG	npxG/Sh	G-xG	np:G-xG	Matches
1	Mateo Retegui	ITA	FW	Atalanta	25-359	1999	21.8	23	80	28	35.0	3.67	1.28	0.25	0.71	12.6	2	3	4	16.0	12.8	0.16	+7.0	+7.2	Matches
2	Moise Kean	ITA	FW	Fiorentina	25-054	2000	27.0	17	88	38	43.2	3.25	1.41	0.18	0.42	15.0	0	1	2	16.0	14.4	0.16	+1.0	+1.6	Matches
3	Marcus Thuram	FRA	FW	Inter	27-260	1997	24.4	14	64	27	42.2	2.62	1.10	0.22	0.52	13.1	0	0	0	9.8	9.8	0.15	+4.2	+4.2	Matches
4	Ademola Lookman	NGA	FW,MF	Atalanta	27-185	1997	21.6	13	71	29	40.8	3.28	1.34	0.17	0.41	14.6	0	1	1	9.4	8.6	0.12	+3.6	+3.4	Matches
5	Romelu Lukaku	BEL	FW	Napoli	31-345	1993	27.0	12	52	17	32.7	1.93	0.63	0.17	0.53	14.4	1	3	4	11.9	8.7	0.17	+0.1	+0.3	Matches
6	Lautaro Martínez	ARG	FW	Inter	27-244	1997	27.5	12	97	34	35.1	3.53	1.24	0.12	0.35	13.8	0	0	0	13.1	13.1	0.14	-1.1	-1.1	Matches
7	Riccardo Orsolini	ITA	FW	Bologna	28-089	1997	16.8	12	58	22	37.9	3.44	1.31	0.16	0.41	17.5	5	3	3	7.9	5.5	0.10	+4.1	+3.5	Matches
8	Artem Dovbyk	UKR	FW	Roma	27-306	1997	24.1	11	58	23	39.7	2.40	0.95	0.16	0.39	14.0	0	2	2	11.4	9.8	0.17	-0.4	-0.8	Matches
9	Nikola Krstović	MNE	FW	Lecce	25-018	2000	30.4	10	127	35	27.6	4.18	1.15	0.06	0.23	18.7	6	2	3	11.2	8.8	0.07	-1.2	-0.8	Matches
10	Lorenzo Lucca	ITA	FW	Udinese	24-225	2000	24.8	10	57	22	38.6	2.30	0.89	0.16	0.41	13.7	0	1	1	7.2	6.4	0.11	+2.8	+2.6	Matches
11	Tijjani Reijnders	NED	MF	Milan	26-268	1998	29.8	10	70	27	38.6	2.35	0.91	0.14	0.37	19.8	2	0	0	7.0	7.0	0.10	+3.0	+3.0	Matches
12	Valentín Castellanos	ARG	FW	Lazio	26-202	1998	20.5	9	84	29	34.5	4.10	1.41	0.08	0.24	16.6	3	2	3	12.1	10.0	0.12	-3.1	-3.0	Matches
13	Scott McTominay	SCO	MF	Napoli	28-136	1996	27.6	9	68	27	39.7	2.46	0.98	0.13	0.33	15.4	1	0	0	6.7	6.7	0.10	+2.3	+2.3	Matches
14	Christian Pulisic	USA	FW,MF	Milan	26-217	1998	22.8	9	42	17	40.5	1.84	0.75	0.14	0.35	17.3	1	3	4	10.1	6.9	0.16	-1.1	-0.9	Matches
15	Dušan Vlahović	SRB	FW	Juventus	25-085	2000	18.9	9	69	24	34.8	3.65	1.27	0.07	0.21	16.1	7	4	4	11.8	8.7	0.13	-2.8	-3.7	Matches
16	Che Adams	SCO	FW	Torino	28-284	1996	24.0	8	47	20	42.6	1.96	0.83	0.17	0.40	16.5	0	0	1	6.4	5.6	0.12	+1.6	+2.4	Matches
17	Santiago Castro	ARG	FW	Bologna	20-206	2004	23.8	8	56	20	35.7	2.35	0.84	0.14	0.40	15.6	0	0	1	7.0	6.2	0.11	+1.0	+1.8	Matches
18	Assane Diao	SEN	FW,MF	Como	19-228	2005	13.9	8	28	13	46.4	2.01	0.93	0.29	0.62	15.5	0	0	0	4.0	4.0	0.14	+4.0	+4.0	Matches
19	Sebastiano Esposito	ITA	FW,MF	Empoli	22-295	2002	22.3	8	53	17	32.1	2.38	0.76	0.13	0.41	19.5	8	1	2	5.5	3.9	0.07	+2.5	+3.1	Matches
20	Dan Ndoye	SUI	FW	Bologna	24-180	2000	21.8	8	46	16	34.8	2.11	0.73	0.13	0.38	17.3	0	2	2	6.7	5.1	0.11	+1.3	+0.9	Matches

Figure 7: The website offers extensive data for all top football leagues. Here is detailed "shot" data for Serie A. These variables could be potentially used as target variables.

Descriptive analysis of FM24 data

Basics

As a basic description, we calculate : (i) the averages of all attributes for the entire league , (ii) the club TMP averages and the club overall average, (iii) for each club the player with the best TMP avg. and best overall, (iv) the top5 players in each TMP category and the best overall player in the league.

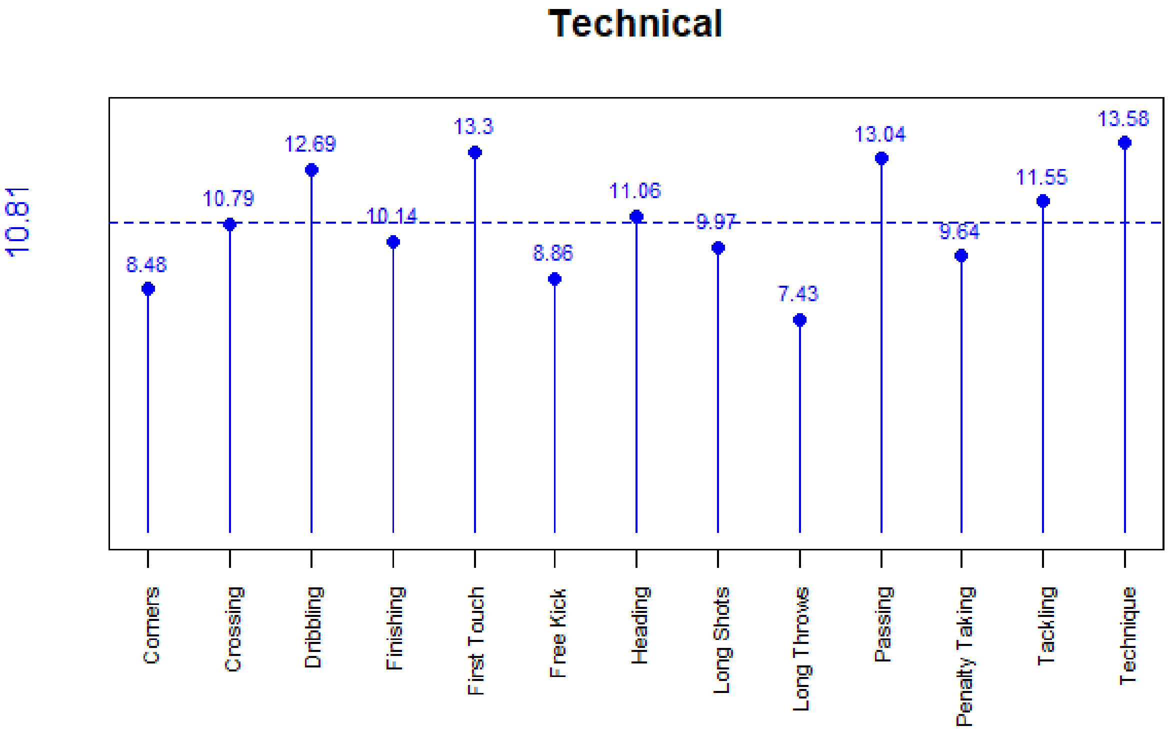


Figure 8: The league averages for each technical attribute. The dotted horizontal line is the average of all the associated attributes, i.e. the average "technical" of the entire league.

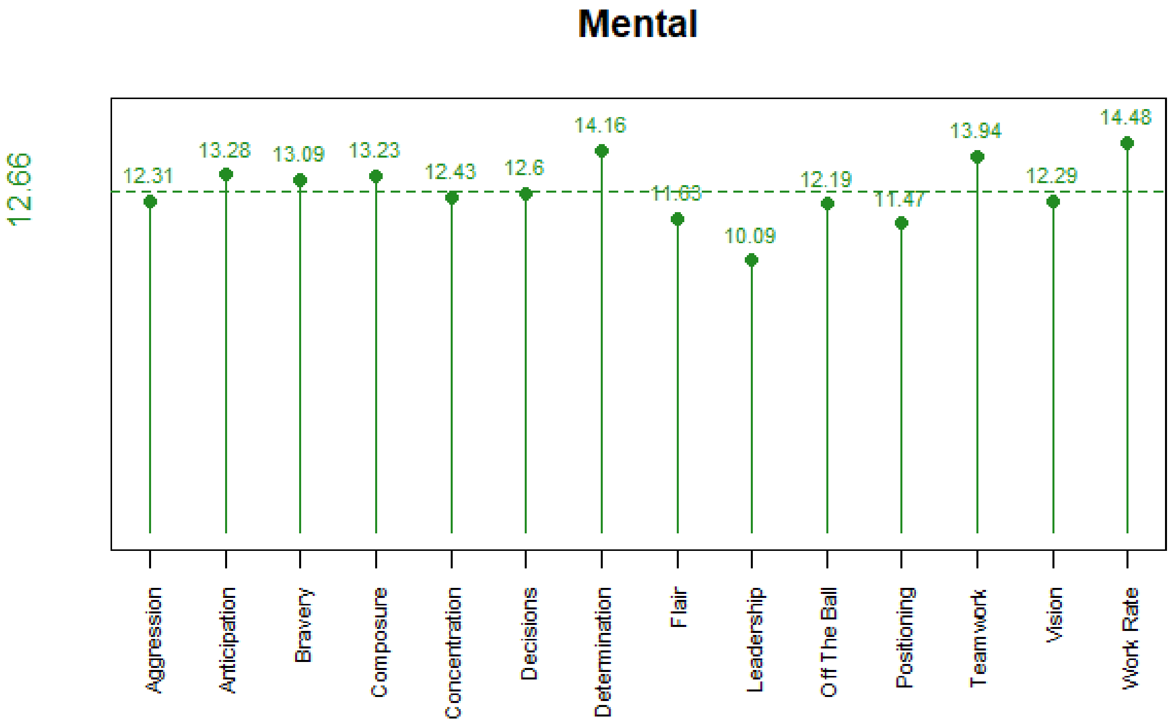


Figure 9: Mental league averages. The "needles" are the league averages of the corresponding attribute, the dotted line is the overall technical attribute average of the league.

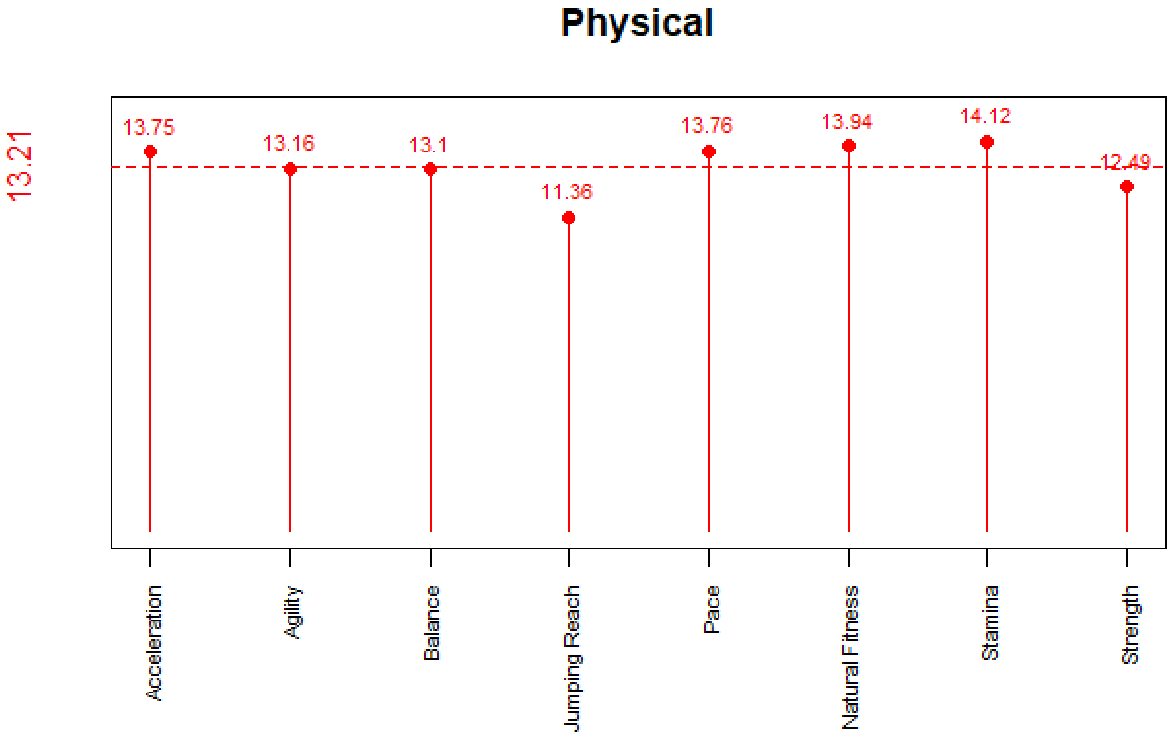


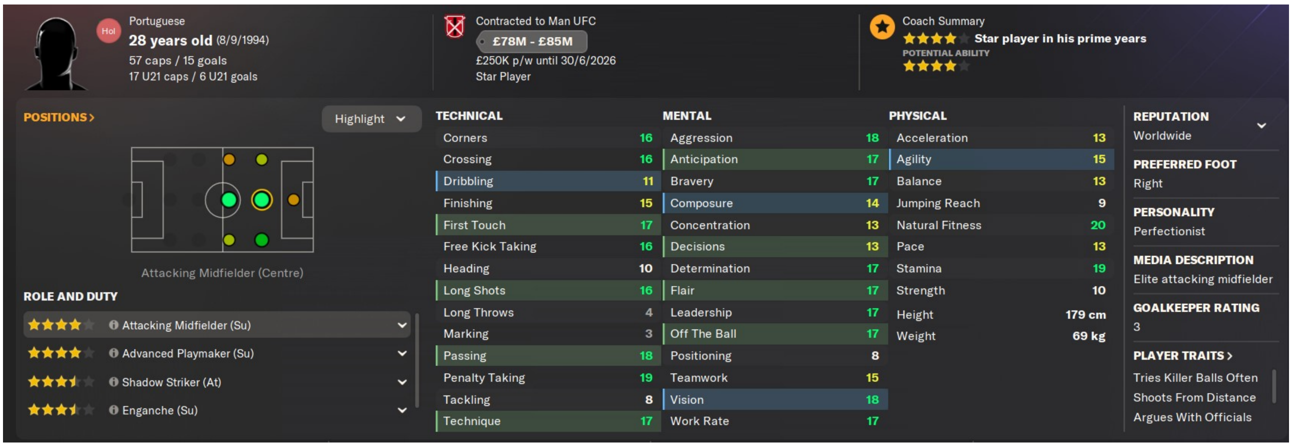
Figure 10: Physical league averages. As a remark, it is said that the EPL the most "physical" leagues in the world. If we had FM24 league attribute averages of other leagues, we could do some hypothesis testing (ANOVA-style for example) and maybe give some evidence to that claim.

Club	Technical	Mental	Physical	Best Technical	Best Mental	Best Physical	Best Overall
Arsenal	13.59	11.40	13.64	Declan Rice 15.86	Martin Ødegaard 13.77	Bukayo Saka 15	Declan Rice 14.31
Aston Villa	12.80	11.11	13.65	Tyrone Mings 14.14	Douglas Luiz 13.31	Ollie Watkins 14.75	Douglas Luiz 13.51
Bournemouth	12.16	10.10	13.16	Lewis Cook 14	Philip Billing 11.92	Dominic Solanke 14.25	Philip Billing 12.51
Bournemouth	12.16	10.10	13.16	Lewis Cook 14	Romain Faivre 11.92	Dominic Solanke 14.25	Philip Billing 12.51
Brentford	12.80	10.95	12.91	Christian Nørgaard 14.43	Mathias Jensen 12.62	Ivan Toney 14.25	Ivan Toney 13.14
Brighton	13.02	11.02	12.94	Pascal Groß 14.86	Pascal Groß 13.62	Igor 14.38	Pascal Groß 13.74
Brighton	13.02	11.02	12.94	Adam Lallana 14.86	Pascal Groß 13.62	Igor 14.38	Pascal Groß 13.74
Chelsea	13.00	10.88	13.86	Thiago Silva 15.14	Christopher Nkunku 13.31	Raheem Sterling 15.25	Reece James 13.86
City	14.02	11.62	14.12	Bernardo Silva 15.64	Kevin De Bruyne 14.46	Erling Haaland 17.12	Erling Haaland 14.63
Crystal Palace	12.43	10.41	12.99	Will Hughes 13.93	Michael Olise 12.92	Marc Guéhi 14.5	Michael Olise 12.77
Everton	12.34	10.48	12.94	Seamus Coleman 13.93	André Gomes 12.31	Dominic Calvert-Lewin 15.12	Abdoulaye Doucouré 12.66
Forest	12.05	10.32	13.11	Ibrahim Sangaré 14	Giovanni Reyna 12.23	Ibrahim Sangaré 15	Ibrahim Sangaré 13
Fulham	12.50	10.82	13.34	João Palhinha 14.36	Willian 12.92	Adama Traoré 15.12	João Palhinha 13.06
Ipswich	11.64	10.11	12.51	Sam Morsy 13.5	Conor Chaplin 12.38	Ali Al-Hamadi 13.75	Kieffer Moore 12.23
Leicester	12.11	10.67	12.46	Conor Coady 13.64	Dennis Praet 12.15	Stephy Mavididi 14.25	Kiernan Dewsbury-Hall 12.86
Liverpool	13.28	11.51	13.42	Thiago 14.93	Alexis Mac Allister 14	Mohamed Salah 15.38	Mohamed Salah 14.14
Liverpool	13.28	11.51	13.42	Virgil van Dijk 14.93	Alexis Mac Allister 14	Virgil van Dijk 15.38	Mohamed Salah 14.14
Newcastle	13.12	10.97	13.10	Bruno Guimarães 15.36	Kieran Trippier 13.23	Alexander Isak 14.62	Bruno Guimarães 14.06
Newcastle	13.12	10.97	13.10	Joelinton 15.36	Kieran Trippier 13.23	Alexander Isak 14.62	Bruno Guimarães 14.06
Southampton	11.84	10.10	12.55	Ryan Manning 13.36	Stuart Armstrong 12.69	Jan Bednarek 14.25	Stuart Armstrong 12.57
Tottenham	12.73	11.21	13.64	Pierre-Emile Højbjerg 14.57	Pedro Porro 13.38	Micky van de Ven 15	Pedro Porro 13.63
United	12.81	11.01	13.68	Bruno Fernandes 15.57	Bruno Fernandes 14.08	Marcus Rashford 15.75	Bruno Fernandes 14.66
West Ham	13.08	11.01	13.55	Edson Álvarez 14.79	James Ward-Prowse 13.31	Kurt Zouma 15.38	James Ward-Prowse 13.74
Wolves	12.31	10.90	13.01	João Gomes 13.79	Pablo Sarabia 12.92	Nélson Semedo 14.12	Matt Doherty 12.66
Wolves	12.31	10.90	13.01	João Gomes 13.79	Pablo Sarabia 12.92	Nélson Semedo 14.12	João Gomes 12.66
Wolves	12.31	10.90	13.01	João Gomes 13.79	Pablo Sarabia 12.92	Nélson Semedo 14.12	Pablo Sarabia 12.66

Table 1: Best players by club with their technical, mental and physical ratings. If some players share the first place in any of these categories they appear multiple times.

Top Mental Attribute Ratings		
Club	**Player**	**Mental Rating**
Arsenal	Declan Rice	15.9
Manchester City	Bernardo Silva	15.6
Manchester United	Bruno Fernandes	15.6
Newcastle	Bruno Guimarães	15.4
Newcastle	Joelinton	15.4
Top Technical Attribute Ratings		
Club	**Player**	**Technical Rating**
Manchester City	Kevin De Bruyne	14.5
Manchester United	Bruno Fernandes	14.1
Liverpool	Alexis Mac Allister	14.0
Arsenal	Martin Ødegaard	13.8
Liverpool	Trent Alexander-Arnold	13.7
Top Physical Attribute Ratings		
Club	**Player**	**Physical Rating**
Manchester City	Erling Haaland	17.7
Manchester United	Marcus Rashford	16.0
Manchester City	Kyle Walker	15.7
Fulham	Adama Traoré	15.7
Liverpool	Mohamed Salah	15.5
Top Overall Ratings		
Club	**Player**	**Overall Rating**
Manchester United	Bruno Fernandes	14.7
Manchester City	Erling Haaland	14.6
Manchester City	Bernardo Silva	14.5
Manchester City	Kevin De Bruyne	14.5
Arsenal	Declan Rice	14.2

Table 2: Top 5 players in each attribute (TMP) and top 5 players overall. This is similar to the previous table, but here it is allowed for the players of the same club to appear. For example, three players out of the overall top 5 play for Manchester City (Haaland, Silva, De Bruyne).



Portuguese
28 years old (8/9/1994)
57 caps / 15 goals
17 U21 caps / 6 U21 goals

Contracted to Man UFC
£78M - £85M
£250K p/w until 30/6/2026
Star Player

Coach Summary
★★★★☆ **Star player in his prime years**
POTENTIAL ABILITY
★★★★☆

Attacking Midfielder (Centre)

ROLE AND DUTY

- ★★★★☆ Attacking Midfielder (Su)
- ★★★★☆ Advanced Playmaker (Su)
- ★★★½☆ Shadow Striker (At)
- ★★★½☆ Enganche (Su)

TECHNICAL		MENTAL		PHYSICAL	
Corners	16	Aggression	18	Acceleration	13
Crossing	16	Anticipation	17	Agility	15
Dribbling	11	Bravery	17	Balance	13
Finishing	15	Composure	14	Jumping Reach	9
First Touch	17	Concentration	13	Natural Fitness	20
Free Kick Taking	16	Decisions	13	Pace	13
Heading	10	Determination	17	Stamina	19
Long Shots	16	Flair	17	Strength	10
Long Throws	4	Leadership	17	Height	179 cm
Marking	3	Off The Ball	17	Weight	69 kg
Passing	18	Positioning	8		
Penalty Taking	19	Teamwork	15		
Tackling	8	Vision	18		
Technique	17	Work Rate	17		

REPUTATION
Worldwide

PREFERRED FOOT
Right

PERSONALITY
Perfectionist

MEDIA DESCRIPTION
Elite attacking midfielder

GOALKEEPER RATING
3

PLAYER TRAITS >
Tries Killer Balls Often
Shoots From Distance
Argues With Officials

Figure 11: The best overall player according to FM24 (mod the "Marking" attribute). In-game screen shot.

PCA

We preform a PCA on data (a), we keep seven principle components (PC) for a total of roughly %70 inertia. The data is "easy enough" to give a direct interpretation of the PC's. In R we use the function prcomp() and view the resulting eigenvectors. We can furthermore sort those eigenvectors to reveal the biggest "weights". So for example

---

$$PC1 \overset{\text{def}}{=} X_1 \approx 0.25(Y_{\text{jump}} + Y_{\text{tackle}} + Y_{\text{bravery}} + \dots) - 0.25(Y_{\text{agility}} + Y_{\text{dribbling}} + Y_{\text{technique}} \; + \dots)$$

$$PC2 \overset{\text{def}}{=} X_2 \approx 0.25(Y_{\text{decisions}} + Y_{\text{composure}} + Y_{\text{passing}} + Y_{\text{first touch}}) + 0.1(Y_{\text{speed}} + Y_{\text{strength}} + Y_{\text{jumping}} \; + \dots)$$

It is clear that $PC1 = X_1$ represents the difference of "Brute Force (physical and mental)" vs "Dexterity/Sophistication". A player with high X_1 is a physical specimen maybe lacking mobility.

$PC2$ seems to represent general "Play-making ability , creativity" since the physical stats are weighted low. Players with high X_2 are chance creators , but not necessarily physically weak.

The first two PCs account for %30 , %20 of the total inertia respectively, and we will stop the interpretation there. The remaining five PCs are interpreted similarly.

To confirm our interpretation we make two PC1 vs PC2 plots.

- **Plot 1:** the top 3 players with the highest PC1 values and the bottom 3 players with the lowest PC1 values.
- **Plot 2:** the top 3 players with the highest PC2 values and the bottom 3 players with the lowest PC2 values.

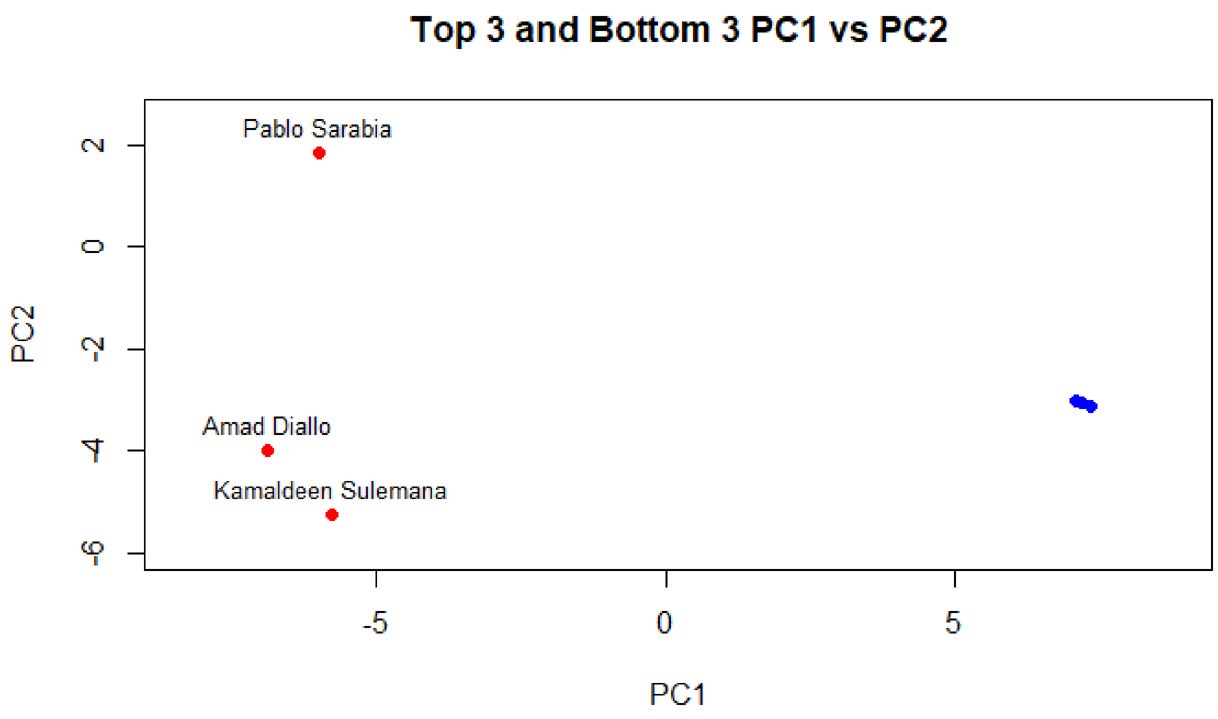


Figure 12: The red dots are the players with lowest PC1, the blue dots are players with highest PC1. Effectively, all the red dot players are "small, agile speedsters".

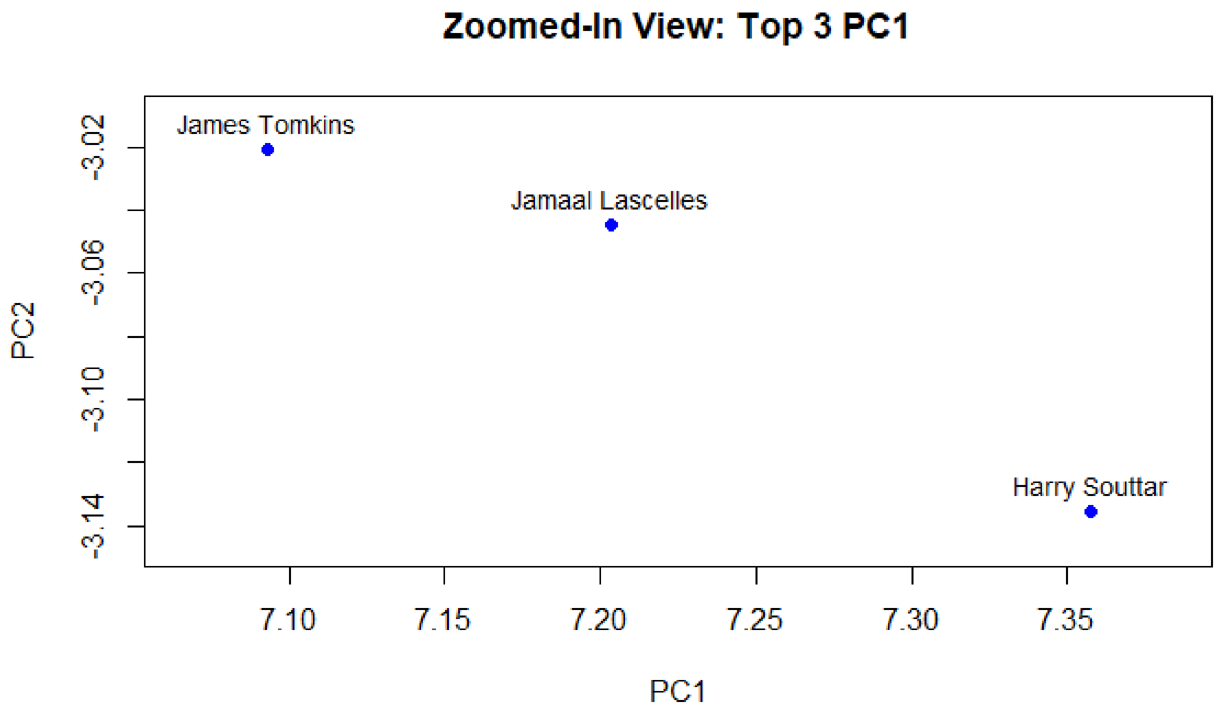


Figure 13: Zooming in on the blue cluster; the players here are big strong defenders, this confirms our interpretation.

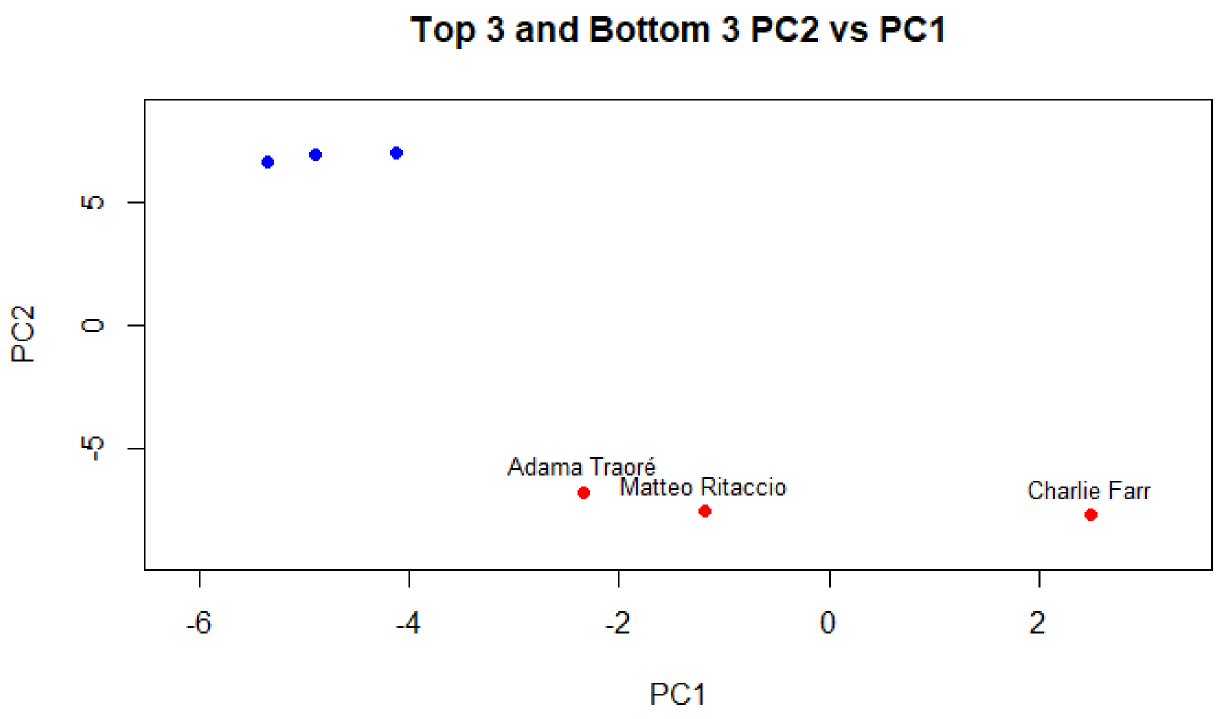


Figure 14: Same for PC2. The red dots are indeed the worst play-makers in the league.

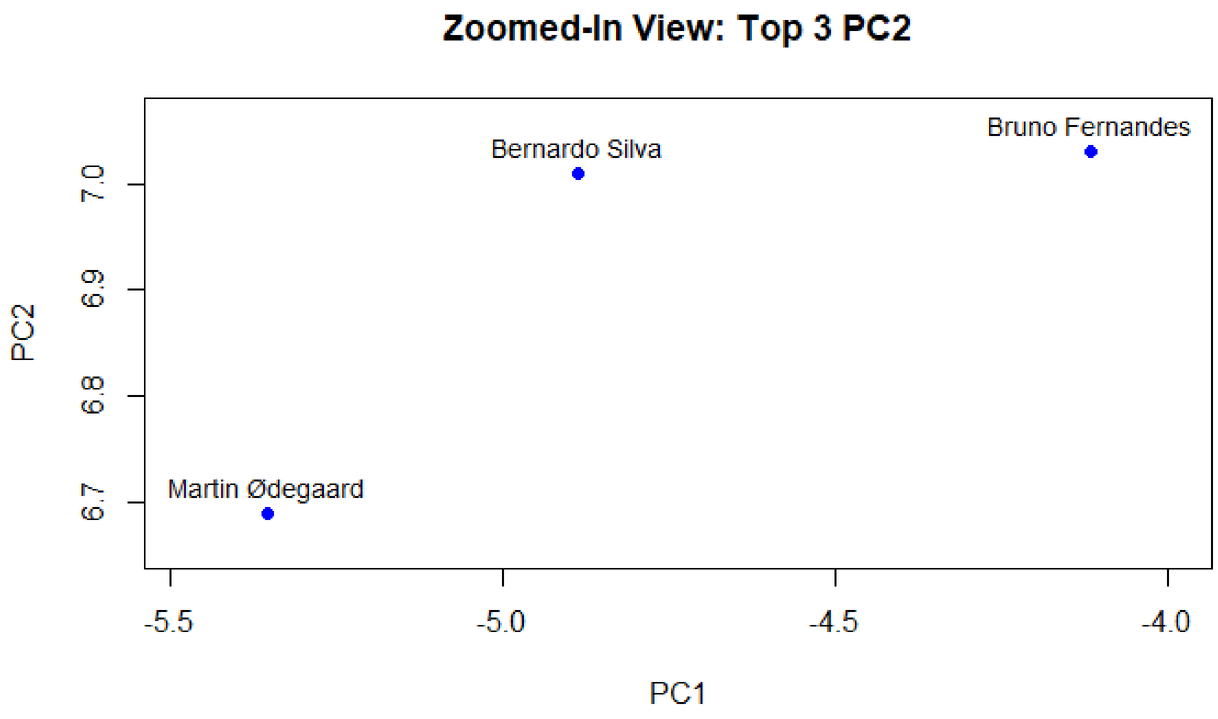


Figure 15: Zooming in : the blue cluster are indeed the best play-makers in the league.

We end with some PC1 vs PC2 plots of different clubs. PC1 means "physically powerful" , PC2 means "very creative".

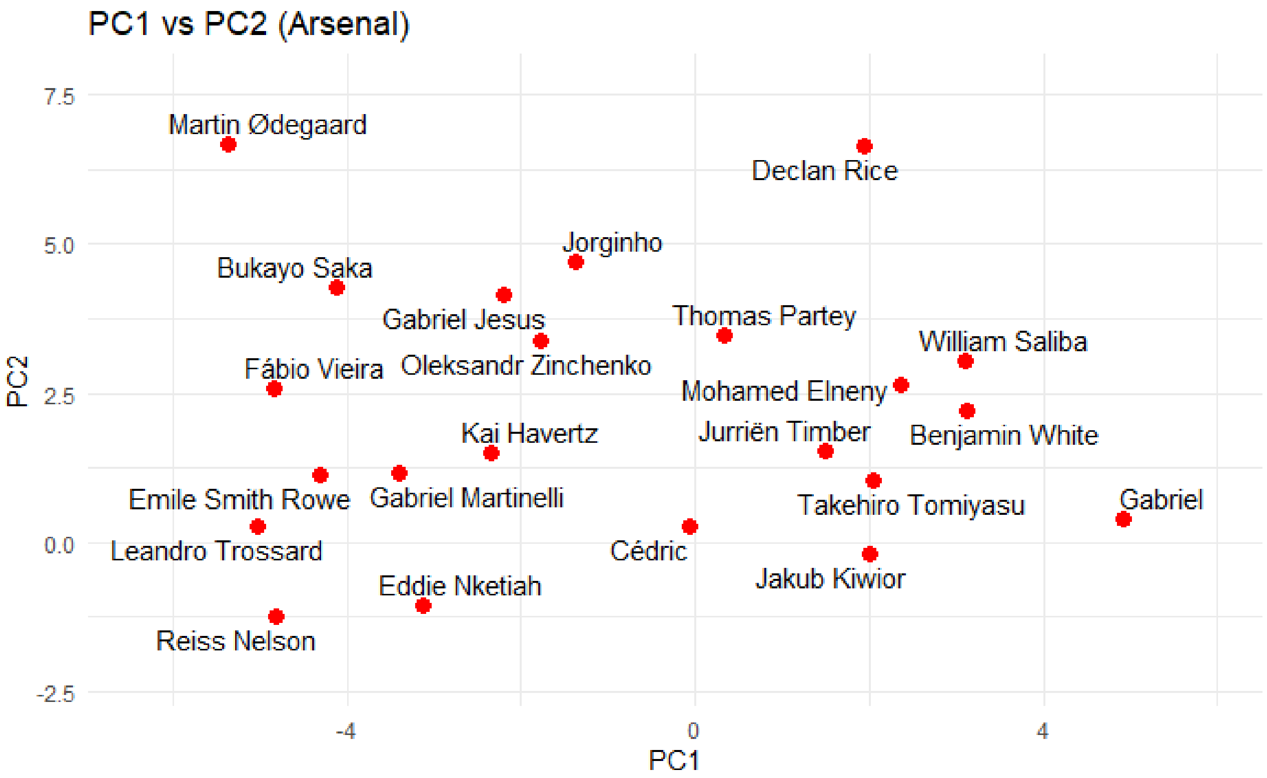


Figure 16: Notice Declan Rice with high PC1 and PC2. From table 2. we know that he is the 5th best player in the league overall and the first in the league in terms of "mental" stats.

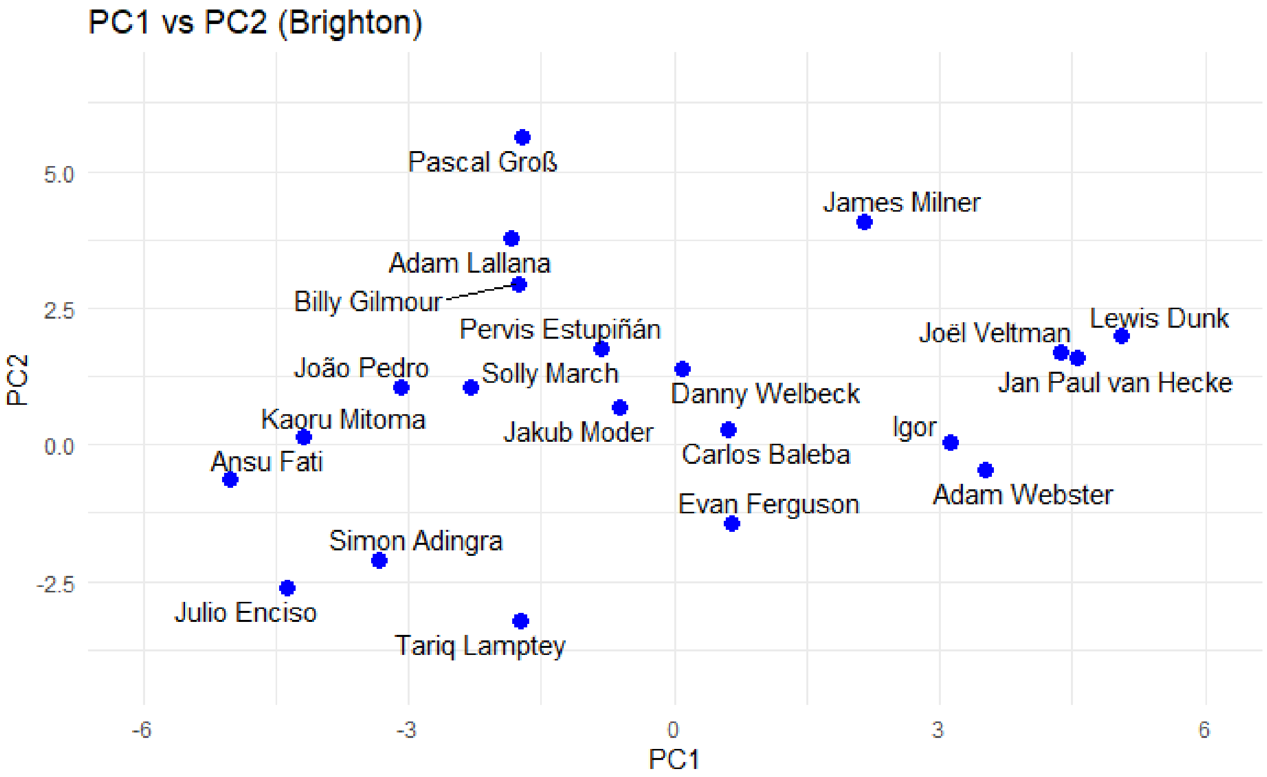


Figure 17: It is known that the Brighton defenders Dunk, Veltman and van Hecke are one of the most creative defenses in the league. The high PC2 confirms that.

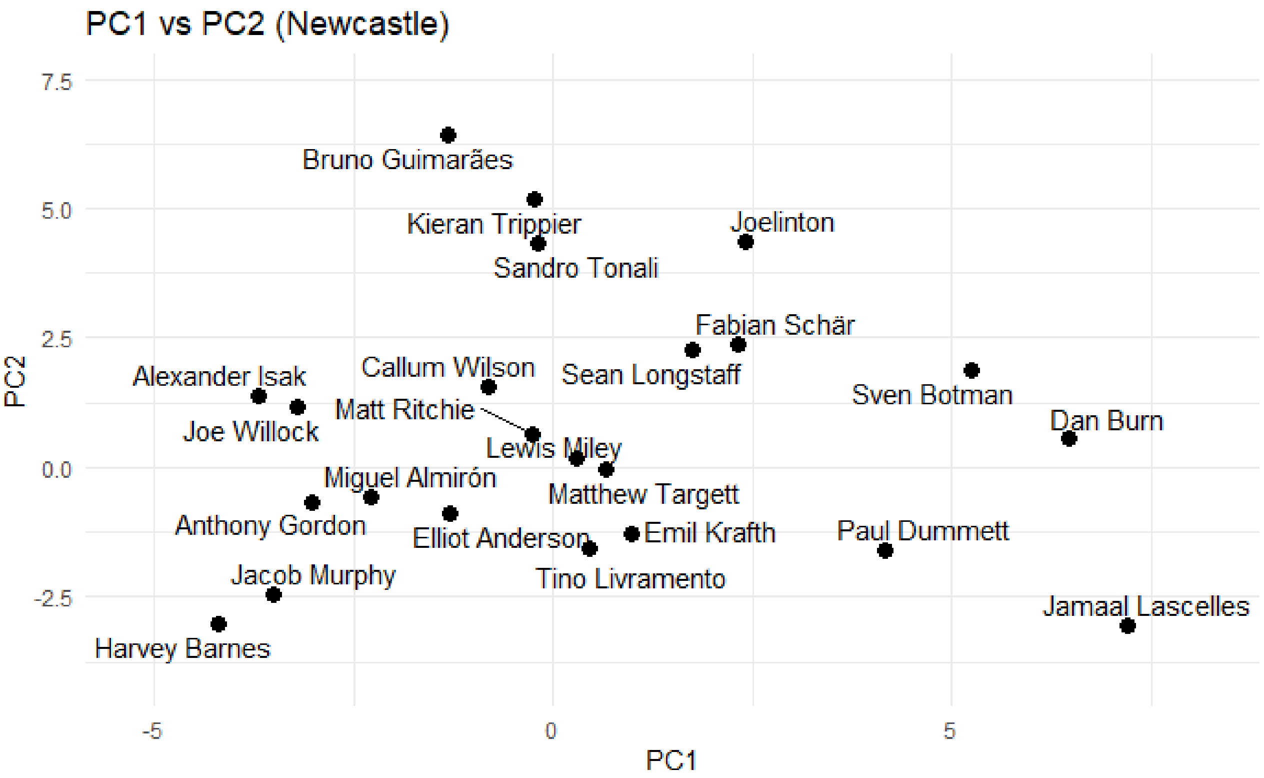


Figure 18: We see that both Joelinton and Giumaraes are creative , with Joelinton being also "strong". According to table 1, they are of "equal technique".

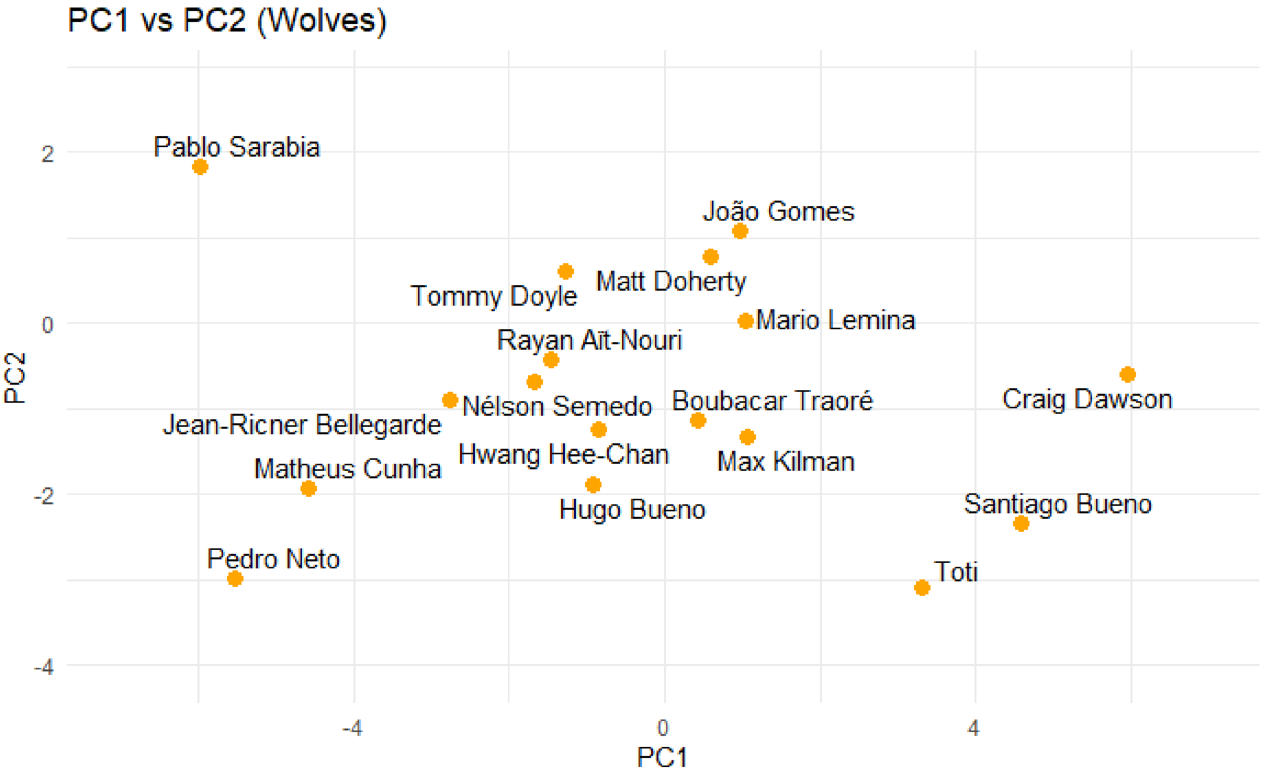
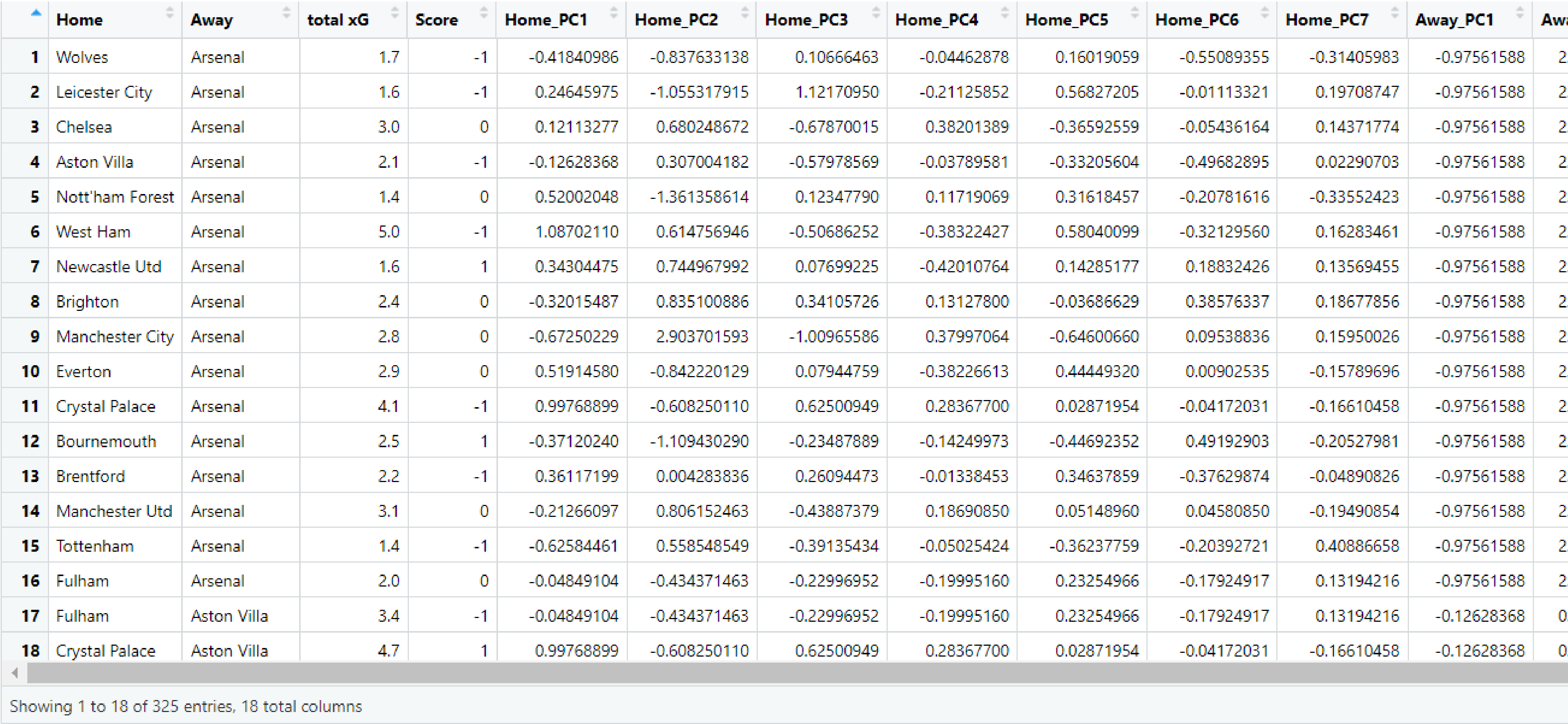


Figure 19: Wolverhampton. Indeed the Wolves star player (Sarabia), detected in table 1, is the most creative player in the squad.

The data (c) - PCA reduced FM24 combined with FBref

Data (c) contains the teams,results,xG of data (b) and the seven + seven (home and away) PC components calculated from data (a). Here, each PC is the average of the individual player PC in the corresponding team , i.e. the PC are team averages; (c) is the "training data".

The testing data (d) contains the remaining 55 matches that are to be played.



	Home	Away	total xG	Score	Home_PC1	Home_PC2	Home_PC3	Home_PC4	Home_PC5	Home_PC6	Home_PC7	Away_PC1	Awa
1	Wolves	Arsenal	1.7	-1	-0.41840986	-0.837633138	0.10666463	-0.04462878	0.16019059	-0.55089355	-0.31405983	-0.97561588	2.
2	Leicester City	Arsenal	1.6	-1	0.24645975	-1.055317915	1.12170950	-0.21125852	0.56827205	-0.01113321	0.19708747	-0.97561588	2.
3	Chelsea	Arsenal	3.0	0	0.12113277	0.680248672	-0.67870015	0.38201389	-0.36592559	-0.05436164	0.14371774	-0.97561588	2.
4	Aston Villa	Arsenal	2.1	-1	-0.12628368	0.307004182	-0.57978569	-0.03789581	-0.33205604	-0.49682895	0.02290703	-0.97561588	2.
5	Nott'ham Forest	Arsenal	1.4	0	0.52002048	-1.361358614	0.12347790	0.11719069	0.31618457	-0.20781616	-0.33552423	-0.97561588	2.
6	West Ham	Arsenal	5.0	-1	1.08702110	0.614756946	-0.50686252	-0.38322427	0.58040099	-0.32129560	0.16283461	-0.97561588	2.
7	Newcastle Utd	Arsenal	1.6	1	0.34304475	0.744967992	0.07699225	-0.42010764	0.14285177	0.18832426	0.13569455	-0.97561588	2.
8	Brighton	Arsenal	2.4	0	-0.32015487	0.835100886	0.34105726	0.13127800	-0.03686629	0.38576337	0.18677856	-0.97561588	2.
9	Manchester City	Arsenal	2.8	0	-0.67250229	2.903701593	-1.00965586	0.37997064	-0.64600660	0.09538836	0.15950026	-0.97561588	2.
10	Everton	Arsenal	2.9	0	0.51914580	-0.842220129	0.07944759	-0.38226613	0.44449320	0.00902535	-0.15789696	-0.97561588	2.
11	Crystal Palace	Arsenal	4.1	-1	0.99768899	-0.608250110	0.62500949	0.28367700	0.02871954	-0.04172031	-0.16610458	-0.97561588	2.
12	Bournemouth	Arsenal	2.5	1	-0.37120240	-1.109430290	-0.23487889	-0.14249973	-0.44692352	0.49192903	-0.20527981	-0.97561588	2.
13	Brentford	Arsenal	2.2	-1	0.36117199	0.004283836	0.26094473	-0.01338453	0.34637859	-0.37629874	-0.04890826	-0.97561588	2.
14	Manchester Utd	Arsenal	3.1	0	-0.21266097	0.806152463	-0.43887379	0.18690850	0.05148960	0.04580850	-0.19490854	-0.97561588	2.
15	Tottenham	Arsenal	1.4	-1	-0.62584461	0.558548549	-0.39135434	-0.05025424	-0.36237759	-0.20392721	0.40886658	-0.97561588	2.
16	Fulham	Arsenal	2.0	0	-0.04849104	-0.434371463	-0.22996952	-0.19995160	0.23254966	-0.17924917	0.13194216	-0.97561588	2.
17	Fulham	Aston Villa	3.4	-1	-0.04849104	-0.434371463	-0.22996952	-0.19995160	0.23254966	-0.17924917	0.13194216	-0.12628368	0.
18	Crystal Palace	Aston Villa	4.7	1	0.99768899	-0.608250110	0.62500949	0.28367700	0.02871954	-0.04172031	-0.16610458	-0.12628368	0.

Showing 1 to 18 of 325 entries, 18 total columns

Figure 20: Screenshot of data (c) from R-studio. The scores have been converted to $-1, 0, +1$ factors corresponding to win-draw-lose of the home team. The definition of "xG" can be found here: Opta – Expected Goals (xG) definition. FBref uses Opta definitions for its statistics. The "away PCs" continue further to the right and are not shown here.

ANOVA: top half vs bottom half

Methodology

We introduce ANOVA (and its more complicated variants) through a regression lens, opting to avoid the watered down version so commonly found online. ANOVA then appears as the special case where the predictors $\boldsymbol{X}$ are categorical but numerically encoded. For convenience we summarize the goals of regression; the theory is perhaps most elegantly laid out in the course STAT-F406 (prof. Paindaveine) and can be found on the professor's website.

Note: At some point we drop the bold font notation for the vectors, when that is done should be clear from context.

Definition 1. (Regression in a nutshell)

Consider $\mathcal{H}$, the vector space of all random variable with finite mean. On this vector space we introduce the scalar product $\langle X, Y\rangle = \mathbb{E}[XY]$ and then $\mathcal{H}$ becomes a *Hilbert space.*

Let $Y \in \mathbb{R}$ the target variable, let $\boldsymbol{X} \in \mathbb{R}^k$ be k-dimensional predictor. The goal is to have Y explained/approximated by $\boldsymbol{X}$, i.e. $Y \approx \boldsymbol{\beta}'\boldsymbol{X}$.

Letting

$$\boldsymbol{\beta} = \min_{\alpha \in \mathbb{R}} ||Y - \boldsymbol{\alpha X}||$$

defines $Y - \boldsymbol{\alpha X}$ uniquely. If we additionally suppose that the $\boldsymbol{X}$ are linearly independent, $\boldsymbol{\beta}$ is uniquely determined and is given by:

$$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{X X}']^{-1}\mathbb{E}[\boldsymbol{X}Y]$$

The quantity

$$\varepsilon = Y - \boldsymbol{\beta X}$$

unambiguously defines the "approximation" of Y by $\boldsymbol{X}$ (i.e. the link between the error and the coefficients).

ε is the orthogonal projection of Y onto the space spanned by $\boldsymbol{X}$.

Definition 2. (Strong model with normality - SMN)

$$\varepsilon \perp X \text{ and } \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

We then have

$$Y \mid X = x \sim \mathcal{N}(\beta' x, \sigma^2)$$

Y conditionally on X is normally distributed, with the mean being the projection of Y onto the span of X.

Definition 3. ($\hat{\beta}$ estimator of β in the SMN)

Let

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

We have that:

$$\hat{\beta} \to \beta, a.s.$$

Furthermore, if the model is SMN (strong with normality)

$$\hat{\beta} \mid \mathcal{X} \sim \mathcal{N}_k\left(\beta, \frac{\sigma^2}{n} Q^{-1}\right).$$

$Q = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$. Also $\hat{\beta}$ is *the maximum likelihood estimator* for β.

Definition 4. (Hypothesis testing on β_j)

For the hypothesis testing problem

$$\begin{cases} \mathcal{H}_0 : \beta_j = c \\ \mathcal{H}_1 : \beta_j \neq c, \end{cases}$$

the natural test at level α rejects $\mathcal{H}_0$ if and only if

$$|T| > t_{n-k,1-\alpha/2}, \quad \text{where} \quad T = \frac{\sqrt{n-k}(\hat{\beta}_j - c)}{\hat{\sigma}\sqrt{(Q^{-1})_{jj}}}.$$

Statistical software most often returns the p-value

$$P_{\mathcal{H}_0}\left(|T| > |T_{\text{observed}}|\right),$$

and we reject at level α if and only if this p-value is less than α.

Definition 5. (Linear hypothesis testing)

Consider the hypothesis testing problem

$$\begin{cases} \mathcal{H}_0 : R\beta = r \\ \mathcal{H}_1 : R\beta \neq r, \end{cases}$$

where R is a fixed $p \times k$ matrix ($p \leq k$), and $r \in \mathbb{R}^p$ is a fixed vector.

$$W = \frac{(n-k)(R\hat{\beta} - r)'(RQ^{-1}R')^{-1}(R\hat{\beta} - r)}{p\hat{\sigma}^2} \sim F_{p,n-k} \tag{1}$$

The resulting test rejects $\mathcal{H}_0 : R\beta = r$ at level α if and only if

$$W > F_{p,n-k,1-\alpha}.$$

Statistical software most often returns the p-value

$$P_{\mathcal{H}_0}\left(W > W_{\text{observed}}\right),$$

and we reject at level α if and only if this p-value is less than α.

ANOVA?

The X appearing in definitions 1:5 can be categorical, it suffices to "encode them numerically". For simplicity suppose Y is continuous and suppose that the only X is a categorical one with $\{C_1, C_2\}$ as categories (this of course also works for more complicated cases).

Writing $X = (X_1, X_2)'$ as the random vector that takes values $(1,0)', (0,1)'$ if "$X = C_1, C_2$" allows us to keep the regression formalism:

$$Y = \beta_1 X_1 + \beta_2 X_2$$

$E[Y \mid X_{1,2}] = \mu_1, \mu_2$ appears as the average of category C_1, C_2.

We can also keep the intercept, then it suffices to use only X_1, then

$$Y = \beta_1 + \beta_2 X_1$$

and:

$E[Y \mid X_1 = 0] = \beta_1$, $E[Y \mid X_1 = 1] = \beta_1 + \beta_2$

In other words, β_1 is the average of C_2 and $\beta_2 + \beta_1$ the average of C_1. The hypothesis test $\mathcal{H}_0 : \mu_1 = \mu_2$ vs $\mathcal{H}_1 : \mu_1 \neq \mu_2$ is equivalent to $\mathcal{H}_0 : \beta_2 = 0$ vs $\mathcal{H}_1 : \beta_2 \neq 0$. The appropriate test is given by definition 4, and it coincides with the usual ANOVA t-test $T = \frac{\bar{Y_2} - \bar{Y_1}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

A simple application

We illustrate the above an application. Let C_1 be the top half of the EPL (top 10) and C_2 the bottom half (bottom 10) - these are encoded via X_1 ("belongs to top half") and the intercept. Let Y be the average of the physical variables, we then have $n \approx 450$ player physical averages as the target and X_1 as the predictor.

$$Y = \beta_0 + \beta_1 X_1$$

β_0 is the average of the bottom 10 teams ("the reference") , β_1 is the difference in average of the top and bottom teams.

We wish to determine if $\mu_1 = \mu_2$. This can be done with the hypothesis test: $\mathcal{H}_0 : \beta_1 = 0$.

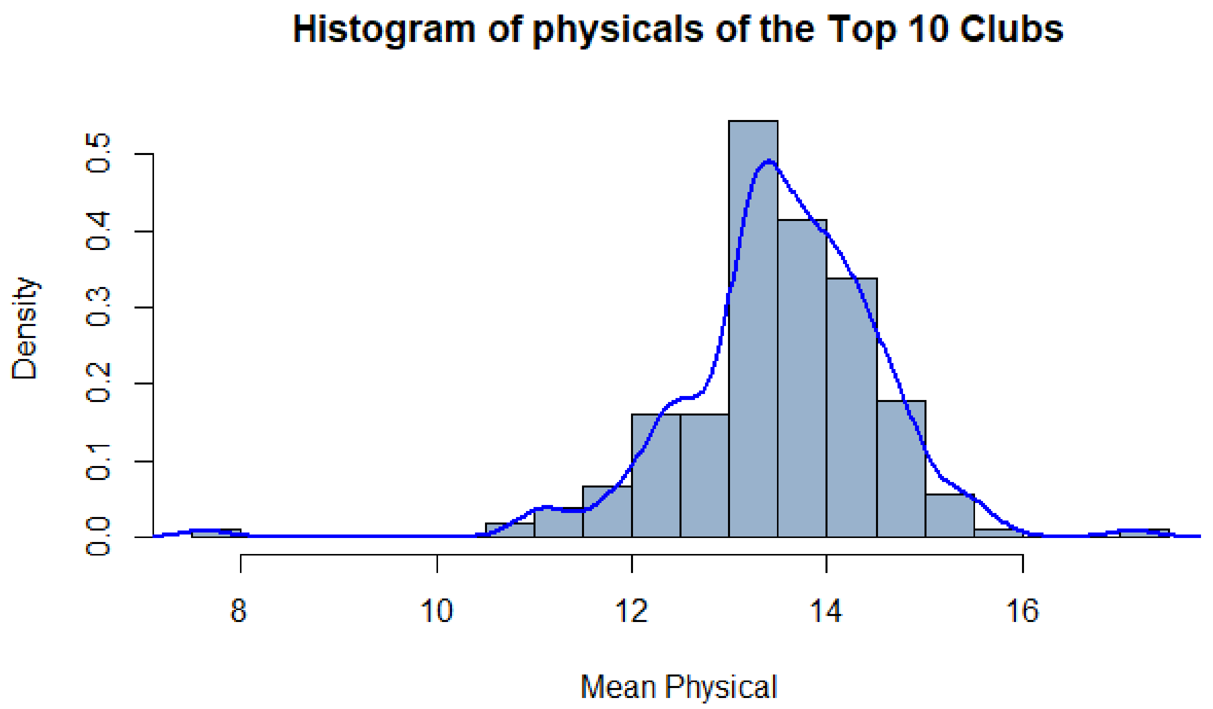


Figure 21: Histogram and kernel density approximations of the physicals distribution for the top 10. The average is 13.5 and the variance ≈ 1 (1.03). There is an outlier at 8, it is "Ritaccio" a young reserve player for Liverpool.

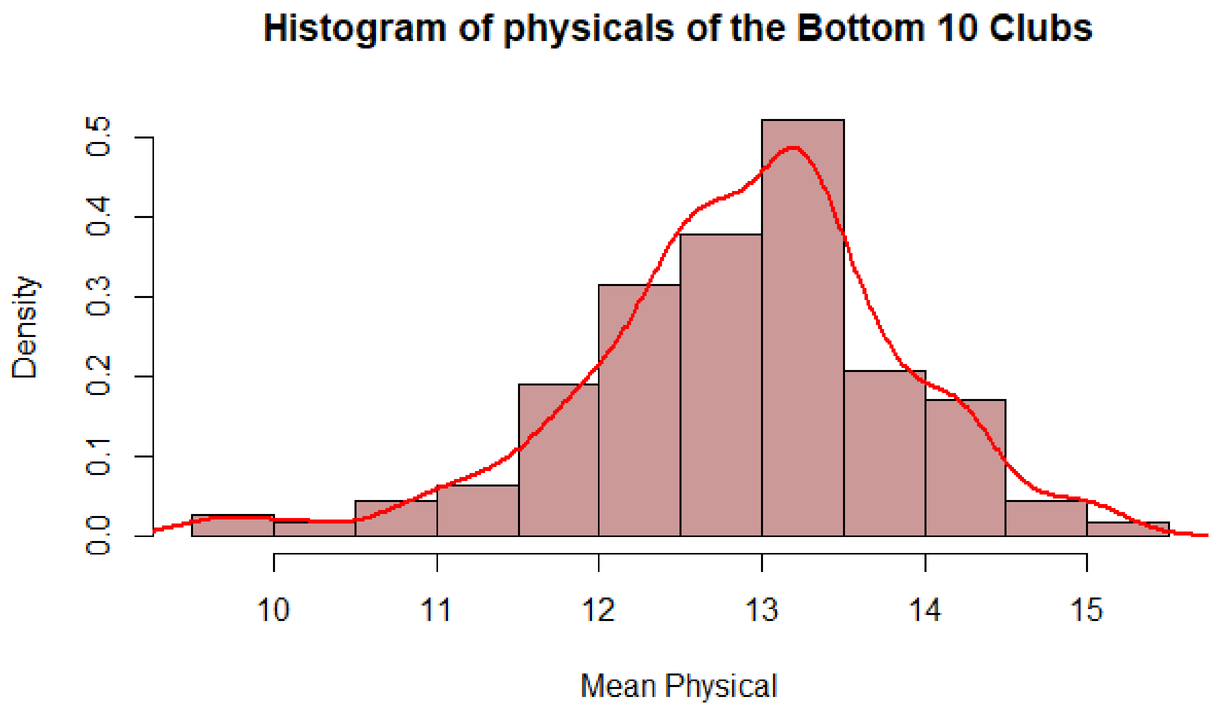
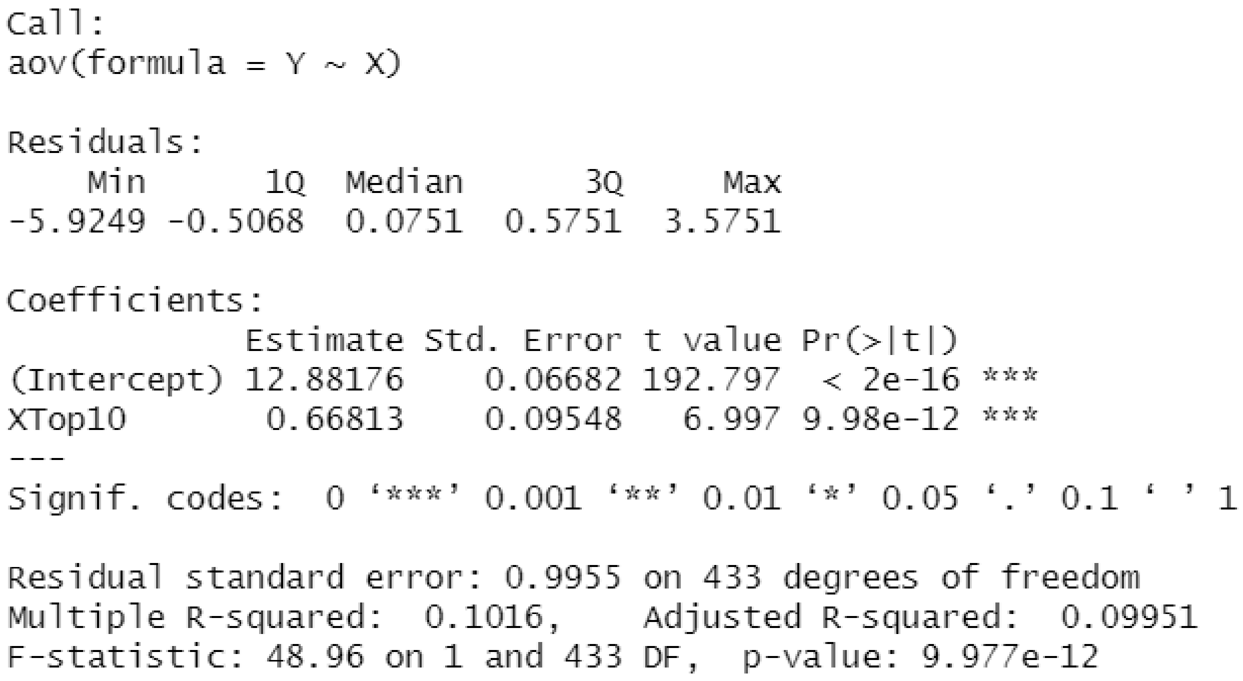


Figure 22: Histogram and kernel density approximations of the physicals distribution for the bottom 10. The average is 12.8 and variance ≈ 1 (0.95).



```
Call:
aov(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9249 -0.5068  0.0751  0.5751  3.5751 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.88176    0.06682 192.797  < 2e-16 ***
XTop10       0.66813    0.09548   6.997 9.98e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9955 on 433 degrees of freedom
Multiple R-squared:  0.1016,    Adjusted R-squared:  0.09951 
F-statistic: 48.96 on 1 and 433 DF,  p-value: 9.977e-12
```

Figure 23: The model produces $\hat{\beta}_0 = 12.88$ bottom averages ; $\hat{\beta}_1 = 0.66$, i.e. $\hat{\mu}_{top} = \hat{\beta}_1 + \hat{\beta}_0 = 13.54$. Both the coefficients are significant since the p-values are fantastically low. The hypothesis test for $\beta_1 = 0$ is rejected - the top teams are physically stronger than the bottom teams.

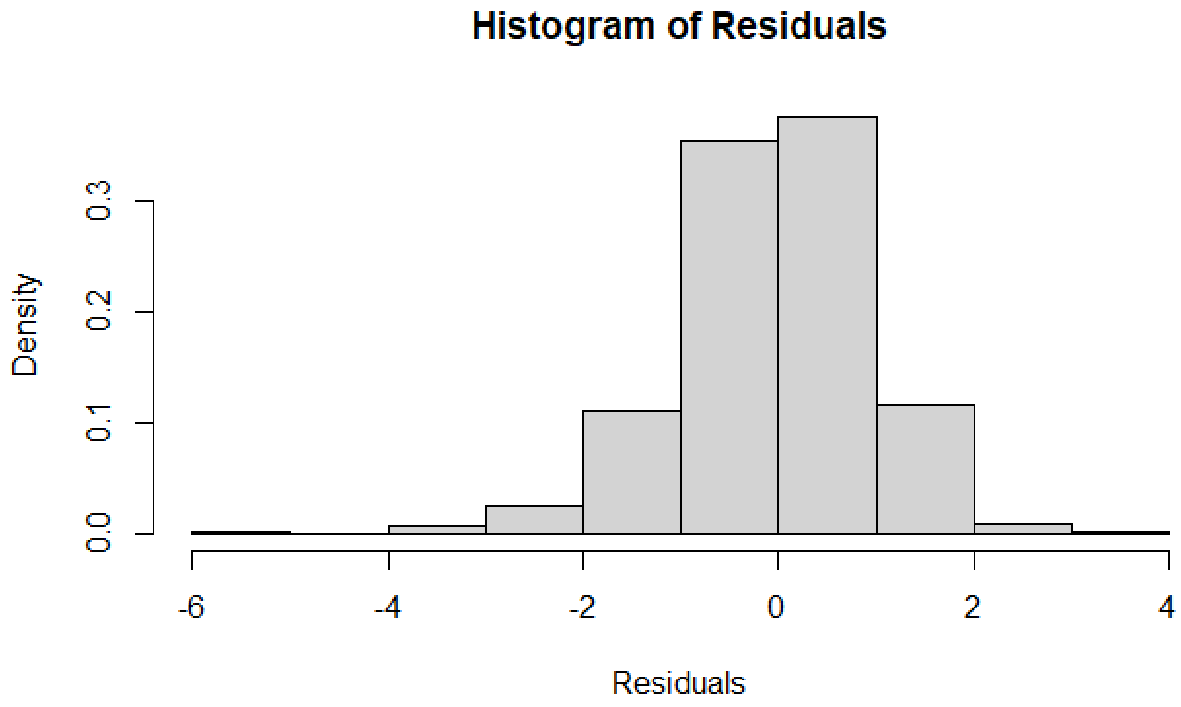


Figure 24: Plotting the residuals of our model we see that they are normal, as was already suggested by the histograms of figures 21. and 22.

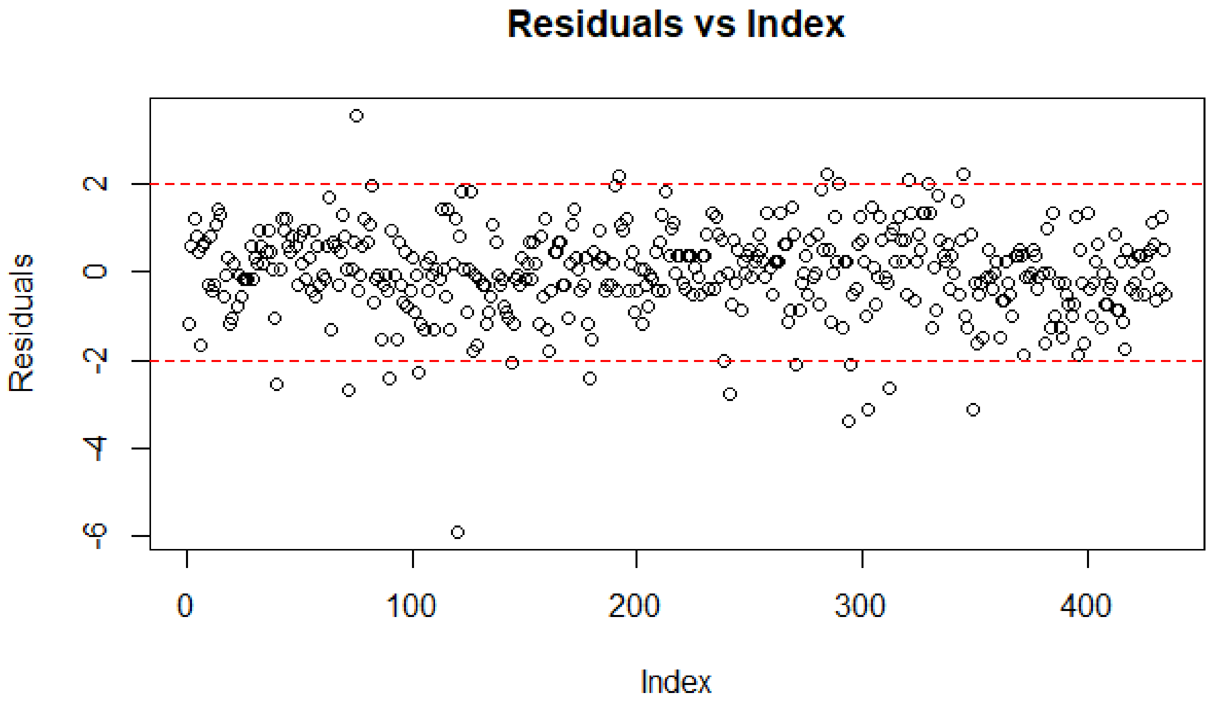


Figure 25: Plot of the standardized residuals with the index representing each player. The dotted red lines are the usual cutoffs for outlier detection. The biggest outlier is the same player as mentioned before.

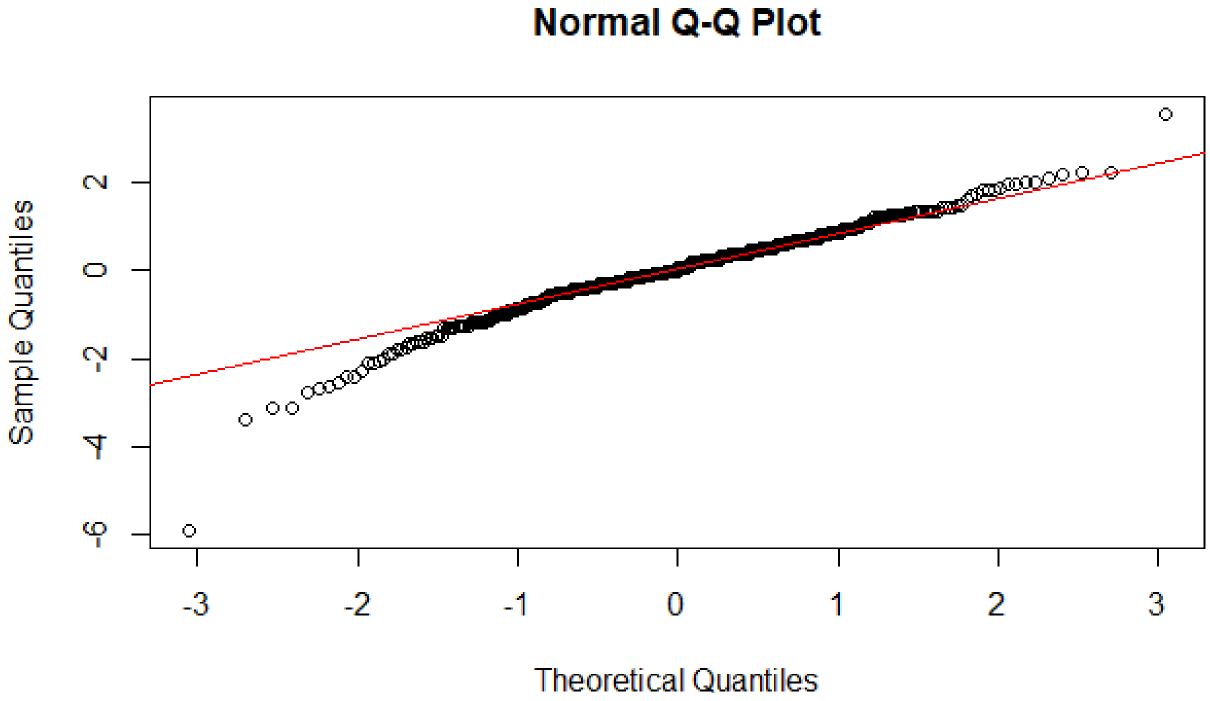


Figure 26: QQ-plot looks reasonable; most of the points "stick" to the line.

Predictive modeling on data (c) with random forests

Methodology

We follow the standard machine learning assumptions, i.e. the "Vapnik" formulation of statistical model training.

In short,

- We chose a model (linear regression, kNN, random forests, xgb-trees,etc.)
- Our training data is (c), our predictors are the PC components, our target is: (i) the win-draw-lose class , (ii) the total xG (sum of home and away xG). (Assuming that our model makes sense as a classifier/regressor of course).
- For (i), (ii) we apply l-o-o cross validation and find the hyperparameter with the least MME/generalization error. This step can potentially also be done using the "out-of-bag-error" in case of random forests.
- We apply the trained model on the test data (d) to get the final predictions.

See the book Bontempi (2020) for more details of the ML-formulation.

The model of our choice are *random forests.* This is because they are easy to "tune", can act as classifiers and predictors, and have an "built-in" variable importance assignment (the out-of-bag error).Rigorous consistency results have been established, but only for special cases (additive model) see Scornet et al. (2015). .

We won't describe random forests in detail here, a detailed explication with R coding in mind can be found in the book Genuer and Poggi (2020), or a simpler explanation in the previously mentioned Bontempi (2020). The o-o-b error is also discussed in the two; it can be used for tuning.

For convenience, we give a quick introduction similar to Bontempi (2020):

Decision trees, on their own, are weak learners. The solution is to combine many different decision trees and make them as uncorrelated as possible in order to reduce variance - this combination (sum) is called a **random forest**.

To be more precise, suppose we have B trees with negligible bias and comparable variance, and mutual correlation ρ. It can be shown that

$$\text{Variance of random forest} \propto \frac{1}{B} \times \rho$$

i.e., by either increasing the number of trees, or by decreasing the mutual correlation, we reduce the total variance.

The random forest algorithm:

- generates a bootstrap sample (with replacement) of the data B times,
- fits each decision tree $b = 1, \ldots, B$, where the set of variables for each split is a random subset (of size n'),
- stores at each split the improvement of the cost function,
- gives the final prediction as an average over B of all the decision trees,
- also gives an importance measure for each variable.

Some very nice heuristics about model tuning in general (in R) , and the random forest parameters in particular, can be found in the book Bartz et al. (2023).

From Bartz et al. (2023), the relevant hyperparameters for the random forest function in R :

- **Number of Trees**: `ntree` — the number of trees in the forest.
- **Tree Depth**: Controlled by:
  - `maxnodes` — the maximum number of terminal nodes (controls tree depth).
  - `nodesize` — the minimum number of observations in a terminal node.
- **Features**: `mtry` — the number of features randomly selected at each split.
- **Subsample Size**: `samplesize` — the number of observations sampled (with or without replacement) for each tree.

In general, it seems best to take the number of trees as high as your computer allows, and not to tune both the subsample size and tree depth, i.e. just tuning one of them while keeping the other fixed.

We will tune the model using the built in o-o-b error. In particular, it will be revealed that the tuning was not really necessary, in the sense that we get a negligible error decrease as compared to the default.

The final model uses the default hyperparameters. More precisely the defaults for the "rf()" function in R. So `ntree` = 500 , etc.

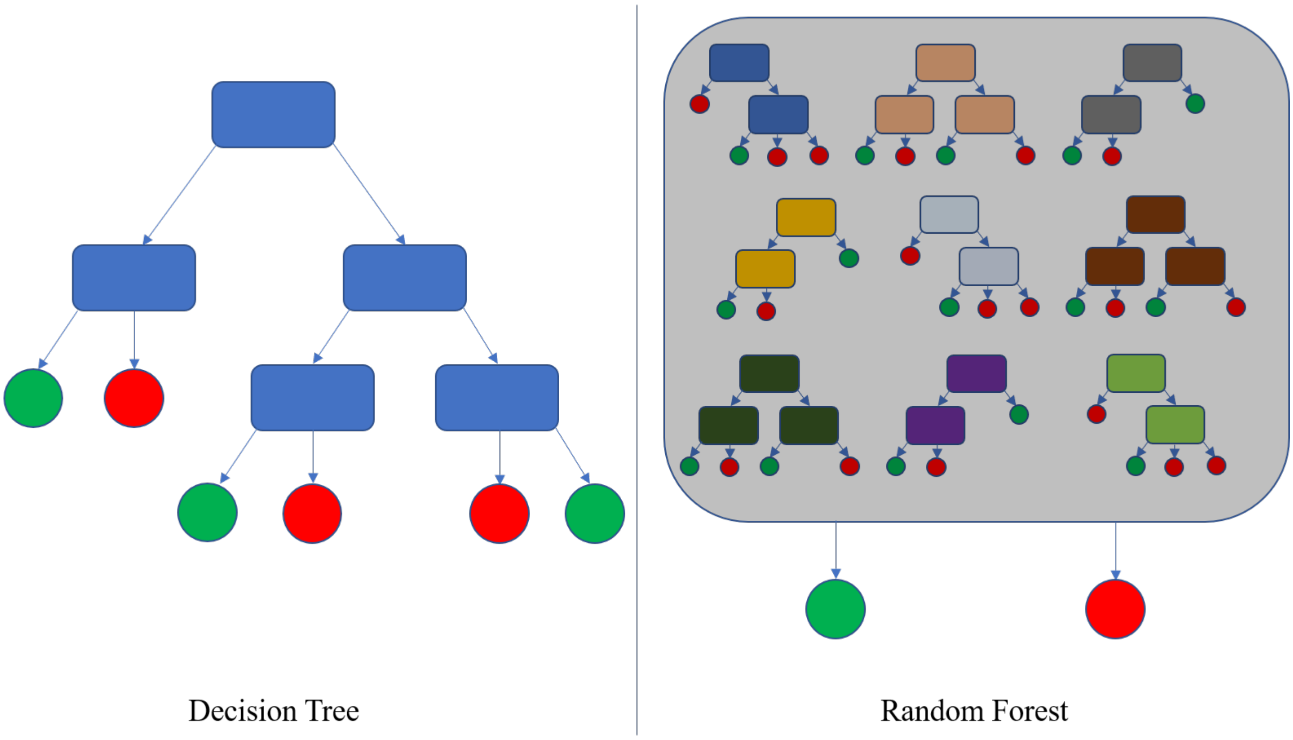


Figure 27: Random forest visual depiction. The interpretability of lone decision trees is sacrificed in favor of the predictive power of the "forest". Taken from Wikipedia.

Classification

Home	**Away**	**Date**	$p_{\mathbf{win}}$	$p_{\mathbf{draw}}$	$p_{\mathbf{loss}}$
Tottenham	Nott'ham Forest	2025-04-21	0.516	0.270	0.214
Manchester City	Aston Villa	2025-04-22	0.690	0.108	0.202
Arsenal	Crystal Palace	2025-04-23	0.528	0.352	0.120
Chelsea	Everton	2025-04-26	0.512	0.402	0.086
Southampton	Fulham	2025-04-26	0.068	0.114	0.818
Newcastle Utd	Ipswich Town	2025-04-26	0.600	0.200	0.200
Wolves	Leicester City	2025-04-26	0.580	0.142	0.278
Brighton	West Ham	2025-04-26	0.366	0.342	0.292
Bournemouth	Manchester Utd	2025-04-27	0.444	0.212	0.344
Liverpool	Tottenham	2025-04-27	0.754	0.088	0.158
Nott'ham Forest	Brentford	2025-05-01	0.490	0.138	0.372
Manchester City	Wolves	2025-05-02	0.762	0.140	0.098
Arsenal	Bournemouth	2025-05-03	0.672	0.286	0.042
Aston Villa	Fulham	2025-05-03	0.528	0.296	0.176
Everton	Ipswich Town	2025-05-03	0.400	0.412	0.188
Leicester City	Southampton	2025-05-03	0.272	0.322	0.406
West Ham	Tottenham	2025-05-03	0.468	0.182	0.350
Chelsea	Liverpool	2025-05-04	0.076	0.476	0.448
Brentford	Manchester Utd	2025-05-04	0.524	0.266	0.210
Brighton	Newcastle Utd	2025-05-04	0.414	0.278	0.308
Crystal Palace	Nott'ham Forest	2025-05-05	0.148	0.206	0.646
Bournemouth	Aston Villa	2025-05-10	0.438	0.080	0.482
Ipswich Town	Brentford	2025-05-10	0.022	0.222	0.756
Wolves	Brighton	2025-05-10	0.404	0.128	0.468
Newcastle Utd	Chelsea	2025-05-10	0.386	0.322	0.292
Tottenham	Crystal Palace	2025-05-10	0.284	0.326	0.390
Fulham	Everton	2025-05-10	0.356	0.472	0.172
Southampton	Manchester City	2025-05-10	0.054	0.040	0.906
Manchester Utd	West Ham	2025-05-10	0.460	0.240	0.300
Liverpool	Arsenal	2025-05-11	0.372	0.494	0.134
Nott'ham Forest	Leicester City	2025-05-11	0.738	0.106	0.156
Chelsea	Manchester Utd	2025-05-16	0.456	0.340	0.204
Manchester City	Bournemouth	2025-05-18	0.466	0.370	0.164
Brentford	Fulham	2025-05-18	0.352	0.238	0.410
Leicester City	Ipswich Town	2025-05-18	0.272	0.470	0.258
Arsenal	Newcastle Utd	2025-05-18	0.700	0.214	0.086
West Ham	Nott'ham Forest	2025-05-18	0.414	0.140	0.446
Everton	Southampton	2025-05-18	0.558	0.278	0.164
Aston Villa	Tottenham	2025-05-18	0.662	0.182	0.156
Crystal Palace	Wolves	2025-05-18	0.318	0.314	0.368
Brighton	Liverpool	2025-05-19	0.194	0.418	0.388
Southampton	Arsenal	2025-05-25	0.074	0.070	0.856
Manchester Utd	Aston Villa	2025-05-25	0.446	0.096	0.458
Wolves	Brentford	2025-05-25	0.348	0.162	0.490
Tottenham	Brighton	2025-05-25	0.400	0.280	0.320
Nott'ham Forest	Chelsea	2025-05-25	0.418	0.396	0.186
Liverpool	Crystal Palace	2025-05-25	0.604	0.254	0.142
Newcastle Utd	Everton	2025-05-25	0.564	0.194	0.242
Bournemouth	Leicester City	2025-05-25	0.610	0.134	0.256
Fulham	Manchester City	2025-05-25	0.310	0.338	0.352
Ipswich Town	West Ham	2025-05-25	0.244	0.278	0.478

Table 3: Win-draw-lose probabilities (p_{win}, p_{draw} , p_{lose}) for the home side, for the remaining matches in the 24/25 EPL season; produced by our *random forest classifier*. The R-code can be found in the appendix. There are 35 matches remaining (as of the 21st of April) and the probabilities are very reasonable.

Regression

Home	Away	Date	total_xG
Tottenham	Nott'ham Forest	2025-04-21	3.387371
Manchester City	Aston Villa	2025-04-22	3.477398
Arsenal	Crystal Palace	2025-04-23	2.379684
Chelsea	Everton	2025-04-26	2.932179
Southampton	Fulham	2025-04-26	3.131498
Newcastle Utd	Ipswich Town	2025-04-26	3.290784
Wolves	Leicester City	2025-04-26	2.890796
Brighton	West Ham	2025-04-26	2.761888
Bournemouth	Manchester Utd	2025-04-27	2.672122
Liverpool	Tottenham	2025-04-27	3.326949
Nott'ham Forest	Brentford	2025-05-01	2.222263
Manchester City	Wolves	2025-05-02	3.397278
Arsenal	Bournemouth	2025-05-03	2.730782
Aston Villa	Fulham	2025-05-03	2.773393
Everton	Ipswich Town	2025-05-03	2.993500
Leicester City	Southampton	2025-05-03	2.413821
West Ham	Tottenham	2025-05-03	3.048535
Chelsea	Liverpool	2025-05-04	3.162260
Brentford	Manchester Utd	2025-05-04	2.871305
Brighton	Newcastle Utd	2025-05-04	2.768877
Crystal Palace	Nott'ham Forest	2025-05-05	2.587943
Bournemouth	Aston Villa	2025-05-10	3.003100
Ipswich Town	Brentford	2025-05-10	2.321794
Wolves	Brighton	2025-05-10	2.913702
Newcastle Utd	Chelsea	2025-05-10	3.072687
Tottenham	Crystal Palace	2025-05-10	3.791163
Fulham	Everton	2025-05-10	2.258398
Southampton	Manchester City	2025-05-10	4.388700
Manchester Utd	West Ham	2025-05-10	2.611000
Liverpool	Arsenal	2025-05-11	2.947274
Nott'ham Forest	Leicester City	2025-05-11	2.651472
Chelsea	Manchester Utd	2025-05-16	3.188172
Manchester City	Bournemouth	2025-05-18	3.509709
Brentford	Fulham	2025-05-18	2.754651
Leicester City	Ipswich Town	2025-05-18	3.392650
Arsenal	Newcastle Utd	2025-05-18	2.353877
West Ham	Nott'ham Forest	2025-05-18	2.681606
Everton	Southampton	2025-05-18	2.409886
Aston Villa	Tottenham	2025-05-18	3.079465
Crystal Palace	Wolves	2025-05-18	2.616853
Brighton	Liverpool	2025-05-19	3.122958
Southampton	Arsenal	2025-05-25	3.443691
Manchester Utd	Aston Villa	2025-05-25	3.115733
Wolves	Brentford	2025-05-25	2.667477
Tottenham	Brighton	2025-05-25	3.623726
Nott'ham Forest	Chelsea	2025-05-25	2.436500
Liverpool	Crystal Palace	2025-05-25	3.184077
Newcastle Utd	Everton	2025-05-25	2.823159
Bournemouth	Leicester City	2025-05-25	3.264138
Fulham	Manchester City	2025-05-25	2.610190
Ipswich Town	West Ham	2025-05-25	2.572928

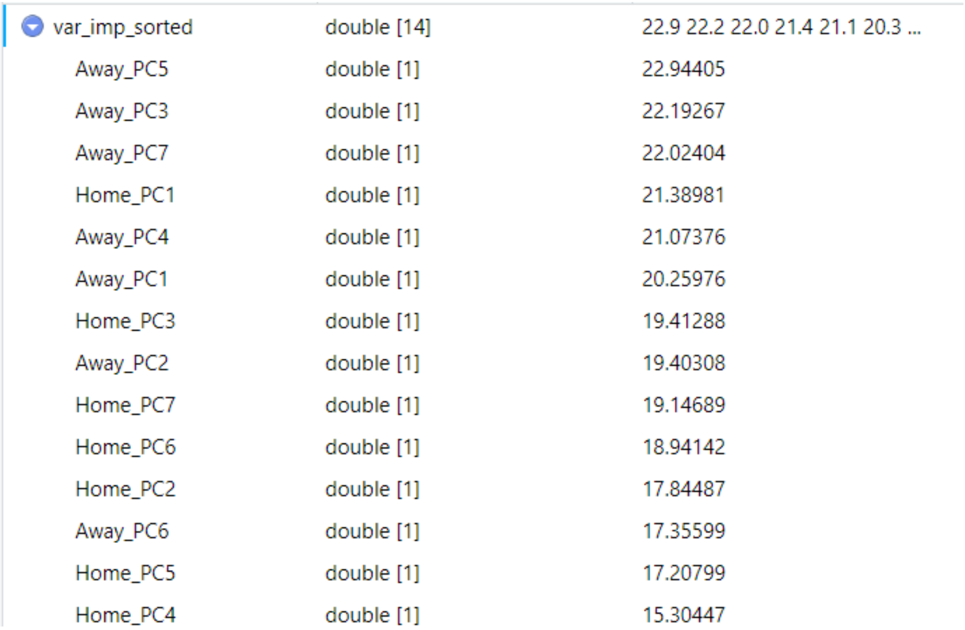
Table 4: Predicted total xG values for remaining matches in the 2024/25 EPL season.

Variable importance with o-o-b error

The *out-of-bag-error* of the random forest makes it possible to quantify *variable importance* (equivalently the predictive power) of each of our predictors. We expect the "lesser" PCs (e.g. PC7) to be less important; one can possibly do without them.

Alternatively, given that our training data set (c) is small (about ≈ 300 rows with 14 columns), we can directly work with $35 + 35 = 70$ attributes, train the model, assign importance to each attribute, then eliminate the unimportant ones. This is easy, since the actual tuning process for random forests can be "neglected", in the sense that usually taking the default number of trees ($n = 500$) will lead to a generalization error G such that going beyond n does not alter G very much (the other hyperparameters also don't require tuning). We did not pursue this approach.

In determining whether the home team will win/draw/lose the o-o-b error suggests the away's team PC stats (in order: PC5,PC3,PC7,PC4,etc.) as most important, along side with the home teams PC1 (which from our previously analysis represents "brute strength"). None of the 14 PC are unimportant in the sense that they can be neglected. Least important seem to be the home PC stats (PC4,PC5 being bottom two).



```
var_imp_sorted    double [14]    22.9 22.2 22.0 21.4 21.1 20.3 ...
    Away_PC5      double [1]     22.94405
    Away_PC3      double [1]     22.19267
    Away_PC7      double [1]     22.02404
    Home_PC1      double [1]     21.38981
    Away_PC4      double [1]     21.07376
    Away_PC1      double [1]     20.25976
    Home_PC3      double [1]     19.41288
    Away_PC2      double [1]     19.40308
    Home_PC7      double [1]     19.14689
    Home_PC6      double [1]     18.94142
    Home_PC2      double [1]     17.84487
    Away_PC6      double [1]     17.35599
    Home_PC5      double [1]     17.20799
    Home_PC4      double [1]     15.30447
```

Figure 28: Print from R studio. Our intuition that only PC1,PC2 matter was incorrect.

This merits a further investigation on the interpretation of the PCs. From the eigenvector decomposition done in the PCA part, we find:

$$PC3 = X_3 \approx -0.5(Y_{\text{speed}} + Y_{\text{acceleration}}) - 0.25(Y_{\text{agility}} + Y_{\text{balance}} + Y_{\text{stamina}}) + \dots$$

$$PC4 = X_4 \approx 0.25(Y_{\text{jumping}} + Y_{\text{heading}} + Y_{\text{composure}}) - 0.30(Y_{\text{work rate}} + Y_{\text{crossing}} + Y_{\text{stamina}}) + \dots$$

$$PC5 = X_5 \approx 0.5(Y_{\text{finishing}} + Y_{\text{penalties}} + Y_{\text{long shots}}) + \dots$$

$$PC6 = X_6 \approx 0.2(Y_{\text{team work}} + Y_{\text{work rate}} + Y_{\text{anticipation}}) - 0.3(Y_{\text{long throws}} + Y_{\text{crossing}} + Y_{\text{corners}} + Y_{\text{free kicks}}) \dots$$

$$PC7 = X_7 \approx -0.5(Y_{\text{aggression}} + Y_{\text{determination}}) + \dots$$

(We ignored the smaller weights)

These we interpret as follows:

$$PC3 \iff \text{raw speed}$$

$$PC4 \iff \text{aerial threat sustainability}$$

$$PC5 \iff \text{goal threat}$$

$$PC6 \iff \text{team cohesion minus set piece reliance}$$

$$PC7 \iff \text{intensity of play}$$

That said, it seems obvious that: **the opponents "goal threat" (PC5), the opponents "raw speed" (PC3), the opponents "intensity of play" (PC7) are the best predictors of the *home teams* win-draw-lose probabilities.**

An application to sports gambling

Let S_0 be our starting wealth, we adopt the *multiplicative dynamic*, i.e. the dynamic defined by:

$$S_n = S_0 \prod_{i=1}^{n} Y_j(\boldsymbol{b})$$

where $Y(\boldsymbol{b}) = \boldsymbol{b'X}$ are betting rounds , X_i are individual binary bets (as described in the abstract), $\boldsymbol{b}$ are betting fractions contained in some simplex.

The X_i we can take as the win-lose-draw bets of the form

$$X = \begin{cases} \tilde{P} \text{ prob. } p \\ 0 \text{ prob. } 1 - p \end{cases}$$

with the $\tilde{P}$ corresponding to the betting casino odds.

Using our *random classifier probabilities* we identify the *weakest casino* , in the sense that $\tilde{P}p = E[X]$ is the biggest (also > 1 to be favorable) . So for example: in the first game Tottenham vs Nott'ham Forest we identify weakest casino with respect to the predicted probabilities. We take that casino as the bet X_i. Depending on the date, there are simultaneous games played on the same day, in that case we take the sum of these X_i as the bet. For each simultaneous round $j = 1, \ldots, m$ we form

$$Y_j(\boldsymbol{b}) = \boldsymbol{b'X} = b_1 X_1 + \ldots b_k X_{k-1} + b_k 1$$

where b_k corresponds to the fraction we keep, i.e. the cash we don't gamble. $\boldsymbol{b}$ acts as a free parameter representing the fraction of our wealth we gamble. The inclusion of b_s prevents ruin (but does not prevent asymptotic ruin).

Assuming the $Y_j(\boldsymbol{b})$ to be independent, **Algoet's theorem** guarantees that the optimal betting fractions $\boldsymbol{b}_j$ for each round j is:

$$\boldsymbol{b}_j^* = \sup_{\boldsymbol{b} \in \Delta^k} E[\log Y_j(\boldsymbol{b})]$$

With $k - 1$ equal to the number of X_i taken at round j and Δ^k the k-dimensional simplex. (the k the X_i is taken to be $X_k = 1$ - the "safety bet").

So, for example, on round $j = 1$ there is only one game played : Tottenham vs Nott'ham Forestm, so for this round $k = 1$. The weakest casino concerns the event "Tottenham to win" ,has odds $\tilde{P} = 2.5$ and our random forest probability is $p_{win} = 0.5$. The round is

$$Y(\boldsymbol{b}) = b_1 X + b_2 = \begin{cases} b_1 \tilde{P} + b_2 & \text{prob. } p_{win} \\ b_2 & \text{prob. } 1 - p_{win} \end{cases}$$

The expected log maximization in that case has the simple solution

$$b_1^* = \frac{\tilde{P} p_{win} - 1}{\tilde{P} - 1}, \quad b_2^* = 1 - b_1^*$$

which for our values gives $b_1^* \approx 0.3$.

This goes on for other dates j and each time we gamble the optimal fractions as given by Algolet's theorem. This strategy is guaranteed to asymptotically dominate any other competing strategy. If for a certain j the number of X_i k is greater than 1 the optimization problem becomes non-trivial, in the sense that there isn't an easy analytic solution and we must solve it numerically.

As a more complicated example, we take $j = 3$ - the round to be played on the 26th of April, i.e. we the consider $k = 5$ games (the X_i):

Home	**Away**	**Date**	$p_{\textbf{win}}$	$p_{\textbf{draw}}$	$p_{\textbf{loss}}$
Chelsea	Everton	2025-04-26	0.512	0.402	0.086
Southampton	Fulham	2025-04-26	0.068	0.114	0.818
Newcastle Utd	Ipswich Town	2025-04-26	0.600	0.200	0.200
Wolves	Leicester City	2025-04-26	0.580	0.142	0.278
Brighton	West Ham	2025-04-26	0.366	0.342	0.292

Table 5: The $k - 1 = 5$ EPL games to be played on the 26th of April, these are assembled into binary bets X_i. The $k + 1$th bet is the safety $X_{k+1} = 1$.

We identify the most favorable casino bets as:

$$\text{"Chelsea vs Everton to end in draw"} \iff X_1 = \begin{cases} \tilde{P_1} = 4.20 & \text{with probability} \quad p_{\text{draw}} = 0.40 \\ 0 & \text{with probability} \quad 1 - p_{\text{draw}} \end{cases}$$

$$\text{"Southampton to lose to Fulham"} \iff X_2 = \begin{cases} \tilde{P_1} = 1.68 & \text{with probability} \quad p_{\text{lose}} = 0.82 \\ 0 & \text{with probability} \quad 1 - p_{\text{lose}} \end{cases}$$

$$\text{"Newcastle to lose to Ipswich"} \iff X_3 = \begin{cases} \tilde{P_3} = 14.00 & \text{with probability} \quad p_{\text{lose}} = 0.20 \\ 0 & \text{with probability} \quad 1 - p_{\text{lose}} \end{cases}$$

$$\text{"Wolves to lose to Leicester"} \iff X_4 = \begin{cases} \tilde{P_4} = 6.25 & \text{with probability} \quad p_{\text{lose}} = 0.28 \\ 0 & \text{with probability} \quad 1 - p_{\text{lose}} \end{cases}$$

$$\text{"Brighton to draw West Ham"} \iff X_5 = \begin{cases} \tilde{P_4} = 4.20 & \text{with probability} \quad p_{\text{draw}} = 0.34 \\ 0 & \text{with probability} \quad 1 - p_{\text{draw}} \end{cases}$$

We then form the round

$$Y(\boldsymbol{b}) = \boldsymbol{b}'\boldsymbol{X} = (b_1, \dots, b_6)' \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_5 \\ 1 \end{pmatrix} = \left(\sum_{i=1}^{5} b_i X_i \right) + b_6$$

Explicitly $Y(\boldsymbol{b})$ is the random variable with 2^5 outcomes:

$$Y(\boldsymbol{b}) = \begin{cases} b_6 & \text{with probability} & \prod_{i=1}^{5}(1 - p_i) \\ P_1 b_1 + b_6 & \text{with probability} & p_1 \prod_{\{i \neq 1\}}^{5}(1 - p_i) \\ P_2 b_1 + b_6 & \text{with probability} & p_2 \prod_{\{i \neq 2\}}^{5}(1 - p_i) \\ P_1 b_1 + P_2 b_2 + b_6 & \text{with probability} & p_1 p_2 \prod_{\{i \neq 1,2\}}^{5}(1 - p_i) \\ P_3 b_3 + b_6 & \text{with probability} & p_3 \prod_{\{i \neq 3\}}^{5}(1 - p_i) \\ P_1 b_1 + P_2 b_2 + P_3 b_3 + b_6 & \text{with probability} & p_1 p_2 p_3 \prod_{\{i \neq 1,2,3\}}^{5}(1 - p_i) \\ \vdots & \vdots & \vdots \\ P_1 b_1 + P_2 b_2 + P_3 b_3 + P_4 b_4 + P_5 b_5 + b_6 & \text{with probability} & p_1 p_2 p_3 p_4 p_5 \end{cases}$$

And we must find $\boldsymbol{b}$ that maximizes $E[\log Y]$ under the constraint $\boldsymbol{b} \in \Delta^5$:

$$\sum_{s=1}^{32} p_s \log P_s \quad , s = \text{ "possible outcome"}, P_s \text{ the corresponding payout}$$

We apply the method above for a starting wealth of $S_0 = 100$eur. The final j of the league is on the 25th of May ; by the time of the presentation we will have the result of the gambling application.

More details about this gambling setup can be found in Algoet and Cover (1988), Thorp (2006) , Busseti et al. (2016),Breiman (1962) or in my MA-thesis (which can be provided at request).

Note: gambling on the xG predictions is more difficult as we aren't in possession of the "true" probabilities. xG gambling would require of us to make additional assumptions as to make it compatible with the above scheme. Alternatively , one could use the xG as "assistance" to guide one in any betting endeavor.

Appendix - Pretty graphs

Bubble graphs

These graphs were created with the UI tools available in SAS ; so no manual programming.

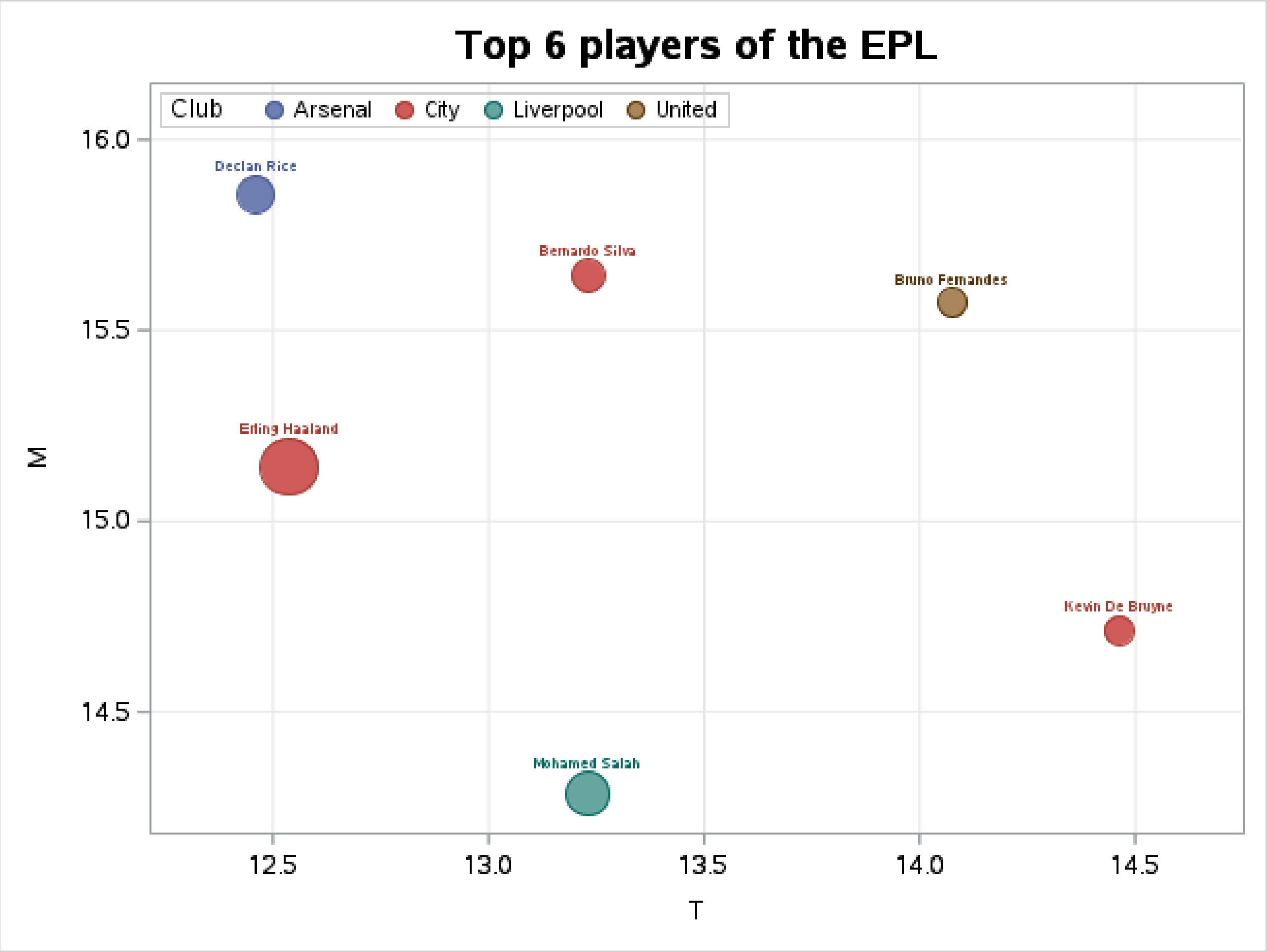


Figure 29: T is the technical mean, M the mental mean , P the physical mean is proportional to the radius of the disk. The players were found with the UI filter $T + M + P > 42.5$.

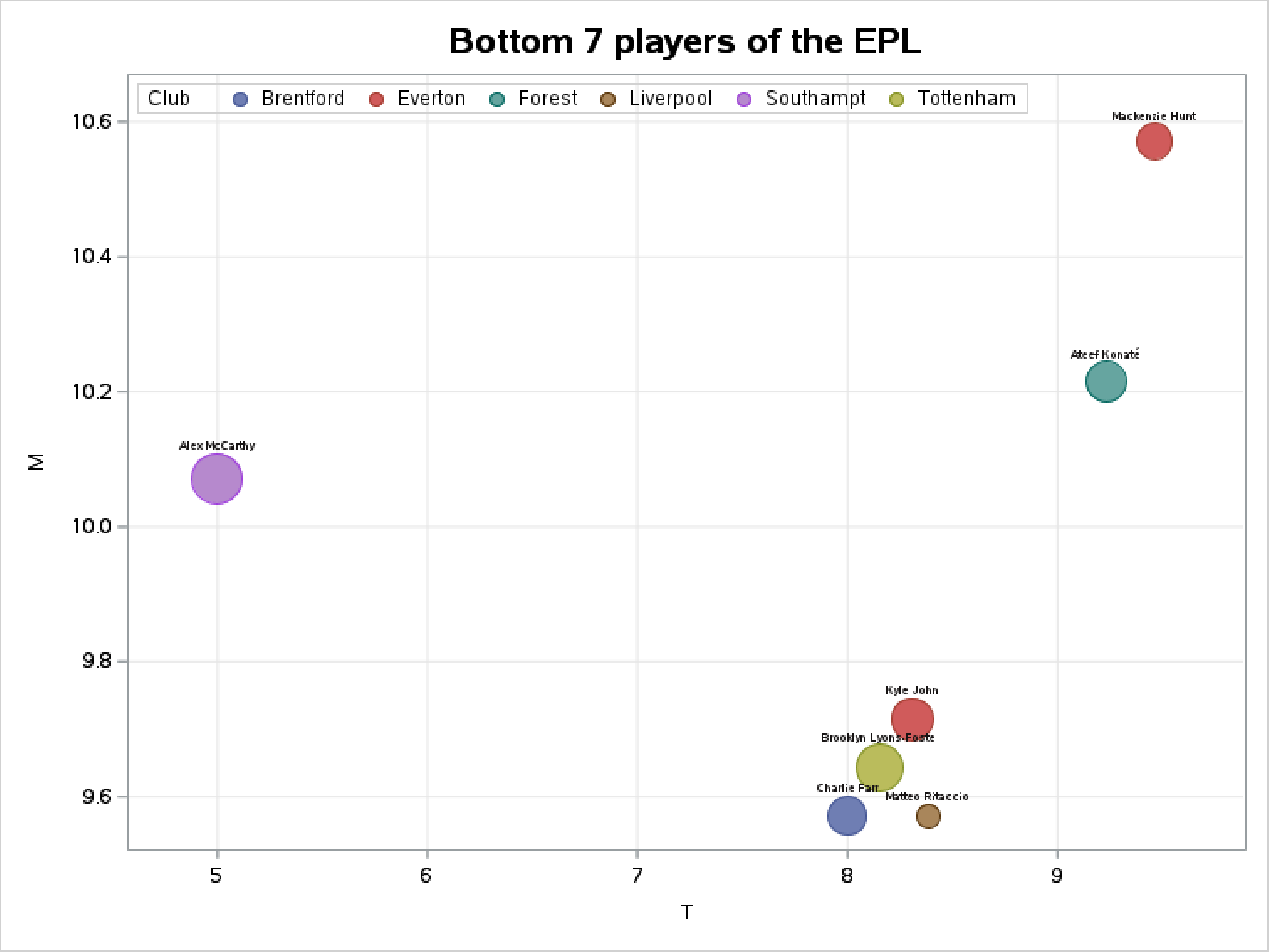


Figure 30: Bottom 7 players of the league; found with the restriction $T + M + P < 28.5$. The worst player is Ritaccio who appeared as an outlier in our ANOVA analysis.

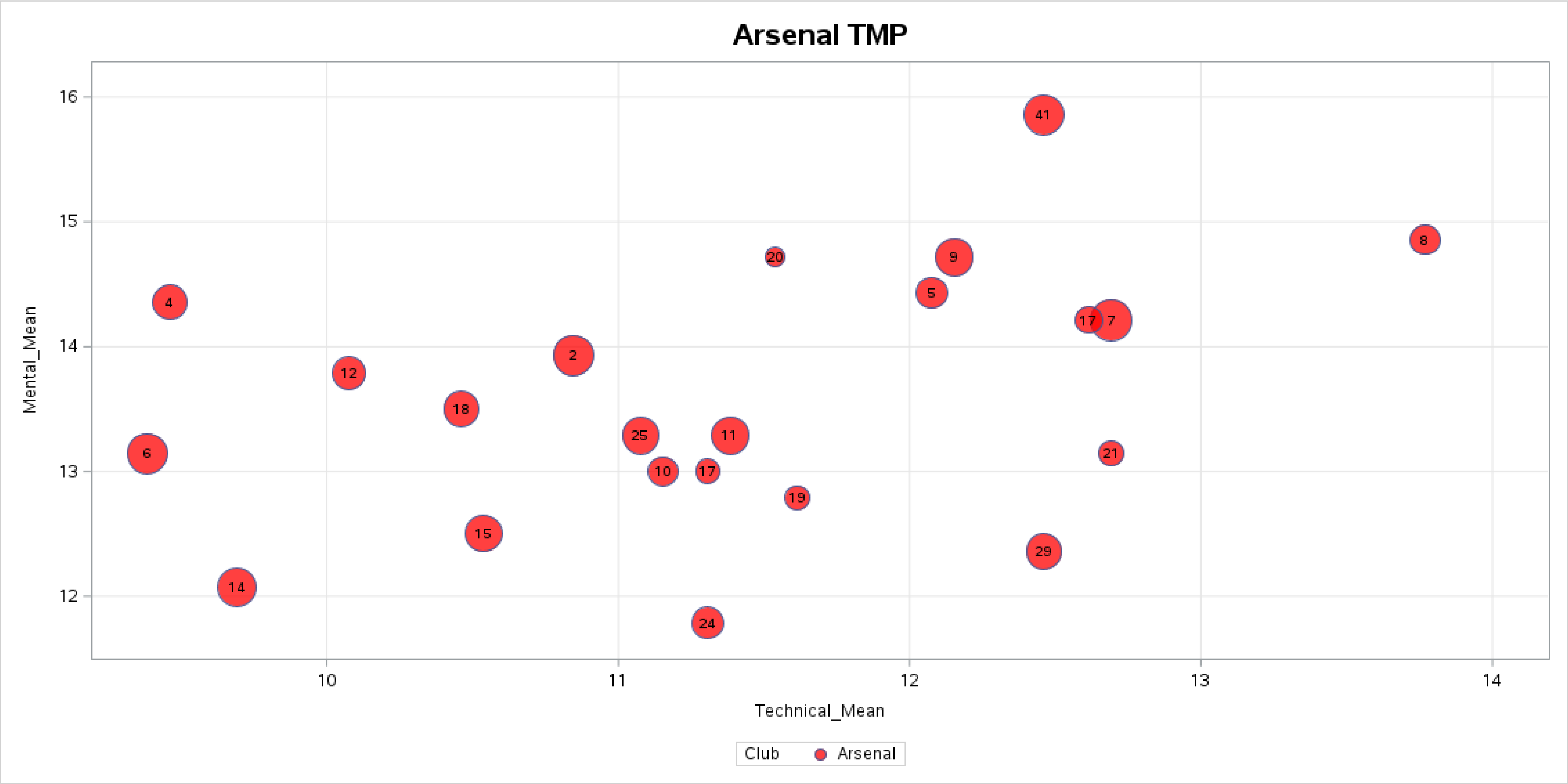


Figure 31: 17 Cédric, 25 Elneny, 6 Gabriel, 9 Gabriel Jesus, 29 Havertz, 20 Jorginho, 15 Kiwior, 11 Martinelli, 24 Nelson, 14 Nketiah, 8 Ødegaard, 5 Partey, 41 Rice, 7 Saka, 2 Saliba, 10 Smith Rowe, 12 Timber, 18 Tomiyasu, 19 Trossard, 21 Vieira, 4 White, 17 Zinchenko. To push the bubble plot to its maximum: one could reserve the color for another variable (height say), the players and club would be represented as "number,letter"; for example Declan Rice, Arsenal would be 41A.

References

P. H. Algoet and T. M. Cover. Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *The Annals of Probability*, 16(2), 1988. doi: 10.1214/aop/1176991793. URL `https://doi.org/10.1214/aop/1176991793`.

Eva Bartz, Thomas Bartz-Beielstein, Martin Zaefferer, and Olaf Mersmann, editors. *Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide*. Springer, Cham, 2023. doi: 10.1007/978-981-19-5170-1.

Gianluca Bontempi. *Statistical Foundations of Machine Learning: The Book*. ULB–Machine Learning Group, Brussels, 2020.

L. Breiman. Optimal gambling systems for favorable games, 1962. URL `https://doi.org/10.21236/ad0402290`.

E. Busseti, E. K. Ryu, and S. Boyd. Risk-constrained kelly gambling. *The Journal of Investing*, 25(3):118–134, 2016. doi: 10.3905/joi.2016.25.3.118. URL `https://doi.org/10.3905/joi.2016.25.3.118`.

Robin Genuer and Jean-Michel Poggi. *Random Forests with R*. Springer, London, 2020. doi: 10.1007/978-3-030-45051-6.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015. doi: 10.1214/15-AOS1321.

E. O. Thorp. Chapter 9: The kelly criterion in blackjack, sports betting, and the stock market. In *Handbook of Asset and Liability Management*, pages 385–428. Elsevier, 2006. doi: 10.1016/S1872-0978(06)01009-X. URL `https://doi.org/10.1016/S1872-0978(06)01009-X`.