Max Mitchell
MSDS 456

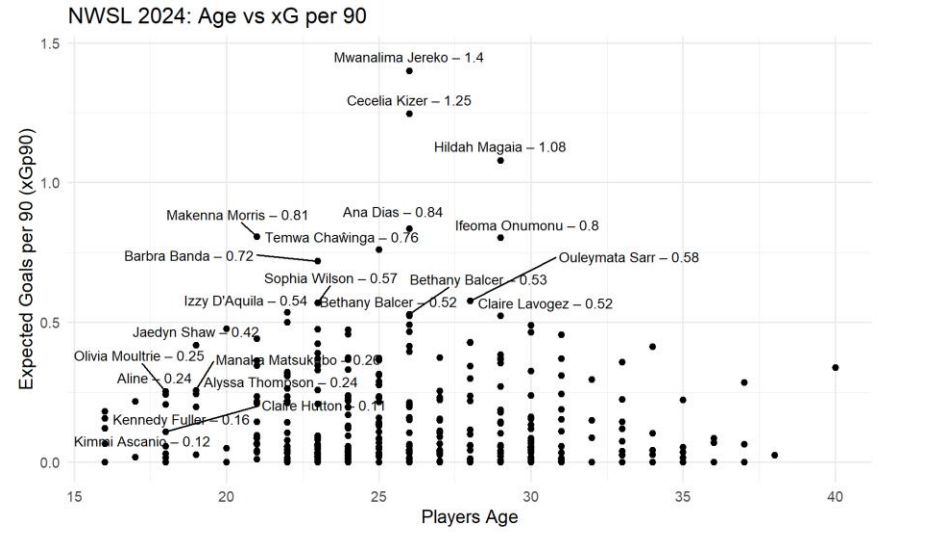
Module 7: Coding Assignment 2

Scenario 1: Data Collection and summarization

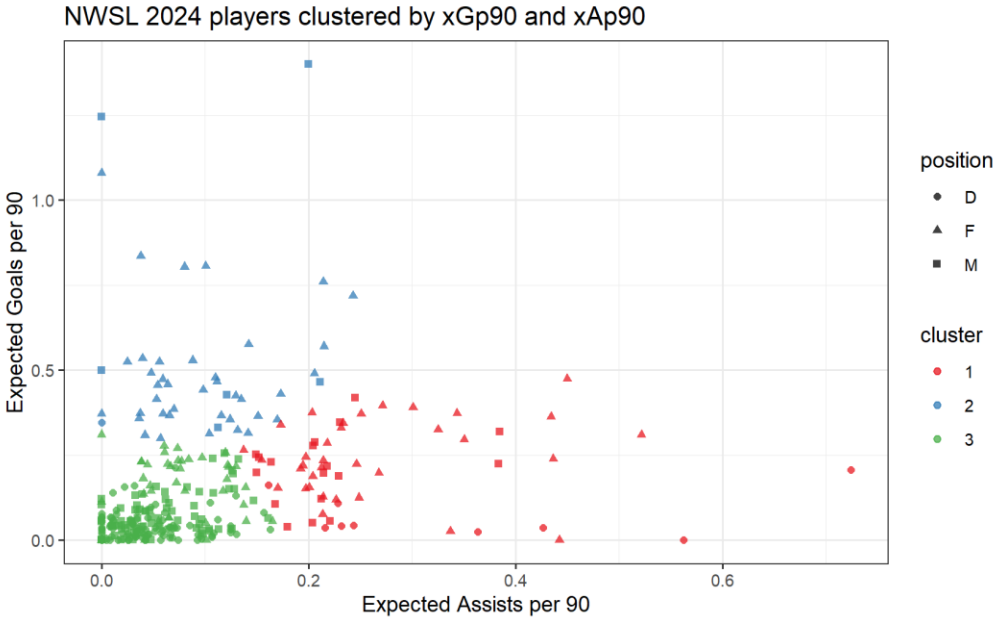
Similar to the nwsl_player_stats file within the ISAR package, I collected data for the 2024 NWSL season via FBref. Although this process was somewhat difficult as I had trouble finding the right table to scrape, as my initial code kept gathering squad related data instead of player, I was eventually able to output the 2024 NWSL player data and make a corresponding master excel file. To do this, instead of writing out a scrape, I simply copied in the entire 2024 NWSL player stats table into an excel spreadsheet, renamed the columns to be the exact same as that in the nwsl_player_stats spreadsheet already in ISAR for the 2022 NWSL season, and then begun subsequent analysis. Within this dataset contains 376 players across the 14 teams. The variables contained in this dataset that I will be using for the remainder of the assignment are player (players name), squad (the NWSL team), pos (players position), age (players age), min (total minutes players), starts (games started), xG (expected goals), xAG (expected assists), xGp90 (expected goals per 90 minutes), xAp90 (expected assists per 90), and xGxAp90 (sum of expected goals and assists per 90 minutes).

The EPL 2024 goalkeeping stats are already downloaded through ISAR and didn't require additional scraping. There are 44 goalkeepers in the dataset who played during the 2024/2025 Premier League season across all 20 teams. The variables used for this assignment include player (goalkeeper's name), nation (goalkeepers nationality),cPSxG/SoT (post-shot expected goals per shot on target), PSxG+/minus(Post-shot expected goals minus goals allowed), LaunchPCT(Launch completion percentage, Att (attempted passes), AvgLen (Average pass length), GK_LaunchPCT (Goal kick launch percentage), and DefActions_OPA_90 (Number of defensive actions outside penalty area per 90 minutes).

Scenario 2: Clustering NWSL Players
Updated Figure 8.3:



Updated Figure 8.10:



Within the updated 8.10 figure, it looks like a high proportion of defenders are in cluster 3 which have low amounts of expected goals per 90 and assists per 90. There is then a higher proportion of forwards in group two, who have a higher expected goals per 90 versus expected assist per 90. Lastly, there are a lot of midfielders in cluster 1, as well as forwards, who have a higher rate of expected assists per 90 than expected goals per 90.

Table summarizing # of athletes assigned to each position and cluster

clusters_three <int>	D <int>	F <int>	M <int>
1	9	34	18
2	1	39	6
3	112	46	85

The table above confirms my prediction that cluster 3 is dominated by defenders. However, not all midfielders and forwards can score goals and assists, leading a solid proportion of them also in cluster 3. As forwards are the highest up the pitch, they dominate clusters 1 and 2, which contain players with higher rates of expected goals and assists per 90. Also, as cluster 1 is dominated by players with higher rates of expected assist per 90, there are 12 more midfielders in cluster 1 than cluster 2.

Updated Figure 8.11:

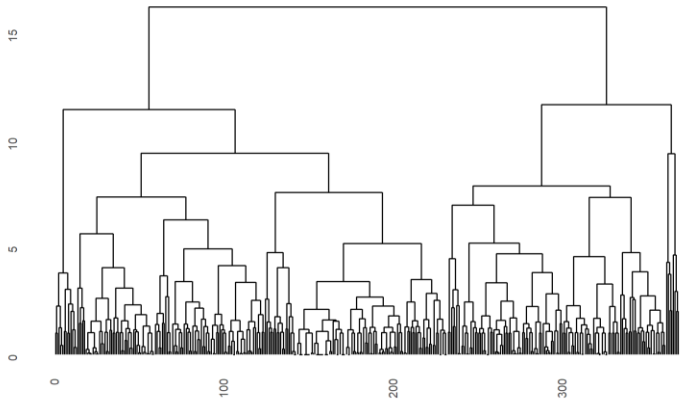


Table summarizing # of players assigned to each cluster:

cluster_k5 <int>	cluster_k4 <int>	count <int>
1	1	129
2	2	108
3	2	111
4	3	13
5	4	8

As shown in the table above, some clusters overlapped, leading to a reduction in assignments. This main change and overlap occurred during the reduction of the clusters from 5 to 4, resulting in the merging of clusters 2 and 3 (k=5) into a single cluster 2 (k=4). This merge combined 108 and 111 players from two separate k=5 clusters into a single k=4 cluster. The remaining clusters maintained a one-to-one ratio: k=5 cluster 1 became k=4 cluster 1, k=5 cluster 4 became k=4 cluster 3, and k=5 cluster 5 became k=4 cluster 4 (8 players). The only difference is that k=4 cluster 2 is now a combination of k=5 cluster 2 and 3. This consolidation reduced the total cluster count, with the most notable shift occurring between clusters 2 and 3.

Scenario 3: Evaluating EPL Goalkeepers

Recreate 8.9 step 1:
Table Summarizing Cluster Means for k=2

cluster <fctr>	PSxG_SoT <dbl>	PSxG_minus <dbl>	LaunchPct <dbl>	Att <dbl>	AvgLen <dbl>	GK_LaunchPct <dbl>	DefActions_OPA_90 <dbl>
1	0.3232143	0.425	32.87143	94.82143	31.48571	52.09643	1.110357
2	0.3033333	0.160	37.19333	499.00000	34.04667	62.44000	1.148667

Table Summarizing Cluster Means for k=3

cluster <fctr>	PSxG_SoT <dbl>	PSxG_minus <dbl>	LaunchPct <dbl>	Att <dbl>	AvgLen <dbl>	GK_LaunchPct <dbl>	DefActions_OPA_90 <dbl>
1	0.3040000	1.6600000	48.60000	673.60000	38.66000	69.24000	0.894000
2	0.3009091	-0.3727273	30.96364	397.63636	31.64545	58.72727	1.368182
3	0.3248148	0.3740741	33.13704	88.81481	31.51481	51.96667	1.066667

3 rows

Percentage of Variation Explained:

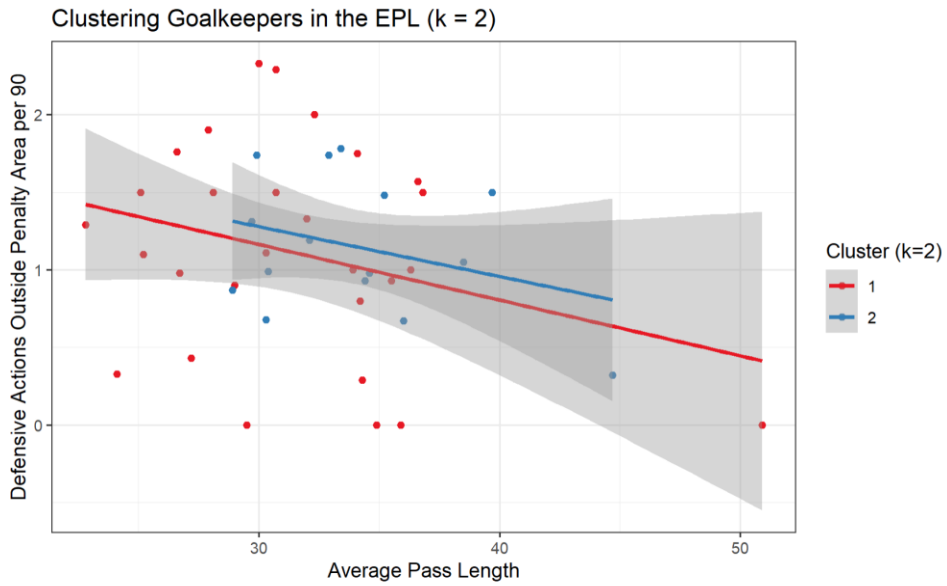
k = 2: 77.78 %
k = 3: 89.26 %
Adding a third cluster captures roughly 90% of the variation, whereas including just two clusters captures on 78% of the variance in the goalkeepers data.

Recreate 8.9 step 2:
Updated dataframe includes player name, nation, and k=2 and k=3 cluster

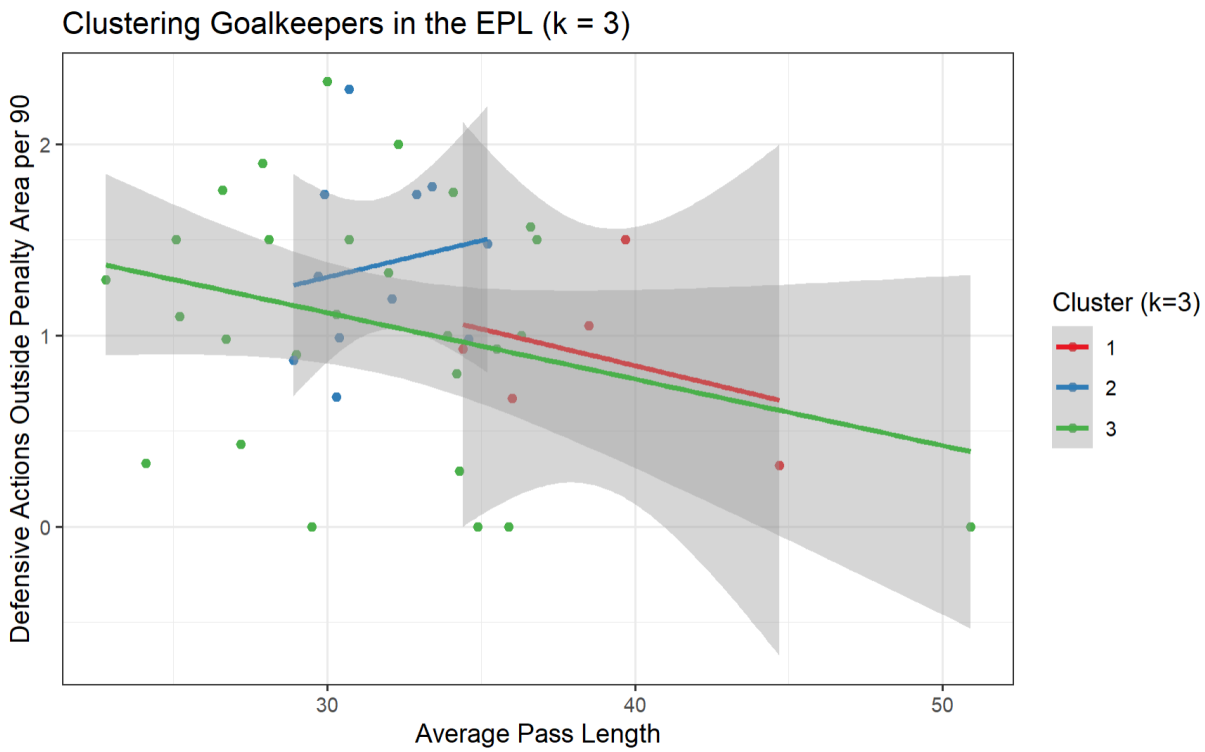
Player	Nation	cluster_k2	cluster_k3
	<chr>	<int>	<int>
Nick Pope	eng ENG	1	2
Alisson	br BRA	1	3
Brandon Austin	eng ENG	1	3
Altay Bayındır	tr TUR	1	3
Daniel Bentley	eng ENG	1	3
Martin Dúbravka	sk SVK	1	3
Ederson	br BRA	1	3
Łukasz Fabiański	pl POL	1	3
Fraser Forster	eng ENG	1	3
Vitezslav Jaros	cz CZE	1	3
Sam Johnstone	eng ENG	1	3
Filip Jørgensen	dk DEN	1	3
Caoimhin Kelleher	ie IRL	1	3
Antonín Kinský	cz CZE	1	3
Joe Lumley	eng ENG	1	3
Alex McCarthy	eng ENG	1	3
Arijanet Muric	xk KVX	1	3
Neto	br BRA	1	3
Robin Olsen	se SWE	1	3
Stefan Ortega	de GER	1	3
Alex Palmer	eng ENG	1	3
Hákon Rafn Valdimarsson	is ISL	1	3
Jason Steele	eng ENG	1	3
Jakub Stolarczyk	pl POL	1	3
Mark Travers	ie IRL	1	3
Guglielmo Vicario	it ITA	1	3
Christian Walton	eng ENG	1	3
Danny Ward	wls WAL	1	3
Mark Flekken	nl NED	2	1
Dean Henderson	eng ENG	2	1
Jordan Pickford	eng ENG	2	1
Aaron Ramsdale	eng ENG	2	1
Matz Sels	be BEL	2	1
Alphonse Areola	fr FRA	2	2
Kepa Arrizabalaga	es ESP	2	2
Mads Hermansen	dk DEN	2	2
Bernd Leno	de GER	2	2
Emiliano Martínez	ar ARG	2	2

1-38 of 43 rows Previous 1 2 Next

Recreate the second plot in Section 8.9 Step 4:
Plot for k=2:



Plot for k=3



Based on the two plots above clustering goalkeepers in the English Premier League for k = 2 and k =3, we can identify different types of goalkeepers in the EPL. First off, the k = 2 chart groups keepers into two clusters: one with short passes and more defensive actions and a second with longer passes and fewer defensive actions. One expects a player with longer passes to be a traditional goalkeeper who boots it up the fielder versus a "sweeper-keeper" who can play in possession and come out of the box to intercept and make short passes. I find Manuel Neuer to be the perfect example for this cluster. The k=3 chart groups keepers into 3 different clusters. This time, there is a more neutral keeper with a medium pass length and about average defensive actions outside the penalty area per 90. Again, there is a sweeper keeper and a more traditional player