Goalkeeper Clustering (2022-23 Season)

Eamon Gara Grady

c. (860) 406-2440

e. eamongaragrady@gmail.com

Background

Motivation

Identify and describe the main goalkeeper categories within Europe's top five leagues[1] to help clubs find potential transfer targets who fit their desired profile

1. Premier League (England), La Liga (Spain), Bundesliga (Germany), Serie A (Italy), Ligue 1 (France)

Why Goalkeepers?

The role of goalkeeper is the most consistent of all position groups

across clubs which facilitates easier comparison

Process

Process

- Gather, clean, and standardize data
- Runs k-means clustering
- Visualize clusters with principal components
- Analyze clusters
- Apply findings

Data

Summary

- Source: Football Reference
- Description: Advanced goalkeeper stats for all who played at least 450 minutes in Europe's top-five leagues in the 2022-23 season (146 players in total[2])
- Notes:
  - Data only includes play in domestic leagues. Data from domestic cups and European competitions are not included.
  - There are four duplicates[3] for players who transferred during the season. I decided to include both observations of these players and treat them as separate players.

2. Including duplicates

3. Benjamin Lecomte, Jonas Omlin, Ionuț Radu, and Yann Sommer

Cleaning

- Cut out unwanted variables, such as own goals against and penalty kicks against
- Converted all counting metrics to per 90 minutes if they were not in this form already
- Standardized each variable (centering at zero and shrinking standard deviation to one) to prepare for k-means clustering
- This left me with 19 variables recorded for each of the 146 players

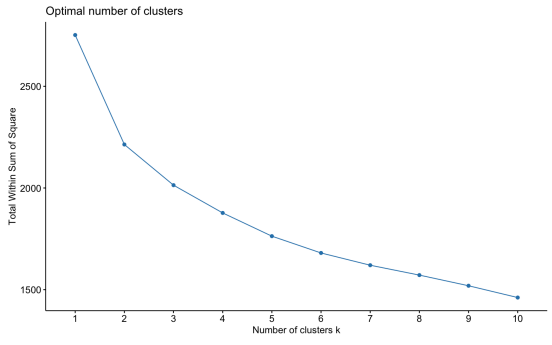
Variables Used

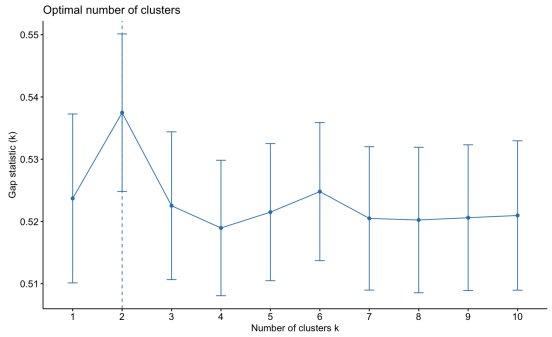
- **90sPlayed:** Minutes played divided by 90
- **Age**
- **AvgPassLen:** Average pass length (not including goal kicks)
- **CornerKickGA:** Goals against from corner kicks per 90 minutes
- **CrossesFaced:** Crosses faced per 90 minutes
- **CrossesStopped%:** Percentage of crosses faced that were stopped
- **DefActionsAvgDist:** Average distance of defensive actions from goal
- **DefActionsOPA/90:** Defensive actions outside of penalty area per 90 minutes
- **FreeKickGA:** Goals against from free kicks per 90 minutes
- **GA**: Goals against per 90 minutes
- **GoalKicksAtt:** Goal kicks taken per 90 minutes
- **GoalKicksAvgLen:** Average distance of pass on goal kicks
- **GoalKicksLaunch%:** Percentage of goal kicks that are launched[4]
- **Launch%:** Percentage of passes that are launched (not including goal kicks)
- **LaunchCmp:** Launched passes that are completed per 90 minutes
- **LaunchCmp%:** Percentage of launched passes that are completed
- **PassAtt:** Passes attempted per 90 minutes (not including goal kicks)
- **PSxG-GA/90:** Post-shot expected goals minus goals allowed per 90 minutes (positive values indicate above average shot stopping ability or good luck, vice versa)
- **ThrowAtt:** Throws attempted per 90 minutes

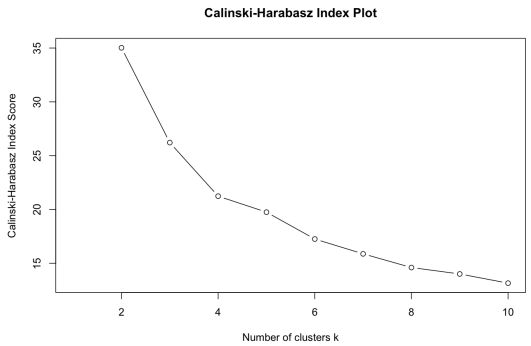
4. Passes of at least 40 yards are defined as “launched”

K-Means Clustering

Choosing Number of Clusters

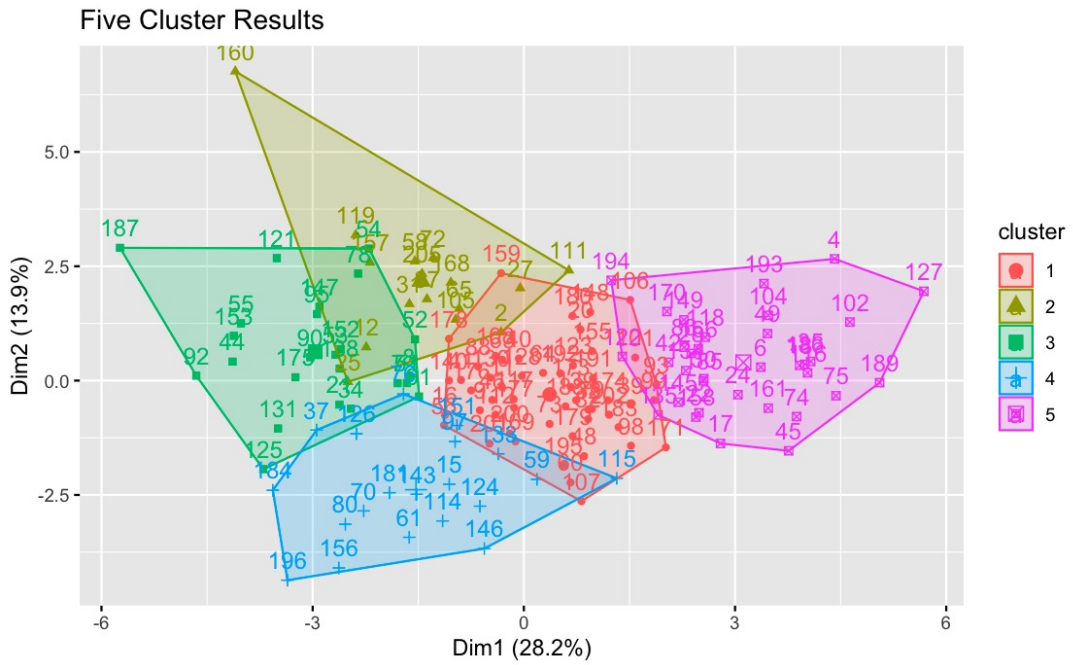






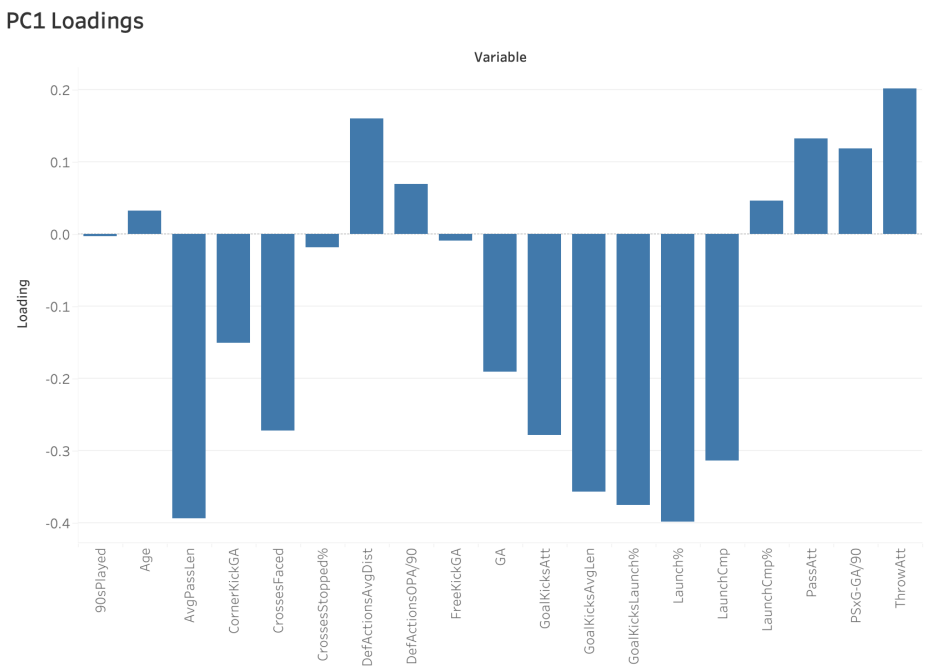
- Produced elbow, gap statistic, and Calinski-Harabasz Index plots to see if a certain number of clusters was consistently recommended as optimal
- Two clusters was most recommended, but since I wanted to create more stratification, I opted to arbitrarily find five clusters

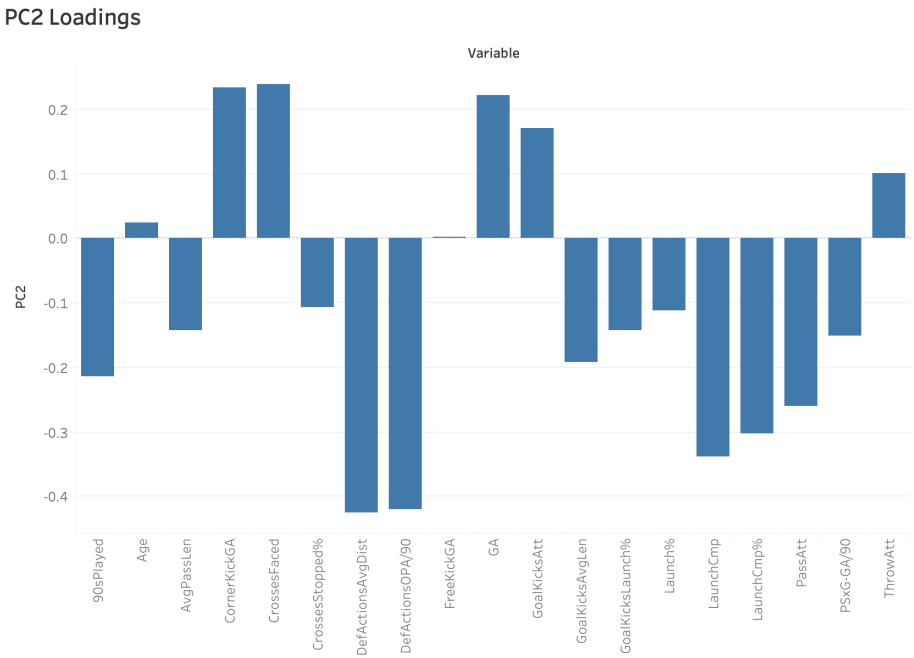
Visualizing Clusters



- Clustering was performed in 19-dimensional space but is visualized in two dimensions here using principal components
- The first two principal components cumulatively explain over 42% of the total variance in the data
- Some overlap between clusters when plotted on these two principal components, but not much

Making Sense of the Visualization: Principal Component Loadings





Launch%, AvgPassLen, and **GoalKicksLaunch%** have the greatest magnitude → GKs who go long will have a more negative value of PC1

DefActionsAvgDist, DefActionsOPA/90, and **LaunchCmp** have the greatest magnitude → GKs whose defensive actions occur further from goal have a more negative value of PC2

Results & Discussion

Cluster One: Overview

- **Size:** 34 players (23.29% of dataset)
- **Centroid Description:** Highly effective goalkeeper. Excellent shot stopper who concedes few goals. Very involved in build up, playing lots of short passes. Accurate when playing long. Faces very few crosses, not great at stopping them. Very good at defending corners, about average with free kicks.
- **Notable Players:** Alisson, Kepa Arrizabalaga, Thibaut Courtois, Gianluigi Donnarumma, Ederson, Manuel Neuer, André Onana, and Marc-André ter Stegen

Variable	Standardized Value
90sPlayed	-0.27
Age	-0.02
AvgPassLen	-1.14
CornerKickGA	-0.60
CrossesFaced	-0.87
CrossesStopped%	-0.16
DefActionsAvgDist	0.73
DefActionsOPA/90	0.43
FreeKickGA	-0.07
GA	-0.69
GoalKicksAtt	-0.85
GoalKicksAvgLen	-1.13
GoalKicksLaunch%	-1.23
Launch%	-1.17
LaunchCmp	-0.80
LaunchCmp%	0.33
PassAtt	0.68
PSxG-GA/90	0.41
ThrowAtt	0.42

Centroid Coordinates

Cluster One: Players

Name	Country	Club	League
Alisson	BRA	Liverpool	Premier League
Kepa Arrizabalaga	ESP	Chelsea	Premier League
Rubén Blanco	ESP	Marseille	Ligue 1
Janis Blaswich	GER	RB Leipzig	Bundesliga
Fabian Bredlow	GER	Stuttgart	Bundesliga
Juan Carlos	ESP	Girona	La Liga
Michele Cerofolini	ITA	Fiorentina	Serie A
Lucas Chevalier	FRA	Lille	Ligue 1
Thibaut Courtois	BEL	Real Madrid	La Liga
Yehvann Diouf	FRA	Reims	Ligue 1
Gianluigi Donnarumma	ITA	Paris S-G	Ligue 1
Ederson	BRA	Manchester City	Premier League
Péter Gulácsi	HUN	RB Leipzig	Bundesliga
Samir Handanović	SVN	Inter	Serie A
Gregor Kobel	SUI	Dortmund	Bundesliga
Andriy Lunin	UKR	Real Madrid	La Liga
Mike Maignan	FRA	Milan	Serie A
Alex Meret	ITA	Napoli	Serie A
Alexander Meyer	GER	Dortmund	Bundesliga
Florian Müller	GER	Stuttgart	Bundesliga
Manuel Neuer	GER	Bayern Munich	Bundesliga
Jonas Omlin	SUI	M'Gladbach	Bundesliga
André Onana	CMR	Inter	Serie A
Mattia Perin	ITA	Juventus	Serie A
Ivan Provedel	ITA	Lazio	Serie A
Pepe Reina	ESP	Villarreal	La Liga
Rémy Riou	FRA	Lyon	Ligue 1
Gerónimo Rulli	ARG	Villarreal	La Liga
Brice Samba	FRA	Lens	Ligue 1
Yann Sommer	SUI	M'Gladbach	Bundesliga
Yann Sommer	SUI	Bayern Munich	Bundesliga
Jason Steele	ENG	Brighton	Premier League
Marc-André ter Stegen	GER	Barcelona	La Liga
Pietro Terracciano	ITA	Fiorentina	Serie A

Cluster Two: Overview

- **Size:** 52 players (35.62% of dataset)
- **Centroid Description:** Effective goalkeeper. Above average shot stopper who concedes slightly fewer goals than average. Older than average. Not very involved in build up, unsuccessful long balls. Faces few crosses, not great at stopping them. Susceptible to conceding from free kicks.
- **Notable Players:** David De Gea, Hugo Lloris, Jan Oblak, Aaron Ramsdale, Unai Simón, Kasper Schmeichel, and Wojciech Szczęsny

Variable	Standardized Value
90sPlayed	0.56
Age	0.20
AvgPassLen	-0.16
CornerKickGA	0.00
CrossesFaced	-0.11
CrossesStopped%	-0.15
DefActionsAvgDist	-0.15
DefActionsOPA/90	-0.28
FreeKickGA	0.12
GA	-0.14
GoalKicksAtt	-0.14
GoalKicksAvgLen	-0.12
GoalKicksLaunch%	-0.10
Launch%	-0.19
LaunchCmp	-0.29
LaunchCmp%	-0.12
PassAtt	-0.30
PSxG-GA/90	0.08
ThrowAtt	0.20

Centroid Coordinates

Cluster Two: Players

Name	Country	Club	League
Édgar Badía	ESP	Elche	La Liga
Oliver Baumann	GER	Hoffenheim	Bundesliga
Marco Bizot	NED	Brest	Ligue 1
Yassine Bounou	MAR	Sevilla	La Liga
Andrea Consigli	ITA	Sassuolo	Serie A
Michele Di Gregorio	ITA	Monza	Serie A
Mory Diaw	SEN	Clermont Foot	Ligue 1
Stole Dimitrievski	MKD	Rayo Vallecano	La Liga
Bartłomiej Drągowski	POL	Spezia	Serie A
Maxime Dupé	FRA	Toulouse	Ligue 1
Fernando	ESP	Almería	La Liga
Fraser Forster	ENG	Tottenham	Premier League
Paulo Gazzaniga	ARG	Girona	La Liga
David de Gea	ESP	Manchester Utd	Premier League
Vicente Guaita	ESP	Crystal Palace	Premier League
Lukáš Hrádecký	fi FIN	Leverkusen	Bundesliga
Sam Johnstone	ENG	Crystal Palace	Premier League
Alban Lafont	FRA	Nantes	Ligue 1
Bernd Leno	GER	Fulham	Premier League
Hugo Lloris	FRA	Tottenham	Premier League
Anthony Lopes	POR	Lyon	Ligue 1
Pau López	ESP	Marseille	Ligue 1
Steve Mandanda	FRA	Rennes	Ligue 1
Vito Mannone	ITA	Lorient	Ligue 1
Agustín Marchesín	ARG	Celta Vigo	La Liga
Jordi Masip	ESP	Valladolid	La Liga
Illan Meslier	FRA	Leeds United	Premier League
Juan Musso	ARG	Atalanta	Serie A
Alexander Nübel	GER	Monaco	Ligue 1
Jan Oblak	SVN	Atlético Madrid	La Liga
Rui Patrício	POR	Roma	Serie A
Jiří Pavlenka	CZE	Werder Bremen	Bundesliga
Nick Pope	ENG	Newcastle Utd	Premier League
Aaron Ramsdale	ENG	Arsenal	Premier League
Álex Remiro	ESP	Real Sociedad	La Liga
José Sá	POR	Wolves	Premier League
Robert Sánchez	ESP	Brighton	Premier League
Kasper Schmeichel	DEN	Nice	Ligue 1
Marvin Schwäbe	GER	Köln	Bundesliga
Matz Sels	BEL	Strasbourg	Ligue 1
Luigi Sepe	ITA	Salernitana	Serie A
Rui Silva	POR	Betis	La Liga
Marco Silvestri	ITA	Udinese	Serie A
Unai Simón	ESP	Athletic Club	La Liga
Łukasz Skorupski	POL	Bologna	Serie A
Marco Sportiello	ITA	Atalanta	Serie A
Wojciech Szczęsny	POL	Juventus	Serie A
Ciprian Tătărușanu	ROU	Milan	Serie A
Kevin Trapp	GER	Eint Frankfurt	Bundesliga
Guglielmo Vicario	ITA	Empoli	Serie A
Iván Villar	ESP	Celta Vigo	La Liga
Danny Ward	WAL	Leicester City	Premier League

Cluster Three: Overview

- **Size:** 20 players (13.7% of dataset)
- **Centroid Description:** Ineffective goalkeeper. Poor shot stopper who concedes far more goals than average. Younger than average. Scarcely involved in build up, plays long more than average. Faces high volume of crosses, effective at dealing with them. Vulnerable to corner kicks, decent against free kicks.
- **Notable Players:** Dean Henderson, Edouard Mendy, and Neto

Variable	Standardized Value
90sPlayed	-1.03
Age	-0.43
AvgPassLen	0.17
CornerKickGA	1.25
CrossesFaced	1.10
CrossesStopped%	0.34
DefActionsAvgDist	-1.05
DefActionsOPA/90	-0.65
FreeKickGA	-0.30
GA	1.18
GoalKicksAtt	0.81
GoalKicksAvgLen	0.05
GoalKicksLaunch%	0.20
Launch%	0.31
LaunchCmp	-0.29
LaunchCmp%	-0.86
PassAtt	-0.56
PSxG-GA/90	-0.72
ThrowAtt	0.13

Centroid Coordinates

Cluster Three: Players

Name	Country	Club	League
Paul Bernardoni	FRA	Angers	Ligue 1
Finn Dahmen	GER	Mainz 05	Bundesliga
Yahia Fofana	FRA	Angers	Ligue 1
Gauthier Gallon	FRA	Troyes	Ligue 1
Ivo Grbić	CRO	Atlético Madrid	La Liga
Dean Henderson	ENG	Nott'ham Forest	Premier League
Daniel Iversen	DEN	Leicester City	Premier League
Mateusz Lis	POL	Troyes	Ligue 1
Alex McCarthy	ENG	Southampton	Premier League
Edouard Mendy	SEN	Chelsea	Premier League
Yvon Mvogo	SUI	Lorient	Ligue 1
Neto	BRA	Bournemouth	Premier League
Jonas Omlin	SUI	Montpellier	Ligue 1
Patrick Pentz	AUT	Reims	Ligue 1
Samuele Perisan	ITA	Empoli	Serie A
Ionuț Radu	ROU	Cremonese	Serie A
Nicola Ravaglia	ITA	Sampdoria	Serie A
Tobias Sippel	GER	M'Gladbach	Bundesliga
François-Joseph Sollacaro	FRA	Ajaccio	Ligue 1
Mark Travers	IRL	Bournemouth	Premier League

Cluster Four: Overview

- **Size:** 16 players (10.96% of dataset)
- **Description:** Average goalkeeper and shot stopper. Very young. Highly involved in build up, plays long well above average and is somewhat effective at doing so. Faces average cross volume, elite at dealing with them. Vulnerable to free kicks, average against corner kicks.
- **Notable Players:** Emiliano Martínez and David Raya

Variable	Standardized Value
90sPlayed	0.33
Age	-0.63
AvgPassLen	0.61
CornerKickGA	-0.04
CrossesFaced	0.05
CrossesStopped%	1.16
DefActionsAvgDist	0.70
DefActionsOPA/90	1.14
FreeKickGA	0.24
GA	0.17
GoalKicksAtt	0.06
GoalKicksAvgLen	1.15
GoalKicksLaunch%	0.99
Launch%	0.63
LaunchCmp	1.23
LaunchCmp%	0.19
PassAtt	0.97
PSxG-GA/90	-0.02
ThrowAtt	-0.09

Centroid Coordinates

Cluster Four: Players

Name	Country	Club	League
Julen Agirrezabala	ESP	Athletic Club	La Liga
Gavin Bazunu	IRL	Southampton	Premier League
Marco Carnesecchi	ITA	Cremonese	Serie A
Koen Casteels	BEL	Wolfsburg	Bundesliga
Oliver Christensen	DEN	Hertha BSC	Bundesliga
Mark Flekken	NED	Freiburg	Bundesliga
Rafał Gikiewicz	POL	Augsburg	Bundesliga
Lennart Grill	GER	Union Berlin	Bundesliga
Tomáš Koubek	CZE	Augsburg	Bundesliga
Giorgi Mamardashvili	GEO	Valencia	La Liga
Emiliano Martínez	ARG	Aston Villa	Premier League
Vanja Milinković-Savić	SRB	Torino	Serie A
David Raya	ESP	Brentford	Premier League
Manuel Riemann	GER	Bochum	Bundesliga
Frederik Rønnow	DEN	Union Berlin	Bundesliga
Robin Zentner	GER	Mainz 05	Bundesliga

Cluster Five: Overview

- **Size:** 24 players (16.44% of dataset)
- **Centroid Description:** Slightly below average goalkeeper and shot stopper. Older than average. Scarcely involved in build up. When on the ball, plays long and is effective at doing so. Faces high volume of crosses and struggles to handle them. About average against set pieces.
- **Notable Players:** Claudio Bravo, Keylor Navas, Guillermo Ochoa, Jordan Pickford

Variable	Standardized Value
90sPlayed	-0.12
Age	0.38
AvgPassLen	1.41
CornerKickGA	-0.15
CrossesFaced	0.51
CrossesStopped%	-0.51
DefActionsAvgDist	-0.29
DefActionsOPA/90	-0.23
FreeKickGA	-0.07
GA	0.13
GoalKicksAtt	0.78
GoalKicksAvgLen	1.04
GoalKicksLaunch%	1.13
Launch%	1.39
LaunchCmp	1.18
LaunchCmp%	0.38
PassAtt	-0.50
PSxG-GA/90	-0.16
ThrowAtt	-1.08

Centroid Coordinates

Cluster Five: Players

Player	Country	Club	League
Sergio Asenjo	ESP	Valladolid	La Liga
Emil Audero	ITA	Sampdoria	Serie A
Claudio Bravo	CHI	Betis	La Liga
Benoît Costil	FRA	Auxerre	Ligue 1
Marko Dmitrović	SRB	Sevilla	La Liga
Łukasz Fabiański	POL	West Ham	Premier League
Ralf Fährmann	GER	Schalke 04	Bundesliga
Wladimiro Falcone	ITA	Lecce	Serie A
Aitor Fernández	ESP	Osasuna	La Liga
Álvaro Fernández	ESP	Espanyol	La Liga
Sergio Herrera	ESP	Osasuna	La Liga
Benjamin Lecomte	FRA	Espanyol	La Liga
Benjamin Lecomte	FRA	Montpellier	Ligue 1
Jeremías Ledesma	ARG	Cádiz	La Liga
Benjamin Leroy	FRA	Ajaccio	Ligue 1
Lorenzo Montipò	ITA	Hellas Verona	Serie A
Keylor Navas	CRC	Nott'ham Forest	Premier League
Guillermo Ochoa	MEX	Salernitana	Serie A
Fernando Pacheco	ESP	Espanyol	La Liga
Jordan Pickford	ENG	Everton	Premier League
Ionuț Radu	ROU	Auxerre	Ligue 1
Predrag Rajković	SRB	Mallorca	La Liga
Alexander Schwolow	GER	Schalke 04	Bundesliga
David Soria	ESP	Getafe	La Liga

Cluster Comparisons

Variable	Standardized Value				
	Cluster One	Cluster Two	Cluster Three	Cluster Four	Cluster Five
90sPlayed	-0.27	0.56	-1.03	0.33	-0.12
Age	-0.02	0.20	-0.43	-0.63	0.38
AvgPassLen	-1.14	-0.16	0.17	0.61	1.41
CornerKickGA	-0.60	0.00	1.25	-0.04	-0.15
CrossesFaced	-0.87	-0.11	1.10	0.05	0.51
CrossesStopped%	-0.16	-0.15	0.34	1.16	-0.51
DefActionsAvgDist	0.73	-0.15	-1.05	0.70	-0.29
DefActionsOPA/90	0.43	-0.28	-0.65	1.14	-0.23
FreeKickGA	-0.07	0.12	-0.30	0.24	-0.07
GA	-0.69	-0.14	1.18	0.17	0.13
GoalKicksAtt	-0.85	-0.14	0.81	0.06	0.78
GoalKicksAvgLen	-1.13	-0.12	0.05	1.15	1.04
GoalKicksLaunch%	-1.23	-0.10	0.20	0.99	1.13
Launch%	-1.17	-0.19	0.31	0.63	1.39
LaunchCmp	-0.80	-0.29	-0.29	1.23	1.18
LaunchCmp%	0.33	-0.12	-0.86	0.19	0.38
PassAtt	0.68	-0.30	-0.56	0.97	-0.50
PSxG-GA/90	0.41	0.08	-0.72	-0.02	-0.16
ThrowAtt	0.42	0.20	0.13	-0.09	-1.08

White (Negative) → Green (Positive)

Discussion

- Many of the variables used are affected by the system that the goalkeeper plays in and the relative quality of their team within their league, so it would be misleading to permanently box a player into a specific cluster
    - This is evidenced by two of the four duplicates changing clusters as they switched clubs in the January transfer window[5]
- While there are some similarities in specific attributes, k-means does a good job finding five largely distinct clusters, making it a useful profiling tool for scouts
- Changing the random seed[6] used can slightly alter the clusters' centroids and compositions which is a drawback of the k-means approach

5. Jonas Omlin went from cluster three at Montpellier to one at Borussia Mönchengladbach; Ionuț Radu went from cluster three at Cremonese to five at AJ Auxerre

6. I used `set.seed(100)` in R

Example Use Case

Manchester United: Replacing David De Gea

- De Gea leaving after 12 years at United, where he won eight trophies and made over 500 appearances, after not receiving a suitable contract offer from the club
- Reports are circulating that manager Erik ten Hag was frustrated with De Gea's struggles in build out[7], which is supported by his placement in cluster two (effective goalkeeper, not involved in build out)
- Current front runner for replacement is André Onana, Inter Milan's standout goalkeeper who led his team to the 2023 Champions League Final
- Onana is in cluster one, meaning he fits the profile of many of the goalkeepers at Europe's top clubs, and is particularly adept with the ball at his feet while also being a top-notch shot stopper
- The potential replacement of De Gea with Onana signals that United are, unsurprisingly, recruiting to better fit Ten Hag's desired style of play

7. "Man Utd messed David de Gea around but ruthless Erik ten Hag's made the right call to get rid" – GOAL.com

Appendix

Software Used

- Programs:
  - R: data cleaning, k-means clustering, and principal component analysis
  - Tableau: visualizing principal component loading vectors
  - Excel: creating tables
- R packages:
  - cluster
  - dplyr
  - factoextra
  - fpc
  - ggplot2