Clustering Analysis on Football Players stats 2022-2023 using K-means Clustering with Principal Component Analysis

By: Syed Norman Daniel Bin Syed Mahadir P125754

Introduction

Football, known as the "beautiful game," crosses international boundaries and brings together people from all walks of life in an orchestra of fervor and talent. Despite having a long history, its magnetic attraction continues to captivate hearts all around the world. This sport creates a tapestry of happiness, collaboration, and pure willpower in every setting, from boisterous stadiums to dusty street corners.

In this dataset contains the stats and name of player in five different league (Premier League, Ligue 1, Bundesliga, Serie A and La Liga leagues) and background with 124 variables. This dataset has been cleaned before doing the clustering analysis. Below shows the variables meaning:

- Rk : Rank
- Player : Player's name
- Nation : Player's nation
- Pos : Position
- Squad : Squad’s name
- Comp : League that squat occupies
- Age : Player's age
- Born : Year of birth
- MP : Matches played
- Starts : Matches started
- Min : Minutes played
- 90s : Minutes played divided by 90
- Goals : Goals scored or allowed
- Shots : Shots total (Does not include penalty kicks)
- SoT : Shots on target (Does not include penalty kicks)
- SoT% : Shots on target percentage (Does not include penalty kicks)
- G/Sh : Goals per shot
- G/SoT : Goals per shot on target (Does not include penalty kicks)
- ShoDist : Average distance, in yards, from goal of all shots taken (Does not include penalty kicks)
- ShoFK : Shots from free kicks
- ShoPK : Penalty kicks made
- PKatt : Penalty kicks attempted
- PasTotCmp : Passes completed
- PasTotAtt : Passes attempted
- PasTotCmp% : Pass completion percentage
- PasTotDist : Total distance, in yards, that completed passes have traveled in any direction

- PasTotPrgDist : Total distance, in yards, that completed passes have traveled towards the opponent's goal
- PasShoCmp : Passes completed (Passes between 5 and 15 yards)
- PasShoAtt : Passes attempted (Passes between 5 and 15 yards)
- PasShoCmp% : Pass completion percentage (Passes between 5 and 15 yards)
- PasMedCmp : Passes completed (Passes between 15 and 30 yards)
- PasMedAtt : Passes attempted (Passes between 15 and 30 yards)
- PasMedCmp% : Pass completion percentage (Passes between 15 and 30 yards)
- PasLonCmp : Passes completed (Passes longer than 30 yards)
- PasLonAtt : Passes attempted (Passes longer than 30 yards)
- PasLonCmp% : Pass completion percentage (Passes longer than 30 yards)
- Assists : Assists
- PasAss : Passes that directly lead to a shot (assisted shots)
- Pas3rd : Completed passes that enter the 1/3 of the pitch closest to the goal
- PPA : Completed passes into the 18-yard box
- CrsPA : Completed crosses into the 18-yard box
- PasProg : Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area
- PasAtt : Passes attempted
- PasLive : Live-ball passes
- PasDead : Dead-ball passes
- PasFK : Passes attempted from free kicks
- TB : Completed pass sent between back defenders into open space
- Sw : Passes that travel more than 40 yards of the width of the pitch
- PasCrs : Crosses
- TI : Throw-Ins taken
- CK : Corner kicks
- CkIn : Inswinging corner kicks
- CkOut : Outswinging corner kicks
- CkStr : Straight corner kicks
- PasCmp : Passes completed
- PasOff : Offsides
- PasBlocks : Blocked by the opponent who was standing it the path
- SCA : Shot-creating actions
- ScaPassLive : Completed live-ball passes that lead to a shot attempt
- ScaPassDead : Completed dead-ball passes that lead to a shot attempt
- ScaDrib : Successful dribbles that lead to a shot attempt
- ScaSh : Shots that lead to another shot attempt
- ScaFld : Fouls drawn that lead to a shot attempt
- ScaDef : Defensive actions that lead to a shot attempt
- GCA : Goal-creating actions
- GcaPassLive : Completed live-ball passes that lead to a goal
- GcaPassDead : Completed dead-ball passes that lead to a goal
- GcaDrib : Successful dribbles that lead to a goal
- GcaSh : Shots that lead to another goal-scoring shot

- GcaFld : Fouls drawn that lead to a goal
- GcaDef : Defensive actions that lead to a goal
- Tkl : Number of players tackled
- TklWon : Tackles in which the tackler's team won possession of the ball
- TklDef3rd : Tackles in defensive 1/3
- TklMid3rd : Tackles in middle 1/3
- TklAtt3rd : Tackles in attacking 1/3
- TklDri : Number of dribblers tackled
- TklDriAtt : Number of times dribbled past plus number of tackles
- TklDri% : Percentage of dribblers tackled
- TklDriPast : Number of times dribbled past by an opposing player
- Blocks : Number of times blocking the ball by standing in its path
- BlkSh : Number of times blocking a shot by standing in its path
- BlkPass : Number of times blocking a pass by standing in its path
- Int : Interceptions
- Tkl+Int : Number of players tackled plus number of interceptions
- Clr : Clearances
- Err : Mistakes leading to an opponent's shot
- Touches : Number of times a player touched the ball. Note: Receiving a pass, then dribbling, then sending a pass counts as one touch
- TouDefPen : Touches in defensive penalty area
- TouDef3rd : Touches in defensive 1/3
- TouMid3rd : Touches in middle 1/3
- TouAtt3rd : Touches in attacking 1/3
- TouAttPen : Touches in attacking penalty area
- TouLive : Live-ball touches. Does not include corner kicks, free kicks, throw-ins, kick-offs, goal kicks or penalty kicks.
- ToAtt : Number of attempts to take on defenders while dribbling
- ToSuc : Number of defenders taken on successfully, by dribbling past them
- ToSuc% : Percentage of take-ons Completed Successfully
- ToTkl : Number of times tackled by a defender during a take-on attempt
- ToTkl% : Percentage of time tackled by a defender during a take-on attempt
- Carries : Number of times the player controlled the ball with their feet
- CarTotDist : Total distance, in yards, a player moved the ball while controlling it with their feet, in any direction
- CarPrgDist : Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal
- CarProg : Carries that move the ball towards the opponent's goal at least 5 yards, or any carry into the penalty area
- Car3rd : Carries that enter the 1/3 of the pitch closest to the goal
- CPA : Carries into the 18-yard box
- CarMis : Number of times a player failed when attempting to gain control of a ball
- CarDis : Number of times a player loses control of the ball after being tackled by an opposing player
- Rec : Number of times a player successfully received a pass

- RecProg : Completed passes that move the ball towards the opponent's goal at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area
- CrdY : Yellow cards
- CrdR : Red cards
- 2CrdY : Second yellow card
- Fls : Fouls committed
- Fld : Fouls drawn
- Off : Offsides
- Crs : Crosses
- TklW : Tackles in which the tackler's team won possession of the ball
- PKwon : Penalty kicks won
- PKcon : Penalty kicks conceded
- OG: Own goals
- Recov : Number of loose balls recovered
- AerWon : Aerials won
- AerLost : Aerials lost
- AerWon% : Percentage of aerials won

For the cluster analysis part will be separated into four parts which is Full data, Forwards, Midfielder and Defender. For the Full data part will be removed non relevant variables which has no related with our main data like “Rk”,” Nation” and etc.

```
Football=Football.drop(["Rk","Nation","Age","Born","Squad","Comp","Player"],axis=1)
```

```
: X.shape
: (2689, 116)
```

The figure 1.1 shows the variables and the data after removing unwanted variables. Next, for the selection variables for forwards, midfielder and defender position will be selected based on the relation with position football. In football, the position "forward" describes a player who typically plays in the attacking third of the pitch and oversees scoring goals and generating scoring opportunities for his side. The position selected as forward is "FW","FWMF" and "MFFW".

Player	Goals	Shots	SoT	G/Sh	G/SoT	Assists	Crs	TouAtt3rd	TouAttPen	PKwon	MP	ShoDist	ShoFK	ShoPK
Brenden Aaronson	1	1.53	0.28	0.04	0.20	0.11	2.54	21.5	2.49	0.00	20	19.0	0.11	0.0
Himad Abdelli	0	1.05	0.35	0.00	0.00	0.00	1.05	17.4	1.16	0.00	14	19.2	0.00	0.0
Matthis Abline	1	4.29	1.43	0.11	0.33	0.00	0.95	12.9	6.67	0.00	5	20.4	0.00	0.0
Matthis Abline	1	5.00	1.67	0.17	0.50	0.00	1.67	18.3	12.50	0.83	11	12.2	0.00	0.0
Zakaria Aboukhlal	5	2.75	1.02	0.11	0.29	0.24	0.96	18.7	4.97	0.00	22	18.1	0.00	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Joshua Zirkzee	1	3.08	0.77	0.06	0.25	0.19	0.00	19.4	4.04	0.00	9	14.4	0.00	0.0
Hakim Ziyech	0	1.93	0.35	0.00	0.00	0.18	9.47	32.5	1.75	0.00	12	20.3	0.00	0.0
Simon Zoller	3	1.93	0.59	0.12	0.38	0.07	1.19	11.3	3.41	0.00	18	12.8	0.00	0.0
Milan ?uri?	1	1.03	0.26	0.13	0.50	0.00	0.00	19.7	4.36	0.00	16	11.0	0.00	0.0
Filip ?uri?i?	2	1.16	0.48	0.12	0.29	0.00	1.56	17.4	1.43	0.00	20	19.4	0.00	0.0

861 rows × 14 columns

Figure 2.1 shows the selected variables for the position forwards.

The position of “midfielder” referred to as the team's "engine" since they are so important to both the attacking and defensive parts of the game. Controlling the flow of the game, connecting the defence and the attack, and supporting both their defensive colleagues and their offensive forwards are their responsibilities. The variables will focus on variables that contains stats passing and defending variables. The position selected as forward is 'MF', 'DFMF', "MFDF" and "FWDF".

Player	Assists	MP	PasTotCmp%	PasTotPrgDist	PasShoCmp%	PasShoAtt	PasMedAtt	PasLonCmp%	PasAss
Salis Abdul Samed	-0.231668	1.188761	1.042464	0.492229	0.538941	1.189115	1.169089	0.968059	-0.467735
Laurent Abergel	-0.029774	0.468817	0.712522	0.470052	0.488789	-0.139396	0.120624	0.521513	-0.356658
Oliver Abildgaard	-0.231668	-1.547025	-0.868130	-0.887784	0.965229	-1.146493	-0.852951	-2.137828	-0.930560
Tyler Adams	-0.231668	1.044772	0.305850	0.431746	0.206687	0.910557	0.382740	0.254383	0.189475
Lucien Agoume	-0.231668	-0.971070	0.390254	0.252314	0.150267	0.364153	0.557485	0.708903	-0.347401
...	...	...	...	...	...	...	...	...	...
Nadir Zortea	-0.231668	-0.395115	-1.290148	0.632347	-0.188256	-0.589375	-0.353682	-0.602826	0.180218
Martín Zubimendi	0.045937	1.044772	0.474657	0.736176	0.275645	0.235587	0.832083	0.310201	-0.458479
Szymon ? urkowski	6.077532	-1.403036	0.881330	-1.265800	0.068770	-0.342958	0.083179	1.849190	1.383561
Szymon ? urkowski	-0.231668	-1.547025	-2.149532	-1.492610	-1.122328	0.996267	-2.101124	-2.137828	-0.930560
Martin Ødegaard	0.449725	1.188761	0.083331	0.031553	0.175342	0.192732	0.020770	0.433799	1.198431

801 rows × 18 columns

PasAss	Pas3rd	PPA	CrsPA	CkIn	CkOut	CkStr	GCA	GcaPassLive	GcaPassDead
-0.467735	0.902113	-0.254486	-0.178696	-0.274039	-0.371189	-0.131818	-0.365029	-0.354513	-0.198754
-0.356658	0.124848	-0.383751	-0.351524	-0.274039	-0.371189	-0.131818	-0.241083	-0.179415	-0.198754
-0.930560	-1.297693	-0.698835	-0.351524	-0.274039	-0.371189	-0.131818	-0.365029	-0.354513	-0.198754
0.189475	0.304499	-0.189853	-0.351524	-0.274039	-0.371189	-0.131818	-0.287563	-0.245077	-0.198754
-0.347401	-0.150128	0.060598	0.184241	0.067321	-0.371189	-0.131818	-0.365029	-0.354513	-0.198754
...	...	...	...	...	...	...	...	...	...
0.180218	-0.857732	0.270654	0.339786	-0.274039	-0.371189	-0.131818	0.254700	0.520976	-0.198754
-0.458479	0.513480	-0.286802	-0.351524	-0.274039	-0.371189	-0.131818	-0.024178	0.127006	-0.198754
1.383561	-0.381108	1.320935	-0.351524	-0.274039	-0.371189	-0.131818	3.508281	5.117297	-0.198754
-0.930560	-1.297693	-0.698835	-0.351524	-0.274039	-0.371189	-0.131818	-0.365029	-0.354513	-0.198754
1.198431	-0.003474	1.329014	-0.161413	-0.042795	-0.034216	1.420922	0.626538	0.805511	-0.198754

Figure 2.2 shows the selected variables for the position midfielder.

Lastly, the position "defender" is used to describe players that spend the most of their time in the field's defensive third, where their main purpose is to keep the other side from scoring goals. Defenders are the final line of defence and are essential to preserving the shape of the squad and defending their own goal. The variables selection will focus on defensive role(Tackles,intercepts) and Positioning and Marking. The position selected as defender is 'DF', 'DFFW', "FWDF".

Player	Tkl	TklWon	TklDef3rd	TklMid3rd	TklAtt3rd	TklDri	TklDriAtt	TklDri%	TklDriPast	Blocks	BlkSh	BlkPass	In
Yunis Abdelhamid	2.50	1.59	1.45	1.00	0.05	1.32	1.68	78.4	0.36	2.23	0.77	1.45	2.0
Abner	2.67	2.33	2.33	0.33	0.00	2.00	2.33	85.7	0.33	1.33	0.67	0.67	2.0
Francesco Acerbi	1.30	0.57	1.14	0.16	0.00	0.81	0.98	83.3	0.16	0.57	0.33	0.24	0.9
Marcos Acuña	3.30	1.74	1.93	1.01	0.37	1.83	2.57	71.4	0.73	1.01	0.46	0.55	1.4
Tosin Adarabioyo	1.07	0.41	0.82	0.16	0.08	0.49	0.74	66.7	0.25	0.82	0.66	0.16	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	..
David Zima	2.00	1.11	1.11	0.89	0.00	0.67	0.89	75.0	0.22	1.56	0.44	1.11	2.0
Oleksandr Zinchenko	1.92	0.75	0.50	0.75	0.67	0.92	1.83	50.0	0.92	1.08	0.00	1.08	0.7
Nadir Zortea	2.00	2.00	2.00	0.00	0.00	2.00	2.00	100.0	0.00	1.00	0.00	1.00	0.0
Kurt Zouma	0.28	0.14	0.14	0.14	0.00	0.21	0.42	50.0	0.21	0.49	0.49	0.00	0.9
Igor Zubeldia	1.43	0.86	1.00	0.43	0.00	0.86	1.07	80.0	0.21	1.00	0.79	0.21	0.7

893 rows × 22 columns

BlkSh	BlkPass	Int	Tkl+Int	Clr	Recov	AerWon	AerLost	AerWon%	Fls	CarMis	CarDis
0.77	1.45	2.00	4.50	2.91	6.64	2.18	1.23	64.0	1.32	0.73	0.68
0.67	0.67	2.00	4.67	2.67	6.00	1.00	2.00	33.3	0.67	1.00	0.67
0.33	0.24	0.98	2.28	3.33	4.96	2.76	1.63	63.0	0.73	0.57	0.24
0.46	0.55	1.47	4.77	1.28	7.52	1.19	0.92	56.5	1.65	1.56	1.38
0.66	0.16	1.07	2.13	5.33	4.67	2.13	0.41	83.9	0.66	0.16	0.25
...	...	...	...	...	...	...	...	...	...	...	...
0.44	1.11	2.00	4.00	2.44	5.33	3.33	2.00	62.5	0.89	0.22	0.44
0.00	1.08	0.75	2.67	1.58	7.00	1.75	0.58	75.0	0.17	1.00	1.00
0.00	1.00	0.00	2.00	1.00	7.00	1.00	2.00	33.3	2.00	0.00	0.00
0.49	0.00	0.92	1.20	5.42	3.94	2.61	0.35	88.1	0.49	0.28	0.00
0.79	0.21	0.79	2.21	4.07	4.93	2.00	1.57	56.0	1.64	0.21	0.07

Figure 2.3 shows the selected variables for the position defender.

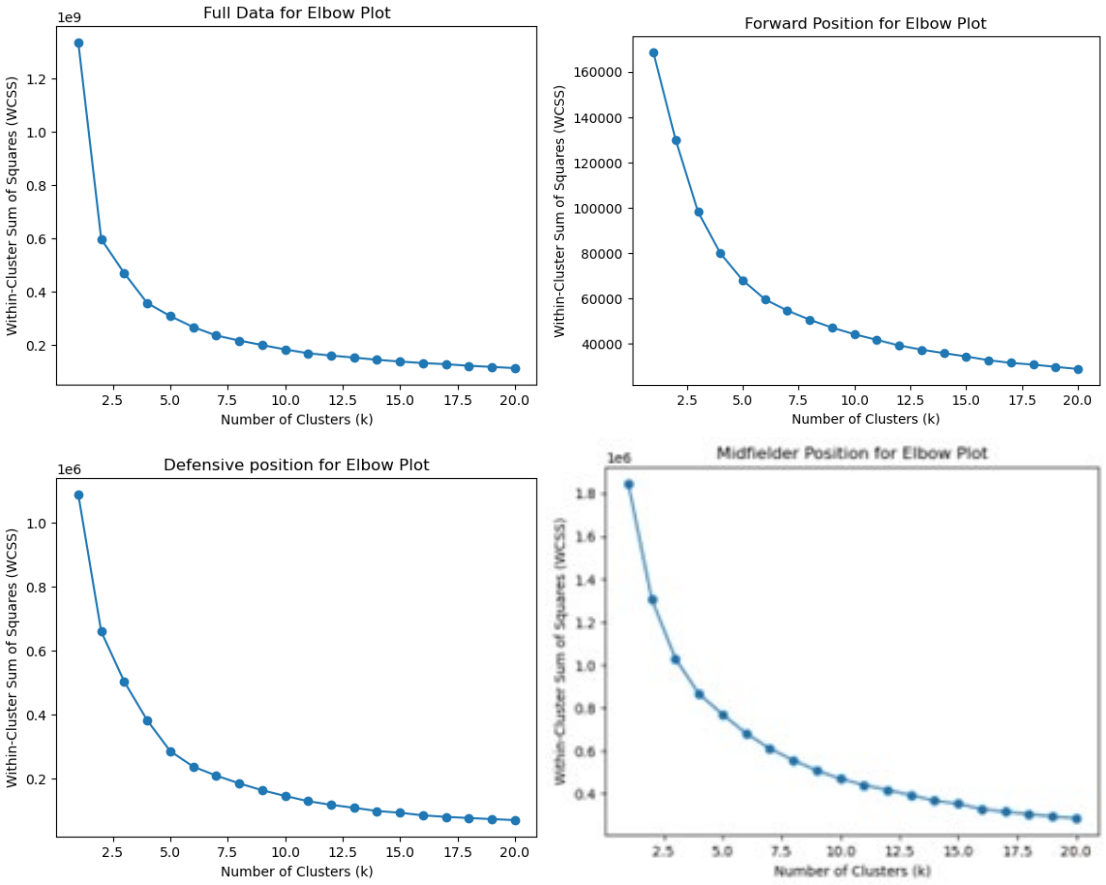
The objective of this topic is to categorizing players or teams into distinct groups based on their performance attributes.

Data Preprocessing using Min Max Scaling

Min-Max scaling is a technique for transforming numerical data so that all characteristics have a common scale within a certain range and because of the Min-Max scaling, each feature will be converted to a range between 0 and 1. This technique will be used before using PCA algorithm. The data will be scale before using this technique.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \textbf{(1.1)}$$

Selecting Best k cluster (Elbow Plot)



The picture 3.1 shows the Elbow Plot for All Position, Forward, Defensive and Midfielder position which determines the best k for each plot.

To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the Within Cluster Sum of Squares known as WCSS starts decreasing in a linear fashion. Thus, for the given data, we conclude that the optimal number of clusters for the data is:

Full Data: 4

Forward Position: 4

Defensive Position: 5

Midfielder Position :4

The determine of k is depends on the huge reduction in WCSS and after that we can see it doesn’t go down quickly.

Clustering Analysis (PCA with K-means Clustering)

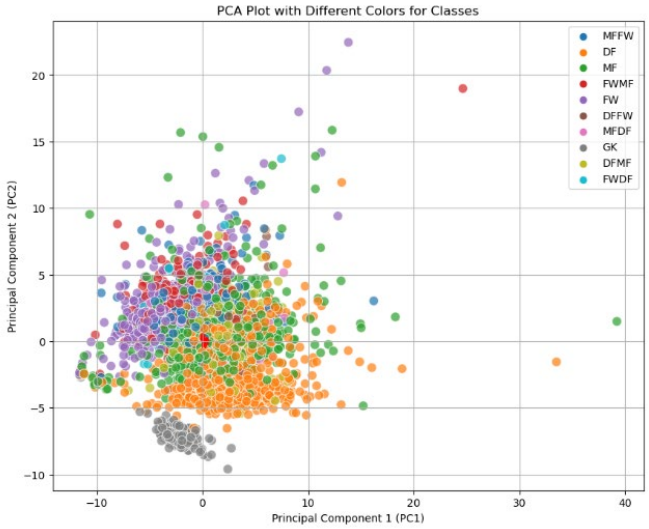


Figure 4.1 shows PCA Plot with Different Position

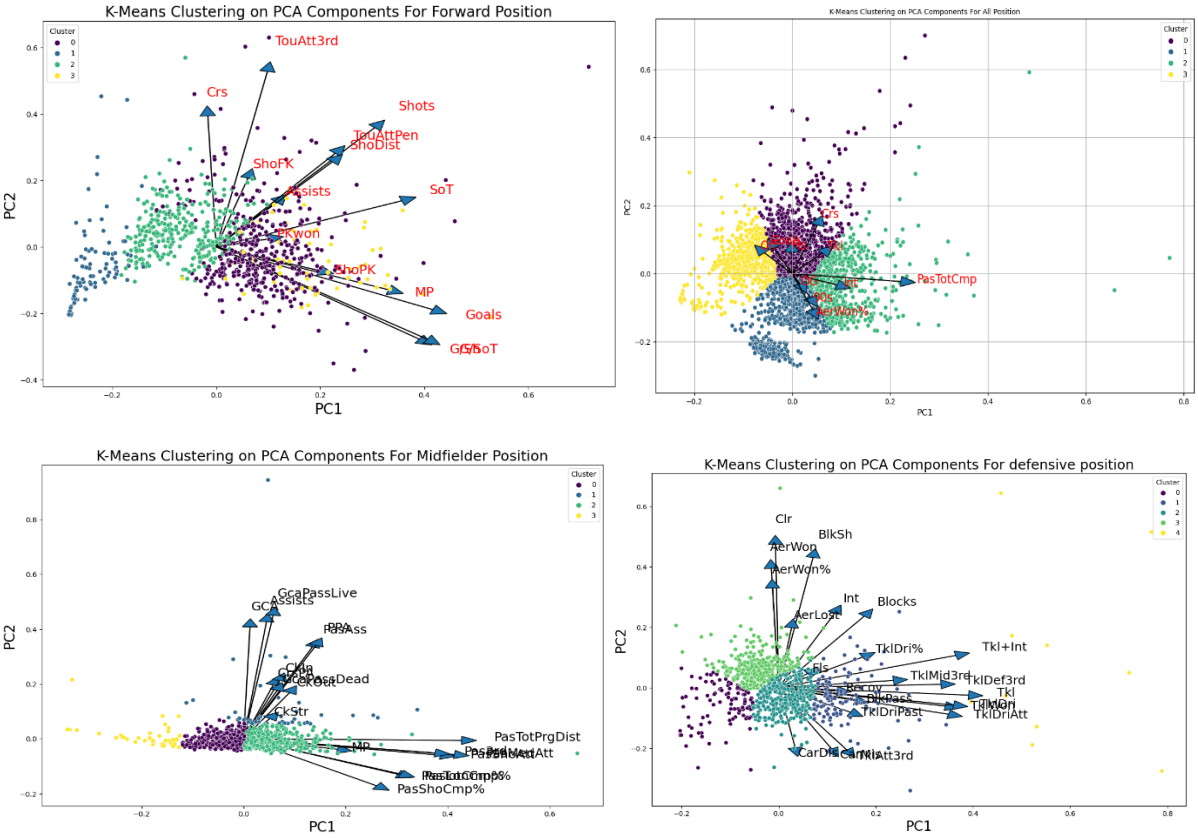


Figure 4.2 shows K-means clustering using PCA Plot with Different Position

From graph 4.1 we can see clearly that position “GK” has better cluster area than another while another position like “MF” and “DF” are to see mixed up little the scatter point with another. It is due to some midfielders and defender having maybe same plays tail like, some midfielders are often to play defensive or some defender like often to play aggressive. For example, from

the observation light blue dots, the defender which plays as “FWDF” often creates chances like forward position although the position is from defender. Those playstyles can give effect where the scatter point will be plot.

For figure 4.2, for all position k-means clustering, we cannot include all the variables with calibrated axes into the graphs because it can show messy plot. However, we select certain variables for the axes. For purple cluster we can conclude that, those areas are often creating goals and cross which the playstyle are mostly aggressive. The green cluster to see that playstyle are often to has been dominant, fit, and good passer than another cluster. Next. The blue cluster indicates that players can play 90 minutes, good in contest a ball in then air but also create their own goals which give the teams lose in their match. Lastly, the yellow cluster shows the players are often to be offside but also creates a chance for goals.

The forward position shows graph with four cluster, the yellow and purple cluster indicates that the players playstyle is often to create goals, very good on shots and can get a penalty kick while the green cluster are often creating chance to teammate for scoring like GCA.

The midfielder position graph shows four cluster, the blue cluster with upwards axis indicate that the playstyle player is often very creative on dead ball situation and can create assists for team while the axis going right with green cluster are often to be very good in passing more importantly completed passes have traveled towards the opponent's goal.

The defender position graph shows five clusters. The blue cluster often has playstyle solid defending technique with good tackle and able to recover loose balls. The turquoise color is balance with all axes that display in the graph, the green cluster playstyle is good in won air ball and intercepts ball while the yellow cluster are often playstyle outside from his position which often helps other position in the match.

Conclusion

In summary, K-means clustering using PCA is an effective method for examining football player performance and playstyles. The clusters make it easier to spot players who have similar traits and can help coaches and analysts better grasp the advantages and disadvantages of certain positions. The method analysis offers insightful information for player development, team building, tactical planning, and talent scouting.