Project Report

Group 6 | DANA 4840

Topic:

Exploring the FIFA players data to assist managers for decision making.

Team Members:

Bhargavkumar Kakadiya	: 100350875
Gaurav	: 100350679
Guneesh Bhatia	: 100346420
Ronak Batra	: 100348475

Index

Contents

1. Introduction

FIFA 19 is a football simulation video game developed by EA Vancouver as part of Electronic Arts' FIFA series. The use of analytics — using data and statistics to better understand something — is growing across most sports. This is especially true in soccer, where the most successful teams are also frequently the most dedicated to analytics. This project will focus on uncovering the insights on how the decisions are made by the management of the reputed clubs (or the creators of the game, in this case EA) based on players' mental and physical attributes and other factors given in the dataset. It can be downloaded from Kaggle dataset here and used as required with preferred tools.

The data contains basic information on every player in the game such as; Name, Club, Nationality, and Age. More importantly, for this purpose, it contains all the physical and skill attributes; Sprint Speed, Free Kick Accuracy, Heading Accuracy etc. Also, ratings for each playing position are included - e.g. CM, RB, ST, GK etc.

The report involves various ways about applying K-means clustering algorithm, and since we want to compare and focus on different physical attributes and skills of the players, we have divided the players into two categories as outfield players and the goalkeepers. Further, we delve deep into the analysis of each category of the players to determine and answer the research questions. We also have applied hierarchical clustering on the same and compared the two to understand which clustering algorithm gives us the best results. Moreover, stepwise-regression and classification algorithms were used to understand the important variables to be considered to predict the wages/value of a player and classify the players based on the position they should be playing at respectively.

Research Questions: -

Further, this report answers the study on the following three key questions which are as follows:

1) How the management of the club scout player replacements based on their playing style, physical attributes such as Age, Stamina, strength, speed etc. and their market value. Recommendations will be made for the following players segment -
    a. All Players
    b. Forward Players
    c. Goalkeepers
2) Predict the wage of a player based on many mental and physical attributes.

Dataset: -

The dataset consists of 18207 rows and 89 columns. The dataset consists is mainly divided into three different type of variables: -

1) Skills based variables: - Crossing, Finishing, Heading Accuracy, Short Passing, Dribbling etc.
2) Position Based variables: -. CM, RB, ST, GK etc.
3) Miscellaneous: - Name, Club, Joining Date, Release Clause etc.

The data dictionary (description of all the variables) is attached in the appendix.

2. Descriptive Analysis

2.1.Players Distribution by Age

Most players are between the ages 20-25. The decreasing number of players after the age 25 points to the fact that only high skilled players continue to play as they get older.

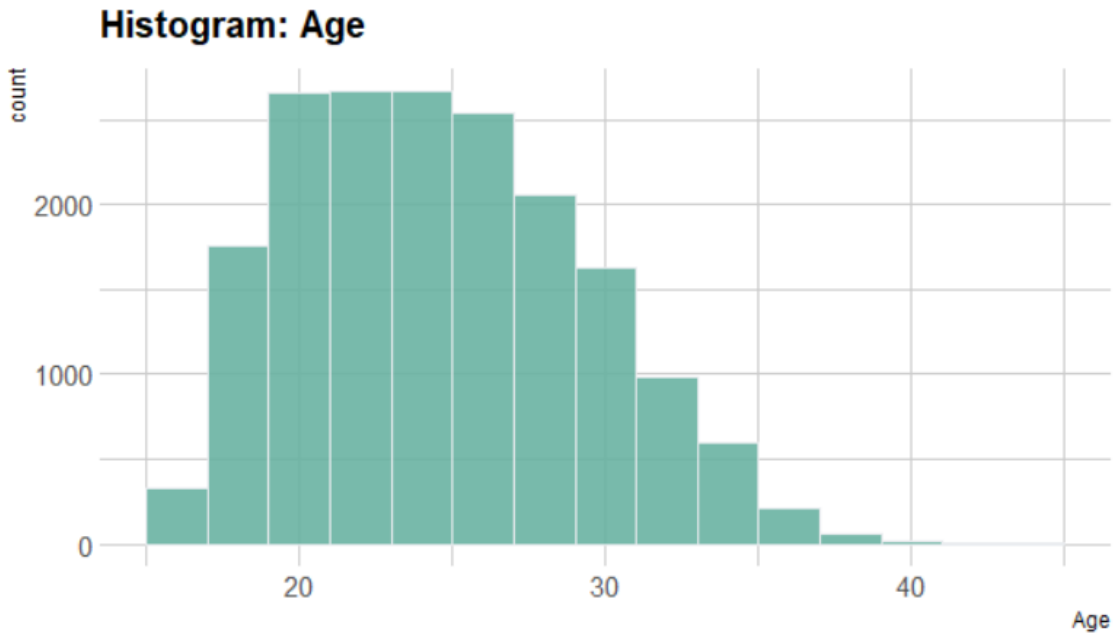
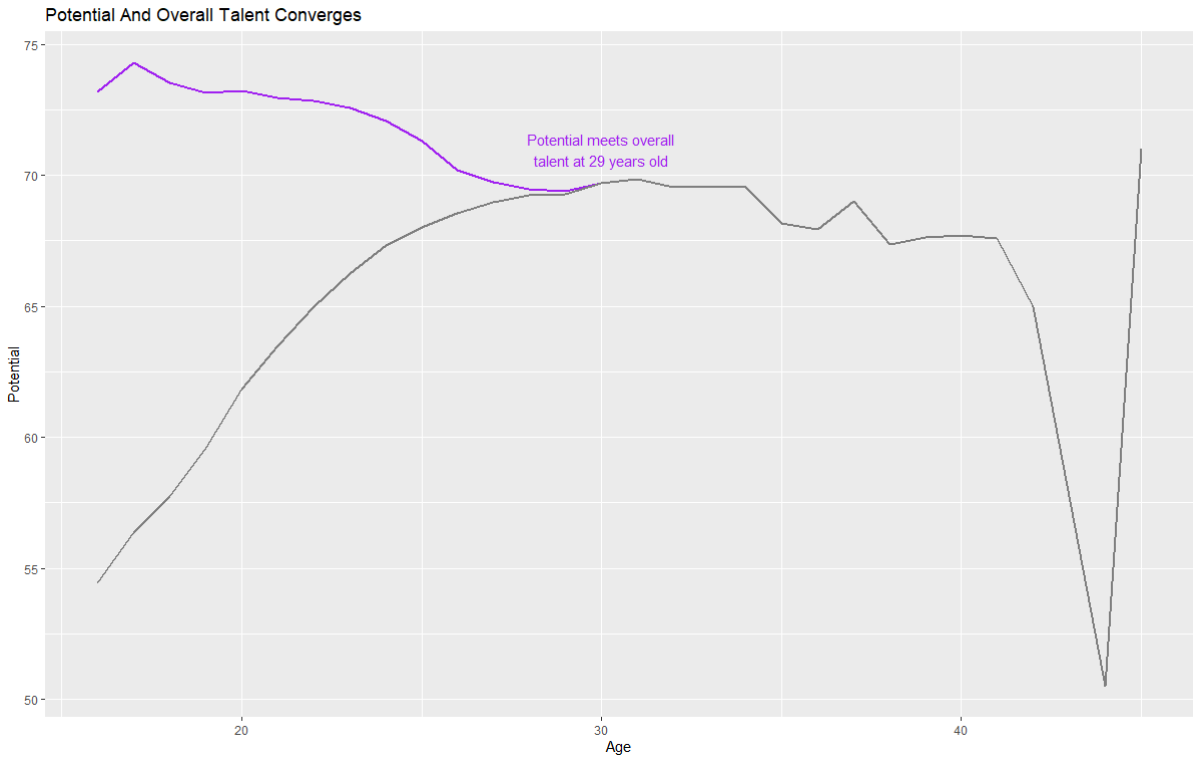


Table 2.1 Distribution of players by Age

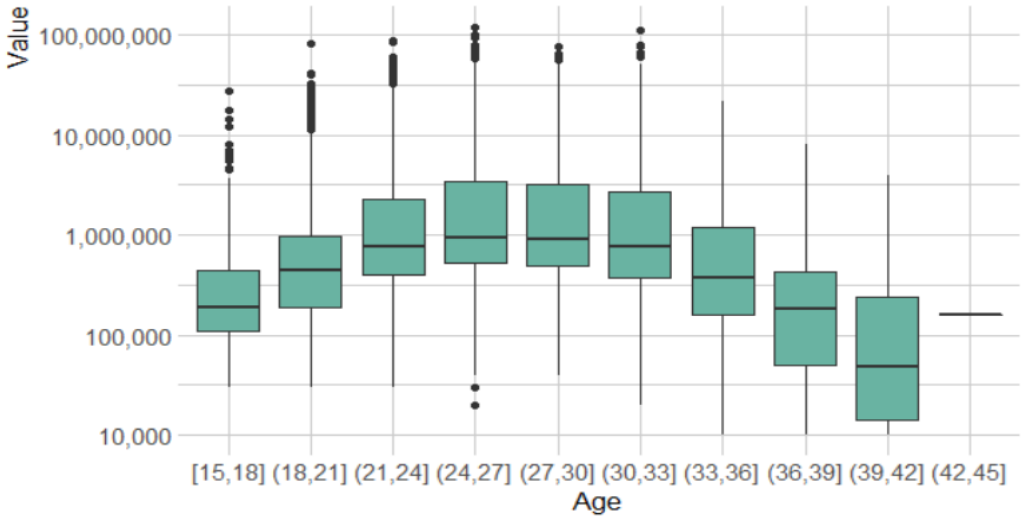
2.2.Mean Overall rating by Age

The mean rating of players tends to improve as age increases, reaches the maximum around age 29, stays constant for a few years, and then reduces. It is also inferred from the graph that the potential of players also converges when their overall ratings reach maximum.



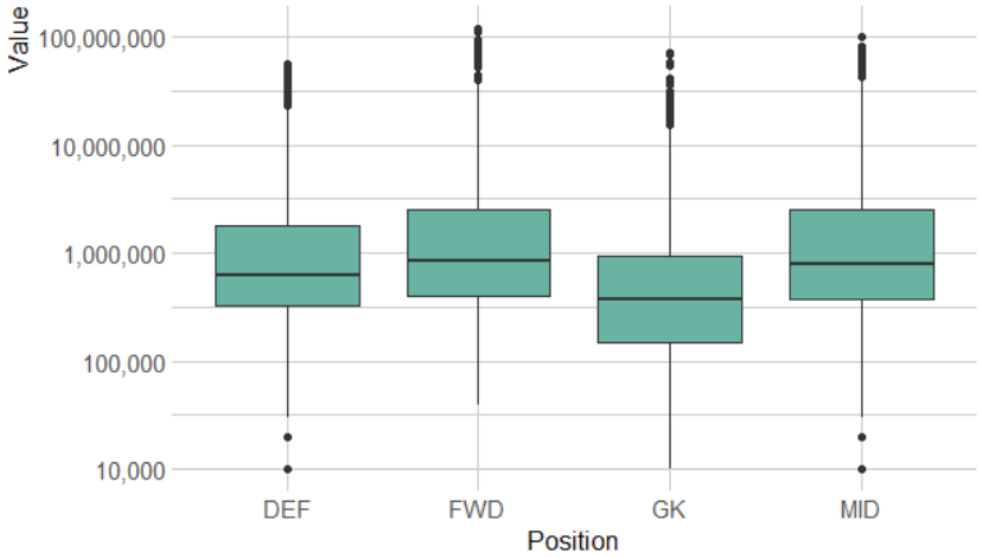
2.3.Age vs Valuations

The player's market value is the highest in their prime playing years (24-30) and reduces as their age increases.



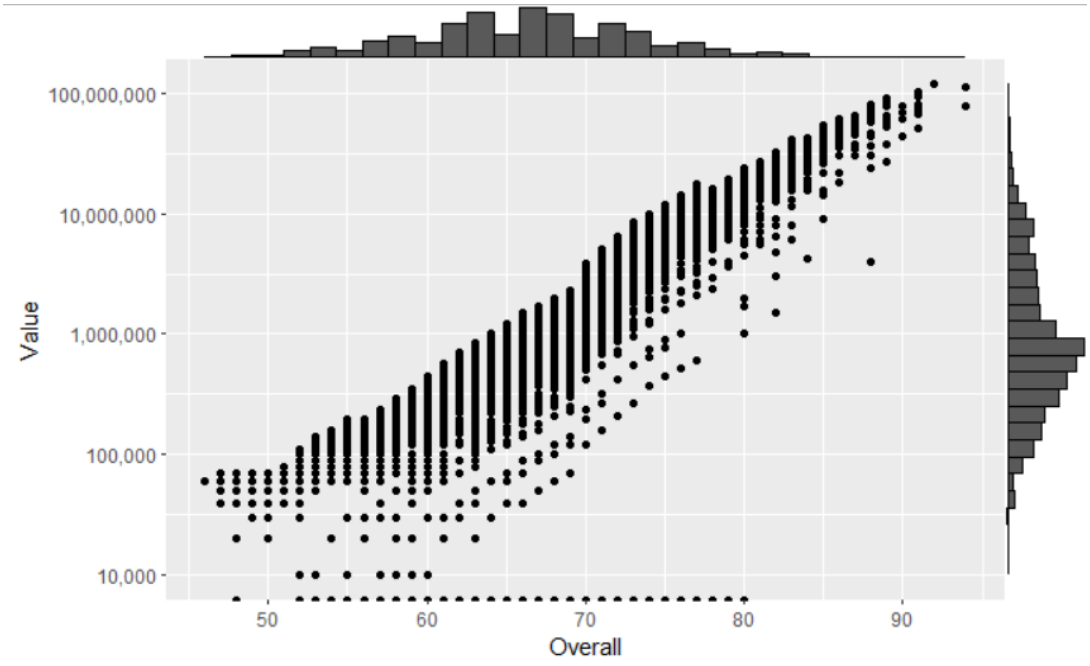
2.4.Position vs Value

The players' market value varies with the position they play at. Goalkeepers are valued at the lowest and midfielders at the highest. Forward players are also valued at approximately as high as midfielders.



2.5.Overall rating vs Value

The value of a player and their overall rating have a high correlation. Some of the players at the same rating have different market value due to the positions they play and the country and club they belong to.



3. Data cleaning and Preprocessing

3.1 Data Cleaning

The dataset which was obtained from Kaggle had very few issues. The data types were observed to be the same throughout each column. However, for our analysis, the data was preprocessed to as followed:

1) Rows with most **missing values** were found and removed from our dataset, these players were novice and were not that significant for analysis.
2) Dataset has 26 columns explaining ratings of each player at that particular position. They are

"LS"	"ST"	"RS"	"LW"	"LF"	"CF"	"RF"	"RW"	"LAM"	"CAM"	"RAM"	"LM"	"LCM"
"CM"	"RCM"	"RM"	"LWB"	"LDM"	"CDM"	"RDM"	"RWB"	"LB"	"LCB"	"CB"	"RCB"	"RB"

These columns had values which were the combination of their actual rating + the potential they had in that playing position as shown below in the left figure. These were transformed to the actual rating as shown in the figure on the right.

LS	ST	RS
88+2	88+2	88+2
91+3	91+3	91+3

LS	ST	RS
88	88	88
91	91	91

s

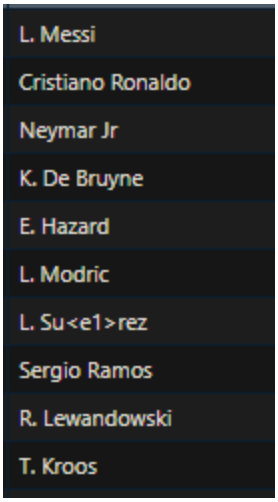
3) Weight column was a string type with lbs as a suffix, the suffix were removed and transformed to numeric.
4) Height of the players was in feet and inches. They were transformed to inches only.

3.2 Outliers

The outliers were calculated using Mahala Nobis distance. At 1% significance level, the following metrics were calculated: -

1. **Outfield Players –**
    a. Cut-off Distance – 72.44
    b. Outliers – 1101/15021
2. **Goalkeepers –**
    a. Cut-off Distance – 68.70
    b. Outliers – 165/1860

To get a better understanding of the outliers, the outfield players that were outliers were summarized.



These players currently are the most exceptional football players. Due to their exceptionally high ratings and value, they seem to be outliers, but their high values are completely justified with the real-life achievements and records they have obtained. Therefore, for further analysis these data points are not removed from the dataset.

3.3 Feature Engineering

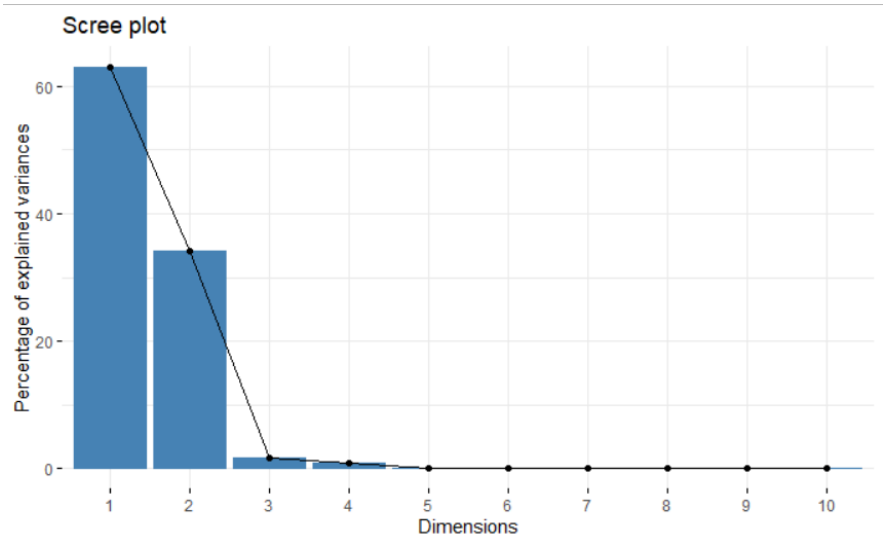
1) The position variable contains the best suited playing position of a player. It has 26 different positions. The new variable called 'Position Group' was formed based on the position variable after mapping as shown in the table

Position Group	Playing Positions
Forward (FWD)	LS, RS, ST, CF, RF, LF, LW, RW.
Midfielders (MID)	LAM, RAM, CAM, LDM, RDM, CDM, LM, CM, RM.
Defenders (DEF)	RB, CB, LB, LCB, RCB, LWB, RWB.
Goalkeepers (GK)	GK.

3.4 PCA

3.4.1 With position ratings

There are 26 positions where a player can play in football. Each outfield player (goalkeepers were not rated and it makes sense) in our dataset is rated how well he would play at those positions. The figure 3.4.1 in the appendix shows us the correlation between the position ratings. It can be observed that most of the forward and mid fielding positions are highly correlated and all defending positions are highly correlated. Next, PCA was run on these 26 variables to reduce the dimensionality.



2 components were able to explain 97% Variance of the 26 position rating variables. After studying the factor loadings, all the forward and mid fielding positions fall into PC1 and all the defending positions fall into PC2.These two components were included in our dataset with names 'Pos_F_M' and 'Pos_D' respectively.

3.4.2 With physical strengths and playing skills

PCA ran on these 39 variables and were not that conclusive so all the variables were used as it is for clustering.

3.5 Final datasets after pre-processing.

The first dataset had these 61 columns with a total of 16122 outfielders:

```
"Name"                    "Age"                     "Nationality"
"Overall"                 "Potential"               "Club"
"Value"                   "Wage"                    "Special"
"Preferred Foot"          "International Reputation" "Weak Foot"
"Skill Moves"             "Work Rate"               "Body Type"
"Real Face"               "Position"                "Jersey Number"
"Joined"                  "Loaned From"             "Contract Valid Until"
"Height"                  "Weight"                  "pos_F_M"
"pos_D"                   "Crossing"                "Finishing"
"HeadingAccuracy"         "ShortPassing"            "Volleys"
"Dribbling"               "Curve"                   "FKAccuracy"
"LongPassing"             "BallControl"             "Acceleration"
"SprintSpeed"             "Agility"                 "Reactions"
"Balance"                 "ShotPower"               "Jumping"
"Stamina"                 "Strength"                "LongShots"
"Aggression"              "Interceptions"           "Positioning"
"Vision"                  "Penalties"               "Composure"
"Marking"                 "StandingTackle"          "SlidingTackle"
"GKDiving"                "GKHandling"              "GKKicking"
"GKPositioning"           "GKReflexes"              "Release Clause"
"PositionGroup"
```

The second dataset contains a total of 59 variables with 2025 goalkeeprs.

```
"Name"                    "Age"                     "Nationality"
"Overall"                 "Potential"               "Club"
"Value"                   "Wage"                    "Special"
"Preferred Foot"          "International Reputation" "Weak Foot"
"Skill Moves"             "Work Rate"               "Body Type"
"Real Face"               "Position"                "Jersey Number"
"Joined"                  "Loaned From"             "Contract Valid Until"
"Height"                  "Weight"                  "Crossing"
"Finishing"               "HeadingAccuracy"         "ShortPassing"
"Volleys"                 "Dribbling"               "Curve"
"FKAccuracy"              "LongPassing"             "BallControl"
"Acceleration"            "SprintSpeed"             "Agility"
"Reactions"               "Balance"                 "ShotPower"
"Jumping"                 "Stamina"                 "Strength"
"LongShots"               "Aggression"              "Interceptions"
"Positioning"             "Vision"                  "Penalties"
"Composure"               "Marking"                 "StandingTackle"
"SlidingTackle"           "GKDiving"                "GKHandling"
"GKKicking"               "GKPositioning"           "GKReflexes"
"Release Clause"          "PositionGroup"
```

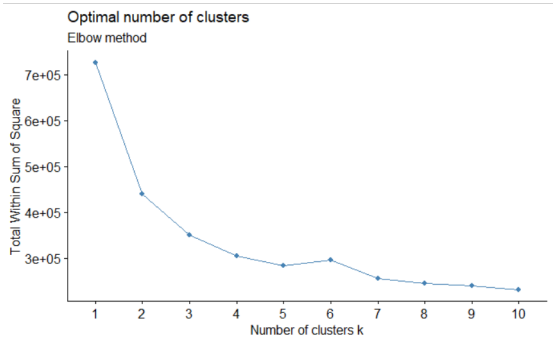
4. Cluster analysis

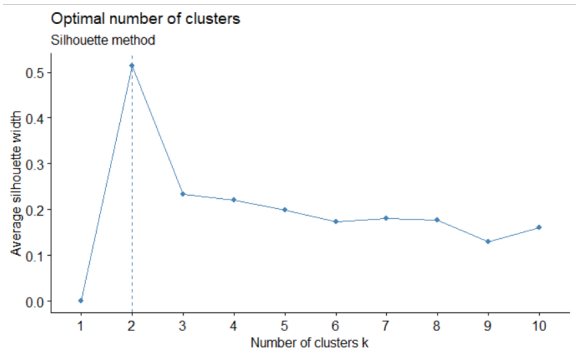
4.1 Clustering on all players based on physical attributes and playing skills.

Our first aim is to cluster every player (FWD, MID, DEF, GK) based on their physical attributes and playing skills. Below variables were selected for the clustering.

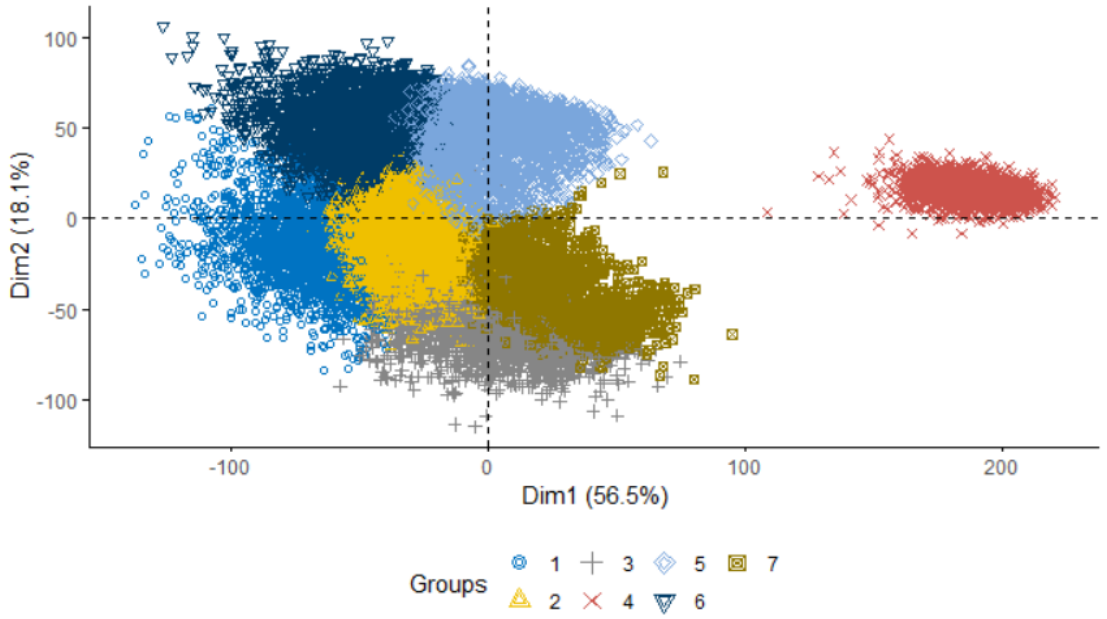
```
"Age"              "Overall"        "Potential"      "Weak Foot"
"Skill Moves"      "Height"         "Crossing"       "Finishing"
"HeadingAccuracy"  "ShortPassing"   "Volleys"        "Dribbling"
"Curve"            "FKAccuracy"     "LongPassing"    "BallControl"
"Acceleration"     "SprintSpeed"    "Agility"        "Reactions"
"Balance"          "ShotPower"      "Jumping"        "Stamina"
"Strength"         "LongShots"      "Aggression"     "Interceptions"
"Positioning"      "Vision"         "Penalties"      "Composure"
"Marking"          "StandingTackle" "SlidingTackle"  "GKDiving"
"GKHandling"       "GKKicking"      "GKPositioning"  "GKReflexes"
```

a. Assessing number of clusters
   - Elbow method = 7
   - Silhouette Method = 2





b. Quality of clusters
   - Kmeans
     i. WCSS – 64.9%
     ii. Silhouette Width – 0.24
     iii. Total Sum of Squares – 255108
     iv. Total within-cluster Sum of Squares – 470731



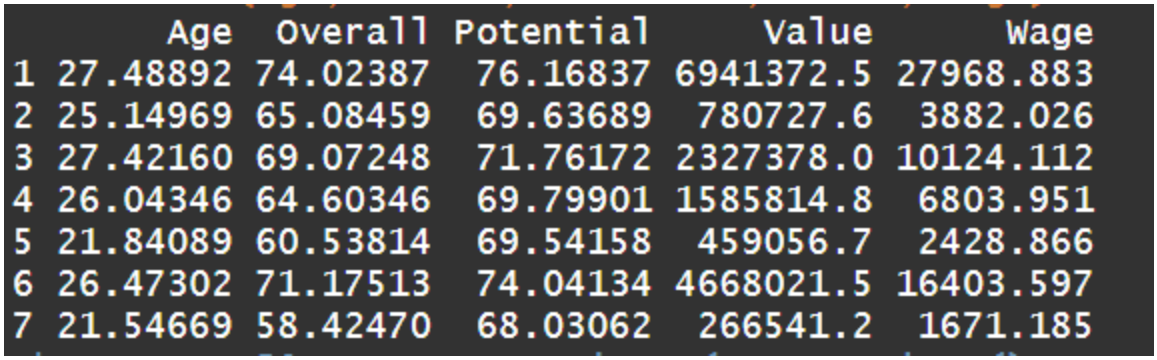
c. Cluster Label and Inferences

The seven clusters formed were analysed, and the confusion matrix was created for mapping cluster label to the number of players playing in that each position.

	DEF	FWD	GK	MID
1	769	63	0	1514
2	1618	87	0	1889
3	1867	27	0	217
4	0	0	2025	0
5	16	1635	0	1259
6	6	1598	0	1565
7	1590	8	0	394

From the above confusion matrix, we can infer that **cluster 4** consists entirely of goalkeepers only. **Cluster 3** and **cluster 7** consists majorly of defenders. **Cluster 2** consists of defenders and defense-minded midfielders, whereas **cluster 5** and **cluster 6** have forward players and attacking midfielders. **Cluster 1** on the other hand consists of players that play mid-midfield, that is, they play on positions such as CM, RM or LM.

We have dived deeper into the cluster results to get a better understanding of the cluster that were formed, and the inferences are tabulated below.



	Age	Overall	Potential	Value	Wage
1	27.48892	74.02387	76.16837	6941372.5	27968.883
2	25.14969	65.08459	69.63689	780727.6	3882.026
3	27.42160	69.07248	71.76172	2327378.0	10124.112
4	26.04346	64.60346	69.79901	1585814.8	6803.951
5	21.84089	60.53814	69.54158	459056.7	2428.866
6	26.47302	71.17513	74.04134	4668021.5	16403.597
7	21.54669	58.42470	68.03062	266541.2	1671.185

Cluster Number	Cluster Label	Inference	Recommendation
1	Midfielders	These are the most valued players, with the highest overall rating and potential. They play in the midfield, and as the domain knowledge confirms, they are the most important players in the team. Aside from having high stamina, speed, and ball control, they are equally good at playing both defensive and offensive.	Managers should consider these players if they are looking to strengthen the core of the team. They can be added to the team if coordination between the forward and defenders needs to be increased.
2	Defensive Midfielders	Players from this group play as defenders or defensive midfielders and are on the low spectrum of the skills. They have an average experience.	These players should be considered by managers that do not want to spend a lot and are looking for defensive midfielders for their team.
3	High Skilled Defenders	This cluster consists of players that play as defenders and have high experience and are highly rated. They are valued high as well, but they are valued lower than the high skilled midfielders and forward players.	If the manager is looking to strengthen their defense and are ready to pay a large sum of money to acquire the player, they should look through this cluster.

4	Goalkeepers	This cluster consists entirely of goalkeepers. It is separate from all other clusters and can be clearly interpreted that no other players can play as goalkeepers, or these players can't play at any other position.	While looking for goalkeepers for the team, managers should browse through this cluster.
5	Inexperienced Forward Players	Amateur players that play in the forward position majorly form this cluster. They have very high potential, but since they are inexperienced, their market value and overall rating are low.	Managers should look through this group if they want the players with highest potential for the lowest price. These players will need more guidance and training, than others, but at the same time they have much more scope to improve.
6	Experienced forward and attacking midfielders.	Midfielders that usually play in attacking formations such as "attacking midfield" and forward players such as strikers and wingers form this cluster. They are highly skilled and have a long-playing experience. As a result, they are valued at a high price and have high overall ratings.	Since these players are highly skilled and bring a lot of experience with them, managers of clubs that have no budget restraint and with a weak attacking line should consider them.
7	Low Skilled Defenders.	Amateur players that play in the defensive positions majorly form this cluster. They have very high potential, but since they are inexperienced, their market value and overall rating are low.	Managers should look through this group if they want the players with highest potential for the lowest price. These players will need more guidance and training, than others, but at the same time they have much more scope to improve.

Limitations

1. The clusters regarding the forward players had a lot of noise and overlapped a lot with midfielders. Arguably, it can be said that the players that play in forward and attacking midfielder positions have the same skill set, but these clusters fail to provide recommendation to the club managers.

2. The cluster that consists entirely of goalkeepers gives very limited insight as to what different types of goalkeepers are there and as a result, we cannot give proper recommendation to the club manager.

   To overcome these limitations and get a deeper insight about the goalkeepers and forward players, clustering was done again based on position.

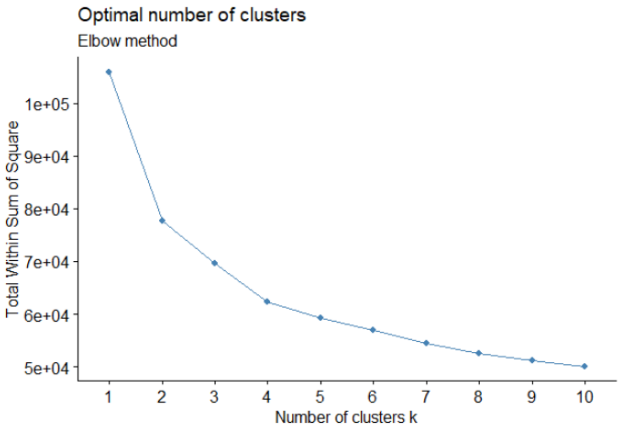
4.2 Clustering on forward players and goalkeepers based on position ratings, physical attributes and playing skills.

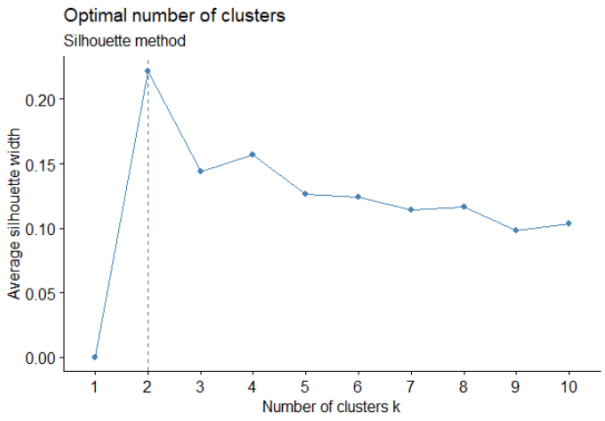
4.2.1 Forward Players

The clustering was performed on below variables for all FWD players

```
 [1] "Height"          "pos_F_M"         "pos_D"           "Crossing"        "Finishing"
 [6] "HeadingAccuracy" "ShortPassing"    "Volleys"         "Dribbling"       "Curve"
[11] "FKAccuracy"      "LongPassing"     "BallControl"     "Acceleration"    "SprintSpeed"
[16] "Agility"         "Reactions"       "Balance"         "ShotPower"       "Jumping"
[21] "Stamina"         "Strength"        "LongShots"       "Aggression"      "Interceptions"
[26] "Positioning"     "Vision"          "Penalties"       "Composure"       "Marking"
[31] "StandingTackle"
```

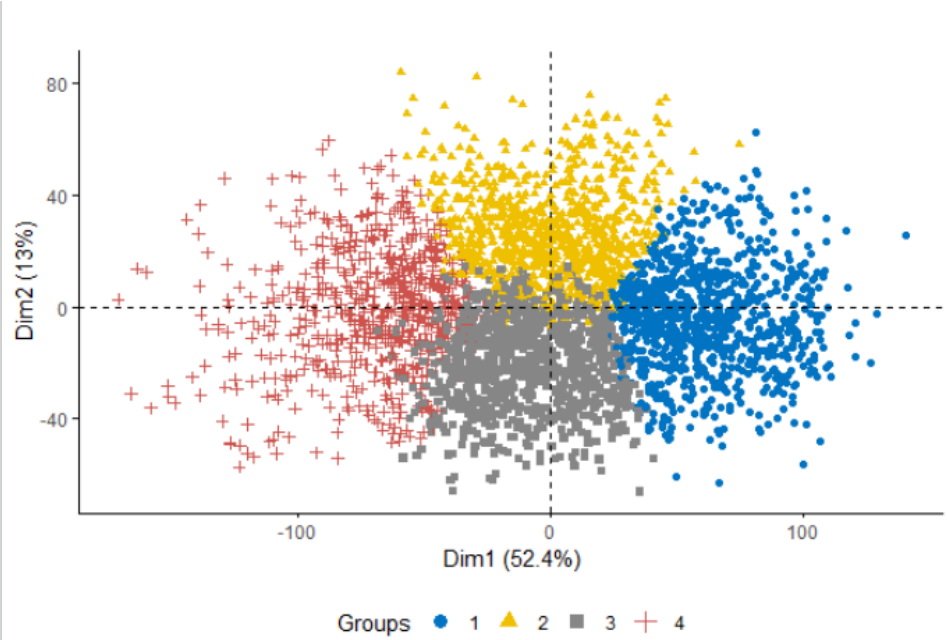
a. Assessing number of clusters
   - Elbow method = 4
   - Silhouette Method = 2

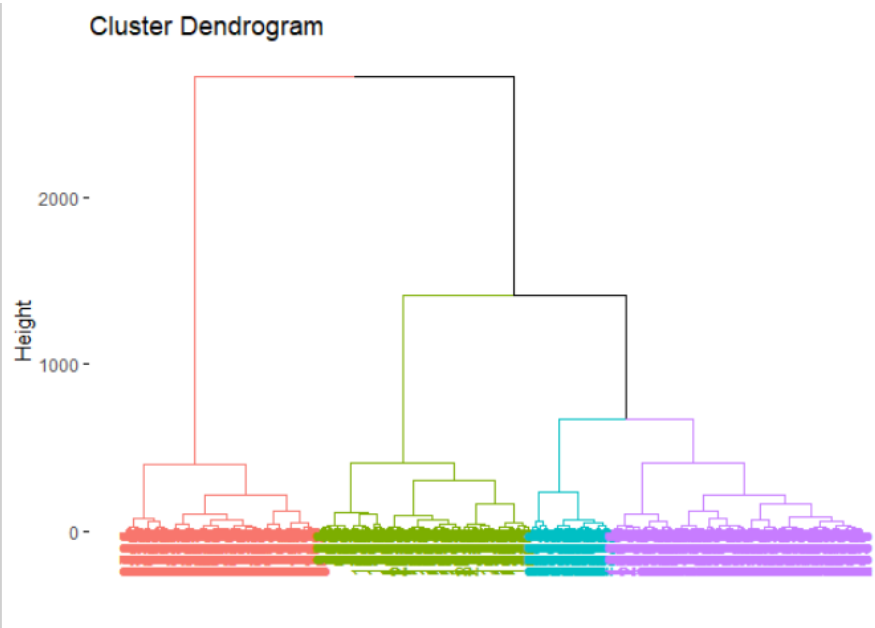




b. Quality of clusters
   - Kmeans
     i. WCSS – 41.2%
     ii. Silhouette Width – 0.16
     iii. Total Sum of Squares – 105927
     iv. Within Sum of Squares – 62256.36

- Hierarchical
    i. Agglomerative Coefficient – 0.98





c. Accessing different methods.
Due to the high Agglomerative coefficient and low silhouette score for k-means, hierarchical clustering was found to be a better clustering method. Moreover, the same was verified by clValid method for 4 clusters.

```
Optimal Scores:

              Score    Method       Clusters
Connectivity 107.3810 hierarchical 4
Dunn           0.1306 hierarchical 4
Silhouette     0.1792 kmeans       4
```

d. Cluster Label and Inferences

The four clusters formed were analysed, and the aggregate values for different attributes were tabulated.

	Group.1	Age	Overall	Potential	Value	Wage	Height	pos_F_M	pos_D	Crossing	Finishing
1	1	20.88652	58.24270	68.84045	301129.2	1871.910	71.13596	46.296812	45.56757	37.75393	59.62247
2	2	24.23890	65.78964	71.27061	1130333.0	5349.894	69.64376	-1.488135	34.73342	57.65433	62.71670
3	3	27.38060	75.46866	77.44179	10554343.3	37329.851	70.71194	-43.036850	31.84210	65.23134	74.10149
4	4	26.76206	68.32018	71.23026	1963963.8	9117.325	73.28399	4.185484	37.71212	45.67763	69.16667

	HeadingAccuracy	ShortPassing	Volleys	Dribbling	Curve	FKAccuracy	LongPassing	BallControl	Acceleration
1	54.70787	49.37865	46.77865	56.28989	42.04494	34.82921	36.85056	56.23034	67.80787
2	51.95666	61.53277	55.33404	67.73256	55.68605	48.04123	52.28436	66.34567	78.42918
3	65.79701	70.99254	70.03284	75.44478	69.10299	62.45373	60.87761	75.62537	76.28955
4	69.62171	59.84539	61.80811	63.22478	51.47917	44.49123	44.91228	65.87829	63.32566

	SprintSpeed	Agility	Reactions	Balance	ShotPower	Jumping	Stamina	Strength	LongShots	Aggression
1	68.10225	63.64045	52.85281	65.21461	57.31461	62.95393	56.68989	60.34607	51.52360	39.88652
2	77.93446	76.23784	60.82241	73.31078	64.40169	63.79810	65.53805	59.21670	58.21142	48.73890
3	76.06716	76.11791	73.36418	71.95821	75.80896	70.27164	71.68657	68.99254	70.70448	61.77910
4	65.59759	61.68202	64.43202	57.26535	70.27083	69.24890	65.51206	77.33662	61.46820	58.50768

	Interceptions	Positioning	Vision	Penalties	Composure	Marking	StandingTackle	SlidingTackle
1	19.25618	55.86629	47.87640	57.66404	50.11011	24.65955	19.93258	19.17079
2	31.15645	63.02537	58.64799	57.45032	59.71142	33.61734	31.04863	29.24524
3	36.72239	75.59254	69.08507	69.83433	72.88060	37.79254	35.17463	30.76716
4	26.89803	69.01096	55.34868	64.91228	63.57237	29.86952	25.97807	22.69298

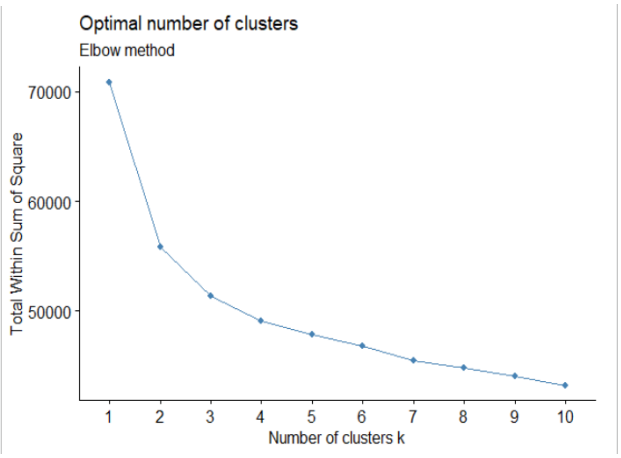
Cluster Number	Cluster Label	Inference	Recommendation
1	Low skilled, highly experienced players	These players have high experience but lower overall score and value than the other cluster with the same experience. Their skills that directly affect their performance in forward playing positions such as shot power, finishing, dribbling range between second and third highest of the 4 clusters.	Players from these clusters are great where managers are looking for highly experienced players to nurture younger players on the team.

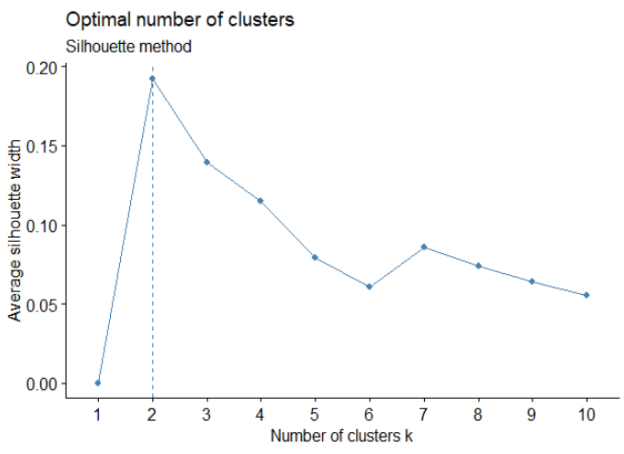
2	Highly skilled, low experienced players	These players have a mean age of 24 and have second highest skills that directly affect their forward playing skills. They can be considered as the players that have developed their skills and the next step for them is to hone their skills further.	Players from this cluster should be considered when managers don't want players that need to be nurtured from the ground up but are still looking for young players that have the potential to be among the top players.
3	Highly skilled, highly experienced players.	These players are the top of the food chain, they have the highest value, highest overall rating and the highest skills score that directly affect playing in forward position.	These players should be considered when the managers want the best of the best players and could afford to pay a huge sum of money.
4	No experience, high potential players	This cluster consists of players that are just starting up in the field. Their skills that directly affect playing in forward position are low, but their core skills such as strength, speed etc are comparable to the players of other three clusters.	These players should be considered if the managers are ready to gamble on the players. They show the greatest potential of all other players and with proper guidance and nurturing could become great players.

4.2.2 Goalkeepers

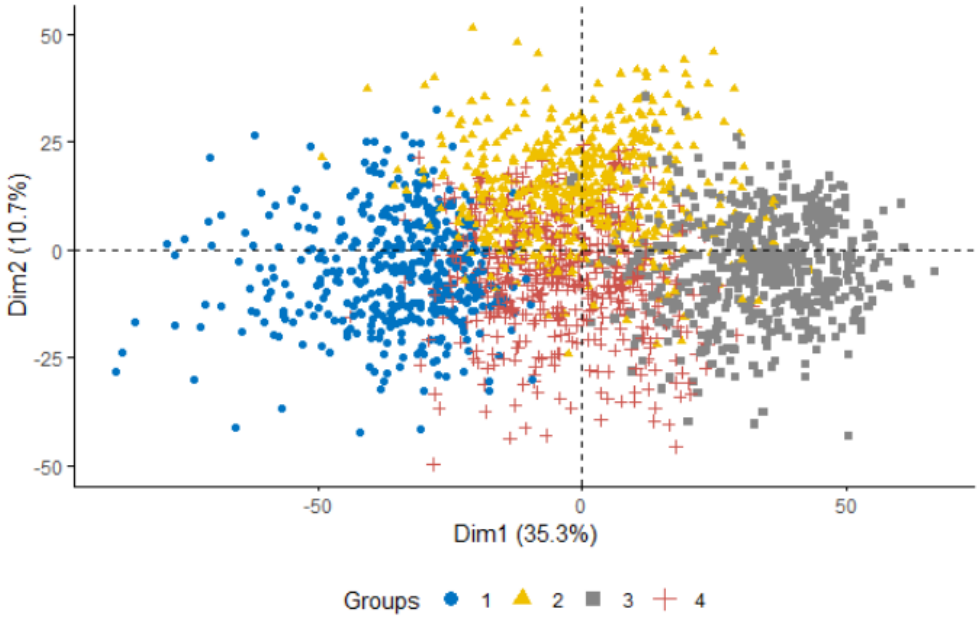
```
"Height"          "Crossing"        "Finishing"       "HeadingAccuracy" "ShortPassing"
"Volleys"         "Dribbling"       "Curve"           "FKAccuracy"      "LongPassing"
"BallControl"     "Acceleration"    "SprintSpeed"     "Agility"         "Reactions"
"Balance"         "ShotPower"       "Jumping"         "Stamina"         "Strength"
"LongShots"       "Aggression"      "Interceptions"   "Positioning"     "Vision"
"Penalties"       "Composure"       "Marking"         "StandingTackle"  "SlidingTackle"
"GKDiving"        "GKHandling"      "GKKicking"       "GKPositioning"   "GKReflexes"
```

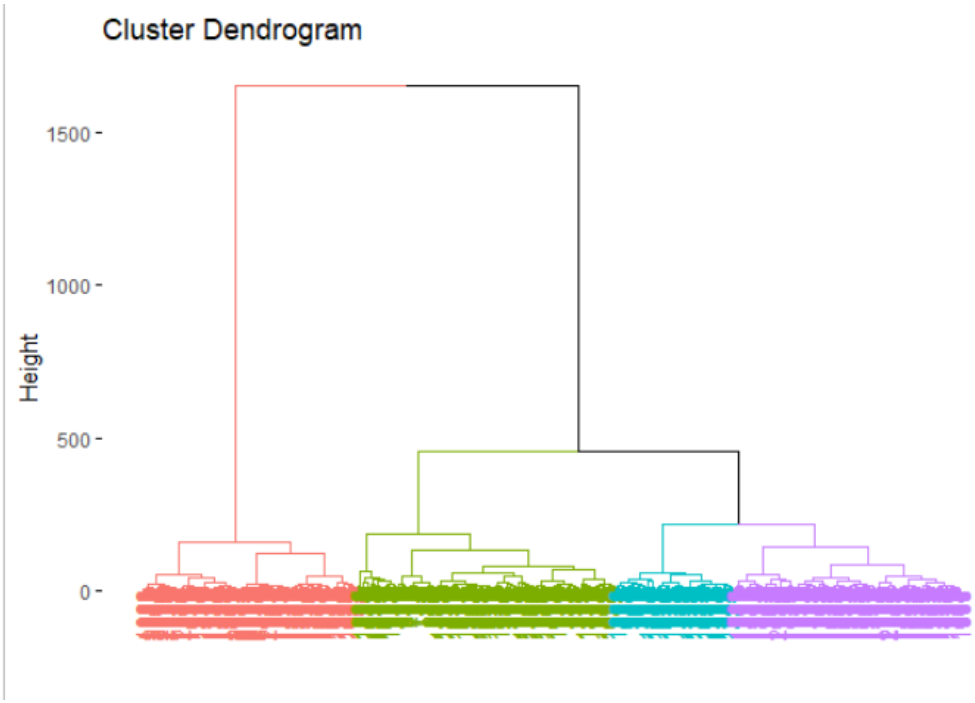
a. Assessing number of clusters
    a. Elbow method = 4
    b. Silhouette Method = 2





b. Quality of clusters
    a. Kmeans
        i. WCSS – 30.8%
        ii. Silhouette Width – 0.11
        iii. Total Sum of Squares – 70840
        iv. Within Sum of Squares – 49040
    b. Hierarchical
        i. Agglomerative Coefficient – 0.98





c. Cluster Label and Inferences

The four clusters formed were analysed.

	cluster	Age	Overall	Potential	Value	GKDiving	GKHandling	GKKicking	GKPositioning	GKReflexes
1	1	29.23502	74.01382	76.00230	5490737.3	74.57834	71.48157	69.41475	72.76959	75.75806
2	2	25.44693	61.87896	67.20484	357076.4	62.58659	59.90317	59.09870	60.10428	63.39106
3	3	21.62734	57.21161	67.29026	205149.8	58.51311	56.68727	55.63670	55.08801	58.63858
4	4	28.53077	67.15385	69.87692	1013451.9	67.42115	65.08846	63.30769	66.14615	68.50577

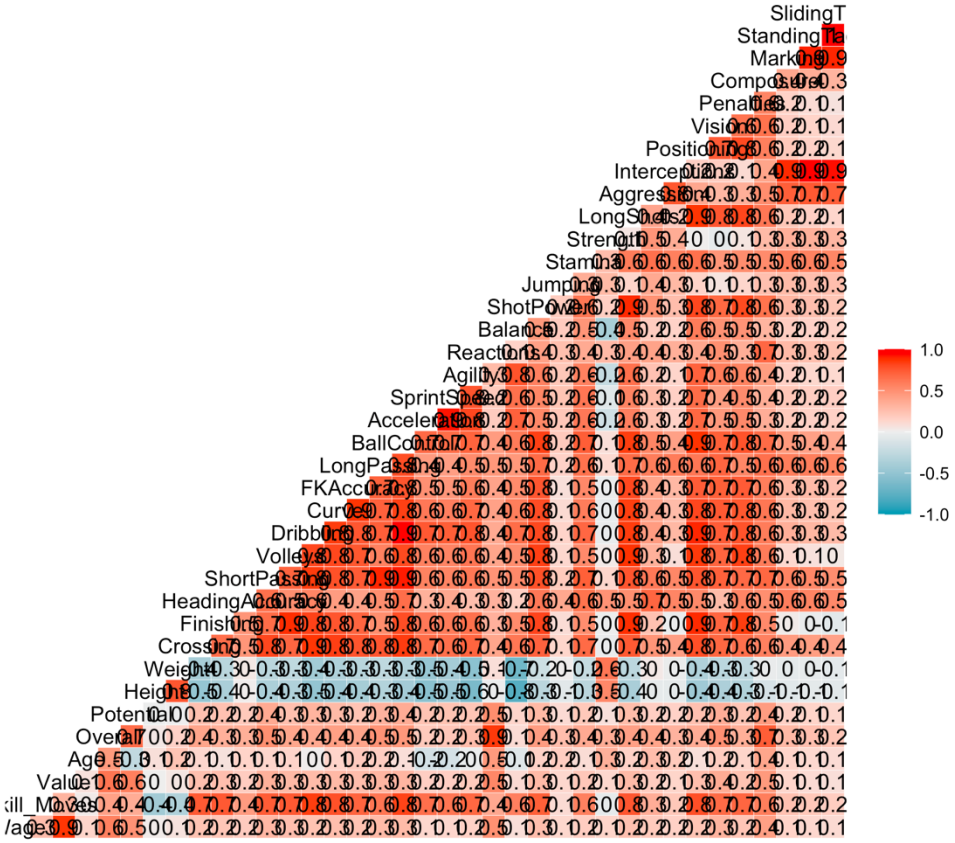
Cluster Number	Cluster Label	Inference	Recommendation
1	Highly experienced, high skills players	These players have high experience but high overall score and value than the other cluster. Also, their goalkeeping skills are higher than other clusters.	Managers should opt for this cluster if they are willing to spend huge amounts of money and highly talented players.
2	Mediocre experienced, mediocre skills	The players in this cluster have mediocre potential, low value,	This cluster should be opted by the manager if they are

		and mediocre goalkeeping skills.	looking for young players that need not to be nurtured a lot.
3	Low skilled, less experienced players.	Players in this cluster have a mean age of 21 and have the lowest goalkeeping skills and value when compared to other clusters.	Players in this cluster should be considered by managers when they are looking for young and cheap talented players whom they can train for the future.
4	Highly experienced and mediocre skills	The players in this cluster have high potential, value, and mediocre goalkeeping skills.	Managers should consider this cluster when they do not want players to be nurtured from ground but have potential to be a top player.

4.3 Stepwise Regression

The stepwise regression is performed on the dataset to understand how the wages of the football players are affected by other physical and playing skill capabilities.

- We start off with checking the correlations between the variables and getting insights of whether any variables in the data have multicollinearity.



- From the correlation matrix, it is very evident that there are many highly correlated variables indicating early signs of multi- collinearity. Although, we can proceed further by removing variables having high correlation based on our subject knowledge, we prefer the model to deal with multi- collinearity and gives us the best possible solution.
- **Splitting the data into training and testing: -** Split the data into training (70%) and testing (30%). When fitting the data, we get various models, the best (chosen by low AIC and high R square value) is shown below: -

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-36221.912	5497.194		-6.589	0.000	-46997.242	-25446.582
Value	0.003	0.000	0.840	138.133	0.000	0.003	0.003
Overall	-131.162	43.548	-0.041	-3.012	0.003	-216.523	-45.802
Potential	197.021	39.644	0.055	4.970	0.000	119.314	274.729
Age	504.187	45.707	0.108	11.031	0.000	414.595	593.779
SlidingTackle	64.084	7.328	0.062	8.746	0.000	49.721	78.447
Stamina	-80.869	11.526	-0.059	-7.016	0.000	-103.462	-58.277
Penalties	45.576	12.375	0.033	3.683	0.000	21.320	69.832
FKAccuracy	-62.387	11.047	-0.050	-5.647	0.000	-84.041	-40.733
Crossing	63.376	12.225	0.053	5.184	0.000	39.414	87.339
ShotPower	35.436	12.311	0.028	2.879	0.004	11.306	59.567
LongPassing	-45.872	13.373	-0.032	-3.430	0.001	-72.084	-19.659
`Skill Moves`	-563.149	250.532	-0.019	-2.248	0.025	-1054.230	-72.069
BallControl	39.206	18.421	0.030	2.128	0.033	3.097	75.315
Height	231.721	72.358	0.028	3.202	0.001	89.889	373.554
Balance	43.241	14.462	0.028	2.990	0.003	14.893	71.590
Strength	-45.842	13.560	-0.026	-3.381	0.001	-72.423	-19.262
Acceleration	-31.505	12.246	-0.022	-2.573	0.010	-55.508	-7.501
Jumping	21.175	9.831	0.011	2.154	0.031	1.906	40.445
Weight	20.072	11.105	0.014	1.807	0.071	-1.696	41.841

Step	Variable	Removed	R-Square	R-Square	C(p)	AIC	RMSE
1	Value	addition	0.746	0.746	328.6860	273361.3951	11284.1652
2	Overall	addition	0.750	0.750	157.1860	273192.9754	11209.2321
3	Potential	removal	0.750	0.750	155.7870	273191.5692	11209.0528
4	Reactions	addition	0.750	0.750	119.0680	273155.2259	11192.6036
5	Composure	addition	0.751	0.751	96.0800	273132.4185	11182.1295
6	Age	removal	0.751	0.751	94.6470	273130.9816	11181.9371
7	Composure	addition	0.751	0.751	68.9190	273105.3983	11170.2535
8	SlidingTackle	addition	0.752	0.751	61.7250	273098.2389	11166.6700
9	Stamina	addition	0.752	0.752	48.8440	273085.4010	11160.5951
10	Reactions	addition	0.752	0.752	40.5200	273077.0956	11156.5120
11	Penalties	addition	0.752	0.752	33.4520	273070.0379	11152.9777
12	FKAccuracy	addition	0.752	0.752	27.7570	273064.3460	11150.0433
13	Crossing	addition	0.753	0.752	25.2570	273061.8457	11148.5090
14	Height	addition	0.753	0.752	22.4920	273059.0793	11146.8583
15	Balance	addition	0.753	0.752	19.1280	273055.7106	11144.9439
16	`Skill Moves`	addition	0.753	0.753	16.6260	273053.2036	11143.4075
17	ShotPower	addition	0.753	0.753	15.4480	273052.0206	11142.4515

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	4.802527e+12	17	282501604878.662	2275.409	0.0000
Residual	1.576262e+12	12696	124154225.456		
Total	6.378789e+12	12713			

As we can observe, the R square value for the model is satisfactory. We move forward and test it on our testing data.

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-26761.350	5802.680		-4.612	0.000	-38136.929	-15385.772
Value	0.003	0.000	0.872	95.783	0.000	0.003	0.004
Overall	-192.308	63.975	-0.060	-3.006	0.003	-317.724	-66.892
Potential	168.198	59.535	0.047	2.825	0.005	51.486	284.910
Age	516.525	68.128	0.107	7.582	0.000	382.967	650.083
SlidingTackle	111.885	32.368	0.106	3.457	0.001	48.431	175.339
Stamina	-72.168	16.051	-0.051	-4.496	0.000	-103.635	-40.700
Penalties	60.299	16.222	0.042	3.717	0.000	28.498	92.100
`Skill Moves`	-1889.110	382.336	-0.064	-4.941	0.000	-2638.642	-1139.577
Dribbling	61.001	22.661	0.052	2.692	0.007	16.575	105.426
LongPassing	-36.654	18.356	-0.025	-1.997	0.046	-72.640	-0.669
Height	206.852	69.766	0.025	2.965	0.003	70.081	343.622
StandingTackle	-62.568	32.394	-0.060	-1.931	0.053	-126.074	0.938
Jumping	30.179	14.199	0.016	2.125	0.034	2.342	58.015
Crossing	34.542	17.883	0.028	1.932	0.053	-0.516	69.599

Final Model Output

Model Summary

R	0.868	RMSE	11142.452
R-Squared	0.753	Coef. Var	113.203
Adj. R-Squared	0.753	MSE	124154225.456
Pred R-Squared	0.750	MAE	5028.305

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Value	addition	0.753	0.753	169.2920	116947.3744	11130.1874
2	Overall	addition	0.757	0.757	84.6260	116864.5139	11044.8232
3	Potential	removal	0.757	0.757	66.5350	117211.1424	11027.5035
4	Reactions	addition	0.758	0.757	70.4980	116850.5540	11030.6764
5	Age	addition	0.758	0.758	62.5120	116842.6651	11021.6793
6	Reactions	addition	0.759	0.758	48.8000	116829.0692	11006.9209
7	SlidingTackle	addition	0.760	0.759	29.3680	116809.7240	10986.3818
8	Stamina	addition	0.760	0.760	24.0670	116804.4333	10980.0401
9	Penalties	addition	0.760	0.760	19.8120	116800.1796	10974.7472
10	`Skill Moves`	addition	0.760	0.760	16.7980	116797.1610	10970.7009
11	Dribbling	addition	0.761	0.760	14.7130	116795.0688	10967.5892
12	LongPassing	addition	0.761	0.760	13.0560	116793.4036	10964.9084
13	Height	addition	0.761	0.760	11.3280	116791.6638	10962.1533

As the model performs equally good at testing data, we conclude that 74-75% of the variance in the wage of the player data can be explained by the following variables shown above in the attachments.

5 Clustering Validation Analysis

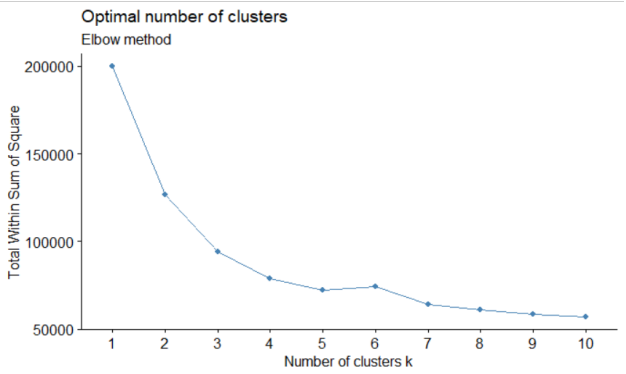
5.1 Clustering on all players based on significant physical attributes, playing skills, reputation

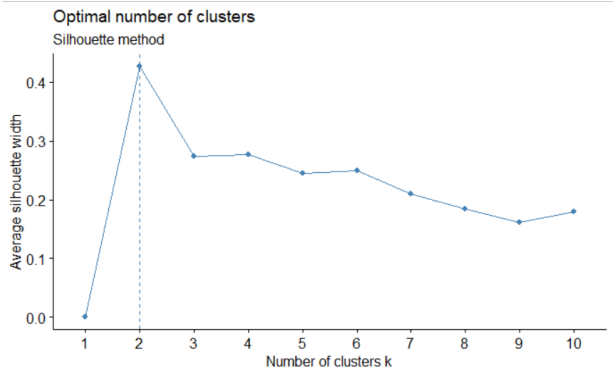
The significant variables for a player playing ability and rating were recognised from stepwise regression and these are the features:

```
"Age"                    "International Reputation"  "Crossing"
"HeadingAccuracy"        "Volleys"                   "Curve"
"LongPassing"            "BallControl"               "Reactions"
"ShotPower"              "Stamina"
```

a. Assessing number of clusters

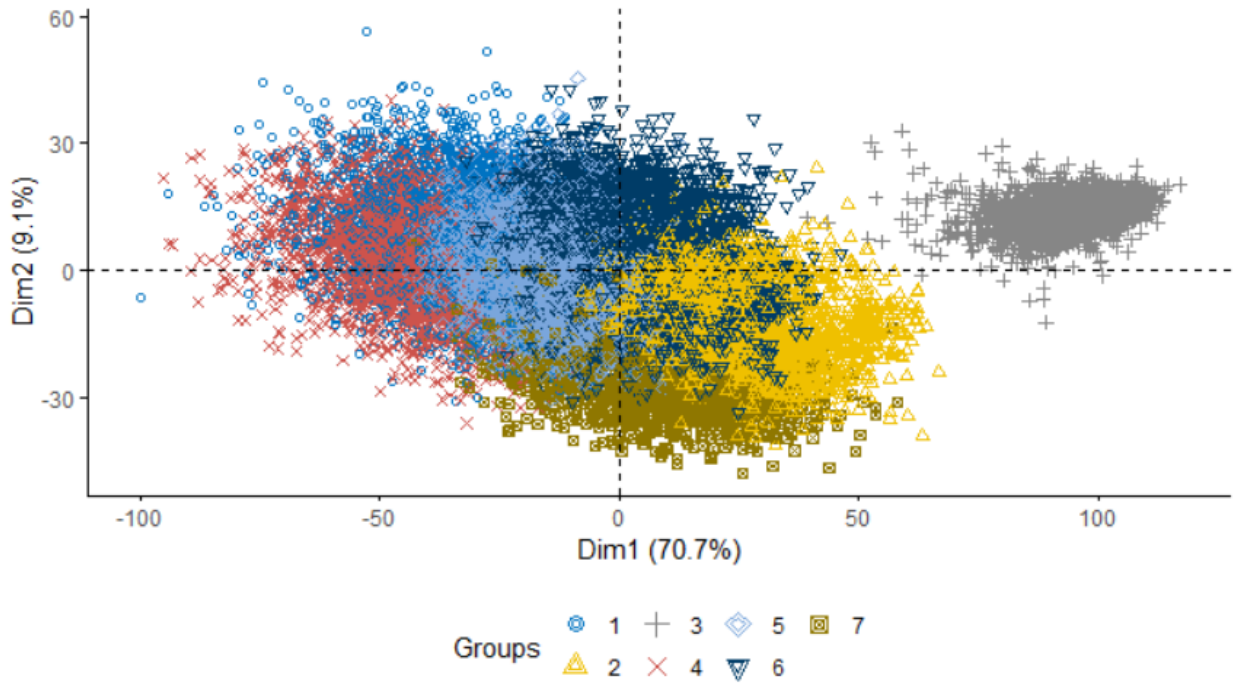
- Elbow method = 7
- Silhouette Method = 2





b. Quality of clusters

- Kmeans
    1. WCSS – 68%
    2. Silhouette Width – 0.26
    3. Total Sum of Squares – 199738
    4. Total within-cluster Sum of Squares – 63425



c. Cluster Label and Inferences

The seven clusters formed were analysed, and the confusion matrix was created for mapping cluster label to the number of players playing in that each position.

	DEF	FWD	GK	MID
1	377	353	1	636
2	1098	1149	0	2685
3	5	0	1015	0
4	0	0	1007	0
5	1554	609	2	545
6	2113	292	0	543
7	731	1015	0	2429

From the above confusion matrix, we can infer that **cluster 3 and 4** consists entirely of goalkeepers only. **Cluster 1, 2** and **7** consists majorly of midfielders along with defenders and forwards. **Cluster 5** consists of defenders whereas **cluster 6** have forward players and attacking midfielders.

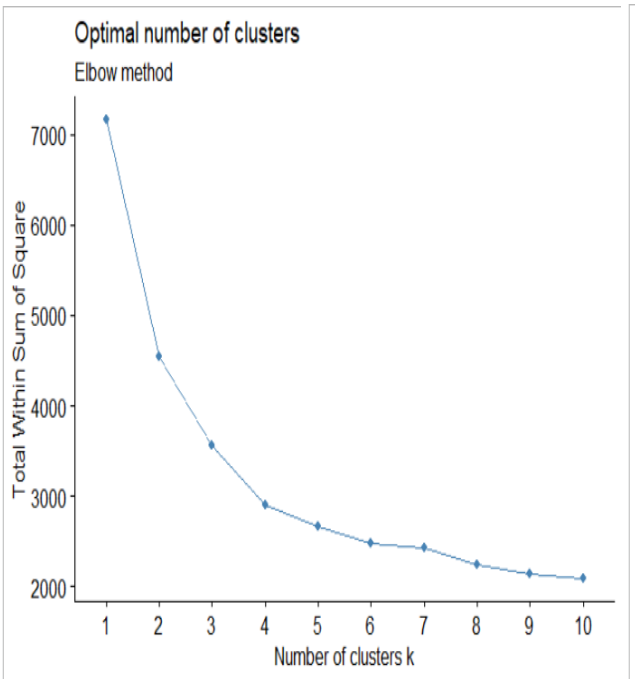
We dived deeper into the cluster results to get a better understanding of the cluster that were formed, and the inferences are tabulated below.

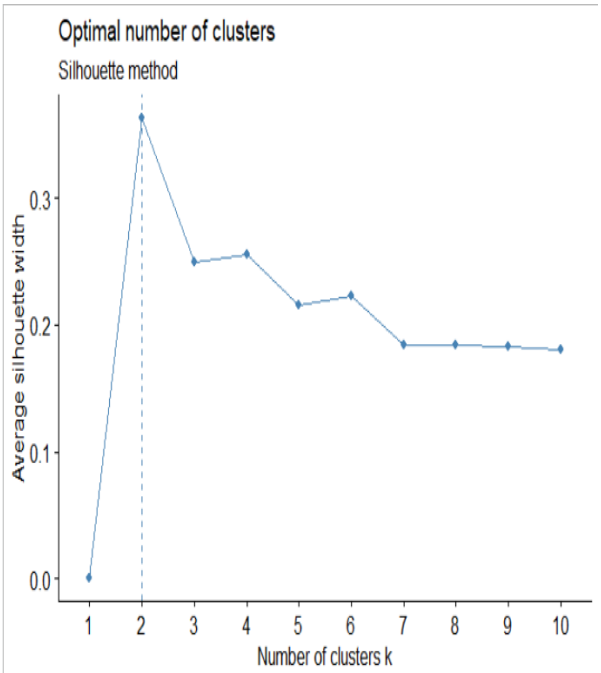
Group.1	Name	Age	Nationality	Overall	Potential	Club	Value	Wage
1	NA	28.86832	NA	77.23628	78.29554	NA	12488152.9	49410.388
2	NA	27.34712	NA	70.15349	72.11071	NA	2997724.0	11286.091
3	NA	30.10980	NA	69.50000	70.84118	NA	2713906.9	11397.059
4	NA	21.96723	NA	59.58590	68.66435	NA	401847.1	2011.917
5	NA	20.77417	NA	58.06568	68.80701	NA	306660.5	1825.092
6	NA	27.60041	NA	67.27782	69.37720	NA	1575880.3	6776.119
7	NA	21.88359	NA	63.44120	71.85868	NA	806912.6	3668.743

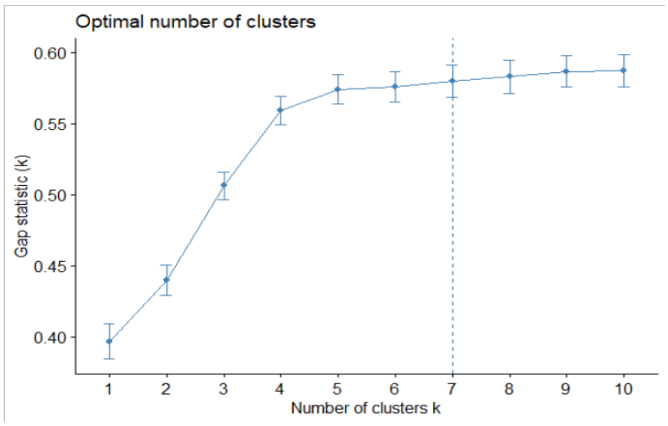
Comparing these results with our complete clustering performed in 4.1, It is observed that clustering based on only significant features results in clusters more evenly distributed among different playing positions than the clusters formed with all features.

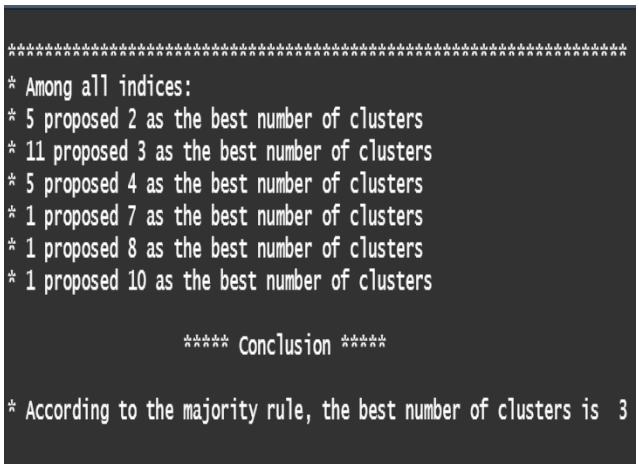
5.2 Cluster validation using sampled data.

Here, we took 50 random samples from each playing positions i.e. FWD, MID, DEF, GK and performed clustering to observe how it performs. It includes the same features for clustering as used in 4.1.







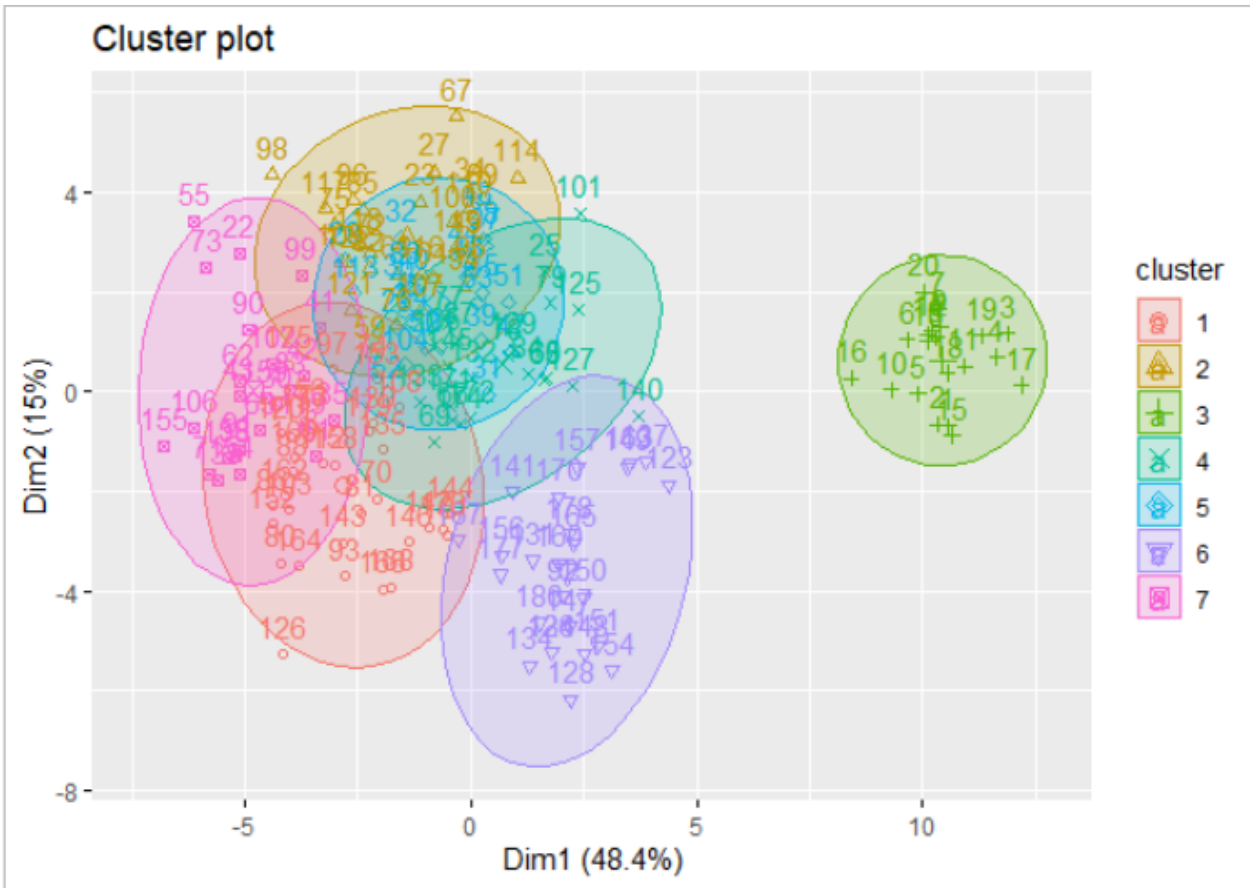


a. Assessing number of clusters
    a. Elbow method = 4
    b. Silhouette Method = 2
    c. Gap Stat Method = 7
    d. NBClust Method= 2

b Quality of clusters

i. Kmeans
   1. WCSS – 67.3%
   2. Silhouette Width – 0.22
   3. Total Sum of Squares – 7160
   4. Within Sum of Squares – 2343.098

ii. Hierarchical
   1. Agglomerative Coefficient – 0.85



	DEF	FWD	GK	MID
1	20	0	0	15
2	0	4	0	25
3	0	0	20	0
4	10	1	0	11
5	0	21	0	3
6	24	0	0	1
7	4	8	0	13

From the above confusion matrix, it is evident that **cluster 3** consisted entirely of goalkeepers. **Cluster 6** was majorly formed by defenders only, whereas **cluster 5** consisted of forward players. **Cluster 1** and **cluster 4** are made of defenders and defending midfielders. Also,

Also, cluster 2 comprised entirely of mid-fielders, that is, the players that play mostly in the central positions such as CM, RM, or LM. **Cluster 7** can be considered of players that are attacking midfielders.

Comparing this analysis with the 4.1, the sampled data signified 7 clusters, it may not be clustered numerically correct however the meanings of clusters are almost the same as before. The clusters are more well-defined than before as the sample size was small.

6 Conclusion

The cluster analysis performed for the full dataset (refer section 4.1) infers that **cluster 4** consists entirely of goalkeepers only. **Cluster 3** and **cluster 7** consists majorly of defenders. **Cluster 2** consists of defenders and defense-minded midfielders, whereas **cluster 5** and **cluster 6** have forward players and attacking midfielders. **Cluster 1** on the other hand consists of players that play mid-midfield, that is, they play on positions such as CM, RM or LM. These results were validated with 200 random sampled data. The validation analysis shows clear distinction among clusters and were comparable to the results obtained with clustering on full data. Thus cluster analysis performed here should help managers and clubs to scout player replacements based on their playing style, physical attributes such as Age, Stamina, strength, speed etc. and their market value. However, this technique solely should not be used as others factors depending on the requirements might change and should also be considered.

In our regression analysis, we performed a stepwise regression to predict a players wage based on the 17 significant features such as age, weight, FKAccuracy, Strength and Stamina etc. This method should predict the wage of the player accurately for the average rated player, as our data included mostly these players and also most players fall under this category. The wage for the players which are highly rated like Messi, Ronaldo, etc. should be predicted considering market prices of similar players.

Appendix

Data Dictionary

Column Name	Description	Range
1. Name	Name of player	*Messi, Ronaldo etc.*
2. Age	Age of player	*16 to 45*
3. Photo	Link for player's photo	
4. Nationality	Nationality of player	*England, Spain, Germany etc.*
5. Flag	Link for player nationality flag	
6. Overall	Player's overall rating	*46 to 94*
7. Potential	Player's potential	*48 to 95*
8. Club	Club for which a player play	*Arsenal, Dortmund etc.*
9. Club Logo	Link for club logo	
10. Value	Sum of value of player	*100K, 105K, 10K etc.*
11. Wage	Wage of player	*1.1M, 375K, 425K etc.*
12. Special		
13. Preferred	Left or Right foot	
14. International	International reputation	*1 to 5*
15. Weak Foot	Rating of player's weaker foot	*1 to 5*
16. Skill Move	Rating of player's skill level	*1 to 5*
17. Work Rate	Rate of player's behaviour on pitch	*Medium/Medium, High/Medium etc.*
18. Body Type	Player's type of body	*Normal, Lean etc.*
19. Real Face	Player's real face	*Yes/No*
20. Position	Preferred position	*ST, GK, LF, CF etc.*
21. Jersey Number	Jersey number	*1 to 99*
22. Joined	Date of joining	*Apr 1, 2008, Apr 1, 2011 etc.*

23. Loaned from	Team from which player is loaned from.	*West Ham United, Walsall etc*
24. Contact Valid Until	Expiring agreement of player	*2019,2021,2020,2022 etc*
25. Height	Height of player	*5'1 – 6'9*
26. Weight	Weight of player	*110lbs – 243lbs*
27. LS	Left Striker	*61+2, 60+2, 59+2 etc.*
28. ST	Striker	*61+2, 60+2, 59+2 etc.*
29. RS	Right Striker	*61+2, 60+2, 59+2 etc.*
30. LW	Left winger	*61+2, 62+2, 63+2 etc.*
31. LF	Left forward	*61+2, 62+2, 63+2 etc.*
32. CF	Center forward	*61+2, 62+2, 63+2 etc.*
33. RF	Right forward	*61+2, 62+2, 63+2 etc.*
34. RW	Right winger	*61+2, 62+2, 63+2 etc.*
35. LAM	Left attacking midfield.	*61+2, 62+2, 63+2 etc.*
36. CAM	Center attacking midfield.	*61+2, 62+2, 63+2 etc.*
37. RAM midfield.	Right attacking	*61+2, 62+2, 63+2 etc.*
38. LM	Left midfield	*61+2, 62+2, 63+2 etc.*
39. LCM	Left center midfield	*61+2, 60+2, 59+2 etc.*
40. CM –	Center midfield	*61+2, 60+2, 59+2 etc.*
41. RCM –	Right center midfield	*61+2, 60+2, 59+2 etc.*
42. RM –	Right midfield	*61+2, 62+2, 63+2 etc.*
43. LWB –	Left wing back	*61+2, 60+2, 59+2 etc.*
44. LDM –	Left defensive midfield.	*61+2, 60+2, 59+2 etc.*
45. CDM –	Center defensive midfield	*61+2, 60+2, 59+2 etc.*
46. RDM –	Right defensive midfield	*61+2, 60+2, 59+2 etc.*
47. RWB –	Right wing back	*61+2, 60+2, 59+2 etc.*

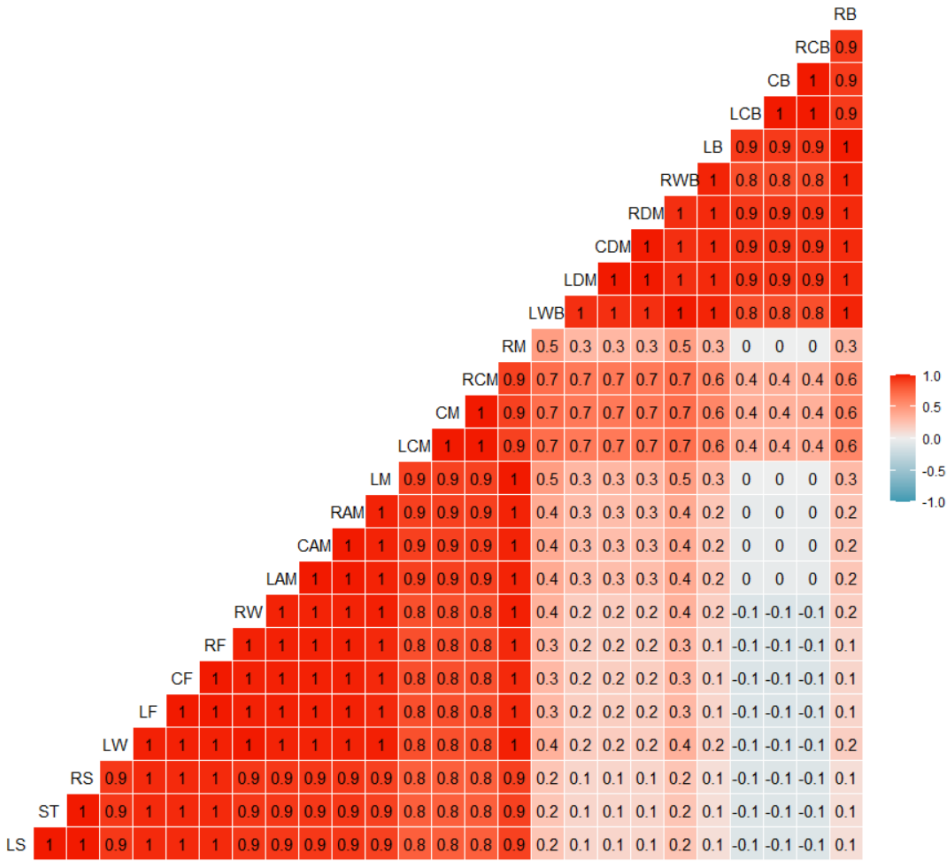
48. LB –	Left back	*61+2, 58+2, 59+2 etc.*
49. LCB –	Left center back	*61+2, 62+2, 63+2 etc.*
50. CB –	Center back	*61+2, 62+2, 63+2 etc.*
51. RCB –	Right center back	*61+2, 62+2, 63+2 etc.*
52. RB –	Right back	*61+2, 64+2, 63+2 etc.*
53. Crossing	Passing the ball from side to middle of the field	*5 to 93*
54. Finishing	Timed shooting	*2 to 95*
55. Heading Accuracy	Ability to head onto target.	*4 to 94*
56. Short Passing	Player's ability to performs short pass.	*7 to 93*
57. Volleys	Player's ability to performs volley.	*4 to 90*
58. Dribbling	Player's ability to dribble the ball.	*4 to 97*
59. Curve	Player's ability to curve the ball.	*6 to 94*
60. FK Accuracy	Player's accuracy to perform free kick.	*3 to 94*
61. Long Passing	Player's ability to perform a long pass.	*9 to 93*
62. Ball Control	Player's ability to control the ball.	*5 to 96*
63. Acceleration	Player's running speed on pitch.	*12 to 97*
64. Sprint Speed	Speed rate of player's sprinting	*12 to 96*
65. Agility	Player's ability to control the ball	*14 to 96*
66. Reaction	Player's ability to respond to situation around him.	*21 to 96*
67. Balance	Player's ability to remain steady when running,	*16 to 96*

	carrying, and controlling the ball.	
68. Shot Power	Rate of power to shoot the ball	*2 to 95*
69. Jumping	Ability and quality of jump	*15 to 95*
70. Stamina	Rate at which player will get tired during the match.	*12 to 96*
71. Strength	Quality of being physically strong.	*17 to 97*
72. Long Shots	Accuracy of shot from outside the penalty area.	*3 to 94*
73. Aggression	Rate of aggression of jostling, tacking and slit tackling.	*11 to 95*
74. Interception	Ability to read the game and passes.	*3 to 92*
75. Positioning	Ability to position during the game.	*2 to 95*
76. Vision	Player's awareness of the position of his teammates and opponents around him.	*10 to 94*
77. Penalties	Accuracy of shot inside the penalty area	*5 to 92*
78. Composure	. Distance at which player starts feeling pressure from the opponent.	*3 to 96*
79. Marking	Ability to mark an opposition player.	*3 to 94*
80. Standing Tackle	Ability to win the ball through stand tackle.	*2 to 93*
81. Sliding Tackle	Ability to win the ball through slide tackle.	*3 to 91*
82. GK Diving	Goalkeeper's ability to dive in.	*1 to 90*
83. GK Handling	Goalkeeper's able to handle the ball.	*1 to 92*

84. GK Kicking	Goalkeeper's ability to perform kicking.	*1 to 91*
85. GK Positioning	Goalkeeper's ability to position themselves.	*1 to 90*
86. GK Reflexes	Goalkeeper's ability to catch/safe the ball.	*1 to 94*
87. Release Clause	Provision within a mortgage contract.	*1.1M, 1.3M, 1.5M etc*

Figures

1) **Fig 3.4.1**

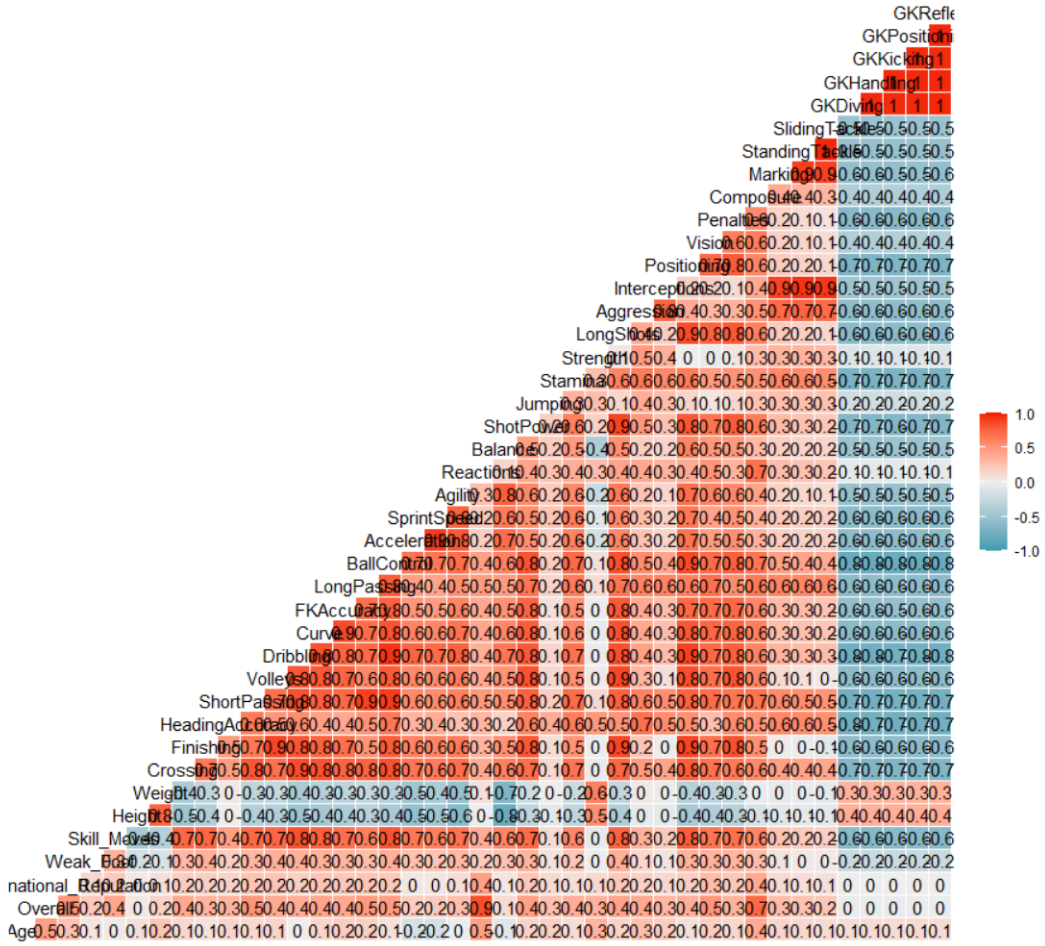


Analysis on Finding Significant Features

To find the significant features among all which contributes most to players rating, below all features were examined.

```
> colnames(df)
 [1] "Age"                  "Overall"              "International Reputation"
 [4] "Weak Foot"            "Skill Moves"          "Height"
 [7] "Weight"               "Crossing"             "Finishing"
[10] "HeadingAccuracy"      "ShortPassing"         "Volleys"
[13] "Dribbling"            "Curve"                "FKAccuracy"
[16] "LongPassing"          "BallControl"          "Acceleration"
[19] "SprintSpeed"          "Agility"              "Reactions"
[22] "Balance"              "ShotPower"            "Jumping"
[25] "Stamina"              "Strength"             "LongShots"
[28] "Aggression"           "Interceptions"        "Positioning"
[31] "Vision"               "Penalties"            "Composure"
[34] "Marking"              "StandingTackle"       "SlidingTackle"
[37] "GKDiving"             "GKHandling"           "GKKicking"
[40] "GKPositioning"        "GKReflexes"
```

This matrix shows the correlation among the features.



After running stepwise regression on all features related to physical strengths, playing skills, and popularity with Overall rating as dependent variable. The final model stats are:

Final Model Output

Model Summary

R	0.934	RMSE	2.477
R-Squared	0.872	Coef. Var	3.739
Adj. R-Squared	0.872	MSE	6.135
Pred R-Squared	0.871	MAE	1.951

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	756969.676	34	22263.814	3628.738	0.0000
Residual	111198.261	18124	6.135		
Total	868167.937	18158			

The model equation is

Overall = Age + International reputation + Crossing + Heading Accuracy + Volleys+ Curve + Long passing + Ball Control + Reactions + Shot power + Stamina