UNSUPERVISED LEARNING

FIFA 23 PLAYER ATTRIBUTES DATASET

1. Abstract:

The aim of this report is to demonstrate how to apply an unsupervised learning algorithm on a given data set, using different clustering and dimension reduction techniques.

The selected dataset is FIFA 23 Player attributes, where the aim is to observe the variables and determine how players can be clustered together. The dataset is composed of 44 variables, 2 of which are categorical (player name and player's best position) and the rest of the variables are players attributes, how each player is rated on a given attribute.

The Curse of Dimensionality is a serious problem, and our dataset suffers from high-dimensionality. To avoid such a problem, Principal Component Analysis will be used before clustering, in order to reduce number of dimensions, but also to reduce unnecessary noise of data. This will possibly improve the clustering performance for the following methods that are going to be used:

- K-means clustering
- Hierarchical clustering

2. Data Description:

The dataset is composed of 44 variables which are the following:

Full Name	Crossing	Acceleration	Aggression	Goalkeeper Handling
Best Position	Finishing	Sprint Speed	Interceptions	
Weak Foot Rating	Heading Accuracy	Agility	Positioning	Goalkeeper Positioning
Skill Moves	Short Passing	Reactions	Vision	
Pace Total	Volleys	Balance	Penalties	
Shooting Total	Dribbling	Shot Power	Composure	Goalkeeper Reflexes
Passing Total	Curve	Jumping	Marking	
Dribbling Total	Freekick Accuracy	Stamina	Standing Tackle	Goalkeeper Kicking
Defending Total	Long Passing	Strength	Sliding Tackle	
Physicality Total	Ball Control	Long Shots	Goalkeeper Diving	

Firstly, the dataset has been checked see if it suffers from any na/null values and there was none. Then some adjustments have been made where the player names have been replaced by number (nrows) since the names will be irrelevant to our analysis.

3. Principal Component Analysis:

PCA was used in order to have a better visualization with the aim of producing a low dimensional representation of a dataset. PCA' goal is to find a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated. PCA finds the best fitting line by maximizing the sum of squared distances from the projected points to the origin, once found that line is considered as a principal component.

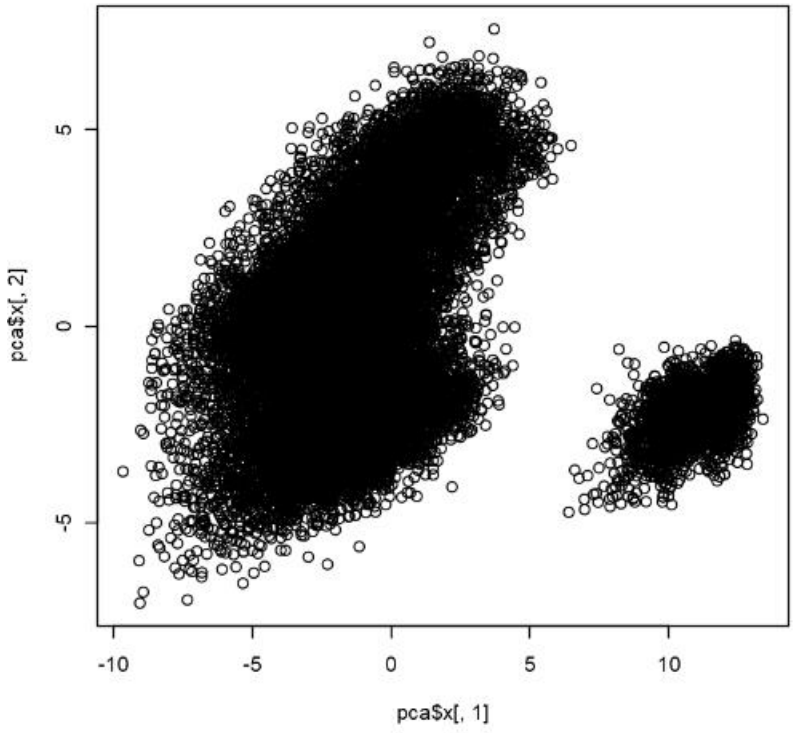

Figure 1

Figure 1 shows how the data points are spread across the highest 2 dimensions (PC1, PC2), from here two main clusters are identified. However, a more graphical representation of the percentages of variation that each PC accounts for is required in order to determine the optimal number of dimensions that would best explain our dataset.

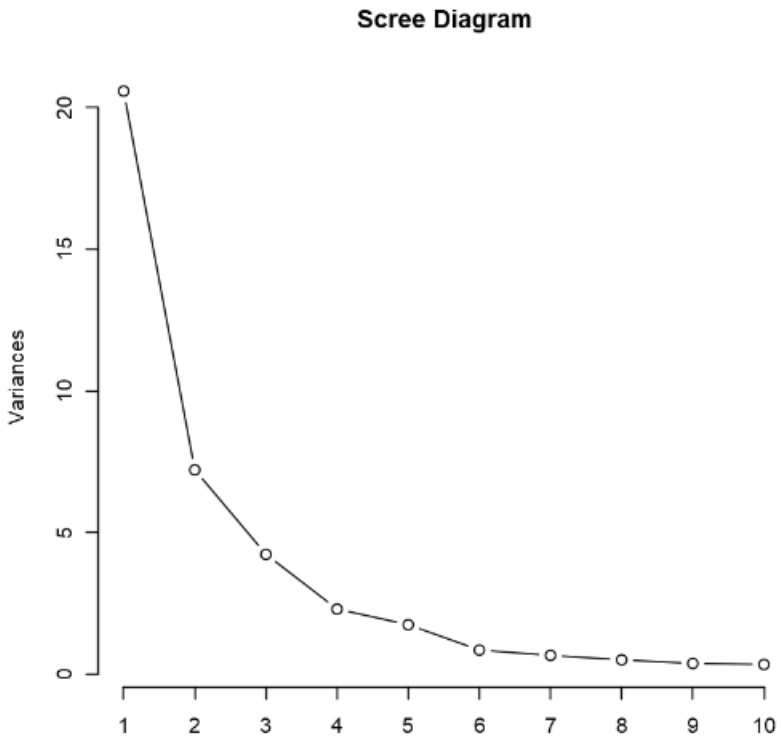

Figure 2

Figure 2 represents a scree plot which demonstrates the percentage of variation of each component.

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6    PC7
Standard deviation     4.5371 2.6848 2.0558 1.51426 1.31921 0.92019 0.8120
Proportion of Variance 0.4901 0.1716 0.1006 0.05459 0.04144 0.02016 0.0157
Cumulative Proportion  0.4901 0.6617 0.7624 0.81696 0.85840 0.87856 0.8943
                          PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.7071 0.6114 0.58185 0.57273 0.51982 0.49747 0.48883
Proportion of Variance 0.0119 0.0089 0.00806 0.00781 0.00643 0.00589 0.00569
Cumulative Proportion  0.9062 0.9151 0.92312 0.93093 0.93737 0.94326 0.94895
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.45504 0.45256 0.43571 0.40781 0.39891 0.36713 0.35387
Proportion of Variance 0.00493 0.00488 0.00452 0.00396 0.00379 0.00321 0.00298
Cumulative Proportion  0.95388 0.95875 0.96327 0.96723 0.97102 0.97423 0.97721
                          PC22    PC23    PC24   PC25    PC26    PC27   PC28
Standard deviation     0.33374 0.32774 0.29297 0.2593 0.25844 0.25317 0.2421
Proportion of Variance 0.00265 0.00256 0.00204 0.0016 0.00159 0.00153 0.0014
Cumulative Proportion  0.97987 0.98242 0.98447 0.9861 0.98766 0.98918 0.9906
                          PC29    PC30    PC31    PC32    PC33    PC34    PC35
Standard deviation     0.23701 0.22793 0.21842 0.18874 0.18119 0.17850 0.16803
Proportion of Variance 0.00134 0.00124 0.00114 0.00085 0.00078 0.00076 0.00067
Cumulative Proportion  0.99192 0.99315 0.99429 0.99514 0.99592 0.99668 0.99735
                          PC36    PC37    PC38    PC39    PC40    PC41    PC42
Standard deviation     0.16138 0.15691 0.14968 0.13776 0.10897 0.07132 0.04780
Proportion of Variance 0.00062 0.00059 0.00053 0.00045 0.00028 0.00012 0.00005
Cumulative Proportion  0.99797 0.99856 0.99909 0.99954 0.99982 0.99995 1.00000
```

Figure 3

Figure 2 and 3 provides us with a good indication that the optimal number of dimensions that would best explain the total variation is 3.

However, more analysis is performed using only 2 dimensions to have a general idea about how the dataset is represented.

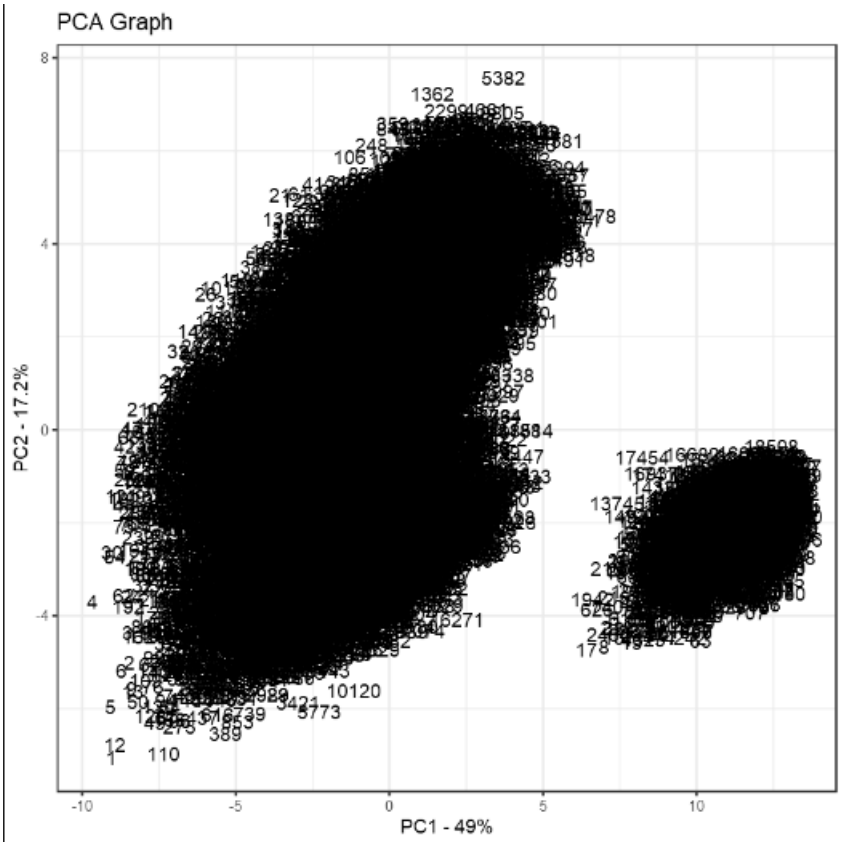

Figure 4

In Figure 4, The x-axis tells us what percentage of the variation in the original data that PC1 accounts for (49%), while the y-axis tells us what percentage of the variation in the original data that PC2 accounts for (17.2%). Another observation that was checked is the loading scores of each dimension, which were the top 10 per PC. The loading scores describe how the values are projected onto the PCs, general example: Interceptions is 4 times more important than crossing in PC2.

For instance, for PC1, the top 10 attributes were 'BallControl' 'Dribbling', 'Short.Passing', 'Positioning', 'Crossing', 'Curve', 'Long.Shots' 'LongPassing', 'Goalkeeper.Diving', 'Freekick.Accuracy'. From this a general observation can be made that some of the top 10 attributes describe a player's ability with football, in other words how comfortable the player is with the ball.

As per PC2, the top 10 attributes were 'Shooting.Total', 'Sliding.Tackle', 'Defending.Total', 'Standing.Tackle', 'Interceptions', 'Marking', 'Dribbling.Total', 'Aggression', 'Pace.Total',

'Passing.Total'. Another general observation can be made here that some of the top 10 attributes here describe the player's defensive ability.

Regarding PC3, the top 10 attributes were 'Reactions', 'Passing.Total', 'Physicality.Total', 'Dribbling.Total', 'Strength', 'Goalkeeper.Positioning', 'Shooting.Total', 'Goalkeeper.Handling', 'Goalkeeper.Reflexes', 'GoalkeeperKicking', here we can make the observation that some of the top attributes here describe the player's goalkeeping abilities.

To visualize this a biplot was used (only on two dimensions):

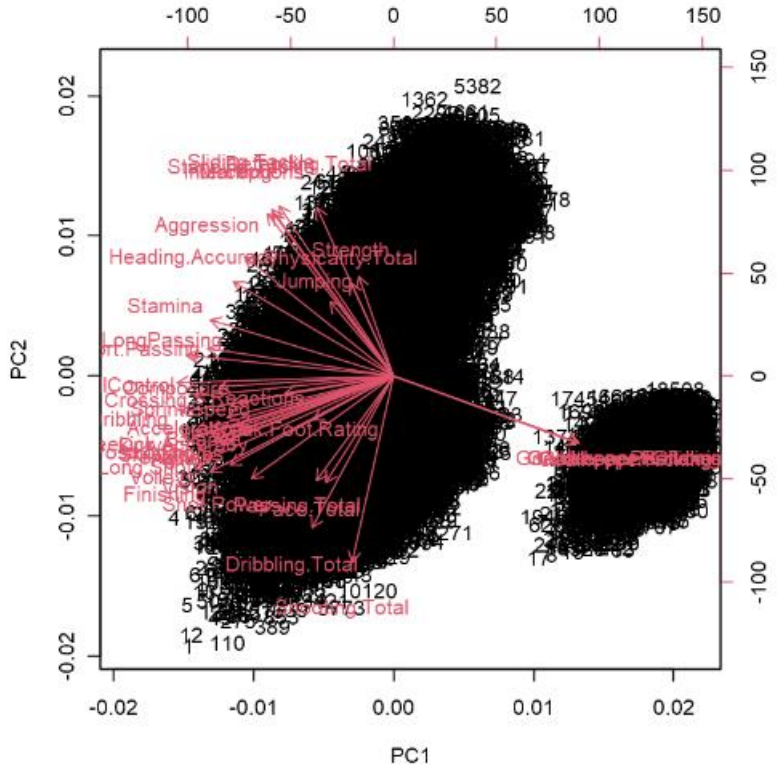

Figure 5

From the biplot graph (Figure 5) few interesting observations can be made:

- On the right-hand corner, we can see that the goalkeepers are clustered together.
- The other main cluster we can observe as well is on the left, which represents the field players Looking at the loading vectors we can also notice the following:
- on the top left corner, we can see the following attributes: sliding tackle, defending total, standing tackle, interceptions, marking, aggression, strength... All of these attributes mainly describe defenders; hence we can infer that the top part of the left cluster is mostly composed of defenders

- As we go down the left clusters, we can also determine that the players become classified as more offensive, based on the following attributes we can see on the bottom left corner: finishing, shooting total, passing total, pace total, vision, long shots, dribbling...
- An interesting way to confirm what we've just observed is to plot the PCA according to the players positions

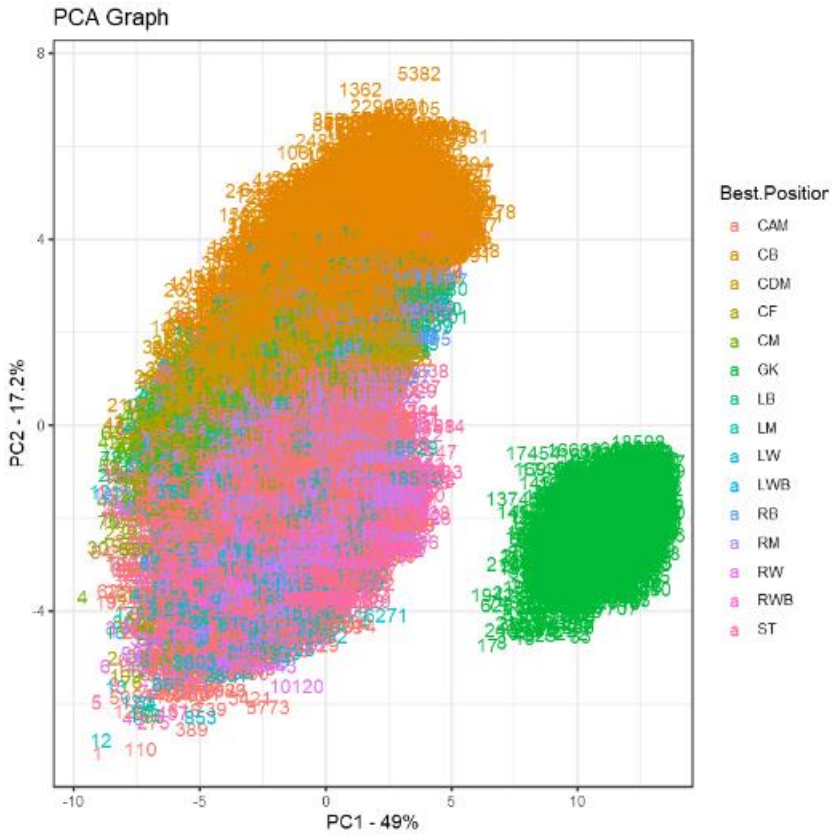

Figure 6

Again, we can clearly see that the goalkeepers are grouped in one cluster while in the other cluster, the top part is mainly formed with defenders and the bottom part is mainly attackers. What can also be interesting to note is that we have wingbacks all over the cluster, which makes sense because in football usually there would be wingbacks who are more offensive and other wingbacks are more defensive. How a player is categorized as offensive or defensive is based on how they play in real life, so the style of play could be influenced by several factors, such as formation, league... It would be interesting to retrieve a dataset where we can identify which leagues for example have the most players who prefer to play an offensive style of play, but that's an analysis to be conducted some other time.

A final interesting 3D visualization of our PCA plot, showing PC1, PC2 and PC3 to get a better idea of the variation in the dataset.

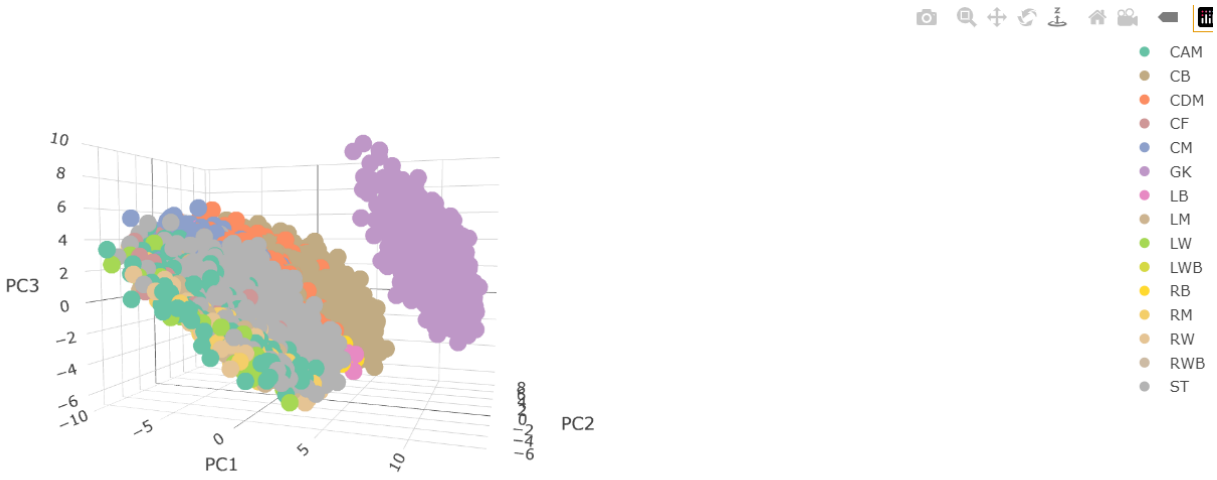

Figure 7

4. Clustering:

a. K-means Clustering:

Based on our general observation from running the PCA algorithm, we observe that the players are separated into two groups: players(attackers, midfielders, and defenders) and goalkeepers, and the players are separated into two main parts, offensive and defensive players. Let's try and apply the k-means clustering at first using this observation to our advantage.

However, as we've decided, the optimal number of dimensions we selected after running the PCA algorithm is 3, we will come to that later...

10004	6474	2061

As we can see right away, the distribution of data points between the clusters is not even, so that might be an indicator telling us that we might need to change the number of our clusters. However, let's see where this will lead us.

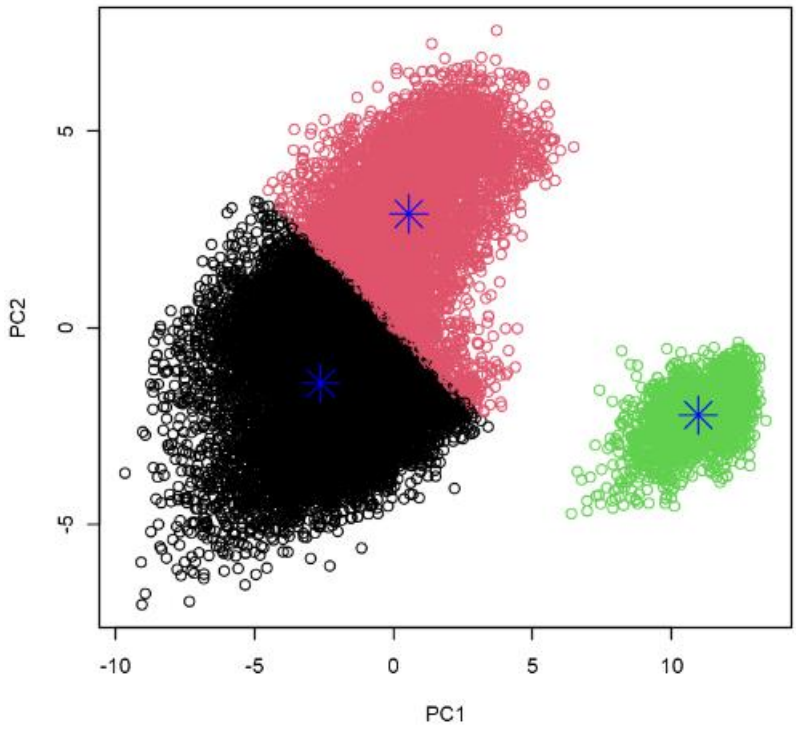

Figure 8

we can see in Figure 8 that the big cluster is divided into three subparts: offensive(black), defensive(red) and goalkeepers(green).

What we just did was select the number of clusters (k) based on our intuition. However, the question is, how can we make this more accurate? i.e., how to choose the best number of clusters? one way we can do that is to minimize the total within-cluster variation. Within-cluster variation for a single cluster can simply be defined as the sum of squares from the cluster mean, which in this case is the centroid we defined in the k-means algorithm. The total within-cluster variation is then the sum of within-cluster variations for each cluster. Using an Elbow plot, we can determine which is the best number of clusters that we think we will be satisfied with.

515223.41	82940.74	41057.52	29796.10	22806.43
224685.31	58879.42	36006.31	27255.02	21089.82
112342.36	47869.10	32737.52	24957.88	19875.32

Figure 9

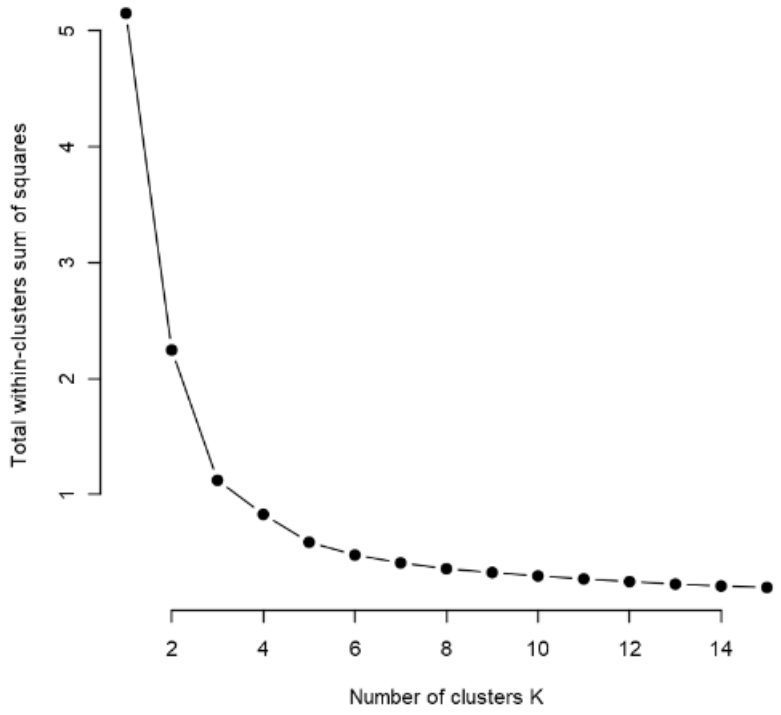

Figure 10

Looking at both the elbow plot & the total within-cluster variation scores (Figures 9 & 10), It seems like the Total_ss tends to change slowly starting from k = 6.

3339	3937	2061	2918	2847	3437

It seems like now we have an average number of datapoints in each cluster.

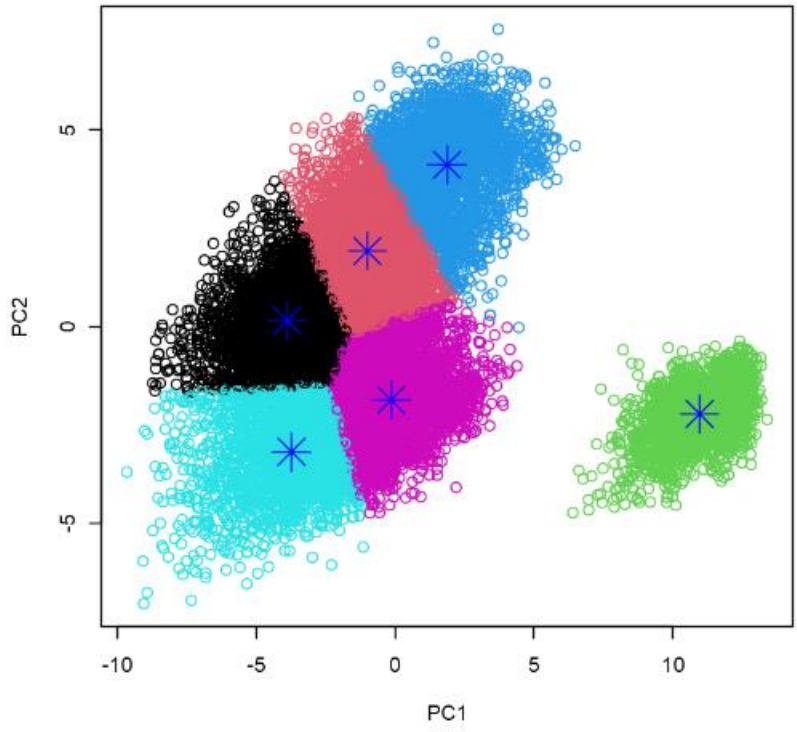

Figure 11

Seeing how the clusters are formed, we can clearly see the goalkeepers are identified as one cluster and for the others, they represent how players are less defensive and more offensive gradually. For instance, the very top cluster seems to be representing the defenders, and as we go down, the defensive attributes of the players decrease gradually while the offensive attributes increase gradually as well.

The same thought process we just did but using the optimal number of dimensions (3).

Starting with just three clusters at the beginning.

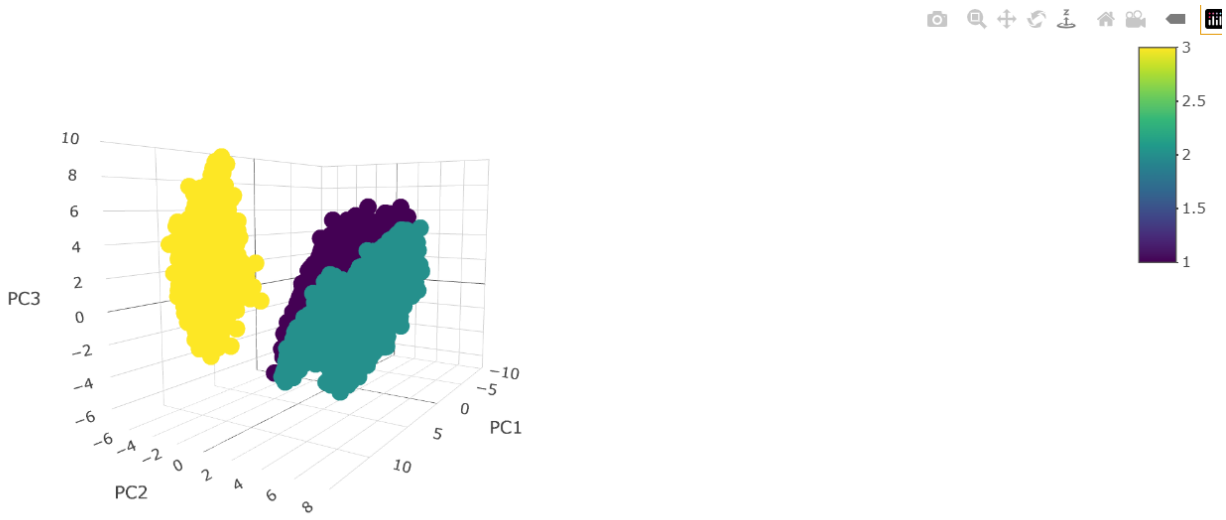

Figure 12

Determining the optimal number of clusters using both the elbow plot & the total within-cluster variation scores.

593573.70	122928.05	73235.80	52837.47	41434.48
291672.49	98027.75	65144.95	47742.42	38802.89
177417.20	82648.21	58534.96	44399.07	36457.62

Figure 13

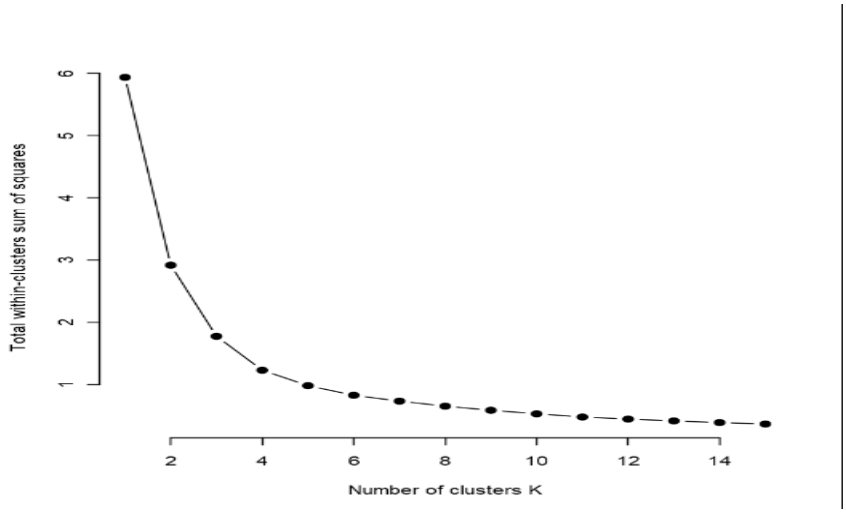

Figure 14

Looking at both the elbow plot & the total within-cluster variation scores, and again it seems like the Total_ss tends to change slowly starting from k = 6. We can observe that we have again a some-what average number of datapoints in each cluster.

2512	4375	2061	3135	3439	3017

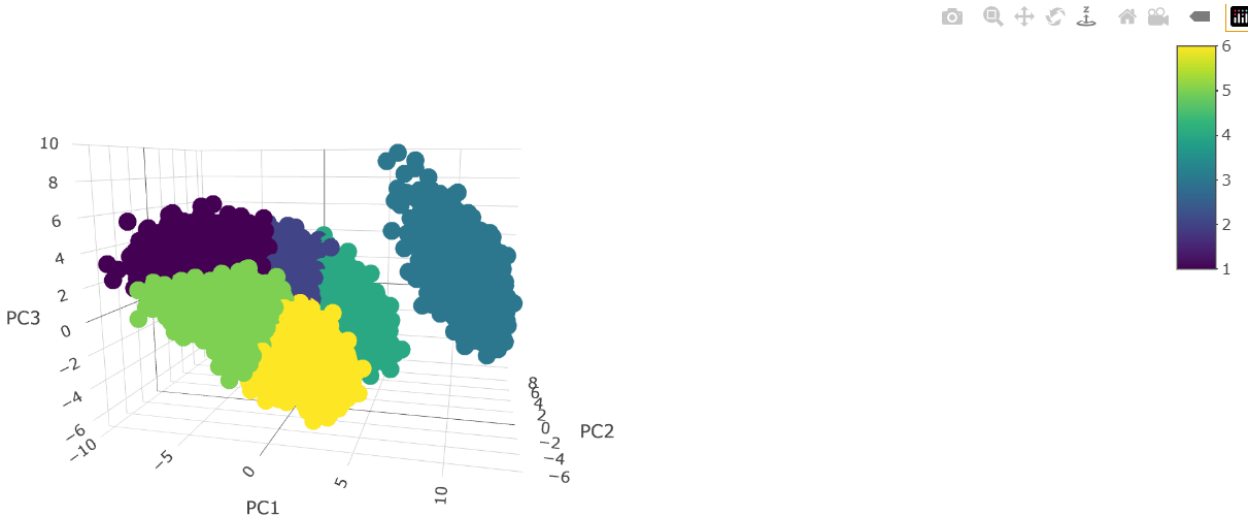

Figure 15

Figure 15 gives us the same observation as the observation that was made under Figure 11. Another clustering method will be applied but without pre-specifying the number of clusters and in order to compare results.

b. Hierarchical Clustering:

We're going to use the PCA scores to our advantage, where we have been able to reduce the dimensionality of our dataset to just 3 dimensions. First let's apply hierarchical clustering using only two dimensions to have a general idea on how the dataset will be clustered.

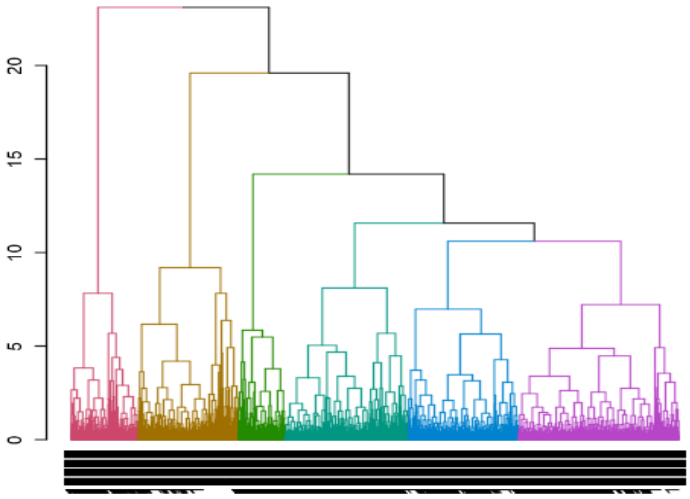

Figure 16

we can observe from the dendrogram (Figure 16) that there are 6 main clusters (same as what we found out using k-means clustering).

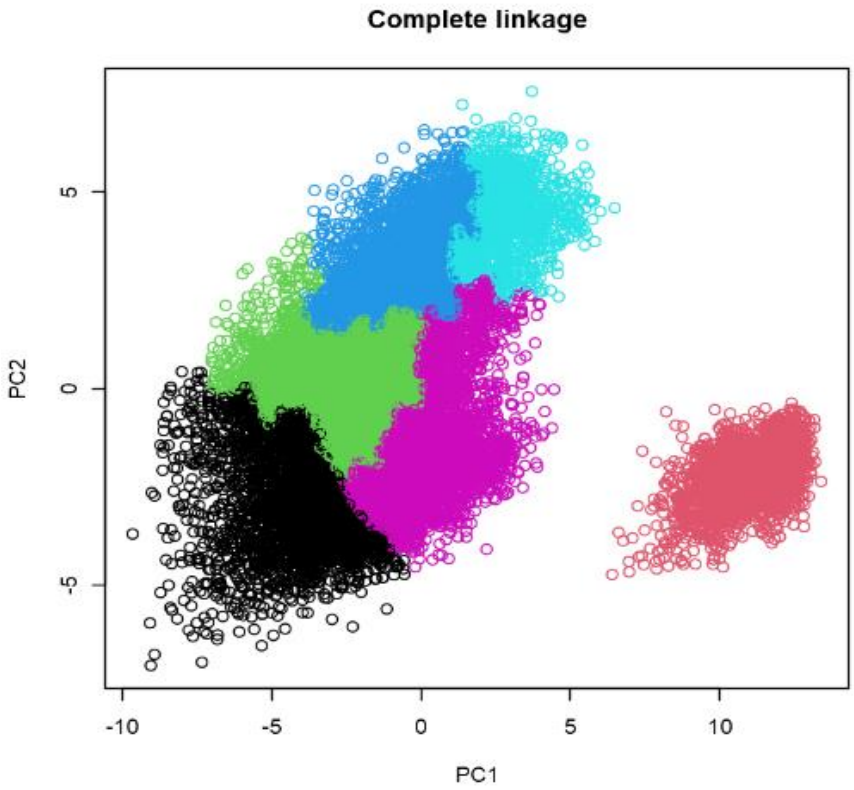

Figure 17

Similar to figure 11, we can see how the clusters are formed, where the goalkeepers are identified as one cluster and for the others, they represent how players are less defensive and more offensive gradually. As discussed earlier, the very top cluster seems to be representing the defenders, and as we go down, the defensive attributes of the players decrease gradually while the offensive attributes increase gradually as well. Now let's shift to using 3 dimensions and see how our data points (players) are clustered.

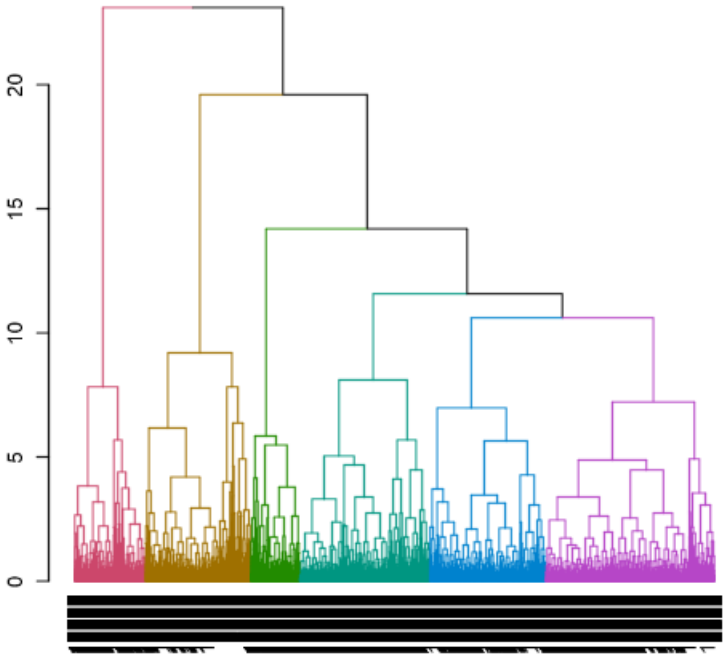

Figure 18

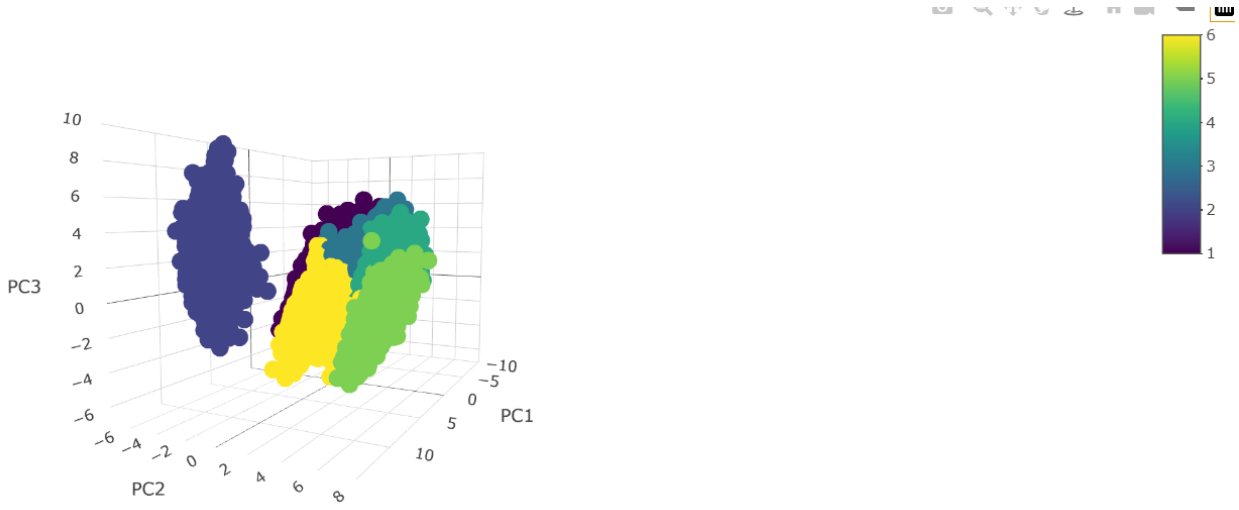

Figure 19

Figures 18 and 19 give us the same expected output as we have seen when only using two dimensions.

5. Conclusion:

We can distinguish that using k-means clustering and hierarchical clustering we end up with somewhat the same conclusion, which is, our datapoints are best divided into 6 main clusters. Where the first one is the goalkeepers, and the other 5 clusters represent the rest of the players. Each cluster represent a set of players with certain attributes as their strong features. For example, some players have higher scores in their defensive attributes than others, which means they're more suited to playing a more defensive playstyle, for example defenders or defensive midfielders. While others are more suited to playing a more offensive playstyle...

However, looking at the biplot we drew earlier in the PCA section, we can actually observe that being offensive or defensive player is a very general/broad observation. There are other measurements that describe the player's style of play, for example how creative is s/he, how versatile is s/he, how clinical is s/he. All of these measurements/metrics explain to us the meaning behind how the players are being clustered using both of the techniques we talked about earlier.