FIFA Player Clustering

BY

K Nikitha Reddy

1. BUSINESS UNDERSTANDING

Domain: Gaming

Sofifa is a website that provides information and statistics about players in the FIFA video game series. It is a fan-made website that is not affiliated with the FIFA video game franchise or its publisher, EA Sports. On the website, users can search for and view detailed information about players in the game, including their ratings, positions, nationalities, and attributes. The website also allows users to compare players and create custom lists and teams. Sofifa gets its data from the FIFA video games and updates its database when new versions of the game are released. It is a popular resource for players of the FIFA series who want to stay up to date on the latest player statistics and build the best possible teams in the game.

Player statistics can be an important factor in building a good team in Sofifa. By looking at a player's ratings and attributes, you can get an idea of their strengths and weaknesses and how they might fit into your team's tactics and style of play.

For example, if a gamer looking for a striker, they might want to look for a player with high ratings in attributes like shooting, pace, and dribbling. If you are looking for a central midfielder, you might prioritize attributes like passing, tackling, and positioning. It's also important to consider a player's overall rating, as this is an indication of their overall quality and how they are likely to perform in the game. However, overall ratings are just one factor to consider, and it's important to look at a player's individual attributes as well.

a. Problem Identification

There is a need to identify patterns and trends in the attributes and overall ratings of players in the FIFA video game series, and group players into clusters based on these characteristics in order to better understand the strengths and weaknesses of different types of players and how they might fit into different tactics and formations.

By clustering of players in FIFA video game series, we can get to know which player belongs to which profile since people might want to have that player in their team to win the game. So, this report identifies different clusters which the player belongs to by considering different series i.e. (FIFA20, FIFA21, FIFA22, FIFA23 (recently released)) and by looking into how cluster sizes change over a period of time.

In the context of players in a video game like Sofifa, clustering can be useful for a number of reasons:

- **To identify patterns and trends:** By clustering players into groups with similar characteristics, it can be easier to identify patterns and trends in the data. This can help understand the strengths and weaknesses of different types of players and how they might fit into different tactics and formations.
- **To make comparisons:** Clustering players can allow to compare and contrast different groups of players, and to identify which players are the most similar or dissimilar to each other. This can be useful when building teams or making decisions about which players to sign.
- **To identify outliers**: Clustering players can help identify players that are unusual or stand out in some way (e.g. because they have exceptionally high or low ratings in certain attributes). These players may be worth paying closer attention to or may require further analysis to understand why they are different from other players.

b. Variables

The independent variables here are:

Sl. No	Variable	Description
1	NAME	Player Name in FIFA
2	AGE	Age of the player
3	OVA	Overall rating
4	POT	Player's Potential
5	HEIGHT	Height of the player
6	WEIGHT	Weight of the player
7	VALUE(IN M)	Financial value of the player
8	WAGE(IN K)	Amount of pay per week
9	CROSSING	Ability to deliver accurate passes
10	FINISHING	Ability to score goals
11	HEADINGACCURACY	Ability to direct the ball with their head
12	SHORTPASSING	Ability to make accurate short passes to their teammates
13	VOLLEYS	Ability to score goals with a shot taken while the ball is in the air.
14	DRIBBLING	Ability to control the ball and maintain possession while moving with it
15	CURVE	Ability to bend the flight of the ball
16	FKACCURACY	Ability to accurately take free kicks
17	LONGPASSING	Ability to make accurate long passes to their teammates
18	BALLCONTROL	Ability to control the ball and maintain possession
19	ACCELERATION	Ability to start and stop quickly, and to change direction rapidly
20	SPRINTSPEED	Ability to run at high speeds over short distances
21	AGILITY	Ability to change direction quickly and move with control
22	REACTIONS	Ability to react quickly to situations on the pitch
23	BALANCE	Ability to maintain their balance and control while moving with the ball
24	SHOTPOWER	Ability to hit powerful shots
25	JUMPING	Ability to jump high and win aerial duels
26	STAMINA	Ability to maintain their energy and endurance over the course of a match
27	STRENGTH	Skills that a player excels in different areas in game
28	LONGSHOTS	Ability to score goals from long range
29	AGGRESSION	Player's physicality and determination on the pitch
30	INTERCEPTIONS	Ability to anticipate and intercept passes and cut out opposition attacks
31	POSITIONING	Ability to get into good positions on the pitch
32	VISION	Ability to see and anticipate passes and scoring opportunities
33	PENALTIES	Ability to score goals from penalty kicks
34	COMPOSURE	Ability to remain calm and focused under pressure
35	MARKING	Ability to defend and mark opponents
36	STANDINGTACKLE	Ability to win tackles while standing up
37	SLIDINGTACKLE	Ability to win tackles by sliding on the ground
38	GKDIVING	Goalkeeper's ability to make diving saves to prevent goals
39	GKHANDLING	Goalkeeper's ability to catch and handle the ball
40	GKKICKING	Goalkeeper's ability to kick the ball accurately and with power
41	GKPOSITIONING	Goalkeeper's ability to position themselves correctly to make saves and prevent goals
42	GKREFLEXES	Goalkeeper's ability to make reflex saves to prevent goals
43	W_F	Ability of playing with a weaker foot
44	SM	Skill moves
45	IR	International Reputation
46	PAC	Pace/Diving
47	SHO	Shooting/Handling
48	PAS	Passing/Kicking
49	DRI	Dribbling/Reflexes
50	DEF	Defending/Pace
51	PHY	Physical/Positioning

c. Objectives

The objectives here are:

- To identify similar players in FIFA game series.
- To group players by position in FIFA game series.
- To compare players by various attributes and FIF series wise.
- To improve team performance by selecting players which belong to a particular cluster.

2. DATA UNDERSTANDING

The Data Understanding phase is where we focus on understanding the data. There are total of 51 columns in the dataset and among that 50 are continuous variables.

a. Data collection

The data was collected from the website https://sofifa.com/ . The fields were selected and scraped by using the webscraping tool Octoparse. It is commonly used for web scraping, which is the process of extracting data from websites in an automated manner. Since Octoparse has just a trail version of it for 10 runs, there were many duplicates generated which limits the data extracted. But for rest of the population the data was entered manually in the excel.

b. Data exploration

Since, different excel sheets are created because its FIFA series wise data. So, currently we will look at the EDA of FIFA 23 game series which was released on September 2022. Different player stats change for different series.

```
'data.frame':   326 obs. of  50 variables:
 $ AGE             : int  21 18 35 23 37 22 23 23 26 22 ...
 $ OVA             : int  80 73 91 77 88 70 83 80 77 79 ...
 $ POT             : int  87 88 91 82 88 78 88 86 80 86 ...
 $ HEIGHT          : int  180 176 169 176 187 185 189 187 188 170 ...
 $ WEIGHT          : int  77 71 67 72 83 75 76 73 80 71 ...
 $ VALUE.IN.M.     : num  42.5 7 54 16 0 3.6 55 33.5 14 36 ...
 $ WAGE.IN.K.      : int  13 19 195 52 0 16 29 34 40 110 ...
 $ CROSSING        : int  66 70 84 79 78 69 83 45 70 75 ...
 $ FINISHING       : int  66 64 90 77 91 62 82 79 80 80 ...
 $ HEADINGACCURACY: int  66 28 70 61 89 46 66 78 78 68 ...
 $ SHORTPASSING    : int  83 69 91 81 78 67 76 76 78 74 ...
 $ VOLLEYS         : int  75 57 88 73 85 59 75 70 77 75 ...
 $ DRIBBLING       : int  78 86 95 76 81 72 86 85 79 82 ...
 $ CURVE           : int  74 77 93 83 79 68 78 69 76 80 ...
 $ FKACCURACY      : int  69 73 93 82 75 62 76 54 60 74 ...
 $ LONGPASSING     : int  82 59 90 77 70 67 68 61 76 74 ...
 $ BALLCONTROL     : int  80 79 93 79 87 70 84 82 81 83 ...
 $ ACCELERATION    : int  77 74 87 73 76 78 86 89 63 82 ...
 $ SPRINTSPEED     : int  72 68 76 66 82 81 87 85 66 77 ...
 $ AGILITY         : int  76 82 91 73 77 74 81 82 69 87 ...
 $ REACTIONS       : int  79 69 92 75 90 56 81 78 74 78 ...
 $ BALANCE         : int  77 83 95 72 67 56 71 51 66 87 ...
 $ SHOTPOWER       : int  83 76 86 78 93 73 86 75 76 84 ...
 $ JUMPING         : int  70 31 68 64 95 55 75 82 71 69 ...
 $ STAMINA         : int  85 56 70 75 70 65 76 68 66 79 ...
 $ STRENGTH        : int  79 67 68 67 75 59 77 69 74 70 ...
 $ LONGSHOTS       : int  78 68 91 81 88 62 83 63 72 80 ...
 $ AGGRESSION      : int  82 57 44 54 62 65 60 32 66 83 ...
 $ INTERCEPTIONS   : int  75 22 40 61 29 42 39 38 60 46 ...
 $ POSITIONING     : int  79 69 93 77 93 67 81 82 80 82 ...
 $ VISION          : int  84 68 94 78 76 69 79 65 77 78 ...
 $ PENALTIES       : int  70 61 75 80 90 51 68 62 80 73 ...
 $ COMPOSURE       : int  80 75 96 76 94 58 78 79 78 82 ...
 $ MARKING         : int  76 19 20 68 24 43 43 32 63 60 ...
 $ STANDINGTACKLE  : int  75 23 35 67 32 48 38 32 65 53 ...
 $ SLIDINGTACKLE   : int  74 15 24 60 24 43 28 31 62 46 ...
 $ GKDIVING        : int  8 7 6 7 7 13 15 7 7 6 ...
 $ GKHANDLING      : int  6 14 11 10 11 7 11 14 6 10 ...
 $ GKKICKING       : int  7 6 15 14 15 14 7 15 9 8 ...
 $ GKPOSITIONING   : int  8 8 14 15 14 8 9 8 11 15 ...
 $ GKREFLEXES      : int  10 8 8 10 11 11 5 5 8 6 ...
 $ W_F             : int  4 5 4 3 4 3 4 3 3 4 ...
 $ SM              : int  3 5 4 3 5 4 4 4 4 4 ...
 $ IR              : int  1 1 5 1 5 1 2 1 3 1 ...
 $ PAC             : int  74 71 81 69 79 80 87 87 65 79 ...
 $ SHO             : int  73 67 89 78 91 64 82 74 77 80 ...
 $ PAS             : int  79 68 90 80 76 68 77 64 75 75 ...
 $ DRI             : int  79 82 94 76 83 69 84 81 78 83 ...
 $ DEF             : int  74 21 34 65 34 45 42 38 64 55 ...
 $ PHY             : int  81 60 64 66 72 62 73 62 70 75 ...
```

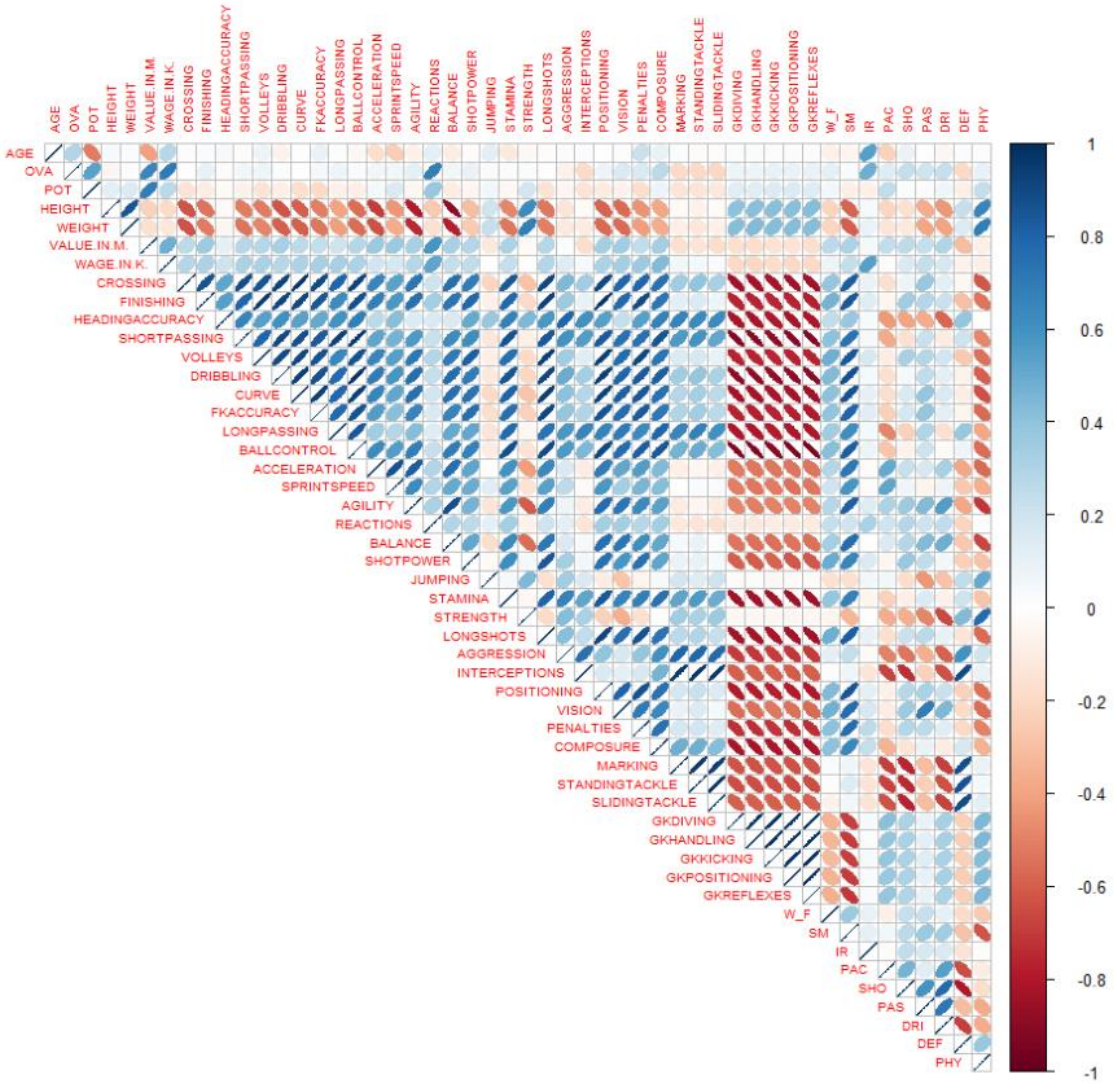
The str function gives us the datatypes in the dataset. There are 50 continuous variables and total of (120+120+120+326= 686) observations in FIFA series (FIFA20,21,22,23).

```
summary(data1)
```

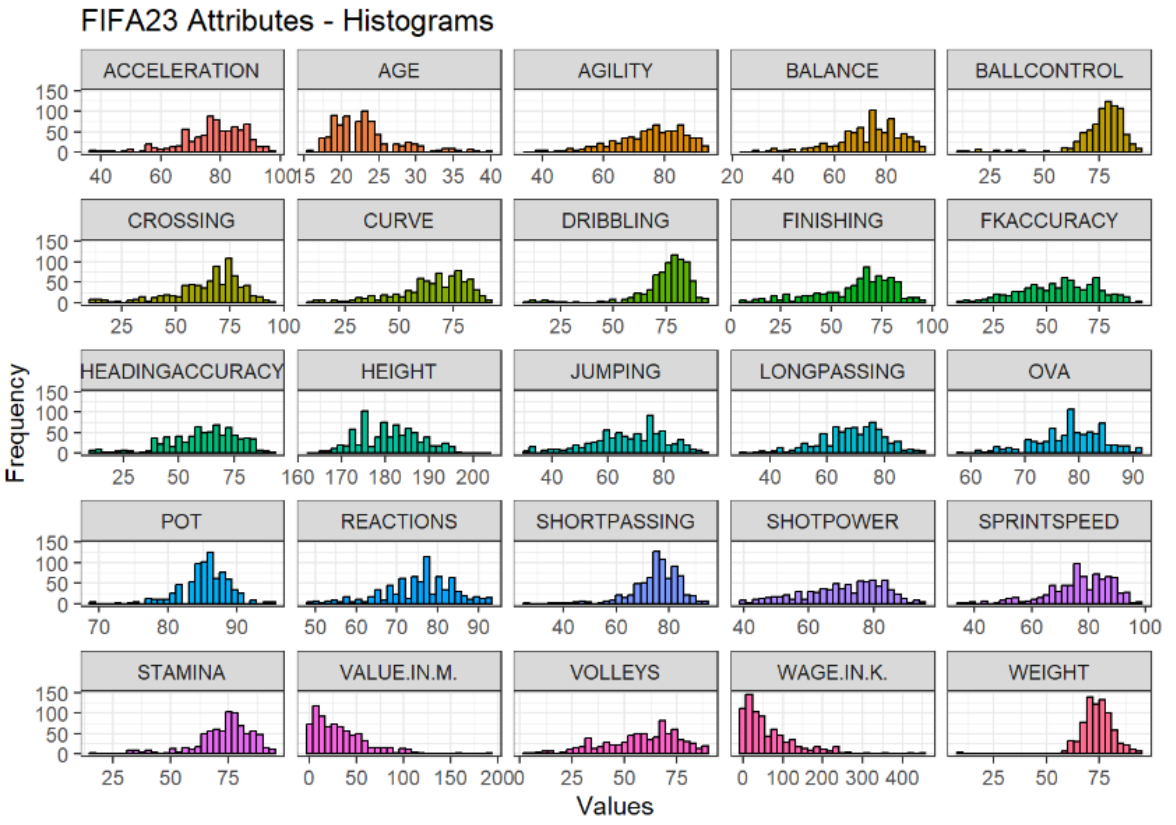
```
##       AGE             OVA             POT            HEIGHT        WEIGHT
##  Min.   :16.00   Min.   :58.00   Min.   :69.00   Min.   :163   Min.   :10.00
##  1st Qu.:20.00   1st Qu.:75.00   1st Qu.:84.00   1st Qu.:175   1st Qu.:70.00
##  Median :22.00   Median :79.00   Median :86.00   Median :181   Median :74.00
##  Mean   :23.03   Mean   :78.42   Mean   :85.42   Mean   :181   Mean   :73.66
##  3rd Qu.:25.00   3rd Qu.:83.00   3rd Qu.:88.00   3rd Qu.:185   3rd Qu.:78.00
##  Max.   :40.00   Max.   :91.00   Max.   :95.00   Max.   :204   Max.   :95.00
##   VALUE.IN.M.       WAGE.IN.K.        CROSSING       FINISHING
##  Min.   :  0.00   Min.   :  0.00   Min.   :11.00   Min.   : 5.00
##  1st Qu.: 10.00   1st Qu.: 16.00   1st Qu.:57.00   1st Qu.:52.00
##  Median : 26.50   Median : 41.00   Median :69.00   Median :67.00
##  Mean   : 32.47   Mean   : 64.98   Mean   :64.97   Mean   :62.41
##  3rd Qu.: 47.50   3rd Qu.: 90.00   3rd Qu.:75.00   3rd Qu.:76.00
##  Max.   :190.50   Max.   :450.00   Max.   :94.00   Max.   :94.00
##  HEADINGACCURACY  SHORTPASSING      VOLLEYS        DRIBBLING
##  Min.   :10.00   Min.   :25.00   Min.   : 4.00   Min.   :10.00
##  1st Qu.:51.00   1st Qu.:71.00   1st Qu.:46.00   1st Qu.:71.00
##  Median :63.00   Median :76.00   Median :61.00   Median :78.00
##  Mean   :61.15   Mean   :75.12   Mean   :58.25   Mean   :75.05
##  3rd Qu.:73.00   3rd Qu.:81.00   3rd Qu.:72.00   3rd Qu.:83.00
##  Max.   :91.00   Max.   :93.00   Max.   :90.00   Max.   :95.00
##      CURVE          FKACCURACY     LONGPASSING     BALLCONTROL
##  Min.   :10.00   Min.   : 8.00   Min.   :30.00   Min.   :12.00
##  1st Qu.:58.00   1st Qu.:43.00   1st Qu.:63.00   1st Qu.:73.00
##  Median :69.00   Median :58.00   Median :70.00   Median :79.00
##  Mean   :65.37   Mean   :55.75   Mean   :68.95   Mean   :76.37
##  3rd Qu.:77.00   3rd Qu.:69.00   3rd Qu.:77.00   3rd Qu.:83.00
##  Max.   :93.00   Max.   :93.00   Max.   :93.00   Max.   :94.00
##   ACCELERATION    SPRINTSPEED      AGILITY        REACTIONS       BALANCE
##  Min.   :37.00   Min.   :34.0   Min.   :35.00   Min.   :49.0   Min.   :25.00
##  1st Qu.:72.00   1st Qu.:72.0   1st Qu.:69.00   1st Qu.:71.0   1st Qu.:66.25
##  Median :78.00   Median :78.0   Median :78.00   Median :77.0   Median :75.00
##  Mean   :77.62   Mean   :77.3   Mean   :75.74   Mean   :75.6   Mean   :73.12
##  3rd Qu.:85.00   3rd Qu.:85.0   3rd Qu.:85.00   3rd Qu.:82.0   3rd Qu.:82.00
```

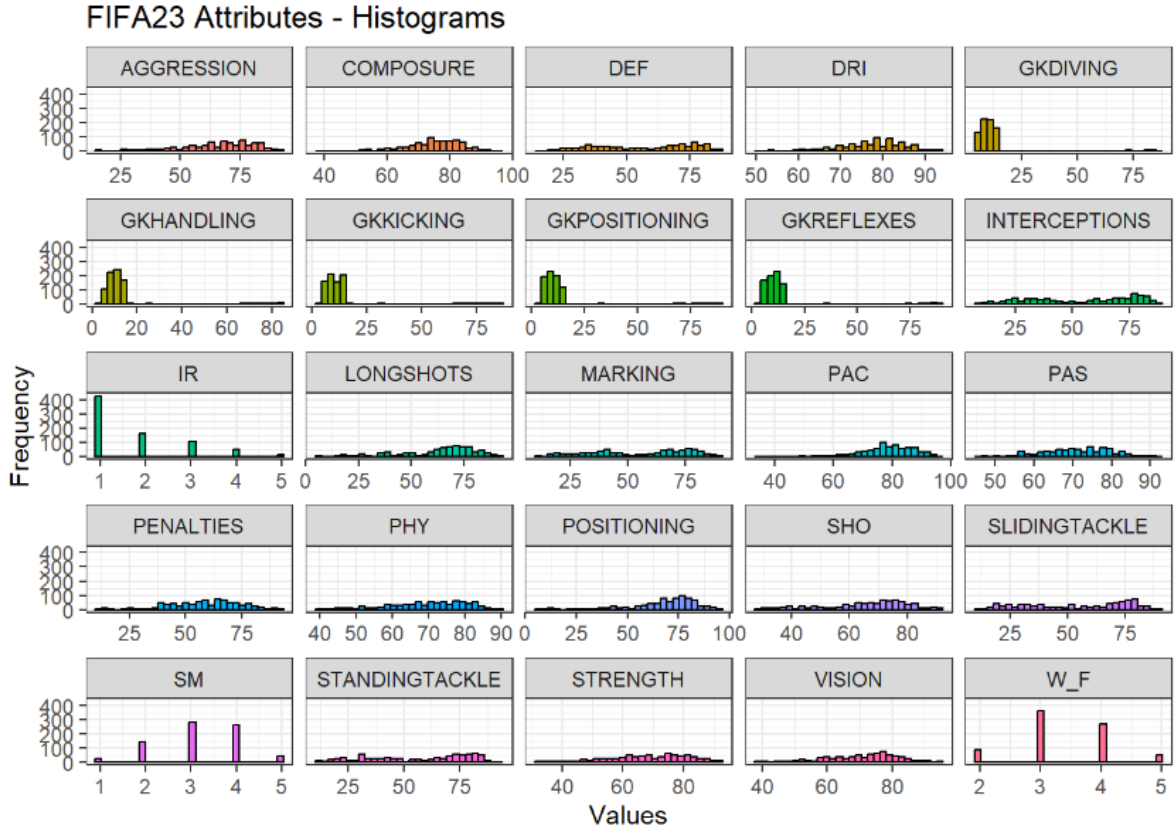
```
##    SHOTPOWER        JUMPING         STAMINA         STRENGTH       LONGSHOTS
##  Min.   :39.00   Min.   :31.0   Min.   :16.00   Min.   :31.00   Min.   : 5.00
##  1st Qu.:63.00   1st Qu.:59.0   1st Qu.:68.00   1st Qu.:60.00   1st Qu.:51.00
##  Median :73.00   Median :69.0   Median :75.00   Median :69.00   Median :67.00
##  Mean   :70.75   Mean   :67.6   Mean   :73.61   Mean   :68.38   Mean   :62.65
##  3rd Qu.:80.00   3rd Qu.:77.0   3rd Qu.:81.00   3rd Qu.:79.00   3rd Qu.:75.00
##  Max.   :94.00   Max.   :95.0   Max.   :94.00   Max.   :93.00   Max.   :91.00
##    AGGRESSION    INTERCEPTIONS    POSITIONING        VISION
##  Min.   :16.00   Min.   : 9.00   Min.   : 5.00   Min.   :38.00
##  1st Qu.:57.00   1st Qu.:34.00   1st Qu.:61.00   1st Qu.:63.00
##  Median :69.00   Median :61.00   Median :72.00   Median :73.00
##  Mean   :65.83   Mean   :55.34   Mean   :67.74   Mean   :70.94
##  3rd Qu.:78.00   3rd Qu.:76.00   3rd Qu.:79.00   3rd Qu.:79.00
##  Max.   :92.00   Max.   :89.00   Max.   :94.00   Max.   :94.00
##    PENALTIES       COMPOSURE        MARKING      STANDINGTACKLE
##  Min.   :11.00   Min.   :38.00   Min.   :11.00   Min.   :10.00
##  1st Qu.:48.25   1st Qu.:70.00   1st Qu.:38.00   1st Qu.:36.00
##  Median :60.00   Median :76.00   Median :62.00   Median :64.00
##  Mean   :58.91   Mean   :75.16   Mean   :55.48   Mean   :57.17
##  3rd Qu.:69.75   3rd Qu.:82.00   3rd Qu.:74.00   3rd Qu.:77.00
##  Max.   :92.00   Max.   :96.00   Max.   :90.00   Max.   :92.00
##  SLIDINGTACKLE      GKDIVING       GKHANDLING      GKKICKING    GKPOSITIONING
##  Min.   :11.00   Min.   : 5.00   Min.   : 4.00   Min.   : 2.0   Min.   : 4.00
##  1st Qu.:32.00   1st Qu.: 7.00   1st Qu.: 8.00   1st Qu.: 8.0   1st Qu.: 7.00
##  Median :58.00   Median :10.00   Median :10.00   Median :11.0   Median :10.00
##  Mean   :53.61   Mean   :12.04   Mean   :12.14   Mean   :12.4   Mean   :12.17
##  3rd Qu.:74.00   3rd Qu.:12.75   3rd Qu.:13.00   3rd Qu.:13.0   3rd Qu.:13.00
##  Max.   :90.00   Max.   :86.00   Max.   :85.00   Max.   :85.0   Max.   :90.00
##    GKREFLEXES        W_F              SM              IR             PAC
##  Min.   : 3.0   Min.   :2.000   Min.   :1.000   Min.   :1.000   Min.   :35.00
##  1st Qu.: 8.0   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:74.00
##  Median :11.0   Median :3.000   Median :3.000   Median :1.000   Median :79.00
##  Mean   :12.8   Mean   :3.362   Mean   :3.213   Mean   :1.765   Mean   :78.46
##  3rd Qu.:13.0   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:85.00
##  Max.   :89.0   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :97.00
##       SHO             PAS             DRI             DEF             PHY
##  Min.   :28.00   Min.   :46.00   Min.   :51.00   Min.   :16.00   Min.   :40.0
##  1st Qu.:59.00   1st Qu.:65.00   1st Qu.:72.00   1st Qu.:40.00   1st Qu.:63.0
```

The summary function gives us the statistics of the data. It gives us the minimum, 1st qu., median, mean, 3rd qu., maximum value of each column. This tells us how the data is distributed.



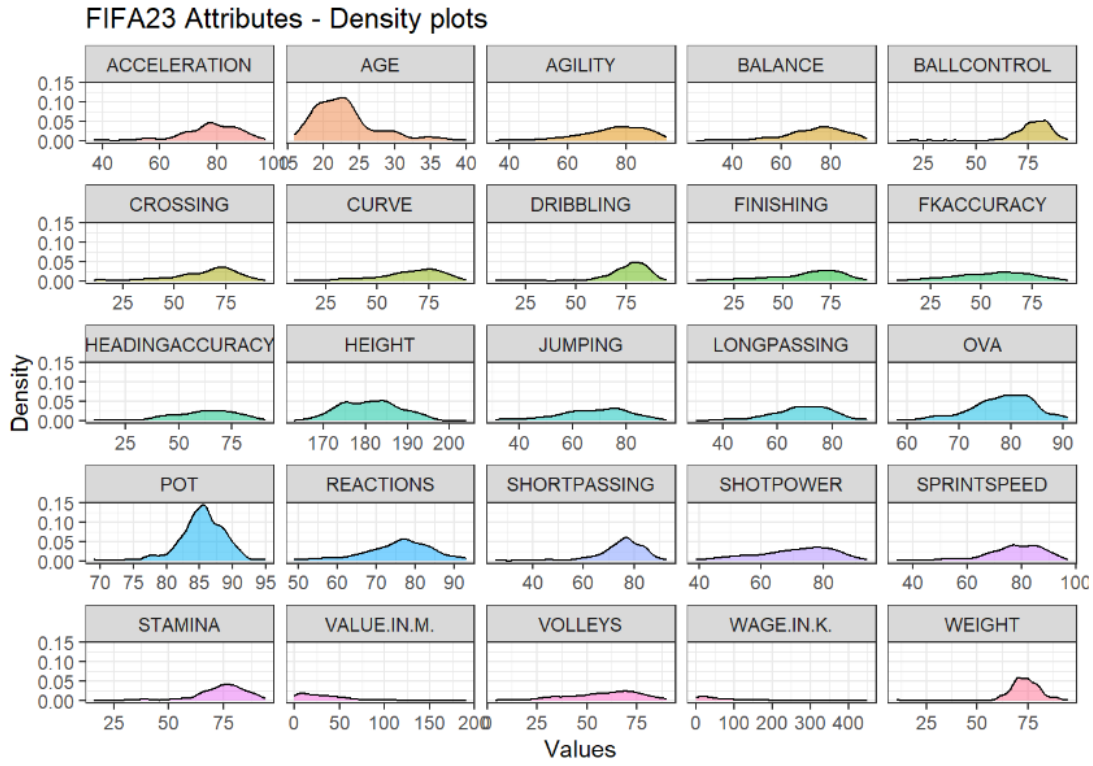
From the above heatmap, there are many variables which are strongly positively and negatively correlated. Height and weight are strongly negatively correlated with balance of the player. If one of it is affected then the player's balance in the game will be affected. Likewise overall rating of the player is strongly positively correlated with the reactions that the player gets. Thus, the color palate on the above figure gives the information about the correlation between all the continuous variables.

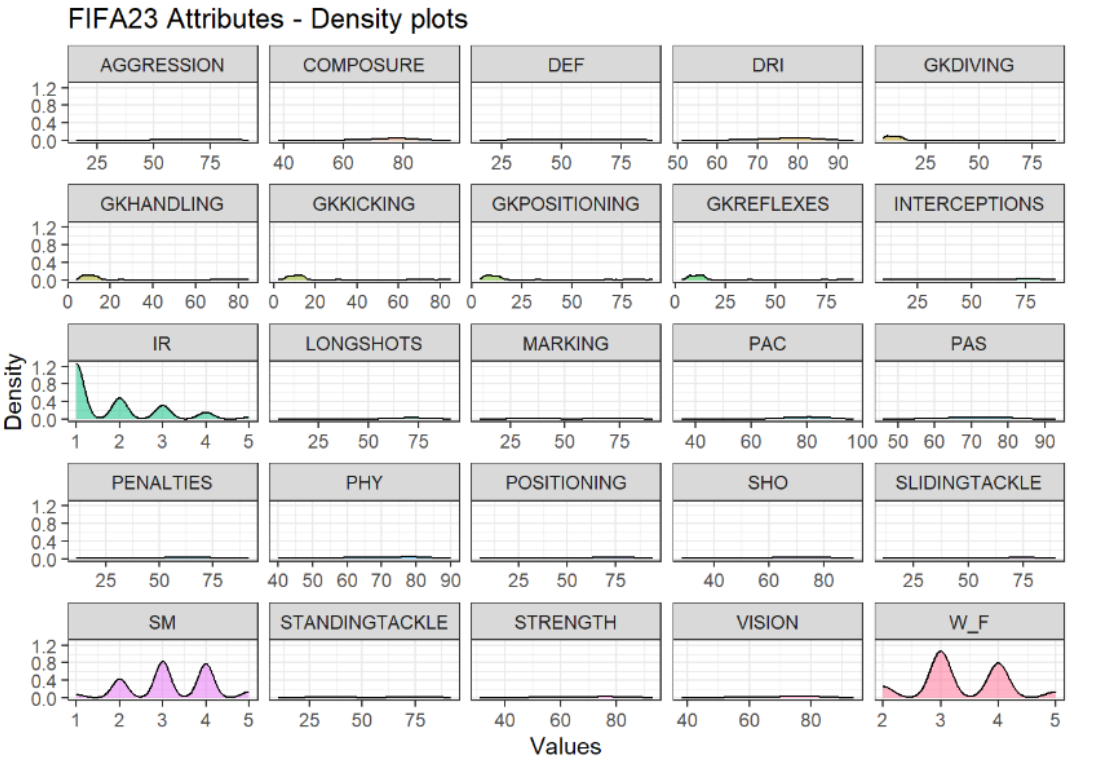




It is observed from the above histograms, we can tell that the most of the variable's data is skewed. Most of the players in this series has a weak foot. Ball control is high which

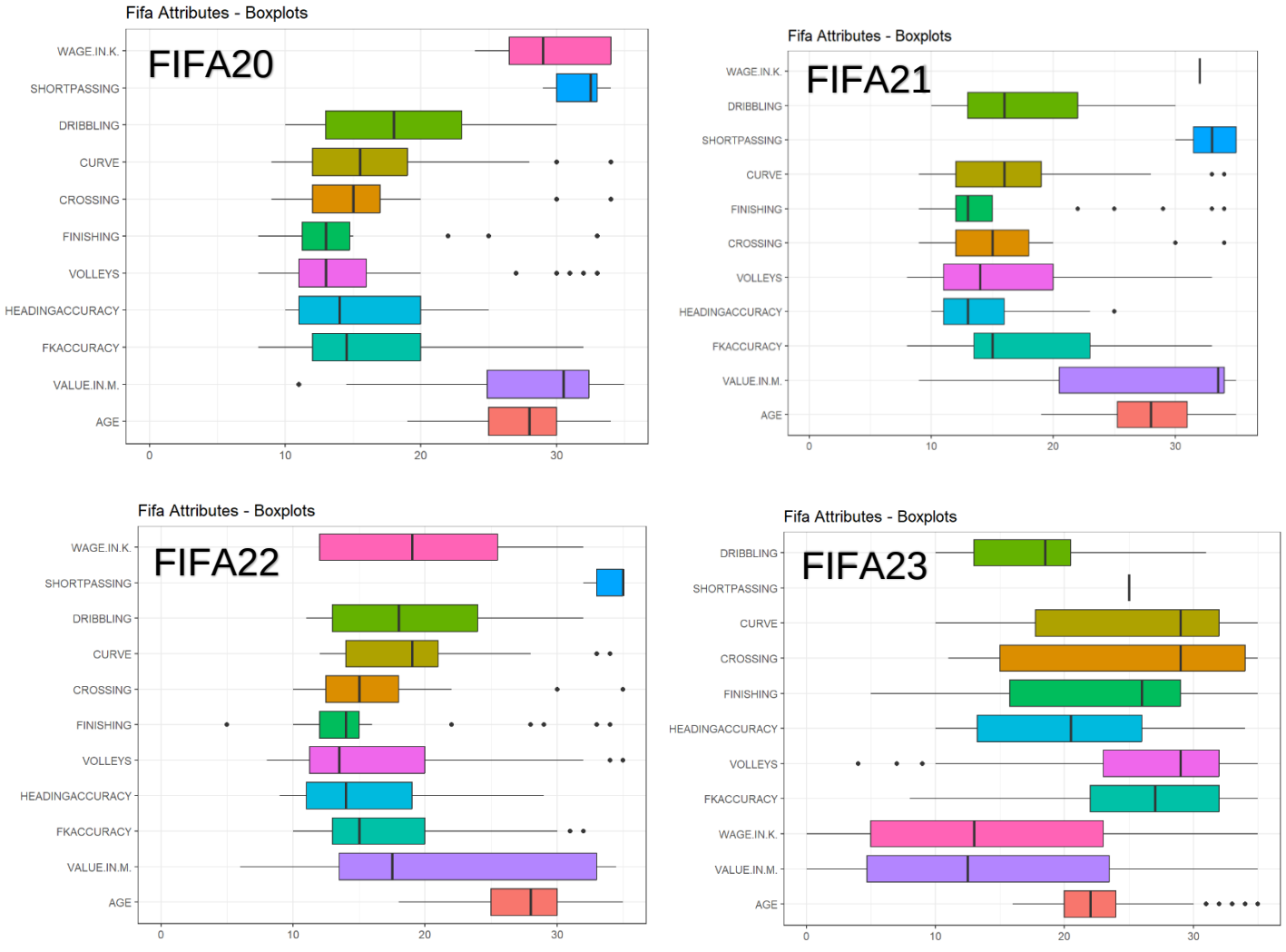
indicates that players ability to control is high which is good enough in this series. Many players are in the age 20-25 years that means there are quite young players who have got a higher accuracy of playing the game. Weight of the players tells us that most of the players are well built which helps to have a higher strength in the game. The players are well paid and most of the players are paid a wage between €20K- €50K per week.





By plotting the density plots, we get to know from the above figures that variables such as wage in K, aggression, composure, DEF, DRI, GKDiving, GKReflexes, Interceptions, Longshots, Marking, PAC, PAS, Strength, Vision, Slidingtackle and Penalties have an even distribution over a range of values. This in turn implies that the density of the values is constant across the entire range.

Comparison of Box Plots over different series of FIFA



By comparing the boxplots in the above figure, we observe that the distribution of value of the player has increased over the different series of FIFA. We can also observe that the wage of the players, or the median wage which they are paid per week has decreased over different series. In FIFA23 we can see that there are many young players coming up for the matches. There is a presence in the outliers in few of the variables over the series, which implies there is unusual or unexpected values in the data.

3. DATA PREPARATION

a. Data integration

The data is obtained from a single source https://sofifa.com/. And a web scraping tool Octoparse was used in order to scrape the required data. While scraping since it's a free version some of the data points were missing. The missing data points were again manually collected by checking the website and filling those points in order to assure data quality.

b. Data cleaning

Initially when the data was scraped, the data looked something like the picture shown below:

HEIGHT	WEIGHT	VALUE	WAGE
170cm	72kg	€95.5M	€560K
187cm	83kg	€58.5M	€410K
175cm	68kg	€105.5M	€290K
193cm	92kg	€90M	€240K
188cm	87kg	€77.5M	€125K
181cm	70kg	€90M	€370K
184cm	80kg	€86M	€300K
175cm	74kg	€90M	€470K
191cm	91kg	€64.5M	€160K
175cm	71kg	€80.5M	€240K
175cm	69kg	€80.5M	€240K
187cm	85kg	€67.5M	€250K
173cm	70kg	€69M	€310K
178cm	73kg	€93.5M	€150K
168cm	70kg	€59.5M	€230K
188cm	89kg	€83M	€220K
176cm	73kg	€69M	€370K

The following formulas were used in excel to extract only the numeric values from the data fields:

=LEFT(A2,3) (for height)

=LEFT(B2,2) (for weight)

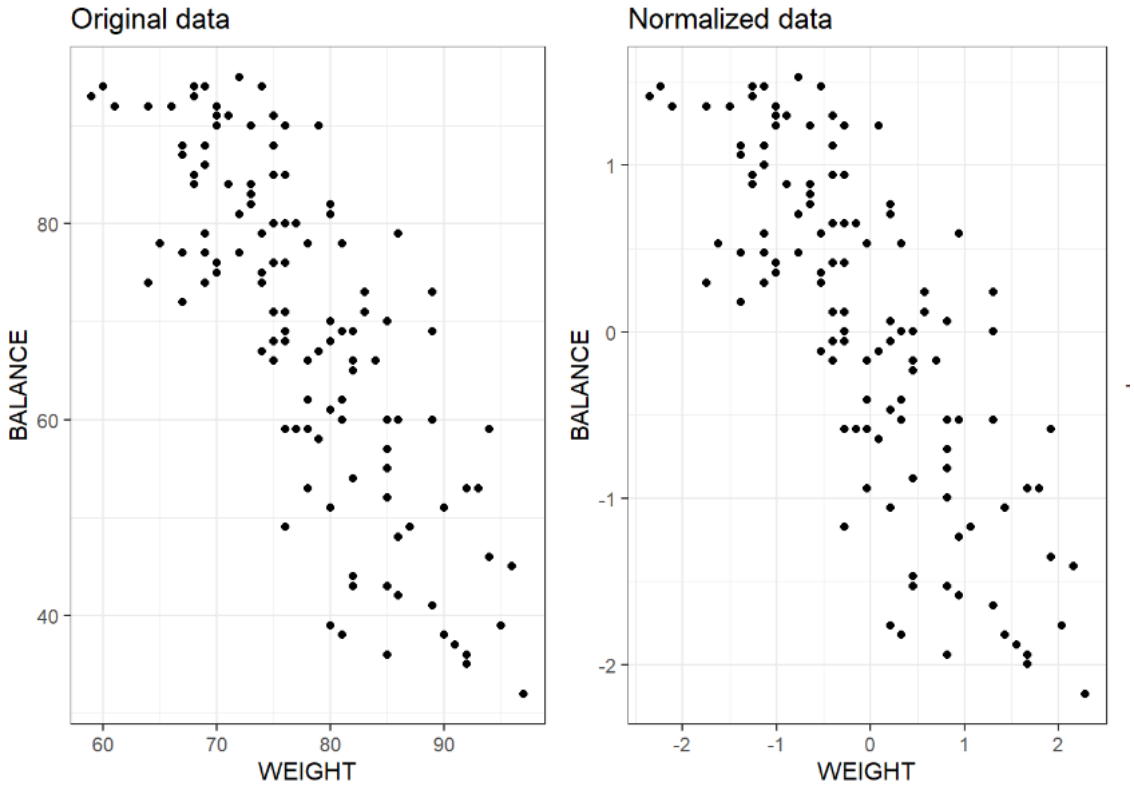
=FIND("M",C2) (for value and wage)

=LEFT(C2,G2-1)

=RIGHT(H2,LEN(H2)-SEARCH("€",H2))

Variable standardization

Standardization is a technique that is often used as a preprocessing step for clustering algorithms. It involves transforming the variables in a dataset so that they have a mean of zero and a standard deviation of one. This can be helpful for clustering algorithms because it ensures that all variables are on the same scale, which can make it easier for the algorithm to identify patterns and group similar observations together. So, **normalization** is done by scaling all the numeric variables.

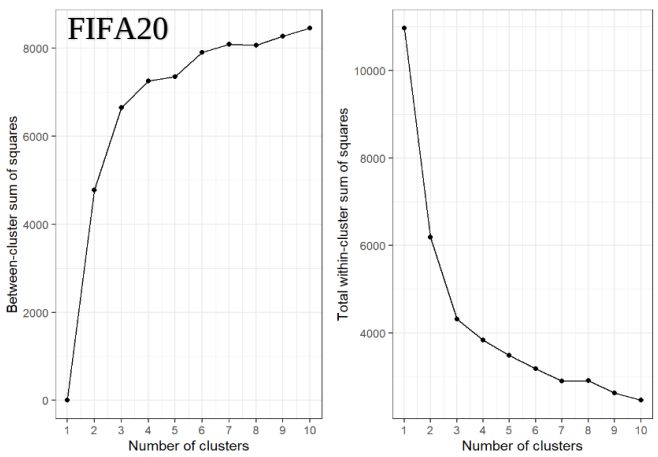


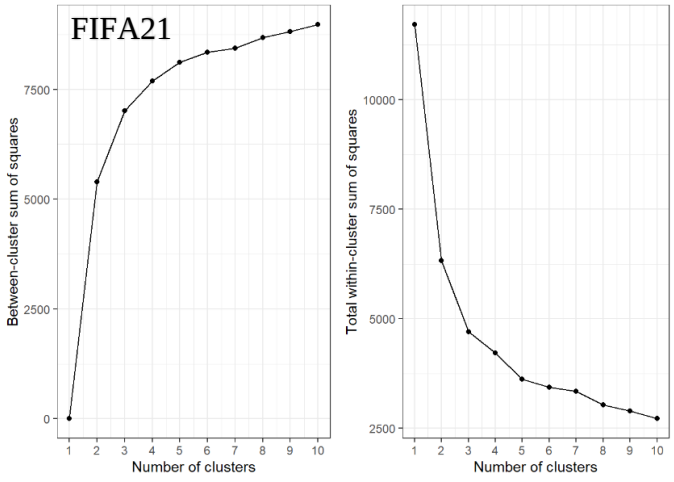
5. CLUSTERING

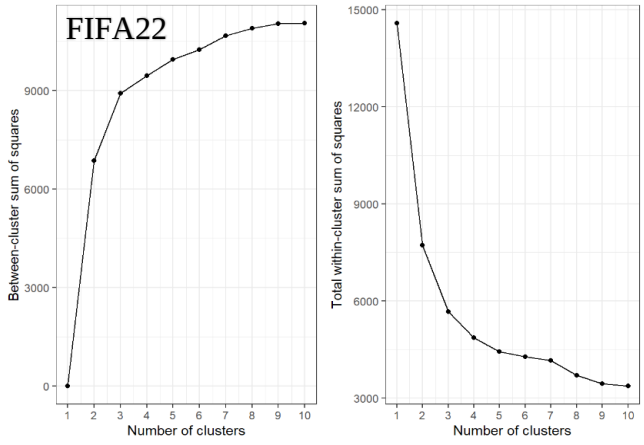
K-Means

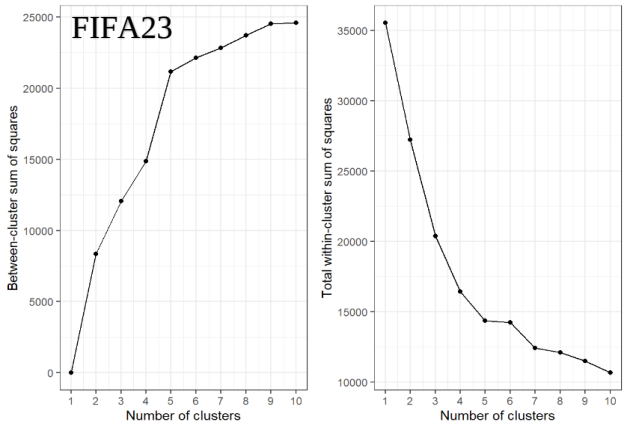
There are several assumptions that are often made when using k-means clustering:

- **The data is linear:** K-means works best when the data is linearly separable, meaning that the clusters can be separated by a straight line.
- **The data is standardized:** K-means is sensitive to the scale of the variables, so it is often recommended to standardize the data before applying k-means.
- **The data is homoscedastic:** This means that the variance of the data is constant across all observations. If the variance is not constant, the results of k-means may be distorted.
- **The data is independent:** K-means assumes that the observations in the data are independent of each other, meaning that the value of one observation does not affect the value of another observation.
- **The number of clusters is known:** K-means requires the user to specify the number of clusters in advance, and this assumption is often made in order to apply the algorithm.





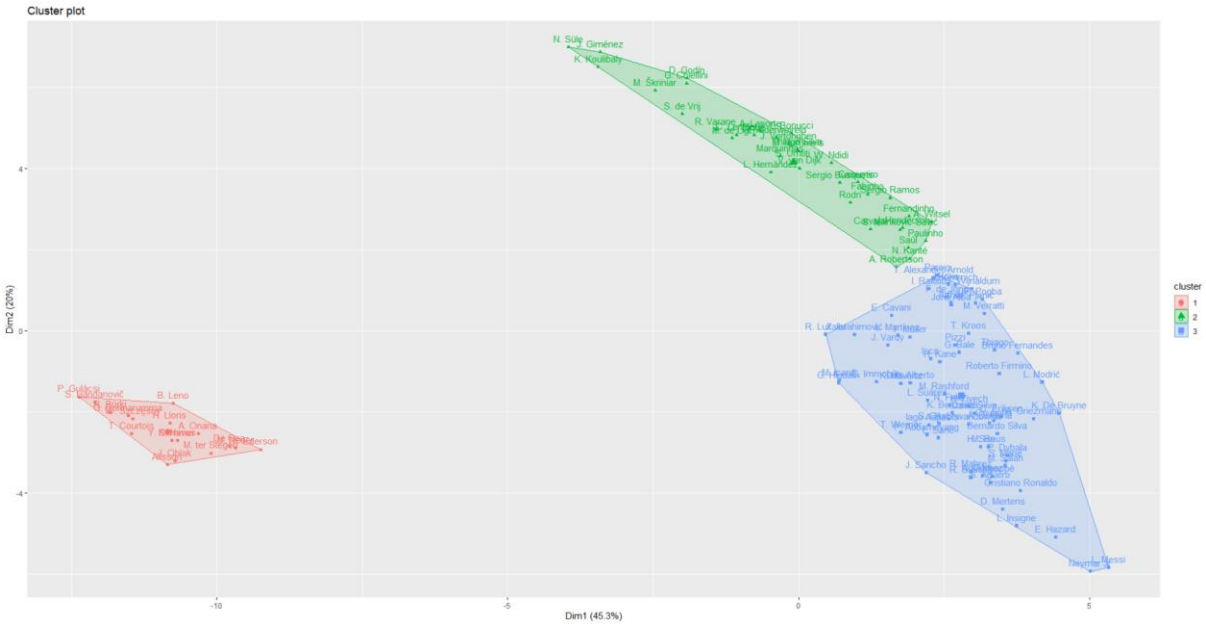




The above picture shows the optimal number of clusters for each series by considering Elbow method. The optimum number of clusters for FIFA23 changes to 4 and rest of them remains 3. The clusters can be visualized as follows:

FIFA20

```
K-means clustering with 3 clusters of sizes 17, 36, 67
```



FIFA21

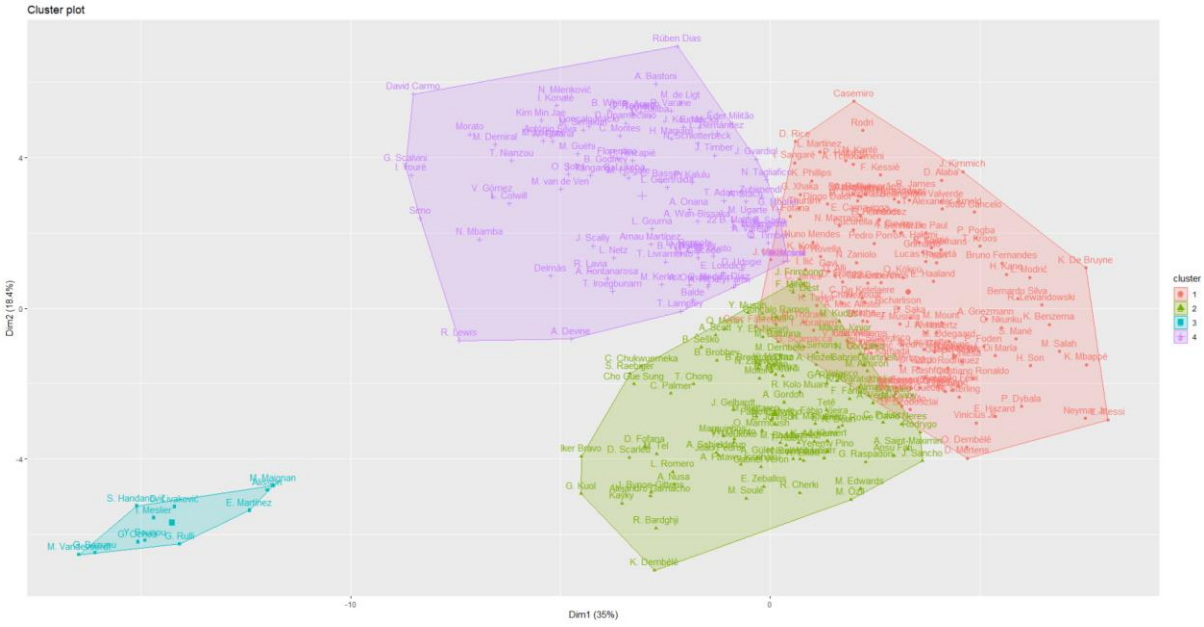
```
K-means clustering with 3 clusters of sizes 72, 32, 16
```

FIFA22

```
K-means clustering with 3 clusters of sizes 77, 18, 25
```

FIFA23

```
K-means clustering with 4 clusters of sizes 129, 95, 11, 91
```



From the above clusters we can tell how the clusters are changing in different FIFA series. The clusters are also highlighted to show how many players are in a particular cluster.

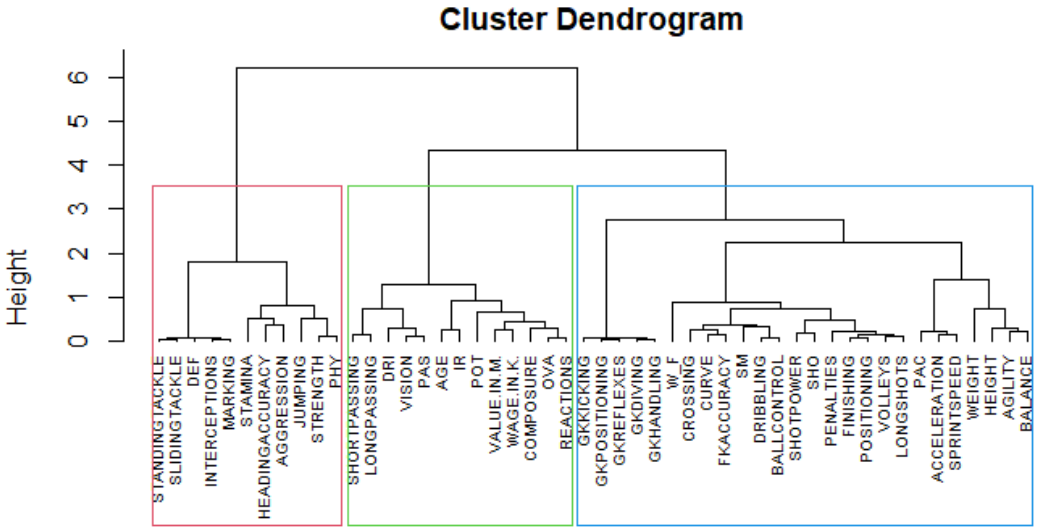
Hierarchical Clustering

There are several assumptions that are often made when using hierarchical clustering:

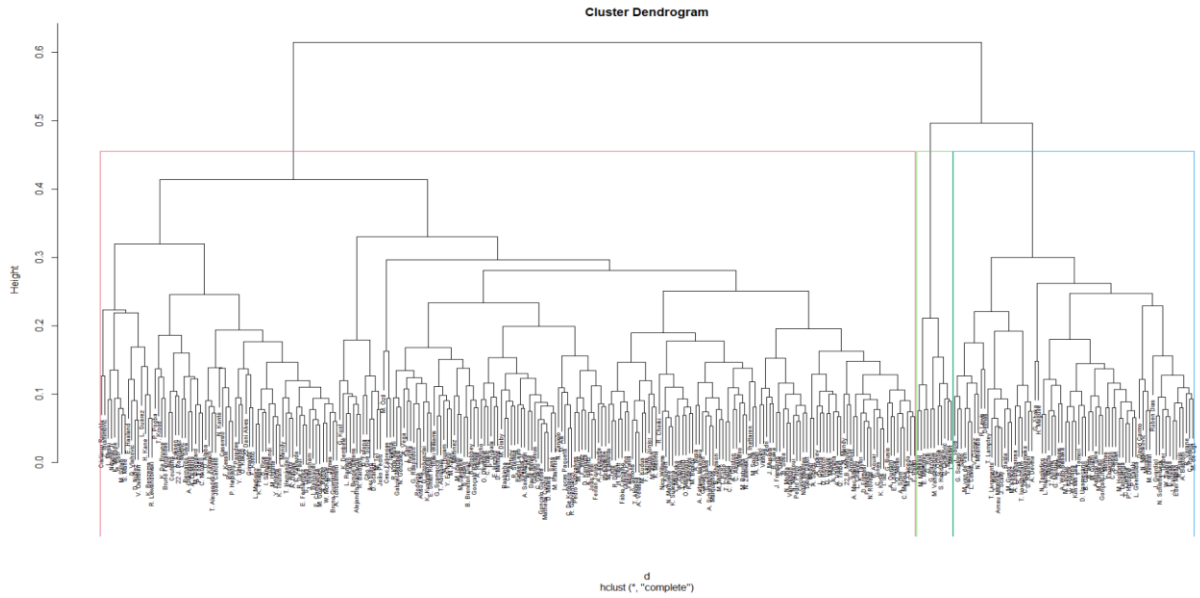
- **The data is linear:** Hierarchical clustering works best when the data is linearly separable, meaning that the clusters can be separated by a straight line.
- **The data is standardized:** Hierarchical clustering is sensitive to the scale of the variables, so it is often recommended to standardize the data before applying the algorithm.
- **The data is homoscedastic:** This means that the variance of the data is constant across all observations. If the variance is not constant, the results of hierarchical clustering may be distorted.
- **The data is independent:** Hierarchical clustering assumes that the observations in the data are independent of each other, meaning that the value of one observation does not affect the value of another observation.
- **The distance measure is appropriate:** Hierarchical clustering relies on a distance measure to calculate the similarity between observations, and it is important to select an appropriate measure for the data.

FIFA20

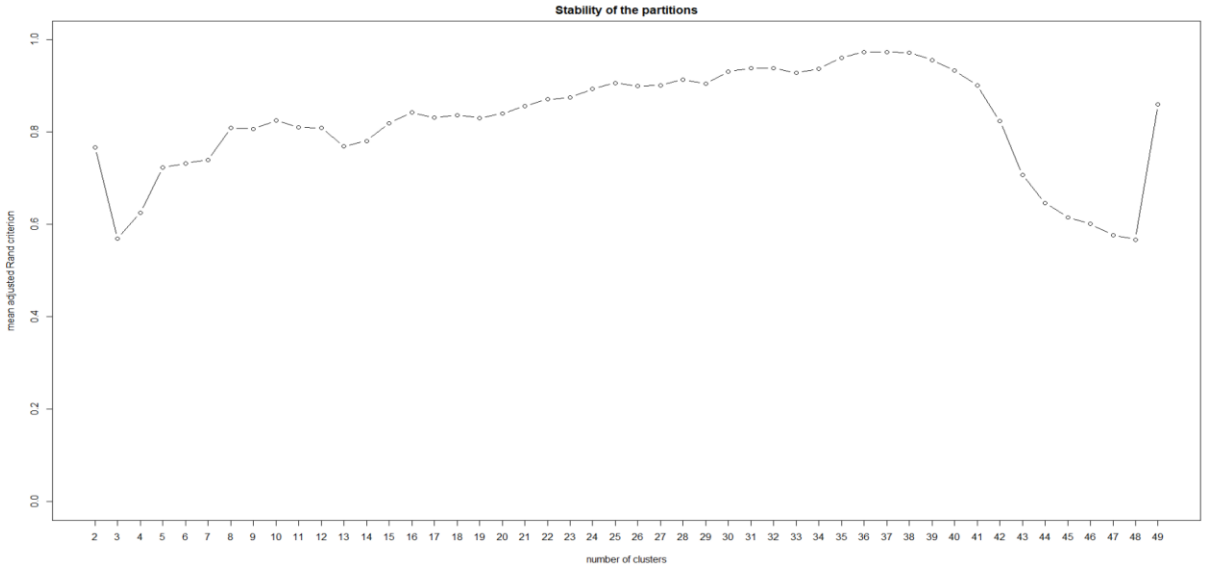
This is how the cluster dendrogram looks. There has to be optimal k value from the plots above. And we use them to separate this cluster as three. The cluster 1 contributes to the High Physical Strength. In this cluster we observe that the players are usually having good stamina, aggressive and defensive power.



The cluster 2 here contributes to the Overall reactions and Wage. This cluster tells us more about players age, vision, value, wage and International reputation the player has. This can be possible by the attributes long passing and shortpassing, this passing the ball towards their teammates has a high correlation with the overall rating of the player. The cluster 3 contributes to the Ability of Goalkeeping and total power. This cluster tells us about the various attributes that are related to Goalkeeping such as positioning, reflexes, diving,etc.. The height and weight determine the Balance that the player has, which inturn will contribute to the Total power.



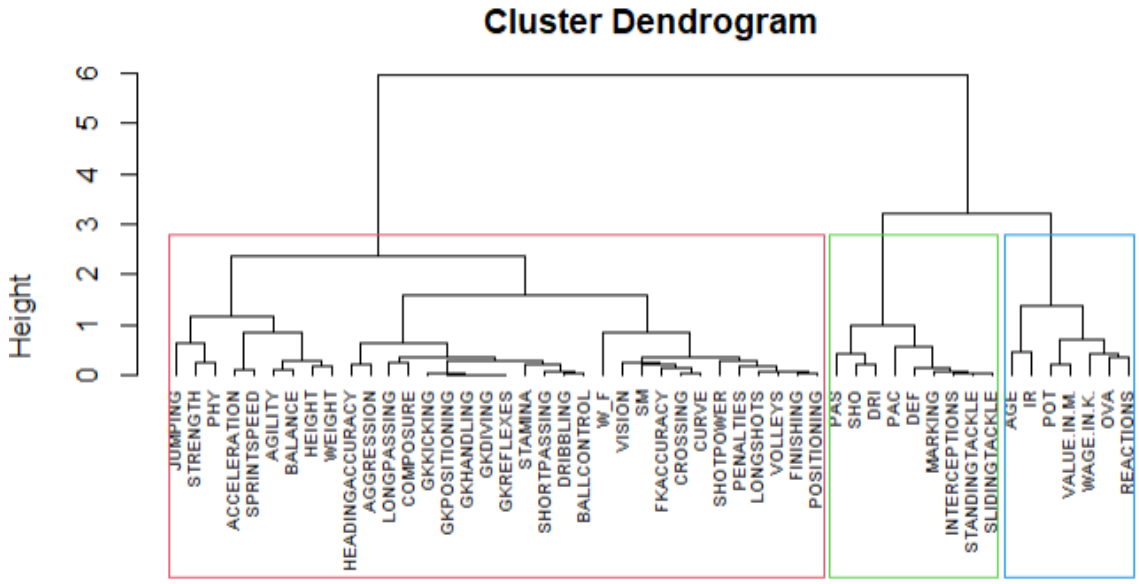
These are the players which fall into that particular cluster.



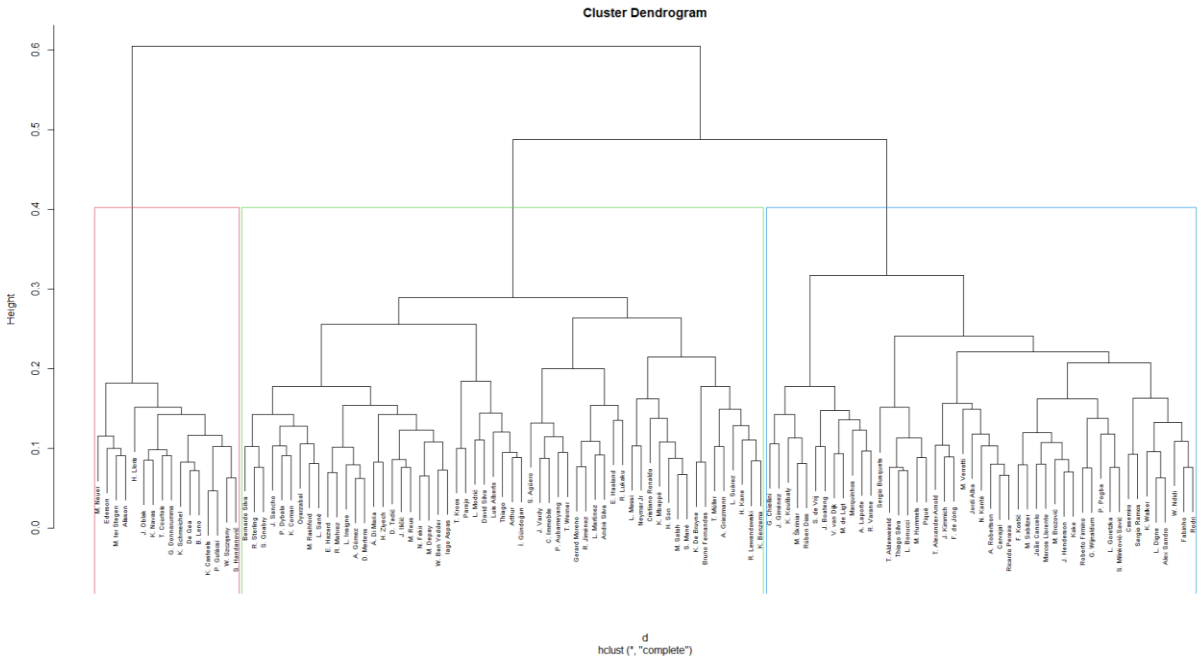
Stability of partitions in clustering refers to the consistency or reproducibility of the cluster assignments of the data points when the clustering algorithm is applied multiple times. If the cluster assignments are consistent across different runs of the algorithm, the partition is considered to be stable. This graph looks consistent.

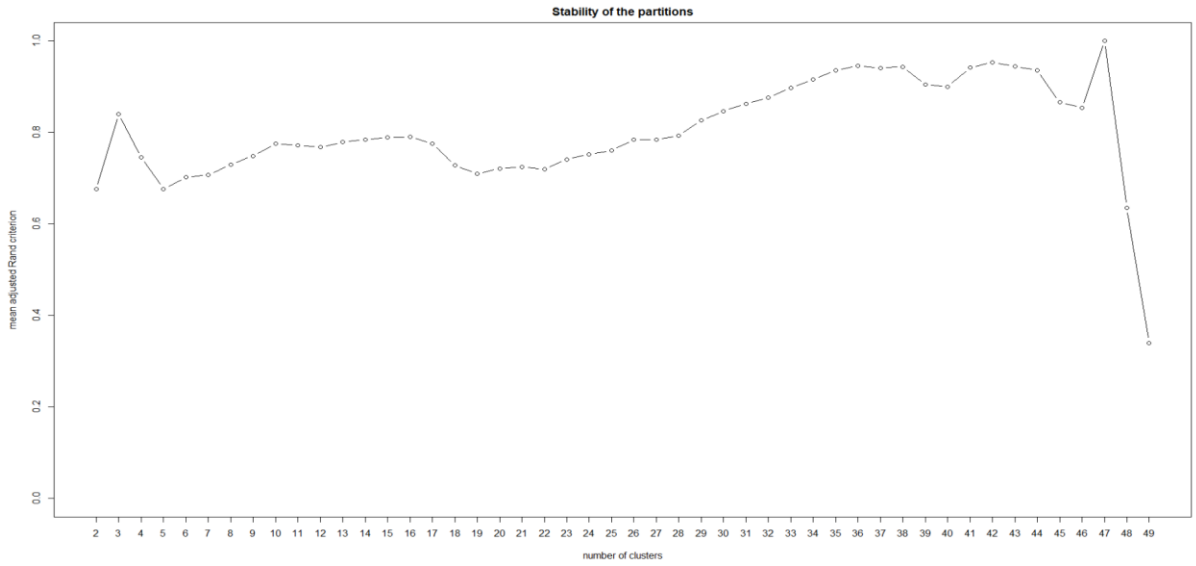
FIFA21

Similarly for FIFA21 series, the cluster dendrogram looks something as shown below. There are three clusters selected by the optimum k value. Cluster 1 here corresponds to the Physical strength and goal keeping ability. Cluster 2 corresponds to the Shooting and Passing Ability. Cluster 3 corresponds to the Overall Rating and Potential. The attributes for each cluster s are clearly combined in a single cluster to understand the cluster well.



These are the players which fall into that particular cluster.

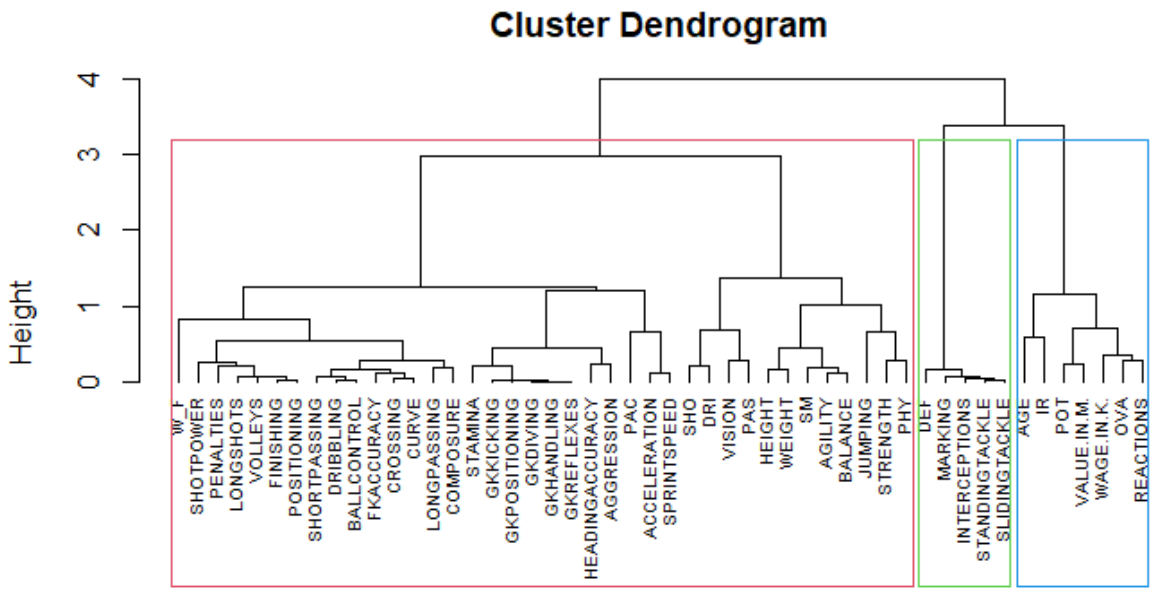


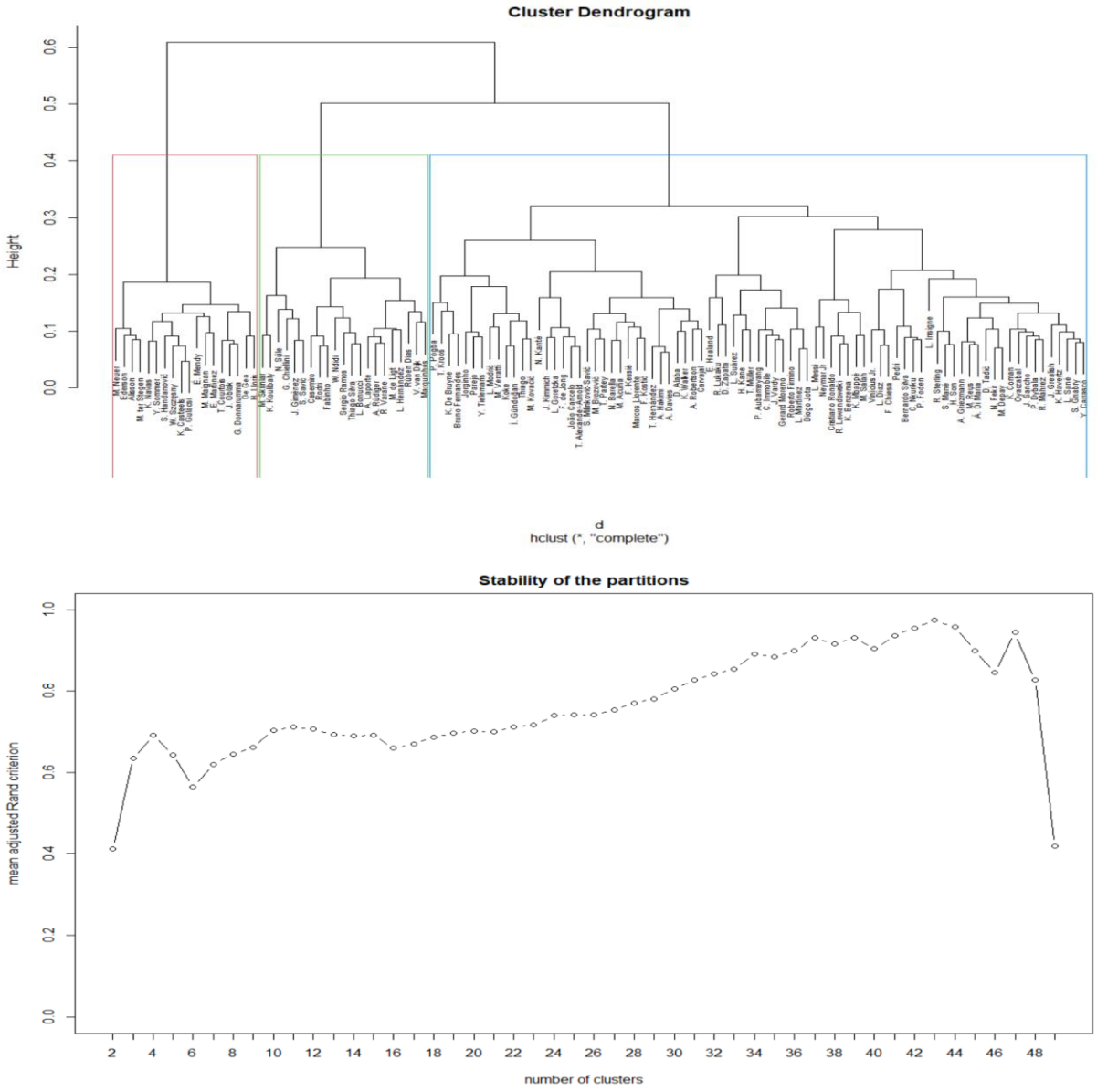


Stability of partitions in clustering refers to the consistency or reproducibility of the cluster assignments of the data points when the clustering algorithm is applied multiple times. If the cluster assignments are consistent across different runs of the algorithm, the partition is considered to be stable. This graph looks consistent.

FIFA22

Similarly, for FIFA22 series, the dendrogram looks same like FIFA21.

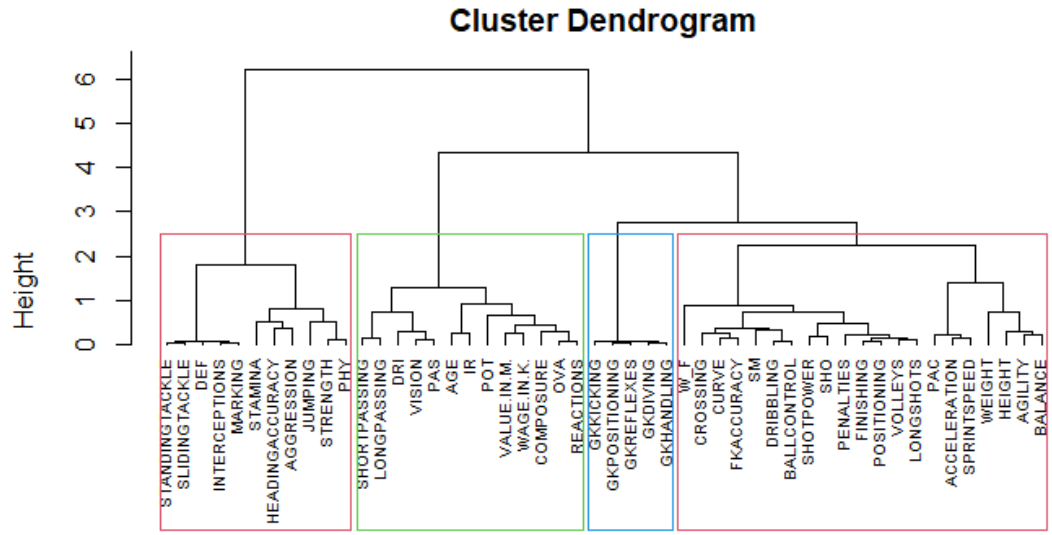


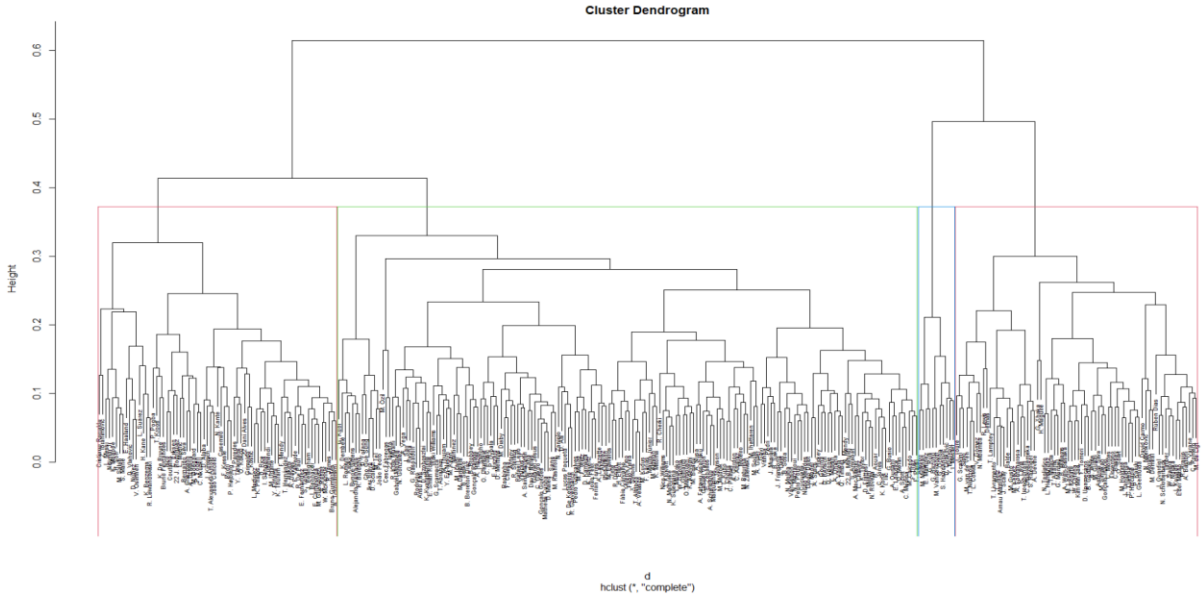


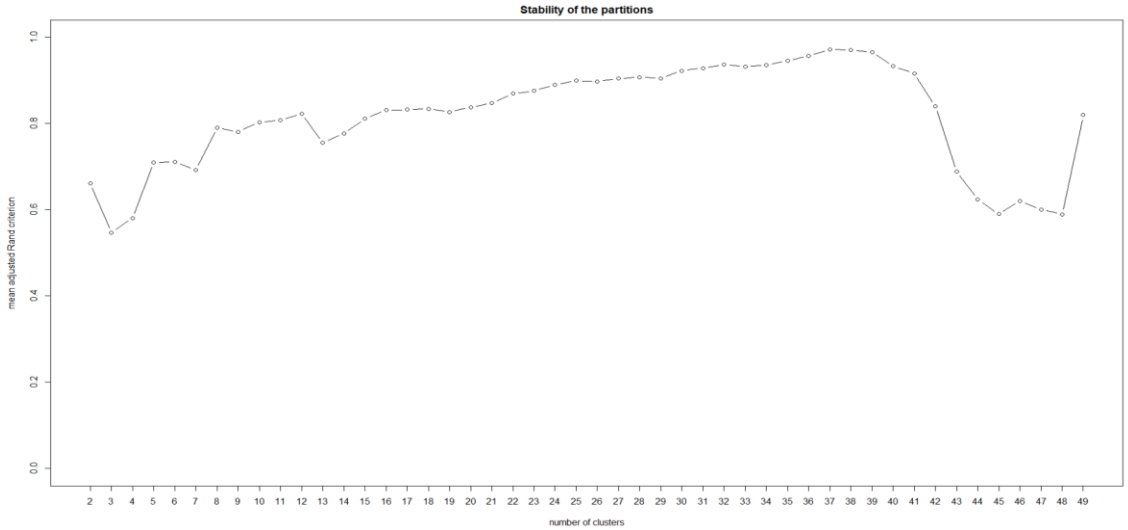
Stability of partitions in clustering refers to the consistency or reproducibility of the cluster assignments of the data points when the clustering algorithm is applied multiple times. If the cluster assignments are consistent across different runs of the algorithm, the partition is considered to be stable. This graph looks consistent.

FIFA23

For FIFA23, the optimum number of clusters selected are 4. So, the cluster dendrogram looks like the picture shown below. Cluster 1 corresponds to the Physical Strength. Cluster 2 corresponds to the Overall Rating and Wage. Cluster 3 corresponds to the Goalkeeping Ability. Cluster 4 corresponds to the Mentality and total power. The attributes for each cluster are clearly combined in a single cluster to understand the cluster well.







If the cluster assignments are consistent across different runs of the algorithm, the partition is considered to be stable. This graph looks somewhat consistent.

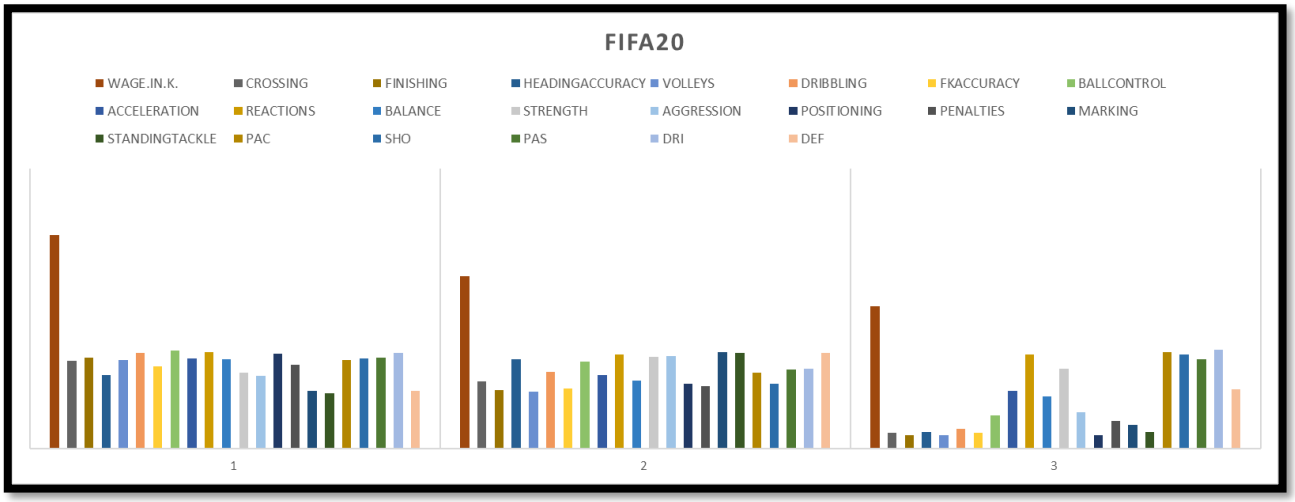
5. BUSINESS PROFILING

The business profiling for each series of FIFA would look something like this:

FIFA20

Cluster	AGE	OVA	POT	HEIGHT	WEIGHT	VALUE.IN.	WAGE.IN.K	CROSSING	FINISHING	HEADINGA	SHORTPAS	VOLLEYS	DRIBBLIN	CURVE	FKACCURA	LONGPASS	BALLCONT	ACCELERA	SPRINTSPE	AGILITY	REACTION	BALANCE	SHOTPOW	JUMPING	STAMINA
1	27.41791	86.98507	88.23881	178.806	74.22388	53.84328	190.5075	78.77612	81.25373	65.79104	83.97015	78.9403	85.71642	82.14925	73.62687	76.64179	87.43284	80.28358	77.76119	82.71642	86.16418	79.83582	81.68657	68.8209	80.34328
2	27.44444	86.19444	88.11111	185.8333	80.38889	42.66667	153.8056	60.22222	52.63889	80.08333	80.11111	51.11111	68.66667	59.72222	53.66667	77.25	77.69444	65.88889	70.11111	63.91667	84.02778	61.11111	72.33333	80.38889	77.94444
3	28	87.29412	88.82353	190.3529	86.94118	42.55882	126.7647	14.23529	11.94118	15.29412	42.05882	12.41176	18.17647	15.47059	14.23529	40.47059	29.52941	51.41176	54.52941	54.41176	84.11765	46.41176	59.82353	71.94118	39.47059
Cluster	STRENGTH	LONGSHO	AGGRESSI	INTERCEP	POSITIONI	VISION	PENALTIES	COMPOSU	MARKING	STANDING	SLIDINGTA	GKDIVING	GKHANDLI	GKKICKIN	GKPOSITI	GKREFLEX	W_F	SM	IR	PAC	SHO	PAS	DRI	DEF	PHY
1	67.98507	79.8806	65	50.76119	85.01493	83.35821	74.97015	85.56716	51.35821	49.76119	43.77612	10.53731	10.64179	10.71642	10.40299	10.56716	3.761194	3.895522	3.208955	78.89552	80.86567	81.14925	85.65672	51.53731	70.55224
2	82.27778	59.33333	82.36111	85.83333	58.08333	68.27778	56.11111	83.33333	86.22222	85.86111	83.58333	10.02778	9.861111	9.5	9.833333	10.19444	3.277778	2.527778	2.861111	68.16667	58.25	71.02778	71.30556	85.16667	81.13889
3	71.35294	14.47059	32.76471	21.52941	12.47059	56.29412	25.11765	64.58824	21.41176	14.70588	14.41176	86.47059	84.05882	79.52941	85.47059	88.17647	2.882353	1	3.176471	86.47059	84.05882	79.52941	88.17647	53.05882	85.47059

The business profile in the above picture is obtained from the Kmeans output. Since there are 50 attributes, the profile is take for optimal number of clusters i.e. 3. By considering few attributes which have a significant different mean difference in three clusters, we plot a bar plot for those three clusters to understand those clusters well and profile them well. The barplot for three clusters is shown as below:



The cluster 1 here stands for **Tier 1: Elite**

The cluster 2 here stands for **Tier 2: Pro**

The cluster 3 here stands for **Tier 3: Beginner**

Elite players: Elite players are typically considered to be the best or most skilled players in FIFA. They may have a high overall rating, excel in multiple areas of the game, and consistently outperform other players. The mean values of cluster 1 tells us that the players are performing well as well as paid well and in certain areas like accuracy, volleys, positioning, etc. they differ from other clusters.

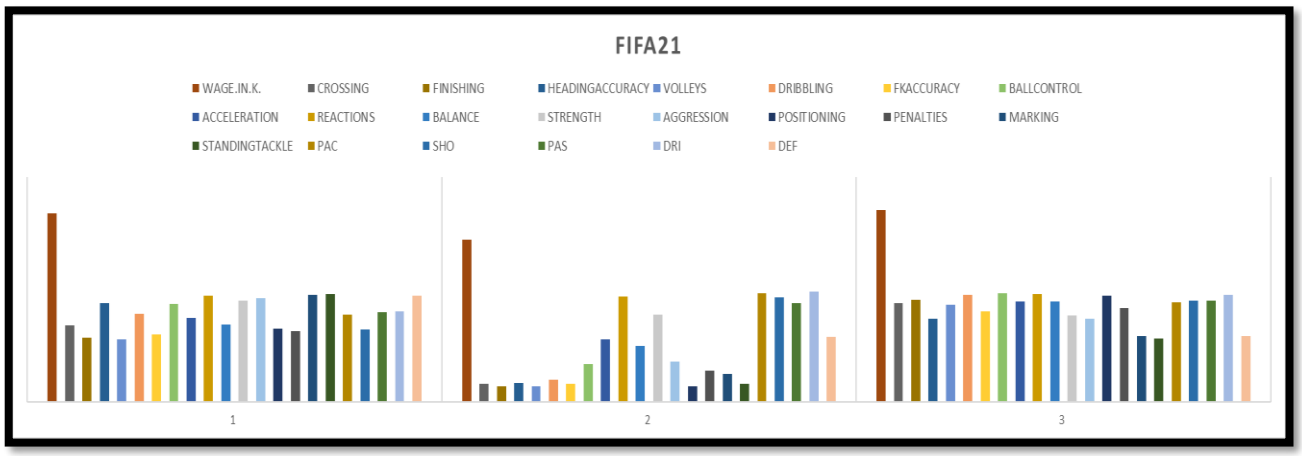
Pro players: Pro players are typically considered to be highly skilled and competitive players, but they may not necessarily be at the same level as elite players. They may have a high overall rating, excel in specific areas of the game, and regularly perform well in competitive matches. So, these players excel in certain areas passing and shooting.

Beginner players: Beginner players are typically new to a game or activity and are still learning the basics. They may have a low pay and less recognition, they struggle with certain aspects of the game, and need more practice in order to improve their skills. Beginners are poor in crossing, finishing and ball control power.

FIFA21

Cluster	AGE	OVA	POT	HEIGHT	WEIGHT	VALUE.IN.	WAGE.IN.	CROSSING	FINISHING	HEADINGA	SHORTPAS	VOLLEYS	DRIBBLIN	CURVE	FKACCURA	LONGPASS	BALLCONT	ACCELERA	SPRINTSPE	AGILITY	REACTION	BALANCE	SHOTPOW	JUMPING	STAMINA
1	28.1875	85.8125	87.03125	185.5625	80.59375	58.32813	150.9375	61.5625	51.53125	79.09375	79.84375	50.21875	70.28125	60.78125	53.96875	78.21875	78.21875	67.03125	71.75	64.46875	84.8125	61.90625	73.46875	80.34375	79.125
2	29.375	87.375	88.5	191.375	87.3125	56.0625	129.9375	14.5	12.3125	15.1875	43.8125	12.8125	18	15.0625	14.25	44.4375	30.0625	50.3125	53.375	54.1875	84	44.75	59.375	69.625	39.3125
3	27.72222	86.19444	87.22222	179.0417	74.40278	71.34028	153.5556	78.94444	81.61111	66.29167	84.26389	77.98611	85.73611	81.77778	72.58333	77.33333	87.06944	80.47222	78.98611	82.59722	86.33333	80.375	82.19444	69.18056	80.22222
Cluster	STRENGTH	LONGSHO	AGGRESSI	INTERCEP	POSITIONI	VISION	PENALTIES	COMPOSU	MARKING	STANDING	SLIDINGTA	GKDIVING	GKHANDLI	GKKICKING	GKPOSITIO	GKREFLEX	W_F	SM	IR	PAC	SHO	PAS	DRI	DEF	PHY
1	81.34375	58.9375	83.03125	85.5	58.4375	69.78125	56.46875	83.75	85.8125	86.15625	83.8125	9.5625	9.65625	9.84375	9.6875	9.9375	3.21875	2.5	2.875	69.65625	57.9375	71.75	72.34375	85.03125	81.0625
2	70.0625	14.6875	32.5	21.75	12.75	55.75	25.125	64.125	22.3125	14.5625	14.5625	86.6875	83.875	78.9375	86	88.5625	2.875	1	3.0625	86.6875	83.875	78.9375	88.5625	52	86
3	69.34722	79.73611	66.70833	52.51389	84.875	83.47222	74.98611	85.26389	52.52778	50.68056	45.15278	10.47222	10.75	10.73611	10.75	10.79167	3.694444	3.847222	3.041667	79.66667	81.04167	81.31944	85.63889	52.70833	71.55556

The business profile in the above picture is obtained from the Kmeans output. Since there are 50 attributes, the profile is taken for optimal number of clusters i.e. 3. By considering few attributes which have a significant different mean difference in three clusters, we plot a bar plot for those three clusters to understand those clusters well and profile them well. The barplot for three clusters is shown as below:



As mentioned above, the optimum clusters remain same in the next series. Just the clusters are changed here.

Cluster1: Here it exhibits the characteristics of **Tier2: Pro**

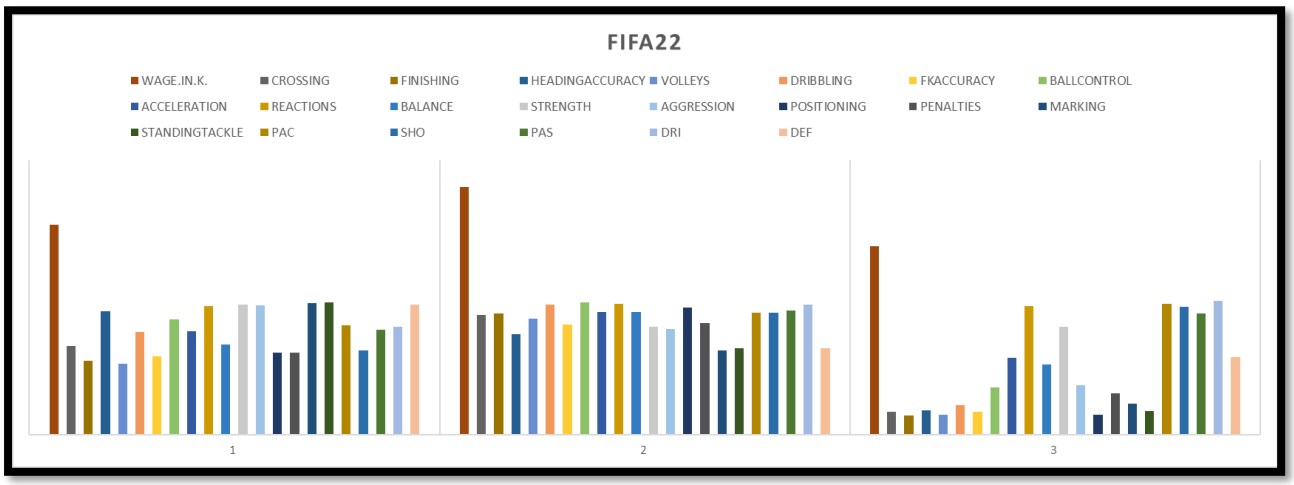
Cluster2: Here it exhibits the characteristics of **Tier3: Beginner**

Cluster3: Here it exhibits the characteristics of **Tier1: Elite**

FIFA22

Cluster	AGE	OVA	POT	HEIGHT	WEIGHT	VALUE.IN.	WAGE.IN.	CROSSING	FINISHING	HEADINGA	SHORTPAS	VOLLEYS	DRIBBLIN	CURVE	FKACCURA	LONGPASS	BALLCONT	ACCELERA	SPRINTSPE	AGILITY	REACTION	BALANCE	SHOTPOW	JUMPING	STAMINA
1	28.16	85.64	86.76	186.6	82.68	55.84	137.76	58.04	48.2	80.6	78.52	46.6	67.32	55	51.4	76.08	75.32	67.88	74.76	61.6	84.04	58.8	72.92	83	76.08
2	27.31169	86.24675	87.61039	179.5195	74.72727	73.14286	162.4026	78.51948	79.51948	65.79221	84.03896	75.94805	85.20779	80.67532	71.90909	77.92208	86.44156	80.55844	78.87013	82.07792	85.8961	80.51948	81.88312	69.74026	81.98701
3	29.83333	86.94444	87.94444	191.6111	86.77778	49.88889	123.3889	14.88889	12.5	15.88889	45.16667	12.77778	19.05556	17.38889	14.88889	44.72222	31	50.05556	52.33333	52.55556	84.22222	45.72222	59.5	68.11111	39.72222
Cluster	STRENGTH	LONGSHO	AGGRESSI	INTERCEP	POSITIONI	VISION	PENALTIES	COMPOSU	MARKING	STANDING	SLIDINGTA	GKDIVING	GKHANDLI	GKKICKIN	GKPOSITI	GKREFLEX	W_F	SM	IR	PAC	SHO	PAS	DRI	DEF	PHY
1	85.4	56.72	84.88	85.04	53.88	64.56	53.68	82.44	86.08	86.76	84.08	10.28	10.08	9.68	9.76	10.04	3.08	2.36	2.92	71.56	55.28	68.68	70.44	85.4	82.88
2	70.77922	79	69.38961	55.96104	83.33766	83.15584	73.01299	84.93506	55.24675	56.35065	51.07792	10.03896	10	10.72727	10.66234	10.75325	3.714286	3.805195	3.272727	79.62338	79.67532	81.11688	85.07792	56.37662	73.25974
3	70.83333	14.66667	32.44444	20.94444	12.83333	57.66667	26.72222	62.38889	20.22222	15.11111	14.77778	85.61111	83.72222	79.16667	85.88889	87.72222	3	1	3.333333	85.61111	83.72222	79.16667	87.72222	50.88889	85.88889

The business profile in the above picture is obtained from the Kmeans output. Since there are 50 attributes, the profile is taken for optimal number of clusters i.e. 3. By considering few attributes which have a significant different mean difference in three clusters, we plot a bar plot for those three clusters to understand those clusters well and profile them well. The barplot for three clusters is shown as below:



As mentioned above, the optimum clusters remain same in the next series. Just the clusters are changed here.

Cluster1: Here it exhibits the characteristics of **Tier2: Pro**

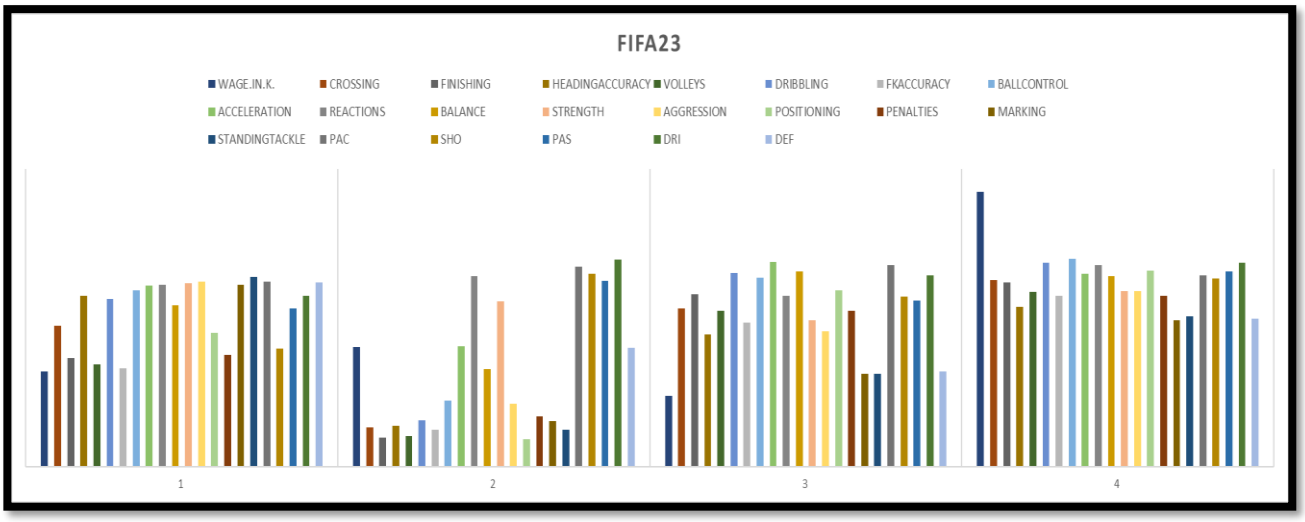
Cluster2: Here it exhibits the characteristics of **Tier1: Elite**

Cluster3: Here it exhibits the characteristics of **Tier3: Beginner**

FIFA23

Cluster	AGE	OVA	POT	HEIGHT	WEIGHT	VALUE.IN.	WAGE.IN.	CROSSING	FINISHING	HEADINGA	SHORTPAS	VOLLEYS	DRIBBLIN	CURVE	FKACCURA	LONGPASS	BALLCONT	ACCELERA	SPRINTSPE	AGILITY	REACTION	BALANCE	SHOTPOW	JUMPING	STAMINA
1	21.76923	75.95604	84.25275	183.8242	75.42857	21.11648	38.37363	56.97802	43.87912	68.78022	73.12088	41.30769	67.59341	52.31868	39.62637	67.50549	71.13187	73.12088	75.82418	67	73.3956	65.12088	59.13187	72.83516	73.85714
2	27.90909	80.81818	84.45455	190.9091	82.81818	25.71818	48.36364	15.81818	11.81818	16.45455	44.27273	12.36364	18.72727	17.45455	15	46.63636	26.63636	48.63636	47.27273	49.90909	76.81818	39.45455	56.45455	63.18182	35.45455
3	20.68421	74.03158	84.58947	177.3368	70.52632	15.18526	28.54737	63.82105	69.58947	53.30526	71.27368	62.88421	78.04211	67.11579	57.97895	61.84211	76.27368	82.57895	80.34737	81.69474	68.94737	78.63158	70.98947	61.46316	68.32632
4	25.50388	82.84496	86.42636	180.969	74.09302	50.65891	110.8372	75.22481	74.42636	64.57364	81.73643	70.5969	82.37209	77.17054	69.04651	76.82946	83.84496	77.92248	76.72093	78.69767	81.44186	77	80.09302	68.1938	79.52713
Cluster	STRENGTH	LONGSHO	AGGRESSI	INTERCEP	POSITIONI	VISION	PENALTIES	COMPOSU	MARKING	STANDING	SLIDINGTA	GKDIVING	GKHANDLI	GKKICKIN	GKPOSITI	GKREFLEX	W_F	SM	IR	PAC	SHO	PAS	DRI	DEF	PHY
1	74.04396	45.96703	74.6044	74.12088	54.03297	60.93407	45.04396	71.82418	73.25275	76.63736	74.28571	9.758242	9.813187	9.681319	9.956044	10.65934	2.989011	2.384615	1.32967	74.65934	47.8022	63.86813	68.97802	74.15385	74.07692
2	66.54545	13.81818	25.36364	17.27273	11.09091	57.54545	20.27273	59.72727	18.45455	15	16.45455	80.63636	77.72727	75.09091	79.90909	83.36364	2.909091	1	1.909091	80.63636	77.72727	75.09091	83.36364	48	79.90909
3	59.03158	65.4	54.55789	34.73684	70.98947	68.68421	62.92632	71.11579	37.50526	37.55789	35.18947	9.694737	9.905263	10.28421	9.842105	10.29474	3.473684	3.6	1.231579	81.38947	68.45263	67.02105	77.14737	38.32632	60.62105
4	70.97674	76.77519	70.85271	59.7907	79.17829	80.46512	68.90698	81.82946	59.10853	60.6124	55.10078	9.968992	10.26357	10.74419	10	10.41085	3.581395	3.697674	2.472868	77.27132	75.8062	78.6124	82.17054	59.86822	72.96124

The business profile in the above picture is obtained from the Kmeans output. Since there are 50 attributes, the profile is taken for optimal number of clusters i.e. 3. By considering few attributes which have a significant different mean difference in three clusters, we plot a bar plot for those three clusters to understand those clusters well and profile them well. The barplot for three clusters is shown as below:



In the recently released series FIFA23, we see that a lot of new players are coming in and their performance mean stats for each cluster is shown above. The clusters can be profiled as follows:

The cluster 4 here stands for **Tier 1: Elite**

The cluster 3 here stands for **Tier 2: Pro**

The cluster 1 here stands for **Tier 3: Amateur**

The cluster 2 here stands for **Tier 4: Beginner**

Tier 3: Amateur players have an average rating and they are generally proficient in their game. Like in cluster 1, we observe that they are performing well in terms of their aggression, passing or shooting. But not as much as the players in Tier1 or Tier 2.