Assignment-2: Clustering Project

Team:

Gayatri Purigilla - 20171309

Keshavan Seshadri - 20171051

Problem

In this project, we visualise and cluster the data using K-means, Hierarchical clustering (Agglomerative and divisive) and DBSCAN.

Antoine Griezmann in FIFA 18.

Data Cleaning

Before we got to building the model, we had cleaned the data in the following ways.

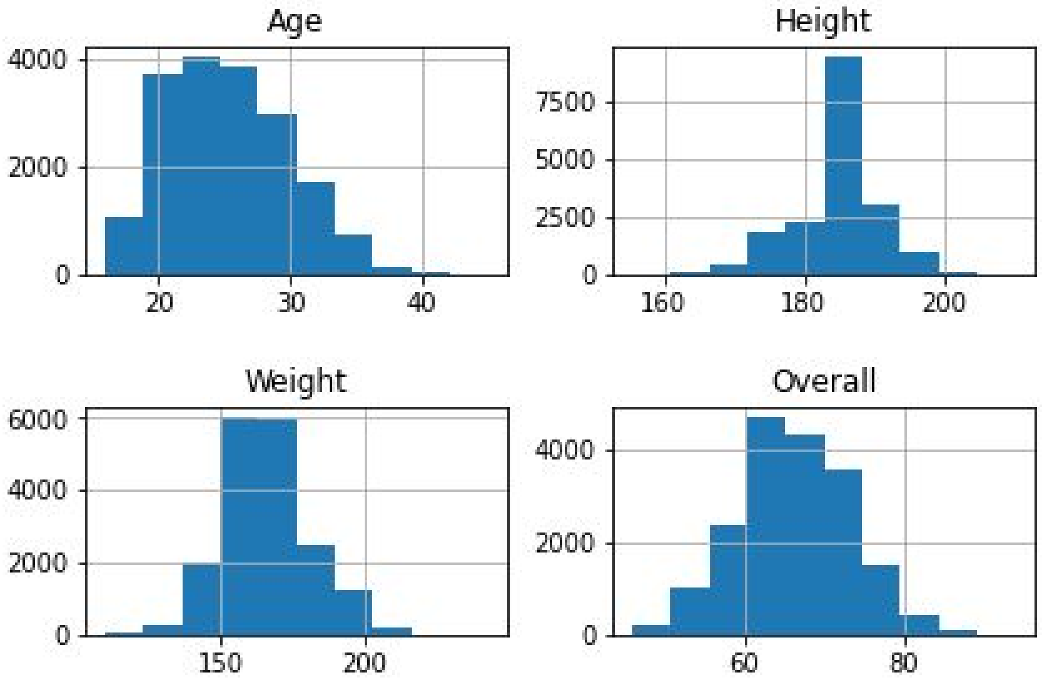
First, we removed the units mentioned in ‘Height’ and ‘Weight’ columns and changed the height from Ft to cms thus converting both the columns into numerical data. Next, the null values in columns with numerical data were replaced with the mean of the values. And finally, the rows in the ‘Club’ column with entries as NaN were filled with ‘No Club’.

Task 1: Data Visualisation

HISTOGRAMS

The following are some of the ways the given data can be visualised using seaborn and matplotlib.

- Histograms of count of players on the basis of age, height, weight, etc.



Age Histogram: The age histogram shows that most players are in between 18 to 30. It also shows that most players retire before 40.

Height Histogram: The height histogram shows that most players are above 6 feet(183 cms). Very few players cross the 200cm.

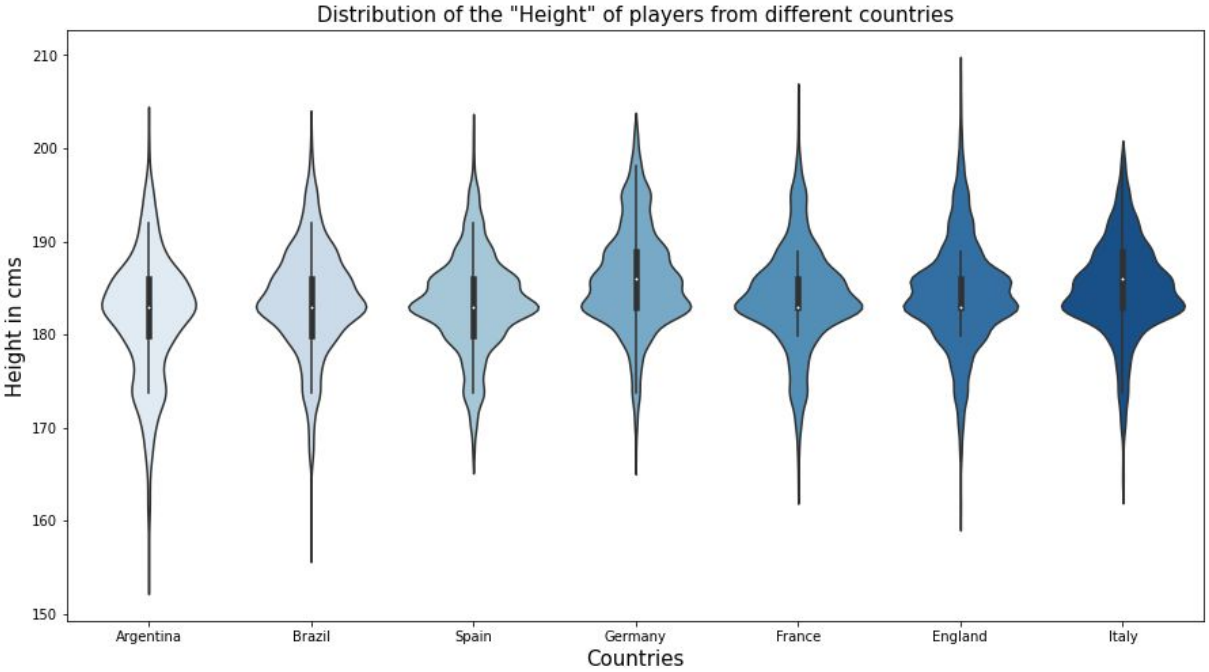
Weight Histogram: Most players weigh between 150 lbs and 175 lbs. Very few players are above 200 lbs. The weight distribution is almost normal.

Overall Histogram: Most players have their overall scores between 60 and 75. Very few players cross 80 and even fewer players cross 90.

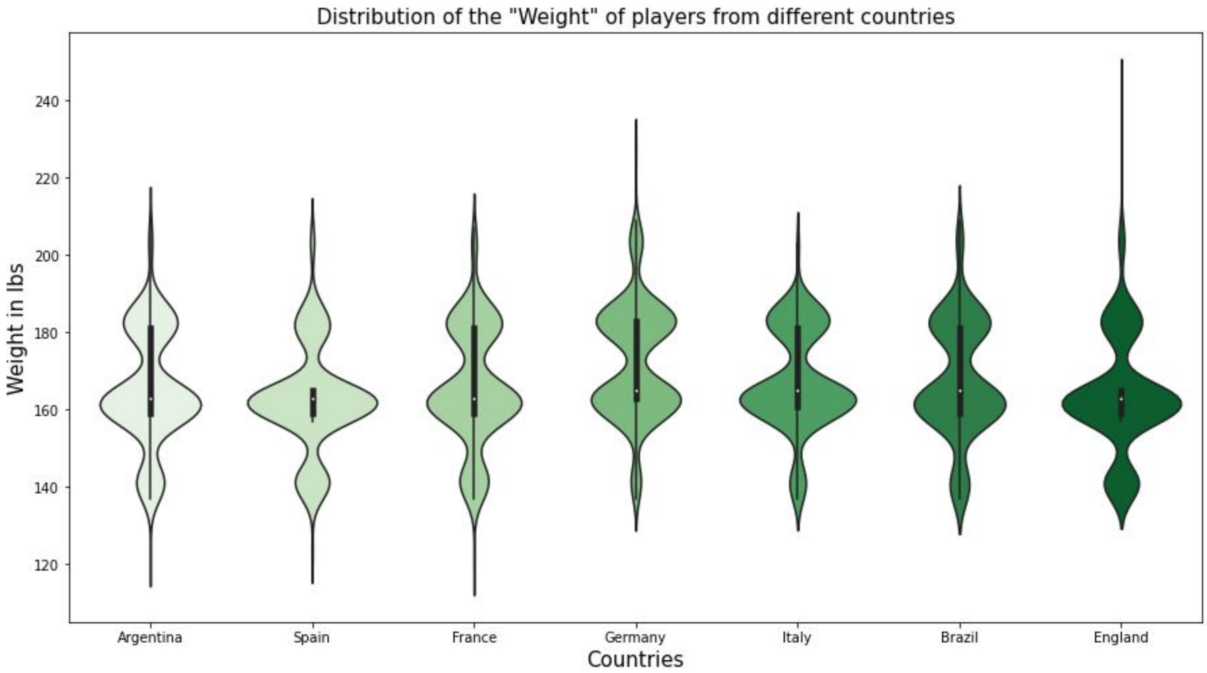
Distribution w.r.t Countries

- The following are the distribution of the features w.r.t to different countries.

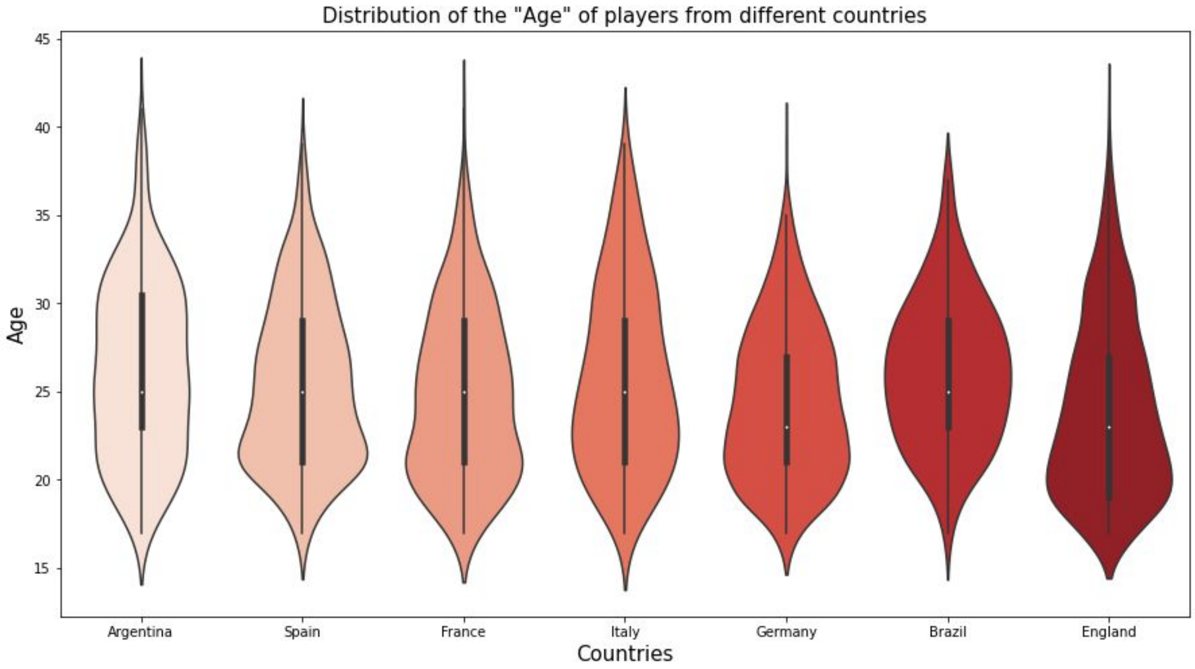
  The height variation is largest in Argentina(vertical length) from 152cms to 202cms. England has one of the tallest players reaching 210 cms.



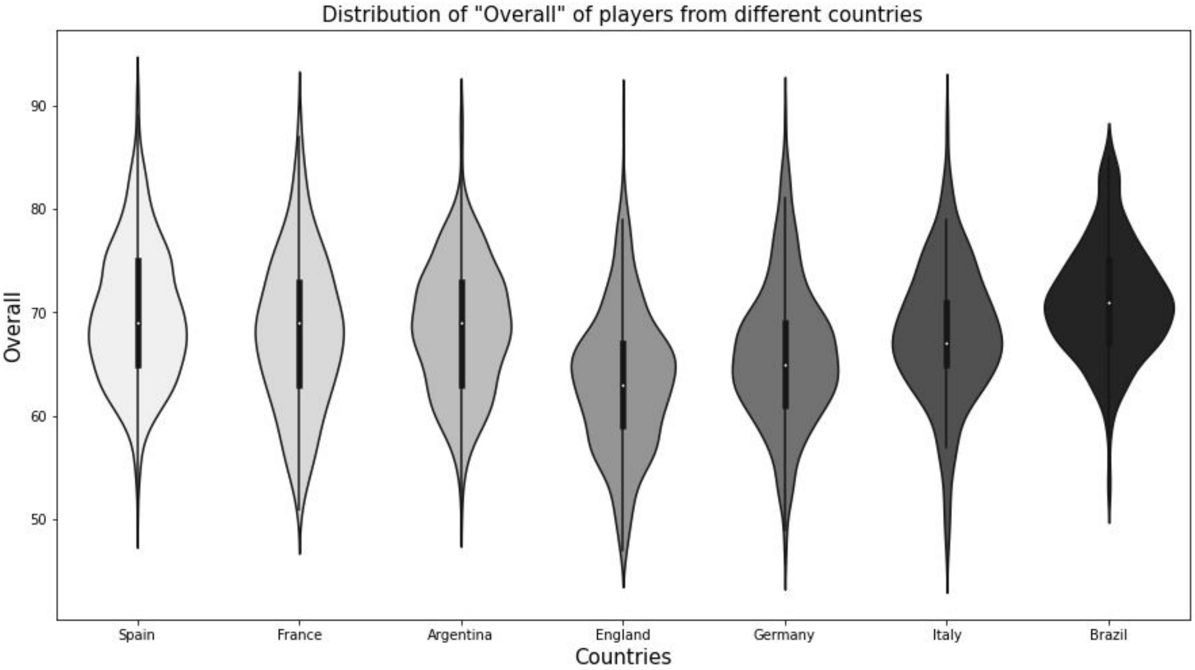
- The French have the least weighed players and the English have the heaviest players. We can observe 3 peaks in the weight distributions in each of these countries.



- Argentina, England and France have the oldest players at around 44 yrs. The age distribution in brazil looks almost normal.

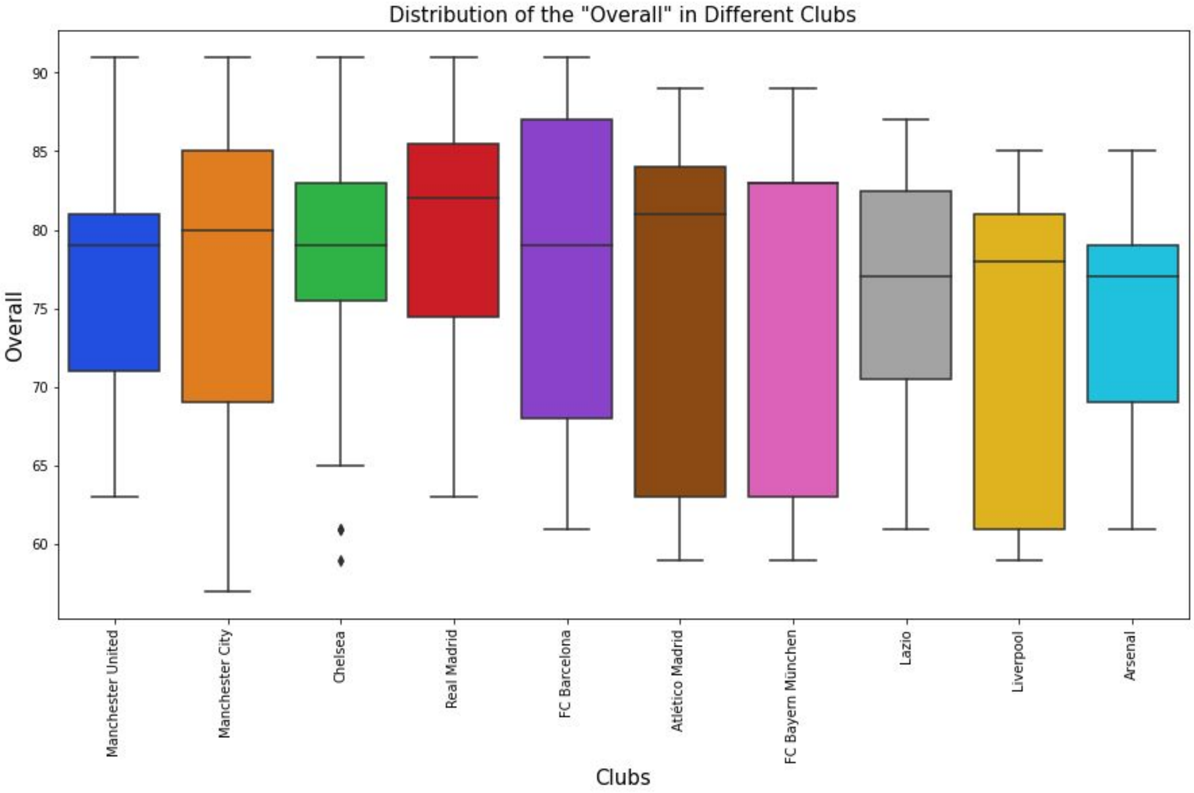


- Except Brazil, all other countries have more players above 90 however the peak of the overall scores is above 70 for brazil. This means that an average player from Brazil is better than an average player from any other country. England and Italy have the largest distribution of overall scores(see vertical length).

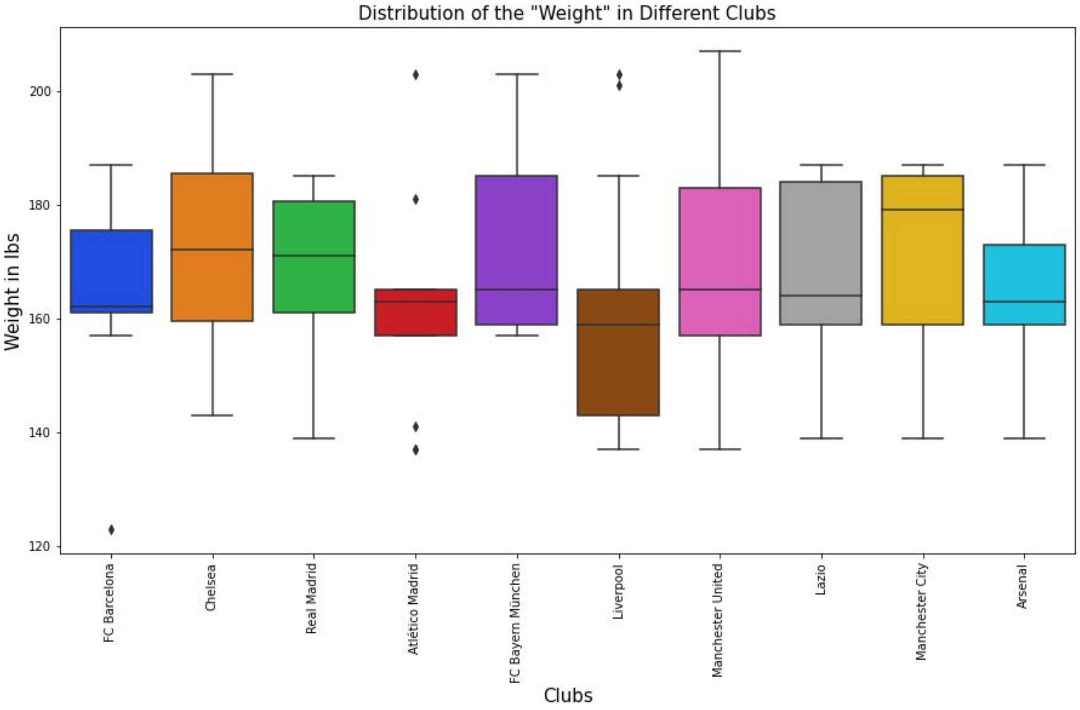


Distribution w.r.t Clubs

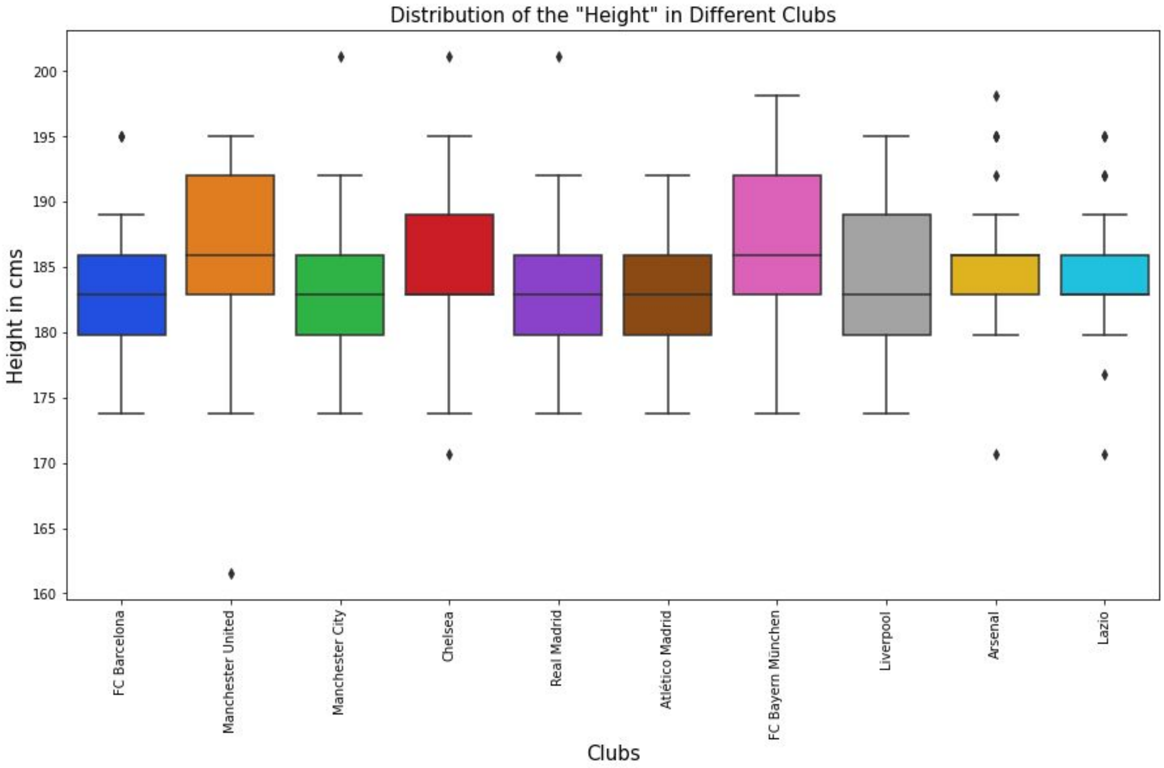
- Chelsea has the smallest distribution(see vertical length). No player from Atletico Madrid , FC Bayern Munchen, Lazio, Liverpool, Arsenal have their ratings above 90. But this could be due to the fact that this data was taken in 2018. Now in 2020, these teams will have better player ratings.



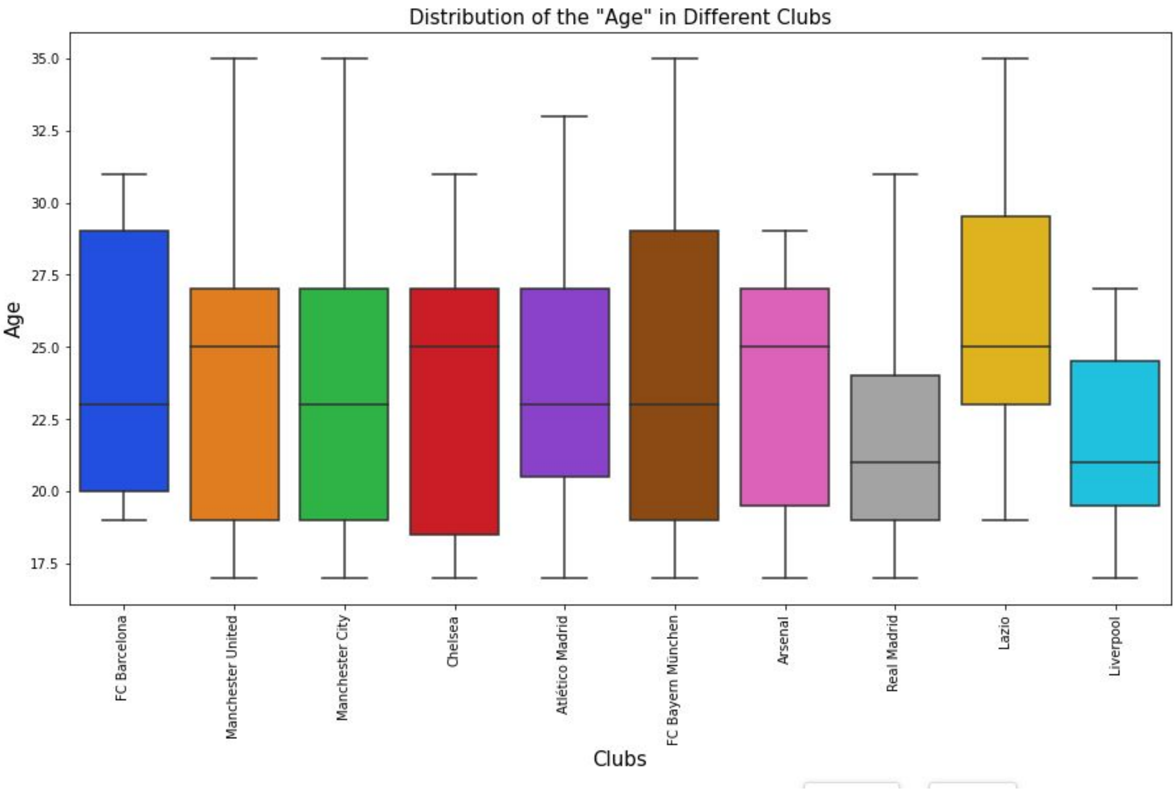
- Manchester United have one of the heaviest and least weighing players. Liverpool players on an average weight less than their counterparts.



- FC Bayern Munchen has the largest height distribution(see vertical length).



- Liverpool have one of the youngest squads with no player above 27.5 years. From the distribution, it can be seen that FC Barcelona and Lazio have very few youth players and therefore show the least focus on developing the youth.



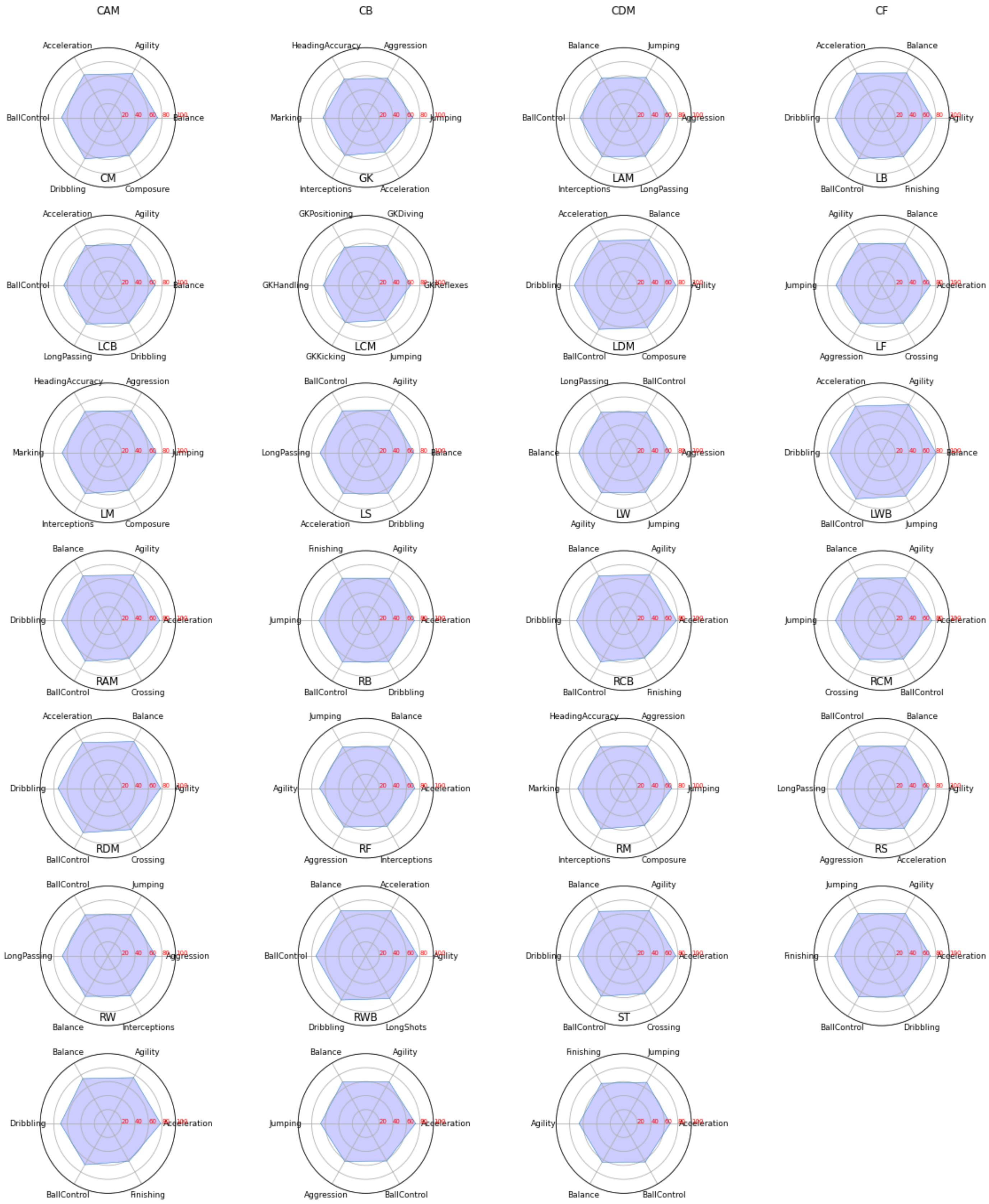
Features w.r.t Position

- These are the top 6 Features for each Position. This is done by calculating the mean of all features and selecting the top 6 features corresponding to each position.

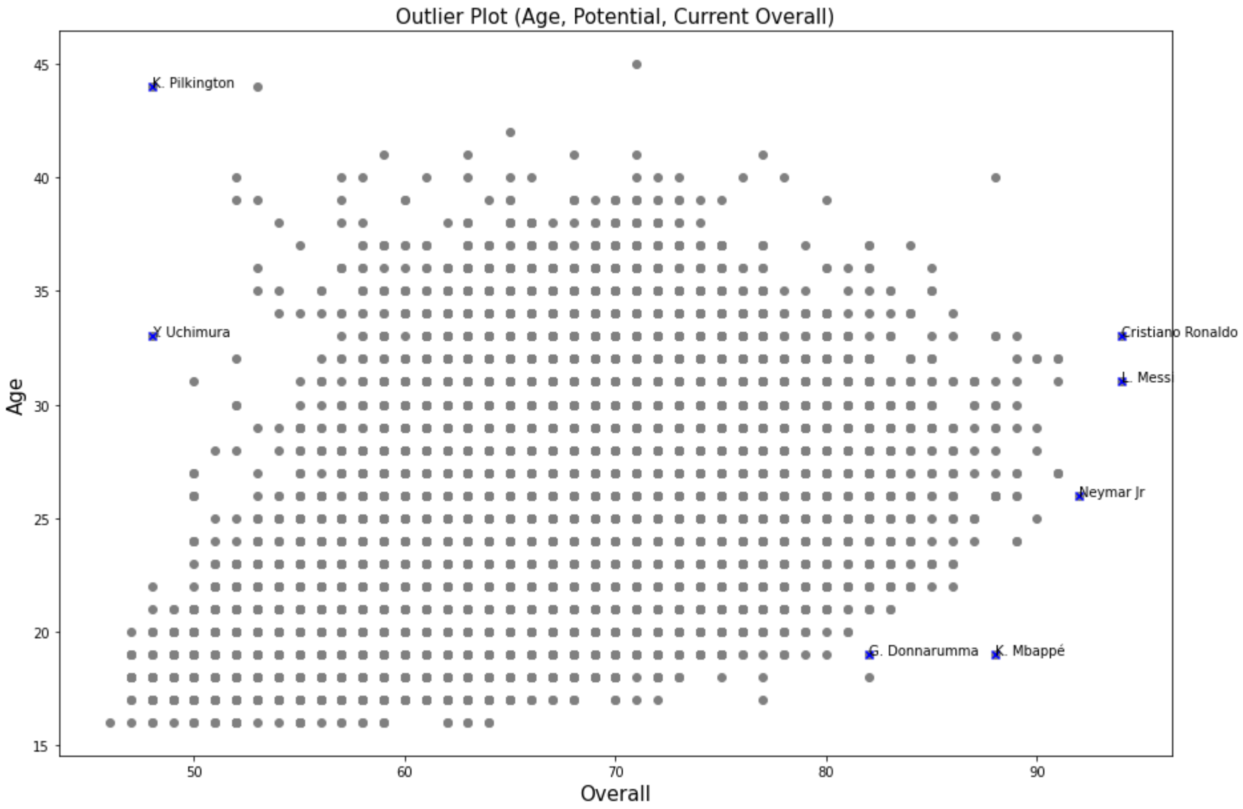
Position CAM: Balance, Agility, Acceleration, BallControl, Dribbling, Composure
Position CB: Jumping, Aggression, HeadingAccuracy, Marking, Interceptions, Acceleration
Position CDM: Aggression, Jumping, Balance, BallControl, Interceptions, LongPassing
Position CF: Agility, Balance, Acceleration, Dribbling, BallControl, Finishing
Position CM: Balance, Agility, Acceleration, BallControl, LongPassing, Dribbling
Position GK: GKReflexes, GKDiving, GKPositioning, GKHandling, GKKicking, Jumping
Position LAM: Agility, Balance, Acceleration, Dribbling, BallControl, Composure
Position LB: Acceleration, Balance, Agility, Jumping, Aggression, Crossing
Position LCB: Jumping, Aggression, HeadingAccuracy, Marking, Interceptions, Composure
Position LCM: Balance, Agility, BallControl, LongPassing, Acceleration, Dribbling
Position LDM: Aggression, BallControl, LongPassing, Balance, Agility, Jumping
Position LF: Balance, Agility, Acceleration, Dribbling, BallControl, Jumping
Position LM: Acceleration, Agility, Balance, Dribbling, BallControl, Crossing
Position LS: Acceleration, Agility, Finishing, Jumping, BallControl, Dribbling
Position LW: Acceleration, Agility, Balance, Dribbling, BallControl, Finishing
Position LWB: Acceleration, Agility, Balance, Jumping, Crossing, BallControl
Position RAM: Agility, Balance, Acceleration, Dribbling, BallControl, Crossing

Position RB: Acceleration, Balance, Jumping, Agility, Aggression, Interceptions
Position RCB: Jumping, Aggression, HeadingAccuracy, Marking, Interceptions, Composure
Position RCM: Agility, Balance, BallControl, LongPassing, Aggression, Acceleration
Position RDM: Aggression, Jumping, BallControl, LongPassing, Balance, Interceptions
Position RF: Agility, Acceleration, Balance, BallControl, Dribbling, LongShots
Position RM: Acceleration, Agility, Balance, Dribbling, BallControl, Crossing
Position RS: Acceleration, Agility, Jumping, Finishing, BallControl, Dribbling
Position RW: Acceleration, Agility, Balance, Dribbling, BallControl, Finishing
Position RWB: Acceleration, Agility, Balance, Jumping, Aggression, BallControl
Position ST: Acceleration, Jumping, Finishing, Agility, Balance, BallControl

We now visualize the importance of these top 6 features vs position using Radar charts.


CAM
Acceleration
Agility
BallControl
Balance
Dribbling
Composure
20 40 60 80 100
CB
HeadingAccuracy
Aggression
Marking
Jumping
Interceptions
Acceleration
20 40 60 80 100
CDM
Balance
Jumping
BallControl
Aggression
Interceptions
LongPassing
20 40 60 80 100
CF
Acceleration
Balance
Dribbling
Agility
BallControl
Finishing
20 40 60 80 100
CM
Acceleration
Agility
BallControl
Balance
LongPassing
Dribbling
20 40 60 80 100
GK
GKPositioning
GKDiving
GKHandling
GKReflexes
GKKicking
Jumping
20 40 60 80 100
LAM
Acceleration
Balance
Dribbling
Agility
BallControl
Composure
20 40 60 80 100
LB
Agility
Balance
Jumping
Acceleration
Aggression
Crossing
20 40 60 80 100
LCB
HeadingAccuracy
Aggression
Marking
Jumping
Interceptions
Composure
20 40 60 80 100
LCM
BallControl
Agility
LongPassing
Balance
Acceleration
Dribbling
20 40 60 80 100
LDM
LongPassing
BallControl
Balance
Aggression
Agility
Jumping
20 40 60 80 100
LF
Acceleration
Agility
Dribbling
Balance
BallControl
Jumping
20 40 60 80 100
LM
Balance
Agility
Dribbling
Acceleration
BallControl
Crossing
20 40 60 80 100
LS
Finishing
Agility
Jumping
Acceleration
BallControl
Dribbling
20 40 60 80 100
LW
Balance
Agility
Dribbling
Acceleration
BallControl
Finishing
20 40 60 80 100
LWB
Balance
Agility
Jumping
Acceleration
Crossing
BallControl
20 40 60 80 100
RAM
Acceleration
Balance
Dribbling
Agility
BallControl
Crossing
20 40 60 80 100
RB
Jumping
Balance
Agility
Acceleration
Aggression
Interceptions
20 40 60 80 100
RCB
HeadingAccuracy
Aggression
Marking
Jumping
Interceptions
Composure
20 40 60 80 100
RCM
BallControl
Balance
LongPassing
Agility
Aggression
Acceleration
20 40 60 80 100
RDM
BallControl
Jumping
LongPassing
Aggression
Balance
Interceptions
20 40 60 80 100
RF
Balance
Acceleration
BallControl
Agility
Dribbling
LongShots
20 40 60 80 100
RM
Balance
Agility
Dribbling
Acceleration
BallControl
Crossing
20 40 60 80 100
RS
Jumping
Agility
Finishing
Acceleration
BallControl
Dribbling
20 40 60 80 100
RW
Balance
Agility
Dribbling
Acceleration
BallControl
Finishing
20 40 60 80 100
RWB
Balance
Agility
Jumping
Acceleration
Aggression
BallControl
20 40 60 80 100
ST
Finishing
Jumping
Agility
Acceleration
Balance
BallControl
20 40 60 80 100

Outliers



The conditions used to mark outliers are as follows:

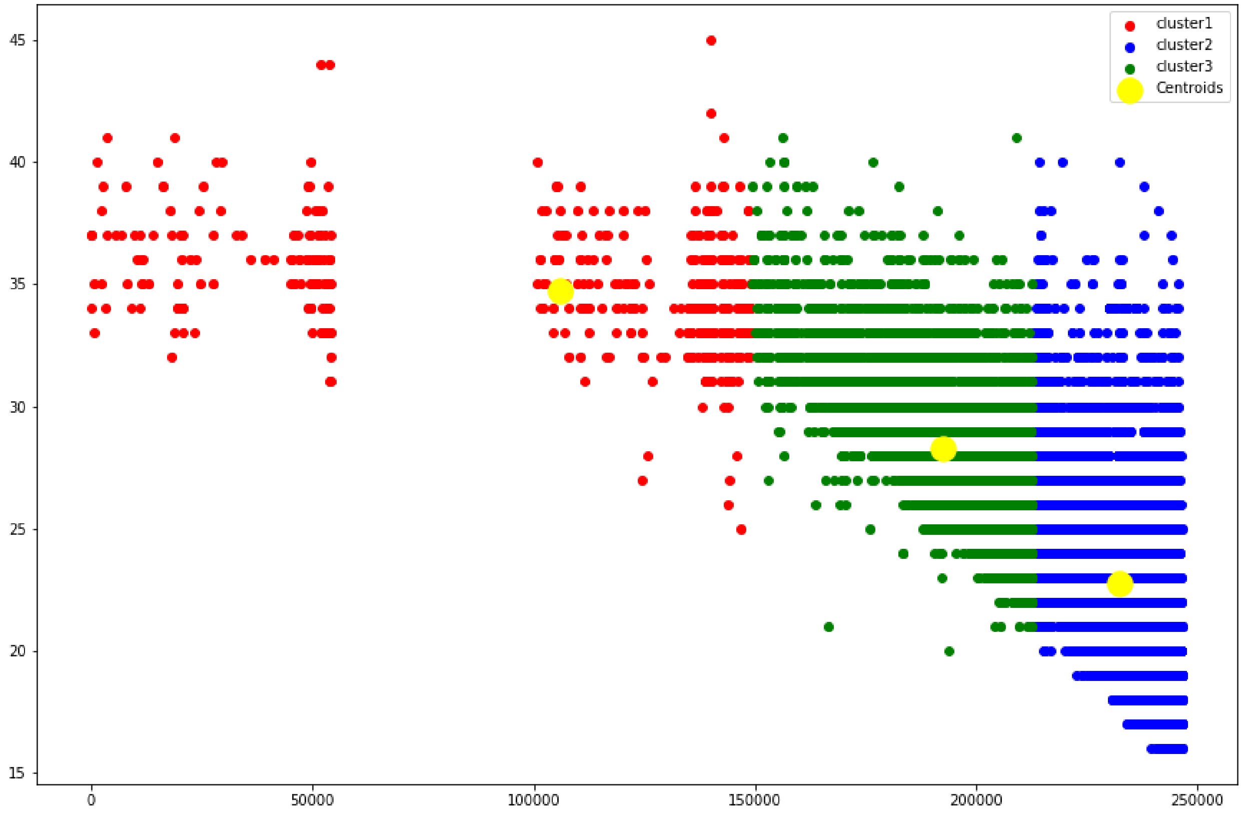
- Overall >= 92
- Age > 40 & Overall > 85
- Age < 20 & Potential > 92
- Overall < 50 & Age > 30

C. Ronaldo, L. Messi, and Neymar Jr are the top 3 stars of the game with overall >=92.
G. Donnarumma and K.Mbappe have the potential to become the future stars.
Y. Uchimura and K. Pilkington are old(age>30) and have overall < 50 making them bad players in the game.

TASK 2: K-means

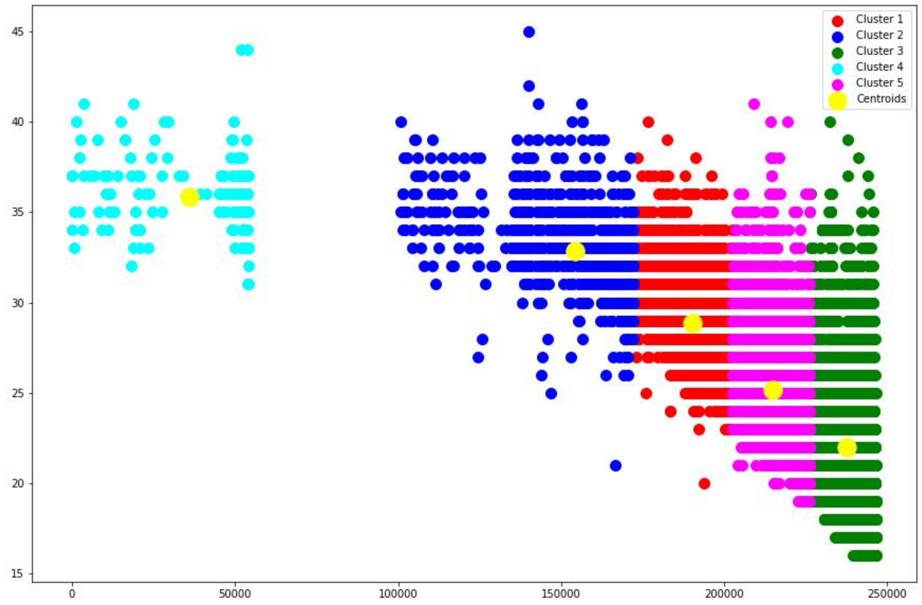
This is the graph obtained on running our K-means the clustering algorithm on the raw dataset with 44 features.
The plots for K=3,5 and 7 are as follows

- K = 3



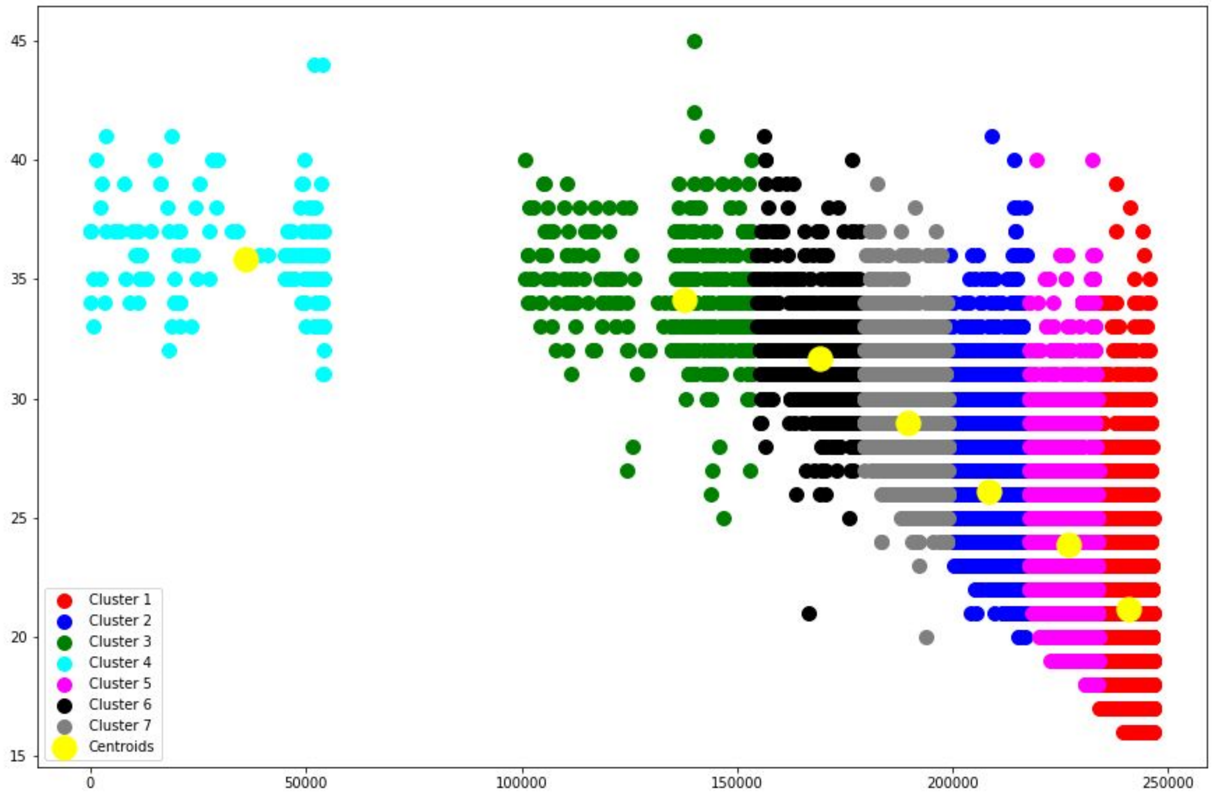
We could infer the following from the above graph after calculating the mean of attributes in each cluster:
Blue - Forwards.
Green - Midfielders
Red - Goalkeepers and Defenders.

- K = 5



We could infer the following from the above graph after calculating the mean of attributes in each cluster:
Cyan - Goalkeepers
Blue - Defenders
Red - Midfielders
Pink: Wingers
Green: Strikers

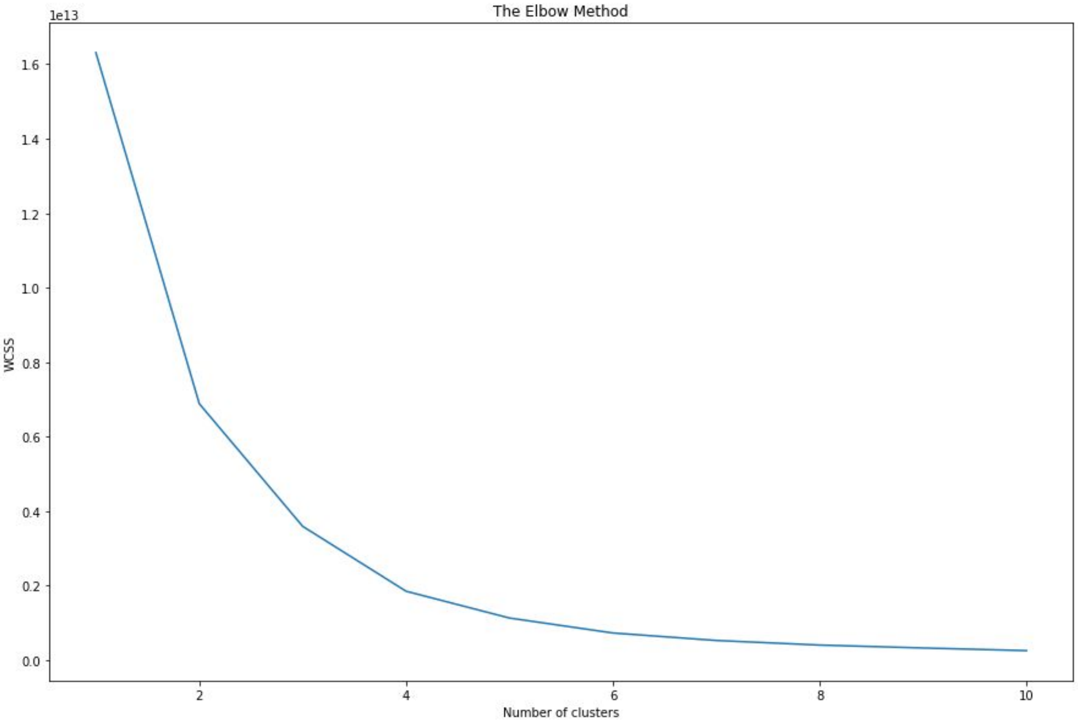
- K = 7

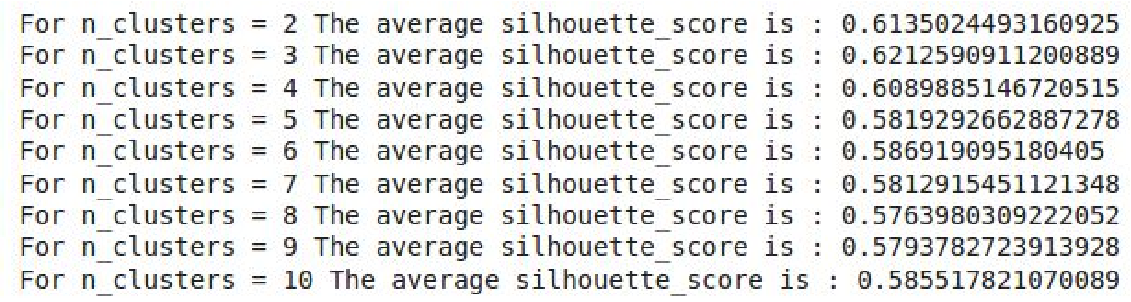


We could infer the following from the above graph after calculating the mean of attributes in each cluster:
Cyan - Goalkeepers
Green - Defenders
Black - Defensive Midfielders
Grey - Midfielders
Blue , Pink- Wingers
Red : Strikers

The plot obtained by elbow method and the average silhouette scores for different number of clusters are as follows

- Elbow Method





```
For n_clusters = 2 The average silhouette_score is : 0.6135024493160925
For n_clusters = 3 The average silhouette_score is : 0.6212590911200889
For n_clusters = 4 The average silhouette_score is : 0.6089885146720515
For n_clusters = 5 The average silhouette_score is : 0.5819292662887278
For n_clusters = 6 The average silhouette_score is : 0.586919095180405
For n_clusters = 7 The average silhouette_score is : 0.5812915451121348
For n_clusters = 8 The average silhouette_score is : 0.5763980309222052
For n_clusters = 9 The average silhouette_score is : 0.5793782723913928
For n_clusters = 10 The average silhouette_score is : 0.585517821070089
```

-

For the elbow method we use WCSS i.e. Within-Cluster Sums of Squares, which is defined as Sum of squares of distances of every data point from its corresponding cluster centroid. Upon plotting the value of WCSS with the number of clusters, The location of a bend in the plot is considered as an indicator of the appropriate number of clusters. I.e. the point after which WCSS doesn't decrease very rapidly with increasing K value.

- Avg silhouette scores

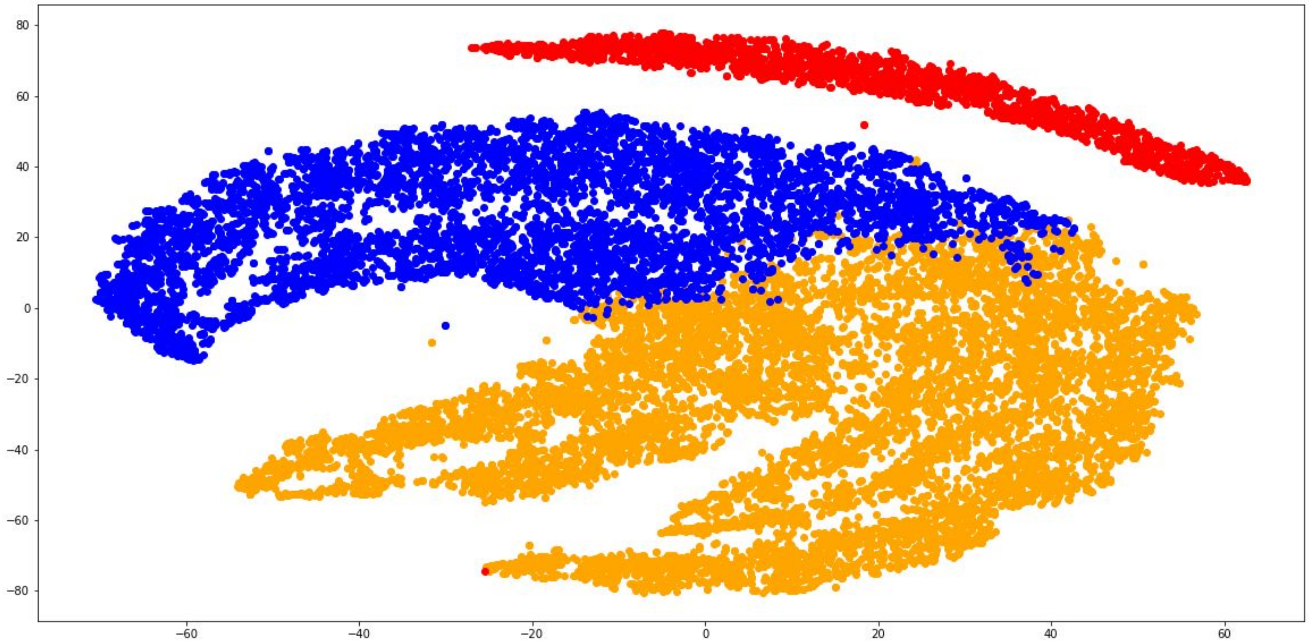
The Silhouette score is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette score for a sample is (b - a) / max(a, b).

The best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

From both the above methods, it can be concluded that the optimum value of k is equal to 3 since it is the location of the bend in the Elbow plot and has the highest average silhouette score.

TASK 3

Agglomerative Clustering



Agglomerative clustering is a hierarchical clustering method using bottom-up strategy. It is clear from the plot that Kmeans gives better results than Agglomerative Clustering.

Positions inferred from the above graph:

RED - Goalkeepers

BLUE - Defenders and Defensive Midfielders

YELLOW - Forwards and Attacking Midfielders

Divisive Clustering

Divisive clustering is another hierarchical clustering method but this uses a top down strategy. This implies that all the data points are considered as one single cluster in the beginning and with each iteration, they are broken down into different clusters.

For divisive clustering, we followed an iterative K-means approach.

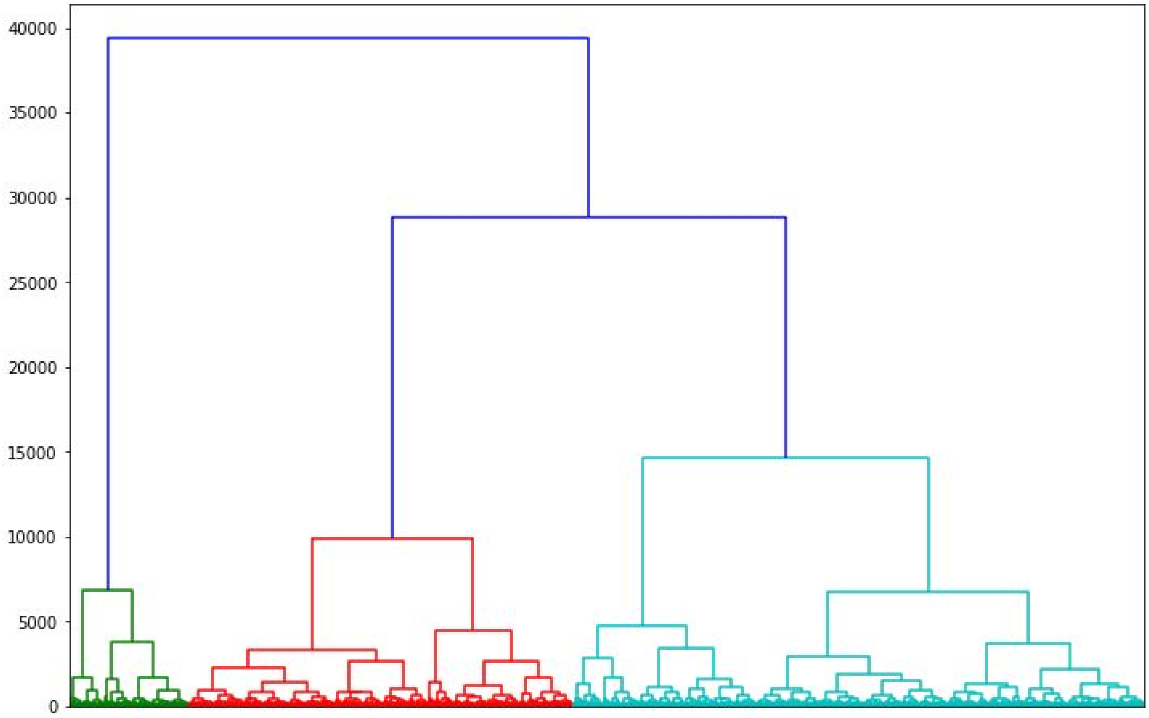
We applied the K-means algorithm on the best cluster and then iteratively partitioned it until each data is in its own singleton cluster.

The divisive clustering is more complex, more efficient and more accurate than agglomerative clustering.

The time complexity is Linear O(N): Divisive clustering given a fixed number of top levels, using an efficient flat algorithm like K-Means, the algorithm is linear in the number of patterns and clusters.

Dendrogram

The dendrogram illustrates how each cluster is composed by drawing a U-shaped link between a non-singleton cluster and its children. The top of the U-link indicates a cluster merge. The two legs of the U-link indicate which clusters were merged. The length of the two legs of the U-link represents the distance between the child clusters.



TASK 4: DBSCAN

Here we use DBSCAN(Density-Based Spatial Clustering of Applications with Noise) to cluster the data.
It has two important parameters ε (epsilon), and minPts.

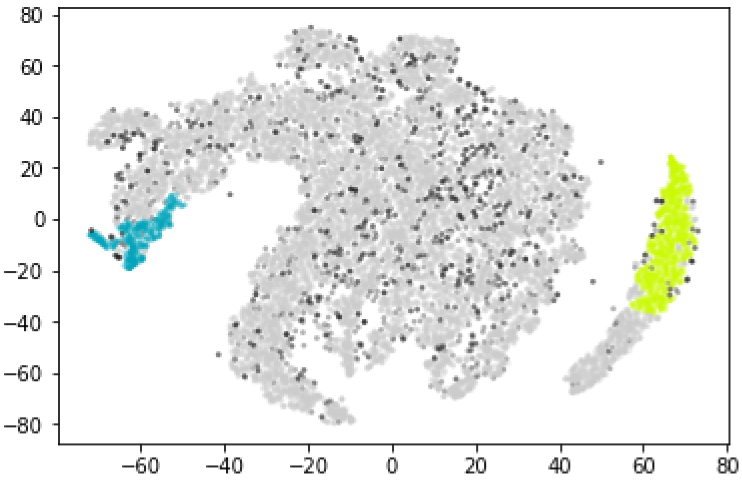
1. ε - The parameter specifying the radius of a neighborhood with respect to some point.
2. MinPts - The minimum number of points required to form a dense region.

The DBSCAN algorithm consists of the following steps (ref. Wikipedia):

3. It finds the points in the ε (epsilon) neighborhood of every point, and identifies the core points with more than minPts neighbors.
4. Then it finds the connected components of *core* points on the neighbor graph, while ignoring all non-core points.
5. Afterwards, it assigns each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assigns it to noise.

SUBTASK 1: Plot

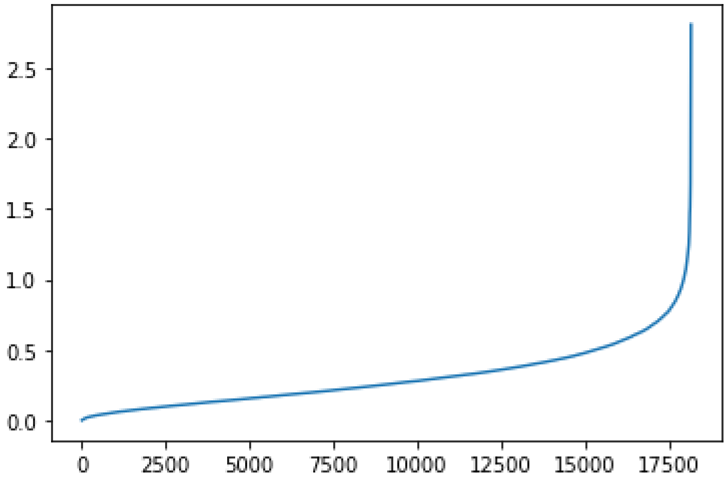
This is the DBSCAN plot obtained on running the data with $\varepsilon = 1.0,$ minPts = 5.



This clearly does not identify the clusters properly. Therefore we have to optimize the values of ε and minPts.

SUBTASK 2: Optimize ε and MinPts

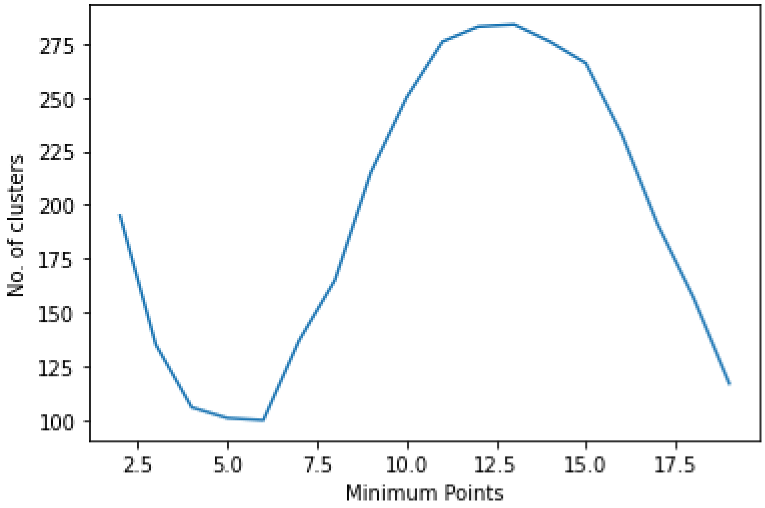
In order to find the optimum ε value, we find the distance to the nearest n points for each point. Then we sort and plot the results obtained.
Then we look at the plot to see an "elbow". The value at which epsilon occurs is selected as epsilon value.



From the plot, we can see that an epsilon value of 1.25 would be most optimum.

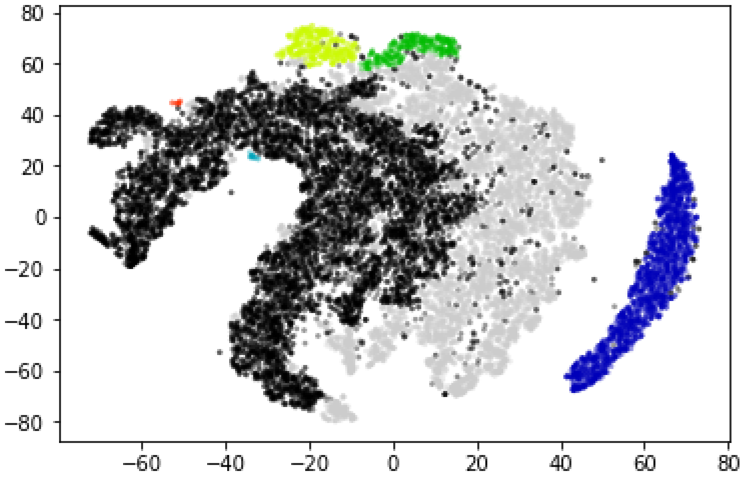
Now we need to optimize minPts. Generally the minPts can be derived from the number of dimensions D in the dataset, as minPts ≥ D + 1.

In this case, due to the noise and the size of the data we do the following.
So we run the DBSCAN algorithm from minPts = 2 to 20 and plot the number of clusters obtained.



We select the value with the least number of clusters formed. In this case, we get minPts as 6.

Now we plot for the optimum values of epsilon = 1.25 and minPts = 6



SUBTASK 3 - Analysis of DBSCAN Clusters:

1. The meaning of the clusters w.r.t to position of the players (determined using mean of attributes inside the cluster):
    a. Blue colored cluster: Goalkeeper(GK) : Since the attributes of Goalkeepers are very different from the rest of the players.
    b. Black colored cluster: Midfielder(CAM, CM, CDM, RM, LM)
    c. Grey colored cluster: Defender(CB, RB, LB)
    d. Yellow: Strikers(ST).
    e. Green: Wingers(RW, LW)

2. DBSCAN does not require one to specify the number of clusters in the data a priori, as opposed to k-means clustering algorithm.
3. DBSCAN can find arbitrarily-shaped clusters and also has the notion of noise, thus robust to outliers.
4. However, DBSCAN does not work well with clusters of varying densities. Regardless of the shape(convex or concave), DBSCAN clusters well within a particular density. But it suffers when the density varies.
5. There are multiple small clusters formed in our graph since DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data are processed.
6. The average Silhouette score we obtained is : -0.47441062 which indicates that the clusters overlap with one another. The negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.

   Therefore, DBSCAN is ***NOT*** a good algorithm for clustering this particular dataset.

FINAL ANALYSIS: CONCLUSION

In the report, we present various data visualization plots and methods to look and understand the data. Then, we use different clustering techniques to cluster the data.

For each of the clustering techniques, we look at the intra and inter cluster distance and also the silhouette score to judge how “good” the clustering technique is.

For k-Means clustering technique we use both Elbow Method and Silhouette score to calculate the optimum value of k. k=3 has the highest average silhouette score.

For n_clusters = 3 The average silhouette_score is : 0.6212590911200889

The best algorithms are K-MEANS and Agglomerative clustering. Since they have the highest silhouette scores and form meaningful clusters.

The DBSCAN is the worst performing algorithm with a negative Silhouette Score of -0.47441062 which indicates wrong clustering and huge overlaps between clusters.

Within each cluster formed, we can calculate the mean of the attributes to determine the position of the player.

The optimum number of clusters for any algorithm should either be 3, 4, or 5.
For 3 clusters: Forwards, Midfielders, Defenders and Goalkeepers
For 4 clusters: Goalkeepers, Defenders, Midfielders, Attackers/Forwards

For 5 clusters: Goalkeepers, Defenders, Midfielders, Wingers, Strikers. (still slightly confusing and can change from one dataset to another)

However for the number of clusters greater than 5 it becomes more and more confusing.

We found outliers in the dataset with the following conditions:

- Overall >= 92
- Age > 40 & Overall > 85
- Age < 20 & Potential > 92
- Overall < 50 & Age > 30

C. Ronaldo, L. Messi, and Neymar Jr are the top 3 stars of the game with overall >=92. They had the highest scores for their positions.
G. Donnarumma and K.Mbappe have the potential to become the future stars. They had the highest potential despite being young(<20 yrs)
Y. Uchimura and K. Pilkington are old(age>30) and have overall < 50 making them bad players in the game.

Hidden Patterns

1. Goalkeepers do not have common attributes with the rest of the players.
2. An average player from Brazil is better than an average player from any other country despite as seen from the distribution.
3. England has one of the tallest players and Argentina has one of the shortest players.
4. Acceleration is the most important attribute for players playing in the wide area of the field.
5. There are 3 peaks in the weight distribution of countries whereas age and overall distribution is almost normal.