Clusturing Report

September 27, 2020

0.0.1 Data Cleaning

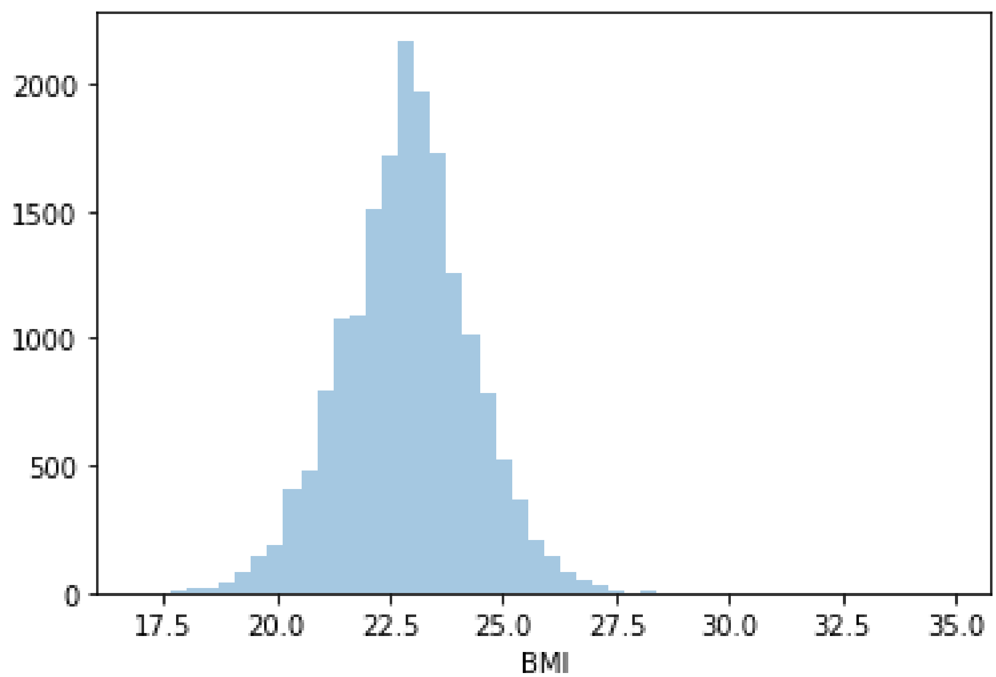
- First we cleaned the data by removing coloums and rows with NaN depending on number of occrence of it
- Next we tried to convert non numeric data to numeric one as much as possible (eg. left and right foot to '0' and '1'. And removed coloums as Name,Images,Flag etc.
- Further we converted all the string numerical data to float or int (eg. Income form €3M to 3000000)

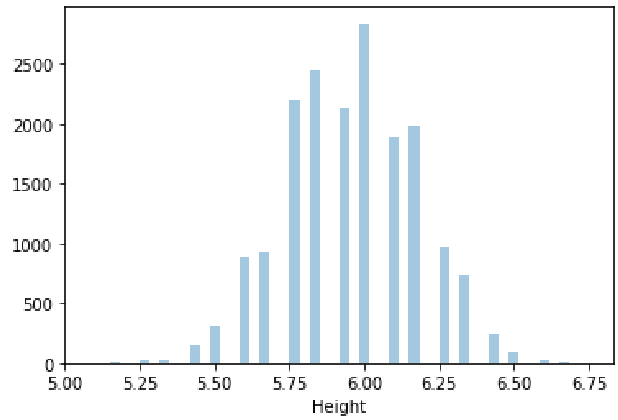
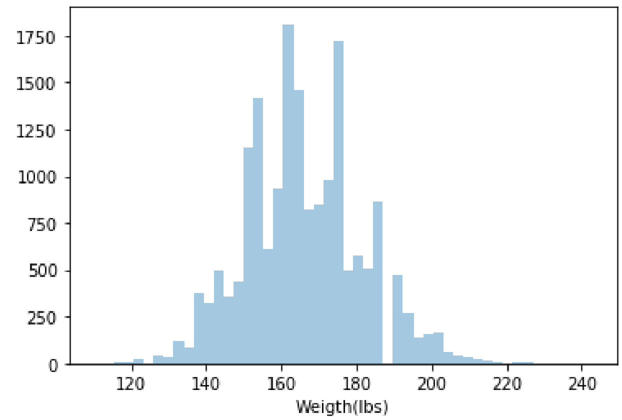
1 1.Analysis from Data Visualisation

1.0.1 A. Weight,Height,BMI

We observe that weight and height does not follow normal distribution perfectly but the BMI follows it much more accurataly.

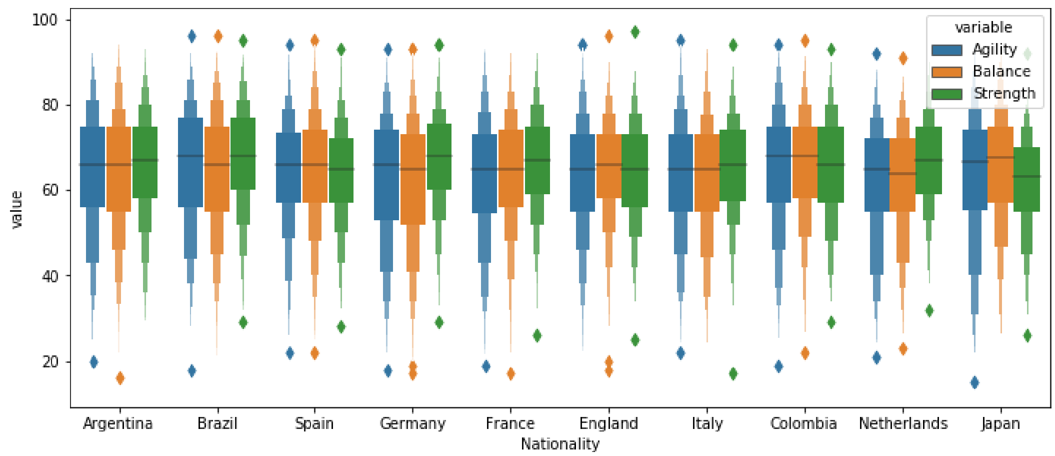
The avarage BMI of all the players is ~23 and range of BMI is 17.5 to 28, which indicates most of players have normal/healthy BMI.



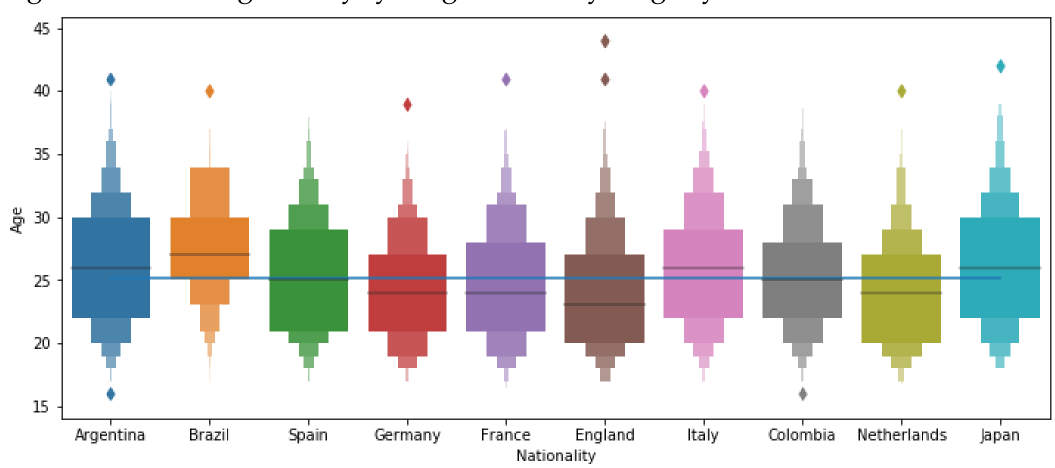
.	.
	

1.0.2 B. Distribution of players in different country

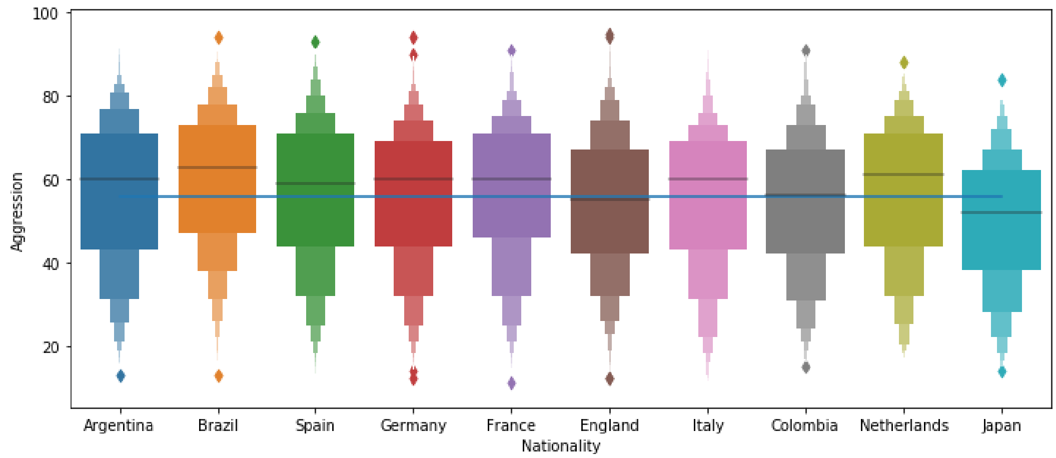
- The Aglity,balance and Strength of Top 10 performing nations is shown below which shows very similar trends.



- The Age of Players of Top 10 countries. We can see that either top performing coutries are generally younger or only slightly older than the world avarage .

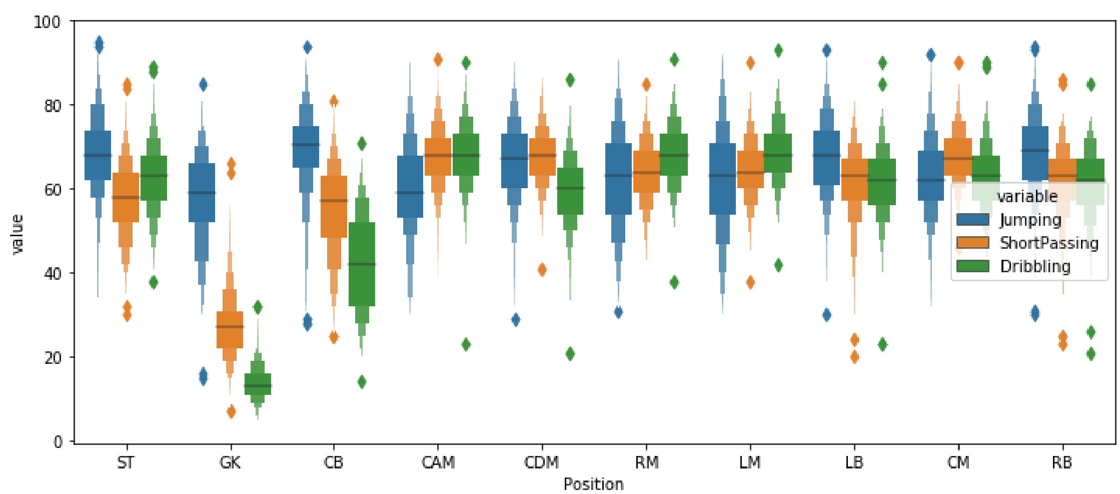


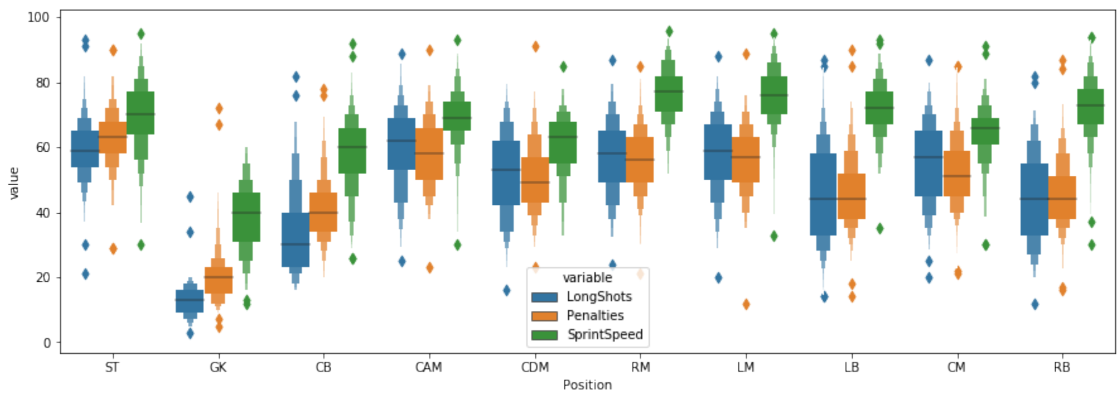
- Aggression of Players of Top 10 countries. Its clear top performers are much more aggressive.



1.0.3 C. Features of players according to their position.

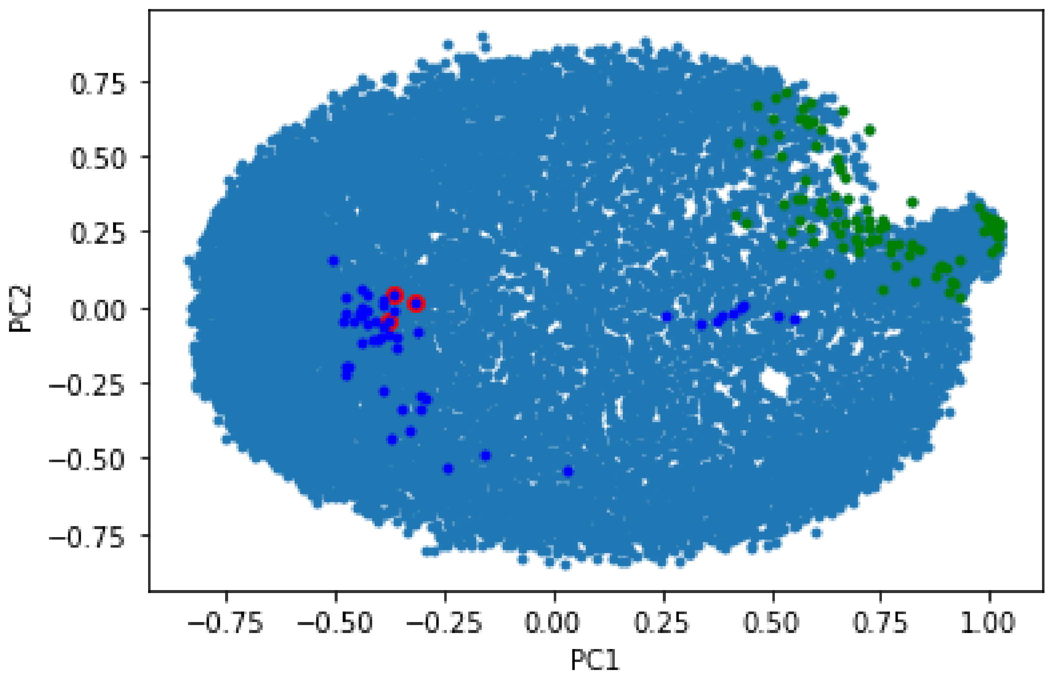
1. Goal Keepers can be easily identified by their low sprint speed.
2. Both Goal Keeper and Center Back(Defender) have low dribbling skill, which is intututive too as they jusr have to pass the ball forward.
3. Both the striker and Defender have high Jumps which is a nesseasarry skill for scoring and defending a corner.
4. The mid fielders(RM and LM) have high sprint speed as they are require to assist both the front and back.
5. The Skills are generally independt of left and right posision (exception may be the primary foot)





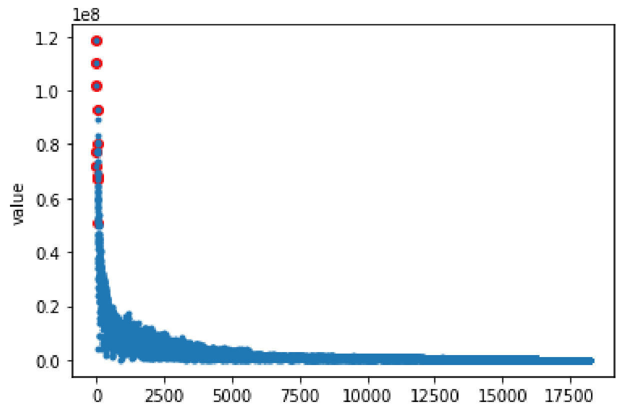
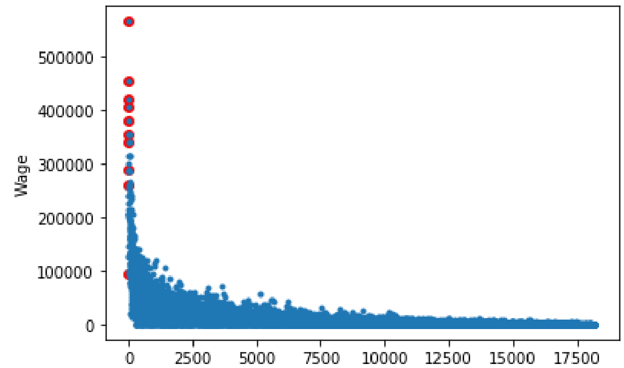
1.0.4 D.Outliers

We have reduced the dimention of data using PCA to get a 2D plot. The Pink Dots are top 3 Players (Messi, Ronado,Neyamer). The violet Dots are top 50 Players. Green are last 100 Players.



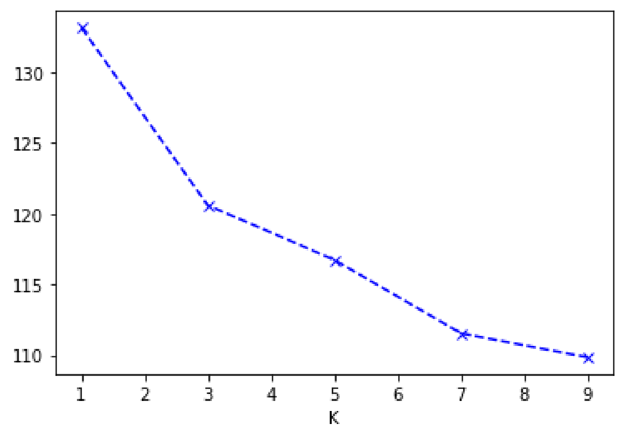
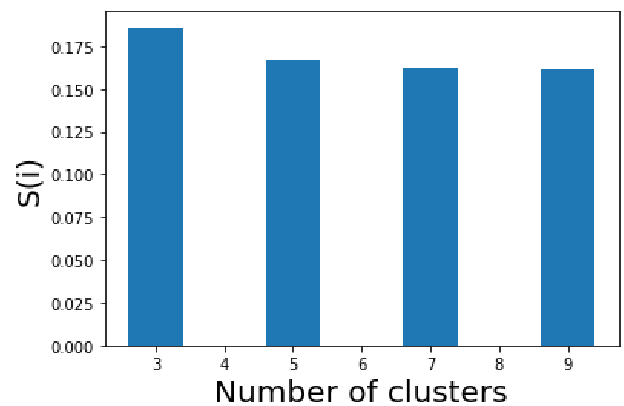
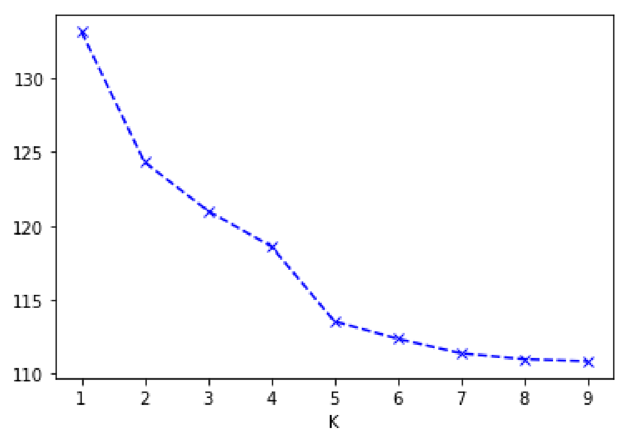
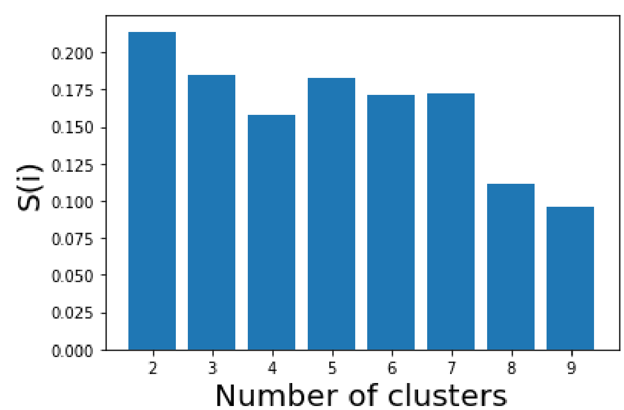
One obervation is the top players are around y=0(PC2) and x=-0.33(PC1). But we cant say something explisitly about a praticular top player as this is dense area. And least performing players are arround Green Area.

As we were not getting outliers by standard approch ,we decided to follow materialistic approch and ploted players wages and value. And results were as expected.

<table>
<tr><th>Value</th><th>Wage</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

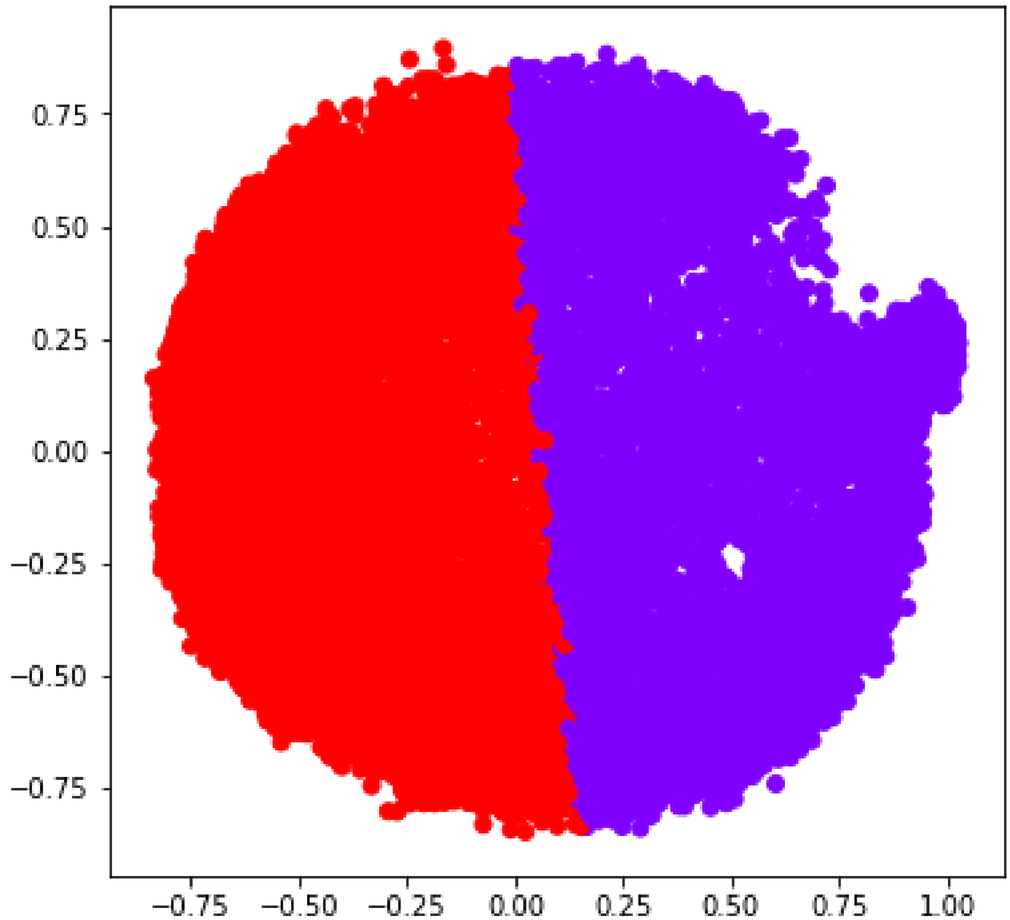
2 2.Analysis from K-means

If we chose k = 3,5,7 we get elbow at 3 and silhouette score is also higest at 3.This indicates apropriate number of clusters is 3.

<table>
<tr><th>Elbow</th><th>Silhouette</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

But when we iterate from k=[2,3. . . 9] we get above values. Here we see higest silhouette score at k=2 BUT elbow is at k=5. As silhouette score at k=5 is also not that bad k should be 5.

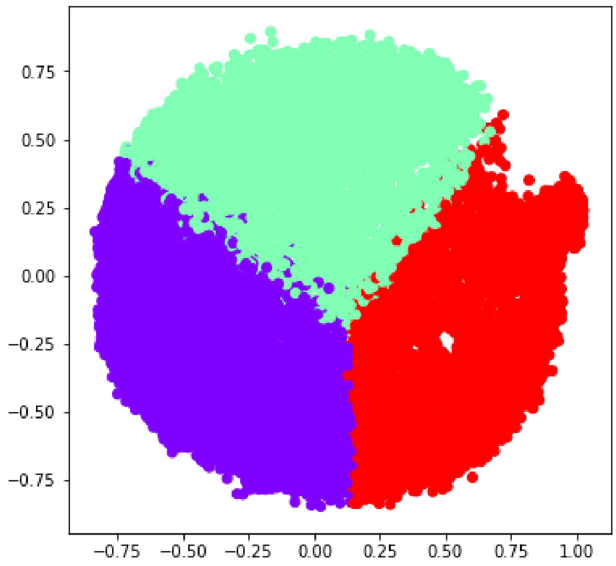
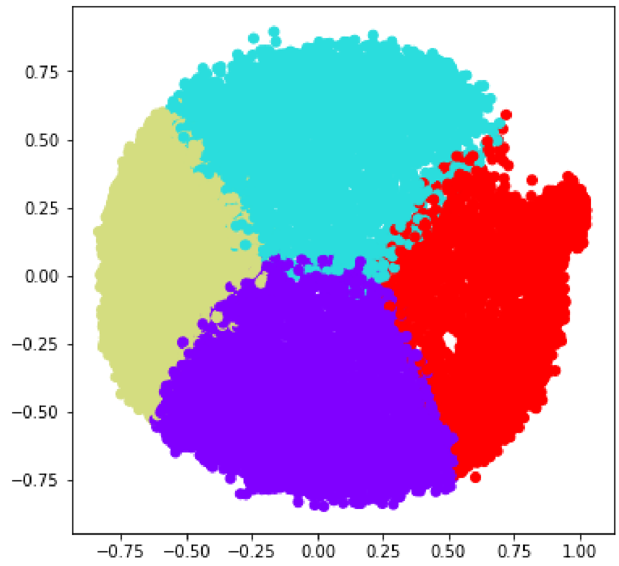
2.0.1 For K=2



For K=2 it has very clearly divided the data into 2 parts.As most of the good players were on the left side and not so good player were on right.This has divided the data into above avarage and below avarage.

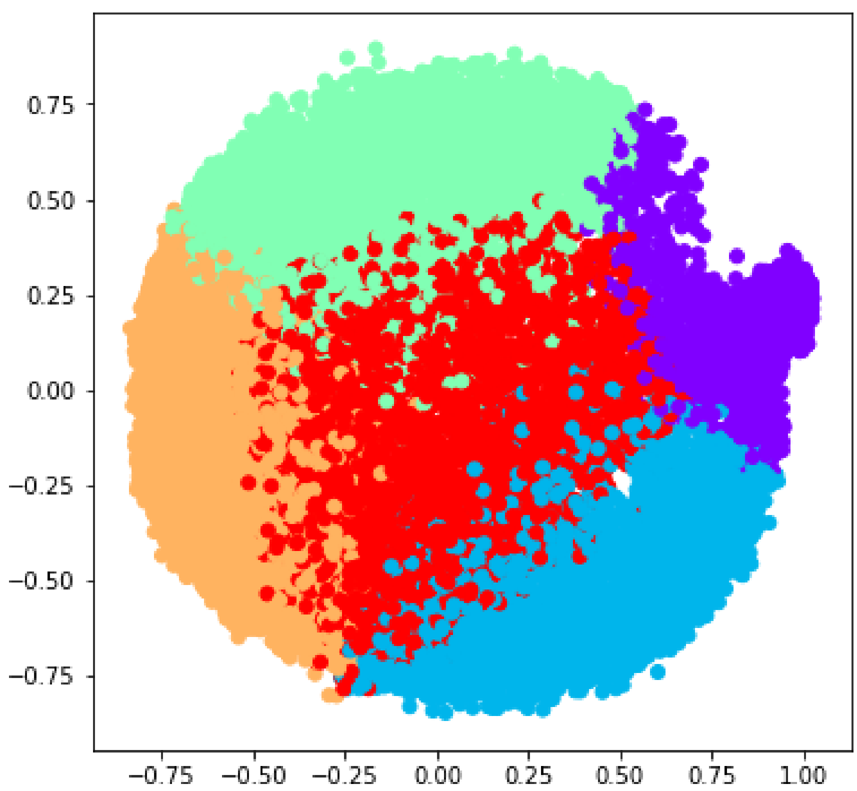
2.0.2 For K=3 and K=4

Both of these have clear decison boundries.For k=4 the clustring have lost the division at x~0 which was in k=2 and k=3.

k=3	k=4
	

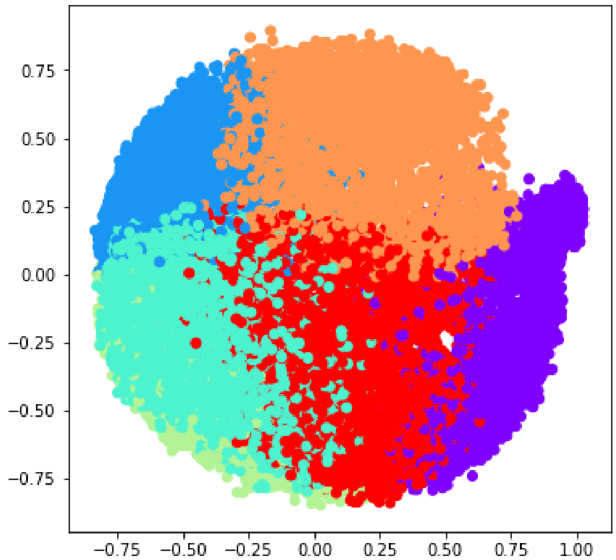
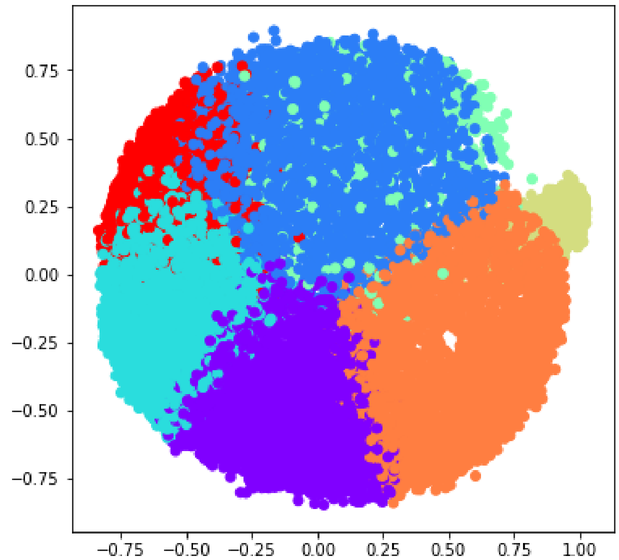
2.0.3 For K=5

For K=5 it has came up with a sperate cluster in between which looks like a extention to the pattren we observed in top players. with this the last rank playered are also properly clusterd in right.



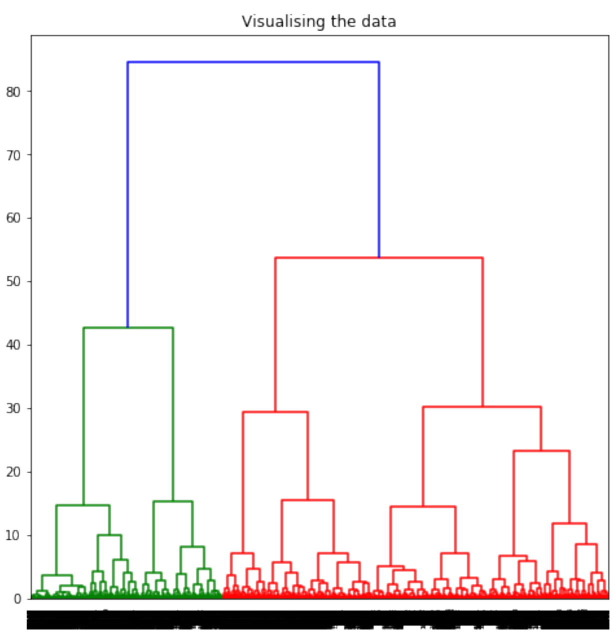
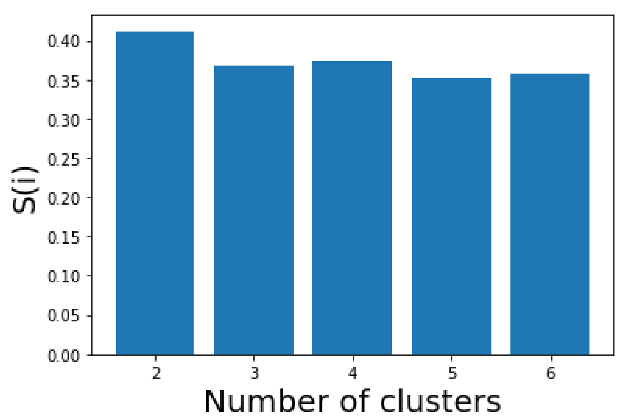
2.0.4 For K=6 and K=7

For k=6,7(Higher than 5) the new cluster formed have very less data and are not that significant. Also the clusters have merged ,epecially for k>=7.

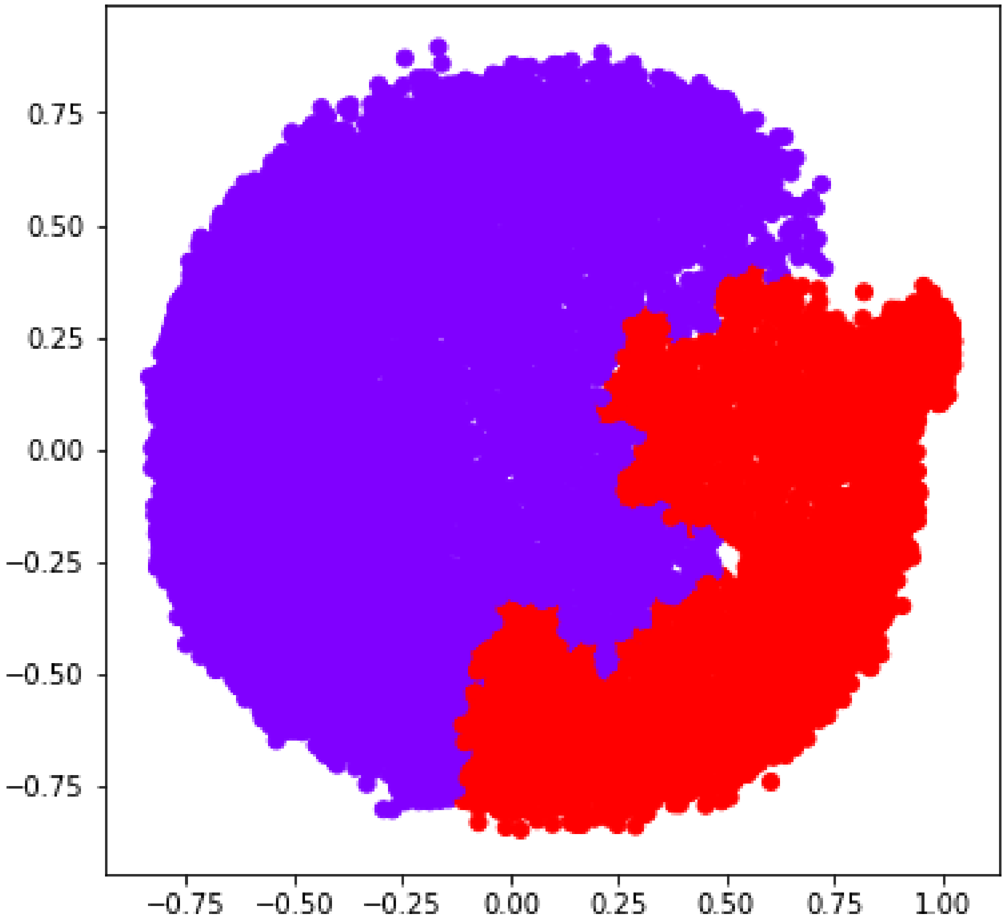
k=6	k=7
	

3 3.Hierarchical Clustering

3.1 Agglomerative(bottom-up strategy)

Dendogram	Sil. Score
	

Following dendogram suggests that we should have 2 clusters,which is also suggested by Silhouette Score.



4 4.DBSCAN

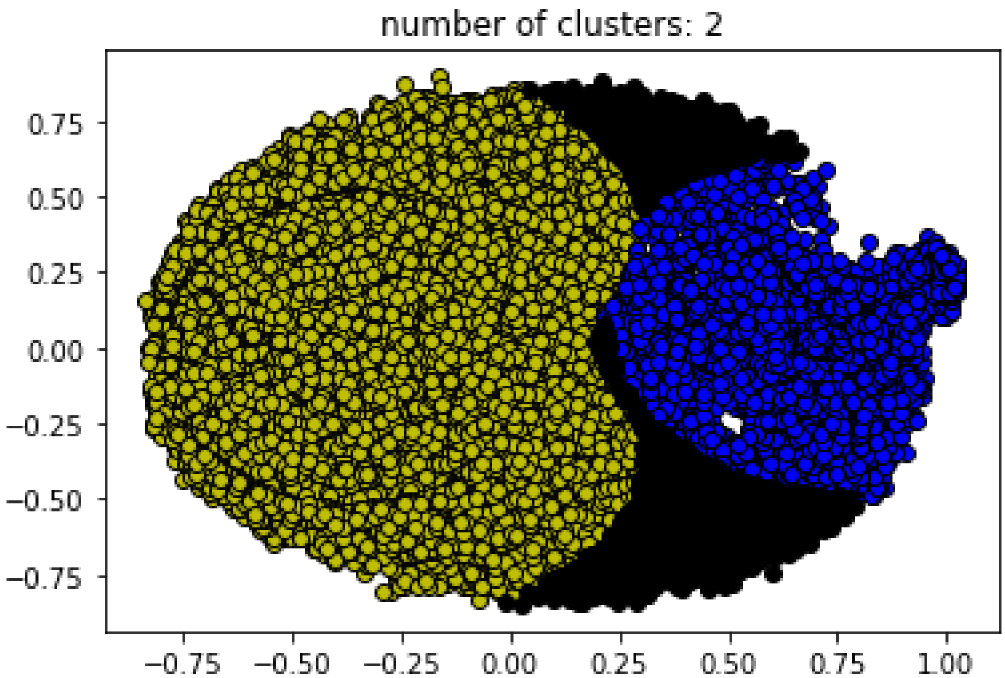
For calulating the epsilon and minPoints we first loop through a number of values for both of them and observed number of clusters formed , ideally they should nither be too high or low.

esp=[0.001,0.01,0.1,1,10,100] min_samples=[2,10,100,500,1000,2000,4000,8000]

From here we concluded that the values should be of order (0.1,500),

esp=[0.1,0.2,0.3,0.4,0.5] min_samples=[200,400,800,1600,3200,6400]

Then we looped arround these values and found that (0.4,3200) gives the higest Silhouette Score and number of cluster=2.



The cluster are well defined and there is no overlaping among them. As stated before low ranked players were in right and good players were in left. DBSCAN have also clustred according to overall above and blow avarage.