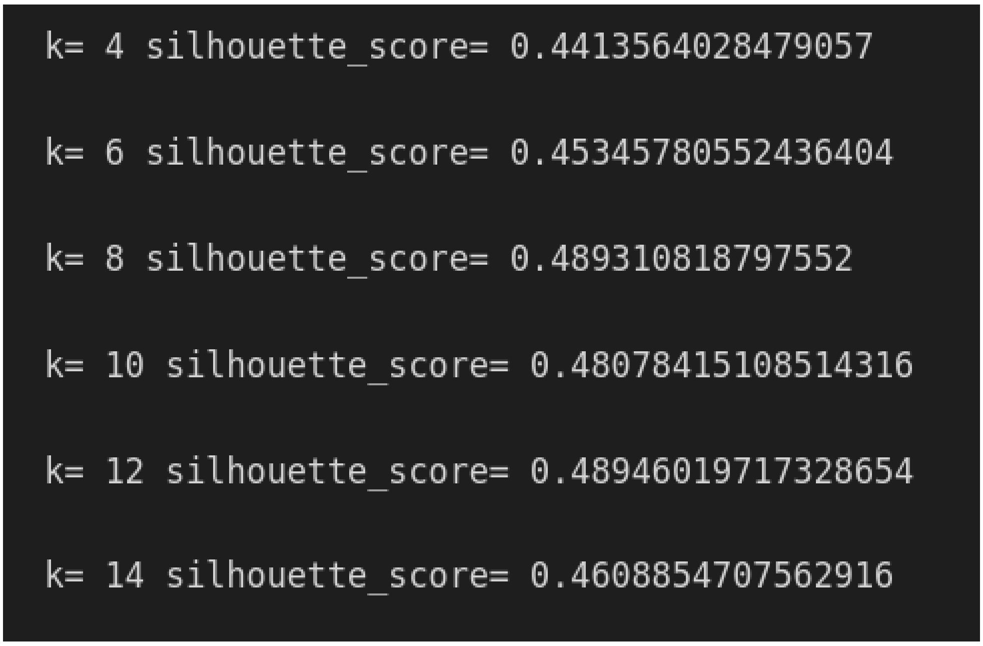
ASSIGNMENT 5: CLUSTERING
TEAM: DATA PIRATES

Q1: Kmeans Clustering:-

1.1: Silhouette score and elbow method

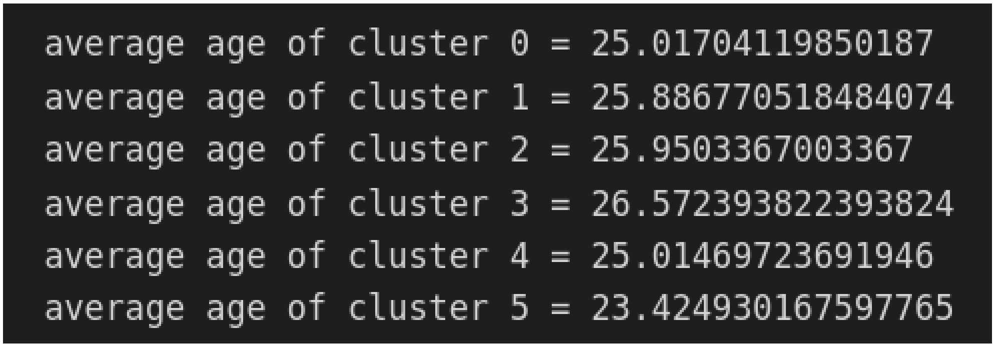


```
k= 4 silhouette_score= 0.4413564028479057

k= 6 silhouette_score= 0.45345780552436404

k= 8 silhouette_score= 0.489310818797552

k= 10 silhouette_score= 0.48078415108514316

k= 12 silhouette_score= 0.48946019717328654

k= 14 silhouette_score= 0.4608854707562916
```

Above are the values obtained for the Kmeans function we created. we can observe that after k=8. the score seems to settle down. So by applying elbow method here we can say that 8 is the optimal number of clusters for this method.

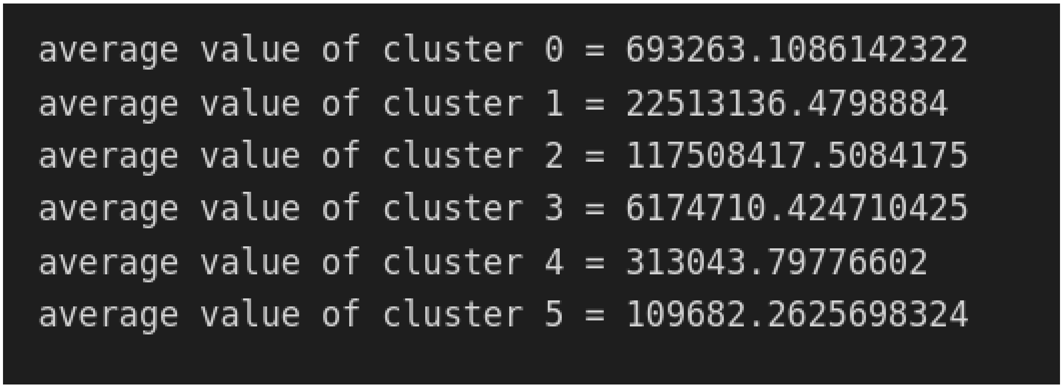
1.2: Finding the label of each Cluster

Average age:



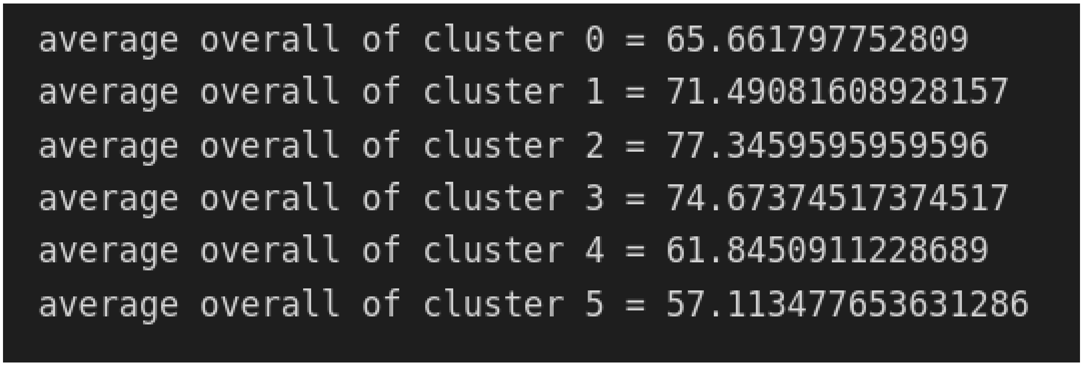
```
average age of cluster 0 = 25.01704119850187
average age of cluster 1 = 25.886770518484074
average age of cluster 2 = 25.9503367003367
average age of cluster 3 = 26.572393822393824
average age of cluster 4 = 25.01469723691946
average age of cluster 5 = 23.424930167597765
```

Average Value:



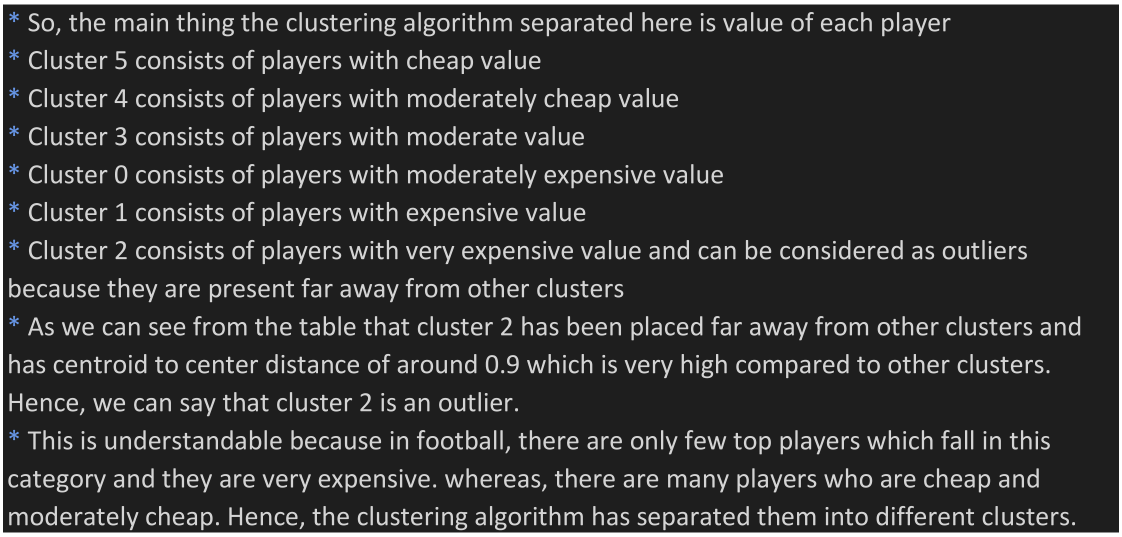
```
average value of cluster 0 = 693263.1086142322
average value of cluster 1 = 22513136.4798884
average value of cluster 2 = 117508417.5084175
average value of cluster 3 = 6174710.424710425
average value of cluster 4 = 313043.79776602
average value of cluster 5 = 109682.2625698324
```

Average Overall



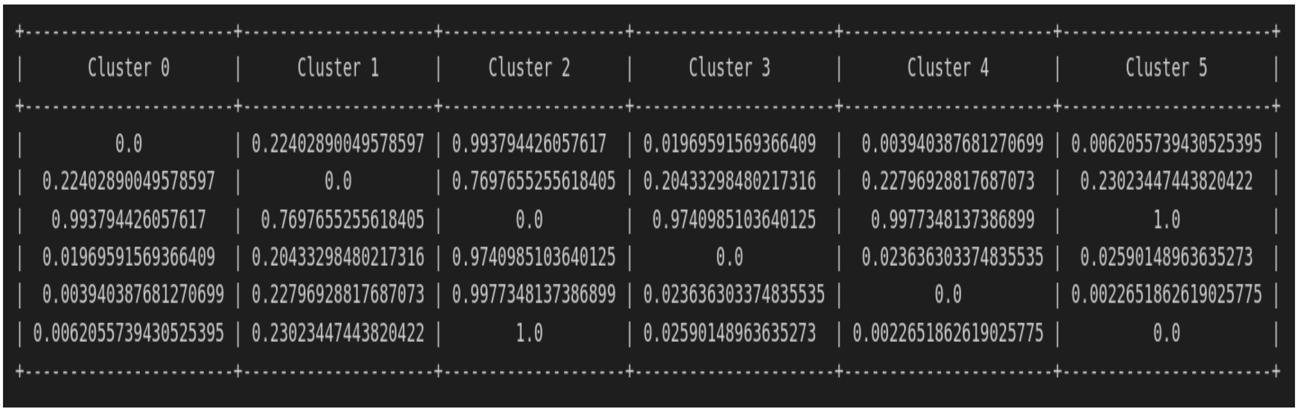
```
average overall of cluster 0 = 65.661797752809
average overall of cluster 1 = 71.49081608928157
average overall of cluster 2 = 77.3459595959596
average overall of cluster 3 = 74.67374517374517
average overall of cluster 4 = 61.8450911228689
average overall of cluster 5 = 57.113477653631286
```

We considered 3 features as the vectors to find clusters in the data namely - age, overall and value of a player.
From the above figures we can conclude the following:-



* So, the main thing the clustering algorithm separated here is value of each player
* Cluster 5 consists of players with cheap value
* Cluster 4 consists of players with moderately cheap value
* Cluster 3 consists of players with moderate value
* Cluster 0 consists of players with moderately expensive value
* Cluster 1 consists of players with expensive value
* Cluster 2 consists of players with very expensive value and can be considered as outliers because they are present far away from other clusters
* As we can see from the table that cluster 2 has been placed far away from other clusters and has centroid to center distance of around 0.9 which is very high compared to other clusters. Hence, we can say that cluster 2 is an outlier.
* This is understandable because in football, there are only few top players which fall in this category and they are very expensive. whereas, there are many players who are cheap and moderately cheap. Hence, the clustering algorithm has separated them into different clusters.

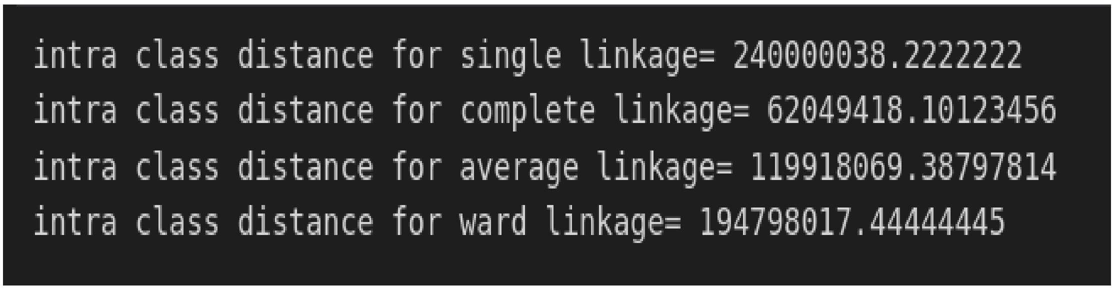
The table supporting the statement above:-



```
+-----------------------+---------------------+--------------------+----------------------+-----------------------+-----------------------+
|       Cluster 0       |      Cluster 1      |     Cluster 2      |      Cluster 3       |       Cluster 4       |       Cluster 5       |
+-----------------------+---------------------+--------------------+----------------------+-----------------------+-----------------------+
|          0.0          | 0.22402890049578597 | 0.993794426057617  | 0.01969591569366409  |  0.003940387681270699 | 0.0062055739430525395 |
|  0.22402890049578597  |         0.0         | 0.7697655255618405 | 0.20433298480217316  |  0.22796928817687073  |  0.23023447443820422  |
|   0.993794426057617   |  0.7697655255618405 |        0.0         |  0.9740985103640125  |   0.9977348137386899  |          1.0          |
|  0.01969591569366409  | 0.20433298480217316 | 0.9740985103640125 |         0.0          |  0.023636303374835535 |  0.02590148963635273  |
|  0.003940387681270699 | 0.22796928817687073 | 0.9977348137386899 | 0.023636303374835535 |          0.0          | 0.0022651862619025775 |
| 0.0062055739430525395 | 0.23023447443820422 |        1.0         | 0.02590148963635273  | 0.0022651862619025775 |          0.0          |
+-----------------------+---------------------+--------------------+----------------------+-----------------------+-----------------------+
```

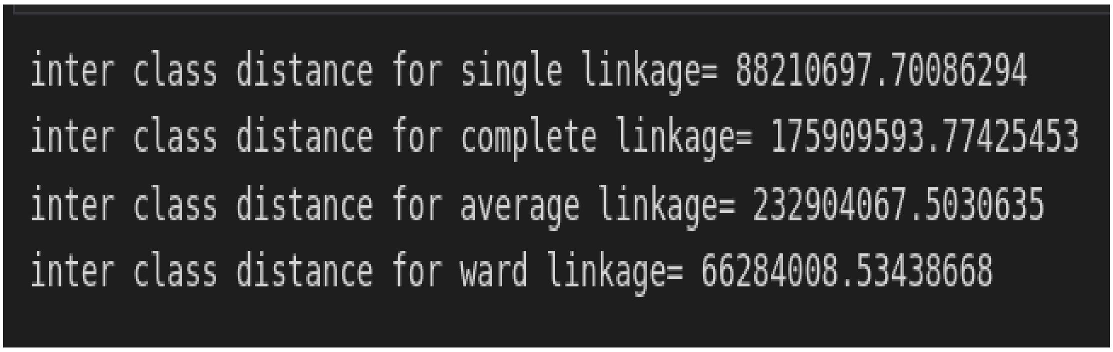
Q2: Agglomerative method:-

The average intra class distance obtained through different hierarchical clustering by having different linkage methods is given below



```
intra class distance for single linkage= 240000038.2222222
intra class distance for complete linkage= 62049418.10123456
intra class distance for average linkage= 119918069.38797814
intra class distance for ward linkage= 194798017.44444445
```

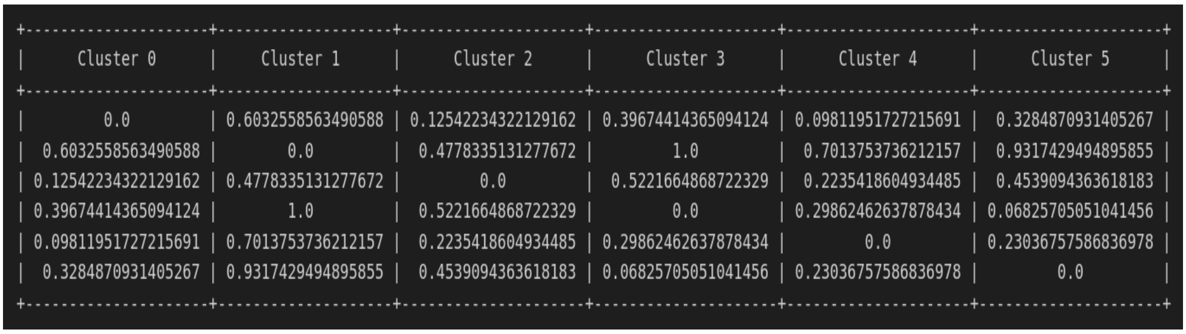
The average inter class distance obtained through different hierarchical clustering by having different linkage methods is given below



```
inter class distance for single linkage= 88210697.70086294
inter class distance for complete linkage= 175909593.77425453
inter class distance for average linkage= 232904067.5030635
inter class distance for ward linkage= 66284008.53438668
```

Complete linkage seem to produce best result here as it has low intra class distance and high inter class distance.

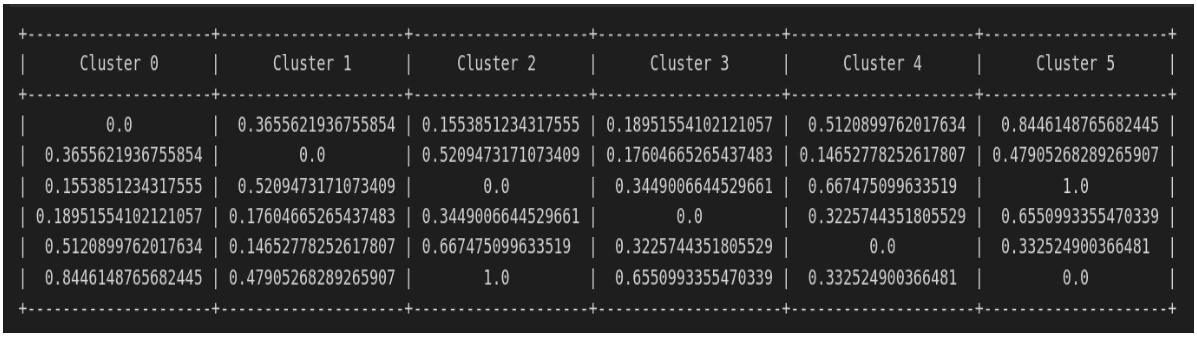
The distance matrix between clusters for single linkage method is given below



```
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|      Cluster 0      |      Cluster 1      |      Cluster 2      |      Cluster 3      |      Cluster 4      |      Cluster 5      |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|         0.0         |  0.6032558563490588 | 0.12542234322129162 | 0.39674414365094124 | 0.09811951727215691 |  0.3284870931405267 |
|  0.6032558563490588 |         0.0         |  0.4778335131277672 |         1.0         |  0.7013753736212157 |  0.9317429494895855 |
| 0.12542234322129162 |  0.4778335131277672 |         0.0         |  0.5221664868722329 |  0.2235418604934485 |  0.4539094363618183 |
| 0.39674414365094124 |         1.0         |  0.5221664868722329 |         0.0         | 0.29862462637878434 | 0.06825705051041456 |
| 0.09811951727215691 |  0.7013753736212157 |  0.2235418604934485 | 0.29862462637878434 |         0.0         | 0.23036757586836978 |
|  0.3284870931405267 |  0.9317429494895855 |  0.4539094363618183 | 0.06825705051041456 | 0.23036757586836978 |         0.0         |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
```

From above we can see that cluster 1 seem to be far away from other clusters which makes it as cluster of outliers.

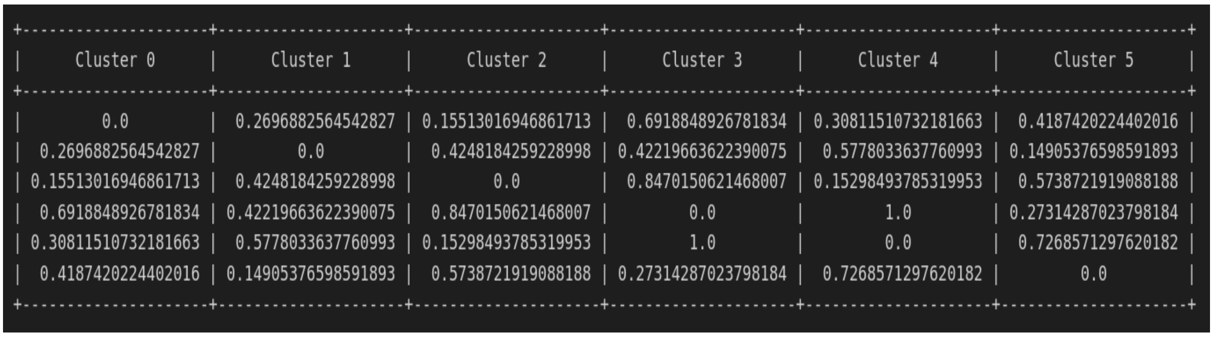
The distance matrix between clusters for complete linkage method is given below



```
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|      Cluster 0      |      Cluster 1      |      Cluster 2      |      Cluster 3      |      Cluster 4      |      Cluster 5      |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|         0.0         |  0.3655621936755854 |  0.1553851234317555 | 0.18951554102121057 |  0.5120899762017634 |  0.8446148765682445 |
|  0.3655621936755854 |         0.0         |  0.5209473171073409 | 0.17604665265437483 | 0.14652778252617807 | 0.47905268289265907 |
|  0.1553851234317555 |  0.5209473171073409 |         0.0         |  0.3449006644529661 |  0.667475099633519  |         1.0         |
| 0.18951554102121057 | 0.17604665265437483 |  0.3449006644529661 |         0.0         |  0.3225744351805529 |  0.6550993355470339 |
|  0.5120899762017634 | 0.14652778252617807 |  0.667475099633519  |  0.3225744351805529 |         0.0         |  0.332524900366481  |
|  0.8446148765682445 | 0.47905268289265907 |         1.0         |  0.6550993355470339 |  0.332524900366481  |         0.0         |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
```

From above we can see that cluster 5 seem to be far away from other clusters which makes it as cluster of outliers.

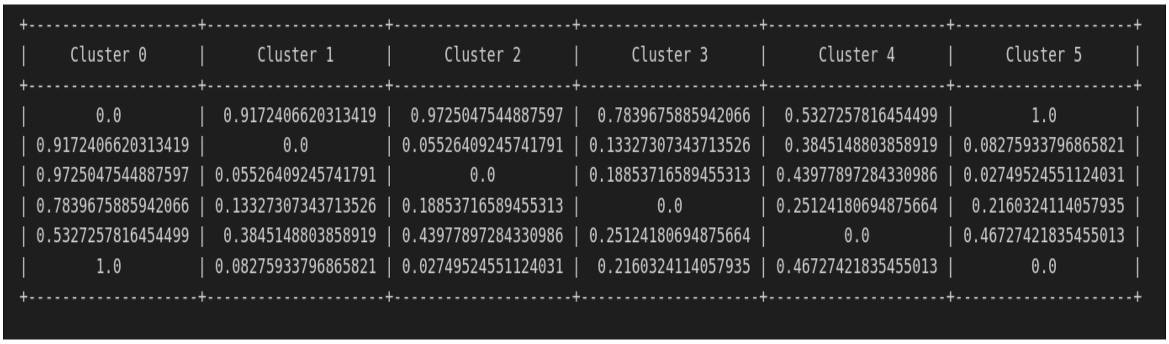
The distance matrix between clusters for Average linkage method is given below



```
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|      Cluster 0      |      Cluster 1      |      Cluster 2      |      Cluster 3      |      Cluster 4      |      Cluster 5      |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
|         0.0         |  0.2696882564542827 | 0.15513016946861713 |  0.6918848926781834 | 0.30811510732181663 |  0.4187420224402016 |
|  0.2696882564542827 |         0.0         |  0.4248184259228998 | 0.42219663622390075 |  0.5778033637760993 | 0.14905376598591893 |
| 0.15513016946861713 |  0.4248184259228998 |         0.0         |  0.8470150621468007 | 0.15298493785319953 |  0.5738721919088188 |
|  0.6918848926781834 | 0.42219663622390075 |  0.8470150621468007 |         0.0         |         1.0         | 0.27314287023798184 |
| 0.30811510732181663 |  0.5778033637760993 | 0.15298493785319953 |         1.0         |         0.0         |  0.7268571297620182 |
|  0.4187420224402016 | 0.14905376598591893 |  0.5738721919088188 | 0.27314287023798184 |  0.7268571297620182 |         0.0         |
+---------------------+---------------------+---------------------+---------------------+---------------------+---------------------+
```

From above we can see that cluster 3 seem to be far away from other clusters which makes it as cluster of outliers.

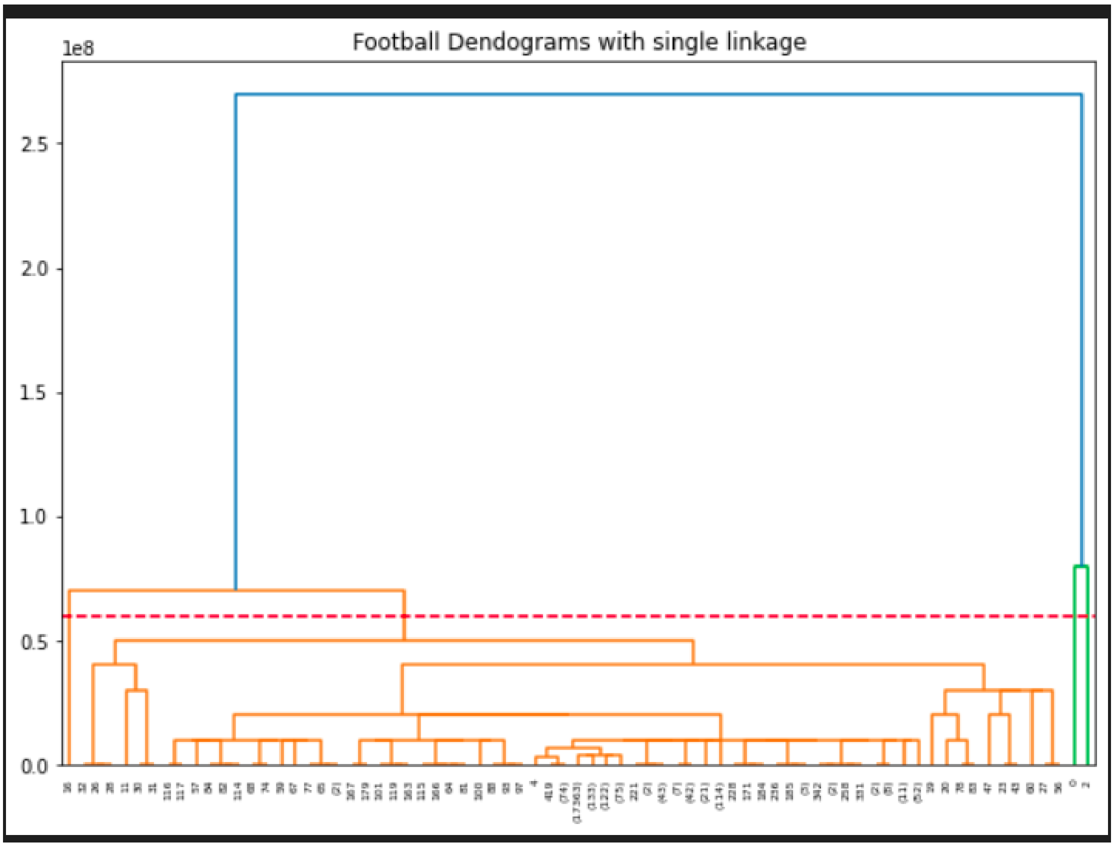
The distance matrix between clusters for mean distance method is given below

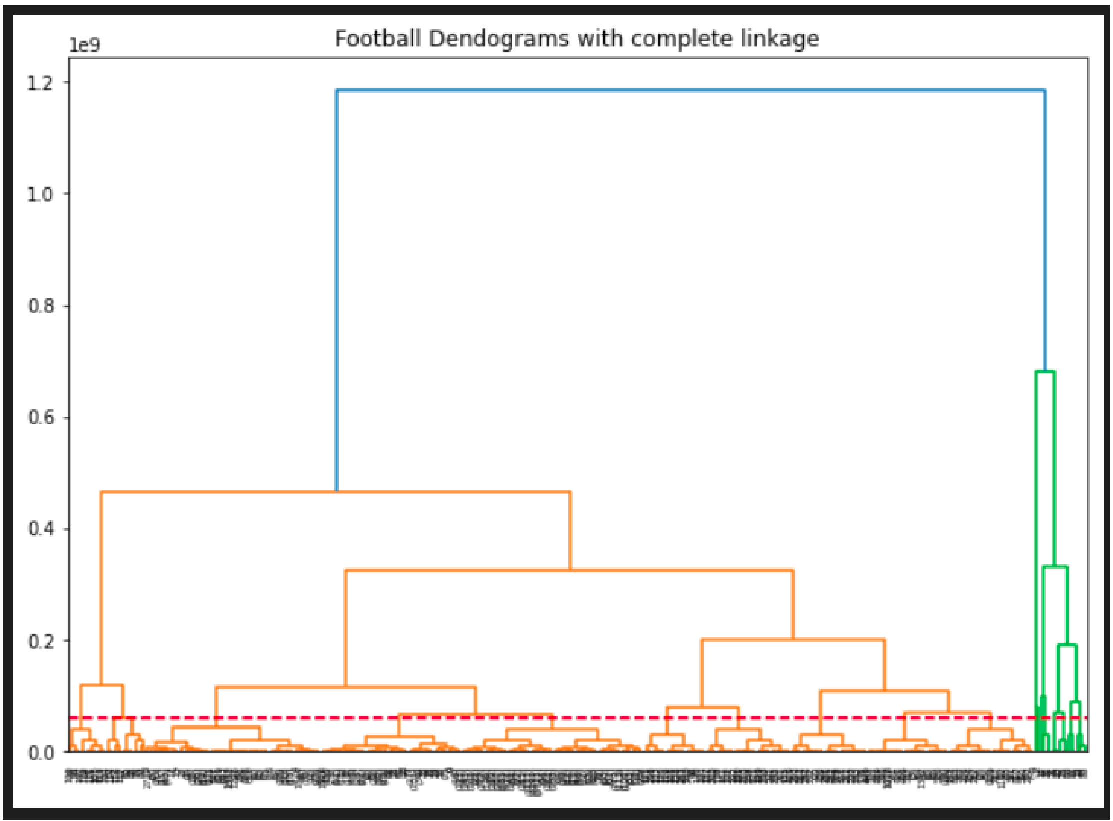


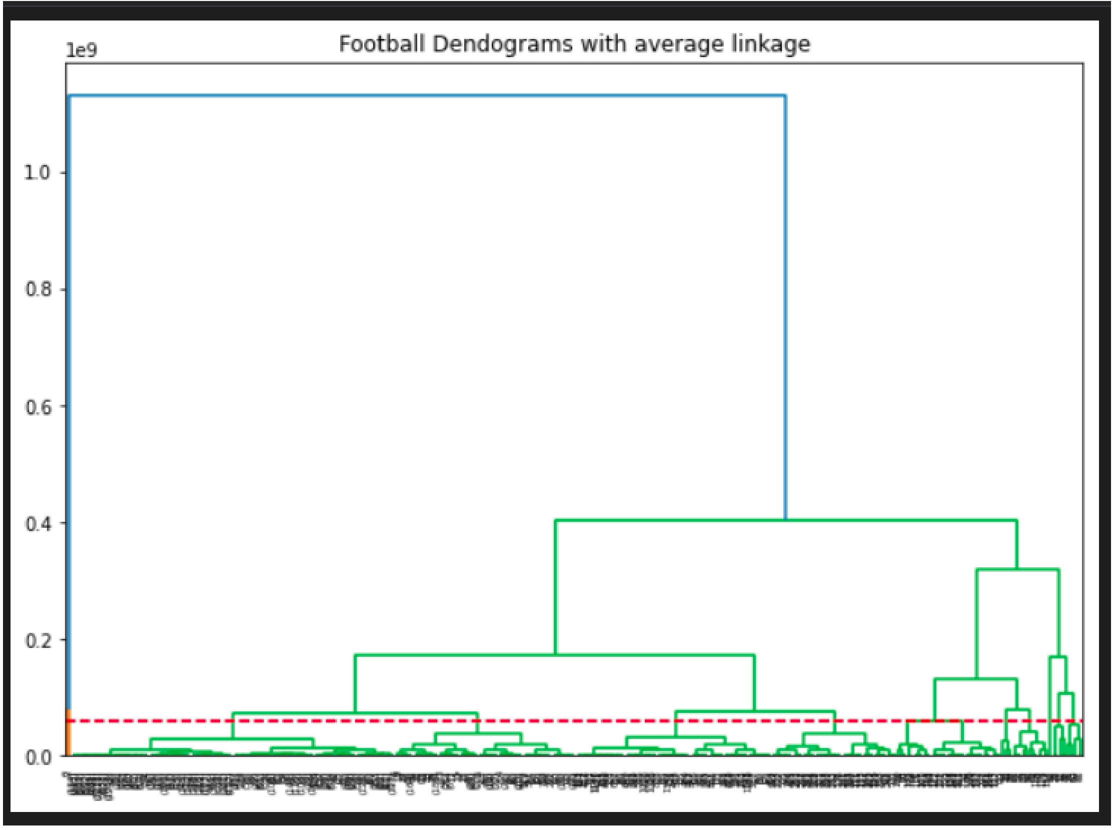
```
+--------------------+--------------------+--------------------+--------------------+--------------------+---------------------+
|     Cluster 0      |     Cluster 1      |     Cluster 2      |     Cluster 3      |     Cluster 4      |      Cluster 5      |
+--------------------+--------------------+--------------------+--------------------+--------------------+---------------------+
|        0.0         | 0.9172406620313419 | 0.9725047544887597 | 0.7839675885942066 | 0.5327257816454499 |         1.0         |
| 0.9172406620313419 |        0.0         | 0.05526409245741791 | 0.13327307343713526 | 0.3845148803858919 | 0.08275933796865821 |
| 0.9725047544887597 | 0.05526409245741791 |        0.0         | 0.18853716589455313 | 0.43977897284330986 | 0.02749524551124031 |
| 0.7839675885942066 | 0.13327307343713526 | 0.18853716589455313 |        0.0         | 0.25124180694875664 |  0.2160324114057935 |
| 0.5327257816454499 | 0.3845148803858919 | 0.43977897284330986 | 0.25124180694875664 |        0.0         | 0.46727421835455013 |
|        1.0         | 0.08275933796865821 | 0.02749524551124031 | 0.2160324114057935 | 0.46727421835455013 |         0.0         |
+--------------------+--------------------+--------------------+--------------------+--------------------+---------------------+
```

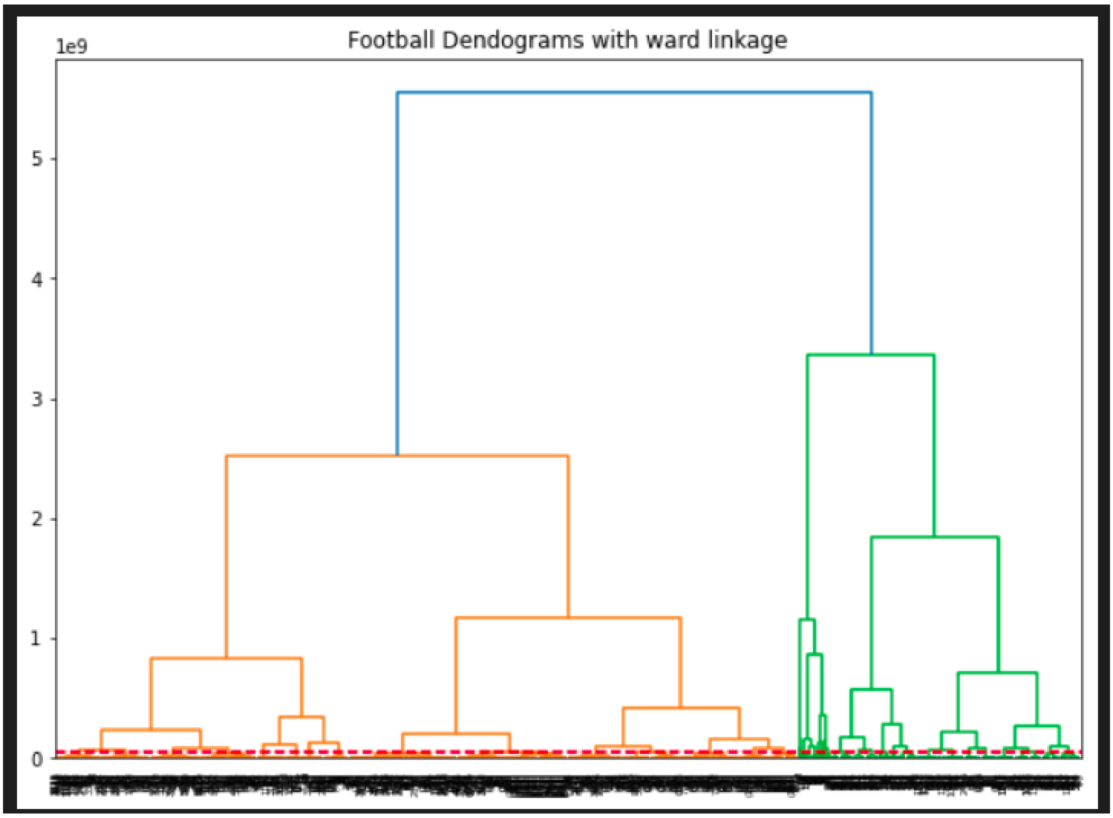
From above we can see that cluster 0 seem to be far away from other clusters which makes it as cluster of outliers.

The Dendograms for different linkage methods are given below.




Football Dendograms with complete linkage
1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0


Football Dendograms with average linkage
1e9
1.0
0.8
0.6
0.4
0.2
0.0


Football Dendograms with ward linkage
1e9
5
4
3
2
1
0