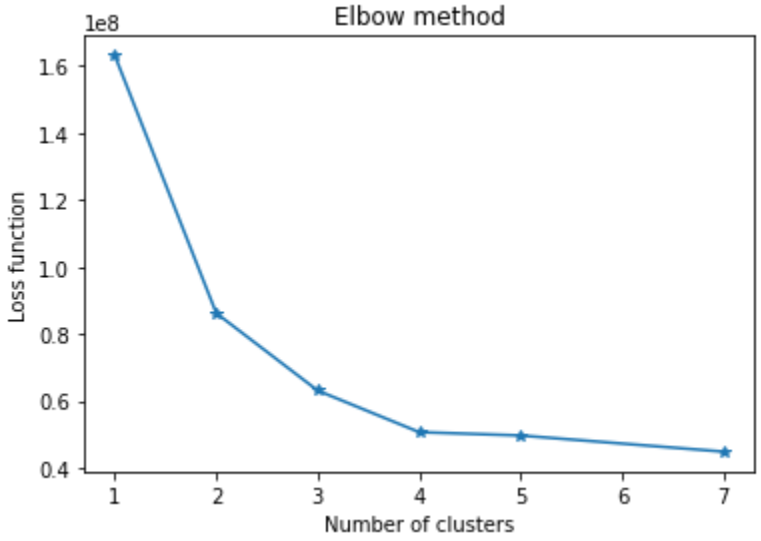
Data Analytics

Assignment 5 Report/Analysis

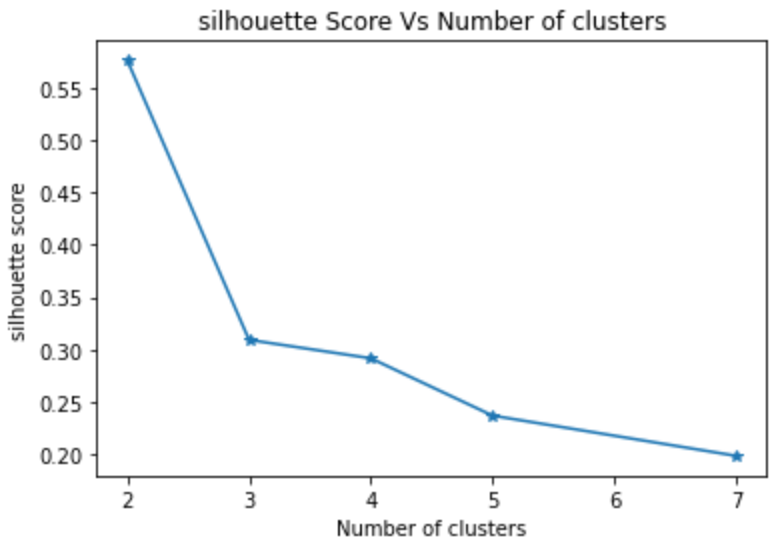
P Shiridi Kumar
Karnati Jahnavi

Features used for building the models:
["ID","Preferred Foot","Weak Foot","Skill Moves","Work Rate",'Body Type', 'Real Face', 'Position', 'Height', 'Weight',
'HeadingAccuracy', 'ShortPassing', 'Volleys', 'Dribbling',
'Curve', 'FKAccuracy', 'LongPassing', 'BallControl', 'Acceleration',
'SprintSpeed', 'Agility', 'Reactions', 'Balance', 'ShotPower',
'Jumping', 'Stamina', 'Strength', 'LongShots', 'Aggression',
'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Composure',
'Marking', 'StandingTackle', 'SlidingTackle', 'GKDiving', 'GKHandling',
'GKKicking', 'GKPositioning', 'GKReflexes']

KMeans:
We have experimented our KMeans model on by variating different number of clusters [2,3,4,5,7]. And we found that number of clusters =3 would be optimal as it has better metrics scores in terms of both the loss and silhouette coefficient .
We have found a sharp bend near number of cluster =3 and number of cluster =4 in elbow plot and hence one of them could be a better number of clusters.



We found that number of clusters=3 has a better silhouette score compared to others



Avg interclass and intraclass distances: We have analysed the avg interclass distance and intraclass distances:

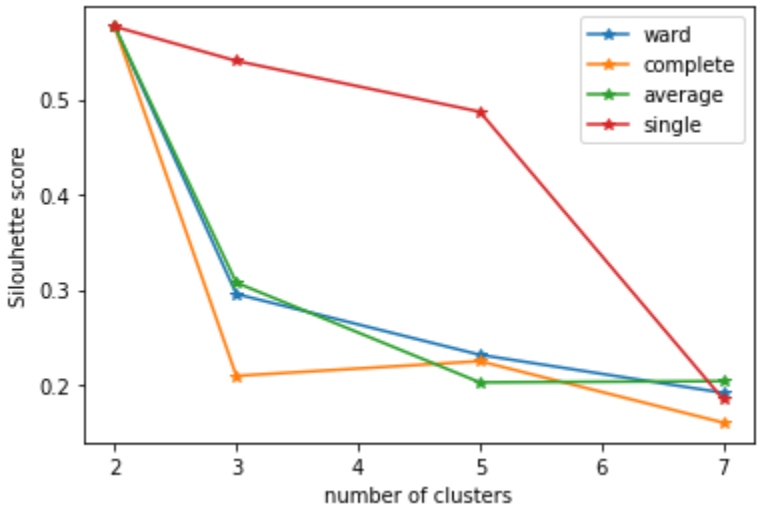
	Avg_Inter_Class_distance	Avg_Intra_class_distance
2	10702.167591	3584.407436
3	7862.796084	3117.255582
4	6442.104015	2670.024408
5	7383.599960	2351.057721
7	7129.021764	2107.184417

Ideally we can infer that for lower number of clusters the ratio of intra class distance and interclass distance is good and notably number of clusters =3 has good loss term as well (can infer from elbow method)

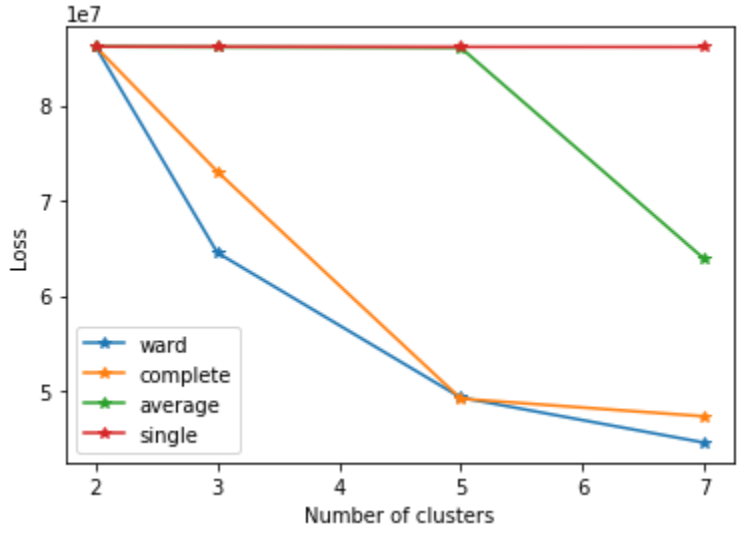
Attribute with highest variability across clusters in Kmeans : We have found out that the attribute GKReflexes has the highest variability across clusters(for all k=2,3,4,5,7) and this attribute might be possibly used as the name for the cluster based on the cluster mean on the attribute GKReflexes:
Naming for three clusters:
Ultra high GKReflexes: 66.1141868
Fairly High : ~ 11
High : ~ 10

Aggglorомative Clustering:

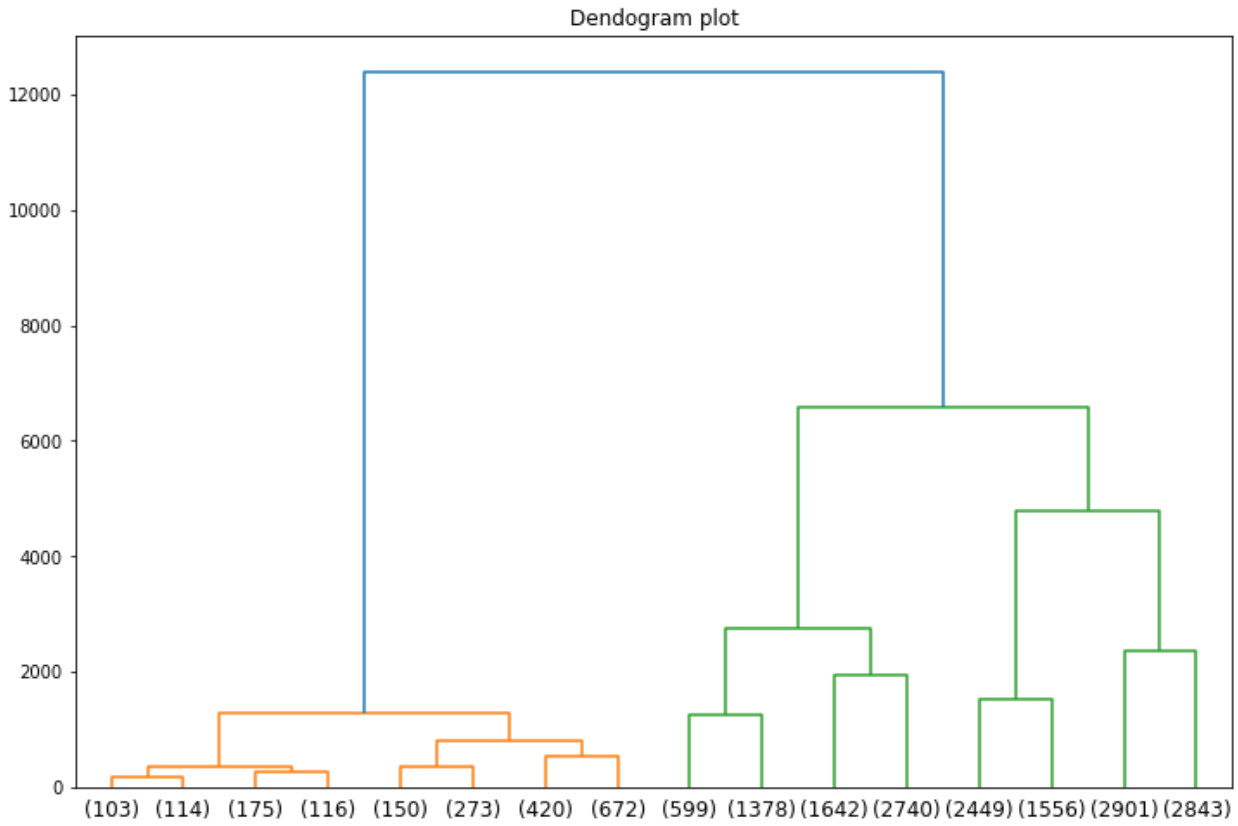
Silhouette scores for various number of clusters and various distance metrics:



Elbow plot for various number of clusters and various distance metrics:



Dendogram Plot:



Comparison and conclusion :

Metrics for kmeans:

	Loss	silhouette_score
2	8.617206e+07	0.576408
3	6.304597e+07	0.309067
4	5.063483e+07	0.291449
5	4.963669e+07	0.236594
7	4.481001e+07	0.198122

Metrics for Aggloromative:

	ward :Silhouette score	complete :Silhouette score	average :Silhouette score	single :Silhouette score	ward : loss	complete : loss	average : loss	single : loss
2	0.576408	0.576408	0.576408	0.576408	8.617206e+07	8.617206e+07	8.617206e+07	8.617206e+07
3	0.295300	0.209299	0.307268	0.540317	6.452812e+07	7.304440e+07	8.612170e+07	8.616310e+07
5	0.231453	0.224705	0.202457	0.486909	4.930565e+07	4.918783e+07	8.605627e+07	8.613800e+07
7	0.191506	0.160149	0.203833	0.184546	4.457570e+07	4.733938e+07	6.386484e+07	8.612639e+07

As we can infer from both the above tables kmeans slightly has higher Sihouette score which means there is good tradefoof between interclass and intraclass distance . and In aggloromative clustering we can infer that interms of silhouette score single link performs well but its loss functions is high and ward distance measure which is also mean distance measure is good in terms of loss .

Intra class and inter class variance tradeoff betwee both methods:
Kmeans:

	Avg_Inter_Class_distance	Avg_Intra_class_distance
2	10702.167591	3584.407436
3	7862.796084	3117.255582
4	6442.104015	2670.024408
5	7383.599960	2351.057721
7	7129.021764	2107.184417

Aggloromative (Mean distance):

	Avg_Inter_class_distance	Avg_Intra_class_distance
2	7134.778394	3584.407436
3	5851.502543	3154.997521
5	4720.965860	2610.468256
7	3913.215744	2403.712757

Since KMeans has better inter class and intra class distances it makes more meaningful clusters as the cluster distances will be high and the points within the cluster will be close. And Kmeans also has better Silhouette coefficient as shown in the above plots.

Failed attempts:
Tried to normalize all values between 0 and 1 but its actually deceased the silhouette score , so only normalized height and weight attributes .