TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Introduction to Data Science

Capstone Project Name: FIFA 2021 Clustering
Assistant Lecturer: PhD. Muriel Visani
Group number: 14

Nguyen Nho Trung – 20204894

Nguyen Thi Thu Giang - 20194750

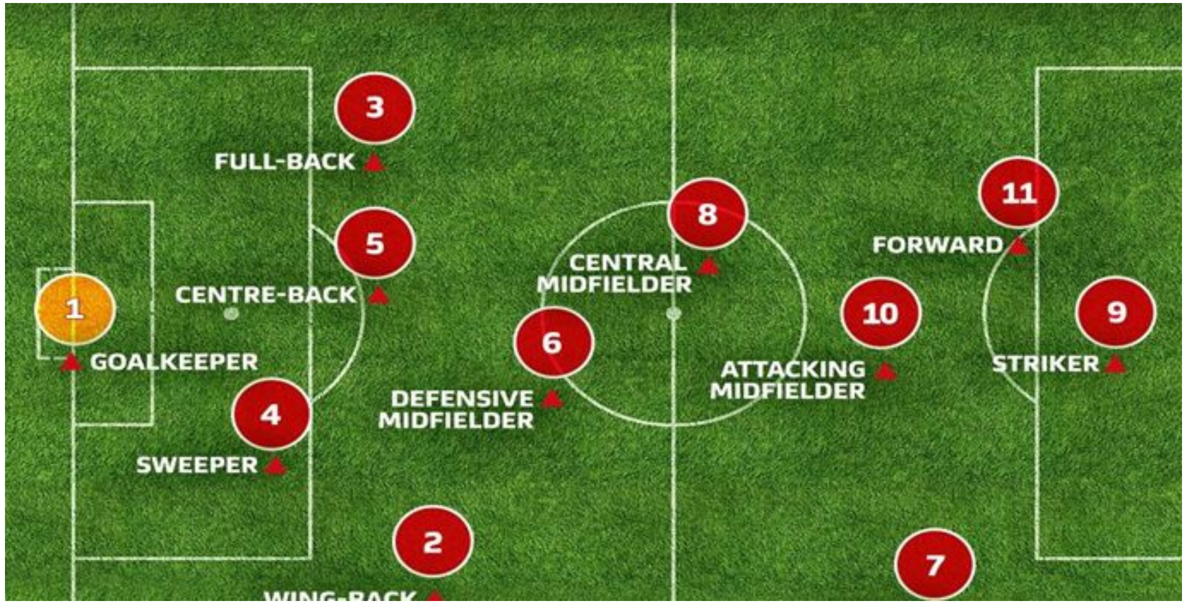
Tong Thi Thu Anh – 20194728

Introduction

Our subject for the Capstone is applying Machine Learning to cluster football players. In detail, our project aims to look at the class of young players and cluster them into groups based on their attributes. We will filter and consider the players who are promised based on their potential attributes.

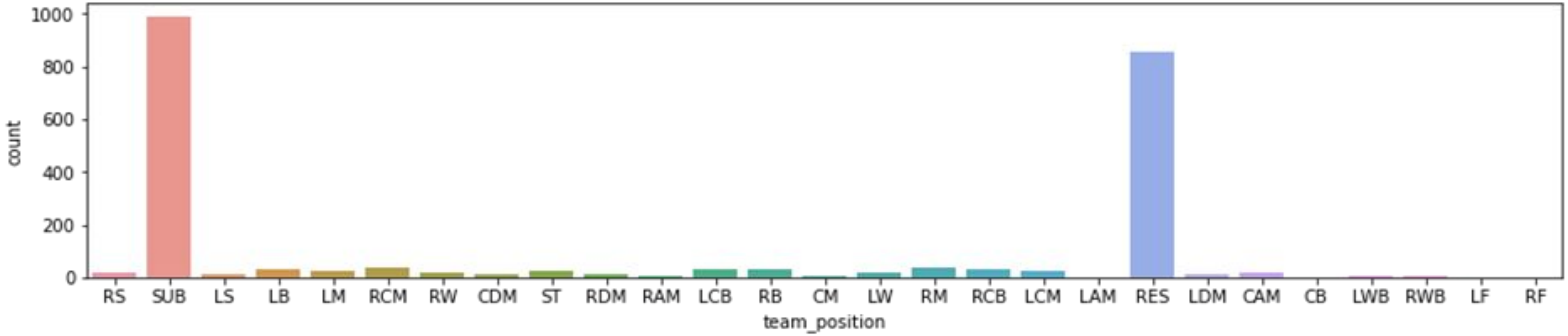
DATASET

- Name
- Age
- Height
- Overall: Rated between 1-99
- Potential: Rated between 1-99
- Work Rate: Effort the player puts in the game
- Team Position
- Playing Skills: Rated between 1-99
- Positional Skills: Rating of a player when play at a specified position rated between 1-99

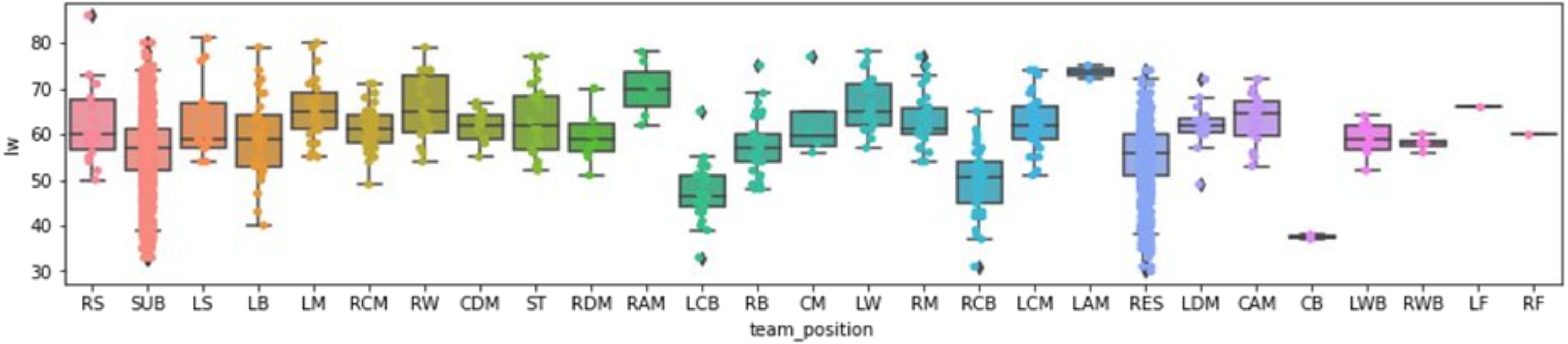


Exploring Data Analysis

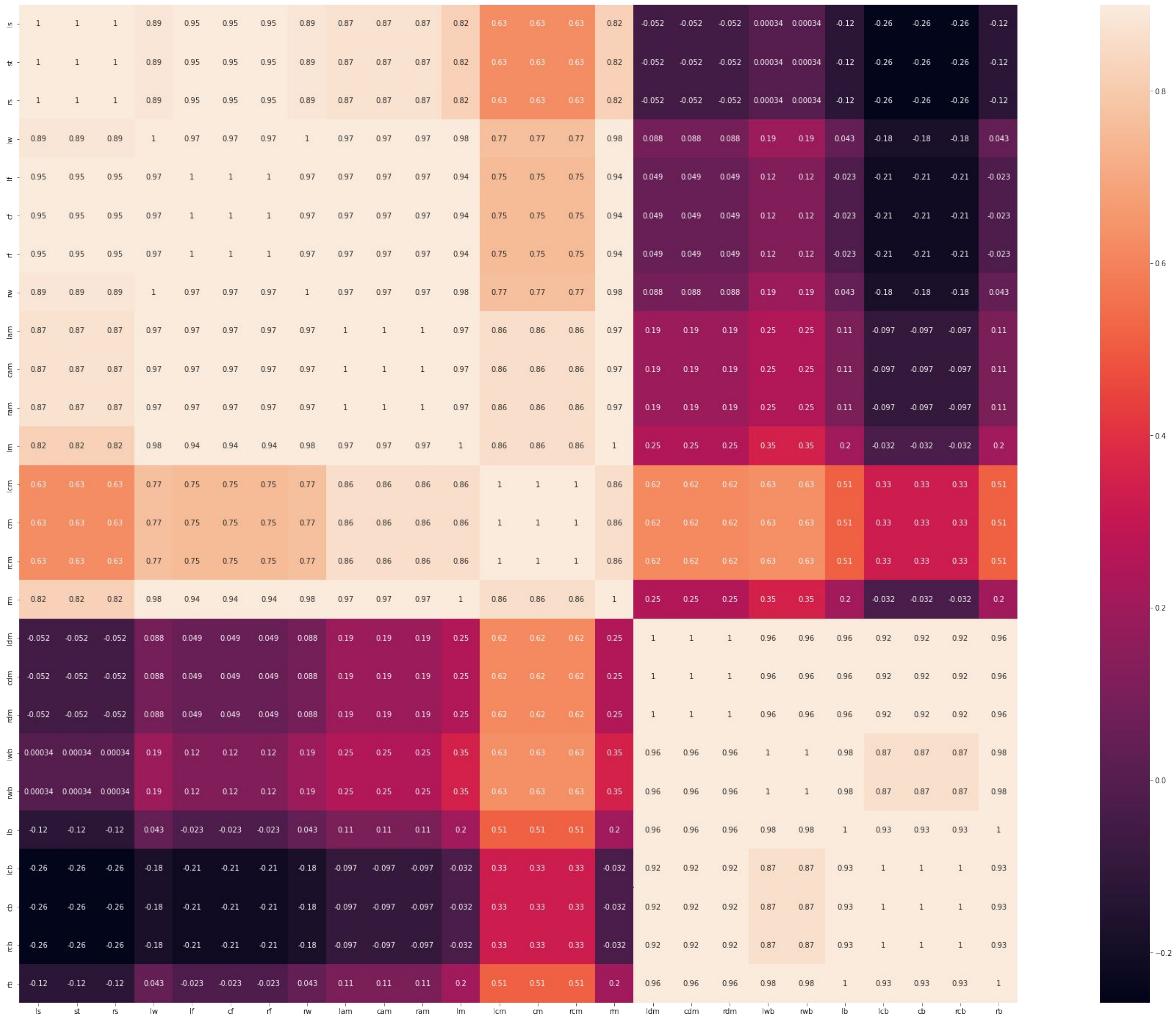
Plot the number of players in each position:



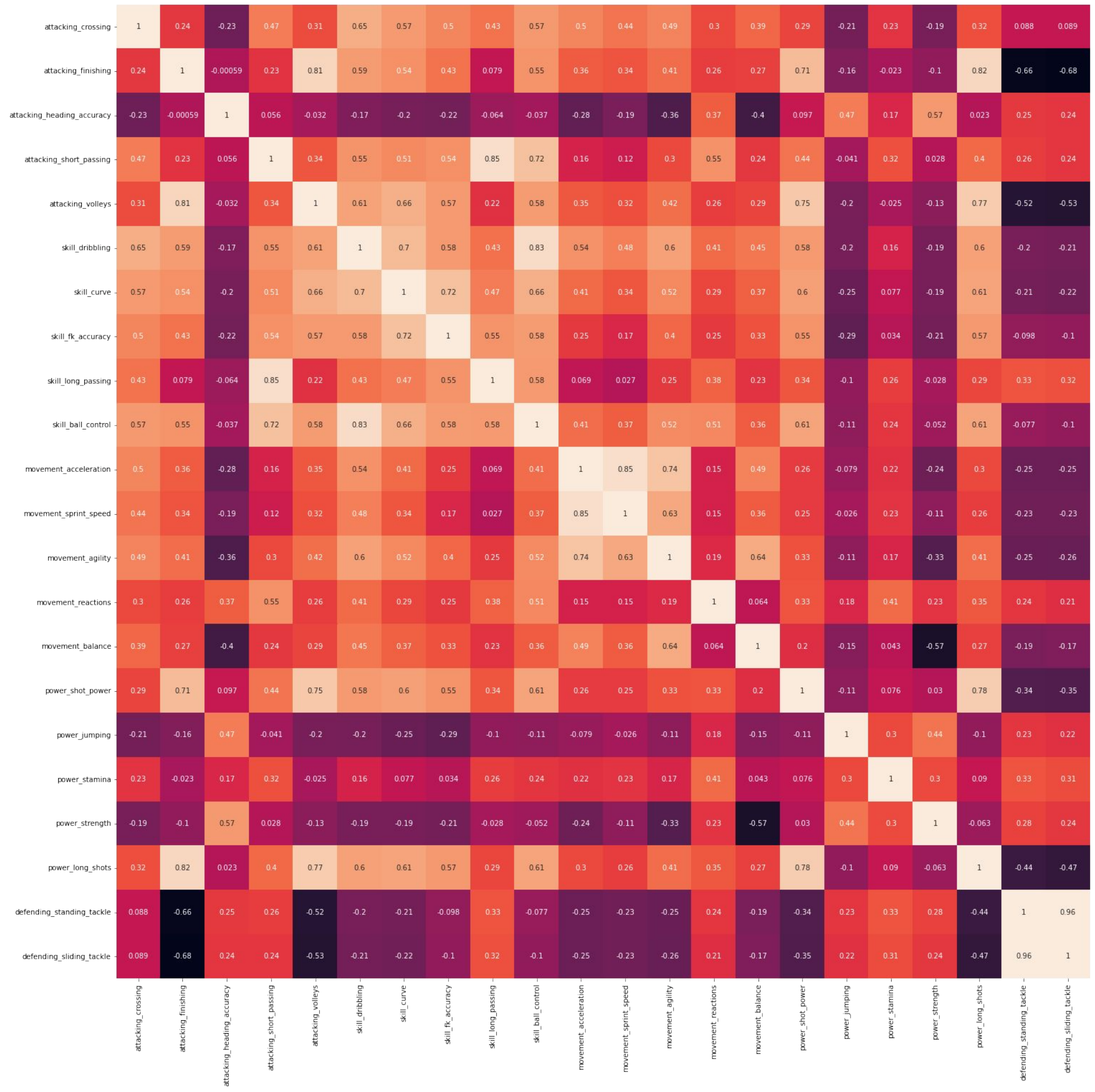
Plot the relationship between team positions and positional skills.



Correlation between the position and playing skills features



	ls	st	rs	lw	lf	cf	rf	rw	lam	cam	ram	lm	lcm	cm	rcm	rm	ldm	cdm	rdm	lwb	rwb	lb	lcb	cb	rcb	rb
ls	1	1	1	0.89	0.95	0.95	0.95	0.89	0.87	0.87	0.87	0.82	0.63	0.63	0.63	0.82	-0.052	-0.052	-0.052	0.00034	0.00034	-0.12	-0.26	-0.26	-0.26	-0.12
st	1	1	1	0.89	0.95	0.95	0.95	0.89	0.87	0.87	0.87	0.82	0.63	0.63	0.63	0.82	-0.052	-0.052	-0.052	0.00034	0.00034	-0.12	-0.26	-0.26	-0.26	-0.12
rs	1	1	1	0.89	0.95	0.95	0.95	0.89	0.87	0.87	0.87	0.82	0.63	0.63	0.63	0.82	-0.052	-0.052	-0.052	0.00034	0.00034	-0.12	-0.26	-0.26	-0.26	-0.12
lw	0.89	0.89	0.89	1	0.97	0.97	0.97	1	0.97	0.97	0.97	0.98	0.77	0.77	0.77	0.98	0.088	0.088	0.088	0.19	0.19	0.043	-0.18	-0.18	-0.18	0.043
lf	0.95	0.95	0.95	0.97	1	1	1	0.97	0.97	0.97	0.97	0.94	0.75	0.75	0.75	0.94	0.049	0.049	0.049	0.12	0.12	-0.023	-0.21	-0.21	-0.21	-0.023
cf	0.95	0.95	0.95	0.97	1	1	1	0.97	0.97	0.97	0.97	0.94	0.75	0.75	0.75	0.94	0.049	0.049	0.049	0.12	0.12	-0.023	-0.21	-0.21	-0.21	-0.023
rf	0.95	0.95	0.95	0.97	1	1	1	0.97	0.97	0.97	0.97	0.94	0.75	0.75	0.75	0.94	0.049	0.049	0.049	0.12	0.12	-0.023	-0.21	-0.21	-0.21	-0.023
rw	0.89	0.89	0.89	1	0.97	0.97	0.97	1	0.97	0.97	0.97	0.98	0.77	0.77	0.77	0.98	0.088	0.088	0.088	0.19	0.19	0.043	-0.18	-0.18	-0.18	0.043
lam	0.87	0.87	0.87	0.97	0.97	0.97	0.97	0.97	1	1	1	0.97	0.86	0.86	0.86	0.97	0.19	0.19	0.19	0.25	0.25	0.11	-0.097	-0.097	-0.097	0.11
cam	0.87	0.87	0.87	0.97	0.97	0.97	0.97	0.97	1	1	1	0.97	0.86	0.86	0.86	0.97	0.19	0.19	0.19	0.25	0.25	0.11	-0.097	-0.097	-0.097	0.11
ram	0.87	0.87	0.87	0.97	0.97	0.97	0.97	0.97	1	1	1	0.97	0.86	0.86	0.86	0.97	0.19	0.19	0.19	0.25	0.25	0.11	-0.097	-0.097	-0.097	0.11
lm	0.82	0.82	0.82	0.98	0.94	0.94	0.94	0.98	0.97	0.97	0.97	1	0.86	0.86	0.86	1	0.25	0.25	0.25	0.35	0.35	0.2	-0.032	-0.032	-0.032	0.2
lcm	0.63	0.63	0.63	0.77	0.75	0.75	0.75	0.77	0.86	0.86	0.86	0.86	1	1	1	0.86	0.62	0.62	0.62	0.63	0.63	0.51	0.33	0.33	0.33	0.51
cm	0.63	0.63	0.63	0.77	0.75	0.75	0.75	0.77	0.86	0.86	0.86	0.86	1	1	1	0.86	0.62	0.62	0.62	0.63	0.63	0.51	0.33	0.33	0.33	0.51
rcm	0.63	0.63	0.63	0.77	0.75	0.75	0.75	0.77	0.86	0.86	0.86	0.86	1	1	1	0.86	0.62	0.62	0.62	0.63	0.63	0.51	0.33	0.33	0.33	0.51
rm	0.82	0.82	0.82	0.98	0.94	0.94	0.94	0.98	0.97	0.97	0.97	1	0.86	0.86	0.86	1	0.25	0.25	0.25	0.35	0.35	0.2	-0.032	-0.032	-0.032	0.2
ldm	-0.052	-0.052	-0.052	0.088	0.049	0.049	0.049	0.088	0.19	0.19	0.19	0.25	0.62	0.62	0.62	0.25	1	1	1	0.96	0.96	0.96	0.92	0.92	0.92	0.96
cdm	-0.052	-0.052	-0.052	0.088	0.049	0.049	0.049	0.088	0.19	0.19	0.19	0.25	0.62	0.62	0.62	0.25	1	1	1	0.96	0.96	0.96	0.92	0.92	0.92	0.96
rdm	-0.052	-0.052	-0.052	0.088	0.049	0.049	0.049	0.088	0.19	0.19	0.19	0.25	0.62	0.62	0.62	0.25	1	1	1	0.96	0.96	0.96	0.92	0.92	0.92	0.96
lwb	0.00034	0.00034	0.00034	0.19	0.12	0.12	0.12	0.19	0.25	0.25	0.25	0.35	0.63	0.63	0.63	0.35	0.96	0.96	0.96	1	1	0.98	0.87	0.87	0.87	0.98
rwb	0.00034	0.00034	0.00034	0.19	0.12	0.12	0.12	0.19	0.25	0.25	0.25	0.35	0.63	0.63	0.63	0.35	0.96	0.96	0.96	1	1	0.98	0.87	0.87	0.87	0.98
lb	-0.12	-0.12	-0.12	0.043	-0.023	-0.023	-0.023	0.043	0.11	0.11	0.11	0.2	0.51	0.51	0.51	0.2	0.96	0.96	0.96	0.98	0.98	1	0.93	0.93	0.93	1
lcb	-0.26	-0.26	-0.26	-0.18	-0.21	-0.21	-0.21	-0.18	-0.097	-0.097	-0.097	-0.032	0.33	0.33	0.33	-0.032	0.92	0.92	0.92	0.87	0.87	0.93	1	1	1	0.93
cb	-0.26	-0.26	-0.26	-0.18	-0.21	-0.21	-0.21	-0.18	-0.097	-0.097	-0.097	-0.032	0.33	0.33	0.33	-0.032	0.92	0.92	0.92	0.87	0.87	0.93	1	1	1	0.93
rcb	-0.26	-0.26	-0.26	-0.18	-0.21	-0.21	-0.21	-0.18	-0.097	-0.097	-0.097	-0.032	0.33	0.33	0.33	-0.032	0.92	0.92	0.92	0.87	0.87	0.93	1	1	1	0.93
rb	-0.12	-0.12	-0.12	0.043	-0.023	-0.023	-0.023	0.043	0.11	0.11	0.11	0.2	0.51	0.51	0.51	0.2	0.96	0.96	0.96	0.98	0.98	1	0.93	0.93	0.93	1



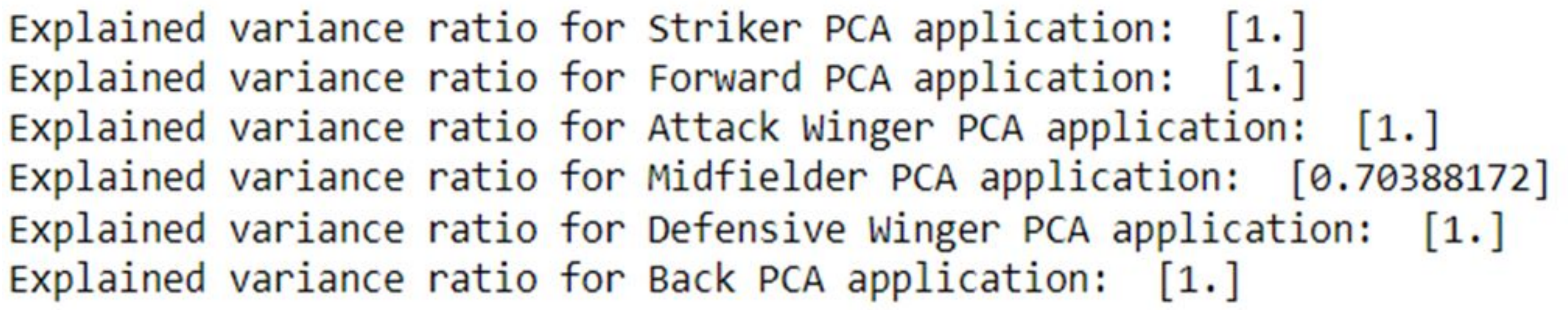
	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_short_passing	attacking_volleys	skill_dribbling	skill_curve	skill_fk_accuracy	skill_long_passing	skill_ball_control	movement_acceleration	movement_sprint_speed	movement_agility	movement_reactions	movement_balance	power_shot_power	power_jumping	power_stamina	power_strength	power_long_shots	defending_standing_tackle	defending_sliding_tackle
attacking_crossing	1	0.24	-0.23	0.47	0.31	0.65	0.57	0.5	0.43	0.57	0.5	0.44	0.49	0.3	0.39	0.29	-0.21	0.23	-0.19	0.32	0.088	0.089
attacking_finishing	0.24	1	-0.00059	0.23	0.81	0.59	0.54	0.43	0.079	0.55	0.36	0.34	0.41	0.26	0.27	0.71	-0.16	-0.023	-0.1	0.82	-0.66	-0.68
attacking_heading_accuracy	-0.23	-0.00059	1	0.056	-0.032	-0.17	-0.2	-0.22	-0.064	-0.037	-0.28	-0.19	-0.36	0.37	-0.4	0.097	0.47	0.17	0.57	0.023	0.25	0.24
attacking_short_passing	0.47	0.23	0.056	1	0.34	0.55	0.51	0.54	0.85	0.72	0.16	0.12	0.3	0.55	0.24	0.44	-0.041	0.32	0.028	0.4	0.26	0.24
attacking_volleys	0.31	0.81	-0.032	0.34	1	0.61	0.66	0.57	0.22	0.58	0.35	0.32	0.42	0.26	0.29	0.75	-0.2	-0.025	-0.13	0.77	-0.52	-0.53
skill_dribbling	0.65	0.59	-0.17	0.55	0.61	1	0.7	0.58	0.43	0.83	0.54	0.48	0.6	0.41	0.45	0.58	-0.2	0.16	-0.19	0.6	-0.2	-0.21
skill_curve	0.57	0.54	-0.2	0.51	0.66	0.7	1	0.72	0.47	0.66	0.41	0.34	0.52	0.29	0.37	0.6	-0.25	0.077	-0.19	0.61	-0.21	-0.22
skill_fk_accuracy	0.5	0.43	-0.22	0.54	0.57	0.58	0.72	1	0.55	0.58	0.25	0.17	0.4	0.25	0.33	0.55	-0.29	0.034	-0.21	0.57	-0.098	-0.1
skill_long_passing	0.43	0.079	-0.064	0.85	0.22	0.43	0.47	0.55	1	0.58	0.069	0.027	0.25	0.38	0.23	0.34	-0.1	0.26	-0.028	0.29	0.33	0.32
skill_ball_control	0.57	0.55	-0.037	0.72	0.58	0.83	0.66	0.58	0.58	1	0.41	0.37	0.52	0.51	0.36	0.61	-0.11	0.24	-0.052	0.61	-0.077	-0.1
movement_acceleration	0.5	0.36	-0.28	0.16	0.35	0.54	0.41	0.25	0.069	0.41	1	0.85	0.74	0.15	0.49	0.26	-0.079	0.22	-0.24	0.3	-0.25	-0.25
movement_sprint_speed	0.44	0.34	-0.19	0.12	0.32	0.48	0.34	0.17	0.027	0.37	0.85	1	0.63	0.15	0.36	0.25	-0.026	0.23	-0.11	0.26	-0.23	-0.23
movement_agility	0.49	0.41	-0.36	0.3	0.42	0.6	0.52	0.4	0.25	0.52	0.74	0.63	1	0.19	0.64	0.33	-0.11	0.17	-0.33	0.41	-0.25	-0.26
movement_reactions	0.3	0.26	0.37	0.55	0.26	0.41	0.29	0.25	0.38	0.51	0.15	0.15	0.19	1	0.064	0.33	0.18	0.41	0.23	0.35	0.24	0.21
movement_balance	0.39	0.27	-0.4	0.24	0.29	0.45	0.37	0.33	0.23	0.36	0.49	0.36	0.64	0.064	1	0.2	-0.15	0.043	-0.57	0.27	-0.19	-0.17
power_shot_power	0.29	0.71	0.097	0.44	0.75	0.58	0.6	0.55	0.34	0.61	0.26	0.25	0.33	0.33	0.2	1	-0.11	0.076	0.03	0.78	-0.34	-0.35
power_jumping	-0.21	-0.16	0.47	-0.041	-0.2	-0.2	-0.25	-0.29	-0.1	-0.11	-0.079	-0.026	-0.11	0.18	-0.15	-0.11	1	0.3	0.44	-0.1	0.23	0.22
power_stamina	0.23	-0.023	0.17	0.32	-0.025	0.16	0.077	0.034	0.26	0.24	0.22	0.23	0.17	0.41	0.043	0.076	0.3	1	0.3	0.09	0.33	0.31
power_strength	-0.19	-0.1	0.57	0.028	-0.13	-0.19	-0.19	-0.21	-0.028	-0.052	-0.24	-0.11	-0.33	0.23	-0.57	0.03	0.44	0.3	1	-0.063	0.28	0.24
power_long_shots	0.32	0.82	0.023	0.4	0.77	0.6	0.61	0.57	0.29	0.61	0.3	0.26	0.41	0.35	0.27	0.78	-0.1	0.09	-0.063	1	-0.44	-0.47
defending_standing_tackle	0.088	-0.66	0.25	0.26	-0.52	-0.2	-0.21	-0.098	0.33	-0.077	-0.25	-0.23	-0.25	0.24	-0.19	-0.34	0.23	0.33	0.28	-0.44	1	0.96
defending_sliding_tackle	0.089	-0.68	0.24	0.24	-0.53	-0.21	-0.22	-0.1	0.32	-0.1	-0.25	-0.23	-0.26	0.21	-0.17	-0.35	0.22	0.31	0.24	-0.47	0.96	1

Data Modeling

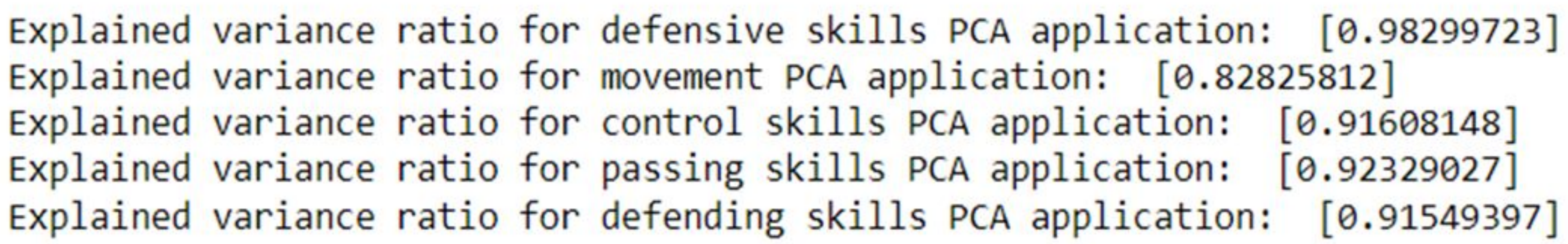
Step 1: Drop unused attributes for clustering

Step 2: Reduce Dimension

Step 3: Standardized data.

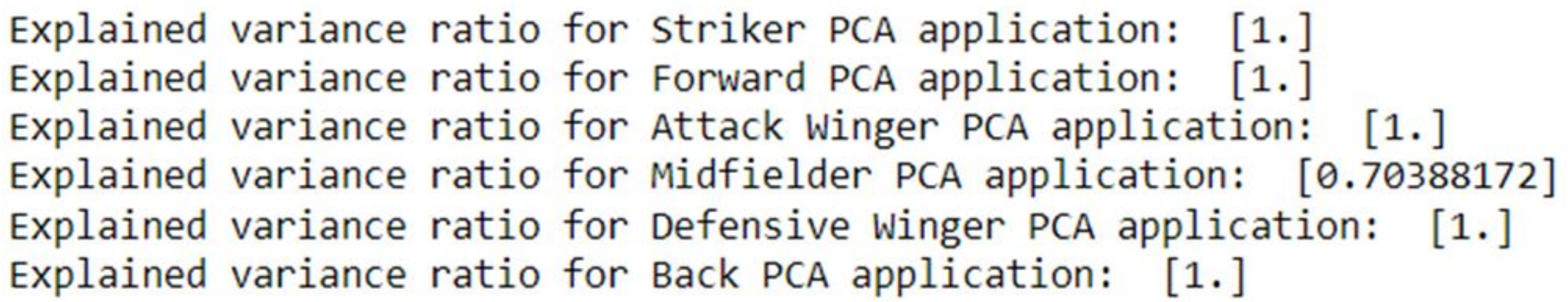


```
Explained variance ratio for Striker PCA application:  [1.]
Explained variance ratio for Forward PCA application:  [1.]
Explained variance ratio for Attack Winger PCA application:  [1.]
Explained variance ratio for Midfielder PCA application:  [0.70388172]
Explained variance ratio for Defensive Winger PCA application:  [1.]
Explained variance ratio for Back PCA application:  [1.]
```

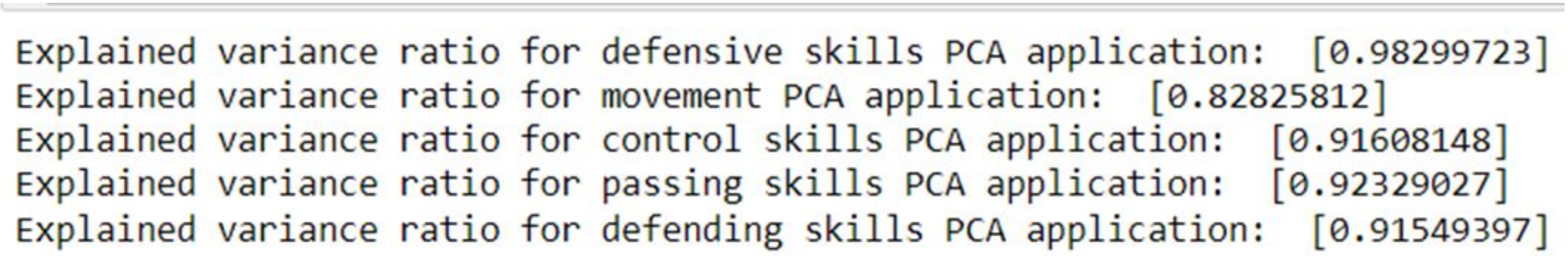


```
Explained variance ratio for defensive skills PCA application:  [0.98299723]
Explained variance ratio for movement PCA application:  [0.82825812]
Explained variance ratio for control skills PCA application:  [0.91608148]
Explained variance ratio for passing skills PCA application:  [0.92329027]
Explained variance ratio for defending skills PCA application:  [0.91549397]
```

Principal Component Analysis (PCA)

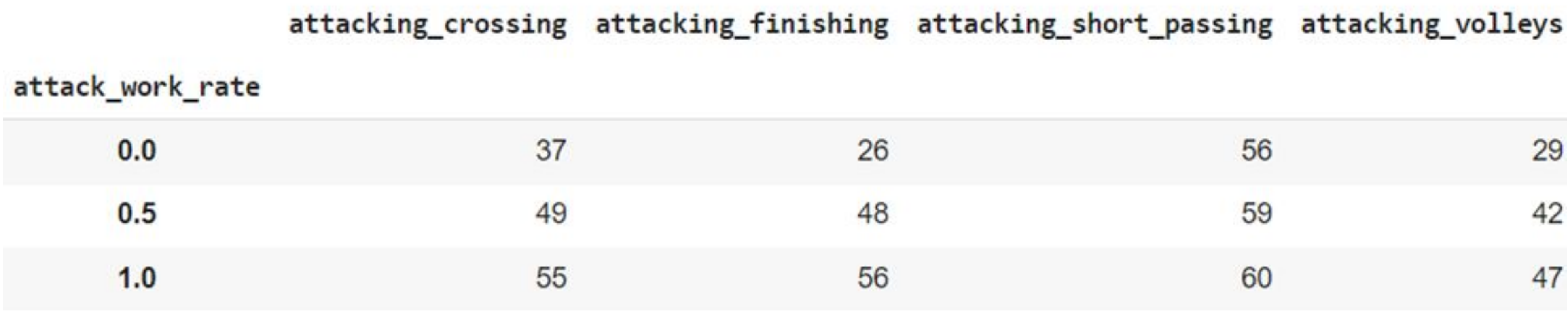


```
Explained variance ratio for Striker PCA application:  [1.]
Explained variance ratio for Forward PCA application:  [1.]
Explained variance ratio for Attack Winger PCA application:  [1.]
Explained variance ratio for Midfielder PCA application:  [0.70388172]
Explained variance ratio for Defensive Winger PCA application:  [1.]
Explained variance ratio for Back PCA application:  [1.]
```

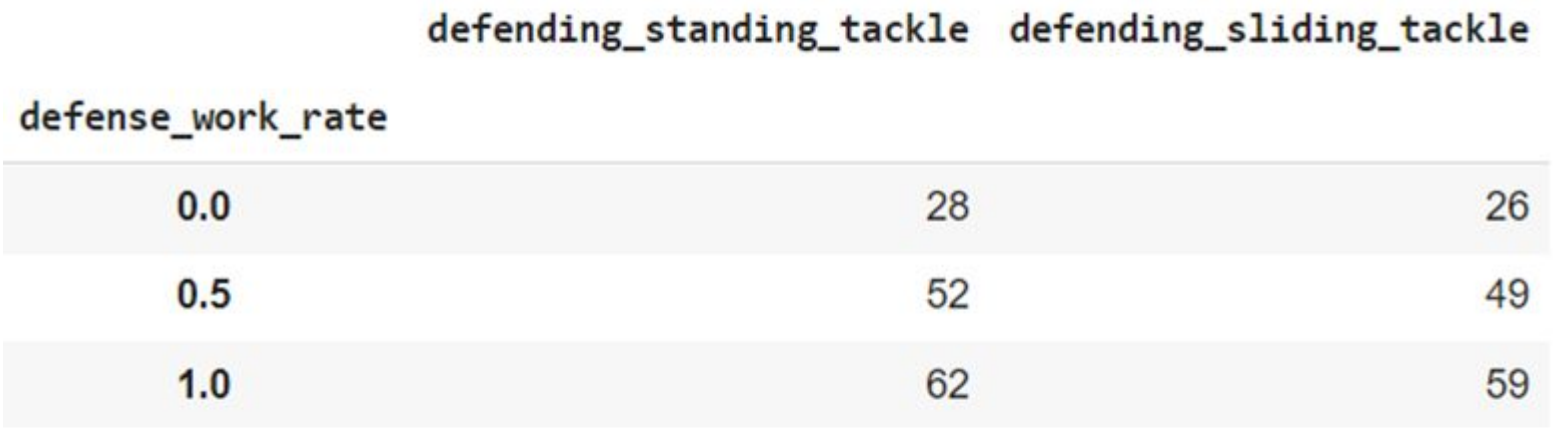


```
Explained variance ratio for defensive skills PCA application:  [0.98299723]
Explained variance ratio for movement PCA application:  [0.82825812]
Explained variance ratio for control skills PCA application:  [0.91608148]
Explained variance ratio for passing skills PCA application:  [0.92329027]
Explained variance ratio for defending skills PCA application:  [0.91549397]
```

Discover the relationship between Work Rate and other skills

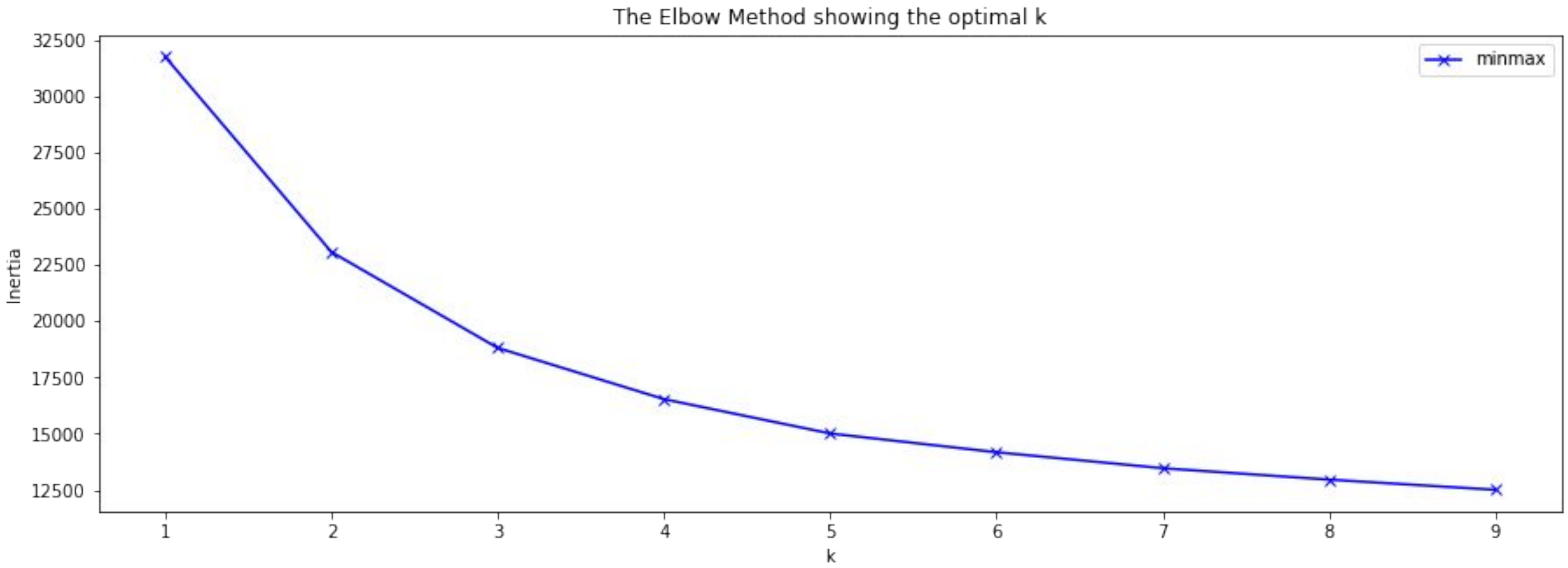


	attacking_crossing	attacking_finishing	attacking_short_passing	attacking_volleys
attack_work_rate				
0.0	37	26	56	29
0.5	49	48	59	42
1.0	55	56	60	47

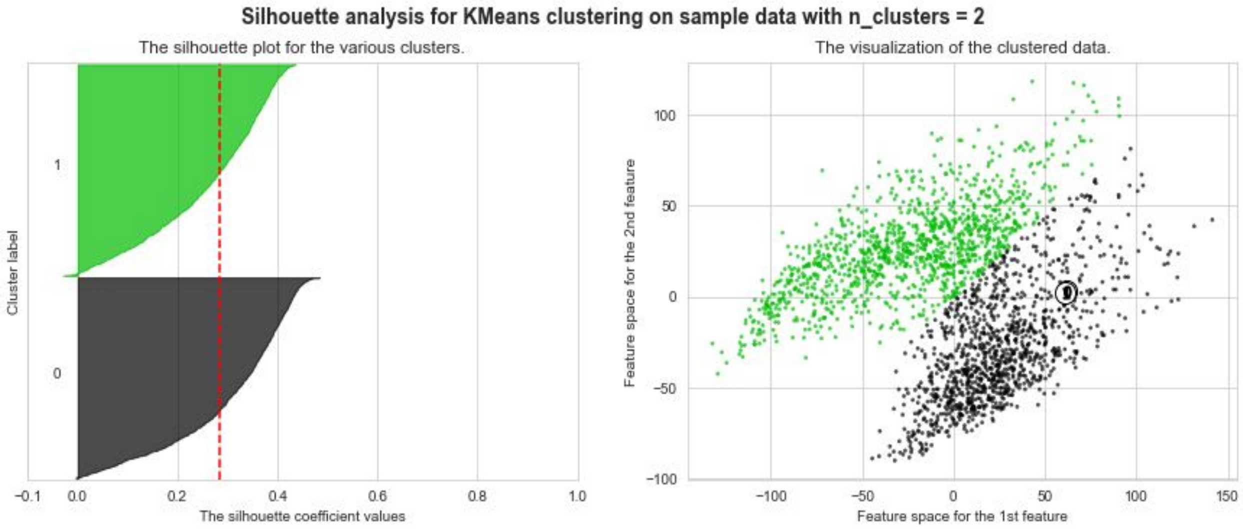


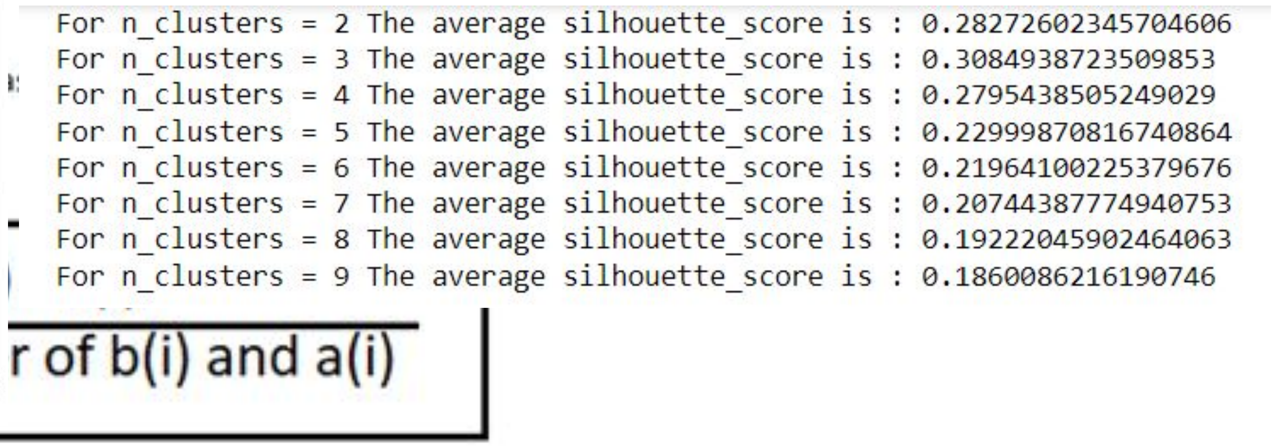
	defending_standing_tackle	defending_sliding_tackle
defense_work_rate		
0.0	28	26
0.5	52	49
1.0	62	59

Method to find optimal K



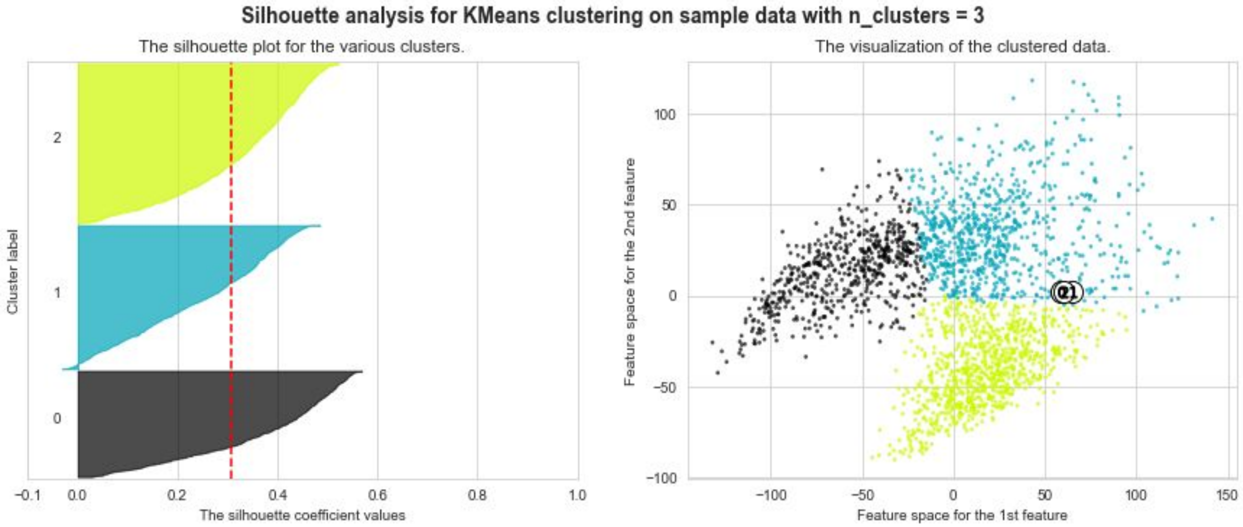
Silhouette method

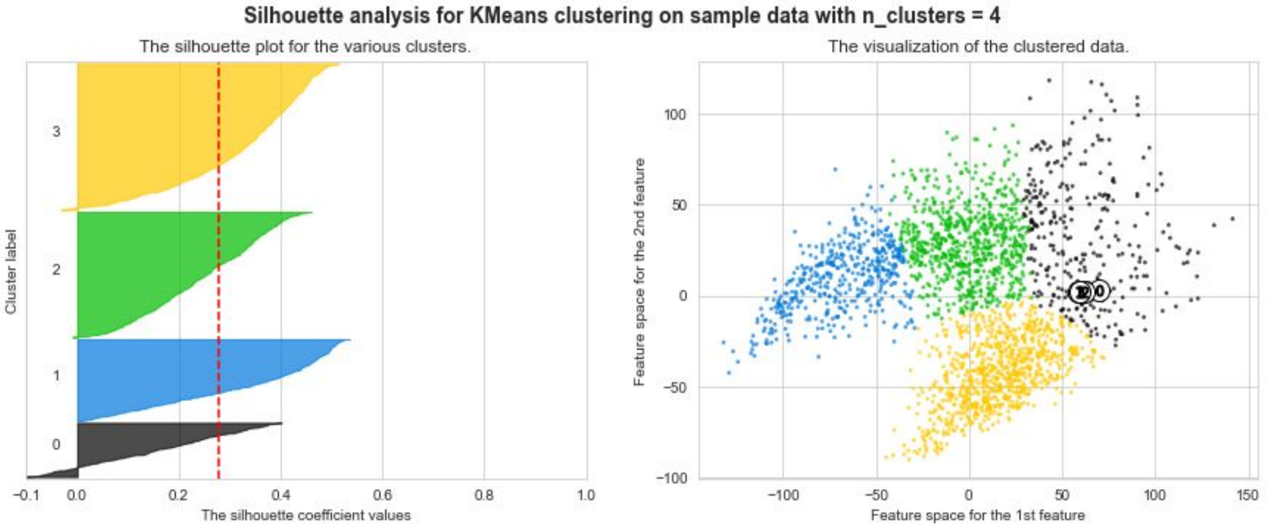




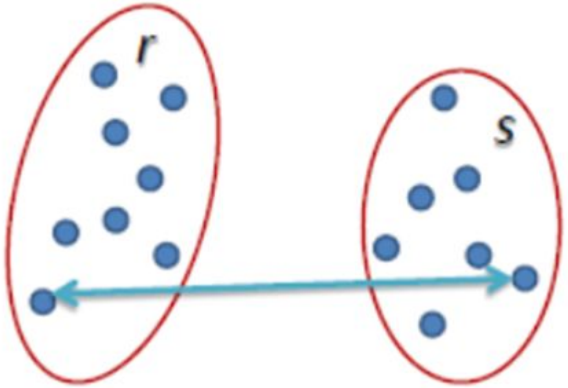
```
For n_clusters = 2 The average silhouette_score is : 0.28272602345704606
For n_clusters = 3 The average silhouette_score is : 0.3084938723509853
For n_clusters = 4 The average silhouette_score is : 0.2795438505249029
For n_clusters = 5 The average silhouette_score is : 0.22999870816740864
For n_clusters = 6 The average silhouette_score is : 0.21964100225379676
For n_clusters = 7 The average silhouette_score is : 0.20744387774940753
For n_clusters = 8 The average silhouette_score is : 0.1922045902464063
For n_clusters = 9 The average silhouette_score is : 0.1860086216190746
```

r of b(i) and a(i)

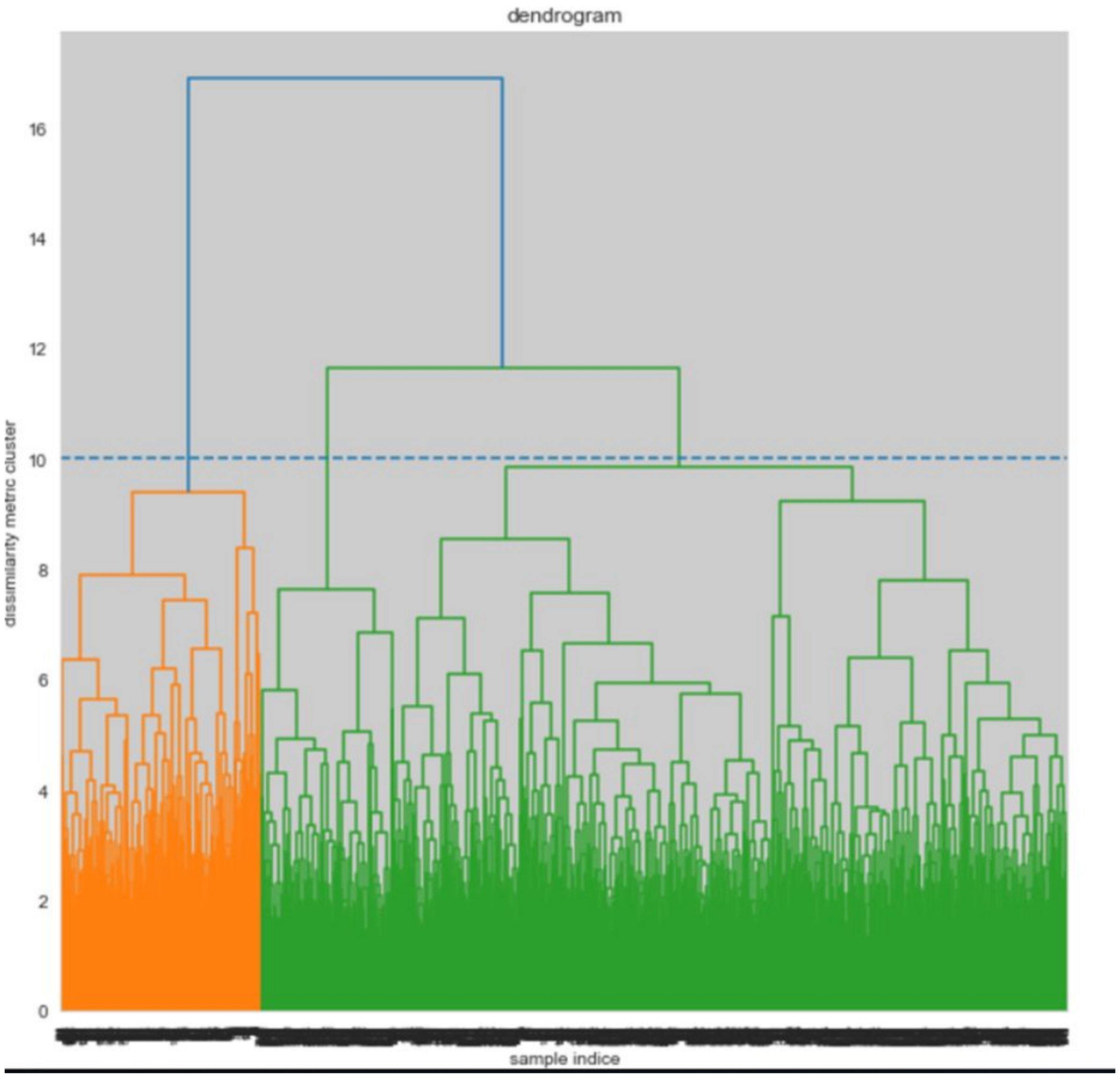




Hierarchical Clustering



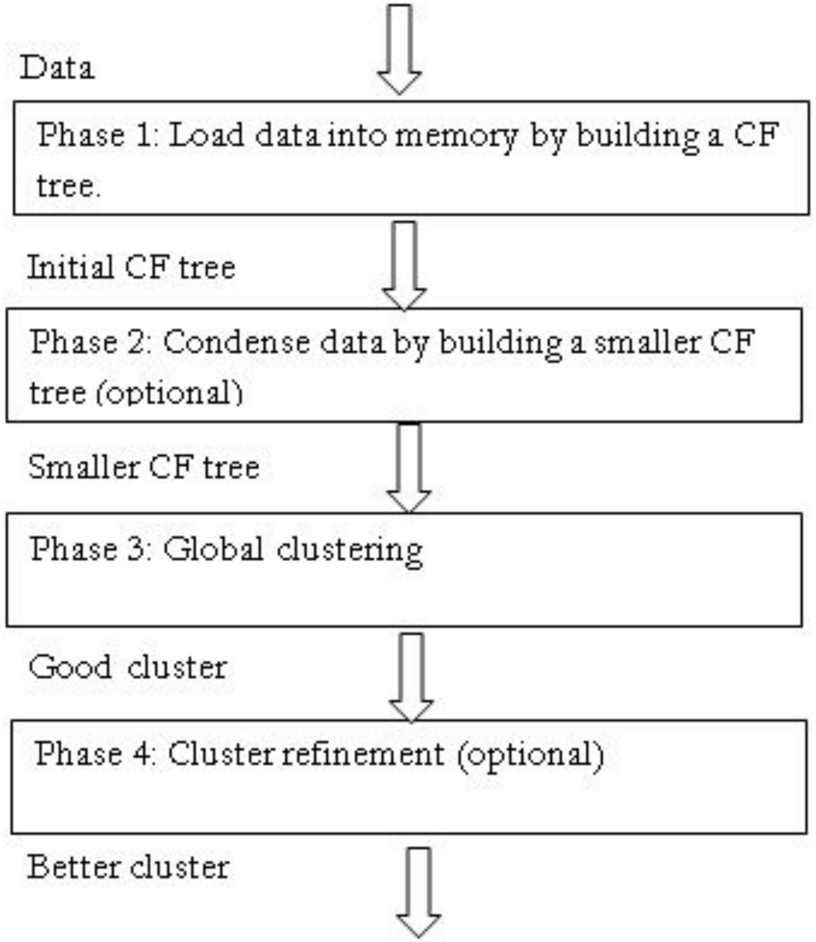
$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$



Birch Algorithm

Birch is a Hierarchical clustering algorithm using top-down

- Idea: Each successful iteration, a cluster is split into smaller clusters according to the value of some similarity measure until each object is a cluster or until the stopping condition is satisfied.
- This approach into using divide and conquer strategy in the clustering process



Data

Phase 1: Load data into memory by building a CF tree.

Initial CF tree

Phase 2: Condense data by building a smaller CF tree (optional)

Smaller CF tree

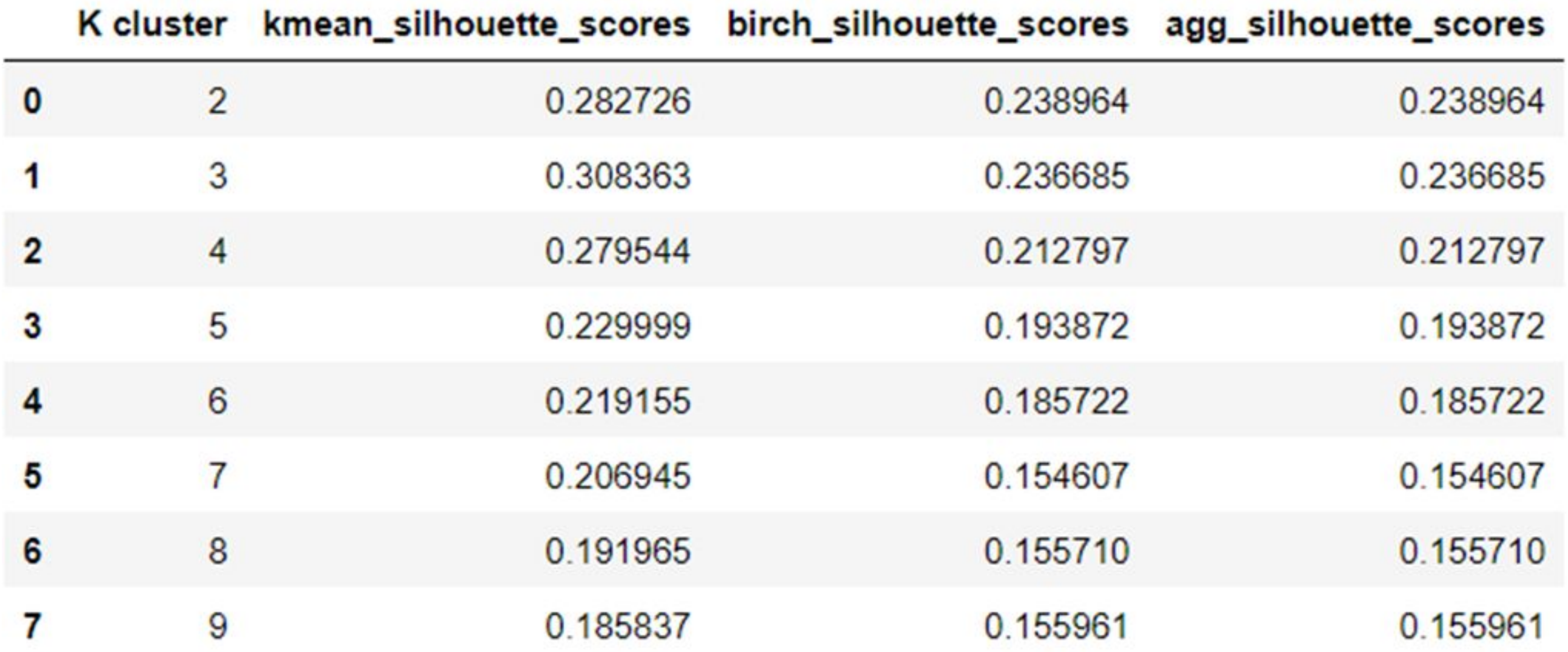
Phase 3: Global clustering

Good cluster

Phase 4: Cluster refinement (optional)

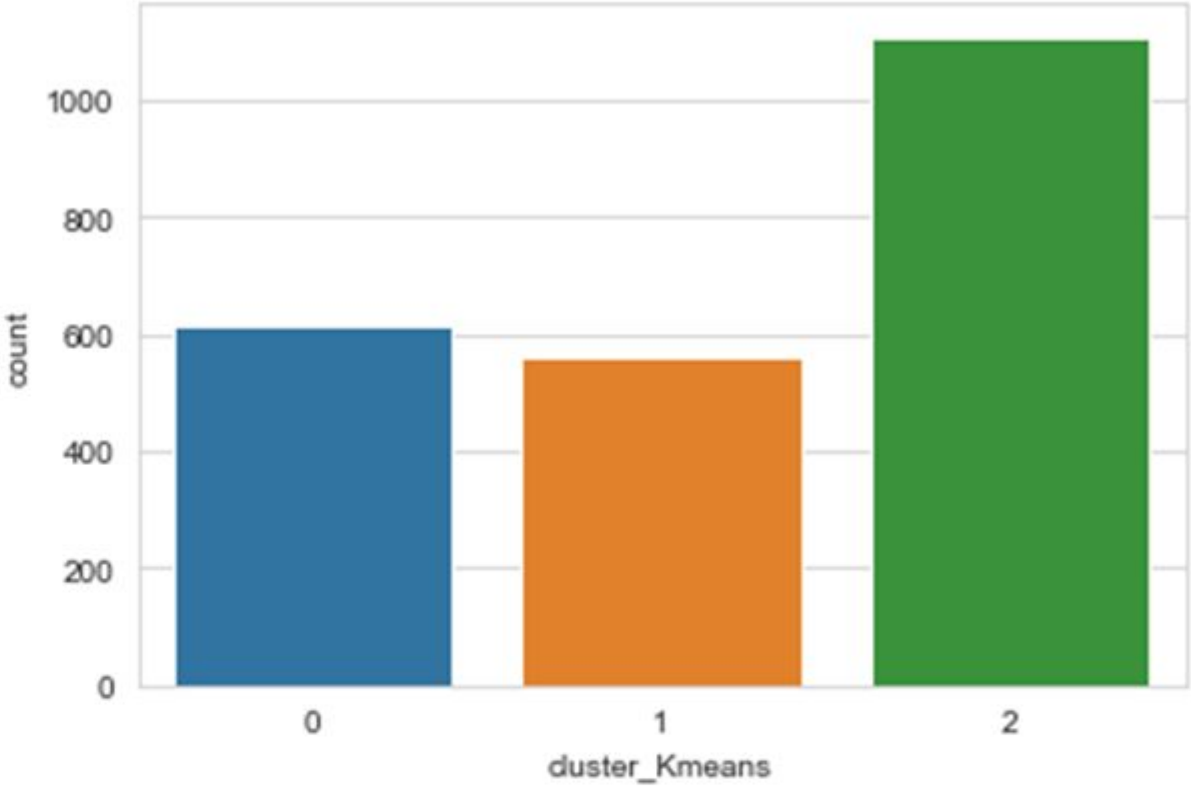
Better cluster

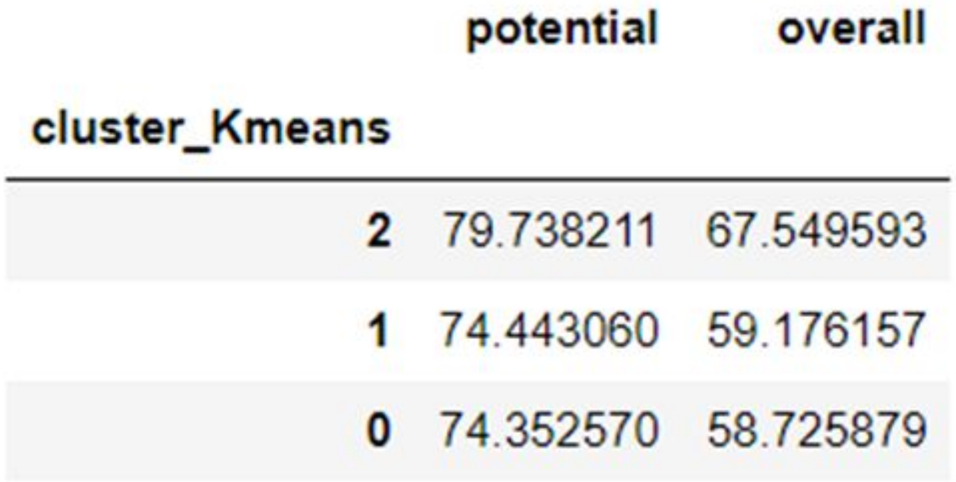
Comparision three Algorithms



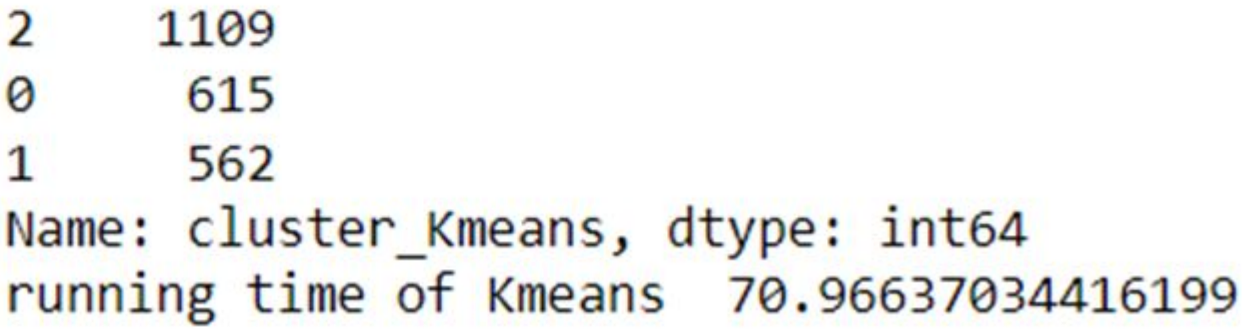
	K cluster	kmean_silhouette_scores	birch_silhouette_scores	agg_silhouette_scores
0	2	0.282726	0.238964	0.238964
1	3	0.308363	0.236685	0.236685
2	4	0.279544	0.212797	0.212797
3	5	0.229999	0.193872	0.193872
4	6	0.219155	0.185722	0.185722
5	7	0.206945	0.154607	0.154607
6	8	0.191965	0.155710	0.155710
7	9	0.185837	0.155961	0.155961

Result



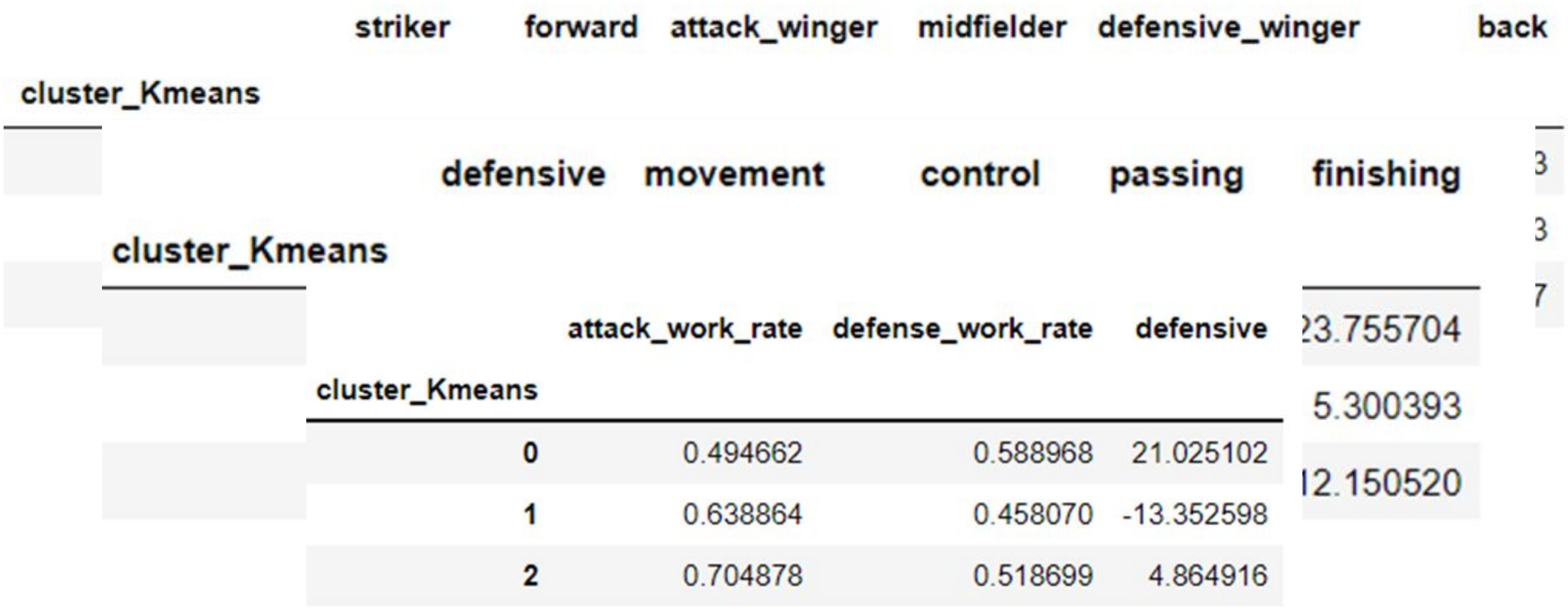


	potential	overall
cluster_Kmeans		
2	79.738211	67.549593
1	74.443060	59.176157
0	74.352570	58.725879



```
2    1109
0     615
1     562
Name: cluster_Kmeans, dtype: int64
running time of Kmeans  70.96637034416199
```

Result



	striker	forward	attack_winger	midfielder	defensive_winger	back
cluster_Kmeans						
						3
						3
						7

	defensive	movement	control	passing	finishing
cluster_Kmeans					
					?3.755704
					5.300393
					!2.150520

	attack_work_rate	defense_work_rate	defensive
cluster_Kmeans			
0	0.494662	0.588968	21.025102
1	0.638864	0.458070	-13.352598
2	0.704878	0.518699	4.864916

THANK YOU FOR LISTENING