Analysis Project: Uncovering Soccer Player Roles

Contents

Analysis Project: Uncovering Soccer Player Roles

Section 1: Main objective of the analysis

This project will demonstrate the use of unsupervised learning techniques, specifically clustering and dimensionality reduction, to identify distinct player roles from a soccer dataset.

Stakeholders of this data will be able to:

- **General:** Identify player types to help Identify distinct player roles: Beyond traditional positions (defender, midfielder, attacker), can we find more nuanced roles based on their on-field actions (e.g., "pressing forward," "playmaking defender," "ball-winning midfielder")?

- **Coaches/Tactical Analysts:** Provide data-driven insights into player roles for recruitment and tactical setup.

- **Scouts:** Help identify players fitting specific, nuanced roles that might be missed by traditional scouting.

Analysis Project: Uncovering Soccer Player Roles
Section 2: Brief description of the data

PlayerStats. This dataset is rich in detailed event information (passes, shots, dribbles, pressures, etc.) with positional data for players over the course of a season.

The event data allows for the creation of data frames that describe player actions and physical statistics. This richness is ideal for both clustering (to find groups of similar players within positional characteristics) and dimensionality reduction (to distil complex player attributes into fewer, interpretable components).

Detailed player attribute fields can be seen on the next slide:

Analysis Project: Uncovering Soccer Player Roles

Section 2: Brief description of the data

Original Player Attributes (from players_df):

- player_id
- player_name
- team_id
- team_name
- height_cm
- weight_kg
- dominant_foot

Per-Match Player Statistics (from player_stats):

- match_id
- minutes_played
- passes
- pass_success_rate
- progressive_pass_distance
- passes_into_final_third
- key_passes
- short_passes
- long_passes
- through_balls
- shots
- shots_on_target
- xg_per_shot
- tackles
- interceptions
- clearances
- pressures
- dribbles
- successful_dribbles
- progressive_carries
- carries_into_final_third
- fouls_committed
- fouls_won
- goals
- assists
- saves
- aerial_duels_won
- player_average_x
- player_average_y

Aggregated & Engineered Player Statistics (used for clustering, generally suffixed with _p90_avg or _avg):

- passes_p90_avg
- progressive_pass_distance_p90_avg
- passes_into_final_third_p90_avg
- key_passes_p90_avg
- short_passes_p90_avg
- long_passes_p90_avg
- through_balls_p90_avg
- shots_p90_avg
- shots_on_target_p90_avg
- tackles_p90_avg
- interceptions_p90_avg
- clearances_p90_avg
- pressures_p90_avg
- dribbles_p90_avg
- successful_dribbles_p90_avg
- progressive_carries_p90_avg
- carries_into_final_third_p90_avg
- fouls_committed_p90_avg
- fouls_won_p90_avg
- goals_p90_avg
- assists_p90_avg
- saves_p90_avg
- aerial_duels_won_p90_avg
- pass_success_rate_avg
- xg_per_shot_avg
- player_average_x_avg
- player_average_y_avg
- most_frequent_position (derived)
- total_minutes_played (derived)

Analysis Project: Uncovering Soccer Player Roles
Section 3: Summary of data exploration

Initial Analysis
The available dataset tables and headers were reviewed to gain an understanding of what data was initially available. This helped identify that 7 player attributes were available and a further 29 player match attributes were available for each game.

Data Cleaning
One of the key pieces of data to cleanse was to ensure that the stats were comparable over the length of each game (as not all players involved in the game played a full 90 mins).

Feature Engineering
Further attributes were then created based on these "per 90 min" calculations, as well as measures to identify the most frequent position that a player played in for each game.

```python
# Calculate per 90 minute stats
player_stats_filtered['p90_multiplier'] = 90 / player_stats_filtered['minutes_played']

stats_to_normalize = [
    'passes', 'progressive_pass_distance', 'passes_into_final_third', 'key_passes',
    'short_passes', 'long_passes', 'through_balls', 'shots', 'shots_on_target',
    'tackles', 'interceptions', 'clearances', 'pressures', 'dribbles',
    'successful_dribbles', 'progressive_carries', 'carries_into_final_third',
    'fouls_committed', 'fouls_won', 'goals', 'assists', 'saves', 'aerial_duels_won'
]

for col in stats_to_normalize:
    player_stats_filtered[f'{col}_p90'] = player_stats_filtered[col] * player_stats_filtered['p90_multiplier']

# Aggregate by player_id and player_name to get season averages of P90 stats
player_season_stats_p90 = player_stats_filtered.groupby(['player_id', 'player_name']).agg(
    **{f'{col}_p90_avg': (f'{col}_p90', 'mean') for col in stats_to_normalize},
    pass_success_rate_avg=('pass_success_rate', 'mean'),
    xg_per_shot_avg=('xg_per_shot', 'mean'),
    player_average_x_avg=('player_average_x', 'mean'),
    player_average_y_avg=('player_average_y', 'mean'),
    most_frequent_position=('position_name', lambda x: x.mode()[0] if not x.mode().empty else 'Unknown'),
    total_minutes_played=('minutes_played', 'sum')
).reset_index()
```

Analysis Project: Uncovering Soccer Player Roles

Section 4: Summary of training

The training applied to the data consisted of:

1. K-Means Clustering (Appendix 1 for details)

- **Application:** Clustered players on their playing statistics (e.g., tackles, passing types, assists per 90 minutes).
- **Hyperparameter Tuning:** Experimented with different values of 'k' (number of clusters) and used the Elbow method to help determine an optimal 'k'.
- **Interpretation:** Analysed the centroids of each cluster to define the "average" player profile within that group, thereby identifying distinct player roles.

2. Hierarchical Agglomerative Clustering (Appendix 2 for details)

- **Application:** Applied to the same playing statistics dataset using Ward Linkage (Euclidean Distance).
- **Interpretation:** Examined the resulting dendrogram to understand the hierarchy of player similarities. The dendrogram was cut at different levels to obtain varying numbers of clusters and compared the resulting player groupings with those from K-Means. This revealed more nuanced relationships and sub-roles.

3. Dimensionality Reduction (PCA) followed by K-Means Clustering (Appendix 3 for details)

- **Application:** Applied PCA to the playing statistics to reduce the number of features while retaining most of the variance. Selected several attributes that explained a significant portion of the variance. Reapplied K-Means to the reduced dataset (on these principal attributes).
- **Interpretation:** Analysed the principal attributes to understand what combinations of original features they represent. Then analysed the clusters formed in this lower-dimensional space.

The goal was to explore how different unsupervised techniques reveal different underlying structures in the data.

Analysis Project: Uncovering Soccer Player Roles

Section 5: Recommended model

The dimension reduction method (+ K Means) was the most suitable model for this data.

Given the large variety of player attributes in the data the earlier K-means models struggled with the sparseness of the data points (as illustrated within Appendix 1). By leveraging dimension reduction I was able to address the curse of dimensionality whilst still retaining accuracy over the clustering proposals.

By selecting only the principal components that explained a significant portion of the total variance (e.g., 90%), the PCA model was able to effectively filter out the less informative "noisy" dimensions. This allowed K-Means to cluster based on the most dominant patterns in the data (see slide 41 in Appendix 3).

With K-Means applied to PCA-reduced data, the distances were more meaningful because the data is denser, less noisy, and the features are less correlated. This lead to clearer, more distinct, and more robust clusters (player archetypes).

PCA acted as an intelligent preprocessing step that cleansed, simplified, and highlighted the most important underlying patterns in the complex player data, enabling K-Means to then effectively group players into meaningful and distinct archetypes.

Analysis Project: Uncovering Soccer Player Roles

Section 6: Summary Key Findings and Insights

The original objective was to Identify player types to help Identify distinct player roles: Beyond traditional positions (defender, midfielder, attacker).

The untrained models were able to successfully sub-categorise footballers beyond their traditional interpreted "positions" of defenders, midfielders and attackers.

I was able to observe defenders who are forward thinking in terms of assists, midfielders with solid defensive attributes, and attackers with strong pressing attributes. These results can be observed in detail in Appendix 1.

Analysis Project: Uncovering Soccer Player Roles

Section 7: Suggestions for next steps

- The model can be improved by further classifying dimension reduction for each of the different "traditional" playing positions (as dimension reduction was only applied to across "All players" rather than the individual groups of defenders, midfielders and attackers).
- This will then lead to even greater insight into the sub categorisation of "new" roles within the data.

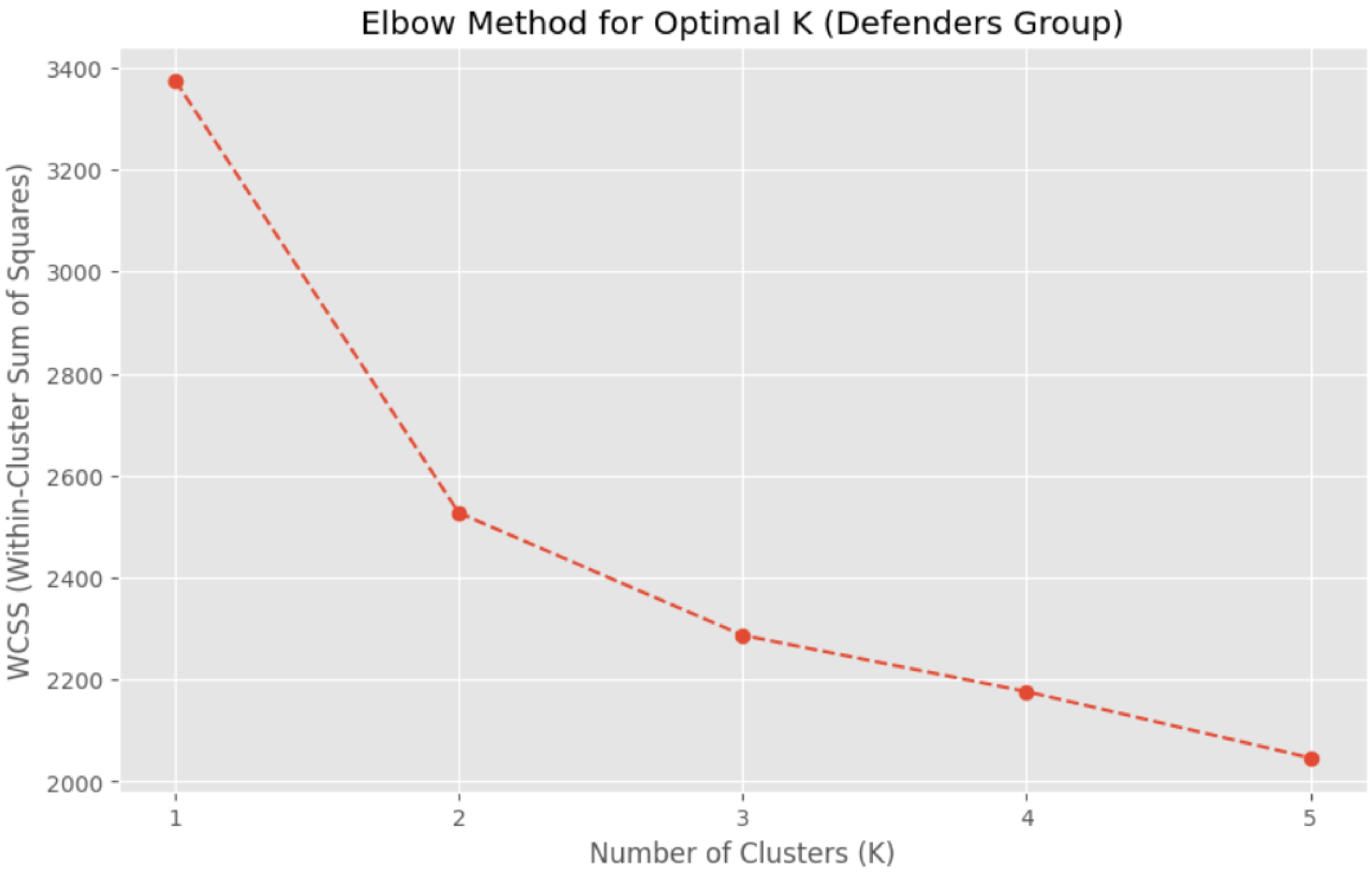
- Analysis of how traditional roles have changed over time would be an interesting further project. Football adapts regularly to how successful teams play with lesser teams copying winning styles.

- I would also like to analyse the playing styles of teams within the data to understand how the clustering of average player styles has an influence on the overall team – but I would need a much larger data set beyond just 1 league to do this.

Appendix

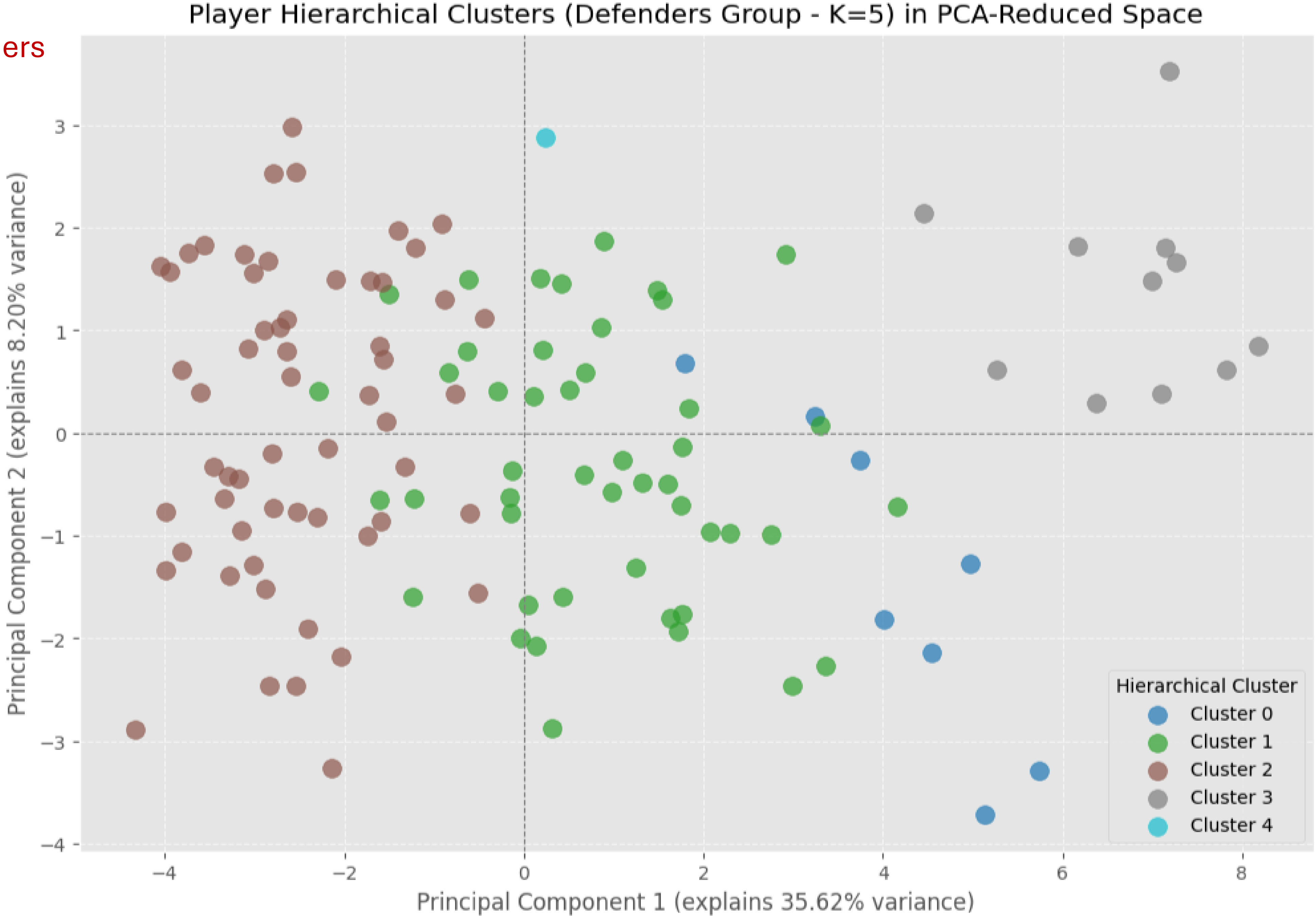
Appendix 1. K-Means Clustering

Defenders



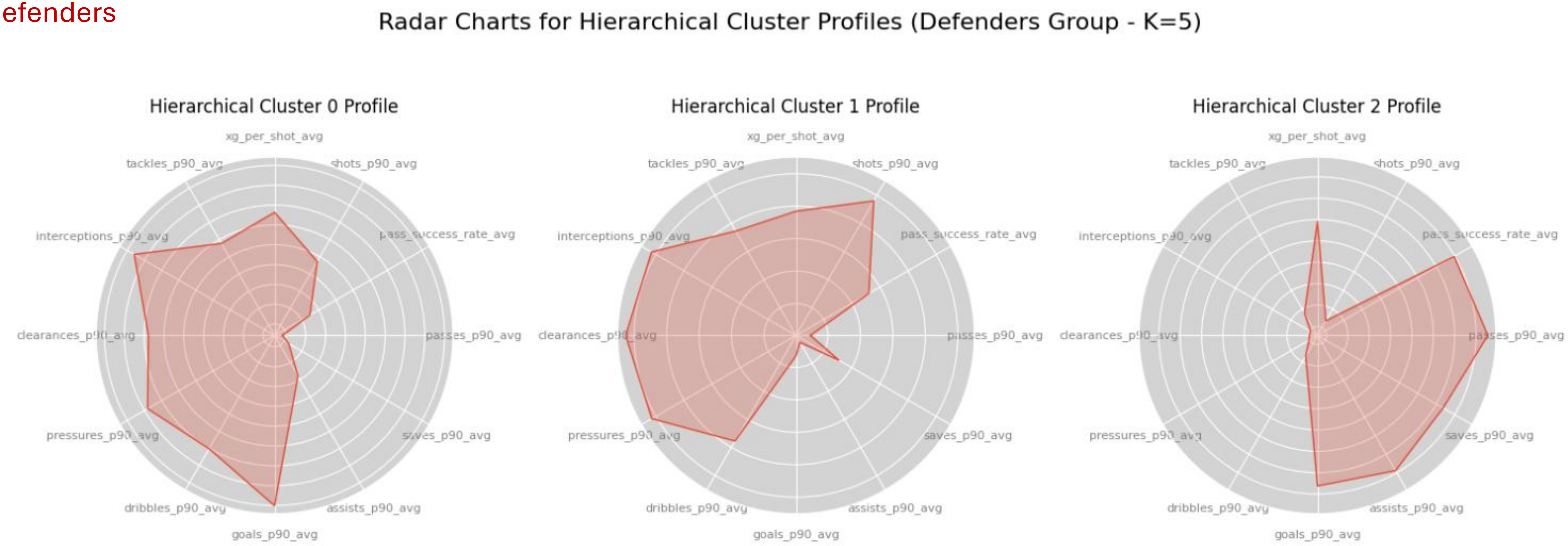
Elbow Method for Optimal K (Defenders Group)
WCSS (Within-Cluster Sum of Squares)
3400
3200
3000
2800
2600
2400
2200
2000
1
2
3
4
5
Number of Clusters (K)

Defenders



Radar Charts for Hierarchical Cluster Profiles (Defenders Group - K=5)

Defenders

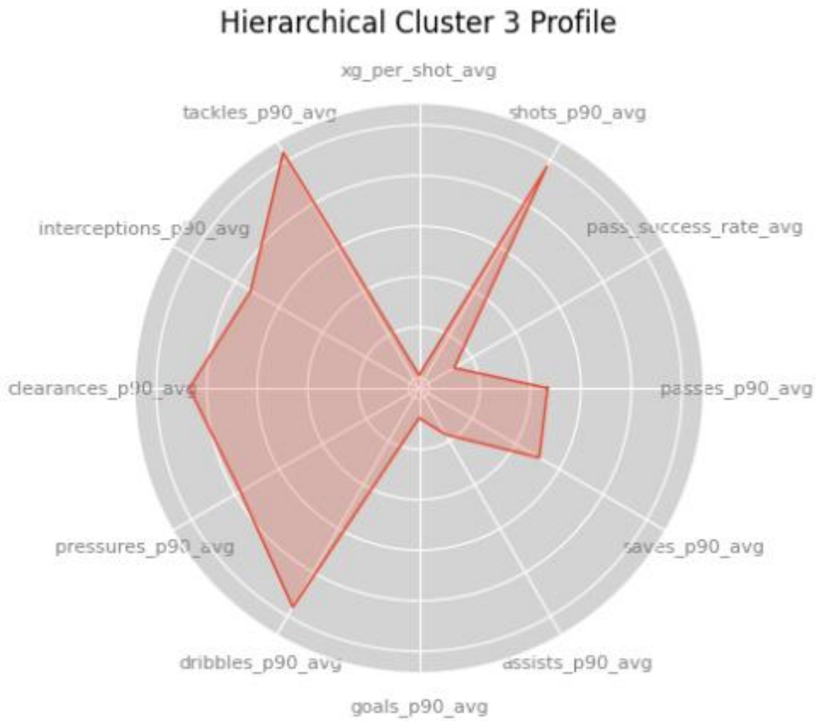


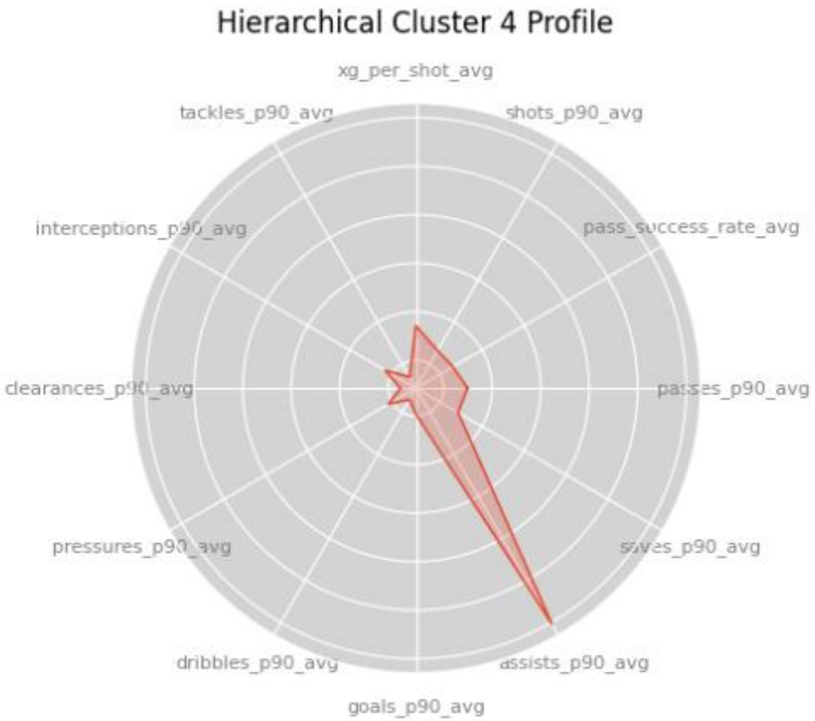
Cluster 0 Profile =
Goalscoring Defenders

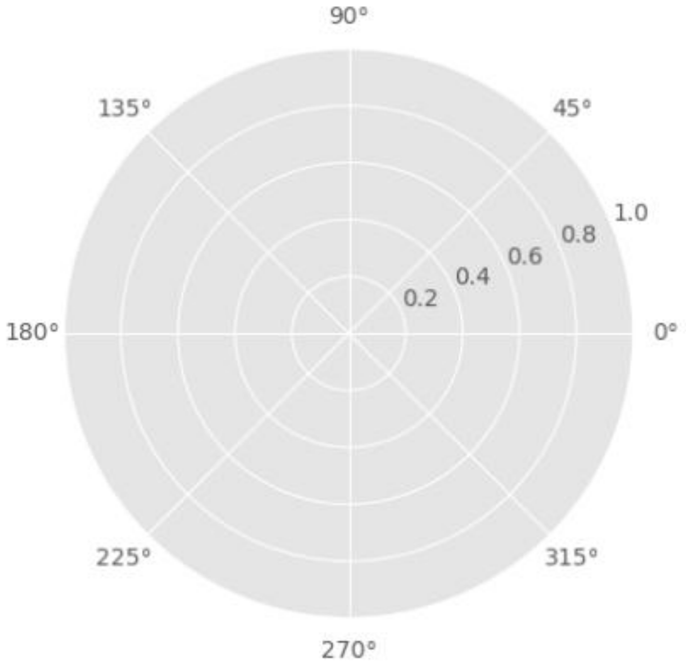
Cluster 1 Profile =
Shooting Defenders

Cluster 2 Profile =
Assisting Defenders

Defenders



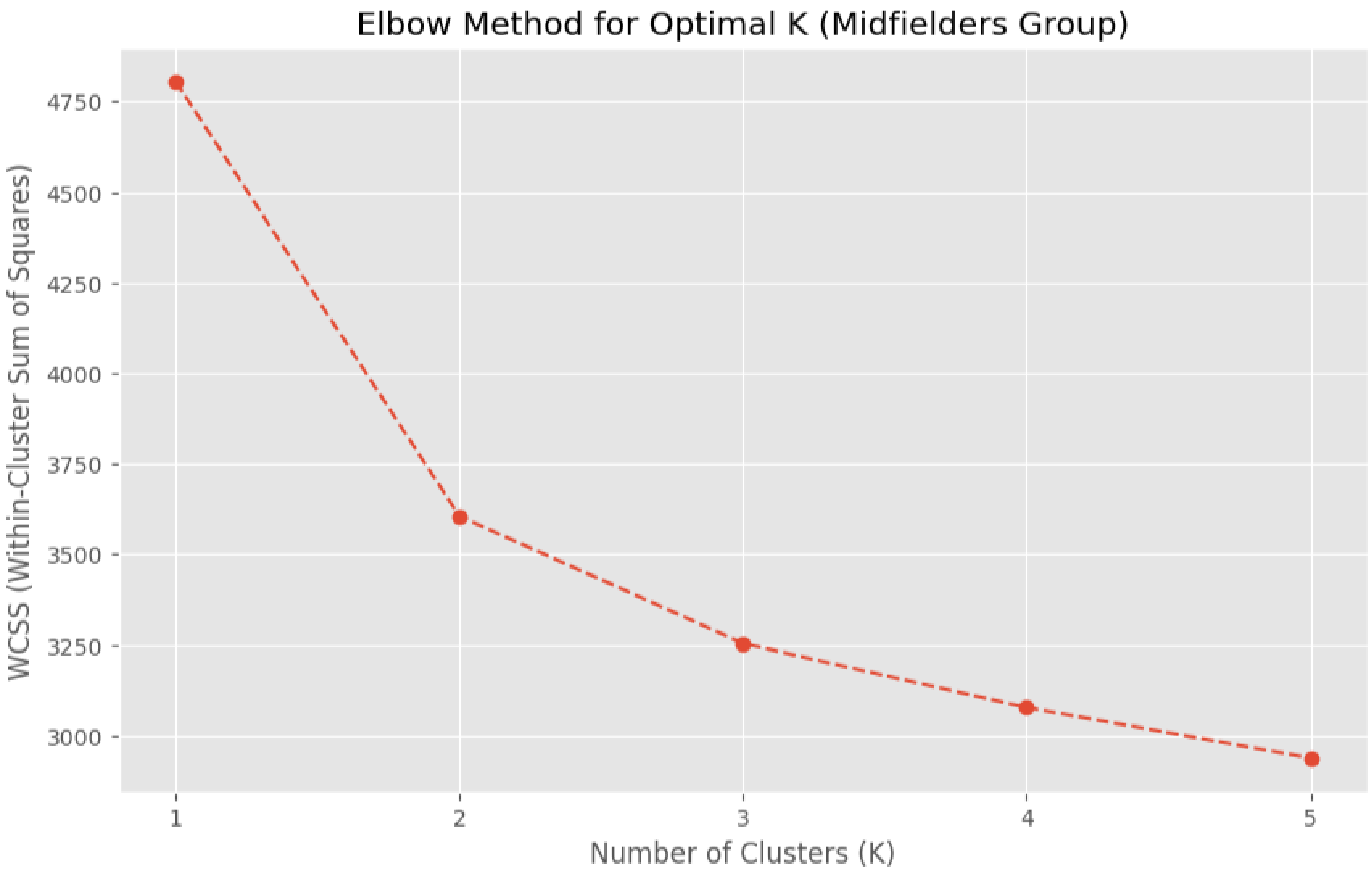




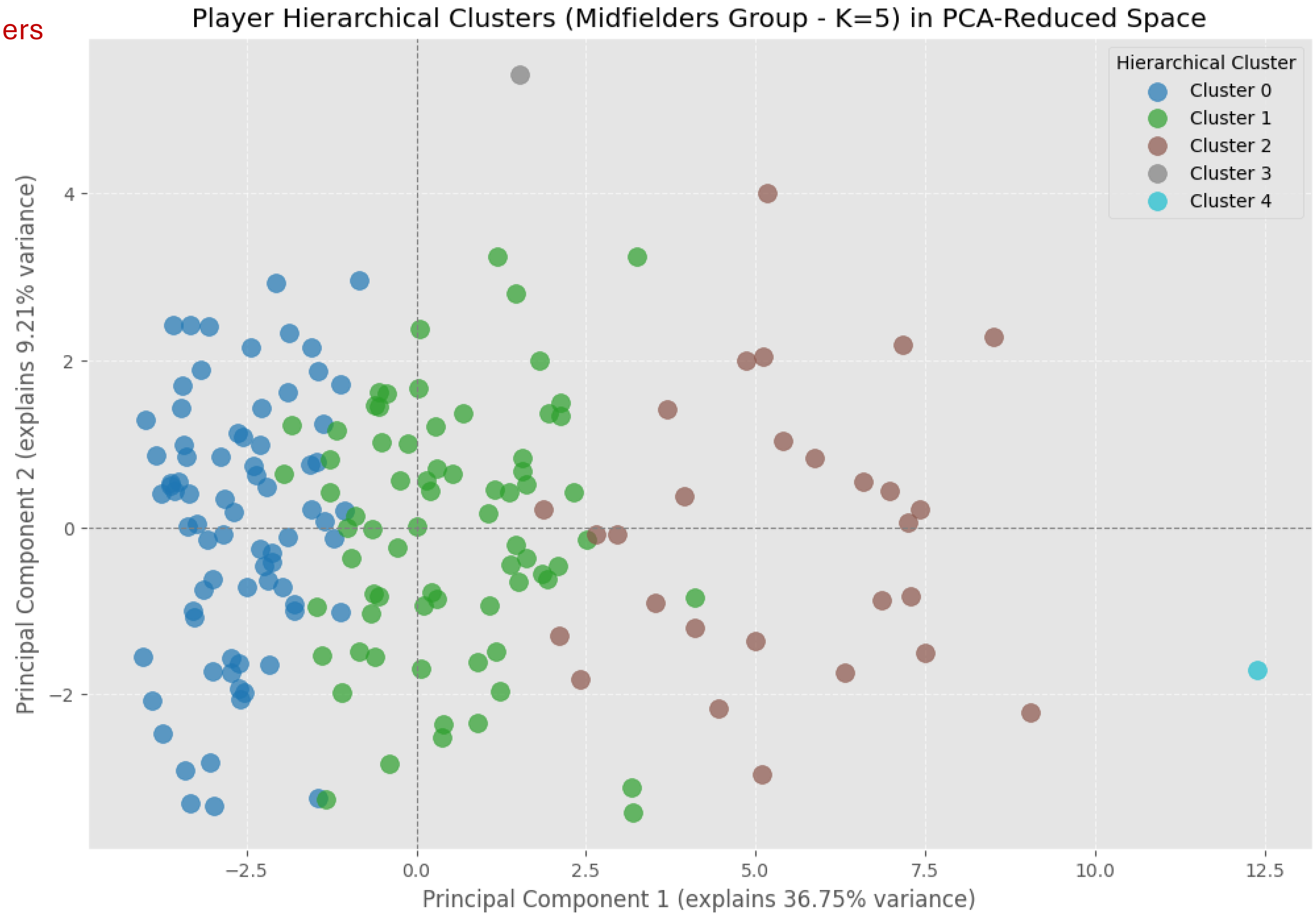
Cluster 3 Profile =
Shooting & Tackling Defenders

Cluster 4 Profile =
Outlier

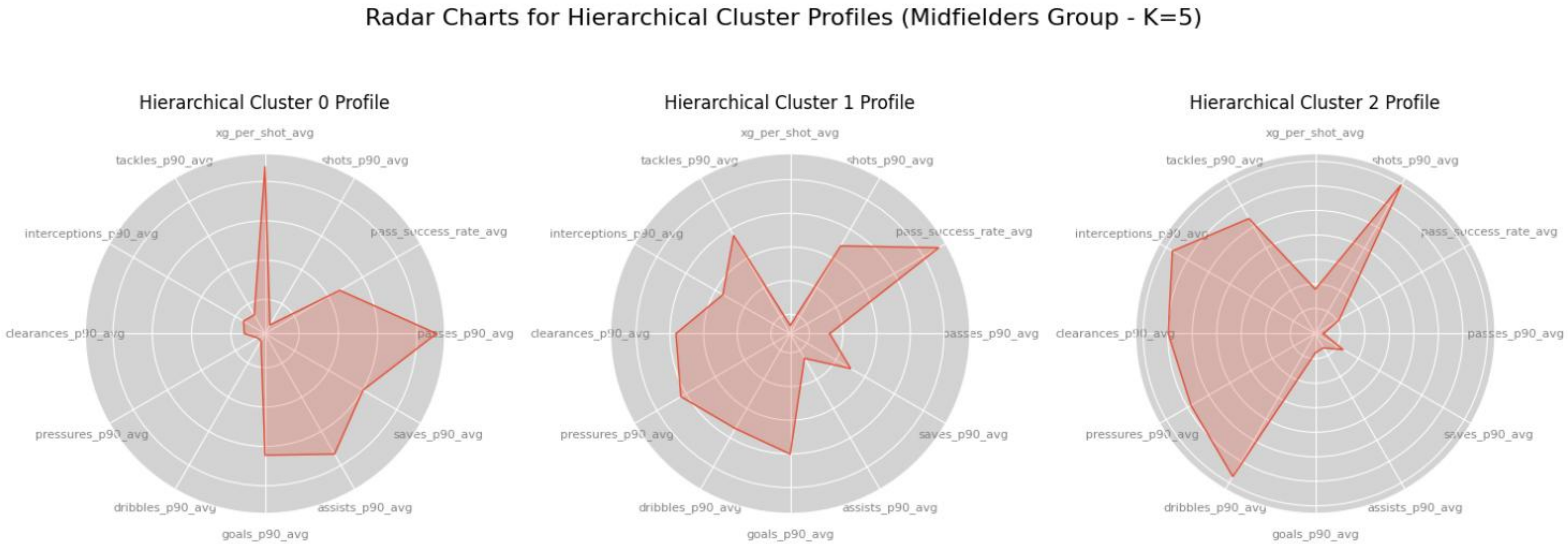
Midfielders



Midfielders



Midfielders

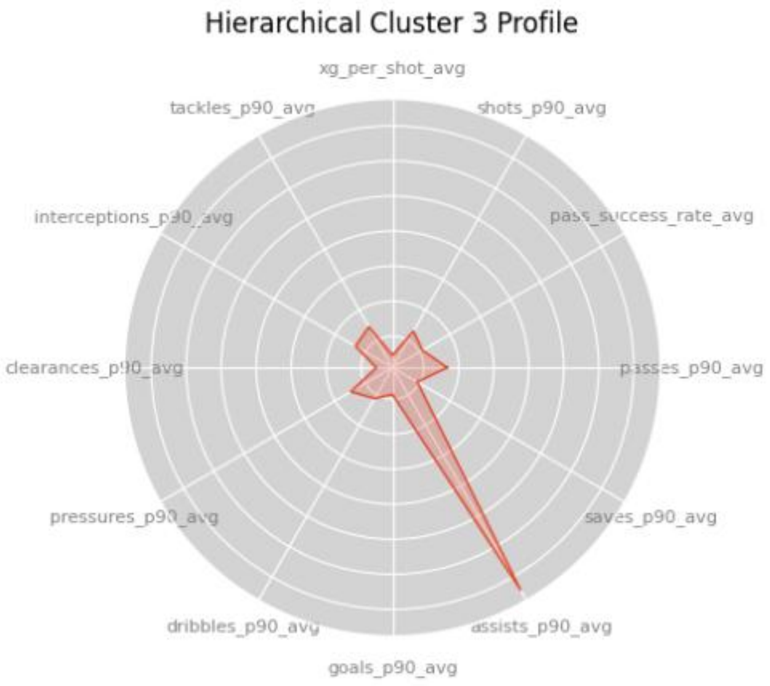


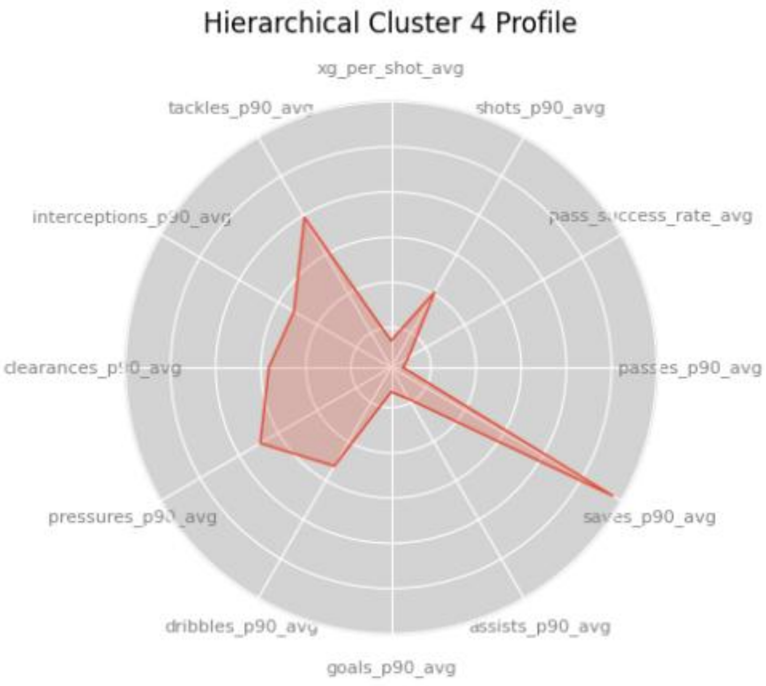
Cluster 0 Profile =
Passing Midfielders

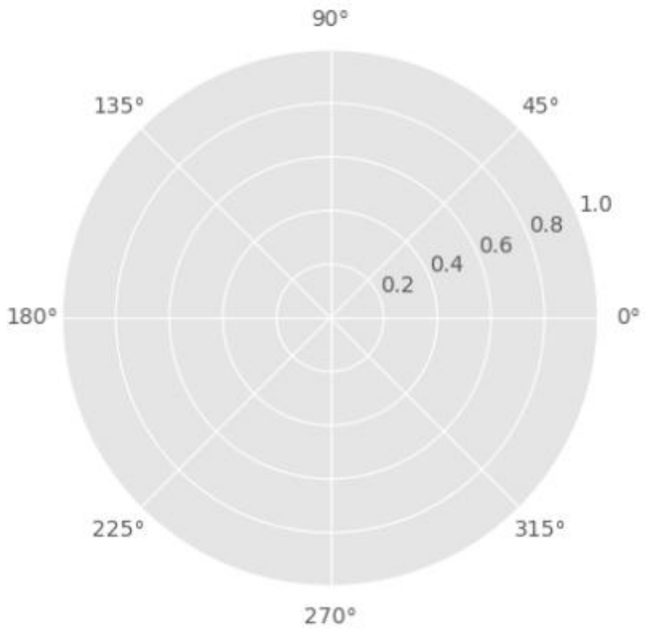
Cluster 1 Profile =
Successful Passing Midfielders

Cluster 2 Profile =
Defensive & Shooting Midfielders

Midfielders





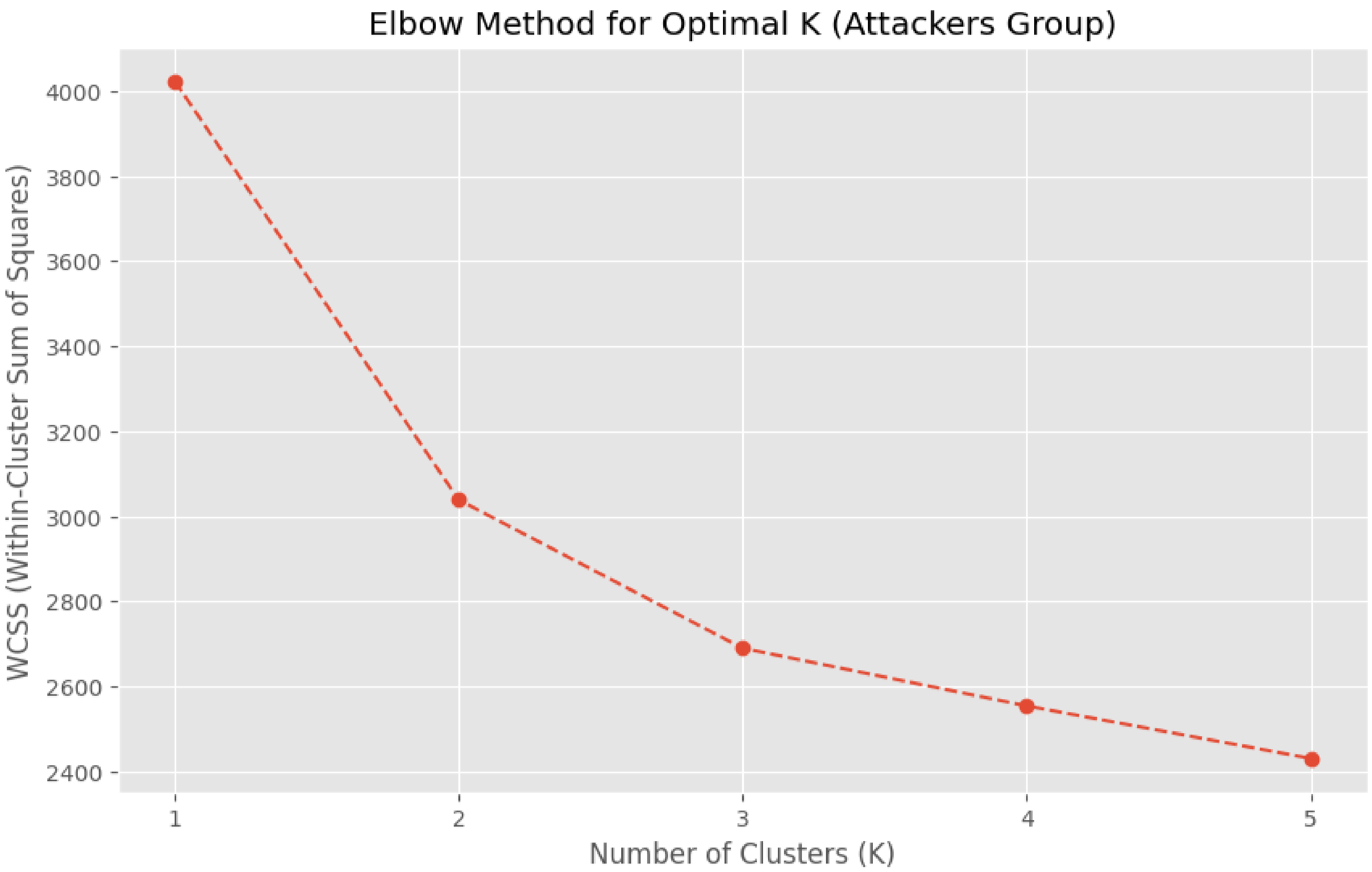


Cluster 3 Profile =
(Outlier) Assisting Midfielders

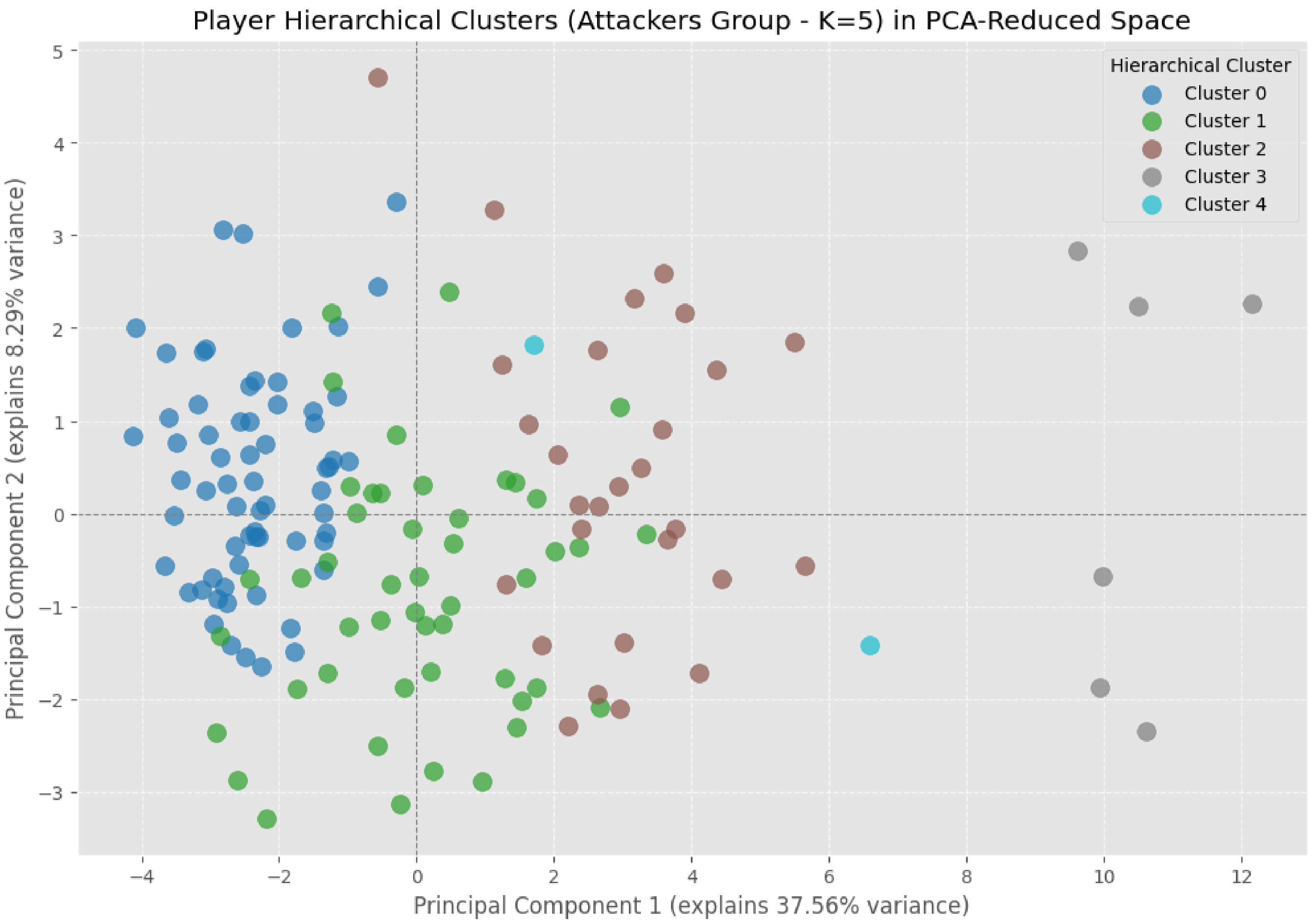
Cluster 4 Profile =
(Outlier)

Attackers

--- Step 3: Determining Optimal K using Elbow Method for Attackers ---

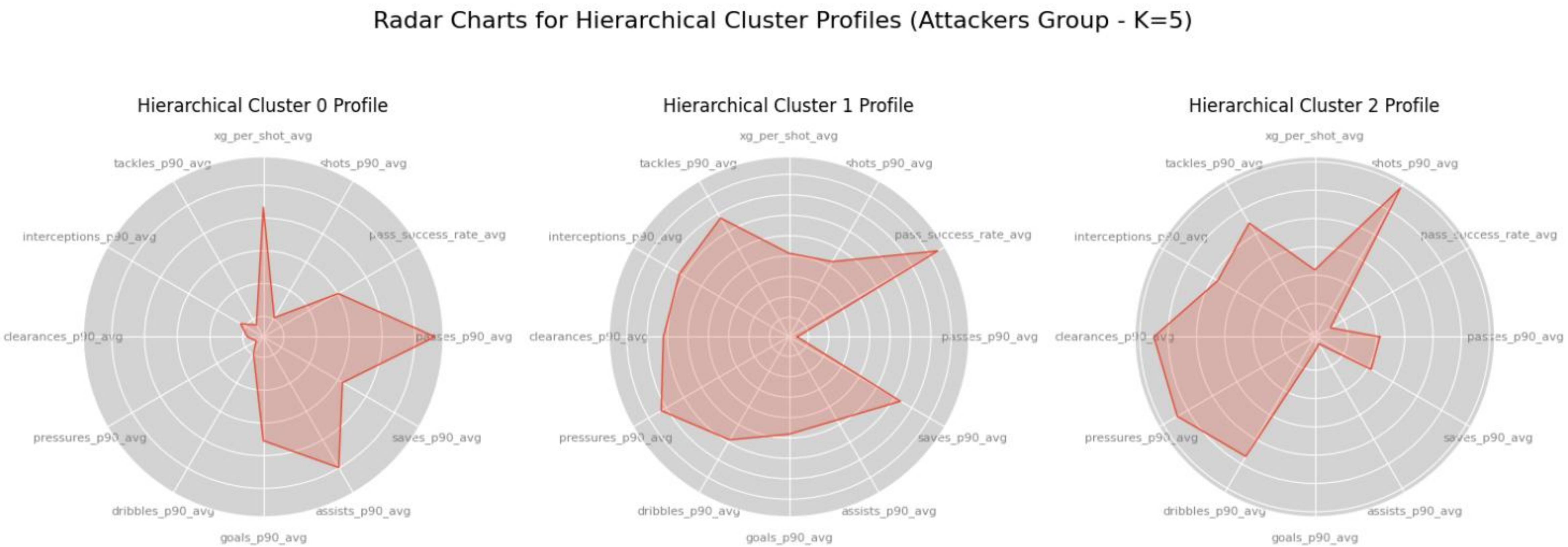


Attackers



Player Hierarchical Clusters (Attackers Group - K=5) in PCA-Reduced Space
Hierarchical Cluster
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Principal Component 1 (explains 37.56% variance)
Principal Component 2 (explains 8.29% variance)
−4
−2
0
2
4
6
8
10
12
5
4
3
2
1
0
−1
−2
−3

Attackers

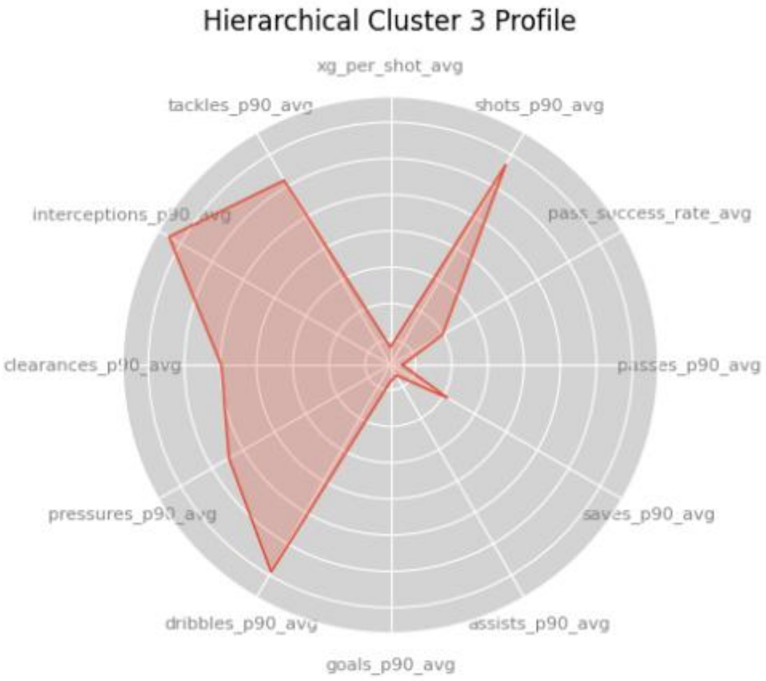


Cluster 0 Profile =
Passing Attackers

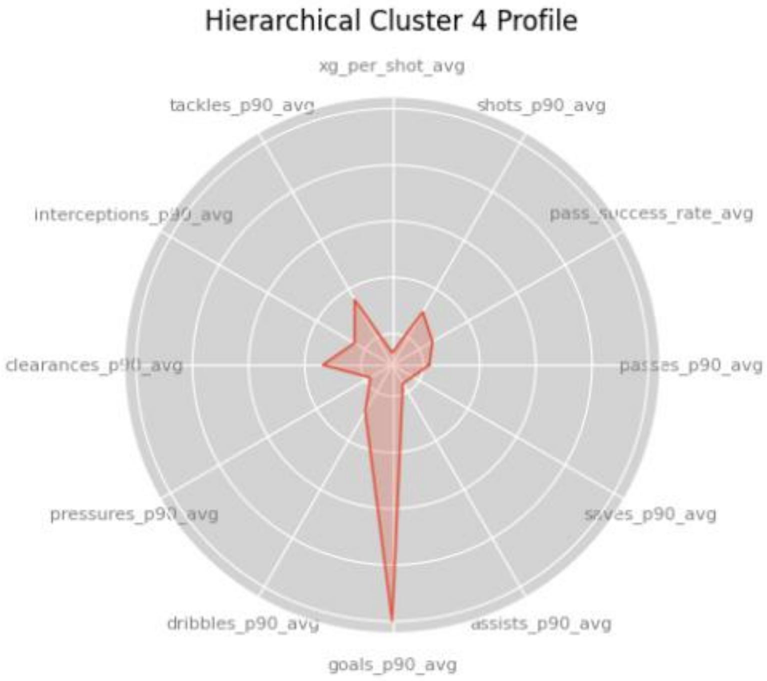
Cluster 1 Profile =
High Press Attackers

Cluster 2 Profile =
Pressure Attackers

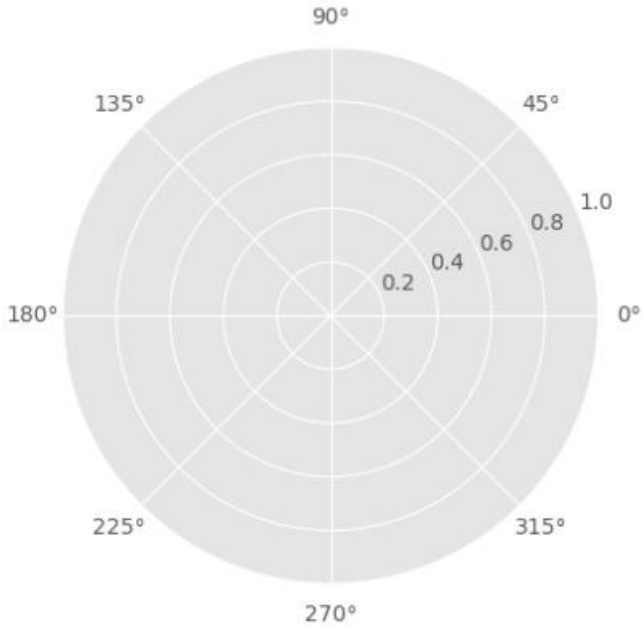
Attackers



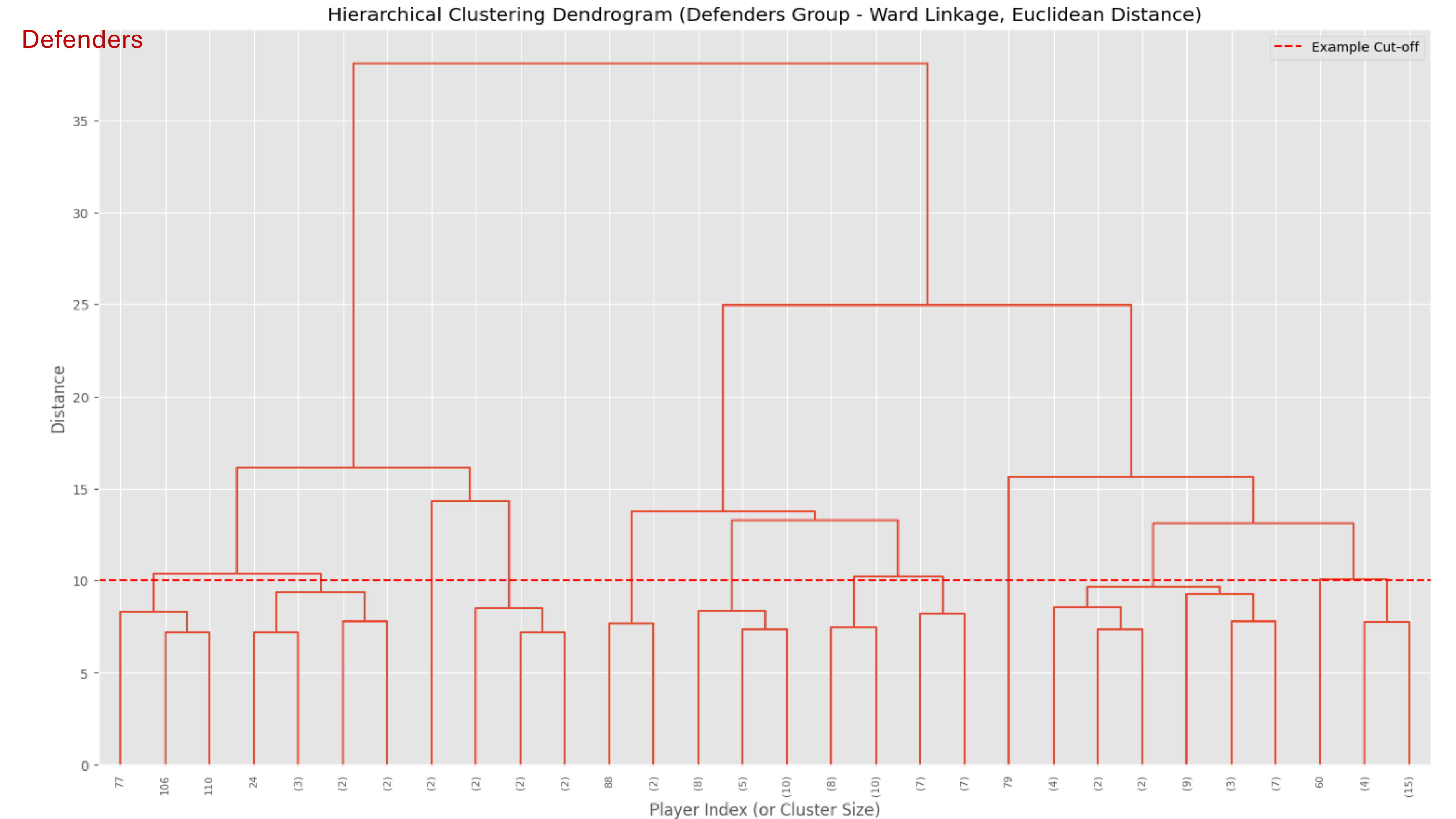
Cluster 3 Profile =
Dribbling Attackers



Cluster 4 Profile =
Outlier



Appendix 2. Hierarchical Agglomerative Clustering



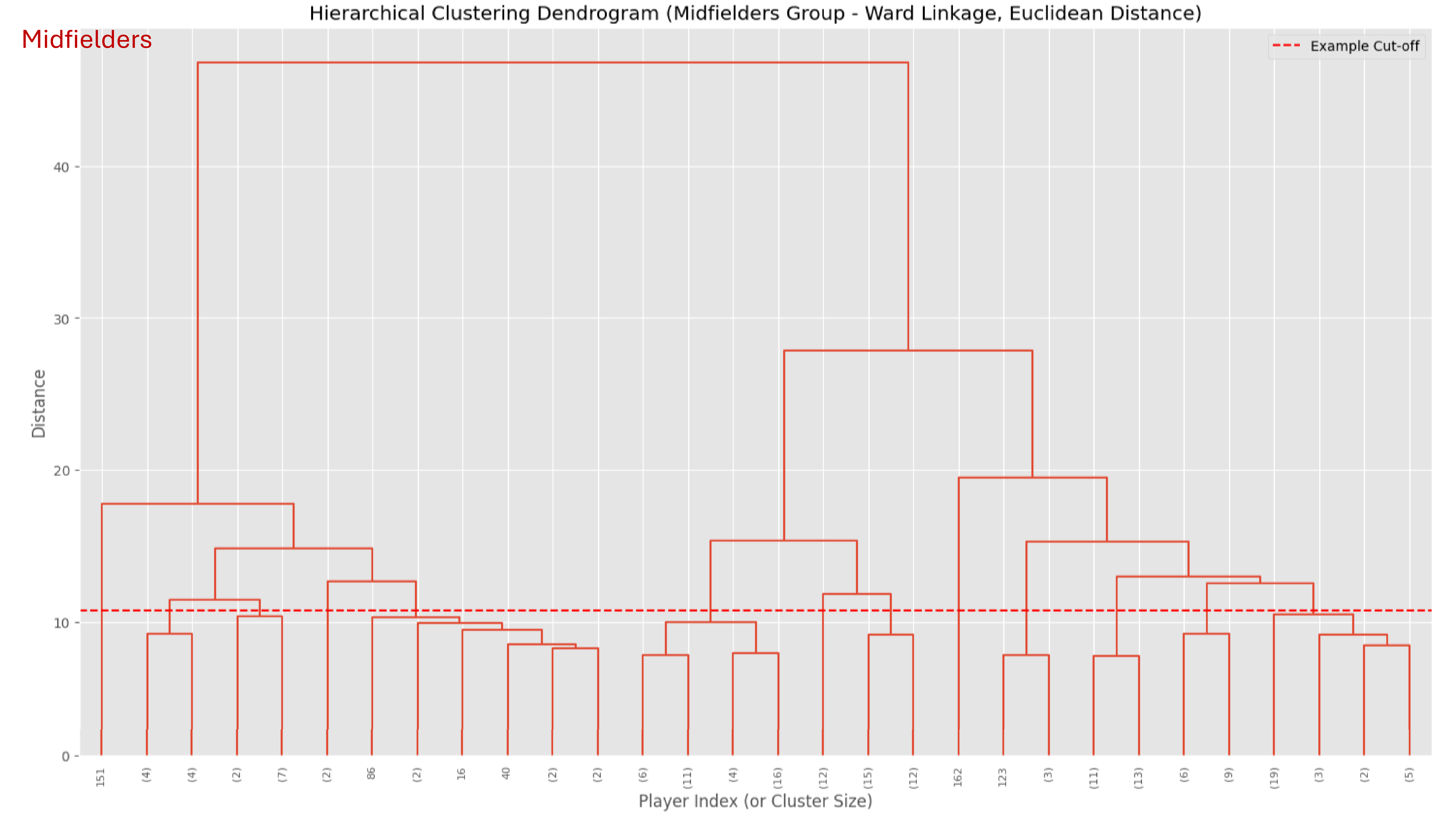
Hierarchical Clustering Dendrogram (Defenders Group - Ward Linkage, Euclidean Distance)
Defenders
Example Cut-off
Distance
0
5
10
15
20
25
30
35
77
106
110
24
(3)
(2)
(2)
(2)
(2)
(2)
(2)
88
(2)
(8)
(5)
(10)
(8)
(10)
(7)
(7)
79
(4)
(2)
(2)
(9)
(3)
(7)
60
(4)
(15)
Player Index (or Cluster Size)

Defenders

```
--- Step 5: Applying Hierarchical Clustering with K = 5 for Defenders ---
Hierarchical clustering applied for Defenders group. 5 clusters identified.

--- Step 6: Interpreting Hierarchical Clusters - Player Role Profiles for Defenders ---

Average (Unscaled) P90 Stats for Each Hierarchical Cluster (Defenders Group Profiles):
   hierarchical_cluster  passes_p90_avg  progressive_pass_distance_p90_avg  \
0                     0       87.102505                        1425.726255
1                     1       88.545461                        1386.378126
2                     2       89.320804                         913.776136
3                     3       91.610137                        2292.819426
4                     4       92.776153                        1055.399808

   passes_into_final_third_p90_avg  key_passes_p90_avg  short_passes_p90_avg  \
0                        11.662796           10.043173             63.645328
1                        12.534991            6.577231             66.078320
2                        13.193213            4.423827             66.630947
3                        12.895684            9.723841             69.741204
4                        14.652253            4.507071             69.228707

   long_passes_p90_avg  through_balls_p90_avg  shots_p90_avg  \
0            11.374600              12.082578       6.994133
1            11.717968              10.749173       6.622470
2            11.525853              11.164004       4.322319
3            11.200284              10.668649      10.646584
4            13.325816              10.221629       7.501914

   shots_on_target_p90_avg  ...  fouls_committed_p90_avg  fouls_won_p90_avg  \
0                 3.199872  ...                 8.457678           8.647107
1                 3.539317  ...                 5.252385           5.242966
2                 2.087121  ...                 3.545957           3.381322
3                 6.525876  ...                10.021730           8.660163
4                 3.644636  ...                 6.334642           7.693968
```

Defenders

```
   goals_p90_avg  assists_p90_avg  saves_p90_avg  aerial_duels_won_p90_avg  \
0       0.296801         0.013594       0.447174                 14.553094   
1       0.035432         0.004059       0.932029                  9.731207   
2       0.051079         0.016920       0.910753                  6.240501   
3       0.007880         0.012000       1.945818                 16.034476   
4       0.045202         0.818182       2.058703                  6.089853   

   pass_success_rate_avg  xg_per_shot_avg  player_average_x_avg  \
0               0.796545         0.178501             49.810718   
1               0.798766         0.165173             47.670630   
2               0.798015         0.157943             46.116756   
3               0.795664         0.153033             45.464773   
4               0.805273         0.187818             41.969091   

   player_average_y_avg  
0             36.606629  
1             33.946672  
2             34.460962  
3             34.966430  
4             36.750909  

[5 rows x 28 columns]
```



Hierarchical Clustering Dendrogram (Midfielders Group - Ward Linkage, Euclidean Distance)
Midfielders
Example Cut-off
Distance
0
10
20
30
40
Player Index (or Cluster Size)
151
(4)
(4)
(2)
(7)
(2)
86
(2)
16
40
(2)
(2)
(6)
(11)
(4)
(16)
(12)
(15)
(12)
162
123
(3)
(11)
(13)
(6)
(9)
(19)
(3)
(2)
(5)

Midfielders

```
--- Step 6: Interpreting Hierarchical Clusters - Player Role Profiles for Midfielders ---

Average (Unscaled) P90 Stats for Each Hierarchical Cluster (Midfielders Group Profiles):
   hierarchical_cluster  passes_p90_avg  progressive_pass_distance_p90_avg  \
0                     0       89.980474                         904.107450   
1                     1       89.439772                        1397.804538   
2                     2       88.920768                        2039.054159   
3                     3       94.802082                        2157.971300   
4                     4       84.798772                        3566.810465   

   passes_into_final_third_p90_avg  key_passes_p90_avg  short_passes_p90_avg  \
0                        13.033393            4.512564             67.425813   
1                        12.737943            6.746100             66.413392   
2                        12.476677           10.553899             66.366778   
3                        15.166080           10.058174             71.047063   
4                        11.180620           16.895618             64.133107   

   long_passes_p90_avg  through_balls_p90_avg  shots_p90_avg  \
0            11.240793              11.313868       4.094475   
1            11.489306              11.537075       6.568862   
2            11.428182              11.125808      10.430148   
3            10.038010              13.717008       7.899591   
4            10.137598              10.528066      12.108950   

   shots_on_target_p90_avg  ...  fouls_committed_p90_avg  fouls_won_p90_avg  \
0                 2.135139  ...                 3.516987           3.716409   
1                 3.197267  ...                 5.590773           5.575776   
2                 4.814188  ...                 8.483806           8.656715   
3                 6.446880  ...                 4.043466           4.664597   
4                 6.455650  ...                11.908975           8.105487   

   goals_p90_avg  assists_p90_avg  saves_p90_avg  aerial_duels_won_p90_avg  \
0       0.025250         0.011987       0.715626                  6.376367   
1       0.041883         0.011181       1.068345                  9.296787   
2       0.029714         0.008991       1.098142                 15.137588   
3       0.020284         1.764706       0.777561                 13.747513   
4       0.000000         0.000000      14.603226                 19.388801   
```

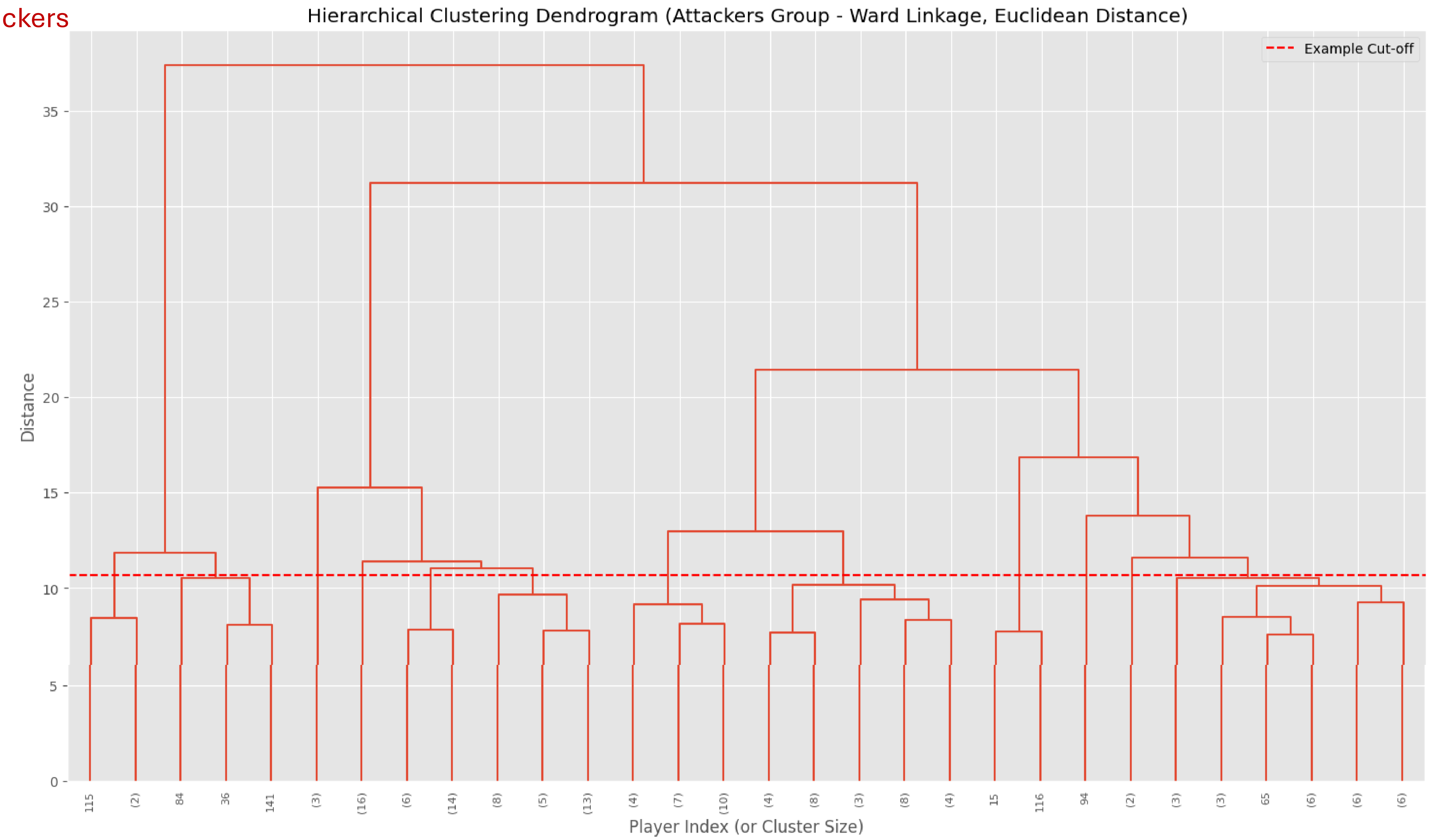
Midfielders

	pass_success_rate_avg	xg_per_shot_avg	player_average_x_avg	\
0	0.796814	0.165267	49.363108	
1	0.804674	0.161904	49.303111	
2	0.800922	0.167220	50.073379	
3	0.802745	0.146373	47.900000	
4	0.790400	0.154700	50.656000	

	player_average_y_avg
0	34.026311
1	34.453742
2	34.488097
3	33.698039
4	36.172000

[5 rows x 28 columns]

Attackers



Hierarchical Clustering Dendrogram (Attackers Group - Ward Linkage, Euclidean Distance)
Example Cut-off
Distance
35
30
25
20
15
10
5
0
115
(2)
84
36
141
(3)
(16)
(6)
(14)
(8)
(5)
(13)
(4)
(7)
(10)
(4)
(8)
(3)
(8)
(4)
15
116
94
(2)
(3)
(3)
65
(6)
(6)
(6)
Player Index (or Cluster Size)

Attackers

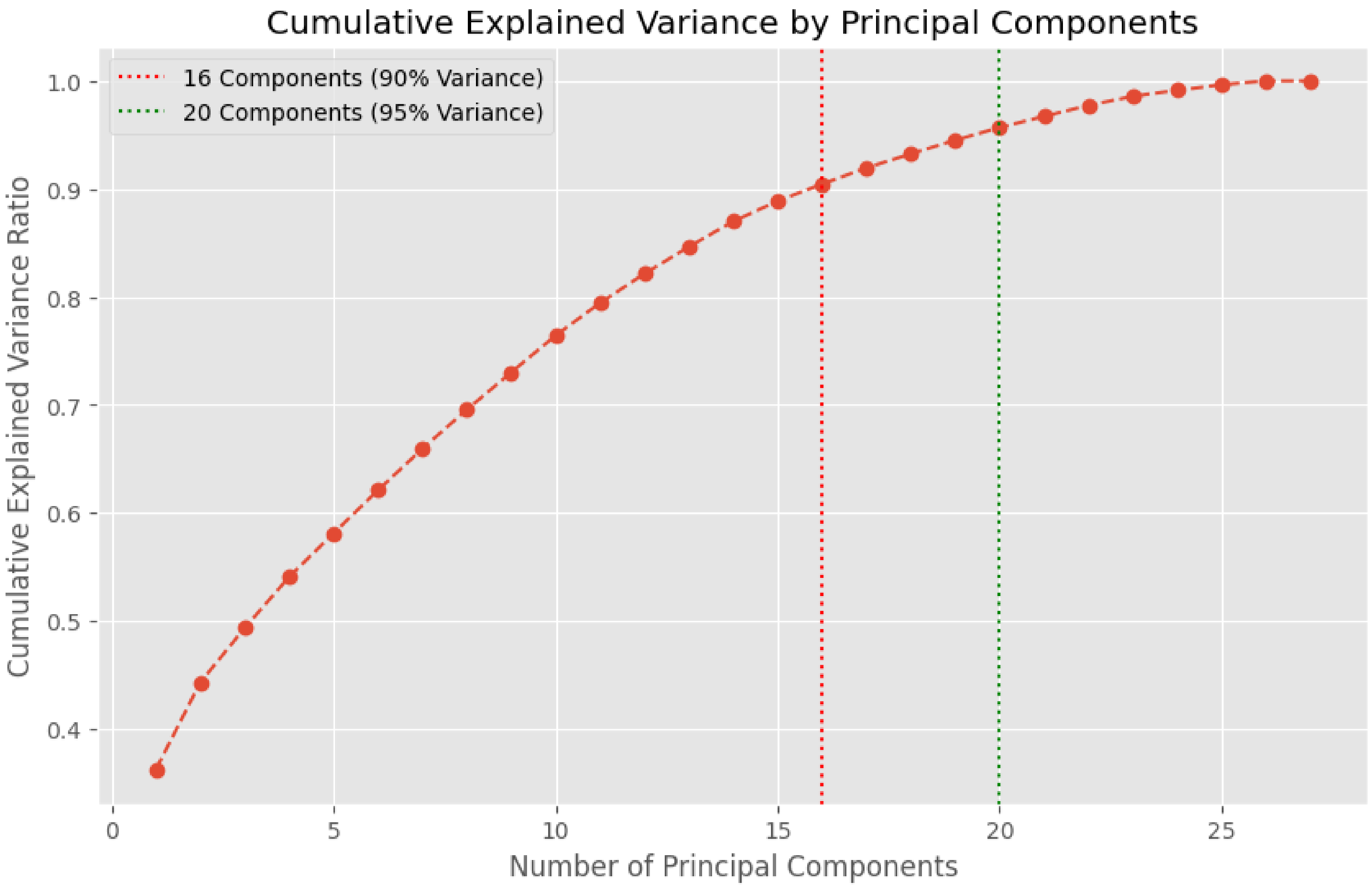
```
--- Step 6: Interpreting Hierarchical Clusters - Player Role Profiles for Attackers ---

Average (Unscaled) P90 Stats for Each Hierarchical Cluster (Attackers Group Profiles):
   hierarchical_cluster  passes_p90_avg  progressive_pass_distance_p90_avg  \
0                     0       90.460286                         906.203501
1                     1       87.186536                        1299.249205
2                     2       90.089376                        1826.975808
3                     3       88.588289                        2618.439728
4                     4       90.021179                        1635.759964

   passes_into_final_third_p90_avg  key_passes_p90_avg  short_passes_p90_avg  \
0                        13.476219            4.752736             67.430429
1                        12.761032            6.658470             65.759500
2                        12.865527            7.673762             66.110395
3                        12.952518           16.543581             64.887488
4                        11.526287            9.095740             66.505595

   long_passes_p90_avg  through_balls_p90_avg  shots_p90_avg  \
0            11.544201              11.485656       4.424309
1            10.825426              10.601610       5.701106
2            11.790350              12.188631       9.189757
3            12.476955              11.223847      14.862116
4            11.369690              12.145894       9.442343

   shots_on_target_p90_avg  ...  fouls_committed_p90_avg  fouls_won_p90_avg  \
0                 2.111839  ...                 3.726141           3.760920
1                 2.764844  ...                 5.014848           5.121832
2                 5.202430  ...                 7.101836           6.785717
3                 7.875861  ...                11.028391          11.526195
4                 5.052271  ...                 5.226972           8.187361

   goals_p90_avg  assists_p90_avg  saves_p90_avg  aerial_duels_won_p90_avg  \
0       0.039217         0.011897       0.850750                  6.395642
1       0.034437         0.004933       1.245038                  9.392391
2       0.032300         0.003802       1.614254                 12.695045
3       0.034847         0.003797       2.272862                 23.120552
4       1.458698         0.000000       0.642797                 17.439435
```

Attackers

```
   pass_success_rate_avg  xg_per_shot_avg  player_average_x_avg  \
0               0.797026         0.167283             51.483591   
1               0.801840         0.162856             52.142158   
2               0.797986         0.170276             53.254496   
3               0.804405         0.164342             52.187077   
4               0.805046         0.154179             53.514424   

   player_average_y_avg  
0             35.354846  
1             33.550242  
2             34.892962  
3             35.040776  
4             36.126866  

[5 rows x 28 columns]
```

Appendix 3. Dimensionality Reduction (PCA) followed by K-Means Clustering

All Players



The graph indicates that reducing the analysis down to 16 player attributes will cover 90% of the data.

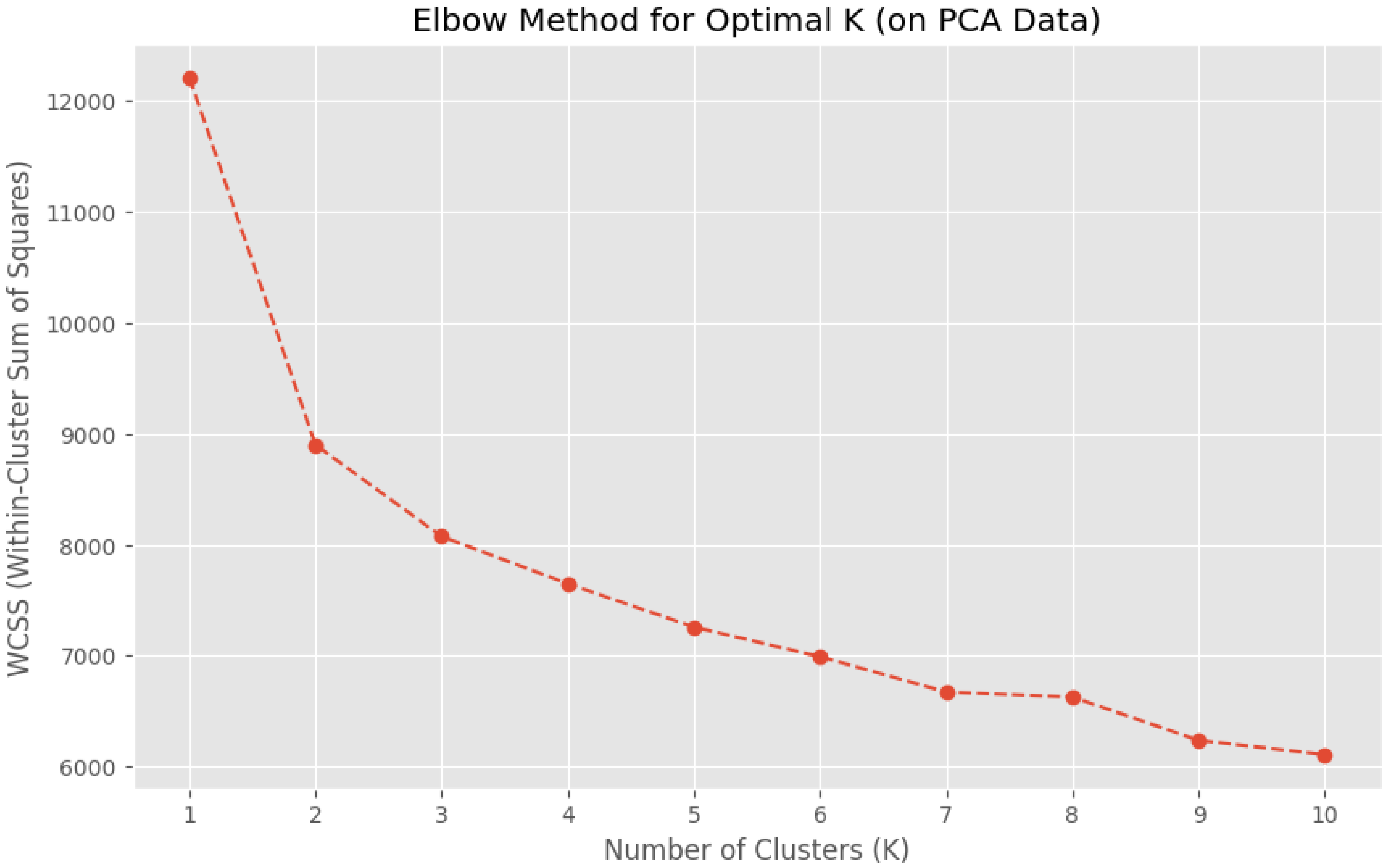
Proposed reduced attributes:

All Players

```
Choosing 16 components to explain at least 90% of the variance.
Data reduced to 16 principal components.
Total variance explained by 16 components: 90.45%

First 5 rows of PCA-transformed data:
        PC1       PC2       PC3       PC4       PC5       PC6       PC7  \
0 -1.355154  0.542808 -0.568607 -0.740069  0.960271 -0.492207  0.904291
1 -2.479231 -1.636707  0.171132 -1.559970  0.216546 -1.415808  1.990674
2 -1.219831  0.003976 -0.078722 -1.399139 -0.616900 -2.273230 -1.105552
3  2.898293 -0.201247 -0.550764  0.038199 -0.776348  0.171285 -1.719444
4 -1.840498  0.965142  0.184873 -0.508857  0.142508 -0.400520 -0.612956

        PC8       PC9      PC10      PC11      PC12      PC13      PC14  \
0 -1.389154 -0.415963  0.281390  0.836873 -0.617269 -0.244182  0.445545
1 -1.582835  0.784925  0.363578  1.988404  0.397200  0.805389  0.260309
2  0.004633 -0.023092  1.475459  0.976967  0.922590 -0.179838  1.093242
3  1.010762 -0.332277  0.376879  0.206040  0.747590 -1.210135 -1.220878
4 -0.763550 -0.715268  1.362252  0.728449 -0.356206 -0.842242  0.109863

       PC15      PC16
0  0.306044 -0.343308
1  0.234577  0.289136
2  0.724470  0.253484
3 -1.216191 -0.143869
4 -0.085528  0.210919

--- Interpretation of Principal Components (Loadings) ---
Top 5 features contributing to each Principal Component:
PC1: ['progressive_carries_p90_avg', 'dribbles_p90_avg', 'aerial_duels_won_p90_avg', 'shots_p90_avg', 'interceptions_p90_avg']
PC2: ['passes_p90_avg', 'short_passes_p90_avg', 'passes_into_final_third_p90_avg', 'long_passes_p90_avg', 'through_balls_p90_avg']
PC3: ['saves_p90_avg', 'player_average_x_avg', 'xg_per_shot_avg', 'goals_p90_avg', 'player_average_y_avg']
PC4: ['through_balls_p90_avg', 'saves_p90_avg', 'player_average_y_avg', 'long_passes_p90_avg', 'goals_p90_avg']
PC5: ['long_passes_p90_avg', 'assists_p90_avg', 'goals_p90_avg', 'pass_success_rate_avg', 'xg_per_shot_avg']
PC6: ['pass_success_rate_avg', 'assists_p90_avg', 'xg_per_shot_avg', 'player_average_x_avg', 'saves_p90_avg']
PC7: ['long_passes_p90_avg', 'short_passes_p90_avg', 'player_average_y_avg', 'assists_p90_avg', 'pass_success_rate_avg']
PC8: ['xg_per_shot_avg', 'pass_success_rate_avg', 'player_average_y_avg', 'goals_p90_avg', 'through_balls_p90_avg']
PC9: ['goals_p90_avg', 'pass_success_rate_avg', 'player_average_y_avg', 'assists_p90_avg', 'player_average_x_avg']
PC10: ['assists_p90_avg', 'through_balls_p90_avg', 'player_average_y_avg', 'goals_p90_avg', 'pass_success_rate_avg']
PC11: ['player_average_x_avg', 'xg_per_shot_avg', 'player_average_y_avg', 'long_passes_p90_avg', 'assists_p90_avg']
PC12: ['passes_into_final_third_p90_avg', 'carries_into_final_third_p90_avg', 'through_balls_p90_avg', 'successful_dribbles_p90_avg', 'shots_on_target_p90_avg']
PC13: ['passes_into_final_third_p90_avg', 'shots_on_target_p90_avg', 'carries_into_final_third_p90_avg', 'shots_p90_avg', 'passes_p90_avg']
PC14: ['successful_dribbles_p90_avg', 'shots_on_target_p90_avg', 'shots_p90_avg', 'dribbles_p90_avg', 'carries_into_final_third_p90_avg']
PC15: ['saves_p90_avg', 'key_passes_p90_avg', 'clearances_p90_avg', 'fouls_committed_p90_avg', 'aerial_duels_won_p90_avg']
PC16: ['key_passes_p90_avg', 'saves_p90_avg', 'aerial_duels_won_p90_avg', 'successful_dribbles_p90_avg', 'progressive_pass_distance_p90_avg']
```

All Players

```
--- Step 2: K-Means Clustering on PCA-reduced Data ---
```



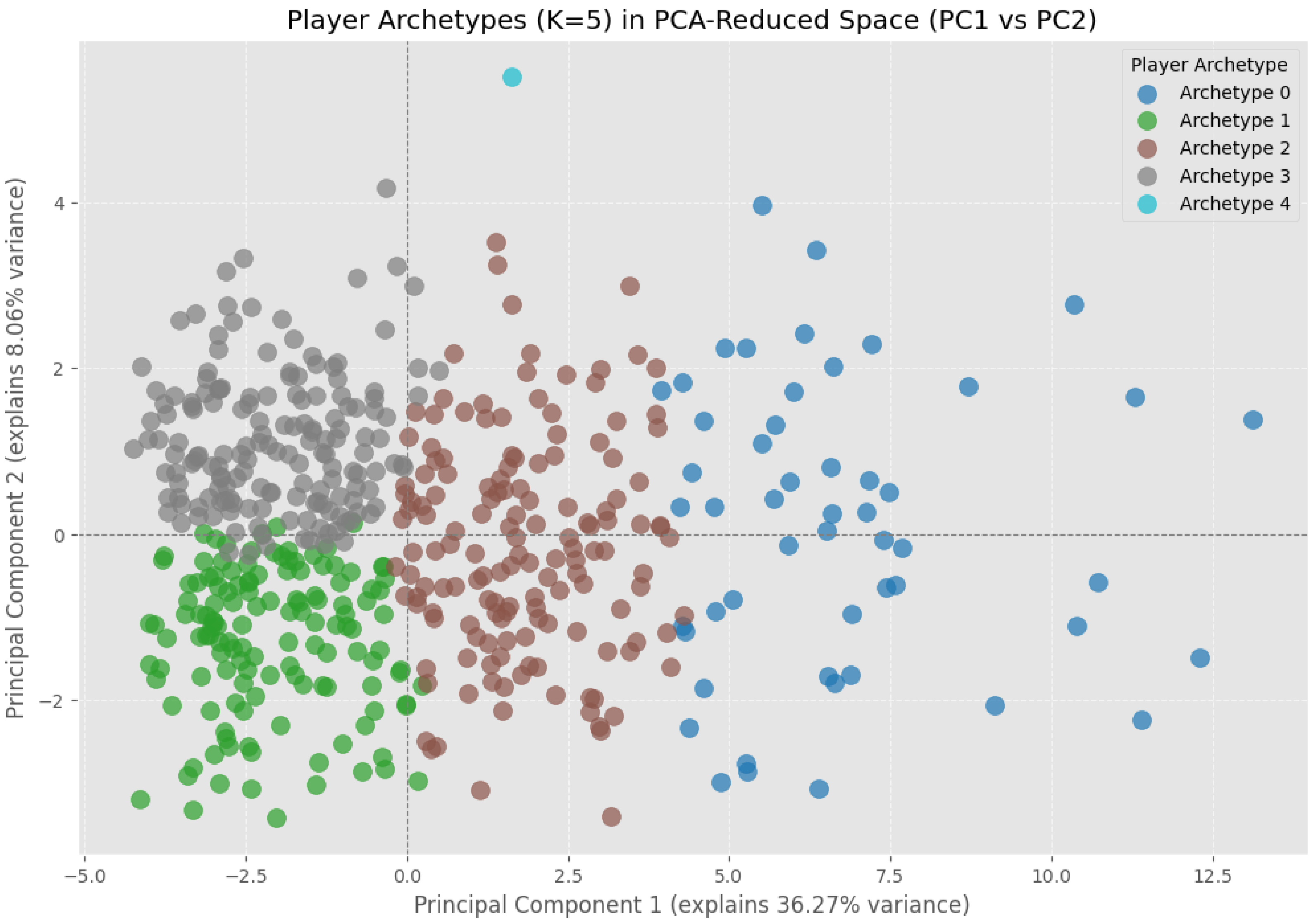
All Players

```
Average (Scaled) P90 Stats for Each Player Archetype:
                   passes_p90_avg  progressive_pass_distance_p90_avg  \
archetype_cluster
0                       -0.099692                           1.605678
1                       -0.808830                          -0.599380
2                       -0.103856                           0.527788
3                        0.787649                          -0.508171
4                        1.463906                           1.612586

                   passes_into_final_third_p90_avg  key_passes_p90_avg  \
archetype_cluster
0                                        -0.298327            1.720747
1                                        -0.390013           -0.552369
2                                        -0.226431            0.432235
3                                         0.614061           -0.494412
4                                         1.811050            1.277189

                   short_passes_p90_avg  long_passes_p90_avg  \
archetype_cluster
0                             -0.085617             0.129194
1                             -0.725666            -0.267141
2                             -0.142084            -0.037427
3                              0.750215             0.219572
4                              1.302581            -0.867721

                   through_balls_p90_avg  shots_p90_avg  \
archetype_cluster
0                              -0.189478       1.779364
1                              -0.106679      -0.549101
2                               0.099236       0.473915
3                               0.048177      -0.550052
4                               1.615479       0.646811

                   shots_on_target_p90_avg  tackles_p90_avg  ...  \
archetype_cluster                                            ...
0                                 1.427451         1.759310  ...
1                                -0.481191        -0.541569  ...
2                                 0.452989         0.422486  ...
3                                -0.482734        -0.504627  ...
4                                 1.923700         0.931448  ...
```

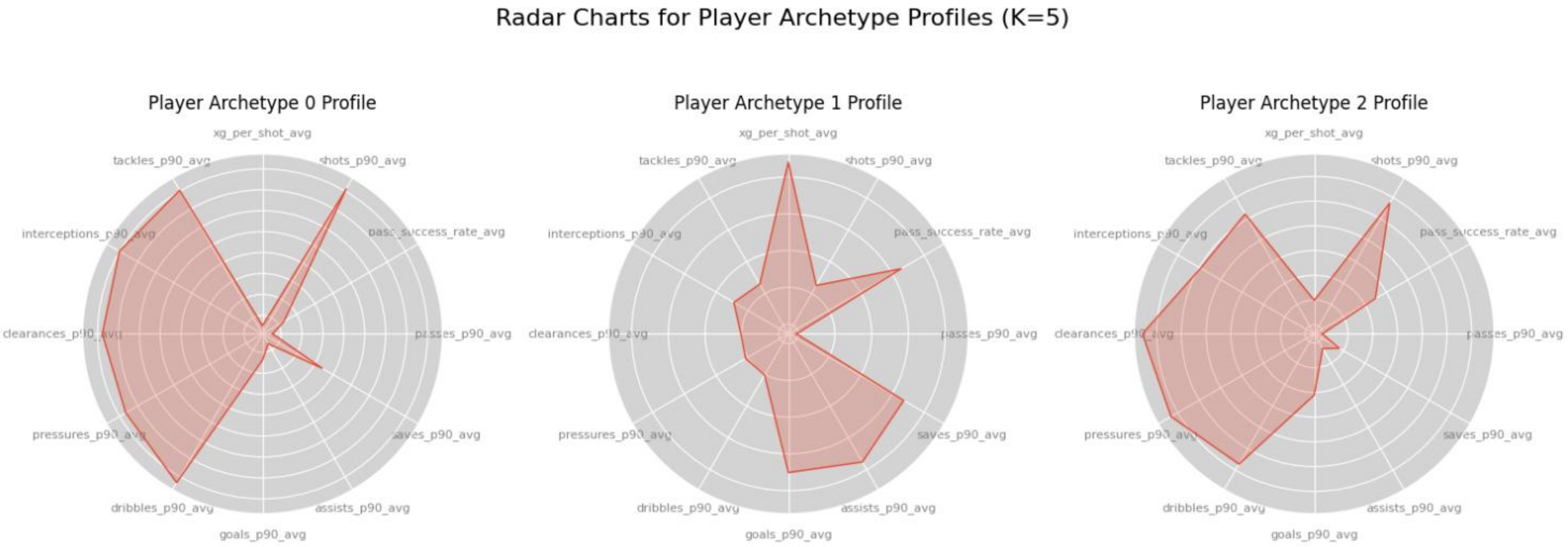
All Players

```
                  fouls_committed_p90_avg  fouls_won_p90_avg  goals_p90_avg  \
archetype_cluster
0                                1.799234           1.764301       0.090823
1                               -0.538729          -0.600140      -0.098462
2                                0.377468           0.488286       0.110638
3                               -0.470096          -0.510889      -0.047421
4                               -0.448668          -0.204380      -0.185853

                  assists_p90_avg  saves_p90_avg  aerial_duels_won_p90_avg  \
archetype_cluster
0                       -0.085507       0.602681                  1.826328
1                       -0.048066      -0.128441                 -0.584612
2                       -0.067918      -0.019960                  0.424471
3                        0.011837      -0.066312                 -0.493142
4                       19.125507      -0.217007                  1.089651

                  pass_success_rate_avg  xg_per_shot_avg  \
archetype_cluster
0                              0.072048        -0.123917
1                             -0.144239         0.081363
2                              0.148776        -0.001682
3                             -0.041325        -0.020108
4                              0.293932        -1.011979

                  player_average_x_avg  player_average_y_avg
archetype_cluster
0                            -0.067882              0.154632
1                            -0.107269              0.105727
2                             0.188722             -0.045522
3                            -0.060089             -0.093177
4                            -0.313929             -0.343430

[5 rows x 27 columns]
```

All Players

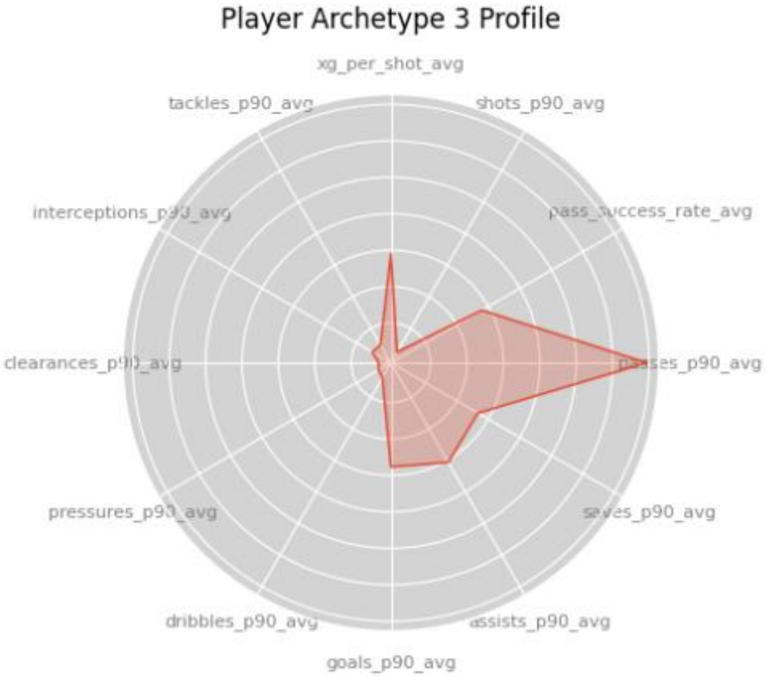

Player Archetypes (K=5) in PCA-Reduced Space (PC1 vs PC2)
Player Archetype
Archetype 0
Archetype 1
Archetype 2
Archetype 3
Archetype 4
Principal Component 2 (explains 8.06% variance)
Principal Component 1 (explains 36.27% variance)
4
2
0
−2
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5

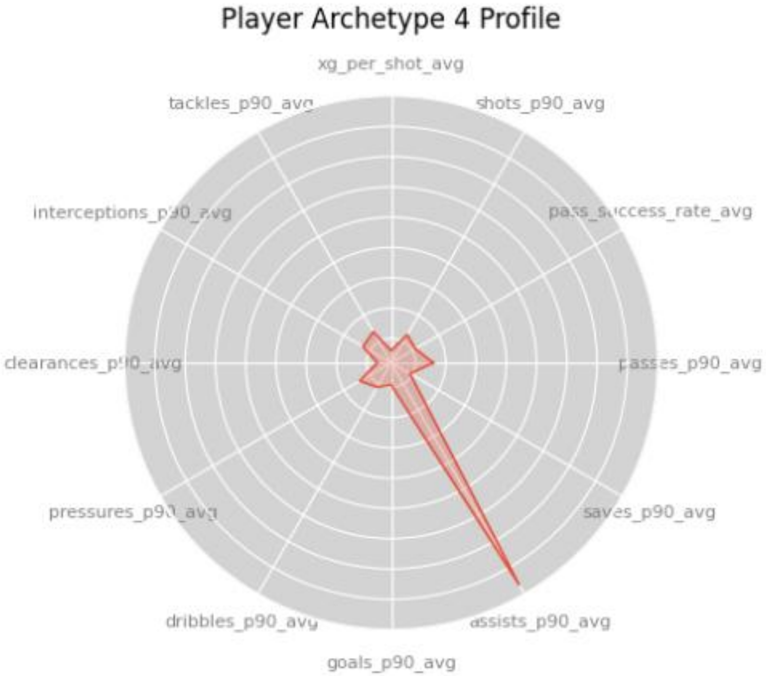
All Players

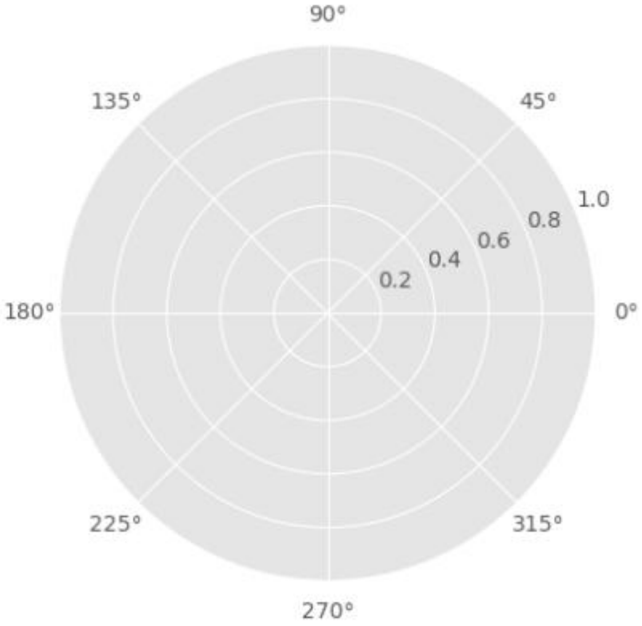


Radar Charts for Player Archetype Profiles (K=5)
Player Archetype 0 Profile
Player Archetype 1 Profile
Player Archetype 2 Profile
xg_per_shot_avg
shots_p90_avg
pass_success_rate_avg
passes_p90_avg
saves_p90_avg
assists_p90_avg
goals_p90_avg
dribbles_p90_avg
pressures_p90_avg
clearances_p90_avg
interceptions_p90_avg
tackles_p90_avg

All Players








```
--- Summary of PCA and K-Means for Player Archetypes ---
PCA reduced the dimensionality of player performance data to 16 principal components, explaining 90.45% of the total variance.
K-Means clustering on these principal components identified 5 distinct player archetypes.

To interpret these archetypes:
1. Examine the 'Interpretation of Principal Components (Loadings)' section to understand what the principal components represent (e.g., PC1 = 'attacking strength', PC2 = 'defensive work rate').
2. Review the 'Average (Scaled) P90 Stats for Each Player Archetype' and the radar charts. These show the average performance profile of players within each archetype across the original scaled metrics.

Consider adjusting `n_components` for PCA and `optimal_k_archetypes` for K-Means to explore different levels of detail in player archetypes.
```

Defenders

```
================================================================================
1) Distinct Defensive Roles
================================================================================

  Archetype 0 (Defensive Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Progressive Carries: 1.85 (scaled)
        - Dribbles: 1.83 (scaled)
        - Aerial Duels Won: 1.83 (scaled)
        - Fouls Committed: 1.80 (scaled)
        - Shots: 1.78 (scaled)
      Lower than average in:
        - Short Passes: -0.09 (scaled)
        - Passes: -0.10 (scaled)
        - Xg Per Shot: -0.12 (scaled)
        - Through Balls: -0.19 (scaled)
        - Passes Into Final Third: -0.30 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Dominik Casemiro (Pos: Centre Back, Score: 0.00)
      - Raphael Tielemans (Pos: Centre Back, Score: 0.00)
      - Liam Zinchenko (Pos: Left Back, Score: 0.00)
      - Lewis Enzo (Pos: Left Back, Score: 0.00)
      - Declan Rice (Pos: Right Back, Score: 0.00)

  Archetype 1 (Defensive Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Player Average Y: 0.11 (scaled)
        - Xg Per Shot: 0.08 (scaled)
      Lower than average in:
        - Clearances: -0.60 (scaled)
        - Progressive Pass Distance: -0.60 (scaled)
        - Fouls Won: -0.60 (scaled)
        - Short Passes: -0.73 (scaled)
        - Passes: -0.81 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Ethan Guimaraes (Pos: Right Back, Score: 0.00)
      - Jack Mount (Pos: Right Back, Score: 0.00)
      - Dominik Mount (Pos: Left Back, Score: 0.00)
      - Federico Onana (Pos: Right Back, Score: 0.00)
      - Jordan Ronaldo (Pos: Left Back, Score: 0.00)

  Archetype 2 (Defensive Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Clearances: 0.56 (scaled)
        - Pressures: 0.53 (scaled)
        - Progressive Pass Distance: 0.53 (scaled)
        - Progressive Carries: 0.52 (scaled)
        - Fouls Won: 0.49 (scaled)
      Lower than average in:
        - Player Average Y: -0.05 (scaled)
        - Assists: -0.07 (scaled)
        - Passes: -0.10 (scaled)
        - Short Passes: -0.14 (scaled)
        - Passes Into Final Third: -0.23 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Declan Cancelo (Pos: Centre Back, Score: 0.00)
      - Lisandro White (Pos: Right Back, Score: 0.00)
      - Mason Awoniyi (Pos: Left Back, Score: 0.00)
      - Dominik Sterling (Pos: Left Back, Score: 0.00)
      - Diogo Silva (Pos: Left Back, Score: 0.00)

  Archetype 3 (Defensive Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Passes: 0.79 (scaled)
        - Short Passes: 0.75 (scaled)
        - Passes Into Final Third: 0.61 (scaled)
        - Long Passes: 0.22 (scaled)
        - Through Balls: 0.05 (scaled)
      Lower than average in:
        - Dribbles: -0.52 (scaled)
        - Pressures: -0.54 (scaled)
        - Shots: -0.55 (scaled)
        - Clearances: -0.55 (scaled)
        - Progressive Carries: -0.57 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Richarlison Casemiro (Pos: Right Back, Score: 0.00)
      - Bukayo Tielemans (Pos: Left Back, Score: 0.00)
      - Mason Chilwell (Pos: Left Back, Score: 0.00)
      - Trent Onana (Pos: Right Back, Score: 0.00)
      - Thiago Saka (Pos: Left Back, Score: 0.00)
```

Midfielders

```
================================================================================
2) Distinct Midfield Roles
================================================================================

  Archetype 0 (Midfield Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Progressive Carries: 1.85 (scaled)
        - Dribbles: 1.83 (scaled)
        - Aerial Duels Won: 1.83 (scaled)
        - Fouls Committed: 1.80 (scaled)
        - Shots: 1.78 (scaled)
      Lower than average in:
        - Short Passes: -0.09 (scaled)
        - Passes: -0.10 (scaled)
        - Xg Per Shot: -0.12 (scaled)
        - Through Balls: -0.19 (scaled)
        - Passes Into Final Third: -0.30 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Mason Garnacho (Pos: Defensive Midfielder, Score: 0.00)
      - Kevin Antony (Pos: Defensive Midfielder, Score: 0.00)
      - Marcus Guimaraes (Pos: Central Midfielder, Score: 0.00)
      - William Watkins (Pos: Central Midfielder, Score: 0.00)
      - Lewis Bissouma (Pos: Central Midfielder, Score: 0.00)

  Archetype 1 (Midfield Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Player Average Y: 0.11 (scaled)
        - Xg Per Shot: 0.08 (scaled)
      Lower than average in:
        - Clearances: -0.60 (scaled)
        - Progressive Pass Distance: -0.60 (scaled)
        - Fouls Won: -0.60 (scaled)
        - Short Passes: -0.73 (scaled)
        - Passes: -0.81 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Lewis Szoboszlai (Pos: Defensive Midfielder, Score: 0.00)
      - Trent Pickford (Pos: Defensive Midfielder, Score: 0.00)
      - Julian Son (Pos: Defensive Midfielder, Score: 0.00)
      - Thiago Sterling (Pos: Central Midfielder, Score: 0.00)
      - Joao Kulusevski (Pos: Central Midfielder, Score: 0.00)
```

```
Archetype 2 (Midfield Focus):
  Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
    Higher than average in:
      - Clearances: 0.56 (scaled)
      - Pressures: 0.53 (scaled)
      - Progressive Pass Distance: 0.53 (scaled)
      - Progressive Carries: 0.52 (scaled)
      - Fouls Won: 0.49 (scaled)
    Lower than average in:
      - Player Average Y: -0.05 (scaled)
      - Assists: -0.07 (scaled)
      - Passes: -0.10 (scaled)
      - Short Passes: -0.14 (scal
      - Passes Into Final Third:

  Top 5 Players in this Archetype
    - Declan Eze (Pos: Central Mi
    - Nicolas Son (Pos: Defensive
    - Ruben Pope (Pos: Attacking
    - Cody Sarr (Pos: Defensive M
    - Jordan Colwill (Pos: Attack

Archetype 3 (Midfield Focus):
  Key Characteristics (Top 5 Posi
    Higher than average in:
      - Passes: 0.79 (scaled)
      - Short Passes: 0.75 (scale
      - Passes Into Final Third:
      - Long Passes: 0.22 (scaled
      - Through Balls: 0.05 (scal
    Lower than average in:
      - Dribbles: -0.52 (scaled)
      - Pressures: -0.54 (scaled)
      - Shots: -0.55 (scaled)
      - Clearances: -0.55 (scaled)
      - Progressive Carries: -0.57 (scaled)

  Top 5 Players in this Archetype (by Composite Score):
    - Virgil Trippier (Pos: Defensive Midfielder, Score: 0.00)
    - Trent Foden (Pos: Defensive Midfielder, Score: 0.00)
    - Richarlison Arteta (Pos: Central Midfielder, Score: 0.00)
    - Oscar Martinez (Pos: Attacking Midfielder, Score: 0.00)
    - Martin Leno (Pos: Attacking Midfielder, Score: 0.00)
```

```
Archetype 4 (Midfield Focus):
  Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
    Higher than average in:
      - Assists: 19.13 (scaled)
      - Shots On Target: 1.92 (scaled)
      - Passes Into Final Third: 1.81 (scaled)
      - Through Balls: 1.62 (scaled)
      - Progressive Pass Distance: 1.61 (scaled)
    Lower than average in:
      - Successful Dribbles: -0.81 (scaled)
      - Long Passes: -0.87 (scaled)
      - Clearances: -0.87 (scaled)
      - Carries Into Final Third: -0.93 (scaled)
      - Xg Per Shot: -1.01 (scaled)

  Top 5 Players in this Archetype (by Composite Score):
    - Nathan Casemiro (Pos: Defensive Midfielder, Score: 0.00)
```

Attackers

```
================================================================================
3) Distinct Attacking Roles
================================================================================

  Archetype 0 (Attacking Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Progressive Carries: 1.85 (scaled)
        - Dribbles: 1.83 (scaled)
        - Aerial Duels Won: 1.83 (scaled)
        - Fouls Committed: 1.80 (scaled)
        - Shots: 1.78 (scaled)
      Lower than average in:
        - Short Passes: -0.09 (scaled)
        - Passes: -0.10 (scaled)
        - Xg Per Shot: -0.12 (scaled)
        - Through Balls: -0.19 (scaled)
        - Passes Into Final Third: -0.30 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Lewis Mac Allister (Pos: Left Winger, Score: 0.00)
      - Ollie Palmer (Pos: Striker, Score: 0.00)
      - Dominik Leno (Pos: Left Winger, Score: 0.00)
      - Nathan Van Dijk (Pos: Second Striker, Score: 0.00)
      - Declan Garnacho (Pos: Second Striker, Score: 0.00)

  Archetype 1 (Attacking Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Player Average Y: 0.11 (scaled)
        - Xg Per Shot: 0.08 (scaled)
      Lower than average in:
        - Clearances: -0.60 (scaled)
        - Progressive Pass Distance: -0.60 (scaled)
        - Fouls Won: -0.60 (scaled)
        - Short Passes: -0.73 (scaled)
        - Passes: -0.81 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Pierre-Emile Odegaard (Pos: Striker, Score: 0.00)
      - Aaron Salah (Pos: Right Winger, Score: 0.00)
      - Fabinho Nunez (Pos: Left Winger, Score: 0.00)
      - Manuel Pedro (Pos: Right Winger, Score: 0.00)
      - Lisandro Chilwell (Pos: Striker, Score: 0.00)

  Archetype 2 (Attacking Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Clearances: 0.56 (scaled)
        - Pressures: 0.53 (scaled)
        - Progressive Pass Distance: 0.53 (scaled)
        - Progressive Carries: 0.52 (scaled)
        - Fouls Won: 0.49 (scaled)
      Lower than average in:
        - Player Average Y: -0.05 (scaled)
        - Assists: -0.07 (scaled)
        - Passes: -0.10 (scaled)
        - Short Passes: -0.14 (scaled)
        - Passes Into Final Third: -0.23 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Raphael Pedro (Pos: Right Winger, Score: 0.00)
      - Marcus Dias (Pos: Second Striker, Score: 0.00)
      - Son Mac Allister (Pos: Second Striker, Score: 0.00)
      - Jarrod Kane (Pos: Second Striker, Score: 0.00)
      - Mohamed Maguire (Pos: Second Striker, Score: 0.00)

  Archetype 3 (Attacking Focus):
    Key Characteristics (Top 5 Positive & Negative Contributions on Scaled Data):
      Higher than average in:
        - Passes: 0.79 (scaled)
        - Short Passes: 0.75 (scaled)
        - Passes Into Final Third: 0.61 (scaled)
        - Long Passes: 0.22 (scaled)
        - Through Balls: 0.05 (scaled)
      Lower than average in:
        - Dribbles: -0.52 (scaled)
        - Pressures: -0.54 (scaled)
        - Shots: -0.55 (scaled)
        - Clearances: -0.55 (scaled)
        - Progressive Carries: -0.57 (scaled)

    Top 5 Players in this Archetype (by Composite Score):
      - Kai Colwill (Pos: Striker, Score: 0.00)
      - Kevin Ronaldo (Pos: Left Winger, Score: 0.00)
      - Ethan Eze (Pos: Right Winger, Score: 0.00)
      - Marcus Chilwell (Pos: Striker, Score: 0.00)
      - Jack Arteta (Pos: Left Winger, Score: 0.00)
```