UNSUPERVISED LEARNING
BUSINESS REPORT-PGP-DSBA

By: Parthasarathi Behura

Define the problem

Fantasy sports are online gaming platforms where participants draft and manage virtual teams of real professional sports players. Based on the performance of the players in the real world, players are allotted points in the fantasy sports platform every match. The objective is to create the best possible team with a fixed budget to score maximum fantasy points, and users compete against each other over an entire sports league or season. Some of these fantasy sports require actual financial investments for participation, with the chances of winning monetary rewards as well as free matchday tickets on a periodic basis.

The fantasy sports market has seen tremendous growth over the past few years, with a valuation of $18.6 billion in 2019. The football (soccer) segment led in terms of market share in 2019, with over 8 million participants worldwide, and is expected to retain its dominance over the next couple of years. Digitalization is one of the primary factors driving the growth of the fantasy sports market as it allows participants the opportunity to compete on a global level and test their skills. With an increase in smartphone usage and availability of fantasy sports apps, this market is expected to witness a globe surge and reach a $48.6 billion valuation by 2027.

Goal & Objective

OnSports is a fantasy sports platform which has fantasy leagues for many different sports and has witnessed an increasing number of participants globally over the past 5 years. For each player, a price is set at the start, and the price keeps changing over time based on the performance of the players in the real world. With the new English Premier League season about to start, they have collected data of the past season and want to analyze it to determine the price of each player for the start of the new season. OnSports have hired you as a data scientist and asked you to conduct a cluster analysis to identify players of different potentials of each player based on previous season performance. This will help them understand the patterns in player performances and fantasy returns and decide the exact price to be set for each player for the upcoming football season.

Data dictionary:

- Player_Name: Name of the player
- Club: Club in which the player plays
- Position: Position in which the player plays
- Goals_Scored: Number of goals scored by the player in the previous season
- Assists: Number of passes made by the player leading to goals in the previous season
- Total_Points: Total number of fantasy points scored by the player in the previous season
- Minutes: Number of minutes played by the player in the previous season
- Goals_Conceded: Number of goals conceded by the player in the previous season
- Creativity: A score, computed using a range of stats, that assesses player performance in terms of producing goalscoring opportunities for other players
- Influence: A score, computed using a range of stats, that evaluates a player's impact on a match, taking into account actions that could directly or indirectly affect the match outcome
- Threat: A score, computed using a range of stats, that gauges players who are most likely to score goals
- Bonus: Total bonus points received (The three best performing players in each match receive additional bonus points based on a score computed using a range of stats. 3 points are awarded to the highest scoring player, 2 to the second best, and 1 to the third.)
- Clean_Sheets: Number of matches without conceding a goal in the previous season

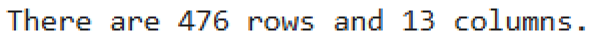
1. Exploratory Data Analysis-

Problem definition - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

1.a Defining problem statement:

OnSports is a fantasy sports platform which has fantasy leagues for many different sports and has witnessed an increasing number of participants globally over the past 5 years. For each player, a price is set at the start, and the price keeps changing over time based on the performance of the players in the real world. With the new English Premier League season about to start, they have collected data of the past season and want to analyze it to determine the price of each player for the start of the new season. OnSports have hired you as a data scientist and asked you to conduct a cluster analysis to identify players of different potentials of each player based on previous season performance. This will help them understand the patterns in player performances and fantasy returns and decide the exact price to be set for each player for the upcoming football season.

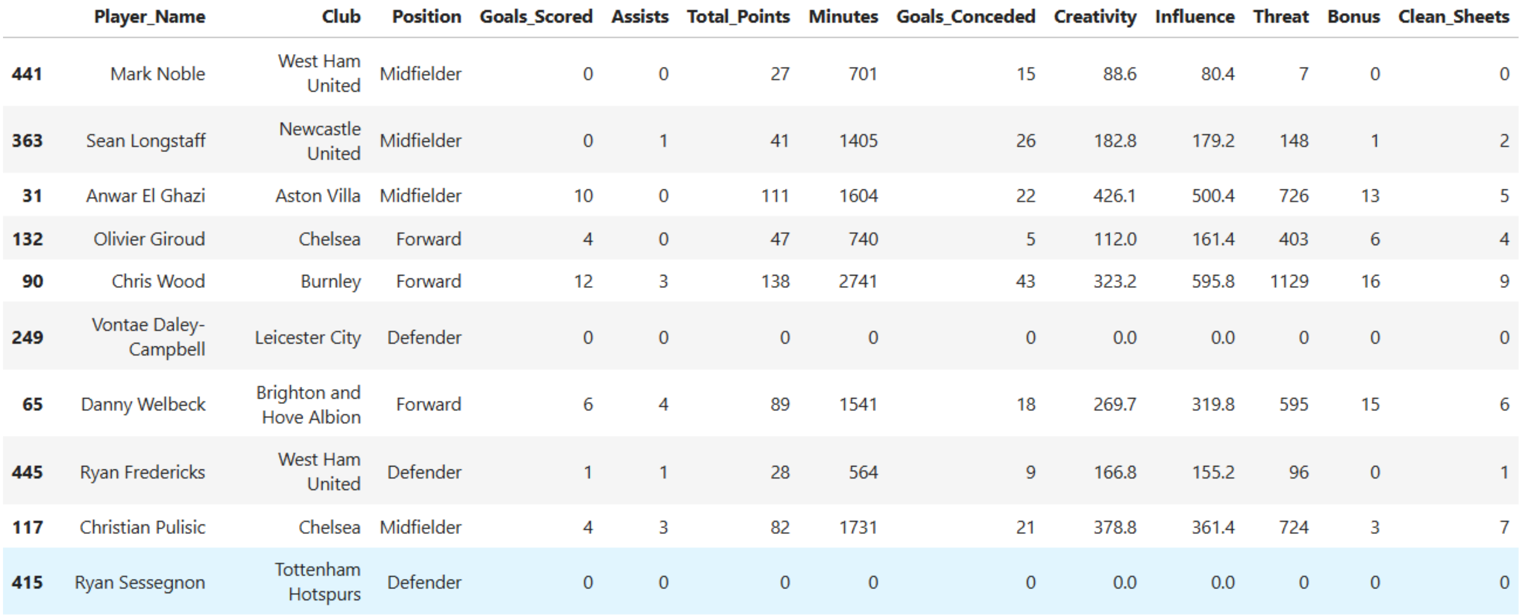
Checking the shape of the data:



```
There are 476 rows and 13 columns.
```

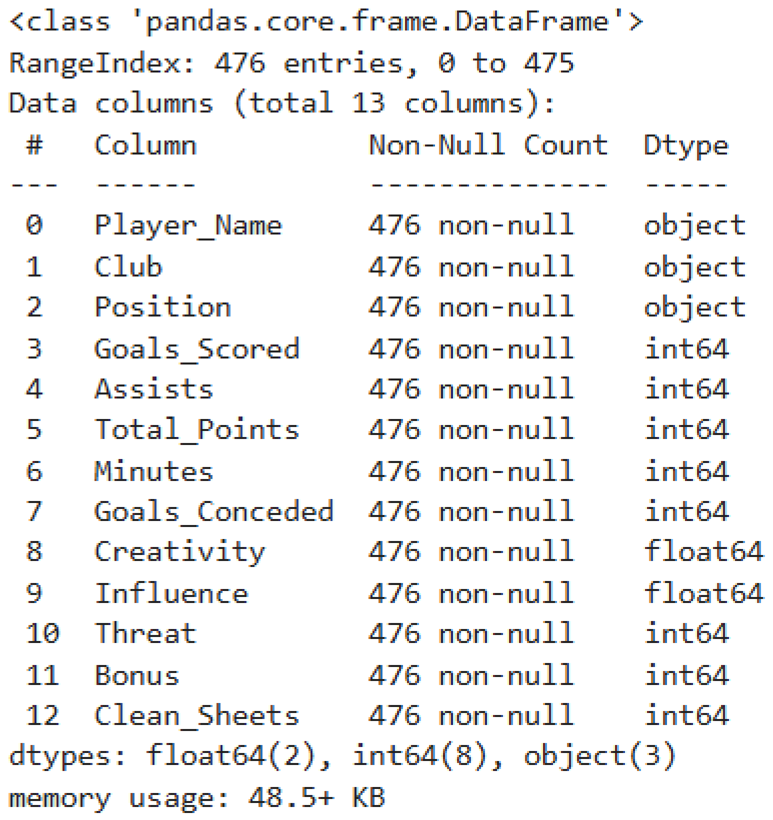
Shows the 476 players with 13 features

Checking random data type:



	Player_Name	Club	Position	Goals_Scored	Assists	Total_Points	Minutes	Goals_Conceded	Creativity	Influence	Threat	Bonus	Clean_Sheets
441	Mark Noble	West Ham United	Midfielder	0	0	27	701	15	88.6	80.4	7	0	0
363	Sean Longstaff	Newcastle United	Midfielder	0	1	41	1405	26	182.8	179.2	148	1	2
31	Anwar El Ghazi	Aston Villa	Midfielder	10	0	111	1604	22	426.1	500.4	726	13	5
132	Olivier Giroud	Chelsea	Forward	4	0	47	740	5	112.0	161.4	403	6	4
90	Chris Wood	Burnley	Forward	12	3	138	2741	43	323.2	595.8	1129	16	9
249	Vontae Daley-Campbell	Leicester City	Defender	0	0	0	0	0	0.0	0.0	0	0	0
65	Danny Welbeck	Brighton and Hove Albion	Forward	6	4	89	1541	18	269.7	319.8	595	15	6
445	Ryan Fredericks	West Ham United	Defender	1	1	28	564	9	166.8	155.2	96	0	1
117	Christian Pulisic	Chelsea	Midfielder	4	3	82	1731	21	378.8	361.4	724	3	7
415	Ryan Sessegnon	Tottenham Hotspurs	Defender	0	0	0	0	0	0.0	0.0	0	0	0

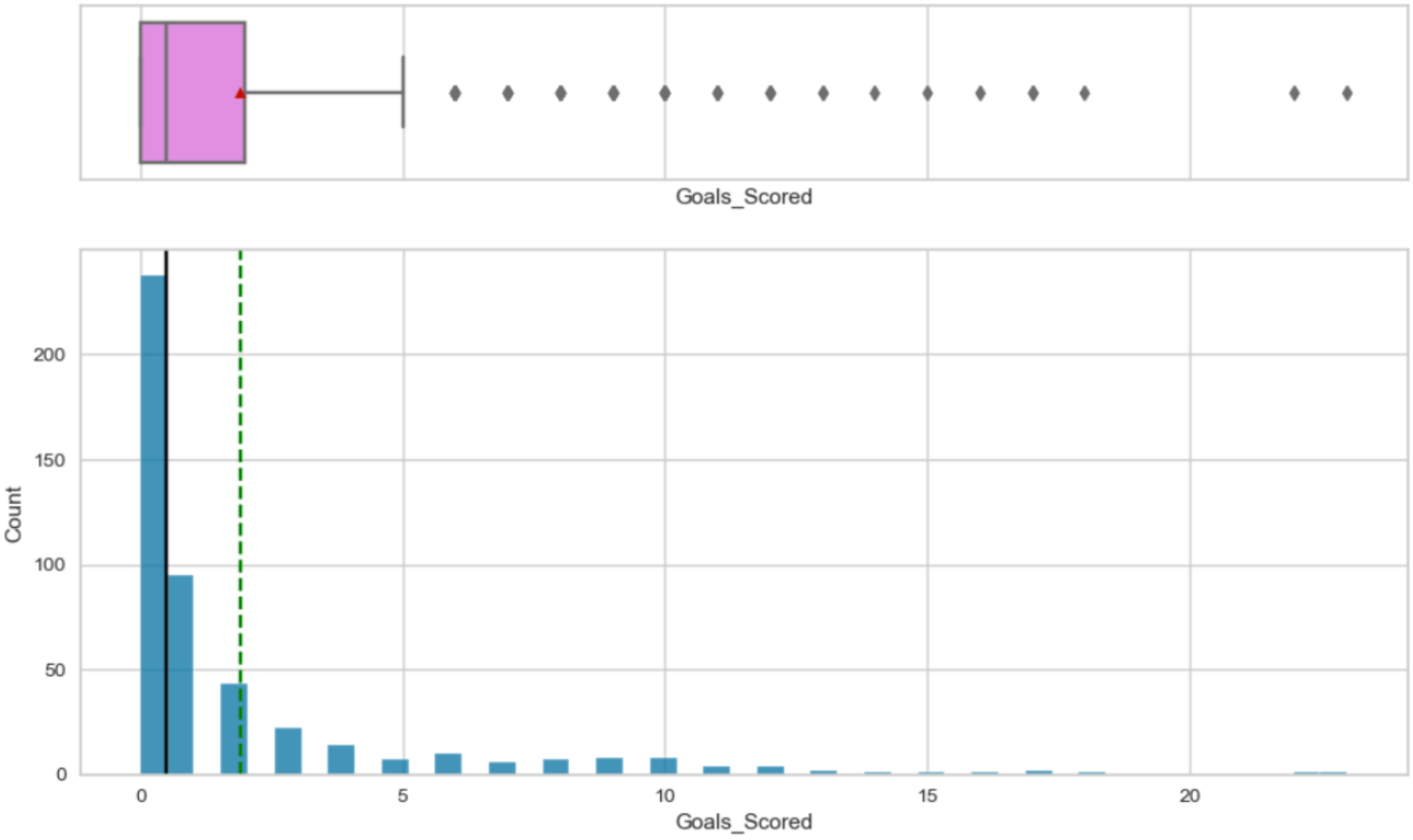
Checking column names & data type:



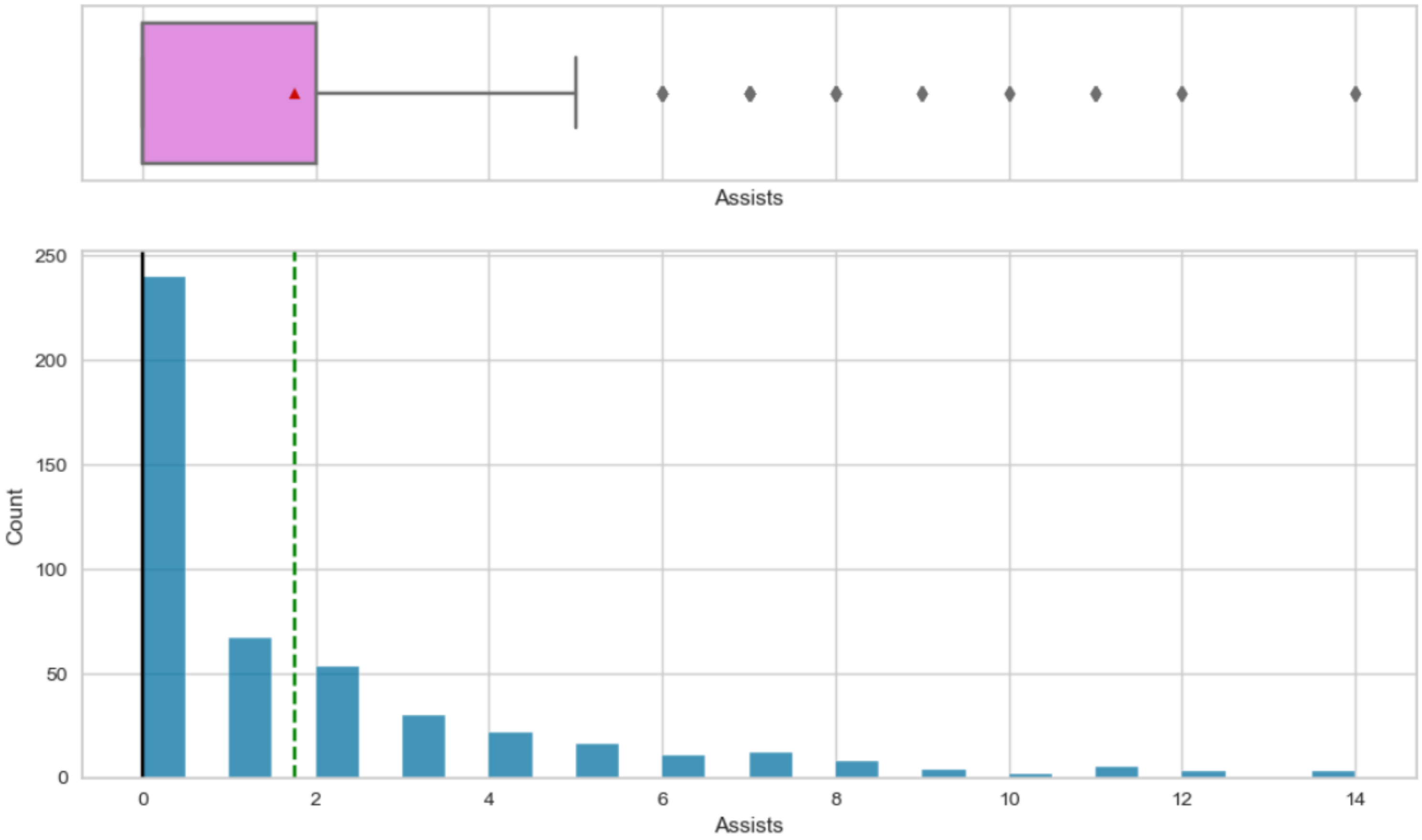
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 476 entries, 0 to 475
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Player_Name     476 non-null    object 
 1   Club            476 non-null    object 
 2   Position        476 non-null    object 
 3   Goals_Scored    476 non-null    int64  
 4   Assists         476 non-null    int64  
 5   Total_Points    476 non-null    int64  
 6   Minutes         476 non-null    int64  
 7   Goals_Conceded  476 non-null    int64  
 8   Creativity      476 non-null    float64
 9   Influence       476 non-null    float64
 10  Threat          476 non-null    int64  
 11  Bonus           476 non-null    int64  
 12  Clean_Sheets    476 non-null    int64  
dtypes: float64(2), int64(8), object(3)
memory usage: 48.5+ KB
```

- Player_Name, Club, Position, are the only object type columns the rest are numerical
- Of the 476 players and 13 columns there are no missing values

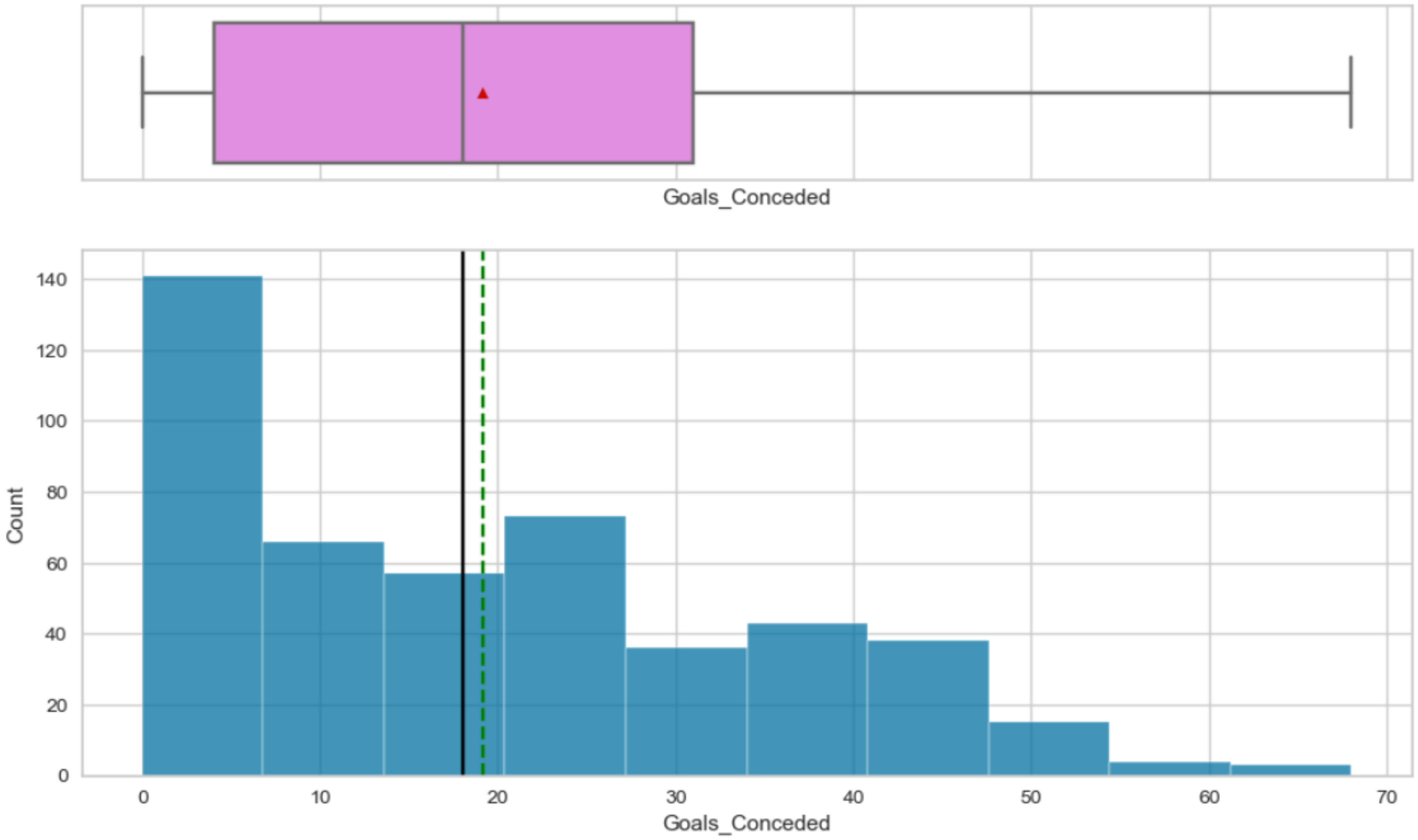
1.b Univariate Analysis:



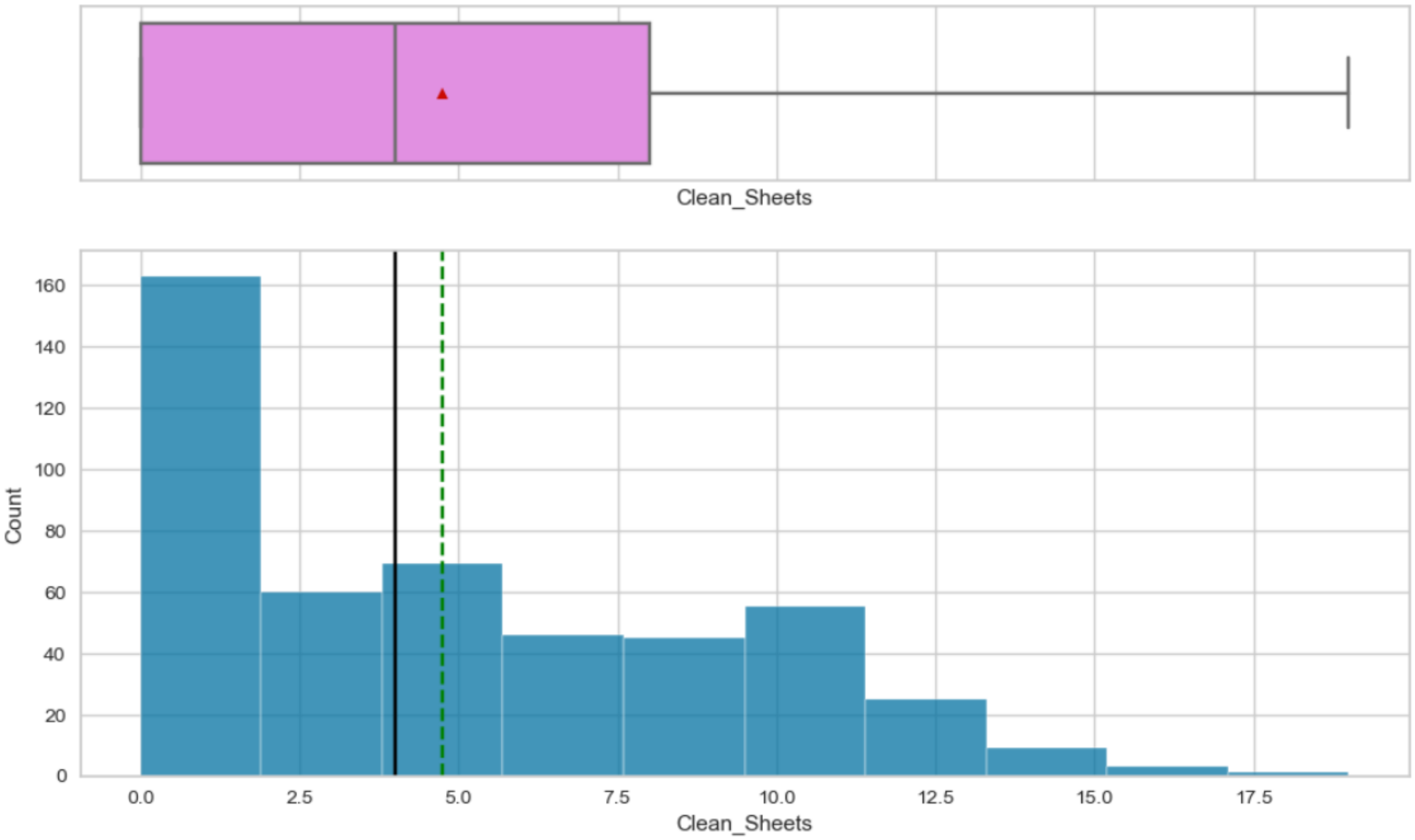
Histogram box plot of Goals_Scored



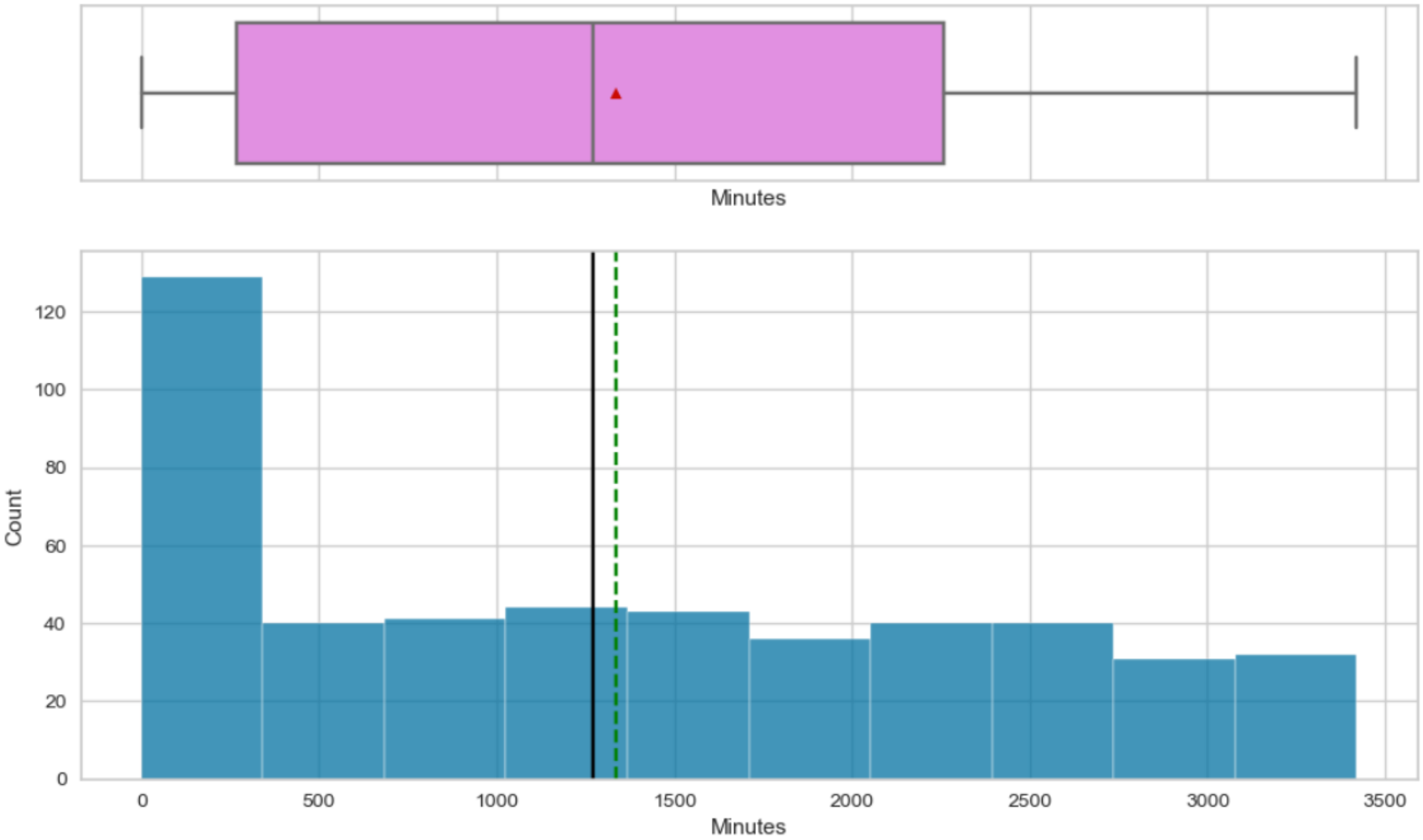
Histogram box plot of Assists



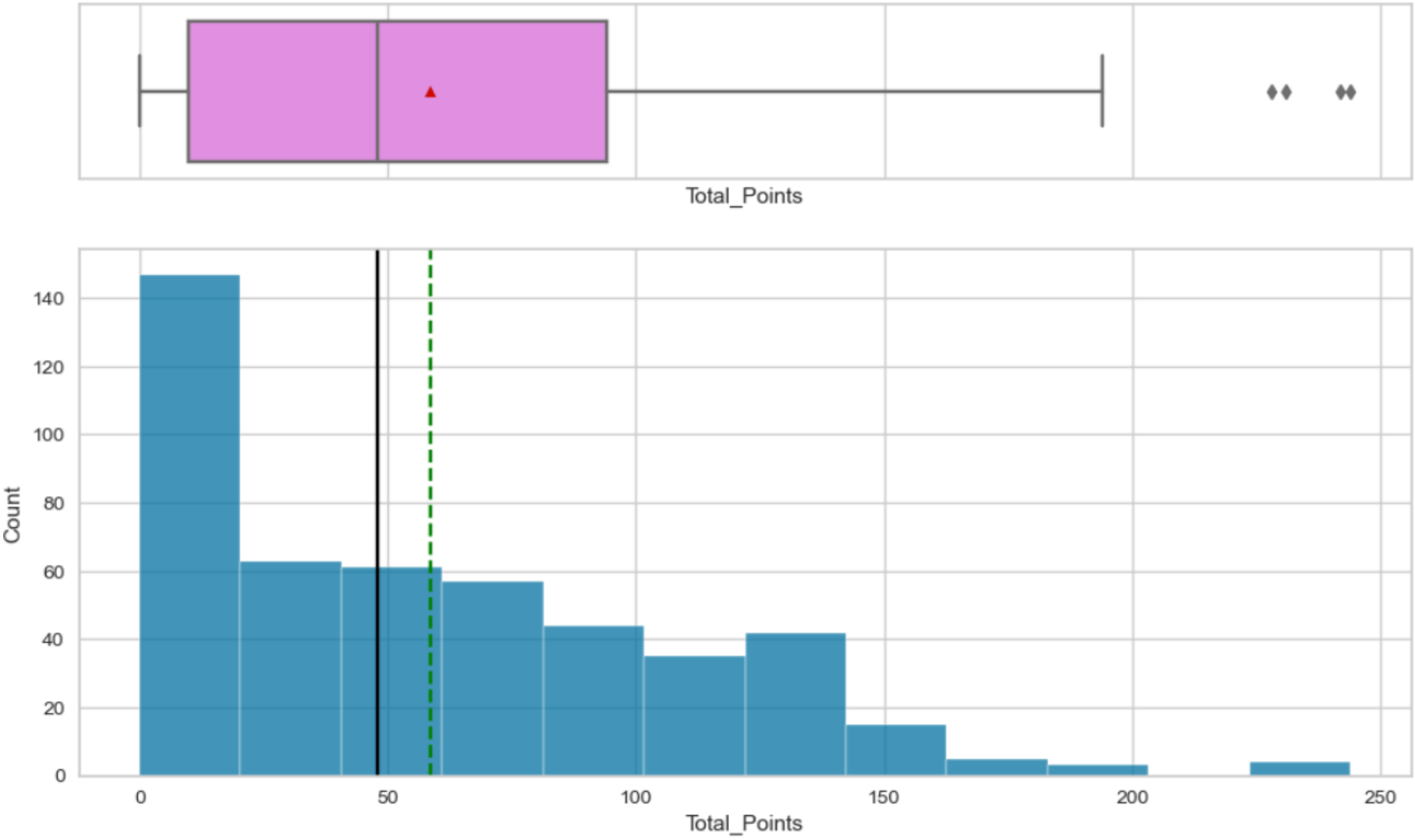
Histogram box plot of Goals_Conceded



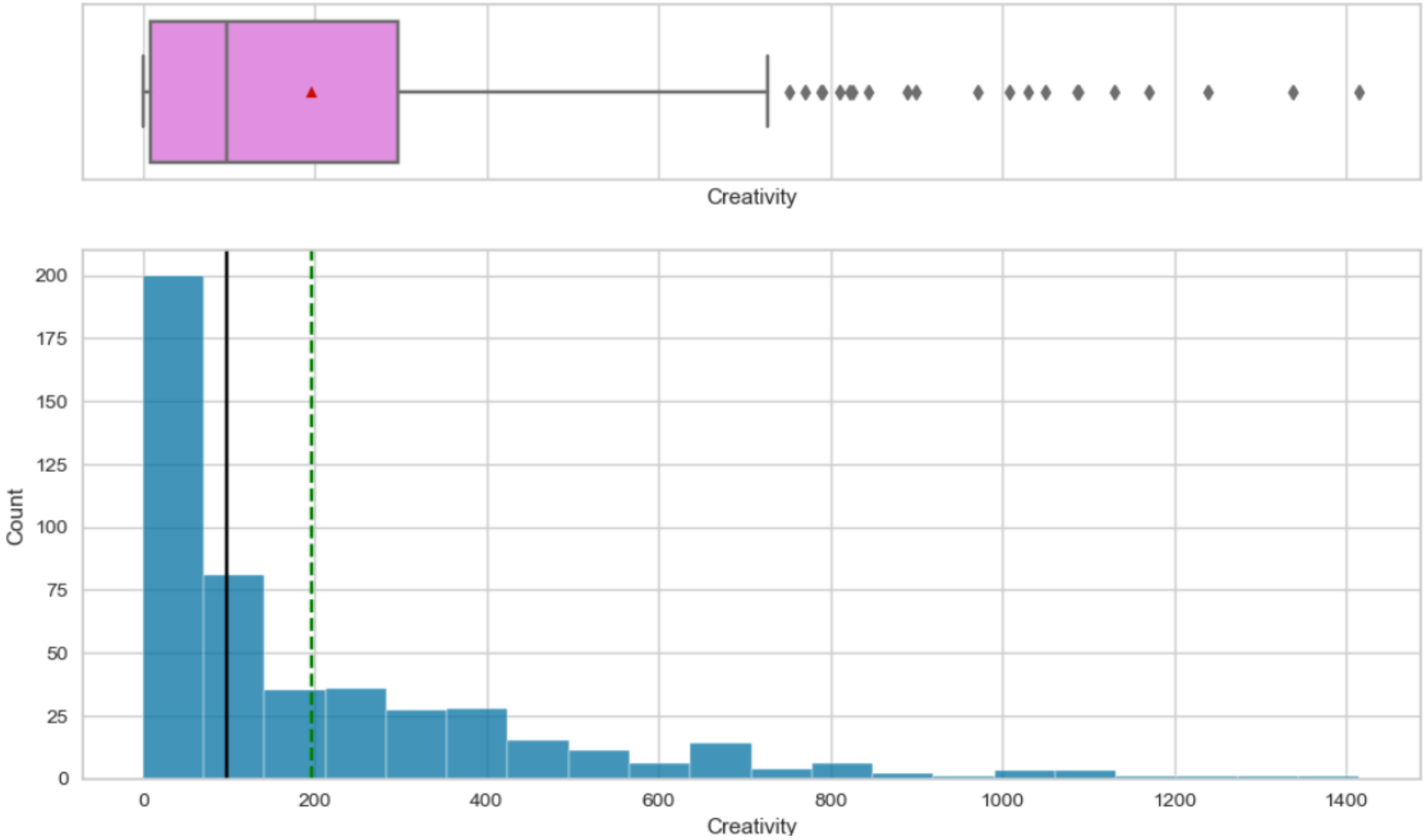
Histogram box plot of Clean_Sheets



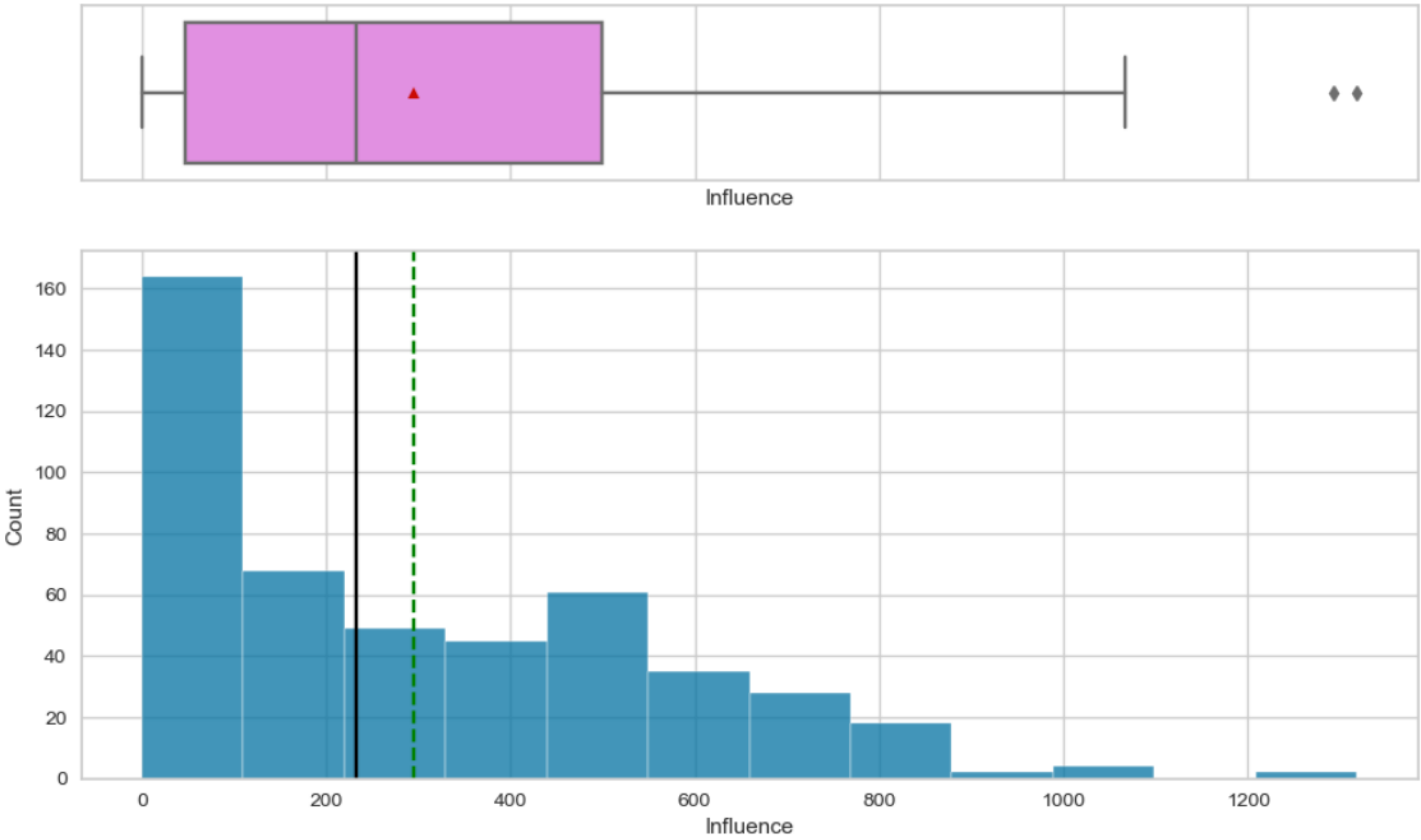
Histogram box plot of Minutes



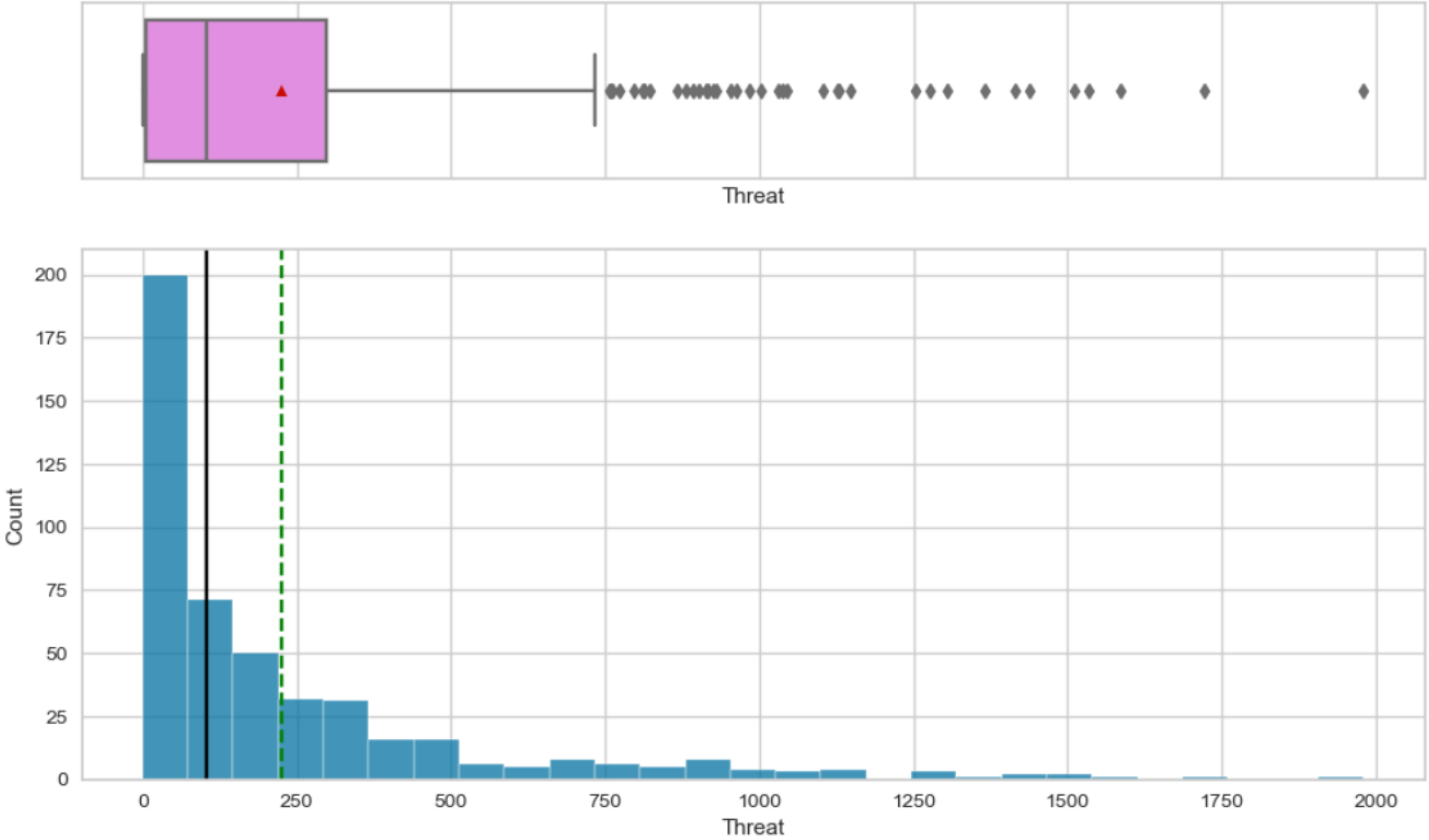
Histogram box plot of Total_Points



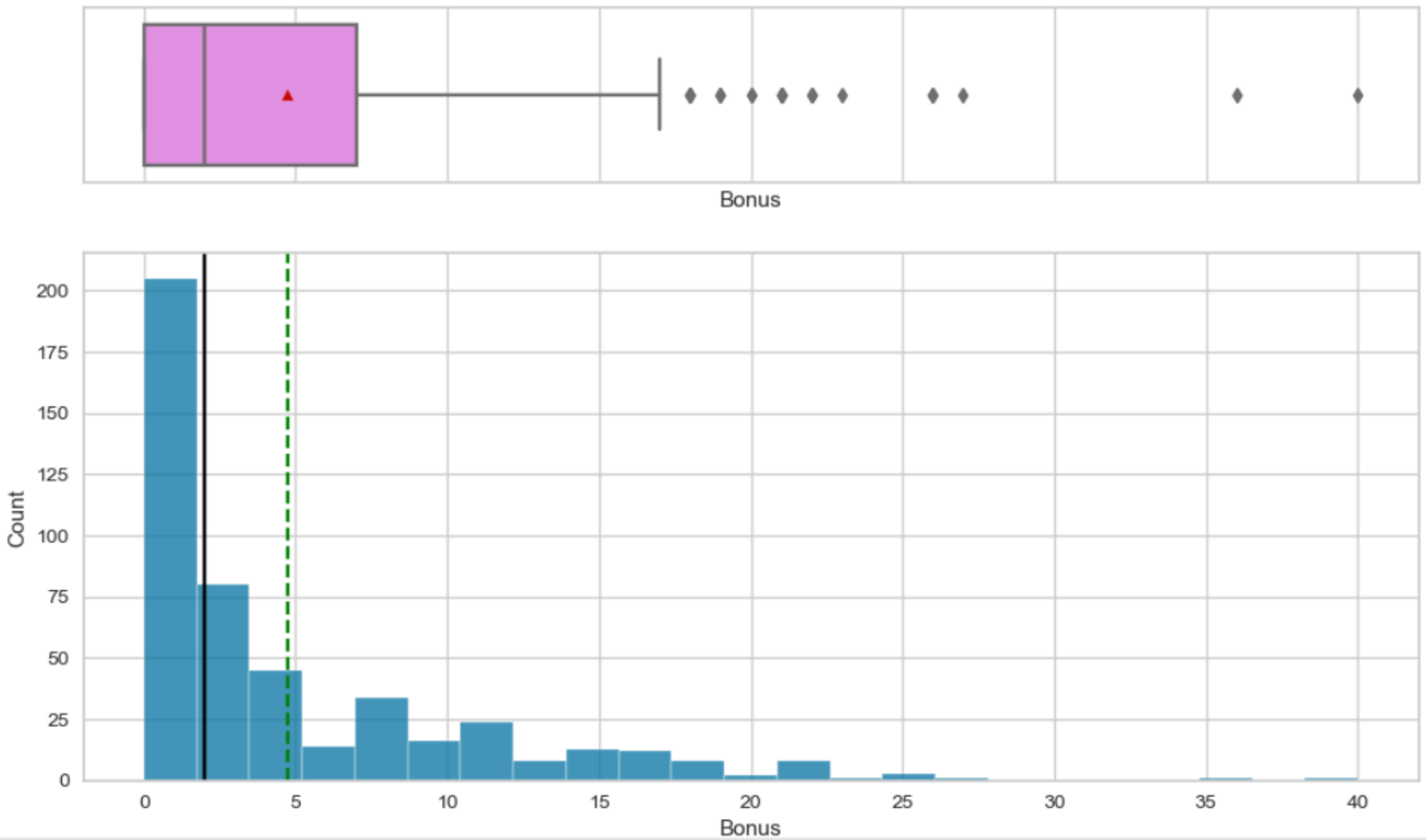
Histogram box plot of Creativity



Histogram box plot of Influence



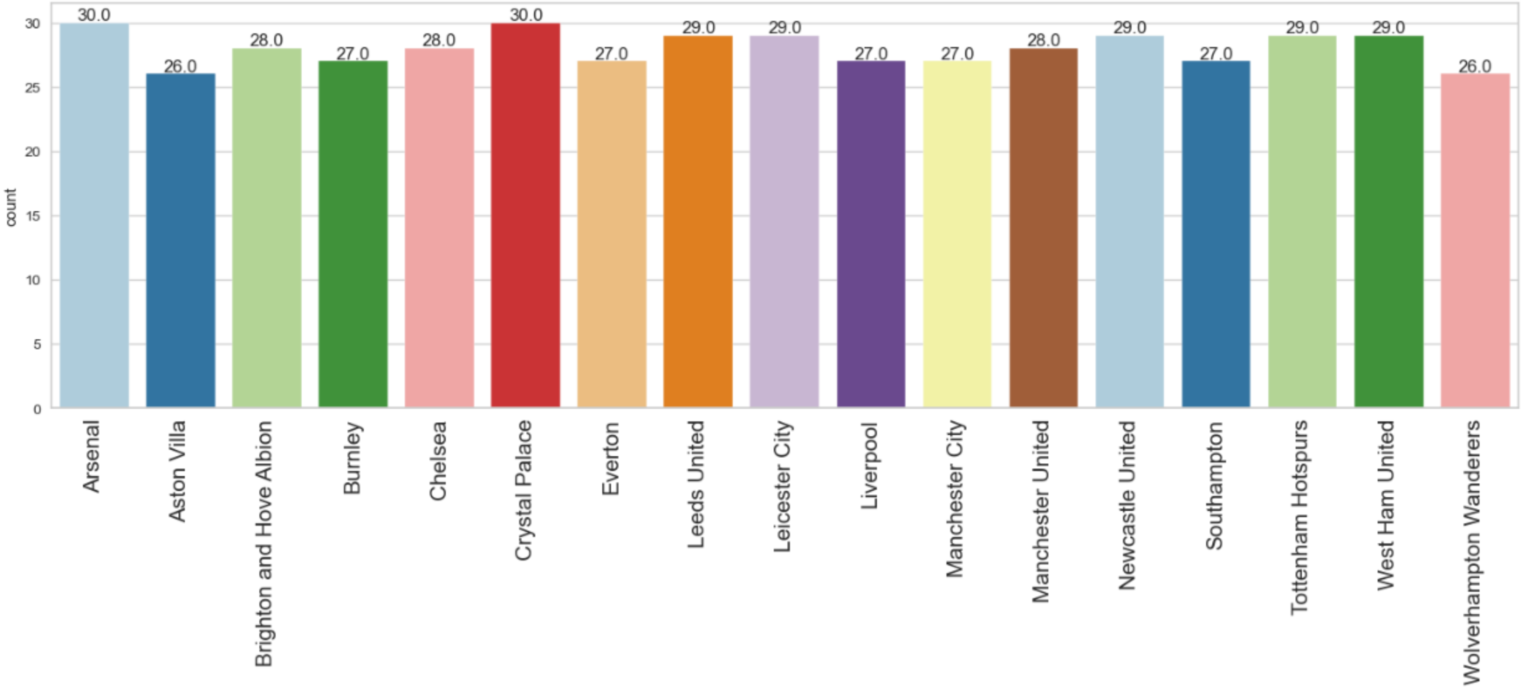
Histogram box plot of Threat



Histogram box plot of Bonus

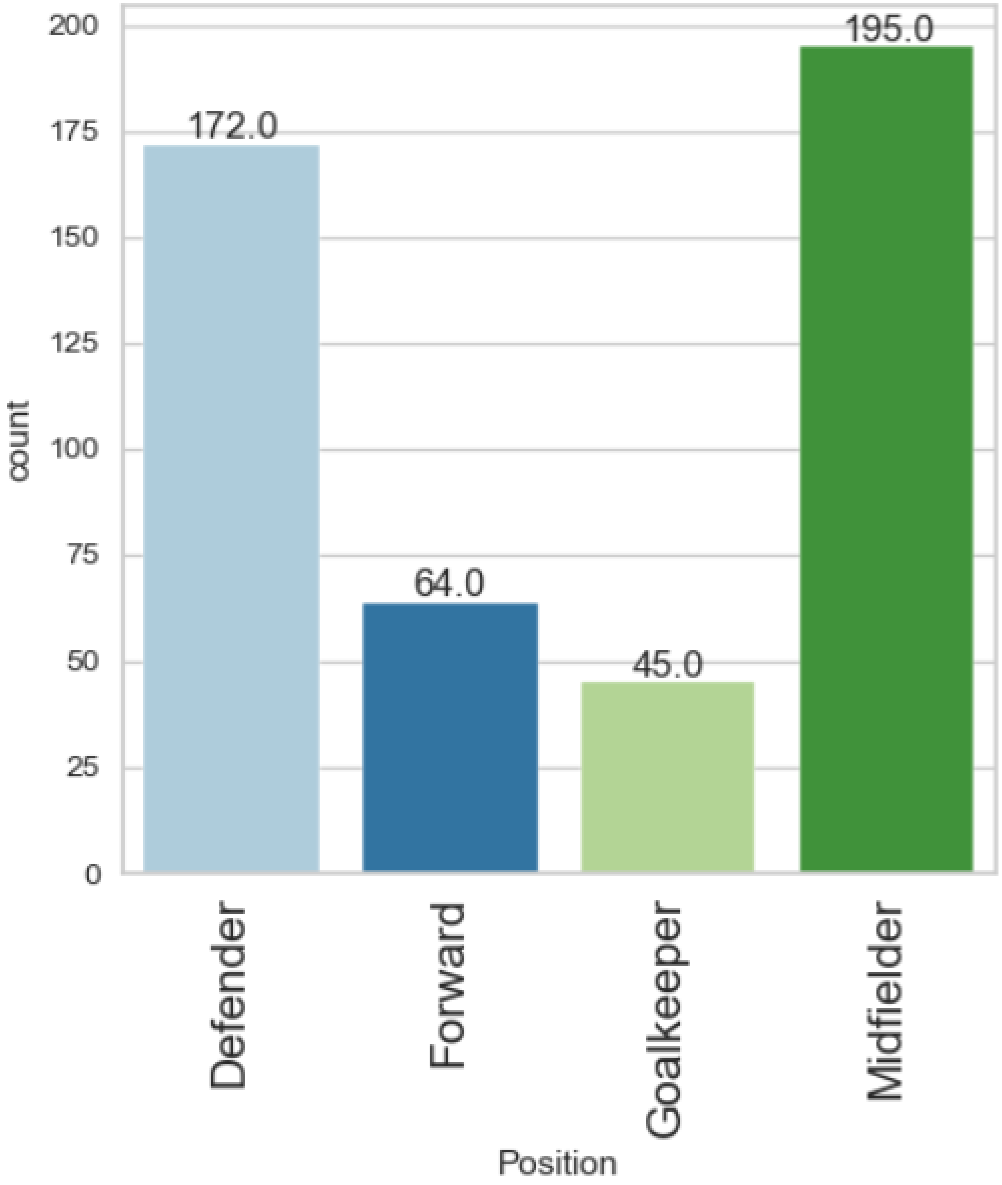
Observations and Insights: The right skewed nature consistent through all plots indicate this is not likely due to outliers but rather an natural imbalance in the players. This imbalance likely stems from one of two factors:

- Players who are higher performers.
- Certain positions that tend to rank higher on the measurable features.



Bar plot of Club

Reletively **uniform distribution** of players from each club, should help to minimize potential errors from imbalanced data.

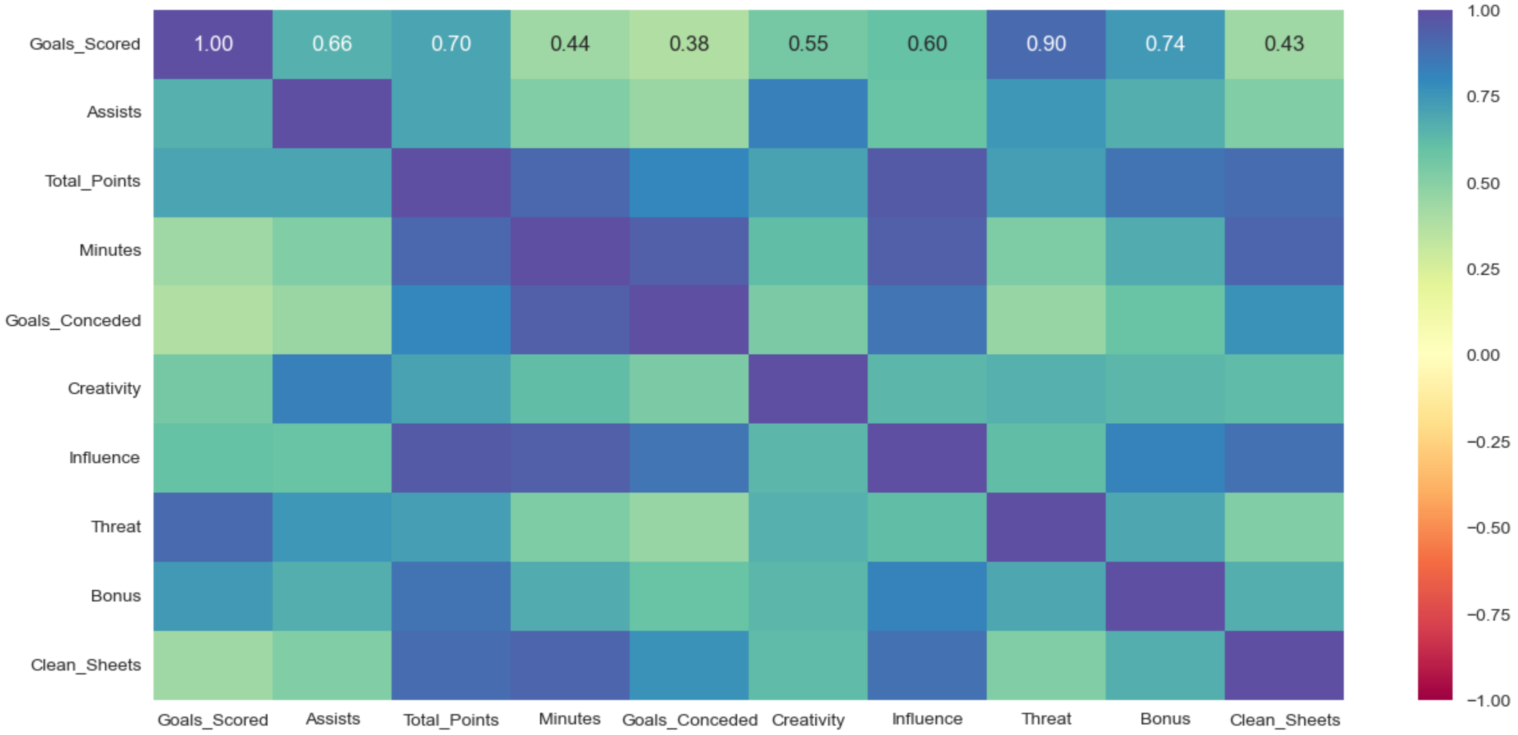


Bar plot of Position

Of the 11 players on the field depending on the formation there is 1 goalie, 3-5 defenders, 4-5 midfielders, 1-3 forwards.

- The split shown above matches those ratios with positions ranked as:
    - Number of Midfielders > Defenders > Forwards > Goalkeepers.
- Given the number of Clubs and Goal Keepers, each club has on average 2-3 goal keepers.

1.c Bivariate Analysis:

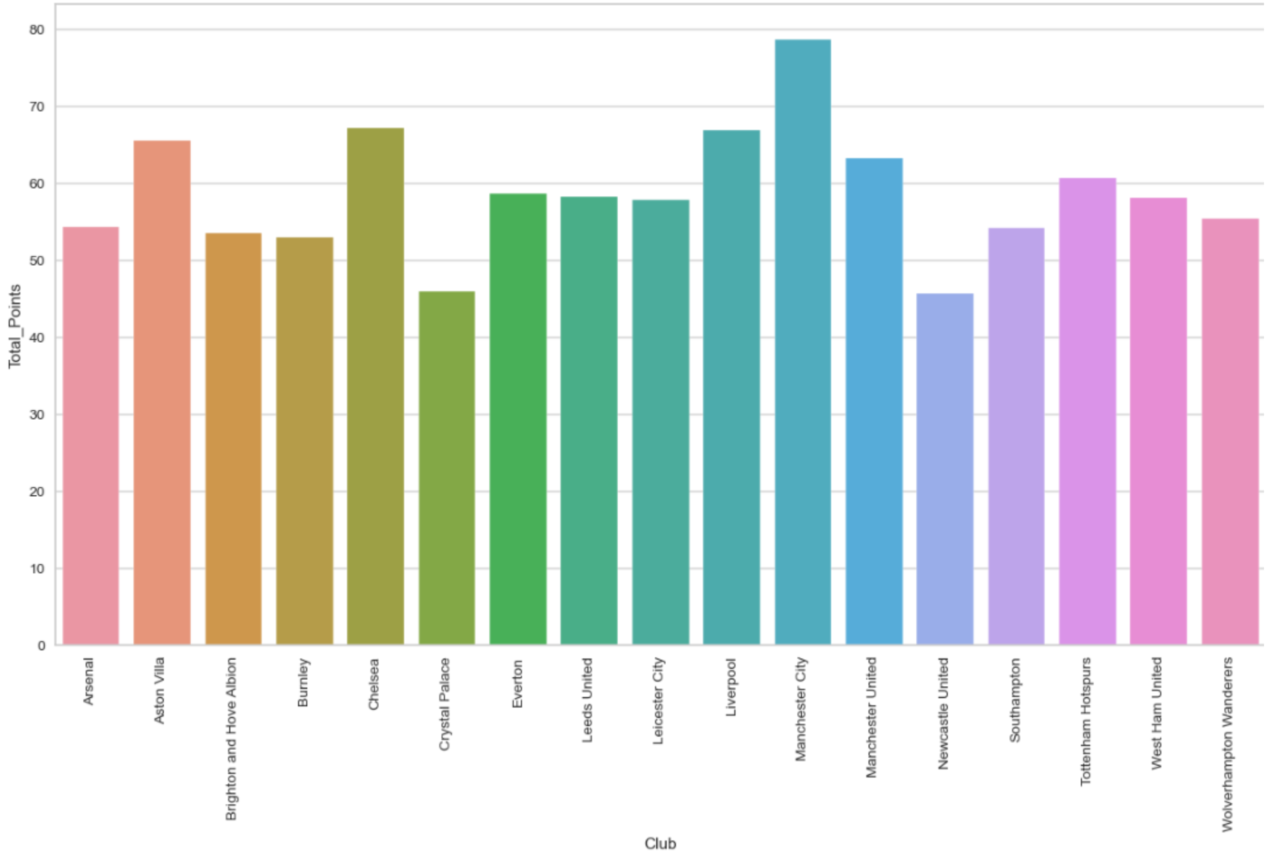


Heat Map Correlation

Observations and Insights: There is a high correlation (>= 0.7 or <= -0.7) between:

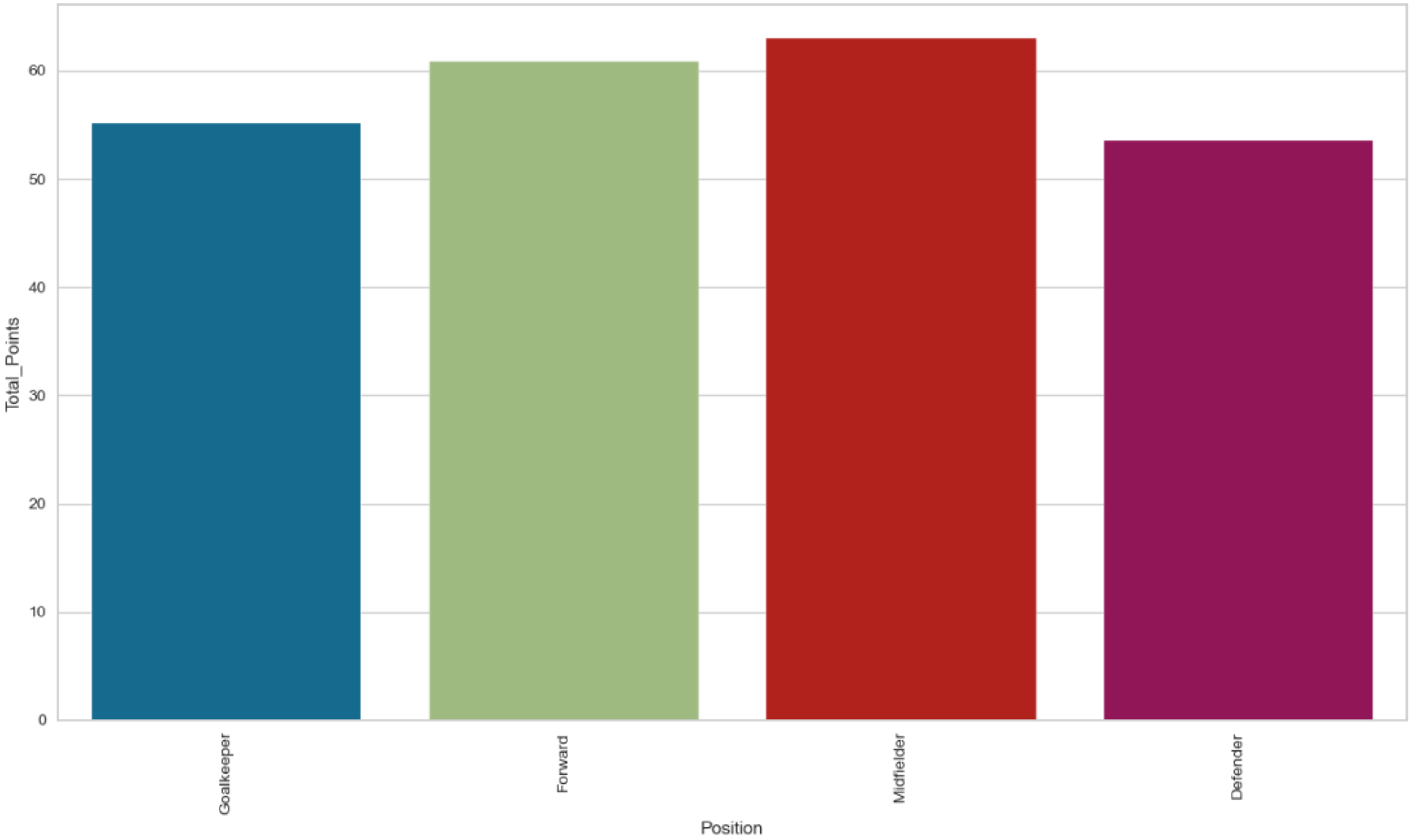
- Correlation between Assists, Goals_Scored, and Total_Points, which makes sense given the first 2 contribute to the 3rd and those likely to score are also likely to get assists.
- Big correlation (.91) between minutes played and total_points, which makes sense as this gives players more chances.
- Correlation between Goals_Conceded, Total_ Points, Minites, which echos our above observation that those withough goals conceded are likely not getting game time.
- Correlation between Creativity and Assits, given creativity is a measure of, " A score, computed using a range of stats, that assesses player performance in terms of producing goalscoring opportunities for other players.", that is likely mostly measured by assits.

Could continue, but the **most relevant observation is that many of these features are highly correlated**.



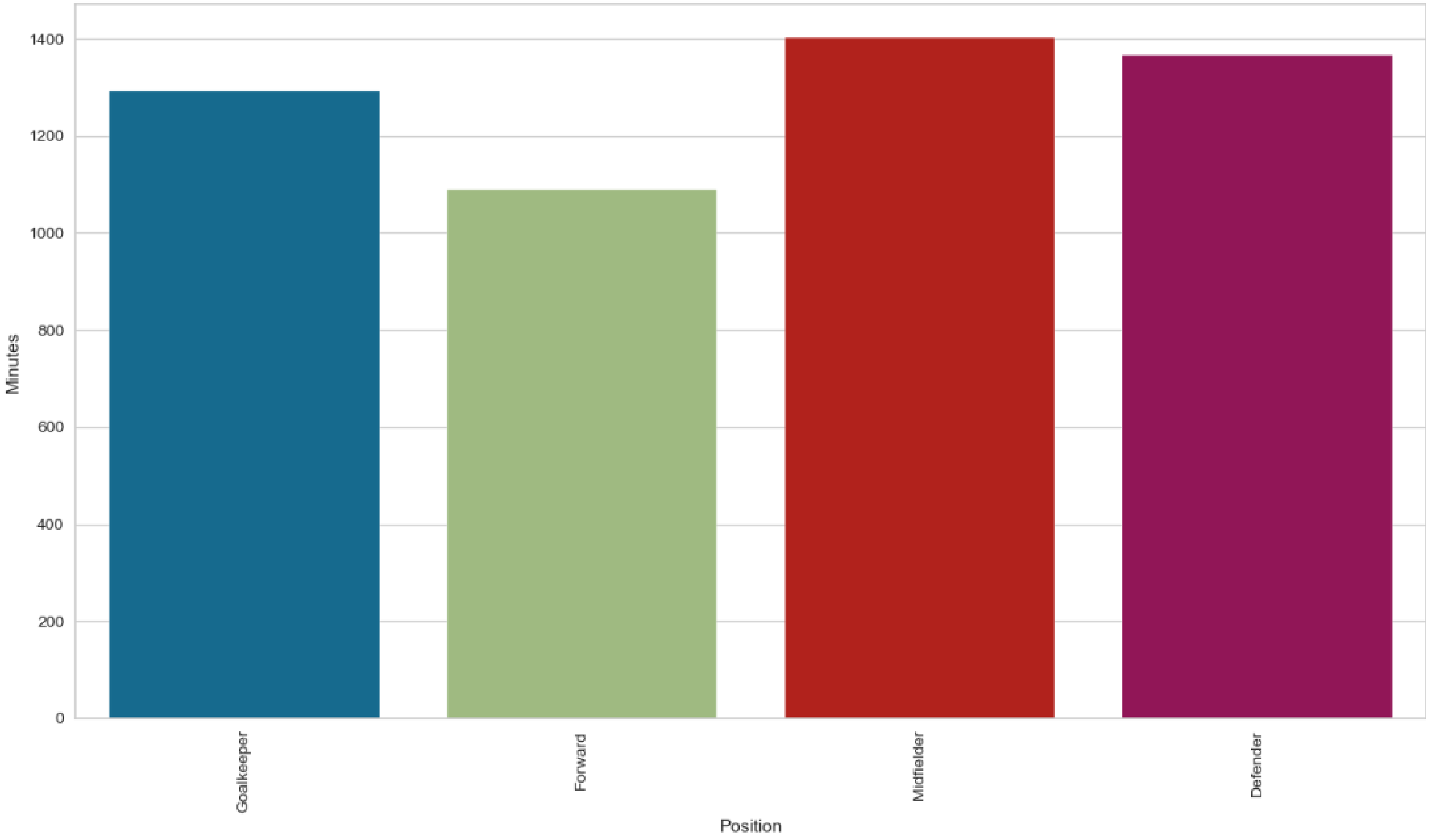
Club vs Total Points

- Manchester City is the leader in points while Crystal Palace, and Newcastle United have the lowest number of points.



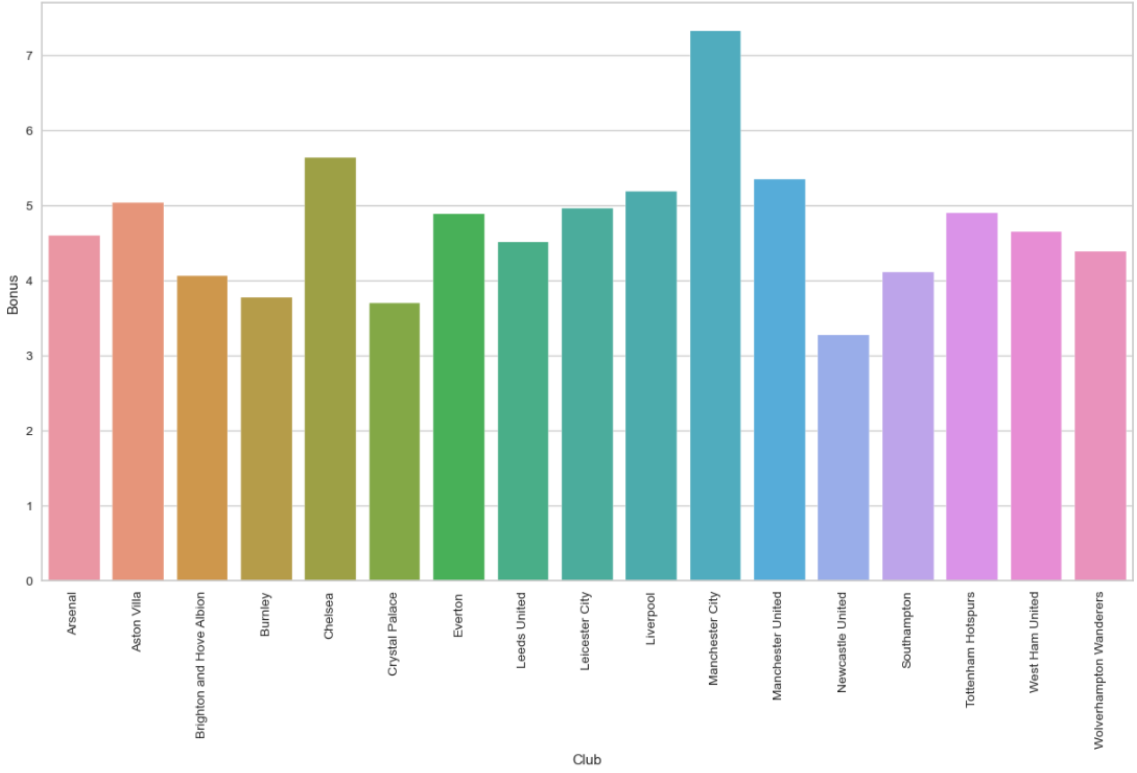
Position vs Total Points

- As expected Midfielders score the most total points followed by Forward more than Defenders and Goalkeepers.



Position vs Minutes

- Midfielders played most the minutes than anyother position players.



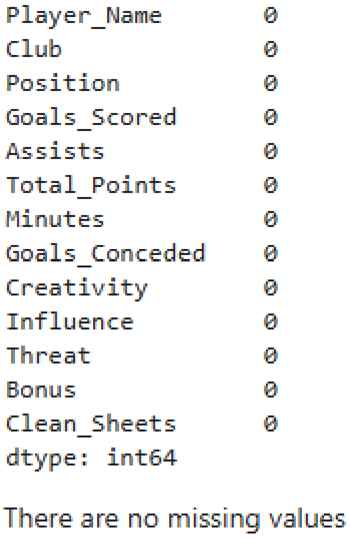
Club vs Bonus

- Manchester City players have scored most bonus points than anyother club players.

2. Data Preprocessing:

Missing Value Treatment (with rationale if needed) - Outlier Detection and Treatment (with rationale if needed) - Feature Engineering (with rationale if needed) - Data Scaling (with rationale if needed) - Train-test split

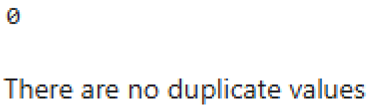
2.a Missing Value:



```
Player_Name       0
Club              0
Position          0
Goals_Scored      0
Assists           0
Total_Points      0
Minutes           0
Goals_Conceded    0
Creativity        0
Influence         0
Threat            0
Bonus             0
Clean_Sheets      0
dtype: int64
```

There are no missing values

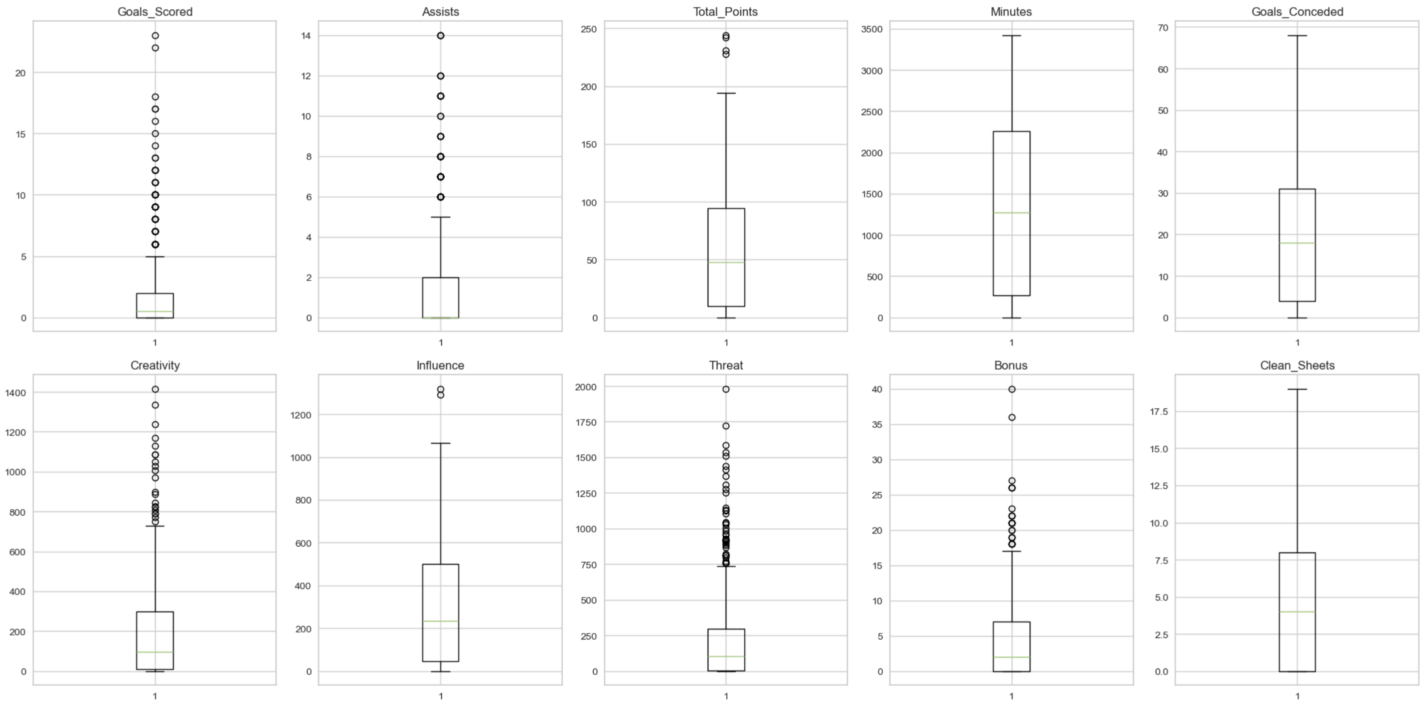
Checking Duplicate Values:



```
0
```

There are no duplicate values

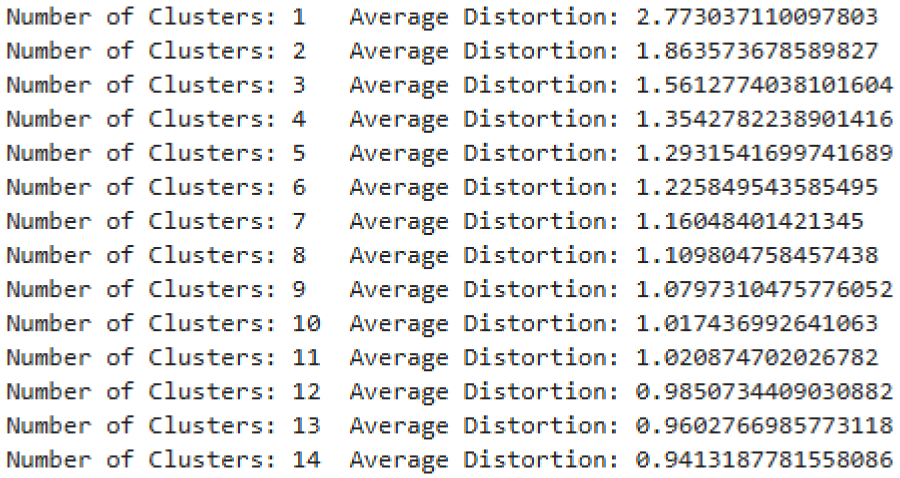
2.b Outlier detection:



Subplots of all numerical variables

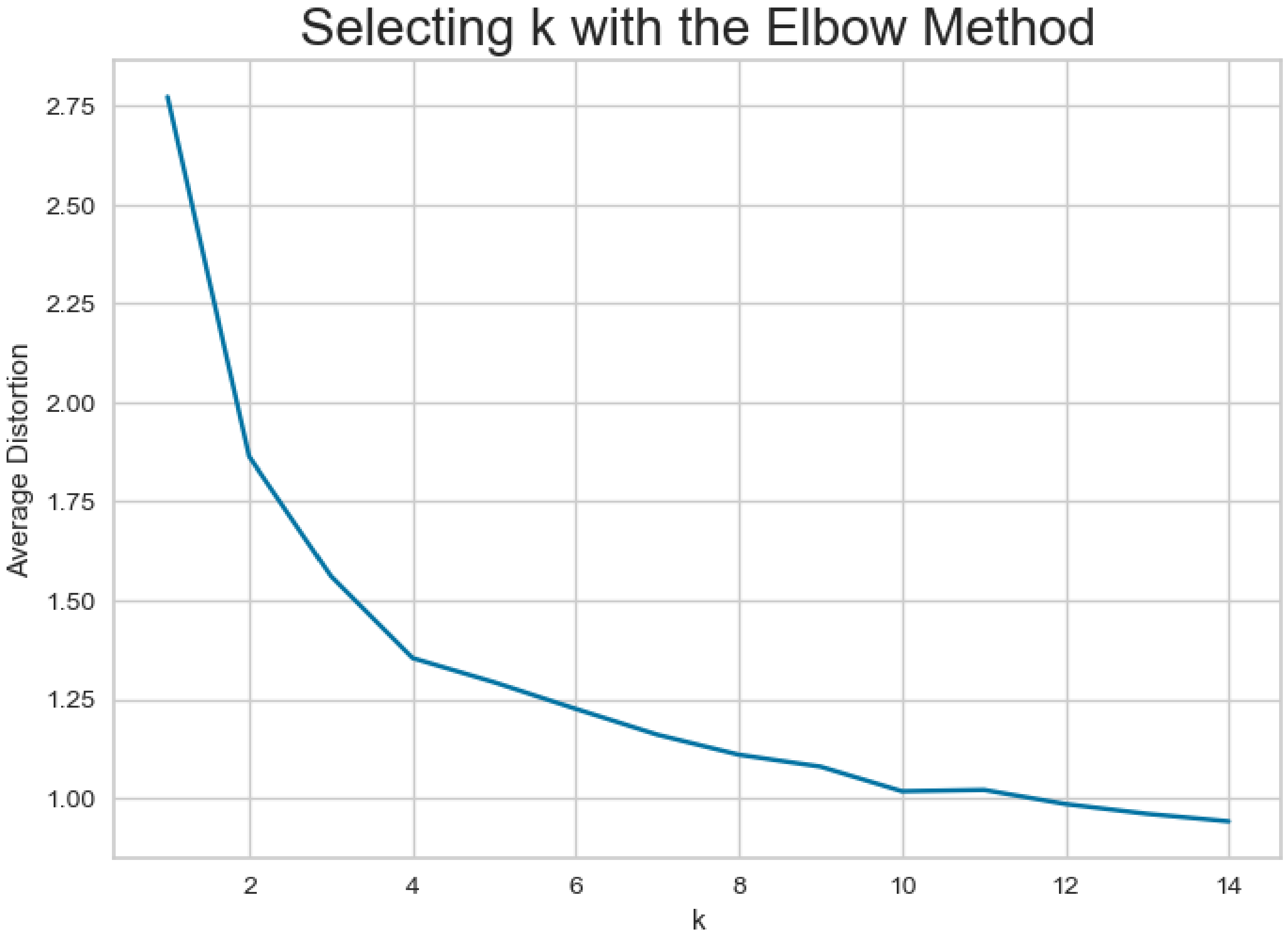
3. Applying K-means Clustering:

Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling



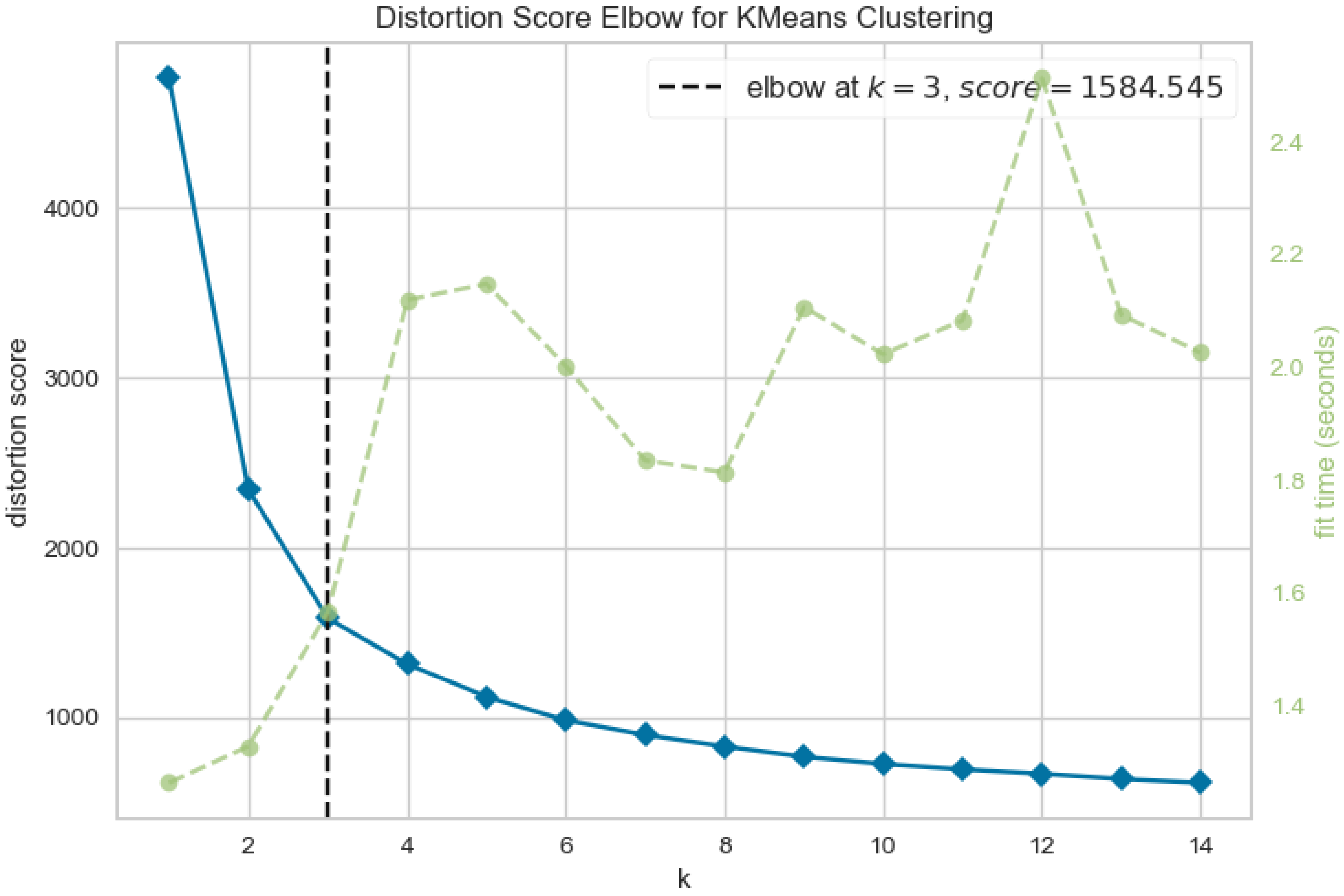
```
Number of Clusters: 1   Average Distortion: 2.773037110097803
Number of Clusters: 2   Average Distortion: 1.863573678589827
Number of Clusters: 3   Average Distortion: 1.5612774038101604
Number of Clusters: 4   Average Distortion: 1.3542782238901416
Number of Clusters: 5   Average Distortion: 1.2931541699741689
Number of Clusters: 6   Average Distortion: 1.225849543585495
Number of Clusters: 7   Average Distortion: 1.16048401421345
Number of Clusters: 8   Average Distortion: 1.109804758457438
Number of Clusters: 9   Average Distortion: 1.0797310475776052
Number of Clusters: 10  Average Distortion: 1.01743699264l063
Number of Clusters: 11  Average Distortion: 1.020874702026782
Number of Clusters: 12  Average Distortion: 0.9850734409030882
Number of Clusters: 13  Average Distortion: 0.9602766985773118
Number of Clusters: 14  Average Distortion: 0.9413187781558086
```

Distoration

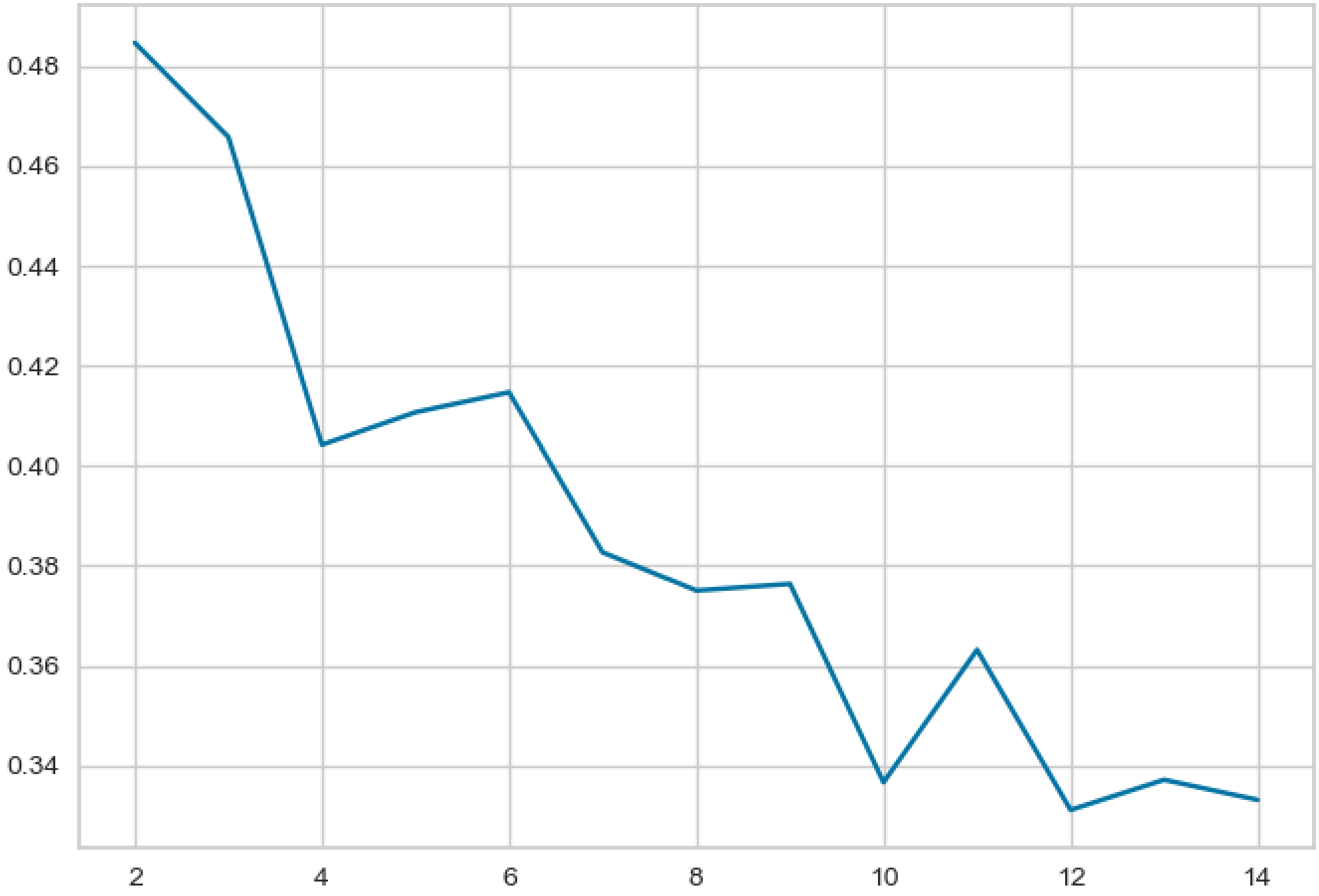


Elbow curve

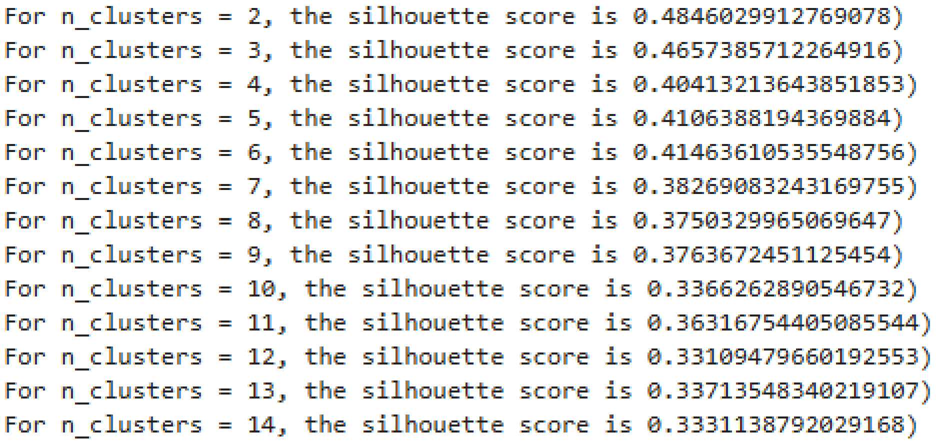
We can take 3 clusters. After 3 we have the clusters going parallel to X-axis.



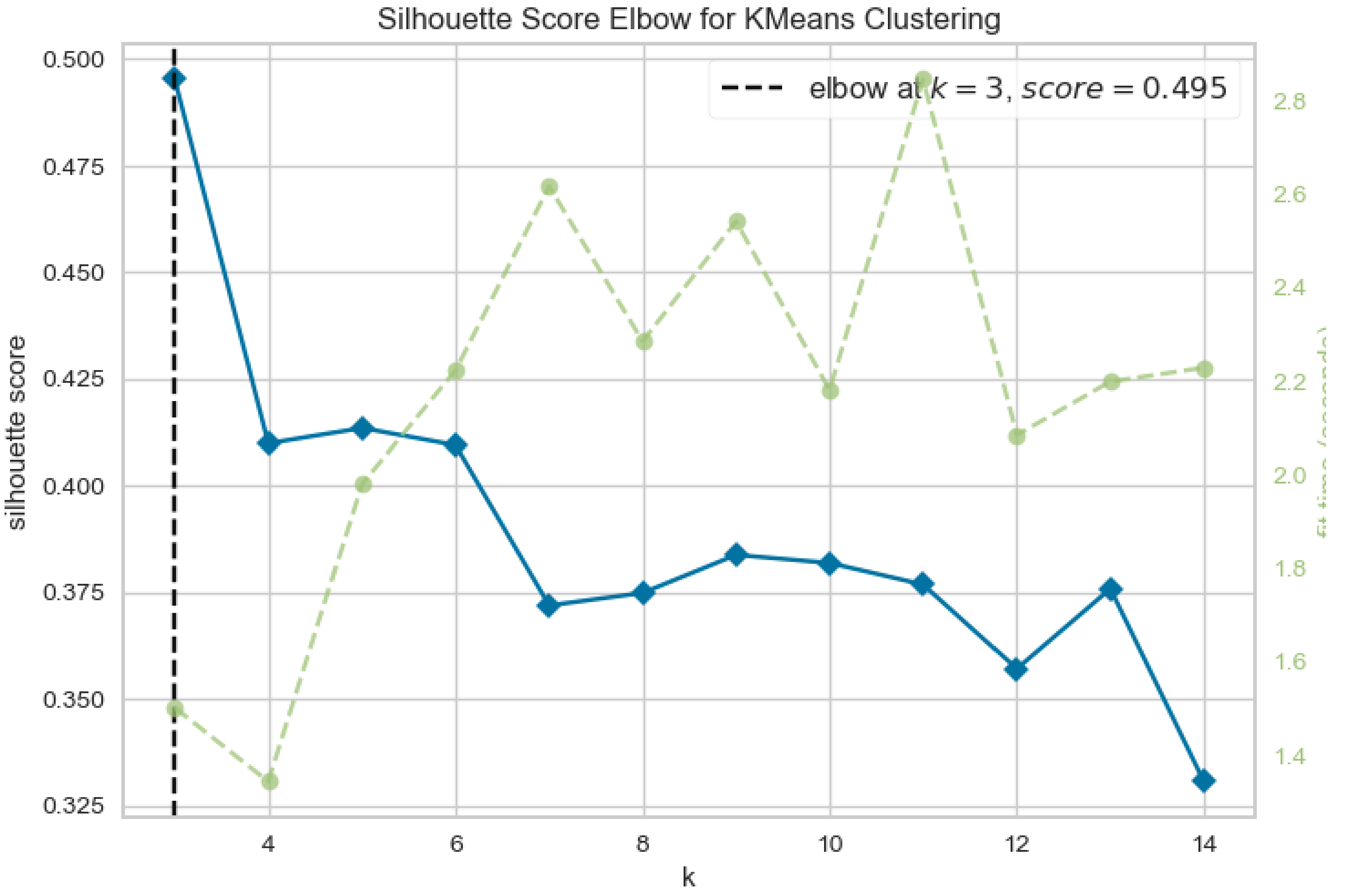
Distoration score for 3 clusters



Silhouette score



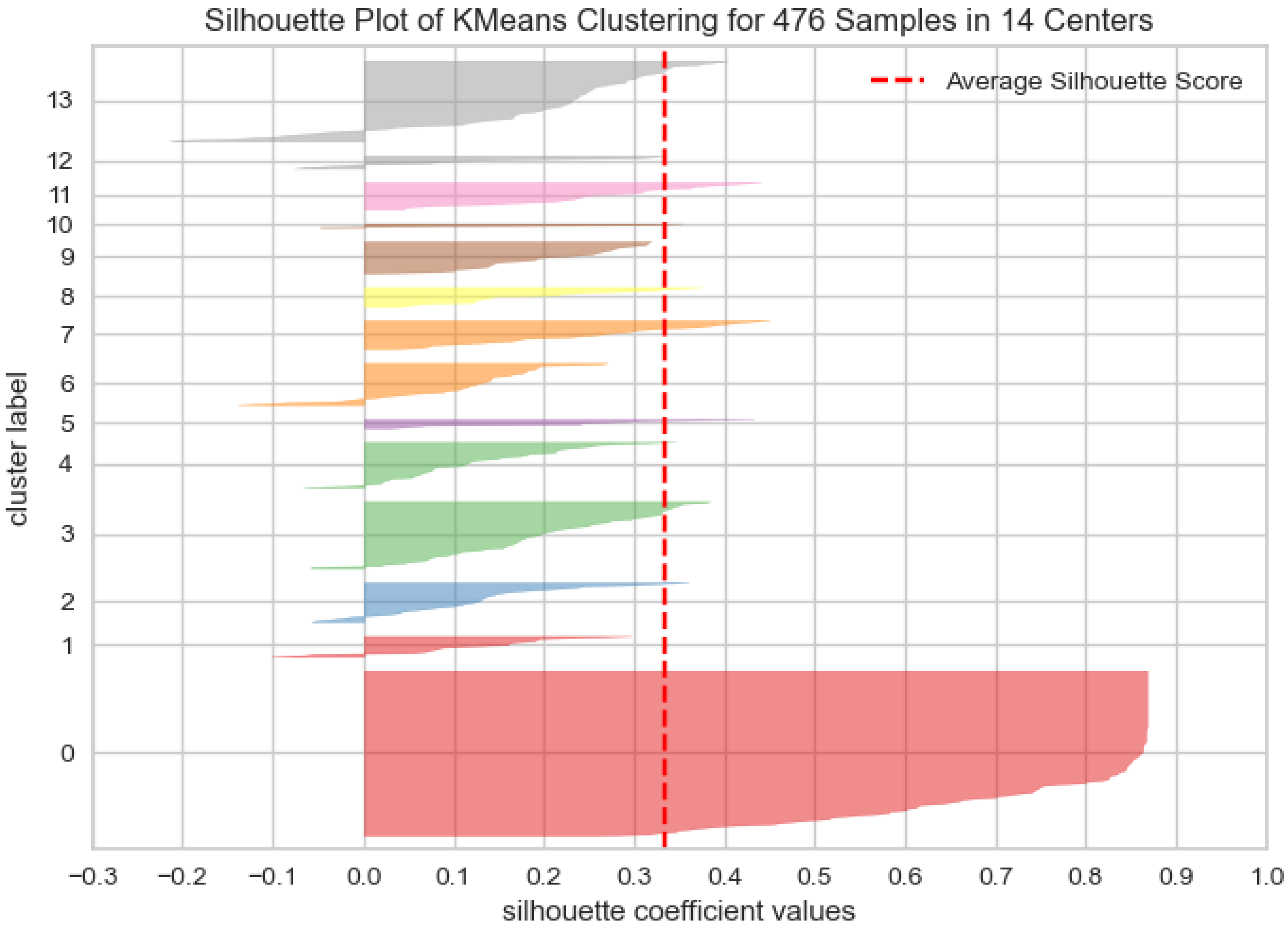
```
For n_clusters = 2, the silhouette score is 0.4846029912769078)
For n_clusters = 3, the silhouette score is 0.4657385712264916)
For n_clusters = 4, the silhouette score is 0.40413213643851853)
For n_clusters = 5, the silhouette score is 0.4106388194369884)
For n_clusters = 6, the silhouette score is 0.41463610535548756)
For n_clusters = 7, the silhouette score is 0.38269083243169755)
For n_clusters = 8, the silhouette score is 0.3750329965069647)
For n_clusters = 9, the silhouette score is 0.3763672451125454)
For n_clusters = 10, the silhouette score is 0.336626289054673 2)
For n_clusters = 11, the silhouette score is 0.36316754405085544)
For n_clusters = 12, the silhouette score is 0.33109479660192553)
For n_clusters = 13, the silhouette score is 0.33713548340219107)
For n_clusters = 14, the silhouette score is 0.3331138792029168)
```



Silhouette score elbow graph

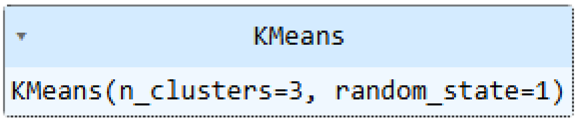
Observations:

- After analyzing the silhouette scores for different k values, it seems 3 is a good value for the number of clusters.



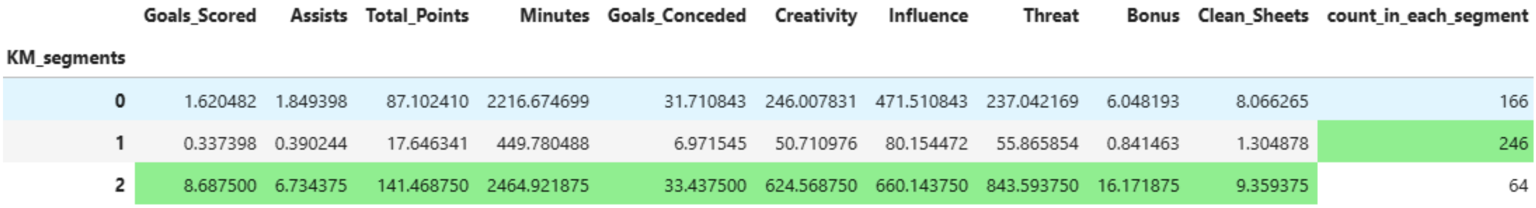
Silhouette Plot of KMeans Clustering for 476 Samples in 14 Centers

Applying KMeans clustering:



```
KMeans
KMeans(n_clusters=3, random_state=1)
```

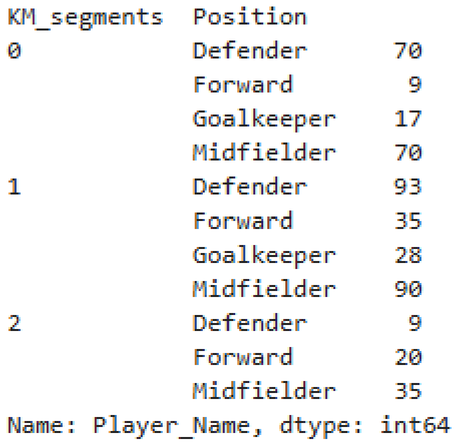
Cluster Profiling:



	Goals_Scored	Assists	Total_Points	Minutes	Goals_Conceded	Creativity	Influence	Threat	Bonus	Clean_Sheets	count_in_each_segment
KM_segments											
0	1.620482	1.849398	87.102410	2216.674699	31.710843	246.007831	471.510843	237.042169	6.048193	8.066265	166
1	0.337398	0.390244	17.646341	449.780488	6.971545	50.710976	80.154472	55.865854	0.841463	1.304878	246
2	8.687500	6.734375	141.468750	2464.921875	33.437500	624.568750	660.143750	843.593750	16.171875	9.359375	64

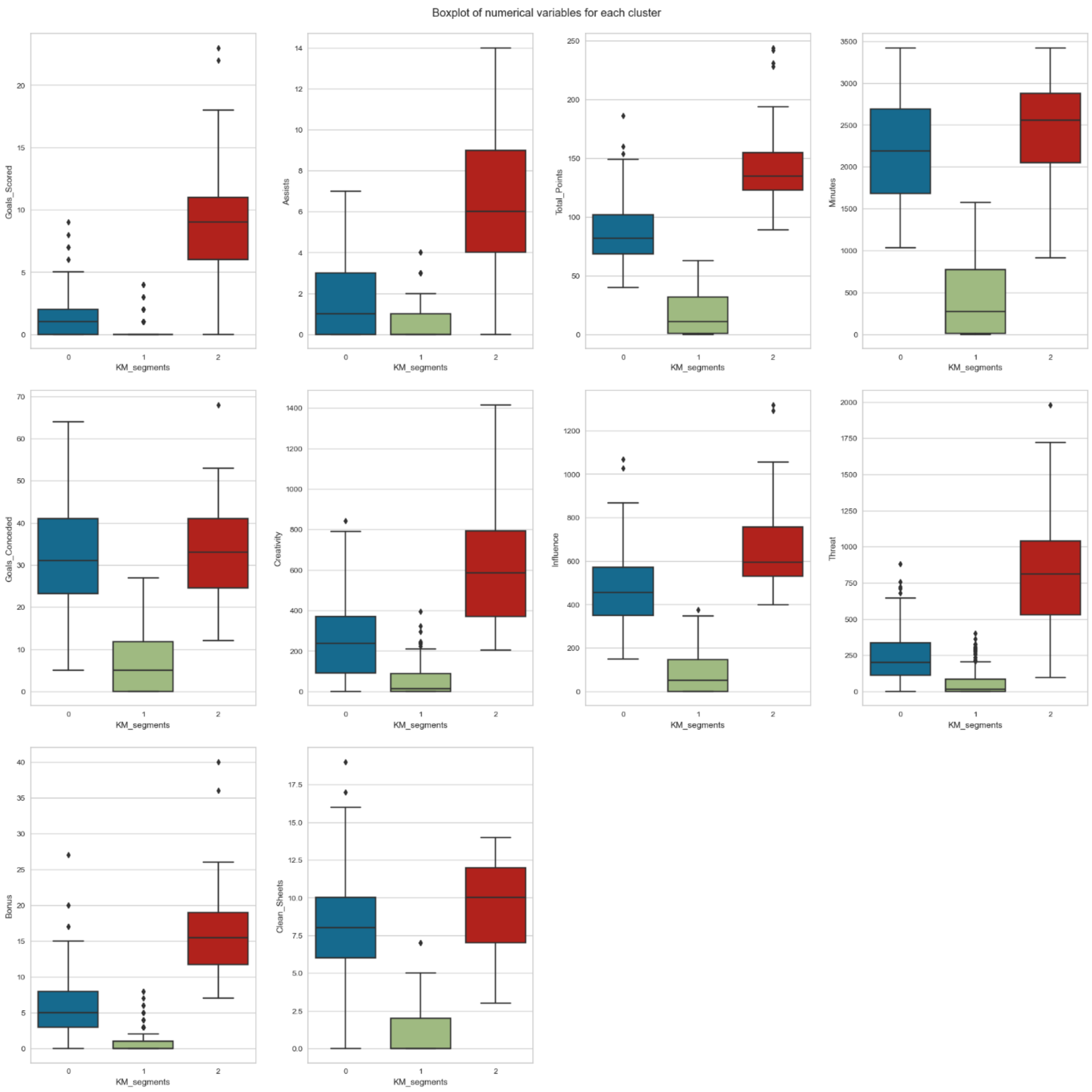
Players in each cluster:

```
In cluster 1, the following players are present:
['Alex Runnarsson' 'Calum Chambers' 'Cedric Soares' 'David Luiz'
 'Edward Nketiah' 'Gabriel Teodoro Martinelli Silva' 'Martin Odegaard'
 'Matt Macey' 'Miguel Azeez' 'Mohamed Naser El Sayed Elneny' 'Pablo Mari'
 'Reiss Nelson' 'Sead Kolasinac' 'Shkodran Mustafi'
 'Sokratis Papastathopoulos' 'William Saliba' 'Ahmed El Mohamady'
 'Carney Chukwuemeka' 'Conor Hourihane' 'Henri Lansbury' 'Jacob Ramsey'
 'Jaden Philogene-Bidace' 'Jose Peleteiro Romallo' 'Keinan Davis'
 'Kortney Hause' 'Marvelous Nakamba' 'Morgan Sanson' 'Orjan Nyland'
 'Wesley Moraes' 'Aaron Connolly' 'Alexis Mac Allister'
 'Alireza Jahanbakhsh' 'Andi Zeqiri' 'Bernardo Fernandes da Silva Junior'
 'Davy Propper' 'Jakub Moder' 'Jason Steele' 'Jayson Molumby'
 'Jose Izquierdo' 'Mathew Ryan' 'Percy Tau' 'Reda Khadra' 'Steven Alzate'
 'Tariq Lamptey' 'Anthony Driscoll-Glennon' 'Bailey Peacock-Farrell'
 'Dale Stephens' 'Erik Pieters' 'Jack Cork' 'Jeff Hendrick' 'Jimmy Dunne'
 'Joel Mumbongo' 'Johann Berg Gudmundsson' 'Josh Benson' 'Kevin Long'
 'Lewis Richardson' 'Phil Bardsley' 'Robbie Brady' 'Will Norris'
 'Billy Gilmour' 'Emerson Palmieri dos Santos' 'Faustino Anjorin'
 'Fikayo Tomori' 'Karlo Ziger' 'Kepa Arrizabalaga' 'Marcos Alonso'
 'Olivier Giroud' 'Valentino Livramento' 'Willy Caballero'
 'Connor Wickham' 'Jack Butland' 'James McCarthy' 'James Tomkins'
 'Jean-Philippe Mateta' 'Mamadou Sakho' 'Martin Kelly' 'Michy Batshuayi'
 'Nathan Ferguson' 'Nathaniel Clyne' 'Reece Hannam' 'Ryan Inniss'
 'Sam Woods' 'Scott Dann' 'Stephen Henderson' 'Andre Tavares Gomes'
 'Anthony Gordon' 'Bernard Caldeira Duarte' 'Cenk Tosun' 'Fabian Delph'
 'Joao Virginia' 'Jonjoe Kenny' 'Joshua King' 'Moise Kean'
 'Nathan Broadhead' 'Niels Nkounkou' 'Robin Olsen' 'Tom Davies'
 'Adam Forshaw' 'Diego Llorente' 'Francisco Casilla' 'Gaetano Berardi'
 'Ian Carlo Poveda-Ocampo' 'Jack Jenkins' 'Jamie Shackleton'
 'Jay-Roy Grot' 'Jordan Stevens' 'Kamil Miazek' 'Leif Davis'
 'Mateusz Bogusz' 'Niall Huggins' 'Pablo Hernandez' 'Robin Koch'
 'Tyler Roberts' 'Cengiz Under' 'Christian Fuchs' 'Daniel Amartey'
 'Demarai Gray' 'Dennis Praet' 'Filip Benkovic' 'Hamza Choudhury'
 'Islam Slimani' 'Luke Thomas' 'Nampalys Mendy'
 'Ricardo Domingos Barbosa Pereira' 'Sidnei Tavares' 'Thakgalo Leshabela'
 'Vontae Daley-Campbell' 'Wes Morgan' 'Adrian Castillo'
 'Alex Oxlade-Chamberlain' 'Caoimhin Kelleher' 'Curtis Jones'
 'Dean Henderson' 'Divock Origi' 'James Milner' 'Joel Matip'
 'Joseph Gomez' 'Naby Keita' 'Neco Williams' 'Ozan Kabak' 'Rhys Williams'
 'Virgil van Dijk' 'Xherdan Shaqiri' 'Aymeric Laporte' 'Benjamin Mendy'
 'Eric Garcia' 'Liam Delap' 'Luke Mbete' 'Nathan Ake' 'Nicolas Otamendi'
 'Scott Carson' 'Sergio Aguero' 'Taylor Harwood-Bellis' 'Zack Steffen'
 'Alex Nicolao Telles' 'Amad Diallo' 'Anthony Elanga' 'Axel Tuanzebe'
 'Brandon Williams' 'Daniel James' 'Donny van de Beek' 'Eric Bailly'
 'Hannibal Mejbri' 'Juan Mata' 'Nathan Bishop' 'Nemanja Matic'
 'Odion Ighalo' 'Shola Shoretire' 'William Fish' 'Andy Carroll'
 'DeAndre Yedlin' 'Dwight Gayle' 'Elliot Anderson' 'Emil Krafth'
 'Fabian Schar' 'Florian Lejeune' 'Javier Manquillo' 'Kelland Watts'
 'Martin Dubravka' 'Matthew Longstaff' 'Paul Dummett' 'Ryan Fraser'
 'Sean Longstaff' 'Yoshinori Muto' 'Caleb Watts' "Daniel N'Lundulu"
 'Fraser Forster' 'Ibrahima Diallo' 'Jake Vokins' 'Kgaogelo Chauke'
 'Michael Obafemi' 'Mohammed Salisu' 'Moussa Djenepo' 'Nathan Tella'
 'Shane Long' 'Takumi Minamino' 'William Smallbone' 'Yan Valery'
 'Bamidele Alli' 'Ben Davies' 'Cameron Carter-Vickers'
 'Carlos Vinicius Alves Morais' 'Dane Scarlett' 'Danny Rose'
 'Davinson Sanchez' 'Erik Lamela' 'Giovani Lo Celso' 'Harry Winks'
 'Japhet Tanganga' 'Joe Rodon' 'Juan Foyth' 'Matt Doherty'
 'Moussa Sissoko' 'Paulo Gazzaniga' 'Ryan Sessegnon' 'Steven Bergwijn'
 'Ademipo Odubeko' 'Albian Ajeti' 'Andriy Yarmolenko' 'Arthur Masuaku'
 'Ben Johnson' 'Darren Randolph' 'Fabian Balbuena'
 'Felipe Anderson Pereira Gomes' 'Frederik Alves' 'Issa Diop'
 'Jamal Baptiste' 'Jordan Hugill' 'Manuel Lanzini' 'Mark Noble'
 'Roberto Jimenez Gago' 'Ryan Fredericks' 'Sebastian Haller'
 'Fernando Marcal' 'John Ruddy' 'Jonathan Castro Otto' 'Ki-Jana Hoever'
 'Max Kilman' 'Morgan Gibbs-White' 'Oskar Buur' 'Owen Otasowie'
 'Patrick Cutrone' 'Raul Jimenez ' 'Ruben Vinagre' 'Vitor Ferreira'
 'Willian Jose']

In cluster 2, the following players are present:
['Alexandre Lacazette' 'Bukayo Saka' 'Nicolas Pepe'
 'Pierre-Emerick Aubameyang' 'Anwar El Ghazi' 'Bertrand Traore'
 'Jack Grealish' 'Ollie Watkins' 'Leandro Trossard' 'Neal Maupay'
 'Pascal Gross' 'Chris Wood' 'Benjamin Chilwell' 'Mason Mount'
 'Timo Werner' 'Christian Benteke' 'Eberechi Eze' 'Wilfried Zaha'
 'Dominic Calvert-Lewin' 'Gylfi Sigurdsson' 'James Rodriguez'
 'Lucas Digne' 'Richarlison de Andrade' 'Jack Harrison' 'Patrick Bamford'
 'Raphael Dias Belloli' 'Rodrigo Moreno' 'Stuart Dallas' 'Harvey Barnes'
 'James Maddison' 'Jamie Vardy' 'Kelechi Iheanacho' 'Youri Tielemans'
 'Andrew Robertson' 'Mohamed Salah' 'Roberto Firmino' 'Sadio Mane'
 'Trent Alexander-Arnold' 'Gabriel Fernando de Jesus' 'Ilkay Gundogan'
 'Joao Cancelo' 'Kevin De Bruyne' 'Phil Foden' 'Raheem Sterling'
 'Riyad Mahrez' 'Bruno Fernandes' 'Edinson Cavani' 'Luke Shaw'
 'Marcus Rashford' 'Callum Wilson' 'Che Adams' 'Danny Ings'
 'James Ward-Prowse' 'Gareth Bale' 'Harry Kane' 'Heung-Min Son'
 'Aaron Cresswell' 'Jarrod Bowen' 'Jesse Lingard' 'Michail Antonio'
 'Pablo Fornals' 'Tomas Soucek' 'Vladimir Coufal' 'Pedro Lomba Neto']

In cluster 0, the following players are present:
['Bernd Leno' 'Daniel Ceballos' 'Emile Smith Rowe' 'Gabriel Maghalaes'
 'Granit Xhaka' 'Hector Bellerin' 'Kieran Tierney' 'Rob Holding'
 'Thomas Partey' 'Willian Borges Da Silva' 'Douglas Luiz Soares de Paulo'
 'Emiliano Martinez' 'Ezri Konsa Ngoyo' 'John McGinn'
 'Mahmoud Ahmed Ibrahim Hassan' 'Matt Targett' 'Matthew Cash'
 'Ross Barkley' 'Tyrone Mings' 'Adam Lallana' 'Adam Webster' 'Ben White'
 'Dan Burn' 'Danny Welbeck' 'Joel Veltman' 'Lewis Dunk' 'Robert Sanchez'
 'Solomon March' 'Yves Bissouma' 'Ashley Barnes' 'Ashley Westwood'
 'Ben Mee' 'Charlie Taylor' 'Dwight McNeil' 'James Tarkowski'
 'Jay Rodriguez' 'Josh Brownhill' 'Matej Vydra' 'Matthew Lowton'
 'Nick Pope' 'Andreas Christensen' 'Antonio Rudiger' 'Callum Hudson-Odoi'
 'Cesar Azpilicueta' 'Christian Pulisic' 'Edouard Mendy' 'Hakim Ziyech'
 'Jorge Luiz Frello Filho' 'Kai Havertz' 'Kurt Zouma' 'Mateo Kovacic'
 "N'Golo Kante" 'Reece James' 'Tammy Abraham' 'Thiago Silva'
 'Andros Townsend' 'Cheikhou Kouyate' 'Gary Cahill' 'Jairo Riedewald'
 'James McArthur' 'Jeffrey Schlupp' 'Joel Ward' 'Jordan Ayew'
 'Luka Milivojevic' 'Patrick van Aanholt' 'Tyrick Mitchell'
 'Vicente Guaita' 'Abdoulaye Doucoure' 'Alex Iwobi'
 'Allan Marques Loureiro' 'Ben Godfrey' 'Jordan Pickford' 'Mason Holgate'
 'Michael Keane' 'Seamus Coleman' 'Yerry Mina' 'Ezgjan Alioski'
 'Helder Costa' 'Illan Meslier' 'Kalvin Phillips' 'Liam Cooper'
 'Luke Ayling' 'Mateusz Klich' 'Pascal Struijk' 'Ayoze Perez'
 'Calgar Soyuncu' 'James Justin' 'Jonny Evans' 'Kasper Schmeichel'
 'Marc Albrighton' 'Timothy Castagne' 'Wesley Fofana' 'Wilfred Ndidi'
 'Alisson Becker' 'Diogo Jota' 'Fabio Henrique Tavares'
 'Georginio Wijnaldum' 'Jordan Henderson' 'Nathaniel Phillips'
 'Thiago Alcantara' 'Bernardo Silva' 'Ederson Moares' 'Fernando Luiz Rosa'
 'Ferran Torres' 'John Stones' 'Kyle Walker' 'Oleksandr Zinchenko'
 'Rodrigo Hernandez' 'Ruben Dias' 'Aaron Wan-Bissaka' 'Anthony Martial'
 'David de Gea' 'Frederico Rodrigues de Paula Santos' 'Harry Maguire'
 'Mason Greenwood' 'Paul Pogba' 'Scott McTominay' 'Victor Lindelof'
 'Allan Saint-Maximin' 'Ciaran Clark' 'Federico Fernandez' 'Isaac Hayden'
 'Jacob Murphy' 'Jamaal Lascelles' 'Jamal Lewis' 'Joelinton de Lira'
 'Jonjo Shelvey' 'Joseph Willock' 'Karl Darlow' 'Matt Ritchie'
 'Miguel Almiron' 'Alex McCarthy' 'Jack Stephens' 'Jan Bednarek'
 'Jannik Vestergaard' 'Kyle Walker-Peters' 'Nathan Redmond'
 'Oriol Romeu Vidal' 'Ryan Bertrand' 'Stuart Armstrong' 'Theo Walcott'
 'Eric Dier' 'Hugo Lloris' 'Lucas Moura' 'Pierre-Emile Hojbjerg'
 'Serge Aurier' 'Sergio Reguilon' 'Tanguy Ndombele' 'Toby Alderweireld'
 'Angelo Ogbonna' 'Craig Dawson' 'Declan Rice' 'Lukasz Fabianski'
 'Said Benrahma' 'Adama Traore' 'Conor Coady' 'Daniel Castelo Podence'
 'Fabio Silva' 'Joao Santos Moutinho' 'Leander Dendoncker' 'Nelson Semedo'
 'Rayan Ait Nouri' 'Romain Saiss' 'Ruben Neves' 'Rui Pedro Patricio'
 'Willy Boly']
```



```
KM_segments  Position
0            Defender      70
             Forward        9
             Goalkeeper    17
             Midfielder    70
1            Defender      93
             Forward       35
             Goalkeeper    28
             Midfielder    90
2            Defender       9
             Forward       20
             Midfielder    35
Name: Player_Name, dtype: int64
```

Cluster segmentation position wise



Cluster box plots

Observations:

Characteristics of each cluster:

Cluster 0

- There are 166 players in this cluster.
- Most of the players in this cluster have a few goals and assists, and the total fantasy points scored in the previous season are midium.
- Most of the players in this cluster had a moderate game time, midium creativity score, midium influence score, and a moderate threat score.
- Most of the players in this cluster received midium bonus points.

Cluster 1

- There are 246 players in this cluster.
- Most of the players in this cluster have a few goals and assists, and the total fantasy points scored in the previous season are lowest.
- Most of the players in this cluster had lowest game time, lowest creativity score, lowest influence score, and lowest threat score.
- Most of the players in this cluster received lowest bonus points.

Cluster 2

- There are 64 players in this cluster.
- Most of the players in this cluster have a lots of goals and assists, and the total fantasy points scored in the previous season are high.
- Most of the players in this cluster had a high game time, a high creativity aand influence scores.
- Most of the players in this cluster received high bonus points.

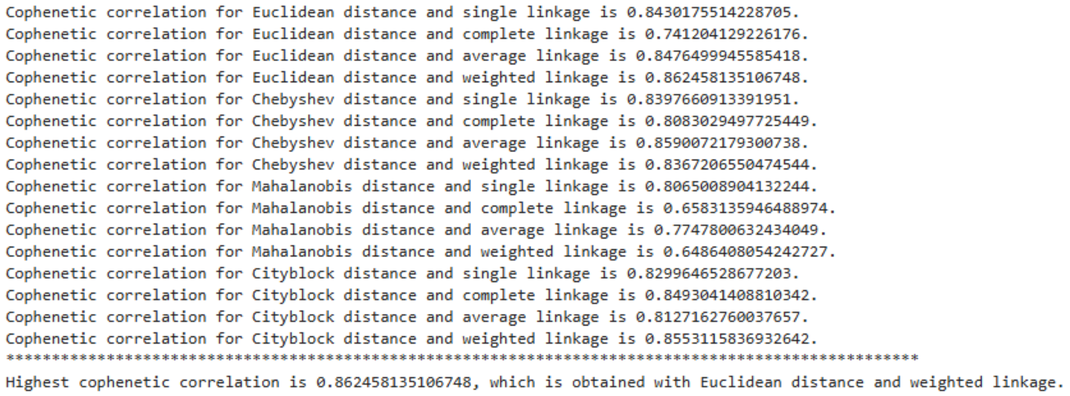
From this we can say that:

- Cluster 2 are the high value players who have performed exeptionally well last season.
- Cluster 0 are the moderate value players who have performed well last season.
- Cluster 1 are the low value players who have performed poorly last season despite getting game time last season.

4. Applying Hierarchical Clustering:

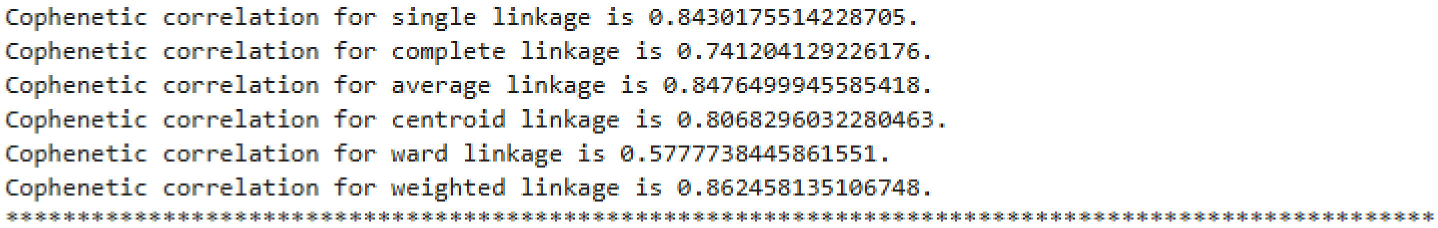
Apply Hierarchical clustering with different linkage methods - Plot dendrograms for each linkage method - Check cophenetic correlation for each linkage method - Figure out the appropriate number of clusters - Cluster Profiling

4.a



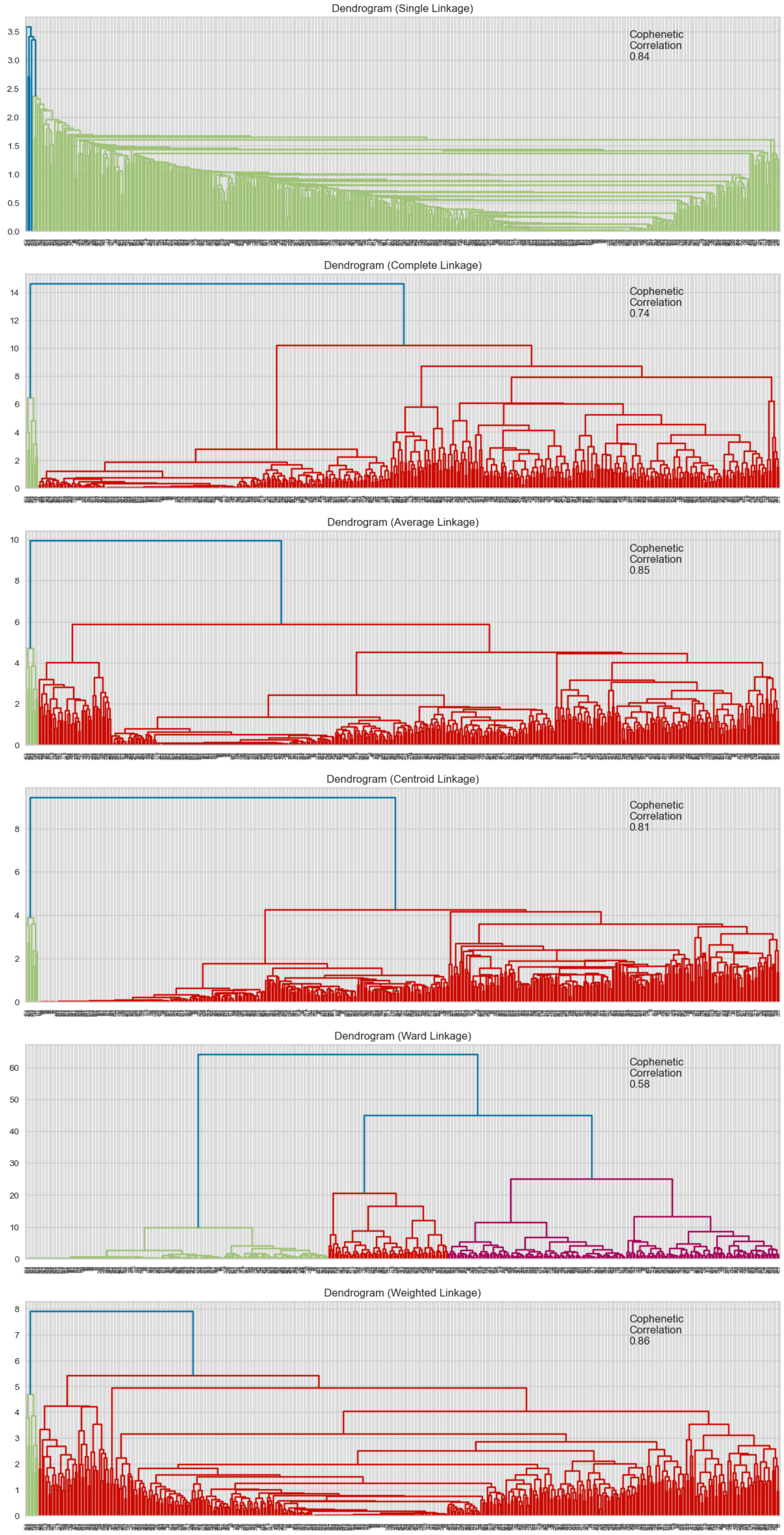
```
Cophenetic correlation for Euclidean distance and single linkage is 0.8430175514228705.
Cophenetic correlation for Euclidean distance and complete linkage is 0.741204129226176.
Cophenetic correlation for Euclidean distance and average linkage is 0.8476499945585418.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.862458135106748.
Cophenetic correlation for Chebyshev distance and single linkage is 0.8397660913391951.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.8083029497725449.
Cophenetic correlation for Chebyshev distance and average linkage is 0.8590072179300738.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.8367206550474544.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.8065008904132244.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.6583135946488974.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.7747800632434049.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.6486408054242727.
Cophenetic correlation for Cityblock distance and single linkage is 0.8299646528677203.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8493041408810342.
Cophenetic correlation for Cityblock distance and average linkage is 0.8127162760037657.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.8553115836932642.
****************************************************************************************************
Highest cophenetic correlation is 0.862458135106748, which is obtained with Euclidean distance and weighted linkage.
```

- Highest cophenetic correlation is 0.862458135106748, which is obtained with Euclidean distance and weighted linkage.

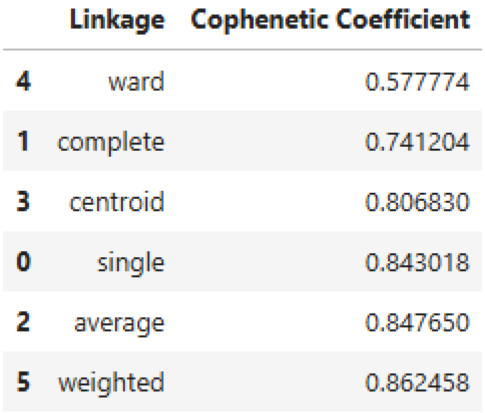


```
Cophenetic correlation for single linkage is 0.8430175514228705.
Cophenetic correlation for complete linkage is 0.741204129226176.
Cophenetic correlation for average linkage is 0.8476499945585418.
Cophenetic correlation for centroid linkage is 0.8068296032280463.
Cophenetic correlation for ward linkage is 0.5777738445861551.
Cophenetic correlation for weighted linkage is 0.862458135106748.
****************************************************************************************************
```

- Highest cophenetic correlation is 0.862458135106748, which is obtained with weighted linkage.

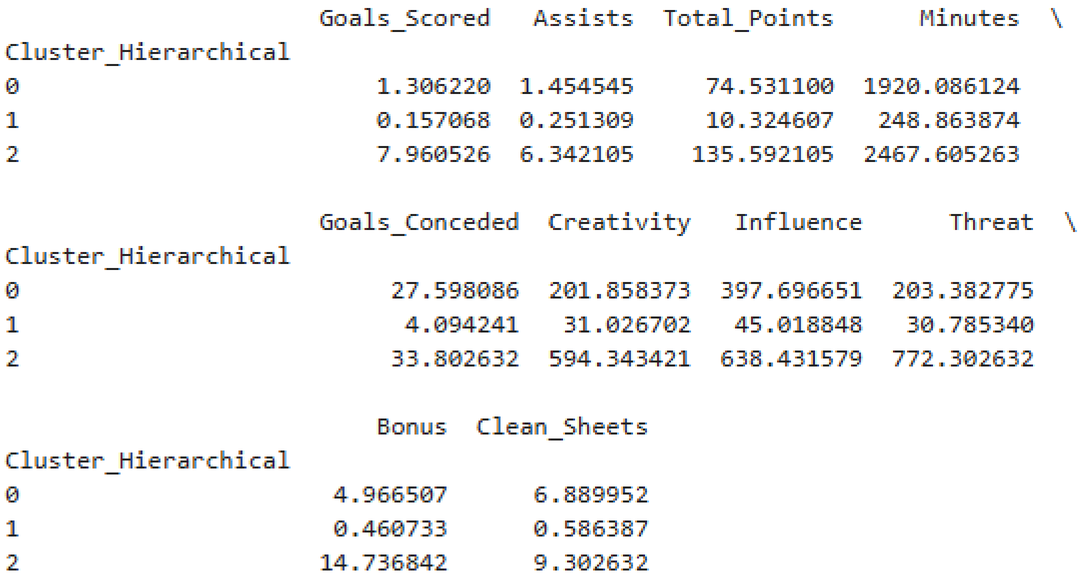

Dendrogram (Single Linkage)
Cophenetic
Correlation
0.84
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Dendrogram (Complete Linkage)
Cophenetic
Correlation
0.74
14
12
10
8
6
4
2
0
Dendrogram (Average Linkage)
Cophenetic
Correlation
0.85
10
8
6
4
2
0
Dendrogram (Centroid Linkage)
Cophenetic
Correlation
0.81
8
6
4
2
0
Dendrogram (Ward Linkage)
Cophenetic
Correlation
0.58
60
50
40
30
20
10
0
Dendrogram (Weighted Linkage)
Cophenetic
Correlation
0.86
8
7
6
5
4
3
2
1
0

Dendogram



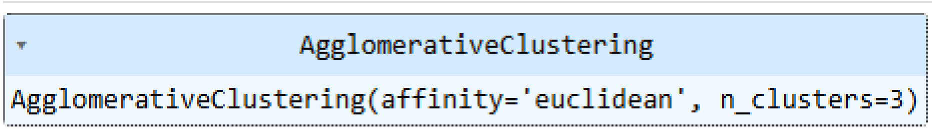
	Linkage	Cophenetic Coefficient
4	ward	0.577774
1	complete	0.741204
3	centroid	0.806830
0	single	0.843018
2	average	0.847650
5	weighted	0.862458

Linkage CC



```
                      Goals_Scored   Assists  Total_Points      Minutes  \
Cluster_Hierarchical
0                         1.306220  1.454545     74.531100  1920.086124
1                         0.157068  0.251309     10.324607   248.863874
2                         7.960526  6.342105    135.592105  2467.605263

                      Goals_Conceded  Creativity   Influence      Threat  \
Cluster_Hierarchical
0                          27.598086  201.858373  397.696651  203.382775
1                           4.094241   31.026702   45.018848   30.785340
2                          33.802632  594.343421  638.431579  772.302632

                          Bonus  Clean_Sheets
Cluster_Hierarchical
0                      4.966507      6.889952
1                      0.460733      0.586387
2                     14.736842      9.302632
```

Applying Hierarchical Clustering with optimal linkage



```
AgglomerativeClustering(affinity='euclidean', n_clusters=3)
```

Creating model using SK learn

Actionable Insights and Recommendations:

Recommendations:

- Cluster 2 players are the top players for fantasy. They fetch more points and have a higher chance of getting bonus points too. These players should be priced higher than the others so that it will be difficult to accommodate too many of them in the same team (because of the fixed budget) and fantasy managers have to make wise choices.
- Cluster 1 players are players who do not play many minutes, most likely come on as substitutes and fetch lesser fantasy points as a result. These players should be priced low and can be good differential picks.
- Cluster 0 are the players who are influential in their team's play but do not tend to score or assist much, resulting in lesser fantasy points than the Cluster 2 players. These players should be priced somewhere between the Cluster 2 and Cluster 1 players.
- Cluster 1 has the players who are in the squad to provide backup in case any of the starting 11 players get injured. They get lower game time and barely get any fantasy points. These players should be priced the lowest amongst the 3 clusters.
- Player performances from previous seasons should be taken into account and fantasy prices from the previous season should be referred to as a benchmark to determine the price for the upcoming season.
- OnSports should conduct cluster analysis separately for each of the playing positions to arrive at a better fantasy pricing strategy, given that football is heavily biased towards offensive players.

END