OnSport Fantasy Sports Cluster Analysis

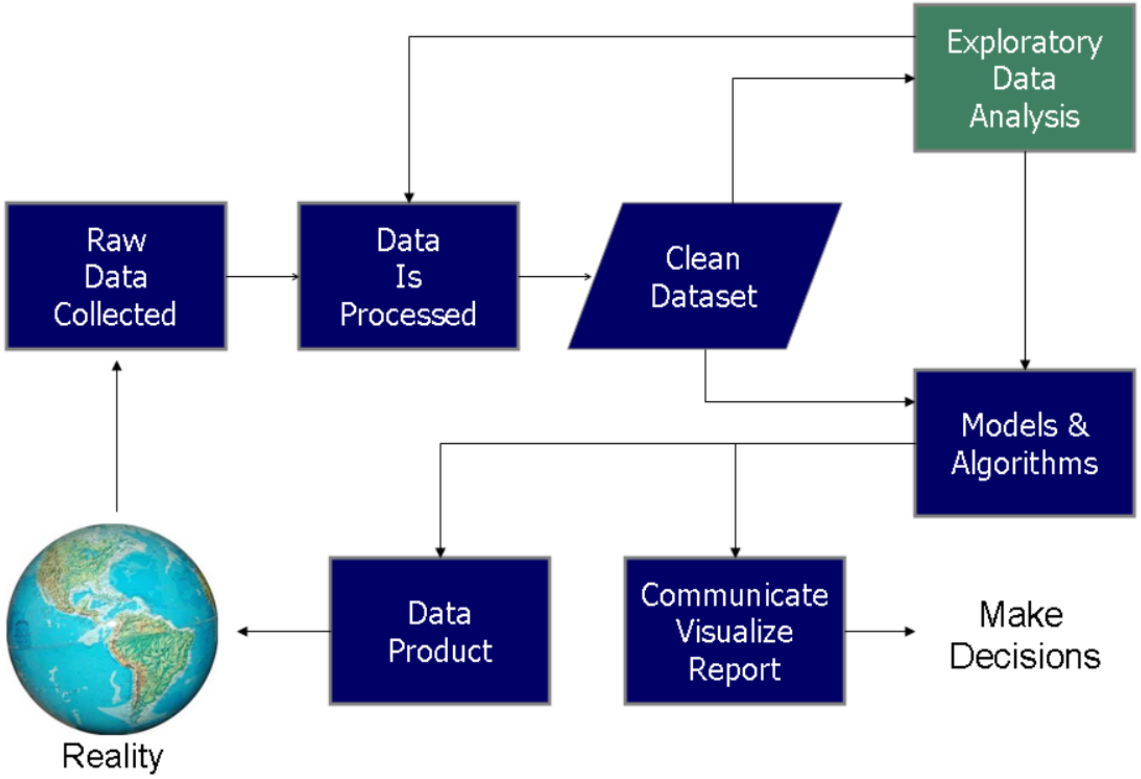
Contents

- Problem Definition
- Data Preprocessing
- Exploratory Data Analysis
- K-Means Clustering
    - Elbow Curve
    - Silhouette plot
    - Cluster Profiling
- Hierarchical Clustering
    - Linkages
    - Cophenetic Correlation
    - Cluster Profiling
- Comparison between Hierarchical and K-means Clustering
- Key Takeaways

Problem Definition

- OnSports is a fantasy sports platform
- Each player is given price at start which depends upon real world performance
- For the next season, data has been provided for player’s performance in previous season
- As a data scientist it was asked to perform cluster analysis
  - To identify players of different potentials
  - To understand patterns in player performance and fantasy returns
  - To help OnSports set price for each player

Data Science Process



Data Preprocessing

- To properly work with our data, we must clean it.

1. The Player names are normalized to remove unwanted characters
2. The Data had no missing values
3. Some outliers in the data was taken care by scaling the data using a min max scaler

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- Feature engineering was applied to manipulate the data to make it work better for ML Models.
  - The Positions of the players was changed to a numerically scaled value
  - Two features were added to the dataset

1. Effectiveness: It gives the understanding of number of points scored by a player per minute

$$\text{Effectiveness} = \frac{Total\ Points\ * 100}{Minutes\ Playes}$$

2. Match Performance: It explains the actual performance of the player in a match without the bonus points

$$Match\ performance = Total\ Point\ - Bonus$$

Exploratory Data Analysis(EDA)

- EDA is an approach to analyze data to summarize their main characteristics
- Usually visual methods are used
- We perform analysis on data that we collected
  - To find important metrics/features
  - By using some nice and pretty visualizations

Exploratory Data Analysis(EDA)

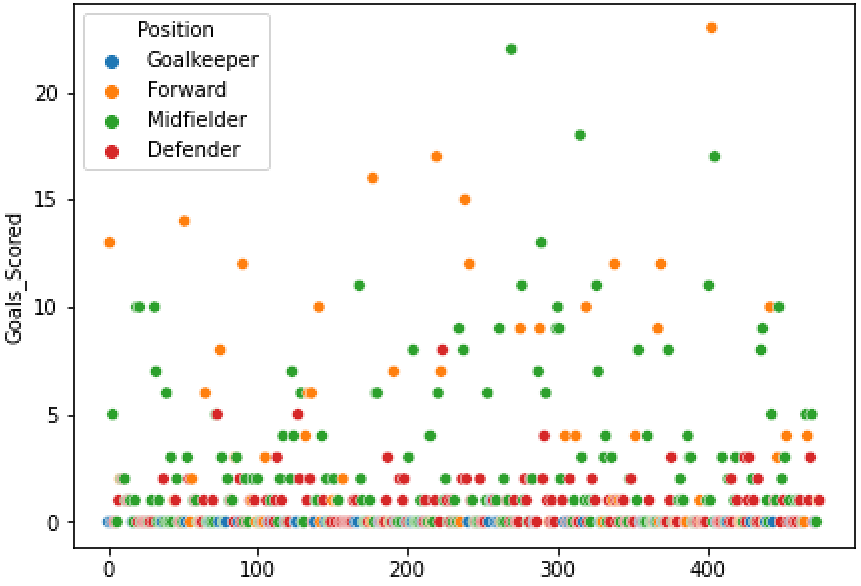
- EDA is majorly performed using two methods:

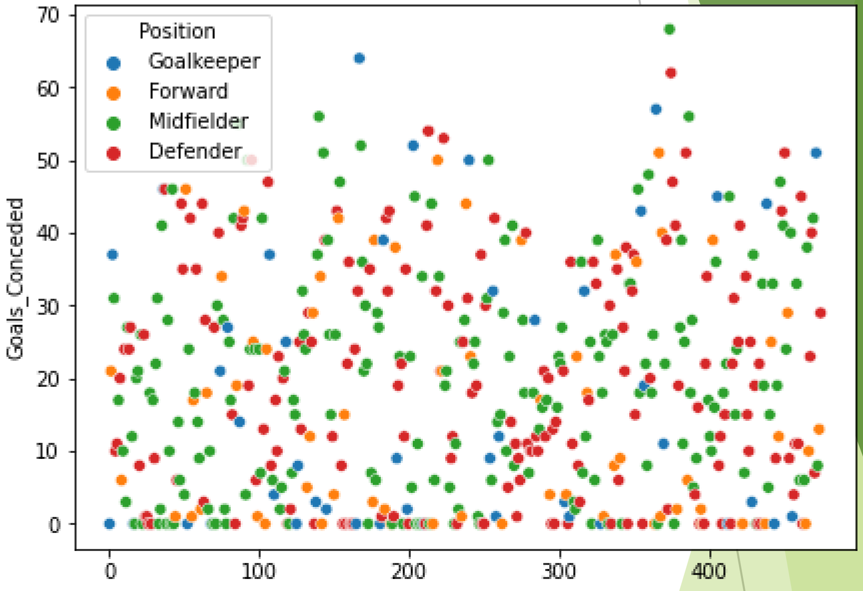
1. Univariate Analysis: Provides summary statistics for each field in the raw data set.
2. Bivariate Analysis: Performed to find the relationship between each variable in the dataset and the target variable of interest.

- The Following plots were used for EDA

1. Pair Plot: It shows a clear and nice view of all variables and their relationship with all other variables
2. Scatter Plot: Plots different observation of the same variable corresponding to index

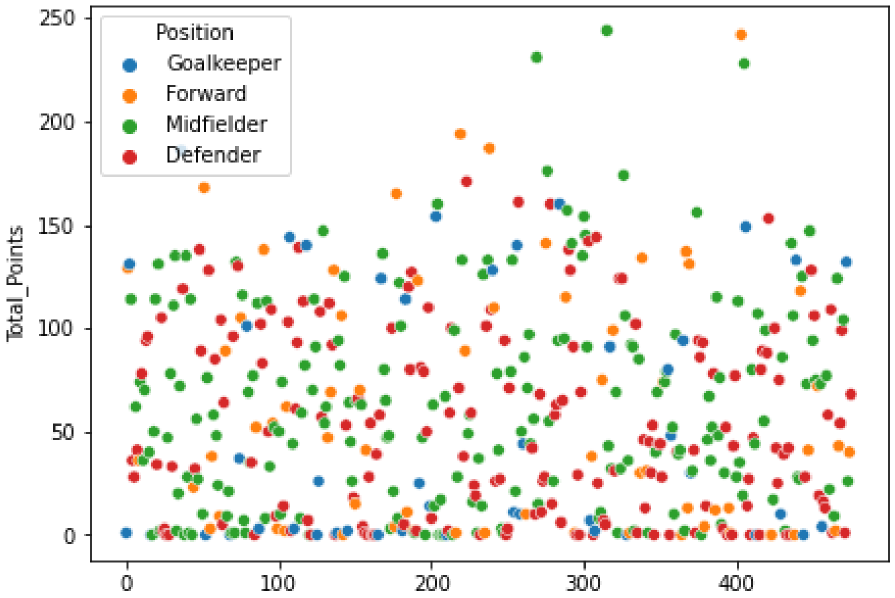
Univariate Analysis



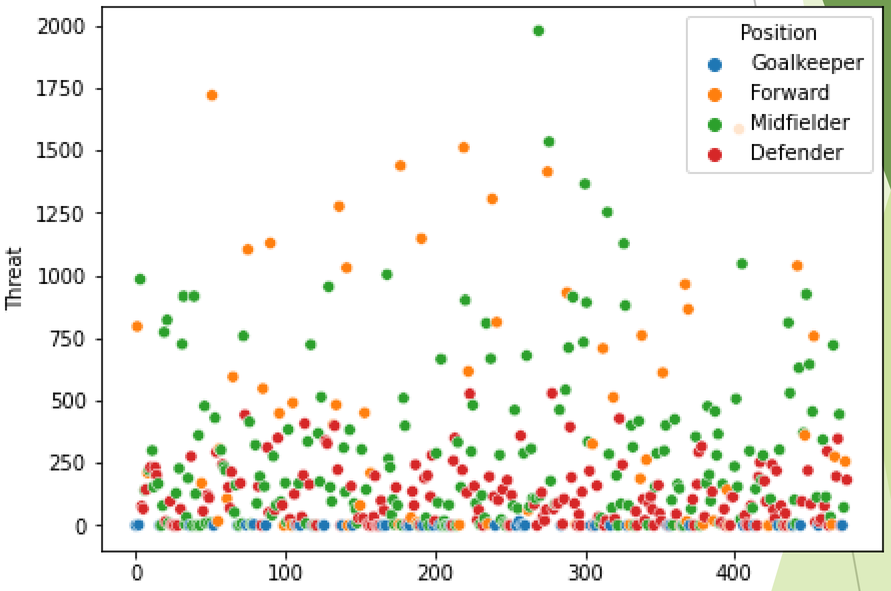


- From the graph, It can be seen that most goals are scored by forward and midfielder
- Highest goal scorer is forward, while lowest is goalkeeper

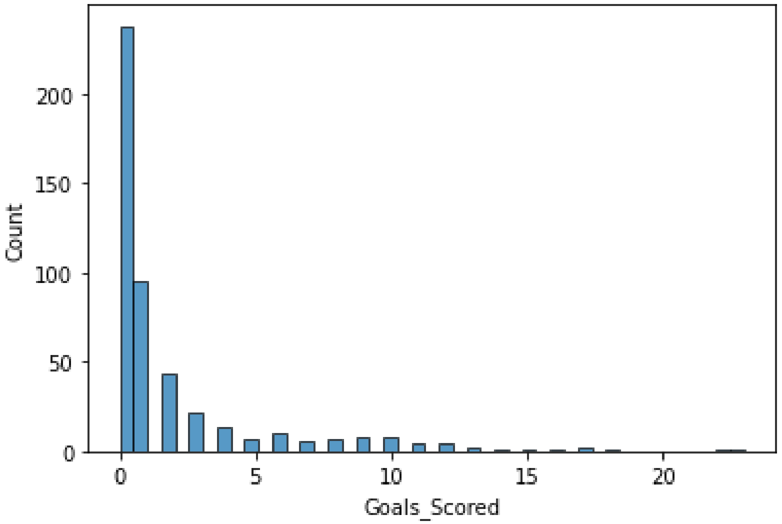
- From the graph, No pattern can be observed in no of goals conceded



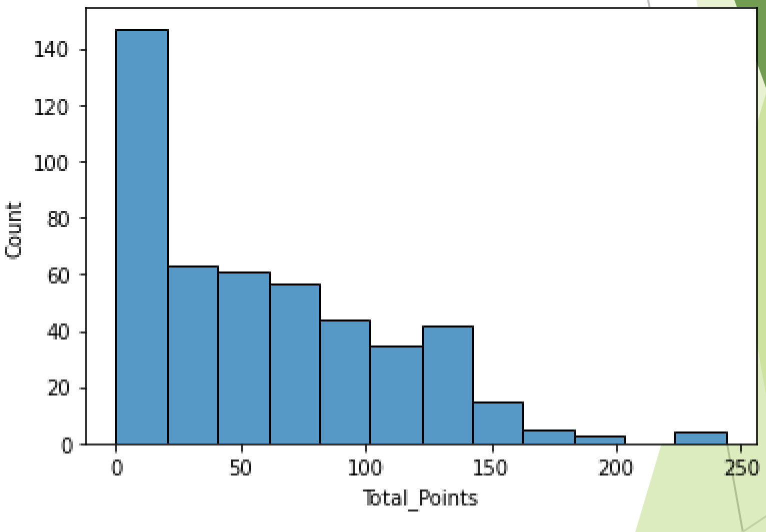
- More than 150 points are obtained by midfielder and forward players, and below 150 there is a no particular pattern



- Highest threat to other team is given by forward and midfielder
- Defender also poses some threat but goalkeeper poses least threat

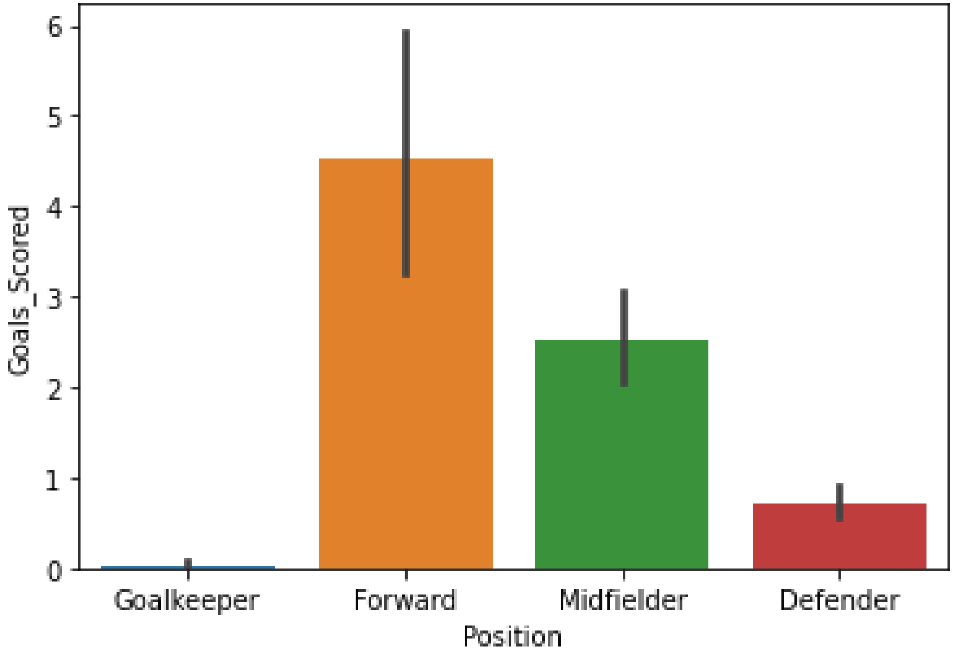


- From the graph, It can be seen that most players have not scored any goal
- Players scoring more that 10 goals are rare

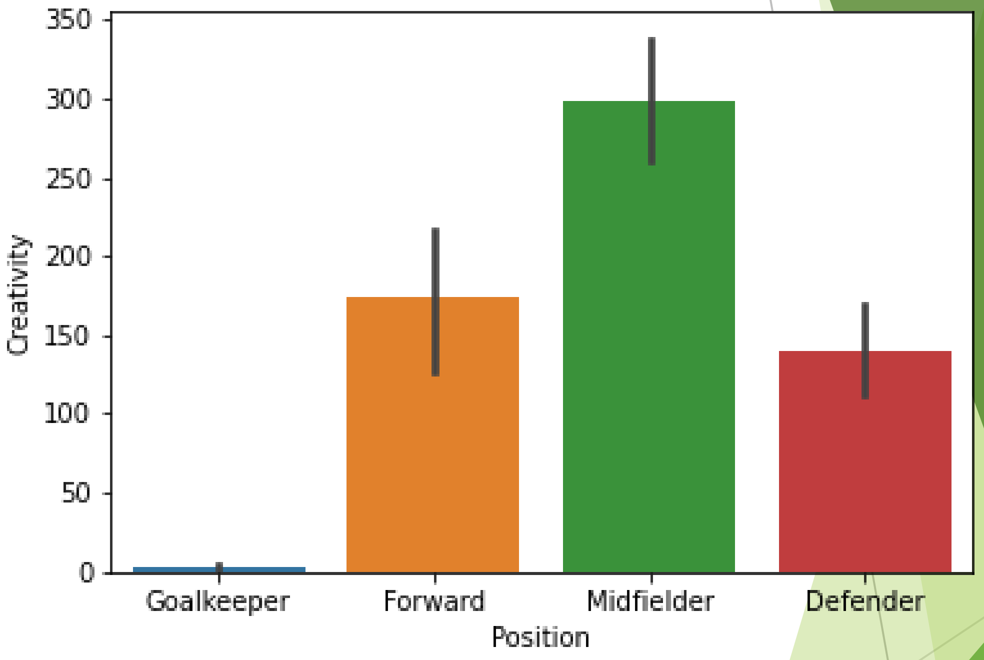


- From the graph, It can be seen that majority of players score between 0-150
- Players scoring above 150 are rare

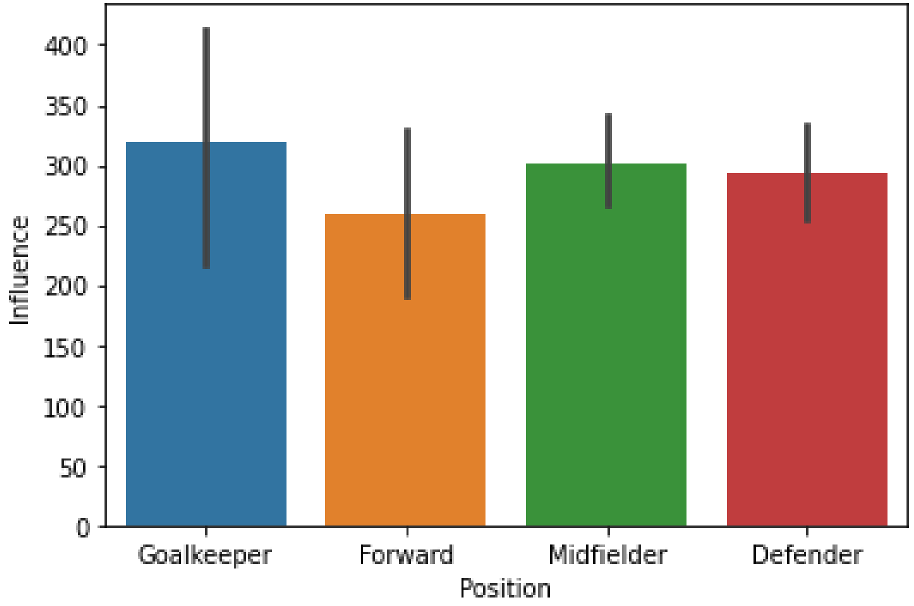
Bivariate Analysis



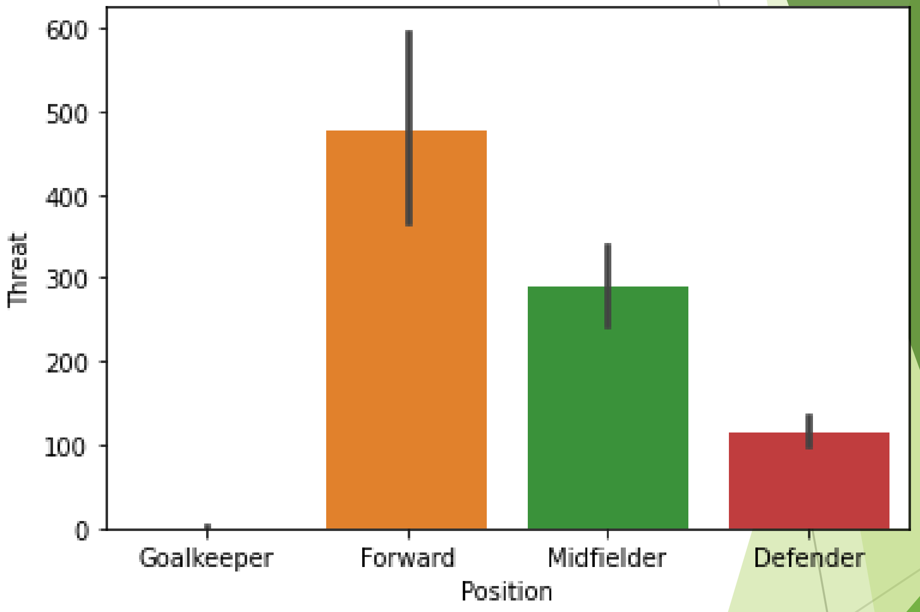
- From the plot, it can be seen that most of the goals are scored by forward and midfielder
- Goalkeeper scored nearly zero goal
- Average goal scored by forward is 4.5



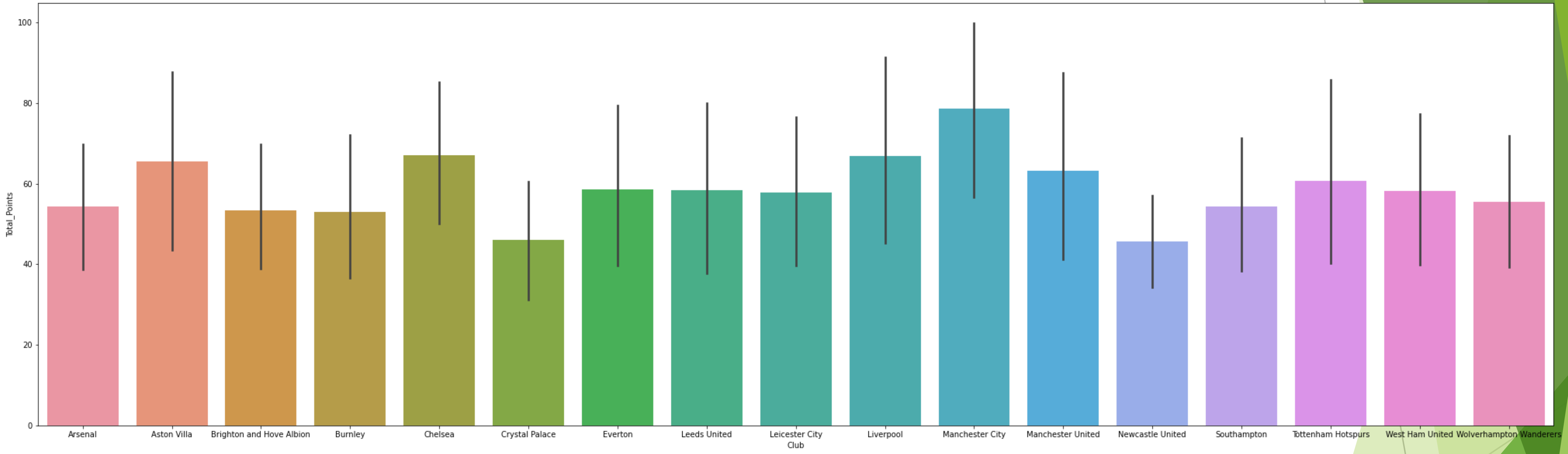
- From the graph, It can be seen that midfielder shows most creativity as it produces opportunities for the forward players
- Goal keeper shows least creativity in a match



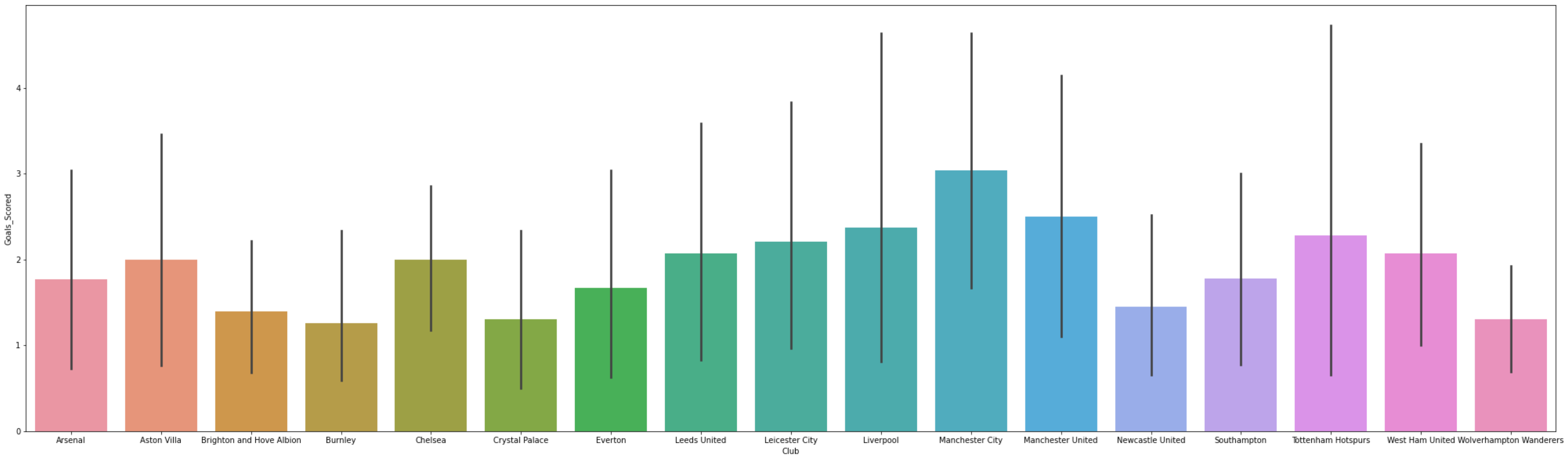
- From the graph, It can be seen that influence of all the players are nearly equal but goalkeeper has most influence



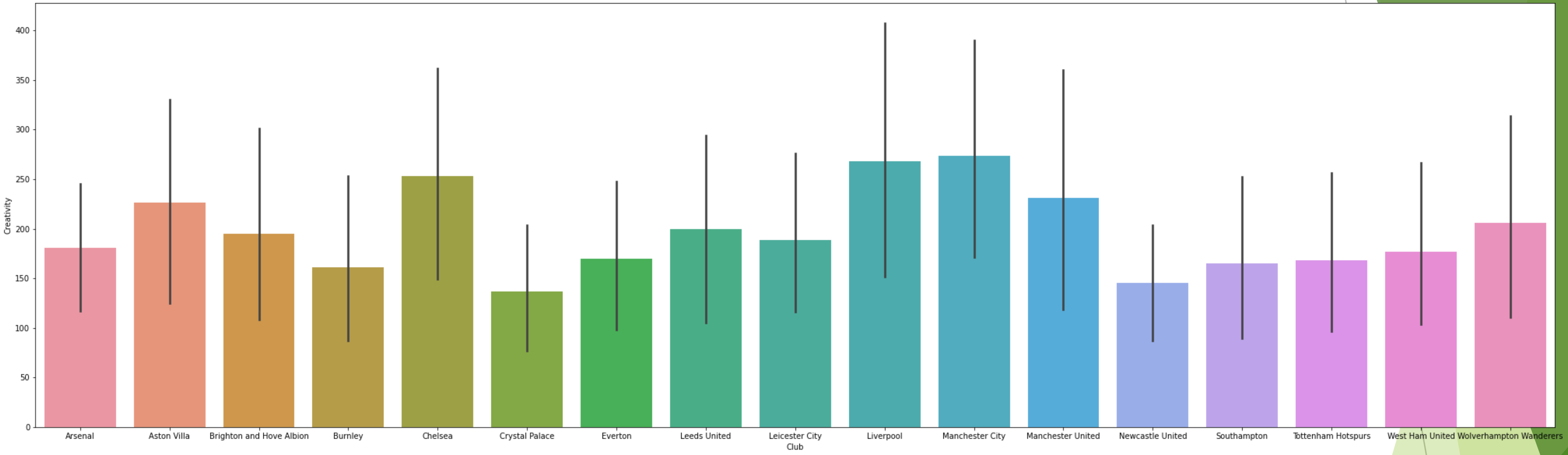
- From the graph, it can be seen that forward poses most threat as seen earlier too



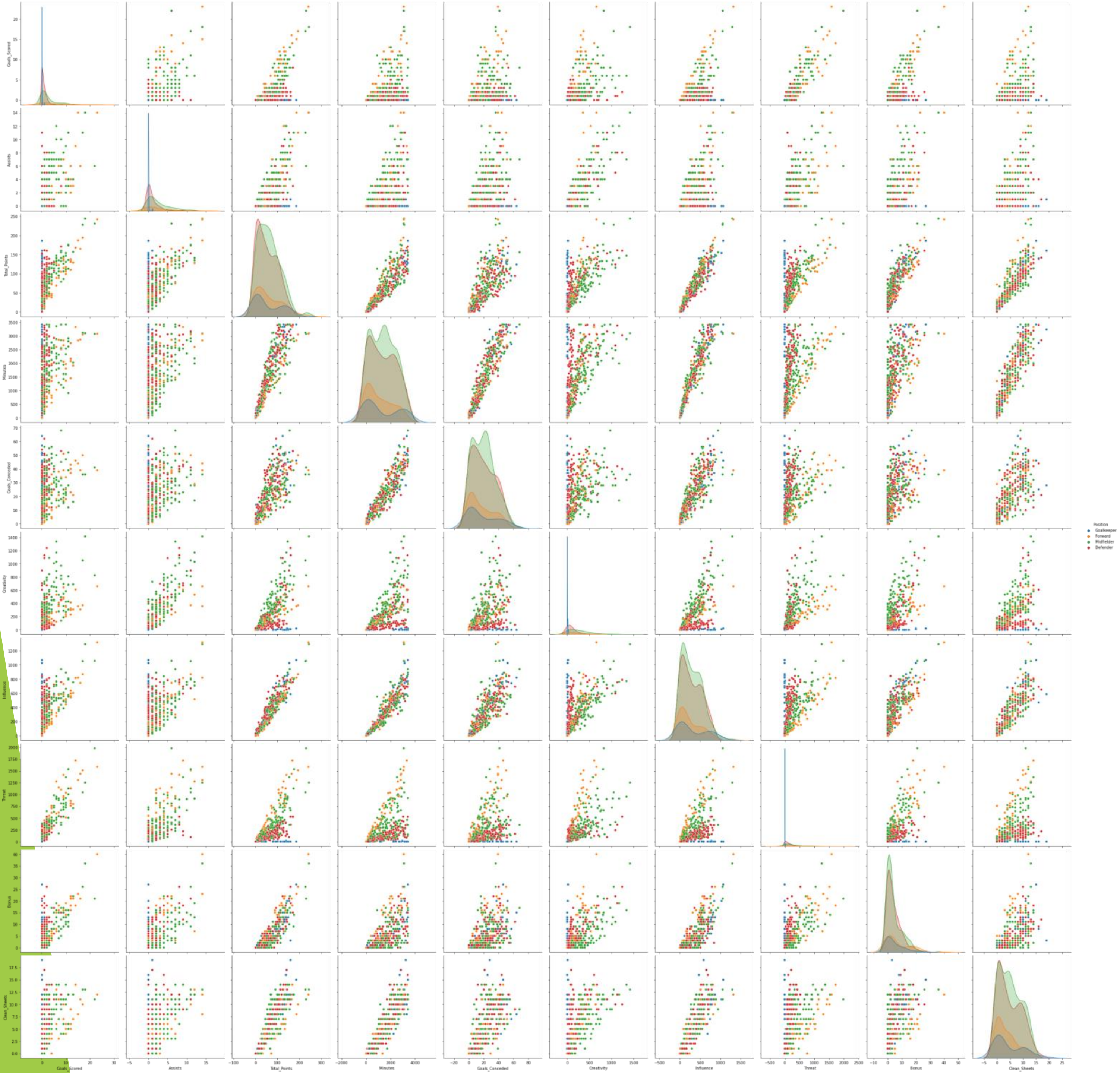
- From the graph, total points scored by teams are nearly equal with Manchester city having highest points
- Least points scored is by crystal palace



- From the graph, total goals scored by teams are nearly equal with Manchester city having highest no. of goals
- Tottenham hotspurs and Liverpool have some players which are really good scorers



- From the graph, it can be seen that Manchester city’s player have the highest influence in a match with crystal city being the least
- Highest variance is seen in Liverpool team, meaning some player in Liverpool have very high influence in a match



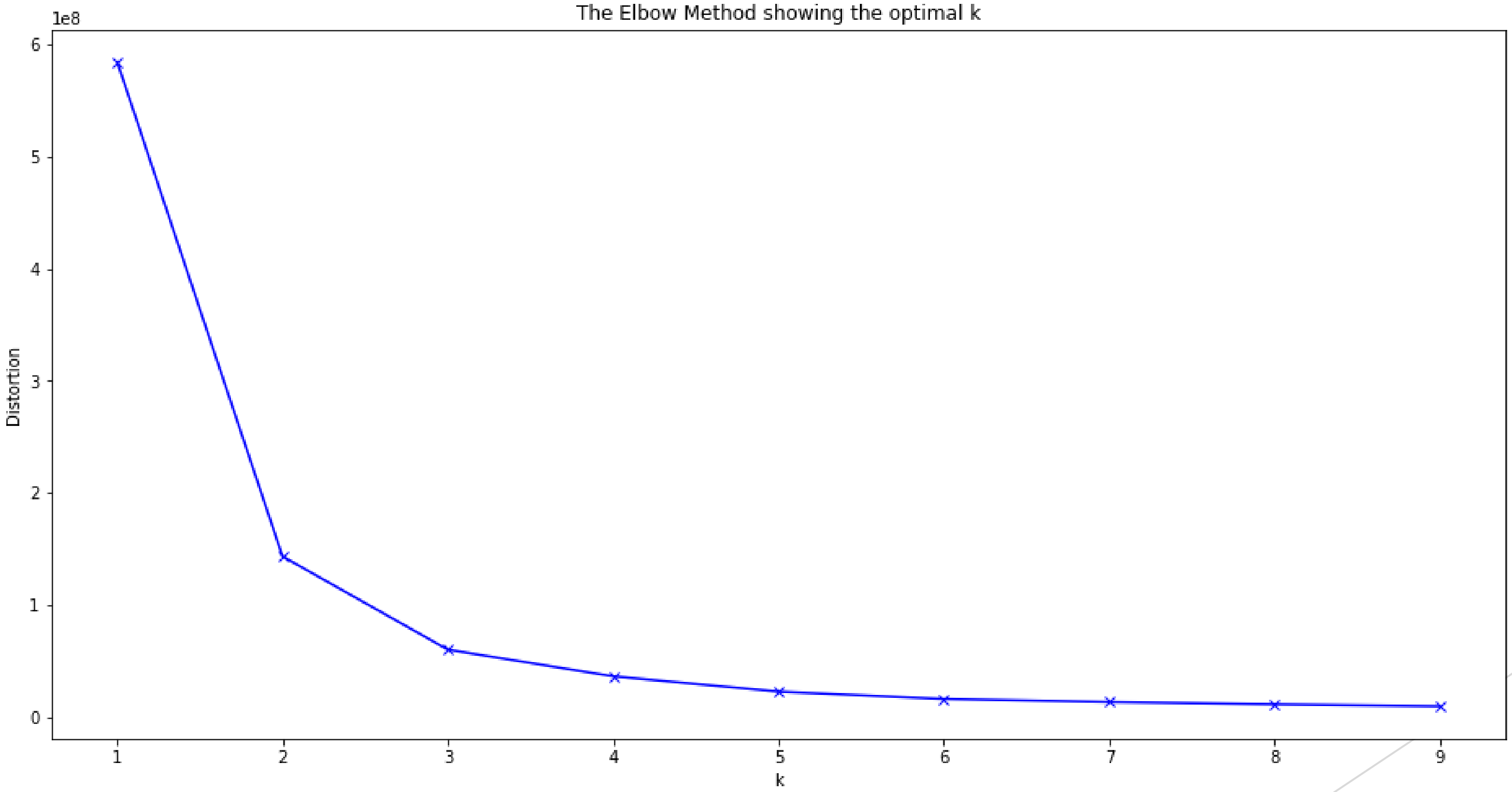
- This figure shows scatter plots of every two variables used in our analysis

K-Means Clustering

- The goal here is to group players based on a variety of scores
- K-means clustering can be explained as:
  - We define the number of clusters which refers to the number of centroids
  - Data points are assigned to the clusters by reducing the in-cluster sum of squares.
  - Process repeats until clusters have been minimized.
- In real world, we don’t know the number of clusters to be chosen
- One way to determine the optimum number of groups is using elbow Method

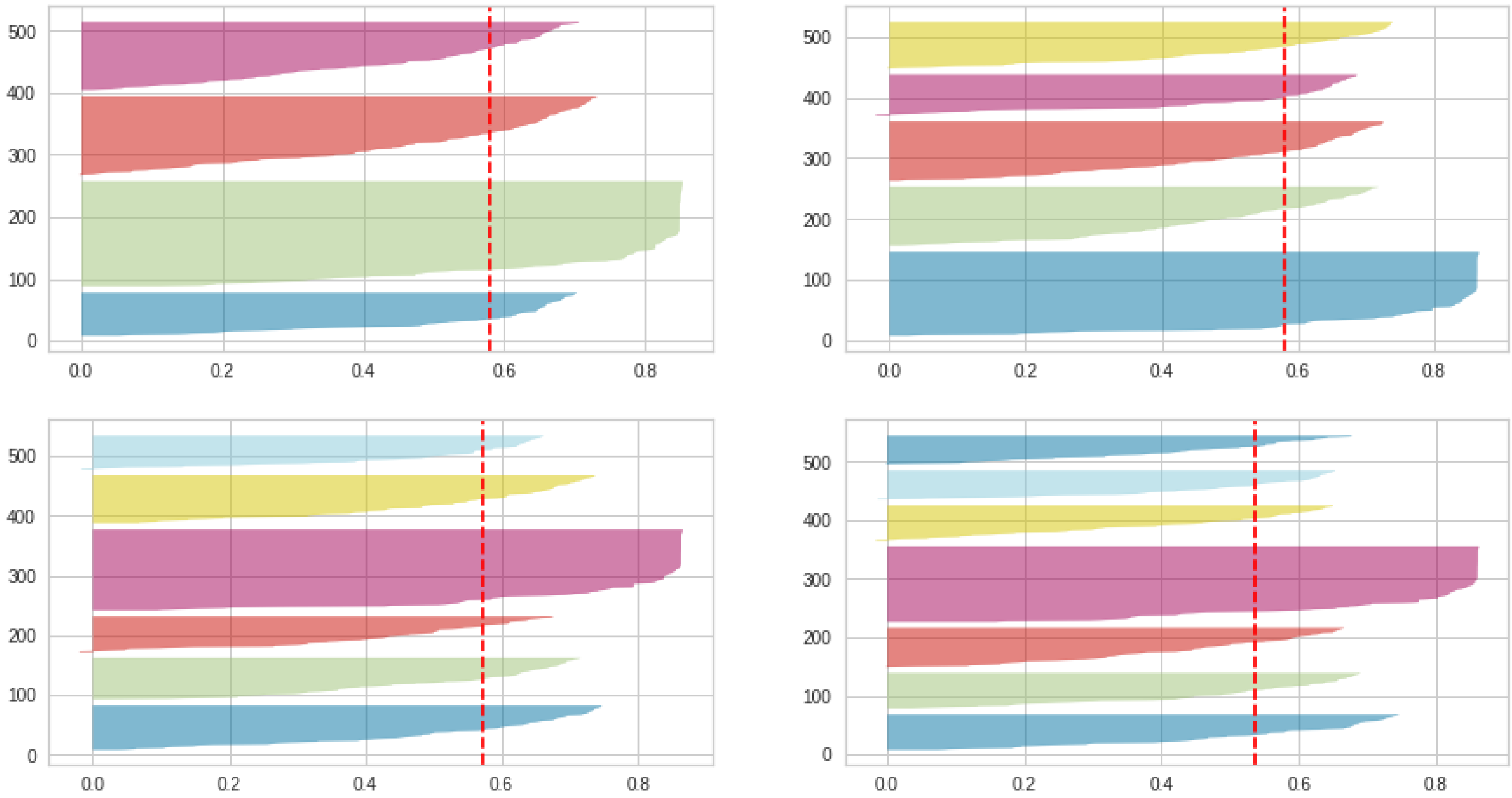
Elbow Curve

- One way to determine the optimum number of groups is using elbow Method
- It compares the in-cluster sum of squares across different centroid configurations



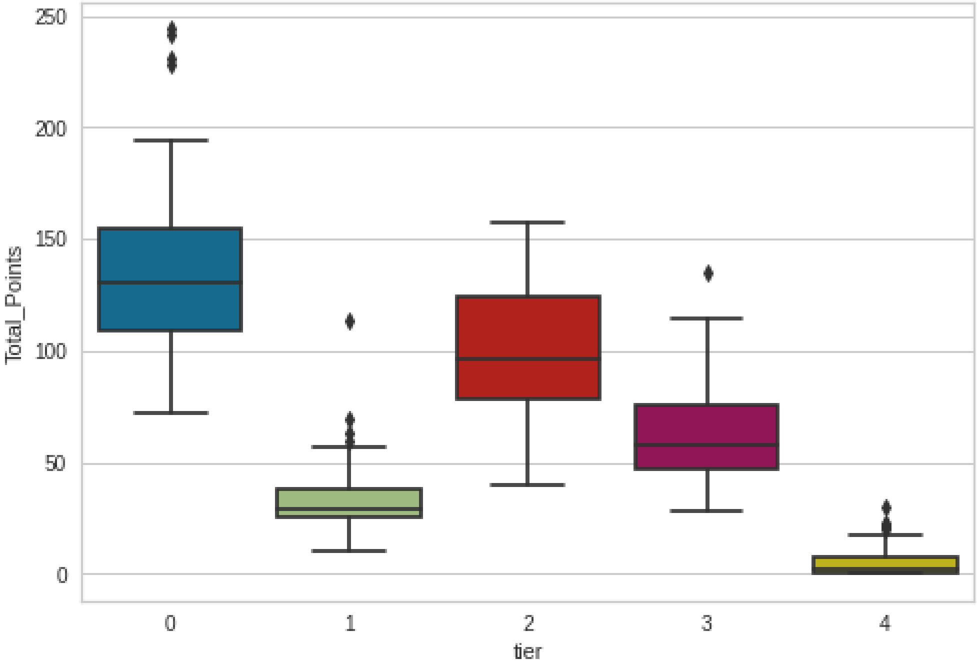
Silhouette Plot

- It is a method to interpret and validate the consistency of the clustering algorithm
- It provides a graphical representation of how well each object has been classified



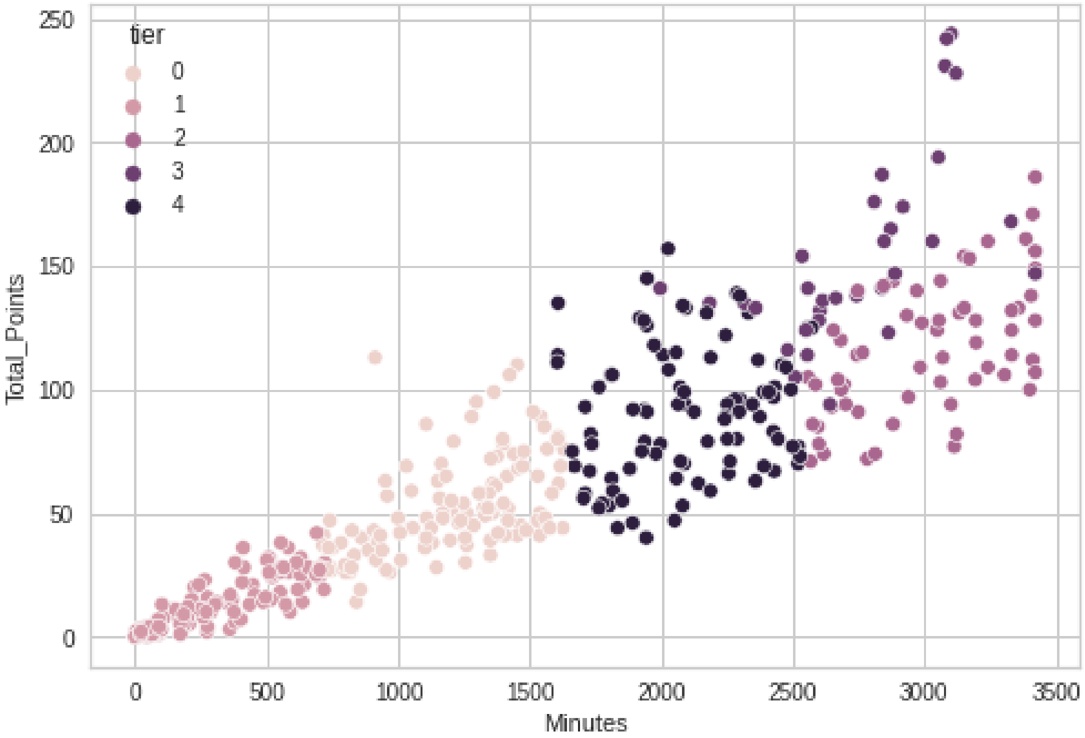
Optimal Number of Clusters

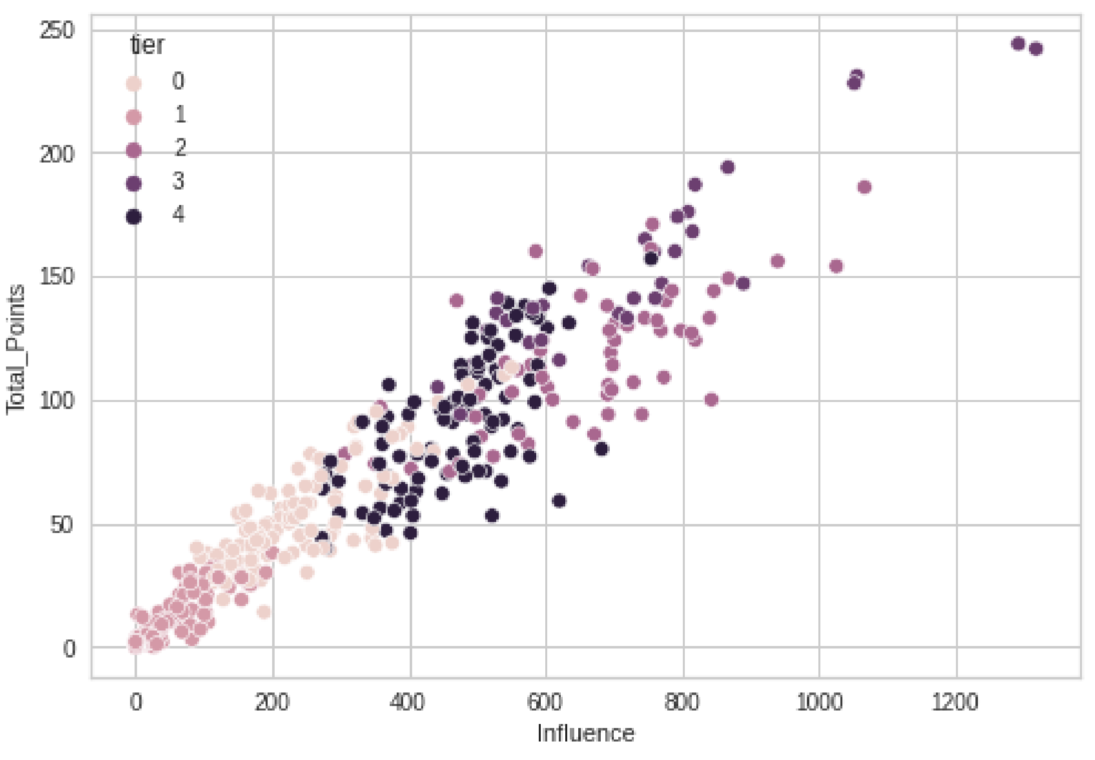
- From the silhouette scores and Elbow curves the optimal number of clusters were chosen
- The maximum silhouette score of 0.58 was obtained for 5 clusters



Fantasy Points scored by tier group

Cluster Profiling





5 Clear cluster can be seen which is consistent with both minutes played by a player and their influence in a match

Hierarchical Clustering

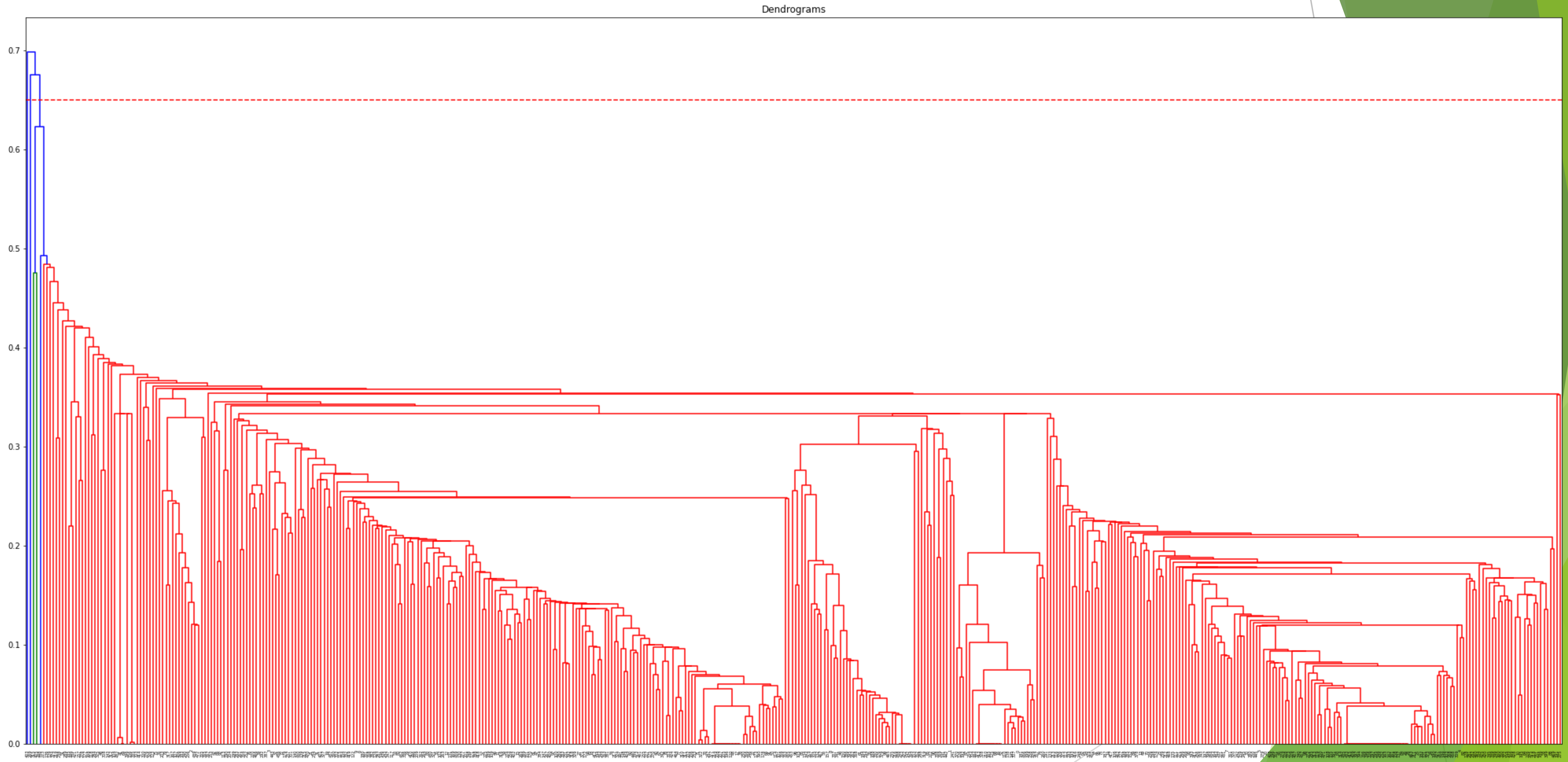
- There are certain challenges with K-means
  - It always tries to make clusters of same size
  - The number of clusters need to be selected at start
  - Ideally, the number of clusters is not known at the start
- This gap is bridged by hierarchical clustering.
- Hierarchical clustering can be explained as:
  - It assigns each data point as a cluster
  - The most similar clusters are combined
  - This process is repeated until we have only one cluster

Linkages

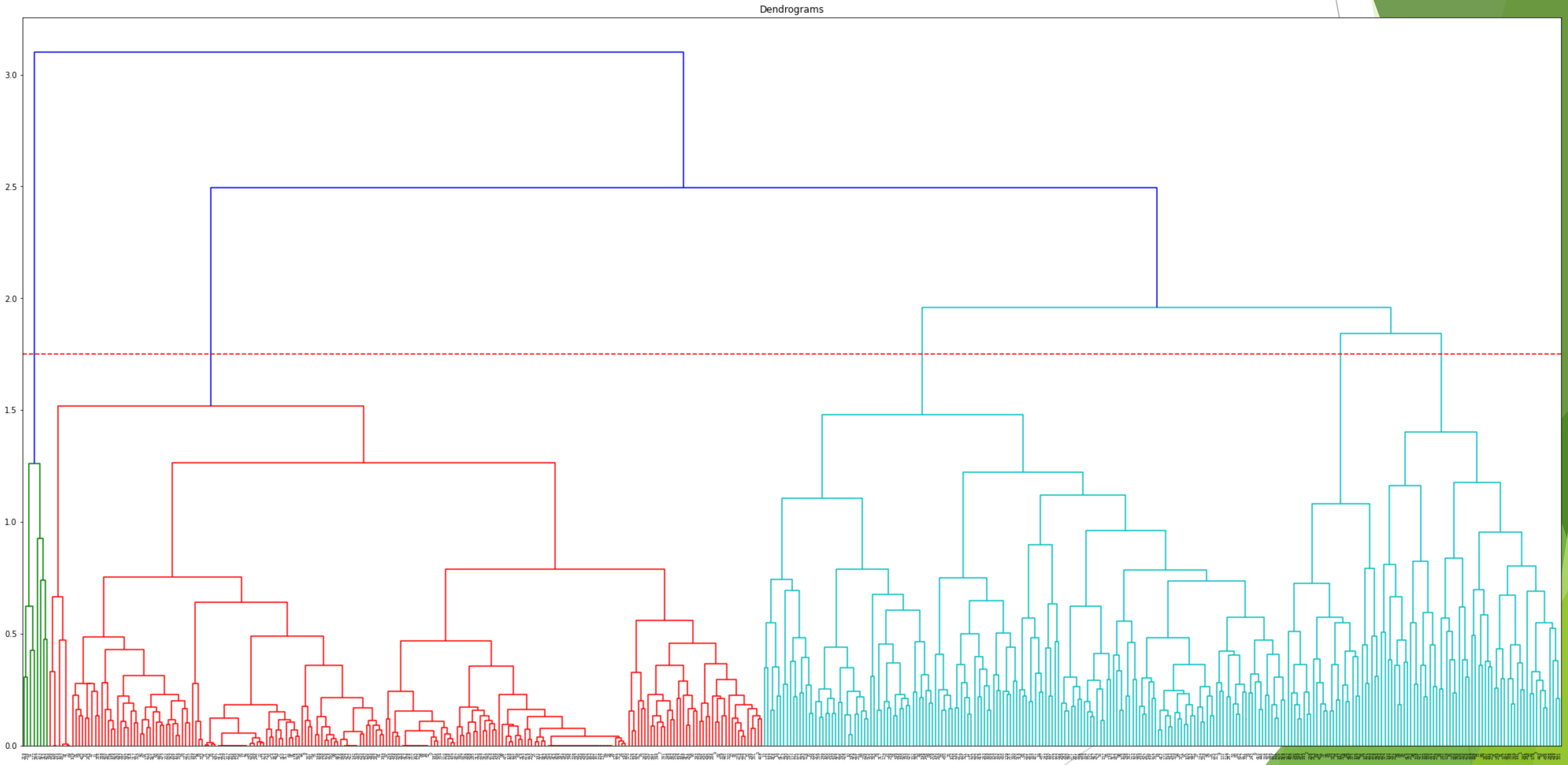
- During hierarchical clustering, two sub-cluster are combined
  - For that distance between them is required
- The different linkage define the different approaches to measure the distance between them

1. Single – It returns the minimum distance between any two points in the clusters
2. Complete- It returns the maximum distance between any two points in the clusters
3. Average- It returns the average distance between all the points in the clusters

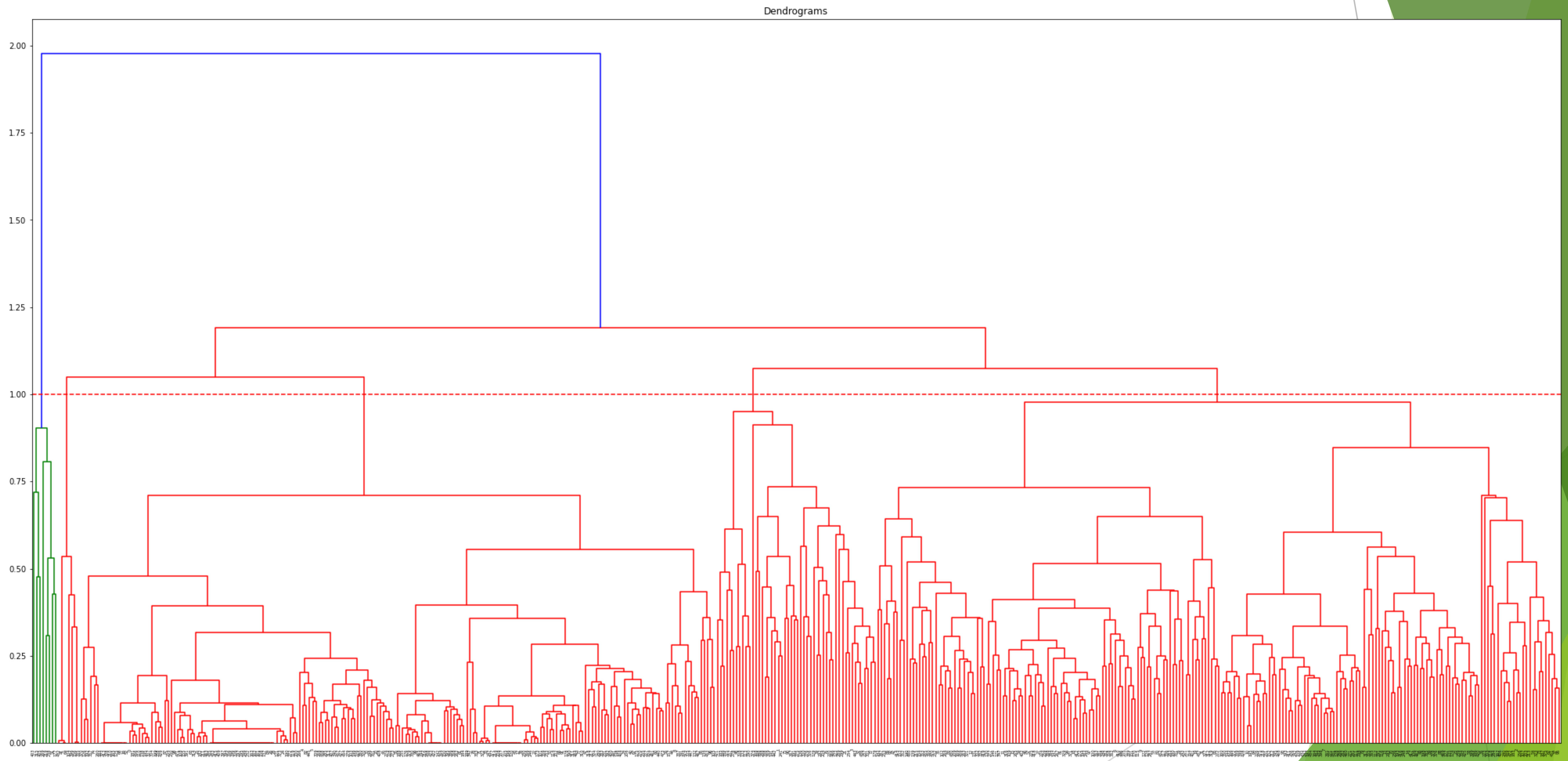
Single Linkage



Complete Linkage

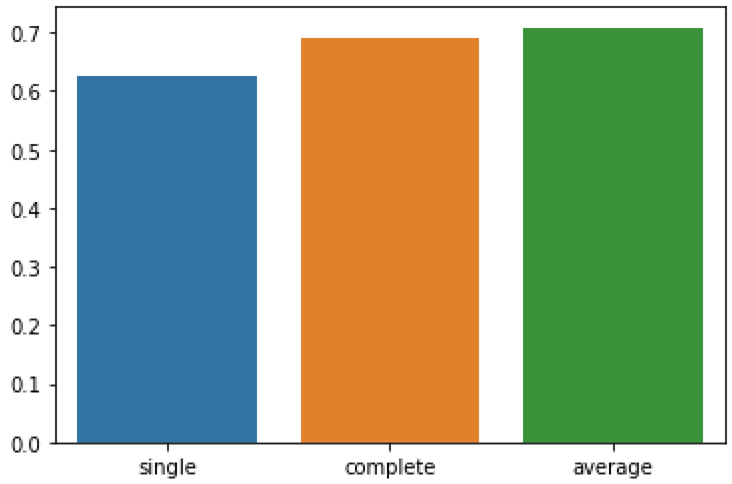


Average Linkage



Cophenetic Correlation

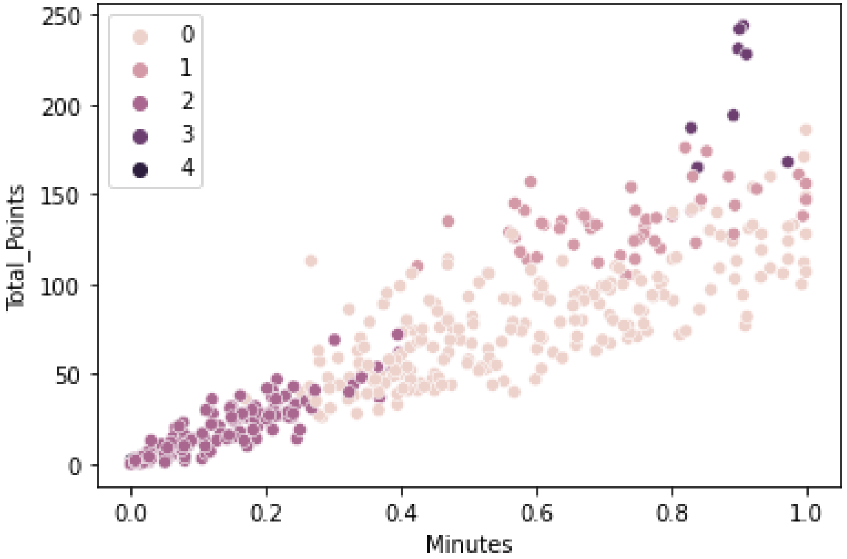
- Cophenet index is a measure of the correlation between the distance of points in feature space and distance on the dendrogram



Cophenet index of different Linkage Methods in hierarchical clustering

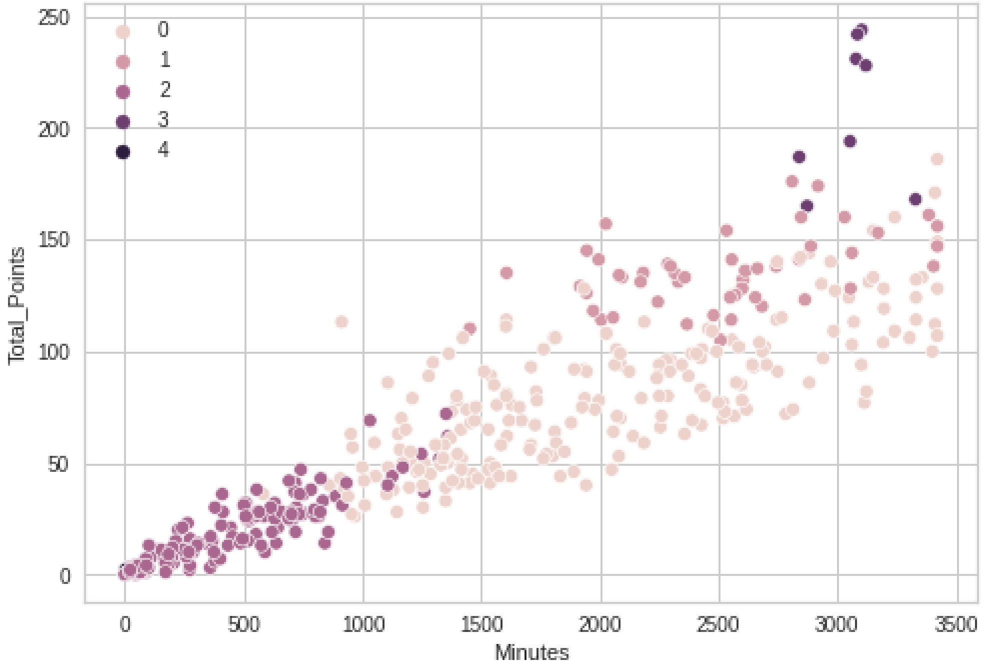
Optimal Hierarchical Clustering

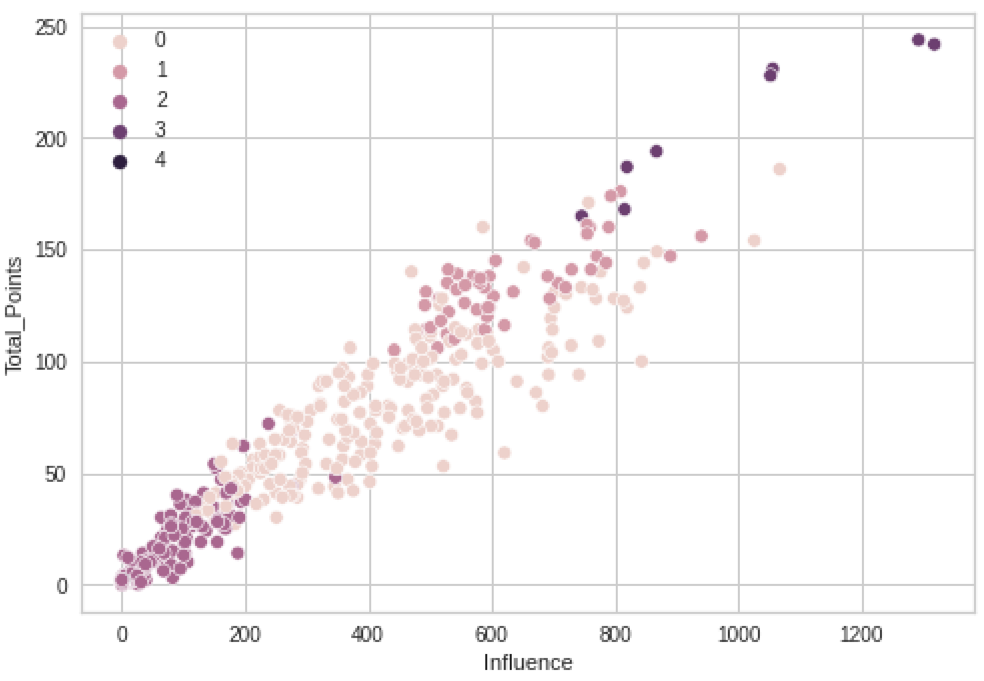
- The highest cophenet index was obtained for average linkage hence we use that.
- Using the dendrogram plot for average linkage and threshold distance as 1.00 we got optimal number of clusters as 5



Scatter plot of total points versus minutes with 5 clusters

Cluster Profiling

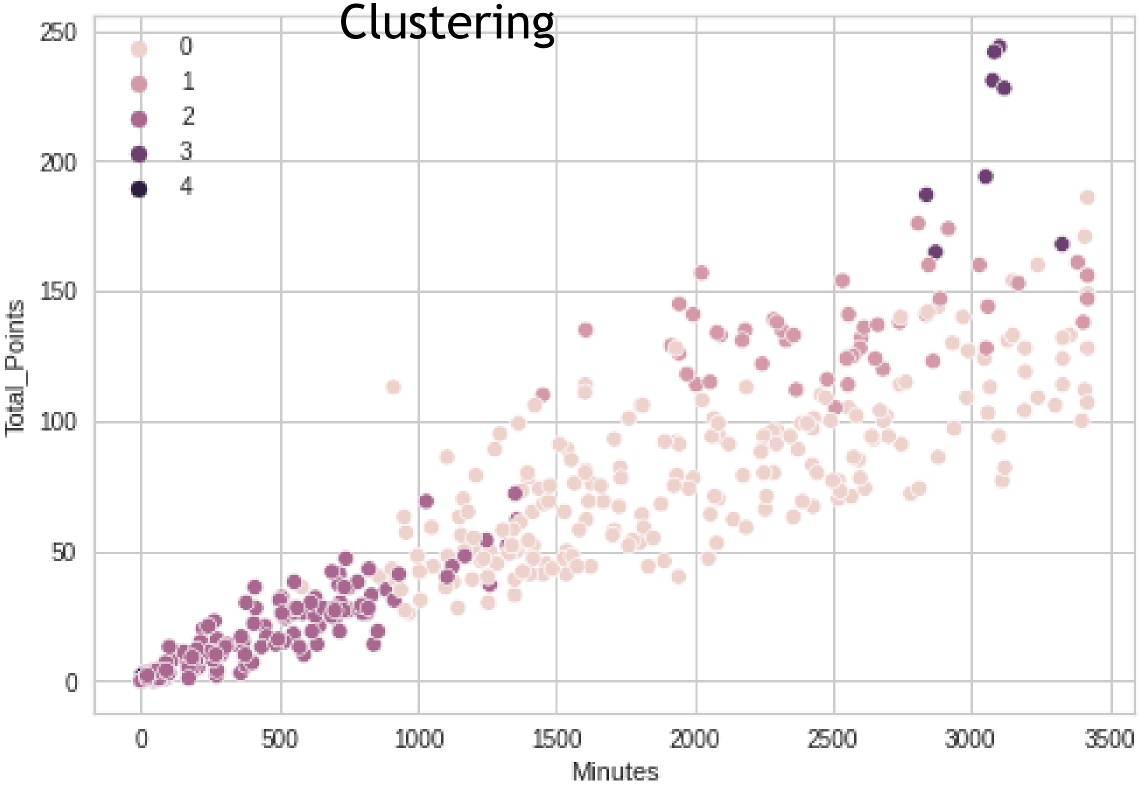




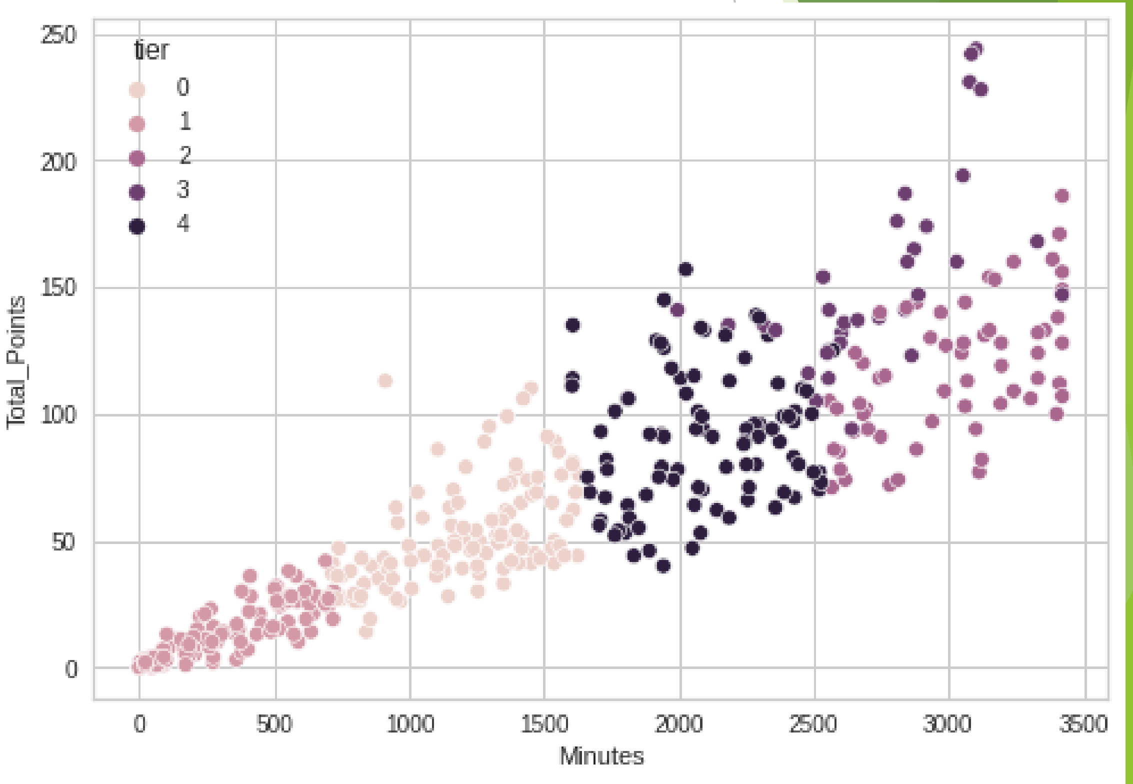
5 Clear cluster can be seen which is consistent with both minutes played by a player and their influence in match

Comparing Clusters

Hierarchical Clustering



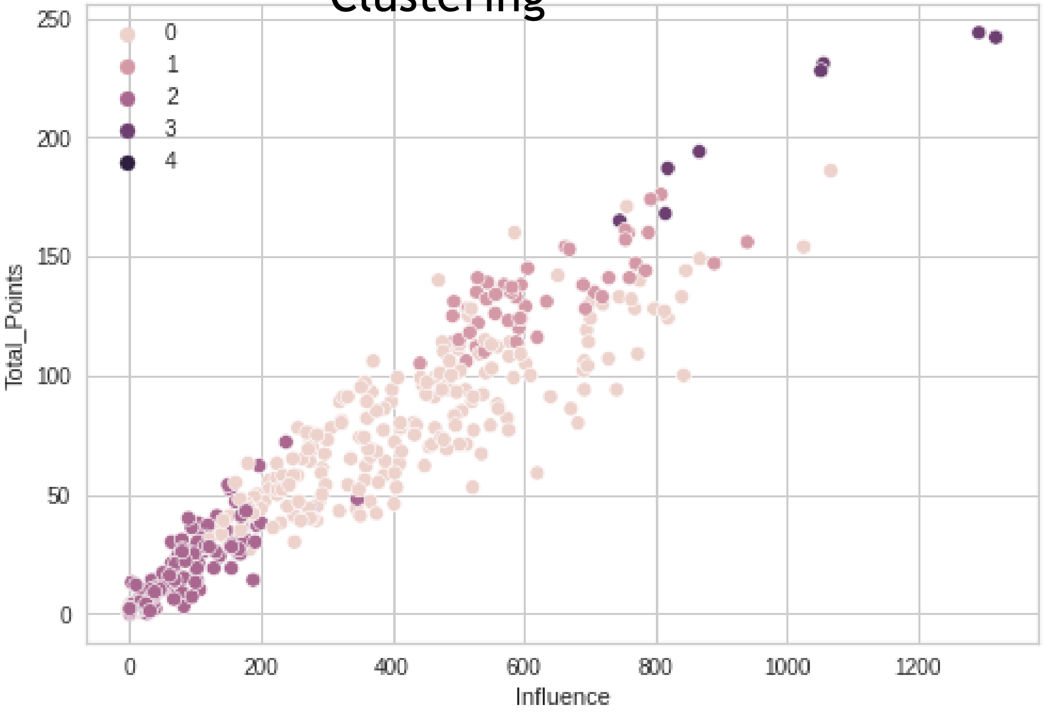
Kmeans Clustering



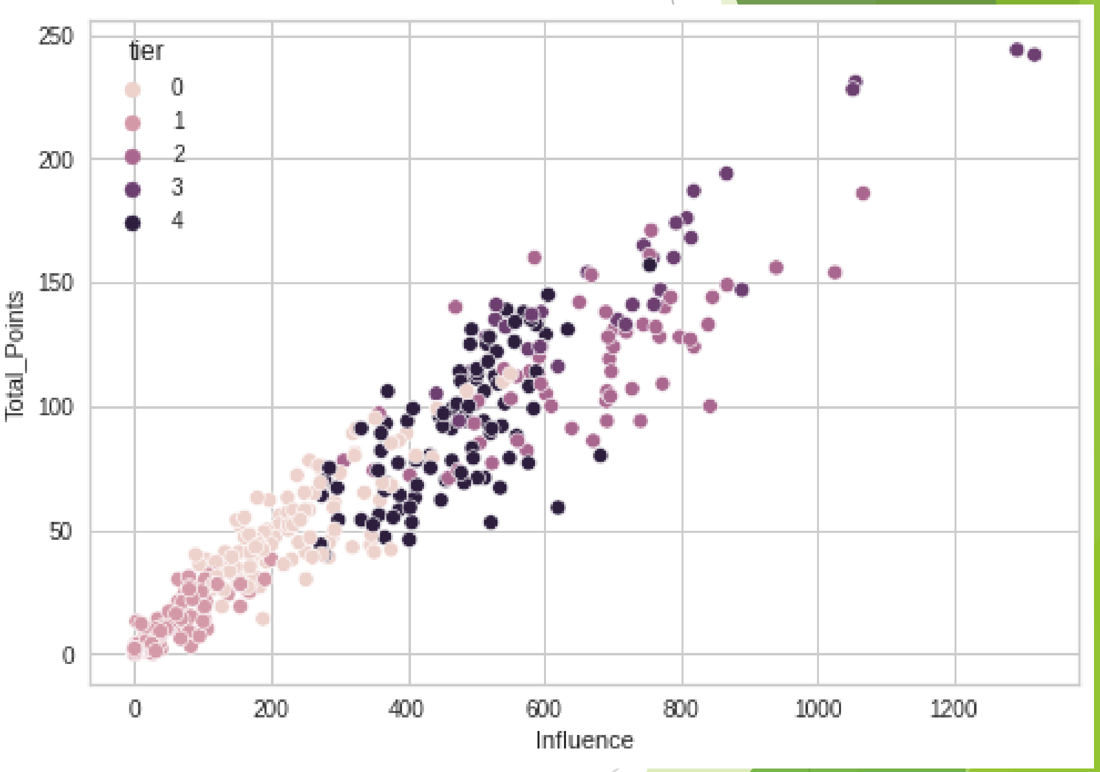
- In hierarchical clustering the total number of points in a cluster in uneven as compared to kmeans clustering

Comparing Clusters

Hierarchical Clustering



Kmeans Clustering



- In hierarchical clustering the clusters can be seen based on the total points scored by a player while in kmeans it depends more on influence of the player
- In hierarchical clustering the clusters are formed with player potentials distributed in a range of 50

Key Takeaways

- In view of scoring more goals the midfielder and forward players should be priced higher
- Manchester City has the highest no of points, hence its players can be priced higher
- Some players in Liverpool have very high influence in a match whom can be priced higher
- Tottenham hotspurs and Liverpool have some players which are really good goal scorers

Key Takeaways

- Hierarchical Clustering provides better way to cluster players based on their total points(potential)
- A particular cluster of players should be priced nearly equal
- Players in different cluster should be priced differently