Machine Learning assignment 4

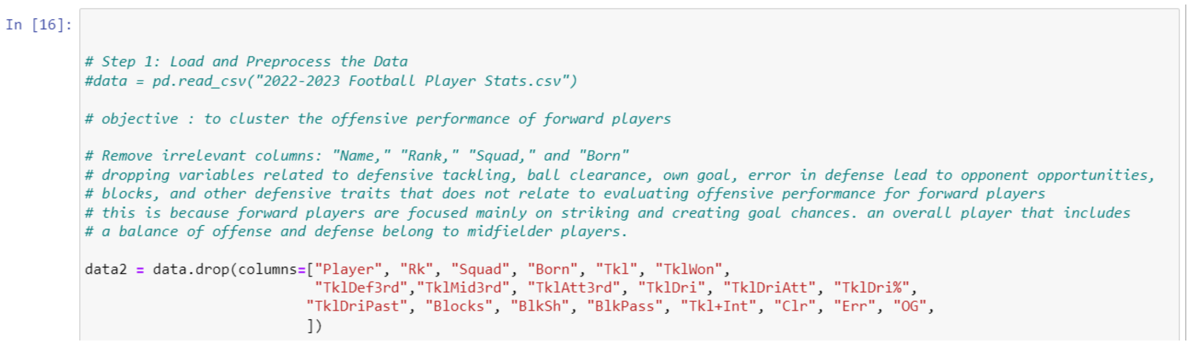
Muhammad Hilman Bin Rozaini (P121535)

Data pre-processing

Firstly, the data consist of 2022-2023 football player stats per 90 minutes. It only includes players from Premier League, Ligue 1, Bundesliga, Serie A and La Liga. There are around 124 variables and around 2500 rows in this dataset.

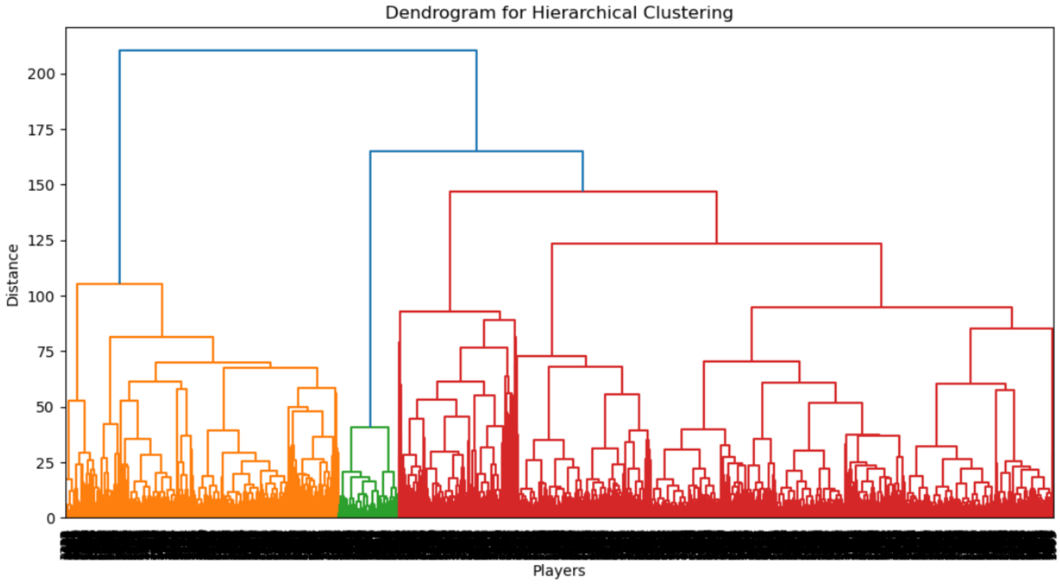
The objective of this project is to perform clustering analysis on the performance of forward players. In analysing the performance of forward players, there are some variables were removed that they are irrelevant to the objective of the analysis. The columns removed are mainly on the defending tackles, defensive ball interceptions, ball clearance, own goal and others that are not directly attributable to performance of offensive playing in a football.

The details of the columns dropped are as the below image:



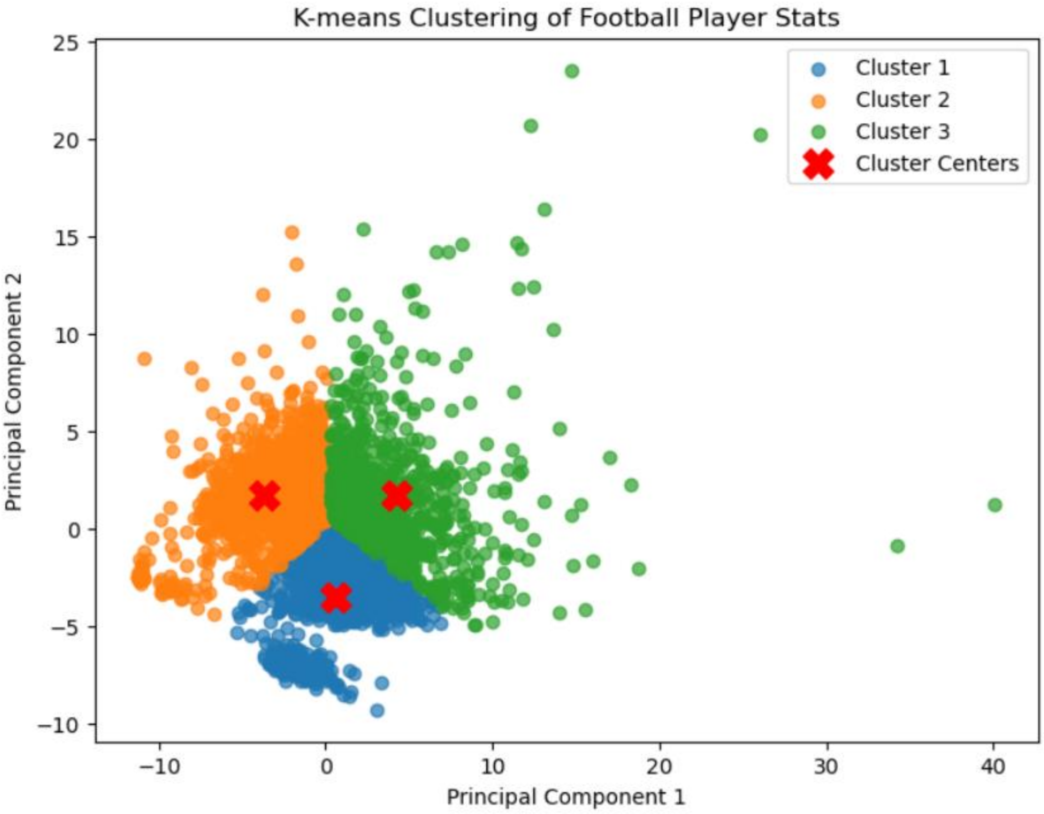
```
In [16]:

# Step 1: Load and Preprocess the Data
#data = pd.read_csv("2022-2023 Football Player Stats.csv")

# objective : to cluster the offensive performance of forward players

# Remove irrelevant columns: "Name," "Rank," "Squad," and "Born"
# dropping variables related to defensive tackling, ball clearance, own goal, error in defense lead to opponent opportunities,
# blocks, and other defensive traits that does not relate to evaluating offensive performance for forward players
# this is because forward players are focused mainly on striking and creating goal chances. an overall player that includes
# a balance of offense and defense belong to midfielder players.

data2 = data.drop(columns=["Player", "Rk", "Squad", "Born", "Tkl", "TklWon",
                            "TklDef3rd","TklMid3rd", "TklAtt3rd", "TklDri", "TklDriAtt", "TklDri%",
                           "TklDriPast", "Blocks", "BlkSh", "BlkPass", "Tkl+Int", "Clr", "Err", "OG",
                           ])
```

Clustering output



According to the dendrogram, the hierarchical clustering suggests that there are three cluster groups in the dataset for forward players.

From the identified number of cluster group through dendrogram, principal component analysis is applied to reduce the dimensionality of the data from its original high-dimensional space to a lower-dimensional space, where in this case it is 2-dimension.



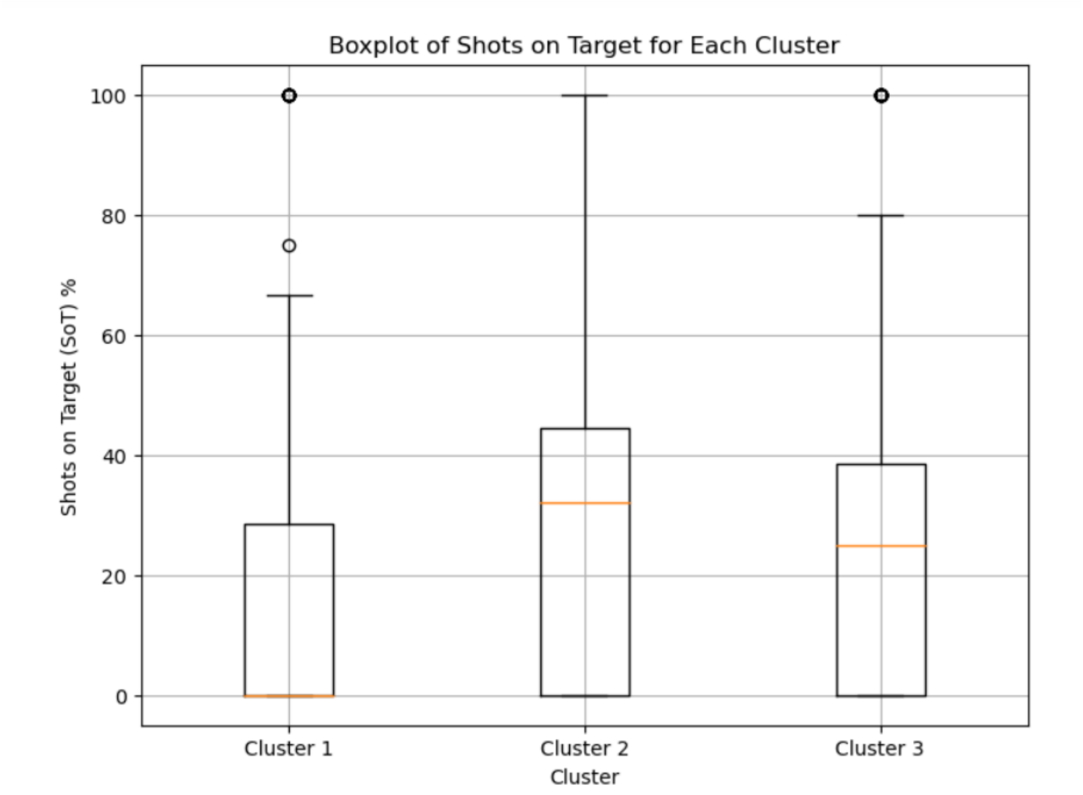
Then, in reference to the above dendrogram, K-means clustering is further applied at k= 3 as shown in the above plot.

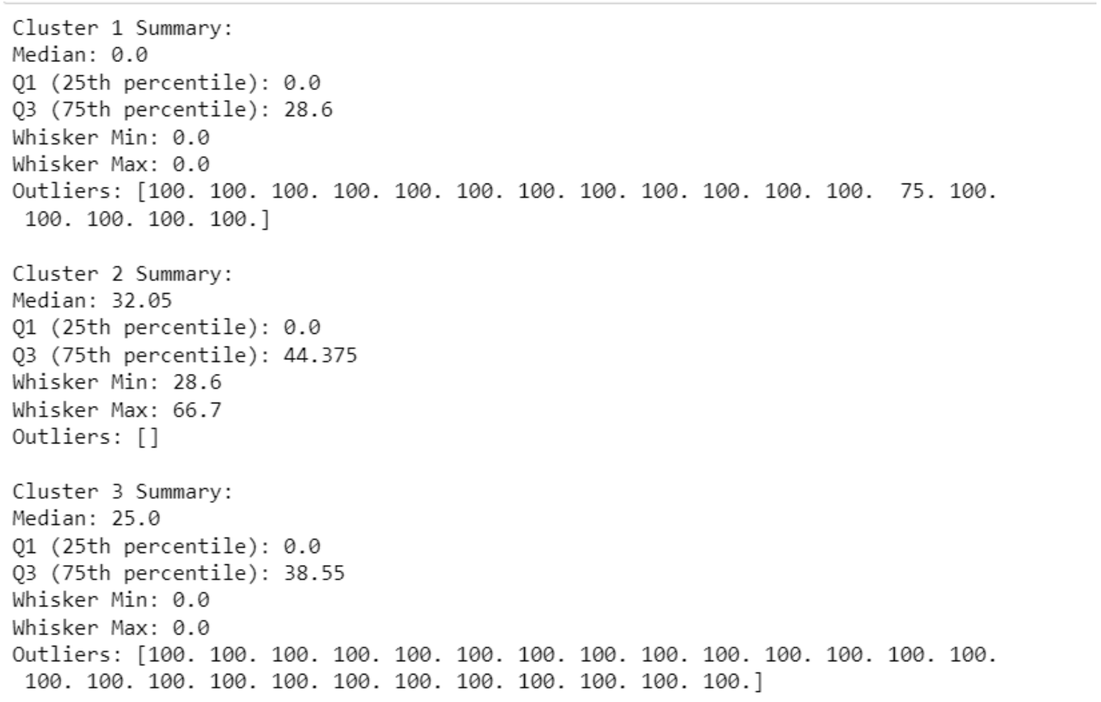
From the k-means clustering plot, the data points are grouped into three clusters based on their similarities in the lower-dimensional PCA space. The cluster separates the players into 3 distinctive groups, where each groups is represented by a different colour. The “X” mark is the cluster centroids, where it is a determining point to indicate the distance between the data points and the cluster center and how well each data points fit into their respective clusters.

Then, players with the similar attributes are clustered within the same cluster, where the attributes includes the variables considered during the data pre-processing phase.

In contrast, players in different clusters have different attributes, suggest that the clustering algorithm has identified different player profiles or playing styles.

Further analysis (Boxplot)





```
Cluster 1 Summary:
Median: 0.0
Q1 (25th percentile): 0.0
Q3 (75th percentile): 28.6
Whisker Min: 0.0
Whisker Max: 0.0
Outliers: [100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100.  75. 100.
 100. 100. 100. 100.]

Cluster 2 Summary:
Median: 32.05
Q1 (25th percentile): 0.0
Q3 (75th percentile): 44.375
Whisker Min: 28.6
Whisker Max: 66.7
Outliers: []

Cluster 3 Summary:
Median: 25.0
Q1 (25th percentile): 0.0
Q3 (75th percentile): 38.55
Whisker Min: 0.0
Whisker Max: 0.0
Outliers: [100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100.
 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100. 100.]
```

To further analyse the characteristics and differences between each clusters, we select variable "Shot on Target Percentage (SoT%)" to further analyse the attacking performance of forward players in each clusters.

Boxplot is used in the further analysis as it is a graphical representation of the variable distribution.

For cluster 1, we can observe that the median and the lower boundary is the same at value of 0%, while the upper boundary players in this cluster only obtain 28.6% accuracy on their shot on target. It means that at least 50% of the players in cluster 1 has a shot of target percentage at 0%. This shows that their shooting accuracy is at least 50% not on target to the goal post.

Moving to cluster 2, the median of the players having shot on target are at 32.05%, while the upper boundary of the players in this cluster 2 have a shot of target of around 44.38%. This indicates that the cluster 2 players are performing better at shot on target as compared to cluster 1, as the percentage shot on target is much better at median and upper boundary of this cluster.

Lastly moving to cluster 3, we can see that the median shot on target is around 25%, while the upper boundary is only 38.55% accuracy.

In overall, we can infer that the shot on target for forward players in cluster 1 has the worst performance, while players in cluster 2 has the best performance in their shot on target accuracy. Players performance in cluster 3 is in between cluster 1 and cluster 2. However, all three clusters have the same lower boundary at 0% shot on target accuracy, and this suggest that there are some forward players in cluster 2 and 3 performs as bad as players in cluster 1.