FIFA 23

ML Project

Supervised By: Dr. Doaa Mahmoud

Supervised By: Eng. Mohamed Abdullah

Presented By:

Mira Ehab

LinkedIn

Github

Kaggle

CONTENTS

01 Problem Definition and Introduction to the Data

02 Objectives

03 Data Analysis

04 Data Preprocessing

05 Modeling

- A. Predict the Position of the Player
- B. Group the Similar Players

01

Problem Definition and Introduction to the Data

Problem Definition and Introduction to the Data

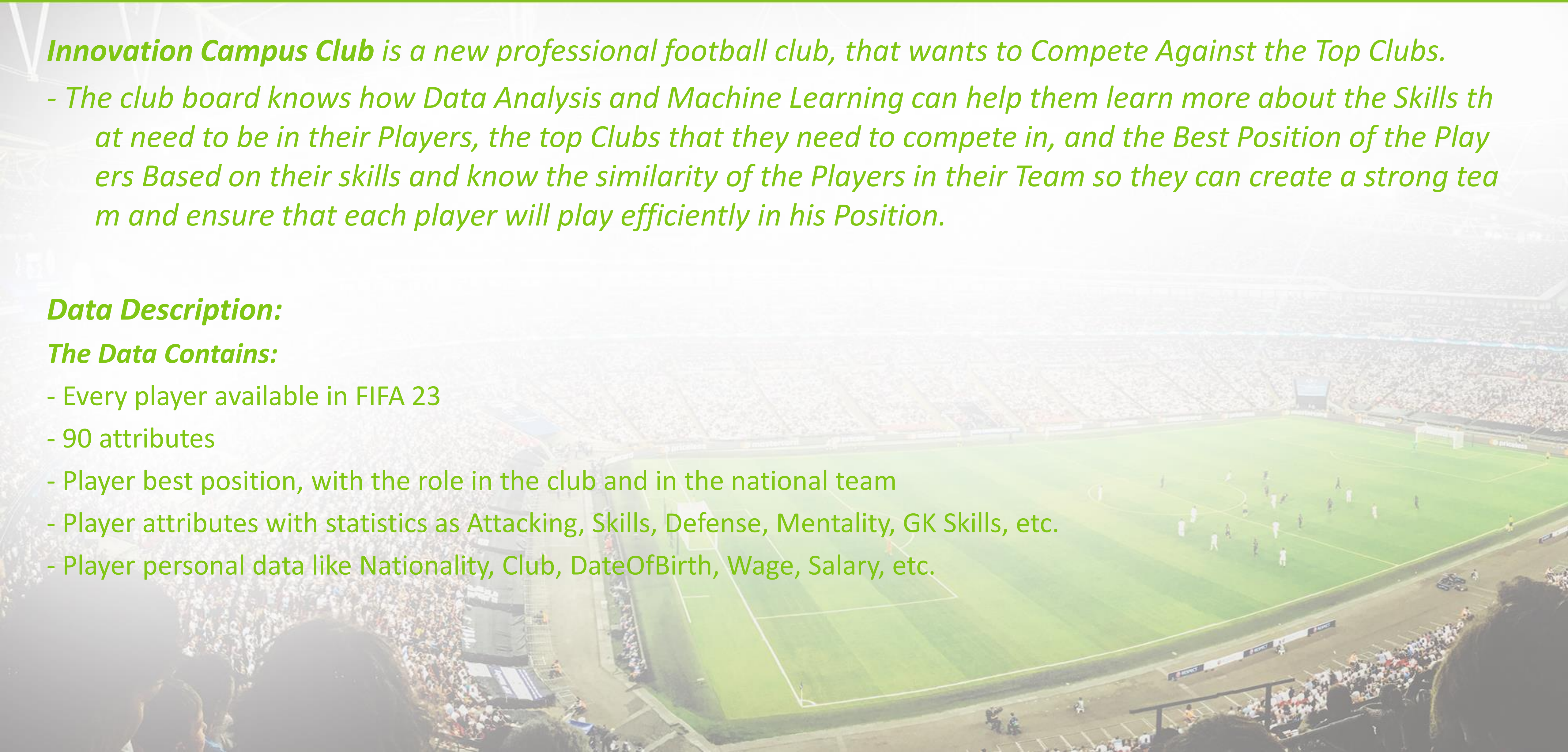
***Innovation Campus Club** is a new professional football club, that wants to Compete Against the Top Clubs.*

- *The club board knows how Data Analysis and Machine Learning can help them learn more about the Skills that need to be in their Players, the top Clubs that they need to compete in, and the Best Position of the Players Based on their skills and know the similarity of the Players in their Team so they can create a strong team and ensure that each player will play efficiently in his Position.*

Data Description:

The Data Contains:

- Every player available in FIFA 23
- 90 attributes
- Player best position, with the role in the club and in the national team
- Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc.
- Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, etc.



02
Objectives

Objectives

Help the ***Innovation Campus Club*** *by:*

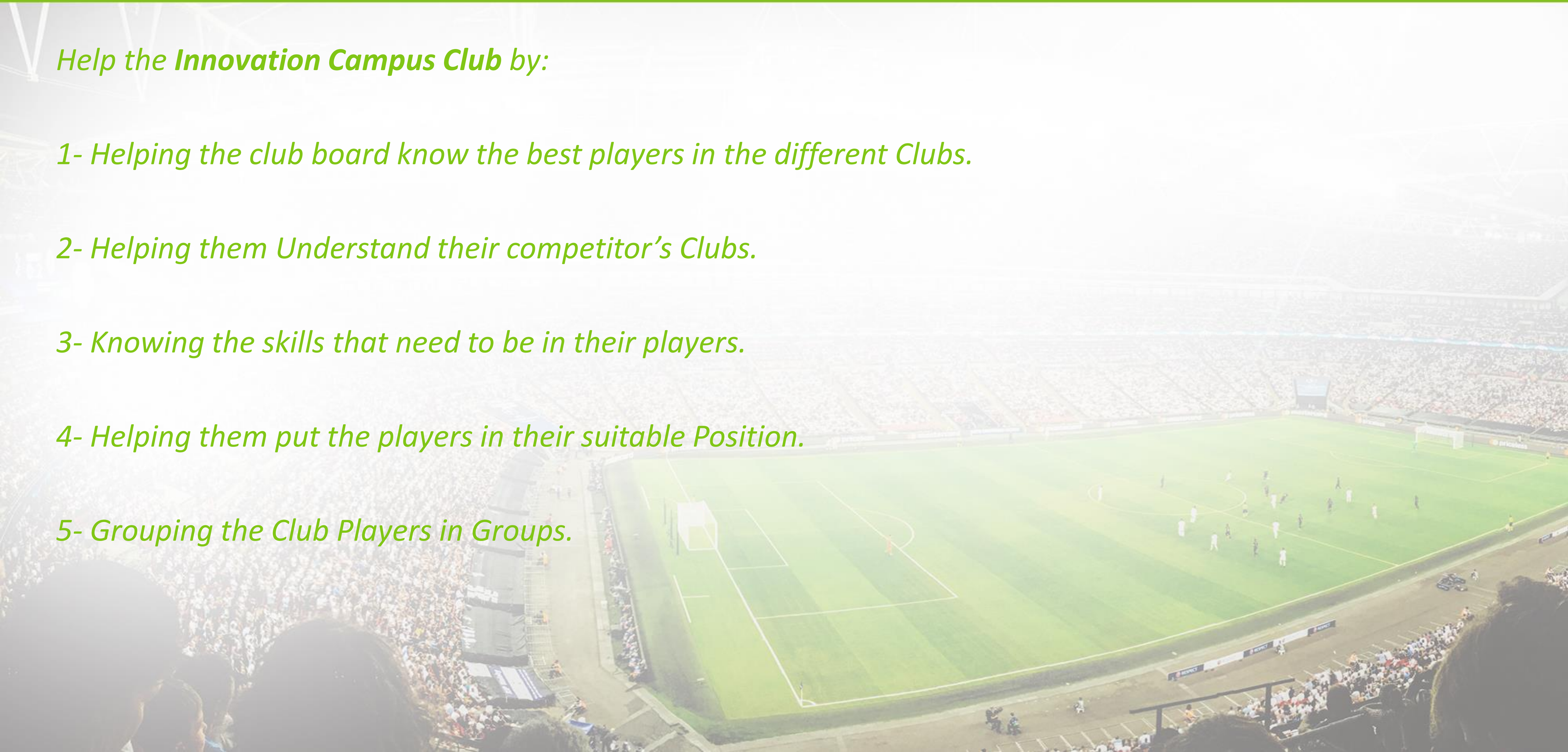
1- Helping the club board know the best players in the different Clubs.

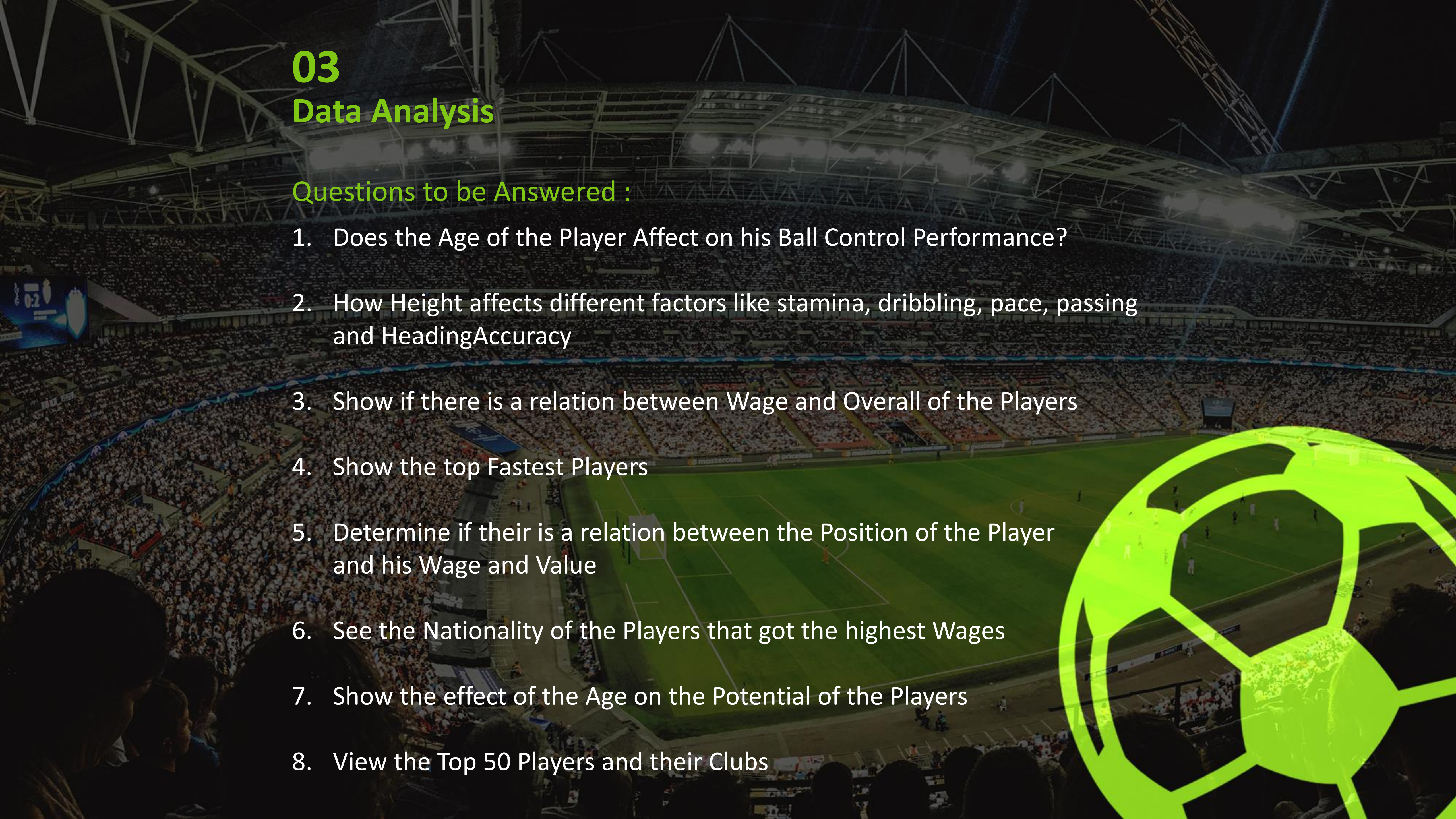
2- Helping them Understand their competitor’s Clubs.

3- Knowing the skills that need to be in their players.

4- Helping them put the players in their suitable Position.

5- Grouping the Club Players in Groups.



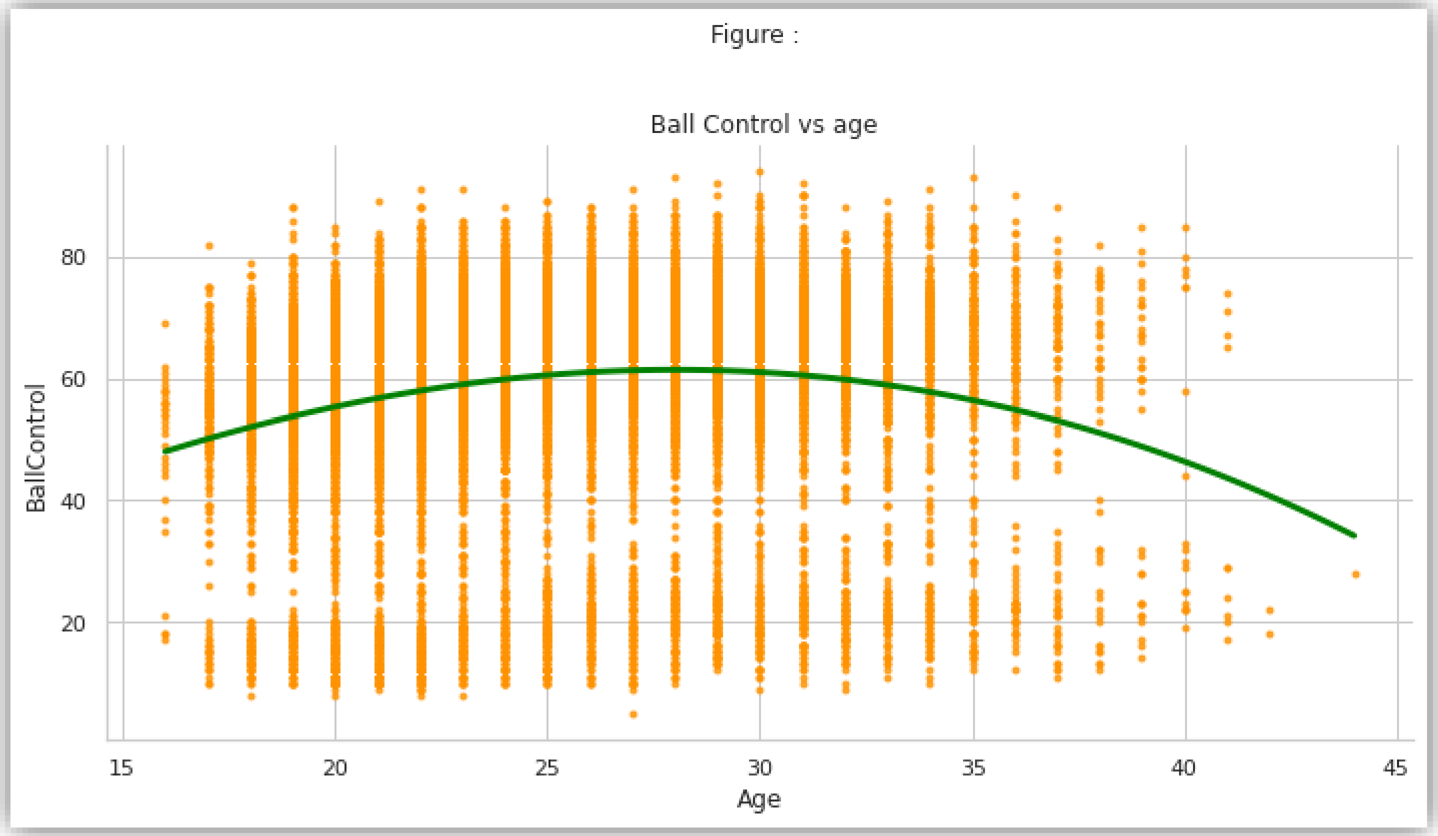


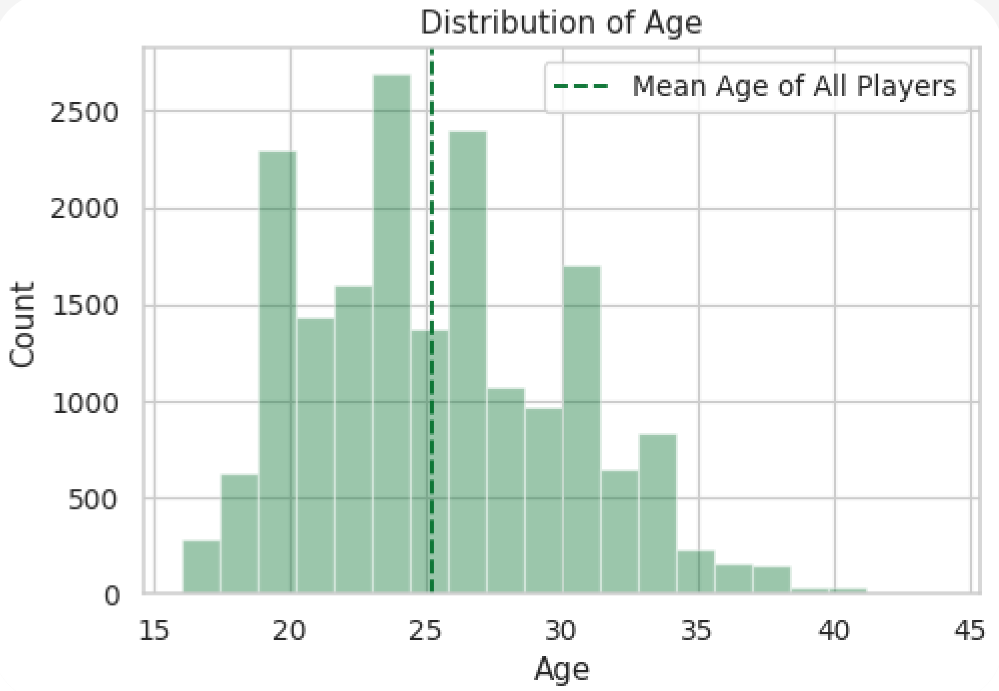
03
Data Analysis

Questions to be Answered :

1. Does the Age of the Player Affect on his Ball Control Performance?
2. How Height affects different factors like stamina, dribbling, pace, passing and HeadingAccuracy
3. Show if there is a relation between Wage and Overall of the Players
4. Show the top Fastest Players
5. Determine if their is a relation between the Position of the Player and his Wage and Value
6. See the Nationality of the Players that got the highest Wages
7. Show the effect of the Age on the Potential of the Players
8. View the Top 50 Players and their Clubs

Does the Age of the Player Affect on his Ball Control Performance?

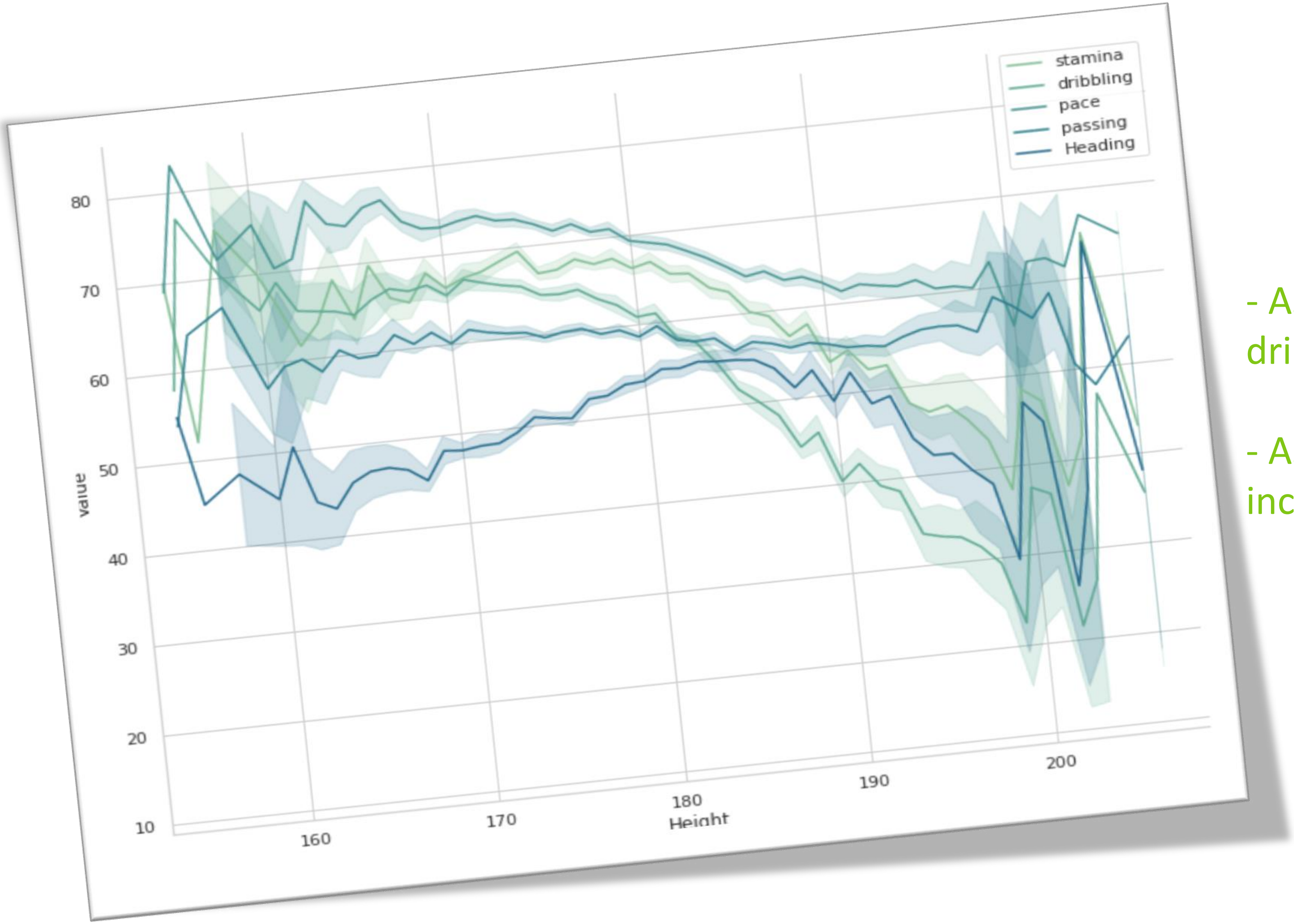




- So We can deduce that the age has an effect on the Player's Ball Control

- While the Age is increasing, the Ball Control decreases.

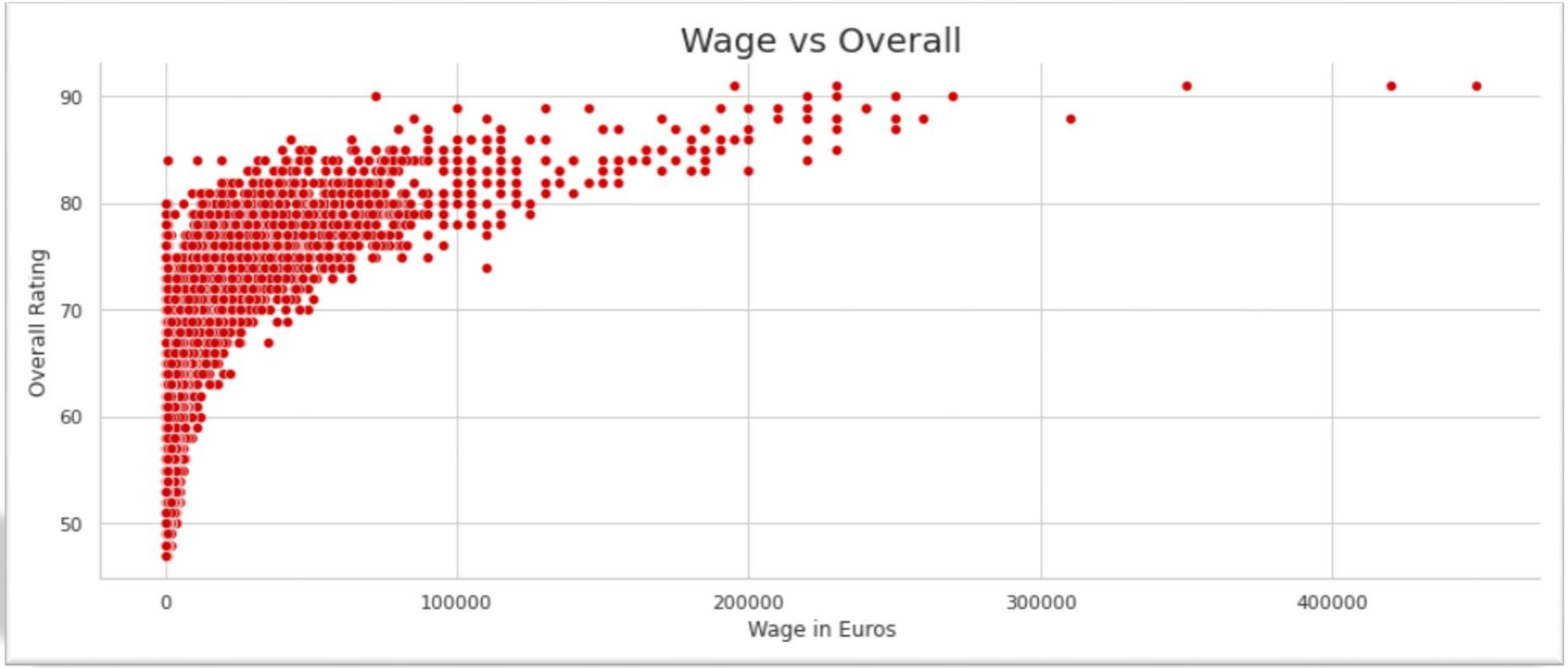
How Height affects different factors like stamina, dribbling, pace, passing and HeadingAccuracy ?



- As height increases, features like stamina, dribbling, pace, passing decreases.

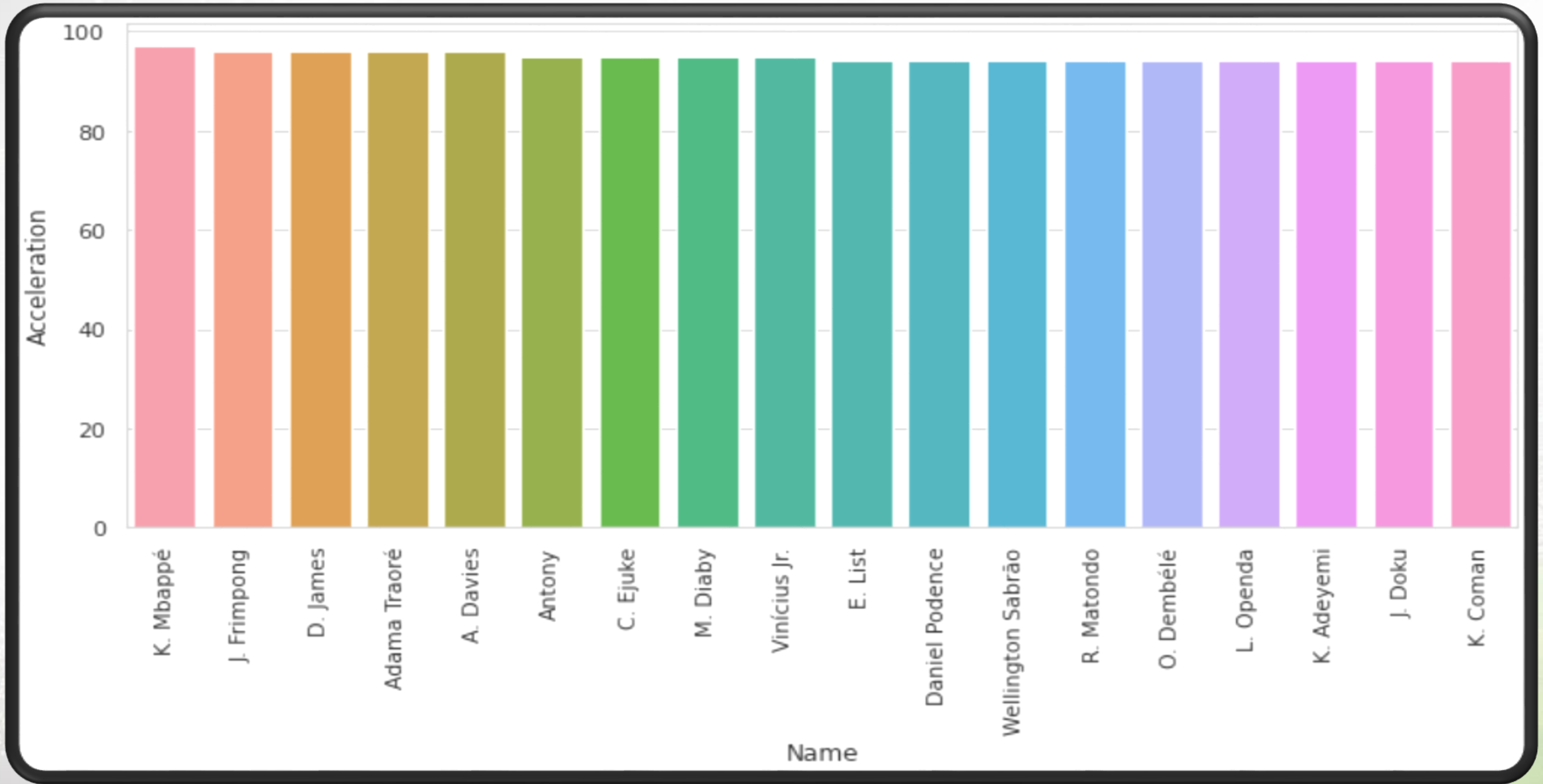
- As height increases, features like Heading increase.

Show if there is a relation between Wage and Overall of the Players

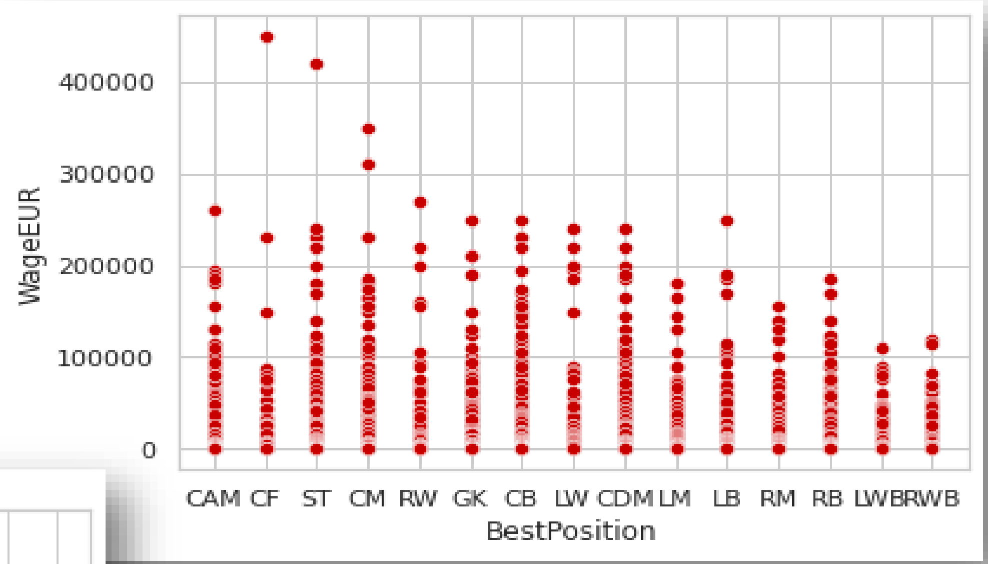


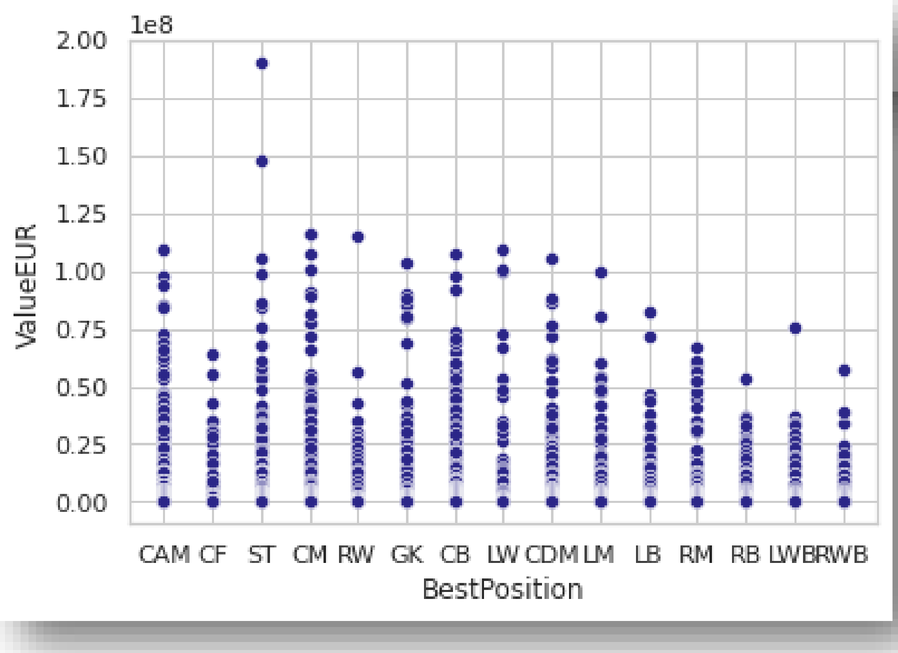
As the Overall Rating Increase, the Wage of the Player Increases too.

Show the top Fastest Players



Determine if their is a relation between the Position of the Player and his Wage and Value

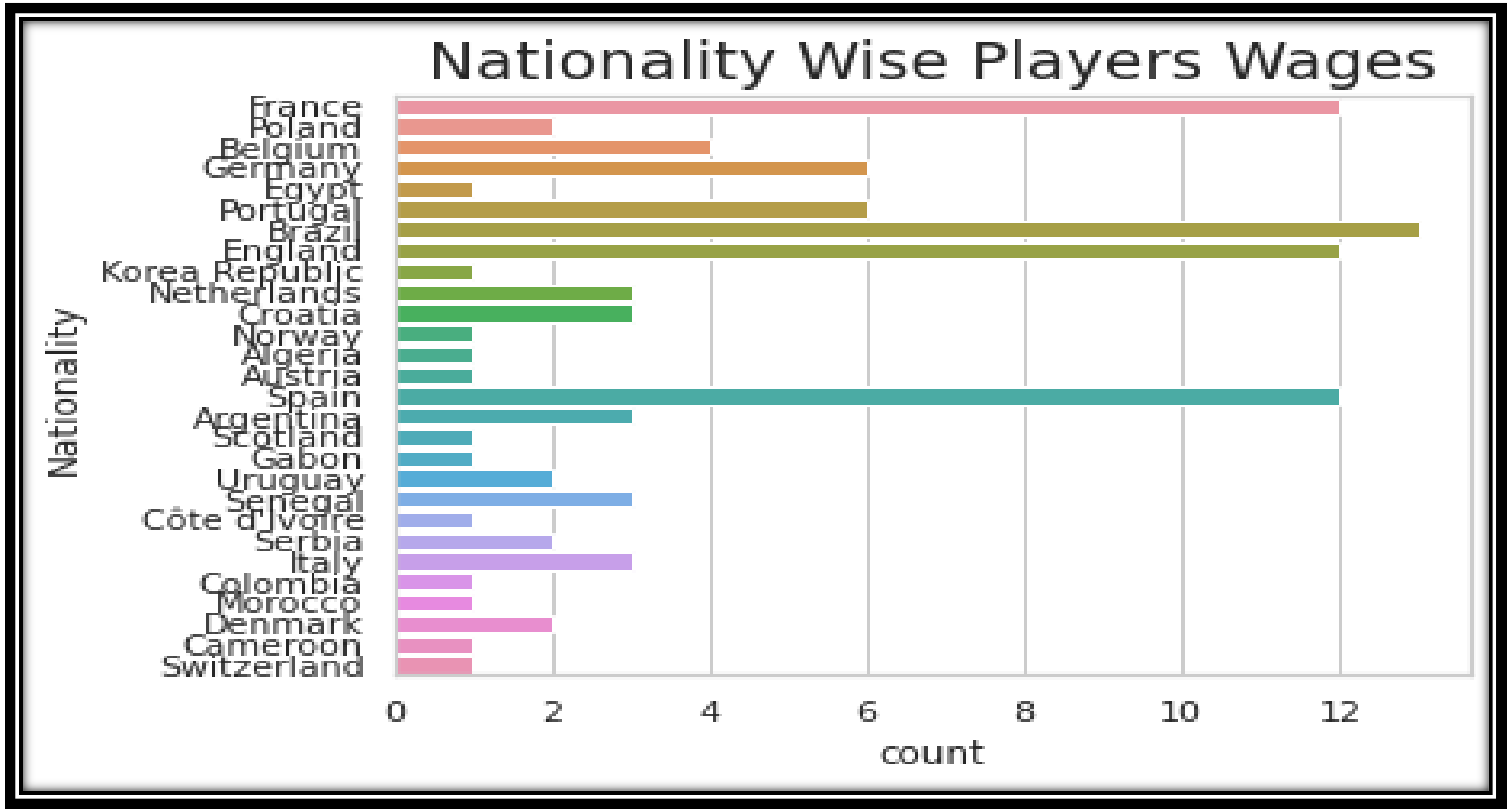




- So we can see that the Players in Positions LM, RM, RB, LWB, RWB got the lowest Wages.

- And the Players With Positions LB, RB, LWB, RWB, CF , RW have the lowest Values.

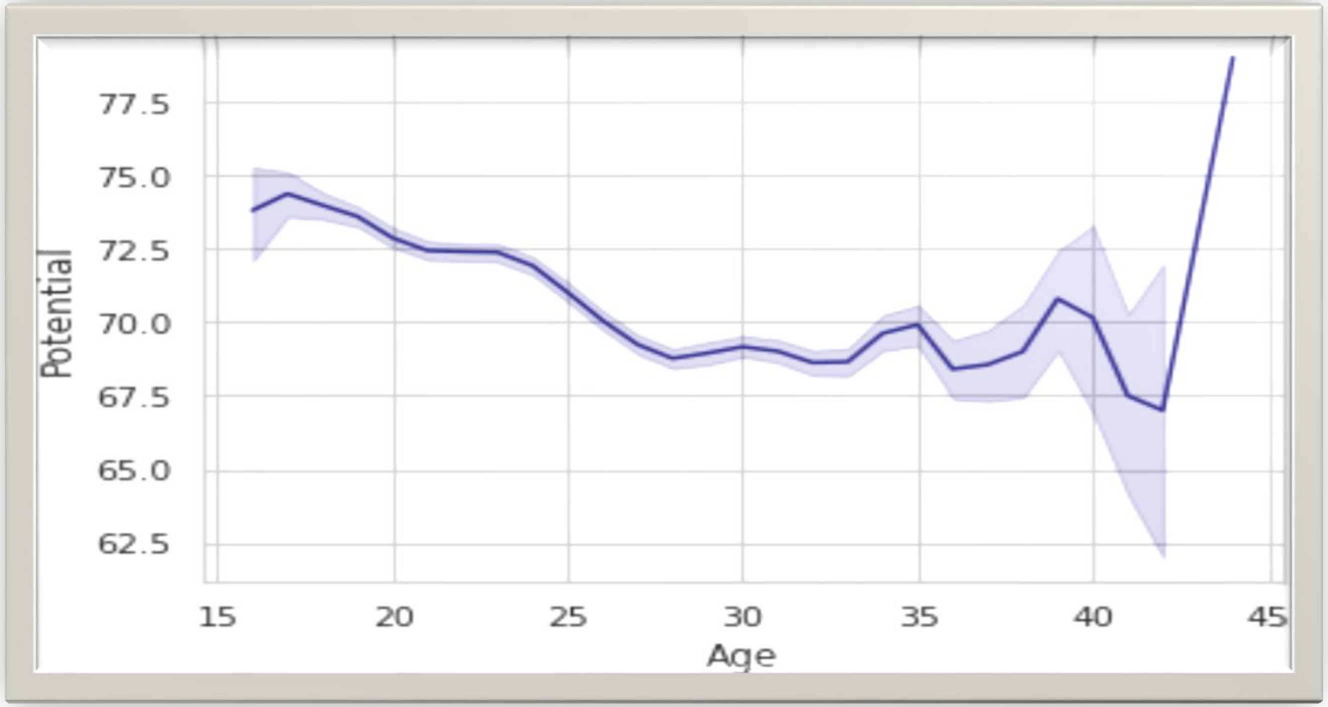
See the Nationality of the Players that got the highest Wages



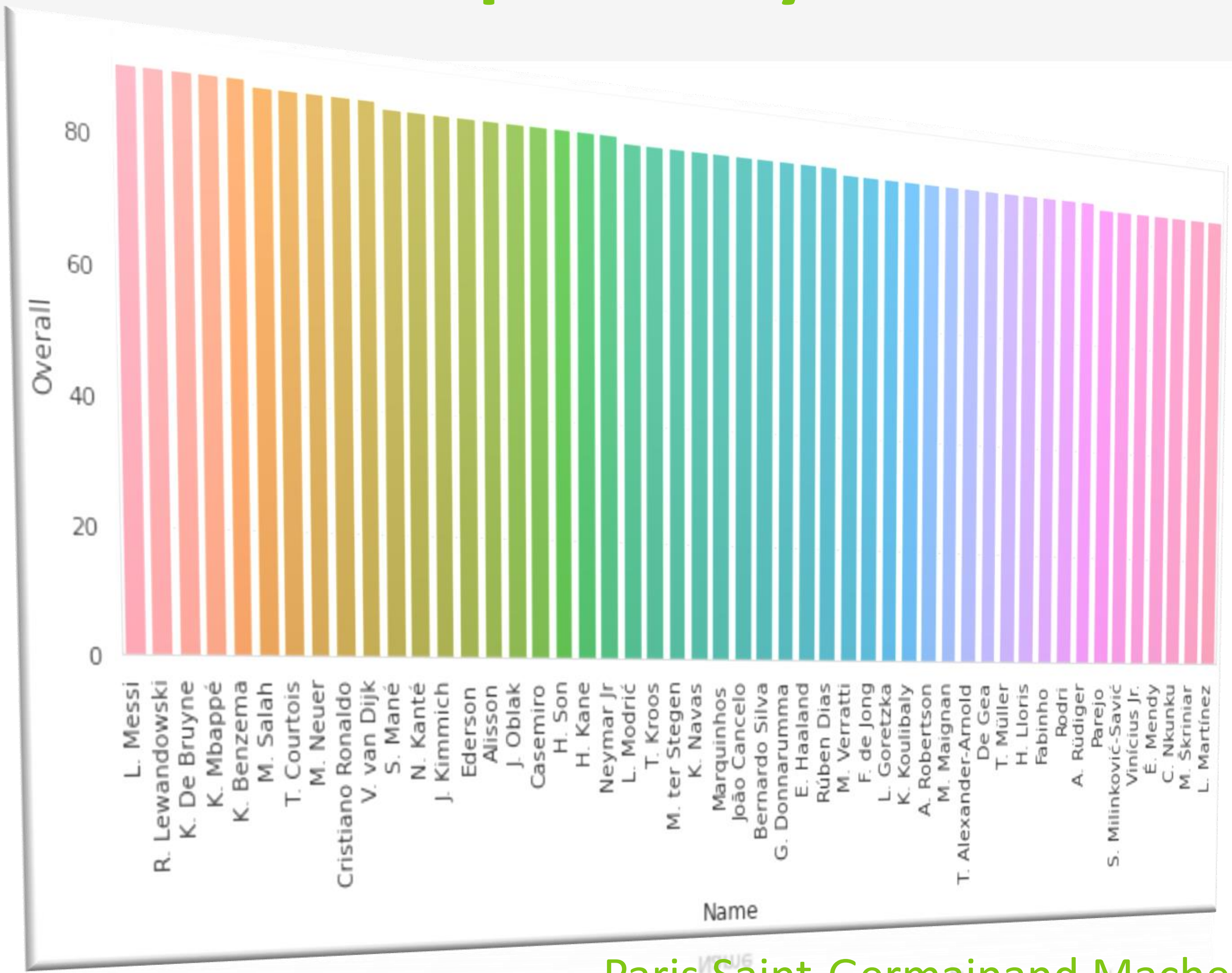
So we can deduce that the Players that got the Maximum Wage are from Brazil , France, England and Spain.

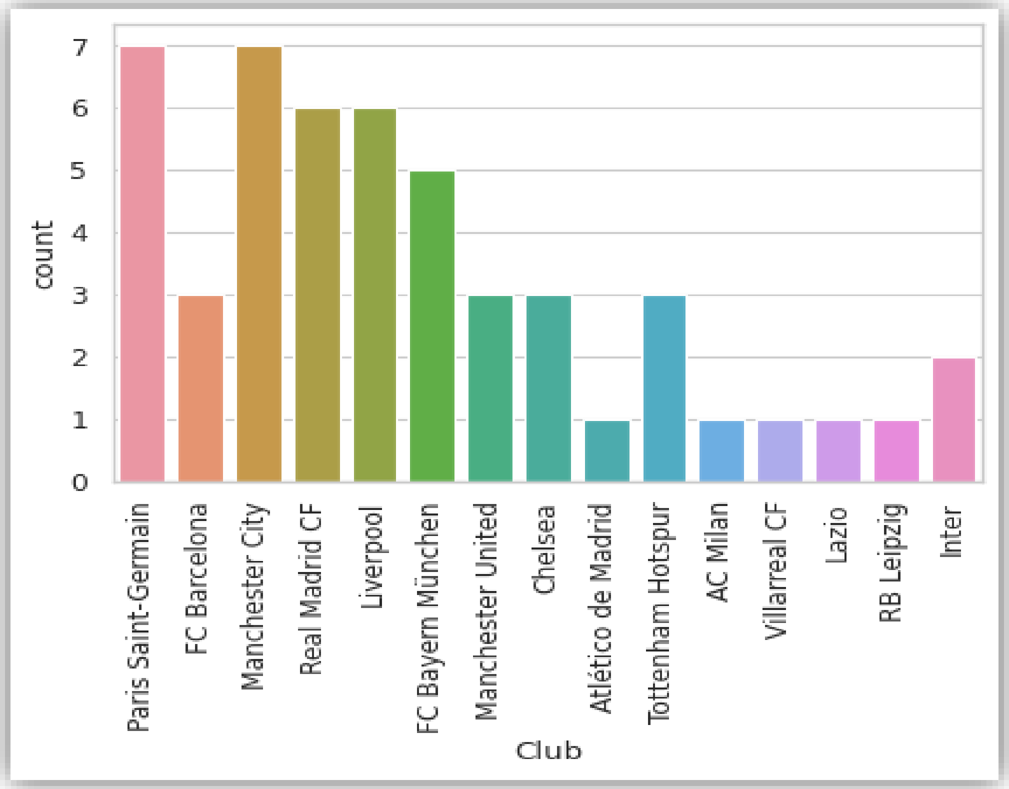
Show the effect of the Age on the Potential of the Players

While the Age Increases the Potential of the Player Decreases.



View the Top 50 Players and their Clubs





- Paris Saint-Germainand Machester City have the maximum top Players numbers

- Liverpool and Real Madrid have the second Maximum top Players numbers.

04

Data Preprocessing

Steps :

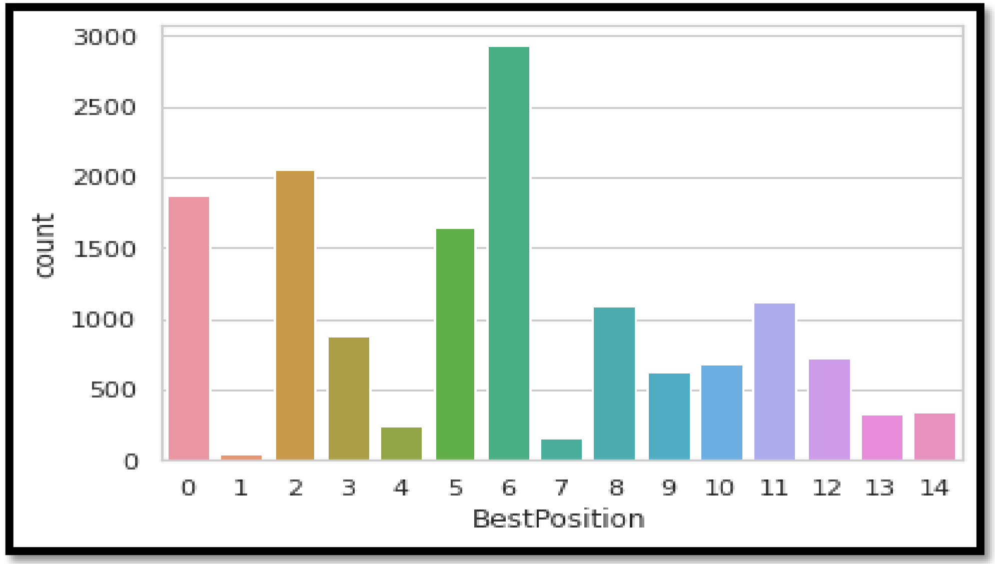
1. Handle the missing values

2. Handle The Categorical Columns

3. Handle the Imbalanced Data

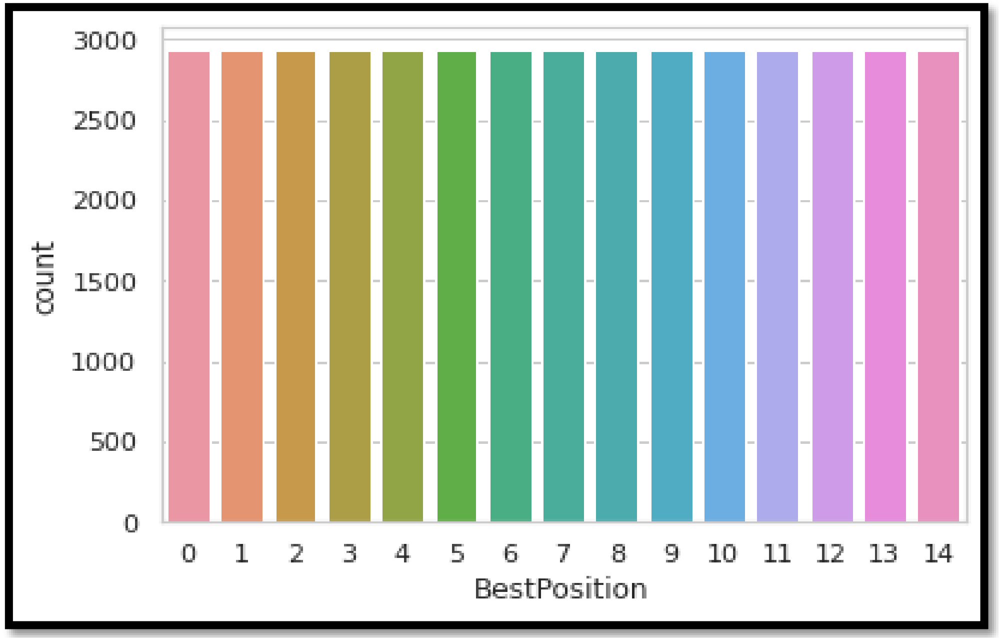
4. Feature Scaling

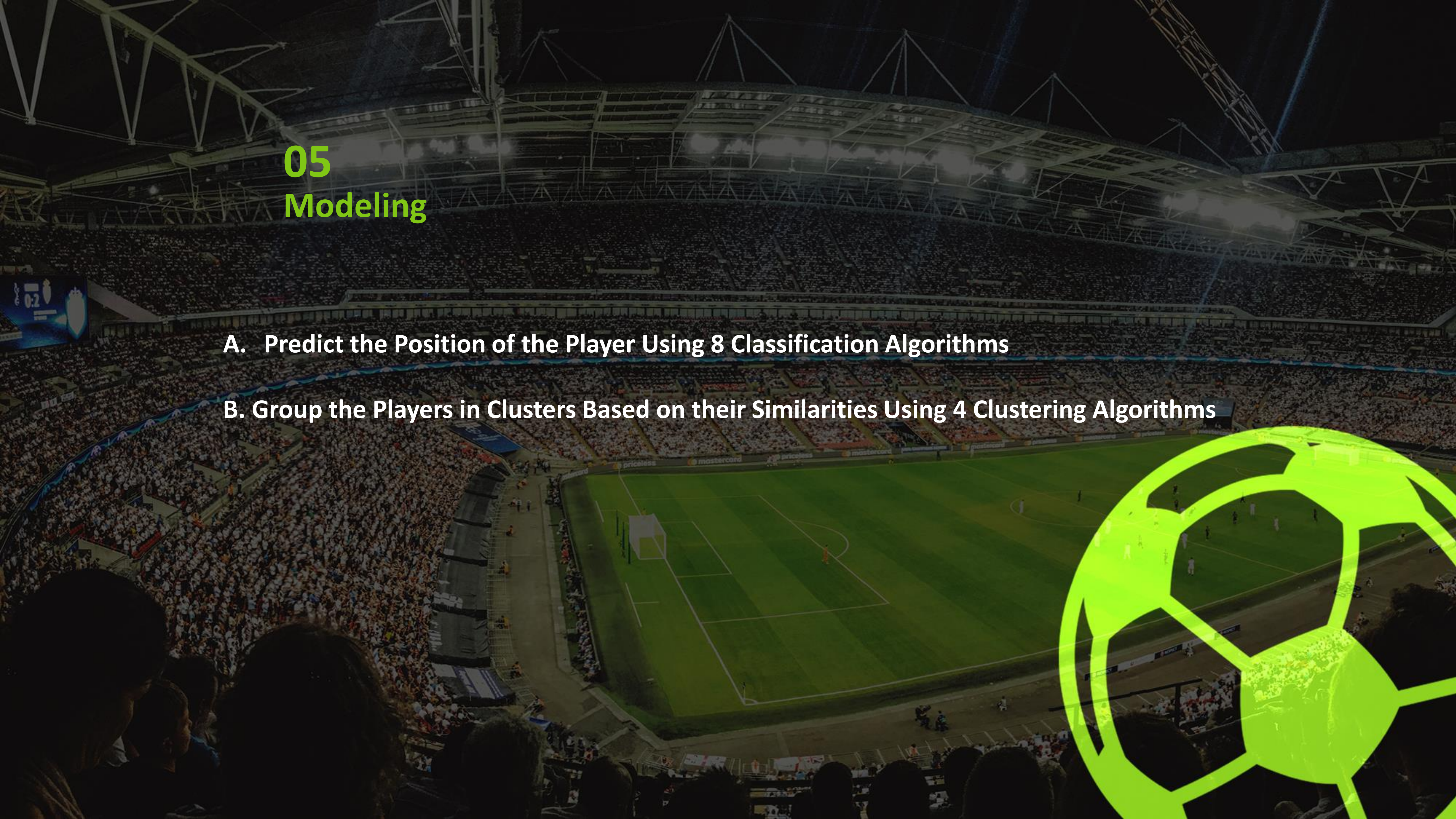
Handle the Imbalanced Data



As We can see Here the Data is Imbalanced s o we need to fix this issue.

Used the SMOTE method to Balance the Training Data





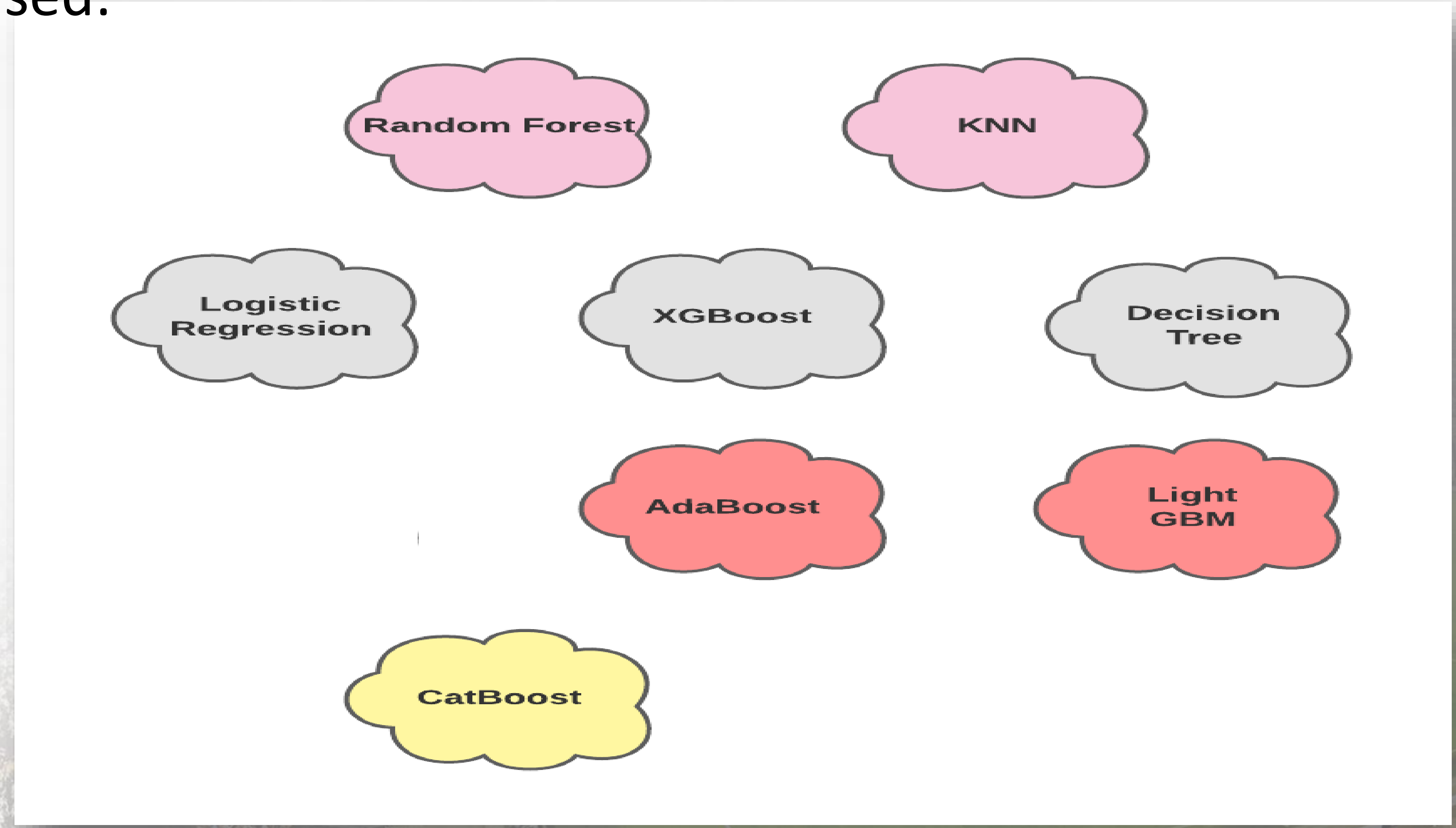
05
Modeling

A. Predict the Position of the Player Using 8 Classification Algorithms

B. Group the Players in Clusters Based on their Similarities Using 4 Clustering Algorithms

A.Predict the Position of the Player

Models Used:



Random Forest

KNN

Logistic Regression

XGBoost

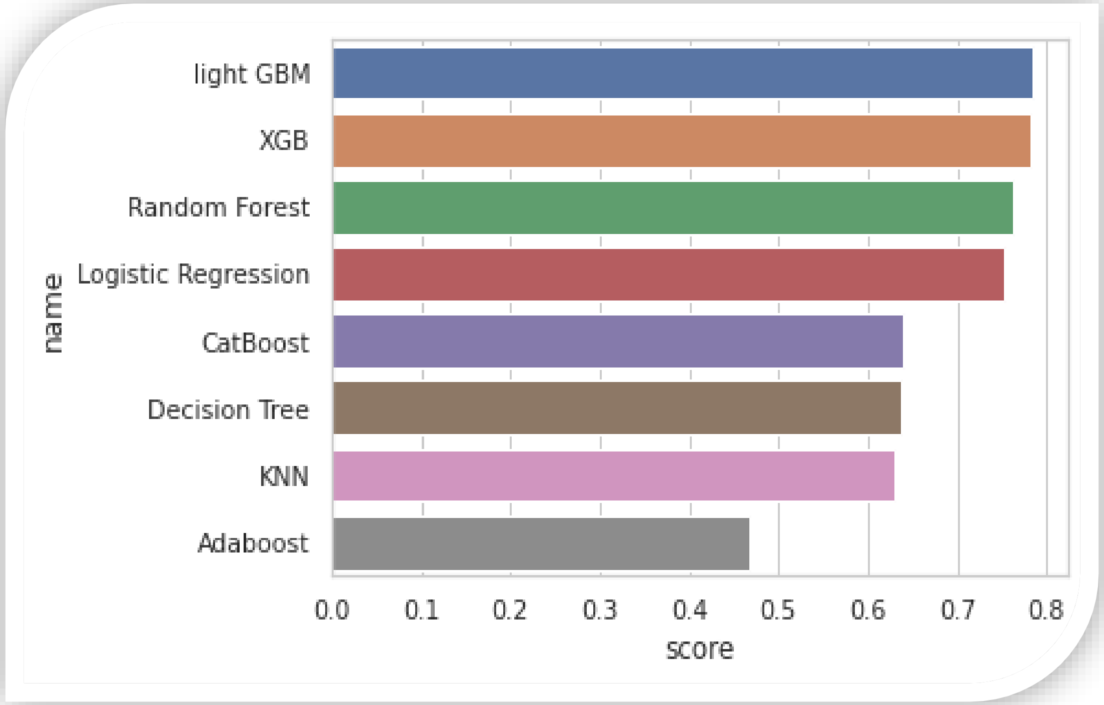
Decision Tree

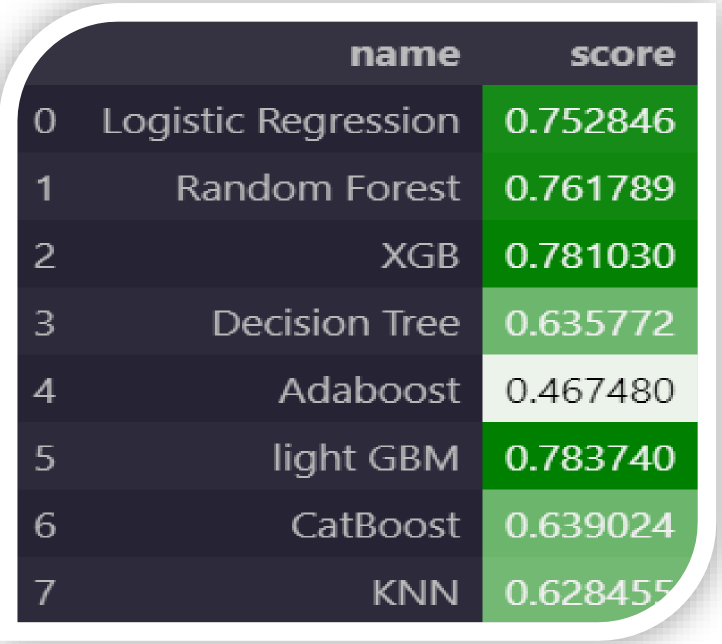
AdaBoost

Light GBM

CatBoost

Comparing the test accuracy of the 8 Algorithms

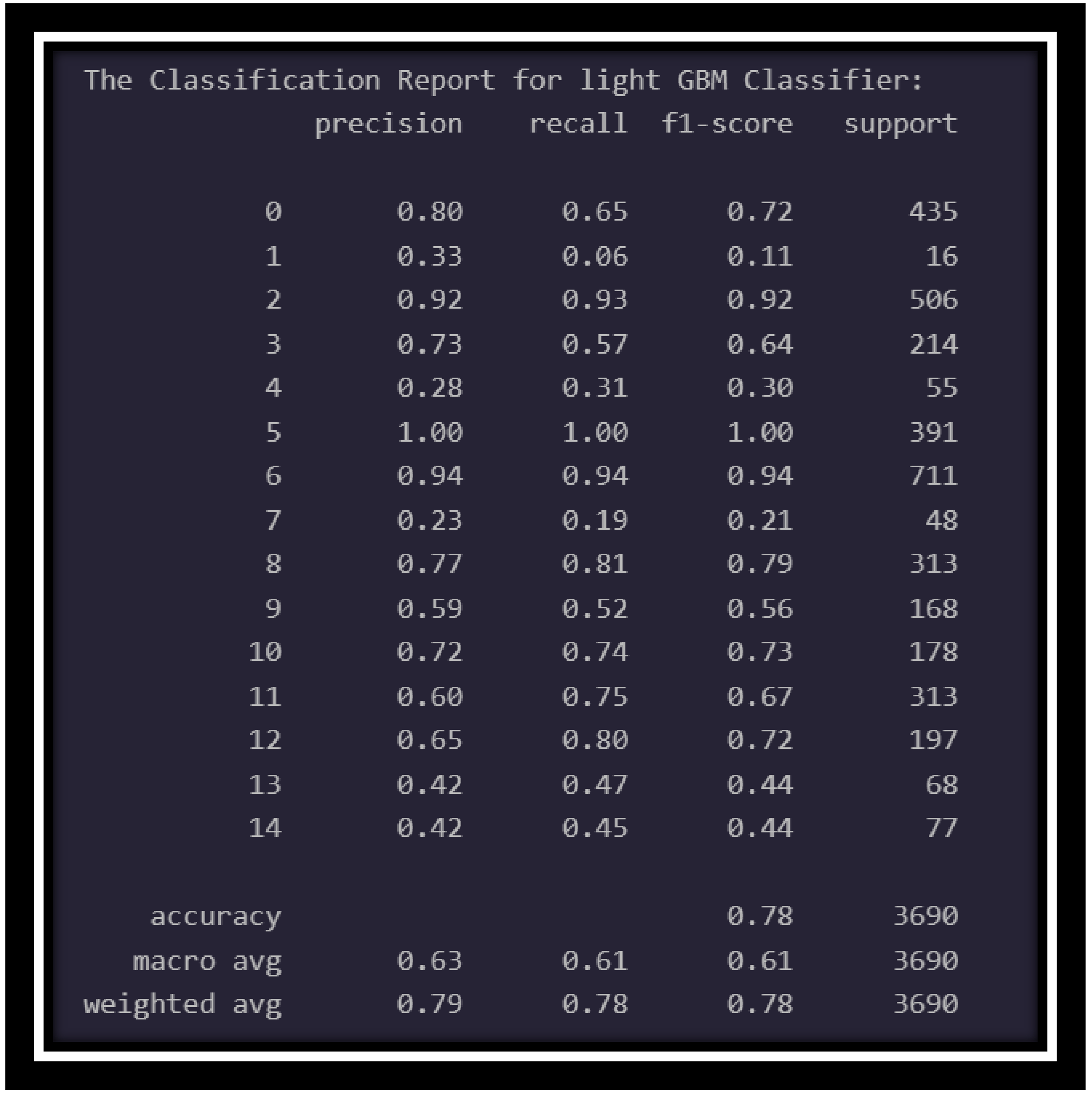




	name	score
0	Logistic Regression	0.752846
1	Random Forest	0.761789
2	XGB	0.781030
3	Decision Tree	0.635772
4	Adaboost	0.467480
5	light GBM	0.783740
6	CatBoost	0.639024
7	KNN	0.628455

So We Can Say that the light GBM and the XGB Algorithms are the best 2 Algorithms for that problem.

Light GBM

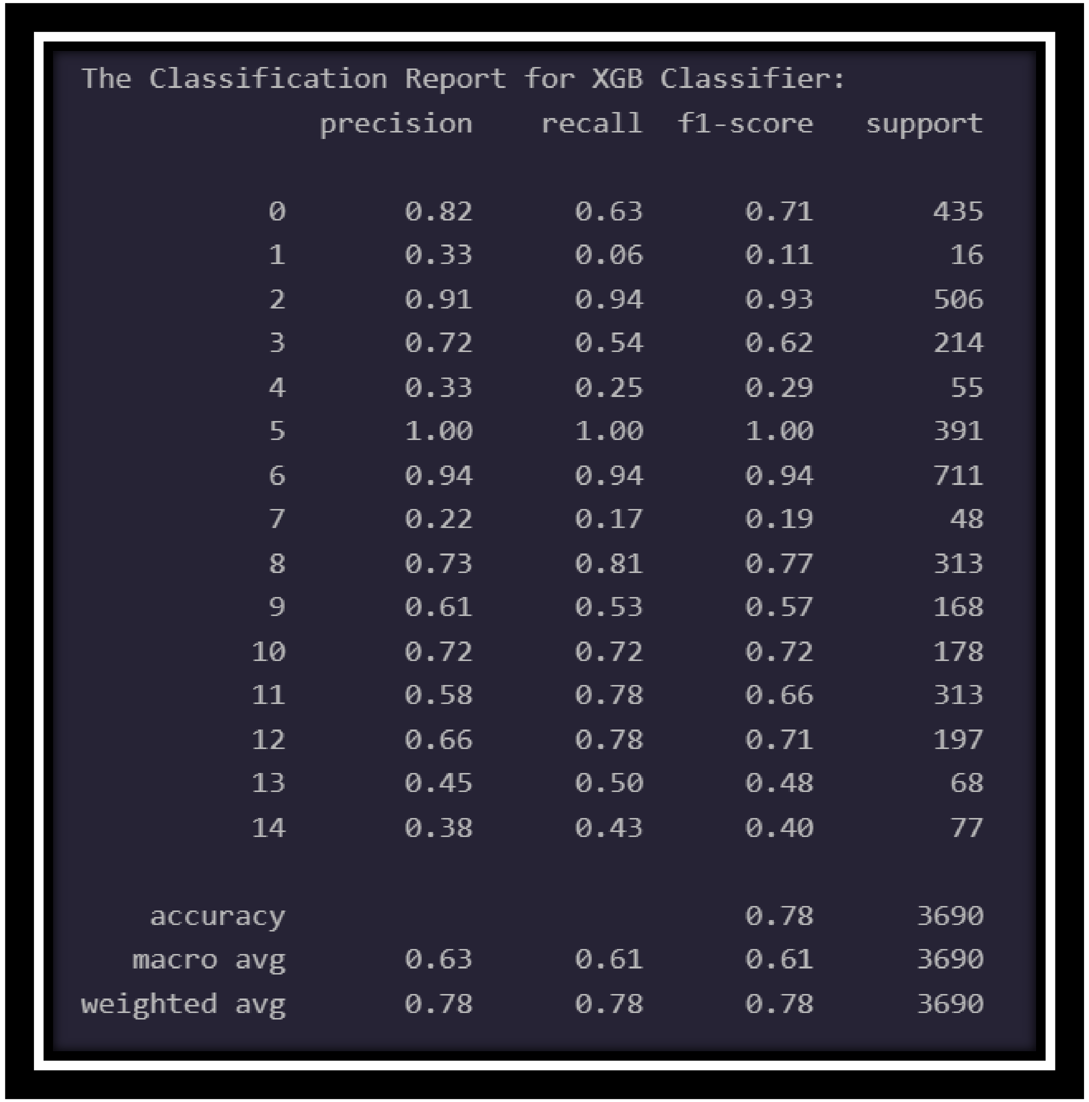


```
The Classification Report for light GBM Classifier:
              precision    recall  f1-score   support

           0       0.80      0.65      0.72       435
           1       0.33      0.06      0.11        16
           2       0.92      0.93      0.92       506
           3       0.73      0.57      0.64       214
           4       0.28      0.31      0.30        55
           5       1.00      1.00      1.00       391
           6       0.94      0.94      0.94       711
           7       0.23      0.19      0.21        48
           8       0.77      0.81      0.79       313
           9       0.59      0.52      0.56       168
          10       0.72      0.74      0.73       178
          11       0.60      0.75      0.67       313
          12       0.65      0.80      0.72       197
          13       0.42      0.47      0.44        68
          14       0.42      0.45      0.44        77

    accuracy                           0.78      3690
   macro avg       0.63      0.61      0.61      3690
weighted avg       0.79      0.78      0.78      3690
```

XGB



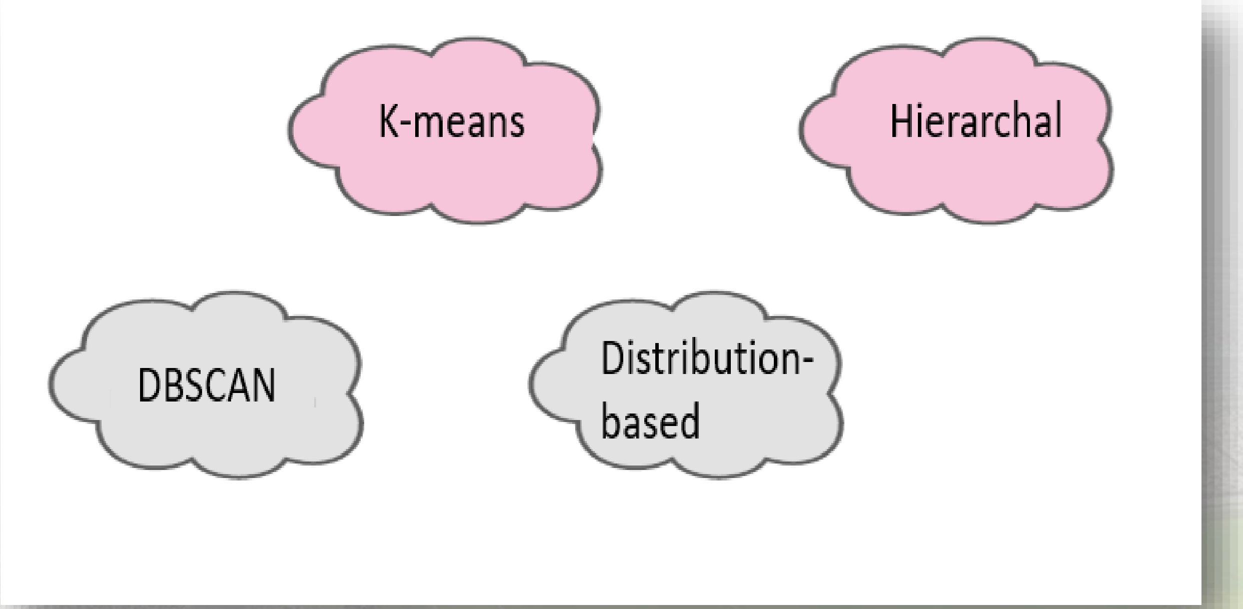
```
The Classification Report for XGB Classifier:
              precision    recall  f1-score   support

           0       0.82      0.63      0.71       435
           1       0.33      0.06      0.11        16
           2       0.91      0.94      0.93       506
           3       0.72      0.54      0.62       214
           4       0.33      0.25      0.29        55
           5       1.00      1.00      1.00       391
           6       0.94      0.94      0.94       711
           7       0.22      0.17      0.19        48
           8       0.73      0.81      0.77       313
           9       0.61      0.53      0.57       168
          10       0.72      0.72      0.72       178
          11       0.58      0.78      0.66       313
          12       0.66      0.78      0.71       197
          13       0.45      0.50      0.48        68
          14       0.38      0.43      0.40        77

    accuracy                           0.78      3690
   macro avg       0.63      0.61      0.61      3690
weighted avg       0.78      0.78      0.78      3690
```

B. Group the Players in Clusters (with Overall > 86)

Models Used:



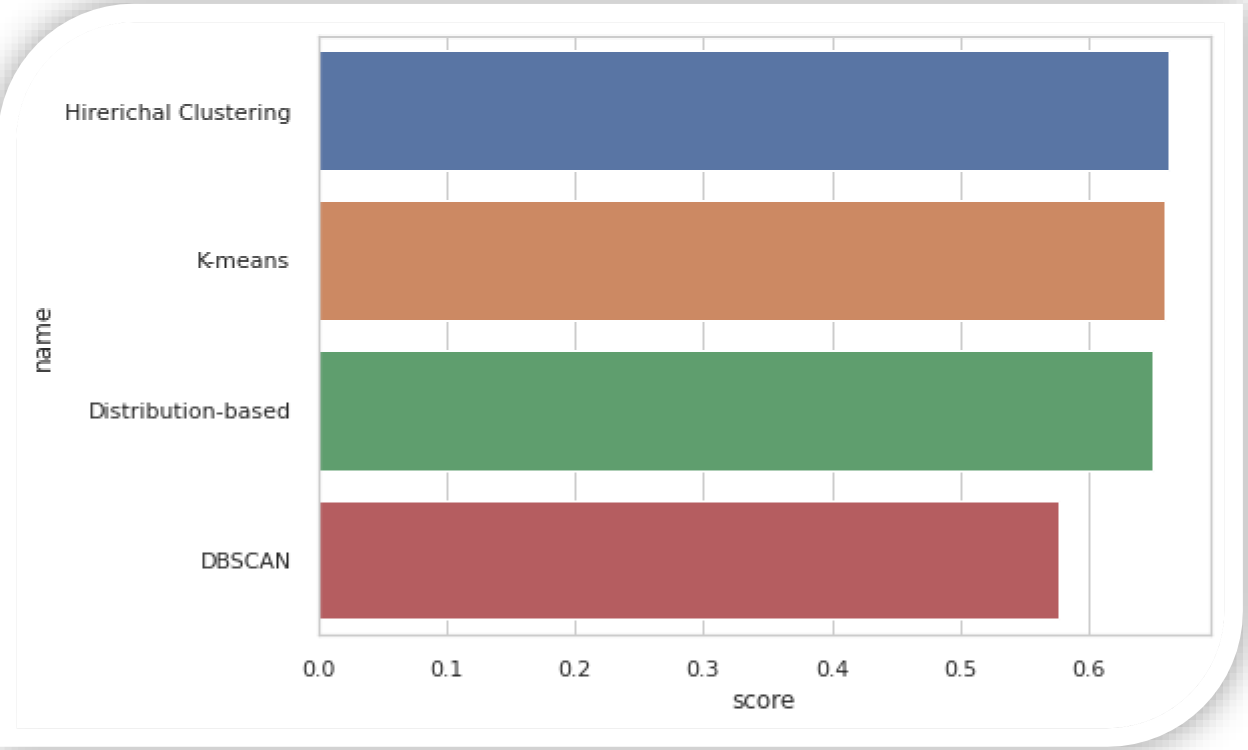
K-means

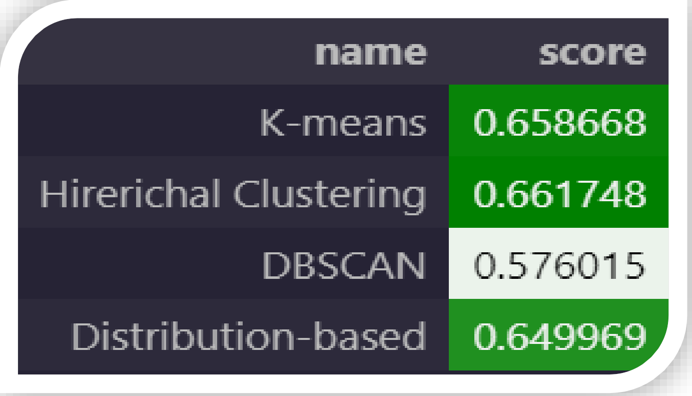
Hierarchal

DBSCAN

Distribution-based

Comparing the 4 Algorithms based on the Silhouette Score

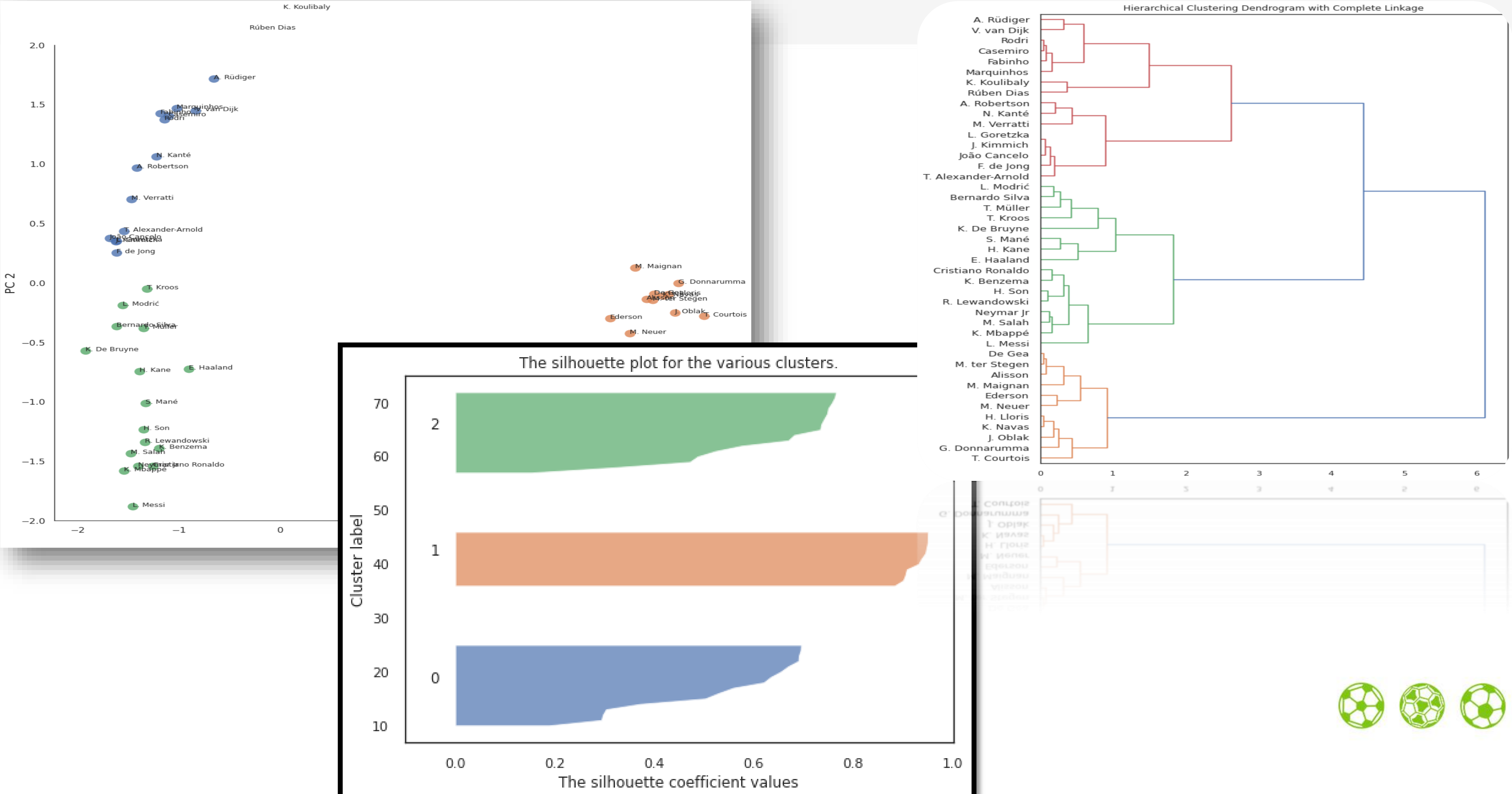




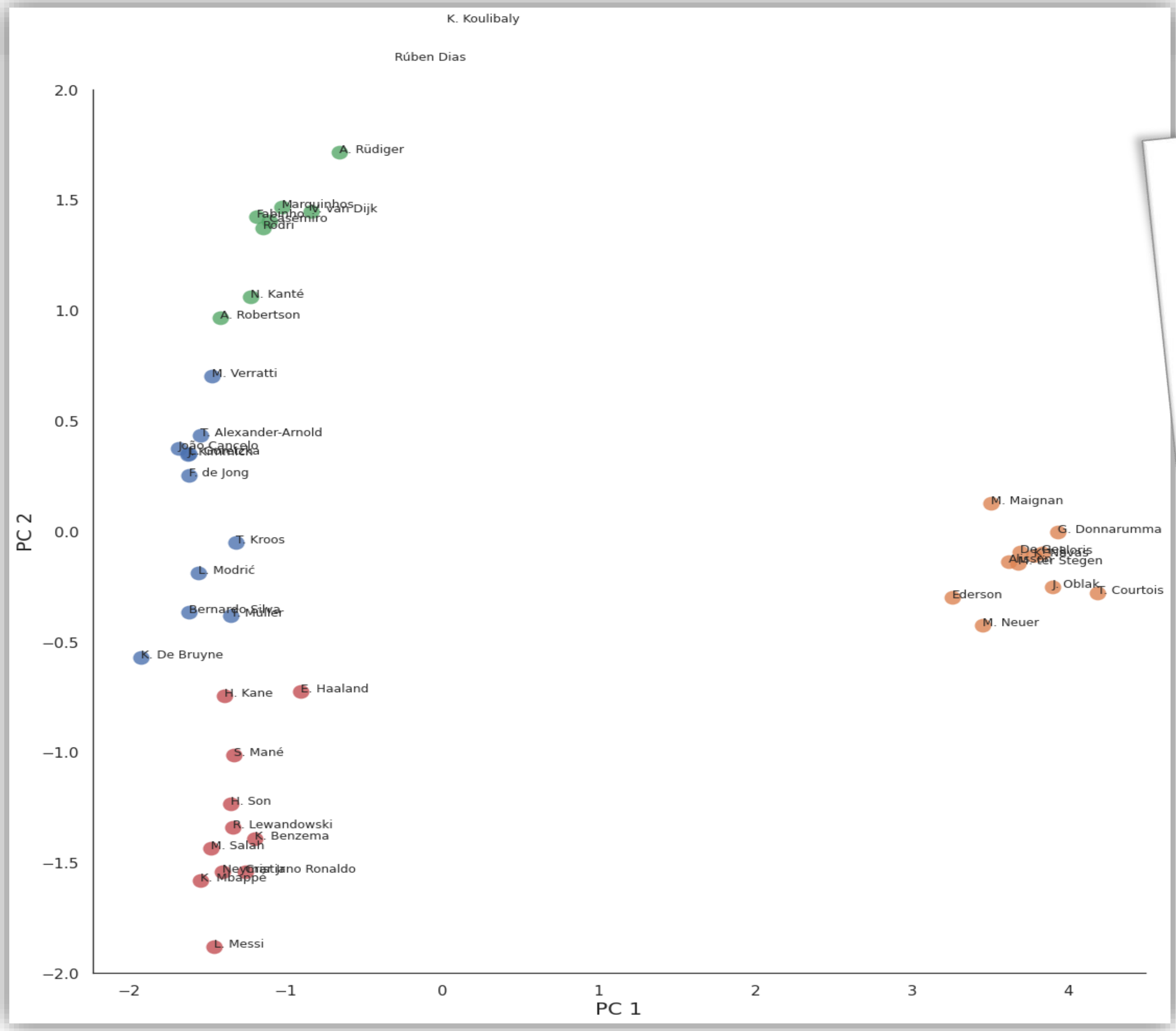
name	score
K-means	0.658668
Hirerichal Clustering	0.661748
DBSCAN	0.576015
Distribution-based	0.649969

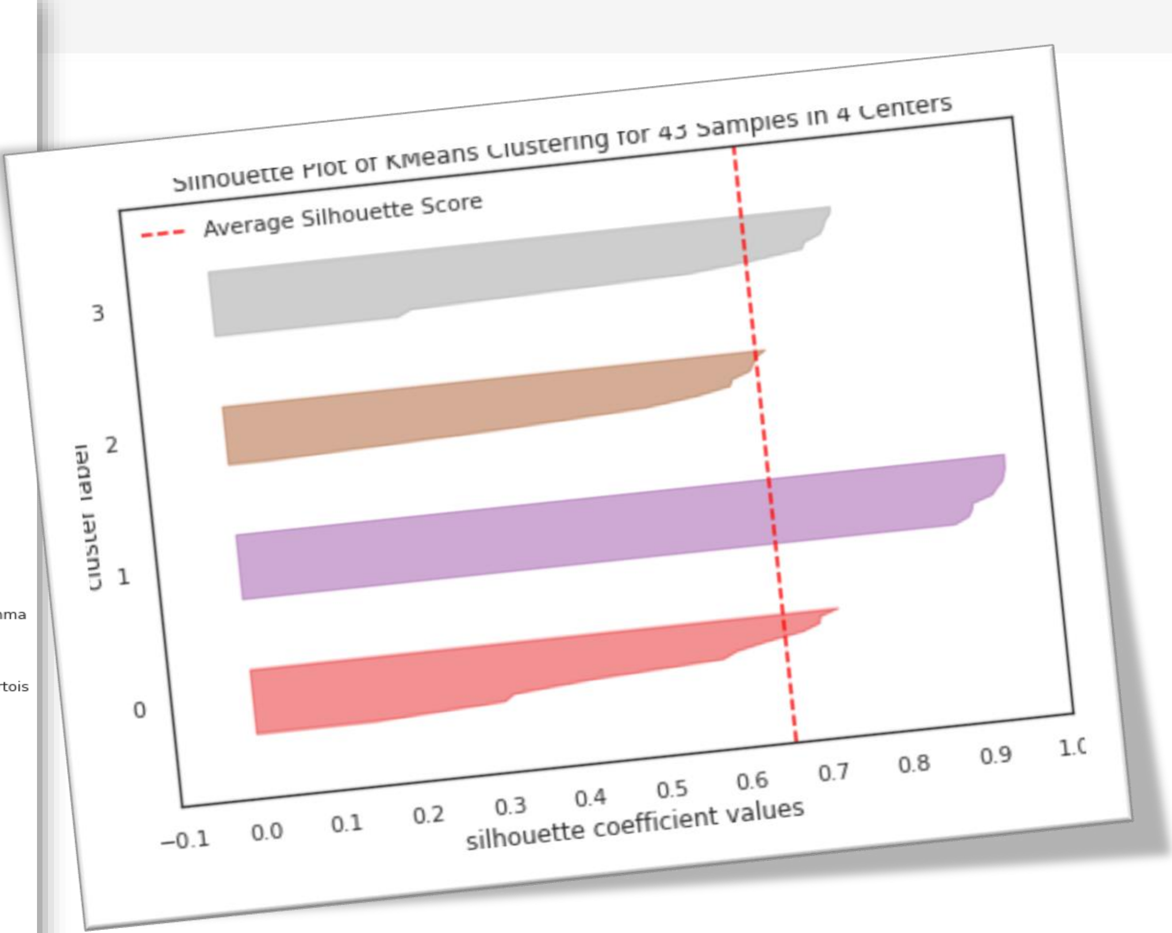
So We Can Say that the Hierarchal Clustering and the K-means Algorithms are the best 2 Algorithms for that problem.

Hierarchal Clustering



K-means


K. Koulibaly
Rúben Dias
A. Rüdiger
Marquinhos
Fabinho
V. van Dijk
Casemiro
Rodri
N. Kanté
A. Robertson
M. Verratti
T. Alexander-Arnold
João Cancelo
J. Kimmich
L. Goretzka
F. de Jong
T. Kroos
L. Modrić
Bernardo Silva
T. Müller
K. De Bruyne
H. Kane
E. Haaland
S. Mané
H. Son
R. Lewandowski
K. Benzema
M. Salah
Neymar
Cristiano Ronaldo
K. Mbappé
L. Messi
M. Maignan
G. Donnarumma
De Gea
H. Lloris
K. Navas
Alisson
M. ter Stegen
J. Oblak
T. Courtois
Ederson
M. Neuer
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
PC 2
−2
−1
0
1
2
3
4
PC 1


Silhouette Plot of KMeans Clustering for 43 Samples in 4 Centers
Average Silhouette Score
3
2
1
0
cluster label
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
silhouette coefficient values

THANK YOU