Gebze Technical University
Computer Engineering
CSE 454 Data Mining
Fall 2020 Project Assignments

Project Report

Gökhan Has
161044067

Lecturer : Dr. Burcu Yılmaz

Running Clustering Algorithms and Visualizing Data on Fifa 2021 Dataset

Summary

In this study, it provides information about the implementation of DBScan, Hierarchical and Mean Shift Clustering algorithms on Fifa 2021 data. The aim of the study is to classify the approximately 17000 players in the Fifa 2021 game according to their characteristics in 80 different categories and to identify players with anormal data, if any. In this study, it is also planned to investigate whether the same players in the specified clustering algorithms will be included in the same clusters or in different clusters.

Introduction

FIFA 2021 is a football simulation game developed by EA Sports company. Game data includes detailed information about the players and teams. There is a data set consisting of 80 columns for each player. The main ones are the salary, the transfer fee, which foot, etc. are information. The data set includes information of 17 126 players from 164 countries, 651 teams. It has sufficient content for data set classification algorithms and we need to do some preliminary studies before starting operations on the data set.

PREPROCESS STEP

These clustering algorithms are algorithms that work on numerical data. Therefore, non-numerical data in the data set has been turned into numerical data. It has been decided which data will be effective in clustering the players. Most of these data are variables of the type of continuous variable and these data are turned into categorical variables because they increase the granules. Variables such as M, K used in the data actually mean millions and thousands, and the places where these values are corrected by multiplying millions and thousands. To use the data more efficiently, the standard scaling method was used to pull the data into a more meaningful range. Thus, data such as wage and transfer price (value) were included in a meaningful scale.

MODELS

A) DBScan

DBSCAN algorithm was presented by Ester, Kriegel, Sander and Xu at the conference KDD'96. This algorithm calculates the distance of objects from their neighbors and performs clustering by grouping areas with more objects than a predetermined threshold in a given region. The DBSCAN algorithm has brought many new terms and approaches to data mining.

Since DBScan is a density base algorithm, it is affected by the proximity of data to each other, so it can discover randomly shaped clusters. It can find clusters that are completely surrounded by different clusters. It is sensitive in detecting outlier data. Since it is a density-based algorithm, it may not be able to cluster if the data is very sparse. Sampling affects density measures.

B) Hierarchical Clustering

Hierarchical clustering is a clustering algorithm as the name suggests. It has two different variations as Agglomerative (from part to whole) and Divisive (from whole to part). In this work, the Agglomerative version was studied.

In this variation, first all the data is made into a set, that is, if there are N elements, N sets are formed. Then, clusters that are close to each other in distance merge to form a new cluster. This situation continues until the system is stable or the number of clusters the programmer enters.

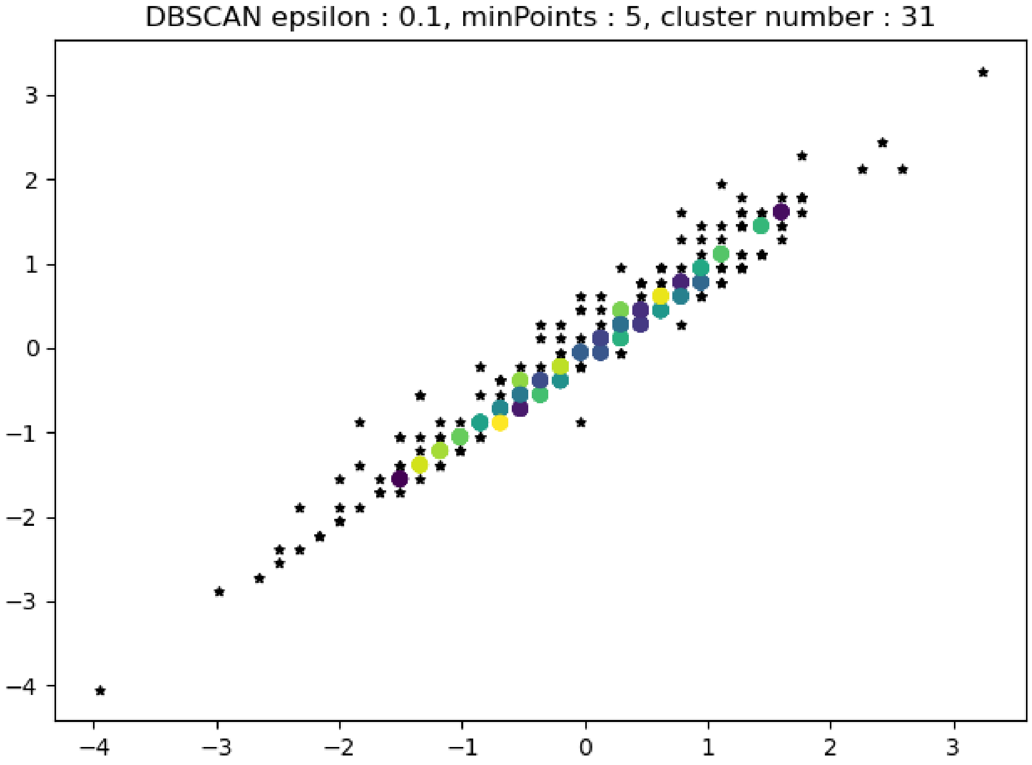
C) Mean Shift Clustering

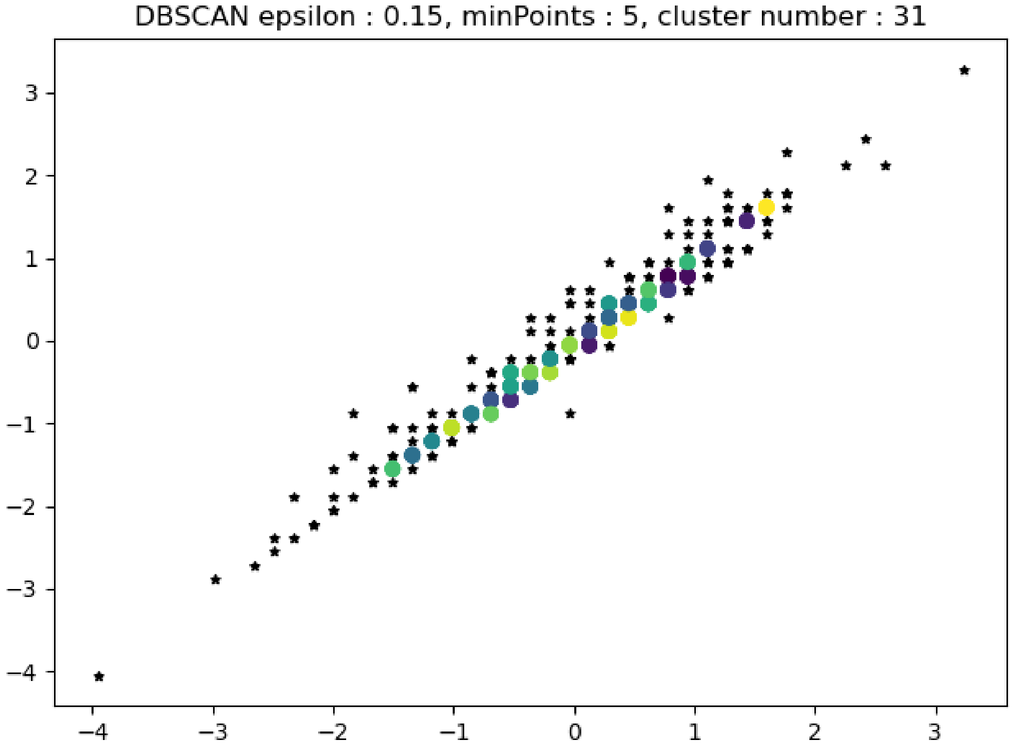
Meanshift falls under a category of clustering algorithms, unlike Unsupervised learning, which recursively allocates data points to clusters by shifting points towards mode (mode is the highest density of data points in the region in the context of Meanshift). Hence, it is also known as the Mod search algorithm. The average shift algorithm has applications in image processing and computer vision.

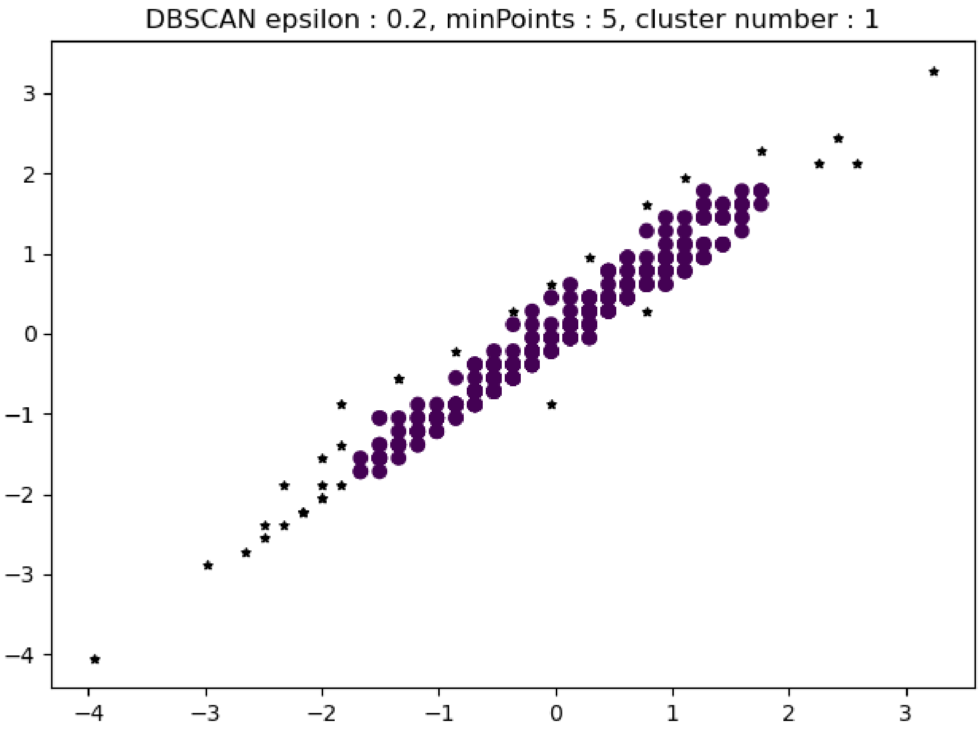
Given a set of data points, the algorithm recursively assigns each data point toward the nearest cluster center, and the direction to the closest cluster center point is determined by where most of the nearby points are. Therefore, each iteration, each data point will be closer to where the most points are located, which will reach the cluster center. When the algorithm stops, each point is assigned to a cluster.

MODELS RESULTS

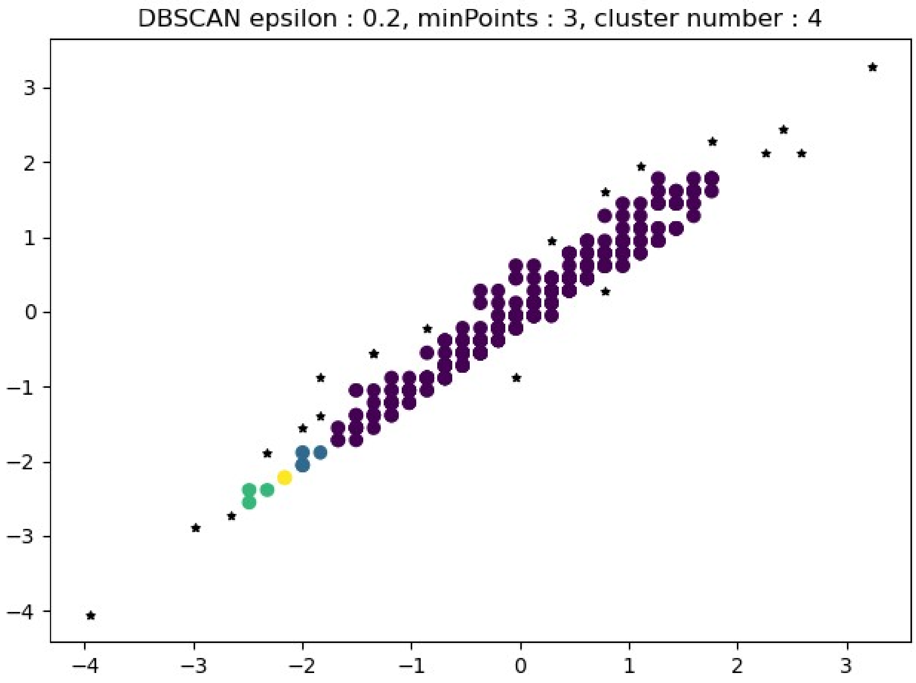
A) DBScan

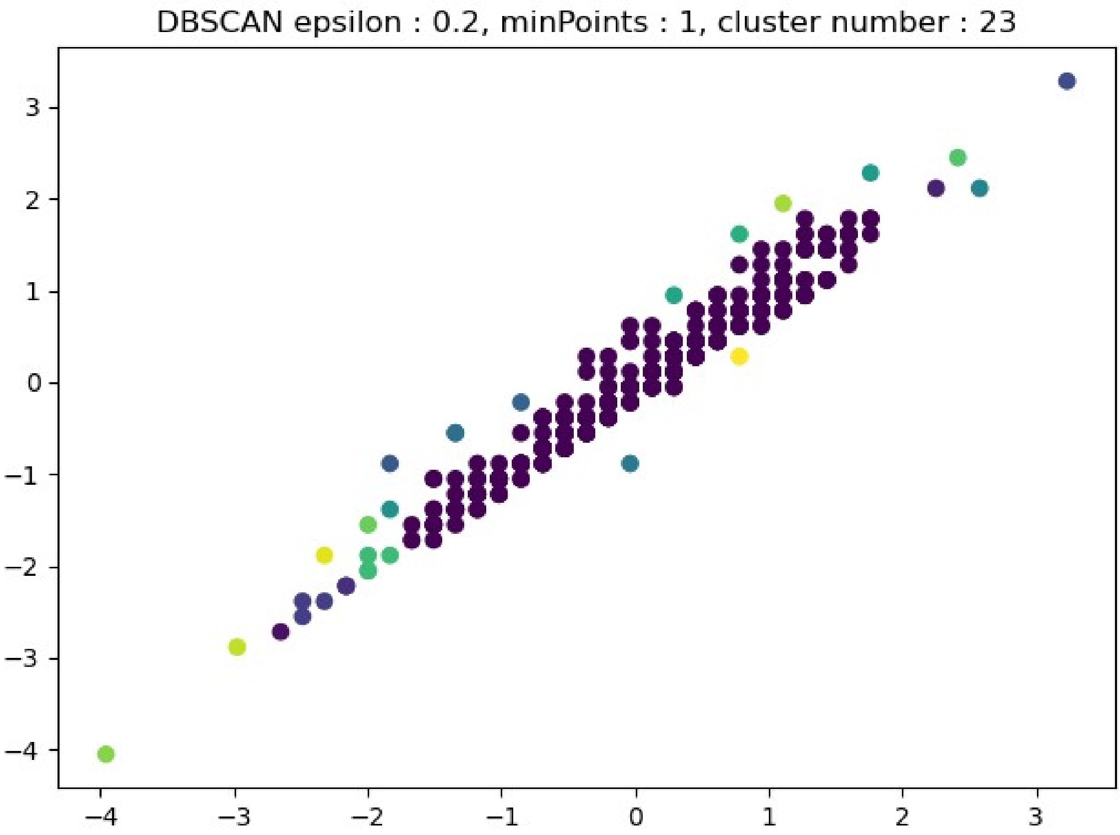






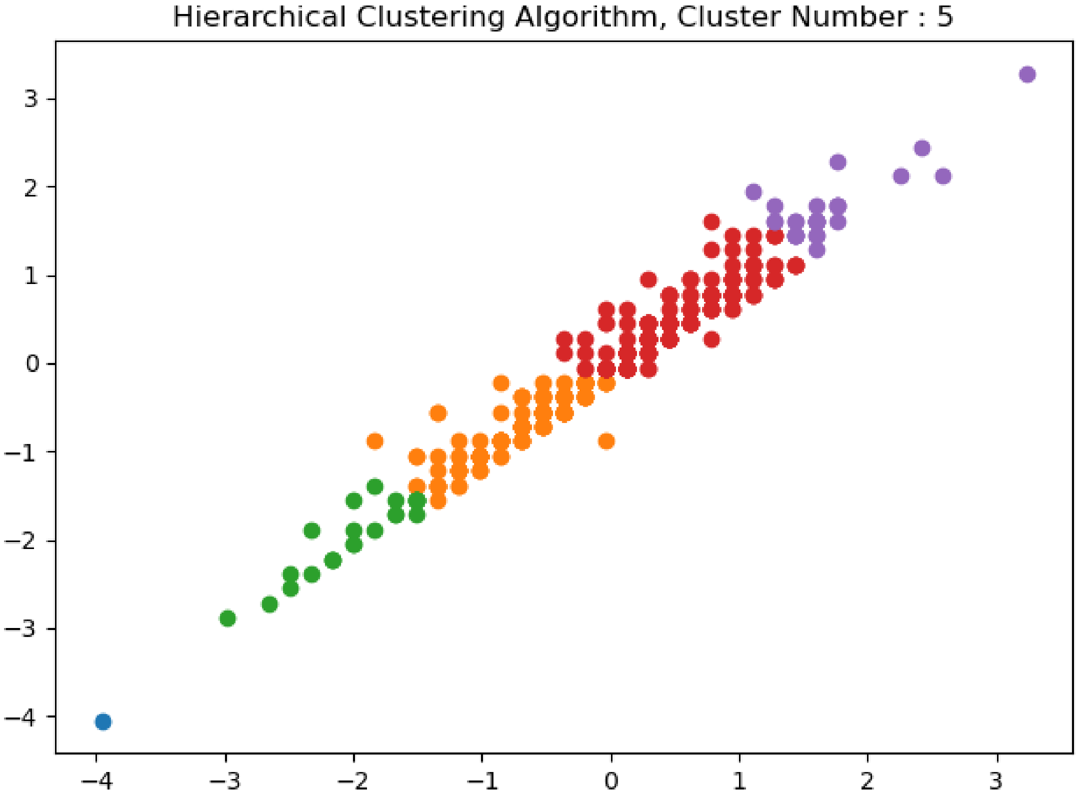
Now the minpoint variable will be kept constant at 5 and the epsilon value will be increased. As I value the Epsilon value relatively low, separate clusters were formed as seen in the plot. In the dataset I use, the data are collected on the middle side of the plot, so the clusters are on that side. Noise values have emerged on the right, where there is less data. And the areas where objects are relatively far from each other are these areas. I would like to point out that at 0.1 and 0.15 the same number of clusters is formed but different points are included in the cluster.

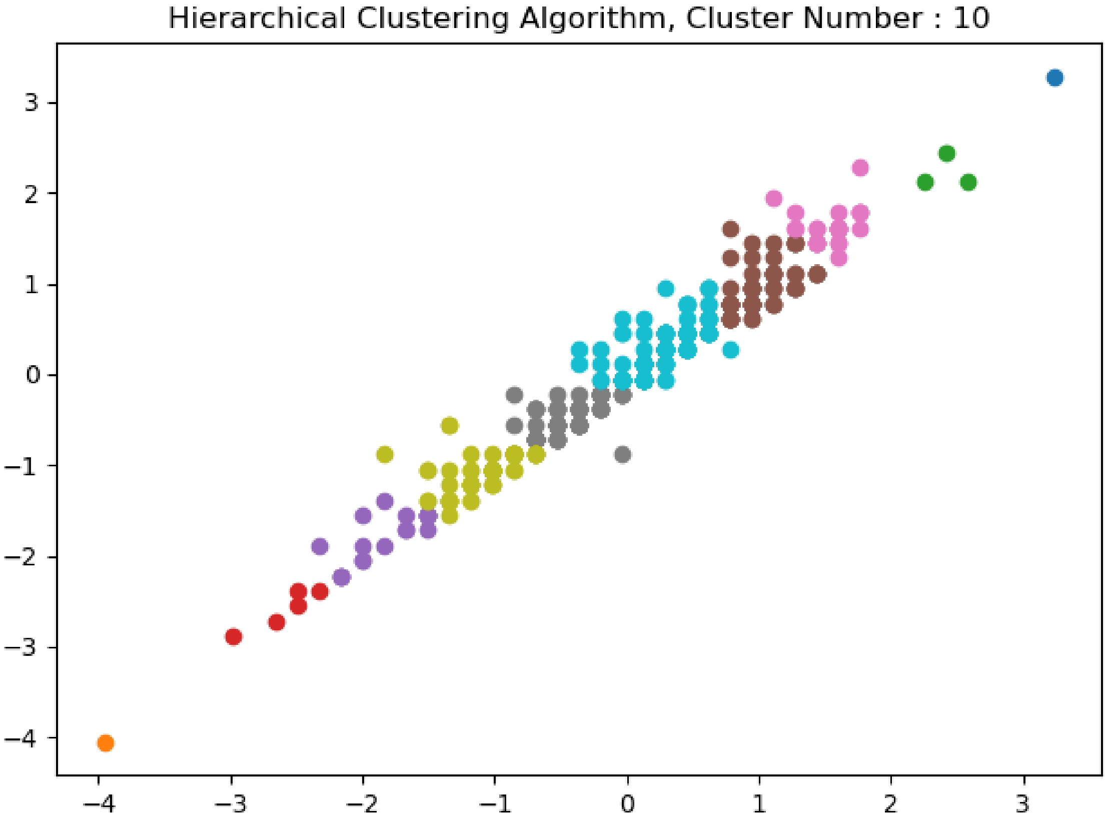


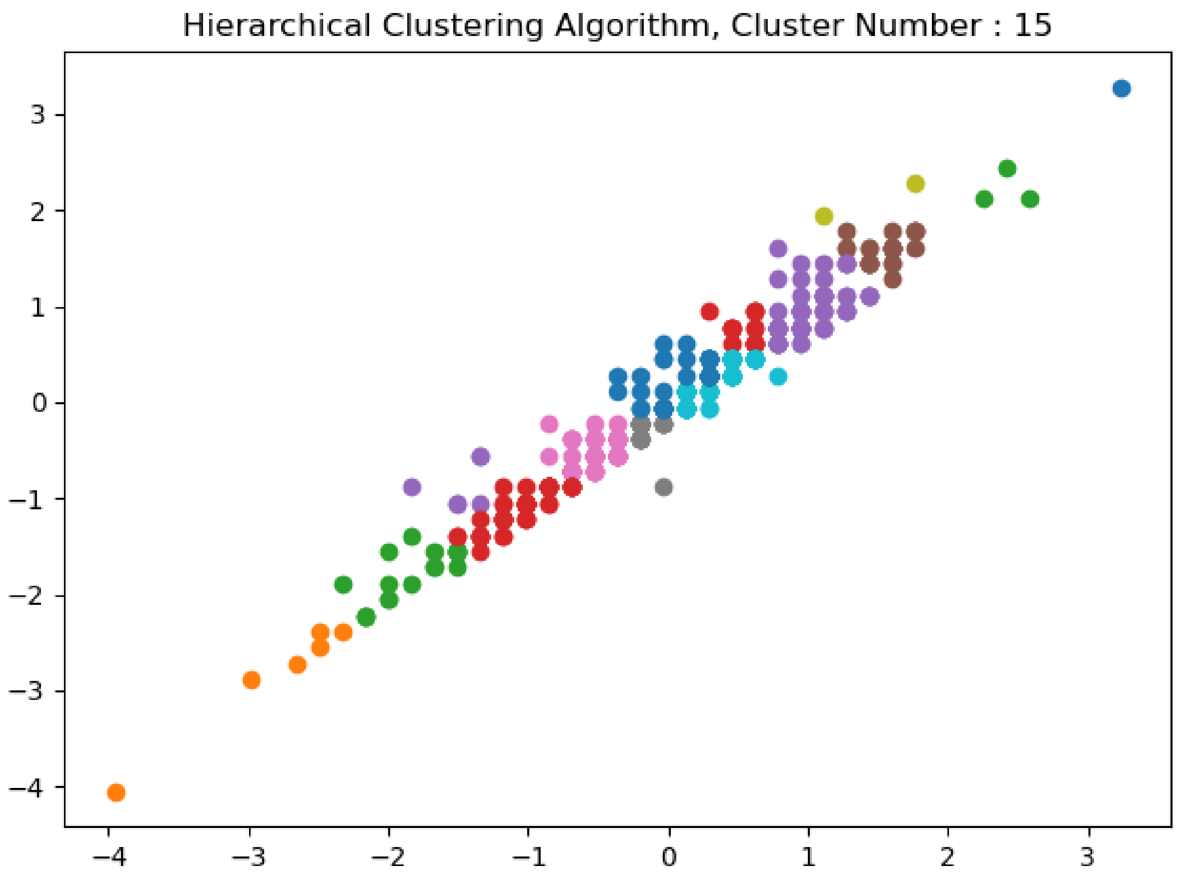


When I decreased the minpoints, I expect the number of clusters increased. Many clusters split. That's why it increased. Because now, theoretically, it is not provided that objects that have more distances to each other are in the same set. Likewise, it was observed that new clusters were formed on the right side where data was relatively less (distance between objects is relatively greater). Where objects are relatively close to each other, clusters are split, the number of clusters has increased.

B) Hierarchical Clustering

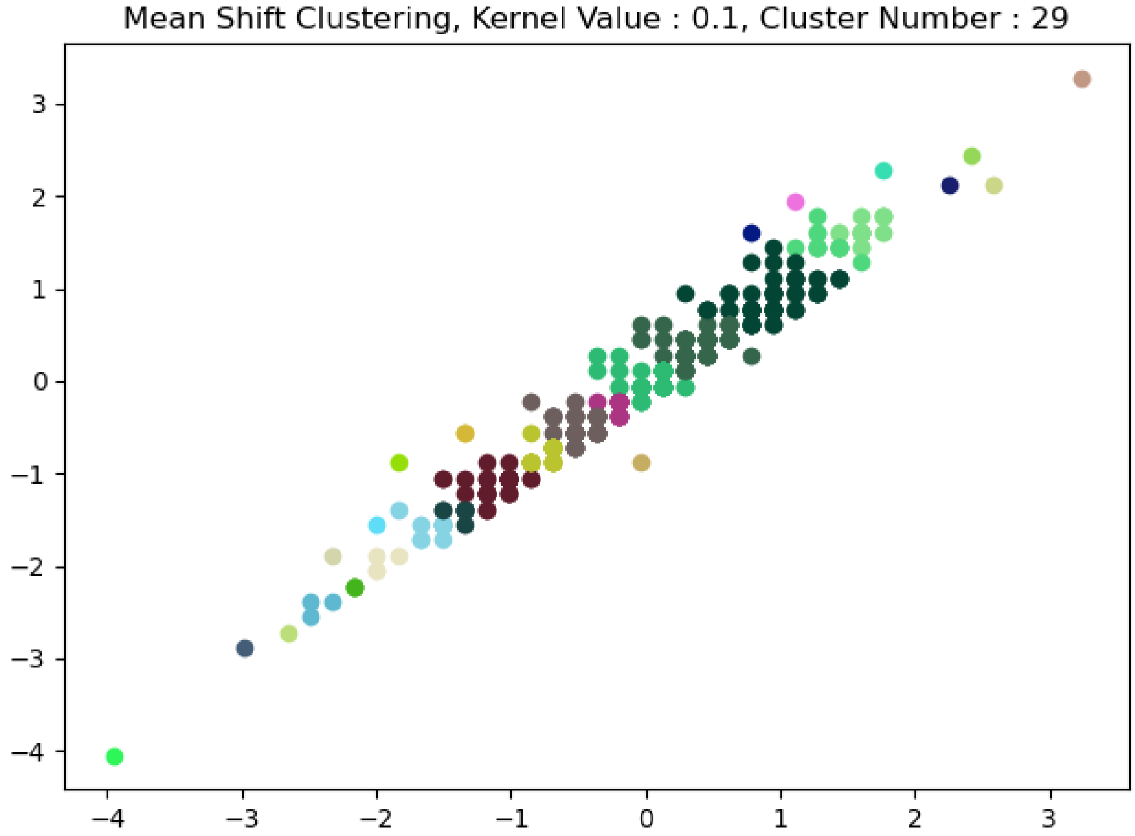


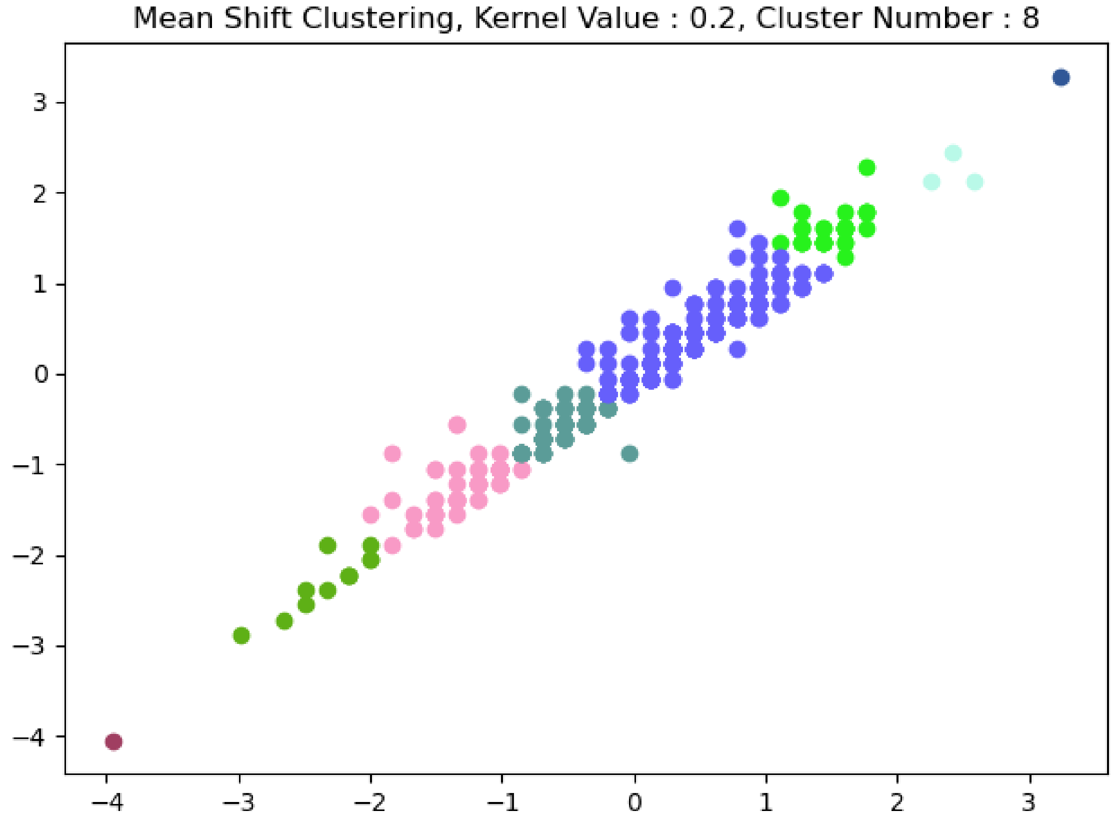


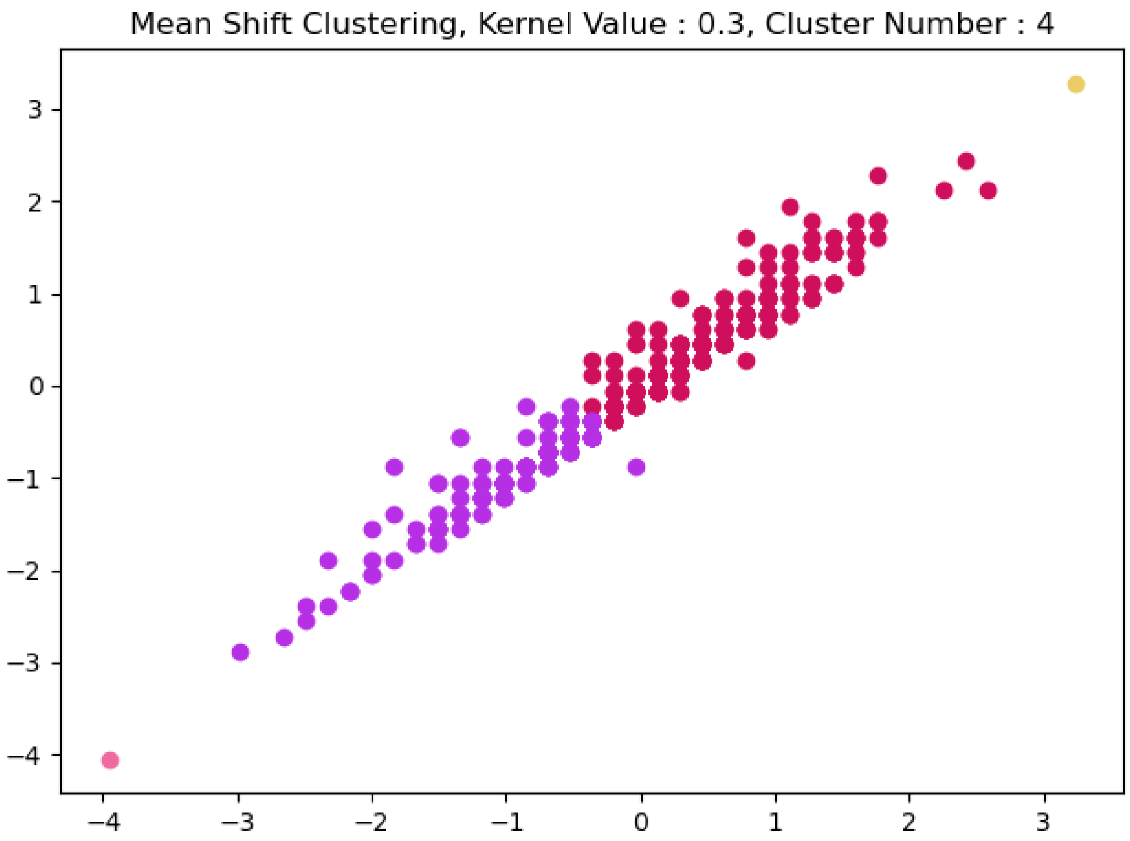


Single linkage is used in this algorithm. Single linkage is also sometimes called 'friends of friends' linkage method. The single linkage method always picks the minimum distance when it updates. Therefore, two points might be considered close under the single linkage scheme if they can be connected by a chain of points with small distances between any two consecutive points down the chain. The distance between the two endpoints of the chain is only as big as the longest link along the chain. The number of links along the chain does not matter. Therefore, it is expected that more clusters will be observed as the number of N increases. Players who play too well or those who play too bad are in a cluster.

C) Mean Shift Clustering







This algorithm starts from random data. And this data becomes the center of the kernel. The kernel is shifted as kernel value. This is the summary of the implementation of the algorithm. Here, it is observed that as you increase this value, the number of clusters decreases. Considering that the kernel has increased its hiccup, we can connect more easily. In this implementation I calculated according to the gauss distance.

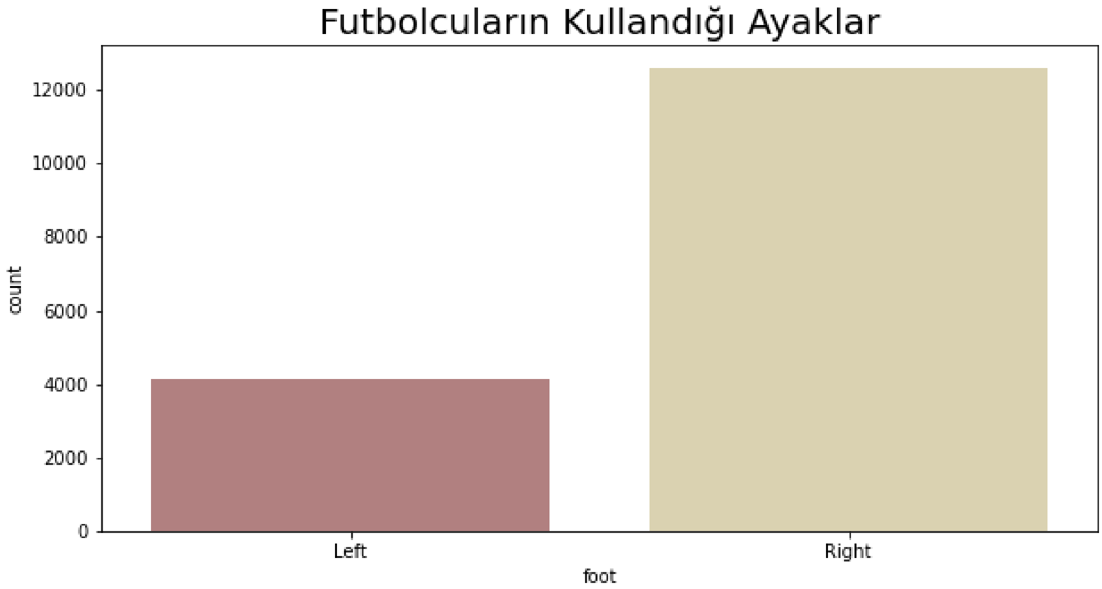
GENERAL RESULT

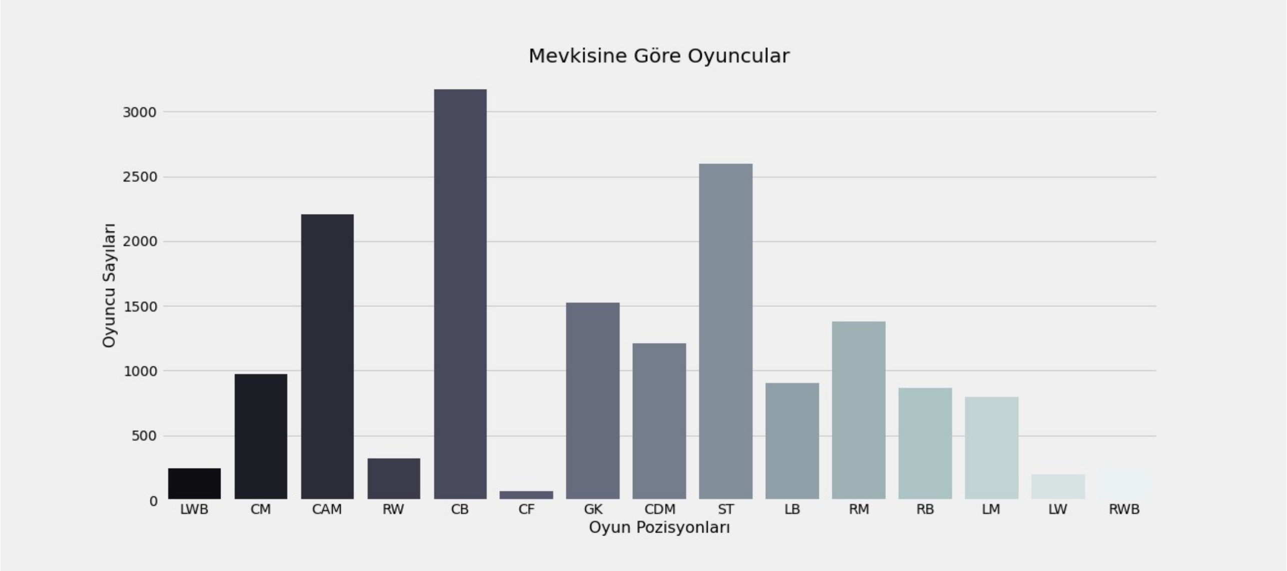
Thanks to the clustering algorithms applied on Fifa 2021 data, the data could be clustered. As a result of the applied algorithms, different results were obtained and similar results were obtained. The results were observed as expected and consistent with those observed in real life.

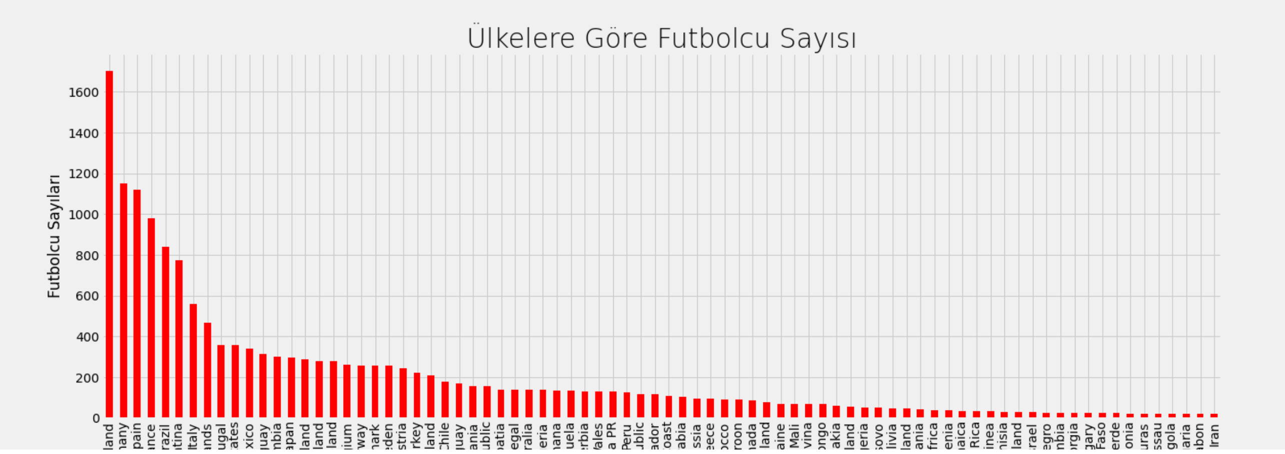
As a result of the algorithms applied, the data were classified successfully and consistently. As a result of the classification, samples that all algorithms gather in the same class were observed and it was concluded that this was a metric showing the accuracy of the study.

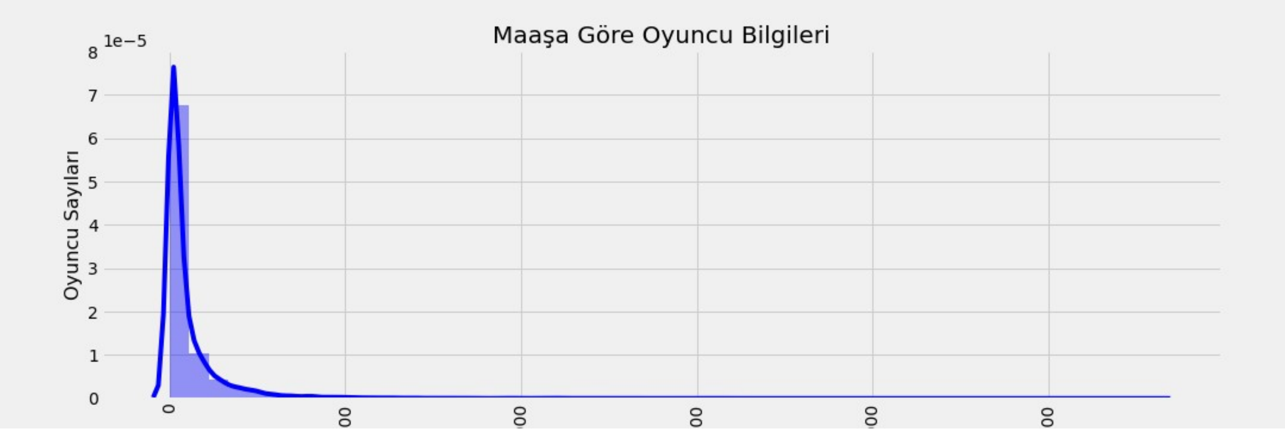
VISUALIZATION EXAMPLES

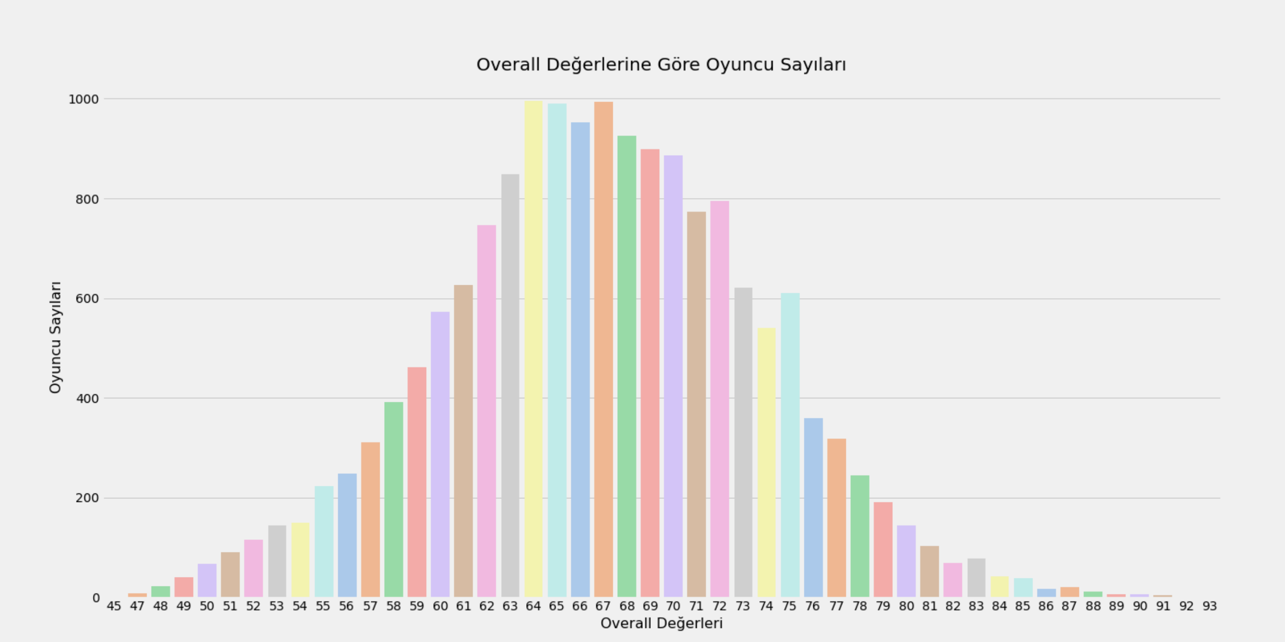
Images of the images are available in the result_pictures folder. It can be examined in detail from that folder.

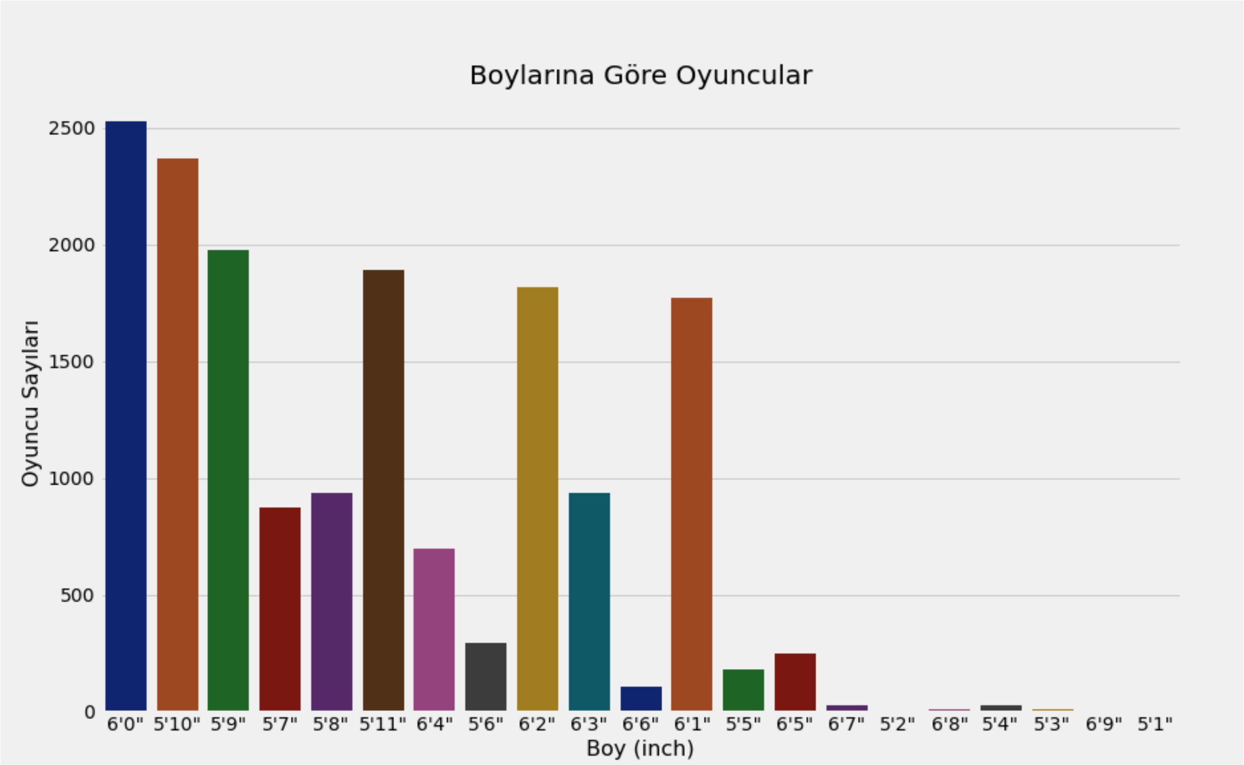


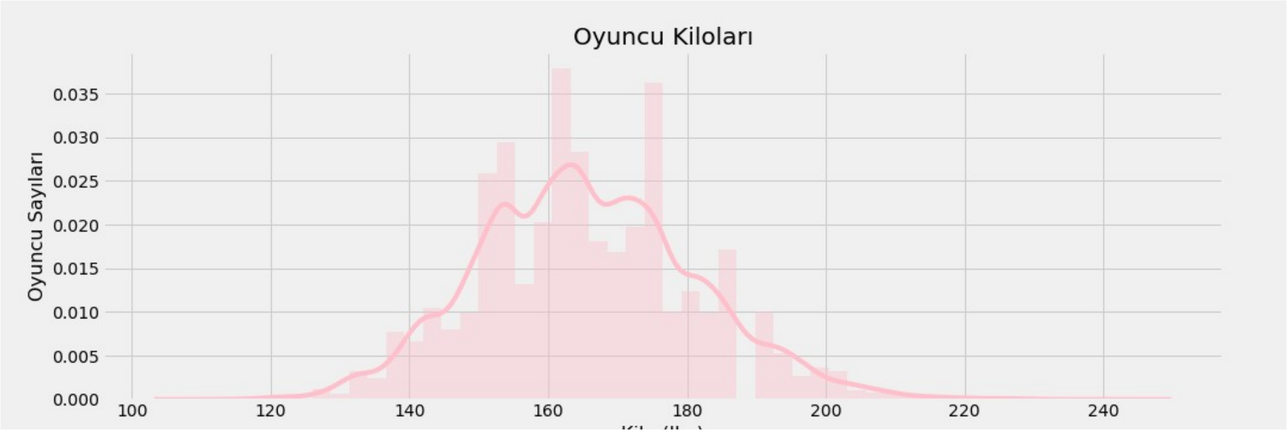


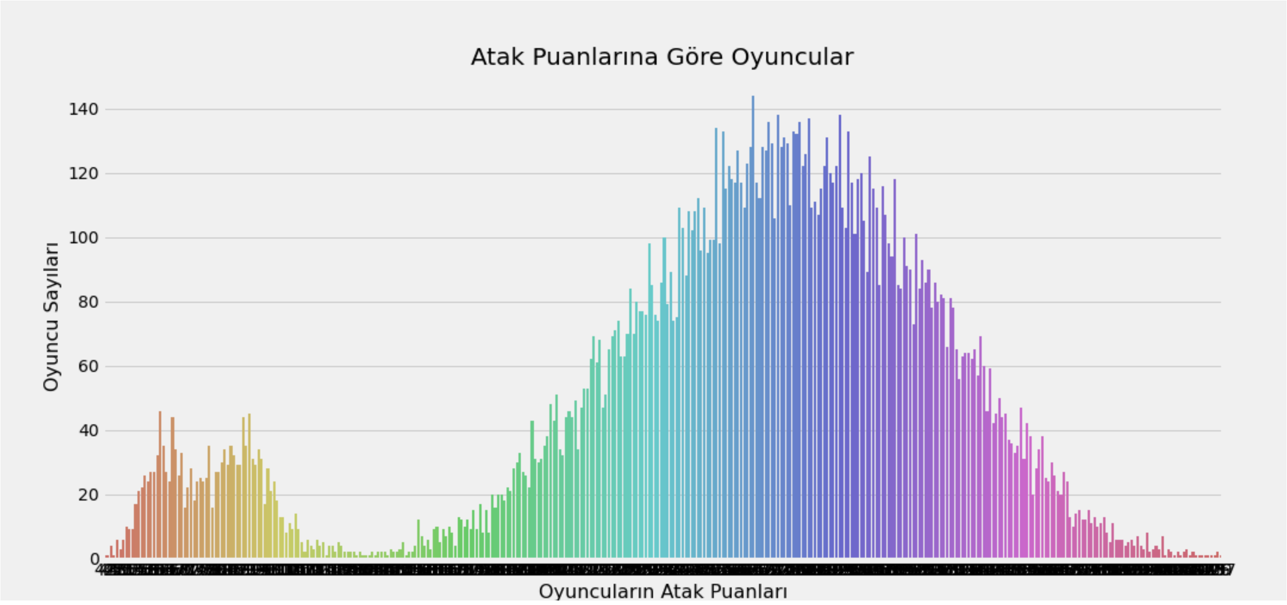

Ülkelere Göre Futbolcu Sayısı
Futbolcu Sayıları
0
200
400
600
800
1000
1200
1400
1600
land
many
pain
ance
razil
ntina
Italy
ands
ugal
ates
xico
guay
mbia
apan
land
land
land
gium
way
mark
eden
stria
rkey
land
Chile
guay
ania
ublic
oatia
egal
ralia
eria
hana
uela
erbia
Wales
a PR
Peru
ublic
ador
oast
abia
ssia
eece
occo
roon
nada
land
aine
Mali
vina
ongo
akia
land
geria
sovo
livia
land
ania
frica
enia
aica
Rica
inea
nisia
land
srael
egro
mbia
orgia
gary
Faso
erde
onia
uras
ssau
gola
aria
abon
Iran

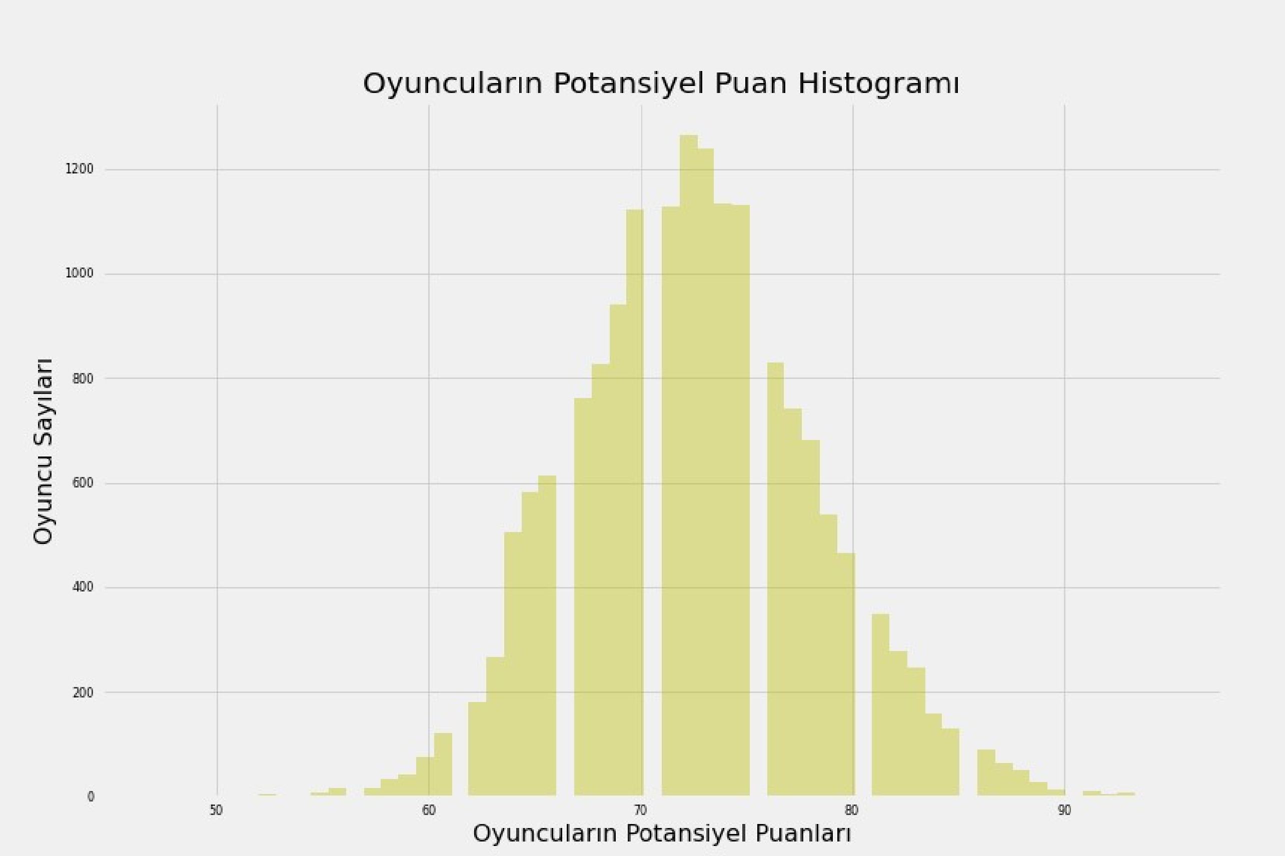

Maaşa Göre Oyuncu Bilgileri
1e−5
Oyuncu Sayıları
0
1
2
3
4
5
6
7
8
0
00
00
00
00
00

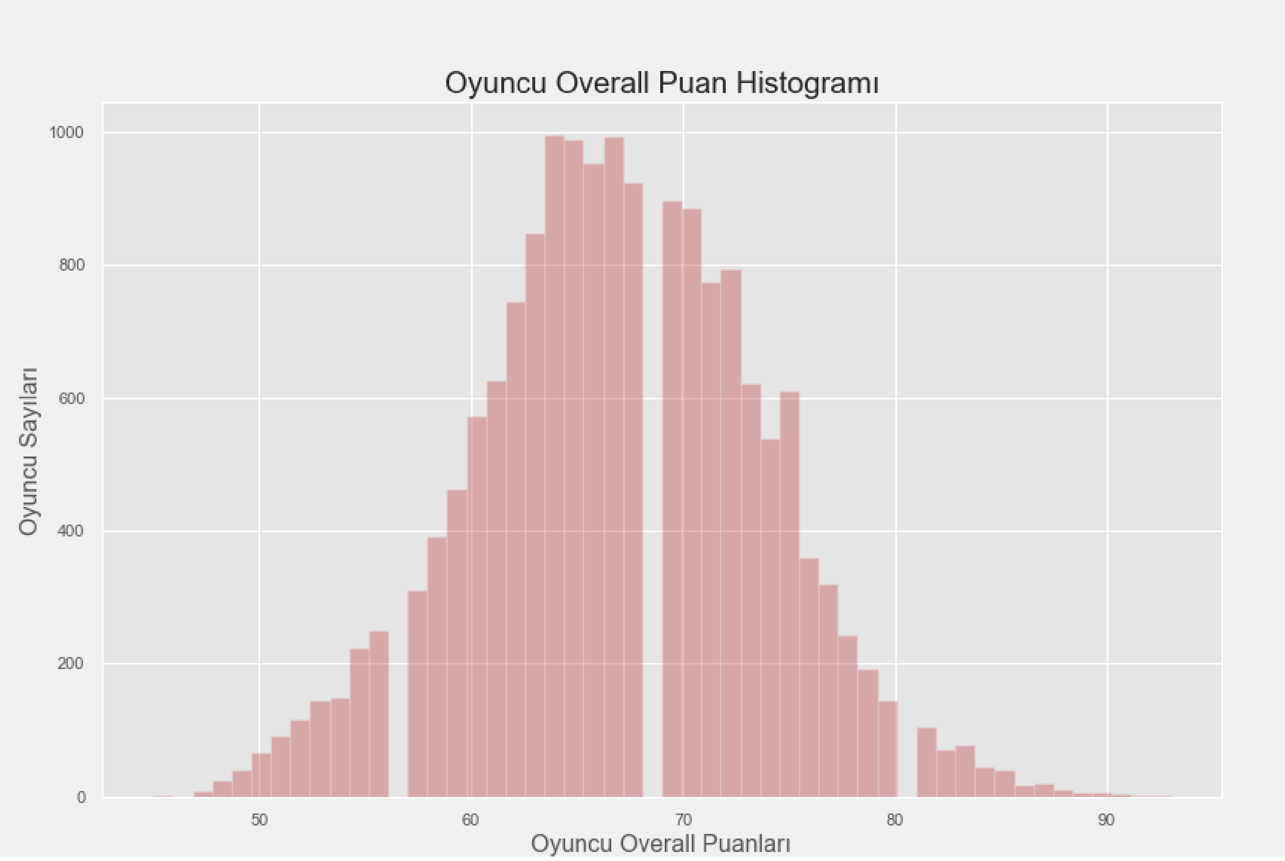

Overall Değerlerine Göre Oyuncu Sayıları
Oyuncu Sayıları
0
200
400
600
800
1000
45
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
Overall Değerleri

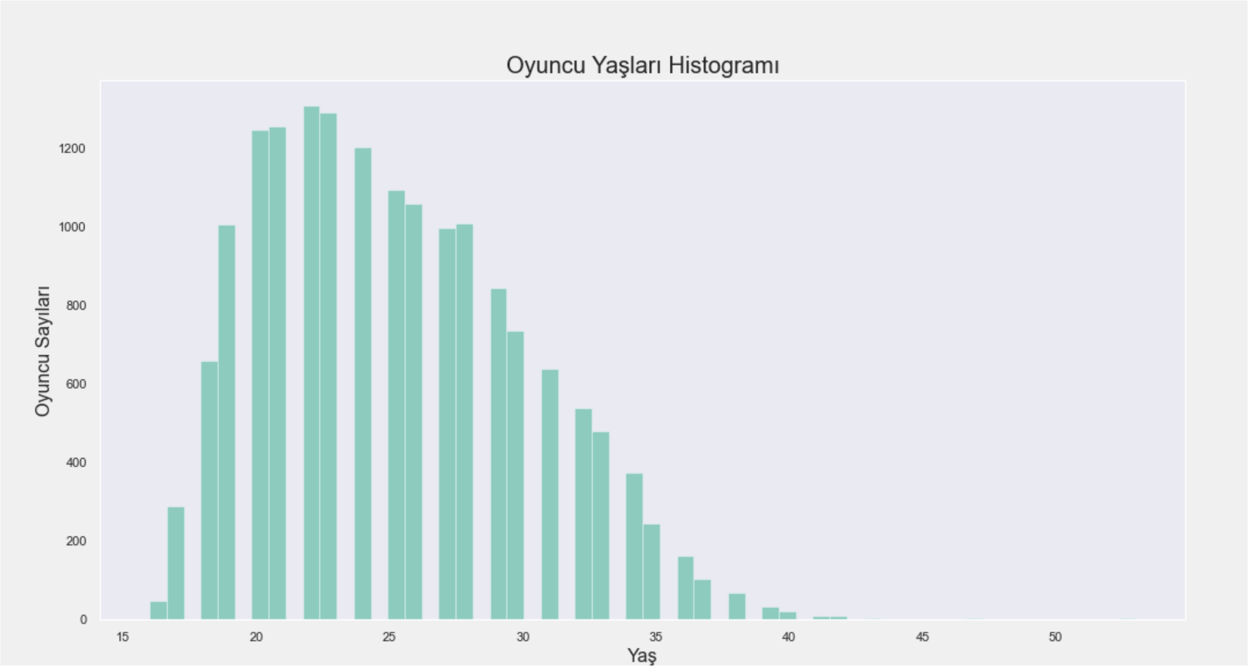

Boylarına Göre Oyuncular
Oyuncu Sayıları
0
500
1000
1500
2000
2500
6'0"
5'10"
5'9"
5'7"
5'8"
5'11"
6'4"
5'6"
6'2"
6'3"
6'6"
6'1"
5'5"
6'5"
6'7"
5'2"
6'8"
5'4"
5'3"
6'9"
5'1"
Boy (inch)

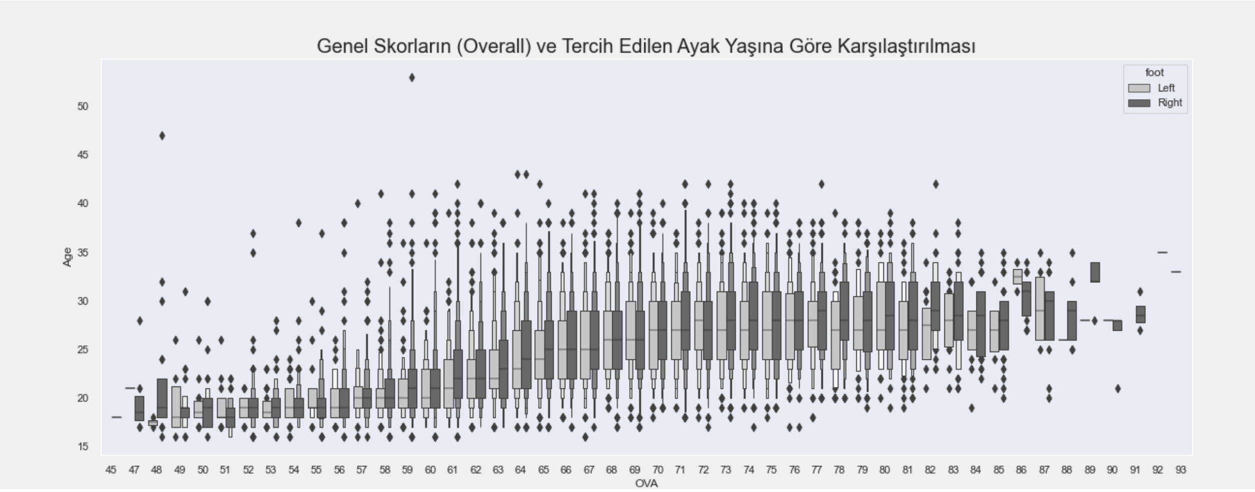

Oyuncu Kiloları
Oyuncu Sayıları
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
100
120
140
160
180
200
220
240

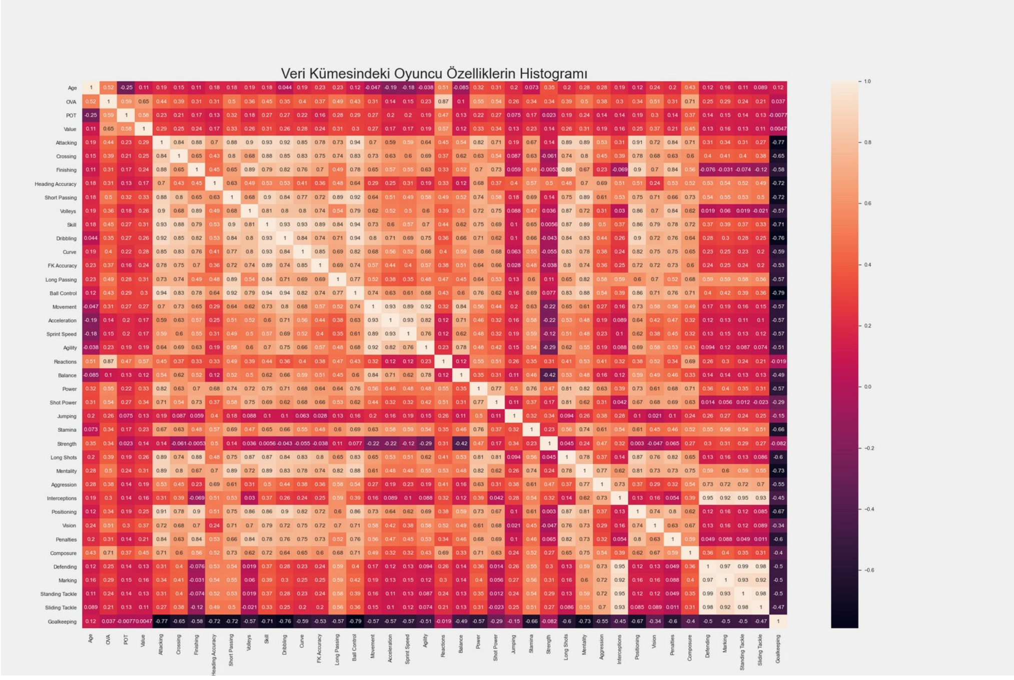

Atak Puanlarına Göre Oyuncular
Oyuncu Sayıları
0
20
40
60
80
100
120
140
Oyuncuların Atak Puanları

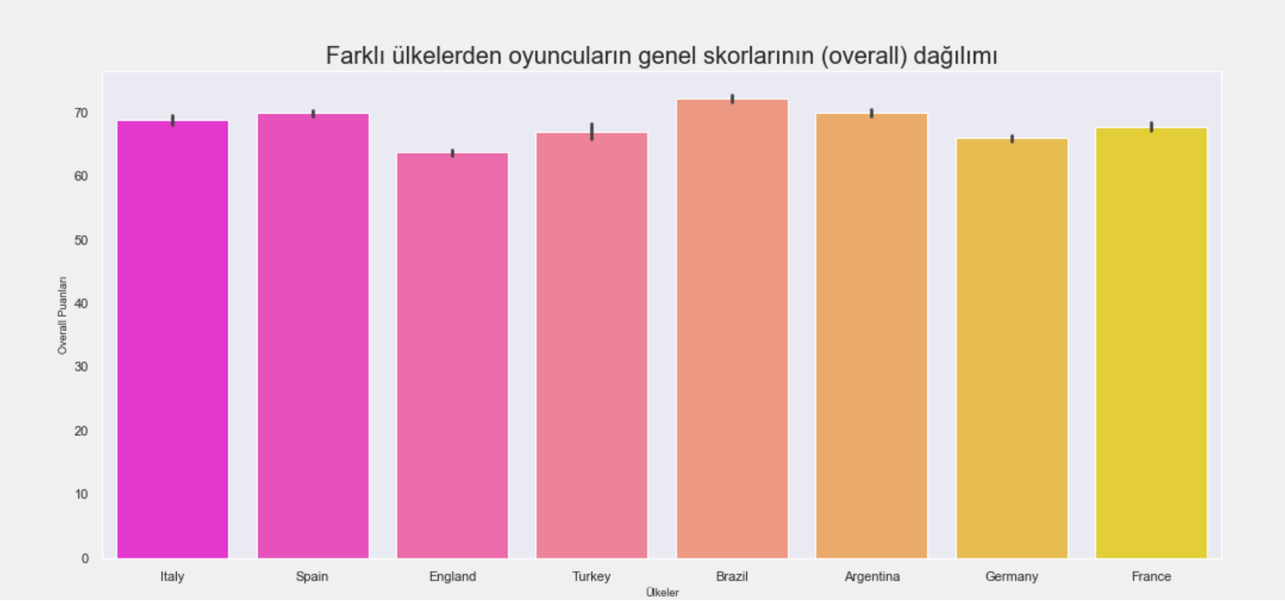

Oyuncuların Potansiyel Puan Histogramı
Oyuncu Sayıları
1200
1000
800
600
400
200
0
50
60
70
80
90
Oyuncuların Potansiyel Puanları

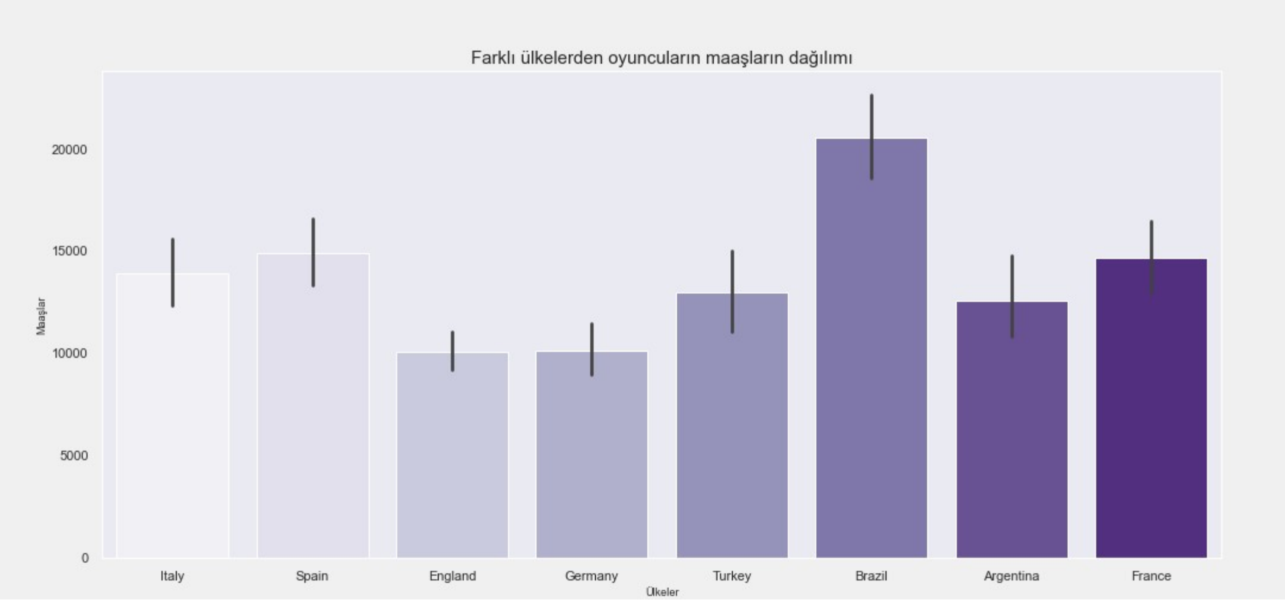

Oyuncu Overall Puan Histogramı
Oyuncu Sayıları
1000
800
600
400
200
0
50
60
70
80
90
Oyuncu Overall Puanları

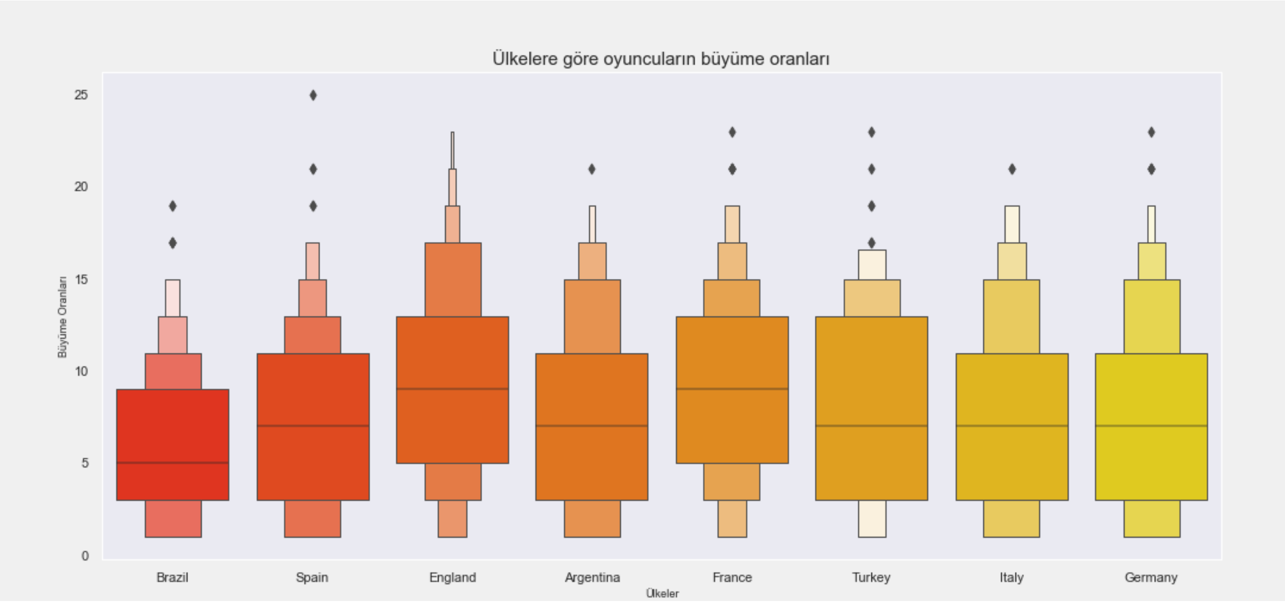

Oyuncu Yaşları Histogramı
Oyuncu Sayıları
0
200
400
600
800
1000
1200
15
20
25
30
35
40
45
50
Yaş

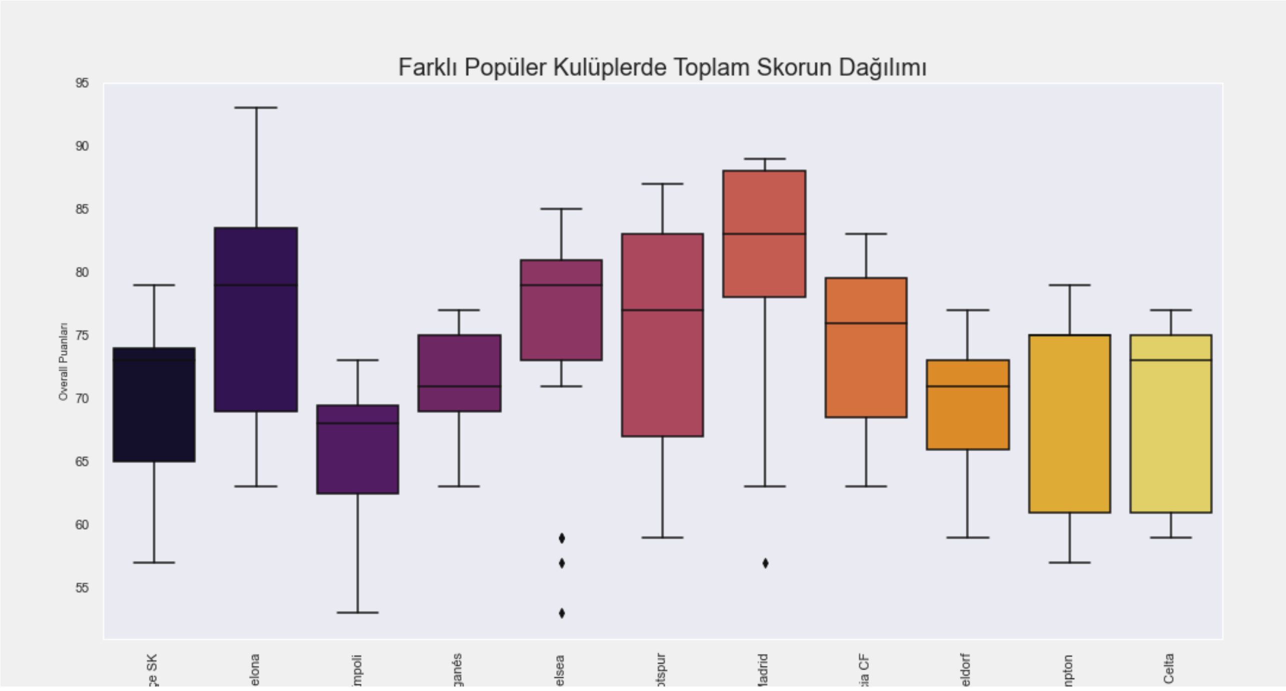

Genel Skorların (Overall) ve Tercih Edilen Ayak Yaşına Göre Karşılaştırılması
foot
Left
Right
Age
15
20
25
30
35
40
45
50
45 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93
OVA

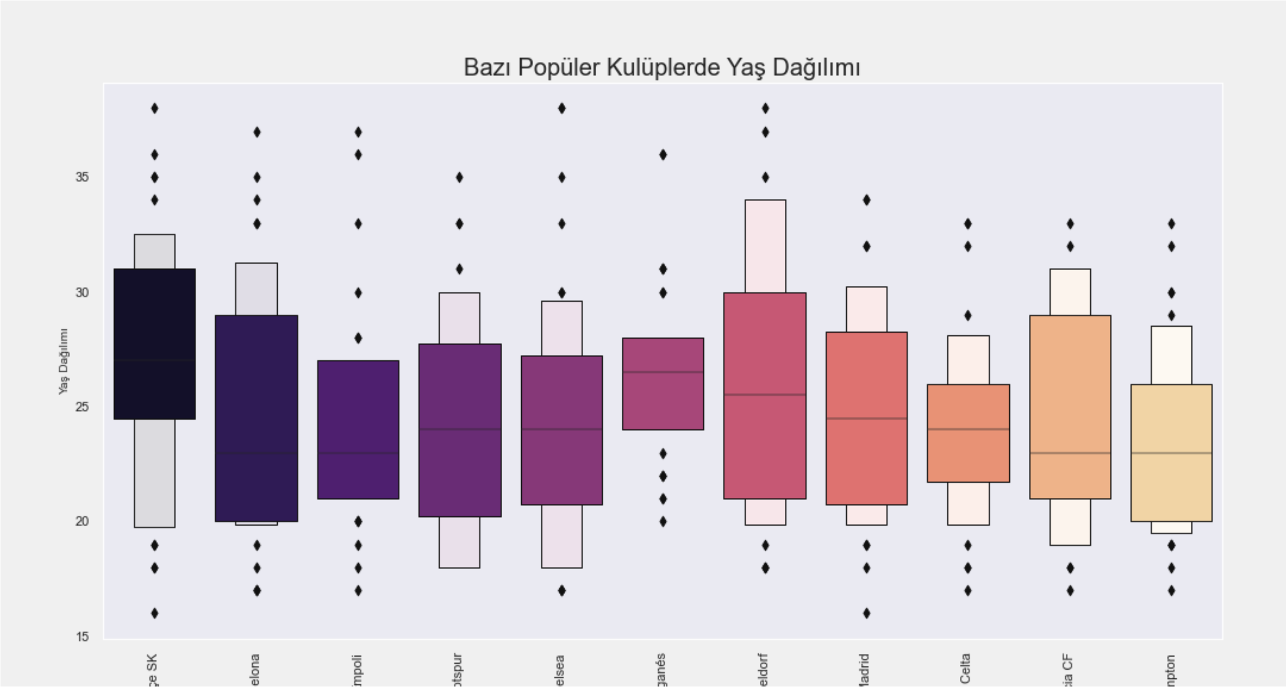

Veri Kümesindeki Oyuncu Özelliklerin Histogramı
Age
OVA
POT
Value
Attacking
Crossing
Finishing
Heading Accuracy
Short Passing
Volleys
Skill
Dribbling
Curve
FK Accuracy
Long Passing
Ball Control
Movement
Acceleration
Sprint Speed
Agility
Reactions
Balance
Power
Shot Power
Jumping
Stamina
Strength
Long Shots
Mentality
Aggression
Interceptions
Positioning
Vision
Penalties
Composure
Defending
Marking
Standing Tackle
Sliding Tackle
Goalkeeping
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

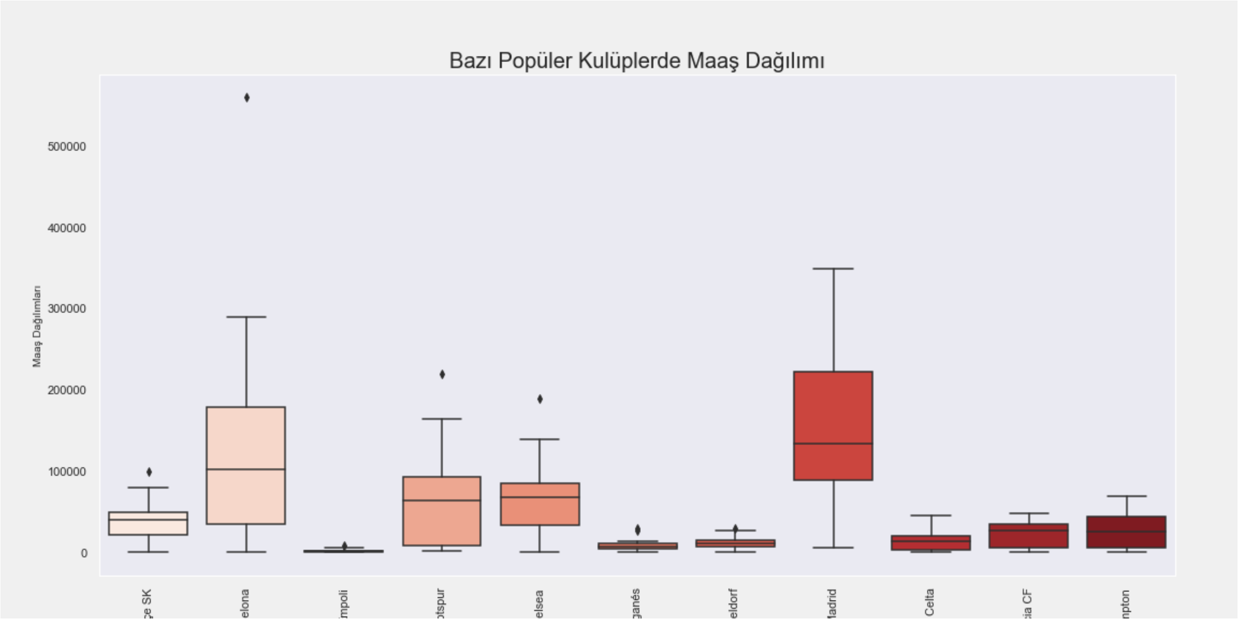

Farklı ülkelerden oyuncuların genel skorlarının (overall) dağılımı
Overall Puanları
0
10
20
30
40
50
60
70
Italy
Spain
England
Turkey
Brazil
Argentina
Germany
France
Ülkeler


Farklı ülkelerden oyuncuların maaşların dağılımı
Maaşlar
0
5000
10000
15000
20000
Italy
Spain
England
Germany
Turkey
Brazil
Argentina
France
Ülkeler

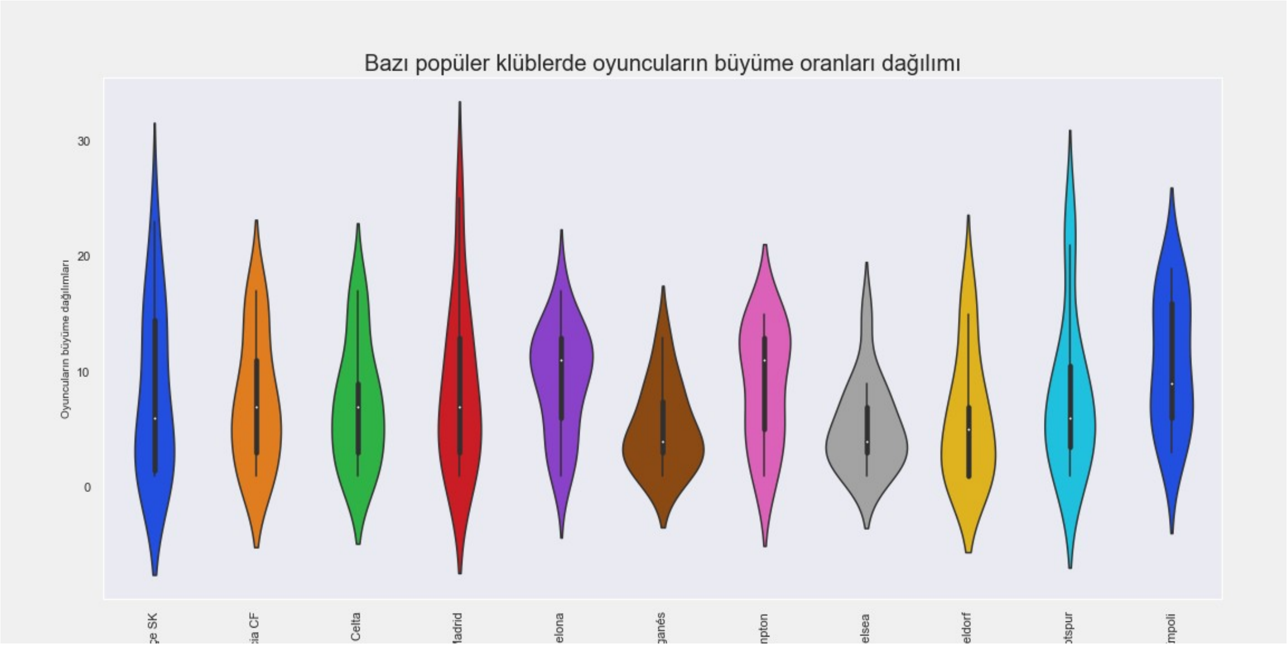

Ülkelere göre oyuncuların büyüme oranları
Büyüme Oranları
0
5
10
15
20
25
Brazil
Spain
England
Argentina
France
Turkey
Italy
Germany
Ülkeler


Farklı Popüler Kulüplerde Toplam Skorun Dağılımı
Overall Puanları
95
90
85
80
75
70
65
60
55
ҫe SK
elona
mpoli
ganés
elsea
otspur
Madrid
ia CF
eldorf
npton
Celta

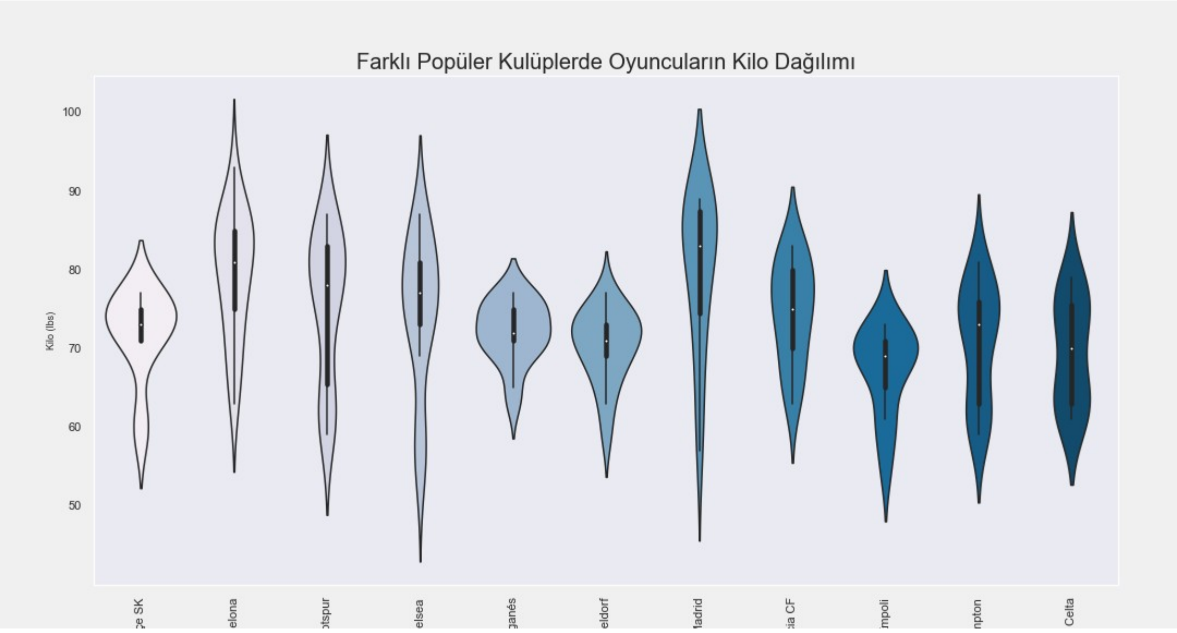

Bazı Popüler Kulüplerde Yaş Dağılımı
Yaş Dağılımı
35
30
25
20
15
ҫe SK
elona
mpoli
otspur
elsea
ganés
eldorf
Madrid
Celta
ia CF
npton

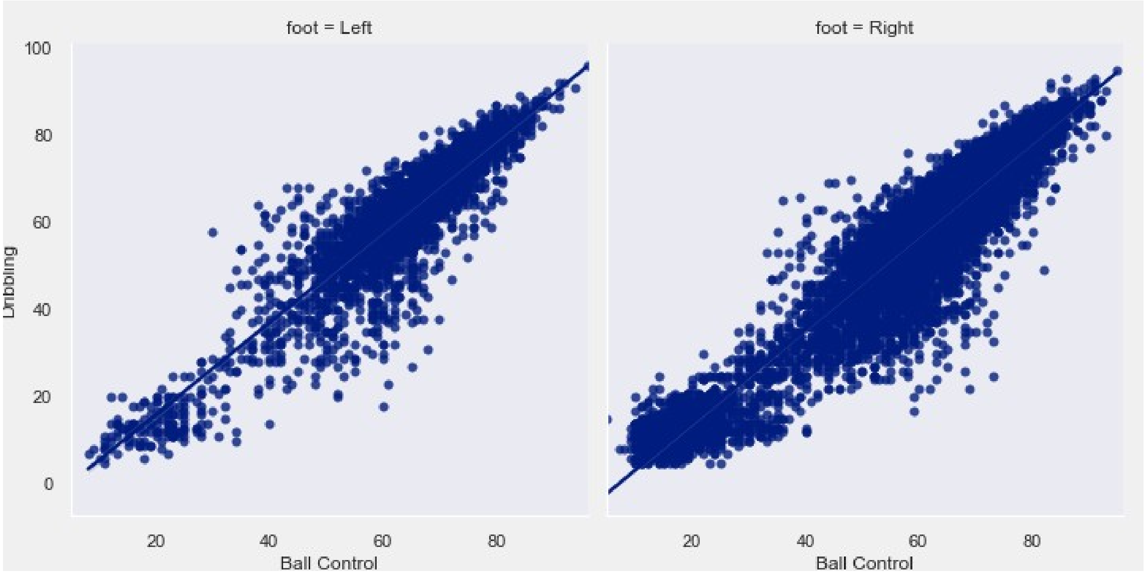


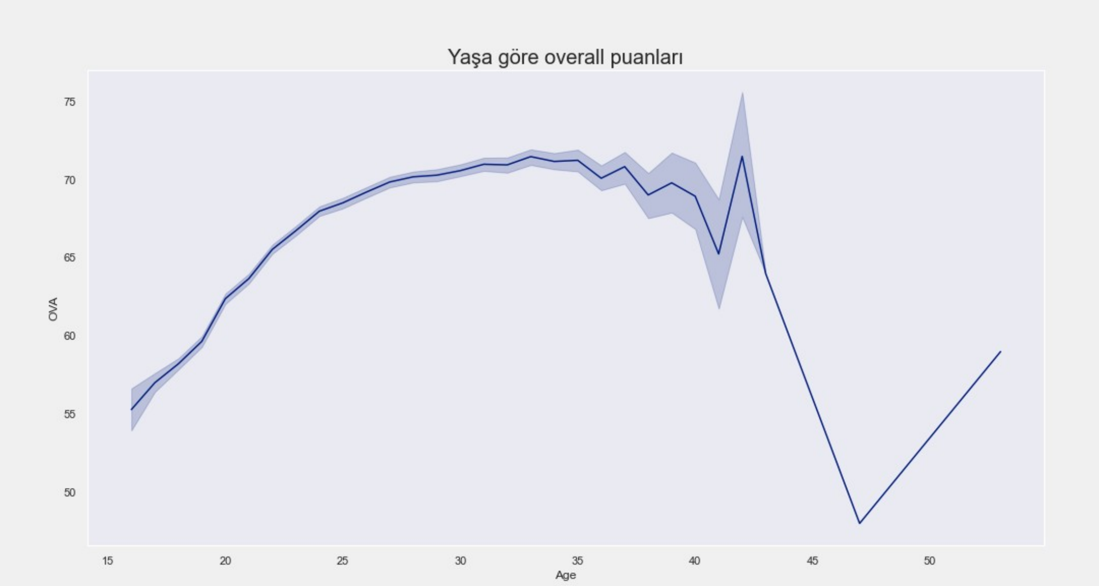
Bazı Popüler Kulüplerde Maaş Dağılımı



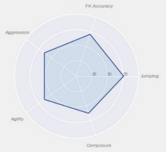
Bazı popüler klüblerde oyuncuların büyüme oranları dağılımı


Farklı Popüler Kulüplerde Oyuncuların Kilo Dağılımı
Kilo (lbs)
100
90
80
70
60
50
ce SK
elona
otspur
elsea
ganés
eldorf
Madrid
cia CF
mpoli
npton
Celta

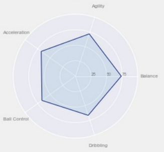

foot = Left
foot = Right
Dribbling
100
80
60
40
20
0
20
40
60
80
Ball Control
20
40
60
80
Ball Control


Yaşa göre overall puanları
OVA
75
70
65
60
55
50
15
20
25
30
35
40
45
50
Age

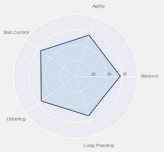
ACB



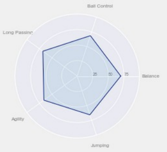
CAM



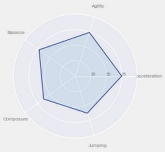
CAM CDM



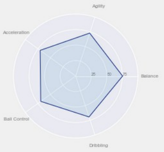
CAM CDM CM



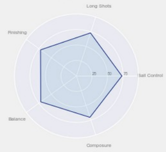
CAM CDM LM



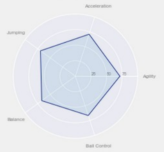
CAM CF



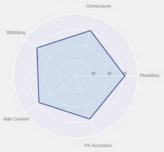
CAM CF CM



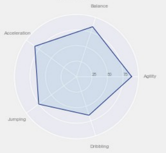
CAM CF LM



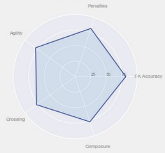
CAM CF LW



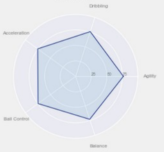
CAM CF RM



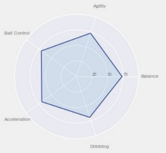
CAM CF RW



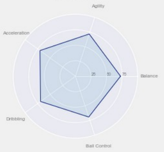
CAM CF ST



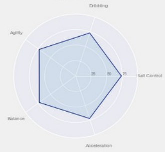
CAM CM



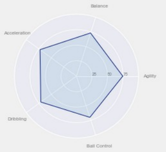
CAM CM CDM



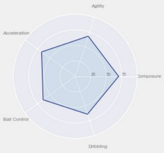
CAM CM CF



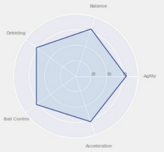
CAM CM LM



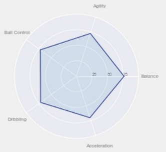
CAM CM LM RM



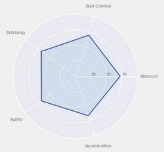
CAM CM LW



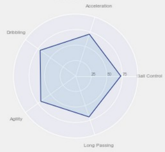
CAM CM RM



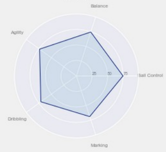
CAM CM RW



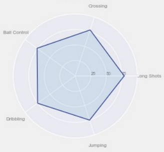
CAM CM ST



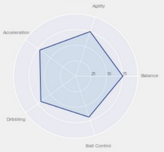
CAM LB



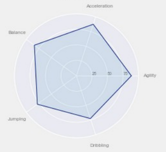
CAM LB LM



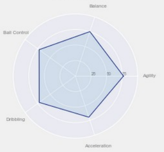
CAM LM



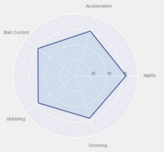
CAM LM CF



CAM LM CM



CAM LM LB


Acceleration
Agility
Crossing
Dribbling
Ball Control