**Task 1**

Prepare a complete data analysis report on the given data

* Many attributes like height, potential etc follows ‘**Normal Distribution**’
* Top 10 players who tend to ‘**get paid the most’** by their countries



* Top 10 countries which **monthly spends** most of their money in euro

|  |  |
| --- | --- |
| **Countries** | **Cost in euro** |
| Spain | 16279000 |
| England | 15691000 |
| France | 13802000 |
| Brazil | 13717000 |
| Germany | 12032000 |
| Argentina | 10540000 |
| Italy | 7974000 |
| Portugal | 4686000 |
| Belgium | 4549000 |
| Netherlands | 4479000 |

* **Average Budget** for creating a good team of 11 players is **1233000** euros
* Top 10 countries having **average overall rating of top 11 players**

|  |  |
| --- | --- |
| **Countries** | **Average overall rating of top 11 players** |
| France | 87.18 |
| Spain | 87.18 |
| Brazil | 87.09 |
| Germany | 86.55 |
| Belgium | 86.36 |
| Argentina | 85.82 |
| Italy | 85.00 |
| Portugal | 85.00 |
| England | 84.00 |
| Netherlands | 84.00 |

* Top 10 countries having **average agility rating of top 11 players**

|  |  |
| --- | --- |
| **Countries** | **Average Agility rating of top 11 players** |
| Brazil | 92.82 |
| Italy | 92.36 |
| Argentina | 92.27 |
| Japan | 92.18 |
| England | 92.00 |
| France | 92.00 |
| Colombia | 91.91 |
| Belgium | 91.82 |
| Portugal | 91.73 |
| Spain | 91.45 |

* Top 10 countries having average ‘**BallControl**’ rating of top 11 players

|  |  |
| --- | --- |
| **Countries** | **Average Agility rating of top 11 players** |
| Spain | 88.45 |
| Brazil | 88.45 |
| Argentina | 87.91 |
| France | 87.73 |
| Germany | 87.45 |
| Portugal | 86.36 |
| Italy | 86.36 |
| Belgium | 86.27 |
| England | 84.45 |
| Croatia | 84.45 |

* Top 10 countries having average ‘**Physical**’ rating of top 11 players

|  |  |
| --- | --- |
| **Countries** | **Average Physical rating of top 11 players** |
| Brazil | 85.64 |
| England | 85.27 |
| Spain | 84.91 |
| Senegal | 84.82 |
| France | 84.64 |
| Argentina | 83.82 |
| Netherlands | 83.55 |
| Germany | 83.45 |
| Uruguay | 83.36 |
| Portugal | 83.27 |

* **According to above analysis**
* **we can see that compared to overall rating some attributes like Agility, Ball-control, Physical ratings are in trending**
* **because they are also determining which team can perform good**
* Some Highly correlated attributes which can be eliminated

0 ('value\_eur', 'release\_clause\_eur', 0.98)

1 ('defending\_standing\_tackle', 'defending\_sliding\_tackle', 0.9758)

2 ('goalkeeping\_diving', 'goalkeeping\_reflexes', 0.9739)

3 ('goalkeeping\_diving', 'goalkeeping\_handling', 0.9713)

4 ('goalkeeping\_handling', 'goalkeeping\_reflexes', 0.9712)

5 ('goalkeeping\_diving', 'goalkeeping\_positioning', 0.9712)

6 ('goalkeeping\_positioning', 'goalkeeping\_reflexes', 0.9711)

7 ('goalkeeping\_handling', 'goalkeeping\_positioning', 0.9705)

8 ('goalkeeping\_kicking', 'goalkeeping\_reflexes', 0.9679)

9 ('goalkeeping\_diving', 'goalkeeping\_kicking', 0.9671)

10 ('goalkeeping\_handling', 'goalkeeping\_kicking', 0.9665)

11 ('goalkeeping\_kicking', 'goalkeeping\_positioning', 0.9661)

12 ('mentality\_interceptions', 'defending\_standing\_tackle',0.9453)

13 ('skill\_dribbling', 'skill\_ball\_control', 0.9435)

14 ('mentality\_interceptions', 'defending\_sliding\_tackle', 0.9339)

15 ('movement\_acceleration', 'movement\_sprint\_speed', 0.9237)

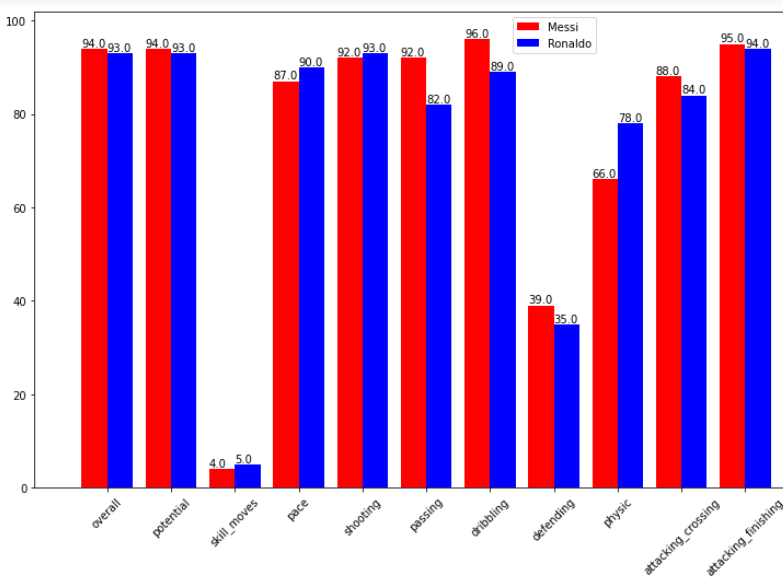
16 ('defending\_marking', 'defending\_standing\_tackle', 0.9186)

17 ('attacking\_short\_passing', 'skill\_ball\_control', 0.9177)

18 ('gk\_diving', 'gk\_reflexes', 0.917)

19 ('defending\_marking', 'defending\_sliding\_tackle', 0.9092)

* **Lionel Messi Vs Cristiano Ronaldo**



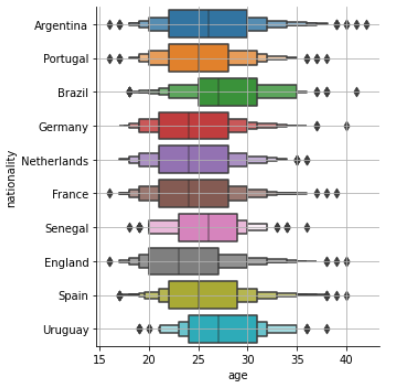
**Overall Points**

* **Lionel Messi ---- 3015.0**
* **Cristiano Ronaldo ---- 2981.0**

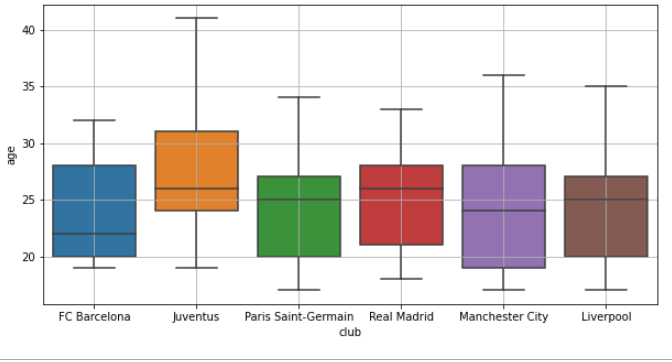
### Top **2% clubs** having **Average Overall rating** of Top **20%** players



* **Age** Distribution of top 10 countries



* **Age** Distribution of top clubs



**Task 2** Explore football skills and cluster football players based on their attributes

-Data is first Normalized using MinMax Scaler

-After scaling, Number 60 columns were reduced to 3 using PCA (principle component analysis)

|  |  |  |  |
| --- | --- | --- | --- |
| **Unsupervised Algorithm** | |  |  | | --- | --- | | Number of clusters | Silhouette Score | |
| 1. Means Algorithm   For Number of cluster=2 we get clear classification | |  |  | | --- | --- | | 2 | 0.611 | | 3 | 0.408 | | 4 | 0.433 | | 5 | 0.413 | | 6 | 0.386 | | 7 | 0.389 | |  |  | |
| Hierarchical Clustering Ward Linkage | |  |  | | --- | --- | | 2 | 0.639 | | 3 | 0.474 | | 4 | 0.466 | | 5 | 0.428 | | 6 | 0.384 | | 7 | 0.367 | |
| Hierarchical Clustering Complete Linkage | |  |  | | --- | --- | | 2 | 0.325 | | 3 | 0.466 | | 4 | 0.448 | | 5 | 0.412 | | 6 | 0.381 | | 7 | 0.372 | |
| Hierarchical Clustering Single Linkage | |  |  | | --- | --- | | 2 | 0.639 | | 3 | 0.476 | | 4 | 0.328 | | 5 | 0.263 | | 6 | 0.262 | | 7 | 0.242 | |
| Hierarchical Clustering Average Linkage | |  |  | | --- | --- | | 2 | 0.639 | | 3 | 0.465 | | 4 | 0.403 | | 5 | 0.419 | | 6 | 0.399 | | 7 | 0.368 | |

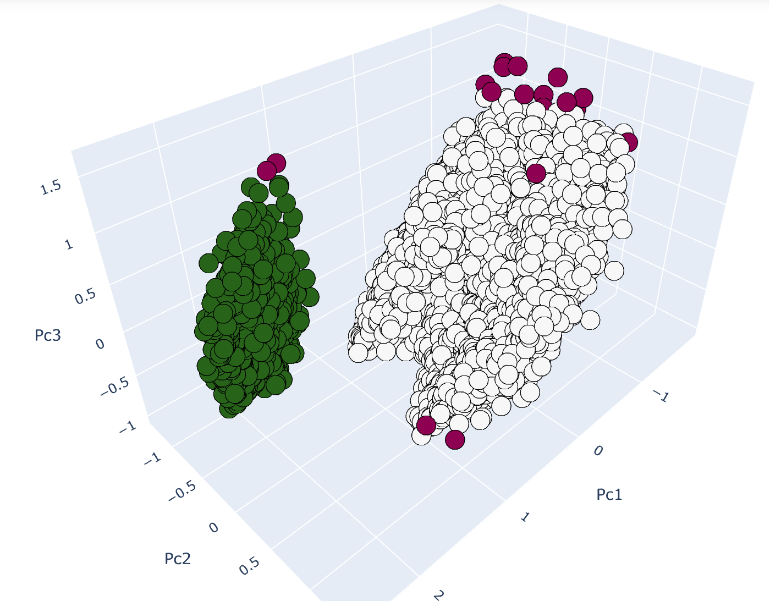
* DBSCAN Result

We don’t need to specify the number clusters in case of DBSCAN algorithm

This clustering algorithm is good for analyzing outliers

eps=0.3, min\_samples=50

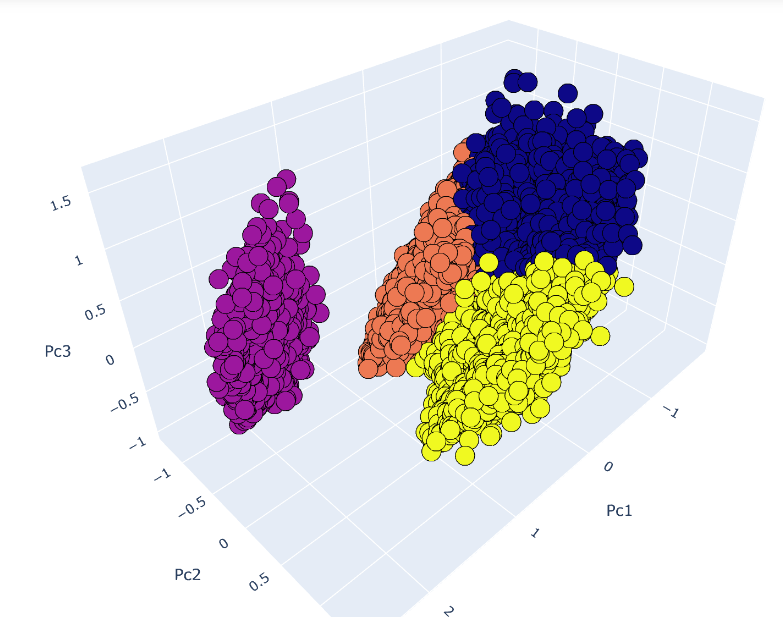
Epsilon is the radius under which min\_samples(minimum data points) come



* KMEANS Result

We need to specify the number clusters in case of K-Means algorithm

Number of clusters = 4



Hierarchical Clustering Result on attributes **wage\_eur, value\_eur**

{'single': {2: 0.956, 3: 0.957, 4: 0.954, 5: 0.954, 6: 0.947, 7: 0.946, 8: 0.935, 9: 0.935},

'ward': {2: 0.904, 3: 0.779, 4: 0.780, 5: 0.655, 6: 0.661, 7: 0.661, 8: 0.661,9: 0.662},

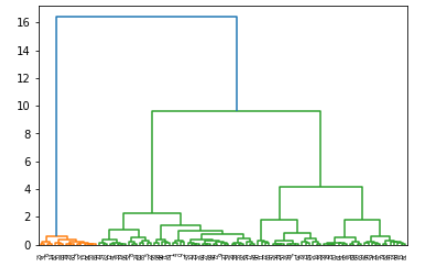
'average': {2: 0.959, 3: 0.925, 4: 0.925, 5: 0.913, 6: 0.913, 7: 0.856, 8: 0.856, 9: 0.855},

'complete': {2: 0.932, 3: 0.930, 4: 0.924, 5: 0.919, 6: 0.878, 7: 0.878, 8: 0.870, 9: 0.870}}

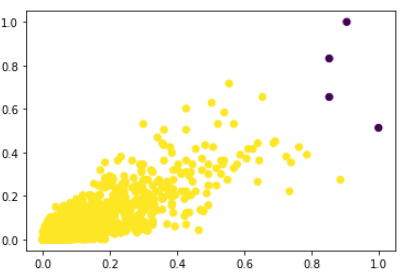
Hierarchical Clustering performs well on salary data

It basically finds in hierarchies

Dendogram plot for finding Number of Clusters

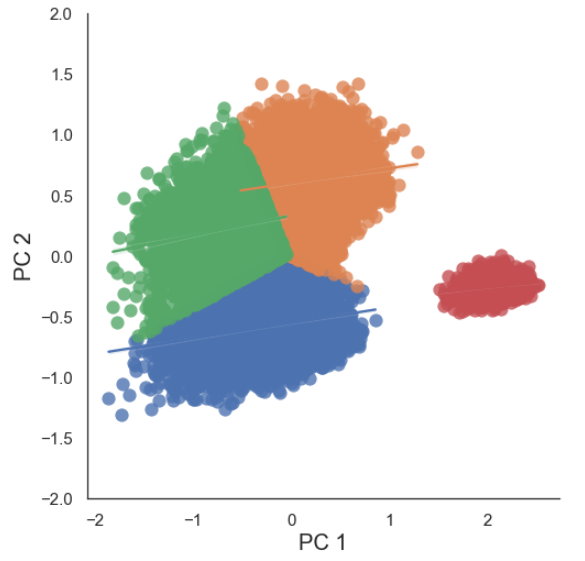


for two clusters average clustering look like



Kmeans Clustering For **k = 4**

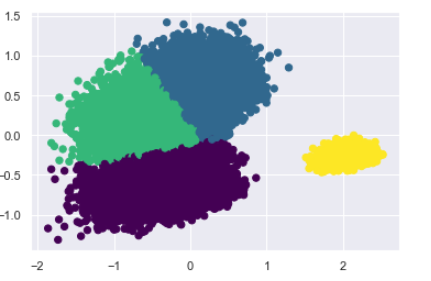
Result :-



**Gaussian Mixture is the advance of Kmeans which uses probabilistic method**

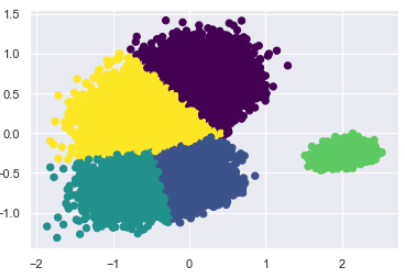
**Gaussian Mixture For Number Of Components = 4 (Covariance Type=Full)**

Result **:-**



**Gaussian Mixture For Number Of Components = 5 (Covariance Type=Full)**

Result **:-**



**Conclusion on Clustering :-**

Kmeans performs good here for number of clusters =2

But we can’t say that 2 is the best we have seen that 4,5 and some 13,19,22 also give good result.

Gaussian Mixture

Same as kmeans only uses probabilistic method to find cluster

DBSCAN clustering is mainly used for clustering outliers

Hierarchical Clustering is used here for clustering players according to their wages

And it gives good result on salary

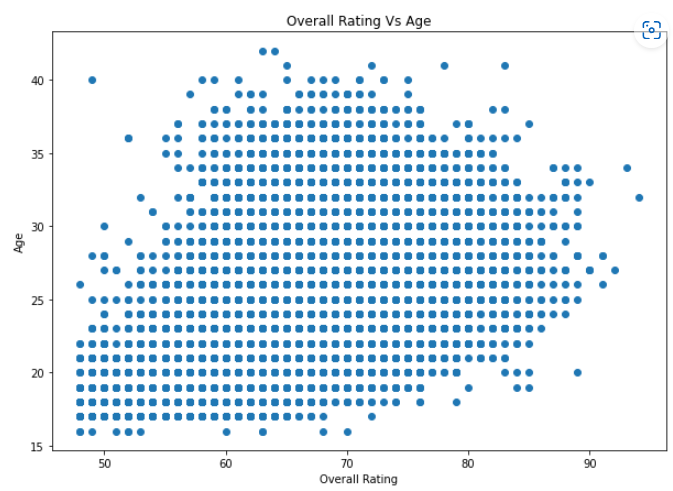
**Task 3**

1. **Prepare a rank ordered list of top 10 countries with most players. Which countries are producing the most footballers that play at this level?**

|  |  |
| --- | --- |
| **Nationality** | **Number of Players** |
| England | 1667 |
| Germany | 1216 |
| Spain | 1036 |
| France | 984 |
| Argentina | 886 |
| Brazil | 824 |
| Italy | 732 |
| Colombia | 591 |
| Japan | 453 |
| Netherlands | 416 |

**Plot the distribution of overall rating vs. age of players. Interpret what is the age after which a player stops improving?**

**Result -**



Conclusion :-

After age 35, player stop improving

3)**Which type of offensive players tends to get paid the most: the striker, the right-winger, or the left-winger?**

**Result -**

**Striker and Right winger tend to get paid the most**

**Challenges faced :-**

**Which attribute should be removed , it was a big question for us but after talking to a person who was good at playing football. We were able to eliminate columns which weren’t required .**

**Talking about clustering algorithms it has always been a difficult to choose the best algorithm.**

**Like here in our case Kmeans performed good but it wasn’t able to identify outliers**

**And Here DBSCAN comes in play, it was able to identify outliers.**

**But still** we can’t say which number of clusters is good here some good results are 2,4,5 and some 13,19,22 also give good result.