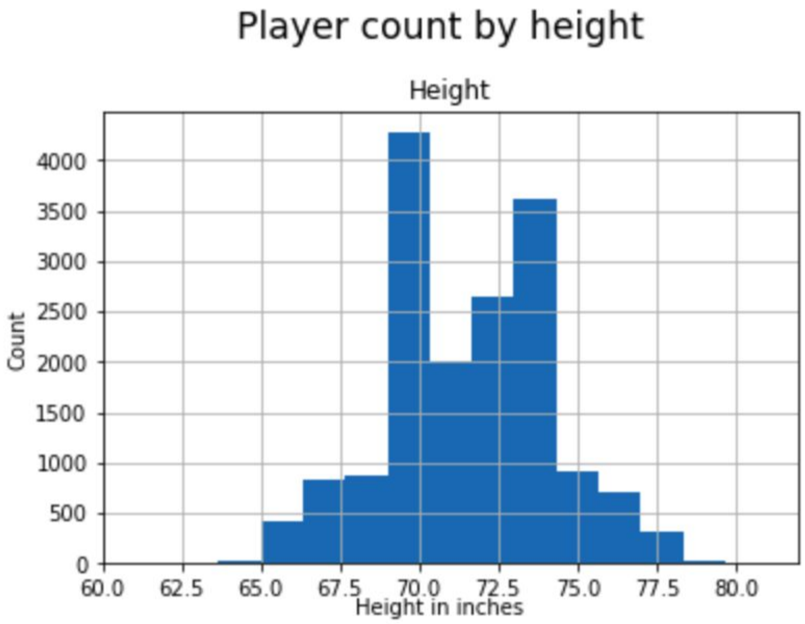
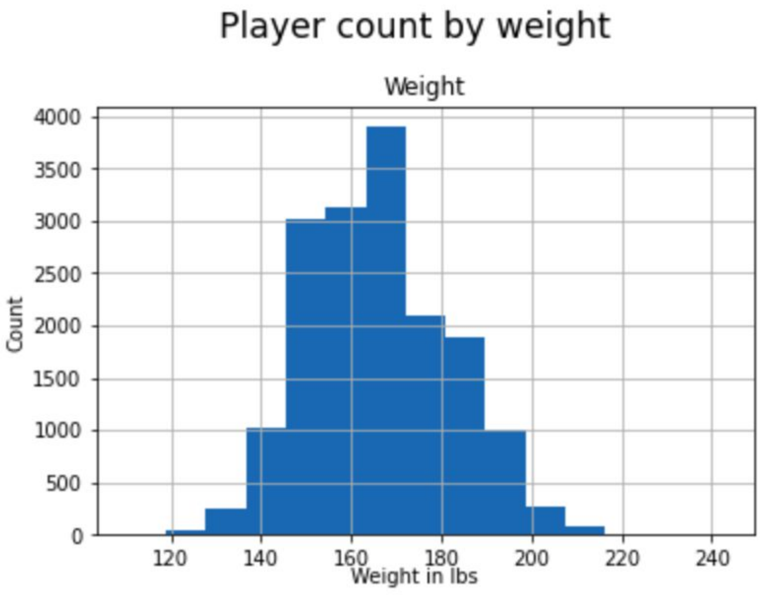
# DATA VISUALIZATION

## Histograms for distribution of players

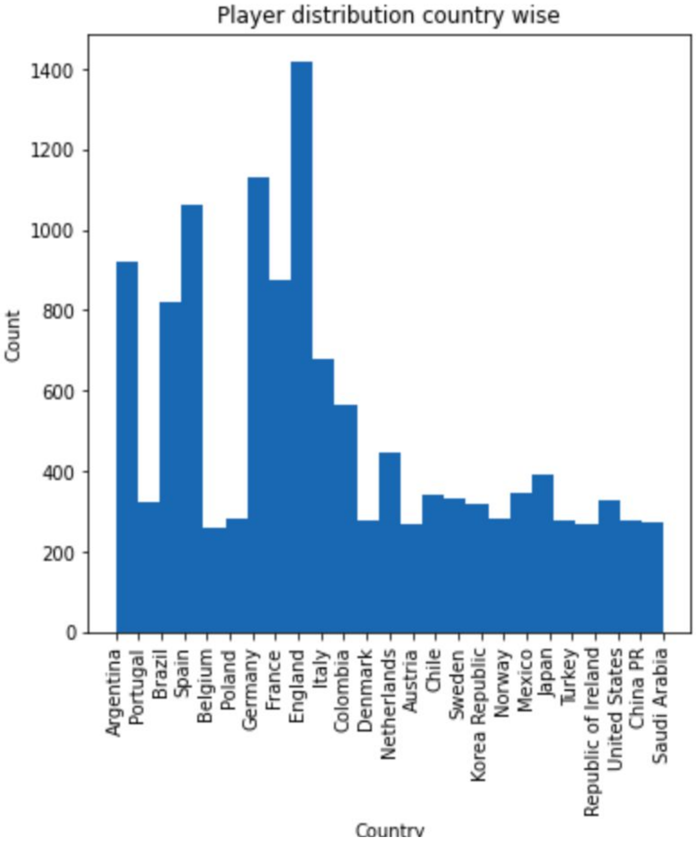
### 1) Player count by height



### 2) Player count by weight

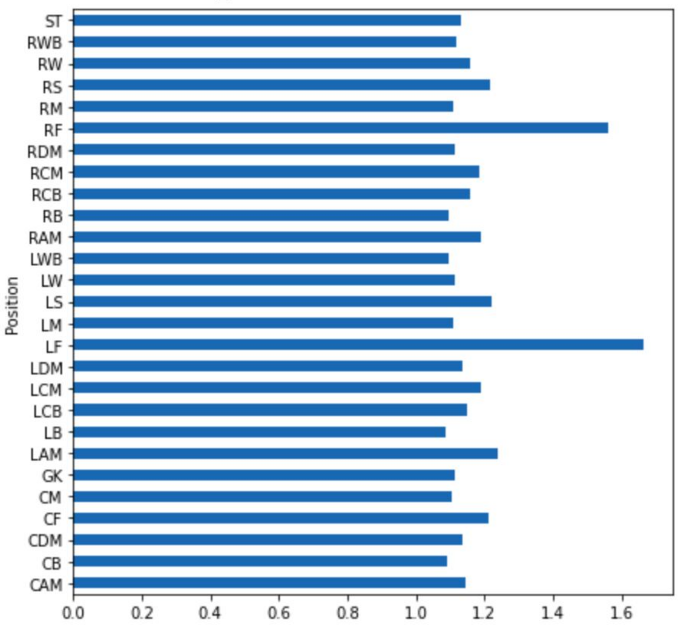


### 3) Player count by country

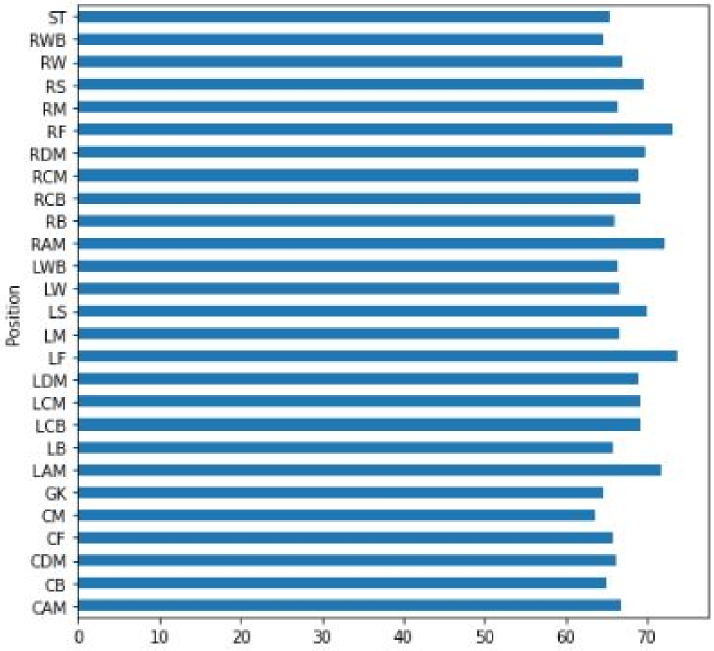


## Features of players according to their position

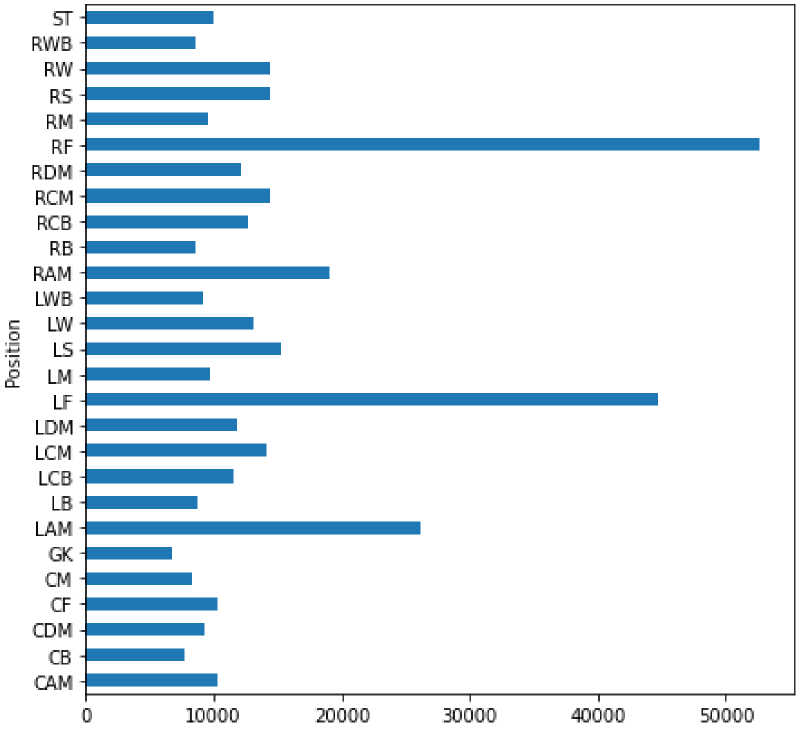
1) **Average "International reputation" by position:** As we can see from the plot, international reputation for LF(Left Forward) and RF(Right Forward) is better than all other positions.



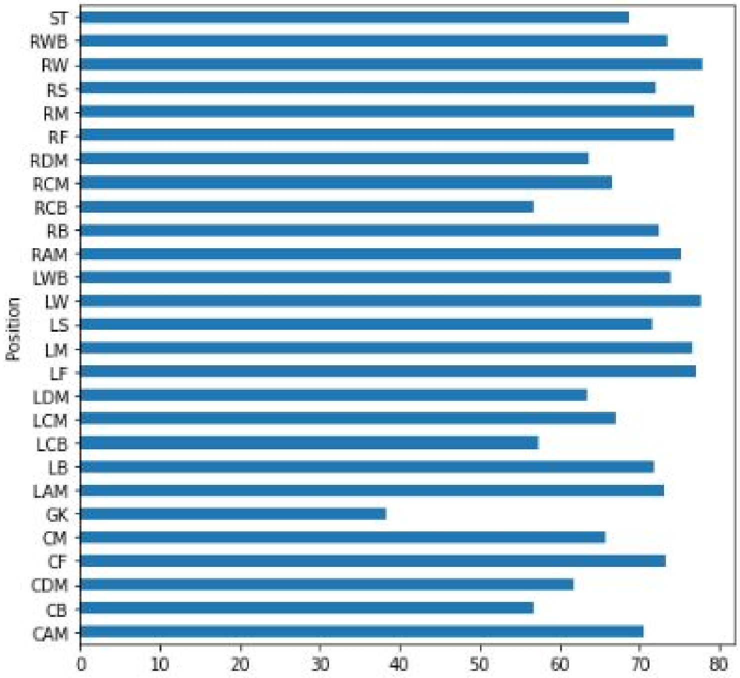
2) **Average "Overall" by position:** Overall for LF(Left Forward), RF(Right Forward), RAM (Right attacking midfield), LAM(Left attacking midfield) is better than other positions.



3) **Average "Wage" by position:** Wage for LF(Left Forward), RF(Right Forward) is much more than rest of the positions. Hence footballers playing at these positions earn the most.

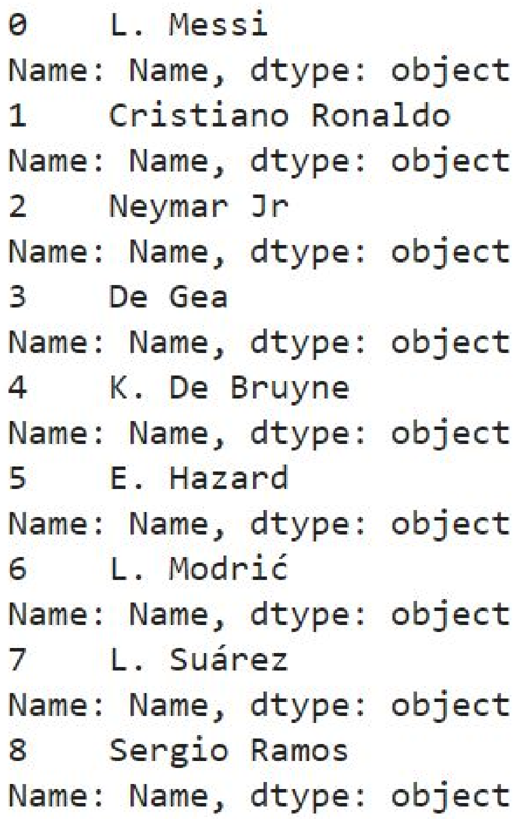


4) **Average "Acceleration" by position:** Acceleration for GK(Goal Keeper) is lowest among all positions, which is expected.



## Finding outliers using Z-score:

The data points which are way too far from zero will be treated as the outliers. Here taking z value = 3.5 for attribute "Overall" , we found below outliers:

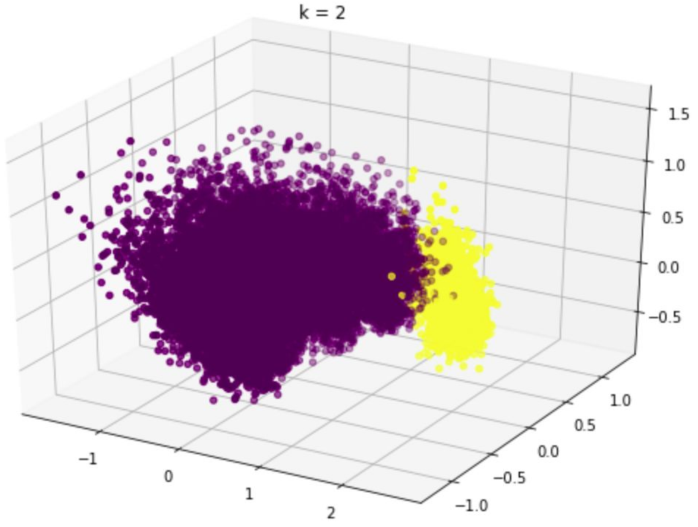


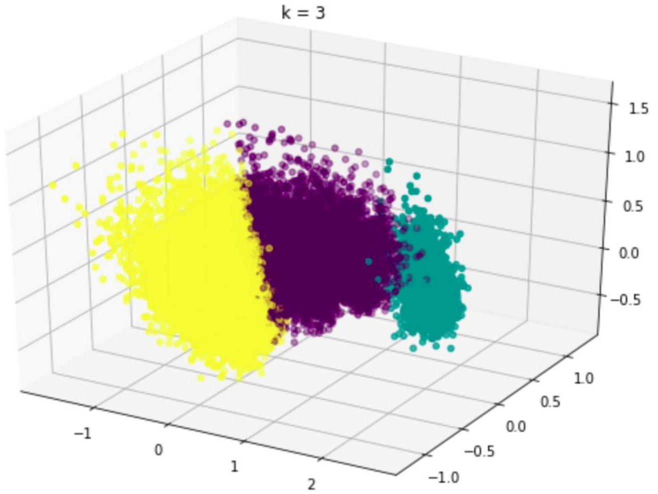
```
0    L. Messi
Name: Name, dtype: object
1    Cristiano Ronaldo
Name: Name, dtype: object
2    Neymar Jr
Name: Name, dtype: object
3    De Gea
Name: Name, dtype: object
4    K. De Bruyne
Name: Name, dtype: object
5    E. Hazard
Name: Name, dtype: object
6    L. Modrić
Name: Name, dtype: object
7    L. Suárez
Name: Name, dtype: object
8    Sergio Ramos
Name: Name, dtype: object
```

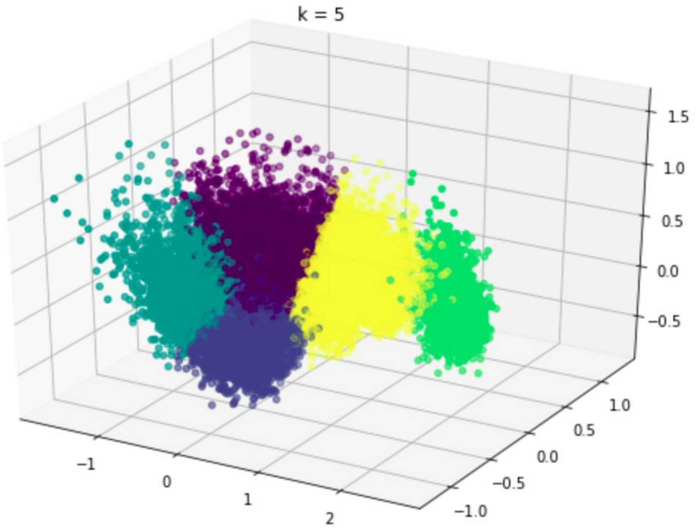
# K MEANS CLUSTERING

Here we apply the k-means clustering algorithms to cluster the data for values of k=2,3,5 and 7. For plotting the data we first apply PCA to reduce the dimension to 3 and plot the points using 3d scatter plots. We can see the clusters formed with different colours in the scatter plot.

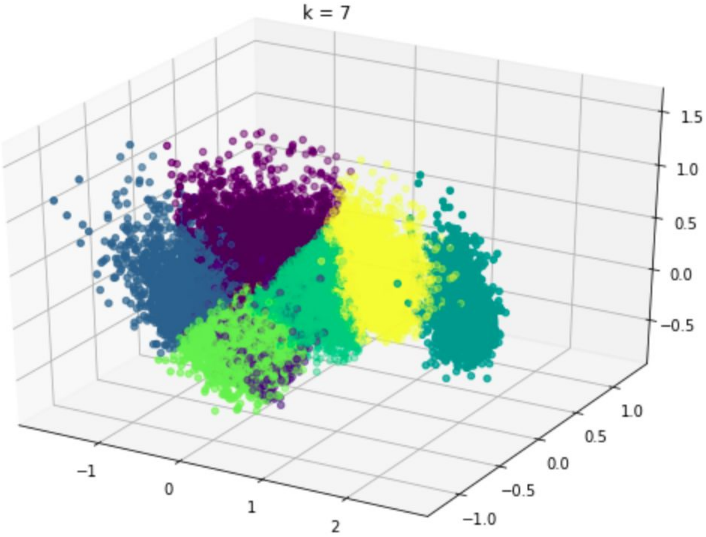
Visualizing clusters for different values of k:







k = 5

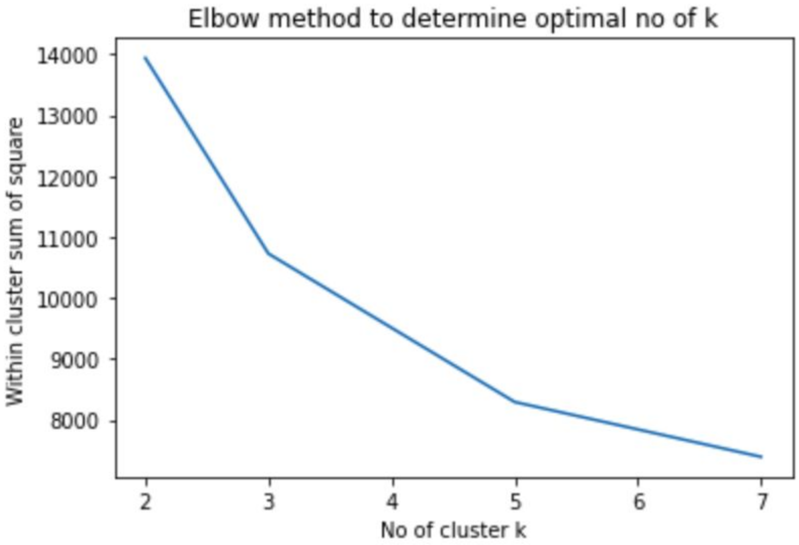


k = 7

**Selecting k(Optimal number of clusters):**

We use these 2 methods to find k in k-Means:

1) Elbow Method: Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes the first to diminish. In the plot of WSS-versus-k, this is visible as an elbow.
   We are getting an elbow at k=5.



2) Silhouette Method: The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).

   For K= 2 we get Silhouette score as 0.5338305201049188
   For K= 3 we get Silhouette score as 0.2539951147339688
   For K= 5 we get Silhouette score as 0.2302291888016746
   For K= 7 we get Silhouette score as 0.19517216383264804

**Evaluating clusters:**

Here we can only use "Internal Evaluation" to evaluate the clusters formed, as there are no ground truth values provided.
We use following methods to assess the quality of clusters formed based on internal criterion:

1) Davies–Bouldin index: For k=5 we get Davies bouldin score as 1.4652018638797104

2) Silhouette coefficient: For K= 5 we get Silhouette score as 0.2302291888016746

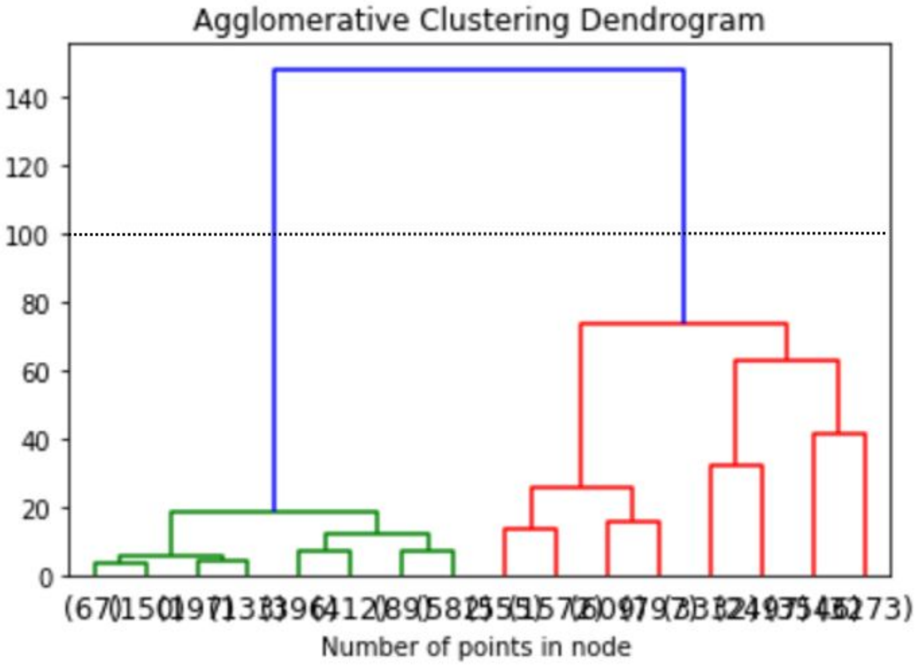
# Agglomerative Clustering

The agglomerative clustering is the most common type of bottom up hierarchical clustering used to cluster the data.

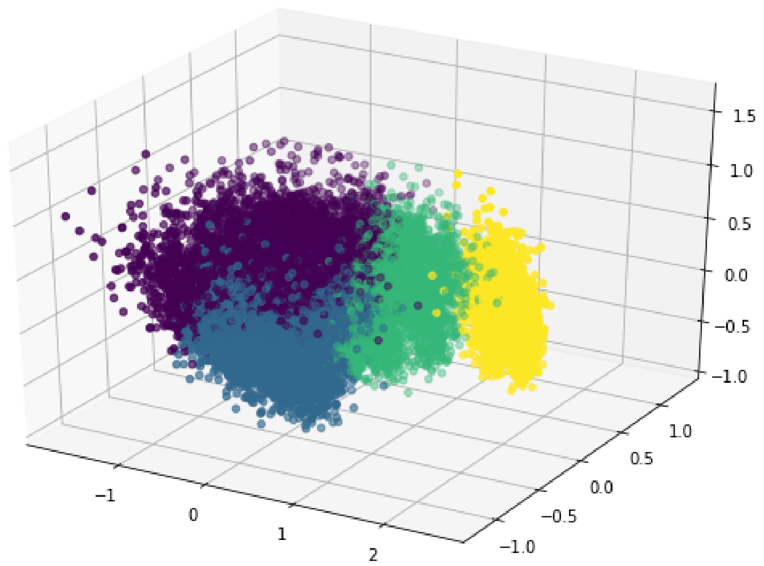
**Selecting no. of clusters:**
Dendrogram is a way to visualize the clusters and then decide the suitable number of clusters. The height of the dendrogram indicates the order in which the clusters were joined.

The longest vertical distance without any horizontal line passing through it is selected and a horizontal line is drawn through it. The number of vertical lines this newly created horizontal line passes is equal to the number of clusters.

Here dotted horizontal line represent the best cut. It cuts 2 vertical lines, hence no. of clusters is 2.



Below is the scatter plot of the clusters obtained using agglomerative clustering. Here we first apply PCA to get 3 features out of all of the features and plot it using scatter plot.



**Evaluating clusters:**

Here we can only use "Internal Evaluation" to evaluate the clusters formed, as there are no ground truth values provided.
We use following methods to assess the quality of clusters formed based on internal criterion:

1) Davies–Bouldin index:0.5338719851481576.

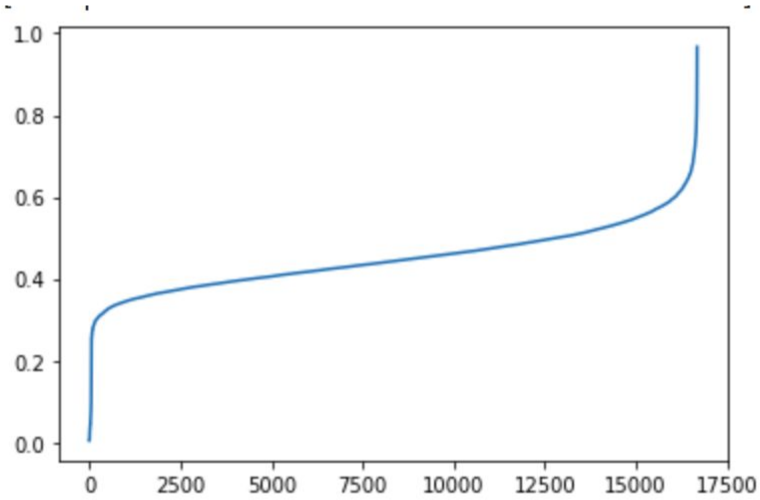
2) Silhouette coefficient:  0.6279093007162538

# DBSCAN

**Selecting epsilon:**

There are various aspects for choosing an epsilon. One way is to find a suitable value for epsilon by calculating the distance to the nearest n points(using knn) for each point, sorting and plotting the results. Then we look to see where the change is most pronounced and select that as epsilon. The optimal value for epsilon will be found at the point of maximum curvature(knee point).

Here we get knee at epsilon = 0.7

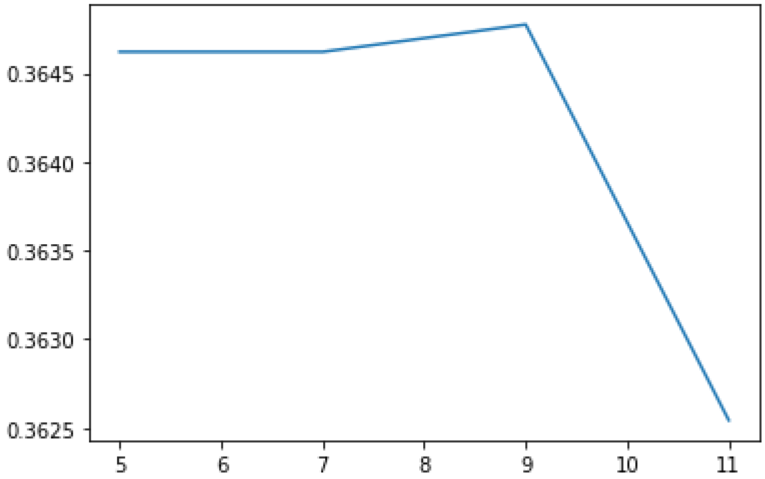


**Selecting minPts:**

There is no general way of choosing minPts, however low minPts means it will build more clusters from outliers, therefore we don't choose a too small value for it.

Here we get the best silhouette score for min points 9.

Plot for silhouette score vs minimum no of points



**Evaluating clusters:**
Here we can only use "Internal Evaluation" to evaluate the clusters formed, as there are no ground truth values provided.
We use following methods to assess the quality of clusters formed based on internal criterion:

1) Davies–Bouldin index: 2.8316199369250348

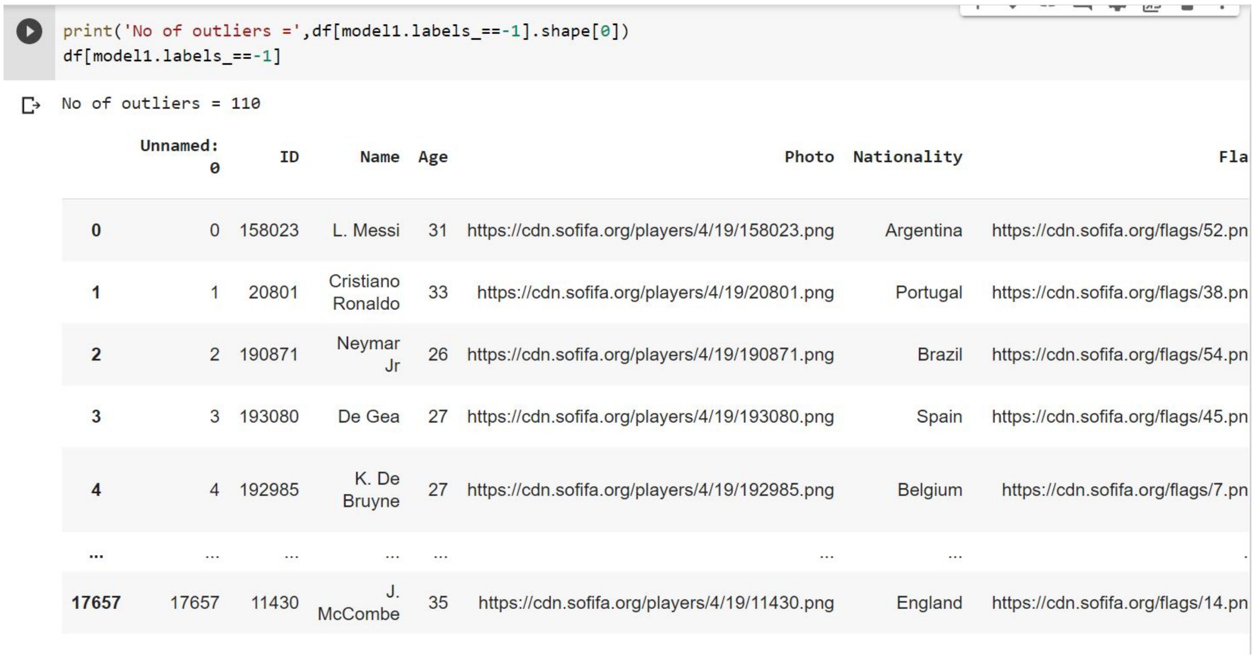
2) Silhouette coefficient: 0.36462021486452323

Below is the scatter plot of the clusters obtained using DBSCAN clustering. Here we first apply PCA to get 3 features out of all of the features and plot it using scatter plot.

# OUTLIERS

Using the DBSCAN algorithm(for epsilon=0.7 and minPts=9), we found 110 outliers( samples labelled as "-1"). Analyzing these outliers, we found that best football players like Messi and Ronaldo belong to the outlier samples. This is expected as stats of these players are much better than rest of the players and hence these players are outliers.

We get the following entries in the dataset as outliers



```
print('No of outliers =',df[model1.labels_==-1].shape[0])
df[model1.labels_==-1]
```

```
No of outliers = 110
```

| | Unnamed: 0 | ID | Name | Age | Photo | Nationality | Fla |
|---|---|---|---|---|---|---|---|
| **0** | 0 | 158023 | L. Messi | 31 | https://cdn.sofifa.org/players/4/19/158023.png | Argentina | https://cdn.sofifa.org/flags/52.pn |
| **1** | 1 | 20801 | Cristiano Ronaldo | 33 | https://cdn.sofifa.org/players/4/19/20801.png | Portugal | https://cdn.sofifa.org/flags/38.pn |
| **2** | 2 | 190871 | Neymar Jr | 26 | https://cdn.sofifa.org/players/4/19/190871.png | Brazil | https://cdn.sofifa.org/flags/54.pn |
| **3** | 3 | 193080 | De Gea | 27 | https://cdn.sofifa.org/players/4/19/193080.png | Spain | https://cdn.sofifa.org/flags/45.pn |
| **4** | 4 | 192985 | K. De Bruyne | 27 | https://cdn.sofifa.org/players/4/19/192985.png | Belgium | https://cdn.sofifa.org/flags/7.pn |
| **...** | ... | ... | ... | ... | ... | ... | . |
| **17657** | 17657 | 11430 | J. McCombe | 35 | https://cdn.sofifa.org/players/4/19/11430.png | England | https://cdn.sofifa.org/flags/14.pn |