A Needle in a Data Haystack - 67978

Football Players Analysis, Similarity and Transfer Benefit Prediction

Group 21

Names:	Email	id	CS id
Omer Jacobi	omer.jacobi@mail.huji.ac.il	301583696	omer.jacobi
Artyom Abramovich	artyom.abramovich@mail.huji.ac.il	324326719	artyomab
Aviv Kotek	aviv.kotek@mail.huji.ac.il	203973490	kotek

Table of Contents

Problem description:

Our project consists of multiple parts:

Data exploration:

Clustering players into groups and insights.

Transfer Benefit Prediction:

Implement a model that given a player, his current team and a team he wants to move to, predict whether the player will improve during the time in his new team, or not. Using this model we can suggest a player which team to move to in order to improve his skills.

Player Similarity:

Imagine a successful player decides to leave his club in order to pursue his dreams elsewhere. He was the perfect player for your club. As a manager you really liked this player, and now your task is to find a suitable substitution. There are thousands of players around the world. How can you know which player should you choose to replace your leaving player? We've thought about an algorithm that given a 2 players p_1, p_2 outputs a grade of how similar they are.

Data used:

We have used 2 datasets from kaggle that were taken from FIFA video games. In the beginning we wanted to use real player data (their stats and achievements during a season), but this data needed crawling which was hard to do. In the end we've decided to use FIFA datasets instead as their creation was inspired by the real game. Two of the datasets contain almost similar features.
Some of the features in the datasets were of special significance for us and we will explain them further in the report. The information about the rest of the features is present in the dataset source links provided below.

FIFA19 Dataset:

Containing FIFA19 players and their attributes.
https://www.kaggle.com/karangadiya/fifa19 The dataset is available in .csv format. It contains X 18207 players and 89 features for every club.

FIFA08-FIFA16 Dataset:

This dataset is in sqlite format and included multiple tables to work with.
https://www.kaggle.com/hugomathien/soccer

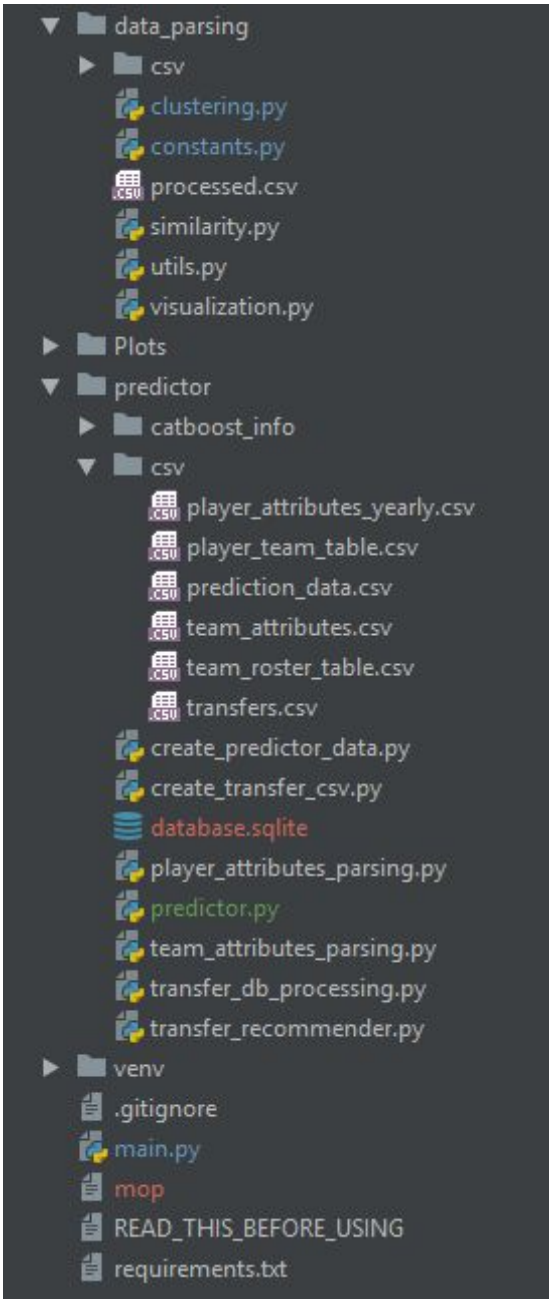
- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers

- Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches

Data pre-processing:

- Some of the features contained lots of missing values. These features were dropped and were not included in our analysis.
- Some of the features were the same for every row in the dataset, therefore they were not useful were removed.
- In FIFA08-FIFA16 dataset, the years 2008 and 2009 were missing from the Team_Attributes table so our analysis used only the data from 2010 - 2016 years.
- Some of the features were categorical, we used dummy variables technique to convert them to numeric values and use the in our analysis.
- Some of the categorical variables were converted to numerical using heuristics.
- To improve our analysis, all the features were normalized to mean=0 and std=1

Project Code



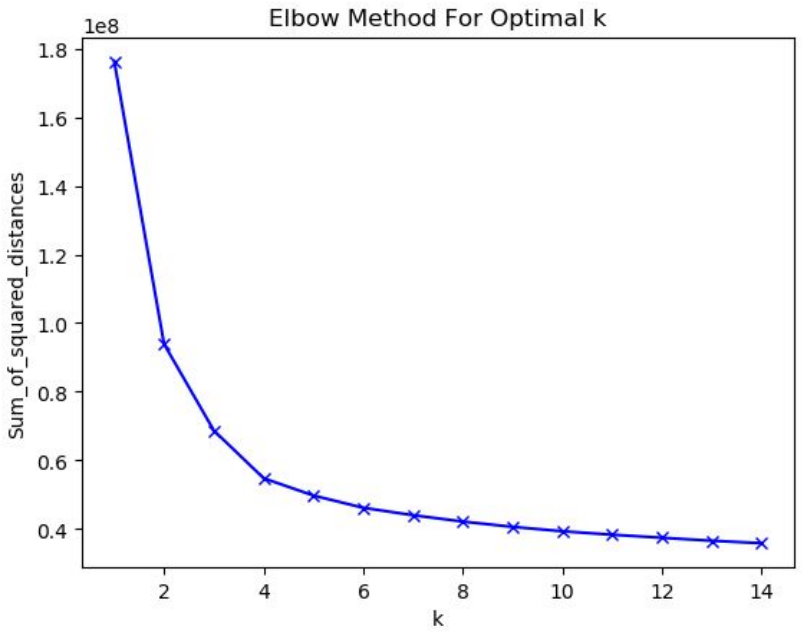
Please visit our repository in the following link. We've provided a detailed information about every file in our project there to keep this report shorter.
https://github.com/akotek/data_final_project

Data Analysis and Clustering:

In football every player learns to play a specific role (position), usually there's one main position that player sticks to throughout his career. In our data there are 27 different positions. In our project we will implement slightly different models for every different kind of player. 27 different groups is too much, we would like to cluster the players into just few clusters. From our experience watching football games there are 4 distinct types of positions: Attackers, defenders and midfielders, and also one more special position, the goalkeeper. So we would expect to see 4 clusters in our data.

Optimal Number of Clusters:

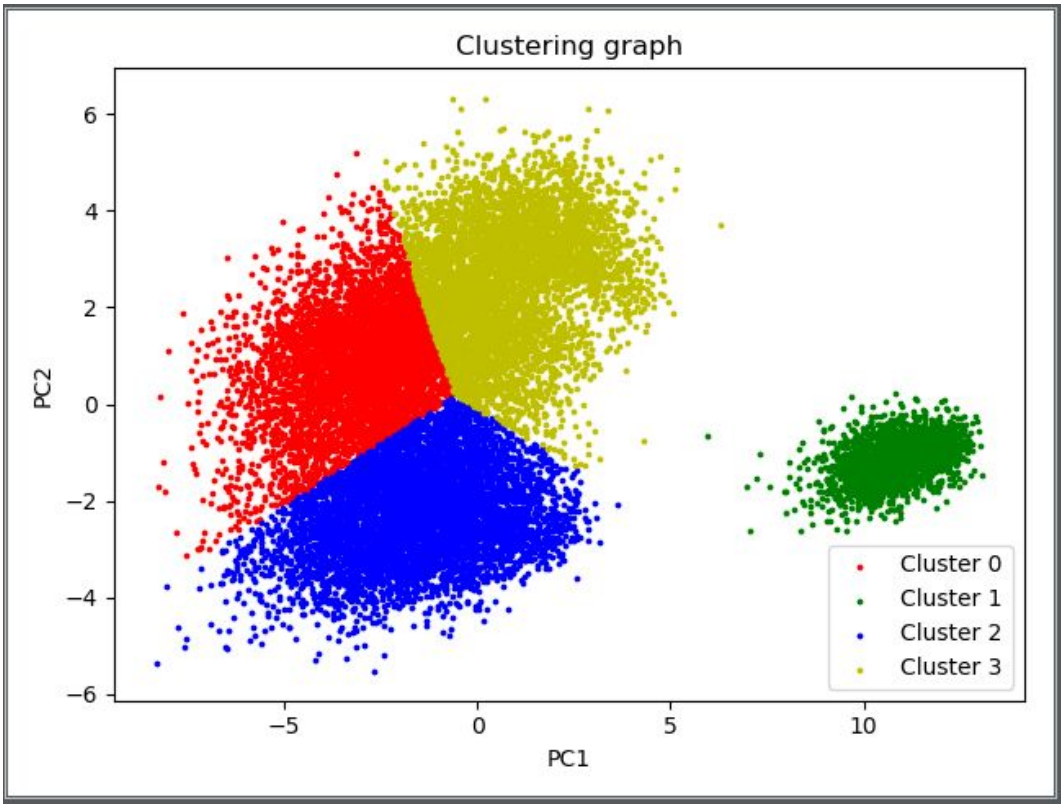
Let's check our theory using the **Elbow Method** for optimal number of clusters:
Elbow method: For each k value, we initialised k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster centre. As k increases, the sum of squared distance tends to zero. Imagine we set k to its maximum value n (where n is number of samples) each sample will form its own cluster meaning sum of squared distances equals zero. Below is a plot of sum of squared distances for k in the range specified above. If the plot looks like an arm, then the elbow on the arm is optimal k:



We can deduce visually that the elbow is located between k=3 and k=4. Which means that our experience watching football games helped us to guess correctly, we will stick with k=4 for next sections.

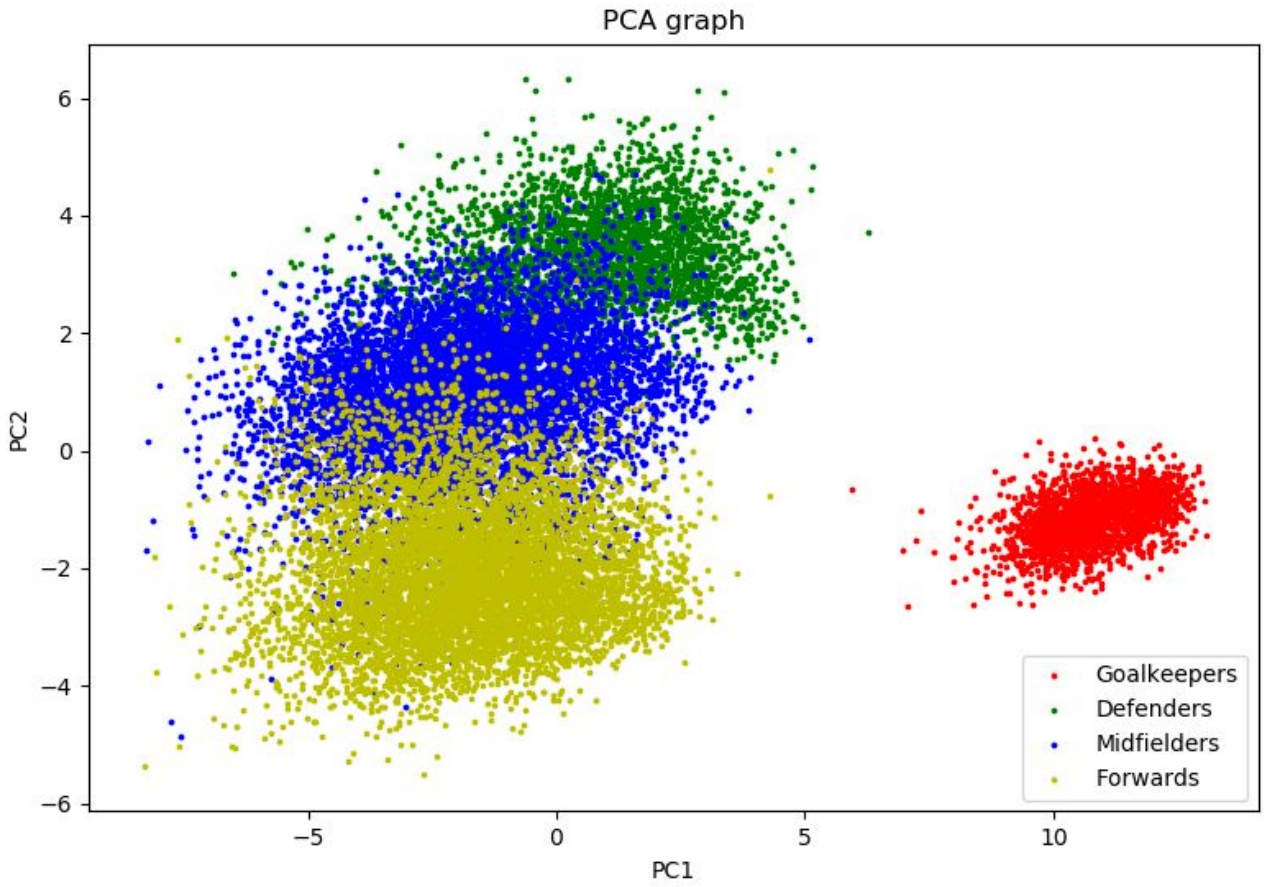
Clustering

Lets cluster our data using k-means algorithm to 4 clusters and see what we get:



We can see a clear separation, looking back to our PCA graph we can see that groups in the clustering are also separated by the logic we have encountered before.
Exploring the given clusters by plotting cluster number histogram for every position we can match the positions to their classes in the following way:
Cluster 0: LB, LCM, LDM, LF, CDM, CM, LWB, RB, RCM, RDM, RWB
Cluster 1: GK
Cluster 2: LCB, CB, RCB
Cluster 3: LAM, CAM, CF, LM, LS, LW, RAM, RF, RM ,RS, RW, ST

To understand the meaning behind the positions' abbreviations please refer here.
And from this we can deduce that:
Cluster 0 = Midfielders, Cluster 1 = Goalkeepers, Cluster 2 = Defenders, Cluster 3 = Forwards.
Using this information, we will use PCA to get the importance of features for every group of players:



As can be seen, this group distribution makes sense even though we can observe outliers id Defenders, Forwards and Goalkeepers groups. We've found that for PC1 the importance of features was distributed pretty much evenly. Reactions, Jumping and Strength features were less important for PC1. Also we've observed a negative

correlation between goalkeeping features and the rest of the features. For PC2 the top4 important features were: SlidingTackle, StandingTackle, Interceptions and Marking.

Transfer Benefit Prediction:

General idea:

If for every **player** transfer from **team_a** to **team_b** we have the attributes of the player and attributes of teams **team_a** and **team_b**. We call it **player_vector**, **previous_team_vector** and **next_team_vector** respectively. We can build a vector consisting of all of these attributes together. Also for players that made the transfer we know their attributes after they've played some time in **team_b**, by comparing the attributes of before the transfer to after the transfer we can deduce if his skill improved. Specifically we've looked on the 'overall_rating' feature of the player. If the 'overall_rating' value is higher after the transfer, then the transfer was beneficial for the player.

Getting the transfers:

Using FIFA08-FIFA16 dataset we managed to get the transfers dataset which is saved in transfer.csv file. This is how we built it:

1. Because in 'Player_Attributes' table there was no 'Club' feature, we had to find out their club using the 'Matches' table. We've iterated over all the matches and got the 'Club' feature of every player as well the dates he was playing in these clubs. We've saved this data in 'player_team_table.csv'.
2. Using 'player_team_table.csv' we found all the transfers that this player made. Using this we've created 'transfer.csv'.
3. Now for every transfer we've added the relevant player attributes and the attributes of the 2 relevant teams. If the player played in one of the teams more than one year before or after the transfer, then we took the average values for every feature values over the years he played in (average of **player_vector**, **previous_team_vector** and **next_team_vector** over the years). The true label of every row was computed in the following way:

   If 'overall_rating' after the transfer is greater than before the transfer, label=1, else label=0.

   Doing this we've created 'prediction_data.csv'.

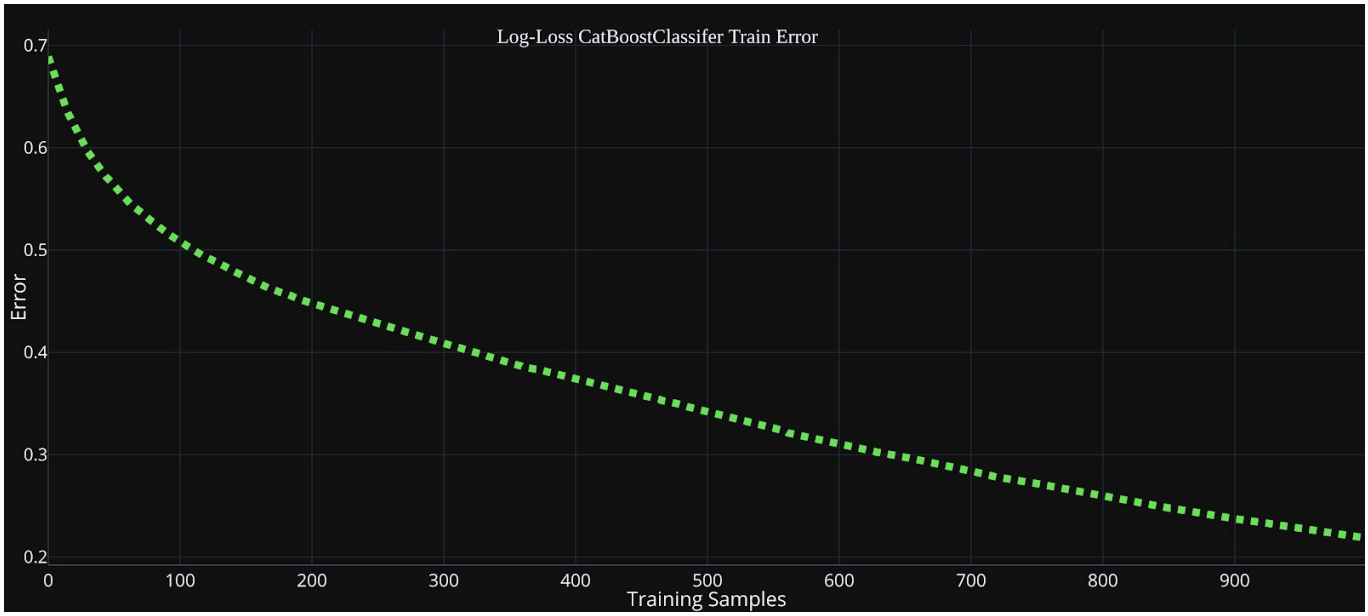
Analyzing ‘predicition_data.csv’:

- In total we’ve got 1363 transfers during 2010 - 2016 years. In fact we’ve found much that much more transfers happened during these years, but some of the players had missing information between those years so unfortunately we had to drop these transfers.
- Every transfer has 79 features (before transforming categorical features to dummy variables).
- 63% of the transfers were beneficial for the players involved.

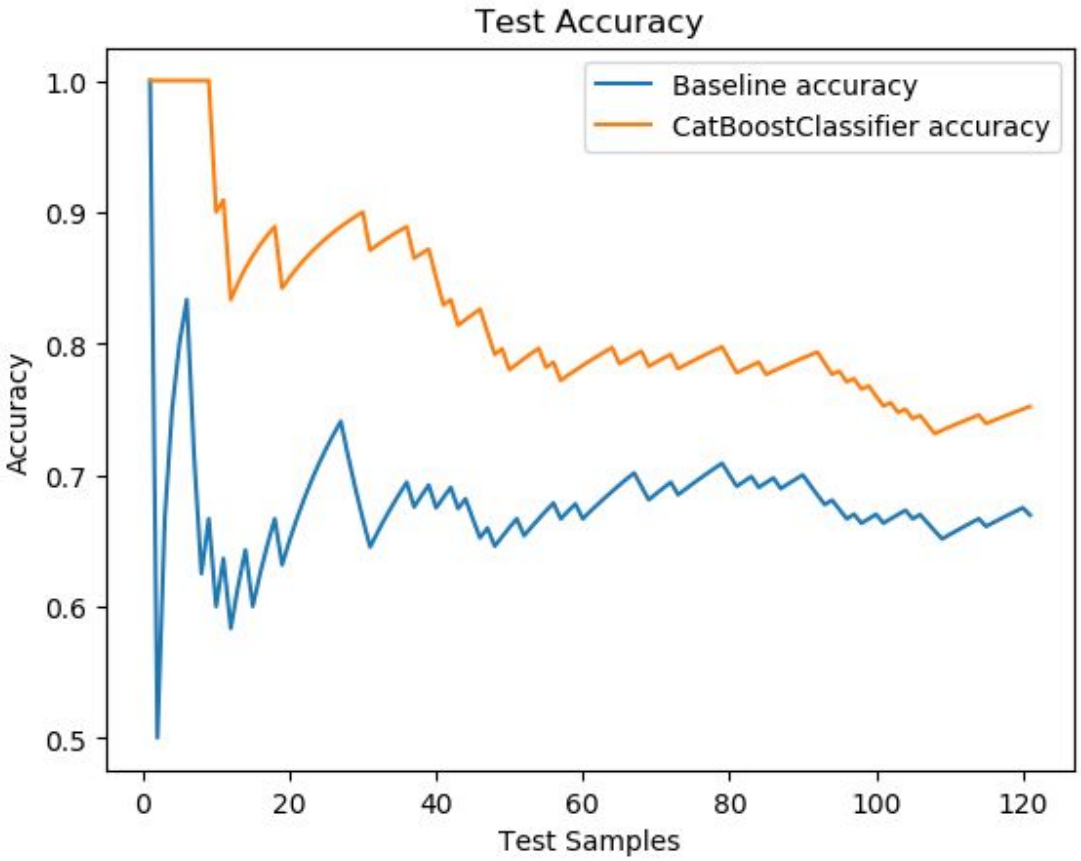
Prediction:

We’ve separated the dataset, 90% for train and 10% for test.
We’ve used a model called CatBoostClassifier for predictions.
To evaluate our results, we’ve compared the performance of our model to a baseline model. Because 63% of the transfers in our dataset are beneficial, we’ve define our baseline model to return 1 for every transfer in the dataset.Our Model Performance:

Training error:



Test error:



As can be seen our model accuracy on test data is 75% vs the baseline that has 66%. It means that our model is not worthless, but in our opinion better results can be achieved. Maybe some other manipulations on transfer data can improve our model.

Feature Importances:

The top 3 ranked features by our model are:

1. Overall_Rating = 10.084, 2. Potential = 4.373, 3. Reactions = 3.523

Overall_Rating and Potential features are directly connected to its future Overall_Rating when the player will be playing in his next team, so it's logical.

Improving Prediction Performance:

We can use our results from the first chapter (clustering). We've seen that there are 4 clusters of playes. That means that some of the features of player attributes might be more relevant for a specific player. So we can try to do feature selection depending on what player we're dealing with (goalkeeper, defender, midfielder or attacker). By feature names, or by analysing our clusters further we can get the most relevant features for every cluster. That means that we can build 4 different models for every group of players. Unfortunately we don't have enough transfer data to make 4 different models to make quality predictions.

Players similarities:

Setup: Data separation and logic:

1. We created vectors for each player, as we separated GK (goalkeepers) from the rest of the players. We created a Vector of 35x dimension

```
['Dribbling', 'Strength', 'Volleys', 'Positioning', 'ShortPassing', 'LongPassing',
 'BallControl', 'HeadingAccuracy', 'Vision', 'Reactions', 'Finishing', 'Aggression',
 'Potential', 'Curve', 'SprintSpeed', 'Composure', 'Marking', 'Stamina', 'Height',
 'FKAccuracy', 'Crossing', 'Acceleration', 'Balance', 'LongShots', 'Penalties', 'Agility',
 'ShotPower', 'Interceptions', 'SlidingTackle', 'StandingTackle', 'Jumping', 'Weak Foot',
 'defensive work rate', 'attacking work rate', 'Skill Moves']
```

   As some of the data was given to us in scale of [1, 100], other was needed to pre-processed (as noted in pre-processing stage).
2. Using the clustering results from the first chapter, we've separated our modules/algorithms to 4x groups.
   [GK, DEFENDERS, MIDFIELDERS, FORWARDS] as we believe each of these groups have similar attributes and should be checked together. We used normal positioning and gathered players to those groups ('if player X position is 'CB' - put him in DEFENDERS group').

```
# ----------------------------------------
# Positions
# ----------------------------------------
# https://gaming.stackexchange.com/questions/167318/what-do-fifa-14-position-acronyms-mean

DEFENDERS = ['CB', 'LCB', 'RCB', 'LB', 'RB', 'LWB', 'RWB', 'LB']

MIDFIELDERS = ['CM', 'LDM', 'LAM', 'RDM', 'RAM', 'CDM', 'CAM', 'LM', 'RM', 'LCM', 'RCM']

FORWARDS = ['ST', 'CF', 'LW', 'RW', 'LS', 'RS', 'LF', 'RF']

GOALKEEPERS = ['GK']
```

We then wanted to check similarities of players, we created an algorithm that for a given player, finds most similar players (from its 'group' of players - gk, df, mid, etc..)

and shows their similarity. The similarity between player p_1, p_2 is $distance\_func(p_1, p_2)$. This will help clubs to find players in the position they need with specific abilities but in a lower price.

Input: 'player_name', 'similarity_function', 'num_of_rec'

Output: 'num_of_rec' most 'similar' players to given 'player_name' with their name and value.

We've tried to use different similarity function: cosine, manhattan and euclidean distance functions After examining the results (next section) we've concluded that cosine similarity worked best for us, which is logical because we're dealing with high dimensional vectors (35).

Example usage: Finding similar players to Cristiano Ronaldo:

```
C:\Python\python.exe C:/Users/avivko/PycharmProjects/data_final_project1/main.py
Starting similarity algorithm....
Enter player/s name you want to compute, spare them by comma
Cristiano Ronaldo
which distance function you want to use: Cosine, Manhattan or Euclidean?
Cosine
you chose Cosine
```

	distance	Name	Selected Player	Release Clause	Overall
190871	0.990439	Neymar Jr	Cristiano Ronaldo	€228.1M	92
153079	0.979589	S. Agüero	Cristiano Ronaldo	€119.3M	89
158023	0.979273	L. Messi	Cristiano Ronaldo	€226.5M	94
183277	0.976002	E. Hazard	Cristiano Ronaldo	€172.1M	91
211110	0.972278	P. Dybala	Cristiano Ronaldo	€153.5M	89
188545	0.965259	R. Lewandowski	Cristiano Ronaldo	€127.1M	90
202126	0.908104	H. Kane	Cristiano Ronaldo	€160.7M	89

This makes sense, as all those players (as we know) play in similar positions, are high rated (overall) and have high forwards abilities.

Then we wanted to check our algorithm results and compare it to 'known' similar players. We used an article that have compared and found similarity of famous international players.

7 of the Most Uniquely Similar Players From Two of the Best Football Clubs on the Planet

By Tushaarsachdeva
31 DEC 2016

477 1

http://www.90min.in/posts/4343003-7-of-the-most-uniquely-similar-players-from-two-of-the-best-football-clubs-on-the-planet

This article presents 7 pairs of players that it claims are similar in the way they play.

Results:

We've compared 4x players in 4x groups (gk, df, mid, forward) and show results of **'cosine'** using **'tag cloud' viz and results** we're almost the same as thought:

Neymar would be most similar, then right after Aguero, Messi, etc..

Same results for defenders - Pique + Sergio ramos

As can be seen in the word clouds, we really get similar players appear together in the word cloud.

For the rest 5 pairs of players in the article we've found similar results.

Impediments:

1. Visualizing the data from a 'data frame' of similar players and their distance, we had troubles thinking how to visualize this. Using Tag clouds as counting frequencies (and not text occurrences), we managed to solve and visualize this issue.
2. Deciding on how to measure similarity- should we measure each defender with all players? Should we check it only with defenders? What about goalkeepers? After testing and playing with data, similarity metrics and other criterias we have found that testing each player with his own group gives better results. We will use this concept of separating to 4x groups in all of the project.
3. We've failed to find a dataset that compares between players, so we could compare our results only to 7 pairs of similar players. This means that we can't

know for sure if our approach works and can rely solely on our experience watching football matches, which too can be biased.

Future Work:

1. The same analysis can be done solely with real world data.
2. Get more transfer data to check our Transfer Benefit Predictor approach with several different models depending on the players' position.
3. Transfer Benefit Predictor may work better if we take into consideration the opposition teams participating in the same league and competitions with the next team the player in question transfers to.
4. During our work we didn't notice major difference between right-handed and left-handed players. But we think that it needs to be analyzed more thoroughly to find an effect for that.

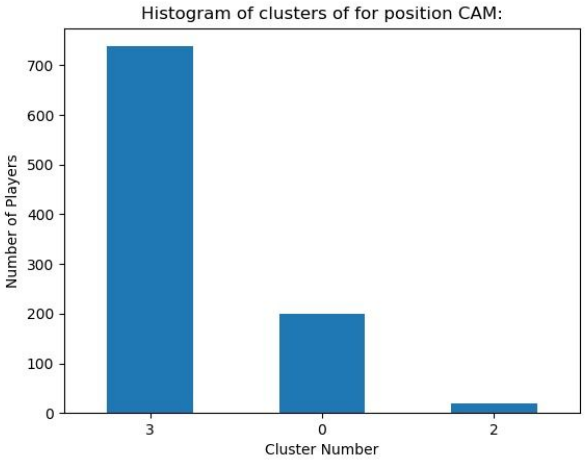
Conclusion:

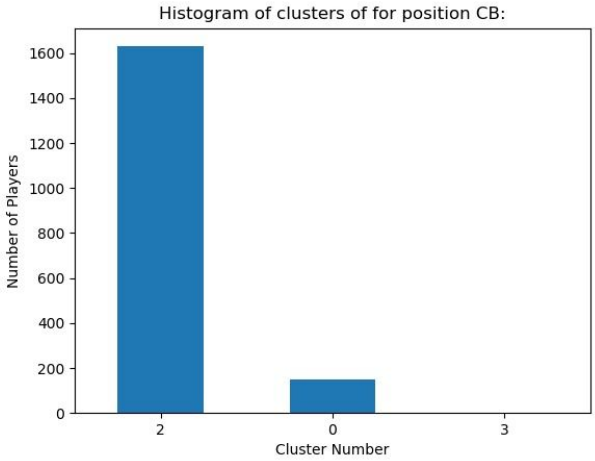
Thanks to this project we can advice every football player that is thinking to agree to a transfer to another club and our advice will be 70%-75% accurate.

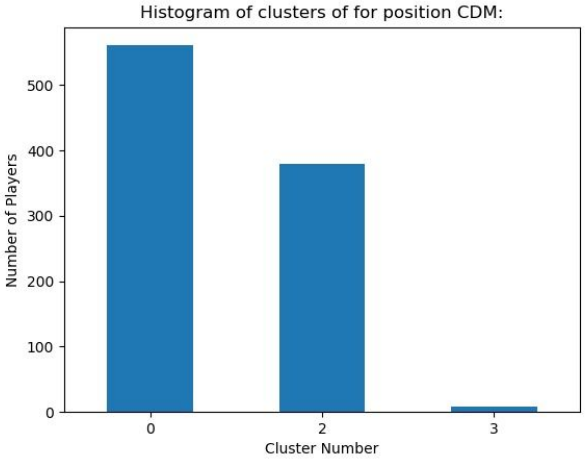
Also we can help a football team manager to find suitable football players to join his team.

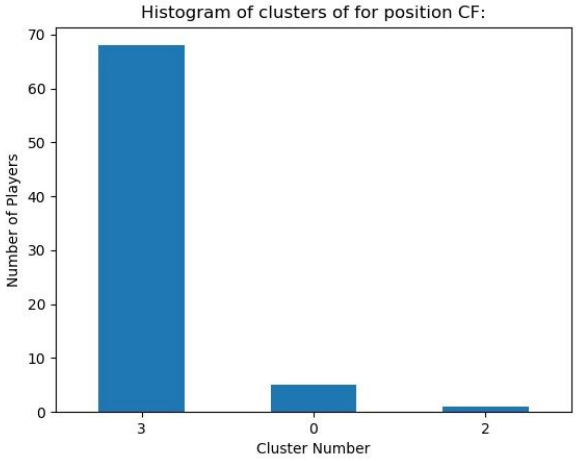
Appendix:

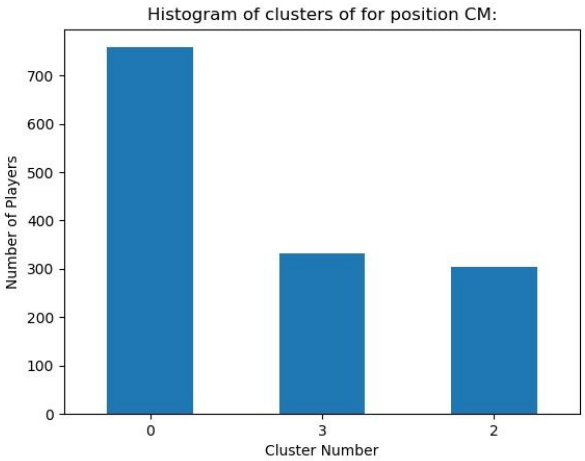
Histogram of clusters by positions:

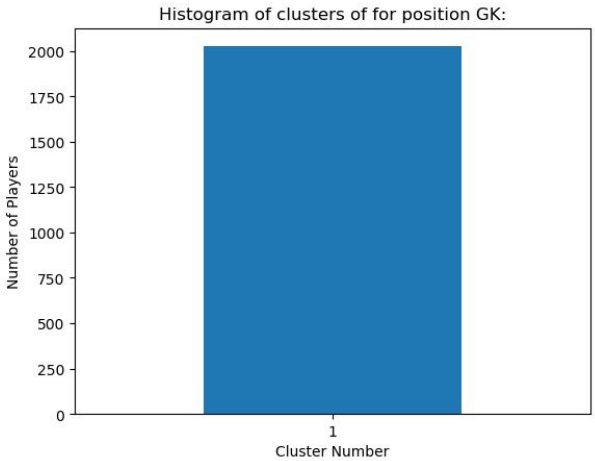


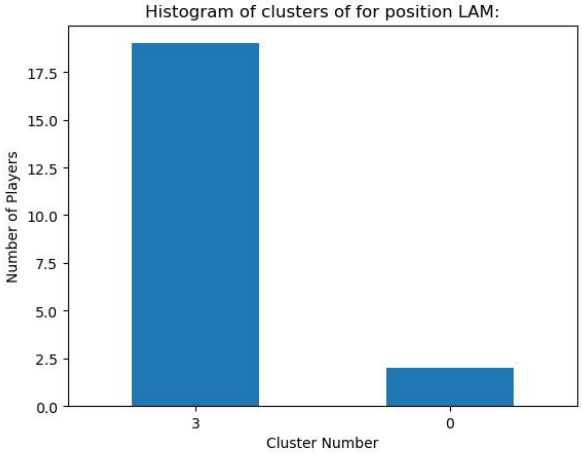


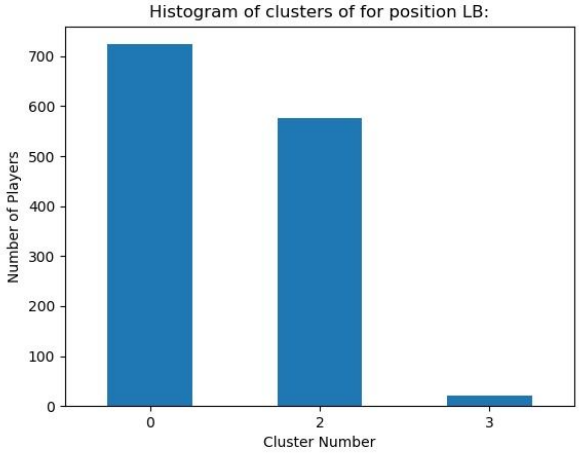


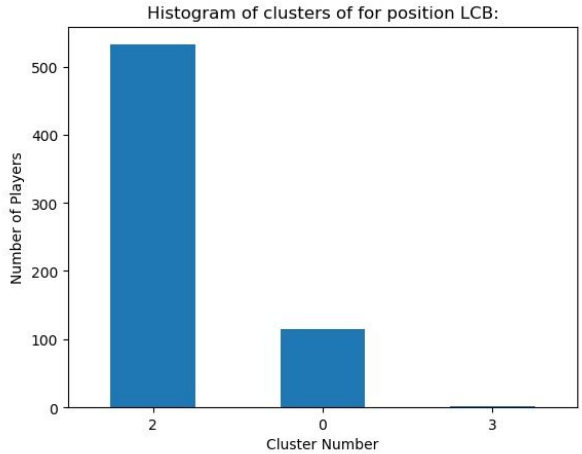


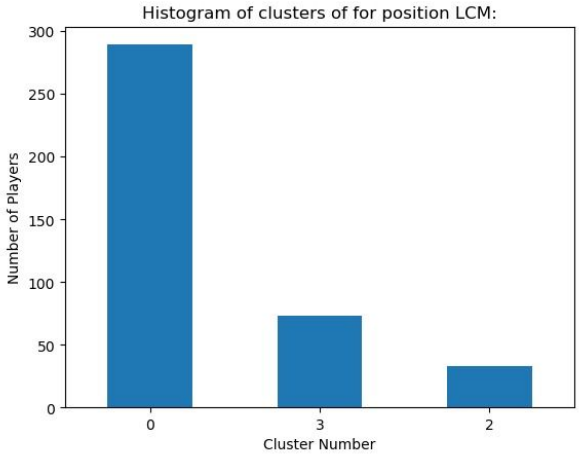


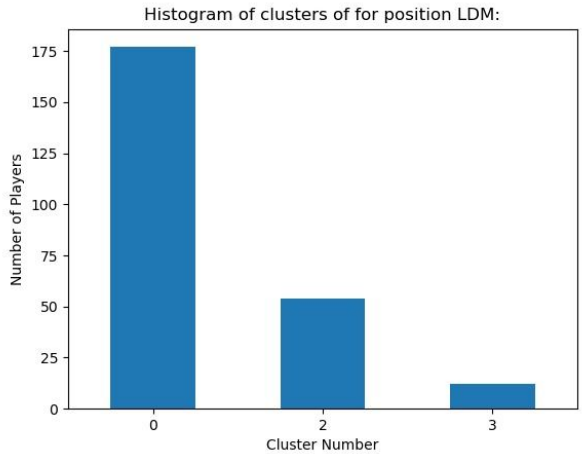


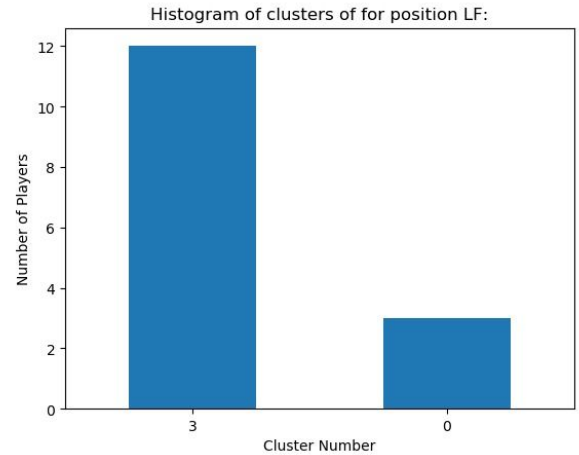

Histogram of clusters of for position LAM:
Number of Players
Cluster Number
3
0

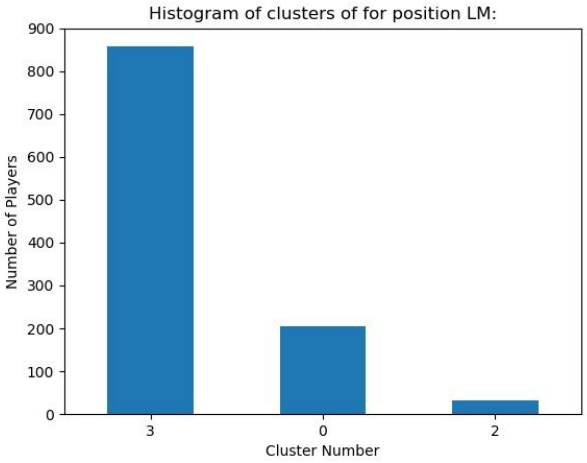

Histogram of clusters of for position LB:
Number of Players
Cluster Number
0
2
3

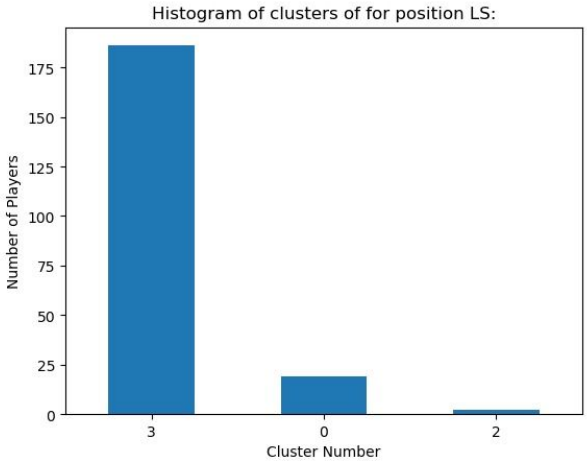

Histogram of clusters of for position LCB:
Number of Players
Cluster Number
2
0
3

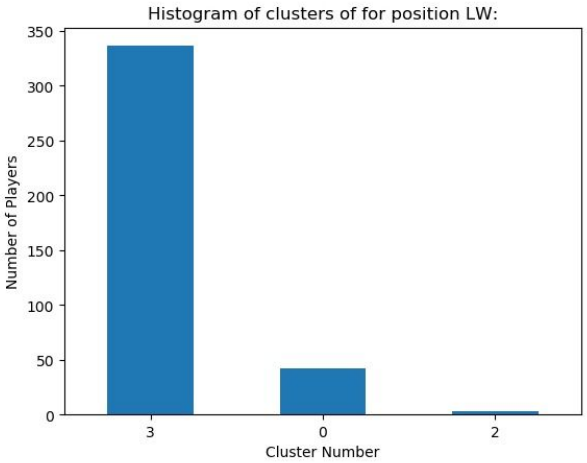

Histogram of clusters of for position LCM:
Number of Players
Cluster Number
0
3
2

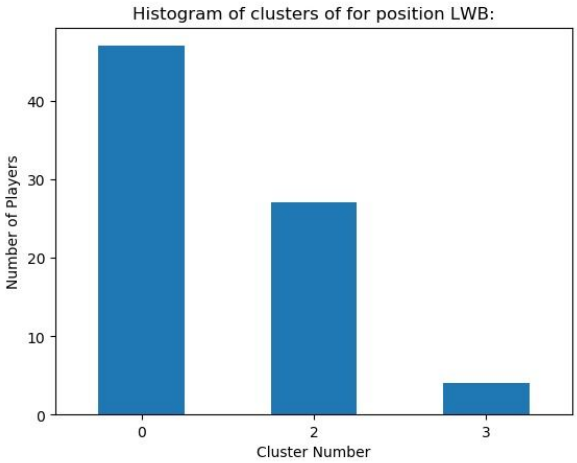

Histogram of clusters of for position LDM:
Number of Players
Cluster Number
0
2
3

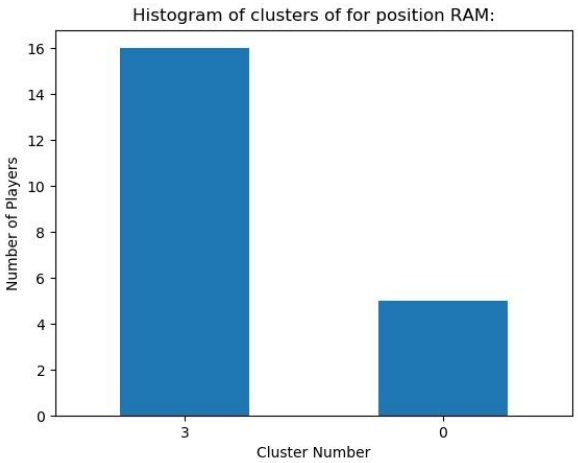

Histogram of clusters of for position LF:
Number of Players
Cluster Number
3
0

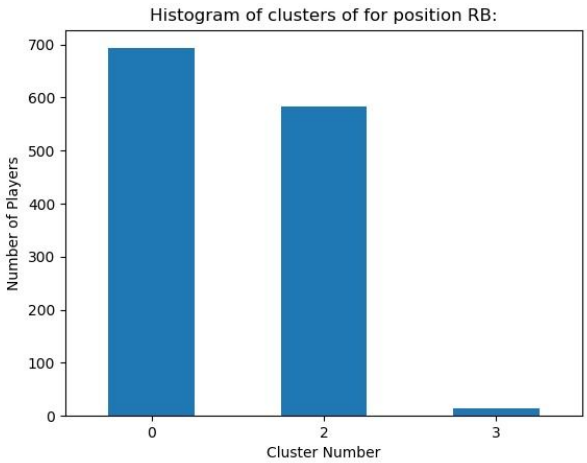

Histogram of clusters of for position LM:
Number of Players
Cluster Number
900
800
700
600
500
400
300
200
100
0
3
0
2

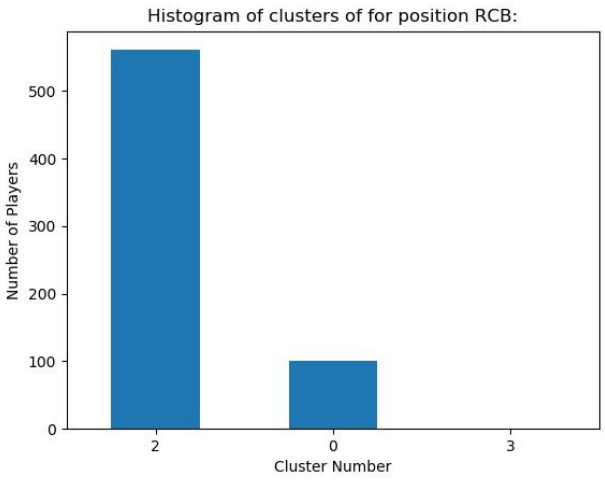

Histogram of clusters of for position LS:
Number of Players
Cluster Number
175
150
125
100
75
50
25
0
3
0
2

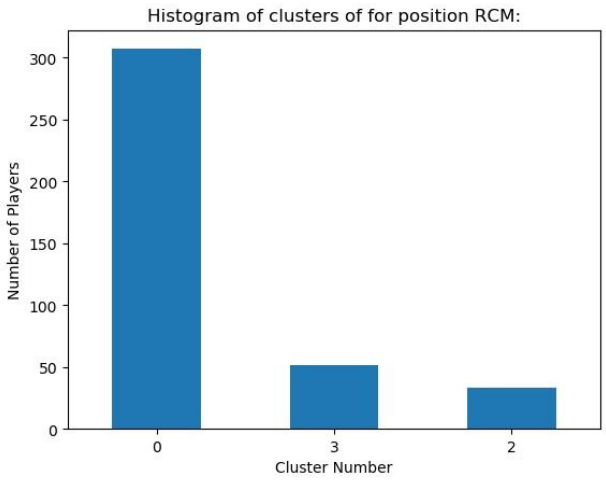

Histogram of clusters of for position LW:
Number of Players
Cluster Number
350
300
250
200
150
100
50
0
3
0
2

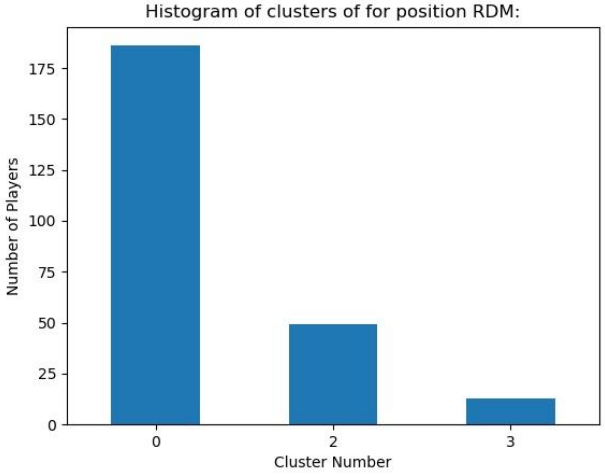

Histogram of clusters of for position LWB:
Number of Players
Cluster Number
40
30
20
10
0
0
2
3

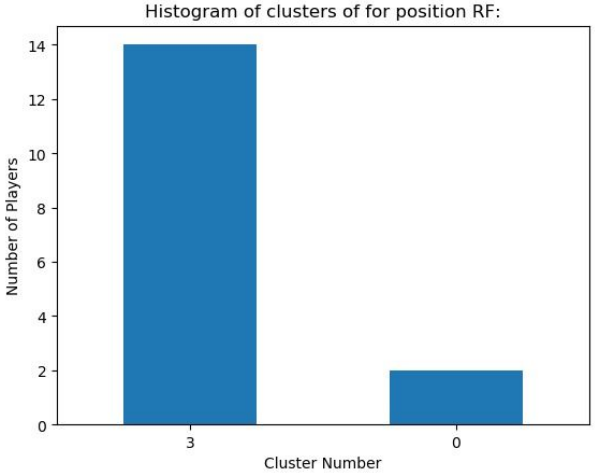

Histogram of clusters of for position RAM:
Number of Players
Cluster Number
16
14
12
10
8
6
4
2
0
3
0

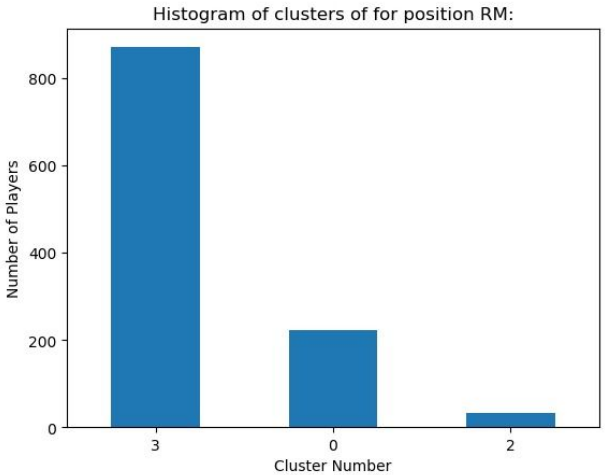

Histogram of clusters of for position RB:
Number of Players
Cluster Number
700
600
500
400
300
200
100
0
0
2
3

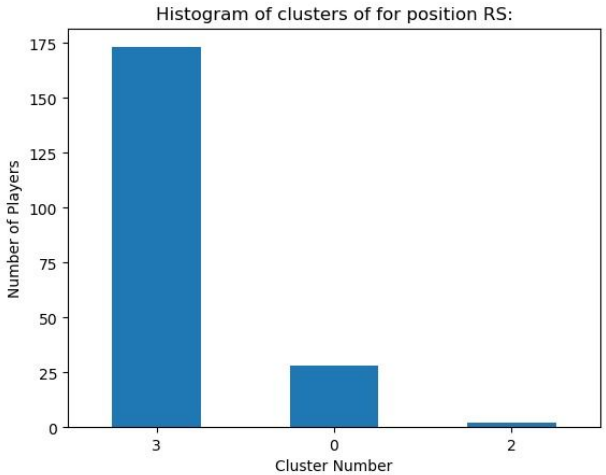

Histogram of clusters of for position RCB:
Number of Players
0
100
200
300
400
500
2
0
3
Cluster Number

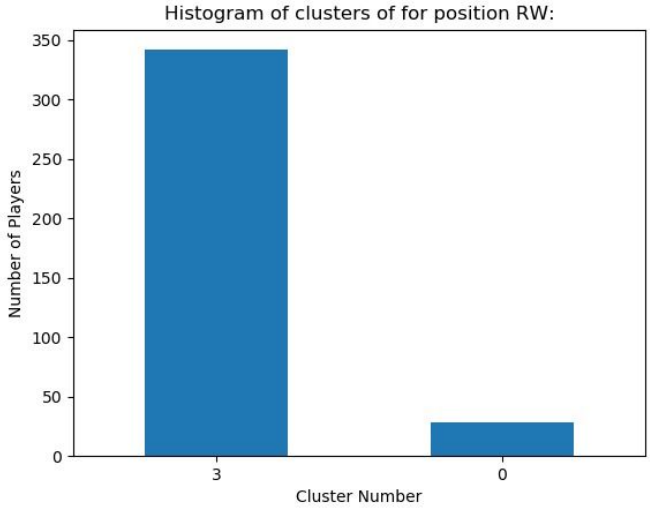

Histogram of clusters of for position RCM:
Number of Players
0
50
100
150
200
250
300
0
3
2
Cluster Number

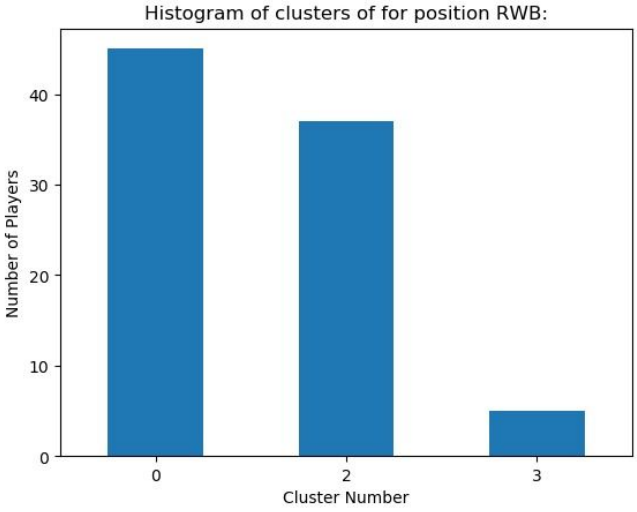

Histogram of clusters of for position RDM:
Number of Players
0
25
50
75
100
125
150
175
0
2
3
Cluster Number

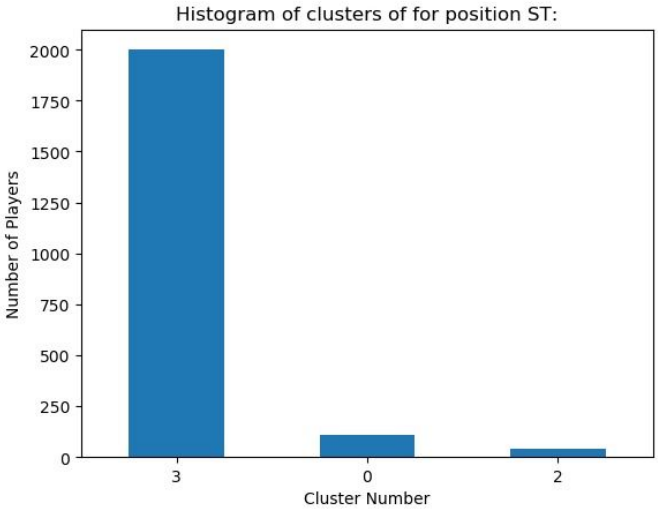

Histogram of clusters of for position RF:
Number of Players
0
2
4
6
8
10
12
14
3
0
Cluster Number


Histogram of clusters of for position RM:
Number of Players
0
200
400
600
800
3
0
2
Cluster Number


Histogram of clusters of for position RS:
Number of Players
0
25
50
75
100
125
150
175
3
0
2
Cluster Number


Histogram of clusters of for position RW:
Number of Players
Cluster Number
0
50
100
150
200
250
300
350
3
0


Histogram of clusters of for position RWB:
Number of Players
Cluster Number
0
10
20
30
40
0
2
3


Histogram of clusters of for position ST:
Number of Players
Cluster Number
0
250
500
750
1000
1250
1500
1750
2000
3
0
2

Player position reference:

For more information about football players' position please refer here.