2023

Is A “Centre-Forward” Footballer’s Market Value Determined only by Their Statistics?

BY GAJAN KANAGENTHIRAN 190282527

Data analytics ECS784U/ECS784P

Hypothesis

The problem this report tackles is whether a 'Centre-Forwards' football player is priced mainly on their performance statistics (goals + assists per game). In my opinion, I believe that football players' goals and assists only have little influence on their market value, as there would be other factors that would have an effect on their value.

Introduction and Literature review

A very controversial topic in football is a player's market value, as in football, a player's market value is not determined by one unbiased group of people instead it is determined by the club the player plays for. The price reflects what the club values for that player. Many Footballing fans believe that footballers are priced incorrectly and are inflated as the clubs are biased toward making a profit from a player instead of correctly evaluating a player's worth. Fans believe this is ruining the sport as the sport is shifting its focus to a more business-oriented focus than it being about the entertainment and performance aspect of football.

For example, a common quality desired by a forward, which is an attacking-focused player who plays further up the field, is the amount of goal contribution they have especially their goals and assists. However, many fans would argue and discuss how 'Centre-Forward' football players are not priced fairly as clubs increase their value when they get very fluky goals and assists during a season.

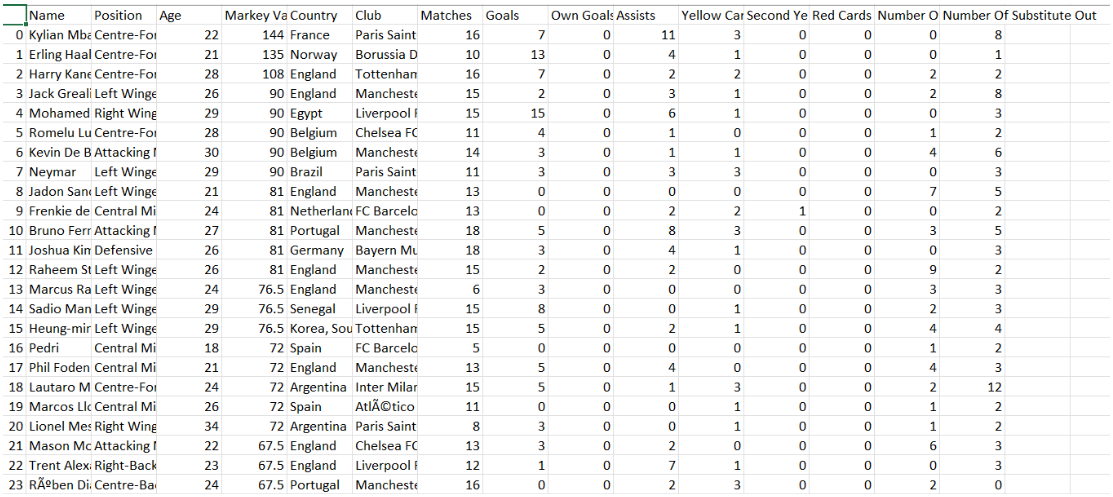
Fans believe that some players with similar goals and assists per game are getting priced the same when they do not have the same impact on the sport or team as another player who offers much more than just lucky goals and assists. ***'part of the reason players have "poor stats" is that the output they produce on the field is not considered within the context of their role in the team.'*** (Worville, 2020). Fans believe other factors such as playing style, contribution, age, and the league of the player as some leagues may be harder to play in than others should hold more importance than just statistics alone.

Fortunately, modern football has now been innovated to a level where all players' statistics are recorded, making it easier to analyse the performance of a footballer. This project aims to determine the relationship between a "Centre-Forward's" market value and their statistics to see how impactful and influential a "Centre-Forward's" statistics is to their market value.

Data set

The data set being examined is "Most Expensive Footballers 2021" (Naik, 2021). The dataset aims to provide a club's more accurate reflection of a player's worth in a free market by considering various factors such as future prospects, performance at their club and national team, and the level and status of their league.

From the dataset, only the top 50 "Centre-Forward" position players with the highest market value would be considered. This is because different positions in football are analysed based on different statics therefore, only one position is being considered to remove any unfairness. Furthermore, "Left-Forward" and "Right-Forward" would also not be considered due to the argument that one side could be easier to perform well in than the other, eliminating any selection bias.



	Name	Position	Age	Markey Va	Country	Club	Matches	Goals	Own Goals	Assists	Yellow Car	Second Ye	Red Cards	Number O	Number Of Substitute Out
0	Kylian Mba	Centre-For	22	144	France	Paris Saint	16	7	0	11	3	0	0	0	8
1	Erling Haal	Centre-For	21	135	Norway	Borussia D	10	13	0	4	1	0	0	0	1
2	Harry Kane	Centre-For	28	108	England	Tottenham	16	7	0	2	2	0	0	2	2
3	Jack Greali	Left Winge	26	90	England	Mancheste	15	2	0	3	1	0	0	2	8
4	Mohamed	Right Wing	29	90	Egypt	Liverpool F	15	15	0	6	1	0	0	0	3
5	Romelu Lu	Centre-For	28	90	Belgium	Chelsea FC	11	4	0	1	0	0	0	1	2
6	Kevin De B	Attacking I	30	90	Belgium	Mancheste	14	3	0	1	1	0	0	4	6
7	Neymar	Left Winge	29	90	Brazil	Paris Saint	11	3	0	3	3	0	0	0	3
8	Jadon San	Left Winge	21	81	England	Mancheste	13	0	0	0	0	0	0	7	5
9	Frenkie de	Central Mi	24	81	Netherlan	FC Barcelo	13	0	0	2	2	1	0	0	2
10	Bruno Ferr	Attacking I	27	81	Portugal	Mancheste	18	5	0	8	3	0	0	3	5
11	Joshua Kin	Defensive	26	81	Germany	Bayern Mu	18	3	0	4	1	0	0	0	3
12	Raheem St	Left Winge	26	81	England	Mancheste	15	2	0	2	0	0	0	9	2
13	Marcus Ra	Left Winge	24	76.5	England	Mancheste	6	3	0	0	0	0	0	3	3
14	Sadio Man	Left Winge	29	76.5	Senegal	Liverpool F	15	8	0	0	1	0	0	2	3
15	Heung-mir	Left Winge	29	76.5	Korea, Sou	Tottenham	15	5	0	2	1	0	0	4	4
16	Pedri	Central Mi	18	72	Spain	FC Barcelo	5	0	0	0	0	0	0	1	2
17	Phil Foden	Central Mi	21	72	England	Mancheste	13	5	0	4	0	0	0	4	3
18	Lautaro M	Centre-For	24	72	Argentina	Inter Milar	15	5	0	1	3	0	0	2	12
19	Marcos Llo	Central Mi	26	72	Spain	AtlÃ©tico	11	0	0	0	1	0	0	1	2
20	Lionel Mes	Right Wing	34	72	Argentina	Paris Saint	8	3	0	0	1	0	0	1	2
21	Mason Mc	Attacking I	22	67.5	England	Chelsea FC	13	3	0	2	0	0	0	6	3
22	Trent Alex	Right-Back	23	67.5	England	Liverpool F	12	1	0	7	1	0	0	0	3
23	RÃºben Di	Centre-Ba	24	67.5	Portugal	Mancheste	16	0	0	2	3	0	0	2	0

Figure 1: Original data set' (Worville, 2020).

From the data set the columns that were of interest were the:

- Position
- Market Value
- Goals
- Assists
- Matches

Initially, the data from the CSV file was inputted into python and then filtered so that only the top 50 "Centre-Forward" were taken into the data frame. From that data frame, four more data frames were derived: Market Value, Goals, Assists and Matches.

The goals and assists data frames were added together and then the result was divided by the matches data frame to find the goals and assists per match which was renamed to the player performance statistics data frame. Then a final data frame was created with the market value and the goals and assists per match as seen in table 1. This final data frame was then used to plot a graph.

Table 1: Part of the Market Value vs Player Performance statistics (Goals and Assists Per Match) Played table.

Name	Market Value (£)	Player Performance statistics (Goals and Assists per Match played)
Kylian Mbappé	144.0	1.1
Erling Haaland	135.0	1.7
Harry Kane	108.0	0.6
Romelu Lukaku	90.0	0.5
Lautaro Martínez	72.0	0.4
Victor Osimhen	54.0	0.8
Gabriel Jesus	54.0	0.6
Robert Lewandowski	54.0	1.5
Richarlison	49.5	0.4
Timo Werner	49.5	0.3
Dušan Vlahovic	45.0	1.0
Dominic Calvert-Lewin	40.5	1.0
Memphis Depay	40.5	0.5
Álvaro Morata	40.5	0.4
Cristiano Ronaldo	40.5	0.8
Youssef En-Nesyri	36.0	0.6
Alexander Isak	36.0	0.5

Learning Methods:

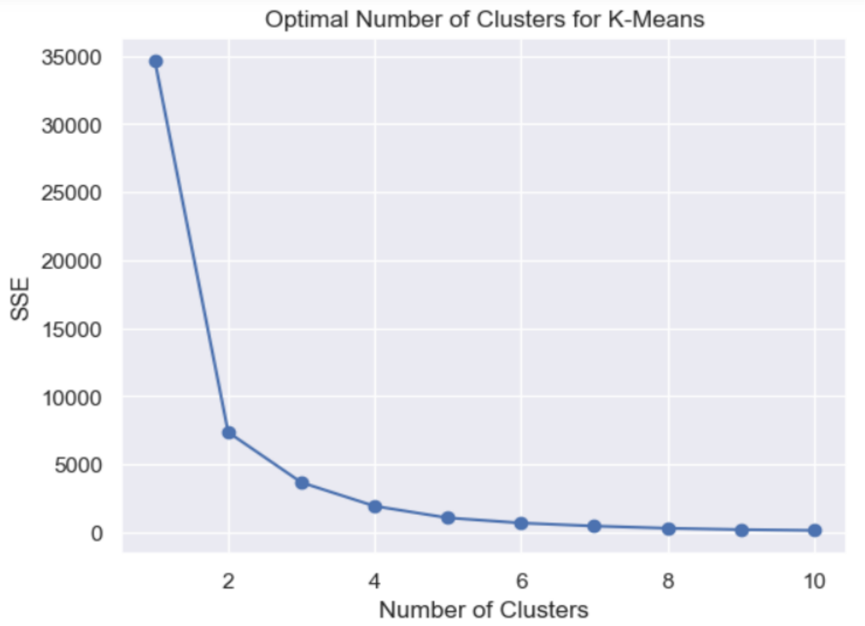
K-Means:

K-means is an unsupervised data analytic technique where it splits the dataset into K-clusters, grouping each data point into a cluster based on its proximity to the mean. The algorithm then iteratively updates the position of each centroid based on the newly assigned data points until convergence. This was the main reason K-Means was chosen to analyse the data as each cluster can be analysed to give a more detailed understanding of the relationship between the two variables. (Code can be found in Juptyer notebook).

Based on the optimisation of how many clusters would be most favourable, it was clear 4 clusters or more provided a lower SSE (Sum of Squared Errors) as seen in graph 1. The graph had shown the closer the number of clusters is to 10, the lower the SSE value is. However, the number of clusters chosen was four. The reasoning being is that even if having a high number of clusters would reduce the SSE it can also complicate the results:

- With fewer clusters, it would be easier to interpret the results.
- With more clusters, it would be harder to identify key underlying patterns and relationships in the data, especially with a small dataset being used.

Graph 1: Optimal Number of Clusters for K-Means graph



(Code for graph 1 can be found in jupyter notebook file)

Linear Regression

Linear Regression is a supervised data analytic technique that is used to model the relationship between two variables using a linear equation, and then test the model created using another dataset to see how well the model can predict the test data. This works by splitting the dataset to obtain two datasets. One is used to train the model and the other is used to compare against the model to see how accurate the model is.

Linear Regression was also chosen to analyse the data over other unsupervised data analytics techniques such as GMMs (Gaussian Mixture Models) as GMMs would produce a similar result to K-Means especially, as there are such few data points. Therefore, it would not make sense to do GMM. As the relationship between Market Value and player performance statistics is being investigated, linear regression would be used in an unorthodox method. For this method to work a model would be created with some of the data points and then plotting the training data against the model, the regression line can be used to inspect the relationship between the two variables. When the training data is put against the model we can see if the relationship is linear and how strong the correlation is between the two variables.

Results:

K-means:

Graph 2: "K-Means of the Market Value of A 'Centre-Forward' Compared to Their Player Performance statistic"

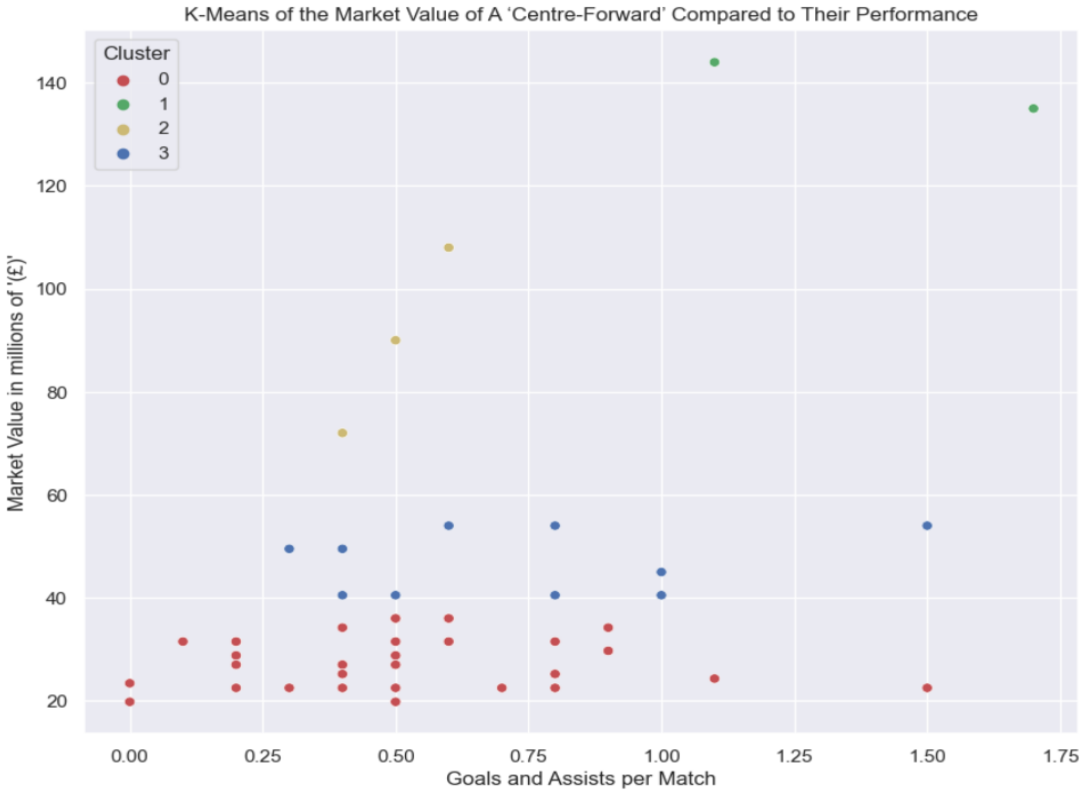


Table 2: Table with the Correlation Values of Each cluster

Cluster Colour	R-Value (5.S.F)	Market Value Range in Millions (£)	Average Player Performance Statistic Range (2.d.p)
Red	0.0053404	20-40	0.49
Blue	0.19665	40-60	0.73
Yellow	1.0000	70-110	0.50
Green	-1.0000	130-150	1.40

(Code for graph 2 and table 2 can be found in jupyter notebook file)

When looking at the green cluster it has the most expensive market value ranging from £130-150 million and has the highest average player performance statistics of 1.4. However, the R-Value of the green cluster is -1.0 which is a strong negative correlation indicating the better the player performance statistic the cheaper they are. But as there are only two data points in this cluster it would be very difficult to come to a reliable conclusion for it as the correlation can either only be 1.0

or -1.0. Therefore, the green cluster should not have much significance or weight to the overall hypothesis.

Both the Red and Blue Clusters have an R-Value below 0.2 with the Red Cluster consisting of an R-Value of 0.00534. An R-Value this close to zero demonstrates an extremely weak to no Linear correlation between market value and player performance statistics, showing that the average player performance statistic has only a small influence in determining market value.

The Yellow Cluster however is the exact opposite of the green cluster with an R-Value of 1.0 showing a perfect positive linear correlation between the two variables when the market value ranges between £70-110 million. However, the average player performance statistics of the yellow cluster was just 0.01 more than the red cluster which has the lowest-valued players and has a difference of 0.23 average player performance statistics compared to the blue cluster which has cheaper players ranging between £40-60 million. From this data, it demonstrates that even in the higher price range, performance statistics are not considered as heavily because the average market value in the yellow cluster was higher than the other two clusters while the player performance statics was low when compared as well. But the yellow cluster has only three data points present in it meaning it would heavily bias an R-value or 1.0 or -1.0 which needs to be considered when figuring out the relationship between the two variables.

Linear Regression:

Graph 3: "Linear Regression of the Market Value of A 'Centre-Forward' Compared to Their Player Performance statistic"

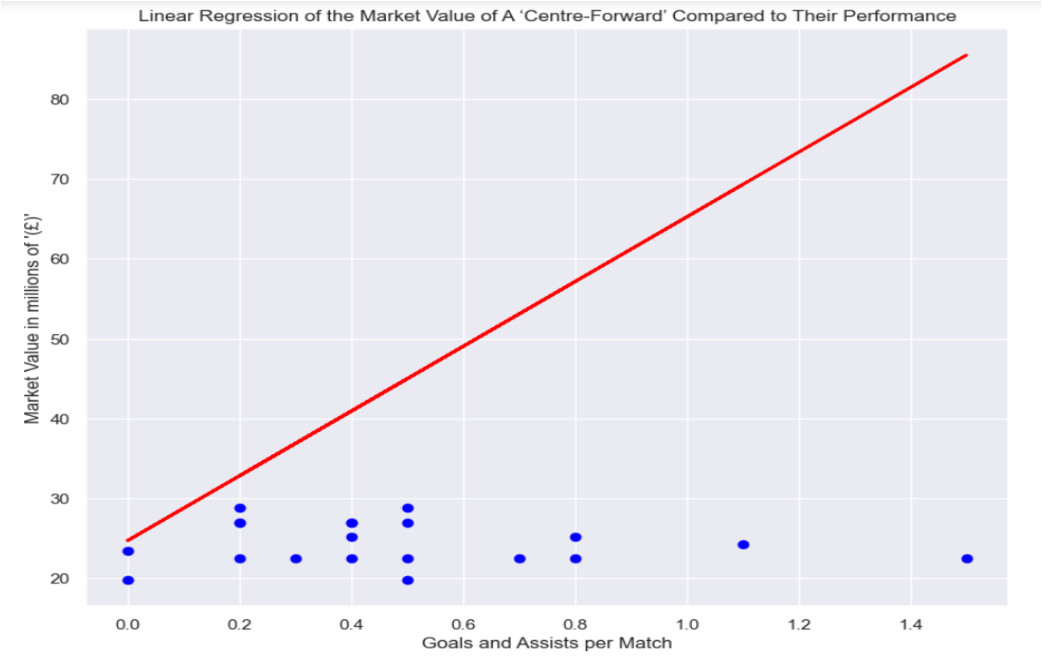


Table 3: Table with the Correlation Values for Linear Regression Model

Linear Regression Result	
R-Value (5.S.F)	0.24308

(Code for graph 3 and table 3 can be found in jupyter notebook file)

If only the correlation of the model was considered, the R-Value obtained from the regression model is 0.24308 suggesting that there is a small positive correlation between the two variables.

Looking at graph 2 it is evident that only a few test data points are close to the linear regression model. Nevertheless, the model has a poor fit as most of the test data points are not very close to the regression line potentially showing that the variables Market value and Player performance statistics do not have a linear relationship with one another as the model is not accurate as the model is underfitting and not complex enough.

Choosing The Optimal Data Analytic Technique

K-Means:

One main challenge when determining the relationship between the two variables was that some clusters had significantly fewer points in them compared to the rest. Especially when observing the yellow and green clusters as there are only three or fewer points in those clusters. This was a challenge as only a few data points it would make the correlation very one-sided as the clusters would have correlations that would be closer to -1 or 1 and not fall in between.

Another problem faced when doing K-Means is when the optimisation was analysed it had shown the SSE would drop the more clusters present. Due to the low number of data points, it would not be possible as explained in the **learning methods** section, resulting in the K-Mean having a larger SEE than desired.

Linear Regression:

The sample size used was not ideal for the linear regression method as the larger the data set is, the better the model would be at predicting the test data. With only 50 points available, 30 were used to train the model and 20 points were used to test the model, meaning the model would be much less accurate. Furthermore, the model produced from the linear regression method had a lot of underfitting and was not complex enough as the test data used were not very close to the regression line as seen in graph 3.

After looking at the drawbacks of each data analysis method and considering why each method was chosen as mentioned in the **learning method** section, the data analysis method chosen was the K-Means method. When comparing the two methods, Linear regression looks at the data set as a whole unlike, K-Means as K-Means breaks the data into clusters, allowing more detailed analysis, and making it easier to find hidden trends in the data. With K-Means, it was easier to identify the underlying patterns highlighting the different types of relations the two variables can have with each other. Furthermore, the Regression model was underfitting and was too simple as it does not highlight any underlying patterns in the data, it only provided an unreliable correlation between the two variables across the whole dataset.

Key Findings:

When looking at all the K-mean clusters excluding the yellow cluster it shows that higher Market Value results in a higher average player performance statistic. This is evident when looking at table 2 as when the average player performance statistics increased the Market value range in Millions (£) also increased. Considering this fact, it makes it clear players that who have higher statistics are valued higher, than those who do not tally as many goals or assists as well. This can also be further supported if the results for the linear regression were considered as it had an R-Value of 0.24308 for the regression line.

Although, when investigating each cluster in K-Means individually, it can be determined that player performance statistics have a small impact on the market value of a “Centre-Forward” which the R-values of the blue and red clusters prove, as they are below 0.2. Therefore, it can be considered players valued 60 million or less are somewhat affected by player performance statistics. This may be because many low market-valued players may be overlooked due to having less exposure because they played a lower number of games or play in a league that is not very popular, so their player performance statistics would have some impact on how much they can bring to the sport as they are not watched often.

Yet, when considering the R-Value of the yellow cluster -1.0 it can be determined after a certain price range above 120 million, the player performance statistics and market value are indirectly related. However, in practical terms, if a player performs better, it should not reduce his market value; therefore, it can be deduced that player performance does not impact market value when the market value is above £130 million. Therefore, it can be concluded that 'Centre-Forward' players are not priced higher solely based on goals and assist but could be priced higher due to other factors which could be age, league and playing style, etc. after their value passes £130 million. As typically more expensive players play in more important games that are watched by more people.

Therefore, based on the results, it can be concluded that higher-valued players with similar statistics are not priced based on statistics only. While low-valued "Centre-Forward" players' price is influenced by their statistics. But overall, as player performance statistics (goals and assists per match) increased a "Centre-Forward's" market value in millions (£) would also increase.

Conclusion

The report aimed to deduce how impactful and influential a ”Centre-Forward's" statistic is to their market value. Based on the K-Means analysis it was found that the higher a player's statistics are the more impactful to market values below a market value of £60 million. However, higher-valued players over £130 million players that have similar player statistics are not priced solely on their statistics. Furthermore, it was clear that the use of K-Means to analyse the data was more superiorly suited to analyse this dataset compared to the linear regression method due to the more in-depth analysis it provided. As it broke down the relationship for different market values and performance statistic ranges.

References:

Naik, S.S. (2021) *Most expensive footballers 2021*, *Kaggle*. Available at: https://www.kaggle.com/datasets/sanjeetsinghnaik/most-expensive-footballers-2021 (Accessed: March 10, 2023).

Worville, T. (2020) *Explained: How can someone with bad stats be a good player?*, *The Athletic*. Available at: https://theathletic.com/1763286/2020/04/23/bad-stats-but-good-player-analysis/ (Accessed: March 10, 2023).