Uncovering Trends in Football Transfers to Determine League Status

Dylan Fox

10/13/2022

Football (soccer) is the most popular sport in the world and drives a significant amount of data in various aspects - from watching the sport to playing the sport. One exciting aspect of the sport that has fans experience a wide array of emotions is the transfer of players between teams. With thousands of transfers spanning across several leagues every season, it opts for an opportunity to analyze and understand the underlying patterns within the transfer market. By completing clustering analyses, comparing distributions of data over time, and connecting the data with contextual knowledge of the football world, we will be able to uncover unique trends in the transfer of players. These can be centered around the nationalities of players being transferred, the transfer fee for a player, the market amount they are valued at, and much more.

Introduction

Football players are continuously ridiculed for the transfers and teams they sign for. Often times it can sway the public opinion of the player. Not only do teams play a vital role in the transfer process, but the league in which the team plays does as well. This data set we look at contains thousands of transfer observations from Europe's top seven leagues from the years 2009 to 2021. By grouping the transfer observations by league, we can start to understand underlying patterns and attributes about the league. For example, the typical age of players in a league, the number of transfers, the most frequent nationalities of players, and so on. In addition, we can start to understand the attributes that make up variables such as transfer fee amount and market value amount. These two metrics play a crucial role in the transfer process - how are they determined?

These are the sorts of questions and patterns we can opt to uncover with this data. By grouping the data by league and season, we can perform interesting cluster analyses on the data to discover if there are other attributes that cluster similarly to the league in which transfers take place. Overall, we'll look to identify trends that can help us determine the status of a league over time.

Dataset

The dataset used for this analysis comes to us from Transfermarkt.com and was collected by GitHub User Dmitrii Antipov (d2ski). It contains roughly 70,000 observations of Europe's top seven football leagues from 2009-2021. Each observation is a single player transfer. The seven leagues are:

- English Premier League - `GB1`
- La Liga - `ES1`
- Serie A - `IT1`
- Bundesliga - `L1`
- French Ligue 1 - `FR1`
- Liga Portual BWIN - `PO1`
- Dutch Eredivisie - `NL1`

Each observation of a transfer has 23 variables with it. Some examples of these variables are: season (year), transfer window (summer or winter), team name, player name, transfer fee amount, market value amount, player position, etc.

Methods

Preprocessing

The data came to us with a considerable amount of observations missing values. Thankfully, the missing values were contained to three columns: player age, transfer fee amount, and market value amount. Before we can begin to utilize data mining techniques, we need to impute the missing the data.

We start with imputing data for the observations missing player age values. The amount of observations missing this value was very small compared to the other columns missing data. Therefore, it was deemed wise to simply impute the missing values with the average age of the rest of the observations. This was done because it is a quick way to impute data and works well since there are so few observations missing this column value.

The other two columns missing values, transfer fee amount and market value amount, were missing significantly more. Because of this reason, it was best to take more time to impute accurate values into the observations. This was accomplished by building and testing several multivariate linear models to predict each value.

For imputing both columns, we randomly split our data into a train and test set (70/30 split). We followed this with our feature selection by creating models with various combinations of columns from the data set. In both of the final models, we ultimately used similar, but slightly different, columns to impute our missing data.

In terms of imputing the transfer fee amount, we learn that the following columns produce the most accurate results in predicting transfer fee amount: league, season, player age, market value amount, and whether a transfer was free. If we examine these variables ourselves, we deduce that these make sense to include in our model to predict transfer fee amount. Each league in our data set has various levels of playing - it's known that the English Premier League is arguable the best league in the world. If you compare the Premier League with Portugal's league, you are bound to find higher profile players England. The season is also important because a transfer value in 2009 is vastly different than 2021 due to inflation. The player's age often correlates with a player's value and transfer fees. If an observation had a market value amount, it was by far the most significant factor in predicting transfer fee amount. Finally, the column "is_free" always denoted that the transfer fee amount is worth zero.

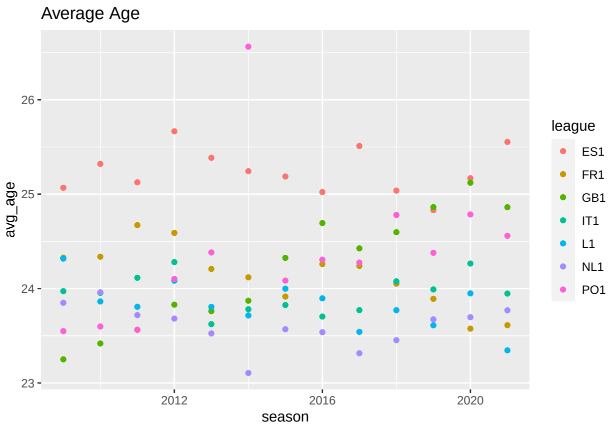
After building a multivariate linear model with the aforementioned columns, we can test the model on our test data set. Then we can take the difference between the actual and predicted result. After doing this, we discover that the average difference of the actual and predicted results was about 4,000,000 EUR.

Next, we move on to imputing the market value amount column. This process was very similar to the imputing the transfer fee column. The following columns produced the most accurate results in predicting market value: league, season, player age, transfer fee, whether a transfer was free, and whether the transfer a loan. Again, using these columns makes sense for similar reasons they made sense to use in predicting transfers fees – often times these are the metrics used to determine a player's worth which is correlated with market value amount.

After building the multivariate linear model for market value amount, we tested it on our test data set. We took the difference between the actual and predicted result. After doing this, we discover that the average difference of the actual and predicted results was slightly lower than the transfer fee value results – just shy of 4,000,000 EUR.

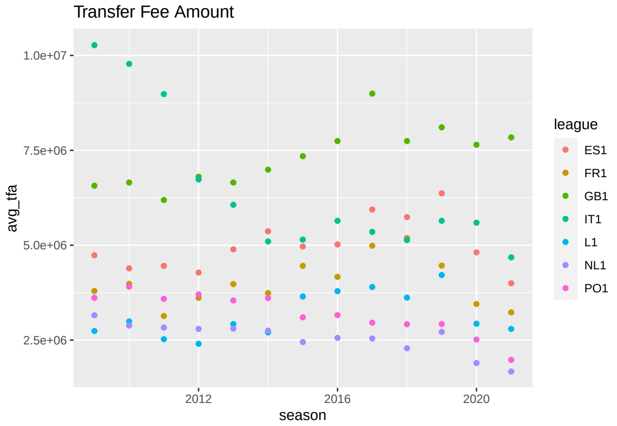
Exploratory Analysis

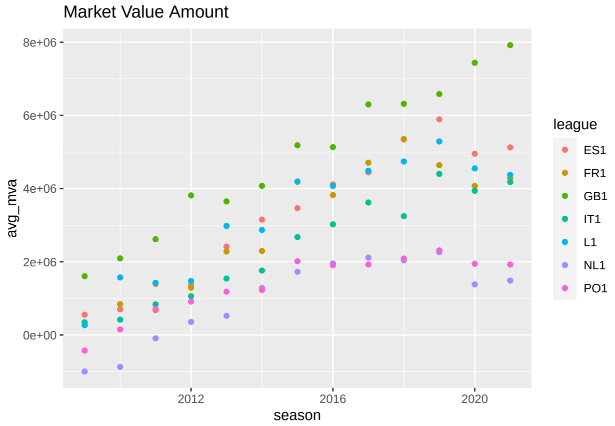
With the data set no longer missing any values, we can start to explore trends within the data. We can visualize the average age of each transfer coming in to a league over the years to start.



Some of these trends match with historical knowledge. For example, the Netherlands league has a very low average age each year. This is not surprising as this league is known to promote youngsters and then transfer them out. The French league started with a higher average age than how it ended in 2021; this also matches with how the league behaves currently. More and more young players are transferring to the French league. On the other side, it’s surprising to see La Liga (Spain) consistently have one of the highest average age of its players. This can most likely be attributed to the fact that some of the most prominent teams play in Spain – FC Barcelona, Real Madrid, Atletico Madrid. As young players age and mature, they will look to make a transfer to these talented teams. It makes sense that La Liga has a higher average age with that in mind.

Next, we can take a look at transfer fee amount and market value amount. The graphs below illustrates the average of each metric over time by league.





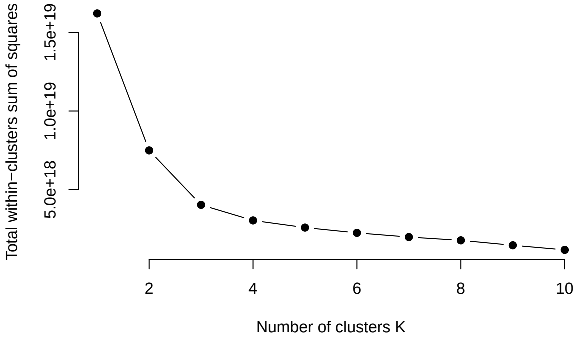
These graphs illustrate interesting results. The Premier League (Great Britain) can be seen consistently having one of the highest average transfer fee each year. This is not surprising since this league is considered the best in the world. In addition, the Premier League goes onto have the highest average market value amount in a transfer each as well. Again, not surprising for the same reason.

Overall, this preliminary analysis is very useful. We are able to gather information very quickly that give us a good synopsis on how each league compares to one another. We can also see how trends relate to world events in this time frame. For example, each league has their average market value amount of a transfer increase over time. This most likely has a high correlation with inflation. It’s interesting we do not see that same trend in the graph depicting transfer amount – but, part of this can attributed to the fact that it’s very common for a transfer fee to be 0 EUR. This can skew the averages.

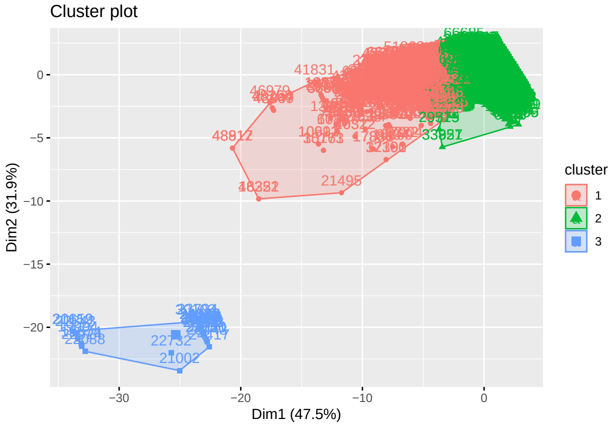
This information leads us into diving deep within the data. We are able to narrow our search in providing comparative analyses of the same league over time by looking at specific metrics, such as transfer fees and market value amounts.

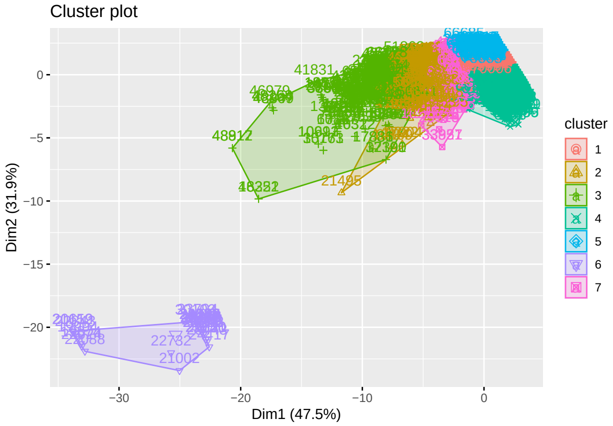
Results

Since our data contains seven distinct categories (each league), we can conduct a K-Means clustering algorithm on the data to see how the optimal number of clusters compares. Judging from the elbow plot below, the optimal number of clusters is actually two or three.



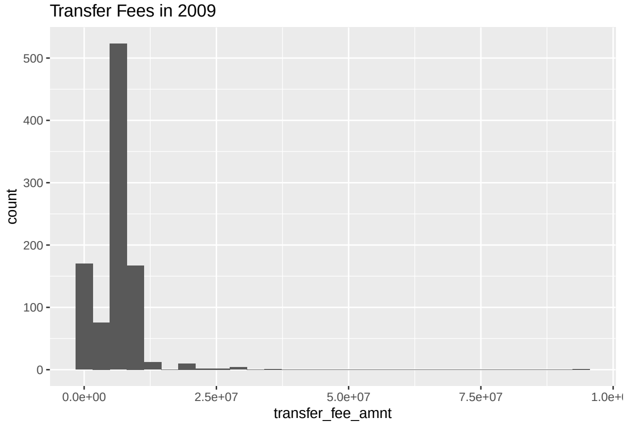
Here is a comparison of using k-means clustering when our K equals three and K equals seven.

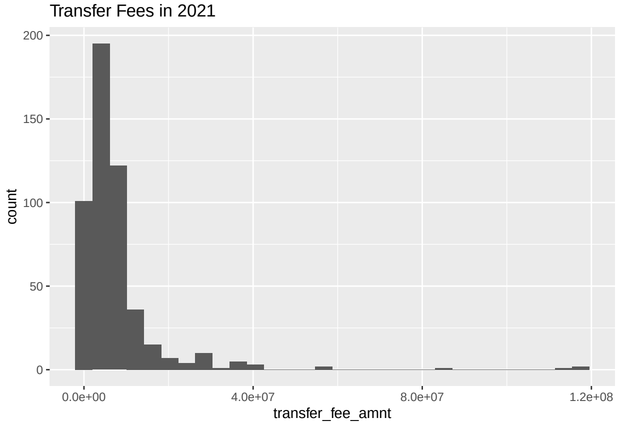




We know for a fact that there are seven possible clusters in our data set – one for each league. While this may be true, it's evident that the leagues share similarities when we base the clustering off of player age, transfer fees, and market value amount. This is the reason why the elbow plot leans toward a value of three for K.

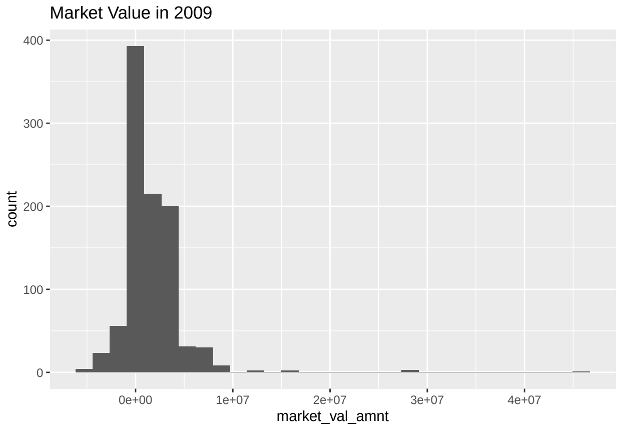
We can narrow down our analysis and focus on single league, the Premier League. If we analyze trends over time, we'll be able to determine the status of the league and how it's changed. Below are two histograms that illustrate the difference in distributions of transfer fees in 2009 versus 2021 of the Premier League.

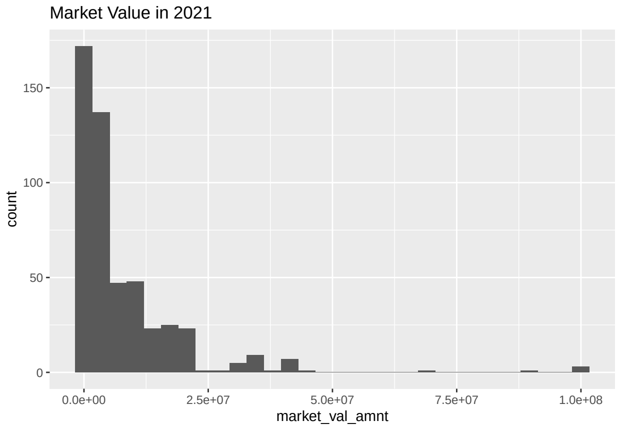




In 2009, it's evident more transfers were similar in transfer fee values. The peak of the distribution is the same as 2021; but there's over double the amount of occurrences in 2009 than in 2021. In 2021, the distributions are more distributed across values. This is because 2009 sports nearly double the amount of transfers than 2021. While 2009 may have double the amount of transfers, 2021 has an average transfer fee of 7,840,395 EUR, while 2009 has an average of 6,567,565 EUR. Based on these trends, it's evident that transfer fees have increased significantly for the Premier League over time.

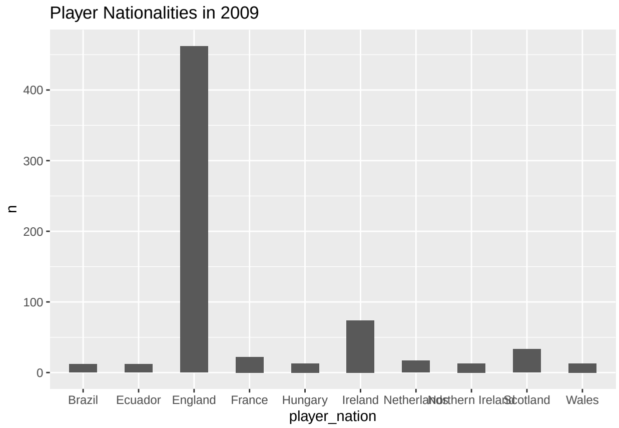
Here are graphs showing the distributions of market value in 2009 versus 2021.

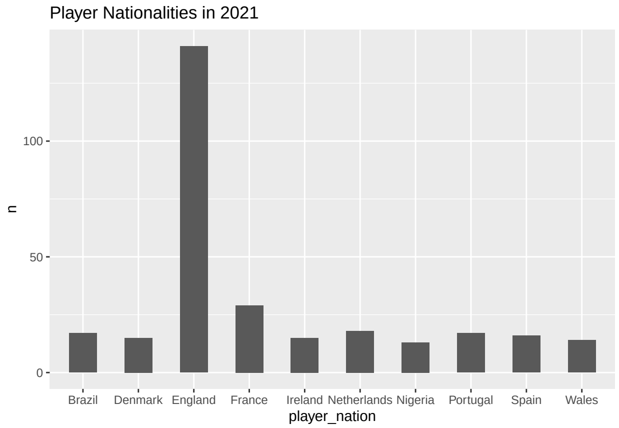




Comparing 2009 to 2021, there is a similar story as the transfer fee amount. This is expected to happen because transfer fees and market valuations are typically correlated together. In 2009, the distribution is highly dense in the center. In 2021, the peak is similar to 2009 but there are many more occurrences of higher market value amounts. The average market value amount in 2009 was 1,604,642 EUR and the average in 2021 7,919,594 EUR. Based on these metrics, it's clear that the Premier League has grown significantly in the last decade in terms of player valuations.

Furthermore, the Premier League has expanded on the diversity of its players. Shown below are plots highlighting the top ten nationalities from each year.





It's evident from these plots that England continues to be the primary source of players in the Premier League. This makes sense as England is home to the Premier League. But, we start to see more players of different nationalities in the more recent years. This can be attributed to other metrics we have looked at – transfer fee and market value. As leagues grow and become more wealthy, they are able to attract higher profile players. In 2021, we see countries like France, Spain, and Portugal with more players in the Premier League. These are countries where some of the best players come from. It makes sense for them to be attracted to arguably the best league in Europe.

Discussion

This dataset has provided an interesting analysis. By examining trends and comparing leagues over time, we are able to determine the health of a league currently and in the past.

As the example with the Premier League has shown, transfer fees, market valuations, player nationalities vastly differ from the beginning of this dataset and the most recent year. Inflation aside, this makes sense when we compare it with recent history. The most expensive transfers have taken place in the more recent years. Teams are more likely to pay a high price for young players. This is all due to the fact that teams and leagues tend to acquire wealth as they perform well – and the Premier League has performed well. Many leagues have two to four teams that consistently compete for the league champion spot; but the Premier League often times have more than four contenders playing well and challenging each other for the top spot. This is evident in the data as we transfer fees and market values increase from 2009 to 2021.

Another example that supports this is the Portuguese league. Although it isn't illustrated here, it sees a similar progression as the Premier League. In 2009, most transfer fees and market valuations were very low. In 2021, we see these transfer fees and market valuations take a significant jump. The Portuguese league isn't considered one of Europe's top five leagues – why did it witness such a huge jump? This can be likely attributed to the rise Cristiano Ronaldo. Ronaldo is one of football's greatest, and he's from Portugal. With his rise, he likely brought a lot of attention to Portuguese talent. We see this in the player nationalities of the Premier League in 2021. While Portugal does not make the list in 2009, they're a steady performer in 2021.

Focusing on the K-Means clustering analysis, I believe three clusters makes the most sense. Although there are seven leagues (potentially seven clusters), there are leagues that are very similar in skill level and talent. Thus, this would produce similar transfer observations. We can break down the list of leagues into three categories – the high performers, the up and coming, and the leagues that aren't quite there yet. Each league used in this analysis will fall within one of those categories and most football fans would tend to agree. Leagues like the Premier League or La Liga are the high performers. Leagues like Serie A or the Bundesliga could be considered the up and coming. While leagues such as Eredivisie or Liga Portugal could be considered at the bottom. Despite this, there is a lot of debate centered around which leagues boast the best competition and best players. This is why we see such close similarities in the clusters, regardless how many clusters we introduce to the algorithm.

In future iterations of this analysis, it would be beneficial to continue down this path of clustering. A hierarchical clustering could prove interesting in building to determine metrics that decide what sorts of transfers fall within each league. For example, some player nationalities are more dominant in specific leagues, such as English in the Premier League. A transfer observation that records a player as English has a high chance of coming from the Premier League. The same is true for the other leagues, as well. Another example is market value amounts. The top leagues tend to boast the highest market valuations. If a transfer is valued at a higher market amount, it's likely coming from one of the top leagues.

Incorporating other attributes, such as player position, into the analyses could provide some unique insight as well. Each team across each league need players in all positions; but an analysis done using player position might highlight if a specific position is favored in each league.

Overall, this analysis has concluded with multiple ways to determine the health status of a league and how it compares over time. Using the Premier League is a prime example because of the significant growth is has experienced over the last decade. In addition, patterns such as the increase in Portuguese players in the Premier League were found unexpectedly – but make tremendous sense when looking at the football environment as a whole. There are still a lot more directions this analysis could go, as discussed, and exciting patterns to uncover.