Proposal

Stats is 'Fun' - Dav King, Luke Thomas, Thomas Barker, Harry Liu

Introduction

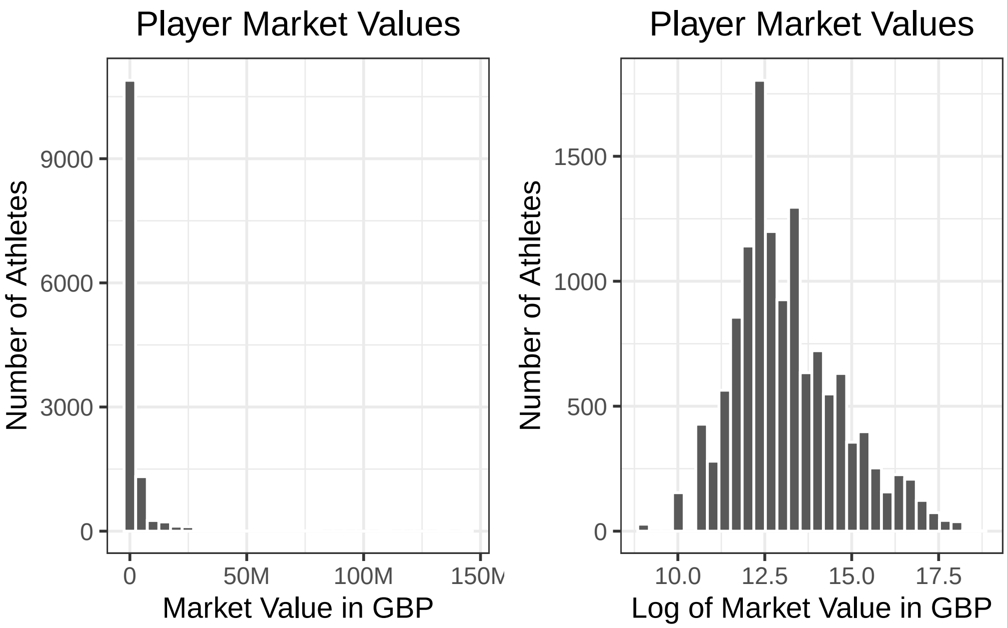
Every year, tens of thousands of soccer players are transferred on the international transfer market. Soccer clubs pay other clubs, sometimes hundreds of millions of dollars, to sign players to their team in hopes of improving club performance. These astronomical transfer prices are often determined by clubs with the help of Transfermarkt, a popular online soccer database that uses data scouts to collect data from the match sheets of soccer games. Our goal with this research is to take a deeper look into Transfermarkt's valuations of soccer players around the globe. Specifically, we are asking: how can we use a player's individual characteristics (nationality, age, position, etc.) and performance (goals, assists, minutes, red cards, etc.) to predict their market value? We predict that factors like being younger, playing as a forward, having more goals and assists, and playing more minutes will all contribute positively/increase a player's market value.

Data description

There are several different datasets present in the overall data. In the players dataset, each observation is a player and contains information such as their name, player id, country of citizenship, position, and date of birth. In the player valuations dataset, each observation is a market value of a player (of which there can be multiple for one player over time), and also includes the player id and date of the valuation. In the appearances dataset, each observation is a player's appearance in a soccer game, and it also includes the player id, number of goals scored in that appearance, number of assists in that appearance, number of minutes played in that appearance, and the number of yellow and/or red cards obtained in that appearance. As mentioned in the introduction, these data were collected by Transfermarkt's data scouts who look at the match sheets of soccer games and collect statistics like these.

Exploratory data analysis

Distribution of Response Variable: End-of-Season Market Value

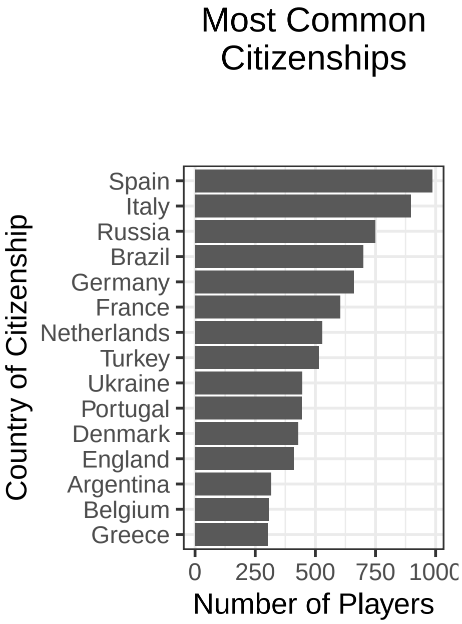


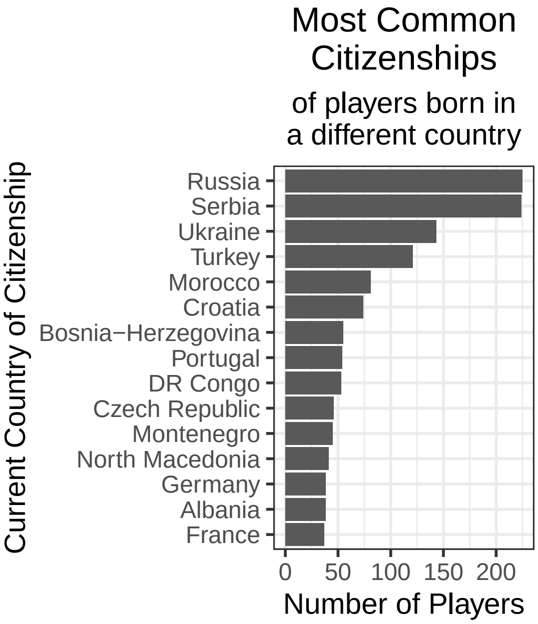
Our initial look at the univariate distribution of player market values shows a unimodal and extremely right-skewed distribution, with a range from \$9000 to \$135000000 and a peak close to \$0. They have a median at \$293000 (Q1 = \$180000, Q3 = \$720000) but a mean of \$1155740 (sd = \$3837886), which is well beyond the third quartile of the data. In fact, using the Median Absolute Deviation (Median +/- 3MAD) definition of outliers, this means that the mean, as well as all players with a market value above \$1093604, are considered outliers. The skew of 12.45 and the Kurtosi value of 262.36 both confirm what is clearly seen in the plot - this is clearly a non-normal distribution, and it may well be better considered a logarithmic distribution instead.

If we log-transform the response variable, however, it roughly follows a normal distribution. It is unimodal, with a center around 12.5 and minimal skew. It has a median at 12.59 (Q1 = 12.1, Q3 = 13.49), which is very close to the mean of 12.79 (sd = 1.33). It ranges from 9.1 to 18.72, and its skew and Kurtosi values suggest that it is not significantly different from normal. Using the very robust Median +/- 3MAD definition of outliers, it still has 231 outliers (28 below a log value of 9.183, 203 above a log value of 15.993) among its 8599 observations, but this is much closer to a normal distribution than the untransformed variable. While we will not necessarily run our regression using the log of our response variable, this is still important to note.

Distribution of Potential Predictor Variables

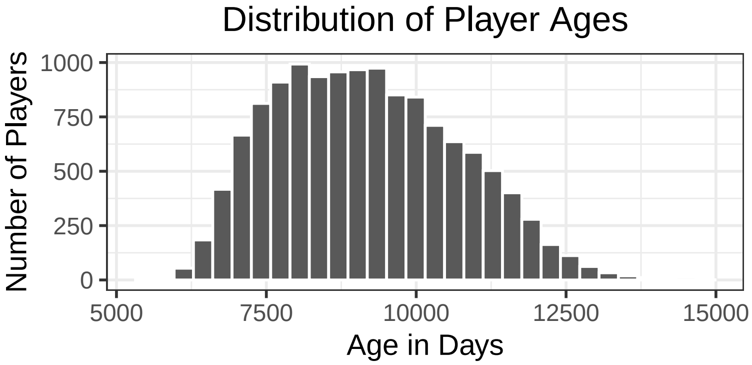
Nationality





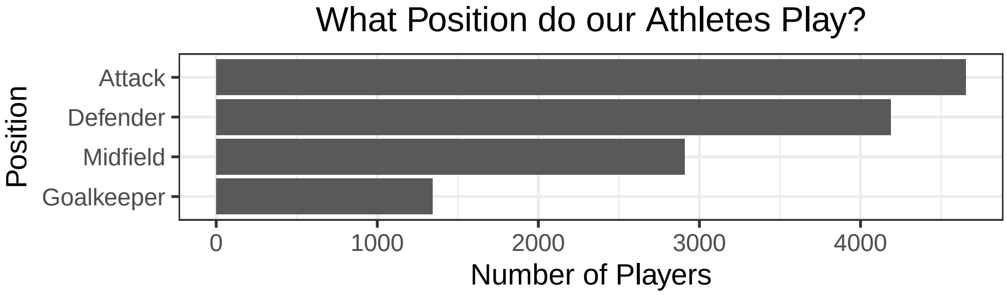
The first predictor variable to look at here is the nationality of different soccer players. The first graph shows us the 15 most common of the 160 countries in which players in this dataset were born. We see a big dominance in European players, with Spain (1006), France (904), Italy (890), Germany (807), and Brazil (718) making up the top 5 most common of these countries. The second graph shows us the 30 most common of the 141 countries in which players who were not born there hold citizenship. This graph reveals a large number of players who were likely born in the USSR, with Russia (225), Serbia (224), Ukraine (143), Turkey (121), and Morocco (81) serving as the most common countries on that list. This could be interesting to consider where players are moving to, and whether that has any impact on their market value. As a note, in order for nationality to be relevant in our analysis, we will need to collapse it down to a much smaller number of countries.

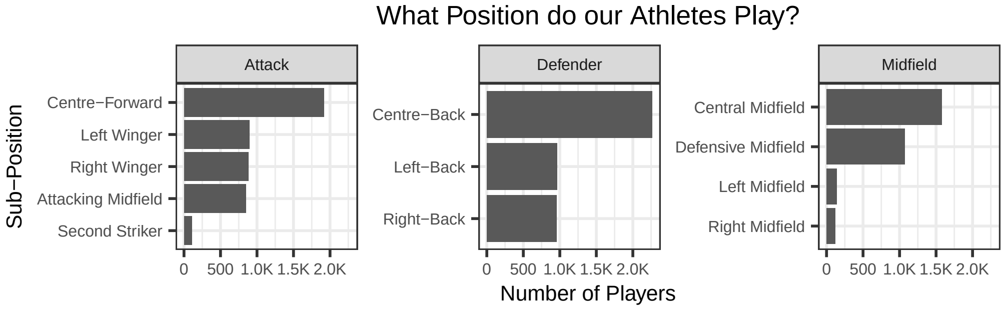
Age



The next predictor to look at is a player's age. Looking at the histogram, it appears to be mostly normally distributed with a slight right-skew, unimodal, with a center around 9000 days (~24.6 years) and a range from 5486 days (~15 years) to 15346 days (~42 years). The data have a median at 9246 days (~25.3 days, Q1 = 8030 days/~22 years, Q3 = 10515.25 days/~29.1 years), with a mean at 9405.33 days (~25.7 years, sd = 1687.5 days / ~4.6 years). The skew (0.38) and Kurtosi (-0.56) values do not suggest that the data are significantly non-normally distributed. Using the robust Median +/- 3MAD defnition of outliers, only 4 players are outliers (those above 14892.982 days / ~40.8 years).

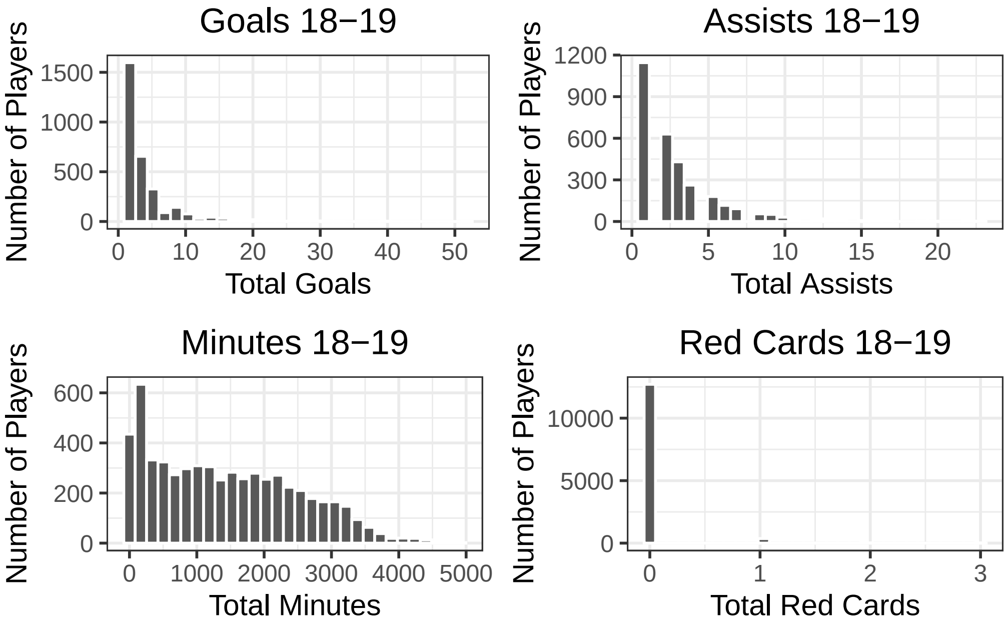
Position





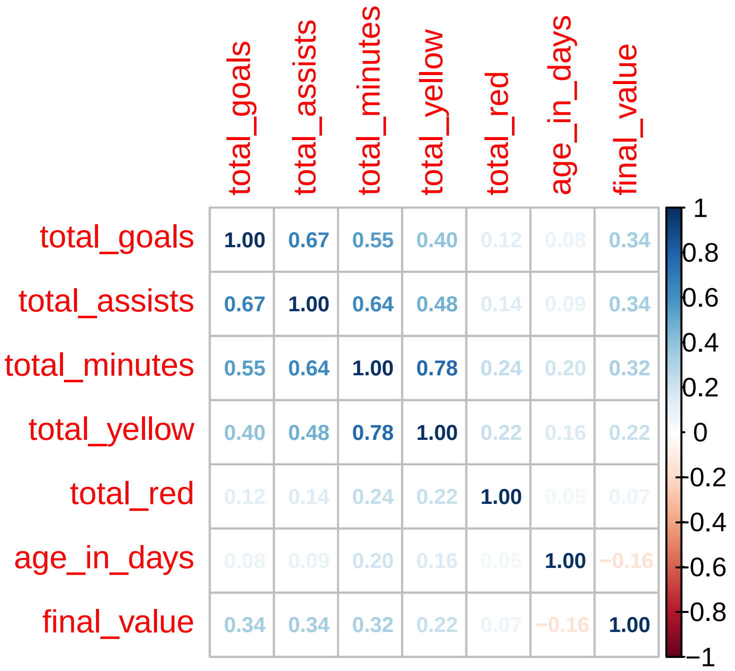
The next predictor we will consider is the position that someone plays. The first plot simply allows us to look at the general position that an athlete plays - attack, midfield, defender, or goalkeeper. The most common position for an athlete to play is Attack (4654), followed by Defender (4189), Midfield (2909), and Goalkeeper (1342). This roughly corresponds to the number of players at each of those positions who are on the field at any given time. The second plot allows us to look at the most common sub-positions within each position. Far and away the most common position is centre-back (2267), followed (unsurprisingly) by centre-forward (1915) and central midfield (1577). Between the two of these variables, given that they do not have too many different categorical levels, we should be able to use this as an interesting variable in the prediction of market value (as well as a potential interaction with goals, assists, and other variables).

On-Field Performance



The remaining predictors that we will consider are a player's personal stats. Given that all of these variables are heavily right-skewed and any observations greater than 0 were considered outliers, we elected to present the graphs only when the values were present (i.e., greater than 0); however, we have considered the data both with and without these values. The first plot shows us the total goals scored by players who scored any (removing 10019 players who did not score a goal), which shows a unimodal but incredibly right-skewed distribution, centered around 2 with a range from 1 to 51. The median is 2 (Q1 = 1, Q3 = 5), with a mean at 3.83 (sd = 4.36). The skew (3.16) and Kurtosi (14.94) values verify that this is still a skewed distribution. The second plot shows us the total assists earned by players (excluding 10019 players who did not earn any), which shows a unimodal but still right-skewed distribution, again centered around 2 with a range from 1 to 23. The median is 2 (Q1 = 1, Q3 = 4), with a mean at 2.94 (sd = 2.7). The skew (2.39) and Kurtosi (7.92) values reflect that this remains a skewed distribution. The third plot shows us the total minutes played (excluding 7222 players who did not play), which shows a unimodal and right-skewed (but less so) distribution, with a mode around 250 and a range from 1 to 4913. The median is 1168.5 (Q1 = 348, Q3 = 2141), with a mean at 1338.91 (sd = 1066.15). The skew (0.55) and Kurtosi (-0.63) values actually suggest that this is a reasonable distribution of data, and that it is not concerningly distributed. The final plot shows us the total number of red cards that players received. This reflects 15,213 players receiving 0 red cards, while 399 received 1, 23 received 2, and only 3 received 3. Given this lack of distribution, we may be better served by considering the total number of yellow cards received by a player instead.

Numeric Predictor and Response Relationships



The plot above shows the correlations between our key numeric predictor variables and our response variable of final market value. The most noticeable correlation coefficients are between a player's total minutes played, and their total goals, assists, and yellow cards, producing values of 0.55, 0.64, and 0.78 respectively. These each indicate a moderately strong correlation, which logically makes sense, since a player who plays more minutes would expect to accumulate more stats. Another noticeable value is the moderately strong correlation of 0.67 between a player's total goals and total assists, since these stats are both accumulated more by attacking players and less by defensive players. From this information, it may be interesting to explore the potential interaction effect between a player's position and their total goals and assists, since attackers are expected to have more goals and assists than defenders. Lastly, there does not appear to be strong correlations between total red cards, age, or final market value and any of the other numeric predictors.

Important Data Transformations

While many of the necessary data transformations have been discussed already, they will be summarized here. We have already transformed our data to consider market value only at the end of the 2018-19 season, accumulated stats across said season, and transformed our data to contain one observation per player. We will need to collapse our categorical variable of nationality into a meaningful number of categories. Additionally, we may well be served by limiting our analysis only to players who actually played in the 2018-19 season, or at least creating some sort of a dummy variable that accounts for this.

Analysis approach

For this analysis, the response variable is final_value, a player's final market value at the end of the 2018-19 season. Potential predictors include country of birth, country of citizenship, position, sub-position, dominant foot, height, age, and player statistics from the 2018-19 season, such as total goals, assists, minutes, yellow cards, and red cards. We plan to use multiple linear regression to conduct our analysis.

Data dictionary

The data dictionary can be found here.