Data Analysis : Term Project

Rauhan Nazir

Overview and Goal

There is a common perception around the footballing world that when it comes to the market value of English football players, they are over-rated compared to footballers from other countries. There are quite a few examples that come to my mind that support this perception. One would be the transfer of Harry Maguire from Leicester City to Manchester United. The goal of this assignment is to find out or come closer to finding out whether clubs actually do pay a premium when it comes to buying English players or is it just a misconception. So the y variable from the data that I chose is the market value of the players and the x variable is the country that they belong to. While there are several other confounding variables that were introduced to make sure that the relation between x and y variables was not an exaggerated one and it was as close to reality as possible.

Quality of Data and Data Munging

This data set contained information about the top 500 highest valued players in the market and it is a quite representative data set as there is no selection bias. There is almost no chance that the variables contain any measurement error as most of them are categorical variables such as the name of the players or what country they belong to or what club do they represent. Numerical variables such as goals, number of assists and the cards they received while playing are also the kind of data that do not have measurement error as it is quite factual. The only variable that might contain measurement error is actually the market value, as it is an estimation and there is no way to find out the actual value unless these players actually make a transfer. However, I am quite confident about these numbers as well as this value took into account several metrics such as Future prospects, Age, Performance at the club and national team, Level and status of the league, both in sporting and financial terms, Reputation/prestige, Development potential,League-specific features, Marketing value, Number & reputation of interested clubs, Performance potential, Experience level, Injury susceptibility, Different financial conditions of clubs and leagues, General demand and "trends" on the market,General development of transfer fees, External factors such as the corona virus pandemic and its consequences.

The data set that I used did not require much cleaning, however there were some changes that I made.

Adding Goals per Match and Assists per Match Columns

The columns that I added were the *Goals per Match*(Goals/Matches) and *Assists per Match*(Assists/Matches). Instead of using the absolute values of goals and assists, it made more sense to normalize these variables by divinding them by number of matches each player played. Giving us more accurate measure of their performance.

Changing column names

One minor thing that I did was change the column name of Market Value of Players for simplicity as it contained special characters.

Creating a binary variable for English Players

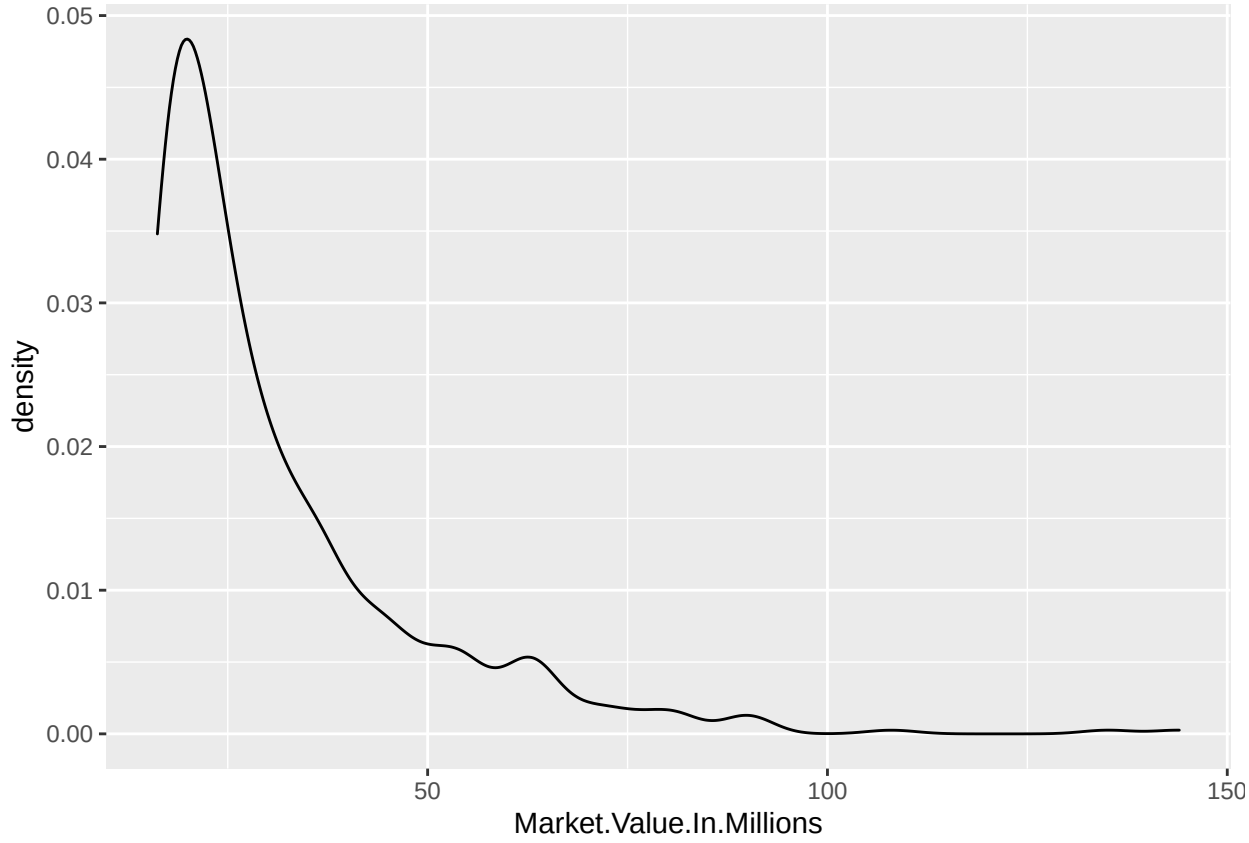
To answer the main question about the market value of English players I decided to assign binary values to the column of country, where it is going to be 1 if the player is English and 0 for all of the other nationalities, which is the x variable that our focus is one . Market Value of players will be regressed on this binary variable.

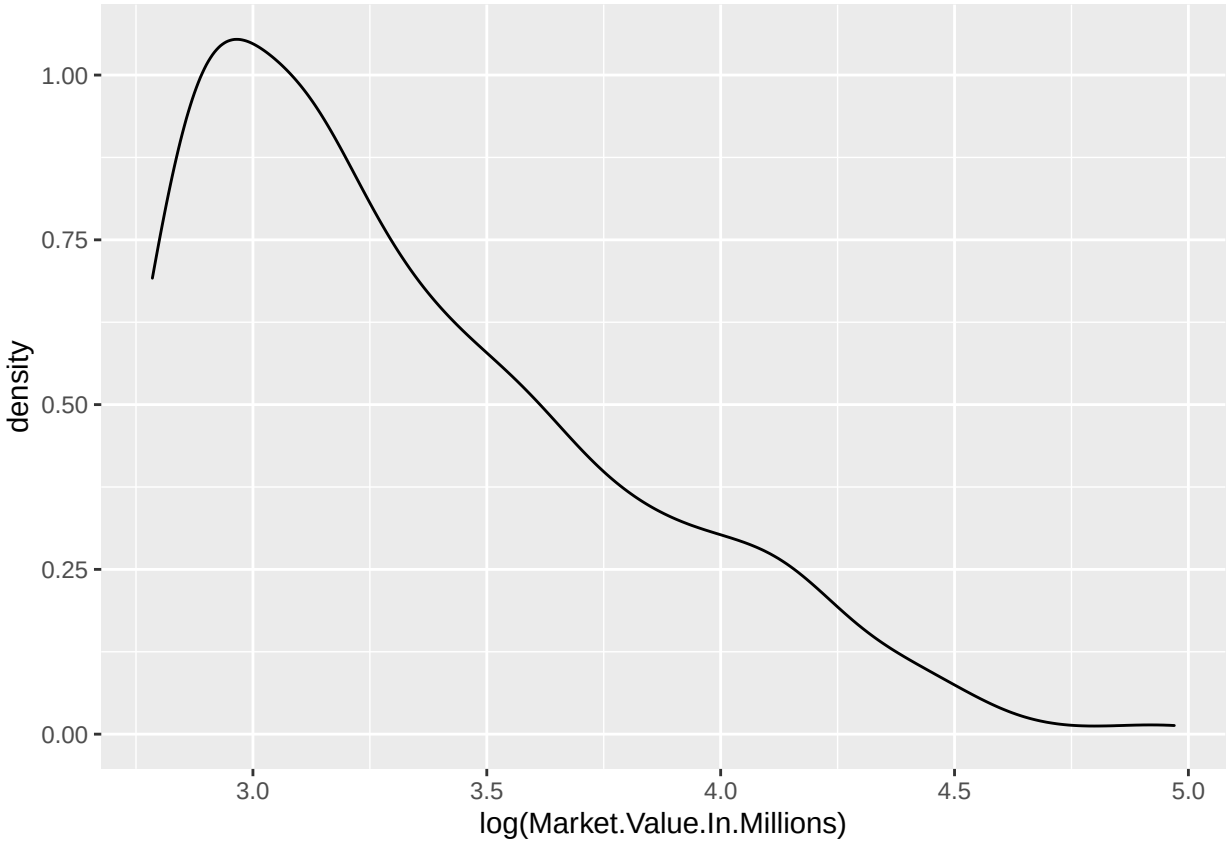
Changing the position of the players

One other thing that I decided to do while cleaning the data was to take right wingers and left wingers as just wingers as there is no major difference between the two other than the side. Same was done for right backs and left backs as I just took them as full backs. Even though some teams have a preference to attack more from one of the sides depending on the formation and the tactics however for the kind of analysis that I am doing, it makes more sense to treat them similarly. Finally it was conscious decision to not treat all midfielders the same as it is more likely that certain midfielders are more valued than others due to their position, for instance attacking midfielders could be more valued than defensive midfielders.

Checking for Skewness and deciding what variable to use (log/absolute)

The other decision that I had to make was to decide whether to use *absolute values* of the market value of players or take the *log values* instead. For that I checked the distribution through a geom density curve. The absolute values were right skewed so for that reason I took the log and the distribution of that was relatively more normally distributed.Hence that was the variable that I decided to go ahead with and use in my regressions.





Data Summary Table

I also created a data summary table to know more about the values and also the data summary skim function that gives more insight into the distribution of variables.

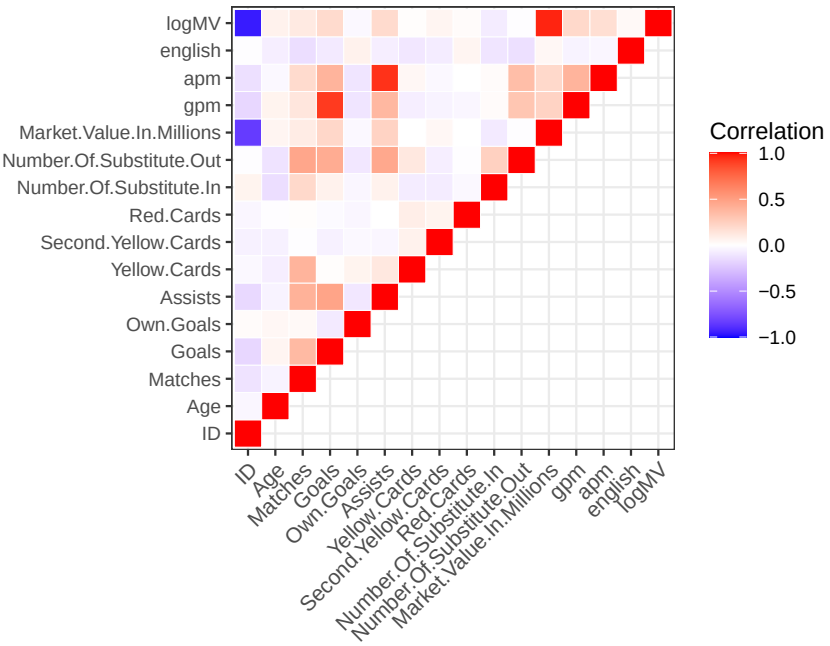
Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P5	P95
Matches	12.40	13.00	4.34	0	24	4.95	18.00
Goals	2.16	1.00	2.88	0	23	0.00	8.00
Own Goals	0.03	0.00	0.17	0	1	0.00	0.00
Assists	1.51	1.00	1.85	0	12	0.00	5.00
Red Cards	0.05	0.00	0.21	0	1	0.00	0.00
Assists	1.51	1.00	1.85	0	12	0.00	5.00
Yellow Cards	1.59	1.00	1.45	0	7	0.00	4.00
Market Value Millions	31.54	25.20	17.58	16.20	144.00	16.20	63.22
Goals per Match	0.17	0.11	0.20	0.00	1.35	0.00	0.54
Assists per Match	0.11	0.08	0.13	0.00	0.69	0.00	0.38
Log of MArket Value	3.34	3.23	0.45	2.79	4.97	2.79	4.15

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
ID	500	0	249.5	144.5	0.0	249.5	499.0	
Age	20	0	25.0	3.2	16.0	25.0	36.0	
Matches	24	0	12.4	4.3	0.0	13.0	24.0	
Goals	17	0	2.2	2.9	0.0	1.0	23.0	
Own.Goals	2	0	0.0	0.2	0.0	0.0	1.0	
Assists	12	0	1.5	1.9	0.0	1.0	12.0	
Yellow.Cards	8	0	1.6	1.4	0.0	1.0	7.0	
Second.Yellow.Cards	2	0	0.0	0.2	0.0	0.0	1.0	
Red.Cards	2	0	0.0	0.2	0.0	0.0	1.0	
Number.Of.Substitute.In	14	0	2.4	2.5	0.0	2.0	13.0	
Number.Of.Substitute.Out	18	0	3.7	3.3	0.0	3.0	20.0	
Market.Value.In.Millions	34	0	31.5	17.6	16.2	25.2	144.0	
gpm	77	1	0.2	0.2	0.0	0.1	1.4	
apm	62	1	0.1	0.1	0.0	0.1	0.7	
english	2	0	0.1	0.3	0.0	0.0	1.0	
logMV	34	0	3.3	0.4	2.8	3.2	5.0	

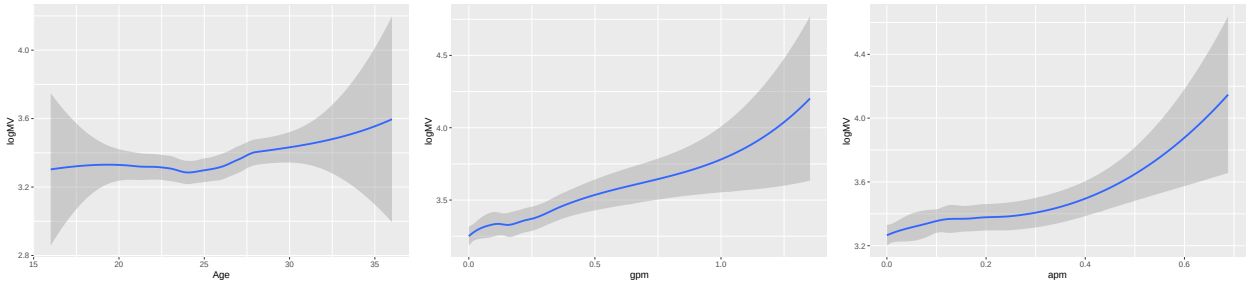
Correlation Matrix

Before running the regressions I made a correlation matrix to explore the correlations between different variables. At first glance it seems like there is no significant relation between a player being English and Log of Market value. While variables like Goals per match and assists per match has a much stronger correlation. However this will be explored further with the regressions to be more sure about the relation between nationality and log of market value.



Deciding whether to use Splines

Final thing to consider before running the regressions was figuring out if I needed knots at different points based on the relations of x variables (numeric) with the market value in different ranges. For that I used loess. It showed that there was no need for any knots as there was no significant change in the trend throughout.



Regressions

For the first one I regressed log of Market Value on English (Binary variable). There was no significant relation between the two variables as 0 was in the range of the confidence interval. However, to further explore this, I added confounding variables, adding one after each regression to find out what the exact impact of controlling for other x variables is and the impact of them on the regression model. The relation between the goals per match and logMV was highly significant, for goal per match to be higher by one unit, on average the logMV is going to be higher than almost 45%, as shown in the results of regression 2. For every confounding variable added, adjusted R squared increased, so it made sense to include them in the regression. The most significant increase came after adding the Club as a factor in the regression model. In the final regression, goals per match were significant at a confidence interval of 99% while the assists per match were significant at a confidence interval of 95%, while the relation of logMV was still insignificant. Below is the final regression that was run.

$$MarketValue(log) := \beta_0 + \beta_1 English + \beta_2 Gpm + \beta_3 Apm + \beta_4 Age + \beta_5 Position + \beta_6 Club$$

Conclusion

While our analysis did not support the perception that English players are over valued and that you have to pay a premium, it does not provide us with conclusive evidence. I say that because firstly, the data set only contained data on the players in 2021, so it does not analyze the trends for a long enough time aand secondly, as mentioned earlier that there could be a measurement error in the market value of players as it is an estimation and it could be the case that some important variable was not taken into account while estimating this value. The perception is so strong and the examples are so frequent that follow the trend of English players being over valued, this hypothesis is worth exploring further to move closer to reality.

Appendix

Exhibit 1

The final regression results

Table 2: Regression Model Summary

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
(Intercept)	3.33**	3.26**	3.23**	2.97**	2.94**	3.41**
	(0.02)	(0.03)	(0.03)	(0.17)	(0.18)	(0.22)
english	0.04	0.05	0.05	0.06	0.07	0.04
	(0.07)	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)
gpm		0.45**	0.35**	0.34**	0.51**	0.55**
		(0.12)	(0.13)	(0.13)	(0.15)	(0.17)
apm			0.38	0.40*	0.44*	0.41*
			(0.20)	(0.20)	(0.22)	(0.21)
Age				0.01	0.01	−0.01
				(0.01)	(0.01)	(0.01)
as.factor(Position)Central Midfield					0.07	0.03
					(0.09)	(0.09)
as.factor(Position)Centre-Back					0.07	0.02
					(0.10)	(0.09)
as.factor(Position)Centre-Forward					−0.11	−0.04
					(0.10)	(0.09)
as.factor(Position)Defensive Midfield					0.07	0.11
					(0.11)	(0.10)
as.factor(Position)Full-Back					−0.03	−0.11
					(0.10)	(0.09)
as.factor(Position)Goalkeeper					0.08	0.02
					(0.15)	(0.13)
as.factor(Position)Left Midfield					−0.17	−0.05
					(0.16)	(0.18)
as.factor(Position)Right Midfield					−0.25*	−0.10
					(0.12)	(0.22)
as.factor(Position)Second Striker					−0.03	−0.24
					(0.20)	(0.18)
as.factor(Position)Winger					−0.02	−0.03
					(0.09)	(0.09)
as.factor(Club)ACF Fiorentina						−0.22
						(0.15)
as.factor(Club)Ajax Amsterdam						−0.26
						(0.15)
as.factor(Club)Al-Rayyan SC						−0.50**
						(0.15)
as.factor(Club)Arsenal FC						−0.00
						(0.15)
as.factor(Club)AS Monaco						−0.14
						(0.15)
as.factor(Club)AS Roma						0.02
						(0.15)
as.factor(Club)Aston Villa						0.02
						(0.13)
as.factor(Club)Atalanta BC						−0.09
						(0.13)
as.factor(Club)Athletic Bilbao						−0.18
						(0.12)
as.factor(Club)AtlÃ©tico de Madrid						0.50**
						(0.17)
as.factor(Club)Bayer 04 Leverkusen						0.05
						(0.18)
as.factor(Club)Bayern Munich						0.52**
						(0.14)
as.factor(Club)Bologna FC 1909						−0.54**
						(0.13)