Do football players get more or less valuable as they get older?

Adam Jozsef Kovacs

2021-12-22

Introduction

Nowadays it is more and more common in the world of football to apply rigorous data analysis for both performance analysis and recruitment advisory. The aim of this project, related to the second goal, is to uncover the relationship between the age of football players (x variable) and their estimated market values (y variable). The relevance of this question is twofold: advances in medicine and coaching techniques on the one hand, earlier maturing youngsters on the other hand changed the length and shape of footballers careers urging scouts and analysts to reconsider their practices. To understand the pattern of association between age and market value, several linear regression models are estimated including multivariate ones that consider a variety of potential confoundres (e.g. position, nationality, goals, assists etc. as z variables).

Data

The dataset analyzed contains footballers in the European big 5 championships and was constructed from two sources. For the two main variables, market value and age as well as basic information on players (e.g. nationality, team, preferred foot), **transfermarkt.com** is the source of data. As for statistics on the performance of players (e.g. goals, assists etc.), **fbref.com** is the source of data. More information on how the dataset was built is available in the Data section of the Appendix.

In a bit more detail about the variables: market value and player information downloaded from transfermarkt contain data of players before the beginning of the 2021/22 season. As a consequence, football statistics for the preceding season are considered from fbref (they contain up-to-date information potentially driving the valuation of the players). When merging the datasets from the two sources, inner joins were used for most variables (lost observations did not play a single game in the 2020-21 season). The only exceptions were goalkeeper statistics where left join was applied and NA values of outfield players were replaced by zeros (goalkeepers also have 0 for e.g. goals). An issue in data preparation were players that changed clubs during the season. To avoid duplicate records, either the sum or the mean of their performance measures in their two clubs were considered (based on if it is a total or a ratio variable).

Further data transformation steps included scaling market value of players to million EUR instead of simply EUR for easier interpretation. Also, from the number of matches and minutes played, the combined variable of (average) minutes per game was created. For a more detailed description of all variables please see the Data section of the Appendix. Finally, a filtering of the data was also applied to improve data quality: Those players who did not make at least 2 appearances and a total of 90 minutes over the entire past season were dropped, as we consider their valuation (and football stats) unreliable.

Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Age	26.90	27.00	4.22	17.00	40.00	20.00	34.00
Market value (mn EUR)	13.46	7.00	16.93	0.20	160.00	0.90	50.00
Minutes per game	66.18	69.25	19.09	10.27	90.00	29.70	90.00

Table 1: Descriptive statistics

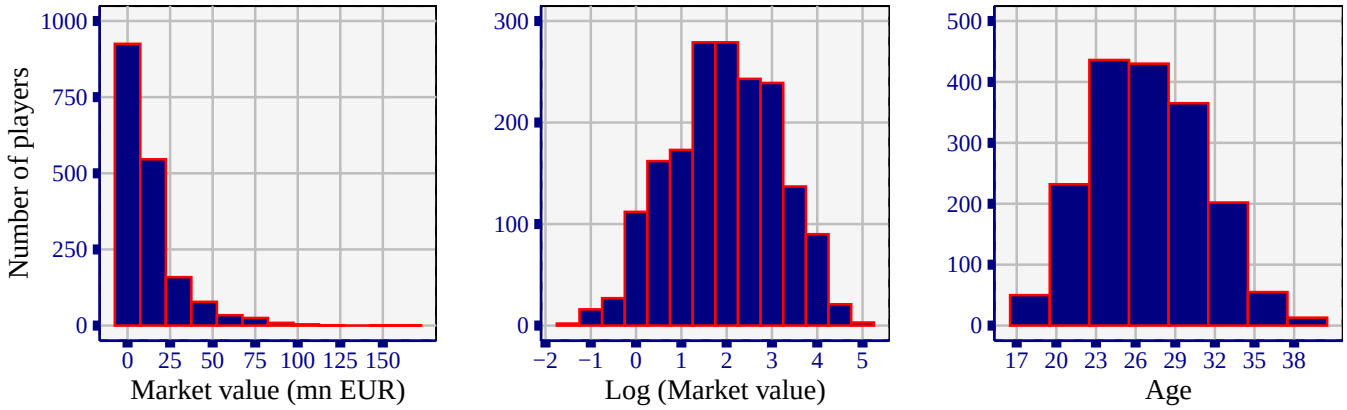
Table 1 contains descriptive statistics on the age (x) and market value (y) variables as well as the minutes per game for the players. The number of observations in the dataset is 1783. Average minutes per game is presented here as a

potentially very important covariate. Players are valued based on their performance on the pitch and for that they have to play. Its value in the data ranges from around 10 to the maximum of 90. The mean and median numbers are both around 65-70 minutes. This means that the variable has a close to normal distribution.

The age of players ranges from 17 to 40 and has its median and mean value very close to each other at around 27 years. The distribution is thus close to a normal distribution. As for market value, it ranges from 200 thousand to 160 million euros. Its mean at around 13.5 million euros is much higher than the median at 7 million. This shows that the distribution is left skewed (has a long right tail). It is also notable that the 160 million valuation is an extreme value given that the 95th percentile is almost half this value (90 million).

The distribution of age and market value are also visualized on Figure 1. We can see the same patterns that we could read out from the descriptive statistics. Based on the distribution of market value (skewed distribution with long right tail) log transform of the y variable is taken as the main explained variable. Looking at the distribution of the log transform in the middle, it is close to a normal distribution. Besides the statistical reasoning, the fact that it is likely affected by age in multiplicative, not additive ways also points toward this transformation.

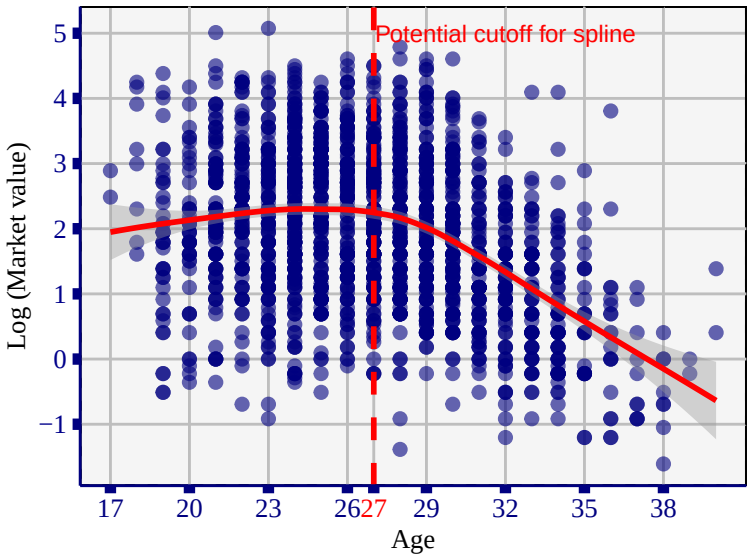
Figure 1: Distribution of main variables



Modelling

To grasp the relationship between x and y variables, a non-parametric regression (lowess) is visualized on Figure 2.

Figure 2: Lowess on age and log market value



What we can infer from this chart is that until the age of around 27, there is a slight positive relationship between

age and market value. From then on, there is a negative relationship between the two variables. Based on this chart, in a linear spline model, a knot at 27 years of age looks reasonable.

Regarding control variables, all player information including nationality, team (instead of league that would capture the same, yet less information), position and preferred foot are used as factor variables (see frequencies on Figure 4 in the Appendix). As for football performance statistics, descriptives and scatters with the outcome variable are on Figrure 3 of the Appendix. A correlation matrix is also analyzed to identify useful covariates. Based on the matrix, all variables are significantly correlated with market value, and the majority is practically uncorrelated with age. Some of them however have high multicollinearity. As a result, *shots on target per 90 minutes* and *shots per 90 minutes* are dropped since the complex *expected goals per 90* variable already captures the variance of these. Goalkeeper statistics are correlated and with themselves and age, but are kept to capture the relationship for them as well (rest of the statistic completely uninformative for keepers). As for defensive stats, they are also kept because though correlated with each other, they contain different information and are uncorrelated with age.

We estimate a total of 5 models to learn about the general pattern of association between the age of players and their market value. The first is a simple linear regression with age as the only explanatory variable. We can interpret it as being one year older is associated with a 9.9% lower market value on average in our data.

In the second model, we introduce a linear spline with the cutoff at 27 years as discussed in the descriptive statistics part. This model confirms the pattern on the scatter: below 27 years of age, being one year older is associated with a 3.3% higher expected market value. After turning 27, however, an additional year goes together with on average 22% lower market value for the players in our dataset.

In the third model, we include the average minutes per game as a covariate. It is statistically significant even at 1% level and shows that more game time is associated with higher market value. Its effect on the age variables is also notable: it takes away the significance of age below 27 years.

In the fourth model, we also include the factor variables on basic player information. This has a huge effect on the goodness of fit of the model, R^2 rises up to 74% (from 26%). Finally, we also include all the performance metrics from the preceding season in model 5.

[Dependent variable: Log (Market value)]

	(1)	(2)	(3)	(4)	(5)
Intercept	4.617***	1.469***	1.261***	2.256***	1.855***
	(0.1803)	(0.3272)	(0.3065)	(0.3413)	(0.3949)
Age	-0.0994***				
	(0.0066)				
Age (<27)		0.0335*	-0.0059	-0.0157	-0.0247**
		(0.0133)	(0.0128)	(0.0098)	(0.0085)
Age (>=27)		-0.2200***	-0.2193***	-0.1949***	-0.1997***
		(0.0112)	(0.0111)	(0.0078)	(0.0068)
Minutes per game			0.0181***	0.0259***	0.0106***
			(0.0013)	(0.0011)	(0.0016)
Player info	No	No	No	Yes	Yes
Attacker stats	No	No	No	No	Yes
Midfielder stats	No	No	No	No	Yes
Defender stats	No	No	No	No	Yes
Goalkeeper stats	No	No	No	No	Yes
Observations	1,783	1,783	1,783	1,783	1,783
R2	0.12181	0.18683	0.26401	0.74373	0.80216

* Heteroskedasticity-robust standard errors in parantheses
* Signif. codes: '***': 0.001 '**': 0.01 '*': 0.05

Table 2: Models to uncover relation between age of players and their market value

The preferred model is the fifth one. The explanatory power of this model is the best with R^2 above 80%. Furthermore, our main explanatory variable (age) is statistically significant at even 1% level for both tiles of the spline. The model in a mathematical format:

$Log(Market\ value) = 1.855\ -0.025\ (Age < 27)\ -0.2\ (Age \geq 27) + \delta Z$

where Z stands for the control variables, which includes minutes per games played, competition, team, position, nationality, preferred foot and a selection of football statistics relevant for attackers, midfielders, defenders and goalkeepers.

The interpretations of the estimated parameters are the following: When every covariate is zero, the log market value of footballers is expected to be 1.855. Keeping all other covariates in the model constant, each time a footballer gets one year older below the age of 27, they are expected to lose around 2.5% of their market value. As for footballers that turned 27 already, keeping all other covariates in the model constant, when they get one year older, players are expected to lose around 20% of their market value.

Robustness check, generalization and external validity

To ensure that our results are robust, we use heteroskedastic robust standard errors. With their help, we check if the main parameters are statistically significantly different from zero. We run two two-sided hypothesis tests for the variables of the linear spline of age:

$$H_0 := \beta_1 = 0 \; and \; H_0 := \beta_2 = 0$$
$$H_A := \beta_1 \neq 0 \; and \; H_A := \beta_2 \neq 0$$

For age below 27 years old, the t-statistic is -2.92, for above 27 years old, it is -29.21 which are both higher than 1.96 in absolute value. Below 27 years, the p-value of the test is 0 and above it is also 0. The 95% confidence interval of β_1 is [-0.04, -0.01], while that of β_2 is [-0.21,-0.19], both of which do not contain zero. This means that we can refuse the null-hypothesis in both cases, they are statistically significantly different from 0. Furthermore, we can also claim with 95% confidence that holding all other covariates in the model constant, below 27 years of age being one year older is associated with 4% to 1% lower market value in the general pattern represented by our data. And we can also claim with 95% confidence that holding all other covariates in the model constant, above 27 years of age being one year older is associated with 21% to 19% lower market value in the general pattern represented by our data.

In order to discuss external validity, it is important to return to the nature of the data. The dataset analyzed has good internal validity as it has very good coverage of the players in the top 5 leagues (for the exact numbers see Figure 4 of the Appendix). But this is no guarantee for external validity. To assess that, two approaches could be taken: The first would be to test the model for players of other leagues. If the model performs well on e.g. the Dutch championship as well, then it has high external validity. The second approach would be running the model on other years. If it performs well for 2019 as well, that would also indicate high external validity. Unfortunately, these would require a lot of time-consuming data preparation and are out of scope for this research. Personally, I would expect the model to have high external validity for other championships.

Summary and conclusion

In this project, the relationship between age and market value of players was analyzed with regression analysis using a dataset of players in the top 5 European leagues. Findings suggest that before turning 27, players in general tend to be valued slightly higher by the market as they get older. However, applying multivariate regression we have shown that among two very similar players (control for position, goals, nationality etc.), the players that is one year older is valued on average at 2% less. After the age of 27, this effect is even more drastic, the player that is one year older is valued on average at 20% less.

Though the model giving these parameters explains 80% of the variation in the market value and thus brings us close to causality, stating that there is a causal link from x to y would be irresponsible. Not only because it is observational data, but there are further confounders not controlled for, two of the most notable being former clubs and performance (beyond one season). Also, by making the initial restriction of at least 2 matches and 90 minutes, young players yet to prove themselves but having great potential were disproportionately excluded.

Revisiting the research question, there are two main takeaways from this analysis. First, the market values potential and while gaining experience is important, if there are two players with the same profile, the market will value higher the one with more seasons left to play in his career. Second, footballers tend to reach their peak at around 27 years of age and the downslide in market value from then on is drastic. Thus, the likes of Ibrahimovic or C. Ronaldo seem to be the exception rather than the rule.

Appendix

Data

To get the data from the two sources presented in the main text, Jason Zivkovic's R package called **worldfootballR** is used. The worldfootballR package has a function to download the valuations of players from transfermarkt (get_player_market_values) as well as a huge collection of these urls matched to the corresponding urls of players on fbref (player_dictionary_mapping). Finally, a function of the package that is written to query data from fbref on football stats (fb_big5_advanced_season_stats) was also used. Further information on how the worldfootballR package works other than the github page is available **here** and **here**.

During the data preparation phase, from the variables available for players from both transfermarkt and fbref, only a selection of variables were kept that are potential confounders. From transfermarkt, the following variables were used:

- **player_market_value_euro**: market value of player
- **player_age**: age of player
- comp_name: competition in which player plays
- squad: team in which player plays
- player_name: name of player
- player_position: primary position player plays
- player_nationality: nationality of player
- player_foot: preferred foot of player

From the fbref data, the following variables were selected:

- matches: matches played in the season
- minutes: minutes played in the season
- **minutes_per_game**: average minutes played per matches played during the season
- Attacking stats (most relevant for midfielders and attackers)
  - goals_per90: goals scored per 90 minutes
  - assists_per90: assists provided per 90 minutes
  - xg_per90: expected goals generated per 90 minutes
  - xa_per90: expected assists generated per 90 minutes
  - shots_per90: shots taken per 90 minutes
  - shots_on_target_per90: shots taken that were on target per 90 minutes
- Midfielder stats
  - completed_passes_total: total completed passes over the season
  - completed_passes_percent: ratio of passes completed from all attempted
  - progressive_passes: number of passes that move the ball towards the opponents goal at least 10 yards
- Defensive stats
  - tackles: number of tackles over the season
  - blocks: number of blocks over the season
  - interceptions: number of interceptions over the season
  - pressures: number of times pressure placed on opponent over the season
- Goalkeeper stats
  - goals_against_per90: goals conceded per 90 minutes
  - save_percent: ratio of saves made to total number of shots faced
  - clean_sheet_percent: ratio of games where goalkeeper did not concede

The raw data (and all the codes) are available in the github repository of the project **here**.

Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Goals per 90	0.13	0.06	0.18	0.00	1.50	0.00	0.52
Assists per 90	0.09	0.05	0.12	0.00	1.10	0.00	0.31
Expected goals per 90	0.13	0.06	0.16	0.00	1.16	0.00	0.45
Expected assists per 90	0.09	0.06	0.08	0.00	0.57	0.00	0.25
Shots per 90	1.13	0.81	0.97	0.00	5.66	0.00	3.00
Shots on target per 90	0.38	0.23	0.40	0.00	2.53	0.00	1.19
Completed passes	699.09	608.00	493.21	8.00	2598.00	85.10	1631.90
% of passes completed	78.67	79.40	8.62	29.25	96.40	64.01	90.70
Progressive passes	56.75	46.00	48.34	0.00	317.00	0.00	148.90
Tackles	28.38	24.00	23.33	0.00	160.00	0.00	74.00
Blocks	25.23	22.00	19.03	0.00	113.00	0.00	60.00
Interceptions	18.32	15.00	15.98	0.00	92.00	0.00	49.00
Pressures	236.14	216.00	164.87	0.00	958.00	4.00	548.80
Goals conceded per 90	0.11	0.00	0.41	0.00	3.33	0.00	1.26
% of shots saved	5.51	0.00	18.74	0.00	100.00	0.00	66.70
% of matches clean sheet	2.06	0.00	8.20	0.00	100.00	0.00	21.60

Table 3: Descriptive statistics of covariates

Figure 3: Scatter of covariates and log market value

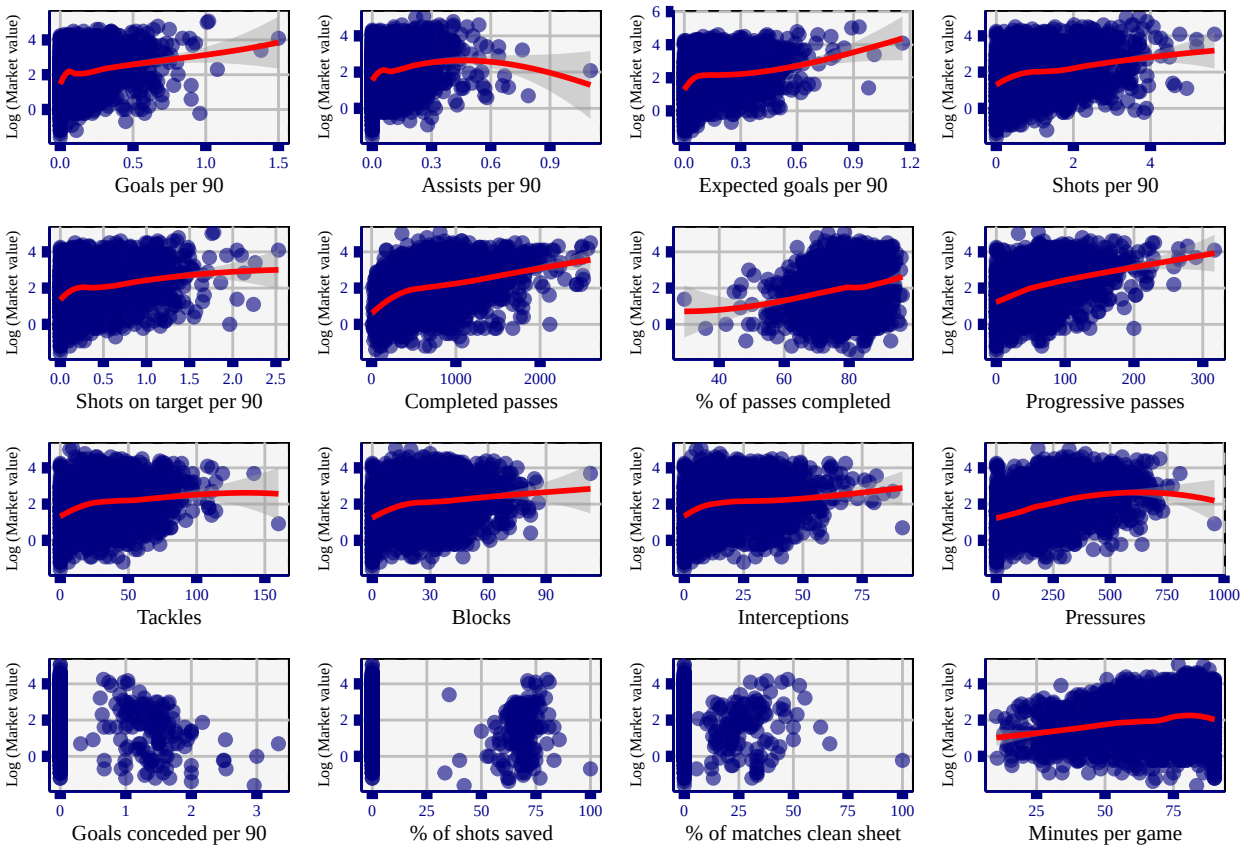
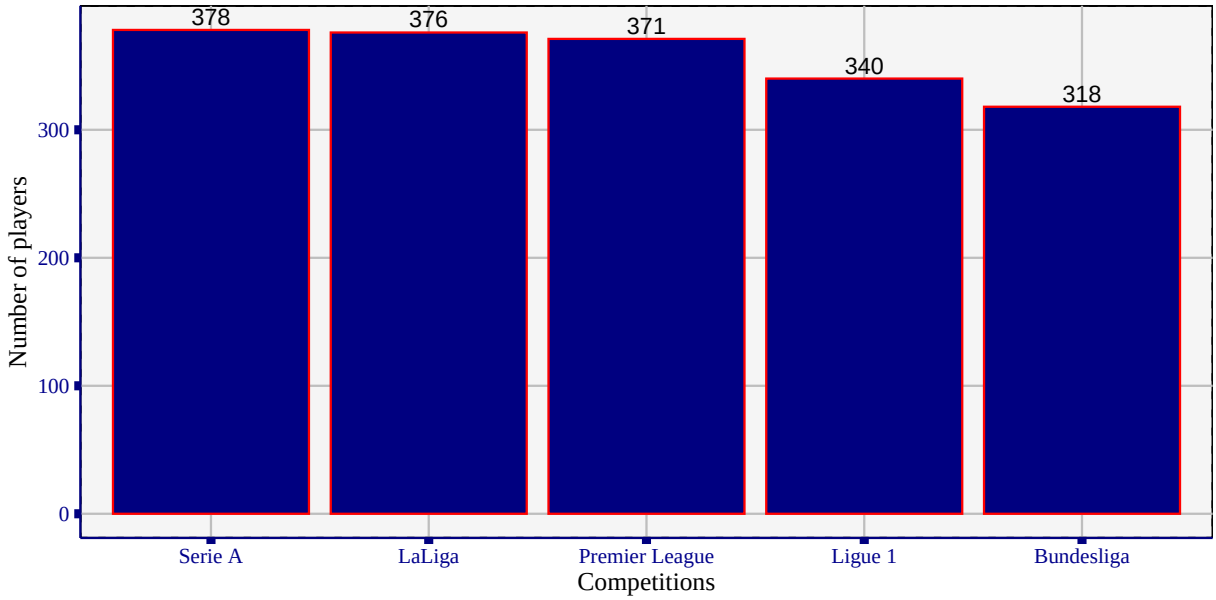
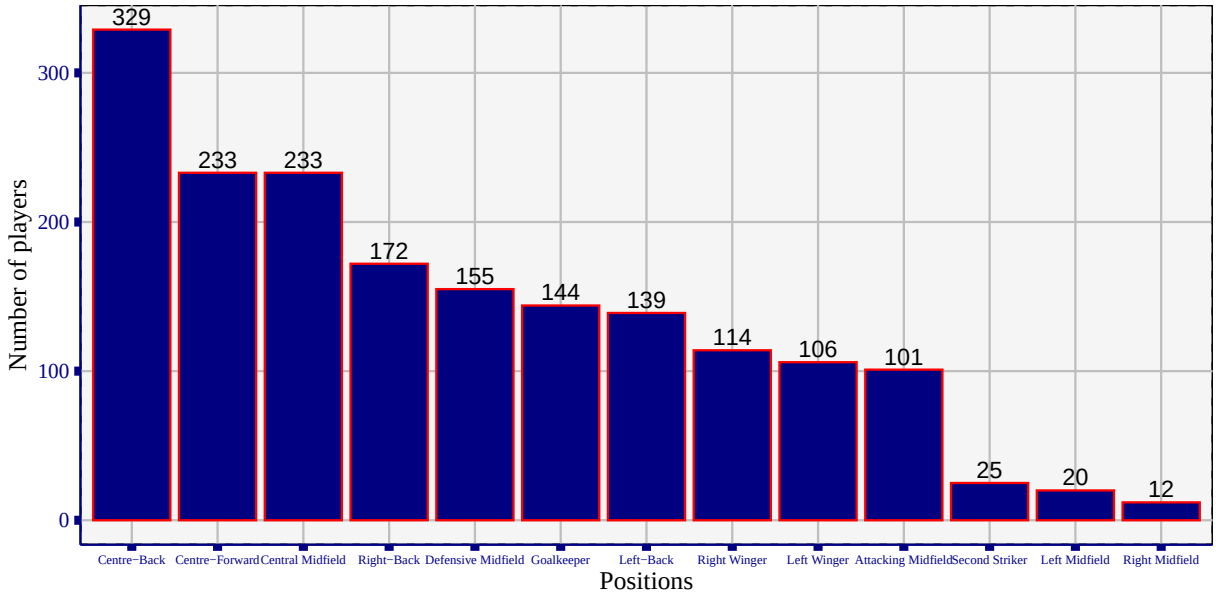


Figure 4: Frequencies of factor variables as covariates





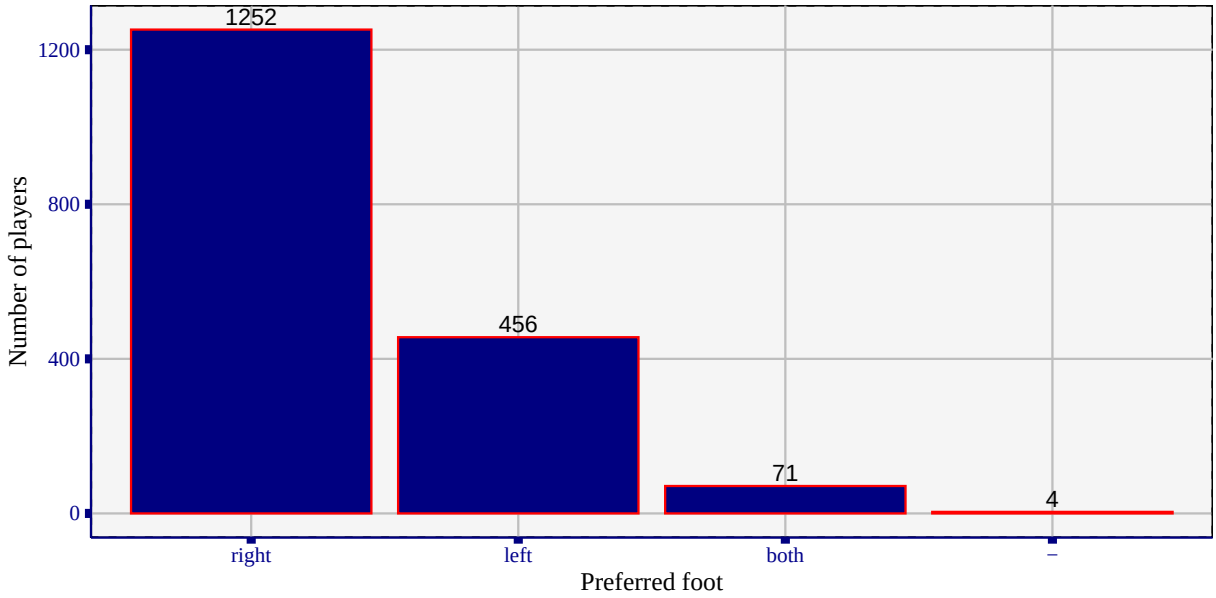


Figure 5: Correlation matrix