Report of MA678 Final Project

Xu Luo

2022-12-06

Load the packages

Absract

Football is the most popular sport in the world and the transfer market of football players is the most valuable market among all sports. In 2022, the player with the highest transfer value is Kylian Mbappe, whose value is up to 205 millions euros. The football clubs buy in players for many reasons, including enhancing their attack or defend ability, marketing, and so on. Thus, many factors could affect the transfer value of the players. To find out the effect of this factors, I used the FIFA players data and built a multilevel model with group level Position and League. The results show that the value of the players on different positions are affected more or less by different factors. For example, the value of central back players are more affected by their defending ability. This report includes 4 parts: Introduction, Methodology, Result, and Discussion.

Introduction

When a club decides whether to buy a player, the attributes of the players are always considered first (e.g. age, pace, shooting, defending, passing). Moreover,the clubs prefer different strengths when they consider the players on different positions. For example, the clubs may demand the striker has more pace and shooting skills while they need the midfielders pass and dribble well. They are willing to pay more transfer fees for the players who have corresponding abilities on the exact positions.

Besides, players from better leagues are usually more valuable in the deals between the clubs, since the players who play well in the low level league may perform bad in the upper level leagues. Players in the upper level league also gain more exposure to the media and has higher reputation, which may lead to higher business and marketing value.

In addition, the contract length and the wage in different leagues are also business factors that affect transfer value.A great example of this is Eduardo Camavinga. In 2020, his value at that time was reported as being as high as £90m. But Real Madrid only pay £35m to buy him in 2022 because Camavinga had just a year left on his deal and could have penned a deal with someone for free.

Therefore, I try to built multilevel models to figure out the effect of fixed effects(e.g. age, wage, international reputation and so on) and random effects(Positions and Leagues)

Methodology

Data

In the past decades, football players and football matches has generated a huge amount of data, which has been used by the clubs to improve their tactics or strategies. However, most of these data are on the team-levels. To analyse the factors that affect the player's value, more specific data about the individual player are important. In this research, I proposed the use of FIFA 22 game data from EA Sports for the study. Since 1995 the FIFA football video games provide an extensive and coherent scout of players worldwide. Player's information like wage and contract expiration date are collected clearly by EA. Other attributes like pace, shooting and defending are also included. Thus, the FIFA 22 game data from Kaggle(https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv) is appropriate for the analysis.

But data cleaning is still necessary before the modelling. Here I select the columns I need and subset a new data frame "val_columns". Then I exclude all the Goal Keepers from the data set since the evaluation of GK's value is totally different from other positions. NA values are also removed from the data. For the player's position, I noticed that some players could play multiple positions. So I split the "player_positions" columns into three and choose the first column as player's official position. Additionally, the I add a new column called "contract_due_in" by subtracting the contract expiration date by 2021. Last, I filter the top 9 leagues of the data since the original data set includes too many players and low level leagues, which may leads to outliers.

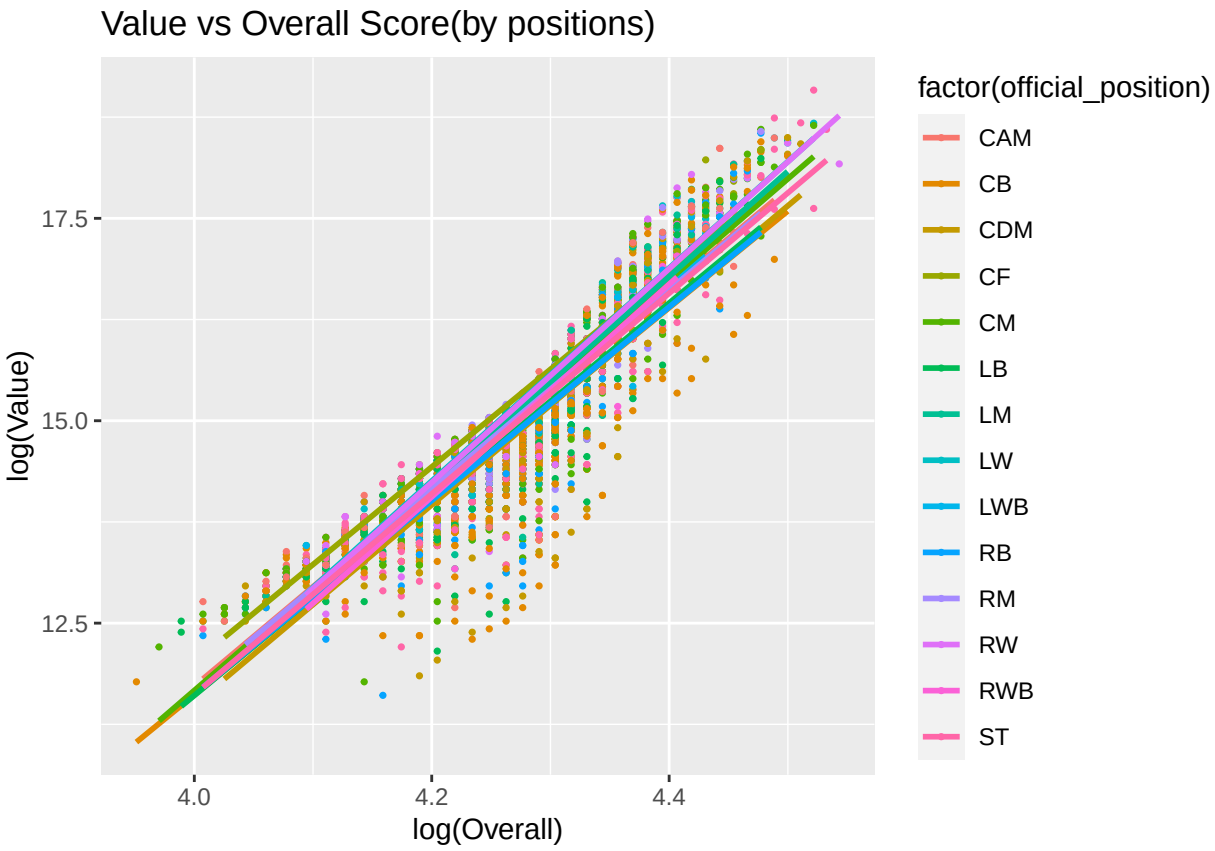
After the manipulations above, I form a data set of 4139 players and 17 columns:

column names	explanation
sofifa_id	player ID on sofifa
short_name	player short name
official_position	Position of player
overall	player current overall attribute
potential	player potential overall attribute
value_eur	player value (in EUR)
wage_eur	player weekly wage (in EUR)
age	player age
club_contract_until	contract expiration date
international_repu..	international reputation
pace	player speed attribute
shooting	player shooting attribute
passing	player passing attribute
dribbling	player dribbling attribute
defending	player defend attribute
physic	player heading accuracy
contract_due_in	contract expiration date

EDA

After the data cleaning, I got a data set of 4139 players and 17 variables. But which variables could be used in analysis is still needed the following analysis.

Value v.s. overall scores in different groups: positions, league_level



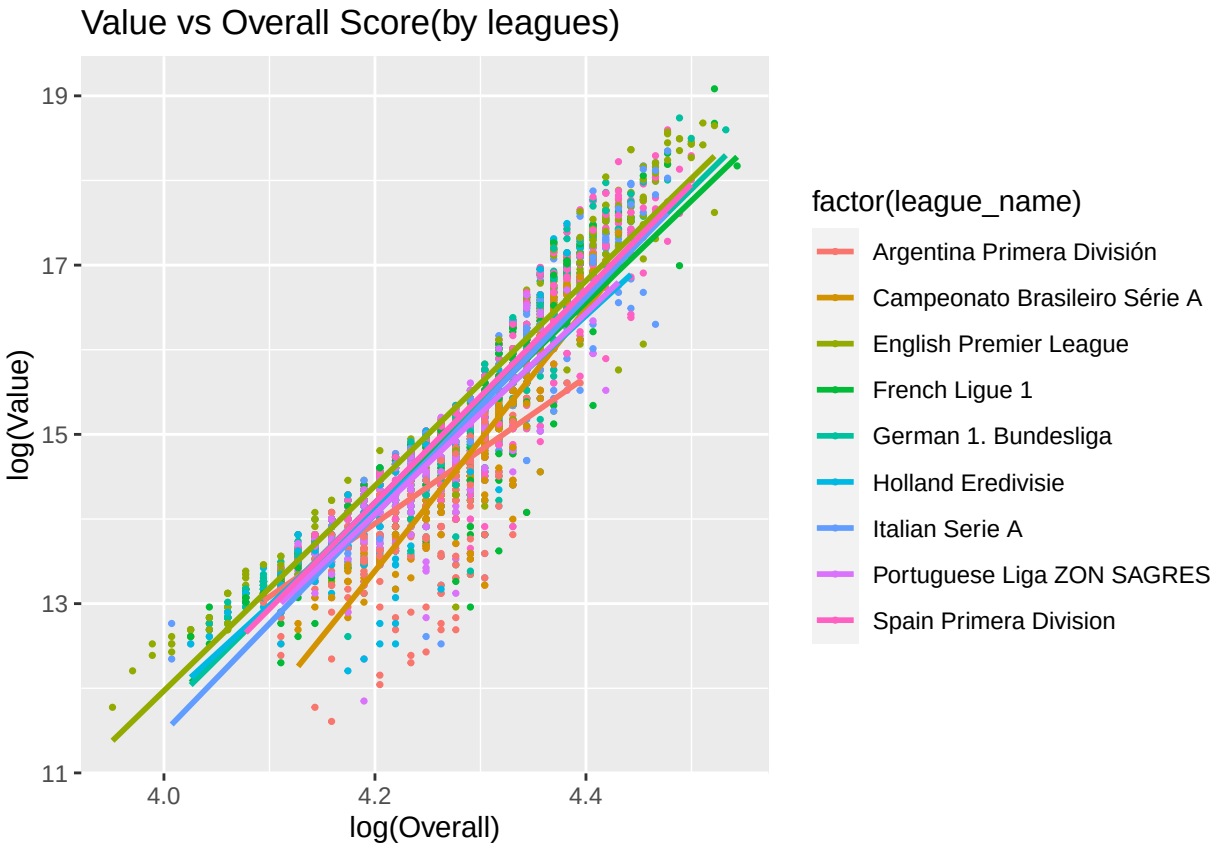


Figure 1 illustrates the relationship between value and overall score, while fig(a) is in position level and fig(b) is in league level. However, whatever the level, value show the increasing trend as points going up. In different positions and leagues, the intercepts and slopes show little differences. I also draw the graph of value versus wage_eur, age, potentials, contract_due_in, reputation, pace,shooting,passing,defending,dribbling,physic, and the figures are similar. Thus I put them in the appendix.

Model fitting

As I mentioned in the Introduction, positions and leagues may have random effects on the model, I decide to use multilevel model. Since all variables are more or less skewed and have heavy tails, I adjust all variable to log(variable) to create log data frame. Next, the Pearson correlation matrix is created to decide the predictors selection.

pearson coefficient matrix

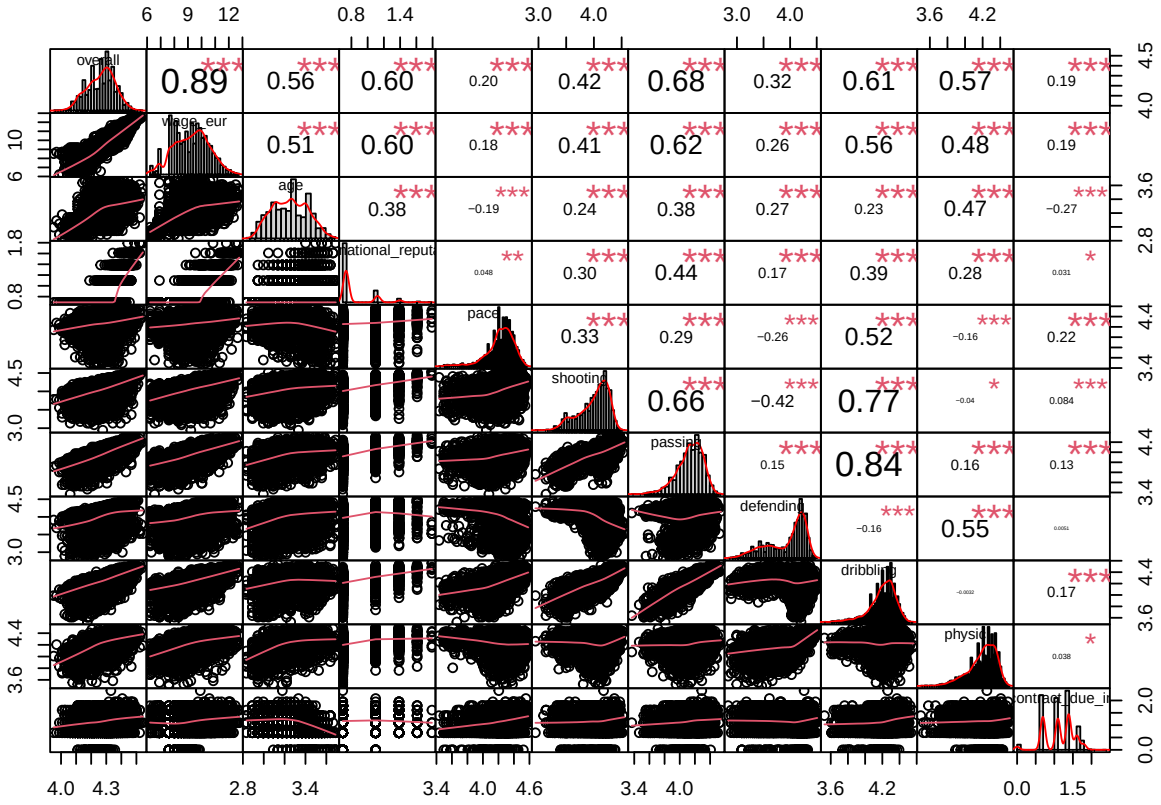


Figure 1: Correlation Matrix

The Pearson correlation matrix clearly shows Pearson relationship between variables and I use .6 as a significant level to check whether variables are highly related. Here, I wipe out the predictor "overall" since it has coefficients larger than .6 with many other predictors. Then, the high correlation appears between shooting & passing and passing & dribbling as well as shooting & dribbling. Hence, I decided to drop passing and dribbling. I keep "shooting" since it is a crucial factor evaluating the attack ability.

Besides, players on different positions have quite different strengths and abilities, random effect of positions is important for variables: pace + shooting + defending+ physic. On the other hand, "wage, reputation, and

contract" are easily affected by the leagues . Thus, I varied the slopes and intercepts of "wage, reputation, and contract" on different leagues.

The model and the results are below, all variables here are considered as statistically significant at = 0.5 level.: ## FIT THE MODEL

```
fit_FIFA <- lmer(value_eur ~  age  + wage_eur + international_reputation + pace + shooting
                 + defending+ physic + contract_due_in
              + ( 1+ pace + shooting + defending+ physic | official_position)
              + (1 + wage_eur + international_reputation + contract_due_in| league_name),
              data = log_FIFA_players,
              REML = FALSE)
summary(fit_FIFA)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: value_eur ~ age + wage_eur + international_reputation + pace +
##     shooting + defending + physic + contract_due_in + (1 + pace +
##     shooting + defending + physic | official_position) + (1 +
##     wage_eur + international_reputation + contract_due_in | league_name)
##    Data: log_FIFA_players
##
##      AIC      BIC   logLik deviance df.resid
##   3957.0   4178.5  -1943.5   3887.0     4104
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6873 -0.6499  0.0004  0.6134  4.1608
##
## Random effects:
##  Groups            Name                     Variance  Std.Dev. Corr
##  official_position (Intercept)              8.4603962 2.90868
##                    pace                     0.3542405 0.59518  -0.02
##                    shooting                 4.5008130 2.12151   0.10 -0.08
##                    defending                4.2599670 2.06397  -0.30 -0.10
##                    physic                   0.1715542 0.41419  -0.55 -0.51
##  league_name       (Intercept)              0.5667497 0.75283
##                    wage_eur                 0.0051283 0.07161  -0.89
##                    international_reputation 0.1075983 0.32802   0.28 -0.63
##                    contract_due_in          0.0006969 0.02640  -0.97  0.94
##  Residual                                   0.1402216 0.37446
##
##
##
##
##  -0.96
##  -0.40  0.52
##
##
##
##  -0.34
##
## Number of obs: 4139, groups:  official_position, 14; league_name, 9
##
## Fixed effects:
```

```
##                             Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)               -10.79299    1.00249  150.34622 -10.766  < 2e-16 ***
## age                        -2.88431    0.05350 4048.92030 -53.916  < 2e-16 ***
## wage_eur                    0.60251    0.02724   12.07446  22.115 3.87e-11 ***
## international_reputation    0.79974    0.12071    6.30635   6.626 0.000463 ***
## pace                        1.32446    0.18162   14.27175   7.293 3.53e-06 ***
## shooting                    2.76316    0.57347  155.80312   4.818 3.41e-06 ***
## defending                   2.33829    0.55845  161.56604   4.187 4.63e-05 ***
## physic                      0.39213    0.12913   17.40913   3.037 0.007297 **
## contract_due_in             0.15632    0.02088   21.90254   7.488 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) age    wage_r intrn_ pace   shotng dfndng physic
## age         -0.121
## wage_eur    -0.016 -0.139
## intrntnl_rp  0.075 -0.027 -0.532
## pace        -0.180  0.081 -0.063  0.006
## shooting     0.040 -0.016 -0.032 -0.007 -0.068
## defending   -0.284 -0.008 -0.030 -0.005 -0.080 -0.930
## physic      -0.396 -0.057 -0.024  0.014 -0.416 -0.358  0.428
## contrct_d_n -0.097  0.365  0.319 -0.118 -0.026 -0.011 -0.011 -0.019
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

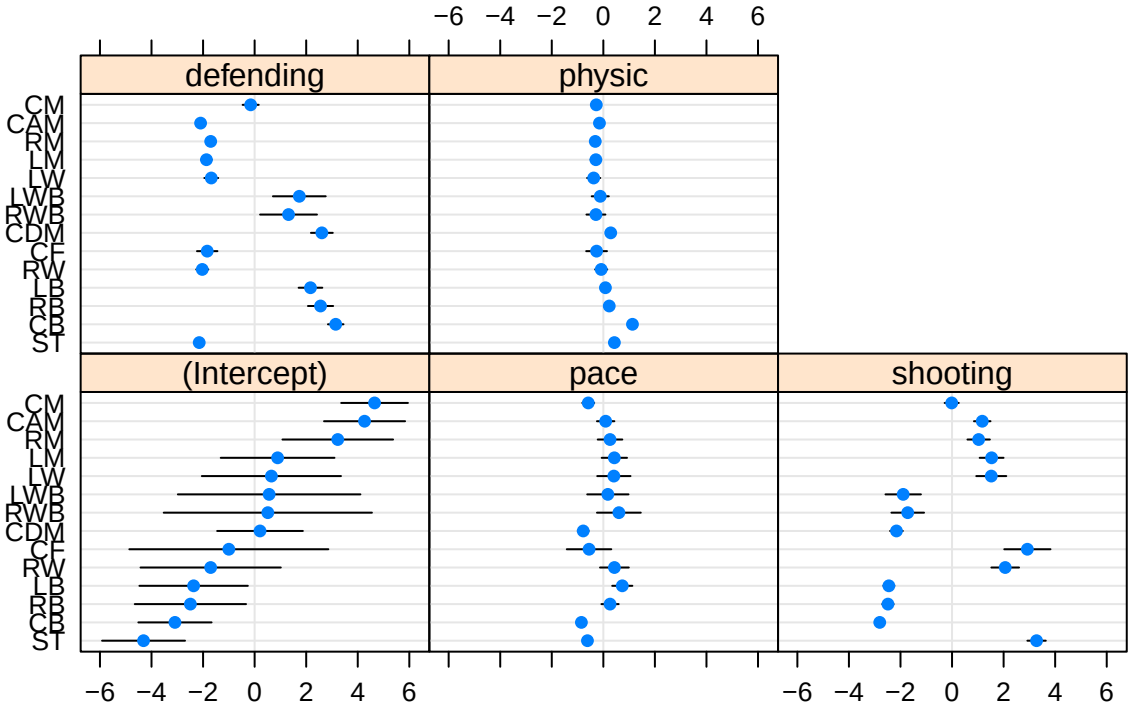
The following tables show the random effect of positions and leagues:

```
##     (Intercept)  pace shooting defending physic
## CAM        4.26  0.09     1.17     -2.10  -0.15
## CB        -3.09 -0.85    -2.80      3.15   1.13
## CDM        0.21 -0.78    -2.16      2.61   0.29
## CF        -1.00 -0.55     2.92     -1.84  -0.26
## CM         4.65 -0.59    -0.01     -0.15  -0.27
## LB        -2.37  0.74    -2.45      2.17   0.08
## LM         0.89  0.43     1.54     -1.87  -0.29
## LW         0.65  0.41     1.52     -1.68  -0.38
## LWB        0.56  0.18    -1.89      1.73  -0.12
## RB        -2.49  0.26    -2.49      2.56   0.23
```

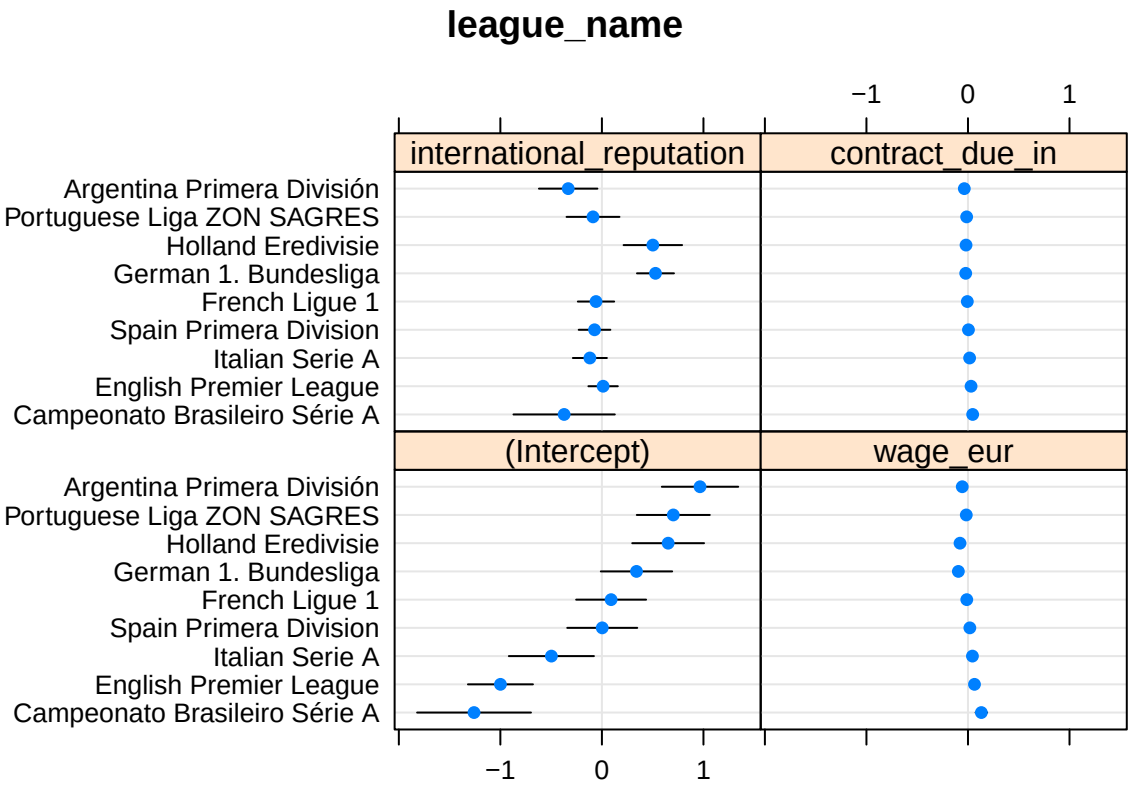
```
##                                (Intercept) wage_eur international_reputation
## Argentina Primera División            0.97    -0.06                     -0.33
## Campeonato Brasileiro Série A        -1.26     0.13                     -0.37
## English Premier League               -1.00     0.06                      0.01
## French Ligue 1                        0.09    -0.01                     -0.06
## German 1. Bundesliga                  0.34    -0.09                      0.53
## Holland Eredivisie                    0.65    -0.08                      0.50
## Italian Serie A                      -0.50     0.04                     -0.12
## Portuguese Liga ZON SAGRES            0.70    -0.02                     -0.09
## Spain Primera Division                0.00     0.02                     -0.07
## NA                                      NA       NA                        NA
##                                contract_due_in
## Argentina Primera División               -0.04
## Campeonato Brasileiro Série A             0.05
## English Premier League                    0.03
```

```
## French Ligue 1                          -0.01
## German 1. Bundesliga                    -0.02
## Holland Eredivisie                      -0.02
## Italian Serie A                          0.02
## Portuguese Liga ZON SAGRES              -0.01
## Spain Primera Division                   0.00
## NA                                         NA
```

```
## $official_position
```




```
##
## $league_name
```


league_name
international_reputation
contract_due_in
(Intercept)
wage_eur
Argentina Primera División
Portuguese Liga ZON SAGRES
Holland Eredivisie
German 1. Bundesliga
French Ligue 1
Spain Primera Division
Italian Serie A
English Premier League
Campeonato Brasileiro Série A
−1
0
1

The following plot could illustrate that the clubs prefer to offer high price to players with good defending and shooting skills regardless the positions:

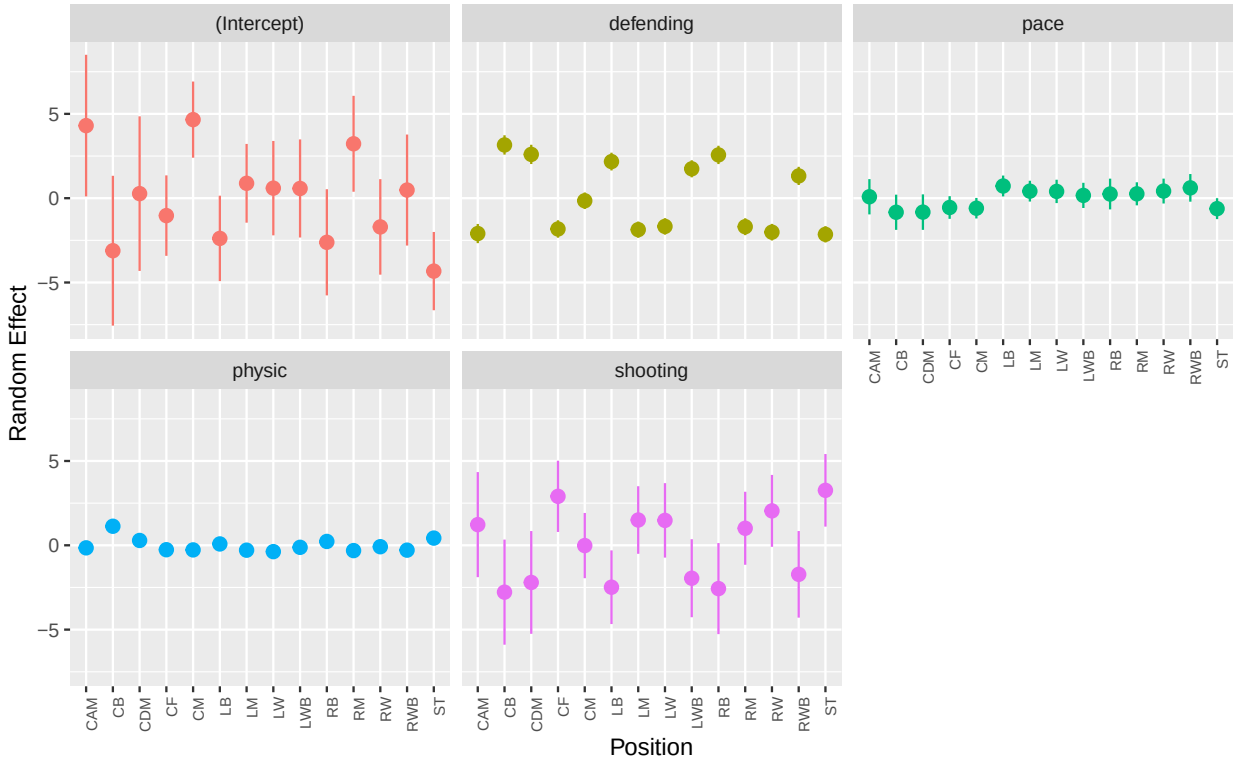


Figure 2: Random Effect of NBA Model

Result

Interpretation

From the model fitting above, I get the following formula of fixed effect:

$$log(value+1) = -10.79 - 2.88 \times log(age+1) + 0.60 \times log(wage+1) + 0.80 \times log(internationalreputation+1)$$

$$+1.32 \times log(pace+1) + 2.76 \times log(shooting+1) + 2.33 \times log(defending+1) + 0.39 \times log(physic+1) + 0.15 \times log(contractdue+1)$$

Then add the random effect of position to the intercepts and slopes and get the estimated formula, here I take the "CAM" position English Premier League as an example:

$$log(value+1) = -6.53 - 2.88 \times log(age+1) + 0.66 \times log(wage+1) + 0.81 \times log(internationalreputation+1)$$

$$+1.41 \times log(pace+1) + 3.96 \times log(shooting+1) + 0.23 \times log(defending+1) + 0.24 \times log(physic+1) + 0.11 \times log(contractdue+1)$$

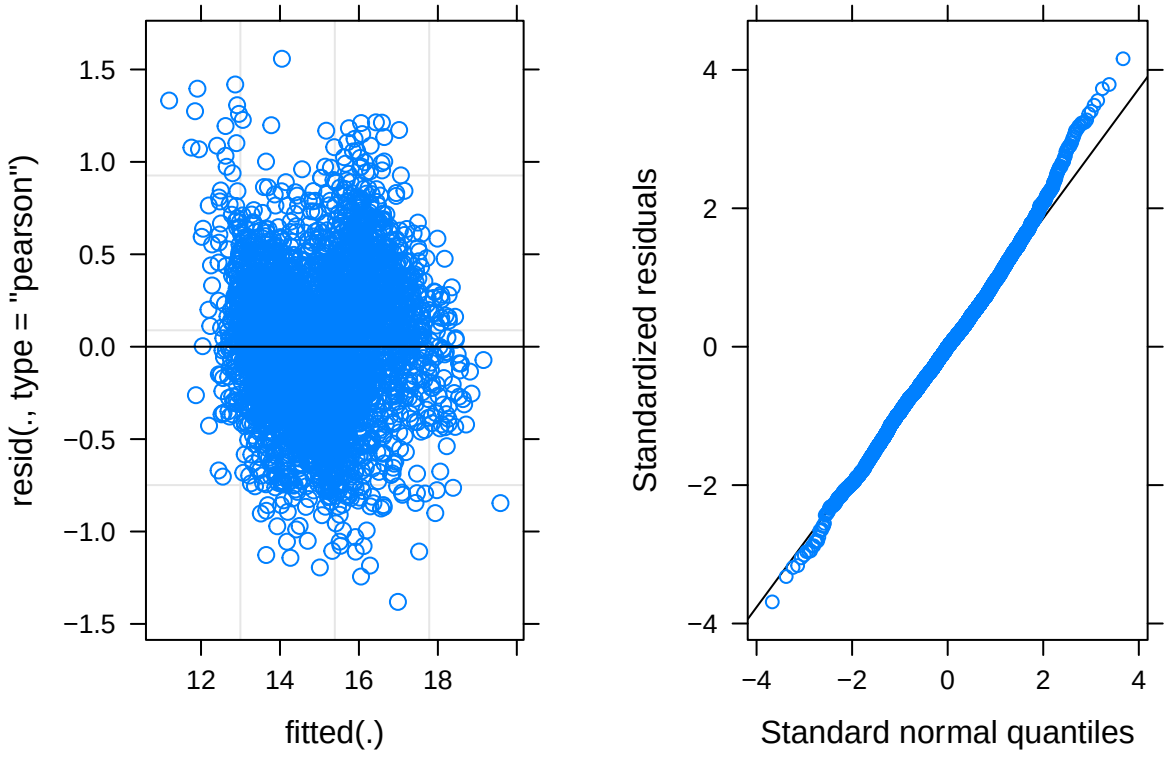
In the formula, all the parameters are positive except age. It's natural that the parameter of age is negative since the performance of players usually decrease since he becomes older. In the model, for every increase 1% of player's defending ability, the prediction of transfer market value increase 23% for the players in "CAM" position in English Premier League. All the parameters are changing from positions to position, league to league.

Besides, I also find that CAM in English Premier League has higher value compared to the average (-6.53 v.s. -10.79) among all leagues. This is understandable since the market value of the English Premier League is highest in the world, so their players are more famous.

Model checking

The residual plot and the Q-Q plot really show that the model fitted well above.

```
residual_plot <- plot(fit_FIFA)
qq_plot       <- qqmath(fit_FIFA)
grid.arrange(residual_plot, qq_plot, ncol = 2)
```



Discussion

In this report, multilevel model is used to figure out the relationship between players' value and their abilities, contract as well as reputation. Random effects from positions and leagues are also included in the model. The results basically show that the value of the players on different positions and leagues are affected more or less by different factors.

Appendix

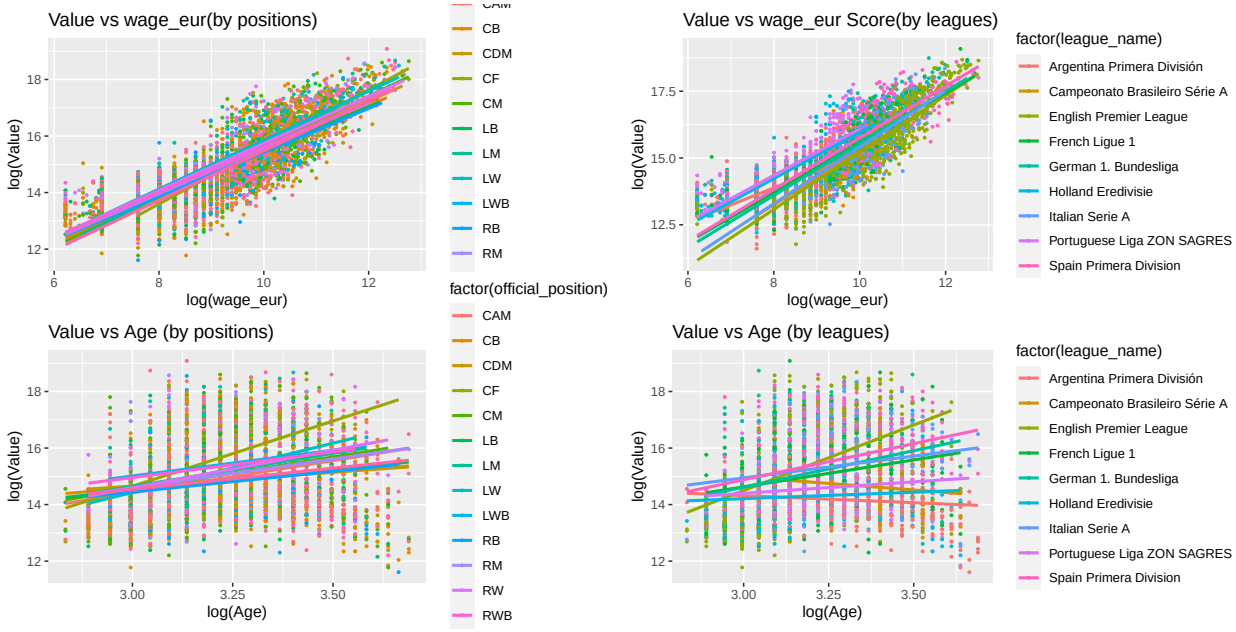


Figure 3: EDA: random effects of league and positions(1)

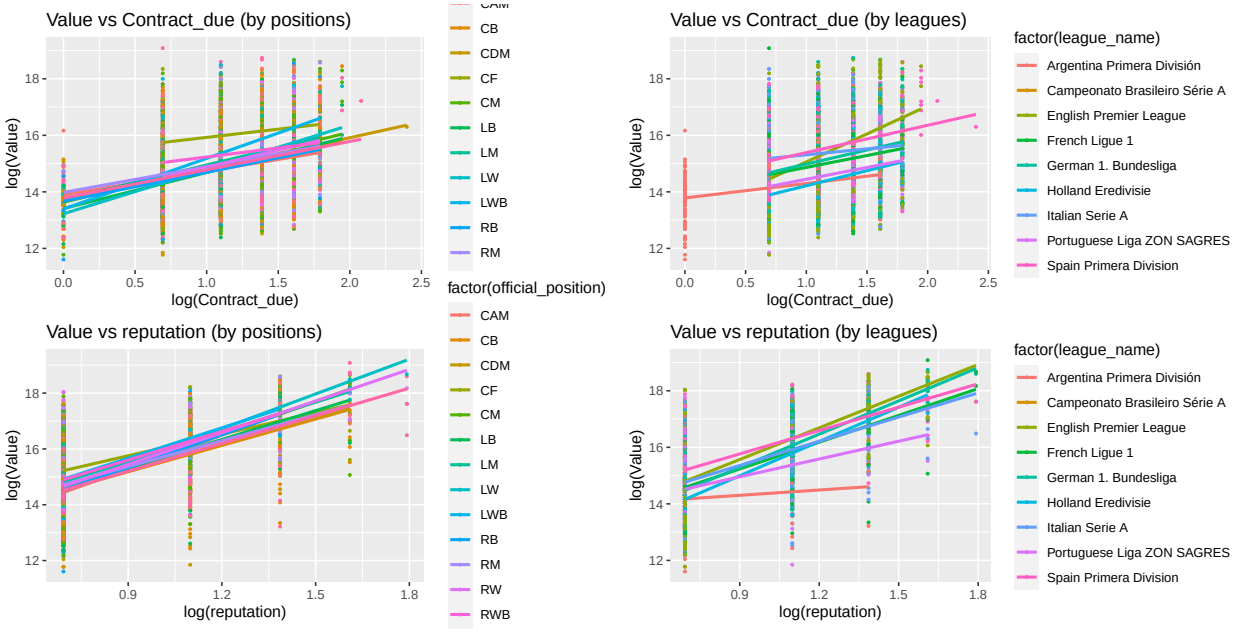


Figure 4: EDA: random effects of league and positions(2)

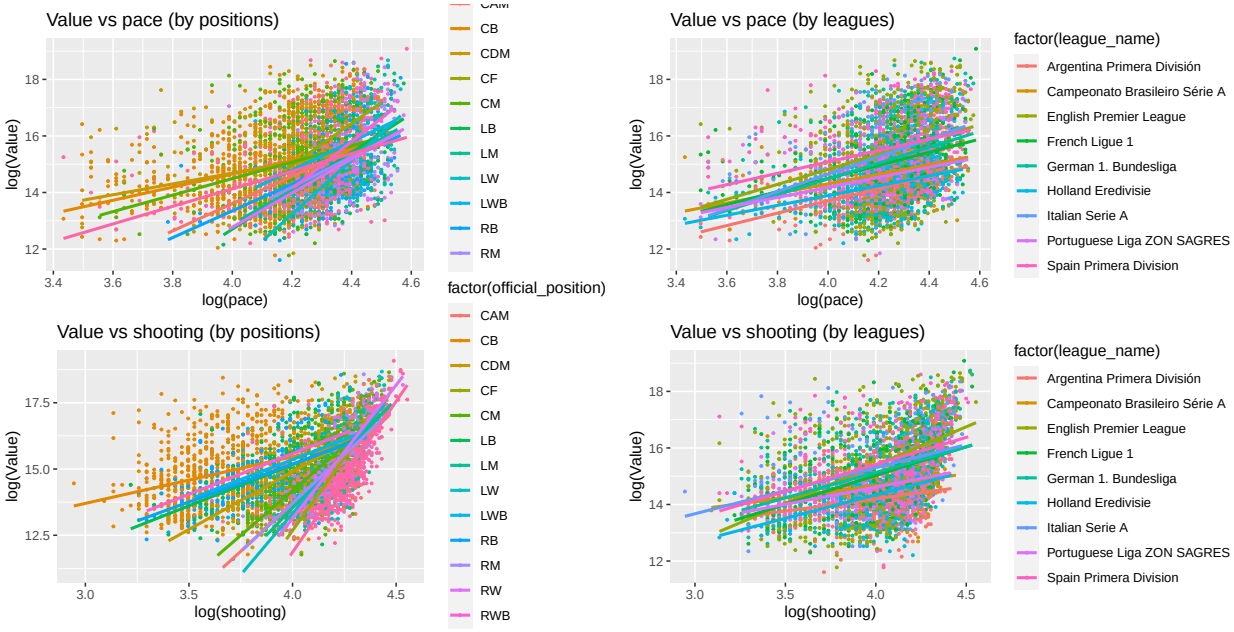


Figure 5: EDA: random effects of league and positions(3)

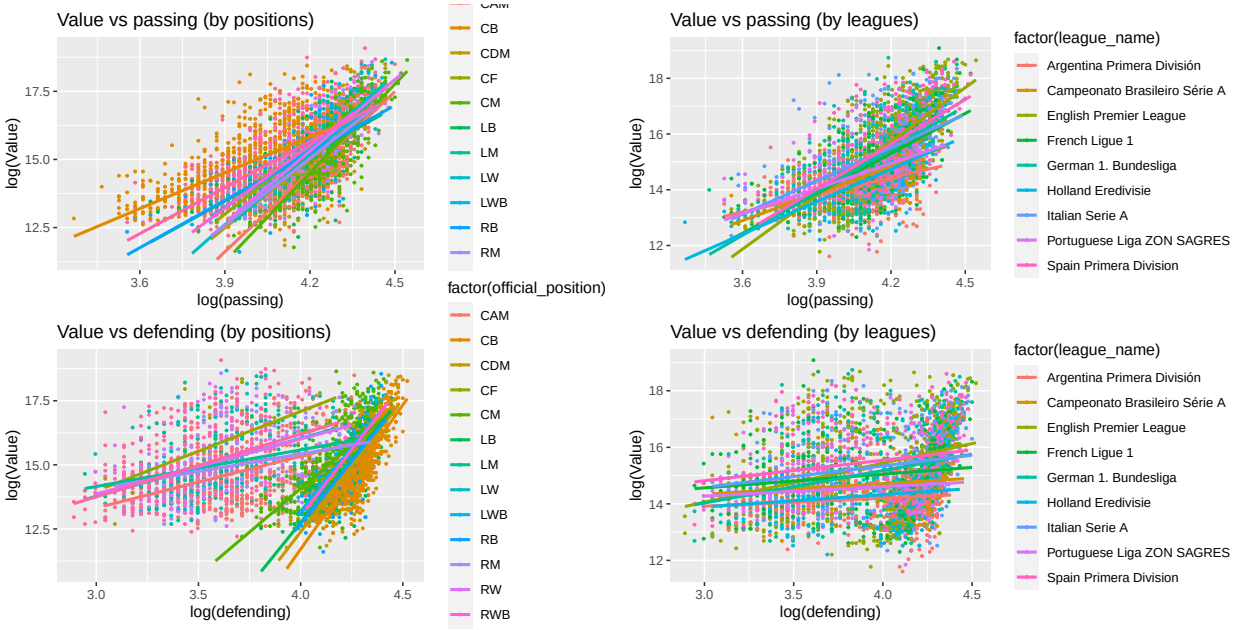


Figure 6: EDA: random effects of league and positions(4)

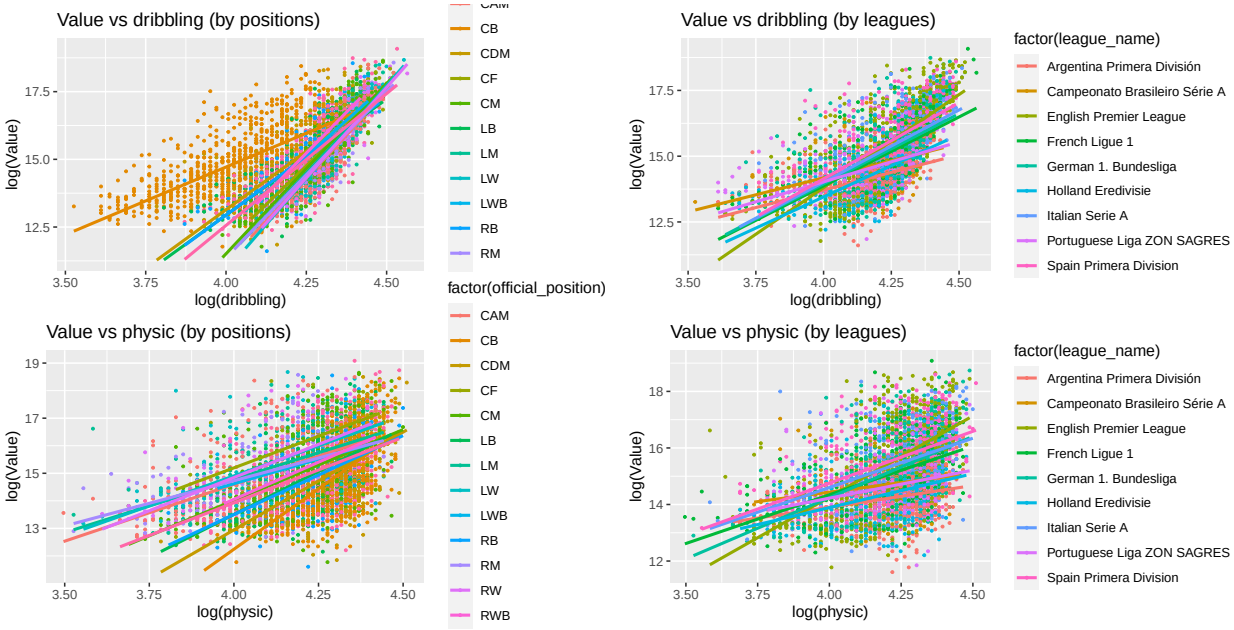


Figure 7: EDA: random effects of league and positions(5)