2017-2018 FIFA Player Market Value Analysis

678 Midterm project for Samuel Luo

Abstract:

This is a statistical report about the market value of soccer players. The goal for this report is to explore the factors that would probably affect a player's market value, and use them to perform the model analysis. The method I used is applying a linear regression model with four related variables (clubs, reputation, positions, and age) to predict log of player's value. The result is that player's age has a very big impact on reducing their market value. Players in position Attack and Midfield tend to gain more market value than plays in position Defense and Goalkeeper. The more reputation would result in more market value a player will get. Thus, the four related variables do have effects on a player's market value and they might be used for predicting the player's market value.

Introduction:

#1.1 Background:

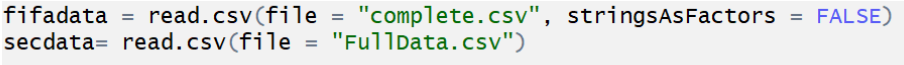
In recent years, plenty of new young football (Soccer) players come into people's view and show their skills and strengths. Presenting in top leagues, they become famous and their personal values are increasing gradually and rapidly. We know that soccer players' personal market value might largely depend on their ability to strike goals or defense against opponents. But are there any other factors play roles to affect football player's personal value? How it could be changed because of these factors? I am willing to perform model analysis to represent and interpret the relationship of athletes' value and other effects which are highly connected with it.

#1.2 Data Sources:

The datasets I used for performing this analysis, 'Complete FIFA 2017 Player Dataset Global' and 'FIFA18 More Complete Player Dataset', are obtained from Kaggle website, but they are originally scraped from a reputed soccer website called Sofifa.com. I believe that the variables of these datasets are relatively reliable and credible, especially the reputation grading, which are computed by precision operations, is of certain referential values. Once again, the expected goal for me is to create models which would effectively represent what influents player's value and if applicable, this model can do a prediction for new player's value.

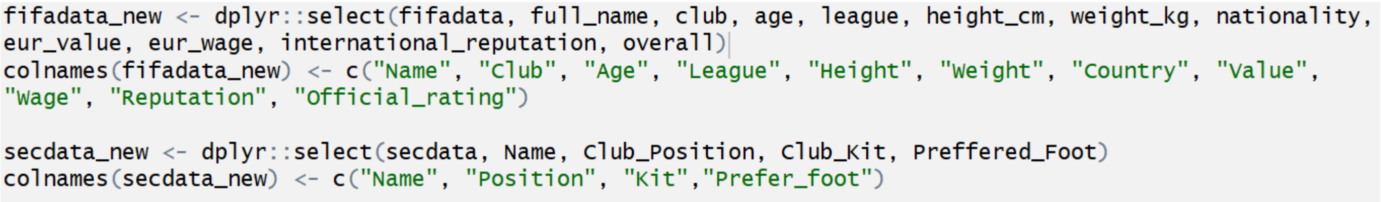
#1.3 Previous Work: Data combining and cleaning

Firstly, I imported two datasets into R.



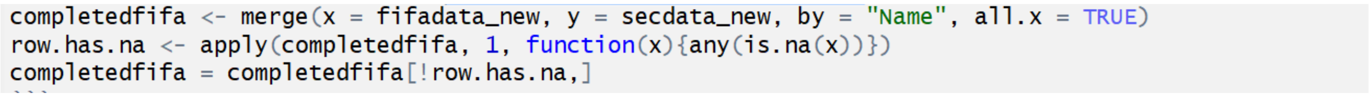
```
fifadata = read.csv(file = "complete.csv", stringsAsFactors = FALSE)
secdata= read.csv(file = "FullData.csv")
```

And then I selected the potential variables in both datasets that are related to this future analysis and rename them to be clarified easily.



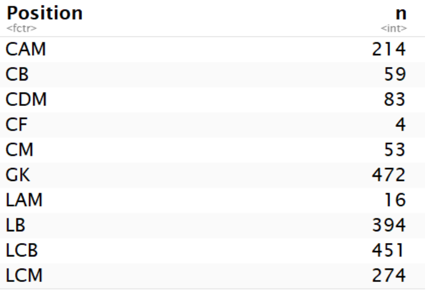
```
fifadata_new <- dplyr::select(fifadata, full_name, club, age, league, height_cm, weight_kg, nationality,
eur_value, eur_wage, international_reputation, overall)
colnames(fifadata_new) <- c("Name", "Club", "Age", "League", "Height", "Weight", "Country", "Value",
"Wage", "Reputation", "Official_rating")

secdata_new <- dplyr::select(secdata, Name, Club_Position, Club_Kit, Preffered_Foot)
colnames(secdata_new) <- c("Name", "Position", "Kit","Prefer_foot")
```

After that, I merged these two new datasets by player's names and make it be a new data frame, which is called 'completedfifa.' To polish this data, I wiped out the abnormal values of each column such as NAs, so that finally it can be used for our analysis.



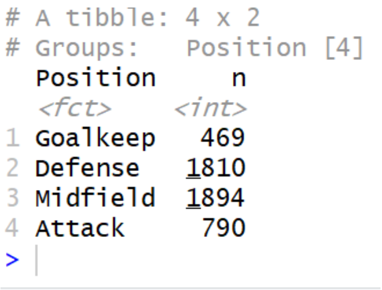
```
completedfifa <- merge(x = fifadata_new, y = secdata_new, by = "Name", all.x = TRUE)
row.has.na <- apply(completedfifa, 1, function(x){any(is.na(x))})
completedfifa = completedfifa[!row.has.na,]
```

Because that there are too many kinds of soccer player positions, and part of them sometimes are similar, I delete the positions which stand for substitutes players and reserves players and then put other positions into four general groups: Goalkeeper, Defense, Midfield, and Attack.



Position <fctr>	n <int>
CAM	214
CB	59
CDM	83
CF	4
CM	53
GK	472
LAM	16
LB	394
LCB	451
LCM	274

Figure1: 29 kinds of positions



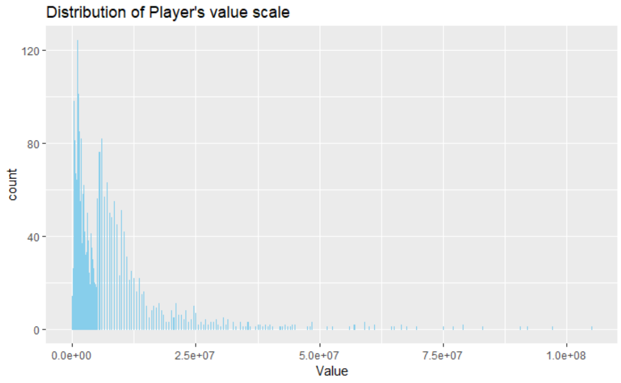
```
# A tibble: 4 x 2
# Groups:   Position [4]
  Position      n
  <fct>     <int>
1 Goalkeep    469
2 Defense    1810
3 Midfield   1894
4 Attack      790
>
```

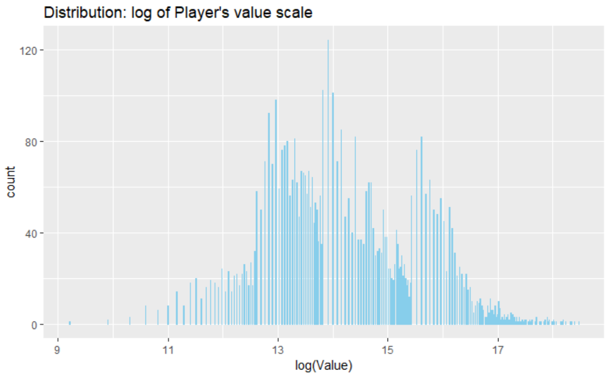
Figure2: 4 general groups of positions

After completing data cleaning, I finally choose the variables clubs, positions, reputation and age as objects of study for a player's market value.

#1.4 EDA Visualization

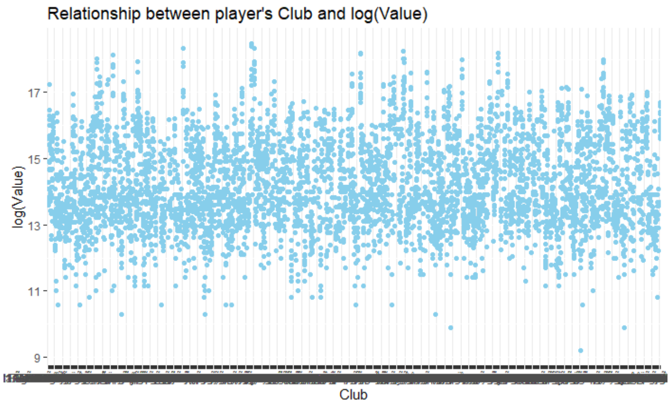
A. Market Value





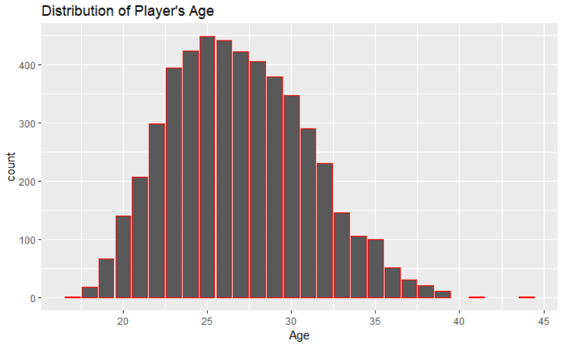
From the simple distribution plot of Player's value, we can find out that most of the player's value is much lower than 2.5e+07 EURO. But for a few players, their value could be two times or even around four times 2.5e+07 EURO. Based on the fact that most of the values are lay on the left corner and a few values' plots are too far from the majority, it would be better to do log transformation for player's value during fitting model.

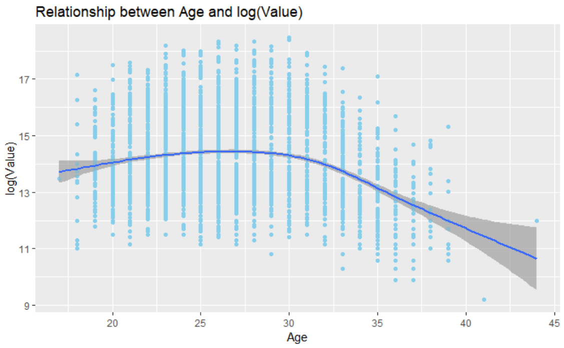
B. Club



In fact, players from some clubs do have more values in EURO compare to others, which we can assume that a soccer player's value might partly depend on which club the player works for.

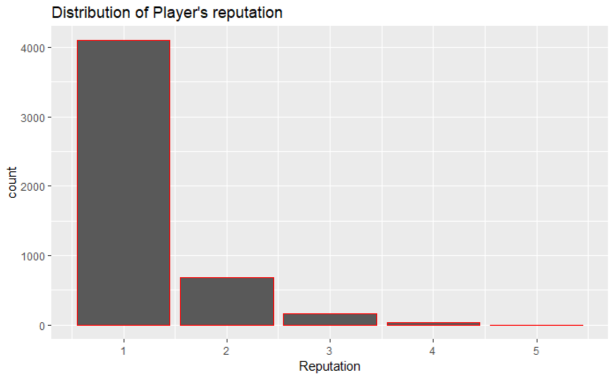
C. Player's age

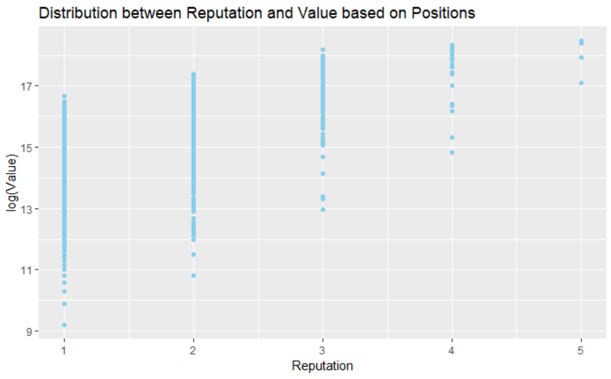




By looking at the simple distribution of Player's age, players that in their 20+ make up most of the population of soccer players. After taking player's log(value) into account, the second plot a linear relationship that log(value) has slightly increase when young players are approaching their 20s, and then log(value) begins to decrease more and more after around 30 years old. Since the regression line looks like a parabola, I might use poly transformation for this variable in model.

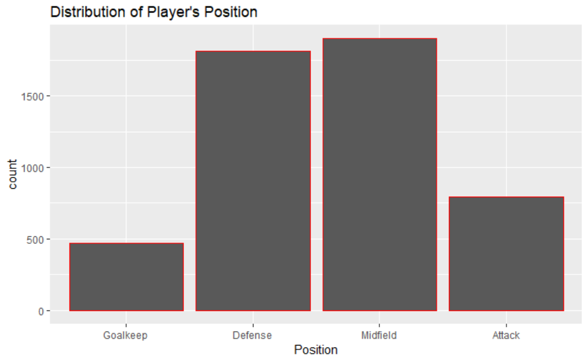
D. Player's reputation

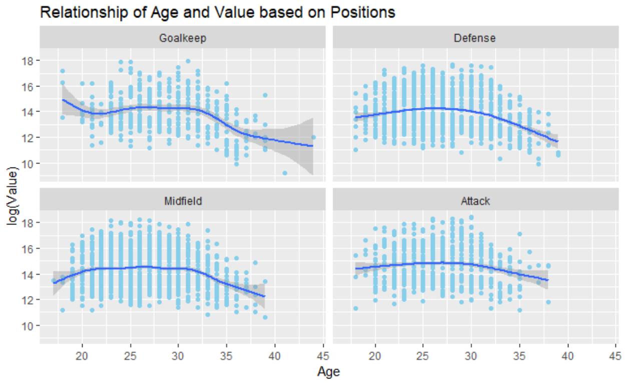


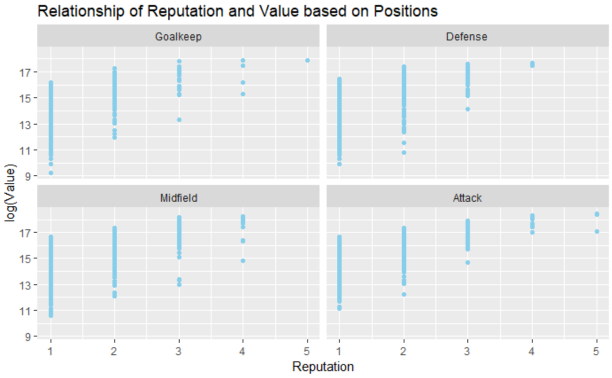


About the players' reputation, most of them do not have much popularity. As the grading of reputation increases, the number of corresponding players is obviously decreasing. After adding the variable log(Value) into the visual, reputation seems to have certain effects on the players' market value. The more reputation will result in the more market value a player has.

E. Player’s positions







Right now make Player's positions come into account. First, the population of goalkeeper and attack is relatively lower than that of defense and midfield. Then, for the relationships of value and other variables based on player's positions. This relationship distribution reveals that players whose position are attack and midfield make relatively more chance to have higher market value.

Method:

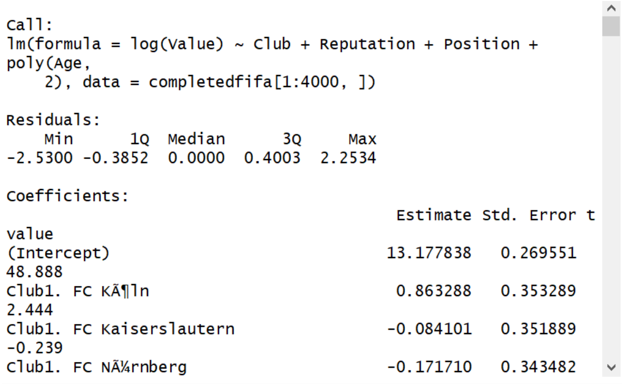
#2.1 Model choices

I use nearly 4/5 of the total data for fitting models and then predict the rest of the data in the model checking part.

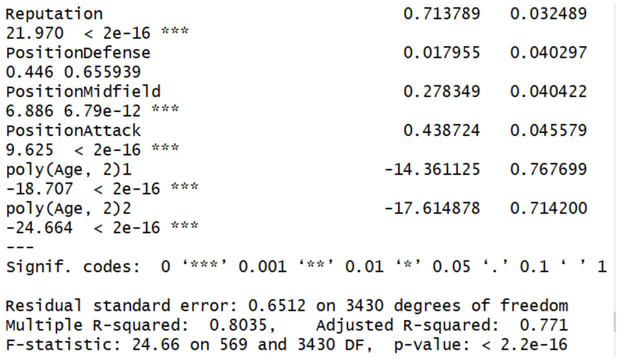
A. Choice 1: Linear regression

```
soccer1 <- lm(data = completedfifa[1:4000,], log(Value) ~ Club + Reputation + Position +
poly(Age,2))
summary(soccer1)
```

Summary of this model

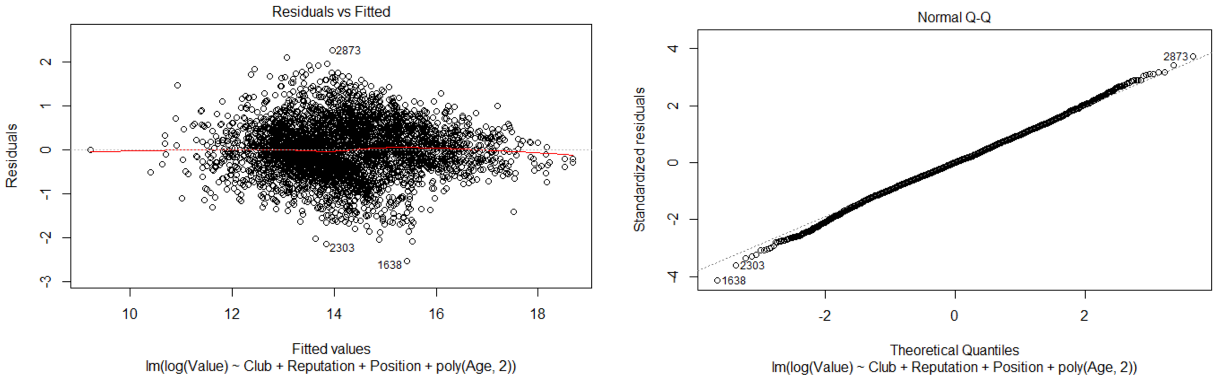


```
Call:
lm(formula = log(Value) ~ Club + Reputation + Position +
poly(Age,
    2), data = completedfifa[1:4000, ])

Residuals:
    Min      1Q  Median      3Q     Max
-2.5300 -0.3852  0.0000  0.4003  2.2534

Coefficients:
                                       Estimate Std. Error t
value
(Intercept)                           13.177838   0.269551
48.888
Club1. FC KÃ¶ln                         0.863288   0.353289
2.444
Club1. FC Kaiserslautern              -0.084101   0.351889
-0.239
Club1. FC NÃ¼rnberg                    -0.171710   0.343482
```

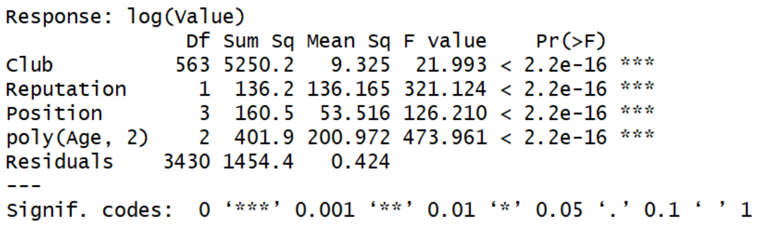


```
Reputation                             0.713789   0.032489
21.970  < 2e-16 ***
PositionDefense                        0.017955   0.040297
0.446 0.655939
PositionMidfield                       0.278349   0.040422
6.886 6.79e-12 ***
PositionAttack                         0.438724   0.045579
9.625  < 2e-16 ***
poly(Age, 2)1                        -14.361125   0.767699
-18.707  < 2e-16 ***
poly(Age, 2)2                        -17.614878   0.714200
-24.664  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6512 on 3430 degrees of freedom
Multiple R-squared:  0.8035,    Adjusted R-squared:  0.771
F-statistic: 24.66 on 569 and 3430 DF,  p-value: < 2.2e-16
```

Residual plot for this model



ANOVA test



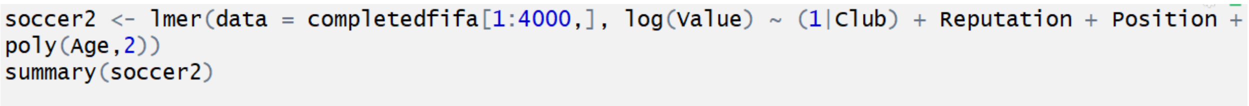
```
Response: log(Value)
               Df Sum Sq Mean Sq F value    Pr(>F)    
Club          563 5250.2   9.325  21.993 < 2.2e-16 ***
Reputation      1  136.2 136.165 321.124 < 2.2e-16 ***
Position        3  160.5  53.516 126.210 < 2.2e-16 ***
poly(Age, 2)    2  401.9 200.972 473.961 < 2.2e-16 ***
Residuals    3430 1454.4   0.424                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This linear regression model fits well. In this model’s summary, the R-squared for this model is around 80% and the p-value is very small which is about 2.2e-16. By looking at the residual plot, plots seem like equally distributed and symmetric along the horizontal line. In addition, QQplot looks well too, since all of the plots are close to the straight line. Finally, by using ANOVA to test the coefficients, the result shows that all of the coefficients are significant. Thus, this might be an appropriate model for doing a deeper analysis.

B. Choice 2: Multilevel model

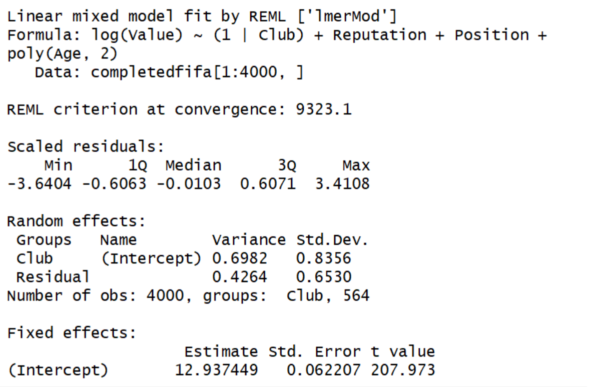
Here I fitted three multilevel models, and later I would evaluate these models by using ANOVA test comparison.

a) One random effect: Club

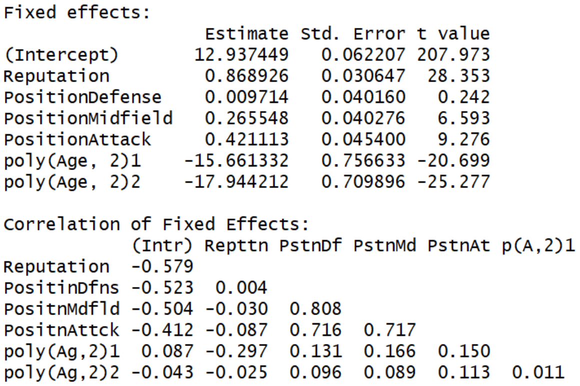


```
soccer2 <- lmer(data = completedfifa[1:4000,], log(Value) ~ (1|Club) + Reputation + Position +
poly(Age,2))
summary(soccer2)
```

Summary of this model

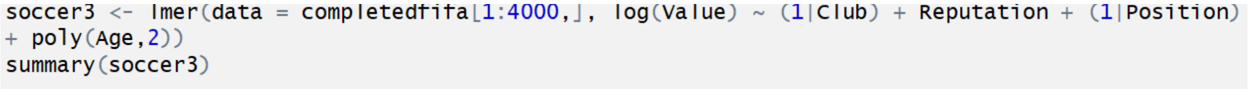


```
Linear mixed model fit by REML ['lmerMod']
Formula: log(Value) ~ (1 | Club) + Reputation + Position +
poly(Age, 2)
   Data: completedfifa[1:4000, ]

REML criterion at convergence: 9323.1

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.6404 -0.6063 -0.0103  0.6071  3.4108 

Random effects:
 Groups   Name        Variance Std.Dev.
 Club     (Intercept) 0.6982   0.8356  
 Residual             0.4264   0.6530  
Number of obs: 4000, groups:  Club, 564

Fixed effects:
                  Estimate Std. Error t value
(Intercept)      12.937449   0.062207 207.973
```



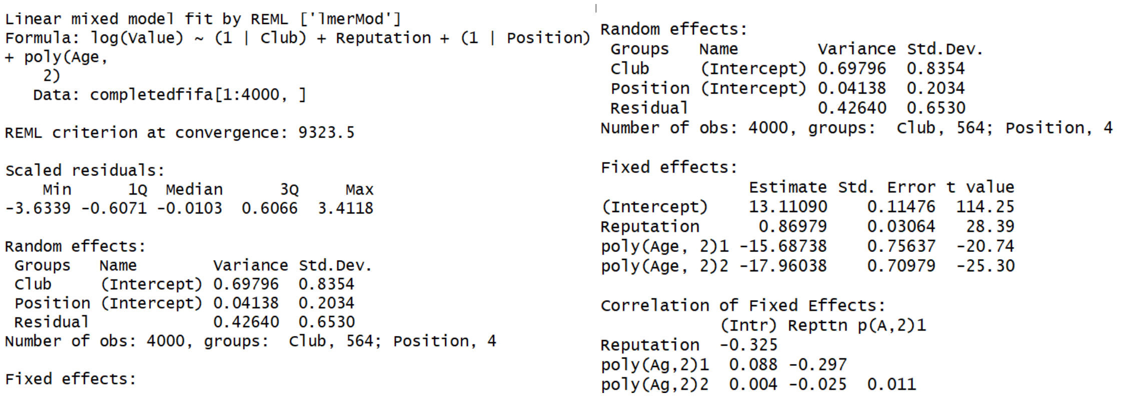
```
Fixed effects:
                  Estimate Std. Error t value
(Intercept)      12.937449   0.062207 207.973
Reputation        0.868926   0.030647  28.353
PositionDefense   0.009714   0.040160   0.242
PositionMidfield  0.265548   0.040276   6.593
PositionAttack    0.421113   0.045400   9.276
poly(Age, 2)1   -15.661332   0.756633 -20.699
poly(Age, 2)2   -17.944212   0.709896 -25.277

Correlation of Fixed Effects:
            (Intr) Repttn PstnDf PstnMd PstnAt p(A,2)1
Reputation  -0.579                                    
PositinDfns -0.523  0.004                             
PositnMdfld -0.504 -0.030  0.808                      
PositnAttck -0.412 -0.087  0.716  0.717               
poly(Ag,2)1  0.087 -0.297  0.131  0.166  0.150       
poly(Ag,2)2 -0.043 -0.025  0.096  0.089  0.113  0.011
```

b) Two random effect: Club and Position



```
soccer3 <- lmer(data = completedfifa[1:4000,], log(Value) ~ (1|Club) + Reputation + (1|Position)
+ poly(Age,2))
summary(soccer3)
```

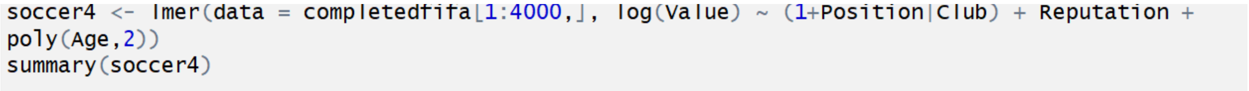
Summary of the model



```
Linear mixed model fit by REML ['lmerMod']
Formula: log(Value) ~ (1 | Club) + Reputation + (1 | Position)
+ poly(Age,
    2)
   Data: completedfifa[1:4000, ]

REML criterion at convergence: 9323.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6339 -0.6071 -0.0103  0.6066  3.4118

Random effects:
 Groups   Name        Variance Std.Dev.
 Club     (Intercept) 0.69796  0.8354
 Position (Intercept) 0.04138  0.2034
 Residual             0.42640  0.6530
Number of obs: 4000, groups:  Club, 564; Position, 4

Fixed effects:
```

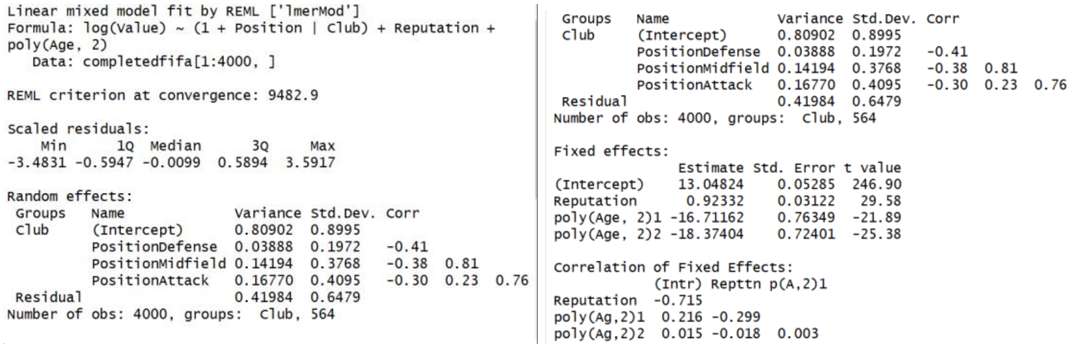
```
Random effects:
 Groups   Name        Variance Std.Dev.
 Club     (Intercept) 0.69796  0.8354
 Position (Intercept) 0.04138  0.2034
 Residual             0.42640  0.6530
Number of obs: 4000, groups:  Club, 564; Position, 4

Fixed effects:
               Estimate Std. Error t value
(Intercept)    13.11090    0.11476  114.25
Reputation      0.86979    0.03064   28.39
poly(Age, 2)1 -15.68738    0.75637  -20.74
poly(Age, 2)2 -17.96038    0.70979  -25.30

Correlation of Fixed Effects:
            (Intr) Repttn p(A,2)1
Reputation  -0.325
poly(Ag,2)1  0.088 -0.297
poly(Ag,2)2  0.004 -0.025  0.011
```

c) Random slope model



```
soccer4 <- lmer(data = completedfifa[1:4000,], log(Value) ~ (1+Position|Club) + Reputation +
poly(Age,2))
summary(soccer4)
```

Summary of this model

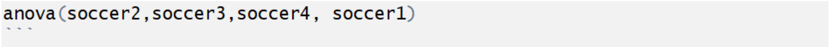


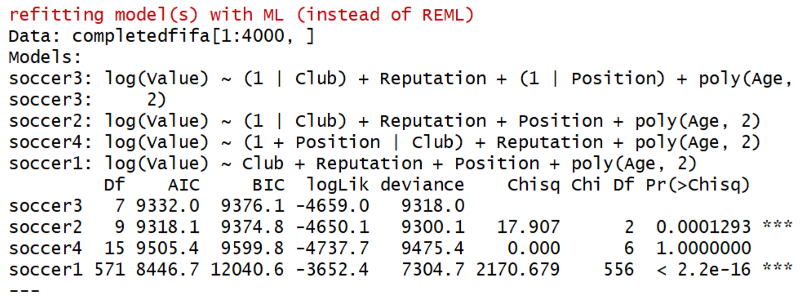
```
Linear mixed model fit by REML ['lmerMod']
Formula: log(Value) ~ (1 + Position | Club) + Reputation +
poly(Age, 2)
   Data: completedfifa[1:4000, ]

REML criterion at convergence: 9482.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4831 -0.5947 -0.0099  0.5894  3.5917

Random effects:
 Groups   Name             Variance Std.Dev. Corr
 Club     (Intercept)      0.80902  0.8995
          PositionDefense  0.03888  0.1972   -0.41
          PositionMidfield 0.14194  0.3768   -0.38  0.81
          PositionAttack   0.16770  0.4095   -0.30  0.23  0.76
 Residual                  0.41984  0.6479
Number of obs: 4000, groups:  Club, 564
```

```
 Groups   Name             Variance Std.Dev. Corr
 Club     (Intercept)      0.80902  0.8995
          PositionDefense  0.03888  0.1972   -0.41
          PositionMidfield 0.14194  0.3768   -0.38  0.81
          PositionAttack   0.16770  0.4095   -0.30  0.23  0.76
 Residual                  0.41984  0.6479
Number of obs: 4000, groups:  Club, 564

Fixed effects:
               Estimate Std. Error t value
(Intercept)    13.04824    0.05285  246.90
Reputation      0.92332    0.03122   29.58
poly(Age, 2)1 -16.71162    0.76349  -21.89
poly(Age, 2)2 -18.37404    0.72401  -25.38

Correlation of Fixed Effects:
            (Intr) Repttn p(A,2)1
Reputation  -0.715
poly(Ag,2)1  0.216 -0.299
poly(Ag,2)2  0.015 -0.018  0.003
```

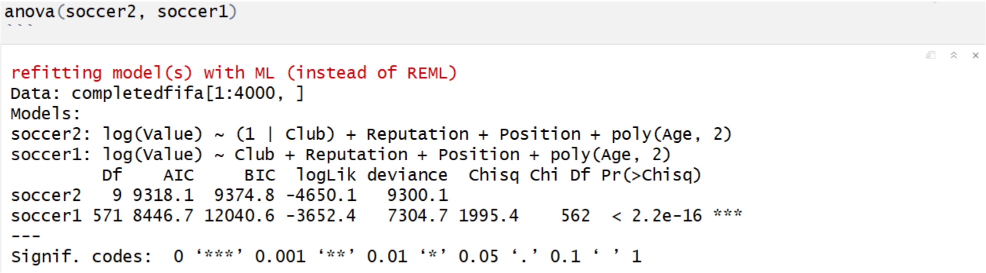
We cannot find out which one is the best-fitted model by looking at their model summaries. And since they are all multilevel models with the same variables, their residual plots looks like the same. In order to evaluate these models, I used the ANOVA test to do the comparison.

ANOVA test: Four models.



```
anova(soccer2,soccer3,soccer4, soccer1)
`````
```



```
refitting model(s) with ML (instead of REML)
Data: completedfifa[1:4000, ]
Models:
soccer3: log(Value) ~ (1 | Club) + Reputation + (1 | Position) + poly(Age,
soccer3:     2)
soccer2: log(Value) ~ (1 | Club) + Reputation + Position + poly(Age, 2)
soccer4: log(Value) ~ (1 + Position | Club) + Reputation + poly(Age, 2)
soccer1: log(Value) ~ Club + Reputation + Position + poly(Age, 2)
         Df    AIC     BIC  logLik deviance    Chisq Chi Df Pr(>Chisq)
soccer3   7 9332.0  9376.1 -4659.0   9318.0
soccer2   9 9318.1  9374.8 -4650.1   9300.1   17.907      2  0.0001293 ***
soccer4  15 9505.4  9599.8 -4737.7   9475.4    0.000      6  1.0000000
soccer1 571 8446.7 12040.6 -3652.4   7304.7 2170.679    556  < 2.2e-16 ***
---
```

According to the ANOVA test, among the multilevel models, soccer 2 with one random effect is more appropriate since it fits significantly better but others are totally not. Here comes the comparison of linear model soccer1 and multilevel model soccer2.
```

ANOVA test: soccer1 (linear model), soccer2 (multilevel model)



```
anova(soccer2, soccer1)
```

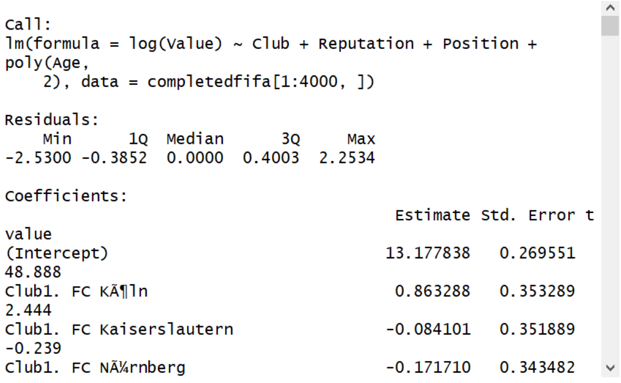
```
refitting model(s) with ML (instead of REML)
Data: completedfifa[1:4000, ]
Models:
soccer2: log(Value) ~ (1 | Club) + Reputation + Position + poly(Age, 2)
soccer1: log(Value) ~ Club + Reputation + Position + poly(Age, 2)
         Df    AIC     BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
soccer2   9 9318.1  9374.8 -4650.1   9300.1
soccer1 571 8446.7 12040.6 -3652.4   7304.7 1995.4    562  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As the result of ANOVA test shows, soccer1 fits significantly better than soccer2, and the P-value is really small. In addition, the median residual of linear model soccer 1 is 0.000, while the median residual of multilevel model soccer2 is -0.0042, which means that soccer 1 fits slightly better than soccer2. So finally the linear model might be more appropriate for interpretation.

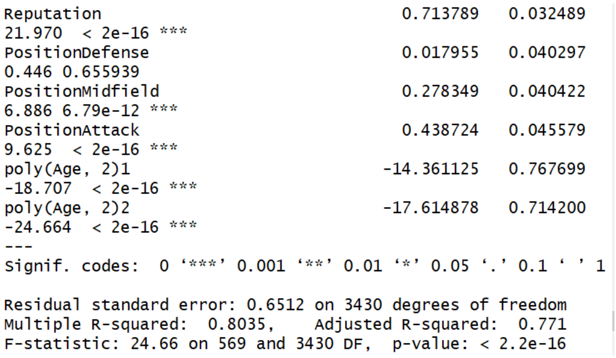
Result:

#3.1 Model used and interpretation

Let’s recall the summary of linear model soccer1.



```
Call:
lm(formula = log(Value) ~ Club + Reputation + Position +
poly(Age,
    2), data = completedfifa[1:4000, ])

Residuals:
    Min      1Q  Median      3Q     Max
-2.5300 -0.3852  0.0000  0.4003  2.2534

Coefficients:
                                        Estimate Std. Error t
value
(Intercept)                            13.177838   0.269551
48.888
Club1. FC KÃ¶ln                          0.863288   0.353289
2.444
Club1. FC Kaiserslautern               -0.084101   0.351889
-0.239
Club1. FC NÃ¼rnberg                    -0.171710   0.343482
```



```
Reputation                                0.713789    0.032489
21.970   < 2e-16 ***
PositionDefense                           0.017955    0.040297
0.446 0.655939
PositionMidfield                          0.278349    0.040422
6.886 6.79e-12 ***
PositionAttack                            0.438724    0.045579
9.625   < 2e-16 ***
poly(Age, 2)1                           -14.361125    0.767699
-18.707   < 2e-16 ***
poly(Age, 2)2                           -17.614878    0.714200
-24.664   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6512 on 3430 degrees of freedom
Multiple R-squared:  0.8035,    Adjusted R-squared:  0.771
F-statistic: 24.66 on 569 and 3430 DF,  p-value: < 2.2e-16
```

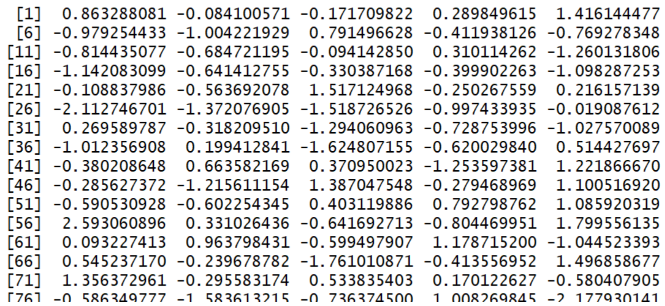
Reputation: We can clearly see that if a player’s reputation increases by one unit, then his log(value) increases by 0.713789.

Positions: The coefficient of Defense is 0.003312. The coefficient of Midfield is 0.266148. The coefficient of Attack is 0.425964. We can see that, for Players in the same club with the same age and reputation, players whose positions are Defense would have slightly more personal value in EURO than Goalkeepers. But for players in the same club with same age and reputation whose positions are Midfield and Attack, they would have obviously more market value compared to other positions, which may be because audience and club managers pay more attention to players who can strike the goal.

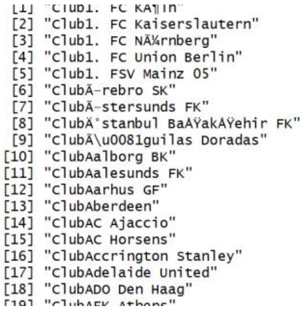
Age: For variables of age: poly(Age, 1) and poly(Age, 2), which coefficients are -16.216151 and -19.075873, both shows that age has a great influence on players' market value. As players get old, their market value gradually decreases.

#3.2 Interpretation for coefficients of Clubs.

In order to have a better interpretation for clubs, it would be a good way to divide the clubs into several groups based on their similarities of coefficients. Firstly, I made a list of coefficients corresponding to a list of club names.



```
 [1]  0.863288081 -0.084100571 -0.171709822  0.289849615  1.416144477
 [6] -0.979254433 -1.004221929  0.791496628 -0.411938126 -0.769278348
[11] -0.814435077 -0.684721195 -0.094142850  0.310114262 -1.260131806
[16] -1.142083099 -0.641412755 -0.330387168 -0.399902263 -1.098287253
[21] -0.108837986 -0.563692078  1.517124968 -0.250267559  0.216157139
[26] -2.112746701 -1.372076905 -1.518726526 -0.997433935 -0.019087612
[31]  0.269589787 -0.318209510 -1.294060963 -0.728753996 -1.027570089
[36] -1.012356908  0.199412841 -1.624807155 -0.620029840  0.514427697
[41] -0.380208648  0.663582169  0.370950023 -1.253597381  1.221866670
[46] -0.285627372 -1.215611154  1.387047548 -0.279468969  1.100516920
[51] -0.590530928 -0.602254345  0.403119886  0.792798762  1.085920319
[56]  2.593060896  0.331026436 -0.641692713 -0.804469951  1.799556135
[61]  0.093227413  0.963798431 -0.599497907  1.178715200 -1.044523393
[66]  0.545237170 -0.239678782 -1.761010871 -0.413556952  1.496858677
[71]  1.356372961 -0.295583174  0.533835403  0.170122627 -0.580407905
[76] -0.586349777 -1.583613215 -0.736374500  1.008269845 -2.177930141
```

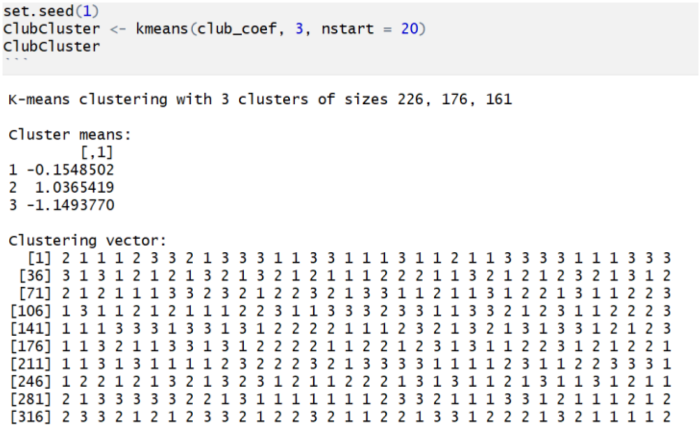


```
 [1] "Club1. FC KÃ¶ln"
 [2] "Club1. FC Kaiserslautern"
 [3] "Club1. FC NÃ¼rnberg"
 [4] "Club1. FC Union Berlin"
 [5] "Club1. FSV Mainz 05"
 [6] "ClubÃ–rebro SK"
 [7] "ClubÃ–stersunds FK"
 [8] "ClubÄ°stanbul BaÅŸakÅŸehir FK"
 [9] "ClubÃ\u0081guilas Doradas"
[10] "ClubAalborg BK"
[11] "ClubAalesunds FK"
[12] "ClubAarhus GF"
[13] "ClubAberdeen"
[14] "ClubAC Ajaccio"
[15] "ClubAC Horsens"
[16] "ClubAccrington Stanley"
[17] "ClubAdelaide United"
[18] "ClubADO Den Haag"
[19] "ClubAEK Athens"
```

List of coefficients

List of club names

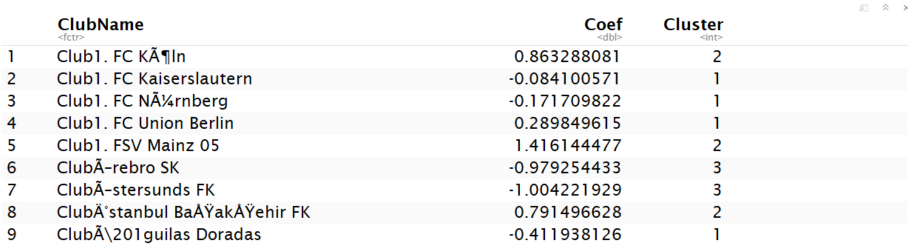
Then I clustered the clubs based on their coefficients, by using the method called k-means clustering.



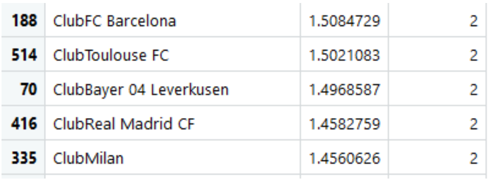
~~~
set.seed(1)
ClubCluster <- kmeans(club_coef, 3, nstart = 20)
ClubCluster
```
~~~

```
K-means clustering with 3 clusters of sizes 226, 176, 161

Cluster means:
        [,1]
1 -0.1548502
2  1.0365419
3 -1.1493770

Clustering vector:
  [1] 2 1 1 1 2 3 3 2 1 3 3 3 1 1 3 3 1 1 1 3 1 1 2 1 1 3 3 3 3 1 1 1 3 3 3
 [36] 3 1 3 1 2 1 2 1 3 2 1 3 2 1 2 1 1 1 2 2 2 1 1 3 2 1 2 1 2 3 2 1 3 1 2
 [71] 2 1 2 1 1 1 3 3 2 3 2 1 2 2 3 2 1 3 3 1 1 2 1 1 3 1 2 2 1 3 1 1 2 2 3
[106] 1 3 1 1 2 1 2 1 1 1 2 2 3 1 1 3 3 3 2 3 3 1 1 3 3 2 1 2 3 1 1 2 2 2 3
[141] 1 1 1 3 3 3 1 3 3 1 3 1 2 2 2 2 1 1 1 2 3 2 1 3 2 1 3 1 3 3 1 2 1 2 3
[176] 1 1 3 2 1 1 3 3 1 3 1 2 2 2 2 1 1 2 2 1 2 3 1 3 1 1 2 2 3 1 2 1 2 2 1
[211] 1 1 3 1 3 1 1 1 1 2 3 2 2 2 3 2 1 3 3 3 3 1 1 1 1 2 3 1 1 2 2 3 3 3 1
[246] 1 2 2 1 2 1 3 2 1 3 2 3 1 2 1 1 2 2 2 1 3 1 3 1 1 2 1 3 1 1 3 1 2 1 1
[281] 2 1 3 3 3 3 3 2 2 1 3 1 1 1 1 1 1 1 2 3 3 2 1 1 1 3 3 1 2 1 1 1 2 1 2
[316] 2 3 3 2 1 2 1 2 3 3 2 1 2 2 3 2 1 1 2 2 1 3 3 1 2 2 2 1 3 2 1 1 1 1 2
```

From this image, we can find that the coefficients of clubs have been divided into three clusters: (1. -0.1548502) (2. 1.0365419) (3. -1.1493770). Next, I merged the cluster number with corresponding club's name and coefficient to create a new data table 'cluster_df', which is shown below:



	ClubName <fctr>	Coef <dbl>	Cluster <int>
1	Club1. FC KÃ¶ln	0.863288081	2
2	Club1. FC Kaiserslautern	-0.084100571	1
3	Club1. FC NÃ¼rnberg	-0.171709822	1
4	Club1. FC Union Berlin	0.289849615	1
5	Club1. FSV Mainz 05	1.416144477	2
6	ClubÃ–rebro SK	-0.979254433	3
7	ClubÃ–stersunds FK	-1.004221929	3
8	ClubÄ°stanbul BaÅŸakÅŸehir FK	0.791496628	2
9	ClubÃ\201guilas Doradas	-0.411938126	1

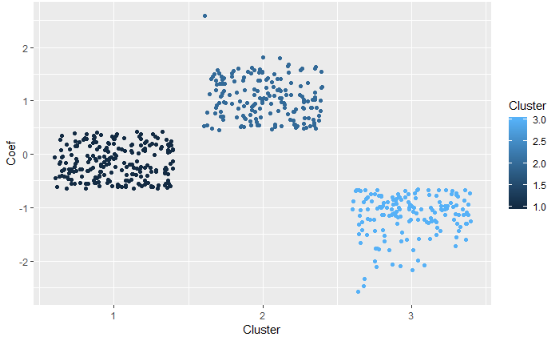


188	ClubFC Barcelona	1.5084729	2
514	ClubToulouse FC	1.5021083	2
70	ClubBayer 04 Leverkusen	1.4968587	2
416	ClubReal Madrid CF	1.4582759	2
335	ClubMilan	1.4560626	2

The data table clearly show that many famous soccer clubs are in cluster 2.

After that I use ggplot to visual the clusters, which is easier to interpret.

```
ggplot(data=cluster_df, aes(x=Cluster, y=Coef, color=Cluster))+geom_jitter()
```



From this visual, we can easily find out that the clubs have been divided into three clusters based on their coefficient value. The coefficient of clubs in cluster 1 is near equals to 0, which means that the market value of players who work for these clubs would not be largely affected by their clubs. Similarly, the market value of players who work for clubs in cluster 3 would have a decrease affected by their clubs, while the market value of players who work for clubs in cluster 2, such as Real Madrid and FC Barcelona, would be higher than average because of their clubs.

#3.3 Model Checking

Right now I use the linear model ‘soccer1’ to predict the rest of the data for checking the accuracy of this model. And then we use the formula for computing the MAPE (mean absolute percentage error):

MAPE = ABS (Actual – predict)/Actual)

```
soccer1.predict <- predict(soccer1,completedfifa[4001:4963,])

accuracy <-mean(abs(soccer1.predict -
log(completedfifa[4001:4963,]$Value))/log(completedfifa[4001:4963,]$Value))
accuracy

```
```

```
 [1] 0.03703817
```

Here we can see that the MAPE value is 3.7%, which means that the model makes a good prediction and it fits well.
```

## Discussion:

### #4.1 Implication

The model, as well as interpretations, can help people such as club managers and soccer fans roughly predict a player's market value, and understand how it could be changed if the player transfers to another club, makes his reputation decrease for some reasons or becomes older.

This model refers that players' age has deeply impacted their market value, which was not I expected. And I thought that positions do not make a difference of market value, but the truth is that the average market value of Attack and Midfield is much higher than Defense and Goalkeeper.

### #4.2 Limitation and Further Direction

The variables might be not enough for fitting a model of player's market value, because there are too many other possibilities, except for their abilities, might affect their market values such as health condition, club manager's and coach's appreciation, or advertising value. I might find more related variables for fitting a more concrete model, which can be credibly used for predict a soccer player's personal market value.

## Acknowledgement:

First, please allow me to express my sincere appreciation to our professor Yajima, who always gave me useful suggestions and encouragement when I met with difficulties in my project, such as model selection and improvement. Also, I would like to thank my teaching assistant and friend, Alice Tao, who kindly offered me suggestions about my method using. With their help, I felt great confidence to accomplish my work.

## Reference:

KevinH. (2017, December 26). Fifa 18 More Complete Player Dataset. Retrieved from https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset/data

Agarwal, S. (2017, April 13). Complete FIFA 2017 Player dataset (Global). Retrieved from https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global

Verma, A. (2018, May 24). Exploratory Analysis of FIFA 18 dataset using R – Towards Data Science. Retrieved from https://towardsdatascience.com/exploratory-analysis-of-fifa-18-dataset-using-r-ba09aa4a2d3c