Final project data analysis

Abstract. In this paper, Jiyoon (Clover) Jeong and Jin Kweon are trying to inspect the official FIFA 2017 data. Our goal is to find out how the FIFA ratings on the players were decided. The better rating indicates how valuable the players are. We were curious whether the rating well indicates the players' stats. Our data contains 17588 players with 53 different variables.

Q.

What should we write for executive summary…?

Do we need to double-space…?

Does everything look good…?

Can I include codes under appendix….?

How do we really make results and conclusion parts, as we explained all under Method section…?

Executive summary:

Background:

Two of us have pretty solid prior knowledge of soccer, and we used some of our intuitions when we do analysis. For example, When we do test out the PCAs, it is possible for us not just test all the quantitative variables, but select some of the variables that we thought would be reasonable to test. (which is more efficient)

Another important thing we decide to do for our analysis is to focus on clubs. Players play either on their national team or clubs. And, it is definitely possible for players to play for different positions and kit number in national and clubs. That way, it makes much harder for us to analyze and draw conclusions if we consider both. As players are selected to represent their countries based on their performance in their teams and players spend much more time in clubs, we decide to let players' club profiles as our major variables. (however, depends on the situation, we used national information, and we are going to explicitly say it if so. Also, please refer to appendix for the R codes.

1. Introduction

a) Problem (the question I want to address)

1. We want to know how the ratings (and ages) are different based on club positions (goalkeeper, defense, midfielder, attacker) and preferred foot of the players.

Reason: There has been enough arguments that FIFA has brought more and unfair attentions onto attacker (sometimes, midfielder as well), and other positions are treated/rated unfairly. And, we want to check whether the argument is true. Also, we are just curious how ratings are different amongst the players with different preferred foot. (also at the same time, the variable "Age" is really important to soccer players, so I want to include this for our quantitative variable. And, we assumed ratings and ages are closely related.)

Hypothesis: The means for each group of position-effect, preferred-foot-effect, and interaction are the same.

Aim/objective: Subjects will be 3656 soccer player (but different groups of population). As we stated above under the "Data" section, we filtered out the unspecified positions and modified the data, and 3656 samples are pretty decent size of samples. (around 20% of entire data) The quantitative measurement/variable are ratings and ages. There are 2 factors: club positions and preferred foot. There are 4 levels for club position: goalkeeper, defense, midfielder, and attacker. There are 2 levels for preferred foot: left and right. More than two different populations with two factors implies that I need to use "two-way MANOVA."

2. We are going to find out what quantitative variables are related with the rating.

Reason: It is reasonable to assume that the overall ratings of professional soccer players are proportional to their score variables such as 'Weak_foot', 'Skill_Moves', 'Ball_control', … , 'GK_Reflexes'. The assumption that the higher these score variable means the higher overall ratings is reasonable and we want to check which predictors are more stronger than other predictors by linear regression and transformation of variables.

Hypothesis: There should exist some quantitative variables that have a linear relationship with rating.

Aim/objective: Detect the predictors which have strong linear relationship with variable 'Ratings' and find proper transformation of variables if needed.

3. I want to test how much each variable is correlated with the rating. PCA will also help me find co-linearity issue.

Reason: We aim to find PCs to best summarize the variables, and see how players' rankings are plotted.

Hypothesis: Different positions have different variables correlated with the rating, and I should find the skills that are important to the position should have high correlation with the rating.

Aim/objective: I hope to make good interpretation of the components to explain the the relationships between variables, and eventually help us how rating can be explained by other variables.

b) Data (summary of the data, the study design, data collection)

Please refer to the data dictionary we made: https://github.com/yjkweon24/public-health-245/blob/master/dictionary.csv

We collected our data in Kaggle website (please refer to the reference). The original data has 17588 rows and 53 columns.

It is important how we sample the data. Players are selected to play in the national team if they perform well in the club. It is true nowadays in the soccer world, players spend more time playing for the club. So, we are focusing on inspecting players' club profiles only.

Although national related information is not our major variables, we our not going to take them out, as these information help us as some of the players do not have enough club informations. In this case, we replace national information with club's. For example, many players do not have specific positions ("Sub") for their clubs, and we tried to find these missed informations from their national positions, if possible.

To do the better analysis, we tried to be really careful of choosing the right sample. As we care about club positions more, we replace club positions with national positions if club positions are "Sub," "Res," and empty(" "), and this will help us a lot when we want to analyze the relationship between positions and ratings. Whenever we do analysis that has to do with positions, we needed to work on the extracted samples of the size 3656, as the others do not show clear positions.

We examined the raw data and found that variable 'National_Position' has 16513 missing values and variable 'National_Kit' has 16513 NA values. Variable 'Club_Position' and 'Club_Joining' has one missing value which is 384th observation, and variable 'Contract_Expiry' and 'Club_Kit' has one NA values which is also 384th observation.

c) Purpose of the study

The purpose of this study is to examine how the ratings were decided, and we hope to correct them if some of the players are either over- or under- rated.

2. Methods

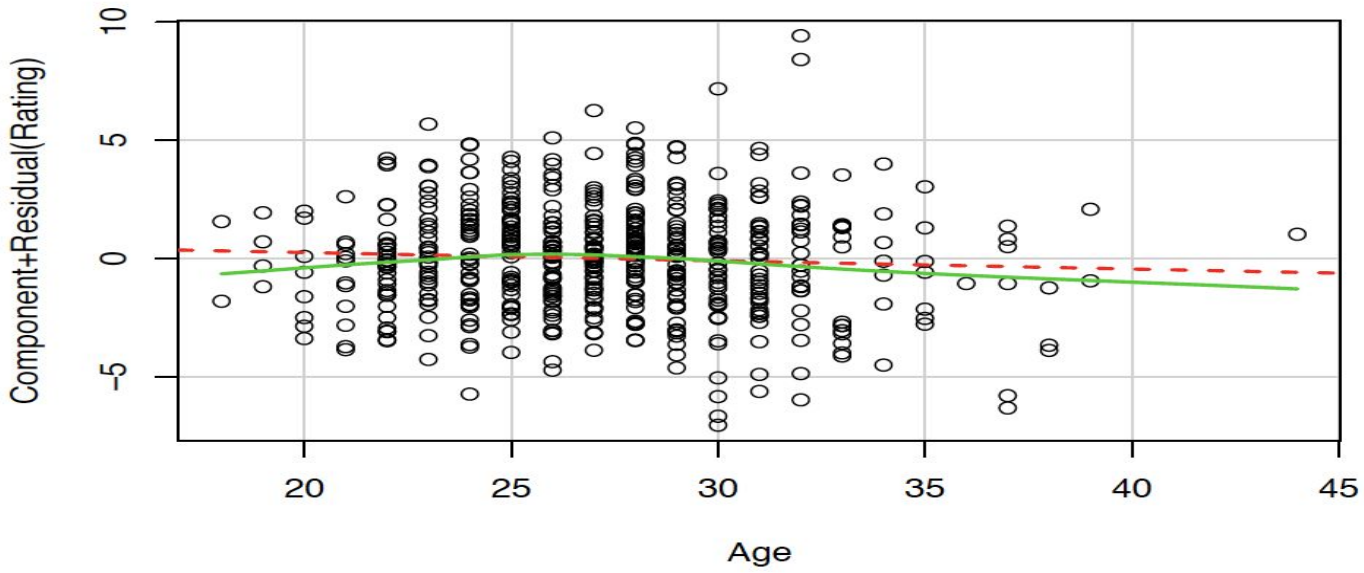
a) Method (my choice of model, analytic method, why)

1. Multivariate test:

As I mentioned above under the “Problem,” our subject will be 3656 soccer players (but different groups of population). The quantitative variables are ratings and ages (and as we assume ratings and ages are somewhat closely related, we can say that our measurement is generally just ratings. Intuitively, for most of the cases, it is true for ratings and ages. To prove my points, I include the lm() summary. Again, this is not 100% correlated, but I assumed to be), and there are 2 factors: club positions (4 levels) and preferred foot (2 levels). So, I need to use Two-way MANOVA. Our team decides to conduct this test to see if there any difference of means from many groups.

Here is the lm() output, below: (as you can see the variable “Age” is significant, but this is not perfect way, as this is just a linear model with marginal effect. For example, there might be some other variables affecting the relationship between Rating and Age variables, so I decided to Component Plus Residual plot to see better linear relationship between them.)

```
## Call:
## lm(formula = Rating ~ Age, data = soccer_positioncleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7383  -3.9101  -0.5628   3.4372  21.5670
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.47142    0.65801  99.499   <2e-16 ***
## Age          0.21755    0.02379   9.144   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



As we can see the test output, we can ambiguously argue that ratings and ages are kind-of related.

Here is our test, below:

My null hypotheses H_0 are: (μ and β are means for each group. For example, μ_{11} will be the mean of both first level of both factors)

1. $H_0^{int}: \mu_{11} = \mu_{21} = \mu_{31} = \mu_{41} = \mu_{21} = \mu_{22} = \mu_{23} = \mu_{24} = 0$ (no interaction effect)
2. $H_0^{fac1}: \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$
3. $H_0^{fac2}: \beta_1 = \beta_2 = 0$ (I am following the notations from the textbook)

Test statistics for each H_0 are (where Λ^* is Wilk's lambda and SSP = Sum of Squares and cross Products):

1. $\Lambda^*_{int} = \frac{SSP_{res}}{SSP_{int} + SSP_{res}}$
2. $\Lambda^*_{fac1} = \frac{SSP_{res}}{SSP_{fac1} + SSP_{res}}$
3. $\Lambda^*_{fac2} = \frac{SSP_{res}}{SSP_{fac2} + SSP_{res}}$

I will do a test for interaction before the tests for main factor effects, because if interaction effects exist, the factor effects do not have a clear interpretation. Thus, we do not need to proceed additional multivariate tests (pg.316)

Here is our test summary outputs, below:

```
##                                                    Df   Wilks approx F
## as.factor(Club_Position)                            3 0.95995  25.0976
## as.factor(Preffered_Foot)                           1 0.99788   3.8678
## as.factor(Club_Position):as.factor(Preffered_Foot)  3 0.99648   2.1463
## Residuals                                        3648
##                                                  num Df den Df  Pr(>F)
## as.factor(Club_Position)                               6   7294 < 2e-16
## as.factor(Preffered_Foot)                             2   3647 0.02099
## as.factor(Club_Position):as.factor(Preffered_Foot)     6   7294 0.04515
## Residuals
##
## as.factor(Club_Position)                          ***
## as.factor(Preffered_Foot)                        *
## as.factor(Club_Position):as.factor(Preffered_Foot) *
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  Response Rating :
##                                                    Df  Sum Sq Mean Sq
## as.factor(Club_Position)                             3    1260  419.98
## as.factor(Preffered_Foot)                           1      15   14.76
## as.factor(Club_Position):as.factor(Preffered_Foot)    3     174   58.02
## Residuals                                         3648  125462   34.39
##                                                   F value    Pr(>F)
## as.factor(Club_Position)                          12.2117 5.993e-08 ***
## as.factor(Preffered_Foot)                          0.4291    0.5125
## as.factor(Club_Position):as.factor(Preffered_Foot)  1.6869    0.1676
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Age :
##                                                    Df  Sum Sq Mean Sq
## as.factor(Club_Position)                             3    1783  594.36
## as.factor(Preffered_Foot)                           1     105  105.11
## as.factor(Club_Position):as.factor(Preffered_Foot)    3     109   36.30
## Residuals                                         3648   57985   15.90
##                                                   F value  Pr(>F)
## as.factor(Club_Position)                          37.3930 < 2e-16 ***
## as.factor(Preffered_Foot)                          6.6127 0.01016 *
## as.factor(Club_Position):as.factor(Preffered_Foot)  2.2834 0.07701 .
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our conclusion is that we reject all of three null hypothesis. Thus, there is no position effect, preferred foot effect, and position-preferred foot interaction effects on ratings (based on the assumption that ratings and ages are correlated well enough).

2. Linear regression:

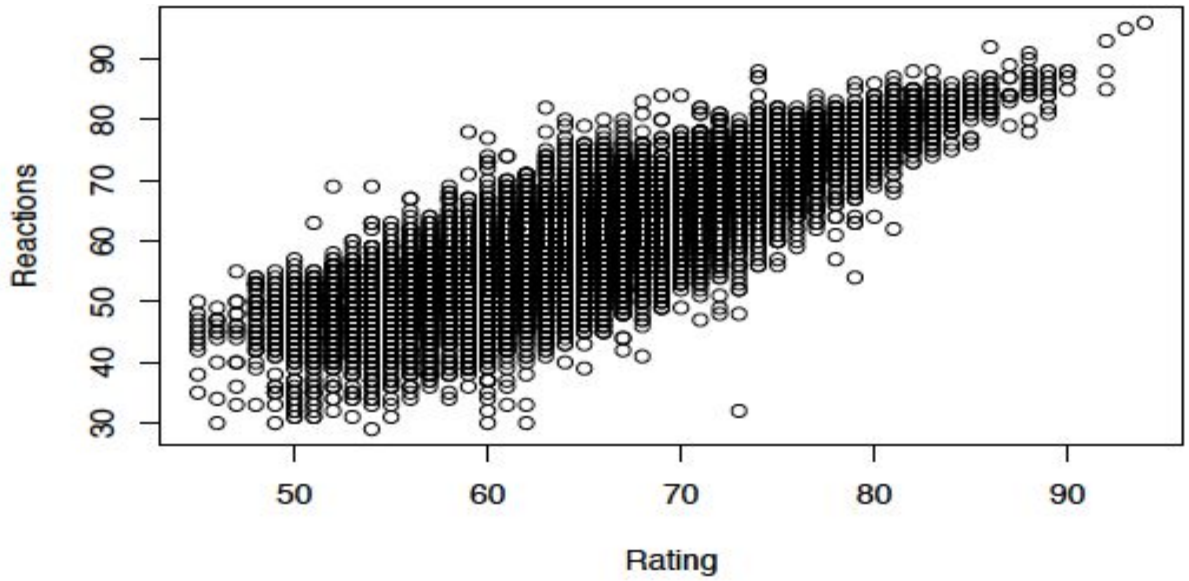
We will use linear regression coefficients, T-test, and F-test to find which predictors(quantitative and qualitative variables in factors) are significant by regressing YYYYY on some XXXXX (you can fill in -Jin).

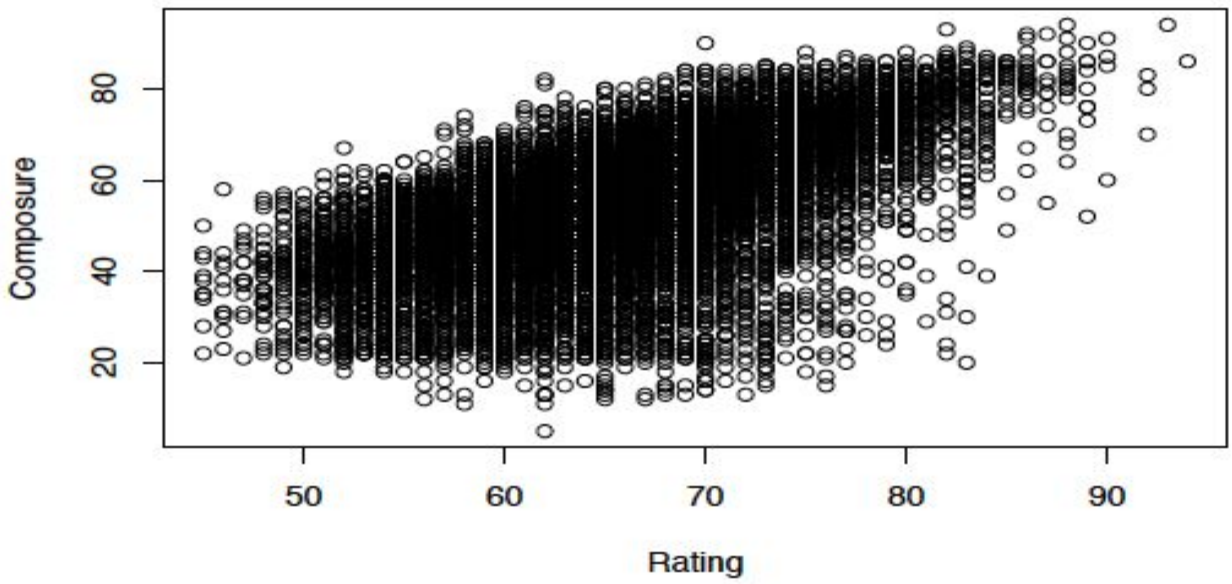
(a) Fit a linear regression model with percent body fat using rating as the response and other variable ‘ ‘ as the predictors. Check significant predictors & t-test for significance of each variable

```
##
## Call:
## lm(formula = formula, data = soccer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0853 -1.7691  0.0305  1.7416 13.7478
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.016e+02  1.090e+02   0.932 0.351772
## Contract_Expiry    -4.783e-02  5.373e-02  -0.890 0.373625
## Height              2.827e-02  2.547e-02   1.110 0.267408
## Weight              2.026e-02  2.051e-02   0.988 0.323516
## Age                -3.613e-02  2.639e-02  -1.369 0.171329
## Weak_foot           1.091e-01  1.205e-01   0.905 0.365693
## Skill_Moves         9.184e-01  1.858e-01   4.944 8.95e-07 ***
## Ball_Control        1.373e-01  2.083e-02   6.588 7.07e-11 ***
## Dribbling           2.512e-02  1.770e-02   1.419 0.156257
## Marking            -2.711e-02  1.619e-02  -1.674 0.094394 .
## Sliding_Tackle      2.619e-02  1.793e-02   1.461 0.144348
## Standing_Tackle     4.545e-02  1.883e-02   2.413 0.015982 *
## Aggression         -6.374e-05  9.087e-03  -0.007 0.994404
## Reactions           3.849e-01  1.638e-02  23.500  < 2e-16 ***
## Attacking_Position -5.958e-02  1.323e-02  -4.504 7.41e-06 ***
## Interceptions      -2.431e-02  1.319e-02  -1.843 0.065663 .
## Vision             -2.057e-02  1.169e-02  -1.760 0.078668 .
## Composure           7.975e-02  1.102e-02   7.236 9.00e-13 ***
## Crossing           -1.575e-03  1.151e-02  -0.137 0.891180
## Short_Pass          7.323e-02  2.017e-02   3.631 0.000296 ***
## Long_Pass           1.151e-02  1.505e-02   0.764 0.444768
## Acceleration        2.489e-02  1.864e-02   1.335 0.182161
## Speed               1.689e-02  1.737e-02   0.972 0.331058
## Stamina            -1.441e-02  1.175e-02  -1.226 0.220610
## Strength            3.492e-02  1.207e-02   2.895 0.003877 **
## Balance            -9.942e-03  1.293e-02  -0.769 0.442133
## Agility             1.790e-02  1.301e-02   1.376 0.169141
## Jumping             1.398e-02  9.300e-03   1.503 0.133052
## Heading             7.721e-02  1.106e-02   6.980 5.27e-12 ***
## Shot_Power          2.194e-02  1.242e-02   1.767 0.077581 .
## Finishing          -2.083e-03  1.392e-02  -0.150 0.881087
## Long_Shots         -2.653e-02  1.369e-02  -1.937 0.052971 .
## Curve              -1.889e-03  1.194e-02  -0.158 0.874330
## Freekick_Accuracy   2.100e-02  9.540e-03   2.201 0.027953 *
## Penalties          -8.935e-03  1.025e-02  -0.871 0.383708
## Volleys             8.667e-03  1.165e-02   0.744 0.457269
## GK_Positioning      2.915e-02  2.429e-02   1.200 0.230337
## GK_Diving           2.824e-02  2.472e-02   1.142 0.253548
## GK_Kicking          6.343e-02  2.219e-02   2.859 0.004340 **
## GK_Handling         1.110e-01  2.410e-02   4.605 4.64e-06 ***
## GK_Reflexes         6.201e-02  2.437e-02   2.545 0.011075 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.66 on 1034 degrees of freedom
##   (16513 observations deleted due to missingness)
## Multiple R-squared:  0.8592, Adjusted R-squared:  0.8538
## F-statistic: 157.8 on 40 and 1034 DF,  p-value: < 2.2e-16
```

We converted Height and Weight variable into numeric variable and fit a linear regression model using Ratings as the response and the all other quantitative variables as the predictors. The summary of this fit shows interesting result. We expected that most of the score variables('Weak_foot', 'Skill_moves', …. , 'GK_Reflexes') will have at least 0.01 significance in the beginning. However, the lm results clearly shows that only 10 variables are strongly related to their ratings. (T test) If we set significant level as 0.001, variable 'Skill_Moves', 'Ball_Control', 'Reactions', 'Attacking_Position', 'Composure', 'Short_Pass', 'Heading', 'GK_Handling' are significant among 41 predictors.

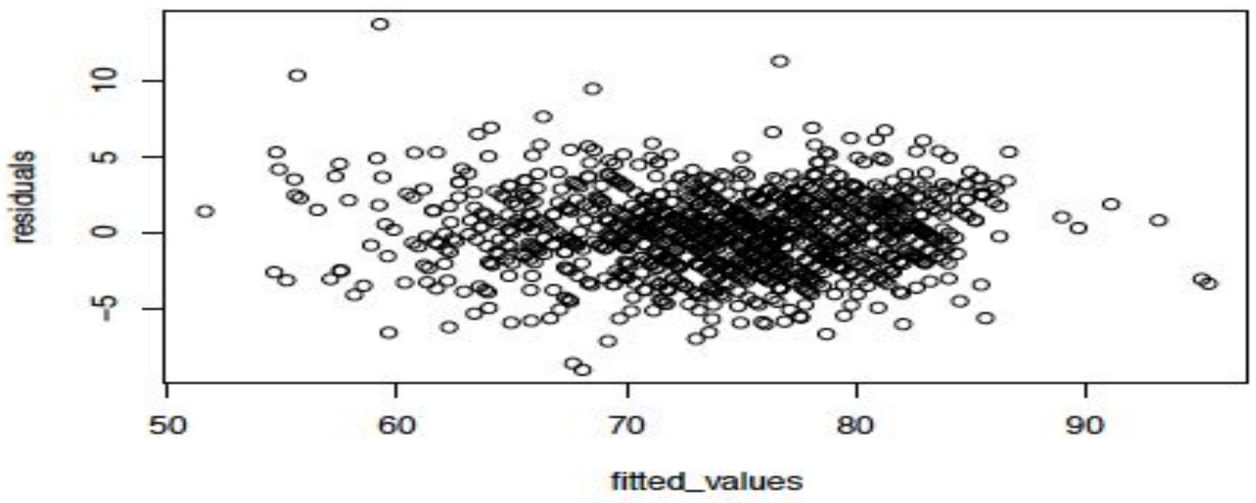
(b) Data visualization - shows several plots of significant variables in part (a) VS ratings





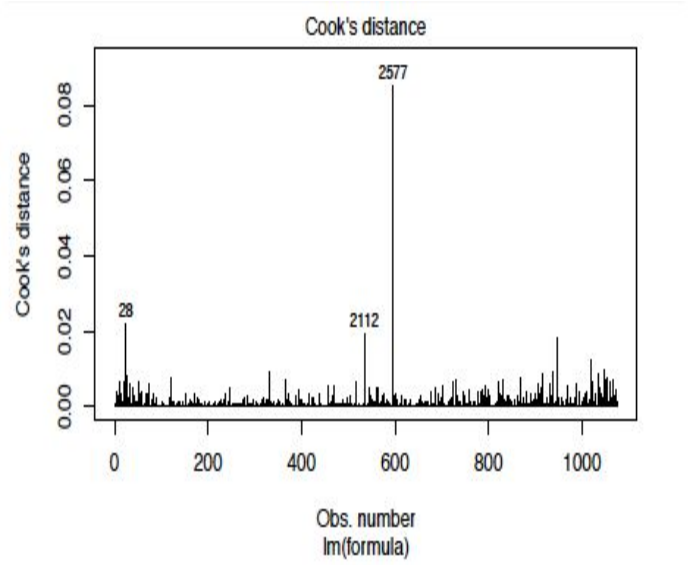
These variable shows clear linear relationship between variable and the response (ratings)

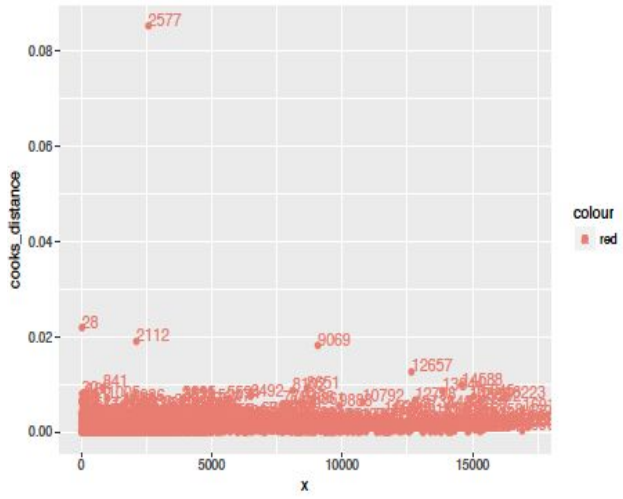
(c) Draw a residual plot, with the fitted values on the x-axis, and the residuals on the y-axis to check if there is any violation of assumptions of the linear model. (linear or not, variance constant, normal or not)

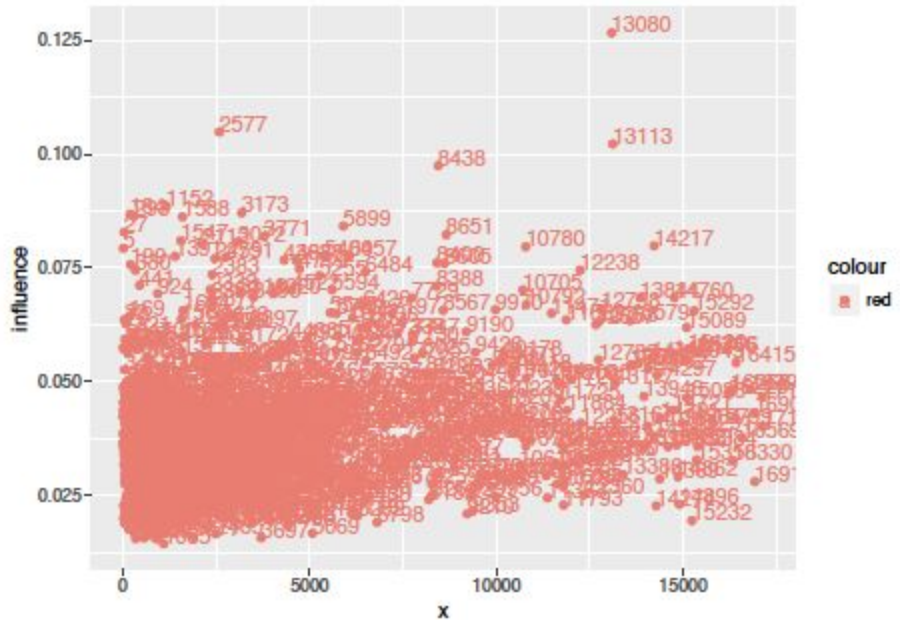


The residual plot shows that it does not particularly violate linear model assumption since the residuals are symmetric to y = 0 axis and does not show specific patterns. Also, variance seems like a constant too.

(d) Plots of Cook's distance and influential points to detect outliers







2577 28 2112 13080

Let’s remove 2577th, 28th, 2112, and 13080th players and fit again.

```
##
## Call:
## lm(formula = formula, data = X_refit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8500 -1.7297  0.0207  1.7321 10.6755
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.144e+02  1.064e+02   1.075 0.282671
## Contract_Expiry    -5.489e-02  5.246e-02  -1.046 0.295666
## Height              3.505e-02  2.487e-02   1.410 0.158957
## Weight              1.484e-02  1.999e-02   0.743 0.457950
## Age                -2.939e-02  2.573e-02  -1.142 0.253582
## Weak_foot           1.124e-01  1.175e-01   0.956 0.339109
## Skill_Moves         8.717e-01  1.812e-01   4.812 1.72e-06 ***
## Ball_Control        1.555e-01  2.066e-02   7.524 1.15e-13 ***


## Dribbling           1.841e-02  1.756e-02   1.048 0.294798
## Marking            -2.763e-02  1.578e-02  -1.751 0.080283 .
## Sliding_Tackle      2.691e-02  1.747e-02   1.541 0.123642
## Standing_Tackle     4.367e-02  1.834e-02   2.381 0.017432 *
## Aggression         -8.024e-04  8.852e-03  -0.091 0.927793
## Reactions           3.956e-01  1.642e-02  24.090  < 2e-16 ***
## Attacking_Position -5.416e-02  1.304e-02  -4.153 3.55e-05 ***
## Interceptions      -2.469e-02  1.289e-02  -1.915 0.055761 .
## Vision             -2.129e-02  1.140e-02  -1.868 0.062112 .
## Composure           7.897e-02  1.074e-02   7.355 3.88e-13 ***
## Crossing           -5.519e-03  1.122e-02  -0.492 0.622917
## Short_Pass          6.638e-02  1.991e-02   3.334 0.000888 ***
## Long_Pass           1.503e-02  1.468e-02   1.024 0.306193
## Acceleration        2.851e-02  1.818e-02   1.568 0.117126
## Speed               2.426e-02  1.703e-02   1.424 0.154672
## Stamina            -1.260e-02  1.147e-02  -1.098 0.272512
## Strength            3.346e-02  1.175e-02   2.847 0.004501 **
## Balance            -3.421e-03  1.267e-02  -0.270 0.787224
## Agility             3.438e-03  1.289e-02   0.267 0.789764
## Jumping             1.186e-02  9.076e-03   1.306 0.191781
## Heading             7.448e-02  1.095e-02   6.803 1.73e-11 ***
## Shot_Power          1.218e-02  1.219e-02   0.999 0.317898
## Finishing          -3.958e-03  1.357e-02  -0.292 0.770668
## Long_Shots         -1.966e-02  1.337e-02  -1.471 0.141650
## Curve              -6.818e-03  1.165e-02  -0.585 0.558605
## Freekick_Accuracy   2.469e-02  9.309e-03   2.652 0.008119 **
## Penalties          -1.007e-02  9.987e-03  -1.008 0.313697
## Volleys             8.041e-03  1.135e-02   0.708 0.478997
## GK_Positioning      4.095e-02  2.375e-02   1.724 0.084998 .
## GK_Diving           2.898e-02  2.407e-02   1.204 0.228786
## GK_Kicking          5.316e-02  2.169e-02   2.450 0.014431 *
## GK_Handling         1.034e-01  2.351e-02   4.400 1.19e-05 ***
## GK_Reflexes         6.401e-02  2.378e-02   2.692 0.007218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 1030 degrees of freedom
##   (16513 observations deleted due to missingness)
## Multiple R-squared:  0.8662, Adjusted R-squared:  0.861
## F-statistic: 166.7 on 40 and 1030 DF,  p-value: < 2.2e-16
```

The significant variables did not changed. It suggests that the linear relationship between the significant variables that we stated in part (a) is still strong even though we removed potential outliers and influential points.

(e) Lasso for variable selections…

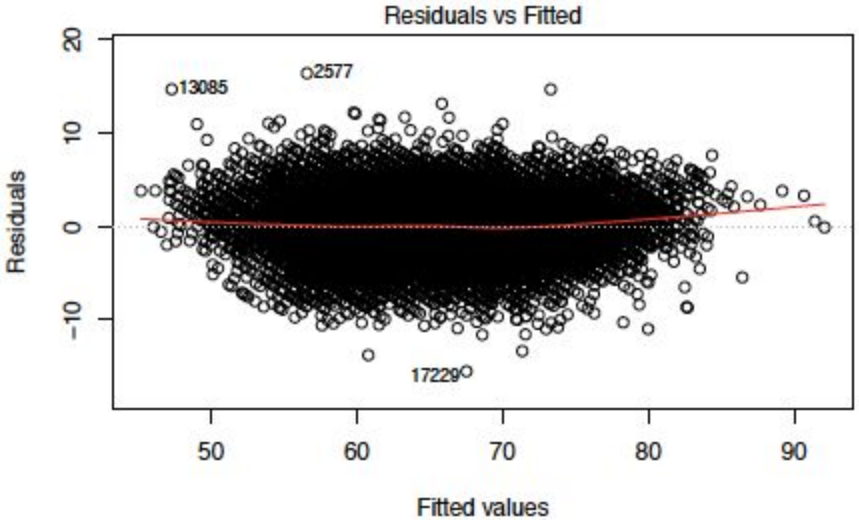
Did it but looks quite insignificant

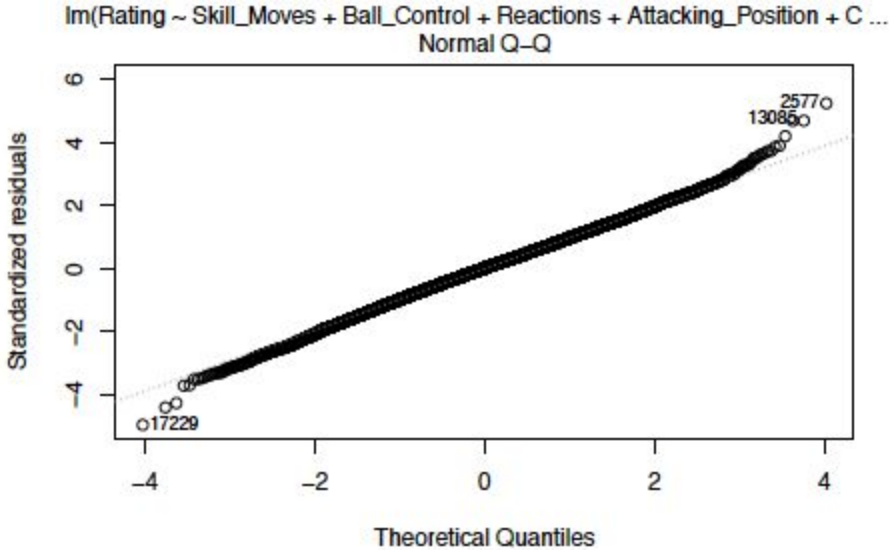
(f) Transform variables using square root and polynomials and find if relationship other than linear is appropriate. Check with residual plot and Component plus residual plot.

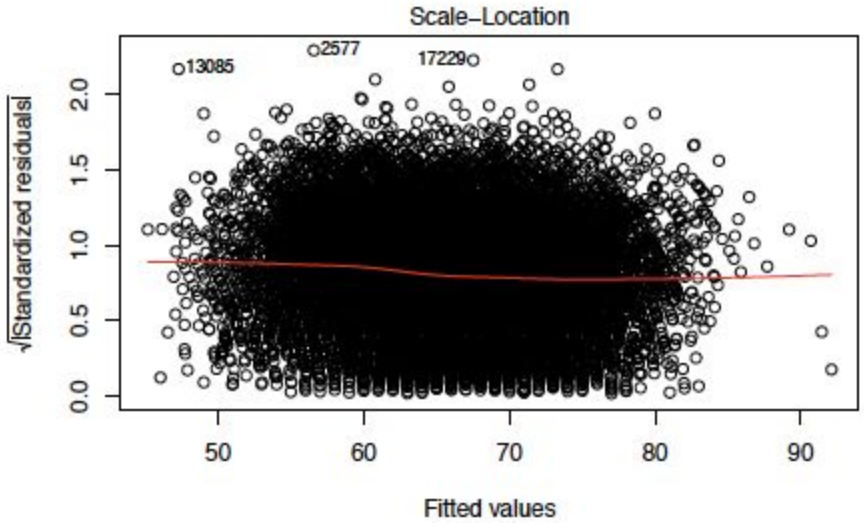
(g) Compare the model that we build with significant variables in part (a) in with the original model → F - statistics

We picked eight most significant variables in part(a) and re-fit the linear regression with only those predictors. The resulted F statistics is -498.6338, and consequently, P value is clearly equal to 1. Therefore, we do not reject the null hypothesis and choose the small model that we made over full model.

```
##
## Call:
## lm(formula = Rating ~ Skill_Moves + Ball_Control + Reactions +
##     Attacking_Position + Composure + Short_Pass + Heading + GK_Handling,
##     data = soccer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5182  -2.0370   0.0571   2.0690  16.4261
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       15.075079   0.211542   71.26   <2e-16 ***
## Skill_Moves        0.690857   0.049748   13.89   <2e-16 ***
## Ball_Control       0.180794   0.004760   37.98   <2e-16 ***
## Reactions          0.363997   0.003818   95.33   <2e-16 ***
## Attacking_Position -0.055912   0.002464  -22.69   <2e-16 ***
## Composure          0.063303   0.002837   22.32   <2e-16 ***
## Short_Pass         0.077611   0.003893   19.94   <2e-16 ***
## Heading            0.125185   0.002344   53.41   <2e-16 ***
## GK_Handling        0.283945   0.003277   86.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.129 on 17579 degrees of freedom
## Multiple R-squared:  0.8049, Adjusted R-squared:  0.8048
## F-statistic:  9064 on 8 and 17579 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
13085
2577
17229
20
10
0
−10
50
60
70
80
90


lm(Rating ~ Skill_Moves + Ball_Control + Reactions + Attacking_Position + C ...
Normal Q–Q
Standardized residuals
Theoretical Quantiles
2577
13085
17229
6
4
2
0
−2
−4
−4
−2
0
2
4



(h) Conclude association…

i) also lm onto each position

We divided club position into 4 types which are midfielder, goalkeeper, attacker, and defense. As our intuition, the significant variables for each club position were various from position to position.

<For midfielder>

```
##
## Call:
## lm(formula = Rating ~ ., data = mid[, -c(1:8, 13, 14, 16, 17)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1045 -1.2298 -0.0474  1.1767 12.3326
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.013e+02  6.756e+01  -2.980 0.002932 **
## Contract_Expiry    1.023e-01  3.326e-02   3.075 0.002147 **
## Height            -7.848e-03  1.647e-02  -0.476 0.633887
## Weight             1.793e-02  1.416e-02   1.266 0.205607
## Age                4.431e-02  1.651e-02   2.684 0.007362 **
## Weak_foot         -8.437e-02  8.264e-02  -1.021 0.307461
## Skill_Moves        1.860e-01  1.079e-01   1.725 0.084842 .
## Ball_Control       2.179e-01  1.782e-02  12.226  < 2e-16 ***
## Dribbling          6.099e-02  1.421e-02   4.294 1.89e-05 ***
```

```
## Marking            -8.766e-04  8.684e-03  -0.101 0.919606
## Sliding_Tackle     -2.637e-03  1.003e-02  -0.263 0.792682
## Standing_Tackle     6.521e-03  1.059e-02   0.616 0.538109
## Aggression          1.392e-02  5.389e-03   2.583 0.009890 **
## Reactions           1.887e-01  1.128e-02  16.734  < 2e-16 ***
## Attacking_Position -1.070e-02  9.514e-03  -1.125 0.260939
## Interceptions      -1.352e-03  7.048e-03  -0.192 0.847888
## Vision              1.974e-02  1.056e-02   1.869 0.061853 .
## Composure           1.378e-02  7.761e-03   1.776 0.075966 .
## Crossing            3.976e-02  7.986e-03   4.978 7.24e-07 ***
## Short_Pass          2.576e-01  1.753e-02  14.698  < 2e-16 ***
## Long_Pass          -5.485e-03  1.230e-02  -0.446 0.655745
## Acceleration        4.423e-02  1.156e-02   3.825 0.000137 ***
## Speed               2.874e-02  1.051e-02   2.735 0.006311 **
## Stamina             4.368e-02  6.812e-03   6.411 1.99e-10 ***
## Strength            6.273e-03  7.760e-03   0.808 0.419009
## Balance            -1.068e-02  9.374e-03  -1.139 0.254723
## Agility            -2.612e-02  9.224e-03  -2.832 0.004698 **
## Jumping             1.693e-02  5.311e-03   3.187 0.001469 **
## Heading             2.485e-02  6.217e-03   3.997 6.77e-05 ***
## Shot_Power          6.747e-03  9.647e-03   0.699 0.484412
## Finishing           1.669e-02  8.367e-03   1.994 0.046304 *
## Long_Shots          6.451e-03  9.116e-03   0.708 0.479329
## Curve              -4.240e-04  7.196e-03  -0.059 0.953021
## Freekick_Accuracy  -2.109e-02  6.294e-03  -3.352 0.000826 ***
## Penalties           1.469e-02  6.866e-03   2.140 0.032543 *
## Volleys            -1.939e-02  6.606e-03  -2.935 0.003396 **
## GK_Positioning     -3.304e-02  1.647e-02  -2.006 0.045018 *
## GK_Diving           1.076e-02  1.632e-02   0.659 0.509872
## GK_Kicking          2.167e-02  1.605e-02   1.350 0.177240
## GK_Handling        -6.242e-03  1.647e-02  -0.379 0.704658
## GK_Reflexes         2.061e-02  1.642e-02   1.255 0.209731
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.878 on 1338 degrees of freedom
## Multiple R-squared:  0.8921, Adjusted R-squared:  0.8889
## F-statistic: 276.6 on 40 and 1338 DF,  p-value: < 2.2e-16
```

<For goalkeeper>

```
##
## Call:
## lm(formula = Rating ~ ., data = goal[, -c(1:8, 13, 14, 16, 17)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88050 -0.26246 -0.04162  0.24796  1.75055
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.341e+01  2.913e+01   0.460  0.64557
## Contract_Expiry    -6.702e-03  1.437e-02  -0.466  0.64122
## Height             -1.958e-04  6.588e-03  -0.030  0.97630
## Weight              3.248e-03  2.771e-03   1.172  0.24198
```

```
## Age                -4.437e-03  5.965e-03  -0.744  0.45746
## Weak_foot          -8.626e-03  3.479e-02  -0.248  0.80435
## Skill_Moves                NA         NA      NA       NA
## Ball_Control        1.348e-03  4.253e-03   0.317  0.75139
## Dribbling           1.980e-03  6.459e-03   0.307  0.75933
## Marking            -1.824e-02  7.782e-03  -2.344  0.01965 *
## Sliding_Tackle     -6.578e-03  5.223e-03  -1.259  0.20884
## Standing_Tackle     3.948e-03  7.364e-03   0.536  0.59227
## Aggression         -4.838e-03  2.785e-03  -1.737  0.08329 .
## Reactions           1.179e-01  4.544e-03  25.943  < 2e-16 ***
## Attacking_Position  1.683e-03  7.738e-03   0.218  0.82792
## Interceptions       2.065e-03  5.337e-03   0.387  0.69905
## Vision              7.441e-04  1.901e-03   0.391  0.69577
## Composure           5.364e-03  1.790e-03   2.997  0.00294 **
## Crossing            1.606e-02  6.971e-03   2.303  0.02189 *
## Short_Pass         -6.398e-03  4.093e-03  -1.563  0.11900
## Long_Pass           1.384e-03  3.749e-03   0.369  0.71220
## Acceleration       -4.850e-04  3.970e-03  -0.122  0.90285
## Speed               1.831e-03  3.861e-03   0.474  0.63568
## Stamina             1.387e-03  3.284e-03   0.422  0.67303
## Strength           -6.157e-04  2.563e-03  -0.240  0.81035
## Balance            -1.739e-05  2.672e-03  -0.007  0.99481
## Agility             4.621e-04  2.419e-03   0.191  0.84859
## Jumping            -5.835e-04  2.822e-03  -0.207  0.83631
## Heading             3.267e-03  6.338e-03   0.515  0.60660
## Shot_Power          1.287e-03  3.611e-03   0.356  0.72182
## Finishing           7.022e-03  8.081e-03   0.869  0.38551
## Long_Shots         -1.801e-04  7.413e-03  -0.024  0.98063
## Curve              -1.104e-02  5.738e-03  -1.924  0.05517 .
## Freekick_Accuracy   9.505e-03  3.622e-03   2.624  0.00908 **
## Penalties           4.160e-03  3.336e-03   1.247  0.21330
## Volleys            -1.405e-02  6.858e-03  -2.048  0.04134 *
## GK_Positioning      2.150e-01  6.681e-03  32.187  < 2e-16 ***
## GK_Diving           2.168e-01  7.125e-03  30.422  < 2e-16 ***
## GK_Kicking          5.231e-02  4.048e-03  12.923  < 2e-16 ***
## GK_Handling         2.129e-01  6.113e-03  34.834  < 2e-16 ***
## GK_Reflexes         1.998e-01  7.068e-03  28.262  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4091 on 328 degrees of freedom
## Multiple R-squared:  0.9965, Adjusted R-squared:  0.996
## F-statistic:  2364 on 39 and 328 DF,  p-value: < 2.2e-16
```

<For attacker >

```
##
## Call:
## lm(formula = Rating ~ ., data = attack[, -c(1:8, 13, 14, 16,
##     17)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6278 -0.7074 -0.0308  0.6839  3.7782
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       34.0080463 67.4588306   0.504  0.61436
## Contract_Expiry   -0.0157547  0.0332438  -0.474  0.63574
## Height             0.0015055  0.0153600   0.098  0.92195
## Weight             0.0062090  0.0102071   0.608  0.54323
## Age               -0.0194596  0.0162270  -1.199  0.23093
## Weak_foot         -0.0098843  0.0744319  -0.133  0.89440
## Skill_Moves       -0.0113886  0.0983477  -0.116  0.90785
## Ball_Control       0.1865387  0.0171619  10.869  < 2e-16 ***
## Dribbling          0.0788754  0.0158421   4.979 8.44e-07 ***
## Marking           -0.0094134  0.0082646  -1.139  0.25517
## Sliding_Tackle     0.0169487  0.0090271   1.878  0.06094 .
## Standing_Tackle    0.0123302  0.0085759   1.438  0.15104
## Aggression        -0.0026639  0.0042581  -0.626  0.53182
## Reactions          0.1081880  0.0111468   9.706  < 2e-16 ***
## Attacking_Position 0.1512162  0.0139185  10.864  < 2e-16 ***
## Interceptions      0.0011275  0.0059177   0.191  0.84895
## Vision             0.0020472  0.0081584   0.251  0.80196
## Composure          0.0158741  0.0073626   2.156  0.03149 *
## Crossing           0.0074954  0.0068631   1.092  0.27523
## Short_Pass         0.0710995  0.0111818   6.358 4.12e-10 ***
## Long_Pass          0.0003652  0.0071895   0.051  0.95950
## Acceleration       0.0360193  0.0113260   3.180  0.00155 **
## Speed              0.0361724  0.0110717   3.267  0.00115 **
## Stamina            0.0071693  0.0063282   1.133  0.25771
## Strength           0.0241791  0.0073408   3.294  0.00105 **
## Balance           -0.0096980  0.0078075  -1.242  0.21469
## Agility            0.0028954  0.0082875   0.349  0.72694
## Jumping           -0.0110688  0.0049559  -2.233  0.02590 *
## Heading            0.0291663  0.0074130   3.935 9.34e-05 ***
## Shot_Power         0.1125516  0.0115427   9.751  < 2e-16 ***
## Finishing          0.1227016  0.0138316   8.871  < 2e-16 ***
## Long_Shots         0.0205525  0.0103572   1.984  0.04768 *
## Curve             -0.0065526  0.0069769  -0.939  0.34803
## Freekick_Accuracy  0.0003279  0.0056311   0.058  0.95358
## Penalties          0.0023008  0.0072071   0.319  0.74966
## Volleys           -0.0038818  0.0087410  -0.444  0.65714
## GK_Positioning     0.0043233  0.0150414   0.287  0.77389
## GK_Diving          0.0119551  0.0151930   0.787  0.43167
## GK_Kicking         0.0111589  0.0152081   0.734  0.46340
## GK_Handling       -0.0357996  0.0155763  -2.298  0.02190 *
## GK_Reflexes       -0.0037424  0.0151008  -0.248  0.80435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.177 on 583 degrees of freedom
## Multiple R-squared:  0.9659, Adjusted R-squared:  0.9636
## F-statistic:   413 on 40 and 583 DF,  p-value: < 2.2e-16
```

<For Defense>

```
##
## Call:
## lm(formula = Rating ~ ., data = defense[, -c(1:8, 13, 14, 16,
```

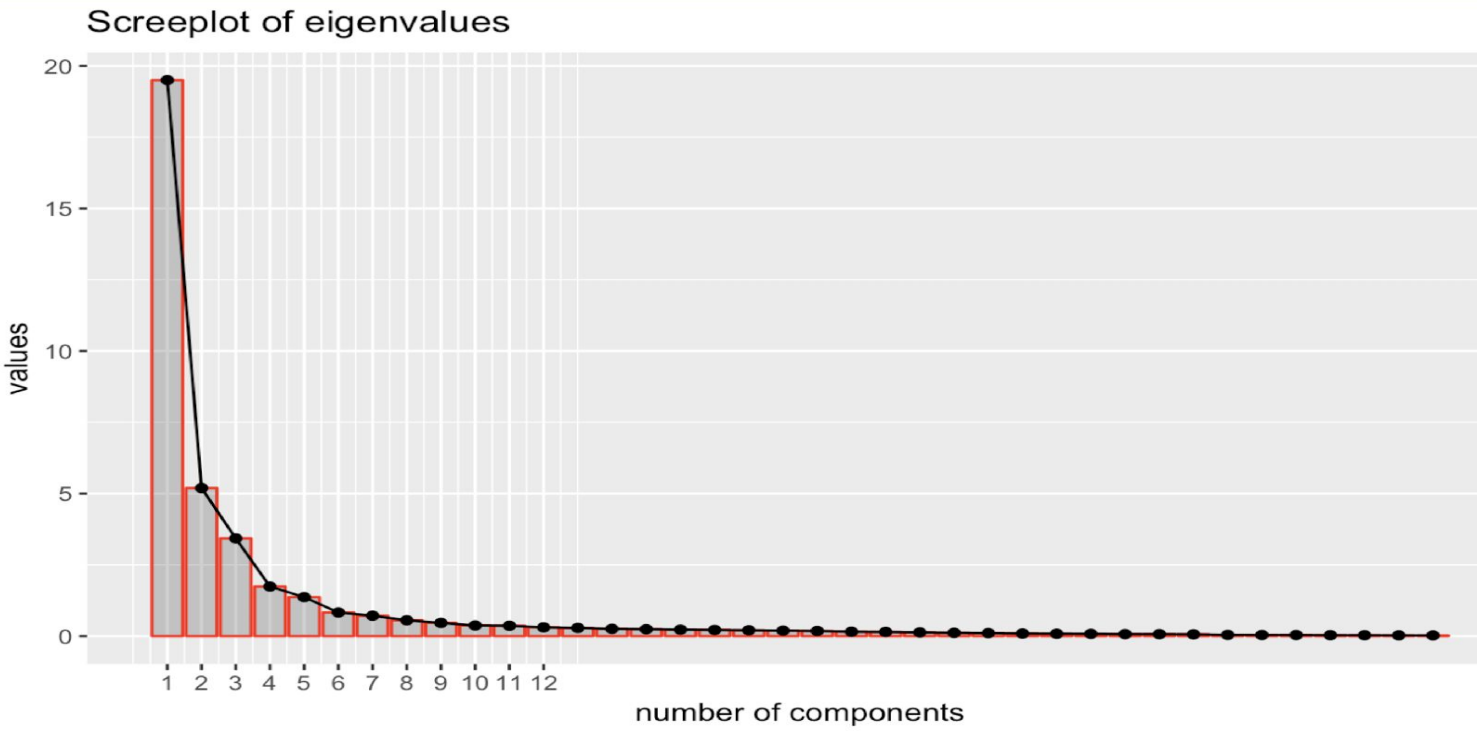
```
##     17)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8546 -0.8569 -0.0594  0.7838 17.9679
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -52.534805  58.671289  -0.895 0.370742
## Contract_Expiry     0.030833   0.028921   1.066 0.286590
## Height             -0.018939   0.015189  -1.247 0.212662
## Weight              0.008360   0.011552   0.724 0.469392
## Age                 0.002216   0.014381   0.154 0.877548
## Weak_foot           0.031655   0.073719   0.429 0.667704
## Skill_Moves         0.695939   0.138324   5.031 5.59e-07 ***
## Ball_Control        0.062709   0.010203   6.146 1.07e-09 ***
## Dribbling           0.004588   0.007075   0.648 0.516787
## Marking             0.065687   0.013180   4.984 7.11e-07 ***
## Sliding_Tackle      0.093025   0.014264   6.522 1.01e-10 ***
## Standing_Tackle     0.168890   0.017014   9.926  < 2e-16 ***
## Aggression          0.037652   0.005624   6.695 3.26e-11 ***
## Reactions           0.123303   0.009403  13.114  < 2e-16 ***
## Attacking_Position -0.004075   0.005486  -0.743 0.457792
## Interceptions       0.097812   0.010925   8.953  < 2e-16 ***
## Vision              0.003119   0.005920   0.527 0.598424
## Composure           0.012489   0.005935   2.104 0.035546 *
## Crossing           -0.020463   0.005275  -3.879 0.000110 ***
## Short_Pass          0.077702   0.010692   7.268 6.45e-13 ***
## Long_Pass          -0.007921   0.007737  -1.024 0.306122
## Acceleration        0.001851   0.008955   0.207 0.836255
## Speed               0.044850   0.008466   5.298 1.39e-07 ***
## Stamina             0.024160   0.006298   3.836 0.000131 ***
## Strength            0.040849   0.007954   5.135 3.26e-07 ***
## Balance            -0.007090   0.007502  -0.945 0.344811
## Agility            -0.008391   0.006703  -1.252 0.210877
## Jumping             0.014556   0.005057   2.878 0.004064 **
## Heading             0.074955   0.007870   9.524  < 2e-16 ***
## Shot_Power          0.014356   0.005101   2.815 0.004960 **
## Finishing           0.014567   0.005834   2.497 0.012663 *
## Long_Shots         -0.010539   0.005652  -1.865 0.062482 .
## Curve               0.009872   0.005111   1.932 0.053634 .
## Freekick_Accuracy  -0.007608   0.004708  -1.616 0.106333
## Penalties           0.001631   0.005196   0.314 0.753633
## Volleys             0.005446   0.005242   1.039 0.299061
## GK_Positioning      0.005312   0.014058   0.378 0.705607
## GK_Diving          -0.012240   0.014161  -0.864 0.387590
## GK_Kicking         -0.006346   0.013602  -0.467 0.640895
## GK_Handling        -0.023382   0.014039  -1.666 0.096064 .
## GK_Reflexes        -0.019395   0.013911  -1.394 0.163490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 1244 degrees of freedom
## Multiple R-squared:  0.9289, Adjusted R-squared:  0.9266

## F-statistic: 406.2 on 40 and 1244 DF,  p-value: < 2.2e-16
```

Summary : As we can see in the summary(goalfit), the significant variables of goalkeeper are quite different than the significant variables of other club position. For example, goalkeeper position does not have strong linear relationship with ball control, short pass, and heading ability but all other three position (midfielder, attacker, and defense) have strong linear relationship with those predictors. This is intuitively true since goalkeeper's most important ability is to block and catch the ball instead of passing and moving the ball to other players or to goal net.

3. PCA:

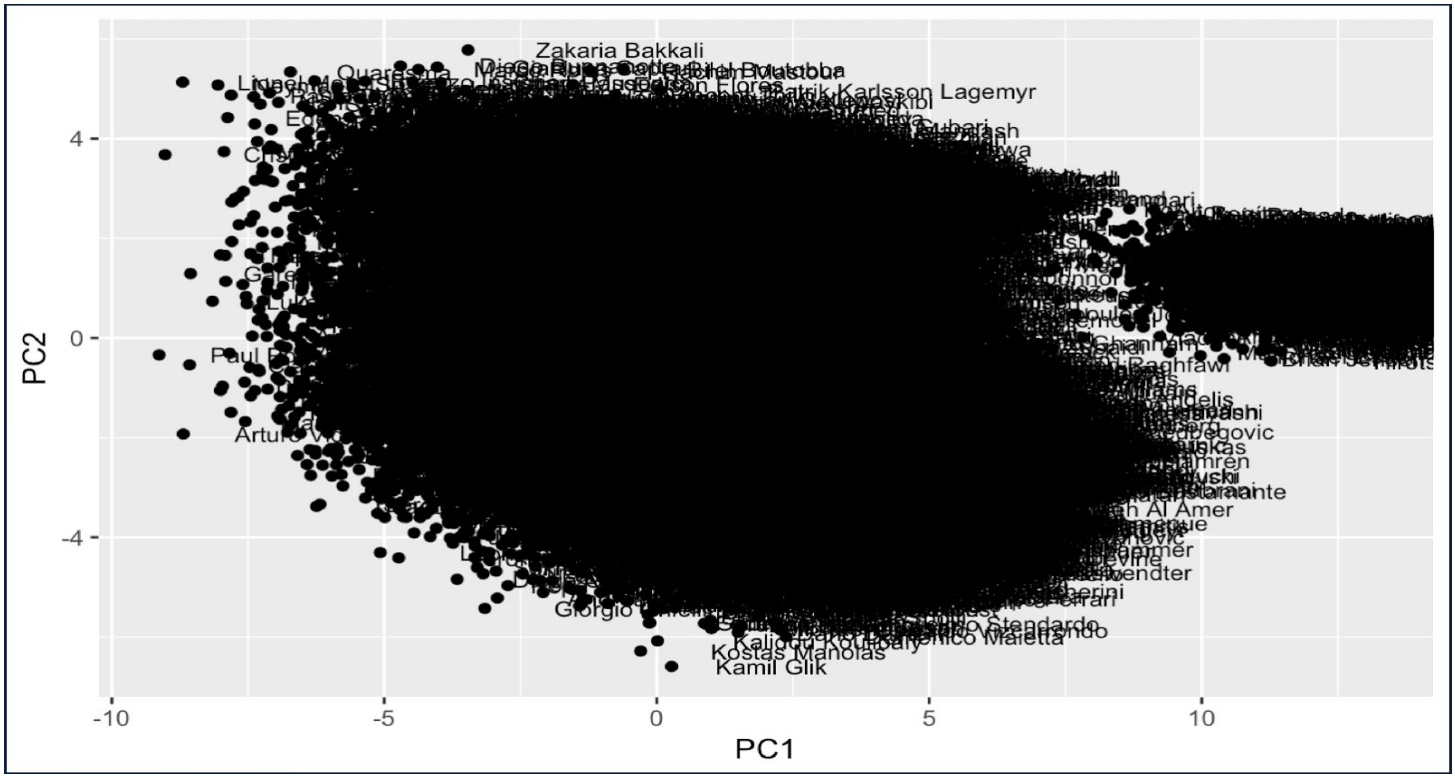
First of all, we performed the PCA with the correlation matrix. And, we used the 38 variables out of 53 variables, as PCA performs on continuous variables. (again, please refer to the data dictionary for more details)

Here is a scree-plot and percentage of the total sample variance explained by the top (in terms of variance explained) 5 components, we got:

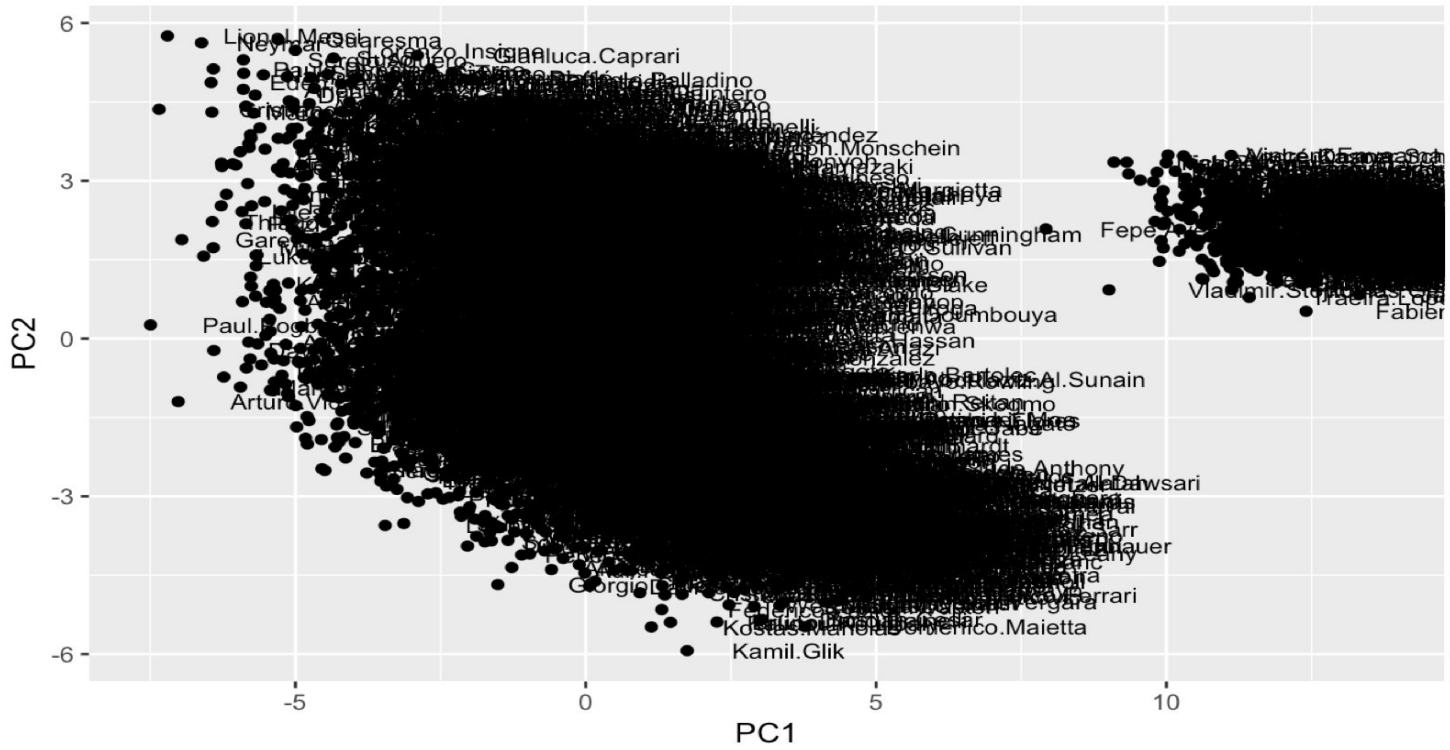


	eigenvalue	percentage	cumulative.percentage
comp1	19.50103815	51.31852145	51.31852
comp2	5.19363027	13.66744809	64.98597
comp3	3.42873946	9.02299859	74.00897
comp4	1.73872644	4.57559589	78.58456
comp5	1.36788830	3.59970607	82.18427

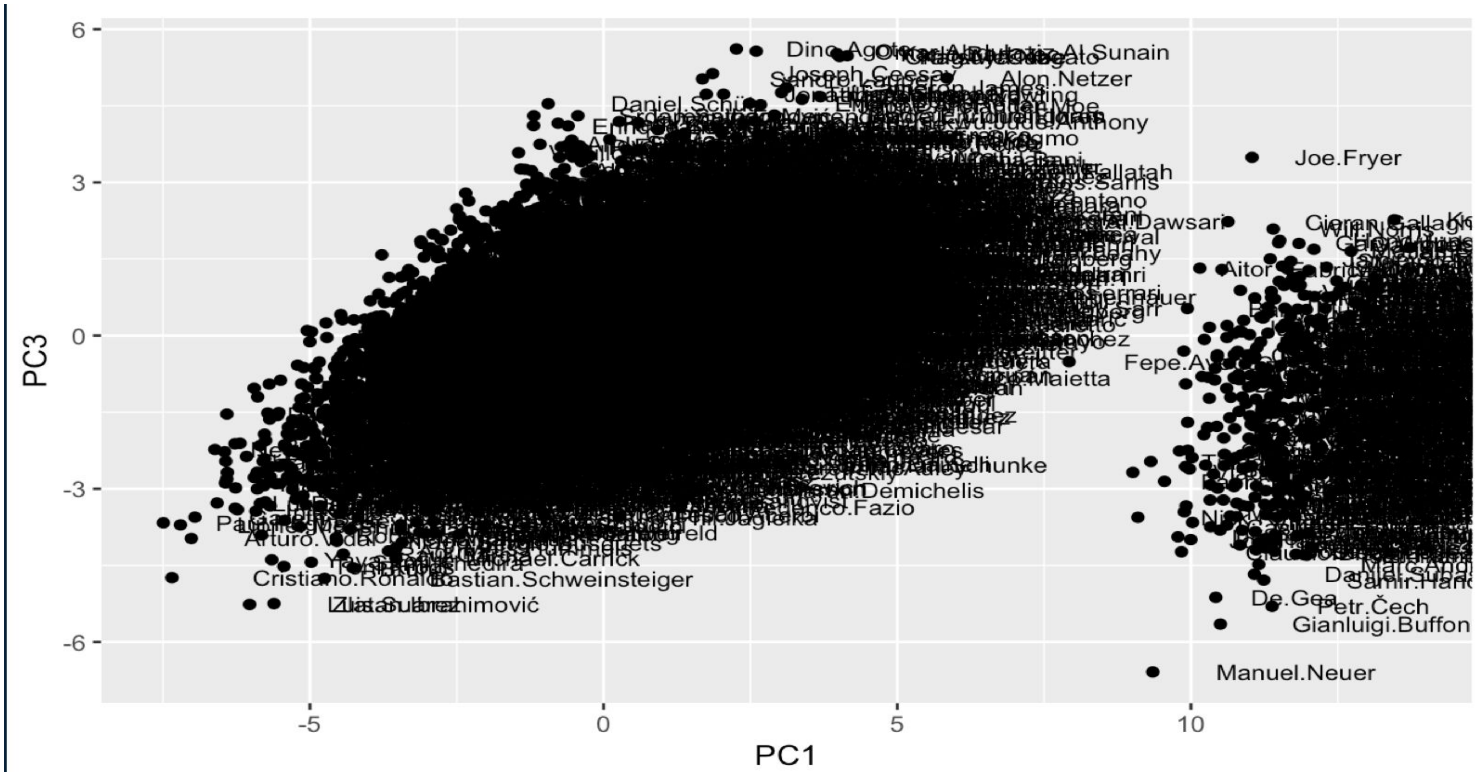
It seems like (based on the elbow on scree-plot and Kaiser or Jolliffe's rules, it is recommended to keep up to around 7 or 8 components, but I will mostly use up to three components for our analysis. The problem of this method is that since we are using a huge data, the correlation unit circle and individuals on the PCA map are hard to be interpreted. (picture below)



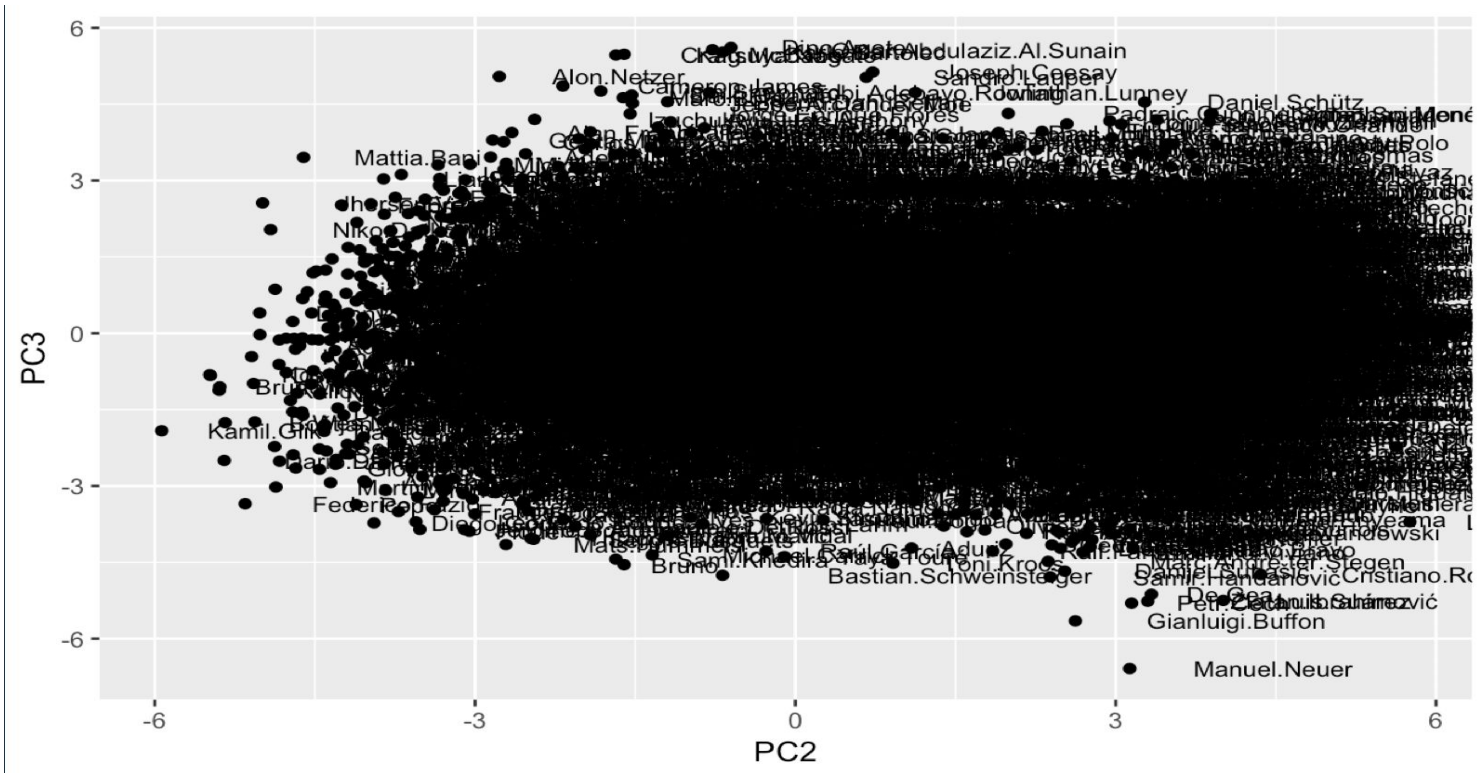
So, we decided to also use the sample we cleaned (3656 players). Now here is a new individual plotted on PC1 v.s. PC2, below:



As you can see, the patterns are both the same, but it is easier for us to do analysis on the second diagram.
Here is a plot for PC1 v.s. PC3, below:

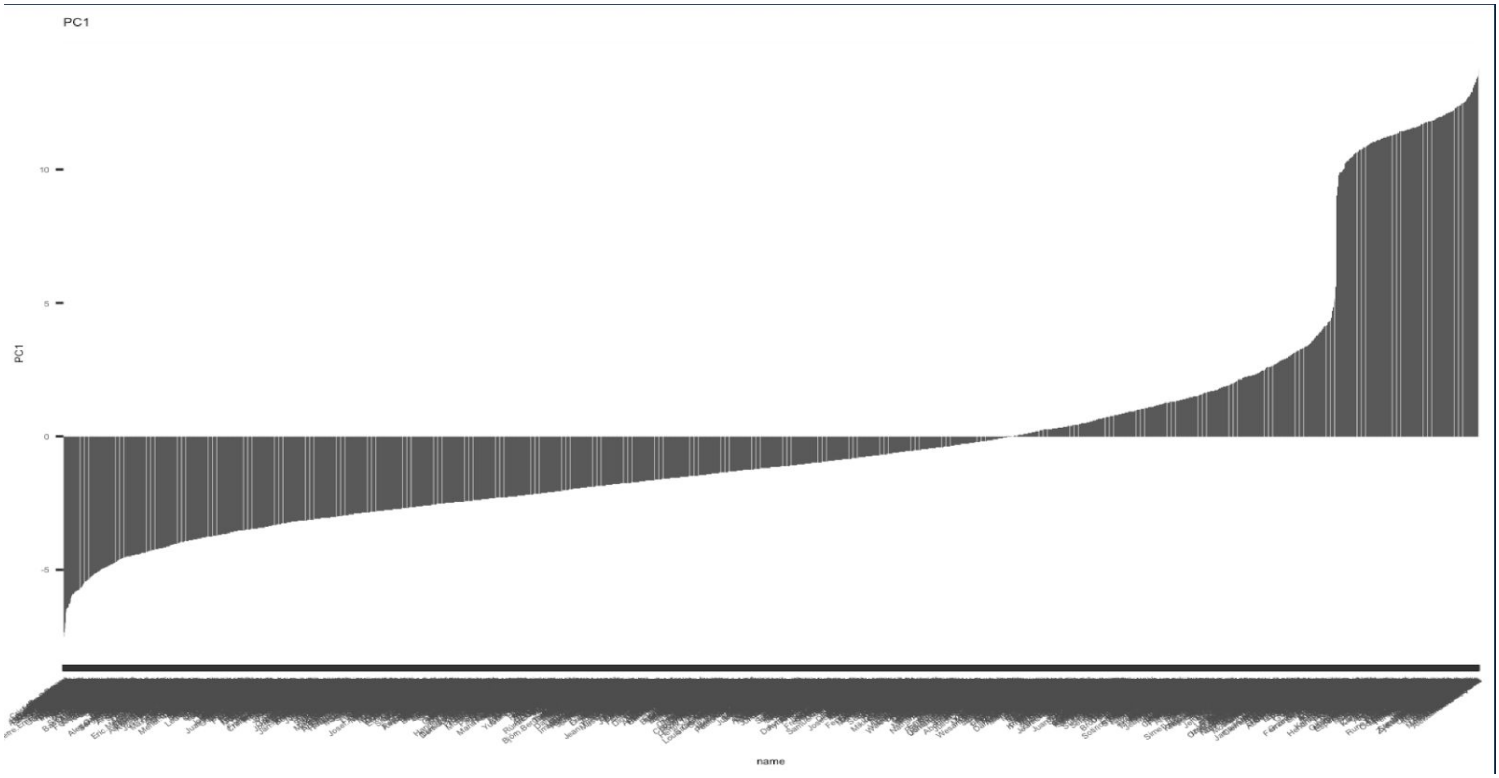


Here is a plot for PC2 v.s. PC3, below:



To do better interpretation, we decided to rank the players based on the scores on the first three PCs. As we have stated above, since both of us strong background knowledges of soccer, we believed rankings could enormously help us a lot.

Here is the rankings for PC1, below (I also included the ranking bar plots. Although there are many players and hard to see each player's name, it is always good to see the plot to see how the patterns of rankings are distributed):

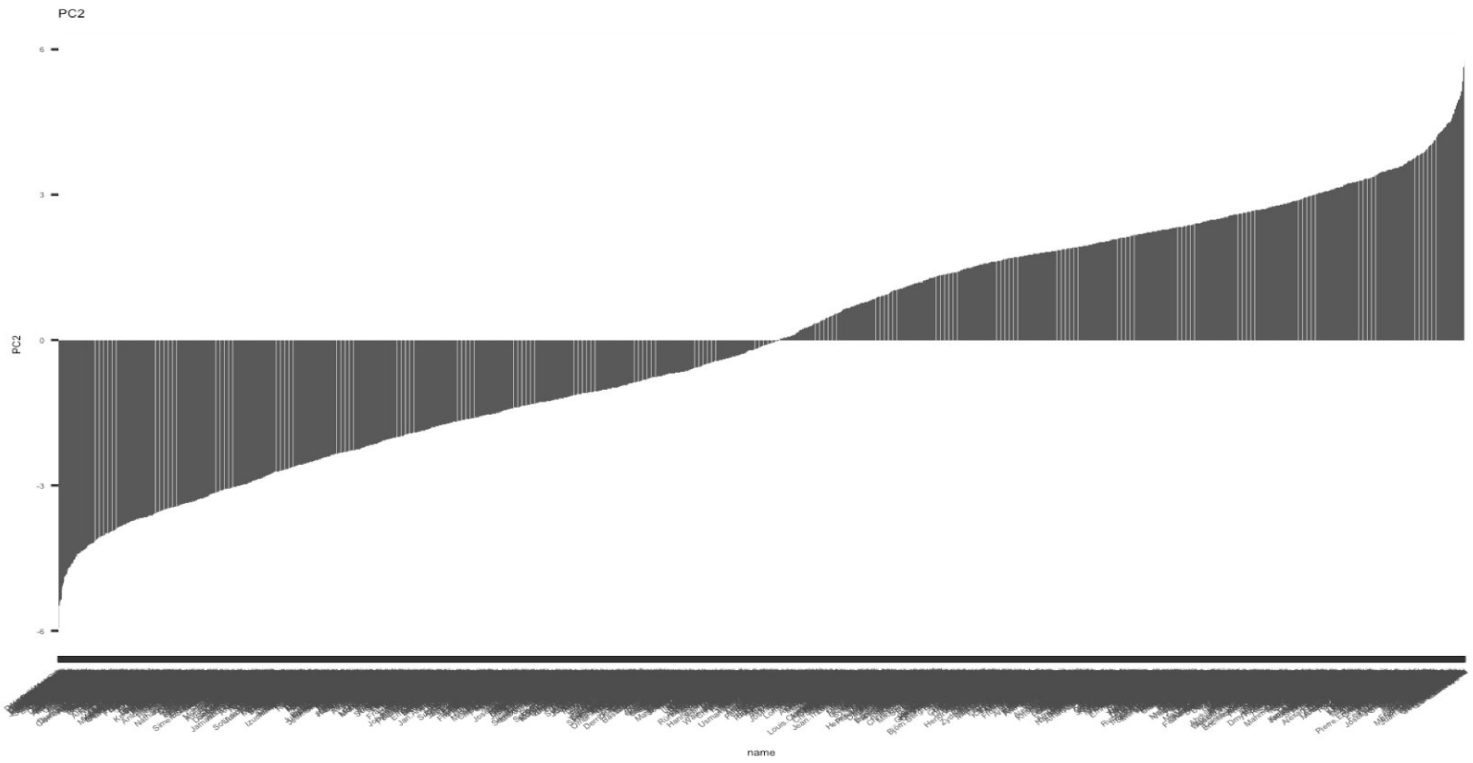


##	PC1	Rank	name
## Paul.Pogba	-7.494345	1	Paul.Pogba
## Cristiano.Ronaldo	-7.348817	2	Cristiano.Ronaldo
## Lionel.Messi	-7.203099	3	Lionel.Messi
## Arturo.Vidal	-7.017833	4	Arturo.Vidal
## Gareth.Bale	-6.957777	5	Gareth.Bale
## Neymar	-6.616190	6	Neymar
## Luka.Modrić	-6.580567	7	Luka.Modrić
## Eden.Hazard	-6.451492	8	Eden.Hazard
## Marco.Reus	-6.437124	9	Marco.Reus
## Thiago	-6.430138	10	Thiago
## Miralem.Pjanić	-6.415132	11	Miralem.Pjanić
## Paulo.Dybala	-6.412915	12	Paulo.Dybala
## David.Alaba	-6.405251	13	David.Alaba
## Henrikh.Mkhitaryan	-6.277171	14	Henrikh.Mkhitaryan
## Kevin.De.Bruyne	-6.272761	15	Kevin.De.Bruyne
## Alexis.Sánchez	-6.265620	16	Alexis.Sánchez
## Radja.Nainggolan	-6.233487	17	Radja.Nainggolan
## Ángel.Di.María	-6.188500	18	Ángel.Di.María
## Alexandre.Lacazette	-6.080171	19	Alexandre.Lacazette
## Zlatan.Ibrahimović	-6.026668	20	Zlatan.Ibrahimović

```
##                              PC1 Rank                     name
## Michal.Pešković          12.86967 3637          Michal.Pešković
## Julian.Pollersbeck       12.89347 3638       Julian.Pollersbeck
## Tomohiko.Murayama        12.90020 3639        Tomohiko.Murayama
## Ole.Söderberg            12.91422 3640            Ole.Söderberg
## Álvaro.Villete           12.97138 3641           Álvaro.Villete
## Masaaki.Higashiguchi     12.98207 3642     Masaaki.Higashiguchi
## Sergiy.Pogorilyi         13.06364 3643         Sergiy.Pogorilyi
## Connor.Ripley            13.13448 3644            Connor.Ripley
## Guillaume.Faivre         13.14898 3645         Guillaume.Faivre
## Tatsuya.Morita           13.16693 3646           Tatsuya.Morita
## Víctor.García            13.16973 3647            Víctor.García
## Josip.Posavec            13.22660 3648            Josip.Posavec
## Roy.Kortsmit             13.25871 3649             Roy.Kortsmit
## Jasmin.Fejzić            13.32119 3650            Jasmin.Fejzić
## Ailton.Cardenhas         13.39412 3651         Ailton.Cardenhas
## Takuto.Hayashi           13.43534 3652           Takuto.Hayashi
## Ko.Shimura               13.46060 3653               Ko.Shimura
## Aly.Keita                13.46305 3654                Aly.Keita
## Ahmet.Şahin              13.57598 3655              Ahmet.Şahin
## Aleksandar.Jovanović     13.72681 3656     Aleksandar.Jovanović
```

We think that PC1 indicates the “star” players. Although not all of them have super high ratings, they were “hot potatoes” in 2016 transfer markets. Please remember that the data is collected at the end of 2016.

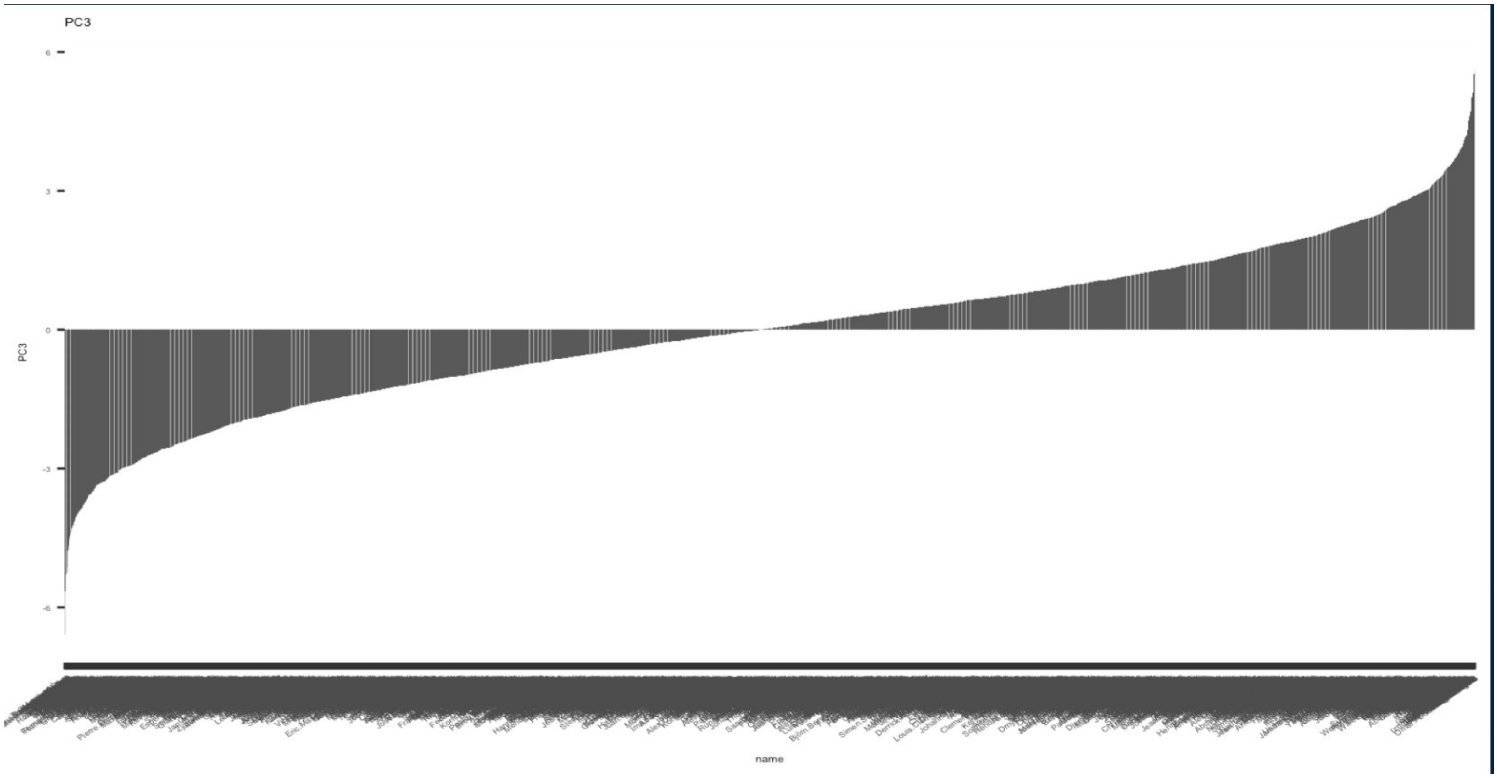
Here is the rankings for PC2, below:



	PC2	Rank	name
Kamil.Glik	-5.934239	1	Kamil.Glik
Kostas.Manolas	-5.484392	2	Kostas.Manolas
Domenico.Maietta	-5.478857	3	Domenico.Maietta
Kalidou.Koulibaly	-5.395775	4	Kalidou.Koulibaly
Bruno.1	-5.390837	5	Bruno.1
Dario.Dainelli	-5.348768	6	Dario.Dainelli
Boštjan.Cesar	-5.340772	7	Boštjan.Cesar
Federico.Fazio	-5.152711	8	Federico.Fazio
Kendall.Waston	-5.092467	9	Kendall.Waston
Michaël.Ciani	-5.074149	10	Michaël.Ciani
Wes.Morgan	-5.062524	11	Wes.Morgan
Danny.Batth	-5.016826	12	Danny.Batth
Magnus.Troest	-5.014742	13	Magnus.Troest
Jherson.Vergara	-4.988994	14	Jherson.Vergara
Niko.Datković	-4.914075	15	Niko.Datković
Sebastián.Coates	-4.876368	16	Sebastián.Coates
Gian.Marco.Ferrari	-4.872075	17	Gian.Marco.Ferrari
Martín.Demichelis	-4.865523	18	Martín.Demichelis
Mikel.Villanueva	-4.835783	19	Mikel.Villanueva
Cristian.Zapata	-4.833008	20	Cristian.Zapata

```
##                              PC2 Rank                 name
## Deulofeu                4.868137 3637              Deulofeu
## Ioannis.Fetfatzidis     4.890564 3638 Ioannis.Fetfatzidis
## Raffaele.Palladino      4.924336 3639  Raffaele.Palladino
## Ousmane.Dembélé         4.959498 3640     Ousmane.Dembélé
## Malcom                  4.963970 3641              Malcom
## Geoffrey.Bia            4.966117 3642        Geoffrey.Bia
## Riyad.Mahrez            4.978207 3643        Riyad.Mahrez
## Nani                    5.015088 3644                Nani
## Marco.Sau               5.022318 3645           Marco.Sau
## Sebastian.Giovinco      5.044004 3646  Sebastian.Giovinco
## Tana                    5.062047 3647                Tana
## Gerso                   5.125732 3648               Gerso
## Paulo.Dybala            5.128308 3649        Paulo.Dybala
## Sergio.Agüero           5.300084 3650       Sergio.Agüero
## Suso                    5.334608 3651                Suso
## Gianluca.Caprari        5.384609 3652    Gianluca.Caprari
## Lorenzo.Insigne         5.480603 3653    Lorenzo.Insigne
## Neymar                  5.624290 3654              Neymar
## Quaresma                5.680732 3655            Quaresma
## Lionel.Messi            5.753872 3656        Lionel.Messi
```

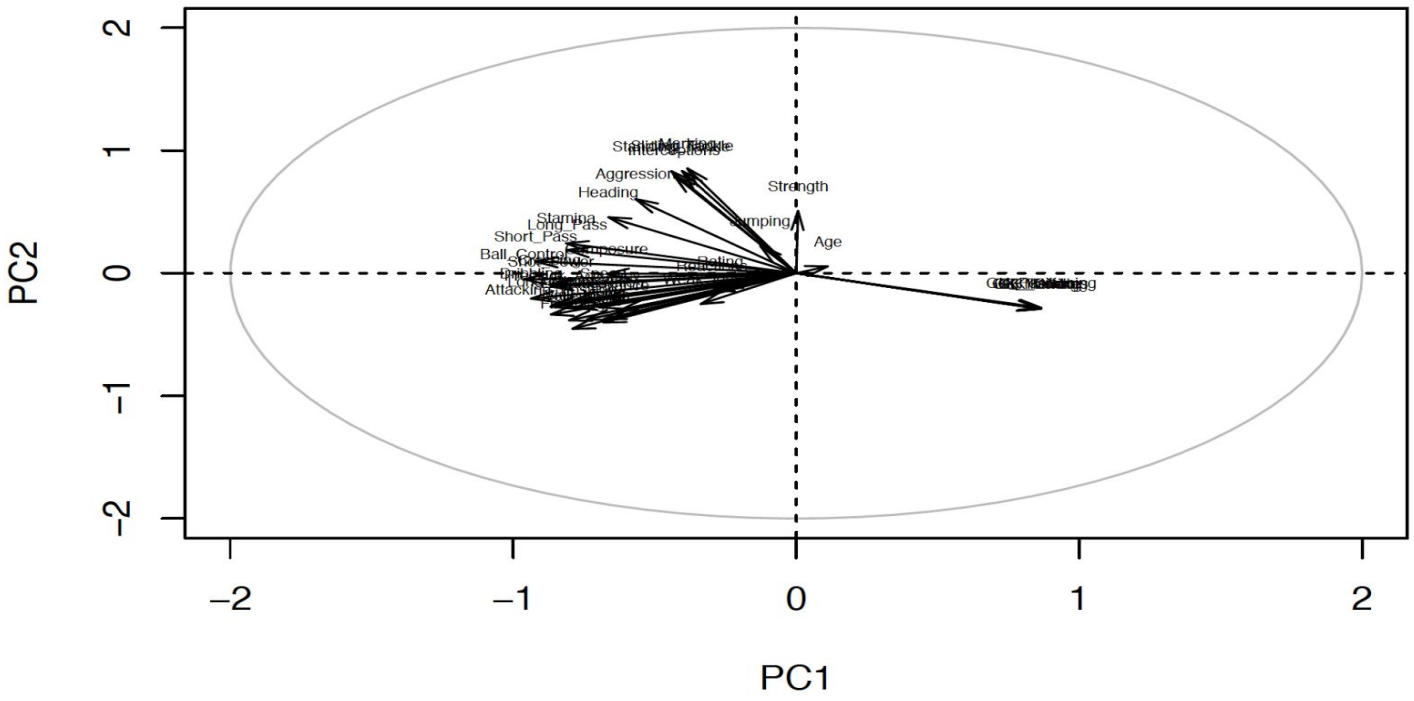
It seems like PC2 indicates how the players are good at dribbling and assisting.
Here is the rankings for PC3, below:

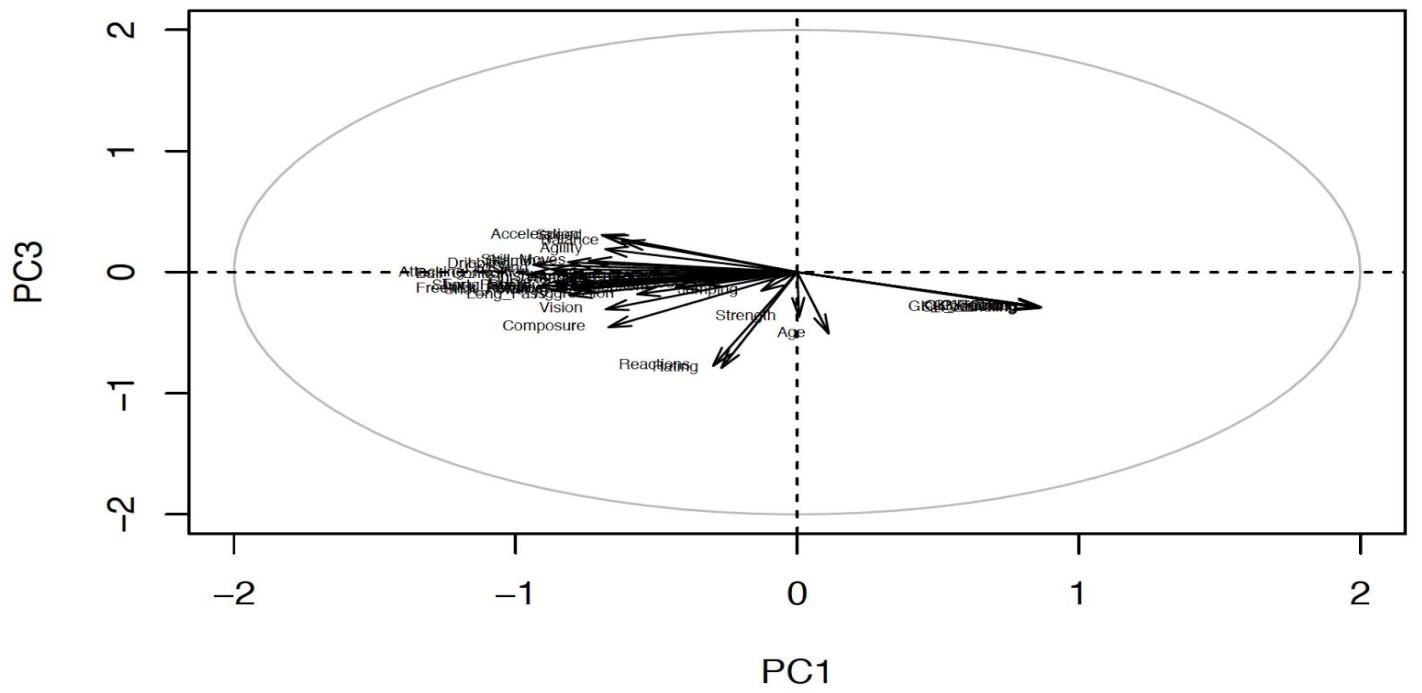


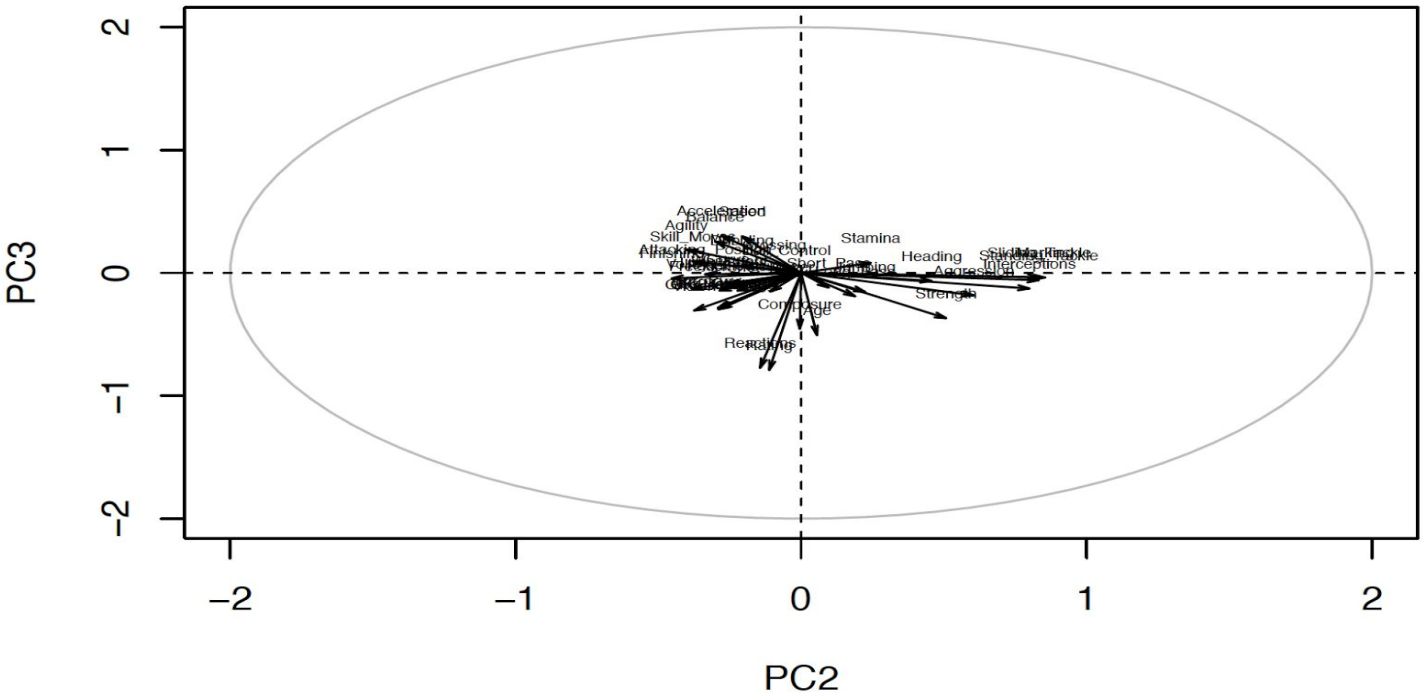
	PC3	Rank	name
Manuel.Neuer	-6.586977	1	Manuel.Neuer
Gianluigi.Buffon	-5.650170	2	Gianluigi.Buffon
Petr.Čech	-5.302168	3	Petr.Čech
Zlatan.Ibrahimović	-5.264714	4	Zlatan.Ibrahimović
Luis.Suárez	-5.247691	5	Luis.Suárez
De.Gea	-5.126426	6	De.Gea
Samir.Handanovič	-4.787075	7	Samir.Handanovič
Bastian.Schweinsteiger	-4.755098	8	Bastian.Schweinsteiger
Cristiano.Ronaldo	-4.740520	9	Cristiano.Ronaldo
Danijel.Subašić	-4.673793	10	Danijel.Subašić
Bruno	-4.546841	11	Bruno
Toni.Kroos	-4.519086	12	Toni.Kroos
Marc.André.ter.Stegen	-4.480644	13	Marc.André.ter.Stegen
Sami.Khedira	-4.437868	14	Sami.Khedira
Yaya.Touré	-4.389973	15	Yaya.Touré
Michael.Carrick	-4.357239	16	Michael.Carrick
Emiliano.Viviano	-4.300733	17	Emiliano.Viviano
Ralf.Fährmann	-4.281651	18	Ralf.Fährmann
Raúl.García	-4.279134	19	Raúl.García
Claudio.Bravo	-4.233416	20	Claudio.Bravo

```
##                                  PC3 Rank                          name
## Jonathan.Menéndez           4.306049 3637          Jonathan.Menéndez
## Jorge.Enrique.Flores        4.310541 3638       Jorge.Enrique.Flores
## Padraic.Cunningham          4.317841 3639         Padraic.Cunningham
## Jeppe.Arctander.Moe         4.519018 3640        Jeppe.Arctander.Moe
## Daniel.Schütz               4.540592 3641              Daniel.Schütz
## Erlend.Dahl.Reitan          4.547530 3642         Erlend.Dahl.Reitan
## Marc.Bola                   4.626368 3643                 Marc.Bola
## Daiki.Ogawa                 4.679683 3644               Daiki.Ogawa
## Tobi.Adebayo.Rowling        4.726900 3645      Tobi.Adebayo.Rowling
## Jonathan.Lunney             4.727113 3646           Jonathan.Lunney
## Liam.Shephard               4.758431 3647           Liam.Shephard
## Cameron.James               4.857167 3648           Cameron.James
## Sandro.Lauper               5.026763 3649           Sandro.Lauper
## Alon.Netzer                 5.043123 3650             Alon.Netzer
## Joseph.Ceesay               5.131417 3651           Joseph.Ceesay
## Craig.McCabe                5.464761 3652            Craig.McCabe
## Katsuya.Nagato              5.479827 3653         Katsuya.Nagato
## Karlo.Bartolec              5.522548 3654         Karlo.Bartolec
## Omar.Abdulaziz.Al.Sunain    5.569618 3655 Omar.Abdulaziz.Al.Sunain
## Dino.Agote                  5.614001 3656               Dino.Agote
```

It seems like PC3 indicates the players who have good stamina (compared to the players in their own positions). Lastly, I want to see how the variables are correlated, so I made the correlation circle between variables and PCs.







Based on what I could see from PCA, definitely, “ratings” are taking popularity into account. However, the positions, ages, and strength itself are not the important facts for ratings. However, all other skills such as jumping, heading, agility, agility, vision, etc are all important factors. On top of that, some of the variables might not be important for some players, but that could be really important on some players’ ratings. For example, the variable “Position” itself does not have huge influence on ratings; however, based on players’ positions, the way how the rankings are assigned to the players are different. Furthermore, our

correlation circles and PCs will be different if we perform extra-analysis on each position. (Remember, we only did PCA into original data)

Last but not least, based on what we found from PCA, there could be multicollinearity issues arose, as many variables are often related each other. For example, the variables "Acceleration" and "Agility" are really correlated.

3. Result and Summary

a) Results (summary of numerical analysis, interpretation, assumption check)

Multivariate tests:

Our conclusion is that ratings are not different based on club positions, preferred foot of the players, and interaction effects of these two.

Linear regression:

Linear regression using all quantitative variables shows us that there are particular variables which are linear related to rating for each players. For overall ratings, the summary of lm fit of full model suggests that variable Skill_Moves, Ball_Control, Reactions, Attacking_Position, Composure, Short_Pass, Heading, and GK_Handling have strong linear relationship with ratings. As our assumption in the beginning of the analysis, we found that FIFA rating is in favor of the players who serve attacking position since variable Attacking_Position much more significant than other predictors.

However, as we shown in the last analysis in linear regression part, we gain fairly different results for significant predictors by club position. Especially the lm fit from the data consists of players who have goalkeeper position shows quite different significant predictors than the other three positions. It suggests that FIFA assess goalkeeper's rate quite differently than the other position.

PCA:

There are many variables correlated each other, and these can be really different based on positions and which PCs we are mapping with. Our PCA is used to support the results of linear regression, how and when each variable are correlated and significant.

b) Conclusion

Our goal was to find out how the FIFA ratings on the players were decided. We were curious whether the rating well indicates the players' stats.

We concluded that the ratings were not evaluated 100% fairly (for example, we found that the ratings put more weights on forwards and star players), but they were good scales to show how good the players were in overall in their positions. We found that it was not a good idea for FIFA to compare players in different positions and rated them together. They might be able to improve the index as to create separate ratings for different positions.

c) How it can be further developed

In our analysis, the variable "Position" is doing an important role. We get the samples based on the players who can be categorized into four big positions, and we sometimes separate the data into these four positions, and work on another analysis, to see how the each position behaves on ratings. Some other groups might be able to go into deeper based on other variables.

We separate data into four big positions, and this is pretty decent categorization. However, this might need further improvement for better analysis. It is because every team has different tactic and philosophy, and player positions can be more complex than what it is. For example, RWB (right wing back) players are required to play as midfielder and even attacker for some teams. One great example you could find if you are interested in: https://en.wikipedia.org/wiki/Total_Football. So, our grouping is arbitrary, and this can be different.

Also, remember that most of the data seems like coming from the players playing in European leagues. This can be further developed, when we can collect all the players' statistics in every continent.

References

https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global

Appendix

Please refer to the link for the R codes:
https://github.com/yjkweon24/public-health-245/blob/master/code/project%20(Jin%20Kweon%2C%20Jiyoong%20Clover%20Jeong).Rmd