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.

Executive summary:

* Our team has believed that FIFA rating has been somewhat biased, and we decided to inspected 17588 players' ratings that we downloaded from Kaggle.
* We decided to whether ratings are different based on positions and preferred foot of the players using Two-way MANOVA, how quantitative variables are related with the rating using linear regression, and find how each variable is correlated each other using PCA. (we believe the order of our analysis should be Two-way MANOVA -> linear regression -> PCA)
* We concluded that FIFA ratings are not 100% fairly evaluated, and it is better for people not use FIFA ratings to compare players with different positions. So, FIFA ratings should be used to compare players in the same positions.

Background:

Two of us have solid prior knowledge of soccer, and we used some of our intuitions when we performed analysis. For example, When we tested out the PCAs, we did not use all the quantitative variables as out testing set, but select some of the variables that we thought would be reasonable to test. (which is more efficient)

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

1. Introduction

a) Problem (the question I want to address)

1. We wanted to know how the ratings (and ages) are different based on club positions (goalkeeper, defense, midfielder, attacker) and preferred foot of the players. -> Two Way MANOVA

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 wanted to check whether the argument was true. Also, we were curious how ratings would be different amongst the players with different preferred foot. (also at the same time, the variable "Age" is really important to soccer players, so we wanted to include this for our quantitative variable. And, we assumed ratings and ages would be closely related.)

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

Aim/objective: Our group used "two-way MANOVA."

2. We wanted to find out what quantitative and qualitative variables would be related with the rating. -> Linear Regression

Reason: It is reasonable to assume that the overall ratings of professional soccer players are proportional to their score (important) variables (these variables can be varied based on player's position, and again, we used our prior knowledge) such as 'Weak foot,' 'Skill Moves,' 'Ball control,' 'GK Reflexes,' etc. We wanted to test how ratings would change when each predictors changed, using linear regression and transforming variables.

Hypothesis: There should exist some quantitative variables that have linear relationships with the rating variable.

Aim/objective: Our group detected the predictors which have strong linear relationship with the variable 'Ratings' and found proper transformation of variables if needed.

3. We tested how each variable would be correlated with the rating. PCA will also help me find co-linearity issue. -> PCA

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

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

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

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

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 sampled 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 their club teams. So, we were focusing on inspecting players' club profiles. Although national related information was not our major variables, we did not take them out, as these information could help us when some of the players did not have enough club information. In this case, we replaced empty club information with national information. For example, many players did not have specific positions ("Sub," "Res," and empty) in their clubs, and we tried to find these missed information from their national positions, if possible. Whenever we did analysis that had to do with positions, we needed to work on the extracted samples of the size 3656, as the others did not show clear positions.

We examined the raw data and found that variable 'National Position' had 16513 missing values and variable 'National Kit' had 16513 NA values. Variable 'Club Position,' 'Club Joining,' 'Contract Expiry,' and 'Club Kit' had one missing/NA value which at 384th observation. (data dictionary: https://github.com/yjkweon24/public-health-245/blob/master/dictionary.csv)

2. Methods

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

1. Multivariate test:

Our subject would be 3656 soccer players: around 20% of our entire set. (but different groups of population) The quantitative variables are ratings and ages (and as we assumed ratings and ages were somewhat closely related, we could say that our measurement was just ratings. To prove my points, we drew the linear model, component plus residual plot, and get correlation, and we found out they were pretty correlated. Again, this was not 100% correlated, but we assumed to be), and there are 2 factors: club positions (4 levels - goalkeeper, defense, midfielder, and attacker) and preferred foot (2 levels - left and right). So, we needed to use Two-way MANOVA. Our team decided to conduct this test to see if there any difference of means from different groups.

Our conclusion was that we reject all of three null hypothesis. (p-values are all pretty small, meaning all are less than 0.05)

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)

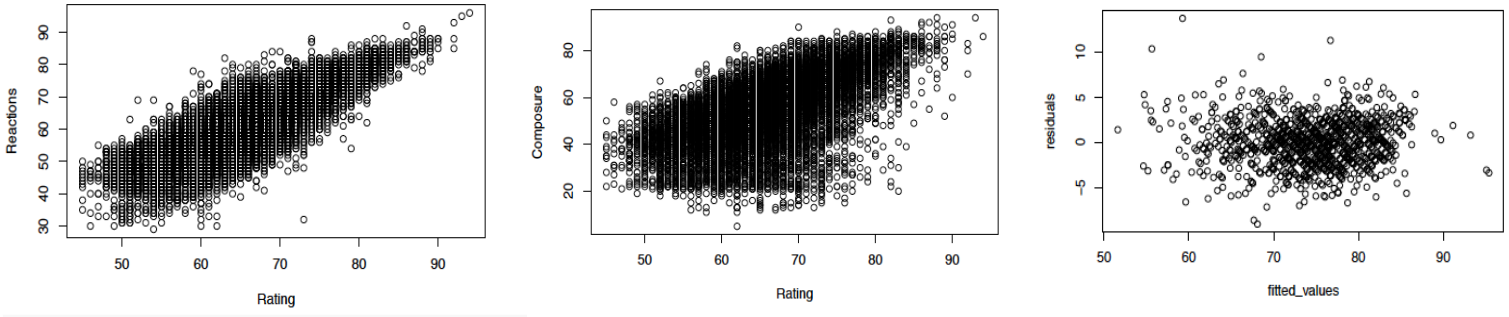
Thus, there were position effect, preferred foot effect, and position-preferred foot interaction effects on ratings (based on the assumption that ratings and ages are correlated well enough), meaning that the means for different groups were different. So, we could say that it would be likely that different positions and different preferred foot made different ratings.

2. Linear regression:

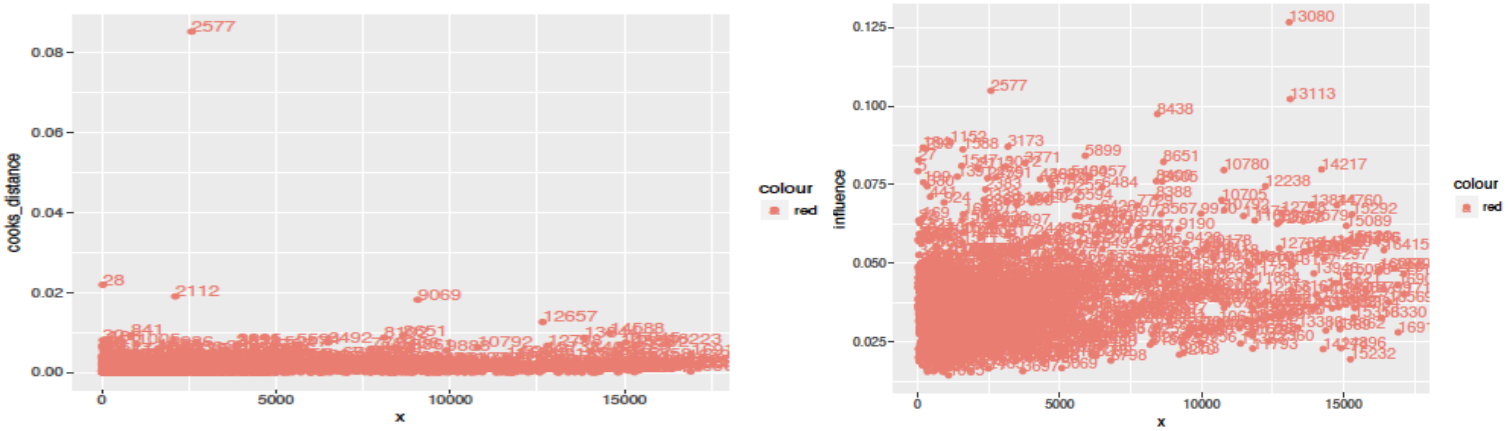
We used linear regression coefficients, T-test, and F-test to find which predictors (quantitative and qualitative variables in factors) contributed significantly in determining ratings, by regressing variable "Rating" on other explanatory variables.

First, we used lm() function and fitted a linear regression model, using rating as the response and other variable as the predictors and check significant predictors. (t-test for significance of each variable)

We converted Height and Weight variable into numeric variable and fit a linear regression model using Ratings as the response and all the other quantitative variables as the predictors. The summary of this fit showed interesting result. We expected that the most of the score variables ('Weak foot,' 'Skill moves,' …. , 'GK Reflexes') would have at least 0.01 significance in the beginning. However, the lm() function results clearly showed that only 10 variables were considered significant in the linear modeling. (t-test) If we set significant level as 0.001, variable 'Skill Moves,' 'Ball Control,' 'Reactions,' 'Attacking Position,' 'Composure,' 'Short Pass,' 'Heading,' and 'GK Handling' would be significant among 41 predictors. The first and the second plot below showed the relationship between variable 'Reaction' and variable 'Composure,' which would be the two most significant variables when we build a linear model with the response variable 'Ratings.'



From the first and the second plots, we could see that these variables had clear linear relationship with the response. (Ratings) We also made a residual plot (the third one from above), 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) The residual plot showed that it did not particularly violate the linear model assumptions since the residuals were symmetric to y = 0 axis and did not show specific patterns. Also, variance seemed constant as well. Next, we plotted of Cook's distance and influential points to detect outliers.



We can see that 2577th, 28th, 2112th, 9069th, and 13080th observations are potential outliers and influential points, so, we removed those points and fitted again. The lm() fit showed that the significant variables did not change, and it suggested that the linear relationship between the significant variables that we stated in the beginning would be still strong even though we removed potential outliers and influential points. In fact, the adjusted R-squared increased.

```
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  3.979e-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
```

Goalkeeper

```
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            0.0291563  0.0074139   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
```

Attacker

After we fitted all the players with different positions at the same time using lm() function, we tried to fit lm() function onto each position of the players since we wanted to know if FIFA rating has been somewhat biased to certain position as we stated in the executive summary. We divided club position into 4 types, which are midfielder, goalkeeper, attacker, and defense. As we expected, the significant variables for each club position were various from position to position.

We could see that in the summary, the significant variables of goalkeeper are quite different than the significant variables of other club positions. (we only compared attacker and goalkeeper summaries here due to the page limits) For example, goalkeeper position did not have strong linear relationship with ball control, short pass, and heading ability but all other three positions (midfielder, attacker, and defense) had strong linear relationships with those predictors. This result intuitively made sense since goalkeeper's most important abilities would be to block and catch the ball instead of passing the ball to other players or to bring the ball to opponent's 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 performed on continuous variables. (again, please refer to the data dictionary for more details)

We decided to use the first top five components - these five components' cumulative percentage is around 82%. (based on the elbow on scree-plot, Kaiser, or Jolliffe's rules, it was recommended to keep up to 7 components, but we decide to mostly use three ~ five components for most of our analysis.

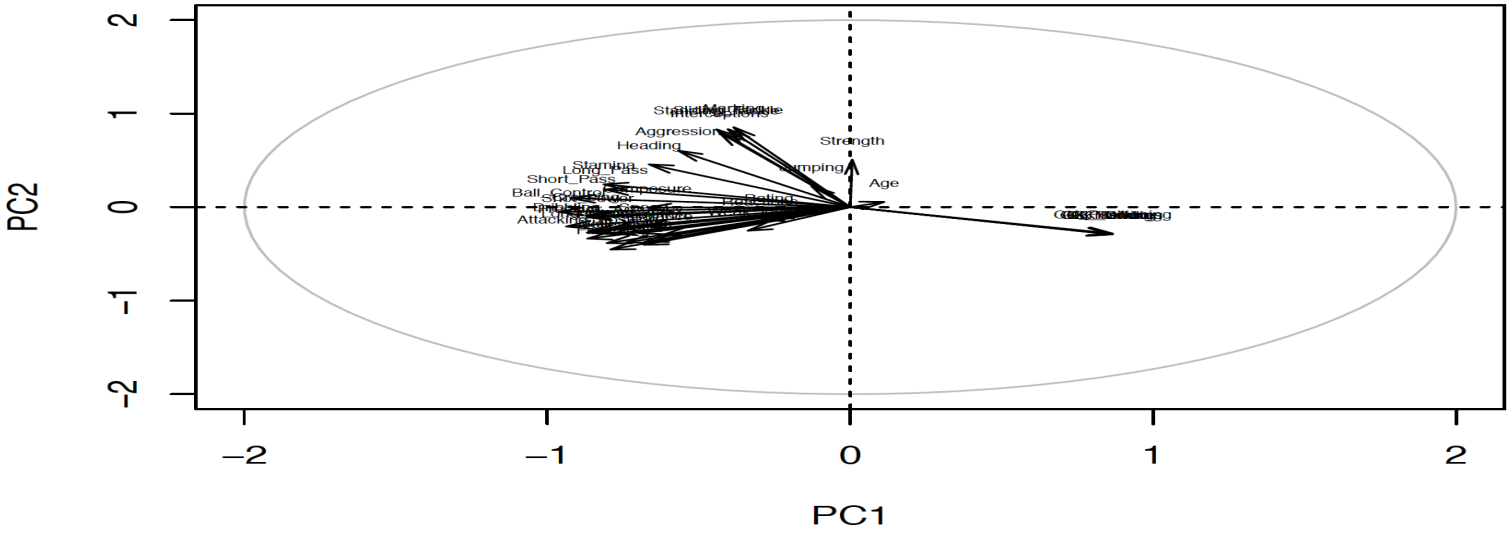
Since we were using a huge data, the correlation unit circle and individuals on the PCA map were hard to be interpreted, so, we also decided to use the sample we cleaned up before. (3656 players) To get better interpretation, we decided to rank the players based on the scores on the first three PCs.

The players placed on the top rankings for PC1 are "Paul Pogba," "Cristiano Ronaldo," "Arturo Vidal," "Gareth Bale," "Neymar," "Luca Modric," "Eden Hazard," "Marco Reus," "Thiago Alcantara," "Paulo Dybala," "Kevin De Bruyne," "Alexis Sanchez," etc. Compared to the players placed on bottom rankings, we could easily realize that PC1 indicated the "star" players. Although not all of them had super high ratings, they were "hot potatoes" in 2016 transfer markets. Please remember that the data we were using for the analysis was collected at the end of 2016.

For PC2, there were "Deulofeu," "Ousmane Dembele," "Riyad Mahrez," "Nani," "Paulo Dybala," "Sergio Aguero," "Neymar," "Quaresma," "Messi," etc on the top rank. By looking at the players' stats, we could see that PC2 might indicate the stats of dribbling and assisting.

PC3 had "Manuel Neuer," "Gianluigi Buffon," "Petr Cech," "De gea," "Samir Handanovic," Claudio Bravo," "Ter Stegen," "Toni Kroos," "Luis Suarez," "Zlatan Ibrahimovic," "Michael Carrick," "Raul Garcia," "Bastian Schweinsteiger," etc on the top ranks. And, it was easy to see that most of them are either good goalkeepers or equipping good stamina (compared to the players in their own positions).

Lastly, we wanted to see how the variables were correlated, so we made the correlation circle between variables and PCs. (we only showed PC 1 v.s. PC2, here due to the page limits)



Based on what we could see from correlation circles (we also inspected PC1 v.s. PC3 and PC2 v.s. PC3), "ratings" were taking popularity into account. However, the positions, ages, and strength itself would not be the important facts to have high ratings. However, all other skills such as jumping, heading, agility, agility, vision, etc were important factors, for most of the players. 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 did not have huge influence on ratings; however, based on players' positions, the

way how the rankings were assigned to the players would definitely be different. (for example, players who were on the top on PC3 had high composure, vision, reading, long passing, strength, age, but they did not necessarily have acceleration and agility, which were important factors to attackers) Furthermore, our correlation circles and PCs would be different if we perform extra-analysis on each position.

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

3. Result and Summary

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

Multivariate tests:

Our conclusion is that the mean of ratings were different based on different positions, preferred foot of the players, and interactions of these two.

Linear regression:

Linear regression using all quantitative variables showed us that there were particular variables which were linear related to rating for each players. For overall ratings, the summary of lm() function's fit of full model suggested that variables: Skill Moves, Ball Control, Reactions, Attacking Position, Composure, Short Pass, Heading, and GK Handling had strong linear relationship with ratings. As our assumption in the beginning of the analysis, we found that FIFA rating was in favor of the players who served in attacking position since the factor Attacking Position (after we dummified) was much more significance than other predictors.

However, as we showed in the last analysis in linear regression part, we gained fairly different results of significant predictors by club positions. Especially the lm() function's fit from the data consists of players who had goalkeeper position showed quite different significant predictors than the other three positions. It suggested that FIFA assessed goalkeeper's rate quite differently than the other positions.

PCA:

There were different variables affecting FIFA ratings, and these could be really different based on positions and the types of players. (the players who are in the same "type" are not necessarily in the same positions, but in similar rankings on some PCs)

b) Conclusion

Since we had spent a lot of time on EDA, all of our three methods could be worked out smoothly without no error message. Each of our method was necessary to come up with our conclusion. (please refer to "Problem" under Introduction)

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 if they were the same type (two players could be regarded as dribblers but one could be midfielder and the other could be attacker) and positions. We found that it was not a good idea for FIFA to compare players in different positions/types and rated them together. They might be able to improve the index by creating separate ratings for different positions and types.

c) How it can be further developed

In our analysis, the variable "Position" was doing an important role. We categorized the samples into four positions, and we sometimes separated the data into these four, and worked on another analysis to see how the each position behaved on ratings. Other researchers working on this data might be able to go into deeper by categorizing other variables.

We separated data into four big positions, and this was pretty decent categorization. However, this might need further improvement for better analysis. We would say it because every team had different tactic and philosophy, and player positions could be more complex than what it was recorded. For example, RWB (right wing back) players were required to play as midfielder and even attacker for some teams. One great example you could find if you would be interested in, below: https://en.wikipedia.org/wiki/Total_Football.

Also, please remember that most of the data seemed like coming from the players playing in European leagues. This could be further developed, if we could collect all the players' statistics in different continents.

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