FIFA Salary Prediction: A Multi-Method Approach

Executive Summary:

The data used for analysis comes from FIFA 2019. FIFA video games are soccer video games where a player can play as various real-life soccer teams. Each player has various attributes in the game that combine to create an overall rating for each player. The rating scale is from 0-100. The dataset contains the individual attribute ratings as well as the overall player rating. It also contains categorical data about each individual player. These categorical variables range from nationality of the player to which foot the player kicks with. Lastly, the dataset contains a value for each player, which is how much a player is worth, and a wage for each player. These last two are the variables we are interested in.

Our overall goal with this dataset is to be able to construct a model that will be able to predict a player's value or wage using the attributes from the game. Essentially, we are looking at the relationship between attributes and how they impact a player's wage. Are there certain attributes that are more important when it comes to wages? Are there some that don't really matter? Ultimately if there is a brand-new player, using their attributes we should be able to give a range for the value and the wage.

In order to analyze the dataset, preprocessing was done to be able to analyze useful information. Preprocessing is essentially cleaning the data in order to make it usable and reduce any problems with analysis. We used multiple methods in order to analyze the dataset. The methods we used were Principal component analysis (PCA), Exploratory factor analysis (FA), Canonical correlation analysis (CCA), Multiple regression, and Correspondence Analysis.

We used PCA to group variables into components that would most greatly affect a player's wage. Since the dataset had many variables, this helped group up different player attributes to help condense the model. Since players have many different rankings in game, grouping them up to help predict their wage was the purpose of CFA as well. Ultimately 4 groups of variables were chosen. They are Attacking, Defending, Physicality, and Speed. Canonical correlation analysis was used to see how the groups of player attributes related to each other. Using CCA, the dataset was split into 5 sets which are: Position, Skill, Balance, and Worth. Multiple regression was used to build a model to predict wage based on the variables in the dataset. Lastly, Correspondence analysis was used in an exploratory fashion. CA was used to see if there were any interesting relationships between a players nationality and the team they play for.

We can see from figures 4 and 9 that the correct number of components according to the scree plots are around 4. From the analysis that we have performed in this paper we can deduce that factors like Age, Passing, Ball Control and 16 others predict a player's wage. These factors can be mainly categorized in 4 groups i.e. Attack, Defense, Physicality and Speed. This model is quite useful for applications when a manager wants to buy a player, he can see the stats and have a rough approximation about his wage. This model can also be used to compare different players while buying or making a loan offer.

There are some few limitations. Firstly, we do not know how a certain player performs in critical times. The data we have in our dataset contains the current or a single year's attribute and it fails to display his overall performance from the start of his career. Because of which experience factor is not considered over here. If two players have the same stats, then the experience factor will strongly determine his quality of play hence determining the wage of the player. More research will have to be done in order to include a time variable. Other literature analyzing soccer player wage includes a previous season component, which our dataset lacks

Secondly, this database does not have any data telling how a certain player is as a team player. Even if a certain player might have good stats, but if he is unable to perform as a team player then his wage is surely going to affect. Playing as a team may consist of data like passes, assists, keeping team morale high etc. This is more subjective and harder to measure. Sometimes a team could consist of superstars and not perform well due to locker room issues and things of that nature.

From our research we can conclude that a player's wage is hugely impacted by 4 groups which we got by using different types of analysis. The 4 groups are named as Attack, Defense, Physicality and Speed. These 4 groups are resulted from the PCA and CFA. Attack is the largest component that explains a player's wage. Attributes like accuracy, finishing and passing ability go into this component. Most of the variables in the attacking component can also correlate with goalscoring. Future research might need to be done to uncover such an effect. Moreover, from the CA we can say that a player's wage might affect a player's nationalities. If a certain player has good stats but is from a country where football is not that popular (For ex. China, Uzbekistan, Hong Kong etc.) then his wage might be affected. While the CA was ultimately just for exploratory purposes, this is also an interesting find.

Technical Summary:

Abstract

We focus on exploring the factors that determine the salary of the FIFA player. We tried to predict the salary of the football player. For the analysis we used FIFA-19 data containing the 86 variables and 18,900 observations in other words 18,900 football players records. Firstly, multiple regression is applied on the dataset to explain the wage and value of the player. To add to it Principal component analysis(PCA) and Exploratory factor analysis(FA) techniques are applied on the data for further analysis and to find similar kinds of skills/positions and transform them into groups such as to group same kind of positions or same kind of abilities like attack or defense etc. Canonical correlation analysis (CCA) is also applied to the dataset to find inter correlations present between different positions or the abilities or certain types of skills of a player. Finally, as the data also possesses non-numerical type variables Correspondence analysis (CA) is also applied to achieve relationships among the nationality, clubs with salary of the player. Applying the multiple regression almost with 75% accuracy is new Football player salary can be predicted accurately. To add to it using other techniques suggested above interesting groups are achieved among various types of variables.

Introduction/Literature Review:

In the 20th century with the boom of digital media it also influenced sports and FIFA is one of them. In 2018 total revenue generated by FIFA is over $4.6 Billion [FIFA Financial Report 2018]. With this big sport also professional players earn in millions and one of the biggest challenges faced by clubs is to determine the salary of the Professional player based on player's performance, skills, and previous seasons and many more. A few decades ago, football data was not collected continuously, and the player's salary was decided based on qualitative analysis (Frick, B. 2006). But in the 1990s the football committee started collecting data of each player. One effort has been investigated by a player has an insignificant effect on the salary (Wicker et al., 2013), this study says the determination of the player's salary depends on various factors such as games played internationally, type of performance in previous seasons, and in attracting side how many goals scored by the player (Frick, 2011). As this factor affects players' salary, we can also consider that skills like accuracy, passing skills etc. are also reasonable factors to contribute to a player's salary.

In this study, we will use the FIFA-19 dataset taken from Kaggle (A Public data repository) and originally taken from sofifa.com consist of skills and positions and other and wage and value of players. The wage provided in the dataset is the gross salary of each player which did not include other earnings of the player such as advertising or sponsorship of brands.Since the data is multidimensional we will use multiple regression as well as conditionality reduction techniques such as Principal component analysis and Exploratory factor analysis also Canonical correlation analysis being used to check cross variable variance. The data also possesses categorical variables such as Nationality, clubs, etc. due to that reason correspondence analysis is also being used to extract some useful information which will help to explain how the prediction of the player's salary can be made efficient.

Previous literature finds that there are certain factors that contribute to a soccer players salary more than others. Lucifora and Simmons find that there is a “superstar effect” (2003). They use Italian league soccer data. The main technique they use is a multiple regression where the dependent variable is the natural log of wages. They essentially find the more people are willing to see the player play live, and the number of spectators significant to the players earnings. These two factors contribute to the superstar effect defined in the paper. Other factors include performance and reputation. They control for team influences using a team fixed effect. Battre et al. (2008) find that region of birth and leadership skills influence salaries using German professional soccer data. While those two variables are the strongest, they find that other data, like goals scored and position also influence soccer player wage. Battre et al. also find what seems to be a “superstar effect” in their data. In 2017 a study ”Computational Estimation of Football Player Wages” (Yaldo, L., & Shamir, L. (2017) aimed to predict the football player salary to help football clubs in contract negotiation study found some more interesting facts which affect the salary of a player. The study found that some players are overpaid who lack in some skills because of their physicality, on contrary some players are underpaid who possess those important skills and are not physically attractive. To add to it this study also finds that just abilities cannot contribute in predicting the salary other factors also matter like fan following.

Ultimately, our goal is to be able to predict player wage based on our FIFA dataset. Summarizing previous research indicates that player wage is not something that is completely subjective. This means that there are some components that could actively be affecting player wages within professional soccer.

Methods

Model Building and Multiple Regression

Raw data was preprocessed by deleting the unwanted rows and performing log-transformation. Also multicollinearity was checked by using VIF to get the final model for this dataset. Also stepwise regression was performed to get the best fit model to this dataset.

Principal Component Analysis

Principal component analysis (PCA) was conducted using the prcomp function, and principal function from the Psych package in RStudio. For the latter function, varimax rotation was used. The correct number of components were determined by considering both the eigenvalue and scree methods.

Confirmatory Factor Analysis

To get the deep knowledge of the FIFA-19 data confirmatory factor analysis (CFA) was conducted using the psych package using the factanal function applied using R in RStudio. The number of factors were determined using eigenvalue and the scree plot methods.

Canonical Correlation Analysis

Canonical correlation analysis is used to identify the relations among two sets of variables. The purpose of this analysis is data interpretation by finding the canonical variates that are more significant in terms of explaining the relationship between the two datasets. This analysis is an overview of hoe FIFIA players .

Correspondence Analysis

Correspondence analysis is an extension of PCA and usually done on categorical data. Since the research question is prediction of wages/value, the CA was done in an exploratory fashion.

Model Building and Multiple Regression

Looking at Figure 1, we can see that the model is passing the global f-test. Also, looking at individual p-values of each variables we can see that 14 variables are considered significant for this model i.e., Age, Value, International-Reputation, Skill Moves, Crossing, Heading Accuracy, Dribbling, FKAccuracy, Long Passing, Ball Control, Shot Power, Stamina, Penalties and Sliding Tackle.

As R-square value is 75.68%, we can say that this model is a good fit for our dataset. Looking at the multiple R-squared, we can say that 75.68% of the variability in wage is explained by the model. These metrics are close in value, which suggests that overfitting is not an issue here.

For checking multicollinearity, We firstly checked the beta values and their values were good (not in millions or billions). Also, I used VIF from CAR Package and saw that the values of 5 variables were having a VIF value more than 10. Below table shows the VIF values of the final model after deleting the variables having huge VIF values.

As you can see in Figure 2, all values are below 10 and there are 14 variables which can be considered for the final model. 5 variables were having VIF values greater than 10 and were deleted from the model.

The final model will contain 14 variables named as Age, Value, International-Reputation, Skill Moves, Crossing, Heading Accuracy, Dribbling, FKAccuracy, Long Passing, Ball Control, Shot Power, Stamina, Penalties and Sliding Tackle.

The final model is shown in Figure 3. We can see that, out of 14 variables, 5 variables i.eCrossing, HeadingAccuracy, Dribbling, ShotPower, Penalties have high beta-coefficient values.These skills are strongly related to an attacking player which implies that attacking players have a higher wage/value as compared to the defending and midfielding players.

Principal Component Analysis

The original dataset contained 84 variables and 18,900 observations. After cleaning and preprocessing, the resulting dataset contained 58 variables and 14,744 observations. The dependent variables related to player salary were removed from the dataset as well, leaving 56 total variables. Before starting the PCA process, the following techniques were used to verify assumptions about the dataset: Chronach’s Alpha, Kaiser-Meyer-Olkin (KMO), and Bartlett’s Test of Sphericity.

Cronbach's Alpha assesses the consistency of each summated scale in the dataset. A value of 0.81 was recorded, suggesting good reliability for the dataset. Next, KMO was checked, recording a value of 0.72, indicating a relatively strong relationship and durable sample size. Lastly, Bartlett’s Sphericity Test was checked to evaluate shared variance. A value of $p < 0.001$ was recorded. The null hypothesis that there are no correlations present in the dataset was rejected, accepting the alternative hypothesis that there are correlations, shared variance between the variables.

In order to choose the number of components for analysis to best represent the variance in the dataset, both the eigenvalue and scree methods were considered. As seen in Table 1, the eigenvalue method suggested that 9 components be used. The scree method (Figure 4), in contrast, suggested 5 or fewer components.

To test the components suggested by the eigenvalue method, 6-9 components were used, using varimax rotation and a final cutoff value of 0.6, in four separate analyses. The resulting components were problematic, as they contained significant cross-loadings (when first checked with lower cut-off values), and were not readily interpretable. This indicated that fewer components should be used.

Further, the 5 or fewer components suggested by the scree method were tested. More specifically, 3-5 components were tested, using varimax rotation and a final cutoff value of 0.6. All of the analyses had some issues with cross loadings when lower cut-off values were used. Using 5 components, the fifth component was questionable, as it did not provide much useful information, and contained two contradictory variables with opposite signs. Using 3 components, the third component was questionable from an interpretability standpoint, as the variables did not give a clear picture of what the component meant in terms of application. Lastly, analysis was conducted and evaluated using 4 components. The main difference here, compared to the analyses with 3 and 5 components, was that all of the components were easily interpretable. Components 3 and 4, respectively, contained variables that made sense together, in terms of the application. Thus, 4 components were determined to be ideal for analysis.

After configuring the PCA model with 4 components, the results are shown in Table 2. Using 4 components, 63.3% of the variance was accounted for overall. Components 1 and 2, respectively, made up 33.7% and 17.5%, and together explained 51.3% of the variance. Components 3 and 4 added 6.6% and 5.5%, respectively. While the latter two components were much lower than the previous two, they were still well above 1, and added additional information about the variance of the independent variables in the dataset that was relevant to the application.

After consulting the loadings (Figure 5), Component 1 was identified as being strongly associated with positions and skill sets that are related to offensive parts of the game: positions like striker and forward, and skills like finishing and shotpower. Component 1, therefore, was labelled “attack.” Component 2 was identified as being strongly associated with positions and skill sets that are related to defensive parts of the game: positions like back and sweeper, and skills like marking and tackling. Component 3, with correlations such as strength and weight, was labelled “physicality.” Lastly, component 4 was correlated with acceleration and sprint speed, and was labelled “speed.”

Confirmatory Factor Analysis

Originally data consists of 84 variables and 18,900 observations. After preprocessing and removing some unwanted variables and cleaning data. The data remained for further processing consisting of 58 variables and 14,744 observations.

Prior to applying the factor analysis all the assumptions are verified using various tests. First test applied on the selected dataset is the **Cronbach's alpha** it tests the internal consistency of the data; in other words, it tests how closely they related the set of variables to each other. Cronbach's alpha came out to be 0.81 which is between acceptable and the good on the scale. Secondly another test applied is **Kaiser-Meyer-Olkin (KMO)** Test for verifying that the selected data is suitable for factor analysis the minimum should be 0.6. KMO came out to be 0.72 which is a positive sign for applying factor analysis. Third and last test to assess the assumption is **Bartlett's Test of Sphericity** which verifies whether the variables possess correlations or not. Selected data for factor analysis give $p<0.001$ which means it rejects the null hypothesis (no correlations present among the variables) and accepts the alternative hypothesis (correlation is present among the variables).

In order to perform factor analysis on data to get some interesting results the first step is to get how many factors to be utilized for that two methods were applied on the data were eigenvalue and the scree plot. Applying the eigenvalues chart (Figure 6) gives a total of 9 factors which has eigenvalue greater than 1. Scree plot (Figure 4) gives the similar kind of result but with fewer factors of 7. For further analysis it is needed to try and check what factors are perfect for the further analysis. Firstly, using the varimax rotation total 8 factors were rotated with the cutoff 0.40. Unfortunately, due to lower cutoff and more factors a huge cross loading is founded with 8 factors. In the second iteration 7 factors were tested with same rotation and cutoff 0.4 but same issue followed of cross loadings to improve issue of cross loadings cutoff was raised to 0.5 but it didn’t help.

For the final iteration 4 factors found to be the best fit for the factor analysis but for 4 factors also cutoff is to be raised to 0.55 in order to remove cross loadings issue. To add to it factor rotation with 3 factors were also applied to check if it reduces the cross loadings with the lower cutoff but the 4 factors with cutoff of 0.55 gives the best result.

Total of 56 variables were applied to the exploratory factor analysis, the dependent variables were removed from the factor analysis to get a better interpretation. Out of total 56 variables 26 variables are contributing to the first factor 9 factors are contributing to the second factor 3 variables are contributing to the third factor and for the fourth factor just 2 variables are contributing to it. To add to it from the 56 total variables 40 of them are contributing to 4 factors with 0.55 cutoff and remaining 16 are not contributing to any one of the factors.

The table (Table 3) gives information of Sum of squared loadings Proportion variance and the cumulative variance. As all the factors sum of squares loadings are greater than 1 we will keep all the factors for the analysis. Proportion of variance is the contribution of the variance by each factor individually and the 1st factor contributes a total of 32.1% variance, 2nd factor contributes a total of 17% variance, 3rd factor contributes a total of 6.7% variance and 4th factor contributes a total of 5% variance. Finally, the cumulative variance is the total variance contributed by all the factors so in total by considering 4 factors a total of 60.8% variance is explained using 4 factors.

Rotating the factors gives better interpretation, so from the analysis of the variables present in 1st factor it can be said the 1st factor is named as **ATTACK**. Similarly, 2nd factors possess a defensive type of variables, and so it is named as **DEFENCE**. The 3rd factors consist of physical strength type of variables and for that reason it is named as **PHYSICALITY** and the 4th factor is named as **SPEED** as it consists of acceleration and speed type of variables.

Canonical Correspondence Analysis

Canonical correlation analysis is used to identify the relations among two sets of variables. The purpose of this analysis is data interpretation by finding the canonical variates that are more significant in terms of explaining the relationship between the two datasets. The purpose of this analysis is to figure out the number of Canonical variables that are highly correlated to players' income and value, and how they can help us to predict the upcoming player's wage or value.

In this section we are going to use FIFA-19 data containing the 86 variables and 18,900 observations. The data was cleaned to have 58 variables which were separated into 5 sets.·

- **Position**: Set 1 consisted of all the variables associated with the defense and offense positions of the players in the field.
- **Skills**: Set 2 consisted of all the skills that were accounted as players' talents such as ball control, strength etc.
- **Balance and Accuracy**: Set 3 was related to player's speed and balance
- **Merits**: Set 4 was the players overall reputation and performance in FIFA -19
- **Worth**: The last one is our Dependent Variables , Income and Value.

To find the canonical variables that are most significant explaining the relationship between our first four sets and the dependent variables, we ran the CCA package to check out the associate within and in between skills and Balance and Accuracy. Because the factors in both of these sets have similar connotations to define players' talents.

The interpretation of raw canonical coefficients is the same as the linear regression models. Consider the Figure 8 above, and we want to interpret HeadingAccuracy on CV1. The interpretation would go as follows:

A one unit increase in HeadingAccuracy would have a .66 units increase in the value of CV1 for the Set 3, when other variables are held constant.

Furthermore, from figure 7 we get most of the correlation from the first two canonical variates. Wilk's Lambda test statistics in figure 9, determines the combined dimensions from 1 to 5. Since the p-value is less than .05 level of significance, it concludes that all the dimensions are statistically significant. Figure 10 indicates that most of the attributes of Set 2 and Set 3 are negatively correlated to each other.

Since we have different sets of attributes that have some significance in calculating the players' worth, we decided to do a canonical correlation analysis with each set separately to find their significance with our dependent variables. Using the Ycca package in R, since we have two dependent variables , we get two canonical variates. For all four sets of data, we are going to use only the first canonical variate for our interpretations and results since it is the one correlating the most among each set. Subsequently Position, Skills, BalanceandAccuracy have 56%, 62%,39% canonical correlation with Worth respectively. Since player's Metrics is just players overall performance in FIFA which highly depends on players attributes such as skills, position and BalanceandAccuracy, I used cbind to put together the three aforementioned sets to run an analysis with worth. The top 6 Player's attributes were LCM, LM, LAM, LF, LW and BallControl with a canonical correlation of 94%, 87%, 85%, 82%, 82% and 80% respectively. From the structural loadings we can deduce that players' positions in the field had the most significance when choosing players merits and overall performance in FIFA. Additionally, we can also see the significance of canonical variate from Bartlett's Chi-Squared Test shown in figure 11.

Since we have shown the relationship between the variables of each set with one another, we precede back to our research question about finding canonical variables to find an association between players worth and its attributes. If we break down the players' attributes with respect to their interconnected sets, then we can see their correlations in detail.

For instance, in set1 we are looking at different positions, the players hold in the field with correspondence to their wages and value in FIFA-19. We can see that the first 6 positions have a higher correlation with players' worth. One of the perks of ycca is that it gives us the redundancy among the canonical variables of the sets. As we said before looking at the canonical correlations of each set between the two canonical variates, we are only going to use the first canonical variate because that gives us the most information as shown in the table 4. We can interpret the fraction of total variance explained by each CV across sets. We also need to see how the variables interact within the set. ycca package gives us canonical communalities that construes how much of the total variance for each variable within sets is explained. For instance, reactions explain 82% of the variance within the attributes.

Last but not least, you can also see how much of the variance of the first set is explained in the other set and vice versa. Furthermore, it helps us in determining whether we would be able to drive something new from the given sets of variables or not. In figure 13, we can construe that 37% of players' worth is described by the given set of player's attributes.

In conclusion, if we were to look at significant skills that would aid the players in increasing its worth and wages and to predict the upcoming players then Reaction Composure, LCM, BallControl, LM and LAM 90%, 80%, 79% 77%, 72% respectively would be the top 6 attributes that correlated the most to players income and value, based on the data in FIFA-19.

Correspondence Analysis

After doing the CA, most of the dimensions explain roughly the same amount of variance as seen in figure 15. This might signify that further dimension reduction might have to be done.

Figure 16 is the symmetric plot of the first two dimensions. Red is a soccer club and blue are the nationalities of the players. The distance between each color dots represents their correspondence. We can see some outliers that show that there are very little players from certain nationalities that play in top level soccer. Countries like China or Uzbekistan are some that look that way.

Conclusion - in terms of application

The primary goal of our analysis was to create a model to predict player wage in the FIFA 19 video game, given their relevant player characteristics. In creating this model, through our analysis, we were also able to recognize certain player characteristics that were very similar, that could be grouped into broader categories to describe player attributes. This simplified the process of predicting wage by reducing the number of characteristics used, and made the model easier to understand and interpret.

Using four different techniques to group the player characteristics, we were able to identify four distinct groups to explain wage: Attack, Defense, Physicality, and Speed. Additionally, we concluded that a player's nationality could also have an impact on wage, based on the popularity of the sport in their respective country. Altogether, these insights provide a simple, efficient way to model player wage.

In the future, we may conduct further research, such as sensitivity analysis, to refine our groups and test our model. We could apply our analyses to other datasets, such as other years in the video game series, to ensure its applicability. Finally, we could compare our results with other studies that have modeled similar data, to verify our results, and the efficacy of our model.

References:

Battré, M., Deutscher, C., & Frick, B. (2009). Salary determination in the German Bundesliga: a panel study. In *No 0811, IASE Conference Papers, International Association of Sports Economists*.

Christiansen, N. A., Sievertsen, H. H., & Seminar, S. E. (2008). Superstar Effect in Italian Football–. In *Seminar Paper*.

Frick, B. (2006). *Salary determination and the pay-performance relationship in professional soccer: evidence fron Germany* (pp. 125-146). Ediciones de la Universidad de Oviedo.

Frick, B. (2007). THE FOOTBALL PLAYERS'LABOR MARKET: EMPIRICAL EVIDENCE FROM THE MAJOR EUROPEAN LEAGUES. *Scottish Journal of Political Economy*, *54*(3), 422-446.

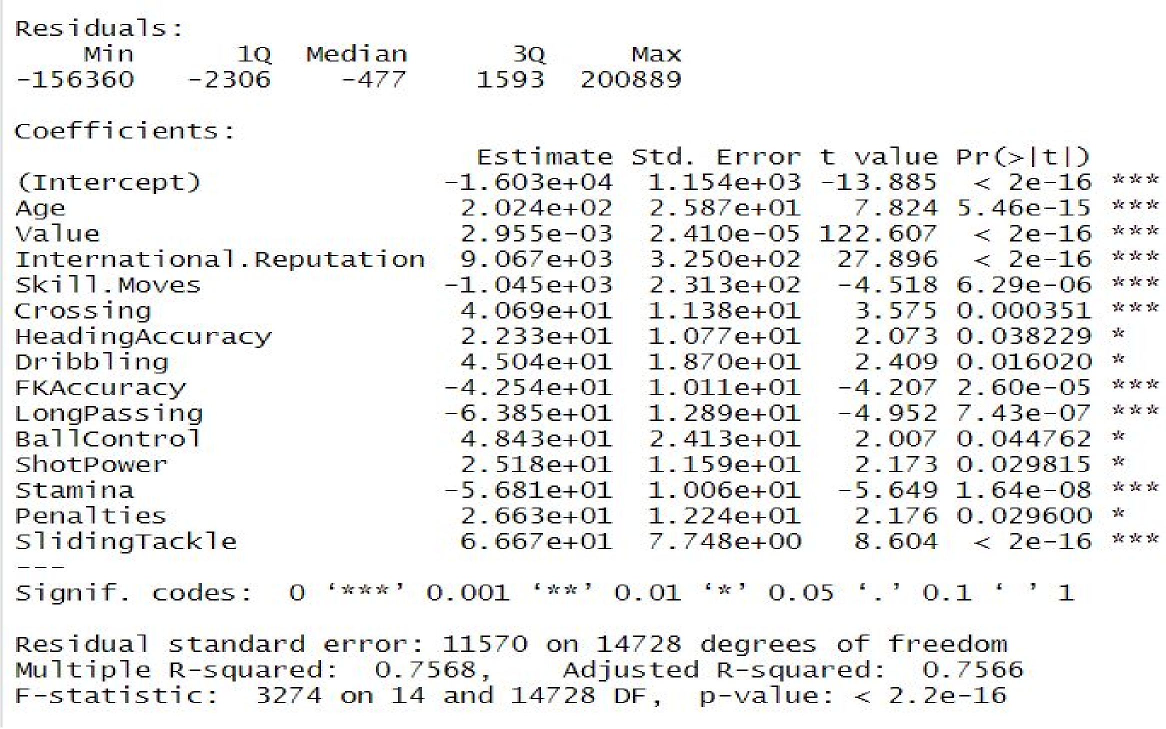
Frick, B. (2011). Performance, salaries, and contract length: Empirical evidence from German soccer. *International Journal of Sport Finance*, *6*(2), 87.

Lucifora, C., & Simmons, R. (2001). *Superstar effects in Italian football: An empirical analysis*. Università cattolica del Sacro Cuore.

Wicker, P., Prinz, J., Weimar, D., Deutscher, C., & Upmann, T. (2013). No pain, no gain? Effort and productivity in professional soccer. *International Journal of Sport Finance*, *8*(2), 124.

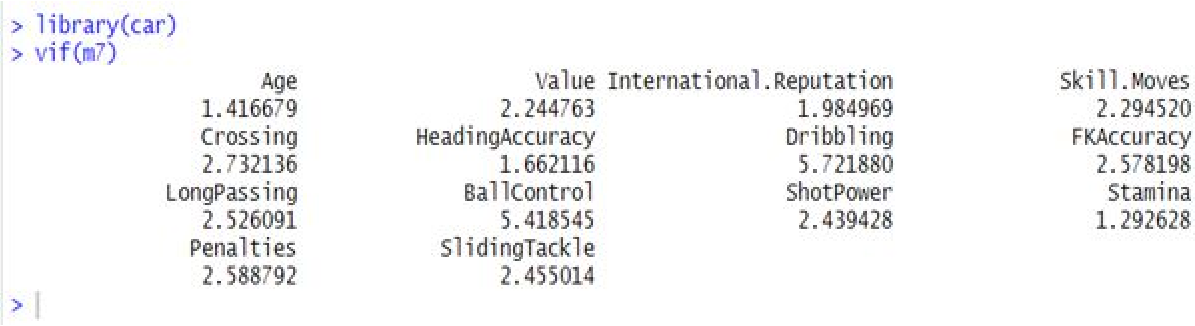
Yaldo, L., & Shamir, L. (2017). Computational Estimation of Football Player Wages, *International Journal of Computer Science in Sport*, *16*(1), 18-38. doi: https://doi.org/10.1515/ijcss-2017-0002

Tables and Figures:



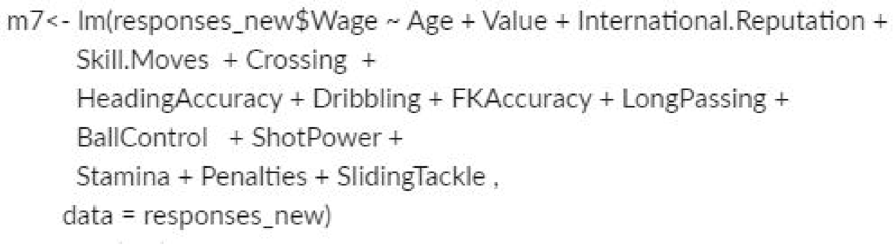
```
Residuals:
    Min      1Q  Median      3Q     Max 
-156360   -2306    -477    1593  200889 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -1.603e+04  1.154e+03 -13.885  < 2e-16 ***
Age                        2.024e+02  2.587e+01   7.824 5.46e-15 ***
Value                      2.955e-03  2.410e-05 122.607  < 2e-16 ***
International.Reputation  9.067e+03  3.250e+02  27.896  < 2e-16 ***
Skill.Moves               -1.045e+03  2.313e+02  -4.518 6.29e-06 ***
Crossing                   4.069e+01  1.138e+01   3.575 0.000351 ***
HeadingAccuracy            2.233e+01  1.077e+01   2.073 0.038229 *  
Dribbling                  4.504e+01  1.870e+01   2.409 0.016020 *  
FKAccuracy                -4.254e+01  1.011e+01  -4.207 2.60e-05 ***
LongPassing               -6.385e+01  1.289e+01  -4.952 7.43e-07 ***
BallControl                4.843e+01  2.413e+01   2.007 0.044762 *  
ShotPower                  2.518e+01  1.159e+01   2.173 0.029815 *  
Stamina                   -5.681e+01  1.006e+01  -5.649 1.64e-08 ***
Penalties                  2.663e+01  1.224e+01   2.176 0.029600 *  
SlidingTackle              6.667e+01  7.748e+00   8.604  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11570 on 14728 degrees of freedom
Multiple R-squared:  0.7568,	Adjusted R-squared:  0.7566 
F-statistic:  3274 on 14 and 14728 DF,  p-value: < 2.2e-16
```

Figure 1 : Regression Model Output



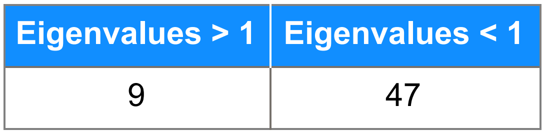
```
> library(car)
> vif(m7)
                     Age                    Value International.Reputation              Skill.Moves 
                1.416679                 2.244763                 1.984969                 2.294520 
                Crossing          HeadingAccuracy                Dribbling               FKAccuracy 
                2.732136                 1.662116                 5.721880                 2.578198 
             LongPassing              BallControl                ShotPower                  Stamina 
                2.526091                 5.418545                 2.439428                 1.292628 
               Penalties            SlidingTackle 
                2.588792                 2.455014 
> |
```

Figure 2: Multicollinearity Check



```
m7<- lm(responses_new$Wage ~ Age + Value + International.Reputation +
      Skill.Moves  + Crossing  +
      HeadingAccuracy + Dribbling + FKAccuracy + LongPassing +
      BallControl   + ShotPower +
      Stamina + Penalties + SlidingTackle ,
    data = responses_new)
```

Figure 3 : Regression Model



Eigenvalues > 1	Eigenvalues < 1
9	47

Table 1: Eigenvalues greater than 1

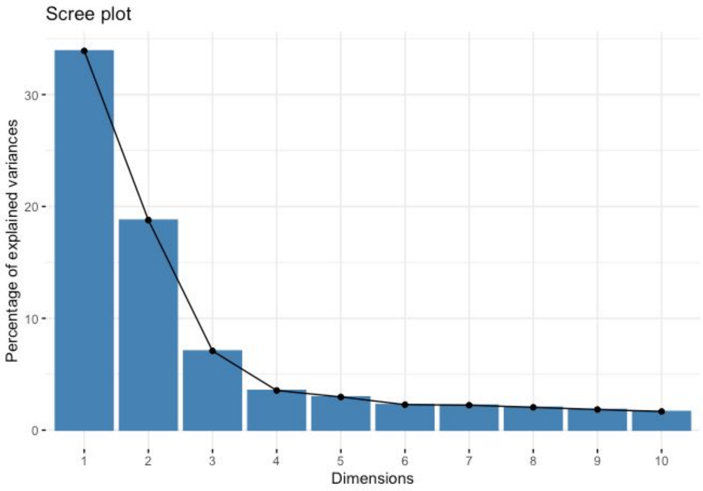
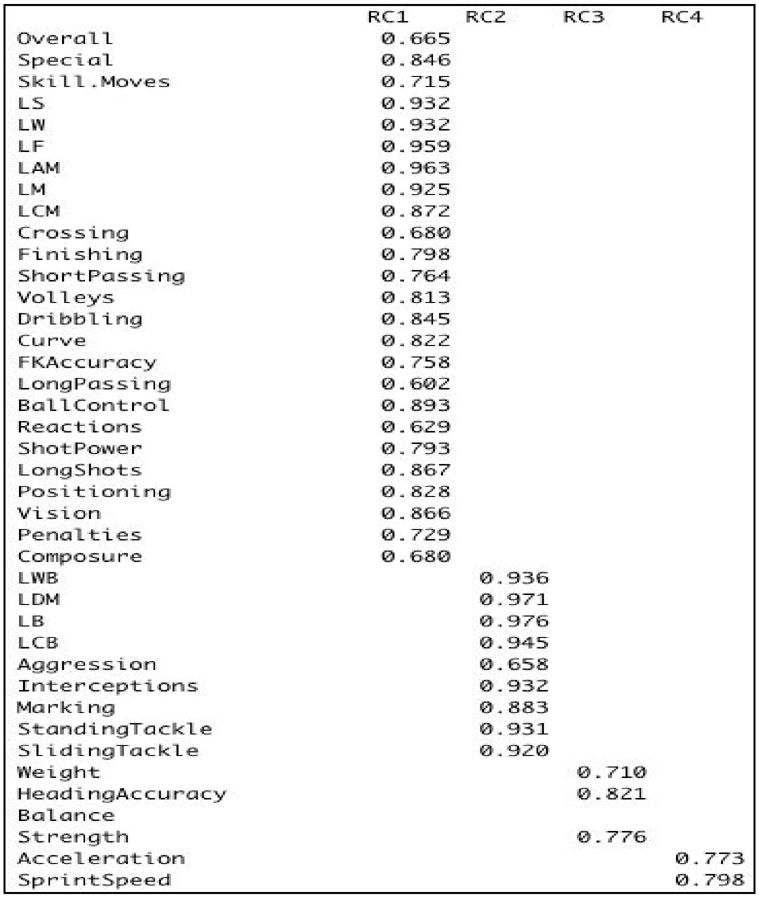


Figure 4: Scree Plot

	Attack	Defense	Physicality	Speed
Loadings	18.9	9.8	3.7	3.1
Proportion of Variance	33.7%	17.5%	6.6%	5.5%
Cumulative Variance	33.7%	51.3%	57.8%	63.3%

Table 2: Factors Contributing Variance for PCA



```
                           RC1    RC2    RC3    RC4
Overall                     0.665
Special                     0.846
Skill.Moves                 0.715
LS                          0.932
LW                          0.932
LF                          0.959
LAM                         0.963
LM                          0.925
LCM                         0.872
Crossing                    0.680
Finishing                   0.798
ShortPassing                0.764
Volleys                     0.813
Dribbling                   0.845
Curve                       0.822
FKAccuracy                  0.758
LongPassing                 0.602
BallControl                 0.893
Reactions                   0.629
ShotPower                   0.793
LongShots                   0.867
Positioning                 0.828
Vision                      0.866
Penalties                   0.729
Composure                   0.680
LWB                                0.936
LDM                                0.971
LB                                 0.976
LCB                                0.945
Aggression                         0.658
Interceptions                      0.932
Marking                            0.883
StandingTackle                     0.931
SlidingTackle                      0.920
Weight                                    0.710
HeadingAccuracy                           0.821
Balance
Strength                                  0.776
Acceleration                                     0.773
SprintSpeed                                      0.798
```

Figure 5: Component Loadings

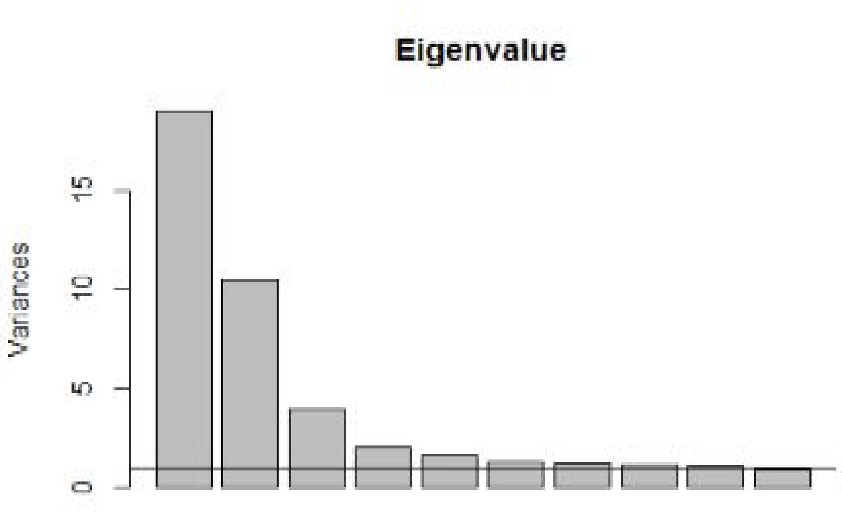
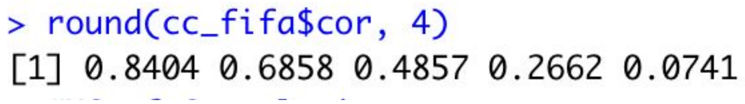


Figure 6: Eigenvalues

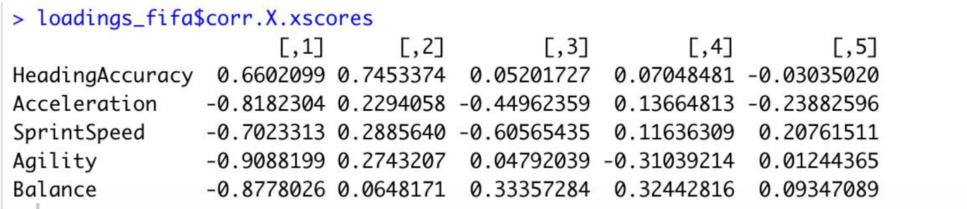
	Attack	Defense	Physicality	Speed
SS Loadings	17.979	9.533	3.747	2.809
Proportion Variance	0.321	0.170	0.067	0.050
Cumulative Variance	0.321	0.491	0.558	0.608

Table 3: Factors Contributing Variance for FA



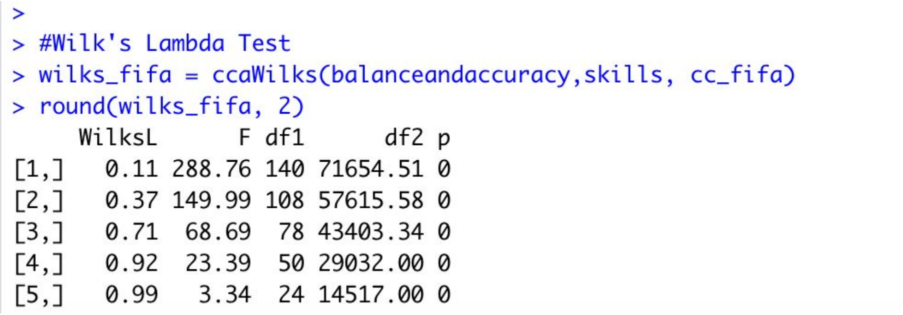
```
> round(cc_fifa$cor, 4)
[1] 0.8404 0.6858 0.4857 0.2662 0.0741
```

Figure 7: Canonical Correlation of Set 2 and 3



```
> loadings_fifa$corr.X.xscores
                      [,1]      [,2]        [,3]        [,4]        [,5]
HeadingAccuracy  0.6602099 0.7453374  0.05201727  0.07048481 -0.03035020
Acceleration    -0.8182304 0.2294058 -0.44962359  0.13664813 -0.23882596
SprintSpeed     -0.7023313 0.2885640 -0.60565435  0.11636309  0.20761511
Agility         -0.9088199 0.2743207  0.04792039 -0.31039214  0.01244365
Balance         -0.8778026 0.0648171  0.33357284  0.32442816  0.09347089
```

Figure 8: Set 3 loadings



```
>
> #Wilk's Lambda Test
> wilks_fifa = ccaWilks(balanceandaccuracy,skills, cc_fifa)
> round(wilks_fifa, 2)
     WilksL      F df1      df2 p
[1,]   0.11 288.76 140 71654.51 0
[2,]   0.37 149.99 108 57615.58 0
[3,]   0.71  68.69  78 43403.34 0
[4,]   0.92  23.39  50 29032.00 0
[5,]   0.99   3.34  24 14517.00 0
```

Figure 9: Set 2 and Set 3 Wilk’s Lambda Test

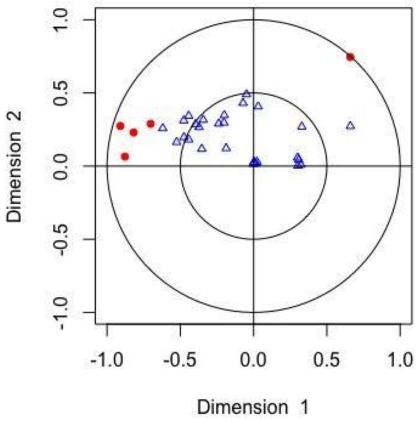
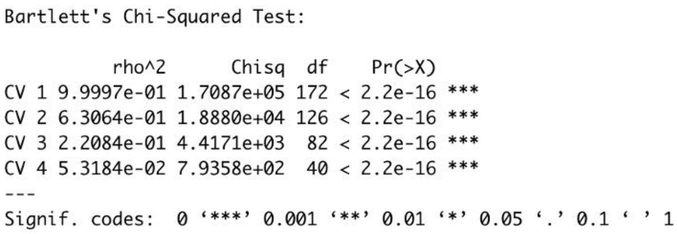


Figure 10: CC Plot of Skills and BalanceandAccuarcy



```
Bartlett's Chi-Squared Test:

          rho^2      Chisq  df    Pr(>X)
CV 1 9.9997e-01 1.7087e+05 172 < 2.2e-16 ***
CV 2 6.3064e-01 1.8880e+04 126 < 2.2e-16 ***
CV 3 2.2084e-01 4.4171e+03  82 < 2.2e-16 ***
CV 4 5.3184e-02 7.9358e+02  40 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Players’ attributes and Metrics

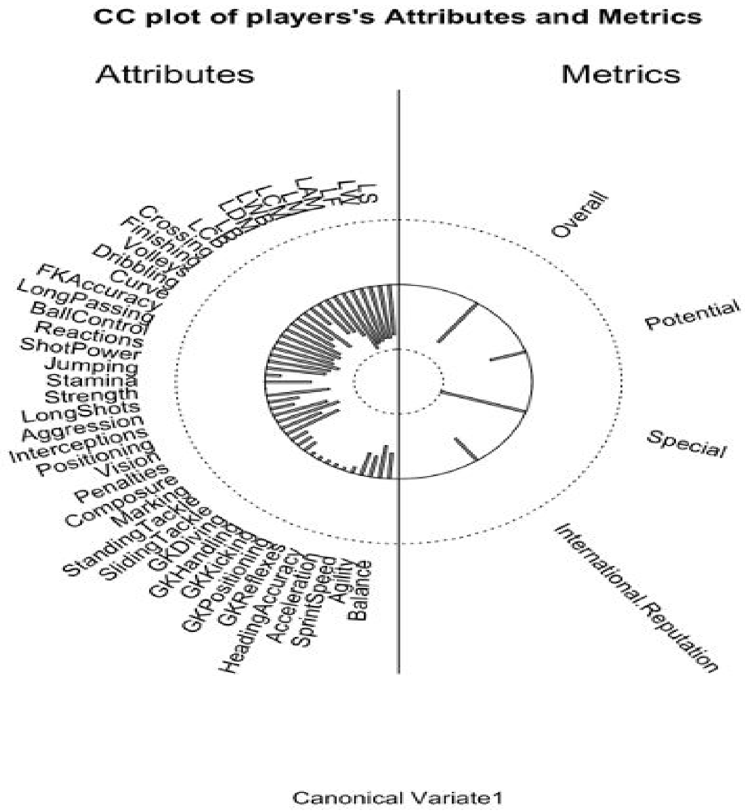


Figure 12

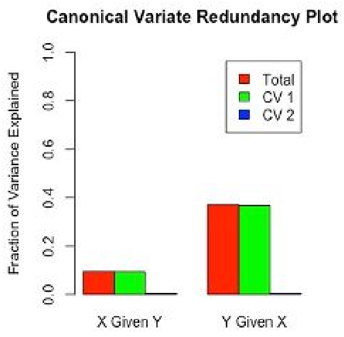
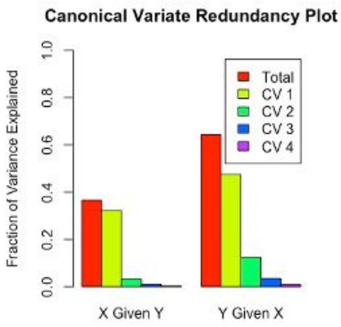
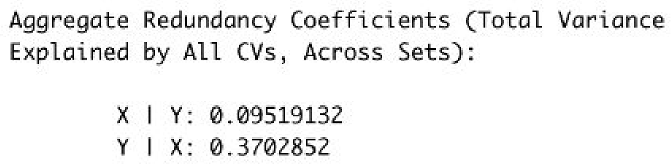
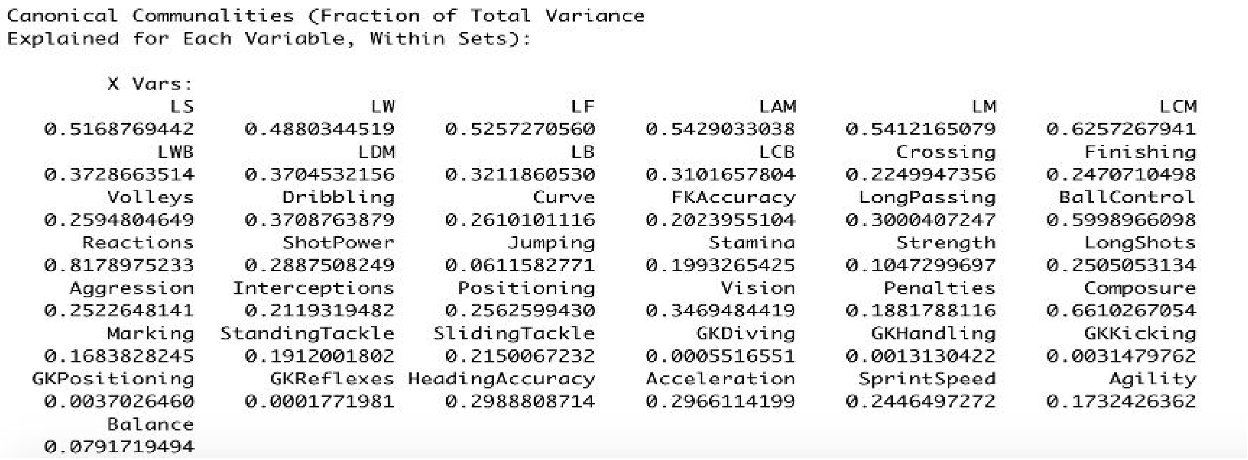
Players’ Attributes and Worth	Players’ Attributes and Metrics
	

Table 4



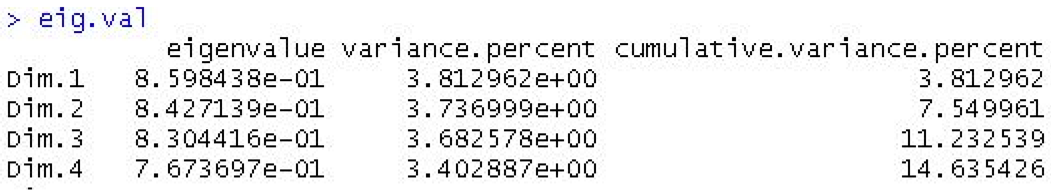
```
Aggregate Redundancy Coefficients (Total Variance
Explained by All CVs, Across Sets):

        X | Y: 0.09519132
        Y | X: 0.3702852
```

Figure 13: Redundancy of Players' Attributes and Worth



```
Canonical Communalities (Fraction of Total Variance
Explained for Each Variable, Within Sets):

        X Vars:
           LS              LW              LF             LAM              LM             LCM
 0.5168769442    0.4880344519    0.5257270560    0.5429033038    0.5412165079    0.6257267941
          LWB             LDM              LB             LCB        Crossing       Finishing
 0.3728663514    0.3704532156    0.3211860530    0.3101657804    0.2249947356    0.2470710498
      Volleys       Dribbling           Curve      FKAccuracy     LongPassing     BallControl
 0.2594804649    0.3708763879    0.2610101116    0.2023955104    0.3000407247    0.5998966098
    Reactions       ShotPower         Jumping         Stamina        Strength       LongShots
 0.8178975233    0.2887508249    0.0611582771    0.1993265425    0.1047299697    0.2505053134
   Aggression   Interceptions     Positioning          Vision       Penalties       Composure
 0.2522648141    0.2119319482    0.2562599430    0.3469484419    0.1881788116    0.6610267054
      Marking  StandingTackle    SlidingTackle        GKDiving      GKHandling       GKKicking
 0.1683828245    0.1912001802    0.2150067232    0.0005516551    0.0013130422    0.0031479762
GKPositioning     GKReflexes HeadingAccuracy    Acceleration     SprintSpeed         Agility
 0.0037026460    0.0001771981    0.2988808714    0.2966114199    0.2446497272    0.1732426362
      Balance
 0.0791719494
```

Figure 14: Players' Attributes



```
> eig.val
         eigenvalue variance.percent cumulative.variance.percent
Dim.1  8.598438e-01     3.812962e+00                    3.812962
Dim.2  8.427139e-01     3.736999e+00                    7.549961
Dim.3  8.304416e-01     3.682578e+00                   11.232539
Dim.4  7.673697e-01     3.402887e+00                   14.635426
```

Figure 15: CA eigenvalues

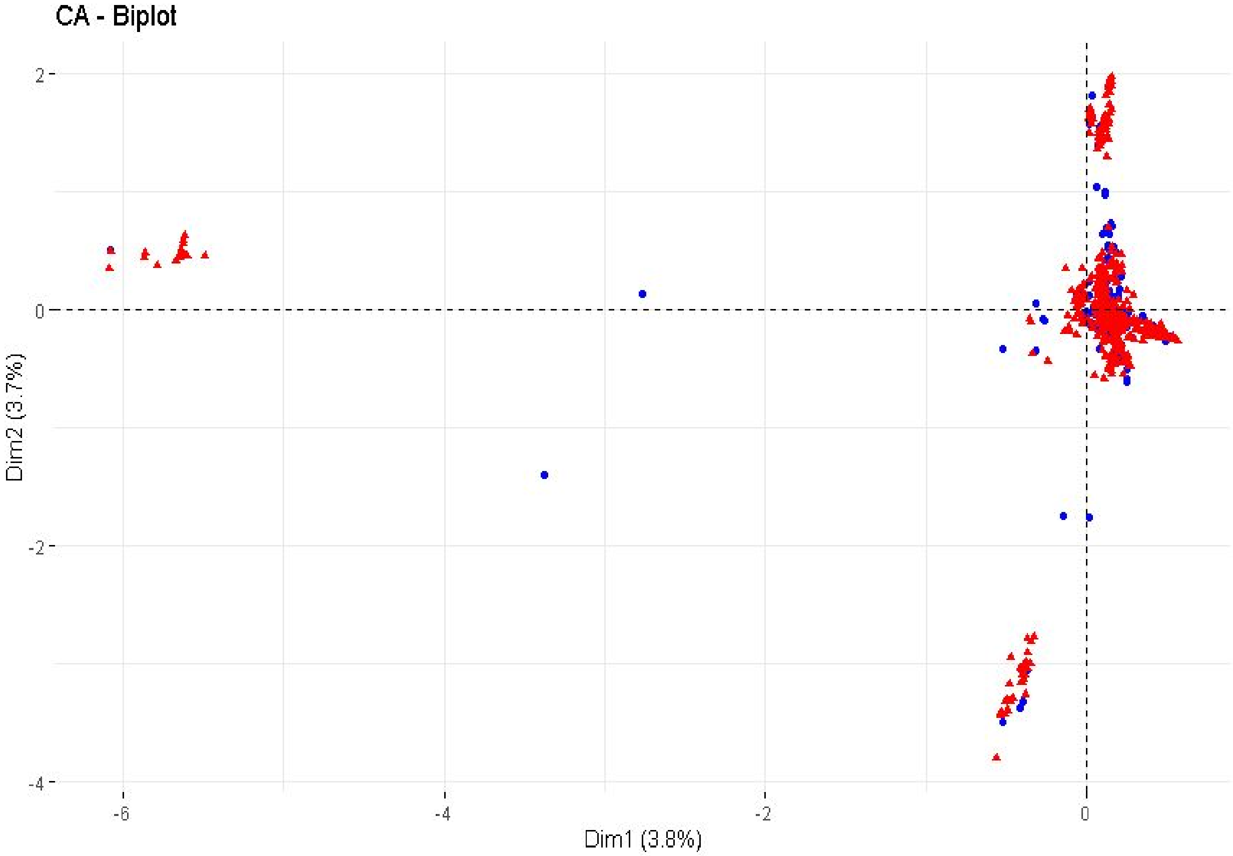


Figure 16: Symmetric plot of team and nationality