Exploratory Multivariate Data Analysis on the Premier League

S&DS 363 – Multivariate Statistical Methods for the Social Sciences

May 10th, 2022

Table of Contents

Introduction

FIFA 20 is a video game which simulates soccer. Leagues and countries that are recognized by FIFA, the international football organization, are included in the game. Gamers first start the game with a random set of soccer players. They have to compete in matches or complete challenges in order to purchase more players from the transfer market, which allows gamers to buy soccer players from each other. Soccer players are given statistics based off their real-life gameplay in the prior year. For example, players who scored many goals are typically given a higher shooting score.

The Premier League the highest level in the English football system. It consists of 20 teams that compete with each other yearly. Within this league, there are some clubs that are considered to be stronger than others. The goal of this project is to perform multivariate exploratory analysis on the Premier League using player statistics from FIFA 2020.

Design and Primary Questions

The purpose of this analysis is to see if players can be categorized by position using overall skill, wage, mentality composure, and movement agility. The soccer positions are forwards, midfielders, and defenders. Forwards focus more on attacking and goals. Midfielders focus on bringing the ball from defenders to forwards. Defenders keep the other team from scoring. A common belief that soccer players hold is that forwards are considered the best players since they score the most goals. Moreover, better players tend to get paid more because of their talent. Mentality composure measures how well a player performs under stress, and movement agility measures how well a player can adjust their speed and direction on the field.

Is there a significant difference in overall skill, wage, mentality composure, and movement agility across positions? To answer this question, one-way MANOVA will be performed. And if there is a difference, is it possible to categorize players by these variables? Discriminant Analysis will be performed to explore that topic. Lastly, are there other ways that we can group different variables from the dataset? Principle Component Analysis will be performed to see if there are groupings of the variables that explain variability better than these four variables.

The FIFA 20 dataset contains statistics for each player in the game. We will use a subset of dataset and use data from clubs who were apart of the Premier League during the 2019 – 2020 season. It contains a multitude of statistics, from age, height, weight, overall scores, reputations, attacking skills, defensive skills, and strengths. We will use the Premier League data to see if it is possible to answer such questions, and see if these answers can be applied to other leagues as well.

Data

The data is collected from the player statistics in FIFA 20. Every soccer player that plays for a league that is recognized by FIFA is included into the video game. Statistics that represent the virtual gameplay of a player, such as shooting, movement agility, balance, and crossing are determined by the actual gameplay of that soccer player from the year prior. Officials from FIFA met to discuss their gameplay and assign numbers from 1-99 based on it. Although many officials are a part of the decision of scoring and they follow a pre-defined rubric, scoring is still biased as it is subjective. There is really no way to know if the officials are consistent with their scoring and this could produce some error within the analysis.

Name	Type	Description
sofifa_id	Categorical	Unique ID given to each player
short_name	Categorical	Shortened name of the player
club	Categorical	Club that the player plays for
nationality	Categorical	Country the player was born in
overall	Continous (1-99)	The overall rating of the player; higher ratings indicate the player is better
potential	Continous (1-99)	The overall rating of the player if the player trains a lot
value_eur	Continous (Euros)	The value of the player in the transfer market as of 10/14/19
wage_eur	Continous (Euros)	The salary of the player in 2019
mentality_composure	Continous (1-99)	How well the player performs under preassure; higher ratings indicate they perform better under preassure
movement_agility	Continous (1-99)	How well the player changes speed and direction; higher ratings indicate higher agility
age	Continous (Years)	Age of the player
height_cm	Continous (Cm)	Height of the player
weight_kg	Continous (Kg)	Weight of the player
international_reputation	Continous (1-5)	How well-known a player is; higher ratings indicate the player is well-known internationally
weak_foot	Continous (1-5)	How good a player is at using their weak foot; higher ratings indicate the player is good at using their weak foot
skill_moves	Continous (1-5)	How good a player is at performing difficlut moves; higher ratings indicate the player is good at doing complex techniques
attacking_crossing	Continous (1-99)	How good a player is at crossing the ball; higher ratings indicate the player is good at crossing
attacking_finishing	Continous (1-99)	How good a player is at setting up an opportunity to score; higher ratings indicate the player is good at finishing
attacking_heading_accuracy	Continous (1-99)	How good a player is at using their head to pass the ball; higher ratings indicate the player is good at heading
attacking_short_passing	Continous (1-99)	How good a player is at passing the ball for short distances; higher ratings indicate the player is good at short passing
attacking_volleys	Continous (1-99)	How good a player is at hitting the ball in the air; higher ratings indicate the player is good at volleys
defending_marking	Continous (1-99)	How good a player is at preventing a member of the other team from stealing the ball; higher ratings indicate the player is good at marking
defending_standing_tackle	Continous (1-99)	How good a player is at doing standle tackles; higher ratings indicate the player is good at standing tackles
defending_sliding_tackle	Continous (1-99)	How good a player is at doing slide tackles; higer ratings indicate the player is good at sliding tackles

Figure 1: Variables

Descriptive Plots and Summary Statistics

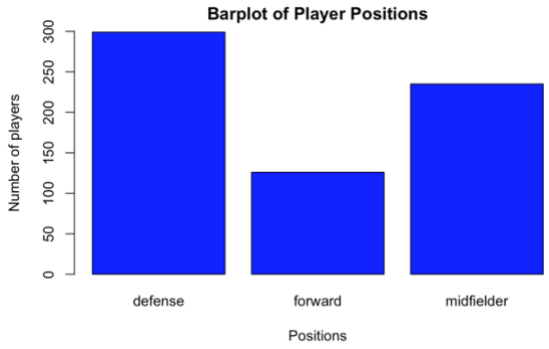


Figure 2: Bar Plot of Player Positions

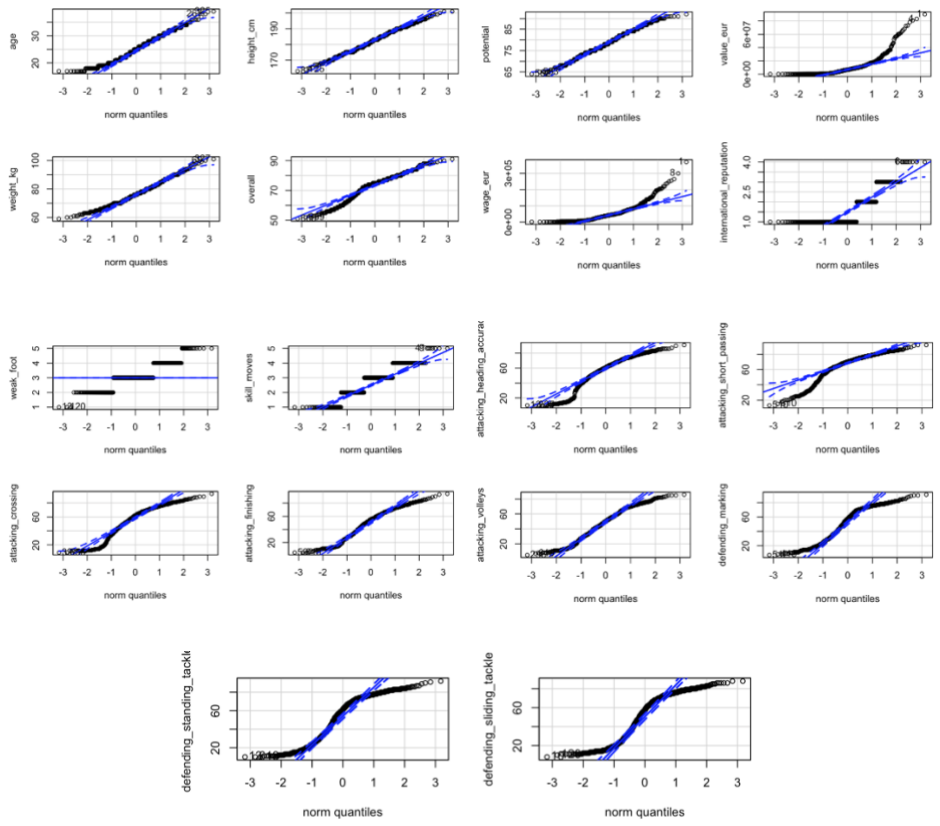


Figure 3: Norm Quantiles of Continuous Variables

Figure 4 reveals that not all of the variables are normally distributed. Wages and value experience the most skew since they have the most extreme outliers.

Multivariate Analysis

Examining Differences in Selected Variables by Position Using One-Way MANOVA

We first begin by observing differences in the means of overall score, wage, movement agility and mentality composure by positions. The three categories for the player position are forward, midfielder, and defense. A common belief that is held among many soccer fans is that forward players tend to get paid more since they are the ones who score the goals.

In order to perform MANOVA, the assumptions for it must be checked. The most important assumption of MANOVA is that the residuals of the MANOVA model are multivariate normally distributed. Thus, I will begin by not checking the rest of the assumptions and fit the data into a MANOVA model and look at the residuals before looking at the results of the model.

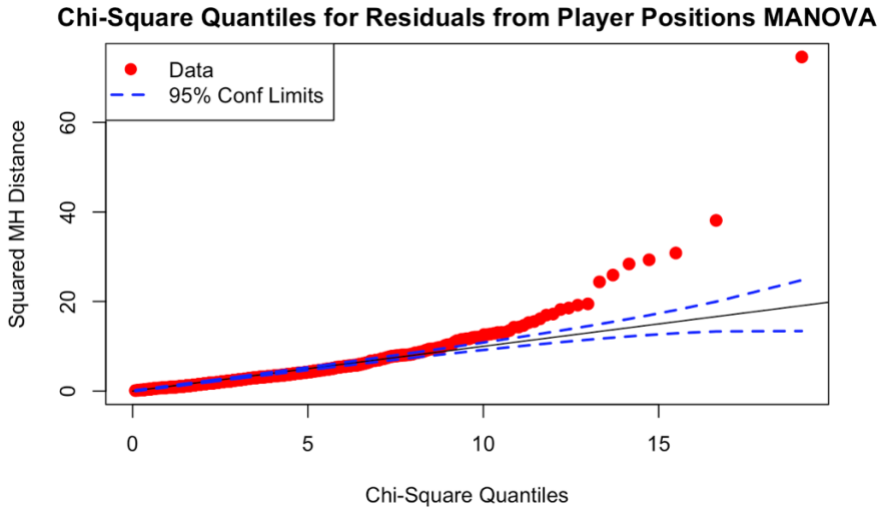


Figure 4: Chi-Square Quantiles for Residuals from Player Positions MANOVA

The residuals of the MANOVA model are not multivariate normally distributed, thus transformations will need to be applied to the response variables.

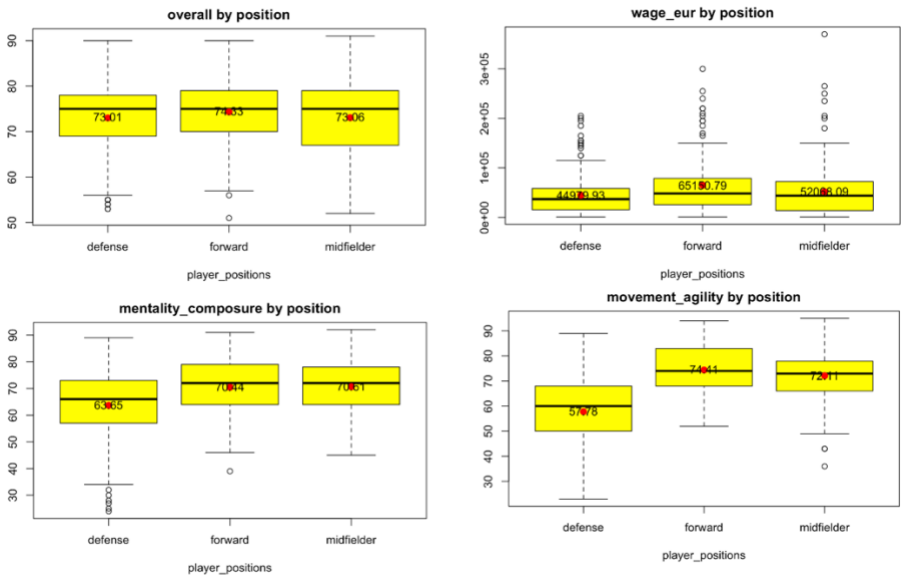


Figure 5: Boxplots for Dependent Variable by Position

Figure 5 shows boxplots for each variable by position. Wage appears to be right-skewed, and overall, mentality composure, and movement agility appear to be left-skewed. A cube-root transformation will be applied to the wage data and a squared transformation will be applied to the rest of the variables. Boxplots for each variable by position after transformations are shown below in Figure 6, and we can see that the data appears to look more normal.

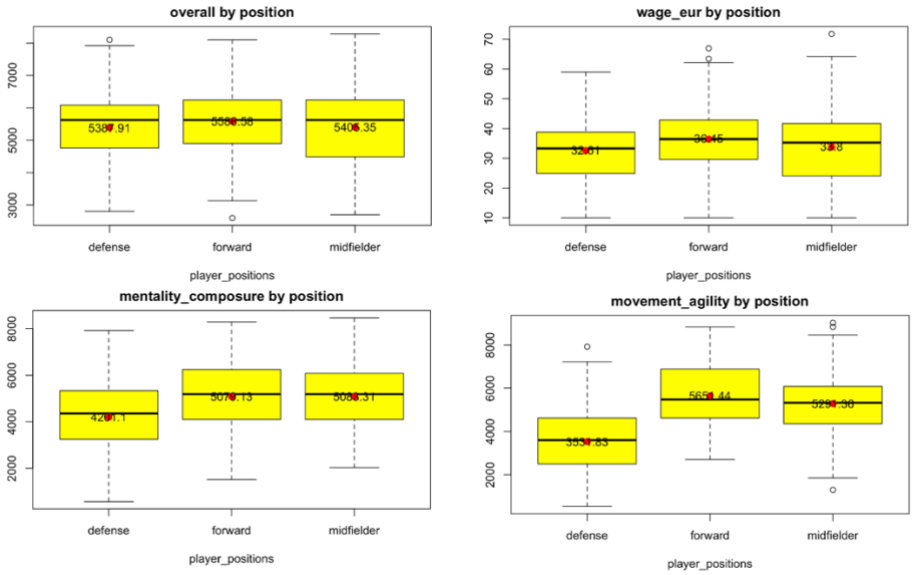


Figure 6: Boxplots for Transformed Dependent Variables by Position

Afterwards, chi-square quantile plots for observations within each group are made to ensure the dependent variables are multivariate normally distributed within each position. Figure 7 shows us that the transformed data does meet this assumption.

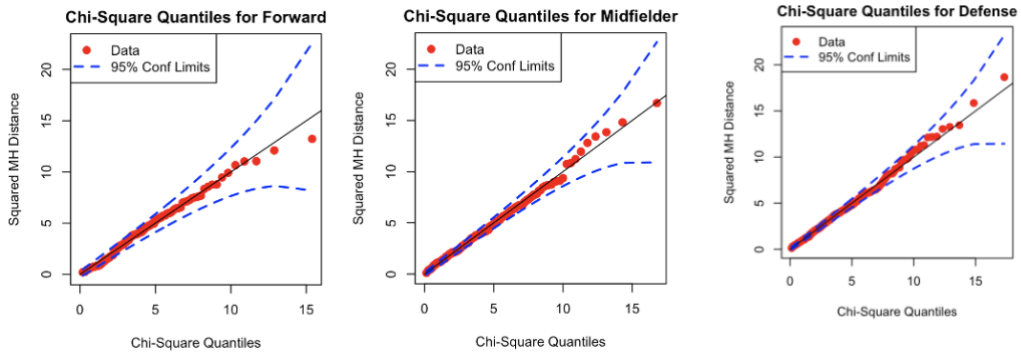


Figure 7: Chi-Square Quantiles by Position

Lastly, all observations are independent. One-way MANOVA by position is performed and Figure 8 shows the residuals are normally distributed.

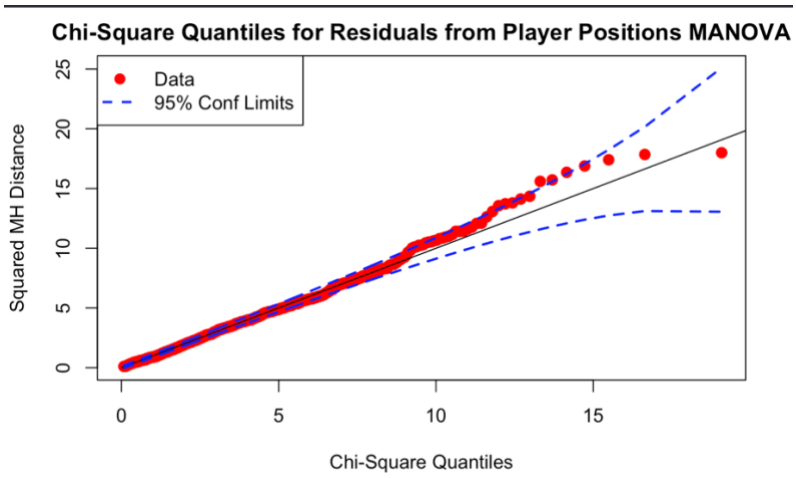
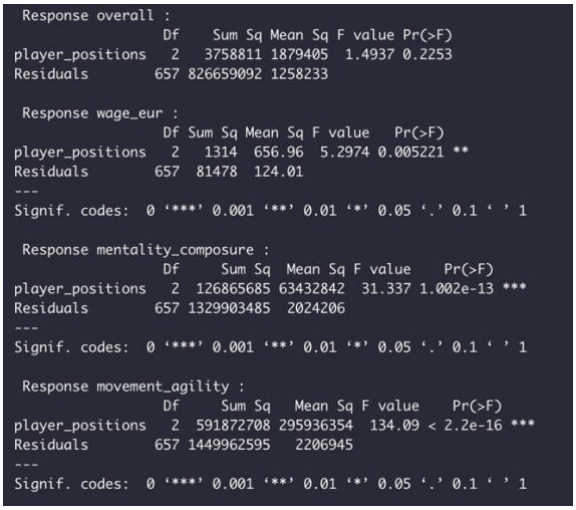


Figure 8: Chi-Square Quantiles for Residuals from Player Positions MANOVA

Figure 9 shows the test of significance of variables at the univariate level. We see that the overall score has a p-value of 0.2253, thus we fail to reject the null hypothesis and cannot conclude that there is a significant different in the overall score by position. This is reasonable considering how the overall score is calculated. Although a forward player may have excellent

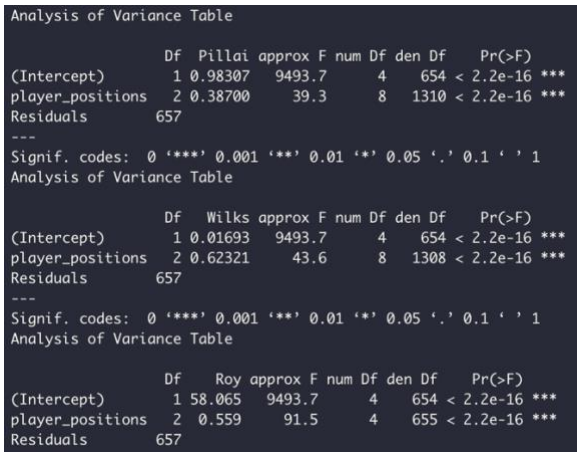
shooting skills, they probably have lower tackling skills. Meanwhile, a defense player may be great at tackling but not as good at shooting. In short, there is no player that is great at every skill thus the distribution of the overall score and the mean is similar at each position. This can be confirmed by looking at the Boxplots in Figure 6. The wage has a p-value of 0.005221. When alpha is set to 0.05, this p-value rejects the null hypothesis, thus there is evidence that there is a significant difference of wage across positions. MANOVA does not tell us which positions have higher wages than the others, but when looking at the Boxplot in Figure 6, it can be seen that forward players do have the highest average, which lines up with the prior belief mentioned earlier. The mentality composure has a p-value of 1.002e-13, thus the null hypothesis is rejected so it can be concluded that there is evidence that supports the claim that there is a significant difference of mentality composure across groups. The boxplot in Figure 6 shows that midfielders have the highest average in mentality composure, with the average for forward players being slightly lower than it. This makes sense, as one the most important jobs for midfielders are to pass the ball from the defense to the offense, and to help the offense not lose the ball once it is in their possession. They must always be alert of the soccer ball; hence they should be able to perform well under stress. The movement agility has a p-value of 2.2e-16, thus the null hypothesis is rejected, so there does exist evidence that shows there is a significant difference of movement agility across groups. The boxplot in Figure 6 show that forwards have the highest movement agility, meanwhile defenders have the lowest movement agility. It is reasonable to see that forwards tend to be faster since they must avoid defenders in order to score. Defenders are often times already positioned by the goal, so they do not need to rely on speed as much.



```
 Response overall :
                  Df    Sum Sq Mean Sq F value Pr(>F)
player_positions   2   3758811 1879405  1.4937 0.2253
Residuals        657 826659092 1258233

 Response wage_eur :
                  Df Sum Sq Mean Sq F value   Pr(>F)
player_positions   2   1314  656.96  5.2974 0.005221 **
Residuals        657  81478  124.01
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response mentality_composure :
                  Df     Sum Sq  Mean Sq F value    Pr(>F)
player_positions   2  126865685 63432842  31.337 1.002e-13 ***
Residuals        657 1329903485  2024206
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response movement_agility :
                  Df     Sum Sq   Mean Sq F value    Pr(>F)
player_positions   2  591872708 295936354  134.09 < 2.2e-16 ***
Residuals        657 1449962595   2206945
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: Test of Significance at the Univariate Level

Figure 10 shows the results of using Pillai's Trace, Wilks' Lambda, and Roy's Largest Root to test the significance at the multivariate level. Pillai's Trace has a value of 0.98307 and a p-value of less than 2.2e-16, which means that the positions play an impact on the responding variables. Wilks' Lambda has a value of 0.01693, which means approximately 1.7% of the variance percentages in the dependent variables are not explained by positions. Moreover, it has a p-value less than 2.2e-16, thus the positions explain most of these differences. Roy's test has a value of 58.065, which is a quite large, and a p-value less than 2.2e-16. Thus, all three tests show that the multivariate mean is different across all three positions.



```
Analysis of Variance Table

                  Df  Pillai approx F num Df den Df    Pr(>F)
(Intercept)        1 0.98307   9493.7      4    654 < 2.2e-16 ***
player_positions   2 0.38700     39.3      8   1310 < 2.2e-16 ***
Residuals        657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

                  Df   Wilks approx F num Df den Df    Pr(>F)
(Intercept)        1 0.01693   9493.7      4    654 < 2.2e-16 ***
player_positions   2 0.62321     43.6      8   1308 < 2.2e-16 ***
Residuals        657
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

                  Df    Roy approx F num Df den Df    Pr(>F)
(Intercept)        1 58.065   9493.7      4    654 < 2.2e-16 ***
player_positions   2  0.559     91.5      4    655 < 2.2e-16 ***
Residuals        657
```

Figure 10: Statistical Tests for MANOVA Model on the Multivariate Level

Figure 11 shows the coefficients of the MANOVA model. The coefficients for forward and midfielders are relative to the defenders. Since all the coefficients are positive, it can be concluded that overall, wage, mentality composure, and movement agility are generally lower for defenders. Forwards and midfielders have similar mentality composures and movement agilities, but it appears that their wages and overalls are different.

```
Coefficients:
                          overall   wage_eur  mentality_composure  movement_agility
(Intercept)              5387.913    32.605  4201.104             3531.829
player_positionsforward   198.666     3.849   878.031             2119.615
player_positionsmidfielder 17.436     1.191   882.207             1759.528
```

Figure 11:

Predicting the Position of Each Player Through Discriminant Analysis

Since there are significant differences across the multivariate means, Discriminant Analysis will be performed in order to see if these variables are accurate at categorizing the positions of each player. Discriminant Analysis has assumptions that are similar to MANOVA; thus, we will proceed using the transformed data from earlier. Linear Discriminant Analysis assumes similarity of covariance matrices across groups.

```
[1] "Covariance Matrix for Forward Players"
                          overall   wage_eur mentality_composure movement_agility
overall             1327506.97 13670.8505          1558229.51       808640.676
wage_eur              13670.85   151.5577            15912.37         8695.643
mentality_composure 1558229.51 15912.3717          2180834.18       762002.892
movement_agility     808640.68  8695.6428           762002.89      2514445.529
[1] "Covariance Matrix for Midfielders"
                          overall   wage_eur mentality_composure movement_agility
overall             1354152.23 12817.8552          1389677.52        404333.81
wage_eur              12817.86   132.3116            13304.86          3754.71
mentality_composure 1389677.52 13304.8646          1832936.23        384896.32
movement_agility     404333.81  3754.7105           384896.32       1825710.27
[1] "Covariance Matrix for Defenders"
                          overall   wage_eur mentality_composure movement_agility
overall             1153856.0 10409.0002          1176689.9       271951.707
wage_eur              10409.0   105.9464            11200.5         3454.001
mentality_composure 1176689.9 11200.5017          2108698.4       922131.256
movement_agility     271951.7  3454.0014           922131.3      2377317.786
[1] "Ratio of Largest to Smallest Covariance Elements for Forwards and Midfielders"
```

```
[1] "Ratio of Largest to Smallest Covariance Elements for Forwards and Midfielders"
                   overall wage_eur mentality_composure movement_agility
overall                1.0      1.1                 1.1              2.0
wage_eur               1.1      1.1                 1.2              2.3
mentality_composure    1.1      1.2                 1.2              2.0
movement_agility       2.0      2.3                 2.0              1.4
[1] "Ratio of Largest to Smallest Covariance Elements for Forwards and Defenders"
                   overall wage_eur mentality_composure movement_agility
overall                1.2      1.3                 1.3              3.0
wage_eur               1.3      1.4                 1.4              2.5
mentality_composure    1.3      1.4                 1.0              1.2
movement_agility       3.0      2.5                 1.2              1.1
[1] "Ratio of Largest to Smallest Covariance Elements for Midfilders and Defenders"
                   overall wage_eur mentality_composure movement_agility
overall                1.2      1.2                 1.2              1.5
wage_eur               1.2      1.2                 1.2              1.1
mentality_composure    1.2      1.2                 1.2              2.4
movement_agility       1.5      1.1                 2.4              1.3
```

Figure 12: Covariance Matrix by Position and Ratios of Matrices

All of the ratios of largest to smallest covariance elements for each position are less than 4, thus this assumption is met. A Box's M-test statistic will not be computed since it is sensitive to outliers and large datasets thus it is likely it may falsely reject the null hypothesis.

When there are no prior probabilities, Figure 13 shows that the raw results have an accuracy of 63% and Figure 14 shows the cross-validated results have an accuracy of 62%. The cross-validated results have a slightly slower accuracy than the raw results, thus suggesting that this model is overfit.

```
            defense forward midfielder
  defense       228       2         69
  forward        33      19         74
  midfielder     53      13        169
[1] 0.63
```

Figure 13: Raw Results With No Priors

```
            defense forward midfielder
  defense       221       3         75
  forward        27      18         81
  midfielder     58       8        169
[1] 0.62
```

Figure 14: CV Results With No Priors

We are given two models, LD1 and LD2, as shown in Figure 15. LD1 has a proportion of trace of 0.9501, which is significantly higher than LD2, thus it seems to be the better model.

```
[1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"       "call"
Call:
lda(prim[, c("overall", "wage_eur", "mentality_composure", "movement_agility")],
    grouping = prim$player_positions)

Prior probabilities of groups:
   defense    forward midfielder
 0.4530303  0.1909091  0.3560606

Group means:
            overall wage_eur mentality_composure movement_agility
defense    5387.913 32.60537            4201.104         3531.829
forward    5586.579 36.45478            5079.135         5651.444
midfielder 5405.349 33.79629            5083.311         5291.357

Coefficients of linear discriminants:
                               LD1           LD2
overall              0.0011107276 -0.0003596454
wage_eur            -0.0361549333  0.1677935568
mentality_composure -0.0006091071 -0.0009761319
movement_agility    -0.0005165919  0.0002761078

Proportion of trace:
   LD1    LD2
0.9501 0.0499
```

Figure 15: Linear Analysis Functions

Figure 16 reports that in LD1, overall has the highest weight (when looking at absolute values of the standardized coefficients). In LD2, wage and mentality composure have the highest weights (when looking at absolute values of the standardized coefficients). Overall has the lowest weight but it still plays a role in LD2.

```
[1] "Raw (Unstandardized) Coefficients"
                      LD1  LD2
overall              0.00 0.00
wage_eur            -0.04 0.17
mentality_composure  0.00 0.00
movement_agility     0.00 0.00
[1] "Normalized Coefficients"
                      LD1   LD2
overall              0.01  0.00
wage_eur            -0.21  0.98
mentality_composure  0.00 -0.01
movement_agility     0.00  0.00
[1] "Standardized Coefficients"
                      LD1   LD2
overall              1.25 -0.40
wage_eur            -0.41  1.88
mentality_composure -0.91 -1.45
movement_agility    -0.91  0.49
```

Figure 16: Coefficients of Linear Discriminant Functions

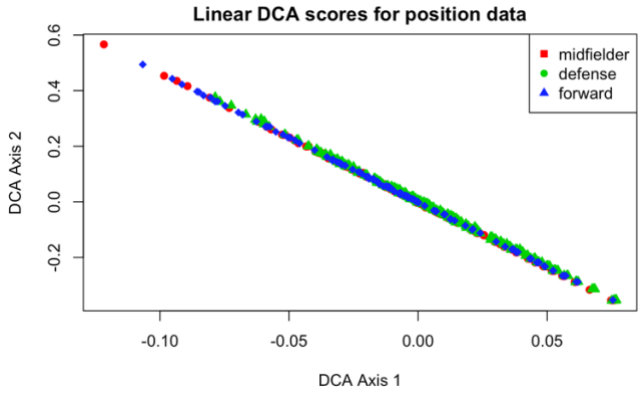


Figure 17: Linear DCA Scores for Premier League

Linear DCA scores are shown in Figure 17. There are not clear clumps of each position on the plot. This is not surprising considering that the model only had an accuracy of 63%. Moreover, since we have a diagonal line, it implies that each player received a similar score by both functions. Many scores overlap with each other from all groups, thus suggesting that this model is not a great predictor.

Figure 18 shows the partition plots. Wage and movement agility has the lowest error rate among all the possible pairs of the variables. This suggests that using a model with only wage and movement agility will have the highest accuracy among all the possible pairs of variables. Notice that mentality composure and movement agility follow close behind, with an error rate of

0.418. This is interesting to note since these two had the lowest p-values in the MANOVA model. Moreover, the rest of the pairs that include either mentality composure or movement agility has an error rate that ranges from 0.415 – 0.464. However, the partition plot for overall and wage has an error rate of 0.541, thus suggesting that these two variables without either mentality composure or movement agility is not as accurate. Since movement agility is in the pair with the lowest error rate, it may be the best predictor.

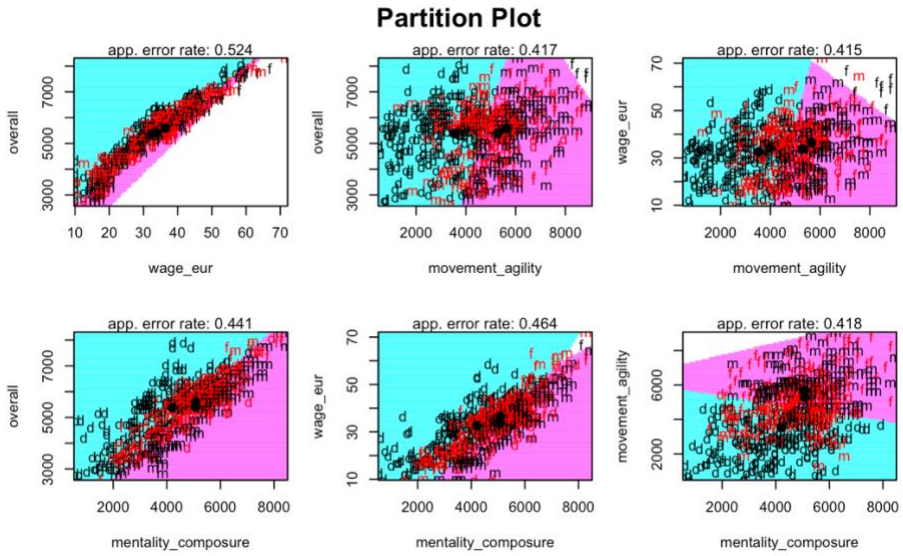


Figure 18: Partition Plots for Linear DCA

Stepwise Discriminant Analysis will be performed to determine which variables are important. Figure 19 shows that when Stepwise regression is preformed, only movement agility is kept. This lines up with our results of our previous Linear Discriminant Analysis. As explained earlier, different positions rely on agility more often than others, hence why there is a difference of averages between the groups. This has a lower accuracy than the previous model with all variables, thus suggesting that the rest of the variables do help a little bit with categorizing teams, but they are not as important as movement agility.

```
 `stepwise classification', using 660-fold cross-validated correctness rate of method lda'.
660 observations of 4 variables in 3 classes; direction: both
stop criterion: improvement less than 5%.
correctness rate: 0.57727;  in: "movement_agility";  variables (1): movement_agility

 hr.elapsed min.elapsed sec.elapsed
      0.000       0.000       8.121

method      : lda
final model : player_positions ~ movement_agility
<environment: 0x7fc135de9028>

correctness rate = 0.5773
[1] "call"                "method"              "start.variables"
[4] "process"             "model"               "result.pm"
[7] "runtime"             "performance.measure" "formula"
crossval.rate      apparent
    0.5772727            NA
```

Figure 19: Stepwise Discriminant Analysis

Reducing Dimensions by Principal Components Analysis (PCA)

The following variables will be considered for PCA: age, height_cm, weight_kg, overall, potential, value_eur, wage_eur, international reputation, weak foot, skill moves, attacking crossing, attacking finishing, attacking heading accuracy, attacking short passing, attacking volleys, defending marking, defending standing tackle, and defending sliding tackle. PCA does not make assumptions of multivariate distribution. However, value and wage have extreme outliers, so the previous transformations that were used in the MANOVA analysis will be used here. The overall score will also go through a square transformation since it went through one in the MANOVA model. Figure 20 shows a chi-square quantile plot for the data after wage, value, and overall are transformed. The data does not entirely follow a multivariate normal distribution. The data appears to be slightly left skewed since there are a few outliers with large values.

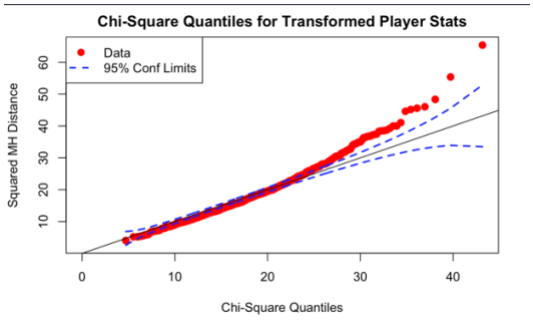


Figure 20: Chi-Square Quantiles for Transformed Data

The correlation plot in Figure 21 shows that many of the variables have a somewhat strong correlation with each other. The attacking statistics somewhat strongly with each other. This is because if someone is good at one attacking skill, there are more likely to be good at another attacking skill. The defensive statistics also correlate strongly with each other for the same reason. Overall, potential, value, and wage also all correlate with each other. This is because players are paid more because they have higher overall scores. Height and weight negatively correlate with attacking statistics.

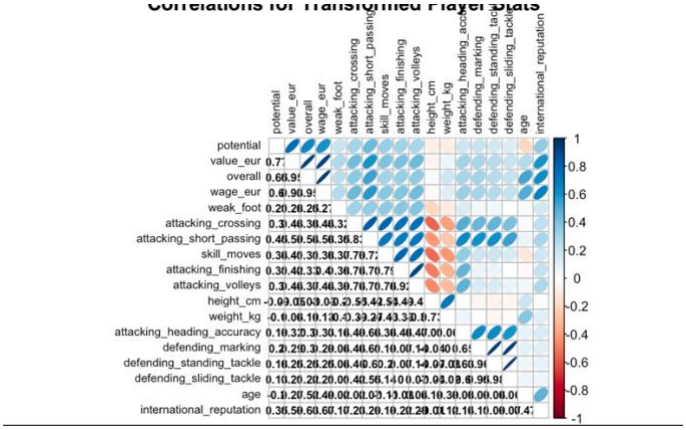


Figure 21: Correlation Matrix

Figure 22 shows that there does appear to be a linear relationship among the variables.

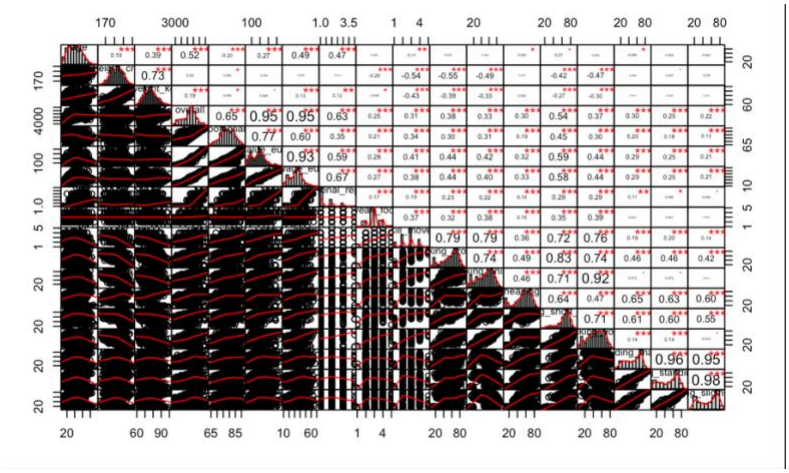


Figure 22: Linear and Correlation Matrix

PCA is performed on the data, and different methods suggest there are only four components. Figure 23 shows that component one explains approximately 49.8% of the total variance, component two explains 17.8% of the total variance, component three explains 15.9% of the total variance, and component four explains 6.9% of the total variance.

```
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
Standard deviation     2.6767728 1.7868205 1.6939325 1.11781601 0.99640061 0.88242520
Proportion of Variance 0.3980618 0.1773738 0.1594115 0.06941737 0.05515634 0.04325968
Cumulative Proportion  0.3980618 0.5754356 0.7348471 0.80426445 0.85942080 0.90268047
                          Comp.7     Comp.8     Comp.9     Comp.10     Comp.11
Standard deviation     0.67215779 0.57551992 0.48853301 0.423733174 0.374669569
Proportion of Variance 0.02509978 0.01840129 0.01325914 0.009974989 0.007798738
Cumulative Proportion  0.92778026 0.94618155 0.95944068 0.969415673 0.977214411
                          Comp.12     Comp.13     Comp.14     Comp.15     Comp.16
Standard deviation     0.341550849 0.333436599 0.256176249 0.218498631 0.199037611
Proportion of Variance 0.006480943 0.006176665 0.003645904 0.002652314 0.002200887
Cumulative Proportion  0.983695355 0.989872020 0.993517923 0.996170237 0.998371125
                          Comp.17      Comp.18
Standard deviation     0.137039418 0.1026642709
Proportion of Variance 0.001043322 0.0005855529
Cumulative Proportion  0.999414447 1.0000000000
```

Figure 23: Summary of PCA

When looking at the eigenvalues of each component, Figure 24 shows that components one, two, three, and four all have eigenvalues greater than 1.

```
Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 Comp.10 Comp.11
  7.17    3.19    2.87    1.25    0.99    0.78    0.45    0.33    0.24    0.18    0.14
Comp.12 Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18
  0.12    0.11    0.07    0.05    0.04    0.02    0.01
```

Figure 24: PCA Eigenvalues

When looking at scree plots, Figure 25 shows that there are elbows at component two and component four.

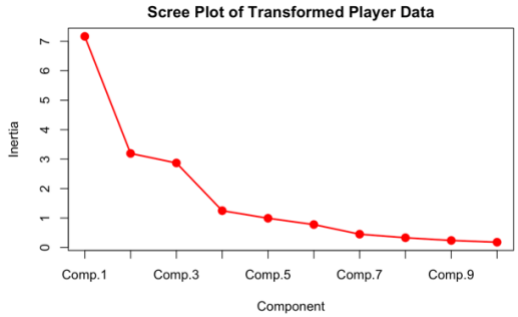


Figure 25: Score Plot of PCA

Figure 26 shows the loadings for each of the components. Component one has high loadings (defined as greater than 0.24) for overall, value, wage, skill moves, and the attacking statistics. Recall earlier that it was explained that many soccer enthusiasts perceive attackers to be the best players because they generally score the goals. Moreover, players have a higher value and wage because they are in demand for how well they play. Players with higher skill moves do a better job at doing difficult soccer moves. Thus, these variables can be grouped together as variables that impact how a player is viewed. Component two includes height, and weight. These variables can be grouped into health. Component three includes the defending statistics. The individual defending statistics of soccer players tend to not vary too much from each other. For example, if one player is really good at doing sliding tackles, they are probably really good at doing standing tackles. Thus, component three is the variables that explain defending. Component four includes age and potential. Younger players have more potential as they are still young. Meanwhile older players have less potential, especially as they become closer to retirement. These two variables describe the growth in a career can be grouped into growth.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
age	0.07	0.32	0.15	0.54	0.37	0.05	0.26	0.17
height_cm	-0.14	0.39	0.00	0.05	-0.47	-0.09	-0.07	-0.16
weight_kg	-0.08	0.41	0.05	0.24	-0.42	-0.05	0.09	-0.43
overall	0.27	0.29	0.23	-0.10	0.08	0.00	0.26	0.11
potential	0.22	0.08	0.16	-0.62	-0.18	-0.05	-0.05	0.05
value_eur	0.29	0.21	0.23	-0.25	-0.01	-0.03	0.15	0.07
wage_eur	0.28	0.25	0.23	-0.05	0.10	-0.02	0.15	0.07
international_reputation	0.18	0.22	0.25	0.08	0.31	0.01	-0.84	-0.21
weak_foot	0.15	-0.10	0.13	0.10	-0.29	0.92	-0.06	0.07
skill_moves	0.27	-0.28	0.06	0.06	-0.06	-0.08	0.10	-0.52
attacking_crossing	0.31	-0.19	-0.07	0.11	0.09	-0.04	0.12	-0.36
attacking_finishing	0.27	-0.27	0.14	0.21	-0.21	-0.19	-0.02	0.11
attacking_heading_accuracy	0.24	0.07	-0.25	0.21	-0.37	-0.22	-0.25	0.51
attacking_short_passing	0.35	-0.07	-0.10	0.03	-0.03	-0.01	0.10	-0.02
attacking_volleys	0.28	-0.23	0.12	0.24	-0.17	-0.16	-0.09	0.08
defending_marking	0.22	0.17	-0.43	-0.05	0.07	0.07	-0.01	-0.02
defending_standing_tackle	0.21	0.15	-0.45	-0.05	0.08	0.09	-0.02	-0.07
defending_sliding_tackle	0.19	0.16	-0.46	-0.05	0.11	0.10	-0.02	-0.09

	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15	Comp.16	Comp.17	Comp.18
age	0.10	0.07	0.07	0.48	0.06	0.01	0.04	0.01	0.03	0.29
height_cm	0.63	0.38	-0.08	0.01	-0.11	0.03	-0.01	0.00	-0.01	0.00
weight_kg	-0.60	-0.18	0.03	0.00	0.01	-0.01	0.01	0.02	0.00	0.01
overall	0.02	-0.01	0.01	0.03	0.06	0.01	-0.05	-0.36	-0.08	-0.74
potential	-0.13	0.09	-0.02	0.63	0.11	-0.03	0.09	0.20	0.01	0.13
value_eur	0.01	-0.04	-0.02	-0.38	0.05	0.04	-0.12	-0.48	0.06	0.57
wage_eur	0.11	-0.13	-0.02	-0.42	0.03	-0.01	0.16	0.73	0.01	-0.03
international_reputation	-0.02	-0.02	-0.01	0.00	-0.08	0.04	-0.03	-0.06	0.00	-0.04
weak_foot	0.04	-0.02	-0.04	0.01	0.03	0.02	0.00	0.00	0.00	0.01
skill_moves	0.38	-0.42	0.43	0.13	0.18	-0.02	-0.01	-0.04	0.00	0.02
attacking_crossing	0.03	0.06	-0.82	0.07	0.08	-0.05	-0.01	-0.02	0.02	0.00
attacking_finishing	-0.12	0.22	0.07	0.00	-0.07	0.79	0.03	0.04	-0.04	-0.02
attacking_heading_accuracy	0.10	-0.51	-0.17	0.08	0.09	-0.12	0.04	-0.06	0.01	-0.01
attacking_short_passing	-0.05	0.04	0.11	0.07	-0.88	-0.24	-0.01	0.01	-0.02	0.03
attacking_volleys	-0.16	0.52	0.21	-0.12	0.31	-0.52	0.00	0.01	-0.03	-0.01
defending_marking	-0.04	0.09	0.10	0.02	0.13	0.10	-0.80	0.20	-0.11	0.01
defending_standing_tackle	-0.04	0.15	0.14	-0.04	0.08	0.09	0.29	-0.06	0.75	-0.08
defending_sliding_tackle	-0.02	0.10	0.09	-0.06	0.11	0.08	0.47	-0.09	-0.64	0.08

Figure 26: PCA Components

Figure 27 shows the PC score plots for each pair of components, colored by the position of the player. Defenders are in black text, forwards are in red text, and midfielders are in the green text. The PC Score Plot for components one and two show two clumps, one for midfielders and forwards, and another for defenders. This suggests that defenders tend to weigh more and are taller, which makes sense given that they tackle other players to reclaim possession of the ball. Moreover, they score lower in the component that measures reputation, which lines up with our previous MANOVA model. The PC score plot for component one and two has two clumps. The first clump includes some defenders, and the second clump includes the rest of the defenders, along with the forwards and midfielders. There appears to be a difference of statistics even within the defenders. This may be because defenders include goalies, which may have significantly lower attacking statistics than the rest of the defenders because their training is different from the rest of the players. The PC score plot for component one and three shows these same clumps. The PC score plot for component one and component four does not seem to have any clumps, although defenders to tend to have lower scores on the x-axis. This may be because component four includes age, and it is likely it is normally distributed among the positions. The PC score plot for component two and three does not show any distinct clumps, although defenders tend to score higher on the y-axis. This makes sense as the third component focuses on defending statistics. The PC score plot for component two and four does not show any distinct clusterings. However, it is interesting to point out that many players fall outside of the 95% confidence interval when compared to the other score plots. Component two and four include height, weight, and age suggesting that these variables are on their own are not accurate estimators. The PC score plot for component three and four does not show distinct clumps.

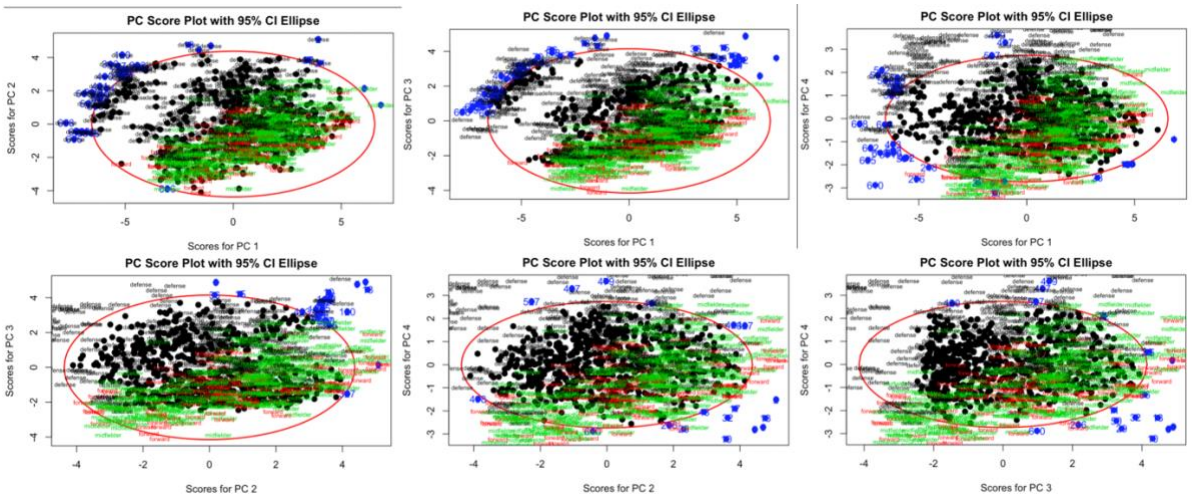


Figure 27: PC Score Plots for Each Pair of Components

The biplots for component one and two and component one and three are shown in Figure 28. It is interesting to note that the distinct clumps that were seen in the PC score plots in Figure 27 are also visible in the biplots. As mentioned earlier, it appears that goalies tend to have lower statistics in component one and higher statistics in component two and three when compared to the rest of the players.

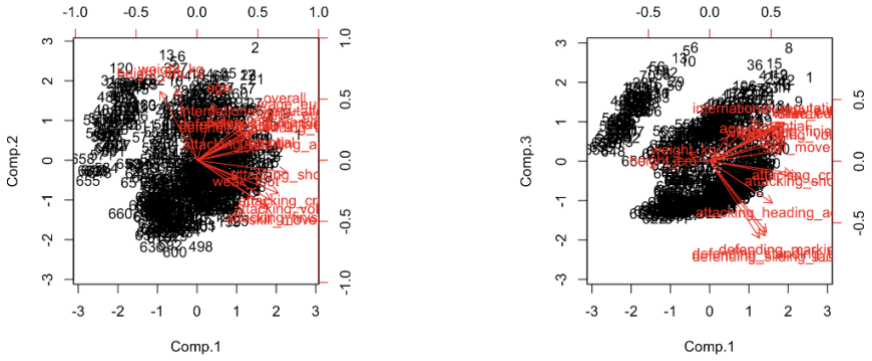
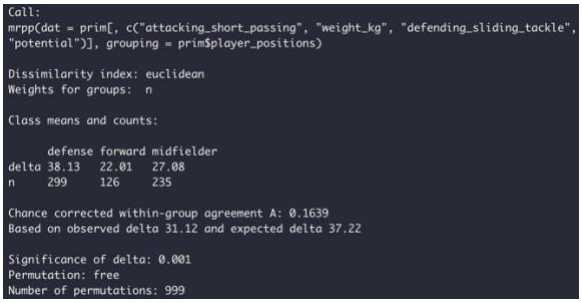


Figure 28: Biplot for Component 1 and 2 and Biplot for Component 1 and 3

Parallel Analysis is not appropriate to use since the data is not multivariate normally distributed. As mentioned earlier, the correlation plot in Figure 21 shows that many of the variables are correlated to each other. Moreover, there is a lot of data, thus these components are significant and did not happen by chance. The highest variable in each of the four components

will be used to perform Multiple Response Permutation Procedure (MRPP). MRPP is a test that determines if there is a difference between groups. Unlike MANOVA, it has no assumptions.



```
Call:
mrpp(dat = prim[, c("attacking_short_passing", "weight_kg", "defending_sliding_tackle",
"potential")], grouping = prim$player_positions)

Dissimilarity index: euclidean
Weights for groups:  n

Class means and counts:

      defense forward midfielder
delta 38.13   22.01   27.08
n     299     126     235

Chance corrected within-group agreement A: 0.1639
Based on observed delta 31.12 and expected delta 37.22

Significance of delta: 0.001
Permutation: free
Number of permutations: 999
```

Figure 29: MRPP on Attacking Short Passing, Weight, Defending Sliding Tackle, and Potential

Figure 29 reports the p-value of this test is 0.001, thus there is evidence that attacking short passing, weight, defending sliding tackle, and potential is different across the positions. This lines up with our analysis of the PC score plots.

Conclusions and Discussion

A one-way MANOVA model was used to determine if there are any differences in the means of overall, wage, mentality composure, and movement agility by player positions. The overall score was the only dependent variable that did not have a significant impact on the position. The significant difference in wage suggests that some positions are paid more than others, with Figure 6 revealing that it is the forwards since they have the highest average. The significant difference in mentality composure suggests that some positions are more demanding than others in terms of performing under pressure. The difference in movement agility suggests that some positions require faster response to speed and direction than others. Because of these results, discriminant analysis was performed to see if the positions can be categorized.

Despite having significant differences in mean by wage, mentality composure, and movement agility, the models from the linear discriminant analysis struggled to categorize players. It only had an accuracy of 63% with the raw results, and an accuracy of 62% with cross-validated results, suggesting the model is overfit. Seeing how low the accuracy is, this suggests there may be other variables that were not considered that are contributing to these differences. PCA was performed to see if there are better grouping of all the variables.

The PCA grouped 18 different variables into four components: reputation, health, defending, and growth. The variable with the highest loading in each component was selected to perform a MRPP test on it to see if there is a significant difference of these four variables across positions. The results of the PCA suggest that the original variable of overall, mentality, and movement agility may not have been the best variables to use categorizing players. Attacking and defending statistics make more sense, as forwards are more likely to be better at attacking and defenders are more likely to be better at defending.

The models from the Linear Discriminant Analysis show some evidence of it being overfit since the cross-validated results had a lower accuracy than the raw results. This suggests that using overall, wage, movement agility, and mentality composure may not be the best variables for predictions, and moreover, they are specific to this season of the Premier league and the model probably cannot be used to categorize players in other seasons or clubs. This suggests that the differences in movement agility and mentality composure was unique to the 2019 – 2020 season of the Premier League.

It is natural to believe that there are differences among the three positions, given that different skills are required for each one. It is quite possible that there are not just specific statistics that vary significantly, rather it is the interaction of statistics that vary significantly across the positions. Overall, our analysis supports that there are differences among the three positions, and given that these differences may be difficult to identify, more complex tools and methods will need to be used in further analysis.

Points for Further Analysis

Although the one-way MANOVA showed that there is a significant difference in means for wage, mentality composure, and movement agility by position, it did not control for other variables that may have contributed to this. For example, it is reasonable to have a prior belief that age impacts movement agility and mentality composure. If all else is equal, an older player will probably have a lower movement agility than a younger one. Testing can be done to see if this holds. If this ends up being true, then it is worth considering if age should be controlled for and perhaps a one-way MANCOVA model that treats age as an interaction with some of the responding variables could be used to test this.

In a broader sense, many of the variables tend to correlate with each other, which could create some issues of multilinearity. An in-depth analysis of Principal Component Regression can be performed in order to develop a model that predicts positions based on the relevant components. Moreover, models that consider interactions amongst variables could be utilized as well to spot the complex relationships among the variables. It is possible that the difference of the positions is that easily described by simple statistics, which could explain the overfitting of our Linear Discriminant models. Further analysis can be done to weigh the significance of these interactions, and perhaps the new models developed can be applied to other leagues as well. In conclusion, it is possible that a more complex model that can categorize players based on not only variables but their interactions as well.