MULTIVARITE ANALYSIS OF NBA PAY-OFFS OF 2021-22 SEASON

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR COURSE
STAT 467 – MULTIVARIATE ANALYSIS
DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

Bilgehan Aydoğdu - 2428886
Beyza Demet - 2428985

January 2024

ABSTRACT

This research was made for the multivariate analysis of player statistics from the NBA Play-offs in 2021-22 by using statistical methods to identify trends and connections in the dataset. The main objectives of the analysis include exploring mean differences across player performance metrics, conducting one-way and two-way MANOVA to understand the impact of categorical variables, and applying principal component analysis (PCA) and factor analysis to identify latent structures in the data. For player classification, linear discriminant analysis (LDA) offers insights into the discriminative strength of the chosen variables. Furthermore, player groups based on their statistical profiles are revealed with cluster analysis, providing a more detailed understanding of player similarities and differences. Lastly, we grouped players based on their stats using cluster analysis. Our observations do not only help us understand how players and teams performed in the NBA Playoffs but also show how multivariate statistical analysis can teach us more about the game dynamics. This study might be a useful guide for future basketball analysis and how to look at player performances in the NBA.

1. Introduction

NBA play-offs contains 8 teams in each East and West conference, which means 16 teams in total. An elimination model is used in the NBA Playoffs, which begin with regular season. Through the first round, conference semifinals, and conference finals, teams play each other in best-of-seven series. In the NBA Finals, the winners of each conference finals face with each other.

In the NBA Playoffs, how well players perform really matters. Individual player performances become critical components as the level of competition rises, influencing not just the success of the team but also the quality of basketball. Their team's success or failure is greatly influenced by their statistics, such as assists and points received.

Throughout the 2021-2022 campaign, the NBA Play-offs showcased intense competition among sixteen top NBA teams and their players. Beyond the excitement of the matches, there is a wealth of statistical information that offers important insights into team and player performances. For all the teams which compete in NBA Play-offs, performance on the field depends on the players' influence on the match. In this research, we mainly focused on analysis of player statistics that might affect the game. To measure these influence various statistical methods were conducted such as multivariate hypothesis testing to check the difference of means of the response variables. One-way and Two-way MANOVA were conducted to measure the effects of categorical variables on response variables. Moreover, many statistical analyses were performed to measure the effects the players' stats on the game such as Principal Component Analysis and Regression, Factor Analysis, Linear Discriminant Analysis and Cluster Analysis.

1.1. Data description

The 2021-2022 NBA Players Stats Dataset consists of 217 observations of NBA basketball players, encompassing 31 variables, including 26 numerical and 5 categorical variables.

- Player: Player's name
- Pos: Position
- Age: Player's age
- Tm: Team of the player
- G: Games played
- GS: Games started
- MP: Minutes played per game
- FG: Field goals per game
- FGA: Field goal attempts per game
- FG%: Field goal percentage
- 3P: 3-point field goals per game
- 3PA: 3-point field goal attempts per game
- 3P%: 3-point field goal percentage
- 2P: 2-point field goals per game
- 2PA: 2-point field goal attempts per game
- 2P%: 2-point field goal percentage
- eFG%: Effective field goal percentage
- FT: Free throws per game
- FTA: Free throw attempts per game
- FT%: Free throw percentage
- ORB: Offensive rebounds per game
- DRB: Defensive rebounds per game
- TRB: Total rebounds per game
- AST: Assists per game
- STL: Steals per game
- BLK: Blocks per game
- TOV: Turnovers per game
- PF: Personal fouls per game
- PTS: Points per game
- AGE_LEVEL: Age levels of players (Rookie, Prime, Mature,Old)
- CONFERENCE: conference of the team (East-West)

This Data contains 2021-22 NBA play-offs player statistics per game.

1.2. Research questions

- Is there any difference between players' average expected Field Goal and Total rebound of players on season and game?
- Is there a difference between the player's age level in terms of player's Field Goal and Total Rebound stats?
- Is there a difference between the player's age level and teams in terms of player's Field Goal and Total Rebound stats?
- Is it possible to classify the players' Age, G, GS, TRB, AST and PTS variables according to their conference East and West?

1.3. Aim of the Study

The purpose of this study is to determine whether there is a statistically significant correlation between each player's performance on an individual basis and the team's overall performance. In addition to helping to better understand individual player performances, the results will give coaches, commentators, and basketball fans important information with which to base their judgment calls, plan of attack, and projections for potential playoff scenarios.

2. Methodology/Analysis

In the exploratory data analysis part, First, the "ryston" method is used to check normality. Square root transformation of all data is also used to make a distribution closer to normal. In addition, chi-square plots, normal Q-Q plots, histograms, correlation plot, and scatter plot are used. In the inferences about a mean vector part, one- population hypothesis testing is conducted to estimate the mean. Hotelling's one-sample T2-test is applied. In this test, FG and TRB variables are taken as responses.

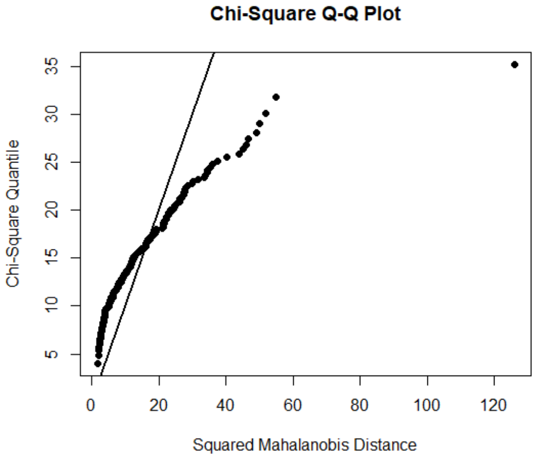
In the comparisons of several multivariate means part, MANOVA assumptions are checked. It was checked whether the dependent variables conformed to multivariate normal distribution within each group. The Levene test was performed to test the homogeneity of variance. It is assumed that the relationships between the independent and dependent variables are linear. After the assumptions were satisfied, the first ONE-WAY MANOVA was conducted. In this analysis, FG (Field Goal) and DRB (Defensive Rebound) were taken as response variables, and age level was taken as a categorical variable. Secondly, a TWO-WAY MANOVA was conducted. In two-way MANOVA, we select response variables as FG (Field Goal) and DRB (Defensive Rebound) and factors as Age Level and Team.

In the principal components analysis, scatter plot matrix, correlation plots, scree plot, polar plot and some figures are used to visualize the data. In the principal components' regression part, A Principal Component Regression model was established between FG taken as the response variable and predictors. For the factor analysis and factor rotation part, correlation plot and parallel analysis scree plot is used. Also, the Bartletts and KMO test were conducted to check whether factor analysis was applied. In the Linear Discriminant Analysis and Classification part, events showing the analysis of players divided into East and West are included. In addition, to evaluate the performance of the model, training and test performances were conducted. As a last part, the clustering part, K means clustering is used. The distribution of the variables was visualized by displaying the scatterplot matrix of the data. After the data was standardized, it was clustered, and a plot showing the sum of squares within the group was drawn.

3.Results and Findings

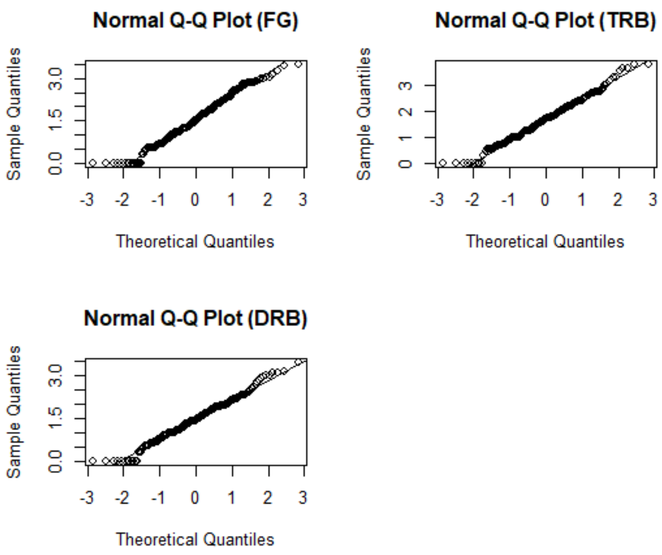
3.1 Exploratory Data Analysis

To check the multivariate normality, a (Q-Q) plot and normality testing with the "ryston"method are used.

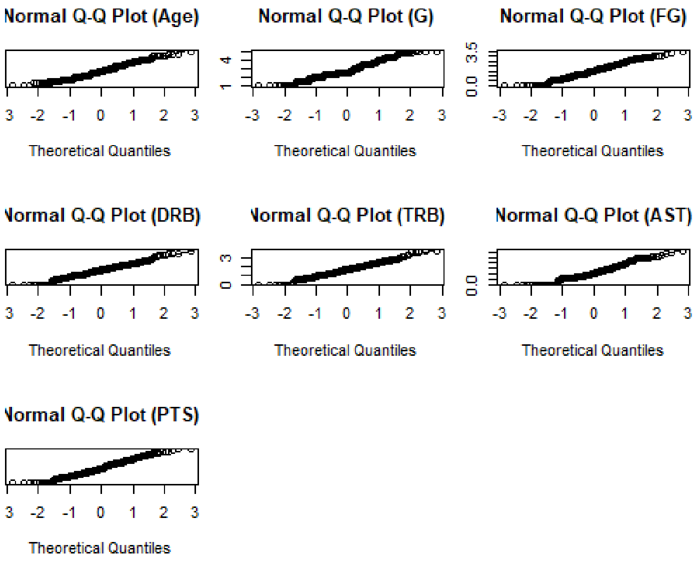


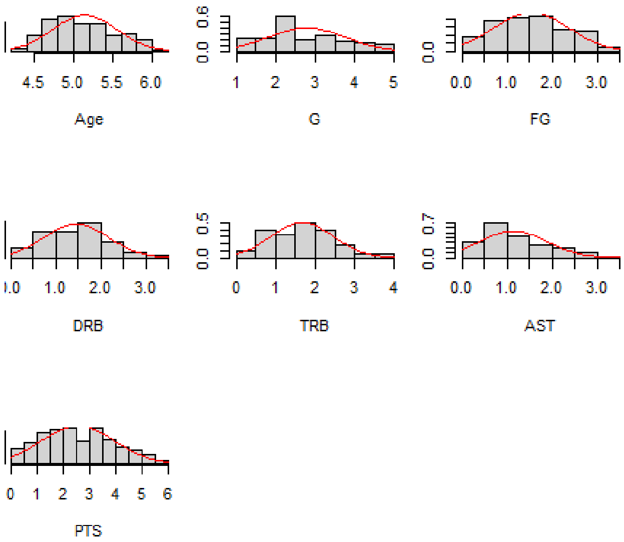
The Q-Q plot which is given above shows as a graphical representation, aiding in evaluating the normality assumption, and, likely, the data is not normally distributed. In addition, as a result of the normality test (shown in TABLE 1), it is seen that the data is not normally distributed. When the multivariate normality is checked, besides the univariate non-normality of three response variables, it can be observed that non-normality in multivariate non-normality for all variables, as shown in the TABLE 2.

After applying the square root transformation, three variables meet the requirements for univariate normality. After using square root transformation, it can be seen that the "FG," "DRB," and "TRB" variables satisfy the normality assumption. (shown in TABLE 3). Then, multivariate nomality checked but it can not be satisfied. The following is a normal Q-Q plot of 3 response variables, which satisfy univariate normality after necessary transformations.



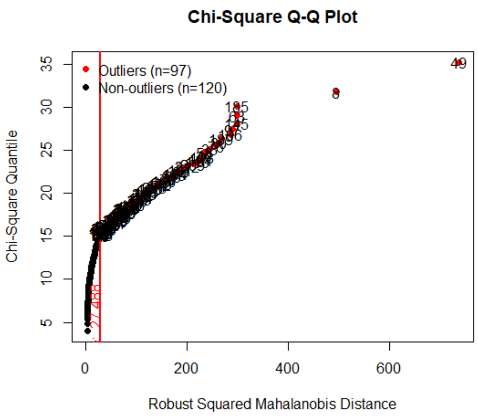
From now on, not every variable in the data will be used, "Player", "Pos", "Age", "Tm", "G", "FG", "DRB", "TRB", "AST", "PTS" variables will be subset and analyzed. It will be done through these. The following plots represent a normal Q-Q plot of the variables and distributions for univariate normality. that will be used in the analysis.

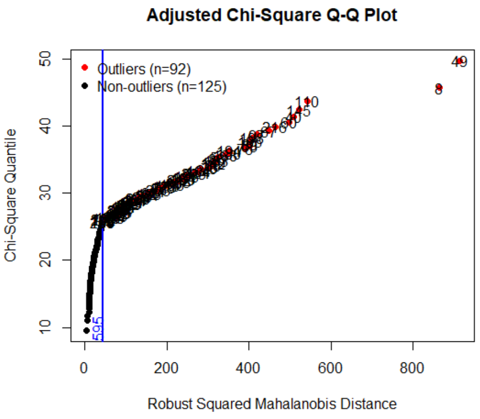




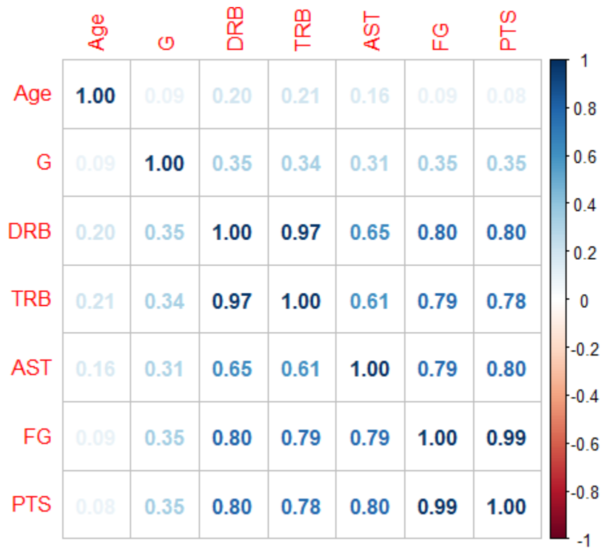
Looking at the normal Q-Q graphs and histogram, it can be said that "FG," "DRB," and "TRB" meet the univariate normality condition, while the other variables do not. After checking univariate normality, bivariate normality can be prevented for each pair of variables.

Before starting the multivariate analysis, robust Mahalanobis distances and adjusted Mahalanobis distance methods are applied to check whether the data have multivariate outliers.



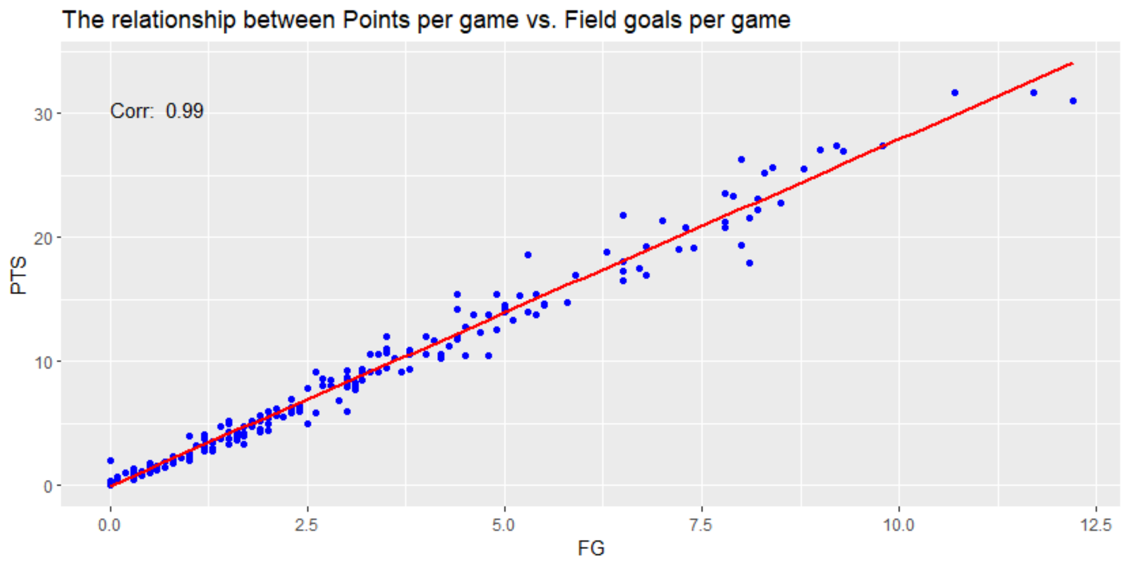


As seen, there are 97 outliers observations proved by Mahalanobis Distance while 92 outliers observations confirmed by Adjusted Mahalanobis Distance in this dataset.
A correlation Plot was created to interpret the direction and strength of the relationship between the variables in the data.



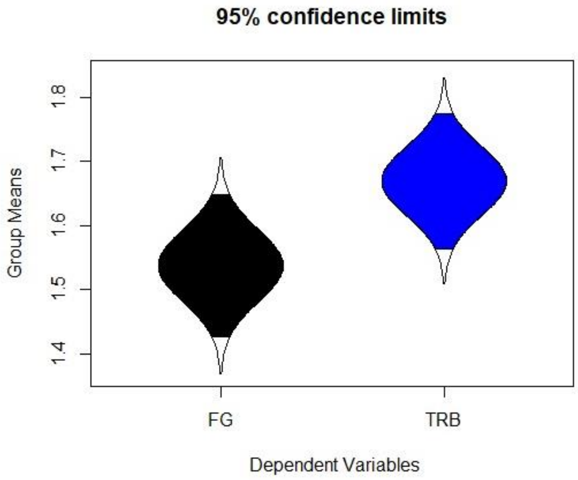
	Age	G	DRB	TRB	AST	FG	PTS
Age	1.00	0.09	0.20	0.21	0.16	0.09	0.08
G	0.09	1.00	0.35	0.34	0.31	0.35	0.35
DRB	0.20	0.35	1.00	0.97	0.65	0.80	0.80
TRB	0.21	0.34	0.97	1.00	0.61	0.79	0.78
AST	0.16	0.31	0.65	0.61	1.00	0.79	0.80
FG	0.09	0.35	0.80	0.79	0.79	1.00	0.99
PTS	0.08	0.35	0.80	0.78	0.80	0.99	1.00

When the correlation graph is examined, it is seen that the “FG” and “PTS” variables have a high positive correlation with variables other than “Age” and “G”. To make a general comment, it can be concluded that “Age” and “G” are more independent in the data, and the remaining variables are related. It can be noted that there is a very strong correlation between “FG” and “PTS”. This can be seen more clearly in the plot below.



3.2. Inferences About a Mean Vector

In this one population mean hypothesis testing, there are two responses that wanted to be model which are "FG" and "TRB". It was previously mentioned that these variables provide normality assumptions. It is desired to test the null hypothesis that the observations come from the mean vector of the response variables by using Hotelling's T-squared test. For better visualization, the following graph was represented.



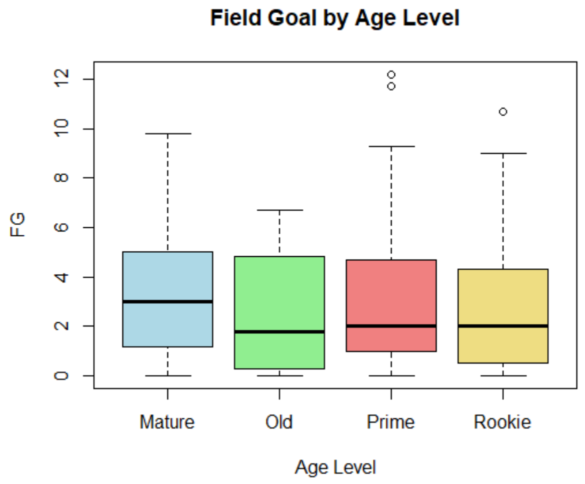
Confidence intervals are given at TABLE 4 for each variable on the side. Notice that the individual 95% CI for "FG" and "TRB" do not include mu0. Hotelling's T-squared test output is given at TABLE 5.

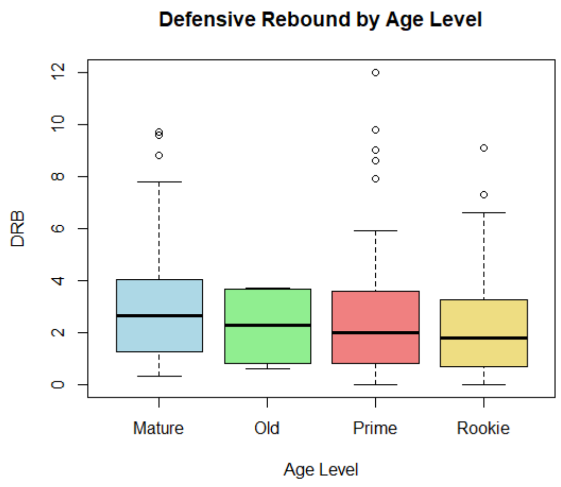
Since the p value< alpha, the null hypothesis is rejected. Therefore, there is not enough evidence to conclude that the mean vector equals (15,150).When the confidence ellipse for mean vector differences below is examined, it can be understood that the null hypothesis will be rejected. This interpretation can be deduced since the confidence interval does not include mu values. (shown in Figure 1)

3.3 COMPARISONS OF SEVERAL MULTIVARIATE MEANS

3.3.1 ONE-WAY MANOVA

The response variables FG (Field Goal) and DRB (Defensive Rebound) were selected from data by using R. The following box-plots of response variables were created according to age levels to visualize whether there is any difference.





The necessary transformations are conducted in Table.6. And univariate normality is satisfied for response variables.

After checking normality assumptions, the one-way MANOVA test was conducted by using R. As it seen in Table 7, summary of the test indicates that We reject the null hypothesis. Therefore, we are 95% confident that at least one age level is significantly different than others since p < a0.05. We need to hold a post-hoc analysis to see which one causes the difference. To do so,

The results of analysis (Table 8) demonstrate that there is a highly significant difference in the TRB variable between age groups. We are unable to draw a comparable conclusion for FG, though.

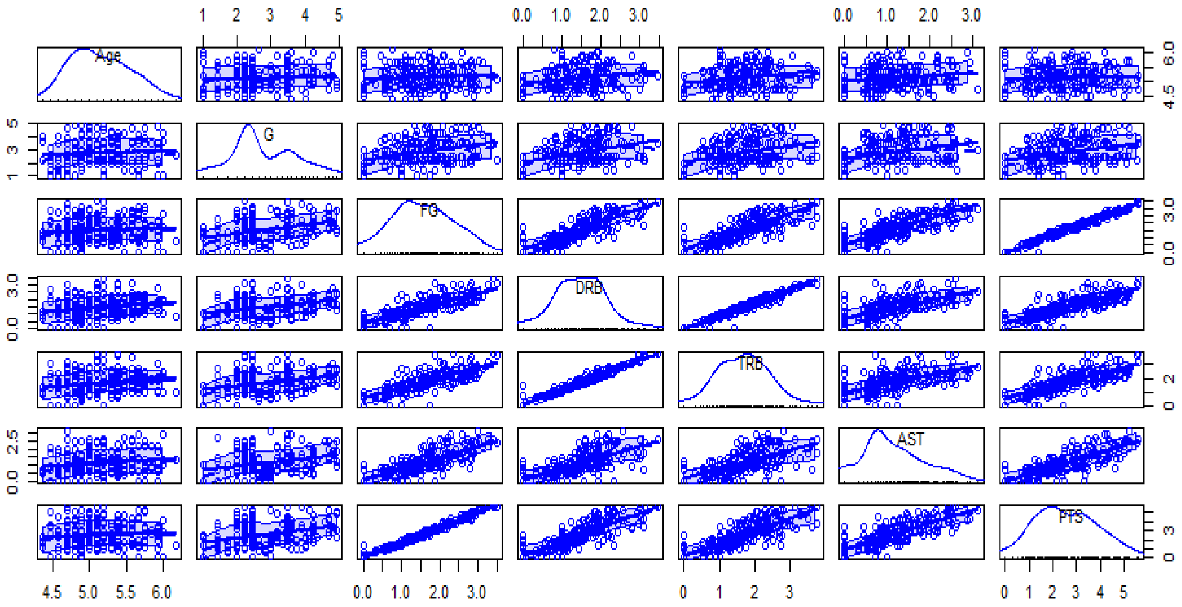
3.3.2 TWO-WAY MANOVA

In two-way MANOVA, we select response variables as FG (Field Goal) and DRB (Defensive Rebound) and factors as Age Level and Team. Multivariate normality tested for response variable but multivariate normality is not satisfied by using transformations such as log, box-cox and sqrt transformation as it seen in Table 9. After applying the necessary transformations are conducted in Table xx. And univariate normality is satisfied for response variables (shown in Table 10).

After checking normality assumptions, the two-way MANOVA test was conducted by using R. As shown in Table 11, summary of the test indicates that the factor levels clearly since the interaction variable is not significant. For factor1 (Age Level), there is little difference between the mean values of FG and DRB. Similarly, factor 2 (Team)'s mean values remain unchanged.
Interaction plots of Field Goals and Defensive Rebound are given at Figure 2 and Figure 3.

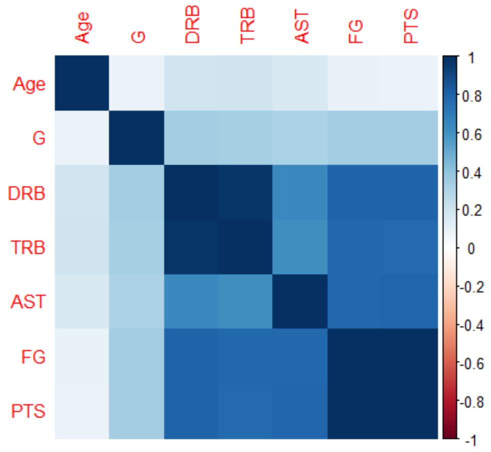
3.4. Principal Component Analysis and Regression

3.4.1. Principal Component Analysis



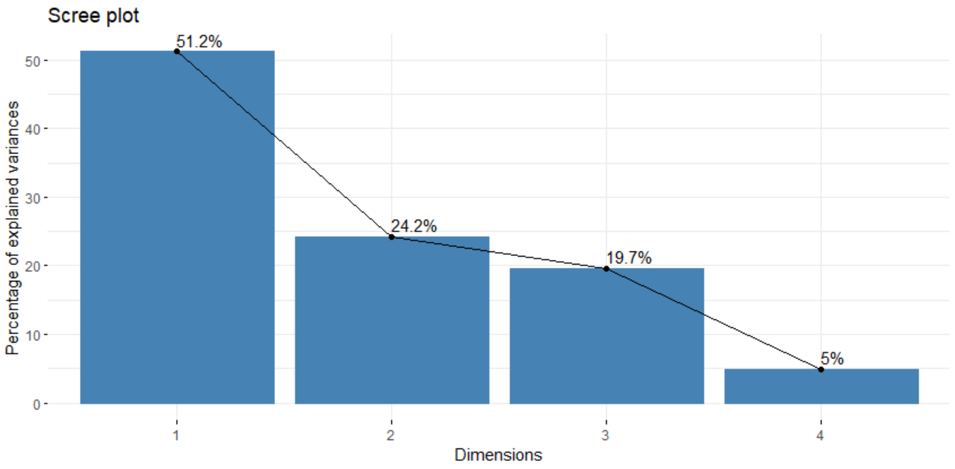
It can be seen a matrix of scatterplots for the numeric variables in the data. In this plot, the variables Player's age, Games played, Field goals per game, Defensive rebounds per game, Total rebounds per game, Assists per game, Points per game are focused on. When the histograms are examined, it is clearly seen that the variables Field goals per game, Defensive rebounds per game, and Total rebounds per game have a normal distribution. When the histograms of other variables are examined, it cannot be said that they have a normal distribution. It is seen that there is a strong correlation among the Assists per game and Points per game in the data which is very suitable condition for PCA.

A correlation matrix can be visualized to represent the correlation among the numerical variables in the data set as an alternative way.



When the correlation plot is examined, it can be said that “the player's age” and “the games played” are independent of each other and that there is very little correlation between these variables and other variables. As mentioned in the previous plot, it is clear that there is a very strong correlation between “field goals per game” and “Points per game”. There is a moderate correlation for these two variables with variables other than the player's age and the games played.

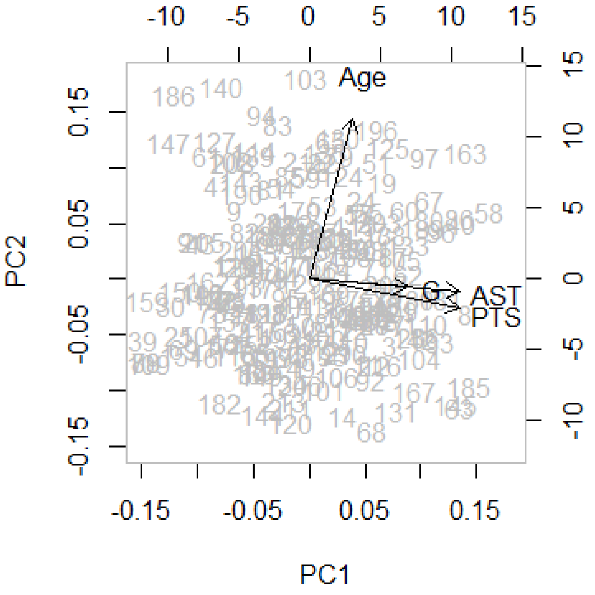
The result of importance of components (shown in Table 12) gives the standard deviation, the proportion of variance explained by each principal component, and the cumulative proportion of variance explained. For example, we can see that the first three components explain the 95.04% variability in data which is very good.
To decide how many components should be included in the analysis, one should look at the contribution of each component to the cumulative ratio. Scree plots can be used in a visual way.



When the scree plot is examined, three components seem OK. It can be seen that the first three components explain almost 95.04% of the variability, which is very good, as stated above. The analysis can be continued by removing the first three components.

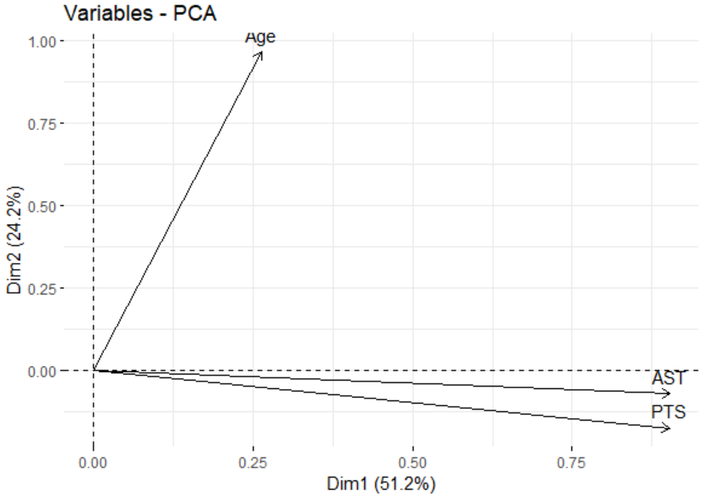
As a result of the components output (shown in Table 13), the first component is positively correlated with "AST" and "PTS". The component is most correlated with "AST" at 0.9033 and could be considered as primarily a measure of "AST". The second component is positively correlated with "AST" and "PTS". The component is most associated with "Age" at 0.9639 and could be considered primarily a measure of "Age". The third component is also negatively correlated with "G", and could be regarded as mainly a measure of "G".

After removing the first three components, components must be linearly independent. As it can be seen, all components are linearly independent in the Figure 4.
A binary plot is created to interpret how strongly the loading of a particular variable contributes to a particular principal component.



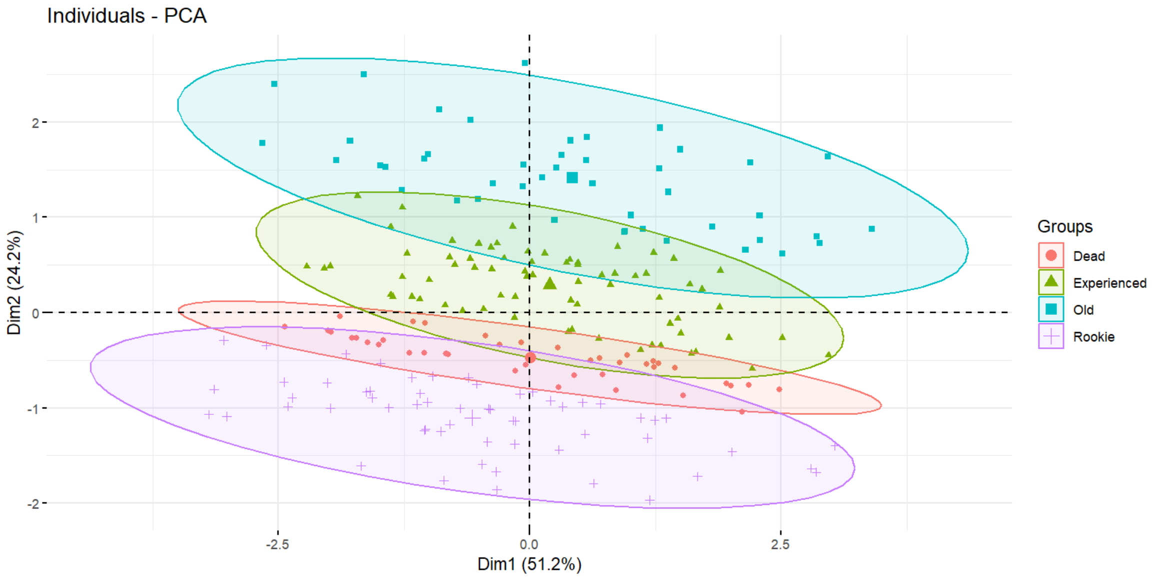
In this bipolar plot, each project value shows how much weight they have on that PC. In this example, it can be said that "G", "AST" and "PTS" affect PC1, while "Age" affects PC2. Vectors that point in the same direction correspond to similar response profiles. The fact that "AST " and "G" are in the same direction may indicate that the variables in the data set are similarly affected by these two factors. Additionally, it may show a high positive correlation between these factors. The fact that the "age" variable is far away from these factors may indicate that age is more independent or less related to these factors.

Figure 5 shows the contributions of variables to the principal components. In addition, in the following table, the ones being the first three variables with high contribution were shown. AST, PTS, and Age are the first three variables having the highest contribution to the first two components.



The visual representation of how individual observations are positioned in the PCA given in Figure 6. It can be observed which components are suitable for explaining the cases.

The observations have been classified with respect to one categorical variable which is players' conference level. The visual representation of this has been given in the following plot.

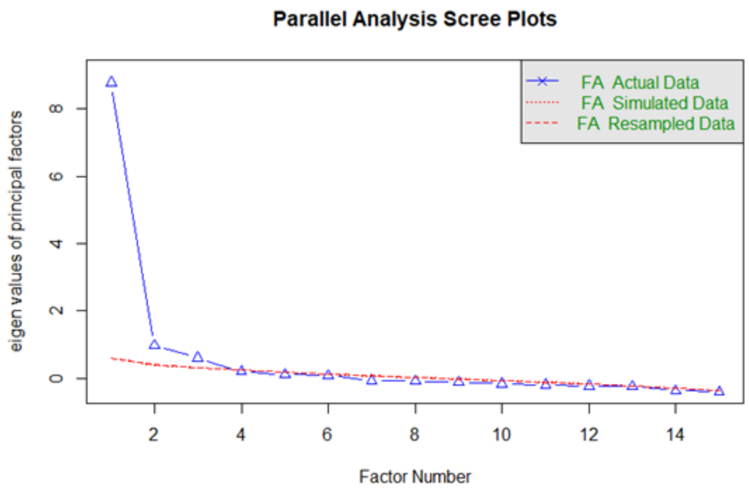


3.4.2. Principle Component Regression:

According to the output given in Table 14, the model is significant. The Multiple R-squared value is 0.8302, indicating the proportion of variance in the response variable (FG) that is explained by the predictors. Also, the model suggests that the predictors PC1, PC2, and PC3 are significantly associated with the response variable (FG), and the highest variance belongs to the third component. The overall model fits the data well. With MSE, the performance of the model on the data can be checked. Calculating this value yields an MSE of approximately 1.231851, indicating a reasonable level of performance.

3.5 Factor Analysis

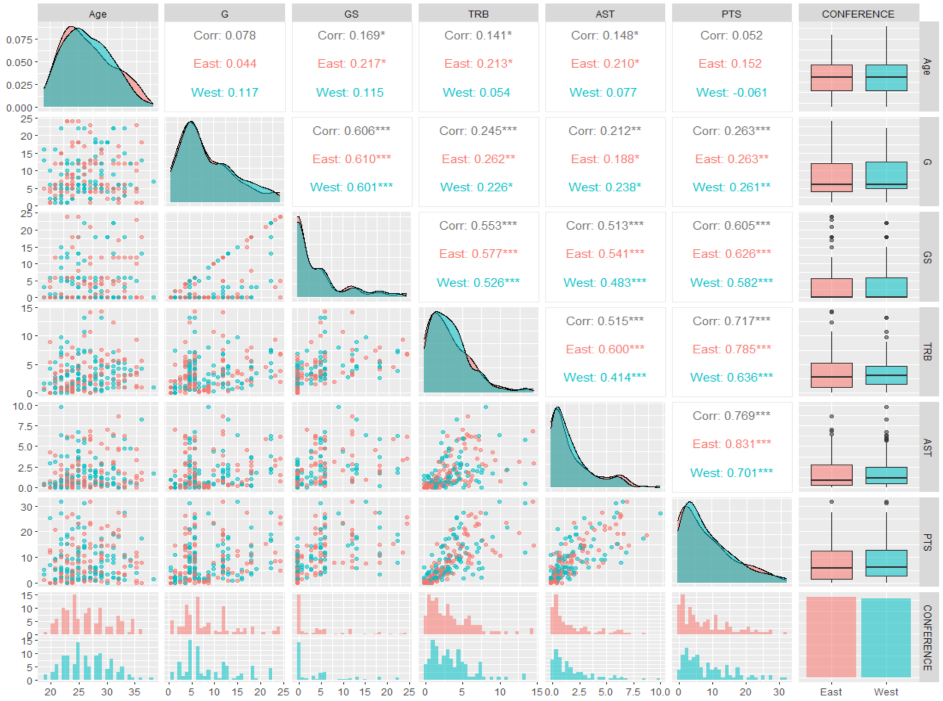
As the first step, the variables that we want to focus on were selected from dataset. Then, the correlation plot (shown in Figure 7) examined and the factorability of the variables in the dataset was tested by using R. The independent variables were selected and KMO test was conducted. According to the result of KMO test (shown in Table 15), the MSA value (0.83) is greater than 0.5. After that, Bartletts Test was conducted (shown in Table), and factor analysis may be performed for this dataset.



However, the p-value doesn't provide sufficient evidence to support the hypothesis, even after attempting several adjustments with the factor analysis function. This indicates that the variables in the dataset do not group together in a similar way because they are so dissimilar from one another. We are unable to conclude that healthy formation of these groups, it might be explained by any other unidentified reason.

3.6 Linear Discriminant Analysis and Classification

The Linear Discriminant Analysis was conducted to measure differences between East and West Conferences. The players are separated East and West according to the region that they play. 6 variables were used in this analysis such as Age, Games, Games Started, Total Rebound, Assists and Points.

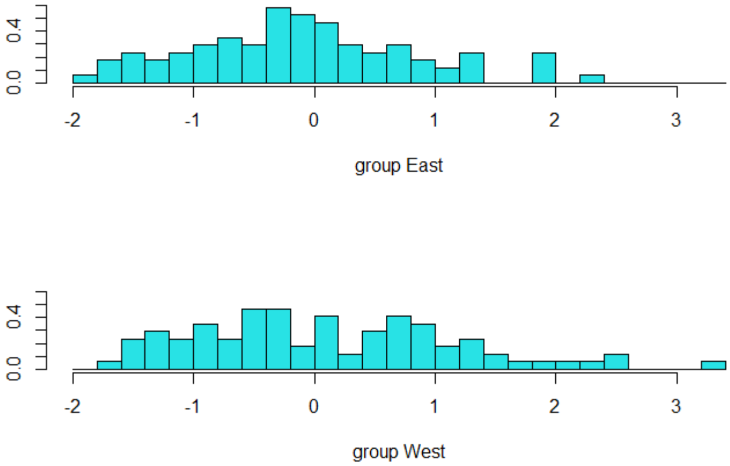


14

The LDA output (shown in Table 17) indicates that players from the East Conference account for 50% of the training observations. Moreover, the model was conducted by using the same output.

$$-0.015 \ast Age + 0.203 * G - 0.126 * GS - 0.0002 * TRB - 0.11 * AST + 0.103 \ast PTS.$$

To help visualize the model better, the following plot was provided.

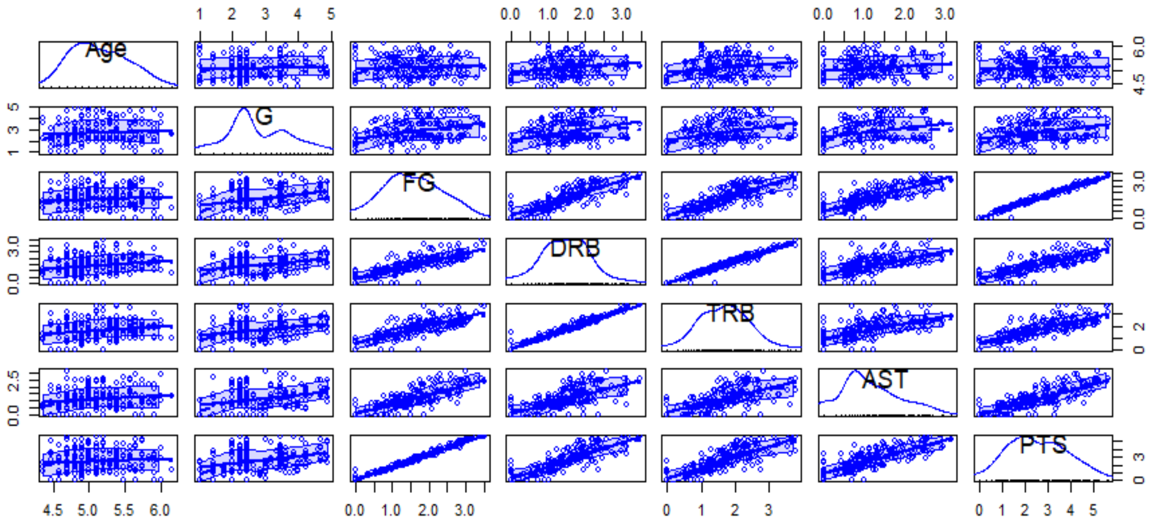


The partition plot was used to provide a better visualization of the classification of each and every combination in the training data set. (shown in Figure 8)

To evaluate the performance of the model train and test performances were conducted. As a result of train performance, the model correctly classifies the players with 0.54 probability for the training data. As a result of the test performance, the model correctly classifies the players with 0.49 probability for the test data.

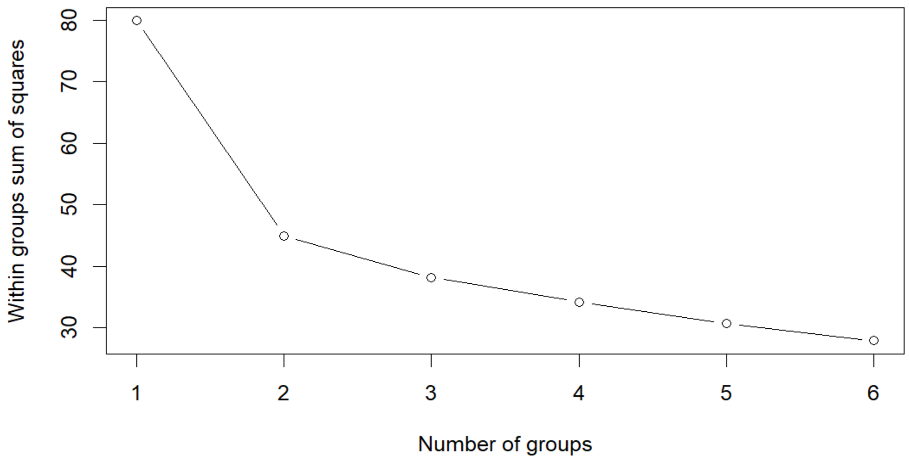
3.7. Cluster Analysis

K-means clustering can be used for the dataset. First, look at the scatterplot matrix of the data for some numerical variables.



When performing multivariate analysis or clustering based on distances, standardizing the data makes interpreting scatterplots easier. This is particularly useful when variables have different units or measurement scales. Thus, we normalize each variable by its range since variances are too high. After such standardization, the variances became as shown in Table 18.

Since the variances of the standardized data are very similar, we can progress with clustering the data. First, we plot the within-groups sum of squares for one- to six-group solutions to see if we can get any indication of the number of groups.



In the above plot, the only "elbow" in the plot occurs for two groups, thus we will look at the two-group solution. The group means for two groups are computed in the output given in Table 19 and the cluster number for each state is given in Table 20.

4. Discussion/Conclusion

The primary goal of this study is to provide readers with a helpful guide by evaluating player statistics from the NBA playoffs in the 2021–2022 season. At the beginning of the examinations, it was found that the data was multivariate and not normal. Many transformations methods were performed to solve this problem such as Log, Box-Cox and square root transformations. As a result of these transformations, some response variables were made univariate. However, despite these variables, many variables did not meet the normality assumptions.

For inference about a mean vector, the variables that satisfy the normality conditions were selected. To purpose of this question was comparing the average means. So, testing the null hypothesis that observations come from the mean vector of response variables found a p-value <alpha. Therefore, the null hypothesis is rejected. Thus, it can be concluded that there is not enough evidence to conclude that the mean vector equals (15,150).

To compare several multivariate means, in one-way MANOVA, one categorical variable called as “Age Level” added to the dataset and the multivariate means compared according to levels of this categorical variable. As a result of the one-way MANOVA, we may say that between age groups, there is a statistically significant variation in the TRB variable. For FG, however, we are unable to reach a similar conclusion.

In two-way MANOVA, one more categorical variable has been added to the previous part. This categorical variable is called “Conference” levels, demonstrating which player comes from which conference, East or West. Multivariate means of FG and DRB variables were compared according to the factor levels of Age Level and Conference. As a result of this analysis, we may conclude that there is not much of a difference between the mean values of FG and DRB for component 1 (Age Level). The mean values of component 2 (Team) also stay the same.

For factor analysis, according to the correlation plot, the variables have high correlation were selected. The KMO and Barlett tests were performed to check if the data is appropriate to do factor analysis or not. However, in every possible scenario, the p-value does not satisfy the condition. As conclusion, this suggests that there is no discernible pattern among the variables in the dataset.

For Linear Discriminant Analysis and Classification, the players are divided into East and West groups based on the regions in which they participate. This analysis made use of six variables: age, games played, games started, total rebound, assists, and points. As a result, it is observed that the 50% of training observations comes from East Conference.

For prıncıpal components, AST, PTS, and Age were the first three variables that contributed the most to the first two components. In addition, the prıncıpal components regressıon model suggests that the predictors PC1, PC2, and PC3 are significantly associated with the response variable (FG). In Cluster analysis, K-means clustering was performed, since variances are very high, we normalize each variable using its range.

To conclude, many healthy and unhealthy results were obtained. Since the dataset has not been distributed normally, it caused some problems in many tests and results. We made many transformations to make this data normal. Then, we used the normalized variables and try to reach reliable results. As researchers, this project had many useful insights about multivariate analysis.

References

[1] ESPN. (2022). 2022 NBA Playoffs: Bracket. ESPN.

https://www.espn.com/nba/bracket/_/year/2022

[2] Basketball Reference. (2022). NBA 2022 Playoffs Per Game Stats. Basketball Reference.

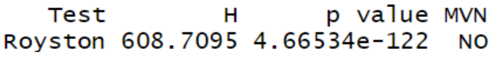
https://www.basketball-reference.com/playoffs/NBA_2022_per_game.html

[3] Complete 2022 NBA Playoffs Schedule, Results, Game Times, TV, Point Spreads: Updated Daily. Sports Illustrated FanNation. (2022)

https://www.si.com/fannation/nba/fastbreak/news/complete-2022-nba-playoffs-schedule-results-gametimes-tv-point-spreads-updated-daily

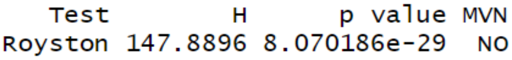
Appendices

TABLE 1:



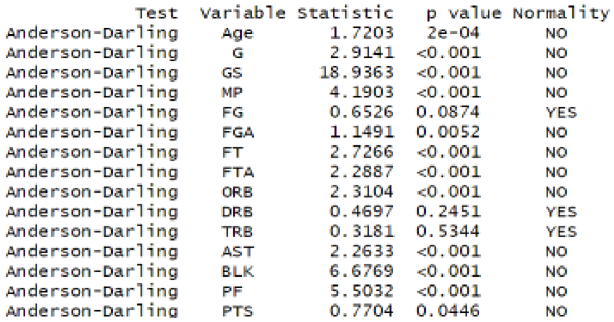
```
   Test        H       p value MVN
Royston 608.7095 4.66534e-122  NO
```

TABLE 2:



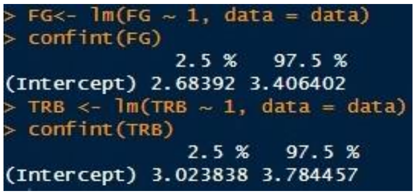
```
   Test        H       p value MVN
Royston 147.8896 8.070186e-29  NO
```

TABLE 3:



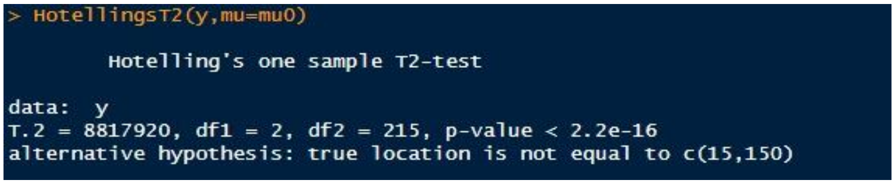
```
             Test  Variable Statistic   p value Normality
 Anderson-Darling    Age       1.7203   2e-04      NO
 Anderson-Darling     G        2.9141  <0.001      NO
 Anderson-Darling    GS       18.9363  <0.001      NO
 Anderson-Darling    MP        4.1903  <0.001      NO
 Anderson-Darling    FG        0.6526  0.0874      YES
 Anderson-Darling    FGA       1.1491  0.0052      NO
 Anderson-Darling    FT        2.7266  <0.001      NO
 Anderson-Darling    FTA       2.2887  <0.001      NO
 Anderson-Darling    ORB       2.3104  <0.001      NO
 Anderson-Darling    DRB       0.4697  0.2451      YES
 Anderson-Darling    TRB       0.3181  0.5344      YES
 Anderson-Darling    AST       2.2633  <0.001      NO
 Anderson-Darling    BLK       6.6769  <0.001      NO
 Anderson-Darling    PF        5.5032  <0.001      NO
 Anderson-Darling    PTS       0.7704  0.0446      NO
```

TABLE 4:



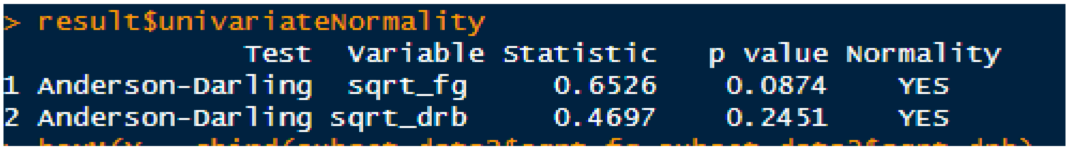
```
> FG<- lm(FG ~ 1, data = data)
> confint(FG)
               2.5 %   97.5 %
(Intercept) 2.68392 3.406402
> TRB <- lm(TRB ~ 1, data = data)
> confint(TRB)
                2.5 %   97.5 %
(Intercept) 3.023838 3.784457
```

TABLE 5:



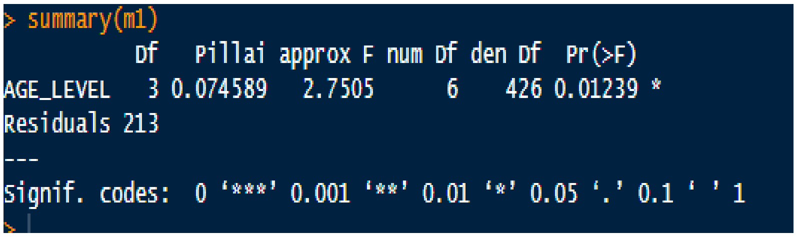
```
> HotellingsT2(y,mu=mu0)

        Hotelling's one sample T2-test

data:  y
T.2 = 8817920, df1 = 2, df2 = 215, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(15,150)
```

TABLE.6:



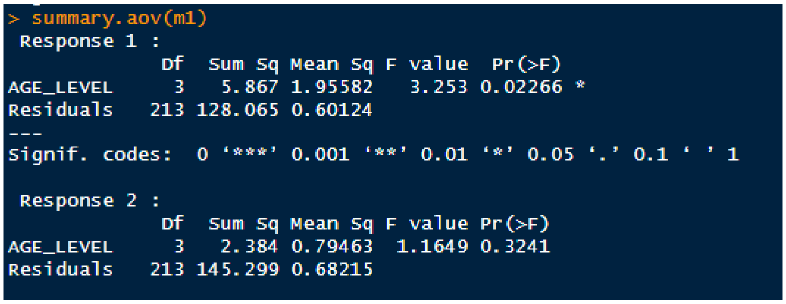
```
> result$univariateNormality
              Test  Variable Statistic   p value Normality
1 Anderson-Darling  sqrt_fg     0.6526    0.0874    YES
2 Anderson-Darling sqrt_drb     0.4697    0.2451    YES
```

TABLE.7:



```
> summary(m1)
           Df   Pillai approx F num Df den Df  Pr(>F)
AGE_LEVEL   3 0.074589   2.7505      6    426 0.01239 *
Residuals 213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

TABLE.8:

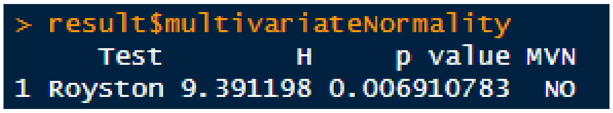


```
> summary.aov(m1)
 Response 1 :
             Df  Sum Sq Mean Sq F value  Pr(>F)  
AGE_LEVEL     3   5.867 1.95582   3.253 0.02266 *
Residuals   213 128.065 0.60124                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response 2 :
             Df  Sum Sq Mean Sq F value Pr(>F)
AGE_LEVEL     3   2.384 0.79463  1.1649 0.3241
Residuals   213 145.299 0.68215
```

response1 == TRB

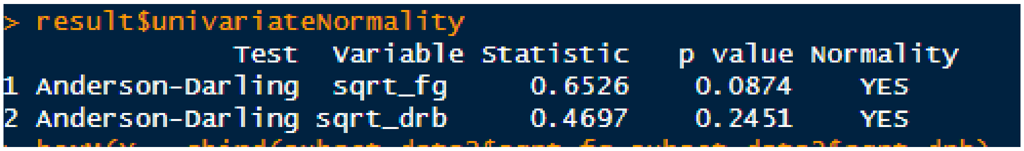
response2 == FG

TABLE 9:



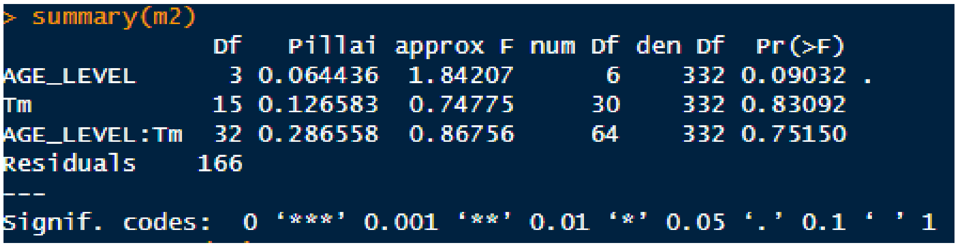
```
> result$multivariateNormality
     Test        H      p value MVN
1 Royston 9.391198 0.006910783  NO
```

TABLE 10:



```
> result$univariateNormality
              Test  Variable Statistic   p value Normality
1 Anderson-Darling  sqrt_fg      0.6526    0.0874    YES
2 Anderson-Darling sqrt_drb      0.4697    0.2451    YES
```

TABLE 11:



```
> summary(m2)
              Df   Pillai approx F num Df den Df  Pr(>F)  
AGE_LEVEL      3 0.064436  1.84207      6    332 0.09032 .
Tm            15 0.126583  0.74775     30    332 0.83092  
AGE_LEVEL:Tm  32 0.286558  0.86756     64    332 0.75150  
Residuals    166                                          
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 1:

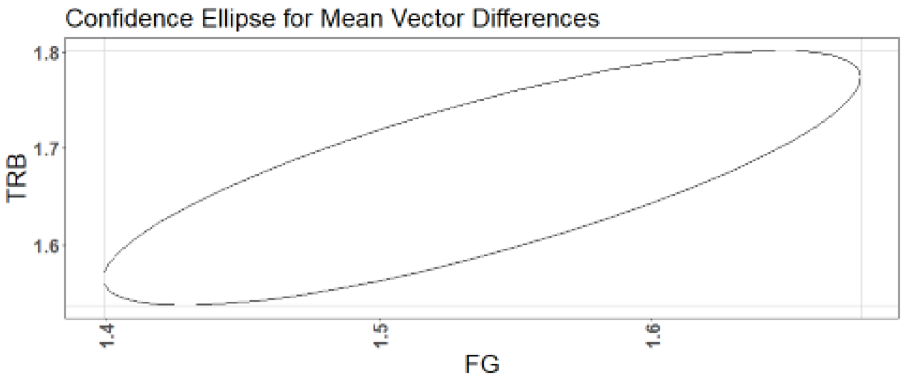


FIGURE 6:

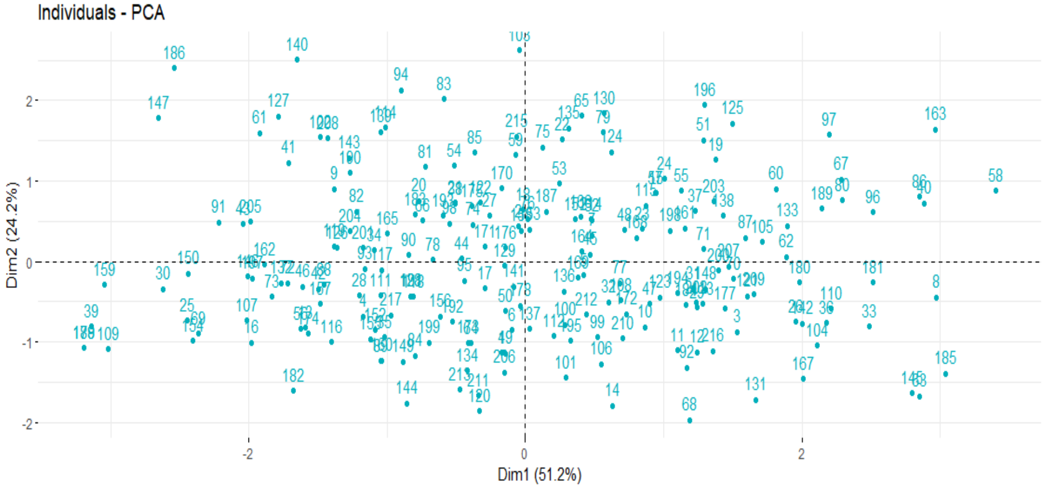
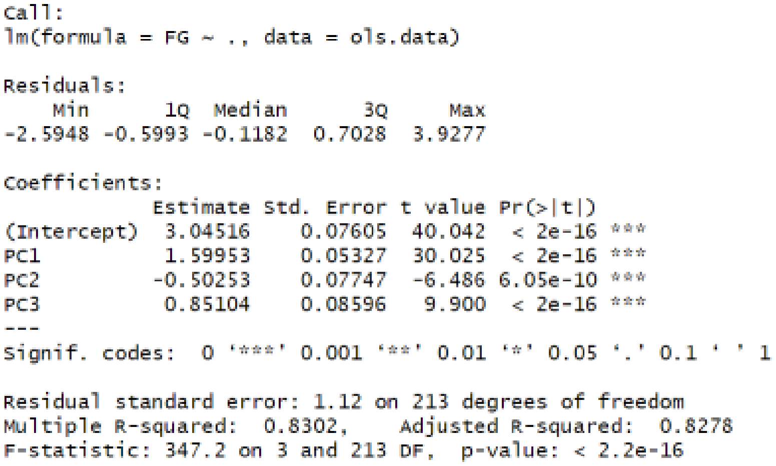
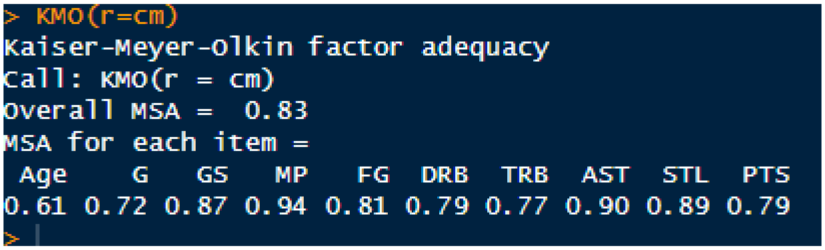


TABLE 14:



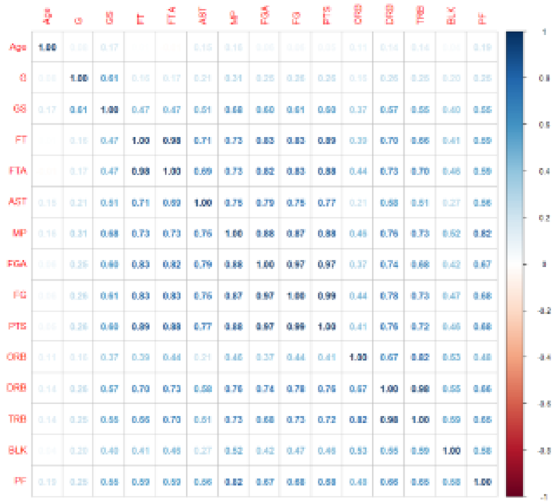
```
Call:
lm(formula = FG ~ ., data = ols.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5948 -0.5993 -0.1182  0.7028  3.9277 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.04516    0.07605  40.042  < 2e-16 ***
PC1          1.59953    0.05327  30.025  < 2e-16 ***
PC2         -0.50253    0.07747  -6.486 6.05e-10 ***
PC3          0.85104    0.08596   9.900  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.12 on 213 degrees of freedom
Multiple R-squared:  0.8302,	Adjusted R-squared:  0.8278 
F-statistic: 347.2 on 3 and 213 DF,  p-value: < 2.2e-16
```

TABLE 15:



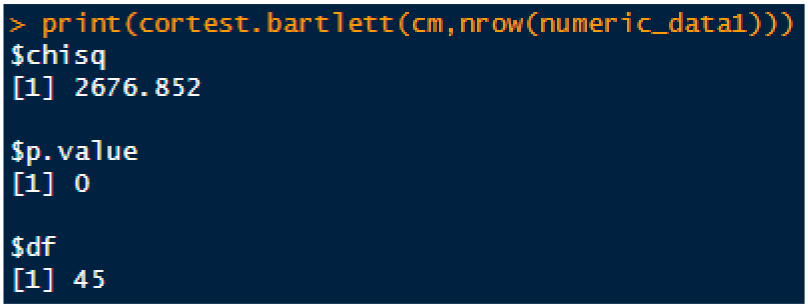
```
> KMO(r=cm)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.83
MSA for each item = 
 Age    G   GS   MP   FG  DRB  TRB  AST  STL  PTS 
0.61 0.72 0.87 0.94 0.81 0.79 0.77 0.90 0.89 0.79 
> 
```

FIGURE 7:



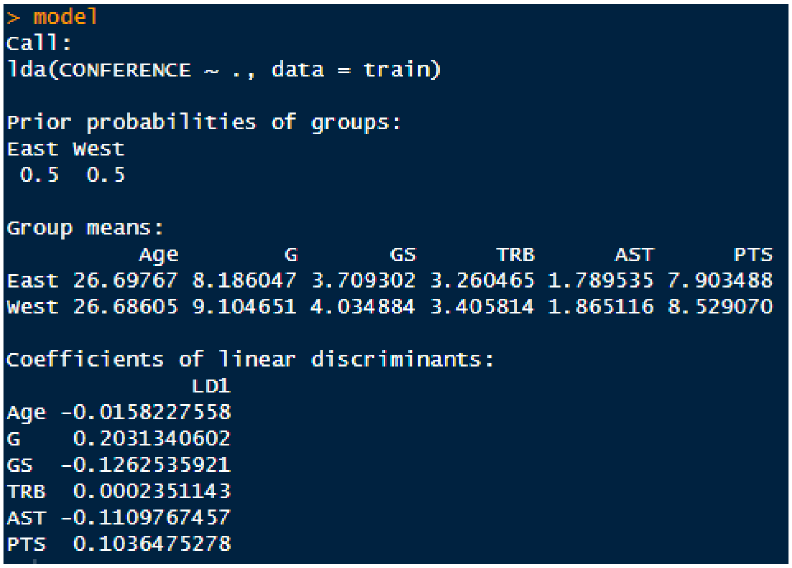
	Age	G	GS	FT	FTA	AST	MP	FGA	FG	PTS	ORB	DRB	TRB	BLK	PF
Age	1.00	0.08	0.17	[illegible]	[illegible]	0.15	0.16	0.06	0.06	0.05	0.11	0.14	0.14	0.06	0.19
G	0.08	1.00	0.61	0.16	0.17	0.21	0.31	0.25	0.26	0.26	0.16	0.26	0.25	0.20	0.25
GS	0.17	0.61	1.00	0.47	0.47	0.51	0.68	0.60	0.61	0.60	0.37	0.57	0.55	0.40	0.55
FT	[illegible]	0.16	0.47	1.00	0.98	0.71	0.73	0.83	0.83	0.89	0.39	0.70	0.66	0.41	0.59
FTA	[illegible]	0.17	0.47	0.98	1.00	0.69	0.73	0.82	0.83	0.88	0.44	0.73	0.70	0.46	0.59
AST	0.15	0.21	0.51	0.71	0.69	1.00	0.75	0.79	0.75	0.77	0.21	0.58	0.51	0.27	0.56
MP	0.16	0.31	0.68	0.73	0.73	0.76	1.00	0.88	0.87	0.88	0.45	0.76	0.73	0.52	0.82
FGA	0.06	0.25	0.60	0.83	0.82	0.79	0.88	1.00	0.97	0.97	0.37	0.74	0.68	0.42	0.67
FG	0.06	0.26	0.61	0.83	0.83	0.76	0.87	0.97	1.00	0.99	0.44	0.78	0.73	0.47	0.68
PTS	0.05	0.26	0.60	0.89	0.88	0.77	0.88	0.97	0.99	1.00	0.41	0.76	0.72	0.46	0.68
ORB	0.11	0.16	0.37	0.39	0.44	0.21	0.45	0.37	0.44	0.41	1.00	0.67	0.82	0.53	0.48
DRB	0.14	0.26	0.57	0.70	0.73	0.58	0.76	0.74	0.78	0.76	0.67	1.00	0.98	0.55	0.66
TRB	0.14	0.25	0.55	0.66	0.70	0.51	0.73	0.68	0.73	0.72	0.82	0.98	1.00	0.59	0.65
BLK	0.06	0.20	0.40	0.41	0.46	0.27	0.52	0.42	0.47	0.46	0.53	0.55	0.59	1.00	0.58
PF	0.19	0.25	0.55	0.59	0.59	0.56	0.82	0.67	0.68	0.68	0.48	0.66	0.65	0.58	1.00

TABLE 16:



```
> print(cortest.bartlett(cm,nrow(numeric_data1)))
$chisq
[1] 2676.852

$p.value
[1] 0

$df
[1] 45
```

TABLE 17:



```
> model
Call:
lda(CONFERENCE ~ ., data = train)

Prior probabilities of groups:
East West 
 0.5  0.5 

Group means:
          Age        G       GS      TRB      AST      PTS
East 26.69767 8.186047 3.709302 3.260465 1.789535 7.903488
West 26.68605 9.104651 4.034884 3.405814 1.865116 8.529070

Coefficients of linear discriminants:
              LD1
Age -0.0158227558
G    0.2031340602
GS  -0.1262535921
TRB  0.0002351143
AST -0.1109767457
PTS  0.1036475278
```

FIGURE 2:

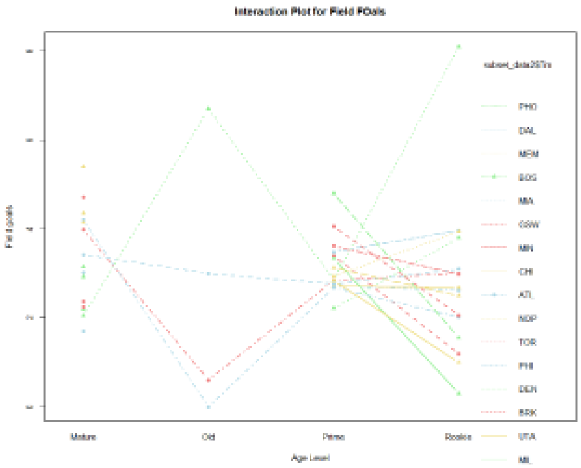


FIGURE 3:

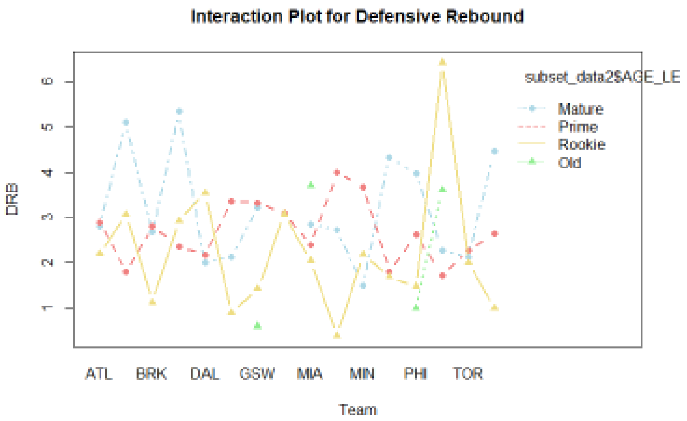


TABLE 12:

```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.4308 0.9839 0.8867 0.44561
Proportion of Variance 0.5118 0.2420 0.1966 0.04964
Cumulative Proportion  0.5118 0.7538 0.9504 1.00000
```

TABLE 13:

```
          PC1         PC2         PC3
Age 0.2629834  0.96391893  0.02426502
G   0.5896673 -0.04817297 -0.80602245
AST 0.9033969 -0.07257840  0.28398673
PTS 0.9023090 -0.17679263  0.23534187
```

FIGURE 4:

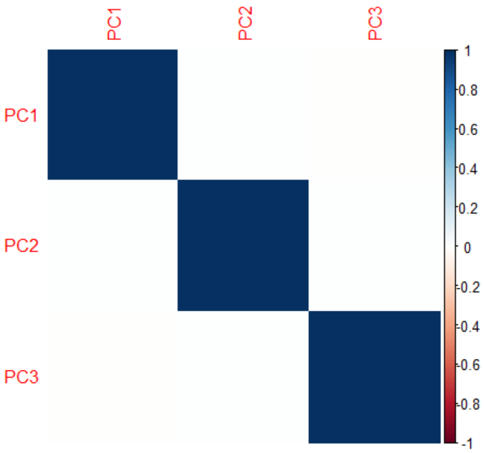


FIGURE 5:

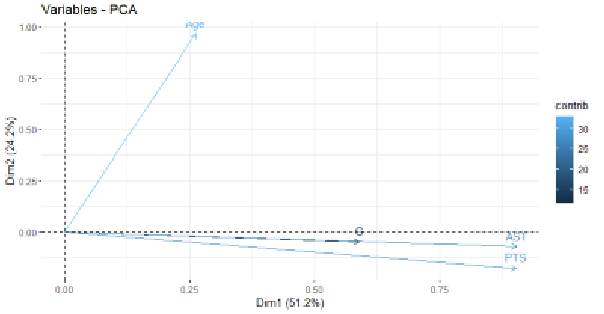


FIGURE 8:

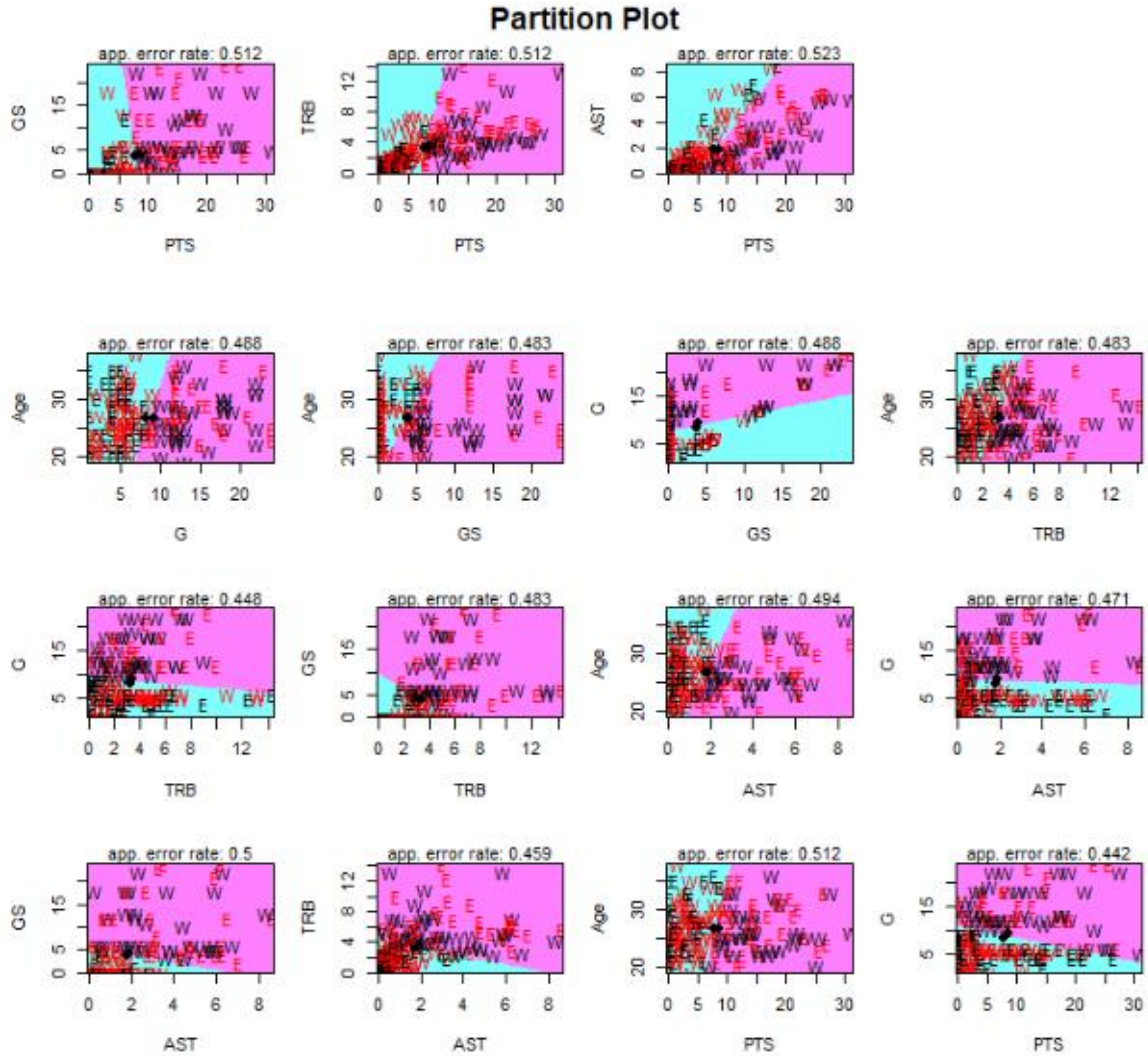


TABLE 18:

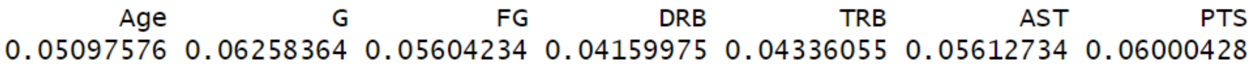

```
       Age          G         FG        DRB        TRB        AST        PTS
0.05097576 0.06258364 0.05604234 0.04159975 0.04336055 0.05612734 0.06000428
```

TABLE 19:

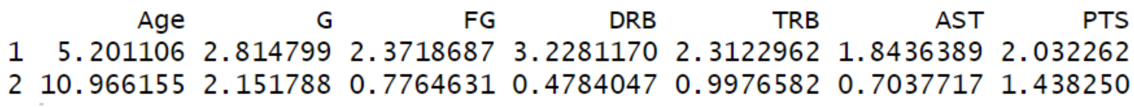
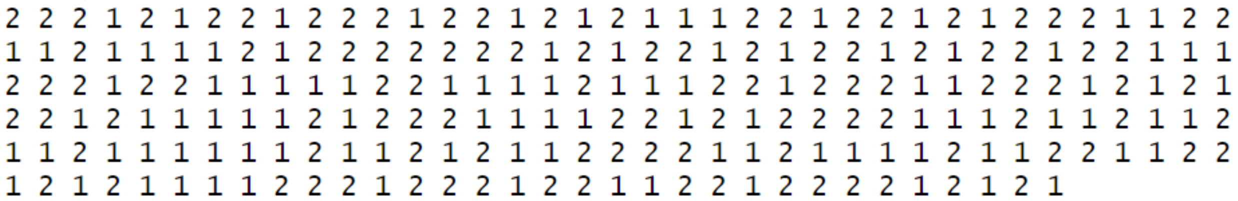

```
        Age        G        FG       DRB       TRB       AST      PTS
1  5.201106 2.814799 2.3718687 3.2281170 2.3122962 1.8436389 2.032262
2 10.966155 2.151788 0.7764631 0.4784047 0.9976582 0.7037717 1.438250
```

TABLE 20:



```
2 2 2 1 2 1 2 2 1 2 2 2 1 2 2 1 2 1 2 1 1 1 2 2 1 2 2 1 2 1 2 2 2 1 1 2 2
1 1 2 1 1 1 1 2 1 2 2 2 2 2 2 2 1 2 1 2 2 1 2 1 2 2 1 2 1 2 2 1 2 2 1 1 1
2 2 2 1 2 2 1 1 1 1 1 2 2 1 1 1 1 2 1 1 1 2 2 1 2 2 2 1 1 2 2 2 1 2 1 2 1
2 2 1 2 1 1 1 1 1 2 1 2 2 2 1 1 1 1 2 2 1 2 1 2 2 2 2 1 1 1 2 1 1 2 1 1 2
1 1 2 1 1 1 1 1 1 2 1 1 2 1 2 1 1 2 2 2 2 1 1 2 1 1 1 1 2 1 1 2 2 1 1 2 2
1 2 1 2 1 1 1 1 2 2 2 1 2 2 2 1 2 2 1 1 2 2 1 2 2 2 2 1 2 1 2 1
```