A Multivariate Analysis of NBA Player Data

Julia Stiller

5/4/2024

Introduction

The National Basketball Association (NBA) serves as a dynamic arena where talent, strategy, and statistics intersect to shape the events of a game. Statistical analyses play a pivotal role in unraveling the complexities of player performance, offering insights that can inform strategic decisions and drive competitive advantage. This project delves into the intricate web of relationships among diverse performance metrics, leveraging multivariate statistical techniques to extract meaningful patterns from player-level NBA data.

In an era where data-driven decision-making reigns supreme, insights gleaned from statistical models can provide a competitive edge for players, coaches, analysts, and even fans alike. By dissecting the multidimensional nature of player performance, these analyses offer a holistic view of an athlete's contributions on the court, transcending traditional box score metrics to uncover nuanced insights. Statistical modeling empowers teams to optimize player utilization, refine game strategies, and forecast outcomes with greater accuracy. In essence, the fusion of statistics and sports not only enhances our understanding of the game but also revolutionizes how we approach player evaluation, team management, and strategic planning in the ever-evolving landscape of professional basketball.

Design and Primary Questions

Through this exploration, I hope to answer pressing questions such as:

- What are the key dimensions driving variations in player performance metrics?
- Which combinations of performance metrics are most indicative of a player's overall impact on the game?
- What predictive models can we construct to forecast player or team outcomes based on historical data?
- Can we classify teams into distinct categories based on their statistical profiles, and what insights do these clusters offer?

I will broach the analysis of these questions with the following tests:

1. Principal Component Analysis (PCA)
2. Quadratic Discriminant Analysis (QDA)
3. Multivariate Analysis of Variance (MANOVA)
4. Cluster Analysis

The study will involve analyzing player-level NBA data spanning the 2021 - 2022 season, including scoring, rebounding, assists, and other performance metrics. After data preprocessing, PCA will uncover underlying patterns, followed by QDA for player classification. MANOVA will then assess performance differences across teams. Finally, Cluster Analysis will identify teams clusters, providing insights for talent evaluation and team strategy.

Data

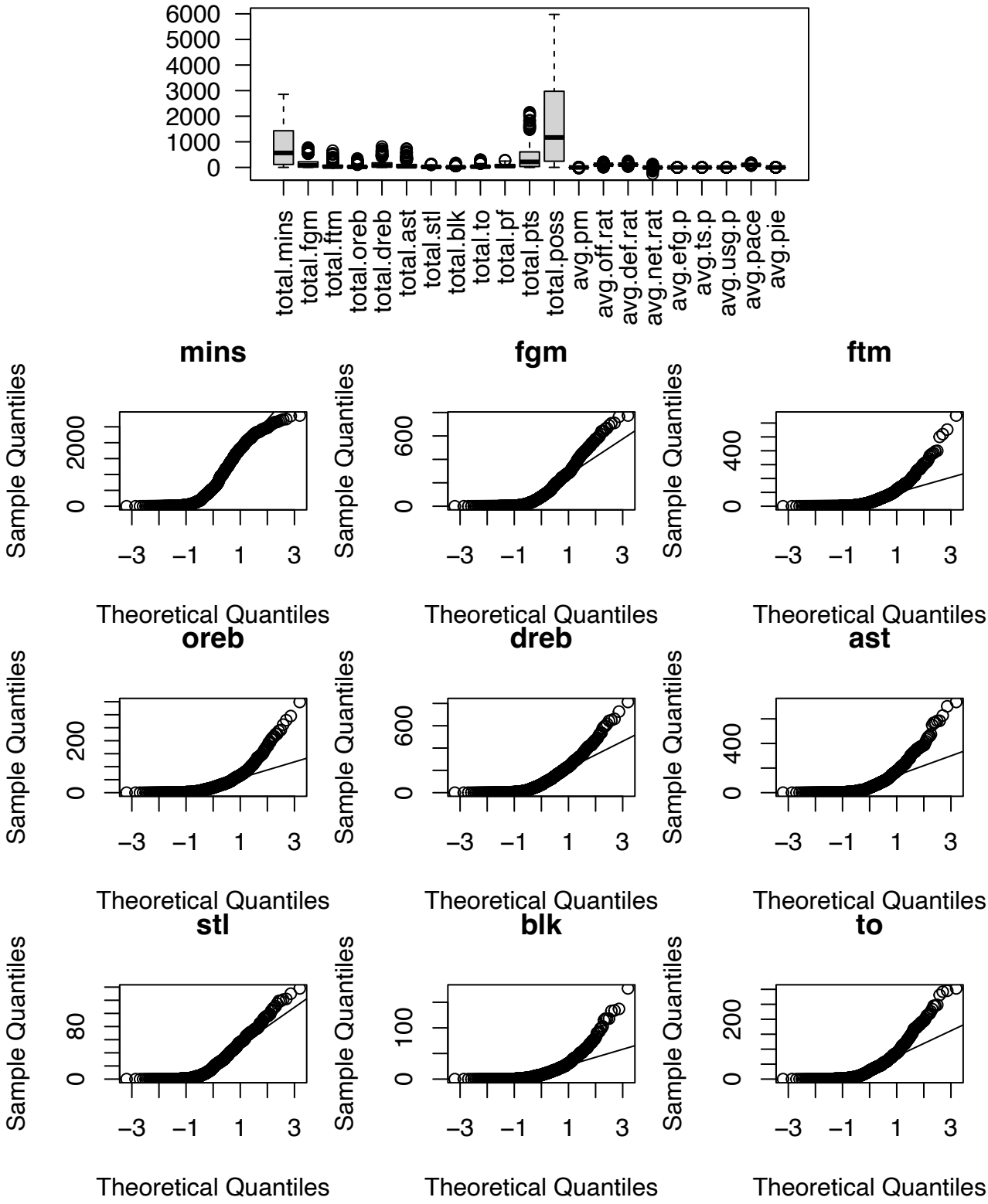
My dataset is NBA player-level data from the 2021 - 2022 season. I have 21 variables related to the performance of 715 players. The variables are as follows:

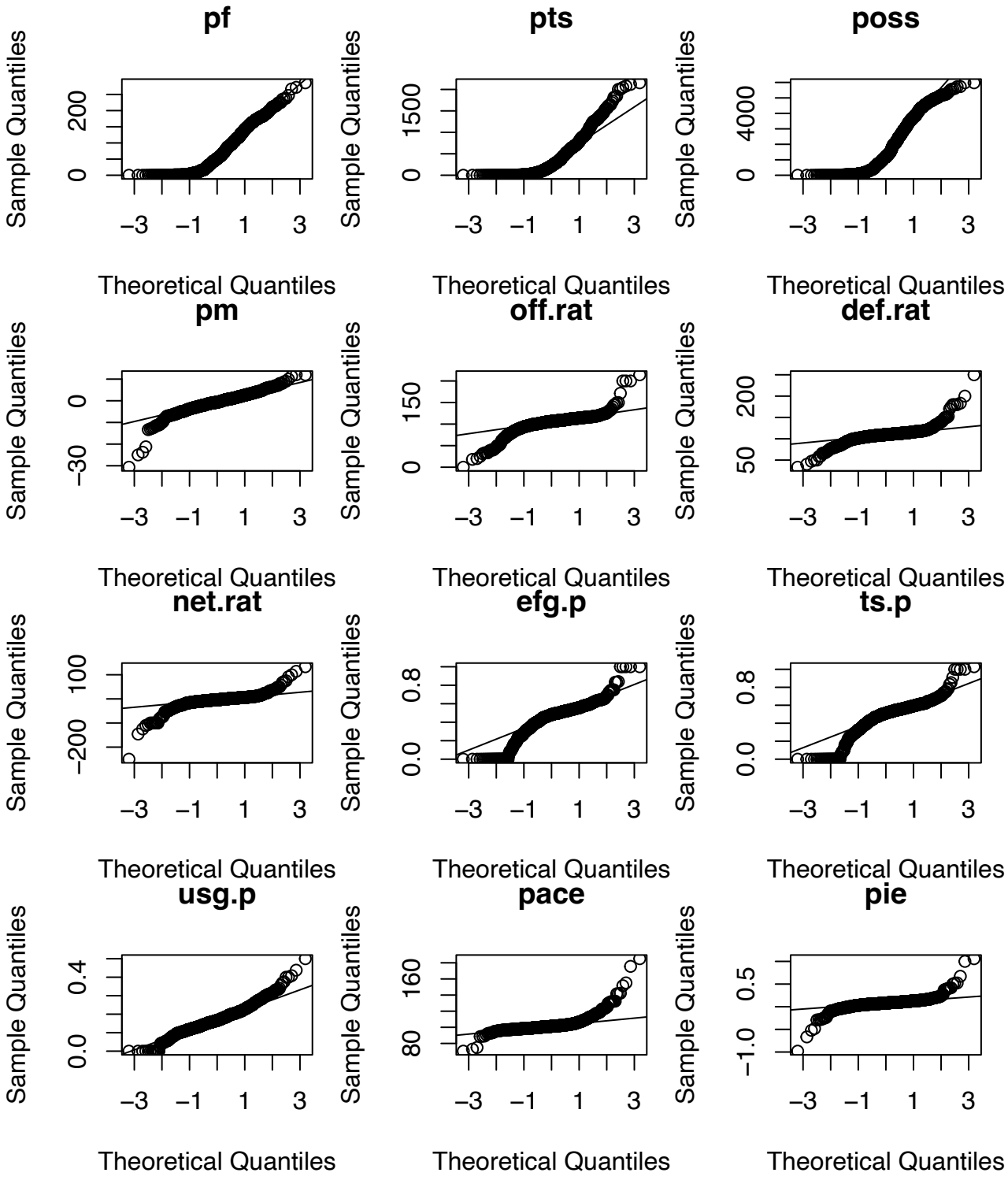
Variable	Description	Type
player	Player	Categorical
team	Team	Categorical
season	Season	Categorical
pid	Player ID	Categorical
tid	Team ID	Categorical
total.mins	Total Minutes Played	Continuous
total.fgm	Total Field Goals Made	Continuous
total.ftm	Total Free Throws Made	Continuous
total.oreb	Total Offensive Rebounds	Continuous
total.dreb	Total Defensive Rebounds	Continuous
total.ast	Total Assists	Continuous
total.stl	Total Steals	Continuous
total.blk	Total Blocks	Continuous
total.to	Total Turnovers	Continuous
total.pf	Total Personal Fouls	Continuous
total.pts	Total Points	Continuous
total.poss	Total Possessions	Continuous
avg.pm	Average Plus-Minus	Continuous
avg.off.rat	Average Offensive Rating	Continuous
avg.def.rat	Average Defensive Rating	Continuous
avg.net.rat	Average Net Rating	Continuous
avg.efg.p	Average Effective Field Goal Percentage	Continuous
avg.ts.p	Average True Shooting Percentage	Continuous
avg.usg.p	Average Usage Percentage	Continuous
avg.pace	Average Pace	Continuous
avg.pie	Average Player Impact Estimate	Continuous

This data was originally collected by experts starting in 2011 and going all the way up to 2022. Computers calculated aggregate metrics such as the ratings and percentages. I narrowed down the data to just be the 2021 - 2022 season for this analysis since the dataset was so large. Expanding to additional seasons will definitely be a point for further analysis in the future. Since the data was collected by humans, that is always a possible source of error, but there is so much data that small deviations should get averaged out. I did not notice any questionable points in the data, but I will be looking for outliers and other issues throughout my analysis.

Descriptive Plots and Summary Statistics

First, I will examine the univariate distributions of my variables by looking at the boxplots and normal quantile plots for each variable. I will then make transformations as appropriate.


6000
5000
4000
3000
2000
1000
0
total.mins
total.fgm
total.ftm
total.oreb
total.dreb
total.ast
total.stl
total.blk
total.to
total.pf
total.pts
total.poss
avg.pm
avg.off.rat
avg.def.rat
avg.net.rat
avg.efg.p
avg.ts.p
avg.usg.p
avg.pace
avg.pie
mins
fgm
ftm
oreb
dreb
ast
stl
blk
to
Sample Quantiles
Theoretical Quantiles
0
2000
600
400
200
100
80
−3
−1
1
3



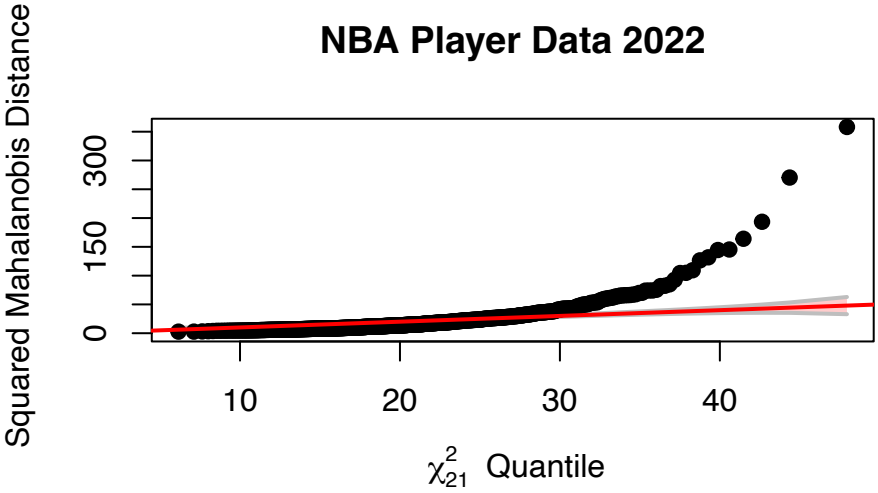
A lot of my "total" variables (field goals made, free throws made, offensive rebounds, defensive rebounds, assists, blocks, turnovers, points, possessions) have a lot of outliers and are not normally distributed. This makes sense because count data often follows Poisson distributions. Additionally, player-level basketball data like this is going to vary a lot by the position of the player. For example, a center is going to have a lot more rebounds and blocks than a point guard. I also see that my "average" variables tend to be more normally distributed, but the values of these variables are large and therefore could be overbearing in my analysis. Because of all of this information, I am going to transform some of my variables by taking the square root.

I'll start by transforming field goals made, free throws made, offensive rebounds, defensive rebounds, assists, steals, blocks, turnovers, personal fouls, points, possessions, offensive rating, defensive rating, and pace.

Multivariate Analysis and Results

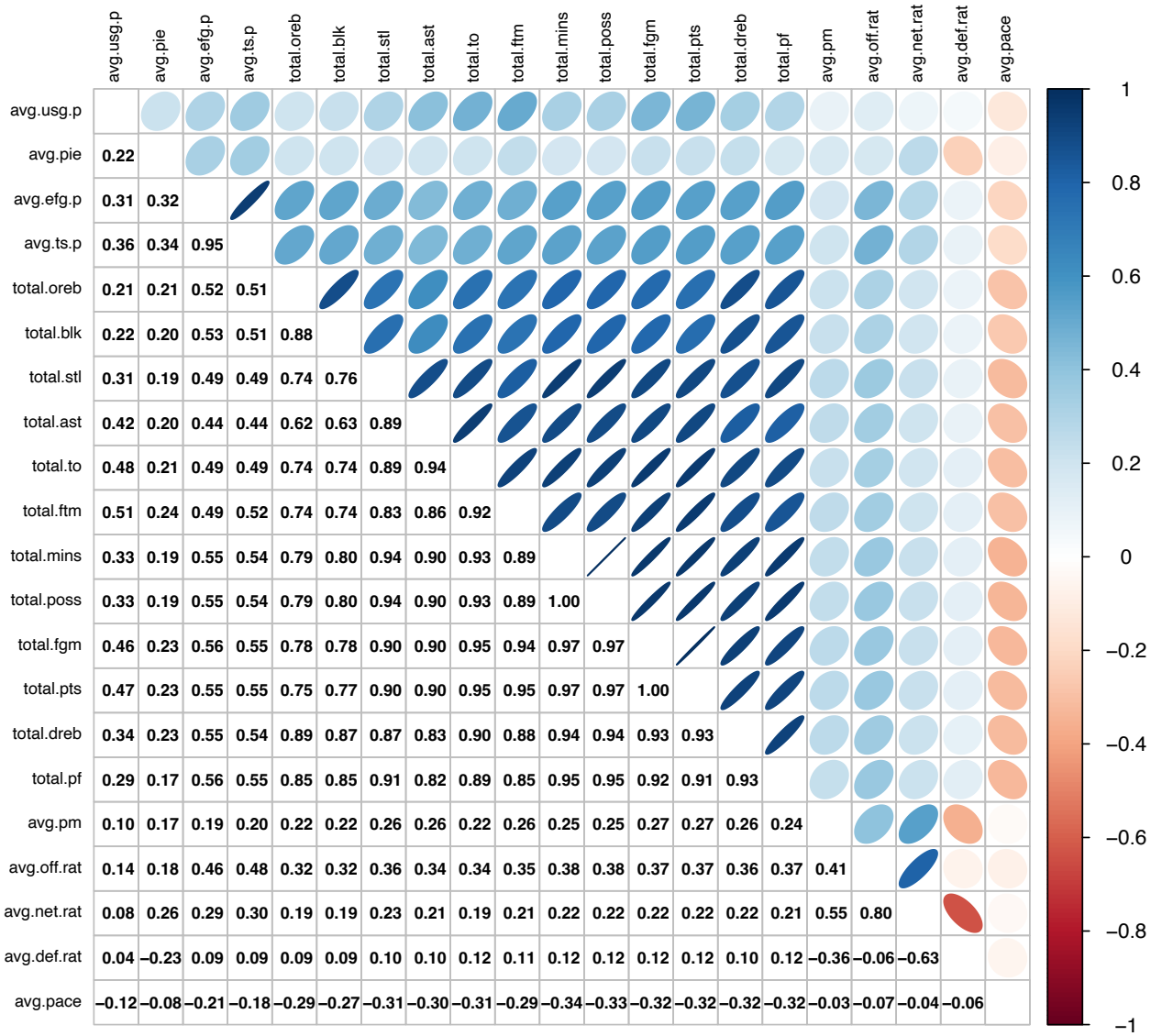
1. Principal Components Analysis (PCA)

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset by transforming the original variables into a set of linearly uncorrelated variables called principal components. These components capture the maximum amount of variance in the data, allowing for a more concise representation of the underlying patterns and relationships among the variables. First, I will examine whether the data seems to have a multivariate normal distribution.



The data does not appear to have a multivariate normal distribution (although this plot looks better than it did before I made the transformations above), which means that my data is skewed. This is not a requirement for PCA, but it is good to know for conducting my other analyses and interpreting them.

Now we can look at the correlation matrix between all variables. In PCA, we want to see if there are strong relationships between the variables because PCA will identify directions where most of the variability in the data occurs. If there are strong relationships between the variables, then PCA will work well. If there are not strong relationships between the variables, then PCA will not work well.



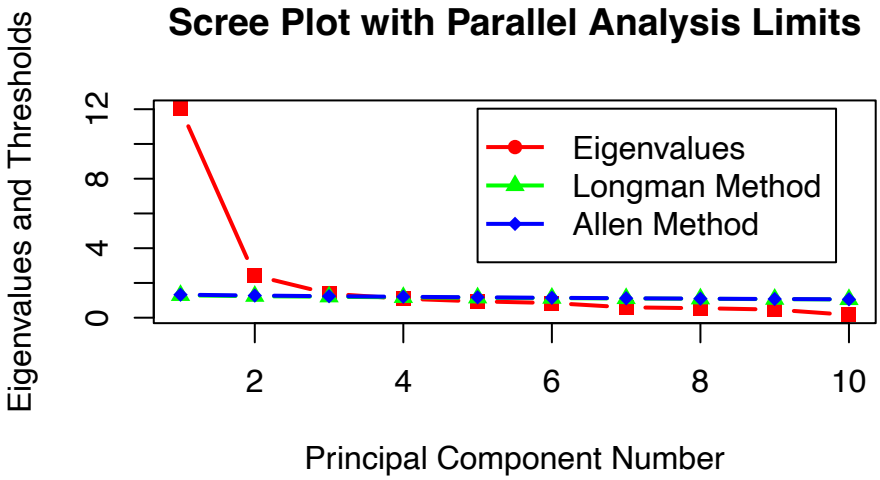
Based on this correlation visual, PCA will work well because a lot of the variables are strongly correlated, so it will be able to identify directions where most of the variability in the data occurs. For example, there is a strong positive correlation between field goals made, free throws made, and points, which makes sense because field goals and free throws literally constitute the total number of points. Interestingly, defensive rating is very weakly correlated with almost all of the other variables, except it is pretty negatively correlated with net rating. This also makes sense because net rating is offensive rating MINUS defensive rating, so when defensive rating is big, it pulls the net rating down, whereas it does not have as much of an effect on the net rating when defensive rating is small. Overall, PCA should work well because there are a lot of relationships between the variables which can definitely be reduced down into fewer dimensions.

Next, PCA was conducted on the correlation matrix of the data, which standardizes the variables so that they are all on the same scale. This is important because PCA is looking for directions where most of the variability in the data occurs, and if the variables are on different scales, then PCA will be biased towards the variables with the largest scale. I will examine how many principle components to retain based on the cumulative proportion of variance explained by a given number of PC's, the eigenvalues, a scree plot, and parallel analysis.

	Cumulative Proportion of Variance	Eigenvalues
Comp.1	57.29%	12.03
Comp.2	68.83%	2.42
Comp.3	75.43%	1.39
Comp.4	80.63%	1.09
Comp.5	85.11%	0.94
Comp.6	89.14%	0.85
Comp.7	91.97%	0.6
Comp.8	94.57%	0.55
Comp.9	96.79%	0.47
Comp.10	97.55%	0.16
Comp.11	98.14%	0.12
Comp.12	98.68%	0.11
Comp.13	99.04%	0.08
Comp.14	99.34%	0.06
Comp.15	99.56%	0.05
Comp.16	99.74%	0.04
Comp.17	99.87%	0.03
Comp.18	99.95%	0.02
Comp.19	99.99%	0.01
Comp.20	99.99%	0
Comp.21	99.99%	0

If I want to account for 80% of the variance in my principle components, based on the cumulative proportion displayed above, I would need to keep the first five PC's. These would account for 84.02% of the variance in my data. If I want to keep eigenvalues greater than 1, I would need to keep the first four PC's, as the fifth one is 0.95. Next, I will check a scree plot and perform a parallel analysis to see whether this helps determine how many PC's to examine. However, as mentioned in the summary statistics section above, my data does not have a multivariate normal distribution, so I will be cautious about interpreting the results of the parallel analysis.

```
##    pcompnum  longman    allen
## 1         1 1.286781 1.325017
## 2         2 1.237118 1.275676
## 3         3 1.204850 1.236918
## 4         4 1.170669 1.204394
## 5         5 1.146002 1.174077
## 6         6 1.124936 1.146425
## 7         7 1.106372 1.119762
## 8         8 1.089169 1.095362
## 9         9 1.066136 1.075780
## 10       10 1.049412 1.056897
```



To interpret the scree plot, I am looking for an "elbow" in the red line to see where I should cut off the number of principle components. I would say that the most obvious elbow is at 2 PC's, meaning I would only keep the first one. However, taking the parallel analysis into account, I am looking for where the blue and green lines intersect with the red line. I would say that the lines cross paths between the 3rd and 4th principal components, which would suggest that I keep the first 3 PC's. Combining all of the observations together, I will keep the first 4 PC's to find a middle ground between this parallel analysis, the cumulative proportion of variance explained, and the eigenvalues.

Now that I have decided to keep four principal components, I will examine the loadings for these retained components and think about how to interpret them.

	Comp.1	Comp.2	Comp.3	Comp.4
total.mins	0.28	0.07	0.09	0.05
total.fgm	0.28	0.06	0.06	-0.08
total.ftm	0.27	0.06	0.05	-0.16
total.oreb	0.24	0.04	0.02	0.22
total.dreb	0.28	0.06	0.07	0.06
total.ast	0.26	0.06	0.14	-0.15
total.stl	0.27	0.06	0.13	0.03
total.blk	0.24	0.04	0.02	0.21
total.to	0.27	0.08	0.09	-0.13
total.pf	0.27	0.07	0.07	0.13
total.pts	0.28	0.06	0.06	-0.09
total.poss	0.28	0.07	0.09	0.04
avg.pm	0.09	-0.39	0.25	0.00
avg.off.rat	0.13	-0.40	-0.10	0.29
avg.def.rat	0.03	0.45	-0.33	0.24
avg.net.rat	0.09	-0.57	0.14	0.09
avg.efg.p	0.19	-0.15	-0.55	0.14
avg.ts.p	0.19	-0.16	-0.56	0.10
avg.usg.p	0.12	0.00	-0.21	-0.66
avg.pace	-0.10	-0.08	0.01	0.00
avg.pie	0.08	-0.24	-0.25	-0.43

1. The first PC is loaded the most on minutes, field goals made, points, and possessions, which suggests that this principal component represents time with the ball. This is because the more minutes a player is on the court, the more likely they are to possess the ball, and thus they are likely to get points during these possessions (i.e. field goals made).

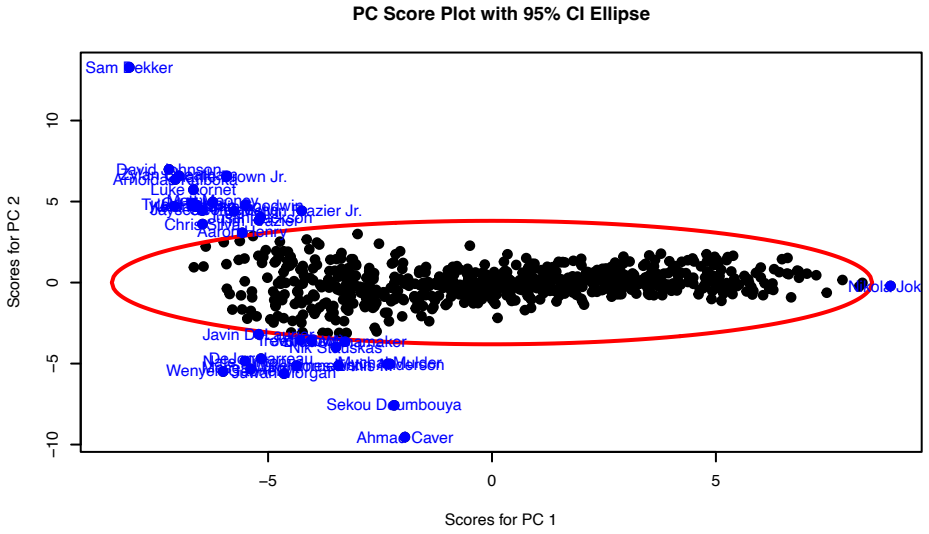
2. The second PC is loaded the most on defensive rating, net rating, and offensive rating which suggests that this principal component represents the player's overall ability to score and prevent their opponent from scoring. As previously mentioned, the net rating is the difference between the offensive and defensive ratings, so a player with a high net rating is likely to have a high offensive rating and a low defensive rating, which is why the loading for the defensive rating is the opposite sign from the other two.

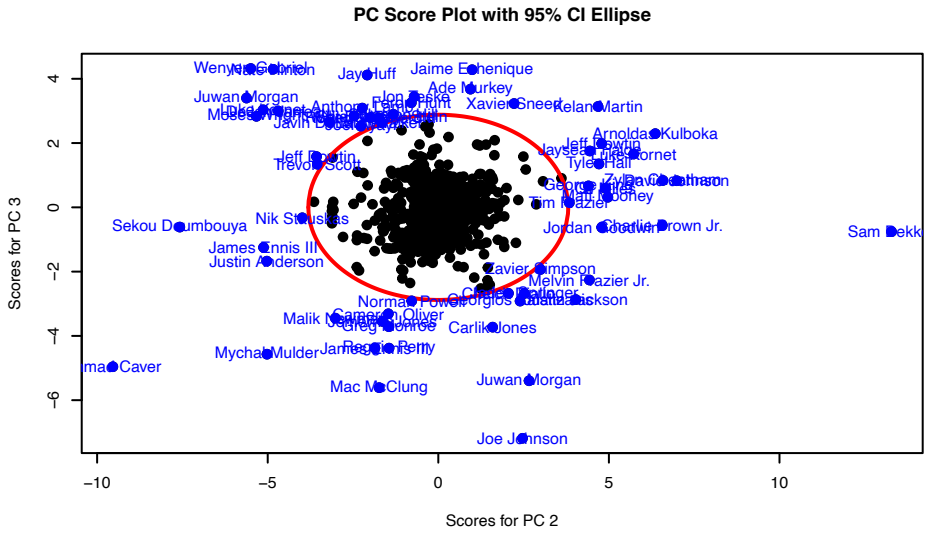
3. The third PC is loaded the most on effective field goal percentage and true shooting percentage which suggests that this principal component represents the player's shooting accuracy. This again makes sense that they would be loaded the most on the same principal component.

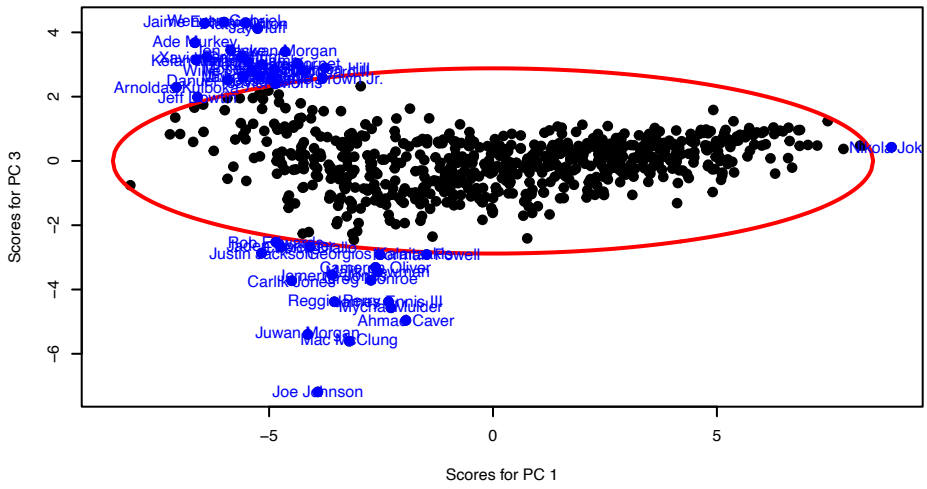
4. The fourth PC is loaded the most on usage percentage and player impact estimate which suggests that this principal component represents the player's overall impact on the game. This is because the usage percentage represents how many times a player ends one of his team's possessions (generally with field goal attempts, but could also be due to turnovers, etc.) and the player impact estimate measures their contribution overall.

Therefore, in summary, the majority of the variability in my data can be explained by the player's time with the ball, their ability to score while preventing their opponent from scoring, their shooting accuracy, and their overall impact on the game. These are all key aspects of a player's performance in a basketball game, so the results of PCA make sense in the context of this topic.

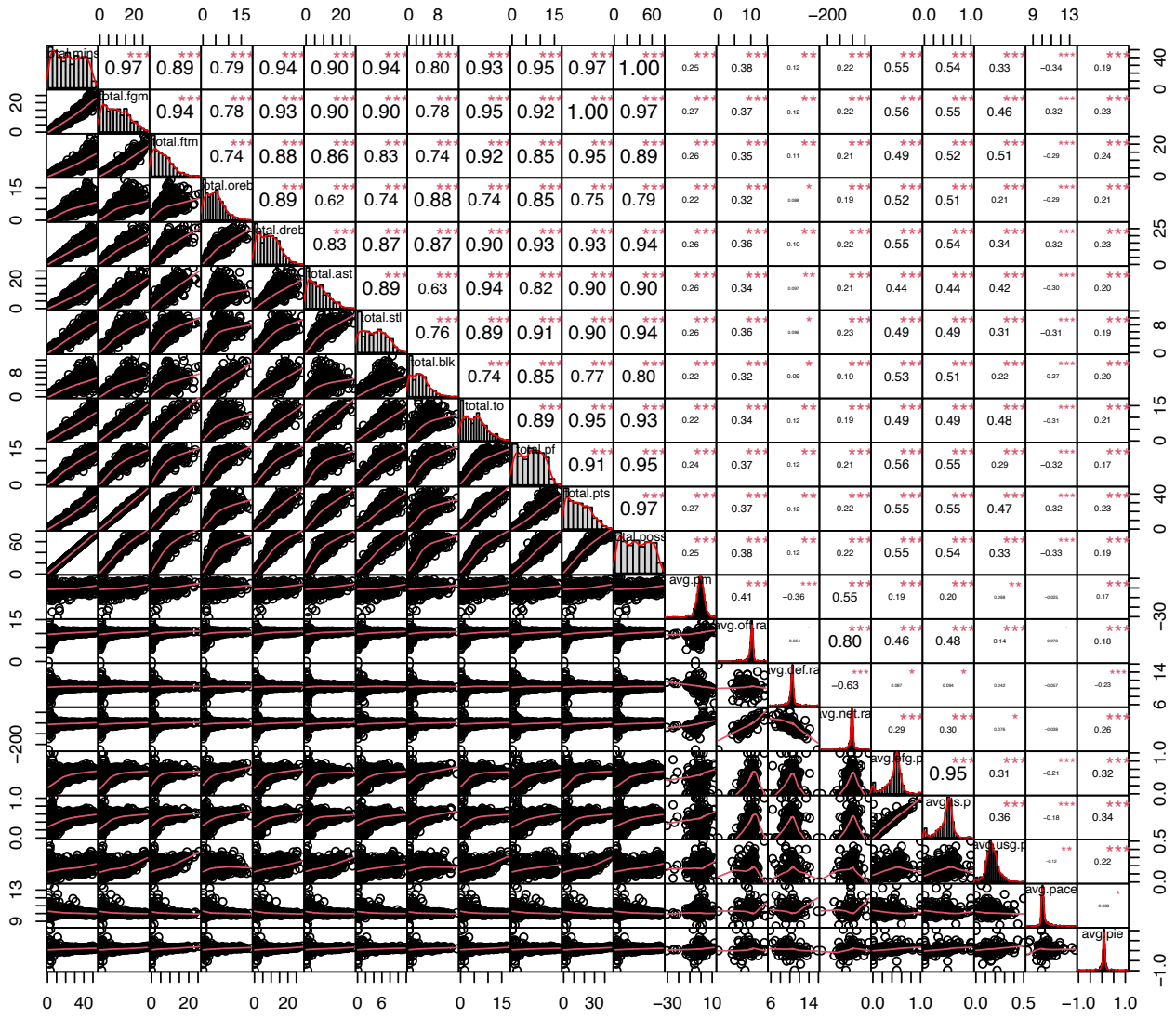
Next, I made score plots for a few pairs of component scores (one and two, one and three, two and three). The score plots include a 95% Confidence Ellipse for the two components.







I can see some groupings based on position- like I talked about in the beginning, with player-level basketball data, a lot of the differences come from the fact that each player has an assigned position and therefore has different priorities when it comes to their actions on the court. This is then reflected in their statistics being higher in some areas while lower in others. For example, in the plots for PC's 1 and 2, I see a lot of point guards in the top left corner which I think makes sense because they are trying to prevent the other team from scoring, which is consistent with a high PC 2. They might not possess the ball as much or score very often which is consistent with a low PC 1. In the plots for PC's 2 and 3, there is definitely much less of a pattern because the confidence ellipse is more centered with outliers in all directions. This makes sense because we know that the majority of the variability in the data is captured by PC 1, which is not included in this graph. Finally in the plots for PC's 1 and 3, I see some forwards in the bottom left corner which I think makes sense because this position is known for not scoring very much. So this would be consistent with the low points of PC 1 as well as the low shooting accuracy of PC 3. These are just a few examples but are not necessarily the case for every point outside of the confidence ellipse.



To conclude my Principal Component Analysis, I examine the scatter plots above and see that some of my variables have strong linear relationships, but most of them have weak linear relationships, if it all. Because PCA assumes linearity, I do think there are limitations to the effectiveness of using PCA on this data. That being said, I was able to reduce the number of dimensions from 21 to 4, which is a significant reduction. The first 4 PC's were able to capture a lot of the variability in the data, as I was able to account for 79.5% of the variance, and the loadings for these 4 PC's were highly interpretable to capture very important aspects of a player's performance. The outliers on the score plots were able to show some groupings based on position, which is consistent with the fact that player-level sports data is going to vary a lot depending on their role. Lastly, I have a lot of observations relative to the number of variables which is a good thing for this analysis. In summary, I think that PCA was effective on this data because it was able to reduce the number of dimensions, capture a lot of the variability in the data, and show some relationships between the variables. However, due to the linearity assumption not being met, I would want to revisit my transformations to try to adjust this and see if I can get a better result.

2. Discriminant Analysis

Discriminant Analysis is a technique used to classify observations into groups based on their characteristics. I will be using Discriminant Analysis to classify players into teams based on the values of their metrics (which represent the player's performance). The first step is to consider whether the assumptions have been met. Unfortunately, after examining the chi-square quantile plots, I see that the data does not have a multivariate

normal distribution WITHIN each group. I will also examine the covariance matrices for each team to see if they are similar.

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  dd[, -c(1:5)]
## Chi-Sq (approx.) = 11248, df = 5775, p-value < 2.2e-16
```

The p-value is very small ($p<0.05$), so I reject the null hypothesis that the covariance matrices are equal. Unfortunately, this means there is evidence that the covariance matrices are different across my groups. Because of this, I will be using quadratic discriminant analysis. I will also perform linear discriminant analysis just to be able to compare the classification accuracies (since the covariance assumption is not met).

LDA	QDA
24.14%	94.08%

The quadratic discriminant analysis model accurately classifies 94% of the data whereas the linear model accurately classifies only 24% of the data. I will now proceed with checking whether there is statistical evidence that the multivariate group means are different.

```
##            Df   Wilks approx F num Df den Df    Pr(>F)
## dd$team    25 0.19327   1.9818    525 9691.1 < 2.2e-16 ***
## Residuals 616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value is very small ($p<0.05$), I reject the null hypothesis that the multivariate group means are equal. This means that there is statistical evidence that the multivariate group means are different! Furthermore, this suggests that discriminant analysis is effective in classifying the players into their respective teams because the teams have different average values for the 21 metrics that I have in my dataset. Next, I will look into how many discriminant functions are significant and what the relative discriminating power of each function is.

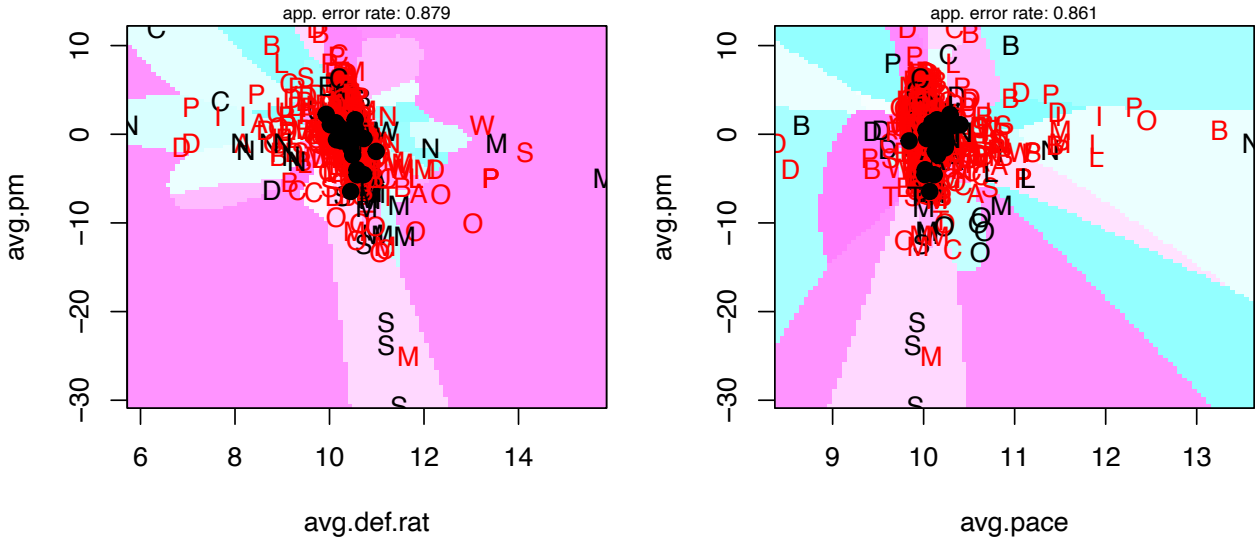
Test of Function(s)	Wilks Lambda	Approximate F	p-value
1 through 21	0.1933	1.9818	0.0000
2 through 21	0.3250	1.4622	0.0000
3 through 21	0.4893	1.0103	0.4330
4 through 21	0.5786	0.8502	0.9844
5 through 21	0.6375	0.7749	0.9993
6 through 21	0.6935	0.7019	1.0000

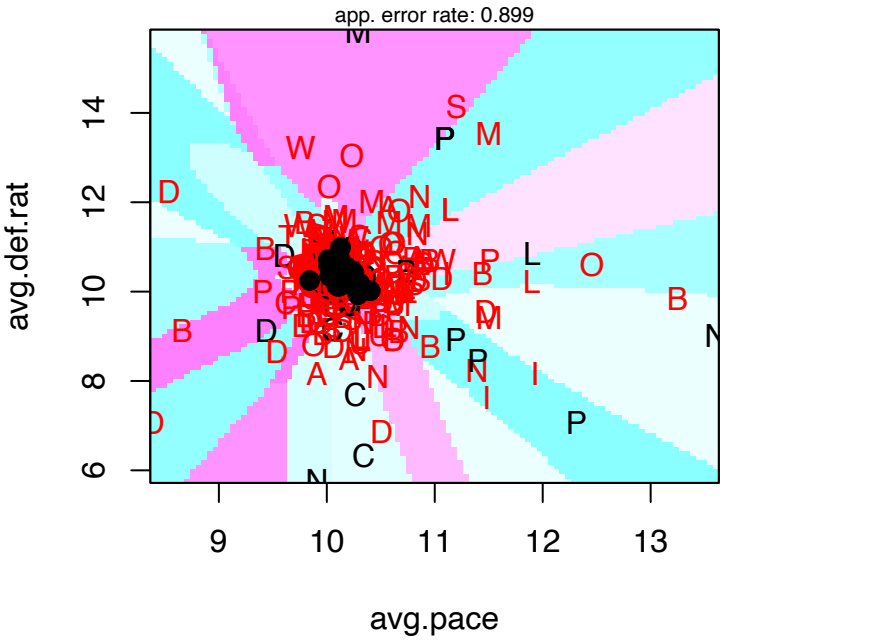
There are 21 discriminant functions since there are 21 variables, however, I only output the top 6 rows because the majority of the functions are not significant. Only the first two discriminant functions are significant ($p<0.05$), and the functions after that quickly become very insignificant. The first function has the most relative discriminating power and is about 30% better at discriminating than the second function. Now I will look into which variables are the most discriminating.

```
##  Response avg.pm :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## dd$team      25 2833.7  113.35  8.9601 < 2.2e-16 ***
## Residuals   616 7792.6   12.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Response avg.off.rat :
##              Df Sum Sq Mean Sq F value  Pr(>F)
## dd$team      25  48.62  1.9448   1.744 0.01433 *
## Residuals   616 686.90  1.1151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avg.def.rat :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## dd$team      25  33.01 1.32056  2.5271 7.012e-05 ***
## Residuals   616 321.90 0.52257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avg.net.rat :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## dd$team      25  29929 1197.17  1.9748 0.003367 **
## Residuals   616 373427  606.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avg.pace :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## dd$team      25  7.836 0.31345   1.954 0.003856 **
## Residuals   616 98.815 0.16041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I ran MANOVA on all of the variables, but these are the significant discriminating variables: average plus-minus, average offensive rating, average defensive rating, average net rating, and average pace. So lastly, I will now use this knowledge to get plots of my data in the space spanned by average plus-minus, average defensive rating, and average pace that show which regions are assigned to each group.

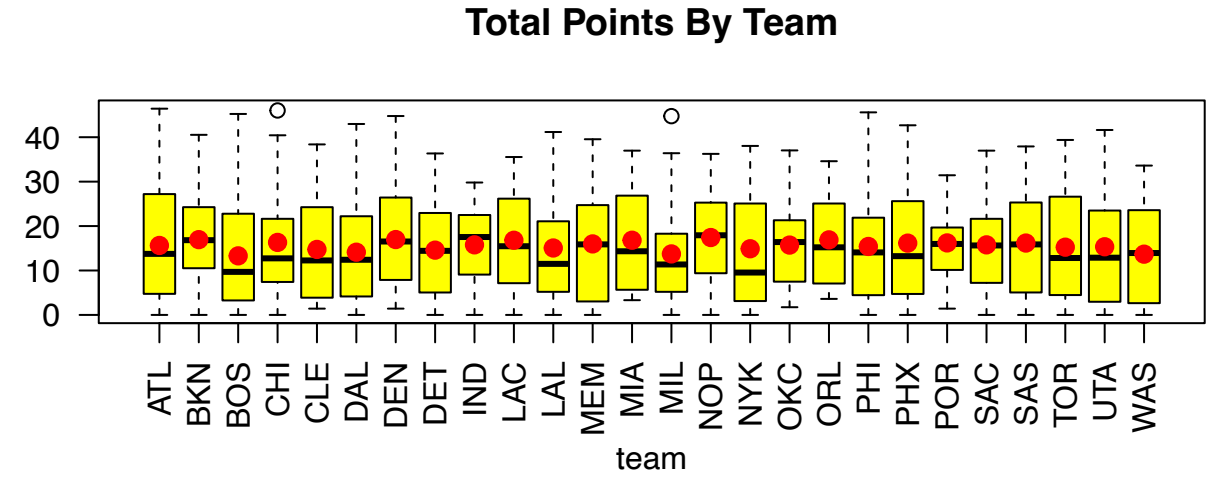


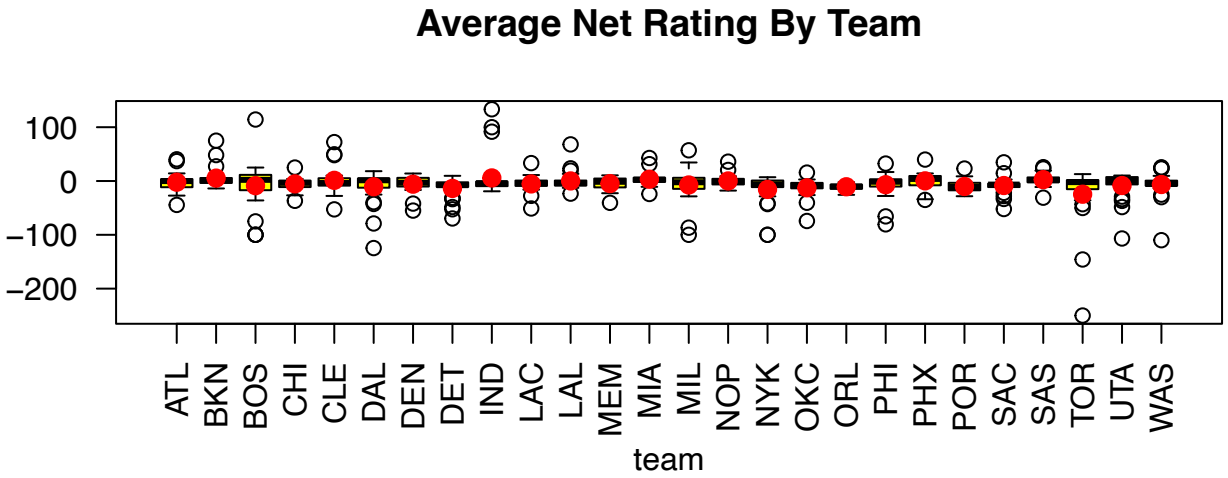


Although my data does not meet the two assumptions of discriminant analysis (multivariate normality and equal covariance matrices), I was able to use quadratic discriminant analysis to classify the players into their respective teams with 94% accuracy. This is much better than the 24% accuracy I got from linear discriminant analysis. I also found that there is statistical evidence that the multivariate group means are different, which suggests that discriminant analysis is effective in classifying the players into their respective teams. The first two discriminant functions are significant, and the first function has the most relative discriminating power. The most discriminating variables are average plus-minus, average offensive rating, average defensive rating, average net rating, and average pace. I will be able to leverage this information in my next multivariate analysis: MANOVA. Lastly, I was able to plot the results in the space spanned by these variables to show which regions are assigned to each group. This is a very interesting result because it shows that the players are being classified into their teams based on their performance metrics.

3. Multivariate Analysis of Variance (MANOVA)

MANOVA is a technique used to determine whether there are statistically significant differences between the means of two or more groups. I will be using MANOVA to determine whether there are statistically significant differences in the team-level means of total points and average net rating. While I could run MANOVA on all of my metrics, I found in my discriminant analysis that average net rating is a defining characteristic of players. I also know that points are inherently a defining characteristic because it is literally what leads teams to win or lose games. So I will only be analyzing these two metrics and will start by creating boxplots of each of these metrics for all teams (note: this analysis is excluding HOU, GSW, MIN, and CHA because they have fewer than 22 observations which is required for some of the functions).





It looks like BOS, MIL, and WAS may have fewer points scored on average than the other teams, while BKN, DEN, LAC, and NOP may have more points scored. Overall, the square root (because the data is transformed) of the points scored is pretty comparable across all the teams. Average net rating is even harder to say which teams might be different just by looking at the box plot. I would saw that DET and TOR look like they have lower average net ratings, while BKN and IND may be a bit on the higher end. I can now run univariate ANOVA, as well as one-way MANOVA to see if there are statistically significant differences in the combination of total points and average net rating by team.

```
##  Response total.pts :
##              Df Sum Sq Mean Sq F value Pr(>F)
## team         25    787  31.465  0.2438      1
## Residuals   616  79509 129.073
##
##  Response avg.net.rat :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## team         25  29929 1197.17  1.9748 0.003367 **
## Residuals   616 373427  606.21
## ---
## 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 2.23899   688.49      2    615 < 2.2e-16 ***
## team         25 0.08082     1.99     25    616  0.003018 **
```

```
## Residuals   616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

My univariate results say that total.pts do not significantly differ by team (p=1), but avg.net.rat does (p=0.003). This is interesting because just looking at the boxplots, I would have guessed the opposite, but running statistical analysis tells me something different. This is however consistent with my findings from the discriminant analysis. My multivariate results say that there is a significant difference in the combination of total.pts and avg.net.rat by team (p=0.003) using Roy's (which is based on the direction of maximum discrimination). So I will not perform a univariate contrast for total points since it is statistically insignificant, but I will perform a univariate contrast for average net rating, as well as a multivariate contrast for both of the metrics.

```
##
## Sum of squares and products for the hypothesis:
##              total.pts avg.net.rat
## total.pts    0.8673953    57.53196
## avg.net.rat 57.5319615  3815.93798
##
## Sum of squares and products for error:
##             total.pts avg.net.rat
## total.pts    79508.72    35785.07
## avg.net.rat  35785.07   373426.89
##
## Multivariate Tests:
##                   Df test stat approx F num Df den Df   Pr(>F)
## Pillai             1 0.0104357 3.242825      2    615 0.039722 *
## Wilks              1 0.9895643 3.242825      2    615 0.039722 *
## Hotelling-Lawley   1 0.0105458 3.242825      2    615 0.039722 *
## Roy                1 0.0105458 3.242825      2    615 0.039722 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Sum of squares and products for the hypothesis:
##              total.pts avg.net.rat
## total.pts     3.082561    161.6733
## avg.net.rat 161.673324   8479.3987
##
## Sum of squares and products for error:
##             total.pts avg.net.rat
## total.pts    79508.72    35785.07
## avg.net.rat  35785.07   373426.89
##
## Multivariate Tests:
##                   Df test stat approx F num Df den Df    Pr(>F)
## Pillai             1 0.0228303 7.184345      2    615 0.0008237 ***
## Wilks              1 0.9771697 7.184345      2    615 0.0008237 ***
## Hotelling-Lawley   1 0.0233637 7.184345      2    615 0.0008237 ***
## Roy                1 0.0233637 7.184345      2    615 0.0008237 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I ran multivariate contrasts on all of the teams, but only included the output from those that were statistically significant. Very interestingly, the combination of total.pts and avg.net.rat slightly sets IND apart from the other teams (p=0.03) and TOR is VERY statistically different from the other teams (p<0.001).

I would have never guessed this from the boxplots, but it is cool to see that the statistical analysis is able to pick up on these differences in multi-dimensional space. Now for the univariate contrast for average net rating alone.

```
## Linear hypothesis test
##
## Hypothesis:
## 25 teamBKN - teamBOS - teamCHI - teamCLE - teamDAL - teamDEN - teamDET - teamIND - teamLAC - teamLAL
##
## Model 1: restricted model
## Model 2: avg.net.rat ~ team
##
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    617 376512
## 2    616 373427  1      3085 5.089 0.02443 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Linear hypothesis test
##
## Hypothesis:
## - teamBKN - teamBOS - teamCHI - teamCLE - teamDAL - teamDEN - teamDET  + 25 teamIND - teamLAC - teaml
##
## Model 1: restricted model
## Model 2: avg.net.rat ~ team
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    617 377243
## 2    616 373427  1    3815.9 6.2947 0.01237 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

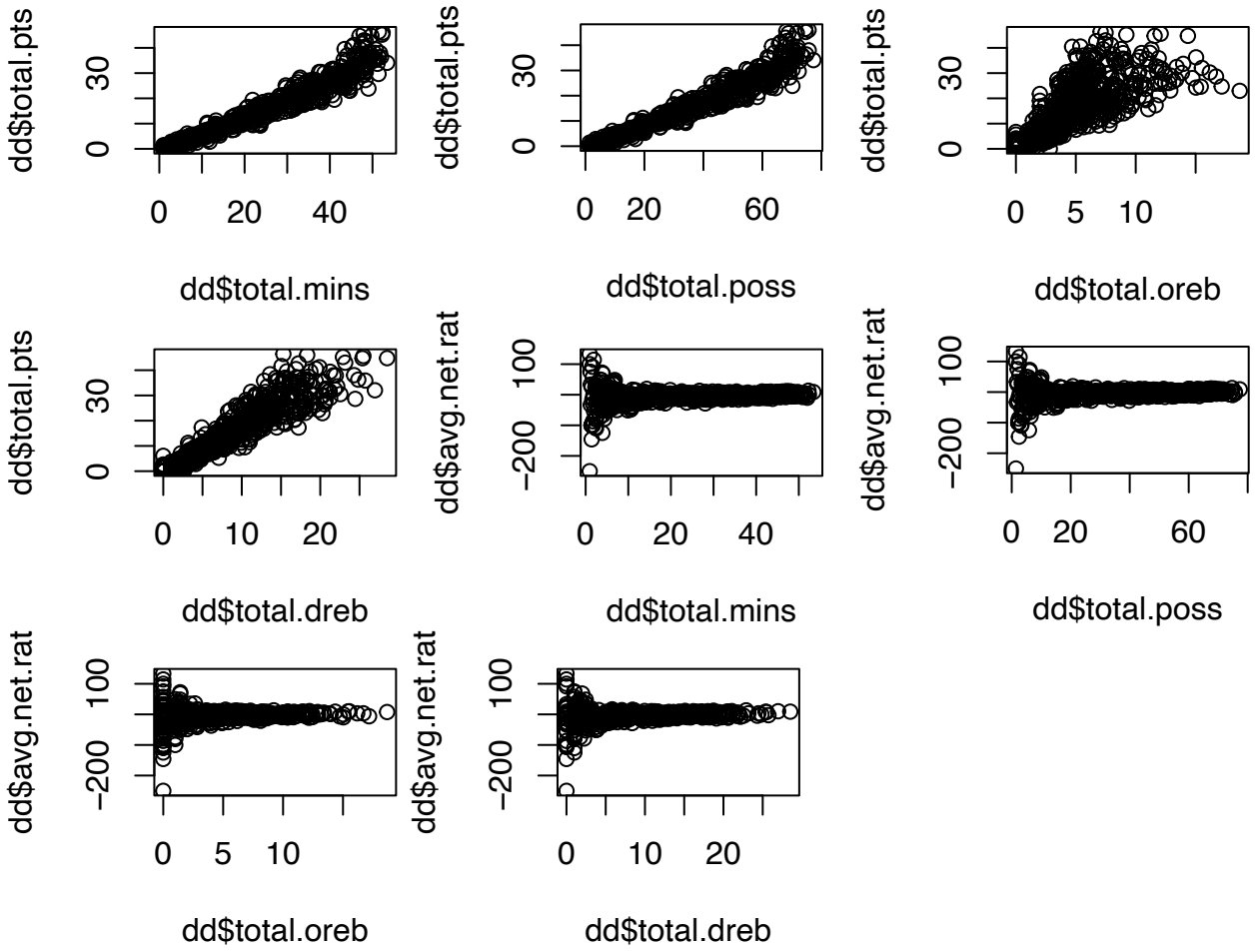
```
## Linear hypothesis test
##
## Hypothesis:
## - teamBKN - teamBOS - teamCHI - teamCLE - teamDAL - teamDEN - teamDET - teamIND - teamLAC - teamLAL -
##
## Model 1: restricted model
## Model 2: avg.net.rat ~ team
##
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    617 375768
## 2    616 373427  1      2341 3.8616 0.04985 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Linear hypothesis test
##
## Hypothesis:
## - teamBKN - teamBOS - teamCHI - teamCLE - teamDAL - teamDEN - teamDET - teamIND - teamLAC - teamLAL -
##
## Model 1: restricted model
## Model 2: avg.net.rat ~ team
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    617 381906
```

```
## 2    616 373427  1    8479.4 13.988 0.0002012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, I ran univariate contrasts for avg.net.rat on all of the teams, but only included the output from those that were statistically significant: BKN, IND, NYK, and TOR are different from the other teams with respect to avg.net.rat. This makes sense based on my observations that TOR looked like a lower average net rating and BKN looked a bit on the higher end, so it is cool to validate that hypothesis. Because this test is the one that had the most significant number of p-values, these are the ones that I will feed into a multiple comparison correction in a few paragraphs. Since I just ran 3*26=78 tests, I do think it is very important to make some corrections because it is possible that some of these were significant just by chance.

Before I get into assumptions and corrections, I am going to add total minutes played, total possessions, total offensive rebounds, and total defensive rebounds to my model and fit as a multiple-response linear model. The reason I am adding the former two is because I think they could be predictors of total points, and the reason I am adding the latter two is because I think they could be predictors of average net rating. I will start with some plots to see if there are linear relationships between my covariates and my responses.



The relationships that look the most linear are between total minutes played and total points, total possessions and total points, and total defensive rebounds and total points. I will now perform univariate and multivariate analyses of variance by team.

```
##  Response total.pts :
##              Df Sum Sq Mean Sq    F value    Pr(>F)
## team         25    787      31     4.6978 1.553e-12 ***
## total.mins    1  74682   74682 11150.3735 < 2.2e-16 ***
```

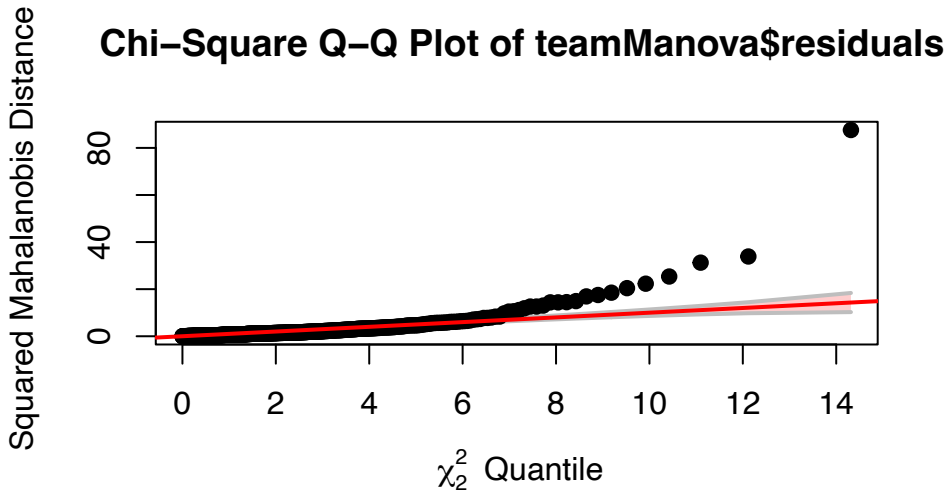
```
## total.poss    1     61      61     9.0937  0.002671 **
## total.oreb    1     12      12     1.7590  0.185244
## total.dreb    1    655     655    97.7909 < 2.2e-16 ***
## Residuals   612   4099       7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response avg.net.rat :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## team         25  29929  1197.2  2.0650  0.001856 **
## total.mins    1  16112 16111.8 27.7915 1.876e-07 ***
## total.poss    1   2197  2196.7  3.7892  0.052042 .
## total.oreb    1    312   311.8  0.5378  0.463625
## total.dreb    1      6     6.0  0.0104  0.918994
## Residuals   612 354801   579.7
## ---
## 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 37.986  11604.8      2    611 < 2.2e-16 ***
## team         25  0.203      5.0     25    612 1.463e-13 ***
## total.mins    1 18.222   5566.8      2    611 < 2.2e-16 ***
## total.poss    1  0.022      6.7      2    611  0.001369 **
## total.oreb    1  0.004      1.2      2    611  0.306308
## total.dreb    1  0.160     48.9      2    611 < 2.2e-16 ***
## Residuals   612
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

My univariate results say that team, total.mins, total.poss, and total.dreb are all significant predictors of total.pts ($p<0.01$ for all of them), i.e., total.pts significantly differ by each of those predictors (not total.oreb). I find this very interesting because team was not a significant predictor of total.pts in the previous model, but now that I added continuous variables to the model, it is. My univariate results also say that team and total.mins are significant predictors of avg.net.rat ($p<0.01$ for both), i.e., avg.net.rat significantly differs by each of those predictors. While it is good to see that team has remained a significant predictor of avg.net.rat, it is interesting to see that total.mins is the only significant continuous predictor which is not what I expected.

My multivariate results are now very different from last time as well. All three methods say that there is a significant difference in the combination of total.pts and avg.net.rat by team, total.mins, total.poss, and total.dreb ($p<0.01$ for all of them). It is cool to see how adding different predictors changes which combinations are significant!

I will now check model assumptions by making a chi-square quantile plot of the residuals.



Unfortunately, the residuals for my MANOVA model do not appear to be multivariate normal. I did attempt to remove the teams that did not follow a multivariate normal distribution, but the residuals still did not appear to be multivariate normal, so I am not including that part in this analysis. This means that I have to be careful about the interpretability of my results since the model assumptions are not met. I will now perform a multiple comparison correction on the p-values from the univariate contrasts for avg.net.rat. These tests produced 4 significantly different teams so I think it is important to try to adjust the p-values. I will use the Bonferroni, Holm, and Hochberg methods to adjust the p-values.

```
##  [1] 1.0000000 0.6351800 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [8] 1.0000000 0.3216200 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [15] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [22] 1.0000000 1.0000000 0.0052312 1.0000000 1.0000000
```

```
##  [1] 1.0000000 0.5863200 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [8] 1.0000000 0.3092500 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [15] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [22] 1.0000000 1.0000000 0.0052312 1.0000000 1.0000000
```

```
##  [1] 0.9791000 0.5863200 0.9791000 0.9791000 0.9791000 0.9791000 0.9791000
##  [8] 0.9791000 0.3092500 0.9791000 0.9791000 0.9791000 0.9791000 0.9791000
## [15] 0.9791000 0.9791000 0.9791000 0.9791000 0.9791000 0.9791000 0.9791000
## [22] 0.9791000 0.9791000 0.0052312 0.9791000 0.9791000
```

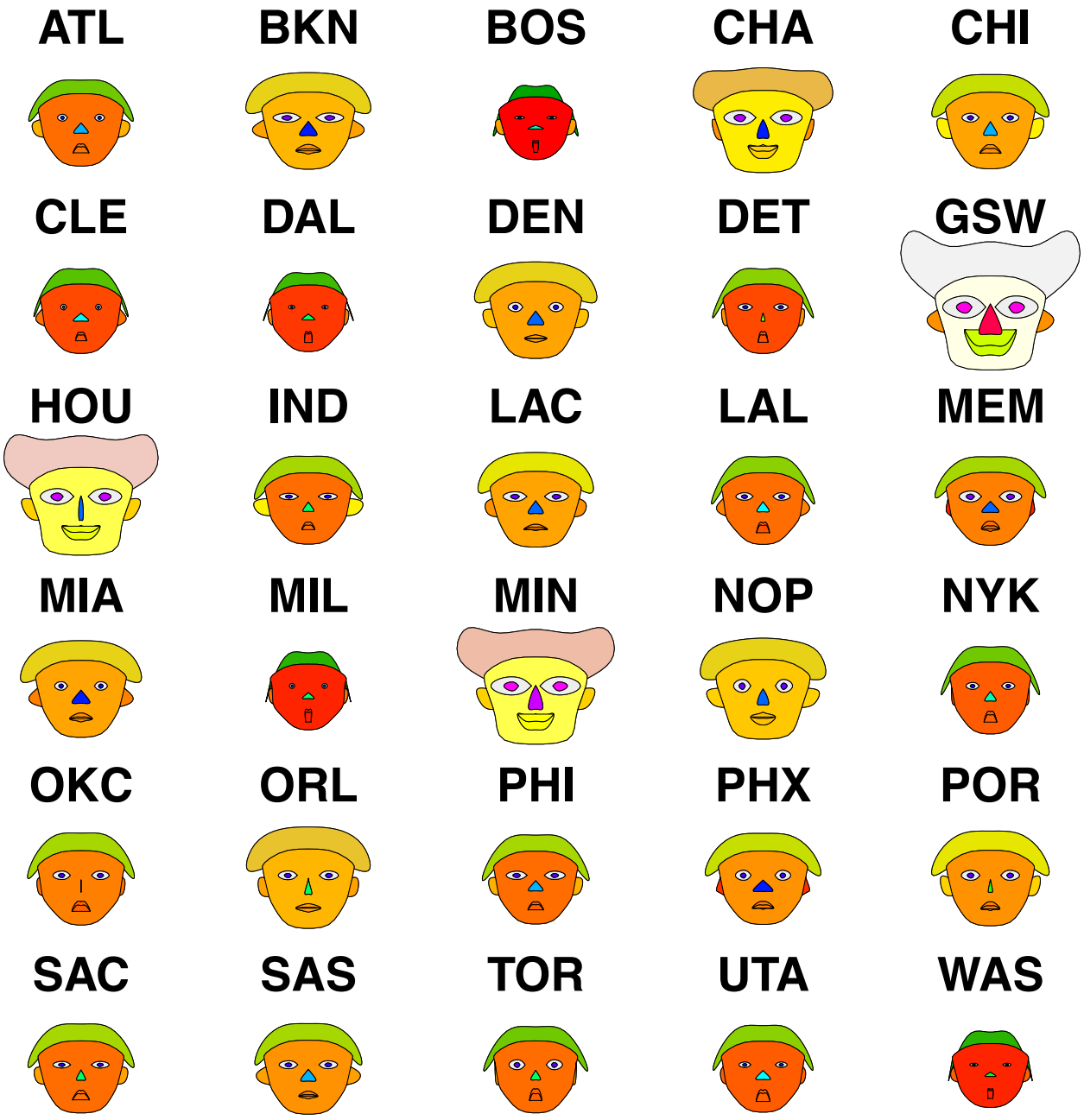
For all three of the adjustments, the only one that remains significant is TOR. This makes sense because of how low the p-value for TOR was with respect to avg.net.rat. After adjusting the p-values, I would now say that TOR is the only team that is statistically different from the other team's average net rating.

To summarize my multivariate analysis of variance, I found that the combination of total.pts and avg.net.rat significantly differ by team, total.mins, total.poss, and total.dreb. I performed a multiple comparison correction on the p-values from the univariate contrasts for avg.net.rat, and after adjusting the p-values, I found that TOR's average net rating is the only one that is statistically different from the other teams. I also found that the residuals from my MANOVA model do not appear to be multivariate normal, so I have to be careful about the interpretability of these results.

4. Cluster Analysis

Cluster analysis is a technique used to group observations into clusters based on their characteristics. I will be using cluster analysis to group NBA teams based on their performance metrics. I will start by examining the data to see if there are any patterns that I can identify visually. I will then perform hierarchical clustering with Euclidean distance, Manhattan distance, and squared maximum distance, as well as complete

agglomeration, average agglomeration, and Ward's agglomeration methods to group the teams. I will start with the whimsical technique.




```
## effect of variables:
##  modified item       Var
##  "height of face   " "total.mins"
##  "width of face    " "avg.fgm"
##  "structure of face" "avg.ftm"
##  "height of mouth  " "avg.oreb"
##  "width of mouth   " "avg.dreb"
##  "smiling          " "avg.ast"
##  "height of eyes   " "avg.stl"
##  "width of eyes    " "avg.blk"
```

```
##  "height of hair   " "avg.to"
##  "width of hair   "  "avg.pf"
##  "style of hair   "  "avg.pts"
##  "height of nose  "  "avg.poss"
##  "width of nose   "  "avg.pm"
##  "width of ear    "  "avg.off.rat"
##  "height of ear   "  "avg.def.rat"
```

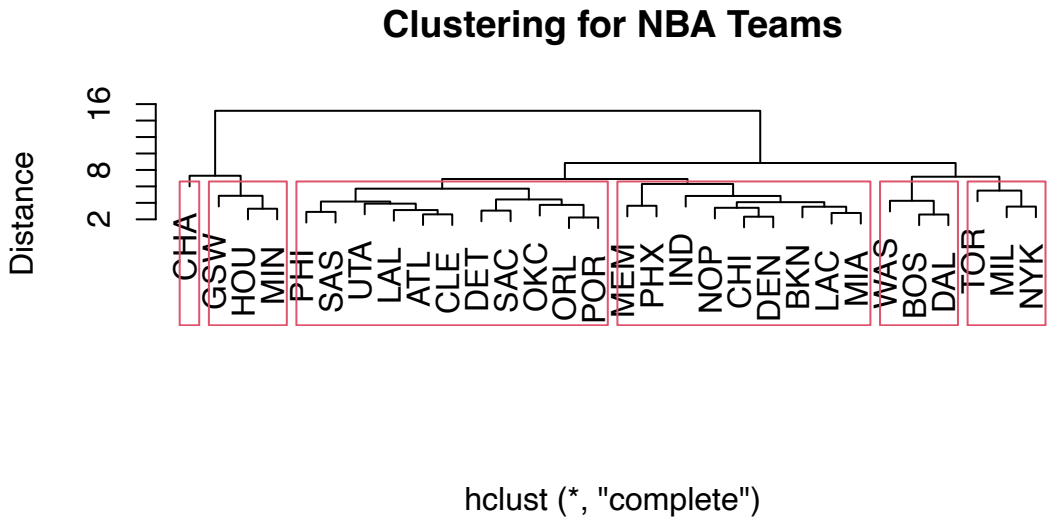
Groupings I notice from the faces plot (there are some faces that are similar to the grouping but I didn't include - these are just some of my initial observations I will try to validate later in my analysis):

1. Red faces with green, concave down hair and oval mouths: BOS, CLE, DAL, DET, MIL, WAS. Based on variable effects that I can see in the output, these teams have distinct structure of face (avg.ftm), smiling (avg.ast), width of hair (avg.pf), and style of hair (avg.pts).

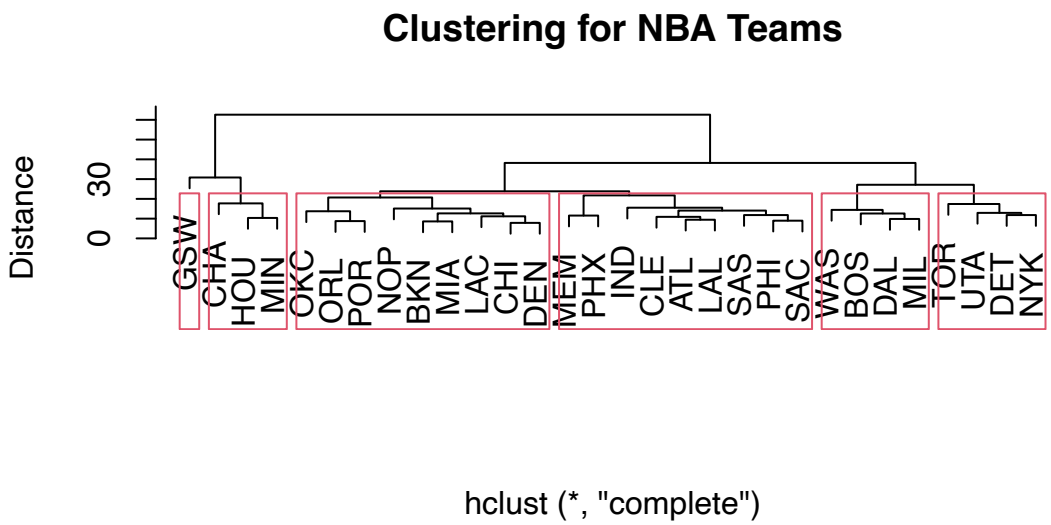
2. Orange/golden faces with tannish/yellow hair and blank expressions: BKN, DEN, LAC, MIA, NOP, ORL. These teams have distinct structure of face (avg.ftm), smiling (avg.ast), width of hair (avg.pf), and style of hair (avg.pts).

3. Yellow faces with with big purple eyes and large concave up hair: CHA, HOU, MIN. These teams have distinct structure of face (avg.ftm), smiling (avg.ast), height of eyes (avg.stl), width of hair (avg.pf), and style of hair (avg.pts).

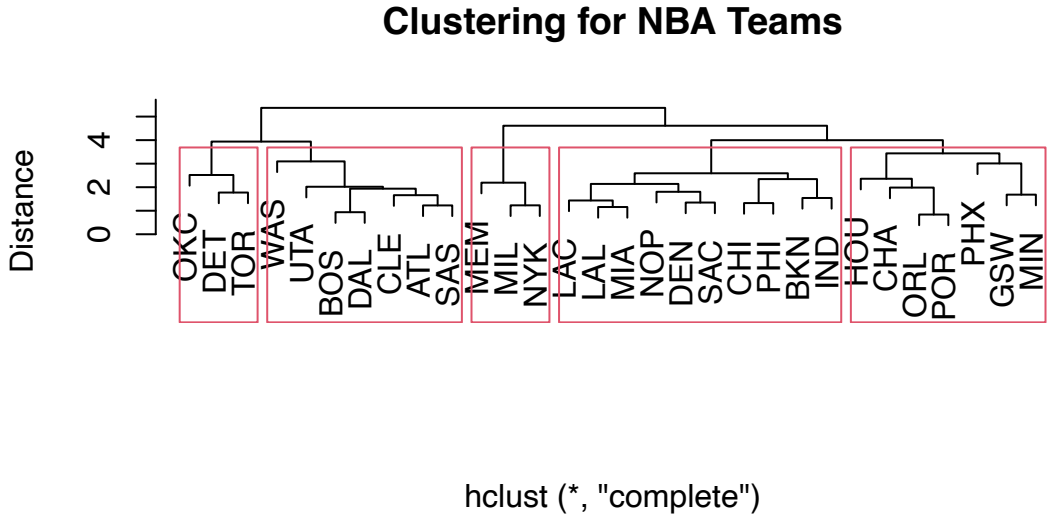
Now, I will try hierarchical clustering with Euclidean distance (default) and complete agglomeration (default).



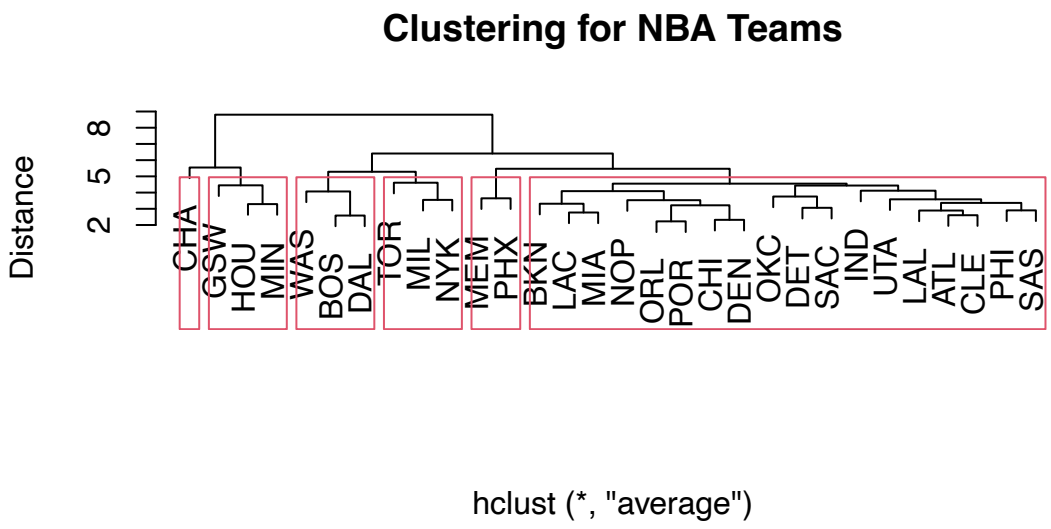
I think I can see roughly six main groups, that could definitely be broken down into more groups depending on what I find in the rest of my analysis. First I notice that CHA is in a group on its own. Based on my initial theories of groupings, I notice that BOS & DAL & WAS are in the same group, BKN & DEN & LAC & NOP & MIA are in the same group, and HOU & MIN are in the same group. This is similar to some patterns I saw in the faces. Next, I will try hierarchical clustering with Manhattan distance and complete agglomeration (default).



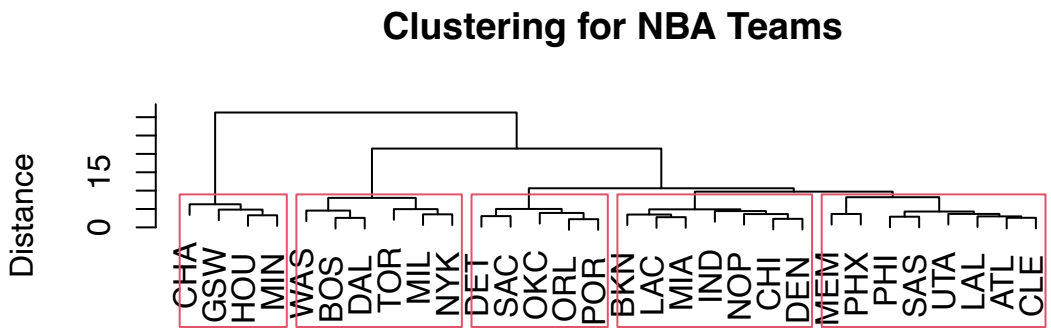
I notice GSW is an outlier on its own this time. BOS & DAL & MIL & WAS are in the same group, and also are identified as the only 4 teams of that group. For my last distance method, I will try hierarchical clustering with squared maximum distance and complete agglomeration (default).



This time, no team is a unique group on its own. BOS & DAL & WAS are still in the same group but there are other teams in that group. Now, I will change the agglomeration methods to try average and Ward's (with Euclidean distance as default) because they were suggested as possibly the better methods to use.

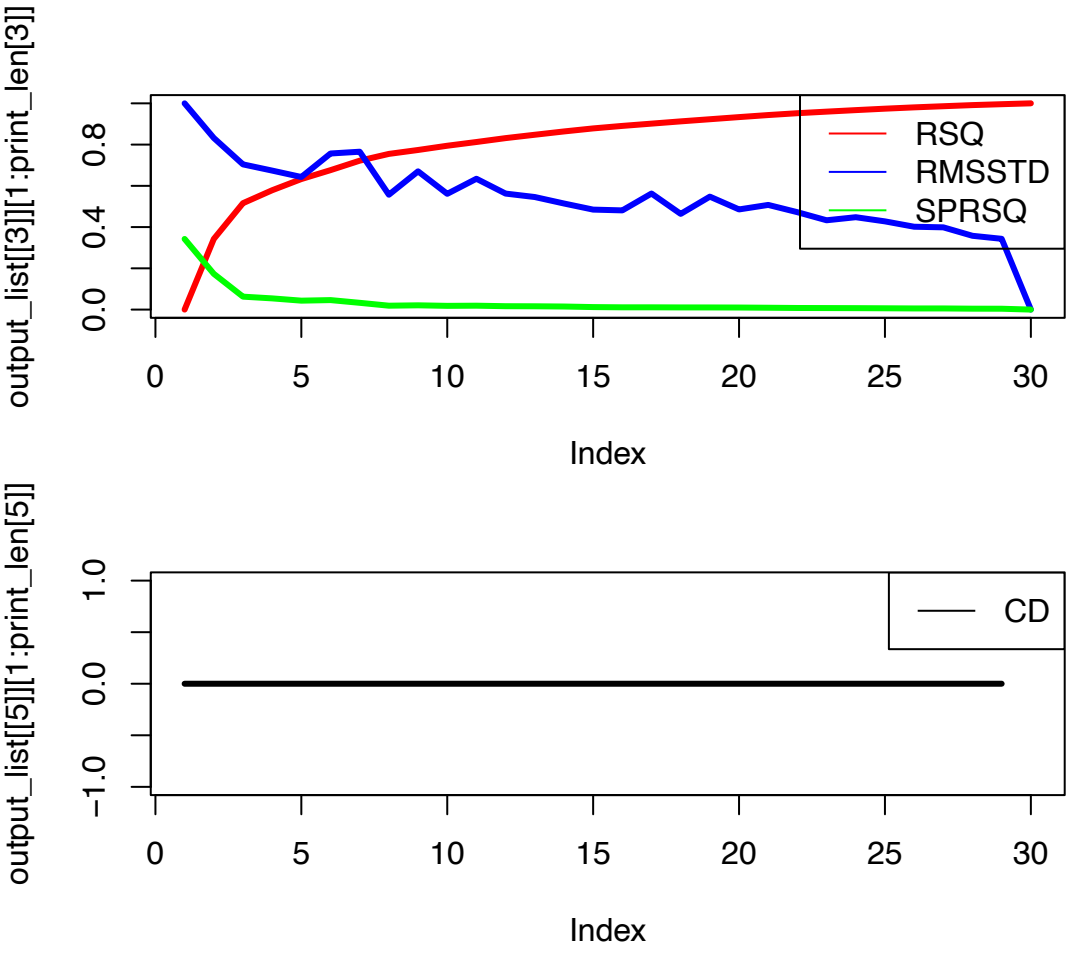


This one was interesting because I could see it being five or six groups but the rectangles are not grouping main groups that I would have considered just by looking at it. CHA is back to being its own group.



This looks like the most evenly disbursed grouping. There is not a group that consists of only 1, 2, or even 3 teams and the groupings visually make sense on the dendrogram. Because the groupings are so evenly disbursed, I am going to use the Euclidean distance and the Ward's agglomeration method to plot cluster distance.

```
## [1] "Creating Distance Matrix using euclidean"
## [1] "Clustering using ward.D"
## [1] "Clustering Complete. Access the Cluster object in first element of output"
## [1] "Calculating RMSSTD"
## [1] "RMSSTD Done. Access in Element 2"
## [1] "Calculating RSQ"
## [1] "RSQ Done. Access in Element 3"
## [1] "Calculating SPRSQ"
## [1] "SPRSQ Done. Access in Element 4"
## [1] "Calculating Cluster Dist. "
## [1] "CD Done. Access in Element 5"
```



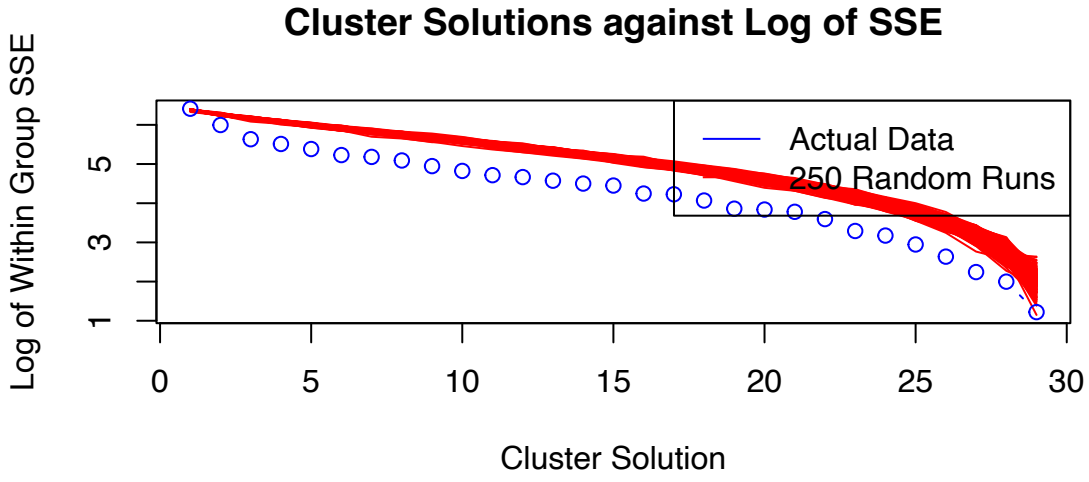

```
## [[1]]
##
## Call:
## hclust(d = dist1, method = clus_m)
##
## Cluster method   : ward.D
## Distance         : euclidean
## Number of objects: 30
##
##
## [[2]]
##  [1] 1.0000000 0.8321069 0.7040233 0.6739420 0.6428406 0.7567717 0.7654518
##  [8] 0.5569411 0.6703415 0.5613516 0.6343008 0.5624146 0.5458363 0.5143823
## [15] 0.4847617 0.4809415 0.5626806 0.4640519 0.5475747 0.4856348 0.5074622
## [22] 0.4722595 0.4330356 0.4478739 0.4276677 0.4017609 0.3987180 0.3577585
## [29] 0.3432016 0.0000000
##
## [[3]]
##  [1] 0.0000000 0.3424898 0.5159105 0.5785941 0.6329172 0.6761293 0.7220666
##  [8] 0.7549313 0.7738180 0.7944690 0.8125793 0.8314468 0.8477794 0.8637492
## [15] 0.8785230 0.8906678 0.9016635 0.9125811 0.9230189 0.9333581 0.9433161
## [22] 0.9521961 0.9598867 0.9672532 0.9741701 0.9804770 0.9860429 0.9915249
## [29] 0.9959384 1.0000000
##
```

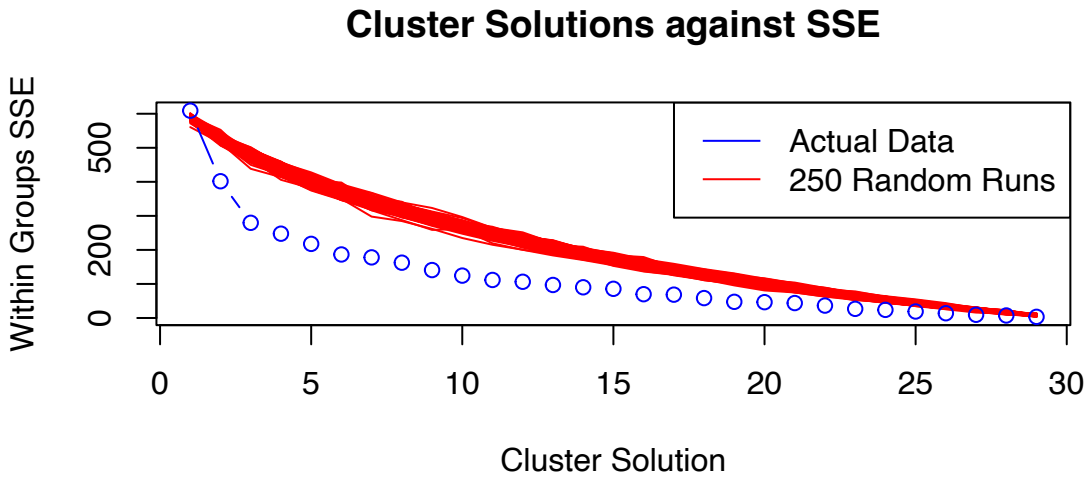
```
## [[4]]
##  [1] 0.342489830 0.173420719 0.062683601 0.054323026 0.043212163 0.045937295
##  [7] 0.032864624 0.018886790 0.020650944 0.018110337 0.018867492 0.016332561
## [13] 0.015969772 0.014773808 0.012144845 0.010995680 0.010917567 0.010437832
## [19] 0.010339244 0.009958020 0.008879925 0.007690657 0.007366474 0.006916931
## [25] 0.006306886 0.005565926 0.005481934 0.004413487 0.004061631 0.000000000
##
## [[5]]
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

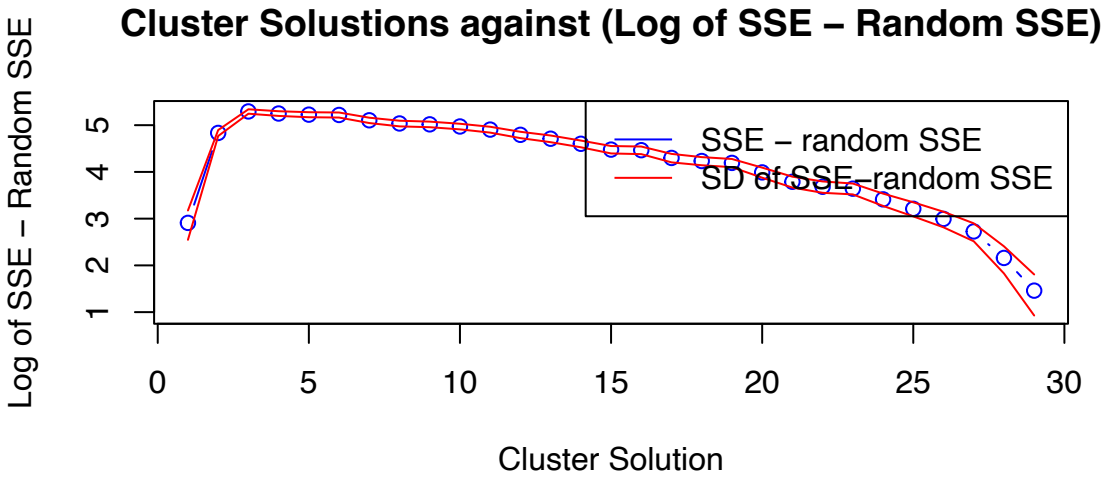
Based on this visual, I think I want to retain 5 groups. I want the root-mean-square standard deviation to be small, and there is a local minimum at 5. I want the R-squared to be close to 1, and it crosses the 60% threshold at 5. I also want the semi-partial R-squared to be near 0, and it looks like there is an elbow at 4, so it is more leveled-out at 5. Cluster distance is 0 everywhere for this method, which I think is good because I want this to be small as well. I will now run k-means clustering on my data to see how it compares to this.

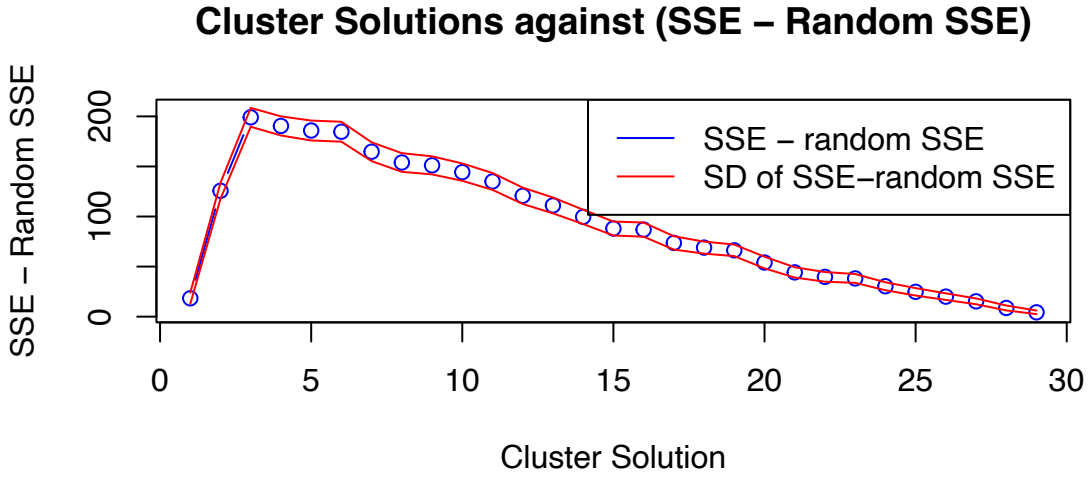
```
## [1] "Teams in Cluster  1"
## [1] "DET" "NYK" "TOR"
## [1] " "
## [1] "Teams in Cluster  2"
## [1] "BOS" "DAL" "MIL" "UTA" "WAS"
## [1] " "
## [1] "Teams in Cluster  3"
## [1] "CHA" "GSW" "HOU" "MIN"
## [1] " "
## [1] "Teams in Cluster  4"
##  [1] "ATL" "BKN" "CLE" "IND" "LAL" "MEM" "MIA" "PHI" "PHX" "SAS"
## [1] " "
## [1] "Teams in Cluster  5"
## [1] "CHI" "DEN" "LAC" "NOP" "OKC" "ORL" "POR" "SAC"
## [1] " "
```

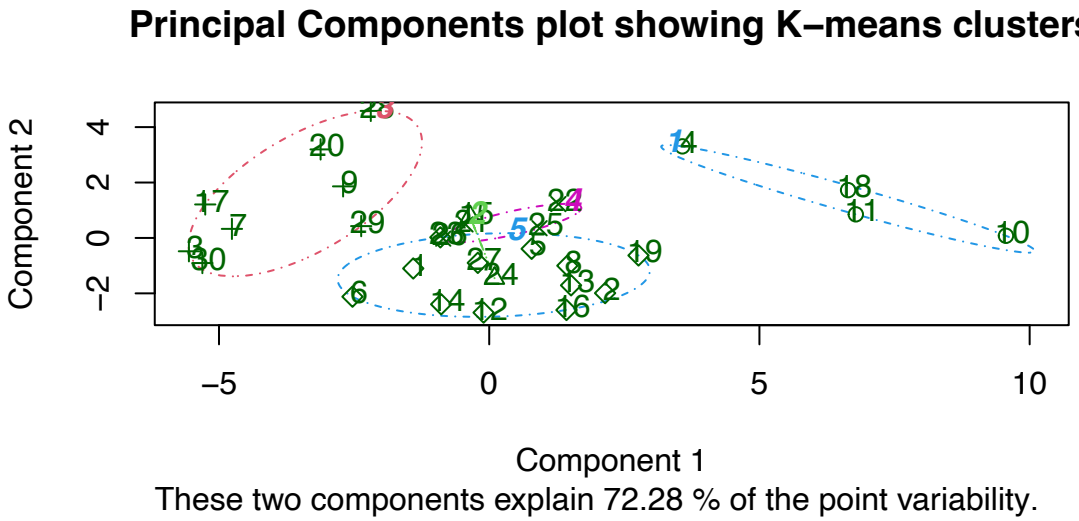
These clusters identified by k-means clustering are relatively similar to the clusters identified in the dendrogram using Euclidean distance and Ward's agglomeration methods. Next I will use the modified script by Matt Peeples to produce a screeplot-like diagram with randomized comparison based on randomization within columns (i.e. as if points had been randomly assigned data values, one from each column). This keeps total internal SS the same.

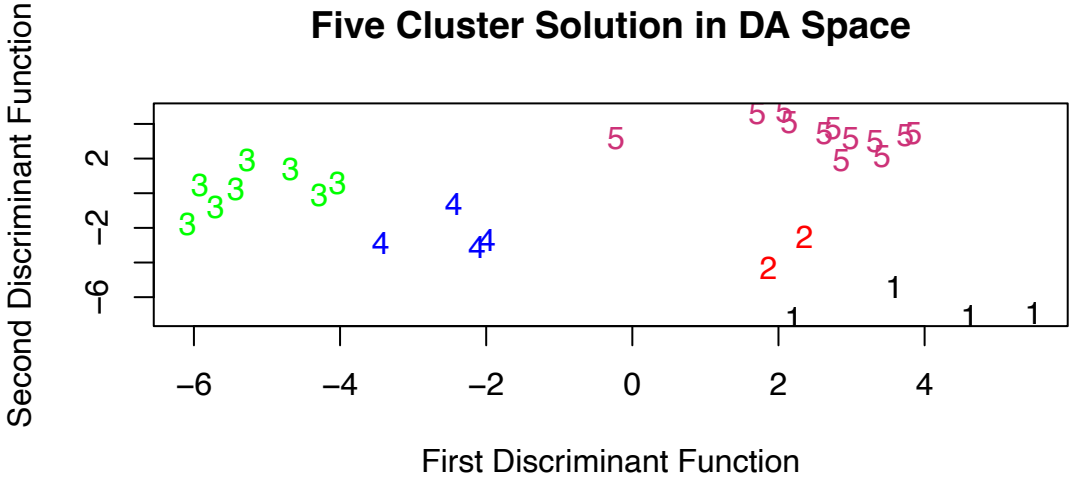



Cluster Solutions against SSE
Within Groups SSE
0
200
500
0
5
10
15
20
25
30
Cluster Solution
Actual Data
250 Random Runs

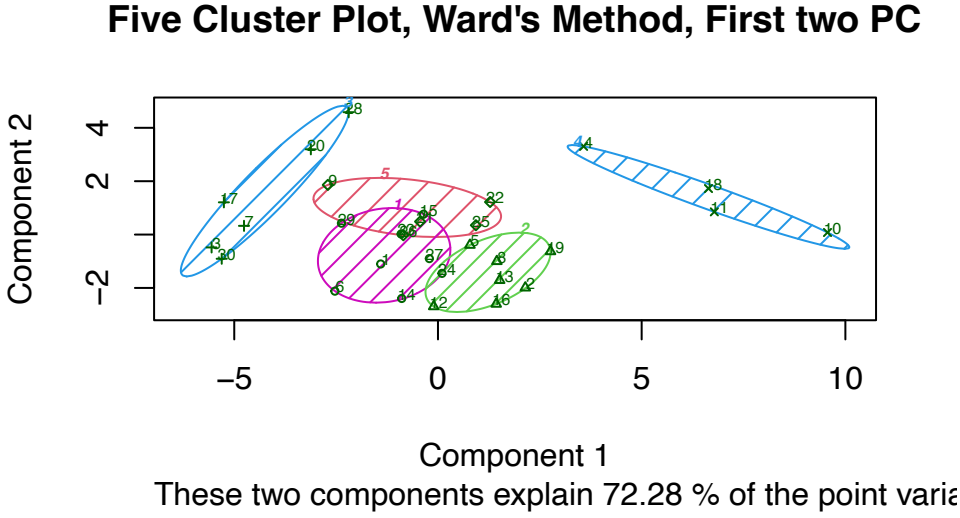

Cluster Solustions against (Log of SSE – Random SSE)
Log of SSE – Random SSE
1
2
3
4
5
0
5
10
15
20
25
30
Cluster Solution
SSE – random SSE
SD of SSE–random SSE

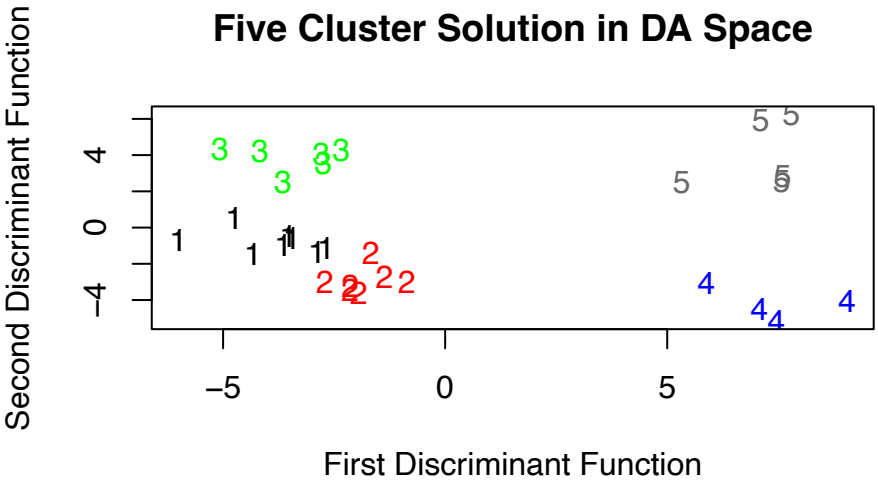

Cluster Solutions against (SSE – Random SSE)
SSE – Random SSE
0
100
200
0
5
10
15
20
25
30
Cluster Solution
SSE – random SSE
SD of SSE–random SSE





The actual data is less than the random runs which is good. Looking at the Within Groups SSE plot, I see that the elbow is at 3 or 4. There is not much change after k=5, so I will continue to keep 5 groups for the rest of the analysis. Lastly, similar to the plots of the K-means clusters in PCA & DA space shown above, I will produce plots of my results in discriminant analysis space and in PCA space for the clusters identified in the dendrogram that used Euclidean distance and Ward's agglomeration.





I am comfortable with the number of groups being 5 based on what I found above. I think that the groups are distinct and make sense based on the variables represented by my first to PC's. In PCA, I found that the first PC represents time with the ball and the second PC represents a team's ability to score and prevent their opponent from scoring. Even though groups 1 and 5 overlap a little on the PC plot, I think that the groups are generally separated pretty well based on these variables. So overall, based on my previous exploration, I am sticking with 5 groups.

I am taking my final groups from the k-means clustering output:

- Teams in cluster 1: DET, NYK, TOR
- Teams in cluster 2: ATL, BKN, CHI, DEN, IND, LAC, LAL, MEM, MIA, NOP, PHI, PHX, SAS
- Teams in cluster 3: CHA, GSW, HOU, MIN
- Teams in cluster 4: OKC, ORL, POR, SAC
- Teams in cluster 5: BOS, CLE, DAL, MIL, UTA, WAS

I can tell that these clusters are not based off of the location or conference of the team since they are all mixed together. So performance would make the most sense as to what separates the teams amongst these clusters. Comparing with my PC plot above:

- Cluster 1 seems to have low time with the ball and a low ability to score. This makes sense with my MANOVA findings that TOR & DET have lower average net ratings (TOR statistically so).
- Cluster 2 has more time with the ball but still a low ability to score. This makes sense with what actually happened in the season because teams like MIA, MEM, and PHX made it far in the season (but GSW won the championship).
- Cluster 3 has the lowest time with the ball and the most variability in their ability to score. This was interesting to me because like I mentioned, GSW won the championship so I would expect them to be in the top right of the plot.
- Cluster 4 is best positioned for both variables- more time with the ball and better ability to score. This does not make sense to me because these four teams did not do well in the 2021-2022 season. So this cluster makes sense that they're grouped together but not when I put them on the PC plot.
- Cluster 5 spans an average amount of time with the ball and has a relatively good ability to score. This also does not makes sense with my MANOVA observations that BOS, MIL, and WAS had fewer points

scored on average.

Overall, I do think that the way the teams are grouped together makes sense based off of their performances in the 2021-2022 season. However, when trying to makes sense of them on the PC plot, some of them make sense based on the variables, but some of them make no sense at all.

Conclusions and Discussion

The comprehensive multivariate analysis of NBA player data conducted in this project provides valuable insights into the dynamics of the game. By leveraging advanced statistical techniques, this study tackled fundamental questions such as the drivers of player variations and the classification of teams based on statistical profiles.

The analyses revealed several key findings about the structure and patterns of player-level basketball data. Principal Component Analysis (PCA) interrogated the underlying dimensions of player performance, highlighting the significance of factors such as time with the ball, scoring ability, shooting accuracy, and overall impact on the game. Discriminant Analysis enabled accurate classification of players into teams based on their metrics, showcasing the effectiveness of statistical modeling in team categorization. Multivariate Analysis of Variance (MANOVA) provided valuable insights into the differences in team-level means of average net rating, shedding light on the distinct characteristics of different NBA teams. Cluster Analysis further examined teams by grouping them based on performance profiles, offering nuanced perspectives on team compositions and competitive strategies.

This project dives deep into the statistical intricacies of basketball games and the resulting data. Anyone with a stake in the game can make more informed and strategic decisions by leveraging the insights generated from this study. Coaches can optimize player rotations and game strategies based on the player's stats, while analysts can identify key performance indicators and predictive models for player outcomes. Team management can leverage the clustering analysis to understand the competitive landscape and develop targeted strategies for team improvement. Overall, the project represents a significant contribution to the intersection of sports analytics and statistical modeling, enhancing our understanding of the game and its complexities.

Points for Further Analysis

The project opens up several avenues for further analysis and exploration in the domain of sports analytics. Some potential directions for future research include:

- Explore additional transformations to help meet the assumptions of multivariate analysis techniques, such as the data following multivariate normal distributions
- Incorporate more seasons worth of data to capture longitudinal trends and performance variations over time
- Integrate position data to analyze player performance based on their roles and responsibilities on the court
- Investigate the impact of game strategies, coaching decisions, and player interactions on team outcomes
- Develop predictive models for performance based on advanced machine learning algorithms and statistical techniques

By delving deeper into these areas, researchers can gain a more comprehensive understanding of player and team dynamics in basketball games, enabling more accurate predictions, strategic insights, and data-driven decision-making in the sports industry.