College Basketball- A deeper review of power rankings

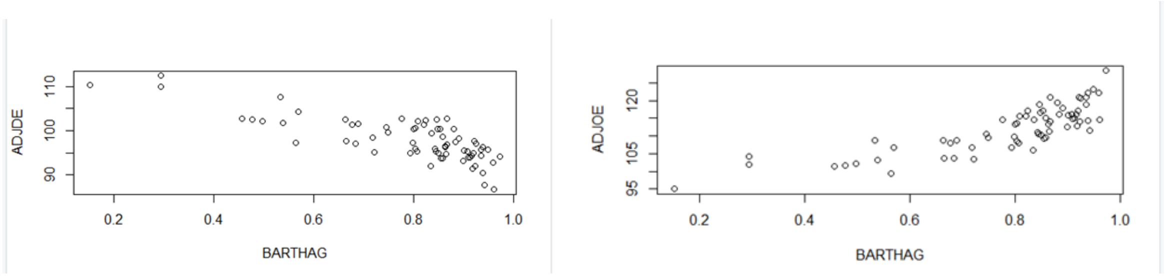
Aakash Patel, Hima Barla, Devindra Sawh, Brooke Bilton, Carrie Stienen

Executive Summary

Advanced Stats are a way to study basketball through objective analysis rather than just looking at wins,losses or box scores.   In any sport, there is a concept of power rankings and the rankings help justify which team is better and should win. But in order to fully understand the power rating, one would want to peel back the layers and determine what statistics are the best predictors. This paper will focus on the statistics from the NCAA College Basketball season of 2018. To determine what stats are the best predictors, we decided to perform a series of analysis on the dataset which include Multiple Regression, Principal Component Analysis(PCA), Common Factor Analysis (CFA), Correspondence Analysis, and K-mean.

A linear relationship exists between the ADJOE and the power rating (BARTHAG) – when the ADJOE rates go up, the power rating (BARTHAG) goes up with a positive slope and when ADJOD and ADJDE decreased and the power rating(BARTHAG) increased with a negative slope.

Figure 0



The variables that corresponded the most strongly to power ranking ones for efficiencies. It's our finding that if a team wants to best improve their power ranking, they should focus very strongly on their efficiency. Offensive and defensive efficiency were shown to be about equally important. That is to say that a team's ability to score points and their ability to keep the other team from scoring points is equally important. Both offense and defense need to be focused on and improved equally in order to improve power ranking. Some of the limitations were that there only one year of data is used. For future research make sure to use multiple years of data to find out the predictor variable.

College Basketball: A deeper review of power rankings

Abstract

This paper takes a look at the most commonly recorded statistics in college basketball for the year 2018 in order to identify the best way to predict a team's power ranking which is basically the chance of beating an average Division I team. Power rankings are a way to rank teams and help to predict which teams are more likely to win or be seeded in the March Madness tournament. Through different methods such as regression, factor analysis, principal component analysis and canonical correlation analysis, this research was intended to figure out which variables were most significant and relevant when it comes to predicting a team's power rating. Knowing this kind of information might help to analyze the winning teams performances and determine what their strengths are. In this paper, we present some background applications that other researchers have conducted on similar data as well as the methods that we used and our results. Canonical correlation was done in order to validate the power ranking was a good predictor variable against all of the basketball statistics that were recorded. Clustering was conducted to see if the data had any patterns or segments. Factor analysis was used to describe the variability among correlated variables as well as principal component analysis as a variable selection method. Lastly, regression allowed for us to narrow down and identify the top variables that contribute to the team's power ranking.

Introduction

"When performance is measured, performance improves. When performance is measured and reported back, the rate of improvement accelerates"

-Pearson's Law

Advanced Stats are a way to study basketball through objective analysis. It is a more in-depth way than just looking at a simple box score, and more accurately evaluates the skill and production of a player or team. This paper will focus on team`s statistics from the NCAA College Basketball season of 2018.

The analytics revolution has changed the way we think and talk about basketball. We cannot have an intelligent conversation about hoops without relevant statistics to back up the thoughts and lend credence to the opinions. All the respectable basketball columnists have come up with a plethora of advanced metrics which support their claims and provide astonishing insight into the game. Knowing the interpretation of numbers truly enhances basketball sophistication.

Win shares estimate an individual player's contribution to their team's win total. Through a complex formula it credits offensive win shares by calculating a player's marginal points from his points produced and offensive possessions and dividing it by the marginal points per win. Defensive win shares are credited by computing a player's marginal defense from his defensive rating and dividing it by the marginal points per win. Adding offensive and defensive win shares together to get total win shares

Basketball is all about efficiencies. Maximizing points scored and minimizing points allowed on each possession is more important than overall totals. Totals are influenced by variables like pace—or the number of possessions a team gets in a game—which can differ depending on coaching philosophies. Offensive and defensive efficiencies are adjusted for pace, calculating points scored and allowed on a per-possession basis. To make the numbers easy to digest, they are reported per 100 possessions, so they look similar to points-per-game figures.

There are four important advanced statistics that winning teams excel at. These factors are Effective Field Goal Percentage (eFG%), Turnover Ratio (TO Ratio), Offensive Rebound Percentage (OREB%), and Free Throw Attempt Rate (FTA Rate). eFG% measures field goal percentage adjusting for the fact that a 3-point field goal is worth one more point than a 2-point field goal. The formula is eFG% = ((FGM + (0.5 * 3PM)) / FGA. TO Ratio is the number of turnovers a player or team averages per 100 possessions used. Highlights good scorers who infrequently pass the ball (i.e. Players who typically catch and shoot.). The formula is (TO * 100) / (FGA + (FTA * 0.44) + AST + TO. OREB% is the percentage of team offensive rebounds grabbed by a player or team. The formula for finding this is OREB / (OREB + OppDREB). This eliminates a player or team’s inflated rebound numbers if the team misses a lot of shots. Free Throw Attempt Rate shows free throws attempted relative to field goals attempted by a player or team. FTA Rate shows how often a player or team goes to the line, and how good that individual is at drawing fouls. The formula is simply FTA/FGA.

In 2018, Villanova had exactly 128.4 adjusted offensive(ADJOE) points and became champions for that season and 2nd runner was Michigan with adjusted offensive points (ADJOE) of 114.4. Villanova had the highest ADJOE and there were about 32 schools that had a better ADJOE than Michigan. Therefore, there isn’t just one statistic that can be used to measure which school is better.

Literature Review

In the article *The Use of Data Mining for Basketball Matches Outcomes Prediction(Miljković et al.,2015)*, the writers were dealing with similar data that looked at predictive analysis on spread of NBA games and classification on game outcomes. Similar to the canonical correlation analysis they split the data into 2 attribute sets: standard basketball statistics and league standings. They used multivariate linear regression for predicting spread and feature selection, decision trees, and K-nearest neighbor for classification. Their model resulted in 67% accuracy for game outcomes, 10% accuracy for spread which was expected because basketball scores can be hard to predict the exact difference and their goal was to provide approximate information on possible differences. With this study they hope to be able to expand this experiment leveraging the model they created for other sports like football and hockey.

Another article examined is *Statistics Free Sports Prediction (Dubbs, 2010)*. It is so named because the author uses a machine learning model to predict sports winners and not traditional statistics that are normally used in sports predictors. This model was used to predict winners in baseball, basketball, football, and hockey. The model was not significantly good at predicting baseball, football or hockey winners, however it was good at predicting basketball winners. The model was built with a least squares regression using Gaussian elimination. Fifty percent of the data was used for training and the other half for testing. Even numbered years were the training data and odd numbered years were the testing data. A straw-man model was constructed for comparison; this model favored the higher rated team 100% of the time. In predicting NBA games, the model constructed in this study performed similar to other models cited by the paper, and significantly better than the straw-man model.

Lastly, the article *Football and Basketball Predictions Using Least Squares (Stefani,1977)* looks to focus on the accuracy of human experts and existing models in regards to football and basketball versus the accuracy of least squares. The existing models included penalties for teams based on how many points they won or lost by so that a good team would have to meet a certain point margin in order to not get penalized for not out-scoring its opponent. The article took the point spread into account but did not use any penalty method but rather used the rating of the team. Meaning if one team was rated 116 and another team was rated 100, then the better rated team would be expected to win by 16 points. However, if the better team one by less than 16 points, there wasn't any penalty to affect its new rating. The conclusion is that the least squares performed similar or better than the experts and models which means statistical analysis will uncover the importance of basketball rankings.

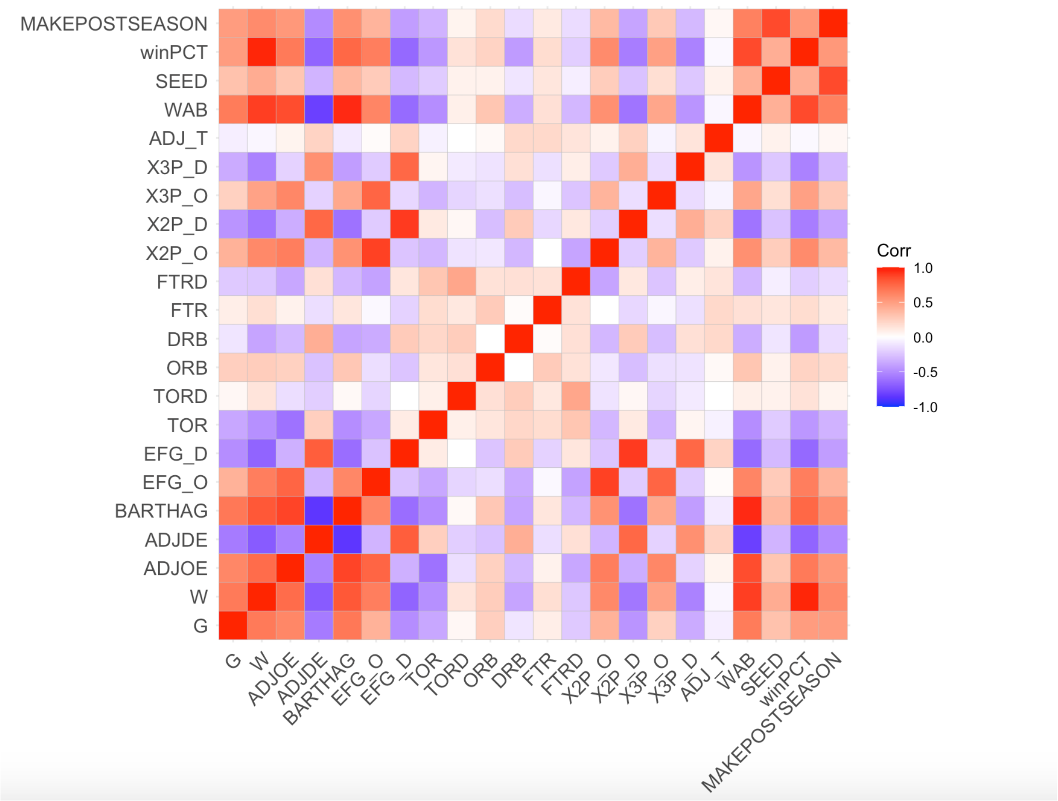
Methods

Exploratory Analysis- Correlations

The correlation between the independent variables shows a strong correlation for multicollinearity, see Figure 1 - Correlation Heatmap. Looking at the association between the power rankings(barthag) and independent variables, power rankings had the strongest associated with Winning Above the Bubble, and efficiencies. Looking at the actual values, there is a .88 correlation for Adjusted Offensive Efficiency and negative correlation of .86 for Adjusted Defensive Efficiency.

Figure 1

Correlation Heatmap



Exploratory Analysis : Clustering via K-means

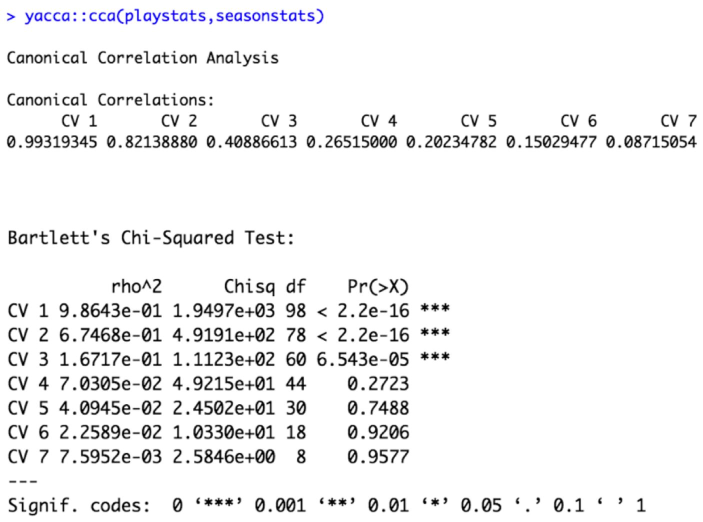
The objective of using clustering was to uncover hidden patterns and to see what schools are similar to focus on. For example, see if there was a pattern for schools that made it far in the tournament, such as the Elite 8 round. Since clustering was an exploratory approach, the dependent variable, Power Ranking, was included with the rest of the independent variables. The approach to determine how many clusters to use was based on the average silhouette width which resulted in using 2 clusters.

CCA

One method of variable selection used was canonical correlation analysis. The two sets of variables were divided up by variables that had to do with the actual game playing such as shooting and rebound percentages for a total of 15 variables, and the variables that summarized the season such as games won or power rating for a total of 7. The purpose of this was to figure out the canonical variables that are significant in explaining the association between play performance measures and actual play performance.

Canonical correlation was done in R and analyzed using cca from the YACCA package. From the test of significance, the first 3 correlations are shown as significant because of the low p-values. This also makes sense because the canonical correlations show a large dip after the third correlation at 0.265 The canonical correlations are shown below in Table 1.1.

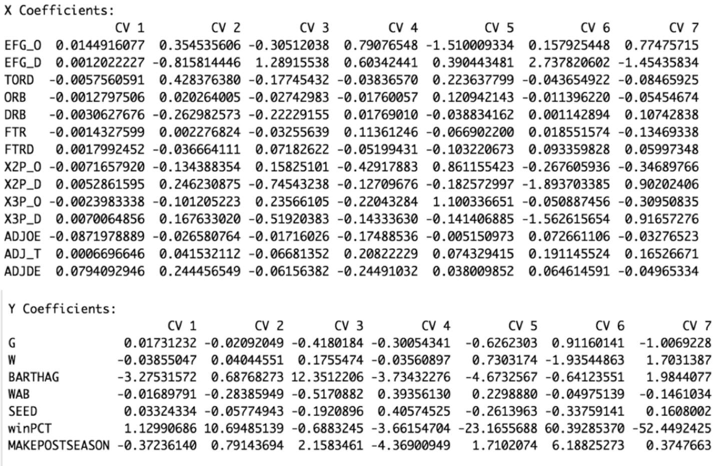
Table 1.1



```
> yacca::cca(playstats,seasonstats)

Canonical Correlation Analysis

Canonical Correlations:
      CV 1       CV 2       CV 3       CV 4       CV 5       CV 6       CV 7
0.99319345 0.82138880 0.40886613 0.26515000 0.20234782 0.15029477 0.08715054



Bartlett's Chi-Squared Test:

            rho^2      Chisq df    Pr(>X)
CV 1 9.8643e-01 1.9497e+03 98 < 2.2e-16 ***
CV 2 6.7468e-01 4.9191e+02 78 < 2.2e-16 ***
CV 3 1.6717e-01 1.1123e+02 60 6.543e-05 ***
CV 4 7.0305e-02 4.9215e+01 44    0.2723
CV 5 4.0945e-02 2.4502e+01 30    0.7488
CV 6 2.2589e-02 1.0330e+01 18    0.9206
CV 7 7.5952e-03 2.5846e+00  8    0.9577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Looking at the coefficients, most of the play coefficients are minimal and have similar impact with the exception of variables such as Effective Field Goal Percentage Shot and Adjusted Defensive Efficiency being the highest but only at around 0.01 and 0.07. On the other hand, for the Y coefficients, Power Rating and winPCT have the highest coefficients meaning these variables would be worth looking at for predictive analysis as the dependent variable.

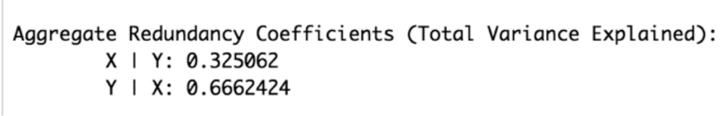
Table 1.2



```
X Coefficients:
               CV 1         CV 2        CV 3        CV 4         CV 5         CV 6        CV 7
EFG_O  0.0144916077  0.354535606 -0.30512038  0.79076548 -1.510009334  0.157925448  0.77475715
EFG_D  0.0012022227 -0.815814446  1.28915538  0.60342441  0.390443481  2.737820602 -1.45435834
TORD  -0.0057560591  0.428376380 -0.17745432 -0.03836570  0.223637799 -0.043654922 -0.08465925
ORB   -0.0012797506  0.020264005 -0.02742983 -0.01760057  0.120942143 -0.011396220 -0.05454674
DRB   -0.0030627676 -0.262982573 -0.22229155  0.01769010 -0.038834162  0.001142894  0.10742838
FTR   -0.0014327599  0.002276824 -0.03255639  0.11361246 -0.066902200  0.018551574 -0.13469338
FTRD   0.0017992452 -0.036664111  0.07182622 -0.05199431 -0.103220673  0.093359828  0.05997348
X2P_O -0.0071657920 -0.134388354  0.15825101 -0.42917883  0.861155423 -0.267605936 -0.34689766
X2P_D  0.0052861595  0.246230875 -0.74543238 -0.12709676 -0.182572997 -1.893703385  0.90202406
X3P_O -0.0023983338 -0.101205223  0.23566105 -0.22043284  1.100336651 -0.050887456 -0.30950835
X3P_D  0.0070064856  0.167633020 -0.51920383 -0.14333630 -0.141406885 -1.562615654  0.91657276
ADJOE -0.0871978889 -0.026580764 -0.01716026 -0.17488536 -0.005150973  0.072661106 -0.03276523
ADJ_T  0.0006696646  0.041532112 -0.06681352  0.20822229  0.074329415  0.191145524  0.16526671
ADJDE  0.0794092946  0.244456549 -0.06156382 -0.24491032  0.038009852  0.064614591 -0.04965334

Y Coefficients:
                     CV 1        CV 2       CV 3        CV 4        CV 5        CV 6        CV 7
G              0.01731232 -0.02092049 -0.4180184 -0.30054341  -0.6262303  0.91160141  -1.0069228
W             -0.03855047  0.04044551  0.1755474 -0.03560897   0.7303174 -1.93544863   1.7031387
BARTHAG       -3.27531572  0.68768273 12.3512206 -3.73432276  -4.6732567 -0.64123551   1.9844077
WAB           -0.01689791 -0.28385949 -0.5170882  0.39356130   0.2298880 -0.04975139  -0.1461034
SEED           0.03324334 -0.05774943 -0.1920896  0.40574525  -0.2613963 -0.33759141   0.1608002
winPCT         1.12990686 10.69485139 -0.6883245 -3.66154704 -23.1655688 60.39285370 -52.4492425
MAKEPOSTSEASON -0.37236140  0.79143694  2.1583461 -4.36900949   1.7102074  6.18825273   0.3747663
```

From the output in Table 1.3, the redundancy coefficients show that 32% of the variance in the play statistic variables can be explained by the season statistic variables and 66% of variation season statistic variables can be explained by play statistic variables meaning the overall statistics that have to do with actual playing give better possible predictability results when it comes to looking at things such as Power Rating, whether or not they make the postseason, and their seed in the March Madness tournament.

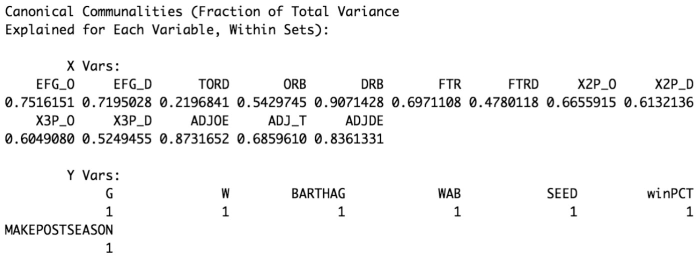
Table 1.3



```
Aggregate Redundancy Coefficients (Total Variance Explained):
        X | Y: 0.325062
        Y | X: 0.6662424
```

Looking at the commonality coefficients in Table 1.4, we can see which variables are contributing to the CC analysis the most or which ones load the best. This shows how much the variable's variance is reproducible from the variates. The variables that may not load the best might be FTRD Free Throw Rate Allowed) and TORD (Turnover Percentage Committed (Steal Rate)).

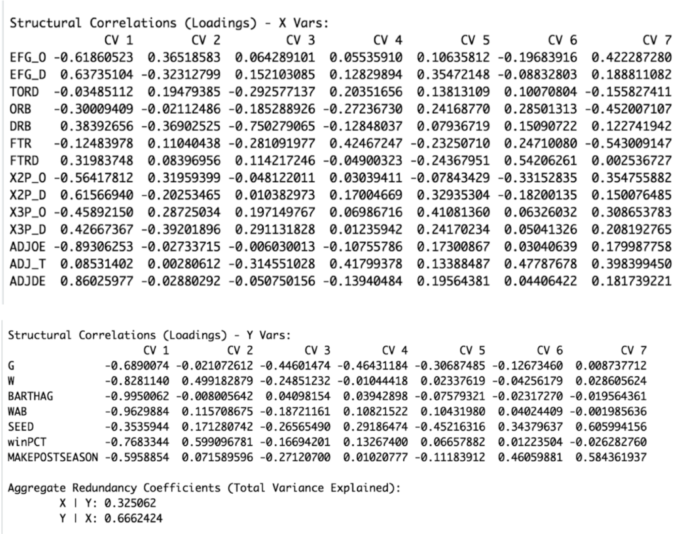
Table 1.4



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

        X Vars:
    EFG_O     EFG_D      TORD       ORB       DRB       FTR      FTRD     X2P_O     X2P_D 
0.7516151 0.7195028 0.2196841 0.5429745 0.9071428 0.6971108 0.4780118 0.6655915 0.6132136 
    X3P_O     X3P_D     ADJOE     ADJ_T     ADJDE 
0.6049080 0.5249455 0.8731652 0.6859610 0.8361331 

        Y Vars:
             G              W        BARTHAG            WAB           SEED         winPCT 
             1              1              1              1              1              1 
MAKEPOSTSEASON 
             1 
```

Now looking at the loadings, it does show that TORD (Turnover Percentage Committed (Steal Rate)) especially, has a low score at -0.03 and may not be necessary to include in a predictive analysis. There are a few other variables that we can see in Table 1.5 that have loadings less than 0.3, which should also be considered when creating a model for predictive analysis.

Table 1.5



```
Structural Correlations (Loadings) - X Vars:
            CV 1        CV 2         CV 3        CV 4        CV 5        CV 6         CV 7
EFG_O -0.61860523  0.36518583  0.064289101  0.05535910  0.10635812 -0.19683916  0.422287280
EFG_D  0.63735104 -0.32312799  0.152103085  0.12829894  0.35472148 -0.08832803  0.188811082
TORD  -0.03485112  0.19479385 -0.292577137  0.20351656  0.13813109  0.10070804 -0.155827411
ORB   -0.30009409 -0.02112486 -0.185288926 -0.27236730  0.24168770  0.28501313 -0.452007107
DRB    0.38392656 -0.36902525 -0.750279065 -0.12848037  0.07936719  0.15090722  0.122741942
FTR   -0.12483978  0.11040438 -0.281091977  0.42467247 -0.23250710  0.24710080 -0.543009147
FTRD   0.31983748  0.08396956  0.114217246 -0.04900323 -0.24367951  0.54206261  0.002536727
X2P_O -0.56417812  0.31959399 -0.048122011  0.03039411 -0.07843429 -0.33152835  0.354755882
X2P_D  0.61566940 -0.20253465  0.010382973  0.17004669  0.32935304 -0.18200135  0.150076485
X3P_O -0.45892150  0.28725034  0.197149767  0.06986716  0.41081360  0.06326032  0.308653783
X3P_D  0.42667367 -0.39201896  0.291131828  0.01235942  0.24170234  0.05041326  0.208192765
ADJOE -0.89306253 -0.02733715 -0.006030013 -0.10755786  0.17300867  0.03040639  0.179987758
ADJ_T  0.08531402  0.00280612 -0.314551028  0.41799378  0.13388487  0.47787678  0.398399450
ADJDE  0.86025977 -0.02880292 -0.050750156 -0.13940484  0.19564381  0.04406422  0.181739221
```

```
Structural Correlations (Loadings) - Y Vars:
                     CV 1         CV 2        CV 3        CV 4        CV 5        CV 6         CV 7
G              -0.6890074 -0.021072612 -0.44601474 -0.46431184 -0.30687485 -0.12673460  0.008737712
W              -0.8281140  0.499182879 -0.24851232 -0.01044418  0.02337619 -0.04256179  0.028605624
BARTHAG        -0.9950062 -0.008005642  0.04098154  0.03942898 -0.07579321 -0.02317270 -0.019564361
WAB            -0.9629884  0.115708675 -0.18721161  0.10821522  0.10431980  0.04024409 -0.001985636
SEED           -0.3535944  0.171280742 -0.26565490  0.29186474 -0.45216316  0.34379637  0.605994156
winPCT         -0.7683344  0.599096781 -0.16694201  0.13267400  0.06657882  0.01223504 -0.026282760
MAKEPOSTSEASON -0.5958854  0.071589596 -0.27120700  0.01020777 -0.11183912  0.46059881  0.584361937

Aggregate Redundancy Coefficients (Total Variance Explained):
        X | Y: 0.325062
        Y | X: 0.6662424
```

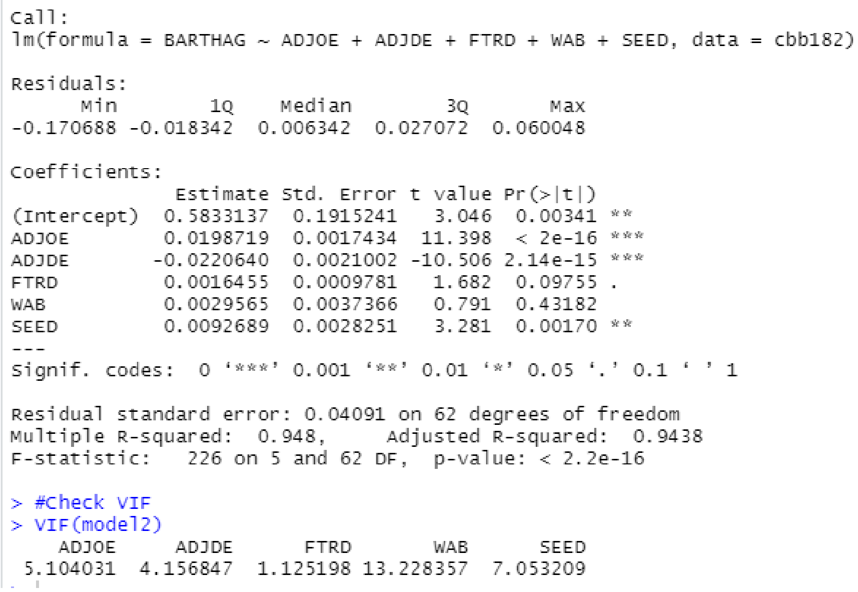
Multiple Regression

In multiple regression analysis, the goal was to predict the power rating BARTHAG, the dependent variable of the basketball team based on twenty four independent variables ADJOE, ADJDE, G,W etc. For decision making I have used ADJOE, one of the independent variables to check its impact and the result was the higher the ADJOE, the higher the levels of power rating.

ADJOE, ADJDE and SEED have a very high significant impact on the power rating when ran the regression analysis on the model, which directly shows what the coach can do to fix to improve the performance of the team to lead the competition; predict this about the future; or to decide what to do. p-value for a variable is less than significance level, so we have enough evidence to reject the null hypothesis for the team and there is a non-zero correlation.

The p-value for other independent variables tests the null hypothesis that the variable has no correlation with the dependent variable. Since there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is insufficient evidence to conclude that there is an effect on power rating.

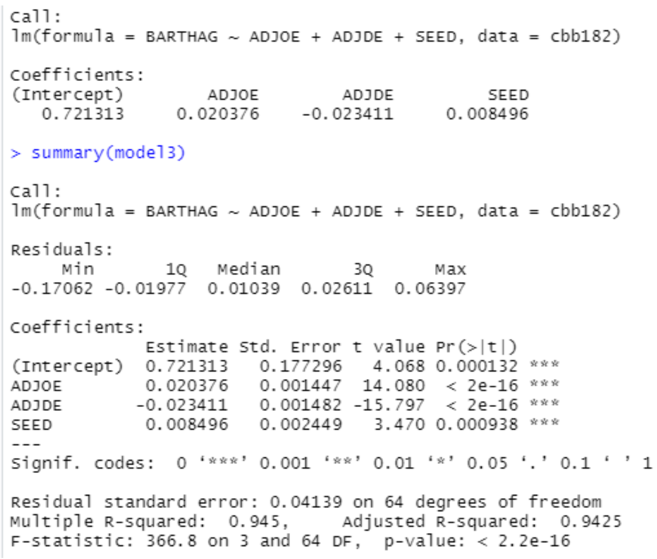
Table 1.6



```
Call:
lm(formula = BARTHAG ~ ADJOE + ADJDE + FTRD + WAB + SEED, data = cbb182)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.170688 -0.018342  0.006342  0.027072  0.060048 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.5833137  0.1915241   3.046  0.00341 ** 
ADJOE        0.0198719  0.0017434  11.398  < 2e-16 ***
ADJDE       -0.0220640  0.0021002 -10.506 2.14e-15 ***
FTRD         0.0016455  0.0009781   1.682  0.09755 .  
WAB          0.0029565  0.0037366   0.791  0.43182    
SEED         0.0092689  0.0028251   3.281  0.00170 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04091 on 62 degrees of freedom
Multiple R-squared:  0.948,	Adjusted R-squared:  0.9438 
F-statistic:   226 on 5 and 62 DF,  p-value: < 2.2e-16

> #Check VIF
> VIF(model2)
    ADJOE     ADJDE      FTRD       WAB      SEED 
 5.104031  4.156847  1.125198 13.228357  7.053209 
```

The regression output above shows that the ADJOE, ADJDE, and SEED predictor variables are statistically significant because their p-values are less than 0.05. On the other hand, FTRD, WAB are not statistically significant because its p-value is greater than the usual significance level of 0.05.

Table 1.7



```
Call:
lm(formula = BARTHAG ~ ADJOE + ADJDE + SEED, data = cbb182)

Coefficients:
(Intercept)        ADJOE        ADJDE         SEED
   0.721313     0.020376    -0.023411     0.008496

> summary(model3)

Call:
lm(formula = BARTHAG ~ ADJOE + ADJDE + SEED, data = cbb182)

Residuals:
     Min       1Q   Median       3Q      Max
-0.17062 -0.01977  0.01039  0.02611  0.06397

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.721313   0.177296   4.068 0.000132 ***
ADJOE        0.020376   0.001447  14.080  < 2e-16 ***
ADJDE       -0.023411   0.001482 -15.797  < 2e-16 ***
SEED         0.008496   0.002449   3.470 0.000938 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04139 on 64 degrees of freedom
Multiple R-squared:  0.945,     Adjusted R-squared:  0.9425
F-statistic: 366.8 on 3 and 64 DF,  p-value: < 2.2e-16
```

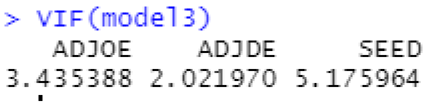
The p- value< 2.2e-16 of the regressor variable is less than 0.05 that means we can reject the null hypothesis, and this implies it passes the F test. The p-values for the coefficients indicate these relationships are statistically significant.

When checked the ANOVA to compare the two models and observed the variation between two models the p value is 0.04<0.05 and that concludes that the predictor variable significantly affects the model and doesn't have to be removed from the model.

The multiple r squared and adjusted r square is 0.94 which tells that power rating has strongly correlation coefficient which indicates better fit for the model.

Lastly performed the VIF on the model and removed variables which were more than 5 as they were multicollinear with other variables.

Table 1.8



```
> VIF(model3)
   ADJOE    ADJDE     SEED
3.435388 2.021970 5.175964
```

PCA

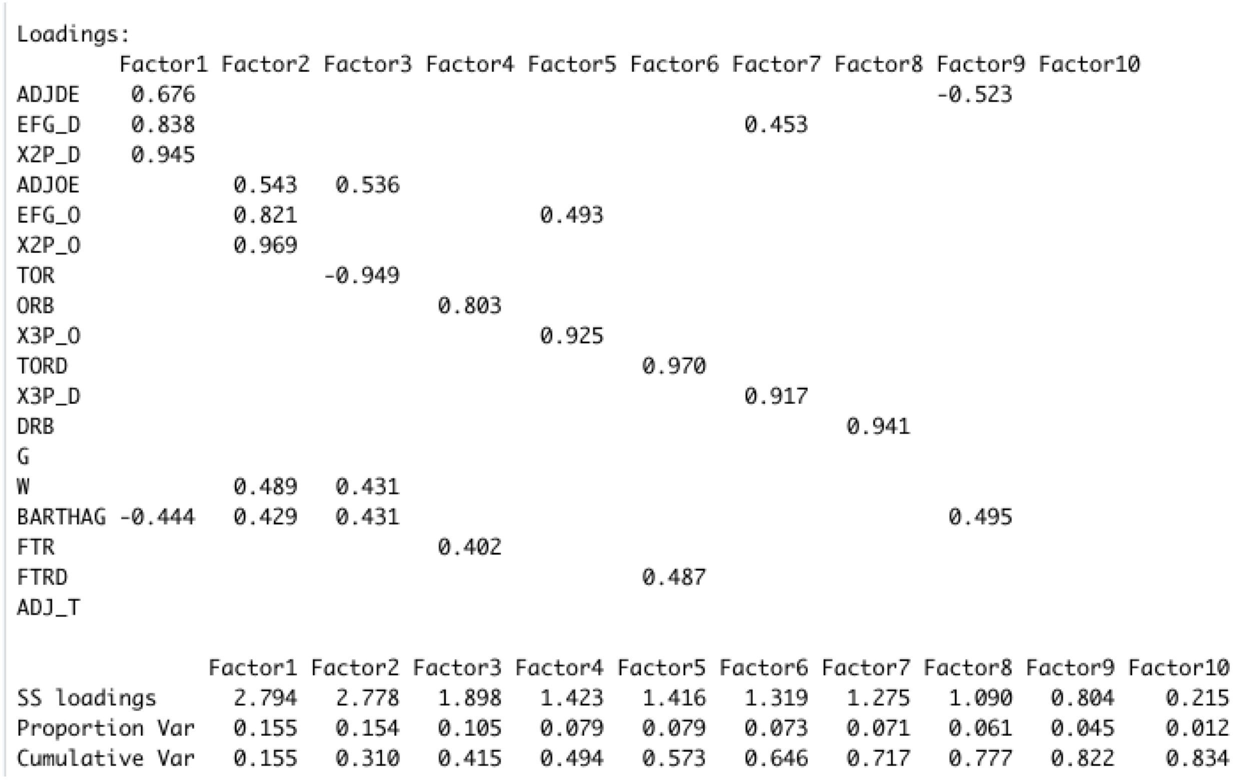
The next method for analysis used was PCA. From the given data there were some NA values that represented the teams that didn’t make it to postseason and they are replaced by 0 for this analysis and are represented by a created variable(make postseason) to keep the original data intact. There was another variable added to provide us with the win percentage of a given team and for the dependent variable for the analysis was power rating represented by the BARTHAG.

Since different variables were used to provide the same data differently the original column was removed from the analysis along with the string column to give us just the numeric data for analysis. For the multicollinearity, the Variance Inflation Factor test was performed on the dataset from the library Car. To provide us with a good understanding on how many variables are correlated. There was another test done to check for Multicollinearity which was multicollinearity with correlations.

Before performing the PCA Bartlett's Test of Sphericity to find out if the sample size for the dataset had enough variance to run the model. There was also a reliability analysis done using Cronbach' Alpha to see how reliable the dataset was for the analysis. After all the tests are done the PCA analysis was run see the discussion section for the results of the analysis.

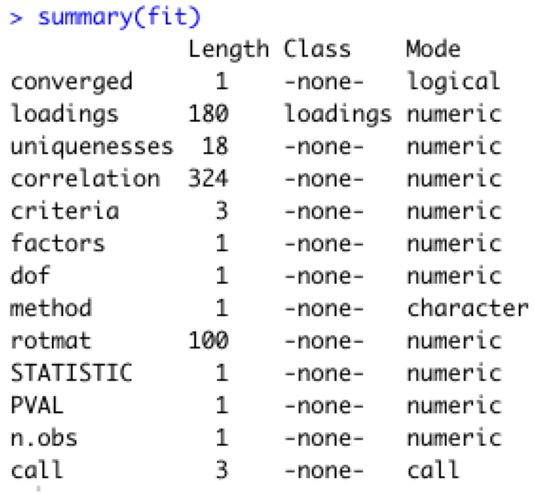
Factor Analysis

Table 1.9



```
Loadings:
        Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8 Factor9 Factor10
ADJDE    0.676                                                          -0.523
EFG_D    0.838                                           0.453
X2P_D    0.945
ADJOE            0.543   0.536
EFG_O            0.821                   0.493
X2P_O            0.969
TOR                     -0.949
ORB                              0.803
X3P_O                                    0.925
TORD                                             0.970
X3P_D                                                    0.917
DRB                                                              0.941
G
W                0.489   0.431
BARTHAG -0.444   0.429   0.431                                           0.495
FTR                              0.402
FTRD                                             0.487
ADJ_T

               Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7 Factor8 Factor9 Factor10
SS loadings      2.794   2.778   1.898   1.423   1.416   1.319   1.275   1.090   0.804    0.215
Proportion Var   0.155   0.154   0.105   0.079   0.079   0.073   0.071   0.061   0.045    0.012
Cumulative Var   0.155   0.310   0.415   0.494   0.573   0.646   0.717   0.777   0.822    0.834
```

Table 1.10



```
> summary(fit)
             Length Class    Mode
converged      1    -none-   logical
loadings     180    loadings numeric
uniquenesses  18    -none-   numeric
correlation  324    -none-   numeric
criteria       3    -none-   numeric
factors        1    -none-   numeric
dof            1    -none-   numeric
method         1    -none-   character
rotmat       100    -none-   numeric
STATISTIC      1    -none-   numeric
PVAL           1    -none-   numeric
n.obs          1    -none-   numeric
call           3    -none-   call
```

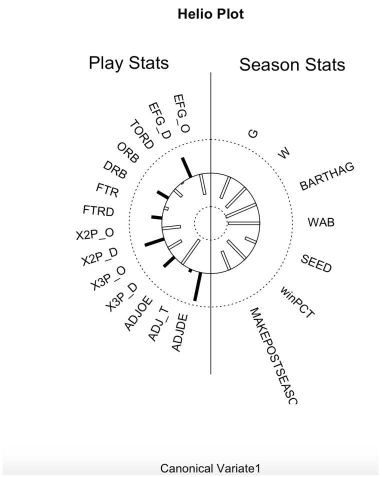
Used factor analysis for ten factors, and omitted the SEED column due to too many null values, and the MAKEPOSTSEASON column due to too much covariance (MAKEPOSTSEASON is 1 for nearly all values.) Ten factors are shown here, although after the first nine the proportional variance is negligible. Given the strong tendency towards defense and offense, respectively, Factor 1 is defense ability and Factor 2 is offense ability. Factor 3 matches strongly with wins and power rating (and efficiency, which is unsurprising given the importance of efficiency.) This means Factor 3 likely measures the chance of winning. Factor 5 matches very strongly to three-pointers, and a low match with efficiency. Since it's so heavily focused on three-pointers, it might just be a measure of three-point shots. Likewise, Factor 6 may just be the steal rate, Factor 7 defense three-pointers, and Factor 8 defense rebounds. Given these findings, I'd say that Factors 1, 2, and 3 are the most important, and the rest of the factors differ only slightly from already known variables.

Discussion and Results

By running the canonical correlation analysis and analyzing both the correlations and the loadings, we were able to validate that power rating was a good predictor variable for our study when looking at and comparing the different basketball statistics since it had the highest coefficient and loading value. The helio plot for the first variate (shown below) visualizes Power Rating as one of the strongest dependent variables for this study.

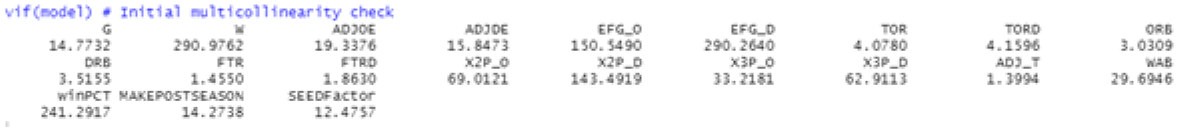
Figure 2

Helio Plot



For the PCA analysis as mentioned in the methods section multiple tests were run before performing the actual analysis the first test that was run was the VIF model to see if there was any multicollinearity between all the variables that was within the dataset. Here is the model of the first model that was run.

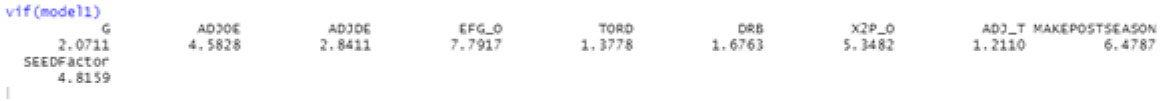
Table 1.11



```
vif(model) # initial multicollinearity check
             G              W          ADJOE          ADJDE          EFG_O          EFG_D            TOR           TORD            ORB
       14.7732       290.9762        19.3376        15.8473       150.5490       290.2640         4.0780         4.1596         3.0309
           DRB            FTR           FTRD          X2P_O          X2P_D          X3P_O          X3P_D          ADJ_T            WAB
        3.5155         1.4550         1.8630        69.0121       143.4919        33.2181        62.9113         1.3994        29.6946
        winPCT MAKEPOSTSEASON     SEEDFactor
      241.2917        14.2738        12.4757
```

From the result shown in table 1.11 a couple of variables seem to have high VIF numbers that were above 10 some that were above 100. So, a process of elimination was done and for this process a variable with the high VIF number was removed first then the analysis was done and if the model yielded another variable with high VIF then it was removed from the analysis. So, to keep this short there was about 11 variables removed and here is the final VIF model

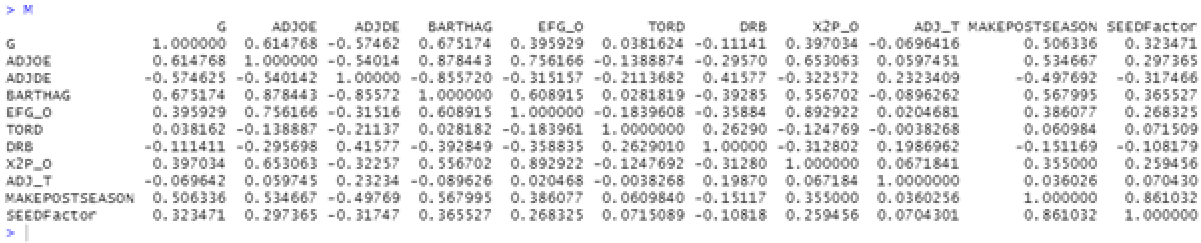
Table 1.12



```
vif(model1)
              G           ADJOE           ADJDE           EFG_O            TORD             DRB           X2P_O           ADJ_T MAKEPOSTSEASON
         2.0711          4.5828          2.8411          7.7917          1.3778          1.6763          5.3482          1.2110         6.4787
     SEEDFactor
         4.8159
|
```

As shown in table 1.12 the VIF model some of variables are above 5 even though they might have multicollinearity with other variables they are still kept in for the analysis. After the VIF analysis was done there was another test done to check for multicollinearity and that was multicollinearity with correlations and this model was run after the variables were removed. Here is the result of that.

Table 1.13

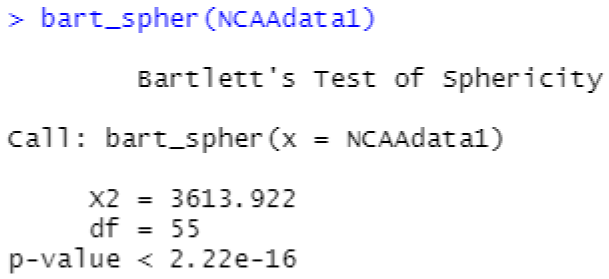


```
> M
                       G     ADJOE    ADJDE   BARTHAG     EFG_O       TORD      DRB     X2P_O      ADJ_T MAKEPOSTSEASON SEEDFactor
G               1.000000  0.614768 -0.57462  0.675174  0.395929  0.0381624 -0.11141  0.397034 -0.0696416       0.506336   0.323471
ADJOE           0.614768  1.000000 -0.54014  0.878443  0.756166 -0.1388874 -0.29570  0.653063  0.0597451       0.534667   0.297365
ADJDE          -0.574625 -0.540142  1.00000 -0.855720 -0.315157 -0.2113682  0.41577 -0.322572  0.2323409      -0.497692  -0.317466
BARTHAG         0.675174  0.878443 -0.85572  1.000000  0.608915  0.0281819 -0.39285  0.556702 -0.0896262       0.567995   0.365527
EFG_O           0.395929  0.756166 -0.31516  0.608915  1.000000 -0.1839608 -0.35884  0.892922  0.0204681       0.386077   0.268325
TORD            0.038162 -0.138887 -0.21137  0.028182 -0.183961  1.0000000  0.26290 -0.124769 -0.0038268       0.060984   0.071509
DRB            -0.111411 -0.295698  0.41577 -0.392849 -0.358835  0.2629010  1.00000 -0.312802  0.1986962      -0.151169  -0.108179
X2P_O           0.397034  0.653063 -0.32257  0.556702  0.892922 -0.1247692 -0.31280  1.000000  0.0671841       0.355000   0.259456
ADJ_T          -0.069642  0.059745  0.23234 -0.089626  0.020468 -0.0038268  0.19870  0.067184  1.0000000       0.036026   0.070430
MAKEPOSTSEASON  0.506336  0.534667 -0.49769  0.567995  0.386077  0.0609840 -0.15117  0.355000  0.0360256       1.000000   0.861032
SEEDFactor      0.323471  0.297365 -0.31747  0.365527  0.268325  0.0715089 -0.10818  0.259456  0.0704301       0.861032   1.000000
> |
```

The result from table 1.13 shows that there is some correlation between MAKEPOSTSEASON and SEEDFactor but one of the suspicions is to why they are correlated is that both variables have 0 as a data so maybe that is why they are correlated. For the power rating there is one variable that is highly correlated and that is ADJOE which is at 87% and there are some other like G and EFFG_O that are above 50%. Even though the rest aren't correlated with power rating the PCA analysis will be done on these variables since they don't have high multicollinearity with power rating. Before the PCA analysis was done there was additional testing done on the data set to better provide us whether there was enough sample size for the data set and whether the dataset was reliable.

Bartlett's Test of Sphericity was performed to find out if there was enough sample size for the data set and for reliability testing Cronbach' Alpha was performed to find out the reliability of the data set. Here is the result for both of them

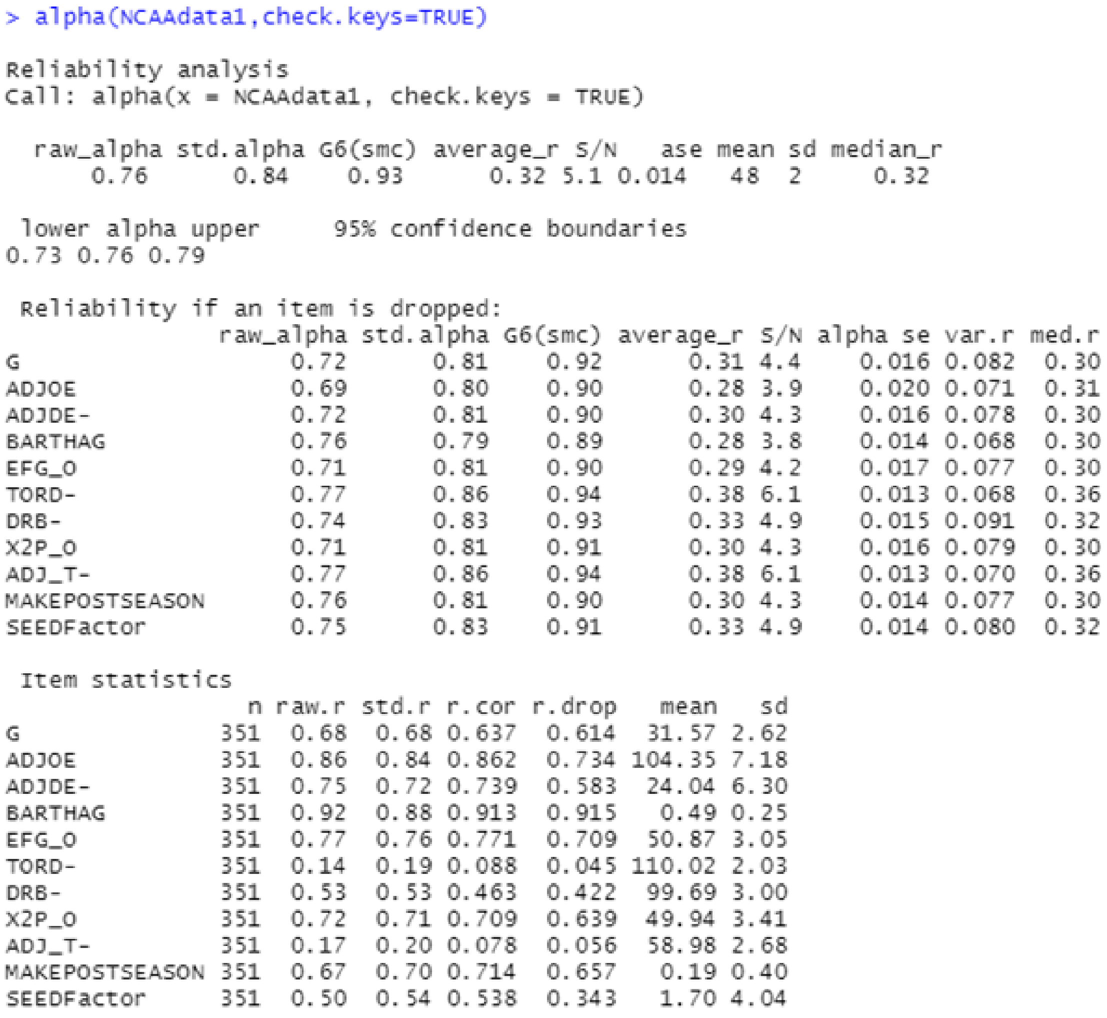
Table 1.14



```
> bart_spher(NCAAdata1)

        Bartlett's Test of Sphericity

Call: bart_spher(x = NCAAdata1)

     X2 = 3613.922
     df = 55
p-value < 2.22e-16
```

For the Bartlett's Test of Sphericity we see that the p-value is 2.22e-16 so it is reasonable in terms of the sample size.

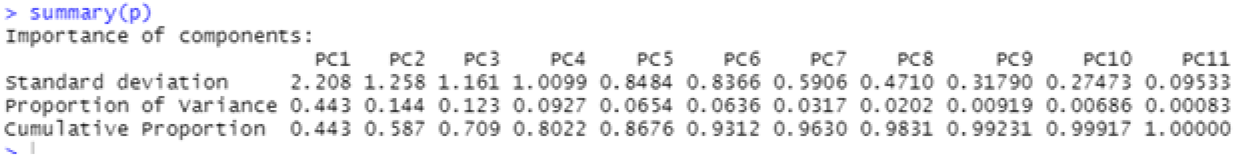
Table 1.15



```
> alpha(NCAAdata1,check.keys=TRUE)

Reliability analysis
Call: alpha(x = NCAAdata1, check.keys = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
      0.76      0.84    0.93      0.32 5.1 0.014   48  2     0.32

 lower alpha upper     95% confidence boundaries
0.73 0.76 0.79

 Reliability if an item is dropped:
               raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
G                   0.72      0.81    0.92      0.31 4.4    0.016 0.082  0.30
ADJOE               0.69      0.80    0.90      0.28 3.9    0.020 0.071  0.31
ADJDE-              0.72      0.81    0.90      0.30 4.3    0.016 0.078  0.30
BARTHAG             0.76      0.79    0.89      0.28 3.8    0.014 0.068  0.30
EFG_O               0.71      0.81    0.90      0.29 4.2    0.017 0.077  0.30
TORD-               0.77      0.86    0.94      0.38 6.1    0.013 0.068  0.36
DRB-                0.74      0.83    0.93      0.33 4.9    0.015 0.091  0.32
X2P_O               0.71      0.81    0.91      0.30 4.3    0.016 0.079  0.30
ADJ_T-              0.77      0.86    0.94      0.38 6.1    0.013 0.070  0.36
MAKEPOSTSEASON      0.76      0.81    0.90      0.30 4.3    0.014 0.077  0.30
SEEDFactor          0.75      0.83    0.91      0.33 4.9    0.014 0.080  0.32

 Item statistics
                 n raw.r std.r r.cor r.drop   mean   sd
G              351  0.68  0.68 0.637  0.614  31.57 2.62
ADJOE          351  0.86  0.84 0.862  0.734 104.35 7.18
ADJDE-         351  0.75  0.72 0.739  0.583  24.04 6.30
BARTHAG        351  0.92  0.88 0.913  0.915   0.49 0.25
EFG_O          351  0.77  0.76 0.771  0.709  50.87 3.05
TORD-          351  0.14  0.19 0.088  0.045 110.02 2.03
DRB-           351  0.53  0.53 0.463  0.422  99.69 3.00
X2P_O          351  0.72  0.71 0.709  0.639  49.94 3.41
ADJ_T-         351  0.17  0.20 0.078  0.056  58.98 2.68
MAKEPOSTSEASON 351  0.67  0.70 0.714  0.657   0.19 0.40
SEEDFactor     351  0.50  0.54 0.538  0.343   1.70 4.04
```

For the reliability analysis the Cronbach' Alpha's raw score is .76 when 95% confidence interval between .73 and .79. Which seems to suggest that the data set is reliable to perform the PCA analysis. For the PCA analysis as mentioned previously the variables that were in the elimination process were kept for the analysis and here is the result of that.

Table 1.16



```
> summary(p)
Importance of components:
                         PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8     PC9    PC10    PC11
Standard deviation     2.208 1.258 1.161 1.0099 0.8484 0.8366 0.5906 0.4710 0.31790 0.27473 0.09533
Proportion of Variance 0.443 0.144 0.123 0.0927 0.0654 0.0636 0.0317 0.0202 0.00919 0.00686 0.00083
Cumulative Proportion  0.443 0.587 0.709 0.8022 0.8676 0.9312 0.9630 0.9831 0.99231 0.99917 1.00000
```

From the analysis performed the result seems to suggest that only 43%of the data is being explained from the given data. From this analysis one can conclude that there might have been some limitation with the given dataset which could be resolved given a larger sample size but since this is comparing college basketball one can utilize other years to increase sample to give us better understanding of the data.

Multiple Regression analysis explained the phenomenon what the coach what to understand from the data eg: tell which team is winning or losing the game by comparing the stats(ADJOE,ADJDE,SEED) which are significantly related to power rating and looking at the multiple r squared value which is 94% ,has a strong correlation coefficient indicates better fit of the model.

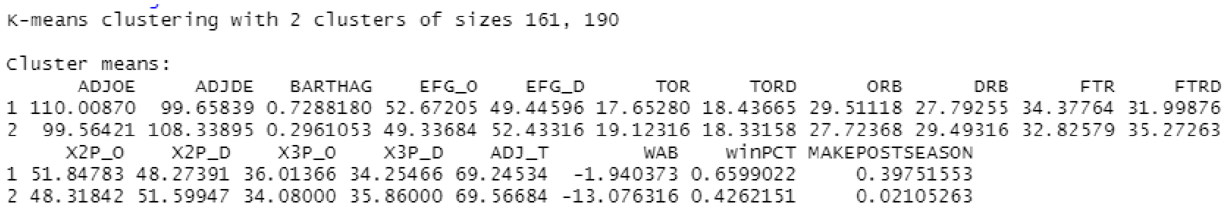
The CFA analysis revealed several factors that explain the model's variance. The two of the most important and most explanatory of which were the first two, one of which corresponded strongly to offense and the other to defense. Factor three also showed a correlation with winning games and power rating, which is the most important in terms of our goal, which is to predict power rating.

The clustering outcome of the 2 clusters did not show a good intra-class as the segments were not spaced enough to see where one began and ended versus the other. Also, the inter-class shows that some of the better teams are away from the center. Even with other clustering methods, there wasn't an improvement. For example, K-Medoids and Fuzzy clustering were attempted but the results were very similar. The Dunn Coefficient was .5 which does not make a case to use soft or hard clustering. One major pattern that was found was that the 1st cluster on the right included mostly all the postseason teams by visually analyzing. Also looking Table 3.0 for the cluster means summary, it shows that cluster 1 has a higher mean for power ranking , win pct, and making the postseason. This further justifies that cluster 1 could be potentially used for classification via machine learning.

In college sports there is a clear difference between the top programs and the majority of the division. So it would make sense to see that the top teams look more like outliers compared to its peers given 64 teams make the postseason and of that 64 only 25 are really considered when experts talk about ranks. Twenty-five is usually the threshold of rankings by different sport orgs (associated press, Sports Illustrated, ESPN, etc). Therefore, using clustering for the entire data set may not be the best approach and looking for a way to subset the date for future work should be considered.

Table 1.17

K-Means Summary



```
K-means clustering with 2 clusters of sizes 161, 190

Cluster means:
      ADJOE     ADJDE   BARTHAG    EFG_O    EFG_D      TOR     TORD      ORB      DRB      FTR     FTRD
1 110.00870  99.65839 0.7288180 52.67205 49.44596 17.65280 18.43665 29.51118 27.79255 34.37764 31.99876
2  99.56421 108.33895 0.2961053 49.33684 52.43316 19.12316 18.33158 27.72368 29.49316 32.82579 35.27263
     X2P_O    X2P_D    X3P_O    X3P_D    ADJ_T        WAB    winPCT MAKEPOSTSEASON
1 51.84783 48.27391 36.01366 34.25466 69.24534  -1.940373 0.6599022     0.39751553
2 48.31842 51.59947 34.08000 35.86000 69.56684 -13.076316 0.4262151     0.02105263
```

Figure 3

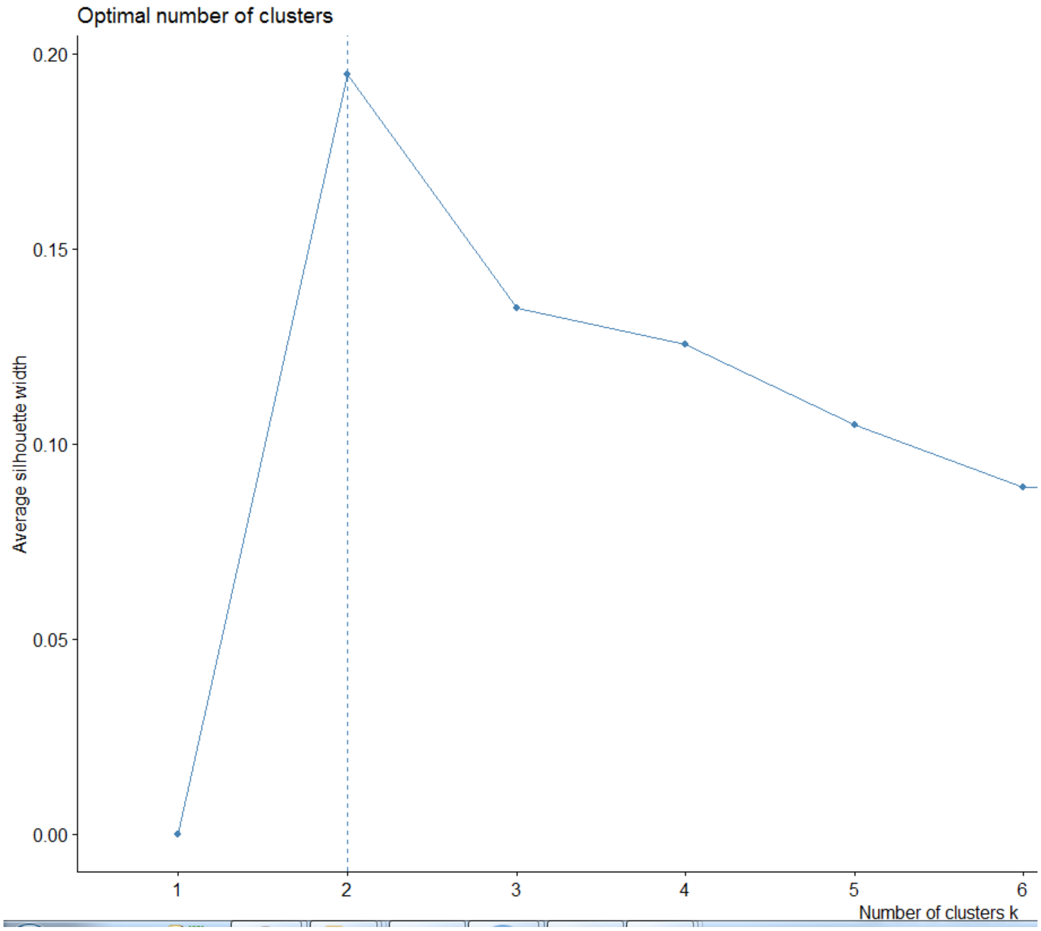


Figure 4
K-means Clustering

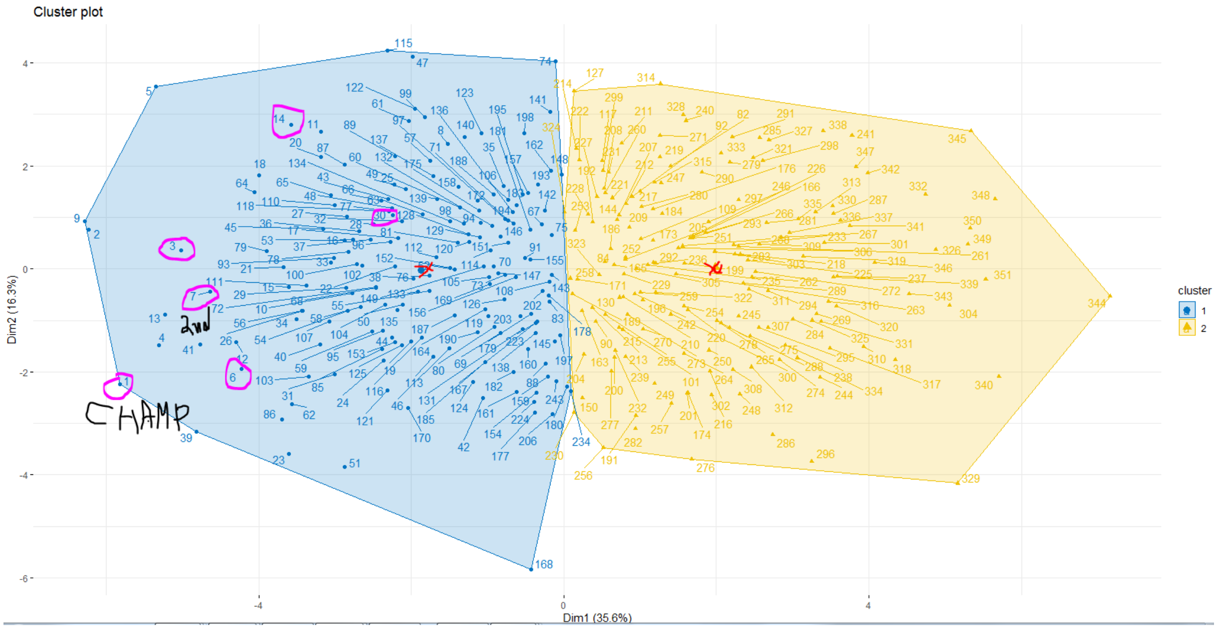
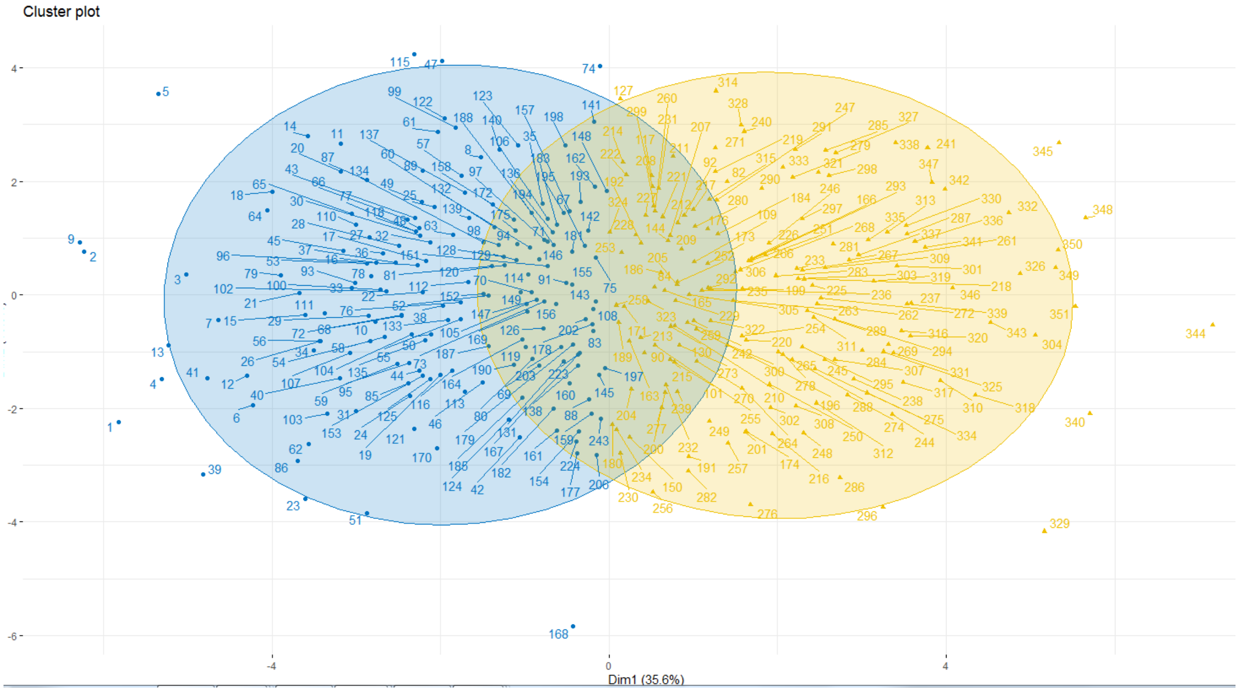


Figure 5

Fuzzy Clustering



Conclusion

Performance measurement and analysis is critical in improving basketball performance. We focused primarily on predicting power ranking, which is a way of measuring a team's chance at winning. The variable we found that corresponded the most strongly to power rating was efficiency. It's our finding that if a team wants to best improve their power rating, they should focus very strongly on their efficiency. Offensive and defensive efficiency were shown to be about equally important. That is to say that a team's ability to score points and their ability to keep the other team from scoring points is equally important. Both offense and defense need to be focused on and improved equally in order to improve power ranking.

In some ways basketball is easier to predict than other sports. Higher ranked teams win more often in basketball than in other sports. College basketball prediction is a popular and growing field. Some people engage as a hobby, some for financial gain, some a mixture of both. Power ranking can affect everything from sports gambling to team sponsorships. As college basketball increases in popularity, so does the desire to be able to accurately predict power rankings. Most college football games have been cancelled for the 2020 season, but the fate of the upcoming college basketball season is still unknown. If the spring college basketball season goes on as usual in 2021, there may be an explosion in popularity due to an absence of sports games for such a long time. It's possible then that power ranking prediction could be even more important. More research and better models are crucial for this prediction.

Appendix

Miljkovic, D., Gajic, L., Kovacevic, A., & Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. IEEE 8th International Symposium on Intelligent Systems and Informatics. doi:10.1109/sisy.2010.5647440

Dubbs, A. (2018). Statistics-free sports prediction. *Model Assisted Statistics and Applications, 13*(2), 173-181. doi:10.3233/mas-180428

R. T. Stefani,(1977), Football and Basketball Predictions Using Least Squares. *IEEE Transactions on Systems, Man, and Cybernetics*, 7(2), 117-121, doi: 10.1109/TSMC.1977.4309667.