College Basketball Data Analysis

Xixi

2022-12-30

## setup

# INTRODUCTION

篮球比赛的广泛性和对于人们的影响之大。

所以研究篮球数据的重要性不言而喻。

This paper will focus on solving the two questions related to the two variables stated above: Do the basketball levels differ between Conferences? If so, what are the differences? And how can we predict a team’s Seed based on the its existing basketball game performance data of the regular season?

Our analysis aims to identify the basketball performance indicators which help the team ace in the regular season and assist the coaches in designing targeted training programs, improving teams’ competitiveness.

# DATA

My analysis was done on the data from [Kaggle](https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset?select=cbb19.csv) named College Basketball Dataset by Andrew Sunberg.

The bulk of our analysis was done on data we found on Kaggle taken from National Agricultural Statistical Services (NASS), which annually surveys farmers nationwide. Due to a funding shortage, some data is missing from 2019 Quarter 2. In the raw data, “Alaska”, “Delaware”, “Nevada”, “New Hampshire”, and “Rhode Island” are all included in “Other”. In order to make the dataset more detailed and precise, we separated the “Other” into those five states and filled those rows with “NA”. There were some empty cells filled with “-”, and some negligibly small values in cells represented by “(Z)”, and these were all replaced with zeroes. There was partial 2021 data that we did not include because it only covered Q1 and Q2. In our Stressor data set, there was a generic category of stressors too small to give their own category that were grouped together under “Other”. These included weather, starvation, insufficient forage, queen failure, and hive damage/destruction. We interpreted all the variables in our dataset in the following table:

需要着重描述一下是如何处理数据的！

seed跟所在的conference有没有关系？或许存在某个conference占据很多席位的现象？与实力的关系？

seed与postseason的关系？ (需要思考出一种方法量化postseason的排名数据 重要‼️

# RESULTS

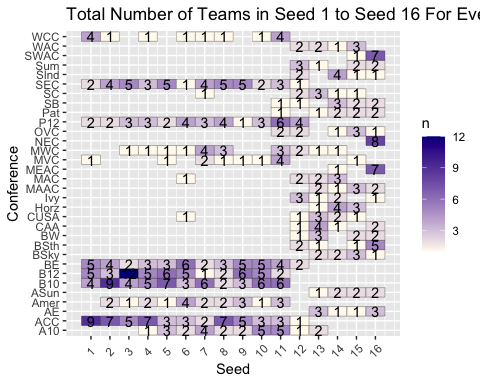
Question 1:

To develop the best approach for reducing loss, we wanted to create effective models that would indicate how other predictors might be used as tools for understanding this phenomenon. Initial investigation into Percent Loss revealed a rather weak correlation to any one stressor in particular. Then we utilized a tile plot to display the correlation of Percent Loss and all the Stressor variables as follows:

cbb\_all\_1 <- cbb\_all %>%   
 count(CONF,SEED) %>%   
 arrange(CONF,SEED) %>%   
 filter(SEED != "NA")  
  
p\_seed <- ggplot(cbb\_all\_1, aes(CONF, SEED, fill = n)) +  
 geom\_tile(color = "black") +  
 geom\_text(aes(label=n)) +  
 scale\_fill\_continuous(low= "floralwhite", high= "darkblue") +  
 theme(axis.text.x=element\_text(angle=45,vjust=0.5)) +  
 scale\_y\_continuous(breaks=seq(1,16,1)) +  
 xlab("Conference")+ylab("Seed")+ggtitle("Total Number of Teams in Seed 1 to Seed 16 For Every Conference Over 2013-2021") + # 2020的数据除外  
 coord\_fixed() +  
 coord\_flip()

## Coordinate system already present. Adding new coordinate system, which will  
## replace the existing one.

p\_seed



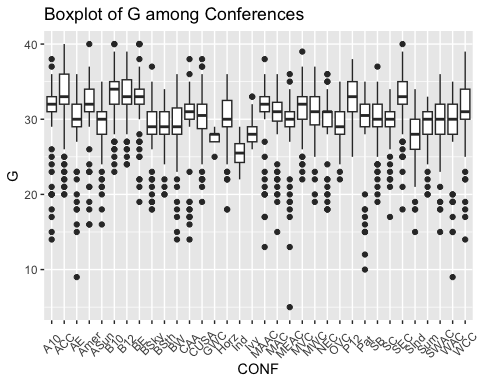
# 在tile上增加数字  
# 改为深色代表“大”，浅色代表“小”,空白代表NA(而并非将NA filter掉)  
# 不能忽略ind组的teams，虽然他们的seed都是NA

从历年数据(2013-2021,2020除外)来看，ACC,B10,B12在seed上有绝对的优势。 MEAC, NEC和SWAC排名比较落后 GWC和Ind不在上述的表格中了，因为这两个conferences从来没有team获得seed席位。一共是33个conference，外加一个independent，共计34个conference；去掉没有排上号的两个conference，还剩下图上的32个。

按照CONF分组，将performance data绘制在box plot上，寻找相差特别大的indicators。 分析..

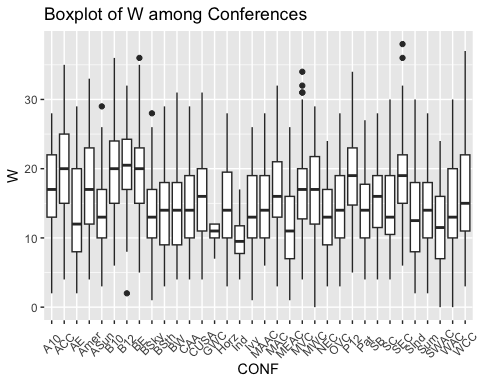
# 只需要选择特征明显的盒图就可以了，有一些波动不大直接忽略就可以了

box\_1 <-ggplot(cbb\_all, aes(x=CONF, y=G)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5)) + labs(title="Boxplot of G among Conferences",x= "CONF", y = "G")  
box\_1



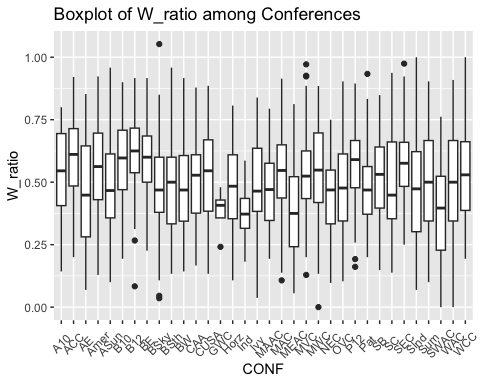
B10 is the highest, ACC, B12, BE, P12, SEC high.. Ind lowest GWC, Ivy, SInd low

box\_2 <- ggplot(cbb\_all, aes(x=CONF, y=W)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5)) + labs(title="Boxplot of W among Conferences",x= "CONF", y = "W")  
box\_2



ACC, B10, B12, BE higher Ind lower

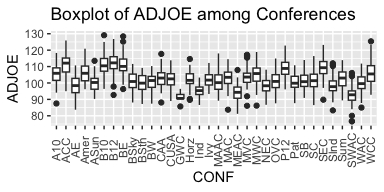
box\_3 <- ggplot(cbb\_all, aes(x=CONF, y=W\_ratio)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of W\_ratio among Conferences",x= "CONF", y = "W\_ratio")  
box\_3



ACC, B10, B12, BE high Ind, MEAC lower

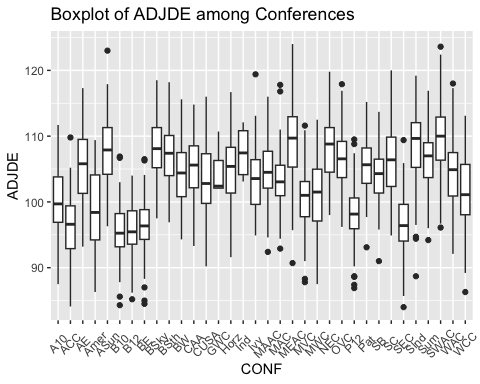
# 似乎这样放大图片在html视图中是有问题的，需要再调整！

box\_4 <- ggplot(cbb\_all, aes(x=CONF, y=ADJOE)) +   
 geom\_boxplot() +   
 theme(axis.text.x=element\_text(angle=90,vjust=0.5))+   
 labs(title="Boxplot of ADJOE among Conferences",x= "CONF", y = "ADJOE")  
box\_4



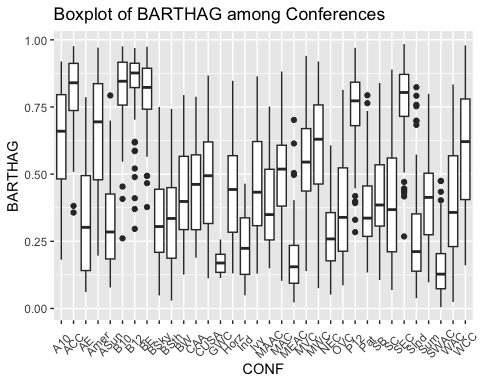
ACC, B12 highest, B10, BE high GWC, SWAC low

box\_5 <- ggplot(cbb\_all, aes(x=CONF, y=ADJDE)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of ADJDE among Conferences",x= "CONF", y = "ADJDE")  
box\_5



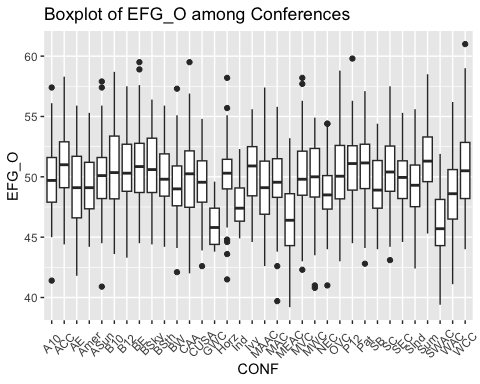
B10, B12 lowest.. ACC, BE, SEC low

box\_6 <- ggplot(cbb\_all, aes(x=CONF, y=BARTHAG)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of BARTHAG among Conferences",x= "CONF", y = "BARTHAG")  
box\_6



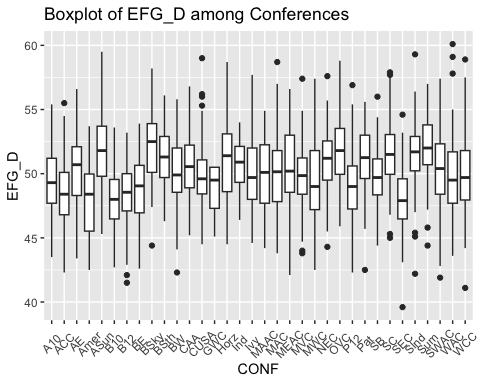
B12 highest B10, ACC high

box\_7 <- ggplot(cbb\_all, aes(x=CONF, y=EFG\_O)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5)) + labs(title="Boxplot of EFG\_O among Conferences",x= "CONF", y = "EFG\_O")  
box\_7



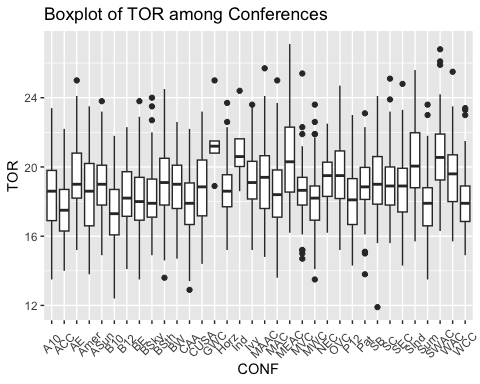
GWC, SWAC lowest 其他比较平均

box\_8 <- ggplot(cbb\_all, aes(x=CONF, y=EFG\_D)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of EFG\_D among Conferences",x= "CONF", y = "EFG\_D")  
box\_8



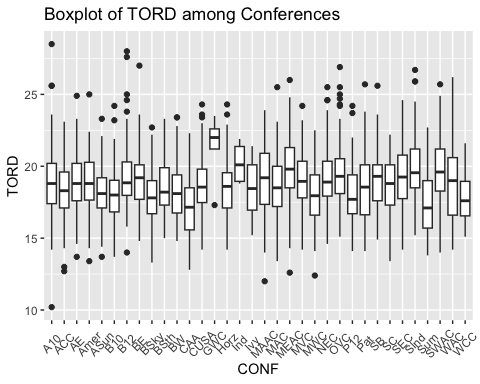
整体比较平均

box\_9 <- ggplot(cbb\_all, aes(x=CONF, y=TOR)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of TOR among Conferences",x= "CONF", y = "TOR")  
box\_9



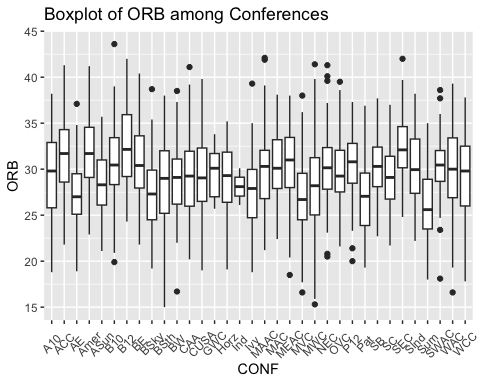
整体比较平均

box\_10 <- ggplot(cbb\_all, aes(x=CONF, y=TORD)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of TORD among Conferences",x= "CONF", y = "TORD")  
box\_10



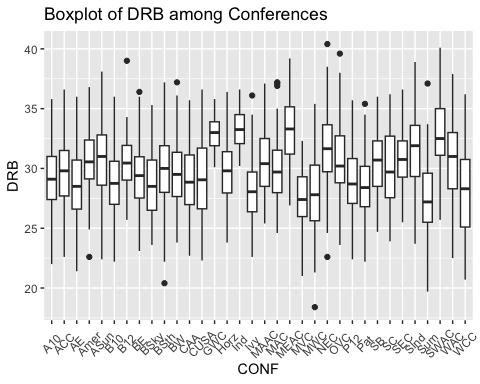
GWC highest (extreme) ind high

box\_11 <- ggplot(cbb\_all, aes(x=CONF, y=ORB)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of ORB among Conferences",x= "CONF", y = "ORB")  
box\_11



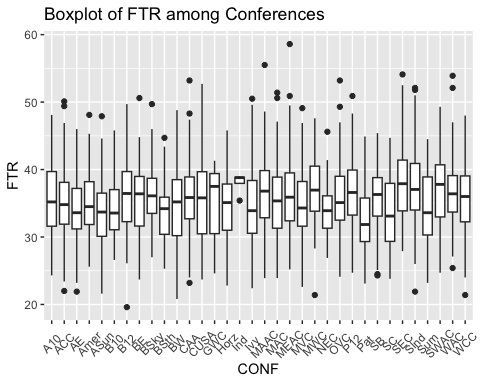
相对平均，Sum略低

box\_12 <- ggplot(cbb\_all, aes(x=CONF, y=DRB)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of DRB among Conferences",x= "CONF", y = "DRB")  
box\_12



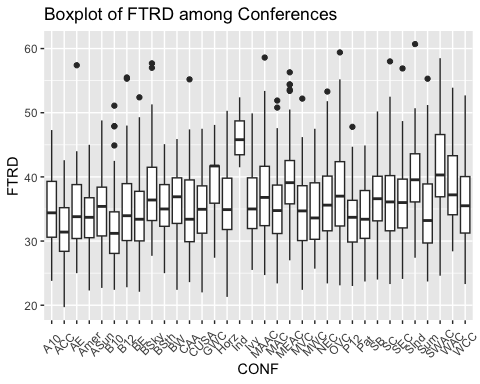
Ind, GWC, MEAC highest 其他比较平均

box\_13 <- ggplot(cbb\_all, aes(x=CONF, y=FTR)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of FTR among Conferences",x= "CONF", y = "FTR")  
box\_13



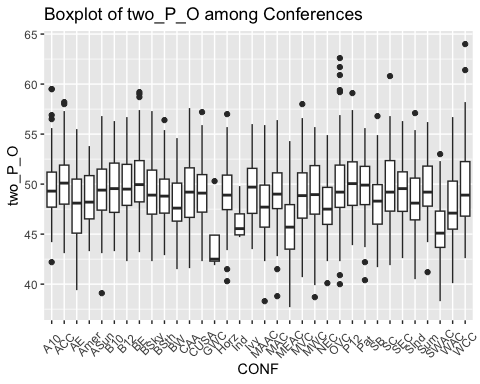
Ind slightly higher Pat lower

box\_14 <- ggplot(cbb\_all, aes(x=CONF, y=FTRD)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of FTRD among Conferences",x= "CONF", y = "FTRD")  
box\_14



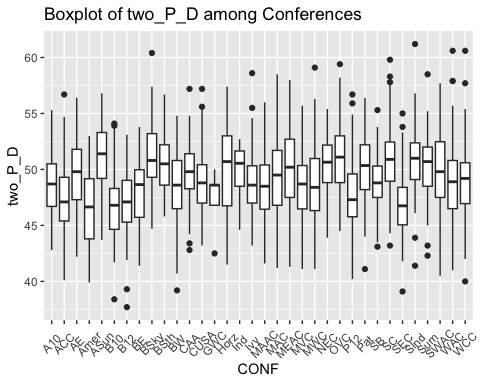
Ind highest (extreme) GWC high 其他比较平均

box\_15 <- ggplot(cbb\_all, aes(x=CONF, y=two\_P\_O)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of two\_P\_O among Conferences",x= "CONF", y = "two\_P\_O")  
box\_15



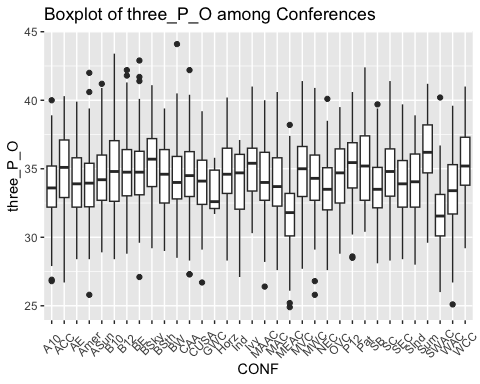
* GWC lowest (extreme).
* SWAC, Ind, MEAC low.
* 其他比较平均.

box\_16 <- ggplot(cbb\_all, aes(x=CONF, y=two\_P\_D)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of two\_P\_D among Conferences",x= "CONF", y = "two\_P\_D")  
box\_16



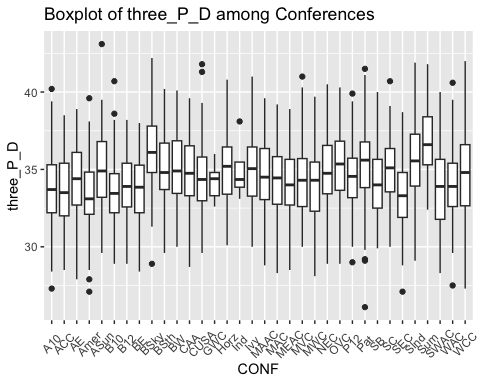
ASun relatively high Amer, SEC, ACC, B10, B12, P12 relatively low 其余波动不大

box\_17 <- ggplot(cbb\_all, aes(x=CONF, y=three\_P\_O)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of three\_P\_O among Conferences",x= "CONF", y = "three\_P\_O")  
box\_17



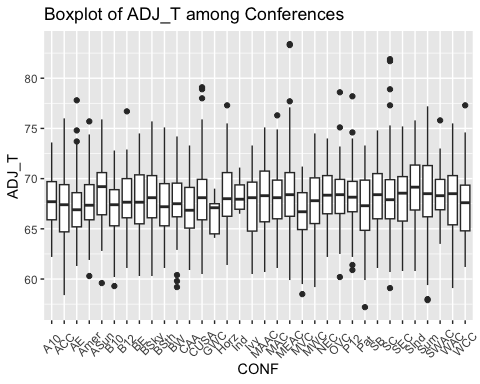
SWAC lowest MEAC low Sum slightly higher

box\_18 <- ggplot(cbb\_all, aes(x=CONF, y=three\_P\_D)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of three\_P\_D among Conferences",x= "CONF", y = "three\_P\_D")  
box\_18



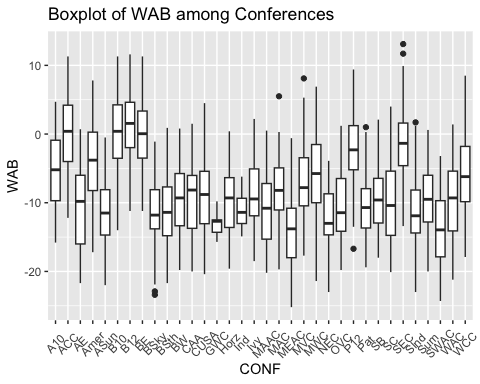
Sum, BSky relatively higher 其余波动不大

box\_19 <- ggplot(cbb\_all, aes(x=CONF, y=ADJ\_T)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of ADJ\_T among Conferences",x= "CONF", y = "ADJ\_T")  
box\_19



Asun slightly higher

box\_20 <- ggplot(cbb\_all, aes(x=CONF, y=WAB)) + geom\_boxplot() + theme(axis.text.x=element\_text(angle=45,vjust=0.5))+ labs(title="Boxplot of WAB among Conferences",x= "CONF", y = "WAB")  
box\_20



分布参差不齐 B12 highest (>0) ACC, B10, BE also very high (around 0) P12, SEC, A10 high (-5 - 0) 其他 (-15 - -5)

## 寻找影响seed的因素

seed跟所在的conference有没有关系？或许存在某个conference占据很多席位的现象？与实力的关系？

seed与postseason的关系？ (需要思考出一种方法量化postseason的排名数据 重要‼️

一些文献依据 The investigation in this area has been focused, traditionally, on men’s basketball teams (Ibáñez et al., 2003; Trninić et al., 2002). These topics have been widely studied and some important trends have been found: fields goals and defensive rebounds are the game performance indicators that best discriminate between winning and losing teams in basketball (Ibáñez et al., 2003; Ittenbach and Esters, 1995; Karipidis et al., 2001). Sampaio et al. (2010) suggest that winning teams performance is due to the achievement of more opportunities to attempt field-goals and also to the improvement in the decision making of the players of the winning teams as well as a better strategic and tactical environment. More recent studies have been exploring contextual factors such as game location (home and away), game type (regular season and playoff), game final score differences (close, balanced and unbalanced games), players’ gender (men and women), level of competition (Euroleague, National Basketball Association, etc.) and age (senior and junior) to establish better understanding of performance analysis (Lorenzo et al., 2010).

The study of game performance as a function of the differences in the final score of the game is becoming an important variable to consider. Available literature has analyzed categories separately (Sampaio and Janeira, 2003; Gómez et al., 2008). In balanced games (equal or below 12 points) the game performance indicators that discriminated between winners and losers were defensive rebounds (Gómez et al., 2008), missed 3-point field goals (Gómez et al., 2009), field goals percentages and defensive rebounds (Janeira et al., 1996) and in unbalanced games (between 13 and 28 points of final score difference). Sampaio and Janeira (2003) concluded that losing teams performed poorly in all performance indicators. Gómez et al. (2008) found that successful 2 points field-goals, defensive rebounds and assists discriminated between winning and losing teams.

Conversely, only a few studies have analyzed performance in function of a competition phase and never with high-level basketball competition. In fact, most of professional leagues are organized in two phases: firstly, all the teams compete against each other twice in a double round, with the aim of add up victories to classification. Finally, only the top classified teams (usually the best 8 teams of the regular season) play-off for the championship. The qualifying round is organized in such a way that best classified teams compete against the worst classified team and so on in a knockout round the best of three or five games. Thus, winning a game has a different importance: while in the regular season you can lose three games in a row and remain in competition, if you lose two of three, or three of five playoff games in a row you will be eliminated. Sampaio and Janeira (2002) studied Portuguese Professional Basketball League, which is a lower level competition according to the International Basketball Federation (Gómez et al., 2008). They pointed out that in the playoff games there were less ball possessions (71.16) than in the regular season games (74.75). Hence, the game rhythm was slower with a direct impact on the points per game, reducing the final score. There are no more studies inspecting this contrast between regular season and playoff games. Therefore, the aim of the present study was to identify basketball game performance indicators, which best discriminate between winners and losers in regular season and playoff games.

# 做一个Correlation Analysis

library(DT)  
cbb\_noteam <- subset(cbb\_all, select = -c(TEAM,CONF))  
cbb\_noteam

## YEAR SEED POSTSEASON G W W\_ratio ADJOE ADJDE BARTHAG EFG\_O EFG\_D TOR  
## 1: 2013 1 16 36 29 0.8055556 121.0 89.7 0.9692 54.7 44.0 19.3  
## 2: 2013 1 32 34 31 0.9117647 118.9 90.2 0.9599 54.9 44.9 17.2  
## 3: 2013 1 16 37 31 0.8378378 111.6 86.2 0.9514 53.3 41.5 20.3  
## 4: 2013 1 1 40 35 0.8750000 115.9 84.5 0.9743 50.6 44.8 18.3  
## 5: 2013 2 64 32 25 0.7812500 107.6 85.0 0.9381 51.1 43.0 20.1  
## ---   
## 2798: 2021 NA NA 27 21 0.7777778 102.7 93.4 0.7479 50.5 45.6 18.4  
## 2799: 2021 NA NA 29 21 0.7241379 113.3 101.8 0.7743 54.3 48.3 15.5  
## 2800: 2021 NA NA 27 22 0.8148148 102.5 94.6 0.7153 48.6 47.0 15.6  
## 2801: 2021 NA NA 27 22 0.8148148 101.5 102.3 0.4749 51.2 51.0 16.6  
## 2802: 2021 NA NA 29 26 0.8965517 108.5 101.6 0.6786 56.3 49.3 16.6  
## TORD ORB DRB FTR FTRD two\_P\_O two\_P\_D three\_P\_O three\_P\_D ADJ\_T WAB  
## 1: 20.9 39.0 31.4 45.8 27.0 52.0 43.2 40.3 30.4 67.8 7.8  
## 2: 20.8 37.8 29.8 40.8 29.9 55.0 42.1 36.5 32.9 65.1 7.6  
## 3: 18.4 33.8 29.3 39.5 32.0 52.9 39.3 36.4 30.3 67.7 7.5  
## 4: 27.0 38.2 33.3 40.0 34.9 50.8 43.4 33.3 31.8 67.1 9.0  
## 5: 22.4 30.4 31.0 36.8 35.3 50.2 41.4 35.3 30.7 62.5 6.6  
## ---   
## 2798: 18.8 27.1 23.4 35.3 26.4 49.7 46.4 34.6 29.6 69.6 -1.7  
## 2799: 17.4 29.6 30.7 28.5 23.4 52.5 51.0 37.7 29.5 69.1 -2.1  
## 2800: 23.6 29.0 25.4 29.9 27.1 49.4 46.1 31.1 32.4 67.5 -2.7  
## 2801: 24.3 26.8 30.2 25.6 32.2 49.6 51.8 35.8 33.0 75.1 -3.8  
## 2802: 20.3 26.1 27.2 27.9 24.6 59.4 48.2 34.7 34.0 71.1 -1.1

cbb\_corr <- na.omit(cbb\_noteam)  
# cor(cbb\_corr) %>% as.data.frame() %>% datatable()  
  
# 转换成dataframe  
# KableExtra - package  
# DT - package datatable()  
# Hmisc - package rcorr()

First of all, correlation ranges from -1 to 1. It gives us an indication on two things:

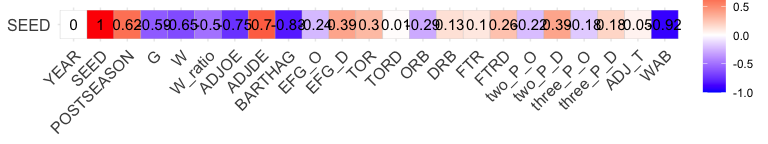
The direction of the relationship between the 2 variables The strength of the relationship between the 2 variables

Regarding the strength of the relationship: The more extreme the correlation coefficient (the closer to -1 or 1), the stronger the relationship. This also means that a correlation close to 0 indicates that the two variables are independent, that is, as one variable increases, there is no tendency in the other variable to either decrease or increase.

(不用特意将年份分出来 # 可以考虑一下只画seed和所有变量的关系，重点关注seed和postseason的关系 # knit之后效果不好，需要调整；或许还是采取之前的策略

# fig.align='center',fig.height=4, fig.width=4  
  
ggcorrplot(cor(cbb\_corr[,"SEED"], cbb\_corr),lab = TRUE) + coord\_flip()

## Coordinate system already present. Adding new coordinate system, which will  
## replace the existing one.



年份(year)那一行/最下面的一行，颜色特别淡，说明year和其他的数据几乎是independent的关系(no correlation) 主要关注点在seed(主要是想通过regular season的数据来预测seed)和postseason(探究seed和postseason的关系)上！

positively correlated: seed and postseason (r= 0.620798514), seed and ADJDE (r= 0.697441023), EFG\_D (r= 0.38708232), two\_P\_D (r= 0.38995196)

negatively correlated: seed and WAB (r= -0.918860161强相关), seed and BARTHAG (r= -0.826583963强相关), seed and ADJOE (r= -0.745242957), seed and W (r= -0.649496188), G (r= -0.58732506) ,W\_ratio (r= -0.49766516)

## Modeling

set.seed() train data test data lm() function with summary() goal: minimize MSE and MAE

K-fold cross-Validation <https://www.geeksforgeeks.org/cross-validation-in-r-programming/>

# setting seed to generate a reproducible random sampling  
cbb\_model = cbb\_corr  
set.seed(6)  
  
# creating training data as 80% of the dataset  
random\_sample <- createDataPartition(cbb\_model$SEED,  
 p = 0.8, list = FALSE)  
# generating training dataset from the random\_sample  
training\_dataset <- cbb\_model[random\_sample, ]  
# generating testing dataset from rows which are not included in random\_sample  
testing\_dataset <- cbb\_model[-random\_sample, ]  
# Building the model  
model <- lm(SEED ~., data = training\_dataset)  
# Predicting the target variable  
predictions <- predict(model, testing\_dataset)  
# Computing model performance metrics  
data.frame( R2 = R2(predictions, testing\_dataset$SEED),  
 RMSE = RMSE(predictions, testing\_dataset$SEED),  
 MAE = MAE(predictions, testing\_dataset$SEED))

## R2 RMSE MAE  
## 1 0.8829255 1.642301 1.33846

# K-fold Cross-Validation  
train\_control <- trainControl(method = "cv",  
 number = 10)  
model <- train(SEED ~., data = cbb\_model,  
 method = "lm",  
 trControl = train\_control)  
print(model)

## Linear Regression   
##   
## 476 samples  
## 22 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 428, 428, 428, 428, 429, 428, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.588502 0.8851522 1.27  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Stepwise Regression in R 找更多model分析，用k-fold cross validation计算RMSE, Rsquared, MAE

positively correlated: seed and postseason (r= 0.620798514), seed and ADJDE (r= 0.697441023), seed and EFG\_D (r= 0.38708232), seed and two\_P\_D (r= 0.38995196)

negatively correlated: seed and WAB (r= -0.918860161强相关), seed and BARTHAG (r= -0.826583963强相关), seed and ADJOE (r= -0.745242957), seed and W (r= -0.649496188), seed and G (r= -0.58732506) , seed and W\_ratio (r= -0.49766516)

cbb\_train\_1 <- subset(training\_dataset, select = c(SEED,G,W,W\_ratio,ADJOE,ADJDE,BARTHAG,EFG\_D,two\_P\_O,WAB))  
cbb\_train\_2 <- subset(training\_dataset, select = -c(YEAR,POSTSEASON))  
  
  
#define intercept-only model  
intercept\_only\_1 <- lm(SEED ~ 1, data= cbb\_train\_1)  
#define model with all predictors  
all\_1 <- lm(SEED ~ ., data= cbb\_train\_1)  
#perform backward stepwise regression  
model\_both\_1 <- step(intercept\_only\_1, direction='both', scope=formula(all\_1), trace=1)

## Start: AIC=1179.07  
## SEED ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + WAB 1 7011.3 1311.6 475.21  
## + BARTHAG 1 5689.1 2633.8 741.55  
## + ADJOE 1 4517.0 3805.8 882.17  
## + ADJDE 1 4139.2 4183.6 918.32  
## + W 1 3595.7 4727.2 964.98  
## + G 1 2789.6 5533.2 1025.13  
## + W\_ratio 1 2193.1 6129.7 1064.24  
## + EFG\_D 1 1438.3 6884.6 1108.60  
## + two\_P\_O 1 339.3 7983.6 1165.17  
## <none> 8322.9 1179.07  
##   
## Step: AIC=475.21  
## SEED ~ WAB  
##   
## Df Sum of Sq RSS AIC  
## + W\_ratio 1 155.8 1155.7 428.89  
## + G 1 118.9 1192.7 440.92  
## + W 1 32.4 1279.1 467.65  
## + ADJOE 1 26.9 1284.6 469.30  
## + ADJDE 1 16.7 1294.8 472.31  
## + two\_P\_O 1 13.9 1297.7 473.15  
## + EFG\_D 1 8.4 1303.2 474.76  
## <none> 1311.6 475.21  
## + BARTHAG 1 0.0 1311.5 477.21  
## - WAB 1 7011.3 8322.9 1179.07  
##   
## Step: AIC=428.89  
## SEED ~ WAB + W\_ratio  
##   
## Df Sum of Sq RSS AIC  
## + W 1 97.8 1057.9 397.12  
## + G 1 83.8 1072.0 402.16  
## + BARTHAG 1 19.0 1136.7 424.55  
## + EFG\_D 1 18.0 1137.7 424.90  
## + ADJOE 1 11.0 1144.7 427.24  
## + ADJDE 1 6.5 1149.2 428.74  
## <none> 1155.7 428.89  
## + two\_P\_O 1 1.1 1154.6 430.52  
## - W\_ratio 1 155.8 1311.6 475.21  
## - WAB 1 4974.0 6129.7 1064.24  
##   
## Step: AIC=397.12  
## SEED ~ WAB + W\_ratio + W  
##   
## Df Sum of Sq RSS AIC  
## + G 1 16.6 1041.4 393.09  
## + EFG\_D 1 10.1 1047.9 395.47  
## + BARTHAG 1 9.2 1048.7 395.79  
## <none> 1057.9 397.12  
## + ADJOE 1 4.0 1053.9 397.67  
## + ADJDE 1 3.6 1054.3 397.81  
## + two\_P\_O 1 0.3 1057.6 399.00  
## - W 1 97.8 1155.7 428.89  
## - W\_ratio 1 221.2 1279.1 467.65  
## - WAB 1 3163.8 4221.8 923.79  
##   
## Step: AIC=393.09  
## SEED ~ WAB + W\_ratio + W + G  
##   
## Df Sum of Sq RSS AIC  
## + EFG\_D 1 9.01 1032.3 391.77  
## <none> 1041.4 393.09  
## + ADJDE 1 4.77 1036.6 393.34  
## + BARTHAG 1 4.15 1037.2 393.57  
## + ADJOE 1 3.05 1038.3 393.97  
## + two\_P\_O 1 0.06 1041.3 395.07  
## - G 1 16.56 1057.9 397.12  
## - W 1 30.61 1072.0 402.16  
## - W\_ratio 1 41.41 1082.8 405.99  
## - WAB 1 3127.18 4168.5 920.94  
##   
## Step: AIC=391.77  
## SEED ~ WAB + W\_ratio + W + G + EFG\_D  
##   
## Df Sum of Sq RSS AIC  
## + ADJOE 1 13.12 1019.2 388.89  
## + BARTHAG 1 8.20 1024.2 390.73  
## <none> 1032.3 391.77  
## - EFG\_D 1 9.01 1041.4 393.09  
## + two\_P\_O 1 1.62 1030.7 393.17  
## + ADJDE 1 0.00 1032.3 393.77  
## - G 1 15.50 1047.9 395.47  
## - W 1 28.57 1060.9 400.20  
## - W\_ratio 1 39.45 1071.8 404.10  
## - WAB 1 2967.00 3999.3 907.11  
##   
## Step: AIC=388.89  
## SEED ~ WAB + W\_ratio + W + G + EFG\_D + ADJOE  
##   
## Df Sum of Sq RSS AIC  
## + BARTHAG 1 25.27 993.96 381.30  
## <none> 1019.23 388.89  
## + ADJDE 1 4.73 1014.50 389.11  
## + two\_P\_O 1 0.29 1018.93 390.78  
## - G 1 12.77 1032.00 391.64  
## - ADJOE 1 13.12 1032.35 391.77  
## - EFG\_D 1 19.08 1038.31 393.97  
## - W 1 23.46 1042.69 395.58  
## - W\_ratio 1 32.95 1052.18 399.04  
## - WAB 1 1310.48 2329.71 702.69  
##   
## Step: AIC=381.3  
## SEED ~ WAB + W\_ratio + W + G + EFG\_D + ADJOE + BARTHAG  
##   
## Df Sum of Sq RSS AIC  
## + ADJDE 1 82.90 911.06 350.03  
## - G 1 2.54 996.50 380.27  
## <none> 993.96 381.30  
## - W 1 6.55 1000.51 381.81  
## + two\_P\_O 1 0.28 993.68 383.19  
## - W\_ratio 1 12.96 1006.92 384.25  
## - BARTHAG 1 25.27 1019.23 388.89  
## - ADJOE 1 30.20 1024.15 390.73  
## - EFG\_D 1 37.64 1031.60 393.50  
## - WAB 1 822.79 1816.74 609.68  
##   
## Step: AIC=350.03  
## SEED ~ WAB + W\_ratio + W + G + EFG\_D + ADJOE + BARTHAG + ADJDE  
##   
## Df Sum of Sq RSS AIC  
## - W 1 0.07 911.13 348.06  
## - W\_ratio 1 0.40 911.46 348.20  
## - G 1 0.41 911.47 348.20  
## - EFG\_D 1 1.31 912.37 348.58  
## <none> 911.06 350.03  
## + two\_P\_O 1 0.03 911.04 352.02  
## - ADJDE 1 82.90 993.96 381.30  
## - BARTHAG 1 103.44 1014.50 389.11  
## - ADJOE 1 111.75 1022.81 392.23  
## - WAB 1 573.96 1485.03 534.67  
##   
## Step: AIC=348.06  
## SEED ~ WAB + W\_ratio + G + EFG\_D + ADJOE + BARTHAG + ADJDE  
##   
## Df Sum of Sq RSS AIC  
## - EFG\_D 1 1.27 912.40 346.59  
## <none> 911.13 348.06  
## - G 1 5.82 916.96 348.49  
## + W 1 0.07 911.06 350.03  
## + two\_P\_O 1 0.03 911.10 350.05  
## - ADJDE 1 89.37 1000.51 381.81  
## - W\_ratio 1 123.24 1034.38 394.52  
## - BARTHAG 1 128.56 1039.69 396.48  
## - ADJOE 1 129.96 1041.09 396.99  
## - WAB 1 584.36 1495.49 535.35  
##   
## Step: AIC=346.59  
## SEED ~ WAB + W\_ratio + G + ADJOE + BARTHAG + ADJDE  
##   
## Df Sum of Sq RSS AIC  
## <none> 912.40 346.59  
## - G 1 5.84 918.24 347.03  
## + EFG\_D 1 1.27 911.13 348.06  
## + two\_P\_O 1 0.05 912.35 348.57  
## + W 1 0.03 912.37 348.58  
## - W\_ratio 1 124.19 1036.60 393.34  
## - ADJOE 1 133.98 1046.38 396.93  
## - BARTHAG 1 135.80 1048.20 397.59  
## - ADJDE 1 137.78 1050.18 398.31  
## - WAB 1 594.00 1506.41 536.13

#view results of backward stepwise regression  
model\_both\_1$anova

## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 NA NA 381 8322.8796 1179.0730  
## 2 + WAB -1 7.011327e+03 380 1311.5527 475.2147  
## 3 + W\_ratio -1 1.558366e+02 379 1155.7161 428.8949  
## 4 + W -1 9.779921e+01 378 1057.9169 397.1191  
## 5 + G -1 1.656287e+01 377 1041.3540 393.0912  
## 6 + EFG\_D -1 9.005870e+00 376 1032.3481 391.7732  
## 7 + ADJOE -1 1.311952e+01 375 1019.2286 388.8874  
## 8 + BARTHAG -1 2.527211e+01 374 993.9565 381.2962  
## 9 + ADJDE -1 8.289533e+01 373 911.0612 350.0304  
## 10 - W 1 7.224775e-02 374 911.1334 348.0607  
## 11 - EFG\_D 1 1.269142e+00 375 912.4025 346.5924

#view final model  
model\_both\_1$coefficients

## (Intercept) WAB W\_ratio G ADJOE BARTHAG   
## -4.33766090 -0.82986380 7.99275351 -0.07521933 -0.29202360 14.16200872   
## ADJDE   
## 0.33506762

#define intercept-only model  
intercept\_only\_2 <- lm(SEED ~ 1, data= cbb\_train\_2)  
#define model with all predictors  
all\_2 <- lm(SEED ~ ., data= cbb\_train\_2)  
#perform backward stepwise regression  
model\_both\_2 <- step(intercept\_only\_2, direction='both', scope=formula(all\_2), trace=1)

## Start: AIC=1179.07  
## SEED ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + WAB 1 7011.3 1311.6 475.21  
## + BARTHAG 1 5689.1 2633.8 741.55  
## + ADJOE 1 4517.0 3805.8 882.17  
## + ADJDE 1 4139.2 4183.6 918.32  
## + W 1 3595.7 4727.2 964.98  
## + G 1 2789.6 5533.2 1025.13  
## + W\_ratio 1 2193.1 6129.7 1064.24  
## + EFG\_D 1 1438.3 6884.6 1108.60  
## + two\_P\_D 1 1348.6 6974.3 1113.54  
## + TOR 1 680.1 7642.8 1148.51  
## + ORB 1 644.5 7678.3 1150.28  
## + FTRD 1 554.5 7768.4 1154.74  
## + three\_P\_D 1 439.0 7883.9 1160.38  
## + EFG\_O 1 432.3 7890.5 1160.70  
## + two\_P\_O 1 339.3 7983.6 1165.17  
## + three\_P\_O 1 275.3 8047.6 1168.22  
## + DRB 1 147.3 8175.6 1174.25  
## + FTR 1 48.6 8274.2 1178.83  
## <none> 8322.9 1179.07  
## + ADJ\_T 1 4.9 8318.0 1180.85  
## + TORD 1 4.9 8318.0 1180.85  
##   
## Step: AIC=475.21  
## SEED ~ WAB  
##   
## Df Sum of Sq RSS AIC  
## + W\_ratio 1 155.8 1155.7 428.89  
## + G 1 118.9 1192.7 440.92  
## + W 1 32.4 1279.1 467.65  
## + ADJOE 1 26.9 1284.6 469.30  
## + FTRD 1 23.9 1287.6 470.18  
## + DRB 1 19.5 1292.0 471.48  
## + two\_P\_D 1 19.4 1292.2 471.53  
## + ADJDE 1 16.7 1294.8 472.31  
## + two\_P\_O 1 13.9 1297.7 473.15  
## + ORB 1 12.1 1299.5 473.68  
## + EFG\_D 1 8.4 1303.2 474.76  
## + TOR 1 8.3 1303.2 474.78  
## + EFG\_O 1 6.9 1304.6 475.19  
## <none> 1311.6 475.21  
## + TORD 1 6.1 1305.4 475.42  
## + ADJ\_T 1 3.0 1308.6 476.35  
## + three\_P\_D 1 0.8 1310.7 476.97  
## + FTR 1 0.5 1311.0 477.07  
## + BARTHAG 1 0.0 1311.5 477.21  
## + three\_P\_O 1 0.0 1311.6 477.21  
## - WAB 1 7011.3 8322.9 1179.07  
##   
## Step: AIC=428.89  
## SEED ~ WAB + W\_ratio  
##   
## Df Sum of Sq RSS AIC  
## + W 1 97.8 1057.9 397.12  
## + G 1 83.8 1072.0 402.16  
## + two\_P\_D 1 27.6 1128.2 421.67  
## + FTRD 1 22.2 1133.5 423.48  
## + BARTHAG 1 19.0 1136.7 424.55  
## + EFG\_D 1 18.0 1137.7 424.90  
## + ADJOE 1 11.0 1144.7 427.24  
## + ORB 1 6.9 1148.9 428.62  
## + ADJDE 1 6.5 1149.2 428.74  
## <none> 1155.7 428.89  
## + TOR 1 4.9 1150.8 429.26  
## + three\_P\_O 1 4.1 1151.6 429.54  
## + EFG\_O 1 3.4 1152.3 429.77  
## + ADJ\_T 1 3.4 1152.3 429.77  
## + two\_P\_O 1 1.1 1154.6 430.52  
## + FTR 1 1.1 1154.6 430.53  
## + DRB 1 0.8 1154.9 430.62  
## + three\_P\_D 1 0.2 1155.5 430.81  
## + TORD 1 0.1 1155.6 430.86  
## - W\_ratio 1 155.8 1311.6 475.21  
## - WAB 1 4974.0 6129.7 1064.24  
##   
## Step: AIC=397.12  
## SEED ~ WAB + W\_ratio + W  
##   
## Df Sum of Sq RSS AIC  
## + two\_P\_D 1 17.2 1040.7 392.86  
## + G 1 16.6 1041.4 393.09  
## + FTRD 1 12.5 1045.4 394.58  
## + EFG\_D 1 10.1 1047.9 395.47  
## + BARTHAG 1 9.2 1048.7 395.79  
## + three\_P\_O 1 5.7 1052.2 397.05  
## <none> 1057.9 397.12  
## + TOR 1 5.1 1052.8 397.26  
## + ADJOE 1 4.0 1053.9 397.67  
## + ORB 1 3.8 1054.2 397.76  
## + ADJDE 1 3.6 1054.3 397.81  
## + EFG\_O 1 2.9 1055.0 398.08  
## + ADJ\_T 1 2.7 1055.2 398.13  
## + FTR 1 1.7 1056.2 398.51  
## + two\_P\_O 1 0.3 1057.6 399.00  
## + TORD 1 0.2 1057.7 399.04  
## + DRB 1 0.1 1057.8 399.07  
## + three\_P\_D 1 0.0 1057.9 399.11  
## - W 1 97.8 1155.7 428.89  
## - W\_ratio 1 221.2 1279.1 467.65  
## - WAB 1 3163.8 4221.8 923.79  
##   
## Step: AIC=392.86  
## SEED ~ WAB + W\_ratio + W + two\_P\_D  
##   
## Df Sum of Sq RSS AIC  
## + ADJOE 1 16.56 1024.2 388.73  
## + G 1 16.05 1024.7 388.92  
## + BARTHAG 1 15.81 1024.9 389.01  
## + FTRD 1 13.64 1027.1 389.82  
## + three\_P\_O 1 12.16 1028.6 390.37  
## + ADJ\_T 1 11.29 1029.4 390.69  
## + EFG\_O 1 11.03 1029.7 390.79  
## <none> 1040.7 392.86  
## + FTR 1 4.59 1036.1 393.17  
## + two\_P\_O 1 3.42 1037.3 393.60  
## + ORB 1 0.98 1039.7 394.50  
## + EFG\_D 1 0.93 1039.8 394.52  
## + TOR 1 0.72 1040.0 394.59  
## + ADJDE 1 0.60 1040.1 394.64  
## + three\_P\_D 1 0.16 1040.6 394.80  
## + DRB 1 0.02 1040.7 394.85  
## + TORD 1 0.00 1040.7 394.86  
## - two\_P\_D 1 17.20 1057.9 397.12  
## - W 1 87.44 1128.2 421.67  
## - W\_ratio 1 212.45 1253.2 461.82  
## - WAB 1 2985.57 4026.3 907.68  
##   
## Step: AIC=388.73  
## SEED ~ WAB + W\_ratio + W + two\_P\_D + ADJOE  
##   
## Df Sum of Sq RSS AIC  
## + BARTHAG 1 34.25 989.91 377.74  
## + G 1 13.72 1010.44 385.58  
## + FTRD 1 12.86 1011.30 385.90  
## + ADJ\_T 1 7.84 1016.32 387.79  
## <none> 1024.16 388.73  
## + TOR 1 4.56 1019.60 389.03  
## + FTR 1 3.42 1020.74 389.45  
## + ADJDE 1 3.18 1020.98 389.54  
## + TORD 1 2.87 1021.29 389.66  
## + three\_P\_O 1 2.74 1021.42 389.71  
## + EFG\_O 1 0.80 1023.35 390.43  
## + three\_P\_D 1 0.44 1023.71 390.57  
## + ORB 1 0.28 1023.88 390.63  
## + two\_P\_O 1 0.05 1024.11 390.71  
## + DRB 1 0.02 1024.13 390.72  
## + EFG\_D 1 0.01 1024.15 390.73  
## - ADJOE 1 16.56 1040.72 392.86  
## - two\_P\_D 1 29.76 1053.91 397.67  
## - W 1 70.05 1094.21 412.00  
## - W\_ratio 1 174.01 1198.17 446.67  
## - WAB 1 1368.53 2392.68 710.87  
##   
## Step: AIC=377.74  
## SEED ~ WAB + W\_ratio + W + two\_P\_D + ADJOE + BARTHAG  
##   
## Df Sum of Sq RSS AIC  
## + ADJDE 1 86.85 903.06 344.66  
## + TORD 1 12.51 977.40 374.88  
## + FTRD 1 9.83 980.08 375.93  
## + FTR 1 7.53 982.37 376.82  
## + ADJ\_T 1 6.80 983.10 377.10  
## <none> 989.91 377.74  
## + TOR 1 4.63 985.27 377.94  
## + G 1 3.69 986.22 378.31  
## + three\_P\_D 1 3.64 986.27 378.33  
## + EFG\_D 1 1.82 988.09 379.03  
## + three\_P\_O 1 1.04 988.86 379.33  
## + DRB 1 1.03 988.88 379.34  
## + ORB 1 0.28 989.63 379.63  
## + EFG\_O 1 0.19 989.72 379.67  
## + two\_P\_O 1 0.12 989.79 379.69  
## - BARTHAG 1 34.25 1024.16 388.73  
## - ADJOE 1 34.99 1024.90 389.01  
## - W 1 44.87 1034.77 392.67  
## - two\_P\_D 1 50.95 1040.86 394.91  
## - W\_ratio 1 166.50 1156.41 435.12  
## - WAB 1 891.99 1881.90 621.14  
##   
## Step: AIC=344.66  
## SEED ~ WAB + W\_ratio + W + two\_P\_D + ADJOE + BARTHAG + ADJDE  
##   
## Df Sum of Sq RSS AIC  
## + ADJ\_T 1 15.49 887.57 340.05  
## + FTRD 1 10.33 892.73 342.27  
## + EFG\_D 1 5.13 897.93 344.48  
## <none> 903.06 344.66  
## - W 1 4.80 907.86 344.69  
## + three\_P\_D 1 3.19 899.87 345.31  
## + TOR 1 2.43 900.63 345.63  
## + FTR 1 2.30 900.76 345.69  
## + G 1 0.49 902.57 346.45  
## + three\_P\_O 1 0.42 902.64 346.48  
## + DRB 1 0.23 902.83 346.56  
## + EFG\_O 1 0.17 902.89 346.59  
## + ORB 1 0.03 903.03 346.65  
## + TORD 1 0.01 903.05 346.66  
## + two\_P\_O 1 0.01 903.05 346.66  
## - two\_P\_D 1 9.61 912.68 346.71  
## - W\_ratio 1 58.68 961.74 366.71  
## - ADJDE 1 86.85 989.91 377.74  
## - BARTHAG 1 117.91 1020.98 389.54  
## - ADJOE 1 121.22 1024.28 390.78  
## - WAB 1 599.37 1502.43 537.12  
##   
## Step: AIC=340.05  
## SEED ~ WAB + W\_ratio + W + two\_P\_D + ADJOE + BARTHAG + ADJDE +   
## ADJ\_T  
##   
## Df Sum of Sq RSS AIC  
## + FTRD 1 13.43 874.13 336.22  
## - W 1 3.48 891.05 339.54  
## <none> 887.57 340.05  
## + EFG\_D 1 4.33 883.24 340.18  
## + FTR 1 2.78 884.78 340.85  
## + three\_P\_D 1 2.62 884.95 340.92  
## + TOR 1 2.11 885.46 341.14  
## + three\_P\_O 1 1.57 886.00 341.38  
## + two\_P\_O 1 0.33 887.23 341.91  
## + DRB 1 0.33 887.24 341.91  
## + G 1 0.30 887.27 341.92  
## + EFG\_O 1 0.14 887.43 341.99  
## + TORD 1 0.11 887.46 342.00  
## + ORB 1 0.08 887.49 342.02  
## - two\_P\_D 1 14.37 901.94 344.19  
## - ADJ\_T 1 15.49 903.06 344.66  
## - W\_ratio 1 54.95 942.52 361.00  
## - ADJDE 1 95.54 983.10 377.10  
## - ADJOE 1 123.16 1010.73 387.69  
## - BARTHAG 1 124.24 1011.81 388.10  
## - WAB 1 595.94 1483.51 534.28  
##   
## Step: AIC=336.22  
## SEED ~ WAB + W\_ratio + W + two\_P\_D + ADJOE + BARTHAG + ADJDE +   
## ADJ\_T + FTRD  
##   
## Df Sum of Sq RSS AIC  
## - W 1 2.22 876.35 335.19  
## + EFG\_D 1 5.22 868.91 335.94  
## <none> 874.13 336.22  
## + TORD 1 3.85 870.28 336.54  
## + three\_P\_D 1 3.57 870.56 336.66  
## + TOR 1 2.52 871.62 337.12  
## + two\_P\_O 1 2.17 871.96 337.27  
## + ORB 1 2.01 872.12 337.34  
## + FTR 1 0.58 873.55 337.97  
## + EFG\_O 1 0.54 873.59 337.99  
## + three\_P\_O 1 0.47 873.66 338.02  
## + G 1 0.27 873.86 338.11  
## + DRB 1 0.10 874.04 338.18  
## - FTRD 1 13.43 887.57 340.05  
## - two\_P\_D 1 14.72 888.86 340.61  
## - ADJ\_T 1 18.60 892.73 342.27  
## - W\_ratio 1 48.98 923.12 355.05  
## - ADJDE 1 97.19 971.32 374.50  
## - BARTHAG 1 121.07 995.20 383.77  
## - ADJOE 1 121.27 995.40 383.85  
## - WAB 1 573.94 1448.07 527.04  
##   
## Step: AIC=335.19  
## SEED ~ WAB + W\_ratio + two\_P\_D + ADJOE + BARTHAG + ADJDE + ADJ\_T +   
## FTRD  
##   
## Df Sum of Sq RSS AIC  
## + EFG\_D 1 5.53 870.83 334.78  
## <none> 876.35 335.19  
## + TORD 1 4.26 872.09 335.33  
## + three\_P\_D 1 3.87 872.49 335.50  
## + TOR 1 2.45 873.91 336.13  
## + G 1 2.41 873.94 336.14  
## + two\_P\_O 1 2.36 873.99 336.16  
## + W 1 2.22 874.13 336.22  
## + ORB 1 2.16 874.20 336.25  
## + EFG\_O 1 0.86 875.49 336.82  
## + FTR 1 0.45 875.90 337.00  
## + DRB 1 0.26 876.10 337.08  
## + three\_P\_O 1 0.20 876.16 337.11  
## - FTRD 1 14.69 891.05 339.54  
## - two\_P\_D 1 15.59 891.94 339.93  
## - ADJ\_T 1 19.95 896.31 341.79  
## - ADJDE 1 134.13 1010.48 387.60  
## - W\_ratio 1 142.11 1018.46 390.60  
## - BARTHAG 1 172.43 1048.79 401.81  
## - ADJOE 1 177.33 1053.68 403.59  
## - WAB 1 586.63 1462.98 528.95  
##   
## Step: AIC=334.78  
## SEED ~ WAB + W\_ratio + two\_P\_D + ADJOE + BARTHAG + ADJDE + ADJ\_T +   
## FTRD + EFG\_D  
##   
## Df Sum of Sq RSS AIC  
## <none> 870.83 334.78  
## - EFG\_D 1 5.53 876.35 335.19  
## + two\_P\_O 1 2.37 868.46 335.74  
## + ORB 1 2.34 868.49 335.75  
## + TOR 1 2.32 868.51 335.76  
## + three\_P\_D 1 2.14 868.68 335.84  
## + G 1 2.03 868.80 335.88  
## + W 1 1.92 868.91 335.94  
## + TORD 1 1.88 868.95 335.95  
## + DRB 1 1.08 869.75 336.30  
## + EFG\_O 1 1.04 869.79 336.32  
## + three\_P\_O 1 0.11 870.72 336.73  
## + FTR 1 0.09 870.74 336.74  
## - FTRD 1 15.58 886.41 339.55  
## - two\_P\_D 1 18.32 889.15 340.73  
## - ADJ\_T 1 18.96 889.79 341.01  
## - W\_ratio 1 126.62 997.44 384.63  
## - ADJDE 1 134.76 1005.58 387.74  
## - BARTHAG 1 177.62 1048.44 403.68  
## - ADJOE 1 181.69 1052.51 405.16  
## - WAB 1 558.16 1428.98 521.97

#view results of backward stepwise regression  
model\_both\_2$anova

## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 NA NA 381 8322.8796 1179.0730  
## 2 + WAB -1 7011.326928 380 1311.5527 475.2147  
## 3 + W\_ratio -1 155.836592 379 1155.7161 428.8949  
## 4 + W -1 97.799209 378 1057.9169 397.1191  
## 5 + two\_P\_D -1 17.200240 377 1040.7166 392.8573  
## 6 + ADJOE -1 16.559531 376 1024.1571 388.7302  
## 7 + BARTHAG -1 34.249465 375 989.9076 377.7370  
## 8 + ADJDE -1 86.845108 374 903.0625 344.6618  
## 9 + ADJ\_T -1 15.494059 373 887.5684 340.0508  
## 10 + FTRD -1 13.433691 372 874.1348 336.2249  
## 11 - W 1 2.219737 373 876.3545 335.1937  
## 12 + EFG\_D -1 5.528339 372 870.8262 334.7763

#view final model  
model\_both\_2$coefficients

## (Intercept) WAB W\_ratio two\_P\_D ADJOE BARTHAG   
## -6.44869443 -0.82091537 8.24557772 0.17145355 -0.30972253 15.13874804   
## ADJDE ADJ\_T FTRD EFG\_D   
## 0.36305352 -0.07591390 0.03701627 -0.13172055

# Predicting the target variable  
predictions\_1 <- predict(model\_both\_1, testing\_dataset)  
# Computing model performance metrics  
data.frame( R2 = R2(predictions\_1, testing\_dataset$SEED),  
 RMSE = RMSE(predictions\_1, testing\_dataset$SEED),  
 MAE = MAE(predictions\_1, testing\_dataset$SEED))

## R2 RMSE MAE  
## 1 0.8934268 1.565535 1.263876

# Predicting the target variable  
predictions\_2 <- predict(model\_both\_2, testing\_dataset)  
# Computing model performance metrics  
data.frame( R2 = R2(predictions\_2, testing\_dataset$SEED),  
 RMSE = RMSE(predictions\_2, testing\_dataset$SEED),  
 MAE = MAE(predictions\_2, testing\_dataset$SEED))

## R2 RMSE MAE  
## 1 0.8843884 1.644719 1.329491

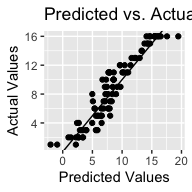
按照model performance metrics来看，both\_1表现最好，其次是both\_2，再是model(full)。

model\_1: SEED ~ -4.33766090 + -0.82986380\* WAB + 7.99275351\* W\_ratio + -0.07521933\* G + -0.29202360\* ADJOE + 14.16200872\* BARTHAG + 0.33506762\* ADJDE model\_2: SEED~ -6.44869443+ -0.82091537\* WAB + 8.24557772\* W\_ratio + 0.17145355\* two\_P\_D + 15.13874804\* BARTHAG + -0.30972253\* ADJOE + 0.36305352\* ADJDE + -0.07591390\* ADJ\_T + 0.03701627\* FTRD + -0.13172055\* EFG\_D

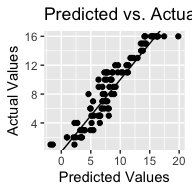
将test model的数据点绘制在图上，观察是否拟合(可参考STOR320的final paper)。

x-axis observed values y-axis predicted values 观察是不是线性相关，加regression line residual plot 是否接近0

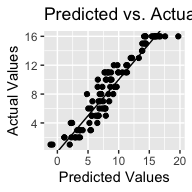
ggplot(testing\_dataset, aes(x=predict(model\_both\_1,newdata = testing\_dataset), y= testing\_dataset$SEED)) +  
 geom\_point() +  
 geom\_abline(intercept=0, slope=1) +  
 labs(x='Predicted Values', y='Actual Values', title='Predicted vs. Actual Values')



ggplot(testing\_dataset, aes(x=predict(model\_both\_2,newdata = testing\_dataset), y= testing\_dataset$SEED)) +  
 geom\_point() +  
 geom\_abline(intercept=0, slope=1) +  
 labs(x='Predicted Values', y='Actual Values', title='Predicted vs. Actual Values')



ggplot(testing\_dataset, aes(x=predict(model,newdata = testing\_dataset), y= testing\_dataset$SEED)) +  
 geom\_point() +  
 geom\_abline(intercept=0, slope=1) +  
 labs(x='Predicted Values', y='Actual Values', title='Predicted vs. Actual Values')



Conclusion and Discussion