NBA Playoff Factors

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

"In this NBA Analysis, we see which variables/factors play a key role in the NBA team wining the playoffs"

```r
nbaAllDf = read.csv("allYears.csv",header=TRUE)

# drop team names
drops <- c("Team")
nbaAllDf <- nbaAllDf[ , !(names(nbaAllDf) %in% drops)]
names(nbaAllDf)[names(nbaAllDf) == 'ELIETE.DEF.SCORE'] <- 'ELITE.DEF'
names(nbaAllDf)[names(nbaAllDf) == 'ELIETE.OFF.SCORE'] <- 'ELITE.OFF'


# summery
head(nbaAllDf)
```

```
##   GP  W  L      W.L.    G   GA DIFF CQF.GP   CQF.W.L. CSF.GP   CSF.W.L.
## 1 53 39 14 0.7358491 5131 4999  132      6 0.6666667      6 0.6666667
## 2 48 27 21 0.5625000 4484 4417   67      6 0.6666667      4 1.0000000
## 3 49 29 20 0.5918367 5075 5087  -12      7 0.5714286      7 0.5714286
## 4 46 21 25 0.4565217 4213 4198   15      7 0.5714286      6 0.6666667
## 5 39 29 10 0.7435897 4093 4065   28      5 0.8000000      7 0.4285714
## 6 41 25 16 0.6097561 4005 3995   10      6 0.6666667      6 0.3333333
##   CF.GP   CF.W.L. F.GP    F.W.L. PF.ELITE.OFF REG.ELITE.OFF REG.DEF
## 1     6 0.6666667    6 0.6666667            5             2       2
## 2     4 1.0000000    6 0.3333333            6             3       1
## 3     6 0.3333333    0 0.0000000            4             2       1
## 4     4 0.0000000    0 0.0000000            2             0       1
## 5     0 0.0000000    0 0.0000000            5             1       2
## 6     0 0.0000000    0 0.0000000            2             0       0
##   PO.ELITE.DEF ELITE.OFF ELITE.DEF Champ Rk  X3P.  X2P.   FT.
## 1            6      4.25      5.00     3  1 0.351 0.463 0.712
## 2            6      5.25      4.75     2  2 0.350 0.439 0.762
## 3            7      3.50      5.50     2  3 0.397 0.469 0.829
## 4            4      1.50      3.25     1  4 0.352 0.430 0.741
## 5            4      4.00      3.50     1  5 0.367 0.484 0.777
## 6            2      1.50      1.50     1  6 0.352 0.458 0.803
```

```r
names(nbaAllDf)
```

```
##  [1] "GP"            "W"             "L"             "W.L."
##  [5] "G"             "GA"            "DIFF"          "CQF.GP"
##  [9] "CQF.W.L."      "CSF.GP"        "CSF.W.L."      "CF.GP"
## [13] "CF.W.L."       "F.GP"          "F.W.L."        "PF.ELITE.OFF"
## [17] "REG.ELITE.OFF" "REG.DEF"       "PO.ELITE.DEF"  "ELITE.OFF"
## [21] "ELITE.DEF"     "Champ"         "Rk"            "X3P."
## [25] "X2P."          "FT."
```

```r
trainDf =read.csv("train.csv",header=TRUE)

# drop team names
drops <- c("Team")
trainDf <- trainDf[ , !(names(trainDf) %in% drops)]
names(trainDf)[names(trainDf) == 'ELIETE.DEF.SCORE'] <- 'ELITE.DEF'
names(trainDf)[names(trainDf) == 'ELIETE.OFF.SCORE'] <- 'ELITE.OFF'

attach(trainDf)
summary(trainDf)
```

```
##        GP              W               L              W.L.       
##  Min.   :29.00   Min.   :10.00   Min.   : 8.00   Min.   :0.3143  
##  1st Qu.:33.00   1st Qu.:16.00   1st Qu.:13.75   1st Qu.:0.4688  
##  Median :36.00   Median :21.00   Median :16.50   Median :0.5736  
##  Mean   :38.21   Mean   :21.95   Mean   :16.26   Mean   :0.5645  
##  3rd Qu.:43.00   3rd Qu.:26.00   3rd Qu.:19.00   3rd Qu.:0.6286  
##  Max.   :56.00   Max.   :45.00   Max.   :25.00   Max.   :0.8333  
##        G              GA            DIFF              CQF.GP     
##  Min.   :2726   Min.   :2769   Min.   :-94.000   Min.   :4.000  
##  1st Qu.:3299   1st Qu.:3340   1st Qu.:-41.000   1st Qu.:5.000  
##  Median :3546   Median :3584   Median :-19.500   Median :6.000  
##  Mean   :3794   Mean   :3796   Mean   : -1.839   Mean   :5.571  
##  3rd Qu.:4220   3rd Qu.:4211   3rd Qu.: 15.250   3rd Qu.:6.000  
##  Max.   :5462   Max.   :5326   Max.   :214.000   Max.   :7.000  
##     CQF.W.L.          CSF.GP        CSF.W.L.          CF.GP      
##  Min.   :0.0000   Min.   :0.00   Min.   :0.0000   Min.   :0.000  
##  1st Qu.:0.2000   1st Qu.:0.00   1st Qu.:0.0000   1st Qu.:0.000  
##  Median :0.5714   Median :4.00   Median :0.0000   Median :0.000  
##  Mean   :0.4937   Mean   :2.92   Mean   :0.2503   Mean   :1.321  
##  3rd Qu.:0.6667   3rd Qu.:6.00   3rd Qu.:0.4286   3rd Qu.:0.000  
##  Max.   :1.0000   Max.   :7.00   Max.   :1.0000   Max.   :7.000  
##     CF.W.L.           F.GP            F.W.L.          PF.ELITE.OFF  
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :2.000  
##  Mean   :0.1244   Mean   :0.7768   Mean   :0.07092   Mean   :2.339  
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :7.0000   Max.   :1.00000   Max.   :6.000  
##  REG.ELITE.OFF      REG.DEF        PO.ELITE.DEF     ELITE.OFF    
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000  
##  1st Qu.:0.000   1st Qu.:0.7500   1st Qu.:0.000   1st Qu.:0.750  
##  Median :1.000   Median :1.0000   Median :2.000   Median :1.625  
##  Mean   :0.875   Mean   :0.9286   Mean   :2.277   Mean   :1.973  
##  3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:3.250  
##  Max.   :4.000   Max.   :2.0000   Max.   :8.000   Max.   :5.250  
##    ELITE.DEF        Champ              Rk              X3P.       
##  Min.   :0.00   Min.   :0.0000   Min.   : 1.000   Min.   :0.2450  
##  1st Qu.:0.25   1st Qu.:0.0000   1st Qu.: 5.000   1st Qu.:0.3167  
##  Median :1.50   Median :0.0000   Median : 8.000   Median :0.3330  
##  Mean   :1.94   Mean   :0.6607   Mean   : 8.348   Mean   :0.3387  
##  3rd Qu.:3.25   3rd Qu.:1.0000   3rd Qu.:12.000   3rd Qu.:0.3650  
##  Max.   :6.25   Max.   :3.0000   Max.   :16.000   Max.   :0.4190  
##       X2P.              FT.
```

```
##  Min.   :0.4050   Min.   :0.5790
##  1st Qu.:0.4520   1st Qu.:0.7240
##  Median :0.4775   Median :0.7540
##  Mean   :0.4768   Mean   :0.7508
##  3rd Qu.:0.4990   3rd Qu.:0.7843
##  Max.   :0.5630   Max.   :0.8600
```

```r
knitr::opts_chunk$set(echo = TRUE)
mat=cor(nbaAllDf)
require(corrplot)
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```r
png(height=1200, width=1500, pointsize=15, file="NBAoverlap.png")
corrplot(mat, method = "color", addCoef.col="grey", order = "AOE")
```

The correlation plot indicates high correlation between elite player variables(ELITE.DEF, ELITE.OFF) and the team's goal difference along with how well they do in all the Conference Finals and Finals some of these variables must be dropped to avoid multicollinearity in the final model. 3 pints shots, 2 point shots, and free throws are don't correlate much with any other variables in the dataset.

```r
model=polr(ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.   , data = trainDf,   Hess = TRUE)

#summary(model)
drop1(model, test="Chisq")
```

```
## Single term deletions
##
## Model:
## ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.
##           Df    AIC    LRT  Pr(>Chi)
## <none>        117.72
## ELITE.DEF  1 151.70 35.984 1.990e-09 ***
## ELITE.OFF  1 123.25  7.539  0.006038 **
## X3P.       1 133.40 17.686 2.605e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
stepAIC(model,direction = "backward" )
```

```
## Start:  AIC=117.71
## ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.
##
##             Df    AIC
## <none>          117.72
## - ELITE.OFF  1 123.25
## - X3P.       1 133.40
## - ELITE.DEF  1 151.70
```

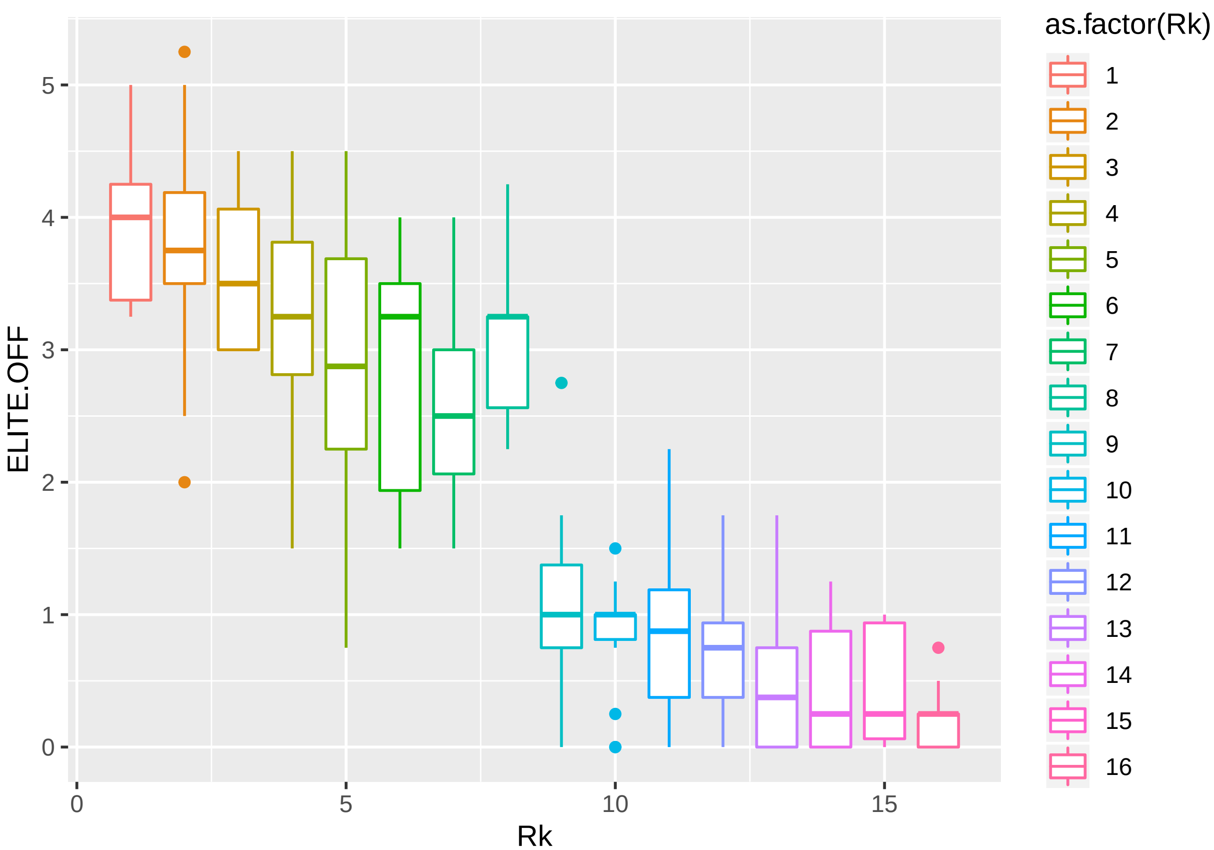
```
## Call:
## polr(formula = ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.,
##     data = trainDf, Hess = TRUE)
##
## Coefficients:
## ELITE.DEF ELITE.OFF      X3P.
##  1.511346  0.830106 42.964031
```

```
##
## Intercepts:
##      0|1      1|2      2|3
## 20.41771 23.95619 26.14537
##
## Residual Deviance: 105.715
## AIC: 117.715
```

The dataset used to create this model contains a total of 15 NBA seasons from year 2003 to 2018. This excludes 2012 as the playoffs were cancelled that year. Each season, 16 teams are selected to the playoffs based on performance, thus we have a dataset of size 240. Logistic regression was used to model the dataset to understand which variables had the most impact on the playoff results. We gave each team a variable between 3 and 0 based on their performance in the playoffs; a variable defined as "Champ". The teams that won the playoffs received 3, and a value of 2 was given to the teams that won the conference finals. A value of 1 was assigned to the teams that won that conference semifinals. All the other teams received a value of 0. The training dataset was created using half of the original dataset which had been randomly selected. After performing backwards elimination and using our intuition several variables were considered insignificant and was dropped. We dropped variables such as final win-loss percentage (F.W.L.) and conference semi-final win-loss percentage (CSF.W.L.), as they offered no significant understanding as to what factors led the teams to win. Hence we chose the model: polr(formula = ordered(Champ) ~ ELITE.DEF + ELITE.OFF + X3P.. From this model we can see that both elite defense and offensive players are very influential to a teams chances of winning the NBA playoffs. Also 2 point shots have make the difference when it comes to 2 points shots does not indicate the winner or loser.

PLOTS

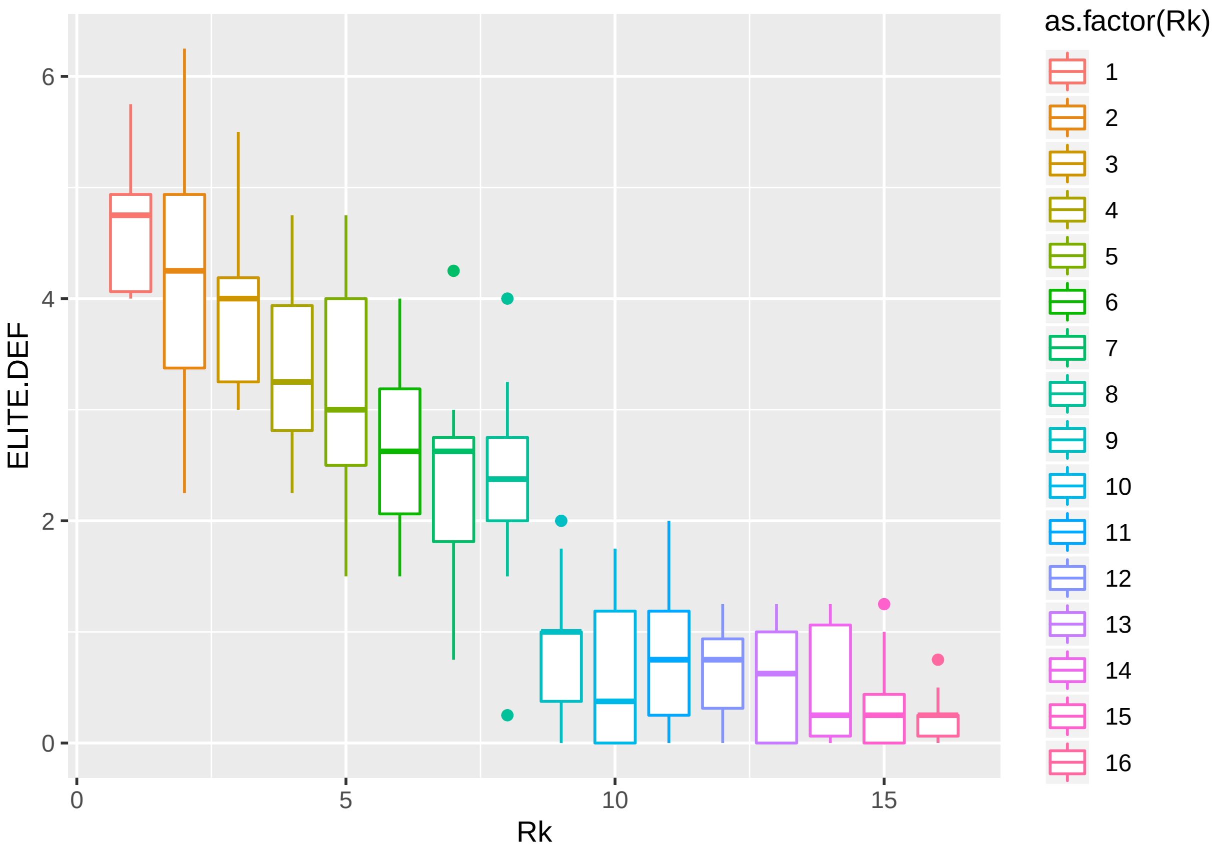
```
ggplot(nbaAllDf,aes(x=Rk,y=ELITE.OFF,colour=as.factor(Rk)))+geom_boxplot()
```



In

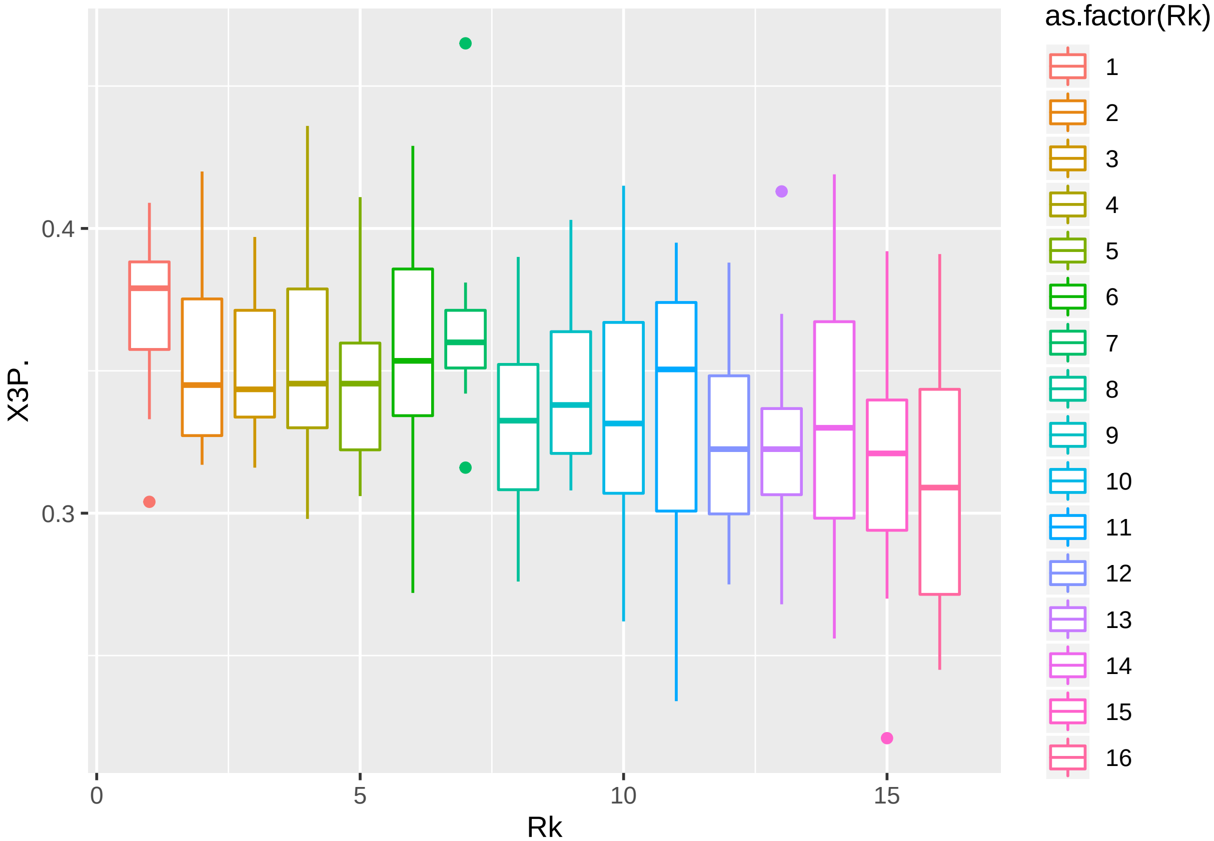
this plot the x-axis represent the rank of the team. In rank 1 represents the team that did the best in the playoffs,and 16 represents the team that did the worst. The y-axis elite offensive score which is calculated using total number to elite forward players for the regular season after the trade deadline and total elite players during the playoffs. The teams that made it past the conference semifinals have significantly more elite players then the teams that only made it till the conference first round. There are outliers presents in Ranks 7,8,9, 15 and 16.

```
ggplot(nbaAllDf,aes(x=Rk,y=ELITE.DEF,colour=as.factor(Rk)))+geom_boxplot()
```



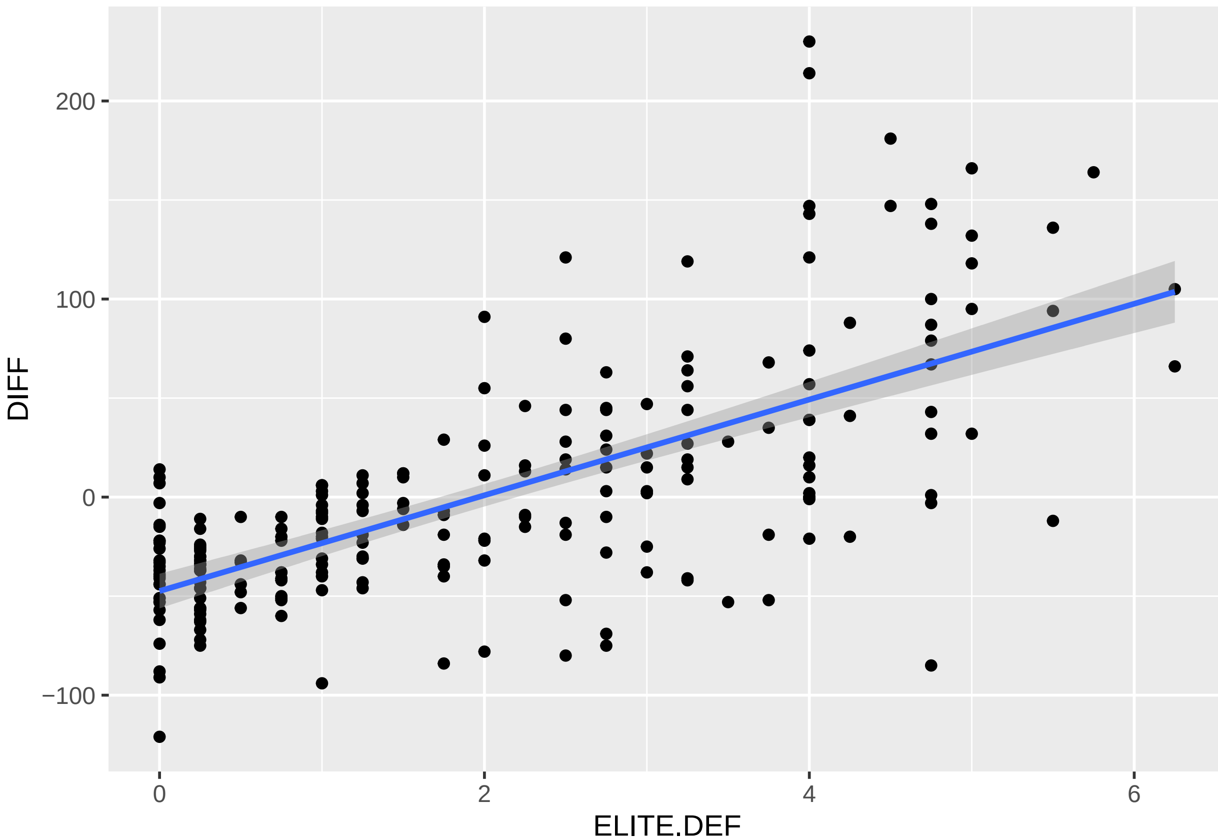
In this plot x-axis represents rank of the team as explained previously. The y-axis represents elite defensive score which is calculated using total number to elite defensive players for the regular season after the trade deadline and total elite defensive players during the playoffs. There are outliers present for ranks 6,7,8,15 and 16. Overall we see the trend of the teams that make it further in the pay off having an increasing number of elite players.

```
ggplot(nbaAllDf,aes(x=Rk,y=X3P.,colour=as.factor(Rk)))+geom_boxplot()
```

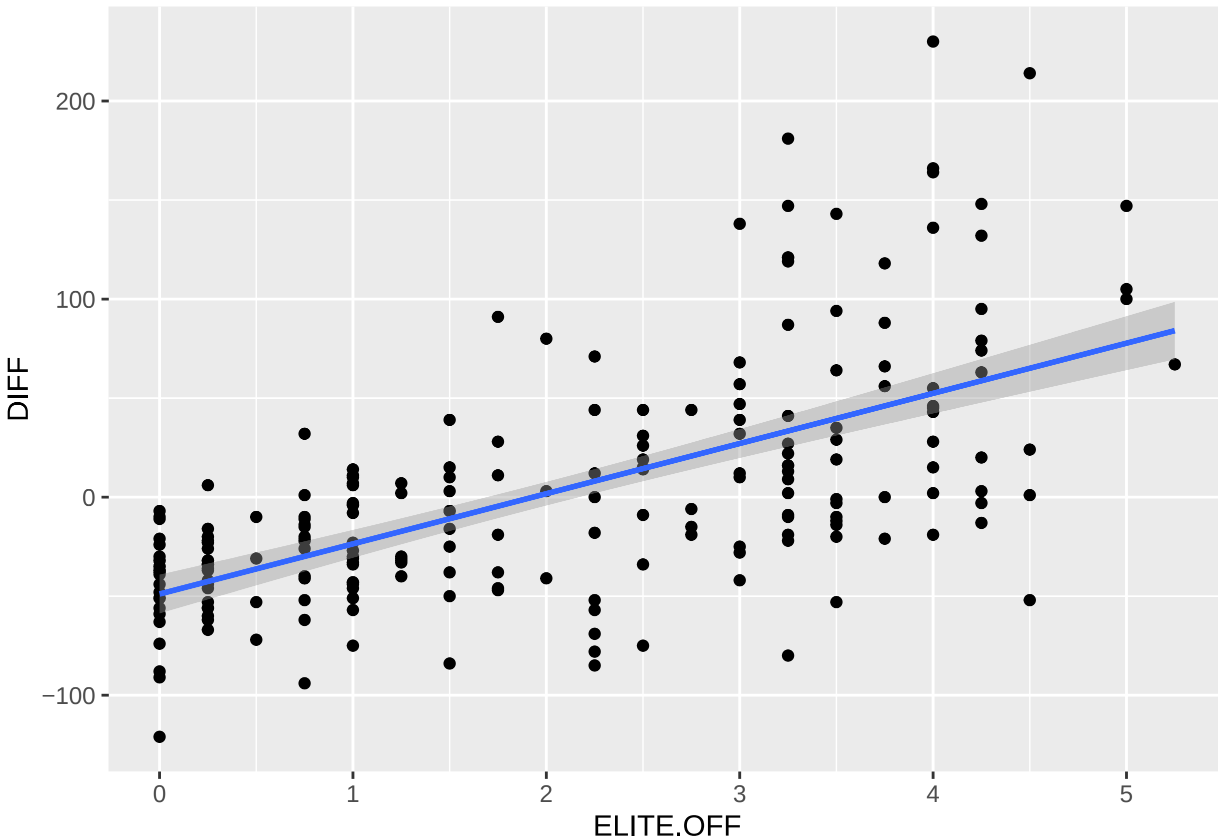


In this plot x-axis represents rank of the team as explained previously. Y-axis represent the percentage of 3 point shots out of all the points scored. There are a few outliers present for the 1st, 7th, 13th, and 14th ranks however the overall trend shows that the teams that rank higher score more 2 point shots.

```
ggplot(nbaAllDf,aes(x=ELITE.DEF,y=DIFF))+geom_point()+geom_smooth(method="lm")
```




```
ggplot(nbaAllDf,aes(x=ELITE.OFF,y=DIFF))+geom_point()+geom_smooth(method="lm")
```




```r
detach(trainDf)
```

These two plots indicate that the major point difference between teams positively correlates with the elite defensive and offensive players that are present on the team. However, we can see from these graphs that increase of elite defensive players tend to create more of a goal difference then increase in elite offensive player.