## Introduction

### *Scientific question*

Can we model an NBA team's season win rate based on summary statistics from the season (e.g. total points, total rebounds, total shots on goal, etc.) as well as additional added variables?

### *Data*

We used data from the following NBA seasons from Basketball Reference to build our model:

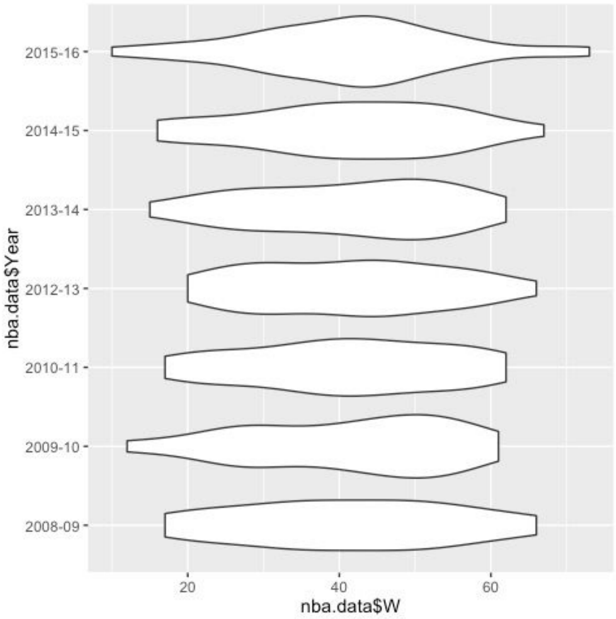
2008-09
2009-10
2010-11
2012-13
2013-14
2014-15
2015-16

We did not use the 2011-12 season because it had fewer games due to a lockout. The lockout was because the Owners and the Players were unable to come to an agreement until Christmas Day 2011. The season started soon after then, when normally the season starts in late October. We did not use seasons before 2008 because of rule changes before 2007 that would affect the season and how the games were played. (There were still rule *clarifications* within the seasons we included but no major rule *changes*). See Appendix for variables measured and variable codes

## Exploratory Data Analysis

The NBA Data we used has 210 observations on 32 variables. We were given 29 of them from Basketball Reference and created 3 of our own variables. One was ABR, which is an abbreviation of the teams to help with better graphics and analysis. The other two were Conference and after Lockout, which will be discussed in the Model Building section.

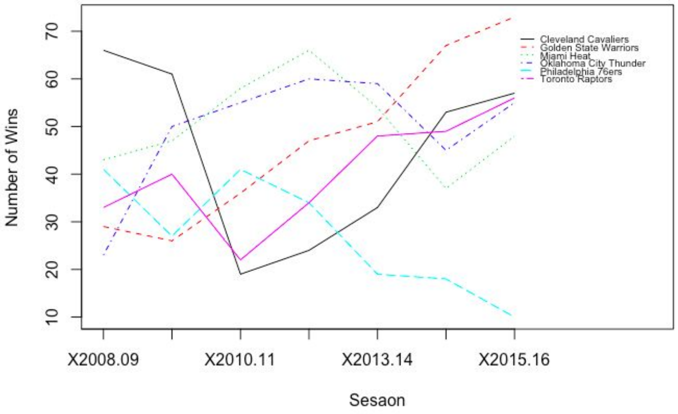
To see how the number of wins has changed from season to season, we created a violin plot. The figure is below:



This plot shows how the number of wins have been distributed throughout each season. In recent years, the distribution has become more of a normal distribution. This is a difference from the 2008-2009 season to the 2013-14 season, which all appear to be more of a uniform distribution. This shows that teams are becoming both better (more wins) and worse (less wins). There could be an argument to say that last season, 2015-16, was a fluke. This is because the Golden State Warriors had the all time wins record in the regular season with 73 wins, and the Philadelphia 76ers had an extremely horrible season with just 10 wins. We will not know if that season is the new trend or is just a fluke without more seasons worth of data from after 2015-16 season. This change in the distribution in the number of wins could affect our model, but should not affect it too much due to the recent change only being one season in our dataset out of 7 seasons.

We calculated a correlation matrix next to see the relationships between our dependent variable, number of wins, and other variables in the data set. This was done to help determine if there were any variables that stand out as highly correlated with number of Wins and should be used in our model. The full correlation matrix can be found in the Appendix under Figure 2.

The three that were the had the highest correlation with number of  wins were Field Goal Percentage, 2 Point Percentage, and 3 Point Percentage. These were 0.65, 0.665, 0.528 respectively. This shows that there is a high positive correlation between each of these variables against number of wins, but they are not perfectly correlated (equaled to 1). These variables show they are the source of some variability in the number of wins per season across all the teams. These correlations are intuitive. They show that if a team takes more efficient shots at a higher conversion rate, then the team will have more wins across the season.



The figure above shows some interesting trends that occurred over the 7 seasons that we observed. A plot of the trends for all 30 teams can be found in the Appendix under figure 13. The first trend that is worth noticing is the massive drop in wins the Cleveland Cavaliers experienced from the 2009-10 season to the 2010-11 season. The Toronto Raptors also experience a somewhat significant drop although it is about half the magnitude of the Cavs drop off. These drop offs were caused by the forming of the “big 3” in Miami. Lebron James left the Cavaliers and Chris Bosh left the Raptors to join Dwayne Wade in Miami during the 2010 offseason. Lebron would return to Cleveland for the 2014-15 season resulting in a significant upswing in the number of wins the Cavs had that season and a subsequent decrease in wins for the Heat. The Oklahoma City Thunder had a steep increase in wins in the 2009-10 season thanks to the addition of key rookies James

Harden and Serge Ibaka to complement Kevin Durant and Russell Westbrook. The Golden State Warriors improved their win totals season to season by drafting allstars Steph Curry in 2009, Klay Thompson in 2011, and Draymond Green in 2012. Finally the Philadelphia 76ers were plagued by injuries to 1st round draft picks, bust of picks, and tanking resulting in the second worst record of all time since the season was expanded to 82 games.

**Model Building**

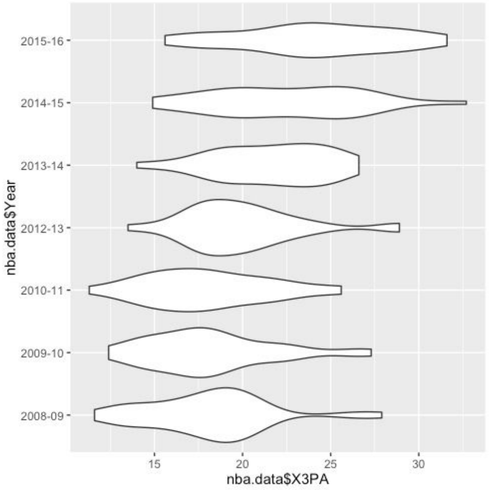
We utilized both the Stepwise Variable Selection and Principal Component Analysis to see if we could narrow down variables in the data set to help predict wins in a season. Using stepwise variable selection in R, we reduced the number of variables from 26 to 13. The 13 variables were MP, FG, 3PA, 2P, 2PA, FT. , DRB, TRB, AST, STL, TOV, CONF, and after Lockout. We will refer to this model as our “Stepwise Model.” Also, descriptions for the variables in our Stepwise Model are referenced in the Appendix under Figure 1. Essentially, stepwise variable selection provided us with a preliminary model that established a useful framework for further model building. The summary output and analysis of deviance table of the Stepwise Model is located in the Appendix under Figures 3 and 4, respectively .

Next, we wanted to see if Principal Component Analysis (PCA) would provide us with any different insights into variable reduction. Principal Component Analysis deals with only numeric variables, so the variables “Conference,” “Year,” and “after Lockout” were not included in the PCA. A full description of the R output from running PCA can be found in the Appendix under Figure 5 . After ensuring the PCA ran properly, the column means and variable means from the PCA were the same, we looked closely at each Principal Component. The first principal component output described 24% of the variation in the data. The first principal component will describe the most variation in the data, all others goes down from the 24%. After taking this into consideration, we decided to prefer the Stepwise Model, rather than the PCA output.

We wanted to build a poisson model, which is used to measure the count of wins across all teams in a season. Using the count of wins, we choose to measure the rate of wins using Games Played as the offset to help with seasons that were either cut short due to a lockout or for the future if the number of games changes throughout the league. This helped to have a more uniform measure of the number of wins across seasons. The Stepwise Model only measured count of wins without an offset, this is taken into account when deciding on a final model.

We wanted to simplify and reduce the number of variables in our Stepwise Model even further. To do this, we tried to see if there were any interaction terms that could be significant. Rather than immediately adding interaction terms to the model, we plotted every variable that was significant against number of wins with splitting on Year and Team. Both Year and Team were used at factor variables, even though Year could be used as a continuous variable. Due to the nature of the data set, all these plots were too hard to locate any interaction terms, because "Year" had 8 different levels and "Team" had 30 different levels. By taking this into consideration, we added a new variable "Conference" (coded CONF), which is associated with whether the team is in the Eastern Conference (E) or the Western Conference (W). We also split on year after the shortened season with the NBA Lockout (2011-2012). We decided to label the NBA seasons after the lockout as "AL" for "After Lockout", and label the seasons before the lockout as "BL" for "Before Lockout". By creating these two new variables, the graphs were easier to determine if there were any significant interaction terms within our data set. We considered this season as our cutoff season due to the figure below. This figure shows the distribution of 3 point attempts across each season.

There is an overall trend over the recent years to take more 3 pointers. This is shown in the spread and center of each distribution across each season. From seasons 2008-09 to 2010-11, the mean of the 3 point attempts decreases. However, immediately after the 2012-13 season, the mean of the 3 point attempts increases significantly and continues to



increase for each following season. We used the 2011-12 season as the cutoff point. This season was the shortened season due to the NBA Lockout, which is why we code the variable as either “After Lockout” or “Before Lockout.” We thought this would lead to a model with an interaction term between 3 point attempts and the “after Lockout” variable.

The figure(Fig 6) shows the number of wins versus 3 point attempts per game, while splitting on “After Lockout” and “Before Lockout.” This figure shows there is a significant difference in the mean of “BL” and the mean of “AL”. Due to this observation, we added the interaction term of 3 point attempts and “After Lockout” to our model.

When creating plots for every variable in the Stepwise Model against wins while splitting on Conference, the plot with “Field Goals” had the most potential of an interaction term with Conference. The figure can be seen in the appendix (Fig 7). Although both the plots show a potential interaction term between 3 point attempts and After Lockout and Conference and Field Goal, the interaction terms proved to be not significant in our model and were ultimately removed from our final model. Further discussion of this is in the next section, Model Selection.

## Model Selection

Similar to our Stepwise Model, we created a derived model that included the terms from our Stepwise Model, along with the interaction term between 3 point attempts and “After Lockout,” as well as the other potential interaction term between field goals and the conference each team is in. We will call this the “Interaction Model.” The full summary output and analysis of deviance table for the Interaction Model is in the Appendix under Figures 8 and 9 respectively.

To refresh, the main effects in both the Stepwise Model and the Interaction Model were Minutes Played, Field Goals, 3 Point Attempts, 2 Points Made, 2 Point Attempts, Free Throw Percentage, Defensive Rebounds, Total Rebounds, Assists, Steals, Turnovers, Conference, and after Lockout. Intuitively thinking about each variable individually, we removed Minutes Played. This variable is the minutes every team plays in each season, which should be the same across each team and does not add additional variation of the numbers of wins. We took out Field Goals because this variable adds both the 3 point and 2 point baskets made, but does not distinguish from the two. We took out the variable and added both 3 pointers and 2 pointers into the model to help distinguish from the two. Knowing the game of basketball, Free Throw Percentage does not help determine wins. Free Throws are only worth 1 point, however, they are arbitrary based off of fouls. There is too much variability in the amount of Free Throws per game to keep it in the model. We took out Total Rebounds and replaced it with Offensive Rebounds. Having all three, Defensive Rebounds, Offensive Rebounds, and Total Rebounds, in the model causes linear dependency among variables. Having the split of Offensive versus Defensive shows a better picture of the team, and makes the model more interpretable.

Both the interaction terms between 3 Point Attempts and after Lockout and Field Goals and Conference proved to not be significant from the Interaction Model and were not kept in the final model.

By switching around variables to make a more interpretable model, taking some variables out that were redundant or not necessary, and adding more intuitive variables, we created our final model.

**Final Model**

Let $Y_i$ denote the number of wins for a single team in a single season where i = 1, …, 210
We assume that $\{Y_i : i = 1, \ldots, 210\}$ is a set of independent Poisson random variables.
We will model the mean number of wins by team *i*, $\mu_i$, per season (82 games), $P_i$.

The win rate for team *i* is

$$rate = R_i = \frac{\mu_i}{P_i}\ ,\ rate \geq 0$$

$$log(\frac{\mu_i}{P_i}) = \alpha + \sum_{j=1}^{6} \beta_j x_{ij} + \gamma(conf)_i + \delta(lock)_i$$

Where

$\boldsymbol{x}_{\mathrm{i}} = (x_{\mathrm{i},1}, \dots , x_{i,6})^{\mathrm{T}}$ are the 6 covariates for observation *i*

*j* = (1 3-pointers attempted, 2 2-pointers attempted, 3 offensive rebounds, 4 defensive rebounds, 4 assists, 5 steals, 6 turnovers)

*(conf)*$_i$ = $\{0 = Eastern\ Conference,\ 1 = Western\ Conference\}$

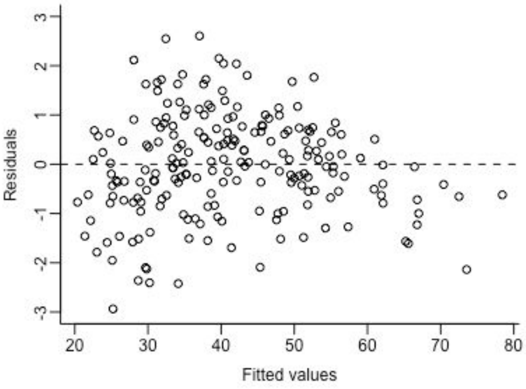
*(lock)*$_i$ = $\{0 = after\ lockout,\ 1 = before\ lockout\}$

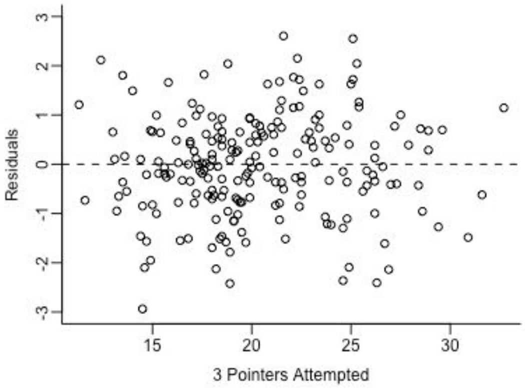
$\boldsymbol{\beta}$ is the coefficient vector estimated by $\boldsymbol{\beta}$

$\gamma$ and $\delta$ are coefficients estimated by $\widehat{\gamma}$ and $\widehat{\delta}$ respectively

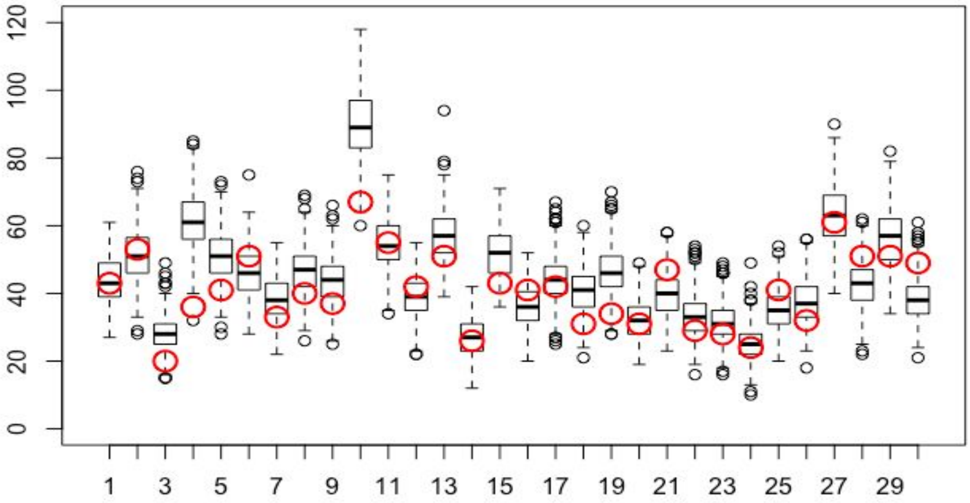
Although a team winning zero games in a season is in the domain of our model, there has never been a team in the NBA that has not won a game. This could happen, but it is extremely unlikely. The summary output and analysis of deviance table is in the Appendix under Figures 10 and 11 respectively. This model will be known as the “Final Model”.

## Diagnostics





Above are the plots for the fitted values versus residuals and 3 pointers versus residuals. As we can see the values are centered around zero and have a constant spread. The residual plots for all nine variables that we included in our data can be found in the Appendix under Figure 12. All of the plots for the continuous variables are centered around zero with a constant spread. The two factor variables that we include have residuals centered around zero.



The above graph is the cross-validation of our model with the NBA results from this season. For a majority of the league our model seem to do a good job of predicting the actual win totals. The one outlier however are the Golden State Warriors. Our model predicts that on average Golden State should win more than 82 games. This is impossible since the season is 82 games. One weakness of our model is that it is allowed to output values that are above 82 wins.

### Interpretation of Our Final Statistical Model

*Interpretation of Coefficients:*

- Intercept: 0.861966 (not very meaningful)
    - A team with zero field goal attempts (2 or 3 points), no rebounds (offensive or defensive), no assists, no steals, no turnovers, in the Eastern Conference, and from a season after the lockout, would have an average win rate of

$e^{0.861966} \approx 2.3678$ wins per season. This is a practically impossible scenario and therefore not a meaningful or useful interpretation.

- X3PA: -0.078272
  - We estimate the win rate to increase by a factor of $e^{-0.078272} \approx 0.9247$ units for a one additional 3-point field goal attempt.
  - 95% Confidence Interval: [0.9142728 , 0.9352722]
- X2PA: -0.095197
  - We estimate the win rate to increase by a factor of $e^{-0.095197} \approx 0.9092$ units for a one additional 2-point field goal attempt.
  - 95% Confidence Interval: [0.8990682, 0.9194225]
- ORB: 0.119155
  - We estimate the win rate to increase by a factor of $e^{0.119155} \approx 1.1265$ units for one additional offensive rebound.
  - 95% Confidence Interval: [1.099428, 1.15433]
- DRB: 0.133814
  - We estimate the win rate to increase by a factor of $e^{0.133814} \approx 1.1432$ units for one additional defensive rebound.
  - 95% Confidence Interval: [1.12396, 1.162729]
- AST: 0.068501
  - We estimate the win rate to increase by a factor of $e^{0.068501} \approx 1.0709$ units for one additional assist.
  - 95% Confidence Interval: [1.055171, 1.086867]
- STL: 0.152732
  - We estimate the win rate to increase by a factor of $e^{0.152732} \approx 1.1650$ units for one additional steal.
  - 95% Confidence Interval: [1.130522, 1.200556]
- TOV: -0.161877
  - We estimate the win rate to increase by a factor of $e^{-0.161877} \approx 0.8505$ units for one additional turnover.

    - 95% Confidence Interval: [0.8320226, 0.8694814]
- CONFW: 0.031710
    - We estimate the win rate to increase by a factor of $e^{0.031710} \approx 0.8505$ units for a team in the Western Conference compared to the Eastern Conference.
    - 95% Confidence Interval: [0.9864405, 1.08012]
- afterLockout: 0.151033
    - We estimate the win rate to increase by a factor of $e^{0.151033} \approx 1.1630$ for a team playing before the lockout compared to a team playing after the lockout.
    - 95% Confidence Interval: [1.10074, 1.228856]

*Answering the Scientific Question*

We were able to successfully create a model to predict win rate for NBA teams. The variables we found to be useful in this model were field goals attempted (2 or 3 points), rebounds (offensive and defensive), assists, steals, turnovers, conference, and before or after the NBA lockout. However, this model does have several limitations, which we discuss below.

**Discussion of Our Results**

*Purpose.* This model is useful in predicting the win rate of NBA teams, and identifying corresponding variables to an increased win rate.

*Implications.* There are some interesting implications from this model. One of the more obvious implications come from looking at the coefficients for 3-point attempts and 2-point attempts, which logically suggest that 3-point attempts correspond to a higher win rate than 2-point attempts. The coefficient for our factor variable CONF (conference) suggests that the Western Conference has a slightly higher win rate than Eastern conference. This may be due to an underlying confounding variable that we did not measure, which would affect Western Conference teams differently than Eastern Conference teams. Another interesting implication is that defensive rebounds correspond to a higher win rate increase than offensive rebounds. This makes sense because a defensive rebound changes the possession of the ball, giving the team that got the rebound a better chance of scoring. *Limitations*. One major limitation of this model is that it models variables that correspond to win rate, but does not prove that these variables cause the win rate to change. Another limitation is that the data we used to build our model consisted of 210 "teams," however even though teams are technically different from season to season, they are closely related to the preceding and following seasons since players and coaches are not randomized every year. Basketball, like sports in general, also has many unknown variables that we were not able to measure such as referees, money,  and player health. Finally, statistical analysis is very popular in basketball right now, and strategies change based on that analysis. Teams can change how they play based on perceived or real trends, which can also affect win rate.

**Ideas for Further Analysis**

Expanding on what we have started, creating a new model that directly correlates to how a team plays would be different than what we have created. Our model predicts the number of wins/win rate, a coach can't derive particular value to what plays the team should run, how it should play against certain teams, and who the coach should play or trade for. A larger data set, with specific player statistics could help build this potential model. This new model could help affect gameplay instead of predicting how a team does at the end of the year.

Diving deeper into the impact of players could be interesting for future analysis. Teams such as the Chicago Bulls from 2015-16 season and the Chicago Bulls from this past season 2016-17 are technically different teams but are very closely related. The Bulls acquired Dwayne Wade in the Summer between these two seasons. By tracking who the team trades for and gets rid of could lead interesting correlations between how the team does and who on the team is different to see the different impact of players and how well a team does in a season.

## Appendix

*Figure 1*

Variable codes:

| | |
|---|---|
| YearEnd | Year the season ends |
| Year | Years of the seaon |
| Rk | Rank at the end of the season |
| UniqueTeam | Unique Team name by season |
| Team | Team name |
| ABR | Abbreviation |
| CONF | Conference (E for East, W for West) |
| G | Games Played in the Season (Always 82) |
| W | Wins |
| MP | Minutes Played |
| FG | Field Goals Made |
| FGA | Field Goals Attempted |
| FG% | Field Goal Percentage |
| 3P | 3 Point Field Goals Made |
| 3PA | 3 Point Field Goals Attempted |
| 3P% | 3 Point Field Goals Percentage |
| 2P | 2 Point Field Goals Made |
| 2PA | 2 Point Field Goals Attempted |
| 2P% | 2 Point Field Goals Percentage |
| FT | Free Throws Made |
| FTA | Free Throws Attempted |
| FT% | Free Throws Percentage |
| ORB | Offensive Rebounds |
| DRB | Defensive Rebounds |

| | |
|---|---|
| TRB | Total Rebounds |
| AST | Assists |
| STL | Steals |
| BLK | Blocks |
| TOV | Turnovers |
| PF | Personal Fouls |
| PTS | Points |
| afterLockout | After the 2011-2012 (the season with the NBA lockout) season is “AL”, Before the 2011-2012 (the season with the NBA lockout) season is “BL” |

*Figure 2*

```
R code: library(corrplot)
        corrplot.mixed(cor(nba.data[,c(9:31)]))
```

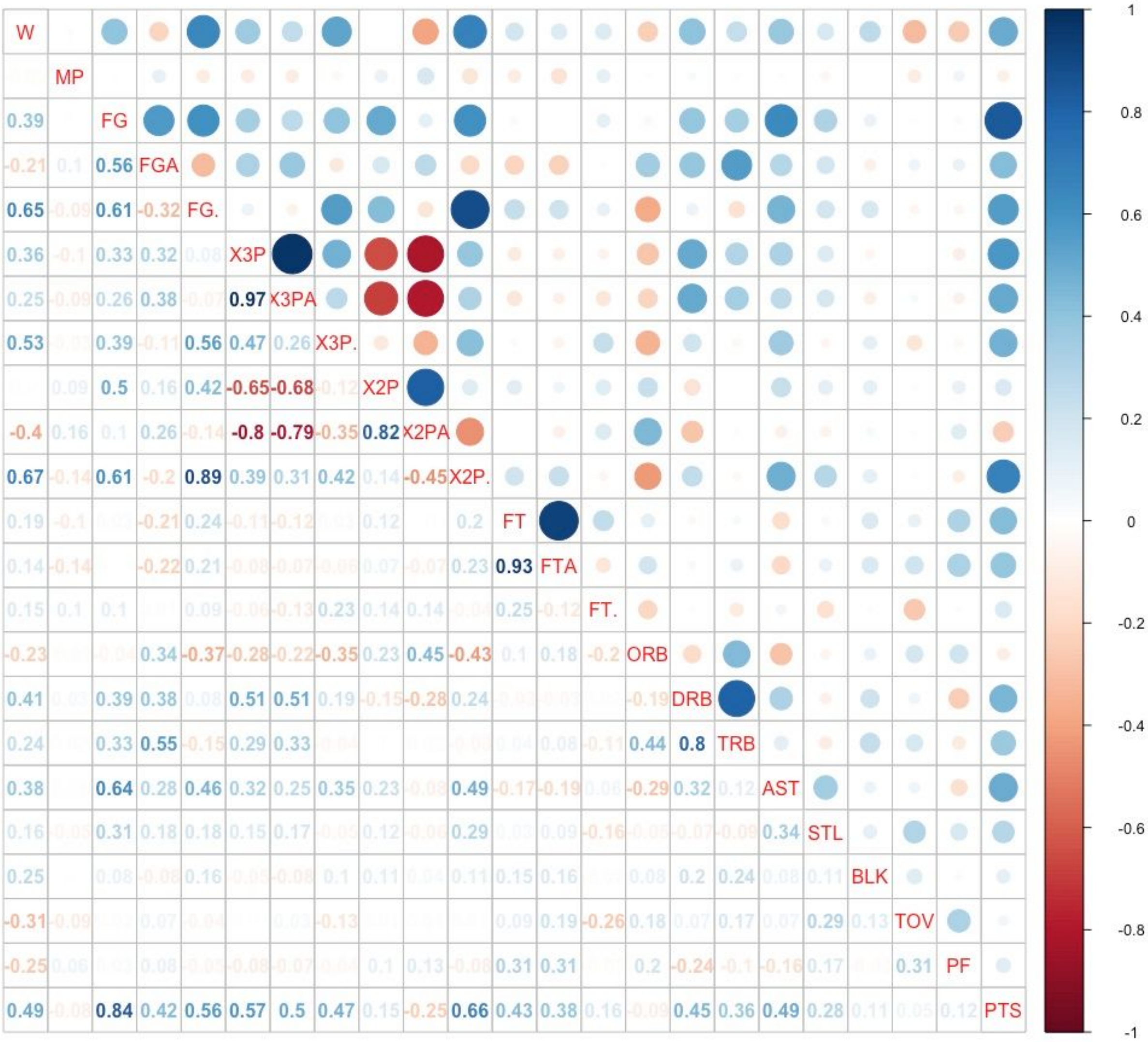


*Figure 3*

Stepwise Model:

```
R code: stepwise.model = glm(formula = nba.data$W ~ MP + FG +
X3PA + X2P + X2PA + FT. + DRB + TRB + AST + STL + TOV + CONF +
afterLockout, family = poisson, data = nba.data)
```

Summary of Stepwise Model:

```
R code: summary(stepwise.model)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.33293  -0.57043   0.01399   0.49381   2.11048
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.107680   3.062288  -1.668  0.09533 .
MP              0.036398   0.012736   2.858  0.00426 **
FG              0.163392   0.032636   5.006 5.54e-07 ***
X3PA           -0.125753   0.012814  -9.814  < 2e-16 ***
X2P            -0.065100   0.036202  -1.798  0.07214 .
X2PA           -0.123983   0.007507 -16.516  < 2e-16 ***
FT.             1.759832   0.418937   4.201 2.66e-05 ***
DRB            -0.034951   0.014456  -2.418  0.01561 *
TRB             0.147653   0.013069  11.298  < 2e-16 ***
AST             0.029080   0.009039   3.217  0.00129 **
STL             0.135819   0.016248   8.359  < 2e-16 ***
TOV            -0.145060   0.011539 -12.571  < 2e-16 ***
CONFW          -0.040822   0.025114  -1.626  0.10406
afterLockoutBL  0.056020   0.030311   1.848  0.06458 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 942.58  on 209  degrees of freedom
Residual deviance: 132.68  on 196  degrees of freedom
AIC: 1314.5

Number of Fisher Scoring iterations: 4
```

*Figure 4*

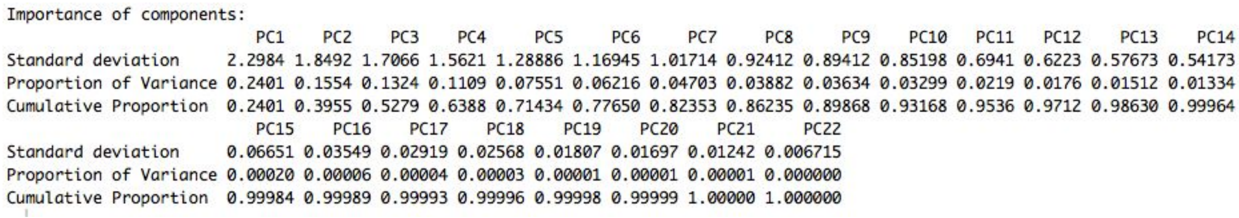
Analysis of Deviance Table of the Stepwise Model:

```
R code: anova(stepwise.model, test=”Chi”)

Model: poisson, link: log

Response: nba.data$W

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           209     942.58
MP            1    0.222       208     942.36 0.6374495
FG            1  133.702       207     808.65 < 2.2e-16 ***
X3PA          1   19.353       206     789.30 1.087e-05 ***
X2P           1  120.776       205     668.53 < 2.2e-16 ***
X2PA          1  194.437       204     474.09 < 2.2e-16 ***
FT.           1   13.753       203     460.34 0.0002085 ***
DRB           1   68.219       202     392.12 < 2.2e-16 ***
TRB           1   60.707       201     331.41 6.623e-15 ***
AST           1    8.663       200     322.75 0.0032472 **
STL           1   20.567       199     302.18 5.758e-06 ***
TOV           1  163.310       198     138.87 < 2.2e-16 ***
CONF          1    2.771       197     136.10 0.0959992 .
afterLockout  1    3.418       196     132.68 0.0644811 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 5*

Summary of PCA:

```
R code: nba.pca <- prcomp(nba.data[,c(10:31)], scale=TRUE)
        summary(nba.pca)
```



```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10   PC11   PC12    PC13    PC14
Standard deviation     2.2984 1.8492 1.7066 1.5621 1.28886 1.16945 1.01714 0.92412 0.89412 0.85198 0.6941 0.6223 0.57673 0.54173
Proportion of Variance 0.2401 0.1554 0.1324 0.1109 0.07551 0.06216 0.04703 0.03882 0.03634 0.03299 0.0219 0.0176 0.01512 0.01334
Cumulative Proportion  0.2401 0.3955 0.5279 0.6388 0.71434 0.77650 0.82353 0.86235 0.89868 0.93168 0.9536 0.9712 0.98630 0.99964
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21     PC22
Standard deviation     0.06651 0.03549 0.02919 0.02568 0.01807 0.01697 0.01242 0.006715
Proportion of Variance 0.00020 0.00006 0.00004 0.00003 0.00001 0.00001 0.00001 0.000000
Cumulative Proportion  0.99984 0.99989 0.99993 0.99996 0.99998 0.99999 1.00000 1.000000
```

*Figure 6*

R code: ggplot(nba.data,

aes(W,X3PA,color=afterLockout,label=afterLockout)) +

geom_text(aes(label=afterLockout),hjust=0,vjust=0)

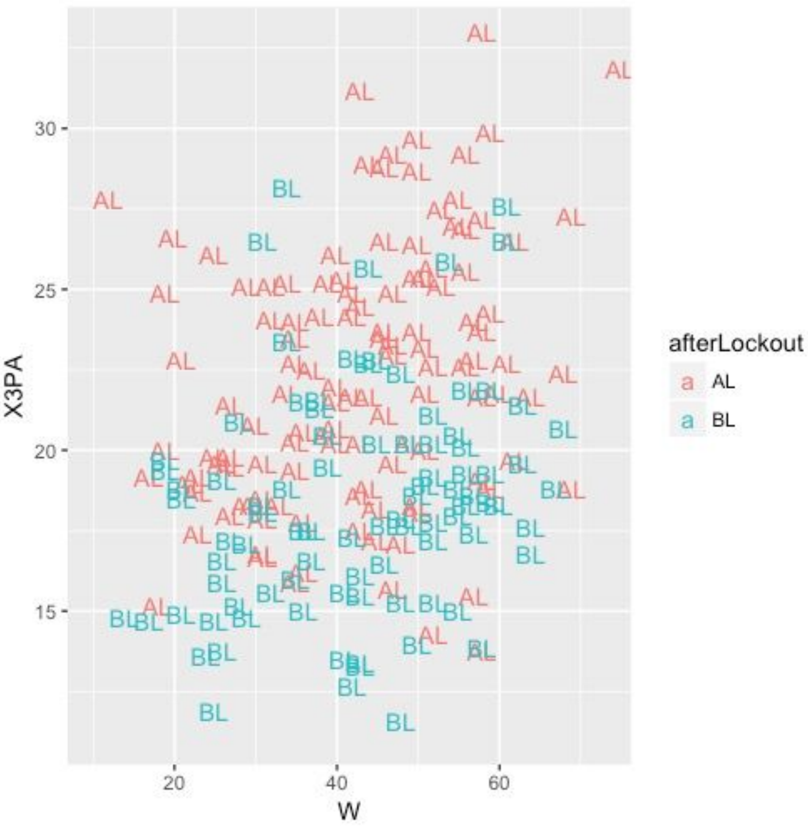


*Figure 7*

R code: ggplot(nba.data, aes(W,FG,color=CONF,label=CONF)) +

geom_text(aes(label=CONF),hjust=0,vjust=0)

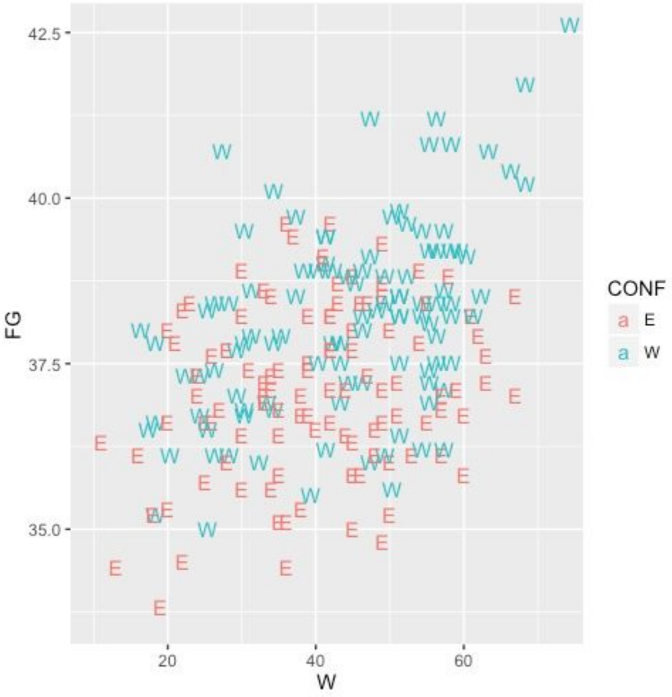


*Figure 8*

Interaction Model:

```
interaction.model = glm(formula = nba.data$W ~ offset(log.games)
+ MP + FG + X3PA + X2P
     + X2PA + FT. + DRB + TRB + AST + STL + TOV + CONF +
afterLockout + FG:CONF +
     X3PA:afterLockout, family = poisson, data = nba.data)
```

Summary Output of the Interaction Model:

```
R code: summary(interaction.model)
Deviance Residuals:
     Min         1Q    Median        3Q       Max
-2.22850  -0.51959   0.03593   0.52050   2.11250
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -9.926829   3.080752  -3.222 0.001272 **
MP                   0.037257   0.012768   2.918 0.003522 **
```

```
FG                  0.170050   0.036750   4.627 3.71e-06 ***
X3PA               -0.128632   0.013180  -9.760  < 2e-16 ***
X2P                -0.066537   0.037228  -1.787 0.073891 .
X2PA               -0.124147   0.007510 -16.531  < 2e-16 ***
FT.                 1.736883   0.420014   4.135 3.54e-05 ***
DRB                -0.035451   0.014470  -2.450 0.014286 *
TRB                 0.148962   0.013156  11.323  < 2e-16 ***
AST                 0.030253   0.009124   3.316 0.000914 ***
STL                 0.138368   0.016398   8.438  < 2e-16 ***
TOV                -0.144673   0.011560 -12.515  < 2e-16 ***
CONFW               0.313818   0.641195   0.489 0.624540
afterLockoutBL     -0.064428   0.121310  -0.531 0.595350
FG:CONFW           -0.009454   0.017080  -0.553 0.579923
X3PA:afterLockoutBL  0.006247   0.006061   1.031 0.302662
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 942.79  on 209  degrees of freedom
Residual deviance: 131.35  on 194  degrees of freedom
AIC: 1317.2
Number of Fisher Scoring iterations: 4
```

*Figure 9*

Analysis of Deviance Table for Interaction Model:

```
R code: anova(interaction.model, test=”Chi”)
Model: poisson, link: log
Response: nba.data$W
Terms added sequentially (first to last)
                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
```

```
NULL                                209     942.79
MP                  1    0.204      208     942.58 0.6511176
FG                  1  133.466      207     809.12 < 2.2e-16 ***
X3PA                1   19.347      206     789.77 1.090e-05 ***
X2P                 1  120.826      205     668.94 < 2.2e-16 ***
X2PA                1  194.593      204     474.35 < 2.2e-16 ***
FT.                 1   13.755      203     460.60 0.0002083 ***
DRB                 1   68.376      202     392.22 < 2.2e-16 ***
TRB                 1   60.692      201     331.53 6.673e-15 ***
AST                 1    8.705      200     322.82 0.0031725 **
STL                 1   20.641      199     302.18 5.541e-06 ***
TOV                 1  163.233      198     138.95 < 2.2e-16 ***
CONF                1    2.796      197     136.15 0.0945181 .
afterLockout        1    3.392      196     132.76 0.0654961 .
FG:CONF             1    0.348      195     132.41 0.5553130
X3PA:afterLockout   1    1.061      194     131.35 0.3030504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Figure 10*

Final Model:

```
final.model = glm(formula = W ~ offset(log.games) + X3PA + X2PA
+ ORB + DRB + AST + STL + TOV + CONF + afterLockout, family =
poisson, data = nba.data)
```

Summary Output of Final Model:

```
summary(final.model)
Deviance Residuals:
    Min         1Q    Median        3Q       Max
```

```
-3.3802  -0.6844  -0.0266   0.6593   2.6055
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.861966   0.444461   1.939   0.0525 .
X3PA           -0.078272   0.005793 -13.512  < 2e-16 ***
X2PA           -0.095197   0.005714 -16.661  < 2e-16 ***
ORB             0.119155   0.012431   9.586  < 2e-16 ***
DRB             0.133814   0.008651  15.467  < 2e-16 ***
AST             0.068501   0.007550   9.073  < 2e-16 ***
STL             0.152732   0.015333   9.961  < 2e-16 ***
TOV            -0.161877   0.011234 -14.409  < 2e-16 ***
CONFW           0.031710   0.023144   1.370   0.1707
afterLockoutBL  0.151033   0.028087   5.377 7.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 942.79  on 209  degrees of freedom
Residual deviance: 231.46  on 200  degrees of freedom
AIC: 1405.3
Number of Fisher Scoring iterations: 4
```

*Figure 11*

Analysis of Deviance Table of the Final Model:

```
R code: anova(final.model, test=”Chi”)
Model: poisson, link: log
Response: W
Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
```

```
NULL                           209     942.79
X3PA          1   52.951       208     889.84 3.420e-13 ***
X2PA          1   94.778       207     795.06 < 2.2e-16 ***
ORB           1    1.110       206     793.95   0.29207
DRB           1  158.693       205     635.26 < 2.2e-16 ***
AST           1  128.698       204     506.56 < 2.2e-16 ***
STL           1   19.882       203     486.68 8.237e-06 ***
TOV           1  222.105       202     264.57 < 2.2e-16 ***
CONF          1    4.178       201     260.39   0.04096 *
afterLockout  1   28.932       200     231.46 7.495e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figure 12*

R code:

```
res <- resid(final.model)
fits <- fitted(final.model)

par(mfrow=c(5,2), cex=0.75, mar=c(4,4,1,1), mgp=c(2,0.5,0),
bty="L")

plot(fits, res,
     xlab="Fitted values", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$X3PA, res,
     xlab="3 Pointers Attempted", ylab="Residuals",
ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$X2PA, res,
     xlab="2 Pointers Attempted", ylab="Residuals",
ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$ORB, res,
     xlab="Offensive Rebounds", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$DRB, res,
     xlab="Defensive Rebounds", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$AST, res,
     xlab="Assists", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$STL, res,
     xlab="Steals", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)

plot(nba.data$TOV, res,
     xlab="Turnovers", ylab="Residuals", ylim=c(-3,3))
abline(h=0, lty=2)
```

```
plot(as.numeric(as.factor(nba.data$CONF)), res, xaxt="n",
     xlab="Conference", ylab="Residuals", ylim=c(-3,3))
mtext(side=1, at=1:2, line=0.5,
sort(as.character(unique(nba.data$CONF))), cex=0.7)
abline(h=0, lty=2)

plot(as.numeric(as.factor(nba.data$afterLockout)), res,
xaxt="n",
     xlab="Before or After Lockout", ylab="Residuals",
ylim=c(-3,3))
mtext(side=1, at=1:2, line=0.5,
sort(as.character(unique(nba.data$afterLockout))), cex=0.7)
abline(h=0, lty=2)
```

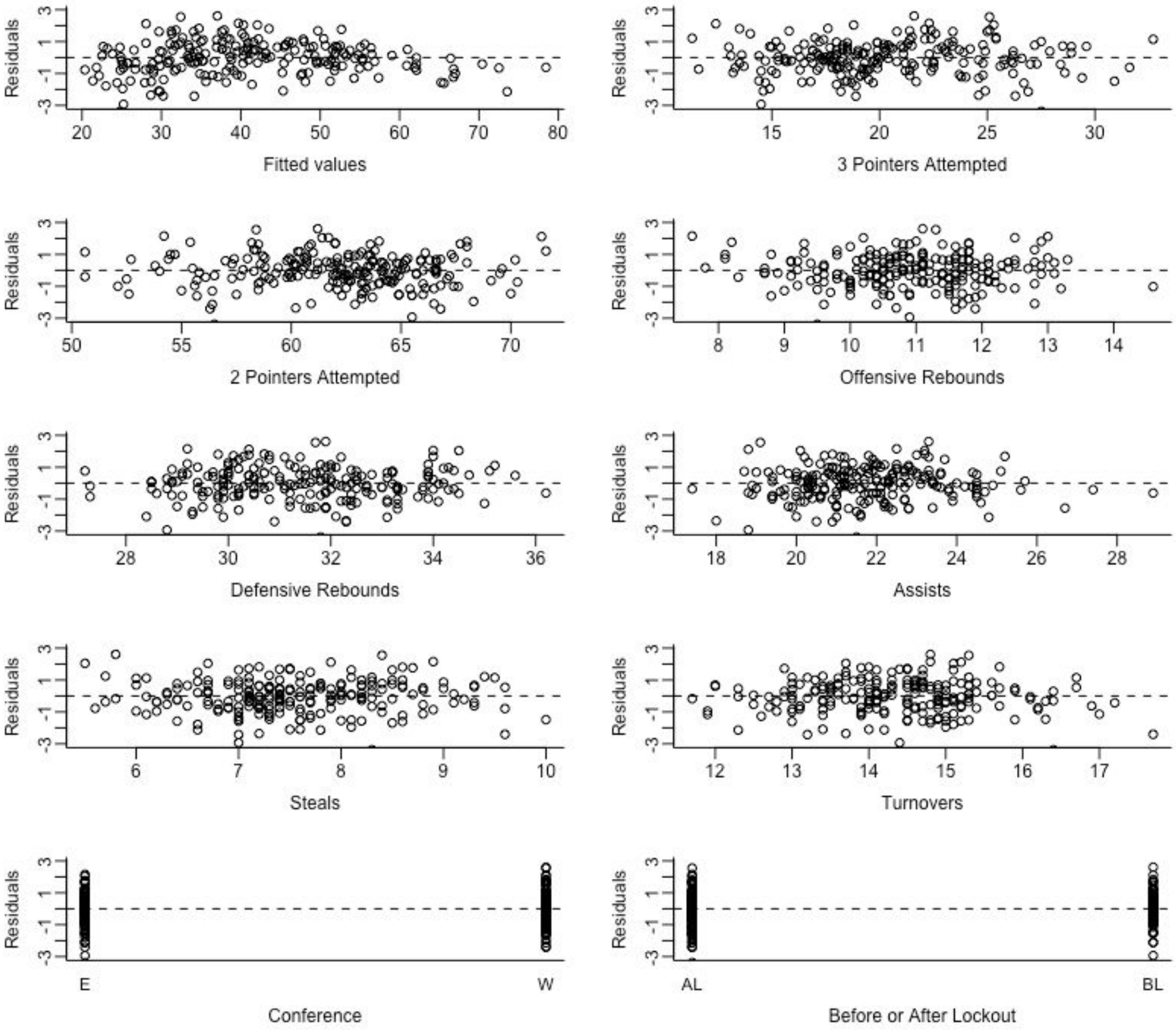
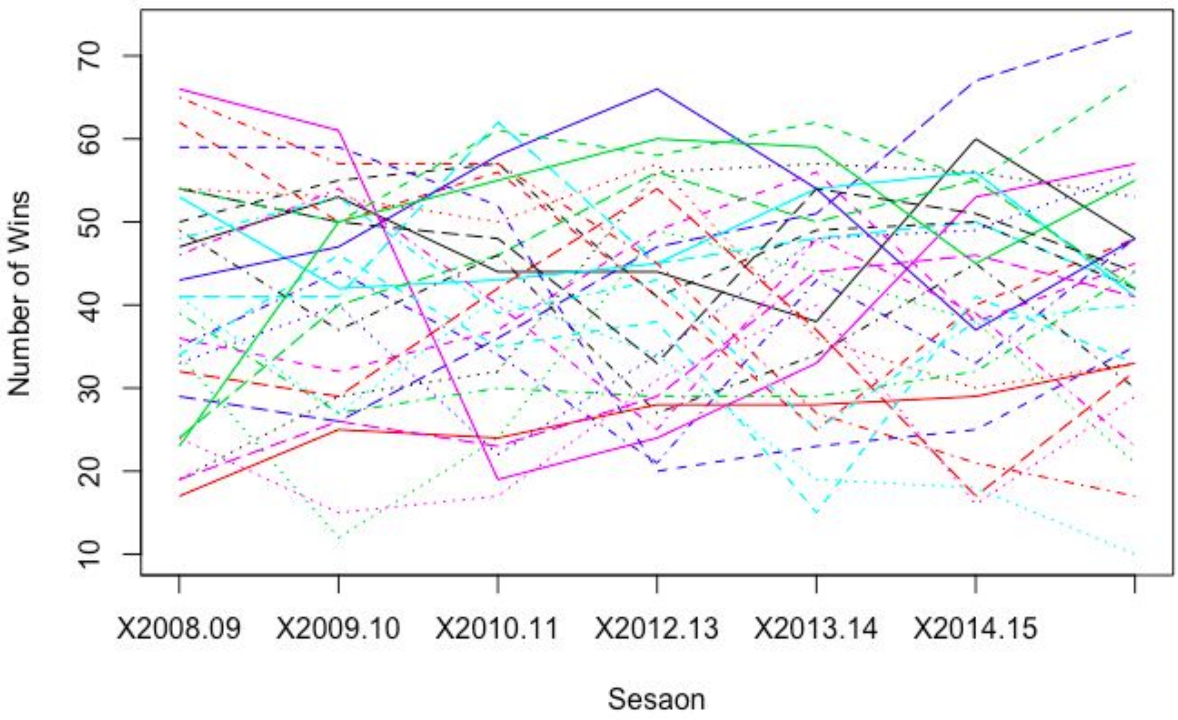


*Figure 13*

Additional R code for Spaghetti Plot, Stepwise Variable Selection, Graph of Win Distributions by Season, Graph of 3PA Distributions by Season, and Cross Validation graph.

```
Spaghetti Plot:
```



```
matplot(t(nba.seasons[,-1]), type="l", xaxt="n", xlab="Sesaon",
ylab="Number of Wins")
axis(side=1, at=0:7, names(nba.seasons))

is <- 1:length(NBA.spag$Team)

matplot(t(NBA.spag[,-1]), type="l", xaxt="n", xlab="Sesaon",
ylab="Number of Wins", xlim = c(1,9))
axis(side=1, at=0:7, names(NBA.spag))
legend(7, 70, NBA.spag$Team, cex=0.6, lty=is, col=is, bty="n")
```

Stepwise Variable Selection:

```
nba.null = glm(nba.data$W ~ 1, family = poisson)
nba.full = glm(nba.data$W ~ ., data = nba.data[,c(2,6:32)],
family = poisson)
library(MASS)
stepAIC(nba.full,direction="both")
```

Graph of Win Distributions by Season:

```
library(ggplot2)
ggplot(nba.data, aes(x=nba.data$Year, y=nba.data$W)) +
geom_violin() + coord_flip()
```

Graph of 3PA Distributions by Season:

```
ggplot(nba.data, aes(x=nba.data$Year, y=nba.data$X3PA)) +
geom_violin() + coord_flip()
```

Cross Validation Graph:

```
preds <- predict(final.model, newdata=NBA.2016.17, se.fit=TRUE)
REPS <- 500
pred.counts <- lapply(1:nrow(NBA.2016.17), function (Team)
  rpois(REPS, exp(rnorm(REPS, preds$fit[Team],
preds$se.fit[Team]))))

par(mfrow=c(1,1), cex=0.75, mar=c(3,3,0.5,0.5), mgp=c(2,0.5,0),
bty="L")

boxplot(pred.counts, names=names(preds$fit), ylim=c(0, 120))
points(NBA.2016.17$W, pch=1, col="red", cex=2, lwd=2)
```