Predicting Champions: Using Season Statistics to Determine Playoff Performance in Sports

Brian Donohue
bdonohue@stevens.edu
Department of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030

Benjamin Rose
brose@stevens.edu
Deptartment of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030

Daniel Bolella
dbolella@stevens.edu
Deptartment of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030

David Fonorow
dfonorrow@stevens.edu
Deptartment of Computer Science
Stevens Institute of Technology
Hoboken, NJ 07030

ABSTRACT

Post-season performance by a sports team is largely dependent on circumstances surrounding the game currently being played. However, it seems that some sports teams continually outperform others on a year by year basis. Therefore circumstantial variables are not the only factors that indicate playoff performance. Considering there is a wealth of data over past sports seasons, it is possible to perform an analysis on the data available in an attempt to determine playoff performance. The outcomes provided by such a study can help coaches, players, and fans alike in determining what team qualities are the most beneficial for playoff performance in a given sport.

1. INTRODUCTION

This study was designed as an application of statistical methods to define a model based on a seemingly chaotic system: professional sports. If successful, the study will yield four separate statistical models for determining the playoff performance of each of the major sports in America: baseball, football, basketball, and hockey.

Attempting to create models that predict sporting event outcomes is by no means a new field of research, and much money can be gained from developing accurate statistical models for predictions. The basis of this study, like previous studies, is that in many sports the same teams make it into the playoffs. The goal of this study is to determine which regular season factors have influenced playoff performance, and develop models using the regular season factors that will help determine an individual team's playoff performance.

By analyzing regular season statistics over the past few years in each respective sport, patterns in the data sets will be analyzed to construct a model for predicting the winner of the playoff season in each sport. On obtaining the model for the sport, the regular season variables for the season in progress will be used in an attempt to determine the champions of this year's season.

2. METHOD

Any statistical analysis must first begin by collecting data relevant to the study. In an attempt to obtain the best model, all possible regular season variables were collected over a fixed period of years to peform the analysis. In addition to the regular season variables, a categorical variable that indicates the team's playoff performance was also collected. This categorical variable will be referred to as *playoff level* from herein. *Playoff level* indicates which round of the playoffs the team reached with 0 signifying the team did not make the playoffs, and a maximum value signifying the team won the championship for that sport (this value varies between the different sports). Because all of the sports are comprised of different rules, different number of games in regular season play, and different datasets it is reasonable to assume the method for obtaining prediction models will vary between sports.

After the data collection was completed, the data was formatted into an Excel spreadsheet, and converted to a comma separated value (CSV) file. The CSV file could then be easily imported into R, a statistical analysis software suite, in order to determine relationships between the regular season variables and the corresponding *playoff level.*

Using the tools available in R, a two-fold analysis was performed for each sport. The first was an analytical study of the relationship between regular season variables and the *playoff level.* The second was comprised of a graphical analysis using boxplots to visualize the relationship between each regular season variable and the corresponding *playoff level*. From this analysis a model for predicting an individual team's playoff performance was obtained, and this model was applied to currently available data to predict the winners for this year.

3. BASEBALL

Baseball is one of the most unique sports played in America because of the amount of games played per season (around 162 games per team). The goal of this study is to be able to determine if it is at all possible for the number of games won in any part (or combination of parts) of a season to help determine who would go on to the playoffs and possibly predict how far they would go. By looking at the past five seasons and breaking them down into four individual quarters, the number of runs scored, and the number of runs allowed at the end of the season we hoped to achieve a suitable model.

3.1 Data

For baseball the *playoff level* categorical variable was defined using the following levels:

0: Did not qualify for post-season
1: Lost in Divisionals
2: Lost in League Championship
3: Lost in the World Series
4: Won the World Series

The regular season variables used were:

- Wins for the first quarter
- Wins for the second quarter
- Wins for the third quarter
- Wins at the end of the season
- Runs scored by the end of the season
- Runs allowed by the end of the season

Each season was divided first by marking the second and last quarter by the All-Star Break date and last regular season game date, respectively. The first quarter dates were generated by finding the middle date between Opening Day and the All-Star Break for the season. The third quarter was done similarly, only using the All-Star Break and the date of the last regular season game for the season.

3.2 Analysis

Linear regressions were applied to the six variables gathered for the analysis. The first regression was with just the four winnings variables. This regression showed there were a few variables that really had a high probability. They were W3 (.855), W1:W4 (.938), and W1:W2:W4 (.947). The second regression was with the RA and RS variables, which did not really produce high probabilities at all. The final regression was with all six variables, which only resulted in W2:RS:RA having the highest probability (.700).

The regressions returned interesting data, but further analysis was conducted. A box plot was made for each quarter of the season against playoff results to see if there were any interesting trends that could be found. In the first quarter (Figure 10), it seemed that only those with 18 or more wins would ever make the playoffs, and that those who had 30 wins were most likely to make it to the World Series. On the chart for the All-Star Break (Figure 11) it was interesting to find that, although most who made it to the playoff fell between 23 and 30 wins for that quarter, there was a huge probability that teams with only 8 wins would make it to the playoffs (and quite far, as a matter of fact). What was also interesting was that no team with 31 or more wins in the second quarter made it to the playoffs. For the third quarter (Figure 12), 18 wins seemed to be the magic number for making it into the playoffs, with few outliers ever making it into the playoffs. Lastly, the chart for the fourth quarter (Figure 13) showed that generally teams who won 18 or more would make the playoffs, with those winning 26 or more games automatically making it in.

3.3 Results

The analysis gave some interesting results as to what trends lead teams into making the playoffs. Based on the regression results, it seems that there are some really strong probabilities of accurately determining who would make the playoffs. However, it would only be helpful once the season was three quarters finished. By that time, it is still uncertain who will win, but it is usually easy to even guess based off intuition which teams are even in the running and which are lost causes.

As for the charts, they provided a little more detail about some more specific trends that could help out earlier in the season. For example, if a team has won 30 games in the first quarter then they will make it to the World Series. If a team has won only 8 games in the second quarter, they will make it to the League Championships.

Despite these trends, there is no concrete, 100% way of accurately predicting champions for a baseball season. Anything could change at any moment (trades, injuries, suspensions). However, the trends may help in coming up with pretty good guesses for future seasons. Considering the 2010 season was already finished by the time the assignment was given, a hypothesis could not be made as to who would win. However, it would be interesting to follow up during the 2011 season.

4. FOOTBALL

The goal of this statistical analysis is to attempt to determine the winner of the SuperBowl in the current NFL season (2010-2011). For Football the model was obtained by evaluating the results and statistics of all 32 NFL teams over the past ten seasons. These variables were used to obtain the best possible model for predicting the SuperBowl winner for the 2010-2011 NFL season.

4.1 Data

In football the *playoff level* categorical variable has six levels, which are defined as follows:

0 = Did not qualify for play-offs
1 = Lost in the wild card round
2 = Lost in the conference finals

3 = Lost in the division finals
4 = Lost in the superbowl
5 = Won the superbowl

Beside this categorical value, which has been designated as "Result" in the dataset file, six other variables were analyzed:

- First half record
- Second half record
- End of season record
- Total points scored
- Total points given up
- Differential in points

4.2 Analysis

In order to determine the effect each variable had on the standings of an NFL team at the end of a season, and if they had a chance at winning the SuperBowl, several trending patterns had to be analyzed or determined. The simplest way to do this was to create several correlation graphs and see which one had the greatest effect on the results of a team. The following variables were the most directly correlated with the post season performace of a team when looking at every team:

- End of season record (0.709)
- Differential in points (0.665)
- First half record (0.618)

The following variables were the most directly correlated with the post season performance of the team when looking at only teams which made the play-offs:

- End of season record (0.425)
- Differential in points (0.370)
- First half record (0.343)

It was very interesting to see that it doesn't matter if the correlation is ran on all teams in the league, or on just the twelve teams that made the play-off each year. Either way the same three statistics are the most significant when trying to determine the results of the play-offs. Although, when looking at just the play-off teams, each of the statistics where far less correlated than when looking at all the teams. With the exception of a few outliers, an analysis of Figure 6 shows that the end record of a team is fairly significant in determining their playoff performance. It seems strange that the correlation between these two values is less than 0.5 (it may be due to the outliers). A similar course of action occurs when looking at Figure 9 and it is easy to see that as the points differential increases the team's chances of going furthur in the play-offs also increases. Although the difference of all the levels is not as drastic as the last comparison.

To further this investigation, and also to test out some interesting theories, I deceided to see if it was better to have a good offense or defense when going into the play-offs. As you can see by Figure 6 and Figure 7 the adage of "the best offense is a good defense" holds true. A team is much more likely to do well in the playoffs if their defense did well during the regular season. Also, a team with a great offense is likely to lose in the SuperBowl and not win it according to the analysis. In order to create a model for predicting the winner of the SuperBowl this year, a logistic regression was leveraged. After the model was obtained, it was possible to determine which of the NFL teams in this season have the best chance of winning the SuperBowl.

4.3 Results

By looking over the results of previous seasons and applying a best fit algorithm obtained from that to the current season, the following teams have the greatest chance of winning the superbowl this year (in decending order, after week 12):

- New England Patriots
- Atlanta Falcons
- New Orleans Saints
- Chicago Bears
- Pittsburgh Steelers

If this statistical analysis ends up holding true then the final four would be New England vs. Pittsburgh, and Atlanta vs. Chicago with the superbowl being New England vs. Atlanta. NOTE: This data set was taken after week 12, so it is not completely up to date but we still believe that it is important to note how much predictions can change over the course of a season.

5. BASKETBALL

The goal of this study, as previously explored, is to discover and use trends that exist in sports history to predict the winner of this year's championship. For basketball, 10 years of seasonal statistical data has been collected. This dataset includes 22 fields which will then be reduced to determine variables are most significant in determining which team is going to win the playoffs. After finding the pertinent variables, a model can be obtained for predicting the playoff winner for this season.

5.1 Data

For basketball the categorical variable *playoff level* was defined with six levels as follows:

0 = Did not qualify for post-season.
1 = Lost in the Sweet Sixteen.
2 = Lost in the Elite Eight.
3 = Lost in the Final Four.
4 = Lost in the Championship.
5 = Won the Championship.

In the dataset the *playoff level* variable was named as "final". Additionally, the following regular season variables were collected and analyzed:

- Team name
- Games won
- Games lost
- Total minutes played
- Field goals made
- Field goals attempted
- Threes made
- Threes attempted
- Free throws made
- Free throws attempted
- Offensive rebounds
- Total rebounds
- Assists
- Steals
- Turnovers
- Blocks
- Personal fouls
- Technical fouls
- Ejections
- Flagrant fouls
- Total points

5.2 Analysis

To determine the effect of each variable on the teams post-season performance, trending patterns need to be identified. To this end, several linear models were constructed and analyzed. In addition, a logistic regression was performed on these models. The variables most directly correlated with post-season performance are, in descending order:

- Games won
- Free throws attempted
- Personal fouls

Common sense may lead one to automatically assume that the most significant variable would be games won, however this was not the case in other sports as previously seen. Fortunately, in every season analyzed, the number of games won during the regular season was the single most significant variable, so this is consistent with those expectations. A visual representation of the trends present is available in Figure 1. Rarely does the team with the most wins also bring home the championship. In fact, if a team wins the most games, there is a high probability they will not advance past the Final Four.

As for the free throws attempted and personal fouls, the two variables are highly related, as committing many fouls leads to many free-throw attempts. In this case, it is better to analyze the independent variable, personal fouls. Figure 2 shows the relationship between post-season performance and number of personal fouls. Teams with lower seasonal foul counts tend to do better in the post-season. This is because in basketball if a team is losing at the end of the game, that team will actively attempt to foul more frequently in an attempt to regain posession of the ball. Therefore, the teams that lose the most games will likely end up with a higher foul count at the end of the regular season, and as such will not have qualified for the playoffs.

In basketball, players are considered more skilled if they can consistently score three-pointers. To determine the significance of a team's overall skill level in relation to their post-season performance, Figure 3 shows a boxplot of post-season performance against total number of three-pointers made by the team during the season. It is clear that there is little to no trending occurring here. Interestingly, in 9 out of 10 years analyzed, the winners of the championship had a 3-pointer accumulation between 400 and 500. In no year did the team with the most 3-pointers win the championship, suggesting that if a team consists purely of super-stars who never pass, they are doomed in the post-season.

5.3 Results

By analyzing previous trends of winners, one sees that the following characteristics should be met to have the highest probability of winning the basketball championship:

- The team does not win the most games, but rather is around the 75% proficiency mark in that category.
- The team does not consist entirely of superstars taking only 3-point shots on the net.
- The team plays by the rules and keeps fouling to a minimum.

So far for the 2010 season, the following teams have been the closest to matching those attributes, and thus have the highest probability of winning the championship. They are, in descending order:

- Chicago Bulls
- Indiana Pacers
- Denver Nuggets
- Phoenix Suns
- New Orleans Hornets

6. HOCKEY

The hocket analysis occurred over the past five years for all NFL teams, and included 9 regular season variables as well as two post season variables. Statistical analysis on this dataset was implored in an attempt to obtain a model for accurately predicting playoff performance.

6.1 Data

The post-season of the NHL is fairly uniform which makes analysis of post-season performance. Each round of the post-season is played in a best of seven format. Therefore, the winner of the Stanley Cup always has sixteen wins in the post-season games. The number of wins a team has for the post-season can then be linearlly converted to their respective *playoff level* using the following equation:

If the team made the playoffs: Let PlayoffWins be the number of post-season wins.
otherwise Let PlayoffWins = -4

Playoff Level = PlayoffWins / 4 + 1

The *playoff level* categorical variable is defined with six levels as follows:

0: Team did not make the playoffs
1: Team did not advance past the first round of the playoffs
2: Team advanced past the first round of the playoffs
3: Team advanced past the second round of the playoffs
4: Team made it to the finals
5: Team won the Stanley Cup

In addition to playoff variables, the following 9 regular season variables were collected to determining playoff performance:

- Regular season wins
- Goals scored
- Goals scored against the team
- Overtimes won
- Overtimes lost
- Penalties
- Penaly time in minutes
- Power play opportunities
- Power play goals

Using these variables, statistical analysis will occur to determine their relation to the corresponding *playoff level.*

6.2 Analysis

The analysis of post-season performance in hockey is rather different due to limitations in the dataset obtained from the NHL. Because data on individual quarters of the season was unavailable, the analysis had to occur over the data obtained from the entire regular season. Given this limitation certain data needs to be eliminated from the set before performing the analysis. If an analysis occurred over the entire season for every team, inaccurate conclusions may arise. For instance, there is a strong correlation between the number of wins in the regular season and the number of wins in the post season. However this correlation arises from the fact that the number of regular season wins determines which teams enter the playoffs. Therefore before the analysis occurs, the teams that did not make the playoffs were eliminated from the data set altogether.

In order to determine which regular season variables are likely to determine playoff performance a correlation between the variables and the number of playoff season wins was obtained. Although the correlation takes place between the number of playoff season wins and the regular season variables, it provides a basis for determining which variables are most influential to *playoff level.* Because the playoff wins and *playoff level* variables are linear, analysis on playoff wins can help draw conclusions on *playoff level* while providing the benefit of using quantitative analysis. The analysis of correlations between the playoff season wins and regular season variables it is found that most variables have very little correlation to number of playoff wins. The largest correlation is 0.287, and the three most correlated variables have been listed in descending order:

- Regular Season Goals Scored
- Power Play Goals
- Regular Season Wins

In order to further obtain a picture about which variables are significant in determining *playoff level* a multivariate ANOVA analysis was performed with playoff wins as the response and regular season variables as the explanatory variables. However, it was found that none of the regular season variables had much significance in determining playoff wins. The following variables were significant:

- Regular Season Wins
- Regular Season Goals

In addition to the multivariate ANOVA analysis, a logistic regression was taken to determine a model for predicting *playoff level.* While providing the model for this study, the logistic regression also gives a means for significance analysis. The following regular season variables were significant in the logistic regression:

- Regular Season Goals Against the Team
- Regular Season Wins
- Regular Season Goals
- Regular Season Penalties

Lastly, two boxplots were obtained using regular season wins as the explanatory variable, and playoff wins as the response for one graph and *playoff level* as the response for the second graph. These graphs can be observed in Figures 4 and 5 of the Appendix. A quick analysis of the graphs show that the Stanley Cup winners are evenly distributed over the range of playoff wins between 45 and 54.

6.3 Results

The results of the hockey analysis are largely inconclusive. The attempts made to find correlation and significance patterns between the regular season variables and the post season data yieled very weak relationships between the datasets. Despite these inconclusive results, another attempt to predict the Stanley Cup winner was made based on the logistic regression model.

Using the formula obtained from the logistic regression model a test was conducted using data from the 2009-2010 NHL season. The test was very simple: substitute the variables from a sample of teams in order to observe the *playoff level* the model yields. If the predicted *playoff level* corresponds to the team's actual *playoff level* then this model provides a basis for predicting the Stanley Cup winner of the 2010-2011 season. The results of this test were completely invalid with the Stanley Cup winner yielding a *playoff level* between 1 and 2, and one of the first round losers yielding a *playoff level* between 4 and 5.

The results of the hockey analysis are largely not useful in respect to predicting *playoff level* based on a team's regular season data. However, a significant conclusion also arises from this study. The conclusion is that a hockey team's regular season data cannot determine that team's *playoff level*. The results of this analysis seem to confirm that, but it must be determined if this conclusion provides accuracy in concrete scenarios. In the 2010 NHL post season such a concrete example can be found. The first seed of the 2009-2010 playoff season (Washington Capitols) lost in the first round of the playoffs. On the other hand, the last seed of the 2009-2010 playoff season (Philadelphia Fliers) made it all the way to the Stanley Cup Finals. Such practical scenarios as the one discussed in the 2009-2010 playoff season make it impossible to obtain a model for predicting *playoff level* based on regular season variables.

7. CONCLUSIONS

Before beginning this analysis on all the major sports in the United States we had high hopes that we would be able to find some way of acurately predicting the future winners of each. Obviously it would not be this easy as many people and companies have been trying to do this for years for profit. Each sport was slightly different, and each sport yielded slightly different results in how accurate our calculations ended up being in predicting the winner.

Even though our final results and analysis of each sport did not yeild perfectly accurate predictions of the winners of each, they did yeild some other interesting facts. We found out that in baseball the number of games won per quarter was highly significant and in order to make the playoffs it is best to have between 23 and 30 wins in the 3rd quarter of the season. In basketball it is best to not win the most amount of the games and to play by the rules. In football having a good start to the season and a good but not great offense was key. In hockey it is just not possible to conclude that any regular season data can provide an accuarate prediction of playoff performance.

It was not all that surprising that our final calculations were not entirely accurate. Professional sports are incredibly random and many outside factors can effect the outcome of games that may not be known before hand. Factors such as injuries, weather, player trades, and luck. Even though we were not able to decisevely say what team would win in each sport this season, we found a lot of good information, and are one step closer to more accurate predictions.

8. REFERENCES

[1] Major League Baseball Statistics
http://www.mlb.com
28 November 2010.

[2] Pro Football Reference
http://www.pro-football-reference.com/
24 November 2010.

[3] NBA Team Stats
http://www.nba.com/statistics/sortable_team_statistics/sortable1.html
28 November, 2010.

[4] NBA Post-Season Ladder Stats
http://dougstats.com/ 28 November, 2010.

[5] NHL Team Stats
http://www.nhl.com/ice/teamstats.htm?navid=nav-sts-teams
1 December, 2010.

APPENDIX

Figure 1: Basketball: Boxplots of games won per season x vs. Post-season performance y

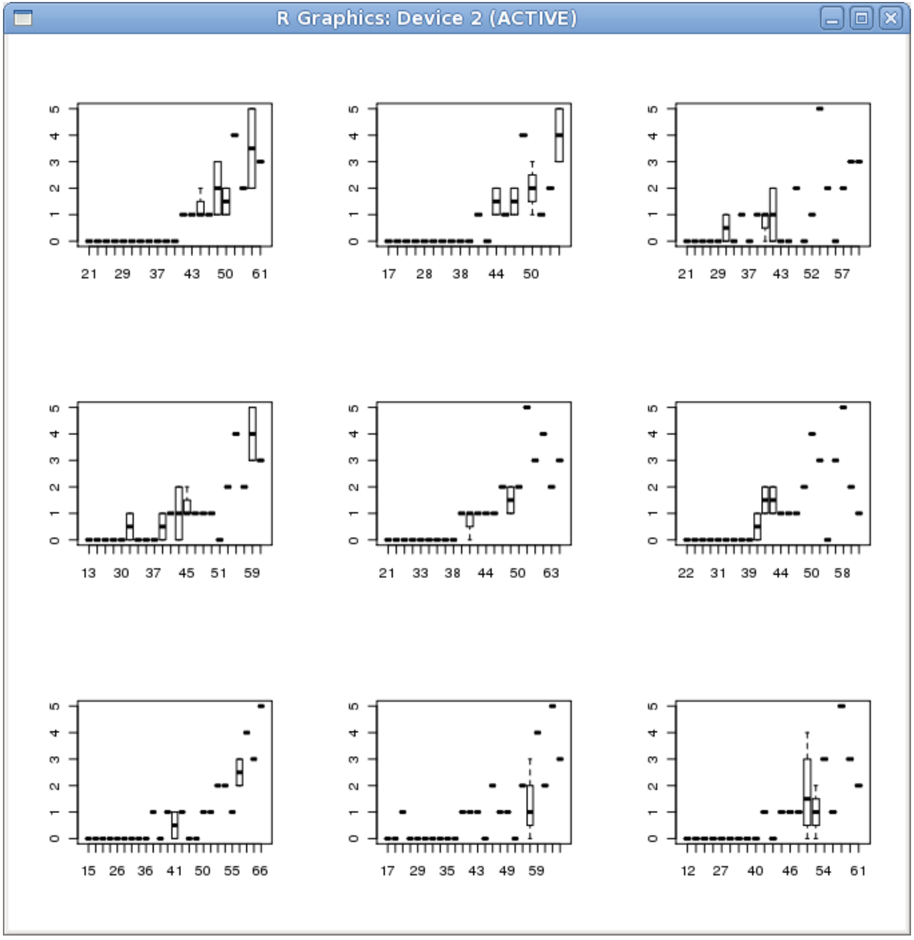


Figure 2: Basketball: Boxplots of personal fouls x vs. Post-season performance y

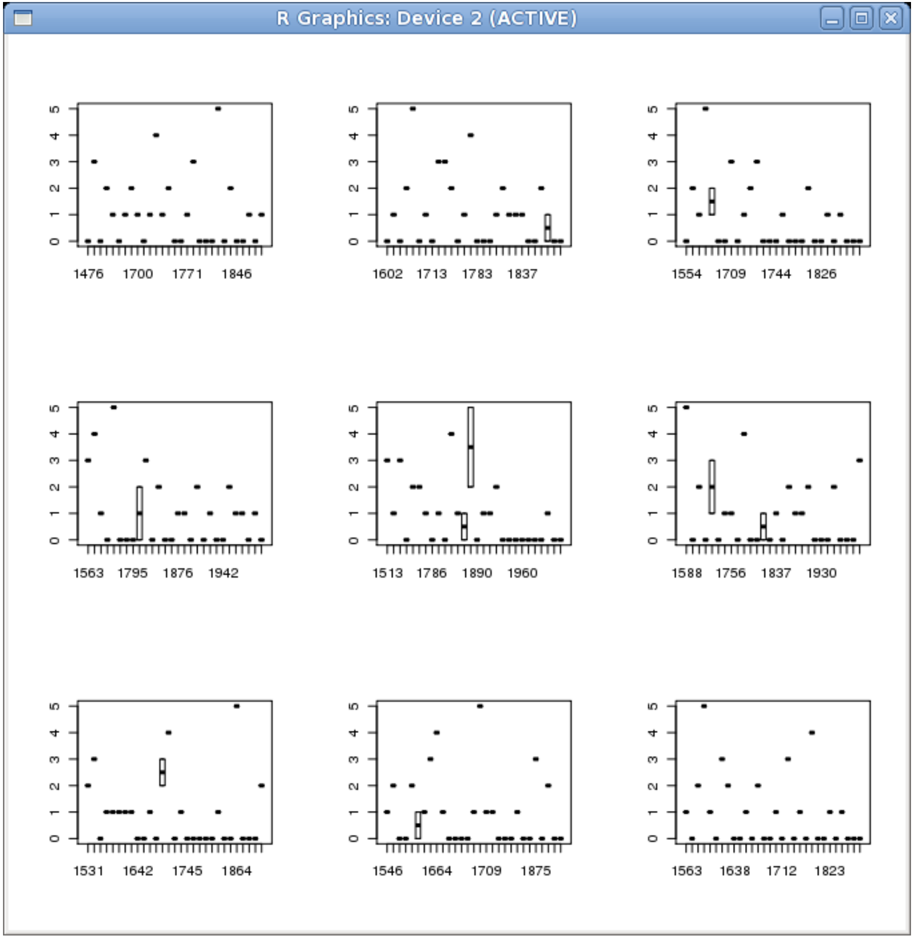


Figure 3: Basketball: Boxplots of 3-pointers made x vs. Post-season performance y

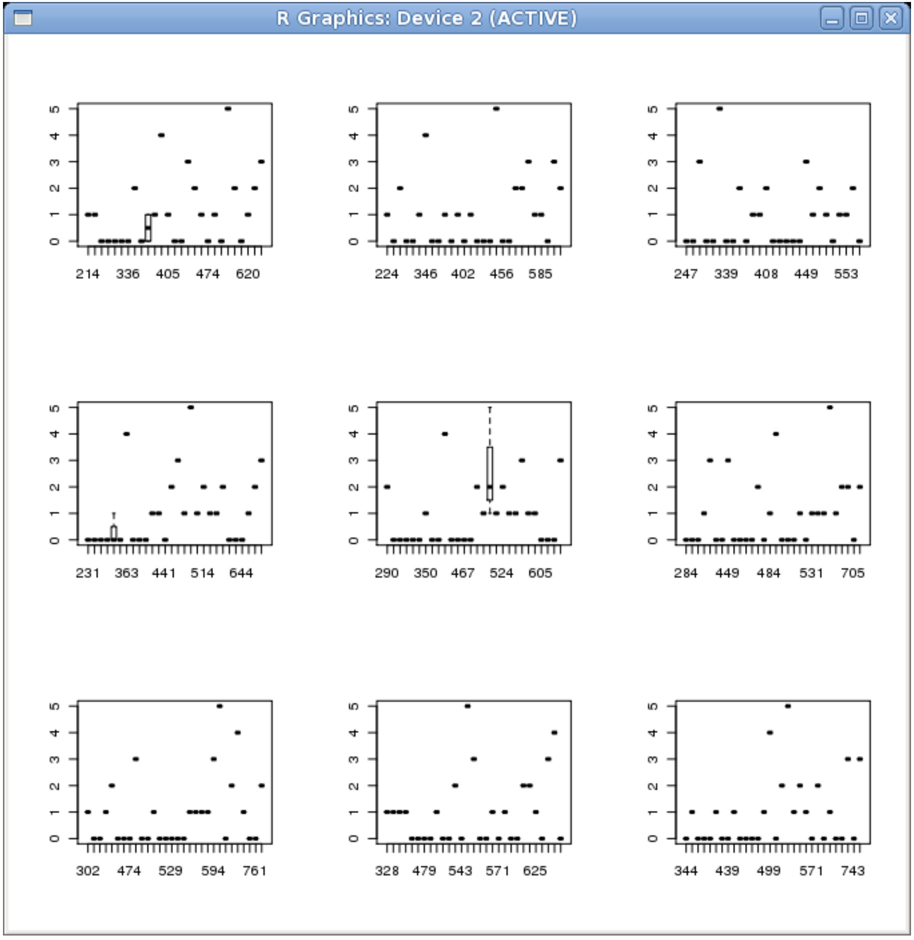


Figure 4: Hockey: Boxplots for regular season wins x vs. Playoff Level y

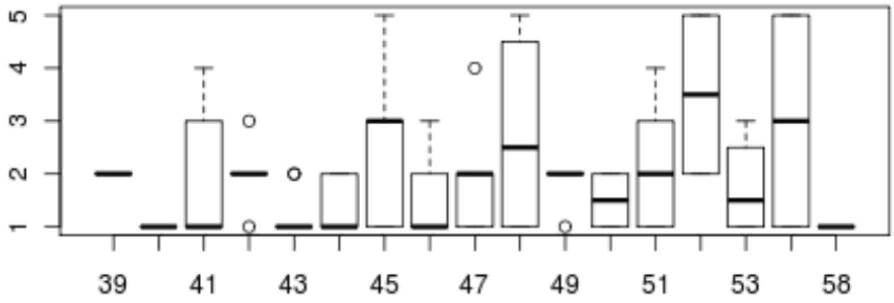


Figure 5: Hockey: Boxplots for regular season wins x vs. Playoff Wins y

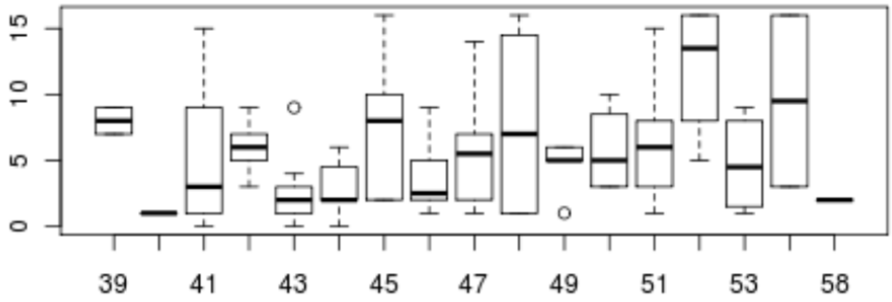


Figure 6: Football: Boxplots for Season wins x vs. Playoff level y

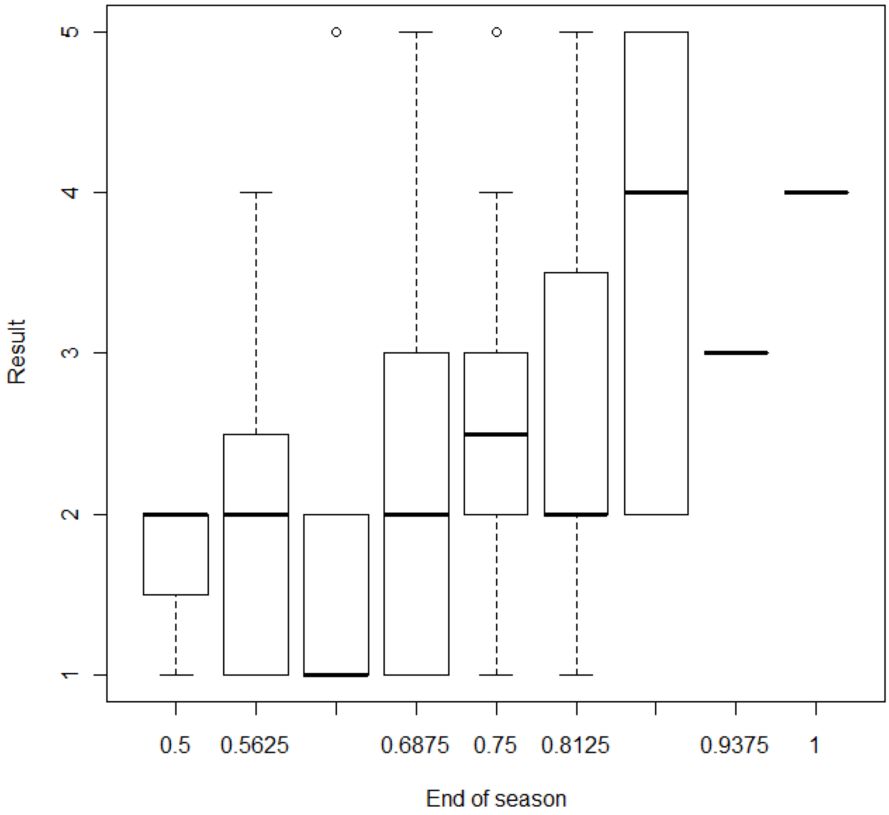


Figure 7: Football: Boxplots for Playoff level x vs. Points scored against y

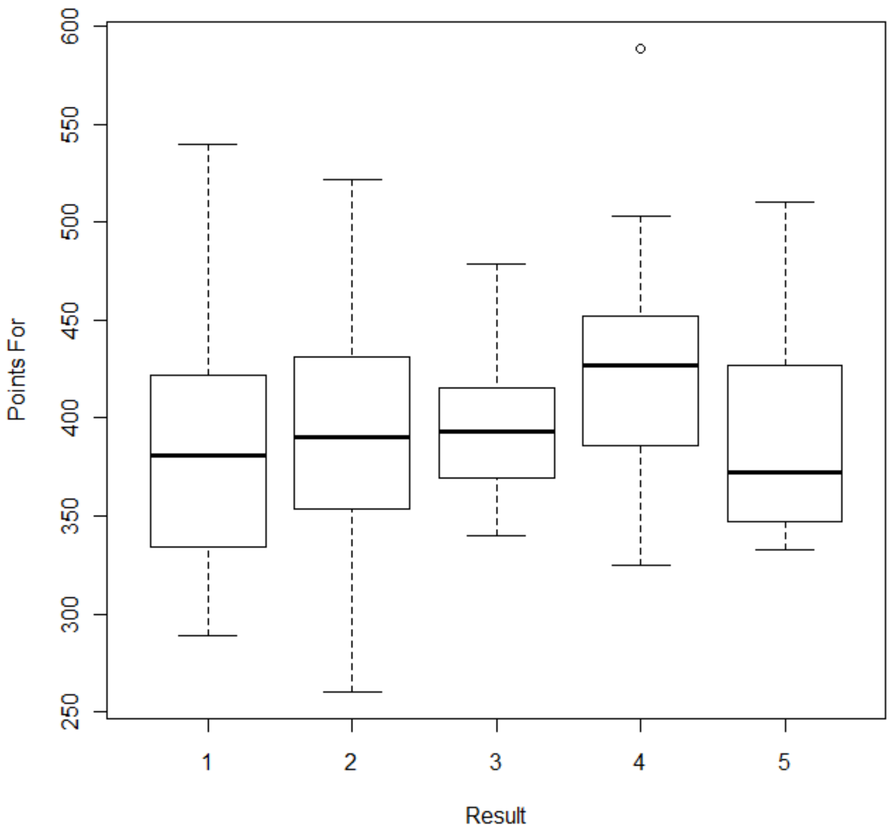


Figure 8: Football: Boxplots for Playoff level x vs. Points scored y

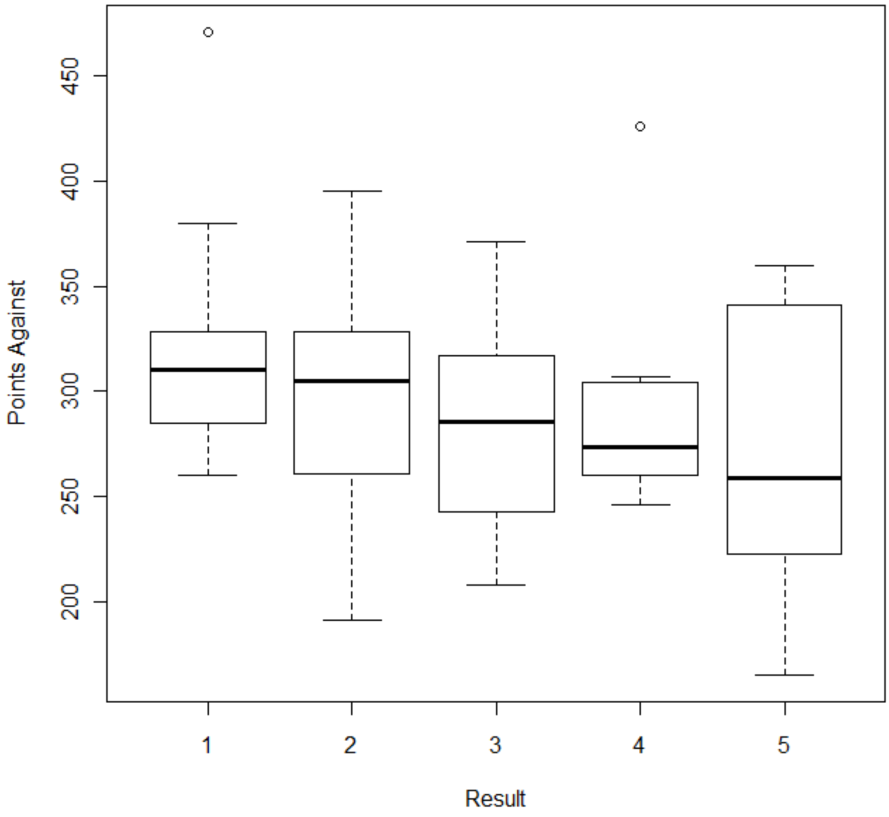


Figure 9: Football: Boxplots for Playoff level x vs. Point differential y

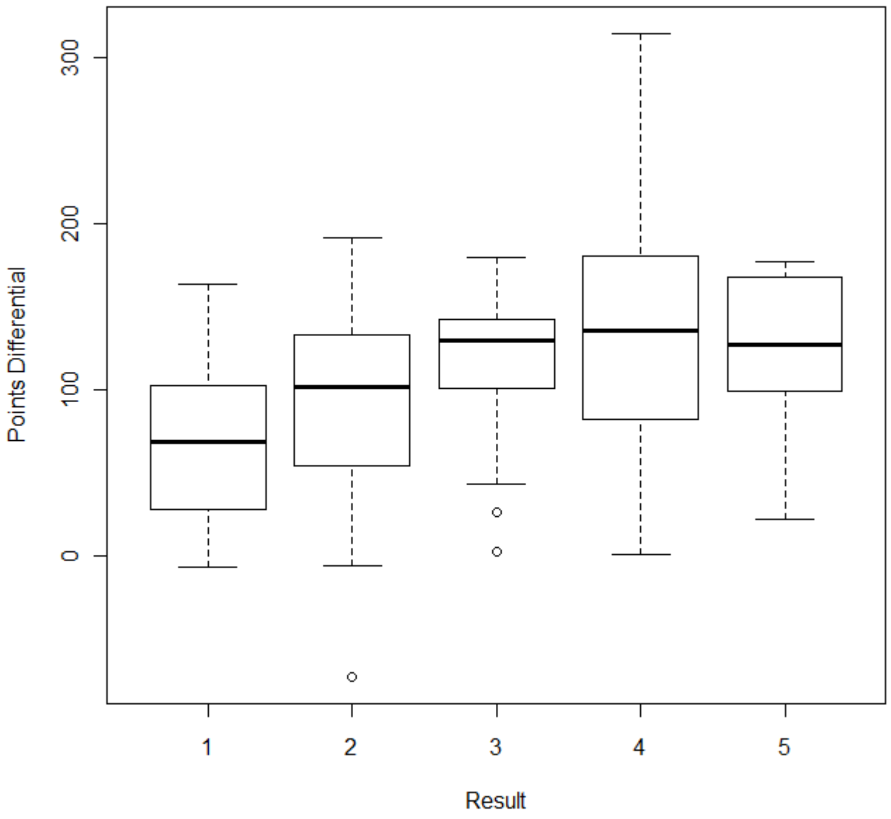


Figure 10: Boxplot for the first quarter wins x vs. Playoff Level y

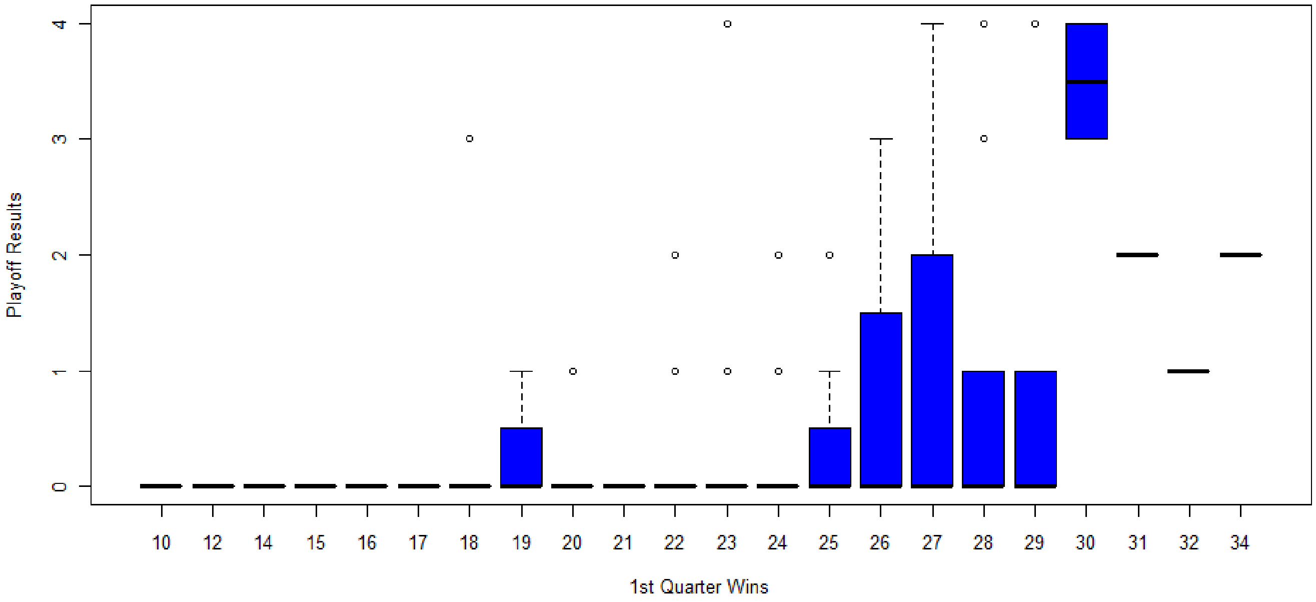


Figure 11: Boxplot for the second quarter wins x vs. Playoff Level y

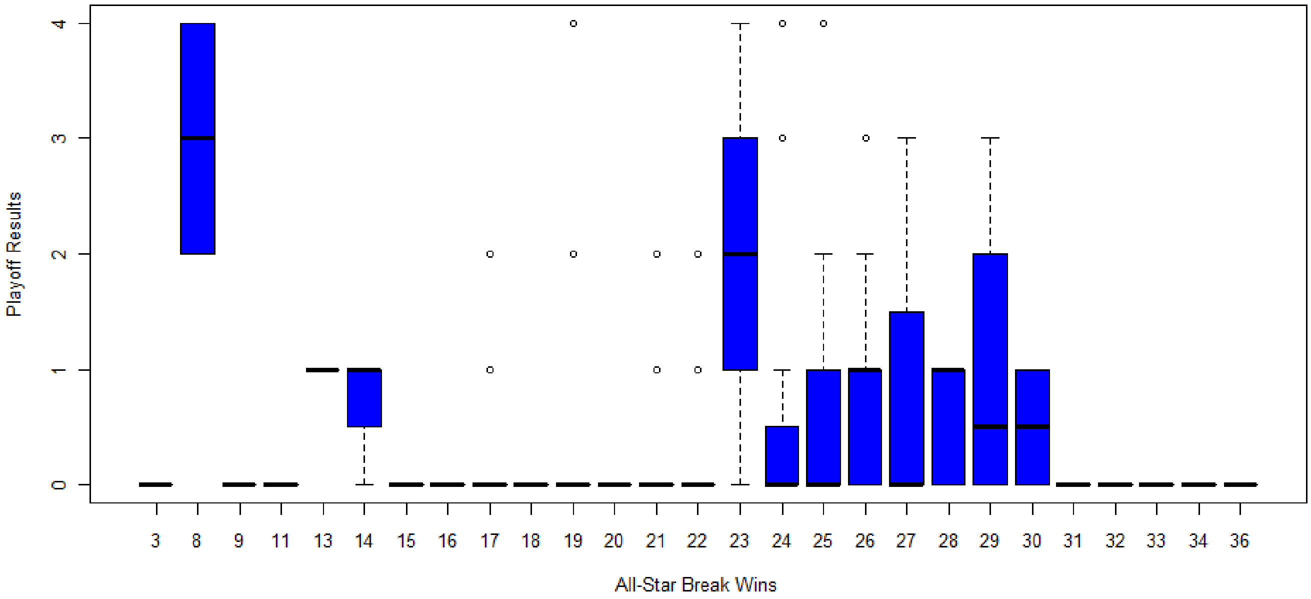


Figure 12: Boxplot for the third quarter wins x vs. Playoff Level y

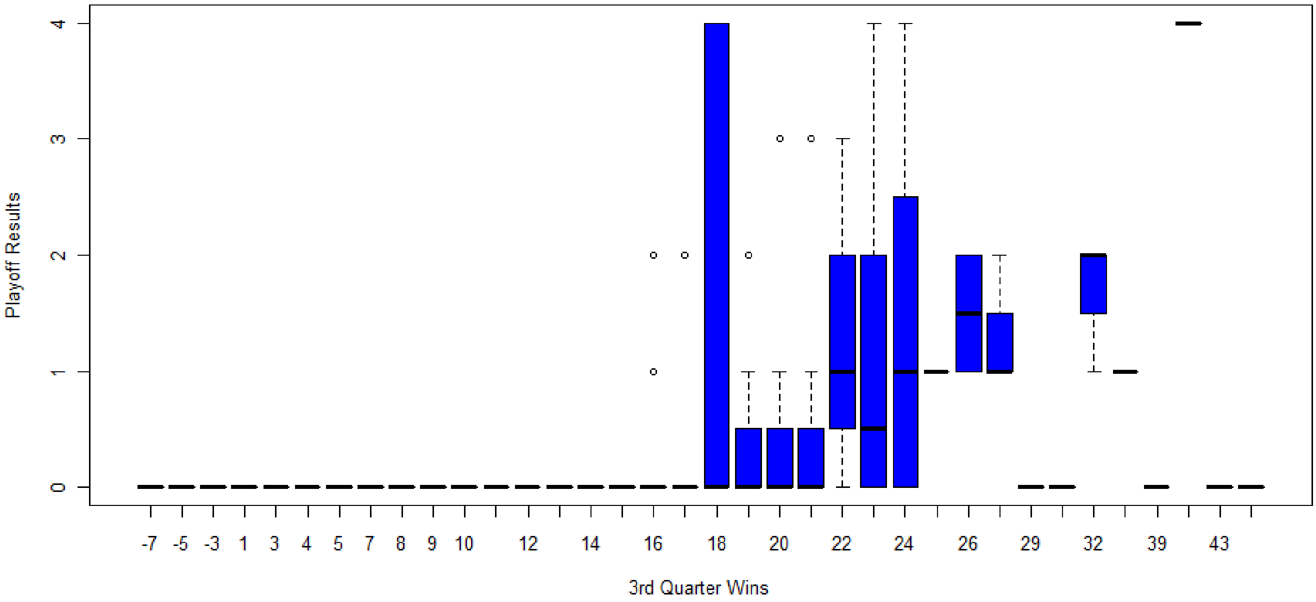
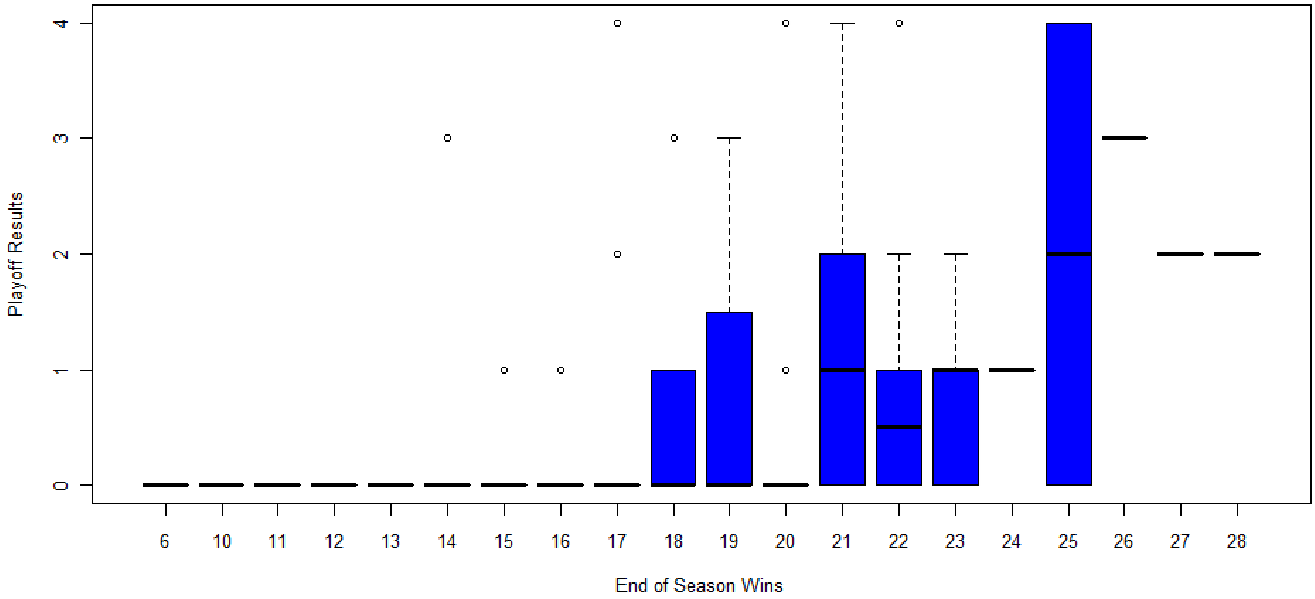


Figure 13: Boxplot for the fourth quarter wins x vs.
Playoff Level y


Playoff Results
End of Season Wins