SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

# SAS & R

## A PERFECT COMBINATION FOR SPORTS ANALYTICS

#SASGF

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

# SAS & R: A Perfect Combination for Sports Analytics

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

## ABSTRACT

Revolution Analytics reports more than two million R users worldwide. SAS has the capability to use R code, but users have discovered a slight learning curve to performing certain basic functions such as getting data from the web. R is a functional programming language while SAS is a procedural programming language. These differences create difficulties when first making the switch from programming in R to programming in SAS. However, SAS/IML enables integration between the two languages by enabling users to write R code directly into SAS/IML. This paper details the process of using the SAS/IML command Submit /R and the R package “XML” to get data from the web into SAS/IML. The project uses public basketball data for each of the 30 NBA teams over the past 33 years, taken directly from Basketball-Reference.com. The data was retrieved from 66 individual web pages, cleaned using R functions, and compiled into a final dataset composed of 48 variables and 895 records. The seamless compatibility between SAS and R provide an opportunity to use R code in SAS for robust modeling. The resulting model provides a clear and concise approach for those interested in pursuing sports analytics, as well as, a performance comparison between SAS and R.

## INTRODUCTION

Moving from one program to another can provide challenges, especially when users have built proficiencies that do not directly translate over to the new program. SAS makes moving from R to SAS simple with SAS/IML’s integration with R. This integration allows users to write R commands directly into SAS/IML, call R packages and functions, and transfer data between the two programs seamlessly. We used this integration with R to illustrate how users can scrape data from the web using the R package XML and the function readHTMLTable within SAS/IML. This process allows for a continuous workflow and streamlined code. Using these procedures, we analyzed 33 years of basketball data and looked at trends in the game over that span.

## OBTAINING THE DATA

The data for this project came from Basketball-Reference.com, a branch of the Sports Reference family, one of the leading sources for sports statistics in the world. Data is organized according to the season in which the statistics occurred. Thus, to analyze 33 seasons of data one would have to pull statistics from 33 different web pages. Additionally, we added opponent statistics to these data sets, requiring two tables to be downloaded from each web page. An example of the data is presented in Figure 1.

**Team Stats** Playoff teams are marked with an asterisk (*) · Glossary · SHARE · Embed · CSV · Export · PRE · LINK · ?

| Rk | Team | G | MP | FG | FGA | FG% | 3P | 3PA | 3P% | 2P | 2PA | 2P% | FT | FTA | FT% | ORB | DRB | TRB | AST | STL | BLK | TOV | PF | PTS | PTS/G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Golden State Warriors* | 82 | 19730 | 3410 | 7137 | .478 | 883 | 2217 | .398 | 2527 | 4920 | .514 | 1313 | 1709 | .768 | 853 | 2814 | 3667 | 2248 | 762 | 496 | 1185 | 1628 | 9016 | 110.0 |
| 2 | Los Angeles Clippers* | 82 | 19730 | 3228 | 6830 | .473 | 827 | 2202 | .376 | 2401 | 4628 | .519 | 1468 | 2067 | .710 | 784 | 2711 | 3495 | 2031 | 640 | 409 | 1012 | 1749 | 8751 | 106.7 |
| 3 | Dallas Mavericks* | 82 | 19880 | 3255 | 7036 | .463 | 732 | 2082 | .352 | 2523 | 4954 | .509 | 1386 | 1843 | .752 | 858 | 2608 | 3466 | 1846 | 663 | 371 | 1062 | 1644 | 8628 | 105.2 |
| 4 | Oklahoma City Thunder | 82 | 19830 | 3184 | 7119 | .447 | 632 | 1864 | .339 | 2552 | 5255 | .486 | 1524 | 2020 | .754 | 1052 | 2844 | 3896 | 1681 | 598 | 454 | 1205 | 1829 | 8524 | 104.0 |
| 5 | Toronto Raptors* | 82 | 19855 | 3108 | 6829 | .455 | 726 | 2060 | .352 | 2382 | 4769 | .499 | 1585 | 2014 | .787 | 881 | 2526 | 3407 | 1701 | 615 | 357 | 1057 | 1712 | 8527 | 104.0 |
| 6 | Houston Rockets* | 82 | 19805 | 3032 | 6832 | .444 | 933 | 2680 | .348 | 2099 | 4152 | .506 | 1525 | 2133 | .715 | 958 | 2624 | 3582 | 1820 | 777 | 407 | 1366 | 1803 | 8522 | 103.9 |
| 7 | San Antonio Spurs* | 82 | 19955 | 3208 | 6854 | .468 | 677 | 1847 | .367 | 2531 | 5007 | .505 | 1368 | 1754 | .780 | 806 | 2772 | 3578 | 2000 | 657 | 444 | 1146 | 1564 | 8461 | 103.2 |
| 8 | Cleveland Cavaliers* | 82 | 19780 | 3089 | 6739 | .458 | 826 | 2253 | .367 | 2263 | 4486 | .504 | 1453 | 1934 | .751 | 911 | 2612 | 3523 | 1814 | 603 | 340 | 1171 | 1510 | 8457 | 103.1 |
| 9 | Portland Trail Blazers* | 82 | 19855 | 3175 | 7049 | .450 | 807 | 2231 | .362 | 2368 | 4818 | .491 | 1272 | 1589 | .801 | 879 | 2881 | 3760 | 1799 | 525 | 372 | 1117 | 1494 | 8429 | 102.8 |
| 10 | Atlanta Hawks* | 82 | 19730 | 3121 | 6699 | .466 | 818 | 2152 | .380 | 2303 | 4547 | .506 | 1349 | 1735 | .778 | 715 | 2611 | 3326 | 2111 | 744 | 380 | 1167 | 1457 | 8409 | 102.5 |

*Figure 1: Basketball-Reference.com 2014 season team statistics table (top 10 teams).*

While Basketball-Reference.com allows users to easily download tables into various formats, downloading 66 of these tables individually would be quite the daunting task. Hence, the R package XML and function readHTMLTable become very useful. By applying readHTMLTable to a list of the 33 web pages, each table from those web pages is almost

# SAS & R: A Perfect Combination for Sports Analytics

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

instantly read into SAS/IML. A preview of this code can be seen below in Figure 2.

```
url.list = sprintf("http://www.basketball-reference.com/leagues/NBA_%s.html",
                   c(1980:1998, 2000:2011, 2013:2014))  #create a list of the links

columnclasses = c("character", "character", rep("integer", 4), "numeric",
                  rep("integer", 2), "numeric", rep("integer", 2),
                  "numeric", rep("integer", 2), "numeric",
                  rep("integer", 9), "numeric")   #vector for setting column names

library(XML)

for(i in 1:33) {  #read in the tables from the urls
  nam = paste0("x", i)
  assign(nam, readHTMLTable(url.list[[i]], colClasses = columnclasses,
                            stringsAsFactors=FALSE))
}
```

*Figure 2: Application of XML function readHTMLTable to Basketball-Reference.com data.*

Additional work is needed to clean and join the team statistics and opponent statistics tables; however, the result is a data set of 48 variables and 836 observations, as seen below in Figure 3.

| | 48 | SeasonEnd | Team | W | Playoffs | FG | oppFG | FGA | oppFGA | Fgpercent | oppFGpercent |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 895 | | Int | Nom | Nom | Int | Int | Int | Int | Int | Int | Int |
| 1 | ■ $\chi^2$ | 1980 | Atlanta Hawks | 50 | 1 | 3261 | 3144 | 7027 | 6872 | 0.464 | 0.458 |
| 2 | ■ $\chi^2$ | 1980 | Boston Celtics | 61 | 1 | 3617 | 3439 | 7387 | 7313 | 0.49 | 0.47 |
| 3 | ■ $\chi^2$ | 1980 | Chicago Bulls | 30 | 0 | 3362 | 3585 | 6943 | 7222 | 0.484 | 0.496 |
| 4 | ■ $\chi^2$ | 1980 | Cleveland Cavaliers | 37 | 0 | 3811 | 3811 | 8041 | 7610 | 0.474 | 0.501 |
| 5 | ■ $\chi^2$ | 1980 | Denver Nuggets | 30 | 0 | 3462 | 3736 | 7470 | 7591 | 0.463 | 0.492 |
| 6 | ■ $\chi^2$ | 1980 | Detroit Pistons | 16 | 0 | 3643 | 3847 | 7596 | 7761 | 0.48 | 0.496 |
| 7 | ■ $\chi^2$ | 1980 | Golden State Warriors | 24 | 0 | 3527 | 3438 | 7318 | 6975 | 0.482 | 0.493 |
| 8 | ■ $\chi^2$ | 1980 | Houston Rockets | 41 | 1 | 3599 | 3658 | 7496 | 7382 | 0.48 | 0.496 |
| 9 | ■ $\chi^2$ | 1980 | Indiana Pacers | 37 | 0 | 3639 | 3693 | 7689 | 7545 | 0.473 | 0.489 |
| 10 | ■ $\chi^2$ | 1980 | Kansas City Kings | 47 | 1 | 3582 | 3328 | 7489 | 6992 | 0.478 | 0.476 |
| 11 | ■ $\chi^2$ | 1980 | Los Angeles Lakers | 60 | 1 | 3898 | 3723 | 7368 | 7921 | 0.529 | 0.47 |
| 12 | ■ $\chi^2$ | 1980 | Milwaukee Bucks | 49 | 1 | 3685 | 3456 | 7553 | 7487 | 0.488 | 0.462 |
| 13 | ■ $\chi^2$ | 1980 | New Jersey Nets | 34 | 0 | 3456 | 3480 | 7504 | 7427 | 0.461 | 0.469 |
| 14 | ■ $\chi^2$ | 1980 | New York Knicks | 39 | 0 | 3802 | 3707 | 7672 | 7492 | 0.496 | 0.495 |
| 15 | ■ $\chi^2$ | 1980 | Philadelphia 76ers | 59 | 1 | 3523 | 3444 | 7156 | 7561 | 0.492 | 0.455 |
| 16 | ■ $\chi^2$ | 1980 | Phoenix Suns | 55 | 1 | 3570 | 3563 | 7235 | 7480 | 0.493 | 0.476 |
| 17 | ■ $\chi^2$ | 1980 | Portland Trail Blazers | 38 | 1 | 3408 | 3349 | 7167 | 7008 | 0.476 | 0.478 |
| 18 | ■ $\chi^2$ | 1980 | San Antonio Spurs | 41 | 1 | 3856 | 4000 | 7738 | 7997 | 0.498 | 0.5 |
| 19 | ■ $\chi^2$ | 1980 | San Diego Clippers | 35 | 0 | 3524 | 3752 | 7494 | 7508 | 0.47 | 0.5 |
| 20 | ■ $\chi^2$ | 1980 | Seattle SuperSonics | 56 | 1 | 3554 | 3408 | 7565 | 7424 | 0.47 | 0.459 |
| 21 | ■ $\chi^2$ | 1980 | Utah Jazz | 24 | 0 | 3382 | 3559 | 6817 | 7182 | 0.496 | 0.496 |
| 22 | ■ $\chi^2$ | 1980 | Washington Bullets | 39 | 1 | 3574 | 3615 | 7796 | 7771 | 0.458 | 0.465 |
| 23 | ■ $\chi^2$ | 1981 | Atlanta Hawks | 31 | 0 | 3291 | 3401 | 6866 | 6867 | 0.479 | 0.495 |
| 24 | ■ $\chi^2$ | 1981 | Boston Celtics | 62 | 1 | 3581 | 3372 | 7099 | 7296 | 0.504 | 0.462 |
| 25 | ■ $\chi^2$ | 1981 | Chicago Bulls | 45 | 1 | 3457 | 3527 | 6903 | 7209 | 0.501 | 0.489 |
| 26 | ■ $\chi^2$ | 1981 | Cleveland Cavaliers | 28 | 0 | 3556 | 3608 | 7609 | 7174 | 0.467 | 0.503 |
| 27 | ■ $\chi^2$ | 1981 | Dallas Mavericks | 15 | 0 | 3204 | 3622 | 6928 | 7060 | 0.462 | 0.513 |
| 28 | ■ $\chi^2$ | 1981 | Denver Nuggets | 37 | 0 | 3784 | 4059 | 7960 | 8017 | 0.475 | 0.506 |
| 29 | ■ $\chi^2$ | 1981 | Detroit Pistons | 21 | 0 | 3236 | 3499 | 6986 | 6869 | 0.463 | 0.509 |
| 30 | ■ $\chi^2$ | 1981 | Golden State Warriors | 39 | 0 | 3560 | 3631 | 7284 | 7204 | 0.489 | 0.504 |

*Figure 3: Complete data set with 33 seasons of aggregated statistics.*

## PROCEDURES IN SAS/IML

To run R code in SAS/IML, users must use a special Submit statement, SUBMIT / R, followed by an ENDSUBMIT statement. All R code found between these two statements is executed in R and then passed back into SAS/IML.

```
submit / R;

{insert R code here}

endsubmit;
```

Inserting the code found in Figure 2 will run the code in R and pass it back to SAS/IML. While it is entirely possible to run the code in R, save the data set, and read the saved dataset into SAS/IML, this integration creates continuity of workflow and allows users to work in a single window within one program. This process provides better access to robust modeling tools in SAS.

## ANALYSIS

When analyzing this data set, we started by looking at trends in the NBA over the past 33 seasons. At this point, it is important to note that lockout-shortened seasons were excluded from the analyses; thus, allowing for more consistent data from season to season.  Each season consisted of 82 games for each team, and the analyses were performed for regular season games.

A popular topic in the NBA has been the rise of 3-point shot attempts over the years. Players such as Ray Allen and Stephen Curry have attributed to this trend by shooting at proficiencies never before seen in the league, and teams like the Houston Rockets continue to shoot more

# SAS & R: A Perfect Combination for Sports Analytics

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

and more 3s every season. The rise of analytics in sports has been a strong driver for this change. Figure 4 (below) shows the rise in 3-point attempts (3PA) since the three point line was introduced in the 1979-1980 season.

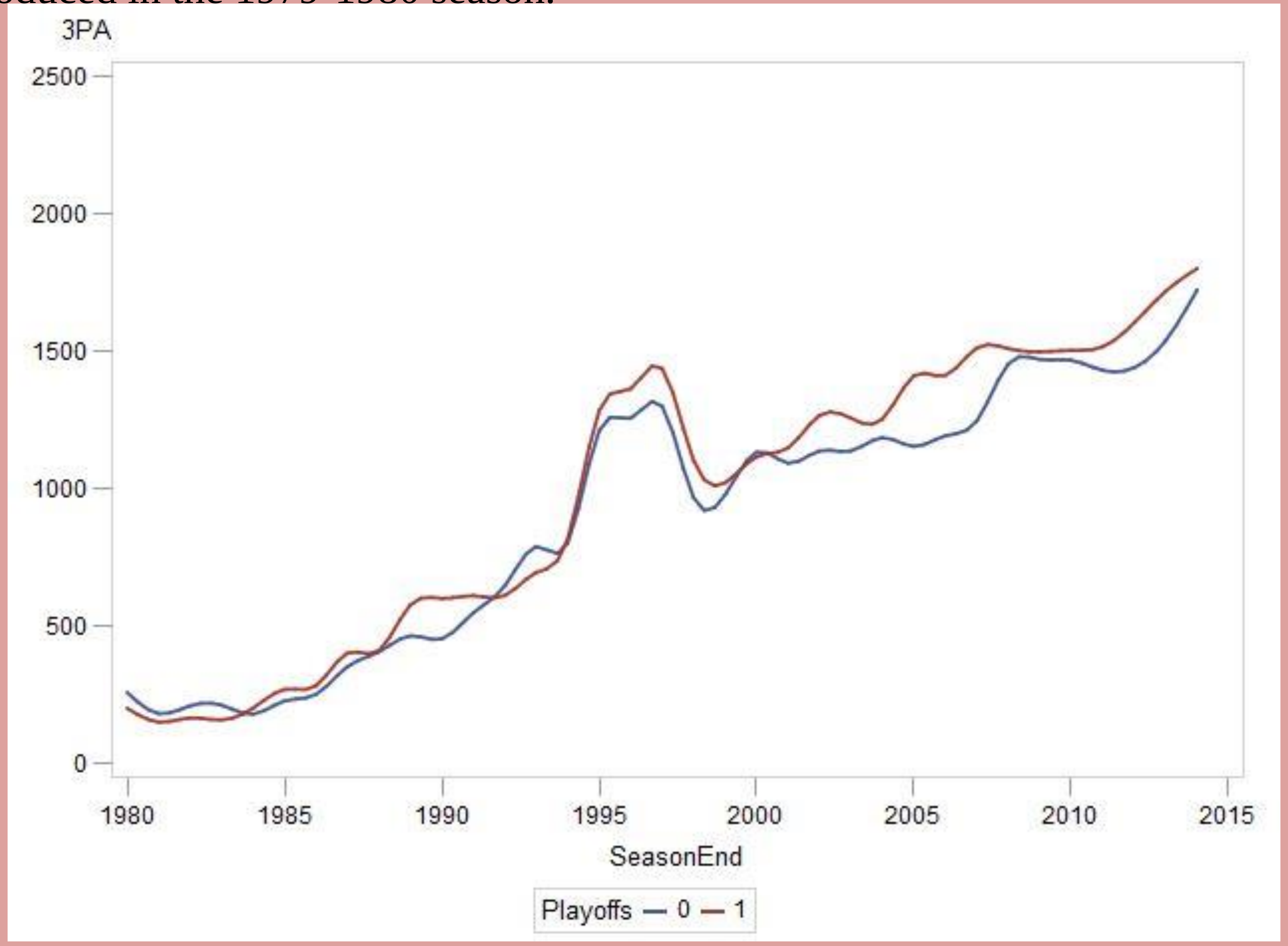


*Figure 4: Rising trend of 3-point shot attempts.*

It is clear that the rate at which 3-point shots are attempted has increased dramatically over this span of 35 years. Additionally, teams that make the playoffs typically attempt more three point shots than teams that do not make the playoffs. This point of interest is likely attributable to the fact that playoffs teams have better shooters, thus, allowing them the freedom to attempt more 3s.

Another topic of discussion in the NBA, recently, has been fouling, specifically the Hack-a-Player phenomenon. When a bad free throw shooter is on the floor, teams will intentionally foul that player hoping they will miss one of two free throws consistently. However, despite the introduction of these new tactics, overall fouls per team per season have declined significantly since 1980.

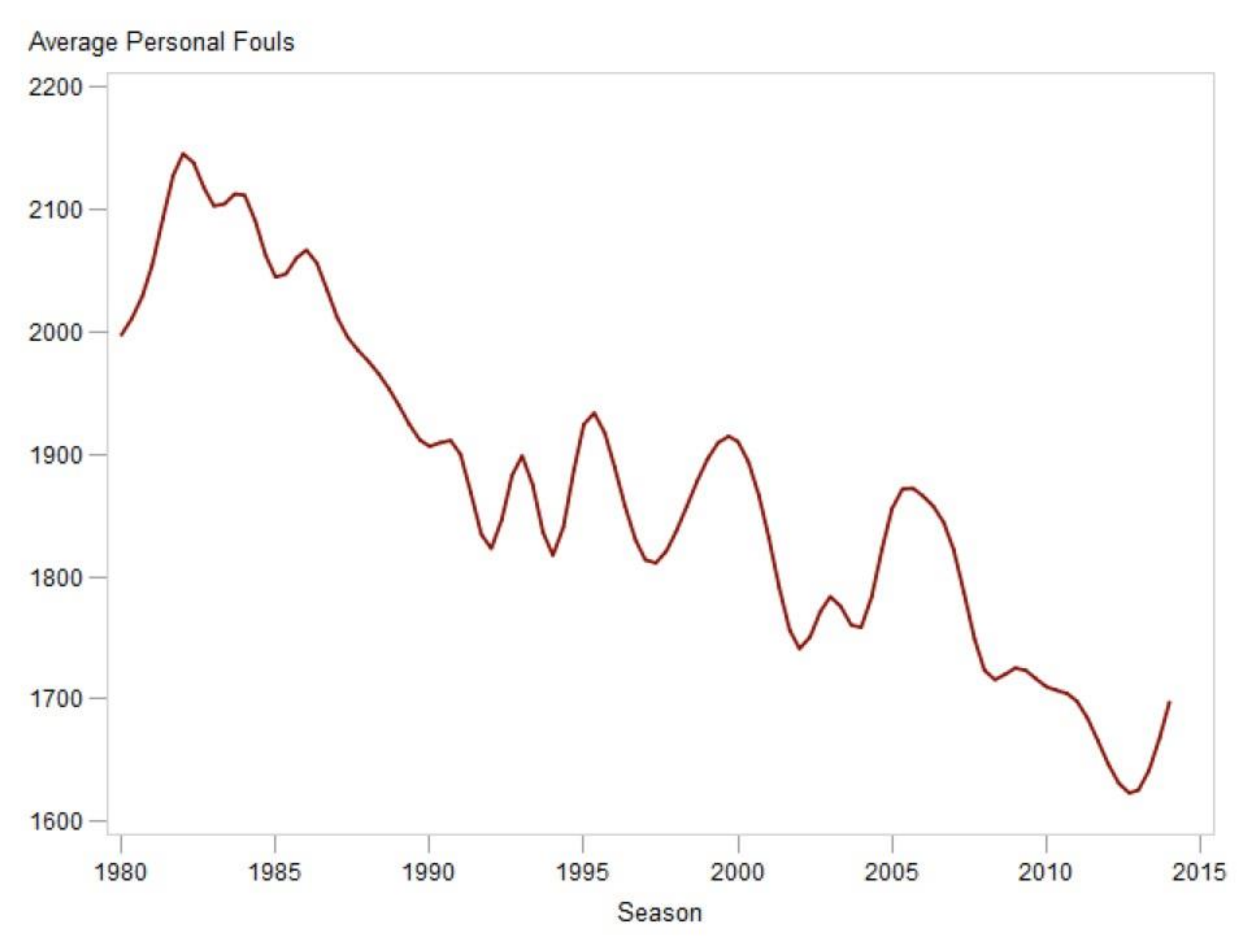


*Figure 5: Declining rate of average fouls per season.*

THE UNIVERSITY OF ALABAMA | Culverhouse College of Commerce

# SAS & R: A Perfect Combination for Sports Analytics

Matthew B. Collins and Taylor K. Larkin

The University of Alabama

As represented in Figure 5, the number of fouls per season has declined greatly, which could be due to the optimization of rules and increased technology and standards among referees.

Using some of the more advanced methods in SAS, the model included Effective Field Goal Percentage (a metric that compensates for 3-point attempts being worth an extra point), Turnover Percentage (the percent of possessions in which a team turns the ball over), and Wins for playoff vs. non-playoff teams in the 2014 season.

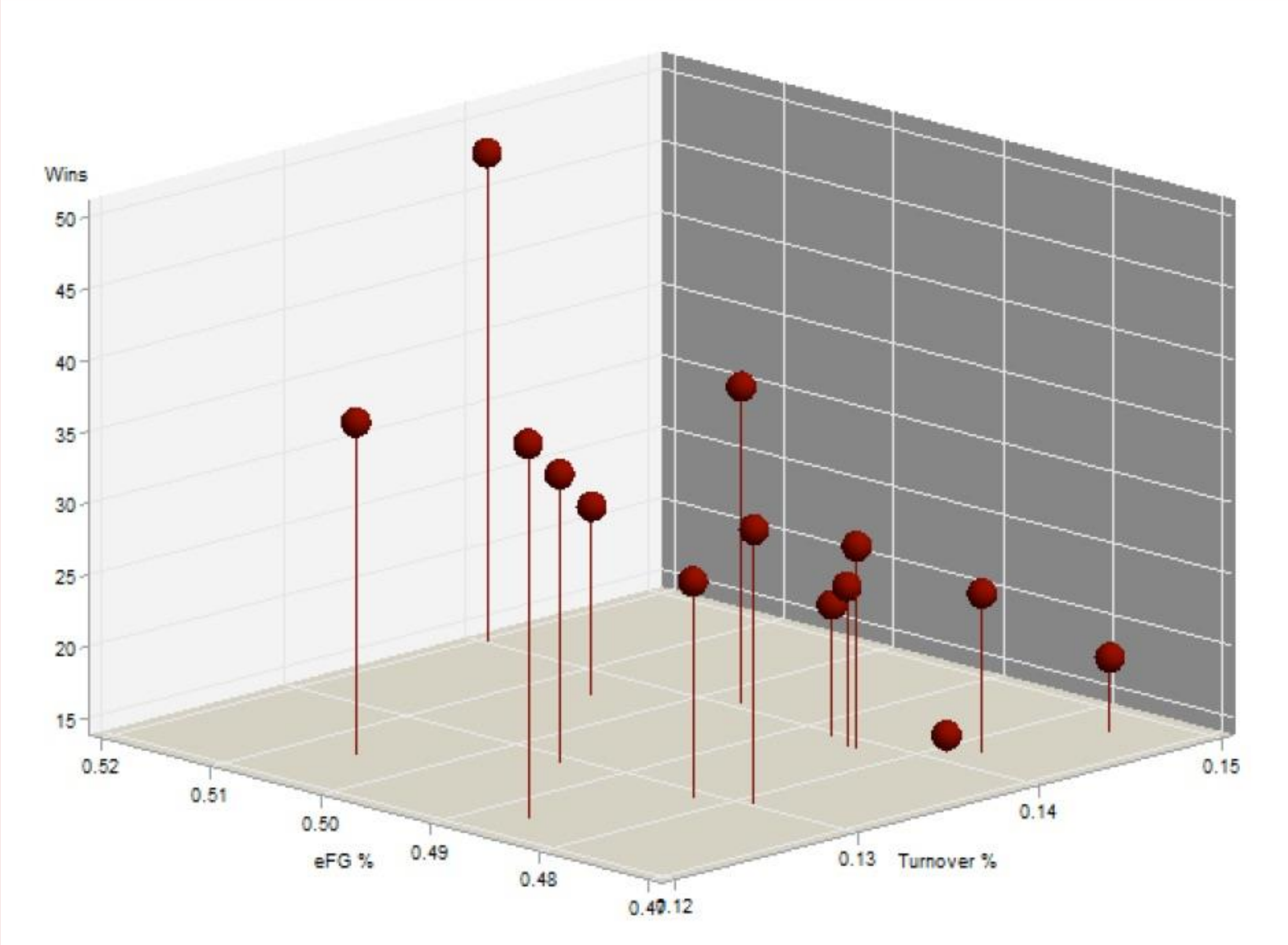


*Figure 6: 3D plot of 2014 season non-playoff teams*

This graph in Figure 6 shows how non-playoff teams with a low number of wins congregate toward the right of the chart, where high Turnover Percentage and low Effective Field Goal Percentage are located. One notable outlier is the team in the upper left of the graph, the Phoenix Suns. Although the Suns had a great Effective Field Goal Percentage and achieved 48 wins, they still missed the playoffs despite having more wins than some teams in the playoffs. The Suns were not in the Playoffs because each of the two conferences in the NBA are guaranteed eight playoffs and the Suns are in the Western Conference, the best of the conference in recent years.

Looking at the playoff teams for 2014, represented in Figure 7, many of the teams are clustered higher in Effective Field Goal Percentage. The team with the highest number of wins in the regular season, the San Antonio Spurs, won the NBA Finals.

The techniques presented in this paper demonstrate how R users can easily use that knowledge to begin performing analysis in SAS/IML and illustrates how the two programs work seamlessly together in the booming field of sports analytics.

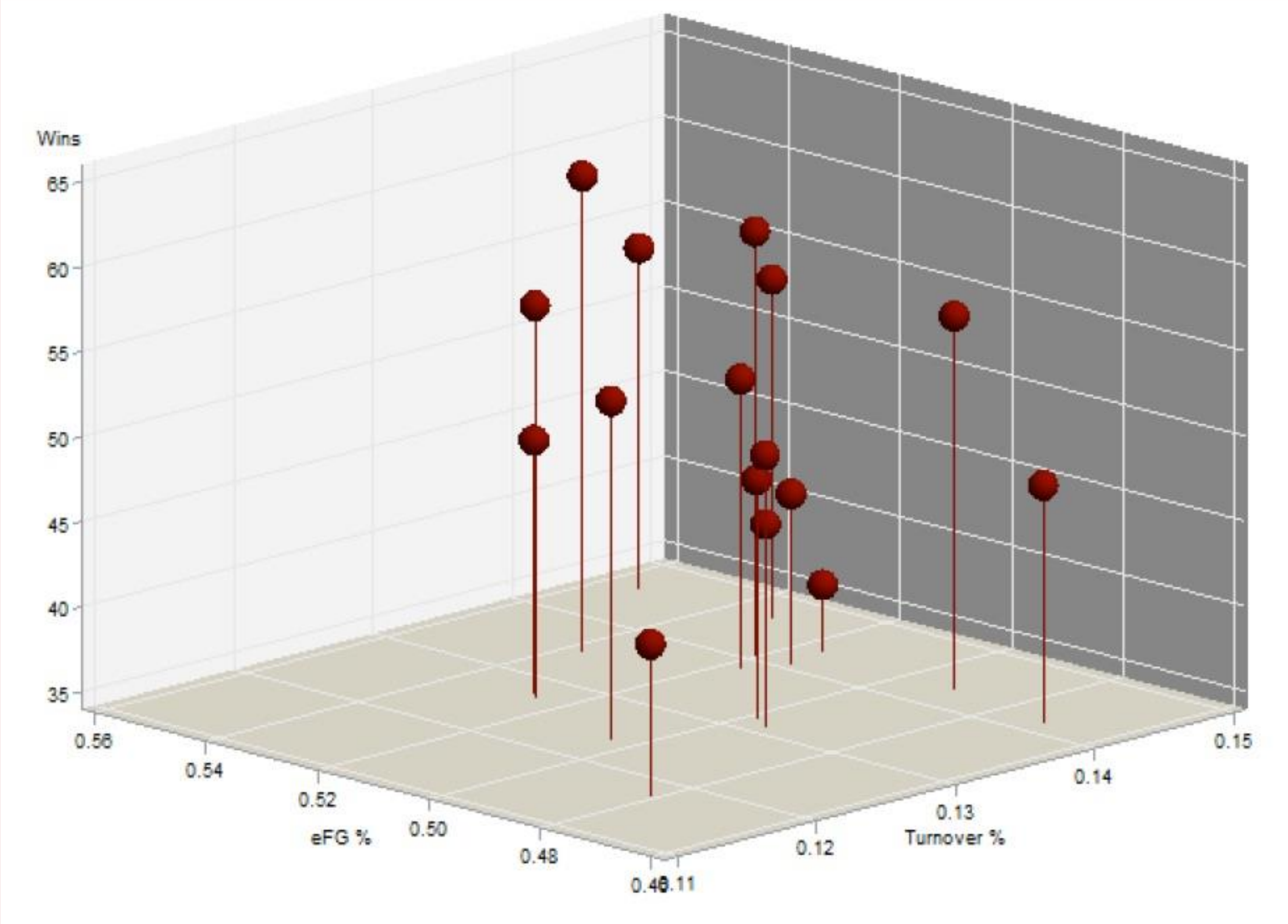


*Figure 7: 3D plot of 2014 season playoff teams*

## REFERENCES

[1] Duncan Temple Lang and the CRAN Team (2015). XML: Tools for Parsing and Generating XML Within R and S-Plus. R package version 3.98-1.3. http://CRAN.R-project.org/package=XML

[2] NBA & ABA League Index | Basketball-Reference.com. (n.d.). Retrieved September 14, 2015, from http://www.basketball-reference.com/leagues/?lid=front_qi_leagues

[3] Sports-Reference.com - Sports Statistics and History. (n.d.). Retrieved from http://www.sports-reference.com

[4] Statistical Programming with SAS/IML Software. (n.d.). Retrieved from http://www.sas.com/en_us/software/analytics/iml.html

[5] SAS/IML(R) 9.22 User's Guide. (n.d.). Retrieved from http://support.sas.com/documentation/cdl/en/imlug/63541/HTML/default/viewer.htm#imlug_r_sect004.htm

SAS® GLOBAL FORUM 2016

IMAGINE. CREATE. INNOVATE.

LAS VEGAS | APRIL 18-21

#SASGF