NFL 2015: A Statistical Analysis

Dave Dyer[1], Jostein Barry-Straume[1], Robert Flamenbaum[1], & Bryan Cikatz[1]

[1] Southern Methodist University

Abstract

This study investigates how average and variance of yards gained impacts a given team's ability to get touchdowns in the NFL. The theory is that teams with high average yards gained, but with low variance, would get more touchdowns than those with higher average yards gained, but higher variance. In other words, consistency is more important than big plays.

Related Work

There are two related works that also use NFL play-by-play data to ascertain trends in yardage and touchdowns. The first, "Underrated NFL Stats" (Iyer, 2011) looks into the big plays allowed by the defense and the teams with the best / worst big plays allowed stats. The second, "How to Quantify the NFL" (Kelly, 2016) looks at big play differentials, but doesn't actually do any statistical analysis. Our analysis focuses on the variance of the yardage per play numbers and measures the efficacy of teams by measuring the touchdowns from each game & team.

Introduction

The data are the 2015 play-by-play records available at Kaggle (https://www.kaggle.com/maxhorowitz/nflplaybyplay2015). These data are comprised of 46,129 rows describing every NFL play run over the 2015 season, by 32 teams. There are 65 columns that include multiple values, including text description of the play / penalty, names of players , touchdown boolean flags, down markers, timestamps, game ids, and yardages. For the purposes of this study, we focused on the average and variance of yards per play, by team, and measured the "success" of the team based on the count of touchdowns per game.

The Tools

For this analysis, we use MongoDB for data mining and basic statistics, and R for data visualization and Markdown. We originally had planned on using MySQL on bluemix, but abandoned it in favor of Mongo after some serious issues whist loading the data.
We used Mongodb hosted on IBM Bluemix and Rstudio with Rmarkdown for the analysis, visualization and writeup. We used many r packages to support our analysis and data mining

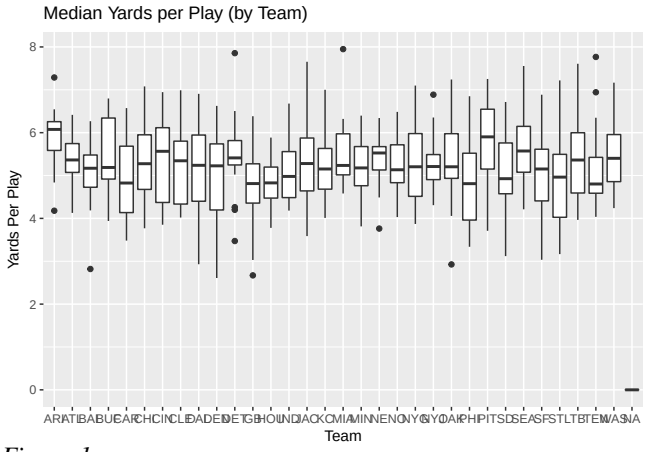


Figure 1

R (3.4.2, R Core Team, 2017) and the R-packages *bindrcpp* (0.2, Müller, 2017), *dplyr* (0.7.4, Wickham, Francois, Henry, & Müller, 2017), *ggplot2* (2.2.1, Wickham, 2009), *knitr* (1.17, Xie, 2015), *mongolite* (1.2, Ooms, 2014), *papaja* (0.1.0.9492, Aust & Barth, 2017, 2017), and *tibble* (1.3.4, Müller & Wickham, 2017). For data mining, we used the mongolite package (https://jeroen.github.io/mongolite/) to connect to the bluemix data set and, where possible, we used mongo to do the statistical analysis in favor of R. Below are the specifications for the Mongo Database:

Bluemix.Technical.Specifications	Specifications
Bluemix Storage	1GB
Data Size	14.3MB
Database Server	Compose for MongoDB-jj
Database Version	3.2.11
Database Location	US South
Cloud Hosting Service	IBM Bluemix
Processors	1 x 2.0 GHz Cores
Memory	1GB RAM

Analysis

To start, we decided to plot number of touchdowns against yards gained (per game). We expect to see a linear trend upwards that illustrates the basic football concept of more yards = more points. As you can see, there appears to be an upward trend – that is, as yards increase, so does the number of touchdowns.

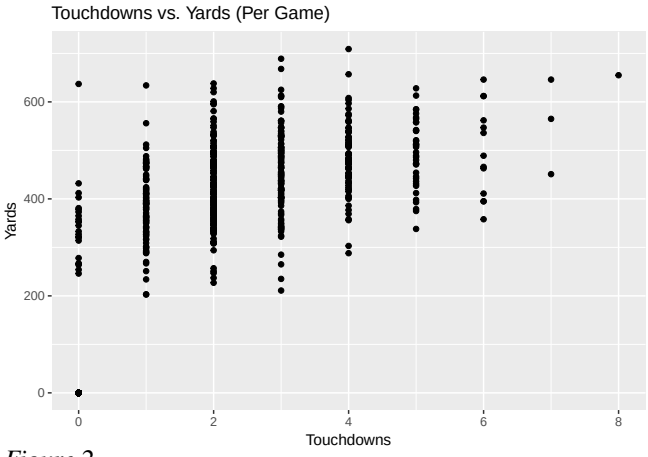
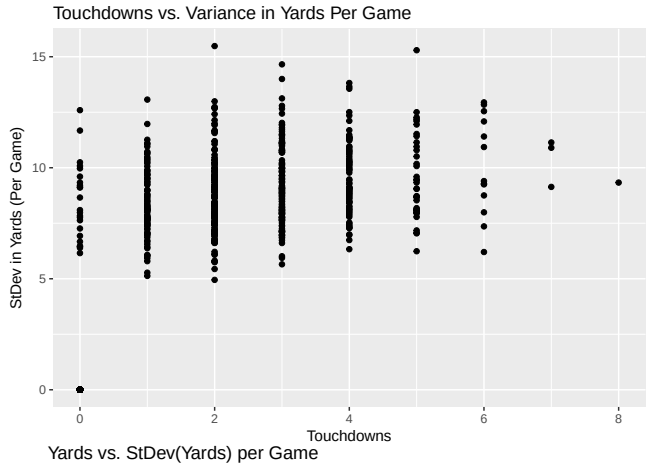
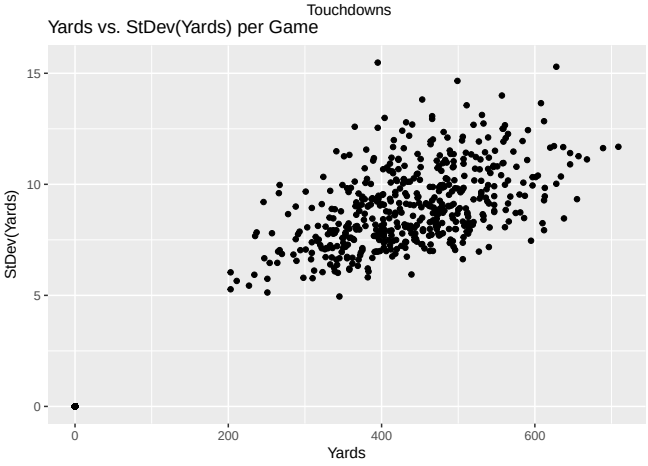


Figure 2





Discussion

We used for all our analyses.

Conclusion

Appendices

References

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Iyer, V. (2011). *Underrated nfl stats: Big plays allowed* (1st ed.). www.sportingnews.com: sportingnews.com. Retrieved from http://www.sportingnews.com/nfl/news/191984-underrated-nfl-stats-big-plays-allowed

Kelly, D. (2016). *How to quantify the nfl: Five stats to better understand football* (1st ed.). www.theringer.com: theringer.com. Retrieved from https://www.theringer.com/2016/8/4/16038580/five-better-nfl-stats-teddy-bridgewater-dwight-freeney-187cb1932

Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from https://CRAN.R-project.org/package=bindrcpp

Müller, K., & Wickham, H. (2017). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [Stat.CO]*. Retrieved from http://arxiv.org/abs/1403.2805

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/