NFL 2015: A Statistical Analysis

Dave, Jostein, Bryan and Robert

November 28, 2017

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] 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] 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

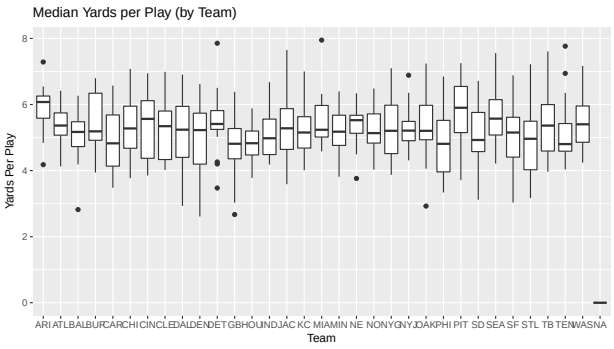
For this analysis, we use MongoDB for data mining and basic statistics, and R for data visualization and Markdown. We originally had planned on using MySQL on bluemix, but abandoned it in favor of Mongo after some serious issues whist loading the data.

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. 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:

```
##                Parameter           Specification
## 1       Bluemix Storage                    1GB
## 2             Data Size                  14.3MB
## 3       Database Server Compose for MongoDB-jj
## 4      Database Version                  3.2.11
## 5     Database Location                US South
## 6 Cloud Hosting Service             IBM Bluemix
## 7            Processors      1 x 2.0 GHz Cores
## 8                Memory                 1GB RAM
```

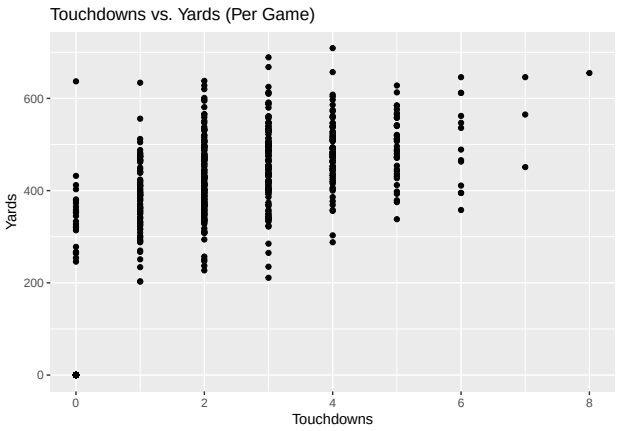
Analysis

When the spread of per-team yards-per-play is viewed in a quartile plot, it is easy to see that there is not a great amount of variance between the medians. This is about what we'd expect, since the NFL is incredibly competitive, and self-normalizing due to constant trading, strategy and balance. If one team was clearly dominant in median yards per play, we would expect that team to dominate the entire season, which is not the case. However, there is quite a bit of difference in the outer quartiles and in the number of outliers. These data support our investigation into whether or not a large variance changes team performance.

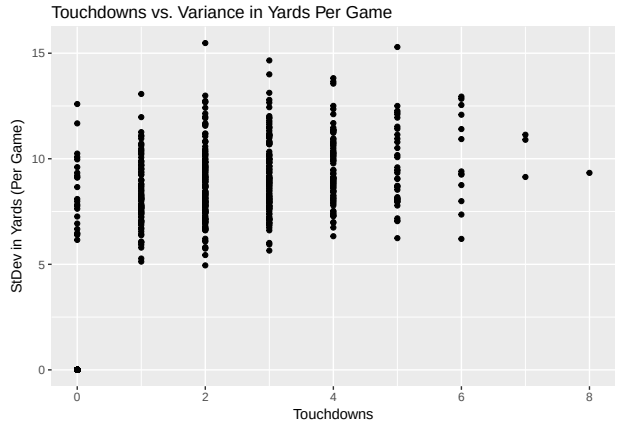


To start, we decided to plot number of touchdowns

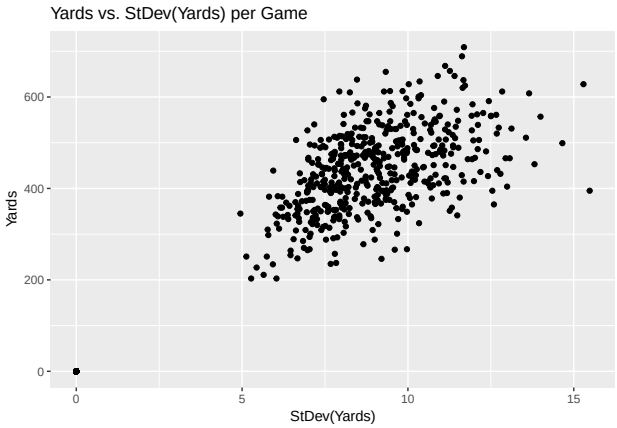
against yards gained (per game). We expect to see a linear trend upwards that illustrates the basic football concept of more yards = more points. As you can see, there appears to be an upward trend – that is, as yards increase, so does the number of touchdowns.



However, the theory of this paper is that, if yards per play is normalized, the lower variance teams will get more touchdowns per game. So it is important to know how the non-normalized standard deviation of yards per play relates to touchdowns per game on average. The below plot shows that it is relatively static, with perhaps a slight rise as standard deviations go up.



However, the data are not normalized, and in order to understand the impact a normalization routine would have on the data, we can examine the relationship between the standard deviation of yards per play and the total yards per game. We expect that, as yards per game goes up, so does total yards, but that is not what's interesting about this plot. Note that it's not uniform linear growth as the standard deviation and the yards per game increase. The spread increases as yards and standard deviation increase, indicating that you get more yards with more variance, but you also run the risk of not getting as many yards (presumably because you're taking more chances, and you get higher rewards for higher risk... sometimes.)



Insert normalization thing here, then talk about it.

Conclusion

This is where we wrap up all the things and make our conclusion.

Appendices

Put code here.

References

Insert references here.

```
#r_refs(file = "r-references.bib")
```