NFL TOUCHDOWNS!

MSDS 7330 – 404

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

INTRODUCTION

ABSTRACT

This study investigates how the 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.

THE DATA

The data are the 2015 play-by-play records available at Kaggle. The uncompressed, raw file is 14.7 MB and is 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.

THE TOOLS

For this analysis, we used 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 whilst loading the data.

TOOLS (CONT.)

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 to connect to the bluemix data set and, where possible, we used mongo to do the statistical analysis in favor of R.

PREVIOUS WORK

- Underrated NFL Stats
  - Based on 2010 season
  - Looks at the relationship between big-plays allowed and touchdowns allowed
  - Big-play defined as 20 or more yards

- How to Quantify the NFL
  - Based on 2015 season
  - Looks at five different statistics – we focus on big-play percentage
  - Big-play defined as runs of more than 10 yards or passes of more than 25 yards

"UNDERRATED NFL STATS"

BEST TEAMS AT PREVENTING BIG-PLAYS

1. **Steelers** (36, 5 TDs)
2. **Chargers** (43, 5 TDs)
3. **Falcons** (48, 9 TDs)
4. **Vikings** (49, 9 TDs)
5. **Buccaneers** (51, 9 TDs)

WORST TEAMS AT PREVENTING BIG-PLAYS

32. Broncos (84, 20 TDs)

31. Seahawks (76, 14 TDs)

30. Cardinals (74, 14 TDs)

29. Jaguars (72, 16 TDs)

28. Redskins (70, 10 TDs)

"UNDERRATED NFL STATS" (CONT.)

KEY INSIGHTS

- "Only six of [the top ten teams] finished .500 or better last season, but those six teams fall in the top seven on the list."
- "Of [the bottom ten teams], only Seattle, with a sub-.500 record, and Philadelphia made the playoffs."

"HOW TO QUANTIFY THE NFL"

KEY INSIGHTS

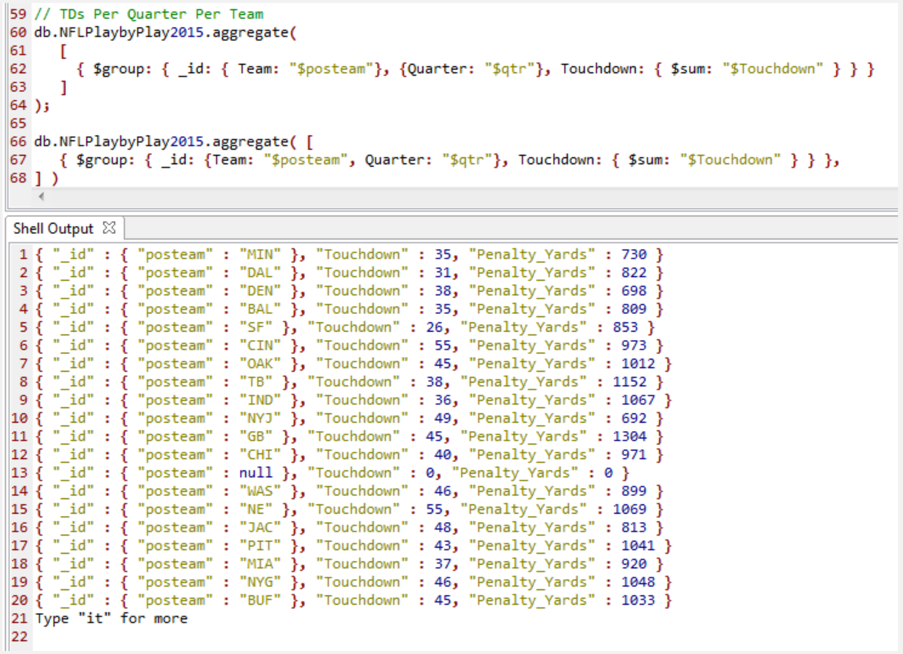
- Teams with the best big-play percentage were the Bills, Vikings, Seahawks, Panthers, Chiefs, and Steelers (roughly 9% of all plays were big-plays).
- Five of these six teams made the playoffs.

EXPANDING UPON PREVIOUS WORK

Ultimately, we are looking to expand upon the role of big-plays in the NFL. Previous work looked more at the defensive side whereas we are focusing on the offensive side of the equation. Furthermore, we seek to discover the importance of consistency in relation to big-plays. We want to know if smaller but more consistent gains are more effective than larger but more irregular gains.

DATA & ANALYSIS

INITIAL QUERIES IN MONGODB



```
// TDs Per Quarter Per Team
db.NFLPlaybyPlay2015.aggregate(
   [
     { $group: { _id: { Team: "$posteam"}, {Quarter: "$qtr"}, Touchdown: { $sum: "$Touchdown" } } }
   ]
);

db.NFLPlaybyPlay2015.aggregate( [
   { $group: { _id: {Team: "$posteam", Quarter: "$qtr"}, Touchdown: { $sum: "$Touchdown" } } },
] )
```

```
{ "_id" : { "posteam" : "MIN" }, "Touchdown" : 35, "Penalty_Yards" : 730 }
{ "_id" : { "posteam" : "DAL" }, "Touchdown" : 31, "Penalty_Yards" : 822 }
{ "_id" : { "posteam" : "DEN" }, "Touchdown" : 38, "Penalty_Yards" : 698 }
{ "_id" : { "posteam" : "BAL" }, "Touchdown" : 35, "Penalty_Yards" : 809 }
{ "_id" : { "posteam" : "SF" }, "Touchdown" : 26, "Penalty_Yards" : 853 }
{ "_id" : { "posteam" : "CIN" }, "Touchdown" : 55, "Penalty_Yards" : 973 }
{ "_id" : { "posteam" : "OAK" }, "Touchdown" : 45, "Penalty_Yards" : 1012 }
{ "_id" : { "posteam" : "TB" }, "Touchdown" : 38, "Penalty_Yards" : 1152 }
{ "_id" : { "posteam" : "IND" }, "Touchdown" : 36, "Penalty_Yards" : 1067 }
{ "_id" : { "posteam" : "NYJ" }, "Touchdown" : 49, "Penalty_Yards" : 692 }
{ "_id" : { "posteam" : "GB" }, "Touchdown" : 45, "Penalty_Yards" : 1304 }
{ "_id" : { "posteam" : "CHI" }, "Touchdown" : 40, "Penalty_Yards" : 971 }
{ "_id" : { "posteam" : null }, "Touchdown" : 0, "Penalty_Yards" : 0 }
{ "_id" : { "posteam" : "WAS" }, "Touchdown" : 46, "Penalty_Yards" : 899 }
{ "_id" : { "posteam" : "NE" }, "Touchdown" : 55, "Penalty_Yards" : 1069 }
{ "_id" : { "posteam" : "JAC" }, "Touchdown" : 48, "Penalty_Yards" : 813 }
{ "_id" : { "posteam" : "PIT" }, "Touchdown" : 43, "Penalty_Yards" : 1041 }
{ "_id" : { "posteam" : "MIA" }, "Touchdown" : 37, "Penalty_Yards" : 920 }
{ "_id" : { "posteam" : "NYG" }, "Touchdown" : 46, "Penalty_Yards" : 1048 }
{ "_id" : { "posteam" : "BUF" }, "Touchdown" : 45, "Penalty_Yards" : 1033 }
Type "it" for more

```

- We ran queries on:
  - Total TD's for the season by team
  - Total TD's for each quarter by team
  - Average yards gained per play by team
  - Average TD's earned per play by team

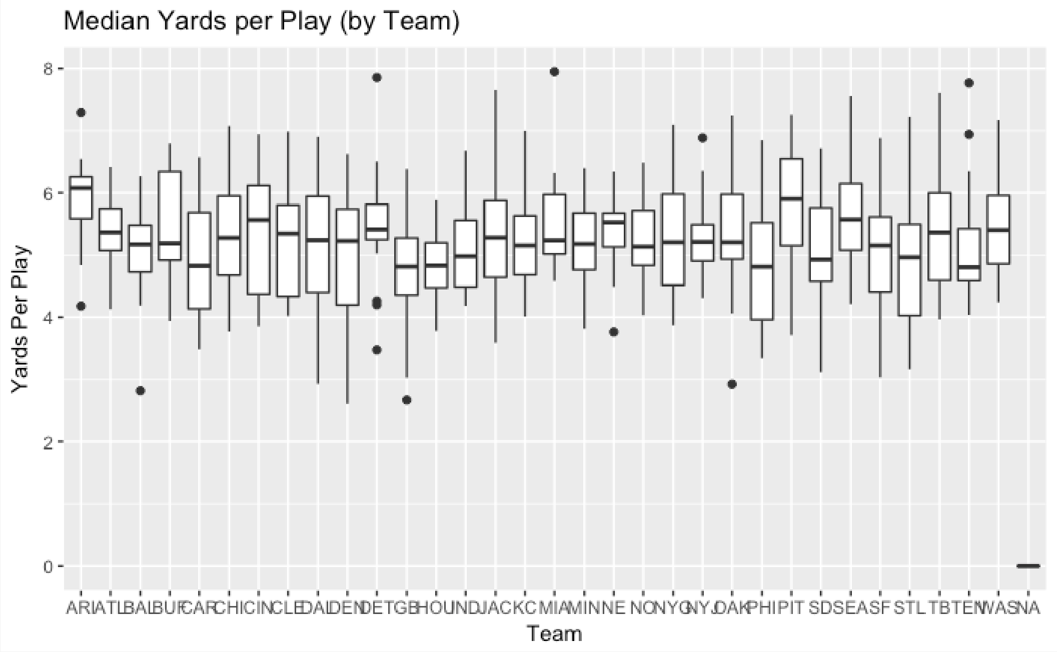
LIST OF R PACKAGES

- library("papaja")
- library("mongolite")
- library("dplyr")
- library("knitr")
- library("tibble")
- library("kableExtra")
- library("ggplot2")

MONGOLITE

```
# Returns average yards gained and average number of touchdowns per team
avgYdsTds <- m$aggregate ( ' [
    { "$group": {
        "_id": "$posteam",
        "ydsGainAvg": { "$avg": "$Yards.Gained" },
        "tdsAvg": { "$avg": "$Touchdown" }
        }
    } ] ' )
```

MEDIAN YARDS PER PLAY



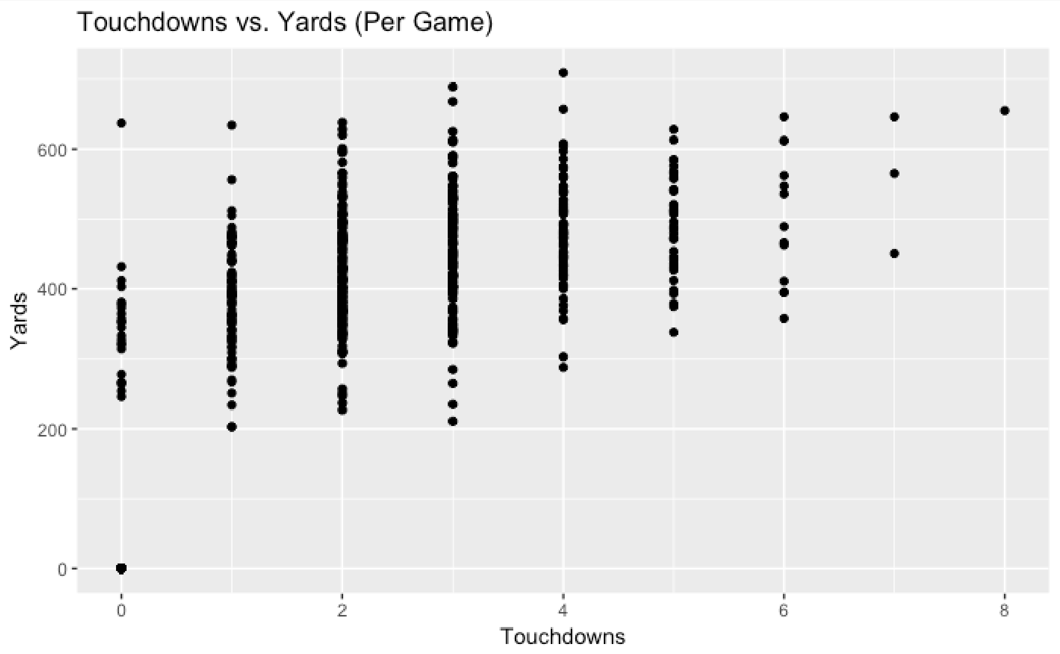
```
bp1 <- ggplot(data = bygame,
aes(posteam, x = posteam, y =
avgYdsPerGamePerTeam))

bp1 + geom_boxplot() +
ylab("Yards Per Play") +
xlab("Team") + ggtitle("Median
Yards per Play (by Team)")
```

INSIGHTS

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.

TD'S VS. YARDS



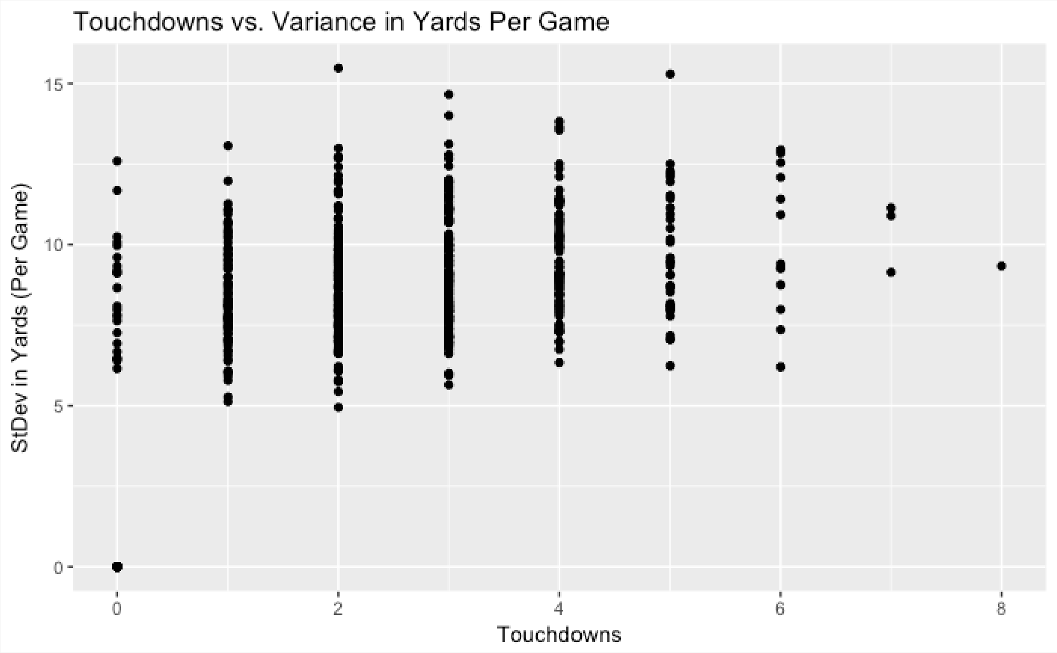
```
p2 <- ggplot(ydsGainVTd,
aes(SumTDs, SumYdsGain)) +
geom_point() + xlab("")
```

```
p2 + xlab("Touchdowns") +
ylab("Yards") +
ggtitle("Touchdowns vs. Yards
(Per Game)")
```

INSIGHTS

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.

TD'S VS. VARIANCE IN YARDS



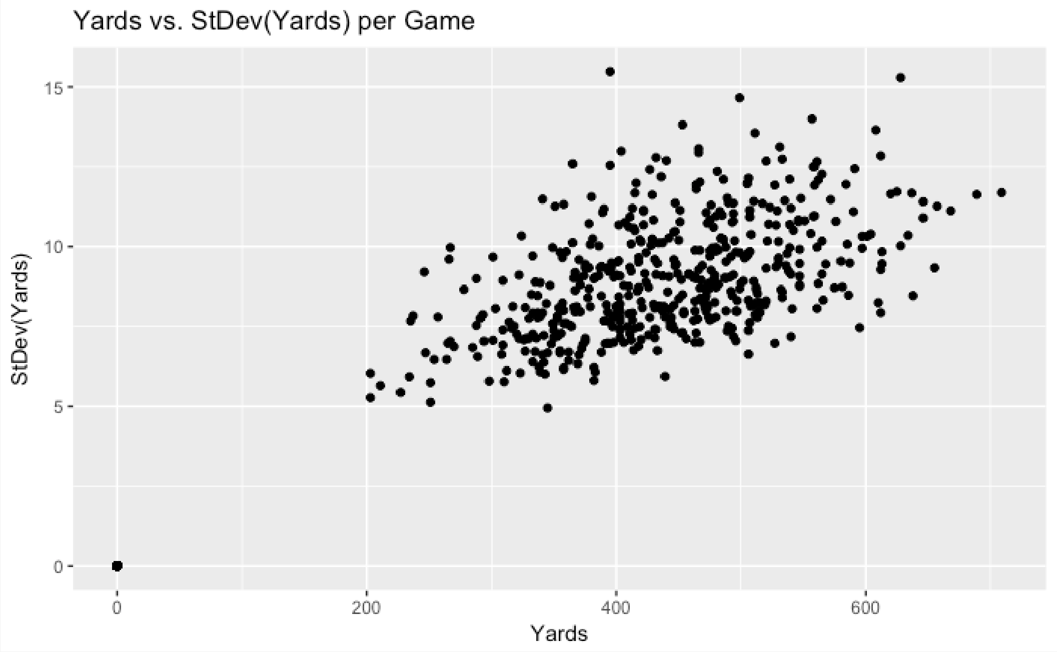
```
ggplot(ydsGainVTd, aes(SumTDs,
SDYdsGain, group =
"_id.GameID")) +

  geom_point() +
xlab("Touchdowns") +
ylab("StDev in Yards (Per Game)")
+ ggtitle("Touchdowns vs. Variance
in Yards Per Game")
```

INSIGHTS

The previous plot shows that the relationship 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 the following plot.

YARDS VS. VARIANCE OF YARDS



ggplot(ydsGainVTd, aes(SumYdsGain, SDYdsGain, group = "_id.GameID")) +

geom_point() + xlab("Yards") + ylab("StDev(Yards)") + ggtitle("Yards vs. StDev(Yards) per Game")

INSIGHTS

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.)

CONCLUSIONS

SPECIFICATIONS & REFERENCES

COMPUTER/PROGRAM SPECIFICATIONS

MONGODB

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

1 x 2.0 GHz Cores

1GB RAM

OS X EL CAPITAN

VERSION 10.11.6

MacBook Pro (Retina 15-inch, Mid 2015)

Processor: 2.2 GHz Intel Core i7

Memory: 16 GB 1600 MHz DDR3

Graphics: Intel Iris Pro 1536 MB

REFERENCES

- 2015 NFL Data from Kaggle
  - https://www.kaggle.com/maxhorowitz/nflplaybyplay2015
- NFL Articles
  - https://www.theringer.com/2016/8/4/16038580/five-better-nfl-stats-teddy-bridgewater-dwight-freeney-187cb19326f1
  - http://www.sportingnews.com/nfl/news/191984-underrated-nfl-stats-big-plays-allowed
- GitHub Link For Our Project
  - https://github.com/tigerninjaproject1/nfl

REFERENCES (CONT.)

- R and R packages
  - https://www.R-project.org/
  - https://github.com/crsh/papaja
  - https://yihui.name/knitr/
  - https://docs.mongodb.com/manual/ (papaja)
  - https://CRAN.R-project.org/package=dplyr
  - http://arxiv.org/abs/1403.2805 (mongolite)
  - https://CRAN.R-project.org/package=bindrcpp
  - https://CRAN.R-project.org/package=tibble
  - http://ggplot2.org

Q & A