Notes for Presentation:

* 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, touchdowns, 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 MongoDB after some serious issues whilst loading the data. We used many R packages to support our analysis. For data mining, we used the mongolite package to connect to the Bluemix data set and, where possible, we used MongoDB to do the statistical analysis in favor of R.​
* mongolite
  + We initially used MongoDB via 3T to load the data, run queries, and export said queries to CSV files. Later we realized we can execute MongoDB queries within RStudio via the Mongolite package. While the syntax of the queries slightly differs, we found this method to be more accessible and efficient to execute our project without violating the spirit of the project itself.
* Median yards per play
  + 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 of the teams – mostly between 5-6 yards per play. 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
  + Here we plotted the number of touchdowns against yards gained (per game). We expected to see a linear upwards trend 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.
* Td’s vs variance in yards
  + This plot shows that the relationship is relatively static, with a slight rise as the standard deviations go up. But again, the data are not normalized. 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.
* Yards vs variance of yards
  + 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 can get more yards with a higher standard deviation, 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.)
* Conclusion
  + The positive slope of the fit line in the StDev vs. Yards per Game plot is indicative that the surprising result in figure 3 -- that is, that *higher*, not lower, standard deviation equals more touchdowns -- could be explained by this strong increasing trend. In order to find out for sure, though, we would need a way to remove the impact of Yards per Game on Touchdowns per Game. Unfortunately, such an exercise is outside of the scope of this class.
  + To put it as simply as possible, our project is an observational study and we can’t design a study to remove the confoundedness because that would involve changing the nature of football, which is not feasible
  + Analyzing the relationship between standard deviation of yardage per play and touchdowns scored in the 2015 NFL season yielded some interesting results. While it seems that higher standard deviation tends to result in more yardage gained, the relationship is not so clear as it relates to touchdowns scored. Exactly What these results might mean to an offensive football strategist is debatable, however it is likely that one would conclude based on the analysis, that a mixed approach would lead to more yardage and more points. The high-level strategy is to carefully balance big plays with an overall conservative offense. When commencing this project, the assumed result was that lower standard deviation would result in higher touchdowns, but this relationship is currently inconclusive and warrants more analysis.
* Lessons Learned
  + At the onset of this project, our group anticipated using MySQL to generate queries that would be visualized in R Studio. This strategy reflected the SQL and R talents of the group members. While it was unfortunate that Bluemix MySQL could not correctly load the NFL data, this setback turned into a fantastic learning experience with MongoDB and R. MongoDB was able to load the NFL data set without incident and after a somewhat short learning curve, proved itself to be a rather effective query environment. Another benefit of using MongoDB was that it easily integrated with R by means of the Mongolite package. The combination of MongoDB and R proved very flexible with our large, highly dimensional data set. Using MongoDB with R should prove to be a powerful tool in future projects.