1. Problem:

The problem I am looking to solve is who are the best players to play from NFL week in and week out. There are tons of variables to sort through with this problem. I need to collect all the relevant data points and find out what enables me to make a good prediction, and what doesn’t

1. Data set:

The data set I am using is from a sports database website. I have a data structure that is extremely detailed and organized into a relational data table structure. These tables have every bit of information you can imagine for every play/game of NFL football dating back to 2000. Major table headings are:

Team, Player, Offense, Defense, Kicker, RedZone, Injury, Game, Drive, Play, and Play-by-Play

Every major stat is housed in a subtable that relates to one of the major table headings.

The major stuff that is not in the data (I believe) is the coaching staff. This will be critical for understanding play calling and personnel decisions. I can find this data, I just don’t have it right now.

Another major topic not in here is salary/bonuses. This data is hard to come by and may not be entirely relevant until a contract year.

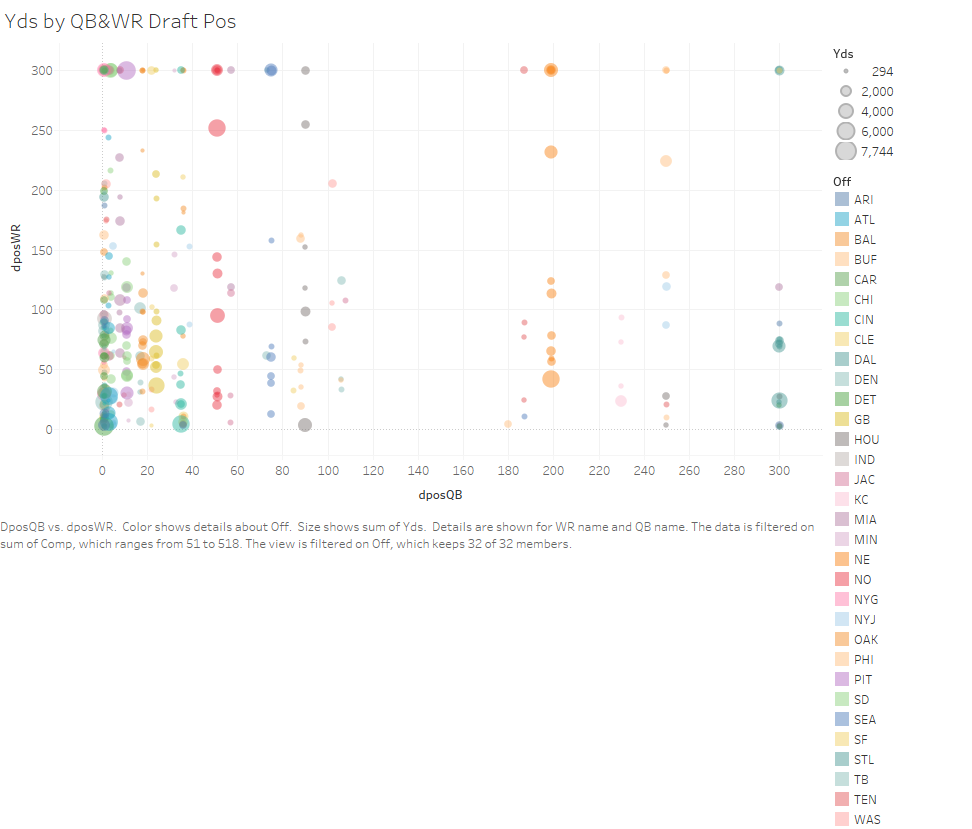
Rule changes are also not noted year by year. The game of football has changed dramatically over the past decade. The game used to be run heavy with stout defenses, to a pass happy league with limitations on how defenses can affect the offense.

Challenges with the data is it is large. I may need to only look back for a limited playing period. The data is pretty well organized, however, if I want to add in any more info (such as coaching) it will make the data messier.

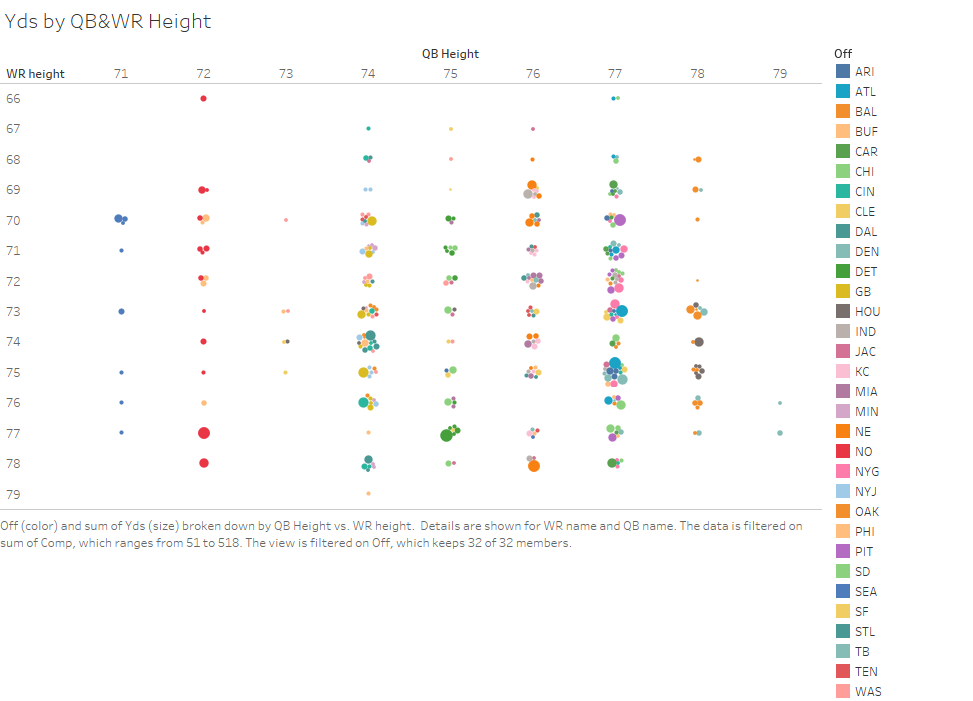
I have the data in a MySQL DB right now, and have been exploring it somewhat.

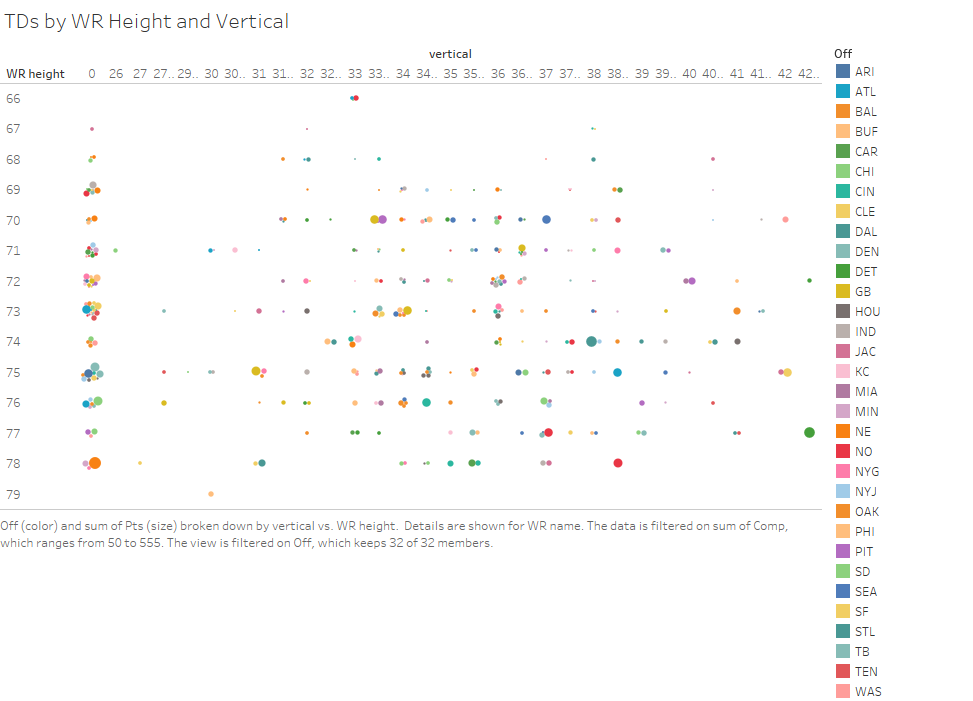
1. The exploration into the data has been purely to understand the basic info that is provided and see if there are any relationships. I have a solid background in Tableau, so I have wired the data into there for some visualizations and plan to migrate these into R.

Examples:

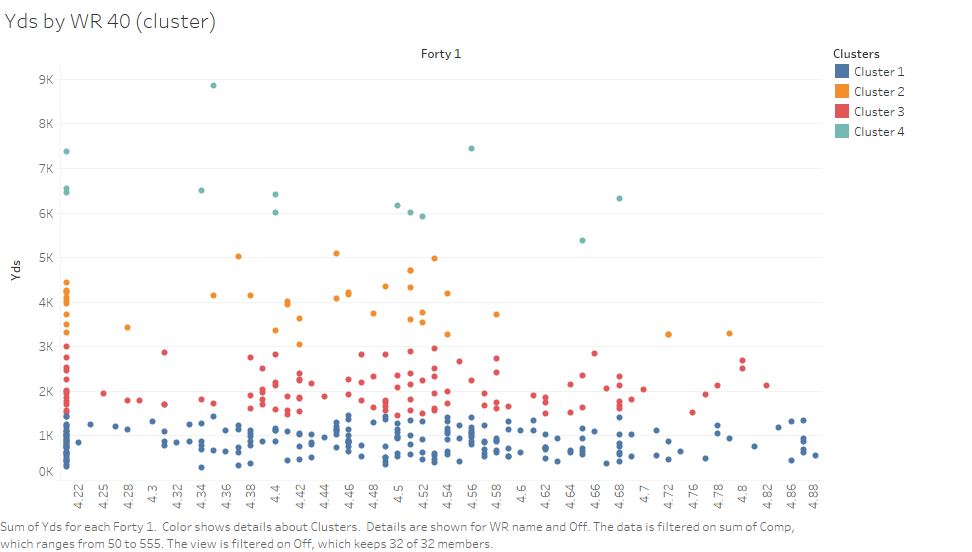


This chart shows the draft positions by QB on X-axis and draft position on Y-axis for WR. The size of the bubbles reflects the Yds accumulated by WR & QB. This is to see if there is any relation to draft position and production.



This view was created to get an idea of how height and vertical leap relate to touchdowns scored. This is on a cumulative level and not a per play or game level. So player who were in their prime during this time period may be over stated.

This shows TDs by WR height and vertical leap. The dimensions are discrete, so the bubbles are clusters at the point of the data. This is just to get a look at the relationship of physical talent and production.



This is a chart in that clusters WR yards. The X-axis is discrete with 40-yd dash times. This is on a cumulative level, so this does not take into account a per-game metric or per play metric.

1. My plan for going forward is to continue the course, learning about linear regression and machine learning. These will help shape my focus. I feel I will have to conduct separate exploratory analyses for each position to discover which variables are meaningful.